### “Umakanth Sai Balguri - (U01403700)”

# Preprocessing

### Loading Library files

rm(list=ls())  
library(rio)  
library(car)

## Loading required package: carData

library(moments)  
library(robustHD)

## Loading required package: ggplot2

## Loading required package: perry

## Loading required package: parallel

## Loading required package: robustbase

### Loading the file into R and

cars.main=import("6304 Module 9 Assignment Data.xlsx")  
colnames(cars.main)=tolower(make.names(colnames(cars.main)))

### Creating subsets for each state and taking sample of 150 using U-number. Adding state column to dataframe U-number as seed

cars.sub.texas = subset(cars.main, cylinders %in% c(4,6,8) & fuel %in% c("gas","diesel") & region %in% c("amarillo, TX", "austin, TX", "brownsville, TX", "college station, TX", "corpus christi, TX", "dallas / fort worth", "el paso, TX", "galveston, TX", "houston, TX", "lubbock, TX", "odessa / midland", "tyler / east TX","waco, TX") )  
set.seed(01403700)  
pcars.t=cars.sub.texas[sample(1:nrow(cars.sub.texas),150,replace=FALSE),]  
pcars.t$state="Texas"  
  
cars.sub.illinois = subset(cars.main, cylinders %in% c(4,6,8) & fuel %in% c("gas","diesel") & region %in% c("champaign urbana", "chicago", "danville", "peoria, IL", "quad cities, IA/IL", "rockford, IL","southern illinois", "springfield, IL") )  
set.seed(01403700)  
pcars.i=cars.sub.illinois[sample(1:nrow(cars.sub.illinois),150,replace=FALSE),]  
pcars.i$state="Illinois"  
  
cars.sub.northcarolina = subset(cars.main, cylinders %in% c(4,6,8) & fuel %in% c("gas","diesel") & region %in% c("asheville, NC", "boone, NC", "charlotte, NC", "eastern NC", "fayetteville, NC","greensboro, NC", "wilmington, NC", "winston-salem, NC") )  
set.seed(01403700)  
pcars.nc=cars.sub.northcarolina[sample(1:nrow(cars.sub.northcarolina),150,replace=FALSE),]  
pcars.nc$state="North Carolina"  
  
pcars = rbind(pcars.t,pcars.i,pcars.nc)

### Creating one primary data set by combining above samples

pcars = rbind(pcars.t,pcars.i,pcars.nc)  
colnames(pcars)=tolower(make.names(colnames(pcars)))  
pcars$state=as.factor(pcars$state)

# Analysis

### Response to Q1

leveneTest(asking.price~state,data=pcars)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)   
## group 2 4.9903 0.007187 \*\*  
## 447   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

As p-value is less than 0.05 (P=0.007187), we can reject the null hypothesis. It means that there is a significant difference between means of asking price between the states.

### Response to Q2

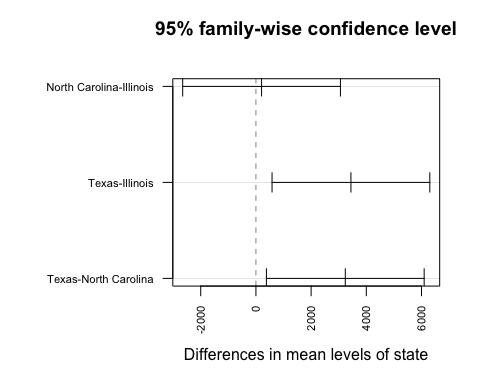
pcars.out=aov(asking.price~state,data=pcars)  
summary(pcars.out)

## Df Sum Sq Mean Sq F value Pr(>F)   
## state 2 1.119e+09 559480732 5.05 0.00678 \*\*  
## Residuals 447 4.952e+10 110788015   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pcars.graph=TukeyHSD((pcars.out))  
pcars.graph

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = asking.price ~ state, data = pcars)  
##   
## $state  
## diff lwr upr p adj  
## North Carolina-Illinois 204.3533 -2653.6926 3062.399 0.9845360  
## Texas-Illinois 3442.5800 584.5340 6300.626 0.0133659  
## Texas-North Carolina 3238.2267 380.1807 6096.273 0.0217518

par(mar=c(5.1,9,4.1,2.1))  
plot(pcars.graph,las=2,cex.axis=.7)



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

From the analysis, P value is 0.00678 (p<0.05), meaning we can reject the null hypothesis. i.e there is a significant difference between means of asking price between the states.  
This can be clearly seen in the plots. However, from the plot data we can only determine that there is a difference in Texas prices when compared to other two states. (Texas has higher asking price than compared to other two states).

Also, we cannot determine the difference between NC and Illinois as the 95% confidence interval runs from positive side to negative side (approx -2500 to +2500 i.e. the difference between NC to Illinois can be either positve or negative, hence its undeterministic from the sampled data)

### Response to Q3

leveneTest(odometer~state,data=pcars)

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 2 0.7327 0.4812  
## 447

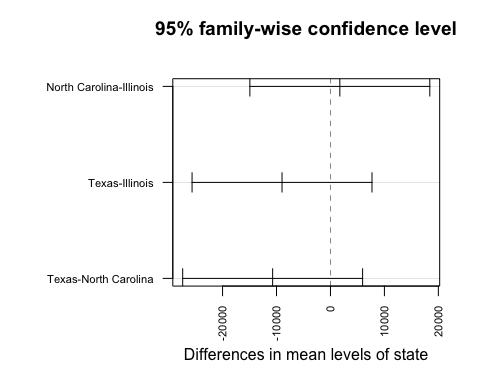
pcars.out2=aov(odometer~state,data=pcars)  
summary(pcars.out2)

## Df Sum Sq Mean Sq F value Pr(>F)  
## state 2 9.923e+09 4.961e+09 1.316 0.269  
## Residuals 447 1.686e+12 3.771e+09

pcars.graph2=TukeyHSD((pcars.out2))  
pcars.graph2

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = odometer ~ state, data = pcars)  
##   
## $state  
## diff lwr upr p adj  
## North Carolina-Illinois 1737.913 -14937.25 18413.080 0.9674331  
## Texas-Illinois -8978.073 -25653.24 7697.093 0.4151512  
## Texas-North Carolina -10715.987 -27391.15 5959.180 0.2866046

par(mar=c(5.1,9,4.1,2.1))  
plot(pcars.graph2,las=2,cex.axis=.7)



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

From the levene test, P value is 0.4812 (p>0.05) meaning we fail to reject the null hypothesis. There is no statistical difference in mean of odometer data between the states. From ANOVA analysis, P value is 0.269 (p>0.05), meaning we fail to reject the null hypothesis. There is no statistical difference in mean of odometer data between the states. This can also be seen in plot as the confidence intervals run thought positive and negative. So none of the state pairs show significant difference to note.

### Response to Q4

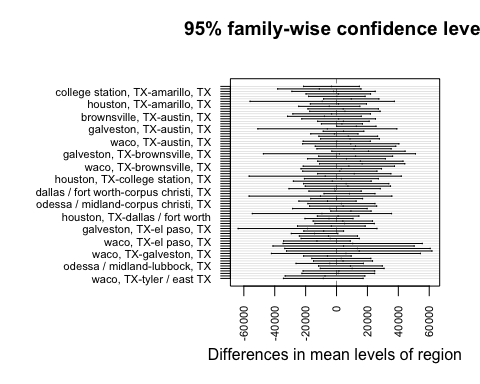
pcars.t$region=as.factor(pcars.t$region)  
pcars.out3=aov(asking.price~region,data=pcars.t)  
summary(pcars.out3)

## Df Sum Sq Mean Sq F value Pr(>F)   
## region 12 3.712e+09 309346679 1.814 0.0514 .  
## Residuals 137 2.336e+10 170505115   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pcars.graph3=TukeyHSD((pcars.out3))  
pcars.graph3

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = asking.price ~ region, data = pcars.t)  
##   
## $region  
## diff lwr upr  
## austin, TX-amarillo, TX -3298.35870 -21372.1197 14775.4024  
## brownsville, TX-amarillo, TX -11142.25000 -38106.7415 15822.2415  
## college station, TX-amarillo, TX -1933.50000 -28897.9915 25030.9915  
## corpus christi, TX-amarillo, TX 1142.75000 -19743.8553 22029.3553  
## dallas / fort worth-amarillo, TX 194.77381 -18099.7437 18489.2913  
## el paso, TX-amarillo, TX 9521.45000 -8364.7702 27407.6702  
## galveston, TX-amarillo, TX -9254.75000 -55958.6193 37449.1193  
## houston, TX-amarillo, TX 1217.38636 -16962.0701 19396.8428  
## lubbock, TX-amarillo, TX -4697.90385 -24484.4311 15088.6235  
## odessa / midland-amarillo, TX 4285.67857 -18503.4761 27074.8333  
## tyler / east TX-amarillo, TX 5422.25000 -17366.9047 28211.4047  
## waco, TX-amarillo, TX -3167.75000 -28270.3255 21934.8255  
## brownsville, TX-austin, TX -7843.89130 -31698.0755 16010.2929  
## college station, TX-austin, TX 1364.85870 -22489.3255 25219.0429  
## corpus christi, TX-austin, TX 4441.10870 -12237.8776 21120.0950  
## dallas / fort worth-austin, TX 3493.13251 -9797.0019 16783.2669  
## el paso, TX-austin, TX 12819.80870 97.5771 25542.0403  
## galveston, TX-austin, TX -5956.39130 -50936.2725 39023.4899  
## houston, TX-austin, TX 4515.74506 -8615.5512 17647.0413  
## lubbock, TX-austin, TX -1399.54515 -16678.4511 13879.3608  
## odessa / midland-austin, TX 7584.03727 -11423.4317 26591.5062  
## tyler / east TX-austin, TX 8720.60870 -10286.8603 27728.0777  
## waco, TX-austin, TX 130.60870 -21596.7381 21857.9555  
## college station, TX-brownsville, TX 9208.75000 -21927.1629 40344.6629  
## corpus christi, TX-brownsville, TX 12285.00000 -13765.1737 38335.1737  
## dallas / fort worth-brownsville, TX 11337.02381 -12684.8545 35358.9021  
## el paso, TX-brownsville, TX 20663.70000 -3048.7048 44376.1048  
## galveston, TX-brownsville, TX 1887.50000 -47342.7008 51117.7008  
## houston, TX-brownsville, TX 12359.63636 -11574.7302 36294.0030  
## lubbock, TX-brownsville, TX 6444.34615 -18732.3877 31621.0801  
## odessa / midland-brownsville, TX 15427.92857 -12171.1084 43026.9655  
## tyler / east TX-brownsville, TX 16564.50000 -11034.5369 44163.5369  
## waco, TX-brownsville, TX 7974.50000 -21563.6205 37512.6205  
## corpus christi, TX-college station, TX 3076.25000 -22973.9237 29126.4237  
## dallas / fort worth-college station, TX 2128.27381 -21893.6045 26150.1521  
## el paso, TX-college station, TX 11454.95000 -12257.4548 35167.3548  
## galveston, TX-college station, TX -7321.25000 -56551.4508 41908.9508  
## houston, TX-college station, TX 3150.88636 -20783.4802 27085.2530  
## lubbock, TX-college station, TX -2764.40385 -27941.1377 22412.3301  
## odessa / midland-college station, TX 6219.17857 -21379.8584 33818.2155  
## tyler / east TX-college station, TX 7355.75000 -20243.2869 34954.7869  
## waco, TX-college station, TX -1234.25000 -30772.3705 28303.8705  
## dallas / fort worth-corpus christi, TX -947.97619 -17865.9287 15969.9763  
## el paso, TX-corpus christi, TX 8378.70000 -8096.8765 24854.2765  
## galveston, TX-corpus christi, TX -10397.50000 -56579.5220 35784.5220  
## houston, TX-corpus christi, TX 74.63636 -16718.8262 16868.0989  
## lubbock, TX-corpus christi, TX -5840.65385 -24361.8420 12680.5343  
## odessa / midland-corpus christi, TX 3142.92857 -18556.6871 24842.5442  
## tyler / east TX-corpus christi, TX 4279.50000 -17420.1157 25979.1157  
## waco, TX-corpus christi, TX -4310.50000 -28428.2744 19807.2744  
## el paso, TX-dallas / fort worth 9326.67619 -3707.2682 22360.6206  
## galveston, TX-dallas / fort worth -9449.52381 -54518.5624 35619.5148  
## houston, TX-dallas / fort worth 1022.61255 -12410.9079 14456.1330  
## lubbock, TX-dallas / fort worth -4892.67766 -20432.0952 10646.7399  
## odessa / midland-dallas / fort worth 4090.90476 -15126.5979 23308.4074  
## tyler / east TX-dallas / fort worth 5227.47619 -13990.0264 24444.9788  
## waco, TX-dallas / fort worth -3362.52381 -25273.8480 18548.8004  
## galveston, TX-el paso, TX -18776.20000 -63681.0521 26128.6521  
## houston, TX-el paso, TX -8304.06364 -21176.0089 4567.8817  
## lubbock, TX-el paso, TX -14219.35385 -29275.9471 837.2394  
## odessa / midland-el paso, TX -5235.77143 -24065.0017 13593.4588  
## tyler / east TX-el paso, TX -4099.20000 -22928.4302 14730.0302  
## waco, TX-el paso, TX -12689.20000 -34260.7932 8882.3932  
## houston, TX-galveston, TX 10472.13636 -34550.3192 55494.5920  
## lubbock, TX-galveston, TX 4556.84615 -41138.1813 50251.8736  
## odessa / midland-galveston, TX 13540.42857 -33532.6470 60613.5041  
## tyler / east TX-galveston, TX 14677.00000 -32396.0756 61750.0756  
## waco, TX-galveston, TX 6087.00000 -42148.5488 54322.5488  
## lubbock, TX-houston, TX -5915.29021 -21319.0808 9488.5004  
## odessa / midland-houston, TX 3068.29221 -16039.7081 22176.2925  
## tyler / east TX-houston, TX 4204.86364 -14903.1367 23312.8639  
## waco, TX-houston, TX -4385.13636 -26200.4841 17430.2114  
## odessa / midland-lubbock, TX 8983.58242 -11659.3268 29626.4916  
## tyler / east TX-lubbock, TX 10120.15385 -10522.7554 30763.0631  
## waco, TX-lubbock, TX 1530.15385 -21641.4543 24701.7620  
## tyler / east TX-odessa / midland 1136.57143 -22399.9664 24673.1092  
## waco, TX-odessa / midland -7453.42857 -33236.4139 18329.5568  
## waco, TX-tyler / east TX -8590.00000 -34372.9853 17192.9853  
## p adj  
## austin, TX-amarillo, TX 0.9999914  
## brownsville, TX-amarillo, TX 0.9736412  
## college station, TX-amarillo, TX 1.0000000  
## corpus christi, TX-amarillo, TX 1.0000000  
## dallas / fort worth-amarillo, TX 1.0000000  
## el paso, TX-amarillo, TX 0.8488055  
## galveston, TX-amarillo, TX 0.9999788  
## houston, TX-amarillo, TX 1.0000000  
## lubbock, TX-amarillo, TX 0.9998527  
## odessa / midland-amarillo, TX 0.9999880  
## tyler / east TX-amarillo, TX 0.9998494  
## waco, TX-amarillo, TX 0.9999999  
## brownsville, TX-austin, TX 0.9962899  
## college station, TX-austin, TX 1.0000000  
## corpus christi, TX-austin, TX 0.9995202  
## dallas / fort worth-austin, TX 0.9995793  
## el paso, TX-austin, TX 0.0463638  
## galveston, TX-austin, TX 0.9999998  
## houston, TX-austin, TX 0.9944205  
## lubbock, TX-austin, TX 1.0000000  
## odessa / midland-austin, TX 0.9800716  
## tyler / east TX-austin, TX 0.9422863  
## waco, TX-austin, TX 1.0000000  
## college station, TX-brownsville, TX 0.9986404  
## corpus christi, TX-brownsville, TX 0.9300643  
## dallas / fort worth-brownsville, TX 0.9297009  
## el paso, TX-brownsville, TX 0.1568909  
## galveston, TX-brownsville, TX 1.0000000  
## houston, TX-brownsville, TX 0.8738919  
## lubbock, TX-brownsville, TX 0.9996790  
## odessa / midland-brownsville, TX 0.8011266  
## tyler / east TX-brownsville, TX 0.7154090  
## waco, TX-brownsville, TX 0.9994485  
## corpus christi, TX-college station, TX 0.9999999  
## dallas / fort worth-college station, TX 1.0000000  
## el paso, TX-college station, TX 0.9176744  
## galveston, TX-college station, TX 0.9999991  
## houston, TX-college station, TX 0.9999998  
## lubbock, TX-college station, TX 1.0000000  
## odessa / midland-college station, TX 0.9999151  
## tyler / east TX-college station, TX 0.9995156  
## waco, TX-college station, TX 1.0000000  
## dallas / fort worth-corpus christi, TX 1.0000000  
## el paso, TX-corpus christi, TX 0.8852604  
## galveston, TX-corpus christi, TX 0.9999159  
## houston, TX-corpus christi, TX 1.0000000  
## lubbock, TX-corpus christi, TX 0.9974886  
## odessa / midland-corpus christi, TX 0.9999994  
## tyler / east TX-corpus christi, TX 0.9999798  
## waco, TX-corpus christi, TX 0.9999932  
## el paso, TX-dallas / fort worth 0.4414020  
## galveston, TX-dallas / fort worth 0.9999608  
## houston, TX-dallas / fort worth 1.0000000  
## lubbock, TX-dallas / fort worth 0.9975254  
## odessa / midland-dallas / fort worth 0.9999538  
## tyler / east TX-dallas / fort worth 0.9994051  
## waco, TX-dallas / fort worth 0.9999988  
## galveston, TX-el paso, TX 0.9710858  
## houston, TX-el paso, TX 0.6105063  
## lubbock, TX-el paso, TX 0.0844124  
## odessa / midland-el paso, TX 0.9992588  
## tyler / east TX-el paso, TX 0.9999412  
## waco, TX-el paso, TX 0.7415763  
## houston, TX-galveston, TX 0.9998813  
## lubbock, TX-galveston, TX 1.0000000  
## odessa / midland-galveston, TX 0.9989638  
## tyler / east TX-galveston, TX 0.9977439  
## waco, TX-galveston, TX 0.9999999  
## lubbock, TX-houston, TX 0.9854579  
## odessa / midland-houston, TX 0.9999979  
## tyler / east TX-houston, TX 0.9999341  
## waco, TX-houston, TX 0.9999752  
## odessa / midland-lubbock, TX 0.9607510  
## tyler / east TX-lubbock, TX 0.9092457  
## waco, TX-lubbock, TX 1.0000000  
## tyler / east TX-odessa / midland 1.0000000  
## waco, TX-odessa / midland 0.9989119  
## waco, TX-tyler / east TX 0.9958149

par(mar=c(5.1,12,4.1,2.1))  
plot(pcars.graph3,las=2,cex.axis=.7)



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

From ANOVA analysis, for none of the pair, P value is <0.05, meaning we fail to reject the null hypothesis. There is no statistical difference in mean of asking price data between the various regions of texas. This can also be seen in plot as the confidence intervals run thought positive and negative. So none of the region pairs show significant difference to note.

### Response to Q5

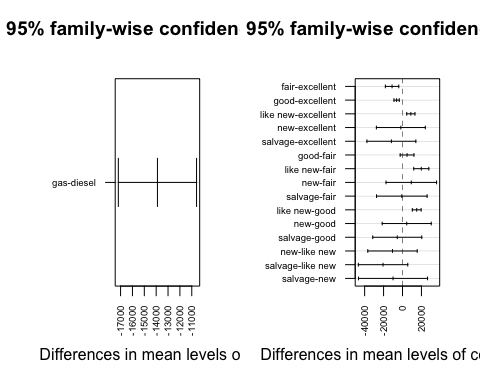
pcars$fuel=as.factor(pcars$fuel)  
pcars$condition=as.factor(pcars$condition)  
pcars.out4=aov(asking.price~fuel+condition,data=pcars)  
summary(pcars.out4)

## Df Sum Sq Mean Sq F value Pr(>F)   
## fuel 1 5.571e+09 5.571e+09 68.61 1.43e-15 \*\*\*  
## condition 5 9.098e+09 1.820e+09 22.41 < 2e-16 \*\*\*  
## Residuals 443 3.597e+10 8.120e+07   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

pcars.graph4=TukeyHSD((pcars.out4))  
pcars.graph4

## Tukey multiple comparisons of means  
## 95% family-wise confidence level  
##   
## Fit: aov(formula = asking.price ~ fuel + condition, data = pcars)  
##   
## $fuel  
## diff lwr upr p adj  
## gas-diesel -13893.06 -17189.41 -10596.72 0  
##   
## $condition  
## diff lwr upr p adj  
## fair-excellent -11029.3460 -18099.623 -3959.069 0.0001475  
## good-excellent -6217.3938 -9016.972 -3417.815 0.0000000  
## like new-excellent 8813.1383 4489.034 13137.243 0.0000002  
## new-excellent -1752.2032 -27592.001 24087.595 0.9999619  
## salvage-excellent -11747.2032 -37587.001 14092.595 0.7843502  
## good-fair 4811.9522 -2459.950 12083.855 0.4071747  
## like new-fair 19842.4843 11858.700 27826.268 0.0000000  
## new-fair 9277.1429 -17420.012 35974.298 0.9195676  
## salvage-fair -717.8571 -27415.012 25979.298 0.9999996  
## like new-good 15030.5321 10384.068 19676.996 0.0000000  
## new-good 4465.1907 -21430.502 30360.884 0.9963896  
## salvage-good -5529.8093 -31425.502 20365.884 0.9902074  
## new-like new -10565.3415 -36669.884 15539.201 0.8562192  
## salvage-like new -20560.3415 -46664.884 5544.201 0.2153567  
## salvage-new -9995.0000 -46470.258 26480.258 0.9701062

par(mfrow=c(1,2))  
par(mar=c(5.1,6,4.1,2.1))  
plot(pcars.graph4,las=2,cex.axis=.6)



par(mar=c(5.1,4.1,4.1,2.1))  
par(mfrow=c(1,1))

From the plot 1, ie the fuel, it can be identified that the mean of diesel is higher than the mean of gas. From plot-2, fair-excellent, good-excellent, like-new-excellent, like-new-fair, like-new-good have a difference when compared to all other pairs.