ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ATHENS UNIVERSITY OF ECONOMICS AND BUSINESS

ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΕΠΙΧΕΙΡΗΣΕΩΝ SCHOOL OF BUSINESS

ΜΕΤΑΠΤΥΧΙΑΚΟ στην ΕΠΙΧΕΙΡΗΜΑΤΙΚΗ ΑΝΑΛΥΤΙΚΗ

MSc in BUSINESS ANALYTICS

MSc in Business Analytics

Course:

Statistics for Business Analytics I

Professor:

Ioannis Ntzoufras

Main Assignment:

Bike Sharing Rental Analysis

Training dataset: bike_50

Student:

Name: Maria Skoli

AM: p2822131

Dec 2021

Table of Contents

Introduction

Nowadays, numerous measures are implemented to eliminate environmental pollution. Due to the high emissions produced by cars, a substantial portion of these measures concern alternative modes of transport. Within this context, bike sharing systems have been designed in an attempt to address micro-mobility needs -like commuting to and from metro stations to enter public transport grid- as well as facilitating transportation between neighboring communities. The first bike-share programs began in the Netherlands as early as the 1960s, but the concept did not take off worldwide until the mid-2000s. Today, there are more than 500 bike sharing programs across Europe and the US. with these systems of alternative transport becoming of great interest as they are growing in popularity.

The spread of bike sharing programs created the need for monitoring and analyzing the data that are often extracted from the bike sensors. More specifically, one of the main goals for many bike sharing providers is to predict the mobility rate that determines the number of bikes rented each hour, month, or year as well as the conditions that may affect this rate. Such predictive capabilities will allow these businesses to optimize the number of bikes available in order to fully satisfy existing demand while maintaining a sustainable fleet size.

To facilitate these needs, historical data from the Capital Bikeshare system is used documenting rental bike usage in Washington D.C. for the years 2011 and 2012. This data is publicly available at http://capitalbikeshare.com/system-data. The corresponding weather and seasonal information extracted from http://www.freemeteo.com are also utilized in the analysis. The data corresponds to the years 2011 and 2012.

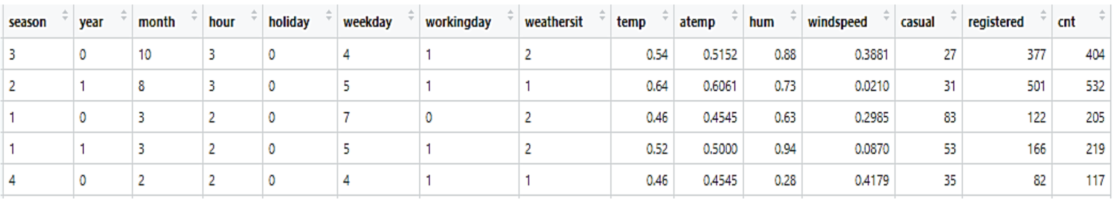
The aim of this assignment is to determine which among the available features influence bike rental counts and use these features to accurately predict the expected demand. In the pages below, certain particularities of the features will first be explored in an exploratory analysis based on visual elements. Then, predictive models are constructed in an attempt to identify the key variables that contribute to or affect the rentals demand. Once the optimal model is determined, it will be used to describe the typical profile of a day for each season.

Descriptive analysis and exploratory data analysis

The analysis begins by loading the data and giving an overview of the structure and dimensions of the dataset.

The dataset contains 1500 observations and for each of them 15 attributes are recorded including weather situation, date, weekday, holiday days, the number of bikes rented on that date and temperature. We have 1500 observations of the hourly data. While this dataset is relatively tidy, some cleaning was still required before moving into the exploratory analysis. For example, windspeed had the zero value for a large number of observations making it likely that this information was missing, and R automatically converted it into zero. We ameliorate this problem by generating random values from the normal distribution into the “zero” values. (Appendix *1*).

Before moving into the analysis, the record code, and the date of the observation (variables: instant and date) are dropped. The date is accounted for through the variables of hour/year/day and the instant is an indexing variable of no interest for the analysis presented here.



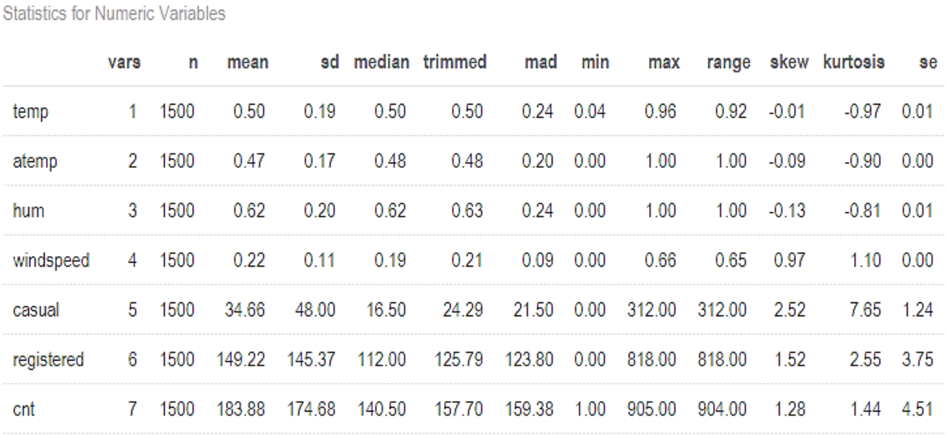
season	year	month	hour	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
3	0	10	3	0	4	1	2	0.54	0.5152	0.88	0.3881	27	377	404
2	1	8	3	0	5	1	1	0.64	0.6061	0.73	0.0210	31	501	532
1	0	3	2	0	7	0	2	0.46	0.4545	0.63	0.2985	83	122	205
1	1	3	2	0	5	1	2	0.52	0.5000	0.94	0.0870	53	166	219
4	0	2	2	0	4	1	1	0.46	0.4545	0.28	0.4179	35	82	117

Table 1: First 5 rows of the dataset

This dataset consists of both numeric and categorical variables. The temperature and humidity are continuous variables and hence numeric while the possible weather conditions (variable *“weathersit”*) are recorded in the form of a categorical variable with 4 possible levels that describe the weather at the day and time of the observation. In the exploratory analysis presented below, the numeric variables will be examined through histograms and Q-Q plots and the categorical variables through frequency tables and bar plots.

Numeric Variables - Descriptive Statistics

As seen in Table 2, for the first 4 variables the mean and median are identical, something to be expected since. it is known that these variables have been normalized. The mean temperature recorded during the hours included in the data set was 20,5 = (0,50*41). Another interesting observation arising from the descriptive statistics in Table 2 is that the number of rides taken by registered riders have a much larger standard deviation than those taken by casual riders. As expected, registered rides are, on average, much more than casual ones since heavy users who account for most of the traffic can be expected to be almost always registered.



Statistics for Numeric Variables

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
temp	1	1500	0.50	0.19	0.50	0.50	0.24	0.04	0.96	0.92	-0.01	-0.97	0.01
atemp	2	1500	0.47	0.17	0.48	0.48	0.20	0.00	1.00	1.00	-0.09	-0.90	0.00
hum	3	1500	0.62	0.20	0.62	0.63	0.24	0.00	1.00	1.00	-0.13	-0.81	0.01
windspeed	4	1500	0.22	0.11	0.19	0.21	0.09	0.00	0.66	0.65	0.97	1.10	0.00
casual	5	1500	34.66	48.00	16.50	24.29	21.50	0.00	312.00	312.00	2.52	7.65	1.24
registered	6	1500	149.22	145.37	112.00	125.79	123.80	0.00	818.00	818.00	1.52	2.55	3.75
cnt	7	1500	183.88	174.68	140.50	157.70	159.38	1.00	905.00	904.00	1.28	1.44	4.51

Table 2: Statistics for the numeric variables

From the histograms below we can observe that the usual rentals are ranged between 0-300 rentals per hour. The temperature seems to be normally distributed. From the histograms of the total count of rides it becomes clear that a lot of the time no rental bike is active. This is more prominent amongst casual users, with registered users showing much more consistent activity.

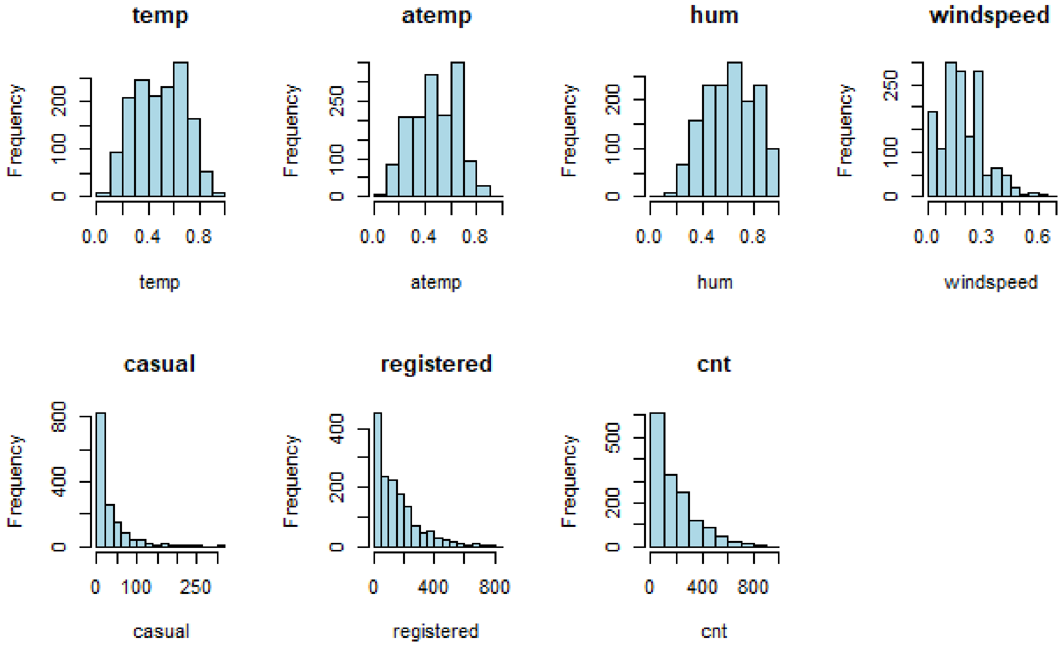


Table 3: Histograms for the numeric variables

Taking a closer look at the QQ- plots we have strong indications that all the variables are not distributed normally. Specifically, for the total count of rentals (cnt variable) high deviation is observed for the edge values.

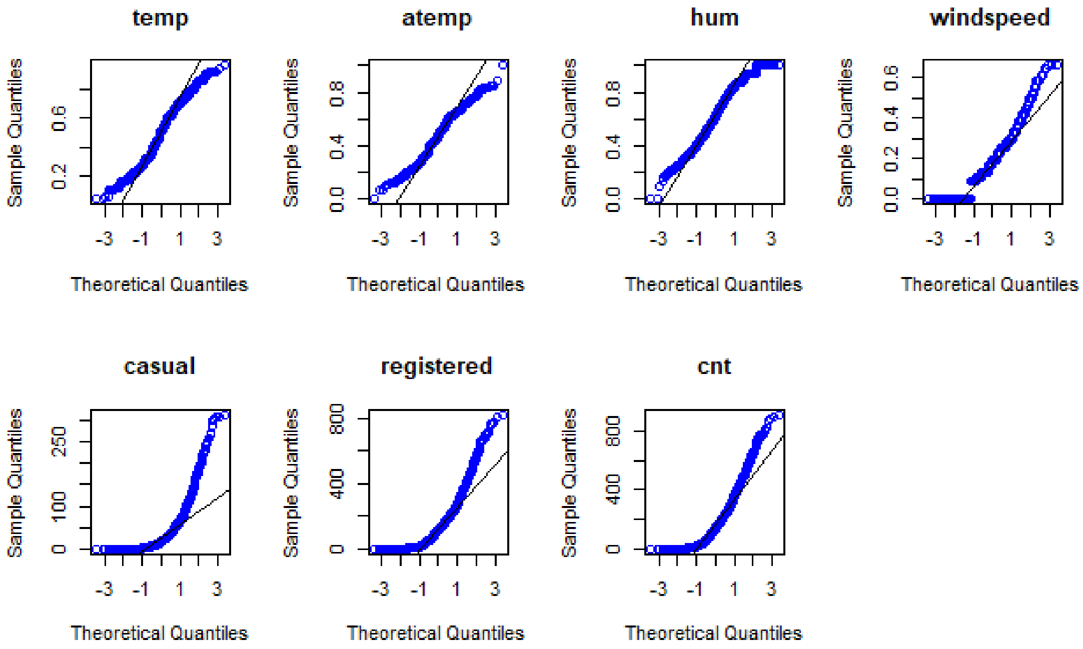


Table 4: Q-Q plots for the numeric variables

The boxplots for the temperature, feeling temperature and humidity seems to be symmetrical and therefore, we were waiting for the distribution to be normal. To test whether this hypothesis is true a normality test is conducted and based on the results it becomes clear that none of the above variables is normally distributed (*Shapiro Wilk test - p-value <0,05 – Appendix* 2.1). Finally, many positive outliers can be spotted in the boxplots corresponding to the variables of wind speed as well as of those of casual, registered, and total count of rides.

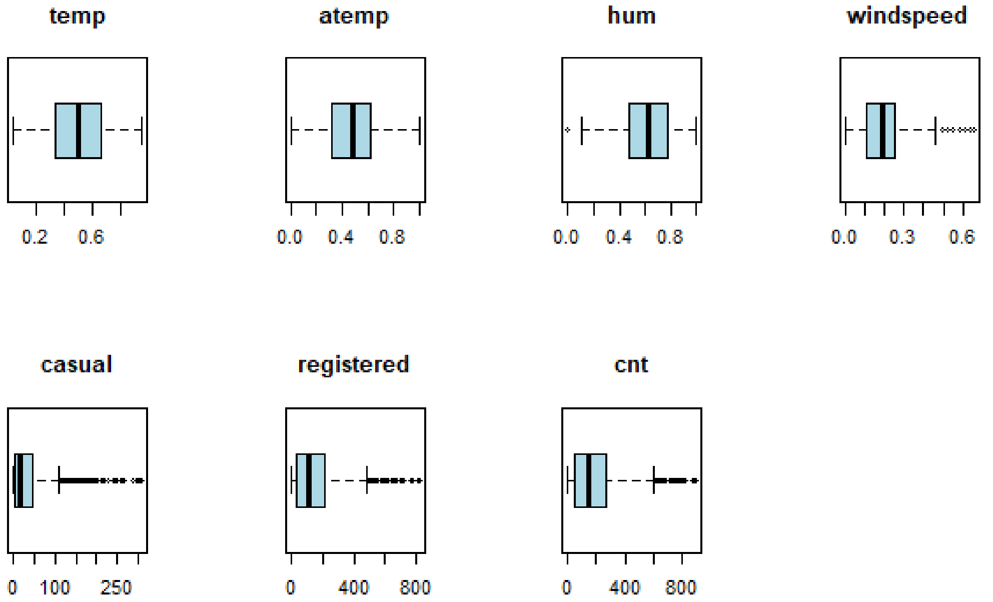


Table 5: Boxplots for the numeric variables

Categorical Variables – Descriptive Statistics

Almost 70% of the observations on the dataset comes from data recorded during a working day. On the other hand, the amount of data originating from holidays is rather limited. This preexisting imbalance in the data might limit the potential of the variable "holiday" to aid in the prediction of total rides per hour.

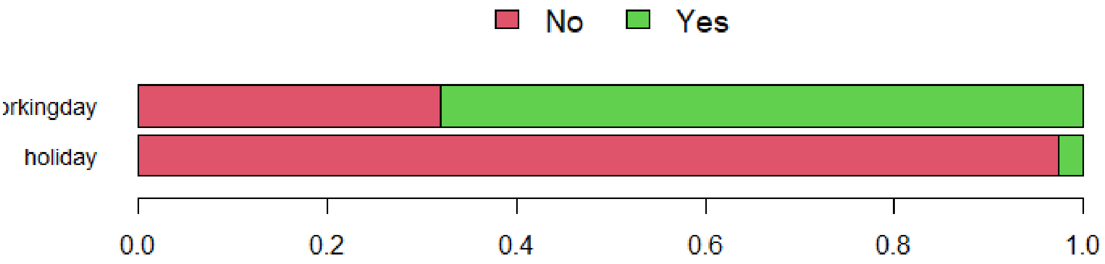


Table 6: Visualize data sufficient in working days and holidays

Moreover, by looking at the frequency table for the rest of the variables, one can understand that enough information is available for all features (e.g., seasons, hours, days) and that the total rides can be affected from the different levels. We should note here that for the weather condition we may not have enough information for the days when the weather is cloudy or rainy. This can be an indication that the different weather conditions may not be a good predictor for the total rental bikes.

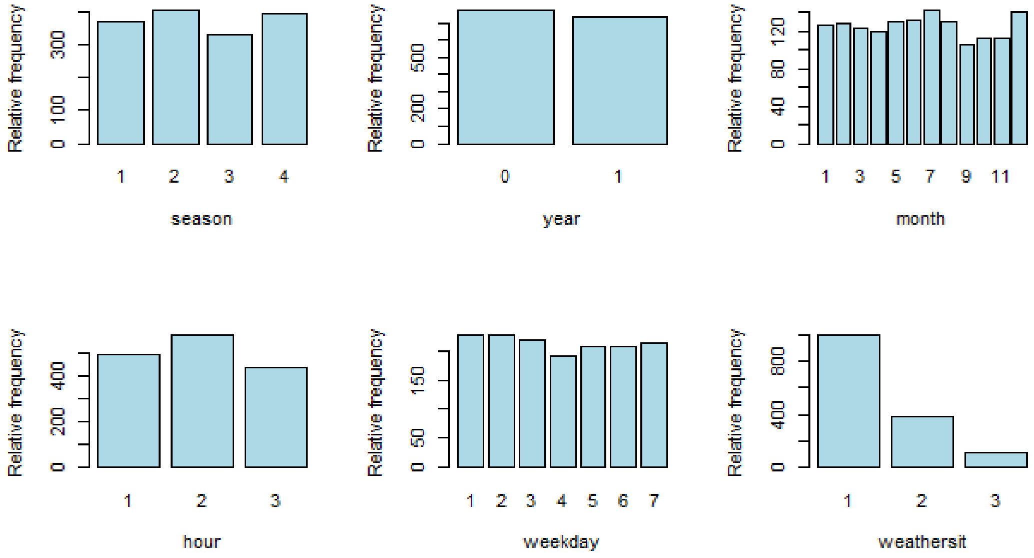
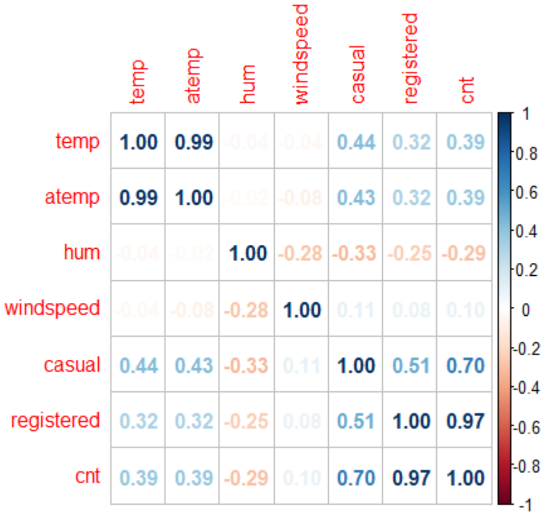


Table 7: Frequency tables for time related features

Pairwise comparisons

In this part the existing relationships among the independent (explanatory) variables will be examined, along with the relationship between the total rental rides (response variable) and the independent variables. In Table 8 (see below) the Pearson correlation between the variables is presented. The Pearson correlation is a linear correlation and as such it will not always capture with accuracy the relationship between variables related in non- linear fashion.



	temp	atemp	hum	windspeed	casual	registered	cnt
temp	1.00	0.99	-0.04	-0.04	0.44	0.32	0.39
atemp	0.99	1.00	-0.02	-0.08	0.43	0.32	0.39
hum	-0.04	-0.02	1.00	-0.28	-0.33	-0.25	-0.29
windspeed	-0.04	-0.08	-0.28	1.00	0.11	0.08	0.10
casual	0.44	0.43	-0.33	0.11	1.00	0.51	0.70
registered	0.32	0.32	-0.25	0.08	0.51	1.00	0.97
cnt	0.39	0.39	-0.29	0.10	0.70	0.97	1.00

Table 8: Corrplot presenting the pearson correlation between variables

Let’s now highlight some key observations based on the above table:

- The temperature and the feeling temperature are highly positively correlated **(0.99).** This is to be expected since these variables correspond to almost identical information.

- Total rental bikes and registered customers are highly positively correlated **(0.97)**. This is also to be expected given that the registered active riders are included when calculating the total active rental bikes. Something very similar is observed between the casual active users and the total count of active users, with a correlation value **(0.70).**

- Total rental bikes are positively correlated with the temperature. **(0.39)**. This means that when the temperature increases, more rentals tend to be recorded.

- Total rental bikes are negatively correlated with the humidity. **(-0.29)**. This means that when higher levels of humidity are present, the bike rentals are generally reduced.

Pairwise comparisons via scatter plots were conducted between the total number of rentals and the other numeric variables; in most of the cases, no linear relationship is observed. The only case in which a linear relationship is observed is between the number of registered active riders and total active riders which is expected as elaborated above. *(The scatter plots are presented in Appendix-2.2)*

Concerning the categorical variables, the following boxplots are constructed in order to examine the existing differences between the medians. Namely, we wanted to extract information for the behavioral characteristics in bikes rental per season, year, month, hour, holiday, weekday, working day and weather condition.

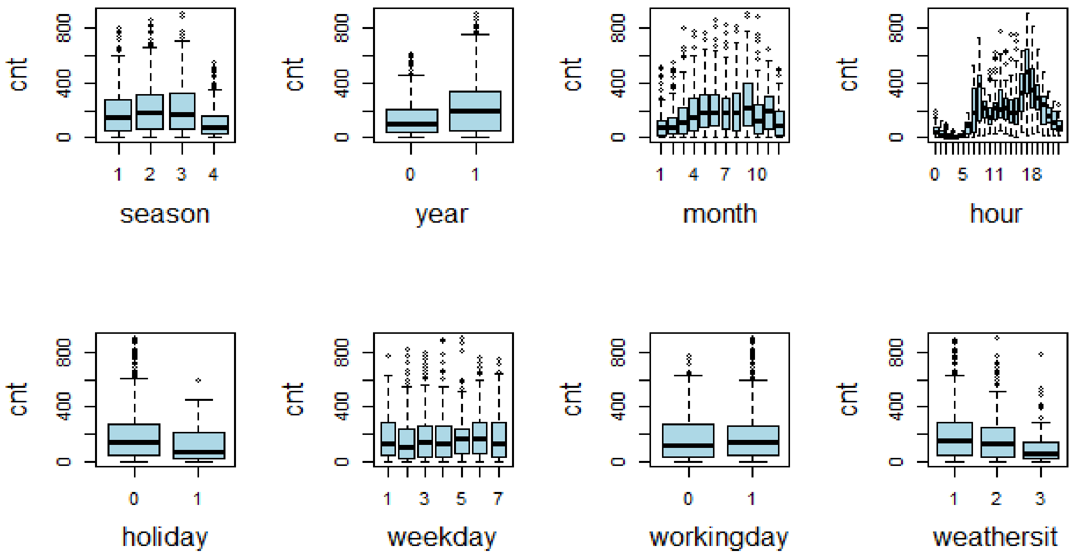


Table 9: The number of rentals per time period

During weekdays, holidays and working days,no major fluctuations are recorded in the count of total rental bikes. This might be an early indication that these variables will not be significantly important when building a predictive model, and they might end up being excluded. On the other hand, more significant fluctuations can be observed in the total number of rentals as one moves through different seasons, months, and hours. Also, an increase in the median number of total rentals is recorded when moving from 2011 compared to 2012. This seems to be plausible as the bike sharing system might be getting more popular as the years go by. Concerning the categorical variables, we observe discrepancies between different hours and therefore, it was deemed advantageous to divide the variable corresponding to hours into 3 categories. "Low Demand Hours", "Average demand hours" and "High demand hours". (*Appendix 3*)

A further step of the analysis is to examine whether there is association between any two of the categorical variables. (*Chi square test – Appendix 4*). (Chi square test – Appendix). Variables that are strongly correlated to each other can cause problems in fitting the model and interpreting the results. As a result, it is preferable for the variables of the dataset to be independent of each other to overcome the issue of multicollinearity. After the Chi tests were conducted, an association was observed between the pairs of variables given below

- Month and season

- Working day and Holiday
- Working day and weekday
- Season and weather sit

Predictive or Descriptive models

Taking into consideration all the above results, it is now possible to make an educated selection of the variables that should be part of the initial model. Concerning the numeric variables, the variables corresponding to casual and registered active users cannot be included in the model because they are directly correlated with what we want to predict. If those two variables were used, then the model will end up in a trivial state of using the sum of the two to predict the target (i.e. total count of active bikes).

Additionally, only the temperature variable is kept between temperature and feeling temperature to avoid introducing multicollinearity to the model.

Considering the above observations, the initial model should be the following:

Cnt ~ season, month, year, hour, holiday, weekday, working day, weathersit, temp, hum and windspeed

As discussed before, a big part of the explanatory variables in this dataset are categorical and it is often difficult to fit a linear model on such data. When the covariates are categorical, the corresponding variables are coded as dummy variables. In this approach, it is wise to use the Lasso method, which is a penalized linear regression. Lasso allows the analyst to estimate coefficients that minimize the RSS (sum of square errors). In lasso regression, a value for λ is selected that produces the lowest possible test MSE (mean squared error). (For the selected λ for this dataset see Appendix).

Lasso also has a variable selection property that can deal with the multicollinearity in the data. This property ensures that only the variables who are not collinear will be selected included in the model. Due to this property, Lasso is a powerful method for factor data analysis, as it takes care of both the selection of the features and the estimation of the model coefficients..

All that said, the variable selection property of Lasso creates the problem of partial selection of dummy variables. A categorical variable with n levels will be included as n different dummy variable in the linear model. This makes it possible for the Lasso selection process to select only some of the dummy variables derived from one categorical variable. Specifically, utilizing the lasso methodology for the given dataset it becomes clear that not all the seasons are statistically significant features of the model. Here, we need to decide either to use seasonality as an explanatory variable in our model or not. The same problem was encountered when inspecting the month variable. However, only 3 months are significant and therefore, that variable can be more easily excluded from the model. Similarly, only one level of the variable corresponding to the weather (weathersit) is important based on the Lasso elimination. As mentioned in the previous chapter, season and weathersit are associated and therefore, we decided to remove the weathersit variable.

The model determined as optimal after using lasso is the following:

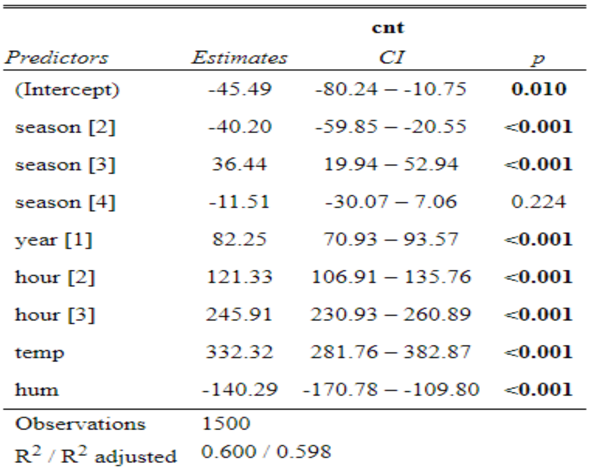
$$\boldsymbol{Cnt = -13.14 + 30.097 * season3 - 3.1439 * season4 + 67.8442 * year1 - 1.7959 * month7 + 3.884525 * month9 + 99.961 * hour2 + 221.22 * hour3 - 26.8244 * weathersit3 + 247.64 * atemp - 104.47 * hum + \varepsilon}$$

Having in mind the above convention about seasons, months and weathersit we are going to implement the stepwise method (AIC) in the following model:

Cnt ~ season, year, hour, temp, hum

The aim of this analysis is to predict the total count of rental bikes and therefore, the AIC methodology is recommended. (*Appendix 5B*). The key difference between lasso and stepwise regression is that the latter shrinks all coefficients towards zero. while the former has the potential to remove predictors from the model by shrinking the coefficients completely to zero. When implemented, the stepwise method did not exclude another variable from the model. Hence, the final model remains as specified using Lasso and is the following:

:



	cnt		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-45.49	-80.24 – -10.75	**0.010**
season [2]	-40.20	-59.85 – -20.55	**<0.001**
season [3]	36.44	19.94 – 52.94	**<0.001**
season [4]	-11.51	-30.07 – 7.06	0.224
year [1]	82.25	70.93 – 93.57	**<0.001**
hour [2]	121.33	106.91 – 135.76	**<0.001**
hour [3]	245.91	230.93 – 260.89	**<0.001**
temp	332.32	281.76 – 382.87	**<0.001**
hum	-140.29	-170.78 – -109.80	**<0.001**
Observations	1500		
R^2 / R^2 adjusted	0.600 / 0.598		

Table 10: Final model summary after Lasso and AIC methods

$\boldsymbol{Cnt = -45,49 - 40,20 * season2 + 36.44 * season3 - 11.51 * season4 + 82.25 * year1 + 121,33 * hour2 + 245.91 * hour3 + 332.32 * temp - 140.29 * hum +}$ **ε, ε~N(0, (110,4)2)**

This linear model fits quite well with the data as indicated by the value of the R^2 adj = 0,598. We are taking into consideration the R^2 adj and not the R^2 because the R^2 adj is not affected by the number of predictors (independent variables) in the model. The adjusted R^2 value of 59,8% for this regression implies that the independent variables explain 59,8% of the variation in the dependent variable (cnt).

Examine the assumptions of the final model

The final step of this analysis is to examine if the statistical assumptions of the final model are satisfied. For the analysis to be fully valid, first, the residuals should be distributed normally with constant variance. Second, the relation between the dependent and the independent variables should be linear. Third, since we are dealing with time series data, the independence of the residuals needs to be ensured.

Given the graphs below, from the three statistical assumptions of the final model, only the third one is satisfied. (*Tests* – *Appendix 6*)

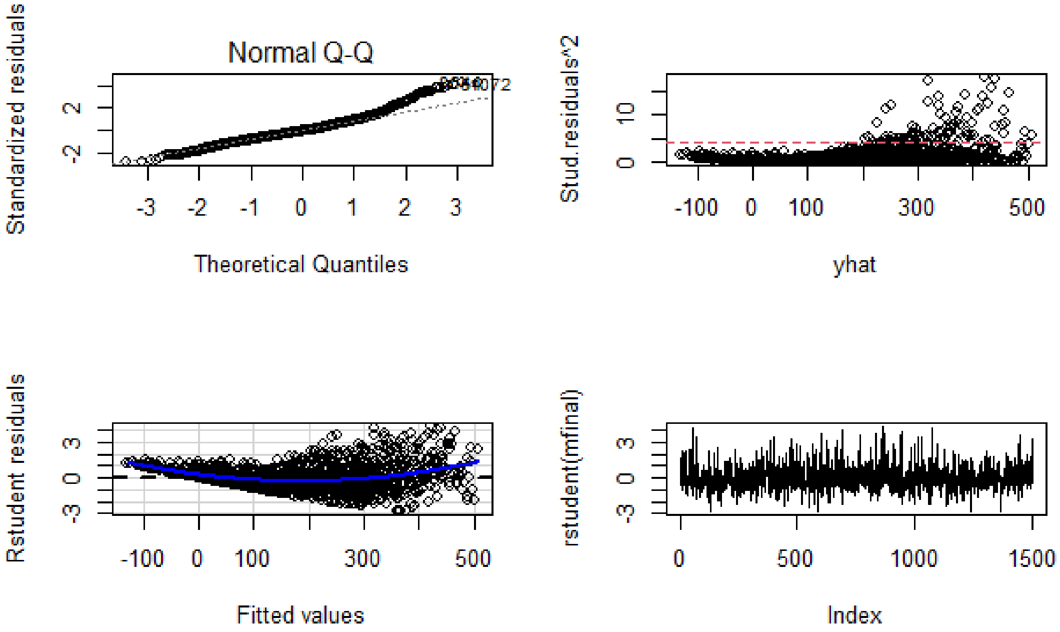
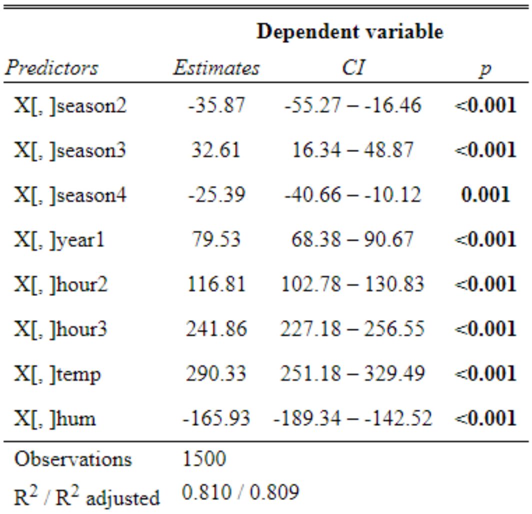


Table 11: Normality plot, Constant variance plot, Linearity plot and Independence plot

We tried to mediate the violation of the first and second assumption using a variety of transformations on the target variable corresponding to the total count of rental bikes (e.g., logarithm or square root) but no useful result were achieved on any of the two assumptions. Logarithm fixes the normality and the constant variance but not the linearity. On the other hand, the square root fixes the linearity, but not the other assumptions. (*Appendix 7*)

Interpretation of the final model

A model that is deemed optimal was created with the use of the AIC methodology. It is now time to try and interpret the parameters of that model. First, to make that interpretation simpler, the intercept was excluded. (*Appendix 8*)



	Dependent variable		
Predictors	*Estimates*	*CI*	*p*
X[,]season2	-35.87	-55.27 – -16.46	**<0.001**
X[,]season3	32.61	16.34 – 48.87	**<0.001**
X[,]season4	-25.39	-40.66 – -10.12	**0.001**
X[,]year1	79.53	68.38 – 90.67	**<0.001**
X[,]hour2	116.81	102.78 – 130.83	**<0.001**
X[,]hour3	241.86	227.18 – 256.55	**<0.001**
X[,]temp	290.33	251.18 – 329.49	**<0.001**
X[,]hum	-165.93	-189.34 – -142.52	**<0.001**
Observations	1500		
R^2 / R^2 adjusted	0.810 / 0.809		

Table 12: The summary table of the final model when we removed the intercept

$$Cnt = -35{,}87 * season2 + 32.61 * season3 - 25.39 * season4 + 79.53 * year1 + 116{,}81 * hour2 + 241.86 * hour3 + 290.33 * temp - 165.93 * hum + \varepsilon$$

We excluded the intercept, but the weighted coefficient was added as weighted coefficient to the remaining variables.

One can interpret the coefficients of this model as follows:

- If we compare two counts of active rentals recorded under conditions with the same characteristics and differ solely in the year during which they are recorded, then the expected difference in the number of rentals will be almost 79 rentals more for the year 2012.
- If we compare two counts of active rentals recorded under conditions with the same characteristics and differ solely in the hour during which they were recorded, then the expected difference in the number of rentals will be almost 116 rentals higher for the average demand hour compared to the low demand hour and 241 rentals more for the high demand hours.
- If we compare two counts numbers of rentals active rentals recorded under conditions with the same characteristics and which differ by exactly one degree of temperature the number of rentals will be almost 7 rentals higher for the higher temperature recording.
- If we compare two counts numbers of rentals active rentals recorded under conditions with the same characteristics and which differ by exactly 1 degree of humidity, then the expected difference in the number of rentals will be almost 1,65 rentals more for the recording made in

lower humidity. If we compare two number of rentals with the same characteristics which differ only by 1 degree humidity, then the expected difference in the number of rentals will be almost **1,65 rentals**[1] in favor of the lower humidity.

Performance of the final model

The error in the predicted value with the calculated covariates is 110,4 which means that the error in the estimated count will be 2x110,4=220,8 rentals around the expected/fitted value (extrapolation)

It seems that the performance of the model (R2) is ameliorated. However, that's not the case as we calculated the R2 once more and we found that the true R2 is still **59.8%.**

Further analysis

The out of sample predictive ability of the final model

Next step to our analysis is to assess the out of sample predictive ability of our final model. Evaluating the model accuracy is an essential part of the process in creating models to describe how well the model is performing in its predictions. The errors is an indicator of how much the model is making mistakes in the prediction. The values for predicted variables are computed from the predictors in the test part of the sample and then compared to their true values in the test part of the sample. Consequently, estimating the mean square error of prediction using metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) and calculating the R2 can be basic indicators to evaluate the model accuracy.

To test if the model we created can predict and interpret another set of similar observations we used a smaller dataset of 500 different observations (test dataset) . The test dataset has been cleaned in the same way as our main dataset.

We calculated the predictive values for the count of rental bikes using the new observations based on the linear model we have created. In other words, we inserted the values of the observations of the test dataset in the following equation:

Cnt = −35,87*season2 +32.61*season3 −25.39*season4+79.53*year1+116,81*hour2 + 241.86*hour3 + 290.33*temp −165.93*hum + ε, ε~N(0, (110,4)2)

[1] The result came from the division of 165 with 100 because the humidity has been normalized.

Then, we found the difference between the predicted count of rentals and the actual one. This calculation helped us to calculate the RMSE and we find out that RMSE is **116,29,** meaning that the error in the estimated count will be 2x116,3=232,6 rentals around the expected/fitted value.

Conserning the R2, namely the percentage of the variation in the count of rentals that is explained by the independent parameters of the model (e.g., temperature, hum) is **59,4%.** That may not be a perfect fit but it is a worthwhile one thinking that we have a small size sample for data that are changing every second.

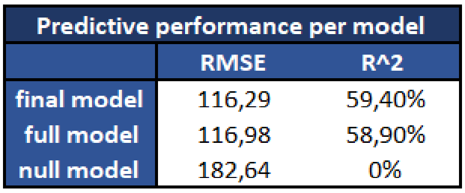
Comparison of the full, null, and final model

We described in the chapter above the accuracy of our final model. Then we thought about looking at whether our data can be described better using the full model or the null model. (*Appendix 9)*

The full model contains all the parameters we have on the dataset (season, month, year, hour, weekday, workingday, holiday, weathersit, temp, atemp, hum, windspeed). Following the same process as previously we concluded that the error in the estimated count of rentals in the full model is 2*116,98 = … (RMSE =**116,98**) and the percentage of the explained variation in the count of rentals is **58,9%**.

We can see that although we added more explanatory variables,

Finally, we examined the performance of the null model. The null model does not contain any explanatory variable. In this way the calculated RMSE = **182,64** meaning that the error in estimated count of rentals is 2*182,64 = …. and the R2 = **0%** which is reasonable because we have no explanatory variables in this model either the standard error or the R2 Adjusted have improved in a significant level.



Predictive performance per model		
	RMSE	R^2
final model	116,29	59,40%
full model	116,98	58,90%
null model	182,64	0%

Table 13: Predictive performance per model

From the table above, we can conclude that the model with the best performance is the final model, which predicts the total count of rentals based on the season, the year, the hour, the temperature and the humidity. There are not big differences between the full model and the final model. However, between the final model and the full model it is reasonable to select the final model because it gives someone the opportunity to predict the count of the total bikes rentals interpreting less parameters.

We can anticipate high errors in this dataset basically for two reasons. The data are not up to date, as they are concerning years 2011 and 2012. Secondly, we trained our model with a non sufficient sample of observations which contains multiple outliers.

Discussion

Regression modeling proved to be a valuable method when attempting to predict the number of total users for a given hour/day etc.. A good prediction can help managers manage their time, total users and investment. This paper mainly focuses on using lasso and stepwise models to predict the bike sharing total rentals. In this paper, the main features include season, year, hour, temperature and humidity. These features were chosen amongst many and identified as optimal by the Lasso method and subsequently the AIC methodology. By analyzing the average RMSEs of the final model, the full model and the null model, the AIC regression model has been shown to achieve the smallest errors in the test data and the best overall performance. This method for predicting the number of total users in bike sharing has shown clear potential and could become a tool for management teams to refer to when makind decisions.

In order to understand the utility of this model, we will present below an example. We are going to describe a typical profile of a day for each season. This can be very helpful for industries and bike sharing companies as they can predict the average demand for each season and take the appropriate actions.

First the average differences in temperature and humidity between the seasons are given.

As mentioned in earlier sections, the levels for the seasons variable are the following: 1: springer, 2:summer, 3:fall, 4:winter. Consequently, from the graph below we can extract the information:

Average Weather conditions per Season				
	Temperature (normalized)	Temperature (Actual)	Humidity (normalized)	Humidity (Actual)
Spring	0.5	20.5	0.6	60
Summer	0.7	28.7	0.6	60
Fall	0.5	20.5	0.7	70
Winter	0.3	12.3	0.6	60

Table 14: Average weather conditions per Season

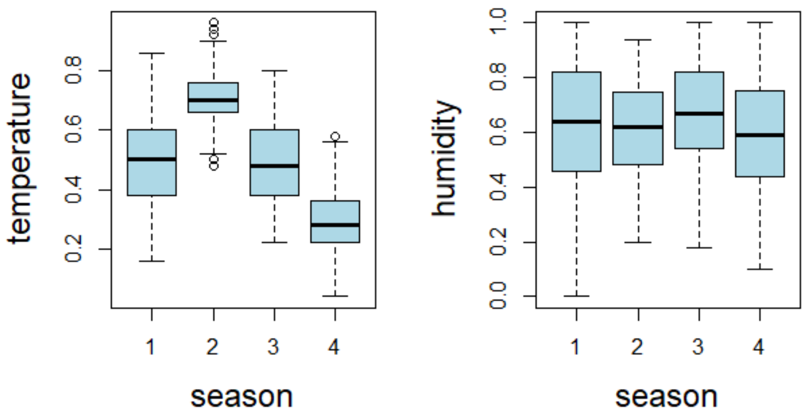


Table 15: Median differences in Temperature and Humidity per season

To add to the above, we examined if there are significant differences between the average temperature and humidity for the two years. As seen in the box plots below, no such differences are observed

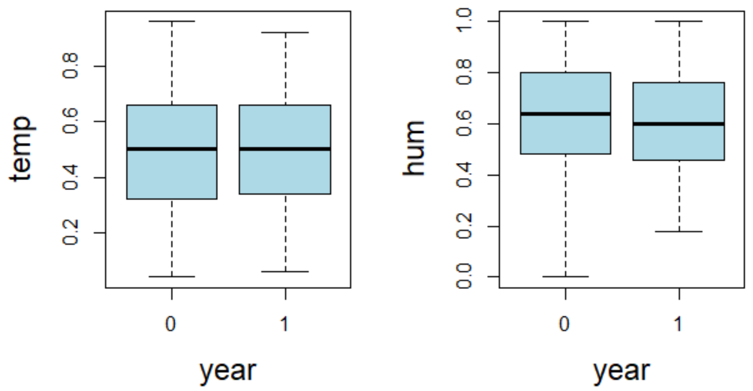
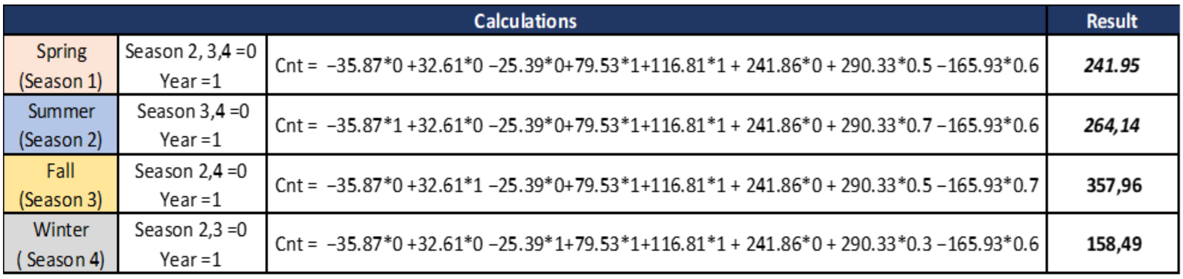


Table 16: Median differences in temperature and humidity per year

Now, we are ready to calculate the expected total count of rental bikes for each season.

A typical day in 2012 would have hour =2 since that’s the average demand hour and the values of temp and humidity would be as shown in the above table.

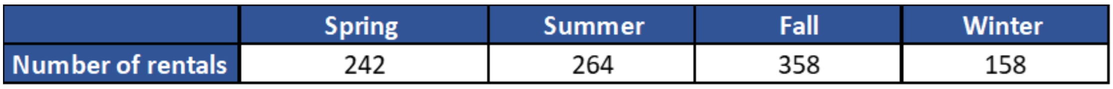
Cnt = −35,87*season2 +32.61*season3 −25.39*season4+79.53*year1+116,81*hour2 + 241.86*hour3 + 290.33*temp −165.93*hum



Calculations			Result
Spring (Season 1)	Season 2, 3,4 =0 Year =1	Cnt = −35.87*0 +32.61*0 −25.39*0+79.53*1+116.81*1 + 241.86*0 + 290.33*0.5 −165.93*0.6	***241.95***
Summer (Season 2)	Season 3,4 =0 Year =1	Cnt = −35.87*1 +32.61*0 −25.39*0+79.53*1+116.81*1 + 241.86*0 + 290.33*0.7 −165.93*0.6	***264,14***
Fall (Season 3)	Season 2,4 =0 Year =1	Cnt = −35.87*0 +32.61*1 −25.39*0+79.53*1+116.81*1 + 241.86*0 + 290.33*0.5 −165.93*0.7	**357,96**
Winter (Season 4)	Season 2,3 =0 Year =1	Cnt = −35.87*0 +32.61*0 −25.39*1+79.53*1+116.81*1 + 241.86*0 + 290.33*0.3 −165.93*0.6	**158,49**

Table 17: Total count of rentals of a typical day for each season (calculations)

We can see the average results in the summary table below.



	Spring	Summer	Fall	Winter
Number of rentals	242	264	358	158

Table 18: Summary results of total number of rentals per season

The season with the greatest demand for bicycles is Fall. One could expect this since it is the period when people return from their summer holidays, and it is also the beginning of the school season.

In contrast, the season with the least demand for bicycles is winter. This makes sense since due to the lower temperatures and the more intense weather phenomena, people may prefer other ways of transportation.

We can also detect a small increase in total rentals between spring and summer which can be explained due to the better weather and to the tourist period.

Conclusion

The model proposed has a relatively good fit and is easy to understand and interpret. The model, however, has a big error and so it would just be useful to see what prices affected the number of rentals for that period.

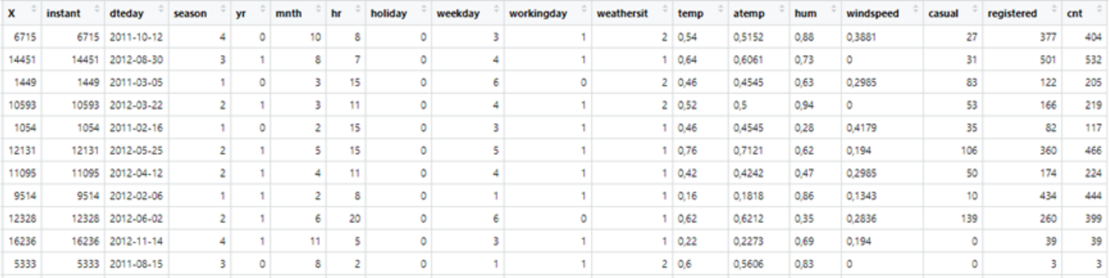
One of the major factors that greatly affects the demand for bicycles is the temperature. This makes sense since people tend to go out more as the temperature rises and - for higher temperatures - bike use can be recreational and not only as a means to commute to work or to other everyday obligations.

In future analysis, it would be useful to include an additional variable that would indicate whether the rental station is located in close proximity to a metro station or not. This would allow researchers to gain insight on the use of sharing bikes as bridges to reach metro stations. Such insight could prove to be highly valuable for businesses looking to expand the bike services and build stations in other regions. Such businesses can profit from identifying the most profitable locations to build stations to maximize reach and demand.

Appendix

1. Data Cleansing

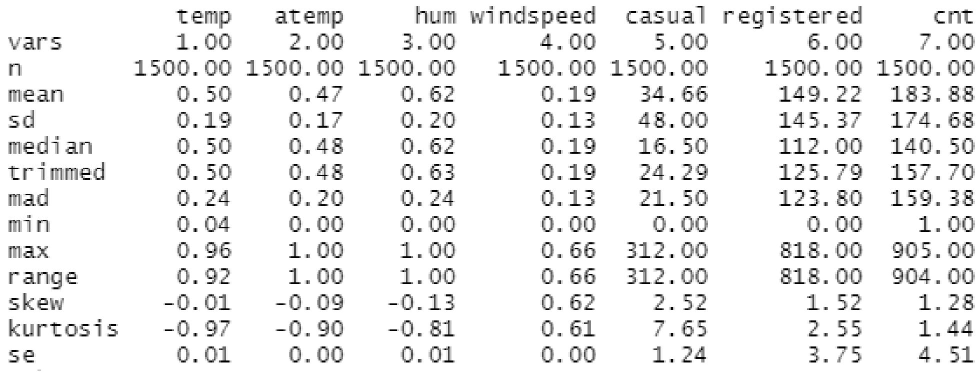
The initial dataset has the following format. As we can see the column “X” is identical with the column “instant” and therefore, we will drop the first one.



X	instant	dteday	season	yr	mnth	hr	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
6715	6715	2011-10-12	4	0	10	8	0	3	1	2	0,54	0,5152	0,88	0,3881	27	377	404
14451	14451	2012-08-30	3	1	8	7	0	4	1	1	0,64	0,6061	0,73	0	31	501	532
1449	1449	2011-03-05	1	0	3	15	0	6	0	2	0,46	0,4545	0,63	0,2985	83	122	205
10593	10593	2012-03-22	2	1	3	11	0	4	1	2	0,52	0,5	0,94	0	53	166	219
1054	1054	2011-02-16	1	0	2	15	0	3	1	1	0,46	0,4545	0,28	0,4179	35	82	117
12131	12131	2012-05-25	2	1	5	15	0	5	1	1	0,76	0,7121	0,62	0,194	106	360	466
11095	11095	2012-04-12	2	1	4	11	0	4	1	1	0,42	0,4242	0,47	0,2985	50	174	224
9514	9514	2012-02-06	1	1	2	8	0	1	1	1	0,16	0,1818	0,86	0,1343	10	434	444
12328	12328	2012-06-02	2	1	6	20	0	6	0	1	0,62	0,6212	0,35	0,2836	139	260	399
16236	16236	2012-11-14	4	1	11	5	0	3	1	1	0,22	0,2273	0,69	0,194	0	39	39
5333	5333	2011-08-15	3	0	8	2	0	1	1	2	0,6	0,5606	0,83	0	0	3	3

Figure 1: The initial dataset

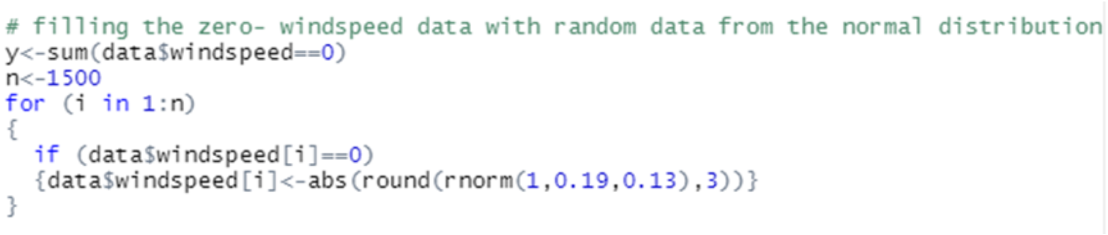
Another thing that we observed is that the “windspeed” has multiple “zero” values. This may be due to missing values that R converted them into “zero”. For this reason and with the knowledge that the values of variable “windspeed” are normalized we filled the zeros with random values from the normal distribution with sd= 0,13 and mean = 0,19 as we can extract from the describe table.



```
             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         0.50    0.47    0.62      0.19   34.66     149.22  183.88
sd           0.19    0.17    0.20      0.13   48.00     145.37  174.68
median       0.50    0.48    0.62      0.19   16.50     112.00  140.50
trimmed      0.50    0.48    0.63      0.19   24.29     125.79  157.70
mad          0.24    0.20    0.24      0.13   21.50     123.80  159.38
min          0.04    0.00    0.00      0.00    0.00       0.00    1.00
max          0.96    1.00    1.00      0.66  312.00     818.00  905.00
range        0.92    1.00    1.00      0.66  312.00     818.00  904.00
skew        -0.01   -0.09   -0.13      0.62    2.52       1.52    1.28
kurtosis    -0.97   -0.90   -0.81      0.61    7.65       2.55    1.44
se           0.01    0.00    0.01      0.00    1.24       3.75    4.51
```

Figure 2: Variable statistics

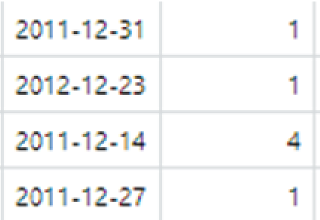
The code we used is the following:



```
# filling the zero- windspeed data with random data from the normal distribution
y<-sum(data$windspeed==0)
n<-1500
for (i in 1:n)
{
  if (data$windspeed[i]==0)
  {data$windspeed[i]<-abs(round(rnorm(1,0.19,0.13),3))}
}
```

Figure 3: Variable windspeed data cleansing code

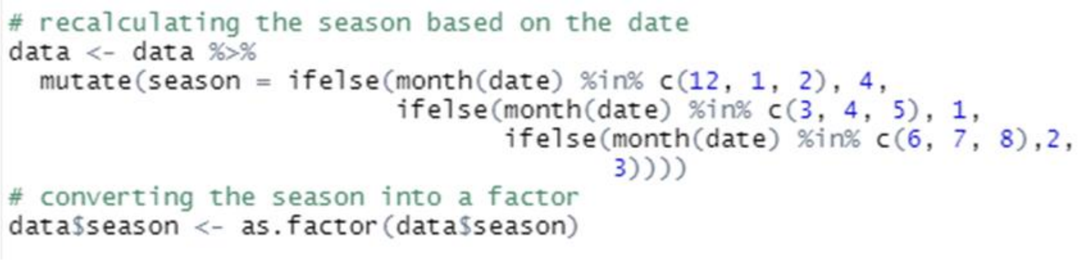
One more problem observed in this dataset is that there is no reconciliation between the date and seasons. We can see an example in the output above.



2011-12-31	1
2012-12-23	1
2011-12-14	4
2011-12-27	1

Figure 4: Season differences in data

For this reason, we recalculated the column season based on the date day column using the following R code.

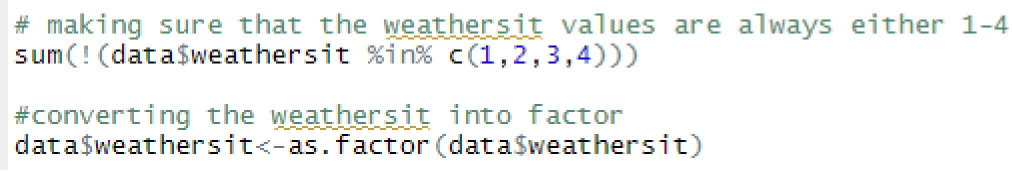


```
# recalculating the season based on the date
data <- data %>%
  mutate(season = ifelse(month(date) %in% c(12, 1, 2), 4,
                         ifelse(month(date) %in% c(3, 4, 5), 1,
                                ifelse(month(date) %in% c(6, 7, 8),2,
                                       3))))
# converting the season into a factor
data$season <- as.factor(data$season)
```

Figure 5: Recalculating season code

The same process we followed and for the variables “year” and “month”, just to be sure that everything is right.

Finally we checked that all the variables contain the appropriate values. For example, that the weathersit variable contains only “1,2,3 or 4” indices based on the corresponding weather condition.



```
# making sure that the weathersit values are always either 1-4
sum(!(data$weathersit %in% c(1,2,3,4)))

#converting the weathersit into factor
data$weathersit<-as.factor(data$weathersit)
```

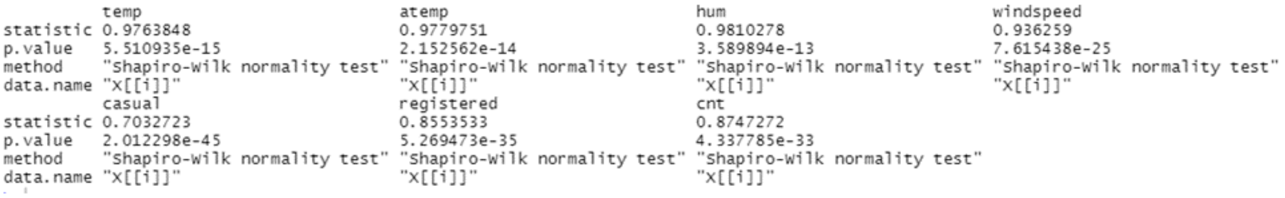
Figure 6: Checking for wrong inputs in weathersit

]

2. Descriptive Analysis

2.1 Normality test (Shapiro Wilk) for the numeric variables

Below we can see the outputs of the normality test Shapiro Wilk testing the numeric variables. For all test p-value <0,05 and therefore, we have strong indications that the variables are not normally distributed.



```
          temp                          atemp                         hum                           windspeed
statistic 0.9763848                     0.9779751                     0.9810278                     0.936259
p.value   5.510935e-15                  2.152562e-14                  3.589894e-13                  7.615438e-25
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test" "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "X[[i]]"                      "X[[i]]"                      "X[[i]]"                      "X[[i]]"
          casual                        registered                    cnt
statistic 0.7032723                     0.8553533                     0.8747272
p.value   2.012298e-45                  5.269473e-35                  4.337785e-33
method    "Shapiro-Wilk normality test" "Shapiro-Wilk normality test" "Shapiro-Wilk normality test"
data.name "X[[i]]"                      "X[[i]]"                      "X[[i]]"
```

Figure 7: Shapiro Wilk tests of normality

2.2 Pairwise comparisons

Additionally, with the corrplot we saw in the main report there is the scatter plots .

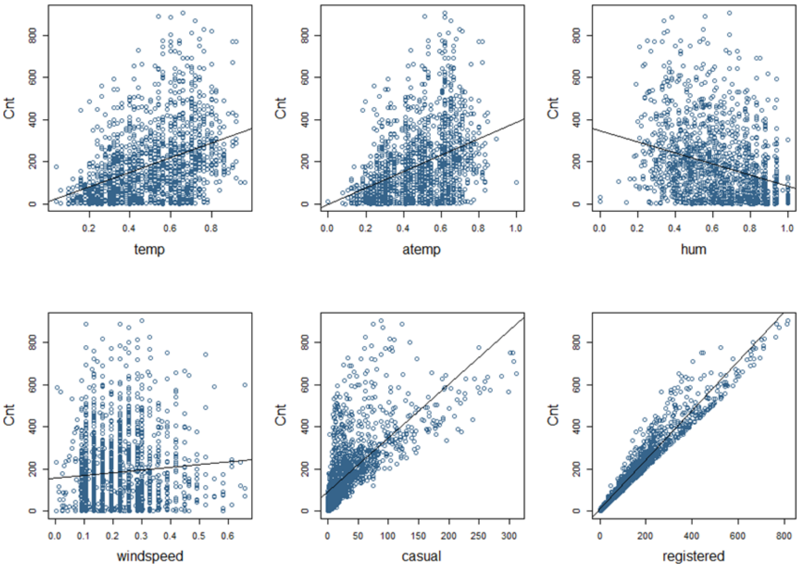


Figure 8: Scatter plots of cnt and numeric variables

3. Split the “hour” variable into categories “Low demand hours”, “Average demand hours” and “High demand hours”

To categorize the hour variable we extracted the relevant information from the following boxplot:

As we can see the high demand hours are between 7-8 in the morning and 4-8 in the afternoon. This can be reasonable if we think that on these hours many people go to their working place or they are returning from it. Also, many students may rent a bike to go to school in the morning. The low hour demand refers to late night hours between 11pm - 6am which was expected as the bars and the cafes are closed and people usually sleep during these hours.

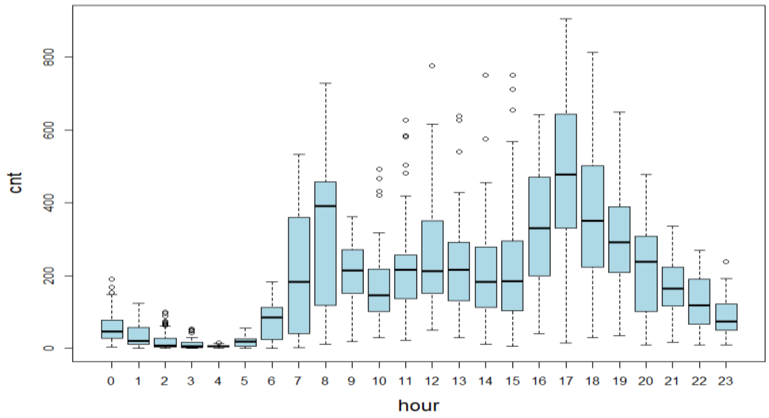
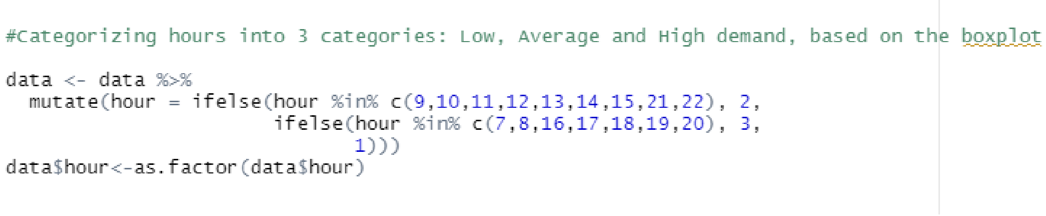


Figure 9: Cnt per hour

The code we used is the following:



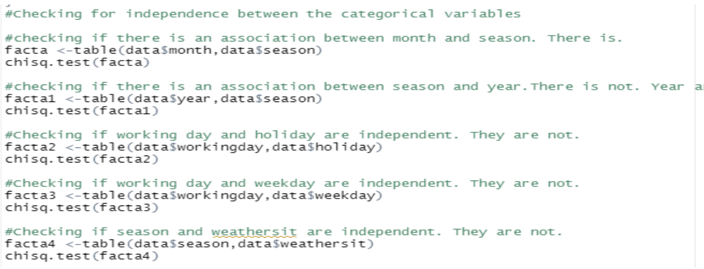
```
#Categorizing hours into 3 categories: Low, Average and High demand, based on the boxplot

data <- data %>%
  mutate(hour = ifelse(hour %in% c(9,10,11,12,13,14,15,21,22), 2,
                       ifelse(hour %in% c(7,8,16,17,18,19,20), 3,
                              1)))
data$hour<-as.factor(data$hour)
```

Figure 10: Dividing hours into categories - code

4. Checking if there is association between categorical variables

We conducted the chi square test to examine whether the categorical variables are independent. The pairs we have examined is the following:

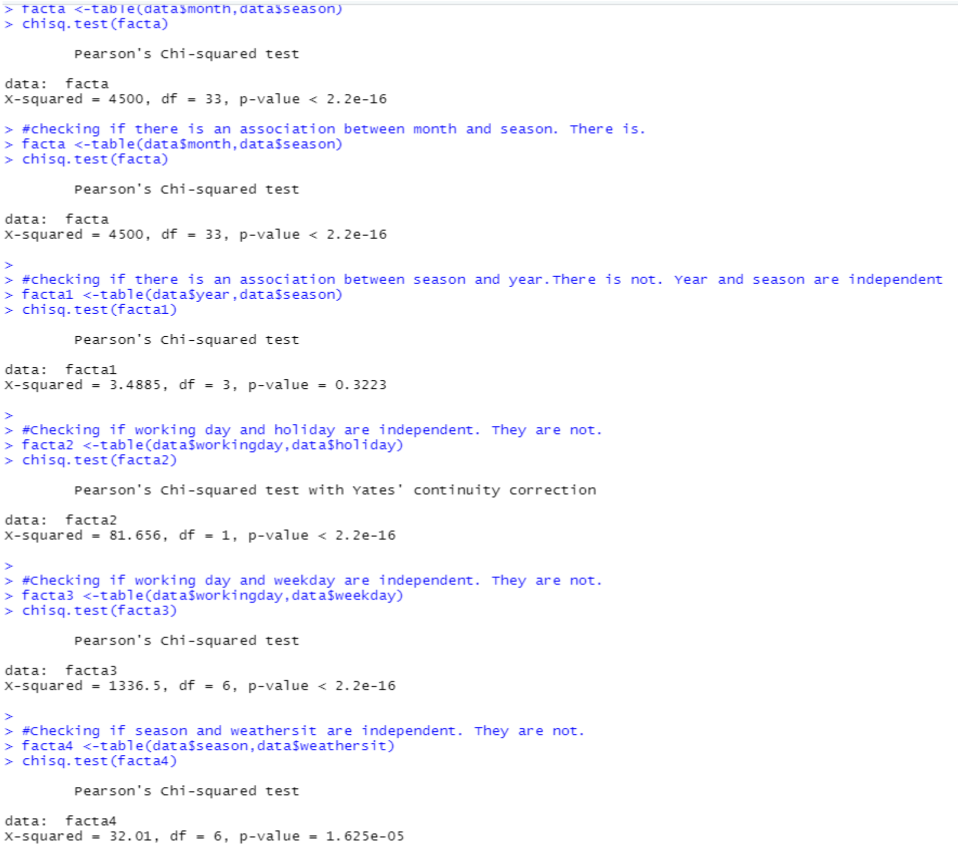
- Month with season
- Year with season
- working day with holiday
- working day with weekday
- season with weathersit



```
#Checking for independence between the categorical variables

#checking if there is an association between month and season. There is.
facta <-table(data$month,data$season)
chisq.test(facta)

#checking if there is an association between season and year.There is not. Year and season are independent
facta1 <-table(data$year,data$season)
chisq.test(facta1)

#Checking if working day and holiday are independent. They are not.
facta2 <-table(data$workingday,data$holiday)
chisq.test(facta2)

#Checking if working day and weekday are independent. They are not.
facta3 <-table(data$workingday,data$weekday)
chisq.test(facta3)

#Checking if season and weathersit are independent. They are not.
facta4 <-table(data$season,data$weathersit)
chisq.test(facta4)
```

Figure 11: Chi square tests for association

The results that the chi square test produced were the following:



```
> facta <-table(data$month,data$season)
> chisq.test(facta)

        Pearson's Chi-squared test

data:  facta
X-squared = 4500, df = 33, p-value < 2.2e-16

> #checking if there is an association between month and season. There is.
> facta <-table(data$month,data$season)
> chisq.test(facta)

        Pearson's Chi-squared test

data:  facta
X-squared = 4500, df = 33, p-value < 2.2e-16

> 
> #checking if there is an association between season and year.There is not. Year and season are independent
> facta1 <-table(data$year,data$season)
> chisq.test(facta1)

        Pearson's Chi-squared test

data:  facta1
X-squared = 3.4885, df = 3, p-value = 0.3223

> 
> #Checking if working day and holiday are independent. They are not.
> facta2 <-table(data$workingday,data$holiday)
> chisq.test(facta2)

        Pearson's Chi-squared test with Yates' continuity correction

data:  facta2
X-squared = 81.656, df = 1, p-value < 2.2e-16

> 
> #Checking if working day and weekday are independent. They are not.
> facta3 <-table(data$workingday,data$weekday)
> chisq.test(facta3)

        Pearson's Chi-squared test

data:  facta3
X-squared = 1336.5, df = 6, p-value < 2.2e-16

> 
> #Checking if season and weathersit are independent. They are not.
> facta4 <-table(data$season,data$weathersit)
> chisq.test(facta4)

        Pearson's Chi-squared test

data:  facta4
X-squared = 32.01, df = 6, p-value = 1.625e-05
```

Figure 12: Outputs of chi Square tests

From the output above we can understand that only year and season are independent variables, meaning that the existence of one does not influence the result of the other. (p-value =0,3223 > 0,05).

We have indications that the pairs: month-season, working day- holiday, working day-weekday and season- weathersit are associated. (p-value <0,05)

5. Predictive models

A. Lasso methodology

The Lasso technique has a goal to achieve the minimum variance and for the prediction to be unbiased.

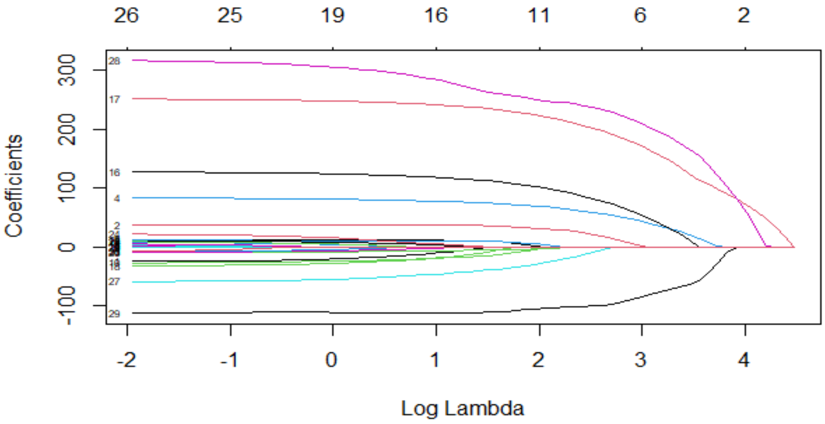


Figure 13: Lasso output

In the graph above we can see the order in which variables stop being significant.

We want to find a lambda that minimizes the MSE (min square error). To achieve that we are going to use the cross-validation method. We found the following two lambdas, lambda min and lamba1se. The lines in the graph below show us the two lambdas as a log. The left line is the lambda.min which achieves the minimum MSE. Accordingly, the right line corresponds to the Lambda 1se which is the largest value of lambda such that error is within 1 standard error of the minimum.

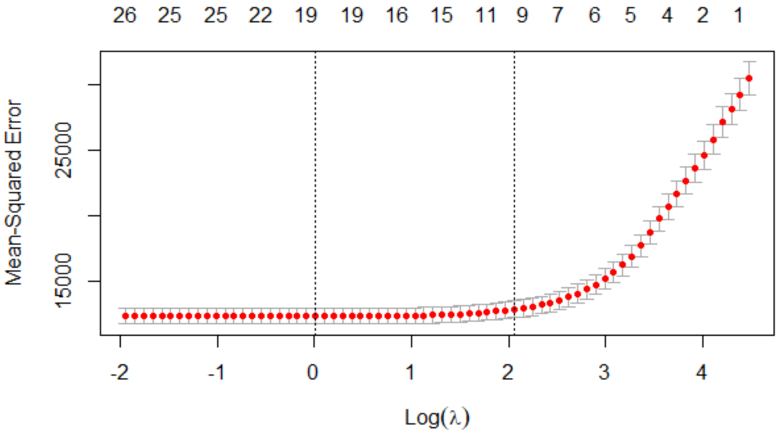


Figure 14: Lambda selection

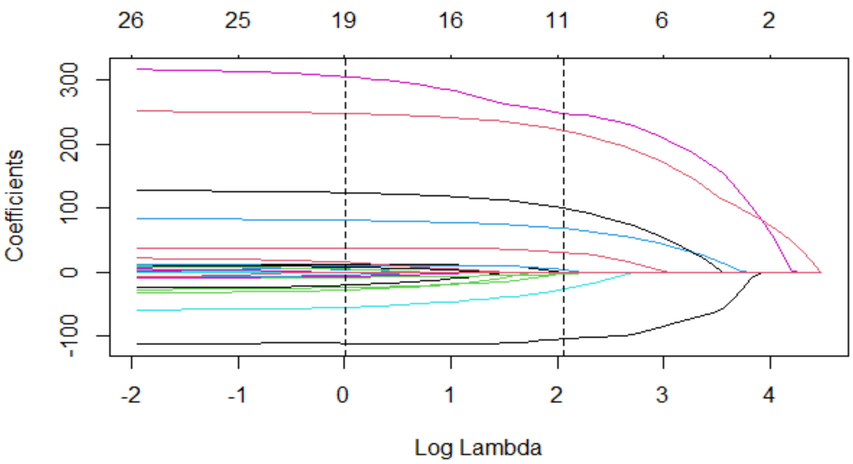
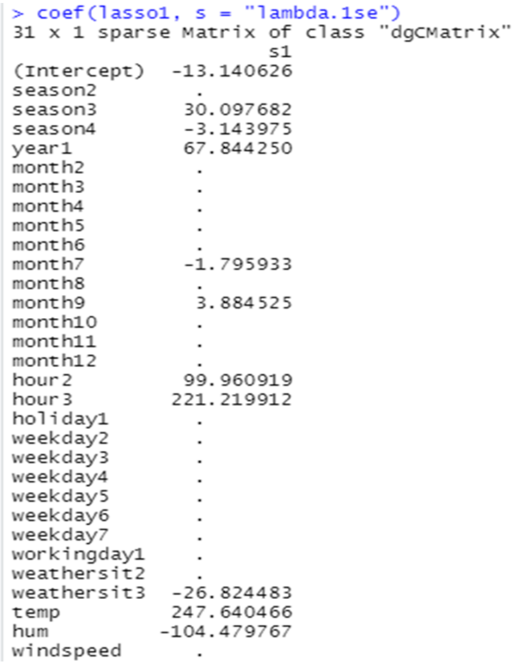


Figure 15: Final lasso output

To select the final model we will select the lambda 1se because is only 1 standard error above the MSE and the model will contain less variables. So, this lambda is parsimonious. We can understand from the graph above that the difference between the selected variables is big between the two lambdas selection. However, the difference in MSE is small and therefore, we select the lambda 1se to eliminate the parameters of the model.

The final model using the lasso technique is the following.



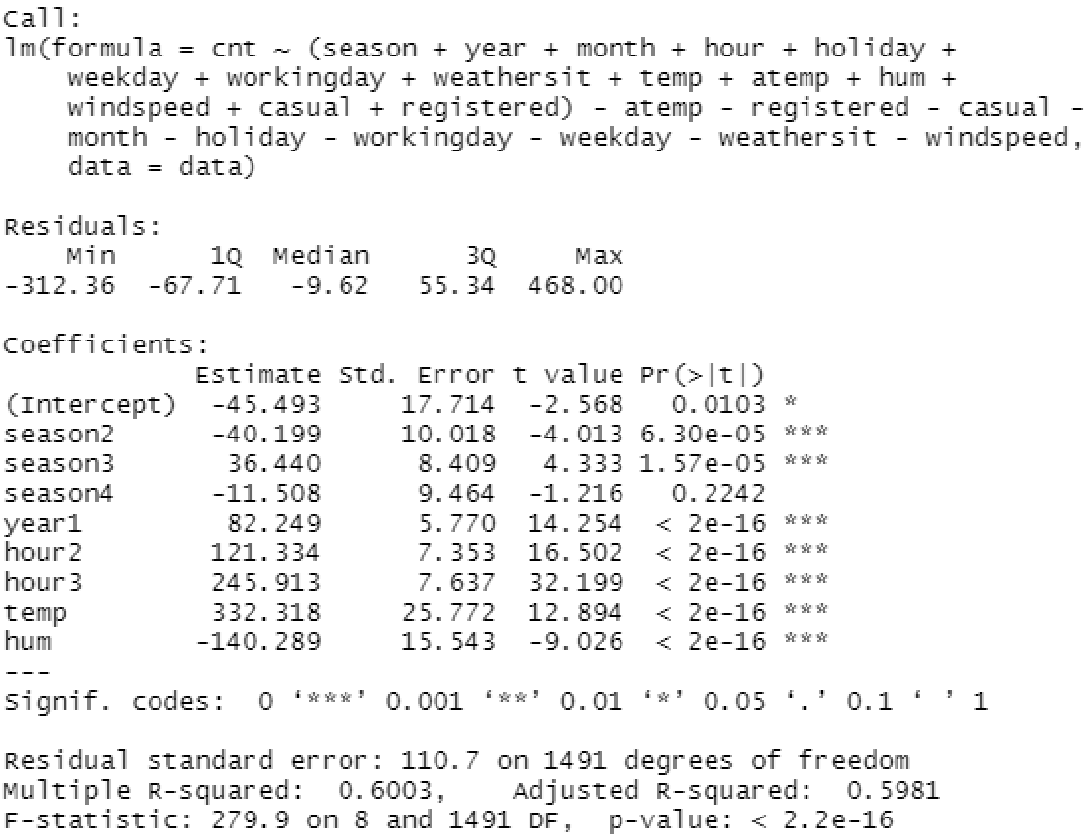
```
> coef(lasso1, s = "lambda.1se")
31 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept)  -13.140626
season2        .
season3       30.097682
season4       -3.143975
year1         67.844250
month2         .
month3         .
month4         .
month5         .
month6         .
month7        -1.795933
month8         .
month9         3.884525
month10        .
month11        .
month12        .
hour2         99.960919
hour3        221.219912
holiday1       .
weekday2       .
weekday3       .
weekday4       .
weekday5       .
weekday6       .
weekday7       .
workingday1    .
weathersit2    .
weathersit3  -26.824483
temp         247.640466
hum         -104.479767
windspeed      .
```

Figure 16: Final model using lasso

We can see that lasso regression shrunk the coefficients of the non-significant variables completely to zero.

B. Stepwise methodology

To determine which variables appear to be more significant we can use stepwise methods. In this assignment we select to use the stepwise method in both directions. In this way the method can add and exclude variables from the model simultaneously. We have also selected the summary() function which corresponds to the AIC criterion because the main goal of this analysis is prediction and not interpretation. In case we wanted to interpret the price value we would have used the BIC criterion. As a result, we end up to the following model:



```
Call:
lm(formula = cnt ~ (season + year + month + hour + holiday + 
    weekday + workingday + weathersit + temp + atemp + hum + 
    windspeed + casual + registered) - atemp - registered - casual - 
    month - holiday - workingday - weekday - weathersit - windspeed, 
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max 
-312.36  -67.71   -9.62   55.34  468.00 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -45.493     17.714  -2.568   0.0103 *  
season2      -40.199     10.018  -4.013 6.30e-05 ***
season3       36.440      8.409   4.333 1.57e-05 ***
season4      -11.508      9.464  -1.216   0.2242    
year1         82.249      5.770  14.254  < 2e-16 ***
hour2        121.334      7.353  16.502  < 2e-16 ***
hour3        245.913      7.637  32.199  < 2e-16 ***
temp         332.318     25.772  12.894  < 2e-16 ***
hum         -140.289     15.543  -9.026  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 110.7 on 1491 degrees of freedom
Multiple R-squared:  0.6003,	Adjusted R-squared:  0.5981 
F-statistic: 279.9 on 8 and 1491 DF,  p-value: < 2.2e-16
```

Figure 17: Stepwise model output

As we can see, the AIC criterion kept the intercept and the variables season, year, hour, temperature and humidity as significant for the model. The same result we have extracted from the lasso regression but here the coefficients are a little different.

6. Assumptions of the final model

A. Normality

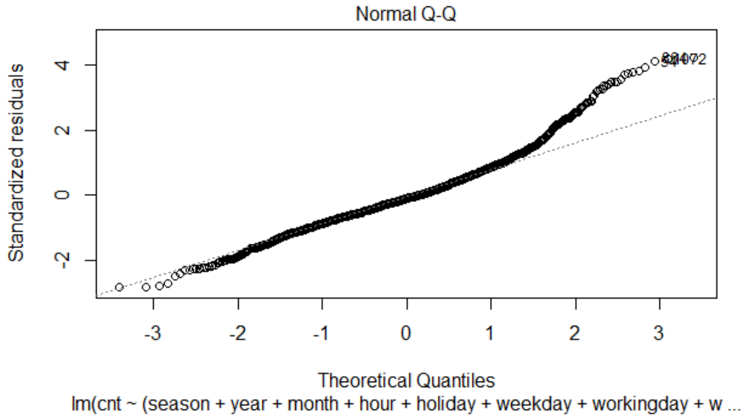
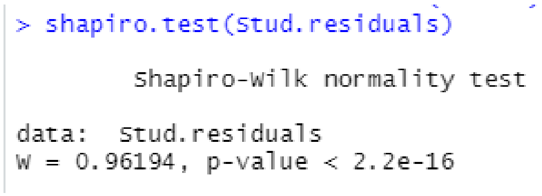


Figure 18: QQ Plot of residuals

We have strong indications from the Q-Q plot that the residuals of the final model are not normally distributed. This hypothesis has been confirmed from the Shapiro Wilk test (p-value <0,05)

Shapiro test for normality



```
> shapiro.test(Stud.residuals)

        Shapiro-Wilk normality test

data:  Stud.residuals
W = 0.96194, p-value < 2.2e-16
```

Figure 19: Shapiro test for the residuals

B. Constant variance

Then we check if the residuals are homoscedastic.

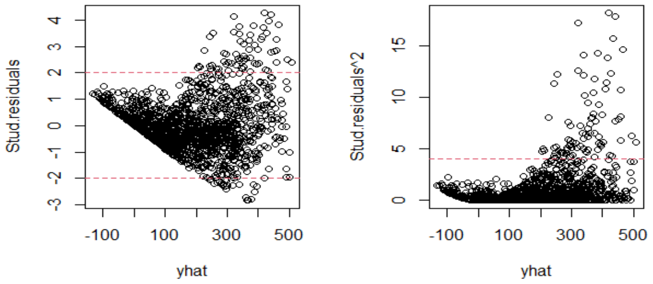
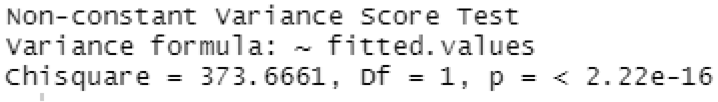


Figure 20: Constant variance plot

The ncvTest computes a score test of the hypothesis of constant error variance against the alternative that the error variance changes. Looking at the ncv test the p-value <0,05 and so the null hypothesis is rejected. So, we have strong indications that the variance is not constant, meaning that the residuals are heteroskedastic. The same result we can observe visually from the above graphs. In the first graph we have a lot of observations out of the red lines.

Ncv test for constant variance



```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 373.6661, Df = 1, p = < 2.22e-16
```

Figure 21: ncv test output

C. Linearity

The third assumption refers to the linearity of the model.Finally, we can observe from the graph that the residuals do not follow a line.

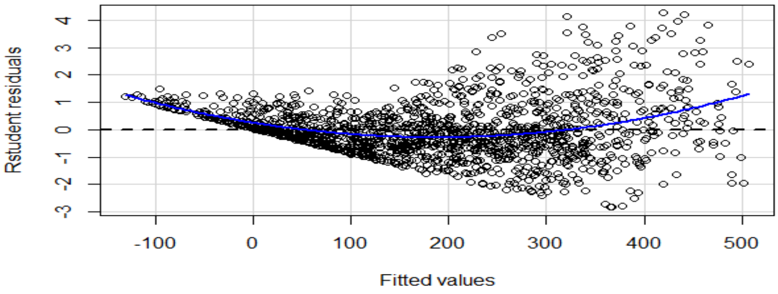
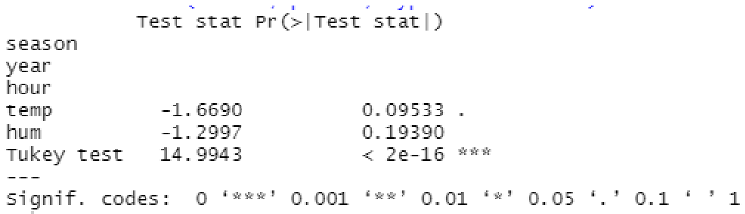


Figure 22: Linearity plot output

We can confirm this assumption with the Tukey test where we reject the null hypothesis of linearity.

Residual Plot test for linearity



```
             Test stat Pr(>|Test stat|)
season
year
hour
temp         -1.6690          0.09533 .
hum          -1.2997          0.19390
Tukey test   14.9943          < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 23: Residuals plot output

D. Independence of the residuals

As our data is time referenced it would be good to check about the independence of the residuals. We can have a visual indication from the graph below that the way the residuals are shown is random and does not follow a pattern.

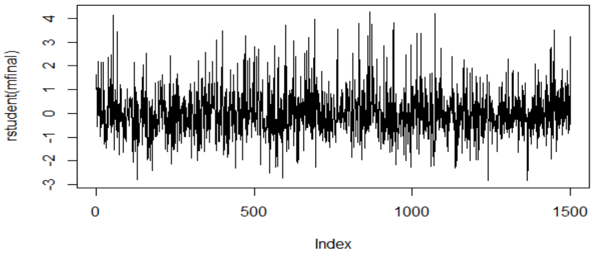
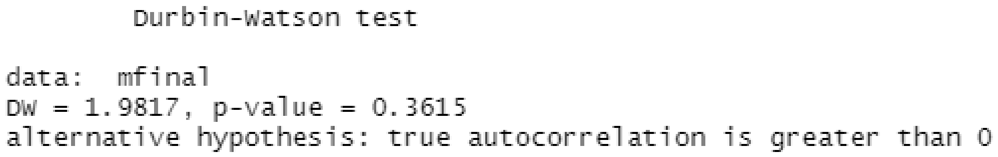


Figure 24: Independence plot output

We can confirm this assumption with the Durbin Watson test where we cannot reject the null hypothesis of independence. (p-value=0.3615>0,05). So, we have strong indications that the residuals are independent.

Durbin Watson test for Independence



```
        Durbin-Watson test

data:  mfinal
DW = 1.9817, p-value = 0.3615
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 25: Durbin Watson test output

E. Leverages points

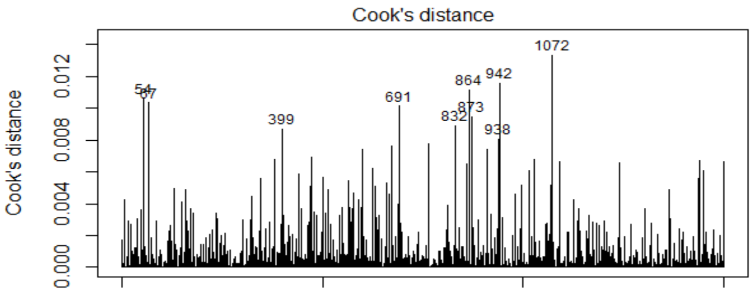


Figure 26: Top 10 leverages points

We found the top 10 leverages points and to exclude them from the model, but nothing really changed. We have only a small increase in the R^2 adjusted of the final model.

7. Alternative models to fix the assumptions

7.1 Square root - model

We constructed the following model:

sqrt(cnt)~.+poly(temp,2)+poly(hum,4)

and the normality and the variance have been improved. However, the assumption of linearity have not been accepted.

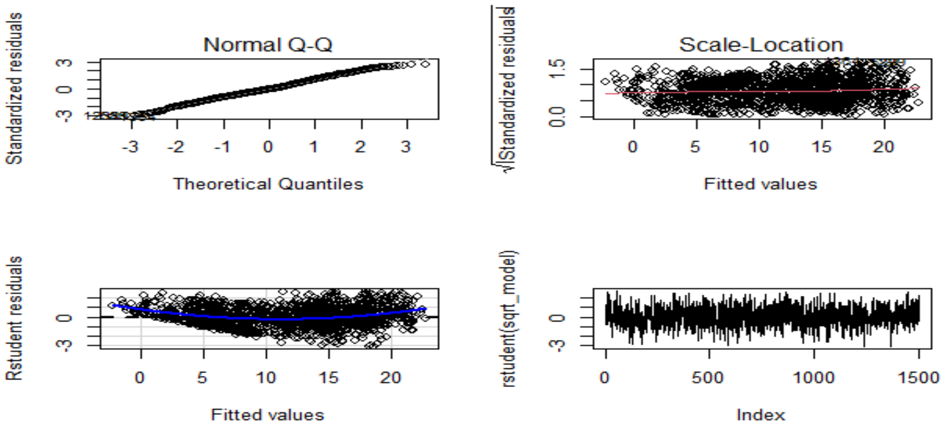


Figure 27: Assumptions visualization

7.2 Logarithm - model

Then we constructed the following model.

log(cnt)~.+exp(temp) -hum

The logarithm may fixes the linearity in the model although the normality and the constant variance have been rejected.

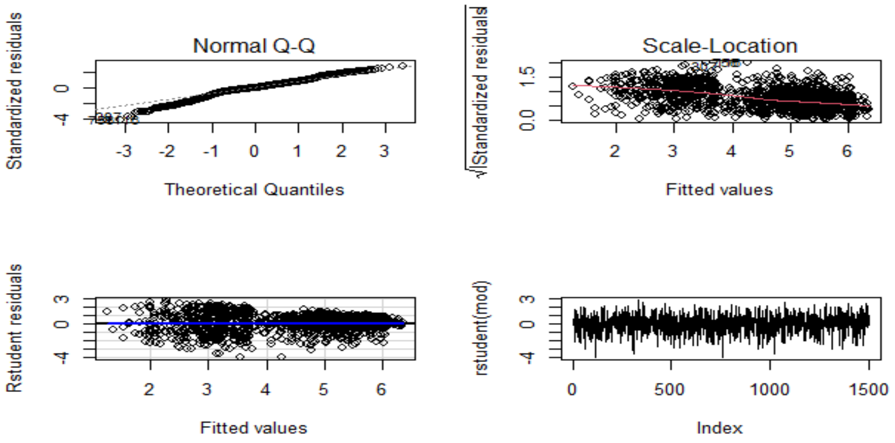
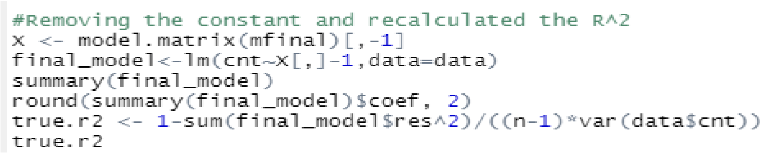


Figure 28: Assumptions Visualization for the log model

8. Removing the constant in order to interpret the final model

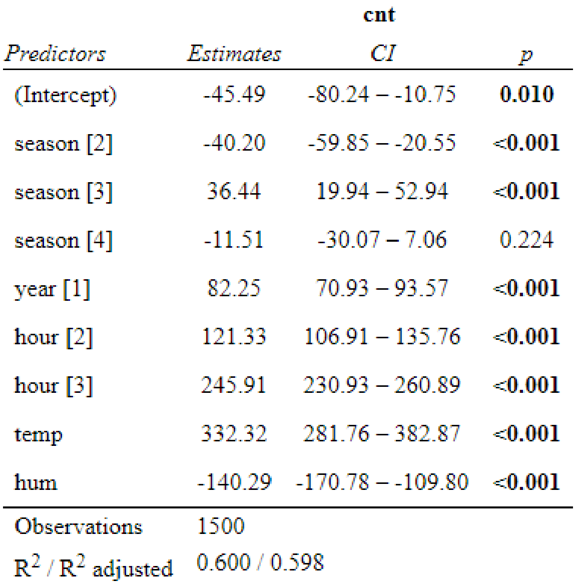


```
#Removing the constant and recalculated the R^2
x <- model.matrix(mfinal)[,-1]
final_model<-lm(cnt~x[,]-1,data=data)
summary(final_model)
round(summary(final_model)$coef, 2)
true.r2 <- 1-sum(final_model$res^2)/((n-1)*var(data$cnt))
true.r2
```

Figure 29: Removing the intercept - code

9. Comparison of the predictive ability of the models

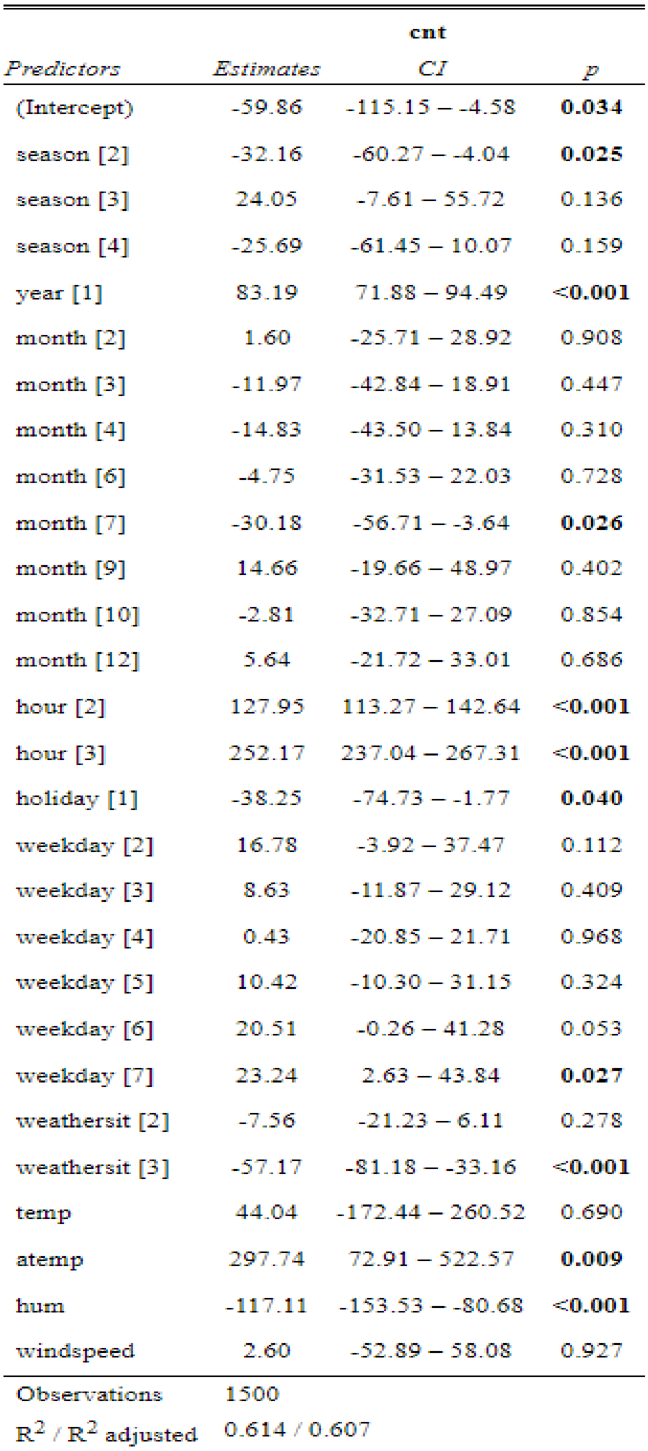
9.1 The final model



		cnt	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-45.49	-80.24 – -10.75	**0.010**
season [2]	-40.20	-59.85 – -20.55	**<0.001**
season [3]	36.44	19.94 – 52.94	**<0.001**
season [4]	-11.51	-30.07 – 7.06	0.224
year [1]	82.25	70.93 – 93.57	**<0.001**
hour [2]	121.33	106.91 – 135.76	**<0.001**
hour [3]	245.91	230.93 – 260.89	**<0.001**
temp	332.32	281.76 – 382.87	**<0.001**
hum	-140.29	-170.78 – -109.80	**<0.001**
Observations	1500		
R^2 / R^2 adjusted	0.600 / 0.598		

Figure 30: Summary output of the final model

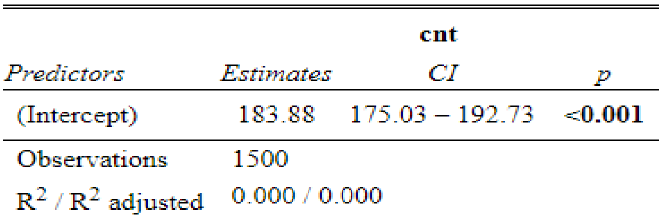
9.2 The full model



		cnt	
Predictors	*Estimates*	*CI*	*p*
(Intercept)	-59.86	-115.15 – -4.58	**0.034**
season [2]	-32.16	-60.27 – -4.04	**0.025**
season [3]	24.05	-7.61 – 55.72	0.136
season [4]	-25.69	-61.45 – 10.07	0.159
year [1]	83.19	71.88 – 94.49	**<0.001**
month [2]	1.60	-25.71 – 28.92	0.908
month [3]	-11.97	-42.84 – 18.91	0.447
month [4]	-14.83	-43.50 – 13.84	0.310
month [6]	-4.75	-31.53 – 22.03	0.728
month [7]	-30.18	-56.71 – -3.64	**0.026**
month [9]	14.66	-19.66 – 48.97	0.402
month [10]	-2.81	-32.71 – 27.09	0.854
month [12]	5.64	-21.72 – 33.01	0.686
hour [2]	127.95	113.27 – 142.64	**<0.001**
hour [3]	252.17	237.04 – 267.31	**<0.001**
holiday [1]	-38.25	-74.73 – -1.77	**0.040**
weekday [2]	16.78	-3.92 – 37.47	0.112
weekday [3]	8.63	-11.87 – 29.12	0.409
weekday [4]	0.43	-20.85 – 21.71	0.968
weekday [5]	10.42	-10.30 – 31.15	0.324
weekday [6]	20.51	-0.26 – 41.28	0.053
weekday [7]	23.24	2.63 – 43.84	**0.027**
weathersit [2]	-7.56	-21.23 – 6.11	0.278
weathersit [3]	-57.17	-81.18 – -33.16	**<0.001**
temp	44.04	-172.44 – 260.52	0.690
atemp	297.74	72.91 – 522.57	**0.009**
hum	-117.11	-153.53 – -80.68	**<0.001**
windspeed	2.60	-52.89 – 58.08	0.927
Observations	1500		
R^2 / R^2 adjusted	0.614 / 0.607		

Figure 31: Summary output for the full model

9.3 The null model



	cnt		
Predictors	*Estimates*	*CI*	*p*
(Intercept)	183.88	175.03 – 192.73	<**0.001**
Observations	1500		
R^2 / R^2 adjusted	0.000 / 0.000		

Figure 32: Summary output for the null model