Module Assignment
Module 9
QMB-6304 Foundations of Business Statistics

Preprocessing

1. Load the file "6304 Module 9 Assignment Data.xlsx" into R. This data set includes several variables on 437 counties in six midwestern states. This will be your master data set
2. Create a single data frame for your analysis. This will be your primary data set. It will meet the following characteristics:
    a. Includes state, popdensity, density.category, percollege, and inmetro variables from the master (N=437) data set.
    b. Be a random sample of n=250, with each state making up 20% of the total sample. For example, 20% of 250 observations will be from the state of Illinois, 20% from Indiana, etc. Use your U number as the random number seed.
    c. The variables state, density.category, and inmetro will be factors.

```
#Carolina Aldana Yabur
#U25124553

#Preprocessing
rm(list=ls())
setwd("C:/Users/calda/Desktop")
library(rio)
library(moments)
library(car)
counties_data= import("6304 Module 9 Assignment Data.xlsx")
colnames(counties_data)=tolower(make.names(colnames(counties_dat
a)))

subset1= subset(counties_data)
subset1$state=as.factor(subset1$state)
is.factor(subset1$state)
subset1$density.category=as.factor(subset1$density.category)
is.factor(subset1$density.category)
subset1$inmetro=as.factor(subset1$inmetro)
is.factor(subset1$inmetro)

subset_state=subset(subset1=="state")
subset_popdensity=subset(subset1=="popdensity")
subset_density.category=subset(subset1=="density.category")
```

```
subset_percollege=subset(subset1=="percollege")
subset_inmetro=subset(subset1=="inmetro")

sample.state=subset_state[sample(1:nrow(subset_state),20,replace
=FALSE),]
sample.popdensity=subset_popdensity[sample(1:nrow(subset_popdens
ity),20,replace=FALSE),]
sample.density.category=subset_density.category[sample(1:nrow(su
bset_density.category),20,replace=FALSE),]
sample.percollege=subset_percollege[sample(1:nrow(subset_percoll
ege),20,replace=FALSE),]
sample.inmetro=subset_inmetro[sample(1:nrow(subset_inmetro),20,r
eplace=FALSE),]

mysample=rbind(sample.state,sample.popdensity,sample.density.cat
egory, sample.percollege, sample.inmetro)
mysample=subset1[sample(1:nrow(subset1),250),]
```

Analysis

Using your primary data set:

1. Show the results of an str() command.

```
#Analysis

#Part 1
> str(mysample)
'data.frame':  250 obs. of  20 variables:
 $ county              : chr  "WOOD" "WAYNE" "GALLIA" "BERRIEN"
...
 $ state               : Factor w/ 5 levels "IL","IN","MI",..: 4
1 4 3 2 1 2 5 1 3 ...
 $ area                : num  617 714 469 571 396 ...
 $ poptotal            : num  113269 17241 30954 161378 797159
...
 $ popdensity          : num  183.5 24.2 66 282.6 2011.8 ...
 $ density.category    : Factor w/ 5 levels "100 to 249","250 to
749",..: 1 5 3 2 4 3 1 3 2 5 ...
 $ popwhite            : num  109303 17141 29831 133259 615039
...
 $ popblack            : num  1168 9 871 24872 169654 ...
 $ popamerindian       : num  197 31 79 685 1698 ...
 $ popasian            : num  1028 44 136 1487 7579 ...
 $ popother            : num  1573 16 37 1075 3189 ...
 $ popadults           : num  64052 11613 19586 102485 511309
...
 $ popchild            : num  49217 5628 11368 58893 285850 ...
 $ percollege          : num  29.1 15.7 14.9 23.7 26.7 ...
 $ percprof            : num  8.34 2.76 4.32 6.3 7.69 ...
```

```
 $ percbelowpoverty    : num  10.6 14.4 22.5 14.7 12.1 ...
 $ percchildbelowpovert: num  8.76 19.69 28.46 22.98 18.16 ...
 $ percadultpoverty    : num  12.22 11.62 19.78 11.85 9.76 ...
 $ percelderlypoverty  : num  6.93 14.78 21.77 11 10.73 ...
 $ inmetro             : Factor w/ 2 levels "0","1": 2 1 1 2 2 1
2 1 2 1 ...
```

2. Show the results of the table() command on the state variable.

```
> #Part 2
> table(mysample$state)

IL IN MI OH WI
71 49 46 45 39
```

3. Determine if percollege has an equal variance across all five states. Briefly interpret your results. If you determine there is a difference in variances across the states, discuss where is/are the differences.

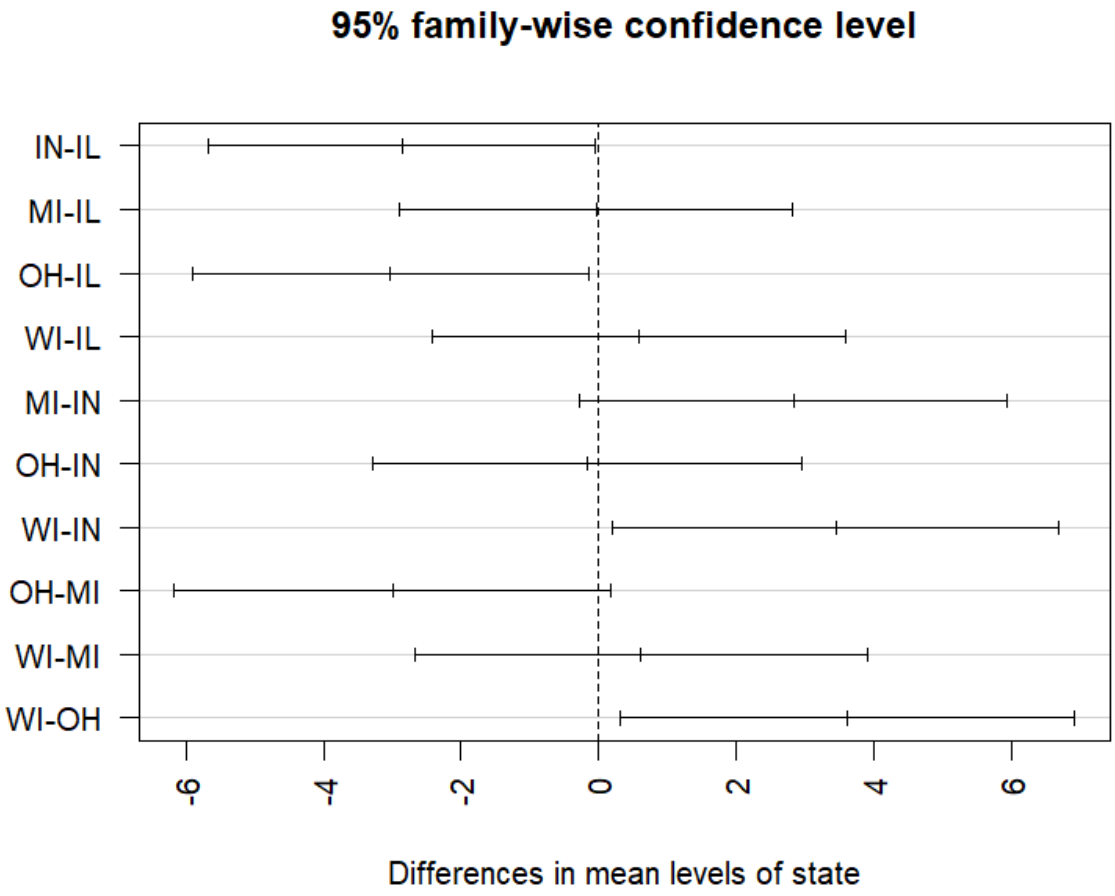
```
> #Part 3
> leveneTest(percollege~state,data = mysample)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value Pr(>F)
group   4  0.7676 0.5472
      245
```

Here the p-value is 0.5472, which is greater than the standard significance level 0.05. Therefore, we don't have enough evidence to reject the null hypothesis that the variance of percollege is equal across all five states.

4. Conduct a one-way analysis of variance with percollege as the dependent variable and state as the independent variable. Plot the results of a Tukey HSD test. Briefly explain the results shown in the plot, stating between which pairs of states do/do not show significant population mean differences in percollege. Make sure factor level names can be clearly and completely read on the appropriate axis of your plot.

```
> #Part 4
> one.way=aov(percollege~state,data = mysample)
> summary(one.way)
             Df Sum Sq Mean Sq F value  Pr(>F)
state         4    567  141.78   4.697 0.00114 **
Residuals   245   7396   30.19
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
> TukeyHSD(one.way)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percollege ~ state, data = mysample)

$state
             diff        lwr         upr     p adj
IN-IL -2.86384159 -5.6681404 -0.05954281 0.0426846
MI-IL -0.03343858 -2.8913299  2.82445278 0.9999998
OH-IL -3.03230209 -5.9093987 -0.15520553 0.0332131
WI-IL  0.57870316 -2.4308065  3.58821281 0.9843735
MI-IN  2.83040301 -0.2694830  5.93028900 0.0918720
OH-IN -0.16846050 -3.2860613  2.94914030 0.9998901
WI-IN  3.44254475  0.2023444  6.68274512 0.0310724
OH-MI -2.99886351 -6.1647577  0.16703069 0.0728847
WI-MI  0.61214174 -2.6745511  3.89883457 0.9861365
WI-OH  3.61100525  0.3075992  6.91441128 0.0243189

> Hs.out=TukeyHSD(one.way)
> Hs.out
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percollege ~ state, data = mysample)

$state
             diff        lwr         upr     p adj
IN-IL -2.86384159 -5.6681404 -0.05954281 0.0426846
MI-IL -0.03343858 -2.8913299  2.82445278 0.9999998
OH-IL -3.03230209 -5.9093987 -0.15520553 0.0332131
WI-IL  0.57870316 -2.4308065  3.58821281 0.9843735
MI-IN  2.83040301 -0.2694830  5.93028900 0.0918720
OH-IN -0.16846050 -3.2860613  2.94914030 0.9998901
WI-IN  3.44254475  0.2023444  6.68274512 0.0310724
OH-MI -2.99886351 -6.1647577  0.16703069 0.0728847
WI-MI  0.61214174 -2.6745511  3.89883457 0.9861365
WI-OH  3.61100525  0.3075992  6.91441128 0.0243189
```




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

plot(Hs.out,las=2)

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

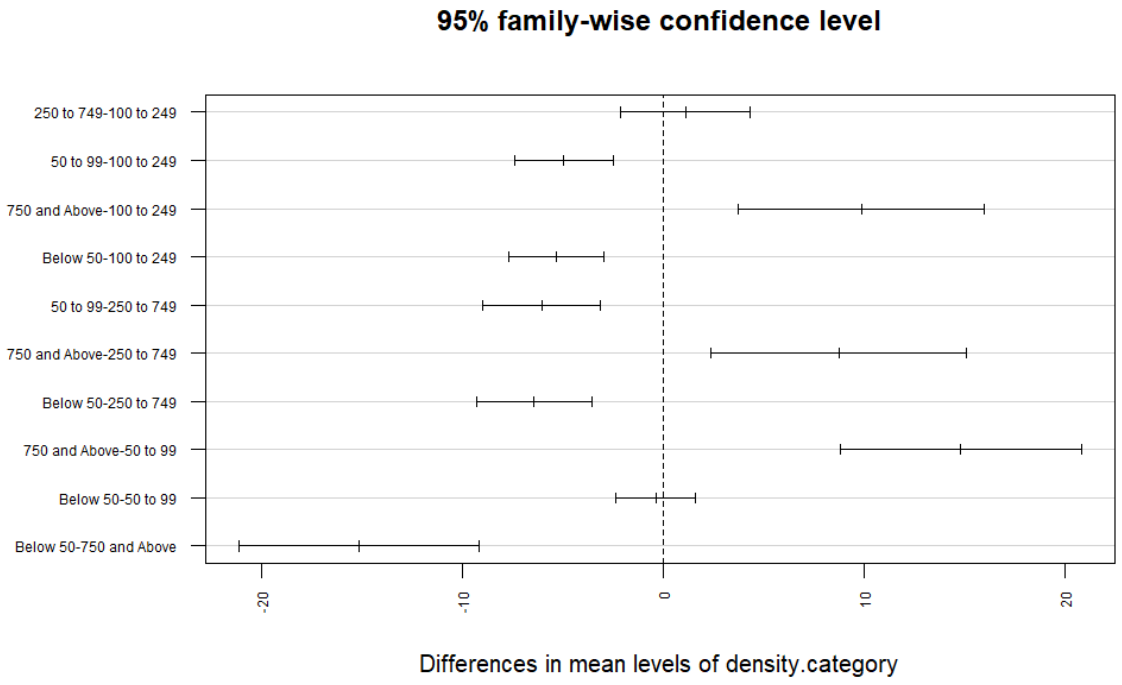
From the plot we can see the confidence intervals that cross the vertical zero line, and we can conclude that there is no statistically significant difference in means between MI-IL, WI-IL, MI-IN, OH-IN, OH-MI, and WI-MI at the 95% confidence level. Two groups (IN-IL, OH-IL) on the left have significantly lower mean than the two groups (WI-IN, WI-OH) on the right side.

5. Repeat Steps 3 and 4 above using percollege as the dependent variable and density.category as the independent variable. Again, briefly explain your analysis results and make sure category names can be clearly and completely read on the appropriate axis of your plot.

```
> #Part 5
> one.way=aov(percollege~density.category,data = mysample)
```

```
> summary(one.way)
                  Df Sum Sq Mean Sq F value Pr(>F)
density.category   4   2465   616.2   27.46 <2e-16 ***
Residuals        245   5498    22.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(one.way)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percollege ~ density.category, data =
mysample)

$density.category
                                  diff        lwr       upr     p
adj
250 to 749-100 to 249      1.1061184  -2.114261  4.326498
0.8793937
50 to 99-100 to 249       -4.9623349  -7.411306 -2.513364
0.0000007
750 and Above-100 to 249   9.8464195   3.702328 15.990511
0.0001536
Below 50-100 to 249       -5.3376196  -7.707754 -2.967485
0.0000000
50 to 99-250 to 749       -6.0684533  -9.011963 -3.124944
0.0000004
750 and Above-250 to 749   8.7403011   2.382892 15.097710
0.0018409
Below 50-250 to 749       -6.4437379  -9.321989 -3.565487
0.0000000
750 and Above-50 to 99    14.8087544   8.805150 20.812359
0.0000000
Below 50-50 to 99         -0.3752846  -2.352896  1.602326
0.9851281
Below 50-750 and Above   -15.1840390 -21.155919 -9.212159
0.0000000

> Hs.out=TukeyHSD(one.way)
> Hs.out
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percollege ~ density.category, data =
mysample)

$density.category
```

```
                                  diff        lwr        upr     p
adj
250 to 749-100 to 249        1.1061184  -2.114261  4.326498
0.8793937
50 to 99-100 to 249         -4.9623349  -7.411306 -2.513364
0.0000007
750 and Above-100 to 249     9.8464195   3.702328 15.990511
0.0001536
Below 50-100 to 249         -5.3376196  -7.707754 -2.967485
0.0000000
50 to 99-250 to 749         -6.0684533  -9.011963 -3.124944
0.0000004
750 and Above-250 to 749     8.7403011   2.382892 15.097710
0.0018409
Below 50-250 to 749         -6.4437379  -9.321989 -3.565487
0.0000000
750 and Above-50 to 99      14.8087544   8.805150 20.812359
0.0000000
Below 50-50 to 99           -0.3752846  -2.352896  1.602326
0.9851281
Below 50-750 and Above     -15.1840390 -21.155919 -9.212159
0.0000000
```




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

```
plot(Hs.out,las=2, cex.axis=.6)

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

From the plot we can see the confidence intervals that cross the vertical zero line, and we can conclude that there is no statistically significant difference in means between 250 to 749-100 to 249, and Below 50-50 to 99 at the 95% confidence level. Five groups on the left have significantly lower mean than the three groups on the right side.

6. Conduct a two-way ANOVA using percollege as the dependent variable and both state and inmetro as the independent variables. Plot the results of a Tukey HSD test to show whether/where there are differences in percollege. Briefly explain the results shown in the plot, stating if state and inmetro together appear to show significant mean differences in percollege. Make sure the names of levels of independent variables can be clearly and completely read on the appropriate axis of your plot. Be sure to include and interpret an appropriate test for equality of variances.

```
> #Part 6
> two.way=aov(percollege~state+inmetro,data = mysample)
> summary(two.way)
             Df Sum Sq Mean Sq F value   Pr(>F)
state         4    567   141.8   6.052 0.000117 ***
inmetro       1   1680  1679.6  71.695 2.37e-15 ***
Residuals   244   5716    23.4
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> TukeyHSD(two.way)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percollege ~ state + inmetro, data =
mysample)

$state
              diff         lwr        upr     p adj
IN-IL -2.86384159 -5.33435051 -0.3933327 0.0139893
MI-IL -0.03343858 -2.55116106  2.4842839 0.9999996
OH-IL -3.03230209 -5.56694382 -0.4976604 0.0101052
WI-IL  0.57870316 -2.07259079  3.2299971 0.9750027
MI-IN  2.83040301  0.09949002  5.5613160 0.0380297
OH-IN -0.16846050 -2.91497975  2.5780587 0.9998180
WI-IN  3.44254475  0.58801871  6.2970708 0.0092704
OH-MI -2.99886351 -5.78792790 -0.2097991 0.0280838
```

```
WI-MI  0.61214174 -2.28334287  3.5076263 0.9777728
WI-OH  3.61100525  0.70079679  6.5212137 0.0067542

$inmetro
      diff      lwr     upr p adj
1-0 5.4916 4.190151 6.79305     0

> Hs.out=TukeyHSD(two.way)
> Hs.out
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = percollege ~ state + inmetro, data =
mysample)

$state
             diff         lwr        upr     p adj
IN-IL -2.86384159 -5.33435051 -0.3933327 0.0139893
MI-IL -0.03343858 -2.55116106  2.4842839 0.9999996
OH-IL -3.03230209 -5.56694382 -0.4976604 0.0101052
WI-IL  0.57870316 -2.07259079  3.2299971 0.9750027
MI-IN  2.83040301  0.09949002  5.5613160 0.0380297
OH-IN -0.16846050 -2.91497975  2.5780587 0.9998180
WI-IN  3.44254475  0.58801871  6.2970708 0.0092704
OH-MI -2.99886351 -5.78792790 -0.2097991 0.0280838
WI-MI  0.61214174 -2.28334287  3.5076263 0.9777728
WI-OH  3.61100525  0.70079679  6.5212137 0.0067542

$inmetro
      diff      lwr     upr p adj
1-0 5.4916 4.190151 6.79305     0
```


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

plot(Hs.out,las=2, cex.axis=.6)

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

From the plot we can see that the confidence interval does not include zero, suggesting that there is no statistically significant difference in means between percollege and state and inmetro at the 95% confidence level. Therefore, the factor variables inmetro and state do not seem to have a meaningful impact on percollege.

Your deliverable will be a single MS-Word file showing 1) the R script which executes the above instructions and 2) the results of those instructions. The first line of your script file should be a "#" comment line showing your name as it appears in Canvas. 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 and submitted by the time stated on Canvas. No collaboration of any sort is allowed on this assignment.**