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

1 Background

The greenhouse gas emissions 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, thus 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 was 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 the 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 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.

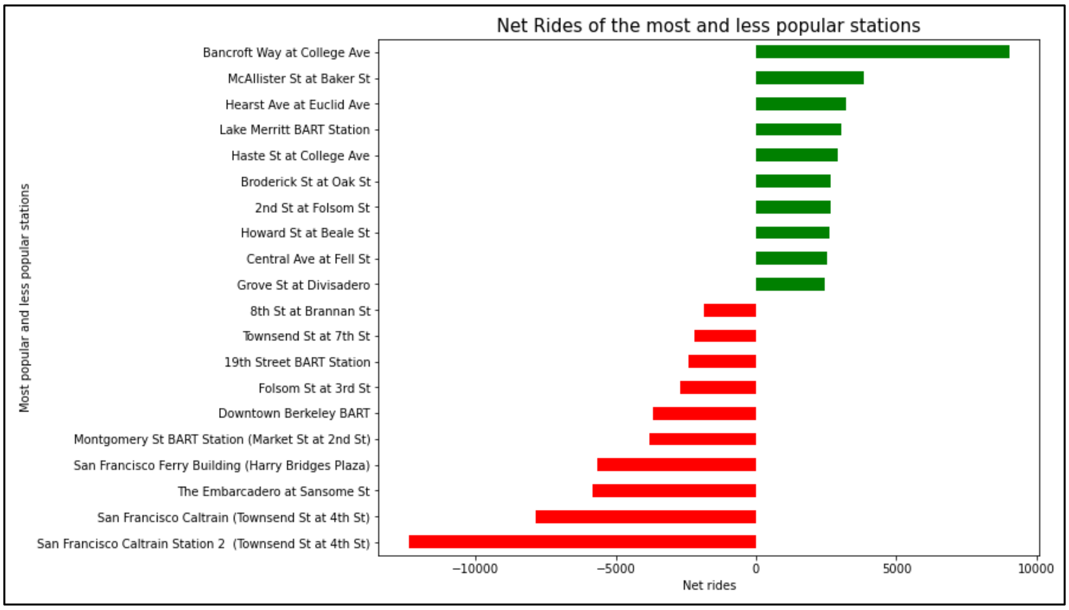
We merged the coordinates of all stations through the months that we found in a dataset on the Bay Wheel website with our dataset from 2018, so that we could better understand and analyze the geographical demand as well as learn more about the cycling habits of customers. This allowed us to learn more 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 started with the demand analytics, we filtered the most and the least 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 least.

Figure 1 Net Rides of the most and least popular stations

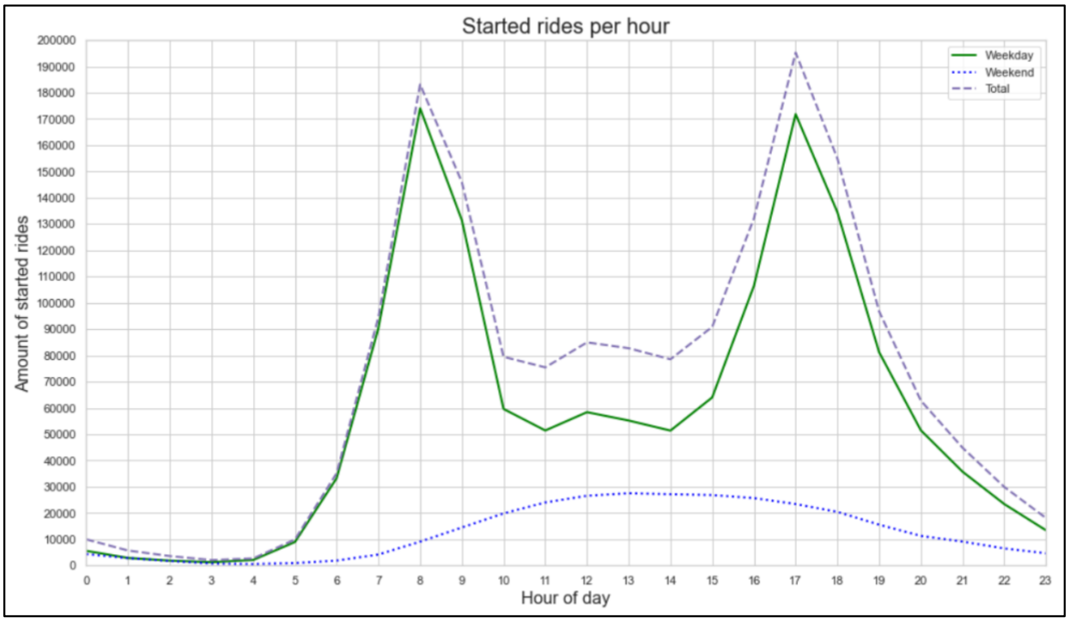


3.1 Hourly Demand *(day and time of departure)*

It is essential for a company that offers bike sharing to have a better awareness of the demand in order to be able to provide 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 influence 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

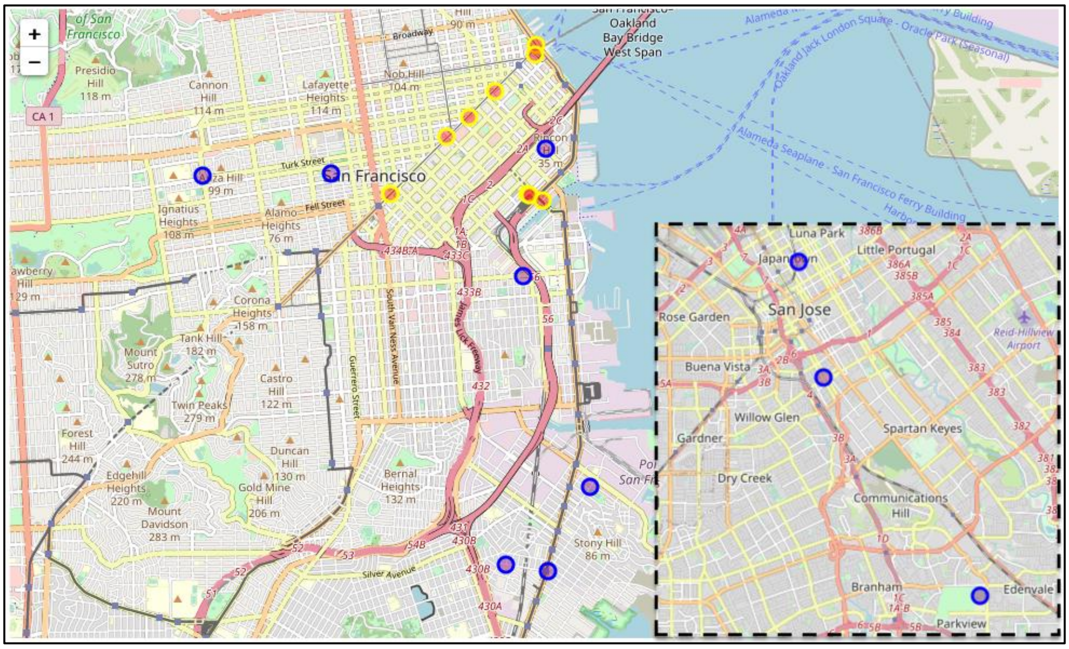


3.2 Geographical Demand *(location of departure/destination)*

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 arrive at places more quickly, all while engaging in greater physical activity and being less harmful to the environment.

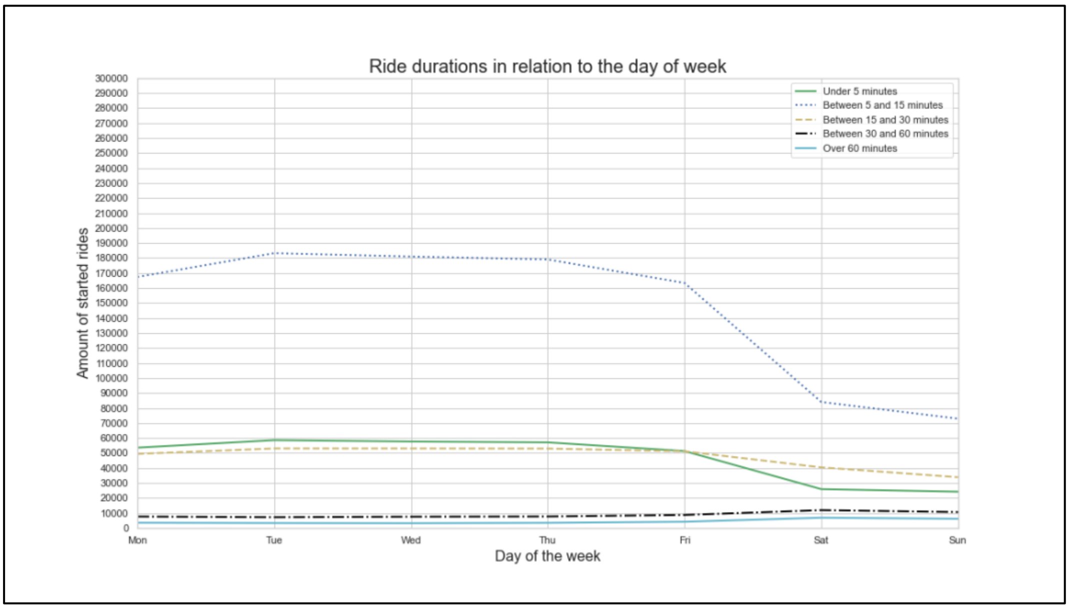
Figure 3 Map of the most and least popular stations



3.3 Daily Ride-Durations

Figure 4 shows a clear change in ride duration patterns, namely a drop-off in some categories activity coinciding with the general drop-off in started rides during the weekend. Interestingly however, there is an increase in started rides that lasted longer than 30 minutes. An explanation could be, that the average weekend customer is more likely to rent one of the bikes not simply for commuting, but rather longer, more casual rides.

Figure 4 Ride durations in relation to the day of week



3.4 Temperature and precipitation

Since a bike offers no protection from the weather, we wanted to see if and how much of an impact factors like temperature and precipitation have on the willingness to start a bike ride.

Figure 5 Monthly mean of temparature

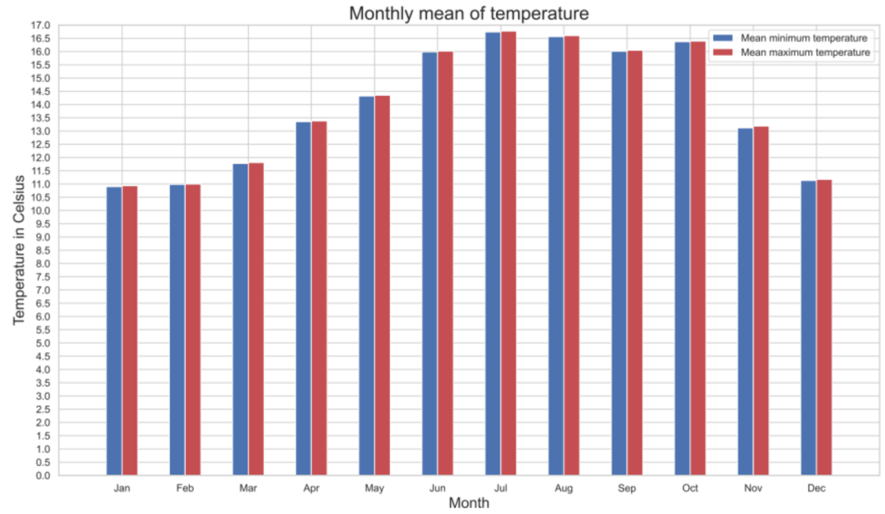
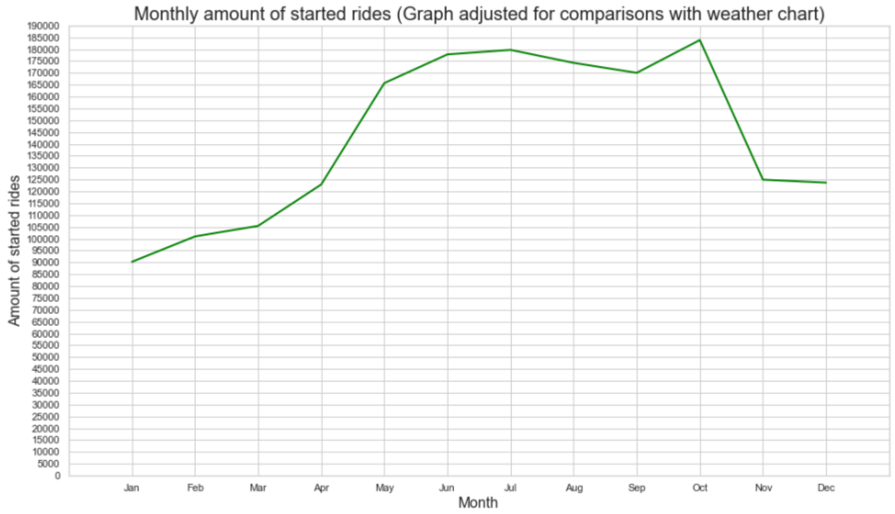


Figure 6 Monthly amount of started rides



First, we created a graph that depicts the mean maximum and minimum monthly temperature. Secondly, we plotted the monthly amount of started rides in order to be able to visually compare the two.

The graphs underline the intuitive hypothesis that warmer temperatures lead to a higher willingness of the customers to rent a bike. Even the dip in temperatures in September compared to the months before and after aligns with the dip in rides around the same time.

Next, we wanted to look at precipitation. Since we can‘t simply compare the raw amount of rides undertaken in hours with/without precipitation, we needed to calculate the mean of rides undertaken with either of those conditions.

3.5 Unsupervised Clustering

We carried out the following cluster analysis in order to more clearly convey the connections that exist between factors such as temperature and the requirements placed on the bike. We are able to ascertain that the correct number of clusters is three by using the elbow approach.

The following are some of the insights we gained:

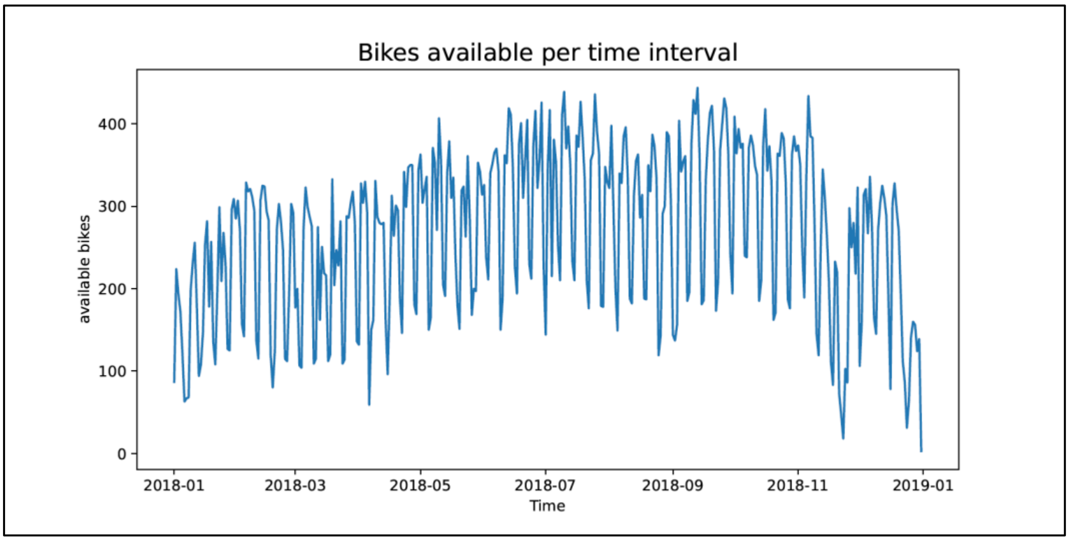
1) The busiest times of day for registered users on weekdays are between the hours of 7-8 am and 5-6 pm. In addition, there was a discernible decrease in the number of people using shared bikes very early and very late in the day. To put it another way, a significant portion of the demand for users is caused by the automobiles used for commuting in the morning and evening.
2) There are more bike rentals available when the weather is nice than when it is terrible. When it is bright or partly overcast, or when there is a light mist in the air, the number of people renting bicycles is substantially greater than when it is raining or snowing.
3) Because there is not a strong association between temperature and rental volume, we can say that the environment in San Francisco is pleasant, the temperature is appropriate throughout the year, and the likelihood of experiencing severe weather is lower.

4 Predictive Analytics

4.1 Availability

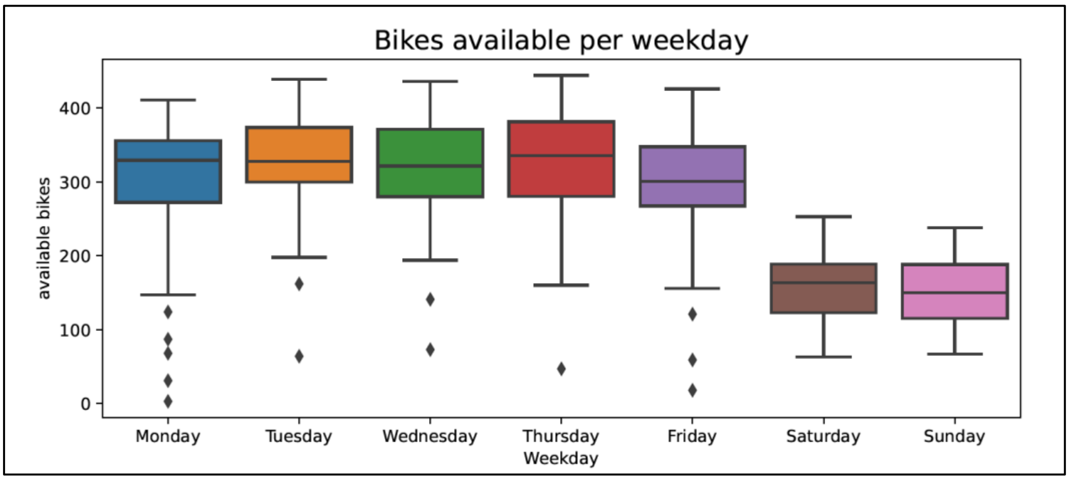
For this part, we decided to only focus on the top 10 most popular stations. We started by counting the size of the bike fleet and then proceeded by plotting the availability of bikes over a year, noticing that during the summer the number of bikes usable seems to be higher than during the winter. This leads to the hypothesis that Bay Wheels reacts to a decreasing demand for bike-rides in the winter and early spring months with a decrease in the number of available bikes.

Figure 7 Bikes available per time interval



Plotting the availability over a week showed that most bikes are in use on Saturdays and Sundays, whilst remaining relatively constant over the rest of the week but with high variation over the whole year.

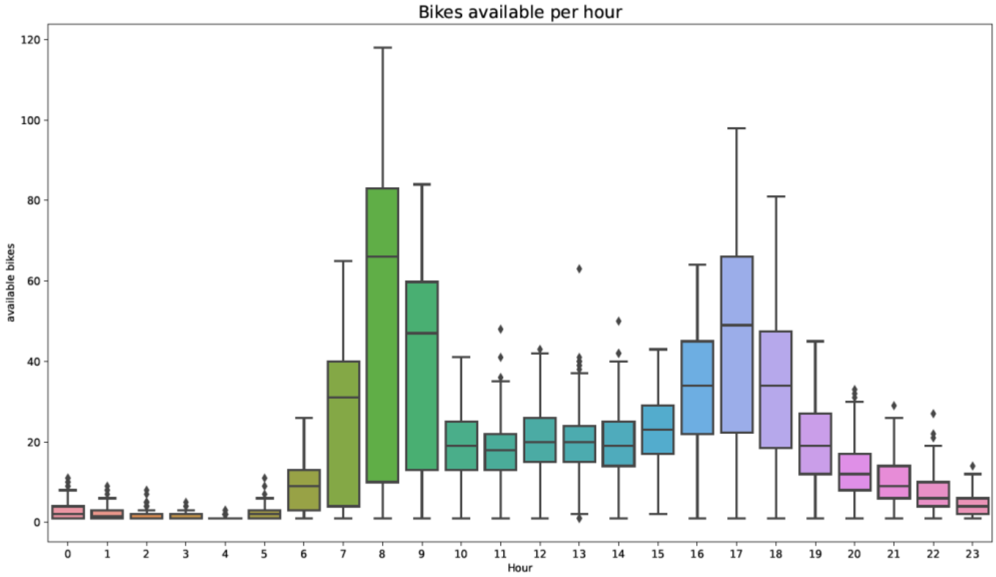
Figure 8 Bikes available per weekday



The plot showing the demand during a day shows that the least number of bikes are in use during the night from 0 o 'clock to 5 o 'clock, while the most bikes are in use during the rush hour in the morning (7-8) and in the evening hours (16-18). It should also be pointed out that the number of bikes in use in the morning is higher than in the evening, leading one to suggest that customers might get to work by bike but get off work via public transit.

This difference may be the result of people being rather exhausted after work, so the idea of taking a train- or bus ride home seems more appealing for some people.

Figure 9 Bikes available per hour



4.2 Unsupervised Clustering

We carried out the following cluster analysis in order to more clearly convey the connections that exist between factors such as temperature and the requirements placed on the bike. We are able to ascertain that the correct number of clusters is three by using the elbow approach.

The following are some of the insights we gained:

4) The busiest times of day for registered users on weekdays are between the hours of 7-8 am and 5-6 pm. In addition, there was a discernible decrease in the number of people using shared bikes very early and very late in the day. To put it another way, a significant portion of the demand for users is caused by the automobiles used for commuting in the morning and evening.

5) There are more bike rentals available when the weather is nice than when it is terrible. When it is bright or partly overcast, or when there is a light mist in the air, the number of people renting bicycles is substantially greater than when it is raining or snowing.

6) Because there is not a strong association between temperature and rental volume, we can say that the environment in San Francisco is pleasant, the temperature is appropriate throughout the year, and the likelihood of experiencing severe weather is lower.

4.3 Regressions and Findings

We chose numerous characteristics that stood out as significant demand-influencing elements, including the following: The highest temperature recorded and the time of day are as follows:

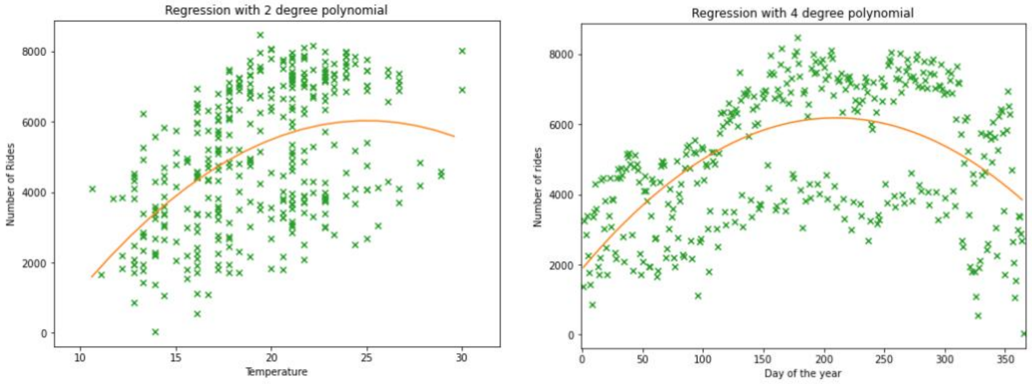
Day of the week, working day, month, week of the calendar, and season: In earlier stages of our research, we saw that the date had an effect on the total number of rides reported; hence, in order to make use of this information, we partitioned the date into numerous classified factors. Nighttime: While we were visualizing the demand over the course of a day, it became evident that throughout the night just a fraction of trips were reported. As a result, we arrived to the conclusion that this would be an essential feature to utilize when estimating future demand.

In order to get our dataset ready for the regressions, we needed to establish a dataframe that had one entry for each of these attributes. This would provide us enough information to work with. By condensing the data such that it just includes the variables that we rated as relevant, we are able to do a more accurate and straightforward regression analysis without needing to aggregate or count the data at a later time.

We began by doing sample regressions using only one variable in order to acquire a better understanding of the effect that variable has on the total number of rides that were registered during that hour. In order to gain an even more comprehensive picture of the situation, we ran some regressions to determine which factors had the most influence on the total number of rides that were logged for a given day.

Although a relationship between the two variables can be seen to exist in the graph of Regression of 2-degree polynomial, this relationship does not follow a linear pattern, therefore the Pearson correlation coefficient does not provide any useful information. It would be more appropriate to do a regression with a broader scope and determine the relevant coefficient of determination.

Figure 10 (left) and 11(right) Regression with 2 and 4 polynomial



4.4 Model Building and Evaluation

Since there are no linear relations between the observed data, we have opted to use the decision tree and the random forest as our regression algorithms.

Decision Tree:

We may use decision trees for both regression and classification, and they are straightforward to read. Additionally, decision trees are quite useful. It is even possible to print it out so that we can examine how each tree-node contributes to the predictive model. The decision tree performed the best out of all the models that we evaluated, with an R^2 value of 0.7302, representing an almost perfect match to the data that was provided.

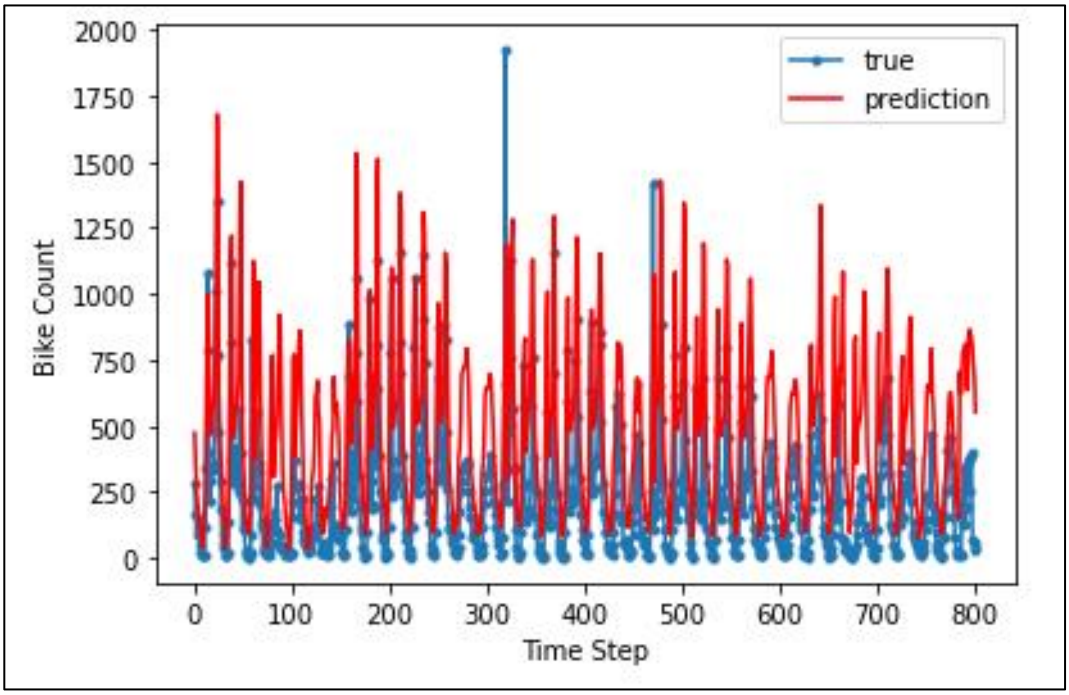
We wanted to know which of our previously selected features had the greatest impact on this, so we visualized the importance of the features. This revealed that we have only have one feature that is significantly more important than the others, and that we even have a few features whose impact is very close to 0. The relevance of the hour (feature 0) is calculated at 0.61724 These are the two features that have the most impact. In addition, you can find below a fragment of the visualization of the decision tree that our model is based upon.

Random Forests:

The random forest algorithm takes the data that is provided and utilizes it to generate a set of decision trees. The method then uses the results of these trees to provide a prediction. With an R^2 value of 0.54 and 100 estimators, the random forest model performed marginally better than the polynomial model when applied to our data.

As a result of the algorithms of regression, we were able to acquire the following prediction:

Figure 12 Bike amount prediction



4.5 Outlook

Our model's capacity to provide more accurate forecasts may be enhanced by using more data gleaned from years gone by, which would be a step in the right direction. If we were able to control the order in which the decision tree takes the input features, then the performance of the model could be improved in another way. This is because certain features have a significantly greater impact than others, and if we took these features into account first, then the performance of the model could improve.

5 Python Code and Team Contributions

Jupyter notebook and executable Python code:
https://github.com/yhsiang25/DSMLRakete

Individual contributions of each team member:

All documents and codes are labeled with names in the repository

Philipp Böhmer, 7380629

Weather data cleaning and merging

Demand patterns analyzing

Yung-Hsiang Chen, 7330269

Data exploration and mining

Stations counting and popularity analyzing

Predictive analytics correcting

Assignment report writing and structuring

Wei Dai, 7349502

Feature engineering

Demand development business goal analyzing

Kai Krause, 7385524

Demand patterns analyzing

Availability analyzing

Xi Wang, 7370695

Unsupervised Clustering analyzing

Jingwen Yi, 7368892

Map of stations plotting

Unsupervised Clustering analyzing