BIKE SHARING DATA ANALYSIS

Gurleen Ahuja, Hardik Munjal

Western University
MDA 9159A

Table of Contents

1.1 Introduction

Public bike sharing has experienced a sharp increase on a global scale as an ingenious mobility solution. Although bike-sharing systems offer quick, affordable, and environmentally beneficial transportation, their unique features have negative effects on both riders and operators. In contrast to traditional public transit (such as buses and subways), which adheres to a set timetable and predetermined routes, bike-sharing offers on-demand transportation. This leads to an uneven distribution of bikes brought on by fluctuating demand and supply.

Effective bike rebalancing solutions are required to address this bicycle imbalance issue, which heavily relies on bicycle mobility modeling and prediction. Due to the imbalance of bicycles, bike-share towns must use expensive redistribution of bikes, which is normally carried out by trucks or trailers traveling throughout the city and relocating bikes between stations. Studies have been done to improve these bike redistribution procedures based on bicycle mobility models and predictions to maximize service availability and decrease redistribution costs.

1.1.1 Dataset

The primary data set comes from the Capital Bikeshare system in Washington, D.C., USA, and is based on a two-year historical log for the years 2011 and 2012 that is publicly accessible at the link. We have used the aggregated at the daily level available here.

The dataset contains per day casual user count with 731 entries, with 14 attributes – 13 independent and one dependent. The dataset contains season, year, month, hour, holiday, weekday, working day, weather conditions, temperature, feeling temperature, casual variable, count, and date information. Although, date does not provide relevant information to generate a model to predict the number of casual users.

We will discuss more on the attributes and their possible values under data understanding.

1.2 Literature Review

Since the establishment of the first bike-sharing system in the Netherlands in the 1960s, there have been four generations of bike-sharing (DeMaio, 2009) (Shaheen, 2010). Since the release of the third generation, bike sharing has grown in popularity. The automatic transaction kiosk at each station and identifiable bike-sharing users can be used to describe the third generation of bike-sharing. Around the world, these methods have achieved a fair amount of success. Fourth-generation bike-sharing programs featuring improved docking stations, bike redistribution, interaction with other means of transportation, and electric bikes have been built in Copenhagen and Madrid (DeMaio, 2009) (Shaheen, 2010).

Numerous studies have recently employed conventional surveys to ascertain the elements that would encourage urban communities to adopt bike sharing (Bikeshare, C., 2013) (Share, A. B., 2011). An invaluable resource for learning more about how bike sharing is used in the city is the automatic data collected from docking stations. Numerous studies have identified factors that affect the use of bike sharing and have attempted to forecast bike sharing flow using various urban factors, including population, jobs, bicycle lanes, proximity to public transportation, the density of bike sharing stations, altitude, retail shops, etc. (Faghih-Imani A. E.-G., 2014) (Rixey, 2013) (Wang, G., & JE, 2012). These studies' use of daily, monthly, or annual aggregated data can obscure the variation of everyday bike-sharing usage (Rixey, 2013) (Wang, G., & JE, 2012). In Barcelona and Seville, Spain, Hampshire used sub-city district-level aggregated hourly arrival and departure rates to study the built environment and bike-sharing utilization (Faghih-Imani A. H., 2017). They discovered that the density of bike-sharing stations, the capacity of the stations, and the number of sites of attraction are crucial variables in explaining the arrival and departure rates of bike-sharing. But rather than using bike-sharing flows at the station level, their study aggregated the flows at the level of sub-city districts, which was less meaningful.

1.3 Research Problem

In this project, we are working towards understanding the system function more effectively given the set of factors and modeling the number of casual users based on various predictors.

Our goal for this project is two-fold:

- **Model Interpretation:** We aim to infer the effect of different predictors on our response variable i.e. casual users to understand and further improve the bike-sharing system by making decisions based on our inferences.
- **Model Prediction:** We aim to develop a superior statistical model to predict the number of casual customers who will rent a bike on a specific day.

1.4 Data Understanding

Name of Dataset: Bike Sharing in Washington D.C. Dataset
Number of Observations: 731

1.4.1 Attributes Information

Attribute	Description	Values
instant	Record index	
dteday	Date	
season	Season	1: Spring 2: Summer 3: Fall 4: Winter
yr	Year	0: 2011 1:2012
mnth	Month	1 to 12
hr	Hour	0 to 23
holiday	extracted from Holiday Schedule	0: not holiday 1: holiday
weekday	Day of the week	
workingday		0: either a weekend or a holiday 1: neither a weekend nor a holiday
weathersit	(extracted from Freemeteo)	1: Clear, Few clouds, Partly cloudy 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
temp	Normalized temperature in Celsius.	0 to 1
atemp	Normalized feeling temperature in Celsius.	0 to 1
hum	Normalized humidity. The values are divided by 100 (max)	0 to 1
windspeed	Normalized wind speed. The values are divided to 67 (max)	0 to 1
casual	count of casual users	

1.5 Data Pre-processing

1.5.1 Checking for Null and Missing Values

We examined our data for NULL and missing values. The results showed that our dataset contained no missing values.

1.5.2 Categorical Variable Transformation

Next, we changed the names of the categories to more meaningful labels. For Instance, the categories of seasons were changed to ‘Spring’, Summer’, ‘Fall’, and ‘Winter’ instead of 1, 2, 3 and 4. We then converted our categorical variables to factor data type instead of numerical

1.5.3 Removing Unnecessary Predictors

We remove predictors which are not useful for us for e.g. registered, cnt and we will remove working day as from data description there seems to be exact collinearity between working day and (holiday + weekday)

1.5.4 Data Denormalization

As a next step in data pre-processing, we have denormalized the data as the data in our data set was normalized.

1.6 Exploratory Data Analysis

1.6.1 Descriptive Statistics

Parameter	Mean	Variance	Standard Deviation	Median	Min Value	Max Value
Actual Temperature	20.31	56.32	7.50	20.43	2.42	25.32
Actual Feel Temperature	23.71	66.39	8.14	24.33	3.95	42.04
Actual Humidity	62.78	202.86	14.24	62.66	0	97.25
Actual Windspeed	12.76	26.96	5.19	12.12	1.5	34
Casual Users	848	471450	686	713	2	3410

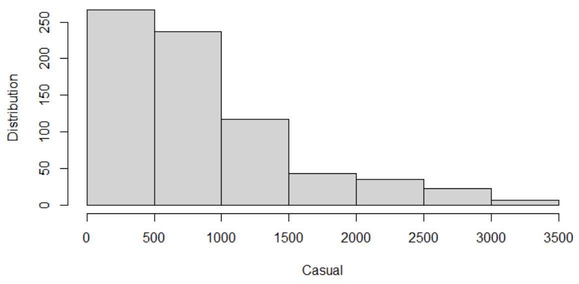
As observed from the table above:

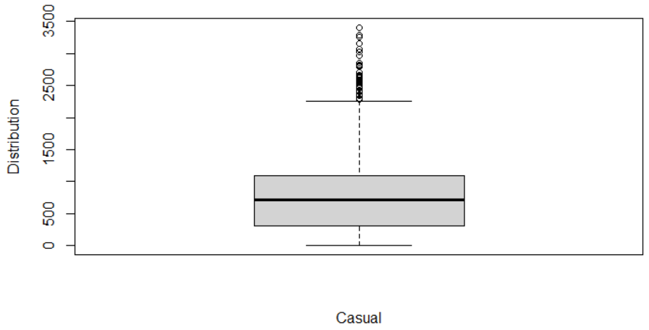
- The values for actual temperature range lie between 2.4 Celsius and 35.3 Celsius with a mean of 20.31
- The feeling temperature has statistics similar to the actual temperature.
- The values for actual humidity range from 0-97.2 with a median value of 62.6.
- The values for windspeed range lie between 1.5 and 34 with mean off 12

1.6.2 Visualization

1.6.2.1 Response Variable

We then checked the distribution of our target variable casual using a histogram and a box plot. Both the plots are shown below.

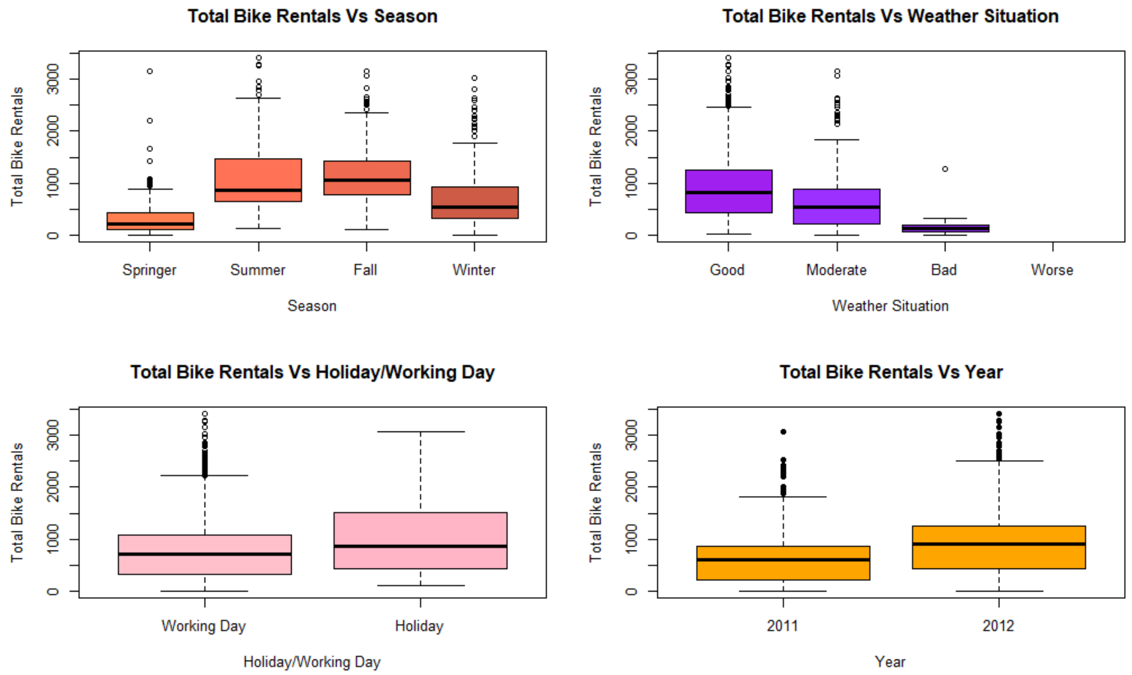




The above histogram suggests that we have a right tailed distribution for casual. The above box plot indicates the presence of outliers in the data.

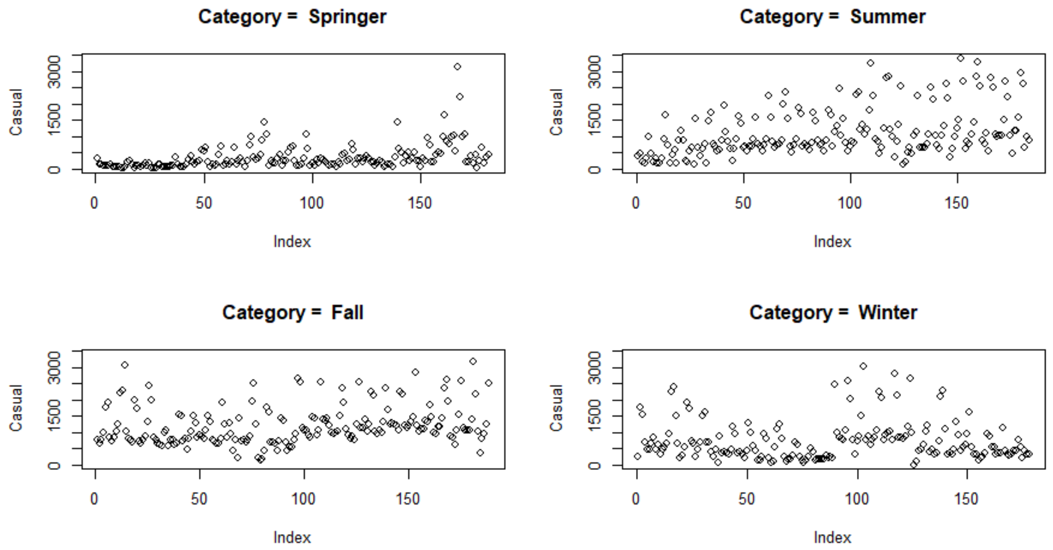
1.6.2.2 Categorical Variables

As a next step to our initial analysis, we visualized our categorical features using scatter plots and box plots. These plots are shown below.

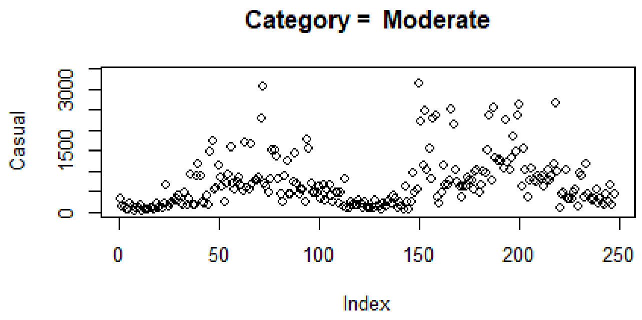


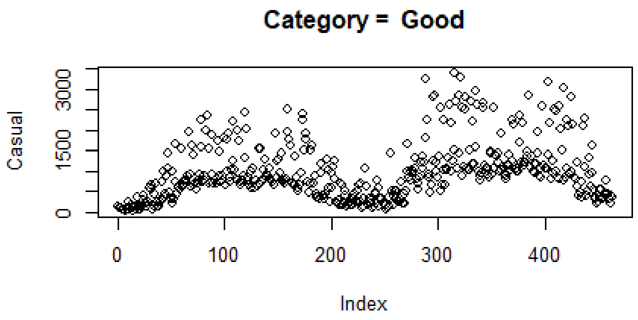
Figure(a)

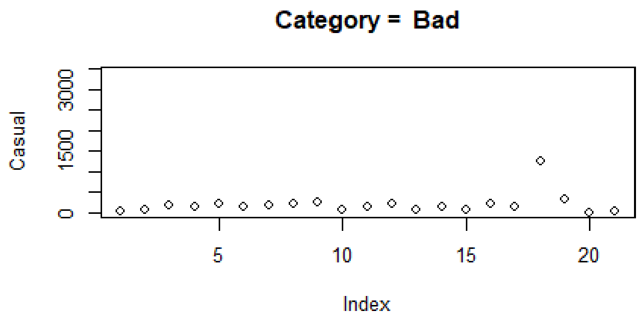
Figure(a) suggests that the season type has a relationship with the number of casual users. The number of casual users is high in Summer and Fall, and lower in Spring and Winter.



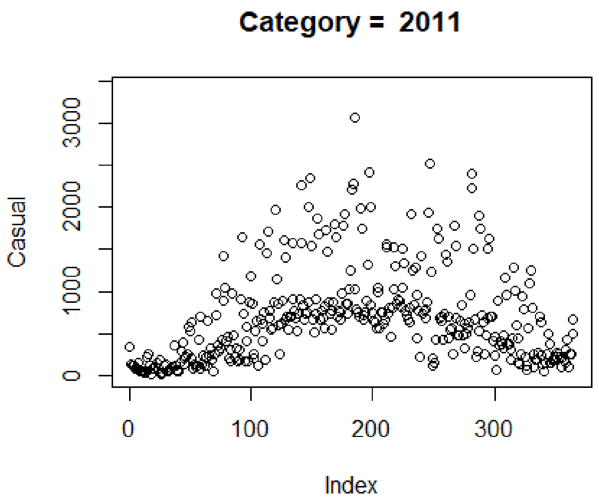
Figure(a) shows that the number of casual users is directly proportional to how good the weather condition is.

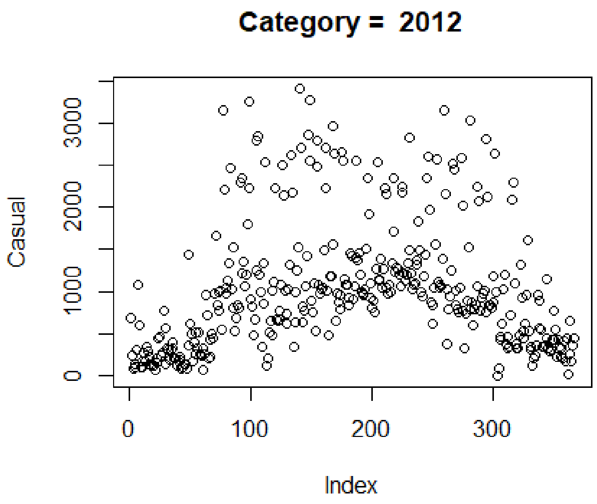


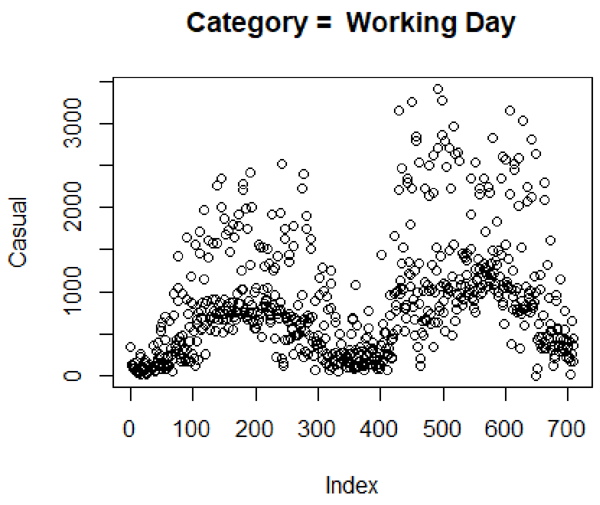


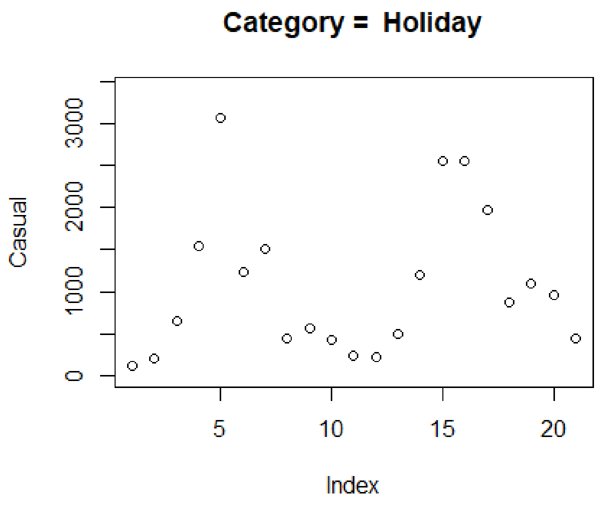


Figure(a) indicates that the number of casual users have increased from 2011 to 2012.









1.6.2.3 Continuous Variables

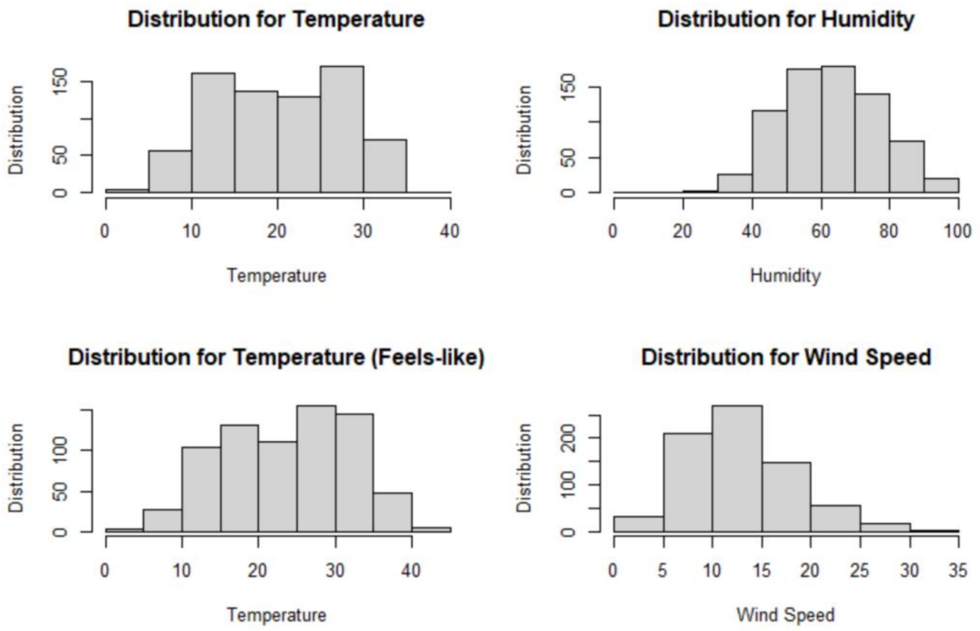


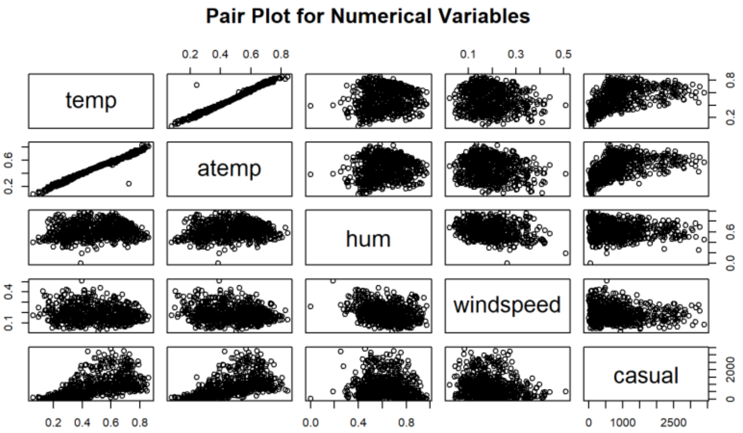
Figure (b)

As per figure (b), The distribution of Humidity is left-skewed. The above distribution for temperature seems to be right skewed. In the next step, we have illustrated the pair plot for our continuous variables.

1.6.2.4 Correlation Analysis

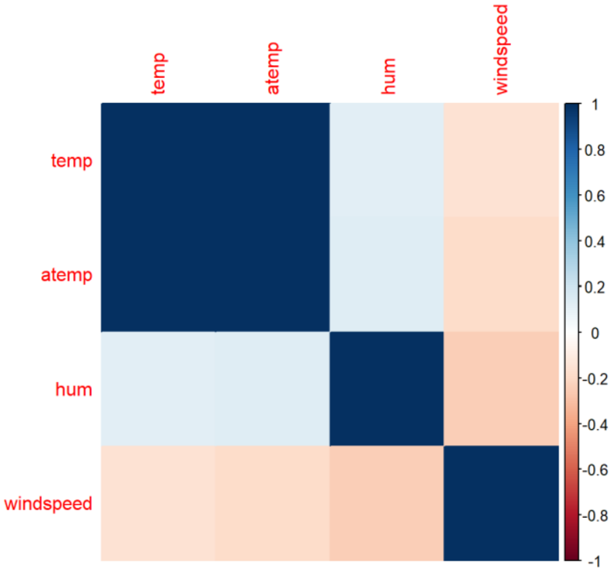
If we look at the graph between temp and casual, from the plot, Casual User Count is distributed as a cloud that is densely centred between 0.5 and 0.8. Nearly all of the data points will have an impact on our regression model, as evidenced by the minor tailing clusters near the higher end of the X axis.

As shown above, the data points in the distribution between windspeed and casual also form a noticeable cloud around the Wind Speed between 0 and 0.3.



It is clear from the plot between humidity and casual that for humidity between 0.2 and 1 form a cloud. We can also see that temp and atemp have a very strong linear relationship. We next find the correlation matrix to test this further.

Temperature and Feeling Temperature are highly associated, as seen in the correlation graph above. As a result, one of the variables would need to be eliminated from our regression model. We will ahead with excluding atemp from our analysis.



1.6.3 Statistical Inference - Experimentation

As observed from visualizing plots, temperature, seasons and workingday have strong association with the response variable - casual. To confirm association we need to run experiments to confirm if our response variable is associated with the identified variables. We will try to answer the below questions through statistical tests and make inferences if possible. Although, if we get significant results, we need to test that all model assumptions are met before making any inferences. A regression model assumes the following four assumptions:

- **Linearity:** Test using scatter plot between independent and response variable
- **Independence:** Test using residual plot to check residuals are independent
- **Normality:** Test using Q-Q plot or Shapiro-Wilk test whether the residuals of the model are normally distributed
- **Equal Variance (Homoscedasticity):** Test using residual vs fitted plot or BP-test if the residuals have constant variance at every level of x

1.6.3.1 Is temperature significant predictor for our response variable casual users?

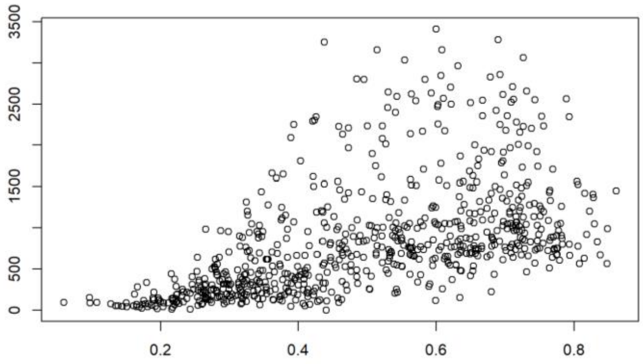
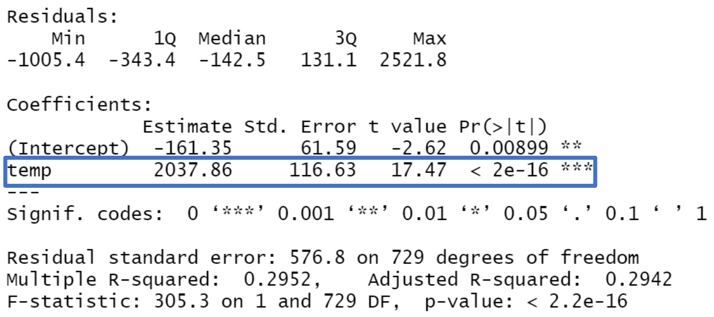


Figure 3 (a)



```
Residuals:
    Min      1Q  Median      3Q     Max
-1005.4  -343.4  -142.5   131.1  2521.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -161.35      61.59   -2.62  0.00899 **
temp         2037.86     116.63   17.47  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 576.8 on 729 degrees of freedom
Multiple R-squared:  0.2952,	Adjusted R-squared:  0.2942 
F-statistic: 305.3 on 1 and 729 DF,  p-value: < 2.2e-16
```

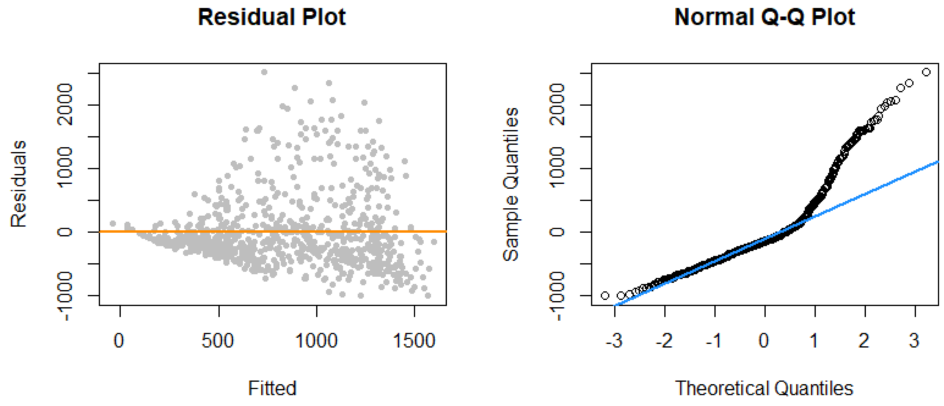
Figure 3(b)

Looking at the Figure 3 (a), it seems that temperature has linear relationship with response variable - casual. Let’s test the causality by running a hypothesis test (t-test):

$$H0: \beta_{temp} = 0;\ H1: \beta_{temp} \neq 0$$

We can see that the p-value is very small. This serves as evidence against the null point hypothesis. This means that we can reject H0. We can thus conclude that temperature is a significant predictor of casual.

Before deriving any further inferences, we will check for model assumptions:

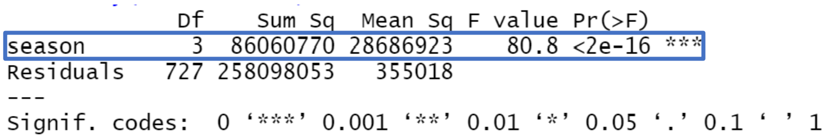


From Figure 3 (a) we observe a linear relationship between temperature and casual users. As observed from the plots above Independence, Normality and Equal Variance assumptions are being violated. As our goal for this analysis was to answer if temperature is a significant predictor rather than how it influences number of casual users we will not try to fix these assumptions and conclude that temperature does influence the number of casual user but we will try to infer relationship later on in the project.

1.6.3.2 Does season influence number of casual users?

We test the hypothesis for pattern observed during visualization of categorical parameters.

$$H0: \beta_{season} = 0; \; H1: \beta_{season} \neq 0$$



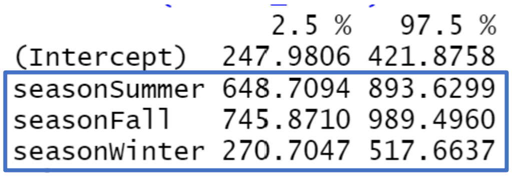
```
             Df    Sum Sq  Mean Sq F value Pr(>F)    
season        3  86060770 28686923    80.8 <2e-16 ***
Residuals   727 258098053   355018                   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 4. F-test Result

Since the p-value is very small, we have against the null point hypothesis. We can hence reject it and conclude that Season is a significant predictor of casual. F-score is much greater than one, this implies that variance between is probably the source of most of the variance in the total sample, and the samples probably come from populations with different means.

Let’s try to calculate the 95% confidence interval of β parameters. Therefore, from the result on the right-hand side we can infer that we are 95% confident about:

- β_{summer} lies between 247.98 and 421.87
- β_{fall} lies between 745.87 and 989.49
- β_{winter} lies between 270.70 and 517.66

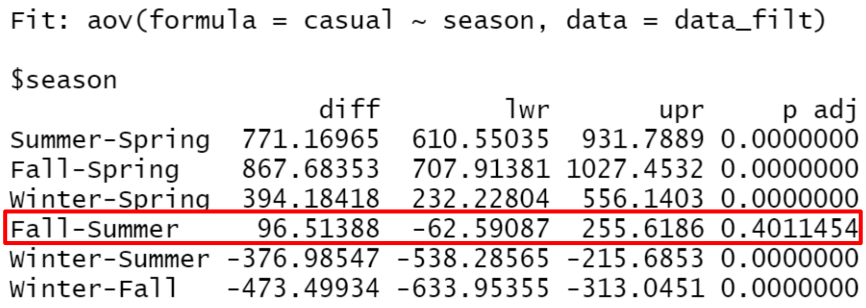


```
                2.5 %   97.5 %
(Intercept)  247.9806 421.8758
seasonSummer 648.7094 893.6299
seasonFall   745.8710 989.4960
seasonWinter 270.7047 517.6637
```

1.6.3.3 Which two seasons show significant difference in average casual users?

As from one-way ANOVA test we observed there is a significant relationship between seasons and number of causal users, we now try to understand whether a change in season has any effect on the number of casual users using bike sharing system. We will perform an TukeyHSD test at a significance level of 0.05 to assess the significance of difference between change in seasons.

As observed from results on righthand side, the p-value for “Fall to Summer” is higher than significance level. Therefore, we can conclude that apart from Fall to Summer all seasons change has significant impact in change of number of users and exact value can be observed under “diff” from the results.



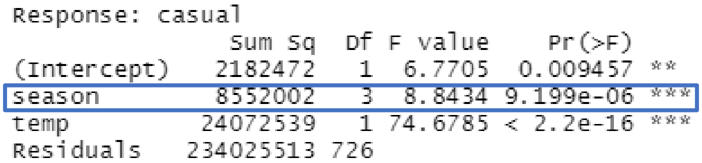
```
Fit: aov(formula = casual ~ season, data = data_filt)

$season
                    diff        lwr       upr     p adj
Summer-Spring  771.16965  610.55035  931.7889 0.0000000
Fall-Spring    867.68353  707.91381 1027.4532 0.0000000
Winter-Spring  394.18418  232.22804  556.1403 0.0000000
Fall-Summer     96.51388  -62.59087  255.6186 0.4011454
Winter-Summer -376.98547 -538.28565 -215.6853 0.0000000
Winter-Fall   -473.49934 -633.95355 -313.0451 0.0000000
```

The pairs of season with highest difference in number of casual users are Summer-Spring and Fall-Spring.

1.6.3.4 What is the effect of season after adjusting for temperature?

From our analysis in 1.6.3.3 we concluded that season has a significant impact on number of casual users. But it is possible that temperature might be influencing this significance. Therefore, let us try to understand if there is a statistically significant difference between different season after removing the effect of temperature. To discount the effect of temperature we will treat it as a covariate. For this analysis we will use the ANCOVA test (Analysis of Covariance) with type III error instead of ANOVA test.

As observed from the results on right hand size season still has a very small p-value which indicates it has a significant impact on number of users even after removing the effect of temperature.



```
Response: casual
               Sum Sq  Df F value    Pr(>F)    
(Intercept)   2182472   1  6.7705  0.009457 ** 
season        8552002   3  8.8434 9.199e-06 ***
temp         24072539   1 74.6785 < 2.2e-16 ***
Residuals   234025513 726                      
```

One thing to note here is that we still do not know how season affects the number of casual user because the assumptions of our model were violated so any inferences we make won't be correct. But we are aware that some family of relationship does exist between season and casual number of users.

1.6.4 Model Fitting – Inference

Now after validating the hypothesis generated from descriptive statistics and domain knowledge we will focus on achieving our first goal to generate inferences on the number of casual users.

To achieve this goal we will try to generate the simplest model i.e. with the least number of predictors and use it to generate inferences. Also, as discussed earlier before making any inferences we need to make sure that our model assumptions are fulfilled, otherwise, any inferences we make won't be significant.

1.6.4.1 Variable Selection

To select the least number of predictors we will perform variable selection using Forward, Backward and Stepwise selection. We started with all the predictors and using both AIC and BIC, we evaluated the error value and predictors returned. The below table summarizes the results:

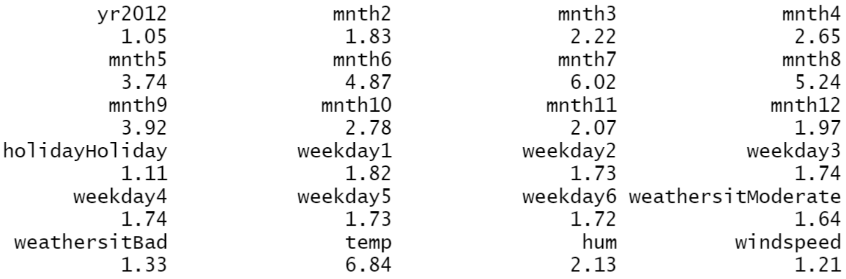
Metric	Selection Method	Error Value	Adjusted R-Squared	p (No. of Predictors)	Predictors
AIC	Forward Selection	8606	0.735	9	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum, season
	Backward Selection	8606	0.735	9	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum, season
	Stepwise Selection	8606	0.735	9	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum, season
BIC	Forward Selection	8725	0.716	8	weekday, year, weathersit, temp, holiday, windspeed, hum, season
	Backward Selection	8723	0.733	8	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum
	Stepwise Selection	8725	0.716	8	weekday, year, weathersit, temp, holiday, windspeed, hum, season

As we can observe from the table AIC remains the same for all selection methods as there is no change in the predictors. But when we consider BIC as an evaluation metric we get the simplest and the best model in terms of Adjusted R-Squared and BIC in backward selection. Therefore, we will use the highlighted model for further analysis.

1.6.4.2 Multi-Collinearity

Now as we have obtained the simplest model we will check for any multi-collinearity amongst these independent variables before working on the inferences. It is important to check for multi-collinearity as if two independent variables are collinear it would result in inflated p-values which would make it difficult to remove them based on the p-values as it would be very hard to reject the hypothesis even if they are not significant.

We calculate VIF (Variance Inflation Factor) to check for multicollinearity. As we can observe all the variables have VIF values less than 10. Therefore, we can move ahead with our analysis without removing any more predictors.



```
         yr2012               mnth2               mnth3               mnth4 
           1.05                1.83                2.22                2.65 
          mnth5               mnth6               mnth7               mnth8 
           3.74                4.87                6.02                5.24 
          mnth9              mnth10              mnth11              mnth12 
           3.92                2.78                2.07                1.97 
 holidayHoliday            weekday1            weekday2            weekday3 
           1.11                1.82                1.73                1.74 
       weekday4            weekday5            weekday6 weathersitModerate 
           1.74                1.73                1.72                1.64 
   weathersitBad                temp                 hum           windspeed 
           1.33                6.84                2.13                1.21 
```

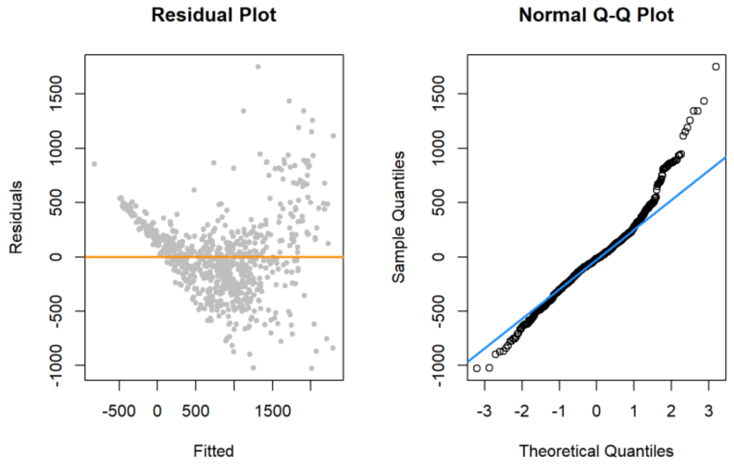
1.6.4.3 Model Assumption Violation

Moving ahead with our analysis, if we want to make inferences using the obtained model, we need to check that no model assumptions mentioned earlier are being violated and if they are being violated, we need to perform transformations to fulfil those assumptions.

Considering our data has been sampled from "day.csv", there is a high chance that this data represents a time-series. Therefore, the assumption of independence will always be violated. But we are moving further and assuming it isn't a concern right now in our analysis.

As observed from the residual plot linearity, and as observed from bp-test equal variance is violated. Furthermore, from Q-Q plot and sw-test we see that normality is also violated as their value is very small compared to the significance value.

Now we will try to fix these model assumptions by checking for Influential points and applying transformations as needed.

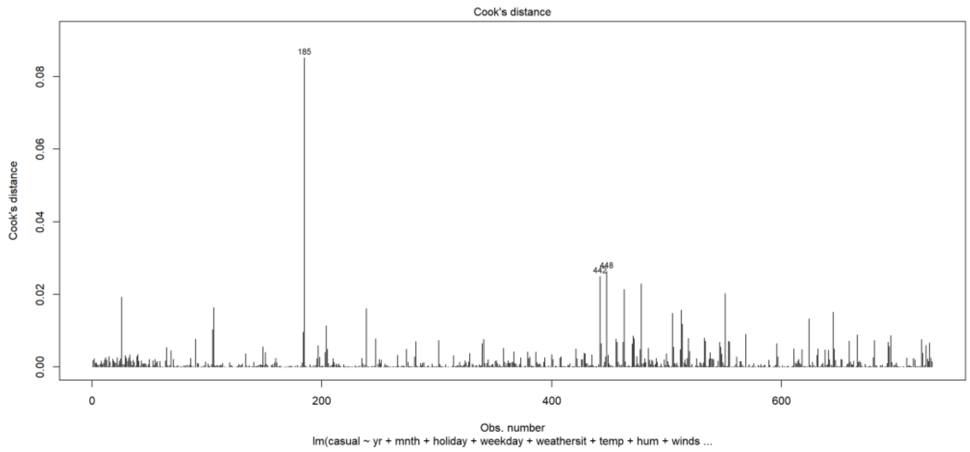


BP-Test	Shapiro-Wilk Test
1.19 x 10^{-20}	3.66 x 10^{-13}

1.6.4.3.1 Unusual Observations

There can be scenarios where a small number of unusual points may have large influence on the fitted model. This might cause us to misinterpret the underlying pattern and often model assumptions are violated because of these points. We will use cook's distance to calculate the influential points in the dataset.

As observed from the plot below, data points with cooks' distance greater than 0.005 are influential, so we get around 50 influential points.

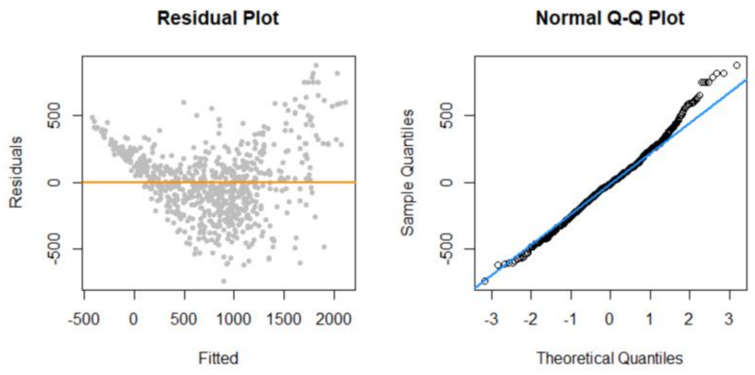


Now we could either remove these data points or apply transformations on response and predictor variables.

1.6.4.3.1.1 Removing Influential Points

After removing the 50 influential points we again fit the model and check for validity of model assumptions

As observed from the residual plot linearity, and as observed from bp-test equal variance is violated. Furthermore, from Q-Q plot and sw-test we see that normality although better than previous fit is still violated as their value is very small compared to the significance value.



BP-Test	Shapiro-Wilk Test
3.96 x 10^{-25}	0.00049

But we cannot remove the influential points as we do not have the information pertaining to incorrectness of these data points. Therefore, we will apply transformation on response and predictor variable to fix the model assumptions.

1.6.4.3.1.2 Transformation

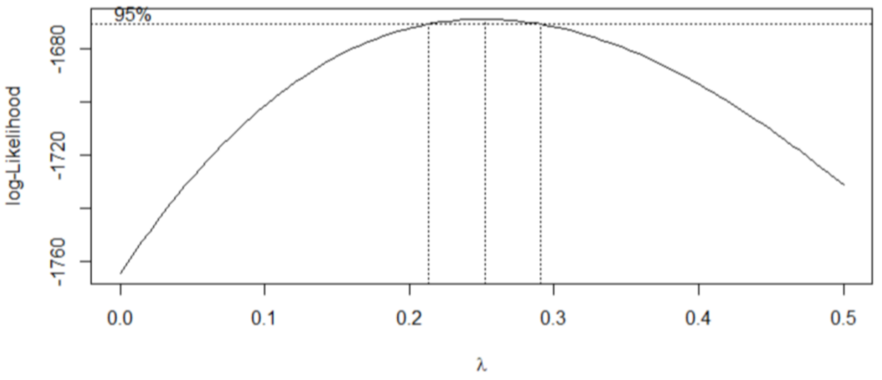
1.6.4.3.1.2.1 Response Transformation

We will use the boxcox method to understand out of all the family of transformations which one would be best for our use case. As we observed during visualization of response variable, casual was right-tailed. We will be able to apply appropriate transformation to fix the skewness of distribution of casual.

As observed from the plot on right side, we can get optimal value of lambda at global maximum. The lambda value seems to be 0.25.

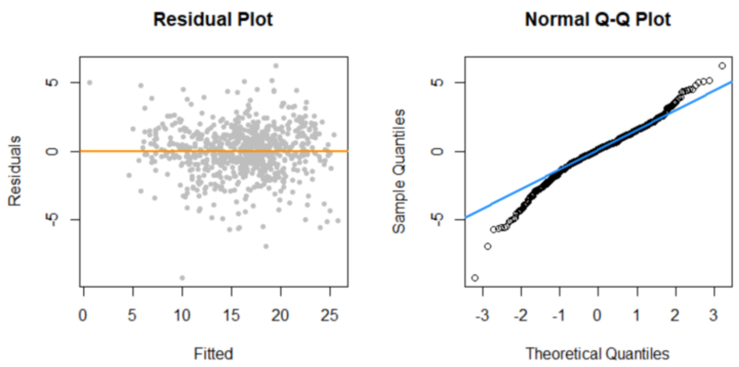
Therefore, we will apply the following transformation on casual:

$$casual = \frac{casual^{\lambda} - 1}{\lambda}$$



After applying the transformation we again fit the model and check for model assumptions:

As observed from the residual plot linearity assumption seems to fulfilled. But, as observed from bp-test equal variance is violated. Furthermore, from Q-Q plot and sw-test we see that normality is worse than previous fit and violated as their value is very small compared to the significance value.

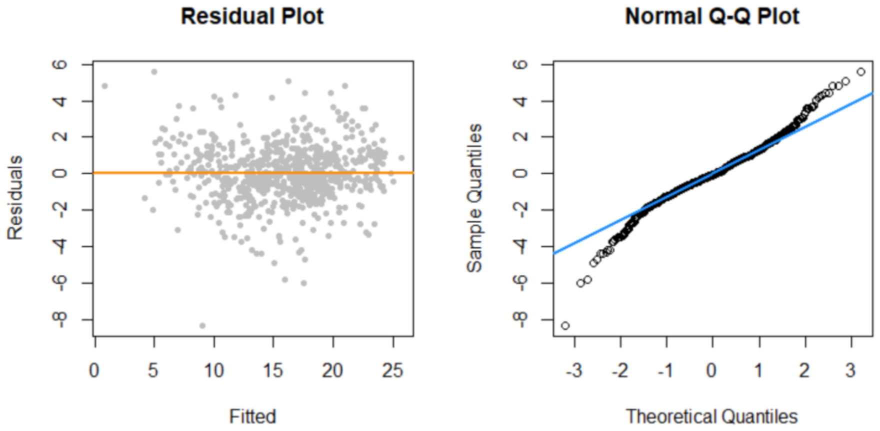


BP-Test	Shapiro-Wilk Test
2.26 x 10^{-12}	1.76 x 10^{-10}

1.6.4.3.1.2.2 *Predictor Transformation*

As we didn't get the desired model assumptions fulfilled from the response transformations we will apply transformation on predictors using Spline. We tested for polynomial transformation but the model assumptions were not satisfied. Therefore, we will proceed with splines and test for model assumptions.

As observed from the residual plot linearity, and as observed from bp-test equal variance is violated. Furthermore, from Q-Q plot and sw-test we see that normality is unchanged compared to the previous fit and violated as their value is very small compared to the significance value.



BP-Test	Shapiro-Wilk Test
2.26 x 10^{-12}	1.52 x 10^{-10}

Therefore, from this analysis we conclude that the model obtained after response transformation can be used to make inferences, but the inferences will not be effective or even insignificant because the model assumptions are violated. We will refer to this model as Model 1 for our further analysis.

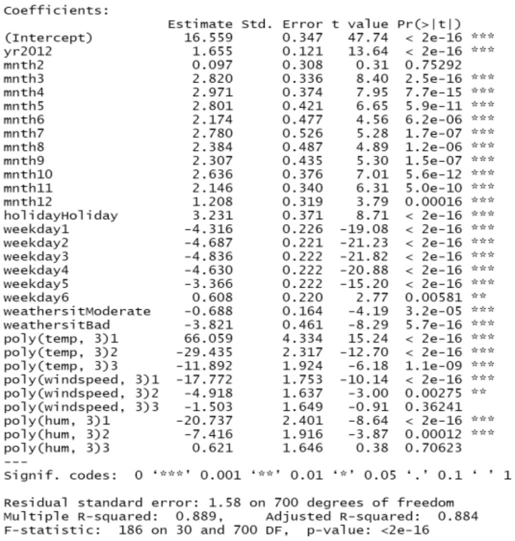
1.6.5 Model Fitting – Prediction

Now working towards our second goal of building a statistical model with high predictive power we will take the Model 1 as our base and add complexity to achieve better prediction. We will fit multiple models and then compare them based on Adjusted R-squared value.

1.6.5.1 Polynomial Terms

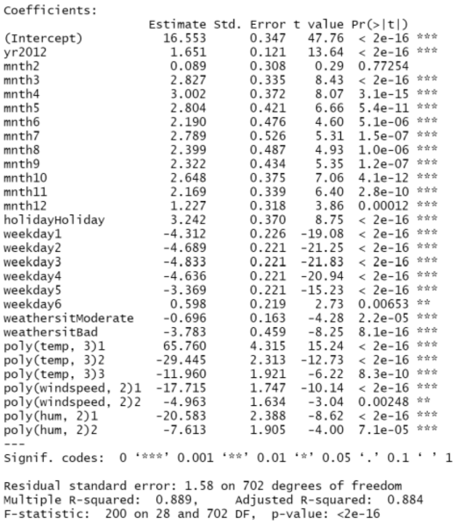
We will try to add polynomial terms to fit our model better to the data. We start by adding polynomial terms to continuous variables with power 3 – temperature, humidity and windspeed. We then iterate and remove values with p-value greater than 0.05. A higher p-value signifies they are not adding any value to the model so therefore, we can remove them, and predictive power of the model is not affected. We repeat this process until all predictors have p-value less then significance level.

We calculate the Adjusted R-squared to be 0.884 and we will refer to this model as Model 2 for future analysis.



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             16.559      0.347   47.74  < 2e-16 ***
yr2012                   1.655      0.121   13.64  < 2e-16 ***
mnth2                    0.097      0.308    0.31  0.75292
mnth3                    2.820      0.336    8.40  2.5e-16 ***
mnth4                    2.971      0.374    7.95  7.7e-15 ***
mnth5                    2.801      0.421    6.65  5.9e-11 ***
mnth6                    2.174      0.477    4.56  6.2e-06 ***
mnth7                    2.780      0.526    5.28  1.7e-07 ***
mnth8                    2.384      0.487    4.89  1.2e-06 ***
mnth9                    2.307      0.435    5.30  1.5e-07 ***
mnth10                   2.636      0.376    7.01  5.6e-12 ***
mnth11                   2.146      0.340    6.31  5.0e-10 ***
mnth12                   1.208      0.319    3.79  0.00016 ***
holidayHoliday           3.231      0.371    8.71  < 2e-16 ***
weekday1                -4.316      0.226  -19.08  < 2e-16 ***
weekday2                -4.687      0.221  -21.23  < 2e-16 ***
weekday3                -4.836      0.222  -21.82  < 2e-16 ***
weekday4                -4.630      0.222  -20.88  < 2e-16 ***
weekday5                -3.366      0.222  -15.20  < 2e-16 ***
weekday6                 0.608      0.220    2.77  0.00581 **
weathersitModerate      -0.688      0.164   -4.19  3.2e-05 ***
weathersitBad           -3.821      0.461   -8.29  5.7e-16 ***
poly(temp, 3)1          66.059      4.334   15.24  < 2e-16 ***
poly(temp, 3)2         -29.435      2.317  -12.70  < 2e-16 ***
poly(temp, 3)3         -11.892      1.924   -6.18  1.1e-09 ***
poly(windspeed, 3)1    -17.772      1.753  -10.14  < 2e-16 ***
poly(windspeed, 3)2     -4.918      1.637   -3.00  0.00275 **
poly(windspeed, 3)3     -1.503      1.649   -0.91  0.36241
poly(hum, 3)1          -20.737      2.401   -8.64  < 2e-16 ***
poly(hum, 3)2           -7.416      1.916   -3.87  0.00012 ***
poly(hum, 3)3            0.621      1.646    0.38  0.70623
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.58 on 700 degrees of freedom
Multiple R-squared:  0.889,	Adjusted R-squared:  0.884
F-statistic:  186 on 30 and 700 DF,  p-value: <2e-16
```

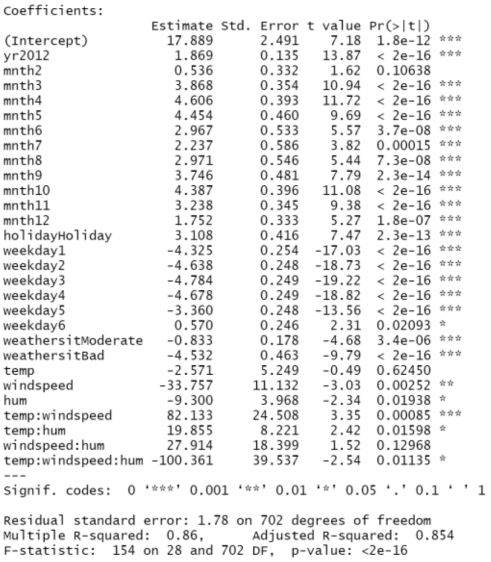
Removed hum^3, windspeed^3



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)             16.553      0.347   47.76  < 2e-16 ***
yr2012                   1.651      0.121   13.64  < 2e-16 ***
mnth2                    0.089      0.308    0.29  0.77254
mnth3                    2.827      0.335    8.43  < 2e-16 ***
mnth4                    3.002      0.372    8.07  3.1e-15 ***
mnth5                    2.804      0.421    6.66  5.4e-11 ***
mnth6                    2.190      0.476    4.60  5.1e-06 ***
mnth7                    2.789      0.526    5.31  1.5e-07 ***
mnth8                    2.399      0.487    4.93  1.0e-06 ***
mnth9                    2.322      0.434    5.35  1.2e-07 ***
mnth10                   2.648      0.375    7.06  4.1e-12 ***
mnth11                   2.169      0.339    6.40  2.8e-10 ***
mnth12                   1.227      0.318    3.86  0.00012 ***
holidayHoliday           3.242      0.370    8.75  < 2e-16 ***
weekday1                -4.312      0.226  -19.08  < 2e-16 ***
weekday2                -4.689      0.221  -21.25  < 2e-16 ***
weekday3                -4.833      0.221  -21.83  < 2e-16 ***
weekday4                -4.636      0.221  -20.94  < 2e-16 ***
weekday5                -3.369      0.221  -15.23  < 2e-16 ***
weekday6                 0.598      0.219    2.73  0.00653 **
weathersitModerate      -0.696      0.163   -4.28  2.2e-05 ***
weathersitBad           -3.783      0.459   -8.25  8.1e-16 ***
poly(temp, 3)1          65.760      4.315   15.24  < 2e-16 ***
poly(temp, 3)2         -29.445      2.313  -12.73  < 2e-16 ***
poly(temp, 3)3         -11.960      1.921   -6.22  8.3e-10 ***
poly(windspeed, 2)1    -17.715      1.747  -10.14  < 2e-16 ***
poly(windspeed, 2)2     -4.963      1.634   -3.04  0.00248 **
poly(hum, 2)1          -20.583      2.388   -8.62  < 2e-16 ***
poly(hum, 2)2           -7.613      1.905   -4.00  7.1e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.58 on 702 degrees of freedom
Multiple R-squared:  0.889,	Adjusted R-squared:  0.884
F-statistic:  200 on 28 and 702 DF,  p-value: <2e-16
```

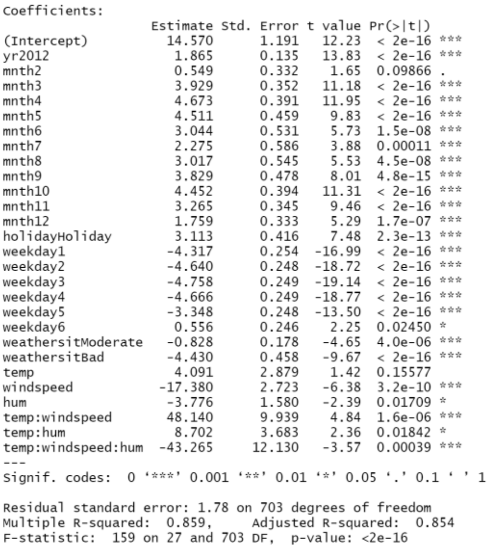
1.6.5.2 Interaction Terms

We will try to add interaction terms to fit our model better to the data. We start by adding interaction terms between continuous variables – temperature, humidity and windspeed. We then iterate and remove predictors with p-value greater than 0.05. A higher p-value signifies they are not adding any value to the model so therefore, we can remove them, and predictive power of the model is not affected. We repeat this process until all predictors have p-value less then significance level.



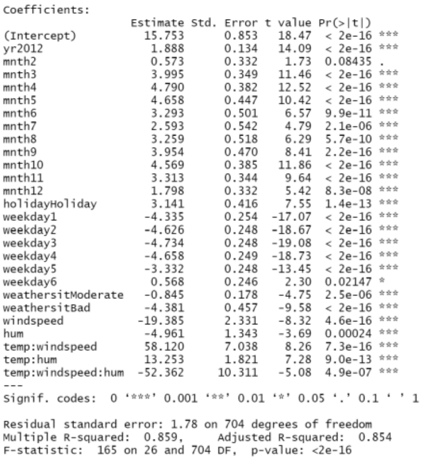
```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            17.889      2.491    7.18  1.8e-12 ***
yr2012                  1.869      0.135   13.87  < 2e-16 ***
mnth2                   0.536      0.332    1.62  0.10638
mnth3                   3.868      0.354   10.94  < 2e-16 ***
mnth4                   4.606      0.393   11.72  < 2e-16 ***
mnth5                   4.454      0.460    9.69  < 2e-16 ***
mnth6                   2.967      0.533    5.57  3.7e-08 ***
mnth7                   2.237      0.586    3.82  0.00015 ***
mnth8                   2.971      0.546    5.44  7.3e-08 ***
mnth9                   3.746      0.481    7.79  2.3e-14 ***
mnth10                  4.387      0.396   11.08  < 2e-16 ***
mnth11                  3.238      0.345    9.38  < 2e-16 ***
mnth12                  1.752      0.333    5.27  1.8e-07 ***
holidayHoliday          3.108      0.416    7.47  2.3e-13 ***
weekday1               -4.325      0.254  -17.03  < 2e-16 ***
weekday2               -4.638      0.248  -18.73  < 2e-16 ***
weekday3               -4.784      0.249  -19.22  < 2e-16 ***
weekday4               -4.678      0.249  -18.82  < 2e-16 ***
weekday5               -3.360      0.248  -13.56  < 2e-16 ***
weekday6                0.570      0.246    2.31  0.02093 *
weathersitModerate     -0.833      0.178   -4.68  3.4e-06 ***
weathersitBad          -4.532      0.463   -9.79  < 2e-16 ***
temp                   -2.571      5.249   -0.49  0.62450
windspeed             -33.757     11.132   -3.03  0.00252 **
hum                    -9.300      3.968   -2.34  0.01938 *
temp:windspeed         82.133     24.508    3.35  0.00085 ***
temp:hum               19.855      8.221    2.42  0.01598 *
windspeed:hum          27.914     18.399    1.52  0.12968
temp:windspeed:hum   -100.361     39.537   -2.54  0.01135 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 702 degrees of freedom
Multiple R-squared:  0.86,	Adjusted R-squared:  0.854
F-statistic:  154 on 28 and 702 DF,  p-value: <2e-16
```

Removed windspeed:hum



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            14.570      1.191   12.23  < 2e-16 ***
yr2012                  1.865      0.135   13.83  < 2e-16 ***
mnth2                   0.549      0.332    1.65  0.09866 .
mnth3                   3.929      0.352   11.18  < 2e-16 ***
mnth4                   4.673      0.391   11.95  < 2e-16 ***
mnth5                   4.511      0.459    9.83  < 2e-16 ***
mnth6                   3.044      0.531    5.73  1.5e-08 ***
mnth7                   2.275      0.586    3.88  0.00011 ***
mnth8                   3.017      0.545    5.53  4.5e-08 ***
mnth9                   3.829      0.478    8.01  4.8e-15 ***
mnth10                  4.452      0.394   11.31  < 2e-16 ***
mnth11                  3.265      0.345    9.46  < 2e-16 ***
mnth12                  1.759      0.333    5.29  1.7e-07 ***
holidayHoliday          3.113      0.416    7.48  2.3e-13 ***
weekday1               -4.317      0.254  -16.99  < 2e-16 ***
weekday2               -4.640      0.248  -18.72  < 2e-16 ***
weekday3               -4.758      0.249  -19.14  < 2e-16 ***
weekday4               -4.666      0.249  -18.77  < 2e-16 ***
weekday5               -3.348      0.248  -13.50  < 2e-16 ***
weekday6                0.556      0.246    2.25  0.02450 *
weathersitModerate     -0.828      0.178   -4.65  4.0e-06 ***
weathersitBad          -4.430      0.458   -9.67  < 2e-16 ***
temp                    4.091      2.879    1.42  0.15577
windspeed             -17.380      2.723   -6.38  3.2e-10 ***
hum                    -3.776      1.580   -2.39  0.01709 *
temp:windspeed         48.140      9.939    4.84  1.6e-06 ***
temp:hum                8.702      3.683    2.36  0.01842 *
temp:windspeed:hum    -43.265     12.130   -3.57  0.00039 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 703 degrees of freedom
Multiple R-squared:  0.859,	Adjusted R-squared:  0.854
F-statistic:  159 on 27 and 703 DF,  p-value: <2e-16
```

Removed temp

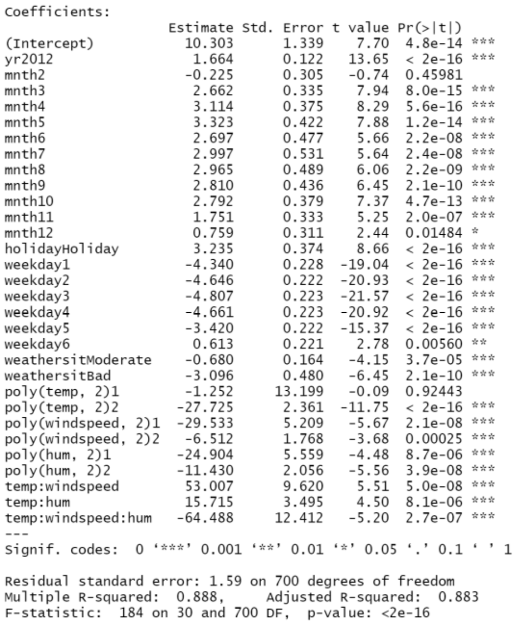


```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)            15.753      0.853   18.47  < 2e-16 ***
yr2012                  1.888      0.134   14.09  < 2e-16 ***
mnth2                   0.573      0.332    1.73  0.08435 .
mnth3                   3.995      0.349   11.46  < 2e-16 ***
mnth4                   4.790      0.382   12.52  < 2e-16 ***
mnth5                   4.658      0.447   10.42  < 2e-16 ***
mnth6                   3.293      0.501    6.57  9.9e-11 ***
mnth7                   2.593      0.542    4.79  2.1e-06 ***
mnth8                   3.259      0.518    6.29  5.7e-10 ***
mnth9                   3.954      0.470    8.41  2.2e-16 ***
mnth10                  4.569      0.385   11.86  < 2e-16 ***
mnth11                  3.313      0.344    9.64  < 2e-16 ***
mnth12                  1.798      0.332    5.42  8.3e-08 ***
holidayHoliday          3.141      0.416    7.55  1.4e-13 ***
weekday1               -4.335      0.254  -17.07  < 2e-16 ***
weekday2               -4.626      0.248  -18.67  < 2e-16 ***
weekday3               -4.734      0.248  -19.08  < 2e-16 ***
weekday4               -4.658      0.249  -18.73  < 2e-16 ***
weekday5               -3.332      0.248  -13.45  < 2e-16 ***
weekday6                0.568      0.246    2.30  0.02147 *
weathersitModerate     -0.845      0.178   -4.75  2.5e-06 ***
weathersitBad          -4.381      0.457   -9.58  < 2e-16 ***
windspeed             -19.385      2.331   -8.32  4.6e-16 ***
hum                    -4.961      1.343   -3.69  0.00024 ***
temp:windspeed         58.120      7.038    8.26  7.3e-16 ***
temp:hum               13.253      1.821    7.28  9.0e-13 ***
temp:windspeed:hum    -52.362     10.311   -5.08  4.9e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.78 on 704 degrees of freedom
Multiple R-squared:  0.859,	Adjusted R-squared:  0.854
F-statistic:  165 on 26 and 704 DF,  p-value: <2e-16
```

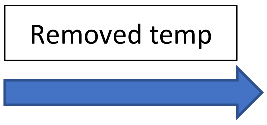
We calculate the Adjusted R-squared to be 0.854 and we will refer to this model as Model 3 for future analysis.

1.6.5.3 Adding both Polynomial Terms and Interaction Terms

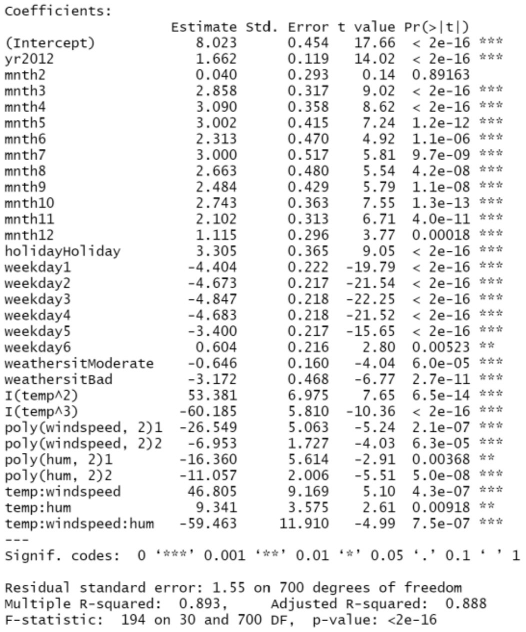
We will try to add both the polynomial and interaction terms to fit our model better to the data. We start by adding interaction terms and polynomial terms between continuous variables – temperature, humidity and windspeed. We use the knowledge obtained from previous analysis to ignore variables with less significance. We then iterate and remove predictors with p-value greater than 0.05. A higher p-value signifies they are not adding any value to the model so therefore, we can remove them, and predictive power of the model is not affected. We repeat this process until all predictors have p-value less then significance level.



```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)            10.303      1.339    7.70  4.8e-14 ***
yr2012                  1.664      0.122   13.65  < 2e-16 ***
mnth2                  -0.225      0.305   -0.74  0.45981    
mnth3                   2.662      0.335    7.94  8.0e-15 ***
mnth4                   3.114      0.375    8.29  5.6e-16 ***
mnth5                   3.323      0.422    7.88  1.2e-14 ***
mnth6                   2.697      0.477    5.66  2.2e-08 ***
mnth7                   2.997      0.531    5.64  2.4e-08 ***
mnth8                   2.965      0.489    6.06  2.2e-09 ***
mnth9                   2.810      0.436    6.45  2.1e-10 ***
mnth10                  2.792      0.379    7.37  4.7e-13 ***
mnth11                  1.751      0.333    5.25  2.0e-07 ***
mnth12                  0.759      0.311    2.44  0.01484 *  
holidayHoliday          3.235      0.374    8.66  < 2e-16 ***
weekday1               -4.340      0.228  -19.04  < 2e-16 ***
weekday2               -4.646      0.222  -20.93  < 2e-16 ***
weekday3               -4.807      0.223  -21.57  < 2e-16 ***
weekday4               -4.661      0.223  -20.92  < 2e-16 ***
weekday5               -3.420      0.222  -15.37  < 2e-16 ***
weekday6                0.613      0.221    2.78  0.00560 ** 
weathersitModerate     -0.680      0.164   -4.15  3.7e-05 ***
weathersitBad          -3.096      0.480   -6.45  2.1e-10 ***
poly(temp, 2)1         -1.252     13.199   -0.09  0.92443    
poly(temp, 2)2        -27.725      2.361  -11.75  < 2e-16 ***
poly(windspeed, 2)1   -29.533      5.209   -5.67  2.1e-08 ***
poly(windspeed, 2)2    -6.512      1.768   -3.68  0.00025 ***
poly(hum, 2)1         -24.904      5.559   -4.48  8.7e-06 ***
poly(hum, 2)2         -11.430      2.056   -5.56  3.9e-08 ***
temp:windspeed         53.007      9.620    5.51  5.0e-08 ***
temp:hum               15.715      3.495    4.50  8.1e-06 ***
temp:windspeed:hum    -64.488     12.412   -5.20  2.7e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.59 on 700 degrees of freedom
Multiple R-squared:  0.888,	Adjusted R-squared:  0.883 
F-statistic:  184 on 30 and 700 DF,  p-value: <2e-16
```



Removed temp



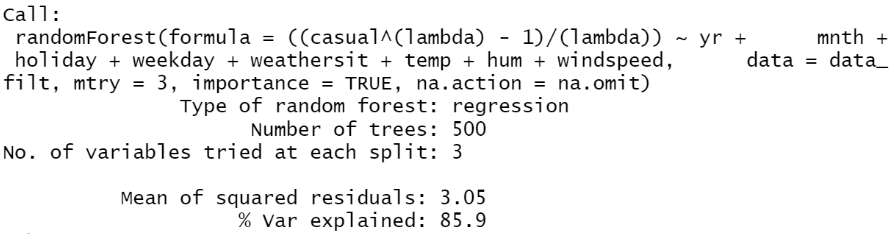
```
Coefficients:
                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)             8.023      0.454   17.66  < 2e-16 ***
yr2012                  1.662      0.119   14.02  < 2e-16 ***
mnth2                   0.040      0.293    0.14  0.89163    
mnth3                   2.858      0.317    9.02  < 2e-16 ***
mnth4                   3.090      0.358    8.62  < 2e-16 ***
mnth5                   3.002      0.415    7.24  1.2e-12 ***
mnth6                   2.313      0.470    4.92  1.1e-06 ***
mnth7                   3.000      0.517    5.81  9.7e-09 ***
mnth8                   2.663      0.480    5.54  4.2e-08 ***
mnth9                   2.484      0.429    5.79  1.1e-08 ***
mnth10                  2.743      0.363    7.55  1.3e-13 ***
mnth11                  2.102      0.313    6.71  4.0e-11 ***
mnth12                  1.115      0.296    3.77  0.00018 ***
holidayHoliday          3.305      0.365    9.05  < 2e-16 ***
weekday1               -4.404      0.222  -19.79  < 2e-16 ***
weekday2               -4.673      0.217  -21.54  < 2e-16 ***
weekday3               -4.847      0.218  -22.25  < 2e-16 ***
weekday4               -4.683      0.218  -21.52  < 2e-16 ***
weekday5               -3.400      0.217  -15.65  < 2e-16 ***
weekday6                0.604      0.216    2.80  0.00523 ** 
weathersitModerate     -0.646      0.160   -4.04  6.0e-05 ***
weathersitBad          -3.172      0.468   -6.77  2.7e-11 ***
I(temp^2)              53.381      6.975    7.65  6.5e-14 ***
I(temp^3)             -60.185      5.810  -10.36  < 2e-16 ***
poly(windspeed, 2)1   -26.549      5.063   -5.24  2.1e-07 ***
poly(windspeed, 2)2    -6.953      1.727   -4.03  6.3e-05 ***
poly(hum, 2)1         -16.360      5.614   -2.91  0.00368 ** 
poly(hum, 2)2         -11.057      2.006   -5.51  5.0e-08 ***
temp:windspeed         46.805      9.169    5.10  4.3e-07 ***
temp:hum                9.341      3.575    2.61  0.00918 ** 
temp:windspeed:hum    -59.463     11.910   -4.99  7.5e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.55 on 700 degrees of freedom
Multiple R-squared:  0.893,	Adjusted R-squared:  0.888 
F-statistic:  194 on 30 and 700 DF,  p-value: <2e-16
```

We calculate the Adjusted R-squared to be 0.888 and we will refer to this model as Model 4 for future analysis.

1.6.5.4 Random Forest

We will try to fit random forest regressor to map the pattern in the data to obtain a model with better predictive power.

As observed from image on right hand side, 500 trees were ensembled together to make the final model.



```
Call:
 randomForest(formula = ((casual^(lambda) - 1)/(lambda)) ~ yr +      mnth +
 holiday + weekday + weathersit + temp + hum + windspeed,      data = data_
filt, mtry = 3, importance = TRUE, na.action = na.omit) 
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 3

          Mean of squared residuals: 3.05
                    % Var explained: 85.9
```

We calculate the Adjusted R-squared to be 0.859 and we will refer to this model as Model 5 for future analysis.

1.6.5.5 Evaluating Performance

We have summarized the Adjusted R-squared obtained from the model fitting as explained above and we choose the best model for evaluating performance and checking for overfitting.

We will use 5-fold cross validation and RMSE to test the performance of the Model 4 on training data. There are 531 datapoints in training set and 220 data points in testing set.

Model	Adjusted R-Squared
Model 1	0.85
Model 2	0.884
Model 3	0.854
Model 4	0.888
Model 5	0.857

The following results are obtained after 5-fold cross validation on Training set and then evaluating final model on testing set:

- RMSE on Training Data: 1013

- RMSE on Testing Data: 1080

As observed, there is not much difference between the training and testing set, therefore we can conclude that our model doesn't overfit the data.

1.7 Results

Using Descriptive statistics and exploratory data analysis we performed hypothesis testing and obtained the following results:

Testing Predictor	p-value	Result
Significance of Temperature	$2\ x\ 10^{-16}$	Reject H_0
Significance of Season	$2\ x\ 10^{-16}$	Reject H_0
Given Temperature, significance of weekday	$2\ x\ 10^{-16}$	Reject H_0

To gain insights into inference of the model, we performed variable selection using forward, backward and stepwise selection and following results were obtained:

Metric	Selection Method	Error Value	Adjusted R-Squared	p (No. of Predictors)	Predictors
AIC	Forward Selection	8606	0.735	9	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum, season
	Backward Selection	8606	0.735	9	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum, season
	Stepwise Selection	8606	0.735	9	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum, season
BIC	Forward Selection	8725	0.716	8	weekday, year, weathersit, temp, holiday, windspeed, hum, season
	Backward Selection	8723	0.733	8	mnth, weekday, year, weathersit, temp, holiday, windspeed, hum
	Stepwise Selection	8725	0.716	8	weekday, year, weathersit, temp, holiday, windspeed, hum, season

Based on the least number of predictors and simplest model we chose highlighted model for our further analysis. For our final model, we tried various algorithms and feature engineering using polynomial terms, interaction terms and random forest regression to obtain the following results on adjusted R-squared:

Model	Adjusted R-Squared
Model 1	0.85
Model 2	0.884
Model 3	0.854
Model 4	0.888
Model 5	0.857

We chose the final model i.e. Model 4 (with interaction and polynomial terms) to evaluate it using 5-fold cross validation using RMSE as evaluation metric. The following results were obtained:

Dataset	*RMSE*
Training Set	1013
Testing Set	1080

1.8 Discussion

We concluded the following from data understanding and exploratory data analysis:

- We removed 'workingday' because of exact collinearity.
- Temp and 'atemp' are highly correlated.

We validated the hypothesis on pattern observed from exploratory data analysis and concluded the following:

- Temperature is a significant predictor of number of casual users.
- Season is a significant predictor of the number of casual users and two season-pairs which show significant difference in average casual users are: Summer to Spring and Fall to Spring.
- Temperature doesn't influence the significance of season on number of casual users.

To work towards our goal of understanding the effect of different predictors on number of casual users we performed Forward, Backward and Stepwise selection using AIC and BIC to get the simplest model with best performance. Depending on least number of predictors and least BIC we chose the following model:

casual ~ yr + mnth + holiday + weekday + weathersit + temp + hum + windspeed

To perform inference, we checked for violation of model assumptions and tried to fix the violated assumptions by performing unusual observation analysis, predictor transformation and response transformation using box-cox methods. We concluded that as model assumptions were not being fulfilled any inference we make would most likely be insignificant.

To work towards our goal of fitting a model to understand the underlying pattern in the data and predict the number of casual users we started with the following model:

(casual^(lambda)-1)/(lambda) ~ yr + mnth + holiday + weekday + weathersit + temp + hum + windspeed

Where lambda = 0.25.

We then added complexity to the model by adding polynomial and interaction terms and evaluated their performance on adjusted R-squared. Similarly, we also used random forest regression to fit the model and test its predictive power by using adjusted R-squared. Finally, we compared all the models and chose the best model as following:

(casual^(lambda)-1)/(lambda)) ~ yr + mnth + holiday + weekday + weathersit + temp*windspeed*hum+ I(temp^2) + I(temp^3) + poly(windspeed, 2) + poly(hum, 2) – hum – windspeed -windspeed:hum - temp

We then performed 5-Fold cross validation to evaluate our final model using RMSE. We obtained training error of 1013 and testing error of 1080 which indicated that our model is not overfitting.

1.8.1 Limitations

- As our data is sampled from time series, the independence assumption for regression may be violated.
- As most of our assumptions were violated even after performing influential point analysis and transformations we cannot completely rely on the inferences generated by model.

Bibliography

Bikeshare, C. (2013). *Capital bikeshare member survey report.* Washington, DC.

DeMaio, P. (2009). Bike-sharing: History, impacts, models of provision, and future. *Journal of public transportation*, 3.

Faghih-Imani, A. E.-G. (2014). How land-use and urban form impact bicycle flows: evidence from the bicycle-sharing system (BIXI) in Montreal. *Journal of transport geography*, 306-314.

Faghih-Imani, A. H. (2017). An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville. *Transportation Research Part A: Policy and Practice*, (pp. 177-191).

Rixey, R. A. (2013). *Station-level forecasting of bikesharing ridership: Station network effects in three US systems.* Transportation research record.

Shaheen, S. A. (2010). *Bikesharing in Europe, the Americas, and Asia: past, present, and future.* Transportation research record.

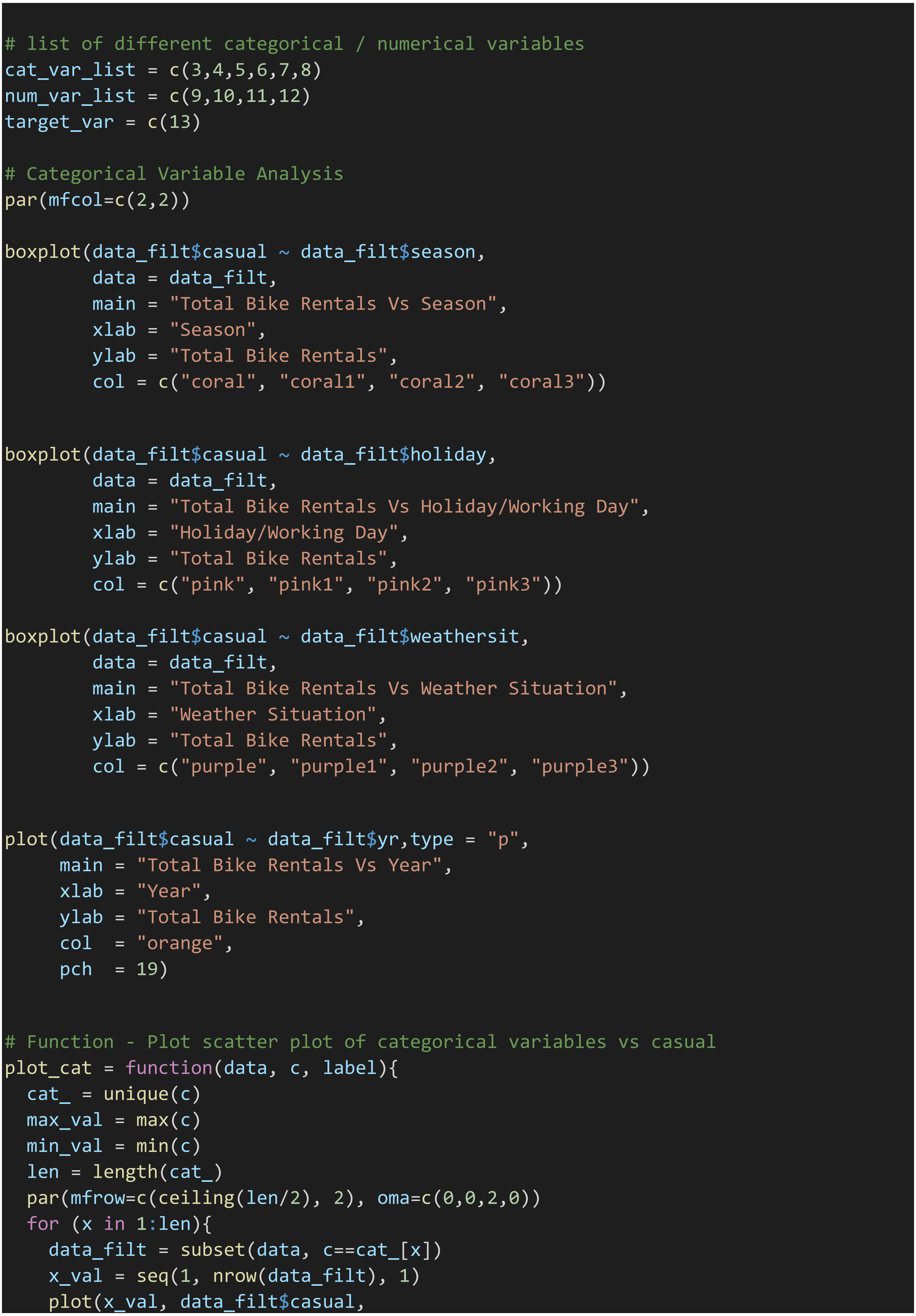
Share, A. B. (2011). *Melbourne bike share survey.* Melbourne: Melbourne: Alta Bike Share.

Wang, X. (., G., S., & JE, H. A. (2012). *Modelling bike share station activity: the effects of nearby businesses and jobs on trips to and from stations.* TRB's 92nd Annual Meeting and Publication in the Transportation Research Record: proceedings.

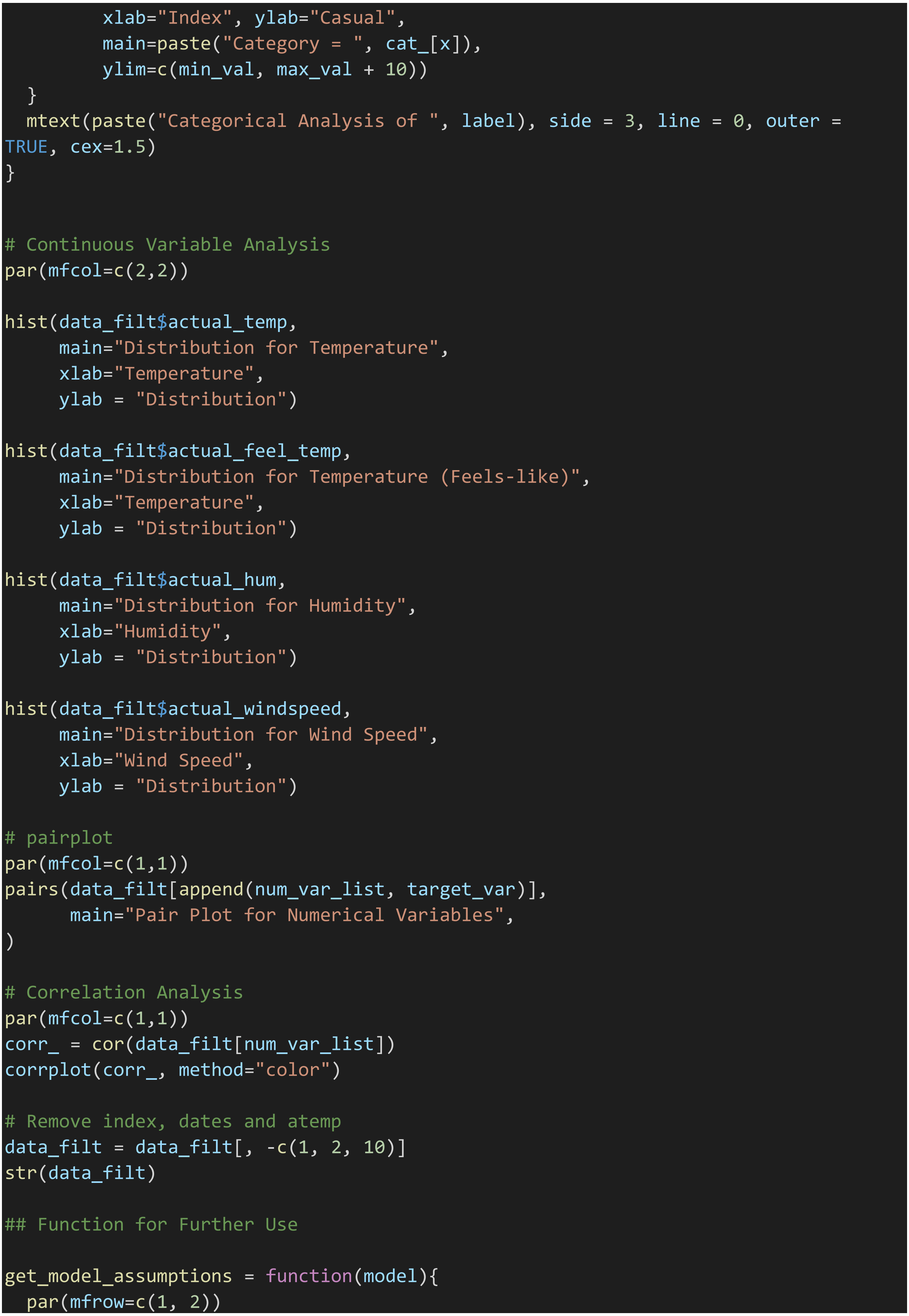
Appendix

```
#######################################
#######################################
## Bike Sharing Data Analysis #########
#######################################
#######################################

# Library Imports
install.packages('corrplot')
install.packages('caret')
install.packages("randomForest")
library(corrplot)
library(faraway)
library(MASS)
library(mgcv)
library(lmtest)
library(car)
library(caret)
library(randomForest)



## Data Exploration
data = read.csv("C:/Users/gurle/OneDrive/Documents/day.csv")
length(data)

str(data)

# Check missing data
colSums(is.na(data))


## Data Transformation

# remove unnecessary variables - workingday, registered and cnt
data_filt = data[, -c(8,15,16)]

# Change categories to more meaningful labels
data_filt$season <- factor(format(data_filt$season, format="%A"),
                           levels = c("1", "2","3","4") ,
                           labels = c("Spring","Summer","Fall","Winter"))

data_filt$holiday <- factor(format(data_filt$holiday, format="%A"),
                            levels = c("0", "1") , labels = c("Working
Day","Holiday"))

data_filt$weathersit <- factor(format(data_filt$weathersit, format="%A"),
                               levels = c("1", "2","3","4") ,
                               labels = c("Good","Moderate","Bad","Worse"))
```

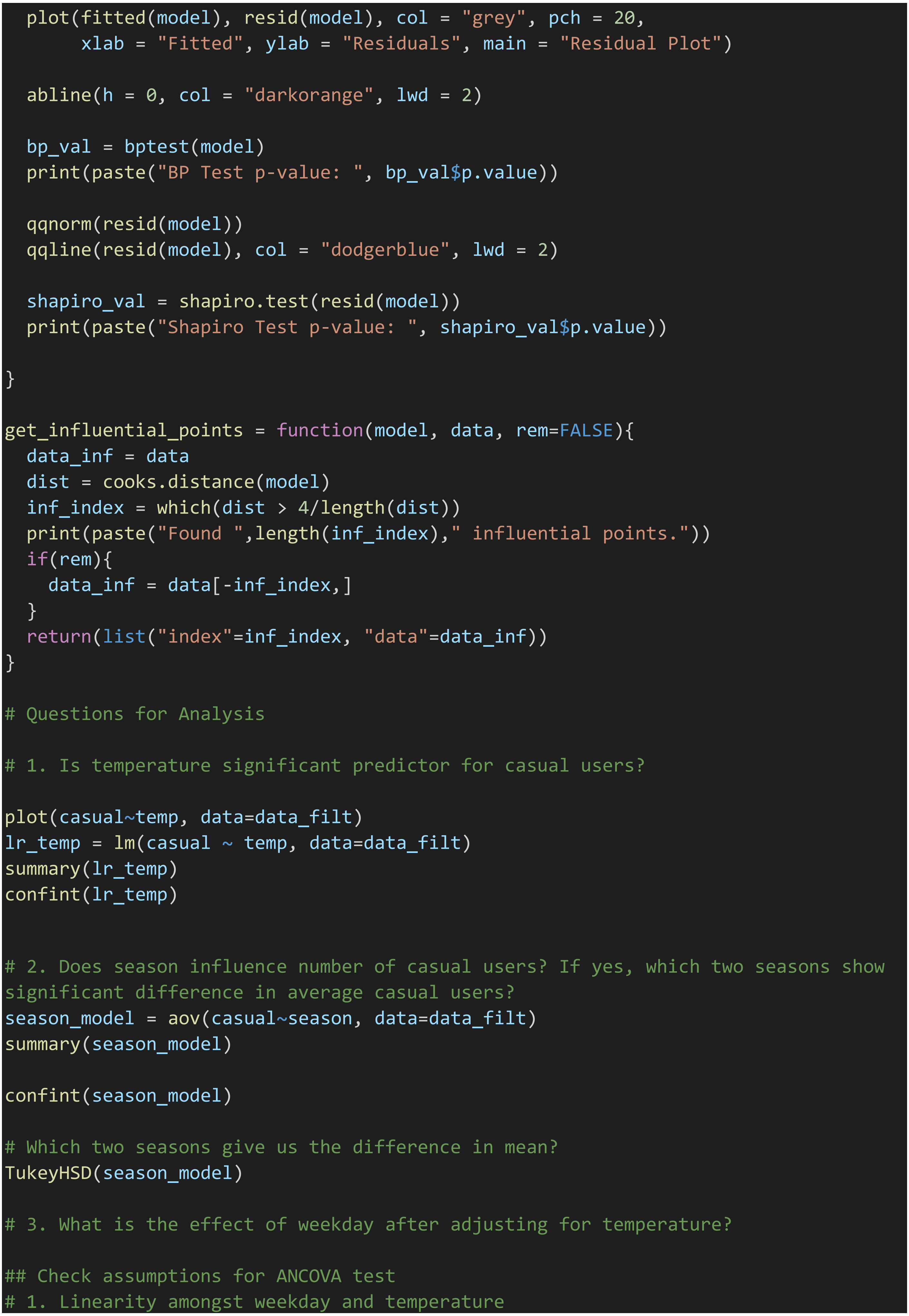
```
data_filt$yr <- factor(format(data_filt$yr, format="%A"),
                       levels = c("0", "1") , labels = c("2011","2012"))


# Add columns with De Normalized value
data_filt$actual_temp <- data_filt$temp*41
data_filt$actual_feel_temp <- data_filt$atemp*50
data_filt$actual_windspeed <- data_filt$windspeed*67
data_filt$actual_humidity <- data_filt$hum*100

# Convert Categorical data to columns
data_filt$mnth = as.factor(data_filt$mnth)
data_filt$weekday = as.factor(data_filt$weekday)
#data_filt$workingday = as.factor(data_filt$workingday)
data_filt$weathersit = as.factor(data_filt$weathersit)

# View final DataFrame and its datatypes
str(data_filt)

## EDA

# Stats of variable
stats = function(var_)
{
  print(paste("Mean: ", mean(var_)))
  print(paste("Variance: ", var(var_)))
  print(paste("S.D.: ", sd(var_)))
  print(paste("Median: ", median(var_)))
  print(paste("Max Value: ", max(var_)))
  print(paste("Min Value: ", min(var_)))
}

stats(data_filt$actual_temp)
stats(data_filt$actual_feel_temp)
stats(data_filt$actual_humidity)
stats(data_filt$actual_windspeed)
stats(data_filt$casual)

par(mfrow=c(1,1))

# Analyse response variable
hist(data_filt$casual,
     main="Distribution for casual",
     xlab="Casual",
     ylab = "Distribution")

boxplot(data_filt$casual,
        main="Distribution for casual",
        xlab="Casual",
        ylab = "Distribution")
```



```
# list of different categorical / numerical variables
cat_var_list = c(3,4,5,6,7,8)
num_var_list = c(9,10,11,12)
target_var = c(13)

# Categorical Variable Analysis
par(mfcol=c(2,2))

boxplot(data_filt$casual ~ data_filt$season,
        data = data_filt,
        main = "Total Bike Rentals Vs Season",
        xlab = "Season",
        ylab = "Total Bike Rentals",
        col = c("coral", "coral1", "coral2", "coral3"))


boxplot(data_filt$casual ~ data_filt$holiday,
        data = data_filt,
        main = "Total Bike Rentals Vs Holiday/Working Day",
        xlab = "Holiday/Working Day",
        ylab = "Total Bike Rentals",
        col = c("pink", "pink1", "pink2", "pink3"))

boxplot(data_filt$casual ~ data_filt$weathersit,
        data = data_filt,
        main = "Total Bike Rentals Vs Weather Situation",
        xlab = "Weather Situation",
        ylab = "Total Bike Rentals",
        col = c("purple", "purple1", "purple2", "purple3"))


plot(data_filt$casual ~ data_filt$yr,type = "p",
     main = "Total Bike Rentals Vs Year",
     xlab = "Year",
     ylab = "Total Bike Rentals",
     col   = "orange",
     pch   = 19)


# Function - Plot scatter plot of categorical variables vs casual
plot_cat = function(data, c, label){
  cat_ = unique(c)
  max_val = max(c)
  min_val = min(c)
  len = length(cat_)
  par(mfrow=c(ceiling(len/2), 2), oma=c(0,0,2,0))
  for (x in 1:len){
    data_filt = subset(data, c==cat_[x])
    x_val = seq(1, nrow(data_filt), 1)
    plot(x_val, data_filt$casual,
```



```
         xlab="Index", ylab="Casual",
         main=paste("Category = ", cat_[x]),
         ylim=c(min_val, max_val + 10))
  }
  mtext(paste("Categorical Analysis of ", label), side = 3, line = 0, outer =
TRUE, cex=1.5)
}


# Continuous Variable Analysis
par(mfcol=c(2,2))

hist(data_filt$actual_temp,
     main="Distribution for Temperature",
     xlab="Temperature",
     ylab = "Distribution")

hist(data_filt$actual_feel_temp,
     main="Distribution for Temperature (Feels-like)",
     xlab="Temperature",
     ylab = "Distribution")

hist(data_filt$actual_hum,
     main="Distribution for Humidity",
     xlab="Humidity",
     ylab = "Distribution")

hist(data_filt$actual_windspeed,
     main="Distribution for Wind Speed",
     xlab="Wind Speed",
     ylab = "Distribution")

# pairplot
par(mfcol=c(1,1))
pairs(data_filt[append(num_var_list, target_var)],
      main="Pair Plot for Numerical Variables",
)

# Correlation Analysis
par(mfcol=c(1,1))
corr_ = cor(data_filt[num_var_list])
corrplot(corr_, method="color")

# Remove index, dates and atemp
data_filt = data_filt[, -c(1, 2, 10)]
str(data_filt)

## Function for Further Use

get_model_assumptions = function(model){
  par(mfrow=c(1, 2))
```

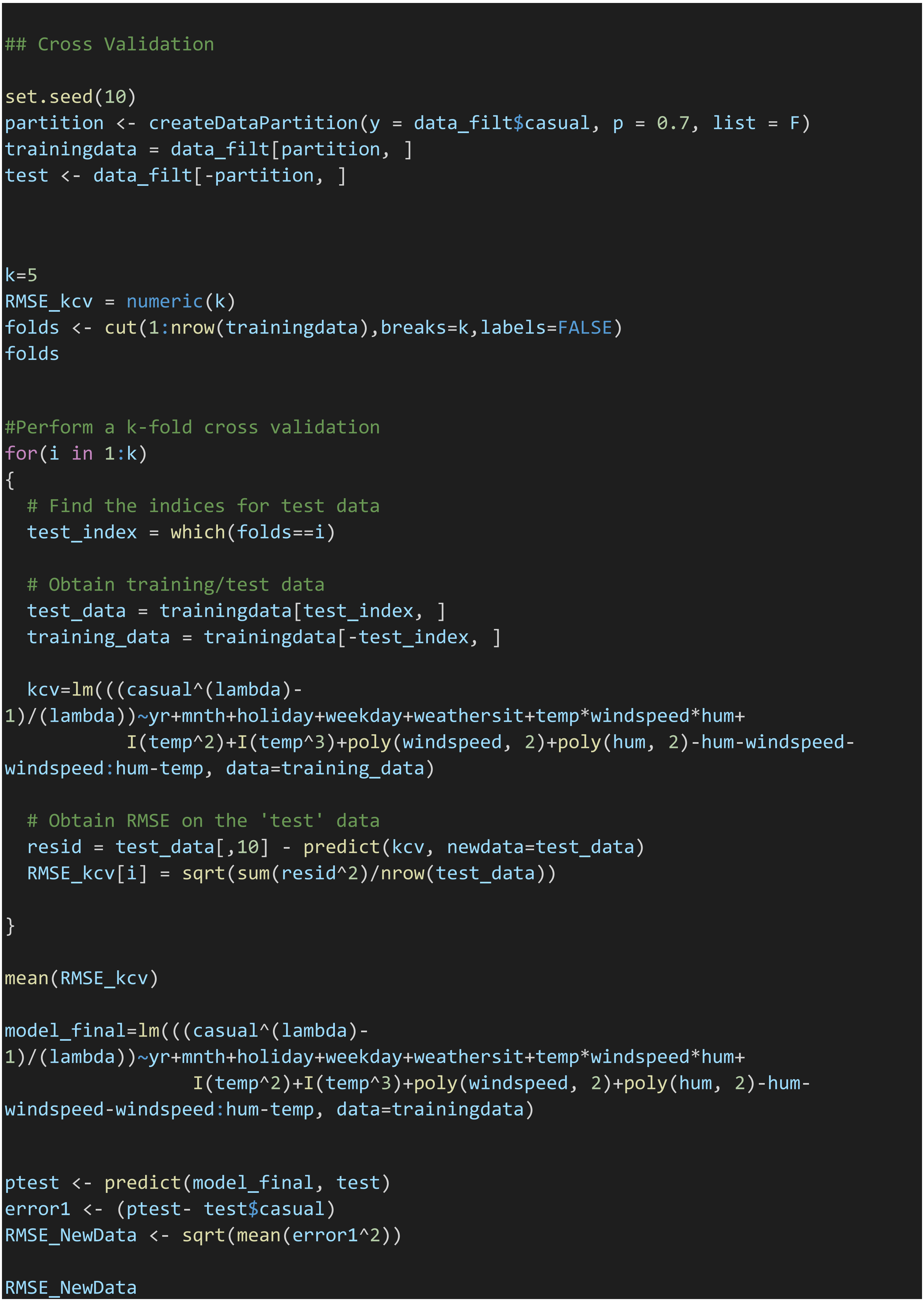


```
  plot(fitted(model), resid(model), col = "grey", pch = 20,
       xlab = "Fitted", ylab = "Residuals", main = "Residual Plot")

  abline(h = 0, col = "darkorange", lwd = 2)

  bp_val = bptest(model)
  print(paste("BP Test p-value: ", bp_val$p.value))

  qqnorm(resid(model))
  qqline(resid(model), col = "dodgerblue", lwd = 2)

  shapiro_val = shapiro.test(resid(model))
  print(paste("Shapiro Test p-value: ", shapiro_val$p.value))

}

get_influential_points = function(model, data, rem=FALSE){
  data_inf = data
  dist = cooks.distance(model)
  inf_index = which(dist > 4/length(dist))
  print(paste("Found ",length(inf_index)," influential points."))
  if(rem){
    data_inf = data[-inf_index,]
  }
  return(list("index"=inf_index, "data"=data_inf))
}

# Questions for Analysis

# 1. Is temperature significant predictor for casual users?

plot(casual~temp, data=data_filt)
lr_temp = lm(casual ~ temp, data=data_filt)
summary(lr_temp)
confint(lr_temp)


# 2. Does season influence number of casual users? If yes, which two seasons show
significant difference in average casual users?
season_model = aov(casual~season, data=data_filt)
summary(season_model)

confint(season_model)

# Which two seasons give us the difference in mean?
TukeyHSD(season_model)

# 3. What is the effect of weekday after adjusting for temperature?

## Check assumptions for ANCOVA test
# 1. Linearity amongst weekday and temperature
```

```
temp_weekday_relation = aov(temp ~ weekday, data=data_filt)
summary(temp_weekday_relation)

# 2. Check variance amongst weekday and temperature
leveneTest(temp ~ weekday, data = data_filt)

# Running ANCOVA test
model = lm(casual ~ weekday + temp, data=data_filt)
summary(model)
Anova(model, type=3)


############## END OF DESCRIPTIVE #######################

# Model Selection

# AIC Backward
fit_all = lm(casual~ season+yr+mnth+holiday+weekday+weathersit+temp+hum+windspeed,
data=data_filt)
summary(fit_all)
fit_back_aic = step(fit_all, direction = "backward")
summary(fit_back_aic)

# AIC Forward
fit_null = lm(casual~1, data=data_filt)

fit_forw_aic = step(fit_null,
scope=casual~season+yr+mnth+holiday+weekday+weathersit+temp+hum+windspeed,
direction="forward")
summary(fit_forw_aic)

# AIC Stepwise
fit_step_aic = step(fit_null,
scope=casual~season+yr+mnth+holiday+weekday+weathersit+temp+hum+windspeed,
direction="both")
summary(fit_step_aic)

n = nrow(data_filt)

# BIC Backward
fit_back_bic = step(fit_all, direction = "backward", k=log(n))
summary(fit_back_bic)

# BIC Forward
fit_forw_bic = step(fit_null,
scope=casual~season+yr+mnth+holiday+weekday+weathersit+temp+hum+windspeed,
direction="forward", k=log(n))
summary(fit_forw_bic)
# BIC Stepwise
```

```
fit_step_bic = step(fit_null,
scope=casual~season+yr+mnth+holiday+weekday+weathersit+temp+hum+windspeed,
direction="both", k=log(n))
summary(fit_step_bic)


# Multi-collinearity Check
vif(fit_back_bic)

get_model_assumptions(fit_back_bic)

# Check for reasons for assumption violation

## Remove influential points
par(mfrow=c(1, 1))
plot(fit_back_bic, which=4)
inf_ob = get_influential_points(fit_back_bic, data_filt, rem=TRUE)
data_inf_rem = inf_ob$data
ind_inf = inf_ob$index




fit_inf = lm(casual ~ yr + mnth + holiday + weekday + weathersit +
               temp + hum + windspeed,data=data_inf_rem)
summary(fit_inf)

get_model_assumptions(fit_inf)

## Transform Y
#### Specify the range of lambda
par(mfcol=c(1,1))
boxcox(fit_back_bic, lambda = seq(0, 0.5, by = 0.05))

lambda = 0.25
fit_tr_y <- lm(((casual^(lambda)-1)/(lambda))~yr + mnth + holiday + weekday +
weathersit +
                 temp + hum + windspeed,data=data_filt)
summary(fit_tr_y)

get_model_assumptions(fit_tr_y)

## Transform X
################## TO DO ############


# Splines
lambda = 0.25
fit_spline <- gam(((casual^(lambda)-1)/(lambda))~yr + mnth + holiday + weekday +
weathersit+s(temp)+s(hum)+s(windspeed),data=data_filt)
summary(fit_spline)
```

```r
get_model_assumptions(fit_spline)



# Polynomial terms
poly_model <- lm(((casual^(lambda)-1)/(lambda))~yr + mnth + holiday + weekday + 
weathersit + 
                   poly(temp, 3)+poly(windspeed, 3)+poly(hum, 3), data=data_filt)
summary(poly_model)
poly_model <- lm(((casual^(lambda)-1)/(lambda))~yr + mnth + holiday + weekday + 
weathersit + 
                   poly(temp, 3)+poly(windspeed, 2)+poly(hum, 2), data=data_filt)
summary(poly_model)


# Interaction
interaction_model<- lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum, 
data=data_filt)
summary(interaction_model)
interaction_model<- lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum-windspeed:hum, 
data=data_filt)
summary(interaction_model)
interaction_model<- lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum-windspeed:hum-
temp, data=data_filt)
summary(interaction_model)

# Interaction and Polynomial term
poly_interaction_model <- lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum+
                               poly(temp, 2)+poly(windspeed, 2)+poly(hum, 2)-hum-
windspeed-windspeed:hum-temp, data=data_filt)
summary(poly_interaction_model)

poly_interaction_model <- lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum+
                             I(temp^2)+I(temp^3)+poly(windspeed, 2)+poly(hum, 2)-
hum-windspeed-windspeed:hum-temp, data=data_filt)
summary(poly_interaction_model)



## Random Forest
# Create random forest for regression
rf <- randomForest(((casual^(lambda)-1)/(lambda))~yr + mnth + holiday + weekday + 
weathersit + 
                     temp + hum + windspeed,data=data_filt, mtry = 3,
                         importance = TRUE, na.action = na.omit)
rf
```



```
## Cross Validation

set.seed(10)
partition <- createDataPartition(y = data_filt$casual, p = 0.7, list = F)
trainingdata = data_filt[partition, ]
test <- data_filt[-partition, ]



k=5
RMSE_kcv = numeric(k)
folds <- cut(1:nrow(trainingdata),breaks=k,labels=FALSE)
folds


#Perform a k-fold cross validation
for(i in 1:k)
{
  # Find the indices for test data
  test_index = which(folds==i)

  # Obtain training/test data
  test_data = trainingdata[test_index, ]
  training_data = trainingdata[-test_index, ]

  kcv=lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum+
           I(temp^2)+I(temp^3)+poly(windspeed, 2)+poly(hum, 2)-hum-windspeed-
windspeed:hum-temp, data=training_data)

  # Obtain RMSE on the 'test' data
  resid = test_data[,10] - predict(kcv, newdata=test_data)
  RMSE_kcv[i] = sqrt(sum(resid^2)/nrow(test_data))

}

mean(RMSE_kcv)

model_final=lm(((casual^(lambda)-
1)/(lambda))~yr+mnth+holiday+weekday+weathersit+temp*windspeed*hum+
                 I(temp^2)+I(temp^3)+poly(windspeed, 2)+poly(hum, 2)-hum-
windspeed-windspeed:hum-temp, data=trainingdata)


ptest <- predict(model_final, test)
error1 <- (ptest- test$casual)
RMSE_NewData <- sqrt(mean(error1^2))

RMSE_NewData
```