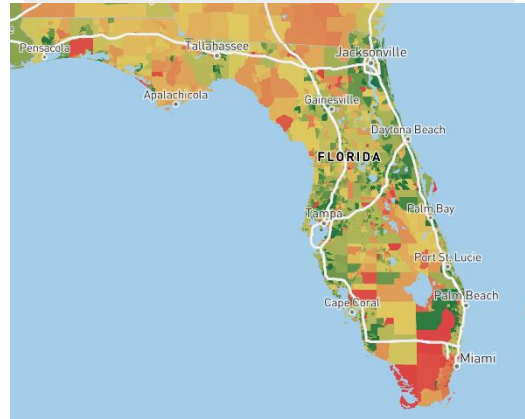# Module Assignment
Module 3
QMB-6304 Foundations of Business Statistics

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 3 Assignment Data.xlsx" into R. This file contains information on crime in each of the 67 counties in Florida. This will be your master data frame.
2. Subset your master data frame, taking only counties with populations below the county population 75th percentile for the entire state. This will be your intermediate data frame.
3. Using the numerical portion of your U number as a random number seed and the method demonstrated in this course, take a random sample of 15 counties from the intermediate data frame. This will be your primary data frame for analytics in this assignment.

```
#Carolina Aldana Yabur
#UID: U25124553
rm(list=ls())
install.packages("rio")
install.packages("moments")
library(rio)
library(moments)
masterdf= import("6304 Module 3 Assignment Data.xlsx")
colnames(masterdf)= tolower(make.names(colnames(masterdf)))
percentile_75th= quantile(masterdf$population,0.75)
intermediatedf= subset(masterdf, masterdf$population <
percentile_75th)
set.seed(25124553)
primarydf = intermediatedf[sample(1:nrow(intermediatedf),15),]
attach(primarydf)
```

Analysis

Using your primary data frame:

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

```
> str(primarydf)
```

```
'data.frame':  15 obs. of  12 variables:
 $ county                     : chr  "Gilchrist" "Columbia"
"Franklin" "Taylor" ...
 $ population                 : num  18269 70617 11864 22436
158834 ...
 $ total.crimes               : num  121 1849 243 689 2099
...
 $ murder                     : num  0 4 0 3 3 1 1 32 1 1
...
 $ rape                       : num  2 26 3 8 14 3 32 225 13
23 ...
 $ robbery                    : num  0 48 0 9 28 1 21 376 19
7 ...
 $ aggravated.assault         : num  24 291 29 118 174 ...
 $ burglary                   : num  34 436 44 240 187 ...
 $ larceny                    : num  52 957 160 256 1513 ...
 $ vehicle.theft              : num  9 87 7 55 180 30 128
732 59 74 ...
 $ crime.rate.per.100k.popln  : num  662 2618 2048 3071 1322
...
 $ clearance.rate.per.100.offenses: num  29.8 36.8 37.4 44.3
25.6 54.6 36.6 32.6 39.8 29.1 ...
>
```

2. Construct a 99% confidence interval on the mean of the Total Crimes variable.

```
> results= t.test(total.crimes, conf.level=0.99)
> results

        One Sample t-test

data:  total.crimes
t = 2.6537, df = 14, p-value = 0.01889
alternative hypothesis: true mean is not equal to 0
99 percent confidence interval:
 -258.5788 4505.5121
sample estimates:
mean of x
 2123.467

>
```

3. Assuming the data in your intermediate data frame represents the population, does your 99% confidence interval include the true population mean for Total Crimes?

```
> mean_total_crimes= mean(intermediatedf$total.crimes)
> cat("Population Mean:", mean_total_crimes, "\n")
Population Mean: 1944.18
> cat("99% Confidence interval:", results$conf.int,"\n")
99% Confidence interval: -258.5788 4505.512
```

Yes, it includes the true population mean for Total Crimes, as we can state with 99% confidence that the true population lies between -258.5788 and 4505.512.

4. Again using your primary data frame, construct a 95% confidence interval on the Total Crimes variable. How much wider is the 99% confidence interval than the 95% confidence interval?

```
> ci99_width= results$conf.int[2] - results$conf.int[1]
> ci99_width
[1] 4764.091
> results2= t.test(total.crimes, conf.level = 0.95)
> ci95_width= results2$conf.int[2] - results2$conf.int[1]
> ci95_width
[1] 3432.482
> ci_difference= ci99_width - ci95_width
> cat("The difference in width for the two confidence intervals
is:", ci_difference,"\n")
The difference in width for the two confidence intervals is:
1331.609
>
```

The difference in width for the two confidence intervals is 1331.609, which means the 99% interval is 1331.609 wider than the 95% confidence interval. This makes sense as a higher confidence interval level corresponds to a wider confidence interval.

5. Using your primary data set can you say (α = .05) that the population mean of the Crime Rate per 100K Population variable is less than 1700?

```
> hyp_test_crime_rate= t.test(crime.rate.per.100k.popln,
mu=1700, alternative = "less")
> hyp_test_crime_rate

        One Sample t-test

data:  crime.rate.per.100k.popln
t = 0.30729, df = 14, p-value = 0.6184
alternative hypothesis: true mean is less than 1700
95 percent confidence interval:
       -Inf 2167.177
sample estimates:
mean of x
 1769.399


>
```

Null hypothesis: The population mean of the Crime Rate per 100K Population variable is 1700. **Alternative hypothesis:** The population mean of the of the Crime Rate per 100K Population variable is less than 1700.

The p-value equals 0.6184, which is greater than the standard significance level 0.05. Therefore, it fails to reject the null hypothesis. This suggests that there is not enough evidence to conclude that the population mean of the variable is less than 1700.

6. Referencing Part 5 above, what would be the maximum "test against" (mu) value in a two-tailed hypothesis test which would yield p = .05 in a test of the Crime Rate per 110K Population variable?

```
> hyp_test_crime_rate2= t.test(crime.rate.per.100k.popln,
mu=2167.177, alternative = "two.sided")
> hyp_test_crime_rate2

        One Sample t-test

data:  crime.rate.per.100k.popln
t = -1.7613, df = 14, p-value = 0.1
alternative hypothesis: true mean is not equal to 2167.177
95 percent confidence interval:
 1285.015 2253.782
sample estimates:
mean of x
 1769.399
```

```
> t.test(crime.rate.per.100k.popln, mu=2253.782, alternative =
"two.sided")

        One Sample t-test

data:  crime.rate.per.100k.popln
t = -2.1448, df = 14, p-value = 0.05
alternative hypothesis: true mean is not equal to 2253.782
95 percent confidence interval:
 1285.015 2253.782
sample estimates:
mean of x
 1769.399
```

Referencing Part 5, when this value 2167.177 was used as the "test-against" (mu) in a two-tailed hypothesis test, the resulting P-value was not 0.05 but 0.01, so I did a second analysis using the upper bound 2253.792 from this 95% confidence interval in another two-tailed hypothesis test with a resulting value of p-value of 0.05.

Therefore, the maximum "test-against" (mu) value in a two-tailed hypothesis test which would yield p=0.05 in a test of the Crime Rate per 110K Population variable is 2253.782.

7. State the name of the county in your primary data frame which has the largest value on the Clearance Rate per 100 Offenses variable.

```
>
primarydf[which.max(primarydf$clearance.rate.per.100.offenses),
c("county")]
[1] "Wakulla"
```

8. Using your primary data frame determine if there is a statistically significant difference between the population means of the Aggravated Assault and Burglary variables.

```
> hyp_test_assault_burglary= t.test(aggravated.assault,
burglary)
> hyp_test_assault_burglary

        Welch Two Sample t-test

data:  aggravated.assault and burglary
t = -0.10913, df = 27.811, p-value = 0.9139
alternative hypothesis: true difference in means is not equal to
0
95 percent confidence interval:
 -334.8815  301.0148
```

```
sample estimates:
mean of x mean of y
 301.2667  318.2000


>
```

<mark>The p-value is equal to 0.9139, which is greater than the standard significance level 0.05. Therefore, it fails to reject the null hypothesis. This means that there is not enough evidence to suggest a statistically significant difference between the population means of the Aggravated Assault and Burglary variables.</mark>

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. **No collaboration of any sort is allowed on this assignment.**