**AUFGABENBESCHREIBUNG**

Main deliverable is a 5-page project report (excl. executive summary, figures, references and appendices) [.pdf] via ILIAS (deadline 21.07 12:00). Upload Python [annotated Jupyter notebook (.ipynb)

Report details the team project, from the business problem through the data science problem and solution, to recommendations and practical relevance.

As presentation is a key component of a successful data science project, we will consider it in our evaluation.

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

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 science 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 sufficient detail and screenshots; use Appendix if needed) and appropriate performance evaluation (proper choice of measures, benchmarking)

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

**TITELSEITE**

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)

Team Name: Error 404: Group not found

Supervisor: Univ.-Prof. Dr. Wolfgang Ketter  
  
Co-Supervisor: Karsten Schroer

Department of Information Systems for Sustainable Society

Faculty of Management, Economics and Social Sciences

**FORMATIERUNG**

**GRAFIKEN, INHALTSVERZEICHNIS, ANHANG ETC.**

**TABLE OF CONTENT**

**1 Executive Summary . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1  
2 Data Collection and Preparation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1**  
 2.1. Bike Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1   
 2.2. Weather Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1

2.3. Station Location Dataset . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 1  
**3 Descriptive Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2**   
 3.1. Temporal Demand and Seasonality . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2   
 3.2. Geographical Demand . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2

3.3. Key Performance Indicators . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 2   
**4 Predictive Analysis . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**

4.1. Feature Engineering . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3   
 4.2. Model Building . . . . . . . . . . . .. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

4.3. Model Evaluation . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3

4.4. Outlook . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3 **5 Conclusion and Recommendations . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 3**   
**A Appendix . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 4**

**EXECUTIVE SUMMARY (1)**

Summarize entire report for a non-technical manager (business problem, data, analytics solution, implications and recommendations)

[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.

[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 to 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. To summarize, a 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 decision 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 human 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.

[DATA] We are using four different data sets. The first one 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 whether things like temperature and humidity have an impact of 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 station are. The fourth data set is constructed of twelve raw data sets and provides information about the customer age and gender which is relevant for our KPIs.

[ANALYTICAL SOLUTION] Our analyse results in….

[IMPLICATIONS] Implications we faced are false data tuples. We found some in the data set which holds the trip information which has a negative duration or a duration over a long period of time. As this is clearly not a valid data tuple, we eliminated these on. We did filter the data so that every trip needs to be at least two minutes long and eight hours at is maximum.

[RECOMMENDATIONS] We further recommend to increase 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.

**DATA COLLECTION & PREPARATION (2)**

Cleaning of datasets for use in later analysis stages

[2.1 BASICS] Our data preparation for all three data sets is focused on three things. Firstly, we check for null values, secondly, we check for duplicates and lastly we drop features we do not need for our analysis.

[2.2 BIKE DATASET] The bike data set provides us with operational raw data from “Blue Bikes” with the following attributes: start\_time (datetime), end\_time (datetime), start\_station\_id (int), end\_station\_id (int), start\_station\_name (str), end\_station\_name (str), bike\_id (int), user\_type (str).

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. Furthermore, 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. It was that the raw data set contained partially implausible data, so the shortest rental period was -1 day 23:06:07, i.e. negative, 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, (Korrektur auf richtig Wert notwendig) 1,125,149 evaluable entries remained.

[2.3 WEATHER DATASET] Secondly, we were provided with hourly weather data for Boston. The weather data ranges from the first January of 2015 to the second of January of 2020. It contains the following attributes: date\_time (“datetime”), max\_temp (float), min\_temp (float), precip (float). 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 6667 tuples. Although it should be 8,760 tuples to map a year without gaps. It can therefore be assumed that 93 hours of weather records are completely missing. Hence 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 whole 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 variable “precip\_id”, which indicates whether rain or snow was recorded for the respective period. This is the case when looking at the given values, so there is no need to adjust the weather data.

[2.4 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 the used to for visualisation to get a feel where in Boston we have docking stations. The data set contains following attributes: Number, Name, Latitude, Longitude, District, Public, Total docks. It needed to be pre-processed in multiple ways (Stationen raus die 2017 nicht da, csv file gibt ein falschen tuple).

**DESCRIPTIVE ANALYSIS (3)**

Demonstrate temporal demand patterns and seasonality.

[3.1 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. On the one hand these are in the morning between approx. 6-10 a.m. and on the other hand in the evening between approx. 3 p.m.-7 p.m. 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 school. 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 to school 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. There is hardly any need 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 the autumn, only to return to the minimum level in December. This can be explained by the prevailing weather conditions and the temperature in the respective seasons. 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. This correlation fits our assumptions very well, but only becomes meaningful if the frequencies of the temperatures that have occurred have also been checked.

Demonstrate geographical demand patterns.

[3.2 GEOGRAPHICAL DEMAND] Put geographical demand here. To precisely predict the total bike usage for the next hour 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 to many bikes result in a not valuable utilization, which we want to prevent. If we do not prevent it, we would spend 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 would potentially lead in bad user reputation as they are frustrated if they want to get a bike, but do not get one fast. Our target is to predict the demand of total bike usage for the next hour to check how many bikes there must be in store to satisfy the customer.

Define Key Performance Indicators which provides overview of current fleet operations.

[3.3 KEY PERFORMANCE INDICATORS] Put Key Indicators here. The data we focus on is from the year 2017. We are using three different data sets. The first one provides information about all trips which were made in 2017 with a bike from “Blue Bike” which includes relevant attributes like start time, end time and station names. The second data set delivers weather information which is relevant to check whether things like temperature and humidity has an impact of the amount of bike rentals and if so, how big this impact is. The third data set provides geolocation of all docking stations. This will be used to visualize where the operating docking station are.

**PREDICTIVE ANALYSIS (4)**

Forecast total system-level demand in the next hour.

[4.1 FEATURE ENGINEERING] Put feature engineering here. 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 Temperature/Rentals graphic that on warm days people are more likely to take a bike as on cold days. 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 Weekday/Rentals graphic showed a big difference between total demand on weekdays comparing to the demand on weekends. On weekdays 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. 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 proofed with the Hour/ Rentals graphic.

[4.2 MODEL BUILDING] Put model building here. 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 an unparametric 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 from 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 what means 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 variable. In addition a broad range of function can be 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 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

[4.3 MODEL EVALUATION] Model evaluation: The worst model result was from the Polynomial regression with a mean absolute error of 64.57750735912532 Bikes and a root mean squared error of 110.76415974052735 Bikes. Second best model is the Tree based regression with a mean absolute error of 50.18581764018217 Bikes and a root mean squared error of 88.179863647624 Bikes. Our best model is the KNN Regression with a mean absolute error of 47.5206055508831 Bikes and a root mean squared error of 82.48795571224487 Bikes. This is also the model we would select for deployment.

[4.4 OUTLOOK] We could improve our models by including more suitable features. In addition we could look for more or new data to train the models better. It is also an option to try some more models and look if they perform better.

**CONCLUSIONS AND RECOMMENDATIONS (5)**

**Conclusions (advantages and limitations) and business recommendation**

[CONCLUSIONS] Write conclusion here.

[RECOMMENDATIONS] We further recommend to increase 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. The mean of high operating hour is (only include the hours with the highest utilization and check the mean) however is XXX. 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 is that there should always be 2 bikes on hold at every station, so a pair of customers always have the chance to rent the bikes a station they would. If there are X 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 only X 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 people are using mobility service system. The fourth reason is that it is not the biggest investment you need to buy more bike (The higher investment would be to set up stations). Moreover, it is good to provide enough bikes on bike event as “Blue Bikes” can then be associated with this (keep vandalism in might – concert etc.).

**APPENDIX (A)**

Put Pictures and sources here.

LIST OF ALL RELEVANT VISUALISATION:

Popularity of station with heatmap