Bachelor's thesis

Business Information Technology

2024

Stephan Gu

Predicting Real Estate Value with Data Analysis

TURKU AMK
TURKU UNIVERSITY OF
APPLIED SCIENCES

Bachelor’s Thesis | Abstract

Turku University of Applied Sciences

Business Information Technology

2024 | 40 pages

Stephan Gu

Predicting Real Estate Value with Data Analysis

This thesis explores the dynamics affecting Finland’s real estate market, focusing on factors that influence property values. By using data from KVKL’s Hintaseurantapalvelu (Price Tracking Service) the study identifies key determinants such as location, accessibility, housing type, age, condition, energy efficiency, and building materials. Urban properties are generally priced higher if they are located near public transport because they are easy access points to the shopping facilities and leisure activities.

Challenges in data cleaning, including mismatched postal codes, inconsistent attributes, and language barriers in preprocessing tools, posed difficulties in the analysis. Different techniques such as fuzzywuzzy matching text data were applied; yet the poor quality of available information limited the improvement. Hedonic regression models were considered for multi-attribute property value estimation; however, the local calibration processes led to a high risk of overfitting as was the case with such models.

The results indicate that the properties located in the city, namely in the public transit access area, had a higher price. Moreover, long-term predictions for pricing within 12 months showed higher effectiveness than a short time, which was the case of the data. This thesis explores improving data quality for real estate analysis and sales predictions in the Finnish market through enhanced geocoding and standardized naming.

Keywords:

Artificial intelligence, Real estate, Data processing, Data science

Opinnäytetyö AMK | Tiivistelmä

Turun ammattikorkeakoulu

Tietojenkäsittely

2024 | 40 sivua

Stephan Gu

Kiinteistöarvojen ennustaminen Data-analyysin avulla

Tässä opinnäytetyössä tutkittiin Suomen kiinteistönmarkkinoiden dynamiikkaa keskittyen tekijöihin, jotka vaikuttavat kiinteistöhintoihin, käyttäen KVKL:n Hintaseurantapalvelun dataa. Opinnäytetyössä tunnistettiin keskeiset tekijät, kuten sijainnin, saavutettavuuden, asuntotyypin, iän, kunnon, energiatehokkuuden ja rakennusmateriaalit. Kaupunkikiinteistöt ja erityisesti ne, jotka sijaitsevat lähellä julkista liikennettä, ovat arvokkaampia, koska ne tarjoavat paremman pääsyn palveluihin, kuten ostoksiin ja viihteeseen. Energiatehokkaat kodit, joissa käytetään kestäviä materiaaleja, houkuttelevat myös korkeampia hintoja, mikä heijastaa kasvavaa ympäristötietoisuutta.

Datan puhdistamiseen liittyvät haasteet, kuten virheelliset postinumerot, epätarkat attribuutit ja kielimuuri analyysityökaluissa, vaikeuttivat tutkimusta. Erilaisia strategioita, kuten tekstidatan korjaustyökaluja (esim. fuzzywuzzy), tutkittiin, mutta datan laatuongelmat rajoittivat niiden hyötyjä. Hedoniset regressiomallit otettiin huomioon kiinteistöhintojen arvioimiseksi, mutta ylisovittamisen riski oli huolenaiheena.

Tulokset osoittivat, että kaupunkikiinteistöt, erityisesti julkisen liikenteen yhteyksien alueilla, ovat arvokkaita. 12 kuukauden ennusteet olivat tehokkaampia lyhyempiin jaksoihin verrattuna, koska ne ottavat huomioon pidemmän aikavälin ja enemmän arvoja. Tämä opinnäytetyö tutkii kiinteistöanalyysin ja myyntiennusteiden parantamista Suomen markkinoilla paremmalla geokoodauksella ja yhtenäisemmillä nimikkeillä.

Asiasanat:

Tekoäly, Kiinteistöt, Tietojenkäsittely, Datatiede

Contents

List of abbreviations (or) symbols **6**

Introduction **7**

1 Real estate dynamics **8**
1.1 Difference between location and accessibility 9
1.2 Housing type and size 10
1.3 Age and condition 11
1.4 Energy class and materials 12

2 Machine Learning: A Comprehensive Overview **14**
2.1 Basics and fundamentals 14
2.2 Consideration and Strategies 16
2.3 Future trends and innovative approaches 17

3 Data Processing **18**

4 Methodology **22**
4.1 Cleaning the data 22
4.2 Feature engineering 24
4.3 Grouping the data 30

5 Data Analysis **32**

6 Results **35**

Conclusion **37**

References **38**

Pictures

Picture 1. Code for counting unique values. 23
Picture 2. Example picture of Turku postal code areas (City of Turku, 2023). 27
Picture 3. Turku city part names and numbers (City of Turku, 2017). 28

Tables

Table 1. List of the columns within the data. 19
Table 2. Example of done window formation. 31

Figures

Figure 1. Correlation heatmap of the features. 20
Figure 2. Price inflation over the years. 33
Figure 3. Price predictions accuracy comparison with windowed formation. 36

List of abbreviations (or) symbols

KVKL	Kiinteistönvälitysalan Keskusliitto (Federation of Real Estate Agency)
ML	Machine Learning
AI	Artificial Intelligence
KPI	Key Performance Indicators
GIS	Geographic Information Systems

Introduction

In the recent past, the real estate business has shifted away from being a range in which the average person purchased a house to a broad-spectrum field with corporate, investment, and valuation structures. With the development of such industries, there is a desire to create predictive systems that can assist in making the decision much easier. In the latter models, AI and ML are at the center, transforming the approach of foreseeing values in real estate. Employment of AI in evaluation of real estate properties enhances forecasting of trends in the changes in price by taking into consideration the many aspects influencing the real estate economy, all of which compete to occupy the fixed time.

Machine learning (ML) is a key aspect of Artificial Intelligence(AI), enabling computers to process large amounts of data and identify trends invisible to the human eye. In the real estate market, price fluctuations are influenced by various factors, including location, building type, age, and condition of the property. In Finland, sustainability and energy efficiency further shape property values. This thesis explores these dynamics using historical data from KVKL's Hintaseurantapalvelu (Price tracking service) to improve property valuation. By leveraging AI to process and predict real estate trends, the goal is to enhance data-driven insights for investors and realign investment strategies.

This thesis is structured into six chapters. Chapters 1 and 2 provide a literature review, with Chapter 1 focusing on real estate perspectives and Chapter 2 exploring machine learning and AI applications. Chapters 3 through 6 detail the applied aspects of the study. Chapter 3 introduces the data and outlines the initial handling process. Chapter 4 covers data cleaning and preparation for analysis. Chapter 5 focuses on analyzing the processed data, and Chapter 6 presents the results along with considerations for future improvements.

1 Real estate dynamics

Historically, real estate was primarily associated with individual homeowners but has developed into a field that includes corporate endeavours and financial opportunities. Nowadays the impact has changed market dynamics and affected recent growth in property values. However, rising real estate costs is not a uniquely influenced by just the market trends but it displays the Finland's community's values. This shift in values is evident as more individuals transition from a preference for rural living to a desire for urban amenities. The real estate reflects larger socioeconomic variables, such as the health of a nation's economy.

Estimating real estate values for personal purchases is a significant commitment, influenced by numerous factors that impact the price. Therefore, cautious thoughts become essential when making such deals. The significance of financial aid for acquiring property are necessary. Careful real estate value prediction is necessary to guarantee alignment with getting loans without affecting economic boom.

Certain dynamics play a crucial role in building the real estate market and influencing commercial strategies adopted by estate agents. Understanding certain dynamics is critical for reading market trends and efficiently managing properties (Eldred, 2012).

Diving deeper into real estate values requires consideration of several factors, including the property's appearance, location, and sentimental value. It is also important to note that culture influences a lot of market trends and popularity in the country. On this topic, culture does not imply the historical or architectural importance of the buildings or the cultural backgrounds, but rather the preferences of the inhabitants or potential purchasers. These aspects may differ as the cultural importance of the buildings may not always align with the inhabitants' cultural preferences.

1.1 Difference between location and accessibility

Real estate valuation heavily depends on commercial trends and cycles. Peak example of this can be seen when comparing properties situated downtown against those located in rural areas. This raises the question of how to differentiate between location and accessibility. GIS play a crucial role in addressing this distinction by providing spatial analysis tools to assess proximity to amenities, infrastructure, and other location-based factors. Regardless of the similarities, real estate values question our understanding between these concepts and highlight the need for advanced tools like GIS to refine such analyses.

The city center's attraction does not only depend on its geographical coordinates, but the sheer number of amenities it offers as accessibility (Tiesdell & Adams, 2011). These areas' accessibility transforms mere proximity with amenities to cover the conveniences being shopping malls, vibrant nightlife with multiple dining options and pubs etc. As a result, the value of city centre homes reflects their unrivalled accessibility rather than their geographic centrality. (Calthorpe, 1993).

Public transit considerably improves accessibility, which affects property values. Mobility improvements, particularly those that enable smooth movement between locations, are inextricably tied to real estate asset worth. (Maselli et al., 2022). Advanced transportation infrastructures, particularly those with dedicated lanes such as rapid bus transit, have a positive impact on property prices. Real estate values in regions served by public transit increase significantly, especially in heavily populated urban areas (Primior, 2023).

The integration of public transportation ensures good access to public roadways or internal roads that connect to them, having efficient road infrastructure is mandatory for a working public transportation network. This convenience is especially useful for people commuting between cities and more remote locales. As transportation networks expand around certain real estate areas, different zones start to form. These zoning decisions have a considerable impact on the

future use and development of the area for example, densely populated areas may present difficulties in developing huge shopping malls due to space limits (Wolny-Kucińska, 2016; Lisowska & Grochowski, 2018).

The location value of a property is often inextricably linked with an emotional value when considering the sentiment and desire for that location (Eldred, 2012). A property by the sea is more desirable compared to a property in the middle of nowhere in some remote forest; this shows how the emotional elements indeed come into play regarding the valuation of property. However, one needs to differentiate this emotional value from the practical benefits associated with accessibility, which, even though they are related, constitute two separate entities.

The value of location can be sought to divide into two parts: emotional appeal and practical accessibility. For example, a site facing the water often has more value compared to one in the forest because of the emotional resonance it comes with. In Finland, many people prefer living in rural areas or forests; therefore, many lakes and waterbeds signify the site preference in the area. This is further supported by the wide ownership of cottages in the most remote regions.

1.2 Housing type and size

To understand real estate values thoroughly, diverse factors must be considered, starting with appearance, location, and cultural significance. In this context, cultural references pertain to the backgrounds of inhabitants rather than to architectural history. Property size and type are crucial for investors and occupants' appeal. Cultural norms strongly influence real estate values, shaping preferences for specific living arrangements.

For this instance, communal living arrangements are more popular in countries like Italy and Latin America where families live close together. In comparison, Finland is culturally more independent, with scattered families being the norm.

Furthermore, age distribution can affect a lot of preferences where younger people, drawn to heavily crowded city centres while retirees enjoying the peace and quit in the countryside. Understanding cultural and demographic factors is critical for accurately appraising real estate values and market trends (Levy, 2004).

Efficient space utilization has more importance in real estate, pointing out the functionality over sheer size. Floor plans serve invaluable information when dealing with space allocation within property. While optimal space utilization is ideal, it remains a challenge, especially in older buildings pictured by outdated layouts. Space utilization is a collection of history and trends where the buildings' initial age can have crucial role (Glascock, 2014).

Certain trends can be seen affecting the housing layouts in Finland's older houses where the bathroom size was vastly bigger compared to modern miniature sizes. While older buildings have large bathrooms reflecting previous emphasis on hygiene, modern design trends wanting more practical layouts. However, this trend has grown outlandish where these designs has grown to be for profit purposes compared to actual comfortability.

1.3 Age and condition

Furthermore, the monetary value of modern housing does not necessarily outweigh that of an older house. On the contrary, the condition of the property and the extent of renovations are more instrumental in the fiscal value. It is therefore common to renovate houses, whether for personal use or to improve the return on investments.

The age of a property appears logical, however, but its effective age is beyond mere chronological time. Effective age involves the structural condition, functionality, and expected lifespan, making it a complex evaluation (Bunyan Ünel et al., 2017). In this case, the importance of maintenance practices lies in

the fact that they affect the wear and tear of a property more than its chronological age.

Renovating older homes for personal use or investment is prevalent with a focus of providing increased comfort and functionality or making maximum possible profit. The complexities of effective age are based on the understanding that several years is not enough information to make a realistic assessment.

1.4 Energy class and materials

The issue of energy class and materials is also important in shaping real estate values. With a growing focus on sustainability and the need to conserve energy, renewable materials are becoming more important to consumers than ever before. Just as renewable resources reflect a more comprehensive, more contemporary social trend toward environmental consciousness, the housing sector is a critical area of public concern and development. Homeowners are increasingly trying to increase energy efficiency, reduce monthly payments rather than overall mortgage amounts, and raise self-sufficiency to try to head off future years of escalating fuel expenses (De Paola et al., 2022).

The interaction between energy class and the provision of heating forms another complex component to be considered. It is critical to note that improving insulation brings energy savings even without advanced heating systems. For example, in Finland, the energy class certificate is mandatory when a property is sold, although this requirement may be lifted in the case of older or historically important buildings. The certification is valid for ten years and includes a range from A to G by issue year.

Therefore, materials play a significant role not only in terms of energy but also in their implication of the emotive value within buildings. The materials that possess high thermal capacity such as rocks, metals and minerals might yield extra heating spending and, in parallel, reduce the level of energy efficiency.

Moreover, the selection of materials is based on people's taste and experience – while some people prefer the warming experience provided by wood, others prefer the cold and timeless feel of a stone and brick. Due to growing housing demand, the prefabricated construction is widespread; it is the expeditious process of the construction accompanied by the factor of the purity. Nevertheless, the outcome exhibits the same parameters as the traditional buildings hence the dichotomy of the material point (De Paola et al., 2022).

2 Machine Learning: A Comprehensive Overview

This research aims to utilize various ML models to analyse real estate data, perform training and predict the future values of plots and estates based on data from the past. During this research, it is required to understand and forego different methodologies of data processing and model creation. Especially an excessive amount of time goes for data preprocessing, as the datasets are bound to have many complexities and imperfections. These errors mostly caused by people and low-quality data, which must be organized and formatted in accordance with ML algorithms.

The construction of ML models is a concept that predominantly depends on the pre-determined libraries and technologies because they shape the result taking into consideration the existing data (Choy & Ho, 2023). Due to the lack of basic knowledge, one might not be able to choose the right type of libraries and technologies. Consequently, understanding the core principles is critical in ensuring proper decision-making and effective ML performance. (Géron, 2019)

If there are excessive empty columns, the column can be considered irrelevant for analysis or further filled with the use of the dummy data technique. Dummy data is often binary and consists of values of one and zero, which can be understood as true or false, respectively, to replace categorical variables. This action supports the original structure of the primary dataset, allowing for better model training and statistics. (Park & Bae, 2020)

2.1 Basics and fundamentals

Firstly, starting with the fundamentals in comprehending the structure and operation of ML models is complemented by their respective designs and systems. Understanding the basic differences between supervised, unsupervised, and semi-supervised learning concerning deep learning versus

linear models' capabilities. Deep learning has shown impressive results across different fields, notably not limited to, but including image identification tasks. (Shi et al., 2023)

When building models, supervised learning refers to model construction made with human guidance and unsupervised learning, allowing fully independent ML. Supervised learning consists of creating linear regression models based on identified data pairs. The trained algorithm memorizes how to relate the input characteristics to the output target. Unsupervised learning is based on unidentifiable data generating patterns through methods other than human assistance. (Vora, 2023)

Among the two most typical artifacts in real estate data analysis, a linear regression model and neural networks can be singled out. The strong point of both approaches is that they treated the underlying complexities of real estate valuation very differently. The first uses a very simple, interpretable construction, while the second has complex architecture structures that enable extracting complex patterns of data. In general, both are supposed to bear responsibility for data extraction in the real estate sphere.

One of the primary challenges related to the development of models is overfitting. Overfitting curriculum is a phenomenon where the model becomes too reliant on the training data and, as a result, fails to predict the testing data effectively or at all. Therefore, in general, underfitting is more acceptable for the model than the overfitting is. While the predictions made based on underfitting are simplistic and are likely to miss many possible nuances in the data, it is still more likely to generalize and not make the mistakes associated with the overfitting. (Ren et al., 2020)

2.2 Consideration and Strategies

Model comparisons in this thesis are limited to linear models as opposed to various deep learning disciplines. Linear models are straightforward, making them appropriate only for superficial regression applications. As such, their application is limited because the assumption that exists in complex scenarios where the linearity of the input features and target variables is not possible to achieve allows only low complexity. (Lee, 2019)

Hedonic regression is a type of regression explicitly tailored to obtain a model that understands the multiple attributes of a particular property. This type of regression can produce the best results possible only when accurate data are available (Taltavull de La Paz, 2021). Valuation data do pass this aspect off, which presents a disadvantage when interpreting the synergies of attributes, especially where multiple factors exist.

Deep learning has been considered the future with a great potential due to its ability to learn complex hierarchical features automatically from raw data, which eliminates manual feature engineering. It is especially good at modeling intricate non-linear dependencies between inputs and targets. However, while this power can be an advantage, the high complexity and the amount of data that needs to be trained are challenging for such systems. (Campesato, 2020)

Data quality is paramount in making model development most effective, with accuracy, completeness, consistency, and relevance being key. Measuring KPIs such as precision and recall becomes essential for assessing model effectiveness. Data with accurate measurements, no missing values, consistent, and relevant information are critical elements of quality data. Paying attention to these factors further guarantees trustworthiness and efficiency in ML models.

2.3 Future trends and innovative approaches

While ML can take parameters from already trained models for setting parameters for future events like pandemics, the very unpredictability of each event means an accurate forecast regarding its effects on the real estate market may not be easy to have. Take, for instance, the COVID-19 pandemic. While it really affected property prices, other pandemics in the future may not have this kind of outcome since there could be different economic and governmental responses.

Real estate is among the most volatile markets, as many factors come into play, such as the rates of interest, the level of employment, and the population. The interdependence of these factors makes accurate real-time predictions difficult to achieve. AutoML does promise some things in the sense that models are self-chosen and tuned, but when diverse data sources appear, the combination and processing remain a big challenge.

Different machines act variably on different tasks: some machines are doing very well on temporal trends, while others work well with spatial or nonlinear data. No model can capture the complexity that comes with the forecasting of real estate. As a matter of fact, the future of AI in the domain is integrated models, whereby several model strengths have enabled them to capture the far-reaching adaptive system. This probably would further have an impact on making such predictions more realistic, more robust, and closer to the dynamic nature of real estate markets.

3 Data Processing

The dataset was obtained from the KVKL's website hintaseurantapalvelu.fi, containing real estate records across the entire Finland. The dataset is originally in Finnish, and the dataset comprises approximately 1.7 million instances dating from 1999 to the present day. However, this thesis focuses on analysing the period from 1999 to 2023.

The data is mainly designed for professionals working with real estate and provided for research purposes, it can't be used for marketing purposes. The research area is the Turku city region alone with deliberate exclusion of towns within its vicinity, such as Raisio, Kaarina, and Lieto. Inside Turku city region alone there was around 80 000 instances. However, the findings from this localized analysis can later be applied to the entire country.

KVKL's website offers an inclusive search function for real estate purchases, allowing users to filter results based on various criteria. These would also include but are not limited to age of building, housing type, no. of rooms, area in square meters, year of construction, pricing in dollars, plot area in square meters, postal code, city, and address. All the attributes are easily accessible and supplemented while exporting the data from the website. Further, the website offers an easy export option to PDF, Excel, or for direct printing.

Table 1. List of the columns within the data.

```
RangeIndex: 79681 entries, 0 to 79680
Data columns (total 46 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Estate Type             30003 non-null  object
 1   Preperty identifier     1997 non-null   object
 2   Share numbers           0 non-null      float64
 3   Housing Type            79681 non-null  object
 4   Municipality            79681 non-null  object
 5   City part               75038 non-null  object
 6   Postal Code             79681 non-null  int64
 7   Address                 79681 non-null  object
 8   Living area(m^2)        76910 non-null  float64
 9   Built year              76458 non-null  float64
 10  Rooms                   76450 non-null  float64
 11  Room layout             79369 non-null  object
 12  Floor number            69679 non-null  float64
 13  Floors                  71352 non-null  float64
 14  Price                   79669 non-null  float64
 15  Debt portion            38789 non-null  float64
 16  Unlevered price         79669 non-null  float64
 17  Price/sqm               77393 non-null  float64
 18  New estate              79681 non-null  bool
 19  Condition               79681 non-null  object
 20  Date                    79681 non-null  datetime64[ns]
 21  Ownership               75510 non-null  object
 22  Estate Description      0 non-null      float64
 23  Plot sqm                57704 non-null  float64
 24  Start of sales          76920 non-null  datetime64[ns]
 25  Sale time               75999 non-null  float64
 26  Treatment fee           64725 non-null  float64
 27  Treatment/sqm           63985 non-null  float64
 28  Material                74117 non-null  object
 29  Beach                   77850 non-null  float64
 30  Elevator                78882 non-null  float64
 31  Rental                  79681 non-null  bool
 32  Sauna                   30436 non-null  float64
 33  Balcony                 29579 non-null  float64
 34  Built year Description  293 non-null    object
 35  Construction area       1516 non-null   float64
 36  Entire plot (other)     16423 non-null  float64
 37  Building Rights         0 non-null      float64
 38  Beach Desccription      114 non-null    object
 39  Heat source             14885 non-null  object
 40  Energy class            17127 non-null  object
 41  Redemption Price        17 non-null     float64
 42  Monthly rent            42 non-null     float64
 43  Redeemed share          6854 non-null   float64
 44  Made Renovations        5791 non-null   object
 45  Upcoming Renovations    5704 non-null   object
```

For MDA reasons the exact information from the dataset can't be shown but the first-face evaluation, the KVKL dataset looked neat and effortless to read through. Though quite a bit of work was put into the preparation of the data to ensure its quality and uniformity. This entailed correcting spelling or typographical mistakes and normalizing naming for proper analysis to occur.

Incomplete data entry and abbreviated variables presented challenges during analysis. To address these, missing context was supplemented using secondary materials such as additional documents or books for clarification. A correlation heatmap of features was also created to better understand relationships and improve the overall data analysis process.

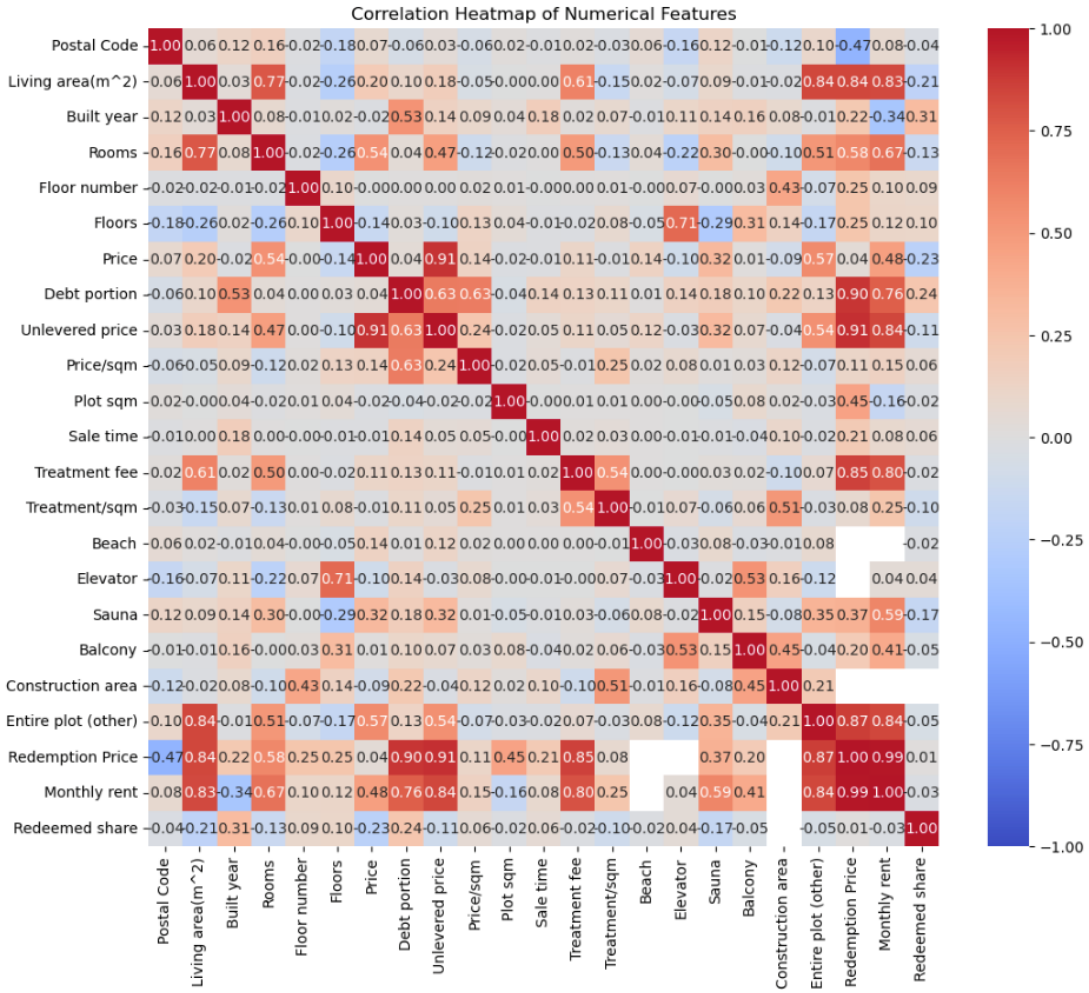


Figure 1. Correlation heatmap of the features.

Another serious problem was how to deal with the missing values from the set of content. In order to keep the data, set unharmed, some pragmatic approaches were used, such as getting rid of the corresponding columns having a lot of missing data or imputation of fake values where applicable. Also, such techniques as imputation, which included, for example, replacing missing values with mean or median values, were applied to make sure the dataset was strong enough for further analysis. Finally, these legwork steps on the data were vital in readying the data for proper ML forecasting and analysis, giving rise to trustworthy predictions and understanding.

4 Methodology

The process of building a ML model begins with excluding unused data and transforming the rest into a format that the model can work with, specifically into numerical values. Thus, ML algorithms work on numeric data; this transformation changes those non-numeric attributes to numbers to enable the model to understand them and connect them to the rest of the features. When the data has been rearranged, it should be visualized. This helps discover trends, outliers, and areas for improvement. Visualization not only helps in understanding the data better but also assists in feature selection, identifying which variables are more likely to contribute meaningfully to the model's performance.

In this project, the data was made up of 45 columns, which means data preparation was a huge, big task that required a lot of detail in accomplishment. Still relatively new to handling raw data and the complexities involved in preprocessing it, at times it felt like an uphill task. Preprocessing data, being the most time-consuming part of building any ML model, is important to ensure high performance and accuracy of the model. Clean, consistent, and relevant data form the backbone of any great model of ML.

4.1 Cleaning the data

Understanding the data structure is crucial, so the initial cleaning focuses on reviewing the columns. Initially, a superficial analysis was conducted in Excel to familiarize with the dataset. Nevertheless, the transition was made to Python and Pandas in Jupyter Notebook to achieve a more comprehensive analysis. The following code demonstrates how the uniqueness of a column was checked:

```
ownership_count = data['Ownership'].nunique()
print(f'Unique ownership values: {ownership_count}')
```

Picture 1. Code for counting unique values.

Counting unique values within the 'Ownership' column facilitated the understanding of useful patterns for the model. Further similar commands were iterated over other columns to check uniqueness and value distribution. Checking uniqueness also provided insight into the extent of preprocessing required for that column. In the 'Ownership' column, one incorrect unique value was discovered and eliminated, leaving just three unique values to transform into numerical form.

After some preprocessing and analysis on the data, the 'Condition' column was identified, categorized as strings. The condition rating varies in range starting from "tolerable" to "excellent," with an additional "new." Values had to be standardized to be analyzed numerically. The unknowns and blank entries have been scored as zero, and the actual conditions have been mapped onto a numeric scale ranging from 6 to 10. "New" is ranked highest in this scale and hence corresponds to the most significant appreciation in the value of the property. This solution is a points-based kind of evaluation; therefore, it works in such a way that the higher the score, the better the general condition.

Subsequent analysis focused on the 'Energy Class' column. Despite its relatively few instances, this column exhibited significant variability with approximately 151 unique values. A deeper look proved that it was badly formatted for the most part, with data whose entries were representative of the same information but differently presented. Common issues included extra spaces before the entry of a character or number, inflating the uniqueness. This process involved unifying similar values and correcting formatting issues, thereby improving the overall consistency and reliability of the dataset for subsequent analysis.

4.2 Feature engineering

Energy class defines the energy efficiency of the buildings. The energy class column in the raw data ranges between G-2007 and A-2018, but in general, all the data in this column was messy with different types of unknown markings, a few contain only a year whereas others contain only the class designation. Some entries only included the year, while others only contained the class designation. To streamline the analysis, entries that provided only the year were removed, retaining only those with energy class designations. Cleaning of this column followed the devising of a point system to quantify the energy efficiency by giving energy classes points depending on how efficient they were. The higher the class the higher points ranging classes from A to G; the best class for efficiency was A and the worst was G. Additionally, the year of the energy class was incorporated as a multiplier to adjust the points according to the recency of the classification. For instances where only the energy class was available without a corresponding year, points were scaled based on the oldest year in the dataset, which was 2007. Consequently, points were assigned as follows: unknown or empty entries received 0 points, while classes G to E typically indicated lower energy efficiency, often associated with older buildings. In this column class D means slightly below average, while class C represented the average energy efficiency for households in Finland. Class B indicating good energy efficiency, and class A signified high efficiency, suggesting that a building could potentially be self-reliant in energy production. Ultimately, the points assigned to energy classes ranged from 0 to 100.

The column ’housing type’, which previously included 17 distinct strings representing other types of properties, was changed into a new one, where each housing type was given a unique integer ID through mapping procedures. As a result, the original string-based entries were replaced with a simpler numerical format, which made it easier for the model to deal with and understand, thus providing a suitable and efficient solution. The logical groups were formed from the mappings, where the distinct ID values were given for the apartment complexes, business types, independent houses, and other

structures like garages and parking spaces.
So the process of ID assignment made it possible to retain the categorical features, which enabled the pure point-based representation not to be applied; moreover, such a method ensured the uniqueness of the classification of each type of house on the input to the model. On top of that, this ID-based method was able to keep data readable and intact in a way that the model could distinguish property types based on various data types correctly.
The other categorical columns were subjected to further transformations:

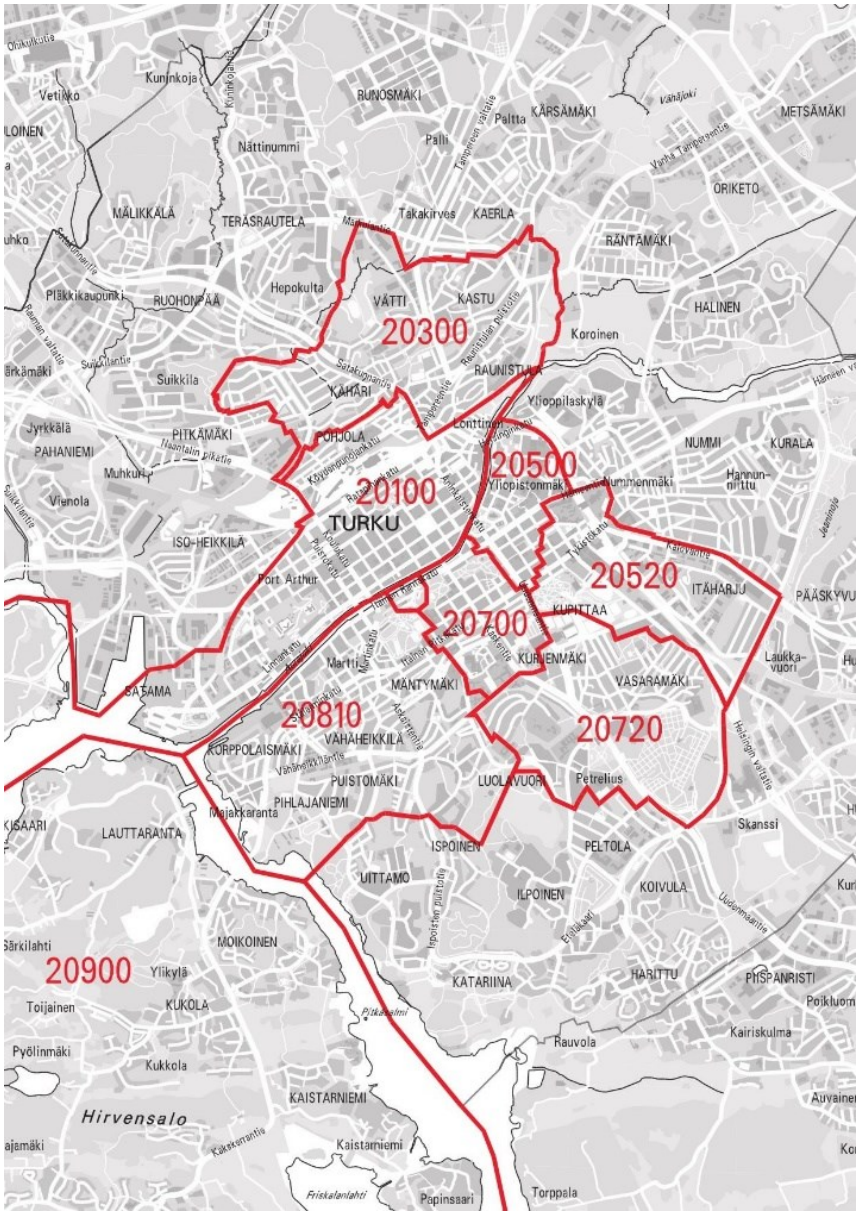
- **New Estate:** In which new estates received the binary “1” while others were marked “0,” the construction status of properties was made the primary predictive factor through this format that was easy for the model to work with.
- **Rental Status:** The "Rental" column was also converted to binary, where "0" was assigned for non-rented properties, while "1" was assigned for the rented properties. This way, the model could use the rental status as a feature while the data format remained unchanged.
- **Time Period:** The conversion of the "Date" column to a monthly period format, which indicates the year and month, enabled the model to identify the temporal patterns without any added layer of complexity which would be inferred from the daily data.

All of these transformations made it possible for all categorical variables to be represented by the corresponding numeric values, so the model could assess all property types and characteristics and, at the same time, avoid extra complexity.

The quality and the quantity of data are two main components when it comes to ML models and significantly affect the results. While using the dataset from KVKL for modeling purposes, several issues related to data collection were encountered. Predictions of real estate especially rely on the location and its accessibility, making factors such as city, postal code, and street address of utmost importance.

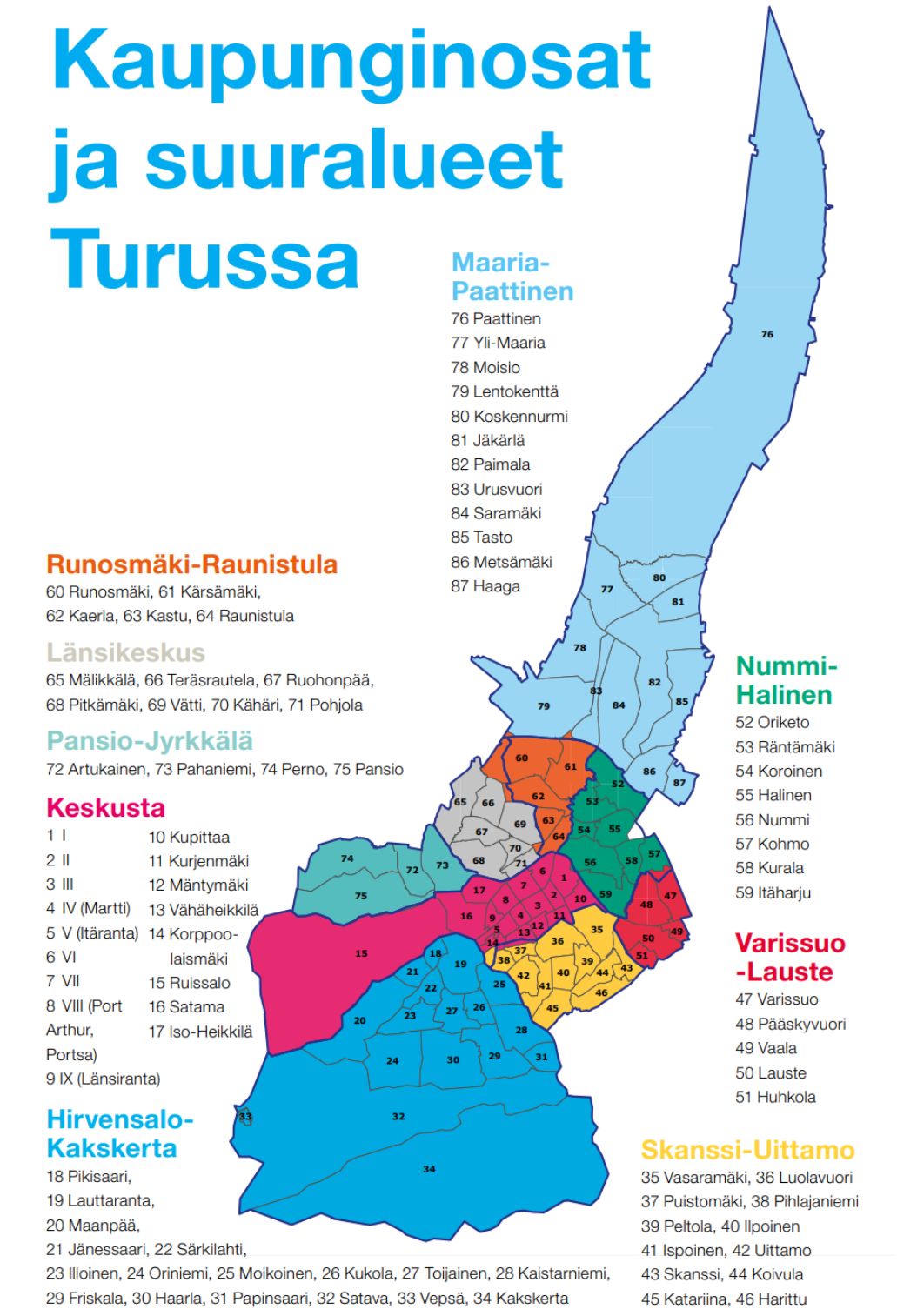
One cause of the issue mentioned above is attributed to postal codes, which provide only a broad sense of direction and geographical areas and can also be quite varying. For instance, some areas with the same postal zone cover different regions, while others with the same postal zone can encompass a single large area. The postal code 20100 for example contains the Turku inner city and Ruissalo, while the postal code 20900 covers a greater portion of Hirvensalo. This aspect is troubling because Hirvensalo has many neighbourhoods, and the name is just a description of the whole island. Such differences in the definition of the postal codes can yield a skew outcome in the analysis of the real estate data. Therefore, there is a need for improvement on the geographic categorization in the following investigations on the analysis.

Considering this, the dataset in question requires further refinement. Approaches may involve additional geographic attributes or the employment of better locational attributes to increase the prediction efficacy. In general, limiting the data quality will remain the major challenge to the creation of the ML models that will be used to make predictions in real estate.



Picture 2. Example picture of Turku postal code areas (City of Turku, 2023).

It is positive, that the postal codes were gathered properly, while the principal problem came with 'city part' data and the column 'street address'. It is cumbersome to collect street addresses of sold houses in the Turku area because there are so many streets that can utilize this data properly. The municipality clearly divided the city parts of Turku, and its map points out which zone a certain area belongs to. However, the 'city part' field is filled in by hand by people, and several problems take place, starting from misspellings up to the usage of unofficial nicknames instead of the formal names of the city parts.



Picture 3. Turku city part names and numbers (City of Turku, 2017).

This column has been disorganized and thus rendered useless, which is unfortunate as even such a column would have had some usefulness because it is not easy to determine real estate values using only postal codes. For example, the city centre and Ruissalo has the same postal code but, such areas are very different, especially if we consider the accessibility. The girth was about 1500 elements which were hard to clean. In most cases, this was accomplished using non-sophisticated and primarily manual techniques. To address this, data on street addresses was utilized, as it allows for the identification of the corresponding city part for each address. With this aid, a compilation of every street address available in Turku was discovered. However, additional problems arose as some of the lengthy streets covered more than one city part. Furthermore, the column of street addresses became untrustworthy given that the entries were few and scattered. Finally, both address and city part columns were omitted, and reliance was placed solely on postal codes for location information.

The establishment of some principles of data entry could be one solution to all the data management problems encountered by KVKL. Rather than stressing the need to fill in several specific details of every real estate entry, it would be better to give a selection of alternatives from which users select the appropriate real estate column, such as city part or postal code (Jha et al., 2020a). Ideally, the system could have contained such automated systems that would fill these information fields, by limiting the possibility of human error to the minimum. When the street address is given, the system can fill area and city postal codes automatically.

Another such case has been encountered when dealing with the ‘building materials’ and ‘warmth source’ columns—the data was too contaminated for further analysis. Due to the absence of standards, these attributes were rendered ineffective. To enhance the organization of the 'sale date' data, the column title was changed to 'time' and scaled to year and month instead of specific dates. By doing so, the data was subjected to time segments, which could later be joined with postal codes to establish better-segmented data for machine-up modeling. (Jha et al., 2020b)

4.3 Grouping the data

After completing the conversion of the viable columns into numbers, the rest of the available data that was absent was replenished with zeros, and then the data was normalized. Following the normalization of data, the next activity was to assimilate the data, constricting the high 80,000 instances to a smaller number. In particular, the focus was on how to manage and arrange the real estate purchase data over time and geolocation.

Upon this partitioning, we were left with around 6,800 instances, approximatively. After this, instances were organized into different Excel files, which corresponded to certain postal codes. The aim was to construct and fit linear models on these files that relate to each specific region. Starting at, for instance, postal code 20100, where the model starts building and learning about area 20100 and then proceeds to 20300 and all the other areas, learning the distinct features of each mail-coded area independently and comparatively.

Before exporting the data to Excel, the windowing method was applied to the price column. That method employed the time series data, which was given from the start to the end of each period, thus allowing future periods to be used in simulating price prediction. This truly assisted in making the data requisite for model training on predicting current prices of real estate.

Price	Price (w=3)	Price (w=6)	Price (w=9)	Price (w=12)
0.213994	0.223221788	0.216888915	0.223379672	0.211784023
0.213468	0.212897977	0.209345603	0.21546358	0.226139706
0.223222	0.216888915	0.223379672	0.211784023	0.213972461
0.212898	0.209345603	0.21546358	0.226139706	0.215108343
0.216889	0.223379672	0.211784023	0.213972461	0.217875686
0.209346	0.21546358	0.226139706	0.215108343	0.216970781
0.22338	0.211784023	0.213972461	0.217875686	0.215815894
0.215464	0.226139706	0.215108343	0.216970781	0.214696092
0.211784	0.213972461	0.217875686	0.215815894	0.217590619
0.22614	0.215108343	0.216970781	0.214696092	0.213077788
0.213972	0.217875686	0.215815894	0.217590619	0.219456659
0.215108	0.216970781	0.214696092	0.213077788	0.216523445
0.217876	0.215815894	0.217590619	0.219456659	0.211743091

Table 2. Example of done window formation.

5 Data Analysis

While reorganizing data, remapping of existing values was an essential step, making it easier to address inconsistency. This task could have turned into a project of its own. Very often, postal codes are mismatched with the addresses or the respective parts of the city sections. To illustrate, sometimes the postal code could not match with a single city part since it could stretch between two or even more different areas. The right postal codes were indicated for the areas, but those were not enough information regarding their location or access. (Alonso, 1993)

Matching addresses with the responding city part is achievable by utilizing publicly available data on Turku's neighbourhoods like Turku Street List and Google Maps. However, the process faces notable challenges, as some addresses can span across multiple city parts, creating overlapping and inconsistent information. Turku consists of 87 distinct city parts, each with dozens addresses, making manually remapping addresses impractical, when many entries are incorrectly categorized or missing altogether.

It would be beneficial to have a more advanced system for matching addresses with the corresponding postal codes and city zones, as this would significantly improve data quality. GIS tools, such as MapInfo Pro or ArcGIS, could be instrumental in this process by more accurately aligning geographic data with postal codes and city zones. However, Google Maps does not provide a comprehensive solution, as there are still issues such as imprecise, outdated, or missing addresses. More frequent updates to Google Maps, particularly for new or modified addresses, would enhance its reliability. Additionally, developing an automated classification system using GIS in combination with machine learning (ML) or fuzzy matching methods could improve both the speed and accuracy of real estate data processing. This approach would ultimately contribute to more precise and efficient real estate data dissemination, particularly in cities like Turku. (KTI Kiinteistötieto, 2022)

Another key factor to take into consideration is the period between the onset of sales and sales. The duration between the commencement of a sale and the sale itself is an indication of the attractiveness of the asset for sale. An asset that sells quickly is in high demand and therefore could have sold for a higher price. On the other hand, real estate is a significant purchase for most people, so longer selling periods might simply reflect the time needed to make such a large decision, rather than low demand. It is also possible to study the trends concerning when it is best to purchase or sell property by comparing the start of the selling dates. As shown in the figure below, inflation trends are evident in the changes in real estate prices over the years. This highlights the importance of considering inflation when analyzing property values, as it directly impacts nominal prices and purchasing power.

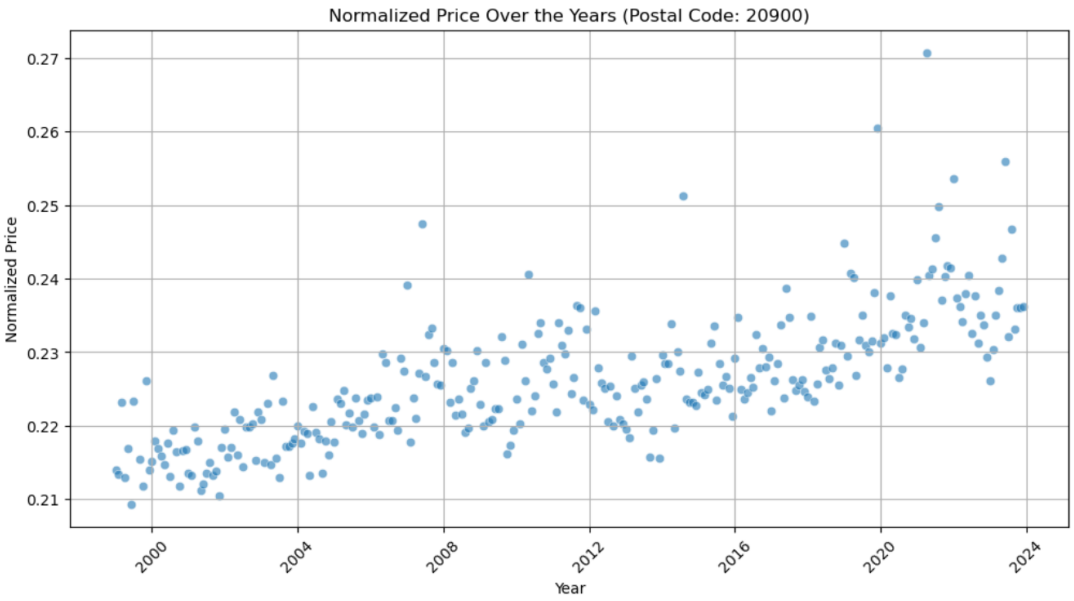


Figure 2. Price inflation over the years.

Incorporation of energy class and building materials into the analysis was intended. However, the attribute for material was highly problematic. Like most of the data, it was incredibly messy, and cleaning would prove tiresome. This attribute was one of many that were so hardly unified.

The optimal process of data preparation and model tuning would be to consider each relevant factor; however, this level of detail is hard to attain, especially when data gathering and labelling is not consistent. During this research, different libraries were investigated to aid in data cleaning and standardization. For instance, Fuzzy Wuzzy was considered for text data matching and cleaning in English; however, a significant portion of the data gathered for this study was in Finnish, which substantially limited the tool's application.

Additionally, problems were encountered with interpreting the date variable because, despite being in a numerical format, the model could not effectively manage this aspect. Time-series analysis could also be caried out using the date variability of customers on the real estate market, in other words, it would be possible to find out that it is best to buy or sell properties during certain months of the year. However, these time-dependent insights were hard to be captured given the difficulty of finding the proper ways to connect the date data with the other ones.

Hedonic regression models, as evidenced by the work of (Sopranzetti, 2015), would have been a useful tool in this case. Such models allow the estimation of property values based on several attributes, therefore it is possible to systematically include different aspects such as the location, type of construction, thermal performance, and so on in the evaluation process. The main emphasis of the data-cleansing effort was to achieve the highest prediction power of the model with yet the most generalizable data possible for new purposes. The risk of overfitting is always present, especially when the data pre-processing is too tailored to the characteristics of the Turku region. Techniques such as issuing IDs or creating a point system for some of the attributes were applied to introduce flexibility and enable model adjustment to other areas. A variety of methods exist, but to have a proper mix of accuracy and adaptability, this framework was implemented.

6 Results

The analysis of the data presents the findings which confirm that location is one of the most significant determinants of property values. Urban property prices are higher than rural property prices owing to better accessibility within cities. One of the major findings is that in the cities where there is an appealing public transport property prices are higher, as residents are able to reach different facilities with ease. Accessibility is more than just place as people are more inclined to places value where there are facilities or services such as shopping and entertainment.

The research also looks at the effects of structure and design on property prices. Attitudes towards shared living facilities in Finland are also not as strong as they are in their more family-oriented counterparts. This trend suggests that more importance will be placed on how space is organized since size has already been rendered obsolete.

The degree of age and level of unimproved properties and their respective worth in the market are quite interesting. Probably the most common price limiting factor is the existence of old constructions. However, the age timeliness of incorporation comes as secondary as the overall state of the facility is the most critical criterion. For instance, homes given an excellent status can command even higher monetary value in the open market, thus bolstering the case for constant upkeep or repair of such facilities. Buyers who hate the hassle of going through renovations will cherish all the possibilities that enhanced refurbishing brings since selling the house increases the personal expenses.

Sustainability and cultural preferences significantly influence real estate trends. Energy-efficient homes with sustainable materials attract higher prices, reflecting growing environmental awareness. In Finland, individualistic preferences for separate cabins contrast with Mediterranean housing trends, emphasizing the role of cultural factors in property valuation.

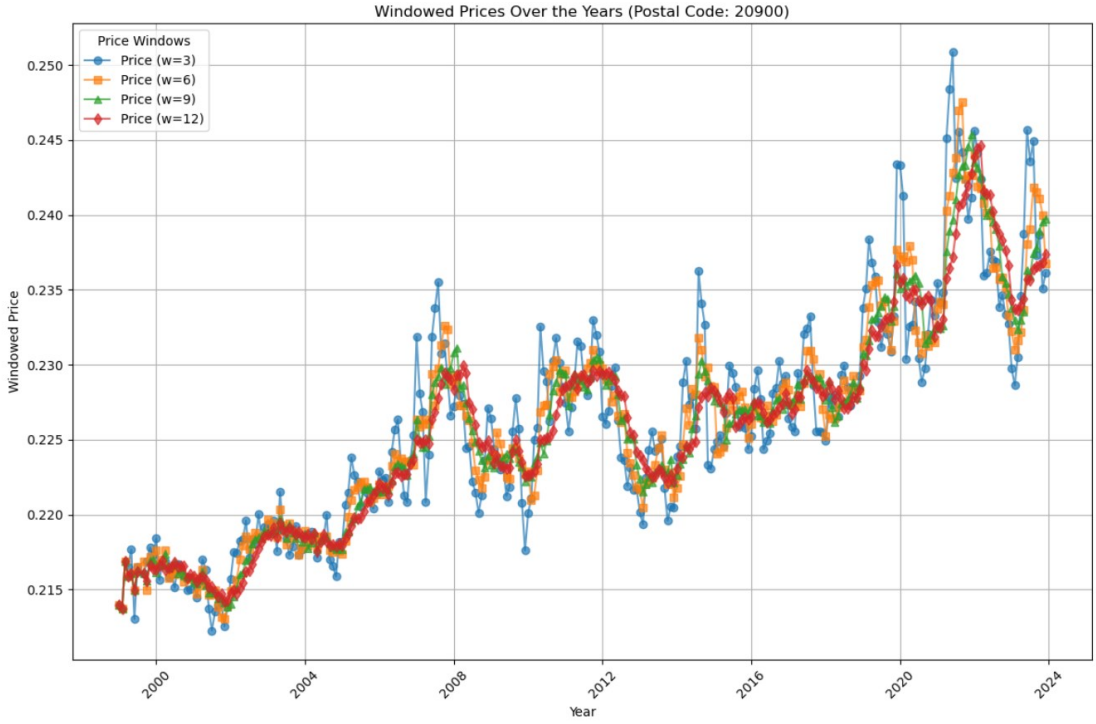


Figure 3. Price predictions accuracy comparison with windowed formation.

Results show that a 12-month window yields more accurate price predictions than a 3-month window, effectively capturing seasonal and market trends.

As such the conducted analysis provides some important aspects affecting the Finnish real estate market. In fact, the correlation between the location, its accessibility, the type of housing, its age and quality, its energy efficiency and the prevalent culture is multi-dimensional in nature in regard to property valuation. KPIs, such as average price per postal code or city part, further enhance the understanding of these dynamics, offering actionable insights. These also provide a guiding theory to the practitioners in the field because they articulate practical issues which decision makers can utilize in the real estate business.

Conclusion

This study explored Finland's real estate dynamics, focusing on a variety of factors significantly influencing market prices. Through a comprehensive analysis of data obtained from KVKL's Hintaseurantapalvelu service, many key trends and relationships was identified among location, accessibility, housing type and size, age and condition, as well as energy efficiency and materials etc. KPIs, such as price trends by postal code and city part, offered valuable metrics for assessing these factors. The findings support known facts about urban properties, especially those near public transportation command higher values. Moreover, as sustainability housing becomes more and more important, it has consequences for the prevailing market trends, indicating that energy-efficient houses will be increasingly appreciated by buyers in the coming years.

While some lessons have been learned, there is still plenty of room for improvement in the processes for collecting and analyzing data in the future research efforts. Moving away from the reliance on property sales in terms of collecting the data and adopting a much better comprehensive approach would be more beneficial in understanding the market. In this case, unified data acquisition techniques focused on geocoding such as addresses and postal codes will significantly enhance the dataset quality. Such an approach may also result in the decrease of the number of collected string values simplifying analysis and increasing its precision.

In addition, this study findings may be the basis for additional research on forecasting prices in the real estate markets more effectively than has yet been achieved. This will be possible by improving data collection techniques and expanding the variables employed in the study. All in all, this study enhances the knowledge on property pricing in Finland and stresses the need to always enhance the quality of data for future research.

References

Alonso, W., 1993. Location and Land Use: Toward a General Theory of Land Rent. Cambridge, MA: Harvard University Press.

Maselli, G., de Luca, S., & Nesticò, A., 2022. Infrastructure Accessibility Measures and Property Values. Journal of Infrastructure, 34(2), pp. 211-225.

Bunyan Ünel, F., Yalpir, Ş. & Gülnar, B., 2017. Preference Changes Depending on Age Groups of Criteria Affecting the Real Estate Value. *International Journal of Engineering and Geosciences*, 2(2), pp. 41-51.

Calthorpe, P., 1993. *Real Estate Development and Urban Form*. San Francisco, CA: Island Press.

Campesato, O., 2020. *Artificial Intelligence, Machine Learning, and Deep Learning*. Dulles, Virginia; Boston, Massachusetts; New Delhi: Mercury Learning and Information.

Choy, L.H.T. & Ho, W.K.O., 2023. The Use of Machine Learning in Real Estate Research. *Land*, 12(4), p. 740.

City of Turku, 2017. *Turku city part names and numbers*. Available at: https://www.turku.fi/sites/default/files/atoms/files//kaupunginosat_ja_suuralueet_turussa.pdf

City of Turku, 2023. Turku city cleaning plans. Available at: https://www.puhdistussuunnitelmat.fi/turku/kartta

De Paola, P., Tajani, F. & Locurcio, M., 2022. *Sustainable Real Estate: Management, Assessment and Innovations*. Basel: MDPI - Multidisciplinary Digital Publishing Institute.

Eldred, G.W., 2012. *Investing in Real Estate*. 7th ed. Hoboken, N.J.: John Wiley & Sons, Inc.

Géron, A., 2019. *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow*. O'Reilly Media.

Glascock, J.L., 2014. *Real Estate*. Bradford, England: Emerald.

KVKL, 2023. Hintaseurantapalvelu. Available at: https://www.hintaseurantapalvelu.fi

Jha, S. B., Babiceanu, R. F., Pandey, V. & Jha, R. K., 2020a. Housing Market Prediction Problem using Different Machine Learning Algorithms: A Case Study. Procedia Computer Science.

Jha, S. B., Pandey, V., Jha, R. K., & Babiceanu, R. F., 2020b. Machine Learning Approaches to Real Estate Market Prediction Problem: A Case Study.

KTI Kiinteistötieto, 2022. The Finnish Property Market 2022.

Lee, W., 2019. *Python Machine Learning*. 1st ed. Indianapolis, Indiana: Wiley.

Levy, D.S., 2004. *Behavioural Real Estate*. Bradford, England: Emerald Group Pub.

Lisowska, M. & Grochowski, P., 2018. Accessibility of Real Estate by Transportation as a Determinant of the Development of Suburban Real Estate Markets – Case Study. Real Estate Management and Valuation, 24(1), pp. 5-18.

Park, B. & Bae, Y.J., 2020. Predicting Property Prices with Machine Learning Algorithms. *Journal of Property Research*, 38(1), pp. 48–70.

Primior, 2023. Importance of Accessibility: How Location Affects Property Value. Available at: https://primior.com/importance-of-accessibility-how-location-affects-property-value/ [Accessed 26 Oct. 2024].

Ren, Y., An, N. & Zhang, X., 2020. The Use of Machine Learning in Real Estate Research. *Land*, 12(4), p. 740.

Shi, D., Zhang, H., Guan, J., Zurada, J., Chen, Z. & Li, X., 2023. Deep Learning in Predicting Real Estate Property Prices: A Comparative Study. In *Proceedings*

of the 56th Hawaii International Conference on System Sciences, pp. 1201–1207.

Sopranzetti, B.J., 2015. Hedonic Regression Models. In *Handbook of Financial Econometrics and Statistics*. Springer Science+Business Media New York, pp. 2119.

Taltavull de La Paz, P., 2021. Machine Learning with Explainability or Spatial Hedonics Tools? An Analysis of the Asking Prices in the Housing Market in Alicante, Spain. Expert Systems with Applications, 168, pp. 113-125.

Tiesdell, S. & Adams, D., 2011. *Urban Design in the Real Estate Development Process*. Chichester; Ames, Iowa: Wiley-Blackwell.

Vora, D.R. & Bhatia, G.S., 2023. *Python Machine Learning Projects*. Los Angeles: BPB Publications.

Wolny Kucińska, A., 2016. Accessibility of Real Estate by Transportation as a Determinant of the Development of Suburban Real Estate Markets – Case Study. *Real Estate Management and Valuation*, 24(1)