Bike Rentals Prediction

Somnath Mahato

September 26, 2018

Contents

Chapter 1

Introduction

1.1 Problem Statement

The objective is to forecast bike rental demand of Bike sharing program in Washington, D.C based on historical usage patterns in relation with weather, environment and other data. We would be interested in predicting the rentals on various factors including season, temperature, weather and building a model that can successfully predict the number of rentals on relevant factors.

1.2 Data

This dataset contains the seasonal and weekly count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding temperature and humidity information. Bike sharing systems are a new way of traditional bike rentals. The wohle process from memberhsip to rental and retrun back has become automatic. The data was generated by 500 bike-sharing programs and was collected by the Laboratory of Artificial Intelligence and Decision Support (LIAAD), University of Porto. Given below is the description of the data which is a (731, 16) shaped data, The variables are:

The table represent the features used in the training and analysis

Table 1.1: Variables and Descriptions

No.	Features	Description
1	instant	Daily customer index
2	dteday	The date index for both years
3	season	Season type 1-Springer, 2-Summer, 3- Fall, 4-Winter
4	year	The year 0-2011, 1-2011
5	month	The months 1-12
6	holiday	0 - Not Holiday, 1- Holiday
7	weekday	Weekdays 0(Monday) - 6(Sunday)
8	workingday	0 - Workingday, 1- Not a Workingday
9	weathersit	Weather type 1- Clear, 2-Misty, 3- Light Rain, 4-Heavy Rain
10	temp	Normalized value of temperatures at every instant
11	atemp	Normalized value of the absolute temperature
12	humidity	Contains the normalized value for the humidity
13	windspeed	contains the normalized value for the windspeed
14	casual	has the number of unregistered users at a given day
15	registered	has the number of registered users at a given day
16	count	Total Rentals with both casual and registered users

count will be used as response variable here, and all other as predictor.

1.3 Data Cleaning

A preliminary data cleaning is performed, converting ”holiday”, ”workingday”, ”weather” to factors to better represent their categorical nature. I only keep the ”temp” variable and removed ”atemp” variable since it is almost repetitive and not a relatively accurate statistic to acquire. I also remove the ”casual” and ”registered” variable from the dataset because they sum up to ”count” and my analysis later will not use them.

1.4 Data Engineering

I have converted the normalized variables such as ”Temp”, ”Windspeed”, ”Humidity” to their de-normalized values to put the analysis on their actual scale. The normalized temperature is calculated as

$$n = \frac{t - tmin}{tmax - tmin}$$

where, tmin = -8 and tmax = 39 and the actual temperature is calculated as,

$$t - tmin = n * tmax - tmin$$
$$t = (n * tmax - tmin) + tmin$$

where t is the actual temperature

$$De - normaliztionforwindspeedcanbegivenas$$
$$norm.windspeed = \frac{windspeed}{maxspeed}$$
$$given, maxspeed = 67$$
$$denorm.windspeed = windspeed * maxspeed$$

$$De - normaliztionforhumidityisgivenas$$
$$norm.humidity = \frac{humidity}{maxhumidity}$$
$$given, maxhumidity = 100$$
$$denorm.humidity = humidity * maxhumidity$$

New variable for daytype which is obtained from workingday and holiday as follows:
if not workingday and it's not a holiday then, its a weekend in daytype
if its' workingday and there is a holiday then, its a holiday in daytype
if its' working day and there is a holiday then, its a workingday in daytype

New variable for temperature type obtained through binning the continuos temperature values,
Temperature between 2-10 is 'Cold Temperature'
Temperature between 10-20 is 'Moderate Temperature'
Temperature between 20-30 is 'Hot Temperature'

Hence, the new variables are:
1. Raw Temperature
2. Raw Windspeed
3. Raw Humidity
4. Day Type
5. Temperature Type

Chapter 2

Methodology

2.1 Exploratory Data Analysis

Without building any model or making any predictions, lets first look at the data by itself.

By looking at data I came across that data is without any missing values however, casual user variable has outliers in it. Visualizations of the bike rental count base on the season, month, day of the week, the type of day, is it a weekday, is it a holiday, and the type of weather, then calculating the mean of temperature, humidity, wind speed and rental count. The purpose of this summarization is to find a general relationship between variables regardless of which year the data is from and hypothesizing the relations with hypothesis validation.

2.1.1 Outlier Analysis

While plotting the outlier using boxplot for all the variables even, most of the variables are static, we came across that number of casual users are increased in several times. We replaced these outliers with their capped values that is replacing the extreme outliers with their 95 quartile value. The boxplot can be as represented,

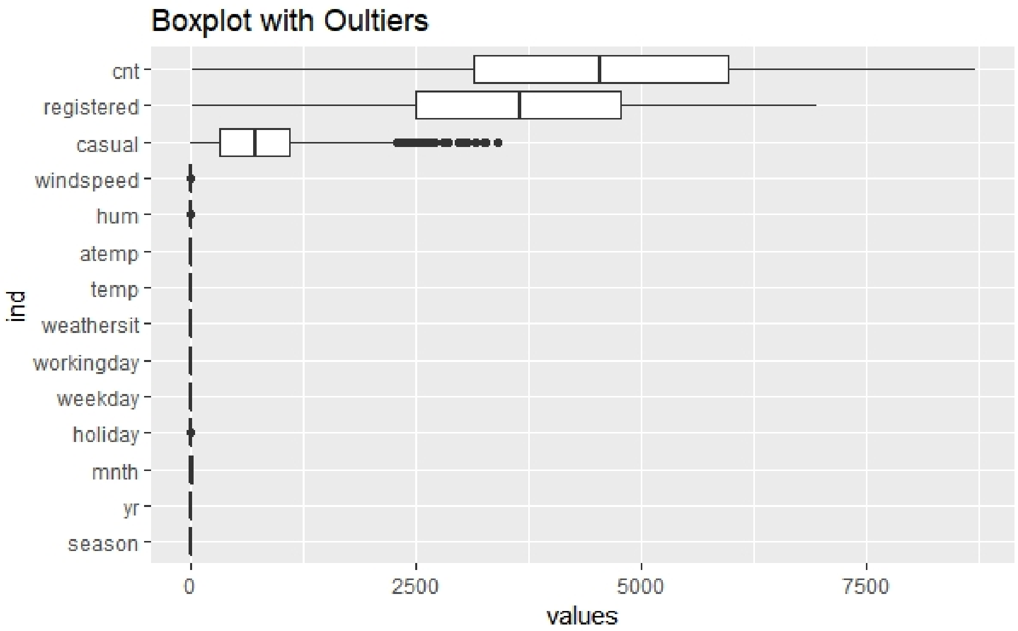

Figure 2.1: Boxplot for outlier Analysis(See R code in appendix B.10)

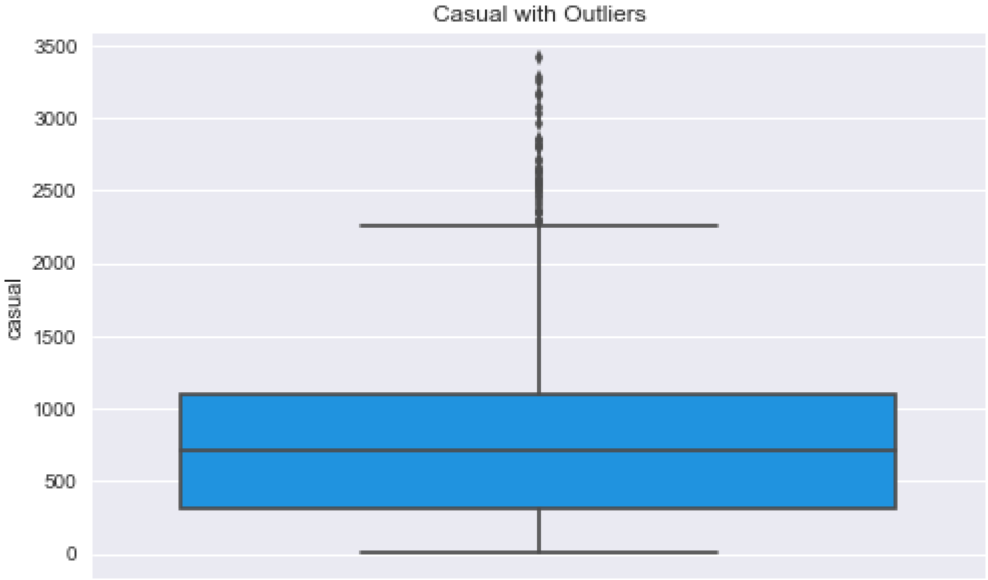

Figure 2.2: Boxplot for Casual(See R code in appendix B.10)

The figure shows that outliers are beyond 2400 and we have to replace them with 2400 itself. The IQRs are as follows:
I Quartile: 0% - 25% value from 2.00 - 315.50
IInd Quartile: 25% - 50% value from 315.50 - 713
IIIrd Quartile: 50% - 75% value from 713.00 - 1096.00
IVth Quartiile: 75% - 100% value from 1096.00 - 2400
We treated the outliers using Capping, the replacement of extreme outliers with the 95%th Quartile value is known as Capping of the outliers . After

treating the outliers the boxplot is as shown,

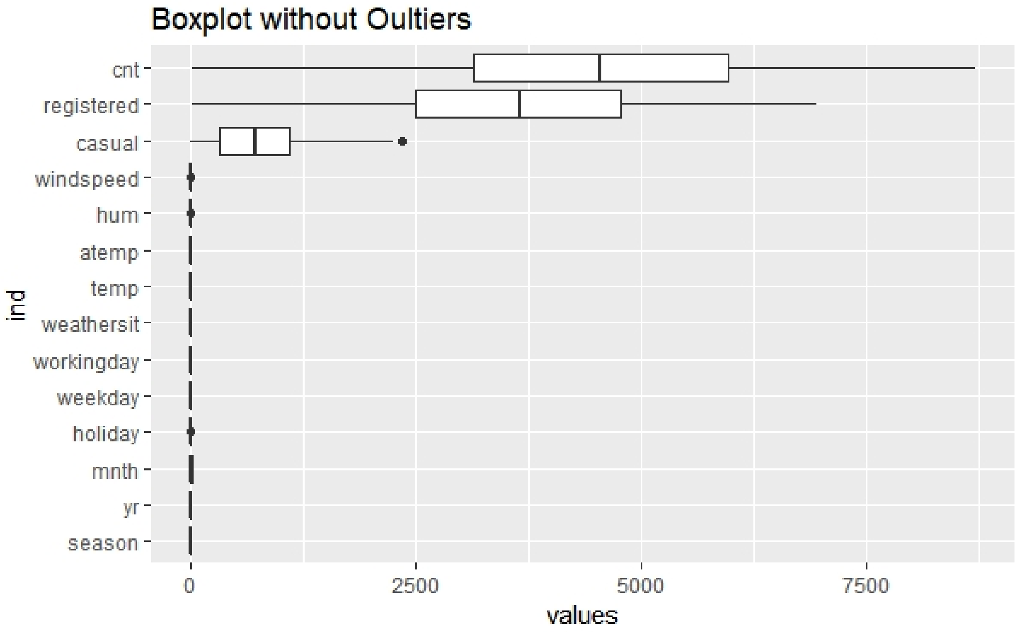

Figure 2.3: Boxplot with outlier treated(See R code in appendix B.10)

2.1.2 Feature Selection

I tried setting up various hypothesis to develop a bi-variate relationship among the response and predictor variables. The relationships are validated using the Analysis of Variance(ANOVA), t-tests and their obtained p-values respectively.
Where,
Null Hypothesis H0 = The variables are independent
Alternative Hypothesis H1 = The variables are not independent

Ist Hypothesis - Is there a relation between the total bike counts and the weekdays?

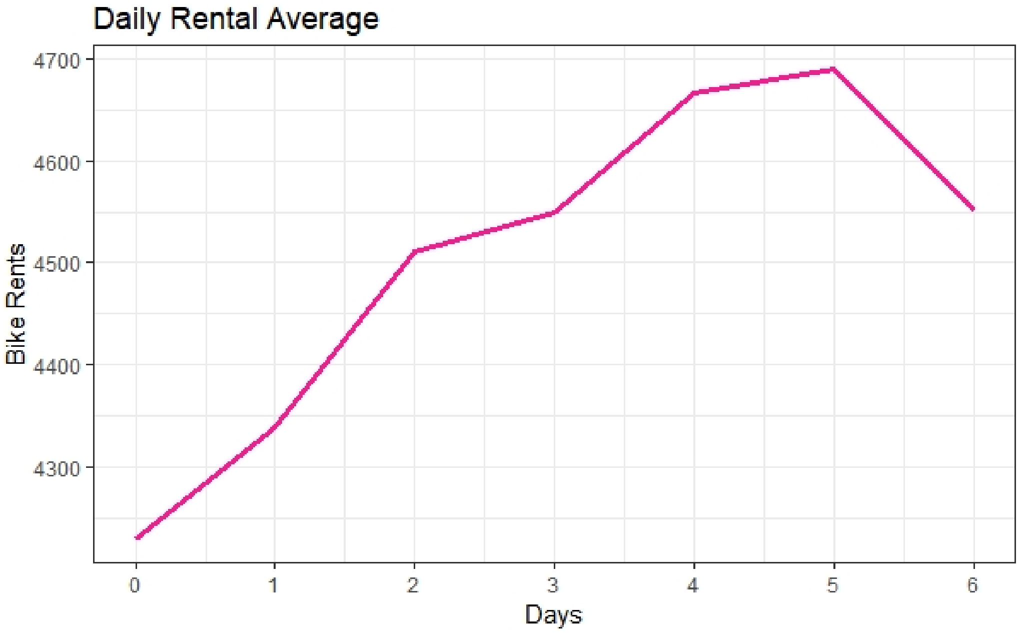

Figure 2.4: Weekdays V/s Average Rental Count(See R code in appendix B.1)

The above plot shows an increasing rental trend towards the weekends from(Thusrday - Saturday) while lesser demands between (Monday - Wednesday) and the obtained p-value: 0.583 with 95% Confidence Interval which means we can't reject our Null Hypothesis and there are chances that this relation may not be fair at times.

IInd Hypothesis - Is there a relation between the season and the rentals average?

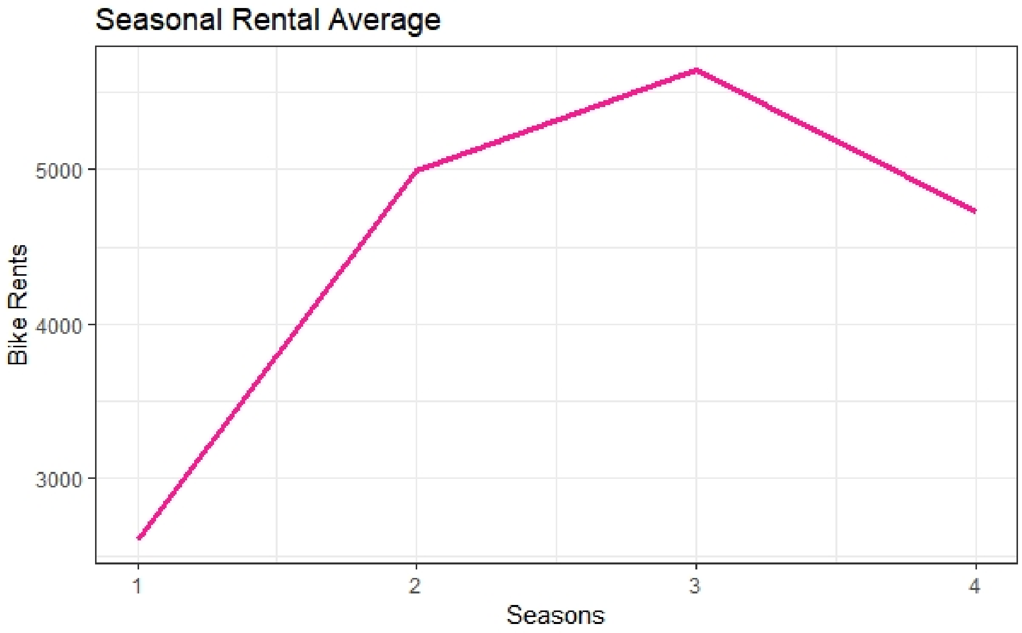

Figure 2.5: Seasons V/s Average Rental Count(See R code in appendix B.2)

The above trend plot shows that Rentals increased significantly between (Summer - Fall) while lowest trends in the Winter Season. Which means people remain cozy during the winters by staying at their homes. The p-value for this relation is ¡ 0.05 with 0.95 CI and thus we rejet the null hypothesis and say the relation between the seasons and the rentals remains true maximum times.

IIIrd Hypothesis - Is there a relation between the months and the rentals average?

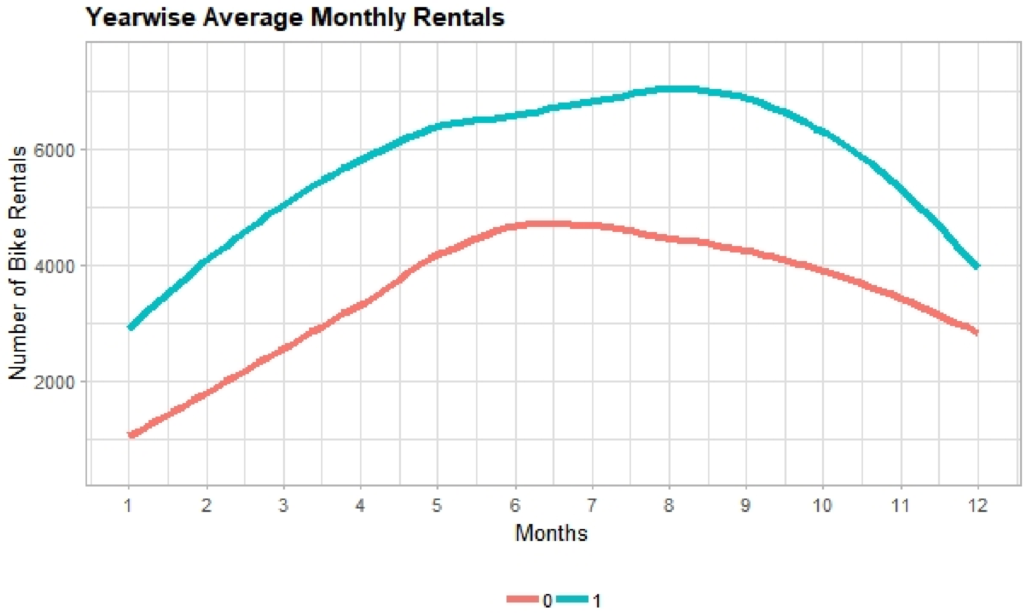

Figure 2.6: Monthly Average Rental Count for both years(See R code in appendix B.3)

The above trend plot shows that Average Rentals increased significantly between (August - September while lowest trends in the (October - January). The p-value for this relation is ¡ 0.05 with 0.95 CI and thus we reject the null hypothesis and say the relation between the months and the rentals remains true maximum times. And the trend also shows that theres' significant increase in the rentals for the year 2012 which says that the Users Index for the company has increased in the years

IVth Hypothesis - Is there a relation between day types and the rentals average? Where,
1 - Weekend
2 - Holiday
3 - Working Day

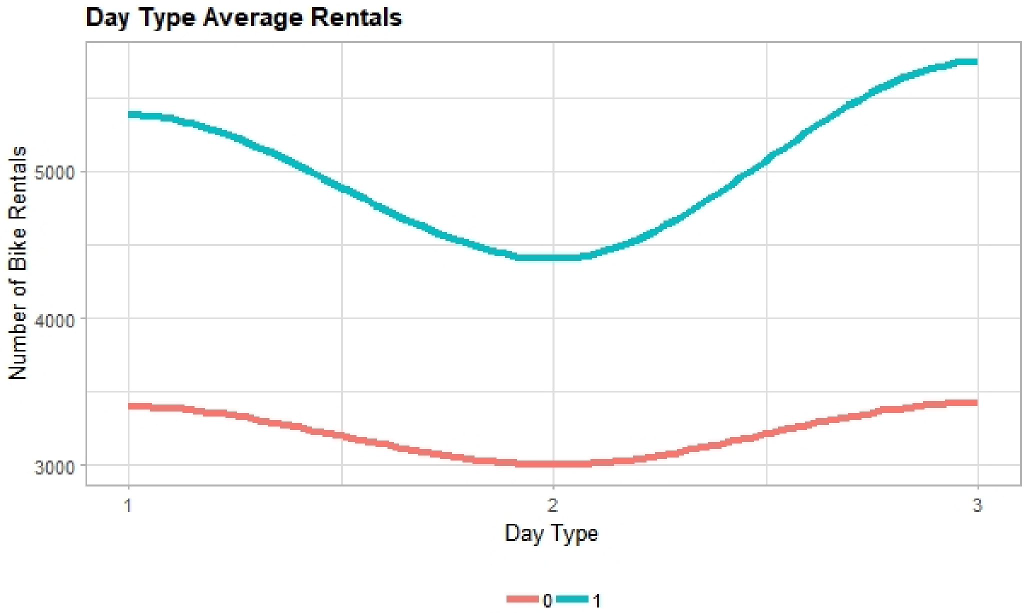

Figure 2.7: Rental Count for all DayTypes in both years(See R code in appendix B.4)

The trend of day types with the average rentals shows that rentals increases on a workingday as there may be chances of rentals requirement in the office commute. However, the holidays has lower rentals in both the years. The obtained p-value for this analysis is 0.08 with 95% CI which implies we can't reject the null hypothesis and there exists biased relation among the variables.

Vth Hypothesis - Is there any relation between the temperature and the bike rentals? Temperature Types
1 - Cool
2 - Moderate
3 - Hot

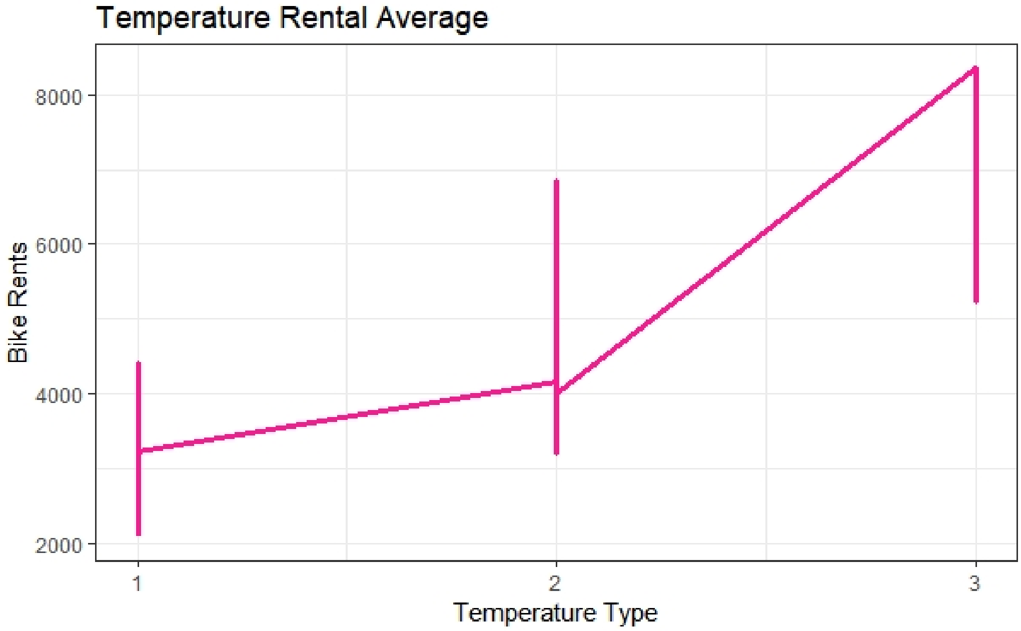

Figure 2.8: Rental Count for the Temperature Types(See R code in appendix B.5)

The trend shows that people are more interested in renting on Hot Days whereas in Cold temperature days people like to be in their home. The obtained p-value for ANOVA test is less than 0.05 which satisfies that there exist a significant relation between the variables with 95% Confidence Interval. The null hypothesis is rejected.

VIth Hypothesis - Is there any relation between the absolute and real temperature? This hypothesis is obtained using the two-sampled t-test where means of both variables are calculated and the if there exists same mean then the variables are dropped. Which means both the variables have the same kind of variances and they are redundant. Obtained p-value is less than 0.02 which implies we can't reject the null hypothesis. The means for both the variables are 0.49 and 0.47 which are almost same.

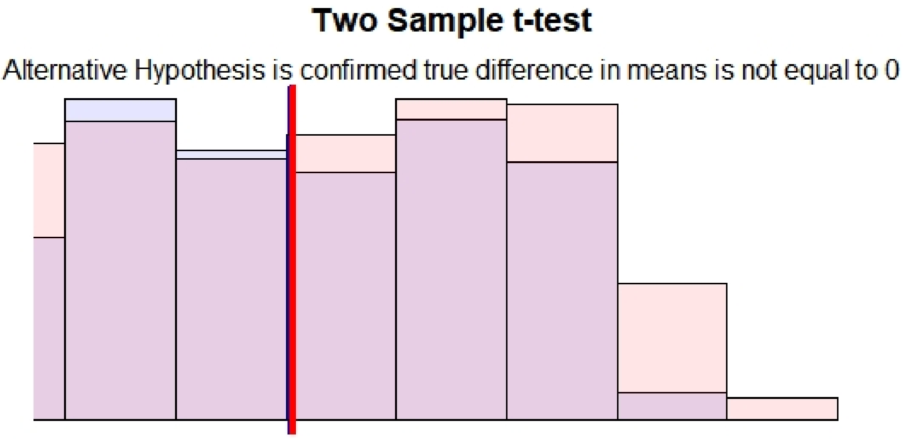

Figure 2.9: Two Sample t-test for absolute and feeled temperature(See R code in appendix B.6)

The above plot is visualizing the two sampled difference between the means and we can see that both the means are overlapped and the mean difference is almost zero which implies both the variables exhibhits the same property and we can drop one among them.

Lets' see the distribution of the continuos variables and check the means of the variables,

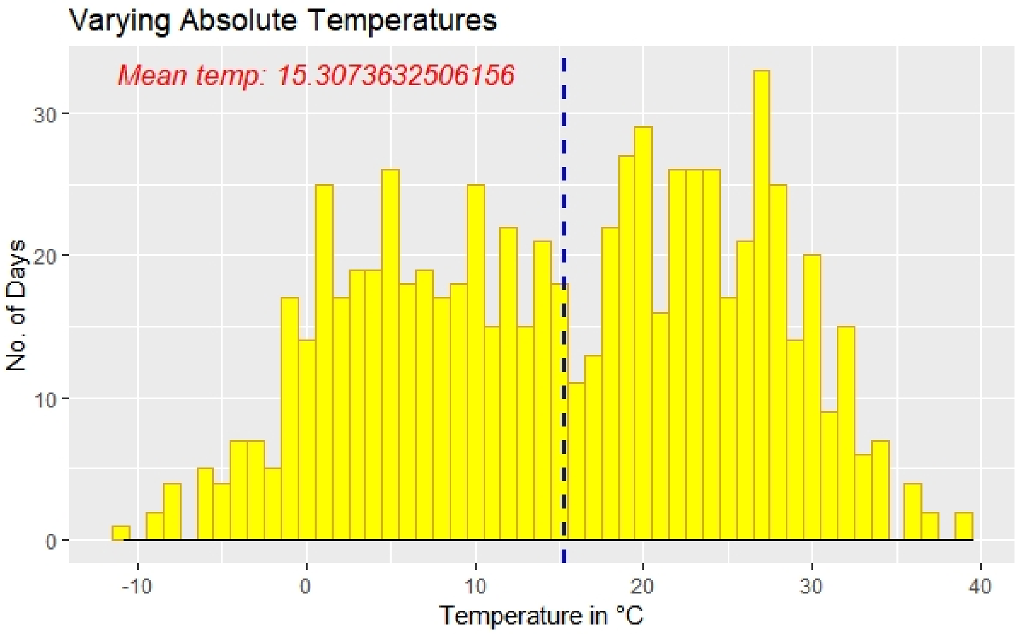

Figure 2.10: Histogram for the absolute temperature(See R code in appendix B.7)

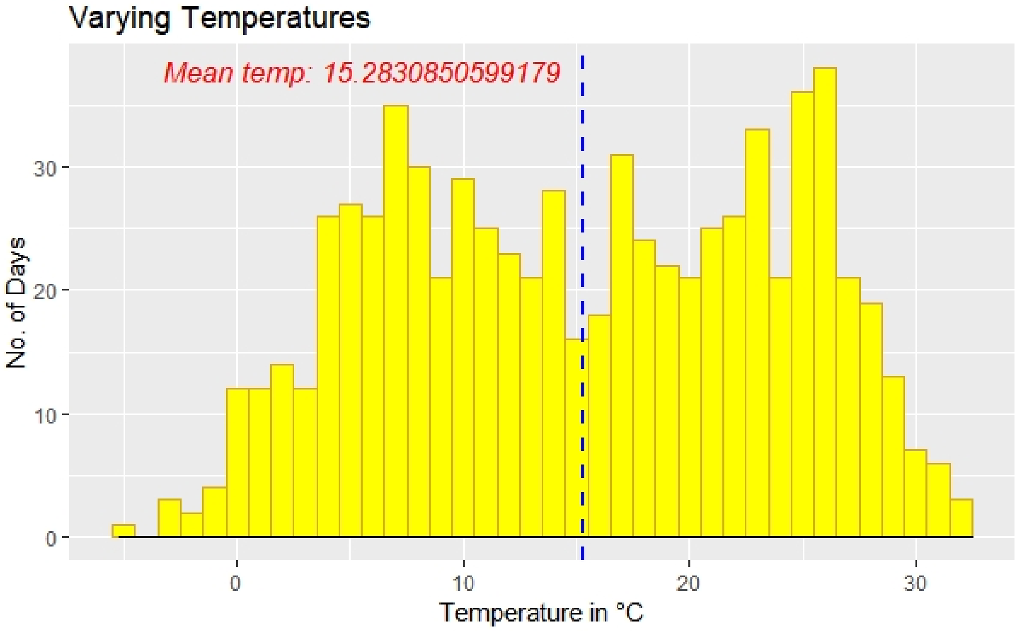

Figure 2.11: Histogram for the temperature(See R code in appendix B.8)

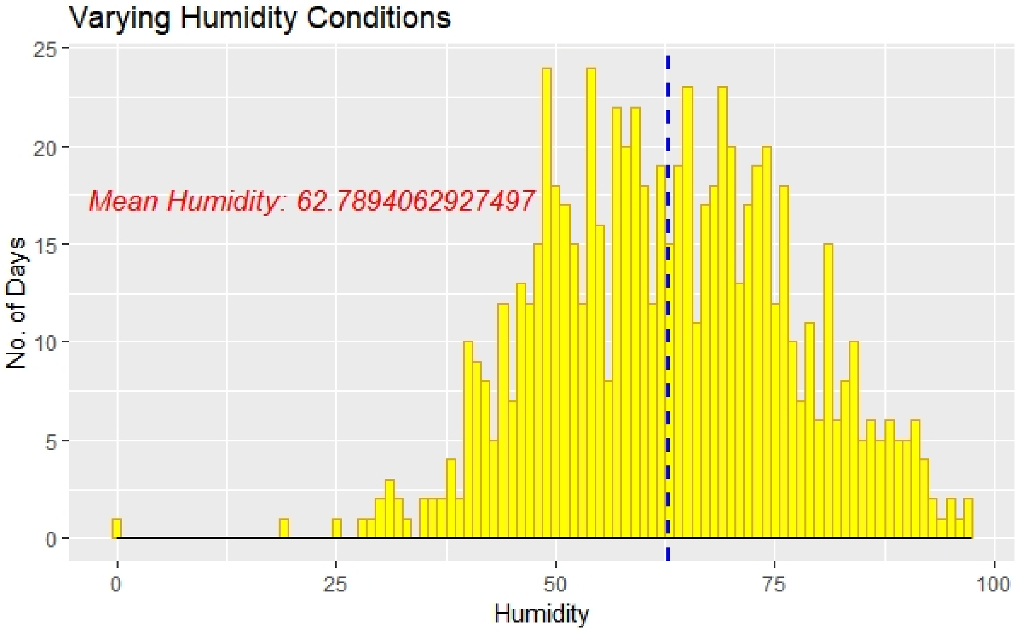

Figure 2.12: Histogram for the humidity(See R code in appendix B.9)

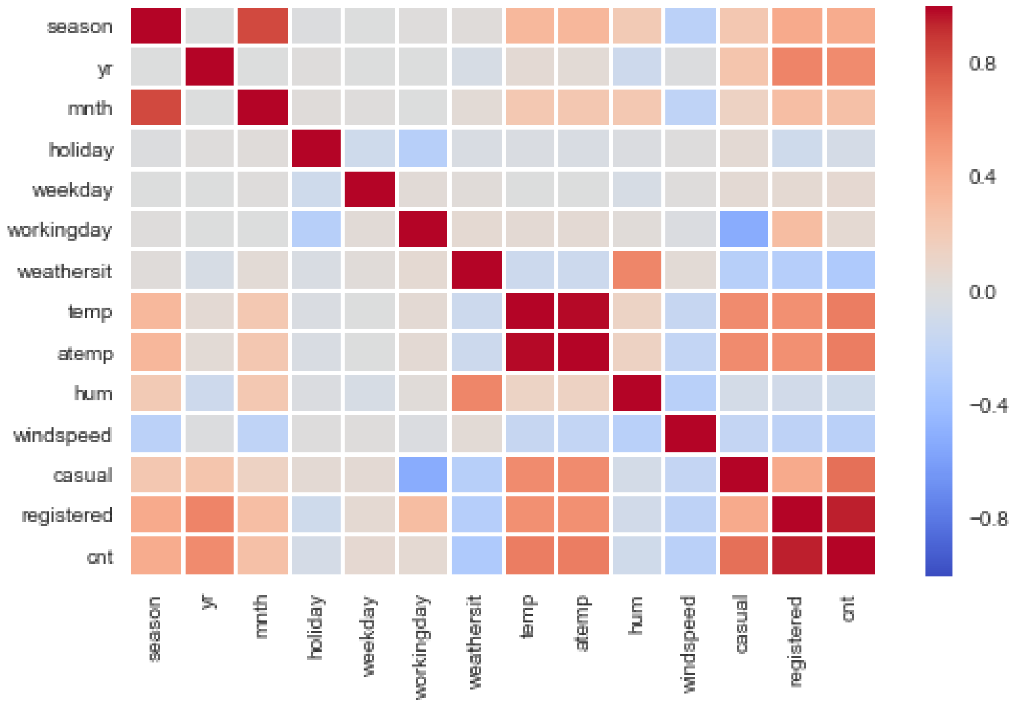

Figure 2.13: Correlation plot for the variables

All the hypothesis and their validations alongwith correlation plot concludes that Season, Months, Temperature, Day Type are the important features that decide the number of rentals on any given day. The casual and registered variables are removed before using any model because they add up to the total counts on a day which is our response variable that we are trying to predict.

2.2 Modeling

2.2.1 Model Selection

We are interested in finding the amount of rentals on a day, given the seasonal and enviromental charachteristics, also we are trying to find the number with all the relevant features, it can be solved by creating a regression among the response and the predictor variables by establishing relevant relations and validating them.

2.2.2 Multiple Linear Regression

```
model1 <- lm(formula = cnt~ .-registered-casual-holiday-
    workingday,rent_df)
summary(model1)

# Call:
#   lm(formula = cnt ~ . - registered - casual - holiday -
    workingday,
#      data = rent_df)
#
#   Residuals:
#   Min          1Q   Median       3Q      Max
#   -4479.9   -461.8             79.0    542.9   2718.6
#
#   Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
#   (Intercept)         3066.81      212.98   14.399   < 2e-16 ***
#   season               496.14       52.29    9.487   < 2e-16 ***
#   yr                  2076.60       62.14   33.420   < 2e-16 ***
#   mnth                 -25.13       16.36   -1.537   0.12485
#   weekday               76.72       15.55    4.935 9.97e-07 ***
#   weathersit          -688.29       74.19   -9.277   < 2e-16 ***
#   hum                -1158.51      299.56   -3.867   0.00012 ***
#   windspeed          -3366.69      432.37   -7.787 2.41e-14 ***
#   day.type2           -426.39      191.83   -2.223   0.02654 *
#   day.type3            120.94       68.71    1.760   0.07881 .
#   temp_typeModerate   1717.90       81.40   21.106   < 2e-16 ***
#   temp_typeHot        2272.05       80.12   28.357   < 2e-16 ***
#   ---
#   Signif. codes:  0    ***    0.001    **    0.01    *    0.05
      .    0.1         1
#
# Residual standard error: 832.4 on 719 degrees of freedom
# Multiple R-squared:  0.8181,  Adjusted R-squared:  0.8154
# F-statistic:  294 on 11 and 719 DF,  p-value: < 2.2e-16
```

Listing 2.1: Model Summary

The above model is trained with excluding registered and casual user variables as they sum up the actual rental counts also the holiday and workingday are removed because we created a new variable, day type which is derived from these two variables.

We can see that **R-Squared and Adjusted R-Squared** value are 0.81 with a p-value less than 0.05 which signifies a fair relationship among the predictors and the response variable.

Lets' check for the Variance Inflation Factor for all the variables in model. Variance inflation factor(VIF) detects multicollinearity in regression analysis. Multicollinearity is when theres correlation between predictors (i.e. independent variables) in a model.

```
vif(model1, threshold = 4, verbose = TRUE)

GVIF Df GVIF^(1/(2*Df))
season      3.554806  1         1.885419
yr          1.018221  1         1.009069
mnth        3.357996  1         1.832484
weekday     1.023189  1         1.011528
weathersit  1.721621  1         1.312105
hum         1.917734  1         1.384823
windspeed   1.182817  1         1.087574
day.type    1.025852  2         1.006401
temp_type   1.218208  2         1.050583
```

Listing 2.2: Model Summary

Lets' tune our model by removing the variables with VIF greater than 3 and then check for the R-Squared Values,

```
#----------- tuned model -----------
tuned_model <- lm(formula = cnt ~ .-registered-casual-season-
    mnth, data = rent_df)
summary(tuned_model)
vif(tuned_model, threshold = 4, verbose = TRUE)
# Residuals:
#    Min       1Q   Median       3Q      Max
# -3773.9   -545.8     41.8    621.2   2417.5
#
# Coefficients:
#   Estimate Std. Error t value Pr(>|t|)
# (Intercept)        3913.75      232.02   16.868  < 2e-16 ***
#   yr                 2078.99       70.43   29.517  < 2e-16 ***
#   weekday              79.96       17.61    4.540 6.58e-06 ***
#   weathersit         -733.68       83.98   -8.736  < 2e-16 ***
#   hum                -690.48      335.28   -2.059   0.0398 *
#   windspeed         -4407.86      483.19   -9.122  < 2e-16 ***
#   day.type2          -424.62      217.28   -1.954   0.0511 .
# day.type3          115.05       77.89    1.477    0.1401
# temp_typeModerate  1995.70       89.15   22.385  < 2e-16 ***
#   temp_typeHot       2555.76       87.49   29.214  < 2e-16 ***
#   ---
#   Signif. codes:  0    ***    0.001    **    0.01    *    0.05
        .    0.1          1
#
# Residual standard error: 943.7 on 721 degrees of freedom
# Multiple R-squared:  0.7656,  Adjusted R-squared:  0.7627
# F-statistic: 261.7 on 9 and 721 DF,  p-value: < 2.2e-16

vif(tuned_model, threshold = 4, verbose = TRUE)

# GVIF Df GVIF^(1/(2*Df))
# yr          1.017901  1        1.008911
# weekday     1.021847  1        1.010865
# weathersit  1.716452  1        1.310135
# hum         1.869143  1        1.367166
# windspeed   1.149291  1        1.072050
# day.type    1.023426  2        1.005806
# temp_type   1.108779  2        1.026151
```

Listing 2.3: Tuned Model Summary

As, we see the R-Squared value reduced to 0.76 from 0.81. Which decreases the regression ability of the model and we lost some crucial information from the analysis. Thus, we can say the previous set of features were better predictors for the Rental Counts. We can conclude that, first set of features were relevant features.

Lets' predict using these features with Linear Regression. Upon correlation of the predicted Rental Count with the Actual Count we got an accuracy of 90.25%.

```
actuals_preds <- data.frame(cbind(actuals=lm_test$cnt, predicteds=pred))
correlation_accuracy <- cor(actuals_preds) #actual vs predicted correlation

correlation_accuracy

#              actuals  predicteds
# actuals     1.000000    0.902521
# predicteds  0.902521    1.000000
```

Listing 2.4: Prediction Accuracy for Linear Models

2.2.3 Decision Tree Regressor

A Regression tree may be considered as a variant of decision trees, designed to approximate real-valued functions, instead of being used for classification methods. A regression tree is built through a process known as binary recursive partitioning, which is an iterative process that splits the data into partitions or branches, and then continues splitting each partition into smaller groups as the method moves up each branch.

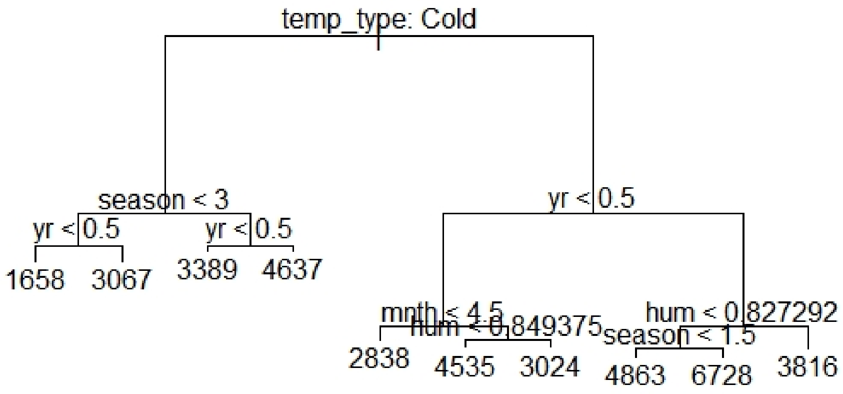

Figure 2.14: Decision Tree Regressor(See R code in appendix **??**)

Chapter 3

Conclusion

3.1 Model Evaluation

We came across two different approaches for estimating the target class. Now, we need to choose which model will be suitable for the future cases and which has a significant efficiency in forecasting the result. Following factors are helpful in determining a models' efficiency:
1. Predictive Performance
2. Interpertability
3. Computational Efficiency

We will be assessing our model with the first factor, Predictive Performance.

3.1.1 Mean Absolute Error

MAE is one of the error measures used to calculate the predictive performance of the model. We will apply this measure to our models that we have generated in the previous section.

```
DMwR::regr.eval(actuals_preds$actuals, actuals_preds$predicteds)
# mae          mse          rmse         mape
# 6.663752e+02 7.047886e+05 8.395169e+02 2.001745e-01
```

Listing 3.1: Evaluation Metrics for Linear Model

The Evaluation Metrics for Linear Regressions are as follows:
Mean Absolute Error = 666.37
Mean Squared Error = 7.04
Root Mean Squared Error = 839

Mean Absolute Percentage Error = 2%

```
DMwR::regr.eval(dt_actuals_preds$actuals, dt_actuals_preds$
    predicteds)
# mae          mse          rmse         mape
# 7.474133e+02 9.580728e+05 9.788119e+02 2.466638e-01
```

Listing 3.2: Evaluation for Decision Tree Regressor

The Evaluation Metrics for Linear Regressions are as follows:
Mean Absolute Error = 747.37
Mean Squared Error = 7.04
Root Mean Squared Error = 978.11
Mean Absolute Percentage Error = 2.4% The correlation, between the predicted and actual values are 85.72% which says this model is capable of prediction at almost 85% of the times.

3.1.2 Root Mean Square

Table 3.1: Root Mean Square Error

Model	Root Mean Square
Linear Regression	839
Decision Tree Regression	978.11

For, Linear Regeression Model we have,
RMSE training = 824.3443
RMSE test = 839.444
The difference between RMSE training and test is low for Linear Regression Model, which validates the model performance and accuracy. The correlation, between the predicted and actual values are 89.88% which says this model is capable of prediction at almost 90% of the times.

3.2 Model Selection

Based on our hypothesis and validations and various VIF iterations, we see that Prediction Accuracy and RMSE for both the test and training are same for Multiple Regression Model and we can select the Linear Regression Model as our final model with 90% prediction accuracy on Day Type, Temperature Type, Season and Weekday as the important predictors in the model.

Appendix A

Extra Figures

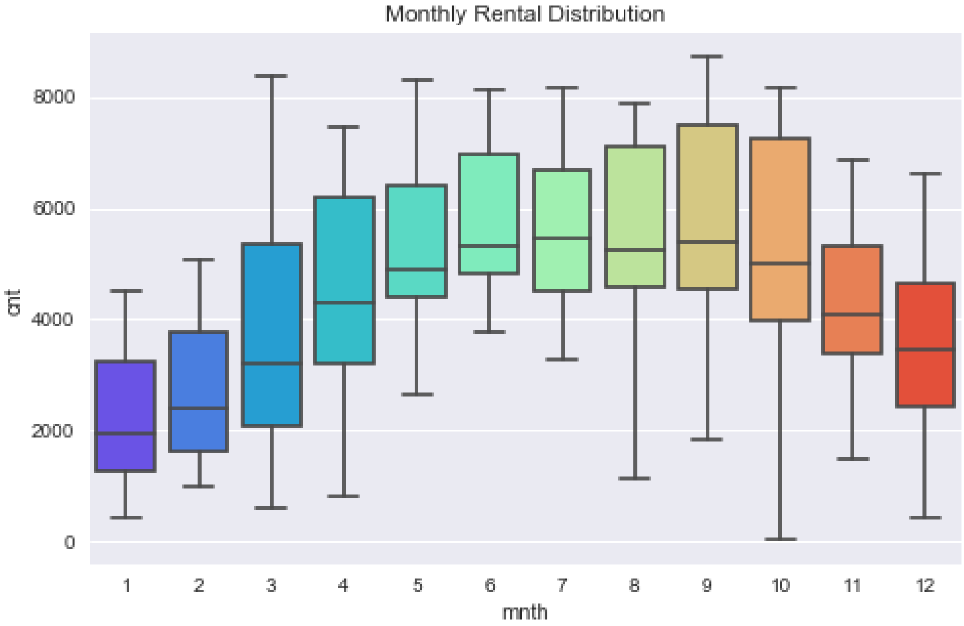

Figure A.1: Boxplot for Monthly Rentals

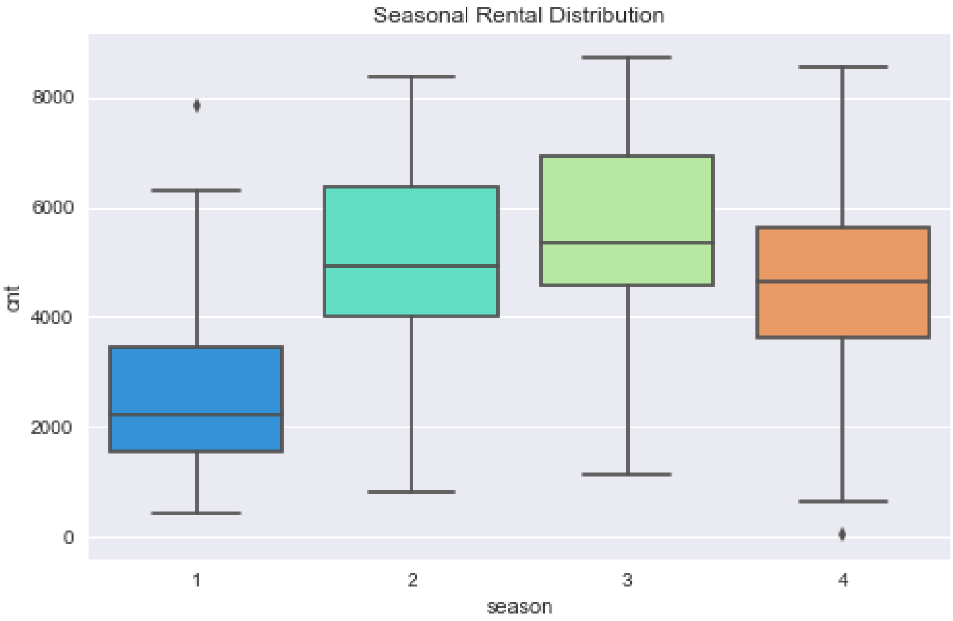

Figure A.2: Boxplot for Seasons Rentals

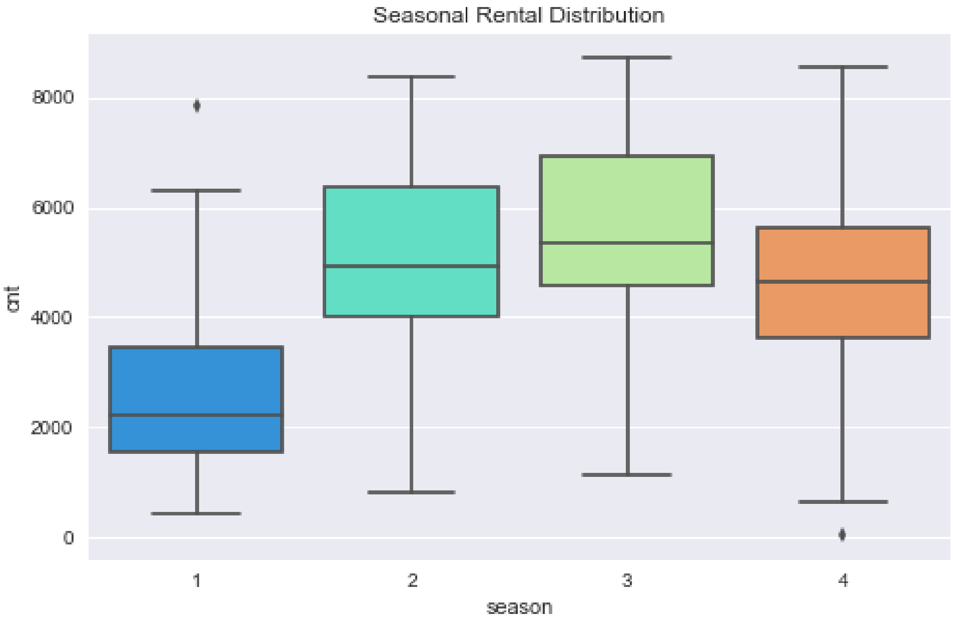

Figure A.3: Boxplot for Seasons Rentals

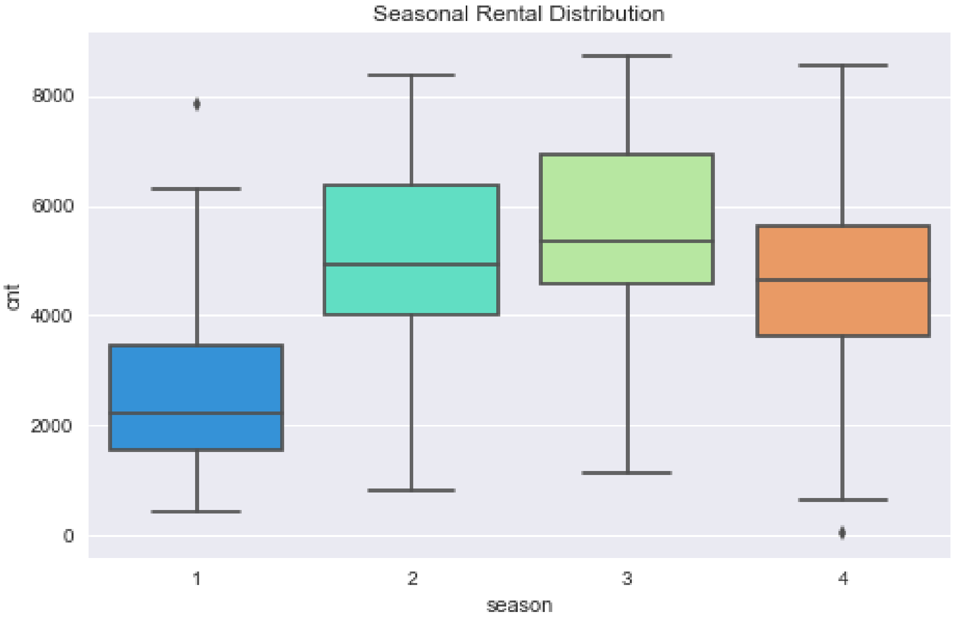

Figure A.4: Boxplot forAnnual Rentals

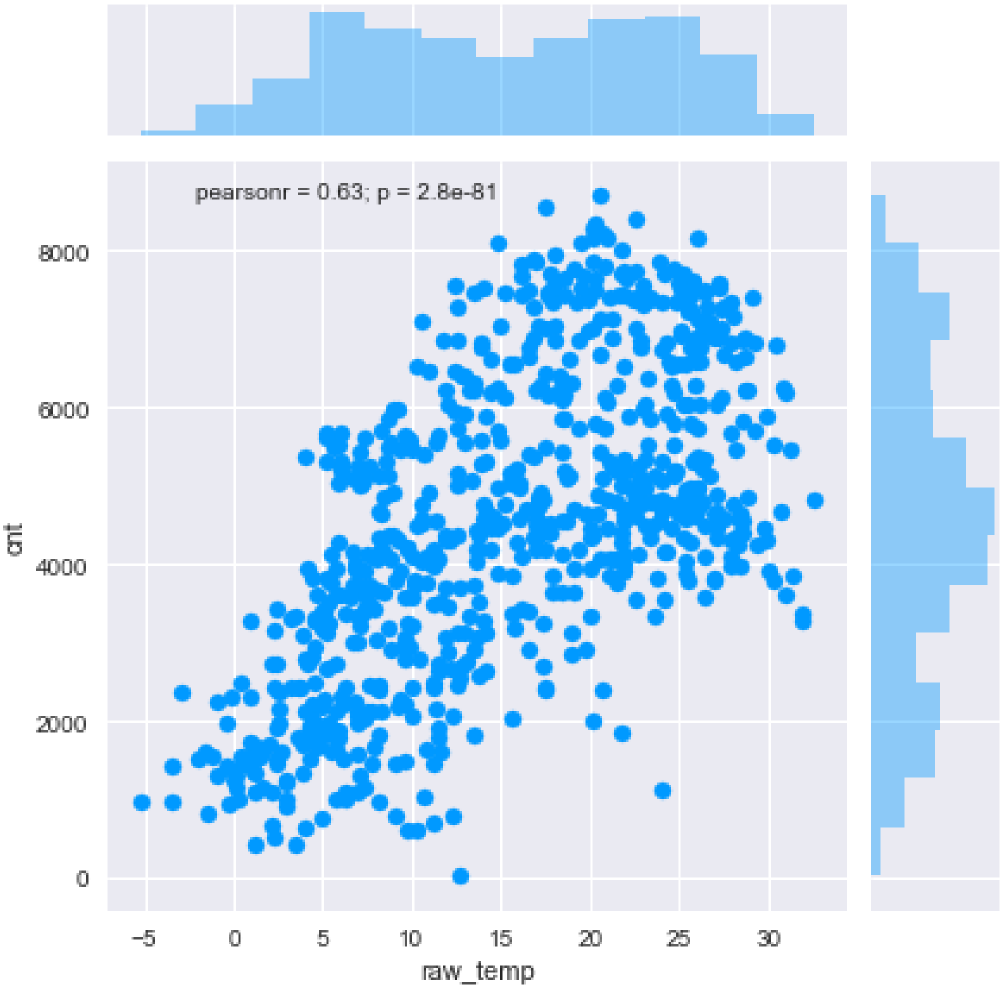

Figure A.5: Scatter plot for Temperature V/s Count

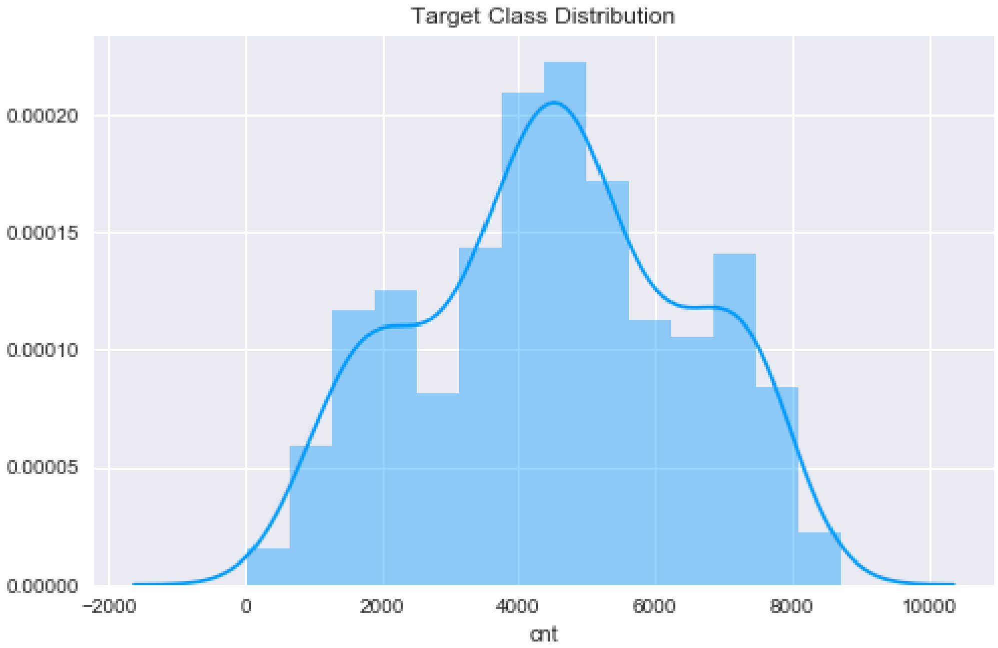

Figure A.6: Target Class Distribution

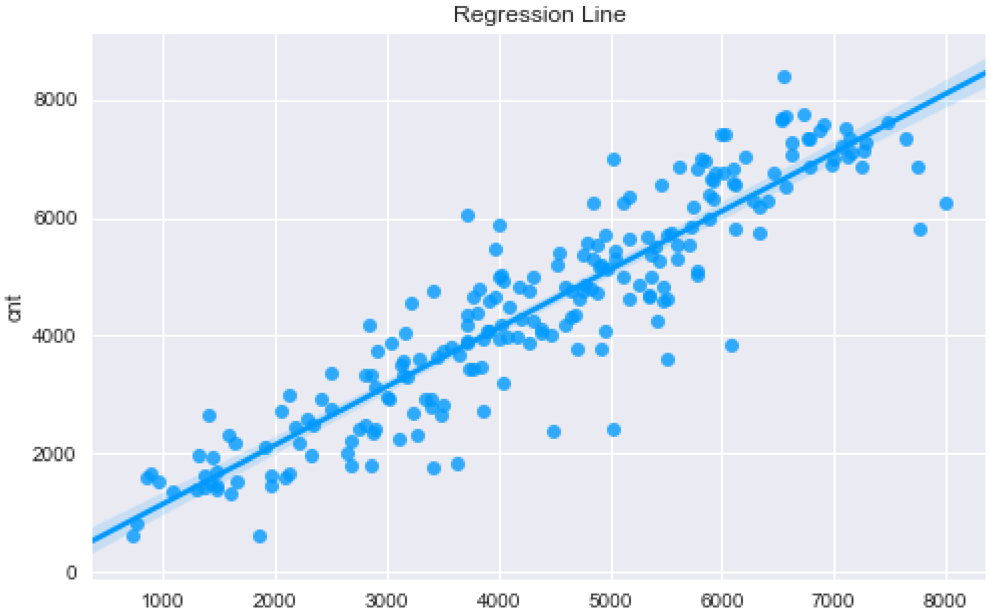

Figure A.7: Regression Plot for Linear Model

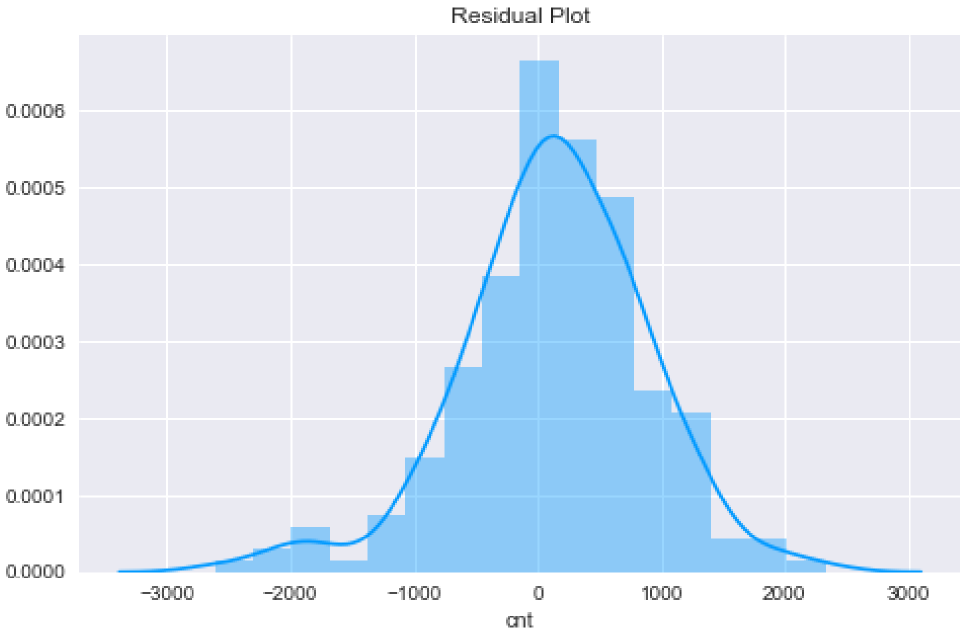

Figure A.8: Residual Plot for Linear Model

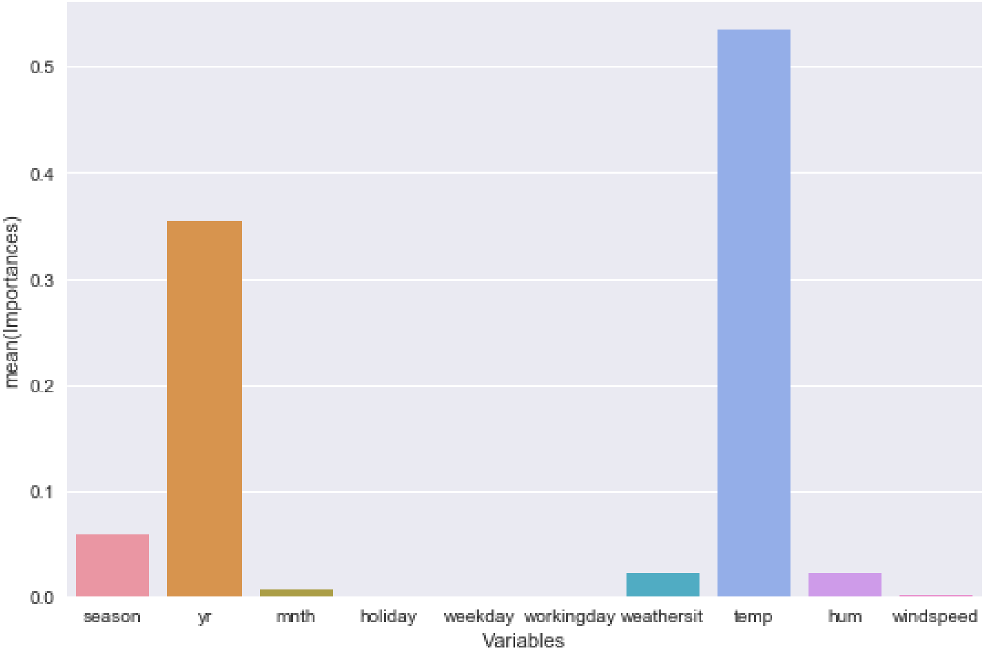

Figure A.9: Variable Importance

Appendix B

Code

```
#----------------  Hypothesis 1----------------
#Is there relation between the weekdays and bike rentals?
summary(aov(cnt ~ as.factor(weekday), data = rent_df))
TukeyHSD(aov(cnt ~ as.factor(weekday), data = rent_df))
#Conclusion: P-value is 0.583, means we cant reject the null hypothesis

#----------------------  Plot 1 --------------------------
week_avg <- rent_df %>%
  group_by(weekday) %>%
  summarize(mean_size = mean(cnt,na.rm = T))
  as.data.frame %>%
  write.table() #grouping data with weekdays

p1 <- ggplot() + theme_bw() +
  geom_line(aes(y = mean_size, x = weekday), size=1.3, data = week_avg,
           stat="identity",colour = 'deeppink') +
  theme(legend.position="bottom", legend.direction="horizontal",
        legend.title = element_blank()) +
  scale_x_continuous(breaks=seq(0,6,1)) +
  labs(x="Days", y="Bike Rents") +
  ggtitle("Daily Rental Average") +
  scale_colour_manual(values=colour)
p1 #lineplot for daily average rents
```

Listing B.1: Hypothesis and Visualization I

```
#----------------  Hypothesis 2----------------
#Is there relation between season and bike rentals?
summary(aov(cnt ~ as.factor(season), data = rent_df))
TukeyHSD(aov(cnt ~ as.factor(season), data = rent_df))
#Conclusion: P-value < 0.05, means we accept the null hypothesis and theres' relation
```

```

#-------------------- Plot 2 ------------------------
season_avg <- rent_df %>%
  group_by(season) %>%
  summarize(mean_size = mean(cnt,na.rm = T))
  as.data.frame %>%
  write.table() #seasonal data grouping

p2 <- ggplot() + theme_bw() +
  geom_line(aes(y = mean_size, x = season), size=1.3, data = season_avg,
            stat="identity",colour = 'deeppink') +
  theme(legend.position="bottom", legend.direction="horizontal",
        legend.title = element_blank()) +
  scale_x_continuous(breaks=seq(1,4,1)) +
  labs(x="Seasons", y="Bike Rents") +
  ggtitle("Seasonal Rental Average") +
  scale_colour_manual(values=colour)
p2 #line plots for average monthly rentals
```

Listing B.2: Hypothesis and Visualization II

‘

```
#---------------- Hypothesis 3----------------
#Is there relation between months and bike rentals?
summary(aov(cnt ~ as.factor(mnth), data = rent_df))
TukeyHSD(aov(cnt ~ as.factor(mnth), data = rent_df))
#Conclusion: P-value < 0.05, means we accept the null hypothesis and theres' relation
#-------------------- Plot 3 ------------------------

monthly_avg <- rent_df %>%
  group_by(mnth,yr) %>%
  summarize(mean_size = mean(cnt,na.rm = T))
  as.data.frame %>%
  write.table() #grouped month and year data

p3 <- ggplot(monthly_avg, aes(x = mnth,y = mean_size,color = as.factor(yr))) +
         geom_smooth(method = "loess", fill = NA, size = 1.5) +
         theme_light(base_size = 10) +
         theme(legend.position="bottom", legend.direction="horizontal",
         legend.title = element_blank()) +
         xlab("Months") +
         ylab("Number of Bike Rentals") +
         ggtitle("Yearwise Average Monthly Rentals") +
         scale_x_continuous(breaks=seq(1,12,1)) +
         theme(plot.title = element_text(size = 11, face="bold")
  )
p3 #lineplot for average rents per months for year 2011 and 2012
```

Listing B.3: Hypothesis and Visualization III

‘

```
#----------------------Hypothesis 4 -----------
#Is there a relation between the day_type and the rental count?
summary(aov(cnt ~ as.factor(day.type), data = rent_df))

#--------------------Plot 4---------
daytype_avg <- rent_df %>%
  group_by(day.type,yr) %>%
  summarize(mean_size = mean(cnt,na.rm = T))
  as.data.frame %>%
  write.table() #grouped daytypes and years

p4 <- ggplot(daytype_avg, aes(x = as.numeric(day.type),y = mean_size,color = as.factor(yr))) +
    geom_smooth(method = "loess", fill = NA, size = 1.5) +
    theme_light(base_size = 10) +
    theme(legend.position="bottom", legend.direction="horizontal",
          legend.title = element_blank()) +
    xlab("Day Type") +
    ylab("Number of Bike Rentals") +
    ggtitle("Day Type Average Rentals") +
    scale_x_continuous(breaks=seq(1,3,1)) +
    theme(plot.title = element_text(size = 11, face="bold"))
p4 #lineplot for average rents on the basis of dayt for year 2011 and 2012
```

Listing B.4: Hypothesis and Visualization IV

‘

```
#---------------- Hypothesis 5-----
#Is the temperature related with the bike rentals
t.test(rent_df_analysis$temp, rent_df_analysis$cnt, paired = FALSE)
#Conclusion: The p-value < 0.05 and the null hypothesis is rejected
summary(aov(cnt ~ as.factor(temp_type), data = rent_df_analysis))
#--------------------Plot 5---------
temptype_avg <- rent_df_analysis %>%
  group_by(temp_type,mnth) %>%
  summarize(mean_size = mean(cnt,na.rm = T))
  as.data.frame %>%
  write.table() #grouped daytypes and years
```

```

p5 <- ggplot() + theme_bw() +
    geom_line(aes(y = mean_size , x = as.numeric(temp_type)), size=1.3, data = temptype_avg,
                stat="identity",colour = 'deeppink') +
    theme(legend.position="bottom", legend.direction="horizontal",
          legend.title = element_blank()) +
    scale_x_continuous(breaks=seq(1,3,1)) +
    labs(x="Temperature Type", y="Bike Rents") +
    ggtitle("Temperature Rental Average") +
    scale_colour_manual(values=colour)
p5 #line plots for average monthly rentals
```

Listing B.5: Hypothesis and Visualization V

‘

```
#---------------Hypothesis 6--------
#Is there any difference between normal temperature and the feeled temp.
#Two sampled t-tests between normal and feeled temp.
t.test(x = rent_df$temp, y = rent_df$atemp, alternative = "two.sided")
#Conclusion: p-Value < 0.05 and the null hypothesis can't be rejected

#---------------Plot 6------------
hist(rent_df_analysis$raw_temp, yaxt = "n", xaxt = "n", xlab = "",
     ylab = "", main = "Two Sample t-test",
     xlim = c(5, 40), col = rgb(0, 0, 1, alpha = .1))
text(x = 50, y = 140, paste("Mean real Temp.\n",round(mean(rent_df_analysis$raw_temp), 2), sep = ""), col = "blue")
abline(v = mean(rent_df_analysis$raw_temp), lty = 1,
       col = rgb(0, 0, 1, alpha = 1), lwd = 4)

par(new = T)
hist(rent_df_analysis$raw_atemp, yaxt = "n", xaxt = "n", xlab = "",
     ylab = "", main = "", xlim = c(5, 40), col = rgb(1, 0, 0, alpha = .1))

abline(v = mean(rent_df_analysis$raw_atemp), lty = 1,
       col = rgb(1, 0, 0, alpha = 1), lwd = 4)


mtext(text = "Alternative Hypothesis is confirmed true difference in means is not equal to 0", line = 0, side = 3)
#Plot represents there is no significant mean difference
```

```
#and the null hypothesis can't be rejected
```

Listing B.6: Hypothesis and Visualization VI

‘

```
#-------------------------- Plot 8 --------------------------------
#Displaying mean on the plot
grob1 <- grobTree(textGrob(paste0("Mean temp: ",mean(rent_df_
    analysis$raw_temp)),
                              x=0.1,  y=0.95, hjust=0,
                              gp=gpar(col="red", fontsize=13,
    fontface="italic")))


#Histogram plot for temperatures
p6 <- ggplot(rent_df_analysis, aes(x=raw_temp) ) +
       geom_histogram(color = "goldenrod",fill="yellow",binwidth
    = 1) +
       labs(x = "Temperature in   C",y = "No. of Days")+
       ggtitle("Varying Temperatures") +
       geom_density(alpha=0.6)+
       geom_vline(aes(xintercept=mean(raw_temp)),
                   color="blue", linetype="dashed", size=1)+
       annotation_custom(grob1)
p6
```

Listing B.7: Histogram for Varying Temperature

‘

```
#-------------------------- Plot 9 --------------------------------
#text on plots
grob2 <- grobTree(textGrob(paste0("Mean temp: ",mean(rent_df_
    analysis$raw_atemp)),
                              x=0.05,  y=0.95, hjust=0,
                              gp=gpar(col="red", fontsize=13,
    fontface="italic")))


#Histogram plot for temperatures
p7 <- ggplot(rent_df_analysis, aes(x=raw_atemp))+
   geom_histogram(color = "goldenrod",fill="yellow",binwidth = 1)
     +
   labs(x = "Temperature in   C",y = "No. of Days")+
   ggtitle("Varying Absolute Temperatures") +
   geom_density(alpha=0.6)+
   geom_vline(aes(xintercept=mean(raw_atemp)),
               color="blue", linetype="dashed", size=1) +
   annotation_custom(grob2)
p7
```

Listing B.8: Histogram for Varying Absolute Temperature

‘

```
#------------------------- Plot 10 -------------------------
#Text on plot
grob3 <- grobTree(textGrob(paste0("Mean Humidity: ",mean(rent_df_analysis$raw_humidity)),
                                x=0.020,  y=0.70, hjust=0,
                                gp=gpar(col="red", fontsize=13, fontface="italic")))

#Histogram plot for temperatures with mean
p8 <- ggplot(rent_df_analysis, aes(x=raw_humidity))+
  geom_histogram(color = "goldenrod",fill="yellow",binwidth = 1) +
  labs(x = "Humidity",y = "No. of Days")+
  ggtitle("Varying Humidity Conditions") +
  geom_density(alpha=0.6)+
  geom_vline(aes(xintercept=mean(raw_humidity)),
                  color="blue", linetype="dashed", size=1) +
  annotation_custom(grob3)
p8
```

Listing B.9: Histogram for Varying Humidity

```
#Outlier Analysis on Required variables
ggplot(stack(rent_df), aes(x = ind, y = values)) +
          geom_boxplot() + coord_flip() +
          ggtitle("Boxplot with Oultiers")

summary(rent_df$casual) #summary of the obtained outliered feature

#Outlier Treatment
x <- rent_df$casual
qnt <- quantile(x, probs=c(.25, .75), na.rm = T)
caps <- quantile(x, probs=c(.05, .95), na.rm = T)
H <- 1.5 * IQR(x, na.rm = T)
rent_df$casual[rent_df$casual < (qnt[1] - H)] <- caps[1]
#replacing extreme low whisker with 25% quartile (Flooring)
rent_df$casual[rent_df$casual > (qnt[2] + H)] <- caps[2]
#replacing extreme high whisker with 75% quartile (Capping)

ggplot(stack(rent_df), aes(x = ind, y = values)) +
  geom_boxplot() + coord_flip() +
  ggtitle("Boxplot without Oultiers")
```

Listing B.10: Outlier Analysis using Boxplot