**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.

**#Debjani Sarma**

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

**library("rio")**

**library("moments")**

**getwd()**

**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.

**crime=import("6304 Module 3 Assignment Data.xlsx")**

**colnames(crime)=tolower(make.names(colnames(crime)))**

**attach(crime)**

1. 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.

**population75=quantile(population,0.75)**

**inter.population=subset(crime,population<population75)**

1. 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.

**set.seed(24173877)**

**my.county=inter.population[sample(1:nrow(inter.population),15),]**

**Analysis**

Using your primary data frame:

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

**> str(my.county)**

**'data.frame': 15 obs. of 12 variables:**

**$ county : chr "Columbia" "Hamilton" "Desoto" "Marion" ...**

**$ population : num 70617 14570 37082 368135 28725 ...**

**$ total.crimes : num 1849 324 736 7569 371 ...**

**$ murder : num 4 2 3 19 2 4 0 5 1 9 ...**

**$ rape : num 26 3 22 184 8 20 3 77 7 94 ...**

**$ robbery : num 48 6 13 136 13 23 0 44 4 43 ...**

**$ aggravated.assault : num 291 53 148 1218 75 ...**

**$ burglary : num 436 81 170 1141 41 ...**

**$ larceny : num 957 160 336 4289 203 ...**

**$ vehicle.theft : num 87 19 44 582 29 97 7 89 19 305 ...**

**$ crime.rate.per.100k.popln : num 2618 2224 1985 2056 1292 ...**

**$ clearance.rate.per.100.offenses: num 36.8 25.9 43.9 43.7 65 24.7 37.4 18.2 28.1 31.5 ...**

**> attach(my.county)**

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

**> ci.totalcrime99=t.test(total.crimes,conf.level = 0.99)**

**> ci.totalcrime99**

**One Sample t-test**

**data: total.crimes**

**t = 3.7754, df = 14, p-value = 0.002048**

**alternative hypothesis: true mean is not equal to 0**

**99 percent confidence interval:**

**426.8931 3609.7736**

**sample estimates:**

**mean of x**

**2018.333**

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

**> ci.total=ci.totalcrime99$conf.int**

**> ci.total**

**[1] 426.8931 3609.7736**

**attr(,"conf.level")**

**[1] 0.99**

**> mean(inter.population$total.crimes)**

**[1] 1944.18**

We are 99% confident that the true population for the true population mean for Total Crimes lies within the range of 426.8931 and 3609.7736. The calculated mean of Total Crimes from the intermediate data frame is 1944.18 , which falls within this range.

1. 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?

**> ci.totalcrime95=t.test(total.crimes, conf.level = 0.95)**

**> width95=ci.totalcrime95$conf.int[2] - ci.totalcrime95$conf.int[1]**

**> width99=ci.totalcrime99$conf.int[2] - ci.totalcrime99$conf.int[1]**

**> difference = width99 - width95**

**> difference**

**[1] 889.6456**

99% confidence interval is 889.6456 units wider than 95% confidence interval.

1. 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?

**> pop.mean=t.test(crime.rate.per.100k.popln, mu=1700, alternative = "less")**

**> pop.mean**

**One Sample t-test**

**data: crime.rate.per.100k.popln**

**t = -0.11733, df = 14, p-value = 0.4541**

**alternative hypothesis: true mean is less than 1700**

**95 percent confidence interval:**

**-Inf 1981.186**

**sample estimates:**

**mean of x**

**1679.931**

Since the p-value is significantly greater than 0.05, we fail to reject the null hypothesis. There is not enough evidence to conclude that the population mean of the Crime Rate per 100K Population variable is less than 1700. The 95% confidence interval of (-Inf 1981.186) indicates that the true population mean could be as high as 1981.186, but we don’t have firm lower limit (-Inf).

1. 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?

**> mean.pop=mean(crime.rate.per.100k.popln)**

**> max.pop=max(crime.rate.per.100k.popln)**

**> for(i in seq(from = mean.pop, to =max.pop))**

**+ {**

**+ my.test=t.test(crime.rate.per.100k.popln, mu= i, alternative = "two.sided")**

**+ if (my.test$p.value <= 0.05)**

**+ {**

**+ i=print(i)**

**+ break**

**+ }**

**+ }**

**[1] 2046.931**

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

**One Sample t-test**

**data: crime.rate.per.100k.popln**

**t = -2.1457, df = 14, p-value = 0.04992**

**alternative hypothesis: true mean is not equal to 2046.931**

**95 percent confidence interval:**

**1313.086 2046.776**

**sample estimates:**

**mean of x**

**1679.931**

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 is 2046.931.

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

**> max.offence=which.max(my.county$clearance.rate.per.100.offenses)**

**> my.county$county[max.offence]**

**[1] "Lafayette"**

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

**> mean.difference=t.test(aggravated.assault,burglary, paired = TRUE)**

**> mean.difference**

**Paired t-test**

**data: aggravated.assault and burglary**

**t = 1.1593, df = 14, p-value = 0.2657**

**alternative hypothesis: true mean difference is not equal to 0**

**95 percent confidence interval:**

**-16.94409 56.81076**

**sample estimates:**

**mean difference**

**19.93333**

Since the p-value is significantly greater than 0.05 , we fail to reject the null hypothesis. This indicates that there is not enough evidence to conclude that there is a significant difference between Aggravated Assault and Burglary variables. The mean difference of

suggests that Aggravated Assault exceeds Burglary by about 19.93333. We are 95% confident that the true difference between the population means of the Aggravated Assault and Burglary variables could be as high as 56.81076, and could also reduce by -16.94409.

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.**