**6304 Assignment 9 Submit Work**

**Surya Teja Pidakala**

**U99286608**

1. Load the file “6304 Module 9 Assignment Data.xlsx” into R. This file contains information on 46,484 vehicles listed for sale on Craig’s List in the United States.

**CODE:**

#Surya Teja Pidakala

#U99286608

rm(list=ls())

library(rio)

library(moments)

Surya\_Exam9=import("C:\\Users\\suria\\Downloads\\6304 Module 9 Assignment Data.xlsx")

**OUTPUT:**

Graphical user interface, application

Description automatically generated

1. Create a single data frame for your 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 you will aggregate to represent the three states. Remember to apply the numerical portion of your U number as the random number seed. This type of sample is referred to as a stratified sample.
   5. Includes 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".

There are several ways this can be accomplished. Carefully plan your method for creating this 3-state n=450 data set.

**CODE:**

suryas\_data.df = subset(Surya\_Exam9, cylinders == 4 |

cylinders == 6 |

cylinders == 8)

suryas\_data.df = subset(suryas\_data.df, fuel =='gas' |

fuel =='diesel')

IllinoisData = subset(suryas\_data.df, region =='champaign urbana' |

region == 'chicago'|

region == 'danville' |

region == 'peoria, IL' |

region == 'quad cities, IA/IL' |

region == 'rockford, IL' |

region == 'southern illinois' |

region == 'springfield, IL')

TexasData = subset(suryas\_data.df, 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')

NCData = subset(suryas\_data.df, 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')

set.seed(99286608)

Illinois\_sset = IllinoisData[sample(1:nrow(IllinoisData), 150, replace = FALSE),]

Texas\_sset = TexasData[sample(1:nrow(TexasData), 150, replace = FALSE),]

NC\_sset = NCData[sample(1:nrow(NCData), 150, replace = FALSE),]

MainData = rbind(Illinois\_sset, Texas\_sset, NC\_sset)

str(MainData)

MainData$region = as.factor(MainData$region)

**OUTPUT:**

Table

Description automatically generated

Text, letter

Description automatically generated

**ANALYSIS**

1. Within your n=450 stratified sample, determine if asking.price has an equal variance across the three states. Briefly interpret your results. If you determine there is a difference in variances across the three factor levels state where the difference(s) is/are.

**CODE:**

library(car)

library(dplyr)

car::leveneTest(asking.price~region, MainData)

**OUTPUT:**

Text

Description automatically generated

**INTERPRETATIONS:**

* From the above results we can see that the P-Value is 0.2849 which is greater than 0.05.
* Since the p value is greater than 0.05. we fail to reject the Null Hypothesis.
* This also opens the possibility that the asking price of the states may vary.

1. Using your sample (n=450) data set 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. Briefly 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. Make sure state names can be clearly and completely read on the appropriate axis of your plot.

**CODE:**

AnovaTest=aov(asking.price~state, MainData)

summary(AnovaTest)

Tukey\_Analysis=TukeyHSD(AnovaTest)

Tukey\_Analysis

plot(Tukey\_Analysis)

**OUTPUT:**

Text

Description automatically generated

**GRAPHS:**

Chart, box and whisker chart

Description automatically generated

**INTERPRETATIONS:**

* The P-Values of Texas-NorthCarolina, Texas-Illinois, NorthCarolina-Illinois are 0.92246, 0.85772, 0.98845.
* None of the p-values of the states are close to 0.05. So, we can consider them as insignificant.

1. Repeat Steps 1 and 2 above using odometer as the dependent variable and state as the independent. Again, briefly explain your analysis results and make sure state names can be clearly and completely read on the appropriate axis of your plot.

**CODE:**

car::leveneTest(asking.price~region, MainData)

AnovaTest1=aov(odometer~state, MainData)

summary(AnovaTest1)

Tukey\_Analysis1=TukeyHSD(AnovaTest1)

Tukey\_Analysis1

plot(Tukey\_Analysis)OUTPUT:

**OUTPUTS:**

Text

Description automatically generated

**GRAPHS:**

Chart, box and whisker chart

Description automatically generated

**INTERPRETATIONS**:

* The P-Values of Texas-NorthCarolina, Texas-Illinois, NorthCarolina-Illinois are 0.1965, 0.4557, 0.8569.
* None of the p-values of the states are close to 0.05. So, we can consider them as insignificant.
* Also, we cannot estimate/compare the odometer reading of these states.

1. Drawing on the 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. Briefly explain the results shown in the plot, stating which regions do appear to show significant mean differences in asking.price. Make sure region names can be clearly and completely read on the appropriate axis of your plot.

**CODE:**

AnovaTest2=aov(asking.price~region, Illinois\_sset)

summary(AnovaTest2)

Tukey\_Analysis2=TukeyHSD(AnovaTest2)

Tukey\_Analysis2

plot(Tukey\_Analysis2)

**GRAPHS:**

Chart

Description automatically generated

**OUTPUT:**

Text

Description automatically generated

**INTERPRETATIONS:**

* Above are the P-Values and 95% confidence intervals of all the regions withing Texas.
* Except for the regions of Chicago-champaign urbana, all the regions have insignificant p-values and can be considered insignificant.
* The p-value of Chicago-champaign urbana is 0.05 and the 95% confidence interval of asking.price lies between -14530.22 to 150.2686.

1. Using your n=450 sample conduct an 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 as demonstrated in class. Make sure names of levels of independent variables can be clearly and completely read on the appropriate axis of your plots.

**CODE:**

AnovaTest3=aov(asking.price~fuel+condition, MainData)

summary(AnovaTest3)

Tukey\_Analysis3=TukeyHSD(AnovaTest3)

Tukey\_Analysis3

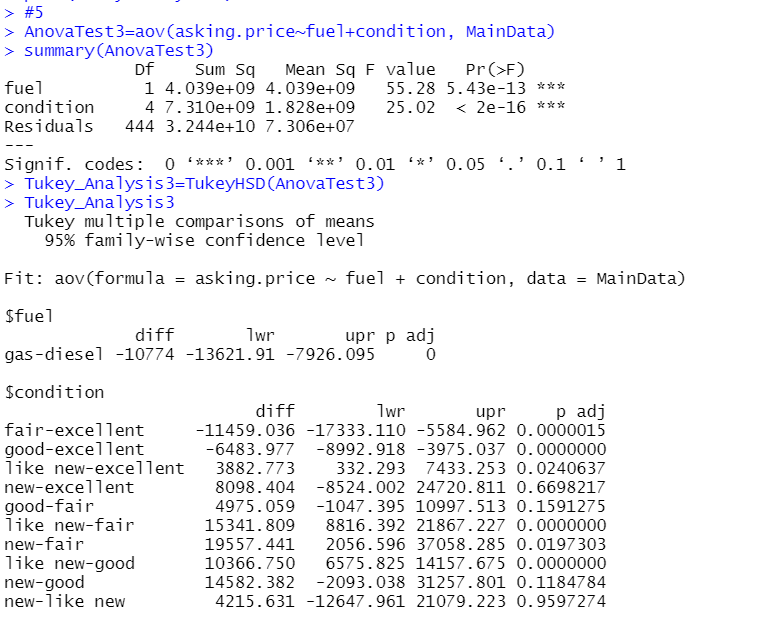
plot(Tukey\_Analysis3)

**GRAPHS:**

Chart, box and whisker chart

Description automatically generated

**OUTPUTS:**

****

**INTERPRETATIONS:**

* Above are the adjusted p-values and confidence intervals of fuel and conditions of cars in 3 states namely Texas, Illinois, and North Carolina.
* Observing the outputs, we can see that none of the p-values are equal or close to 0.05 are insignificant