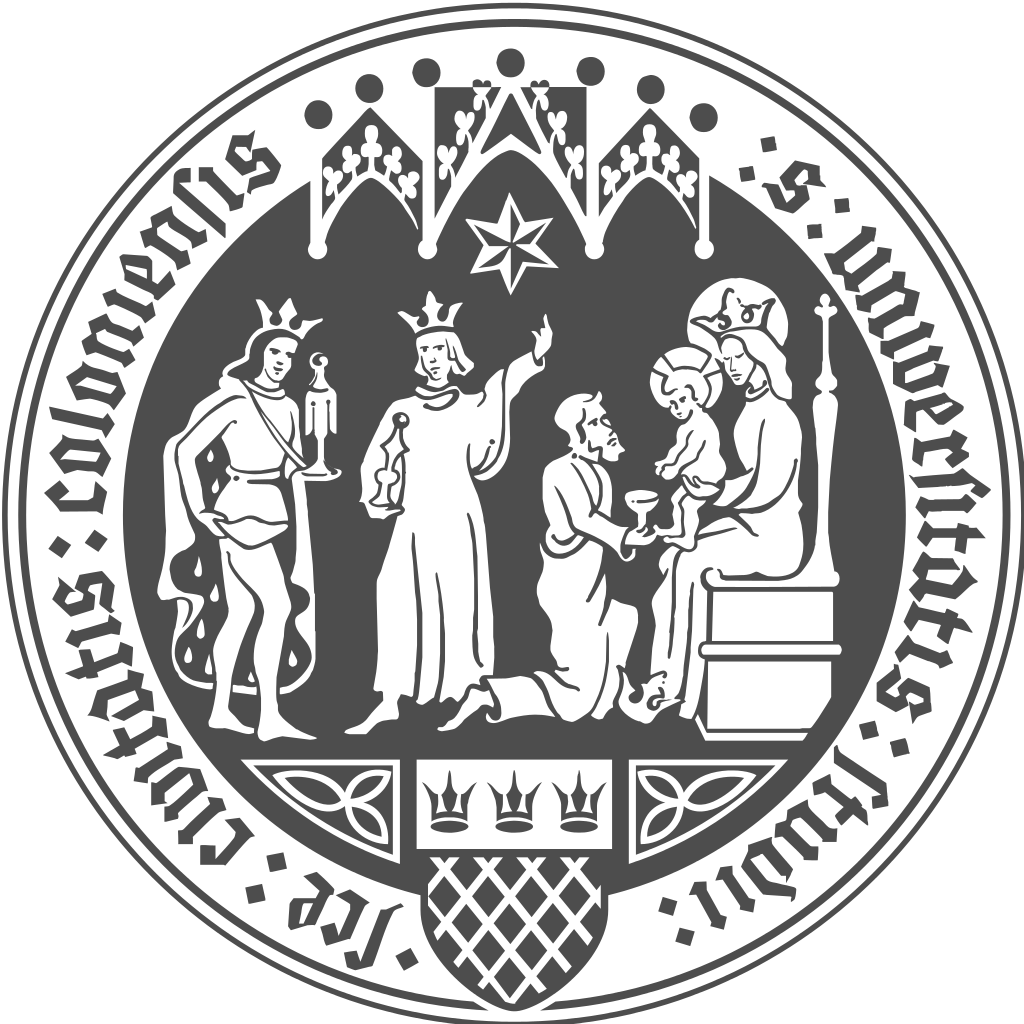
**Data Analysis Of Bike Rental Demand In Frankfurt - Group Black Mamba**

Group Assignment Paper



**Authors**:  
Nicklas Sander (Student ID: 7364184)  
Nina Wohlert (Student ID: 7364123)  
Alexander Abd-Alla (Student ID: 7364394)  
Johanna Berte (Student ID: 7364418)  
Sophia Martin (Student ID:7364316)

**Supervisor**: Univ.-Prof. Dr. Wolfgang Ketter

**Co-Supervisor**: Nastaran Naseri

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

January 30, 2019

# Eidesstattliche Versicherung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne die Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten und nicht veröffentlichten Schriften entnommen wurden, sind als solche kenntlich gemacht. Die Arbeit ist in gleicher oder ähnlicher Form oder auszugsweise im Rahmen einer an- deren Prüfung noch nicht vorgelegt worden. Ich versichere, dass die eingereichte elektronische Fassung der eingereichten Druckfassung vollständig entspricht.

Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

**Nicklas Sander, Nina Wohlert, Alexander Abd-Alla, Johanna Berte, Sophia Martin**

Köln, den 30.01.2020

# Executive Summary

The Deutsche Bahn has a service called ‘Call-A-Bike’ that offers the possibility to rent out bikes immediately via an App. Deutsche Bahn as well offers stations, where the bikes can be returned. In this bike rental domain, a lot of questions and problems occur. A few of them are considered in this report: What kind of customers are using the bike rental service?   
How many bikes are needed during a day, during seasons?   
What factors influence the customers renting behavior?   
How can future bike demand be predicted?   
The resources of the Deutsche Bahn are not used optimally. With this work we tried to design tools with which Deutsche Bahn can work with to understand their customers and bike demand patterns, better. Within the scope of this work, we considered bike bookings in Frankfurt for the years 2015 and 2016 and the corresponding weather at the time. The Data worked with was collected from two different sources. Call-a-Bike data by the ‘Deutsche Bahn’ and Weather Data by the ‘Deutscher Wetterdienst’. To answer these questions, different approaches were made.: Clustering to identify different types of booking stations, customer types and rental zone types, with the help of K-Means, Hierarchical Clustering and Gaussian Mixture algorithms. Furthermore, with developing a prediction model we want to predict Call-A-Bike rental demand in Frankfurt as a function of suitable features available derived from our sources. A polynomial regression with multiple features a LASSO regularization and a normalization was used.   
The presented work shows that the absolute amounts of rented bikes increased strongly from 2015 to 2016. Furthermore, higher numbers of bookings can be observed in the summer months. This is also supported by the fact that most bookings were made at 10-20 degrees Celsius.   
In Frankfurt, there are many trips to work / university school during the week. The main travel times are 7- 9 a.m. in the morning and 4 - 7 p.m. in the evening. The bikes are used less on weekends. However, the trips are longer (on average 31 minutes). During the week the average is 21 minutes.   
We defined different customer types: Occasional drivers, commuters (that drive to work) and leisure time drivers. Matching these drivers there are also leisure time stations, low traffic stations and centric stations. Using a polynomial regression, we have predicted the rentals / hour. After experimenting with relevant features, we achieved an R2 score of 72% with an RMSE of 36 bikes by lasso regularization and normalization of the input features (Degree 12). After aggregation of 2, 6 and 24 hours the best result is 24 hours aggregation with an R2 score of about 80% (Degree 4).

With this information we give Deutsche Bahn a tool to identify the amount of bikes they should supply in Frankfurt. Using our prediction model, they could foresee the development of the bike rental market and get a competitive advantage. With the results of the Clustering they now can identify their customer segment better and can address customers in a targeted manner.

# Table of Contents

[Eidesstattliche Versicherung 0](#_Toc30944749)

[Executive Summary 1](#_Toc30944750)

[Table of Contents 2](#_Toc30944751)

[List of Figures 3](#_Toc30944752)

[List of Tables 4](#_Toc30944753)

[1. Problem Description 5](#_Toc30944754)

[2. Data Description and Preparation 5](#_Toc30944755)

[2.1 Bike Data 5](#_Toc30944756)

[2.2 Weather Data 5](#_Toc30944757)

[3. Data Analytics 6](#_Toc30944758)

[3.1 Descriptive Analysis 6](#_Toc30944759)

[3.2 Clustering 7](#_Toc30944760)

[3.3 Predictive Analytics 8](#_Toc30944761)

[4. Conclusion 9](#_Toc30944762)

[A. Appendix 11](#_Toc30944763)

[A.1 Weather Data Preparation 11](#_Toc30944764)

[A.2 Linear Regression Model 11](#_Toc30944765)

[A.2 Division Of Group Work 11](#_Toc30944766)

# List of Figures

[Exemplary Figure](#_bookmark7) . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

# List of Tables

[Exemplary Table](#_bookmark3) . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

# Problem Description

At Share a Bike stations there are high fluctuations in demand. A lot of factors such as weather, season or the time of the day could have a strong influence on the number of bookings. In order to investigate the customers and their behavior, we used descriptive analytics and clustering. The goal of this project is to understand what the 2015/2016 usage structure of share a bike in Frankfurt looks like. We want to find out by whom and for what the bikes are used. We also study the effects of weather and season on the demand and design a prediction model to determine the hourly to daily demand of customers.

# Data Description and Preparation

## Bike Data

The underlying bike data set was downloaded from the website of the German railway company ‘Deutsche Bahn’ that collected all recorded Call-A-Bike Bookings of bicycles in Germany between 2014 till mid-2017. The corresponding file for our assigned time frame was the booking data set from 05/2017.[[1]](#footnote-1) Each booking entry includes information such as the booking time, rental zones and the Customers’ IDs. The file of 2017 didn’t have any numerical values except for 0.0 in the column ‘Distance’. Hence, calculations regarding the customers’ driven distances couldn’t be executed. Our data preparation included limiting the dataset to 2015 and 2016 and filtering the data set for bookings in Frankfurt. The columns with the names and IDs of the start and end rental zones contained NaN values[[2]](#footnote-2). We suspect that this occurred when the vehicles were dropped off at places that didn’t relate to a specific zone. Dropping the NaN values deleted the entire booking entry. For the analysis of absolute bookings that didn’t refer to rental zones we included these entries since they could still hold enough information for these analyses. Once we analyzed specific stations or routes, we dropped the booking entries with non-specified stations since they didn’t include any valuable information regarding stations anyway. For certain features such as trip durations we could find outliers that did not make a lot of sense, e.g. trips with a duration longer than 3 days. We didn’t drop those for the Descriptive Analysis to get a holistic picture of the data. Some of the outliers were dropped for the Clustering later.

## Weather Data

The source of the weather data is ‘Deutscher Wetterdienst (DWD)’. We chose historical data from one weather station in Frankfurt on an hourly basis. The historical data is already checked by the DWD and has errors marked and described. Because the time in the data is listed in Coordinated Universal Time (UTC) we decided to choose our timeframe from 23:00 31.12.14 until 23:00 31.12.2016 to ensure that we measure the exact time frame for Frankfurt, which lays in the Central European Time (CET). We decided to first upload all possible datasets to get a as much information and options as possible.

We eliminated error values with NaN values and missing data where filled via forward fill. Details are provided in the Appendix A.1   
A Pearson Correlation was implemented. Due to a high correlation (>0.2) with Temperature and more variables we decided to drop Sunshine Duration, Visibility and Relative Air Humidity. This enables us to focus on the future feature target (bike rental demand) and provides us features that are not strongly correlated with each other. Finally, we joined weather data with bike data. The time of the bike data was changed to UTC and the minutes were cut to hours to make sure the right weather is joined with the right Bike data. Heat map 2  
The final prepared dataset consists of Temperature, Total Cloud Cover, Precipitation Height, Mean Wind Speed, Month, Weekday, Hour, Year, amount of bookings. Again, we checked the Pearson Correlation for all variables to get a good overview of the data, as Fig drölf shows.   
The dataset was aggregated to a 2h, 6h, and 24h horizon for the Predictive Analysis later. While aggregating we used the median for those variables, where it was not useful to sum them up in order to avoid that outliers distort the data.

# Data Analytics

## Descriptive Analysis

The Descriptive Analysis was used to describe bike rental demand patterns over the given time interval and give possible explanations for found patterns. The data showed that the year 2016 had approximately 77% more bookings than the year 2015. Seasonal analyzations showed that the summer season had more than 100% bookings than in winter (almost 200% in 2016), whereas spring and autumn showed values in-between those two. Fig 4 Booking amounts per month for both years Figure 4 shows the distribution of bookings over the course of the years 2015 and 2016 with the violet graph representing 2015 and the red graph 2016. It illustrates that bookings continued to increase steadily until August and then declined steadily until January. Figure 5 Booking amounts per week dayAnalyzing the different weekdays showed that the amount of bookings from Monday till Friday is approximately the same with a decrease of bookings during the weekend (see Figure 5). Sunday has the minimum amount when most people don't have to work. Figure 6 Mean trip duration for weekdaysAdditionally, we analyzed the mean duration of trips for the different weekdays. Monday to Friday showed an average duration of around 23 minutes whereas Saturday and Sunday showed around 31 minutes (see Figure 6). An explanation could be that bicycles are used during weekdays to drive to work or school. During the weekend bikes may be used for private trips without a defined destination and without any time pressure.Figure 7 Booking amounts per hour (year 2016)Figure 7 shows that booking rates are highest in the morning (7.00 am - 9.00 am) and in the evening (4.00 pm - 7.00 pm). The shown distribution of bookings could be explained with the typical working rush hours. Though, in the evening the booking time varies more than in the morning. This may indicate that most employees start at a similar time but end the working day at different times. Figure 8Mean Booking duration per hourFigure 8 visualizes the mean trip duration for the hours of the day. At each hour, the average journey time is between 15 and 30 minutes. This suggests that rented bicycles are mostly used to cope with journey times between this interval. It is noticeable that booked bicycles between 5.00 a.m. and 9.00 a.m. have significantly shorter travel times than the other hours. Customers who book a bike during this time may be in a hurry. If the journey time lasts longer, customers might choose a different type of transport. Analyzing rental stations showed that the main train station is the center of traffic with the two rental zones next to the station being the two most used rental zones. The station to the main entrance even had 150% more bookings than the rental zone on the second place. Consequently, the most driven routes were also located around that area.

## Clustering

We used Clustering Algorithms such as K-Means, Hierarchical Clustering and Gaussian Mixture to test whether certain trips, customers or stations would form clusters based on their similarity for specific features. In a first run, we identified outliers that wouldn’t fit into any of the clusters and removed them for the second run. Afterwards, the elbow method was used to approximate the optimal number of clusters. Fig 10 Trip type cluster (Amount of bookings – temperature  
Figure 10 illustrates different clusters with the amount of booking in respect to the temperature. Using K-Means required us to scale the values first with the x-axis showing the temperature and the y-axis showing the amount of bookings. It basically shows that most bookings are made during temperatures between 10 to 20 degree Celsius and bookings decrease the colder or warmer it gets. Hence, it makes sense to form 3 clusters for trip types (low degree trips (yellow), medium degree trips (blue), high degree trips (red)).Figure 11Customer type cluster  
Customer Clustering in terms of average trip duration and number of bookings showed that a large proportion of customers have few bookings at a low average duration. Furthermore, the customers with the most bookings have a low average duration, while those with the fewest bookings have a long average duration. Figure 11 shows this distribution in form of 3 clusters. We named these clusters ‘occasional drivers’ (violet), ‘commuters’ (yellow) and ‘leisure time drivers’ (green). Figure 12 Station types regarding customer types cluster  
The first cluster for the rental zones was built in terms of regular and occasional customers. For this purpose, regular customers were defined as customers who had more than 100 bookings at a station during the selected period. Occasional customers are correspondingly those with less than 100 bookings. We used 100 as a limit because higher limits of up to 500 showed that not many rental zones have customers with that amount of bookings. An explanation for this could be that at some point it gets more favorable for customers to buy an own bike instead of constantly renting one.   
The result in Figure 12 illustrates the optimal amount of 3 clusters with one cluster for stations with only a few regular and occasional customers named ‘low traffic stations’ (blue), second cluster stations for those with medium to many occasional customers named ‘leisure time stations’ (red) and lastly stations that have medium to many regular customers named ‘centric stations’ (yellow).

## Predictive Analytics

To predict the bike rental demand per hour (target variable) for Frankfurt based on the data of 2015 and 2016 we abstracted a feature set out of the data, that we would use to induce, test and evaluate different models. As input features for our hypothesis function we chose the selected set of features that has been described in section 1.2.   
We started with a **linear regression** with several input features. The results were not satisfactory (for details see Appendix A.2) so we moved on to another approach.   
Next, we implemented a **polynomial regression**. We transformed our input features into polynomial components of different degrees to see, which of these more complex hypothesis functions would yield the best results. Therefore, we implemented a cross-validation, that split the data into a training set (50 % of the data points), a validation set (20 %) and a test set (30 %).Figure 14: Training and Validation Error for polynomial model  
Figure 14 shows the training and validation error for the polynomial model over different degrees. The minimum for the validation error function at the degree of 3 is clearly visible. We tested our model performance for a polynomial model with degree of 3 on the test data set. The result was a Root-Mean-Squared Error (RMSE) of 47.2 bikes per hour and an R²-Score of 51%.   
Due to the relative low performance of our model, we decided to reduce the number of features in favor of trying to potentially capture more complexity via higher degrees in the polynomials. The input features were reduced to Temperature, Weekday, Hour, Month. Additionally, due to the high difference between the 2015 and 2016 data set observed in the Descriptive Analysis, we decided to implement year as another feature. Furthermore, we decided to implement a **LASSO regularization** for our hypothesis function to control the sizes of the parameters. The decision for LASSO was made, because LASSO can automatically assign a zero parameter to our input parameters and therein helped us in our feature selection. We ran the same logic for the setting of the hyperparameter (degree) and found, that for polynomial regression with LASSO regularization, the best performance of our model is at degree 18 with a RMSE of 42.7 and an R²-Score of 61.7%.   
To gain even better results we chose to **normalize the input features**. This led to R²-Score of 72% and RSME of 36 bike rental bookings per hour on the test data set. The model performance by degree is depicted in Figure 15. The degree used with the normalized input features was 14.Figure 15 RMSE and R² for Polynomial Model with normalization  
After comparison of all results our best predictive model was the polynomial regression model with LASSO regularization using the normalized input parameters, that were slimmed down to temperature, weekday, month, year and hour. It performed with R²-Score of 72.6% and RMSE of 36.06 bike rentals per hour. We applied the same procedure to the data splits for two, six and 24 hours. The results are shown in table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Most accurate degree | R²-Score on test | RMSE on test |
| One-hour | 12 | 72,6% | 36,1 |
| Two-hour | 8 | 44,7% | 93,0 |
| Six-hour | 4 | 51,7% | 241,2 |
| 24-hour | 4 | 80,3% | 398,1 |

Tab Performance of models depending on timeframe

For comparing the different models, relative evaluation metrics must be considered because the absolute metrics are using different baselines. The different R²-Scores show that our models perform best for either the one hour or the 24-hour time frame. This can be explained by the one-hour-model having a lot of training data and therefore prediction might be more precise. On the other end, the 24-hour split might level hourly variance into a more reliable prediction. The two medium splits fulfill neither of these properties and therefore are performing worse.

# Conclusion

The Deutsche Bahn can use our results to improve the knowledge of its customers. Based on the exposed customer groups, marketing measures can be started to address missing / hardly existing customer segments. One example is the low number of drivers at the weekend. The Deutsche Bahn could promote driving on weekends through special events (e.g. organizing city rallies to the top spots of Frankfurt). In addition, the predicted values of our polynomial regression can be used to calculate how many bikes are needed. Using this knowledge, costs can be saved (by optimal use of resources) and the service quality of the customers increases.   
Nevertheless, our results have to be reflected critically. In the area of lustering and Descriptive Analysis, the following aspects should be noted: According to our current knowledge, one person can book two bikes at the same time. Therefore, a person can be included in the analyses several times, which means the rental demand patterns can be falsified. Furthermore, on the website of the booking data source it was discussed that probably not all stations are included in the data. One station is missing in our dataset.  
In the area of prediction, the biggest criticism is that the records of the rentals did not measure the actual demand. Consequently, if there were not enough bicycles at a location, potential customers could not rent bicycles. This shortage of bike demand is not recorded.   
Furthermore, future weather data was not considered when predicting the rentals. We looked at what the weather was like in the hour when the bike was rented, but depending on the length of the trip, this can have only small / no effect.   
If the prediction is carried out with a station grouped databases, the demand per station can be calculated based on the parameters. It can therefore be revealed how many bikes are needed per station. Using this knowledge, costs can be saved (by optimal use of resources) and the service quality of the customers increases.

In further research, a more complex model could be developed, considering the incoming and outgoing bicycles. This increases the accuracy and the benefit for the company because of the considering of rotation. In addition, a Time Series Analysis could be carried out to clearly show variations in day/night or the seasons (spring, summer, autumn, winter).

# Appendix

## A.1 Weather Data Preparation

Figure 1 Dataset with Error values and missing data  
There were error values in the dataset. E.g. the mean precipitation height -0.57 is in terms of content an unrealistic number as Fig. 1 shows. Therefore, we assigned numerical error values (-999) to 'NaN' values in order to prevent that statistical methods will be biased. Fig 1 shows, only sunshine and precipitation values were missing. Precipitation is only missing 19 values. These were filled via forward fill, because the missing values were < 0.2% and we assume some weather stability, rather than to assume 0 or mean values. The sunshine duration is always missing 6h from 9 pm until 3 am. We imputed them with zeros, because the sun does not shine at night in Frankfurt.

## A.2 Linear Regression Model

Firstly, we ran a linear regression. We implemented a Train-Test-Split, so that we could use 70% of the data for fitting the linear regression and 30% of the dataset for measuring the performance. The linear regression yielded a RMSE (root-mean-squared-error) of 56.55 bikes per hour and an R²-Score of ~ 26%. As the training performance parameters were not significantly better, it became obvious, that our model was underfitted. That result was expectable, because not all our input features were truly linear (e.g. Descriptive Analysis had shown that the demand is not rising through the week from Monday (represented as 0) to Sunday (6) and neither falling). It was proved, that pure linear regression should be discarded for predicting the bike rental demand and we need another approach.

## A.2 Division Of Group Work

For the group work we decided to split into two content related groups. One bike group with Johanna Berte and Alexander Abd-Alla and one weather group with Nicklas Sander, Nina Wohlert and Sophia Martin.

The bike group worked on the Data Collection and Preparation for the Call a Bike data. Also, this group worked out the descriptive Analytics and the Cluster Analytics including the prepared weather Data of the other group.

The Weather group collected and prepared the Weather data and brought together bike and weather data. With the help of the prepared bike demand data, they designed and trained the Predictive Models.

The part Discussion and Outlook was worked on together.

At all times both groups gave each other status reports, feedback and constructive criticism to ensure the quality of our work.

To do:

* ~~Eigennamen bestimmen und groß schreiben~~
* ~~Neue Inhaltsstruktur~~
* Bilder einfügen und Nummerierung checken
* ~~0,5 – 1 satz zur Linearen regression schreiben~~
* ~~Appendix schreiben~~
* Kopfzeile?
* ~~Executive summary schreiben~~

1. <https://data.deutschebahn.com/dataset/data-call-a-bike/resource/0fcce4dd-7fc6-43f8-a59c-983a7945f8ba> [↑](#footnote-ref-1)
2. NaN stands for ‘Not a number’, meaning an undefined value [↑](#footnote-ref-2)