**AUFGABENBESCHREIBUNG**

**TO-DO**

**AUFGABENBESCHREIBUNG**

The main deliverable of the team assignment is a 5-page project report (excl. executive summary, figures, references and appendices) which is to be submitted as a .pdf file via ILIAS no later than the deadline 21.07 12:00. In addition to the report we also expect you to upload your Python code in the form of an 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: Lukas Tempfli (Student ID: 7367097)

Moritz Danhausen (Student ID: 7369413)

Niklas Nesseler (Student ID: 7367375)

Robin Kirch (Student ID: 7364580)

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 [HIER]**

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

This report is based on real world data, which was provided by Blue Bikes, which is an operating bike sharing firm from Boston. We are focusing on data of the year 2017 and we were provided with the task to predict the total bike usage of bikes from “Blue Bikes” for the next hour in Boston. In our project we are transforming the given data sets to our needs and apply modern machine learning techniques to solve the task. [GRUNDBESCHREIBUNG]

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. [BUSINESS PROBLEM]

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. [DATA]

Our analyse results in…. [ANALYTICAL SOLUTION]

Implications we faced were data that is probably false. We found data 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 set we eliminated these on. We did filter the data so that every trip needs to be at least five minutes long and eight hours at is highest. [IMPLICATIONS]

We further would recommend… [RECOMMENDATIONS]

* **Task 1) Data Collection & Preparation:** 
  + Cleaning of datasets for use in later analysis stages
* **Task 2) Descriptive Analysis**:
  + Demonstrate temporal demand patterns and seasonality.
  + Demonstrate geographical demand patterns.
  + Define Key Performance Indicators which provides overview of current fleet operations.
* **Task 3) Predictive Analysis**:
  + Forecast total system-level demand in the next hour.

#TODO

Popularity of station with Heatmap

**3. Detailed report:**

**(a) Problem description (business goal and data science goal)**

In the business context of smart mobility services, it is relevant to monitor the usage of the transport goods, in our case bicycle, to check whether there are enough bicycles available at each time of the day as well as for each station, so the customer is satisfied and relies more on a smart mobility solution. This is of high interest as good functioning smart mobility service reduces typical mobility issue that a big city has. These are traffic in city and place as well as pollution.

**(b) Data description**

The data on which the project relies is constructed of two different files.

Firstly, we were provided with operational raw data from “Blue Bikes” ("boston\_2017.csv") 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).

Secondly, we were provided with hourly weather data for Boston ("weather\_hourly\_boston.csv"). 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).

For our task we are focusing mainly on these data sets. However, we also use further information which was provided by the “Blue Bike” Website (378 docks) -> Stationen manuell löschen die noch nicht 2017 da sind, which is a data set which contain 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 sets contains following attributes: Number,Name,Latitude,Longitude,District,Public,Total docks.

**(c) Brief data preparation details (how your data were created from the raw data) and key charts. Details can be provided in an Appendix.**

**Was ist mit Duplikaten???**

Our data preparation for all three data sets is focused on two things. Firstly, check for null values and secondly drop features we do not need for our analysis.

**(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 recommendation**

**APPENDIX (ANHANG – QUELLENANGABEN) [HIER]**