Statistics for Business Analytics I

Final Assignment

Name: Ilias Dimos

AM: f2822102

Training data file bike_02. csv

Test data file bike_test.csv

Professor: Ioannis Ntzoufras

Academic Period: 2021-2022

Table of Contents

1)Introduction **3**
2)Dataset Characteristics **3**
3)Descriptive analysis and exploratory data analysis **4**
3.1) Univariate analysis for Numeric Variables 4
3.2) Univariate analysis for Factor Variables 5
3.3) Bivariate analysis 5
3.4) Pairwise Comparisons 8
4)Predictive models **8**
4.1) Creation of the Predictive Model 8
4.2) Selecting Covariates with Lasso Technique 9
4.3) Using Stepwise procedure in Lasso model to end up to the Final Model 9
4.4) Assumptions of our Final Model 10
4.5) Interpretation of the Final Model 12
4.6) Out-of-Sample Prediction 13
5)Further analysis **13**
5.1) Typical profile of a day in Winter 14
5.2) Typical profile of a day in Fall 14
5.3) Typical profile of a day in Summer 15
5.4) Typical profile of a day in Springer 16
6)Conclusions **17**
Appendix A **18**
Appendix B **19**
Reference and Bibliography **19**

1)Introduction

Bike sharing systems are new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, the user can easily rent a bike from a particular position and return it back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand bicycles.

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 the important events in the city could be detected via monitoring these data.

Aim: Understanding what influences bike rental count hourly and also predict it in order to satisfy demand.

The Data: Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data. We aggregated the data on hourly basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

2)Dataset Characteristics

To make the dataset more understandable, all non-numeric variables converted to factors and also the mismatch between “season” variable and “mnth” variable has been fixed. Also, the “temp”, “atemp”, “hum”, “windspeed” have been multiplied by 41, 50, 100, 67 respectively in order to normalize their values. The variables “X” “instant” and “dteday” have been removed thus they don’t need for our analysis. Na and missing values could not be identified in the dataset after the proper checks in R programming. For the reader's convenience we renamed the labels of the “weathersit” variable and instead of the numbers 1 ,2, 3 ,4 we used the words “Good”, “Medium”, “Bad “, “Really Bad”. Analytically the label “Good “describes the weather phenomena that consist of: “Clear, few clouds, partly cloudy, partly cloudy”, the label “Medium” refers to the “Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist” weather conditions. The label “Bad“ refers to “Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds” weather phenomena and the label “Really Bad “refers to “Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog” weather phenomena.

The dataset consists of random subsamples of 1500 hour occasions and have the following fields:

• instant: record index

• dteday: date

• season: season (1: springer, 2: summer, 3: fall, 4: winter)

• yr.: year (0: 2011, 1:2012)

• month: month (1 to 12)

• holiday: weather day is holiday or not

• weekday: day of the week

• working day: if day is neither weekend nor holiday is 1, otherwise is 0.

• weathersit: Possible outcomes

- 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
- 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog

• temp: Normalized temperature in Celsius. The values are divided to 41 (max)

• atemp: Normalized feeling temperature in Celsius. The values are divided to 50 (max)

• hum: Normalized humidity. The values are divided to 100 (max)

• windspeed: Normalized wind speed. The values are divided to 67 (max)

• casual: count of casual users

• registered: count of registered users

• cnt: count of total rental bikes including both casual and registered (response)

In the dataset sample there is a variable called “hr” which contains the hour of the day that the bike rental happened, if we look at the data characteristics, we see that there is no mention of her. To just erase the variable would be a massive mistake because holds useful information about our analysis so instead of erasing her we transformed her into factor variable with 24 levels to include her in our analysis.

```
'data.frame':   1500 obs. of  15 variables:
 $ season    : Factor w/ 4 levels "Springer","Summer",..: 2 3 3 4 1 2 2 2 2 2 ...
 $ yr        : Factor w/ 2 levels "2011","2012": 2 1 1 1 2 1 1 2 1 1 ...
 $ mnth      : Factor w/ 12 levels "April","August",..: 2 11 12 4 1 2 2 6 6 2 ...
 $ hr        : Factor w/ 24 levels "0","1","2","3",..: 2 4 5 9 15 16 17 22 11 5 ...
 $ holiday   : Factor w/ 2 levels "Regular Working Day",..: 1 2 1 1 1 1 1 1 1 1 ...
 $ weekday   : Factor w/ 7 levels "Sunday","Monday",..: 2 2 6 7 7 6 1 6 5 2 ...
 $ workingday: Factor w/ 2 levels "No","Yes": 2 1 2 1 1 2 1 2 2 2 ...
 $ weathersit: Factor w/ 4 levels "Good","Medium",..: 1 1 1 3 2 1 1 1 1 1 ...
 $ temp      : num  27.1 18.9 21.3 8.2 26.2 ...
 $ atemp     : num  31.1 22.7 25 11.4 31.1 ...
 $ hum       : num  65 88 83 100 27 55 66 52 55 68 ...
 $ windspeed : num  6 0 13 6 30 ...
 $ casual    : int  3 2 2 0 288 50 116 77 30 0 ...
 $ registered: int  11 9 6 10 372 160 176 223 86 3 ...
 $ cnt       : int  14 11 8 10 660 210 292 300 116 3 ...
```

Table 1. Structure of the Bike Rentals Dataset

In table 1 we observe that our dataset consists of 1500 observations and 15 variables. The variables “season”, “yr”, “mnth”, “hr”, “holiday”, “weekday”, “workingday”, “weathersit” are factors with 4, 2, 12 ,24 ,2, 7 ,2 ,4 levels respectively. The variables “temp”, “atemp”, “hum” and “windspeed are numeric variables and the “casual”, “registered”, “cnt” are integers.

3)Descriptive analysis and exploratory data analysis

3.1) Univariate analysis for Numeric Variables

	temp	atemp	hum	windspeed	casual	registered	cnt
vars	1.00	2.00	3.00	4.00	5.00	6.00	7.00
n	1500.00	1500.00	1500.00	1500.00	1500.00	1500.00	1500.00
mean	20.27	23.67	62.58	13.14	34.99	152.49	187.48
sd	7.84	8.56	19.29	7.95	48.31	150.16	179.33
median	19.68	23.48	62.00	13.00	16.00	115.00	143.50
trimmed	20.23	23.70	62.92	12.87	24.90	127.59	159.83
mad	9.73	10.11	22.24	8.89	22.24	128.99	163.83
min	1.64	3.79	0.00	0.00	0.00	0.00	1.00
max	39.36	49.24	100.00	57.00	350.00	871.00	941.00
range	37.72	45.45	100.00	57.00	350.00	871.00	940.00
skew	0.06	-0.02	-0.11	0.51	2.54	1.54	1.28
kurtosis	-0.93	-0.81	-0.82	0.68	8.00	2.57	1.41
se	0.20	0.22	0.50	0.21	1.25	3.88	4.63

Table 2. Description of Numeric Variables

From the Table 2 above we can see that the mean temperature of the “temp” variable is 20.27 Celsius and the median is 19.68 Celsius also the mean feeling temperature of the “atemp” variable is 23.67 Celsius and the median is 23.70 Celsius. These two variables are symmetrically distributed as the median and the mean values are close to each other. Also, the max temperature was 39.36 Celsius but the real feel temperature was 49.24 and the min temperature was 1.64 but the real feel temperature was 3.79. The average total bike rentals in our dataset are 187 as the max bike rentals are 941 and the average casual users are 35 and the average registered users are 152.

3.2) Univariate analysis for Factor Variables

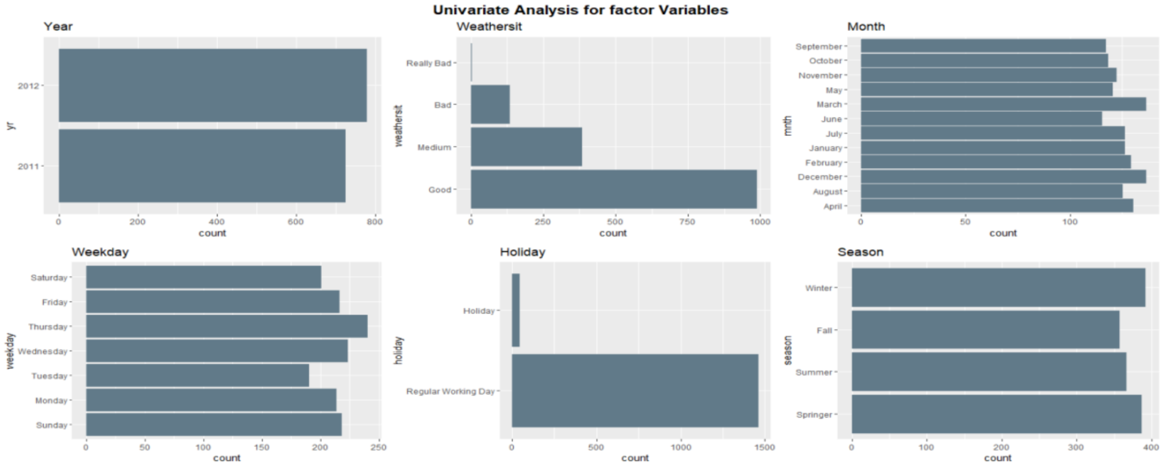

Figure 1. Univariate Analysis Diagrams of the factor variables

By observing the Figure 1 we can see that the year 2012 appears more often in our dataset than the year 2011. So, we expect the total bike rentals in 2012 to be more than in 2011. Also, the majority of “weathersit” in our dataset seem to be good weather phenomena as the label “good“ appears more often than the other labels of the “waethersit” variable. So, we expect most bicycle rentals to be in good weather conditions. Observing the month barplot we can see that the more observed months in the dataset are the March and December. That means that the total bikes rentals will be higher in those months. Also, the “Weekday” diagram in Figure 1 shows that the most observed day of the week is Thursday, so we expect the bike rentals to be higher in this day compared to the others. Last but not least in the Holiday diagram in Figure 1 we see that the most observed day to rent a bike is the Regular Working Day instead of holiday. Finally in the Season barplot we can see that the most observed seasons are the Winter and the Spring but that’s doesn’t mean that the Bike Rentals in those Seasons will be the higher instead of the others.

3.3) Bivariate analysis

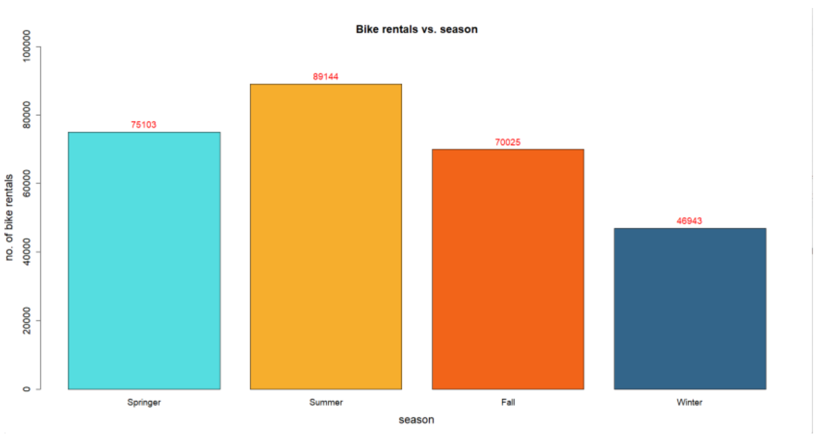

Figure 2. Bikes Rentals Per Season

As you can see in the Figure 2 highest bike rental was recorded in the Summer season and second highest rental was recorded in the Springer season. Total number of bike rentals in the Summer season is 89,144 and the total bike rentals in Spring season is 75,103. We can assume the reason behind this behavior is that the summer and spring seasons provides the most suitable climate for bike riding.

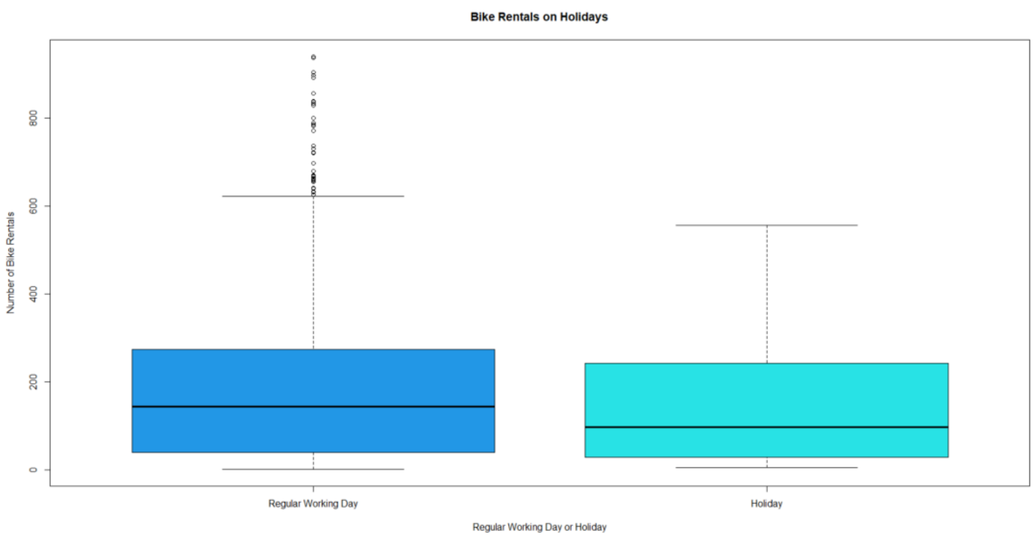

Figure 3. Bike Rentals on Holiday and on Regular Working Day

From the Figure 3 it can be seen that even though there is no huge difference in number of bike rentals per hour on a holiday and a normal working day, the average bike rentals were relatively less on holidays. Also, there were lots of upper end outliers present in working days. Therefore, we can assume that there can be regular bike riders who use the rides to get their workplaces.

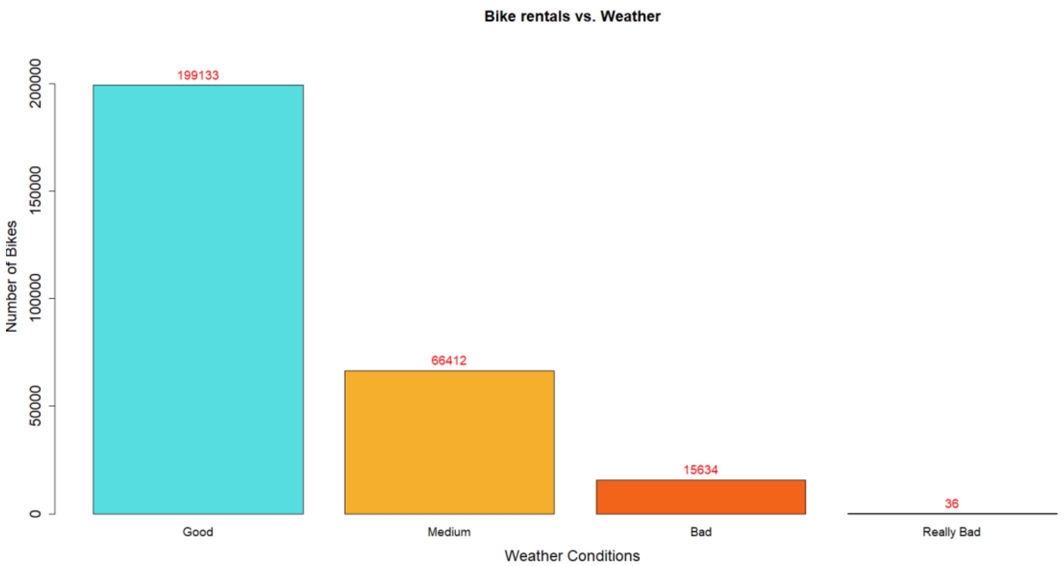

Figure 4. Bike Rentals per Weather Conditions

From the figure 4 it can be clearly seen that highest bike rentals are recorded under clear weather. Compared to clear weather there are very small number of bike rentals happened during mist, light snow, or heavy rain. Since all these bad weather conditions can increase the possibility of road accidents because of low visibility and slippery roads, people rarely choose to ride bikes.

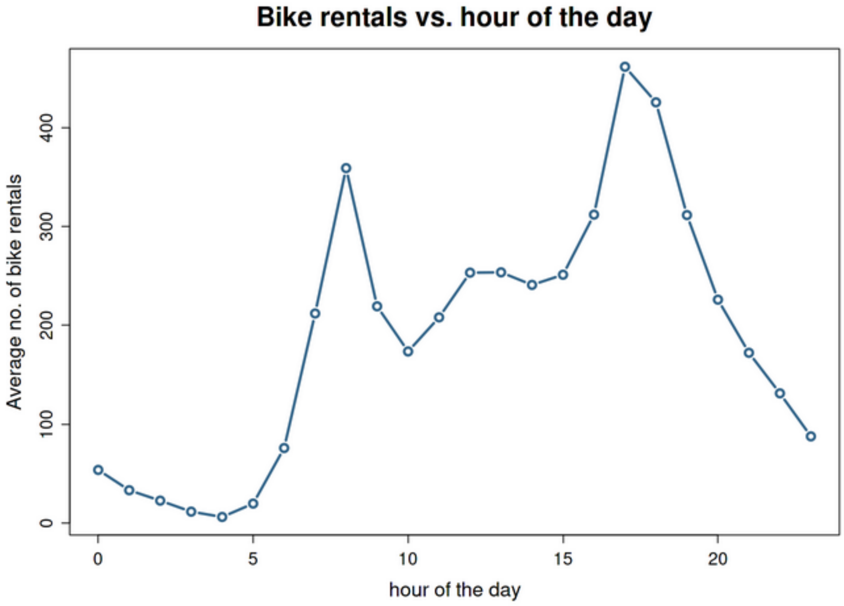

Figure 5. Bike Rentals per hour of the day

As you can see from the Figure 5 there are 2 peaks during 7am to 9am and 4pm to 7pm. These 2 are normal rush times of the day, therefore we can assume this happens because of excess bike rentals of people who are arriving and leaving from workplaces. Apart from this 10 am to 2pm time interval has average bike rentals between 200 to 300 bikes.

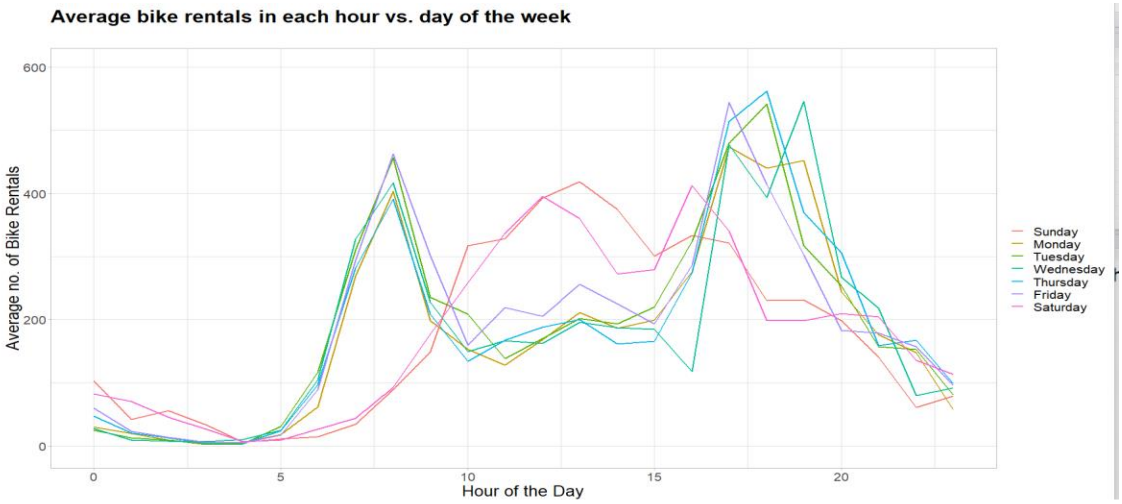

Figure 6. Bike rentals per day of the week

The line graph of the Figure 6 clearly indicates the difference in the patterns of average bike rentals in a weekday and weekend. During the weekdays you can see clear peaks during rush hours from 7am to 9am and 4pm to 7pm. Weekends completely differentiate from this pattern and shows single 12pm to 4 pm. So, we can assume this single peak occurs due to the people who ride bikes as a leisure activity on weekends.

3.4) Pairwise Comparisons

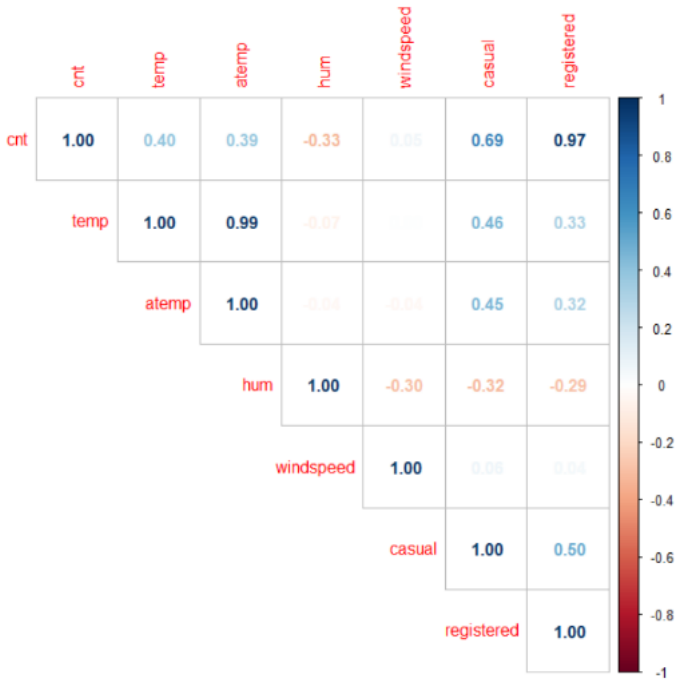

Figure 7. Correlation Plot for numeric variables

From the figure 7 we can see that Temperature and feel temperature as well as registered bike users and total bike rentals show very strong positive relationship with correlation coefficient closer to 1. Total bike rentals and casual users shows moderately strong positive association. Aside from this, rest of the variables do not show any strong inter-dependencies

4)Predictive models

4.1) Creation of the Predictive Model

To be able to identify the best model for predicting the number of bike rentals per hour we have to build the full model (Figure 8) of our analysis. The full model will contain as dependent variable(Y) the variable “cnt” and as independent variables(X) the rest of them except the variables “registered” and “casual”. The reason why we remove these two variables is because they have a strong positive correlation with the depended variable “cnt”. (Refer to Appendix A. Figure 20 for further information about the summary of the full model).

```
Call:
lm(formula = cnt ~ . - registered - casual, data = Bikes)
```

Figure 8. Full regression model

The summary of our full model shows that the model comes with a negative value of the intercept variable. In our case with the Bike Rentals a negative intercept doesn't help us to interpret our model properly. In order to fix the negative intercept problem an efficient solution is to make a new dataset called "Bikes_centered" and centered our covariates at their mean. The Figure 9 shows the new centralized data for our first dataset. (Refer to Appendix B. Figure 23 for further information about the centered Bike Rental model)

```
Call:
lm(formula = cnt ~ . - registered - casual, data = Bikes_centered)
```

Figure 9. Full Regression model with Centered Covariates

4.2) Selecting Covariates with Lasso Technique

It's very important for a model to be fitted with the absolute significant coefficients only. To do that we have to implement the lasso technique first to get rid of some not statistically significant coefficients. Based on the lambda.1se =1.57. We choose the lambda.1se instead of lambda.min because it gives the most regularized model such that the cross – validated error is within one standard error of the minimum. We manage to select 11 coefficients from our full model. Analytically the covariates are: "season", "yr", "mnth", "hr"," holiday", "weekday", "workingday"," weathersit"," temp"," hum", "windspeed". In Figure 10 the form of the Lasso model is represented. (Refer to Appendix A. Figure 21 for further information about lambda.1se from the cross-validation method in Lasso)

```
lm(formula = cnt ~ season + yr + mnth + hr + holiday + weekday +
    workingday + weathersit + temp + hum + windspeed, data = Bikes_centered)
```

Figure 10. Lasso Model based on lambda.1se=1.57

4.3) Using Stepwise procedure in Lasso model to end up to the Final Model

In order to filter our model furthermore to take the absolute significate covariates we implement the "Stepwise" procedure and especially the "both" methods. This method adds and removes covariates based on the AIC criterion until finds the significant ones that match better with our model. The method removed the "season" and the "weekday" covariates.

More specifically we ended up with the following model (Figure 11).

```
lm(formula = cnt ~ yr + mnth + hr + holiday + weathersit + temp +
    hum + windspeed + workingday, data = Bikes_centered)
```

Figure 11. Stepwise Model

It's very important to check if we have multicollinearity among our selected covariates in stepwise model in Figure 10. Multicollinearity is the statically high linear relationship between one explanatory variable with the rest of the explanatories, to identify them we need to use the "Variance Inflation Factors". The criterion that we use to remove the covariates is that if the VIF value for categorical variables with more than 2 factors is greater than 3.16 then we need to remove them because they cause multicollinearity, as far as the other covariates are concerned, we decide if there is need to remove them if their value is greater than 10. In our analysis it was necessary to implement 1 VIF procedure and we managed to remove in the first implementation the "mnth" covariate with *VIF: 7.33 > 3.16*. In Figure 12 we can see our final model after all the proper tests.

```
lm(formula = cnt ~ yr + hr + holiday + workingday + weathersit +
    temp + hum + windspeed, data = Bikes_centered)
```

Figure 12. Final Model

4.4) Assumptions of our Final Model

After finding our final model (Figure 12) with the most significant covariates we need to check if all the assumptions apply to our model. These assumptions are the normality of the model's residuals, the linearity (a linear relationship between the independent variable x, and the dependent variable y, the independence (The residuals must be independent.), and the homoscedasticity (The residuals have constant variance at every level of x).

In order to test the normality assumption of the residuals we did two hypothesis tests for normality (Shapiro-Wilk and Kolmogorov-Smirnov) both tests reject normality of the residuals (*Shapiro-Wilk's $p=1.692e\text{-}15 < 0.05$, KS $p=2.2e\text{-}16<0.05$*) at a significance level 5%. Also, the linearity assumption is violated (Tukey Test $p=2e\text{-}16 < 0.05$) and the homoscedasticity (*ncvTest $p= 2.22e\text{-}16 < 0.05$*). On the contrary independence assumption is not rejected (*durbinWatsonTest $p = 0.354 > 0.05$, Runs Test $p =0.3264 > 0.05$ and Durbin-Watson test $p = 0.8246 > 0.05$*). In total 1 out of 4 assumptions are not rejected.

To cure these problems, we will try to apply logarithm transformation in the depended covariate Y ("cnt"). In Figure 13 we can see the log transformation on the depended variable "cnt".

```
Call:
lm(formula = log(cnt) ~ yr + hr + holiday + workingday + weathersit +
    temp + hum + windspeed, data = Bikes_centered)
```

Figure 13. Log transformed model

For the normality assumption of the residuals, we did two hypothesis tests for normality (Shapiro-Wilk and Kolmogorov-Smirnov) both tests reject normality of the residuals (*Shapiro-Wilk's p= 2.2e-16< 0.05, KS p= 1.87e-11<0.05*) at a significance level 5%. The linearity assumption is not violated (*Tukey Test p= 0.750480 > 0.05*) and the homoscedasticity is violated (*ncvTest p= 2.22e-16< 0.05*). In addition, independence assumption is not rejected (*durbinWatsonTest p = 0.214> 0.05, Runs Test p = 0.5699> 0.05 and Durbin-Watson test p = 0.1076> 0.05*). In total 2 out of 4 assumptions are not rejected. To cure these problems even more, we will try to apply logarithm in the depended covariate Y ("cnt") in combination with polynomials in the numeric variables and weighted least squares. The weighted least squares regression is a method that fixes the assumption of constant variance in the residuals (heteroscedasticity). With the correct weight, this procedure minimizes the sum of weighted squared residuals to produce residuals with a constant variance (homoscedasticity).

In Figure 14 we can see the final multiple regression model with logarithm, weights and polynomial transformations.

```
Call:
lm(formula = log(cnt) ~ +yr + hr + holiday + workingday + weathersit +
    temp + hum + windspeed + I(temp^2) + I(hum^2), data = Bikes_centered,
    weights = wt)
```

Figure 14. Final Model with Transformations

Finally, the linearity assumption and homoscedasticity are not violated (*Tukey Test p=0.99 > 0.05, ncvTest p= 0.15721 > 0.05 respectively*). Also, the independence assumption is not rejected (*durbinWatsonTest p = 0.414 > 0.05, Runs Test p = 0.08 > 0.05*). In total 3 out of 4 assumptions are not rejected. (Refer to Appendix A. Figure 22 for further diagrammatic details about the Assumptions of the final model with transformations in Figure 14)

4.5) Interpretation of the Final Model

```
Coefficients:
                      Estimate  Std. Error t value Pr(>|t|)    
(Intercept)         3.61464721  0.09119882  39.635  < 2e-16 ***
yr2012              0.47026430  0.02905084  16.188  < 2e-16 ***
hr1                -0.55309113  0.12985084  -4.259 2.18e-05 ***
hr2                -1.07841332  0.13316507  -8.098 1.16e-15 ***
hr3                -1.46754839  0.13584193 -10.803  < 2e-16 ***
hr4                -2.18730514  0.15592667 -14.028  < 2e-16 ***
hr5                -0.83749992  0.13626045  -6.146 1.02e-09 ***
hr6                 0.40914628  0.11932358   3.429 0.000623 ***
hr7                 1.36469962  0.10864049  12.562  < 2e-16 ***
hr8                 1.88313954  0.10427080  18.060  < 2e-16 ***
hr9                 1.59276699  0.11132320  14.308  < 2e-16 ***
hr10                1.31396450  0.10619811  12.373  < 2e-16 ***
hr11                1.47094347  0.10311266  14.265  < 2e-16 ***
hr12                1.57013668  0.10929704  14.366  < 2e-16 ***
hr13                1.54641147  0.10190717  15.175  < 2e-16 ***
hr14                1.50215231  0.10507612  14.296  < 2e-16 ***
hr15                1.47162399  0.10875309  13.532  < 2e-16 ***
hr16                1.84547356  0.10643702  17.339  < 2e-16 ***
hr17                2.23397238  0.10051335  22.226  < 2e-16 ***
hr18                2.16942215  0.09910077  21.891  < 2e-16 ***
hr19                1.83655234  0.10072501  18.233  < 2e-16 ***
hr20                1.60547546  0.10526776  15.251  < 2e-16 ***
hr21                1.28750768  0.10676774  12.059  < 2e-16 ***
hr22                0.99017338  0.11405811   8.681  < 2e-16 ***
hr23                0.66392330  0.11774955   5.638 2.06e-08 ***
holidayHoliday     -0.33728251  0.09377127  -3.597 0.000333 ***
workingdayYes       0.08611392  0.03174472   2.713 0.006752 ** 
weathersitMedium   -0.07185375  0.03598591  -1.997 0.046040 *  
weathersitBad      -0.53032933  0.06543324  -8.105 1.10e-15 ***
weathersitReally Bad -0.89827142  0.71998409  -1.248 0.212367    
temp                0.04261565  0.00195424  21.807  < 2e-16 ***
hum                -0.00432091  0.00105225  -4.106 4.24e-05 ***
windspeed          -0.00665524  0.00187679  -3.546 0.000403 ***
I(temp^2)          -0.00207613  0.00023471  -8.846  < 2e-16 ***
I(hum^2)           -0.00014269  0.00003836  -3.719 0.000207 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.311 on 1465 degrees of freedom
Multiple R-squared:  0.7926,    Adjusted R-squared:  0.7878 
F-statistic: 164.7 on 34 and 1465 DF,  p-value: < 2.2e-16
```

Table 3. Summary of the final regression model

After implementing all the proper tests and assumptions we ended up to our final model that consists of logarithm of the depended variable “cnt”, 2nd degree polynomials of the numeric variables “temp” and “hum” and weighted least squares transformations. The final result is summarized in Table 3. The only coefficient that is not statistically significant is “weathersitReallyBad” ($p = 0.21 > 0.05$). The table also tells us that the Residual Standard Error is 1.311 on 1465 degrees of freedom, meaning on average, a prediction will fall outside, exp(1.311) = 3.71% from the actual total bike rentals when this model will make a prediction. Finally, the adj R-squared is 0.7878, meaning that 78,8% of the variance in total daily bike rentals (“cnt”) is explained by the model.

The final regression equation is:

$$\text{Log(cnt)} = 3.61 + 0.47 * \text{yr2012} - 0.55 * \text{hr1} - 1.078 * \text{hr2} - 1.46 * \text{hr3} - 1.46 * \text{hr4} - 0.83 * \text{hr5} - 0.40 * \text{hr6} + 1.36 * \text{hr7} + 1.88 * \text{hr8} + 1.58 * \text{hr9} + 1.31 * \text{hr10} + 1.47 * \text{hr11} + 1.57 * \text{hr12} + 1.54 * \text{hr13} + 1.50 * \text{hr14} + 1.47 * \text{hr15} + 1.84 * \text{hr16} + 2.23 * \text{hr17} + 2.16 * \text{hr18} + 1.83 * \text{hr19} + 1.60 * \text{hr20} + 1.28 * \text{hr21} + 0.99 \text{hr22} + 0.66 * \text{hr23} - 0.337 * \text{holidayHoliday} + 0.086 * \text{workingdayYes} - 0.072 * \text{weathersitMedium} - 0.53 * \text{weathersitBad} + 0.043 * \text{temp} - 0.004 * \text{hum} - 0.007 * \text{windspeed} - 0.002 * \text{temp}^2 - 0.0001 * \text{hum}^2 + \varepsilon,$$

where $\varepsilon \sim N(0, 1.311^2)$.

Figure 15. Final Regression Equation

The final regression equation (Figure 15) describes the behavior of the total bike rentals based on different occasions. The intercept of 3.61 is the log of “cnt” when all the other characteristics are at their mean. Therefore, the exponentiated value is exp(3.61) = 36.9~37 bike rentals when all the other characteristics are at their mean. The exponential coefficient exp(0.47) for the yr2012 is the expected value for the year 2012 over the expected value for the year 2011 . For example, exp(0.47) = 1.59. We can say that the bike rentals will be 59 % increased for the year 2012 compared to 2011 when all the other covariates are constant. For the exponentiated value of the “hr1”, exp(-0.55) = 0.576949, we can say that the bike rentals at 1 am will be 42.3 % decreased compared to the bike rentals at 00:00 when all the other covariates are constant . For the exponentiated value of the “hr18”, exp(2.16) = 8.67, we can say that the bike rentals at 18 pm will be 767% increase compared to 00:00 when all the other covariates are constant. We can say that the big increase to the bike rentals at 18 pm is because many people finish from their work and go home via bike. For the exponentiated value of the “holidayHoliday”, exp(-0.337) = 0.7139, we can say that bike rentals at holiday period are 28.6 % decreased compared to the non-holiday period where the people work normally when all the other covariates are constant. For the exponentiated value of the” weathersitMedium”, exp(-0.072) = 0.93053 , we can say that the bike rentals when the weather has Medium conditions will be 6.94 % decreased compared to the bike rentals when the weather conditions are good when all the other covariates are constant. The coefficient “temp” from the model output tells that a one unit increase in “temp” increases the total bike rentals by 4.39 % when all the other covariates are constant. The coefficient “hum” from the model output tells that a one unit increase in “hum” decreases the total bike rentals by 0.39% when all the other covariates are constant. The coefficient “windspeed” from the model output tells that a one unit increase in “windspeed” decreases the total bike rentals by 0.69% when all the other covariates are constant.

4.6) Out-of-Sample Prediction

In this section we will try to choose the best model for an out of sample prediction. Out-of-sample prediction is the prediction made by the models on data not used during the construction of the models. More specifically we will use a test dataset which contains 500 new observations and will apply here our “lasso model”, “stepwise model”,” full model” and the “null model”. After Appling them into the test dataset we will calculate the “mean absolute error” in order to evaluate the predicting performance of our models and compare them in order to find the best one for out of sample prediction. The Mean absolute error represents the average of the absolute difference between the actual and predicted values in the dataset. The smaller Mean Absolute Error the model has, the better is for prediction. The following table 3 shows the value of the Mean Absolute Error of the models applying on the test dataset.

Model Type	Mean Absolute Error
full_model	183.0609
null_model	145.369
lasso_model	183.1105
final_model	182.1072

Table 4. Mean Absolute Error of each model

By seeing the Table 4 we observe that the “null model” has the smallest Mean Absolute Error in comparison with the other models, so this is the best model for out-of-sample prediction because indicates a better model fit to the test dataset sample.

5)Further analysis

In our further analysis we will describe a typical profile of a day for each season (Autumn, Winter, Spring, Summer) based on the Bikes Rental dataset. To achieve an analysis for each season we have split the dataset into 4 subsets for every season and we will present the average characteristics via diagrams and tables.

5.1) Typical profile of a day in Winter

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
temp	1	391	12.16	4.31	12.30	12.01	4.86	1.64	24.60	22.96	0.31	-0.34	0.22
atemp	2	391	14.84	5.13	14.39	14.71	4.49	3.79	31.06	27.27	0.31	-0.45	0.26
hum	3	391	61.90	20.22	59.00	61.72	23.72	12.00	100.00	88.00	0.13	-0.97	1.02
windspeed	4	391	13.39	8.11	13.00	13.12	8.89	0.00	41.00	41.00	0.43	0.00	0.41
casual	5	391	11.30	16.02	6.00	7.97	8.90	0.00	120.00	120.00	2.92	11.55	0.81
registered	6	391	108.76	111.47	83.00	90.12	94.89	0.00	712.00	712.00	1.92	5.00	5.64
cnt	7	391	120.06	119.28	88.00	100.82	100.82	1.00	731.00	730.00	1.68	3.67	6.03

Table 5. Characteristics of Winter Season

By observing the results of the Table 5 we can say that in a typical day of the Winter season the average temperature is 12.16 degrees Celsius as the max is 24.60 degrees Celsius. Despite the low average of temperature, the Feeling Temperature (atemp) is 14.84 degrees Celsius as the max is 31.06. This may be due to some sunny days the winter may have had. Also, on average in winter there are 120 bike rentals of which 11 are casual users and 109 of them are registered users.

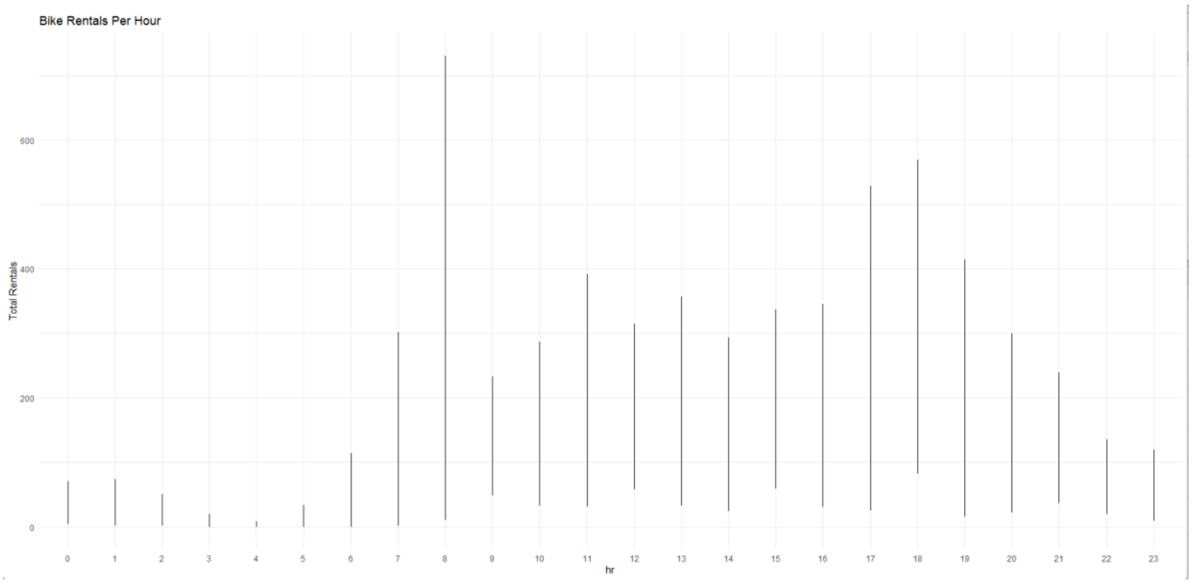

Figure 16. Bike Rentals Per Hour on Winter

By seeing the Figure 16. which shows the Bike rentals per hour on Winter we can say that the highest demand of bikes is at 8 o'clock in the morning when the people are going to carry out their daily obligations and 6 o’clock in the evening when the people are returning home.

5.2) Typical profile of a day in Fall

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
temp	1	357	19.99	5.38	20.50	19.94	6.08	9.02	33.62	24.60	0.04	-0.70	0.28
atemp	2	357	23.57	5.83	24.24	23.57	4.50	10.61	38.64	28.03	-0.02	-0.46	0.31
hum	3	357	67.17	17.53	68.00	67.66	22.24	18.00	100.00	82.00	-0.21	-0.85	0.93
windspeed	4	357	11.96	7.84	11.00	11.63	7.41	0.00	39.00	39.00	0.38	-0.12	0.41
casual	5	357	34.60	49.30	17.00	23.97	22.24	0.00	350.00	350.00	2.78	9.58	2.61
registered	6	357	161.55	157.06	127.00	136.14	133.43	1.00	871.00	870.00	1.57	2.86	8.31
cnt	7	357	196.15	185.72	154.00	167.96	170.50	2.00	938.00	936.00	1.35	1.81	9.83

Table 6. Characteristics of Fall Season

By observing the results of the Table 6 we can say that in a typical day of the Fall season the average temperature is 20 degrees Celsius as the max is 33.62 degrees Celsius. Despite the low average of temperature, the Feeling Temperature (atemp) is 23.5 degrees Celsius as the max is 38.64. This may be due to some sunny days the Fall may have had. Also, on average in Fall there are 196 bike rentals of which 34 are casual users and 162 of them are registered users.

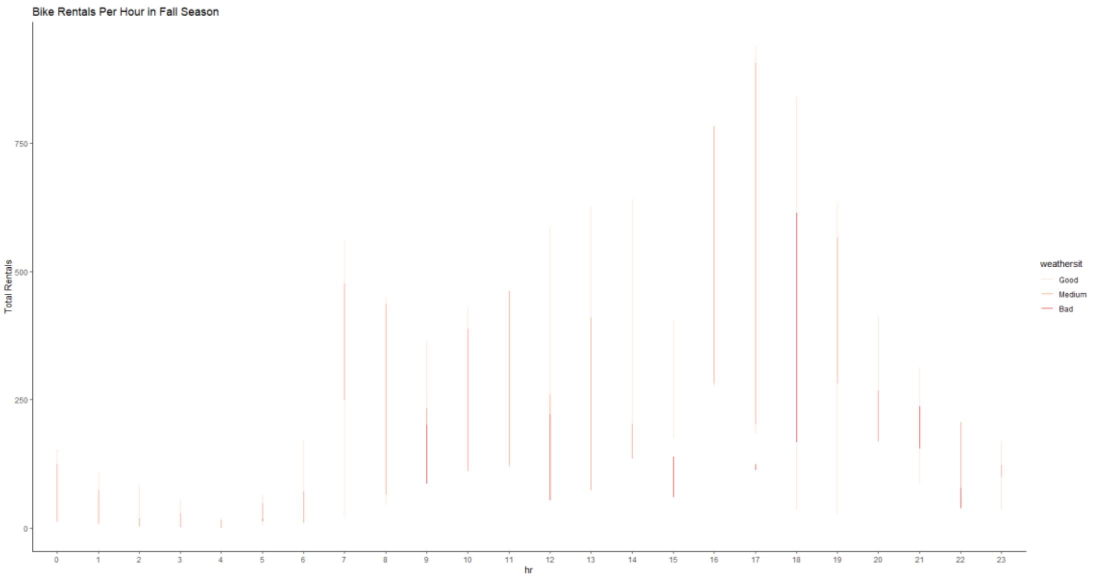

Figure 17. Bike Rentals Per Hour in Fall

By seeing the Figure 17. which shows the Bike rentals per hour on Fall we can say that the highest demand of bikes is at 5 o'clock in the afternoon when the people return home from their work and there are medium weather phenomena.

5.3) Typical profile of a day in Summer

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
temp	1	366	29.63	3.65	29.52	29.67	3.65	18.86	39.36	20.50	-0.10	0.02	0.19
atemp	2	366	33.53	4.33	33.34	33.57	3.37	12.12	49.24	37.12	-0.41	3.01	0.23
hum	3	366	59.99	17.35	59.00	60.29	20.76	23.00	100.00	77.00	-0.06	-0.95	0.91
windspeed	4	366	12.38	7.30	11.00	12.23	5.93	0.00	57.00	57.00	0.91	3.81	0.38
casual	5	366	52.66	52.21	42.00	44.13	45.96	0.00	293.00	293.00	1.73	3.73	2.73
registered	6	366	190.90	170.88	149.50	167.56	159.38	2.00	811.00	809.00	1.18	1.15	8.93
cnt	7	366	243.56	203.80	200.00	221.83	217.20	3.00	941.00	938.00	0.86	0.15	10.65

Table 7. Characteristics of Summer Season

By observing the results of the Table 7 we can say that in a typical day of the Summer season the average temperature is 29.63 degrees Celsius as the max is 39.36 degrees Celsius. The average Feeling Temperature (atemp) is 33.53 degrees Celsius as the max is 49.24. Despite the high temperatures, on average in Summer there are 243 bike rentals of which 52 are casual users and 190 of them are registered users. This is on average an 24 % increase of total bike rentals compared to Fall season and a 50 % increase of total bike rentals compared to Winter Season.

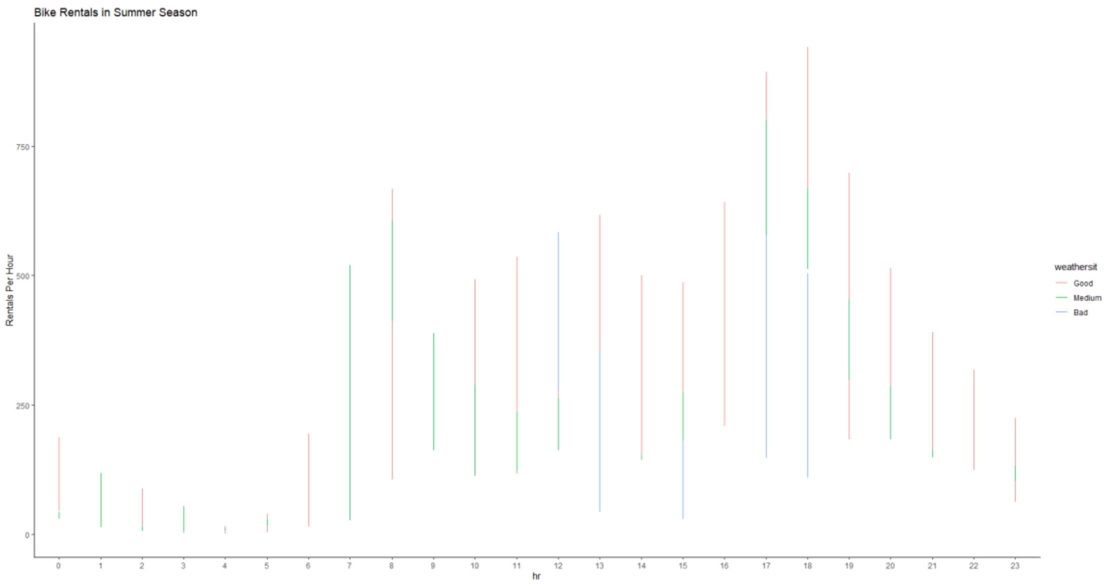

Figure 18. Bike Rentals Per Hour in Summer

By seeing the Figure 18 which shows the Bike rentals per hour on Summer we can say that the highest demand of bikes is at 6 o'clock in the afternoon. Despite that we are at Summer Season we observe that in 6 o’clock most of the bike rentals come with bad weather phenomena.

5.4) Typical profile of a day in Springer

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
temp	1	386	19.86	5.55	19.68	19.85	6.08	6.56	35.26	28.70	0.04	-0.45	0.28
atemp	2	386	23.36	6.12	23.48	23.52	5.62	8.33	40.15	31.82	-0.16	-0.31	0.31
hum	3	386	61.48	20.92	62.00	62.24	25.95	0.00	100.00	100.00	-0.25	-0.80	1.06
windspeed	4	386	14.70	8.21	13.00	14.42	8.89	0.00	43.00	43.00	0.37	0.08	0.42
casual	5	386	42.58	55.39	20.00	31.10	28.17	0.00	311.00	311.00	2.26	5.70	2.82
registered	6	386	151.99	145.91	119.00	129.08	134.92	1.00	700.00	699.00	1.36	1.65	7.43
cnt	7	386	194.57	178.33	159.00	169.15	178.65	1.00	785.00	784.00	1.05	0.50	9.08

Table 8. Characteristics of Spring Season

By observing the results of the Table 8 we can say that in a typical day of the Spring season the average temperature is 19.86 degrees Celsius as the max is 28.70 degrees Celsius. The average Feeling Temperature (atemp) is 23.36 degrees Celsius as the max is 31.82. Despite the high temperatures, on average in Spring there are 194 bike rentals of which 42 are casual users and 151 of them are registered users. This is on average an 1.03 % decrease of total bike rentals compared to Fall season and a 38% increase of total bike rentals compared to Winter Season.

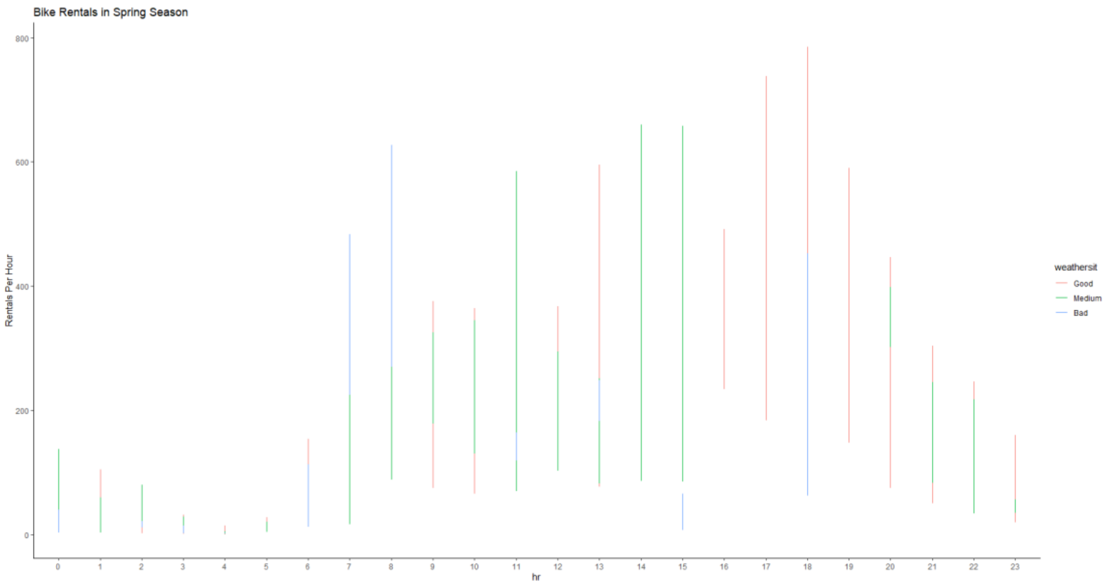

Figure 19. Bike Rentals Per Hour in Spring

By seeing the Figure 19. which shows the Bike rentals per hour on Spring we can say that the highest demand of bikes is at 6 o'clock in the afternoon. Despite that we are at Summer Season we observe that in 6 o'clock most of the bike rentals come with good and bad weather phenomena.

6)Conclusions

The final model of our analysis (Figure 14) seems to have a very good performance as far as the prediction is concerned. The adj R^2 of 78% means that the 78% of the variance of the depended variable "cnt" can be successfully predicted by the model. Therefore we can say that our final model meets the goals of the analysis which was to predict the demand of the bike rentals hourly. Of course, many other good models may exist which can satisfy all the 4 assumptions. In our case the model satisfies 3 out of 4 assumptions (normality of the residuals is violated). That's because we made our analysis based on a dataset which was subseted from a bigger one. In the "interpretation of the final model "section we interpret the performance of some of the covariates of the model. Of course, the same explanation applies to the rest of them. Furthermore, as far as the "hr" variable is concerned it would be a massive mistake to remove her from our analysis even though isn't mentioned in the data characteristics because it provides useful information about the demand of the Bike rentals. Summarizing the report, we can see that no matter the weather conditions and the season the most people prefer to rent bikes especially in the rush hours of the day. The best season to rent a bike is the summer and the not so good season is the winter based on the figure 2. In figure 3 we can see outliers on the "Regular Working Day "boxplot. We can assume that many people prefer to go to their workplace by bike rather than on foot.

Appendix A

```
Call:
lm(formula = cnt ~ . - registered - casual, data = Bikes)

Residuals:
    Min      1Q  Median      3Q     Max 
-320.19  -61.86   -5.00   50.88  426.74 

Coefficients: (4 not defined because of singularities)
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)           -5.6028    25.5934  -0.219 0.826747    
seasonSummer         -24.9808    15.3759  -1.625 0.104450    
seasonFall            -6.6028    14.2687  -0.463 0.643614    
seasonWinter         -67.0795    15.2413  -4.401 1.16e-05 ***
yr2012                82.3898     5.4882  15.012  < 2e-16 ***
mnthAugust           -14.3569    13.5679  -1.058 0.290160    
mnthDecember          42.5255    13.4878   3.153 0.001650 ** 
mnthFebruary          12.5475    13.3513   0.940 0.347479    
mnthJanuary                NA         NA      NA       NA    
mnthJuly             -25.4722    13.5613  -1.878 0.060541 .  
mnthJune                   NA         NA      NA       NA    
mnthMarch            -26.0584    12.9057  -2.019 0.043657 *  
mnthMay              -15.7615    14.0253  -1.124 0.261290    
mnthNovember         -15.2745    15.6428  -0.976 0.329000    
mnthOctober           15.9398    14.0863   1.132 0.257995    
mnthSeptember              NA         NA      NA       NA    
hr1                   -6.9717    18.0791  -0.386 0.699833    
hr2                  -29.1968    17.8056  -1.640 0.101274    
hr3                  -30.7083    17.5626  -1.749 0.080588 .  
hr4                  -30.6056    18.6858  -1.638 0.101659    
hr5                  -10.9890    18.4813  -0.595 0.552204    
hr6                   44.7876    18.0696   2.479 0.013302 *  
hr7                  187.4253    18.2391  10.276  < 2e-16 ***
hr8                  296.0917    18.0770  16.379  < 2e-16 ***
hr9                  160.4963    19.4585   8.248 3.57e-16 ***
hr9                  160.4963    19.4585   8.248 3.57e-16 ***
hr10                 116.0950    17.8127   6.518 9.84e-11 ***
hr11                 151.4004    17.4149   8.694  < 2e-16 ***
hr12                 161.8492    18.9649   8.534  < 2e-16 ***
hr13                 164.5088    17.4351   9.435  < 2e-16 ***
hr14                 158.7299    18.1295   8.755  < 2e-16 ***
hr15                 145.7669    18.8187   7.746 1.77e-14 ***
hr16                 233.3267    19.0534  12.246  < 2e-16 ***
hr17                 374.7796    18.2184  20.571  < 2e-16 ***
hr18                 351.4284    17.5642  20.008  < 2e-16 ***
hr19                 266.3934    17.6862  15.062  < 2e-16 ***
hr20                 178.8539    18.2586   9.796  < 2e-16 ***
hr21                 113.8033    17.9216   6.350 2.87e-10 ***
hr22                  77.1929    18.5701   4.157 3.42e-05 ***
hr23                  39.0287    18.5450   2.105 0.035503 *  
holidayHoliday       -63.4355    17.3075  -3.665 0.000256 ***
weekdayMonday         24.7634    10.2835   2.408 0.016161 *  
weekdayTuesday        19.3706    10.3140   1.878 0.060570 .  
weekdayWednesday      19.9628     9.8843   2.020 0.043605 *  
weekdayThursday       10.3639     9.6886   1.070 0.284931    
weekdayFriday         17.4730     9.9820   1.750 0.080251 .  
weekdaySaturday       14.8753    10.1456   1.466 0.142815    
workingdayYes              NA         NA      NA       NA    
weathersitMedium     -11.7302     6.7061  -1.749 0.080473 .  
weathersitBad        -57.2609    10.9407  -5.234 1.91e-07 ***
weathersitReally Bad -77.4406   104.4902  -0.741 0.458736    
temp                   8.3284     2.4668   3.376 0.000754 ***
atemp                 -2.0773     2.0892  -0.994 0.320231    
hum                   -0.9502     0.1937  -4.905 1.04e-06 ***
windspeed             -1.5207     0.3766  -4.038 5.67e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 102.5 on 1450 degrees of freedom
Multiple R-squared:  0.6842,	Adjusted R-squared:  0.6735 
F-statistic: 64.11 on 49 and 1450 DF,  p-value: < 2.2e-16
```

Figure 20. Summary of the Full Model

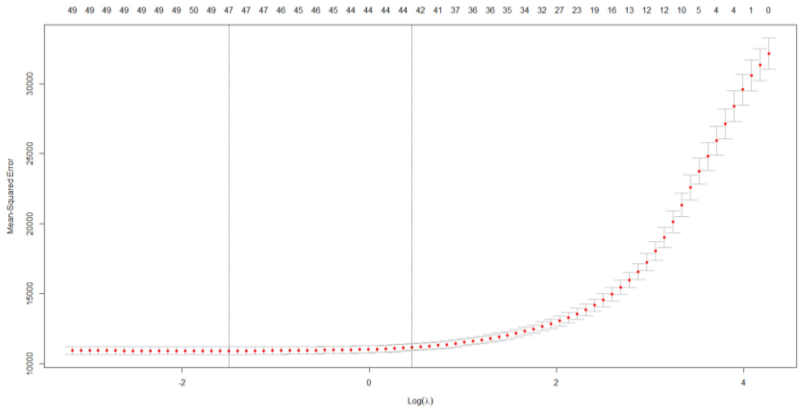

Figure 21. Lasso model for lambda.1se = 1.57

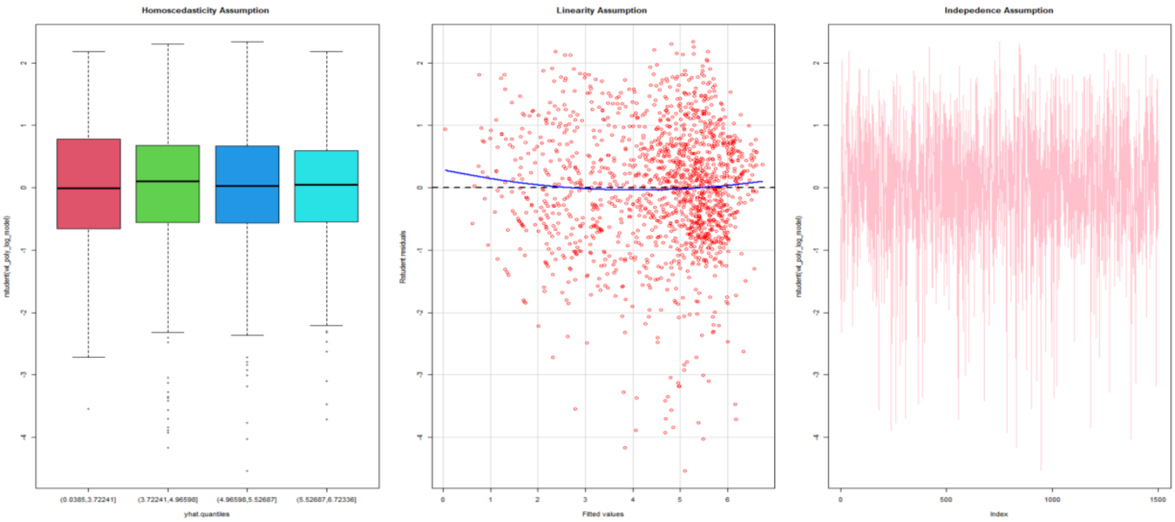

Figure 22. Valid assumptions of the final model with transformations (Figure 14)

Appendix B

```
Coefficients: (4 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          34.5753    17.1070   2.021 0.043451 *
temp                  8.3284     2.4668   3.376 0.000754 ***
atemp                -2.0773     2.0892  -0.994 0.320231
hum                  -0.9502     0.1937  -4.905 1.04e-06 ***
windspeed            -1.5207     0.3766  -4.038 5.67e-05 ***
seasonSummer        -24.9808    15.3759  -1.625 0.104450
seasonFall           -6.6028    14.2687  -0.463 0.643614
seasonWinter        -67.0795    15.2413  -4.401 1.16e-05 ***
yr2012               82.3898     5.4882  15.012  < 2e-16 ***
mnthAugust          -14.3569    13.5679  -1.058 0.290160
mnthDecember         42.5255    13.4878   3.153 0.001650 **
mnthFebruary         12.5475    13.3513   0.940 0.347479
mnthJanuary               NA         NA      NA       NA
mnthJuly            -25.4722    13.5613  -1.878 0.060541 .
mnthJune                  NA         NA      NA       NA
mnthMarch           -26.0584    12.9057  -2.019 0.043657 *
mnthMay             -15.7615    14.0253  -1.124 0.261290
mnthNovember        -15.2745    15.6428  -0.976 0.329000
mnthOctober          15.9398    14.0863   1.132 0.257995
mnthSeptember             NA         NA      NA       NA
hr1                  -6.9717    18.0791  -0.386 0.699833
hr2                 -29.1968    17.8056  -1.640 0.101274
hr3                 -30.7083    17.5626  -1.749 0.080588 .
hr4                 -30.6056    18.6858  -1.638 0.101659
hr5                 -10.9890    18.4813  -0.595 0.552204
hr6                  44.7876    18.0696   2.479 0.013302 *
hr7                 187.4253    18.2391  10.276  < 2e-16 ***
hr8                 296.0917    18.0770  16.379  < 2e-16 ***
hr9                 160.4963    19.4585   8.248 3.57e-16 ***
hr10                116.0950    17.8127   6.518 9.84e-11 ***
hr11                151.4004    17.4149   8.694  < 2e-16 ***
hr12                161.8492    18.9649   8.534  < 2e-16 ***
hr13                164.5088    17.4351   9.435  < 2e-16 ***
hr14                158.7299    18.1295   8.755  < 2e-16 ***
hr15                145.7669    18.8187   7.746 1.77e-14 ***
hr16                233.3267    19.0534  12.246  < 2e-16 ***
hr17                374.7796    18.2184  20.571  < 2e-16 ***
hr18                351.4284    17.5642  20.008  < 2e-16 ***
hr19                266.3934    17.6862  15.062  < 2e-16 ***
hr20                178.8539    18.2586   9.796  < 2e-16 ***
hr21                113.8033    17.9216   6.350 2.87e-10 ***
hr22                 77.1929    18.5701   4.157 3.42e-05 ***
hr23                 39.0287    18.5450   2.105 0.035503 *
holidayHoliday      -63.4355    17.3075  -3.665 0.000256 ***
weekdayMonday        24.7634    10.2835   2.408 0.016161 *
weekdayTuesday       19.3706    10.3140   1.878 0.060570 .
weekdayWednesday     19.9628     9.8843   2.020 0.043605 *
weekdayThursday      10.3639     9.6886   1.070 0.284931
weekdayFriday        17.4730     9.9820   1.750 0.080251 .
weekdaySaturday      14.8753    10.1456   1.466 0.142815
workingdayYes             NA         NA      NA       NA
weathersitMedium    -11.7302     6.7061  -1.749 0.080473 .
weathersitBad       -57.2609    10.9407  -5.234 1.91e-07 ***
weathersitReally Bad -77.4406  104.4902  -0.741 0.458736
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 102.5 on 1450 degrees of freedom
Multiple R-squared:  0.6842,    Adjusted R-squared:  0.6735
F-statistic: 64.11 on 49 and 1450 DF,  p-value: < 2.2e-16
```

Figure 23. Summary of the centered Bike Rentals model

Reference and Bibliography

1) John Virzani. «Εισαγωγή στη στατιστική με τη R»
2) Ιωάννης Ντζούφρας . « Εισαγωγή στον προγραμματισμό και στη στατιστική ανάλυση με R»