OULUN AMMATTIKORKEAKOULU

Markku Kaskenmaa

USING DATA ANALYTICS IN HOCKEY PLAYER TALENT IDENTIFICATION

Master's Thesis

USING DATA ANALYTICS IN HOCKEY PLAYER TALENT IDENTIFICATION

Master's Thesis

Markku Kaskenmaa
Master's Thesis
Spring 2023
Degree programme in Data Analytics
and Project Management
Oulu University of Applied Sciences

ABSTRACT

Oulu University of Applied Sciences
Degree programme in Data Analytics and Project Management

Author(s): Markku Kaskenmaa
Title of the thesis: Using Data Analytics In Hockey Player Talent Identification
Thesis examiner(s): Dr. Teppo Räisänen
Term and year of thesis completion: Spring 2023 Pages: 90 + 2 appendices

---

Goal for the research is to identify relevant data and features which can help the team to find rising future talent to build a data and analytics model and simulate the analysis with real data. The aim and main objectives are to create both data and statistical models for analysing a hockey game data in relation to the defined indicators for identifying better than usual player performance and to run the analysis with real life data.

Data analytics in professional sports can still be seen as a relatively new concept. In professional hockey leagues, like Liiga or NHL, a lot of data related to games are collected, analysed, and even made publicly available.

This research is conducted by analysing the previous research documentation, articles, and results, by collecting information from trusted web sites, blogs, and articles and using them to define the data, features, and statistical models, and building data model and data pipeline and finally executing an example analysis with the build models and collected real data. This research has quantitative and qualitative research parts. Qualitative parts goal is to define the theoretical framework and the quantitative part includes building a data pipeline and visualizations, and running the analysis based on the indicators and model built in the qualitative part.

There are numerous statistics and metrics which all try to explain player performance. Data is collected from games and made available via commercial or public channels, like Wisehockey for Finnish Elite Hockey League, Liiga. In analysing player performance, focus should be put in what happens when player is on ice. With analysing data from games and by using machine learning to detect anomalies in players performance can help teams in player recruitment decisions. Machine learning based predictions and analysing video data from games can help teams in future not only on recruitment but forming winning composition and identifying game patterns during game. The key to realizing these benefits is access to relevant and trustworthy data.

Future research should focus on using novel methods like machine learning to bring full benefits from data and analytics.

Keywords: Data Analytics, Ice Hockey, Data Analytics in Professional Sports, Machine Learning, Ice Hockey Statistics

CONTENTS

1 INTRODUCTION

This research is my master's thesis in OAMK's master's degree programme in data analytics and project management.

The purpose for this research is to help professional hockey team KalPa with their player recruitment for their Finnish Elite League (Liiga) team. Goal for the research is to identify relevant data and features which can help the team to find rising future talent to build a data and analytics model and simulate the analysis with real data. For the sports teams who operate under a moderate financial budget it is important to identify the future team players as early as possible and with cost efficiently.

In professional hockey leagues, like Liiga or NHL, a lot of data related to games are collected, analysed, and even made publicly available. However, the extent to the teams utilizes this data for the player recruitment process varies. In professional sports like soccer, American football and baseball, the data is used for many purposes. Within the NHL, the teams are using the collected data in multiple ways, for example as part of the team coaching and for identifying exceptional talent.

In this research the focus is on game data which is available through Liiga provided data sources and HockeyAllsvenskan league data, available either directly from the league database, or via other trusted public sources. Before starting the research, it has also been identified that in addition to analysing the game data, another important part of the player recruitment is their personal characteristics. These are left out by intention, and focus is put solely on the game data to ensure the research scope is manageable within the scope of the thesis.

This research is conducted by analysing the previous research documentation. These include articles, and results, by collecting information from trusted web sites, blogs, and articles and using them to define the data, features, and statistical models, by interviewing team coaches, scout and building data model and tools using Amazon Web Services and finally executing an example analysis with the build models and collected real data.

2 RESEARCH AIM AND OBJECTIVES

2.1 Aim and objectives

The aim and main objectives are to create both data and statistical models for analysing a hockey game data in relation to the defined indicators for identifying better than usual player performance. In addition to the creation of the models, the objective is also to run the analysis with real life data.

Secondary objectives are to identify relevant indicators which can be used to identify better than usual and exceptional performance of hockey players in their career by analysing the available game data.

2.2 Hypothesis and research questions

The hypothesis for the research is that by using commonly known ice hockey statistics to analyse game data and by using machine learning models, a professional hockey team's management, coaches and scouts can identify and predict players performance and use it to support their player recruitment decisions.

RQ1: How to identify players whose performance development is better than normal by using data and analytics?

Goal for this research question is to identify the indicators, statistical model and data which can identify better than usual and exceptional players (outliers). The players can be young rising talent up to players who have been playing in the different leagues, like Liiga or HockeyAllsvenskan for several seasons. Indicators, statistical and data models should not rely only on goals scored and assist to goal points.

RQ2: What advanced metrics and statistics models are common in hockey, and can they be used to analyse Liiga and HockeyAllsvenskan game data?

Goal for this research question is to identify what advanced metrics and models are common when analysing hockey data in professional leagues such as NHL and how well they fit for purpose in

analysing Liiga and HockeyAllsvenskan game data? What additional data would be needed to use these metrics and models?

RQ3: What data is available?

Goal for this research question is to identify what data is available for a Liiga hockey team and what are the trusted sources to obtain such data? The sources should be such from where the data can be retrieved constantly and automatically preferably via API and should not be only for one time purpose (e.g., by not using exported Excel file at time specific moment).

3 RESEARCH METHODOLOGY

This research is a mixed method research containing both quantitative and qualitative research parts. The research starts with qualitative literature-based research, of which the goal is to define the theoretical framework, to study the relevant previous research and their results to identify the indicators, and data and statistical models to analyse these indicators for identifying player performance. The sources for the qualitative part are previous academic research reports, articles, and results, and trusted publicly available blogs, websites, and articles. In this part I will research how the data is used for talent identification in other relevant professional sports and leagues in hockey (e.g., National Hockey League, NHL and Elitserien), soccer (e.g., Premier League), baseball (e.g., Major League Baseball, MLB) and American football (e.g., National Football Leagues, NFL) and if these are applicable to national league players in Finland.

The qualitative part can include semi structured interviews which aim is to verify the models fit for purpose. Interviewees can include coaches, team management, talent scouts, and other teams' data analysts (for example Nashville Predators data analysts).

The quantitative part includes building a data pipeline and visualizations, collecting the data, and running the analysis based on the indicators and model built in the qualitative part. Data analysis results and their relevance are then evaluated, and the model and methods can be adjusted as a result for future analysis. The data will be obtained from Liiga data sources (Wisehockey) and HockeyAllsvenskan league data by retrieving it directly from available API's. The pipeline, visualizations and analytics will be built using public cloud capabilities and services by using Microsoft Visual Studio Code with Python 3.9 for coding the data extraction, S3, Athena, Glue and QuickSight from Amazon Web Services for the analytics and by programming languages commonly used in data analytics, like Python and SQL.

It is important to note that Finnish national league, Liiga, and Swedish league HockeyAllsvenskan both collect and provide access to data and statistics from the games. However, the leagues in the scope for this study may not provide similar data and statistics used in the references in this theory part, the statistical and data models in this study must meet the data available from selected Liiga and HockeyAllsvenskan.

4 ORGANIZATION

The target organization for this research is KalPa Hockey (KalPa). KalPa is a sports team from Kuopio, Finland and it was founded in 1929. The team specialised in playing ice hockey in the 1950's. KalPa has played in the Finnish Elite League, Liiga since 2005 and in history KalPa has played also in lower divisions. KalPa Hockey Oy's turnover in 2020 was 6,7 million euros and in 2021 4,6 million euros. (KalPa Hockey)

KalPa is active in providing opportunities for both girls and boys in junior level players. Junior KalPa "aims also to develop players for the professional team. Its most important goal is to work closely with the community to make sure local children get a well-rounded start to life". "Junior KalPa has teams for every age group, ranging from the youngest juniors up to 20-year-old players. Junior KalPa also has skating and hockey schools for younger children. Junior KalPa has its own teams for girls and women. There are over 1,000 players and supervisors in Junior KalPa every year." (Junior KalPa)

Thesis researcher is Oulu University of Applied Sciences MBA student Markku Kaskenmaa and supervisor is Dr. Teppo Räisänen. KalPa supervisor is Anssi Laine.

5 THEORY

5.1 Use of data and analytics in professional sports

Today, analytics has become an important tool for any organization to be successful. In professional sports, analytics is often associated with player recruitment decisions. In a data-driven world, professional sport teams are using analytics to confirm or predict answers to questions, like ticket, player recruitment and performance, coaching, pricing and customer relationship. (Mondello, Kamke, 2014)

It's not just teams and the league that is using big data analytics. Sport betting companies have professional hockey as one of the top sports on offer and with in-depth data, they can determine the bets and winning multipliers more precisely. Game producers and game companies may use this data to design their sports games. (Julie, 2014)

Hockey and other professional sports are a huge entertainment business. The fans, media and other analysts can use the data to get and share insights for the player and teams' performance and the game details. There is an analytics business around these sports. Companies that provide data analysis are not just doing this for the fans. They sell their analytics to the teams, players, coaches, opponents and media and it is their business. Media companies like MTV Media, ESPN need the analytics that tracking game and player data provides. (Julie, 2014)

In hockey, the puck and players move on the rink with speed that makes manual statistics impossible to use. Sensors and The Internet of Things have changed data collection for many businesses, including professional sports. There can be sensors in the puck and all over hockey players which all can help to gain better performance overview which coaches and scouts can use to recognize talents under the radar. (Tran, 2021)

Data can be used in many ways in hockey. It can for example indicate fatigue, and coaches can react instantly by for example changing combinations or rest players in some games. Data analysis can bring tools for training and schematics for tactics, game plans and team play. In addition, data can give a more objective view in their development areas for seasoned players and rising young talent. (Hakes, Sauer 2006)

Data analytics in professional sports can still be seen as a relatively new concept. One of the most known stories is the Moneyball book and movie which is a story of how the Major League Baseball (MLB) team Oakland Athletics successfully overcome baseball's player market by using statistics and data. (Hakes, Sauer 2006)

Hakes and Sauer introduced and economic evaluation of the Moneyball story by applying standard econometric procedures to data on player compensation and productivity. Their findings were that specific baseball skills were not valued correctly, and that by using statistical knowledge team manager were successfully able to improve the player valuation. (Hakes, Sauer 2006)

5.2 Use of data and analytics in hockey player recruitment

One of the most important topics in sport analytics is the evaluation of player performance. Analytics can be used by coaches and team management for coaching, managing team, in scouting players, and in entertainment. (Vik, Min-Chun, Jansher, Carlsson, Lambrix, 2021)

Even though there have been major development efforts in past years, hockey still lacks many of the analytical data tools that are in use in other sports like in baseball, basketball, and football. (Tran, 2021)

Hockey teams can use statistics and data to evaluate their current players and use that knowledge then to future player decisions. By using analytics for objectively quantifying and analysing their teams and players contributions can reveal inconsistencies and further give indication where to allocate player salary budgets. Methods for player valuation can be unique to a sport and are not consistent across different sports. The primary means in nearly all professional sports for talent recruitment is through a draft. A draft is a process in which the professional teams take turns in selecting players from the pools of players. (Brower, 2020)

Talent selection and identification in professional sports is complex, and coaches and teams can utilize physiological and technical performance assessments to identify future top players. Identification of these future high valued players can be seen as one of a coach's key tasks and abilities. Talent identification has traditionally been based on viewing and analysing athletes in trial games or training sessions. This approach is based on a subjective view of the ideal player, relies

on viewers personal opinions and experience, and can include bias and limited consistency. (Larkin, O'Connor, 2017)

One of the main goals for talent identification and ice hockey player recruitment is to identify and select the most promising young players with the potential to become a successful professional player. In ice hockey, there are five main variables which influence what level a player can compete: technical, tactical, physical, social, and mental. To be a good player one must be faster, stronger, and smarter. (Kontsas, Lehtola, 2014)

As successful team sport players performance is a combination of multiple skills and qualities, and team coaches and scouts must consider a holistic approach for talent identification (Larkin, O'Connor, 2017). Player comparison requires deep domain knowledge including defining appropriate key metrics for players, and finding a group of players who have playing styles which can match together and make a winning team. Ice hockey leagues provide in-game data and to be agile on player markets teams should learn to use a data-driven approach to evaluate and predict the players performance. (Vik, Min-Chun, Jansher, Carlsson, Lambrix, 2021)

5.3 Using game data to analyze hockey player performance

There are multiple statistical methods that attempt to quantify hockey player's contribution and performance in relation to others. Some of these are more traditional metrics such as goals, assists, points and some are novel and more complex (Pallotta, 2021). These more novel and complex include for example Goals Versus Threshold developed by Tom Awad, Corsi rating, Behindthenet Rating by Gabriel Desjardin, Player Contribution by Alan Ryder, and Even-Strength Total Rating by Timo Seppä (Macdonald, 2011). Other examples of the common metrics are time on ice (TOI), the games played (GP), Shots on Goal, Expected Goals (Piccolo, 2022), Goals For per 60 and Goals Against per 60 (Pallotta, 2021).

Corsi is a statistic commonly used in ice hockey measures even-strength play shot attempt differential (Schulte, Zhao, 2017). NHL.com defines shot attempts as: "any time a player tries to shoot the puck" and "any shot on goal, blocked shot or missed shot is classified as a shot attempt" (Piccolo, 2022). Another commonly known metric is Fenwick in which only missed shots or shots on goal are counted and blocked shots excluded (Piccolo, 2022). Goals Versus Threshold, GVT,

is a statistic which tries to quantify the overall contribution to the game of a given player. This statistic is a combination of Offensive, Defensive, Goalkeeper and Shootout GVTs (Tran, 2021).

Such metrics have strengths and weaknesses. One weakness is that many of the statistics exclude the information about game context where it was measured and assigned (Schulte, Zhao, 2017). Another weakness was concluded by Schuckersa and Argeris in their research paper: “teams were not substantially better at picking optimal players with respect to TOI, GP, or GVT but were markedly better at selecting players within a half standard deviation of optimal”. (Tran, 2021)

To evaluate a players performance, it is more common to use traditional metrics that are related to players the actions, like using a sum of scored goals goal and assists. Some of the statistics used to evaluate a players performance try to consider the game context, like how the event impacted to the game result. A scored goal, when the team is in the clear multiple goal lead at the end of the game does not define the winner of the game. But if a player scores a goal only some seconds before end of the game in a situation where game is tied, many times defines winner. Some players tend to score these important goals than others, and other player may score more often these not so important goals. In example, Washington Capitals player Alex Ovechkin was ranked highest is making these important goals in the NHL season of 2013-2014. In scoring not so important goals his rank was 29th. The game context where the goal is scored should be considered when evaluating players' performance. (Vik, Min-Chun, Jansher, Carlsson, Lambrix, 2021)

Multiple studies exist in the context of using analytics in player evaluation in professional sports. Players are often rated by their observed performance over a series of games. There is an increasing trend for using data analytics and machine learning to model ice hockey. Teams, team management, coaches and fans are interested to evaluate players. For example, for the team management, questions like which players to sign, trade or draft are common. Multiple studies about different models exists, and the most common model is to quantify the total value of players based on their actions. Traditional sports analysis models usually and exclude the actions that lead up to the goal and their focus is on the actions that immediately impact to the scoring goals, as an example passes and pass reception. Values provided by these methods are limited. The relevant context may include for example time on ice, player role or position, current score, and game situation. (Liu, Schulte, 2018)

5.4 Use of machine learning to predict future performance

There are studies for novel methods like machine learning algorithms and models for player evaluation. Liu and Schulte in their study article “Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation” introduced a new method for evaluating game context by using Deep Reinforcement Learning (DRL). In their method they firs use learn an action-value Q function from game events. Then by using a puck possession based Long Short-Term Memory, the neural network integrates continuous history and game signals. Their method then uses the learned Q-function to valuate players actions also considering game contexts and evaluate a player performance. They call the valuation, which aggregates players actions and contribution, as Goal Impact Metric (GIM). They also found that a players GIM stays consistent throughout a season and has high correlation with future player salary and success measures. (Liu, Schulte, 2018)

5.5 Data analytics and analysis

To define the term “data analytics” in the context of this research its common definition will be used. Cambridge Dictionary specifies data analysis as “the process of examining information, especially using a computer, in order to find something out, or to help with making decisions.” (Cambridge Dictionary)

Data analysis is a process of exploring, refining, transforming, and training models with the chosen relevant data to gain useful information for decision making. Data analytics uses methods and tools like machine learning, statistical analysis, and pattern detection to gain insights from data. Its goal is to help organisations in decision making and problem solving by using historical data. Data analysis is often seen as a subset of data analytics. (Sarangam, 2022)

Google defines data analysis as a six phased process. The phases are:

- **Ask.** This first phase is about asking the right questions and understanding why analysis is done and what kind of problem it tries to solve.
- **Prepare.** The second phase is to prepare the data for analysis. This means collecting relevant data to the problem and to define metrics that are needed for the analysis.
- **Process.** This phase is about processing the data. Data is checked and corrected against various inaccuracies, errors, and inconsistencies so that it is consistent and is credible and trustworthy for the analysis.

- **Analyse**. In this phase data is analysed. Primary goals for this phase are finding the relationships, trends, and patterns that will help to solve the problem. This phase is about finding what the data is telling.
- **Share.** The fifth phase is about sharing findings. This can be done with help of visualization, storytelling and using narratives.
- **Act.** In this sixth phase, all things learned from analysis is put in action by providing recommendations to the stakeholder on problem solving and helping them make decisions. (Gautam, 2021)

5.6 Liiga in general

Finnish Elite League, The Liiga (formerly SM-Liiga), is the top professional ice hockey league in Finland. Liiga was founded in 1975 to replace the fundamentally amateur league SM-sarja. Liiga is and Finnish Ice Hockey Association have a cooperation agreement, but Liiga is not directly controlled by latter. (Wikipedia)

In a Liiga season consists of the regular season and playoffs there are currently 15 teams playing. All these teams play 60 matches during a regular season, a quadruple round robin with extra local double rounds. (Wikipedia)

The six highest scored teams per points of the regular season proceed directly to quarterfinals. Teams ranked between seven and ten play preliminary play-offs with best out of three games and the winners continue to quarterfinals. Quarterfinal winners meet at semi-finals and losers of the semi-finals play a bronze medal match and the two winner teams play a gold medal match. In the regular season 2021-2022 there were 632 players in total playing in Liiga. (Wikipedia)

5.7 HockeyAllSvenskan in general

HockeyAllsvenskan is a second highest professional league (previously Allsvenskan and SuperAllsvenskan) in the Swedish ice hockey system (after the SHL). Since the 2009–10 season, the league has consisted of 14 teams. In a regular season, all teams play 52 games, and all teams play each other four times during one season, two home games and two visiting other team. Following the regular season Slutspelserie tournament is played which finals winner wins the championships (Wikipedia)

In the regular season 2021-2022 there were more than 650 players in total playing in HockeyAllsvenskan. (Wikipedia)

5.8 Traditional hockey statistics, metrics, and their meaning

This chapter introduces some known traditional statistics and methods which can be used to explain and evaluate ice hockey players performance.

5.8.1 Plus/Minus

Plus/minus purpose is to measures what effect player has on goals of his team when he is on ice. When plus/minus value is positive for a player it means that the player has been on ice for more goals scored than against, and if it plus/minus is negative, it means player has been on ice for more against. Plus/minus calculates all even strength and short-handed goals, including those where goalkeeper has been pulled from the rink. (Kohl, 2016)

Plus/minus calculation is based on:

- A plus point is given when a player is on the ice and his team scores.
- A minus point is given when a player is on the ice and opponent scores a goal.
- No points are given when a power play goal is scored.
- Players who are on the ice for an empty net goal will receive a plus and a minus unless the team who scores the empty net goal is on a power play. (Jones)

Plus/Minus can be seen as misleading statistics if used alone for evaluating players. This is because the individual player has only little control over their own plus-minus. Players can make mistakes and they can also excel. Ice hockey is a game which includes randomness. This can mean that goals are scored, and a player could have had nothing to do with it, and a player is still given a plus or minus point. Another example of lack of context plus/minus doesn't consider goalkeeper skills. Additionally, there is variation in how many goals a team scores during a season. This means that one player's plus/minus does not necessarily compare to another team's player and that plus/minus requires normalization or to be used in conjunction with other metrics. (Thomas, 2018)

Plus/Minus is available through both data sources and can be used in the analysis part of this thesis.

5.8.2 Time on Ice

Time on ice (TOI), or ice time, is a metric that indicates a player's total ice time over the season or in a single game. For player performance analytics purposes, it is mainly used for normalizing other metrics. (Sans, Carlsson, Lambrix, 2019)

Statistic is available through both data sources and can be used in the analysis part of this thesis.

5.8.3 Points

Points is a common metric summarizing simply goal leading assists and goals a player has performed in a single game, over the season or over his career. (Sans, Carlsson, Lambrix, 2019)

Statistic is available through both data sources and can be used in the analysis part of this thesis.

5.8.4 Goals

Goals is a common metric summarizing goals a player made, in a single game, over the season or over his career. (Sans, Carlsson, Lambrix, 2019)

Statistic is available through both data sources and can be used in the analysis part of this thesis.

5.8.5 Games played

Games played, GP, is a common statistic collected in hockey and it summarizes the number of games where a player has been on ice over a time. Time is often calculated as a single season or over a player's career. (Sans, Carlsson, Lambrix, 2019)

This statistic is available through both data sources and can be used in the analysis part of this thesis.

5.9 Advanced hockey statistics, metrics, and their meaning

This chapter introduces some known advanced statistics and methods which try to explain and evaluate ice hockey players performance. It should be noted that many of the advanced statistics are intended to be used with the data from the National Hockey League, NHL.

5.9.1 Corsi

Corsi, named after Jim Corsi, a former professional goalkeeper, can be seen as one of the foundational concepts in hockey statistics. Corsi is based on shot attempts, and considers any shot attempts, not only shots on goal. This means that any shot which are directed towards the goal is counted, whether it reaches the goalkeeper or not, or is blocked. A positive Corsi for a team indicates they took more shots than the opponent team, which can indicate more offensive pressure and possession. When Corsi is reported on a player it simply indicates how many shot attempts were made when a given player was on ice and can indicate players contribution to the game result. (Lee, 2020)

The Corsi formulas are as follows:

$$Corsi\ (For\ or\ Against) = \text{Shot Attempts For or Shot Attempts Against}$$

$$Corsi\ Percentage\ (C\%) = \frac{\text{Corsi For (CF)}}{Corsi\ Against\ (CA)}$$

As Corsi considers all shot attempts, Corsi For percentage (CF%), which averages over total shot attempts, can indicate a player's performance. CF% is calculated as

$$CF\% = \frac{\text{CF}}{CF\ +\ CA}$$

If players CF% is around 50% this can indicate that a player is effective on creating offensive threat to the opponent but not very active on defence. CF% more than 50 can indicate that a player contributes equally to offense and defence, or is dominant in one, and good enough in the other. A relative metric, Corsi For %, is needed to compare a player to a reference. A higher CF% may can identify more offensive pressure and can lead to more goals. (Lee, 2020) In Corsi, this is CF% relative and is calculated as

$$CF\%\ rel = \frac{\text{CF\%}}{CFoff\%}$$

where CFoff% means the team metric where a player was not on ice. CF% rel can identify important players on a team, or a player to pair with a good scoring player. (Lee, 2020)

Corsi compares only even-strength play situations and does not consider game context, situation, nor the shot attempt quality. Even very low-quality shots are calculated. An underachieving player's Corsi may be affected by better performing players who are on ice with that player. CF% used with average point total could still indicate if a given player is creating opportunities for his team. (Lee, 2020)

Corsi is available through both data sources and can be used in the analysis part of this thesis.

5.9.2 Fenwick

Fenwick also counts any shot that is made on goal during in even-strength game situation. The difference to Corsi is, that it excludes blocked shots. The formula to calculate Fenwick is similar to Corsi. (Lee, 2020)

When measuring Fenwick for a team's it is often presented as a percentage of the total Fenwick of the game. If a Fenwick For % (FF%) is above 50% it indicates that the team made more unblocked shot attempts against goal than their opponent team. Fenwick, when applied to individual players, compares FF% on the ice and FF% when the player is off the ice. A player who is active on offense and is effective in defence is indicated by higher FF% and can indicate a player contributes more when he is on the ice and thus is a potentially high performing player. (Lee, 2020)

Evenly distributed blocked shots between opposing teams is seen as equal Corsi For % and Fenwick For %. Potential issue with Fenwick is that some players tend to block more shots than others. With such players, this can be seen as significant differences in their CF% and FF%. This is related to a player's role; some are playing in defensive roles and as result are blocking more shots. Fenwick counts any shot that make it through the defence and differs from shots on goal (SOG), which counts only shots that make their way to the goalie. Another problem with Fenwick

is that a blocked shot could still be a result of a goal opportunity, and it excludes the total the contribution of a player. (Lee, 2020)

Fenwick is available through both data sources and can be used in the analysis part of this thesis.

5.9.3 PDO

PDO is not an acronym. Luck has an important role on the results in ice hockey and PDO can be said to be a statistic that measures luck. Luck in this context means game results that are outside normal boundaries and variance of the player's normal performance. PDO can be calculated both on a team level and for individual players. (Lee, 2021)

A player cannot maintain a long period of shot percentage that is multiple standard deviations higher or lower than the mean nor any goalkeeper can stop every shot. A better or lower performance for a player can be also referred to as luck. Due to randomness of the game and low number of goals, ice hockey can be said to be more affected by luck than some other sports. Situations caused by randomness during a game can led to a goal and cause the weaker team to win a game. (Lee, 2021)

PDO is the sum of a team's shooting percentage (S%) and its save percentage (Sv%) during even strength play. The combined PDO of all teams in the league is 100%. PDO can be used to evaluate if a team should expect a regression or improvement (Wikipedia). Luckier teams PDO is higher than 100% and a PDO less than 100% means that the team has been unlucky or has been performing below its expected skill level. Over the long term, the luck can be estimated to approach the average (Weissbock, Herna, Inkpen). PDOs tend to approach 100% is often discussed as PDOs regression. The mean is 100% because with PDO any shot on goal result is either a goal or a save. Therefore, a games S% and Sv% will always be 100. Summing PDO over a season will also give 100%. (Lee, 2021)

PDO is often normalized by multiplying it by 10. PDO formulas are:

$$PDO = \frac{\mathrm{S\%}}{Sv\%}$$

where

$$S\% = \frac{\mathrm{Goals}}{Shots\ on\ Goal}$$

and

$$Sv\% = \frac{\text{Saves}}{Shots\ on\ Goal}$$

When evaluating an individual player, PDO can be an important measure as it indicates what happens when a given player is on the ice. (Lee, 2021)

PDO is available through Wisehockey API and can be used in the analysis part of this thesis.

5.9.4 Goals Versus Treshold

Goals Versus Threshold (GVT) is a statistic which uses goals added and aims to evaluate a player's total contribution to the game. GVT is a combination of Offensive and Defensive GVTs. (Turtoro, 2014)

GVT considers both assists and goals. The threshold is a number calculated compared to a threshold player. For example, a player like Oliver Kapanen leaves KalPa for the rest of the season. The questions then are for example who is going to have to replace him on the team? What is the difference between having Kapanen on the team and not having Kapanen on the team? How much can the replacement players cover losing Kapanen? When a team considers whether to replace a player, there are at least two ways for replacement: 1) replacing the player from inside own team or 2) contracting a player outside the team for example other same league teams or from another league (bgred105, 2012). GVT aims to quantify answers to these questions and to understand the value of a player like Kapanen or any other player's GVT. (CLVNNG, 2017)

GVT was created to 1) Compare various players in the NHL, independent of position and 2) provide a way to quantify the importance of defence. (bgred105, 2012)

Calculating GVT uses complex formulas and requires estimation of assists value versus goals value. Calculating GVT requires data how much a player has been on ice and identifying and calculating a threshold value for possible threshold players. GVT is considering both offensive and defensive metrics. (Turtoro, 2014)

According to Turtoro, "The offensive GVT calculates the total value of a player's points and subtracts the expected point value of a threshold player resulting in the total goals added by the player." (Turtoro, 2014)

The defensive GVT is a sum of 1) Goals prevented by preventing shots, 2) plus/minus adjusted to team and position, and 3) position adjusted players contribution to the goalkeeper GVT. Defensive GVT shows the prevented expected goals. The total GVT is calculated by adding Offensive GVT and Defensive GVT together and adding goalkeeper GVT and shootout GVT to the equation. (CLVNNG, 2017)

According to Tom Awad, the five key characteristics of GVT are that it

1. is measured in goals.
2. compares all players over any period.
3. uses only directly goal leading statistics.
4. has built-in accounting.
5. normalises for the strength of the league. (Yost, 2012)

Since GVT is meant to calculate statistics for the NHL players and require specific adjusted parameters, it is not directly applicable for Liiga on HockeyAllSvenskan players. And, that it is more to explain a replacement value of a player than identifying rising talent. It would however be interesting to study this more in the context of Liiga, to try to find out if and with what changes it applies. GVT is not discussed further in this thesis.

5.9.5 Point Shares

Point shares is a statistic originating based on Win Shares in baseball by Bill James and was developed by sports analyst Justin Kubatko in 2011. Aim for Win Shares is to estimate the number of wins a player has been able to create and it originates from basketball. Kubatko decided to apply the same idea to ice hockey after he found the statistic to be working for baseball and basketball. Point share tries to explain players performance at a time in each context. ((Kubatko, 2011)

In Point Shares, marginal goals for (MGF) and marginal goals against (MGA) are linked to teams' points. Team level MGF and MGA are calculated with:

$$MGF = \text{Team Goals} \times \frac{7}{12} \times Team\ Games \times League\ Goals\ Per\ Game$$

$$MGA = \left(1 + \frac{7}{12}\right) \times Team\ Games \times League\ Goals\ Per\ Game - Team\ Goals\ Against$$

At even strength, a team has five skaters and one goalkeeper on ice at a time, six in total. Each of these players contribute to offense and defence. Skaters contribute five parts to the offense and five parts to the defence and goalkeepers both parts will contribute to defence. This gives us in total 12 parts, and 7 of them are defensive. Thus 7/12. (Kubatko, 2011)

Expected points are then calculated from marginal goals for and marginal goals against with:

$$Expected\ Points = \text{League Points Per Goals} \times (\text{MGF} + \text{MGA})$$

Offensive Point Shares are then assigned to players based on goals and with time on ice or games played. Defensive Point Shares are calculated with:

$$Defensive\ Point\ Shares = \frac{\text{Marginal Goals Against}}{Marginal\ Goals\ Per\ Point}$$

Goalkeeper gets 2/7 of the team's defensive point shares. Number 2/7 comes from that there are in total 7 defensive points with 6 players who are on ice at a time and goalkeeper gets two of them. (Kubatko, 2011)

By summing Offensive Point Shares, Defensive Point Shares, and Goalie Point Shares we get the final Point Shares for each player (Kubatko, 2011). This statistic could be suitable for Liiga and HockeyAllSvenskan players, but requires some calculations with data, and is dependent on certain features on the data. As also this statistic is by default part of the datasets offered by neither of the sources I use for this thesis, it is not discussed during the analysis phase. Calculating this statistic would require extra coding and calculations, and as its value is unclear for Liiga and HockeyAllSvenskan, it is not discussed further in this thesis.

5.9.6 Expected goals

Low scoring rate makes analysing team and player performance in ice hockey difficult. This is particularly present when less than a full season's data is used for the analysis. In addition to low scoring rates, the randomness makes evaluating and predicting performance of teams and players harder. (Lukas, 2021)

Expected Goal (xG) is a measure which considers that each shot has a like hood of becoming a goal and not all shots are equal. (Liu, Schulte, 2018). It uses multiple different factors, like shooting position, and game situation, and calculates mathematically the like hood of a shot attempt becoming a goal. xG can be presented as a percentage or directly representing a value of the like hood. xG can also be seen as a measure of shot quality and can present total shot quality a player makes, thus can indicate players contribution to the game. (MacDonald, 2012.)

There are many uses for an expected goals statistic, and it can be used to evaluate players. As an example, expected goals may be used to evaluate offensive and defensive performance of an ice hockey team over a time, or it can be used to evaluate a player's contribution to a single game. (MacDonald, 2012.)

Expected Goals For (xGF) can be used to calculate what is the shot quality of a players team and Expected Goals Against (xGA) what is the opposite teams shot quality when a specific player is on the ice. When xG is normalized with time on ice, it tries to explain the players contribution when he is on ice (Lukas, 2021). Another way to use xGF and xGA is to evaluate the overall shot quality percentage a player has generated. An xGF% higher than 50% can indicate that a player is contributing to create an offensive advantage for their team. (Lukas, 2021)

Expected goals were also highlighted as an important metric by the data analyst of Nashville Predators. Wisehockey calculates expected goals value for each shot recorded and this statistic can be obtained via API thus can be used in the analysis phase of this thesis. However, this statistic is not available from SportContract API.

5.9.7 Zone starts

Zone starts is the number of faceoffs in each end of the rink. It considers the total number of faceoffs, but not the number of faceoffs wins. (Lee, 2021)

Offensive zone faceoff happens when for example the opponents goalkeeper covers the puck, or the opponent gets a penalty. In theory, zone starts can be a measure of offensive pressure and can also indicate player deployment and performance. The problem decreasing the value of this statistic is that an offensive, or defensive faceoff may be awarded to a player even if he had no part how the puck got into the zone. (Lee, 2021)

Christian Lee in his Hockey Stats article wrote that his analysis showed that during NHL season 2019-2020 the distribution of player zone starts is left-skewed with a mean just below 50 and that most players had an offensive zone starts between 35-63%. His analysis revealed a positive correlation between higher offensive zone starts and higher points per game averages. This suggests that more offensive faceoff wins lead to more pressure finally leading to more shots and scoring opportunities. He also concluded that the relationship is much weaker with defensemen because their primary role is preventing the goals and not scoring. (Lee, 2021)

When used as single measure to evaluate players performance, zone starts is not very informative and it should be incorporated with additional variables, like faceoff wins and other factors that can indicate if an offensive faceoff can result as a goal (Lee, 2021). Zone starts is not available from neither of the API's I used as data sources, thus will not be discussed further in this thesis.

5.9.8 Defensive Efficiency Metrics

Many of the available statistics and methods try to explain offensive performance, but a question remains that how to measure defensive performance of a team and a player? To answer to requirement, Raber and Eisenberg introduced a set of efficiency metrics called Defensive Efficiency Metrics (DEMs) that aims to draw a comprehensive picture of a team's performance. (Raber, Eisenberg)
DEMs aim is to indicate how efficient a given team is at defending against following five dangerous possession-types:

1. Defensive zone start

2. Controlled zone entries
3. Transition chance
4. East-West slot pas, and
5. Penalties.

Raber's and Eisenberg's research indicates that if a team is allowing 20 transition chances in 100 possessions, they have an 80% transition prevention score. If the team allows 5 goals on those 20 possessions, they have a 25% transition defence score. Their method used decision tree analysis to find the relationships between the five possession types and found out that and that relation with each of the five possession types were statistically significant and that there is a high correlation with preventing goals. (Raber, Eisenberg)

DEMs is a team level measure and can provide insight into team defensive performance, strengths, and weaknesses. Decision makers can use DEMs to identify which situations a team struggles in defending and use that information in recruiting players that can positively contribute to those. (Raber, Eisenberg)

DEMs, however, are targeted to highlight teams' defensive capabilities, thus are not very effective in predicting a player's career or performance. DEM is not available via the data sources used in this thesis thus is not discussed further.

5.9.9 Game Impact Metric

Game Impact Metric, GIM, were introduced by Guiliang Liu and Oliver Schulte in their research paper "Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation" in 2018. They introduced a new approach for capturing game context with applying Deep Reinforcement Learning (DRL) to learn an action-value Q function from game events. Their method uses a neural network representation which integrates both continuous context signals and game history, using a possession-based Long Short-Term Memory, LSTM. Then the method uses the learned Q-function to valuate players' actions under different game contexts and to evaluate a player's overall performance. (Liu, Schulte, 2018)

Their method introduces a Goal Impact Metric (GIM) which aggregates the values of the player's actions, reflecting other actions than goals. GIM is possession based instead of goal based. They

also found out that GIM can be used to also identify undervalued players, GIM is consistent throughout a season, and that it has high correlation with success measures and future player salary. (Liu, Schulte, 2018)

The method uses a computer vision provided dataset which provides very detailed level information, such as coordinates, about game events and player actions over the entire season (Liu, Schulte, 2018). This may limit its usage on Liiga and HockeyAllsvenskan and GIM is not available in either of the data sources, and calculation would require a lot of effort. However, this method supports my findings later in this thesis about using machine learning for predictions and detecting anomalies.

5.9.10 Players streak duration

In ice hockey, a player's performance varies over the season, and between seasons. Periods where in consecutive games a player makes goals or assists and is assigned points can be called streaks. When streaks are used as a metric, it can indicate players that are having better or less successful periods. Aa a reference to identify a streak (against his own team), for example points from past five games can be used. A streak is a sequence of games over series of games where the player plays during a season, and where the player plays better than usually (hot streak) or below expectations (cold streak). (Fuentes, Carlsson, Lambrix, 2019)

As discussed earlier, in ice hockey goals are rare which means also that long point streaks are rare. Streaks as a metric can only indicate offensive capabilities of a player. To overcome these problems, two other metrics can be used: 1) direct impact, and 2) on-ice impact. (Fuentes, Carlsson, Lambrix, 2019)

Fuentes, Carlsson, and Lambrix present in their research paper: “Player impact measures for scoring in ice hockey” that “While the on-ice metrics results in slightly longer streaks, all four curves show clear straight-line behaviour on lin-log scale suggesting that hot-streak durations when using an individual threshold is exponentially distributed. This itself suggests that hot streaks, when assessed relative to the players’ average performance over a season, may be memoryless and recent performance history (including longer streaks) may not add value compared to just reporting the average performance over the entire season.” (Fuentes, Carlsson, Lambrix, 2019)

This means that a hot streak can primarily indicate a good player at a given time but does not indicate how a player will perform over a season. Fuentes, Carlsson, Lambrix research notes that good players are more likely to have longer hot streaks, and that long hot streaks are more frequent with high salaried players. (Fuentes, Carlsson, Lambrix, 2019)

As a conclusion, player hot-streak duration can help predict the performance in the short term, but not over the season or multiple seasons. Streak duration can be calculated, but as it is not readily available in the datasets. This however is an interesting metric for machine learning based anomaly detection, in other words finding hot and cold streaks in players time series data.

5.9.11 Total Hockey Rating

Total Hockey Ratings (THoR) is a statistic to be used for Forwards and Defensemen in NHL. It was created by Michael Schuckers and Jim Curro in 2013.

THoR uses a statistical model to evaluate players' overall contribution while on the ice. It considers players all on-ice actions as well as quality of their teammates (QoT), their opponents quality (QoC), the current score of the game (Score Effects), and where player starts their shift (Zone Starts). Each game event gets assigned a value which indicates a change that it leads to a goal. (Schuckers, Curro, 2013)

Basis for THoR is data from NHL Real Time Scoring System (RTSS). RTSS records all events that occur in every NHL game, and it uses the Play by Play (PBP) files from every game to obtain the on-ice action events:

- faceoff (FAC)
- hit (HIT)
- giveaway (GIVE)
- takeaway (TAKE)
- blocked shot (BLOCK)
- missed shot (MISS)
- shot on goal (SHOT)
- goal (GOAL)
- penalty (PENL)

With using RTSS, the location for each event is known. Zone information (offensive, neutral, defensive) where the event occurred, is used for HIT, FAC, TAKE, GIVE, MISS, BLOCK or PENL events and the x and y coordinates for SHOT and GOAL events. (Schuckers, Curro, 2013)

THoR aims to create a rating system for NHL forwards and defensemen by valuating them with both scoring goals and preventing goals. THoR uses a data model described above and is then fitted using ridge regression. A high THoR means a player is contributing more to the events that lead to goals. THoR handles the issue in evaluation when the performance change if a player changes a team. (Schuckers, Curro, 2013)

THoR's advantages include that it evaluates players for every event that happens when a player is on the ice, and it aims to highlight good performance on both offense and defence. To overcome the randomness of a game, the model uses two regular season's data. It adjusts for the quality of other players on the ice and values each event for the impact it has on scoring a goal. (Schuckers, Curro, 2013)

In their research paper Schuckers and Curro calculated THoR for the top 15 forwards and the top 15 defensemen. Results of their research indicate that THoR gives a complete evaluation of a given player. (Schuckers, Curro, 2013)

THoR however is again one advanced metric aimed to evaluate NHL players and requires data from the NHL RTSS system. Thus, its usability to evaluate Liiga and HockeyAllsvenskan player is limited. While THoR materials and formulas seem to be still available, it is not widely used.

5.9.12 dCorsi

A study made by Stephen Burst indicated that Corsi For and Corsi Against have a weak correlation with Pearson's R correlation value of -0.13, with a coefficient of determination of 0.02. As a result, he developed in 2014 Delta Corsi, in short, dCorsi which is the differential between a skater's expected and observed Corsi values and it aims to evaluate both offensive and defensive performance for a player (Burtch, 2014). Corsi was discussed earlier in this thesis.

In theory, teams that maintain a greater share of puck possession, are more likely to make goals and as a result win games. dCorsi uses expected Corsi For and expected Corsi Against to evaluate how a player impacts on shot metrics. dCorsi is the residual differential of the expected Corsi For and actual Corsi For and the expected Corsi Against and actual Corsi Against. (Burtch, 2014).

dCorsi uses then a multi-variate linear regression to predict what Corsi results are expected for a given player based on their usage. However, an individual player has only limited control on their time on and because of this the actual usage is determined through weighting several factors that may impact players Corsi results. The factors include for example age, position, team, time on ice, teammates Corsi For and Corsi Against, offensive and defensive faceoffs and faceoff win percentage.

The model then calculates delta Corsi For per 20 minutes (dCF20) and delta Corsi Against per 20 minutes (dCA20). Once expected CF and expected CA are calculated the two values can be compared to the players' observed Corsi results to obtain dCorsi For and dCorsi Against. Finally, to calculate a player's dCorsi over a given season, the expected Corsi is determined by summing dCorsi For and dCorsi Against. (Burtch, 2014)

Values of dCorsi are nearly normally distributed, which means most players dCorsi values are close to zero. The better and worse player can be identified by looking at the players who are at least one sigma away from the centre. (Burtch, 2014)

Stephen Burtch verified the developed metric with selecting 100 player randomly over 7-years population of NHL players. His studies shows that the model is effective and that there is a strong correlation between expected Corsi and actual measured Corsi ($r = 0.7854$ and $r2 = 0.6168$). (Burtch, 2014)

dCorsi is again one of the statistics which is complex to calculate and requires specific datasets. dCorsi is not available through the datasets used for this thesis, nor the dataset containing all needed parameters to calculate it. For this reason, it is not further discussed in this thesis.

5.9.13 Other advanced statistics

In addition to the above statistics, with searching web with Google searches that there are multiple additional statistics which all try to explain players and/or teams' performance and value. Examples of these are Wins-Above-Replacement (WAR) and Regularized Adjusted Plus/Minus, RAPM. Both being targeted to evaluate NHL players and require NHL RTSS data, or similar dataset. Thus, they cannot directly be applied to Liiga or HockeyAllSvenskan players and are not discussed in the thesis.

5.10 Evaluating players

5.10.1 Evaluation of players performance

Rob Found in his 2016 research paper suggests that an optimal strategy for evaluating players performance may be a combination of a player's ability to produce high quality shots and a player's net production of goals. He compares multiple methods to analyse and predict success and his results suggest that when comparing approaches to evaluating team success, and individual contributions to that success, no single metric exists. Goals come from shots, shots can only follow puck possession, and possession is a result of winning faceoffs, getting power plays, and forcing turnovers. (Found, 2016)

Many of the metrics are evaluating team success. For forward players an approach could be to predict team and individual success with goal-based metrics. Neither goal nor shot based metrics are alone effective in evaluating a defenseman. For players in defence roles, statistics such as defensive Efficiency Metric which consider all possessions, resulting in larger samples, can be more effective. (Raber, Eisenberg)

5.10.2 Projecting expected player value

The central element of a hockey team building is to try to measure every player's value with a single number taking into consideration the player salary budget and try to project it to the several seasons ahead. (Vollman, Fyffe, Awad, 2016)

The previously discussed statistics measure may be used to evaluate players' value based on how good the player was in the history, but to predict players future success and value to his team, a model which converts these statistics into players expected future performance in current season and in players remaining career is needed. (Vollman, Fyffe, Awad, 2016)

To predict a future performance of a player, a three-step method, The Marcel Method introduced by hockey analyst Tom Tango in 2005 can be used:

1. Use weighted average of player's three most recent seasons.
2. Use a regression of a players performance against league mean in played games.
3. Adjust against age to differentiate between developing and seasoned players. (Vollman, Fyffe, Awad, 2016)

Random variance in hockey game analytics can be estimated by using for example a simple statistics method by dividing the available data into two and checking how the halves correlate with each other. (Vollman, Fyffe, Awad, 2016)

With this method,

- if there is no correlation, then event is completely random.
- if the correlation is weak, assume that the future performance is closer to the league average than to the actual observed results.
- if the correlation is strong, assume that the actual performance is closer to the observed values than league average.
- perfect correlation means there is no random variance and observed results can be used without any regression. (Vollman, Fyffe, Awad, 2016)

The possibility to use machine learning for predictions exists. This could be for example applied so that several past seasons data is used to train a model and use the current season to create predictions. This approach is discussed later in this thesis.

5.10.3 Laws of ice hockey analytics

In hockey analytics as well as in any business analytics it important to consider various aspects before making any analysis or decisions based on the results. Alan Ryder wrote his article "The

Ten Laws of Hockey Analytics" in 2008. According to his article the "laws" of ice hockey analytics are: (Ryder, 2008)

1. Winning a game matter.

Winning games is all that matters. (Ryder, 2008)

2. Goals for and against affect winning.

A team that scores the most goals wins in a single game wins the game. The team which has a positive average goal differential over the season tends to win more games than it loses. (Ryder, 2008)

3. Randomness is everywhere.

Statical proof exist that randomness is everywhere and goals in ice hockey occur randomly. It means that outcomes do not occur pure by luck, they are uncertain and are influenced of multiple factors such as skill, strategy, and execution. (Ryder, 2008)

4. Winning has a nearly linear relationship to goal differential.

There is a very high correlation between goal differentials and winning the game. Teams that are underperform in a prediction based on goal differential are better and the ones that outperform are worse. Thus, simplified model for winning is:

$$Winning\ \% = \frac{1}{2} + \frac{(goals\ for - goals\ against)}{(total\ goals\ per\ game\ \times\ games\ played)}$$

This can be called "linearity law". It means that game is the sum of its parts, both scored goals, saved goals impact the results and the team's performance is a sum of its individual players' performance. (Ryder, 2008)

5. Sample size matters

When evaluating players using data, the relation in the data amounts is important to understand. If a team plays 82 games in an average NHL season and wins 41 of those games, participates in about 450 goals, 5000 shot-on-goals and 9000 shot-at-goals there is 95% confidence that the team is neither better than 60% or worse than 40%. The statistical information from a sample is proportionate to the square root of the sample size. For one game

outcome there are 2.3 goals, 7.8 shot-on-goals and 10.6 shot-at-goal events; information from eight games of shot-at-goal events is equal to 35 games goals or a full season of wins and losses. In one season a goalkeeper can face 2000 shots and a good forward can make 300 shots. In general, there is more data about goalkeeping than shooting goals. (Ryder, 2008)

6. Expect mean reversion.

Research exists which show that a good luck is indicated by PDO a lot higher than 1 can and bad luck with PDO much less than 1. PDO values are almost normally distributed, and it is one of the easiest predictions in hockey is to find PDO outliers. (Ryder, 2008)

7. Respect the data.

Correct data must be used to answer to the questions. The only unambiguous event in a hockey game is a goal. Shot information may contain errors or bias and ice time may be inaccurate or wrong if recorded by humans. Random errors in the data can average out as measurement noise. Systematic errors or biases should be identified and corrected before data is used in analysis. (Ryder, 2008)

8. Hockey is a team sport.

Team performance is a sum of Individual players performance and there is a much greater variance in individual performance than in team performance. From the team perspective, only thing that matters are the shots that ends as goal. This suggests looking towards statistics like plus/minus. (Ryder, 2008)

9. Puck possession matters.

Analytic efforts should be focused on the shot data. Previous analysis has shown that historic shot at goal data predicts future winning better than historic winning or goal scoring. (Ryder, 2008)

10. Context matters.

In general, it can be assumed that a player who is on more ice does more. Players' time on ice on a game should be adjusted to the game time context (per 60 minutes). Blocking shot attempts during penalty killing and in power play situation getting shot opportunities are the key tasks. The roles of players vary, and the data for these roles look different. Some players move puck on the ice, and some does the shooting goals. A metrics of wingman of a high scoring

player different than a wingman of a different role. Generally, making goals against highly skilled team is more difficult than against lower skill level team, and the metrics from these games look different. (Ryder, 2008)

These rules can be concluded:

- When evaluating player performance using game data, the player's role should be considered.
- Only thing that matter in winning are the goals.
- Puck possession can indicate goal opportunities.
- To predict the future, a large enough data set is needed.
- Understanding the data is crucial for the quality of the analysis.
- Identifying outliers is important. (Ryder, 2008)

6 DATA ANALYSIS

6.1 Data sources

I used two data sources to obtain match level and season level data: 1) Wisehockey for Liiga game level data and 2) SportContract for HockeyAllsvenskan data and full season statistics for Liiga.

6.1.1 Liiga matchlevel data, Wisehockey

As discussed earlier in this thesis, there are numerous metrics and methods which all try to indicate players' performance and can contribute to the projection and prediction of the future. As it is important to select the most valuable methods and metrics, it is important to look at the actual data that can be used for analysis. Even how sophisticated the analysis model and method are, it is only possible to use that if the data available supports that model.

Some leagues like NHL are more open to share their data but Liiga has chosen commercial Wisehockey with closed API as their data platform. Wisehockey has a nice web-based user interface where a user can see the data, replay game situation videos etc. Wisehockey has an API for reading the stored data programmatically, but that is not publicly open. (Wisehockey)

Wisehockey is a commercial professional sports data and analytics platform developed and operated by a Bitwise Oy. It uses multiple sources for collecting data from a hockey match, including sensors on puck and players, and video data. Wisehockey offers a web interface for accessing statistics and video data. Liiga.fi public website provides some basic statistics based on Wisehockey data. (Bitwise Oy) Wisehockey web interface provides an easy yet limited view to the stored data and provides some analytics capabilities. Users of the system are not able to create custom views or queries to the data. This is where the API enters play, however the statistics available through API are limited. (Bitwise Oy)

In addition to web interface, Wisehockey offers API for accessing its data programmatically. (Bitwise Oy). However, the data regarding players performance which can be retrieved from API is limited to basic statistics. Raw data from matches are available but advanced statistics which are viewable through web interface are not available through API. API provides details about every

shot, shift, skaters' etc. The advanced statistics, such as expected goals, xG, are not available over API which limits datasets usability and creates overhead on calculating such statistics. This means that to use those advanced stats, 1) they must be calculated from the raw data or 2) use a data scraper to read the web interface and extract these statistics from scraped raw data. For sake of keeping the scope of the thesis in control, I had to choose option 1 and focus only on those basic statistics, export the expected goals for each shot in each match, and calculate an expected goals for each player by extracting each shot xG and identifying the shooter and finally summing expected goals for a player in one match.

Wisehockey API structure is somewhat complicated. Some statistics which are calculated from raw data are published through it, some are not.

Wisehockey provides ready to use analytics and statistics on players, both skaters and goalkeepers. Wisehockey divides statistics into categories which are: traditional and team play statistics, passing and shooting statistics, skating and puck control statistics, ice time and defence statistics and goalkeeper's statistics. Please see tables 1 - 4 for details about statistics which are available in Wisehockey portal, https://hub.wisehockey.com.

Statistic	Description
+	Plus. A plus is received by the scoring team's players on the ice when an even-strength goal or shorthanded goal is scored.
-	Minus. A minus is received by the players of the allowing team on the ice when an even-strength goal or shorthanded goal is scored.
+/-	Plus/Minus. Total sum of plus and minus
A	Assists. Number of assists received from shooting, passing or deflecting the puck towards the scoring teammate or touching it in any other way which enabled a goal
CA	5v5 Corsi Against. Number of shot attempts against at even strength. Shots + blocks + misses.
CF	5v5 Corsi For. Number of shot attempts for at even strength. Shots + blocks + misses.
CF%	5v5 Corsi For percentage. CF / (CF + CA)
CF% Rel	5v5 relative Corsi For percentage . CF% - CFoff% (CFoff% = team's CF% when a player is off-ice)
FF%	5v5 Fenwick For percentage. (FF) / (FF + FA) (FF = shots on goal for + missed shots on goal for)
FF% Rel	5v5 relative Fenwick For percentage. FF% - FFoff% (FFoff% = team's FF% when a player is off-ice)
G	Goals. Number of goals scored by a player
oiSH%	5v5 team on-ice shot percentage while a player was on the ice
oiSV%	5v5 team on-ice save percentage while a player was on the ice
P	Points. Number of points received by a player for goals scored or assists earned
PDO	5v5 team on-ice shot percentage + save percentage
PIM	Penalty minutes a player received in total
PME/min	Player momentum effect per minute. How much momentum score was affected on average when a player was on the ice.
Screens	Screening the opponent's goalkeeper. How many times a player screened a shot

Table1. Traditional and team play (Wisehockey API documentation)

FP DIST	Forward pass distance. Distance the puck traveled toward the opponent's end of the rink by the player's passes (m)
Pass%	Percentage of successful passes per a player's all passes
Passes	Successful passes per a player's all passes
Rcvd Passes	Received passes. Number of passes received by the player.
S	Number of a player's shots
S%	Percentage of successful shots per a player's all shots
Screens	Number of the times a player screened the opponent's goalkeeper
S TOP	Player's hardest shot
TP DIST	Total pass distance. Distance the puck traveled on the ice by the player's passes (m)
xGA	The sum of the opposing team's expected goals from the total time the player has been on the ice.
xGavg	Average xG value for a player's shot
xGF	xG For. The sum of the whole team's expected goals from the total time the player has been on the ice.
xGF%	xG For percentage. The relation between a team's and the opponent's xG from the total time the player has been on the ice. >50% means that the team creates better xG than the opponent when the player is on the ice.
xGF% Rel	Relative xG For percentage. The relation between a team's xG from the total time the player has been on the ice and xG from the time the player has not been on the ice. A positive value means that the team creates better xG when the player is on the ice.
xGsum	Sum of a player's all shot xG values

Table 2. Passing, shooting and xG (Wisehockey API documentation)

1st Controls	First puck controls in offenses. How many times a player has been the first one to gain control of the puck in an offense.
AVG Shift	Player's average shift duration (min)
BKS	Number of shots blocked by a player
FOW	Number of a player's faceoff wins
FOW%	Faceoff win percentage (per a player's all faceoffs)
PCW	Number of a player's puck contest wins
PCW%	Puck contest win percentage (per a player's all puck contests)
S	Shifts Number of a player's shifts
TOI	Player's time on the ice in total (min)

Table 3. Ice time and defence (Wisehockey API documentation)

GA	Number of goals against
GAA	Goals against average. Number of goals allowed by a goalkeeper per 60 minutes played
Pass%	Percentage of successful passes per a goalkeeper's all passes
Passes	Successful passes per a goalkeeper's all passes
SA	Number of shots against (goals + saves)
SA/min	Average number of shots against per 60 seconds
SA	Lateral Shots against made from lateral passes
Screened GA	Number of screened goals against
Screened S	Number of screened shots (goals + saves + blocked + missed)
SV%	Goalkeeper's save percentage
SV	Lateral Saves from shots made from lateral passes
xGa	xG (expected goals) against
xG Diff	xGa minus goals against
xG Diff Lateral	xGa from lateral pass shots minus goals allowed from lateral pass shots
xG Diff Screened	xGa from screened shots minus goals allowed from screened shots

Table 4. Goalkeeper statistics (Wisehockey API documentation)

Wisehockey API provides the following statistics for skaters in every game. Similar summary statistics are available for the full season. Wisehockey uses the term tournament here. Statistics for players are (Wisehockey API documentation):

- Player's top speed
- Time on ice
- Travelled distance
- Average speed while controlling the puck
- Number of accelerations
- Number of decelerations
- Number of successful passes by the player
- The total number of passes by the player
- Number of successfully received passes
- Total distance of the passes by the player
- Total distance of forward passes by the player
- +/-
- Puck control duration and distance
- Forward puck control distance
- Number of puck contest wins and losses
- Speed zone statistics: player's time spent on different skating speeds
- Number of offensive screens
- Number of blocked shots
- 5v5 Corsi For, 5v5 Corsi Against, 5v5 Corsi For percentage, 5v5 relative Corsi For percentage
- 5v5 Fenwick For percentage, 5v5 relative Fenwick For percentage
- PDO
- 5v5 team on-ice shot and save percentages while the skater was on ice

For goalkeepers, the statistics are (Wisehockey API documentation):

- Saves
- Goals against
- Saves with screen
- Goals against with screen
- Shots against with screen
- Shots against from lateral pass

- Shot area statistics: the following statistics calculated for various shot areas (at goal, close, edges, far):
    - Saves
    - Goals against
    - Saves with screen
    - Goals against with screen
    - Shots against with screen
- Number of successful passes by the goalkeeper
- The total number of passes by the goalkeeper
- Number of successfully received passes
- Total distance of the passes by the goalkeeper
- Total distance of forward passes by the goalkeeper
- Goalkeeper ”shifts”: times when goalkeeper has entered and exited the rink

From the list above it can be observed that for example the expected goals-based statistics for players are not available at the time of extracting the data via API. However, Wisehockey provides detailed level additional statistics for the game, which could be used to calculate such metrics. These detailed statistics include:

- Skater shifts
- Faceoffs
- Offenses
- Puck control
- Shots
- Goals
- Passes
- Blue line crossings (Wisehockey API documentation)

Shot stats include the expected goal value for all shots, which I will use later in the analysis. The statistic contains following features for all any recorded shot (Wisehockey API documentation):

- Period number and time of the shot
- The shooting team
- The start position of the shot
- The speed of the shot
- The result of the shot (goal, missed, saved, blocked)

- The shooter, blocker, and saver
- Screening players
- The shot area (at goal, close, edges, far) where the shot started
- Whether the shot was made directly after a lateral pass
- The direction of the shot at the goal
- Whether the royal road was crossed before the shot or not
- Team strength information during the shot
- Skating speed of the shooter during the shot
- Expected goals value of the shot

Expected goals (xG) was earlier highlighted as one of the valuable statistics when analysing players performance. If expected goals is going to be used, it requires to calculate it using raw data by summarising the xG values in a game for any player for any shot.

Same analogy goes with goals statistics. This creates more complexity to the analysis, requires more coding, and as a result limits the scope of the analysis phase for this thesis. These statistics however are available through the web portal, and the recommendation for Wisehockey is to offer the same statistics for Liiga teams to use via API. Another way to solve this would be using data scraping with the web portal. The limitations of scraping were discussed earlier in this thesis.

Exporting, cleaning, extracting, combining, and summing the datasets was a heavier task than expected during planning of this thesis work and required additional coding and steps during data transformation. Recommendation for the future development and implementation is to use effort to code scripts to export advanced statistics from the web interface to avoid heavy transformation and allow us to use the already existing advanced statistics.

6.1.2 HockeyAllsvenskan data

I found out that HockeyAllSvenskan data can be obtained either directly from HockeyAllSvenskan website or via analytics websites such as SportContract. However, the amount of data points from the games compared to Liiga is less advanced.

6.1.3 Season level data, SportContract

One of the goals for this thesis was to extract and ingest the data programmatically via API. SportContract is yet another commercial professional sports data, video, and analytics platform. SportContract invents and develops new ways to improve professional sports analytics and scouting, built on top of the latest technologies. SportContract does not tell in their public websites their data sources. One possible option is that they read the data from other free and commercial services. They also provide some data via their API, and for the purpose of the thesis I had access to both SportContract web interface and to their API. (SportContract).

Similar problems with statistics are present with SportContract API; many of the statistics shown in their web portal are not programmatically accessible through API calls. Specifically, advanced statistics like expected goals are missing. SportContract API however provides the same features for both Liiga and HockeyAllSvenskan over full regular season 2021-2022, so the full season dataset is obtained via their API. Please see table 5 for a skater full season dataset details (SportContract API documentation).

id	Players id in the system
firstName	Players first name
lastName	Players last name
position	Players role
teamId	Players teams id in the system
teamName	Players teams name
jerseyNumber	Jersey number
GP	Games played
G	Goals scored
A	Assists
P	Points
PM	Penalty Minutes
PlusMinus	PlusMinus points
PP	Power Play time
SH	Short-handed time
AS	All shots
TOIPerGame	Time on ice per game
FOW	Won Faceoffs
FOL	Lost Faceoffs
FO_perc	Faceoffs win percentage
BLS	Block shots
SOG	Shot on goals
Hits	Hits
A1	Primary Assist
A2	Secondary Assist
PPG	Powerplay Goals
SHG	Short-handed Goals

Table 5. SportContract full season dataset.

SportContract has both a web interface and API. Their web-interface has advanced and traditional statistics, but their API is limited to basic statistics over the season. Limited amount of game data can be read per transaction (10 games) and the game statistics are limited. This would have meant that the queries would become complex and the value of that data low. At the time of extracting data, per game stats API was not available. (SportContract API documentation)

Full season statistics are easily accessible over API, and I used both Liiga and HockeyAllSvenskan full season datasets exported from SportContract API. Neither SportContract makes advanced statistics such as expected goals available via their API. (SportContract API documentation)

6.1.4 Statistical model

As shown earlier in this thesis, here are various statistics and models which all try to evaluate and predict player performance. Many of these are developed for NHL, and because of that use of those models would require NHL specific data, such as the RTSS data. Many of them are not generally in use in Finnish and Swedish leagues and their capabilities or feasibility for this purpose cannot be verified in the scope of this thesis.

I interviewed one of the NHL data analysts during my thesis work and based on this interview the most crucial is to select metrics and statistics which indicate how the team is doing when the specific player is on ice, meaning what is their contribution to the team and to the result of a game. This view supports my findings on the researched advanced statistics.

Because of the additional coding that would be needed for the data transformation and to calculate the advanced statistics I needed to limit the scope of the analysis phase.
To keep the scope of my thesis manageable, I limit the statistics used to analyse Liiga players performance-based match data on Wisehockey dataset to the following:

- Time on ice - To be used to normalize the other features. Tells how much time a player is playing.
- Goals - Measures players' contribution to winning the games.
- PDO - Measures “luck” and randomness.
- Expected goals For - Measures the offensive pressure a player has when on ice.
- Plus/Minus - Measures a players contribution to winning the game.

- CorsiFor and Corsi% - Measures how many shot attempts were made when a given player was on ice and can indicate players contribution to the game result.

Based on the theory studied in the thesis earlier, this set of statistics should give a reasonable view on a player's performance using historical data.

6.1.5 Data model and datasets

For this thesis, and to keep the thesis scope manageable, I decided to narrow down the datasets to be used in example analysis in one full regular season. This further means that I decided to use a full season data from the Finnish Elite League, Liiga, from the most recent regular season period, which is the 2021-2022 season. Full season consists of 450 matches in total, with 15 teams playing meaning 60 matches per team (Wikipedia). Full regular season offers variance in games, all teams play against, whereas in playoffs and finals, only certain teams play against each other, which could of course affect the results of the analysis. I used the same analogy for HockeyAllSvenskan data, the most recent data for full season is regular season 2021-2022. In this league, in the regular season, there were 14 teams playing, 52 games per each team, and in total 364 games (Wikipedia).

From the API sources I extracted the following datasets. All datasets contain regular season 2021-2022 data. As a general comment on the datasets, the source from Wisehockey uses feature named id for many of different occasions and to avoid mixing them during analysis, there was a need to rename these features when creating actual analysis queries, tables, and views. I needed also to add additional identifiers to the datasets, like matchid. Prefiltering, for example dropping several features not necessary for my thesis, of the raw data and concatenating the results to single files per dataset was done for easier handling and uploading to the analytics pipeline. Datasets are:

1. Teams
    - Source: Wisehockey API
    - Contents: All teams played in Liiga
    - Features: id, fullName, shortName
2. AllPlayers
    - Source: Wisehockey API
    - Contents: All players played in Liiga
    - Features: id, firstName, lastName, jersey, role, teamId

- Notes: If a player has been playing in more than one team during the season, there are separate entries for each. This created some challenges during analysis.

3. AllMatches
    - Source: Wisehockey API
    - Contents: Information about all matches
    - Features: id, date, status, homeGoals, awayGoals, lastUpdatedhomeTeam.id,homeTeam.fullName, homeTeam.shortNameawayTeam.id, awayTeam.fullName, awayTeam.shortName, venue.name, venue.city
4. AllMatchStatistics
    - Source: Wisehockey API
    - Contents: Detailed match statistics for each match
    - Features: player, team, totalStatistics.skatingStatistics.topSpeed, totalStatistics.skatingStatistics.timeOnIce, totalStatistics.skatingStatistics.distanceTravelled, totalStatistics.skatingStatistics.averageSpeedWithPuck, totalStatistics.skatingStatistics.accelerations, totalStatistics.skatingStatistics.decelerations, totalStatistics.shiftStatisticsSummary.shifts, totalStatistics.shiftStatisticsSummary.averageShiftDuration, totalStatistics.passStatistics.successfulPasses, totalStatistics.passStatistics.allPasses, totalStatistics.passStatistics.receivedPasses, totalStatistics.passStatistics.totalPassDistance, totalStatistics.passStatistics.forwardPassDistance, totalStatistics.plusMinusStatistics.plus, totalStatistics.plusMinusStatistics.minus, totalStatistics.plusMinusStatistics.total, totalStatistics.puckControlStatistics.puckControlTime, totalStatistics.puckControlStatistics.puckControlDistance, totalStatistics.puckControlStatistics.puckControlForwardDistance, totalStatistics.puckContestStatistics.puckContestsWon, totalStatistics.puckContestStatistics.puckContestsLost, totalStatistics.puckContestStatistics.puckContestParticipations,

totalStatistics.shotScreenStatistics.offensiveScreens,
totalStatistics.shotScreenStatistics.blockedShots,
totalStatistics.traditionalStatistics.corsiFor,
totalStatistics.traditionalStatistics.corsiAgainst,
totalStatistics.traditionalStatistics.corsiForPercentage,
totalStatistics.traditionalStatistics.relativeCorsiForPercentage
totalStatistics.traditionalStatistics.fenwickForPercentage,
totalStatistics.traditionalStatistics.relativeFenwickForPercentage
totalStatistics.traditionalStatistics.pdo,
totalStatistics.traditionalStatistics.fullStrengthSavePercentagetotalStatistics.traditi
onalStatistics.fullStrengthShootingPercentage, matchid

5. AllMatchShots
    - Source: Wisehockey API
    - Contents: Detailed shot statistics for each shot in each match
    - Features: team, period, result, shooter, expectedGoals, teamStrength.type, matchid
6. AllPlayerStats
    - Source: Wisehockey API
    - Contents: Full season statistics for players
    - Features: numberOfMatchesPlayed, playerid, skatingStatistics.topSpeed,skatingStatistics.timeOnIce, skatingStatistics.distanceTravelled, skatingStatistics.averageSpeedWithPuck, skatingStatistics.accelerations, skatingStatistics.decelerations, faceoffStatistics.faceoffWins, faceoffStatistics.faceoffCount, plusMinusStatistics.plus, plusMinusStatistics.minus, plusMinusStatistics.total, shotStatistics.shots, shotStatistics.goals, passStatistics.successfulPasses, passStatistics.allPasses, passStatistics.receivedPasses, passStatistics.totalPassDistance, passStatistics.forwardPassDistance, puckControlStatistics.puckControlTime, puckControlStatistics.puckControlDistance, puckControlStatistics.puckControlForwardDistance, puckContestStatistics.puckContestsWon, puckContestStatistics.puckContestsLost, puckContestStatistics.puckContestParticipations,

shotScreenStatistics.offensiveScreens, shotScreenStatistics.blockedShots, traditionalStatistics.corsiFor, traditionalStatistics.corsiAgainst, traditionalStatistics.corsiForPercentage, traditionalStatistics.relativeCorsiForPercentage, traditionalStatistics.fenwickForPercentage, traditionalStatistics.relativeFenwickForPercentage, traditionalStatistics.pdo, traditionalStatistics.fullStrengthSavePercentage, traditionalStatistics.fullStrengthShootingPercentage

7. SkatersLiiga
    - Source: SportContract
    - Contents: Full season statistics for skaters (all players excluding goalies)
    - Features: id, firstName, lastName, position, teamId, teamName, jerseyNumber, GP, G, A, P, PM, PlusMinus, PP, SH, AS, TOIPerGame, FOW, FOL, FO_perc, BLS, SOG, Hits, A1, A2, PPG, SHG
8. SkatersHockeyAllSvenskan
    - Source: SportContract
    - Contents: Full season statistics for skaters (all players excluding goalies)
    - Features: id, firstName, lastName, position, teamId, teamName, jerseyNumber, GP, G, A, P, PM, PlusMinus, PP, SH, AS, TOIPerGame, FOW, FOL, FO_perc, BLS, SOG, Hits, A1, A2, PPG, SHG

6.2 Analytics architecture

This chapter discusses the analytics architecture.

6.2.1 Architecture considerations

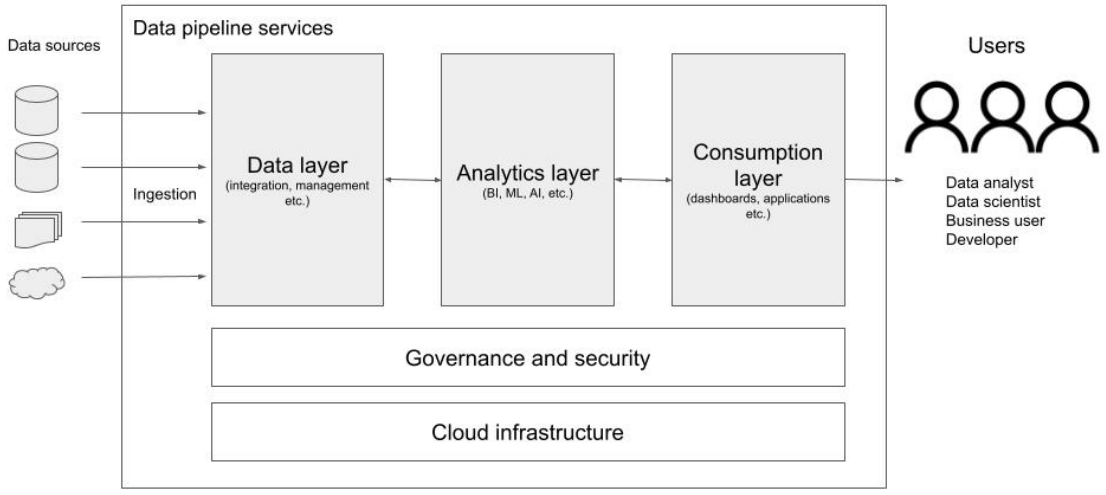
Key architecture considerations for analytics include:

- What data is needed to answer to the question?
- What data is available, where, and how to get it? How to ingest the data?
- Can the data answer to the business question? If not, is there more relevant data, or can the question be modified?, Should exploratory analysis be used to find the insights from that data?

- What is needed to be done to the data before it can be used in analysis? How to transform the data?
- Where to store the raw and transformed data?
- How to query and analyse the data?
- How to present the answer to the business question(s) in a format which is understandable to human, so how to visualize and get meaningful insights and context to the data?

Organizations often have huge amounts of data, which is spread across different platforms and services. As the amount of collected data is increasing, access to it becomes important. Many tools offer in-application statistics and visualizations which may be too simple and are made only to fulfil some simple task (Jarett, Kuo, 2021). Many companies have invested in data practices but still are looking for how to make value from data. Becoming a data driven company requires a holistic data strategy which among other things, considers the right ecosystem. Traditional way has been to include point solutions which provide data services. However, due to the amount of the data, the complexity, and the various needs and use cases these point solutions are not sufficient anymore (Pierce, Tekiner, 2021). By centralizing data sources for cross-platform analytics enables organizations to get an accurate picture of their business (Jarett, Kuo, 2021).

A platform with distributed data warehouse, which supports structured, semi-structured, and streaming data, supporting clients and languages commonly used in data science such as Python, R and SQL, including built-in capabilities for machine learning, and advanced analytics can be called unified data analytics platform. (Paige, 2021)

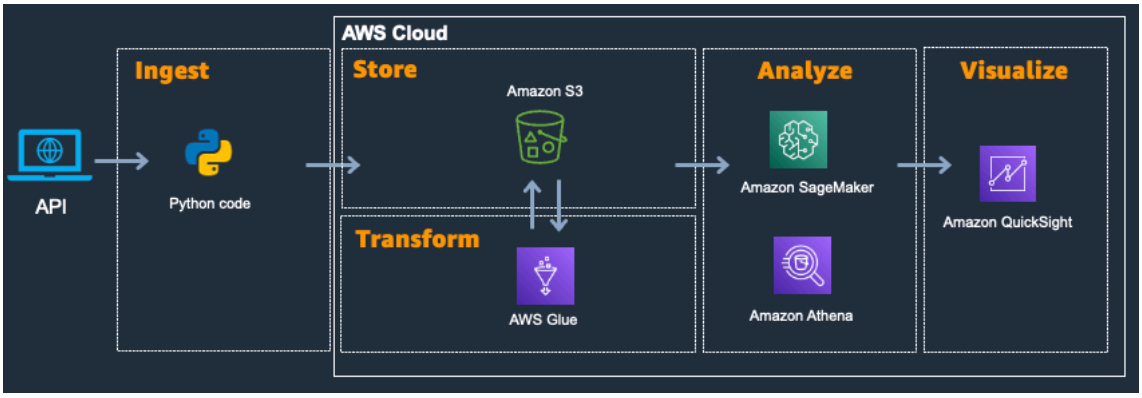


Picture 1. Simplified architecture for data analytics.

Picture 1 illustrates a unified platform which is capable of ingesting diverse data from multiple sources, has pipeline services, a data storage (the data environment), an analytics layer, and a consumption layer. The platform is deployed in the cloud, is governed and secure by design. (Halper, 2021)

6.2.2 Architecture and data processing flow

For this thesis purpose, I built a major part of the analysis tools and pipeline into AWS Cloud. I used Amazon Web Services documentation at https://docs.aws.amazon.com/index.html as instructions for building and configuring the environment and running analytics and "AWS serverless data analytics pipeline reference architecture" for designing the architecture. Only for extracting the data from both API sources, I used a Putin 3.8 with Jupyter notebook instance on my own laptop. Picture 2 presents the analysis pipeline and environment I built to ingest, store, transform, analyse, and visualise the datasets. (Amazon Web Services Documentation)



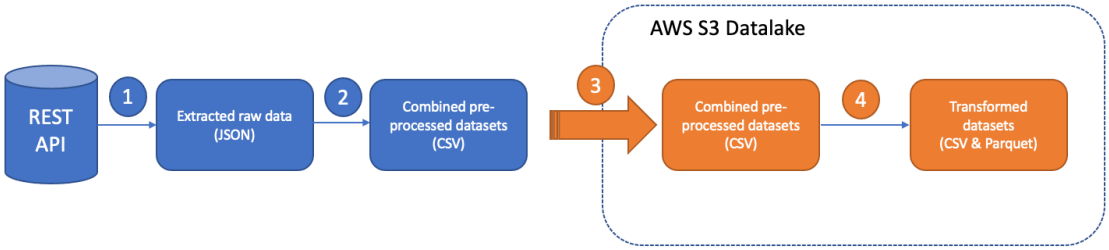
Picture 2. Analysis pipeline in AWS.

The pipeline used for the thesis was:

- Data extraction from API - Microsoft Visual Studio Code with Jupyter Notebook plugin, Python 3.8
- Data storage / Datalake - Amazon S3 bucket
- Raw data ingestion to datalake - Upload raw data files from my computer to S3
- Transforming raw data - Amazon Glue Jobs
- Importing data to data catalog for analysis - Amazon Glue Data Crawler
- Data catalog - Amazon Glue Data Catalog

- Querying text data in datalake and building analysis tables and views - Amazon Athena
- Applying machine learning models - Amazon SageMaker Canvas
- Visualizing, extracting insights and detecting anomalies - Amazon QuickSight

Picture 3 below illustrates the flow for data processing used in this thesis. In this research, steps 1 and 2 are done within researchers computer, steps 3 and 4 are done in the cloud.



Picture 3. Data processing flow.

1. Data is extracted from RESTful API interface with Python code and is stored in local computer in JSON format.
2. Data is then combined and pre-processed into raw comma separated (CSV) datasets presented in 6.1.5
3. Combined and pre-processed CSV datasets are uploaded into AWS S3 data lake (see 6.2.5 for details)
4. Datasets are transformed with AWS Glue ETL jobs into CSV and Parquet formats to be queried and analysed and to be used in training and predictions with machine learning.

6.2.3 Extracting data from API

For this thesis I am using Jupyter notebook and Python to write custom code to fetch the datasets from WiseHockey and SportContract API's. As described earlier, both offer limited data via their API, which indicates that to use advanced analytics already prepared by these sources, the data needs to be ingested with different style; by using data scraping from their web portal.

For extracting and filtering the raw data for the thesis, I used Microsoft Visual Studio Code, Jupyter notebook plugin and Python 3.8 to build the needed code on my laptop computer. Raw data handling was done mainly using Pandas and Numpy libraries. Both APIs were using REST API

format and they had multiple query paths. Structure of API's was such that no complex queries are possible straight to the API meaning a larger portion of data needs to be extracted, and suitable data filtered from the extracted raw datasets. The code is written so that it can be imported with small modification to AWS Lambda functions for constant data fetching. Per season data needs to however be fetched only once after the end of season. If dataset is needed to be updated and kept current with most recent data, "season to date", principle it can also be fetched regularly, for example after each round of matches.

Data extraction was done in JSON (JavaScript Object Notation) and CSV (Comma Separated Values) format. The Python code is portable and can be reused if such a pipeline is built for production use later.

Other possibility is to used data scraping. It is a technique where scraping software basically reads the same web interface, a web page, programmatically which is intended for the human eye. There are multiple scraping tools available, both free open source and commercial licensed tools. Scraping requires programming and the created scrapers can be affected by changes in structure, naming and code of the target website. This means that they may require regular modifications and updates. Another angle to consider with data scraping is license agreements and terms of conditions. I did not investigate whether any of the two data sources allow or deny use of data scrapers. (Wikipedia)

6.2.4 Ingesting data to datalake

Ingesting raw data into the data lake was done by simply uploading raw data files into raw folder in a data lake. Further on, under the raw folder, raw data was organized to multiple sub-folders to keep the data manageable.

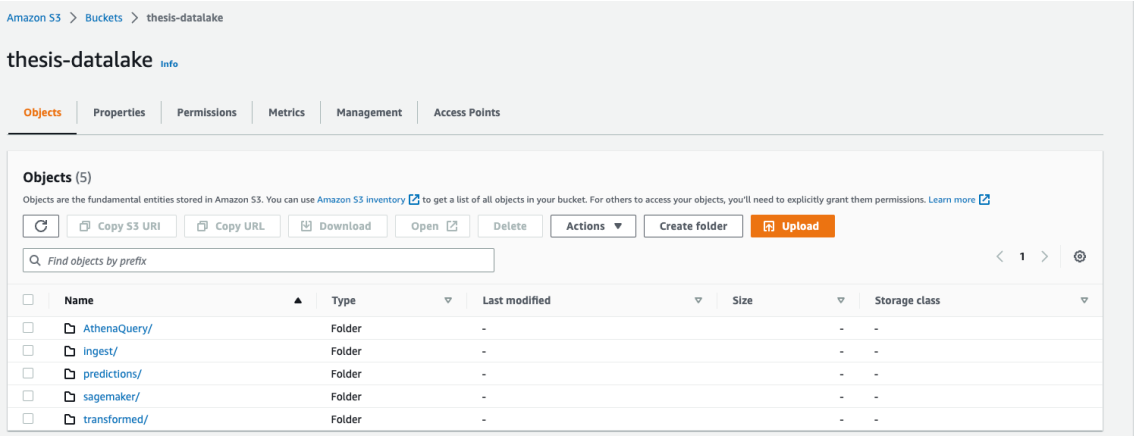
Extracting data and ingesting it to the data lake can be done and automated in AWS by for example using Lambda function written in Python. The Python code written for this thesis can be used as template for the lambda function for future use. Lambda is serverless, meaning code can be executed without running any servers. Lambda function can connect to the API interfaces over the internet, query and extract the data from the API and upload the raw data to the data lake.

In AWS cloud-based data analysis solutions, this could be done by using for example Lambda functions. Python code prepared for the thesis can be transferred to Lambdas.

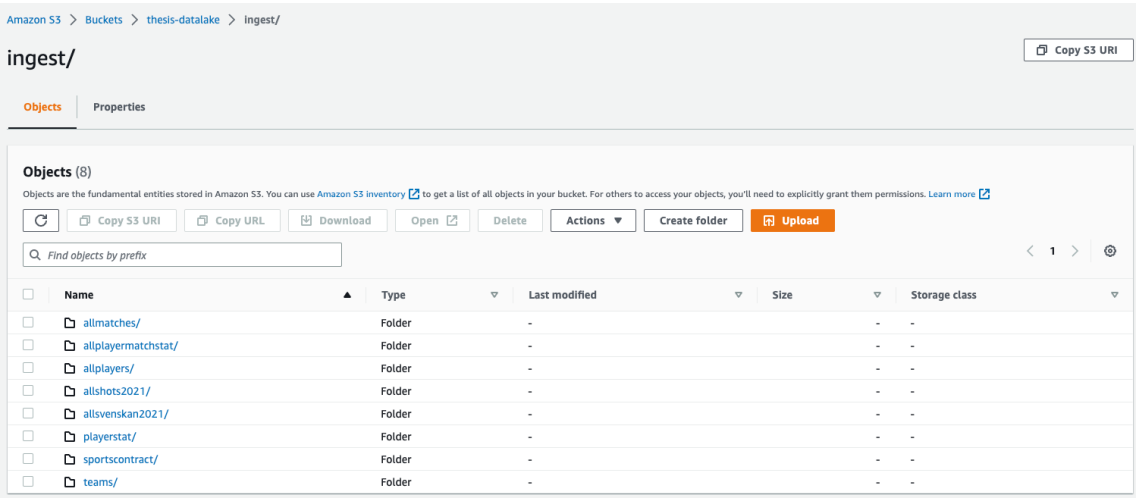
6.2.5 Storing data

The data lake is a core component of the data analytics platform. Modern data warehouses support self-service, advanced analytics, and data sharing. The data warehouse ingests multiple sources of data and makes them available for analysis. (Halper, 2021)

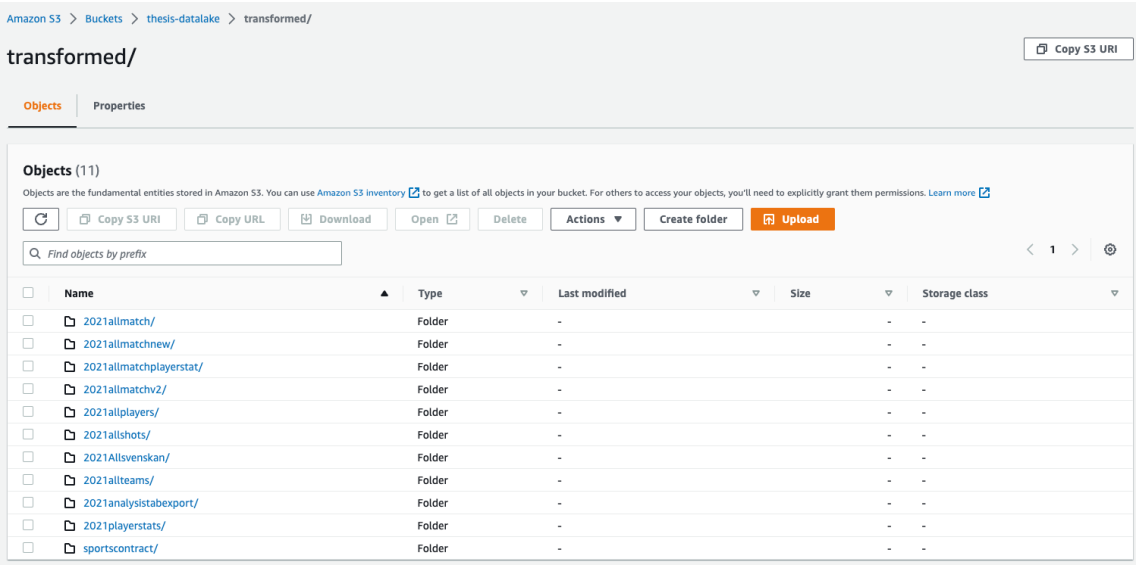
For storing data, I used AWS S3 bucket as data lake which top level structure is shown in pictures 4 - 6. Data stored in S3 can be accessed with the services needed for transforming, analysing and visualization. Raw data was uploaded to the data lake. The data lake used for the thesis has its own folders for each dataset and use.



Picture 4. Data lake top level hierarchy.



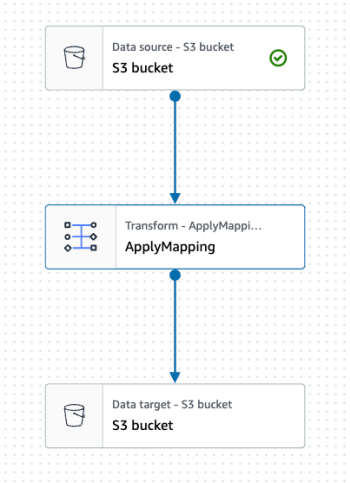
Picture 5. Ingest folder in data lake containing raw uploaded datasets.



Picture 6. Transformed folder in data lake containing transformed datasets.

6.2.6 Processing and transforming

After ingestion, raw data is transformed and cleaned for processing. This is done with ETL tool AWS Glue. Example ETL flow is shown in picture 7 below. As the data from both data sources are pre-processed, it did not contain any major notable errors, thus can be used as such for the analysis. In this phase, from raw data unneeded data features can be dropped, however this part was done with the ingestion python code. Data was further transformed to CSV and Apache Parquet format, which makes querying it more efficient and faster. In this phase datasets could also be combined when needed. CSV format is mainly used for training Machine Learning models, which require structured text format as input.



Picture 7. ETL flow in AWS Glue.

The results of the transformed dataset are then stored to the data lake and metadata is written to AWS Glue data catalogue to enable querying it with Amazon Athena. Athena uses a SQL like language to query the files in the data lake directly without setting any additional databases. With Athena you can query the dataset with easily understandable and common language. The limitation for Athena is that it is a query tool, and the underlying data files cannot be updated with it. It supports using use Data Query Language commands (select), but not for example Data Manipulation Language like update or insert. This made forming data tables and views more complex, because I could not for example change values in underlying data or modify values. This suggests that for a more complex environment, an SQL database, or Data Warehouse for may be needed for data storage, update, and queries. Luckily, I was able to do most of the needed data modification tasks within QuickSight, which again has SQL like language for queries and manipulation.

At the ETL phase the data can be cleaned and combined with other datasets. This is a recommendation for future use cases, to use AWS Glue to combine and edit the datasets where needed. However, because getting all the data and coding to extract and ingest the data, defining structures, tables, and views, due to the structure of extracted data, I did not have extra time to come back, code the ingestion and transformation again and I had to manage the scope by handling the data in analysis and visualization phases.

6.2.7 Building queries and combining datasets

Data is queried and combined with Amazon Athena. I used Athena to build tables and views for visualization. Table created table and view structure is presented in appendix 1.

At this stage complexity of exported and ingested dataset came visible. To calculate expected goal statistic for a player in a single game, I needed to create some complex SQL queries which combines multiple ingested datasets.

One of the issues is that not every player shoots or makes goals, and when this happens, that part of the query per player is missing. One player could have played in more than one team in a season, and in this case the SQL query join returns null value for the Players team. Amazon Athena is a query tools and language, which means with it there is no possibility to update the underlying

datasets stored on S3 data lake. Previously described problems with the dataset could be avoided by creating a SQL database and updating missing values with SQL query language. Using a SQL database such as Amazon Aurora or Redshift data warehouse is recommended for future architecture.

If the API would offer the data in the same structure which is displayed in the WiseHockey web-portal, would also solve this, the table structure would contain only valid tables, the data ingested and analysed data amount would be significantly less, and analysis could have been done without any additional transformation and query building. At the current API exported dataset, we must build all the advanced statistics, meaning we must build a portion of the logic of the application again. Hockey teams already have access to those statistics through a web interface and are paying a license fee for Wisehockey and can expect to have the advanced statistics accessible via API.

6.2.8 Tables and views for analytics

For building the actual analytics selecting relevant features, combining, counting sums, and querying the data is used following Athena tables. As the table structure is not relational, I am not defining or drawing relation formulas nor drawing relation diagrams. Table 6 below describes analysis tables and their content. Please see appendix 1 for detailed description of the tables. In addition, I used a few temporary tables to build and test the final structure.

Table / view	Contents	Purpose
Liiga_2021_analysis_view_v1_1	Final data for analysis	Analysis and visualization in QuickSight
2021allplayersandteams	All players and their teams	Interim phase table
2021allshots	Data about all shots in all matches	Query shot data
2021allteams	Data about all teams	Query team data
2021playerstats	Data about players performance over the season	Query seasonal performance data
v_player_team_count	Data about players team count during season	Used for normalizing data in visualization phase
2021allmatch	Data about matches	Query match data
2021allmatchplayerstat	Data about players performance in a match	Query match level performance data
2021allplayers	Data about players	Query player data
Liiga2021allgamesgoals	Combined data with goal information for all matches	Query goal data
Runkosarja2021allshotsxg	Combined data with sum of expected goals for all games for all players	Query expected goals data
Rs2021playerswithteamsxg	Combined data, expected goals data with players team information	Interim data table for building final view

sc-liiga2021-skaters	Performance data over the whole season from Sportcontract for Liiga	Train ML model
sc-allsvenskan2021-skaters	Performance data over the whole season from Sportcontract for HockeyAllSvenskan	Test ML model

Table 6. Tables and views used for analysis.

6.2.9 Visualization

For visualizing and creating graphs from the data, I used Amazon QuickSight. QuickSight can use the built Amazon Athena tables and views as a data source and it can be used to calculate calculated fields, mean, max, median, and other statistics over the data. I used the built Athena queries as datasets for the visualization.

6.2.10 Machine learning

For creating predictions on full season results, I used Amazon SageMaker canvas. SageMaker canvas makes it possible to use machine learning with no coding. For this thesis purpose, to test if I can predict season results, I used only quick model creation. The accuracy of the model increases if model is further tuned, which means using methods requiring coding.

I trained the model with Liiga full 2021 regular season dataset and the trained model to predict players performance in Liiga based on HockeyAllSvenskan 2021 full regular season dataset. Because I only had a limited number of features and no advanced statistics at all in either dataset, I trained the model to predict players scored goals in a season.

I used Canvas to try if time series prediction is possible based on season match data. This model can predict players future performance in a defined timeframe from the prediction day onwards. As I had only one full season of the data, the model can only give predictions 30 days into the future. I trained a model to predict players' future goals, but the model can be trained to predict any relevant feature which is present in the dataset. Example prediction is presented later in this thesis report.

For third use case for machine learning in this thesis, and with considering the results, I used Amazon QuickSight Anomaly detection to detect unusual patterns in a player's performance. This turned out to be an interesting approach and after discussion with the thesis target organization, is approach they consider very interesting and see potential for future use. Instead of creating predictions with complex statistics, we use simple features and data points from match data to

identify low and high performance. This model does not rely on any specific league, source, or dataset. Any dataset from any league can be used to train a model to predict anomalies, meaning if a player performs better, or worse, plays a longer period with making points or goals or without goals, that can be detected as anomaly and automatic alert can be generated. This could give a hint for scouts and coaches that a player is performing outside their normal performance, regardless of the level, league, team, or role they are playing. Same analogy applies to time series predictions, as I am not trying to compare players against others, between the teams or leagues, and just trying to get an indication of the future performance. The relevant thing is that we have enough data we can use to train the model and verify and make the predictions.

As advanced statistics may be important to monitor and to some extent predict players performance, they only rely on data from history. The use of anomaly detection and machine learning to identify anomalies and to predict future performance with simple match statistics is novel and something to study more. In this model, I used to predict goals, because by making goals you win games, but both cases could be used to predict any parameter in players in-game performance.

The initial trial to train a machine learning model to predict players performance in Liiga based on their performance in other leagues is also something to study more in the future. With the model I trained using this limited feature dataset is probably not very accurate (we cannot verify it because we do not have data from future games in Liiga for all the players). With a more extensive data feature set, models' accuracy could be improved, however this still would give only indication of the direction rather than accurate prediction. With more training data, the model gets better and possibly would offer some value in player decisions. Data can be collected over time on those players selected to play in Liiga and then used create a dataset based on their historical data and real performance in Liiga. This can be then used to train a model. Collecting this data could take time, and still the dataset would have too few data points to train accurate model.

Further on, an assumption can be made that an automatic machine learning models can indicate what parameters are most relevant when they make predictions. This can give an indication what features and statistics are most relevant to look after on that team, league, season, player, or role.

Libraries like Autogluon can be used with SageMaker notebook instance, however, requires some coding with python language. With Autogluon with only small coding effort a model can be selected and trained and then perform predictions based on tabular dataset.

6.2.11 Security and data protection

Security and data protection of the data is important. The dataset used in this thesis contains player names with their performance data, which is in big parts available on public internet, but can be considered data which is regulated with GDPR. The analysis results also form a valuable dataset. These two aspects are key considerations for security, and when such a system and pipeline is put into production must be taken into use. For the thesis purposes, all data was held in a data lake which is encrypted with strong encryption (using AES-256 with SSE-S3 keys) at rest. Access to the data is limited to only to one analysis role and the service roles needed for analysis and pipeline to run.

For the production analytics system, strong at rest and in transit encryption should be applied and enforced, data lake policy to restrict access to the minimum, to enforce only encrypted uploads, to restrict access from outside own organization and deny making data lake bucket public. It is important to not open any access to the analytics environment, except Amazon QuickSight, from the public internet. A fine-grained access policy can be created with AWS Lake Formation and to ensure all components in the pipeline use encrypted data lake bucket as their data storage.

Further consideration is to use an additional layer of protection by protecting the data environment with Threat Detection service like Amazon GuardDuty and identifying sensitive information and possible misconfigurations with Amazon Macie.

To avoid unnecessary contractual and consent management it is important to store the data only within EU. Further, the account used to store and process the data should be separate from other purposes and locked down with proper access and service control policies, network connection filtering and setting access boundaries.

6.3 Analytics results

6.3.1 Anomaly detection

When discussing with my thesis principal KalPa, one of their specific wishes was to find out if there is a way to automatically detect unusual patterns in players performance, in other words is there a way to detect anomalies and get alerted automatically when they are detected.

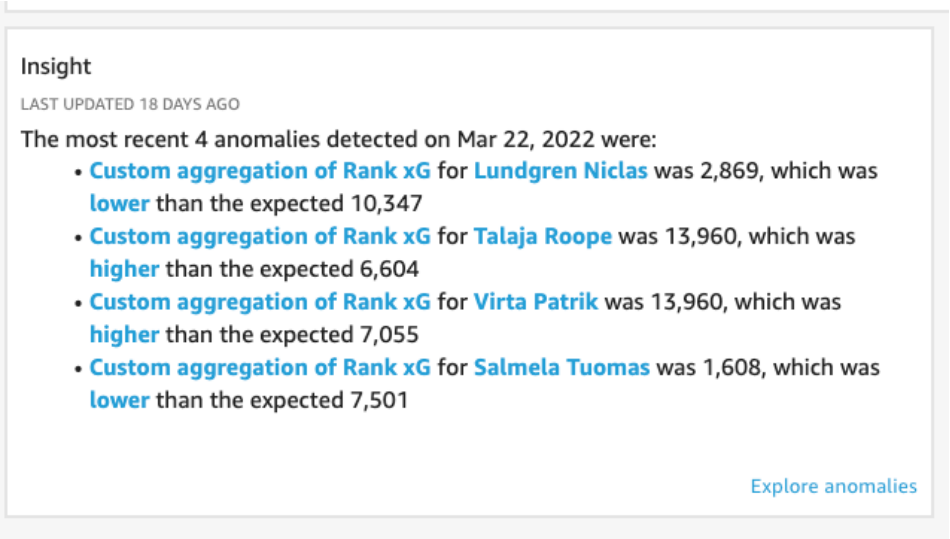
Anomaly detection could give indication on changes on players performance, for example if a player is on longer strike making points. I used Amazon QuickSight for visualizing the results, and that service has built in automatic capability to use Machine Learning for detecting anomalies in the data set. (Amazon)

Anomaly detection has been well studied over the last few decades and is one of the key problems in data mining. Key questions in detecting anomalies are 1) how anomaly is defined and 2) what data structure support to detect anomalies? Generally, it can be said that a data point is an anomaly if the complexity of the model increases substantially with the inclusion of the data point. In supervised learning randomization has been found to be effective l tool. Guha, Misra, Roy and Schrijvers introduced the robust random cut forest sketch. The Robust Random Cut Forest (RRCF) algorithm is an efficient ensemble method for detecting outliers in streaming or time series data. (Guha, Mishra, Roy, Schrijvers, 2016)

A random cut forest (RCF) is a special type of random forest (RF) algorithm which takes a set of random data points, cuts them down to the same number of points, and then builds a collection of models. RCF is similar to a decision tree, is an unsupervised algorithm and uses cluster analysis to detect spikes in time series data, breaks in periodicity or seasonality, and data point exceptions. (Amazon)

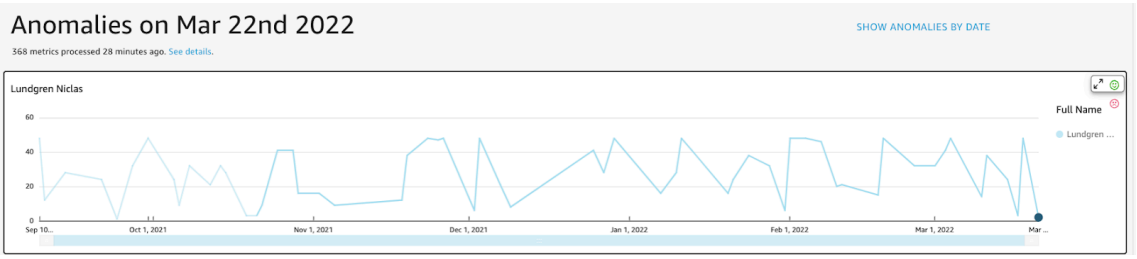
Amazon QuickSight uses RCF for anomaly detection. In addition to detection, it includes a feature to send automatic alerts when anomaly has been found (Amazon). As a test case for anomaly detection, I used a ranking value I calculated using expected goals and PDO. Picture 8 shows anomalies QuickSight identified on data for xG Rank March 2022 and picture 9 detailed view of first identified anomaly.

When looking at the absolute values identified in the anomalies, it should be noted that the Rank xG is reverse rank, meaning the lower value is better. Here It can be observed that for example Niclas Lundgred played better than usual, hence an anomaly was detected.



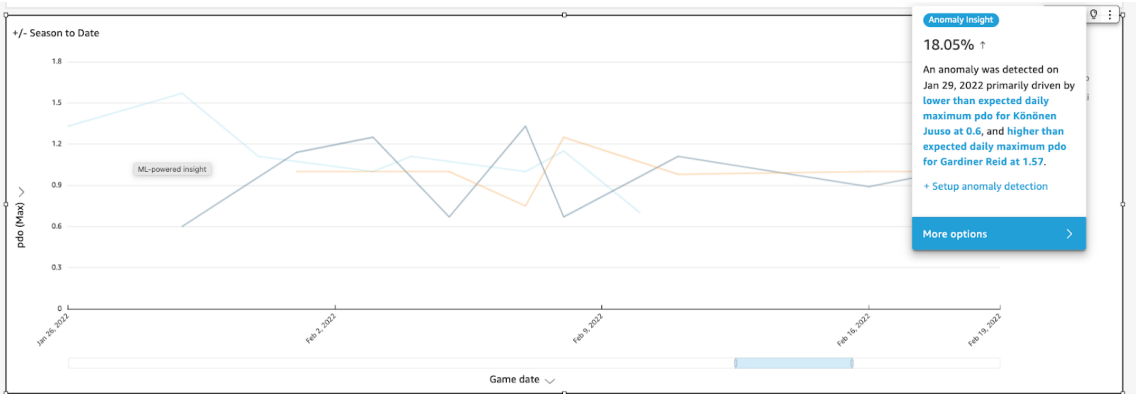
Picture 8. Anomalies on Rank xG on March 22, 2022.

When looking closer at the anomaly on Lundgren, the graphical interpretation of the anomaly can be used. The graphic is scaled, and value should be multiplied with 1000. From the graphic it is observed that Lundgrens xG rank per game have varied around 20000 with regular spikes closer to 1000, which all would have been detected as anomalies. On March 22nd the value was again detected as anomaly, meaning Niclas Lundgrens expected goals in a game was higher than expected, which could indicate better than normal performance.



Picture 9. Detailed view of anomalies for Niclas Lundgren.

As another sample for anomaly detection, I used PDO (picture 10). On Jan 29, 2022, an anomaly was detected for Juuso Könönen and Reid Gardiner. PDO for Könönen was lower than usual, and for Gardiner higher than usual. For Könönen this could indicate lower performance than expected, and for Gardiner it can indicate higher performance than expected.



Picture 10. PDO anomalies for Könönen and Gardiner.

What is notable here is that these anomalies would have been automatically detected and coaches and scouts alerted.

6.3.2 Normalising and summarising

In the dataset extracted from Wisehockey, when a player has been playing in more than one team during the season, a player has a player entry for each team. For example, Reid Gardiner played in two teams during regular season 2021-2022 (JYP and HIFK). This fact caused that in the analytics tables for shot based statistics I created using Amazon Athena queries, such players had multiple entries for each match. This problem was overcome by normalizing the summary values with the count of players teams in that season. This method was used to normalize summarized expected goals and goals values.

6.3.3 Expected goals ranking

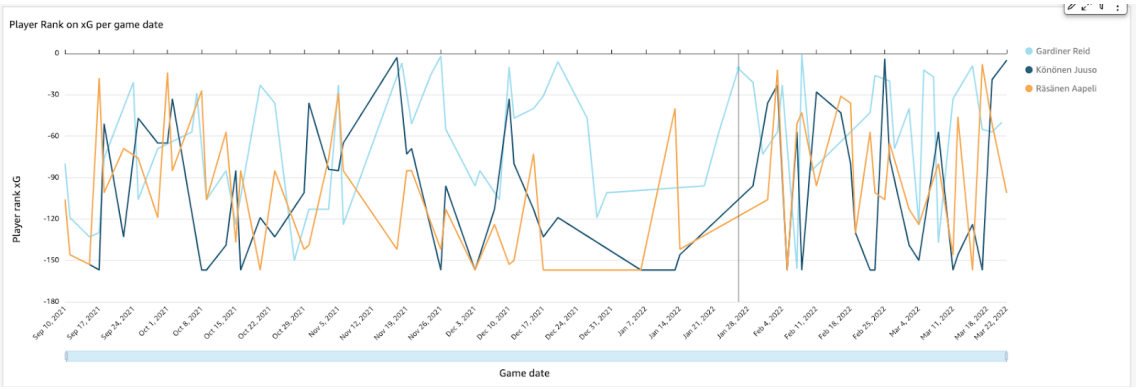
For evaluating players performance, I created a statistical value, Rank xG, which gives a ranking for players performance on a particular game using expected goals. So, for each player in each game where the player made a shot for which expected goal value was given a ranking value was given. This value is descendant, meaning value 1 is the best, 2 second best and so on. By doing so, expected goal statistics in a time series is obtained. This means that players' performance against others using time as a variable can be valuated and that the ranking number directly indicates the order. The highest on the list is the best of all records during regular season 2021-2022. An absolute expected goal value on a given day can be used, but that would not indicate its relation to other players in a single value. This metric gives a single number view to players performance; it measures the offensive pressure a player has when on ice compared to

the reference group. As a reference group, the whole Liiga, players own team or players role or a combination of the earlier can be used. Table 7 shows a sample of calculated rankings by using the whole Liiga as a reference group.

Full name	date	Rank xG	Percentile Rank Role	xG	xG sum over season	plusminus	xG seasontodate	xG Liiga Average	xG Team Average	pdo
Nenonen Markus	Feb 4, 2022	1	99.63	1.61	16.92	0	11.33	0.24	0.33	1
Lappalainen Lasse	Nov 6, 2021	2	99.97	1.38	14.89	1	7.04	0.21	0.19	1.14
Hults Mitch	Dec 15, 2021	3	99.43	1.33	11.94	0	6.66	0.29	0.35	1.09
Nikkilä Petteri	Oct 6, 2021	4	99.97	1.28	9.27	-1	2.65	0.2	0.48	1.25
Kuusela Kristian	Feb 19, 2022	5	99.22	1.27	10.86	1	8.55	0.21	0.33	1.11
Jasek Lukas	Nov 4, 2021	6	99.35	1.21	13.72	-1	5.17	0.17	0.25	0.8
Mäenalanen Saku	Dec 4, 2021	6	99.35	1.21	10.11	2	4.8	0.21	0.31	1.14
Lakatos Dominik	Oct 27, 2021	8	99.01	1.2	4.87	-2	4.58	0.19	0.33	0.75
Turkulainen Jerry	Nov 6, 2021	9	98.76	1.16	13.78	0	5.04	0.21	0.26	1
Liedes Heikki	Oct 6, 2021	10	99.12	1.14	12.68	-1	3.05	0.2	0.48	1.33
Palmu Petrus	Feb 1, 2022	10	98.54	1.14	18.45	1	12.49	0.23	0.32	1
Nurmi Markus	Nov 18, 2021	12	98.51	1.13	14.77	2	5.22	0.26	0.36	1.14
Emanuelsson Petter	Dec 10, 2021	13	98.49	1.12	14.65	1	5.74	0.19	0.2	1.14
Sebok Balázs	Feb 11, 2022	13	99.02	1.12	7.78	2	6.44	0.2	0.2	1.23
Stransky Simon	Oct 22, 2021	13	98.69	1.12	14.51	-1	4.06	0.21	0.3	0.5
Emanuelsson Petter	Feb 5, 2022	16	98.29	1.1	14.65	2	9.93	0.2	0.21	1.33
Innala Jere	Sep 23, 2021	17	98.45	1.07	23.05	-1	2.69	0.2	0.27	0.75

Table 7. xG Ranking.

I compared three players' performance using the calculated Rank xG (Reid Gardiner, HIFK, Juuso Könönen, KalPa and Aapeli Räsänen, KalPa). In addition, I used Rank xG ranking in anomaly detection. When looking at the results on picture 11 should be noted values in y axis should be read as absolute values. Picture shows that a team's coach can possibly use this data as one data source, but for many without detailed understanding of the game on a particular date, it is hard to identify the patterns for better or lower than usual performance. Hence using ML based anomaly detection makes sense.

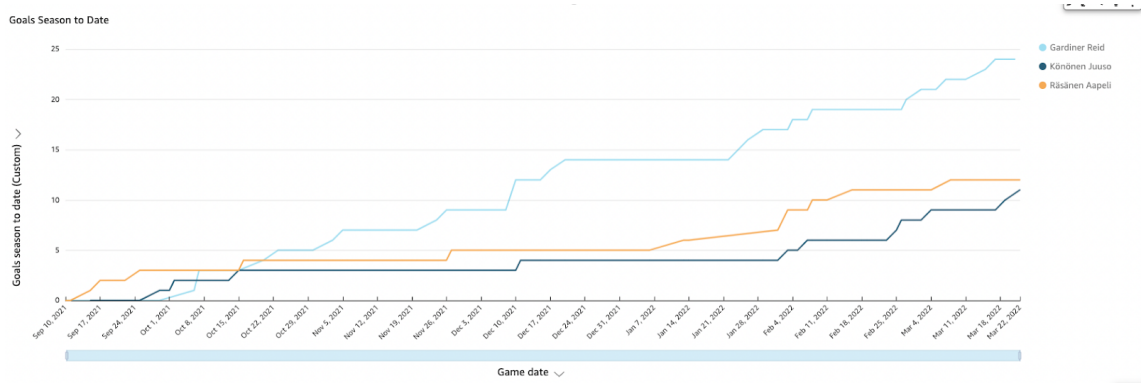


Picture 11. Rank xG for Gardiner, Könönen and Räsänen.

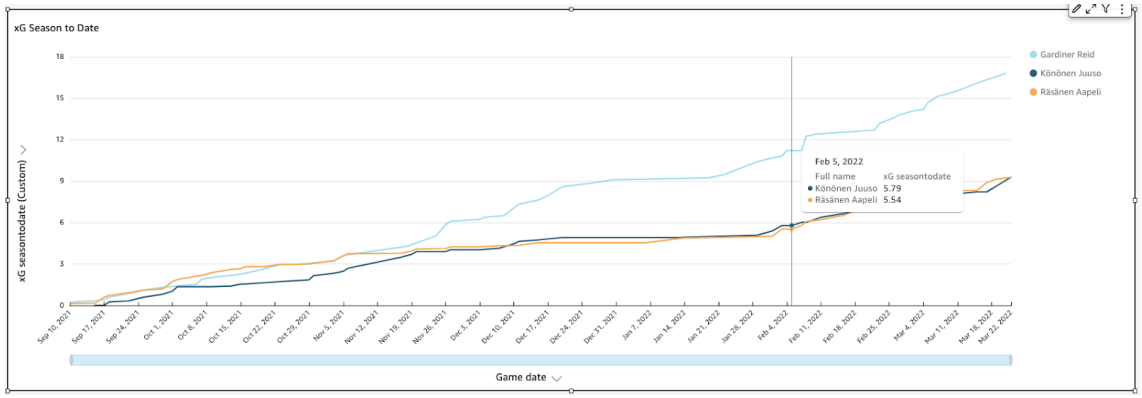
6.3.4 Analysis against reference group

As defined in earlier of this study, players' performance can be compared against each other, when compared within their own reference group. In this example analysis, I used a comparison between three Liiga forward skaters, Reid Gardiner (HIFK), Juuso Könönen (KalPa) and Aapeli Räsänen (KalPa). Comparison visualizes the development and performance over the comparison period, which in this example is Liiga full regular season 2021. From the visualization it is easy to see stages during the season, how players perform during the season and the comparison over the full season. However, it should be noted that playing in different level teams, where better teams win more often, thus players are likely to score more often, should be considered in comparison. But for example, comparing two players within the same team should give reliable indication of their differences.

Picture 12 presents goals players have scored and picture 13 expected goals for a player during the comparison period. It is easy to see the performance difference between players. However, when looking at expected goals comparison, observation is made that both KalPa players have played on the same level almost for the whole season.

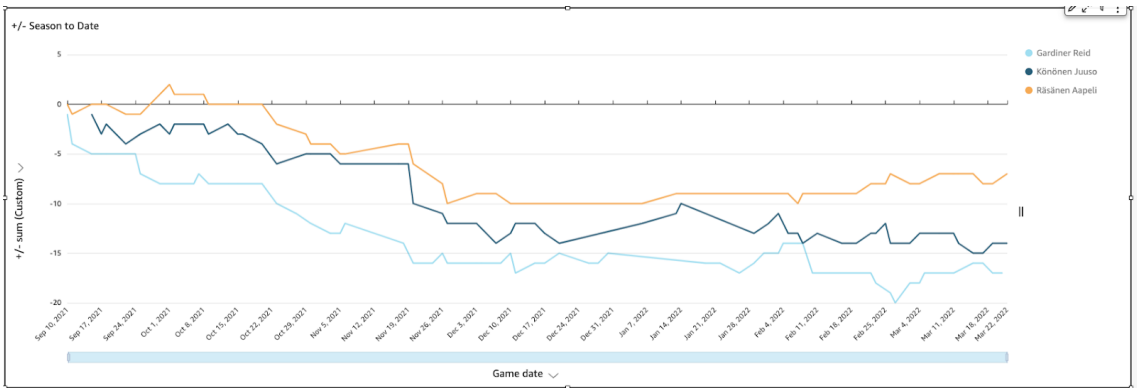


Picture 12. Shots resulting a goal.



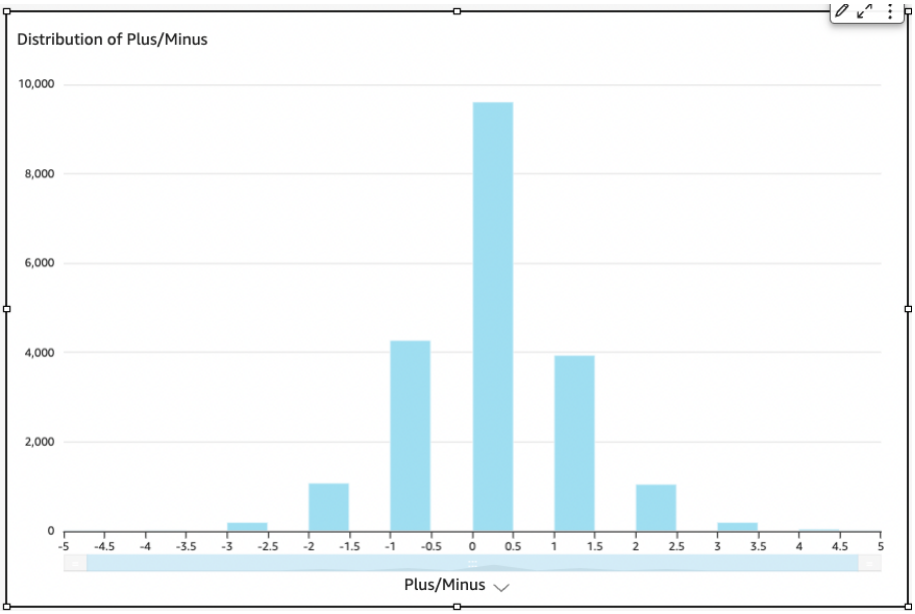
Picture 13. Expected goals.

Visualization of plus/minus points for the same players is a little bit different. As discussed earlier, plus/minus point is given for a player if his team makes goals, or if the opponent team makes goals when a player is on ice (Kohl, 2016). In this comparison observation is made that even if Gardiner has scored a large amount more goals than Räsänen, in the second half of the season their plus/minus statistics are on the same level (picture 14.) This indicates that it is just how many goals a player scores, but how he contributes to the whole team's performance, so what happens when the player is on ice. It is visible from the visual that the player's plus/minus statistic has been decreasing in the first half of the season and been more stable during the second half.



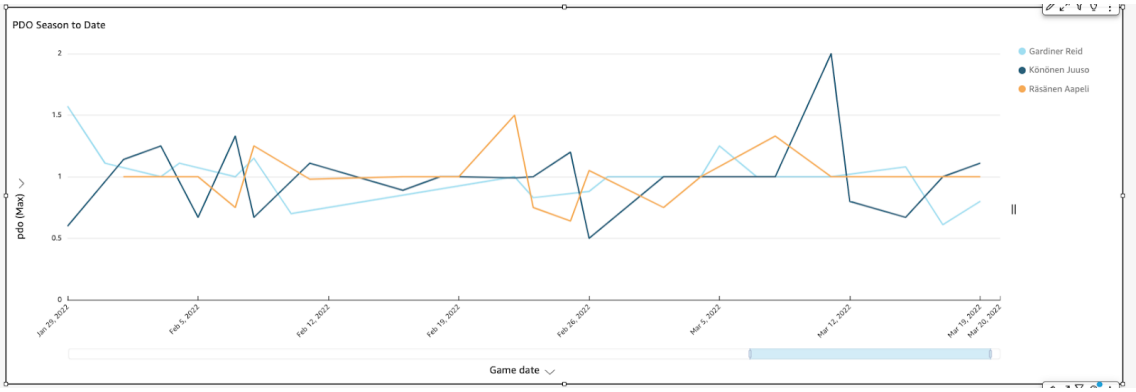
Picture 14. Plus/minus.

In this dataset plus/minus is a normally distributed having centre in 0 (picture 15.) This means that for each game data, most of the players play with zero plus/minus statistics, which gives an indication that performance is better or worse with players who are outside the centre. In a single game, higher plus/minus value can indicate better performance and negative value lower. Interestingly, when a sum is calculated over the season, the better players seem to have a negative plus/minus sum value. This can indicate that they also play more short-handed shifts, and that they simply are more on ice, so the likelihood of the opponent goals during their shifts is higher.

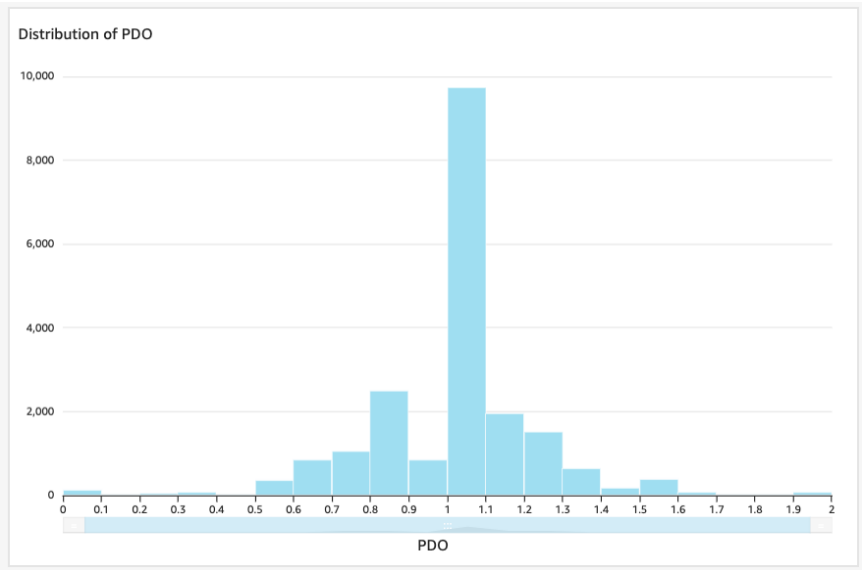


Picture 15. Distribution of plus/minus.

Picture 16 shows visualization of the same three players and their PDO during a season. As defined before, average players PDO should be around 1 (or 100 depending on the scale), above that is better performance and lower values indicate lower performance. This means that PDO is a statistic that is not feasible to use in sum calculation. It is not easy to see the differences over time, PDO is one of the statistics which would benefit from using machine learning based anomaly detection in detecting unusual performance patterns. Picture 16 has been zoomed to illustrate the second half of the season. When looking at how PDO values are distributed in the dataset, it can be observed that it is close to normal distribution, having centre in value 1 (picture 17.) This is aligned with the theory part. This also can indicate that better player PDO is more than 1 and lower performance players value is below 1.

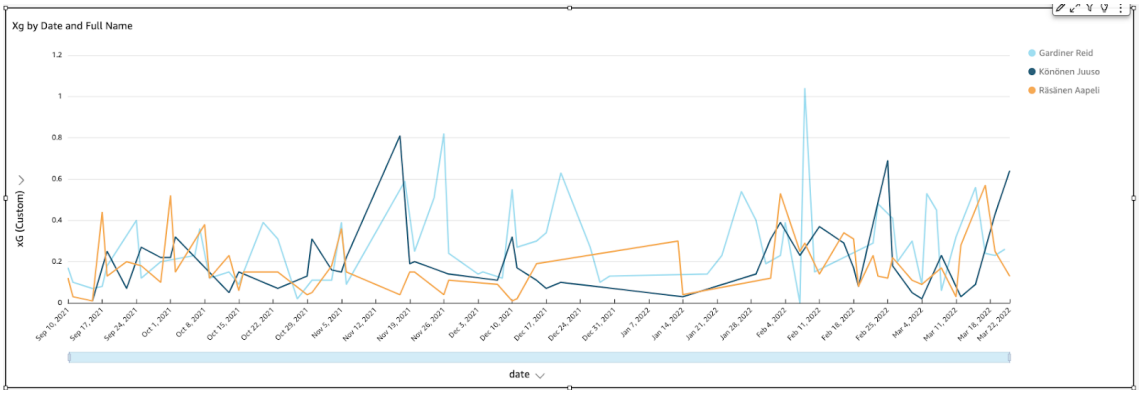


Picture 16. PDO.

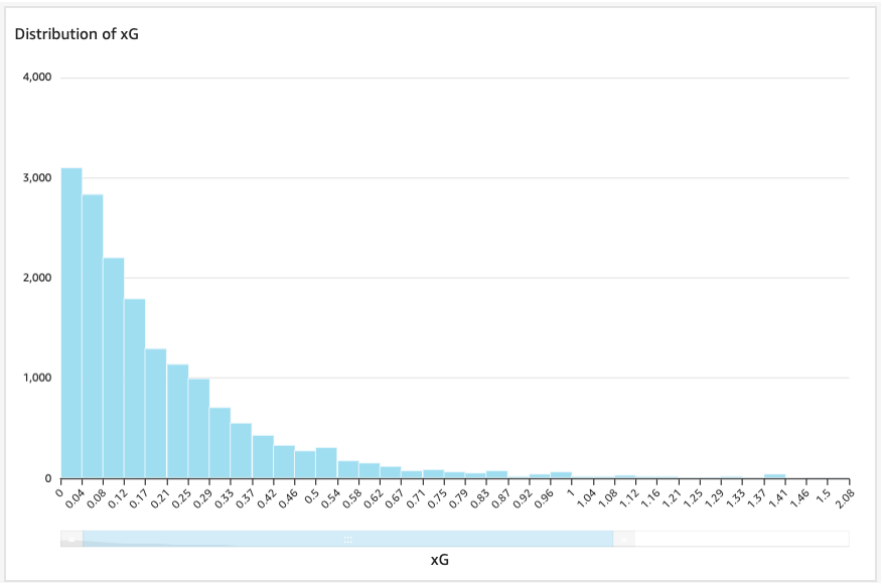


Picture 17. PDO distribution.

When visualizing the expected goals value (xG) over the season (picture 18), similar effect with PDO is seen. It is harder to spot how players performance changes during season, and this time series statistic could be used with anomaly detection. There are spikes in the data for players, which likely will be identified as anomalies. Anomaly detection can identify better or lower performance periods for a player. Using this with the sum over season statistics, and xG Rank discussed earlier can indicate players performance development direction. The distribution for xG in picture 19 is close to exponential cumulative distribution function (CDF), meaning most of the values in the dataset are close to zero, and probability of higher values follow the CDF, the probability approaches zero when the value of the x gets higher. Anomalies may be seen values higher or lower of the function value. From this analysis, it is also possible to see streaks of better and lower performance.

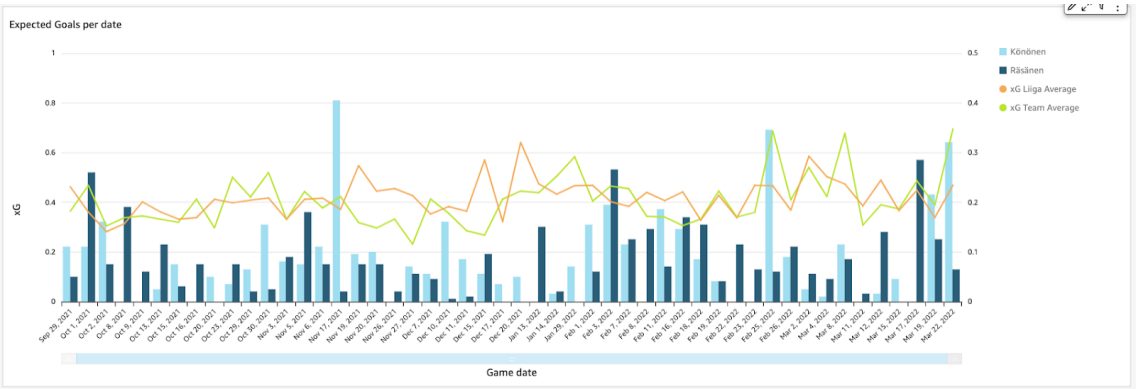


Picture 18. xG by game date.

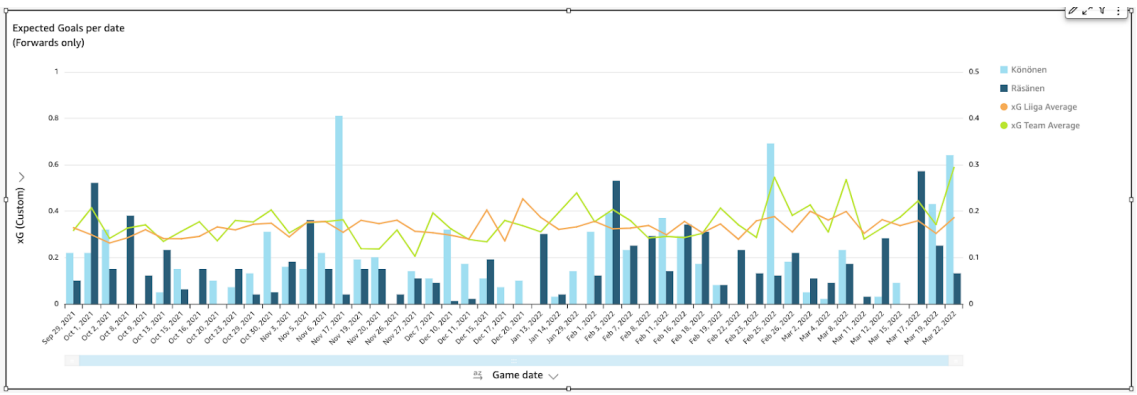


Picture 19. Distribution of xG.

Another way to get indication of performance is to compare a player against the reference group, for example own team. Picture 20 visualizes how Könönen and Räsänen xG values per game are situated in comparison to Liiga average and own team (KalPa) average on a given game day. Picture 21 takes into consideration players per role, it compares the players against forwards.

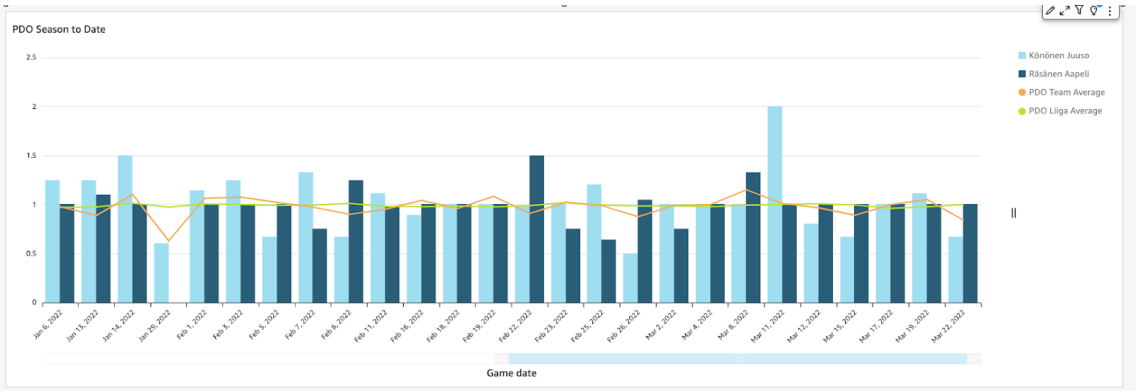


Picture 20. Expected goals comparison.



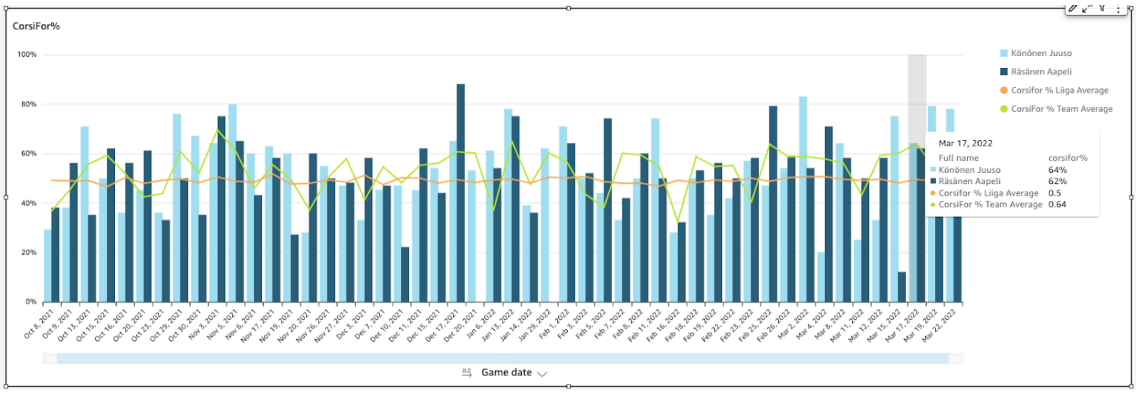
Picture 21. Expected goals comparison, forwards only.

Picture 22 does the same for PDO. In this analysis, it is observed that as PDO average is close to one there is not much value for calculating the average to different groups for comparison. From this analysis, some game dates are highlighted where these two players have been more lucky or unlucky.



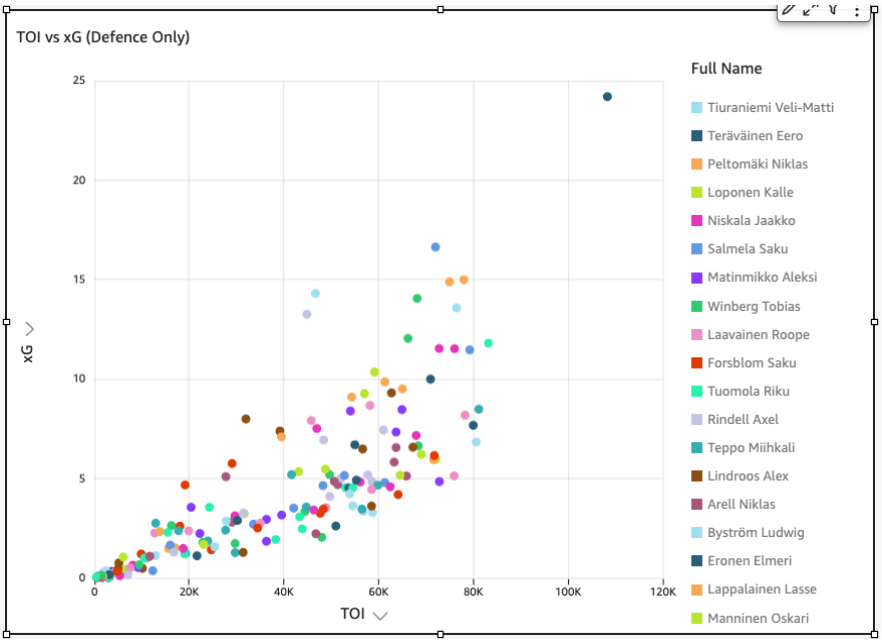
Picture 22. PDO comparison.

Corsi For Percentage (CF%) is the sum of shots on goal, missed shots, and blocked shots over the shots against, missed shots against and blocked shots against at equal strength evaluates a player's or teams puck possession. A CF% value of 45-55% is typical to a hockey player (Puckpedia). In other words, CF% values over 55% indicate better performance, and values below 45% lower performance. Picture 23 indicates CF% comparison for Könönen and Räsänen, as well as comparison against all players in Liiga, and all players in their own team. CF% value for all players in Liiga seems to be close to 50% during the whole regular season 2021-2022. We can see that CF% for Könönen and Räsänen varies above and below leagues and own teams average over the season.

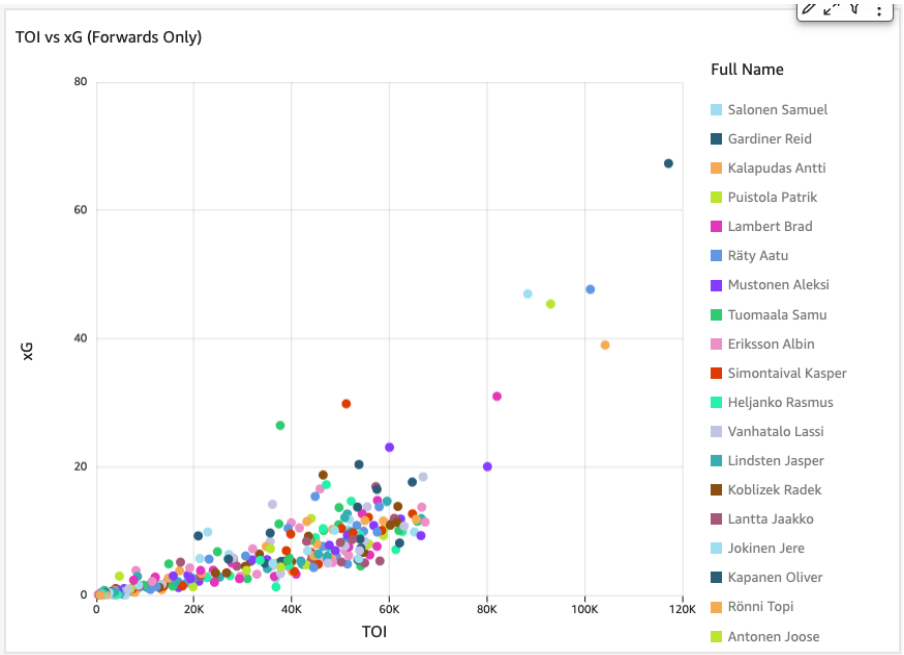


Picture 23. Corsi For%.

Yet another way to illustrate performance is to draw a scatter plot. Machine learning could be applied to find patterns and grouping. Picture 24 and 25 presents a sum of expected goals value in relation to time on ice. Picture 24 presents only players in defence roles and picture 25 the players in forward roles. Higher performing players are in the top right.



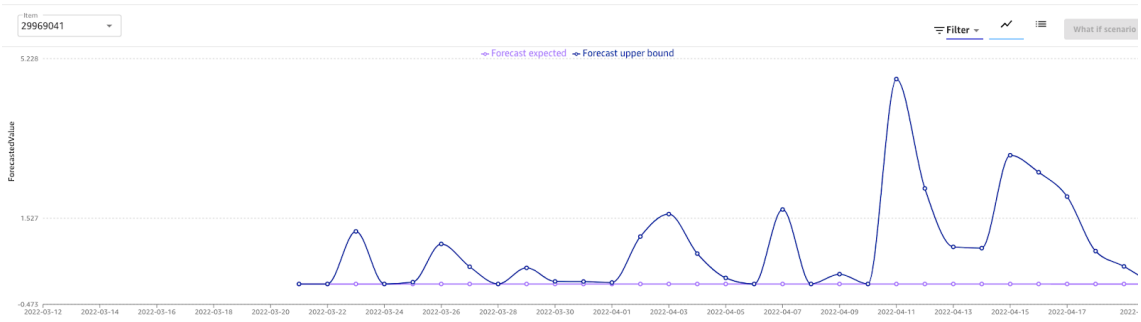
Picture 24. xG and TOI for defence.



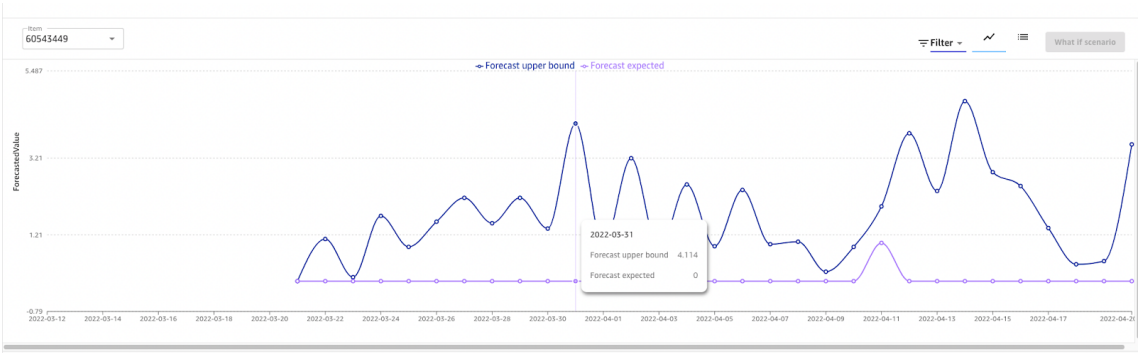
Picture 25. xG and TOI for forwards.

I trained a machine learning model with Amazon SageMaker Canvas autoML to make a time series prediction on Goals. I used an exported dataset from the data I had ingested in the data lake and let SageMaker decide which parameters are important to make such predictions. No further fine tuning for the model was made.

The experiment shows that I was able to make a prediction for 30 days into the future. As a test case, I used two players, Aapeli Räsänen (KalPa) (picture 26) and Reid Gardiner (HIFK) (picture 27.) KalPa did not play in playoffs during season 2022 so Räsänen prediction cannot be compared to the real goal amount. When looking at the prediction for Reid Gardiner, HIFK, and comparing the result to the 2022 playoffs the model predicted 1 goal for Gardiner, and he scored 2 goals during the playoffs (source Liiga.fi). As the data amount used in my thesis is limited to only regular season 2021-2022, and does not include data from previous seasons playoffs, it is not meaningful to continue to interpret the results with other players. The key takeaway is that training a model to create time series predictions is possible, and with tuning the model and with correct data, the model could give indicative predictions on players future performance. Furthermore, a model can be trained to create predictions for the regular season as well. To keep the scope of my thesis in control, I did not dive deep into this topic, and this alone would be an idea for another master's thesis, or even doctoral dissertation.



Picture 26. 30-day prediction for Aapeli Räsänen



Picture 27. 30-day prediction for Reid Gardiner

I trained SageMaker AutoML to predict performance for HockeyAllSvenskan players. I used full season statistics from SportContract API for Liiga from regular season 2021-2022 to train the model, and regular season 2021-2022 data for HockeyAllSvenskan players to create predictions.

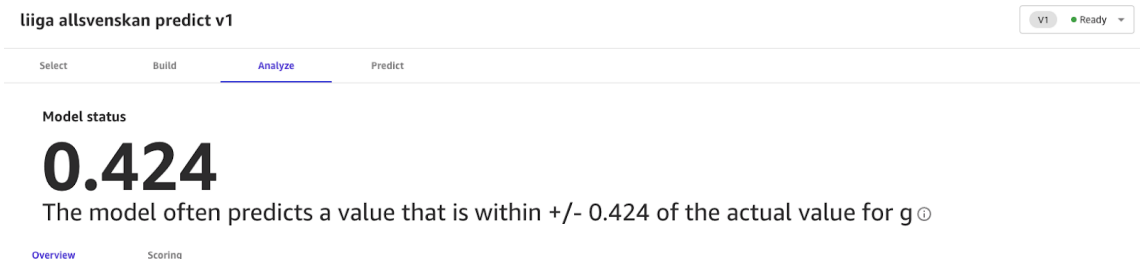
I used this dataset because the features for both datasets are equal, and one can be used to train a model and the other as testing the model. With the predictions we see that the model predicted 7.3 goals for Colby Sissons if he is playing in Liiga. In the regular season 2022-2023 Sissons had made 8 goals by Feb 5th, 2023. It should be noted that predictions were made with earlier season data and are not directly comparable to the current season. Sissons performance in Liiga can be on a higher level so cannot directly be compared. However, the prediction for this one case could be seen as an indicative prediction.

This test indicates following of findings:

- Dataset is too small to train the model for reliable prediction. Dataset should contain multiple seasons to have enough data points.
- With having only one reference point which can be used to check the accuracy against real results, it is impossible to evaluate the reliability. However, this could give indication how a player would perform in Liiga.
- The most important finding is that results directly from different leagues should and cannot be directly compared due to differences in playing styles, levels, and players, but patterns from the data can be identified in any of the leagues where similar data is available. These patterns, like anomalies, are not directly related to absolute data values and possibly can be used for comparing players from any league. By using machine learning in a similar way as I did in this experiment, would probably not give accurate prediction on real performance, but can give indication on the direction of a level of a player.

g	firstname	lastname	position	teamname	jerseynumber
0.6025893688201904	Romans	Semjonovs	FORWARD	Tingsryds AIF	25
9.245505332946777	Maksims	Semjonovs	FORWARD	AIK	27
6.985321044921875	Malte	Setkov	DEFENSEMAN	AIK	83
4.359381675720215	Hunter	Shinkaruk	FORWARD	HV71	12
9.36393928527832	Deven	Sideroff	FORWARD	Kristianstads IK	95
7.320669174194336	Colby	Sissons	DEFENSEMAN	HC Vita Hästen	22
8.925848007202148	Albert	Sjöberg	FORWARD	Södertälje SK	23
6.073116302490234	Carl-Johan	Sjögren	FORWARD	IF Troja-Ljungby	22
0.10184943675994873	Carl	Sjögren	FORWARD	HC Vita Hästen	25
7.496585369110107	Malte	Sjögren	FORWARD	Kristianstads IK	61
2.850355863571167	Victor	Sjöholm	DEFENSEMAN	HV71	5
0.04731038957834244	Hampus	Sjölund	FORWARD	Södertälje SK	40
8.476094245910645	Isac	Skedung	FORWARD	Västerås IK	40
9.314189910888672	Nikolai	Skladnichenko	FORWARD	Kristianstads IK	7
5.03934383392334	Colin	Smith	FORWARD	Södertälje SK	41
3.5695817470550537	Andreas	Söderberg	DEFENSEMAN	HV71	27
3.5695815086364746	Andreas	Söderberg	DEFENSEMAN	Almtuna IS	55
1.1543960571289062	Greg	Squires	FORWARD	Södertälje SK	26
2.889765739440918	Åke	Stakkestad	FORWARD	BIK Karlskoga	15
8.691965103149414	Ludwig	Stenlund	FORWARD	HC Vita Hästen	14

Table 8. Predicting players' scored goals.



Picture 28. Trained model statistics.

7 CONCLUSION AND FINDINGS

The aim and main objectives are to create both data and statistical models for analysing a hockey game data in relation to the defined indicators for identifying better than normal player performance. In addition to the creation of the models, the objective was to run the analysis with real life data.

Secondary objective was to identify relevant indicators which can be used to identify better than normal and exceptional performance of hockey players in their career by analysing the available game data.

I first did qualitative literature-based research to find commonly known metrics and statistics used in analysing hockey players performance. In the second part developed a data and statistical model for analysis, built an analysis pipeline and finally used real life data for Liiga and HockeyAllSvenskan to create example analysis. Data was exported from Wisehockey and SportContract API's. Analysis pipeline was built using Amazon Web Services data analytics capabilities. In addition, I studied the use of machine learning for the evaluation and predictions.

7.1 Findings

First research question was to find out how to use data and analytics to identify players whose performance development is better than normal and to identify the indicators, statistical model and data which can identify better than usual and exceptional players (outliers).

My research shows that by analysing game data from the matches teams can get insights about both teams and players performance. This thesis focused on individual players' performance. Traditional statistical analysis methods can be applied to get insights and information to support teams' development and decision making. By applying analytics, a team's management and coaches can find optimal combinations for their player roster. This can be optimized from a player budget perspective.

In addition to traditional statistics, machine learning can be applied to get insights. In this thesis, I build an example of anomaly detection using Random Forest Cut. With anomaly detection teams

can identify exceptional performance and outliers (anomalies) in players performance which could indicate better or worse than normal development and periods.

I then used time series data to build predictions for players' performance into the future. This could be an interesting approach, to predict how a player or a team will perform in the series of future games, such as in playoffs. This however requires enough data and context to give trustworthy predictions. With only one season worth of data predictions will not be accurate enough. This however could be a future research topic.

I studied using a machine learning model to see if I can train a model with target leagues data and predict another leagues players performance with it. Because the game and playing styles differ between leagues, it cannot give accurate predictions, however, can give indication how a player from another league would compare to another leagues level. Also, this topic be an interesting for future research.

Second research question was to find out what advanced metrics and statistics models are common in hockey, and will they fit for analysing Liiga and Allsvenskan game data and what additional data would be needed to use these metrics and models?

During my research I found out that there are numerous methods and advanced statistics commonly known which all try to evaluate and to predict performance or value. Many of these are targeted to evaluate performance on a team level and on individual player level. Some are complex and their value in analysis for Liiga and HockeyAllSvenskan data could not be proven within the scope of this thesis. Many of them require special types of data, such as NHL RTSS data, and are targeted to evaluate performance in the NHL and most of them were not available in the data sources accessible to me during my research. There are statistics metrics that can used to evaluate and predict performance in Liiga and HockeyAllSvenskan. These statistics include, for example, goals, points, shots, PDO, faceoffs won/lost, time on ice, plus/minus, assists, expected goals, Corsi and Fenwick.

I used a model which combines goals, expected goals, plus/minus, time on ice, Corsi and PDO. In addition to these I calculated an expected goals rank over the full season. This metrics gives a ranking for all the players in all the games during a season using xG sum of their shots in a game. This metric gives a single number view to players performance; it measures the offensive

pressure a player has when on ice compared to the reference group. As a reference group, whole Liiga, players own team or players role or a combination of the earlier can be used.

Third research question was to find out what data is available to support analysis and decision making which could be used as a source for automated data pipeline and continuous analysis?

My research shows that there are both publicly available and free web sites and commercial and non-public API sources where the data can be obtained. I used two APIs to fetch the data for analysis, Wisehockey and SportContract. Both are RESTful APIs and relatively easy to use programmatically with for example using Python and Python libraries and both offer simple documentation of the API contents and structures. Data from public web sites could be extracted by for example using data scraping.

7.2 Hypothesis testing

The hypothesis for the research is that by using commonly known ice hockey statistics to analyse game data and by using machine learning models, a professional hockey team's management, coaches and scouts can identify and predict players performance and use it to support their player recruitment decisions.

My research has indicated that data analytics is common in professional sports and becoming even more important and relevant in ice hockey. There are data which can be used for analysis, tools and methods are commonly known and available. In addition, this research shows that machine learning can help to identify unusual (better or worse than normal) player performance and may support to predict future performance. If a player's performance or potential is evaluated without analysing the supporting data, the view is always biased and subjective. This research shows that data analysis can play a key part in identifying players' performance and provides examples how the data can be used to predict their performance in the future. In leagues like the NHL, using data in the player recruitment process is common and multiple methods and statistics are available to support the decision making. Thus, the hypothesis is true.

7.3 Other findings

An additional finding of this research is related to comparison of players playing in different leagues. As there may be significant differences in the playing style, skill level, and rules, the data and performance between two leagues are not directly comparable, which means the data from the two can't be directly compared. When you find patterns from the data with for example machine learning, like anomalies or other metrics or performance identifiers (like the xG rank used in this thesis), this new data is not anymore directly related to the original data or the leagues where it originates. If generated by identical means, this new data is then naturally normalized and becomes comparable.

Openness of the data is also a finding worth mentioning. In NHL the data seems to be more easily and openly available to anyone who wishes to do analytics, for both professionals and hobbyists. If the data platforms would be more open it would allow data owners, data consumers and third parties to interact and contribute their feedback. (de Reuver, Ofe, Agahari, Abbas, Antragama, Zuiderwijk, 2022)

If data in these researched leagues would be more openly available, it could create both new business opportunities, make the data platforms better with help of the community and create opportunities for players and teams to utilize the data in more efficient ways.

The analysis case in thesis was a good learning experience and it covered the full lifecycle of analysis. It started with data engineering, had phase for data analysis and data science and finally ending into visualizing business insights and presenting them to the stakeholders.

8 DISCUSSION

8.1 Challenges, issues and resolutions

When evaluating the results against the data, I noticed that there are some inconsistencies with the datasets. This can be seen for example in the analysis of goals, the results show that Juuso Könönen has made 11 goals in the analysed time period. However, Liiga.fi statistics indicate that the actual number was 12. When I checked this from the exported raw dataset, it has 11 shots for Könönen which has resulted in a goal. This could mean that for example the problem appears during data extraction from Wisehockey API or error in the data source which then results in a completely missing data entry. One possibility for the error is that I have calculated the goals using shot data, any shot with the result of goal is counted as goals. But if there are goals recorded which did not originate from a shot, those are then not calculated. I checked this from Wisehockey API documentation, and the goal statistic references the shot, so this should be a correct way to calculate. I did not investigate this further, but it should be fixed before doing any real analysis with the data.

An issue I found with the summarized datasets was that because a player could have been playing in a match where they did not shoot any shots, and as result had missing values in goals and expected goals values. This has an effect when the average over a group of players is calculated and compared that to a player. This error in the dataset influences the averages, but as such does not prevent nor introduce significant bias on finding answers to the research questions. This error however in the data model should be fixed if a similar model would be taken into real analysis use. Easy fix could be to design and import the data into a relational database, where the data can be then easily queried and missing values altered. Because this error did not actually affect finding answers to the research questions and considering that data handling took much more effort and time than planned, I decided not to design relations and not to import the data from S3 based data lake to a database. Amazon Athena, which I used for building the analysis tables and views, is a flexible and easy tool for analysing and querying data on structured files. Its limitation however is that it does not allow to be able to directly alter the underlying data files.

Another solution could be to combine datasets in a transformation phase. This however would have been a step back in the process and would have required again too much work to handle within the scope of this thesis. Initially, I did not do this in the transformation phase because I had to explore the data and find out if I could answer the research questions with it.

Yet another challenge I faced is related to the data itself and getting the data from the API's. As mentioned earlier the two sources used both had RESTful API to access the data programmatically. SportContract API contents were easy to extract yet had very limited datasets. Wisehockey API had more data available, however advanced stats they offer via web interface were not among the data. API from Wisehockey was more complex and I needed to extract and handle, transform and query much more data which I in later phases needed to summarize and combine to get my hands into the wanted statistics, like goals and expected goals. The data was rather raw data. As I am not a professional programmer, I had a steep learning curve in coding to extract the data. The fact that I needed to do more coding work in exporting the data meant I needed to combine and summarize and calculate the features in the analysis phase. This turned out to be much more laborious than I expected in the thesis planning phase. I estimate that these two elements together introduced 3 - 4 weeks of more work than what was planned. To keep the thesis scope manageable, I needed to narrow down the statistics I used in the analysis and to do the machine learning part more superficially by only testing concepts and building some examples. However, this does not have a significant effect on the thesis results when assessing them against the research questions and goals. What it means is that I did not get to the analytics automation part as deeply as I wanted to.

I assumed the data sources are providing already cleaned and accurate data which can be ingested into analytics without too much preparation, cleaning, and transformation, and that the data sets are clean and complete. This was not the case for all parts. This challenge was highlighted by the fact that I needed to summarize and calculate statistics from raw data and structures of the exported data. This leads to more work and questions with for example missing value handling, duplicate data entries showing after combining and summarising. This problem was discussed earlier in the thesis and the easiest solution to avoid such difficulties is to use a database where data can be altered as needed. This caused some confusion and extra work but does not provide a significant negative impact on thesis results. Even if the ingested data looks complete, the need for cleaning, supplementing and imputing values to the dataset in the analysis phase should be considered.

Other things to consider with such comprehensive thesis topic and with doing thesis part time along with full time day job, is the thought work it requires. Catching up from the previous day's work, doing a small piece of work, saving all work done and then wrapping up to start same process again the next night is not very effective way. This introduced some limitations on the analysis scope.

8.2 Data and results reliability

Neither of the sources publish publicly the information about the error margins and reliability of the raw data, metrics, and statistics nor how specifically the specific metric is collected or calculated. There has been debate in the public news about the data trustworthiness and inaccuracies in the metrics and the data in Wisehockey platform (MTV Uutiset). This news article was published 16.10.2022 and I exported the data from the API 14.11.2022. My research indicates some inconsistencies which could be caused by the data source but also by a process used for data extraction. In addition, both used data sources limit the statistics available through their API, which means a solution must be developed to calculate these statistics with raw data if those will be used in the analysis. This introduces error possibility which could have a significant effect on the results and in the worst-case lead to wrong decisions.

This means that the good as the analytics capabilities in the platform are, the results should be seen as indicative until source data correctness and trustworthiness can be confirmed. This also means that teams should continue to use subject matter expertise to back up their decision making.

8.3 Future research possibilities

During my research I identified several additional questions to be researched which I could not answer, or they were not in scope for this thesis. These could be for example studied in other thesis projects or used as an idea for doctoral studies and dissertation. Potential questions and issues are:

1. Would hockey leagues like Liiga benefit from more open data strategy and would it provide opportunities for teams and players?

2. How to apply novel methods such as deep learning, neural networks, and computer vision in hockey analytics?
3. Developing and using machine learning models to identify performance in near real time during a match and using video stream analytics.
4. More in-depth research on using machine learning and anomaly detection-based predictions.
5. Using machine learning to evaluate and predict performance between different leagues.
6. Using machine learning to predict season players and teams' performance in ongoing season and use the model to build optimal team compositions.

9 ACKNOWLEDGEMENTS

I would like to acknowledge and give my warmest thanks to my thesis supervisor Dr. Teppo Räisänen. With his guidance I was able to successfully go through all the stages of writing my thesis. I would also like to give special thanks to Anssi Laine and Atte Sacklin and the whole KalPa Hockey team for providing this opportunity to research this very interesting topic.

Special thanks to Dalton Linkus and Janne Kekäläinen from Nashville Predators for their views and insights and to Sami Knuutinen for tipping me this opportunity.

REFERENCES

Amazon Web Services Documentation. https://docs.aws.amazon.com/index.html

Brower, Owen 2020. Applying Analytics to the NFL Draft: Athletic Performance Measures as a Predictor of Future Success. Economics Theses. Search date: 6.9.2022. https://soundideas.pugetsound.edu/economics_theses/111

bgred105, 2012. Sab(re)metrics: What is GVT? Search date: 16.10.2022. https://www.diebytheblade.com/2012/12/7/3411686/sab-re-metrics-what-is-gvt

Bitwise Oy. Wisechockey. Search date: 22.2.2023. https://bitwise.fi/

Burtch, Steve. 2014. What is dCorsi and how does it help us account for "context"? Search Date: 12.10.2022. https://www.pensionplanpuppets.com/2014/7/21/5663440/dcorsi-introduction-what-is-dcorsi-how-do-you-use-dcorsi

Burtch, Steve. 2014. Delta Corsi and Assessment of Individual Player Impacts on Possession Accounting for Usage. https://www.dropbox.com/s/wb10dq5mlpv0uzz/dCorsi-WRITEUP.pdf

Cambridge Dictionary. Search date: 9.9.2022. https://dictionary.cambridge.org/dictionary/english/data-analysis

de Reuver, Mark & Ofe, Hosea Ayaba & Agahari, Wireman & Abbas, Antragama Ewa & Zuiderwijk, Anneke. 2022. The Openness of Data Platforms: A Research Agenda. Search date: 2.3.2023. https://www.researchgate.net/publication/365122960_The_Openness_of_Data_Platforms_A_Research_Agenda

Gautam, Akshay 2021. 6 Phases Of Data Analysis According To Google. Search date: 9.9.2022. https://medium.com/codex/6-phases-of-data-analysis-according-to-google-9e084b89f848

CLVNNG, 2017. Calculating Hockey Analytic's GVT. Search date: 12.10.2022. https://performancedrivenanalytics.wordpress.com/2017/01/04/34/

Found, Rob. 2016. Goal-based Metrics Better Than Shot-based Metrics at Predicting Hockey Success. Search date: 17.12.2022. https://www.researchgate.net/publication/325302195_Goal-based_Metrics_Better_Than_Shot-based_Metrics_at_Predicting_Hockey_Success US Sports Academy.

Halper, Fern, 2021. Unified Platforms for Modern Analytics. TDWI. https://tdwi.org/research/2021/09/arch-all-best-practices-report-unified-platforms-for-modern-analytics.aspx?tc=page0

Jahn K. Hakes & Raymond D. Sauer 2006. An Economic Evaluation of the Moneyball Hypothesis, Journal of Economic Perspectives, American Economic Association, vol. 20(3), pages 173-186, Summer. Search date: 8.9.2022- https://ideas.repec.org/a/aea/jecper/v20y2006i3p173-186.html

Jarett, Leigha & Kuo Micheal, 2021. Creating a unified analytics platform for digital natives. Google. Search date 5.2.2023. https://cloud.google.com/blog/topics/developers-practitioners/creating-unified-analytics-platform-digital-natives

Jones, Wayne. What does plus minus (+/-) mean in hockey? (with stats). Search date: 10.10.2022. https://hockeyanswered.com/what-does-plus-minus-mean-in-hockey-with-stats/

Joshua Weissbock & Herna Viktor & Diana Inkpen. Use of Performance Metrics to Forecast Success in the National Hockey League. University of Ottawa, Ottawa, Canada.

Julie, Kathleen 2014. The Many Ways the NHL Is Using Data Collection This Season to Enhance the League. Search date: 2.9.2022. https://www.sporttechie.com/the-many-ways-the-nhl-is-using-data-collection-this-season-to-enhance-the-league

Junior KalPa web page. Search date: 2.9.2022. https://kalpa.fi/en-gb/article/etusivu/junior-kalpa/43/

KalPa Hockey web page. Search date: 2.9.2022. https://kalpa.fi/?lang=en-gb&langmenu=1

Kohl, Garret, 2016. Behind the Numbers: Why Plus/Minus is the worst statistic in hockey and should be abolished. Search date: 5.10.2022. https://hockey-graphs.com/2016/11/01/behind-the-numbers-why-plusminus-is-the-worst-statistic-in-hockey-and-should-be-abolished/

Kohonen, Iina & Kuula-Luumi, Arja & Spoof, Sanna-Kaisa 2019. The ethical principles of research with human participants and ethical review in the human sciences in Finland Finnish National Board on Research Integrity TENK guidelines. Search date: 15.8.2022. https://tenk.fi/sites/default/files/2021-01/Ethical_review_in_human_sciences_2020.pdf

Kontsas, Jukka & Lehtola, Juha 2014. Goalie and scoring analysis: MOL, Mestis and Liiga. Haaga-Helia University of Applied Sciences. Search date: 5.9.2022. https://www.theseus.fi/bitstream/handle/10024/78338/Kontsas%20Lehtola%20Final.pdf?sequence=1

Kubatko, Justin. 2011. Calculating Point Shares. Search date: 11.10.2022. https://www.hockey-reference.com/about/point_shares.html

Larkin P, O'Connor D 2017. Talent identification and recruitment in youth soccer: Recruiter's perceptions of the key attributes for player recruitment. PLoS ONE. Search date: 5.9.2022 https://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0175716&type=printable

Lee, Chirstian, 2020. Advanced Hockey Stats 101: Corsi. Search date: 11.10.2022. https://medium.com/hockey-stats/advanced-hockey-stats-101-corsi-part-1-of-4-29d0a9fb1f95

Lee, Chirstian, 2020. Advanced Hockey Stats 101: Fenwick. Search date: 11.10.2022. https://medium.com/hockey-stats/advanced-hockey-stats-101-fenwick-part-2-of-4-a8796b8188d

Lee, Chirstian, 2021. Advanced Hockey Stats 101: PDO. Search date: 11.10.2022. https://medium.com/hockey-stats/advanced-hockey-stats-101-pdo-part-3-of-4-d3f319f2e1f1

Lee, Chirstian, 2021. Advanced Hockey Stats 101: Zone Starts. Search date: 11.10.2022. https://medium.com/hockey-stats/advanced-hockey-stats-101-zone-starts-part-4-of-4-1548b894541c

Liu, Guiliang & Schulte, Oliver 2018. Deep Reinforcement Learning in Ice Hockey for Context-Aware Player Evaluation. Simon Fraser University, Burnaby, Canada. Search date: 7.9.2022. https://arxiv.org/pdf/1805.11088.pdf

Lukas, Allison. 2021. Analytics with Alison: Expected Goals. Search date: 16.11.2022 https://www.nhl.com/kraken/news/analytics-with-alison-expected-goals/c-327728890 Seattle Kraken.

Macdonald, Brian 2011. A Regression-Based Adjusted Plus-Minus Statistic for NHL Players, Journal of Quantitative Analysis in Sports: Vol. 7: Iss. 3, Article 4. Search date: 9.9.2022. http://hockeyanalytics.com/Research_files/Regression_Based_Plus_Minus.pdf

MacDonald, Brian. 2012. An Expected Goals Model for Evaluating NHL Teams and Players. http://www.hockeyanalytics.com/Research_files/NHL-Expected-Goals-Brian-Macdonald.pdf MIT Sloan Sports Analytics Conference 2012.

Mondello, Michael & Kamke, Christopher 2014. The Introduction and Application of Sports Analytics in Professional Sport Organizations. Journal of Applied Sport Management: Vol. 6 : Iss. 2. Search date: 3.9.2022. https://trace.tennessee.edu/cgi/viewcontent.cgi?article=1357&context=jasm

MTV Uutiset. 2021. SM-liigan tilastointi huippuluotsien hampaissa – pelkästään tämä esimerkki kertoo raa'asta epätarkkuudesta: "Aikamoisia numeroita". Search date: 12.2.2023. https://www.mtvuutiset.fi/artikkeli/sm-liigan-tilastointi-huippuluotsien-hampaissa-pelkastaan-tama-esimerkki-kertoo-raa-asta-epatarkkuudesta-aikamoisia-numeroita/8549030

Paige, Robert, 2021. What's the Difference Between a Data Lakehouse and a Unified Analyt-ics Platform? Architecture & Governance Magazine. Search date 5.2.2023. https://www.architectureandgovernance.com/digital-transformation/whats-the-difference-between-a-data-lakehouse-and-a-unified-analytics-platform/

Pallotta, Jack 2021. Goals for per 60: The Best Of NHL Statistics. Search date: 14.9.2022.

Piccolo, Nick 2022. Inside the Stats: Corsi, Fenwick and On-Ice Stats. Search date: 14.9.2022. https://insidetherink.com/inside-the-stats-corsi-fenwick-and-on-ice-stats/

Pierce, Susan & Tekiner, Firat, 2021. Building a unified analytics data platform on Google Cloud,. Search date 5.2.2023. Google. https://cloud.google.com/blog/products/data-analytics/building-unified-analytics-data-platform-google-cloud

Pettigrew, Stephen 2015. Assessing the offensive productivity of NHL players using in-game win probabilities. Harvard University. Search date: 6.9.2022. https://global-uploads.webflow.com/5f1af76ed86d6771ad48324b/5fee098cacad0134c9e8d403_1028_rppaper_pettigrew%20sloan%20submission.pdf

Praful Kava and Changbin Gong. 2020. AWS serverless data analytics pipeline reference architecture. Search date: 1.11.2022. https://aws.amazon.com/blogs/big-data/aws-serverless-data-analytics-pipeline-reference-architecture/ Amazon.

Puckpedia. NHL Player Stats. Search datę: 12.1.2023. https://puckpedia.com/stats#:~:text=Corsi%20For%20Percentage%20(CF%25),shots%20against%20at%20equal%20strength

Raber, Matthew & Eisenberg, Daniel. Defensive Efficiency Metrics (DEMs): A Paradigm Shift in Defensive Hockey Analysis Based on Per-Possession Measurements. Big Data Cup.

Ryder, Alan. 2008. Ten Laws of Hockey Analytics. Search date: 10.10.2022. http://hockeyanalytics.com/2008/01/the-ten-laws-of-hockey-analytics/

Sans, C., Carlsson, N., Lambrix, P, 2019. Player impact measures for scoring in ice hockey, In Karlis, D., Ntzoufras, I., Drikos, S. (eds), Proceedings of MathSport International 2019 Conference, pp. 307-317

Sarangam, Ajay 2022. Difference Between Analysis And Analytics | Which Is Better? Search date: 8.9.2022. https://www.jigsawacademy.com/blogs/business-analytics/analysis-vs-analytics/

Schuckers, Michael E. & Argeris, Steven 2015 . You can beat the “market”: Estimating the ROI on NHL scouting. Journal of Sports Analytics 1 (2015) 111–119. Search date: 4.9.2022. https://content.iospress.com/download/journal-of-sports-analytics/jsa0015?id=journal-of-sports-analytics/jsa0015

Schulte, Oliver & Zhao, Zeyu 2017. Apples-to-Apples: Clustering and Ranking NHL Players Using Location Information and Scoring Impact. School of Computing Science, Simon Fraser University. Search date: 4.9.2022. https://www.cs.sfu.ca/~oschulte/files/pubs/sloan-fix.pdf

Schuckers, Michael & Curro, James. 2013. Total Hockey Rating (THoR): A comprehensive statistical rating of National Hockey League forwards and defensemen based upon all on-ice events. https://www.statsportsconsulting.com/wp-content/uploads/Schuckers_Curro_MIT_Sloan_THoR.pdf Lawrence University, Iowa State University, Statistical Sports Consultancy LLC.

SportContract. Search date: 22.2.2023. https://sportcontract.net/#services

SportContract API documentation. Search date: Multiple between November 2022 - February 2023. https://api.sportcontract.net/api-docs/

Sudipto Guha, Nina Mishra, Gourav Roy,Okke Schrijvers. 2016. Robust Random Cut Forest Based Anomaly Detection On Streams. Amazon.

Thomas, G, 2018. Hockey Analytics: Ditching Plus Minus. Search date: 10.10.2022. https://puckprose.com/2018/07/30/hockey-analytics-ditching-plus-minus/

Tran, Bill 2021. How analytics and big data are changing ice-hockey. Search date: 3.9.2022. https://thewincolumn.ca/2021/09/20/how-analytics-and-big-data-are-changing-ice-hockey/

Turtoro CJ, 2014. Catch-all Statistics Part I: GVT versus Point Shares. Search date: 12.10.2022. https://www.allaboutthejersey.com/2014/8/8/5978425/catch-all-statistics-part-i-gvt-versus-point-shares

Vik, Jon & Shih, Min-Chun & Jansher, Rabnawaz & Carlsson, Niklas & Lambrix, Patrick. 2021. Not all goals are equally important - a study for the NHL. Linkoping University. Search date: 6.9.2022.https://www.researchgate.net/publication/285020018_A_Markov_Model_for_Hockey_Manpower_Differential_and_Win_Probability_Added

Vollman, Rob & Fyffe, Iain & Awad, Tom. 2016 The Ultimate Guide to Hockey Analytics. p. 13-19. Hockey Abstract. ECW Press.

What is Random Cut Forest? Amazon. Search date: 2.3.2023. https://docs.aws.amazon.com/quicksight/latest/user/what-is-random-cut-forest.html

Wikipedia. PDO. Search date: 4.1.2023. https://en.wikipedia.org/wiki/Analytics_(ice_hockey)#PDO

Wikipedia. Liiga. Search date: 25.2.2023. https://fi.wikipedia.org/wiki/J%C3%A4%C3%A4kiekon_SM-liiga

Wikipedia. HockeyAllSvenskan. Search date: 25.2.2023. https://en.wikipedia.org/wiki/HockeyAllsvenskan

Wikipedia. Data Scraping. Search date: 25.2.2023. https://en.wikipedia.org/wiki/Data_scraping

Wisehockey API documentation, Wisehockey API Automated real-time statistics for third party services 2021-09-16 V.2.1, Bitwise Oy.

Yost, Travis. 2012. Player Analysis: Goals Versus Threshold. Search date: 11.10.2022. https://www.hockeybuzz.com/blog/Travis-Yost/Player-Analysis-Goals-Versus-Threshold/134/45704

APPENDICES

Table and view structure appendix 1

Jupyter notebook for data extraction appendix 2

TABLE AND VIEW STRUCTURE APPENDIX 1

Liiga_2021_analysis_view_v1_1

1	playerid	int
2	lastname	string
3	firstname	string
4	jersey	int
5	role	string
6	home/away	string
7	matchid	int
8	timeonice	double
9	plus	int
10	minus	int
11	plusminus	int
12	corsifor	int
13	cosrsiagainst	int
14	corsifor%	double
15	pdo	double
16	team	string
17	expectedgoals	double
18	goals	bigint
19	date	date

2021allplayersandteams

1	playerid	int
2	firstname	string
3	lastname	string
4	jersey	int
5	role	string
6	teamid	int
7	teamshortname	string

2021allshots

1	team	string
2	period	bigint
3	result	string
4	shooter	bigint
5	expectedgoals	double
6	teamstrength.type	string
7	matchid	bigint

2021allteams

1	teamid	int
2	teamname	string
3	teamshortname	string

2021playerstats

1	numberofmatchesplayed	int
2	playerid	int
3	skatingstatistics.topspeed	double
4	skatingstatistics.timeonice	double
5	skatingstatistics.distancetravelled	double
6	skatingstatistics.averagespeedwithpuck	double

7	skatingstatistics.accelerations	int
8	skatingstatistics.decelerations	int
9	faceoffstatistics.faceoffwins	int
10	faceoffstatistics.faceoffcount	int
11	plusminusstatistics.plus	int
12	plusminusstatistics.minus	int
13	plusminusstatistics.total	int
14	shotstatistics.shots	int
15	shotstatistics.goals	int
16	passstatistics.successfulpasses	int
17	passstatistics.allpasses	int
18	passstatistics.receivedpasses	int
19	passstatistics.totalpassdistance	double
20	passstatistics.forwardpassdistance	double

v_player_team_count

1	teams played	bigint
2	playerid	int
3	lastname	string

2021allmatch

1	matchid	int
2	matchdate	string
3	status	string
4	homegoals	int
5	awaygoals	int
6	hometeam.id	int
7	hometeam.fullname	string
8	hometeam.shortname	string
9	awayteam.id	int
10	awayteam.fullname	string
11	awayteam.shortname	string
12	venue.name	string
13	venue.city	string

2021allmatchplayerstat

1	playerid	int
2	homeaway	string
3	totalstatistics.skatingstatistics.topspeed	double
4	totalstatistics.skatingstatistics.timeonice	double
5	totalstatistics.skatingstatistics.distancetravelled	double
6	totalstatistics.skatingstatistics.averagespeedwithpuck	double
7	totalstatistics.skatingstatistics.accelerations	int
8	totalstatistics.skatingstatistics.decelerations	int
9	totalstatistics.shiftstatisticssummary.shifts	int
10	totalstatistics.shiftstatisticssummary.averageshiftduration	double
11	totalstatistics.passstatistics.successfulpasses	int
12	totalstatistics.passstatistics.allpasses	int
13	totalstatistics.passstatistics.receivedpasses	int
14	totalstatistics.passstatistics.totalpassdistance	double
15	totalstatistics.passstatistics.forwardpassdistance	double
16	totalstatistics.plusminusstatistics.plus	int
17	totalstatistics.plusminusstatistics.minus	int
18	totalstatistics.plusminusstatistics.total	int
19	totalstatistics.puckcontrolstatistics.puckcontroltime	double

20	totalstatistics.puckcontrolstatistics.puckcontroldistance	double
21	totalstatistics.puckcontrolstatistics.puckcontrolforwarddistance	double
22	totalstatistics.puckconteststatistics.puckcontestswon	int
23	totalstatistics.puckconteststatistics.puckcontestslost	int
24	totalstatistics.puckconteststatistics.puckcontestparticipations	int
25	totalstatistics.shotscreenstatistics.offensivescreens	int
26	totalstatistics.shotscreenstatistics.blockedshots	int
27	totalstatistics.traditionalstatistics.corsifor	int
28	totalstatistics.traditionalstatistics.corsiagainst	int
29	totalstatistics.traditionalstatistics.corsiforpercentage	double
30	totalstatistics.traditionalstatistics.relativecorsiforpercentage	double
31	totalstatistics.traditionalstatistics.fenwickforpercentage	double
32	totalstatistics.traditionalstatistics.relativefenwickforpercentage	double
33	totalstatistics.traditionalstatistics.pdo	double
34	totalstatistics.traditionalstatistics.fullstrengthsavepercentage	double
35	totalstatistics.traditionalstatistics.fullstrengthshootingpercentage	double
36	matchid	int

2021allplayers

1	playerid	int
2	firstname	string
3	lastname	string
4	jersey	int
5	role	string
6	teamid	int

liiga2021allgamesgoals

1	playerid	bigint
2	goals	bigint
3	matchid	bigint
4	lastname	string
5	firstname	string

rs2021playerswithteamsxg

1	playerid	bigint
2	lastname	string
3	firstname	string
4	xg	double
5	matchid	bigint
6	jersey	int
7	role	string
8	teamshortname	string

runkosarja2021allshotsxg

1	shooter	bigint
2	xg	double
3	matchid	bigint

sc-liiga2021-skaters

1	id	string
2	firstname	string
3	lastname	string
4	position	string
5	teamid	string

6	teamname	string
7	jerseynumber	double
8	gp	double
9	g	double
10	a	double
11	p	double
12	pm	double
13	plusminus	double
14	pp	double
15	sh	double
16	as	double
17	toipergame	double
18	fow	double
19	fol	double
20	fo_perc	double
21	bls	double
22	sog	double
23	a1	double
24	a2	double
25	ppg	double
26	shg	double

sc-allsvenskan2021-skaters

1	id	string
2	firstname	string
3	lastname	string
4	position	string
5	teamid	string
6	teamname	string
7	jerseynumber	double
8	gp	double
9	g	double
10	a	double
11	p	double
12	pm	double
13	plusminus	double
14	pp	double
15	sh	double
16	as	double
17	toipergame	double
18	fow	double
19	fol	double
20	fo_perc	double
21	bls	double
22	sog	double
23	a1	double
24	a2	double
25	ppg	double
26	shg	double

```
import requests
from requests.structures import CaseInsensitiveDict
headers = CaseInsensitiveDict()
headers["Accept"] = "application/json"
headers["X-API-KEY"]= "xxxxxxx"
import requests
import json

url_leagues = "https://api.sportcontract.net/leagues"
print (url_leagues)
resp_leagues = requests.get(url_leagues, headers=headers).json()
print(resp_leagues)

resp
type(resp_leagues)
from pandas import json_normalize
df_league=json_normalize(resp_leagues)

df_league
import requests
from requests.structures import CaseInsensitiveDict
headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
parameters["leagueId"] = "5f4fa8e5e933ac773ac51bfd"
parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["X-API-KEY"]= "xxxx"
import requests
url_players = "https://api.sportcontract.net/players"
print (url_players)
print (parameters)
print (headers)
resp_players = requests.get(url_players, params=parameters, headers=headers).json()
from pandas import json_normalize
df_players=json_normalize(resp_players)
df_players.head()
import requests
from requests.structures import CaseInsensitiveDict
headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
#parameters["leagueId"] = "5f4fa8e5e933ac773ac51bfd"
parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["X-API-KEY"]= "xxxx"

import requests
url_games = "https://api.sportcontract.net/leagues/5f4fa8e5e933ac773ac51bfd/games"
print (url_games)
print (parameters)
print (headers)
resp_games = requests.get(url_games, params=parameters, headers=headers).json()
from pandas import json_normalize
df_games=json_normalize(resp_games)
df_games.head()

print(df_games.iloc[0, 0])
```

```
import requests
from requests.structures import CaseInsensitiveDict
headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
parameters["leagueId"] = "5f4fa8e5e933ac773ac51bfd"
parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["X-API-KEY"]= "xxxx"

import requests
url_game =
"https://api.sportcontract.net/games/55c8d10fa7a84b636d94ad61;55c8d10fa7a84b636d94ad57;2022-05-
04"
print (url_game)
print (parameters)
print (headers)
resp_game = requests.get(url_game, params=parameters, headers=headers).json()
from pandas import json_normalize
df_game=json_normalize(resp_game)
df_game#.head()

df_game.to_csv('//xxx/My Drive/OAMK/Thesis/code/game_Example.csv')
df_game.to_json('//xxx/My Drive/OAMK/Thesis/code/game_Example.json')

import requests
from requests.structures import CaseInsensitiveDict
headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
parameters["leagueId"] = "5f4fa8e5e933ac773ac51bfd"
parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["X-API-KEY"]= "xxxx"
import requests
url_skater= "https://api.sportcontract.net/playerstats/skater"
print (url_skater)
print (parameters)
print (headers)
resp_skater = requests.get(url_skater, params=parameters, headers=headers).json()
from pandas import json_normalize
df_skater=json_normalize(resp_skater)
df_skater.head(10)
df_skater.to_csv('//xxx/My Drive/OAMK/Thesis/code/skater_allsvenskan_2021.csv')
df_skater.to_json('//xxx/My Drive/OAMK/Thesis/code/skater_allsvenskan_2021.json')
import requests
from requests.structures import CaseInsensitiveDict
import requests

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
parameters["leagueId"] = "5f4fa8e5e933ac773ac51bfd"
parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["X-API-KEY"]= " xxxx"

url_skater= "https://api.sportcontract.net/playerstats/skater"
print (url_skater)
print (parameters)
print (headers)
resp_skater = requests.get(url_skater, params=parameters, headers=headers).json()
from pandas import json_normalize
```

```
df_skater=json_normalize(resp_skater)
df_skater.head(10)
df_skater.to_csv('//xxx/My Drive/OAMK/Thesis/code/allsvenskan/skater_hasvn_2021.csv')
df_skater.to_json('//xxx/My Drive/OAMK/Thesis/code/allsvenskan/skater_hasvn_2021.json')

# LIIGA and Wisehockey from this point on
from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import requests
import matplotlib

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= " xxxx"

url= "https://api.wisehockey.com/v4.1/tournaments"
print (parameters)
print (headers)
url_resp= requests.get(url, headers=headers).json()
print(url_resp)
from pandas import json_normalize
url_resp
df_tournaments=pd.DataFrame(url_resp)

df_tournaments.to_csv('//xxx/My Drive/OAMK/Thesis/code/tournaments_wisehockey.csv')
df_tournaments.to_json('//xxx/My Drive/OAMK/Thesis/code/tournaments_wisehockey.json')

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
#parameters["leagueId"] = "55c8c0c9a7a84b636d8843e1"
#parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxxx"

url= "https://api.wisehockey.com/v4.1/tournaments"
print (parameters)
print (headers)
url_resp= requests.get(url, headers=headers).json()
print(url_resp)
from pandas import json_normalize
url_resp
df_tournaments=pd.DataFrame(url_resp)

print (df_tournaments)

df_tournaments.to_csv('//xxx/My Drive/OAMK/Thesis/code/tournaments_wisehockey.csv')
df_tournaments.to_json('//xxx/My Drive/OAMK/Thesis/code/tournaments_wisehockey.json')
#fetch all team for tournament id 31
#from urllib.parse import SplitResult
#from cv2 import split
f#rom sqlalchemy import false


headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
headers["Accept"] = "application/json"
```

```
headers["Wisehockey-Api-Key"]= "xxxx"

url= "https://api.wisehockey.com/v4.1/tournaments/31/teams"
url_resp= requests.get(url, headers=headers).json()
url_resp
df_teams=pd.DataFrame(url_resp)

#print (df_teams)
#df_teams.to_csv('//xxx/My Drive/OAMK/Thesis/code/skaters/teams.csv')
#df_teams.to_json('//xxx/My Drive/OAMK/Thesis/code/skaters/teams.json')

jj22= (url_resp['teams'])
print (jj22)
df22=pd.DataFrame(jj22)
print (df22)
df22.to_csv('//xxx/My Drive/OAMK/Thesis/code/skaters/teams.csv', index=False)
df22.to_json('//xxx/My Drive/OAMK/Thesis/code/skaters/teams.json', orient='split', index=False,
force_ascii=False)

extracted_teamids= df22['id'].to_list()
print (extracted_teamids)

#this block will extract all players within teamid (Kalpa) for tournament id 31 =2021-2022 runkosarja

from unittest import registerResult
from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import requests
import matplotlib as pyplot

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()

headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxx"

teamids =['859884935']
results = []

for teamid in teamids:
    results = []
    try:
        res =
requests.get('https://api.wisehockey.com/v4.1/tournaments/31/teams/{}/players'.format(teamid),
headers=headers).json()
        print(res)

        filename=teamid

        df=pd.DataFrame(res)

    except:
        print ('Exception with playerid '+teamid)


#this block will extract all players with all teams for tournament id 31 =2021-2022 runkosarja
```

```
from base64 import encode
from distutils.log import info
from unittest import registerResult
from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import requests
import matplotlib as pyplot

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()

headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxxx"

teamids = extracted_teamids
#teamids=['495643563','859884935','55786244']
#results = []
df=pd.DataFrame([])
df_players=pd.DataFrame([])


for teamid in teamids:
   try:
      res =
requests.get('https://api.wisehockey.com/v4.1/tournaments/31/teams/{}/players'.format(teamid),
headers=headers).json()
      #print(res)
      #filename=teamid
      #new_data=new_data.append(res)
      print (teamid)
      jj22= (res['players'])
      df=pd.json_normalize(jj22)
      df_players=df_players.append(df, ignore_index=True)

   except:
      print ('Exception with teamid '+teamid)


print (df_players)


df_players.to_csv('//xxx/My Drive/OAMK/Thesis/code/skaters/allplayers.csv', index=False)
df_players.to_json('//xxx/My Drive/OAMK/Thesis/code/skaters/allplayers.json', orient= 'split' ,
force_ascii=False, index=False)


extracted_playerids= df_players['id'].to_list()
print (extracted_playerids)
len(extracted_playerids)
#this and next block will extract all players in all teams
from unittest import registerResult
from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import requests
```

```
import matplotlib as pyplot

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()
#parameters["leagueId"] = "55c8c0c9a7a84b636d8843e1"
#parameters["season"] = "2021"
headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxxx"

#playerids =['31046431','60591153', '27471447', '1234567']
playerids =extracted_playerids
results = []

df2=pd.DataFrame([])
for playerid in playerids:
   try:
      res =
requests.get('https://api.wisehockey.com/v4.1/tournaments/31/players/{}/skatersummary'.format(playerid
), headers=headers).json()
      #data=json.loads(res.text)
      y = {"playerid":playerid}
      res.update(y)
      filename=playerid
      print (filename)
      df=pd.json_normalize(res)
      df2=df2.append(df)

      with open('//xxx/My Drive/OAMK/Thesis/code/skaters/'+filename+'.json', 'w') as f:
         json.dump(res, f)


      #df['PlayerId']=playerid
      #print (df)

      #json_out=df.to_json()
      #df.to_json("//xxx/My Drive/OAMK/Thesis/code/skaters/"+filename+"_out.json", index=False,
force_ascii=False, orient="split")
      df2.to_csv("//xxx/My Drive/OAMK/Thesis/code/skaters/allplayerstats_out.csv", index=False)
   except:
      print ('Exception with playerid '+playerid)
#Get match ids
#https://api.wisehockey.com/v4.1/tournaments/31/matches?includeMatchesWithoutStatistics=false


from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import matplotlib

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()

headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxx"

df_matches=pd.DataFrame([])

try:
```

```
   res =
requests.get('https://api.wisehockey.com/v4.1/tournaments/31/matches?includeMatchesWithoutStatistics=
false', headers=headers).json()
   jj22= (res['matches'])
   df_matches=pd.json_normalize(jj22)
except:
   print ('Exception in getting matches.')

len(df_matches)

df_matches.to_csv('//xxx/My Drive/OAMK/Thesis/code/skaters/allmatches2021.csv', index=False)
df_matches.to_json('//xxx/My Drive/OAMK/Thesis/code/skaters/allmatches2021.json', orient= 'split' ,
force_ascii=False, index=False)
extracted_matchids= df_matches['id'].to_list()
#print (extracted_matchids)
len(extracted_matchids)

matchid_chunks = [extracted_matchids[x:x+50] for x in range(0, len(extracted_matchids), 50)]


#print (matchid_chunks[0])

print (len(matchid_chunks))

matchids=(matchid_chunks[0])

#print (matchids)
#get all games - playerstatistics
#https://api.wisehockey.com/v4.1/tournaments/31/matches/219244634/skaterstatistics

from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import matplotlib

headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()

headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxx"

df_skaterstatspergame=pd.DataFrame([])
df2_skaterstat = pd.DataFrame([])
#matchids =extracted_matchids
#matchids= ['448', '367', '445', '446', '449', '450', '447', '444', '442', '364', '440', '439', '443', '441', '437',
'438', '382', '48803', '315', '422', '420', '48802', '48800', '293', '324']
matchids=(matchid_chunks[0])
counter='-0'

for matchid in matchids:
   try:
      res =
requests.get('https://api.wisehockey.com/v4.1/tournaments/31/matches/{}/skaterstatistics'.format(matchid
), headers=headers).json()
      #data=json.loads(res.text)
      y = {"matchid":matchid}
      res.update(y)
      filename=matchid
```

```
        print (filename)
        df_skaterstatspergame=pd.json_normalize(res['skaterStatistics'])
        df_skaterstatspergame['matchid']=matchid
        df2_skaterstat=df2_skaterstat.append(df_skaterstatspergame)
        df2_skaterstat=df2_skaterstat.drop(columns=['periodStatistics', 'totalStatistics.speedZoneStatistics'])
        with open('//xxx/My Drive/OAMK/Thesis/code/skaters/matchstats/'+filename+'.json', 'w') as f:
            json.dump(res, f)

    except:
        print ('Exception with matchid '+matchid)


df2_skaterstat.to_csv("//xxx/My
Drive/OAMK/Thesis/code/skaters/matchstats/allmatch2021"+counter+"_out.csv", index=False)


import pandas as pd
import glob
import os

# setting the path for joining multiple files
files = os.path.join("//xxx/My Drive/OAMK/Thesis/code/skaters/matchstats/", "allmatch2021*.csv")

# list of merged files returned
files = glob.glob(files)

print("Resultant CSV after joining all CSV files at a particular location...");

# joining files with concat and read_csv
df_concat = pd.concat(map(pd.read_csv, files), ignore_index=True)
print(df_concat)


df_concat.to_csv("//xxx/My
Drive/OAMK/Thesis/code/skaters/matchstats/concat/allmatch2021_concat_out.csv", index=False)
#get all games - shotstatistics - including xG per shot!!!
#block 1
#https://api.wisehockey.com/v4.1/tournaments/31/matches/444/goals?normalize=true

from matplotlib.font_manager import json_dump
import requests
import json
import pandas as pd
from requests.structures import CaseInsensitiveDict
import matplotlib


#block 2
headers = CaseInsensitiveDict()
parameters = CaseInsensitiveDict()

headers["Accept"] = "application/json"
headers["Wisehockey-Api-Key"]= "xxx"

df_shotlstats=pd.DataFrame([])
df2_shotlstats = pd.DataFrame([])
#matchids =extracted_matchids
#matchids= ['448', '367', '445', '446', '449', '450', '447', '444', '442', '364', '440', '439', '443', '441', '437',
'438', '382', '48803', '315', '422', '420', '48802', '48800', '293', '324']
#matchids= ['448', '367']
```

```
matchids=(matchid_chunks[0])
counter='-0'
#block 3
for matchid in matchids:
   try:
      res =
requests.get('https://api.wisehockey.com/v4.1/tournaments/31/matches/{}/shots?normalize=true'.format(
matchid), headers=headers).json()
      #data=json.loads(res.text)
      y = {"matchid":matchid}
      res.update(y)
      filename=matchid
      print (filename)
      df_shotlstats=pd.json_normalize(res['shots'])
      df_shotlstats['matchid']=matchid
      df2_shotlstats=df2_shotlstats.append(df_shotlstats)
      df2_shotlstats=df2_shotlstats.drop(columns=['startXPosition',
'startYPosition','secondsFromPeriodStart','speed','blocker','saver','screeningPlayers',
'shotAreaId','fromLateralPass', 'royalRoadCrossed', 'shooterSpeed','shotDirection.horizontal',
'shotDirection.vertical', 'shotDirection.goalHorizontalLimits', 'shotDirection.goalVerticalLimits',
'shotDirection'])
      with open('//xxx/My Drive/OAMK/Thesis/code/skaters/matchstats/shotstats/'+filename+'shots.json',
'w') as f:
         json.dump(res, f)

   except:
      print ('Exception with matchid '+matchid)

df2_shotlstats.to_csv("//xxx/My
Drive/OAMK/Thesis/code/skaters/matchstats/shotstats/allmatch2021"+counter+"shots_out.csv",
index=False)

import pandas as pd
import glob
import os

# setting the path for joining multiple files
files = os.path.join("//xxx/My Drive/OAMK/Thesis/code/skaters/matchstats/shotstats/",
"allmatch2021*.csv")

# list of merged files returned
files = glob.glob(files)

print("Resultant CSV after joining all CSV files at a particular location...");

# joining files with concat and read_csv
df_concat = pd.concat(map(pd.read_csv, files), ignore_index=True)
print(df_concat)


df_concat.to_csv("//xxx/My
Drive/OAMK/Thesis/code/skaters/matchstats/shotstats/concat/allmatch2021shots_concat_out.csv",
index=False)
```