University of Cologne

PROGRAMING DATA SCIENCE

Project Report

June 10th, 2020

Project team:

Manuel Gassner

Irene Xu

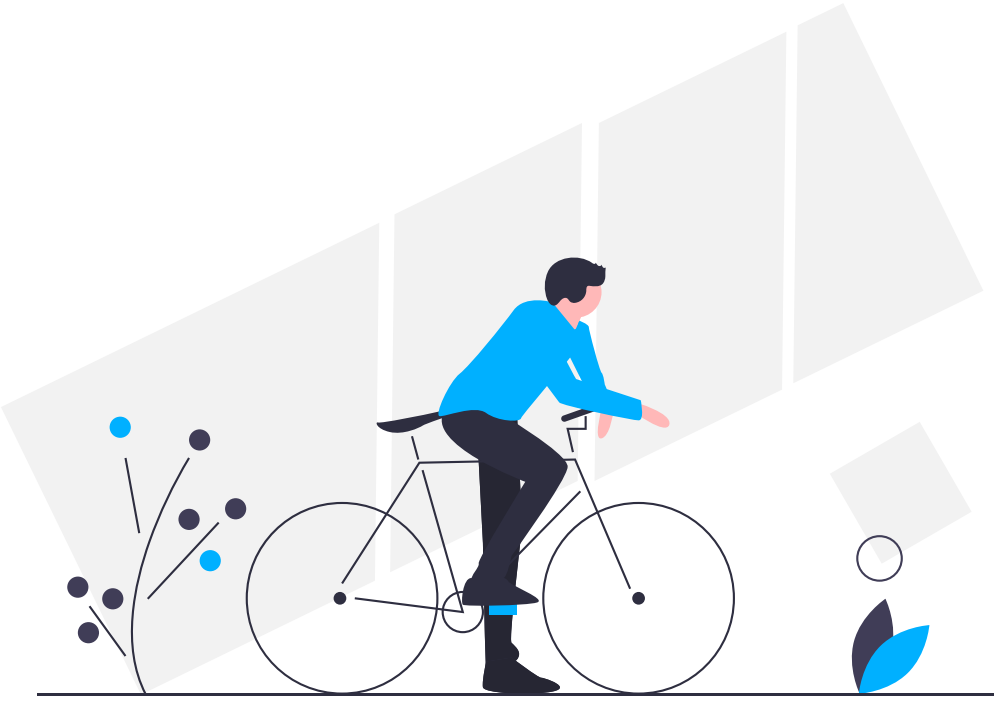
Long Le

EXECUTIVE SUMMARY

This project examines NextBike's trip data of bike usage in Frankfurt, Germany in 2019. Nextbike is one of the key players in the bike-sharing business that produces and operates bike sharing systems around the world. The report includes 5 phases of a data science project: data processing, data exploration, data visualization, statistical analysis, and model development. The goal is to gain insights into the complex bike sharing activity patterns and build a prediction model for bike-using-duration and trip direction.

To construct the prediction model, several analyses are performed on the underlying data set to identify the promising predictors that can forecast the trip duration and the trip direction. The findings emphasize the possible relationship over meaningful visualization. Machine learning algorithms are applied to build both prediction and classification models. A test set is used at the end to assess the performance of the models.

The results for both prediction and classification models show that further improvement is needed. The validity of the classification models fit is questionable. The evaluation of the models indicates that the predictors we selected are not valid indicators for predicting trip duration. The data set should be further explored to find right trends and indicators. In addition, it is also clear that the data set has a large amount of outliers, noises, and measurement errors. Removing those bad data significantly reduces the size of the data set, which greatly affects the accuracy of our models. Thus, the identification of outliers is important for future work on this data set.



THE PROJECT TEAM

The project team contains three individuals who are studying at University of Cologne, Germany. We are a diverse team with multiple nationalities and backgrounds. Working as a team, we shared the common tasks of the project and discussed regularly during the project. Besides, each of the members was aslo reponsible for different tasks in order to boost the speed of the project.

Manuel Gassner

Business understanding, Data understanding, Code refactoring, QA, Descriptive analysis of data, Creation and evaluation of the prediction models, and Handling scm (GIT)

Irene Xu

Business understanding, Data understanding, Code refactoring, QA, Descriptive analysis of data, Creation of report and presentation slides, and Data description

Long Le

Business understanding, Data understanding, Code refactoring, QA, Descriptive analysis of data, Data Visualization, Creation of report and presentation slides, Creation of the prediction models

TABLE OF CONTENTS

1. PROJECT INTRODUCTION

Bike sharing service has evolved fast over the past decades. With the latest technologies, the whole bike rental process has become automatic, users can rent and return bikes at fixed stations or anywhere within the flex zone by using mobile operating systems. These technologies also provide vast amounts of data, which can be used to analyze and develop prediction models, helping companies and services providers to better understand their customers.

Nextbike is a German company that produces and operates bike-sharing systems around the world. The business model of Nextbike is to provide tailor-made mobility solutions to public transport systems, companies, and universities (B2B). It also offers multiple branding and added-value opportunities to its sponsoring partners such as AdvertisingBike and BusinessBike. In this project, we will examine the data provided by Nextbike to explore how their bike sharing system is used, investigate users' behaviors to define possible travel patterns, and develop relevant prediction and classification models to help the company to manage the system in a more efficient and cost-effective manner.

The data set we analyzed consists of trip data of Nextbike's bike in Frankfurt for most of 2019 and it has not been extensively preprocessed yet. We will focus on data manipulation, data cleaning, and data visualization first, and then we will use all relevant information to develop prediction and classification models. The goal of this project is to gain insights into the complex bike sharing activity patterns by using different tools and techniques.

2. INITIAL DATA EXPLORATION

The provided data set consisted of ride information about the bike being rented, bike stations, geographical coordinates and status of each booking. The data set covered approximately 532,000 entries of the year 2019. Originally, there were 13 columns in total. The below table the types and meanings of the columns are stated in Table 1 details all data types and descriptions of the original columns in the provided data set.

To further explain the data, the prefix b represented booking or bikes. The b_number column showed 5-digit bike numbers which started with the leading “0”, (e.g. 00001). In the data set, most of the bike numbers started with “38”, which might indicate the area code of Frankfurt, the city in study. The b_bike_type column indicated the type of the bikes. There were 4 types of bike in total and the most common bike type was type “15”. According to Nextbike’s website at the time we conducted the project, the company offered 4 bike types in city areas: SMARTbike, SMARTbike 2.0, e-SMARTbike, and e-SMARTbike 2.0. Another possible method to categorize bike types was to categorize by business purposes. Nextbike’s website detailed 4 types: BUSINESSbike, ADbike, SPONSORbike, and CAMPUSbike.

The prefix p represented places or positions. Places could be the actual/official stations, where the bikes were available to rent or were allowed to return. They could also represent free-floating areas as the bike operator also allowed customers to rent and return bikes on any public road within a defined zone. The p_bike column helped to determine whether a bike was in a free-floating mode not. If the value of this column showed “true”, then the bike was in a free-floating mode. The p_spot column with ‘True’ or ‘False’ values indicated whether a bike was in an official bike station. The p_spot = “True” parameter indicated the bike was in an official station. The p_name column showed the human readable name of the bike station/area while the p_number column showed the station/area number. Most official stations had a 4-digit station number such as “4250”. The p_uid column provided the unique ID of the location, and the p_bikes column showed the number of available bikes in the station. The p_place_type column indicated the types of places/stations. The value “0” represented the official stations, whereas the value “12” represented the floating areas. Values “3”, “4”, “7”, “9”, “13” represented the education institutions, companies (businessbike), hotels, business areas and aged care institutions (hospitals, school for the elderly) respectively.

Positions were the locations of the bikes. The p_lat column and p_lng column contained the latitude and longitude of the location for the bikes respectively. Together with the date-time data from the datetime column, we could find out the status and the duration of each trip.

The trip column was the most important column for our data cleaning stage. It showed four values: “first”, “last”, “start”, and “end”. The “first” value usually appeared once a day at

Variable Name	Format	Description
b_number	int64	ID of the bike being rented
b_bike_type	float	Types of the bike, there are 4 types of bike
p_spot	bool	If the bike is in the bike station, T = in station, F = not in station
p_bike	bool	If the bike is floating, T = in, F = not floating
p_name	object	Human readable name of the bike station
p_number	float	ID of the bike station
p_place_type	int64	Types of places, e.g., 7 = hotels, 3 = education institutions
p_uid	int64	Unique ID of the location
p_bikes	int64	The number of available bikes in the station
p_lat	float	latitude of the start/end location
p_lng	float	longitude of the start/end location
datetime	object	YY-MM-DD 00:00:00, date-time of the booking
trip	object	Status of the booking (start/end, first/last)

[Table 1]

Table 1: Variables of Nextbike's Data Set

00:00:00, and the "last" value usually appeared once a day at 23:59:00 with the exact same geographical coordinates. An actual ride would not be able to take place if the location of the bike remained unchanged. Therefore, we assumed that the "first" and the "last" values of the trip column represented the first and the last "ping" of each bike that was online and available to use for customers. There were also some cases that the "first" or the "last" value appeared at different times, whereas the geographical coordinates still stayed the same. In those cases, we assumed that the certain bike was offline at a certain time for some reason (e.g. maintenance) and was not available to use. Assuming that a bike was only bookable by one person, and the rows that had the same bike number within a certain period of time are the data from one trip, we found out that in general a "start" value would be followed an "end" value, and they appeared at different times with different geographical coordinates. Thus, we concluded that the "start" represented the start of a bike booking and "end" represented the end of a bike booking. Only the booking that had one "start" entry and one "end" entry could be seen as an effective, valid booking. In other words, the trip that had "start" and "end" values were more informative for the analysis of mobility patterns than others.

3. DATA PROCESSING

After closely examining the data set, we decided to remove all the “first ‘’ and “last” data as they only represented the availability of the bike, but not the actual trips. The data of the actual trips was the most relevant data to analyze sharing bikes’ mobility patterns. We also removed the false trip data that contained double “start” or “end” values. Based on the assumption that once a bike was rented, it would not be able to be rented by someone else. As such, the kind of data showing that the bike was already occupied but had a ‘start’ appeared before it ended was dropped. After that, we arranged the data into a more meaningful order in which one row contained data for one trip. The number of entries was significantly reduced to 154,358 entries.

Then we used the data from the datetime and p_uid columns to calculate the duration of each trip. As the data cleaning process was going, we dropped all the trips that last less than 3 minutes without a change in location. According to Nextbike’s public information on their website, sometimes the frame lock would not open even after several attempts to open it via the app. Nextbike’s system would automatically detect such failure after three minutes and cancel the rental. As such, no actual ride took place and the decision to drop these trips were justified.

We also dropped the trip data that last longer than 2 hours, according to the Census Bureau’s report on commuting in America, the average bicycle commute time is 19.3 minutes, and most bicycle commutes were between 10 and 14 minutes long (McLeod, 2014). We believed that trips with duration longer than 2 hours were unusual trips and that analyzing these trips would not create relevant insights. Moreover, including these trips in the analysis would negatively impact the results. As the trips had relatively low density in the distribution and their duration values were significantly higher than the majority of the remaining trips (see figure 1), their high standard deviation would lead to inaccurate predictions. Therefore, we decided to exclude the trip whose duration was longer than 2 hours and kept them for another further analysis. The number of trip data we had at this point dropped to 76,157 trips.

Finally, we removed the false trip data that had coordinates that were clearly not in Frankfurt. This kind of false measurement could happen if an error occurs at the bike station/spot while returning a bike. After removing all the affected trips, the size of our data set reduced to 72,759 trips.

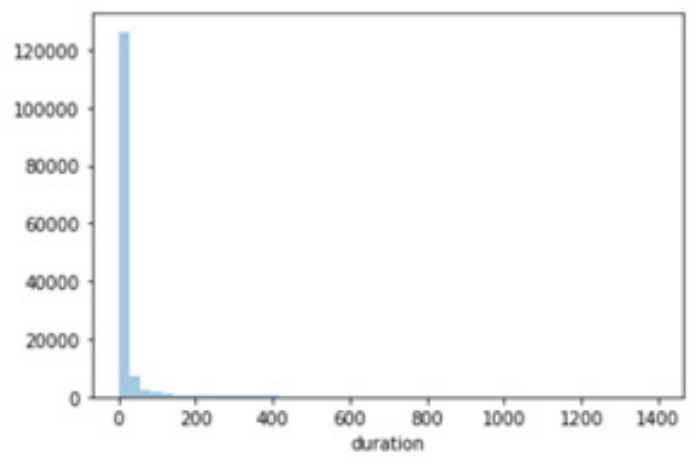


Figure 1: distribution of trip duration

When the new data set was finalized, we created a new DataFrame that stored the following columns:

Bike_number	Start_Time	End_time	Start_Longitude	Start_Latitude	End_Longitude	End_Latitude	Start_Station	End_Station_number	Weekday	Duratic
38999	2019-12-27 12:01:00	2019-12-27 12:13:00	8.692439	50.128915	8.687546	50.115009	4252.0	4247.0	True	12
38999	2019-12-27 21:02:00	2019-12-27 21:16:00	8.687546	50.115009	8.692439	50.128915	4247.0	4252.0	True	14
38999	2019-12-28 08:16:00	2019-12-28 08:27:00	8.692439	50.128915	8.682500	50.115556	4252.0	42006.0	False	11
38999	2019-12-28 16:37:00	2019-12-28 16:53:00	8.682500	50.115556	8.692439	50.128915	42006.0	4252.0	False	16
43400	2019-06-29 12:59:00	2019-06-29 13:14:00	8.768889	50.134167	8.768889	50.134167	42003.0	42003.0	False	15

Table 2: A part of the new DataFrame created after data cleaning

Bike_number: Bike Number

Start_Time: Start Time of the trip

End_Time: End Time of the trip

Start_Latitude: Latitude of the start location

Start_Longitude: Longitude of the start location/position

End_Latitude: Latitude of the end location

End_Longitude: Longitude of the end location

Duration: Duration of the trip

Weekday: Binary variable with 1 indicates weekday

Start_Station: Number of the station where the trip started

End_Station_Number: Number of the station where the trip ended

Month: the month during which the trip took place

Zip_codes: the postal codes of the start locations

Bikes_on_Position: Number of bikes at stations when a bike was rented

End_Bikes: Number of bikes at stations when a bike ended

4. DATA ANALYSIS

4.1 Descriptive analysis

After we dropped every trip that lasts less or equal to 3 minutes with no change of locations, we started our analysis by investigating the distribution of trip duration. As figure 1 shows, the distribution of our trip data is significantly right skewed; the duration is range from 2 minutes to 1394 minutes with the average of 23.55 minutes and the median of 16 minutes. 75% of our trip duration is equal to or less than 27 minutes. This tells us that our data set has a few large values (outliers) that drive the mean upward but does not affect the median.

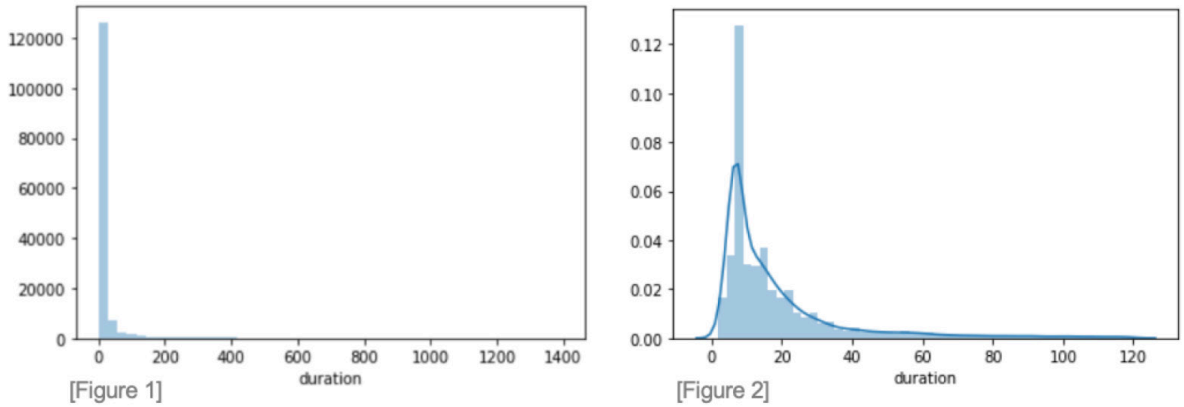


Figure 1 & 2: distribution of duration before or after removing trips that >120 minutes

The outliers could be caused by system failures or unsuccessful log-out attempts, which should be dropped as they are incorrectly entered or measured data. Some of the outliers may come from the legitimate trips and should not be easily dropped. However, as we used common statistical procedures (e.g., linear regression), outliers can affect both our prediction assumptions and results. Thus, we decided to drop some outliers and focus our investigation on a more predictable part of the distribution. From our analysis of the distribution, we found that 93% of our trip data is from 2 minutes to 120 minutes, which means only 7% of our data is larger than 120 minutes. In consideration of the low density of that part of the distribution and its high impact on the standard deviation that will lead to bad results for the predictions, we decided to drop all the trips that have a duration higher than 120 minutes from the data set. Figure 2 above shows the new distribution of trip duration. Since the distribution is still right skewed, we chose to use a log transformation for the duration to get a more normally distributed shape, which could help us when we create our prediction model. Furthermore, since the distribution is not normally distributed, we decide to examine each month's distribution separately. We found out that every month has a duration interval that is more likely to appear, an example is shown below (figure 3: Distribution duration for August):

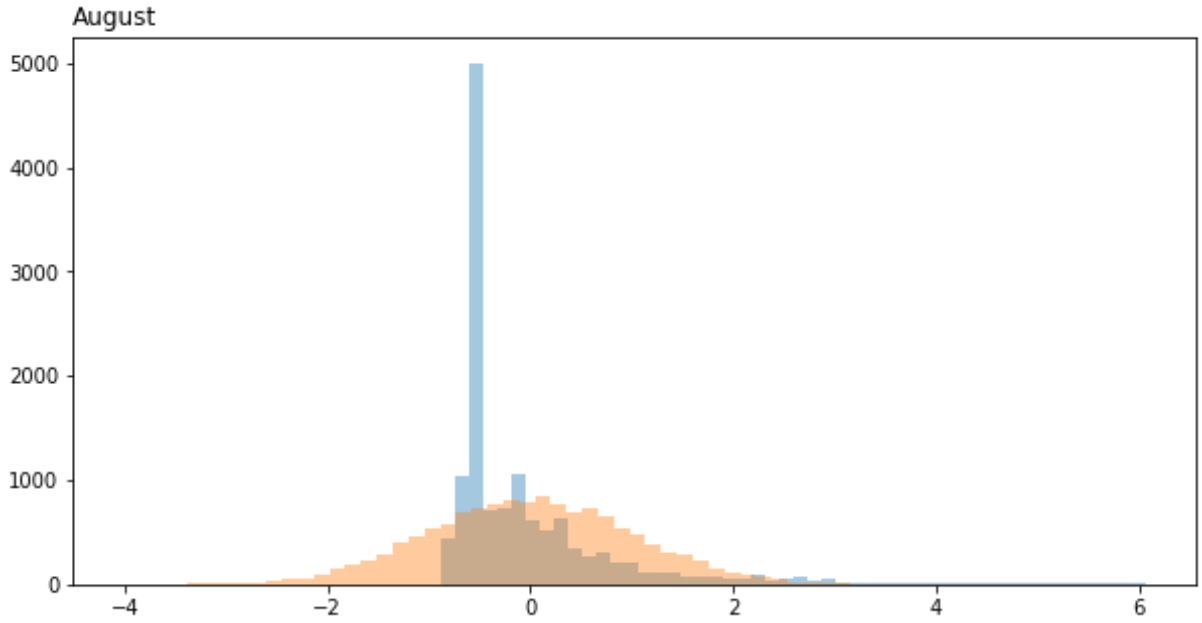


Figure 3: Distribution duration for August

Based on the above, we looked into the weekly trend of the trip duration (see Figure 4) to pinpoint the reason for that peak in the duration of each month. We found an anomaly at the date of 26th november. The anomaly was caused by a “batch” booking of 250 bikes for 384 minutes on the same time at 1 a.m., which seems very unlikely to happen in reality. When we looked at the distribution of each month separately, we could clearly see that this batch bookings influence the distribution of each month greatly (see figure 4).

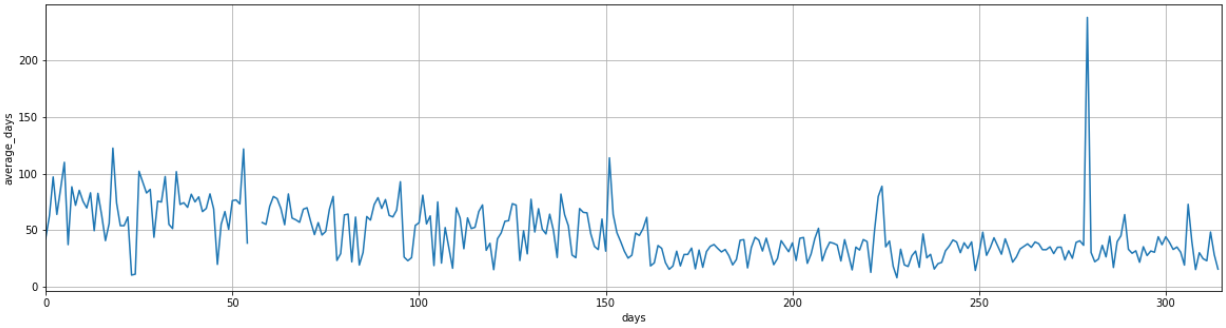


Figure 4: Weekly average duration by days

We believe that these trips are caused by reallocation of bikes to a different station by nextbike employees. This kind of trip must be dropped, because they did not display the customers’ rental behavior. We investigated the whole data set to find all bookings that have a “batch” manner and dropped them from the data set. We also dropped every trip that has the same start and end time more than 4 (boundary for number of bikes that can be booked in parallel) times in the data set. After dropping these bookings, our data set dropped to a size of 51,393 trips. The monthly distribution shape goes closer to a normal distribution (see Appendix A1) by applying a log transformation. The overall distribution shape for the remaining 51,393 trips also became more normally distributed (see figure 5 & 6). Therefore, we will use the data set with a log transformation of the target value for our prediction.

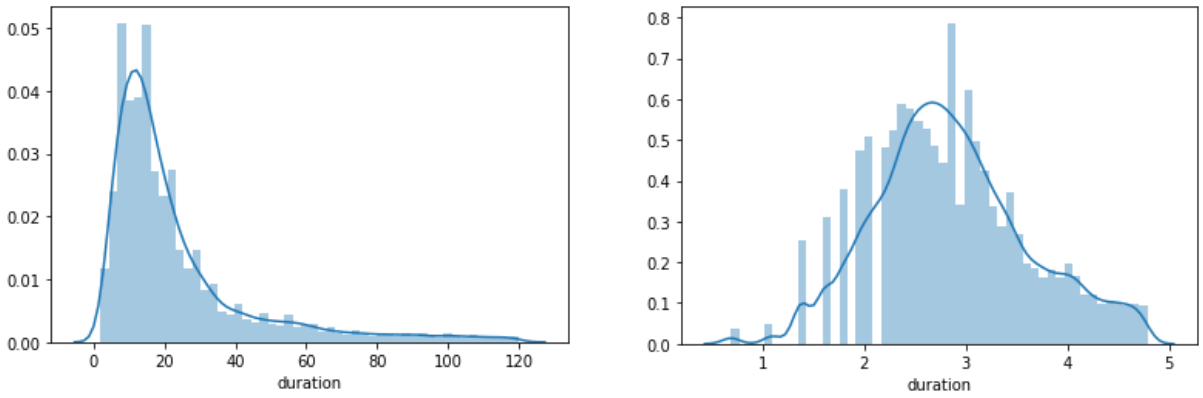


Figure 5 and 6: The distribution of duration <120 before and after log transformation

After dropping all the outliers that would affect our predictions, we visualized aggregate statistics for the trip duration per month, per day of week, and per hour of day to further investigate data. Figure A2 shows the month based analysis. From the number of bookings per month, we can see a clear seasonal trend of more trips in summer month (June, July, August) than winter month (Dec, Jan, Feb). The average trip duration and average standard deviation of the winter month are also lower than other months, which means that the people had a lower demand for bike sharing services in winter month. Besides, by examining the travel duration on weekends vs. weekdays, we found out that for most of the months (except winter month), the mean trip duration on weekends is longer than the trip duration on weekdays, whereas the number of trips occurred on weekdays is significantly higher than the number of trips occurred on weekends. This reveals a possible travel pattern difference between weekdays and weekends: The weekday bike travel is more in favor compared to weekend bike travel when there's a commuting need. However, for people who like cycling or primarily use bikes for transportation (e.g., cyclists), they may tend to ride longer during weekends than weekdays. This could be explained as the purpose of riding is changed from commuting to leisure or exercising, weekend cycling may be less likely to be influenced by weather.

Figure A3 shows the weekly statistical analysis. Although there's no significant difference in numbers of trips between each day of week, weekends travel behavior is slightly different from weekdays travel behavior: the mean duration of each weekday is 17.78, 17.27, 17.49, 18.39, 20.12 minutes respectively, while the mean duration of Saturday and Sunday is 20.20 and 19.99 minutes, so the mean trip duration on weekends (+Friday) are longer than on weekdays.

Figure A4 shows the hourly statistical analysis. We found that most of the trips occurred around 3pm in the afternoon. The average trip duration reaches its highest point in the morning commuting hours between 8-10 o'clock (go to work or go to school) and 12 o'clock (lunch hour), whereas the lowest point occurs at midnight/early morning (2am-6am). When looking at the differences between the weekday and weekend bookings, we once again found that the weekday bike travel is more in favor compared to weekend bike travel in commuting hours (6-8am; 3-5pm). In contrast, compared to weekdays, people tend to use

the shared bike in the afternoon (12-3pm) on weekends.

In addition, as we investigated aggregate statistics for the trip duration per month, per day of week, and per hour of day together, for each weekday on the hourly base, we found out that there were certain time frames that could also be seen as anomalies. An example of 11 o'clock on a monday in January can be found below

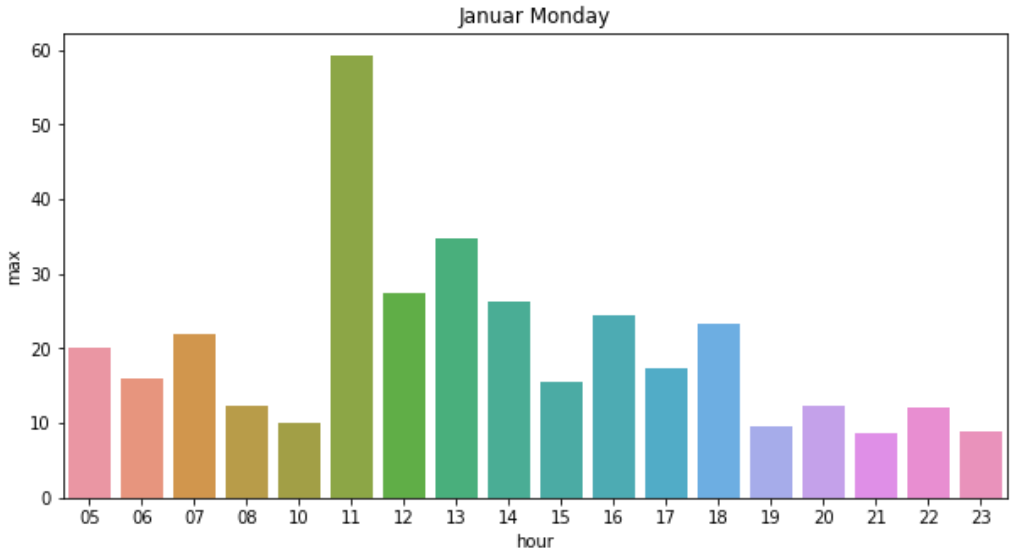


Figure 7: Average duration in weekdays

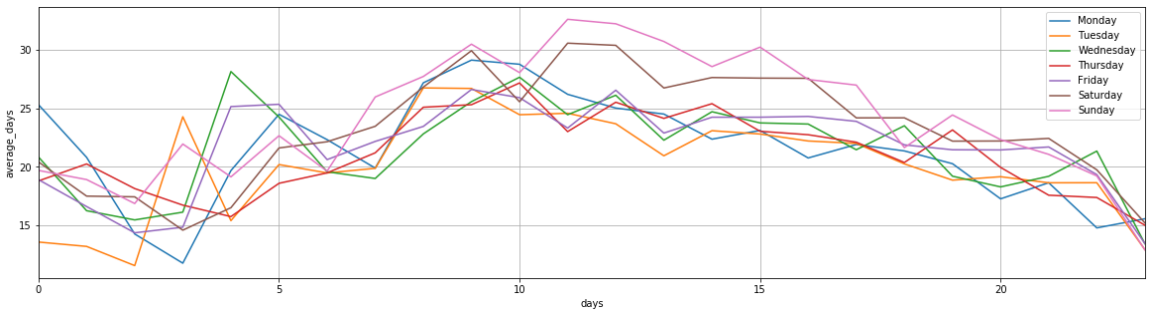


Figure 8: Average duration, January Monday

4.2 Analysis from Visualization

One of the project tasks was to investigate the geographical distribution of starting locations of the trips in summer months. The task required geodata in order to visualize the number of trips by regions. The official geo data for PLZ regions in Frankfurt was available for public download via the link https://www.suche-postleitzahl.org/downloads. The original file for administrative areas by postal code was provided in 'kml' format. The file was then converted into geojson format for the use in this project.

Figure 9,10,11 show a consistent pattern that more trips are generated within the city center than the surrounding areas. In other words, the closer the region is to the city center, the more trips are generated in that region. This pattern holds true for the summer months, which are June, August, and September. To discover if this pattern exists in the winter, we visualized the total trips by the starting PLZ region in the month December (see figure 12). The result shows a similar trend that the further the regions are from the center, the less trip generated.

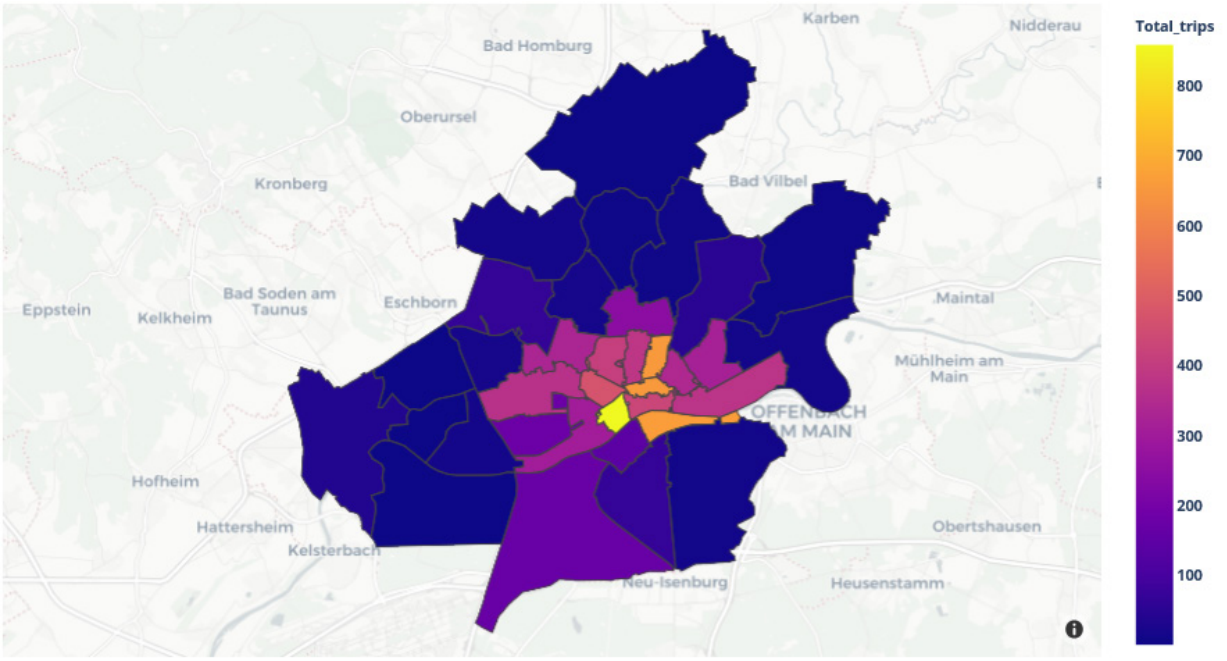


Figure 9: Map of number of started trips by Postal Region in June, 2019

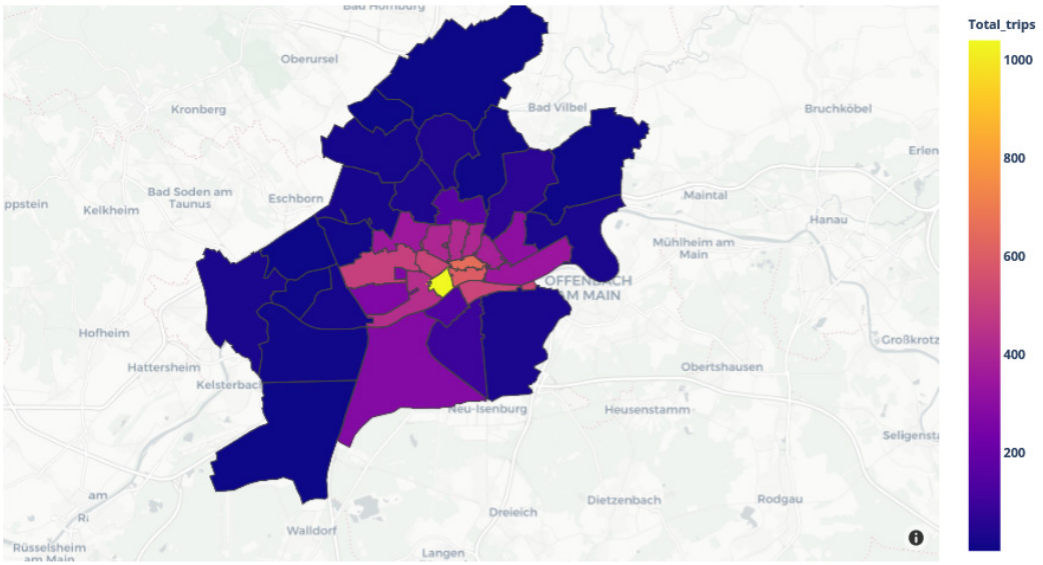


Figure 10: Map of number of started trips by Postal Region in Aug, 2019

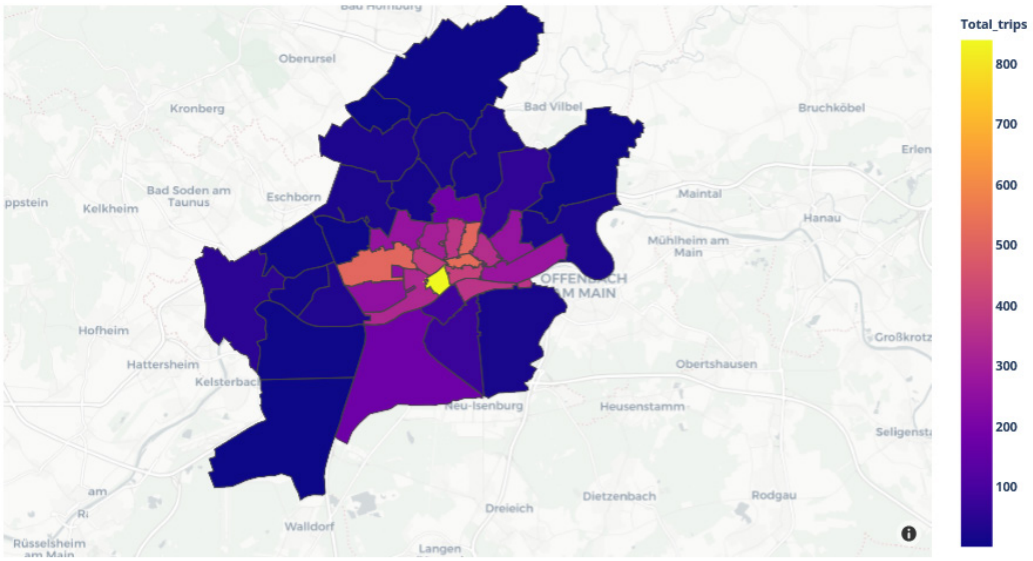


Figure 11: Map of number of started trips by Postal Region in Sep, 2019

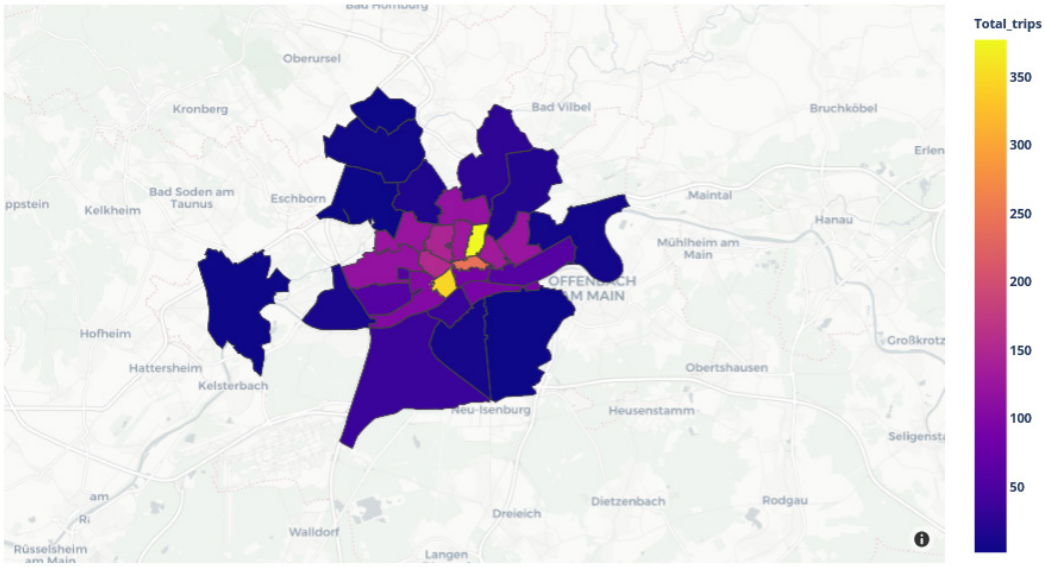


Figure 12: Map of number of started trips by Postal Region in Dec, 2019

As we explored further, we also plotted the average duration of the trips by the region they started on the map by PLZ regions. Figure 13, 14, and 15 display the average duration of the trips by the starting PLZ regions. One consistent pattern was identified. The average duration of trips that started in the center regions were shorter than that of trips started in the border regions. This pattern holds true for both summer months and December, 2019 (see figure 16). Based on this finding, we hypothesized that the region from which the trips started could be a predictor for trip duration. Therefore, we included the Border_district (binary) as a predictor for our prediction model, which will be reported later in this report.

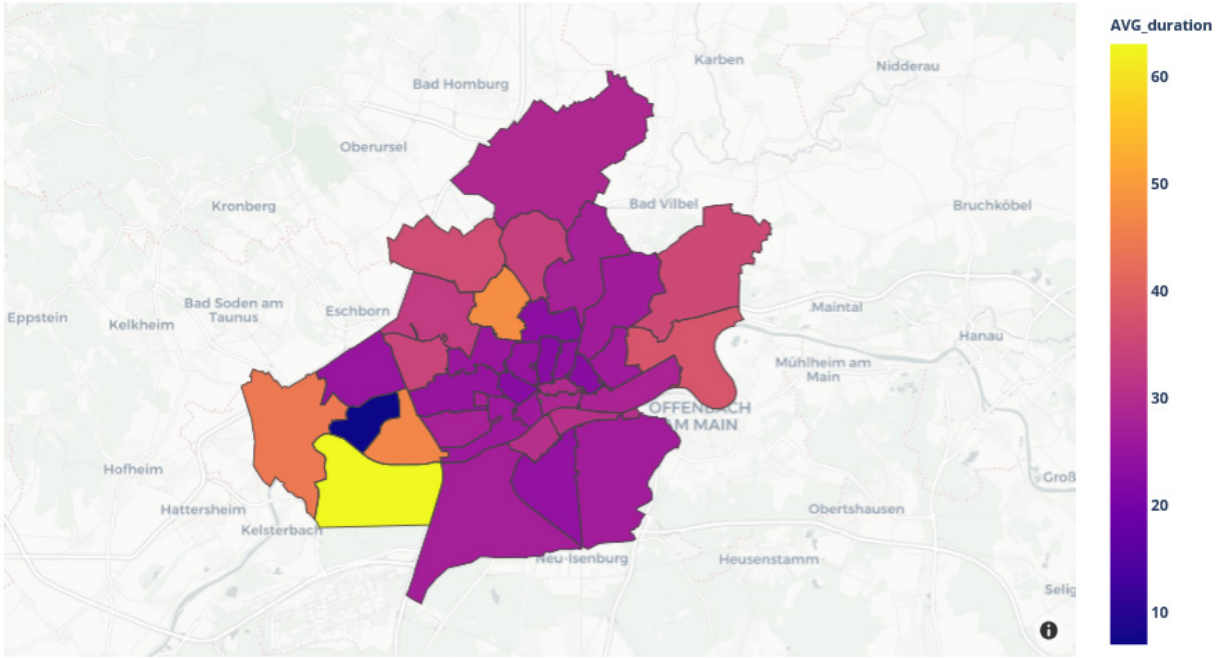


Figure 13: Map of average trip duration by started Postal Region in June, 2019

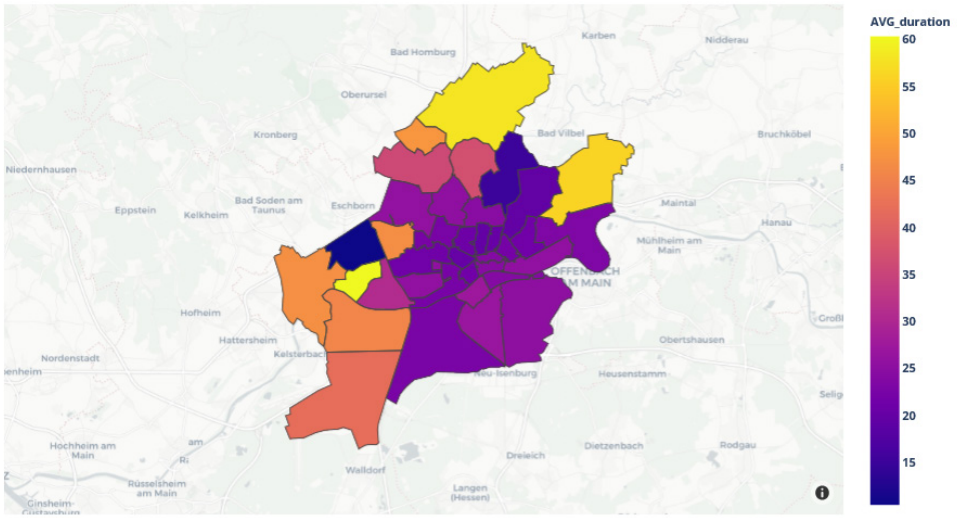


Figure 14: Map of average trip duration by started Postal Region in Aug, 2019

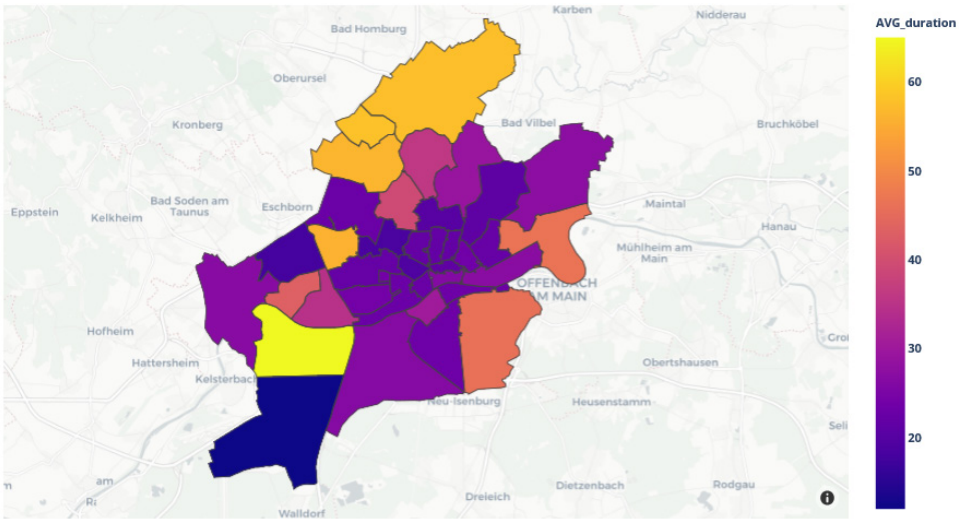


Figure 15: Map of average trip duration by started Postal Region in Sep, 2019

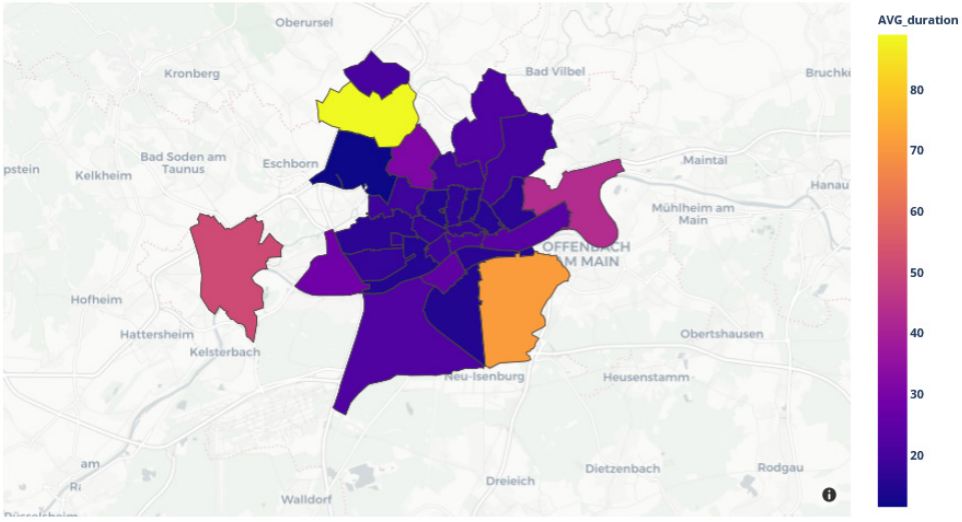


Figure 16: Map of average trip duration by started Postal Region in Dec, 2019

A fraction of the visualization task was to investigate the interesting aspect of the data. As such, we employed the heatmap to discover two parts of the biker using behavior:

Explore if people used bikes to travel to public events.

In order to investigate this, two public events were selected. The first one was the **IAA 2019 event**, one of the leading mobility exhibitions worldwide. The second event was a **football match** between Frankfurt home team, Eintracht Frankfurt, and Dortmund. There are two reasons for the selection of the two mentioned events. Firstly, the locations of the two events have different distances to the city center. While Messe Frankfurt exhibition center is located near the city center area, Commerzbank-Arena is more in the outskirts. Therefore, there might be different patterns in bike usage caused by the location of the event. Secondly, the IAA event was open for visit for a period of 10 days, whereas the football match was a one-time event. As such, this could be a factor that affects user behavior. The followings are the details of the two events and the results of the visualization:

IAA 2019

Location: Messe Frankfurt exhibition center

Visit day: From 12 to 22 September, 2019

Visit hours: Daily from 09:00 am to 07:00 pm, 20th of September from 11 am to 9 pm.

A heatmap with time slider was utilized to investigate if Frankfurt citizens used bikes to commute to the IAA event during its opening hours. Since the event was open 11 days and the opening hours ranged from morning to evening, visitors could decide their time for the visit. As such, we employed the data of the trips that ended 1 hour before the opening time and before the event daily end time for the visualization. The heatmap ran through 5 days before the event started and during the visiting days. The map visualization shows that there were trips that ended next to the exhibition center during the opening hours. Also, Weekends captured more trips to the event location than that of normal. In addition, as the event came closer to its end, more trips that ended near the exhibition center were captured in the map. Interestingly, 2 days before the first opening day, there were trips that ended inside the exhibition center. This could be explained that event staff and volunteers might have organized their preparation meeting before the event started.

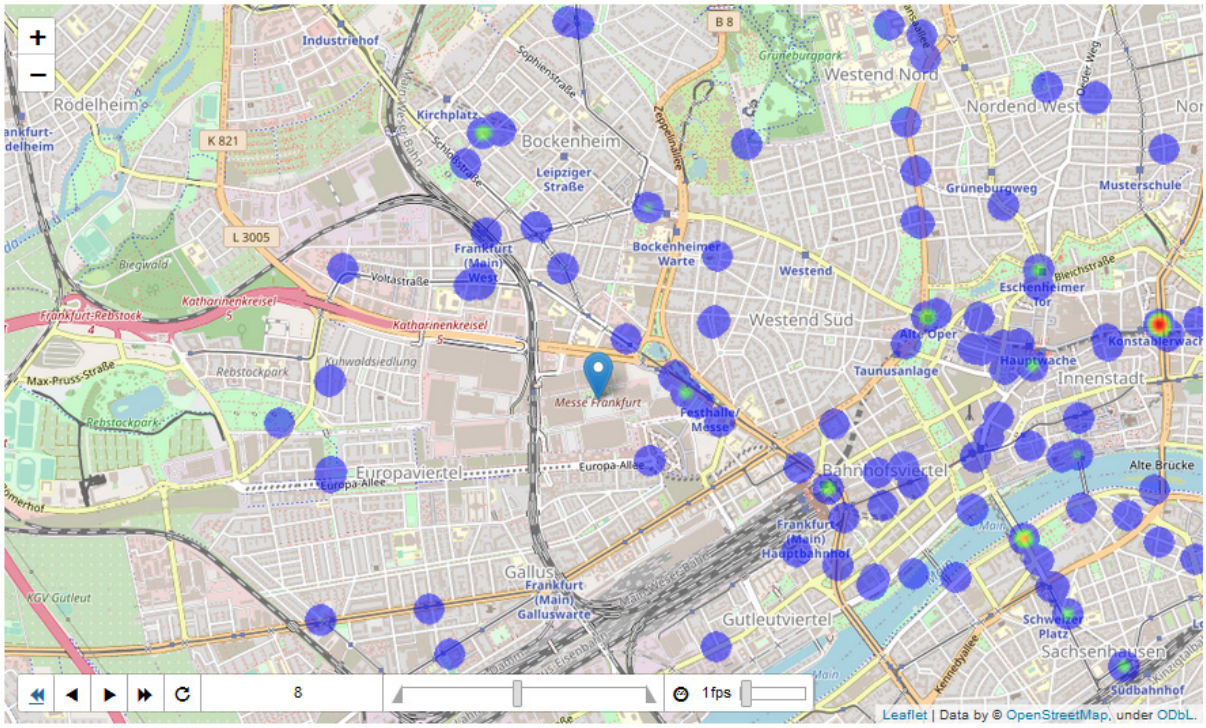


Figure 17: Heatmap of the end trips near the exhibition center on Saturday, 14 September 2019.

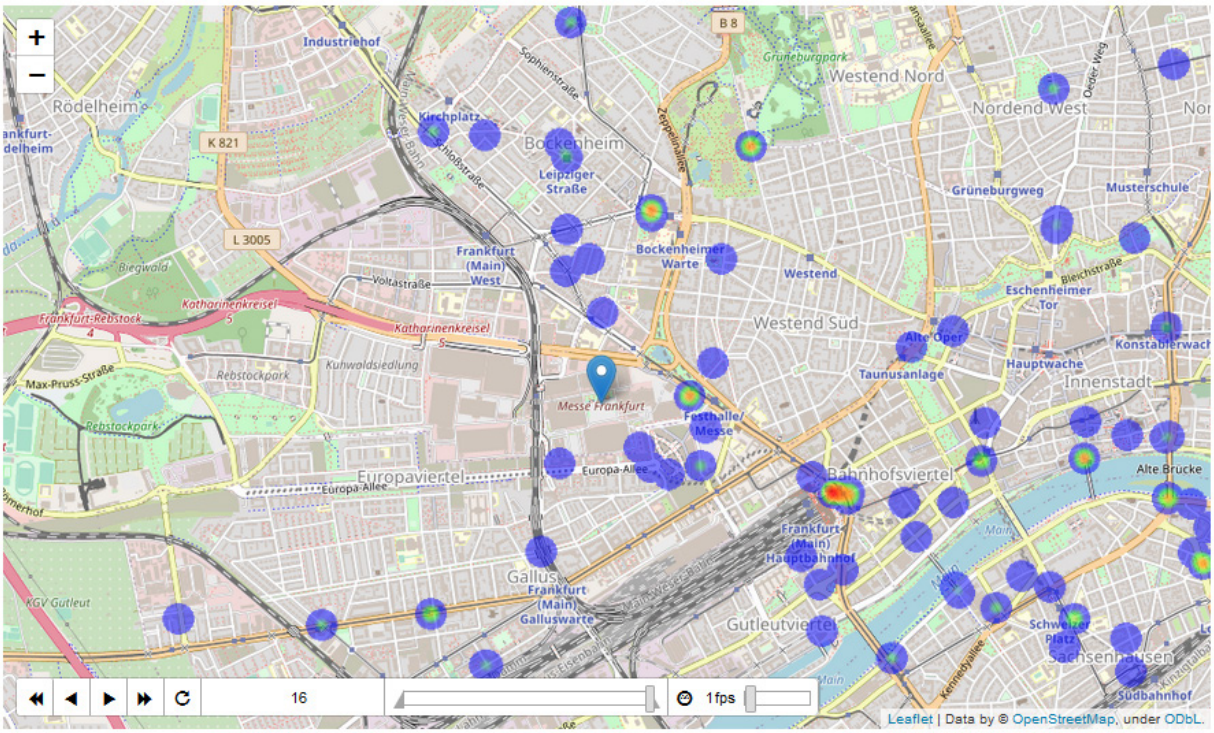


Figure 18: Heatmap of the end trips near the exhibition center on the last day of the event.

Football match between Eintracht Frankfurt and Dortmund

Location: Commerzbank-Arena, Frankfurt

Date and time: 18:00 CEST on 22 September, 2019

Attendance: 51,500

VS

To investigate if people commuted to the football match using bikes, data for the trips that ended 1.5 hour before the match started was employed in the visualization. In contrast to the IAA event, the map shows nearly no bike trips ended near the Commerzbank-Arena before the match started. The location of the Commerzbank-Arena might be a decisive factor in this case. The stadium was far from the city center. Also, the bike route to commute from the city center to the stadium might be difficult.

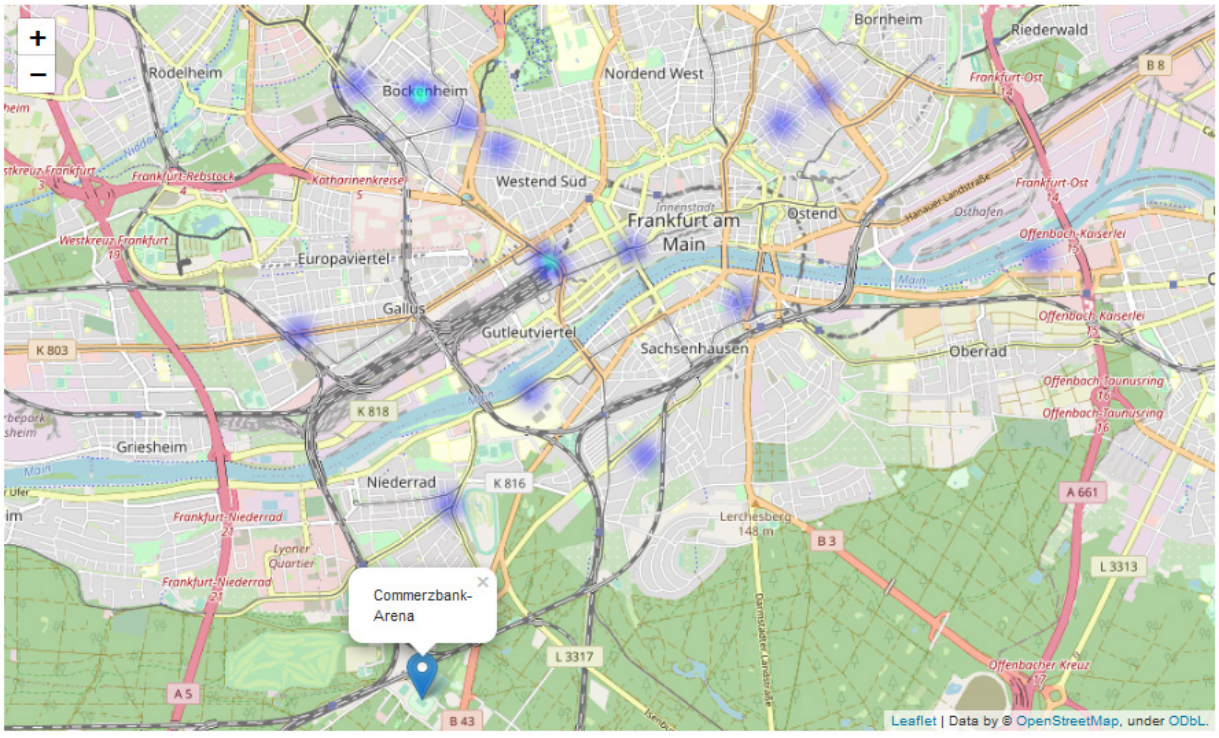


Figure 19: Heatmap of the end trips 1.5 hours before the football match between Eintracht Frankfurt and Dortmund started on 22 September, 2019

Explore if students used bikes to travel to university.

In Germany, the normal lecture time in University is from 8AM. Therefore, we employed the data of trips ending from 7AM to 8AM to investigate if there were bikes that ended near the university campus before the lectures started. We first visualized the number of bikes at fixed stations from 7AM to 8AM on two days May 03 and May 08 2019 (normal weekdays) to check our assumption that students would use bikes to commute from home to school. Using a Folium map with circle markers, whose size represents the number of bikes at the station when the trips ended, we captured the result that was inline with our expectation. From 7AM to 8AM on the selected days, the stations near the University of Applied Sciences were among the top stations with more bikes.

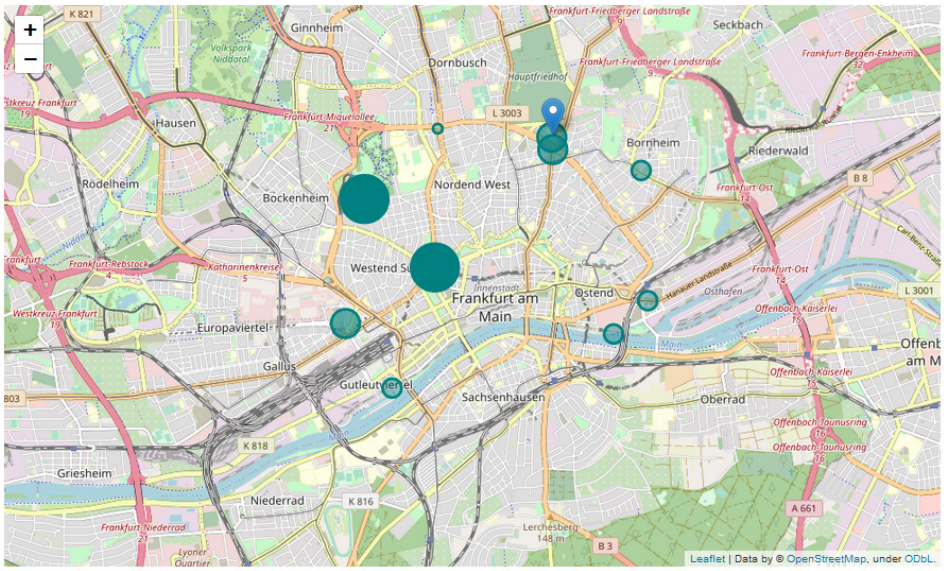


Figure 20: Map of the number of bikes at fixed station from 7AM to 8AM on May 03, 2019.

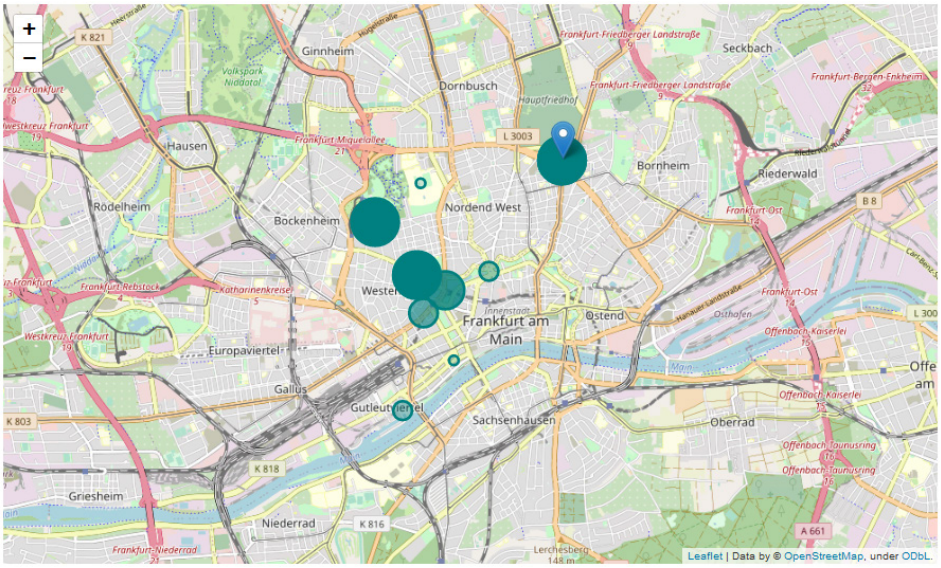


Figure 21: Map of the number of bikes at fixed station from 7AM to 8AM on May 08, 2019.

To further discover this bike using pattern, we utilized the heatmap with time to visualize the end locations of the trips that ended from 7AM to 8AM everyday in May, August, September, October. The heatmaps show that the areas near the University of Applied Sciences were the top popular locations for bike trips to end from 7AM to 8AM in most of the days in May, September, and October. Except for weekends and public holidays, the pattern held true for these three months. In contrast, the heatmap in August shows very few trips ended near the university location from 7AM to 8AM. This could be explained by the fact that August is the summer break time for university students. When the spring semester was ending on August 1st, there were many trips that ended near the university captured. But after that, very few trips ended within the selected time frame.

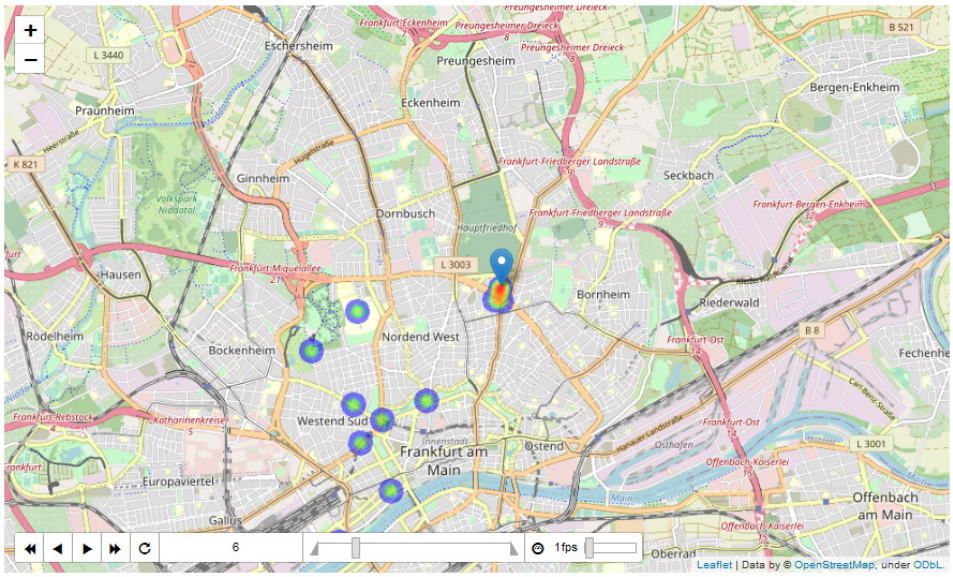


Figure 22: Heatmap of the location of the trips that ended from 7AM to 8AM on May 08, 2019.

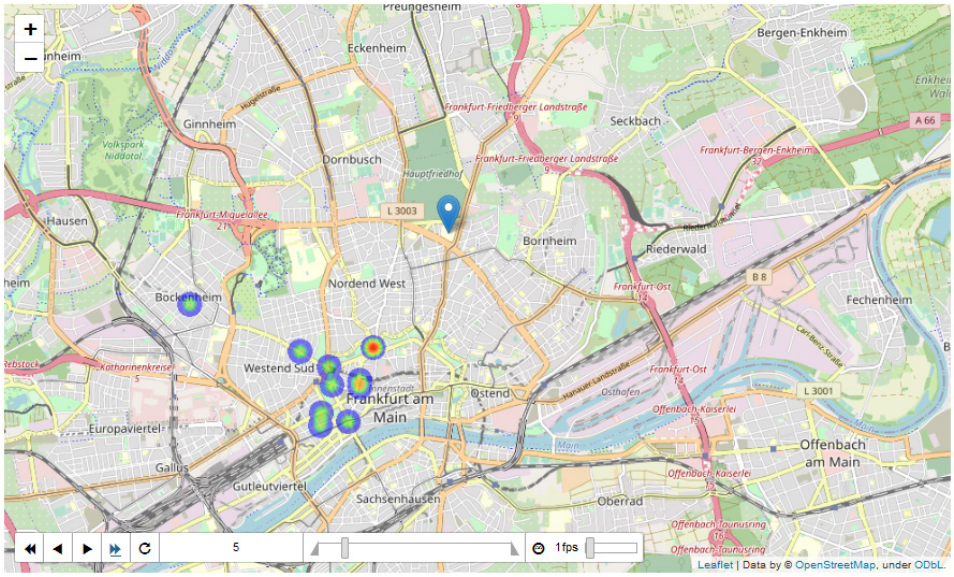


Figure 23: Heatmap of the location of the trips that ended from 7AM to 8AM on August 05, 2019.

5. MODEL DEVELOPMENT

In this project, we developed two models to predict the trip duration and the direction of the trip journey. Based on the preprocessed data set, the findings in the analysis of the underlying statistics, and the business knowledge we gained in this project, we developed the following prodiction and classification models. For the prediction models we also utilized weather data from the DWD and the geodata for the zip codes that were in use in the data visualization.

Feature selection

For feature selection, we followed the following approach. We developed the models and selected the features in parallel. We started with all selected features and trained the model. Then we predicted on the validation set and validated the performance based on the three metrics RMSE, MAE and the R^2. Then we started to reduce the complexity by dropping the "worst" feature and trained again. When the performance went down by a high margin the feature was added again and the next feature would be selected. We continued with that process until the remaining predictors seemed to participate greatly in the success of the model.

5.1 Prediction model for trip duration

For the prediction of the trip duration, we have selected the following predictors as we hypothesized that they would have an impact on the durations of the trips.

5.1.1 Predictors for trip duration

Temperature: Since biking was an outdoor activity, we hypothesized that weather conditions could have certain impacts on bike using behavior, which was the trip duration in this project. We selected the temperature per hour in Celsius degree (°C) as a predictor for the trip duration, because temperature showed a correlation with trip duration in certain time intervals such as in the summer period.

Start Time: According to our descriptive data analysis mentioned above, the trip duration over the day showed a high variation during different day time such as night, morning, evening, and midday. Therefore, we decided to separate the day into 4 time intervals as binary variables and included these variables to the prediction model.

Border_District: Previously as the map visualized, we had many long trips in the border districts of Frankfurt. So we decided to utilize the zip codes of the start positions as a predictor. This Border_District variable is binary and receives the value of 1 if the trip started in a border quarter of Frankfurt, whereas value 0 indicates that the trip started in other

areas.

Last hourly average duration: The weekly analysis showed that trip duration followed a negative trend over the year. After analysing each month on an hourly basis, we could also verify that the average hourly duration also followed a cycle that decreased over time for the year 2019. After eliminating the outliers in the data set based on the median for each hour independently, we hypothesized that the average trip duration of the last 4 hours prior to the trip starts could be seen as an indicator for the trip duration. We set 4 variables with the average trip duration of that hour (e.g variable H1 is the average value of the hour before the trip started).

Last daily average duration: We also utilized the average duration of the last 2 days prior to the trip start because of the presence of a negative trend over the year. For this, 2 variables L1 and L2 with the average trip duration values of the two days before the trip started were used.

Weekday: Our analysis through data visualization captured a difference between trip duration on weekdays and trip duration on weekends. Since bikes could be used both for commuting to work and for daily activities, we hypothesized that day of the week could be a predictor for trip duration. Therefore, the Weekday variable was used as a binary predictor under the assumption that the trips over the weekend were longer on average than trips over the weekdays.

5.1.2 Prediction method for trip duration

For the trip duration prediction we developed a linear regression model and used a log transformation for the target value to issue the right skewed distribution.

5.1.3 Evaluation

We used the following metrics to evaluate our models' performance:

Mean-absolute-error (MAE): The mean-absolute-error is the absolute deviation from the real prediction on average. That clearly stated how many minutes our model missed in the context of the trip duration on average.

Root-mean-square-error (RMSE): Additionally, we will track the root-mean-square-error. The root-mean-square-error will be analyzed and used as a loss metric to see in which direction our prediction went. This is a factor to measure how much the model fits the data and is very sensitive for big outliers.

R square: Lastly, we will use the R^2, which is defined as the margin of the variability that is explained by the model. It is a good indicator for the degree of generalizability of the model.

As mentioned before, we developed a linear regression model with a log transformation of the target value to handle the nonlinear relationship of the duration towards our predictors. After, we trained our model with the training data and then created predictions for the validation set. We measured the performance based on the R^2, MAE, and the RMSE. The validation set showed the following results:

RMSE: 21 minutes

MAE: 13 minutes

The MAE of 13 minutes indicates that our predictions are on average 13 minutes off the real values. This shows that we have had a systematical error (bias) in our model.

Also, the RMSE of 21 minutes is magnificent higher than the MAE. This indicates that our model had many predictions that were very far away from the real values, because they get a higher weight in the RMSE. This is also shown in figure 24 the wide spread of the value points. In addition, the R^2 of 0.051 shows that the model did not predict the duration accurately. This again indicates a systematic error (bias) in the model and we assume that the model is underfitted. This also indicates that our predictors are not good for predicting the trip duration.

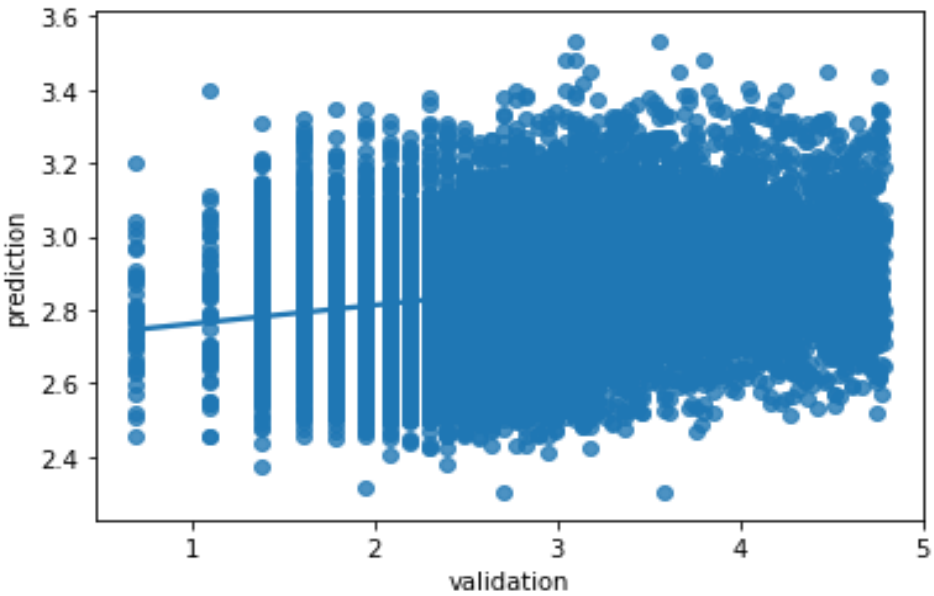


Figure 24: Scatter plot of performance of prediction model (validation set)

We will now verify that by predicting the test set duration based on the trained model. After predicting the values we got an RMSE of 15.91 and a MAE of 15.7, which indicate that our model is not performing well generally. Since the MAE and the RMSE are close to each other it seems that the error is quite constant over the test set. That is clear advice that our model has a bias. That is also proved by the decrease of our models' R^2 from 0,051 to 0,008. The scatter plot of the test set shows that the spread of prediction and the real values

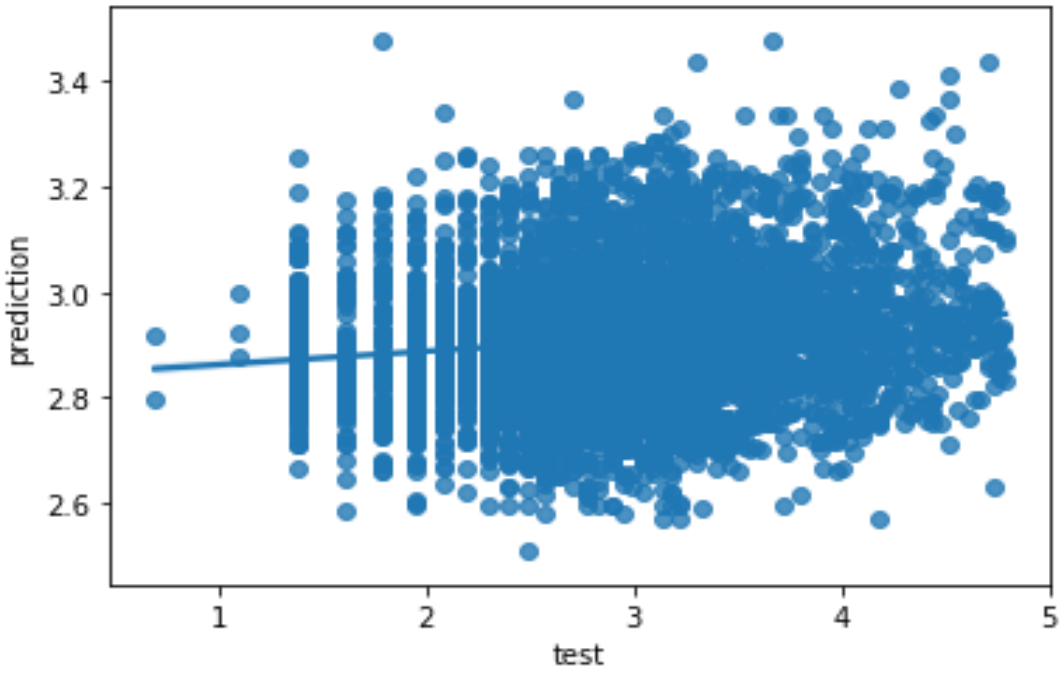


Figure 25: Scatter plot of performance of prediction model (test set)

has decreased on average (lower RMSE), whereas the MAE has increased, which means our prediction had less extreme deviation, but we got a worse result on average. That also strengthens the assumption that our model was not capturing the underlying relationship in the data. That can also be seen throughout the line in the scatter plot.

In summary, it can be seen as proof that the model is not predicting the duration of trips accurately. We assume that this is the result of an underfitting model and could also be a consequence from the drop of the high values and the small data set we used to train the regression. Therefore, the linear regression with even a log transformation, was not the right choice to predict trip durations. To improve the model, it would be useful to analyze other predictors and select another machine learning algorithm that would fit the nonliner prediction better.

5.2 Classification model for trip direction

For the second prediction task, we constructed a model to classify if a trip travelled towards the university or not. We selected the University of Applied Sciences for this task. The following details the predictors included in the model to forecast the direction of the trips.

5.2.1 Predictors for trip trip direction

Border_District: Our previous data visualization showed that trips started in the border districts of Frankfurt had longer average trip duration compared to trips started near the city center. We assumed that as people in the border districts would have to commute to work in the city center, they would possibly take bikes towards the center. Even though the university is not centered, it is more centered than the border districts. Thus, we assumed there could be trips towards the university from people commuting to work.

Student_quarter: We assumed that students were among the major groups of people travelling to the university. Therefore, we hypothesized that trips started in areas where more students lived could have a higher chance that these trips would travel towards the university. We created the feature Student_quarter as a binary predictor, which received the value of 1 if the zip code of the starting position had a student dormitory.

Distance_uni: This variable measured the distance between the starting position of a trip and the university. In addition, we assumed that trips started at locations near the university could have more chances to go towards the university, because it would be very likely that students tended to stay close to school. In this project, we selected the University of Applied Sciences as the university to study. The Distance_uni predictor took the numerical values in kilometers.

Weekday: The Weekday binary variable was used as a predictor under the assumption that there would be more trips towards the university over the weekdays than weekends.

Weather: Similar to what we presented earlier, biking is an outdoor activity which could be influenced by weather conditions. Thus, we hypothesized that weather conditions could influence the possibility of students using bikes to commute to the university. For this task, we employed a binary predictor Weather, which received the value of 1 for good weather and the value of 0 for bad weather. Temperatures above 5°C and the rainfall below 0,1 mm per hour were the indicators for good weather.

Zipcode_pro: We used the past information of the zip codes by counting the number of trips that started in the same zip code zone and went towards the university. This number was divided by the number of all trips that started in that zip code area. We used that as reference to the mean encoding feature engineering method with only using data from the past.

Station_pro: We employed the station numbers in the similar way we utilized zip codes mentioned above to generate a probability with reference to the past trips' start stations.

5.2.2 Classification method for trip direction

Our first step in developing the neuronal network was to decided which hyperparameters to use.

For the neuronal network, we had to select a optimizer, the number of nodes, layers, and the learning rate. Therefore, we started with the simplest neuronal network with one hidden layer, one input layer with one node for each predictor and one output layer for the target. We advanced the neuronal network to a 3-hidden-layer network with each layer containing 7 nodes with a drop out rate of 0.1 to avoid overfitting, a sigmoid activation function, and an Adam optimizer with a learning rate of 0.01.

5.2.3 Evaluation

We developed a neuronal network with the earlier mentioned hyperparameters. For that we used the loss function of our neuronal network to tune our hyperparameter in order to get the best performance out of it. For that purpose, we adjusted the epochs and the batch size of our neuronal network. In order to do that, we started with an epoch of 10 and a batchsize of 500 and stepwise increased them to the point, where our validation set and train set loss were in a certain proportion. That means that it is not the case that the validation loss is not much higher than the training loss (overfitting) or that the training loss is not much higher than the validation loss (underfitting). With that method, we decided to use 20 epochs with a batch size of 4000. Our loss metric was the RMSE to be able to track outliers. Additionally, we tracked the MAE. Since the drop outs prevented us from overfitting, it was not useful with our predictor set to continue the increase of the epochs or the batchsize. Even when the loss was still decreasing, it was very likely that our model would overfit at a certain point.

After we trained the neuronal network with the selected hyperparameters, we predicted the test set. For the evaluation of the test set, we utilized the confusion matrix and the accuracy rate. The confusion matrix contains 4 fields “True/Positive”, “True/Negative”, “False/Positive” and “False/Negative”.

In the validation set, we got a result of 5208 “True/Positive” classifications, 2706 “False/Negative” classification, 3511 “True/Negative” and 3993 “False/Positive” classification. This leads to a classification accuracy of 0.52 ~ 52 %, which is not a good rate. The matrix for our test set has the following values. It predicted 2326 of the trips right as towards the university, whereas it correctly misclassified 1166 trips, 917 trips it predicted wrong as towards the university and 988 trips that went towards the university the model misclassified. Under the fact the model classified 3492 trips out of 5397 correctly and 1905 wrong. That states an accuracy of 0.647 ~65 %, which is a much better value for a classifier in comparison to the validation set rate.

Therefore, the validation loss is lower than the test set's loss. This could be seen as an indicator that the neuronal network could develop a sophisticated model for the trip direction. However, we must take into consideration when working with neuronal network is that neuronal networks are not models that build on assumptions that are developed on knowledge. They are blackboxes that could easily overfit. Furthermore, the small test set (5400 trips) could not be seen as sophisticated enough to assume that the model has no bias. To ensure that this is not the case, a cross validation test would be a good solution to verify that model has no bias and not overfit. In summary, we can conclude that the test result is not adviced to generalization.

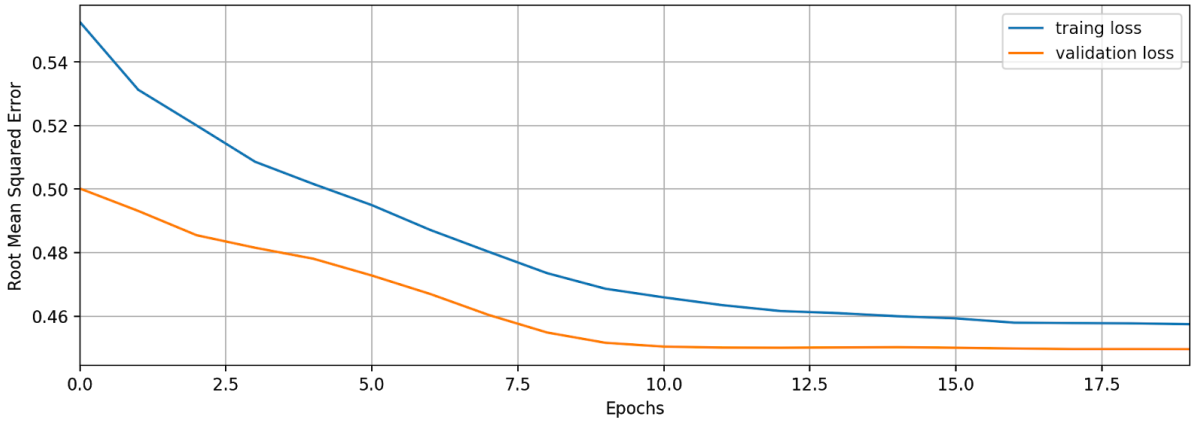


Figure 25: The loss and validation lost of the models with their number of epochs

Confusion matrix	Validation Set	Test Set
True/Positive	5208	2326
True/Negative	3511	1166
False/Positive	3993	917
False/Negative	2706	988
Classification Accuracy	0.52% ~ 52%	0.647% ~ 65%

Table 3: The loss and validation lost of the models with their number of epochs

6. CONCLUSION AND DISCUSSION

About our models

The results of high error values, low R^2, and decreasing accuracy rate with increasing samples show that our models did not perform well and can not be taken as valid predictive models. In general, the models we created did not cover the underlying data structure for the trip duration and the most parts for the journey direction. On the one hand, the unfitting predictors we developed led to the failure of the models, on the other hand, such failure can also be caused by the poor quality of the data, which contains many outliers, noises and measurement errors. Removing those bad data shrunk the data set to a critical size (around 51,000 entries). Therefore, the identification of outliers is even more important for upcoming projects that analyze these target values. Further exploration of neuronal networks might also be worthy to be done for future work.

Unfortunately, our hypothesis on the impact of temperature on trip duration was not supported in this project. We found that temperatures did not correlate linearly to the duration of the trips. In fact, within a certain range of temperature (from 5 degrees to less than 25 degrees), there were significantly more trips than other temperature range. In the 5-25 temperature range, trip durations fluctuated sharply among the trips. This could be explained by the fact during comfortable weather conditions, many occasional bike users started using bikes with various time usage, whereas during less comfortable conditions, more regular bike users commuted using bikes. As regular bike users might have more stable time usage than that of occasional bike users, this diversity in behavior might have caused the inaccurate prediction of our model.

Also the past values of the duration under the assumption that the duration is a trended series, are also shown as not usable for our model. Since the model shows a very low R^2, the past averages did not help to predict the duration values. Therefore, the underlying distribution of the duration even after log transform seems not to be normal distributed.

Further, we could see that the weekly cycles and the trend over the year show that the duration behavior differs greatly in each month. We indicate that the data is dominated by routine bike trips that are most likely the way to work or to the university. This assumption is also strengthened by the fact that the zip code proportion of trips towards the university seems stable over the whole year, except the lecture free time when the number of trips towards the university goes down.

For future projects

For future analysis of the trip duration, we can conclude that the prediction in the manner of mean encoding with a greater train and test set will promise good prediction. If the majority of the trips, especially their duration, are dominated by routine trips, the encoding of features like zip code, station number, hour or weekdays would help to predict the trip duration and the journey direction especially. Furthermore, this type of feature engineering would only be applicable in the time series forecasting, where the scope is about training based on data of the previous year to predict the duration of the present one, since the month and weeks show that big difference in distribution. Also, a potential approach to develop a prediction in the future would be the analysis of events nearby a specific location. We could not use that feature in this project because the city had no such events for the university, the events that could have a huge impact on the trips towards the university.

REFERENCE

McLeod, K. (2014). New: Census Data on Bike Commuting. [online] League of American Bicyclists. Available at: https://bikeleague.org/content/new-census-data-bike-commuting

APPENDIX

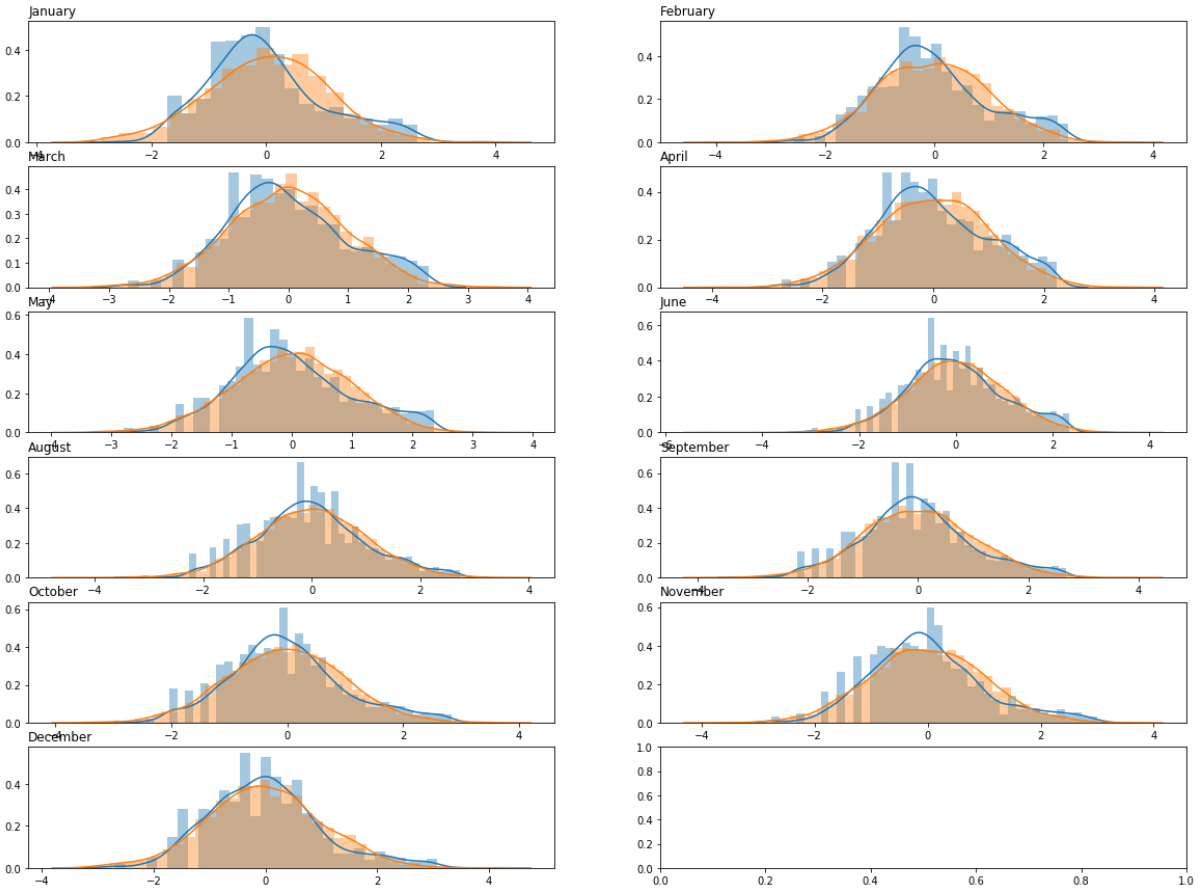
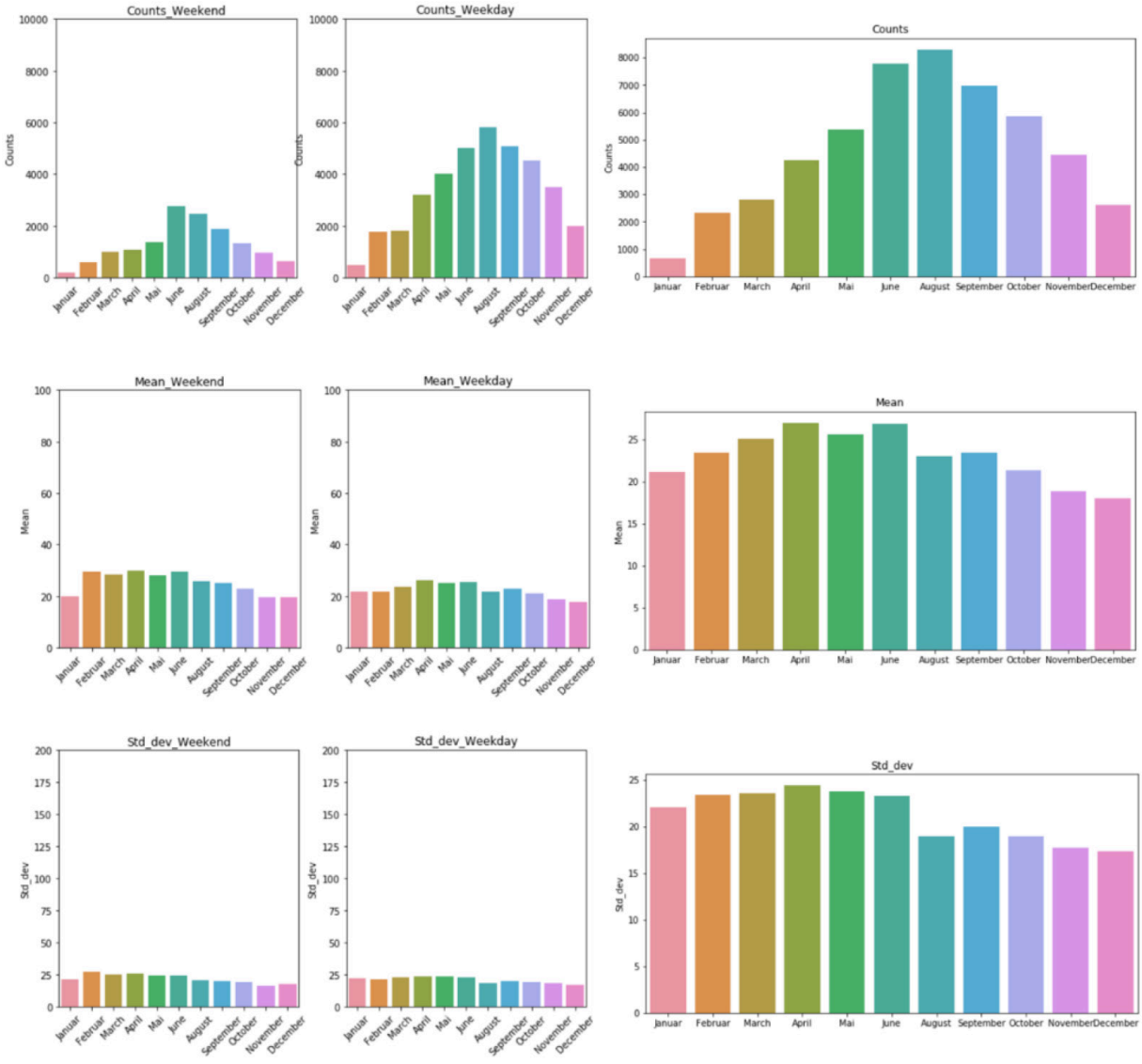


Figure A1: Monthly distribution and normal distribution of the trip duration

Blue columns: Distribution of the duration by month

Yellow columns: Normal distribution



[Figure 3] Monthly analysis

Figure A2: Monthly analysis with comparision between weekend and weekday of the number of trips, average trip duration, and standard deviation

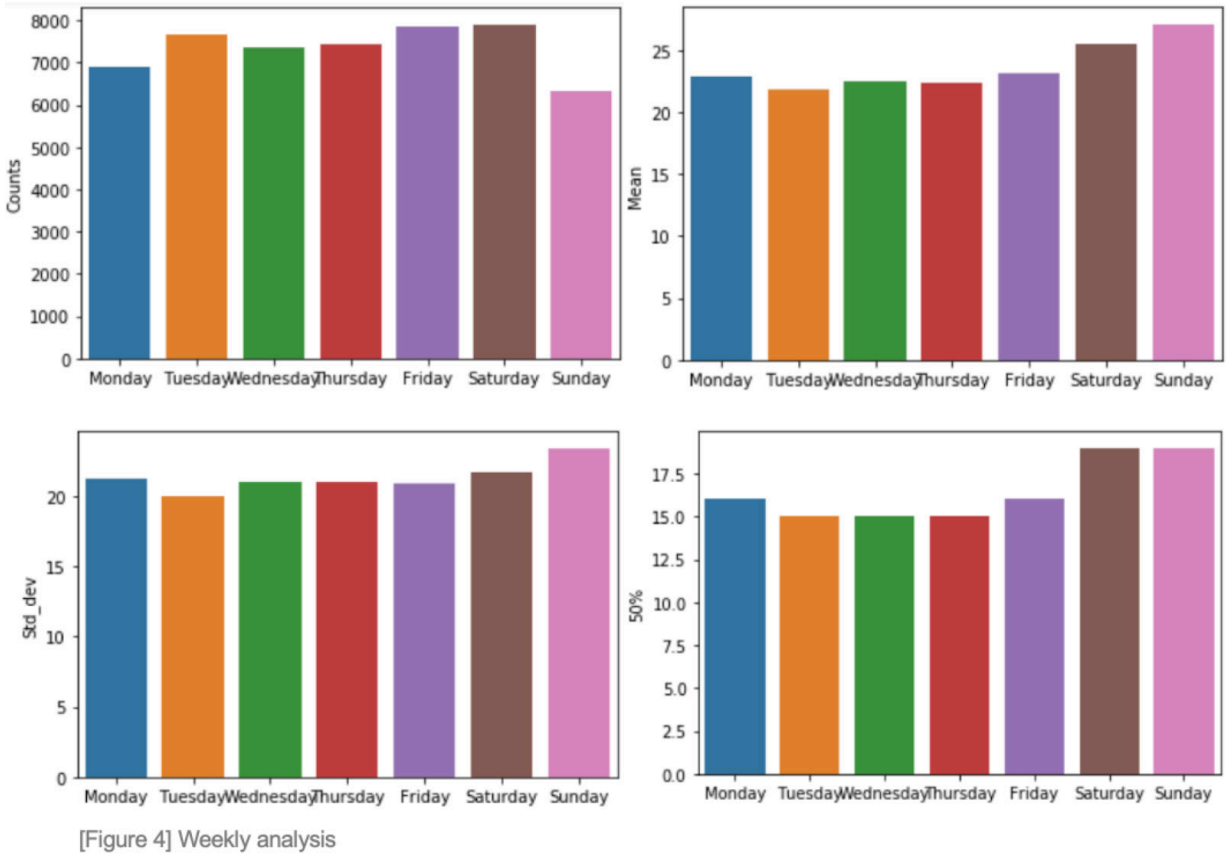


Figure A3: Weekly analysis of the number of trips and standard deviation

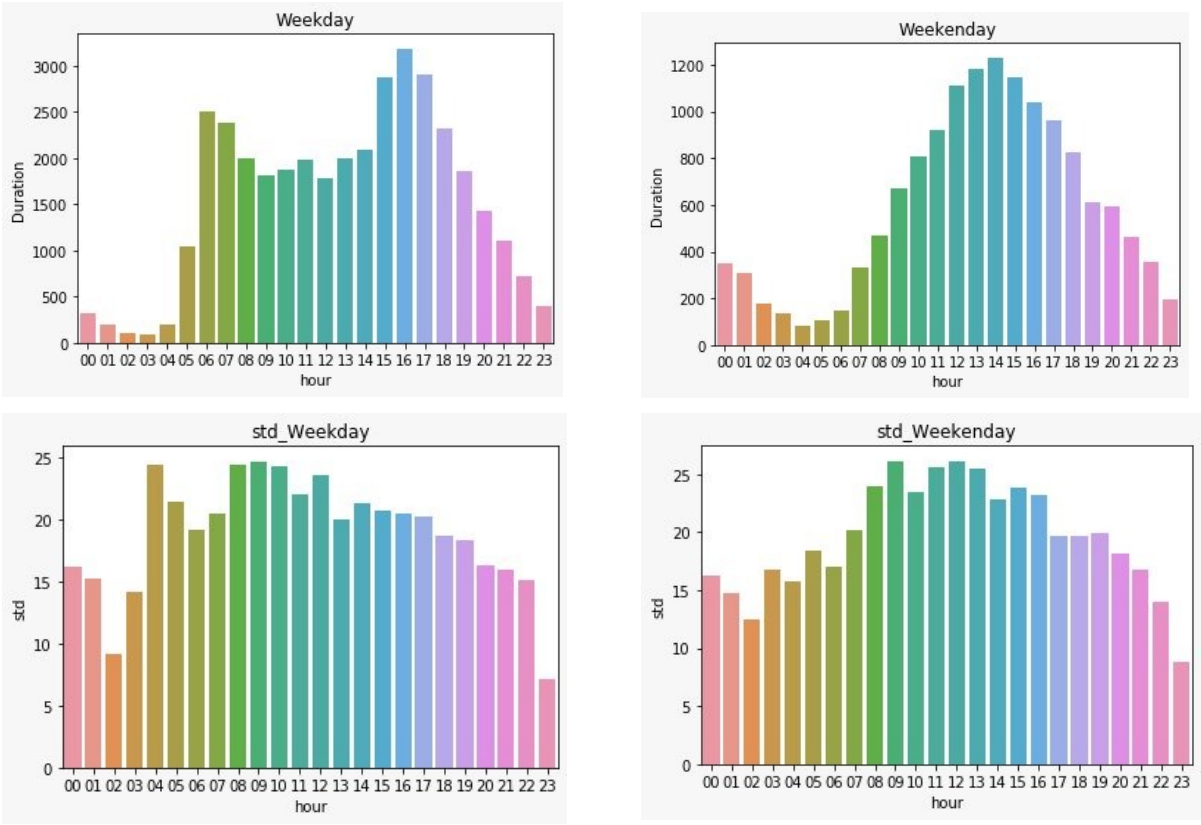


Figure A4.1: Hourly analysis of the number of trips and standard deviation (of weekend and weekdays)

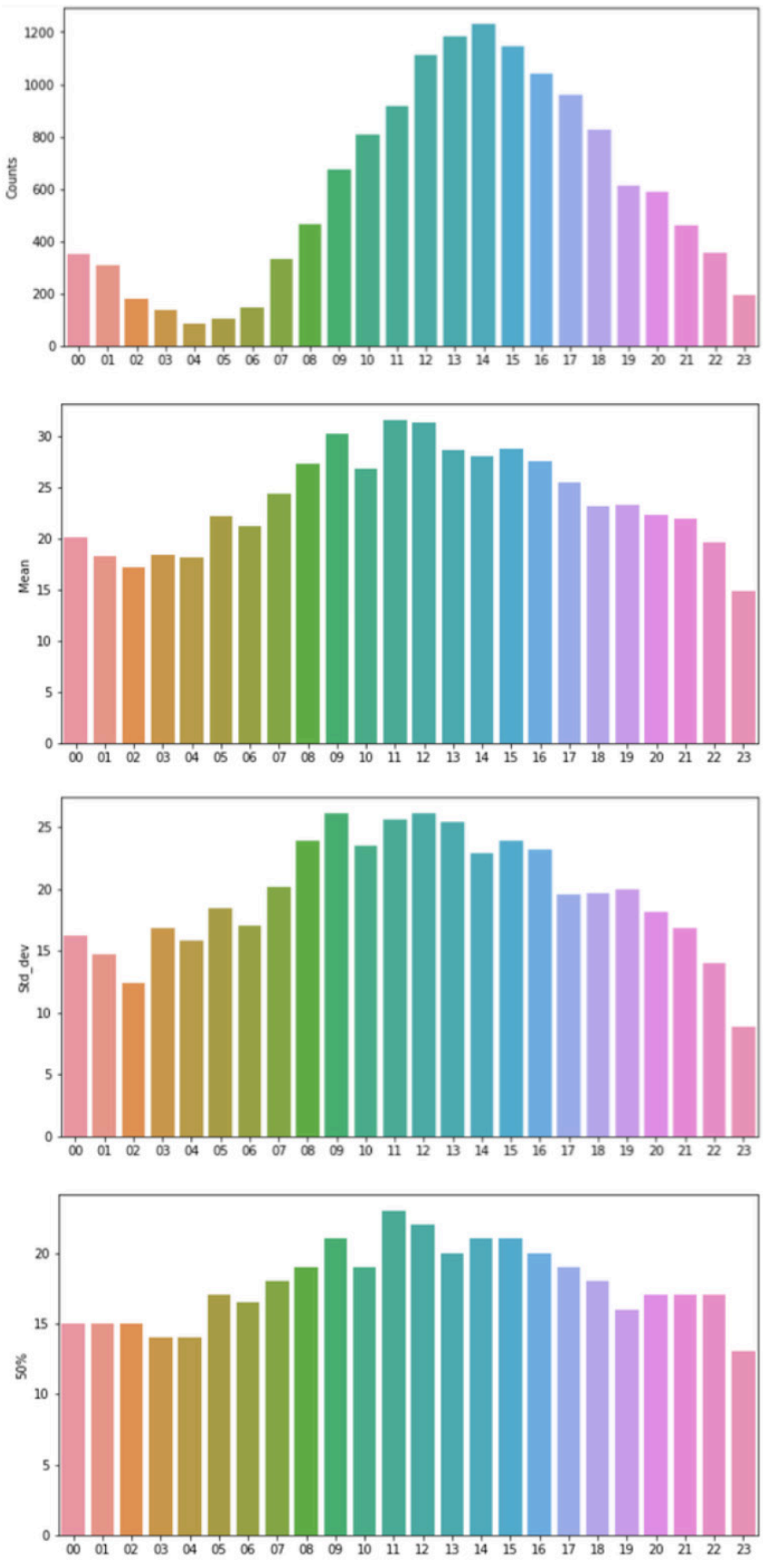


Figure A4.2: Hourly analysis of the number of trips, average duration, and standard deviation