ANOVA Craigs List Automobile Analysis:

Dataset: This file contains information on 46,484 vehicles listed for sale on Craig's List in the United States. Create a single data frame for the analysis which meets the following characteristics:

1. Only includes cars with 4, 6, or 8 cylinder engines.
2. Only includes cars using gasoline or diesel as fuel.
3. Includes all variables appearing in the master (N=46,484) data set.
4. Has a random sample of n=150 cars from each of three states: Illinois, North Carolina, and Texas. The lists at the end of this assignment specify the regions which will aggregate to represent the three states.
5. Include a new variable identifying the state from which a car has been drawn. This will be a factor variable with the levels "Illinois", "Texas", and "North Carolina".

Problem Statement: Do an ANOVA Analysis in R and answer the following questions.

Analysis

1. Within the n=450 stratified sample, determine if asking.price has an equal variance across the three states.
2. Conduct a one-way analysis of variance with asking.price as the dependent variable and state as the independent variable. Plot the results of a Tukey HSD test to show whether/where differences in asking.price among the states exist. Explain the results shown in the plot, stating which pairs of states which do and do not appear to show significant mean differences in asking.price.
3. Repeat above steps using odometer as the dependent variable and state as the independent. Again, briefly explain the analysis results.
4. Drawing on a n=150 sample, use only the vehicles for sale in the state of Texas to conduct a one-way ANOVA using asking.price as the dependent variable and region as the independent variable. Plot the results of a Tukey HSD test to show whether/where there are differences in asking.price among the regions of Texas. Explain the results shown in the plot, stating which regions do appear to show significant mean differences in asking.price.
5. Using the n=450 sample conduct a single ANOVA using asking.price as the dependent variable and fuel and condition as independent variables. Plot the results of a Tukey HSD test to show whether/where there are differences in asking.price by independent variables. Make certain both Tukey plots are visible on the same graphic.

I. Preprocessing

```r
#Author: Suryateja Chalapati

#Importing required libraries
rm(list=ls())
library(rio)
library(moments)
library(dplyr)
library(data.table)
library(car)
library(stringr)

#Setting the working directory and importing the dataset
setwd("C:/Users/surya/Downloads")

df = import("Automobile Data.xlsx", sheet = "Sheet1")
colnames(df)=tolower(make.names(colnames(df)))

#Filtering dataset based on multiple conditions
df$region[df$region == "dallas / fort worth"] <- "dallas / fort worth, TX"
df$region[df$region == "odessa / midland"] <- "odessa / midland, TX"
df$region[df$region == "champaign urbana"] <- "champaign urbana, IL"
df$region[df$region == "chicago"] <- "chicago, IL"
df$region[df$region == "danville"] <- "danville, IL"
```

```r
df$region[df$region == "southern illinois"] <- "southern illinois, IL"
df$region[df$region == "quad cities, IA/IL"] <- "quad cities, IL"

df = dplyr::filter(df, grepl('TX|IL|NC', region))

df1 = df[(df$cylinders == 4) | (df$cylinders == 6) | (df$cylinders == 8) ,]

df1 = df[(df$fuel == "gas") | (df$fuel == "diesel") ,]

df1 = setDT(df1)[, paste0('region',1:2):= tstrsplit(region, ',')]

df1$states = str_sub(df1$region,-2,-1)

df1 <- subset(df1, select = -region2)

set.seed(36991670)
df.sample = data.frame(df1[sample(1:nrow(df1), 450, replace = FALSE),])
df.sample = df1 %>% group_by(states) %>% sample_n(150)

table(df.sample$states)
##
##  IL  NC  TX
## 150 150 150

attach(df.sample)
```

II. Analysis

```r
#Analysis_1
leveneTest(asking.price~states,data=df.sample)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   2  6.8272 0.001201 **
##       447
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Levene Test:

From the output we can see the p value is less than 0.05. So, we reject the Null Hypothesis and say that the Variances across the states is not equal.

```r
#Analysis_2
anova.out = aov(asking.price~states,data=df.sample)
summary(anova.out)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## states        2 5.060e+08 252989726   2.443  0.088 .
## Residuals   447 4.628e+10 103536135
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Result:

p-value is more than 0.05 which means that we fail to reject the Null hypothesis which states that all the states means are the same. So, there is no relationship between states and asking.price.

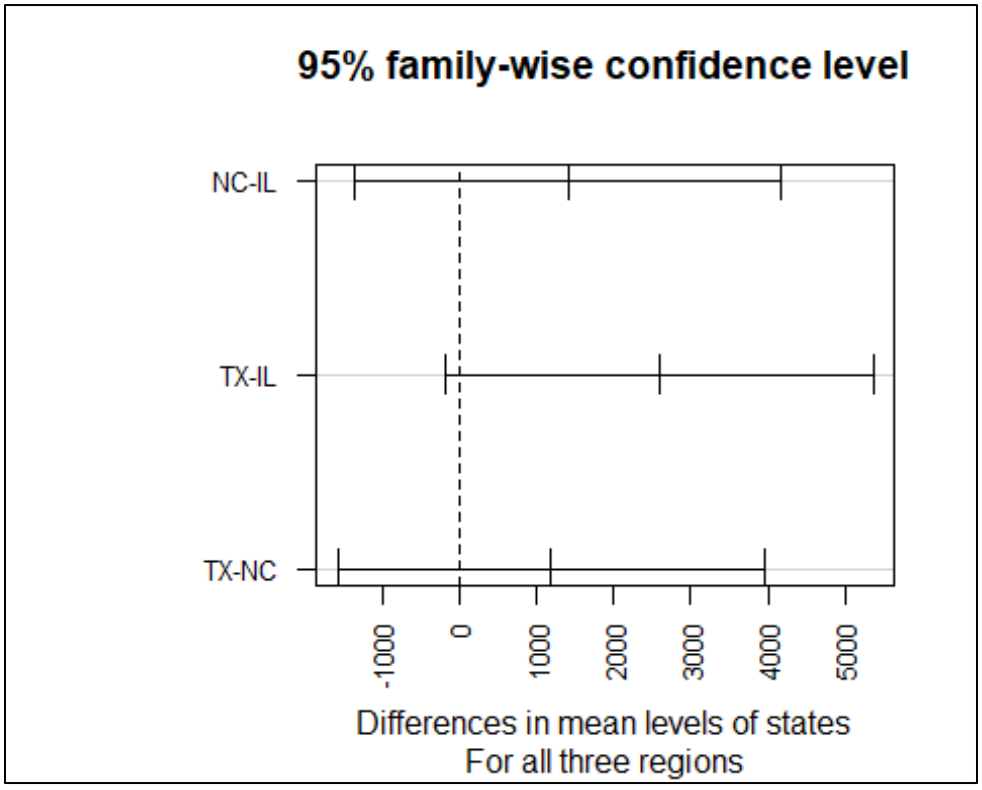
```
tukey.out=TukeyHSD(anova.out)
tukey.out
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = asking.price ~ states, data = df.sample)
##
## $states
##           diff        lwr      upr     p adj
## NC-IL 1404.720 -1358.2031 4167.643 0.4564196
## TX-IL 2594.413  -168.5097 5357.336 0.0708876
## TX-NC 1189.693 -1573.2297 3952.616 0.5693041
```

Tukey Result:

We can see that all rows are not significant as they have p-values which are more than 0.05. We can say there is no relationship between asking.price and states.

```
par(mar=c(5.1,8,4.1,2.1))
plot(TukeyHSD(anova.out), las=2, cex.axis=.8, sub = "For all three regions")
```




```
par(mar=c(5.1,4.1,4.1,2.1))

#Analysis_3
leveneTest(odometer~states,data=df.sample)
```

```
## Warning in leveneTest.default(y = y, group = group, ...): group coerced to
## factor.

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
```

```
## group   2  0.4373 0.6461
##       447
```

Levene Test:

From the output we can see the p value is more than 0.05. So, we fail to reject the Null Hypothesis and say that t he Variances across the states is equal.

```
anova.out1 = aov(odometer~states,data=df.sample)
summary(anova.out1)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## states        2 1.169e+09 5.846e+08   0.152  0.859
## Residuals   447 1.722e+12 3.853e+09
```

ANOVA Result:

p-value is more than 0.05 which means that we fail to reject the Null hypothesis which states that all the states means are the same. So, there is no relationship between states and odometer.

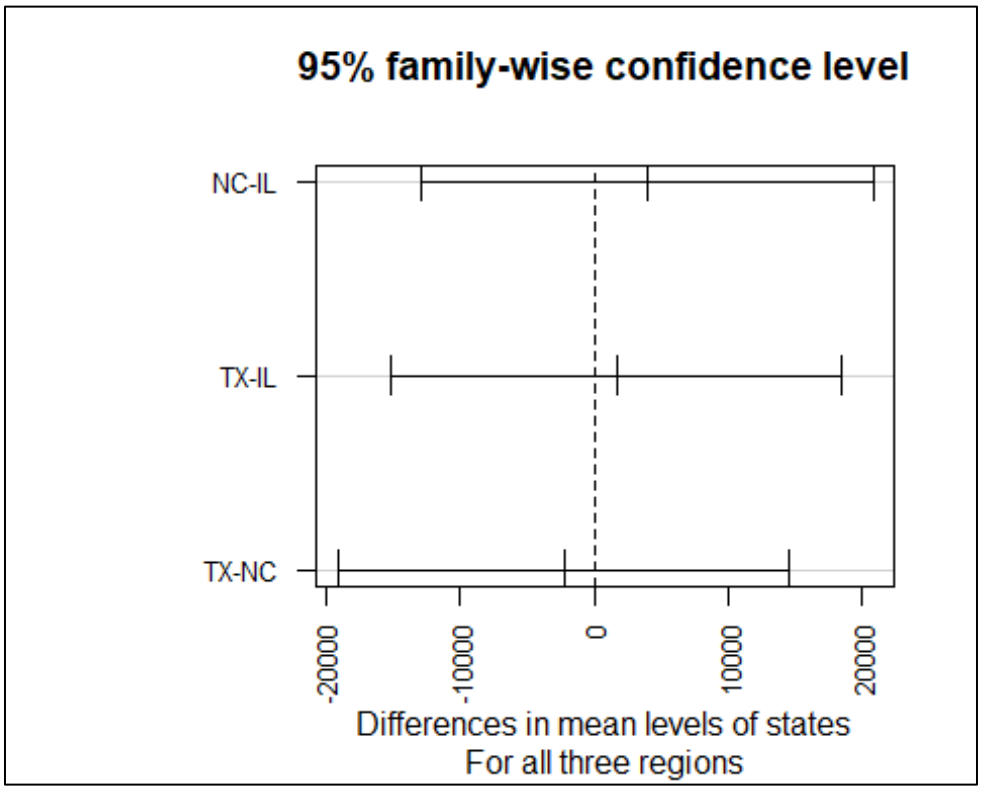
```
tukey.out1=TukeyHSD(anova.out1)
tukey.out1
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = odometer ~ states, data = df.sample)
##
## $states
##            diff       lwr      upr     p adj
## NC-IL  3931.887 -12922.66 20786.44 0.8472598
## TX-IL  1654.847 -15199.70 18509.40 0.9710428
## TX-NC -2277.040 -19131.59 14577.51 0.9458928
```

Tukey Result:

We can see that all rows are not significant as they have p-values which are more than 0.05. We can say there is no relationship between odometer and states.

```
par(mar=c(5.1,8,4.1,2.1))
plot(TukeyHSD(anova.out1), las=2, cex.axis=.8, sub = "For all three regions")
```




```
par(mar=c(5.1,4.1,4.1,2.1))

#Analysis_4
dftx = df.sample[df.sample$states == "TX", ]
```

Taking only Texas as the sample.

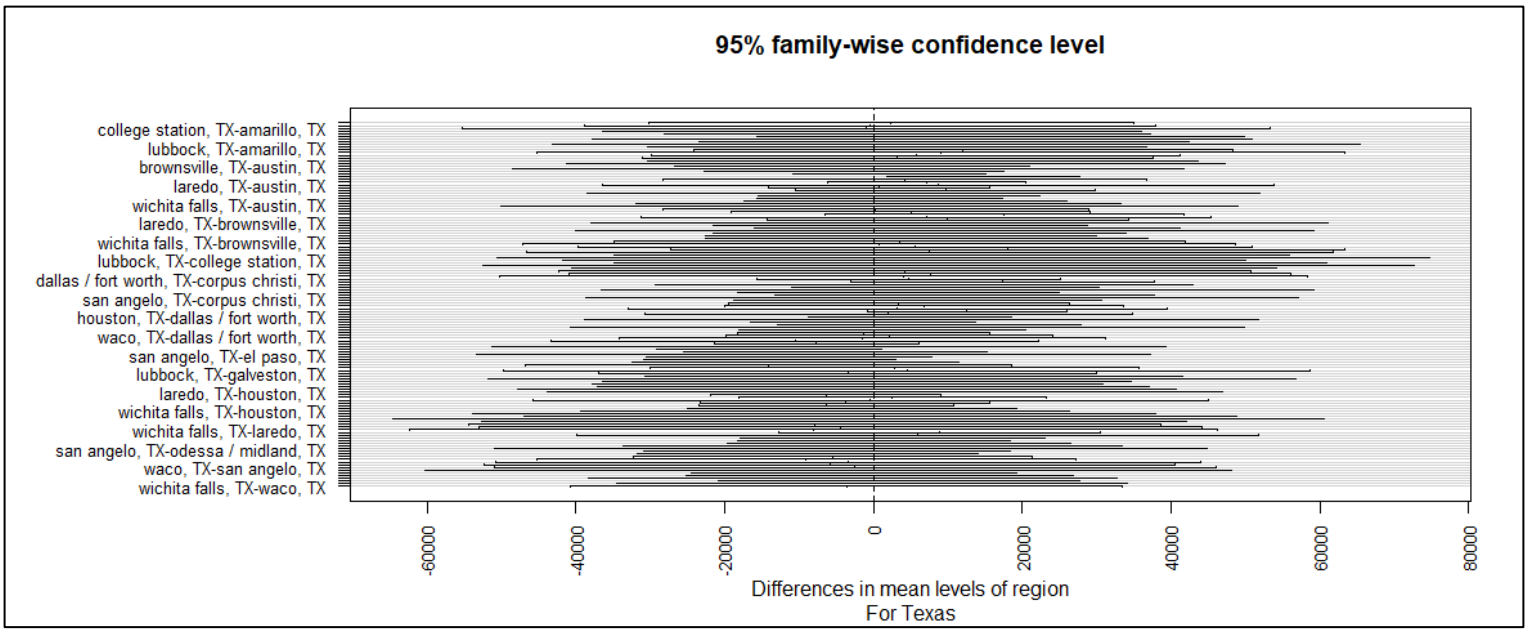
```
anova.tx = aov(asking.price~region, data=dftx)
summary(anova.tx)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## region       16 4.289e+09 268052479   1.698 0.0542 .
## Residuals   133 2.100e+10 157879093
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Result:

p-value is more than 0.05 which means that we fail to reject the Null hypothesis which states that all the regions mean are the same. So, there is no relationship between regions and asking.price.

```
par(mar=c(5.1,15,4.1,2.1))
plot(TukeyHSD(anova.tx), las=2, cex.axis=.8, sub = "For Texas")
```




```
par(mar=c(5.1,4.1,4.1,2.1))

#Analysis_5
anova.two = aov(asking.price~fuel+condition, data=df.sample)
summary(anova.two)
```

```
##              Df    Sum Sq   Mean Sq F value Pr(>F)
## fuel          1 6.966e+09 6.966e+09   95.88 <2e-16 ***
## condition     4 7.565e+09 1.891e+09   26.03 <2e-16 ***
## Residuals   444 3.226e+10 7.265e+07
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

ANOVA Result:

p-value is less than 0.05 which means that we can reject the Null hypothesis which states that all the regions mean are the same. So, there is a relationship between asking.price and fuel and conditions.

```
tukey.two=TukeyHSD(anova.two)
tukey.two
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = asking.price ~ fuel + condition, data = df.sample)
##
## $fuel
##                  diff       lwr       upr p adj
## gas-diesel -15308.42 -18380.91 -12235.92     0
##
## $condition
##                            diff        lwr       upr     p adj
## fair-excellent      -9512.4618 -15077.185 -3947.738 0.0000371
## good-excellent      -6066.5794  -8540.176 -3592.983 0.0000000
## like new-excellent   6111.5256   2450.450  9772.602 0.0000613
## new-excellent       -5783.0934 -29177.723 17611.536 0.9613112
## good-fair            3445.8825  -2257.092  9148.857 0.4631064
## like new-fair       15623.9874   9314.572 21933.403 0.0000000
## new-fair             3729.3684 -20222.958 27681.695 0.9930859
## like new-good       12178.1050   8310.127 16046.083 0.0000000
```

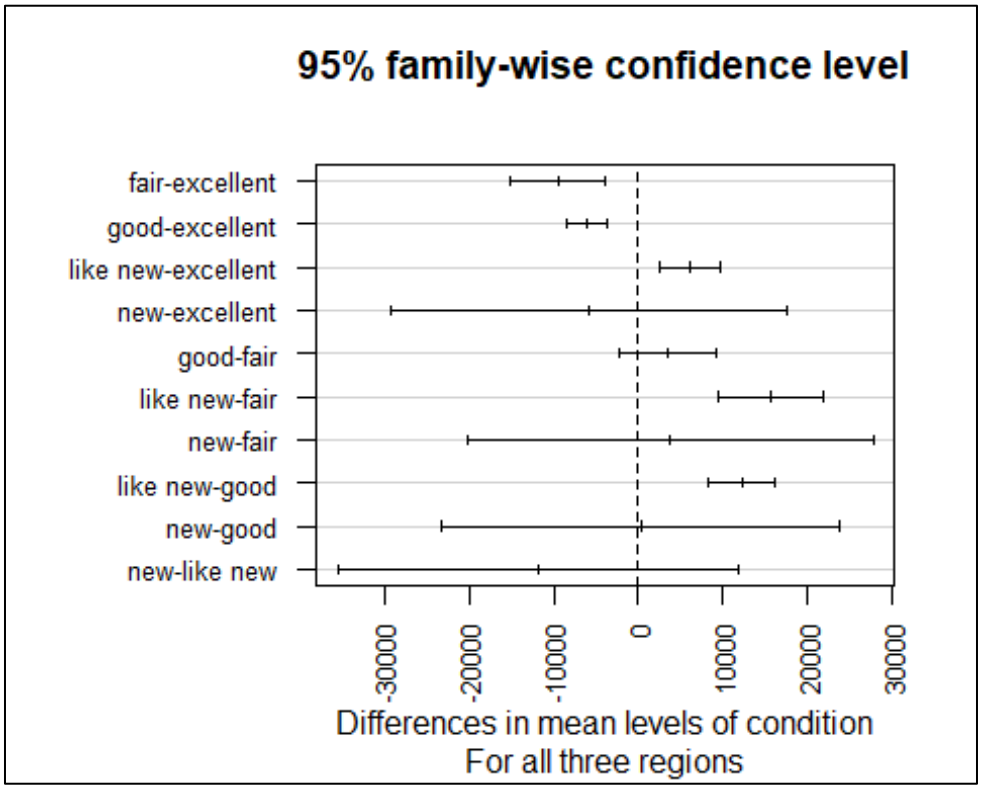
```
## new-good              283.4859 -23144.413 23711.385 0.9999997
## new-like new       -11894.6190 -35477.478 11688.240 0.6398872
```

Tukey Result:

We can see that some rows are significant as they have p-values which are less than 0.05. We can say there is a relationship between asking.price and fuel and conditions. The significant comparisons are as follows:

1. fair-excellent
2. good-excellent
3. like new-excellent
4. like new-fair
5. like new-good

```
par(mar=c(5.1,8,4.1,2.1))
plot(TukeyHSD(anova.two), las=2, cex.axis=.8, sub = "For all three regions")
```




```
par(mar=c(5.1,4.1,4.1,2.1))
```