**Module Assignment**

**Module 9**

**QMB-6304 Analytical Methods for Business**



Write a simple R script to execute the following data preprocessing and statistical analysis. Where required show analytical output and interpretations.

**Preprocessing**

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.
2. 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.

**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.
2. 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.
3. 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.
4. 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.
5. 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.

Your deliverable will be a single MS-Word file created using R Markdown. Your file will show 1) the R script which executes the above instructions and 2) the results of those instructions. The first two lines of your deliverable will state this is “Module 3 Assignment” of our course and your name as it appears in Canvas. Your code chunks and analysis results should be presented in the order in which they are listed here. Deliverable due time will be announced in class and on Canvas. **This is an individual assignment to be completed before you leave the classroom. No collaboration of any sort is allowed on this assignment.**

Regions to be aggregated to represent targeted states.

|  |  |  |
| --- | --- | --- |
| **Texas** | **Illinois** | **North Carolina** |
| amarillo, TX | champaign urbana | asheville, NC |
| austin, TX | chicago | boone, NC |
| brownsville, TX | danville | charlotte, NC |
| college station, TX | peoria, IL | eastern NC |
| corpus christi, TX | quad cities, IA/IL | fayetteville, NC |
| dallas / fort worth | rockford, IL | greensboro, NC |
| el paso, TX | southern illinois | wilmington, NC |
| galveston, TX | springfield, IL | winston-salem, NC |
| houston, TX |  |  |
| lubbock, TX |  |  |
| odessa / midland |  |  |
| tyler / east TX |  |  |
| waco, TX |  |  |

**Preprocessing:**

**#Varun Teja Kolluru**

**#clear all the variable in environment window**

**rm(list=ls())**

**#get all the required libraries**

**library(rio)**

**library(car)**

**#Preprocessing**

**#1**

**#Import the data into R**

**my\_data = import("6304 Module 9 Assignment Data.xlsx")**

**colnames(my\_data)=tolower(make.names(colnames(my\_data)))**

**#2**

**my\_subset=subset(my\_data,**

**(cylinders==4 | cylinders==6 | cylinders==8) &**

**(fuel=="gas" | fuel=="diesel"))**

**my\_texas=subset(my\_subset,**

**(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" ) )**

**#add new state column**

**my\_texas$state =rep("texas",nrow(my\_texas))**

**#Get random sample from the texas subset data**

**set.seed(97)**

**my\_tex\_sample = my\_texas[sample(1:nrow(my\_texas),150,replace=FALSE),]**

**my\_illi=subset(my\_subset,**

**(region=="champaign urbana" |region=="chicago" |region=="danville" |**

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

**region=="southern illinois" |region=="springfield, IL") )**

**#add new state column**

**my\_illi$state=rep('illinois',nrow(my\_illi))**

**#Get random sample from the illinois subset data**

**set.seed(97)**

**my\_ill\_sample = my\_illi[sample(1:nrow(my\_illi),150,replace=FALSE),]**

**my\_nc=subset(my\_subset,**

**(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" ) )**

**#add new state column**

**my\_nc$state =rep("northcarolina",nrow(my\_nc))**

**#Get random sample from the NC subset data**

**set.seed(97)**

**my\_nc\_sample = my\_nc[sample(1:nrow(my\_nc),150,replace=FALSE),]**

**#creating 3 state 450 observation data set**

**my\_sample = rbind(my\_tex\_sample,my\_nc\_sample,my\_ill\_sample)**

**#checking the type of each column**

**str(my\_sample)**

**#change the state column to factor**

**my\_sample$state=as.factor(my\_sample$state)**

**my\_sample$region=as.factor(my\_sample$region)**

**my\_sample$fuel=as.factor(my\_sample$fuel)**

**my\_sample$condition=as.factor(my\_sample$condition)**

Using ‘rm’ command we are clearing all the variables and vectors in the environment window. All the required packages are imported into our R project. Data is imported using the import command.

As per the 2nd question, a subset for cylinders and fuel are taken and then divided the subset into 3 new subset for each state.

After getting the new subsets for each state, a new column ‘state’ is added with respective state name.

Using my last 2 digits U number a random seed is set and a sample of 150 observations are taken for each state.

Using rbind command, a new sample of 450 observations combining all 3states 150 observations are concatenated.

Using ‘str’ command, I have checked for all the data types of the columns which are used in the Analysis part.

‘State’ and ‘region’ columns are ‘char’ data types and these are converted to factors.

**Output in console window:**

> str(my\_sample)

'data.frame': 450 obs. of 11 variables:

$ region : chr "dallas / fort worth" "corpus christi, TX" "el paso, TX" "lubbock, TX" ...

$ asking.price: num 18999 26495 41995 11200 1995 ...

$ year : chr "2015" "2018" "2015" "2013" ...

$ make : chr "dodge" "ford" "gmc" "acura" ...

$ model : chr "charger" "transit passenger wagon" "yukon" "tl" ...

$ condition : chr "excellent" "excellent" "excellent" "excellent" ...

$ cylinders : chr "6" "6" "8" "6" ...

$ fuel : chr "gas" "gas" "gas" "gas" ...

$ odometer : num 84355 25184 73847 53000 255435 ...

$ paint.color : chr "black" "white" "black" "grey" ...

$ state : chr "texas" "texas" "texas" "texas" ...

**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.

**RCode:**

**#Analysis**

**#1**

**leveneTest(asking.price~state,data=my\_sample)**

**boxplot(asking.price~state,**

**main="Price, three state Flavors",**

**col="red",**

**data=my\_sample)**

**Output in console window:**

**#Analysis**

**> #1**

**> leveneTest(asking.price~state,data=my\_sample)**

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

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

**group 2 4.2127 0.0154 \***

**447**

**---**

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

The p-value for levene test is significant. The null hypothesis says H0 is that all factor level variances are equal and alternate hypothesis Ha is at least one factor level variance is different.

The p-value from the levene test is significant, so that, we can reject the null hypothesis and accept the alternate hypothesis saying that ‘at least one factor level variance is different’.

**> boxplot(asking.price~state,**

**+ main="Price, three state Flavors",**

**+ col="red",**

**+ data=my\_sample)**

Chart, box and whisker chart

Description automatically generated

From the above box plot graph also we can say that, at least one factor level variance is different.

For Texas state, the 25 to 75 % is bigger than other states and this has more outliers and the spread is more.

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.

**RCode:**

**#2**

**vari=aggregate(asking.price~state,my\_sample,var)**

**vari**

**max(vari$asking.price)**

**min(vari$asking.price)**

**max(vari$asking.price)/min(vari$asking.price)**

**rm(vari)**

**#one way ANOVA**

**output=aov(asking.price~state,data=my\_sample)**

**summary(output)**

**names(output)**

**my\_mean=aggregate(asking.price~state,my\_sample,mean)**

**my\_mean**

**output$coefficients**

**my\_tukey=TukeyHSD(output)**

**my\_tukey**

**par(mar=c(5.1,8,4.1,2.1))**

**plot(my\_tukey,las=2)**

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

**Output in console window:**

**#2**

**> vari=aggregate(asking.price~state,my\_sample,var)**

**> vari**

**state asking.price**

**1 illinois 57561812**

**2 northcarolina 69297186**

**3 texas 112890353**

**> max(vari$asking.price)**

**[1] 112890353**

**> min(vari$asking.price)**

**[1] 57561812**

**> max(vari$asking.price)/min(vari$asking.price)**

**[1] 1.961202**

**> rm(vari)**

Texas state has more variance than compared to other states. By calculating the min and max, if max variance / min variance is greater than 1.5 then there is a problem.

The max by min value for variance is above 1.5, so that there might be some problem.

**> rm(vari)**

**> #one way ANOVA**

**> output=aov(asking.price~state,data=my\_sample)**

**> summary(output)**

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

**state 2 4.682e+08 234089428 2.929 0.0545 .**

**Residuals 447 3.572e+10 79916450**

**---**

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

**> names(output)**

**[1] "coefficients" "residuals" "effects" "rank"**

**[5] "fitted.values" "assign" "qr" "df.residual"**

**[9] "contrasts" "xlevels" "call" "terms"**

**[13] "model"**

The p-value for analysis of variance is not significant, the null hypothesis for analysis of variance is all factor level means are equal and for alternative hypothesis is at least one factor level mean is different.

As the p-value is not significant, we reject the null hypothesis and accept the alternate hyposthesis. So that, we can say at least one factor level mean is different.

**> my\_mean=aggregate(asking.price~state,my\_sample,mean)**

**> my\_mean**

**state asking.price**

**1 illinois 10836.78**

**2 northcarolina 11396.51**

**3 texas 13225.39**

**> output$coefficients**

**(Intercept) statenorthcarolina statetexas**

**10836.7800 559.7333 2388.6133**

These are the below means for each state and we can clearly see that there is some difference in the mean value of each state.

Texas state has highest mean and in the coefficients output, Illinois state is the intercept and 559.7333 is the difference between Illinois and NC state. Ans 2388.6133 is the difference between Illinois and Texas state.

**> my\_tukey=TukeyHSD(output)**

**> my\_tukey**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = asking.price ~ state, data = my\_sample)**

**$state**

**diff lwr upr p adj**

**northcarolina-illinois 559.7333 -1867.66423 2987.131 0.8504869**

**texas-illinois 2388.6133 -38.78423 4816.011 0.0548974**

**texas-northcarolina 1828.8800 -598.51756 4256.278 0.1803118**

**> par(mar=c(5.1,8,4.1,2.1))**

**> plot(my\_tukey,las=2)**

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

Chart, box and whisker chart

Description automatically generated

Looking at the p-values from the results of TukeyHSD and the above plot, we can interpret the mean asking price of northcarolina-illinois and texas-northcarolina comparisons are statistically not significant and there is a possible difference in asking price of cars sold in texas to illinois due to considerable p-value (0.05)

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.

**RCode:**

**#3**

**leveneTest(odometer~state,data=my\_sample)**

**boxplot(odometer~state,**

**main="Odometer, three state Flavors",**

**col="red",**

**data=my\_sample)**

**o\_vari=aggregate(odometer~state,my\_sample,var)**

**o\_vari**

**max(o\_vari$odometer)**

**min(o\_vari$odometer)**

**max(o\_vari$odometer)/min(o\_vari$odometer)**

**rm(o\_vari)**

**#one way ANOVA**

**output1=aov(odometer~state,data=my\_sample)**

**summary(output1)**

**names(output1)**

**my\_omean=aggregate(odometer~state,my\_sample,mean)**

**my\_omean**

**output1$coefficients**

**my\_otukey=TukeyHSD(output1)**

**my\_otukey**

**plot(my\_otukey)**

**par(mar=c(5.1,8,4.1,2.1))**

**plot(my\_otukey,las=2)**

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

**Output in console window:**

**#3**

**> leveneTest(odometer~state,data=my\_sample)**

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

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

**group 2 2.893 0.05645 .**

**447**

**---**

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

**> boxplot(odometer~state,**

**+ main="Odometer, three state Flavors",**

**+ col="red",**

**+ data=my\_sample)**

The p-value for levene test is significant. The null hypothesis says H0 is that all factor level variances are equal and alternate hypothesis Ha is at least one factor level variance is different.

The p-value from the levene test is significant, so that, we can reject the null hypothesis and accept the alternate hypothesis saying that ‘at least one factor level variance is different’.

Chart, box and whisker chart

Description automatically generated

From the above box plot graph also we can say that, at least one factor level variance is different.

For Texas state, the 25 to 75 % is bigger than other states and this has more outliers and the spread is more.

Texas state has more variance than compared to other states. By calculating the min and max, if max variance / min variance is greater than 1.5 then there is a problem.

The max by min value for variance is below 1.5, there is no problem in the variance.

**> #one way ANOVA**

**> output1=aov(odometer~state,data=my\_sample)**

**> summary(output1)**

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

**state 2 3.643e+09 1.821e+09 0.536 0.585**

**Residuals 447 1.518e+12 3.395e+09**

The p-value for analysis of variance is not significant, the null hypothesis for analysis of variance is all factor level means are equal and for alternative hypothesis is at least one factor level mean is different.

As the p-value is not significant, we fail to reject the null hypothesis. We don’t have enough evidence to say that at least one factor level mean is different.

**> my\_omean=aggregate(odometer~state,my\_sample,mean)**

**> my\_omean**

**state odometer**

**1 illinois 108707.4**

**2 northcarolina 115664.4**

**3 texas 112540.1**

**> output1$coefficients**

**(Intercept) statenorthcarolina statetexas**

**108707.35 6957.04 3832.76**

These are the below means for each state and we can clearly see that there is some difference in the mean value of each state.

NC state has highest mean and in the coefficients output, Illinois state is the intercept and 6957.04 is the difference between Illinois and NC state. Ans 38.32.76 is the difference between Illinois and Texas state.

**> my\_otukey=TukeyHSD(output1)**

**> my\_otukey**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = odometer ~ state, data = my\_sample)**

**$state**

**diff lwr upr p adj**

**northcarolina-illinois 6957.04 -8864.876 22778.96 0.5557908**

**texas-illinois 3832.76 -11989.156 19654.68 0.8363476**

**texas-northcarolina -3124.28 -18946.196 12697.64 0.8879928**

Chart, box and whisker chart

Description automatically generated

Looking at the p-values from the results of TukeyHSD and the above plot, we fail to reject null hypothesis and can interpret that there is no difference in mean odometer readings of cars sold in texas, illinois or north carolina.

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.

**RCode:**

**#4**

**texas\_output= aov(asking.price~region,data=my\_tex\_sample)**

**summary(texas\_output)**

**my\_tex\_tukey=TukeyHSD(texas\_output)**

**my\_tex\_tukey**

**par(mar=c(2.1,12,4.1,2.1))**

**plot(my\_tex\_tukey,las=1,cex.axis=0.6)**

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

**Output in console window:**

**> #4**

**> texas\_output= aov(asking.price~region,data=my\_tex\_sample)**

**> summary(texas\_output)**

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

**region 12 2.296e+09 191353682 1.805 0.053 .**

**Residuals 137 1.452e+10 106017652**

**---**

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

**> my\_tex\_tukey=TukeyHSD(texas\_output)**

**my\_tex\_tukey**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = asking.price ~ region, data = my\_tex\_sample)**

**$region**

**diff lwr upr p adj**

**austin, TX-amarillo, TX -637.1053 -19738.017 18463.807 1.0000000**

**brownsville, TX-amarillo, TX -7925.0000 -34443.893 18593.893 0.9984967**

**college station, TX-amarillo, TX -8522.8000 -31814.610 14769.010 0.9903532**

**corpus christi, TX-amarillo, TX 4478.2727 -15794.646 24751.191 0.9999313**

**dallas / fort worth-amarillo, TX -1720.8889 -20913.854 17472.076 1.0000000**

**el paso, TX-amarillo, TX 4912.0588 -14383.270 24207.387 0.9996964**

**galveston, TX-amarillo, TX -8330.0000 -47149.684 30489.684 0.9999497**

**houston, TX-amarillo, TX 4358.3810 -14583.685 23300.447 0.9998941**

**lubbock, TX-amarillo, TX -2175.7619 -21117.828 16766.304 1.0000000**

**odessa / midland-amarillo, TX 2116.3750 -19146.041 23378.791 1.0000000**

**tyler / east TX-amarillo, TX -2833.6923 -22686.401 17019.017 0.9999995**

**waco, TX-amarillo, TX -6203.8889 -27068.842 14661.064 0.9985686**

**brownsville, TX-austin, TX -7287.8947 -28858.926 14283.136 0.9952457**

**college station, TX-austin, TX -7885.6947 -25337.518 9566.128 0.9482520**

**corpus christi, TX-austin, TX 5115.3780 -8039.429 18270.185 0.9838606**

**dallas / fort worth-austin, TX -1083.7836 -12504.275 10336.708 1.0000000**

**el paso, TX-austin, TX 5549.1641 -6042.532 17140.860 0.9225363**

**galveston, TX-austin, TX -7692.8947 -43316.280 27930.490 0.9999461**

**houston, TX-austin, TX 4995.4862 -5998.129 15989.101 0.9461200**

**lubbock, TX-austin, TX -1538.6566 -12532.272 9454.959 0.9999996**

**odessa / midland-austin, TX 2753.4803 -11880.318 17387.279 0.9999880**

**tyler / east TX-austin, TX -2196.5870 -14694.097 10300.923 0.9999943**

**waco, TX-austin, TX -5566.7836 -19616.836 8483.269 0.9811773**

**college station, TX-brownsville, TX -597.8000 -25954.711 24759.111 1.0000000**

**corpus christi, TX-brownsville, TX 12403.2727 -10212.115 35018.660 0.8208233**

**dallas / fort worth-brownsville, TX 6204.1111 -15448.474 27856.696 0.9990027**

**el paso, TX-brownsville, TX 12837.0588 -8906.313 34580.431 0.7369310**

**galveston, TX-brownsville, TX -405.0000 -40497.797 39687.797 1.0000000**

**houston, TX-brownsville, TX 12283.3810 -9147.120 33713.882 0.7731257**

**lubbock, TX-brownsville, TX 5749.2381 -15681.263 27179.739 0.9994825**

**odessa / midland-brownsville, TX 10041.3750 -13465.111 33547.861 0.9659006**

**tyler / east TX-brownsville, TX 5091.3077 -17148.175 27330.790 0.9998997**

**waco, TX-brownsville, TX 1721.1111 -21426.476 24868.698 1.0000000**

**corpus christi, TX-college station, TX 13001.0727 -5726.247 31728.393 0.4917568**

**dallas / fort worth-college station, TX 6801.9111 -10750.616 24354.438 0.9843164**

**el paso, TX-college station, TX 13434.8588 -4229.540 31099.258 0.3419951**

**galveston, TX-college station, TX 192.8000 -37842.567 38228.167 1.0000000**

**houston, TX-college station, TX 12881.1810 -4396.642 30159.004 0.3739444**

**lubbock, TX-college station, TX 6347.0381 -10930.785 23624.861 0.9900210**

**odessa / midland-college station, TX 10639.1750 -9155.057 30433.407 0.8401333**

**tyler / east TX-college station, TX 5689.1077 -12582.492 23960.707 0.9977732**

**waco, TX-college station, TX 2318.9111 -17047.750 21685.572 0.9999999**

**dallas / fort worth-corpus christi, TX -6199.1616 -19487.277 7088.954 0.9351043**

**el paso, TX-corpus christi, TX 433.7861 -13001.756 13869.328 1.0000000**

**galveston, TX-corpus christi, TX -12808.2727 -49073.572 23457.027 0.9929228**

**houston, TX-corpus christi, TX -119.8918 -13042.970 12803.187 1.0000000**

**lubbock, TX-corpus christi, TX -6654.0346 -19577.113 6269.044 0.8761185**

**odessa / midland-corpus christi, TX -2361.8977 -18495.541 13771.746 0.9999993**

**tyler / east TX-corpus christi, TX -7311.9650 -21536.386 6912.456 0.8773718**

**waco, TX-corpus christi, TX -10682.1616 -26288.262 4923.938 0.5151707**

**el paso, TX-dallas / fort worth 6632.9477 -5109.815 18375.711 0.7897734**

**galveston, TX-dallas / fort worth -6609.1111 -42281.939 29063.716 0.9999899**

**houston, TX-dallas / fort worth 6079.2698 -5073.517 17232.056 0.8268600**

**lubbock, TX-dallas / fort worth -454.8730 -11607.660 10697.914 1.0000000**

**odessa / midland-dallas / fort worth 3837.2639 -10916.486 18591.014 0.9996220**

**tyler / east TX-dallas / fort worth -1112.8034 -13750.558 11524.951 1.0000000**

**waco, TX-dallas / fort worth -4483.0000 -18657.944 9691.944 0.9974193**

**galveston, TX-el paso, TX -13242.0588 -48970.065 22485.947 0.9892278**

**houston, TX-el paso, TX -553.6779 -11881.715 10774.360 1.0000000**

**lubbock, TX-el paso, TX -7087.8207 -18415.858 4240.217 0.6568552**

**odessa / midland-el paso, TX -2795.6838 -17682.353 12090.985 0.9999882**

**tyler / east TX-el paso, TX -7745.7511 -20538.429 5046.927 0.7035182**

**waco, TX-el paso, TX -11115.9477 -25429.187 3197.292 0.3094873**

**houston, TX-galveston, TX 12688.3810 -22850.085 48226.847 0.9922129**

**lubbock, TX-galveston, TX 6154.2381 -29384.228 41692.704 0.9999952**

**odessa / midland-galveston, TX 10446.3750 -26381.211 47273.961 0.9990970**

**tyler / east TX-galveston, TX 5496.3077 -30535.772 41528.387 0.9999989**

**waco, TX-galveston, TX 2126.1111 -34473.438 38725.660 1.0000000**

**lubbock, TX-houston, TX -6534.1429 -17249.394 4181.108 0.6937187**

**odessa / midland-houston, TX -2242.0060 -16667.854 12183.842 0.9999986**

**tyler / east TX-houston, TX -7192.0733 -19445.431 5061.285 0.7443430**

**waco, TX-houston, TX -10562.2698 -24395.599 3271.059 0.3358380**

**odessa / midland-lubbock, TX 4292.1369 -10133.711 18717.985 0.9985594**

**tyler / east TX-lubbock, TX -657.9304 -12911.288 11595.428 1.0000000**

**waco, TX-lubbock, TX -4028.1270 -17861.456 9805.202 0.9988318**

**tyler / east TX-odessa / midland -4950.0673 -20552.415 10652.281 0.9973414**

**waco, TX-odessa / midland -8320.2639 -25191.822 8551.294 0.9056768**

**waco, TX-tyler / east TX -3370.1966 -18426.397 11686.004 0.9999209**

**> par(mar=c(2.1,12,4.1,2.1))**

**> plot(my\_tex\_tukey,las=1,cex.axis=0.6)**

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

Chart

Description automatically generated

From the TukeyHSD test all the p-values are not significant. All the p-values are close to 1.

From the graph, Galveston city in Texas state has very least mean value compared to other cities. Because of this, the lower values of the TukeyHSD test involving Galveston city is very far.

The plot, we fail to reject null hypothesis and can interpret that there is no difference in mean asking prices of cars sold in Texas region.

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.

**RCode:**

**#5**

**output2=aov(asking.price~fuel+condition,data=my\_sample)**

**summary(output2)**

**my\_tot\_tukey=TukeyHSD(output2)**

**my\_tot\_tukey**

**par(mfrow=c(1,2))**

**par(mar=c(5.1,8,4.1,2.1))**

**plot(my\_tot\_tukey,las=1.5,cex.axis=.8)**

**par(mfrow=c(1,1))**

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

**Output in console window:**

**> #5**

**> output2=aov(asking.price~fuel+condition,data=my\_sample)**

**> summary(output2)**

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

**fuel 1 3.124e+09 3.124e+09 47.85 1.62e-11 \*\*\***

**condition 5 4.141e+09 8.281e+08 12.68 1.57e-11 \*\*\***

**Residuals 443 2.893e+10 6.530e+07**

**---**

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

The p-value for analysis of variance is highly significant, the null hypothesis for analysis of variance is all factor level means are equal and for alternative hypothesis is at least one factor level mean is different.

As the p-value is significant, we reject the null hypothesis and accept the alternate hypothesis. So we can say that at least one factor level mean is different.

**> my\_tot\_tukey=TukeyHSD(output2)**

**> my\_tot\_tukey**

**Tukey multiple comparisons of means**

**95% family-wise confidence level**

**Fit: aov(formula = asking.price ~ fuel + condition, data = my\_sample)**

**$fuel**

**diff lwr upr p adj**

**gas-diesel -12785.95 -16418.69 -9153.212 0**

**$condition**

**diff lwr upr p adj**

**fair-excellent -9590.1650 -16171.492 -3008.838 0.0005201**

**good-excellent -5633.5613 -8109.063 -3158.059 0.0000000**

**like new-excellent 1027.7867 -2529.522 4585.095 0.9624517**

**new-excellent 1873.2965 -11560.781 15307.374 0.9986916**

**salvage-excellent -12266.7035 -35441.947 10908.540 0.6547942**

**good-fair 3956.6037 -2759.838 10673.045 0.5417078**

**like new-fair 10617.9517 3432.068 17803.835 0.0004074**

**new-fair 11463.4615 -3350.617 26277.540 0.2329202**

**salvage-fair -2676.5385 -26678.088 21325.011 0.9995581**

**like new-good 6661.3480 2859.879 10462.817 0.0000113**

**new-good 7506.8578 -5993.926 21007.641 0.6046234**

**salvage-good -6633.1422 -29847.117 16580.833 0.9641897**

**new-like new 845.5098 -12894.850 14585.870 0.9999765**

**salvage-like new -13294.4902 -36648.611 10059.631 0.5794931**

**salvage-new -14140.0000 -40846.459 12566.459 0.6545015**

**> par(mfrow=c(1,2))**

**> par(mar=c(5.1,8,4.1,2.1))**

**> plot(my\_tot\_tukey,las=1.5,cex.axis=.8)**

**> par(mfrow=c(1,1))**

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

Condition with fair and good, have high significance p-values, so that we can reject the null hypothesis and accept the alternative hypothesis saying that the mean of fair and good conditions are not equal.

Chart, box and whisker chart

Description automatically generated

In the above graph, we can interpret that salvage has lower value of mean. Records with salvage has very least value of lower state. There is more spread in the negative region.

From the results of Tukey HSD we can say that there is a difference in means of asking process across gas-diesel fuel as it’s p-value is significant.

There is significant p-values of difference in means in levels of conditions for some variables, hence we can reject null hypothesis for above and conclude that there is a difference in the mean level of asking price for those conditions.