SCHOOL OF BUSINESS

DEPARTMENT OF MANAGEMENT SCIENCE & TECHNOLOGY

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ACADEMIC YEAR OF 2021 – 2022

STATISTICS FOR BA I – MAIN ASSIGNMENT

NINAS KONSTANTINOS

f2822108

SUPERVISING PROFESSOR: NTZOUFRAS IOANNIS

Table of Contents

1 – Introduction........ 1
2 – Descriptive Analysis of the Data........ 1
3 – Pairwise Comparisons........ 3
4 – Model Construction........ 4
5 – Further Analysis........ 10
6 – Conclusions and Discussions........ 14
7 – Appendix........ 16

1 – Introduction

Bike sharing systems are constituted as the evolution of the traditional bike renting methods, where actions such as subscription, renting and return of the bike have become fully automatic. Many people prefer bikes as their transportation vehicles due their environmental contribution and, of course, because they are way cheaper than most other transportation vehicles. Also, the bikes not only allow the people to avoid traffic, but they also contribute to reducing it. It should be noted that many people prefer this option of bike renting because it is much faster than the traditional methods, where the customer should go to a certain location, for example a shop, to rent or to return their rented bike. The information extracted from analyses of such systems can provide very valuable insights on many different topics, such as health issues of the citizens of a certain city and even environmental issues. Additionally, a plethora of personalized data can be extracted from such systems. Some examples of personalized data that can be extracted are the duration and the distance travelled during each rental, the location of the user and even the user's preference on his/her transportation vehicle.

The sample that will be used on the following analysis includes 1500 observations (rentals) and 18 variables. The variables of the sample include date related data (such as season, day of the week, hour, month, year) and data related to weather conditions (such as temperature, humidity, windspeed and a general weather category). Furthermore, the sample includes indexes that show whether the record was taken during a working day or a holiday and the daily number of users (in total and divided to registered and casual users). The objective of this analysis is the production of a model that can efficiently predict the number of the total users of this specific bike sharing system per hour, in accordance with the rest of the data in the sample. (Table 1)

```
head(bike_sharing)
```

	X	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
11521		11521	2012-04-30	2	1	4	5	0	1	1	1	0.38	0.3939	0.62	0.2537	1	19	20
7158		7158	2011-10-30	4	0	10	20	0	0	0	1	0.34	0.3636	0.57	0.0000	18	74	92
17075		17075	2012-12-19	4	1	12	5	0	3	1	1	0.26	0.2879	0.75	0.0896	2	29	31
10176		10176	2012-03-05	1	1	3	1	0	1	1	1	0.24	0.2424	0.48	0.1343	3	3	6
8555		8555	2011-12-28	1	0	12	5	0	3	1	1	0.32	0.2879	0.57	0.3582	0	9	9
10485		10485	2012-03-17	1	1	3	23	0	6	0	1	0.50	0.4848	0.77	0.1642	34	151	185

Table 1. Raw data

2 – Descriptive Analysis of the Data

In this section, all the procedures that were conducted to clean the data will be described thoroughly. First, the sample was examined for null and missing values. No missing or null values were found in the data. Then, the variable 'x' was dropped from the sample, because it repeats the information of the variable 'instant'. The variable 'instant' was also dropped since it has no meaning in the analysis that will be conducted. Afterwards, the data type of the variable 'dteday' was dropped since the month and the year of the record are included in other variables. Although, before the variable was dropped, the day of the record was saved in another variable named 'day'. The variables of the season (season), of the year (yr), of the month (mnth), of the hour (hr), of the holiday index (holiday), of the weekday (weekday), of the working day index (workingday) and of the weather category (weathersit) were updated to factors. The variables that show the number of casual, of registered and of total users (casual, registered and cnt respectively) were updated to numeric variables. Also, it was noticed that the seasons were assigned in a wrong manner. To fix this issue, the assignment of the seasons was updated from (1: Springer, 2: Summer, 3: Fall, 4: Winter) to (1: Winter, 2: Spring, 3: Summer, 4: Fall). Finally, it was noticed that the variables that were describing the temperatures, the humidity and the wind speed were normalized. To revert these variables to their original unit of measure, each one of those was multiplied with its maximum non-normalized value. (Table 2)

```
> str(bike_sharing)
'data.frame':   1500 obs. of  16 variables:
 $ day       : num  30 30 19 5 28 17 30 18 29 11 ...
 $ season    : Factor w/ 4 levels "Winter","Spring",..: 2 4 4 1 1 1 4 3 2 3 ...
 $ yr        : Factor w/ 2 levels "2011","2012": 2 1 2 2 1 2 1 2 1 2 ...
 $ mnth      : Factor w/ 12 levels "Jan","Feb","Mar",..: 4 10 12 3 12 3 11 8 4 7 ...
 $ hr        : Factor w/ 24 levels "0","1","2","3",..: 6 21 6 2 6 24 16 8 11 16 ...
 $ holiday   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ weekday   : Factor w/ 7 levels "Sunday","Monday",..: 2 1 4 2 4 7 4 7 6 4 ...
 $ workingday: Factor w/ 2 levels "No","Yes": 2 1 2 2 2 1 2 1 2 2 ...
 $ weathersit: Factor w/ 4 levels "Clear Weather",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ temp      : num  15.58 13.94 10.66 9.84 13.12 ...
 $ atemp     : num  19.7 18.2 14.4 12.1 14.4 ...
 $ hum       : num  62 57 75 48 57 77 46 83 40 41 ...
 $ windspeed : num  17 0 6 9 24 ...
 $ casual    : num  1 18 2 3 0 34 11 12 57 56 ...
 $ registered: num  19 74 29 3 9 151 104 52 99 220 ...
 $ cnt       : num  20 92 31 6 9 185 115 64 156 276 ...
```

Table 2. Clean data

```
> summary(bike_sharing)
      day           season         yr            mnth              hr          holiday         weekday      workingday            weathersit          temp      
 Min.   : 1.00   Winter:369   2011:745   May    :142   15     :  74   No :1454   Sunday   :201   No : 451   Clear Weather    :983   Min.   : 0.82  
 1st Qu.: 8.00   Spring:389   2012:755   Apr    :140   23     :  72   Yes:  46   Monday   :211   Yes:1049   Misty-Cloudy     :377   1st Qu.:13.94  
 Median :16.00   Summer:394              Jul    :134   8      :  70              Tuesday  :241              Light Conditions:140   Median :20.50  
 Mean   :15.85   Fall  :348              Aug    :133   13     :  70              Wednesday:191              Heavy Conditions:  0   Mean   :20.47  
 3rd Qu.:23.00                           Jan    :130   17     :  69              Thursday :198                                     3rd Qu.:27.06  
 Max.   :31.00                           Dec    :130   0      :  68              Friday   :254                                     Max.   :40.18  
                                         (Other):691   (Other):1077              Saturday :204                                                    
     atemp            hum           windspeed          casual         registered          cnt       
 Min.   : 0.00   Min.   :  0.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.00   Min.   :  1.0  
 1st Qu.:16.66   1st Qu.: 47.00   1st Qu.: 7.002   1st Qu.:  4.00   1st Qu.: 35.75   1st Qu.: 41.0  
 Median :24.24   Median : 63.00   Median :12.998   Median : 17.50   Median :120.00   Median :151.5  
 Mean   :23.87   Mean   : 62.77   Mean   :12.758   Mean   : 37.05   Mean   :158.86   Mean   :195.9  
 3rd Qu.:31.06   3rd Qu.: 78.00   3rd Qu.:16.998   3rd Qu.: 52.00   3rd Qu.:230.00   3rd Qu.:291.2  
 Max.   :46.21   Max.   :100.00   Max.   :43.999   Max.   :317.00   Max.   :876.00   Max.   :953.0  

> describe(bike_numerics)
           vars    n   mean     sd median trimmed    mad  min    max  range  skew kurtosis   se
day           1 1500  15.85   8.75  16.00   15.85  10.38 1.00  31.00  30.00  0.00    -1.18 0.23
temp          2 1500  20.47   8.01  20.50   20.48   9.73 0.82  40.18  39.36 -0.03    -0.90 0.21
atemp         3 1500  23.87   8.70  24.24   23.99  10.11 0.00  46.21  46.21 -0.12    -0.84 0.22
hum           4 1500  62.77  19.23  63.00   63.06  22.24 0.00 100.00 100.00 -0.11    -0.86 0.50
windspeed     5 1500  12.76   7.96  13.00   12.44   8.89 0.00  44.00  44.00  0.46     0.26 0.21
casual        6 1500  37.05  49.98  17.50   26.48  24.46 0.00 317.00 317.00  2.35     6.49 1.29
registered    7 1500 158.86 155.20 120.00  133.97 136.40 0.00 876.00 876.00  1.53     2.63 4.01
cnt           8 1500 195.92 184.83 151.50  168.91 175.69 1.00 953.00 952.00  1.20     1.16 4.77
```

Table 3. Summary of clean data

Observing the clean data, first, it is noticed that the total amount of users in 2012 is greater in comparison to that of 2011, while both years have almost the same number of recordings (table 3 & figure 2). Also, the total number of users is almost the same throughout the days of a week or of a month, regardless of whether it is a working day or not. Furthermore, it is noticed that the number of bike rentals is almost equal during non-holidays in comparison to the number of rentals in holidays (table 3 & figure 2). It is observed that the average hour of rental is around 12:00 – 13:00pm (figure 2). The least number of total users is observed between 00:00-06:00 am, while the most users are recorded at 08:00 am and 17:00-18:00pm, which are the hours that people are supposedly commuting from their house to their work and vice versa. In addition, it is remarked that the total number of users tend to rise for days with medium to high levels of humidity, while it drops for days with extreme humidity (>80g/m3) (figure 1). As expected, the number of users tends to fall while the weather conditions are getting more intense, such as very low or very high temperature and humidity and high windspeed (figure 1). In fact, for cases of the most extreme weather conditions, no users have been recorded in the sample (table 3). The most usual number of users recorded is between 150 and 200 for registered and total users, while the average amount of casual users recorded is less than 50. Finally, it is obvious that most of the users in the sample are subscribers of the bike sharing system (table 3).

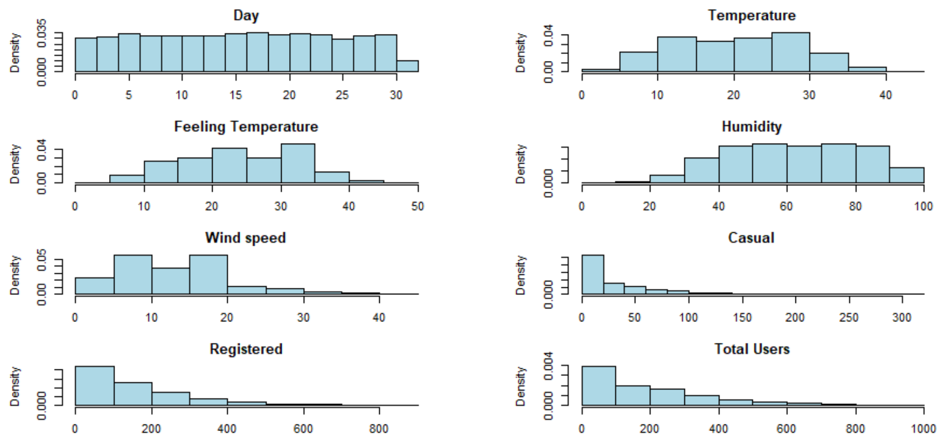

Figure 1. Histograms of numeric variables

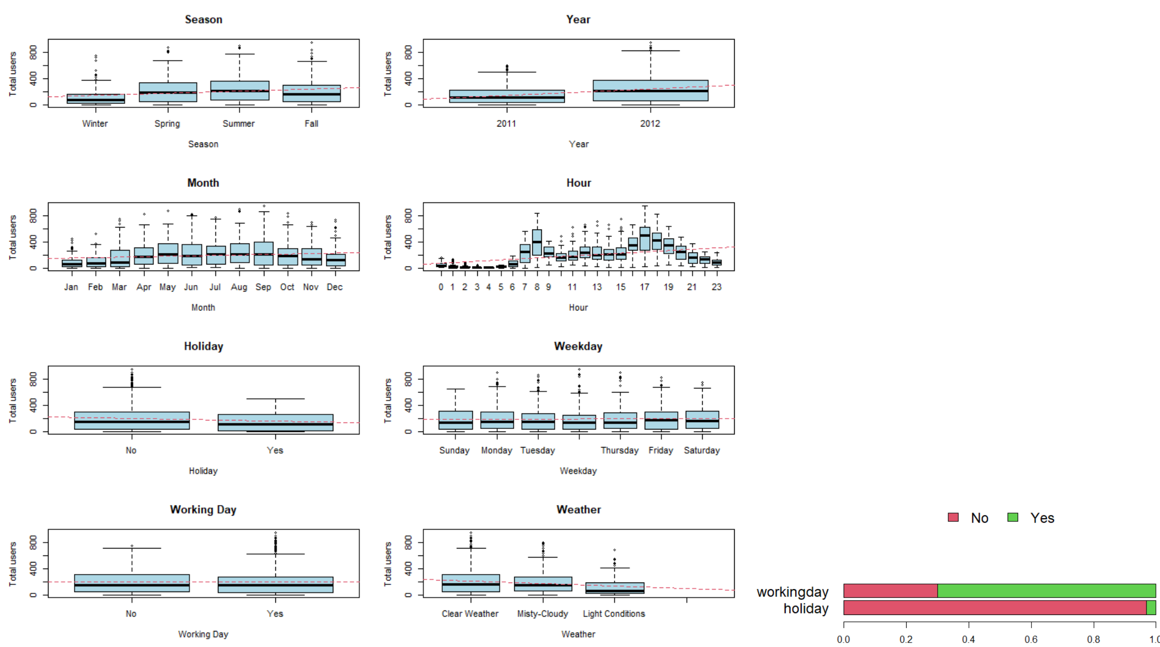

Figure 2. Plots for categorical variables

3 – Pairwise Comparisons

Initially, it is confirmed that the number of users tends to fall while the weather conditions are worsening (figures 2), such as low temperatures and rainfall. These weather conditions are most frequently met during the winter months, during which, indeed the users are less in comparison to the rest of the months. Furthermore, it is observed that the casual users are slightly more tolerant to the weather conditions compared to the registered users, since they react to a smaller degree to changes of the weather conditions (figures 2 & 3). Also, it is perceived that the temperature affects the total number of the daily users to the greatest degree, while the day and the windspeed affect the number of total users to the smallest degree. Specifically, the number of total users increases while the temperature rises, while the number of users doesn't seem to be affected much in any change of the days or the windspeed. It is important to mention that the number of total users seems to drop while the humidity levels increase. (figure 3)

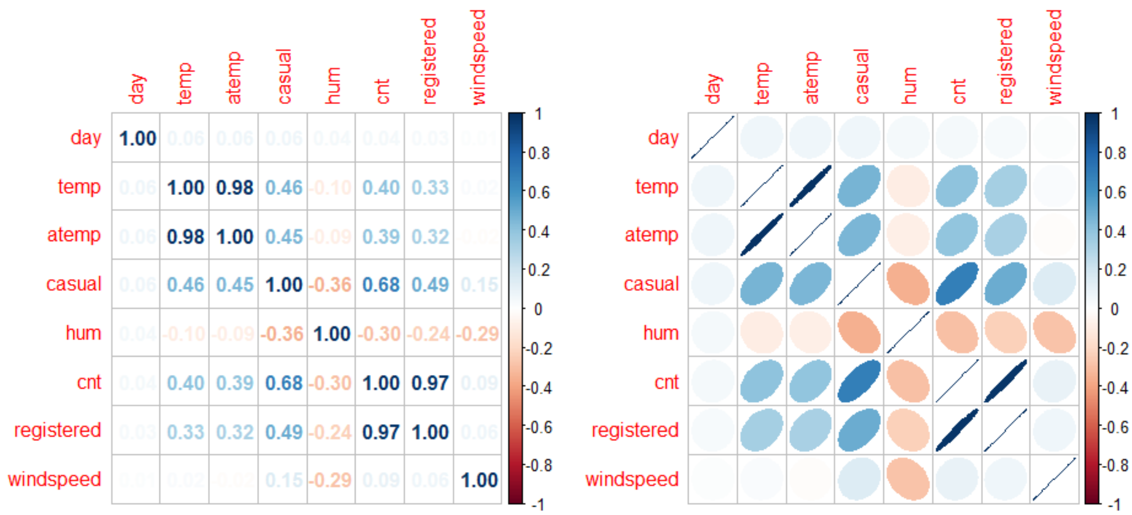

Figure 3. Correlation table of numeric variables

4 – Model Construction

To produce a linear model that predicts the number of total users per hour, first, the lasso method will be conducted using cross validation. Initially, a full model (table 4) is produced to allow the lasso method to drop all the unnecessary variables from it. The variables 'registered' and 'casual', which show the number of subscriber and casual users, are dropped from the full model. This action is executed because those two variables describe fully the number of total users per day, and as a result, they prevent the lasso method to produce good predictive models, since it considers the full model to be perfect, while it most likely is not.

The lasso method shrinks the coefficients of the unnecessary variables to 0 to remove them from the final model it screens based on a tuning parameter λ (lambda). Very big values of that parameter can set a great number of coefficients to zero, while a small value of that parameter can lead to over-fitted models (models with unnecessary explanatory variables). The model that is selected as the best has a lambda value close to one (1) and has a Mean Squared Error (MSE) that is within one standard error of its minimum (figures 4 & 5).

The lasso method screened a model (table 5) that dropped the day, the working day index and the felt temperature variables from the full model. On the model produced from the screening of the lasso method, stepwise methods will be applied to identify the best model to predict the hourly number of users. Since the aim of the analysis is to produce a predictive model, the Akaike Information Criterion (AIC) will be used to explore the possible models for the analysis. Stepwise methods can be applied with a plethora of different techniques. To utilize the full capabilities of the stepwise methods, a constant model is also constructed (table 6). Performing the stepwise method on the model that was produced from the lasso procedure, the weekday variable is also dropped from the model (table 7). In the next steps, the assumptions of the final model will be inspected, to evaluate the model's overall predictive capabilities.

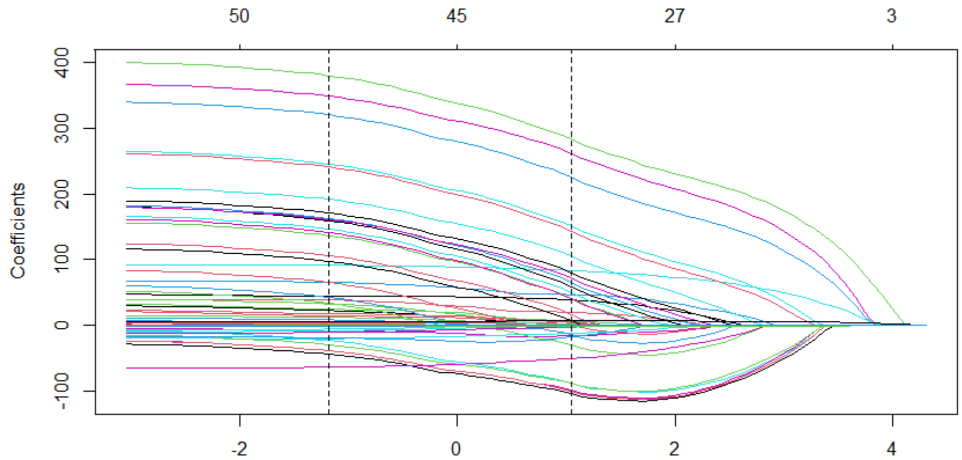

Figure 4. Variable selection – lasso method

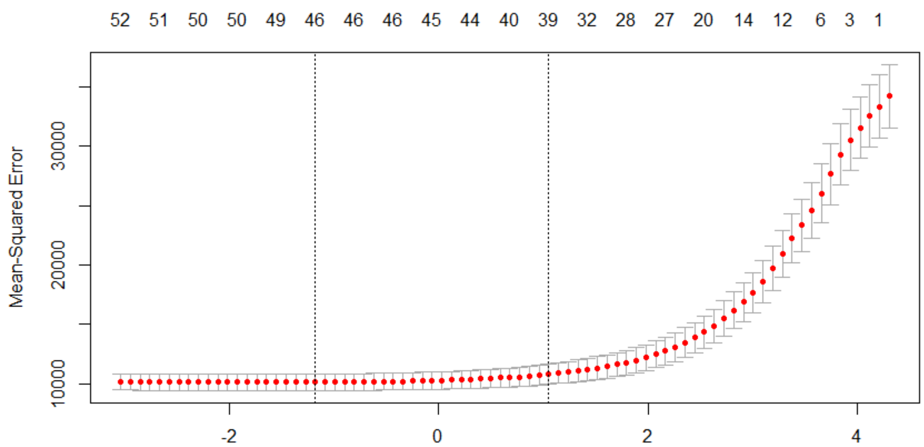

Figure 5. Lasso method using cross validation

The assumptions that will be examined on the final model are the normality and the homoscedasticity and the linearity of its residuals, followed by the independence the sample's observations. Prior to the examination of the above assumptions, the multi-collinearity of the explanatory variables will be examined.

The multi-collinearity of the variables is examined using the General Variance inflation factors method (GVIF). This method identifies linear relationships between explanatory variables. Specifically, the variables that have a GVIF value greater than 3.16 possibly have a multi-collinearity issue. It is obvious that the 'month' and the 'season' variables are heavily correlated with the rest of the variables, while the temperature is correlated with the rest of the variables to a smaller degree (table 8). To fix this issue, the month variable will be attempted to be removed, because it has the biggest GVIF value and the GVIF test will be conducted again. Indeed, it is noticed that after the removal of the month from the model the multi-collinearity problem seems to be fixed (table 9). So, the examination of the rest of the assumptions will proceed with a new model (model 10) that has the same variables as the stepwise model, apart from the month variable.

In the next step, the normality of the residuals will be examined. To examine that assumption, the residuals will be plotted against a normal distribution, and normality tests will be conducted on them. Since both normality tests (Shapiro, Kolmogorov-Smirnov tests) (table 11) conducted have a p-value < 0.05, the null hypothesis that the residuals follow a normal distribution is rejected with a significance level of 5%. The same deduction can be extracted from the plot in (figure 6), in which the residuals slip from the normal line. The violation of the residual's normality may lead to a compromised performance of the hypothesis tests and of the production of the confidence intervals.

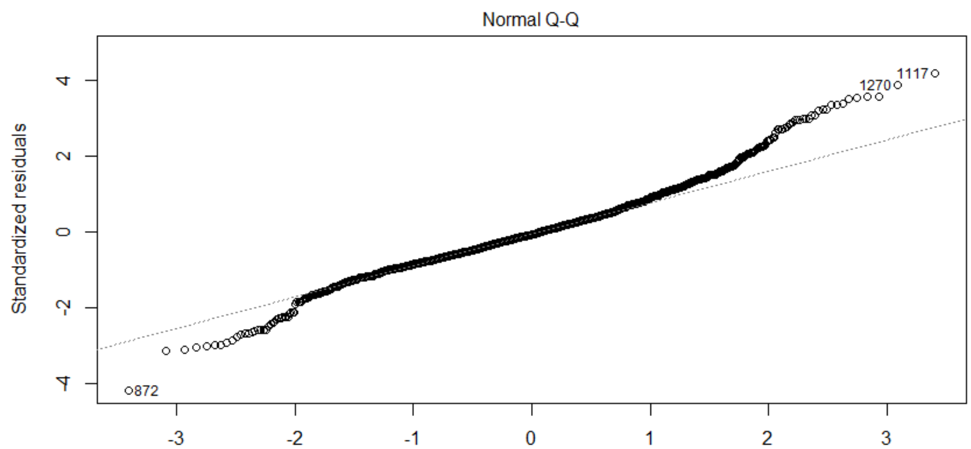

Figure 6. Residual's plot to examine normality

The next assumption that will be examined is the homoscedasticity of the residuals. Specifically, it is examined whether the residuals' variance differs for different values of the dependent (total number of users) variable in the model. To examine that assumption, the residuals are plotted against their fitted values in plots (figure 7), and homoscedasticity tests are conducted. In figure 7, the fitted values are also plotted against the squared residuals and the squared root residuals to reveal possible patterns in the residuals.

To be specific, the tests conducted to examine the residuals' homoscedasticity are the Non-constant variance test (ncv in short) and the Levene test (table 12). The first test examines whether there is a linear effect on the variance of the fitted values. The latter test examines the equality between the variance in the 4 quantiles of the fitted values. It can be observed that both tests have a value that is much smaller than 0.05. So, from the tests alone it can be deducted that with a significance level of 5%, the null hypotheses that the variance between the 4 quantiles of the fitted values is equal and that there is no linear effect on the variance of the fitted values are rejected. Also, by observing the figure 7, the above deductions can be confirmed, since there seems to exist a linear effect on the variance of the fitted values and the variance is not equal in the 4 quantiles of the fitted values in the 'Fitted Values vs Residuals' plot. The latter observation can be also confirmed on the 'Fitted Values vs Squared Residuals' plot and the boxplot of figure 8 as well. The violation of the homoscedasticity assumption may lead to incorrect estimations of the error variance estimator, of the standard errors, and to a compromised performance of the hypothesis tests and the confidence intervals.

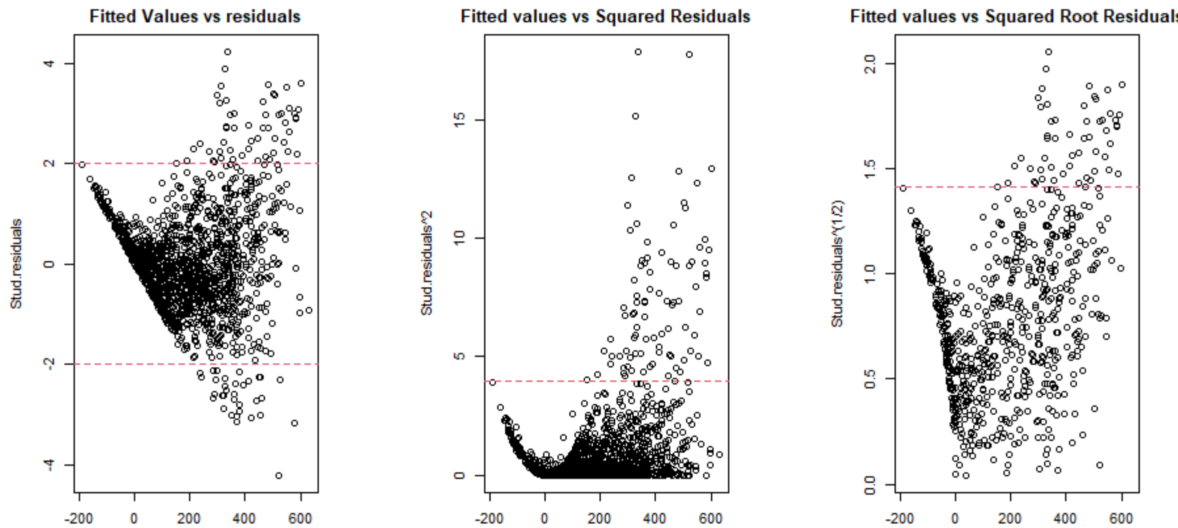

Figure 7. Fitted values vs residuals, squared residuals & squared root of residuals

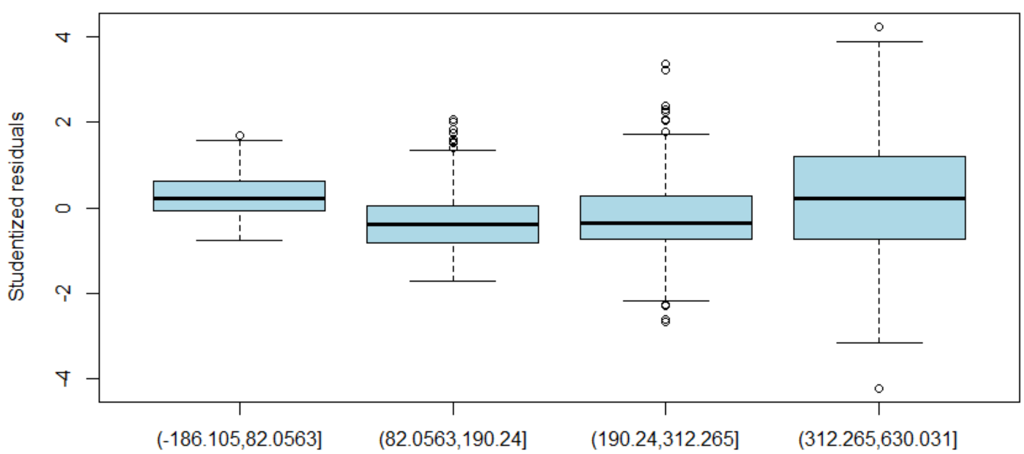

Figure 8. Residuals' variance vs Fitted values

Next, the existence of a linear relationship between the residuals and the fitted values of the model will be examined. To examine that assumption, a Tukey's test will be conducted, and a Residuals' linearity plot will be produced.

Both from the linearity test (table 13) and the plot that examines linearity (figure 9) it can be inferred that there is a violation of the linearity assumption. Specifically, in the Tukey's test it can be observed that there is at least one variable that has a p-value < 0.05, so it can be deducted that with a significance level of 5%, the null hypothesis that there is no linear relationship between the residuals and the fitted values of the model is rejected. The departure of linearity may result to the appearance of the error variance as non-constant, even if it is constant due to the model misspecification. It can also result to an inadequate model in terms of performing the predictions it was produced for.

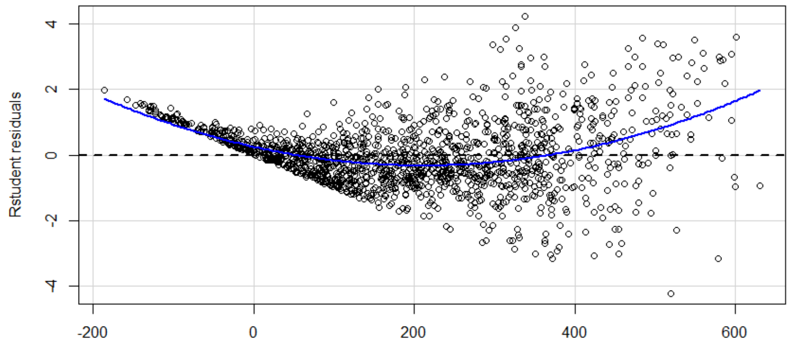

Figure 9. Residuals' linearity plot

Finally, the independence of the observations of the model will be examined. To examine this assumption, a time-sequence plot (figure 10) will be produced, and some critical independence tests (table 14) will be conducted. From the plot in figure 10, there doesn't seem to exist any pattern between the data. The tests conducted are the Runs-test and the Durbin Watson test. The Runs-test examines the randomness between the observations, and the Durbin Watson test examines whether there is an autocorrelation of 1 or more time periods on the observations. Both tests produce a p-value > 0.05, so with a significance level of 5%, we do not reject the null hypotheses of the randomness and of the lack of autocorrelation of 1 or more time periods between the observations.

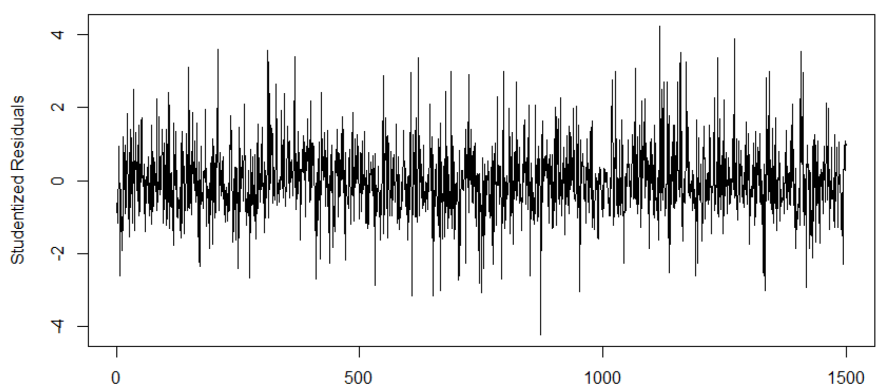

Figure 10. Time-sequence plot

The linearity assumption was fixed by removing the holiday index and the windspeed variables, by adding a polynomial effects of 2nd degree to the humidity variable and of 3rd effect to the temperature variable and by using a logarithmic transformation to the total users 'cnt' and the temperature of the record 'temp' (table 18). The homoscedasticity of the new model was fixed by using a Weighted Least Squares transformation on it (table 17). The independence assumption was not violated after conducting the above procedures (table 19), while the normality assumption is still violated (table 16). As mentioned above, the violation of the normality assumption will result to a compromised performance of the hypothesis tests and of the production of the confidence intervals. The new model produced from the above procedures can be observed in the table 15 and its assumptions in graphical representations in figure 11.

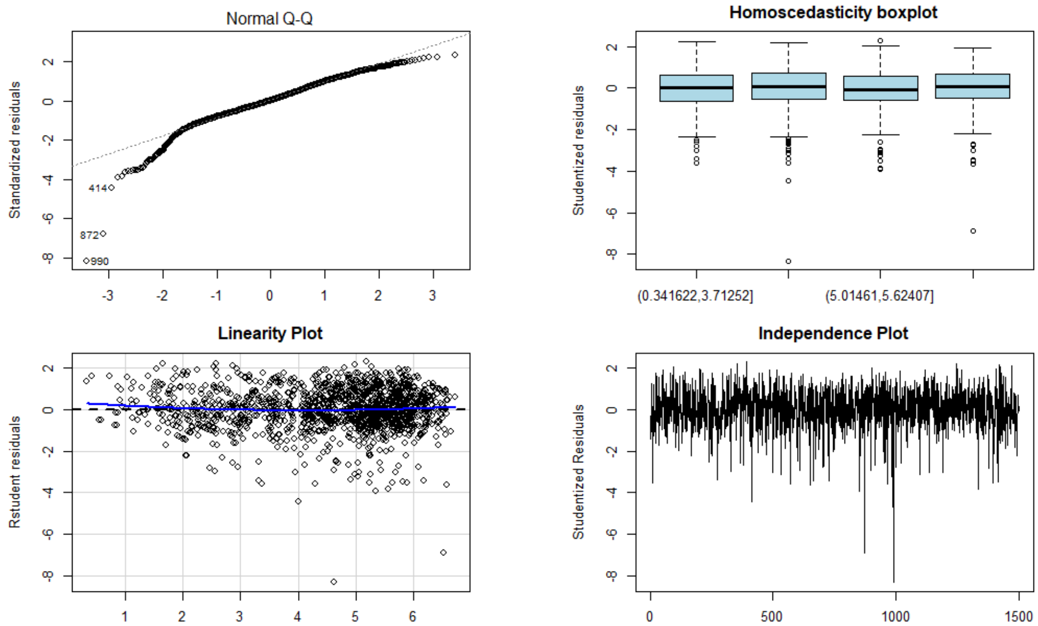

Figure 11. Assumptions of the final model

The final model (table 15) is described from the following equation: $\log(\mathrm{cnt}) = 0.7 + 0.4\mathrm{SeasonSpring} + 0.4\mathrm{SeasonSummer} + 0.5\mathrm{SeasonFall} + 0.5\mathrm{year2012} - 0.6\mathrm{hr1} - 1.2\mathrm{hr2} - 1.7\mathrm{hr3} - 1.9\mathrm{hr4} - 0.8\mathrm{hr5} + 0.3\mathrm{hr6} + 1.5\mathrm{hr7} + 2.2\mathrm{hr8} + 1.9\mathrm{hr9} + 1.3\mathrm{hr10} + 1.5\mathrm{hr11} + 1.6\mathrm{hr12} + 1.5\mathrm{hr13} + 1.5\mathrm{hr14} + 1.5\mathrm{hr15} + 1.9\mathrm{hr16} + 2.2\mathrm{hr17} + 2\mathrm{hr18} + 1.9\mathrm{hr19} + 1.6\mathrm{hr20} + 1.2\mathrm{hr21} + \mathrm{hr22} + 0.8\mathrm{hr23} - 0.05\mathrm{WeatherMisty} - 0.5\mathrm{WeatherLight} + 0.8\log(\mathrm{temp}) + 0.01\mathrm{hum} - 0.00001\mathrm{temp}^3 - 0.0001\mathrm{hum}^2 + \varepsilon$. The distribution of the residuals is not known, since the normality assumption for the model was rejected. As a result, in the model's interpretation, the residuals distribution will remain blank.

Almost 81% of the number of total users can be explained from the model's explanatory variables. Additionally, it must be mentioned that the model's residual standard error cannot be interpreted accurately or compared with that of the rest of the models, due to the logarithmic transformation of the dependent variable (the number of total users) of the model. The model's fit can be regarded as a 'good fit' since it has an Adj. R-Squared value greater than 0.7.

To determine the significance of the variables in terms of their effect to the total number of users, a threshold for a statistical significance equal to a = 5% will be held. The intercept has a significant effect to the number of users. To simplify the interpretation of the effects of the explanatory variables to the total number of users, all variables will be raised to the exponential number 'e'. According to the model's intercept, if it is midnight (00:00), the season is

Winter, the year is in 2011, the humidity is equal to 0 g/m3 and the equation $\exp(0.8\log(temp)) - \exp(0.00001temp^3) = 0$, only a total of $\exp(0.7) = 2$ users will be recorded. The temperature also has a significant effect on the number of users. Each subsequent one-unit increase in the temperature will result to an increase of the total users up to 69 degrees, after which point, each increase in the temperature will result to a decrease of the total number of users. The effect of the season change is significant for the total number of hourly users. If the season is Summer or Spring, the number of total users will increase by (exp(0.4)-1)*100 = 49%, and if the season is Fall, the increase will be equal to 65% in comparison to the number of total users during the Winter season when the rest of the variables remain fixed. The effect of the year variable in the model is also significant. If the year of the record is in 2012, the total users will be 65% more than the users in 2011 when the rest of the explanatory variables remain fixed. The effect of the hour is also significant for the total number of hourly users. In comparison to the number of users at 00:00am, if the recorded hour is 01:00 am the number of rentals will decrease by 82%, if the hour is 02:00 am the number of rentals will decrease by 164% when the rest of the variables remain fixed, and so on and so forth. Regarding the weather variables, the Misty weather is insignificant to the number of total users, so its effect on the number of users can be disregarded. On the other hand, if the weather is ‘light’, the number of total users will decrease by 65% in comparison to the number of users during clear weather when the rest of the variables remain unchanged. The humidity has a significant effect to the number of total users. The subsequent increases of humidity will result to a declining increase of the total users up to 100 g/m3, in which point, the effect of the increase of humidity will be equal to zero, while the rest of the variables remain unchanged.

5 – Further Analysis

To assess the predictive capabilities of the models that were constructed prior to the final model (full model, lasso model, stepwise model, and the constant model), a test sample will be used, so that the models can conduct their predictions on data that are not included in their initial sample. All the procedures that were applied in the initial dataset (removal of variables, update of data types etc.) to clean its data, are also applied on the test dataset. Then, the predictive ability of each model will be examined, by conducting predictions with them on the test sample and by comparing their Root Mean Squared Error (RMSE) (table 20) values of the predictions to identify the model with the best predictive capabilities. The RMSE value calculates the distance of the model’s residuals from the line of best fit, and as a result, models with low RMSE values are considered better. It is observed that the model with the best predictive performance on the out-of-sample test dataset was the stepwise model (table 7), while the model with the worst performance was, as expected, the constant model. It is also worth mentioning that the RMSE values are relatively close for the full, the lasso and the stepwise models, while it increases vastly for the constant model.

Concluding, a typical day in each season, according to the initial dataset will be described, starting from the Spring season based on the figures 12 & 13 and the table 21. At a typical day during spring, the temperature will be approximately 23 degrees and the felt temperature will be approximately 26 degrees, the humidity will be equal to 63 g/m3 and the windspeed will be equal to 13 km/h. In addition, the number of total users will be approximately equal to 219, 168 of which will be registered users, and the rest 51 will be casual users. The hours during which the number of users surge the most, are 08:00 am and 17:00-19:00pm, while the least number of users are met between 02:00-05:00am. Finally, most likely the weather will be either clear or misty and it will be a non-holiday working day.

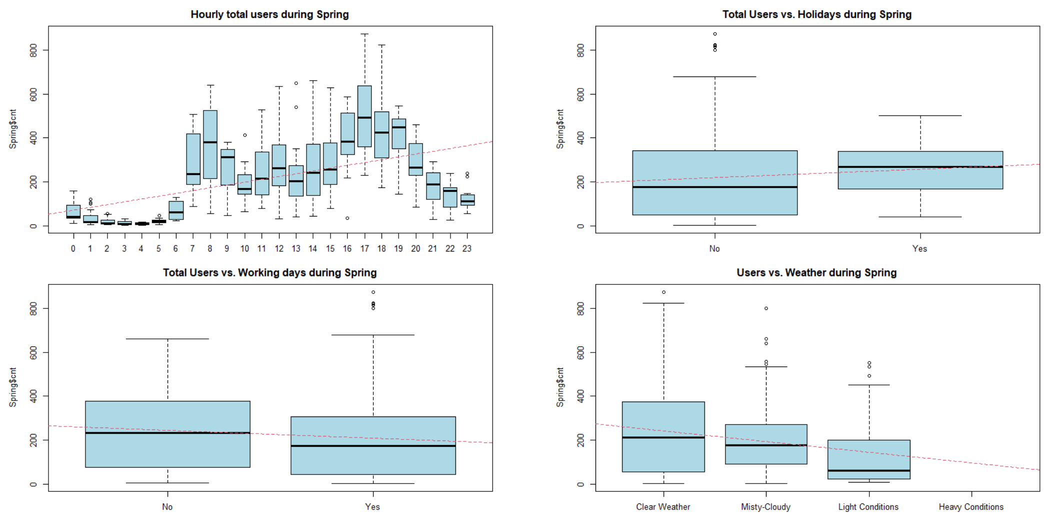

Figure 12. Spring – plots for categorical variables

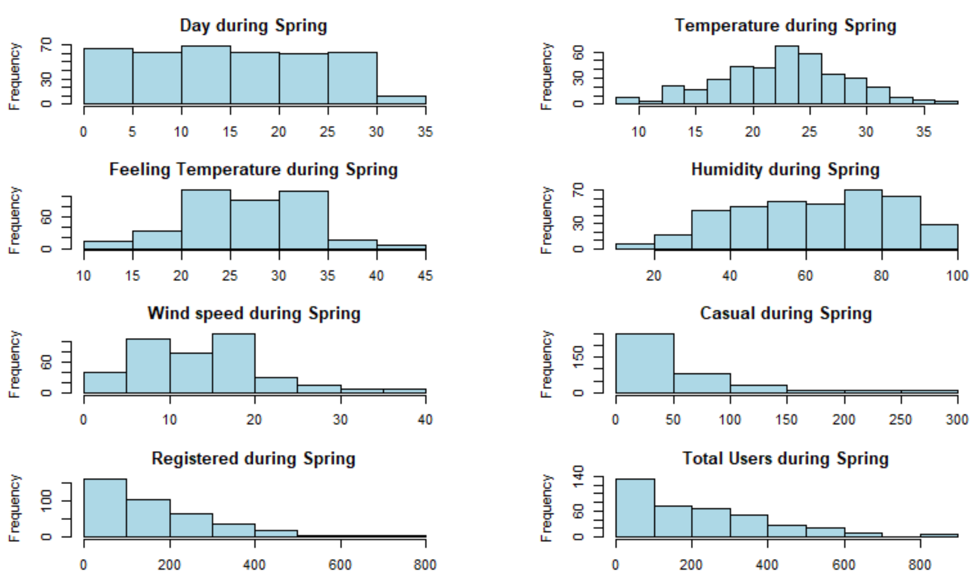

Figure 13. Spring – plots for numeric variables

Next, a typical day in summer will be described by interpreting the table 22 and the figures 13 & 14. At a typical day during summer, the temperature and the felt temperature will be at 29 and 33 degrees respectively, the windspeed will be equal to 12 km/h and the humidity will be 62 g/m3. Furthermore, a total of 247 users will be recorded, of which, 195 will be registered users. The hourly surges of users are at the same times with the hourly users during the spring season, but in a greater degree. Also, it can be noted that there is a significant difference in the number of total users from the surges to the rest of the hours of a day. Finally, a typical day during summer is a non-holiday, working day with clear weather.

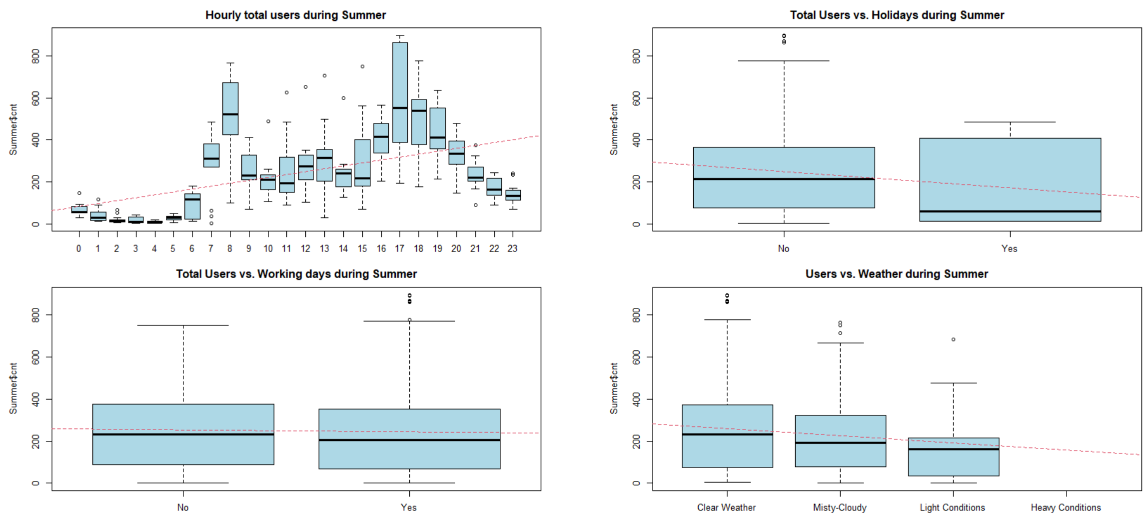

Figure 13. Summer – plots for categorical variables

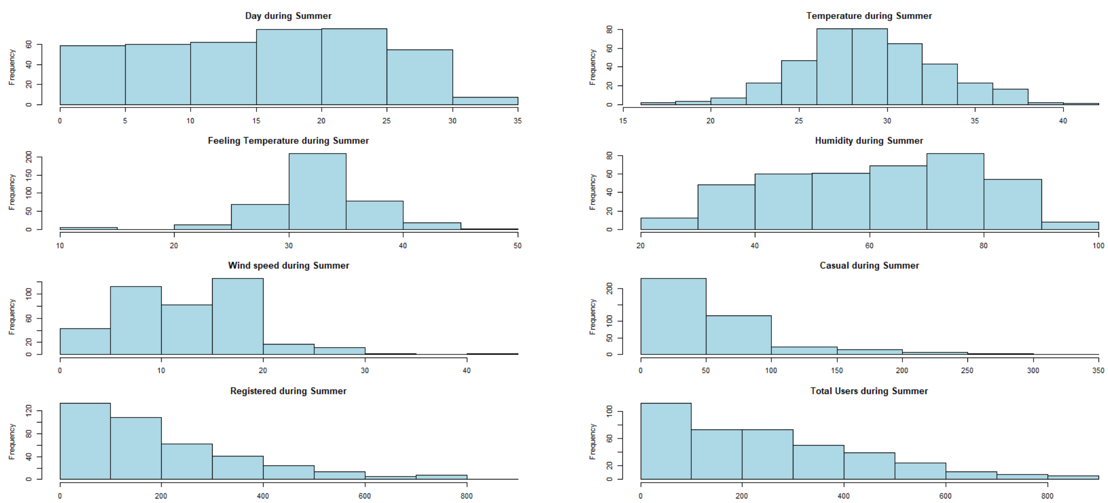

Figure 14. Summer – plots for numeric variables

Following, a typical day during winter will be examined by interpreting the table 23 and the figures 15 & 16. At a typical day during winter, the temperature and the felt temperature will be at 12 and 15 degrees respectively, the windspeed will be equal to 14 km/h and the humidity will be equal to 59 g/m3. The number of total users will be approximately 110, of which, 95 will be registered users. The smallest number of total users is met between 21:00pm-05:00am. The number of total users through a typical winter day is more homogenous in comparison to the above seasons, which means that these rental surges do not exist at the same degree as they do in spring or summer. Also, the fact that there are much less users in comparison to the previous seasons examined, implies that a typical day in winter has heavy weather conditions, whether it is a working day or a holiday or not.

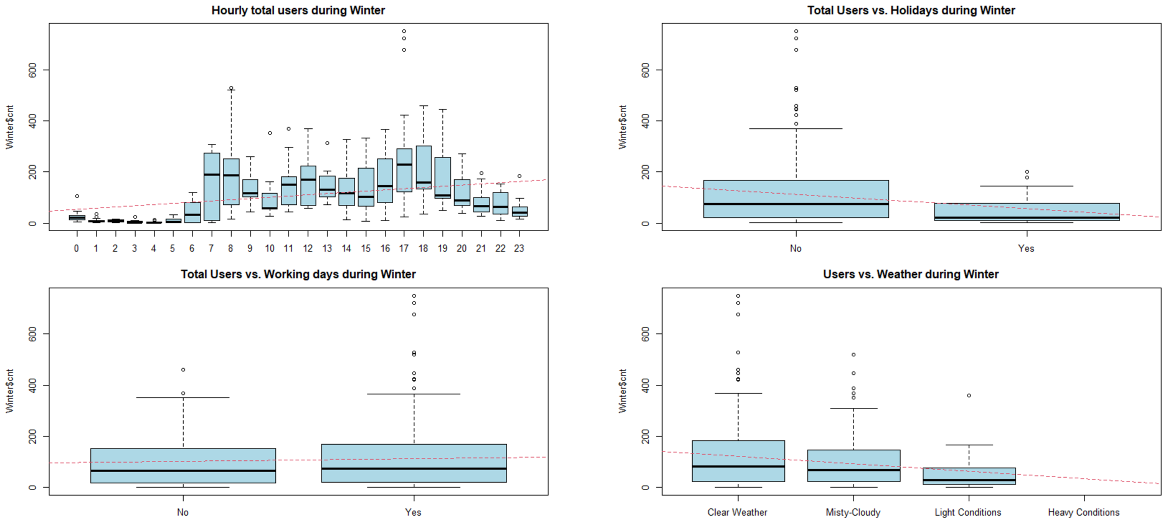

Figure 15. Winter – plots for categorical variables

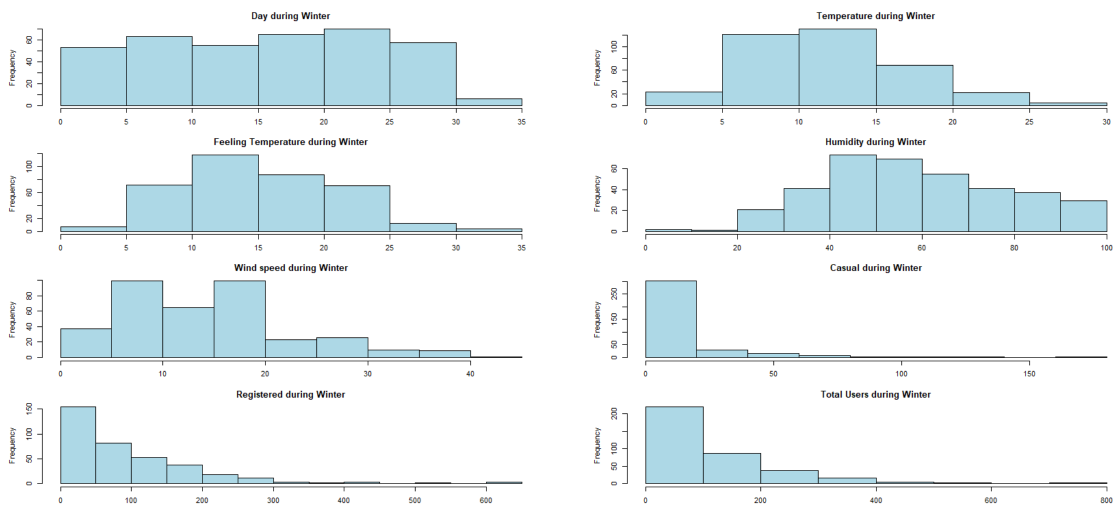

Figure 16. Winter – plots for numeric variables

Finally, a typical day during fall will examined by interpreting the table 24 and the figures 17 & 18. At a typical fall day, the temperature and the felt temperature will be 17 and 20 degrees respectively, the windspeed will be equal to 12 km/h and the humidity will be equal to 67 g/m3. The number of total users will be equal to 203, of which, 175 will be registered users. In the passing of a day, the most users are met at 08:00 am and at 17:00-18:00pm. In the interval between these surges, the number of total users decreases vastly. The weather will be misty-cloudy, and it will be a non-holiday, working day.

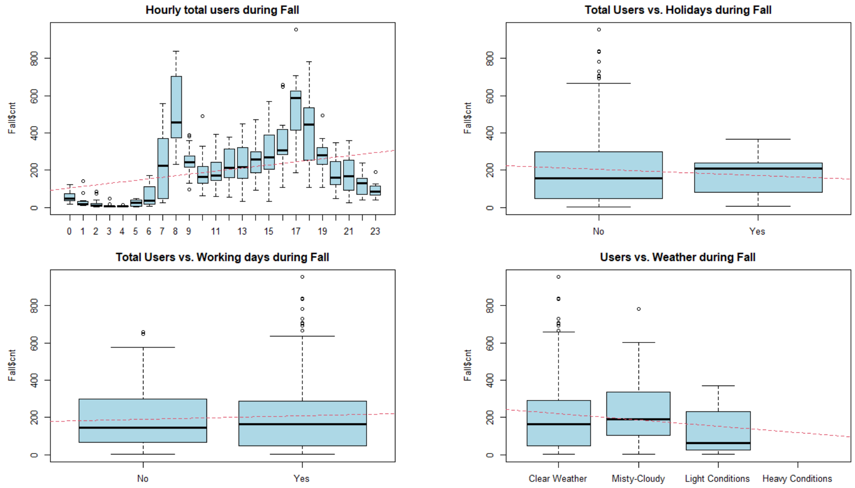

Figure 17. Fall - plots for categorical variables

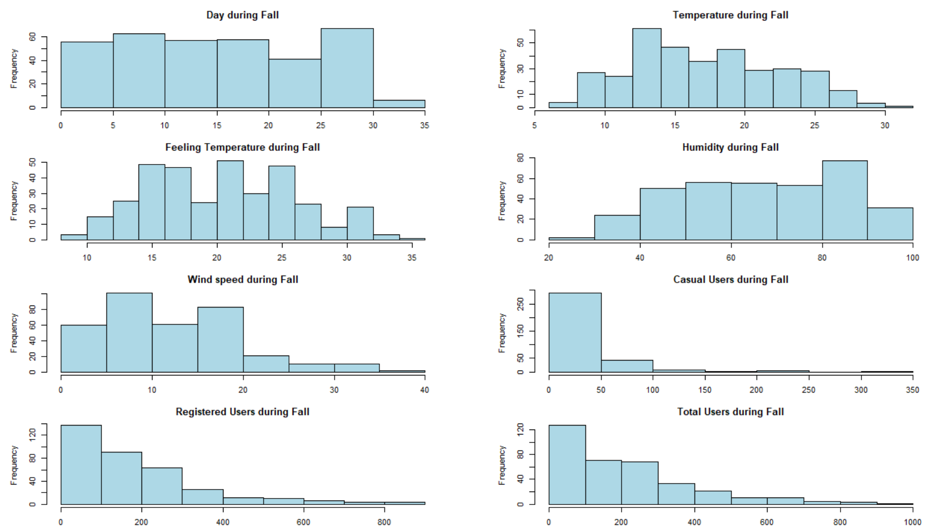

Figure 18. Fall - plots for numeric variables

6 – Conclusions and Discussions

Summarizing the deductions of the analysis, it can be inferred that people generally prefer good weather conditions to use the rent bikes, regardless of whether they are registered users or not. In other words, they prefer clear weather conditions, medium temperature, and relatively high levels of humidity while they are not affected almost at all from the wind speed. As a result, during seasons that usually have worse conditions than the ones stated above, such as winter, the number of daily total users tends to fall. This fact is confirmed from the fact that the smallest number of average total users is met during the winter, while the biggest number of average total users is met during summer. This probably stems from the fact that during those seasons, people prefer to use other means of transportations, such as their personal car or the bus, to avoid those extreme weather conditions. Furthermore, it can be deducted, though without great accuracy, that the number of total users per year tends to rise. It should be mentioned that the number of total users does not seem to vary significantly in holidays in comparison to the rest of the days, and that people seem to use the bike sharing systems to the same extent throughout the days of a week, regardless of whether it is a working day or not. This implies that people tend to use these bikes not only to conduct their daily responsibilities, but also for their own entertainment or exercise. Although, a surge of total bike rentals can be observed in every season during the rush hours that people supposedly go to, or return from, their work, their school etc. In other words, it can be inferred that the main reason that the users prefer this option as a means for transportation is to conduct their daily responsibilities.

To produce a model with predictive capabilities, a cross validation lasso method and a stepwise method were applied to a full model. Initially, the model that was produced from the stepwise method, did not comply with 3 out of 4 model assumptions (Normality, Homoscedasticity, Linearity, and Independence). To fix that issue, the model was transformed using log, polynomial effects to the explanatory variables and a weighted least squares transformation to the model itself. These transformations fixed the homoscedasticity and the linearity assumptions of the model but failed to fix the normality assumption. It is important to mention that due to the transformations it is not possible to compare the final model's predictive capabilities with those of the rest of the models. As a result, it is not currently known whether the final model has better predictive capabilities in comparison to the rest of the models. Concluding, it should be noted that with further research and analysis on the final model and the sample, that it is possible that a model can be constructed that complies with all 4 model assumptions and has good predictive features.

7 – Appendix

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

Residuals:
    Min      1Q  Median      3Q     Max 
-394.36  -59.92   -7.35   51.02  363.22 

Coefficients: (1 not defined because of singularities)
                            Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -90.26156   21.92658  -4.117 4.06e-05 ***
day                         -0.01055    0.29630  -0.036 0.971612    
seasonSpring                37.48919   17.65715   2.123 0.033909 *  
seasonSummer                37.21940   19.74123   1.885 0.059581 .  
seasonFall                  65.76679   16.49395   3.987 7.02e-05 ***
yr2012                      93.14839    5.22011  17.844  < 2e-16 ***
mnthFeb                     -3.03790   12.71886  -0.239 0.811256    
mnthMar                     32.45164   14.95819   2.169 0.030208 *  
mnthApr                     25.16690   22.61074   1.113 0.265871    
mnthMay                     37.04105   24.29997   1.524 0.127646    
mnthJun                     12.76138   24.91334   0.512 0.608568    
mnthJul                    -14.17081   27.37911  -0.518 0.604832    
mnthAug                      4.95031   26.84014   0.184 0.853697    
mnthSep                     51.72333   23.61542   2.190 0.028666 *  
mnthOct                     27.06882   22.00068   1.230 0.218761    
mnthNov                      8.54104   20.83219   0.410 0.681872    
mnthDec                      6.80120   16.76232   0.406 0.684991    
hr1                         -4.73657   17.21590  -0.275 0.783257    
hr2                        -20.79764   18.29156  -1.137 0.255723    
hr3                        -25.82475   17.62200  -1.465 0.143006    
hr4                        -21.15079   18.17555  -1.164 0.244740    
hr5                        -12.10467   17.36168  -0.697 0.485786    
hr6                         63.54488   18.36303   3.460 0.000555 ***
hr7                        212.72741   18.55831  11.463  < 2e-16 ***
hr8                        370.40262   16.93679  21.870  < 2e-16 ***
hr9                        193.23639   17.59691  10.981  < 2e-16 ***
hr10                       127.78222   17.29843   7.387 2.53e-13 ***
hr11                       159.99711   17.67005   9.055  < 2e-16 ***
hr12                       186.42486   18.69861   9.970  < 2e-16 ***
hr13                       169.98745   17.46622   9.732  < 2e-16 ***
hr14                       165.74096   17.90402   9.257  < 2e-16 ***
hr15                       184.27192   17.41536  10.581  < 2e-16 ***
hr16                       265.79281   18.22887  14.581  < 2e-16 ***
hr17                       404.26348   17.47117  23.139  < 2e-16 ***
hr18                       343.55592   17.55952  19.565  < 2e-16 ***
hr19                       269.02982   17.79663  15.117  < 2e-16 ***
hr20                       183.43742   17.44860  10.513  < 2e-16 ***
hr21                       119.74786   17.08142   7.010 3.63e-12 ***
hr22                        87.10396   17.28097   5.040 5.23e-07 ***
hr23                        55.19100   16.76426   3.292 0.001018 ** 
holidayYes                 -27.65922   15.89984  -1.740 0.082143 .  
weekdayMonday                3.81420   10.14421   0.376 0.706974    
weekdayTuesday              -1.34566    9.54964  -0.141 0.887959    
weekdayWednesday            11.37952   10.08437   1.128 0.259325    
weekdayThursday             11.18944   10.04183   1.114 0.265343    
weekdayFriday               15.20233    9.47088   1.605 0.108675    
weekdaySaturday             14.59374    9.87926   1.477 0.139837    
workingdayYes                     NA         NA      NA       NA    
weathersitMisty-Cloudy     -10.45070    6.55354  -1.595 0.111006    
weathersitLight Conditions -66.17164   10.07484  -6.568 7.09e-11 ***
temp                         6.43424    2.01277   3.197 0.001420 ** 
atemp                       -1.15120    1.69032  -0.681 0.495946    
hum                         -0.80622    0.18392  -4.383 1.25e-05 ***
windspeed                   -0.83928    0.35441  -2.368 0.018009 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.26 on 1447 degrees of freedom
Multiple R-squared:  0.7272,	Adjusted R-squared:  0.7174 
F-statistic: 74.17 on 52 and 1447 DF,  p-value: < 2.2e-16
```

```
> summary(lasso_model)

Call:
lm(formula = cnt ~ season + yr + mnth + hr + holiday + weekday + 
    weathersit + temp + hum + windspeed, data = bike_sharing)

Residuals:
    Min      1Q  Median      3Q     Max 
-396.69  -59.90   -7.02   51.95  364.02 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -92.5188    21.4632  -4.311 1.74e-05 ***
seasonSpring                37.3024    17.6370   2.115 0.034599 *  
seasonSummer                37.0954    19.7064   1.882 0.059981 .  
seasonFall                  65.8324    16.4670   3.998 6.71e-05 ***
yr2012                      93.2173     5.2163  17.870  < 2e-16 ***
mnthFeb                     -3.1319    12.7001  -0.247 0.805249    
mnthMar                     32.4668    14.9386   2.173 0.029915 *  
mnthApr                     24.9346    22.5923   1.104 0.269917    
mnthMay                     37.1872    24.2782   1.532 0.125811    
mnthJun                     13.1987    24.8739   0.531 0.595762    
mnthJul                    -13.4465    27.3386  -0.492 0.622899    
mnthAug                      6.7807    26.6888   0.254 0.799480    
mnthSep                     52.3134    23.5869   2.218 0.026716 *  
mnthOct                     26.8124    21.9858   1.220 0.222839    
mnthNov                      8.2918    20.8178   0.398 0.690463    
mnthDec                      6.5196    16.7477   0.389 0.697122    
hr1                         -4.7170    17.2058  -0.274 0.784009    
hr2                        -20.8900    18.2650  -1.144 0.252929    
hr3                        -25.5948    17.6064  -1.454 0.146240    
hr4                        -21.0421    18.1643  -1.158 0.246879    
hr5                        -11.9561    17.3417  -0.689 0.490655    
hr6                         63.7361    18.3511   3.473 0.000530 ***
hr7                        213.0744    18.5358  11.495  < 2e-16 ***
hr8                        370.4603    16.9254  21.888  < 2e-16 ***
hr9                        193.2677    17.5833  10.992  < 2e-16 ***
hr10                       128.0922    17.2436   7.428 1.87e-13 ***
hr11                       160.0680    17.6604   9.064  < 2e-16 ***
hr12                       186.5344    18.6773   9.987  < 2e-16 ***
hr13                       170.1883    17.4229   9.768  < 2e-16 ***
hr14                       165.5745    17.8868   9.257  < 2e-16 ***
hr15                       184.3539    17.4044  10.592  < 2e-16 ***
hr16                       265.5947    18.2167  14.580  < 2e-16 ***
hr17                       404.3395    17.4537  23.166  < 2e-16 ***
hr18                       343.9179    17.5214  19.628  < 2e-16 ***
hr19                       269.1278    17.7866  15.131  < 2e-16 ***
hr20                       183.4608    17.4309  10.525  < 2e-16 ***
hr21                       119.8853    17.0711   7.023 3.34e-12 ***
hr22                        87.3741    17.2559   5.063 4.65e-07 ***
hr23                        55.3341    16.7540   3.303 0.000981 ***
holidayYes                 -27.2258    15.8787  -1.715 0.086630 .  
weekdayMonday                3.7553    10.1308   0.371 0.710931    
weekdayTuesday              -1.1722     9.5412  -0.123 0.902233    
weekdayWednesday            11.3972    10.0773   1.131 0.258254    
weekdayThursday             11.3023    10.0340   1.126 0.260181    
weekdayFriday               15.6356     9.4444   1.656 0.098030 .  
weekdaySaturday             14.6325     9.8728   1.482 0.138529    
weathersitMisty-Cloudy     -10.2466     6.5402  -1.567 0.117401    
weathersitLight Conditions -65.5187    10.0192  -6.539 8.54e-11 ***
temp                         5.1634     0.7578   6.814 1.39e-11 ***
hum                         -0.8138     0.1830  -4.446 9.40e-06 ***
windspeed                   -0.7963     0.3485  -2.285 0.022470 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.21 on 1449 degrees of freedom
Multiple R-squared:  0.7271,	Adjusted R-squared:  0.7177 
F-statistic: 77.21 on 50 and 1449 DF,  p-value: < 2.2e-16
```

Table 4. Full Model

Table 5. Lasso Model

```
> summary(constant_model)

Call:
lm(formula = cnt ~ 1, data = bike_sharing)

Residuals:
    Min      1Q  Median      3Q     Max 
-194.92 -154.92  -44.42   95.33  757.08 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  195.917      4.772   41.05   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 184.8 on 1499 degrees of freedom
```

Table 6. Constant Model

```
Call:
lm(formula = cnt ~ hr + temp + yr + season + weathersit + mnth +
    hum + windspeed + holiday, data = bike_sharing)

Residuals:
    Min      1Q  Median      3Q     Max
-405.67  -58.21   -6.21   51.93  370.11

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)               -83.1001    20.5452  -4.045 5.51e-05 ***
hr1                        -5.0482    17.2021  -0.293 0.769209
hr2                       -21.1799    18.2570  -1.160 0.246198
hr3                       -25.5002    17.5737  -1.451 0.146984
hr4                       -22.1142    18.1238  -1.220 0.222596
hr5                       -11.4466    17.3294  -0.661 0.509018
hr6                        62.7561    18.3477   3.420 0.000643 ***
hr7                       213.0906    18.5334  11.498  < 2e-16 ***
hr8                       370.9583    16.8908  21.962  < 2e-16 ***
hr9                       191.8284    17.5477  10.932  < 2e-16 ***
hr10                      127.7175    17.2407   7.408 2.16e-13 ***
hr11                      159.0597    17.6367   9.019  < 2e-16 ***
hr12                      185.9699    18.6555   9.969  < 2e-16 ***
hr13                      170.2672    17.3957   9.788  < 2e-16 ***
hr14                      164.1036    17.8438   9.197  < 2e-16 ***
hr15                      182.9301    17.3683  10.532  < 2e-16 ***
hr16                      263.8532    18.1427  14.543  < 2e-16 ***
hr17                      403.1968    17.4104  23.158  < 2e-16 ***
hr18                      343.0406    17.4547  19.653  < 2e-16 ***
hr19                      268.4402    17.7436  15.129  < 2e-16 ***
hr20                      183.8776    17.3984  10.569  < 2e-16 ***
hr21                      119.0007    17.0553   6.977 4.55e-12 ***
hr22                       87.7321    17.2516   5.085 4.14e-07 ***
hr23                       55.3302    16.7506   3.303 0.000979 ***
temp                        5.1567     0.7517   6.860 1.02e-11 ***
yr2012                     93.0363     5.2080  17.864  < 2e-16 ***
seasonSpring               38.5318    17.6279   2.186 0.028987 *
seasonSummer               37.6997    19.6978   1.914 0.055828 .
seasonFall                 65.6466    16.4431   3.992 6.87e-05 ***
weathersitMisty-Cloudy     -9.3360     6.5173  -1.433 0.152215
weathersitLight Conditions -66.0743    9.9932  -6.612 5.31e-11 ***
mnthFeb                    -3.8930    12.6837  -0.307 0.758944
mnthMar                    30.5310    14.8966   2.050 0.040589 *
mnthApr                    22.8274    22.5623   1.012 0.311825
mnthMay                    34.0794    24.1920   1.409 0.159137
mnthJun                    12.1573    24.8093   0.490 0.624187
mnthJul                   -15.0082    27.2246  -0.551 0.581531
mnthAug                     5.5666    26.6444   0.209 0.834539
mnthSep                    52.2850    23.5303   2.222 0.026435 *
mnthOct                    25.7250    21.9516   1.172 0.241431
mnthNov                     8.1605    20.7777   0.393 0.694560
mnthDec                     6.1769    16.7314   0.369 0.712047
hum                        -0.8234     0.1825  -4.512 6.95e-06 ***
windspeed                  -0.7651     0.3476  -2.201 0.027900 *
holidayYes                -28.5036    15.1680  -1.879 0.060418 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 98.23 on 1455 degrees of freedom
Multiple R-squared:  0.7259,    Adjusted R-squared:  0.7176
F-statistic: 87.55 on 44 and 1455 DF,  p-value: < 2.2e-16
```

Table 7. Stepwise model

```
> round(vif(model6),2)#multi-collinearity is ok!
           GVIF Df GVIF^(1/(2*Df))
season     3.13  3            1.21
yr         1.03  1            1.02
hr         1.82 23            1.01
holiday    1.03  1            1.02
weathersit 1.40  2            1.09
temp       3.14  1            1.77
hum        1.83  1            1.35
windspeed  1.18  1            1.08
```

Table 9. GVIF for the model without the month variable

```
> round(vif(model5),1) #multi-collinearity
            GVIF Df GVIF^(1/(2*Df))
hr           2.4 23             1.0
temp         5.6  1             2.4
yr           1.1  1             1.0
season     211.5  3             2.4
weathersit   1.4  2             1.1
mnth       464.3 11             1.3
hum          1.9  1             1.4
windspeed    1.2  1             1.1
holiday      1.1  1             1.0
```

Table 8. GVIF on the model extracted from stepwise method

```
> summary(model6)

Call:
lm(formula = cnt ~ season + yr + hr + holiday + weathersit +
    temp + hum + windspeed, data = bike_sharing)

Residuals:
    Min      1Q  Median      3Q     Max
-411.24  -59.99   -7.32   49.80  412.94

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                  -92.7778    19.3685  -4.790 1.84e-06 ***
seasonSpring                  55.2369     9.3455   5.911 4.23e-09 ***
seasonSummer                  37.6426    11.9428   3.152 0.001655 **
seasonFall                    74.9259     8.0876   9.264  < 2e-16 ***
yr2012                        92.8050     5.2115  17.808  < 2e-16 ***
hr1                           -4.6486    17.3195  -0.268 0.788429
hr2                          -24.3453    18.3344  -1.328 0.184435
hr3                          -22.6955    17.6141  -1.288 0.197780
hr4                          -23.2655    18.2179  -1.277 0.201781
hr5                           -9.9891    17.4454  -0.573 0.567006
hr6                           59.9195    18.4385   3.250 0.001181 **
hr7                          211.0695    18.6600  11.311  < 2e-16 ***
hr8                          370.2463    16.9809  21.804  < 2e-16 ***
hr9                          190.7329    17.6555  10.803  < 2e-16 ***
hr10                         131.5183    17.3485   7.581 6.05e-14 ***
hr11                         159.6542    17.6942   9.023  < 2e-16 ***
hr12                         188.3838    18.6667  10.092  < 2e-16 ***
hr13                         173.2375    17.3199  10.002  < 2e-16 ***
hr14                         163.9717    17.7858   9.219  < 2e-16 ***
hr15                         183.8153    17.2815  10.637  < 2e-16 ***
hr16                         263.2406    18.0444  14.588  < 2e-16 ***
hr17                         406.5666    17.3885  23.381  < 2e-16 ***
hr18                         346.1332    17.5072  19.771  < 2e-16 ***
hr19                         272.3177    17.7480  15.344  < 2e-16 ***
hr20                         184.5617    17.5294  10.529  < 2e-16 ***
hr21                         120.2398    17.1555   7.009 3.65e-12 ***
hr22                          84.0929    17.3392   4.850 1.37e-06 ***
hr23                          52.9923    16.8488   3.145 0.001693 **
holidayYes                   -30.6353    15.0820  -2.031 0.042411 *
weathersitMisty-Cloudy        -9.5610     6.5519  -1.459 0.144706
weathersitLight Conditions -68.4489    10.0481  -6.812 1.40e-11 ***
temp                           5.3692     0.5669   9.471  < 2e-16 ***
hum                           -0.6280     0.1803  -3.482 0.000511 ***
windspeed                     -0.6606     0.3489  -1.894 0.058481 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 99.21 on 1466 degrees of freedom
Multiple R-squared:  0.7182,    Adjusted R-squared:  0.7119
F-statistic: 113.2 on 33 and 1466 DF,  p-value: < 2.2e-16
```

Table 10. Stepwise model without month variable (model 6)

```
> shapiro.test(model6$residuals)

        Shapiro-Wilk normality test

data:  model6$residuals
W = 0.97492, p-value = 1.659e-15

> lillie.test(model6$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  model6$residuals
D = 0.057929, p-value = 8.942e-13
```

Table 11. Normality tests

```
> ncvTest(model6)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 389.9911, Df = 1, p = < 2.22e-16
> leveneTest(rstudent(model6)~yhat.quantiles)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value    Pr(>F)
group    3  114.63 < 2.2e-16 ***
      1495
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 12. Homoscedasticity tests

```
> residualPlots(model6, plot=F, type = "rstudent")
           Test stat Pr(>|Test stat|)
season
yr
hr
holiday
weathersit
temp         -1.0742           0.2829
hum          -2.3087           0.0211 *
windspeed    -0.6500           0.5158
Tukey test   22.7499           <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 13. Tukey’s Linearity test

```
> runs.test(model6$res)

        Runs Test

data:  model6$res
statistic = -0.77486, runs = 736, n1 = 750, n2 = 750, n = 1500, p-value = 0.4384
alternative hypothesis: nonrandomness

> durbinWatsonTest(model6)
 lag Autocorrelation D-W Statistic p-value
   1      0.01424665      1.970651     0.6
 Alternative hypothesis: rho != 0
```

Table 14. Independence tests

```
Call:
lm(formula = log(cnt) ~ season + yr + hr + weathersit + log(temp) +
    hum + I(temp^3) + I(hum^2), data = bike_sharing, weights = wt)

Weighted Residuals:
     Min       1Q   Median       3Q      Max
-10.6960  -0.7244   0.0162   0.8944   3.0104

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                7.159e-01  2.179e-01   3.285 0.001045 **
seasonSpring               4.024e-01  4.995e-02   8.056 1.62e-15 ***
seasonSummer               3.585e-01  5.914e-02   6.062 1.71e-09 ***
seasonFall                 4.882e-01  4.453e-02  10.964  < 2e-16 ***
yr2012                     4.878e-01  2.599e-02  18.768  < 2e-16 ***
hr1                       -6.290e-01  1.301e-01  -4.836 1.47e-06 ***
hr2                       -1.242e+00  1.481e-01  -8.387  < 2e-16 ***
hr3                       -1.738e+00  1.494e-01 -11.630  < 2e-16 ***
hr4                       -1.897e+00  1.594e-01 -11.900  < 2e-16 ***
hr5                       -8.084e-01  1.346e-01  -6.007 2.38e-09 ***
hr6                        2.601e-01  1.297e-01   2.006 0.045047 *
hr7                        1.491e+00  1.090e-01  13.683  < 2e-16 ***
hr8                        2.206e+00  9.510e-02  23.199  < 2e-16 ***
hr9                        1.646e+00  1.008e-01  16.326  < 2e-16 ***
hr10                       1.315e+00  1.035e-01  12.699  < 2e-16 ***
hr11                       1.469e+00  1.030e-01  14.262  < 2e-16 ***
hr12                       1.579e+00  1.047e-01  15.075  < 2e-16 ***
hr13                       1.497e+00  1.007e-01  14.862  < 2e-16 ***
hr14                       1.477e+00  1.035e-01  14.270  < 2e-16 ***
hr15                       1.558e+00  1.011e-01  15.413  < 2e-16 ***
hr16                       1.857e+00  9.948e-02  18.663  < 2e-16 ***
hr17                       2.227e+00  9.456e-02  23.547  < 2e-16 ***
hr18                       2.050e+00  9.551e-02  21.463  < 2e-16 ***
hr19                       1.886e+00  9.790e-02  19.262  < 2e-16 ***
hr20                       1.606e+00  1.011e-01  15.886  < 2e-16 ***
hr21                       1.244e+00  1.040e-01  11.957  < 2e-16 ***
hr22                       1.044e+00  1.072e-01   9.743  < 2e-16 ***
hr23                       7.549e-01  1.080e-01   6.989 4.20e-12 ***
weathersitMisty-Cloudy    -4.937e-02  3.232e-02  -1.528 0.126841
weathersitLight Conditions -5.322e-01  5.916e-02  -8.995  < 2e-16 ***
log(temp)                  7.978e-01  6.639e-02  12.017  < 2e-16 ***
hum                        1.291e-02  4.191e-03   3.081 0.002104 **
I(temp^3)                 -8.737e-06  2.146e-06  -4.071 4.92e-05 ***
I(hum^2)                  -1.340e-04  3.487e-05  -3.843 0.000127 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.331 on 1466 degrees of freedom
Multiple R-squared:  0.813,     Adjusted R-squared:  0.8088
F-statistic: 193.2 on 33 and 1466 DF,  p-value: < 2.2e-16
```

Table 15. Final model

```
> shapiro.test(wls_model$residuals)

        Shapiro-Wilk normality test

data:  wls_model$residuals
W = 0.95069, p-value < 2.2e-16

> lillie.test(wls_model$residuals)

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  wls_model$residuals
D = 0.083954, p-value < 2.2e-16
```

Table 16. Normality tests on final model

```
> ncvTest(wls_model)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 2.16755e-05, Df = 1, p = 0.99629
> leveneTest(rstudent(wls_model)~yhat.quantiles)
Levene's Test for Homogeneity of Variance (center = median)
        Df F value Pr(>F)
group    3  1.2666 0.2843
      1495
```

Table 17. Homoscedasticity tests on final model

```
> residualPlots(wls_model, plot=F, type = "rstudent")
           Test stat Pr(>|Test stat|)
season
yr
hr
weathersit
log(temp)     1.7822          0.07492 .
hum           1.2558          0.20939
I(temp^3)    -1.2485          0.21204
I(hum^2)      0.8398          0.40116
Tukey test    1.2725          0.20320
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Table 18. Linearity test on final model

```
> runs.test(wls_model$res)

        Runs Test

data:  wls_model$res
statistic = -1.0331, runs = 731, n1 = 750, n2 = 750, n = 1500, p-value = 0.3015
alternative hypothesis: nonrandomness

> durbinWatsonTest(wls_model)
 lag Autocorrelation D-W Statistic p-value
   1      0.02016849      1.959546   0.432
 Alternative hypothesis: rho != 0
```

Table 19. Independence test on final model

```
data.frame(Full_model = RMSE_full
          ,Lasso_model =  RMSE_lasso,Stepwise_model =  RMSE_step
          ,Constant_model = RMSE_Constant)
Full_model Lasso_model Stepwise_model Constant_model
  104.5207      104.488       104.2883       182.9544
```

Table 20. RMSE values of all models

```
> summary(Spring)
      day              hr        holiday   workingday              weathersit        temp            atemp            hum           windspeed         casual          registered        cnt
 Min.   : 1.00   0       : 22   No :374   No :123   Clear Weather     :241   Min.   : 9.02   Min.   :11.37   Min.   : 17.00   Min.   : 0.000   Min.   :  0.00   Min.   :  2   Min.   :  2.0
 1st Qu.: 8.00   18      : 22   Yes: 15   Yes:266   Misty-Cloudy      :108   1st Qu.:18.86   1st Qu.:22.73   1st Qu.: 46.00   1st Qu.: 7.002   1st Qu.:  7.00   1st Qu.: 45   1st Qu.: 52.0
 Median :15.00   12      : 21                       Light Conditions: 40   Median :22.96   Median :26.52   Median : 65.00   Median :12.998   Median : 29.00   Median :141   Median :183.0
 Mean   :15.65   21      : 20                       Heavy Conditions:  0   Mean   :22.63   Mean   :26.50   Mean   : 62.98   Mean   :13.327   Mean   : 51.31   Mean   :168   Mean   :219.3
 3rd Qu.:23.00   1       : 19                                              3rd Qu.:26.24   3rd Qu.:31.06   3rd Qu.: 78.00   3rd Qu.:19.001   3rd Qu.: 64.00   3rd Qu.:244   3rd Qu.:341.0
 Max.   :31.00   5       : 19                                              Max.   :36.90   Max.   :41.66   Max.   :100.00   Max.   :39.001   Max.   :291.00   Max.   :769   Max.   :873.0
                 (Other):266
> describe(Spring[,sapply(Spring,class) == 'numeric'])
           vars   n   mean     sd median trimmed    mad   min    max  range  skew kurtosis   se
day           1 389  15.65   8.98  15.00   15.60  11.86  1.00  31.00  30.00  0.04    -1.19 0.46
temp          2 389  22.63   5.59  22.96   22.65   4.86  9.02  36.90  27.88 -0.03    -0.25 0.28
atemp         3 389  26.50   6.06  26.52   26.67   6.74 11.37  41.66  30.30 -0.20    -0.10 0.31
hum           4 389  62.98  20.23  65.00   63.58  25.20 17.00 100.00  83.00 -0.21    -0.92 1.03
windspeed     5 389  13.33   7.87  13.00   13.12   8.89  0.00  39.00  39.00  0.39     0.15 0.40
casual        6 389  51.31  61.76  29.00   39.04  37.06  0.00 291.00 291.00  1.84     3.19 3.13
registered    7 389 167.99 149.18 141.00  148.04 148.26  2.00 769.00 767.00  1.27     1.80 7.56
cnt           8 389 219.30 188.56 183.00  197.67 207.56  2.00 873.00 871.00  0.92     0.34 9.56
```

Table 21. Spring descriptives

```
> summary(Summer)
      day             hr       holiday   workingday            weathersit       temp           atemp            hum           windspeed          casual          registered         cnt        
 Min.   : 1.00   13      : 23   No :388   No :105    Clear Weather   :285   Min.   :16.40   Min.   :12.12   Min.   : 25.0   Min.   : 0.000   Min.   :  0.00   Min.   :  1.0   Min.   :  1.00  
 1st Qu.: 9.00   16      : 21   Yes:  6   Yes:289    Misty-Cloudy    : 84   1st Qu.:26.24   1st Qu.:30.30   1st Qu.: 48.0   1st Qu.: 7.002   1st Qu.: 13.00   1st Qu.: 61.5   1st Qu.: 75.25  
 Median :17.00   18      : 21                        Light Conditions: 25   Median :28.70   Median :33.34   Median : 63.5   Median :11.001   Median : 40.50   Median :158.0   Median :212.00  
 Mean   :16.07   19      : 20                        Heavy Conditions:  0   Mean   :29.00   Mean   :32.68   Mean   : 61.7   Mean   :12.054   Mean   : 51.92   Mean   :195.3   Mean   :247.23  
 3rd Qu.:23.00   23      : 19                                               3rd Qu.:31.16   3rd Qu.:34.85   3rd Qu.: 74.0   3rd Qu.:16.998   3rd Qu.: 73.75   3rd Qu.:280.2   3rd Qu.:365.75  
 Max.   :31.00   3       : 18                                               Max.   :40.18   Max.   :46.21   Max.   :100.0   Max.   :43.001   Max.   :307.00   Max.   :810.0   Max.   :897.00  
                 (Other):272                                                                                                                                                                 
> describe(Summer[,sapply(Summer,class) == 'numeric'])
           vars   n   mean     sd median trimmed    mad   min    max  range  skew kurtosis    se
day           1 394  16.07   8.52  17.00   16.08  10.38  1.00  31.00  30.00 -0.05    -1.10  0.43
temp          2 394  29.00   3.95  28.70   28.99   3.65 16.40  40.18  23.78 -0.04     0.18  0.20
atemp         3 394  32.68   4.76  33.34   32.86   3.37 12.12  46.21  34.09 -0.89     3.34  0.24
hum           4 394  61.70  17.68  63.50   62.02  21.50 25.00 100.00  75.00 -0.14    -1.00  0.89
windspeed     5 394  12.05   6.83  11.00   12.08   5.93  0.00  43.00  43.00  0.20     0.34  0.34
casual        6 394  51.92  52.10  40.50   43.10  43.74  0.00 307.00 307.00  1.80     4.00  2.62
registered    7 394 195.31 170.89 158.00  172.31 160.86  1.00 810.00 809.00  1.18     1.25  8.61
cnt           8 394 247.23 202.00 212.00  225.16 212.01  1.00 897.00 896.00  0.85     0.24 10.18
```

Table 22. Summer descriptives

```
> summary(Winter)
      day             hr       holiday   workingday            weathersit       temp           atemp            hum           windspeed          casual          registered          cnt       
 Min.   : 1.00   8       : 24   No :353   No :125    Clear Weather   :244   Min.   : 0.82   Min.   : 0.00   Min.   :  0.0   Min.   : 0.000   Min.   :  0.00   Min.   :  1.00   Min.   :  1.0  
 1st Qu.: 8.00   15      : 24   Yes: 16   Yes:244    Misty-Cloudy    : 88   1st Qu.: 8.20   1st Qu.:10.61   1st Qu.: 44.0   1st Qu.: 7.002   1st Qu.:  1.00   1st Qu.: 19.00   1st Qu.: 20.0  
 Median :17.00   23      : 20                        Light Conditions: 37   Median :12.30   Median :14.39   Median : 57.0   Median :12.998   Median :  5.00   Median : 64.00   Median : 72.0  
 Mean   :16.01   0       : 19                        Heavy Conditions:  0   Mean   :12.22   Mean   :14.90   Mean   : 59.3   Mean   :14.025   Mean   : 14.75   Mean   : 94.74   Mean   :109.5  
 3rd Qu.:23.00   1       : 18                                               3rd Qu.:15.58   3rd Qu.:19.70   3rd Qu.: 75.0   3rd Qu.:19.001   3rd Qu.: 15.00   3rd Qu.:141.00   3rd Qu.:164.0  
 Max.   :31.00   14      : 18                                               Max.   :29.52   Max.   :32.58   Max.   :100.0   Max.   :43.999   Max.   :172.00   Max.   :642.00   Max.   :750.0  
                 (Other):246                                                                                                                                                                 
> describe(Winter[,sapply(Winter,class) == 'numeric'])
           vars   n   mean     sd median trimmed   mad  min    max  range  skew kurtosis   se
day           1 369  16.01   8.55  17.00   16.07 10.38 1.00  31.00  30.00 -0.07    -1.20 0.44
temp          2 369  12.22   5.00  12.30   12.00  4.86 0.82  29.52  28.70  0.47     0.20 0.26
atemp         3 369  14.90   5.79  14.39   14.71  5.62 0.00  32.57  32.57  0.32    -0.14 0.30
hum           4 369  59.30  20.03  57.00   58.98 20.76 0.00 100.00 100.00  0.12    -0.66 1.04
windspeed     5 369  14.02   8.76  13.00   13.64  8.89 0.00  44.00  44.00  0.55     0.19 0.46
casual        6 369  14.75  27.08   5.00    8.17  7.41 0.00 172.00 172.00  3.38    12.83 1.41
registered    7 369  94.74 102.18  64.00   77.61 78.58 1.00 642.00 641.00  2.13     6.49 5.32
cnt           8 369 109.49 117.70  72.00   89.95 88.96 1.00 750.00 749.00  1.96     5.55 6.13
```

Table 23. Winter descriptives

```
> summary(Fall)
      day             hr       holiday   workingday            weathersit       temp            atemp             hum            windspeed          casual          registered         cnt       
 Min.   : 1.00   3       : 22   No :339   No : 98    Clear Weather   :213   Min.   : 6.56   Min.   : 8.335   Min.   : 20.00   Min.   : 0.000   Min.   :  0.00   Min.   :  0.0   Min.   :  1.0  
 1st Qu.: 8.00   10      : 19   Yes:  9   Yes:250    Misty-Cloudy    : 97   1st Qu.:13.12   1st Qu.:15.910   1st Qu.: 53.00   1st Qu.: 6.003   1st Qu.:  4.00   1st Qu.: 45.0   1st Qu.: 48.0  
 Median :15.00   8       : 18                        Light Conditions: 38   Median :16.40   Median :20.455   Median : 68.50   Median :11.001   Median : 16.00   Median :131.0   Median :159.5  
 Mean   :15.66   14      : 18                        Heavy Conditions:  0   Mean   :17.14   Mean   :20.489   Mean   : 67.44   Mean   :11.578   Mean   : 27.93   Mean   :175.4   Mean   :203.3  
 3rd Qu.:24.00   22      : 18                                               3rd Qu.:21.32   3rd Qu.:25.000   3rd Qu.: 83.00   3rd Qu.:16.998   3rd Qu.: 33.00   3rd Qu.:239.0   3rd Qu.:295.5  
 Max.   :31.00   23      : 18                                               Max.   :30.34   Max.   :34.090   Max.   :100.00   Max.   :36.997   Max.   :317.00   Max.   :876.0   Max.   :953.0  
                 (Other):235                                                                                                                                                                     
> describe(Fall[,sapply(Fall,class) == 'numeric'])
           vars   n   mean     sd median trimmed    mad   min    max  range  skew kurtosis    se
day           1 348  15.66   8.99  15.00   15.61  11.86  1.00  31.00  30.00  0.07    -1.26  0.48
temp          2 348  17.14   5.14  16.40   16.99   4.86  6.56  30.34  23.78  0.27    -0.74  0.28
atemp         3 348  20.49   5.63  20.46   20.34   6.74  8.33  34.09  25.75  0.18    -0.76  0.30
hum           4 348  67.44  17.99  68.50   67.84  21.50 20.00 100.00  80.00 -0.16    -1.03  0.96
windspeed     5 348  11.58   8.15  11.00   11.08   7.41  0.00  37.00  37.00  0.44    -0.25  0.44
casual        6 348  27.93  39.79  16.00   19.90  19.27  0.00 317.00 317.00  3.25    14.32  2.13
registered    7 348 175.39 169.63 131.00  147.57 146.78  0.00 876.00 876.00  1.54     2.54  9.09
cnt           8 348 203.32 187.89 159.50  175.90 174.95  1.00 953.00 952.00  1.24     1.36 10.07
```

Table 24. Fall Descriptives