

Data Analysis and Regression Modelling on a Bike Share System

Peter Fitzgerald

Student No: 13119494



SUBMITTED AS PART OF THE REQUIREMENTS FOR
HIGHER DIPLOMA IN SCIENCE IN DATA ANALYTICS
NATIONAL COLLEGE OF IRELAND

28th MAY 2014

SUPERVISOR IOANA GHERGULESCU

Declaration Cover Sheet for Project Submission

SECTION 1 *Student to complete*

Name: PETER FITZGERALD
Student ID: X13119494
Supervisor: IOANA GHERGULESCU

SECTION 2 Confirmation of Authorship

The acceptance of your work is subject to your signature on the following declaration:

I confirm that I have read the College statement on plagiarism (summarised overleaf and printed in full in the Student Handbook) and that the work I have submitted for assessment is entirely my own work.

Signature: _____ Date: _____

NB. If it is suspected that your assignment contains the work of others falsely represented as your own, it will be referred to the College's Disciplinary Committee. Should the Committee be satisfied that plagiarism has occurred this is likely to lead to your failing the module and possibly to your being suspended or expelled from college.

Complete the sections above and attach it to the front of one of the copies of your assignment.

Acknowledgement

I would like to express my gratitude to my supervisor and lecturer in this module, Ioana Ghergulescu, for her advice that helped me to get started on the dissertation and for her support and advice offered throughout the process.

I would also like to thank Jonathan Lambert, NCI staff member, for his advice in helping the direction of this dissertation.

Abstract

The purpose of this study was to undertake a statistical analysis of a data set taken from a bike sharing scheme and provide a step by step guide to understanding the different variables within the dataset. The analysis looked at the variables, individually, and then looked at how they interacted with each other. The main aim of the study was to create a multiple linear regression model that would predict the number of bikes that should be made available at any given hour of the day given a certain set of weather conditions, which would act as the models input variables.

The data comes from a bike sharing system, giving an hourly count of bikes rented over a two year period; these systems have gained increasing support, especially in the last several years. As their popularity grows so does their relevance, and the need to study such systems, as to their validity, becomes ever more important.

The programming language R was used to build the linear regression model, with the end goal of predicting the count of bikes that should be available, given a certain set of weather conditions. The data set was divided into two parts, by year. The first year 2011 was used as the information to build the model and the actual output of 2012 was compared against the predicted results. The results from the model's predictions varied in success, some predictions were extremely close to the actual count; however there were a number of large differences between the predicted count and the actual count.

The model that was developed predicted well in places but more investigation is required to improve the predictions to a more accurate level.

Contents

	Page
Declaration	ii
Acknowledgement	iii
Abstract	iv
1 Introduction	1
1.1 Motivation	1
1.2 Aim	2
1.3 Objective	3
1.4 Solution Overview	3
1.5 Research Structure	3
2 Literature Review	4
2.1 History of Regression	4
2.2 Event Labelling Analysis	5
3 Implementation	7
3.1 Dataset	7
3.2 Preliminary Analysis	8
3.2.1 Range	10
3.2.2 Missing Values	10
3.3 Descriptive Statistics	11
3.4 Normality – Histograms, Shapiro Wilkes, Quantile-Quantile Plots	13
3.5 Correlations	16
3.6 Building Multiple Regression Model	19

4	Evaluation	22
	4.1 Check Normality Assumption in Residuals	22
	4.2 Model Fit	25
	4.3 Compare Predicted Results to 2012 Actual Count	27
5	Conclusions	27
	5.1 Overview	27
	5.2 Further Work	28
6	Bibliography	29
7	Appendix	31
	7.1 Project Proposal	31
	7.2 Requirements Specification	38
	7.3 Management Progress Report 1	51
	7.4 Management Progress Report 2	54
	7.5 Box Cox Transformation	57
	7.6 Additional Code	59
	7.6.1 Modelling Code	59
	7.6.2 Normality Tests	59
	7.6.3 Box Plots Code	60

List of Figures

	Page
3.1 Preliminary Analysis	9
3.2 Summary of Output	12
3.3 Histograms of attributes	14
3.4 Q-Q Plots of attributes	16
3.5 Correlation Scale	17
3.6 Correlation Matrix	18
3.7 Output of Model	21
4.1 Residuals	23
4.2 Residuals: Normality Plot	24
4.3 Summary of Model Output	25

1. Introduction

In recent years bike sharing schemes have become more commonplace in major cities all over the world. The service involves making bikes available for shared use, generally for a short term basis. Bike share eliminates the worries of ownership; the bike is taken from one location in the city and returned to another, generally the user's destination. Bike-sharing has experienced huge growth in recent years and

'as of April 2013 there were around 535 bike-sharing programmes around the world, made of an estimated fleet of 517,000 bicycles. In May 2011 there were around 375 schemes comprising 236,000 bikes.' (earth-policy.org, 2013)

1.1 Motivation

The huge rise in popularity in these schemes has been the motivation for choosing this topic. As their popularity grows, these systems should be evaluated as to their validity and if they make a good addition to cities around the world.

As cities become increasingly more congested, navigating around them has become increasingly difficult and these bike sharing schemes have profited from this modern problem. Local councils have been looking into the grave situation in cities caused by the number of cars travelling through the city:

'New traffic management plans being considered by the National Roads Authority could see a ban on cars driving through Dublin city centre. According to the Irish Times, a new orbital route has been proposed around the city which would mean the pedestrianisation of Suffolk Street and Church Lane.' (Landmark Digital Ltd, 2013)

Similar actions being taken in other cities; London has a congestion charge whereby people travelling in vehicles between 07:00 and 18:00, Monday to Friday incur a daily charge of £10. (Transport For London, n.d.). These actions by local councils will and have had a huge impact on the ever increasing popularity of the bike-share scheme.

Theft and vandalism has also been an influencing factor increasing the number of people that have adopted the bike-share method of transportation rather than using their own bikes which may be targeted. These systems also give people a huge sense of freedom and reduce the degree of worry which may come with leaving your own bicycle unmonitored while you work.

People are becoming ever more knowledgeable about environmental issues and their own carbon footprint; they have become more enthusiastic about transportation that has a less negative effect on the environment we live in.

It may come as no surprise that China's bike share fleets are among the largest in the world. The scheme in Paris comprises around 20,000 bicycles and 1,450 bicycle stations, which is the largest outside of China. Spain has the largest number of systems with 132, then Italy with 104 and China with 79. (Cities Today, 2014)

Healthier lifestyles among urban dwellers could also be attributed to the increase in the popularity of these bike share schemes around the world, fear over obesity levels and the ever increasing emphasis from the media on fitness will only encourage people to invest more time in adopting a more active lifestyle.

1.2 Aim

All these factors discussed have helped us gain an insight into the rise in popularity of these programmes in several major cities. Their rising popularity has been the catalyst for this report as more people are changing their means of transportation; we investigate the viability of these schemes. Our study will look at how varying weather conditions affect the likelihood of someone renting a bike at certain times on a given day. The data set, which we will describe in more detail in section 3.1 under the chapter Implementation, comes from the UCI Machine Learning Repository. The 'dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.' (K. Bache and M. Lichman, 2013)

1.3 Objective

Our main aim is to build a multiple linear regression model that will be used to predict the number of bikes that will be required at any given hour of the day relative to the weather conditions present at that particular time. Our model will take a number of variables, mainly variables attributed to the weather and try to predict the number of bikes that would be needed to supply the demand at that time. Our main objective will be preceded by a statistical analysis of all the attributes within the data set.

1.4 Solution Overview

In our analysis, we will use Microsoft Excel for our initial analysis. The file was downloaded as a Comma Separated file (CSV) and uploaded into R-Studio, the main body of our analysis will be done through use of the programming language R. As mentioned in the objective of the research our main aim is to build a multiple regression model, but before we get to that position there are several issues that must be addressed in our data set; the data must first be checked for any anomalies then any assumptions outlined for a multiple regression model must be adhered to before the process of selecting the independent variables which will be the base behind our predicted dependent variable, i.e. the number of bikes that will be required by the system at a particular time of day, depending on weather conditions . All the steps including building the model will be accomplished through using R and we will also use Microsoft Excel to some extent in dealing with some of the earlier analysis.

1.5 Research Structure

The main structure of the research has been broken up into five main parts. Chapter 1 will look at the introduction to the topic, the motivation behind choosing this topic for the analysis, the aim and objective of the analysis and finally an overview of the solution. Chapter 2 is a review of literature, first looking the history behind multiple regression model building and then we will discuss the analysis that was originally conducted on the bike dataset in the paper 'Event labelling combining ensemble detectors and background

knowledge' by Fanaee-T, Hadi and Gama, Joao. Chapter 3 deals with the Implementation, here we will describe the data set in full and proceed to do the full analysis on the bike data set. This chapter is broken up into several subsections including preliminary analysis, descriptive statistics, and normality then finally building the actual multiple regression model. Chapter 4 Evaluation, will involve analysis of the results from the preceding chapter - our main focus will be on the residuals. Finally, the Conclusion will be discussed in chapter 5. (Note: the words variables and attributes will be used interchangeably throughout the chapters in this report. Any writing in the colour blue that is prefixed with the symbol '>' refers to any R code that was used in the preparation of any required output)

2. Literature Review

In this section we will discuss the history of regression and some of its earliest applications and the people that were credited with first uncovering this statistical practice, and then in section 2.2 – Event Labelling Analysis, we will look at another study that used the same data set that is the basis of this study and see how it was used in their study. But first we will look at the history of regression

2.1 History of Regression

Although Sir Francis Galton is credited with coining the term “regression,” its application was in use by both Carl Gauss and Adrien-Marie Legendre at the beginning of the nineteenth century. Gauss’s work in mathematics and statistics, being the central idea behind regression theory as it is known today, in particular his work in relation to least squares published in 1821. (Guass 1821)

Galton’s work, on the hereditary characteristics of sweet-pea, led to the general idea of regression. The descriptions of his observations were regarding the biological regression of subsequent generations of sweet-pea harvests. The regression he referred to was in relation to the heights of the plants and their tendency toward a central mean. (Galton 1894) His first observations of regression were of plotting the heights of plants against the height of

their offspring. This produced a two dimensional plot which would become the basis of statistical regression.

Galton's work was in some way given a broader meaning by the work of Karl Pearson. He became Galton's biographer and often described the importance of the regression slope which Galton had created, in particular its simplicity of understanding the linear relationship between two "characteristics of arrays." (Pearson 1930) Pearson also published thorough analysis of both regression and correlation. (Pearson 1896) Using an advanced statistical proof, he established that the optimal values of the slope and coefficients could be calculated from the product moment.

Pearson recorded in Galton's biography that Galton had noted that it was possible to influence offspring characteristics, and also that characteristic traits often skipped generations, or even multiple generations. He determined that by looking back in time at the characteristics of past generations, that each generation back had only half as much influence on the possibility of present characteristics. This idea of diminishing influence gave rise to the idea of multiple regression, that a single characteristic could be influenced by a number of factors each with a differing degree of power. Galton was in some way the imagination behind both linear and multiple regression with Pearson expanding these ideas with his mathematical models. Pearson continued to work on multiple regression models throughout his life and also developed other statistical models including work with Chi squared distribution.

2.2 Event Labelling Analysis

The data analysed in this report was an amalgamation of bike rental information, weather information and holiday schedule. This amalgamation was constructed in order to carry out a very different type of analysis to that of a multiple linear regression model. The idea behind rental or the bike sharing scheme is relatively new, but what is more interesting in terms of data analysis is the automated nature of the system. The scheme records accurately durations and times of journey, information which can only be approximated by other forms of public transport. The accuracy and individuality of the recording allows for

the determination of unique events or occurrences that affect the pattern of rentals within the scheme. The architects (Fanaee-T & Gama 2013) of this amalgamated data made great use of this feature in their paper. Their paper focuses on event labelling, a type of artificial intelligence which analyses the patterns in bike rentals and correlates them with local and world events. These events which distort rental patterns can be attributed to physical occurrences such as weather, or social circumstances such as holidays or public disputes. Their work goes far beyond what is being attempted here, a simple predictive model. So where a regression model produces a

n averaging effect, where multiple factors such as climate and holidays can be accounted for to a certain extent, Fanaee and Gama introduce intelligence into their model to account for these events. Their model compiles information from search engines and data on Twitter trends to explain irregularities in the model. Although this is useful in hindsight, its application to future events is only possible when known events are foreseen, such as a scheduled vote, a planned strike or simply a public holiday. "Event Detection" is a human activity, and therefore can be expensive without an automated system. The goal of Fanaee and Gama's intelligent system was to achieve the same result without the expense of human experts. However they too noted that the "bike rental data is highly correlated with environmental and periodicity settings." Settings such as "temperature and hour of the day, month, work of the day (weekend, weekday, and holiday)" are key to the system. They noted that a "regression tree model can make a prediction based on these environmental attributes very close to actual counts." And like the analysis shown here through regression they stated that "bike rental count time series should not be analyzed without taking into account the environmental attributes." So despite the underlying complexity of their intelligent model the same simplicity inherent in linear regression is clearly evident and required for the system to operate correctly.

The advantages of the more complicated system is that event detection on bike sharing data can be "incorporated in a decision support system" for improved scheduling and management of the bike rental scheme. It can also be used in a system for "alarming or suggestion purposes." For example, suggesting individuals remain indoors or use alternative travel methods because of extreme weather conditions or encourage them to participate in a social event in the area.

The model which will be developed here based on multiple linear regression takes its data from the same sources but unlike the analysis that went before our goal will be to create a model which will predict the number of bikes that will be required on the hour given a set weather conditions that are housed in the data set.

3. Implementation

3.1 Data Set

The data set we are using for our analysis is from a two-year usage log of bikes being rented in a bike sharing system in Washington, D.C., USA, known as Capital Bike Sharing (CBS). The reason why this data set is of relevance for our analysis is the data has been compiled over a two year period giving us ample data in which to build a model. The data is available in two formats daily and hourly, during this analysis we will only be looking at the hourly data set.

The data is saved in a comma separated file (CSV) with 17 attributes. All the data has already been converted to numeric values. The following is a list of the attributes from the dataset, and a brief description of their purpose:

1. Instant: The number of the instant (there are 17,379 instances in the hourly data)
2. Dteday: Date of the year (range: 1st Jan 2011 – 31st Dec 2012)
3. Season: 0: Winter, 1: Spring, 2: Summer, 3: Autumn
4. Yr: The year, either: 0: 2011 or 1: 2012
5. Mnth: The month of the year: 1: Jan, 2: Feb, 3: Mar, 4: Apr, 5: May, 6: Jun, 7: Jul, 8: Aug, 9: Sep, 10: Oct, 11: Nov, 12: Dec
6. Hr: Hour of the day, working over a twenty-four hour period. (0-23 hrs)
7. Holiday: This refers to whether the day is a public holiday or not. 0: No, 1: Yes
8. Weekday: Refers to the day of the week, 0: Sunday, 1: Monday, 2: Tuesday, 3: Wednesday, 4: Thursday, 5: Friday, 6: Saturday.

9. Workingday: If day is a working day: 1, if day is weekend/holiday: 0
10. Weathersit: this breaks the day up into 4 weather categories:
- 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 Pellets + Thunderstorm + Mist, Snow + Fog
11. temp : Normalized temperature in Celsius. The values are divided to 41 (max)
12. atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)
13. hum: Normalized humidity. The values are divided to 100 (max)
14. windspeed: Normalized wind speed. The values are divided to 67 (max)
15. casual: count of casual bike users
16. registered: count of registered bike users
17. cnt: count of total rental bikes including both casual and registered (K. Bache and M. Lichman, 2013)

3.2 Preliminary Analysis

After acquiring the data one of the most important issues, and sometimes the most commonly overlooked elements is cleaning the data. Our data has been taken from a study that has been carried out already, but this should not give cause to be complacent and ignore the fact we should check the data first before performing any preliminary analysis on it. We will systematically check the dataset and make sure that the data we have is reasonable, we will then visualise the data to try to gain a better understanding of each attribute, this will help ensure that the model that we build should be absent of errors that could come from unclean data . 'Error prevention is far superior to error detection and cleaning, as it is cheaper and more efficient to prevent errors than to try and find them and correct them later.' (Chapman, 2005)

The first task lies in examining each of the 17 attributes individually, to see if the data under each of these headings has been collected accurately. Our preliminary check on the data is

performed in Microsoft Excel, by looking at the CSV file. We filter each row of the data and make sure that there are no inappropriate entries. All the entries in the data set should correspond with what has been outlined in the list of attributes, and the check should evaluate that each instance of each attribute falls within the parameters given for the particular attribute. For example in figure 3.1 below the attribute 'weekday' clearly shows that all the instances fall within the parameters (0-6) confirming that there are no anomalies with regard to this attribute.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
2	1	01/01/2011							0	1	0.24	0.2879	0.81	0	3	13	16
3	2	01/01/2011							0	1	0.22	0.2727	0.8	0	8	32	40
4	3	01/01/2011							0	1	0.22	0.2727	0.8	0	5	27	32
5	4	01/01/2011							0	1	0.24	0.2879	0.75	0	3	10	13
6	5	01/01/2011							0	1	0.24	0.2879	0.75	0	0	1	1
7	6	01/01/2011							0	2	0.24	0.2576	0.75	0.0896	0	1	1
8	7	01/01/2011							0	1	0.22	0.2727	0.8	0	2	0	2
9	8	01/01/2011							0	1	0.2	0.2576	0.86	0	1	2	3
10	9	01/01/2011							0	1	0.24	0.2879	0.75	0	1	7	8
11	10	01/01/2011							0	1	0.32	0.3485	0.76	0	8	6	14
12	11	01/01/2011							0	1	0.38	0.3939	0.76	0.2537	12	24	36
13	12	01/01/2011							0	1	0.36	0.3333	0.81	0.2836	26	30	56
14	13	01/01/2011							0	1	0.42	0.4242	0.77	0.2836	29	55	84
15	14	01/01/2011							0	2	0.46	0.4545	0.72	0.2985	47	47	94
16	15	01/01/2011							0	2	0.46	0.4545	0.72	0.2836	35	71	106
17	16	01/01/2011							0	2	0.44	0.4394	0.77	0.2985	40	70	110
18	17	01/01/2011							0	2	0.42	0.4242	0.82	0.2985	41	52	93
19	18	01/01/2011							0	2	0.44	0.4394	0.82	0.2836	15	52	67
20	19	01/01/2011							0	3	0.42	0.4242	0.88	0.2537	9	26	35
21	20	01/01/2011							0	3	0.42	0.4242	0.88	0.2537	6	31	37
22	21	01/01/2011							0	2	0.4	0.4091	0.87	0.2537	11	25	36
23	22	01/01/2011							0	2	0.4	0.4091	0.87	0.194	3	31	34
24	23	01/01/2011							0	2	0.4	0.4091	0.94	0.2239	11	17	28
25	24	01/01/2011							0	2	0.46	0.4545	0.88	0.2985	15	24	39
26	25	02/01/2011							0	2	0.46	0.4545	0.88	0.2985	4	13	17
27	26	02/01/2011							0	2	0.44	0.4394	0.94	0.2537	1	16	17
28	27	02/01/2011							0	2	0.42	0.4242	1	0.2836	1	8	9
29	28	02/01/2011							0	2	0.46	0.4545	0.94	0.194	2	4	6
30	29	02/01/2011	1	0	1	4	0	0	0	2	0.46	0.4545	0.94	0.194	2	1	3
31	30	02/01/2011	1	0	1	6	0	0	0	3	0.42	0.4242	0.77	0.2985	0	2	2
32	31	02/01/2011	1	0	1	7	0	0	0	2	0.4	0.4091	0.76	0.194	0	1	1
33	32	02/01/2011	1	0	1	8	0	0	0	3	0.4	0.4091	0.71	0.2239	0	8	8

Fig. 3.1 – Preliminary Analysis

This check is continued for all seventeen attributes, on this preliminary search no irregularities were found in the data. However, the limitations of excel operating on large data sets is realised immediately and it is more challenging to find irregularities in the attributes with continuous variables, such as temperature, humidity and wind speed therefore importing the data into R-Studio will give us more scope in assessing whether our acquired data is accurate.

3.2.1 Range

We were unable to view if all the instances have been accounted for in Excel, to a strong degree of certainty. By looking at the data set in R we can utilise the range function to confirm that all the instances are present:

<code>>range(bike\$instant)</code>	[1] 1 17379
---------------------------------------	-------------

The result confirms that all 17,379 are present and accounted for in the dataset. The range function may also be used to verify that the parameters for the attributes with continuous variables have also been met.

<code>> range(bike\$temp)</code>	[1] 0.02	1.00
<code>> range(bike\$atemp)</code>	[1] 0	1
<code>> range(bike\$hum)</code>	[1] 0	1
<code>> range(bike\$windspeed)</code>	[1] 0.0000	0.8507
<code>> range(bike\$casual)</code>	[1] 0	367
<code>> range(bike\$registered)</code>	[1] 0	886
<code>> range(bike\$cnt)</code>	[1] 1	977

The output for these ranges may appear unusual at first, but remember that temp & atemp have been divided by 41 (41 being the max temperature recorded) and 50 (max temperature) respectively. Likewise, humidity has been divided by 100. So any readings over the value of 1 would have been more unusual. From this, we can imply that the data so far looks valid. The last three ranges refer to the count of bikes which have been used at any hour over the two year period. From our preliminary examination there is no evidence that there has been any recording of unreasonable data thus far.

3.2.2 Missing Values

From viewing the data so far there is very little evidence of any missing values within the dataset, we can establish if that is true in R by using the function below:

> summary(is.na(bike))

Instant	Dteday	season	yr	Mnth
Mode :logical	Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:17379	FALSE:17379	FALSE:17379	FALSE:17379	FALSE:17379
NA's :0	NA's :0	NA's :0	NA's :0	NA's :0

Hr	Holiday	weekday	workingday	weathersit
Mode :logical	Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:17379	FALSE:17379	FALSE:17379	FALSE:17379	FALSE:17379
NA's :0	NA's :0	NA's :0	NA's :0	NA's :0

Temp	Atemp	hum	windspeed	Casual
Mode :logical	Mode :logical	Mode :logical	Mode :logical	Mode :logical
FALSE:17379	FALSE:17379	FALSE:17379	FALSE:17379	FALSE:17379
NA's :0	NA's :0	NA's :0	NA's :0	NA's :0

Registered	cnt
Mode :logical	Mode :logical
FALSE:17379	FALSE:17379
NA's :0	NA's :0

The output clearly confirms that all the attributes in the dataset contain values, there are no missing values. This function has allowed us to search every instance under every attribute and return that there are no missing values (NA's) in the entire dataset.

3.3 Descriptive Statistics

Descriptive statistics are used to summarise a sample, they differ from inferential statistics in that they are not used to infer anything about the population in which the data came from. As their name suggests they help us describe the data. Information regarding the mean, median and the mode are usually discussed at the forefront when talking about descriptive statistics. Descriptive statistics are used mainly to get a feel for the data, secondly for use in statistical tests and thirdly to indicate the error that are associated with

results and sometimes with graphical outputs (Gaten, 2000). We can invoke the summary function in R to give us an output of some of the commonly used descriptive statistics.

> summary(bike)

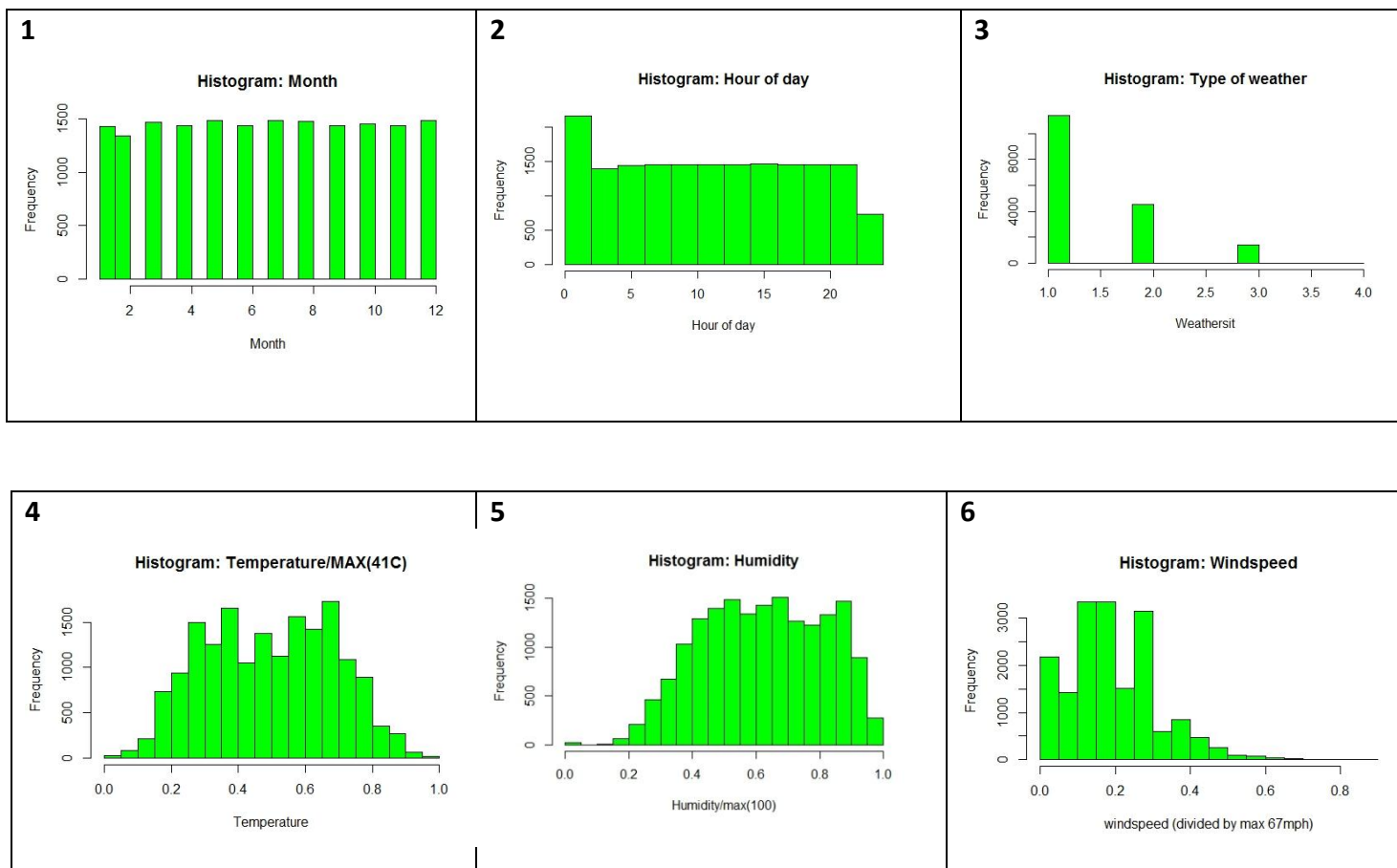
instant	dteday	season	yr	mnth	hr
Min. : 1	01/01/2011: 24	Min. : 1.000	Min. : 0.0000	Min. : 1.000	Min. : 0.00
1st Qu.: 4346	01/01/2012: 24	1st Qu.: 2.000	1st Qu.: 0.0000	1st Qu.: 4.000	1st Qu.: 6.00
Median : 8690	01/02/2012: 24	Median : 3.000	Median : 1.0000	Median : 7.000	Median : 12.00
Mean : 8690	01/03/2011: 24	Mean : 2.502	Mean : 0.5026	Mean : 6.538	Mean : 11.55
3rd Qu.: 13034	01/03/2012: 24	3rd Qu.: 3.000	3rd Qu.: 1.0000	3rd Qu.: 10.000	3rd Qu.: 18.00
Max. : 17379	01/04/2011: 24	Max. : 4.000	Max. : 1.0000	Max. : 12.000	Max. : 23.00
	(Other) : 17235				
holiday	weekday	workingday	weathersit	temp	atemp
Min. : 0.00000	Min. : 0.000	Min. : 0.0000	Min. : 1.000	Min. : 0.020	Min. : 0.0000
1st Qu.: 0.00000	1st Qu.: 1.000	1st Qu.: 0.0000	1st Qu.: 1.000	1st Qu.: 0.340	1st Qu.: 0.3333
Median : 0.00000	Median : 3.000	Median : 1.0000	Median : 1.000	Median : 0.500	Median : 0.4848
Mean : 0.02877	Mean : 3.004	Mean : 0.6827	Mean : 1.425	Mean : 0.497	Mean : 0.4758
3rd Qu.: 0.00000	3rd Qu.: 5.000	3rd Qu.: 1.0000	3rd Qu.: 2.000	3rd Qu.: 0.660	3rd Qu.: 0.6212
Max. : 1.00000	Max. : 6.000	Max. : 1.0000	Max. : 4.000	Max. : 1.000	Max. : 1.0000
hum	windspeed	casual	registered	cnt	
Min. : 0.0000	Min. : 0.0000	Min. : 0.00	Min. : 0.0	Min. : 1.0	
1st Qu.: 0.4800	1st Qu.: 0.1045	1st Qu.: 4.00	1st Qu.: 34.0	1st Qu.: 40.0	
Median : 0.6300	Median : 0.1940	Median : 17.00	Median : 115.0	Median : 142.0	
Mean : 0.6272	Mean : 0.1901	Mean : 35.68	Mean : 153.8	Mean : 189.5	
3rd Qu.: 0.7800	3rd Qu.: 0.2537	3rd Qu.: 48.00	3rd Qu.: 220.0	3rd Qu.: 281.0	
Max. : 1.0000	Max. : 0.8507	Max. : 367.00	Max. : 886.0	Max. : 977.0	

Fig. 3.2 – Summary Output

Our output from Figure 3.2 gives us details on all 17 variables in our dataset: the minimum & maximum, along with the 1st & 3rd quartile. The mean and median are also given and are of particular interest if we are trying to discover any anomalies like outliers that may disrupt our findings when later building our model. We check for any noticeable differences between the mean and median, as the mean is more susceptible to outliers, as it is the average across the total dataset, it could conceivably be distorted greatly by a large unusual outlier or a number of outliers. The median is not affected like the mean by the presence of outliers, as it is simply a calculation of the middle point in our dataset. Therefore we will compare these two values from figure 3.2 - Summary Output to see if there are any large differences between these two figures. Three of the attributes: casual, registered and count are displaying some noticeable differences between their median and mean. Casual users recorded a median and mean of 17.00 and 35.78 respectively, registered users recorded a median and mean of 115.00 and 153.80 respectively, and count (total of casual & registered users) recorded a median and mean of 142.00 and 189.50 respectively. We will investigate these attributes further in the following section on Normality.

3.4 Normality – Histograms, Shapiro Wilkes, Quantile-Quantile Plots

Next we will look at each individual attribute in R and visualise them, this can help to bring an added dimension and a better understanding of the data we are looking at, viewing the shape and dispersion of a data output can be hugely beneficial and is helpful in seeing immediately how change in one variable can drive change in another. Graphs can provide essential meaning to data and help in the future model building process.



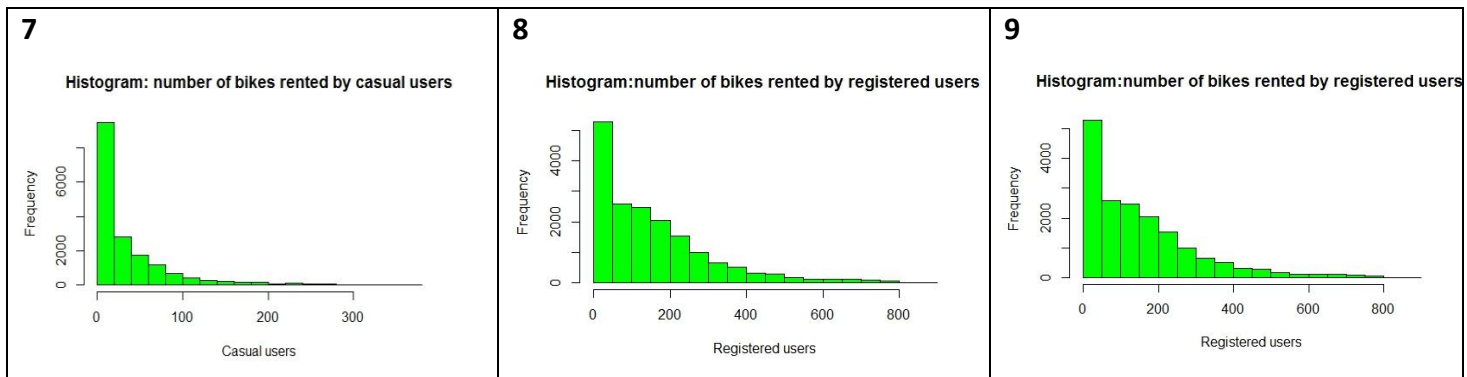


Fig. 3.3 - Histograms of attributes

From figure 3.3 – Histograms of attributes, we can see that most of our variables do not follow a normal distribution. Diagrams 1 - 3 are not normally distributed variables, variables 1 and 3 are types of nominal data. Nominal refers to data that is categorical, ‘nom’ being derived from the latin meaning name. No computations can be performed on nominal data, for example subtracting one month from another has no meaning. Diagrams 4 - 6 do not follow a normal distribution either. Diagrams 4 and 5 are examples of interval data, these types of data have a scale, e.g. Celsius. Diagrams 7, 8 and 9 are positively skewed; these three histograms are examples of discrete data. Histograms can also be used in the identification of outliers. If any outliers are identified we must take careful consideration before rashly dismissing it as an incorrect input. It could in fact be a true correct value which is just highly irregular, it may be caused by an unusual event, and therefore outliers should not be dismissed too quickly without proper examination of the outlier itself and any other variables associated with the outlier e.g. the date which may shed light on its ‘raison d’etre’. For instance there may be a huge uptake in the number of bikes used but if the date was investigated it may indicate that it was a bank holiday with a number of events happening around the city, it may be considered an outlier but the data recorded is correct.

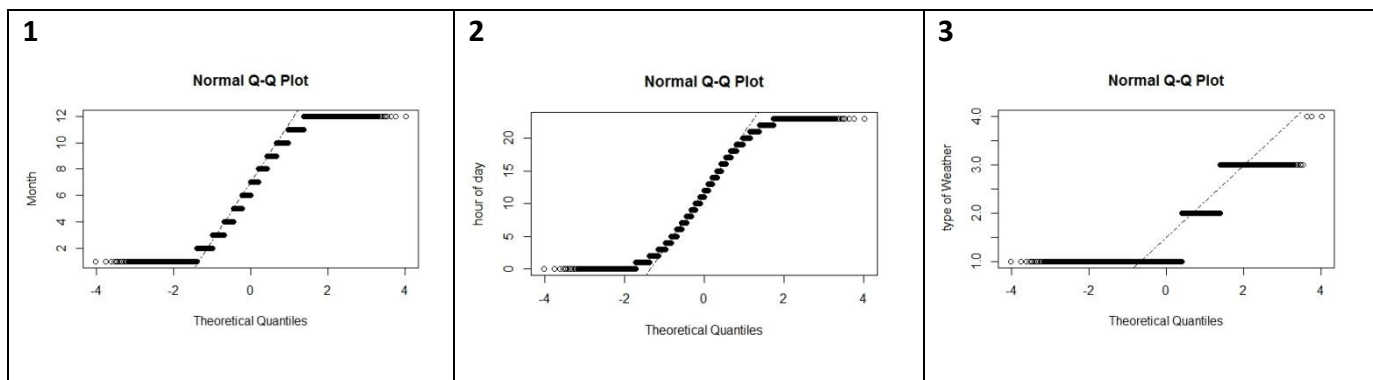
From the histograms above, in particular histograms 7, 8 & 9, which gives a count on the number of casual, registered & total users respectively, which are positively skewed. As a value of less than zero cannot be recorded because the lowest number of bikes that can be used at any time can only be zero, it is not usual to have a distribution that is positively skewed when dealing with this type of data. As we would expect the lower the number used the higher the frequency, and

the frequency drops the higher the usage of bikes, creating the high to low movement in the histogram as we move from left to right in diagrams 7, 8 & 9.

At this point we would usually use the Shapiro-Wilks normality test to establish if the distributions of our variables that we plotted in the above histograms are normally distributed but as our sample is greater than 5000 instances, and when using R this test can only be achieved on samples up to 5000. However, we will explain how the test works: the null hypothesis of the Shapiro-Wilks test is that the population is normally distributed. The test returns a p-value, therefore if a p-value is returned that is less than the agreed alpha level, we will reject the null hypothesis in favour of the alternative hypothesis, i.e. we would have found evidence that our data is not normally distributed. We can however use Quantile-Quantile plots instead to establish if the data is normal or not.

Quantile-Quantile Plots (Q-Q Plots) are usually used to confirm the results from the Shapiro-Wilks Test, but as we did not conduct a Shapiro-Wilks we will use the results from the Q-Q plots solely to confirm if our variables are normally distributed or not.

We will now look at our variables in the bike dataset and confirm if they follow a normal distribution by now plotting Q-Q Plots in R.



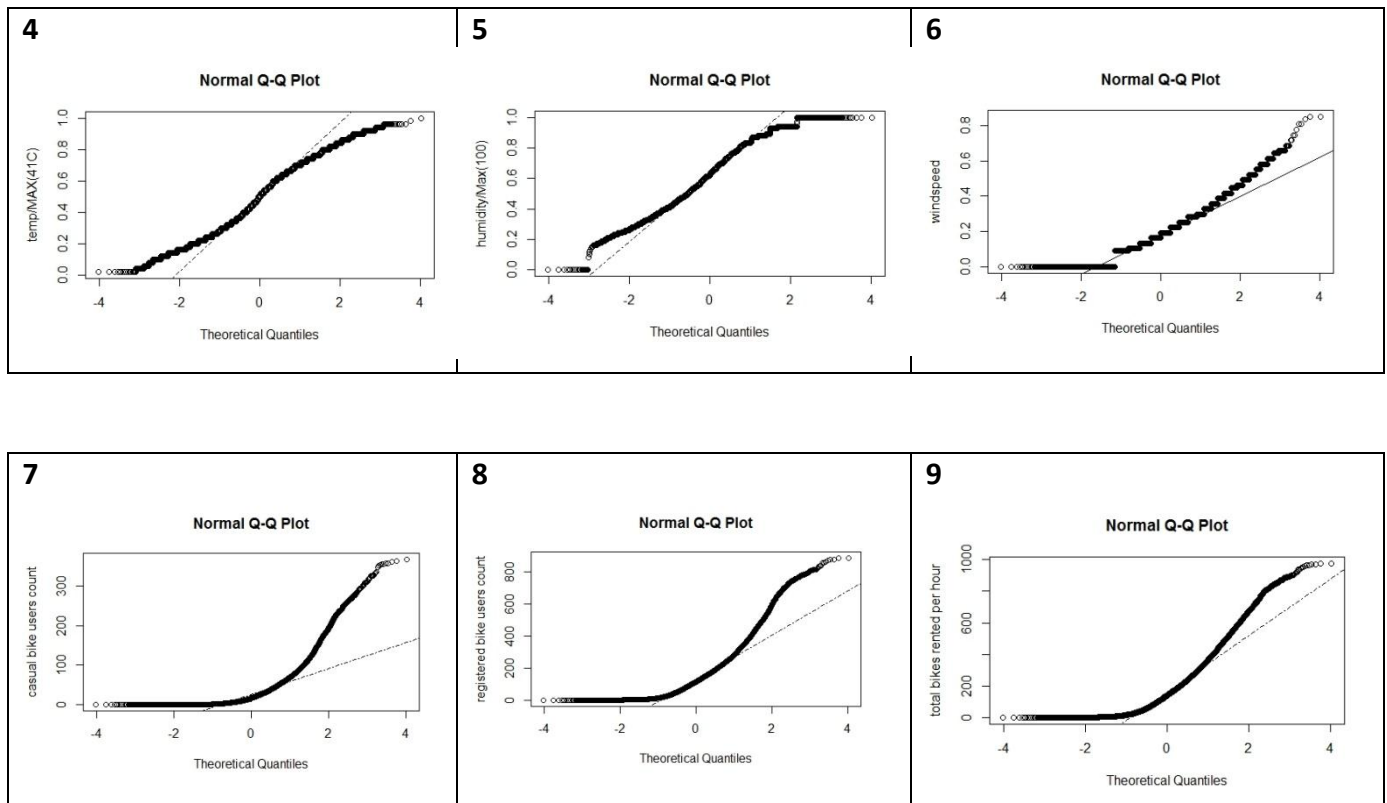


Fig. 3.4 – Q-Q Plots of attributes

The Q-Q plots in figure 3.4 help us distinguish normally distributed variables from non-normally distributed variables, the thick dark line in each of the plots represents the distribution of the actual variable from the bike data set; the straight line is a mapping of what a normal distribution would look like, therefore the closer the actual distribution is to the straight line the more likely the distribution is to being normal.

Q-Q Plots 4, 5 & 6 are examples of interval data, and are of significant interest to us as they will become the major influencing factors for our prediction in the multiple regression model. These plots together with the histograms numbered 4, 5 & 6 from earlier help us in seeing that these variables (temperature, humidity and wind speed) are not normally distributed.

3.5 Correlations

We could look at individual correlations between each variable in our data set, instead we will look at how all the variables within the dataset are correlated by using the correlation

function in R, this will return a correlation matrix, showing how each individual attribute is related to all other attributes. The correlation matrix will give us an insight into which attributes are related, the strength of the relationship is given by the number, as a rule of thumb, values between 0 and 0.25 are considered weak, 0.25 to 0.75 are moderate, and values between 0.75 and 1.00 are strong. If the value is preceded by a negative symbol the relationship is considered indirect/negative, in that, as one variable increases the other falls, i.e. they have a negative relationship. If the value is positive, they have a direct relationship, i.e. as one variable increases so does the other. A value of 1 or -1 is deemed to have a perfect correlation. Figure 3.5 shows a scale of the possible correlations.

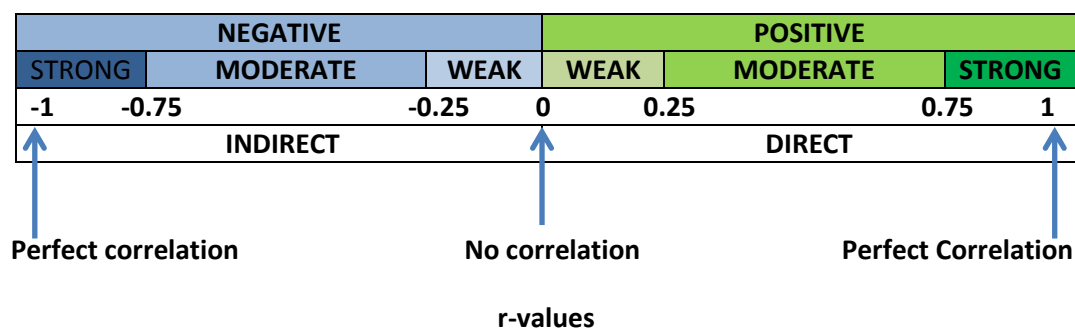


Fig. 3.5 - Correlation Scale

Before we can achieve a correlation matrix we must first remove the attribute 'Dteday', due to the fact this attribute is non-numeric and for correlation to work all the variables must be numeric. We can remove the date attribute from the data frame using the following code in R, as 'Dteday' occurred second in the bike data frame, we create a new data frame excluding the second attribute. With date removed we can generate the correlation matrix.

```
> newbike <- bike[c(1,3:17)]
> cor(newbike)
```


	instant	season	yr	mnth	hr	holiday	weekday	workingday	weathersit
instant	1.000000000	0.404045721	0.866014049	0.489163831	-0.004774815	0.014723494	0.001356820	-0.003415559	-0.014197603
season	0.404045721	1.000000000	-0.010742486	0.830385892	-0.006116901	-0.009584526	-0.002335350	0.013743102	-0.014523552
yr	0.866014049	-0.010742486	1.000000000	-0.010472929	-0.003867005	0.006691617	-0.004484851	-0.002196005	-0.019156853
mnth	0.489163831	0.830385892	-0.010472929	1.000000000	-0.005771909	0.018430325	0.010400061	-0.003476922	0.005399522
hr	-0.004774815	-0.006116901	-0.003867005	-0.005771909	1.000000000	0.000479136	-0.003497739	0.002284998	-0.020202528
holiday	0.014723494	-0.009584526	0.006691617	0.018430325	0.000479136	1.000000000	-0.102087791	-0.252471370	-0.017036113
weekday	0.001356820	-0.002335350	-0.004484851	0.010400061	-0.003497739	-0.102087791	1.000000000	0.035955071	0.003310740
workingday	-0.003415559	0.013743102	-0.002196005	-0.003476922	0.002284998	-0.252471370	0.035955071	1.000000000	0.044672224
weathersit	-0.014197603	-0.014523552	-0.019156853	0.005399522	-0.020202528	-0.017036113	0.003310740	0.044672224	1.000000000
temp	0.136178007	0.312025237	0.040913380	0.201691494	0.137603494	-0.027340477	-0.001794927	0.055390317	-0.102639936
atemp	0.137614610	0.319379811	0.039221595	0.208096131	0.133749965	-0.030972737	-0.008820945	0.054667235	-0.105563108
hum	0.009576774	0.150624745	-0.083546421	0.164411443	-0.276497828	-0.010588465	-0.037158268	0.015687512	0.418130329
windspeed	-0.074504540	-0.149772751	-0.008739533	-0.135386323	0.137251568	0.003987632	0.011501545	-0.011829789	0.026225652
casual	0.158295401	0.120206447	0.142778528	0.068457301	0.301201730	0.031563628	0.032721415	-0.300942486	-0.152627885
registered	0.282045777	0.174225633	0.253684310	0.122272967	0.374140710	-0.047345424	0.021577888	0.134325791	-0.120965520
cnt	0.278378694	0.178055731	0.250494899	0.120637760	0.394071498	-0.030927303	0.026899860	0.030284368	-0.142426138
temp	0.136178007	0.137614610	0.009576774	-0.074504540	0.158295401	0.28204578	0.27837869		
season	0.312025237	0.319379811	0.150624745	-0.149772751	0.12020645	0.17422563	0.17805573		
yr	0.040913380	0.039221595	-0.083546421	-0.008739533	0.14277853	0.25368431	0.25049490		
mnth	0.201691494	0.208096131	0.164411443	-0.135386323	0.06845730	0.12227297	0.12063776		
hr	0.137603494	0.133749965	-0.276497828	0.137251568	0.30120173	0.37414071	0.39407150		
holiday	-0.027340477	-0.030972737	-0.010588465	0.003987632	0.03156363	-0.04734542	-0.03092730		
weekday	-0.001794927	-0.008820945	-0.037158268	0.011501545	0.03272142	0.02157789	0.02689986		
workingday	0.055390317	0.054667235	0.015687512	-0.011829789	-0.30094249	0.13432579	0.03028437		
weathersit	-0.102639936	-0.105563108	0.418130329	0.026225652	-0.15262788	-0.12096552	-0.14242614		
temp	1.000000000	0.987672139	-0.069881391	-0.023125262	0.45961565	0.33536085	0.40477228		
atemp	0.987672139	1.000000000	-0.051917696	-0.062336043	0.45408007	0.33255864	0.40092930		
hum	-0.069881391	-0.051917696	1.000000000	-0.290104895	-0.34702809	-0.27393312	-0.32291074		
windspeed	-0.023125262	-0.062336043	-0.290104895	1.000000000	0.09028678	0.08232085	0.09323378		
casual	0.459615646	0.454080065	-0.347028093	0.090286775	1.000000000	0.50661770	0.69456408		
registered	0.335360849	0.332558635	-0.273933118	0.082320847	0.50661770	1.000000000	0.97215073		
cnt	0.404772276	0.400929304	-0.322910741	0.093233784	0.69456408	0.97215073	1.000000000		

Fig. 3.6 - Correlation Matrix

From the rules given already we can, start to examine our matrix from figure 3.5, and categorise attributes that have strong relationships and those that have moderate and weak relationships. At this point, it should be noted that we are particularly interested in identifying the variables that are strongly correlated, when building our model, we cannot include two variables with a strong correlation as the resulting model may return inexplicable results, and one of the attributes will suffice in helping the prediction. This phenomenon is known as Multicollinearity. To prevent any issues with the model we will only include one of these variables.

In our correlation matrix we can identify a number of highly correlated attributes:

Attribute_1	Attribute_2	Correlation
Year	Instant	0.87
Month	Season	0.83
Temperature	Atemp	0.99
Registered	Count	0.97
Casual	Count	0.69

When it comes to selecting attributes, i.e. the independent variables, to go into our model we must be careful to only include only 1 of the attributes from our list of highly correlated attributes. Some of the results from the correlation matrix are not that surprising, for instance we would expect that month and season would be strongly correlated and we would also expect that temperature and atemp (normalized feeling temperature) would be correlated too. Remember Registered refers to the number of registered users using the bikes and it is not unusual that this figure is extremely highly correlated with the count of bikes taken on an hourly basis.

3.6 Building the Multiple Regression Model

We must now consult the correlation matrix that we constructed. From the attributes that we identified as being highly correlated, we must remember that only one of these attributes will be used in the building of the model.

Let's first focus our attention on what we are trying to achieve when building the multiple regression model: we are building a model that will use a number of attributes from our bike dataset, these will be referred to as our independent variables, these represent the inputs into our model (i.e. month, temperature, humidity etc.) and these will be used in predicting our dependent variable, this is our outcome or effect as a result of the inputs. The dependent variable will be the count attribute, the number of bikes that have been rented at any given hour. Multiple Regression modelling, by its name, uses multiple independent variables to predict one dependent variable. Due to the fact Count has been chosen as our dependent variable, the

attributes casual (number of casual renters per hour) and registered (number of registered users per hour) will be made redundant in our model due to their high correlation with this dependent variable. If we include month we must be sure not to include season, and similarly for any other highly correlated attributes.

Before we move onto designing our multiple regression model, we must first split the dataset, the information from half the dataset will be used to create the multiple regression model. And the other half will be used to test results. We will split the bike dataset by year, the data from 2011 will be the data that is used in creating our Multiple Linear Regression Model. And the 2012 data will be used to confirm our results. Baring in mind that we are avoiding including any variables that are highly correlated, our selection of independent variables includes the following attributes:

Independent Variables	Dependent Variable
Month	Count
Hour	
Weathersit	
Temperature	
Humidity	
Windspeed	

Splitting the dataset is achieved by creating a subset in R, we subset the dataset by selecting all the instances that occur in the year 2011.

```
> split_bike<-subset(bike, bike$yr==0)
```

With the dataset split by year (remember the values in our data set for year are numeric and the year 2011 is represented by the number 0), we can now take the selected variables from the list above and create our model. We split the data due to the fact we are going to use a portion of the data to reconfirm our results from the model, if we included all the data, our model would not be predicting the results as we want it to, hence the decision to split the data in two, the data from 2011 will be used to form the prediction and the results will be compared against the 2012 figures.

The following R code is used to create our multiple linear regression model:

```
>bike_reg_model<-  
lm(split_bike$cnt~split_bike$mnth+split_bike$hr+split_bike$weathersit+split_bike$temp  
+split_bike$hum+split_bike$windspeed)
```

```
Call:  
lm(formula = split_bike$cnt ~ split_bike$mnth + split_bike$hr +  
    split_bike$weathersit + split_bike$temp + split_bike$hum +  
    split_bike$windspeed)  
  
Coefficients:  
            (Intercept)      split_bike$mnth      split_bike$hr  
            12.893         4.981         5.996  
split_bike$weathersit    split_bike$temp    split_bike$hum  
            -3.403        250.736       -142.904  
split_bike$windspeed  
            15.561
```

Fig. 3.7 – Output of Model

From figure 3.7, we can see that in our model we have included 6 independent variables; these come after the tilde character in the formula i.e. month, hour, weather situation, temperature, humidity and wind speed. All these independent variables will play a part in determining the dependent variable, i.e. the count of the bicycles used. The dependent variable is given before the tilde character in the formula.

We should recall at this point that the best fit line formula in simple linear regression is given by the formula:

$$y = a + bx \text{ (where } y \text{ is the dependent variable, } a=\text{intercept and } b=\text{slope of the line)}$$

Similarly, we can extend out this formula for the best fit line in multiple regression:

$$y = a + b_1x_1 + b_2x_2 + \dots + b_nx_n$$

As we are dealing with more than one independent variable, we have more than one x variable in our formula. Each of our independent variables, from our output in figure 3.6, represents one of these X variables:

Variable from formula	Attribute
a	Intercept
X1	Month
X2	Hour
X3	Weathersit
X4	Temperature
X5	Humidity
X6	Windspeed

So the formula for our best fit line is given by the formula:

$$y = 12.893 + 4.981 X_1 + 5.996 X_2 - 3.403 X_3 + 250.736 X_4 - 142.904 X_5 + 15.561 X_6$$

The value preceding each value of X represents the slope; the slope represents the change in y in relation to one unit change in X. In our multiple regression model the coefficient of the X variable is the amount by which the Y variable changes if our X variable increases by a value of one, and the remaining variables in the model remain constant.

4. Evaluation

4.1 Check Normality Assumption in Residuals

We have now created our multiple regression formula that will be used to predict the y-variable/dependent variable = count of total casual & registered users. To analyse the regression results, we must check that the normality assumption holds true on these residuals. Remember residuals are the difference between our predicted values and the actual values that were recorded for our dependent variable.

The following code in R will produce a graphical representation of the residuals:

```
>bike_reg_model.stdRes = rstandard(bike_reg_model)
>plot(bike_reg_model.stdRes, col="red")
>abline(0,0)
```

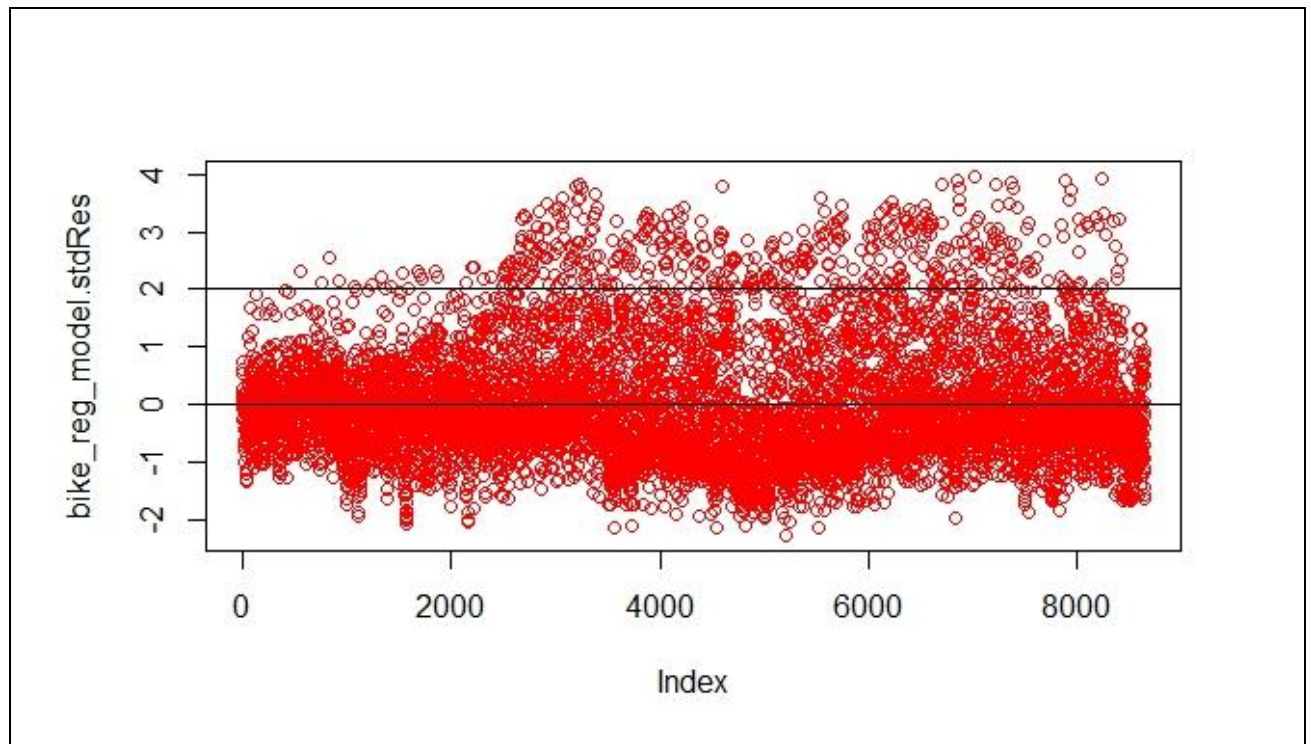


Fig. 4.1 – Residuals

For our model to be a good fit, we would like to see the residuals as close to zero as possible. The closer to zero our residuals are the more likely our model is predicting well. The y-axis, in the Residuals figure above, represents standard deviations. As we can see the majority of residuals fall within 2 standard deviations of the mean, i.e. between -2 and +2. However, there are a number of residuals between +2 and +4 standard deviations which suggests that our model is not predicting well in some instances.

We should now look at a normality plot to see if the residuals confirm to the assumption of normality. Just like we saw earlier when plotting the Q-Q plots we would hope that the residuals fall close to the straight line suggesting that they are normally distributed.

The following code is used to generate the Q-Q plot:

```
>qqnorm(bike_reg_model.stdRes, ylab="standardised residuals", xlab="Normal Scores",  
main="Residuals - Normality Plot", col="red")  
  
>qqline(bike_reg_model.stdRes)
```

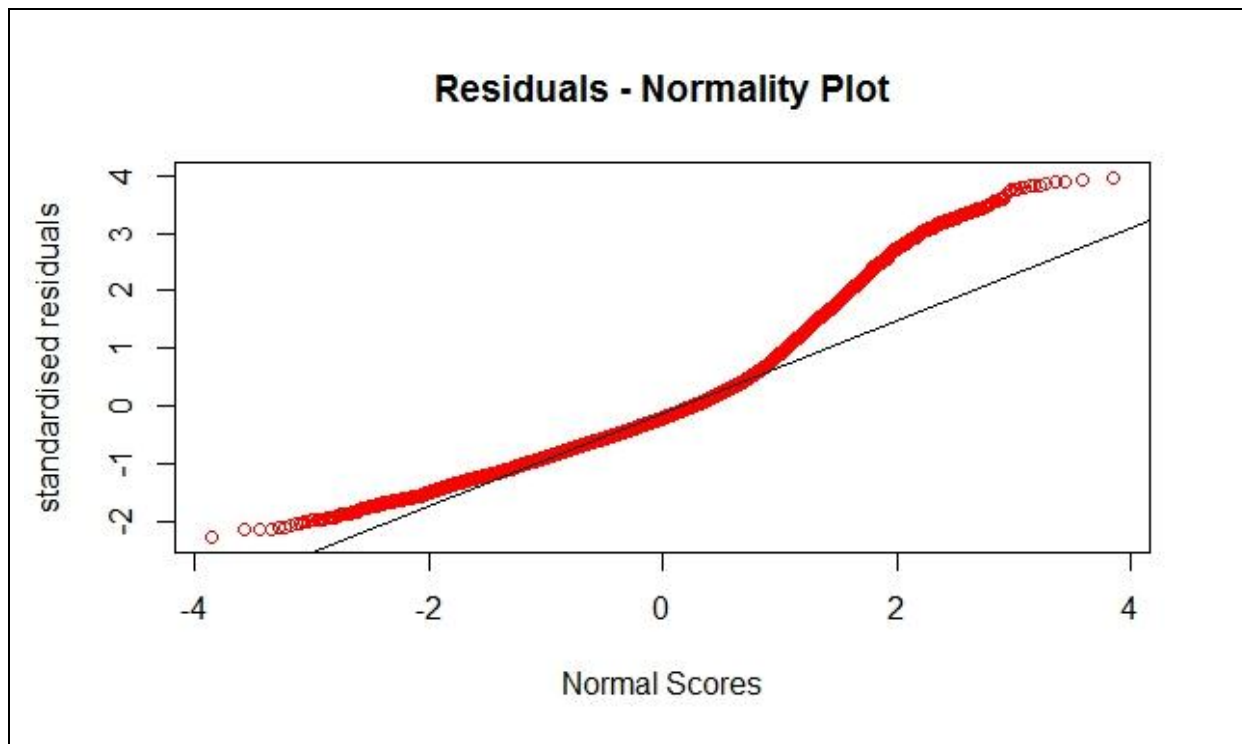


Fig. 4.2 – Residuals: Normality Plot

From figure 4.2, we can clearly see that the residuals do not follow a normal distribution, like before the red line signifies the residuals distribution and the straight line is the path a normal distribution should follow. The closer the red line is to the straight line the more likely the distribution is normally distributed. Our Residuals follow the straight line for a portion of the line; however the ends deviate quite a bit especially in the upper section of the line. At this point, we could opt do perform a transformation to make our residuals follow a normal distribution. However, non-normality of residuals is not a hugely restrictive issue. We would prefer that they would follow a normal distribution but this result should not hinder the model greatly.

4.2 Model Fit

We will now look at how well the model we have produced is fitting the data, by again invoking the summary function, this time it will give us results pertaining to the model we have created. It will help us deduce how well our model is working.

```
>summary(bike_reg_model)
```

```
Call:
lm(formula = split_bike$cnt ~ split_bike$mnth + split_bike$hr +
    split_bike$weathersit + split_bike$temp + split_bike$hum +
    split_bike$windspeed)

Residuals:
    Min       1Q   Median       3Q      Max
-238.62  -69.78  -21.37   44.10  417.47

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    12.8935     6.2808   2.053  0.0401 *
split_bike$mnth  4.9815     0.3530  14.112 <2e-16 ***
split_bike$hr    5.9956     0.1718  34.897 <2e-16 ***
split_bike$weathersit -3.4027     1.9491  -1.746  0.0809 .
split_bike$temp  250.7364     6.0049  41.755 <2e-16 ***
split_bike$hum   -142.9041     6.9765 -20.484 <2e-16 ***
split_bike$windspeed  15.5608     9.7449   1.597  0.1103

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 105.3 on 8638 degrees of freedom
Multiple R-squared:  0.3805, Adjusted R-squared:  0.3801
F-statistic: 884.3 on 6 and 8638 DF, p-value: < 2.2e-16
```

Fig. 4.3 – Summary of Model Output

Let's examine the output from figure 4.3; we can first look at the residuals section of the output. These values will give us the errors in our predictions (difference between actual recorded values and our predicted values). The figures in the residual sections are calculated as the true value less the predicted value. Using this calculation we can see from the maximum error/residual that the model is under predicting the number of bikes that should be available by 417.47, in at least one observation, (but if we go back to figure 4.1 - Residuals, we can see that there were a number of large miss-predictions by this model.

Between the first and third quartiles, which accounts for 50% of where errors fall, we can see that the majority of errors fall between 69.78 greater than the true value and 44.10 less than the true value.

In looking at the co-efficient section of the output, we can see that some of the co-efficients are accompanied by a star, three stars, a dot, or nothing: these indicate the predictive power of each attribute in the model. They each represent different significance levels with three stars showing a significance level of zero, making the argument that the co-efficient is extremely likely to be related to the dependent variable. Therefore the attributes month, hour, temperature and humidity are statistically significant in this model, and they are good features in the prediction of the outcome. Our model has several significant variables and they show signs of being related to the outcome, i.e. the dependent variable, in logical ways.

The Multiple R-Squared value, which is also referred to as the co-efficient of determination, is used to determine how well the values of the dependent variable are explained by the model as a whole. If we return our thoughts to when we were looking at the correlation matrix, we saw that the closer the value was to 1 the stronger the correlation, this same principle can be applied to understanding the co-efficient of determination. This means the closer the value is to 1 the better the model explains the data. With a multiple R-squared value of 0.3805, we know that 38% of the variation in the independent variable is explained by our model. The function of the adjusted R-squared is to penalise model when they use a large number of independent variables, as our model does not contain a large number of independent variables this Adjusted R-squared is quite close to the Multiple R-squared figure.

We can deduce from this information that our model by all accounts is performing quite well. Although the R-Squared value may be perceived as quite low this is not uncommon given real world data, such as our bike data. However, we could argue that the model is slightly weak with regards to the size of some of the errors found in the residuals.

4.3 Compare Predicted Results to 2012 Actual Count

Finally we compare the predicted results from our model to the actual count of the bikes in our data set that were recorded in 2012. The formula we gained from figure 3.6 - Output of Model was used to get the predicted values for each hour of the day in 2012. These comparisons are available on the accompanied disc in the folder data sets in the excel spreadsheet 'hour_edit.xls'. As we saw earlier with the residuals output, there were some large variations between the actual count and the predicted count of bikes. Similarly the model when used to predict the number of bikes that would be needed in 2012 given the hour of the day and a certain number of weather conditions, it predicted well in parts very close to the actual count, but there were some substantial variations between the actual count and the predicted count. Even the process of removing the less predictive features in the model i.e. weather situation and wind s peed, did not help improve the model as a whole.

5. Conclusion

5.1 Overview

We started out with a data set with information from a bike sharing system. In chapter 1, we discussed the aim and motivation behind choosing this topic for a study, and the goal of the project to produce a multiple linear regression model based the information held in the data set. Chapter 2 we used a literature review to gain a brief history behind the inception of regression and some of its earlier uses. During this review we also discussed another study that was conducted using the same data set and how it differed from our own investigation. We then moved into the Implementation, here we looked at each of the variables in the data set individually, to understand how the variables were distributed. Then we looked at how the variables were related to each other or correlated before we eventually built our multiple regression model. Our final step involved the evaluation of the model that was built.

5.2 Further Work

The multiple linear regression model that was developed performed well at predicting some instances when compared against the actual data from 2012. However, there were some predictions that were quite different from the actual counts recorded for that year. Although there were some large discrepancies' when comparing to 2012 counts, even when the predicted results were mapped over the actual data (i.e. the counts from 2011) that were used as the base to form the 'best fit line' equation - some substantial differences were still recorded. An attempt to improve the model could possibly be achieved by removing the less predictive features in the model and adding additional predictive features not currently in this data set. The data we used was only over a two year period, as these schemes have not been in existence for a long period. Collection of more data over several years combined with the accuracy in which the data is currently being collected could help build a more efficient predictive model in the future.

6. Bibliography

Chapman, A. D., 2005. Principles and Methods of Data Cleaning – Primary Species and Species-. 1st ed. s.l.:Global Biodiversity Information Facility.

Cities Today, 2014. Bike-share schemes: what price a healthier city?. [Online]
Available at: <http://cities-today.com/2014/04/bike-share-schemes-price-healthier-city/>
[Accessed 14 May 2014].

earth-policy.org, 2013. US Bike-Sharing Fleet More Than Doubles 2013. [Online]
Available at: www.earth-policy.org/data_highlights/2013/highlights40
[Accessed 10 May 2014].

Fanaee-T, H. & Gama, J., 2013. 'Event labeling combining ensemble detectors and background knowledge. Progress in Artificial Intelligence, pp. 1-15.

Galton, F., 1894. Natural Inheritance. 5th ed. New York: Macmillan and Company.

Gaten, T., 2000. Descriptive Statistics. [Online]
Available at: <http://www.le.ac.uk/bl/gat/virtualfc/Stats/descrip.html>
[Accessed 10 May 2014].

Guass, C., 1821. Theoria combinationis observationum erroribus minimis obnoxiae. s.l.:s.n.

K. Bache and M. Lichman, 2013. UCI Machine Learning Repository. [Online]
Available at: <http://archive.ics.uci.edu/ml>
[Accessed 10 May 2014].

Landmark Digital Ltd, 2013. breakingnews.ie. [Online]
Available at: <http://www.breakingnews.ie/ireland/plans-could-see-ban-on-cars-in-dublin-city-centre-606302.html>
[Accessed 18 Feb 2014].

Osborne, J. & Waters, E., 2002. Four assumptions of multiple regression that researchers should always test. Practical Assessment, Research & Evaluation, VIII(2), pp. 1-9.

Pearson, K., 1896. Mathematical Contributions to the Theory of Evolution. III. Regression, Heredity and Panmixia. Philosophical Transactions of the Royal Society of London, Volume 187, pp. 253-318.

Pearson, K., 1930. The Life, Letters and Labors of Francis Galton. London: Cambridge University Press.

Pedhazur, E. J., 1997. Multiple regression in Behavioural Research. 3rd ed. Orlando: Harcourt Brace.

Southeastern.edu, 2012. ScatterDiagrams & Regression. [Online]

Available at:

<https://www2.southeastern.edu/Academics/Faculty/dgurney/Math241/StatTopics/ScatGen.htm>

[Accessed 01 May 2014].

Transport For London, n.d. Congestion Charge. [Online]

Available at: <http://www.tfl.gov.uk/modes/driving/congestion-charge>

[Accessed 14 May 2014].

7. Appendix

7.1 Project Proposal

Objectives and Contribution to the Knowledge

The objective of this project is to examine the bike sharing system and to determine if the system is a viable means of transportation, as opposed to other forms of more common methods of transportation in relation to changing weather conditions. Through the course of this analysis we will determine the factors influencing people's decision to avail of this mode of transportation on a day by day and hour by hour basis, i.e. by assessing the times of the day individuals use the bikes and if weather conditions play a factor in the decision to use the bike sharing system.

With people around the world becoming more informed in terms of pollution, i.e. individual's carbon footprint along with the amount of energy they consume on a daily basis and with cities leading the race in trying to find more energy efficient modes of transportation, it is important to understand the factors stimulating individuals to choose bike sharing schemes. Large cities around the globe are trying to reduce congestion and some going as far as to ban cars from the city centres.

'New traffic management plans being considered by the National Roads Authority could see a ban on cars driving through Dublin city centre.

According to the Irish Times, a new orbital route has been proposed around the city which would mean the pedestrianisation of Suffolk Street and Church Lane.' (Landmark Digital Ltd, 2013)

It becomes clearer as cities look for a resolution to increasing congestion that analysis of this type of data becomes even more essential as an alternative free moving method of transportation.

Background

Bike sharing systems, sometimes referred to as bike share schemes, is a service whereby individuals can avail of bicycles on a short term basis. Bike sharing systems, as the name suggests eliminates the element of ownership, the individual takes a bike from a certain 'bike share' location and travels for a short period returning the bike at another 'bike share' location/destination.

These schemes have seen huge growth in recent years. 'As of April 2013 there were around 535 bike-sharing programmes around the world, made of an estimated fleet of 517,000 bicycles.' (wikipedia.org, 2014).

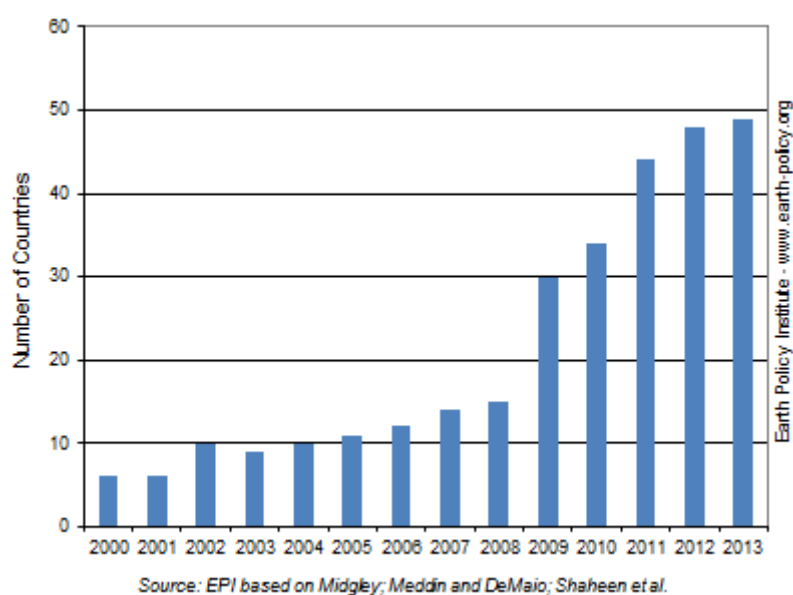


Fig.1 - Countries with bike share programmes, Jan 2000-Apr 2013 (grist.org, 2013)

The rise in popularity of such programmes could be partially attributed to people putting more emphasis on healthy lifestyles also on people becoming more informed on the environment and individuals increased awareness of the growing need for the conservation of energy. Whatever people's personal view point, these systems have seen increased use in cities all over the world. Cities are embracing these systems and are looking at the bicycle as

a means to enhance mobility, reduce pollution in the air, alleviate traffic congestion, and increase healthy living.

Technical Approach

The technical approach will fall into the following cohorts:

- Research
- Literature review
- Requirements Capture
- Implementation

We will research the topic through websites offering information in the area of bike sharing schemes, reviewing the history of these schemes and documenting their rise in popularity in recent years. During the research and literature review we will also look for other data/results available which we can use to compare & contrast to our data findings from our dataset. While reviewing, time will be dedicated to uncovering other methods for analyzing the data that were not deemed necessary at the beginning of the project.

In the requirements capture we will try to gather information from all the relevant stakeholders within the bike sharing system, from the direct customer, to other road users to bicycle distributors and manufacturers. During this process, we will identify all the key stakeholders and try to manage their conflicting interests

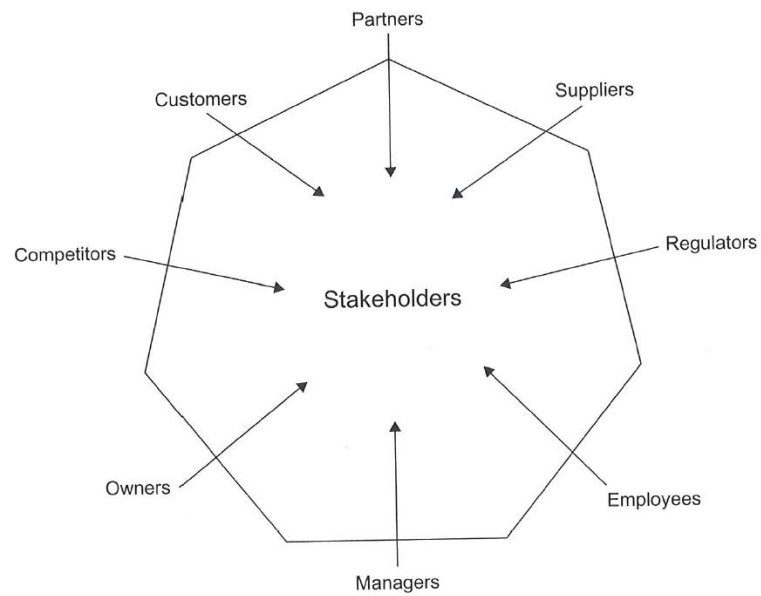


Fig. 2 - Generic Stakeholder categories (Debra Paul, n.d.)

Above Figure 2 displays a generic stakeholder category listing, during our analysis, we will identify who are the actual stakeholders within all the categories listed above, then we will try to ascertain their needs in relation to this project.

Project Plan

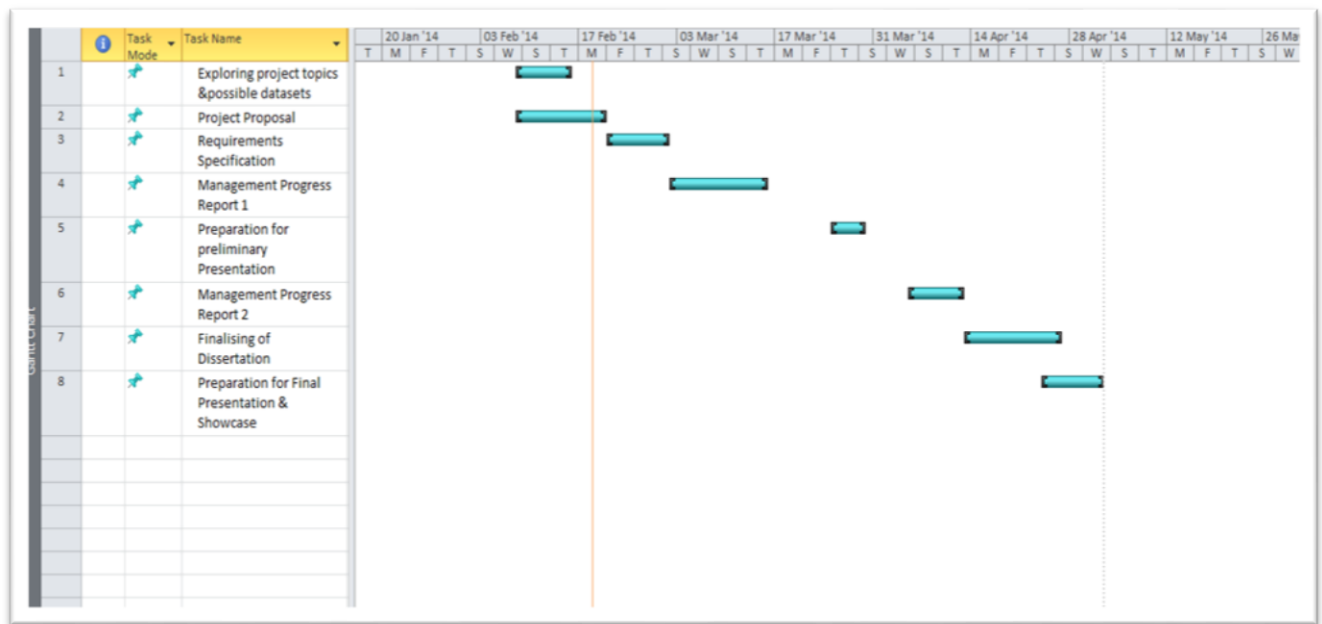


Fig. 3 - Gantt chart for timeline of Project

The gantt chart in figure3 on page 3 gives a timeline for the different deliverables that will be required and their anticipated time for completion.

Technical Details

The implementation language that will be used for the majority of the work in this project will be R-Studio. In addition to R-Studio, Excel will be used, in some capacity, to verify results. For any data mining required, WEKA will be utilized. If required Python and SQL, if the scope of the project changes, may also be incorporated into the overall analysis.

Step 1: Data collection: section 6. Systems/Datasets details the primary dataset that will be utilized, if the need for more data/datasets becomes necessary to find better correlations or build better models, more data will be collected.

Step 2: Data cleaning: It is essential to make sure that the data has no spelling mistakes, and that missing data is handled or any junk/nonsense data is removed. Without this process the results will be incorrect. 'Analyzing this data will result in erroneous conclusions unless the data analysts take steps to validate and clean the data' (Oracle, 2014).

Step3: Data modelling: In this part we will use statistics and R-Studio in conjunction with Excel, to try to establish some meaningful correlations within the data and to help in the prediction of future patterns based on historical data.

Systems/Datasets

The dataset comes from the UCI Machine Learning Repository, center for Machine Learning and intelligent Systems. 'This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bike share system with the corresponding weather and seasonal information' (UCI Machine Learning Repository, 2013)

'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.' (UCI Machine Learning Repository, 2013)



Fig.4 - Bike Sharing in London

Evaluation, Tests and Analysis

At this point we will generate tables, and convert the data into meaningful graphs and other visual displays to give a better visual understanding of the results. Deciding on the statistical methods, e.g. regression, to test the data will be executed at this stage.

Consultation with Specialisation Person(s)

Ioana Ghergulescu, Lecturer, NCI

7.2 Requirements Specification

1 Introduction

Cities around the world are experiencing similar trends, as increasingly more people move to them for employment; congestion is becoming an urgent matter for city councils to resolve. Some cities have tried to action this issue by bringing in a congestion charge and the banning of vehicles at certain times of the day. As cities try to ease congestion and improve the flow of traffic, new methods/systems of transportation are becoming commonplace within our cities. One such system is the bike sharing schemes that are becoming more prevalent in every corner of the globe. There are reportedly '535 schemes worldwide, in 49 countries, including 132 in [Spain](#), 104 in [Italy](#), and 79 in [China](#). The total fleet comprised 517,000 bicycles. This is a sharp increase from 2008, when there were 213 schemes operating in 14 countries using 73,500 bicycles' (Wikipedia, 2014). The popularity and success of these systems is apparent from the explosion of these schemes in different cities in numerous different countries. Our analysis will be limited to one such scheme over a two year period in Portugal.

Purpose

The purpose of this Requirements Specification is to get all the requirements from the different stakeholders, who, to some degree, will have access to the system. Gaining input from the relevant stakeholders will help shape and identify the most important use cases that we will explore in more detail in a later section. Collecting the requirements from just one specific group, e.g. data analysts would give a one-sided view of the system and would only service the requirements of the data analysts. For this reason, the main purpose of the Requirement Specification is to gain the input from several different groups that will help give a well-rounded specification of the systems requirements, which in turn adheres to the needs of all relevant stakeholders.

Project Scope

The scope of this project is to conduct an analysis into a bike sharing scheme from data collected from an active bike sharing scheme run in Portugal, the data is on an hour-by-hour basis and the data is for a two year period. The analysis will involve elements of data mining and statistical analysis. The main scope of this project is to establish the correlation between rental of bikes and the weather with the end goal of predicting the number of bikes that will be required on any given day. The data itself is in the format of a CSV file, it contains 17380 instances and 17 attributes. The scope of the project may change slightly as the project develops incorporating more elements including the possible use of APIs to source data, if it is deemed necessary.

Definitions, Acronyms, and Abbreviations

API – Application Programming Interface: a means by which software components interact with each other

CSV – Comma-Separated Values

USB – Universal Serial Bus

PDF – Portable Document Format

HTML – Hypertext Markup Language

XML – Extensible Markup Language

Web Scraping – This is a technique used to extract information from websites. This process concentrates in the structuring of unstructured HTML, XML so that it can be stored and analysed on a spreadsheet or local database.

Database structure – this is the way in which the data will be stored so it may be accessed in the future for further analysis

IDE - Integrated Development Environment

SQL – Structured Query Language

RDBMS – Relational Database Management System

2 User Requirements Definition

The main users of this system will be:

- Data Analyst
- Statistical Analyst
- Database Administrator
- Manager – decision maker

The information uncovered in this analysis will be mainly used and explored by data analysts; however usage of this system may not be limited entirely to the users listed above.

3 Requirements Specification

A person skilled in the relevant area should be able to navigate the system by following the use case descriptions, which are all outlined below. Prior knowledge of the software being used in the relevant use case is a prerequisite for success in navigating through the system.

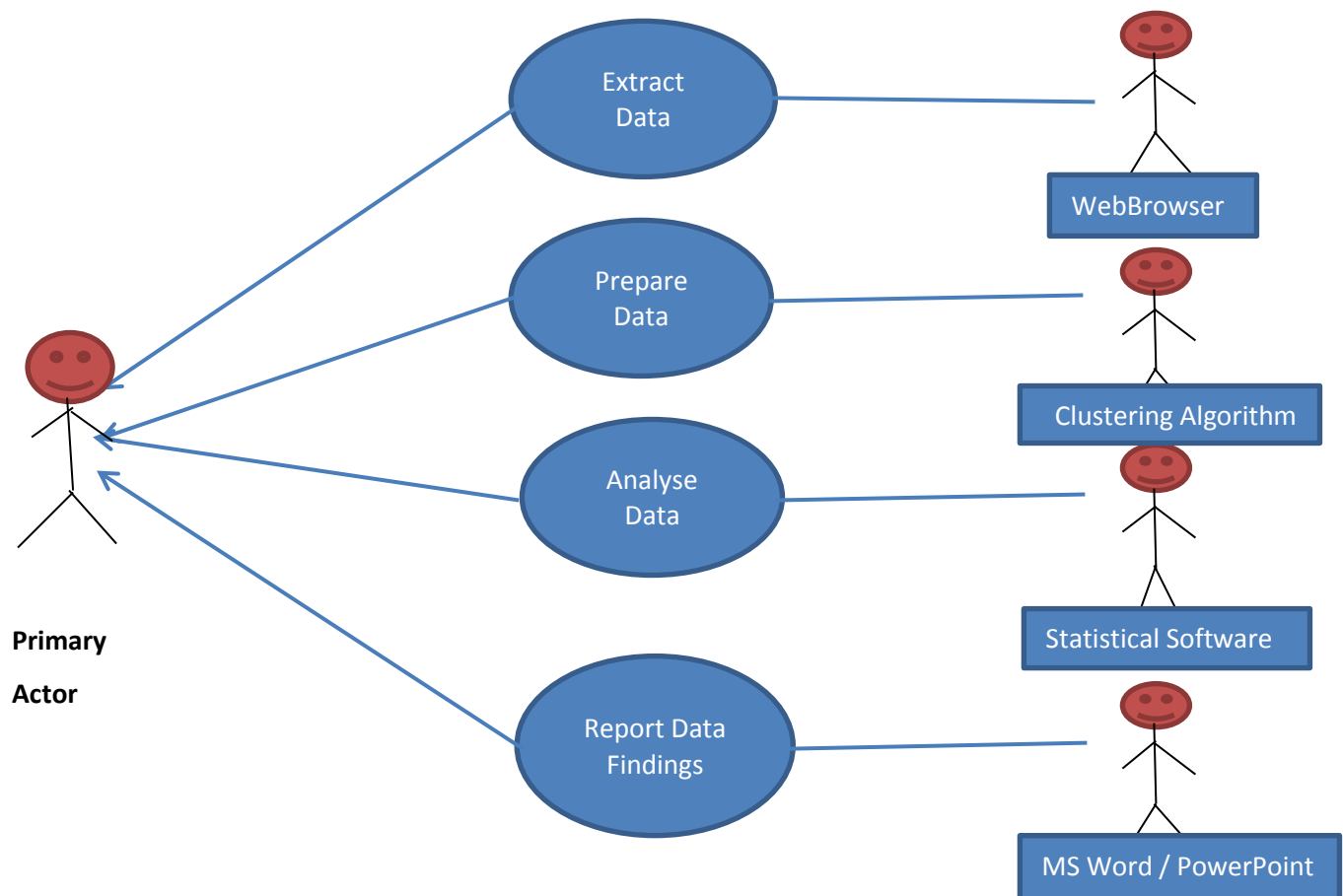
3.1 Functional requirements

Here we will describe the functional requirements of the system:

- Extract Data
- Prepare Data
- Analyse Data
- Report Findings

These are the main functional requirements, they are not ranked, they are of equal importance, each listed requirement is essential for the system to work as a whole.

3.1.1 Use Case Diagram – Bike Sharing System Data Analysis



3.1.2 Requirement 1 – Extract Data

3.1.2.1 Description & Priority

Data Extraction refers to how the data is initially taken from a server/third party, the user will access the internet via a web browser e.g. Internet Explorer/Mozilla/Firefox/Google Chrome etc. and then extract the data to a local hard drive.

3.1.2.2 Use Case

Scope

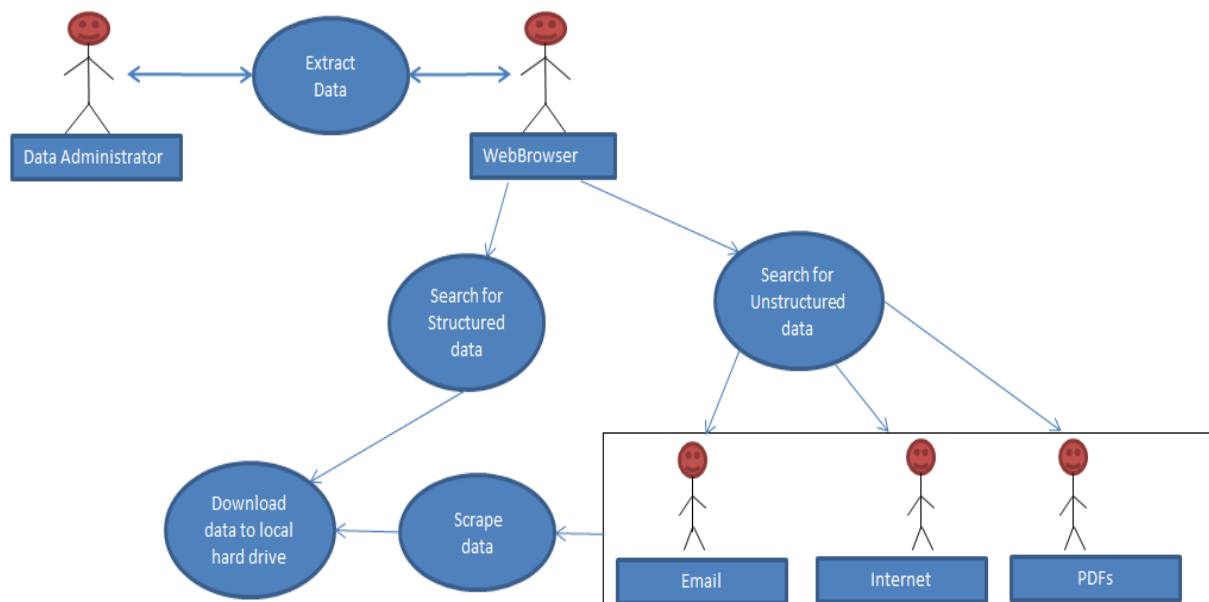
The scope of this use case is to get the data required from the source.

Description

This use case describes the process by which the primary actor, in this instance the Data Analyst, would access the internet via their PC. If the data is available in a structured format e.g. CSV, the data can then be saved directly to the local hard drive. The internet will provide the majority of the data, however third parties (e.g. companies) may have to be approached if the additional data required is not readily available. Depending on the size of the data, this then would be emailed or transferred by USB.

If we are dealing with unstructured data, we would incorporate web scraping to extract the data.

Use Case Diagram



Precondition**Activation**

This use case starts when the actor (Data Analyst) accesses the internet to search for the data.

Main flow

1. The actor accesses the internet via appropriate web browser
2. Search for structured data via search engine
3. Access data: <https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset>
4. Data is downloaded to local hard drive in structured format, CSV file

Alternate flow

1. If additional structured data is required, as the scope of the project increases follow steps from Main flow (ignoring step 3)
2. If unstructured format e.g. email, web pages, PDFs, scanned texts etc., follow from point 2 of the main flow
3. When data is found this information can be extracted by means of web scraping
4. This extracted data is then saved in a reusable format, CSV if possible, then saved to the local hard drive

3.1.3 Requirement 2 – Prepare Data**3.1.3.1 Description & Priority**

Data Preparation is especially required if the data comes from an unstructured source. This requirement is difficult to fully automate into a step by step process, as it involves many different tasks. Data preparation is the most time consuming part of any data mining project.

3.1.3.2 Use Case

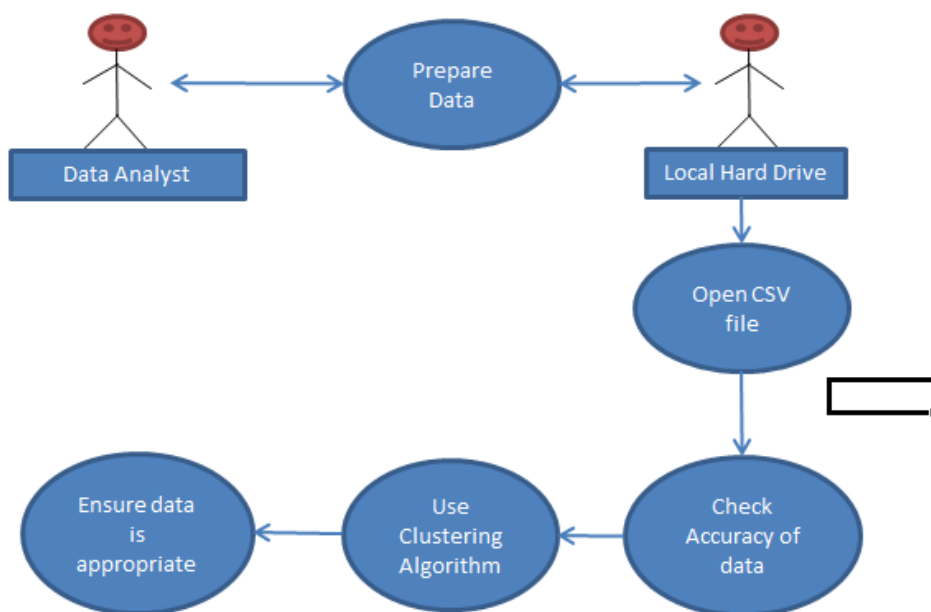
Scope

The scope of this use case is to cleanse the data into quality data so that it may be used for successful analysis through data mining.

Description

This use case describes the stages whereby the data is cleansed; it is the next step after data extraction. This involves checking the accuracy of the data, transforming the data (dealing with missing values) if necessary.

Use Case Diagram



Precondition

The data has been extracted, as in Requirement 1

Activation

This use case starts when the actor (Data Analyst) has finished extracting the data

Main flow

1. Open saved CSV file
2. Check data for accuracy e.g. 'has all relevant information been included?'
3. Make sure data is subjected to standard statistical analysis
4. Use clustering algorithm to evaluate data
5. Ensure the data is appropriate for its column

3.1.4 Requirement 3 – Analyse Data

3.1.4.1 Description & Priority

After ensuring that the data has been cleansed to a high standard and a method for dealing with missing data confirmed. The Statistical analysis should help form meaning to the data.

3.1.4.2 Use Case

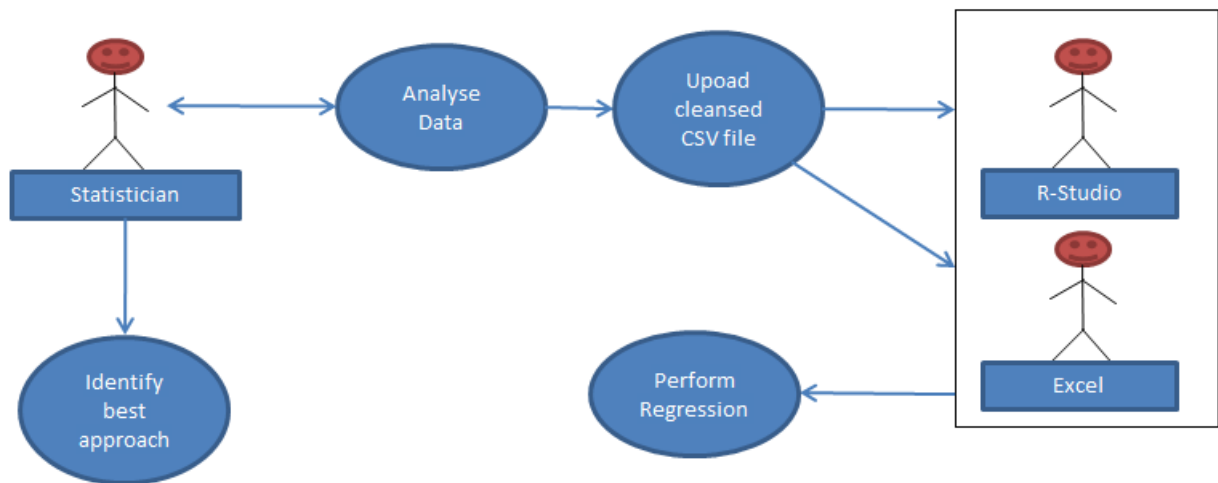
Scope

The scope of this use case is to begin the analysis on the actual 'cleansed' data

Description

This use case describes the steps which will be administered to gain meaning into the data. Which statistical approach to adopt will depend on the nature of the data e.g. Regression Analysis to Predict outcomes

Use Case Diagram



Precondition

Activation

This use case starts when the actor (the statistician) utilises the data for his/her analysis.

Main flow

1. Identify what type of data you are analysing e.g. categorical, numerical
2. Identify the best approach for analysing the data (e.g. Regression analysis, but only if a moderate correlation has been found between at least two variables)
3. Select the software to conduct the analysis (e.g. R-Studio, SAS, SPSS, Excel etc.)
4. Upload cleansed CSV file into statistical software
5. Perform Regression Analysis

3.1.5 Requirement 4 – Report Findings

3.1.5.1 Description & Priority

When the analysis has been completed, we will report the findings in two formats: word and PowerPoint.

3.1.5.2 Use Case

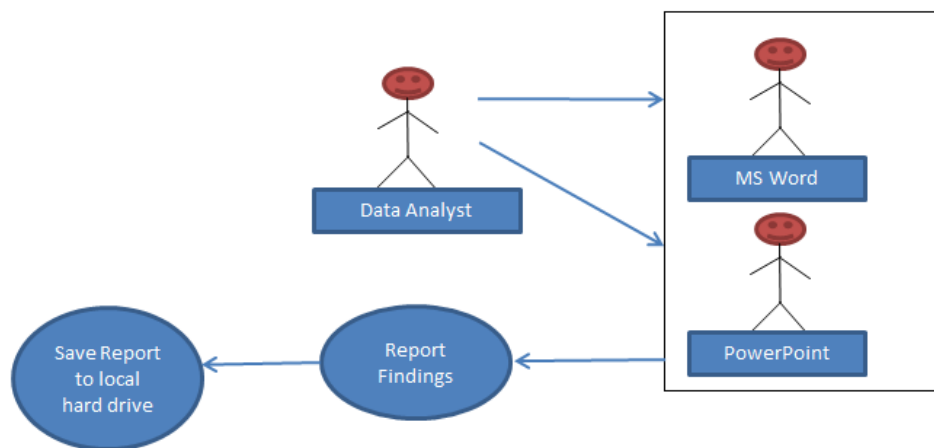
Scope

The scope of this use case is to report the findings via a visual medium

Description

Reporting the findings should come in one of two forms: word and if required to present findings to an audience, PowerPoint can be used to display findings

Use Case Diagram



Flow Description

Word should be opened from PC, along with PowerPoint.

Main flow

1. Report findings by word
2. Report should be saved to local hard drive
3. Report findings by PowerPoint
4. Report should be saved to local hard drive
5. Use graphical display to help reader/audience visualise data findings

3.2 Non-Functional Requirements

3.2.1 Availability requirement

The system will require storage, which were outlined in previous sections. A local hard drive will store the data. And the data will also be backed up on an external hard drive. For analysis the system will require multiple programming languages most of which were mentioned in previous sections, and open source integrated development environment (IDE) for example R. If the scope of the project expands SQL will be used in creating a RDBMS.

3.2.2 Recover requirement

The underlying data should remain unchanged any changes made to the data should be saved to a separate CSV file

3.2.3 Security requirement

Password security will not be required for this project, access will be limited to the four actors outlined in the use cases.

3.2.4 Maintainability requirement

There should not be any issues regarding maintainability, if the system requires more data, this can be saved locally on the hard drive and backed up on the external hard drive.

3.2.5 Portability requirement

The data can be transferred by email, if this becomes an issue regarding size, a USB/flash storage/external hard drive device may be used so others can view the data.

3.2.6 Extendibility requirement

Additional data may be added if it becomes required.

3.2.7 Reusability requirement

The data and any findings will be stored so that further analysis, if required, can be done later with the benefit of the original findings at hand.

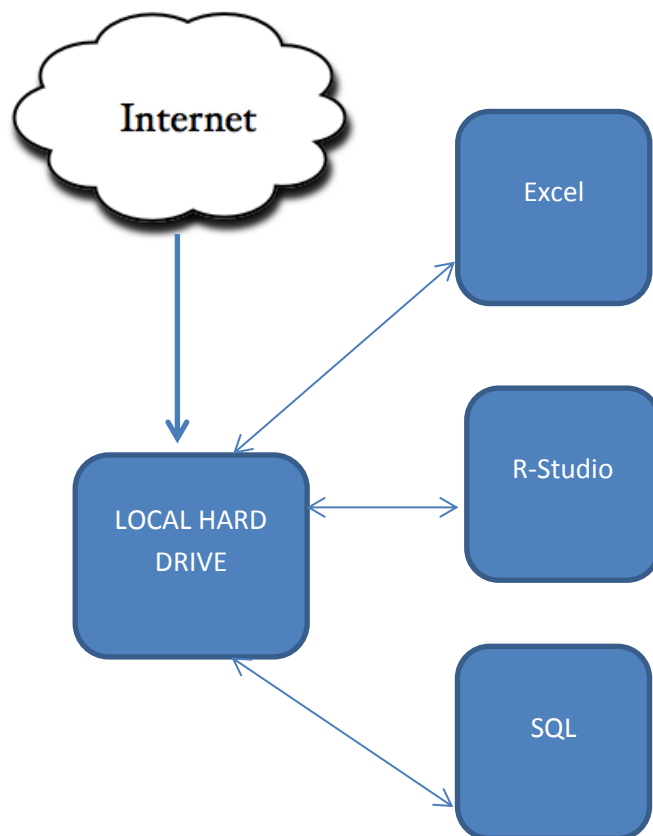
4 Interface requirements

4.1 Application Programming Interfaces (API)

Interfaces which may be used:

- WEKA
- Amazon Web Services
- APIs

5 System Architecture



5 System Evolution

This section describes how the system could evolve over time.

- Users may access the data at anytime
- Access to the dataset should not be limited to one person at any given time
- Users should be able to view the data

- Users should be able to modify the data, without changing the underlying dataset
- New queries can be asked of the dataset
- User are required to save any new queries e.g. via notepad
- The System may require more data for analysis to back up findings of original dataset. This may be implemented by sourcing new structured data or harvesting unstructured data through web scraping.

7.3 Management Progress Report 1

1. Project Description

The purpose of this project is to conduct an analysis into a bike sharing scheme from data collected from an active bike sharing scheme run in Portugal, the data is on an hour-by-hour basis and the data is for a two year period. The analysis will involve elements of data mining and statistical analysis. The main scope of this project is to establish the correlation between rental of bikes and the weather with the end goal of predicting the number of bikes that will be required on any given day.

2. Project Status

The dataset for the project has been sourced, downloaded and saved to the hard drive and backed up on a portable USB. Two academic milestones/deliverables have been passed, i.e. The Project Proposal and The Requirements Specification. We are just over the half way mark of the project. Some activities outlined in the project proposal have gone to schedule, while other activities have proved more time consuming and a rescheduling of activities has become necessary to fulfil the goals of the project. In the proceeding sections we will outline the activities that have been successfully completed to date and the ones that have required more time and some rescheduling of the timetable to complete.

3. Activities Performed During This Reporting Period

The following is a list of activities that have been performed since the beginning of the project

- Exploration of possible datasets to use for project material
- Confirmation of dataset that will be used for the majority of the project work
- Understanding how data is arranged in the dataset
- Direction for project and researching of literature to help achieve project goal

- Project Proposal (Academic Deliverable)
- Extraction of Data from source
- Cleaning & Preparation of Data- partly completed
- Requirements Specification (Academic Deliverable)
- Preliminary analysis of Dataset

4. Activities Planned for Next Reporting Period

The following documents the tasks that have been scheduled to complete or to be started before the next management progress report

- Completion of Cleaning & Preparation of Data
- Exploring other similar projects to help compare and contrast work
- Start full analysis of dataset in R-Studio and excel
- Document all findings from analysis
- Search for results from other datasets that may support finding
- Start dissertation writing
- Think of ideas for structure of final presentation

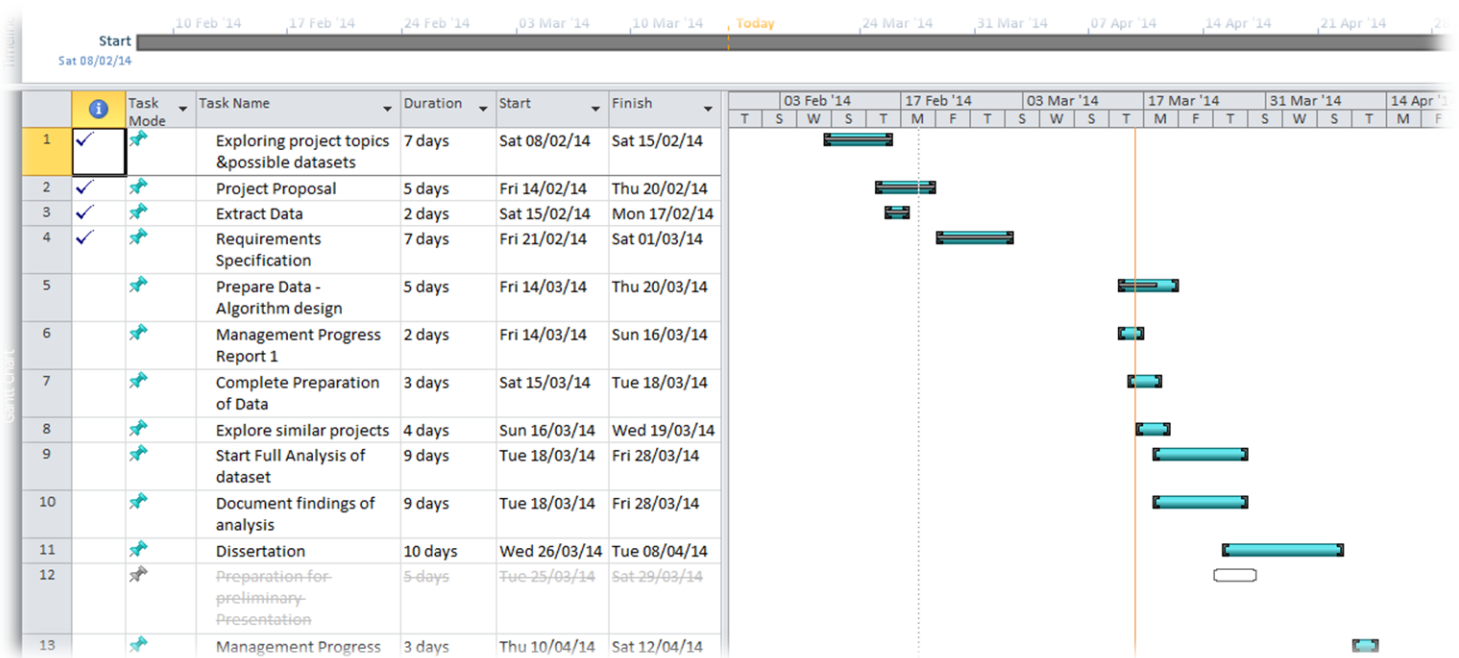
5. Issues/Resolutions

The following shows issues that have surfaced during the project to date and the method used to help resolve the issue:

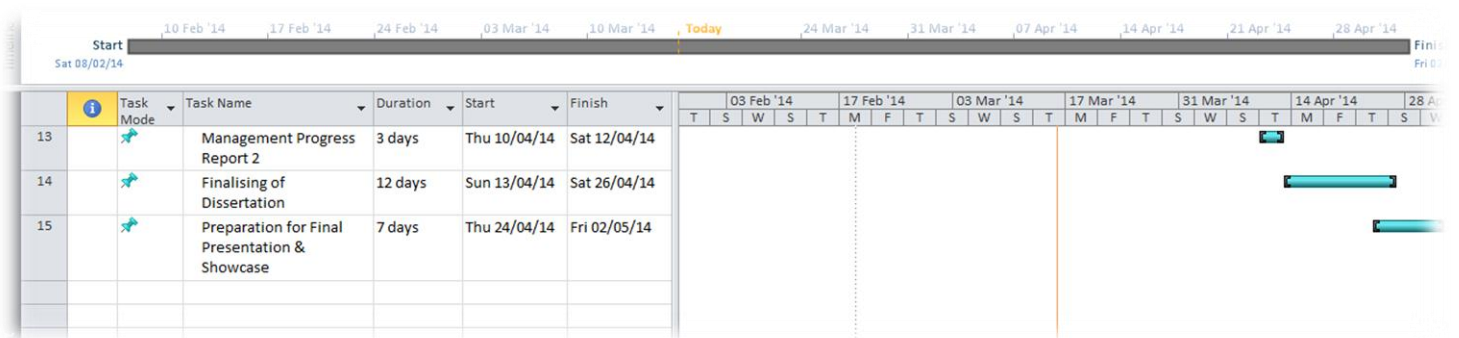
ISSUE	RESOLUTION
Data cleansing-time consuming	More time allocated to process
Algorithm for preparation of data	Further research to get algorithm running correctly

6. Project Change Activity

Below is the updated version of the gantt chart from the project proposal, it displays the tasks which have been fully completed (tasks with tick and a line through the time-line chart) and those which are partially completed. Some tasks have been revised from the original version and some tasks have been added as they became more relevant during the development of the project.



Continued gantt chart



(Note: Preliminary Presentation has a strikethrough as it has been taken off the list of academic deliverables)

7.4 Management Progress Report 2

1. Project Description

The purpose of this project is to conduct an analysis into a bike sharing scheme from data collected from an active bike sharing scheme run in Portugal, the data is on an hour-by-hour basis and the data is for a two year period. The analysis will involve elements of data mining and statistical analysis. The main scope of this project is to establish the correlation between rental of bikes and the weather with the end goal of predicting the number of bikes that will be required on any given day.

2. Project Status

The dataset has been entirely cleaned and prepared for analysis. Partial analysis of the cleaned dataset has completed in excel and SPSS, the dataset has been imported into R and more analysis has been completed comparing results to that found in SPSS. This is the fourth academic milestones/deliverable that has been passed, i.e. The Project Proposal and the Requirements Specification, management report1, and mgmt. report 2. Some activities outlined in the project proposal have gone to schedule, while other activities have proved more time consuming and a rescheduling of activities has become necessary to fulfil the goals of the project. In the proceeding sections we will outline the activities that have been successfully completed to date and the ones that have required more time and some rescheduling of the timetable to complete.

3. Activities Performed During This Reporting Period

The following is a list of activities that have been performed since the beginning of the project

- Cleaning & Preparation of Data- Fully completed
- Preliminary analysis of Dataset using excel

- Imported dataset into SPSS to help with preliminary analysis
- Dataset imported into R-studio, working through mimicking the analysis in SPSS to compare results
- Commenced writing dissertation

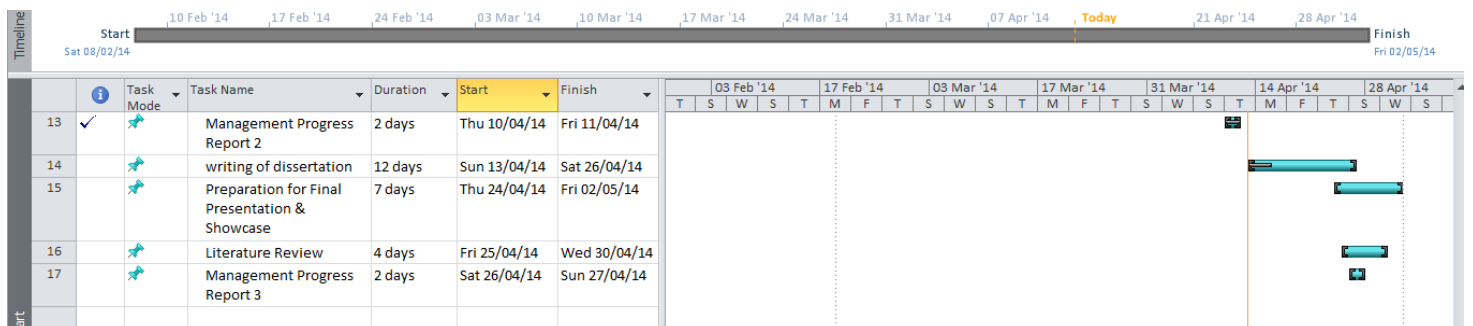
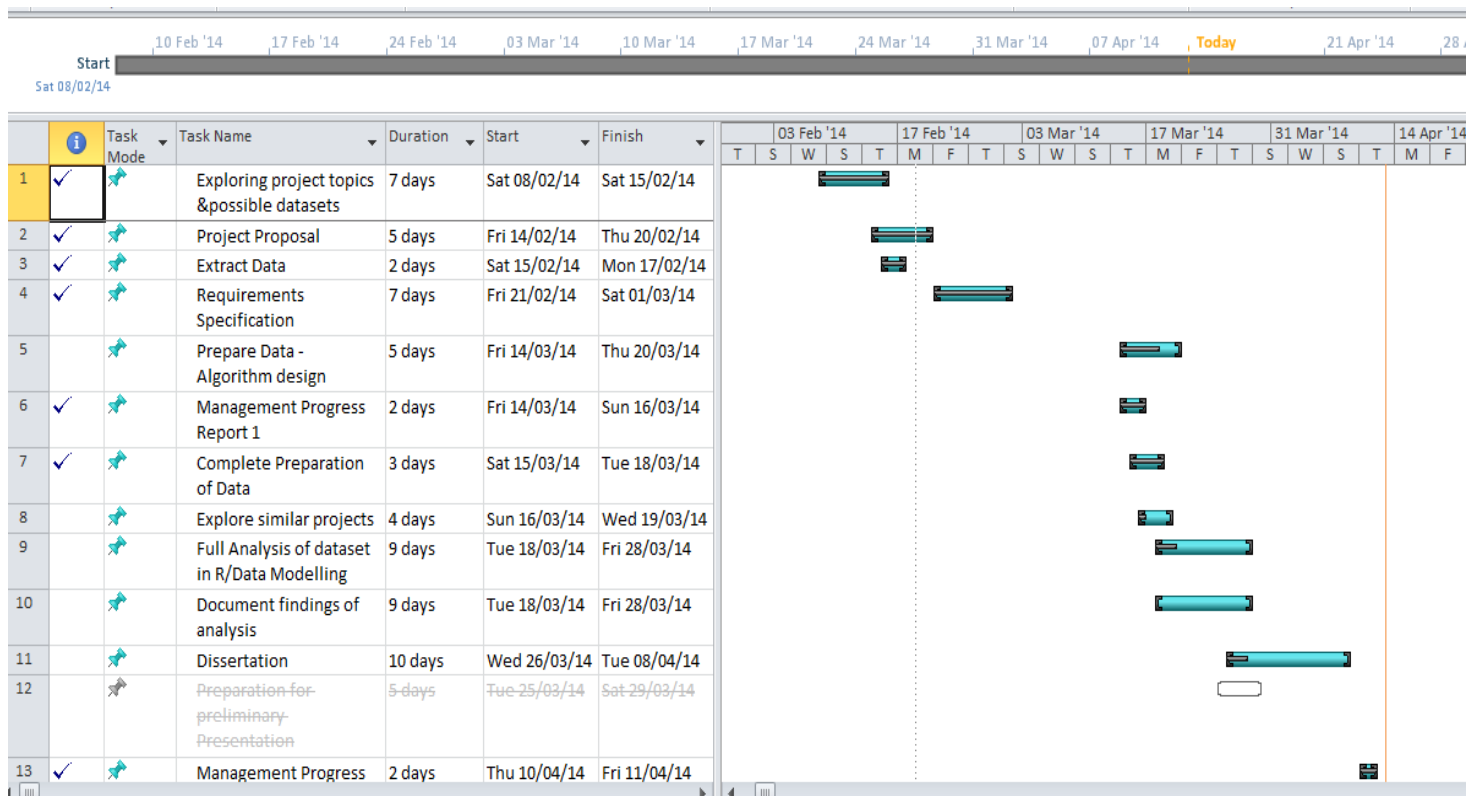
4. Activities Planned for Next Reporting Period

The following documents the tasks that have been scheduled to complete or to be started before the next management progress report

- Completion of analysis in R-Studio
- Exploring other similar projects to help compare and contrast work
- Document all findings from analysis
- Search for results from other datasets that may support finding
- Data Modelling in SPSS and R – Find best fit model
- Start Literature Review
- Continue/Complete dissertation
- Start compiling slides for final presentation

5. Project Change Activity

Below is the updated version of the gantt chart from the Management Report 1, it displays the tasks which have been fully completed (tasks with tick and a line through the time-line chart) and those which are partially completed. Some tasks have been revised from Management Report 1 and some tasks have been added as they became more relevant during the development of the project. The gantt chart runs from the start to the project to the third and final management report.



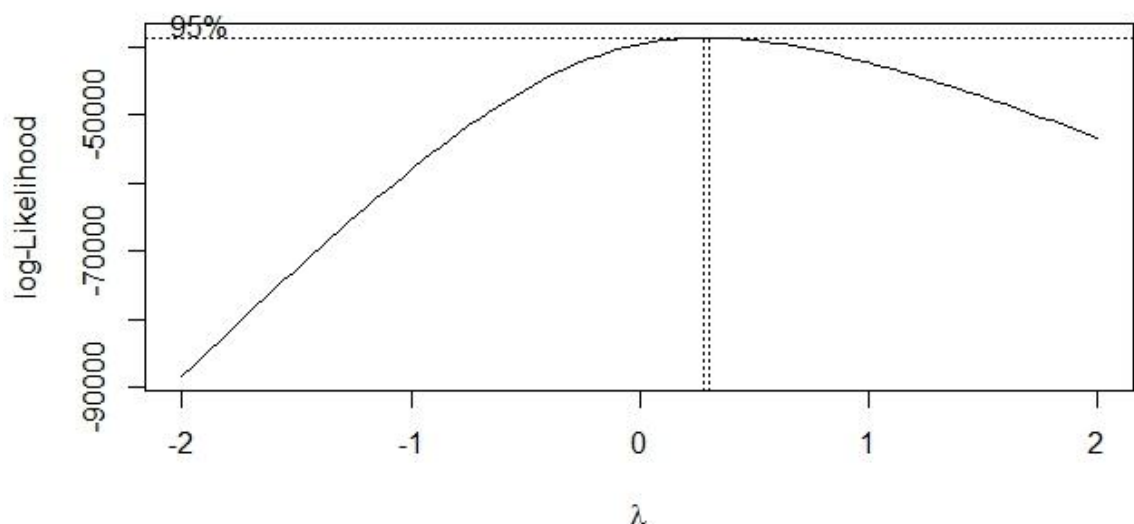
(Note: Preliminary Presentation has a strikethrough as it has been taken off the list of academic deliverables)

7.5 Box-Cox Transformation

- After building the regression model we looked at the residuals and plotted fig 5.2 – Residuals: Normality Plot, from this plot we could see that the distribution of the residuals did not look normal.
- In an attempt to normalise the data, we looked at the Box Cox procedure which ‘chooses an optimal transformation to remediate deviations from the assumptions of the linear regression model’ (Utts, n.d.)
- Using the following code: (the code used was guided by the code used in document (Utts, n.d.))

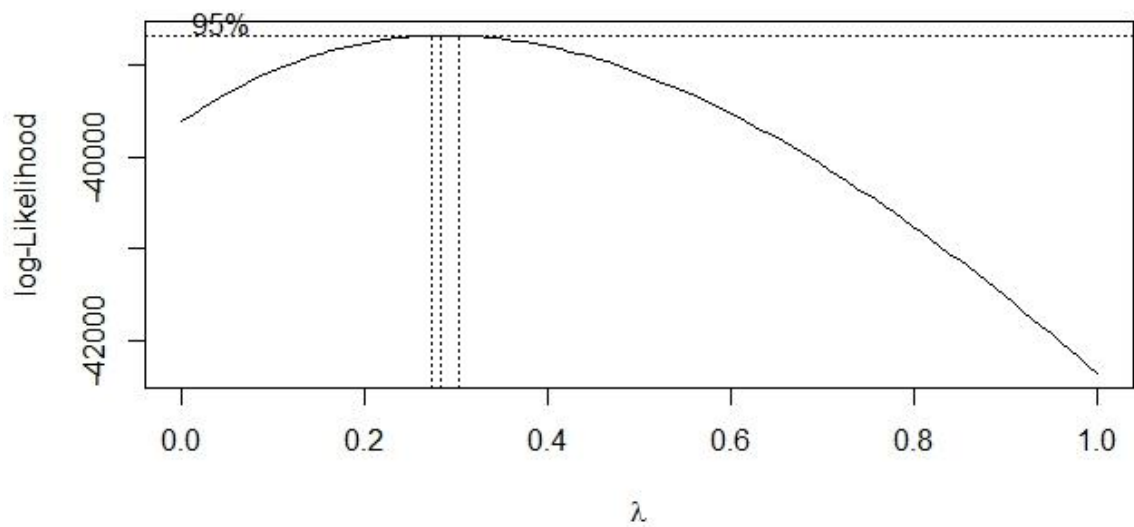
```
>library(MASS)
```

```
>boxcox(bike_reg_model)
```



The above graph is the output we received, it is a plot of the “log likelihood” of the parameter lambda against values of lambda from -2 to 2. The dotted line is indicative of the most ideal value of lambda, around 0.25. We can get a closer approximation for lambda by focusing in on the values between 0 and 1.

```
boxcox(bike_reg_model, lambda = seq(0, 1, 0.10))
```

This graph suggests that the best value of lambda is around 0.3. We use this figure to transform the response variable. We add it to our original data set and then we will establish a new regression model:

```
>split_bike<- cbind(split_bike, split_bike$cnt^0.3)

>names(split_bike)[5] <- "Yprime"

>newModel <-
lm(Yprime~split_bike$mnth+split_bike$hr+split_bike$weathersit+split_bike$temp
+split_bike$hum+split_bike$windspeed, data=split_bike)
```

7.6 Additional Code

7.6.1 Modelling Code

```
boxplots.R x correlations.R x ad-hoc code.R x Normality tests.R x bike_hour x bike x bike modelling.R x
Source on Save Run Source
1 #Model creation
2 split_bike<-subset(bike, bike$yr==0) #subsetting by year
3 dim(split_bike) #confirm number in subset agrees with excel dataset i.e. the number of instances are correct
4
5 bike_reg_model<-lm(split_bike$cnt~split_bike$mnth+split_bike$hr+split_bike$weathersit+split_bike$temp+split_bike$hum+split_bike$windspeed)
6
7 summary(bike_reg_model)
8
9 bike_reg_model.stdRes = rstandard(bike_reg_model)
10 plot(bike_reg_model.stdRes, col="red")
11 abline(0,0)
12 abline(2,0)
13 qqnorm(bike_reg_model.stdRes, ylab="standardised residuals", xlab="Normal Scores", main="Residuals - Normality Plot", col="red")
14 qqline(bike_reg_model.stdRes)
15 summary(p1 <- powerTransform(bike_reg_model))
16 coef(p1, round=TRUE)
17 summary(m1 <- lm(bcPower(split_bike$cnt, p1$roundlam)~split_bike$mnth+split_bike$hr+split_bike$weathersit+split_bike$temp+split_bike$hum+split_bike$windspeed))
18 plot(m1, which=2)
19
20 #box-cox procedure
21 library(MASS)
22 boxcox(bike_reg_model)
23 boxcox(bike_reg_model, lambda = seq(0, 1, 0.10))
24 split_bike<- cbind(split_bike, split_bike$cnt^0.3)
25 names(split_bike)[5] <- "yprime"
26 newModel <- lm(yprime~split_bike$mnth+split_bike$hr+split_bike$weathersit+split_bike$temp+split_bike$hum+split_bike$windspeed, data=split_bike)
27 newModel <- lm(yprime~split_bike$hr+split_bike$weathersit+split_bike$temp+split_bike$hum+split_bike$windspeed, data=split_bike)
28 library(car)
29 newModel.stdRes = rstandard(newModel)
30 plot(newModel.stdRes, col="red")
31 qqnorm(newModel.stdRes, ylab="standardised residuals", xlab="Normal Scores", main="Residuals - Transformed Normality Plot", col="red")
```

7.6.2 Normality tests

```
boxplots.R x correlations.R x ad-hoc code.R x Normality tests.R x bike_hour x bike x bike modelling.R x
Source on Save Run Source
12
13
14 qqnorm(bike$hr, ylab="hour of day")
15 qqline(bike$hr, lty=4)
16
17 hist(bike$weathersit, freq=TRUE, col="green", main="Histogram: Type of weather", xlab="weathersit", ylab="Frequency")
18 qqnorm(bike$weathersit, ylab="type of weather")
19 qqline(bike$weathersit, lty=4)
20
21 hist(bike$temp, freq=TRUE, col="green", main="Histogram: Temperature/MAX(41C)", xlab="Temperature", ylab="Frequency")
22 qqnorm(bike$temp, ylab="temp/MAX(41C)")
23 qqline(bike$temp, lty=4)
24
25 hist(bike$atemp)
26 qqnorm(bike$atemp, ylab="Normalised feeling temperature in C")
27 qqline(bike$atemp, lty=4)
28
29 hist(bike$hum, freq=TRUE, col="green", main="Histogram: Humidity", xlab="Humidity/max(100)", ylab="Frequency")
30 qqnorm(bike$hum, ylab="humidity/MAX(100)")
31 qqline(bike$hum, lty=4)
32
33 hist(bike$windspeed, freq=TRUE, col="green", main="Histogram: windspeed", xlab="windspeed (divided by max 67mph)", ylab="Frequency")
34 qqnorm(bike$windspeed, ylab="windspeed")
35 qqline(bike$windspeed)
36
37 hist(bike$casual, freq=TRUE, col="green", main="Histogram: number of bikes rented by casual users", xlab="casual users", ylab="Frequency")
38 qqnorm(bike$casual, ylab="casual bike users count")
39 qqline(bike$casual, lty=4)
40
41 hist(bike$registered, freq=TRUE, col="green", main="Histogram: number of bikes rented by registered users ", xlab="Registered users", ylab="Frequency")
42 qqnorm(bike$registered, ylab="registered bike users count")
43 qqline(bike$registered, lty=4)
44
45 hist(bike$cnt, freq=TRUE, col="green", main="Histogram: Total number of bikes rented (casual & registered) ", xlab="Total users", ylab="Frequency")
46 qqnorm(bike$cnt, ylab="total bikes rented per hour")
47 qqline(bike$cnt, lty=4)
```

7.6.3 Box Plots Code

```
boxplots.R ✕ | correlations.R ✕ | ad-hoc code.R ✕ | Normality tests.R ✕ | bike_hour ✕ | bike ✕ | bike modelling.R ✕
Source on Save
1 bike<-read.csv("c:/users/peter/Desktop/Data Analytics/Semester 2 - Project/hour.csv", header=TRUE)
2
3 plot(bike$temp)
4 plot(bike$temp, bike$season)
5 plot(bike$season, bike$temp)
6
7 boxplot(bike$season)
8 boxplot(bike$yr)
9 boxplot(bike$mnth)
10 boxplot(bike$hr)
11 boxplot(bike$holiday)
12 boxplot(bike$weekday)
13 boxplot(bike$workingday)
14
15 boxplot(bike$casual, ylab="Number of bikes rented by casual users")
16 boxplot(bike$registered, ylab="Number of bikes rented by registered users")
17
18 boxplot(bike$cnt, ylab="count of total bikes rented")
19 boxplot(bike$registered, ylab="count of total bikes rented by registered users")
20 boxplot(bike$cnt, ylab="count of total bikes rented by casual users")
21 boxplot(bike$windspeed, ylab="windspeed/67(max windspeed registered)")
22
23 boxplot(bike$hum, ylab="humidity (values divided by 100(max))")
24 boxplot(bike$atemp, ylab="normalised feeling temperature in celsius divided by 50(max temp)")
25 boxplot(bike$temp, ylab="normalised temp in Celsius divided by 41(max temp)")
26 save.image("c:/Users/peter/Desktop/Data Analytics/Semester 2 - Project/bike boxplots.RData")
27 view(bike)
28
```