THE INTERNATIONAL JOURNAL OF HUMANITIES & SOCIAL STUDIES

Shared Bike Usage Analysis

Kangyao Xia
Student, Department of Finance, Johns Hopkins University, China

Abstract:
This essay delves into the realm of data analytics to examine patterns and trends in shared bike usage. Leveraging large-scale datasets from shared bike platforms, the analysis encompasses factors influencing usage, including geographical patterns, temporal variations, and user demographics. The essay employs advanced data analytics techniques, such as machine learning algorithms, to uncover hidden insights and predict future usage trends. By scrutinizing the data, the essay aims to contribute valuable insights for optimizing shared bike systems, enhancing urban mobility, and informing sustainable transportation policies.

Keywords: *Data analytics, bike-sharing business, linear regression model, cnt, dataset, Portugal*

1. Introduction

The dataset we selected describes a bike-sharing business that has newly emerged in Portugal. A bike-sharing system is a new form of traditional bike rental business that integrates the use of several information technologies so that it dramatically increases efficiency during the rental process and reduces a huge amount of labour costs. The business runs using a membership program and also allows casual users to rent a bike for single-time use. Members will have to register personal and payment information with the company in order to obtain a membership number and card. During the rental process, it takes only a few seconds to begin the rental session for a registered member, and when they return the bike to any station, the payment process will be automatically done through the cloud center. The new system dramatically reduces the direct labor needed throughout the entire rental process and saves both parties a large amount of time. Until 2012, there were more than 500 bike-sharing companies, with over 500 thousand bikes sharing around the world. Today, modern bike-sharing systems have demonstrated their important role in solving city traffic problems.

2. Questions, Hypothesis and Objectives

Before we started to take a look at the dataset and run any analysis on it, based on common sense, we thought that the factors that would have the most impact on using the rental bike would be temperature and weather conditions. Because normally people would tend to take public transportations during bad weather conditions and lower temperatures. We also assume bad weather conditions and lower temperatures would have a more severe impact on casual users than on registered users.

We are trying to use data analytic skills to answer a few questions:

- Does there exist an explainable relationship between "cnt" variable (count of total rental bikes including both casual and registered) and other variables in the dataset, such as temperature, humidity, weekdays, etc.?
- If yes, how many variables are statistically significant enough to have an effect on the number of shared bikes used?
- What are these statistically significant factors, and to what extent do they affect the number of shared bikes used?
- If we set up a regression model based on these statistically significant factors, how well would the model perform?
- If the models we set up perform well enough, can we make the model more concise by dropping some of the factors and still get a similar output while at the same time making the model more reasonable for the real situation?

Our objective in this project is to find the best-fitting model with sufficient reliability to explain and give a reasonable prediction on critical factors that will have a major impact on the use of shared bikes. We are hoping that based on our analysis and the prediction model, there will be some useful insights which can facilitate the strategic and business planning for shared bike systems/companies.

3. Data Description

The datasets we selected were collected by Hadi Fanaee-T at the University of Porto. The data contains daily counts of rental bikes between 2011 and 2012 in the capital bike-share system in Portugal, along with the corresponding weather and seasonal information. The dataset contains 731 observations, which represents 731 days of the count for bike use. Each observation contains 15 variables, including date, season, month, holiday, weather conditions, temperature, humidity and others.

For weather conditions, variable weathersits are:
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

For the season variable, the dataset assigns a number to each season: 1 for spring, 2 for summer, 3 for fall and 4 for winter. Month and weekday variables follow the same logic as the season. There are two dummy variables in the dataset: holiday and working day. For the working day variable, 1 shows if a day is neither a weekend nor a holiday; otherwise, it is 0.

After examining the data set, we found that the index column and year had no effect on our analysis. Then we found that month is a more detailed method than season to classify a time during a year. Further, we found that the working day is 1 when the weekday is 6 and 0, so we dropped the weekday column. Finally, "atemp" is the feeling temperature it has a high correlation with the temperature, so we delete "atemp" column as well. As a result, we did not take these five columns into consideration during the analysis.

Since the monthly data are 12 numbers assigned by the collection of this dataset, after running a regression for the number set from 1 to 12, the coefficient we get does not make sense. For variable month to work properly, we believe we have to divide individual months into 12 separate dummy variables so that each month's input multiplied by the coefficient would make sense on the effect of total use of bikes during that day. Because we needed to modify the raw dataset and add another 12 variables into the dataset, making our regression model too complex, we decided to discard the variable month.

For our objective variable cnt, there exist a subset contains two variables casual (count of casual users) and registered (count of registered users). These two variables add up to the final amount in cnt in each observation. To analyze these two variables, we will need to run separate regressions against other variables following the same logic and steps we did when analyzing the cnt variable. Since our objective in this project is to study the factors that influence the total number of shared bikes used, we decided to drop these two variables. However, for more thought, if we have a chance to run separate regressions on these two variables, the result may help businesses make marketing decisions based on different approaches towards these two different target groups.

4. Methodology

4.1. Linear Regression Model

4.1.1. Graphic Building with Pairs () and Plot ()

After inputting the csv file in the R, we first run the command pairs () to get the visualized output which could show the possible relationship between cnt and other variables.

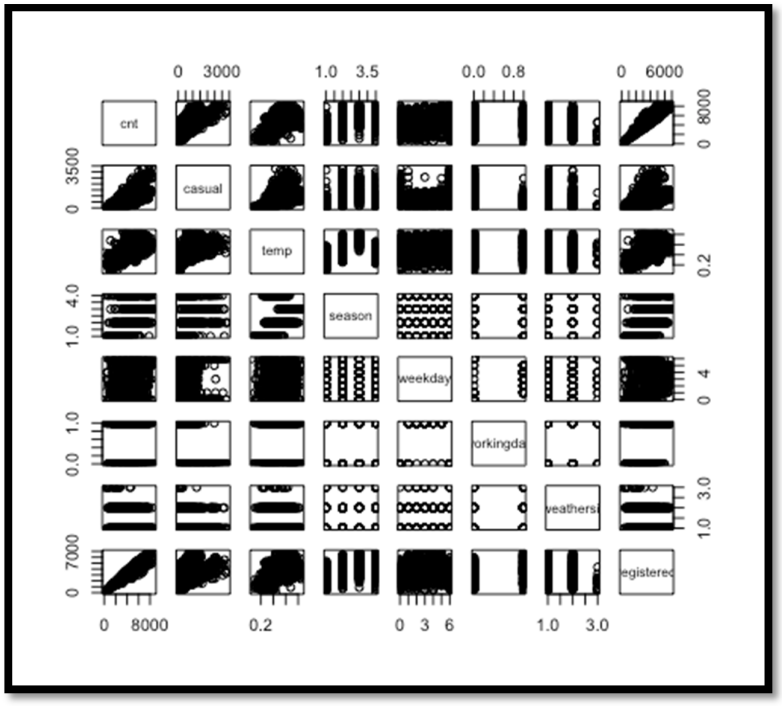


Figure 1

4.1.2. Regression

After seeing the plots, we then run the linear regression in which cnt is the response variable and temp, holiday, working day, weathers, hum; windspeed are the independent variables.

From the summary of this linear regression, we found that the p-values of the variables temp, weathersit, hum, and windspeed are small, which proves their significance. Holidays have a P-value of 0.068, and working days are not significant at all.

```
Call:
lm(formula = cnt ~ temp + weathersit + hum + windspeed + workingday +
    holiday, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4087.5 -1084.5   -99.6  1028.7  3848.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3971.9      343.1  11.578  < 2e-16 ***
temp          6333.4      297.7  21.274  < 2e-16 ***
weathersit    -503.2      125.6  -4.005 6.85e-05 ***
hum          -1850.1      493.4  -3.750 0.000191 ***
windspeed    -4185.3      717.5  -5.833 8.21e-09 ***
workingday     106.7      116.2   0.917 0.359193
holiday       -588.7      322.5  -1.825 0.068382 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1409 on 724 degrees of freedom
Multiple R-squared:  0.4756,    Adjusted R-squared:  0.4713
F-statistic: 109.4 on 6 and 724 DF,  p-value: < 2.2e-16
```

Figure 2

4.1.3. Refined Linear Regression

Then, we deduct the working day from the regression model, and we get an R-square equal to 0.475, and all the variables are significant. However, the P-value of the holiday is 0.034, which is not as small as other factors.

Thus, we deduct holidays to see whether the R-square decreases on a large scale. At this time, we get an R square of 0.4717, and all other factors, including temp, weathersit, hum, and windspeed, are significant. We can see that the holiday does not contribute to the total R square, so we also deduct it from the model.

The R-square is 0.4717, which we think is good enough for a real case.

4.2. Prediction Model Comparison (Regsubsets & Tree)

4.2.1. Regsubsets Method

After getting the results from the regression model, we first use the regsubsets function to test the variables we choose. From the result, we can see that if we choose 4 variables, then we would choose temp, weathersit, hum and windspeed, which is the same as our conclusion.

```
         temp weathersit hum windspeed holiday workingday
1  ( 1 ) "*"  " "        " " " "       " "     " "
2  ( 1 ) "*"  "*"        " " " "       " "     " "
3  ( 1 ) "*"  "*"        " " "*"       " "     " "
4  ( 1 ) "*"  "*"        "*" "*"       " "     " "
5  ( 1 ) "*"  "*"        "*" "*"       "*"     " "
6  ( 1 ) "*"  "*"        "*" "*"       "*"     "*"
```

Figure 3

4.2.2. Tree Method

Besides the linear regression, we also tried to form the tree regression to find the prediction model. Using variables similar to those we use for linear regression, we get the following output.

Later, we decided to compare the tree regression model with the linear regression model to find which model could provide a more concise prediction.

We use set.seed(1) and sample () to set 50% of the dataset as the training data.

Then, with this training set, we re-run the tree regression command and use this model to predict the cnt value of the remaining 50% of the dataset (testing data).

Later, we compute the error between the prediction value and the actual value.

With the same training data and testing data, we then use the linear regression model to get the predictions and then compute the error.

5. Result

From several trials of linear regression, we finally found the most appropriate linear regression model for cnt. The cnt is indeed impacted by the weather condition. The main results are the following:

- Negatively impact the daily usage of the shared bike: weathersit, humidity, and windspeed. The worse the weather is, the higher the humidity and the higher the windspeed is, the fewer people are likely to use the shared bike.
- Positively impact the daily usage of the shared bike: temperature. The higher the temperature, the more people use the shared bikes.
- The regression tree can provide a more concise prediction than the linear regression model.

6. Conclusion

Mining the data from the bike-sharing system can bring many insights. It enables us to learn more about people's travel patterns. Additionally, the data could inform operators and policymakers about the maintenance of a suitable balance of bicycles throughout the system area or help them get better targeting of promotional materials to encourage frequent usage.

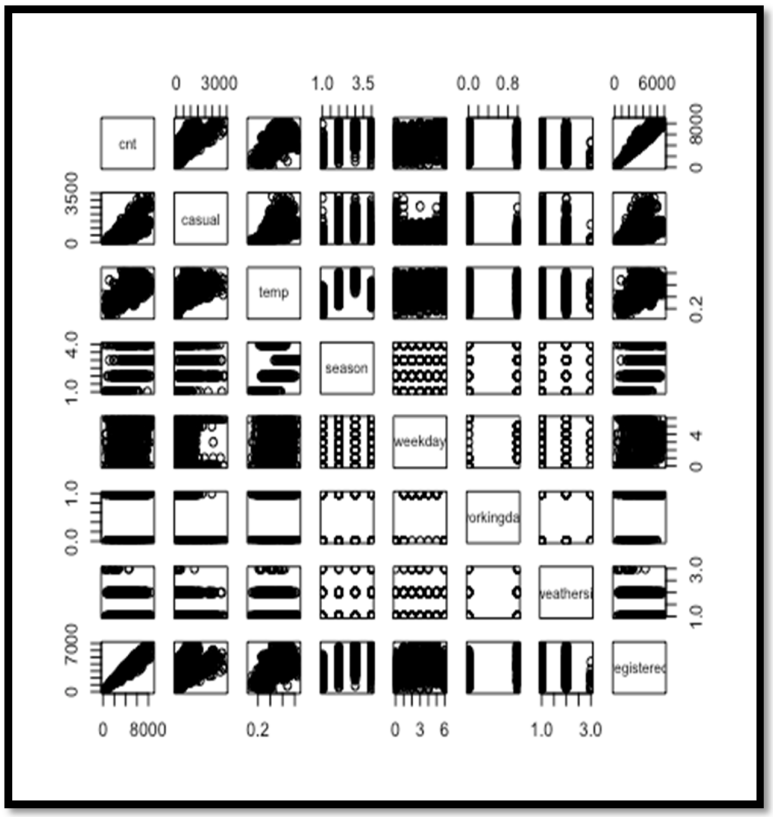
To get more useful insights, further data collection and analysis are required. For example, variables such as the number of total rental bikes per hour, cycling distance/time, and the location of the bike's docking station (near the metro, near the restaurant, near the business central district) can be included.

7. References

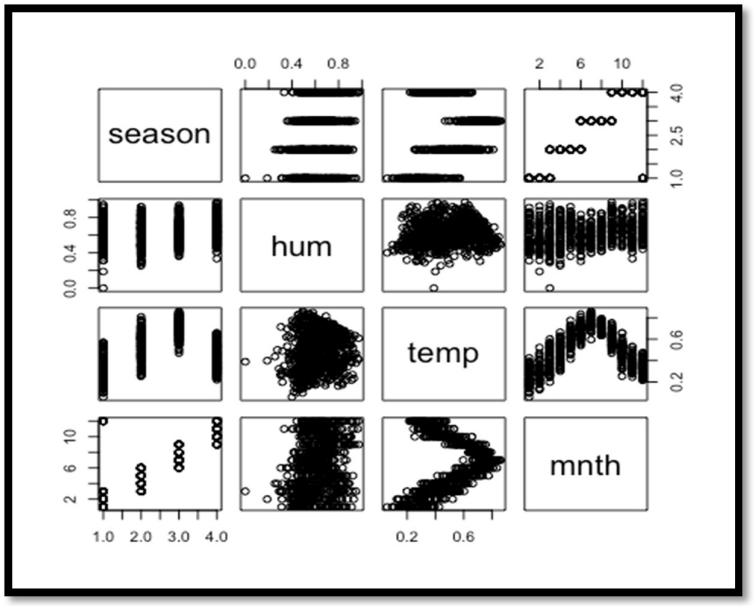
i. Guo Y, Zhou J, Wu Y, Li Z (2017). Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China. *PLoS ONE 12(9): e0185100.* https://doi.org/10.1371/ journal.pone.0185100

ii. O'Brien. O & Cheshire. J & Batty.M (2014). Mining bicycle sharing data to generate insights into sustainable transport systems. *Journal of Transport Geography Volume 34, January 2014, Pages 262–273.*

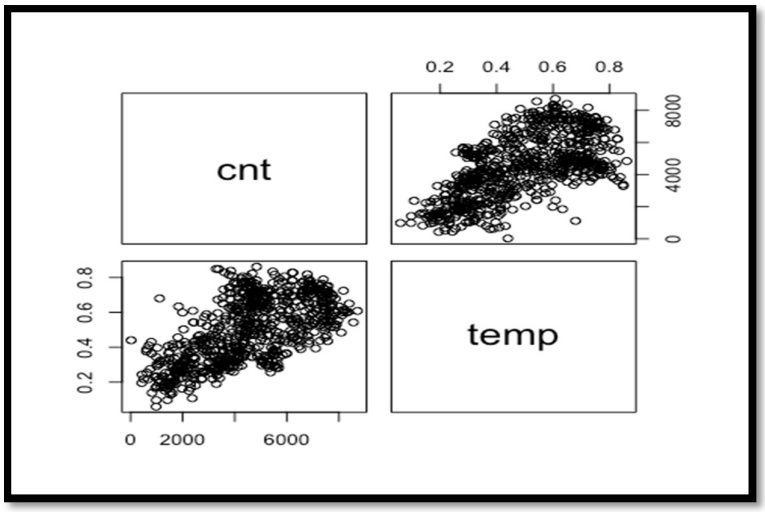
Appendices



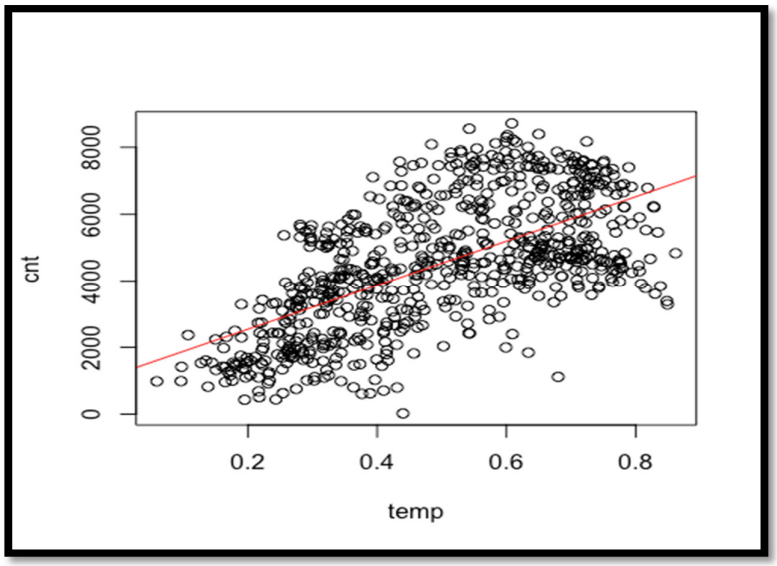
Appendix 1



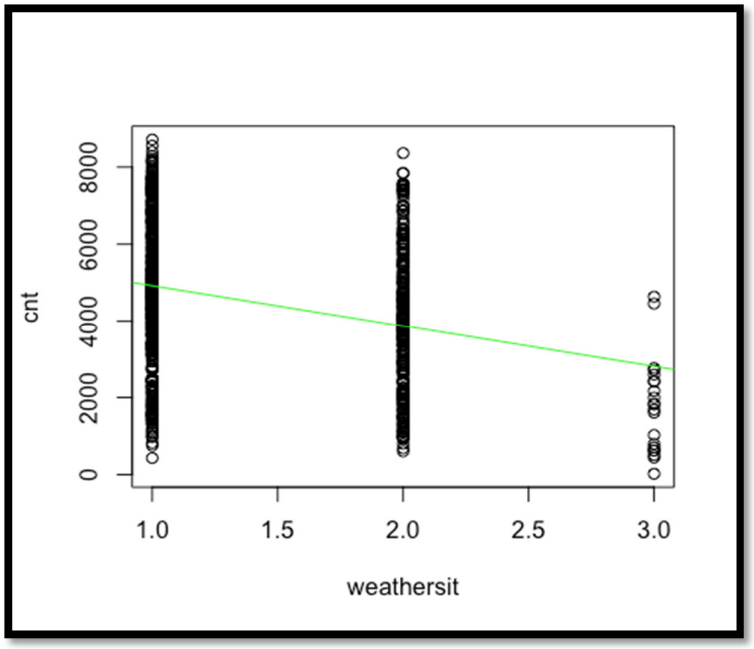
Appendix 2



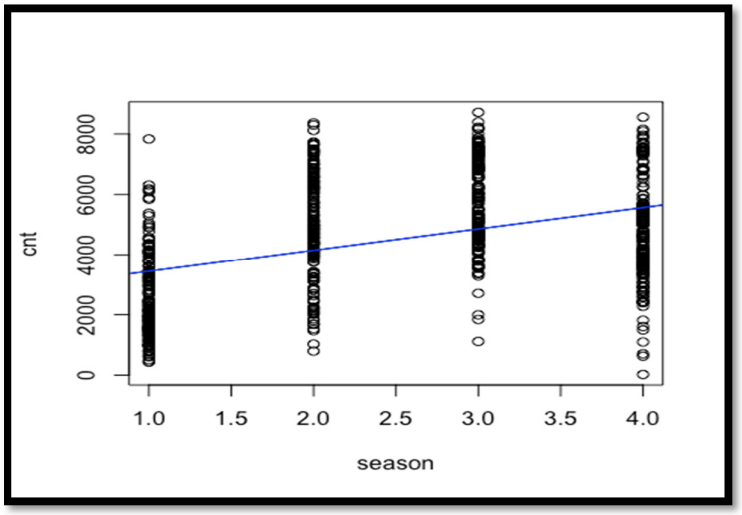
Appendix 3



Appendix 4



Appendix 5



Appendix 6

```
Call:
lm(formula = cnt ~ temp + weathersit + hum + windspeed + workingday +
    holiday, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4087.5 -1084.5   -99.6  1028.7  3848.3

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   3971.9      343.1  11.578  < 2e-16 ***
temp          6333.4      297.7  21.274  < 2e-16 ***
weathersit    -503.2      125.6  -4.005 6.85e-05 ***
hum          -1850.1      493.4  -3.750 0.000191 ***
windspeed    -4185.3      717.5  -5.833 8.21e-09 ***
workingday     106.7      116.2   0.917 0.359193
holiday       -588.7      322.5  -1.825 0.068382 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1409 on 724 degrees of freedom
Multiple R-squared:  0.4756,    Adjusted R-squared:  0.4713
F-statistic: 109.4 on 6 and 724 DF,  p-value: < 2.2e-16
```

Appendix 7

```
Call:
lm(formula = cnt ~ temp + weathersit + hum + windspeed + holiday,
    data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4167.8 -1078.5   -92.7  1048.1  3770.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4040.4      334.8  12.069  < 2e-16 ***
temp          6349.1      297.2  21.365  < 2e-16 ***
weathersit    -495.3      125.3  -3.952 8.51e-05 ***
hum          -1865.3      493.1  -3.783 0.000168 ***
windspeed    -4199.5      717.3  -5.855 7.24e-09 ***
holiday       -662.6      312.3  -2.122 0.034195 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1408 on 725 degrees of freedom
Multiple R-squared:  0.475,     Adjusted R-squared:  0.4714
F-statistic: 131.2 on 5 and 725 DF,  p-value: < 2.2e-16
```

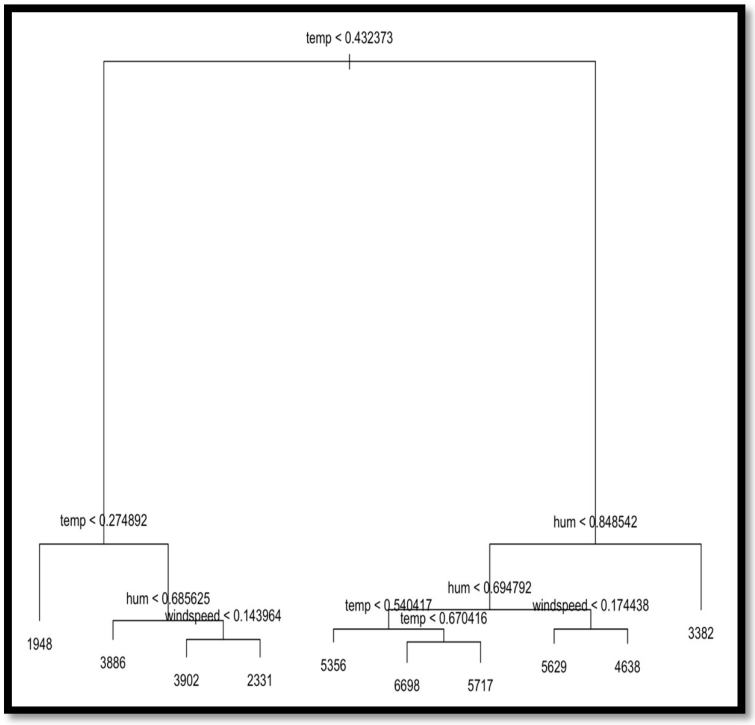
Appendix 8

```
Call:
lm(formula = cnt ~ temp + weathersit + hum + windspeed, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4155.2 -1086.1   -98.3  1054.1  3783.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4008.4      335.3  11.956  < 2e-16 ***
temp          6371.0      297.7  21.400  < 2e-16 ***
weathersit    -484.8      125.5  -3.862 0.000123 ***
hum          -1881.8      494.2  -3.808 0.000152 ***
windspeed    -4210.8      719.0  -5.857 7.16e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1412 on 726 degrees of freedom
Multiple R-squared:  0.4717,    Adjusted R-squared:  0.4688
F-statistic: 162.1 on 4 and 726 DF,  p-value: < 2.2e-16
```

Appendix 9

```
         temp weathersit hum windspeed holiday workingday
1  ( 1 ) "*"  " "        " " " " " "  " "     " "
2  ( 1 ) "*"  "*"        " " " " " "  " "     " "
3  ( 1 ) "*"  "*"        " " " " "*"  " "     " "
4  ( 1 ) "*"  "*"        "*" "*" "*"  " "     " "
5  ( 1 ) "*"  "*"        "*" "*" "*"  "*"     " "
6  ( 1 ) "*"  "*"        "*" "*" "*"  "*"     "*"
```

Appendix 10


temp < 0.432373
temp < 0.274892
hum < 0.848542
hum < 0.685625
windspeed < 0.143964
hum < 0.694792
temp < 0.540417
temp < 0.670416
windspeed < 0.174438
1948
3886
3902
2331
5356
6698
5717
5629
4638
3382

Appendix 11