Name – Manish Sanjay Talekar

U54500765

QMB\_Assignment\_8

#Preprocessing

#q1

**> master\_data <- read\_xlsx("6304 Module 8 Assignment Data.xlsx")**

#q2

**> new\_data <- subset(master\_data, cylinders==4 |**

**cylinders==6 |**

**cylinders==8 )**

**> fuel <- subset(new\_data, fuel=="gas" |**

**fuel=="diesel")**

**> Illinois <- subset(fuel, region=="champaign urbana" |**

**region=="chicago" |**

**region=="danville" |**

**region=="peoria, IL" |**

**region=="quad cities, IA/IL" |**

**region=="rockford, IL" |**

**region=="southern illinois" |**

**region=="springfield, IL")**

**> Illinois$state <- "Illinois"**

**> Texas <- subset(fuel, region=="amarillo, TX" |**

**region=="austin, TX" |**

**region=="brownsville, TX" |**

**region=="college station, TX" |**

**region=="corpus christi, TX" |**

**region=="dallas / fort worth" |**

**region=="el paso, TX" |**

**region=="galveston, TX" |**

**region=="houston, TX" |**

**region=="lubbock, TX" |**

**region=="odessa / midland" |**

**region=="tyler / east TX" |**

**region=="waco, TX")**

**> Texas$state <- "Texas"**

**> North\_Carolina <- subset(fuel, region=="asheville, NC" |**

**region=="boone, NC" |**

**region=="charlotte, NC" |**

**region=="eastern NC" |**

**region=="fayetteville, NC" |**

**region=="greensboro, NC" |**

**region=="wilmington, NC" |**

**region=="winston-salem, NC")**

**> North\_Carolina$state <- "North Carolina"**

**> set.seed(54500765)**

**> illinois\_sample <- sample\_n(Illinois, 150)**

**> texas\_sample <- sample\_n(Texas, 150)**

**> nc\_sample <- sample\_n(North\_Carolina, 150)**

**> primary\_data <- rbind(illinois\_sample, texas\_sample, nc\_sample)**

**> primary\_data$state <- as.factor(primary\_data$state)**

#Analysis

#q1

**> leveneTest(asking.price~state, data=primary\_data)**

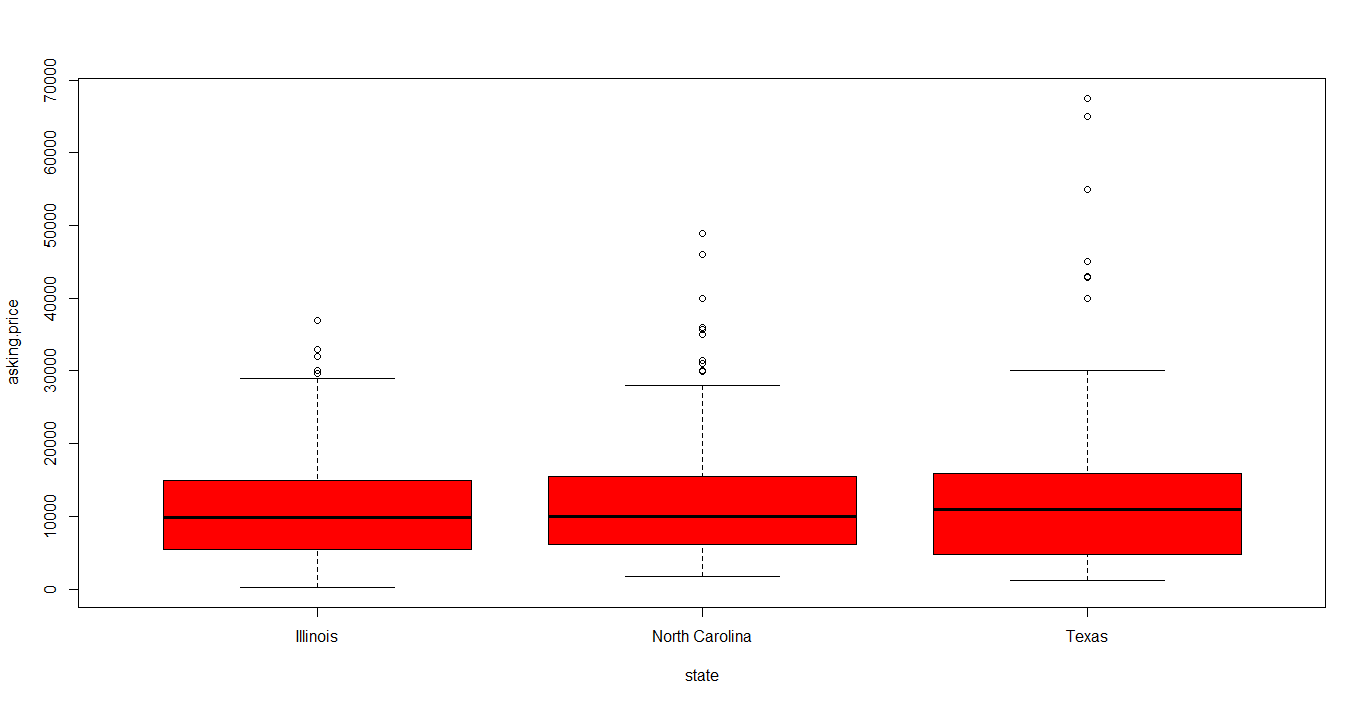
**Levene's Test for Homogeneity of Variance (center = median)**

**Df F value Pr(>F)**

**group 2 1.7967 0.167**

**447**

**> boxplot(asking.price~state, data=primary\_data, col="red")**



INTERPRETATION – Since the p values is greater than the significance level of 0.05, I can conclude that we fail to reject the null hypothesis and therefore all the variances are equal. This can be further confirmed from the above boxplot.

#q2

**> states\_out <- aov(asking.price~state, data=primary\_data)**

**> summary(states\_out)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**state 2 2.684e+08 134218738 1.44 0.238**

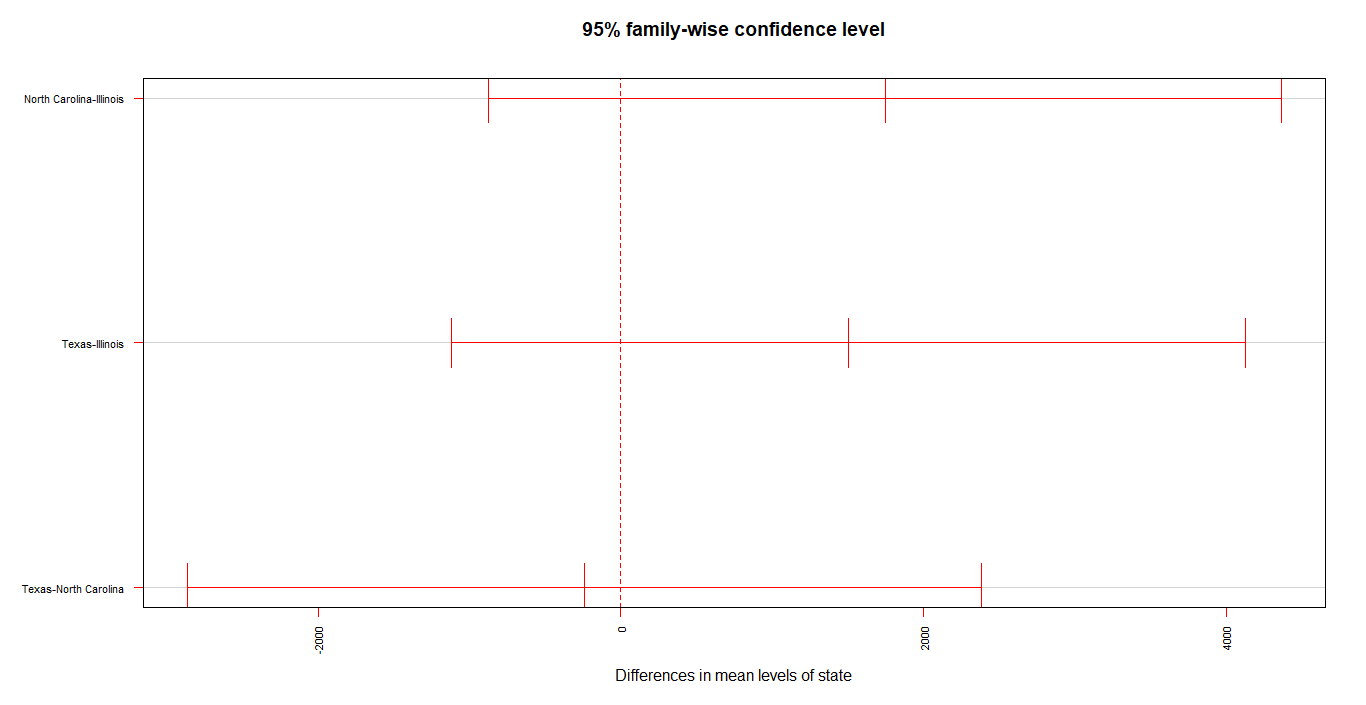
**Residuals 447 4.166e+10 93202868**

**> states1 <- TukeyHSD(states\_out)**

**> par(mar=c(5.1,7.5,4.1,2.1))**

**> plot(states1,las=2,cex.axis=.7, col="red")**

**> par(mar=c(5.1,4.1,4.1,2.1))**



INTERPRETATION – Based on the above graph I can conclude that none of the states pair shows any significant difference in the asking prices. This can be further confirmed from the p value which is greater than the 0.05. Which means we fail to reject the null hypothesis and all the means have the same values.

#q3

**> leveneTest(odometer~state, data=primary\_data)**

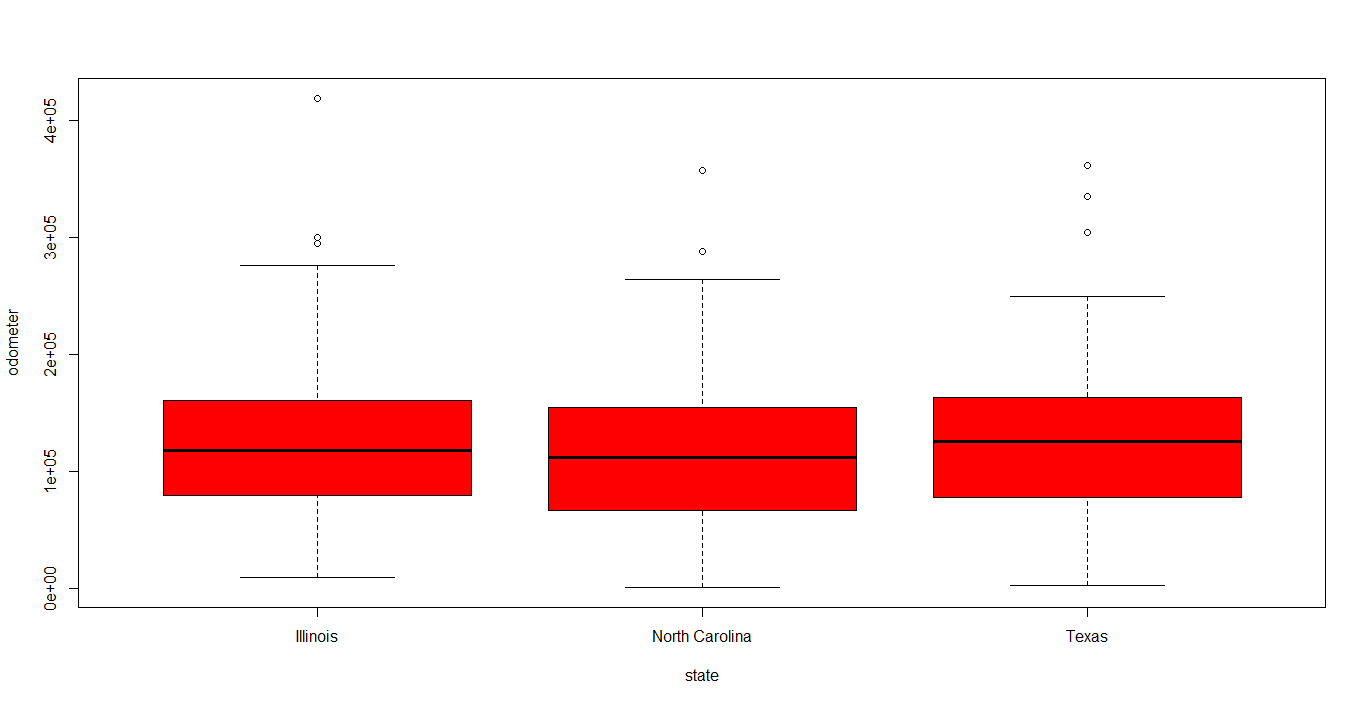
**Levene's Test for Homogeneity of Variance (center = median)**

**Df F value Pr(>F)**

**group 2 0.0254 0.9749**

**447**

**> boxplot(odometer~state, data=primary\_data, col="red")**



**> odometer\_out <- aov(odometer~state, data=primary\_data)**

**> summary(odometer\_out)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**state 2 1.366e+10 6.829e+09 1.703 0.183**

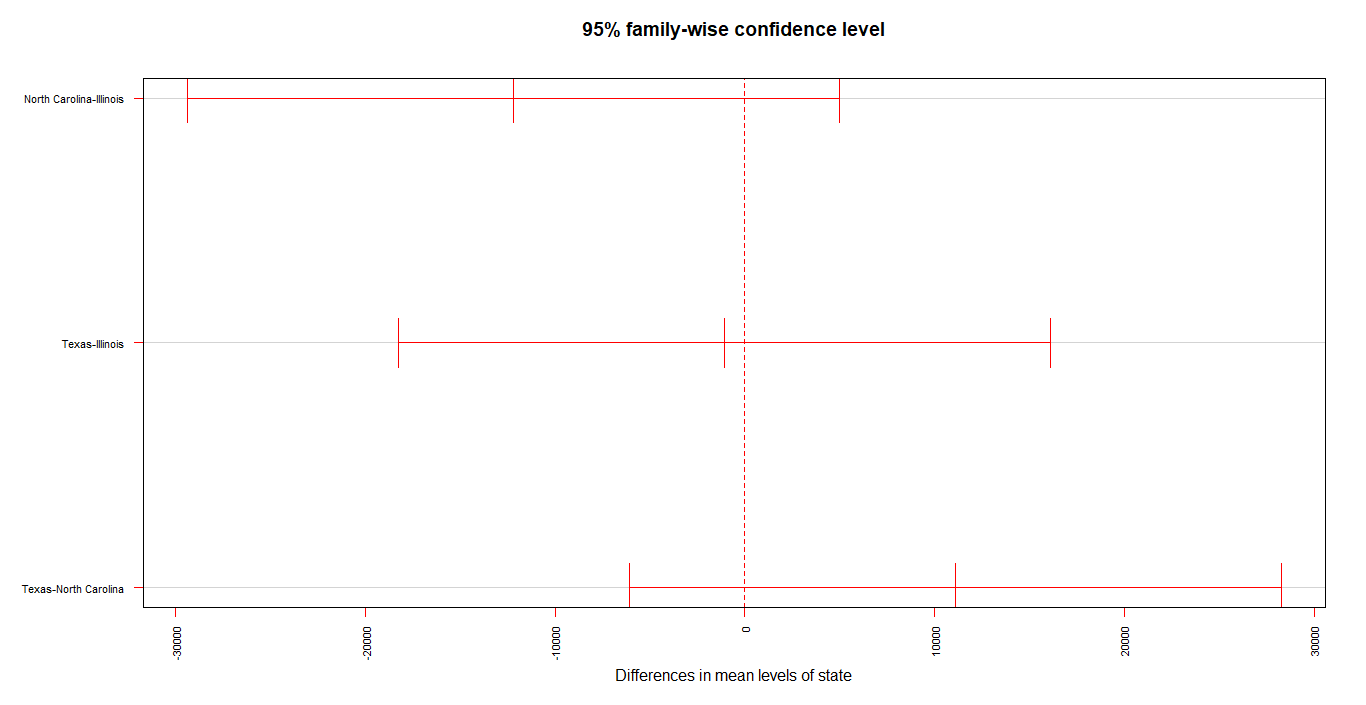
**Residuals 447 1.792e+12 4.009e+09**

**> odometer1 <- TukeyHSD(odometer\_out)**

**> par(mar=c(5.1,7.5,4.1,2.1))**

**> plot(odometer1,las=2,cex.axis=.7, col="red")**

**> par(mar=c(5.1,4.1,4.1,2.1))**



INTERPRETATION – Based on the above graph I can conclude that none of the state’s pair shows any significant difference in the odometer. This can be further confirmed from the p value which is greater than the 0.05. Which means we fail to reject the null hypothesis and all the means have the same values.

#q4

**> texas\_out <- aov(asking.price~region, data=texas\_sample)**

**> summary(texas\_out)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**region 12 1.192e+09 99357721 0.755 0.695**

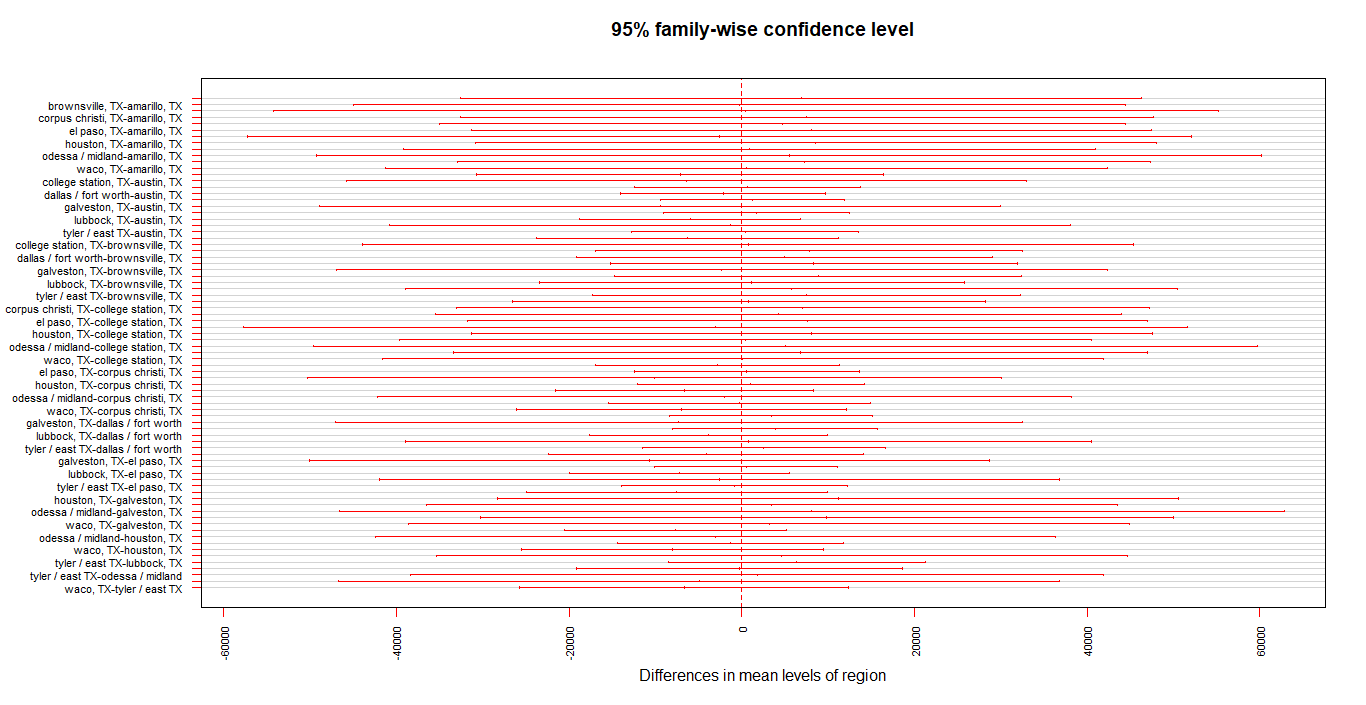
**Residuals 137 1.802e+10 131546452**

**> texas1 <- TukeyHSD(texas\_out)**

**> par(mar=c(5.1,10.5,4.1,2.1))**

**> plot(texas1,las=2,cex.axis=.7, col="red")**

**> par(mar=c(5.1,4.1,4.1,2.1))**



INTERPRETATION – Based on the above graph I can conclude that none of the region’s pair shows any significant difference in the asking price. This can be further confirmed from the p value which is greater than the 0.05. Which means we fail to reject the null hypothesis and all the means have the same values.

#q5

**> all\_out <- aov(asking.price~fuel+condition, data=primary\_data)**

**> summary(all\_out)**

**Df Sum Sq Mean Sq F value Pr(>F)**

**fuel 1 5.204e+09 5.204e+09 74.36 < 2e-16 \*\*\***

**condition 4 5.658e+09 1.415e+09 20.22 2.6e-15 \*\*\***

**Residuals 444 3.107e+10 6.997e+07**

**---**

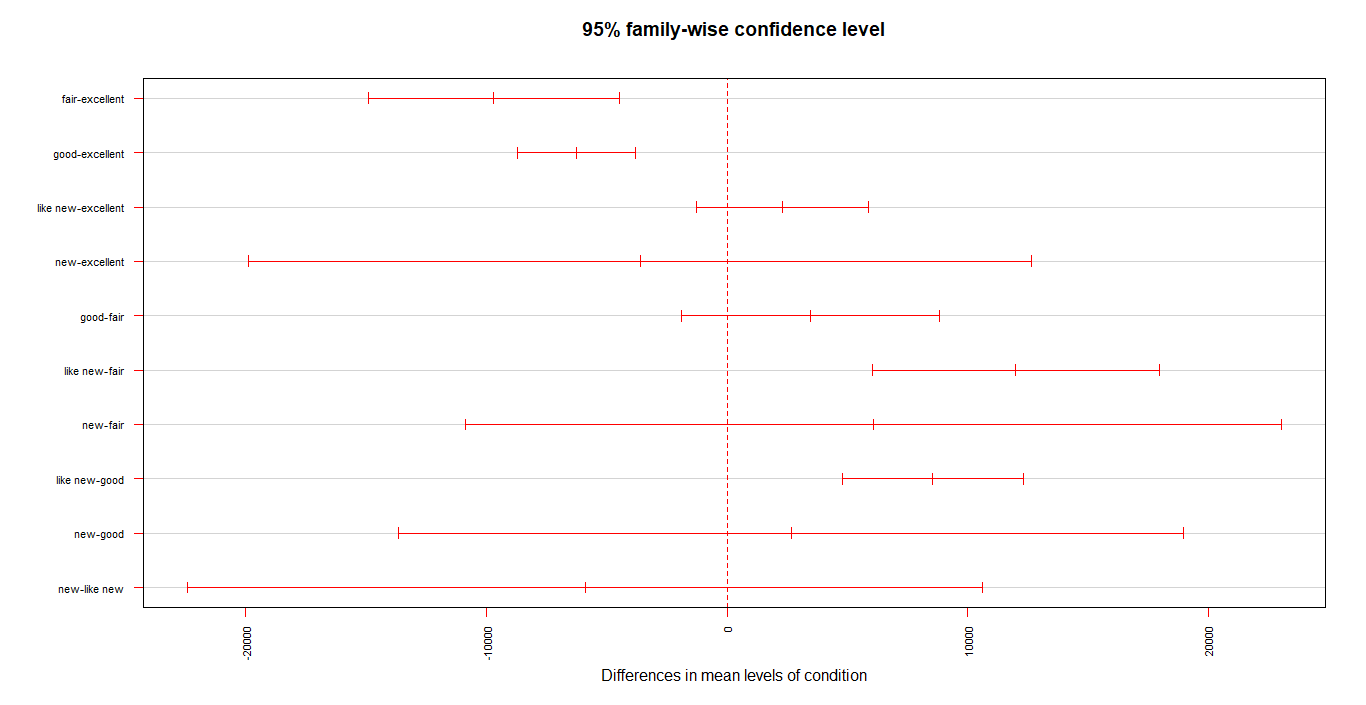
**Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1**

**> all1 <- TukeyHSD(all\_out)**

**> par(mar=c(5.1,7.5,4.1,2.1))**

**> plot(all1,las=2,cex.axis=.7, col="red")**

**> par(mar=c(5.1,4.1,4.1,2.1))**



INTERPRETATION – Based on the p values of fuel and condition we can conclude that we reject the null hypothesis, and it means that at least one of the means is not equal to the other means.

After studying the above plot, we can conclude that the following pairs show significant difference in the means,

1. Fair-excellent
2. Good-excellent
3. Like new-fair
4. Like new-good