Mini Project III - Network Analysis for Helsinki City Bike Dataset

Introduction

Transportation systems are the backbone of cities and regions, providing vital connectivity and shaping the way people move and interact. The inherent complexity of these networks, marked by numerous interconnected elements and diverse stakeholders, necessitates a nuanced understanding to enhance efficiency and effectiveness. Recent strides in computer technology and Geographic Information System (GIS) frameworks have introduced a pragmatic application of graph theoretical concepts, significantly impacting transport planning and spatial mobility studies.

This transformative approach has not only advanced our comprehension of transportation networks but also presented opportunities to optimize their functionality. Traditionally, transportation networks incorporated scheduled services or fixed physical infrastructure, such as roads and railroads. However, the emergence of car and bike-sharing services introduces a new dimension characterized by their ubiquity and capacity for self-organization. Unlike rigid bus and train schedules, bike-sharing operates on-demand, offering enhanced spatial flexibility.

The dynamic nature of bike-sharing networks, exemplified by the ever-changing edges influenced by user activities rather than predetermined routes, highlights a unique aspect of urban mobility. Users, as they traverse the city, organically contribute to the network's structure. This adaptability to user needs, as usage patterns gradually shape the network, creates a symbiotic relationship between the system's evolution and user behaviors.

This project aims to leverage the rich dataset provided by the Helsinki City Bike dataset on Kaggle to conduct an extensive network analysis. The dataset, comprising trip details, bike station information, and geographical data, serves as a valuable resource for unraveling complex relationships within the bike-sharing system. By employing graph theory and network analysis tools, the project seeks to visualize bike stations, identify user travel patterns, and emphasize key routes and popular stations.

Data Overview

The Helsinki City Bike dataset, sourced from Kaggle, represents a comprehensive repository of information crucial for understanding and optimizing the city's bike-sharing system. This dataset encompasses diverse aspects, including trip details, bike station characteristics, and geographic data. The dataset comprises a mix of data types, including datetime, numeric, and object (string) data, ensuring a diverse set of information for analysis. The dataset contains more than 10 million trips made by citizens of Helsinki between 2016-2020. The dataset named database.csv file is downloaded and saved as a bike_data for data processing.

The information of data are:

```
Initial dataset info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 76937 entries, 0 to 76936
Data columns (total 14 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   departure               76937 non-null  object
 1   return                  76937 non-null  object
 2   departure_id            76937 non-null  int64
 3   departure_name          76937 non-null  object
 4   return_id               76937 non-null  float64
 5   return_name             76937 non-null  object
 6   distance (m)            76937 non-null  float64
 7   duration (sec.)         76937 non-null  float64
 8   avg_speed (km/h)        76918 non-null  float64
 9   departure_latitude      76936 non-null  float64
 10  departure_longitude     76936 non-null  float64
 11  return_latitude         76936 non-null  float64
 12  return_longitude        76936 non-null  float64
 13  Air temperature (degC)  76936 non-null  float64
dtypes: float64(9), int64(1), object(4)
memory usage: 8.2+ MB
None
```

Data Processing

In the process of exploring the Helsinki City Bike dataset, first I checked for duplicate rows to ensure data integrity and eliminate any redundancy. The following steps were taken:

Identification of Duplicate Rows:
A boolean series, duplicate_rows, was created to identify duplicate rows within the dataset.

Displaying Duplicate Rows:
Rows identified as duplicates were extracted using the boolean series, and the duplicate data was displayed to provide transparency on the duplicated entries.

Counting Duplicate Rows:
The number of duplicate rows was determined to quantify the extent of redundancy in the dataset. This count is valuable for assessing the impact of duplicate entries on the overall dataset.

Dropping Duplicate Rows:
To ensure data accuracy and streamline further analysis, duplicate rows were removed from the dataset using the drop_duplicates() method.

Updated Dataset:
The final dataset, denoted by bike_data, is now free from duplicate entries, setting the stage for meaningful and accurate network analysis.

Further, after a series of data cleaning, column renaming was implemented. The data type conversion was done for departure and return column representing timestamps, were converted to datetime objects using the pd.to_datetime method. Several column names were refined for clarity and consistency:

- 'distance (m)' was renamed to 'distance'
- 'duration (sec.)' was renamed to 'duration'
- 'avg_speed (km/h)' was renamed to 'average_speed'
- 'Air temperature (degC)' was renamed to 'temperature'

The dataset was filtered to include only records with specific criteria: Distance ranging from 50 meters to 10,000 meters; Duration ranging from 120 seconds to 18,000 seconds; Temperature ranging from -20°C to 50°C.

The dataset, denoted by bike_data, now reflects the refined column names and has undergone range filtering to include high-quality data for analysis.

By cleaning and preparing the dataset for later project phases, this data processing step enhances the overall quality of the data and guarantees that the analysis's conclusions are supported by a trustworthy and clean dataset. Removing duplicate rows improves the findings' robustness within the Helsinki City Bike dataset and is in line with best standards in data preprocessing.

Data Analysis

A distribution analysis was carried out for four important parameters—distance, duration, average speed, and temperature—in an effort to identify the important factors in the Helsinki City Bike dataset.

- Distribution of Distance
- Distribution of Duration
- Distribution of Average Speed
- Distribution of Temperature

Distribution of Distance:

The histogram reveals that the distribution of distances covered by bike ride indicating that

shorter trips are more frequent than longer ones. The majority of bike rides appear to cover distances in the lower range, with a gradual decline as the distance increases.

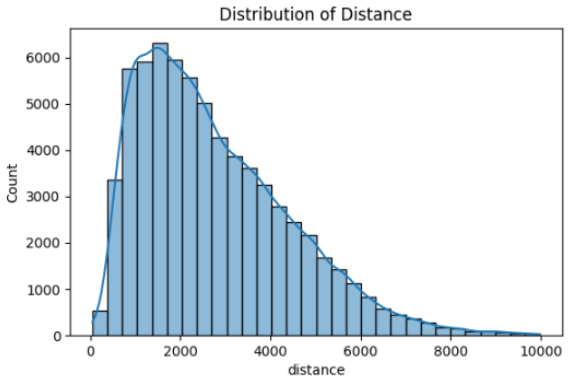

Figure: Distribution of distance(m)

It displays the distribution of the distance variable from the bike_data dataset.

Distribution of Duration:

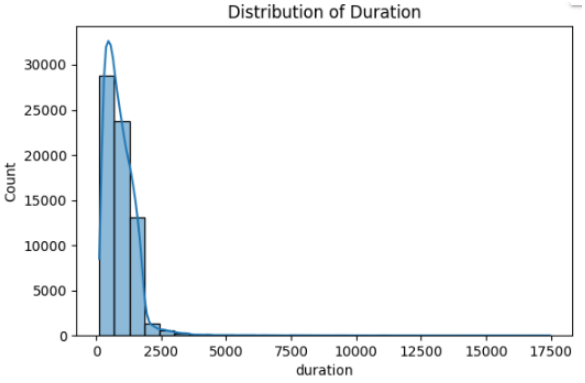

Figure: Distribution of duration(sec)

The histogram for ride durations is similar to the distance distribution. Most bike rides tend to have shorter durations, with a gradual decline in frequency for longer durations. This pattern aligns with the expectation that shorter distances are associated with shorter durations.

Distribution of Average Speed:

The histogram for average speed is suggesting that a significant number of bike rides maintain a moderate speed.

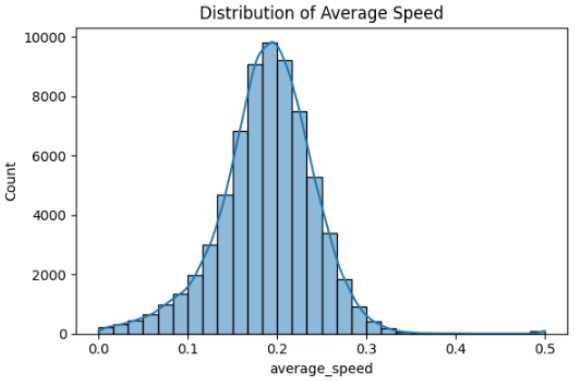

Figure: Distribution of average speed(km/h)

This distribution indicates a balance between slower-paced and faster-paced rides, with a general trend toward the center of the speed spectrum.

Distribution of Temperature:

The temperature distribution appears to follow a roughly symmetric pattern, centered around a specific temperature range.

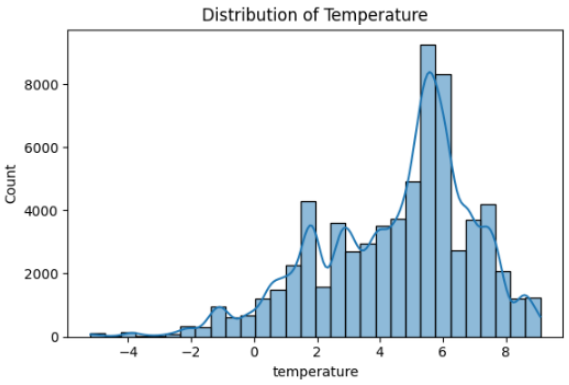

Figure: Distribution of temperature(degC)

This indicates that bike rides occur under a variety of temperature conditions, with a concentration in the middle of the observed temperature spectrum.

In general, these visual aids offer an extensive synopsis of the distribution properties of significant variables found in the Helsinki City Bike dataset. Gaining an understanding of these distributions is essential to identifying trends, anomalies,

and possible correlations that will enhance the complexity and knowledge of the network analysis of the bike-sharing program.

Bike Usage Behavior

With an emphasis on temporal variables like the days of the week and hours of the day, I hope to offer insights that might guide strategic choices about urban mobility and the optimization of bike-sharing systems.

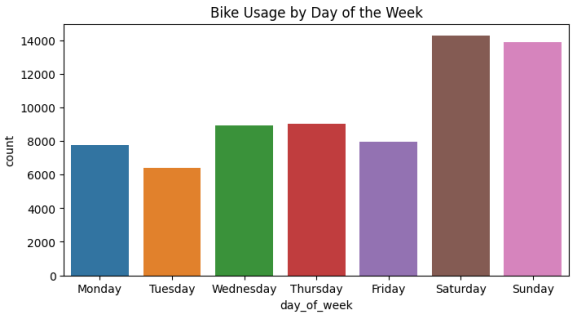

Figure: Bike usage by day of the week

A count plot was created in order to comprehend the weekly fluctuations in bike usage. Plotting the frequency of bike usage by day of the week offers valuable insights into possible associations with specific days of the week, weekends, or special events.

The mostly busy days is saturday followed by sunday. Similarly, tuesday is less consumed day in a week.

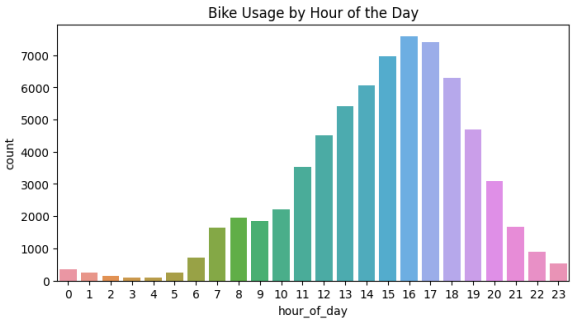

Figure: Bike usage by hour of the day

To explore the hourly patterns of bike usage, a count plot was created. This visualization highlights the distribution of bike trips throughout the day, revealing peak hours and helping to identify periods of high and low demand.

Correlation Analysis

Understanding the connections between various variables in a dataset is mostly dependent on correlation analysis. I examined the correlation matrix of the dataset's important variables from Helsinki City Bike in this part. In order to provide insights into how these variables interact and influence one another, it is intended to find any possible patterns of association or reliance among them.

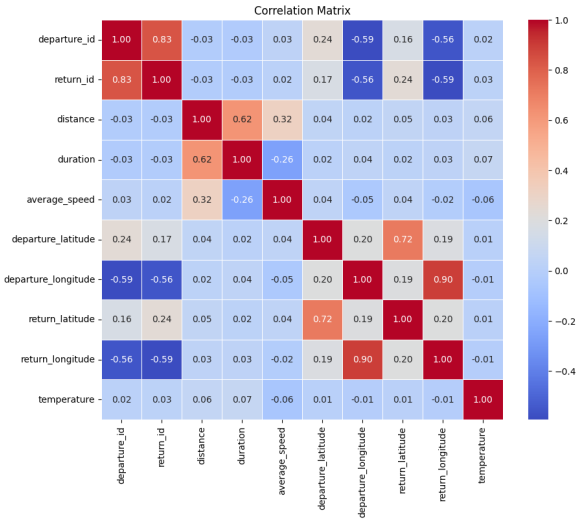

Figure: Correlation Matrix

The correlation matrix is shown visually using a heatmap. The Seaborn library is used to create the heatmap, and for clarity, correlation coefficients have been annotated. The correlation's strength and direction are shown by the heatmap's color intensity, where a cool-to-warm color spectrum denotes negative to positive correlations.

Network Analysis

To understand user travel patterns inside the Helsinki City Bike system, one must grasp

the dynamics of the bike station network. In this section, I use the networkx and matplotlib packages to offer a visual depiction of the bike station network, illuminating the trip patterns and the connectivity between the departure and return stations.

Helsinki City Bike's bike station network visualization provides an overview of the system's connectivity and trip trends. This basic study provides information for future research on station utilization, user behavior, and possible improvements to maximize the effectiveness of the bike-sharing system.

Bike Station Network and User Travel Patterns

Using the NetworkX module, the visualize_bike_network Python program creates a directed graph by processing the bike data. With edge properties like distance and time, each bike ride is represented as an edge connecting the departure and return stations.

Figure: Bike Station Network and User Travel Pattern

Using the kamada kawai layout algorithm, the produced graph is shown with the ideal node spacing. Nodes are represented as circles, and the size of the circles reflects the degree (or total number of connections) of the node. To highlight connections, edges are drawn in a lighter hue, and node labels are added for station identification.

The graphic shows hubs with high user traffic and sheds light on how bike stations are connected to one another. Through the directed edges, travel patterns are revealed, revealing popular bike station pairings as well as frequently traveled routes by users.

Additional Network Statistics for the Helsinki City Bike Network

By calculating and examining more network information, I get deeper into the features of the Helsinki City Bike network in this part. These statistics improve my knowledge of user movement patterns and system performance by offering insightful information about the structural characteristics, connectivity, and centrality of the bike station network.

These statistics include node and edge counts, average degree, network density, triadic closure, and centrality measures such as degree, betweenness, and eigenvector centrality.

```
Number of nodes: 334
Number of edges: 12989
Average degree: 77.77844311377245

Node with Maximum Degree: ('Kalasatama (M)', 173)
Node with Minimum Degree: ('Sateentie', 6)

Network density: 0.23356889823955693
Triadic closure: 0.5951919415269332
```

The bike station network consists of a total of 334 nodes and 12989 directed edges, representing user travel paths between stations.

The network's average degree of connections per node, is 77.7784. This

measure provides information about the network's overall connection of bike stations.

The station with the highest degree is 'Kalasatama (M)' with 173 connections, while 'Sateentie' has the lowest degree with 6 connections.

The network density is 0.23, suggesting that about 23% of possible connections between stations are realized in the network. The triadic closure is 0.59, signifying a relatively high likelihood of forming closed loops (triangles) in the network, highlighting interconnected user travel patterns.

The top 5 nodes by degree are given below and these stations exhibit the highest connectivity in the network, suggesting popular departure and return points.

```
Top 5 nodes by degree:

('Kalasatama (M)', 173)
('Ympyrätalo', 166)
('Itämerentori', 157)
('Pasilan asema', 157)
('Messeniuksenkatu', 156)
```

Betweenness centrality identifies nodes that act as crucial bridges between other nodes in the network. These stations play pivotal roles in facilitating diverse travel routes.

Eigenvector centrality assesses the influence of nodes based on their connections to other well-connected nodes. These stations wield influence due to their connections to other highly connected stations.

```
Top 5 nodes by betweenness centrality:

Node: Haukilahdenkatu, Betweenness Centrality: 0.07792269356771589
Node: Kalasatama (M), Betweenness Centrality: 0.02421085444827892
Node: Kulosaari (M), Betweenness Centrality: 0.015520461623667048
Node: Laajalahden aukio, Betweenness Centrality: 0.0144161833793202
Node: Aalto-yliopisto (M), Korkeakouluaukio, Betweenness Centrality: 0.014168468620836584

Top 5 nodes by eigenvector centrality :

Node: Ympyrätalo, Eigenvector Centrality: 0.10950098757074271
Node: Brahen kenttä, Eigenvector Centrality: 0.10446011186004893
Node: Ooppera, Eigenvector Centrality: 0.10379538492967177
Node: Messeniuksenkatu, Eigenvector Centrality: 0.10374069709324363
Node: Töölönlahdenkatu, Eigenvector Centrality: 0.10367415780669001
```

Important details about the architecture and centrality of the Helsinki City Bike network are revealed by the supplementary network statistics. Further optimizations, such station location, route design, and system resilience enhancements, can be built upon the foundation established by the identified central nodes and important network data.

Feature Engineering and Predictive Modeling for Bike Trip Duration

This section describes how the Helsinki City Bike dataset's feature engineering and predictive modeling techniques were used to estimate bike trip lengths. The objective is to improve our knowledge of the variables affecting travel times and to create a predictive model that can accurately estimate travel times from a subset of attributes.

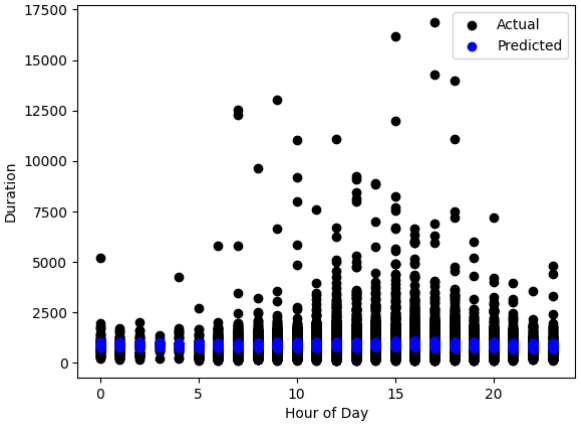

Figure: Feature Engineering

In order to improve the dataset for modeling, we added two more features: "hour_of_day" and "day_of_week." These attributes, which represent the day of the week and hour of the day, were extracted from the 'departure' datetime column.

The features "day_of_week," "hour_of_day," and "temperature" were chosen for the predictive model. It is hypothesised that these characteristics affect how long bike trips last. One environmental component that could affect user behavior is "temperature."

An 80-20 split ratio is used to divide the dataset into training and testing sets. Next, a linear regression model is trained on the training data, initialized, and used to the test data to provide predictions.

The Mean Squared Error (MSE), a statistic that measures the average squared difference between expected and actual values, is used to assess the effectiveness of the model. The calculated MSE for the model is 613463.21, which represents the average squared difference between predicted and actual trip durations in the test set.

A quantitative indicator of the model's correctness, the MSE shows higher performance with lower values. Furthermore, a scatter plot that contrasts the real and estimated travel times for different times of the day provides a visual evaluation of the model's performance.

Visual Analysis of Bike Trip Durations and Popular Stations

I plotted the average journey duration per hour to provide insight into the fluctuation in bike ride durations throughout the day. I can see the trend and distinguish between peak and off-peak hours as well as the fluctuations in the average duration in the line plot below.

The graph reveals distinct patterns in average trip durations throughout the day. Peaks and troughs indicate periods of high and low demand, influencing the average duration.

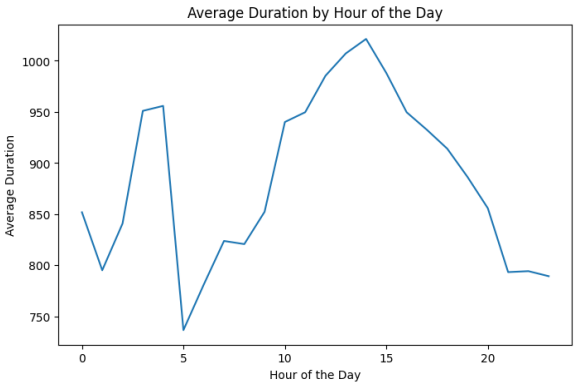

Figure: Average duration by hour of the day

Visual analysis of average durations by the hour provides a nuanced understanding of temporal patterns in bike trip durations. This insight is valuable for optimizing system operations during peak and off-peak hours.

It is essential to comprehend bike station popularity in order to optimize the system. We use a countplot to display the demand for the top 8 most popular stations, which we determined by calculating station frequencies.

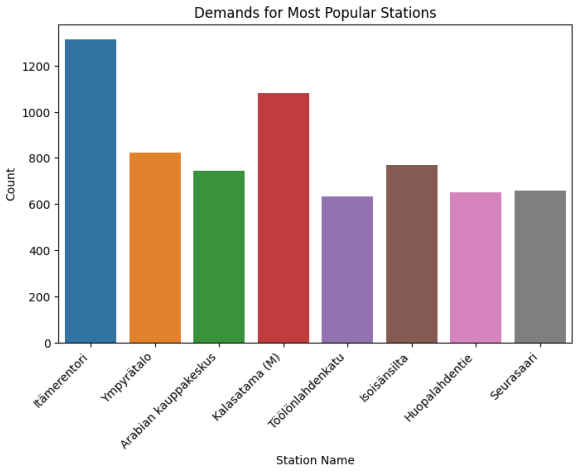

Figure: Demands for most popular stations

The countplot illustrates the distribution of bike departures from the top 8 most popular stations. Identifying stations with high

demand assists in resource allocation and strategic planning.

Itamerentori stations is demamnded highly followed by Kalasatama (M) and Ympyratalo respectively. The less demanded stations is Toolonlahdenkatu.

Displaying Geospatial Visualization

Understanding the distribution and popularity of bike departure stations within the Helsinki City Bike system is made possible by the strong tool that is geographic visualization. The frequency of bike departures from each station is graphically represented in this section's dynamic bubble plot, which offers important insights into the geographical patterns of bike utilization around the city.

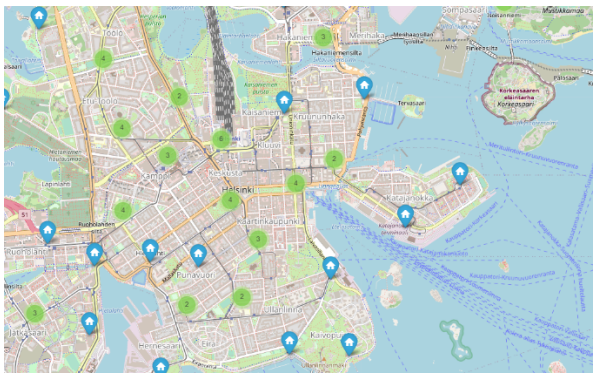

Figure: Geospatial Visualization of Bike Departure Stations

Because station coordinates are taken from an external dataset, stations are placed geographically precisely. To concentrate on relevant city areas, stations with latitude values over 60.254011 are taken off. Helsinki is the center of the base map, which is created using OpenStreetMap tiles. The map incorporates a FastMarkerCluster to improve display efficiency and handle massive datasets quickly.

Every bubble denotes a bike departure station, and its size corresponds to the frequency of departures from that station. High-demand stations receive a visual cue for the altered frequency shown in the color intensity of each bubble. Finally, geographical representation is shown, enabling interactive investigation of Helsinki's bike departure trends.

The spatial distribution of bike departure stations is clearly communicated through the generated bubble map. The size and intensity of color of each bubble provides instantaneous information about the demand and popularity of specific stations.

Degree Centrality Visualization of Helsinki City Bike Network

One important metric used in network analysis to quantify the significance of individual nodes within a network is degree centrality. In this section, I tried to show the degree centrality for the Helsinki City Bike network visually. Each bike station has a degree centrality value assigned to it, which provides information about the stations' relative relevance and impact on the network's overall connectedness.

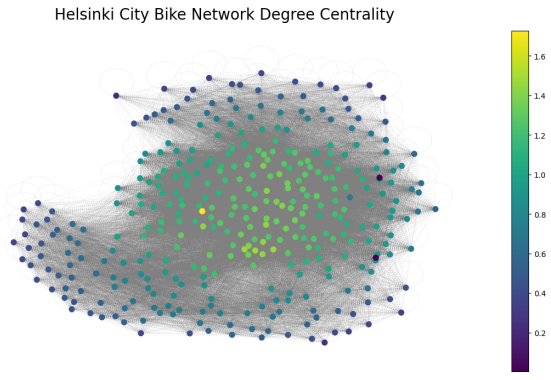

Figure: Helsinki City Bike Network Degree Centrality

Higher degree centrality stations are represented by nodes with more intense colors, signifying their significance in promoting network connectivity. The

geographical layout makes it possible to quickly identify the major hubs and stations that are essential to the operation of the bike-sharing system. The degree centrality visualization offers a useful overview of the significance of each station in the Helsinki City Bike network. This data is essential for strategic planning, system optimization, and the identification of stations that serve as critical nodes in the network as a whole.

Graph Analysis - Degree Distribution and Station Centrality

The Helsinki City Bike network is thoroughly examined in this part, with an emphasis on degree distribution and the identification of stations with high degree centrality. Degree centrality is an essential metric that measures each node's significance within a network, giving information on the prominence and connectivity of bike stations.

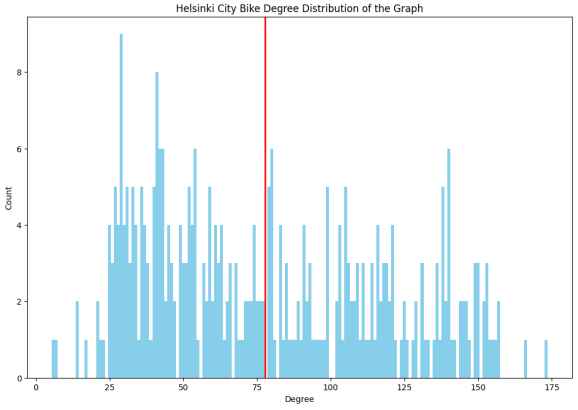

Figure: Helsinki City Bike Degree Distribution of the Graph

A node's average degree in the graph is determined to be 77.78. This measure provides a broad picture of network connectivity by showing the average number of connections made by each station.

Key hubs in the bike-sharing network are located by identifying stations with the highest degree of centrality.

Using their corresponding degree centrality values, the top five stations are as follows:

Kalasatama (M): 0.52 centrality
Ympyrätalo: 0.50 centrality
Itämerentori: 0.47 centrality
Pasilan asema: 0.47 centrality
Messeniuksenkatu: 0.47 centrality

The average degree is shown by the red line, which serves as a point of reference for comprehending the distribution.

Betweenness Centrality Visualization of Helsinki City Bike Network

This section examines the importance of bike stations in the Helsinki City Bike network by delving into the idea of betweenness centrality. The degree to which a station serves as a link between other stations is measured by betweenness centrality, which pinpoints critical locations essential for network connectivity.

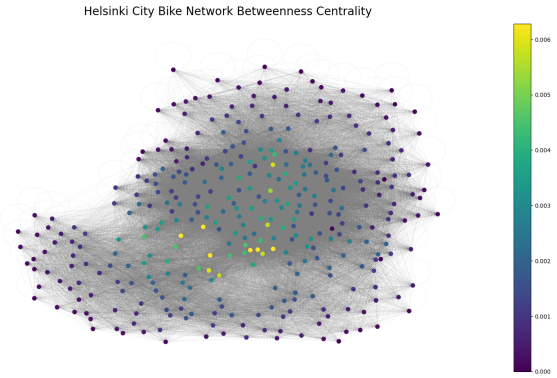

Figure: Helsinki City Bike Network Betweenness Centrality

A measurement of a station's impact on the paths connecting other stations is provided by betweenness centrality, which is computed for every node in the modified network. A color scale is used to depict the betweenness centrality values, highlighting the significance of stations as network links.

To ensure clarity, nodes are scaled uniformly, and the betweenness centrality value is shown by the intensity of the color.

The stations are connected by dot-patterned gray borders that give a visual depiction of network relationships without competing with the centrality visualization.The stations are connected by dot-patterned grey edges that give a visual depiction of network connections without competing with the centrality visualization.

Each bike station's betweenness centrality is highlighted by the color-coded representation, where higher centrality is indicated by more intense colors. Focused investigation of the remaining network is made possible by the removal of the "Haukilahdenkatu" node.

Closeness Centrality Visualization of Helsinki City Bike Network

In network analysis, closeness centrality is a crucial parameter that sheds light on the effectiveness and accessibility of the bike stations in the Helsinki City Bike network. This section shows each station's proximity centrality, highlighting important nodes that provide effective network accessibility.

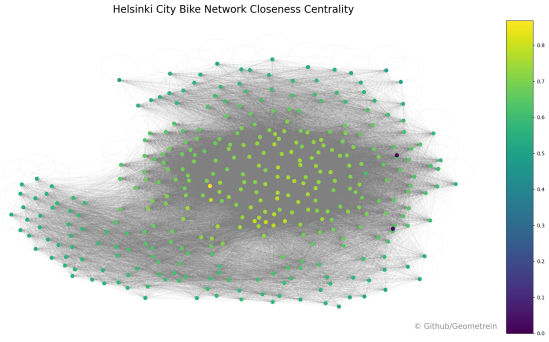

Figure: Helsinki City Bike Network Closeness Centrality

Each node in the graph has its closeness centrality calculated, which measures a station's proximity to every other station in the network. Stations with higher closeness centrality ratings are easier to reach and have shorter mean separations from neighboring stations.

A color scale is used to illustrate the closeness centrality values, highlighting how accessible each bike station is. For clarity, nodes are sized uniformly, and the closeness centrality value is shown by the intensity of the color. The stations are connected by dot-patterned gray borders that give a visual depiction of network relationships without competing with the centrality visualization.

Each bike station's closeness centrality is highlighted by the color-coded representation, where higher centrality is shown by more intense colors. Higher closeness centrality stations make a network easier to access.

Eigenvector Centrality Visualization of Helsinki City Bike Network

A key metric in network research, eigenvector centrality illustrates the significance of bike stations in the Helsinki City Bike network by comparing them to other central stations. The eigenvector centrality of each station is shown in this section, providing information on important nodes that greatly influence the overall structure of the network.

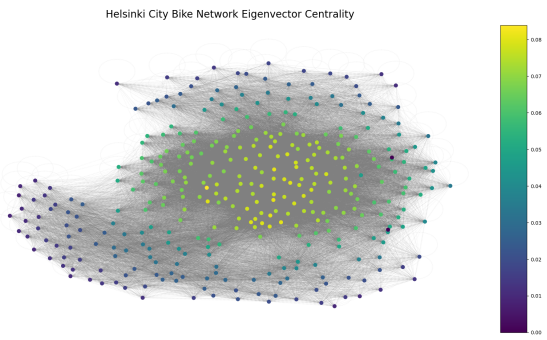

Figure: Helsinki City Bike Network Eigenvector Centrality

For every node in the graph, eigenvector centrality is calculated while taking into account the weighted network depending on the number of bike journeys. Each station's effect is represented by its eigenvector centrality values, where larger values denote a stronger bond with other central stations.

A color scale is used to illustrate the eigenvector centrality values, emphasizing the impact of each bike station. For clarity, nodes are sized uniformly, and the eigenvector centrality value is shown by the intensity of the color. The stations are connected by dot-patterned gray borders that give a visual depiction of network relationships without competing with the centrality visualization

Conclusion

Helsinki City Bike's bike station network visualization provides an overview of the system's connectivity and trip trends. This basic study provides information for future research on station utilization, user behavior, and possible improvements to maximize the effectiveness of the bike-sharing system. These metrics enhance my understanding of the system's complexity, user travel patterns, and the significance of specific bike stations within the network.

Based on particular attributes, feature engineering and predictive modeling have been used to estimate bike journey lengths. After being trained on the constructed features, the linear regression model offers insights into the association between temperature and time-related parameters and bike trip durations. To improve the model's predicted accuracy, more explorations and optimizations can be made, such as adding more features or using sophisticated modeling methods.

The evaluation of predictive modeling offers important information about how well the model performs in estimating the length of bike trips. Further refining and optimization efforts are guided by the MSE, which is a quantitative measure of predicted accuracy.

A detailed knowledge of the temporal patterns in bike trip durations can be obtained by visual examination of average durations broken down by the hour. Optimizing system operations during peak and off-peak hours is made possible by this knowledge. Determining the demand for popular stations also helps with decisions about infrastructure upgrades, bike redistribution, and station upkeep.

The degree centrality visualization offers a useful overview of the significance of each station in the Helsinki City Bike network. This data is essential for strategic planning, system optimization, and the identification of stations that serve as critical nodes in the network as a whole.

The Helsinki City Bike network's centrality depiction sheds important light on each station's function as a connection. Finding stations with strong betweenness centrality is essential to comprehending how they affect network connection as a whole. The closeness centrality map sheds important light on how easily accessible each Helsinki City Bike station is. Determining whether stations have a high closeness centrality is essential to comprehending how they contribute to effective connectivity. The visualization of eigenvector centrality offers significant insights into the impact of individual stations in the Helsinki City Bike network. To comprehend how stations shape the network structure, it is essential to identify those with high eigenvector centrality.