Data Science Project

Bike Rental Analysis

Conceptual Design Report

31 October 2022

Abstract

Increasing traffic, pollution and overcrowded public transport cause major challenges for urban areas and call for new climate-friendly transport options. Public bike rental systems have gained in importance in many cities in recent years. However, modern bike rental systems bring new challenges. Depending on daytime, weekday, weather and season, the demand for bikes can vary greatly at the individual rental stations. Not least for economic reasons, the bike rental providers have a great interest in being able to estimate the respective demand at the individual locations and to be able to provide the required number of bikes.

This report is intended to describe an analysis of data about bike rental systems. This includes the presentation of the objectives, the methodology, the whole process of data collection and the data model of this project. In addition, risks related to the data, preliminary studies and initial findings from this data analysis are discussed.

Table of Contents

List of Figures

List of Tables

1 Project Objectives

The goal of this data science project is the analysis of the data of a bike rental service and to examine the dependency of daily bike rentals on various variables such as weekdays, weather situation and seasons.

The project tries to answer the following questions:

- Is there a difference in the number of daily bike rentals between workdays and non-workdays?
- Does the number of daily bike rentals depend on different weather situations?
- Does the number of daily bike rentals differ among the seasons?
- And if there is a difference among the seasons, is there a difference in the number of daily bike rentals between Spring and Fall?
- Does temperature and wind speed have an influence on the number of daily rented bikes?

The answers to these questions should help a bike rental supplier to provide the correct number of bikes at the rental stations, depending on the factors mentioned. The provision of a suitable number of bikes at the various rental stations helps the provider to increase income on days with high demand and to reduce costs on days with low demand as the effort involved in providing the bikes can be reduced.

The first phase of the project is about finding the right data set for this project. The data must then be prepared and brought into the required format. The distribution of the number of rented bikes depending on the day of the week, the weather and the season is then examined using times series plots and boxplots. In the next step, methods from inferential statistics are used to answer the stated hypothesis.

2 Methods

The data collection is examined by an external party since the data used is extracted from Kaggle (more on the data in Chapter 3). All the data and scripts are stored in Google Drive to ensure collaboration and availability.

For preprocessing and the analysis, the Google Colab environment with Jupyter Notebook is used. The project is executed with Python (v 3.7.15). The necessary packages are the following:

- Google.collab
- Pandas

- Numpy
- Matplotlib
- Scipy
- Seaborn
- Sklearn

In preprocessing data cleaning is done. New features based on the given features are computed such that further analysis can be performed.

The data is analyzed with descriptive and inferential statistics. The descriptive statistics includes univariate methods (summary statistics, contingency tables, time-series plots) and bivariate methods (correlation matrix, boxplots based on groups). For the inferential statistics, hypothesis test statistics and regression analysis are used. Furthermore, to determine the right test-statistic, qq-plots and test for normality (D'Agostino Pearson, Shapiro-Wilk test and Kolmogorow-Smirnow test) are executed.

3 Data

A publicly available data set from Kaggle is used [1]. The set is hosted by Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto [2]. The data is originally from Capital Bikeshare [3]. The data set is a two-year, daily measure of the rented bike demand. This yields a total of 731 observations. Additionally, features about the weather condition for each day are provided and information about the day are included. In Table 1 the raw data is depicted.

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600

Table 1: Raw data

The data is already well formatted. Some of the categorical variables are in a numerical form. Those are formatted to text for a more convenient understanding. Not all columns are needed and analyzed thus those are dropped. The first few lines of the final data set are shown in Table 2. For the regression analysis, also dummy variables are introduced for weather situations. For simplicity those are not depicted in Table 2.

	dteday	season	weekday	workingday	weathersit	temp	windspeed	cnt	workingday_txt	weathersit_txt	season_txt	day_txt
0	2011-01-01	1	6	0	2	0.344167	0.160446	985	No Workingday	Clouds / Mist	1_Winter	6_Sat
1	2011-01-02	1	0	0	2	0.363478	0.248539	801	No Workingday	Clouds / Mist	1_Winter	0_Sun
2	2011-01-03	1	1	1	1	0.196364	0.248309	1349	Workingday	Clear	1_Winter	1_Mon
3	2011-01-04	1	2	1	1	0.200000	0.160296	1562	Workingday	Clear	1_Winter	2_Tue
4	2011-01-05	1	3	1	1	0.226957	0.186900	1600	Workingday	Clear	1_Winter	3_Wed

Table 2: Final data set

The following figures 1 to 3 give an univariate overview of selected features in time series plots.

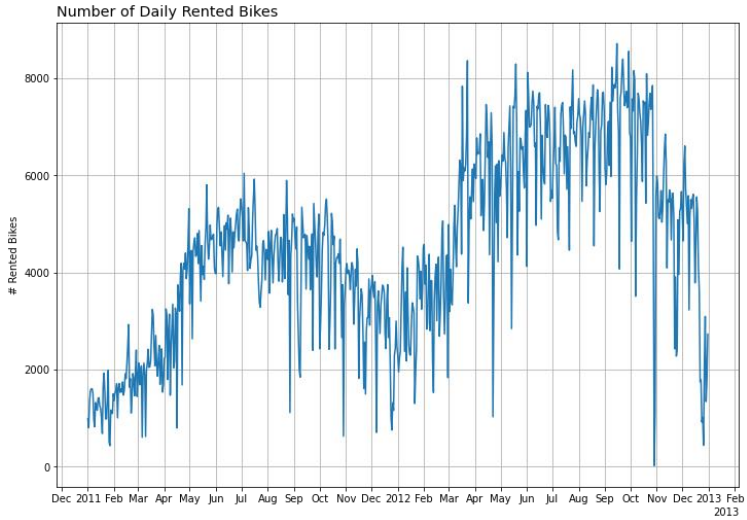

Figure 1: Plot daily rented bikes

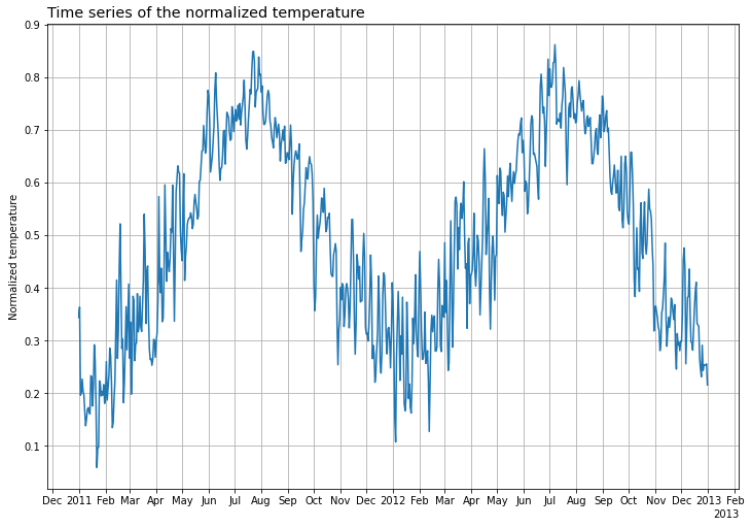

Figure 2: Plot normalized temperatures

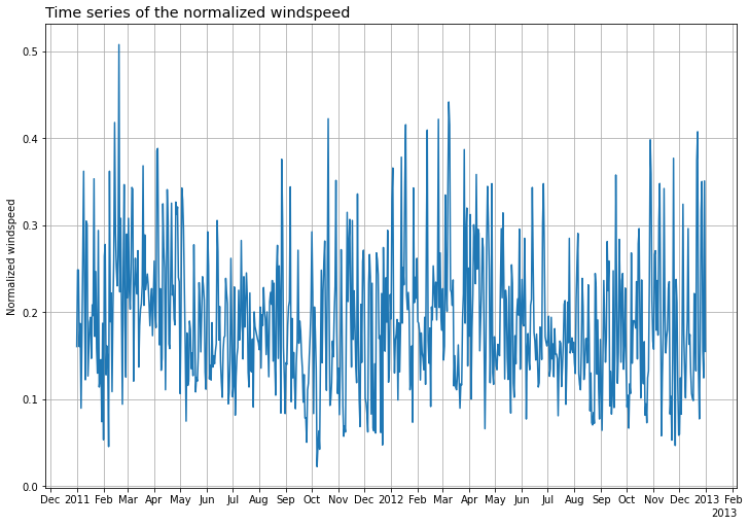

Figure 3: Plot normalized wind speed

4 Metadata

The meta data description relies on the description provided on Kaggle [1]. Table 3 presents the description of the variables. A copy of the metadata is also included in the jupyter notebook for comprehensive purposes.

Variable	Type	Description
dteday	date	hourly date + timestamp
season	integer	1 = spring, 2 = summer, 3 = fall, 4 = winter
yr	integer	1 = 2011, 2 = 2012
mnth	integer	Month, 1 to 12
holiday	integer	if day is holiday is 1, otherwise is 0
weekday	integer	0 = Sunday, ... , 6 = Saturday
working day	integer	if day is neither weekend nor holiday is 1, otherwise is 0
weather sit	integer	1: Clear, Few clouds, Partly cloudy, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Ra Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	float	normalized temperature in Celsius
atemp	float	normalized "feels like" temperature in Celsius
hum	float	normalized relative humidity
wind speed	float	normalized windspeed
casual	integer	number of non-registered user rentals initiated
registered	integer	number of registered user rentals initiated
cnt	integer	number of total rentals

Table 3: Metadata

5 Data Quality

The data quality determines the quality of the analysis and the corresponding conclusions. Wrong or bad data yields wrong decisions. Thus, five quality measures are determined [4] and applied to the data set to estimate the quality. Those are the following:

Accuracy: This measures if the data provided is correct or not. Since there is no information about the collection process of the data, this quality is difficult to assess. But what can be said is that some data change over time (i.e., the temperature rises until noon or to summarize the weather situation in one state for the whole day might be inaccurate). This aspect might put some features to inaccuracy depending on the variety during the day.

Completeness: This characteristic is determined by how many values are missing. If an information is rarely available and can not be collected (or worst case just by a certain group) the feature should not be included. In this data set information is given and no missing values are detected.

Reliability: Answers the question if the data contradict other trusted resources. A brief research did not yield a comparable resource. But the original provider is a bike sharing company, therefore the data should reflect reality. The outcome of the analysis will be discussed to verify its reliability.

Relevancy: Not all data contribute to an analysis and collecting data costs time and money. Thus, data should be expected as relevant enough to collect. In this case, the features seem to be well chosen.

Timeliness: The question is if the data is up to date and reflects the presence state. The provided data is from 2011 and 2012 thus 10 years ago. The features in the data set seem to be time consistent (i.e., temperature, workdays etc.) however, if there is a hidden feature which changed over time the data cannot be considered as up to date. So, it would be better to get a more recent data set.

6 Data Flow

The data flow in Figure 4 shows how the data is generated, collected, prepared and analyzed. First the data is generated when a user is renting a bike. Depending on the location and the time the supplier collects and saves data like date, season, weekday, (non-)workday, weather situation, temperature, wind speed. In addition, the supplier collects the number of daily rented bikes. All the data is stored in the database of the bike supplier.

In a next step the bike supplier provides the date for further analysis. The data scientist imports the data to Google Drive and loads the data into the Jupyter Notebook and transforms it into the Python dataframe. Now the data scientist does the actual analysis which results in several plots and test results.

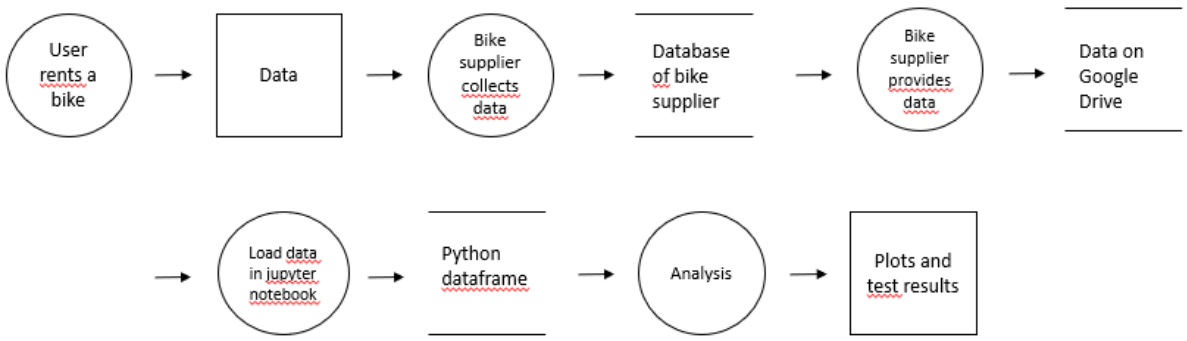

Figure 4: Data flow

7 Data Model

Conceptual Data Model: The daily demand is the only entity in this model. The entity is shown in Figure 5. The Daily Bike Demand has several attributes derived from the date and external factors.

Figure 5: Conceptual data model

Logical Data Model: The Logical Data Model defines the datatypes and the relationship between several entities. The added data types in the single entity are shown in Figure 6. Additional entities on a daily basis (Primary Key is the date) could be added.

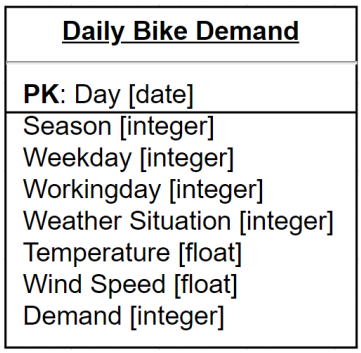

Figure 6: Logical data model

Physical Data Model: In an optimal state the bike sharing company deploys the data directly in Google Drive (main storage). With Python in a jupyter Notebook from Google Collab the data is then further processed (cleaning / enhancing / enriched) and stored in Google Drive as csv for further analysis. This process is depicted in Figure 7. The process highly relies on the cloud services from Google but this leads to basic requirements in hardware. Only a laptop is needed to access the whole infrastructure.

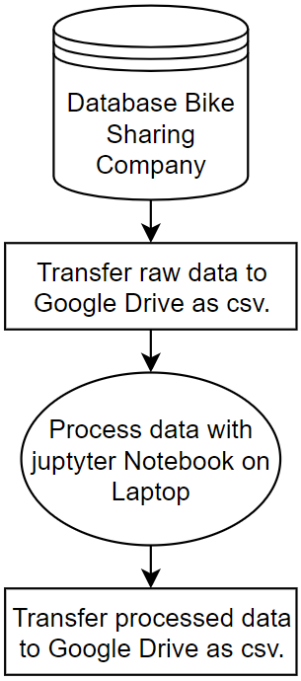

Figure 7: Physical data model

8 Risks

In data quality the timeliness and accuracy aspect might be an issue. Thus, it might lead to wrong statements and therefore to wrong business decisions. If the predicted demand is far away from the reality, there is a risk of unsatisfied customers (to few bikes provided) and loss of revenue. To challenge this issue, periodic updates of the model with the newest data should be done. The statements should be verified with the business.

In this case a third party provided the data. This already is a risk itself since there is no information about the collection nor any transformation to the raw data. The analysis provides insights based on this data and it is not sure if the results apply to all rental bike suppliers. To tackle this drawback the results must be challenged by experts from this business. To get feedback increases the complexity and is time consuming.

The whole infrastructure is based on Google (see Chapter 7). If Google turns off the services or is not reachable, it results in a delay of the analysis. Local copies and backups can be done to ensure constant service and availability.

9 Preliminary Studies

9.1 Descriptive Statistics

Figure 8 shows a time series plot with box plots of the daily demand for each month. More bikes seem to be rented in warmer seasons. December 2012 seems to have some stringer variance.

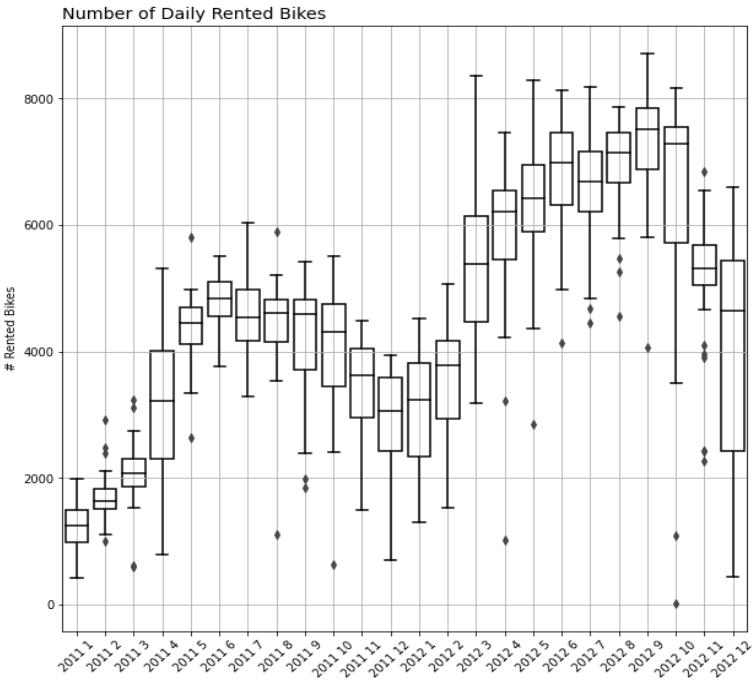

Figure 8: Boxplot daily rented bikes

Figure 9 gives an impression of the variance of daily bike rentals between workdays and non-workdays. There seems to be no big difference between workdays and non-workdays.

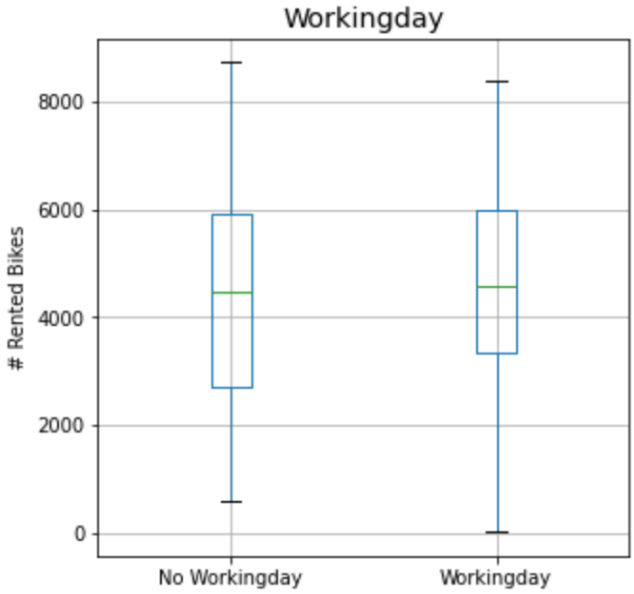

Figure 9: Boxplot rented bikes workdays and non-workdays

Figure 10 gives an impression of the variance of daily bike rentals depending on the weather situation. There seems to be a lower number of daily bike rentals when weather conditions are worse.

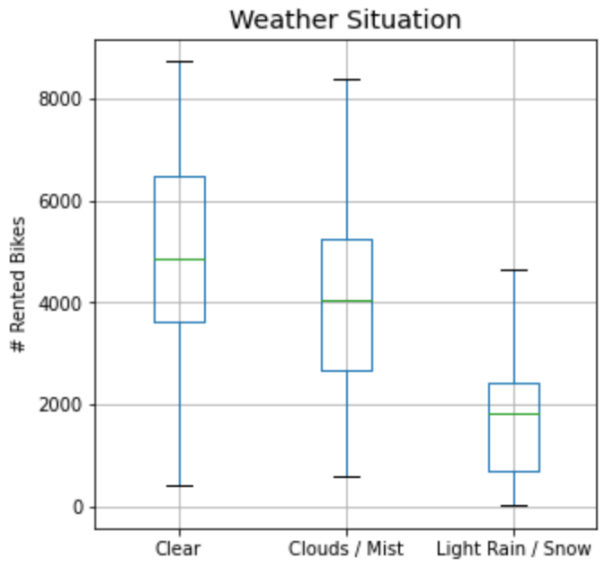

Figure 10: Boxplot daily rented bikes under differing weather conditions

Figure 11 gives an impression of the variance of daily bike rentals depending on seasons. There seems to be a lower number of daily bike rentals in winter.

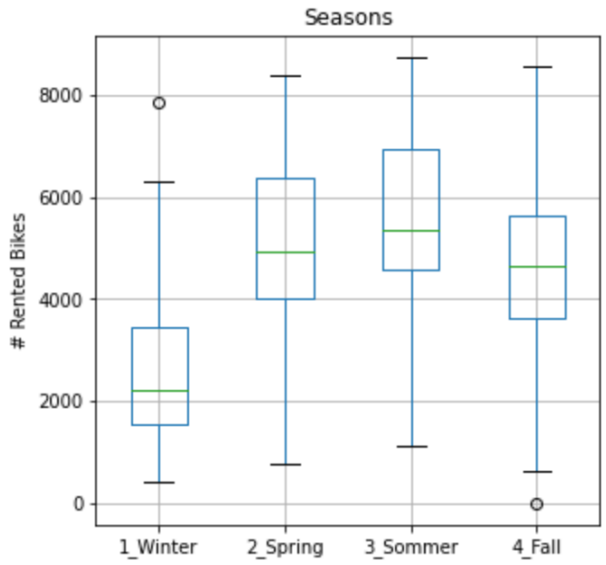

Figure 11: Boxplot daily rented bikes depending on seasons

Figure 12 gives an impression of the variance of daily bike rentals depending on the weekday. There seems to be no significant variance in the number of daily bike rentals depending on the weekday.

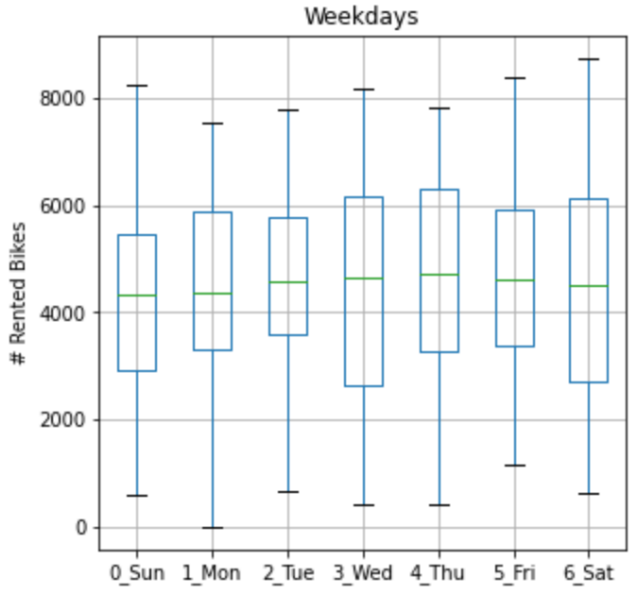

Figure 12: Boxplot daily rented bikes depending on weekday

9.2 Inferential Statistics

Hypothesis 1: On average there are more rentals on workdays than on non-workdays

The examination of hypothesis 1 with Q-Q-Plots shows that there are on average more rentals on workdays than on non-workdays.

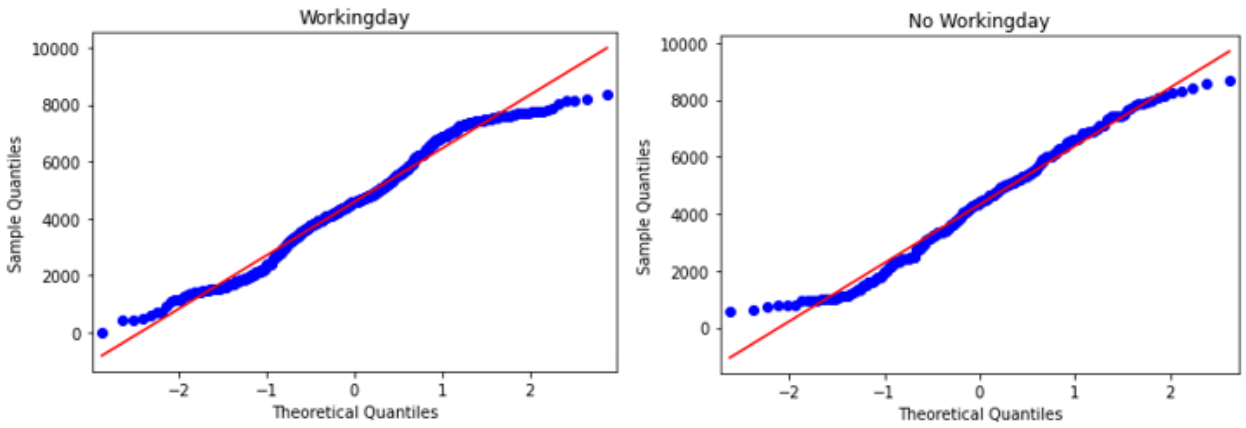

Figure 13: Q-Q-Plot workdays and non-workdays

	Workingday	Non-workingday
D'Agostino-Pearson	0.000	0.000
Shapiro	0.000	0.000
Smirnov	0.000	0.000

Table 4: Test normal distribution workday and non-workday

Since the data is not normally distributed and unpaired, we use the one-sided Mann-Whitney U Test and get the following p-Value:

0.059 > 0.05

Thus, the Null-Hypothesis cannot be rejected. There is no difference between workdays and non-workdays.

Hypothesis 2: The number of daily rented bikes differ for different weather situations

The examination of hypothesis 2 with Q-Q-Plots shows that the number of daily rented bikes differ for different weather situations.

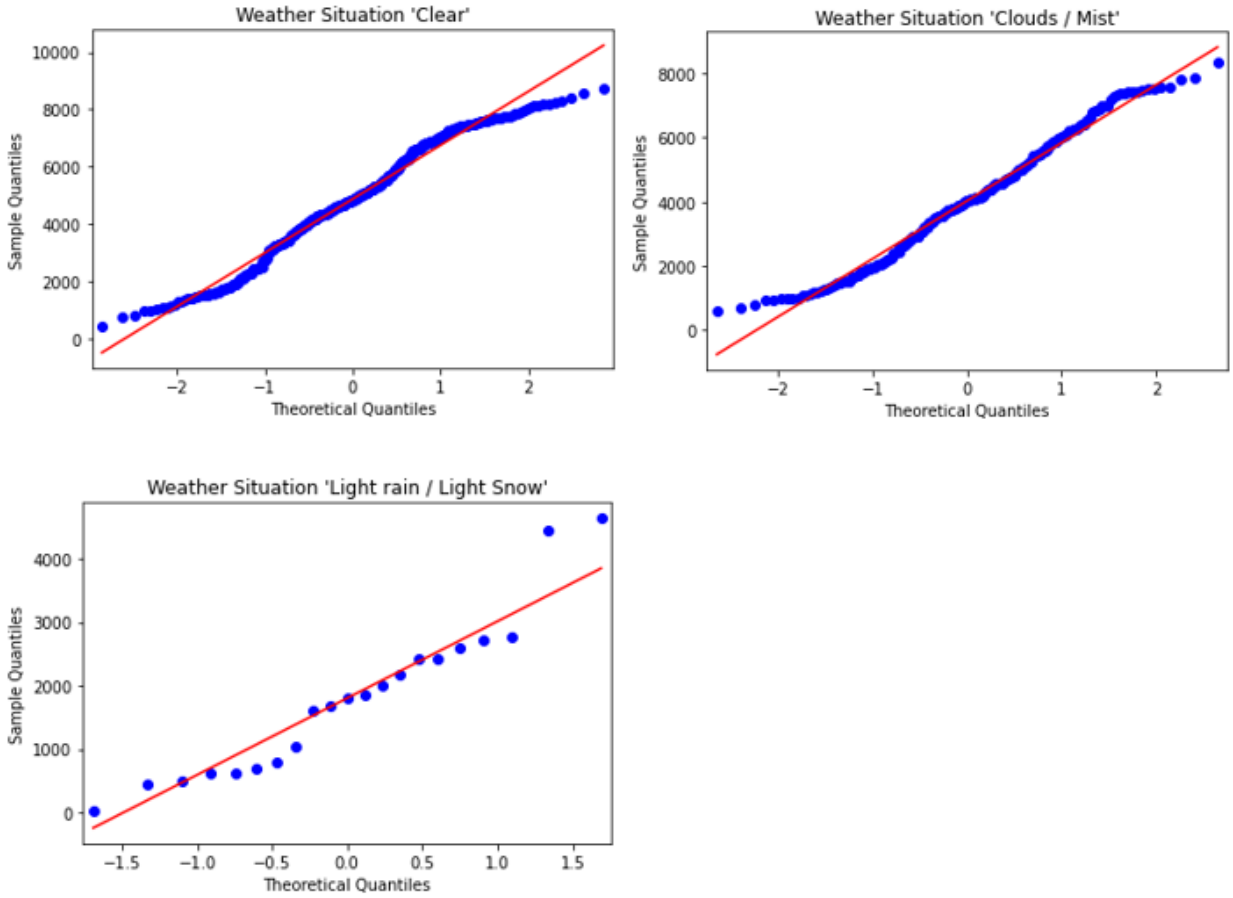

Figure 14: Q-Q-Plots Weather situation

	Clear	Clouds / Mist	Light rain / lignt snow
D'Agostino-Pearson	0.000	0.002	0.210
Shapiro	0.000	0.001	0.092
Smirnov	0.000	0.000	0.000

Table 5: Test normal distribution weather condition

Since the data is not normally distributed and unpaired, we use the Kruskal-Wallis-test which gives the following p-Value:

0.000 < 0.05

With 95% certainty the Null-Hypothesis can be rejected. At least one weather situation differs in the number of daily rentals. Thus, the number of daily rented bikes depends on the weather situation.

Hypothesis 3.1 / 3.2: The number of daily rented bikes differ among the seasons

The examination of hypothesis 3.1 and 3.2 with Q-Q-Plots shows that the number of daily rented bikes differ among the seasons.

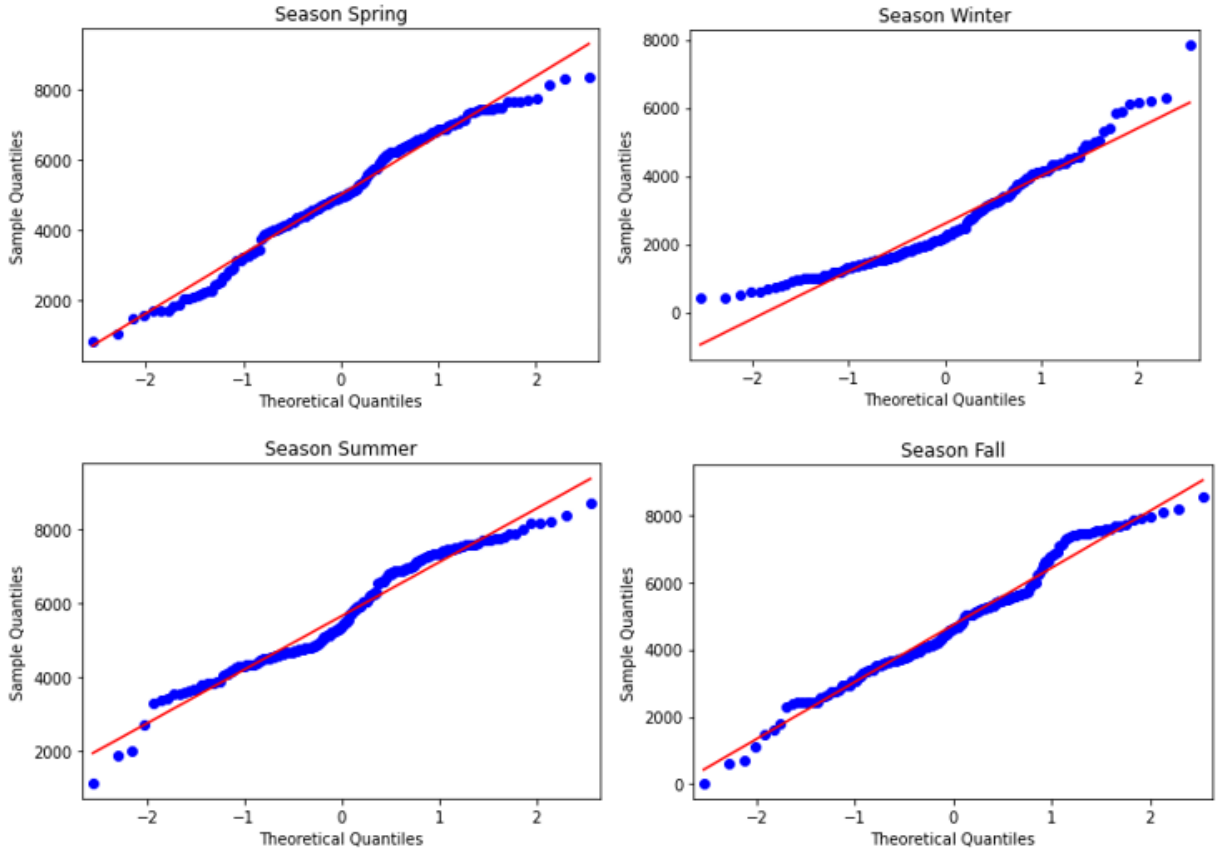

Figure 15: Q-Q-Plots seasons

	Winter	Spring	Summer	Fall
D'Agostino-Pearson	0.000	0.028	0.100	0.691
Shapiro	0.000	0.008	0.000	0.018
Smirnov	0.000	0.000	0.000	0.000

Table 6: Test normal distribution season

Since the data is not normally distributed and unpaired, we use the Kruskal-Wallis-test for hypothesis 3.1 which gives the following p-Value:

0.000 < 0.05

Thus, the Null-Hypothesis can be rejected with 95% certainty. At least one season differs in the number of daily rentals.

	Spring	Fall
D'Agostino-Pearson	0.028	0.691
Shapiro	0.008	0.018
Smirnov	0.000	0.000

Table 7: Test normal distribution spring and fall

Since the data is not normally distributed and unpaired, we use the one-sided Mann-Whitney U test for hypothesis 3.2 which gives the following p-Value:

0.045 < 0.05

Thus, the Null-Hypothesis can be rejected with 95% certainty. Spring has (weak) significant higher daily rentals than Fall.

Regression Test: Temperature and Wind Speed have an influence on daily rented bikes

With the Regression test we check if temperature and wind speed have an influence on daily rented bikes. The result shows that there is quite a strong correlation between rented bikes and temperature. The correlation of wind speed is negligible.

However, as the number of rented bikes is not normally distributed, you have to be cautious with the interpretation of the results of this regression.

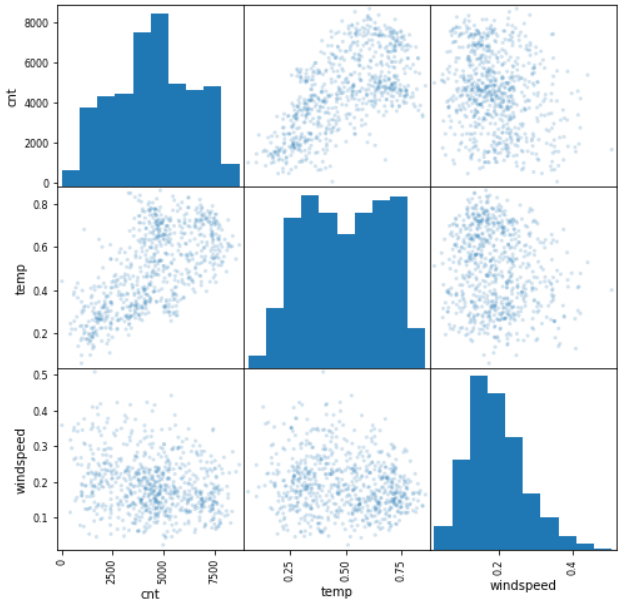

Figure 16: Plot of correlation matrix

	rented bikes (cnt)	temperature	windspeed
rented bikes (cnt)	1.000		
temperature	0.627	1.000	
windspeed	-0.235	-0.158	1.000

Table 8: Correlation matrix

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    cnt   R-squared:                       0.413
Model:                            OLS   Adj. R-squared:                  0.411
Method:                 Least Squares   F-statistic:                     255.6
Date:                Tue, 20 Sep 2022   Prob (F-statistic):           7.99e-85
Time:                        19:36:47   Log-Likelihood:                -6375.2
No. Observations:                 731   AIC:                         1.276e+04
Df Residuals:                     728   BIC:                         1.277e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1991.0459    225.962      8.811      0.000    1547.432    2434.660
temp        6408.5358    304.443     21.050      0.000    5810.844    7006.227
windspeed  -3472.1053    719.099     -4.828      0.000   -4883.861   -2060.350
==============================================================================
Omnibus:                       25.144   Durbin-Watson:                   0.467
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               15.379
Skew:                           0.206   Prob(JB):                     0.000458
Kurtosis:                       2.422   Cond. No.                         15.3
==============================================================================
```

Figure 17: Regression test

41.1% of the variance of the target (number of rented bikes) can be explained with this simple model. This percentage is quite high. Temperature alone explains 39% of the variance (simple test with only temperature). Thus, the result shows that temperature has an influence on the demand, which is not quite surprising.

10 Conclusions

The goal of this project was to identify which variables influence the daily number of rented bikes with the intention to provide bike rentals with accurate information about the number of bikes that must be provided. Our model considered various variables like weekday, weather, season, temperature, and wind speed.

The results of the data analysis showed that the demand seems not be dependent on the weekday. Obviously, the number of demanded bikes is not significantly different on workdays than and non-workdays. However, the demand seems to vary depending on weather situations and seasons. When weather is friendly and not rainy or snowy, the number of daily rented bikes

increases. The same finding could be observed comparing demands among the seasons. Obviously in warmer seasons like Spring, Summer and Fall there is a higher demand for bikes than in winter. Furthermore, the data analysis showed that there seems to be a significantly higher demand in Spring than in Fall.

In further analysis it was examined if temperature or wind speed has the higher impact on the differing demand among the weather conditions and the seasons. The results suggest that temperature seems to explain the biggest part of the variance in demand and daily rented bikes.

Although the referring data was of good quality, a more sophisticated model with more variables should be applied to investigate the other factors which influence the demand the most (e.g., Daytime). Moreover, many variables seem to be correlated (e.g., higher / lower temperature in certain seasons and under certain weather conditions).

References and Bibliography

[1] Kaggle 2022, Bike Sharing Demand, accessed 18.09.2022 https://www.kaggle.com/datasets/imakash3011/rental-bike-sharing

[2] Fanaee-T, Hadi, and Gama, Joao, Event labeling combining ensemble detectors and background knowledge, Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg.

[3] Capital Bikeshare 2022, Website, accessed 29.10.2022 https://ride.capitalbikeshare.com/system-data

[4] Precisely 2022, Website, accessed 29.10.2022 https://www.precisely.com/blog/data-quality/5-characteristics-of-data-quality