Helsinki City Bikes - Network Analysis – Final Project

What are Helsinki City Bikes?

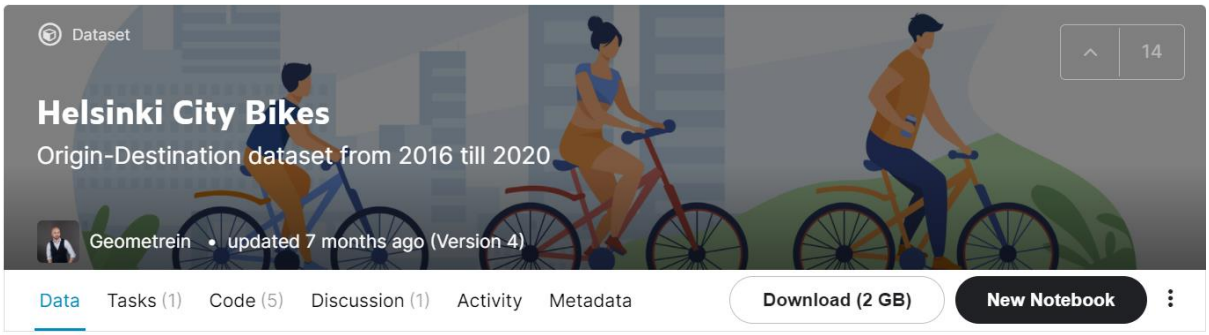
Helsinki City Bikes (HCB) are shared bicycles available to the public in Helsinki and Espoo metropolitan areas, also known as Greater Helsinki. The city bikes were introduced in 2016 as a pilot project with only 46 bike stations available in Helsinki. After becoming popular among the citizens, Helsinki city decided to gradually expand the bike network. In the period between 2017 and 2019, approximately one hundred stations were being added to the network each year. By 2019 the bike network reached its complete state with only 7 stations being added in 2020. As of 2020, there were 3,500+ bikes and 350 stations operating in Helsinki and Espoo.
To use the city bikes, citizens purchase access for a day, week or the entire cycling season that lasts from April to November. All passes include an unlimited number of 30-minute bike rides. For an extra fee of 1€/hour, you can use the bike for longer. Bikes are picked up and returned to stations that are located all around Helsinki and Espoo. Citizens can also use their HSL Transport Cards (which works like London's Oyster Card or the Israeli Rav-Kav).
If the User uses a city bike for more than 30 minutes at a time, an extra fee €1/hour, shall be charged to the User as per the price list. Once a city bike has been successfully returned to a bike station, the User can start a new city bike ride under the same terms. The 30-minute ride time with no extra fee also starts anew.
If a city bike is not successfully returned to a bike station within hours from picking up the bike, the User may be charged an €80 delay fee. If a city bike is not returned to a bike station or to the Service Provider within 24 hours from picking up the bike, the city bike shall be considered as lost.

Project Leading Questions:

We've entered the project with multiple leading questions, we wanted to research a transportation network and see what we can learn from it, and whether we can learn information that could help them make decisions.
The leading questions for our project are:

- **How does the average trip look like?**
- **What can we tell about intended paid trips?**
- **What is the most common route?**
- **Which stations are popular?**
- **Do the most used stations vary yearly?**
- **What are the main areas of the service?**

Source of data and full disclosure:



The source for our data come from Kaggle:
https://www.kaggle.com/geometrein/helsinki-city-bikes
it can also be downloaded from our Google Drive here
In Kaggle there are EDA and NA notebooks for this specific dataset, and we've used some of the graph functions from these notebooks, although we've adjusted them to fit our needs.
Therefore, you will see in our code that we mentioned explicitly when parts of it are used.
It is important for us to let you know that this is an original paper and we've tried to keep it and the research questions as original as possible.
Our data (pre-processing) is consisted of 12,157,457 trips that have 14 features:

departure	date and time of departure
return	date and time of return
departure_id	id of departure station
departure_name	name of departure station
return_id	id of return station
return_name	name of return station
distance	total distance ridden (meters)
duration	total duration of the ride (seconds)
avg_speed	average speed of riding (km per hour)

```
id                          int32
departure                  object
return                     object
departure_id               object
departure_name             object
return_id                  object
return_name                object
distance (m)              float64
duration (sec.)           float64
avg_speed (km/h)          float64
departure_latitude        float64
departure_longitude       float64
return_latitude           float64
return_longitude          float64
Air temperature (degC)    float64
dtype: object
```

As well as some features like ***latitude*** and ***longitude*** for the stations and ***temperature*** (Celsius).

Data Cleaning and EDA

NAs – we've found NAs in 4 features: ***avg_speed, return_latitude, return_longitude & temperature*.** Since most of the NAs (15902 of them) belonged to temperature and there were only 3552 more NAs (3550 of them in ***avg_speed***), we've figured we'd lose at most 0.16% of our data by dropping the NAs which seemed very reasonable to use with the size of our data, so we dropped them.
Outliers – we've then used both common logic and the terms of use for HCB to remove outliers from our data by setting upper and lower limits to distance and duration.

- ***Duration*** - Since HCB don't intend for the users to rent the bikes for a long period ("Return the bike to a bike station within five hours to avoid the €80 delay fee") we've

set the upper limit for our data at 5 hours. As for the lower bound we've decided that 2.5 minutes (150 seconds) is a reasonable time for a person to regret renting the bike and return them, so we've set the minimal duration to be 150 (secs).

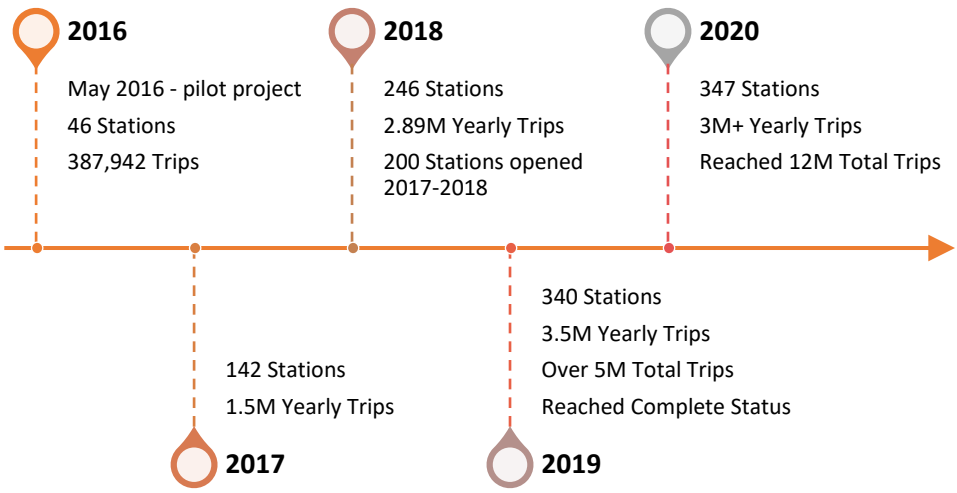
- ***Distance*** – as for the lower bound, there is no reference to the minimal distance between stations, however we've discovered in the HCB's FAQ that when the return station is full, the bike can also be returned in the station's area within 10-20 meters, therefor we we've set the lower limit of distance in our data at 20 (meters). As for the upper limit we've set it to 100km because we wanted to explore to possibility the bicycles are being misused for sport.
- ***Average Speed*** – Most outliers disappeared after setting bounds to duration and distance. We then set the upper limit to 15 km/h (half the speed of a professional cyclist), to remove the single outlier left in the data, while keeping enough data to explore longer rides.

After cleaning the data, we were still left with over 11 million trips:

	distance	duration	avg_speed
count	11197148.00	11197148.00	11197148.00
mean	2298.80	809.19	0.19
std	1652.19	889.77	0.06
min	21.00	151.00	0.00
25%	1142.00	388.00	0.16
50%	1872.00	623.00	0.19
75%	2998.00	1002.00	0.22
max	97833.33	17999.00	14.37

Exploratory Data Analysis

We've used this stage to answer a few of our leading questions, we've also used it to understand the yearly growth of the network in terms of number of stations (who will become nodes in our network) and yearly trips.

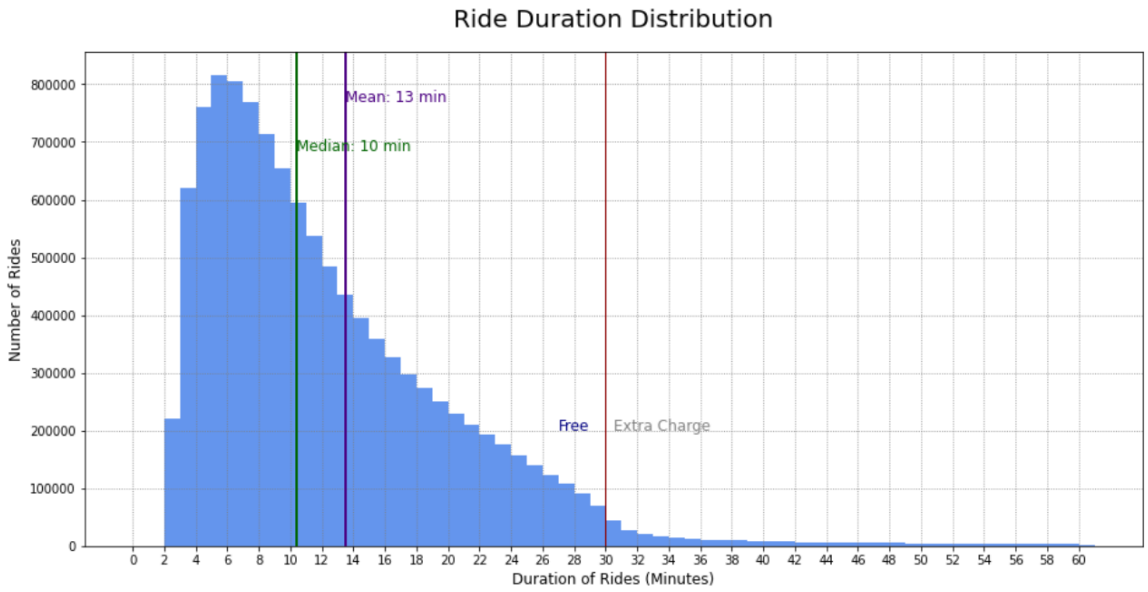


We've also explored the durations and distances of our trips and learned the following:

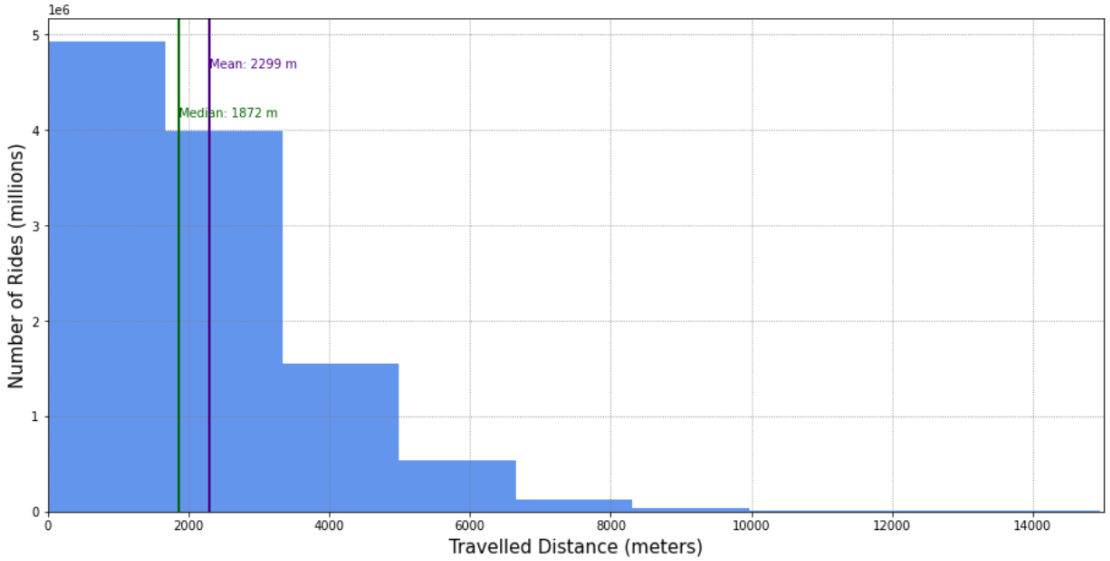
- 69.7% of the trips are under 15 minutes.
- Only 3.51% of the trips in our data were paid trips (over 30 minutes).
- The bikes were used over an hour for only 1.08% of the trips.
- Over the span of 5 years that the service exists, there were 36831 trips over 10km, that means the bikes are used for longer rides at least 1000 time a month (on average) since the service is only available April to October 31st
- More than half of the trips are under 2km.
- Service is used for longer rides (5km): 9.3% in 2020 compared to 6.5% so far

How does the average trip look like?

- The "average" trip will last 13 minutes and be 2.3km long, giving it an average speed of over 10km/h.
- This seems right to us as the median values for both distance and duration are near.
- The short duration and distance of the average trip suggests that the service is being used mostly for shorter rides, probably as a last mile solution to public transport.



Ride Distance Distribution



What is the most common route?

The most common route we found was **Aalto-yliopisto (M), Korkeakouluaukio** to Jämeräntaival (and the other way around) with over 65k trips during 2016-2020. **Aalto-yliopisto,** as we discovered relatively easy, is the University of Aalto, the 2nd most important Academic institution in Finland.

As for Jämeräntaival after we struggled a bit, we were able to find out there are students' residents, including one of the few residential buildings designed by Alvar Aalto in the area. The area (known as the Teekkari Village) housed the athletes competing in the 1952 Summer Olympics before the students moved in.

Another station we thought was worth mentioning since it appeared a lot in our most common route is Itämerentori. We will discuss it when we'll answer, '**Which stations are popular?**'.

```
def common_ride(dataframe):
    route = dataframe["departure_name"] + ' -to- ' + dataframe["return_name"]
    return Counter(route)
```

```
common_ride(df).most_common(10)
```

```
[('Aalto-yliopisto (M), Korkeakouluaukio -to- Jämeräntaival', 33590),
 ('Jämeräntaival -to- Aalto-yliopisto (M), Korkeakouluaukio', 33318),
 ('Itämerentori -to- Tyynenmerenkatu', 19504),
 ('Tyynenmerenkatu -to- Itämerentori', 18368),
 ('Töölönlahdenkatu -to- Baana', 16738),
 ('Itämerentori -to- Salmisaarenranta', 16280),
 ('Itämerentori -to- Töölönlahdenkatu', 16138),
 ('Töölönlahdenkatu -to- Itämerentori', 16030),
 ('Baana -to- Töölönlahdenkatu', 15970),
 ('Itämerentori -to- Itämerentori', 15844)]
```

What can we tell about intended paid trips?

Intended paid trips are trips that went over 30 minutes, that we suspect were intended to pass the 30 minutes free ride mark.

During our EDA we noticed most of the income from Helsinki City Bikes has to come from subscriptions since over 95% of the trips were made under 30 minutes i.e., free.

We've noticed not only that the number of longer trips increases yearly but that 2020 consist of almost 10% of total yearly rides which is the highest percentage among our data.

However, it is worth noting that the increase in longer ride can also be explained due to the COVID-19 pandemic and the fact that people are less willing to use public transportation to avoid exposure to the virus.

We've decided to explore the paid trips and try to figure out some info that could potentially help HCB:

	distance	duration	avg_speed
count	318981.00	318981.00	318981.00
mean	5628.44	4249.78	0.10
std	3892.28	3090.37	0.07
min	501.00	1921.00	0.00
25%	2875.00	2253.00	0.05
50%	4937.00	2967.00	0.09
75%	7345.00	4945.00	0.15
max	97833.33	17999.00	2.11

- Paid trips in 2020 compose 26.21% of the total paid trips.
- Most paid rides are under an hour long. The median is 49 minutes, and the mean duration is 71 minutes (over an hour).
- The mean distance of paid trip is 5.6km and the median is 5km.
- We can see from the ***avg_speed*** that it is very likely that the bikes were unused most of the trip time, as the max ***avg_speed*** is 2.1 km/h which is lower than the average speed of walking for adults.
- The top 10 most common route for paid trips all had the same return as the department station.

```
common_ride(paid).most_common(10)
```

```
[('Itämerentori -to- Itämerentori', 1200),
 ('Unioninkatu -to- Unioninkatu', 1087),
 ('Kaivopuisto -to- Kaivopuisto', 870),
 ('Hietalahdentori -to- Hietalahdentori', 790),
 ('Apollonkatu -to- Apollonkatu', 673),
 ('Liisanpuistikko -to- Liisanpuistikko', 635),
 ('Länsisatamankatu -to- Länsisatamankatu', 627),
 ('Sörnäinen (M) -to- Sörnäinen (M)', 606),
 ('Hakaniemi (M) -to- Hakaniemi (M)', 605),
 ('Isoisänsilta -to- Isoisänsilta', 577)]
```

Network Analysis:

```
Number of nodes: 347
Number of edges: 36287
Average degree: 209.1470

Top 5 nodes by degree:

('Haukilahdenkatu', 328)
('Itämerentori', 314)
('Laajalahden aukio', 305)
('Kamppi (M)', 299)
('Töölönlahdenkatu', 299)

Network density: 0.604471023304626
```

Our network represents the Helsinki city bike network. Nodes are the bike stations while the edges represent bike trips made between different stations.
As our data in very consequent and is made of data from 2016 to 2020, we decide to analyze it in different ways.

- Year by year
- All the years together

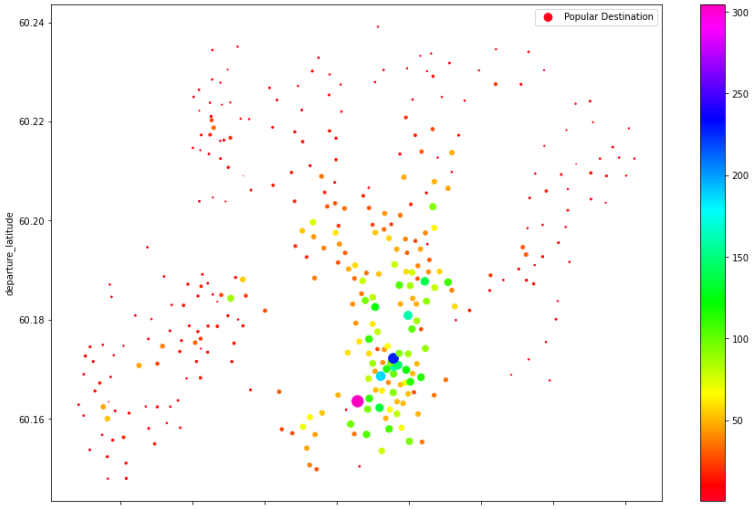
First, we choose to observe the difference between each year:

- In 2016 we had 46 node and 1,081 edges
- In 2017 we had 142 nodes and 8,788 edges
- In 2018 we had 246 nodes and 17,985 edges
- In 2019 we had 340 nodes and 29,189 edges
- In 2020 we had 347 nodes and 31,679 edges

We can observe that the average growth is around 100 new stations per year until 2019, then the network was considered completed and only a few new stations were built.
We decided to draw a network representation to observe our different network by year.(#Index1)
We observed that such a representation will not be helpful for our analysis because of the high density of our network. Instead, we will use other visualisation for our analysis.

Which stations are popular?

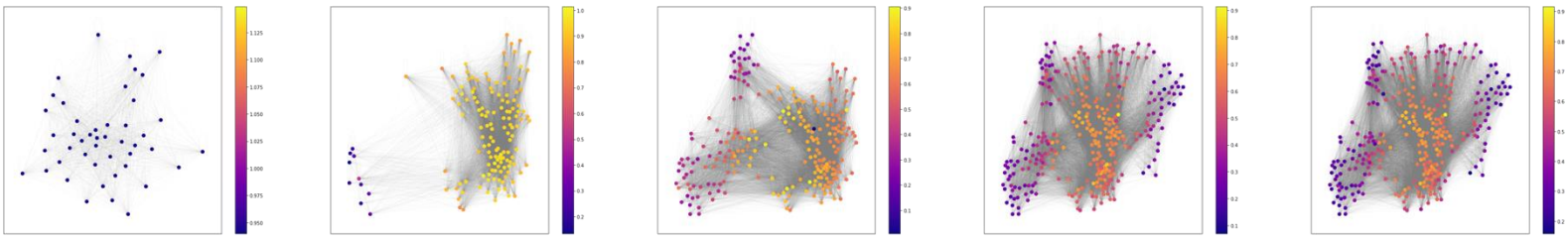
First, we plot the most popular station according to their location.
We can watch that 2 dots are more significant than the others, after we checked on google, those dots are Itämerentori and Haukilahdenkatu, we will talk about them later in this paper.



To answer this question, we decide to make 3 different graphs, based on 3 different measures.

- Degree centrality
- Closeness centrality
- Eigenvector centrality

At first, we built every graph for every year to observe how the network evolve through the years.
We observed a clear evolution of the network, at the beginning while having a small number of stations, every node had the same weight (for degree and closeness centrality), but through the year, we can clearly see that the network slowly centered itself. What was the center of Helsinki (the right side of the map until 2019) became surrounded by others station to connect it accorded to the needs of the population.
There is a graph of the evolution of the degree centrality from 2016 to 2020.



Other graphs for the other measures are in the index.
We can clearly affirm that the bike network has been achieved in 2020, the distribution of each station make sense.
Then, we decided to analyze only the last year for the simple reason that the city bike network get achieved in 2020. Looking at all the years together is likely to give more weight to old stations rather than to "really" popular station that has been built later.

- Degree centrality (index#2): Degree centrality is defined as the number of links incident upon a node (i.e., the number of ties that a node has).
  In our case the degree centrality is the number of bike stations that users have travelled to from the target station.
  We can observe that popular stations are located at the center of our network – the center of Helsinki. As we get away from the center, the degree centrality of the nodes is gradually descending. This fact make sense as we can imagine that people are coming from the peripheric to work in the center of the city where the schools, offices and public places are. The station with the highest Closeness centrality is Haukilahdenkatu. Research on google shows us that Haukilahdenkatu is one of the largest study centers in Helsinki, while students tend to use a lot of bikes, it makes sense that Haukilahdenkatu have a high degree centrality.

```
Top 5 nodes by degree centrality

[('Haukilahdenkatu', 0.9479768786127167),
 ('Itämerentori', 0.907514450867052),
 ('Laajalahden aukio', 0.8815028901734103),
 ('Kamppi (M)', 0.8641618497109826),
 ('Töölönlahdenkatu', 0.8641618497109826)]
```

- Closeness Centrality (index#3): Closeness centrality is a way of detecting nodes that are able to spread information very efficiently through a graph.
  It is calculated as the average of the shortest path length from the node to every other node in the network. In our case closeness centrality will represent how a station is likely to be a good intermediary station. As we know, a ride is free for 30 minutes, after that the client is charged money to continue his ride, people will sometimes prefer to stop their ride at a station, return the bike they had and take a new one to avoid fees and continue to ride freely.

```
Top 5 nodes by closeness centrality

[('Haukilahdenkatu', 0.9177718832891246),
 ('Paciuksenkaari', 0.8398058252427184),
 ('Itämerentori', 0.8357487922705314),
 ('Laajalahden aukio', 0.8357487922705314),
 ('Huopalahdentie', 0.8357487922705314)]
```

- Eigenvector centrality (index#4): Eigenvector centrality is a measure of the influence of a node in a network. Relative scores are assigned to all nodes in the network based on the concept that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. A high eigenvector score means that a node is connected to many nodes who themselves have high scores.
  In our case, this measure will a bit change the definition we made of "popular" node. Eigenvector centrality will indicate stations that are connected to popular stations. Popular stations are not only defined by their geographical position but by their closeness to other important stations. Again, Haukilahdenkatu is the station with the highest Eigenvector centrality.

```
Top 5 nodes by eigenvector centrality

[('Haukilahdenkatu', 0.08020299781378662),
 ('Paciuksenkaari', 0.07663091442344117),
 ('Huopalahdentie', 0.0764357926420222),
 ('Laajalahden aukio', 0.0758488077952895),
 ('Töölöntulli', 0.0757955486224381)]
```

Do the most used stations vary yearly?

To answer this question, we checked for each year, which stations were in the top 5 stations by degree.

```
Best stations that were both in 2020 and in 2019
[['Haukilahdenkatu'], ['Itämerentori'], ['Huopalahdentie']]

Best stations that were both in 2019 and in 2018
[['Itämerentori'], ['Haukilahdenkatu']]

Best stations that were both in 2018 and in 2017
[['Itämerentori']]

Best stations that were both in 2017 and in 2016
[['Erottajan aukio']]
```

One station stays in the top 5 station for 3 years: Itämerentori.
Itämerentori is one of the highest buildings in Helsinki (even though it has only 19 floors), containing offices, shops, supermarkets and more. It makes sense that many people will want to reach this place as it is clearly a central place as we can also see in the most common routes. See index#5, an interactive map we found showing the weight of the Itämerentori node, we can see that it is the huge node in the middle of the map, showing how popular this station is.

Recommendations:

To our understanding, the service is mostly used as a last mile solution and not really to perform long trips, as we can see from the research about the longer rides (specifically the ***avg_speed*** for 75% of those trips is under 0.15 *km/h*) and the fact that most of the trips are under 15 minutes.

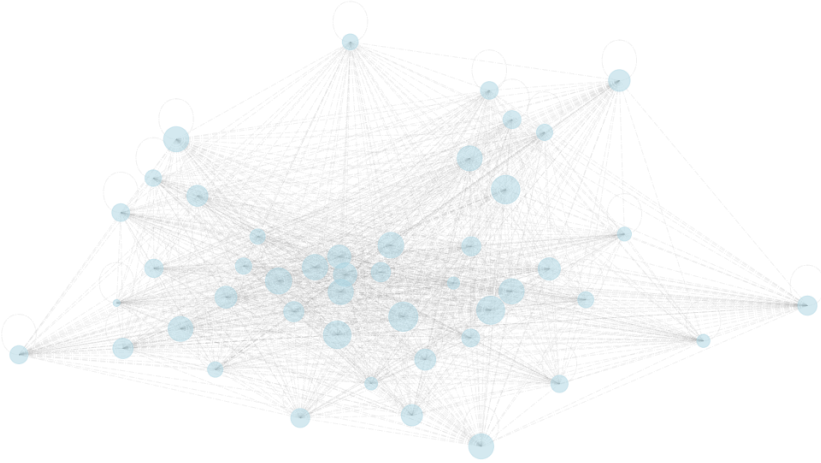
On the other hand, we observed a clear increase in paid trips in 2020, as we explained, possibly due to the COVID-19 pandemic.

We would advise to analyze the trips that occurred in 2021 to understand if this trend is continuing and maybe adapt the service for those longer rides, as we know this service is partly financed by the Helsinki municipality and is made to help Helsinki Citizens and so if there is a clear need for longer trips, the model should be adapted to this fact.

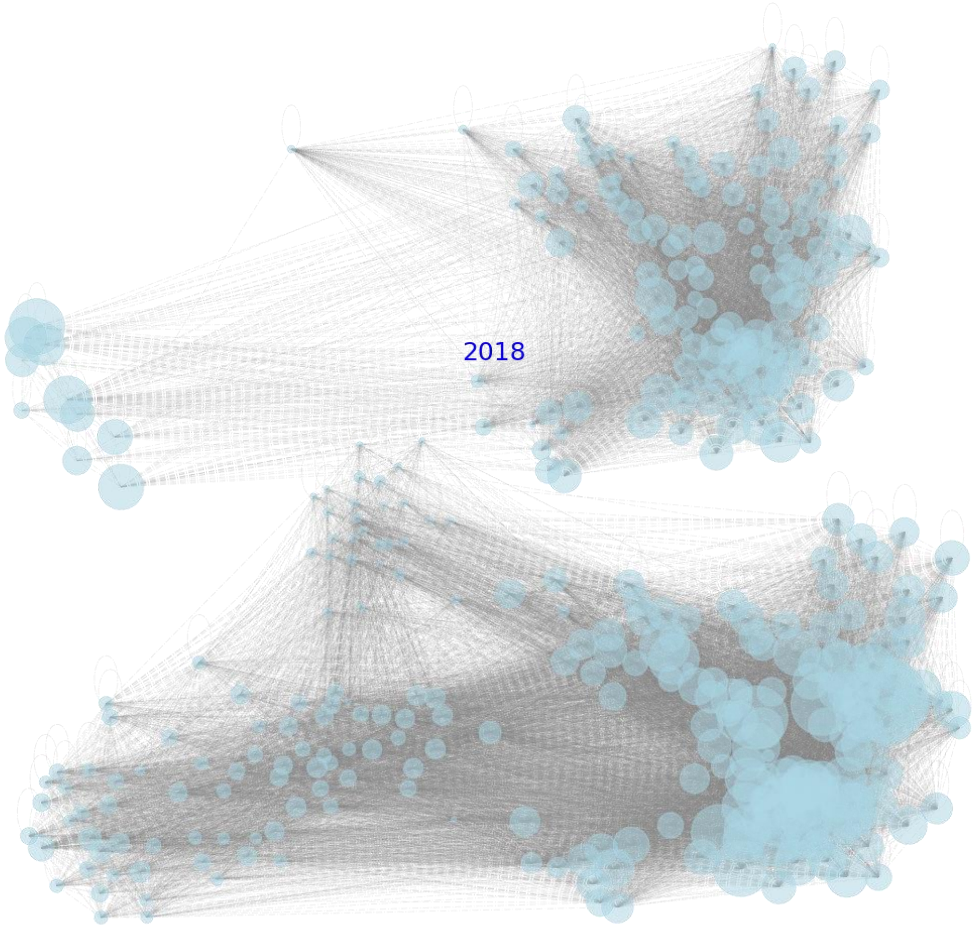
Index:

1. Network representation by year – Link back to the text

2016



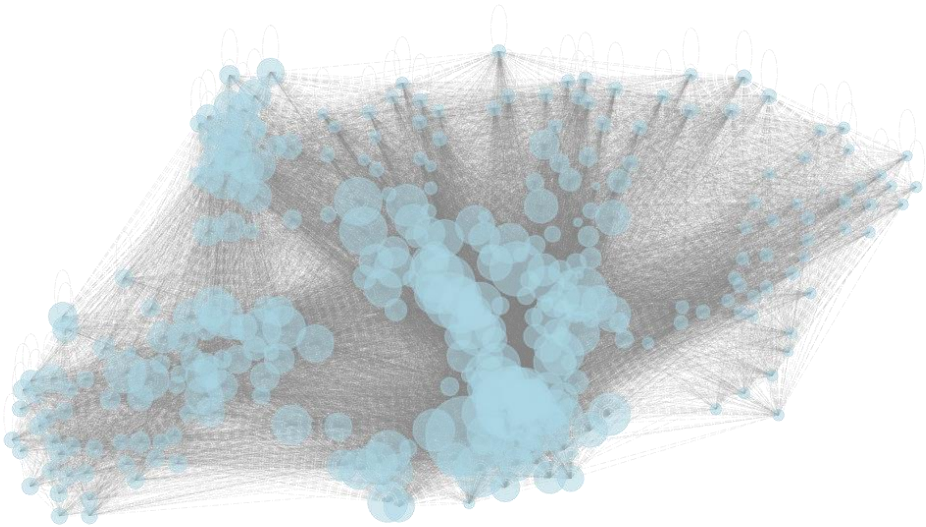
2017



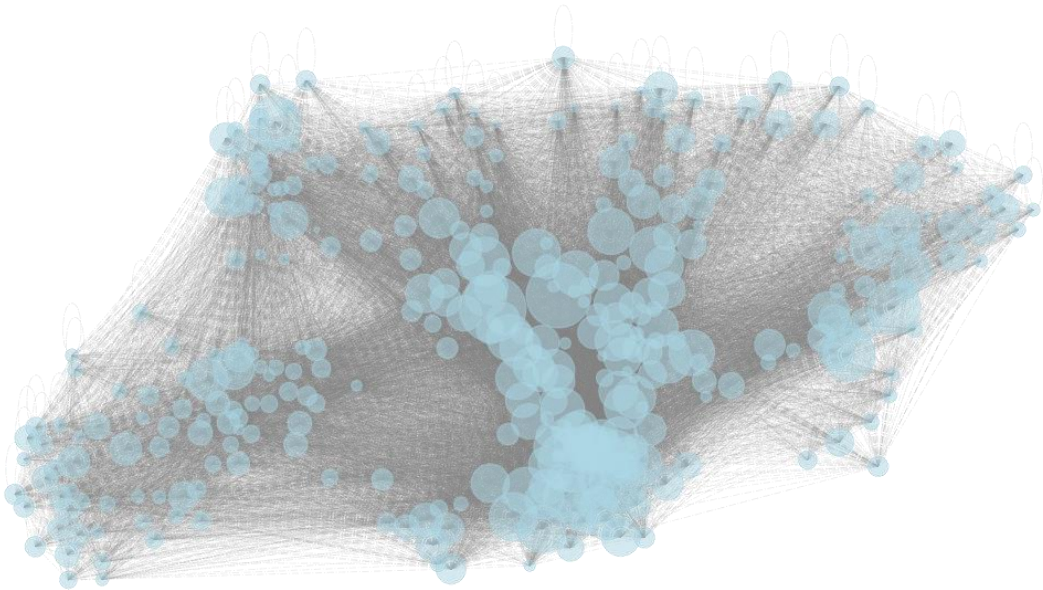
2018

שיווק מתקדם – אנליטיקה של רשתות

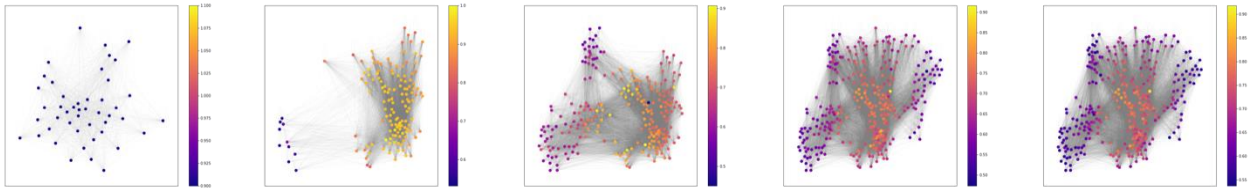
2019



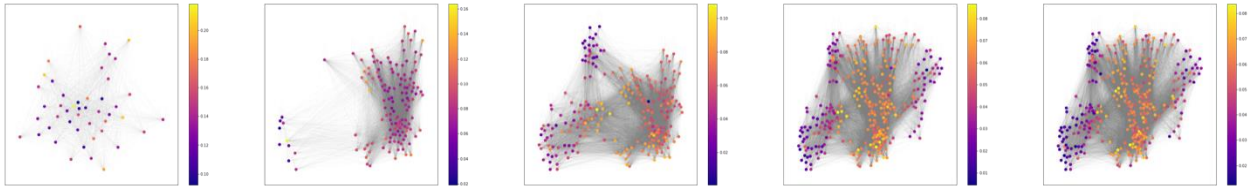
2020



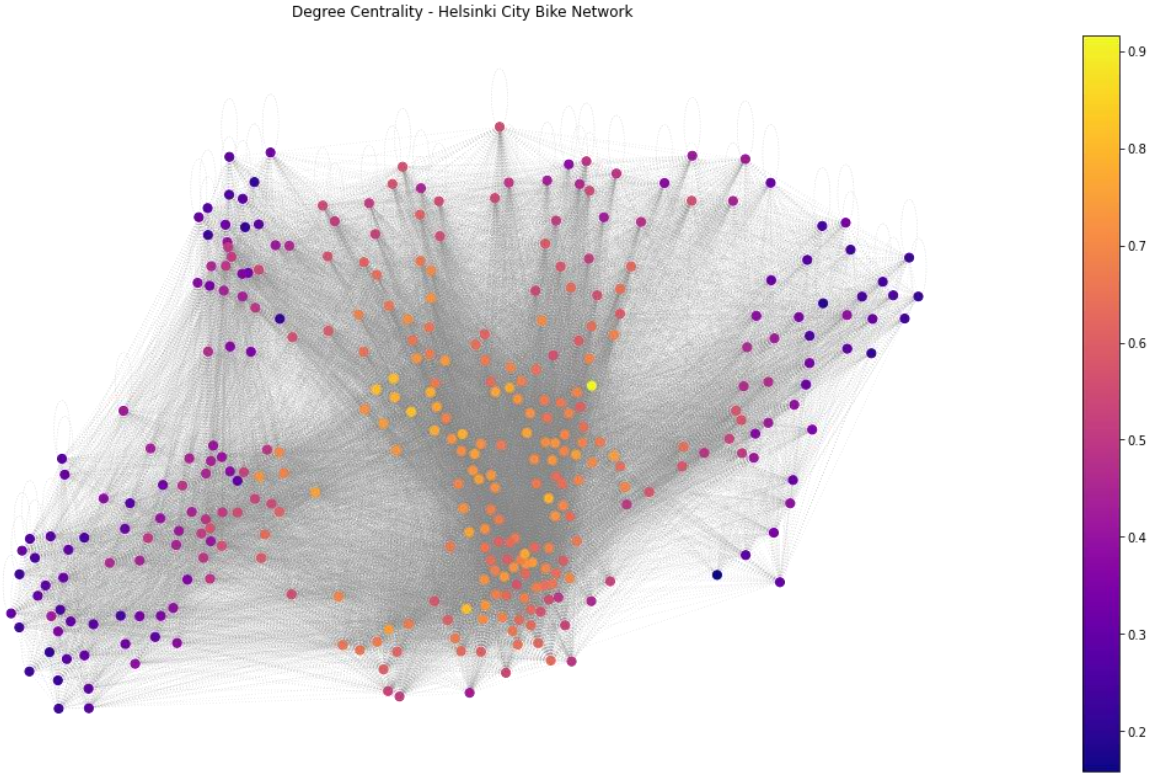
Index 1.1 Evolution of the closeness centrality between 2016 and 2020 – Link back to the text



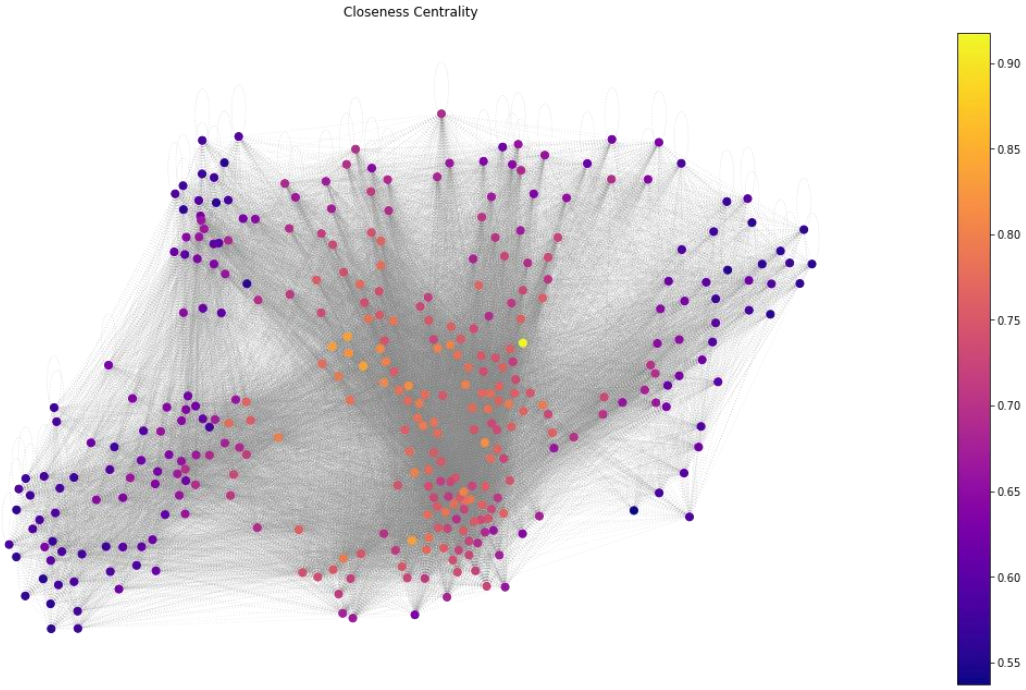
Index 1.2 Evolution of the eigenvector centrality between 2016 and 2020



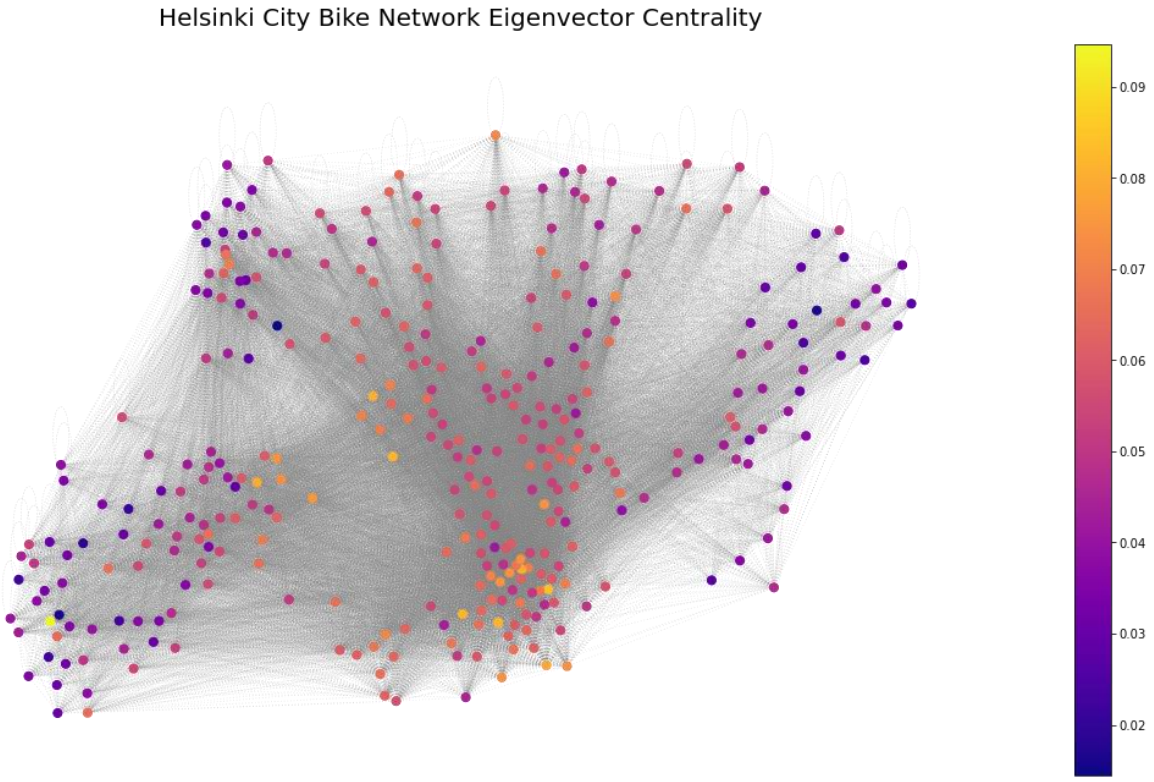
Index2: Degree centrality – Link back to the text



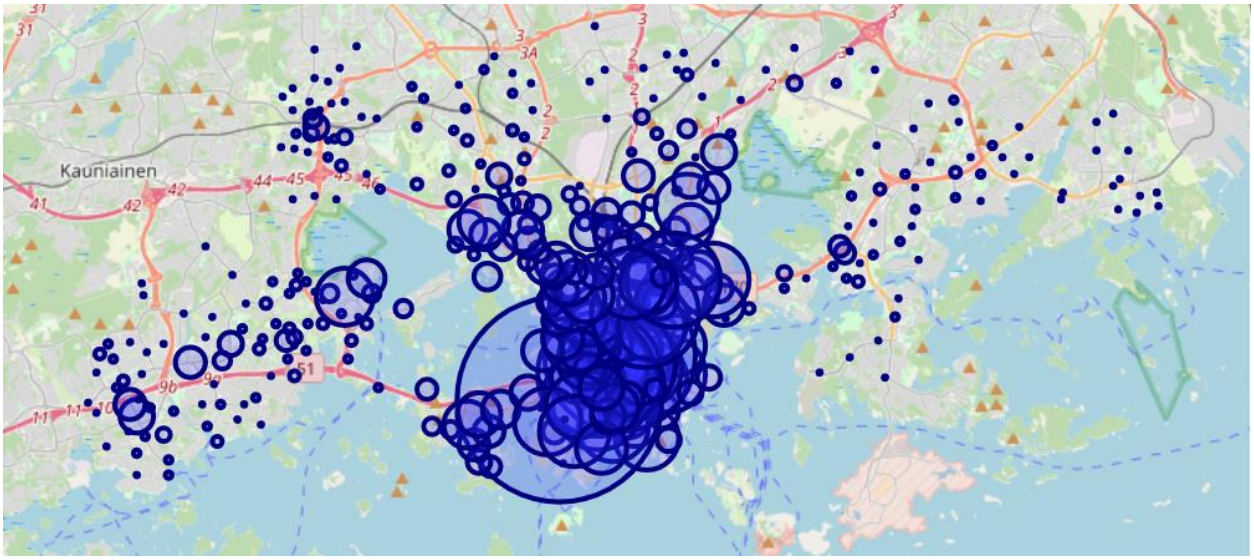
Index3 – Closeness centrality – Link back to the text



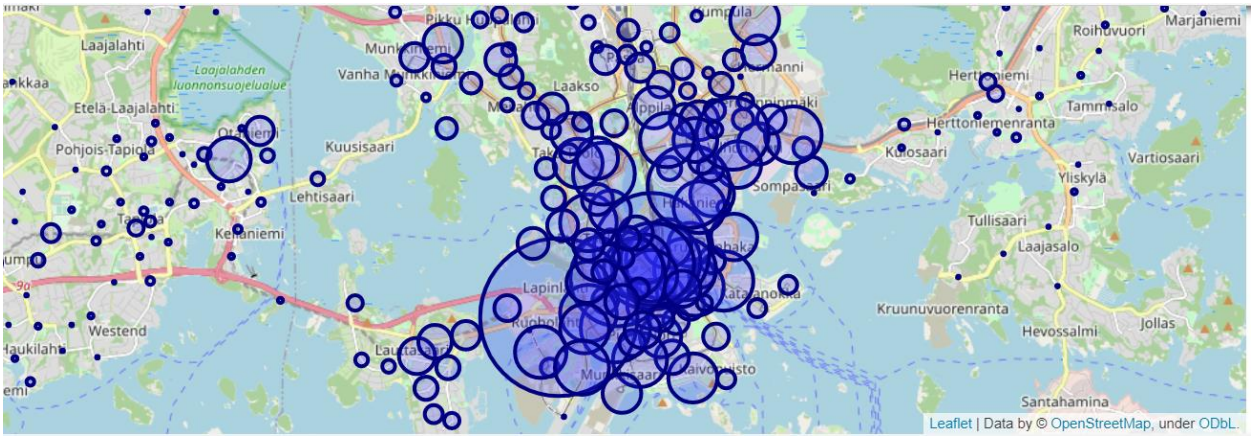
Index4 – EigenVector centrality – Like back to the text



Index#5 Interactive map – Link back to the text



ZOOM IN:



Resources:

https://en.wikipedia.org/wiki/Helsinki_City_Bikes
https://www.hsl.fi/en/citybikes
https://www.hel.fi/hkl/en/by-bike/city-bikes/
https://www.hsl.fi/en/citybikes/helsinki

For Finding Stations we've used Google which led us to the following sites:

https://www.ayy.fi/en/jamerantaival-1
https://www.ayy.fi/en/jamerantaival-3
https://en.wikipedia.org/wiki/Aalto_University
https://www.portofhelsinki.fi/en/cargo-traffic-and-ships/west-harbour

Bonus – why we didn't show use of Gephi in this paper – a Gephi visualization of our smallest network: