Capital Bike-Share Data Analysis

1. Project Description

The dataset contains the bike share rental data from ‘Capital Bikeshare’ company servicing Washington D.C. and surrounding areas since 2010. Bike sharing system provides people with a short distance transportation option without worrying about the traffic hustle and enjoying the city view as well as workout at the same time.
This dataset contains many factors like temperature, weather, season, holidays, working days, humidity, wind-speed that can affect the rental count. We are interested mainly in actual temperature and the different weather change that affect the bike rental count in city.

2. Research Scenario Description (no more than 200 words)

Capital Bikeshare is one of the America’s most successful and largest bikeshare system. They have collected a data of total number of rental counts based on different factors like season, weather, temperature, holidays, working-days, humidity, wind-speed and few more for years 2011 and 2012.
They want to investigate if temperature plays a role in the number of rental counts. Also, they would like to know what is the effect of different weather conditions on the rental count. And at last is there any effect of weather conditions on rental count after adjusting for temperature.

We will be using Simple Linear Regression, ANOVA model, ANCOVA for answering these questions.

3. Describe the data set (no more than 200 words)

The dataset can be downloaded from http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset.

It contains bike rental data containing 731 rows and 16 columns for two years 2010 and 2011 and the rental count based on weather, temperature, humidity, windspeed, holiday, working-day and some other factors.

As a part of some pre-processing, we checked for any missing values in the dataset and found no missing/null values in any of the columns. Since, in order to answer our research questions, we are interested in weather, temperature and count variables, so we have kept these 3 columns only and removed the other unused columns from our dataset.
Also, sampled our dataset and randomly selected 500 rows for our analysis.
At last, we checked for any outliers in our selected variables and found none. So, our final dataset is of dimension (500,3) which we exported to ‘*bike_cleaned.csv*’ file.

3. Research Question (no more than 100 words)

Describe briefly in one or two sentences the main research question. This is similar to the last sentence of our class examples.

1. Is temperature a significant predictor for rental count?
2. Does the rental count varies based on the weather conditions? If yes, which two different weather conditions show a significant difference in average rental count.
3. What is the effect of weather conditions on rental count after adjusting for temperature?

4. Your solution R code

```
# import cleaned data
bike_data <- read.csv('bike_cleaned.csv', header = T)
bike_data$weather <- as.factor(bike_data$weather)
str(bike_data)
head(bike_data)

####################
#           EDA              #
####################

#Categorical variable
aggregate(bike_data$count, by=list(bike_data$weather), summary)
aggregate(bike_data$count, by=list(bike_data$weather), sd)

boxplot(bike_data$count ~ bike_data$weather,
        data = bike_data,
        main = "Total Bike Rentals Vs Weather",
        xlab = "Weather",
        ylab = "Total Bike Rentals",
        col = c("#D6EAF8", "#2ECC71", "#E74C3C", "#F39C12"))

# We can see that most of the bike rentals are in clear weather conditions.
# No-one rents a bike when there is a chance of heavy rains.

# Continuous variable
# Let's see how is the correlation between  temperature and rental count
cor(bike_data$count, bike_data$temperature)   # 0.64   (moderately strong positive correlation)

#Scatter plot
plot(bike_data$temperature, bike_data$count, pch=16, col='blue',
     xlab = 'Temperature in Celsius',
     ylab = 'Rental Count',
     main = 'Rental count based on Temperature')

# the scatter plot shows a positive linear relationship between temperature and rental count.
# There seems to be some points that deviate the plot somewhat from linearity, like the ones showing higher temperature
but less rental counts.
# In order to see if this is due to the outliers, we will check for outliers in our data.

## Outlier Detection:
```

```
outlier_iqr <- function(x){
  iqr <- IQR(x,na.rm = T,type = 7)
  q <- quantile(x)
  upper_bound = q[4]+(iqr*1.5)
  lower_bound = q[2]-(iqr*1.5)
  outliers <- which ((x > upper_bound) | (x < lower_bound))
  return(outliers)
}

# Checking outliers for Rental count and temperature
print(outlier_iqr(bike_data$count))

print(outlier_iqr(bike_data$temperature))

# No outliers are present in the rental count and temperature.
# The deviation we see in plot is might be because the data is recorded like this only.

#####################
#          Research Q-1      #
#####################

### Is temperature a significant predictor for rental count?

# In order to answer this question, will perform a simple linear regression with count as dependent and temperature as
independent variable at alpha = 0.05 level.

# The hypothesis test will be -
# H0: beta_temp = 0 (There is not a linear association between temperature and rental count )
# H1: beta_temp ≠ 0 (There is a linear association between temperature and rental count )
# alpha = 0.05

# Generating model
slr <-  lm(count ~ temperature, data=bike_data)
summary(slr)

# Now, since our model gave significant results, we will check whether all our model assumptions are met before making
any inferences.
# We will check for Linearity, Independence, Constant variance and normality assumptions.


###  Regression Diagnostics
par(mfrow=c(2,2))

plot(bike_data$temperature, bike_data$count, xlab= "Temperature",col = 'blue',
     ylab = 'Rental Count',
     main = 'Temperature vs Rental Count')
abline(slr, col = 'red', lty =1)

plot(bike_data$temperature, resid(slr), axes=TRUE, frame.plot=TRUE, xlab='Temperature', ylab='Residue',
     main = 'Residual Plot', col = 'blue')
abline(h=0, lty=1, col='red')
```

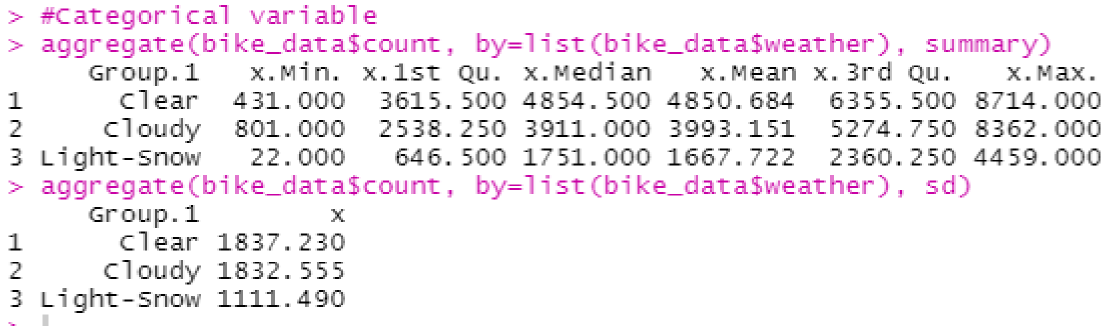
```
plot(fitted(slr), resid(slr), axes=TRUE, frame.plot=TRUE, xlab='Fitted values', ylab='Residue',
     main = 'Fitted values vs Residual plot', col = 'blue')
abline(h=0, lty=1, col='red')

hist(resid(slr), main = 'Histogram of Residuals', xlab = 'Residuals')
par(mfrow=c(1,1))

# Looking at the plots, we can say that all our model assumptions are met.
# Linearity: Residual plot shows that there is a positive linear relationship between the variables.
# Independent: The independence assumption is met as the observations are drawn based on different days of years.
# Constant Variance: From the residual plot, we can see that the variance is almost constant around the regression line.
# Normality: The histogram tells us that the residuals are normally distributed.

## Checking for influence points
cooks.dist <- cooks.distance(slr)
which(cooks.dist > (4/(nrow(bike_data)-2-1)))

# Let's check the effect of these influence points on our model
influence.points <- as.vector( which(cooks.dist > (4/(nrow(bike_data)-2-1))) )

for (i in influence.points) {
  count2 <- bike_data$count[-i]
  temp2  <- bike_data$temperature[-i]

  bike_data2 <- data.frame(count2, temp2)
  cor(count2, temp2)

  lm.2 <- lm(count2 ~ temp2, data = bike_data2)
  model.summary <- summary(lm.2)
  print(paste('Influence Point:', i,
            '  R-squared: ',  round(model.summary$r.squared,4),
            '  beta1: ', model.summary$coefficients[2,1]),)
}

# By removing these influential points for our data, we checked the model's performance.
# Earlier including these points in our data, the adjusted R-square was around 0.41 which is almost the same after removing these points from our data.
# The beta1 estimate also does not a large difference. So these influence points are not much impacting our data. Keeping them in our data won't cause much problem.
# We can say that the model fits well on our data.

# Model Conclusions:
# Reject the Null hypothesis as p-value (2e-16 ) is too small and is less than alpha.
# We can say that, we have significant evidence at alpha = 0.05 level that there is a linear association between temperature and rental count.
# Here the beta1 estimate is positive, which indicates a positive linear relationship.
# Beta1 = 6803, this can be interpreted as for every 1 degree celsius increase in temperature, the rental count increases by around 6803 on average.


######################
```

```
#        Research Q-2        #
#####################

### Does the rental count varies based on the weather conditions.
### If yes, which two different weather conditions show a significant difference in average rental count.

# In order to answer this question, we will perform a global F-test first using anova model to see if there is a
# significant difference in the average rental count due to weather change.
# If the results come significant, we will perform a pairwise comparisons by adjusting for Type-I error using TukeyHSD to
see which two weather conditions show a significant difference.

# Our null hypothesis will be -
# H0: μ_clear = μ_cloudy = μ_lightSnow = μ_heavyRain = 0 (No difference in average rental count due to different
weather conditions)
# H1: μi ≠ μj for some i and j  (Difference in average rental count in at least two of the weather conditions)
# alpha = 0.05

anova.model <- aov(bike_data$count~bike_data$weather)
summary(anova.model)

# Reject the null hypothesis, since p-value (2e-16) is quite small and also less than alpha level.
# From global test, we can say that we have significant evidence at alpha = 0.05 level that there is a significant difference
in average rental count based on different weather conditions.

# In order to see under which two weather conditions are these counts differ, we will perform pairwise comparison test
using TukeyHSD.
TukeyHSD(anova.model)

# Conclusions:
# We can see the p-value (1.3e-14) is significant i.e., less than alpha level.
# So, we can say here that there is a significant difference in average rental count between cloudy - clear, Light Snow -
Clear
# and LighSnow and Cloudy weather conditions.
# All of the different weather conditions show a significant difference in rental count.


#####################
#        Research Q-3        #
#####################

### What is the effect of weather conditions on rental count after adjusting for temperature.

# To answer this question, we will perform ANCOVA test and check whether the differences due to weather conditions are
significant alone
# or does temperature plays a part in it.

Anova(lm(bike_data$count ~ bike_data$weather + bike_data$temperature), type = 3)

### Conclusion:
# After adjusting for temperature using an ANCOVA model, we can see the differences in average rental count due to
different
# weather conditions are still significant.
# Temperature has no effect on the results that we got from ANOVA model.
# So, we can say that temperature is not a co-founding variable here.
```

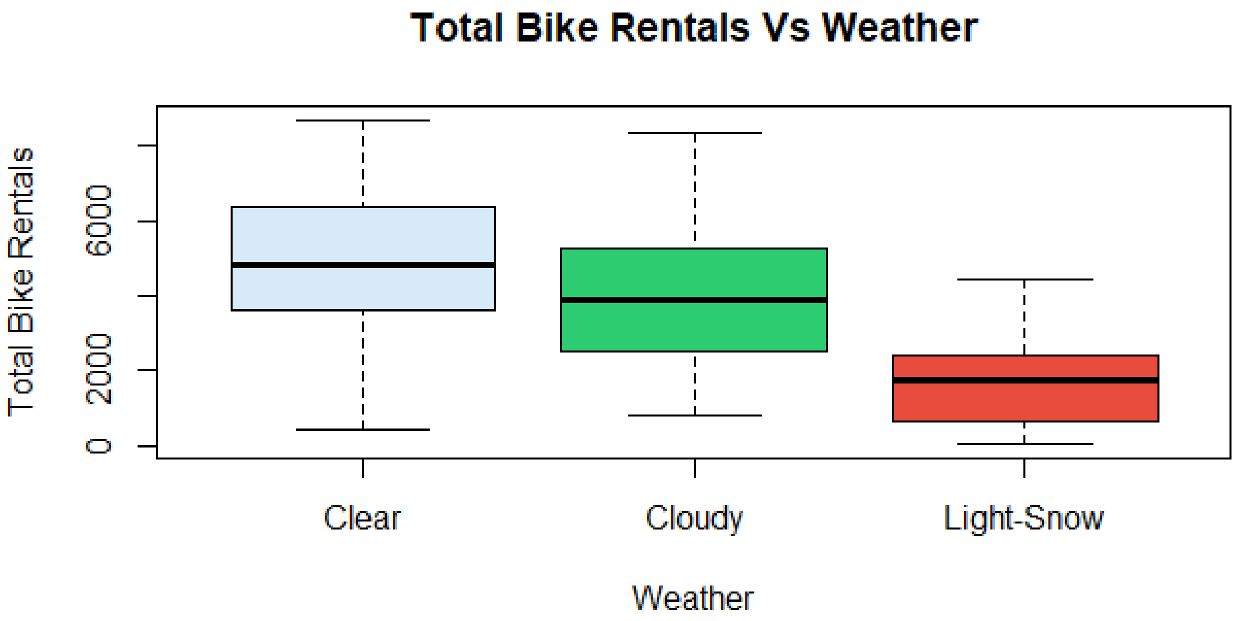
5. Execute your R code, Copy and Paste results here in this Box.

Run your code and copy the output of your code to here.

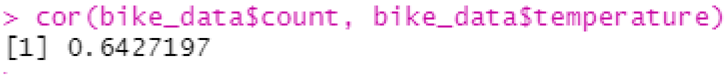
1. EDA



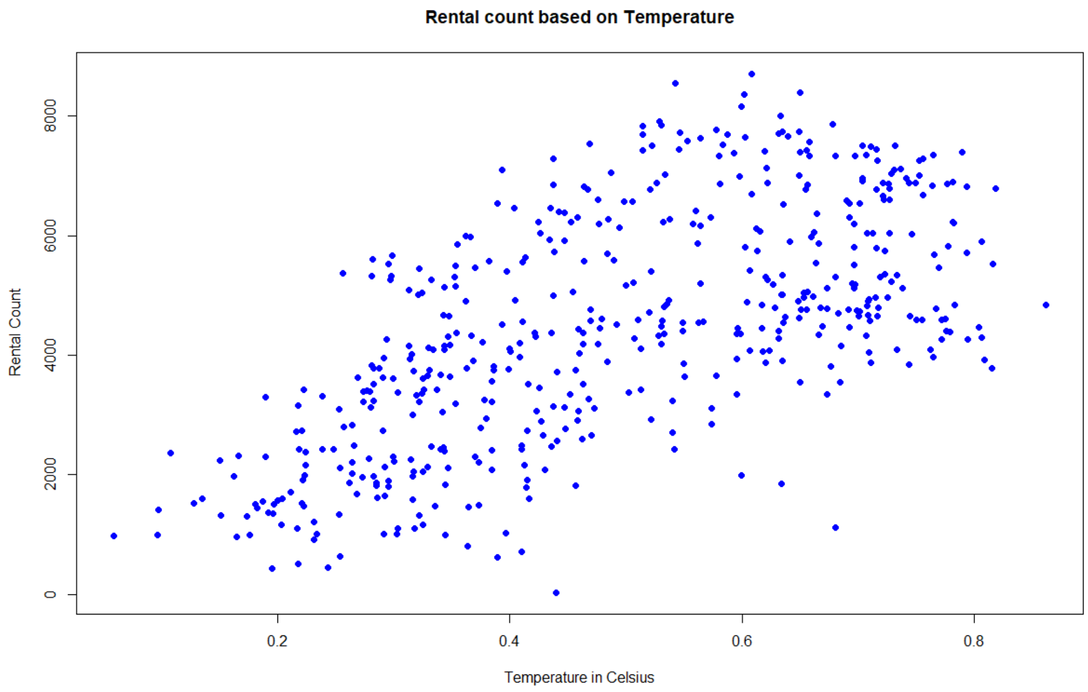
```
> #Categorical variable
> aggregate(bike_data$count, by=list(bike_data$weather), summary)
     Group.1   x.Min.  x.1st Qu. x.Median   x.Mean x.3rd Qu.   x.Max.
1      Clear  431.000  3615.500 4854.500 4850.684  6355.500 8714.000
2     Cloudy  801.000  2538.250 3911.000 3993.151  5274.750 8362.000
3 Light-Snow   22.000   646.500 1751.000 1667.722  2360.250 4459.000
> aggregate(bike_data$count, by=list(bike_data$weather), sd)
     Group.1        x
1      Clear 1837.230
2     Cloudy 1832.555
3 Light-Snow 1111.490
```



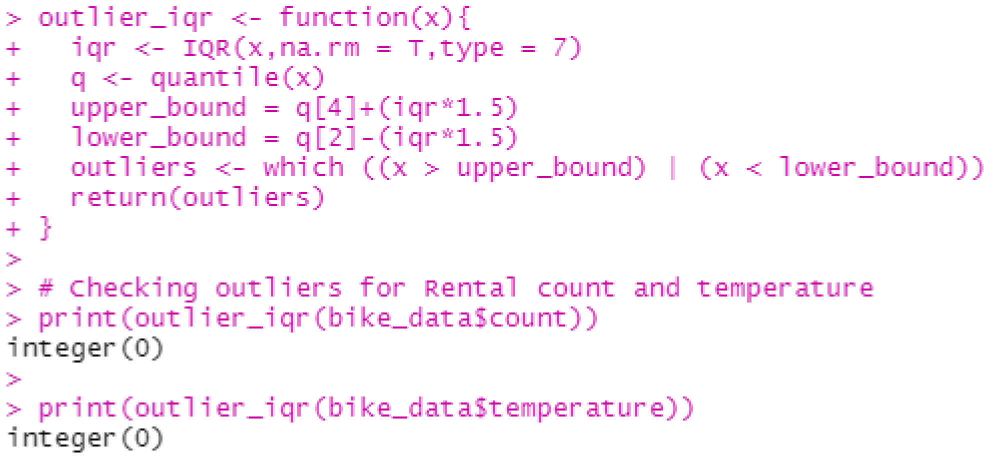
Continuous Variable – EDA



```
> cor(bike_data$count, bike_data$temperature)
[1] 0.6427197
```



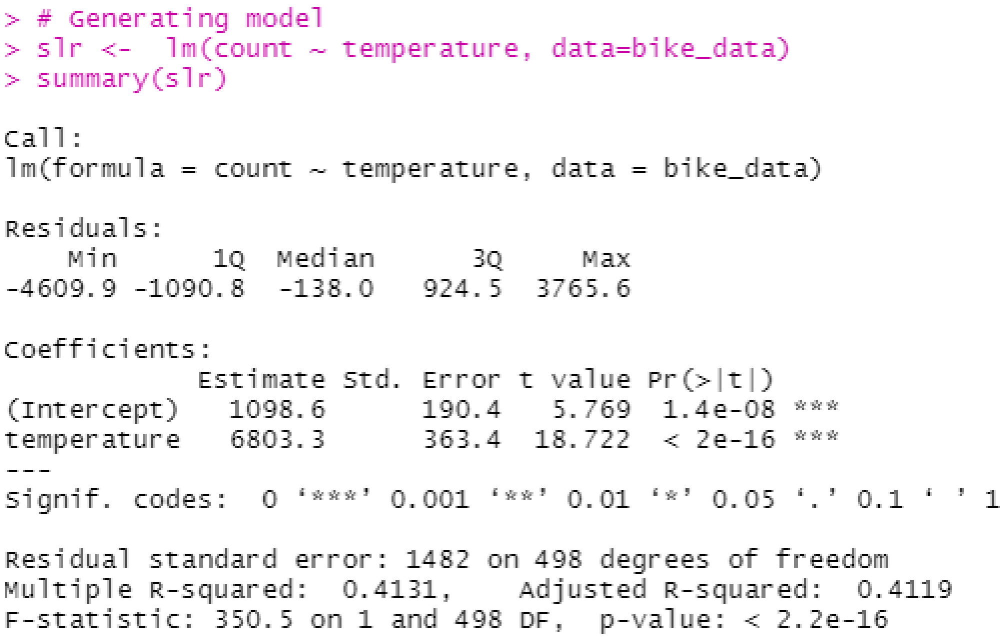
2. Outlier Test



```
> outlier_iqr <- function(x){
+   iqr <- IQR(x,na.rm = T,type = 7)
+   q <- quantile(x)
+   upper_bound = q[4]+(iqr*1.5)
+   lower_bound = q[2]-(iqr*1.5)
+   outliers <- which ((x > upper_bound) | (x < lower_bound))
+   return(outliers)
+ }
>
> # Checking outliers for Rental count and temperature
> print(outlier_iqr(bike_data$count))
integer(0)
>
> print(outlier_iqr(bike_data$temperature))
integer(0)
```

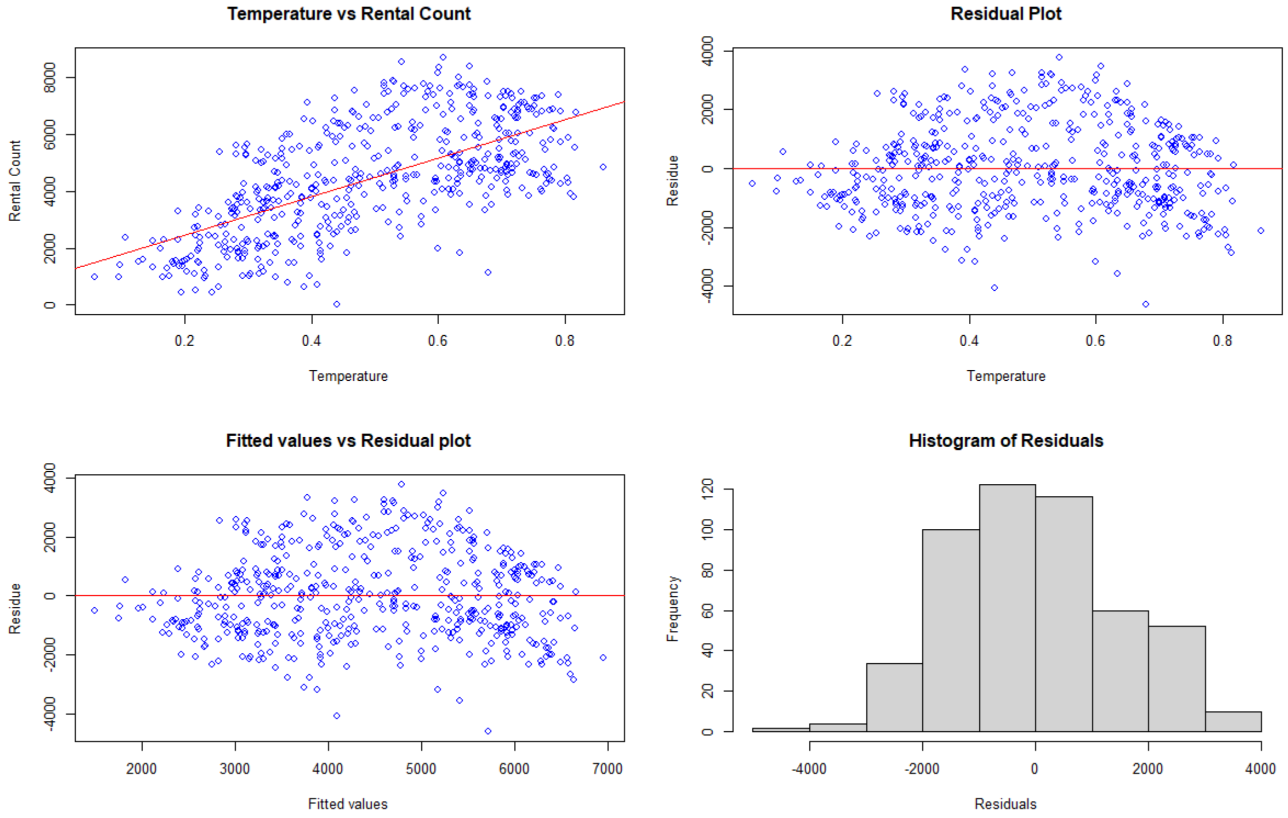
3. Research Question-1

Is temperature a significant predictor for rental count?

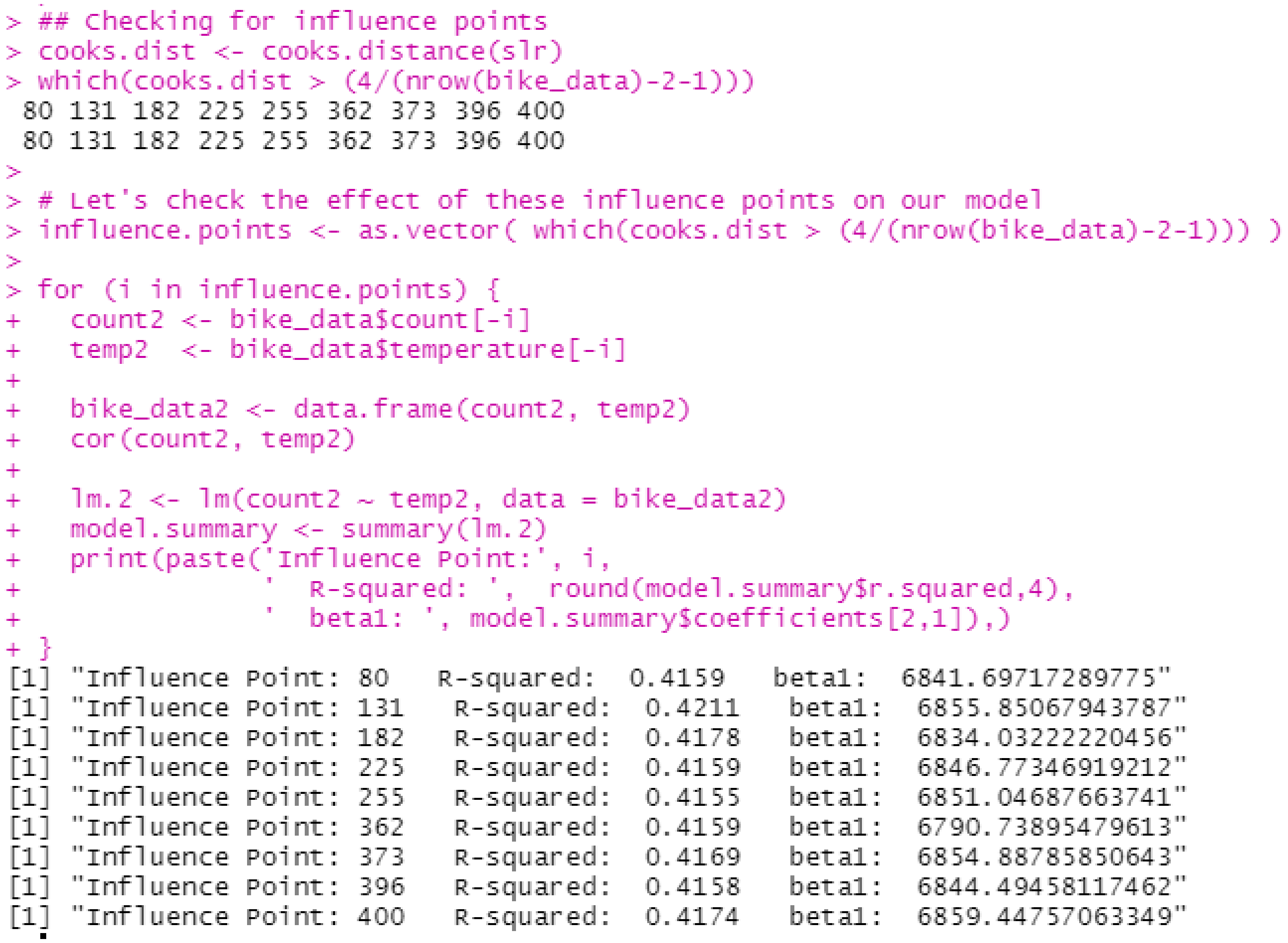


```
> # Generating model
> slr <-  lm(count ~ temperature, data=bike_data)
> summary(slr)

Call:
lm(formula = count ~ temperature, data = bike_data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4609.9 -1090.8  -138.0   924.5  3765.6 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1098.6      190.4   5.769  1.4e-08 ***
temperature   6803.3      363.4  18.722  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1482 on 498 degrees of freedom
Multiple R-squared:  0.4131,	Adjusted R-squared:  0.4119 
F-statistic: 350.5 on 1 and 498 DF,  p-value: < 2.2e-16
```

Regression Diagnostic for Simple Linear regression Model



Check for Influence Points



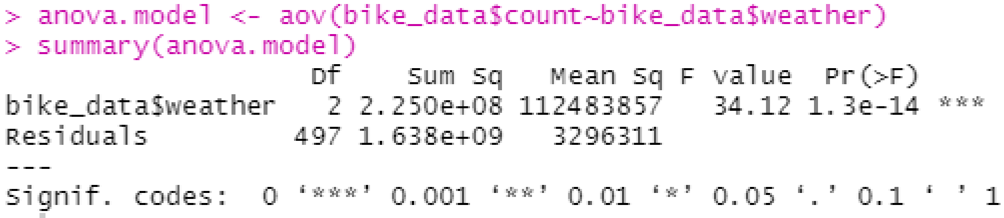
```
> ## Checking for influence points
> cooks.dist <- cooks.distance(slr)
> which(cooks.dist > (4/(nrow(bike_data)-2-1)))
 80 131 182 225 255 362 373 396 400
 80 131 182 225 255 362 373 396 400
>
> # Let's check the effect of these influence points on our model
> influence.points <- as.vector( which(cooks.dist > (4/(nrow(bike_data)-2-1))) )
>
> for (i in influence.points) {
+   count2 <- bike_data$count[-i]
+   temp2  <- bike_data$temperature[-i]
+
+   bike_data2 <- data.frame(count2, temp2)
+   cor(count2, temp2)
+
+   lm.2 <- lm(count2 ~ temp2, data = bike_data2)
+   model.summary <- summary(lm.2)
+   print(paste('Influence Point:', i,
+               '   R-squared: ',  round(model.summary$r.squared,4),
+               '   beta1: ', model.summary$coefficients[2,1]),)
+ }
[1] "Influence Point: 80   R-squared:  0.4159   beta1:  6841.69717289775"
[1] "Influence Point: 131   R-squared:  0.4211   beta1:  6855.85067943787"
[1] "Influence Point: 182   R-squared:  0.4178   beta1:  6834.03222220456"
[1] "Influence Point: 225   R-squared:  0.4159   beta1:  6846.77346919212"
[1] "Influence Point: 255   R-squared:  0.4155   beta1:  6851.04687663741"
[1] "Influence Point: 362   R-squared:  0.4159   beta1:  6790.73895479613"
[1] "Influence Point: 373   R-squared:  0.4169   beta1:  6854.88785850643"
[1] "Influence Point: 396   R-squared:  0.4158   beta1:  6844.49458117462"
[1] "Influence Point: 400   R-squared:  0.4174   beta1:  6859.44757063349"
```

4. Research Question-2

Does the rental count varies based on the weather conditions?

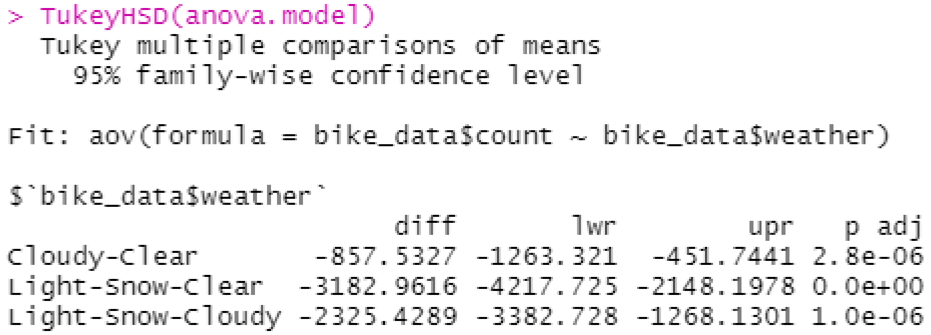
If yes, which two different weather conditions show a significant difference in average rental count.

Global Test



```
> anova.model <- aov(bike_data$count~bike_data$weather)
> summary(anova.model)
                   Df    Sum Sq   Mean Sq F value  Pr(>F)    
bike_data$weather   2 2.250e+08 112483857   34.12 1.3e-14 ***
Residuals         497 1.638e+09   3296311                    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Pairwise Test after controlling for Type-I error

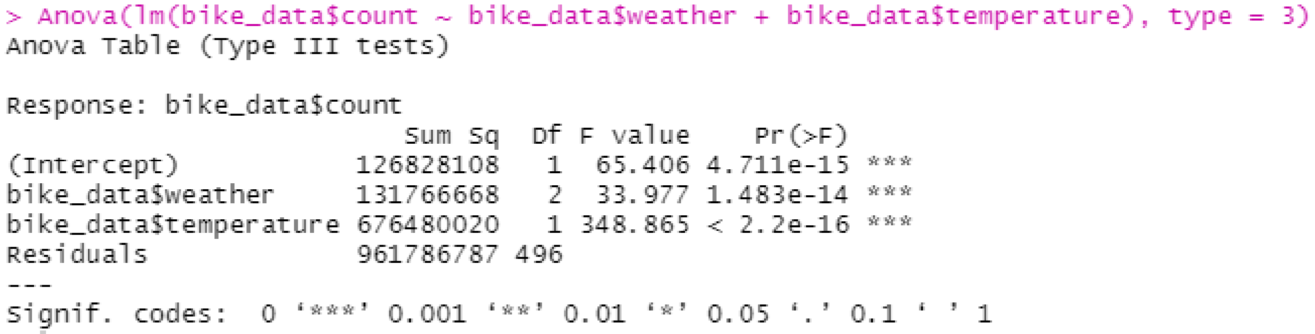


```
> TukeyHSD(anova.model)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = bike_data$count ~ bike_data$weather)

$`bike_data$weather`
                         diff       lwr        upr   p adj
Cloudy-Clear        -857.5327 -1263.321  -451.7441 2.8e-06
Light-Snow-Clear   -3182.9616 -4217.725 -2148.1978 0.0e+00
Light-Snow-Cloudy  -2325.4289 -3382.728 -1268.1301 1.0e-06
```

5. Research Ques- 3:

What is the effect of weather conditions on rental count after adjusting for temperature?

ANCOVA test controlling for temperature



```
> Anova(lm(bike_data$count ~ bike_data$weather + bike_data$temperature), type = 3)
Anova Table (Type III tests)

Response: bike_data$count
                         Sum Sq  Df F value    Pr(>F)    
(Intercept)           126828108   1  65.406 4.711e-15 ***
bike_data$weather     131766668   2  33.977 1.483e-14 ***
bike_data$temperature 676480020   1 348.865 < 2.2e-16 ***
Residuals             961786787 496                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

6. State Your Conclusion (no more than 100 words)

State the conclusion so that a none-statistician can understand.

Conclusion:

1. After performing Simple Linear regression and testing for model assumptions, we concluded that we reject the null hypothesis, since p-value (2e-16) was too small and less than alpha (0.05).
We have significant evidence at alpha = 0.05 level that there is a linear association between temperature and rental count.

2. From the one-way ANOVA analysis, we reject the null hypothesis at alpha = 0.05 level, since p-value came out be (2e-16) which is less than 0.05.
So, we can say that we have significant evidence at alpha = 0.05 level that there is a difference in average rental count due to different weather conditions.
After performing the pair-wise comparisons using TukeyHSD method, we find that all of the different pairs of weather conditions (Clear, Cloudy, Light-Snow) show a significant difference in average rental count.

3. After adjusting for temperature and using ANCOVA model, we concluded that the differences we saw between the weather conditions in the ANOVA model were significant and temperature had no effect on them.
We can still see the differences in average rental count for different weather after adjusting for temperature, since p-value was (1.483e-14) < 0.05. So, we can say temperature is not a co-founding variable in this case.