Trender för användningen av sambrukscyklar i Helsingfors under Covid-19

En utforskande dataanalys av ett öppet dataset

Hannes Holmlund

Lärdomsprov
Informationsteknik
2023

Lärdomsprov

Hannes Holmlund

Trender för användningen av sambrukscyklar i Helsingfors under Covid-19.

En utforskande dataanalys av öppet dataset.

Yrkeshögskolan Arcada: Informationsteknik, 2023

Identifikationsnummer:

8858

Sammandrag:

Helsingforsregionens trafik har en öppen databas som innehåller data på cykel lånesystemet i Helsingfors. I detta examensarbete utförs en utforskande dataanalys för att undersöka trender i Helsingforsregionens Trafiks öppna data. Genom åren har ett par händelser som Coronapandemin och elsparkcyklar påverkat användningen av lånesystemet. Syftet med examensarbetet är att undersöka Coronapandemins påverkan för användningen av lånesystemet och cyklarna. Metoden utforskande dataanalys används för att utnyttja all den tillgängliga öppna data och för att utvinna information om användningen på lånesystemet i Helsingfors. Med processen för utforskande dataanalys utförs också databehandling, dataanalys och datavisualisering. Alla delar av utforskande dataanalysens process utförs i Microsofts datavisualiserings program Power Bi. De visualiseringar som används för att undersöka Helsingforsregionens Trafiks data skapas också med hjälp av Power Bi. Från utforskande dataanalysen syns det att användningen av lånesystemet stiger varje år från 2016 till 2019. Under Coronapandemi åren 2020 och 2021 syns det att användningen sjunker jämfört med 2019. Det syns också att distansen och tiden på cykelturen ökar under Coronapandemin. Målet med examensarbetet är inte att förespråka för andra fordon som elsparkcyklar eller bil. Målet med examensarbetet är inte att bevisa ifall lånesystemet är lönsamt för Helsingforsregionens Trafik. I examensarbetet jämförs inte Helsingfors lånesystemet mot något annat lånesystem i världen.

Nyckelord:

Utforskande dataanalys, Lånesystem, Databearbetning

Degree Thesis

Hannes Holmlund

Trends for usage of bike sharing in Helsinki during Covid-19.

An exploratory data analysis of an open dataset.

Arcada University of Applied Sciences: Information technology, 2023.

Identification number:

8858

Abstract:

Helsinki Region Transport has an open database that holds data for the bike-sharing system in Helsinki. In this thesis an exploratory data analysis was performed to research the trends in Helsinki Region Transports open data. Over the years a couple of events such as the Corona pandemic and launch of E-scooters have affected the use of the bike-sharing system. The purpose of this thesis is to examine what effect the Corona pandemic has had on the bike-sharing system. The method exploratory data analysis is used to utilize all the available open data and to extract information about the use of the bike-sharing system in Helsinki. With the process of exploratory data analysis data processing, data analysis and data visualization are performed. All parts of the exploratory data analysis process are done in Microsoft's data visualization program called Power Bi. The visualizations that are used to research Helsinki Region Transports data are created in Power Bi. From the exploratory data analysis, you can see that usage of the bike-sharing system increases from 2016 to 2019. Under the Corona pandemic during 2020 and 2021 you can see that the usage decreases compared to 2019. You can also see that distance and time traveled per bike ride increases during the Corona pandemic. The goal for this thesis is not to advocate for other vehicles such as e-scooters or cars. The goal of this thesis is not to prove that the bike-sharing system is profitable for Helsinki Region Transport. In this thesis the Helsinki bike-sharing system is not compared to any other bike-sharing system in the world.

Keywords:

Exploratory data analysis, Bike-sharing system, data processing

Innehåll

1 Inledning

Sedan 2016 har Helsingfors haft ett lånesystem i bruk var användare kan betala för att låna cyklar från lånecykelstationer. Sambrukscyklarna använder sedan användaren för att ta sig från startstationen till slutstationen. Lånesystemen har många för- och nackdelar och Helsingfors systemet som baserar sig på stationer begränsar användaren från att avsluta cykelturen var som helst. År 2020 kom coronapandemin till Finland (Stenroos 2020) och ändrade på vardagen för många finländare. Sambrukscyklarna i Helsingfors gav en möjlighet till social distansering i jämförelse med annan lokaltrafik som bussar eller spårvagnar.

1.1 Syfte och mål

Syftet med detta examensarbete är att uppnå en förståelse för hur användningen av Helsingforsregionens Trafiks (HRT) sambrukscyklar förändrades. Genom åren har det varit ett par större händelser, såsom Coronapandemin och lanseringen av elsparkcyklar, som har påverkat användningen av lånesystemet för cyklar i Helsingfors. Denna studie undersöker användningen av sambrukscyklar före och efter början av Coronapandemin med hjälp av utforskande dataanalys. Den huvudsakliga forskningsfrågan är vilken påverkan har Coronapandemin haft på användningen av lånesystemet. Arbetet undersöker om användningen har minskat eller ökat samt om det har skett förändringar i exempelvis rusningstider.

Målet är att utvinna alla data som finns i HRT:s öppna databas (Helsingforsregionens Trafik, 2023) om lånesystem och utföra en utforskande dataanalys för att få insikter i trender. Processen inleds med dataförbehandling för att rensa öppna data från problematiska variabler eller skapa nya variabler för att enklare kunna visualisera informationen. Resultatet är visualiseringar av trender och fenomen i lånesystemets data.

1.2 Avgränsning

Målet med detta examensarbete är inte att bevisa ifall lånesystemet är lönsamt för HRT. Målet med detta examensarbete är inte heller att förespråka användningen av andra fordon så som elsparkbrädorna eller bil. Examensarbete kommer inte heller ta ställning till

ifall HRT borde ändra på sin öppna data eller på hur HRT samlar detta data. I examensarbetet kommer jag inte heller jämföra användningen av Helsingfors lånesystem mot användningen av något annat lånesystem i andra länder.

Viktig poäng för examensarbetet är att det inte är en bekräftande dataanalys utan det är enbart en utforskande dataanalys. Syftet med en utforskande dataanalys är att undersöka de fenomen och trender som finns i data men inte att bevisa en teori över den andra.

1.3 Arbetets uppbyggnad

Examensarbetets uppbyggnad består av sex olika kapitel. Efter det första kapitlet som är indelningen kommer kapitlet som handlar om lånesystem i sin helhet samt historiskt och Helsingforsregionens system. Efter lånesystem kommer kapitlet om data och metoder. I kapitlet diskuteras den data som används i detta examensarbete samt metoderna för databearbetningen och dataanalysen. Examensarbetets fjärde kapitel handlar om implementeringen databearbetningen och dataanalysen. I implementerings kapitlet får läsaren förståelse för hur databearbetningen och dataanalysen utförs med den diskuterade teorin och metoder från föregående kapitel. Det näst sista kapitlet i examensarbetet är resultatredovisning och evaluering av den information som har samlats in från databearbetningen och dataanalysen. I det sista kapitlet av examensarbetet går jag igenom mina egna slutsatser som jag har gjort baserat på valda metoder samt resultat från dataanalysen.

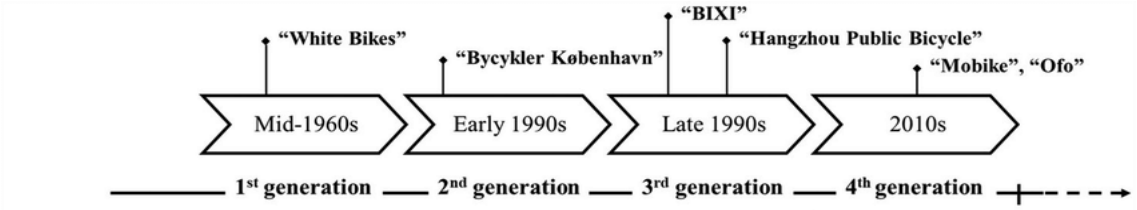
2 Lånesystem

Lånesystem är använda globalt i många storstäder för att underlätta användarens vardagliga transport. Lånesystem använder sig oftast av flera stationer varifrån användaren kan påbörja eller avsluta sin resa.

2.1 Generationer av lånesystem

Det finns flera variationer på lånesystem runt världen och dessa variationer kallas oftast för generationer. Totalt finns det fyra generationer av lånesystem var av några av dessa används runt världen. Varje ny generation av lånesystem har tillfört en eller flera nya funktioner för att underlätta användningen för användarna. Skillnaderna på dessa

generationer är som exempel att den tredje generationen av sambrukscyklar använder sig av stationer och den fjärde generationen av sambrukscyklar är dockningslösa. Dockningslösa lånesystem använder sig inte av stationer. Nedanför finns det en tabell med en visar skillnaderna på dessa generationer.



	Free bikes	Coin-deposit system	Docked IT-based system	Dockless IT-based system
Components	• Bicycles	• Bicycles • Docking stations	• Bicycles • Docking stations • Key/cards connected to payment • Mobile application	• Bicycles • Mobile application
Characteristics	• Free of charge • Users anonymous • Bicycles unlocked • Bicycles located at arbitrary places in a certain region	• Free of charge • Users anonymous • Bicycles locked in specific docking station • Coin as deposit used to unlock bicycles; deposit coin retrieved by returning bicycles to stations	• Usually paid as a membership service, or used temporarily • Users verified with real identity • Bicycles locked in specific docking station • Technology (smartcards, or user interface kiosks) used for picking up and dropping off bicycles in stations	• Pay a deposit for being a member (recently some operators have cancelled the deposit) • User verified with real identity • Charge based on usage time on mobile app or seasonal subscription • Bicycles locked • Bicycles located at arbitrary places • Smart locks, bicycle searching, locking and unlocking via specific mobile app
Cities	1-10	1-10	Over 1000	Over 200
Development paths	Originated in Amsterdam, the Netherlands	Originated in Copenhagen, Denmark, later expended to Europe	Originated in Portsmouth, UK, later expended across the world	First large scale implemented in Beijing, China, later expended to Asia, Europe and the Americas
Main financial resource	Public investment	Public investment	Public investment	Venture capital funding
Collaboration with public transit agency	NA	NA	• Through payment • Clustering the docks around public transport stations	Not yet

Figur 1, Tabell på lånesystem generationer (Chen 2019)

2.1.1 Generation 1

I diskussioner menar Schimmelpennick (2009, refererad i Demaio 2009) att första generationen av lånesystem lanserades i Amsterdam året 1965 och kallades ”vita cyklar”. Cyklarna i Amsterdam var vanliga cyklar som hade blivit målade vita. Idén var att ge allmänheten möjligheten att använda cyklarna fritt så att användaren kunde hitta en cykel och lämna den sedan vid sin destination för nästa användare att hitta. Problemen med systemet

märktes snabbt när en del cyklar blev omvandlade till privat bruk och en del cyklar blev slängda i kanaler. (DeMaio 2009)

2.1.2 Generation 2

Den andra generationen av lånesystem dök upp året 1991 hävdar Nielse (1993, refererad i Demaio, 2009) i de två danska städerna Farsø och Grenå. År 1993 startades ett till lånesystem i den danska staden Nakskov som bestod av fyra stationer och 26 cyklar. Lånesystemen i Danmark var ytterst små jämfört med de system som finns runt jorden i dag men de ledde till viktiga framsteg inom lånesystem funktionaliteten. Fyra år efter lånesystemen i Farsø och Grenå grundades så fick Köpenhamn också ett eget lånesystem med 1100 cyklar. Den andra generationen av lånesystem baserade sig på stationer varifrån användaren kunde betala med en myntdeposition för att låsa upp lånecykeln. Idén var att användaren skulle låna en cykel från en station och sedan föra den tillbaka till en station när hen hade kommit fram. Även om andra generationen av lånesystem tog klara framsteg i funktionaliteten så förekom ändå de samma problem som den första generationen hade. Användarens anonymitet gjorde det möjligt att stjäla eller förstöra cyklarna utan konsekvenser. (DeMaio 2009)

2.1.3 Generation 3

Med introduktionen av tredje generationens lånesystem så löstes problemet med stöld av cyklarna på grund av tekniska framsteg för lånesystemen menar DeMaio och Gifford (2004a, S.10). DeMaio och Gifford påpekar också att användare måste ge kreditkortsinformation så att ifall användaren inte returnerar cykeln så blir de fakturerade för ersättningen av cykeln. DeMaio (2009 s. 2) lyfter också fram att med den tredje generationen av lånesystem så kom tekniska framsteg som mobiltelefontillgång, elektroniska cykelställningar med lås, elektroniska cykellås, datorer ombord cyklarna och smartkort.

2.1.4 Generation 4

Fjärde generationen av lånesystem bygger på de tekniska framsteg som tredje generationen av lånesystem gjorde. Shaheen et al. (2013) beskriver den fjärde generationen av lånesystem som ett multimodalt system. Den fjärde generationen är stations baserade eller dockningslösa som delar smartkort med annan lokaltrafik. Den nya generationen skall

också innovationer i omfördelningen för att främja balansering av cyklar i systemet. (Shaheen et al. 2013 s. 85)

2.2 Helsingfors lånesystem

Helsingfors stad vill öka på antalet cyklister i trafiken och har därför implementerat ett lånesystem. Lånesystemet ansvaras av Stadstrafik Ab och Citybike Finland ansvarar för att upprätthålla tjänsten i Helsingfors. År 2023 har lånesystemet totalt 347 cykelstationer och 3470 sambrukscyklar i bruk. (Stadscyklar 2023)

2.2.1 Historia

Helsingfors tog i bruk sitt lånesystem år 2016 med 50 stationer och 500 cyklar och diskussioner runt sambrukscyklarnas utseende var livlig redan innan de blev använda skrivs det i en Yle artikel (Konttinen 2016). Under åren har cykelstationerna samt mängden cyklar ökat och år 2021 fanns det totalt 3 520 cyklar i Helsingfors och Esbo. Totala mängden lånecykelstationer var 242 i Helsingfors och 110 i Esbo. Under 2021 introducerades 105 nya stationer och 1050 nya sambrukscyklar. (Helsingfors stad 2021)

2.2.2 Funktionalitet

Eftersom Finland är ett nordiskt land så kan Helsingfors BSS system inte användas under vintermånaderna vilket gör att cyklarna är bara tillgängliga under månaderna april till oktober. (Stadscyklar 2023a)

Helsingfors lånesystemet är aktivt i både Helsingfors och Esbo. Lånesystemen i Helsingfors och Esbo är sammankopplade vilket ger möjligheten för användare att cykla mellan Helsingfors och Esbo på en och samma cykel. Vanda har också ett eget lånesystem som inte är sammankopplat med Helsingfors/Esbo systemet. Vanda använder sig av ett skilt lånesystem vilket gör att användaren inte kan färdas med samma cykel från Helsingfors till Vanda eller andra vägen. Helsingfors och Esbos lånesystem samt Vandas lånesystem använder sig av HRT:s mobilapplikation och resekort så användare har lätt tillgång till båda systemen. (Stadscyklar 2023b)

2.3 Problem med lånesystem

Som alla andra system så har också lånesystemen sina egna problem som ägarna måste lösa för att ge användaren den bästa möjliga upplevelsen. Ett av de tydligaste problemen med BSS är cykelkapacitet och tillgänglighet vid stationerna. När cyklarna används under dagen så kommer resorna som användaren gör högst troligen att börja och sluta med två olika stationer. Detta leder till att vissa stationer kommer att fyllas och tömmas i olika takt vilket göra att systemet hamnar i obalans. Det har gjorts många matematiska formler på detta problem för att hitta den bästa lösningen.

Det finns olika lösningar på detta problem som upprätthållarna av lånesystemet kan ta i bruk för att skapa balans igen. En lösning på detta problem är att manuellt flytta på cyklar från station till station. Lånesystemet i Helsingfors och Esbo använder sig av denna lösning. När lånesystemet implementerades så använde sig Stadstrafik två bilar som körde dygnet runt och manuellt flyttade på cyklarna. Bilarna flyttade cyklar från de stationer som var över kapacitet och flyttade cyklarna till stationer som var tomma eller nästan tomma. (Stadstrafik 2016)

Ett annat problem med lånesystem är skadegörelse och stöld av cyklarna. Om sambrukscyklarna blir förstörda med avsikt eller misstag så måste de repareras vilket skapar extra kostnader för ägarna. Skadegörelse av sambrukscyklar är ett ganska vanligt fenomen som förekommer i alla de städer med lånesystem. Stöld är också ett problem av sambrukscyklarna som skapar onödiga extra kostnader för ägarna av lånesystemet. Både stöld och skadegörelse av cyklarna kan lösas på flera olika sätt och storleken på problemet beror på hur lånesystemet fungerar i sin helhet. (Shaheen et al 2010 s. 165)

3 Data och metoder

För databearbetningen och dataanalysen så används Microsofts datavisualiseringsprogram Power Bi. Med Power Bi är det enkelt att komma åt alla data i HRT:s CSV filer för både databearbetningen och datavisualiseringen. Power Bi använder sig av två olika programmeringsspråk för både databearbetning och visualisering. De två programmeringsspråken är Data Analysis Expressions (DAX) och Power Query även känt som ”M”.

Power Query-programmeringsspråket är ett funktionellt och skiftlägeskänsligt språk som används i Analysis service, Excel och Power Bi för att kombinera och filtrera data från olika källor. Power Query används i databearbetningen för att städa upp eller formatera om data. (Microsoft 2023a)

Data Analysis Expressions (DAX) programmeringsspråket är ett bibliotek av olika konstanter, funktioner och operatorer. DAX används för att extrahera mera information genom olika funktioner exempel på dessa funktioner är matematiska funktioner som att räkna medianer och medeltal. (Microsoft 2023b)

3.1 Data

I sin bok beskriver Myatt (2014) rader i en tabell med data som observationer som innehåller information for den specifika observationen. Myatt (2014) lyfter fram ett exempel med en tabell om bilar. En tabell kan innehålla många observationer på olika bilar. En observation för en bil kan sedan ha information om bilens attribut till exempel bilens vikt, antal cylindrar och så vidare. Om dessa attribut sedan kan expanderas och användas på alla observationer i datatabellen så kan man kalla attributen för variabler. (Myatt 2014 s. 18)

Den data som används i detta examensarbete är taget från Helsingforsregionens Trafik (HRT) öppna data. HRT:s öppna data kommer i satser av årliga data börjandes från året 2016 till 2021. Man kan ladda ner data från HRT:s sidor i formen av Comma separated values (CSV) filer. Den tillgängliga data är formaterad på ett sätt som kallas ”Origin-Destination”. Denna formatering betyder att varje punkt av data innehåller information om startstationen, slutstationen, distansen som cyklades i meter, tid för färden samt tid för starten och avslutande av resan. (Helsinforsregionens Trafik 2023a)

Tabell 1, Exempel på data uppbyggnad

	Observation 1	Observation 2
Departure	2017-05-30 23:59:00	31/03/2020 23:55:54
Return	2017-05-31 00:06:00	01/04/2020 00:00:47
Departure station ID	016	249
Departure station name	Liisanpuistikko	Isosaarentie

Return station ID	022	248
Return station name	Rautatientori / länsi	Gunillantie
Covered distance (M)	1539	833
Duration (sec.)	409	288

Till dataanalysen används också öppna data om lånecykelstationerna som finns i Helsingfors och Esbo. Lånecykelstations datasetet är formaterat så att en observation är en station och variablerna är station ID, station namn, stationens adress, stad, operatör, cykelkapacitet samt geografiska X och Y koordinater.

3.2 Bearbetningsmetoder

För både databearbetning samt dataanalys kommer metoden utforskande dataanalys att användas. I sin uppsats om kvantitativ datarensning för stora databaser skriver Hellerstein (2008) att det mänskliga visuella systemet är en sofistikerad dataanalysmotor. Hellerstein (2008) lyfter också fram att det finns några klassiska datavisualiseringstekniker som står ut inom datarensning. Exempel på dessa visualiseringstekniker som Hellerstein lyfter fram är olika typer av histogram och låddiagram. (Hellerstein 2008 s. 29–30)

I sin bok skriver Buttrey (2017) om hur stegen för datarensning alltid beror på den data som man jobbar med och vad man förväntar sig att hitta. Buttrey (2017) skriver att innan man ändrar någonting behöver man titta på hur data ser ut. Generellt menar Buttrey (2017) att man skall söka efter dubbletter i variablerna och se ifall hela observationer är duplicerade. Buttrey (2017) lyfter också fram att man skall titta på varje kolumn för att hitta och räkna antalet tomma variabler i kolumnerna. Det är viktigt att göra dessa steg oberoende av hur stort dataset man jobbar med och det är nyttigt att göra histogram eller sammanfattnings statistik på de tomma variablerna. Buttrey (2017) på pekar att ofta när det saknas variabler i kolumner så fattas det också variabler i andra kolumner för samma observation. Detta beror ofta på att hela observationen har misslyckades att passa ihop med en datakälla när observationen skapades. Buttreys (2017) lösning på detta är att antingen ta bort dessa variabler eller sedan skapa en ny variabel som beskriver ifall observationen saknade dessa variabler eller inte. Buttrey (2017) lyfter också fram att valet att radera variabler eller hela kolumner ska vara en sällsynt taktik och i stället för att radera så borde man ändra på variablerna eller lägga till en variabel för att visa problem. Buttrey

(2017) skriver också om att dom flesta datarensnings projekt innehåller data från flera olika källor och på grund av detta är det bra att skapa en ny kolumn som håller en variabel om datakällan. (Buttrey 2017 s. 199–200)

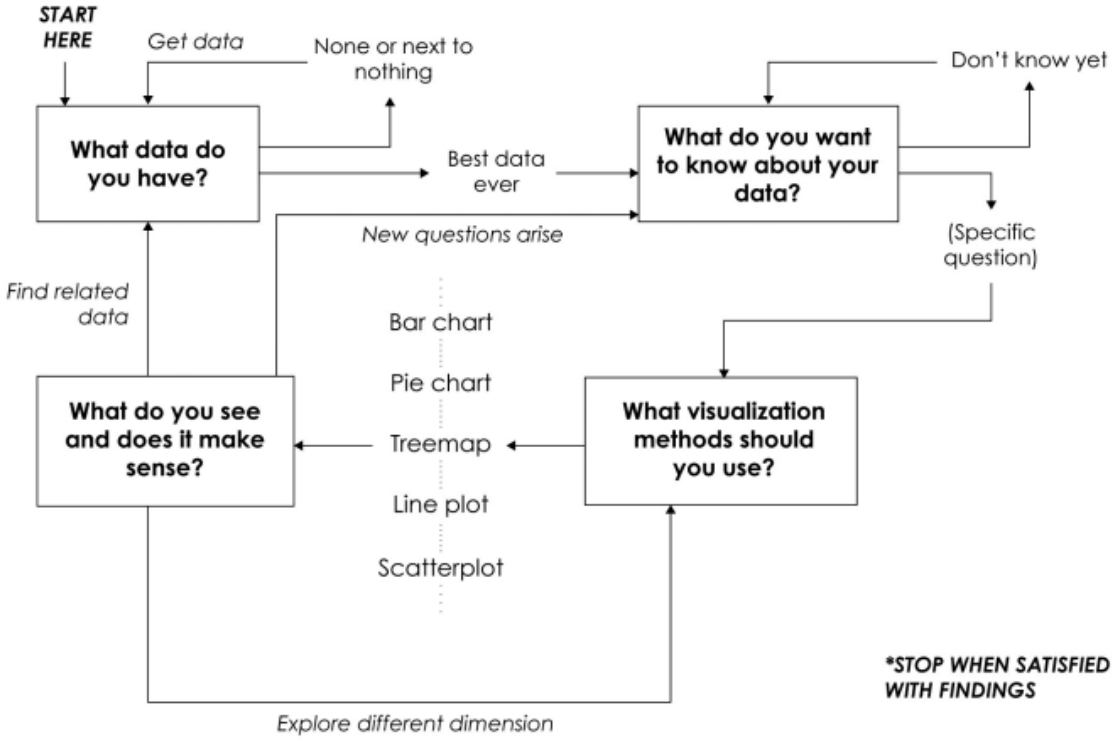
3.3 Analysmetoder

Dataanalysen i detta examensarbete görs med metoden utforskande dataanalys och syftet för dataanalysen är att undersöka användningen av sambrukscyklarna i Helsingfors under coronapandemin. I utforskande dataanalys är grafer gjorda för att lära känna den tillgängliga data så att man kan notera strukturer och trender i data (Cox 2017 s. 50). Seltman (2018 s. 61) beskriver de fem huvudorsakerna för utforskande dataanalys på följande sätt: kontrollerande av antaganden, upptäckande av misstag, preliminärt väljande av lämpliga modeller, fastställa samband mellan variabler samt att bedöma riktningen och storleken av sambanden mellan variablerna.

De steg som man skall ta i processen för utforskande dataanalys varierar alltid för varje dataset och projekt. I sin bok Data Points beskriver Yau (2013 s.136-137) de fyra frågor som man bör beakta när man utforskar data visuellt på följande sätt:

- Vilka data har du?
- Vad vill du veta om dina data?
- Vilka visualiseringsmetoder borde du använda?
- Vad ser du och är det vettigt?

Yau (2013) påpekar också att svaret för varje fråga är beroende av svaret på föregående fråga. Det är vanligt att hoppa mellan dessa frågor och själva processen är väldigt iterativ. Yau (2013) har skapat en figur (Figur 2) som ger en överblick på hela iterativa processen för utforskande dataanalys.



Figur 2, Den iterativa data utforsknings processen (Yau 2013 s. 137)

Yau (2013) lyfter också fram att om man visualiserar en aspekt från en stor mängd data så kan det leda till nyfikenhet av en annan dimension i ens data detta syns också i figur 2.

I deras artikel diskuterar Jebb et al (2017) om skillnaderna mellan utforskande dataanalys och bekräftande dataanalys. En viktig uppgift som Jebb et al (2017) tar upp i artikeln är att definiera skillnaderna mellan utforskande och bekräftande dataanalys. Jebb et al (2017) anser också att utforskande dataanalys har för länge låtits glida över till bekräftande dataanalyssidan. Jebb et al (2017) är förvånade av hur bekräftande dataanalys har blivit normen med tanke på hur begränsat användningen av data är i metoden. I bekräftande dataanalys är endast några förutbestämda hypoteser testade och data är inte undersökt på flera olika sätt. På samma gång som detta är styrkan med bekräftande dataanalys kommer det också med en stor bekostnad på metoden. När en hypotes misslyckas så kan samma data inte användas på nytt om man skall upprätthålla en sann bekräftande dataanalys. För bekräftande dataanalys metodens syfte är utforskande och maximerande av data användningen irrelevant. Jebb et al (2017) påpekar också att dessa begränsningar av bekräftande dataanalys inte nedvärderar metoden utan motiverar i stället behovet för både utforskande och bekräftande dataanalys. På grund av utforskande och bekräftande dataanalysmetodernas komplementära roller så krävs det en synergi mellan metoderna.

Skillnaderna mellan utforskande dataanalys och bekräftande dataanalys leder också till varför det är viktigt att lyfta fram vilken metod som används. Blandandet av utforskande dataanalys i strikt bekräftande sammanhang är olämpligt på grund av metodernas skillnader. Bekräftande dataanalysen syfte är att validera medan utforskande dataanalysen syfte är upptäckt. (Jebb et al 2017 s.)

4 Implementering

4.1 Databearbetning

Databearbetning börjar med att ladda in all data från CSV filerna till Power Bi programmet. I Power Bi har användaren möjligheten att ladda in flera filer samtidigt från en mapp i datorn. Power Bi programmet omvandlar automatiskt CSV filerna till tabeller i programmet. Detta används i databearbetningen för att lätt få in alla månader för ett år i en och samma tabell. När data från alla åren har laddats in så finns det tabeller för varje enskilt år och nästa uppgift är att ta bort onödiga variabler som har kommit med under överföringen av data till Power Bi. Från överföringen av data kommer en ny kolumn i tabellen som håller namnet på ursprungs filen varifrån data har hämtats. Den nya variabeln som håller ursprungsfil namnet omvandlas sedan för att ha en variabel med bara året som cykelturen gjordes. Detta görs med tanke på Buttreys (2017 s. 199–200) kommentar om att det är bra att hålla kvar variabler om källorna för data. Variabeln som innehåller året för när observationen skapades gör det också lättare att sortera åren i dataanalysen.

Nästa bearbetningssteg är att dela på datum- och tidvariabeln så att det blir två skilda variabler för datum och tid. Detta steg görs för både start och slut datum så att det slutgiltiga resultatet är fyra nya variabler som är startdatum och starttid samt slutdatum och slut tid. Kolumnen som innehåller information om startdatumet används också för att skapa en ny kolumn som håller information om vilken veckodag datumet har inträffat. Sedan skapas en ny kolumn som innehåller information om året samt månaden för observationen. Den nya variabeln är formaterad så att först kommer året sedan ett bindestreck och till sist månaden förkortade till tre bokstäver. För att i dataanalysen kunna sortera den nya kolumnen med året och månad skapas en till kolumn som innehåller en variabel om året och månaden som siffra.

Till näst skapades en ny anpassad kolumn för att kategorisera observationer. Den nya variabeln gjordes med booleanska värden så att om observationens distans i meter var tom så fick den nya variabeln värdet sann och om distansen inte var tom så blev värdet falskt. Tanken bakom den nya variabeln var att enkelt kunna se vilka observationer som har tomma värden och sedan kunna visualisera dem.

Under databearbetningen skapades en ny kolumn för tabellen som håller data för alla år. Den nya kolumnen innehåller en textvariabel som beskriver var cykelturen har börjat och slutat. Det finns totalt fem möjliga formationer som variabeln kan ha. Dessa formationer är Espoo, Helsinki, Helsinki-Espoo, Espoo-Helsinki och station ID saknas. Om variabeln är enbart Espoo eller Helsinki betyder det att både startstationen samt slut stationen är i Esbo eller Helsingfors. De två andra möjligheterna är formade så att först kommer start staden och sedan slut staden. Detta betyder att om observationen har Espoo-Helsinki så har cykelturen startat i Esbo och sedan slutat i Helsingfors.

4.1.1 Dataproblem

I datasetet som används för detta examensarbete finns det en del observationer och variabler som måste städas eller helt och hållet tas bort för att kunna utföra en bra dataanalys. Med hjälp av utforskande dataanalys och visualiseringar så hittar man snabbt problem i variablerna för den tillgängliga data. Programmet Power Bi ger också användaren möjlighet att sortera kolumner när man tittar på tabellerna för data. Detta underlättar också sökandet av problematiska variabler inom datasetet. De olika problemen som observationerna har är till exempel distansen för resan är noll meter, tiden det tog att utföra resan saknas, startdestinations ID eller namn saknas, slutdestinationens ID eller namn saknas och resans slut tid saknas. Det finns också observationer där cyklad distans är flera miljoner på minus. Logiskt så är det inte möjligt att cykla flera miljontals meter och det borde inte heller var möjligt att få negativa värden på distansen. Steget som togs i denna analys var att skapa en ny variabel för observationerna som saknar data. Som Buttrey (2017 s. 199–200) konstaterade att man skall helst inte ta bort hela observationer eller variabler utan hellre ändra eller skapa nya variabler.

I databearbetning märktes också att i filen som håller data på oktober månad 2021 saknar information om cykelturens längd i sekunder för alla observationer. Det första som gjordes för denna upptäckt var att granska överföringen av data från CSV filerna till Power Bi. Från granskningen hittades ingen orsak till att oktober 2021 skulle sakna all den informationen. Nästa steg som togs var att granska CSV filen ifall den också saknar informationen vilket den gjorde. Det som gjordes för undersökande var att granska en ny nerladdad fil från HRT. Den nya nerladdade filen öppnades som CSV fil i programmet Notepad var det märktes att samma information saknades.

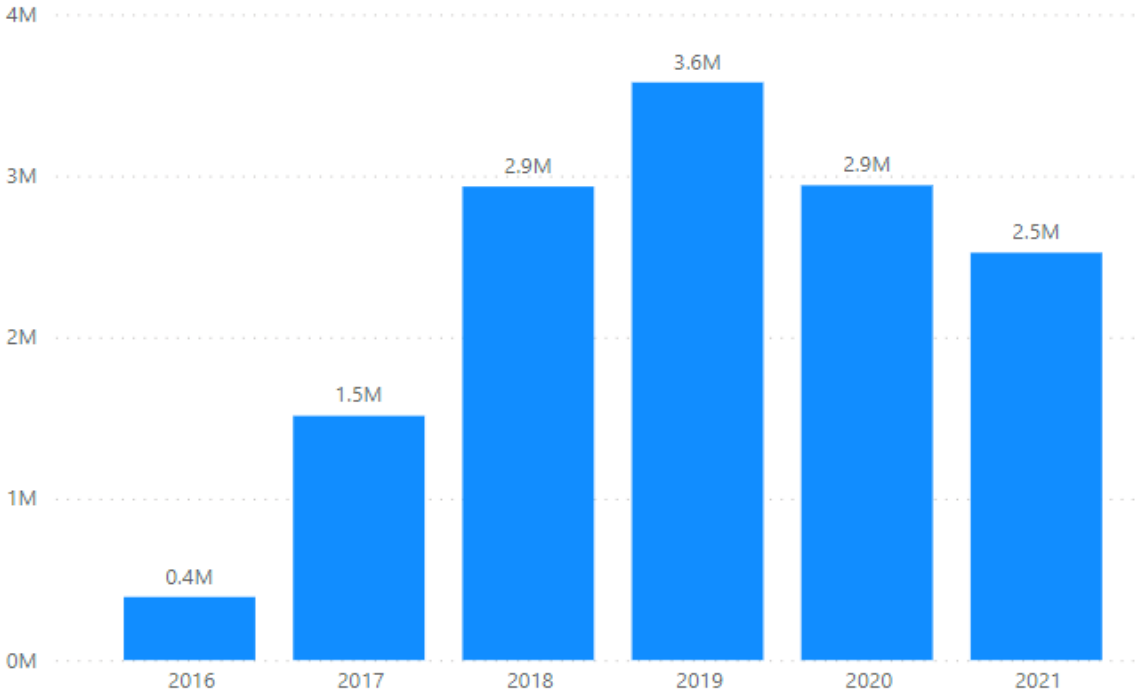
4.2 Dataanalys

Den utforskande analysen av data började redan under databearbetningen för att hitta information om korrelationer och problem i datasetet. Den ursprungliga frågan för utforskande dataanalysen var om användningen av sambrukscyklarna i Helsingfors har ändrats under coronapandemin.

För dataanalysen filtreras bort värden som antas komma från tekniska fel i datasamlingen. Det filtreras också bort sådan information som kan komma från konstig användning av sambrukscyklarna. Ett exempel på dessa så kallade konstig användning är observationer med en distans mellan noll och femtio meter. Dessa värden kan vara tekniska eller komma från att användaren faktiskt har cyklat så kort sträcka. De observationer som filtreras bort har variabler med värden som är under femtio meter resedistans, genomsnittlig kilometer per timme som är under en kilometer i timmen eller över tjugofem kilometer i timmen och sådana observationer var resans tid har tagit längre än fem timmar. Observationer med resedistansen under femtio meter används inte för att stationerna i Helsingfors är inte så nära varandra. Observationer med medelhastigheten över tjugofem kilometer i timmen används inte eftersom i en studie av Virkler och Balasubramanian (1998, refererad i Khan & Raksuntorn 2001 s. 220) räknade de ut att genomsnitt farten för cyklister på olika cykelvägar var mellan 19,6 kilometer i timmen och 24,9 kilometer i timmen. Data filtreras också under fem timmar på grund av att HRT debiterar åttio euro av användaren ifall de överskrider fem timmar (Helsingforsregionens trafik 2023).

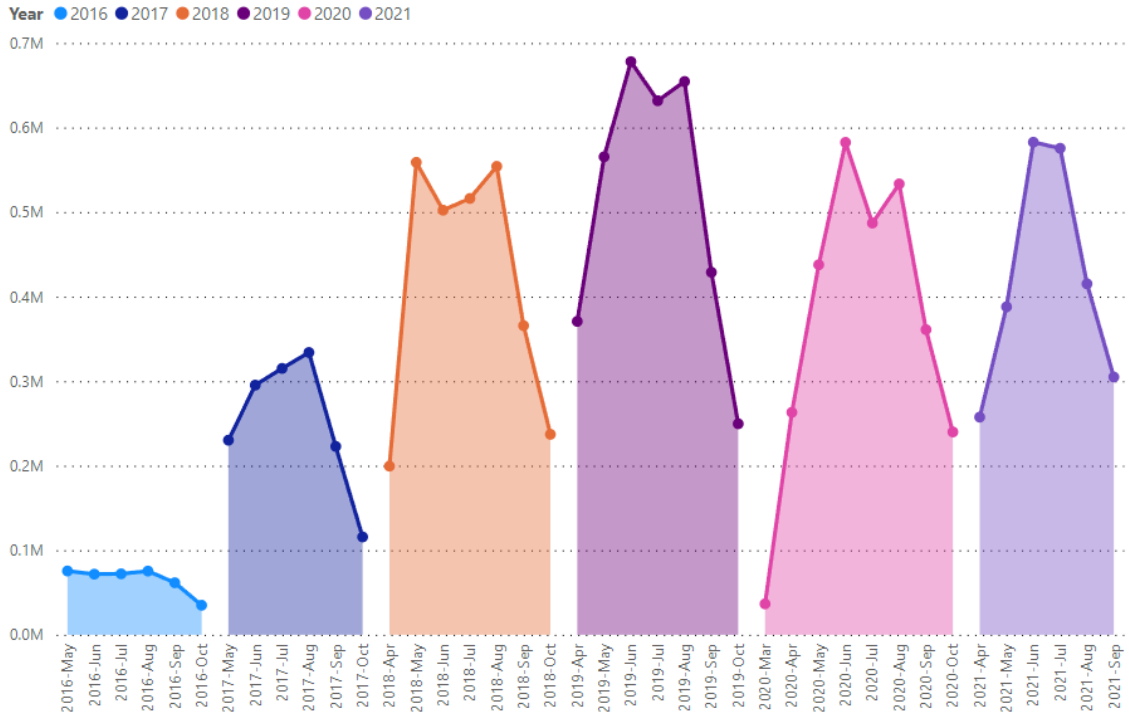
5 Resultatredovisning och evaluering

I den årliga användningen ser man att det klar har minskat på användningen av sambrukscyklarna. Användningen ökades ordentligt mellan 2016 och 2019 men från 2020 till 2021 så minskades användningen båda åren. Coronapandemin kom till Finland i mars 2020 (Stenroos 2020) och totala mängden cykelresor sjönk från 3,6 miljoner till 2,9 miljoner.



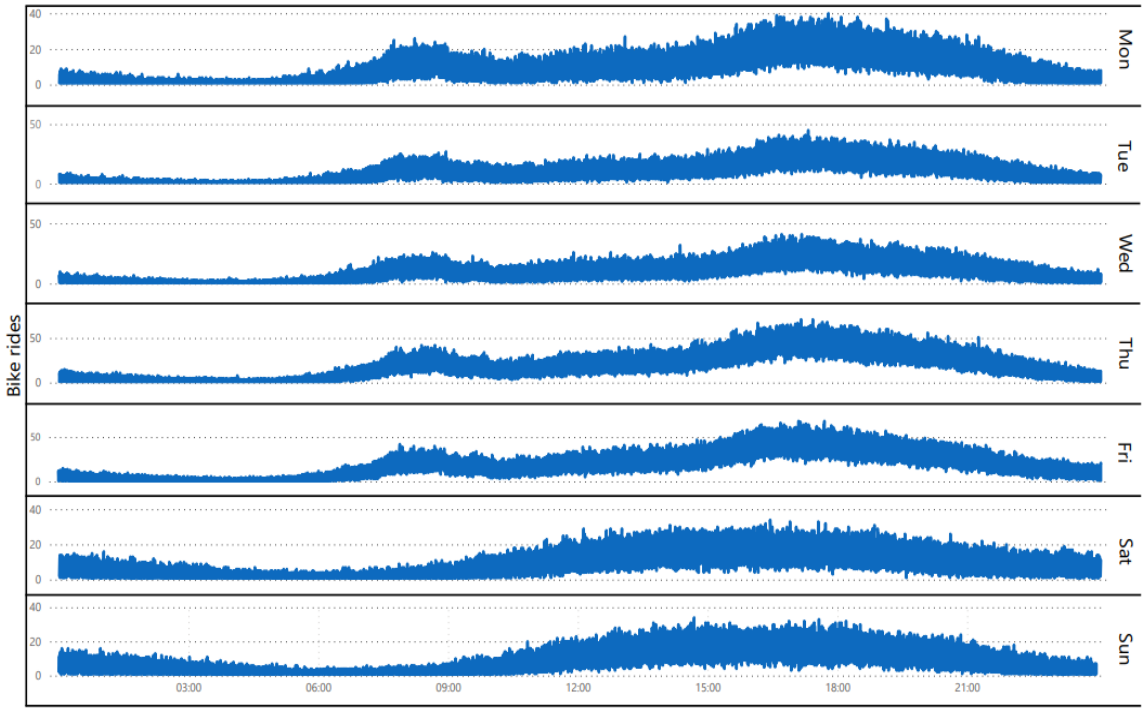
Figur 3, Totala användningen genom åren

Den totala användningen för varje månad (Figur 4) genom åren syns i figur 4. På figuren kan man se att användningen av sambrukscyklarna sjunker vid juni eller juli beroende på året. Efter att användningen har sjunkit under sommaren så ökar den igen till augusti. Året 2021 är enda året när användningen inte stiger mellan juli och augusti. I figuren kan man också se att oktober månad år 2021 saknas efter som hela månaden saknar data på hur långa cykelturerna har varit i sekunder.



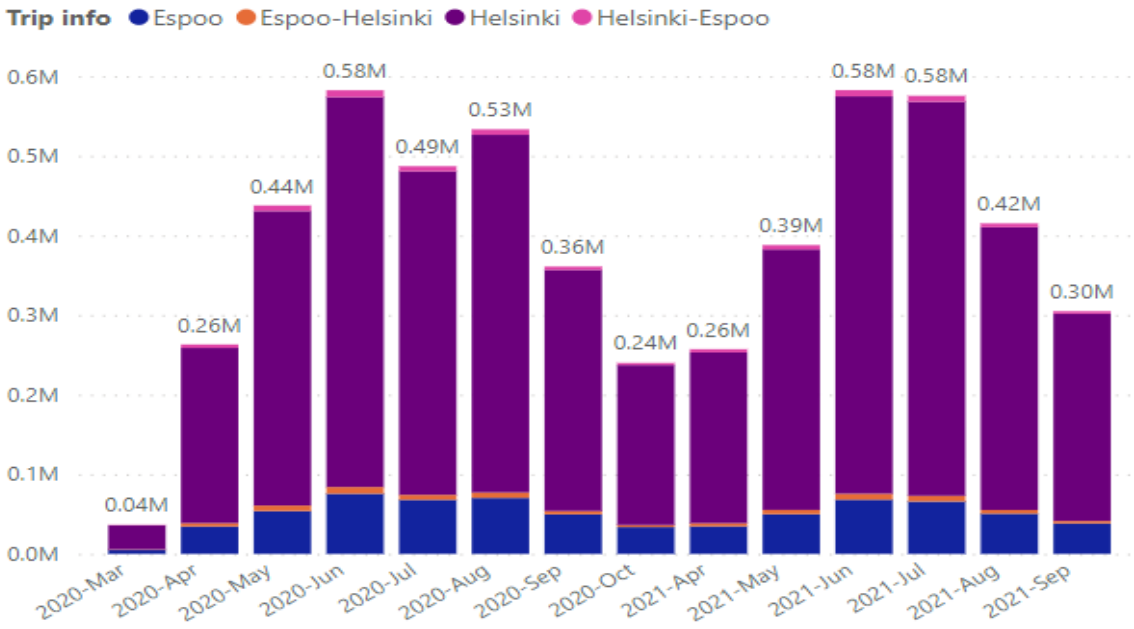
Figur 4, Antal cykelresor gjord för månad och år

I figuren (Figur 5) nedanför texten kan man se användningen av hyrcyklarna genom veckodagarna för åren 2020 och 2021. På vardagarna kan man se att användningen börjar stiga från klockan sex och sedan avtar när klockan närmar sig nio. Användningen av sambrukscyklarna stiger seden sakta genom dagen och toppar användningen mellan klockan tre och sex på kvällen. På helgerna börjar användningen stiga senare på dagen och håller sig ganska stabil ända in till natten. I figuren kan man också se att veckodagarna som användningen av cyklarna har varit högre är tisdag, onsdag, torsdag och fredag. I figuren ser man också att användningen av sambrukscyklarna ökar på nätterna mellan fredag till lördag och lördag till söndag.



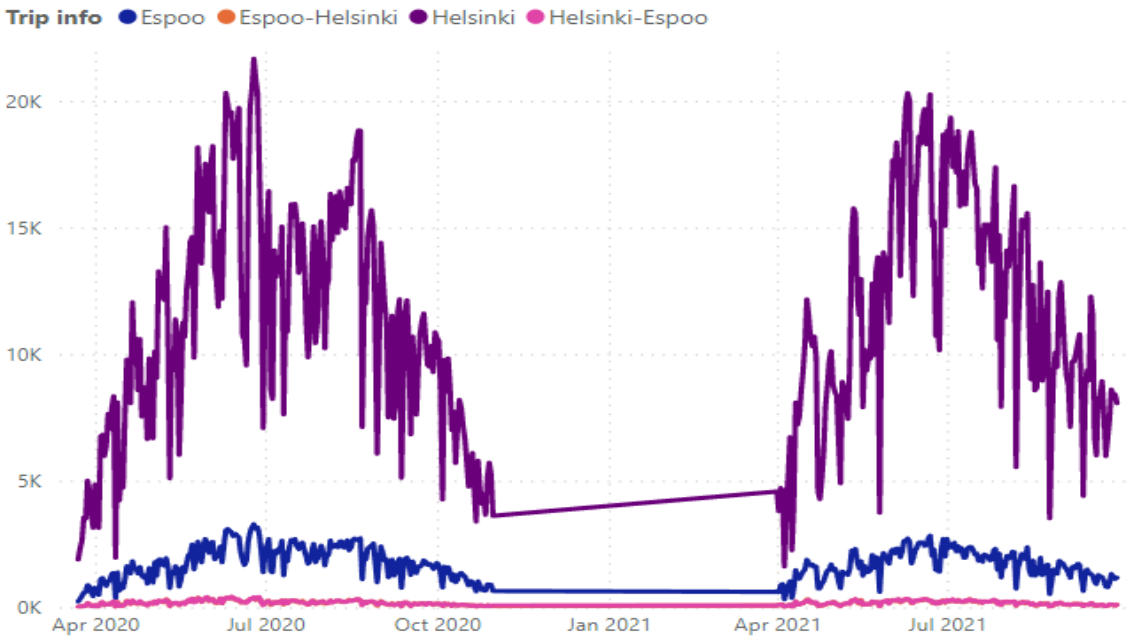
Figur 5, Antal cykelresor genom veckodagarna

I dataanalysen undersöktes också hurudana resor som användarna gjorde med sambrukscyklarna för åren 2020 och 2021. Kolumnen som skapades under databearbetningen som granskar start och slut stationerna för att kolla om resan startade, slutade eller enbart var i Esbo eller Helsingfors användes för figuren (Figur 6) nedanför. Likadant som i figur 4 (Figur 4) ser man att totala användningen av cyklarna minskar juli. Man kan också se från figuren att mängden resor gjorda med sambrukscyklarna från Helsingfors till Esbo och resorna gjorda från Esbo till Helsingfors är nästan samma.



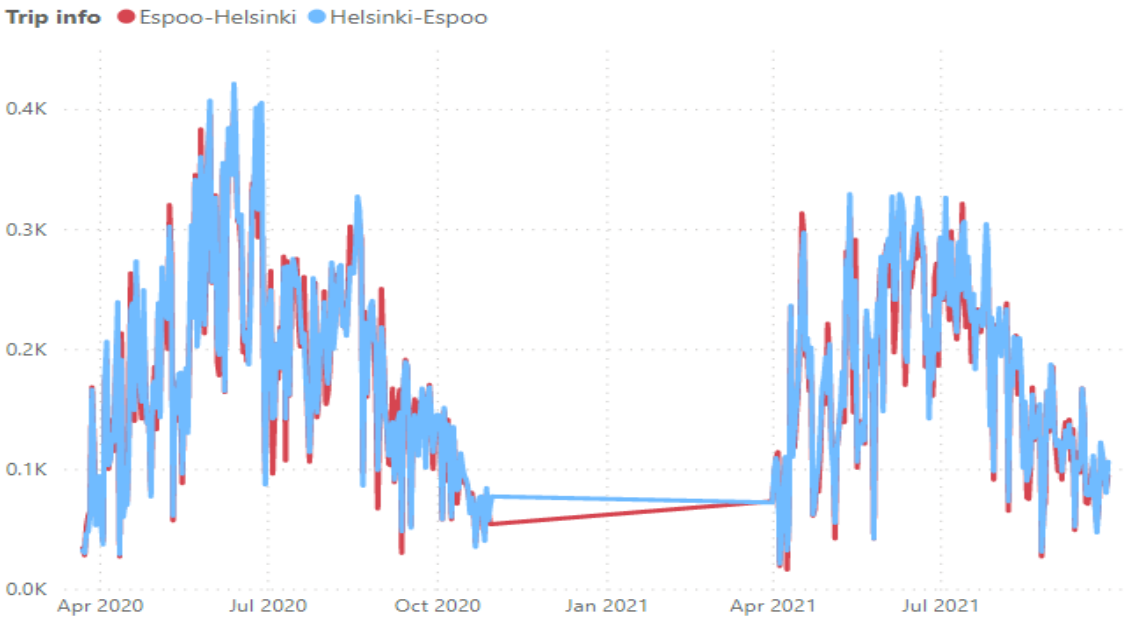
Figur 6, Typer av cykelresor månadsvis för åren 2020 och 2021

Figur 7 visar den dagliga användningen av cyklarna för åren 2020 och 2021. I figuren kan man se att de flesta resor sker innanför Helsingfors. Användningen i Esbo är näst störst för båda åren. Den rosa linjen i figuren visar antal resor som har haft startstation i Helsingfors och slut station i Esbo. I grafförklaringen finns det också en färg för cykelturerna som har startstation i Esbo och slutstation i Helsingfors. Antalet cykelturer som har inträffat med startstation i Esbo och Slutstation i Helsingfors samt de cykelturer som har varit startstation i Helsingfors och slutstation i Esbo ligger så nära varandra att de är svåra att urskilja.



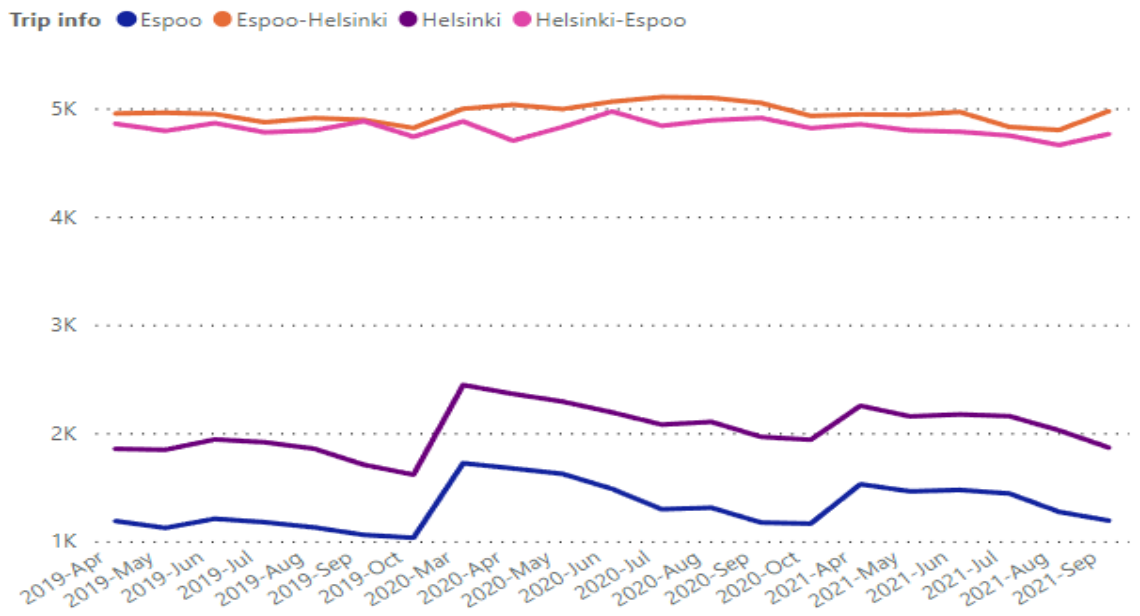
Figur 7, Typer av resor med daglig användning för åren 2020 och 2021

Figur åtta är en in zoomning på föregående figur (Figur 7) var bara cykelturerna mellan Helsingfors och Esbo syns. På figuren (Figur 8) kan man se att mängden på den dagliga användningen för cykelturer från Helsingfors till Esbo och cykelturerna som åker från Esbo till Helsingfors är väldigt nära varandra.



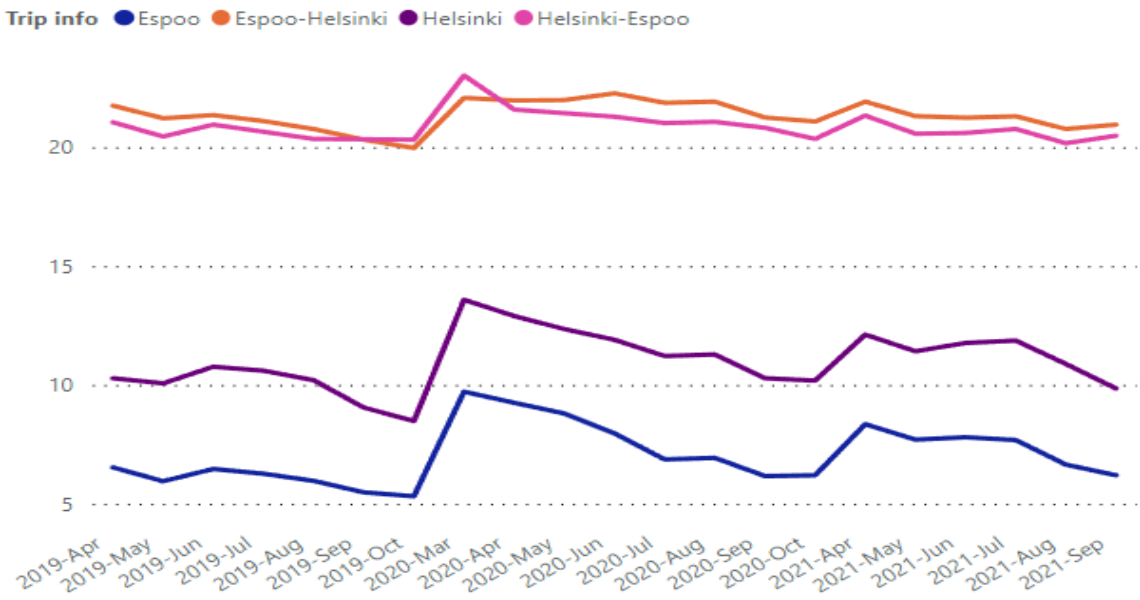
Figur 8, Daglig användning med resor mellan Helsingfors och Esbo för åren 2020 och 2021

I dataanalysen visualiserades också medianen för både cyklad distans i meter (Figur 9) och medianen för cykelturens längd i minuter (Figur 10). I figur nio användes också kategoriserings variabeln som visar vilken typ av cykeltur har gjorts. Median distansen i meter för cykelturer som har gjorts mellan Helsingfors och Esbo ligger runt 5 kilometer. Cykelturerna som har hållits innanför Helsingfors eller Esbo håller sig kring en till två kilometer. I figuren kan man se att medianen för distansen har ökat mellan åren 2019 och 2020. I Helsingfors och Esbo cykelturerna ökar också medianen i början av cykelsäsongen och sjunker sedan under säsongens tid. De cykelturer som cyklas mellan Helsingfors och Esbo hålls ganska stadigt runt 5 kilometer både genom åren 2019, 2020 och 2021 samt cykelsäsongerna. Även om cykelturerna mellan Helsingfors och Esbo inte har lika stora variationer genom åren och säsongerna så ser man att medianen ökar också mellan 2019 och 2020.



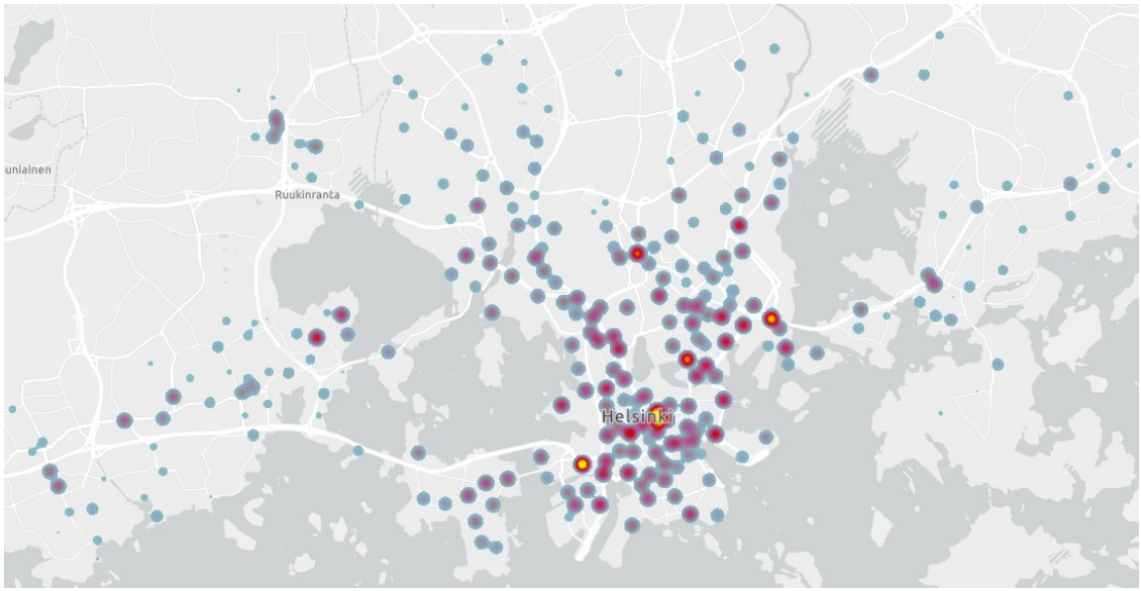
Figur 9, Medianen för distansen cyklad i meter

I figur 10 som visar median för minuterna cyklade kan man se samma trend som i median för meter cyklade. Likadant som med medianen för meter så ökar medianen för de cyklade minuterna mellan åren 2019 och 2020 i alla kategorier. För cykelturerna mellan Helsingfors och Esbo håller sig medianen ovanför 20 minuter. För cykelturerna som håller sig innanför Helsingfors håller sig medianen för tiden mellan tio till femton minuter för åren 2020 och 2021. Under 2019 låg medianen lite över eller under tio minuter. I Esbo var median för tiden på cykelturerna lägst av dom andra kategorierna.



Figur 10, Medianen för cykelturerna i minuter

Figuren nedanför (Figur 11) är en värmekarta över Helsingfors och Esbo. På värmekartan ser man totala användningen per station för båda åren. Bollarna på kartan som är gula i mitten med en stark röd ring är de stationer som har högst användning. Bollarna som är blåa har mindre användning än de röda. Storleken på bollen betyder också på användningen. Exempel på detta är de blåa mindre bollarna som är längre ut från Helsingfors centrum har mycket mindre användning än de röda bollarna mitt i Helsingfors centrum. Eftersom det finns ställen med stationer väldigt nära varandra så klumpar bollarna ihop sig så en boll kan till exempel vara två eller tre stationer.



Figur 11, Värmekarta med totala användningen för åren 2020 och 2021

Under dataanalysen skapades också en tabell som innehåller information om medianen och medeltal för Helsingfors lånesystem genom åren. Medelvärdet för resans längd i minuter sjunker mellan åren 2016 och 2019 för att sedan öka tillbaka mellan åren 2020 och 2021. Medianen på resans längd variera på en minut plus eller minus genom åren men håller sig stadigt runt tio minuter. Både medelvärdet och medianen för cykelturernas längd stiger från 2019 till 2020. Värdena för distansen varierar på både medelvärdet och medianen som högst med några hundra meter och som minst med åtta till trettio meter.

Tabell 2, Medelvärden och medianer för åren 2016 till 2021.

År	Medelvärde på resans längd i minuter	Median på resans längd i minuter	Medelvärde för distans (Meter)	Median för distans (Meter)
2016	15	9	1 965	1594
2017	14	11	2 200	1852
2018	13	10	2 175	1785
2019	12	10	2 183	1752
2020	14	11	2 520	2076
2021	14	11	2 487	2060

6 Slutsatser

Från utförandet av utforskande dataanalysen så har jag hittat intressant information om användningen av sambrukscyklarna i Helsingfors. Den information som jag har hittat har jag sedan visualiserat med en del grafer så att läsaren får en bättre blick över HRT:s öppna data på lånesystemet i Helsingfors. Den huvudsakliga forskningsfrågan om sambrukscyklarnas användning har ändrats under coronapandemin anser jag att har blivit besvarad. Min hypotes i början av arbetet var att användningen av sambrukscyklarna har minskat under coronapandemin var sann. I figur tre (Figur 3) kan man klart se att användningen sjunker från 2020 men inte lika mycket som jag tänkte mig. Med tanke på det undantagstillstånd som kom med coronapandemin trodde jag att man skulle kunna se en mycket större skillnad i användningen. Det som jag tycker är intressant är ökningen på medianen för distansen cyklad och tiden för cykelturerna i början av 2020. Det syns en väldigt tydlig ökning när man tittar på figurerna (Figur 9, Figur 10). Året 2020 är också enda året som lånesystemet har varit i bruk under mars månad och med tanke på att coronapandemin började i Finland samma månad så användes kanske cyklarna längre sträckor. Intressant tycker jag att antalet dagliga cykelturer från Helsingfors till Esbo och antalet dagliga cykelturer från Esbo till Helsingfors är väldigt nära varandra (Figur 8). Hypotetiskt så kan man tänka sig att det är mycket av samma människor som dagligen cyklar fram och tillbaka mellan Esbo och Helsingfors. Tyvärr så fanns det inga data om användare i HRT:s öppna data.

Jag tycker att utforskande dataanalys fungerar bra så länge som man håller sig till metodens syfte. Jag håller med om det som lyftes fram i metoderna att det är viktigt att hålla

utforskande dataanalys och bekräftande dataanalys skilt och att man alltid skall lyfta fram vilken av metoderna som har använts. Idén för utforskande dataanalysen är att undersöka allting som går att hitta i data men bekräftande dataanalys söker svar på en färdigt formulerad hypotes. Detta tycker jag att är en viktig sak att komma ihåg eftersom detta examensarbete är utfört som en utforskande dataanalys på HRT:s öppna data.

De brister som jag själv ser i examensarbetet är att jag borde sett mera på användningen från tidigare år. I examensarbetet tar jag mesta dels upp hur användningen har varit under coronapandemin men jag själv anser att jag kanske borde ha visat mera på användningen genom alla år för att kunna ge en tydligare helhets blick. Jag ser också metodvalet som en av bristerna. Jag nöjd med utforskande dataanalysen som jag har utfört men om jag skulle göra arbetet på nytt så skulle jag byte till en bekräftande dataanalysmetod. Orsaken för bytet är att jag tror att det skulle gynna ämnet mera att föra fram validerade slutsatser som kan användas i framtiden.

6.1 Framtida arbeten

I framtida arbeten skulle det vara intressant att se en bekräftande dataanalys på HRT:s öppna data. Bekräftande dataanalys processen är så annorlunda från utforskande dataanalys processen. Jag tror att det finns många hypoteser man kan skapa från lånesystem datasetet som kan leda till intressanta fynd. Det som också skulle vara intressant att se i framtida arbeten är någon form av artificiell intelligens analys. HRT:s öppna data på sambrukscyklarna är kanske ännu för litet för att kunna göra arbeten med AI.

Källor

Buttrey, S. E., & Whitaker, L. R. (2017). A data scientist's guide to acquiring, cleaning, and managing data in r. John Wiley & Sons, Incorporated.

Cox, V. (2017). Translating statistics to make decisions : A guide for the non-statistician. Apress L. P..

Demaio, P. (2009) Bike-sharing: History, Impacts, Models of Provision, and Future. *Journal of public transportation, 12, 41-56* https://doi.org/10.5038/2375-0901.12.4.3

Demaio, P. & Gifford, J (2004) Will Smart Bikes Succeed as Public Transportation in the United States?, *Journal of public transportation, 7, 1-15* http://doi.org/10.5038/2375-0901.7.2.1

Hellerstein, J.M. (2008). Quantitative Data Cleaning for Large Databases. https://dataresponsibly.github.io/courses/documents/Hellerstein_2008.pdf

Helsingfors stad. (2021). The city bike season will start on 1 April. https://www.hel.fi/en/news/the-city-bike-season-will-start-on-1-april

Helsinforsregionens Trafik. (2023). Öppen data. https://www.hsl.fi/sv/hrt/oppen-data Öppen data 2023

Helsinforsregionens Trafik. (2023). Användarvillkor. https://www.hsl.fi/sv/stadscyklar/helsingfors/anvandarvillkor

Jebb, T, A. Parrigon, S., Woo, S, E. (2017). Exploratory data analysis as a foundation of inductive research. Human Resource Management Review, 27, 265–276 https://doi.org/10.1016/j.hrmr.2016.08.003

Khan, S., I., Raksuntorn, W. (2001). Characteristics of Passing and Meeting Maneuvers on Exclusive Bicycle Paths. Transportation Research Record, 1776(1), 220-228. 10.3141/1776-28

Konttinen, M. (2 Maj 2016). Kaupunkipyörät kurvaavat tänään Helsingin kaduille: "Mitä pidempään katsoo, sitä kauniimmaksi se muuttuu". Yle. https://yle.fi/a/3-8846554

Mircrosoft. (2023a). *Power Query M formula language.* https://learn.microsoft.com/en-us/powerquery-m/

Mircrosoft. (2023b). *Learn DAX basics in Power BI Desktop* https://learn.microsoft.com/en-us/power-bi/transform-model/desktop-quickstart-learn-dax-basics

Myatt, G. J., & Johnson, W. P. (2014). Making sense of data i : A practical guide to exploratory data analysis and data mining. John Wiley & Sons, Incorporated.

Seltman, H. J. (2018) Experimental Design and Analysis https://www.stat.cmu.edu/~hseltman/309/Book/Book.pdf

Shaheen, S. A., Guzman, S., & Zhang, H. (2010). Bikesharing in Europe, the Americas, and Asia: Past, Present, and Future. *Transportation Research Record*, 2143, 159–167. https://doi.org/10.3141/2143-20

Shaheen, S. A., Cohen, A. P., & Martin, E. W. (2013). Public Bikesharing in North America: Early Operator Understanding and Emerging Trends. *Transportation Research Record*, 2387, 83–92. https://doi.org/10.3141/2387-10

Stadstrafik. (2023). *Stadscyklar.* https://kaupunkiliikenne.fi/sv/trafigering/med-cykel/stadscyklar/

Stenroos, M. (21 Mars 2020). Taistelu tuntematonta vastaan – Näin hallitus sulki Suomen seitsemässä päivässä. Yle. https://yle.fi/a/3-11267255

Yau, N. (2013). Data points : Visualization that means something. John Wiley & Sons, Incorporated.