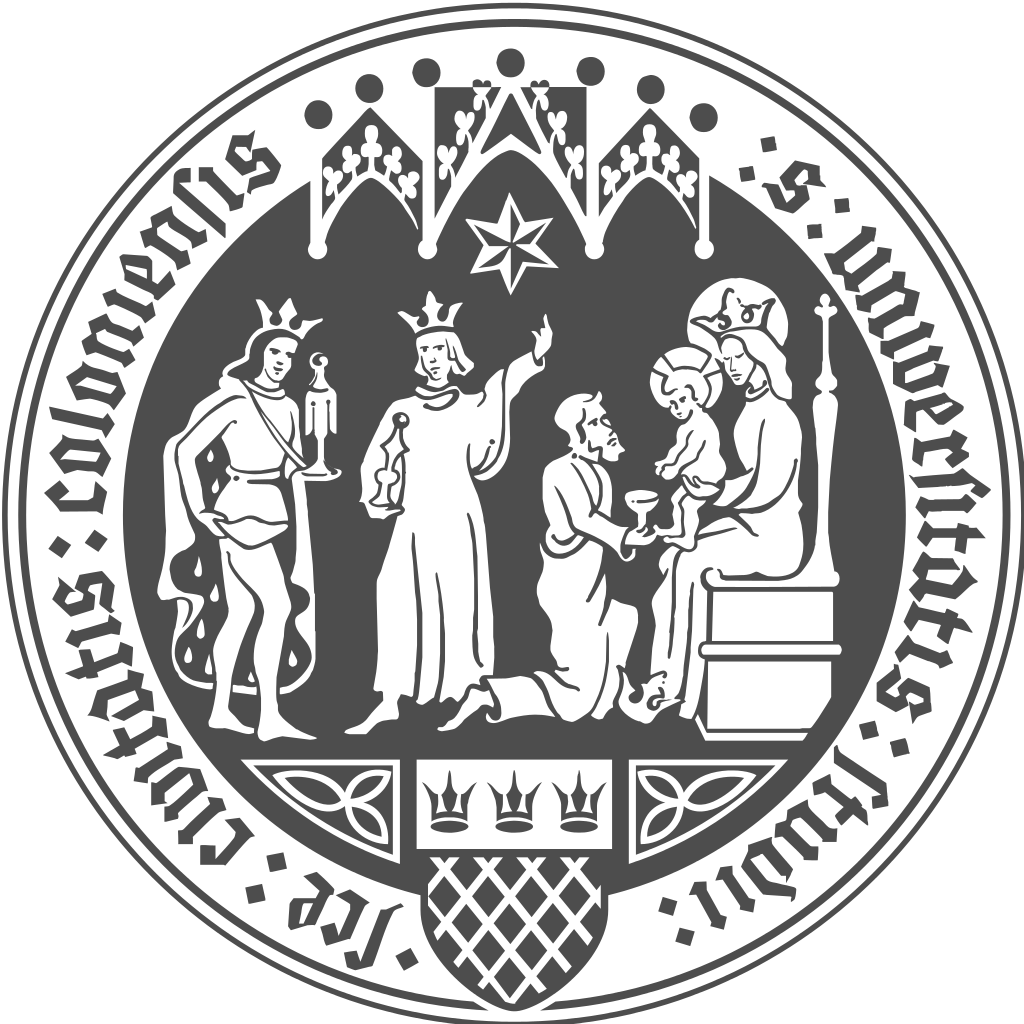
**Data Analysis Of Bike Rental Demand In Frankfurt - Group Black Mamba**

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**: Nasti?

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**

2. One-page executive summary: summarizes the entire report for a non-technical manager (the business problem, data, the analytics solution, implications and recommendations)

Business problem

Data

The Analytics Solution

Implications

Recommendations:

Table of Contents

[**Eidesstattliche Versicherung** 0](#_Toc29827654)

[**Executive Summary** 1](#_Toc29827655)

[**List of Figures** 3](#_Toc29827656)

[**List of Tables** 4](#_Toc29827657)

[1. Data Collection and Preparation 5](#_Toc29827658)

[1.1 Bike Data 5](#_Toc29827659)

[1.2Weather Data 5](#_Toc29827660)

[2. Descriptive Analytics 5](#_Toc29827661)

[3. Cluster Analysis 5](#_Toc29827662)

[4. Predictive Analytics 5](#_Toc29827663)

[5. Discussion & Outlook 5](#_Toc29827664)

[A. Appendix 5](#_Toc29827665)

[A.1 Division Of Group Work 5](#_Toc29827666)

**List of Figures**

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

**List of Tables**

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

# Data Collection and Preparation

## Bike Data

The underlying bike data set was downloaded on the Deutsche Bahn website. For descriptive analysis and clustering the Opendata\_Booking\_Call\_a\_Bike file was relevant. This file contains all recorded Call-A-Bike Bookings of bicycles in Germany. Each booking includes information such as the booking date, start time, end time and city.

The file "Hackathon\_Booking\_Call\_a\_Bike" also contains all booking transactions and carries additional information such as the X and Y coordinates of a station. This information would be useful for analyzing the distances travelled in a trip. However, this file only contains data up to summer 2016, and therefore it does not meet the selection criteria (2015 and 2016) of the task.

Data preparation included limiting the dataset to 2015 and 2016 and filtering on Frankfurt only. Null values were found in columns, such as the Rental\_Zone or the identification keys. Dropping the null values deleted the entire row. In order to be able to measure booking frequencies consistently, the null values were not deleted here, otherwise individual transactions would not have been considered at all.

In the descriptive analysis, the booking frequencies for the booking station at the main station are considered. This requires the deletion of the null values because rows with null values are not analyzable with respect to the station.

## 1.2Weather Data

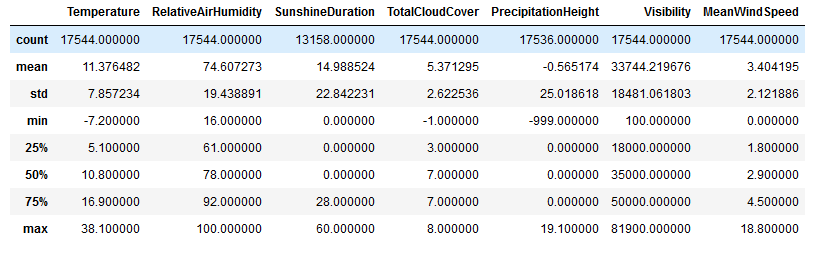
The source of the weather data is Deutscher Wetterdienst (DWD). We chose historical data from one weather station in Frankfurt on an hourly basis. The historical data is already checked by the DWD and has errors marked and described. Because the time in the data is listed in Coordinated Universal Time (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. After uploading all available datasets and joining them, we cleaned the data and handled error values. 

Figure Dataset with Error values and missing data

There are still error values in the dataset. E.g. the mean precipitation height -0.57 is in terms of content an unrealistic number, Fig. 1 shows. Therefore we assigned numerical error values (-999) to 'NaN' so that statistical methods are not biased, what Fig.2 illustrates.

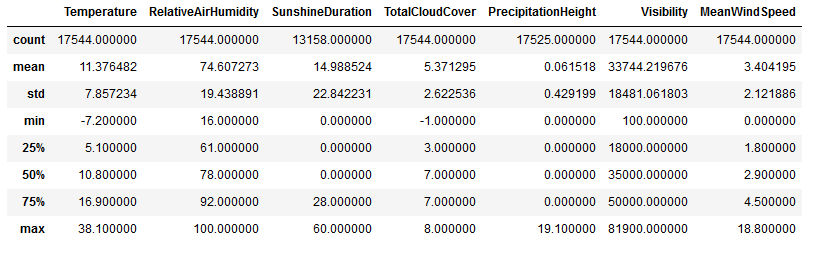
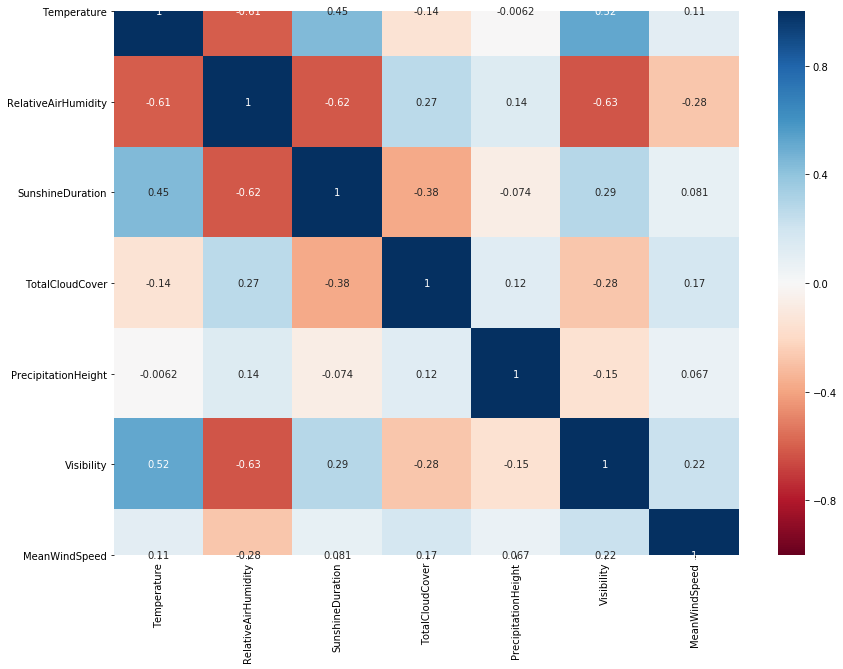


Figure Dataset without errors with missing data

Fig 2 shows, that only sunshine and precipitation values are missing. Precipitation is only missing 19 values. These will be filled via forward fill, because the missing values are < 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 typically does not shine at night in Frankfurt.

To get a feeling for the data some random days were plotted, furthermore a Pearson Correlation was implemented, as seen in Fig.3. Due to a high correlation (>0.2) with Temperature and more variable 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 correlated with each other. This makes all our future models more precise



Finally, we joined weather data with bike data. The time of the bike data was changed to UTC and the minutes where cut to hours, to make sure the right weather is joined with the right Bike data.

The final dataset consists of Temperature, Total Cloud Cover, Precipitation Height, Mean Wind Speed, Month, Weekday, Hour, amount of bookings.The dataset was aggregated to a 2h,6h, and 24h horizon for the Predictive Analysis later, and saved separately.

# Descriptive Analytics

The descriptive analysis was about identifying demand patterns in the data set. We have taken different approaches in this regard.

On the one hand, booking frequencies were displayed by days, months and years.

The main finding was that bookings continued to increase steadily until August and then declined steadily until January. Aggregated by seasons, it can be seen that bookings are most common in summer and lowest in winter. In spring and autumn, the booking figures are almost the same. Furthermore, the booking figures for 2016 are significantly higher than for 2015. With regard to the days, it can be stated that bookings are approximately the same during the week.

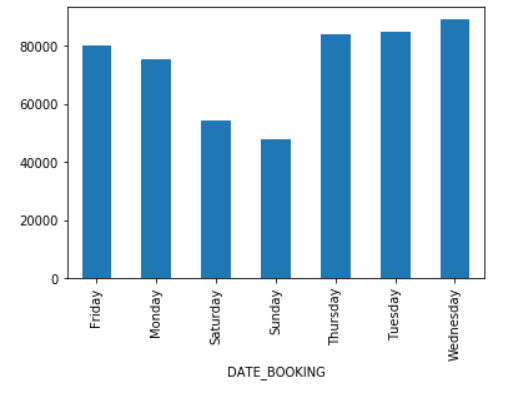


Figure 3 Booking Amounts per Week day

With exception of weekend days, significantly fewer bookings are made. Sunday is the minimum where most people don't have to work. On average, the booking time at weekends is significantly higher than during the week.

Bicycles could be used in the week to drive to and from work.

On weekend the bikes may be used for private rides without a defined start- and final destination or without time pressure.

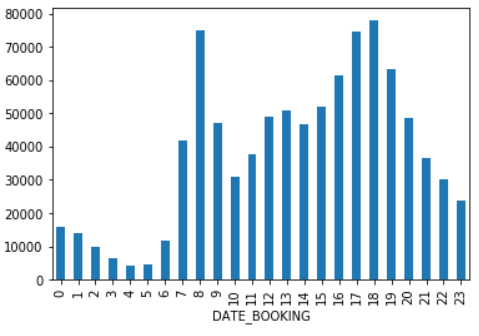


Figure 4 booking amounts per hour

If you look at the booking time depending on the hour, you can see the following.

in the morning (7.00 am - 9.00 am) and in the evening (4.00 pm - 7.00 pm), booking rates are on top

regarding the normal working hours of employees, keyword 'rushhour', one might assume that many customers rent bicycles to get to work and come back.

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 more different times

The following patterns can be seen when visualizing the average booking period.

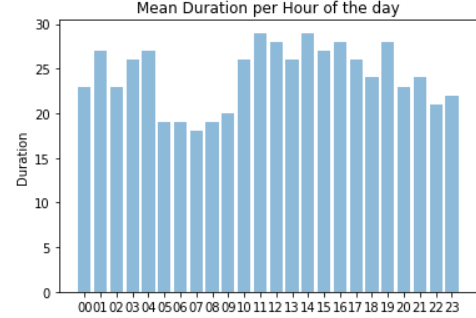


Figure 5Mean Booking duration per hour

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.

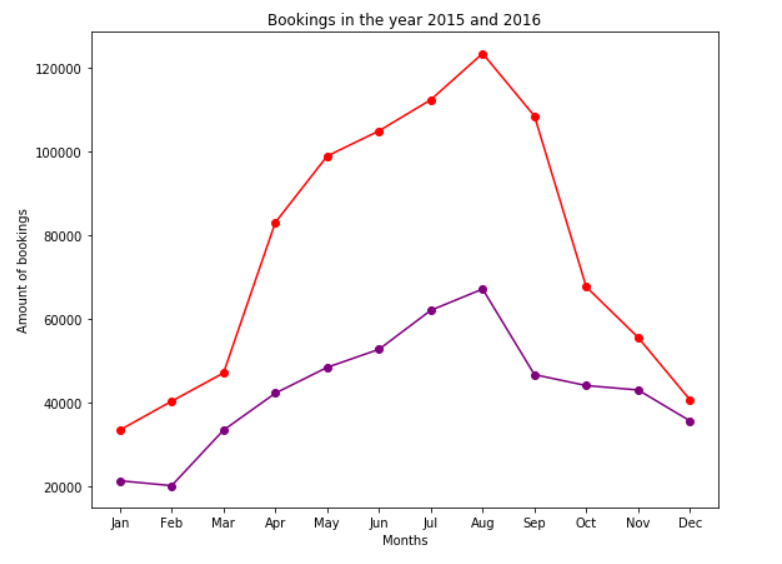
If the journey time lasts longer, customers choose a different mode of Transport

Figure 6 Booking amounts per month for two years

# Cluster Analysis

When clustering, we try to classify data into specific patterns.

HIerzu we used the K Means++ algorithm, Hierarchical CLustering and Gaussian Mixture algorithm. If the data sets are too large, hierachian clustering cannot be used and the K Means ++ algorithm has been applied.

Outliers have been removed for meaningful clustering.

We differentiated between the following attributes: Station, Customer and Trip.

The Elbogen method was used before the cluster was created to identify an approximation to the optimal number of clusters.

For each of the attributes, the documentation details the most meaningful cluster.

Stations:

The first cluster for the stations 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 this station during the selected period. Casual customers are correspondingly those with less than 100 bookings.

The result of clustering shows that there are more stations with few regular customers and occasional customers, then stations with several occasional customers and few regular customers. Other clusters have a higher variance, for example from many casual customers with few regular customers or several to many regular customers with few to several casual customers

In addition, when clustering the number of customers compared to the number of bookings of the stations, it was found that the variance increases as the values increase. There is a cluster for a few bookings and customers, one for multiple bookings and customers and in the outermost area few stations that have few customers and many bookings, or on the other hand many customers and several bookings.

Customers:

Customer clustering in terms of average time and number of bookings showed that a large proportion of customers have few bookings at a low average time. Furthermore, the customers with the most bookings have a low average time, while those with the fewest bookings have long average times.

Trip:

When clustering the bookings, it was shown that most trips were made at temperatures between 10 and 20 degrees Celcius. There will be fewer bookings the colder or the warmer it gets.

# Predictive Analytics

For predicting the bike rental demand for Frankfurt in 2015 and 2016 we abstracted a feature set out of the data, that we would use to induct different models, that we tested and evaluated. The target variable was the total bike rental demand for Frankfurt in a specific hour. As input features for our hypothesis function we chose the selected set of features that has been described in section 1.2.

Linear Model

As a first model we ran a linear regression on the dataset to predict the bike rental demand. 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 is underfitted. That result was expectable, because not all of 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 linear regression should be discarded for predicting the bike rental demand and we need another approach.

Polynomial Model

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 features 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 1: Training and Validation Error for polynomial model

Figure 1 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 RMSE of 47.2 bikes per hour and an R²-Score of 51%.

Polynomial Model + LASSO

With these results 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 X with a RMSE of XX and an R²-Score of XX % (see Figure X).

Normalizing

We also implemented different options of normalizing the input features. With this we checked, whether the performance of the prediction would improve, when the input features are all normalized. In our models, the performance did not improve when the parameters were normalized.

Comparing and evaluating

After comparison of all results our best predictive model was the polynomial regression model with LASSO regularization and without normalized input parameters. It performed at the level X.

# Discussion & Outlook

Summary of the results

The presented work shows that [hier kommt eine kurze Zusammenfassung was in den oberenTeilen steht]

What the company can do with our results

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 potential customer segments. In addition, the prediction values can be used to calculate how many bikes are needed overall in Frankfurt. Thus, these resources can be managed optimally.

[hier kommen dann die konkreten benefits basierend auf unseren ergebnissen]

Critical reflection

The test metrics for the prediction models showed that our prediction is not very precise. Having a high RMSE(XX) in our result must also be interpreted in context with the high variance (XX) of the target feature distribution data.

In the area of clustering and descriptive analysis, the following aspect should be noted: 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 do not measure the actual demand. Consequently, if there were not enough bicycles at a location, potential customers could not rent bicycles, but this is not depictured in our data.

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.

Es gibt noch weitere Aspekte, die die Ergebnisse verfälschen können. Zum einen haben wir das Risiko von unvollständigen Daten. Das kommt in den Datenverarbeitungsteil, oder?

5.4 Possible further research

[Was könnte man im bereich descriptive / clustering noch machen?]

If one provided the terminal stations for the entire period (not only for the year XXX), a more complex model could be developed, considering the incoming and outgoing bicycles. This increases the accuracy and the benefit for the company.

In addition, a Time Series Analysis could be carried out to clearly show variations in day/night or season (spring, summer, autumn, winter) -> What is the benefit?

If the prediction is carried out with our 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.

5.5 Concluding business recommendations

Was sind auf Basis der vorher erarbeiteten Aspekte unsere Empfehlungen?r

# Appendix

## A.1 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.

Nastarans Anforderungen:

The report should be written clearly and professionally and include the following sections:

VI

1. Cover page with informative title, team number and member names

2. One-page executive summary: summarizes the entire report for a non-technical manager (the

business problem, data, the analytics solution, implications and recommendations)

3. Detailed report:

(a) Problem description (business goal and data mining goal)

(b) Data description

(c) Brief data preparation details (how your data were created from the raw data) and key charts.

Details can be provided in an Appendix.

(d) Data analytics: Analytical methods applied (with su\_cient detail and screenshots; use Appendix

if needed) and appropriate performance evaluation (proper choice of measures, benchmarking).

(e) Conclusions (advantages and limitations) and business recommendations

**3 Description of Tasks**

1. **Data Collection and Preparation**: You have been provided with a full dataset of bike sharing rentals. Select the city and year you have been allocated and clean your dataset for use in later stages of your project. To obtain Weather data access the open data portal of the German Weather Service (DWD).

2. **Descriptive Analytics**: Analyze the bike rental demand patterns for the relevant one-year period and city (please check carefully which city your team has been allocated). Specifically show how rental patterns (such as start time, trip length, start and end location) for the given sample varies on a seasonal, weekly and daily level. Give possible reasons for the observed patterns.

3. **Cluster Analysis**: Based on the bike rental demand patterns, can you identify clusters of trip types and/or customer types? How would you label these clusters? Can you cluster the locations based on their demand pattern?

4. **Predictive Analytics**: Develop a prediction model that predicts bike rental demand as a function of suitable features available in or derived from the datasets. **–** Why did you choose a specific regression type? Clearly justify your choice of regression method and describe its advantages over other methods.

**– How good is your model? Evaluate your model’s performance and comment on its shortfalls.**

**–** Show how you model’s performance varies as you increase or decrease temporal resolution for the following period length:1h, 2h, 6h, 24h.

**–** How could the model be improved further? Explain some of the improvement levers that you might focus on in a follow-up project.

5. **Discussion & Outlook**: Discuss the implications of your results for the fleet operator. Which further analysis would you consider useful and could be conducted on the given dataset?