Blue Bikes Data Science Project

Team No. 3: Error 404: Group not found

Authors:

Robin Kirch (Student ID: 7364580)
Lukas Tempfli (Student ID: 7367097)
Moritz Danhausen (Student ID: 7369413)
Niklas Nesseler (Student ID: 7367375)
Sven Dornbrach (Student ID: 7364484)

Supervisor:

Univ.-Prof.Dr.Wolfgang Ketter

Co-Supervisor:

Karsten Schroer

21.07.2021

Department of Information Systems for Sustainable Society Faculty of Management, Economics and Social Sciences

1. List of Content

2. Executive Summary

2.1. Introduction

This report is based on real world data, which was provided by Blue Bikes, which is an operating bike sharing firm from Boston. It focuses on the task to predict the total bike usage of bikes from “Blue Bikes” for the next hour. For this we were provided with data of the year 2017, which we transform to our needs and afterwards apply modern machine learning techniques to solve the task.

2.2. Business Problem

In the business context of smart mobility services it is relevant to precisely predict the total bike usage for the next hour, as this marks a high value process, because it can minimize the cost of assets and increase the total value of the operating workflow. This is eminent, because we want to prevent to have too many bikes as this would result in a not valuable utilization. Otherwise, we are spending investment money, which is not really required for the satisfaction of the operational task. However, the opposite case is even worse, as this means we are lacking bikes and are not fully utilizing the market potential. This also will potentially lead in bad reputation, as customers will be frustrated, if they do not get a bike quickly. Our target is to predict the demand of total bike usage for the next hour to check how many bikes we must keep in store to satisfy the customer needs. As a result good demand prediction results in a better customer service and therefore in higher profit. Furthermore, we can monitor the system to evaluate our performance. This is relevant as our business decisions are highly connected to the evaluation of the current system performance. In general, it is of high interest to build a good reputation for smart mobility services as more people would switch to use them, which will reduce typical mobility and societal issues that a big city has such as emission, pollution, traffic jams and road accidents.

2.3. Data

The raw_data_boston data set provides information about all trips which were made in 2017 with a “Blue Bikes” which includes relevant attributes like start time, end time and station names. This data was collected by “Blue Bikes” themselves and will further show how important ubiquitous real time data is. The second data set delivers weather information which is relevant to check things like temperature and precip which have an impact on the amount of bike rentals and if so, how big this impact is. The third data set provides geolocation of all docking stations, which will be used to visualize where the operating docking stations are.

2.4. Implications

Implications we faced were false data tuples. We found some in the data set with negative duration or duration over a long period of time. Because these were clearly no valid data tuples, we eliminated them. We did filter the data so that every trip needs to be at least two minutes long and eight hours at its maximum.

2.5. Recommendations

We further recommend increasing the size of our bike fleet to 2000 bikes, so we ensure that there are always enough bikes on each station. The explanation for this is provided in detail at the end of the report.

3. Data Collection and Preparation

Our data preparation for all three data sets focuses on four things. Firstly we check for null values, secondly we check for duplicates, furthermore we create new attributes in regard of our needs and lastly, in term of feature selection, we drop features we do not need for our analysis.

3.1. Bike Dataset

The bike data set provides us with operational raw data from “Blue Bikes” with the following attributes: start_time, end_time, start_station_id, end_station_id, start_station_name, end_station_name, bike_id, user_type. This dataset has around 1.3 million entries and does not contain any null values. We added an “hour” attribute, which takes the hours from the start_time and is used later to analyse the number of rentals per hour. The same applies to the date_time column, which contains the date including the full hour. A further attribute “duration” was added for later processing and plausibility checks of the start and end times, which contains the rental period of the bicycles. We discovered that the raw data set contained partially implausible data, as the shortest rental period was -1 day 23:06:07 and the longest rental period in the data set was over 48 days. This made it necessary to consider what time period should be judged as realistic and thus retained. Clearly negative durations mark incorrect data. The same applies to rental periods lasting several days. We also considered it useful to filter out rentals that were too short, for which a lower limit of 2 minutes was set, since a shorter period of time can be assumed that it will probably be an accidentally made or aborted rental. In order to delete as little data as possible, an upper limit of 8 hours was chosen, since such a rental period could possibly still be explained by longer day trips. However, multi-day rentals are extremely unrealistic, and it is therefore assumed that this is incorrect data. In addition, these could also have been problematic for further consideration. After cleaning up the dataset, 1,302,372 evaluable entries remained.

3.2. Weather Dataset

The weather data ranges from the first January of 2015 to the second of January of 2020. It contains the following attributes: date_time, max_temp, min_temp, precip. We filtered the data set to the time period of the year 2017. It does not contain null values for individual attributes. However, from the tuple count it can be seen that it only contains 8,667 tuples, although it should be 8,760 tuples to map a year without gaps. In addition to that we have we found 132 duplicate entries. That leaves us with 8,535 valid entries. We considered adding the missing 225 hours and filling it with the same value of the last available data for the same hour. However we opted to not do so as we still have 97.4% of all hours in the year, which marks a representative data set. Additionally these are not connected they should be unproblematic for further consideration. For this reason and since the weather can usually be very variable, the addition of all tuples was omitted. The temperature information ranges between -16°C and 35°C. This is quite realistic and therefore does not require any adjustment. The same applies to the dummy variables “precip_id”.

3.3. Station Location Dataset

We also use further information which was provided by the “Blue Bike” Website (378 docks), which is a data set containing geographic information about the docking station. This will be used for the visualization to get a feeling where the docking stations are in Boston. The data set contains following attributes: number, name, latitude, longitude, district, public, total docks. It needed to be pre-processed in multiple ways.

4. Descriptive Analysis

4.1. Temporal Demand and Seasonality

If we look at the temporal distribution of the number of rentals during a day, two clear peaks can be seen. The first is in the morning between approx. 6-10 a.m. and the second is in the evening between approx. 3 p.m.-7 p.m. (cf. Figure 3). This means that they have a clear correspondence with the typical rush hour traffic, which is why it can be assumed that the rentals were mainly due to residents of the city of Boston using the Blue Bikes for commuting to work or to university. Furthermore, a somewhat higher demand can generally be observed in the evening compared to the morning, which is reflected in a somewhat higher peak and the later slower decrease in rental numbers. This can possibly be explained by the fact that in the evening, in addition to rush hour traffic, bicycles are also used for leisure activities. A moderate need can be seen during lunchtime. Demand continues to decrease during the night and remains low until 5 a.m. In particular, the assumption that the bicycles are used for trips to work or for trips to university can also be substantiated with regard to the distribution over the week. This shows that a high demand can already be observed on Mondays, it reaches its peak in the middle of the week and levels off sharply on Friday (cf. Figure 4). There is lower demand on the weekend days. With regard to the seasonal fluctuations, the development over the course of the year is very meaningful. It turns out that there is a low demand in the winter months, which increases significantly in the spring. A sharp rise in rents can be seen in April. Subsequently, the demand remains at a high level in the summer months and reaches its peak in August. After that the demand steadily decreases over autumn, only to return to the minimum level in December (cf. Figure 5). This can be explained by the prevailing weather conditions and the temperature in the respective seasons (cf. Figure 6). The predominantly warm, sunny weather in summer will encourage people to see bicycles as an alternative to public transport or the car. In winter and the adjacent months, the weather is usually changeable, which is why other modes of transport are preferred to cycling. This is also underlined by the visualization of the loans made at certain temperatures. It turns out that most loans are made at a mild temperature range of 15-25 ° C (cf. Figure 29). This correlation fits our assumptions very well, but only becomes meaningful if the frequencies of the temperatures that have occurred have also been checked. Turns out that both low temperatures (between 0-10°C) and temperatures around 20°C occur more frequently (cf. Figure 7). However, since no increased demand can be identified at low temperatures, it can be assumed that the statement about higher demand in summer is correct.

4.2. Geographical Demand

In the following we consider a few particularities with regard to the geographical distribution of the use of Blue Bikes and the changes that occur due to different daytimes. It is noticeable that the most popular stations are in close proximity to the universities. The stations on or near the Massachusetts

Institute of Technology campus have 19,000-42,000 annual rents. Those at "Harvard University" are between 7000-22000 annual loans (cf. Figure 13). This also confirms our assumption that students in particular are among the users of the rental bikes. It also shows that there is also a high demand at the various train stations of the city. Around 25,000 bicycles were rented both at the “South Station” and “Central Station”. In general, it can be said that bicycles are particularly popular in the city centre and the neighbouring areas, but their use at the stations outside of the city is rather low.
If you look at the heat maps, which indicate the popularity of the destinations in the morning and in the evening, it is noticeable that the areas around the universities are mainly used in the morning (cf. Figure 11). In the evening there is a shift in the most frequent destinations (cf. Figure 10). The use of the stations on the Charles River, especially those at Massachusetts Institute of Technology, is decreasing somewhat, but remains at a high level. This can be explained on the one hand by means of evening events at the university and on the other hand by personal evening arrangements in the area of the river and the parks.
Furthermore, there is an increased use of the station on Nashua Street in Boston Westend in the evening. On the one hand, there is the PD Garden, a large multifunctional hall in which NHL and NBA events take place and thus attract numerous spectators. On the other hand the North Station, another train station, where bicycles are a great alternative for transportation within the city. In addition, the entire Westend seems to be more frequented than in the morning, this could be due to the fact that it is a mixed residential / commercial area with restaurants and bars which are visited by the residents. In addition, the relocation of the end stations used can be explained by the fact that people return to their residential areas in the evening.

4.3. Key Performance Indicators

We selected three KPIs. Our first KPI is the utilization ratio of the bicycles. We made various graphs that illustrate hourly values over the day and over the year. At peak demand we have a utilization ratio of approximately 40% (cf. Figure 14). We can see a trend that there peak utilization during the rush hours (cf. Figure 15).With this information managers can decide to increase or to decrease the fleet size.

Our second KPI is the break-even point ratio. It measures the percentage of hourly rented bikes to hit the break-even point. To compute this ratio we divide the hourly rented bikes through the number of bikes to make profit. Due to the fact we do not know this number we used the average number of rented bikes hourly instead. Due to analysis results the company makes the most profit at summer during the peak times for work in the morning and at the afternoon (cf. Figure 17). This KPI helps the management to illustrate if they were enough rentals in the last hour to make profit.

The last KPI we selected is the subscriber customer ratio. This KPI shows the percentage of bikes rented by a subscriber or a customer last hour. The analysis shows that the majority of bikes were booked by subscribers which can be explained the bikes are mostly used for the way to work or to the university (cf. Figure 19). This ratio can be useful for the marketing department of a company to adjust the advertisement for example.

5. Predictive Analysis

5.1. Feature Engineering

Based on our results from the descriptive analytics we decided to choose these four features for our prediction models.

Max_temp feature: We observed on the Rentals/Temperature graphics that on warm days people are more likely to take a bike as on cold days (cf. Figure 29). Furthermore, on very warm days the demand

goes down. In conclusion the demand is dependent on the temperature what means that this is a suitable feature for our prediction.

IsWeekday-feature: The Rentals/Weekday graphic shows a big difference between total demand on weekdays comparing to the demand on weekends (cf. Figure 4). On weekends is a much lower demand as on weekdays so we decided that this is an important feature to include in our prediction.

Precip feature: We observed on the „Demand in dependency of the weather“ part that on rainy/ snowy hours the demand is very low and on days with no rain/snow the demand is very high (cf. Figure 2). That means that the weather has a big impact on the demand and is a suitable feature. Hour feature: On nights the demand is logically lower than on days and on rush hour times before and after work the demand is high. This can be proven with the Hour/ Rentals graphic (cf. Figure 3).

First we created an IsWeekday feature in our data frame from the weekday column and dropped duplicates. After that we wanted to get a data frame including all of our features. To achieve that we did a left merge on the weather data and removed the unused features.

5.2. Model Building

For the demand prediction we decided to choose the following three models:

KNN Regression: We started with this regression because it is one of the simplest regressions. Furthermore, it is a non-parametric algorithm, what means it does not make strong assumptions about the form of the mapping function. This includes that KNN is free to learn any functional form the training data. Drawbacks are that the features for this regression need to be scaled before using this algorithm and KNN is sensitive to noise in the dataset which means that you have to delete null values and outliers.

Polynomial Regression: We continued with a Polynomial regression because it provides the best approximation of the relationship between the dependent and independent variables. In addition, a broad range of function can fit under it. It basically fits a wide range of curvature. Drawbacks are that the presence of one or two outliers in the data can affect the results of the nonlinear analysis and we need to put focus on the best degree of the regression.

Tree based Regression: At the end we also wanted to include a model that is easy to understand. Furthermore, data preparation during pre-processing requires less effort and does not require normalization of the data. Another advantage is that tree based regressions are not largely influenced by outliers or missing values, and it can handle both numerical and categorical variables. Drawbacks are that they are relatively expensive as the amount of time taken and the complexity levels are greater and small changes in the data tends to cause a big difference in the tree structure, which causes instability.

5.3. Model Evaluation

Model evaluation: The worst model result was from the Polynomial regression with a mean absolute error of approximately 64.0 bikes and a root mean squared error of approximately 110.0 bikes (cf. Figure 26). Second best model is the Tree based Regression with a mean absolute error of 49.94 bikes and a root mean squared error of approximately 88.0 bikes (cf. Figure 27). Our best model is the KNN Regression with a mean absolute error of approximately 46.0 bikes and a root mean squared error of approximately 80.0 bikes (cf. Figure 25). This is also the model we would select for deployment because it has the lowest error metrics.

5.4. Outlook

We could improve our model by including more suitable features. In addition, we could look for more or new data to train the model better. It is also an option to try some more models and look if they perform better. Moreover it is considerable that the dummy we use for weekend/weekday worsen our prediction in comparison to an input feature that differs between all weekdays.

6. Conclusions and Recommendations

We further recommend increasing the size of our bike fleet to 2000 bikes, so we ensure that there are always enough bikes on each station. Currently we are operating with 1799 bikes and have a maximum of rented bikes at the same hour of 1564. There are three things we need to keep in mind because we should have a higher number of bikes than the mean of the highest operating hours. The first thing is that there should always be 2 bikes on hold at every station, so a pair of customers has always the chance to rent the bikes at any station. If there are not enough bikes at a station a warning is triggered, and bikes are taken of the place where there are too many and transported to the location which has to little bikes left in the respective location. Besides, it is also good for marketing purposes if we always have bikes available at each station as the see our “Blue Bikes”. Secondly there is always a part of our fleet size which is getting repaired and therefore out of service. The third reason is that every year more and more people are using mobility service systems. The fourth reason is that it is not the biggest investment to buy more bikes (The higher investment would be to set up stations). Moreover, it is good to provide enough bikes on bike events as “Blue Bikes” for advertisement reasons (keep vandalism in mind – concert etc.).

7. Supplementary document

Task 1:

Lukas Tempfli: Data preparation
Moritz Danhausen: Data preparation
Robin Kirch: Data preparation
Sven Dornbrach: Data cleaning

Task 2:

Lukas Tempfli: KPI's and visualisations
Moritz Danhausen: Visualisations and descriptions
Robin Kirch: Visualisations
Sven Dornbrach: Heatmaps and trip visualizations

Task 3:

Lukas Tempfli: Features and visualizations
Robin Kirch: Regression models and visualizations

Additional Tasks:

Moritz Danhausen: Executive summary
Niklas Nesseler: Git Hub Repository

Everyone wrote the text segment to the tasks he worked on.

One team member left our group very early which means that only five team members contributed to the report.

Link to the git-Repository:

https://github.com/NiklasNesseler/ProjectNotFound

8. Appendix

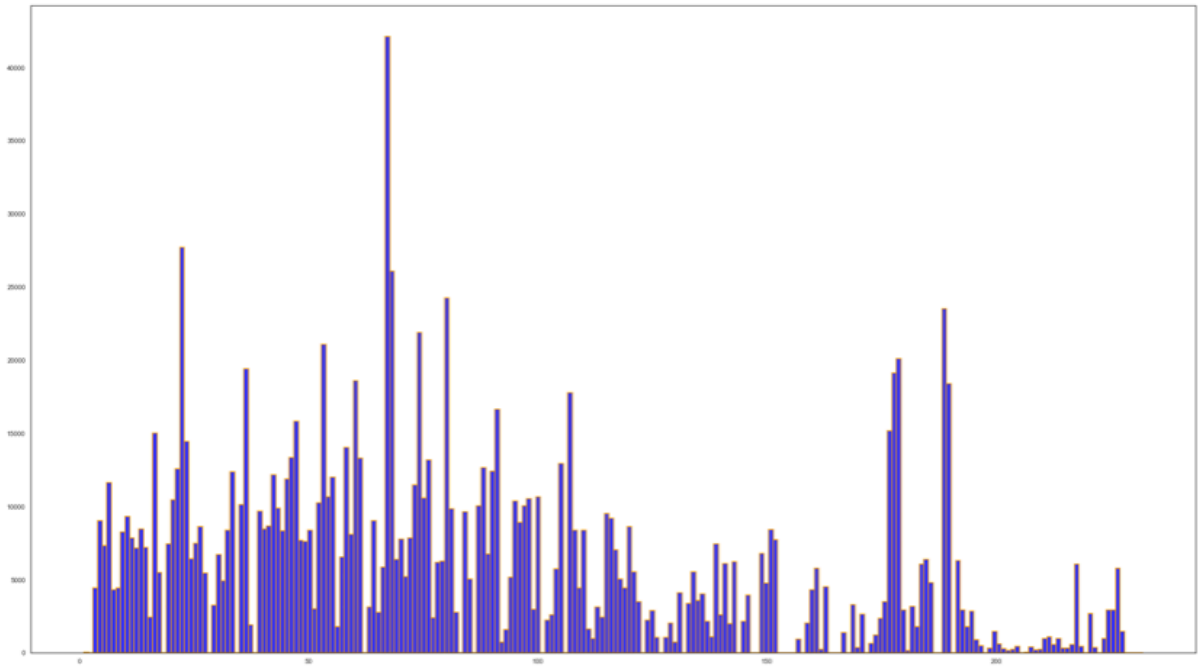

Figure 1

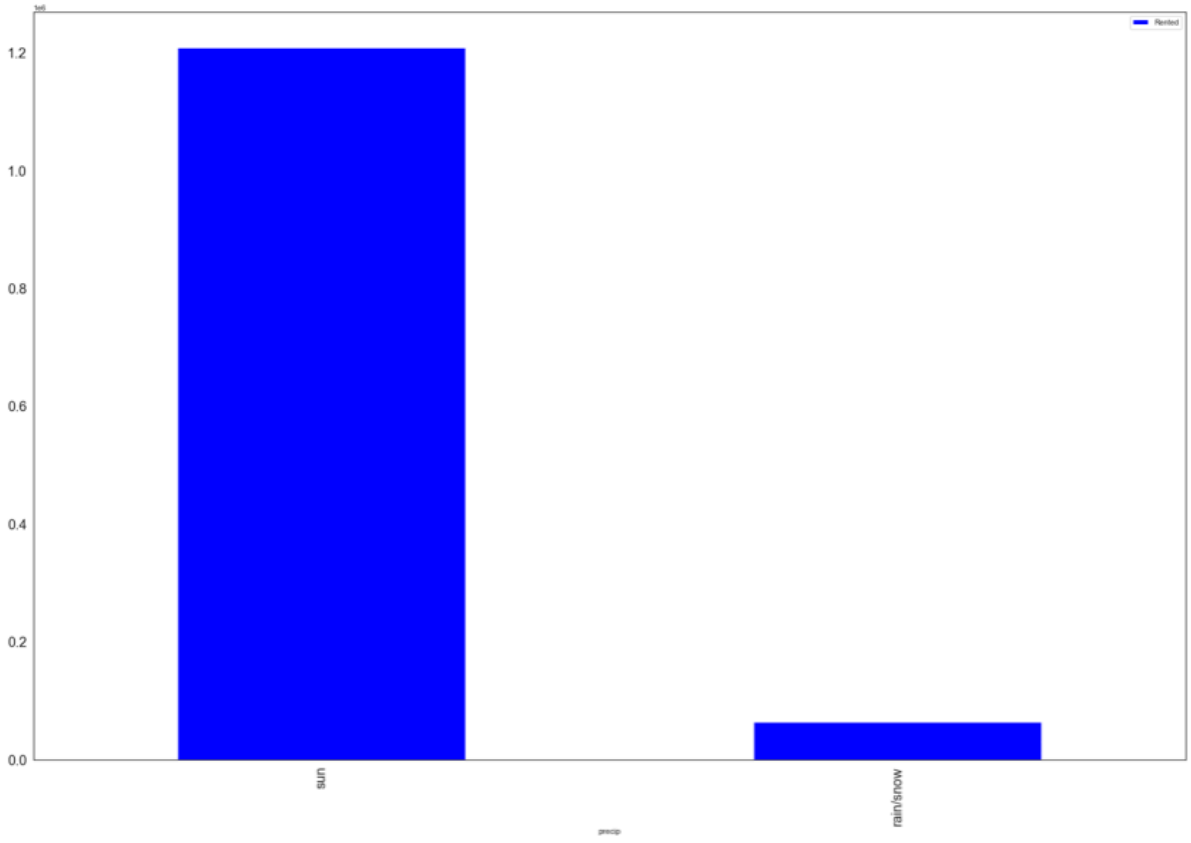

Figure 2

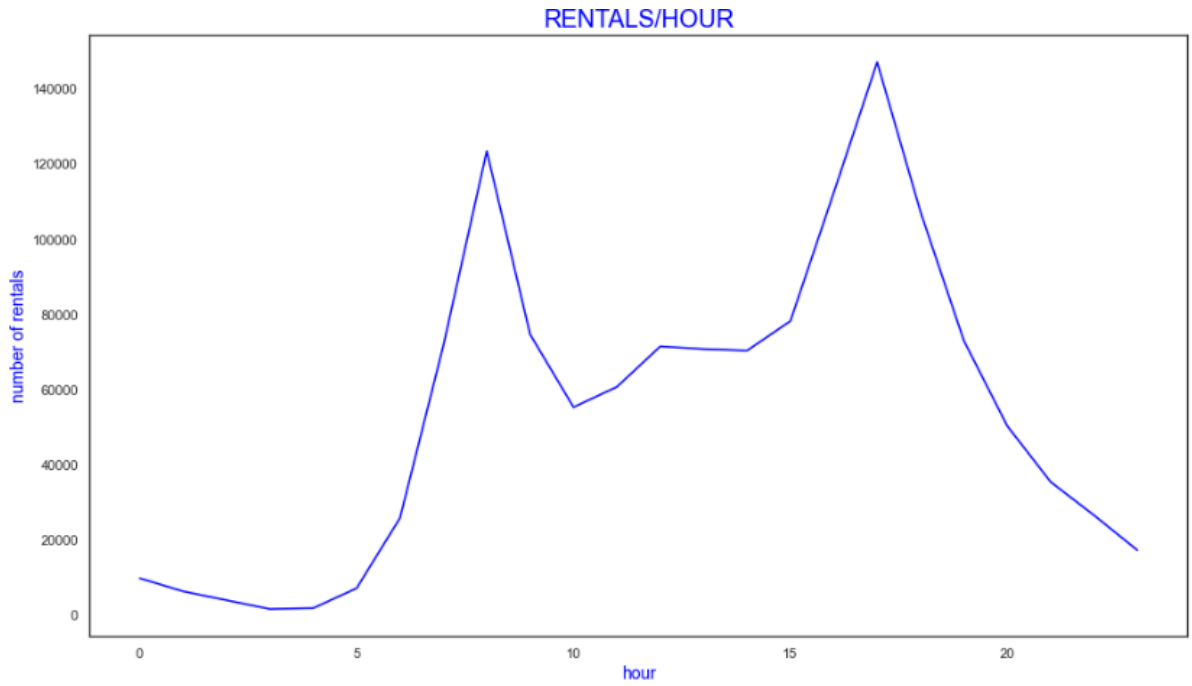

Figure 3

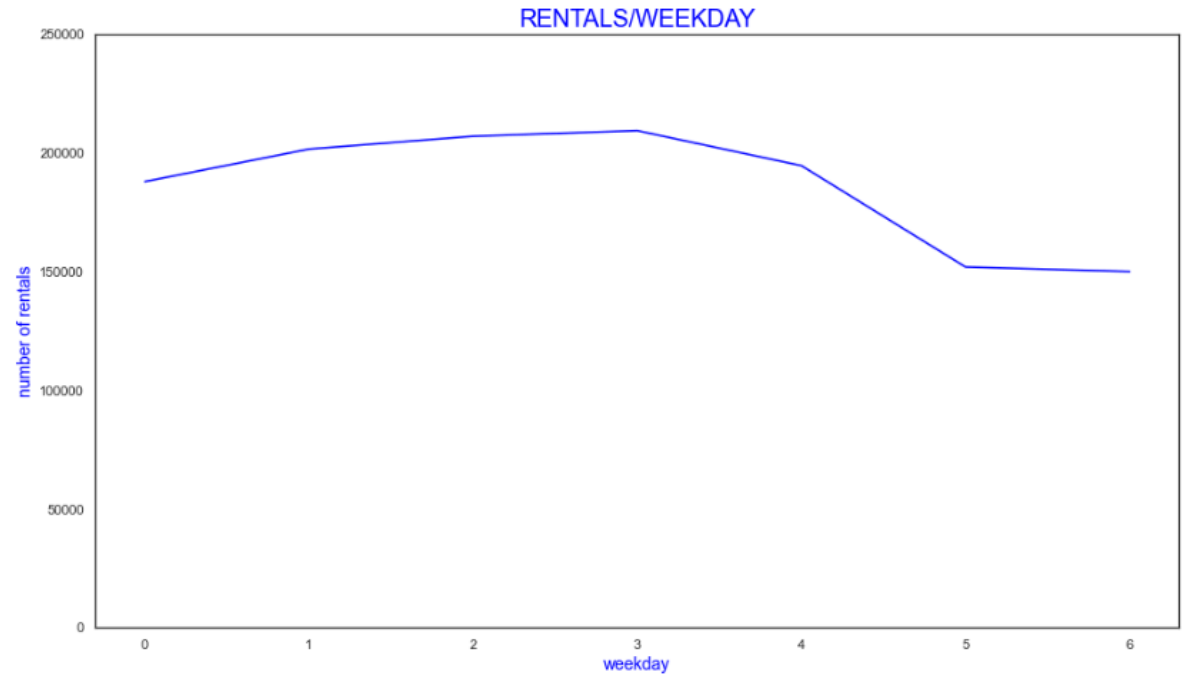

Figure 4

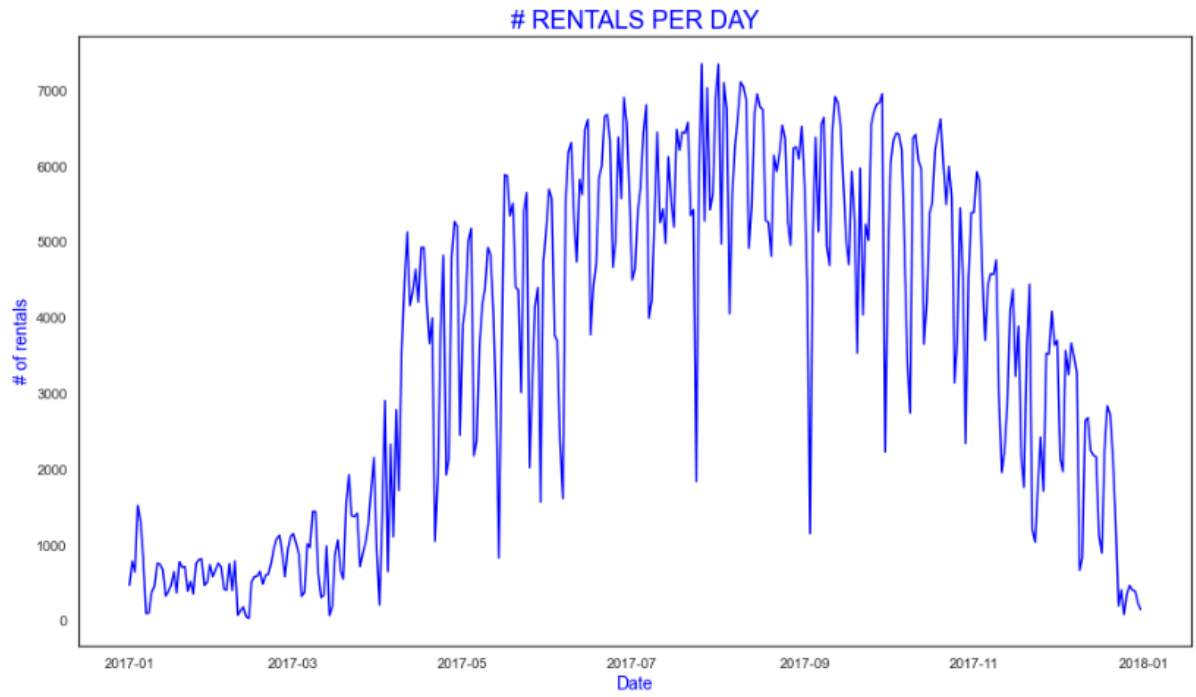

Figure 5

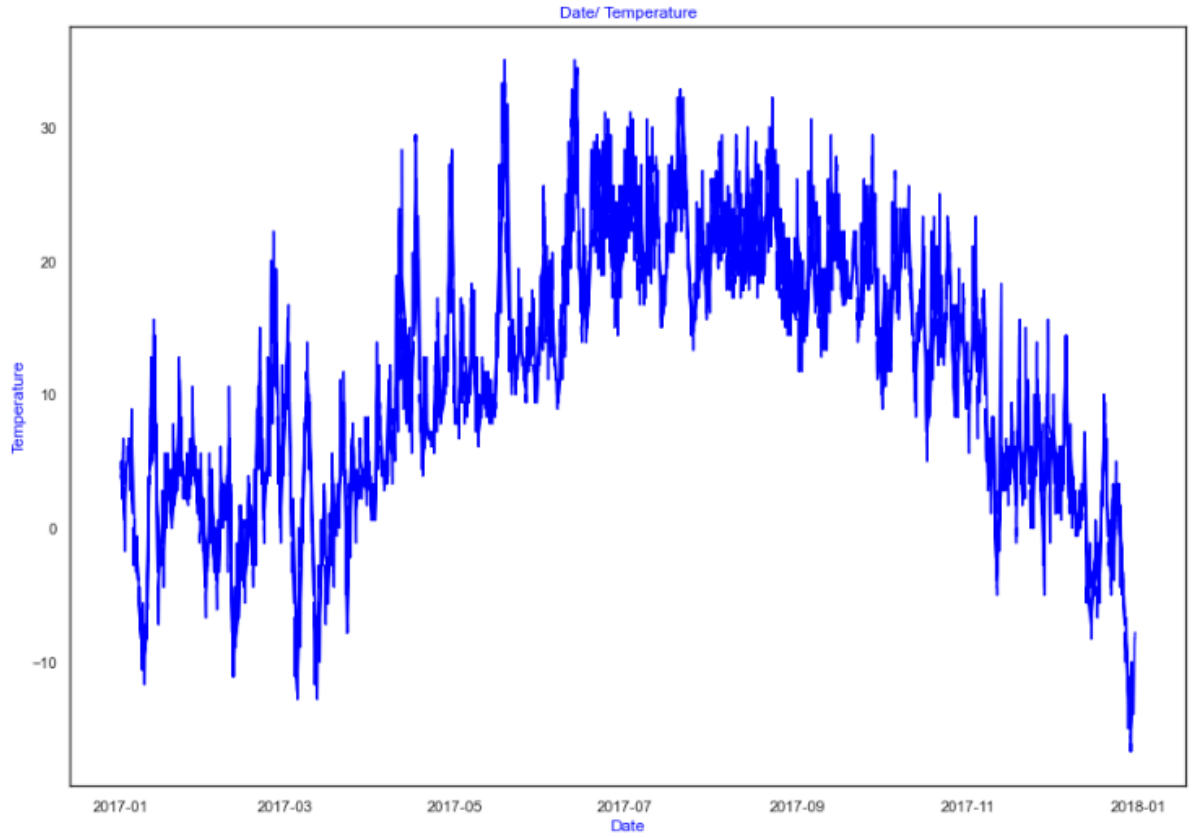

Figure 6

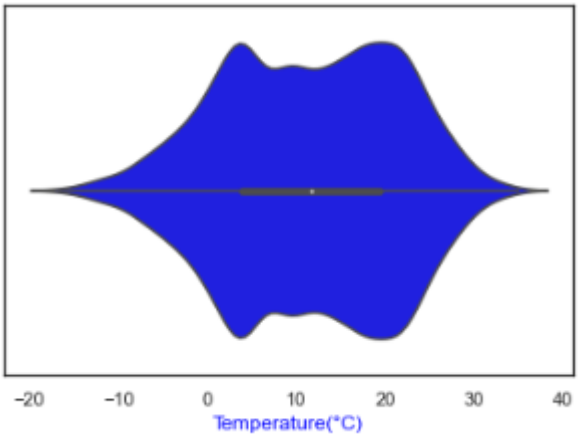

Figure 7

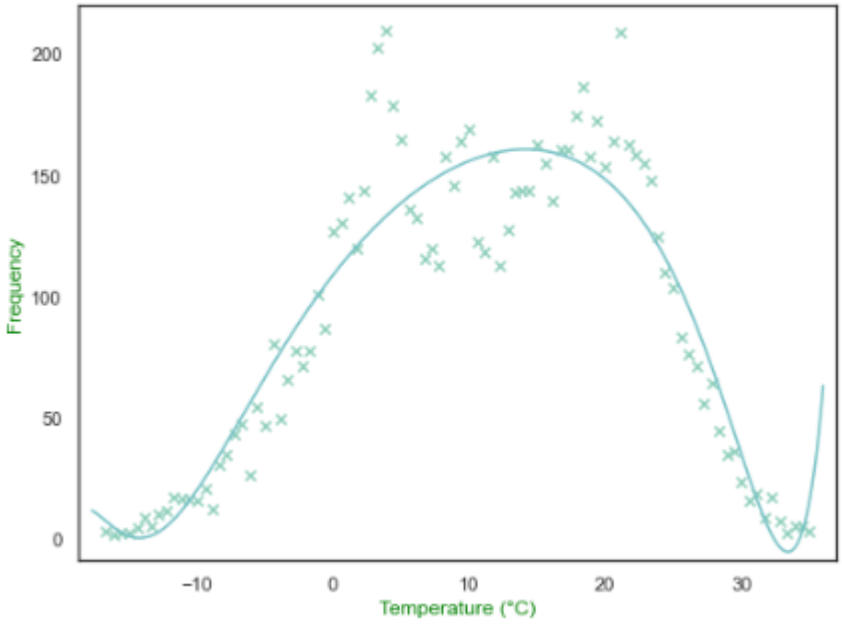

Figure 8

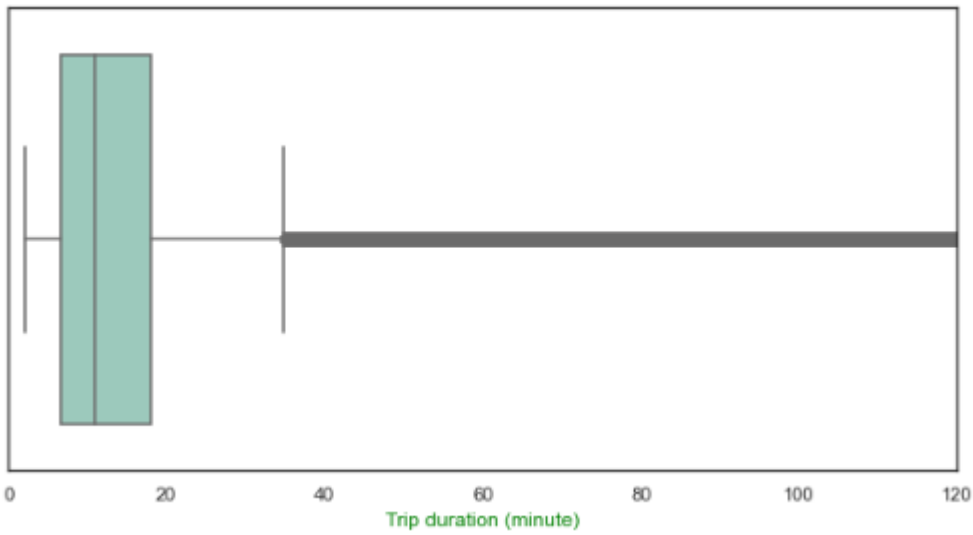

Figure 9

Figure 10

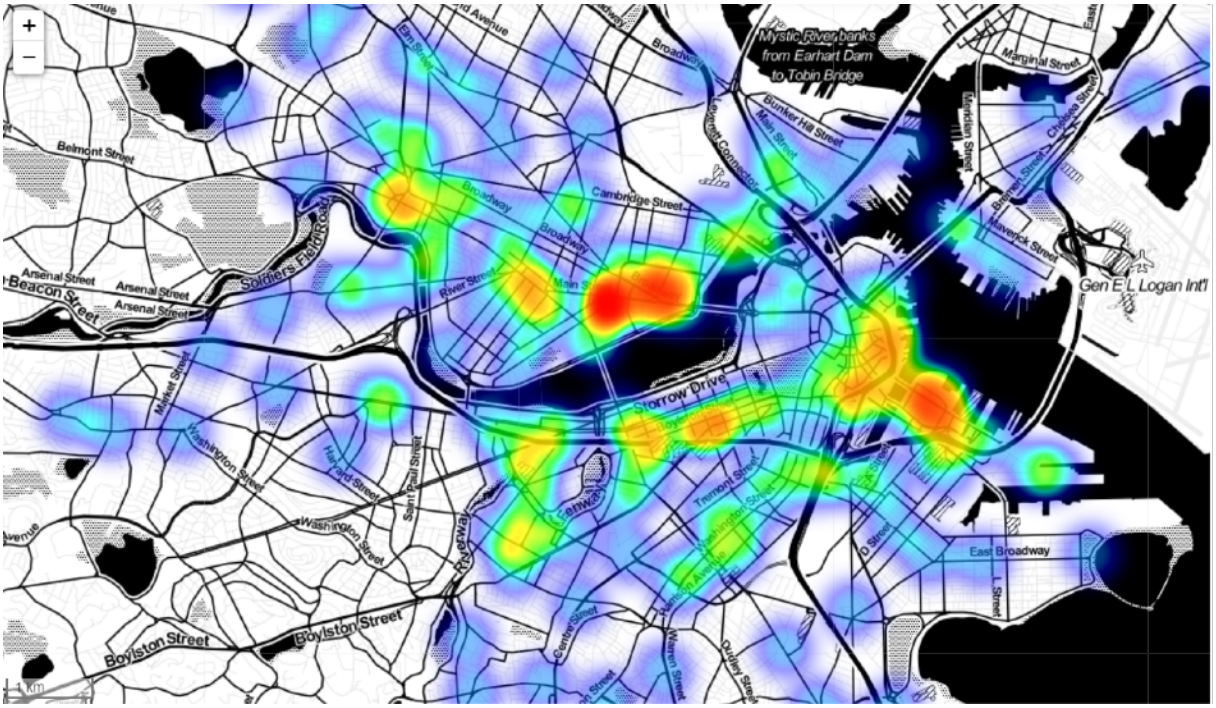

Figure 11

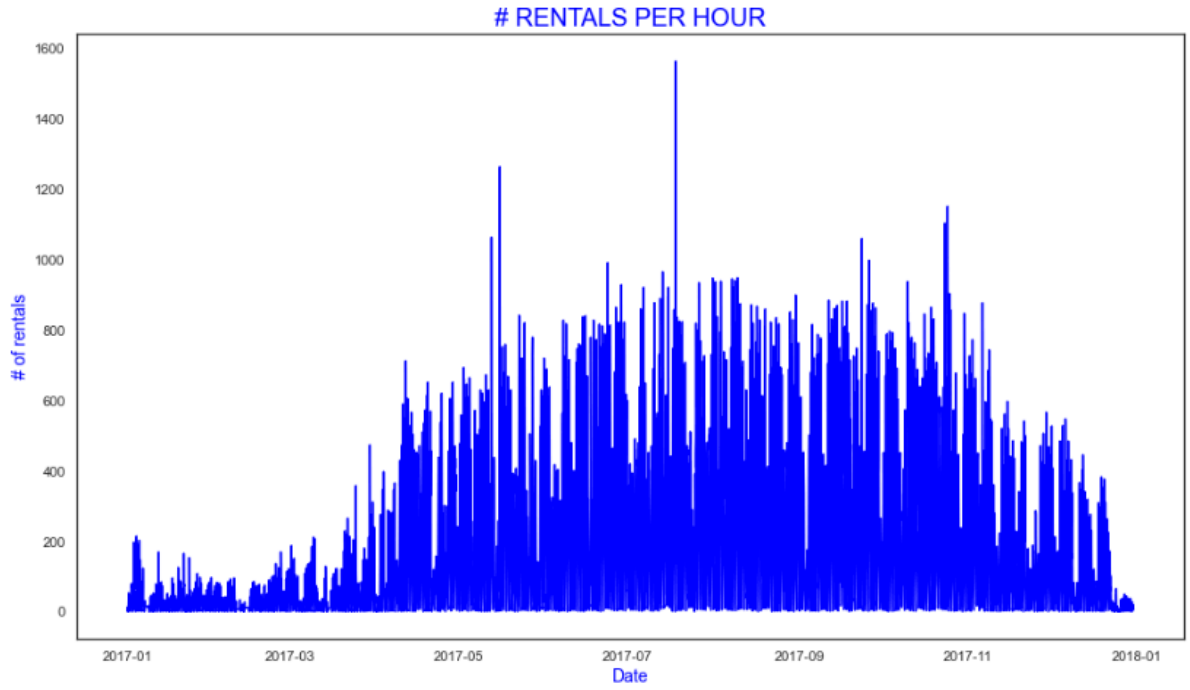

Figure 12

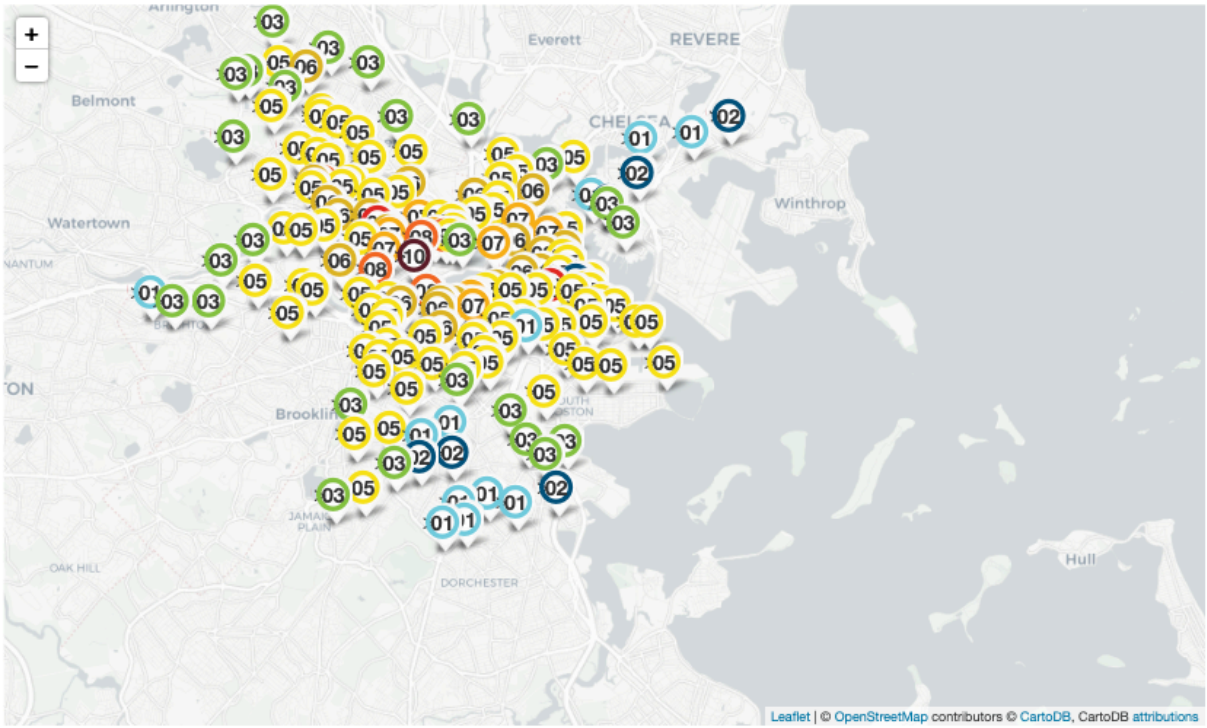

Figure 13

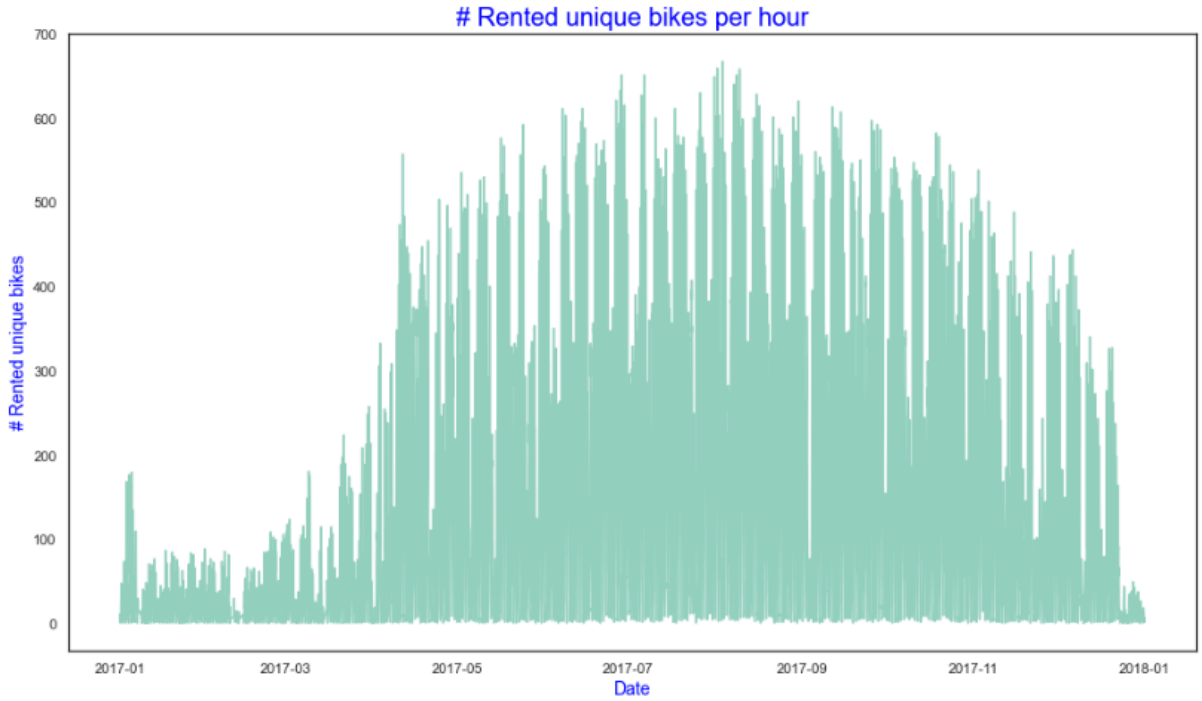

Figure 14

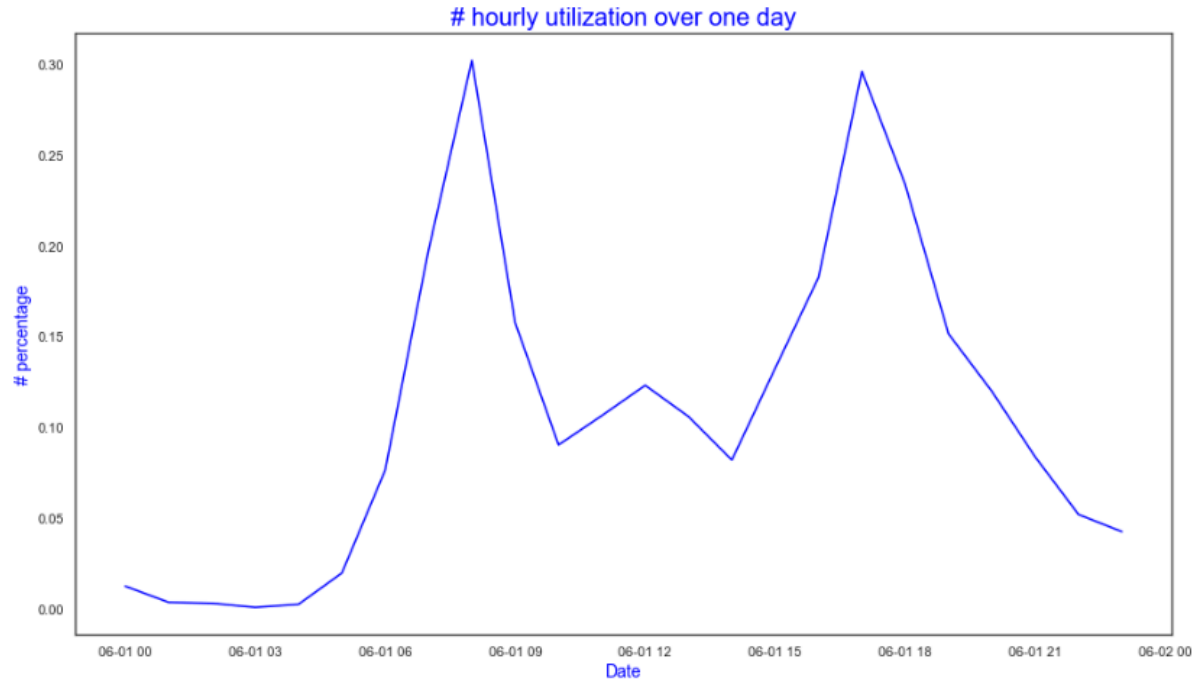

Figure 15

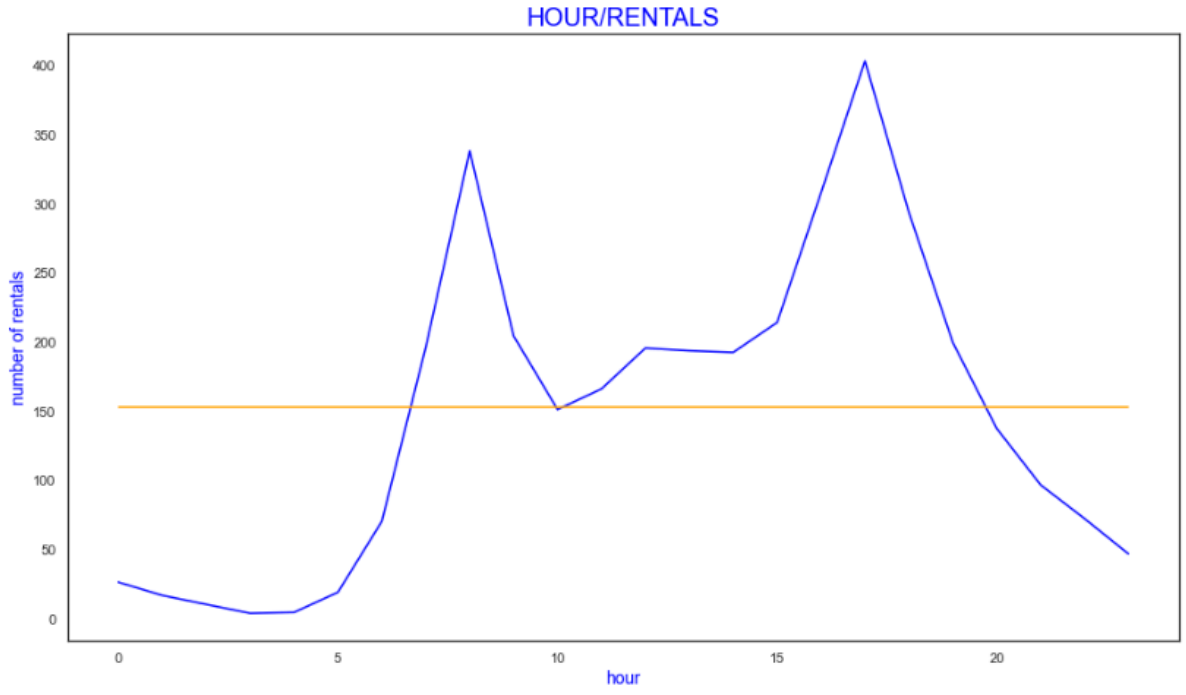

Figure 16

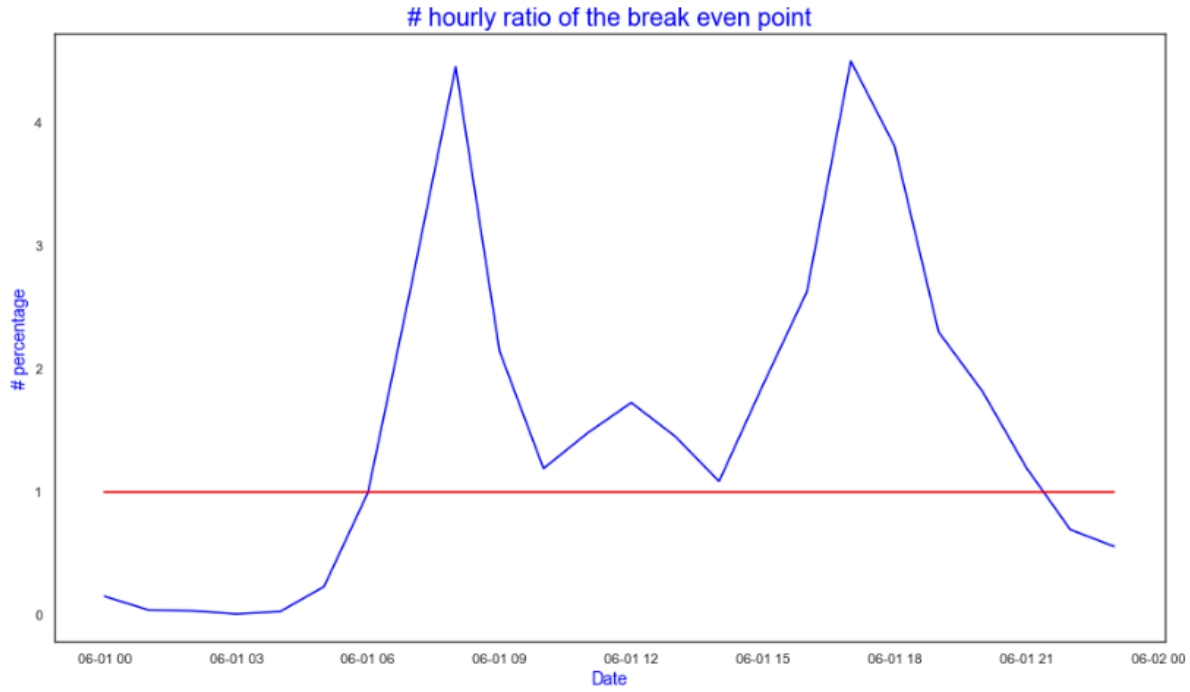

Figure 17

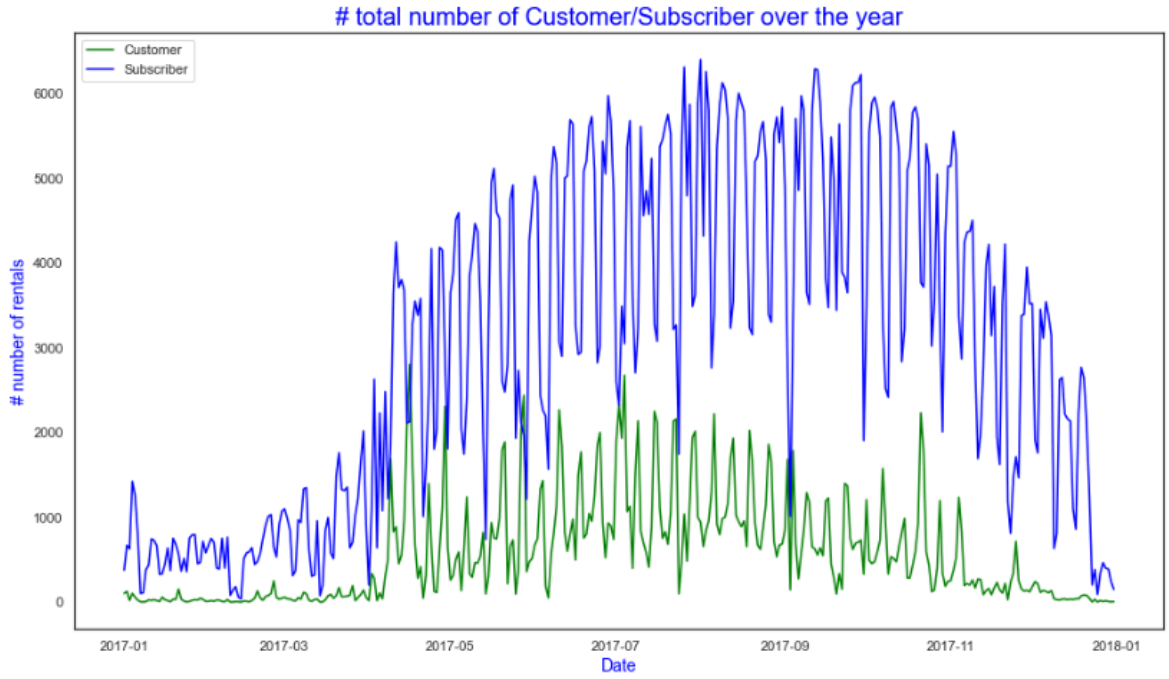

Figure 18

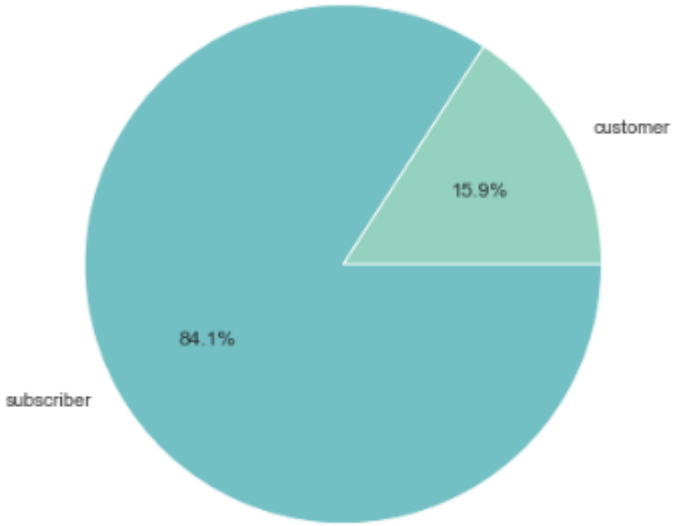

Figure 19

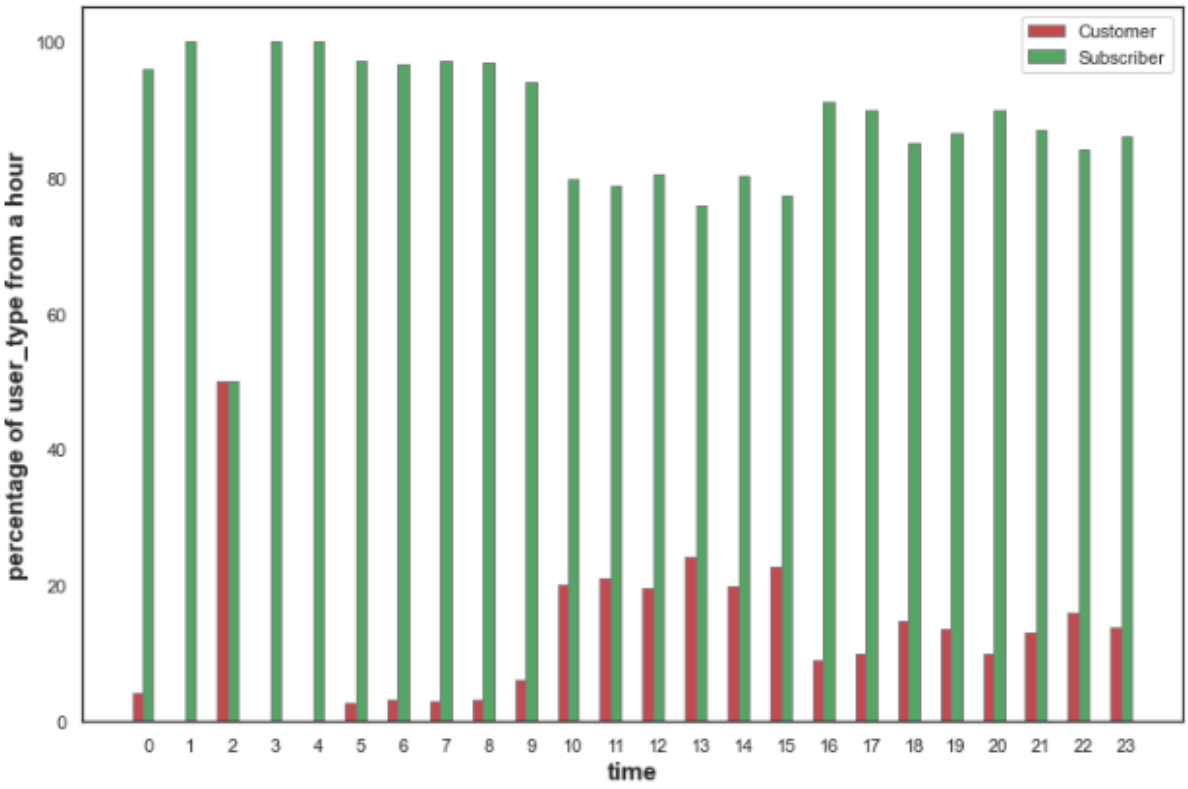

Figure 20

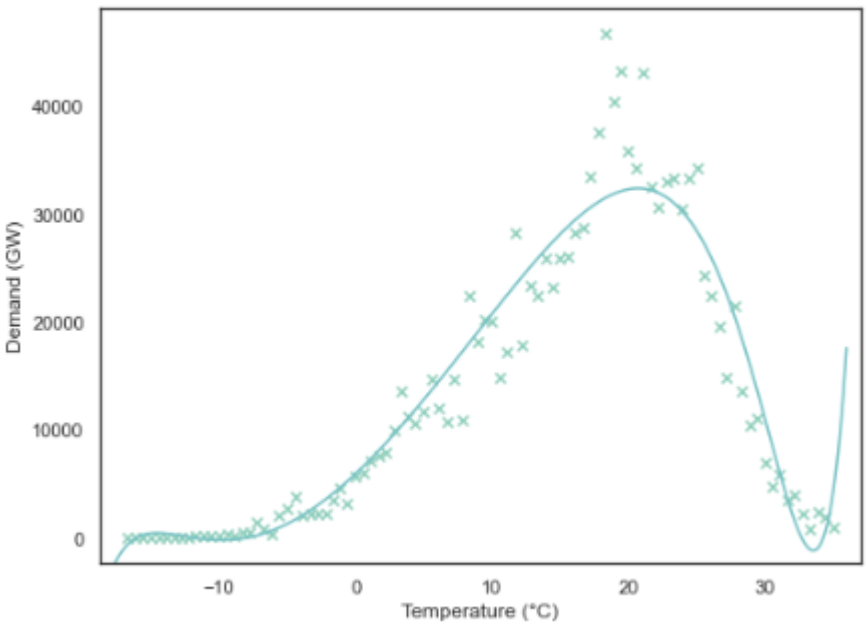

Figure 21

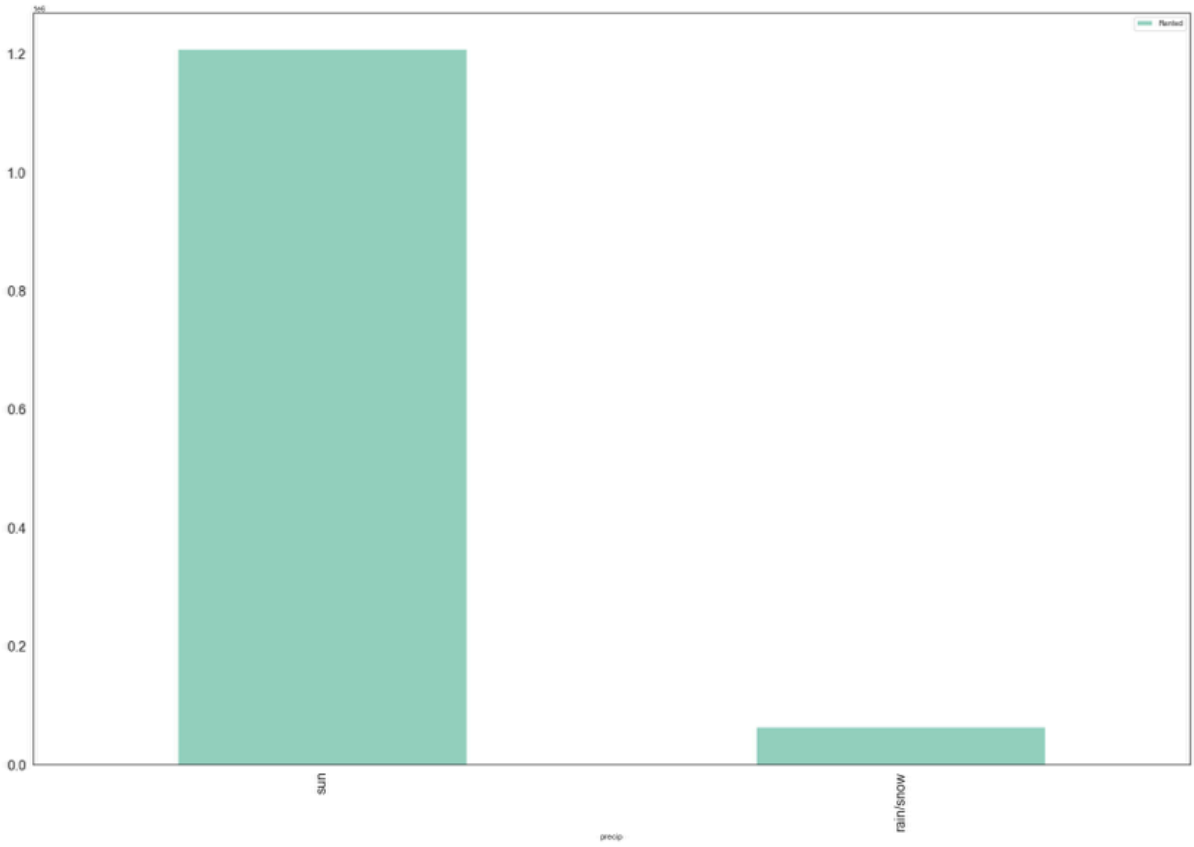

Figure 22

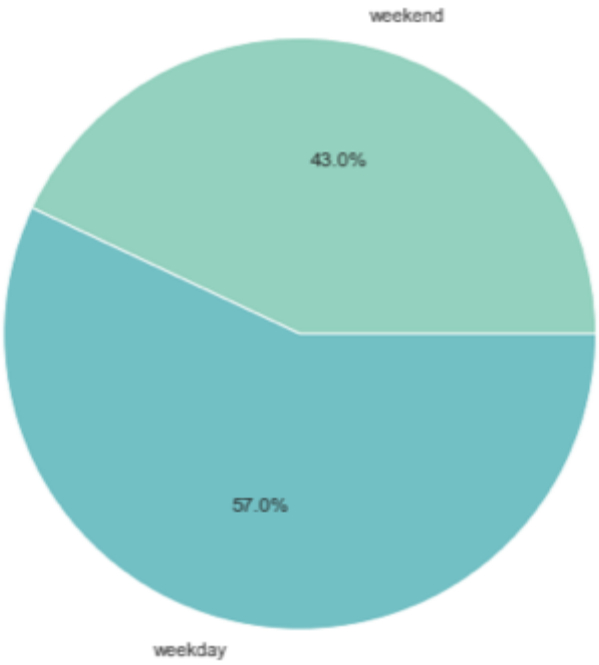

Figure 23

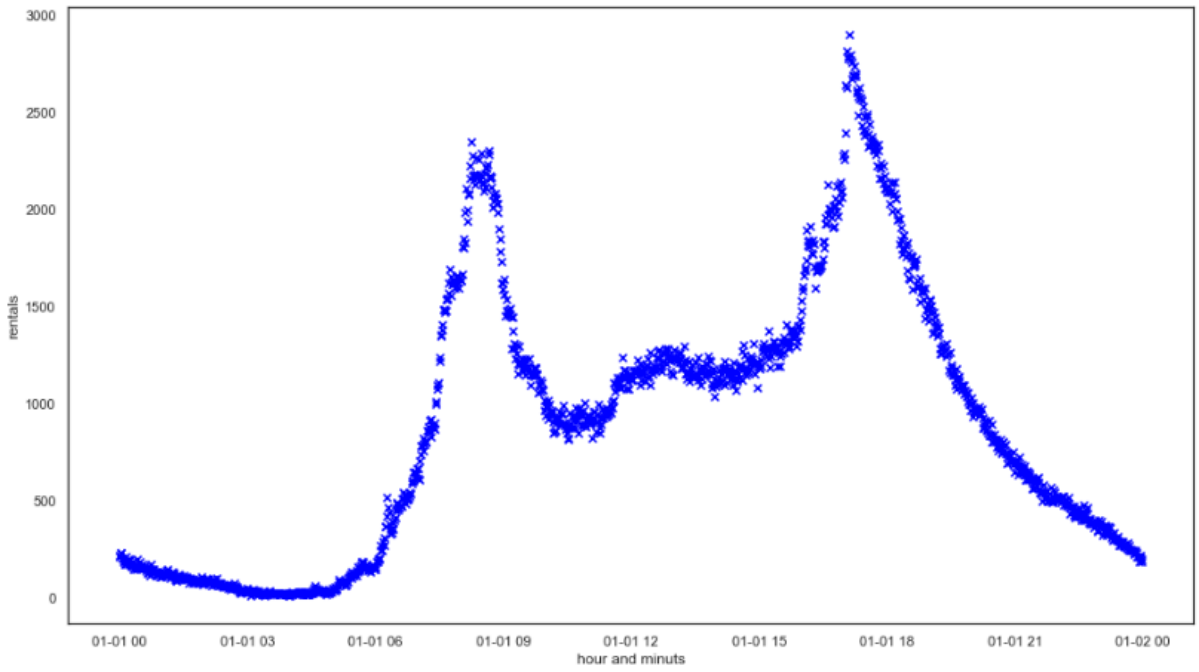

Figure 24

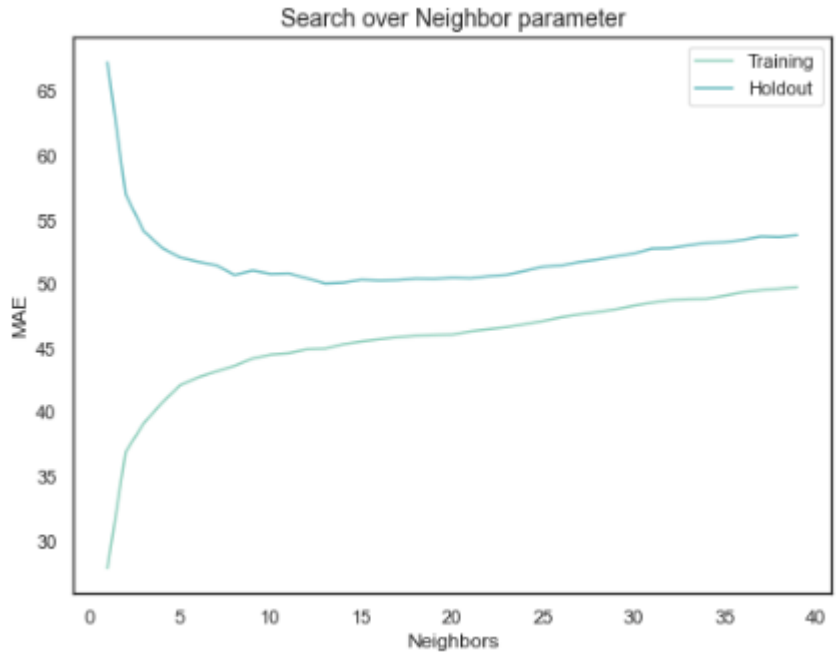

Figure 25

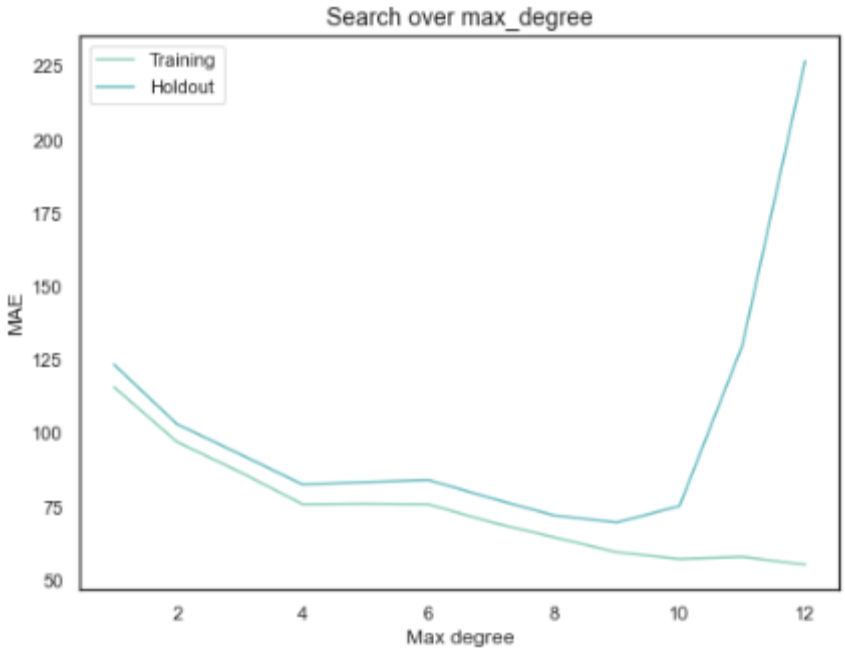

Figure 26

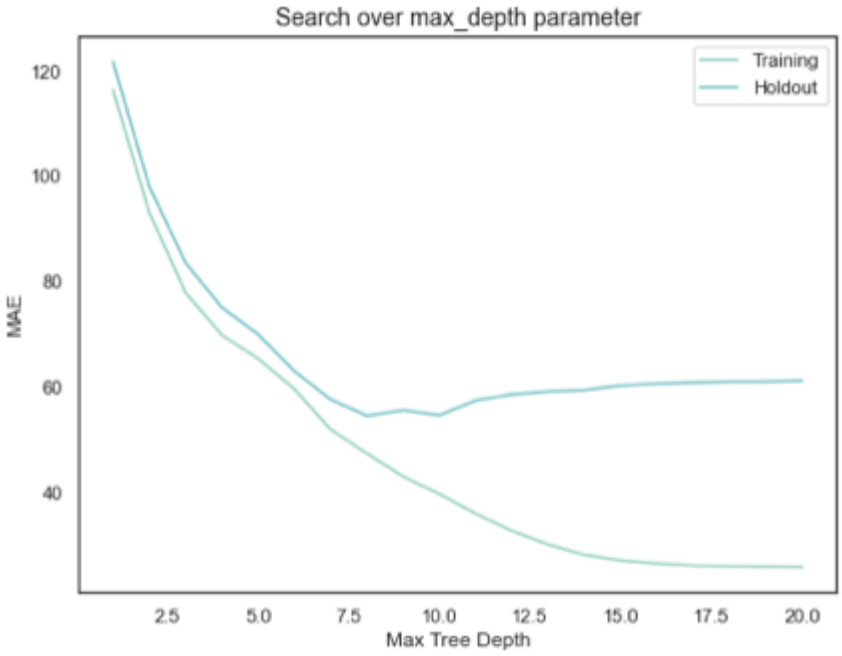

Figure 27

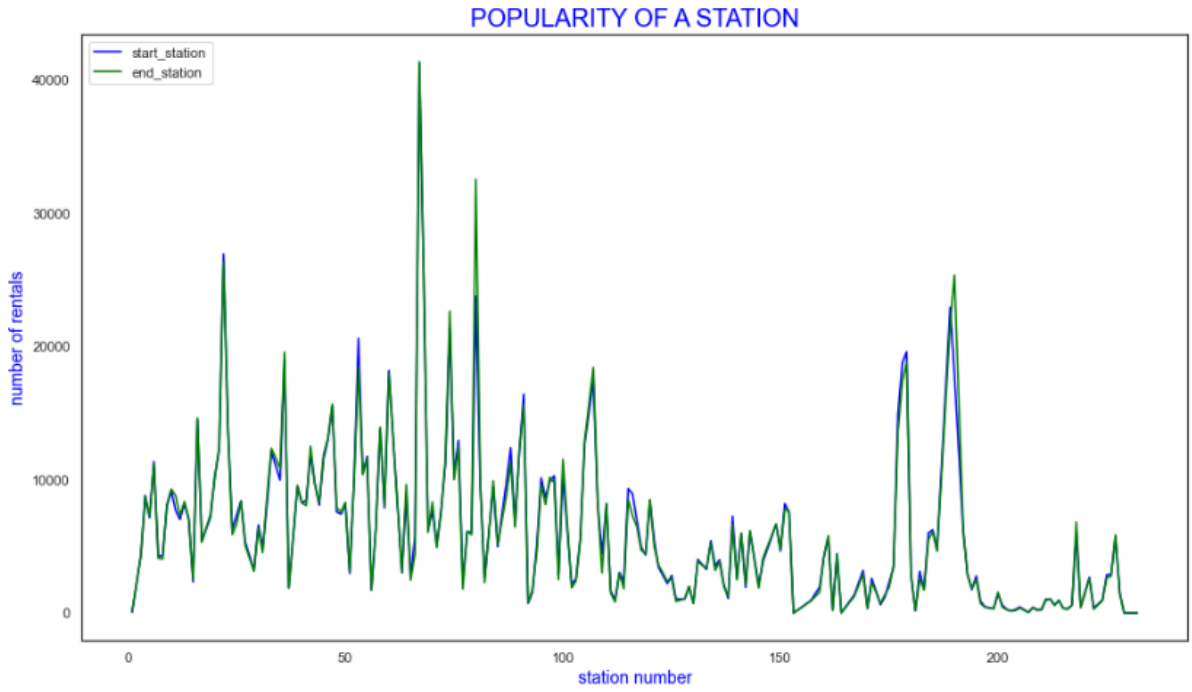

Figure 28

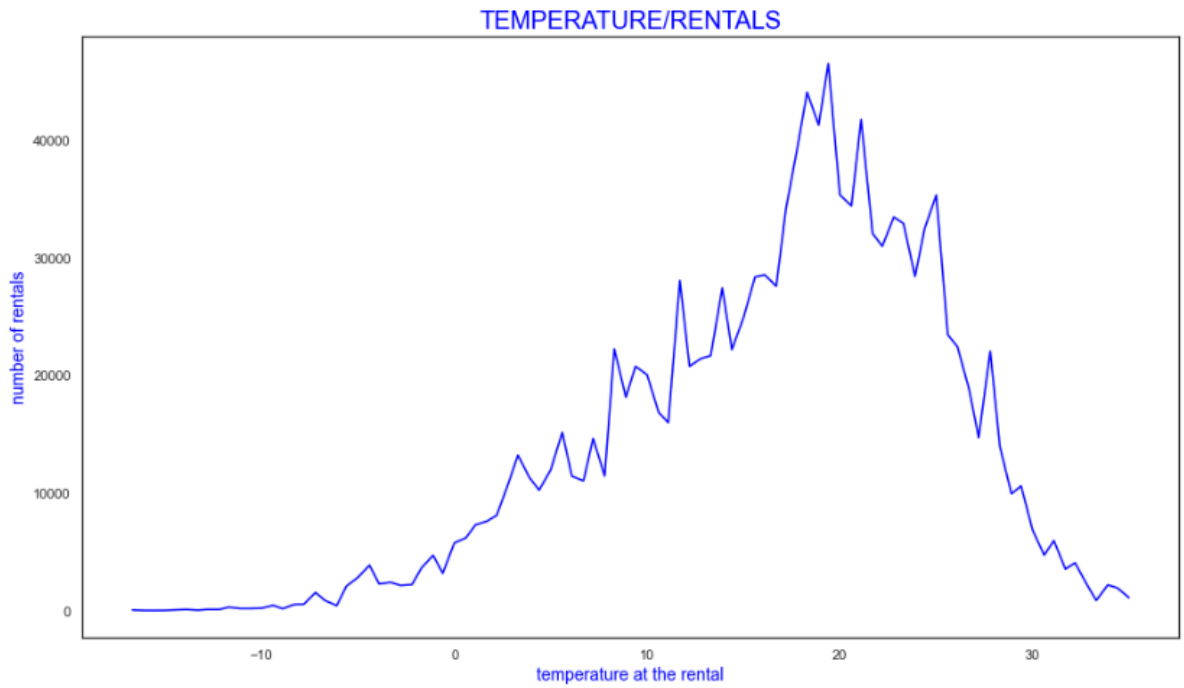
TEMPERATURE/RENTALS
number of rentals
40000
30000
20000
10000
0
−10
0
10
20
30
temperature at the rental

Figure 29