Introduction to Datascience and Machine Learning –

Team Assignment Project Report

Team Rakete:

Philipp Böhmer, 7380629

Yung-Hsiang Chen, 7330269

Wei Dai, 7349502

Kai Krause, 7385524

Xi Wang, 7370695

Jingwen Yi, 7368892

Instructor: Prof. Dr. Wolfgang Ketter

TA: Karsten Schroer, Philipp K. Peter

Bachelor of Science WI / IS

Faculty of Management, Economics, and Social Sciences

Department of Information Systems for Sustainable Society University of Cologne

Summer Term 2022

Contents

Contents I
1 Background ..... - 1 -
2 Data Cleaning ..... - 1 -
3 Customer Behavior Analytics ..... - 2 -
3.1 Hourly Demand (day and time of departure) ..... - 2 -
3.2 Geographical Demand (location of departure/destination) ..... - 3 -
3.3 Daily Ride-Durations (trip qualities such as duration, net distance2 or net speed) - 4 -
4 Predictive Analytics ..... - 4 -
4.1 Availability ..... - 4 -
4.2 - 4 -

1 Background

The emissions of greenhouse gases that are caused by transportation make up the second greatest portion of the EU's overall emissions. Because of this, it has been known for a long time that in order to fulfill our objectives for decarbonization, we would need to adjust how we think about mobility. Vehicles powered by internal combustion (IC) engines continue to be the primary mode of urban transportation in the modern day. This mobility arrangement is associated with four well-known detrimental effects on society. In this project, we study how fleet operators may make use of real-time data streams, which are becoming more commonplace, in order to monitor and improve their operations, therefore boosting their quality of service, as well as their profitability.

The information that we have access to is comprised of the datasets of bike sharing rentals made via Bay Wheels in San Francisco, in 2018, which were gathered from their respective website: https://www.lyft.com/bikes/bay-wheels/system-data

2 Data Cleaning

We began aggregating the time data once we had finished scanning the data in order to acquire a better grasp of the information that was available to us. We then proceeded to get rid of any data that had a duration of minus one and any rides that had a length of less than three minutes and began and terminated at the same station. This was done because we regarded such rides as not being "genuine" rides but rather accidental starts (e.g. a person changed his mind because of bad weather or other reasons). Because of the policy of Bay Wheels, which we discovered on their website and which states that a bike will be reported as stolen if the ride-time surpasses 24 hours and a fee would be levied, we terminated any trip that over that point (filtered as outliers).

After removing duplicates of some IDs, we learned that this was because certain radio stations had changed their names, which led us to our discovery that certain IDs had been used more than once. We were able to resolve this problem by changing the station name to the one that had been used the most recently for the relevant Station ID.

We are able to simply generate demand patterns and do trend analysis by adding a category value for each day of the week as well as each hour of the day.

We merged the coordinates of all stations through months that we found in a datasets from on the Bay Wheel website with our dataset from 2018, so that we could better understand

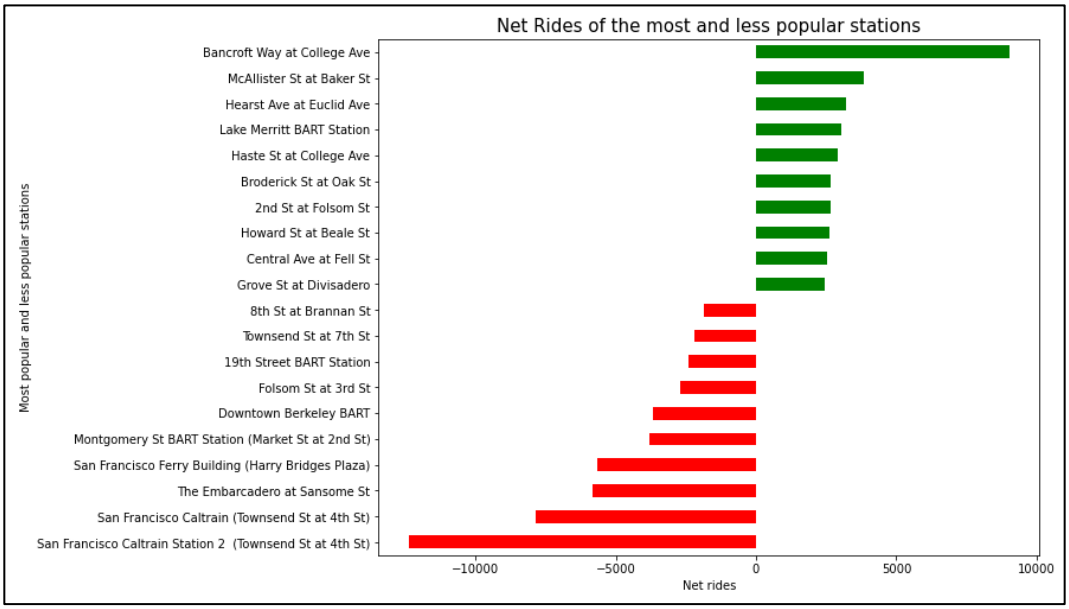
and analyze the geographical demand, as well as learn more about the cycling habits of customers, in order to optimize the bike delivery management strategy. This allowed us to learn more about customers' cycling habits. This also enables us to better graphically depict the data, for example, by highlighting the stations on the map that are used the most and the least often so that trends may be seen.

We also counted the number of unique beginning and ending stations in order to assess the overall use of the stations, as well as to identify the stations that received the greatest traffic in order to increase the number of bikes or investigate other possible station sites. After that, the data set that had been cleaned up was exported so that everyone could use the same information.

3 Customer Behavior Analytics

Before we starting with the Demand Analytics we filtered the most and the less popular stations. We calculate the deviation of the amount of the rides between departures and destinations. By doing so, we are able to narrow down the stations to those that the consumers ride to the most / the less.

Figure 1 Net Rides of the most and less popular stations



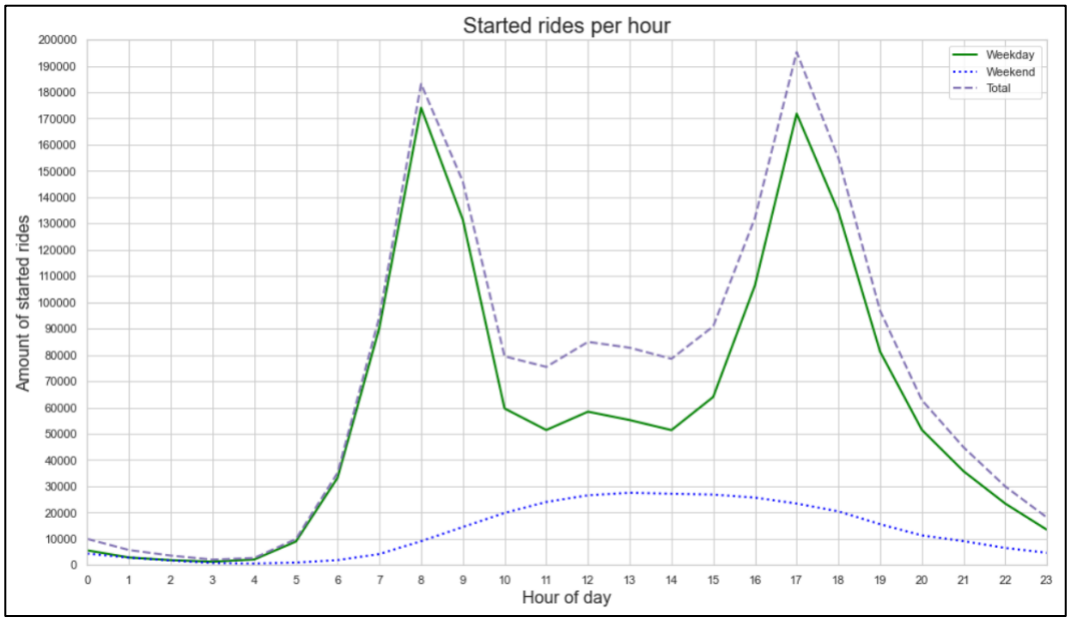
3.1 Hourly Demand *(day and time of departure)*

It is essential for a company that offers bike sharing to have a better awareness of the demand in order to be able to give its clients with an available bike at any time. This may be accomplished by shifting resources and ensuring that the service (an available bike) can be

provided. The purpose of this section of our study was to investigate how the varying seasons and weather conditions play a role in this, as well as how the demand is affected by these factors.

We also want to acquire a more in-depth understanding of the data, particularly with regard to the demand and how it evolves over the course of each hour over the course of the day. As a result, we began by organizing the data according to weekdays and weekends, in order to determine how the demand for rides is affected or altered by the presence or absence of business days. Next, we plotted the number of rides that were started during each hour of the day and compared the results in a single plot.

Figure 2 Started rides per hour

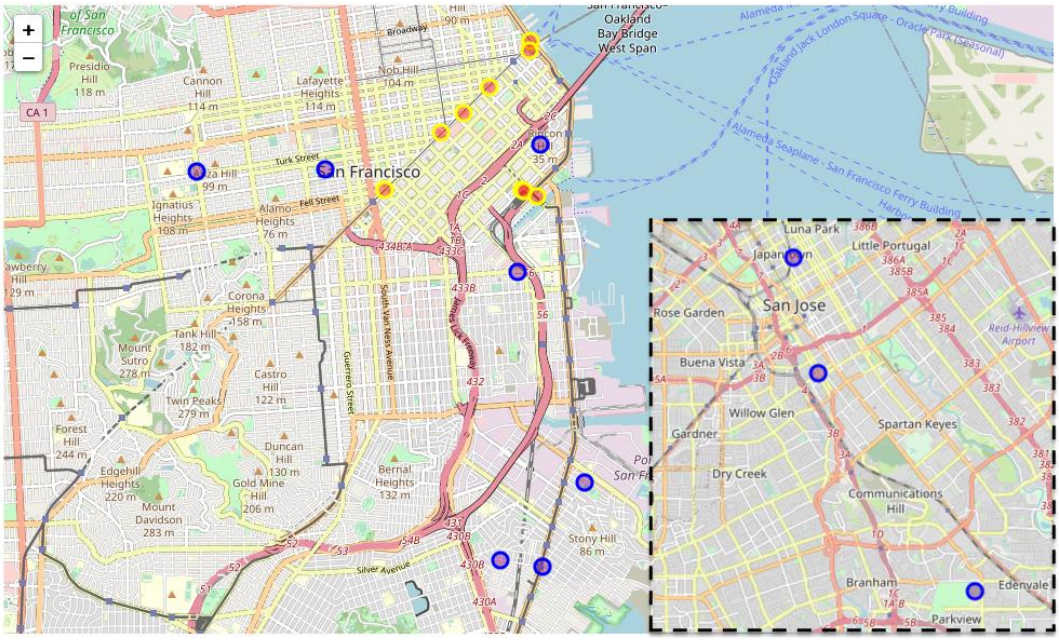


3.2 Geographical Demand *(location of departure/destination)*

We created a map of San Francisco utilizing the available coordinates of the stations that were supplied by the data exploration that came before it so that these results could be presented in a more visual format.

This map provides a very clear visual representation of the pattern that we discovered in the data: The stations that are most frequently used are the ones that are located close to the heart of San Francisco (on the main streets to the peers), whereas the stations that are located in the suburbs, a great distance from the city (or even outside of the City, in San Jose), are the ones that are used the least. This might be because of greater distances to go or because there is less traffic in the suburbs, both of which eliminate the benefits that consumers look for when renting a bike in a more congested area: to avoid being stuck in traffic and arriving at places more quickly, all while engaging in greater physical activity and being less harmful to the environment.

Figure 3 Net Rides of the most and less popular Stations



3.3 Daily Ride-Durations *(trip qualities such as duration, net distance2 or net speed)*

4 Predictive Analytics

4.1 Availability

4.2 The expected number of trips in the next hour

4.3 Model Evaluation

4.4 Outlook

4.5