Introduction to Machine Learning and Data Mining Project 1

Preparation of The Data

Project Report

Wojciech Ciok s182125
Kacper Żyła s182134

Contents

Chapter 1

Introduction

This project was executed by two students, Mr Wojciech Ciok and Mr Kacper Żyła. This is the result of assignment given under the course Introduction to Machine Learning and Data Mining 2018/2019. The data we chose to work with[1] is information collected by a bike sharing system in Washington D.C., USA. As it is the first of three projects, it's focus is mainly on description, understanding and preparation of the data. Therefore this paper includes careful and precise analysis of the dataset which will be crucial in the following projects.

Bike sharing systems are the new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Chapter 2

Data Set Description

2.1 Problem of Interest and Data Origin

Our data set consists of data generated by a bike sharing system in years 2011-2012 in Washington D.C which is publicly available in http://capitalbikeshare.com/system-data combined with weather data from this period collected from http://www.freemeteo.com. This was done by Hadi Fanaee-T and João Gama in paper "Event labeling combining ensemble detectors and background knowledge"[1]

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of important events in the city could be detected via monitoring these data. Since the data was combined with weather and seasonal information it is expected to detect interesting correlations and dependencies within the data.

2.2 Previous Data Application

This data set was previously used by scientists in University of Porto to preform event labeling. The point of the research was to propose a model that could preform event labeling without the need for human experts. Throughout the project many different machine learning models were used as the point of proposed solution was to combine results given by different detectors. The proposed model reached around 70% accuracy .At the end scientists came to the conclusion that it is possible to use trained machine learning models and background knowledge to label events without human experts and that bike rental data is highly correlated with environmental and periodicity settings such as temperature, hour of the day or month.

2.3 Machine Learning Modeling Aim

In this section we will explain our goals and hopes for the course of the experiments specified for every technique.

2.3.1 Classification

In the classification task we are planning to predict season and weather class based on temperature, wind speed, humidity and number of rented bikes on a given day.

2.3.2 Regression

For this method it would be interesting to predict the number of rented bikes based on weather conditions and if the given day was a working day or not. Especially useful would be season, temperature, atmospheric conditions, humidity, wind speed and holiday or weekend information,

2.3.3 Clustering

For the clustering task we think that relevant attributes might include season, weather or temperature as in our opinion there might be visible similarities between data points in, for example the same season.

2.3.4 Association Mining

For the association mining relevant information might be whether a given day was a working day or not. We might also be able to observe association between season and weather conditions.

2.3.5 Anomaly Detection

Since we have information about weather and date, it would be interesting to observe anomalies in rented bikes numbers. Example of such anomaly could be Hurricane Sandy which was in Washington D.C. in 2012-10-30.

Chapter 3

Explanation of the Attributes

In this section we will explain each attribute from our data set, give a brief overview and highlight problems if any.

- instant
  Index of the given record.
- dteday (discrete, nominal)
  Date in format year-month-day.
- season (discrete, nominal)
  Current season, takes values from1 to 4.
  1: spring, 2: summer, 3: fall, 4: winter
- yr (discrete, nominal)
  Year 0: 2011, 1: 2012
- mnth (discrete, nominal)
  Month expressed by a number 1-12.
- hr (discrete, nominal)
  Hour between 0 and 23.
- holiday (discrete, nominal)
  Weather day is holiday or not (0 or 1).
  Extracted from http://dchr.dc.gov/page/holiday-schedule
- weekday (discrete, nominal)
  Which day of the week it is? 0: Monday... 6: Sunday.
- workingday (discrete, nominal)
  If day is neither weekend nor holiday is 1, otherwise is 0.
  68.39% of days are work days
- weathersit (discrete, nominal)
  Weather conditions. Can be one of four:
  - 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 Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
  The most common weather condition is 1: Clear, Few clouds, Partly cloudy, Partly cloudy.

- temp (continuous, interval)
  Normalized temperature in Celsius. The values are divided to 41 (max).
  Mean value is 0.49
  Stantard deviation is 0.18.

- atemp (continuous, interval)
  Normalized feeling temperature in Celsius. The values are divided to 50 (max).

- hum (continuous, interval)
  Normalized humidity. The values are divided to 100 (max).
  Mean value of humidity is 0.62
  Median is 0.62
  Standard deviation 0.14

- windspeed (continuous, interval)
  Normalized wind speed. The values are divided to 67 (max).
  Mean value of wind speed is 0.19
  Median is 0.18
  Standard deviation 0.07

- casual (discrete, ratio)
  Count of casual users.

- registered (discrete, ratio)
  Count of registered users.

- cnt (discrete, ratio)
  Count of total rental bikes including both casual and registered.
  Mean value of rented bikes is 4504 daily.
  Median is 4548.
  Standard deviation 1935.88

No data is missing in the data set.

Chapter 4

Data Visualization

Essential method of description and analysis of data is visualization. Representing information in this form gives the reader an intuition and better understanding of the data. In this chapter we'll use this tool to thoroughly recognize the data set and relations between attributes.

4.1 Outliers in the Data

To discover if there are outliers in our data we plotted box plots for several of the attributes. Below are the plots of the attributes in which we discovered potential outliers.

First one is humidity. We can observe two potential outliers in the box plot. This might be due to measurements or human error but this points shouldn't be a problem in the future as they are not very distant from other measurements.

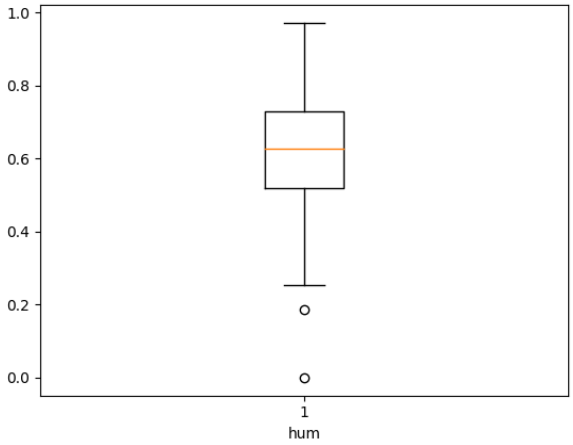

Figure 4.1: Box plot of humidity attribute

In case of wind speed we can observe quite a few outliers. The date of the highest measurement is 2011-02-19. This day in Washington D.C. there was a really strong wind. So these values while being higher than the rest correspond to real recorded values therefore they probably are not just errors.

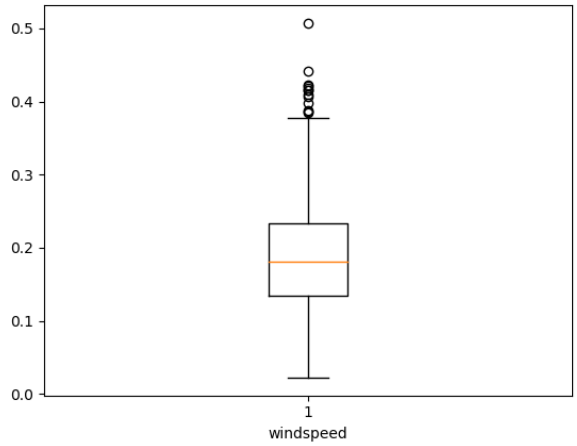

Figure 4.2: Box plot of wind speed attribute

Last one is the number of casual users. As visible on the plot there is a significant number of outliers. These values might be important in future analysis though as they may correspond to some specific conditions or anomalies.

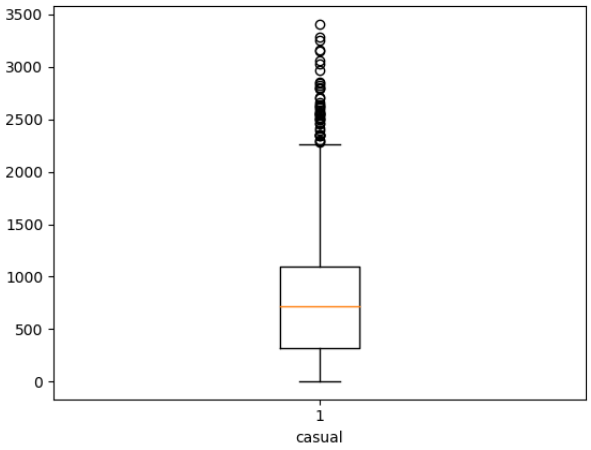

Figure 4.3: Box plot of casual users attribute

4.2 Distribution of the Data

In order to check if the data seems to be normally distributed we will plot a univariate distribution of observations and see how well it fits to a corresponding bell curve.
There are of course attributes upon which this kind of analysis would give us no information like day, month etc.
Here are our results.

Let's start with normalized perceptible temperature. Although probability seems to concentrate in the middle, we can clearly see a drop in the middle of the graph. This data seems not to be normally distributed.

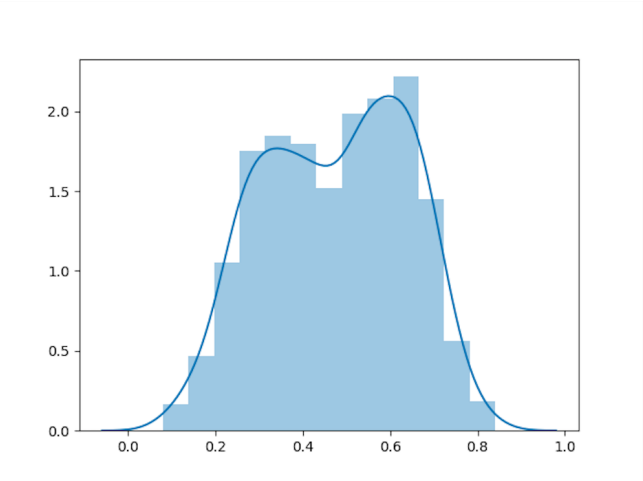

Figure 4.4: Perceptible temperature attribute pdf

In case of actual temperature case seems to be even more definitive.

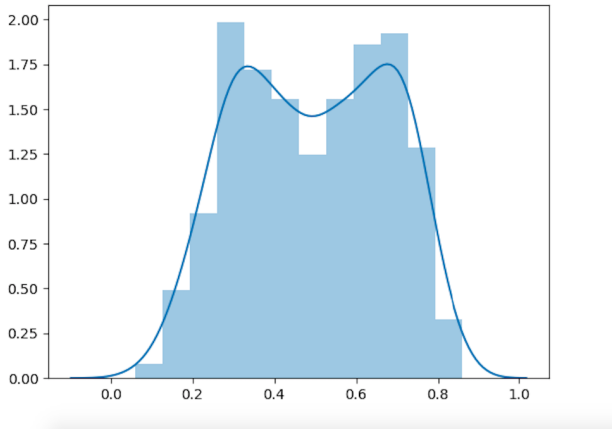

Figure 4.5: Temperature attribute pdf

Number of casual users is clearly not normally distributed but we can make an observation that very the number of such users is often relatively small.

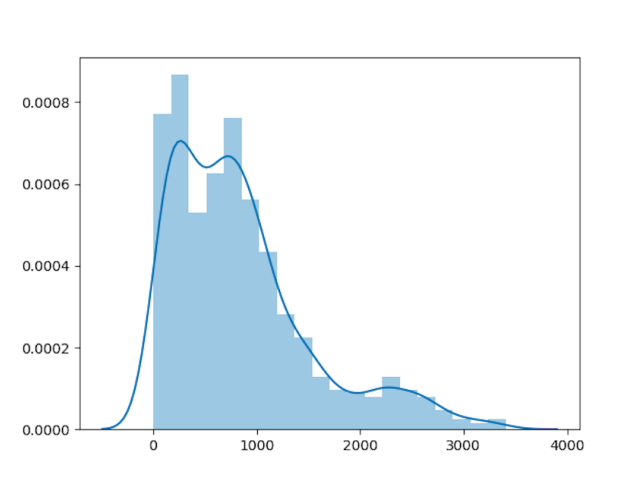

Figure 4.6: Number of casual users attribute pdf

In case of registered users the graph somewhat resembles a bell curve, it's hard to definitely state if this data is normally distributed or not. The data seems to be fairly symmetrical in respect to the middle probability concentration.

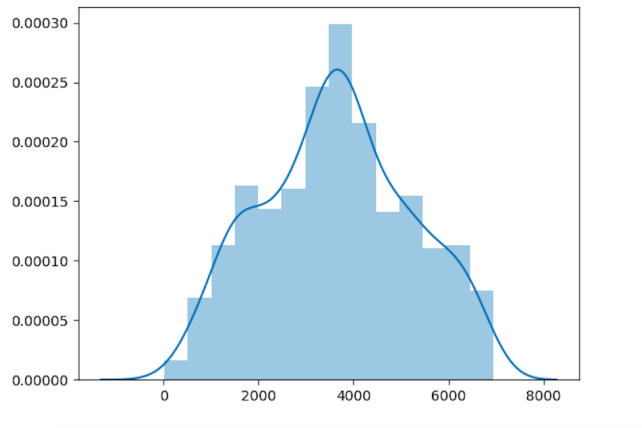

Figure 4.7: Number of registered users attribute pdf

Total number of rented bikes resembles the previous distribution but we can see how the casual users made a difference on the left side. Probably not normally distributed.

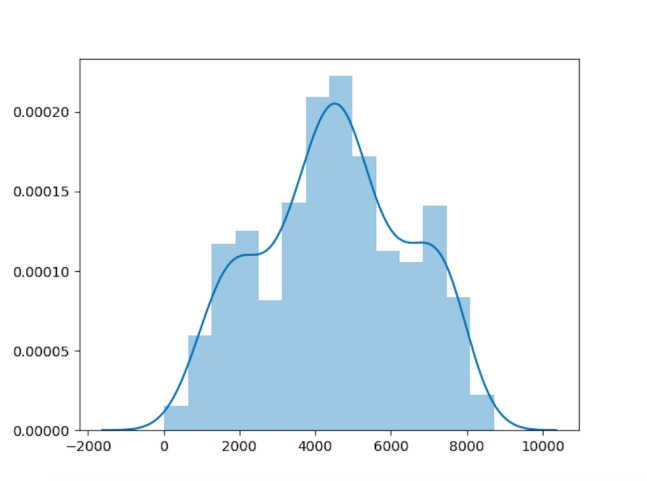

Figure 4.8: Total number of bikes rented attribute pdf

Humidity look very promising, we can see a clear concentration in the middle and symmetry. It might be normally distributed.

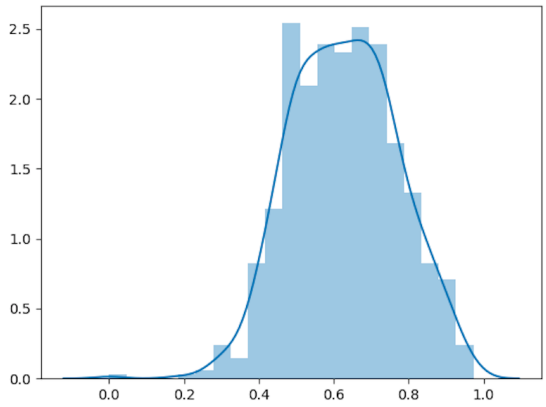

Figure 4.9: Humidity attribute pdf

At the end we will look at the wind speed probability distribution. Its graph fairly resembles the bell curve. We can see how probability get smaller when going away from the middle values.

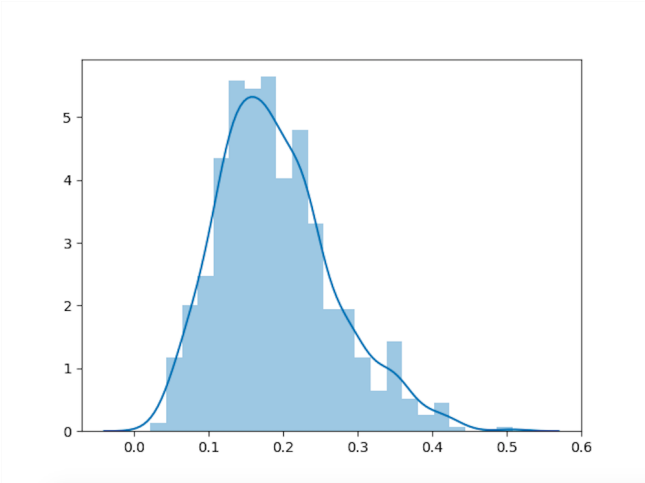

Figure 4.10: Wind speed attribute pdf

4.3 Correlation

We checked our attributes for possible correlations. Apart from obvious cases like correlation between registered users and users total for which correlation coefficient is 0.94, there are 2 interesting cases.

The first one is correlation between month and temperature. While correlation coefficient between these two attributes is only 0.22 we can observe something different on the plot. The temperatures are rising till July and then start to drop as we could expect. Therefor there is correlation between these two attributes that won't be visible if we only look at the correlation coefficient.

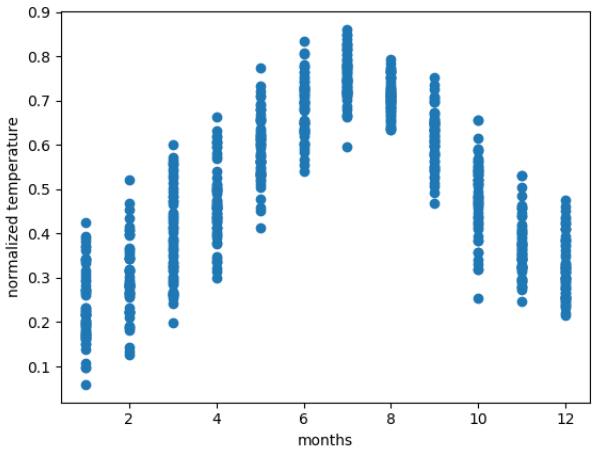

Figure 4.11: Temperatures in given months

Second case is correlation between temperature and total number of rented bikes. The correlation coefficient between the two is 0.627 which is visible on the plot. We can see that as the temperature is rising so in the number of rented bikes. Granted the correlation is not very high it is still visible on the plot.

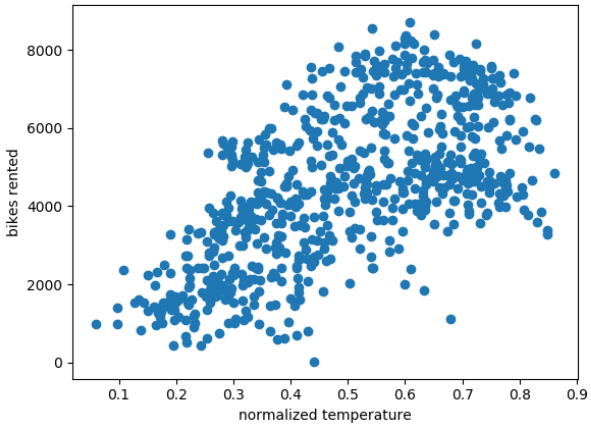

Figure 4.12: Total number of rented bikes based on temperature

4.4 PCA

During our PCA we didn't use all of the attributes as we didn't find some of them relevant. One of them is year as it's only purpose is distinction between year 2011 and 2012.

4.4.1 Variation

As the part of PCA we researched the variation explained as a function of PC included.

This is a plot showing how much of variance is explained by which Principal Component. We can see that PC1 explains about 40% of variation, PC2 about 20% and PC3 about 12%.

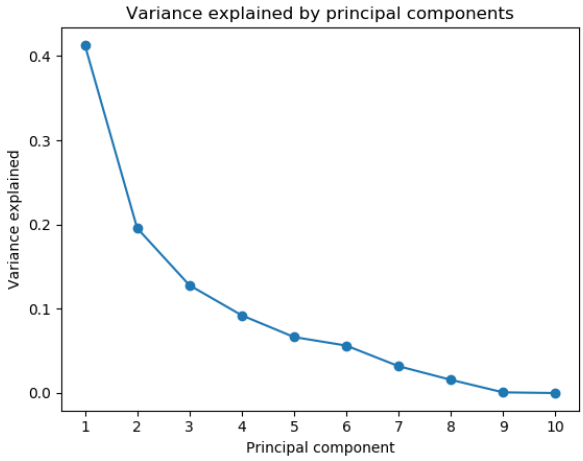

Figure 4.13: Percent of variance explained by the principal components

Next plot shows variance explained based on PCs included. We can observe that first 2 PCs explain most of the variance and first 3 explain about 73% of the variance.

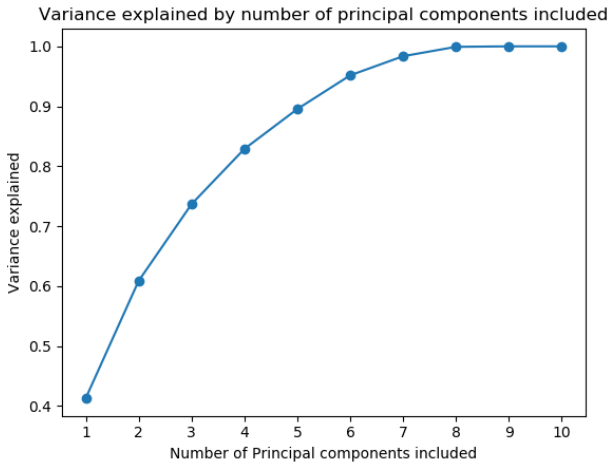

Figure 4.14: Percent of variance explained by the number of principal components

4.4.2 Principal directions

When performing PCA we can take a closer look at the principal directions. These are eigenvectors are directions in the features space which capture the maximal variance. After calculating the first direction we look for one which is orthogonal and so on. Here are the firs and second principal directions from our PCA:

$$v_1 = \begin{bmatrix} 2.98612530e-01 \\ 3.65583218e-01 \\ 4.36139833e-01 \\ -1.83837696e-01 \\ -1.20249367e-01 \\ 3.73638020e-02 \\ -7.64588168e-02 \\ 7.29000455e-01 \\ -4.65164709e-03 \\ -4.10323095e-17 \end{bmatrix}$$

$$v_2 = \begin{bmatrix} 2.39725380e-01 \\ 4.05269677e-01 \\ 4.93058564e-01 \\ -1.85202149e-01 \\ -1.90717402e-01 \\ 6.13139032e-02 \\ -9.61183135e-03 \\ -6.78710980e-01 \\ 2.85378866e-03 \\ 2.11925306e-17 \end{bmatrix}$$

4.4.3 Data projection

We began by plotting data projection on PC1 and PC2. For our classes we chose seasons as there are roughly the same number of points in each class. However the result wasn't satisfying. It might be partially due to the fact that PC1 and PC2 in total explain about 60% of the variance.

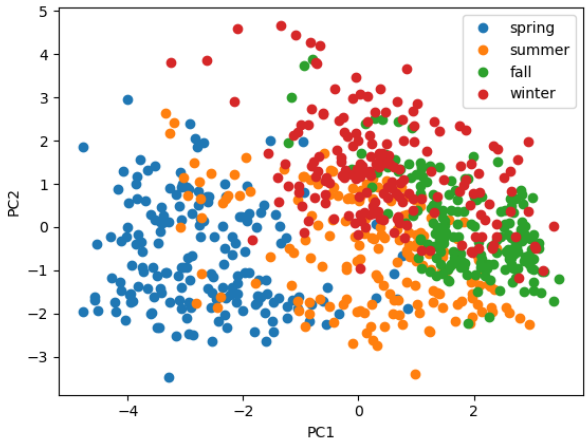

Figure 4.15: PC1 vs PC2

So in our next step we plotted data projection using first 3 PCs. As PC1, PC2, and PC3 explain in total about 72% of variance the result is not ideal but slightly better. We can observe distinctions between classes. They may not be sharp but that is because of information loss due to using only first 3 PCs.

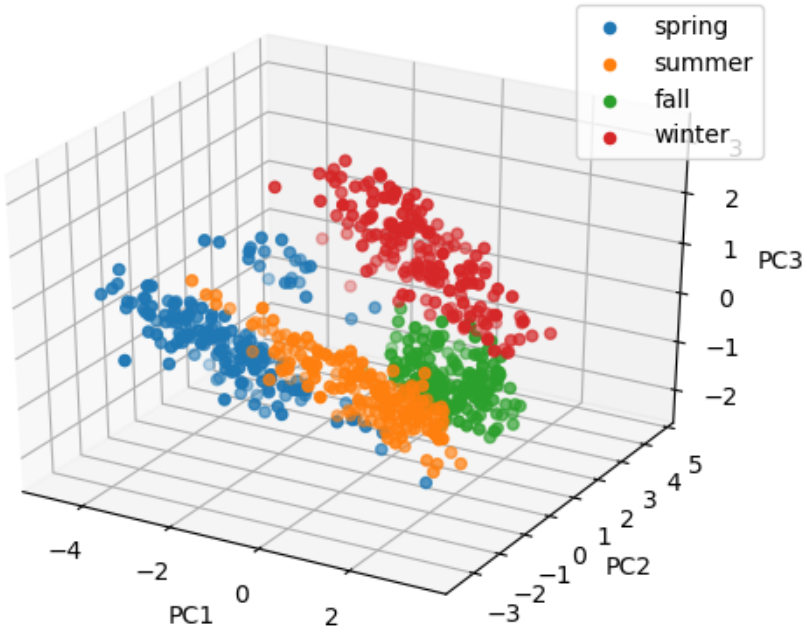

Figure 4.16: PC1 vs PC2 vs PC3

Chapter 5

Conclusions and Contribution

Analysis and visualization gave us a much better understanding of our data set. Not only did we learn the basic assumptions of the attributes, like usage of the numbers to represent the weather information, but also in the mean time we realized that some of the attributes will not affect our considerations in some situations. Date attribute may be problematic since it is the only attribute which is not a number, since we already have year and month attribute we might think about adding month day attribute just for simplicity and consistency. Also, in some experiments we do not want to use some attributes as it would be less interesting. For example we will not use month attribute when trying to guess the season. We would rather restrict ourselves to rented bikes count and some atmospheric information. We learned that there there is no missing data. In our analysis we found some outliers and data that we should be aware of but nothing that could prevent us from executing our goals. We feel fairly confident that we will be able to use the data in the machine learning methods the way we planned.

The contribution table:

Author	Section
Wojciech Ciok	Chapter 1
Wojciech Ciok	Section 2.1
Kacper Żyła	Section 2.2
Kacper Żyła	Section 2.3
Wojciech Ciok	Chapter 3
Kacper Żyła	Section 4.1
Wojciech Ciok	Section 4.2
Kacper Żyła	Section 4.3
Kacper Żyła	Section 4.4
Wojciech Ciok	Chapter 5

Bibliography

[1] Hadi Fanaee-T and Joao Gama. "Event labeling combining ensemble detectors and background knowledge". In: *Progress in Artificial Intelligence* (2013), pp. 1–15. ISSN: 2192-6352. DOI: 10.1007/s13748-013-0040-3. URL: http://dx.doi.org/10.1007/s13748-013-0040-3.