Homework 2
QMB 3200: Advanced and Quantitative Methods
Fall 2019

Descriptive Analysis

Submitted to
Dr. Jim Dewey
Florida Polytechnic University

Submitted by
Luiz Gustavo Fagundes Malpele
Department of Data Science
Florida Polytechnic University

September 9th, 2019

1. Introduction

The purpose of this project is to investigate how relative wages vary across the state of Florida, this project includes descriptive statistics which aim at investigating relation of wages and another variable. The data includes information regarding the wages in each county and the population.

2. Summarized data

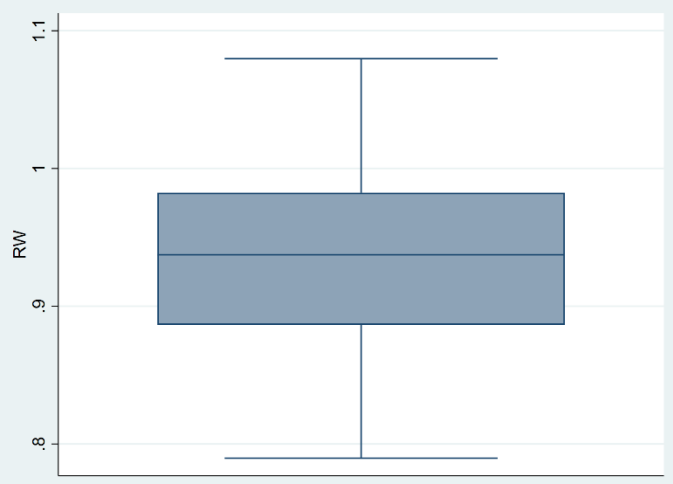
The data of the project is below, summarize statistics, such as mean, number of observations, standard deviation, minimum and maximum, were used to describe the data for 6 different variables: rw (ratio of wage for a specif job in each county relative to the state average wage for that job in 2013), pci (2012 per capita income), pop (2012 population), wden (weighted density, people per squared mile), sh65up (the share of the 2012 population age 65 or older), shlh (the share of 2012 employment in the Leisure and Hospitality sector).

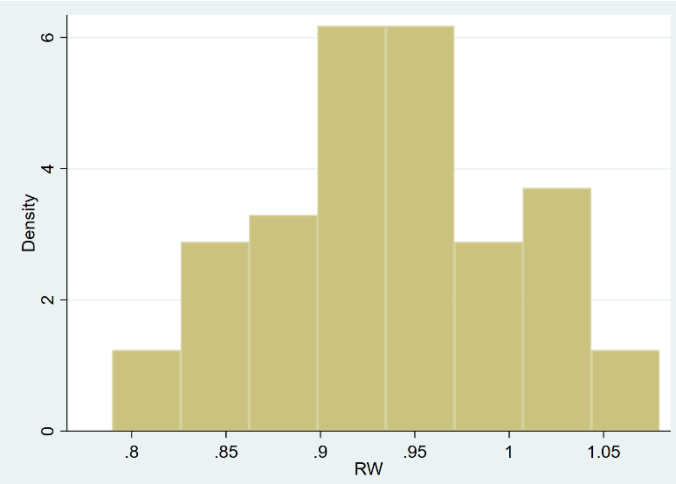
Summarize statistics of the data (1)

Variable	Obs	Mean	Std. Dev.	Min	Max
rw	67	.9367558	.0661804	.7896933	1.079695
pci	67	34921.63	10090.27	19985	65042
pop	67	284693	453786.6	8519	2551290
wden	67	1202.28	1571.373	12.48067	9075.18
sh65up	67	18.68342	6.729126	10.10612	45.4021
shlh	67	10.08139	4.308352	3.800786	26.38741

2. Box plot and frequency distribution for each variable

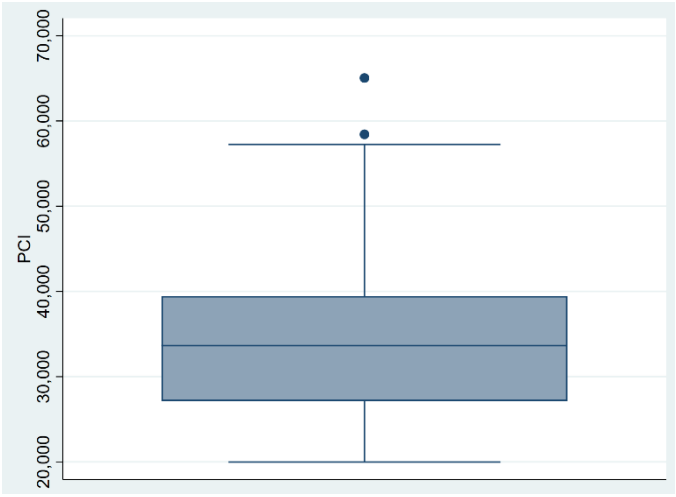
2.1 Box plot and frequency distribution for wage ratio

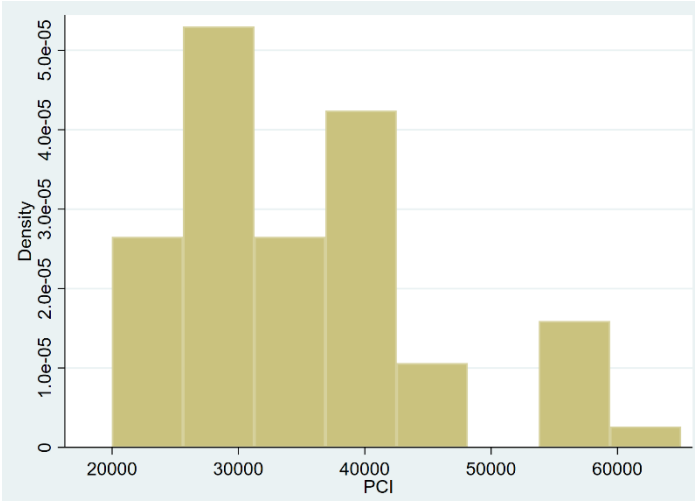




These two graphs show that in most counties the earned was below the state's average, since the 3rd quartile is below 1. This fact can be explained by the fact that in the most populous counties it is earned more than the rest of the other counties, it is noticeable that the 0.90 and 0.95 are the most frequent values on the histogram, so in most of the counties it is earned 5% to 10% less in relation to the Florida state.

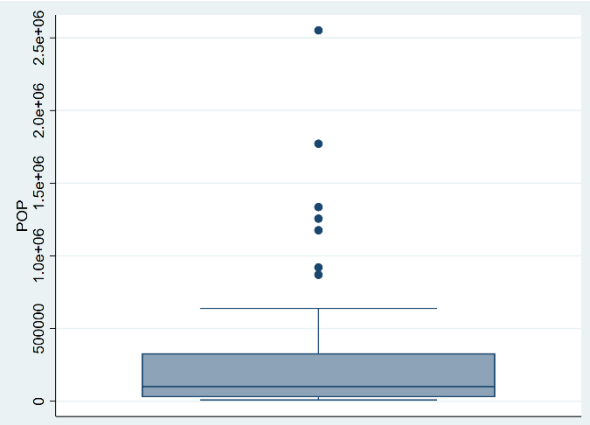
2.2 Box plot and frequency distribution for per capita income in 2012

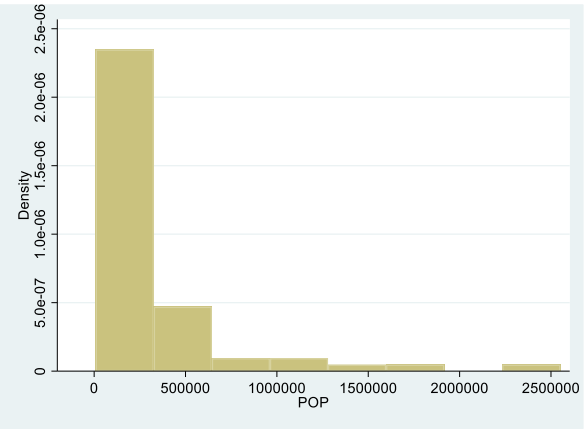




These two graphs show that there are per capita income follows a normal distribution, nonetheless there are outliers which are represented by two points on the box plot with a county average per capita income of approximately U$D 58,000 and U$D 65,000.

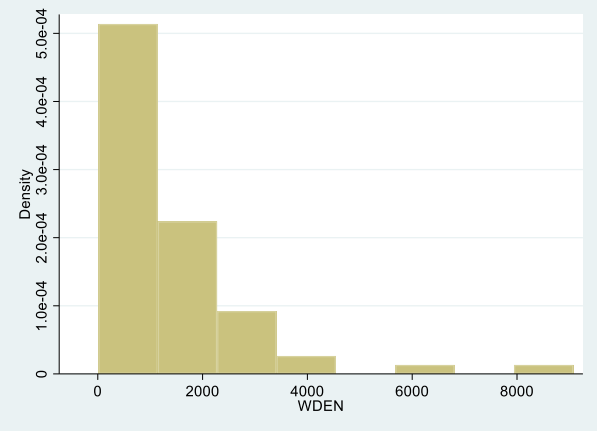
2.3 Box plot and frequency distribution for population





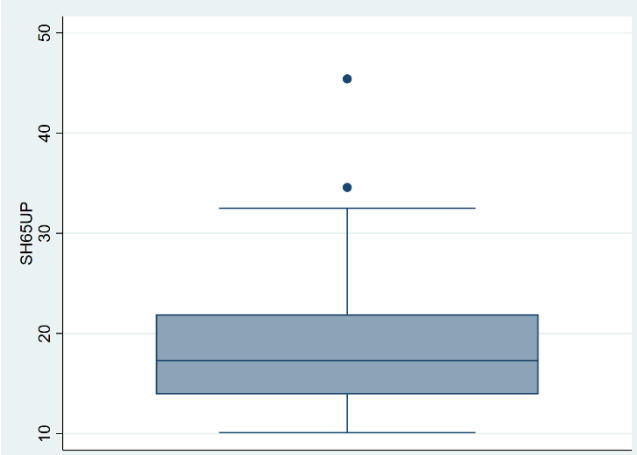
These two graphs show the frequency distribution for population and it can be observed that 75% of counties do not have more than 300,000 habitants. Nonetheless, most of Florida population live on the biggest cities of the states, which are the outliers of the boxplot, such as the 2,500,000-habitant city. The histogram is highly skewed, what illustrates that most cities are around 0 and 250,000 habitants.

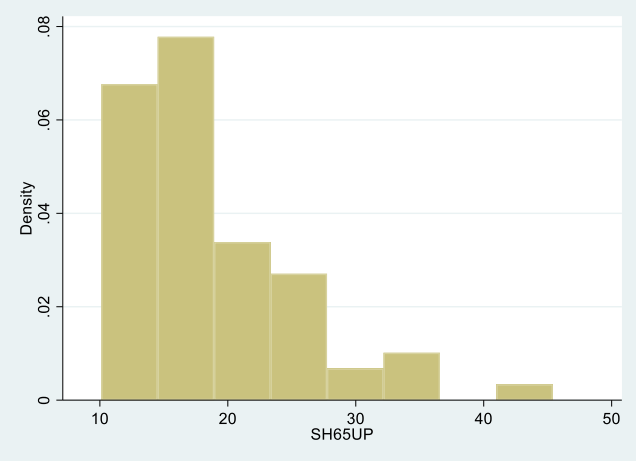
2.4 Box plot and frequency distribution for weighted density



These two graphs show that 75% of counties do not reach 2,000 people per squared mile, this is also illustrated by the histogram which is positively skewed. Furthermore, there are two outliers, both around 6,000 and 9,000 people per squared mile. These two outliers are probably, but not necessarily the two most populous counties, they relation people per area can be influenced by factors such as the unpopulated area of a county, national parks and the number of habitants.

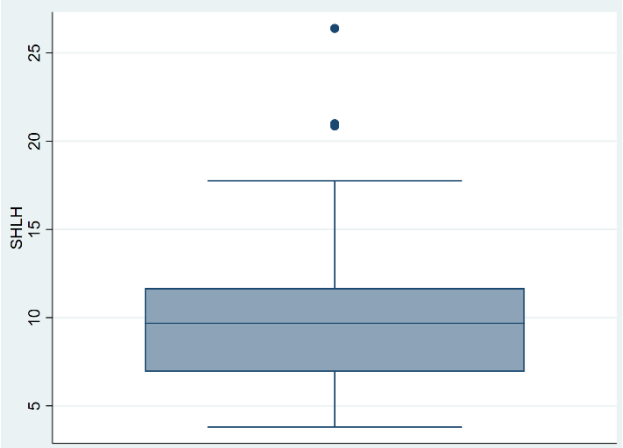
2.5 Box plot and frequency distribution for population aging 65 or more

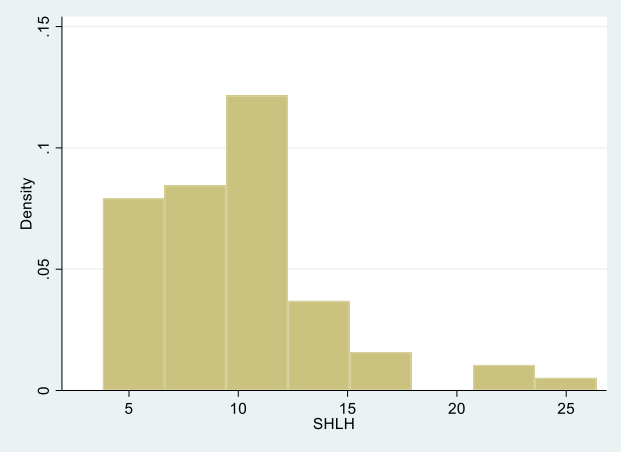




These two graphs show the frequency distribution of people aged 65 or more living on the counties. It can be said that 75% of the states don't have more than 25% percent of people in this category. Nevertheless, there are two outliers that reached 25% and 45% of people aging 65 or more, these two counties are probably known as retirement counties.

2.6 Box plot and frequency distribution for people employed on Leisure or Hospitality sector





These two graphs show that an average of 10% of people in most counties are employed in the leisure or hospitality sector. There are three outliers ranging from 20% to 30% of people employed on these two sectors and these cities main economic activity is related to tourism.

3. Summarize statistics weighed by population

Summarize statistics of the data weighed on population (2)

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
rw	67	19074434	1.001233	.0494401	.7896933	1.079695
pci	67	19074434	41027.27	8038.506	19985	65042
pop	67	19074434	997210.8	796618.6	8519	2551290
wden	67	19074434	3487.37	2628.339	12.48067	9075.18
sh65up	67	19074434	17.83872	6.09767	10.10612	45.4021
shlh	67	19074434	10.80728	3.841975	3.800786	26.38741

This table shows the summary of data weighed on population, instead of counties. There are important points that are useful for the analysis, such as the max values, averages, standard deviations. This table shows that average percentage of people aging 65 or more is 17.8%, average per capita income is U$D 41,000 and that the percentage of people working on leisure or hospitality sector is 10.8%. There are more information that can be taken out from this table, such as minimum, maximum, and standard deviation for each variable.

4. Correlation matrix for logarithmic comparison of each variable

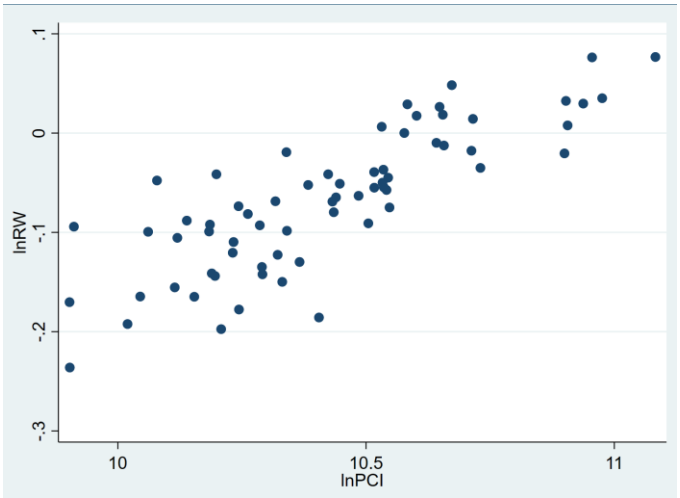
```
. pwcorr lnPOP lnWDEN lnSH65UP lnSHLH lnRW lnPCI
```

	lnPOP	lnWDEN	lnSH65UP	lnSHLH	lnRW	lnPCI
lnPOP	1.0000					
lnWDEN	0.9419	1.0000				
lnSH65UP	0.1365	0.1502	1.0000			
lnSHLH	0.4762	0.5574	0.1220	1.0000		
lnRW	0.8162	0.8312	0.1665	0.4722	1.0000	
lnPCI	0.6962	0.7534	0.3100	0.5748	0.8223	1.0000

These correlation matrix compares the logarithmic value of each variable of the data set, the natural logarithmic (ln) of each variable was taken with the creation of a new variable with the command “gen lnVAR = ln(var)”. The correlation matrix shows a high correlation between the following variables: population and weighed density, population and wage ratio, wage ratio and weighed density, per capita income and wage ratio. After pointing out these strong correlations, it is possible to say that it possible to make more consistent assumptions by referring to these variables’ relations.

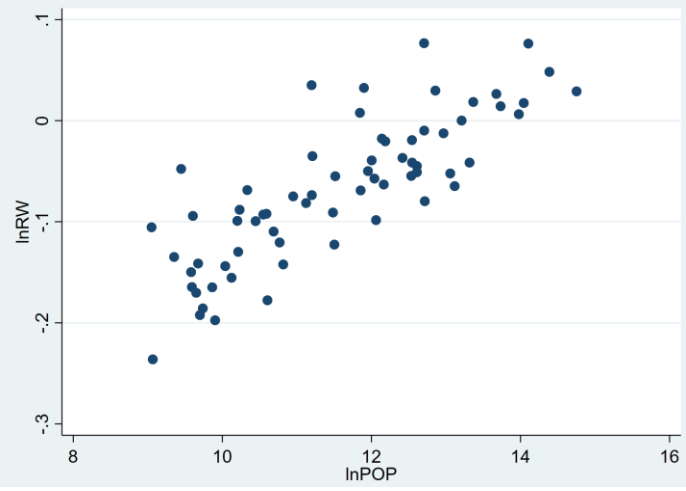
5. Scatter plot for logarithmic comparison of wage ratio and other variables

5.1. Scatter plot for logarithmic comparison of wage ratio x per capita income



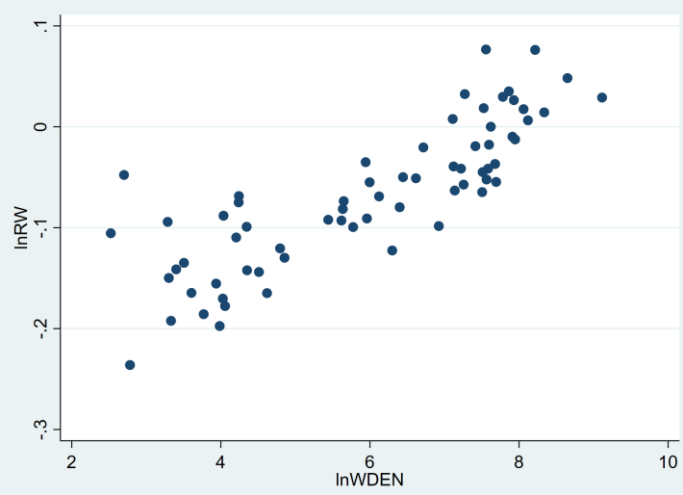
This scatter plot shows that most the points rely between 0 and -0.1 on the y-axis and on the middle of the x-axis, so it shows that most counties' population earn less than average state wage and that population in counties in have a similar per capita annual wage.

5.2. Scatter plot for logarithmic comparison of wage ratio x population



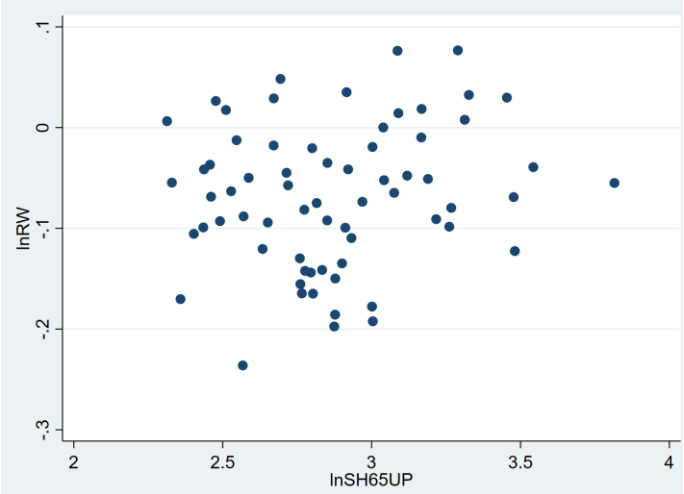
This scatter plot shows a linear pattern that counties with largest population are also the counties with largest wages of the Floridian state. The largest counties rely between 0 and 0.1 on the y-axis, so it means that people on these areas are earning more than the average state wage.

5.3. Scatter plot for logarithmic comparison of wage ratio x weighed density



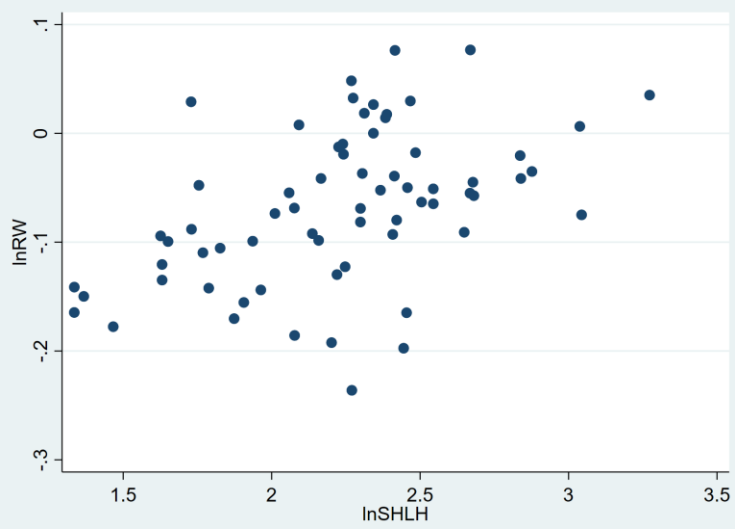
This graph also shows a similar patter to the wage ratio x population, this one shows that the densest counties also have the highest wage ratios of Florida. The graph follows a linear pattern, with a consistent R value.

5.4. Scatter plot for logarithmic comparison of wage ratio x population aging 65 or more



This graph compares rage ration and population aging 65 or more and since it doesn't follow a pattern it is not possible to make assumptions regarding the population. The wages across this population are diverse and do not follow a linear pattern.

5.5. Scatter plot for logarithmic comparison of wage ratio x population aging 65 or more



This scatter plot comparing wage ratio and people employed on leisure and hospitality is inconclusive, there are multiple dots in the middle of x-axis ranging between -0.2 and 0.1 on y-axis. Since the 10.8% of people is the average percentage working on this sector, it is not possible to make direct relation between this variable and wage ratio.

6. Conclusion

Based on the data, it is possible to conclude that most Florida population is located on most populated counties, although 75% of counties have a population size under 300,000 habitants. Furthermore, the most populated counties are also denser and have the highest rage ratios of the state compared to the 75%-smallest counties in terms of population.

Appendix A: Do-file-for-Homework 2

```
*QMB 3200 Homework 2

cd " C:\Users\luizg\Desktop\Homework 2\",

log using “Homework 2.smlc” replace
import delimited "Florida+County+Data.csv"

*Summarizing all variables:
summarize pci
summarize pop
summarize rw
summarize sh65up
summarize shlh
summarize wden

*Descriptive histograms for all variables:
histogram pci
histogram pop
histogram rw
histogram shlh
histogram wden
histogram sh65up

*Descriptive graph box for all variables:
graph box pci
graph box pop
graph box rw
graph box sh65up
graph box shlh
graph box wden
```

```
help summarize

*Summarize with analytical all variables:
summarize pci [aw=pop]
summarize pop [aw=pop]
summarize rw [aw=pop]
summarize sh65up [aw=pop]
summarize shlh [aw=pop]
summarize wden [aw=pop]

*Generate the natural log variables:
gen lnPCI=ln(pci)
gen lnPOP=ln(pop)
gen lnWDEN=ln(wden)
gen lnRW=ln(rw)

*Generate correlation matrix for the natural log variables:
pwcorr lnPCI lnPOP lnRW lnWDEN sh65up shlh

*Scatter plot for the log created
twoway (scatter lnRW lnPOP)
twoway (scatter lnRW lnPCI)
twoway (scatter lnRW lnWDEN)
twoway (scatter lnRW shlh)
twoway (scatter lnRW sh65up)

log close
clear
```

Appendix B :Do-file-for-Homework 2

```
   name:  <unnamed>
       log:  C:\Users\luizg\Desktop\Homework 2.smcl
  log type:  smcl
 opened on:   8 Sep 2019, 20:38:51

. import delimited "C:\Users\luizg\Desktop\Florida+County+Data.csv"
(8 vars, 67 obs)

. summarize rw

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          rw |         67    .9367558    .0661804   .7896933   1.079695

. summarize pci

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pci |         67    34921.63    10090.27      19985      65042

. summarize pop

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
         pop |         67      284693    453786.6       8519    2551290

. summarize wden

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        wden |         67     1202.28    1571.373   12.48067    9075.18
```

```
. summarize sh65up

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      sh65up |         67    18.68342    6.729126   10.10612    45.4021

. summarize shlh

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        shlh |         67    10.08139    4.308352   3.800786   26.38741

. graph box rw

. graph box pci

. graph box pop

. graph box wden

. graph box sh65up

. graph box shlh

. graph box pci

. graph box pop

. graph box wden

. graph box sh65up
```

```
. graph box shlh

. hist rw
(bin=8, start=.7896933, width=.03625021)

. hist pci
(bin=8, start=19985, width=5632.125)

. hist pop
(bin=8, start=8519, width=317846.38)

. hist wden
(bin=8, start=12.48067, width=1132.8374)

. hist sh65up
(bin=8, start=10.106119, width=4.411998)

. sh shlh

. help

. help summarize

. hist pci
(bin=8, start=19985, width=5632.125)

. summarize rw [aw=pop], frac
option frac not allowed
r(198);

. summarize rw [aw=pop]
```

```
    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
          rw |      67    19074434    1.001233   .0494401   .7896933   1.079695

. summarize pci [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
         pci |      67    19074434    41027.27   8038.506      19985      65042

. summarize pop [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
         pop |      67    19074434    997210.8   796618.6       8519    2551290

.  summarize wden [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
        wden |      67    19074434     3487.37   2628.339   12.48067    9075.18

. summarize sh65up [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
      sh65up |      67    19074434    17.83872    6.09767   10.10612    45.4021
```

```
. summarize shlh [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+--------------------------------------------------------------------
        shlh |      67    19074434    10.80728   3.841975   3.800786   26.38741

. gen lnRW = ln(rw)

. gen lnPOP = ln(pop)

. gen lnWDEN = ln(wden)

. gen lnSH65UP = ln(sh65up)

. gen lnSHLH = ln(shlh)

. help cor

. pwcorr lnPOP lnWDEN lnSH65UP lnSHLH lnRW lnPCI

             |    lnPOP   lnWDEN lnSH65UP   lnSHLH     lnRW    lnPCI
-------------+------------------------------------------------------
       lnPOP |   1.0000
      lnWDEN |   0.9419   1.0000
    lnSH65UP |   0.1365   0.1502   1.0000
      lnSHLH |   0.4762   0.5574   0.1220   1.0000
        lnRW |   0.8162   0.8312   0.1665   0.4722   1.0000
       lnPCI |   0.6962   0.7534   0.3100   0.5748   0.8223   1.0000

. twoway (scatter lnRW lnPCI)

. twoway (scatter lnRW lnPOP)
```

```
. twoway (scatter lnRW lnWDEN)

. twoway (scatter lnRW lnSH65UP)

. twoway (scatter lnRW lnSHLH)

. graph box rw

. graph box pci

. graph box pop

. graph box wden

. graph box sh65up

. graph box shlh

. summarize rw[aw=rw]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+---------------------------------------------------------------
          rw |      67  62.7626386    .9413616   .0661518   .7896933   1.079695

. summarize rw [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+---------------------------------------------------------------
          rw |      67    19074434    1.001233   .0494401   .7896933   1.079695
```

```
. summarize pci [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         pci |      67    19074434    41027.27   8038.506      19985      65042

. summarize pop [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
         pop |      67    19074434    997210.8   796618.6       8519    2551290

.  summarize wden [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
        wden |      67    19074434     3487.37   2628.339   12.48067    9075.18

. summarize sh65up [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
      sh65up |      67    19074434    17.83872    6.09767   10.10612    45.4021

. summarize shlh [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+-----------------------------------------------------------------
        shlh |      67    19074434    10.80728   3.841975   3.800786   26.38741
```

```
. summarize rw pci pop wden sh65up shlh

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
          rw |         67    .9367558    .0661804   .7896933   1.079695
         pci |         67    34921.63    10090.27      19985      65042
         pop |         67      284693    453786.6       8519    2551290
        wden |         67     1202.28    1571.373   12.48067    9075.18
      sh65up |         67    18.68342    6.729126   10.10612    45.4021
-------------+---------------------------------------------------------
        shlh |         67    10.08139    4.308352   3.800786   26.38741

. summarize rw pci pop wden sh65up shlh [aw=pop]

    Variable |     Obs      Weight        Mean   Std. Dev.       Min        Max
-------------+---------------------------------------------------------------
--
          rw |      67    19074434    1.001233   .0494401   .7896933   1.079695
         pci |      67    19074434    41027.27   8038.506      19985      65042
         pop |      67    19074434    997210.8   796618.6       8519    2551290
        wden |      67    19074434     3487.37   2628.339   12.48067    9075.18
      sh65up |      67    19074434    17.83872    6.09767   10.10612    45.4021
-------------+---------------------------------------------------------------
--
        shlh |      67    19074434    10.80728   3.841975   3.800786   26.38741
. log close
      name:  <unnamed>
       log:  C:\Users\luizg\Desktop\Homework 2.smcl
  log type:  smcl
 closed on:   9 Sep 2019, 10:01:20
-------------------------------------------------------------------------------
```