National College of Ireland

Hdip

2013/2014

Majid Ghanbari

13115570

Majid.Ghanbari@student.ncirl.ie

# Traffic Prediction of Bike Rental based on Environmental and Seasonal Factors

National
College of
Ireland

# Table of Contents

## Statement of originality

I certify that the content of this thesis is the product of my own work and that all the sources used in preparing this thesis have been acknowledged.

Majid Ghanbari
………………..

## Executive Summary

Rental bike systems provide an alternative type of transport to many people. However, it is largely believed that the regular use of them is largely influenced by the environmental factors. People usually blame the bad weather conditions for not getting to work on a bike. The weather conditions and seasonal effect can contribute to the reduction in using these systems while in some cases; the extensive use of them due to good weather conditions could create issues for terrific management authorities in large cities.

In this project, a solution is proposed to model the number of bike users for a rental bike system based on the environmental and seasonal factors. The relationships between environmental variables such as temperature, humidity, wind speed, holidays, weekdays, etc. which can play part in a user's decision on whether or not to use a bike are investigate. To analyse these variable and their relationships, extensive use of analysis and visualisation tools such as R, Rattle and Weka were used. It is then a linear regression model produced to predict the number of bike users based on weather conditions and seasons and etc.

After performing several tests to determine the quality of the model, it is found that the proposed model can explain around 80% of variations in the target value (number of bike users).

# 1 Introduction

The world population has seen a large increase in the last several decades while demands for access to transportation have increased dramatically. In recent years the possibility of using bikes as public transport system has captured the attention of many researchers and authorities. Nowadays the so called shared bike systems have become an integral part of many metropolitans around the world. Apart from providing customers with quick and flexible access to transport in urban areas, this solution offers many other advantages. Perhaps its second most import factor is the effect it can have on environment and pollution. However the planning and maintenance of such transport network is influenced by many factors and there are still challenges remain to be address in order to exploit its full potentials.

## *1.1 Domain Description*

Bike sharing systems are 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 (Fanaee-T, et. al (2013). These systems could provide an alternative type of transport to many people whether they would be used for work purposes or leisure times. Despite offering many advantages to their users, there have been some drawbacks to them. It has been believed that the performance of them is largely influenced by factors which are usually beyond our control. It not uncommon to hear people to blame the bad weather conditions not to opt for riding bike for getting to work. The weather conditions and seasonal effect can contribute to the reduction in using these systems while in some cases; the extensive use of them due to good weather conditions could create issues for terrific management authorities in large cities.

## 1.2 Motivations

The motivation behind this project was the author’s interest in environmental issues. The deployment of such systems can reduce the negative impacts on our daily lives. Stocked for long hours in terrific jams, noise pollutions and lack of exercise are some of the many unwanted side effects of driving cars in large cities. Dublin city has also a shared bike system and this is the reason why the topic was chosen for this project. Initially it was planned to conduct the research on Dublin city shared bike system, however the only data freely available for the purpose of research was Washington DC shared bike system data, but the approach will be the same and it could be applied to any other data collected from a different system along with data on weather conditions for the corresponded city. The shared bike systems are a great solution to large cities with high number of cars which are the main reason for the pollution in urban areas. However there are still scepticisms about their effectiveness in bad weather conditions or the possibility of their regular use in various seasons throughout the year.

## 1.3 Aims

The aim of this project is to analyse the relationship between the factors which could somehow affect the number of rented bikes and their mobility in a city. It also introduces a model for predication of a bike rental system that can count hourly or daily usage based on the environmental and seasonal settings. Seasonal and environmental factors have profound effect on number of people who use these services.

### 1.3.1 Scope Statements

The scope of the project is to analyse the relationship between the factors which could somehow affect the number of rented bikes and their mobility in a city. Seasonal and environmental factors have profound effect on number of people who use these services. The output information will be presented to the user by means of graphs and visual tools however the finalised prototype does not provide a dashboard.

### 1.3.2 Research Questions

The following research questions are answered in this study:

1. How the distributions of number of bike users for both categories of users have changes in 2011 and 2012?
2. How the distributions of environmental variables have changed in 2011 and 2012?
3. How the distributions of number of bike users and the weather condition are varied in relation to each other in both years?
4. How the numbers of bike users have changed with respect to the seasons in both years?
5. How the weather conditions have changed in various seasons in 2011 in both years with respect to each other?
6. Are there associations among variables used in this study (both independent and dependent)? If there are, how strongly they are related?
7. How a predictive model for number of bike users based on weather conditions and seasonal effects can be produced?
8. How to determine the quality and evaluate the model?

## *1.4 Solution Overview*

The data set under study is related to 2-year usage log of a bike sharing system namely Capital Bike Sharing (CBS) at Washington, D.C., USA. The use of two whole sequential years (2011, 2012) enables the researcher to investigate the effect of seasonal setting on the system. The size of datasets is particularly useful for both supervised and semi-supervised learning. The environmental data such as weather conditions are obtained from a different source while it was also possible to extract data relevant to holidays and week days using some other external resources in order to apply a holistic approach to the analysis.

This project is mainly built based around data mining concepts, each steps mentioned here describes the task that the researcher faced while going through the process of the data mining procedure.

1. Datasets downloaded, organised, and integrated for each specific intended outcome.

2. R data analytics package (R project, 2014) and some particularly important libraries for this project had to be downloaded. These libraries include RWEKA, and RATTLE.

3. The hourly recorded datasets for the two years (2011, 2012) under investigation content large number of instances. A training dataset decided to be used as preliminarily analysis.

4. An exploratory model including visual representation of key information and outcomes is created.

5. A predictive model to describe the behaviour of the system in relation to various influential factors is produced.
6. Test and evaluate the proposed model

## *1.5 Structure*

Chapter 2 provides an overview of the researches and works done within this domain. It includes a literature study of shared bike systems as well as the possible effect of environmental and seasonal factors on them. The most relevant aspects will be covered in more depth.

Chapter 3 describes the design and procedures of data mining approached that were employed as well as tools and technologies used in order to complete this project. It explains in details how the variables have been used and which analytic procedure has been performed in order to achieve the outcome for each step towards building the model. It also discusses some data preparation and transformation tasks which have been required in various scenarios.

Chapter 4 discusses the implementation phase of the project. It demonstrates the steps and analytic approach along with summary of results. It also includes visual representation for outputs of each task as they can be more comprehensive and convenient to the readers. It includes the explanation of results for these analytic tasks as well as the interpretations for the graphical representations of outputs.

Finally chapter 5 describes the methods used to test and evaluate the recommended linear regression model. The result for these test are discussed and used to measure the quality of this model.

## 2. Background

The use of rental bike systems in urban areas can potentially effect the traffic management and infrastructure of cities (Froehlich et al., 2010). However the management of these systems themselves is a challenge of its own. In rental bike systems, the distribution of bikes has to be constantly controlled and managed in order to timely supply number of required bikes to certain stations with respect to their frequency of use (Shu et al., 2011). Lin (2012) purposed a Geo-Aware redistribution system to address the above mentioned issue. Great deals of researches have been done to develop systems to monitor the dynamic and detect pattern of mobility in large cities by studying of these systems (Etienne and Latifa, 2013). From the users' point of view, it has been tried to investigate factors that can effect on users' behaviour and attitude towards conditions surrounding the use of these systems (Borgnat et al., 2011). Traditionally it is believed that the weather conditions can influence the potential users' decisions on whether to use a bike in a particular day (Ahmed et al., 2010). Weather condition continues to be a subject of debate among researchers and stakeholders. According to a research conducted by jdantos (2012) in Washington DC, as the temperature falls below 40 F and hike above 80 F the number of bike users drops dramatically. Between these two extreme weather conditions, temperature plays a smaller role in users' decision making on whether or not to use a bike and it will be more difficult to predict the number of bike users because there are a variety of other factors involved.

## 3. Methods

The implementation phase of this project has been divided into three major sections. The first section involves the data preparations. In the second section some exploratory data analysis is conducted and finally the third section provides solution to a predictive model, and fits the model. However the steps proposed for the completion of the project as a whole involves seven distinctive phase whish are described in the next section.

### *3.1 Analytic Approach*

The following steps demonstrate the structure used for completing this project, beginning with gaining knowledge domain. As shown in FIG.1 identifying and gathering the required data will be the next stage. This phase include researching various resource in order to pull the appropriate type of information which often has to be collated from various sources.

The third stage is deciding on the role of variables. This is where the knowledge domain comes into play and help to identify suitable attributes that could be used in order to address the research questions.

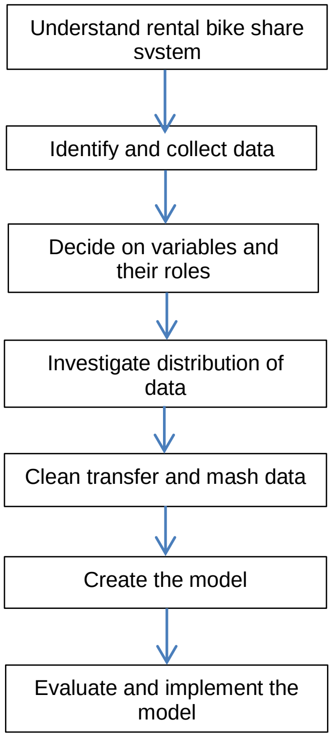

Understand rental bike share svstem
Identify and collect data
Decide on variables and their roles
Investigate distribution of data
Clean transfer and mash data
Create the model
Evaluate and implement the model

**FIG 1**

The fourth steps is the familiarisation with data using important statistical parameters describing our attributes such as standard deviations, maximums and minimums, means, medians and inter quartiles. In doing so, an extensive use of graphical demonstrations such as histogram, scatter plots, bar plots and box plots will be required.

The fifth phase involves the appropriate transformation of attributes in order to build the purposed model.

In the sixth stage of the project, the model with be created.

Finally in the last stage of project (seventh phase), the model will be fitted and evaluated.

### 3.2 Predictive Model

A Multiple Linear Regression (MLR) analysis is a supervised learning (the outcome of learning is predetermined and defined inadvance) (Maimon et al. 2010) is used to create a predictive model for this project. Using this model the number of shared bike used per day/ hour can be determined based on weather conditions as well as seasonal effect. However before using the model to predict anything, the model can tell us how strongly each can affect the usage of shared bike systems or whether they have any influence at all. A multiple linear regression model is an extended version of a simple linear regression model where a dependent variable will be predicted using several independent variables (Kutner et al. 2005). The relationship between the dependent and independent variables can be described with the following mathematical relationship where $Y$ describes the dependent variable and $X_1, X_2, X_3 \dots X_n$ are independent variables, $a_1, a_2, a_3 \dots . a_n$ are called regression coeffiecients and $\mathcal{E}$ is known as the error term or noise.

$$Y = a_0 + a_1X_1 + a_2X_2 + a_3X_3 + \cdots + a_nX_n + \mathcal{E} \quad \text{(Eq 1)}$$

Equation (Eq 1) is also called a response function and all independent variables are called predictors (Kutner et al. 2005).

## 3.3 Data Description

The data set under study is related to 2-year usage log of a bike sharing system namely Capital Bike Sharing (CBS) at Washington, D.C., USA. The use of two whole sequential years (2011, 2012) enables the researcher to investigate the effect of seasonal setting on the system. The size of datasets is particularly useful for both supervised and semi-supervised learning. The environmental data such as weather conditions are obtained from a different source while it was also possible to extract data relevant to holidays and week days using some other external resources in order to apply a holistic approach to the analysis.

- hour.csv: bike sharing counts aggregated on hourly basis. Records: 17379 hours

- day.csv - bike sharing counts aggregated on daily basis. Records: 731 days

The data hour.csv was aggregated to create day.csv dataset.

Both hour.csv and day.csv have the following fields, except hr (hour: 0 to 23) which is not available in day.csv.

The holiday attribute has been extracted from the following source: http://dchr.dc.gov/page/holiday-schedule. The following (Table 1) gives detailed descriptions of variables (UCI 2014) used in the dataset.

**Table1**

| Variable | Description |
|---|---|
| instant | record index |
| dteday | date |
| season | Winter, spring, summer, autumn (numeric 1, 2,3,4) |
| yr | Year (2011, 2012) numeric values(0, 1) |
| mnth | Months (numeric values 1 to 12) |
| holiday | whether a day is holiday or not (1 for holiday, 0 for not holiday) |
| weekday | day of the week (numeric values 1 to 7) |
| workingday | if day is neither weekend nor holiday is 1, otherwise is 0 |
| weathersit | Clear, Few clouds, partly cloudy, partly cloudy (numeric value 1) Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist (numeric value 2)<br>Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds (numeric value 3)<br>Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog (numeric value 4) |
| temp | Normalised temperature in Celsius. The values are divided by 41 (max) |
| atemp | Normalised air temperature in Celsius. The values are divided by 50 (max) |
| hum | Normalised humidity. The values are divided by 100 (max) |
| windspeed | Normalised wind speed. The values are divided by 67 (max) |
| casual | count of casual users |
| registered | count of registered users |
| cnt | count of total rental bikes including both casual and registered |

### *3.4. Tools and Technologies*

In this project, R is used as the main analytic package while the project has benefited from availability of Rattle; one of R very powerful visualisation tools. Rattle is a free graphical interface for data mining with R (Williams, G. 2011). The other data mining tool which has been used in this project is Weka. Weka is a collection of machine learning algorithms for data mining tasks written in Java, containing tools for data pre-processing, classification, regression, clustering, association rules, and visualisation (Hornik, 2013). Package Rweka is an R interface to Weka. Both RATTLE and RWeka are adds on to R package. Once they are added, they can be called by using rattle() and RWeka() functions. Whenever rattle is called, there will be a window opened with various features for data mining such as decision trees (R project, 2014).

# 4. Implementation

In this section, all the steps involving the data preparation process, exploratory analysis of variable and the approaches involving a predictive model are described.

## *4.1 Data Preparation*

Both data sets (2011, 2012) have had originally several categorical attributes. For the purpose of correlation analysis, they had to be converted to numeric variable. Season with four category was changes to (1, 2, 3, 4) numeric levels. Weekdays categorical variable was converted to (1 to 7) level and holiday, workingdays were changed to (0 or 1) instead of (Yes or No). The weathersit attribute with three levels as it was described in section 3.3, was converted to numeric with three levels (1 to 3). The numeric values for environmental variables (temperature, win speed, humidity, and air temperature) had to be normalised for the purpose of building the linear regression model as followed:

- Normalised temperature in Celsius. The values are divided by 41 (max)
- Normalised air temperature in Celsius. The values are divided by 50 (max)
- Normalised humidity. The values are divided by 100 (max)
- Normalised wind speed. The values are divided by 67 (max)

There was no missing data in these datasets.

## *4.2 Exploratory Analysis*

In this section, variables in data set were explored and their important parameters were calculated and summarised. For both years 2011 and 2012 theses summarised parameters and characteristics were compared by means of tabulated values and graphs. Histograms, boxplots and density functions were used wherever could be more convenient in order to demonstrate these features.

### 4.2.1 Number of Rented Bikes

A comparison between the two years (2011, 2012) was conducted to investigate how the number of rented bikes have changed for both category of users namely as registered users and casual users.

#### I. Methods

R was used to load and process the data for both years. The variables summarised in this analysis are number of registered users, number of casual users and the total users (cnt). It was then Rattle was used to visualise some of those statistics to more conveniently looking at parameters in relation to attributes.

#### II. Results

According to table 2, in 2012, on average the rental bikes use has increased dramatically compare to 2011. However the minimum use has been very small in 2012 (20 registered users for a whole day, 2 casual users and total of 22 for that day). Considering the mean value and the median for each category, it can be concluded that these variables are normally distributed. However further analysis will be needed to confirm the normally of these attributes.

**Table 2**

| | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| **Registered** | | | | | | |
| **2011** | 416 | 1730 | 2915 | 2728 | 3632 | 4614 |
| **2012** | 20 | 3730 | 4776 | 4581 | 5663 | 6946 |
| **Casual** | | | | | | |
| **2011** | 9.0 | 222.0 | 614.0 | 677.4 | 871.0 | 3065.0 |
| **2012** | 2.0 | 429.8 | 904.5 | 1018.5 | 1262.0 | 3410.0 |
| **cnt** | | | | | | |
| **2011** | 431 | 2132 | 3740 | 3406 | 4586 | 6043 |
| **2012** | 22 | 4369 | 5927 | 5600 | 7011 | 8714 |

FIG 2 shows a comparison between the number of rental bike users in 20011 and 2012 for each category of users. Clearly there is more variation in number of registered bike users in 2011 while in 2012 there has been a more stable count of registered users. The total number of bike users (cnt) for each has also been compared.

As it can be seen, the number of registered users in 2012 is shifted to the high end, indicating we had more days with large number of registered users.
The casual number of users' density graph shows more variation for both years. Both of them are not smooth, however the average number of casual users still higher in 2012. Finally the total numbers of users in both years follow nearly the same trends as for the registered users which are expected.

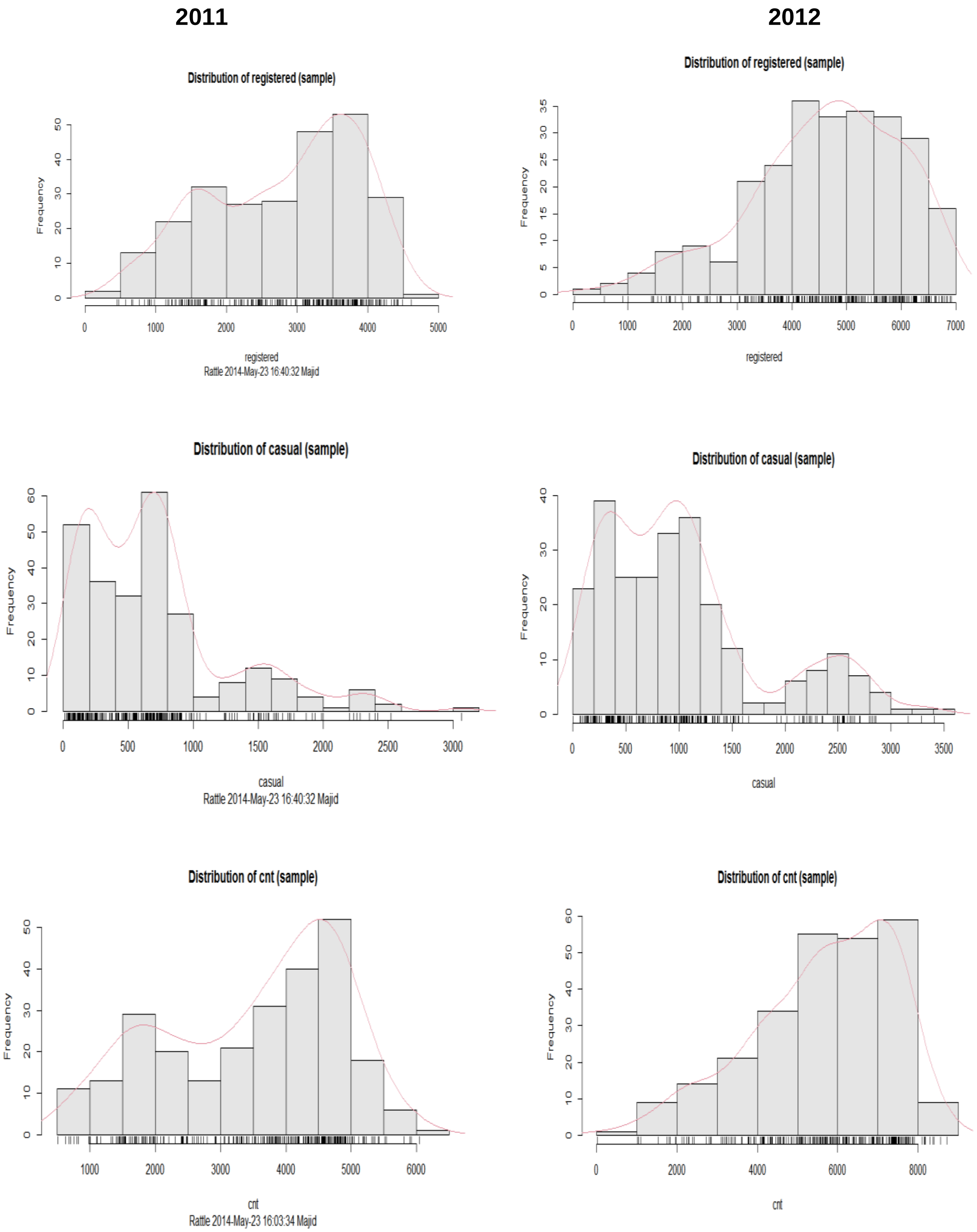


**FIG 2 Comparison of Distributions for number of users for years 2011 and 2012**

### 4.2.2 Environmental Variables

In this section, attributes such as temperature, humidity, wind speed and weather sit will be analysed and the results will be graphically represented. They are factors that are considered to have profound influence on the number of bike users throughout the year.

### I. Methods

Both files (2011(1).csv and 2012(1).csv) were loaded and analysed using R. The weather sit has three levels indicating three different scenarios. For each year the number of days corresponded to a one of three weather conditions has been determined and tabulated as followed (Table 2):

**Table 2**

| **no. of days**<br>**weathersit** | **2011** | **2012** |
|---|---|---|
| Clear-FewClouds-PartlyCloudy | 226 | 237 |
| LightSnow-LightRain+ Thunderstorm+ ScatteredClouds-LightRain + ScatteredClouds | 15 | 6 |
| Mist+Cloudy-Mist+ BrokenClouds-Mist+ FewClouds-Mist | 124 | 123 |

Table 3 has summarised important statistics for the four environmental variables, temperature, wind speed, humidity and the air temperature. These parameters have been compared for the two consecutive years 2011 and 2012.

**Table 3**

| | Min | 1st Qu. | Median | Mean | 3rd Qu. | Max |
|---|---|---|---|---|---|---|
| **temp** | | | | | | |
| **2011** | 2.424 | 13.325 | 19.646 | 19.953 | 26.923 | 34.816 |
| **2012** | 4.407 | 14.256 | 21.081 | 20.667 | 26.812 | 35.328 |
| **atemp** | | | | | | |
| **2011** | 3.953 | 16.098 | 23.642 | 23.342 | 30.619 | 42.045 |
| **2012** | 5.083 | 17.534 | 24.889 | 24.093 | 30.382 | 40.246 |
| **hum** | | | | | | |
| **2011** | 0.00 | 53.83 | 64.75 | 64.37 | 74.21 | 97.25 |
| **2012** | 25.42 | 50.81 | 61.19 | 61.22 | 71.11 | 92.50 |
| **windspeed** | | | | | | |
| **2011** | 1.500 | 9.084 | 12.522 | 12.824 | 15.750 | 34.000 |
| **2012** | 3.126 | 8.959 | 11.708 | 12.701 | 15.490 | 29.585 |

## II. Results

In general, the weather condition has been better than 2011, with more days (11 days more) as Clear-FewClouds-PartlyCloudy, less days (9 days less) with LightSnow-LightRain+Thunderstorm+ScatteredClouds-LightRain+ ScatteredClouds, and with only one day less weather condition as (LightSnow-LightRain+Thunderstorm+ScatteredClouds-LightRain+ScatteredClouds).

FIG 3 shows the distributions of the three important environmental variables, temperature, wind speed and humidity. The air temperature (atemp) has been discarded in this analysis because of its dependency to the temperature. On average the temperature and consequently the air temperature has been higher in 2012, however the humidity has been lower. On average, the wind speed shows an increase in 2012.

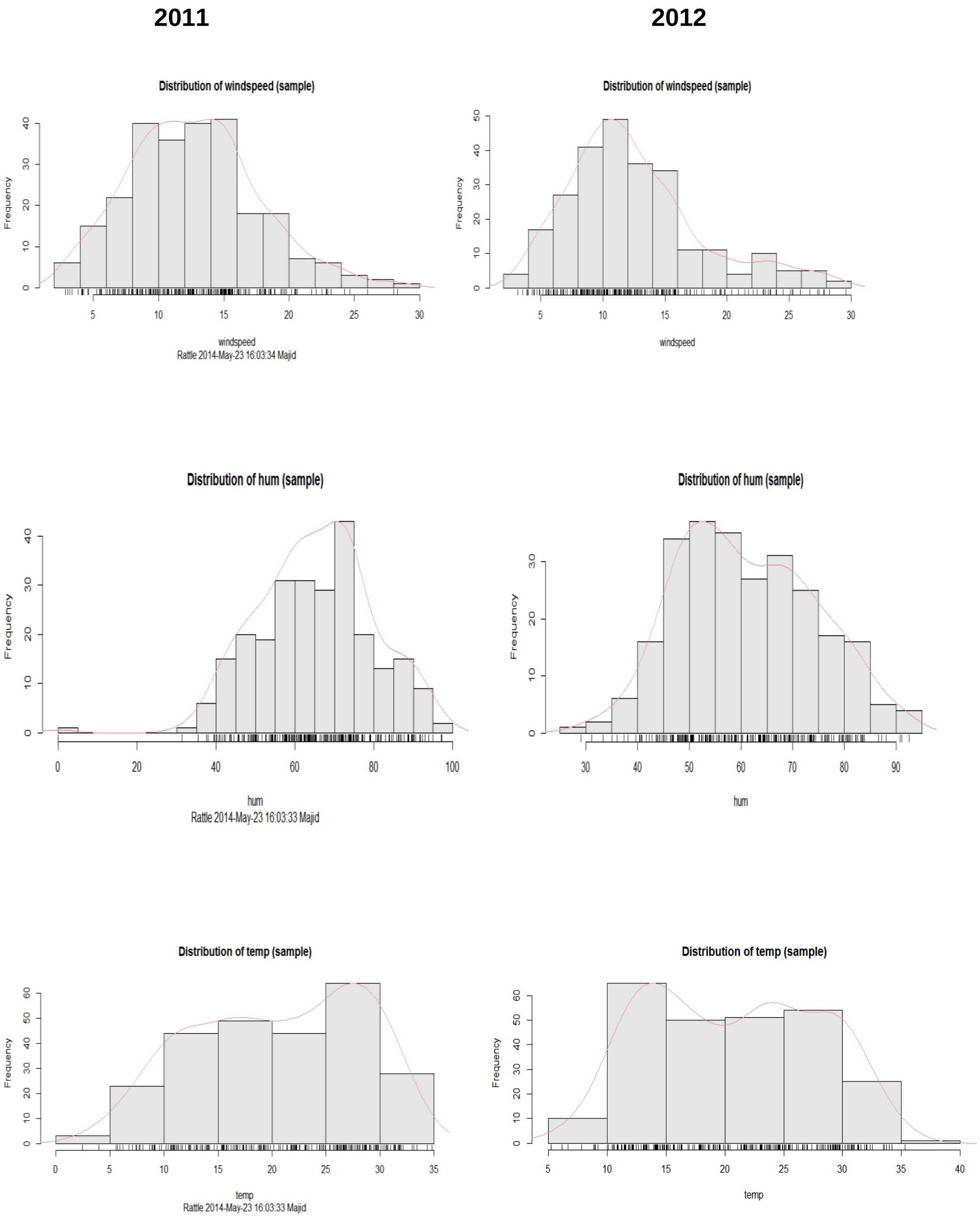


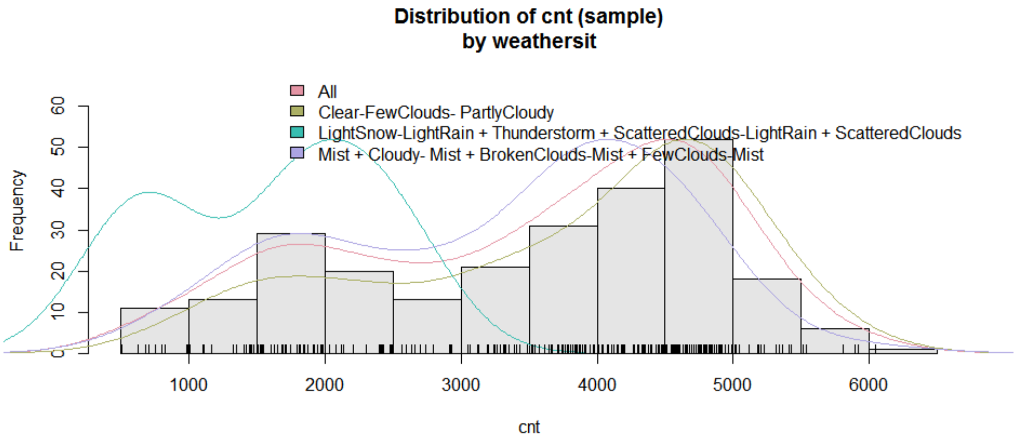
**FIG 3 Comparison of Distributions for Wind speed, temperature and humidity for years 2011 and 2012**

### 4.2.3 Distributions of Number of Bike Users by Weathersit

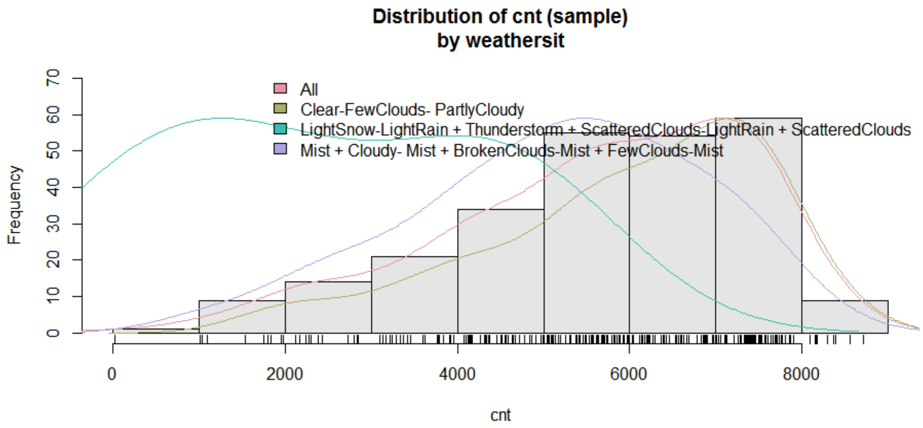
The weathersit is believed to have a huge influence on the number of bike users. In this part of exploratory data analysis, a comparison between the numbers of bike users in both years has been conducted with respect to the weather condition.

#### I. Methods

Using R and Rattle, a pair of density graph for the three levels of weather conditions have been created.



2011



2012

**FIG 4 Density graphs for Weather Situations in years 2011, 2012**

**II. Results**

According to FIG 4, in both years, whenever the weather has been nice and clear (Clear-FewClouds-PartlyCloud),the bike users have been increased dramatically. The number of bike users has been the lowest when the weather conditions considered being bad (LightSnow-LightRain+ThunderStorm+ScatteredClouds-LightRain+ScatteredClouds). The distribution of number of bike users in all kind of weather condition in total has been labelled by “All”.

### 4.2.4 Seasonal Variable

The season variable with its four levels (winter, spring, summer, autumn) is one of the factors which could have some potential effect on the number of rented biked throughout the year.

**I. Methods**

Two datasets (day.2011(1).csv and day.2012(1).csv) were uploaded to R and then the Rattle. The Season has been was chosen as target variable. Density graphs for total number of bikes users for both years have been created. It was then six box plots for two categories of bike users and the total count were created to show the differences of means for each category in different seasons.

**II. Results**

FIG 5 represents the density graphs for total number of bike users in different seasons for both years. In both years the total numbers of rental bike uses have been the lowest during the winter time. Summer and spring have more days with high number of bike users while autumn is about the average of both extremes. Autumn 2011 particularly look normally distributed over the whole range of number of bike users.

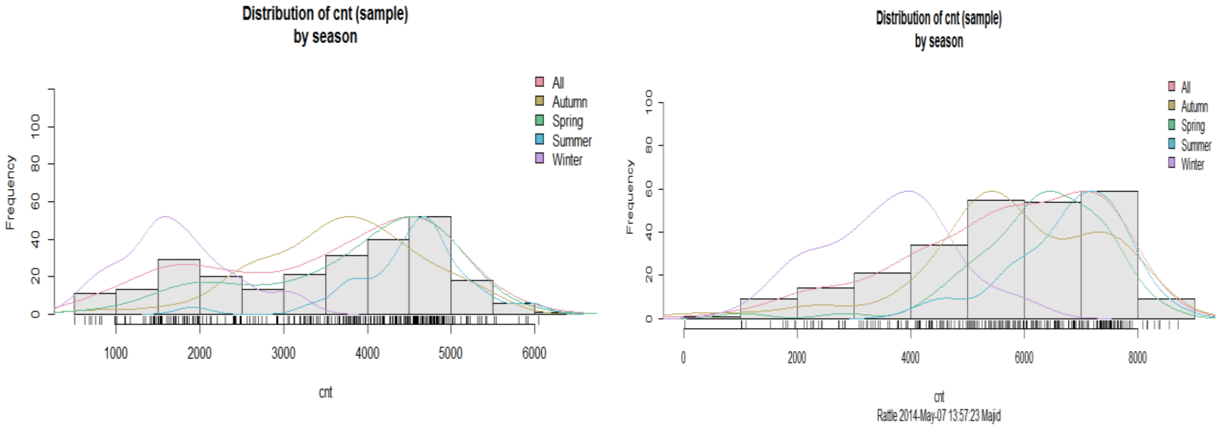


**FIG 5 Density graphs for total number of bike users in different seasons for years 2011 and 2012**

FIG 6 takes a closer look at one of these years (2012) by looking at monthly distribution of total number of bike users using multiple boxplots. As it can be seen, September and October with one and two outliers respectively, have the highest average number of bike users which are two months of autumn. After these two months, July and August have the largest average of total of the bike users. January and February which both belong to winter and December which is partly in winter time and have all the lowest average of bike users.

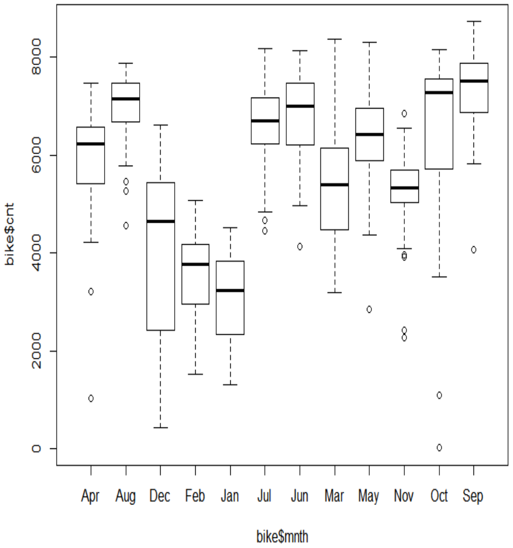


**FIG 6 Monthly Distribution of total number of bike users in 2012**

According to FIG 7 number of registered of bike users in all seasons are higher in 2o12 in comparison to year 2011. The highest number of registered bike users for both years are in summer while the lowest are in winter seasons. This is also the case for number of casual users in each year. These consequently are valid for the total number of bike users for each year.

The total number of bike users in 2012 has increased for every season compare to year 2011, so the increase within each year could be seasonal both the increase between the two years can be due to better weather condition throughout the whole year in 2012.

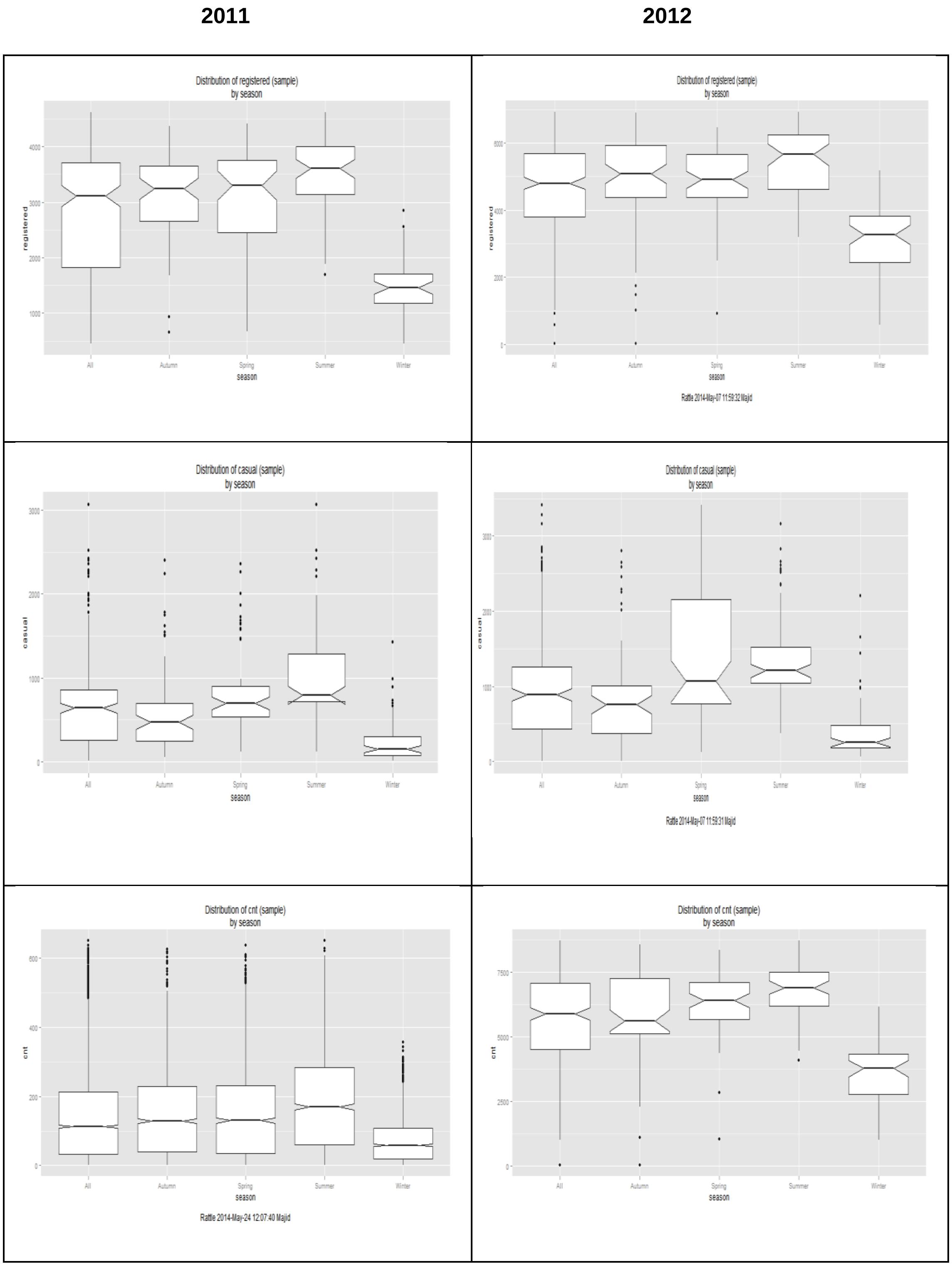


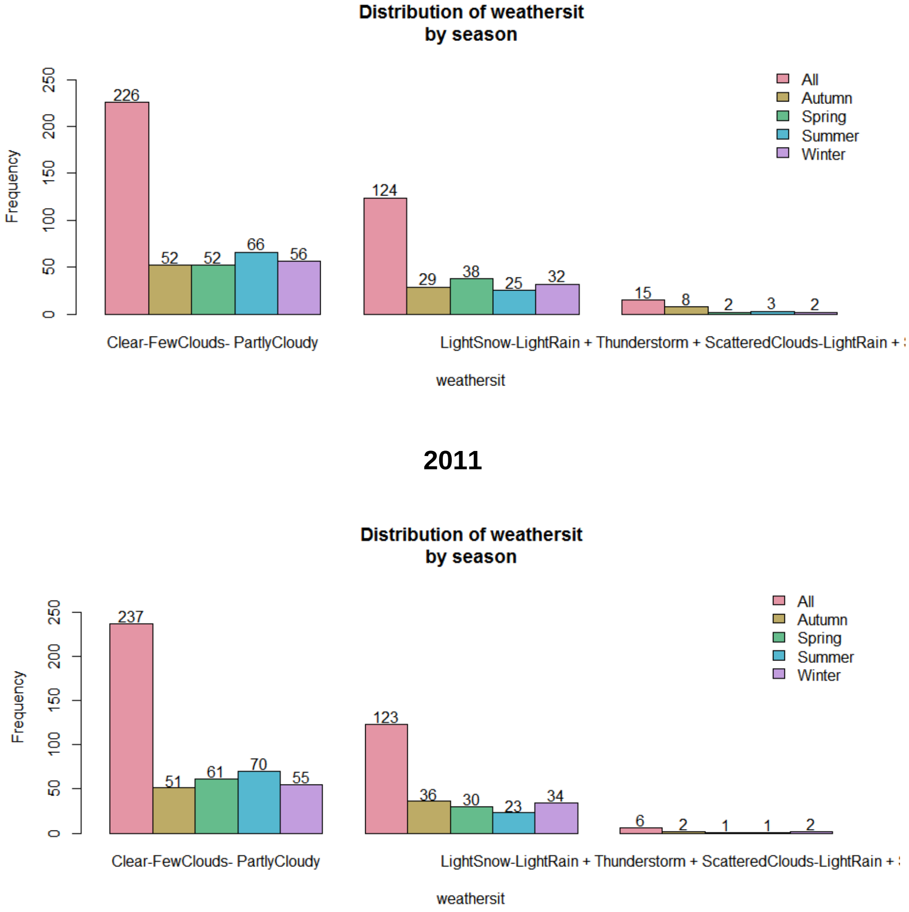
**FIG 7 Comparison of seasonal distribution of bike users in both years (2011, 2012)**

### 4.2.5 Distribution of Weathersit by Season

We already know that the overall temperature has risen in 2012 and overall we had better weather conditions in 2012 from our earlier analysis, however to investigate further, a comparison of the weathersits by season for between the two years was performed.

#### I. Methods

Barplots were used to visualised the difference in number of various type of weather conditions for each year.



**2011**

**2012**

**FIG 8 Distribution of weather conditions by season for years 2011 and 2012**

II. **Results**

The results of the analysis and bar plots shows that in 2012 we had a better summer and spring weather condition with more clear (Clear-FewClouds-PartlyCloudy) days. The number of this type of weather condition was reduced by one in both autumn and winter. In addition, the type of bad weather condition (LightSnow-LightRain+ThunderStorm+ScatteredClouds LightRain+ScatteredClouds) was less present in all seasons except in winter which remained the same (2 days). Thus we can conclude that the overall improvement has been due to an improvement in each individual season and not by a particular season like summer or winter only.

### 4.2.6 Distributions of Environmental Variables by Season

Now that we know we have had an improvement for weather condition almost right throughout the whole year for every season in 2012, we can investigate how the environmental factors have changed within this year in each season compare to the previous year.

I. **Methods**

Using R and the Rattle six cumulative representations are plotted to show best how the three environmental variables have changed from season to season within each individual year. A comparison can be made for variation between the two years.

II. **Results**

Results are displayed in FIG 9. The temperature has been increased in all seasons in 2012 compare to 2011. The wind speed has also been dropped in 2012 overall for all seasons as well as the humidity. Thus, we can conclude the year 2012 was a warmer year and there has been an improvement for other environmental factors such as wind seed and humidity in all seasons.

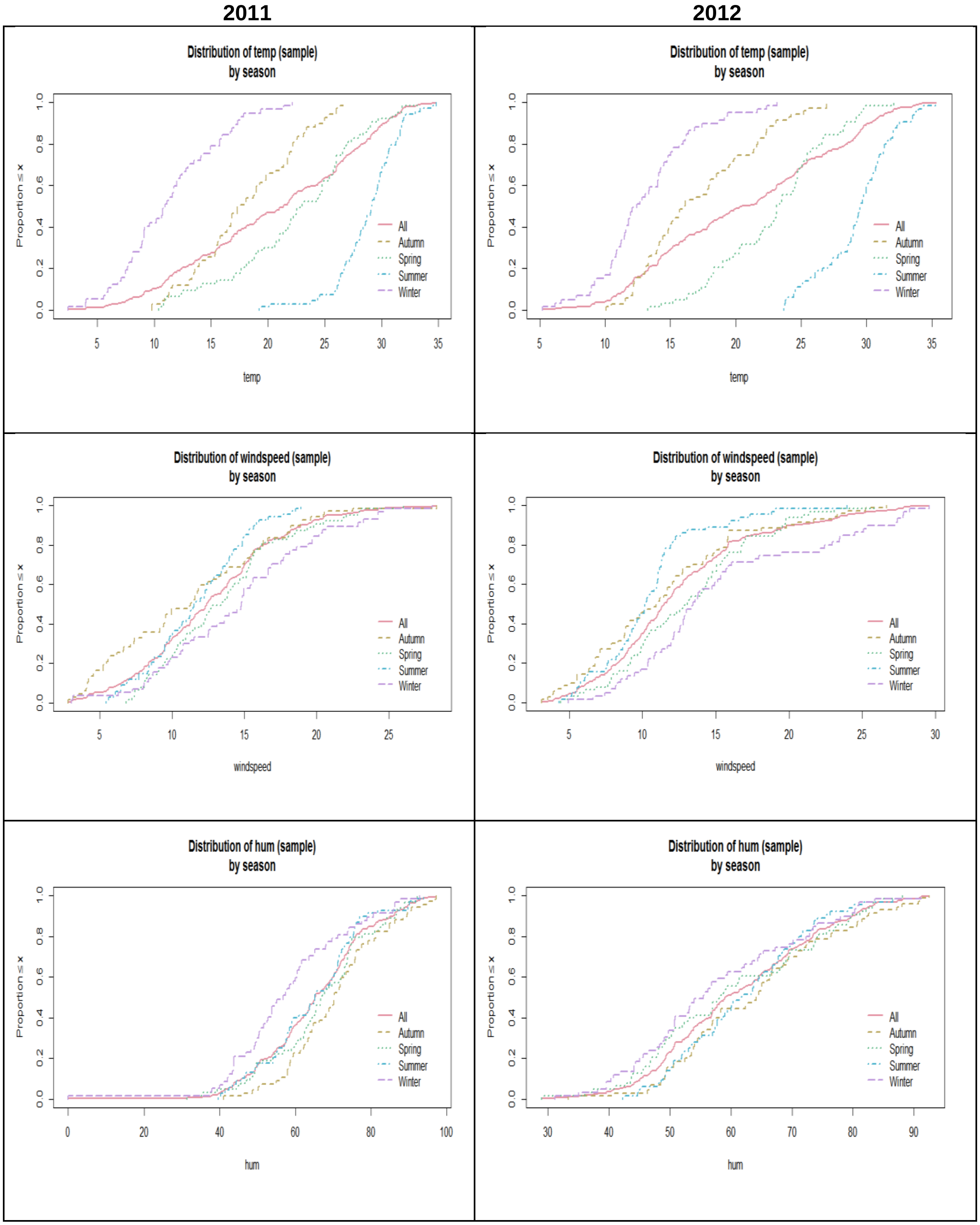


FIG 9 Comparison of Distributions of Environmental Variables by Season for both years (2011, 2012)

## *4.3 Predictive Analysis*

The aim of this section is to investigate the associations as well extends of these associations between variables in this study. A Multi Linear Regression model (MLR) is created to predict the target value (number of rental bikes users) based on weather conditions and seasonal factor. The effects of these factors as well as the combined effect (interaction effect) of some of these factors with respect to their level of contributions to the variations of the target variable are also considered.

For the purpose of predictive analysis (supervised learning) in this section, the two consecutive years have been combined to create a single file (day.csv). This is done to create a file with the appropriate size for better model building and validity of the results. In doing so, the effects of environmental and seasonal variables are investigated over a two years period. The day.csv is the numeric version (with dummy variables) of the day(1).csv file as they were described earlier (section 4.1). Table 3 shows the summary of statistics for number of bike users in file day(1).csv (2 years arranged as a single file):

**Table 3**

| Min | 1st Q | Median | Mean | 3rd Q | Max |
|---|---|---|---|---|---|
| 22 | 3152 | 4548 | 4504 | 5956 | 8714 |

### 4.3.1 The Normality Tests

An investigation is conducted to check whether the target variable is normally distributed. The assumption for normality is a requirement for the regression analysis.

## I. Methods

1. A histogram is one of most popular tests to check whether a data is normality distributed. The hist() function in R is used to create the histogram.

2. Q-Q normality plot

If the shape of the graph approximately follows a straight line, then we can assume that the data is normally distributed.

## II. Results

The histogram is quite symmetric around the mean. This indicates that data for the cnt (total number of bike users) is fairly normally distributed. The Q-Q plot also nearly follows a straight line, so this test confirms the earlier conclusion.

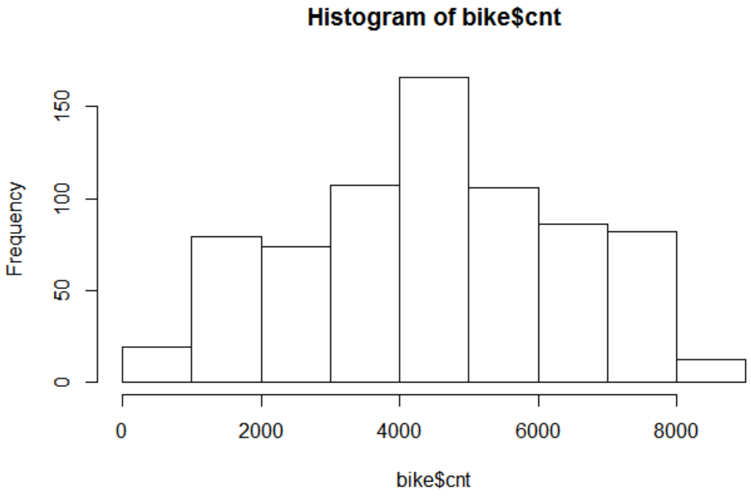


**FIG 9 Histogram of number of bike users**

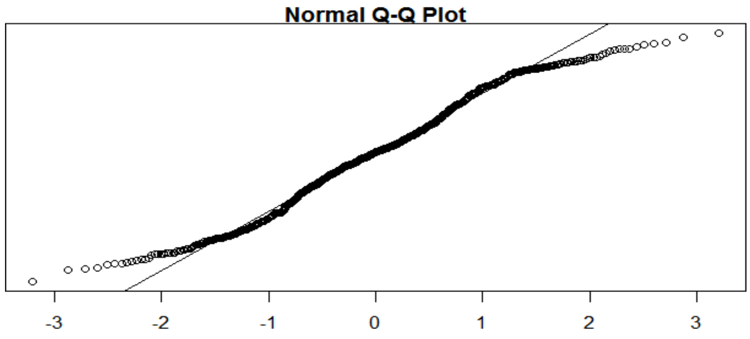


**FIG 10 Q-Q normality plot for cnt (number of bike users)**

### 4.3.2 Correlation Analysis and MultiCoLinearity

A correlation analysis is performed to determine the potential associations between the variable. “Correlations tell us how well two variables relate to each other” (Quick 2010, p.77).

Before creating a MLR model it is necessary to investigate the relationship between not only the dependent variable (target variable) and all other independent variable (predictor variables), but also to check how closely independent variables are related each other among themselves. One of the variables from each pair with strong association between them has to be removed in our model building process; otherwise they affect the credibility of our model. This problem is called “MultiCoLinearity” (Quick 2010).

#### I. Methods

Correlation functions will not accept categorical variables, so all variable have to be numeric in order to perform the correlation analysis. Using R and its cor() function, the numeric version of “day.csv” file was used with dummy variables for season, month, working day, etc. The normalised versions of environmental variables (temperature, wind speed, air temperature and humidity) were used.

## II. Results

Table 4, summarises the result for the analysis.

**Table 4**

| | instant | season | mnth | holiday | weekday | workingday | weathersit | temp | atemp | hum | windspeed | casual | registered | cnt |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| instant | "1.00" | "0.41" | "0.50" | "0.02" | "-0.00" | "-0.00" | "-0.02" | "0.15" | "0.15" | "0.02" | "-0.11" | "0.28" | "0.66" | "0.63" |
| season | "0.41" | "1.00" | "0.83" | "-0.01" | "-0.00" | "0.01" | "0.02" | "0.33" | "0.34" | "0.21" | "-0.23" | "0.21" | "0.41" | "0.41" |
| mnth | "0.50" | "0.83" | "1.00" | "0.02" | "0.01" | "-0.01" | "0.04" | "0.22" | "0.23" | "0.22" | "-0.21" | "0.12" | "0.29" | "0.28" |
| holiday | "0.02" | "-0.01" | "0.02" | "1.00" | "-0.10" | "-0.25" | "-0.03" | "-0.03" | "-0.03" | "-0.02" | "0.01" | "0.05" | "-0.11" | "-0.07" |
| weekday | "-0.00" | "-0.00" | "0.01" | "-0.10" | "1.00" | "0.04" | "0.03" | "-0.00" | "-0.01" | "-0.05" | "0.01" | "0.06" | "0.06" | "0.07" |
| workingday | "-0.00" | "0.01" | "-0.01" | "-0.25" | "0.04" | "1.00" | "0.06" | "0.05" | "0.05" | "0.02" | "-0.02" | "-0.52" | "0.30" | "0.06" |
| weathersit | "-0.02" | "0.02" | "0.04" | "-0.03" | "0.03" | "0.06" | "1.00" | "-0.12" | "-0.12" | "0.59" | "0.04" | "-0.25" | "-0.26" | "-0.30" |
| temp | "0.15" | "0.33" | "0.22" | "-0.03" | "-0.00" | "0.05" | "-0.12" | "1.00" | "0.99" | "0.13" | "-0.16" | "0.54" | "0.54" | "0.63" |
| atemp | "0.15" | "0.34" | "0.23" | "-0.03" | "-0.01" | "0.05" | "-0.12" | "0.99" | "1.00" | "0.14" | "-0.18" | "0.54" | "0.54" | "0.63" |
| hum | "0.02" | "0.21" | "0.22" | "-0.02" | "-0.05" | "0.02" | "0.59" | "0.13" | "0.14" | "1.00" | "-0.25" | "-0.08" | "-0.09" | "-0.10" |
| windspeed | "-0.11" | "-0.23" | "-0.21" | "0.01" | "0.01" | "-0.02" | "0.04" | "-0.16" | "-0.18" | "-0.25" | "1.00" | "-0.17" | "-0.22" | "-0.23" |
| casual | "0.28" | "0.21" | "0.12" | "0.05" | "0.06" | "-0.52" | "-0.25" | "0.54" | "0.54" | "-0.08" | "-0.17" | "1.00" | "0.40" | "0.67" |
| registered | "0.66" | "0.41" | "0.29" | "-0.11" | "0.06" | "0.30" | "-0.26" | "0.54" | "0.54" | "-0.09" | "-0.22" | "0.40" | "1.00" | "0.95" |
| cnt | "0.63" | "0.41" | "0.28" | "-0.07" | "0.07" | "0.06" | "-0.30" | "0.63" | "0.63" | "-0.10" | "-0.23" | "0.67" | "0.95" | "1.00" |

According to this correlation matrix, the variable instant (which is equivalent to date), temperature, air temperature, number of registered users as well as casual users are highly correlated with total number of bike users. The association between registered and casual users is obvious, so we can discard them in our MLR model. We also have to discard the following variables from our regression analysis due to Multicolinearity effect:

Month (mnth) which is strongly correlated with season, air temperature (atemp) which is highly correlated with temperature (temp) to prevent multi-colinearity have to be eliminated in model building process.

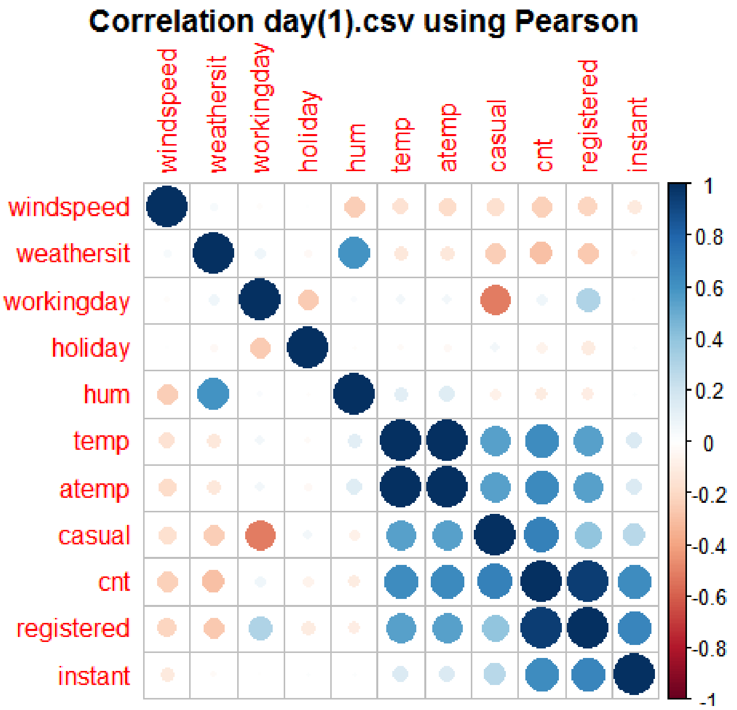


**FIG 10 Demonstration of correlation between variables**

### 4.3.3 Multi Linear Regression (MLR) Model

After determining the possible relationship among variables and discarding few which could cause mutlicolinearity, the multi linear regression model is built.

I. **Methods**

Using categorical day(1)csv file, the lm() function in R is used to create the regression model. This is done in conjunction with factor() function in order to convert the categorical variables to factors as followed:

```
>lm.countbike<-
lm(cnt~instant+factor(season)+factor(holiday)+factor(weekday)+factor(workingda
y)+factor(weathersit)+temp+hum+windspeed, data=bike)
```

It is then summary function used to get the output.

**II. Results**

Table 5 shows the calculated values for the estimated regression coefficients, t-value and p-values for each. The stars indicate the importance of each factor. The p-value for the following factors is smaller than 0.05 (significant level):

factor(weekday)Thursday

factor(weekday)Tuesday

factor(weekday)Wednesday

factor(weekday)Monday

factor(weekday)Saturday

This indicates the probability of those coefficients being equal to zero and means a higher likelihood to be insignificance. This also suggests that the elimination of these factors will not affect the reliability of the model significantly.

**Table 5**

| Coefficient | estimate | P-value | t-value |
|---|---|---|---|
| (Intercept) | 2082.1463 | 2.25e-13 *** | 7.476 |
| instant | 4.9278 | < 2e-16 *** | 28.700 |
| factor(season)Spring | 476.3653 | 1.49e-05 *** | 4.360 |
| factor(season)Summer | -287.1612 | 0.031660 * | -2.153 |
| factor(season)Winter | -423.9959 | 0.000129 *** | -3.850 |
| factor(holiday)Yes | -739.8599 | 0.000305 *** | -3.629 |
| factor(weekday)Monday | -182.9699 | 0.143201 | -1.466 |
| factor(weekday)Saturday | 17.0487 | 0.889279 | 0.139 |
| factor(weekday)Sunday | -401.3475 | 0.001119 ** | -3.272 |
| factor(weekday)Thursday | -27.4823 | 0.822434 | -0.224 |
| factor(weekday)Tuesday | -95.6322 | 0.435908 | -0.780 |
| factor(weekday)Wednesday | -37.5598 | 0.760610 | -0.305 |
| factor(workingday)Yes | NA | NA | NA |
| factor(weathersit)LightSnow-LightRain + Thunderstorm + ScatteredClouds-LightRain + ScatteredClouds | -1959.4988 | < 2e-16 *** | -8.772 |
| factor(weathersit)Mist + Cloudy- Mist + BrokenClouds-Mist + FewClouds-Mist | -400.6528 | 5.24e-06 *** | -4.590 |
| temp | 5188.1382 | < 2e-16 *** | 15.767 |
| hum | -1647.8416 | 2.57e-07 *** | -5.203 |
| windspeed | -2838.1184 | 1.06e-09 *** | -6.183 |

## Summary:

Residual standard error: 881.2 on 714 degrees of freedom

Multiple R-squared:  0.7976, Adjusted R-squared:  0.7931

F-statistic: 175.9 on 16 and 714 DF, p-value: < 2.2e-16

The above summary shows the estimated coefficients. The standard errors for the residuals, R-squared, F-statistic, estimation for σ are all included in this important summary.

R-squared (0.7976) represents the quality of the model. This mean almost 80% of the variations in number of bike users can be explain by this model. Adjusted-R (0.7931) is slightly lower than R-squared because it considers the number of variables used to build the model. The p-value of the F-statistics is quite small compare to the 0.05 significant level and supports the likelihood of the model being insignificant.

In order to confirm the level of importance of each independent variable in this model, an ANOVA test is conducted. Using R command anova(lm.countbike) the following output is produced (Table 6 ):

**Table 6**

Analysis of Variance

| **Variable** | **Df** | **Sum Sq** | **Mean Sq** | **F value** | **Pr(>F)** |
|---|---|---|---|---|---|
| instant | 1 | 1083287662 | 1083287662 | 1395.2144 | < 2.2e-16 *** |
| season | 3 | 646062014 | 215354005 | 277.3640 | < 2.2e-16 *** |
| holiday | 1 | 8579468 | 8579468 | 11.0499 | 0.0009322 *** |
| weekday | 6 | 13586342 | 2264390 | 2.9164 | 0.0081254 ** |
| weathersit | 2 | 210366938 | 105183469 | 135.4705 | < 2.2e-16 *** |
| temp | 1 | 183700495 | 183700495 | 236.5961 | < 2.2e-16 *** |
| hum | 1 | 9902522 | 9902522 | 12.7539 | 0.0003791 *** |
| windspeed | 1 | 29678259 | 29678259 | 38.2240 | 0.000000001061 *** |
| Residuals | 714 | 554371693 | 776431 | | |

According to table 6, the variable (weekday with 2 stars) contributed least to the variation of the response variable (cnt) however its p-value (0.0081254) is still well below 0.05 significant level and should be included in model.

**Note:** As it can be seen, the “workingday” variable is not used by the model and NA indicated the exclusions (Table 5). It was then the Rattle instead of R was used to rebuild the model, but the same result obtained. After investigating this issue, the explanation was found in Rattle as followed:

> Singularities were found in the modelling and are indicated by an NA in the following table. This is often the case when variables are linear combinations of other variables, or the variable has a constant value. These variables will be ignored when using the model to score new data and will not be included as parameters in the exported scoring routine (rattle 2012).

Therefore the author continued to test and evaluate the model and looked into possible ways of improving the quality of model even further.

# 5. Testing and Evaluation

To investigate the quality and performance of the regression model further several diagnostic observations are performed. Weka is then used to predict the number of rental bike users and compare the result to the actual (observed) values in order to evaluate the proposed linear regression model.

## *5.1 Testing the Quality of the Model*

Four diagnosis plots are used to determine the performance of the models.

### 5.1.1 Diagnostic Plots

1. Residuals vs Leverage plot

   It is used to check the points with excessive leverage. Cook's distance estimates the influence of a data point

2. Location-Scale plot

   It is often used conjunctively with Residuals vs Leverage plot to identify points with excessive leverage.

3. Normal Q-Q plot

   It is used to check whether the residuals are normally distributed.

4. Residuals vs Fitted

   It is used to detect non-linearity, unequal error variances, and outliers.

**I. Methods**

Using function plot(lm.count), the four above mentioned plots are produced.

**II. Results**

FIG 10 shows the Residuals vs Leverage and Scale-Location plots. The points in both plots are grouped together with very few that are far from the centre. This is a good indication for the quality of the model.

On the other hand, the Q-Q plot follows the shape of a straight line demonstrating the normality of residuals.

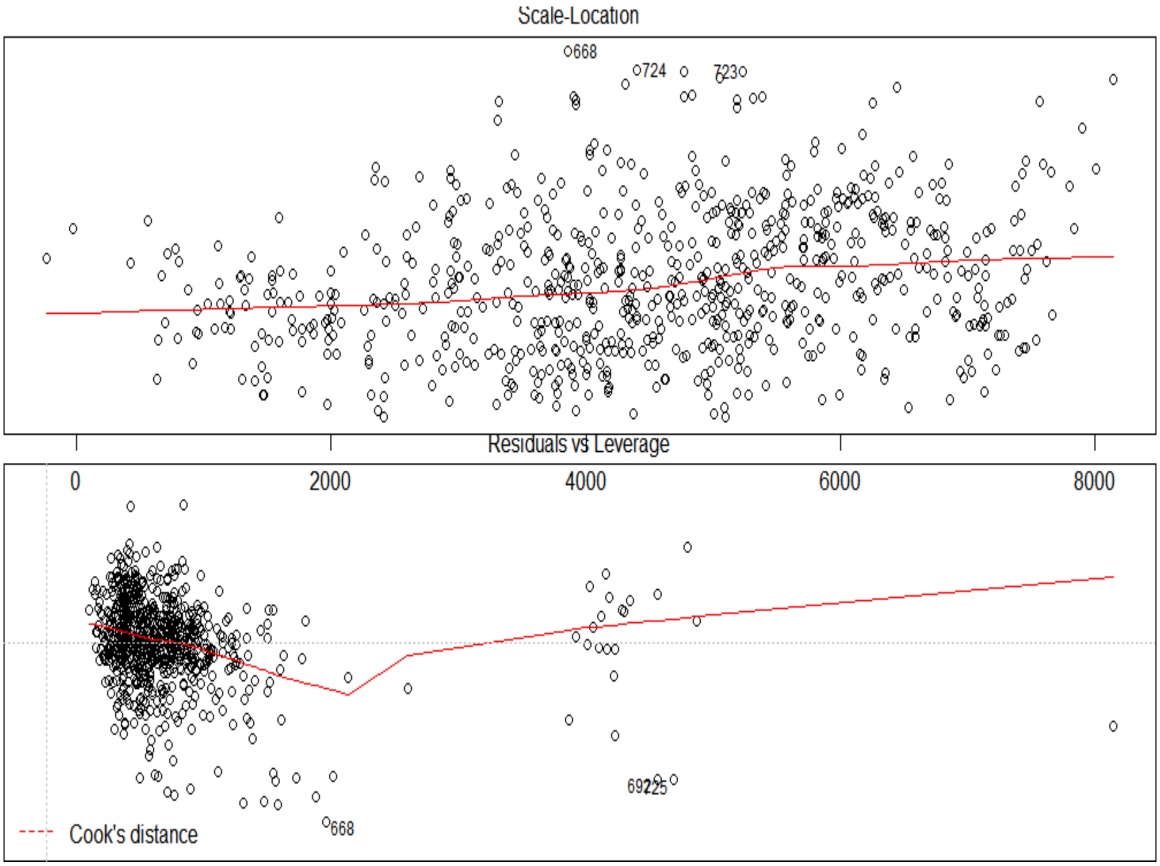


**FIG 10 Scale location plot and Residuals vs Leverage plot**

In FIG 11, the Residuals vs Fitted points to three outliers (724, 723, and 668) as it were referred to in Q-Q plot as well. However, the model overall demonstrates a good level of reliability and quality in modelling the regression relationships among variable and predicting the target variable.

### 5.1.2 Checking for Outliers

Our earlier observation gave us some indication of the existence of few outliers. In order to confirm their presence, a test will be conducted.

Outliers are known to be the source of error and under performance in regression models. Thus it is important to investigate the reason for presence in data.

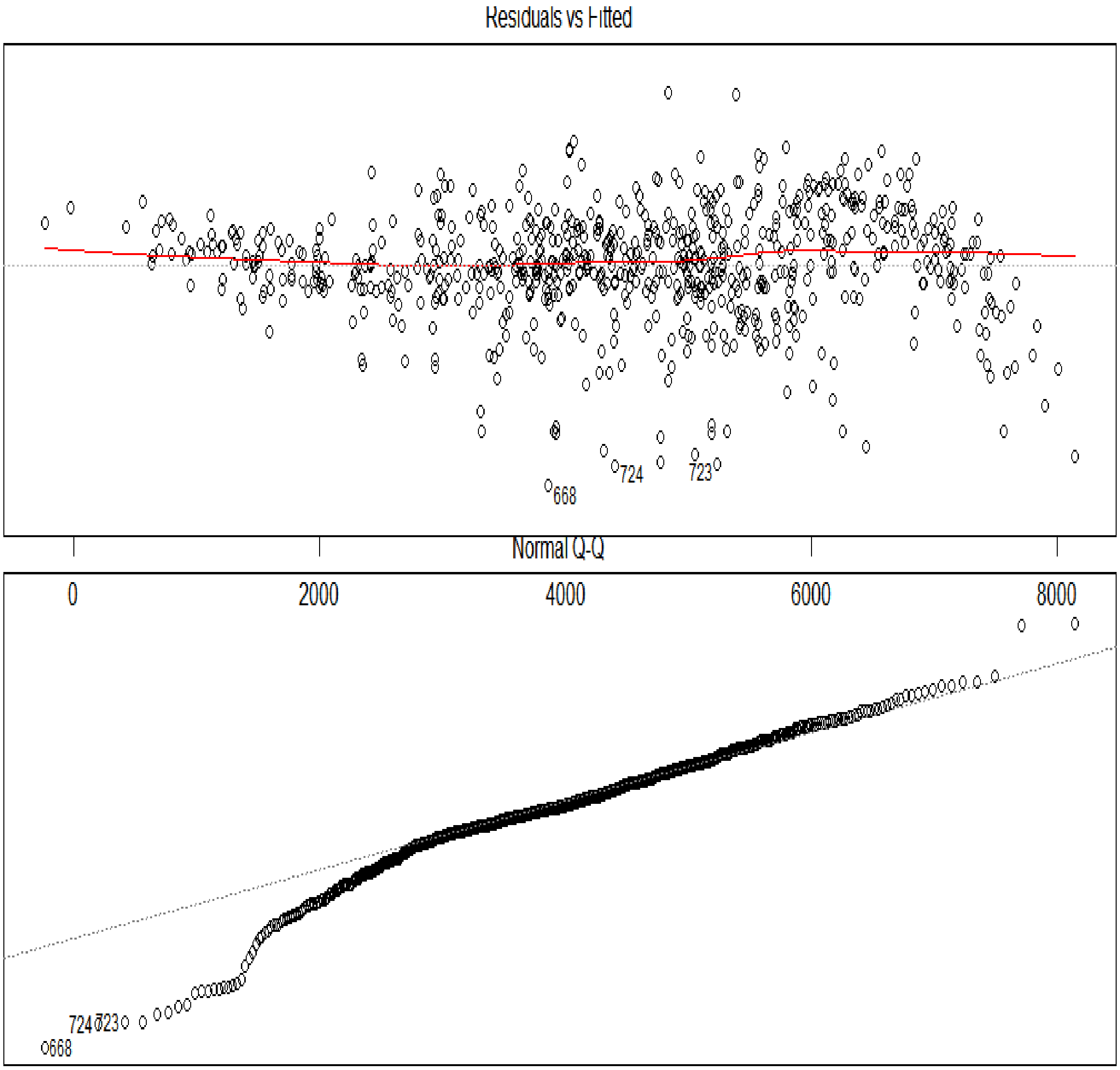


**FIG 11 Normality Q-Q plot and Residuals vs Fitted**

## I. Methods

The outlierTest() in R is used to check on the most extreme points in our model. The model lm.countbike will be passed to this function.

### II. Results

Table 7 summaries the result of the test. It identifies the point (669) as the extreme outlier, confirming somewhere around that region there is a cluster of outliers as it was observed in earlier plots (FIG 11).

**Table 7**

| observation | rstudent | unadjusted p-value | Bonferonni p |
|---|---|---|---|
| 669 | -4.025503 | 0.00006294 | 0.046009 |

## 5.1.3 Testing for Autocorrelation

The model is tested for the possible occurrence of a phenomenon called autocorrelation. Autocorrelation can distort a model and to detect its presence is crucial to performance of each regression model.

The following describes best what autocorrelation is and why it has to be investigated while building regression models:

> Autocorrelation in the residuals is a scourge because it distorts the regression statistics, such as the F statistic and the t statistics for the regression coefficients. The presence of autocorrelation suggests that your model is missing a useful predictor variable or that it should include a time series component, such as a trend or a seasonal indicator (Teetor 2011, p. 294).

### I. Methods

A test called the Durbin–Watson test is performed using dwtest() in R. This test checks the null hypothesis that there is no significant autocorrelation. If the p-value is bigger than the chosen level of significance (0.05 usual standard of significant level), the null hypothesis will be accepted, otherwise we conclude there is an autocorrelation which is not equal to zero.

#### II. Results

The following is the result of the R command dwtest(lm.countbike):

………………………………………………………………………

```
Durbin-Watson test
data:  lm.countbike
DW = 0.8865, p-value = 0.7607
alternative hypothesis: true autocorrelation is greater than 0
```

………………………………………………………………………

The p-value for the test is much higher than 0.05 which strongly supports the null hypothesis for this test. So we can conclude there is no significant autocorrelation in our regression model which means it is unlikely that our model is affected by missing a useful predictor variable. This test could assure us that the elimination of predictor “workingday” which was discarded by R in our model building process has not influenced the quality of our model.

## *5.2 Evaluation of Model*

In this section, Weka is used in order to evaluate the linear regression model. Weka offers several built in functions for selecting appropriate attributes as well as a wide range of data mining algorithms.

### 5.2.1 Prediction of Response Using Weka

Weka can be used to quickly build and visualise data mining models. It can also be used to conveniently evaluate them.

#### I. Methods

The file day(1).csv is uploaded to the Weka. Independent variables which were used in building the model earlier are selected and the ones which could create multicolinearity are removed. Number of bike users (cnt) is selected as target value. “Classify” tab is selected and from “choose” tab’s drop down list, folder option is selected and then the subcategory “LinearRegression”. From test options “Use training set” option is chosen. By clicking start the linear regression model is created. Pressing on “more options” tab the “output prediction” can be ticked and pressing ok and then start, calculate all predicted values for each instance of our data set using the training data set.

## II. Results

The first 10 rows of the output are shown in table 9. The linear regression model is as shown below (eq. 2):

```
cnt =                                                                    (eq. 2)
      4.9276 * instant +
    424.6177 * season=Autumn,Spring,Summer +
    475.4049 * season=Spring,Summer +
   -762.33   * season=Summer +
    772.4697 * holiday=No +
    264.3203 * weekday=Monday,Tuesday,Wednesday,Saturday,Thursday,Friday +
    125.4115 * weekday=Wednesday,Saturday,Thursday,Friday +
   1561.1615 * weathersit=Mist + Cloudy- Mist + BrokenClouds-Mist + FewClouds-Mist,Clear-
FewClouds- PartlyCloudy +
    399.2018 * weathersit=Clear-FewClouds- PartlyCloudy +
   5186.2136 * temp +
  -1653.068  * hum +
  -2833.5303 * windspeed +
  -1473.3135
```

**Table 8**

```
=== Evaluation on training set ===
            === Summary ===
Correlation coefficient          0.893
Mean absolute error            634.2595
Root mean squared error         871.2972
Relative absolute error         40.0975 %
Root relative squared error     45.0077 %
Total Number of Instances       731
```

According to the above summary table 8, the correlation coefficient between the actual values and the predicted values is 89.3% which indicate a strong correlation. The error column shows the difference between the predicted value and actual value for each instance.

**Table 9**

| INST# | ACTUAL | PREDICTED | ERROR |
|---|---|---|---|
| 1 | 985 | 1253.175 | 268.175 |
| 2 | 801 | 900.325 | 99.325 |
| 3 | 1349 | 1130.575 | -218.425 |
| 4 | 1562 | 1150.56 | -411.44 |
| 5 | 1600 | 1599.03 | -0.97 |
| 6 | 1606 | 1628.103 | 22.103 |
| 7 | 1510 | 1001.279 | -508.721 |
| 8 | 959 | 503.43 | -455.57 |
| 9 | 822 | 277.988 | -544.012 |
| 10 | 1321 | 924.439 | -396.561 |

# 6. Conclusions

In this project, the influences of environmental and seasonal factors on a rental bike system were investigated. A variety of approaches were used to study the relationship between those factors. Weather conditions and seasonal effect are considered as most important variables when people decide to choose this type of transportation. According to this study, in warmer weather, the number of bike users increase however with a high level of humidity this number decreases. In winter time and during its high pick and colder month the number reduces dramatically. Windy days have not the same effect as the temperature, however rain and snow causes less people use the system. Comparing the year 2011 and 2012, in general the year 2012 experienced a better weather conditions, with less rain and clearer days. The average temperature has been higher, with less wind and lower humidity. All these could be the reasons behind higher number of bike users in 2012. The model developed in this project, could explain nearly 80% of variations in the target variable (number of bike users n rental bike system) based on weather conditions, seasons and whether a day is weekday, weekend or holiday. The evaluation of the model showed around 89% correlation between the calculated predicted values and observed values. However it is of crucial importance to inspect other various possible elements surrounding these systems when making decision on how best it can be managed and optimised. There are some drawbacks influencing this project and regression model which is caused by how the data was initially collated. The number of bike users could be also hugely influenced by the availability of bikes. There is no information available on how many bike have been available to rent in each instant. For instance, there could be some stations (docks) that were out of bikes for rent while some customers have been willing to use a bike and that could influence the number of users regardless of the weather conditions and seasons.

# 7. Further Development or Research

With having more time and access to more data resources, there would be possibilities for further improvement in the model developed for prediction of the number of bike users in this rental bike system. Firstly there are approaches for the reduction (backward or forward reductions) of number of predictors (independent) variables used in building the model with respect to the level of their contribution to the variations in the target variable. Applying these algorithms is iterative in nature and involves a trial and error process. Another method that might improve the quality of the model is to consider the interaction effect of two or more of independent variables. However, both of these approaches are very time consuming, especially when the number of predictors is large and there is no guarantee in obtaining a better result at the end.

Secondly, with access to more data resources it is possible to create Geo-coded map of the city, visualise the clusters of bike users in and around the city by having the data on when they check in or out of a dock. When receiving data via sensors installed on each rental bike, terrific management authorities can collect live data and monitor the movement of bike users and make just-in-time decision for terrific movements on time.

# 8. References

Ahmad et al. (2010). “Impact of weather on commuter cyclist behaviour and implications for climate change adaptation”. *Capital Bikeshare: Usage and Weathe*r. [Online], 1-2. Available from http://www.atrf.info/papers/2010/2010_Ahmed_Rose_Jacob.pdf [Accessed 2nd March 2014].

Borgnat et al (2011). “SHARED BICYCLES IN A CITY”: *A SIGNAL PROCESSING AND DATA ANALYSIS PERSPECTIVE.* World Scientific Publishing Company. [Online], 1-25. Available from hal.archivesouvertes.fr/docs/00/49/03/25/PDF/velov_acs.pdf [Accessed 01 march 2014].

Capital Bikeshare. (2014).System Data. [Online], Available from http://www.capitalbikeshare.com/system-data [Accessed 28th February 2014].

Etienne, C. and Latifa, O. (2012). ‘Model-based count series clustering for Bike Sharing System usage mining’, a case study with the V´elib’ system of Paris. [Online].1-22 Available From http://www.comeetie.fr/pdfrepos/velibpp.pdf [Accessed 3rd March 2014].

Fanaee-T, et al. (2013). "Event labelling combining ensemble detectors and background knowledge", Progress in Artificial Intelligence (2013): pp. 1-15, Springer Berlin Heidelberg, doi: 10.1007/s13748-013-0040-3

Frulisch, J. et al. 2009. Sensing and Predicting the Pulse of the City through Shared Bicycling. [Online]. Available from www.nuriaoliver.com/bicing/IJCAI09_Bicing.pdf [Accessed 3rd March 2014].

Hornik, K. (2013). RWeka Odds and Ends. [Online],1-3. Available from: http://cran.rproject.org/web/packages/RWeka/vignettes/RWeka.pdf [Accessed on 28th February 2014].

Jdantos (2012) *Notes from Washington D.C. on cities, bicycling, public transit, baseball, music, politics, food*. Available from http://jdantos.wordpress.com/2012/03/01/capital-bikeshare-usage-and-weather/ [Accessed 15th April 2014].

Kutner et al. (2005). *Applied Linear Statistical Models*. Fifth Edition. ISBN: 0-07-238688-6. McGraw-HillfIrwin.

Lin, J. H. (2012). ‘A Geo-Aware and VRP-Based Public Bicycle Redistribution System’ International Journal of Vehicular Technology, 1-15 [Online], Available from http://www.hindawi.com/journals/ijvt/2012/963427/ [Accessed 4th March 2014].

Maimon et al. (2010). *Data Mining and Knowledge Discovery handbook*. 2nd Edition. Springer. ISBN 978-0-387-09822-7.

Quick, J. M. (2010). Statistical Analysis with R. Beginner's Guide. *Take control of your data and produce superior statistical analyses with R.* ISBN 978-1-849512-08-4. Packt Publishing. Birmingham, Uk.

Rattle (2012). Data Mining Tookit in R. [Online], Available from https://code.google.com/p/rattle/source/browse/trunk/src/model.R?r=677 [Accessed 5th March 2014].

R project, (2014). The Comprehensive R Archive Network. [Online], Available from http://CRAN.R-project.org [Accessed 3rd March 2014].

Shu, J. and Chou, M. C. (2011). “Models for Effective Deployment and Redistribution of Bicycles within Public Bicycle-Sharing Systems”, *Submitted to Operations Research manuscript OPRE-2011-02-077.R2* [Online], 1-27.Available from bschool.nus.edu/Staff/bizteocp/SMRT2013R2edited2.pdf [Accessed 3rd March 2014].

Teetor P. (2011). *R Cookbook*. ISBN: 978-0-596-80915-7. Published by O’Reilly Media, Inc. CA

UCI (machine Learning Repository) (2014). Bike Sharing Dataset Data Set. [online], Available from: http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset [Accessed on 20th February 2014].

William, G. (2011). *Data Mining with R and Rattle*: The Art of Excavating Data for Knowledge Discovery. Publisher: Springer. ISBN 978-1-4419-9890-3.

Witten H. and Frank E. (2005) *Data Mining*: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, San Francisco, 2nd edition, 2005.

Zhao, Y. (2012). *R and Data Mining*: Examples and Case Studies. Publisher: Academic Press, Elsevier. ISBN: 978-0-123-96963-7

# 9. Appendix