Report: Group Project in Data Science and Machine Learning (Plottwist)

1. Data Collection and Preparation

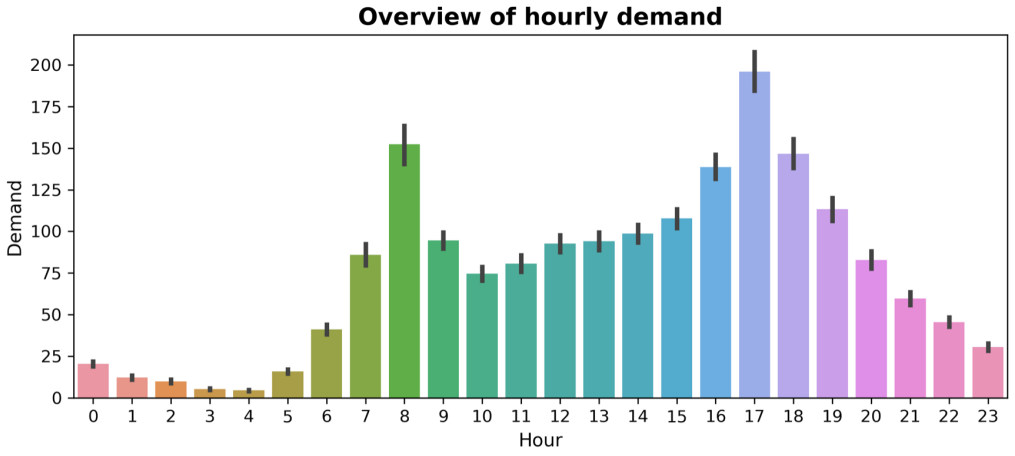
a. GitHub Repository

https://github.com/apfelcake/dsml-plottwist

2. Descriptive analytics

a. Temporal Demand Patterns and Seasonality

When looking at the demand during a day, it becomes obvious that there are great differences. There are two peaks at 8am and 5pm which stand out to the other hours. This could be explained by the usual office hours which go from about 9am to 5pm. The bikes are possibly used to get to work and for the way home after. This may be further supported when looking at the variation between days of the week later.
From midnight to 4am demand remains on a relatively constant dropping, low level before rising to the first peak at 8am. This is followed by a drop and remaining again at a relative constant, though higher level than before. At 4pm, the demand is rising again prior to reaching the all-day-high at 5pm. This is followed by a decline which ends at 5 am.

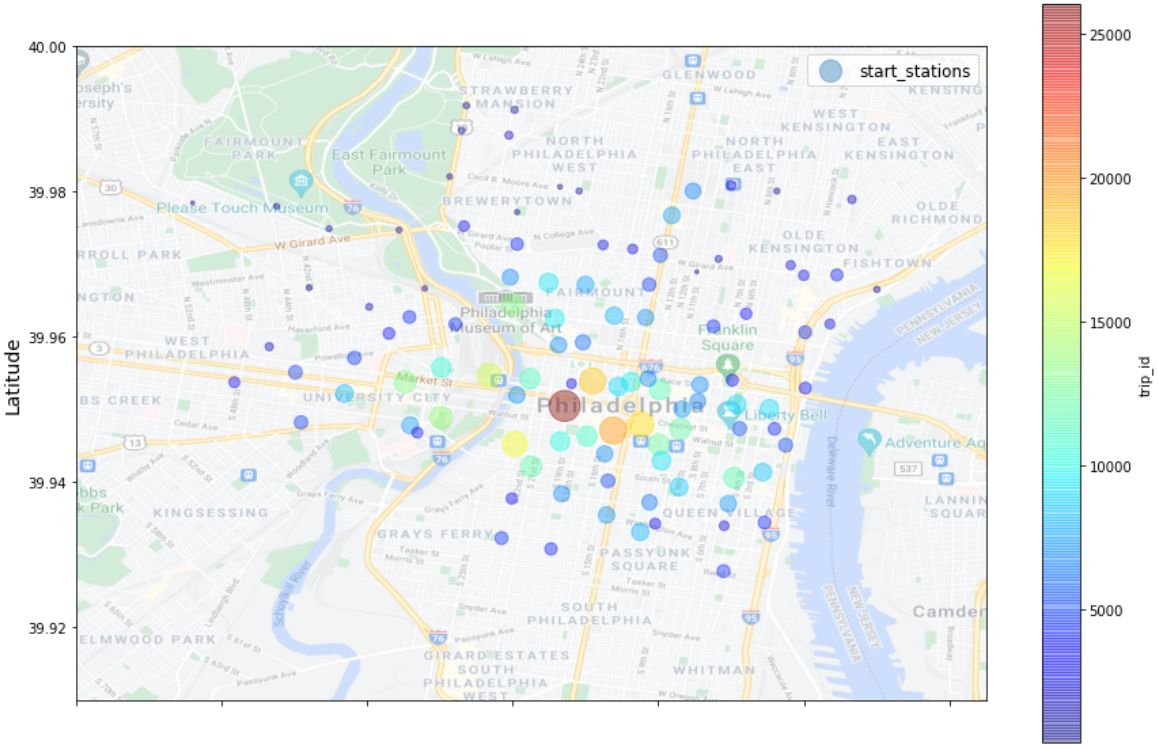


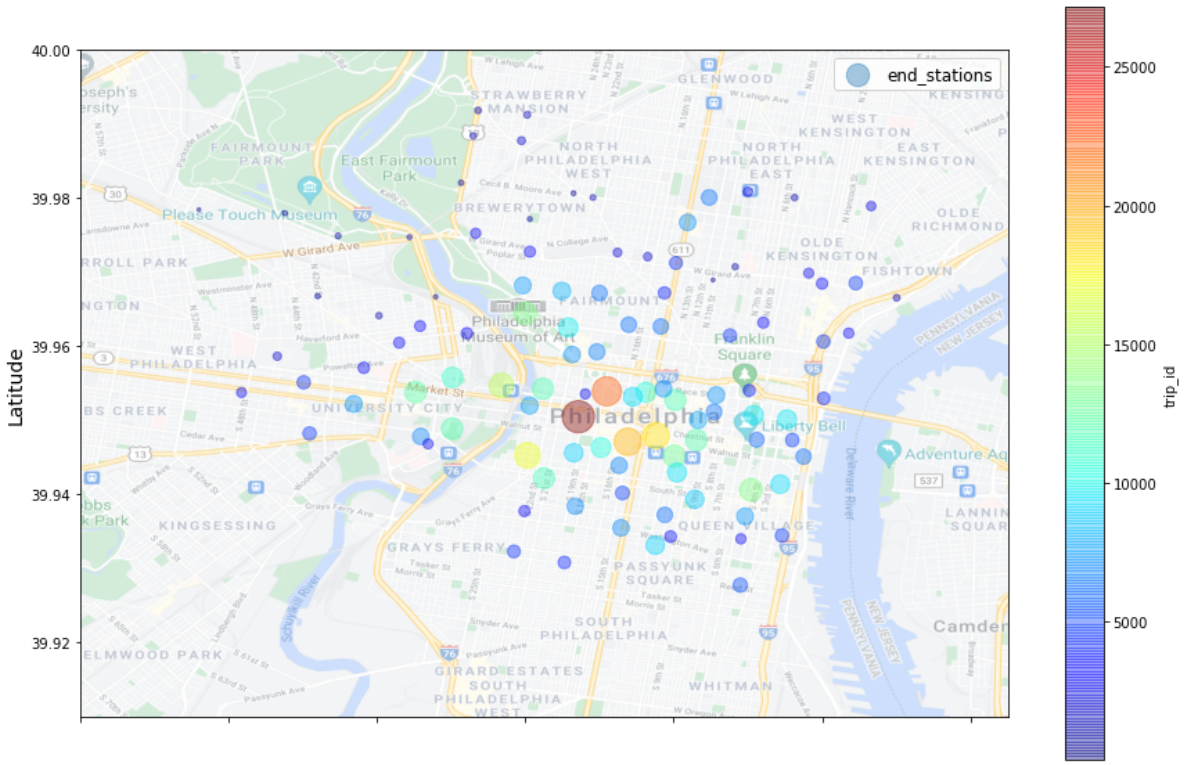
The demand during a week supports the thesis that the bikes are often used for transportation to and from work. From Monday to Tuesday there is a constant demand which does not fluctuate very much showing no outstanding peaks. On Saturday demand is significantly lower than during the working days followed by an even lower demand on Sunday which marks the low of the week. As Saturday and Sunday people do not have to get to work, they do not need the bikes so that demand is the lowest during those days.

Considering the demand during the year, there are some results which could be expected. Demand is the lowest during January and February which are probably the coldest and most snowy months of the year. This makes this time especially unattractive for riding bikes. From March on demand rises steadily to reach its high in September. It has to be noted that the highest demand does not really stand out to the other months as it remains on a relatively constant high level from June to September. After that it is falling a little bit in October and November which can be explained by the temperature getting lower before dropping by half in December probably also due to temperature dropping and the chance of snow. Another effect could be holidays (Christmas Eve, New Years Eve, Boxing day, etc.).

b. Geographical Demand Patterns

The Geographical Demand can be estimated by comparing the most frequented stations in the city of Philadelphia.
We compare the demand of stations where trips were started to the stations where trips have ended. It is important to note that trips can also end where they have started.





As expected the most demand for bike stations is in downtown. The city center is typically the area where the least residents own a car or other way of personal transportation thus making public bike sharing more attractive. The demand in start and end stations is also almost congruent. Meaning rental bikes rarely leave their general area and are mostly used to get around the busy streets of downtown.
In recent years we also see an increase in stations around the south area of Philadelphia (Source: https://www.rideindego.com/stations/) which were not available in 2016. It is to be expected that Ride Indego identified the need of their customers to get to the bay area around Delaware River.

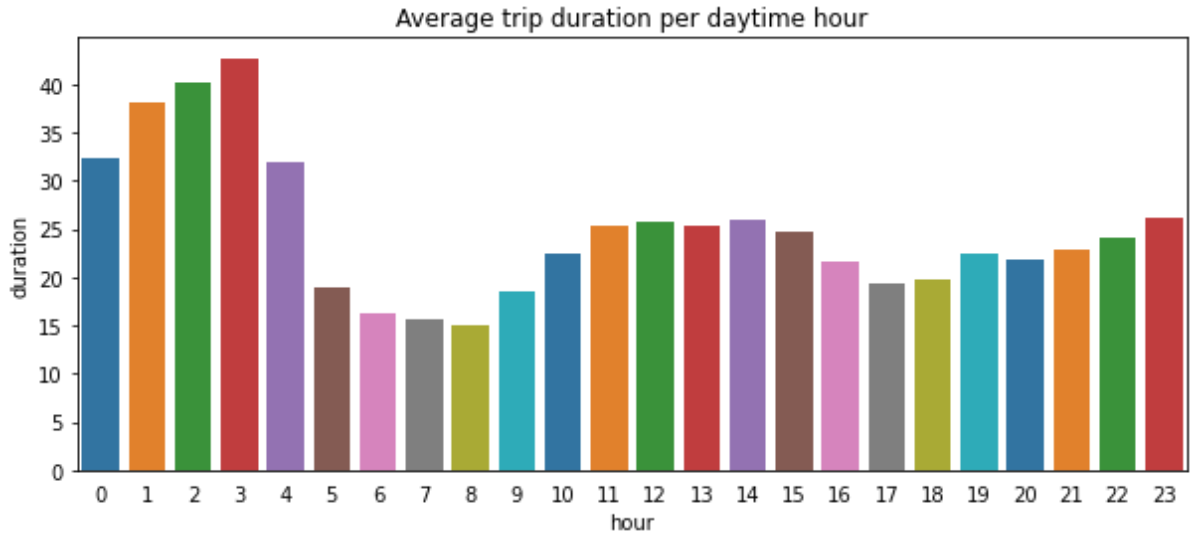
c. Key Performance Indicators (KPIs)

The meaning behind the variable revenue per hour is to calculate the revenue brought in not by sales of passes but by actual usage of the bikes. Since we assume that this has a higher margin than the sales of the passes, it could be a goal for the company to increase this in order to increase their earnings.
When we look at the development of the values during the year, one can see that the values differ greatly, indicating the big fluctuation in revenue that comes on top of pass sales. The peaks could be caused by special events which attract a lot of long usages of bikes, maybe by tourists. The goals for the company should be to smooth this out, making it less dependable on a few days. Also a general increase in these numbers would be desirable. This could be achieved by attracting more people who would ride the bike all day like tourists from outside and not locals who tend to use the bike to get to work. An increase in touristic rides could be achieved e.g. by more marketing at tourist spots and centers as well as offering maps with routes to explore Philadelphia.

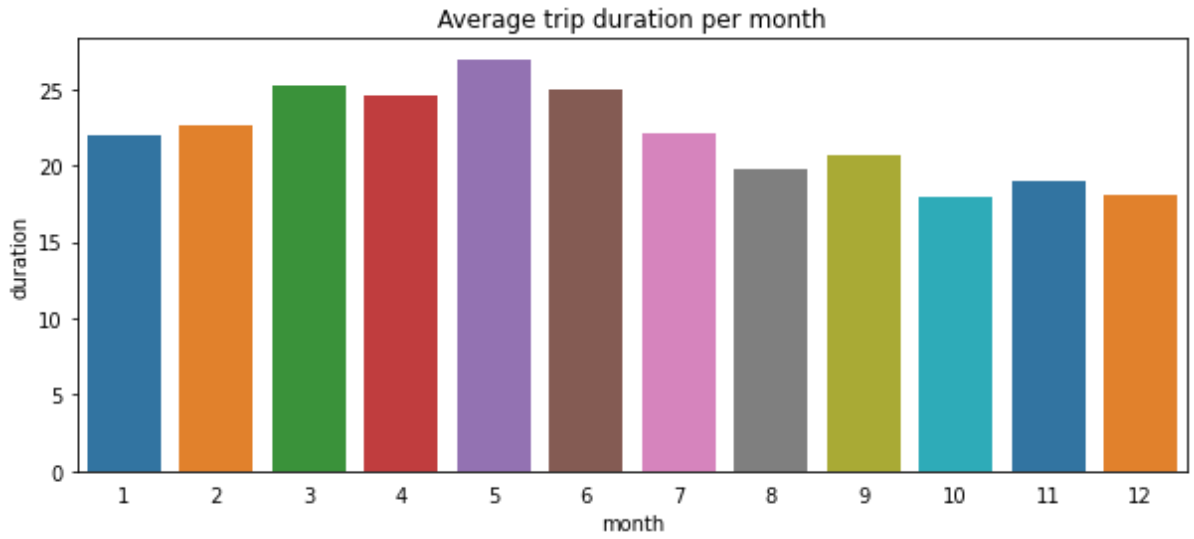
The percentage per passholder type gives information about the customers who use the bikes. Walkups indicate that they are not regular users in contrast to the other types. That is why Walkups are encoded with 0 and the other types with 1 in our code.
When we look at the development during the year, it is striking that passholders are the ones who use the bikes the most. This indicates a strong base of customers bringing in revenue by buying passes and supports our earlier theory that the bikes are commonly used by people every day to get to work. Nevertheless, the low ratio of Walkups leave room for the company to enhance spontaneous rides by people who are only in the city for a couple of days like tourists. Possible actions have already been described earlier.

Trip duration

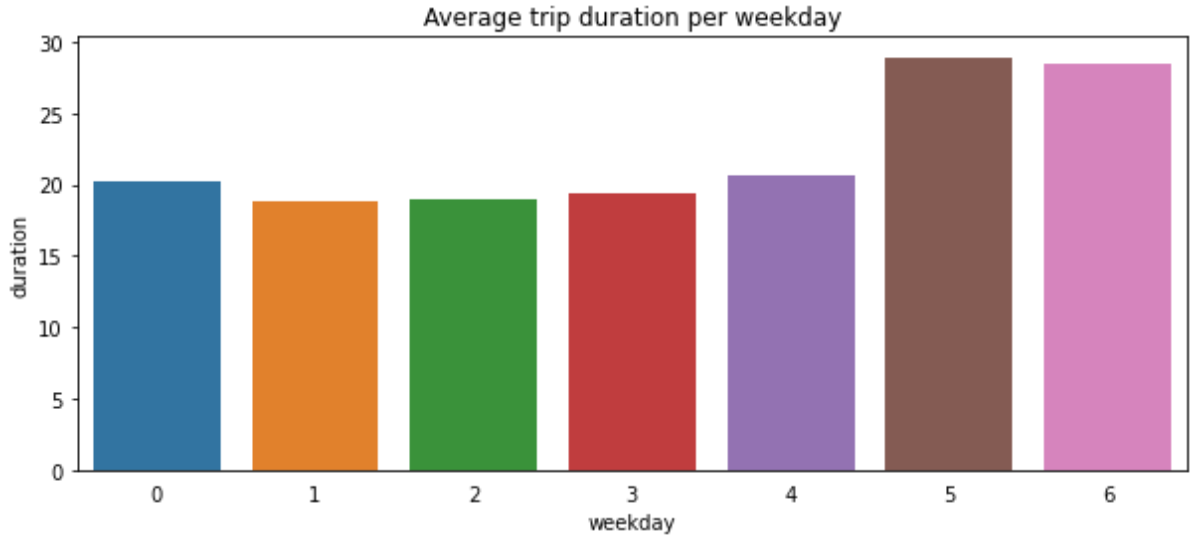
The trip duration is assumed to be an important indicator for usage patterns in bicycle rental. We assume that the trip duration is significantly varying between different daytimes, weekdays seasons. In the following, these three factors are examined concerning its influence on the average bicycle trip duration.



While analyzing the average trip duration in relation to the daytime, it can be seen that bicycle rental customers tend to perform longer trips during the night and during the middle of the day. While the average trip duration is the lowest at 8 am, it reaches its maximum of 43 minutes at 3 am.

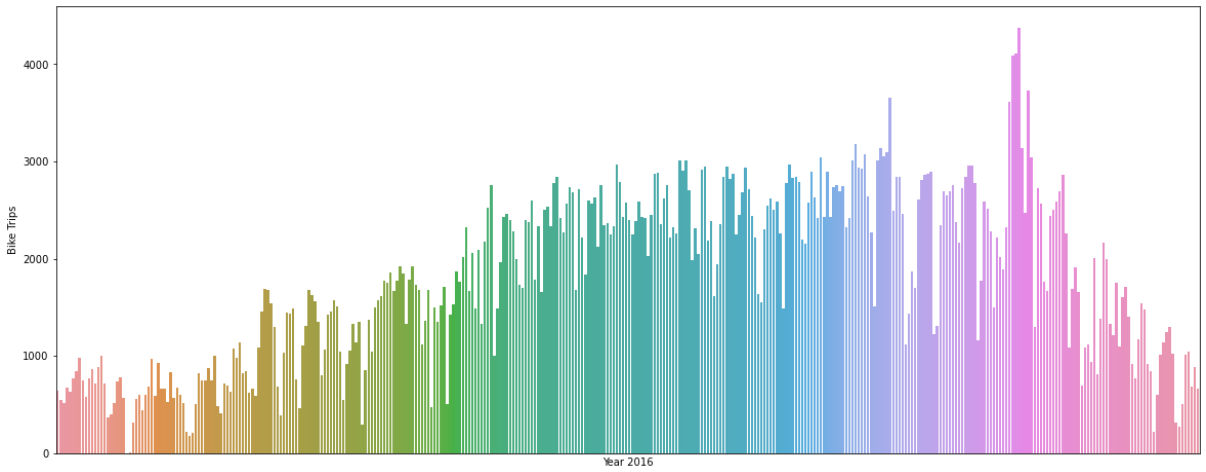


The season seems to have less influence on the average trip duration than the daytime has. As expected, the average trip duration reduces during the cold winter months and reaches its maximum in may, a month with rather mild weather conditions.



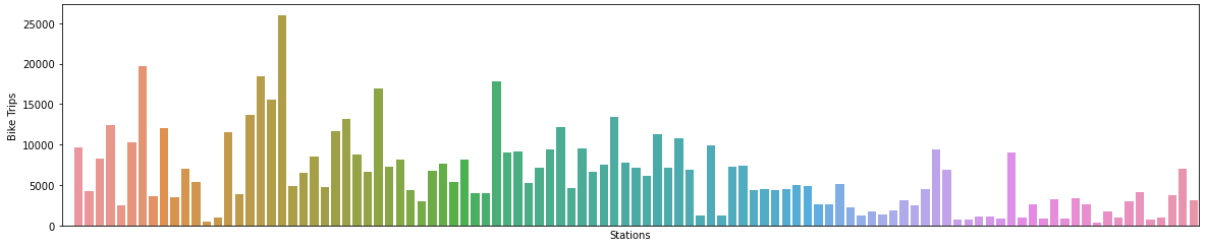
As displayed above, the weekday has a significant impact on the average trip duration. While the customers tend to perform shorter trips during the week that don't vary much between the specific days, a significant increase can be seen during the weekend.

Trips per Day



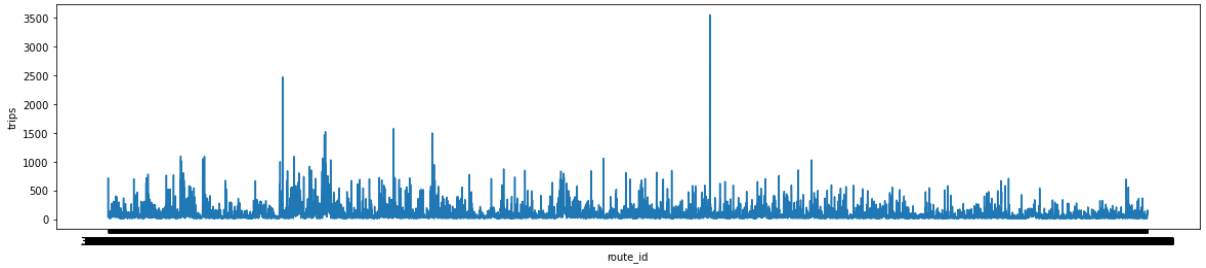
This barplot describes the trips per day in the course over the year. As expected you get more trips per day within the month with mild weather conditions.

Trips per Station / Year



Usage is different from station to station. Most stations are not frequented very much but a small number of stations have high usage. As you can see in the heatmap above, the higher frequented stations are in the middle of the city.

To get a better overview, we made a plot which shows the most used routes. There are just some routes which are taken excessively. Most routes have a usage of around 100 trips per year.



	trips
count	6361.000000
mean	100.222292
std	143.595784
min	11.000000
25%	24.000000
50%	52.000000
75%	116.000000
max	3551.000000

3. Predictive analytics

a. Feature Engineering

A possible correlation between total system-demand in the next hour and the following features could exist:

- Hour of the day

As we have seen in task 2, demand during a day differs greatly from hour to hour. There are certain trends visible as already explained. That is why we selected hour as a feature.

- Day of the week

Also as seen in task 2, there is higher demand during the workdays than during the weekend. That is why we have selected a boolean variable with 1 for a workday and 0 for weekend.

- Time of the year

In task 2 it became visible that demand during summer is on a relatively stable level from July to November. That is why we created another boolean variable for the regression, summer, setting it to 1 during these months and 0 during the others.

- precip (yes/no)

It is likely that precip has a negative influence on the demand for bikes. Since people tend to prefer to not get wet by the rain and riding a bike exposes them to that, one can assume that people rent less bikes when it is raining.

- Temperature

Temperature goes to some point hand in hand with time of the year since it is warmer in the summer and colder in the winter. One can assume that the temperature has a positive influence on demand, as better and warmer weather makes riding a bike and by that renting one more attractive.

- Major events

In July 2016, during the week of July 23 to July 28, there were several events, such as a concert by Lady Gaga, Snoop Dog or a theme party "DNC Rock 'n' roll kickoff at Johnny Brenda's" in Philadelphia. However, surprisingly, this had no effect on demand, as you can see in our graph "Overview of daily demand" in July.

- Holiday

As well as for major events we checked for holidays as a possible feature. We looked at the fourth of July and could not find any significant influence. This is why we decided to not include holiday as a dummy feature.

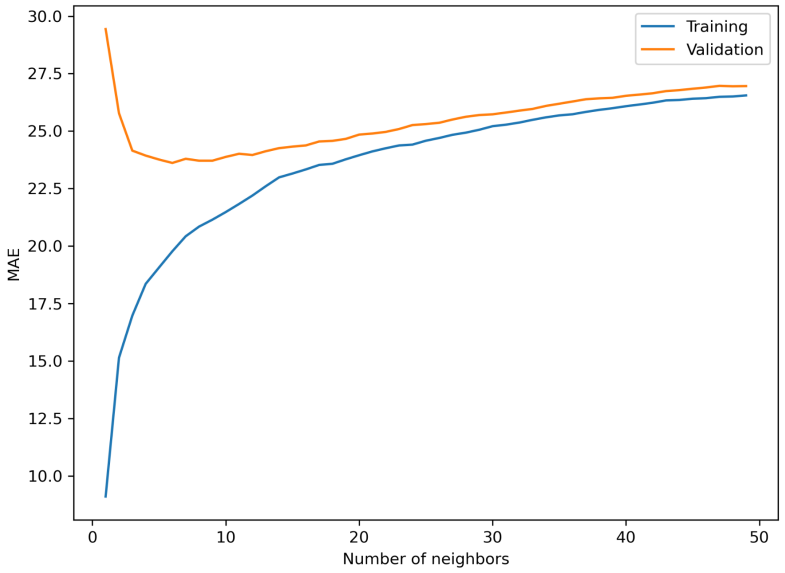
b. Regression Models

- **Decision Tree Regression**

Since decision trees are very prone to overfitting and easy to understand as well as to interpret, we chose this model for our regression. It is also not influenced by outliers which exist in our data set and can capture non-linear relationships. Since at the first look out data is more non-linear, this is also an important advantage for the model.
We controlled the disadvantage of overfitting of the model by graphically determining the optimal tree depth, which we then set to 9.

- **KNN Regression**

We chose the KNN model because when we look at the data there are less linear and more non-linear features in our prediction task. This becomes visible when looking at the influence of the weekday on the demand for example and also explains the bad performance of the linear regression later.
The issue that KNN models can not be used for feature selection is not an issue for us since we pre-selected our features based on common knowledge and a deeper look at the relationship between demand and several features such as weekday and hour. The other disadvantage of the KNN model regarding its speed is also secondary for us since our dataset is not that big.
We also determined the number of neighbours graphically. The k value determines the number of neighbors we look at when we assign a value to any new observation. A very low value would lead to overfitting. When we take a look at our graph, a value of 11 seems to produce the lowest MAE when looking at the training and validation set.



- **Ridge Regression**

We chose Ridge Regression to also demonstrate a linear model.
Also, in Ridge Regression, the coefficients are regularized ("shrunk").
This means that the estimated coefficients are pushed towards zero.

The Ridge Regression is appropriate for our data set because it "optimizes" our prediction and allows us to avoid overfitting.
A disadvantage of ridge regression is that it does not reduce the number of variables because it never results in a coefficient being zero. Rather, ridge regression only minimizes it. Again, this aspect is secondary because our dataset is not that big.

c. Model Evaluation

To evaluate the performance of the aforementioned regression models, R^2 error, also known as Coefficient of Determination, was chosen. This particular metric compares the model with a constant baseline, which is determined by taking the mean of the data and drawing a line at the mean. The value varies between 0 and 1, with 1 representing a perfect fit and 0 an ineffectual one, thus the closer the value is to 1, the better the model. Out of the three models, the linear model fared the worst, with a score of 0.26635. Due to the lack of linear relationships in the prediction task, which are necessary for the linear model to function properly, this is far from a surprising result. The decision tree algorithm fared much better, with a score of 0.75523. Given the decision tree's ability to handle numerical and categorical data and to mirror human decision making, such performance was to be expected. While the KNN algorithm actually did slightly worse, with a score of 0.73245, we opted for it nevertheless for the reasons mentioned in the previous paragraph, hence the KNN algorithm received a closer look under two additional evaluation metrics, Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE).

The KNN algorithm had a MAE of 24.01443, and a RMSE of 38.36756. It should be noted that the RMSE will always be larger or equal to the MAE, and the greater the difference is between them, the greater the variance of the errors in the sample becomes. Since the dataset we are working with includes outliers we would advise to pay closer attention to the MAE, which is more robust to such datasets. The MAE value implies that on average, the forecast's distance from the true value is 24% of the true value, which may sound like a significant difference, but given the arbitrary nature of MAE and the fact that the dataset records human behaviours which may vary greatly, conclusions cannot be drawn from the metric alone until the data is well understood.

d. Outlook

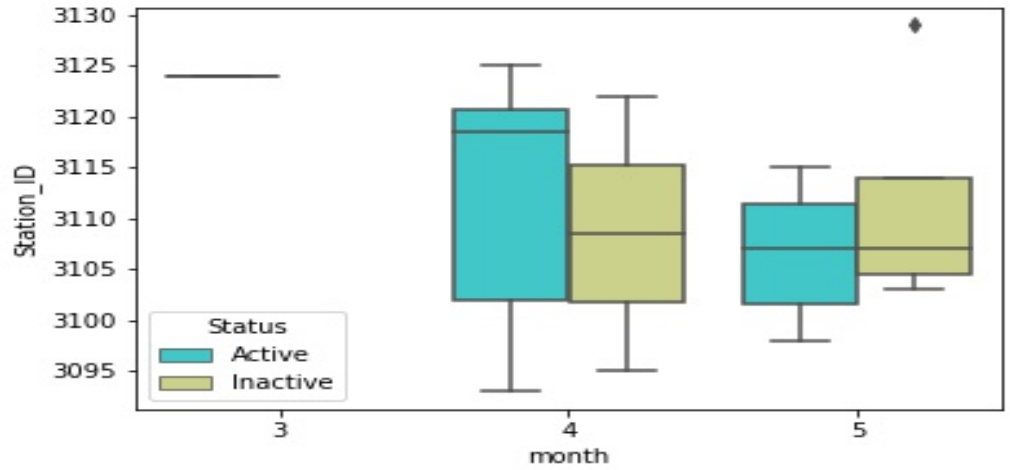
As we have seen in the evaluation of our model, it is not perfect. There is room for improvement when we look at the R^2 error which should be a goal for a follow-up project.
One way to make an improvement possible could be a stronger focus on non-linear regression models, since it became clear that our linear model was by far the worst choice in terms of errors.
Furthermore, one could include more external data such as data about traffic or more weather data. Possible features could be humidity, wind speed and a general status of the weather, meaning weather it is clear, cloudy, foggy or snowy.
Another possibility to improve the model would be looking at a greater time span, e.g. two or three years. This would limit special random effects at certain times and create more data points which could be helpful. There are also some other models that would deserve a closer look, such as random forests.

To improve the usefulness for the business side, more data about the customer would be great, maybe dividing them into groups of age.

4. Extras

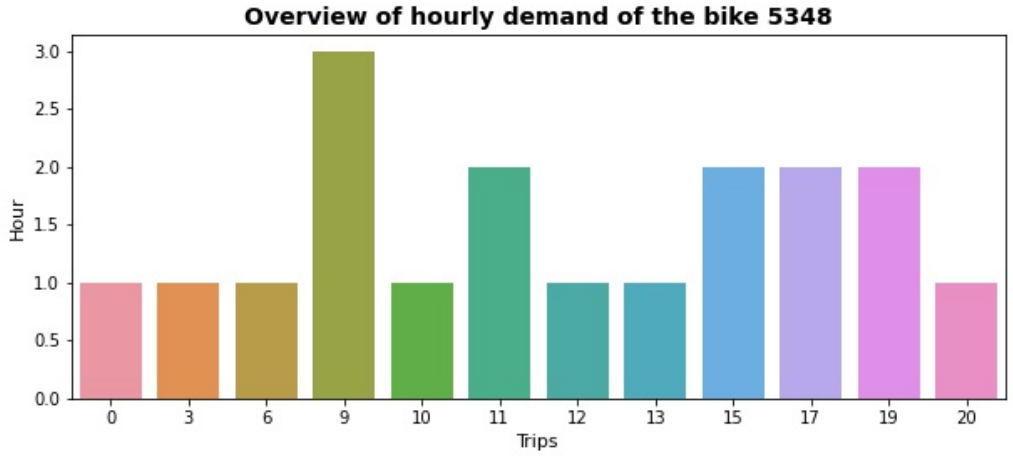
a. Station_Status

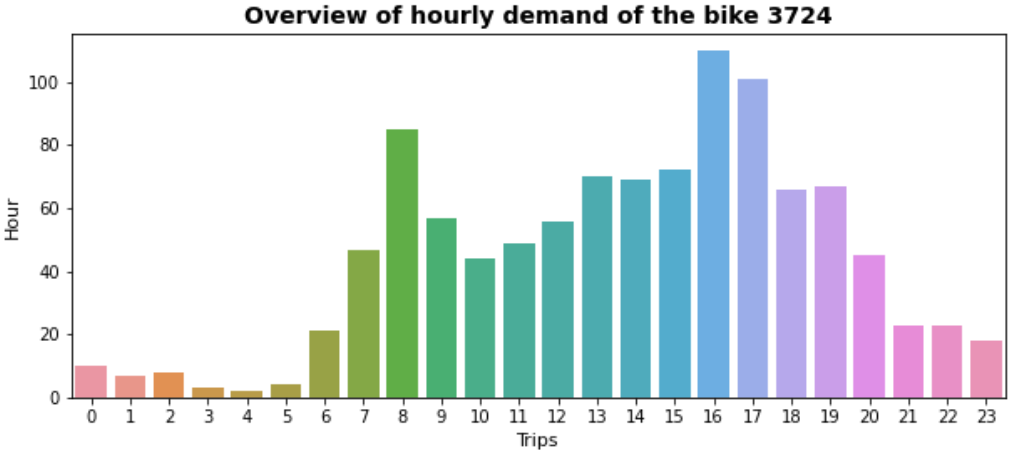
In addition to the “Trip data” we have got more information about stations, for example about their availability and names. For the year 2016 we only have data about 32 stations. This data covers the months March, April and May. As we don’t have information about all stations and all the months of the year, we couldn't draw parallels between the availability of the stations and their popularity and the registered trips per station.



b. Average uses per bicycle

While observing the average trips per bicycle we see a big difference among the bikes. The average bike is used for 639 trips, but looking at the data we see that one third of all bikes are used for less than 639 trips, seven percent are even used for less than 300 trips. It shows that some bikes are more popular than others. Looking at the stations where the bikes are registered and their geographical demand can give us the reasons why partitions between the bikes are so uneven and how to change it. This could help us to improve the bike sharing system and make more trips possible. To illustrate the unevenness between the bikes, we have diagrams from two bikes.


Overview of hourly demand of the bike 5348
Hour
Trips
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0
3
6
9
10
11
12
13
15
17
19
20


Overview of hourly demand of the bike 3724
Hour
Trips
0
20
40
60
80
100
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23