Assignment 3

Gus Lipkin

QMB 3200 ~ Advanced Quantitative Methods

Executive Summary

Introduction and Background

Assignment 3 was created as a way for students in QMB 3200 Advanced Quantitative Methods to practice using the statistical software Stata. As the first such assignment, it is relatively short and simple. This data exploration looks at how population, the population density, per capita income, and the wage ratio of each of the sixty-seven counties in Florida are related and correlate with each other.

Data

The data used was supplied by the professor of the class and is called “Florida County Data.csv” The file includes six variables for each of the 67 Florida counties for the year 2012. RW is the typical ratio of the wage for a specific job relative to the state average wage for that job. PCI is the per capita income in the county and POP is the population. WDEN is the weighted population density which takes into account large uninhabited spaces such as forests or farms. SH65UP is the percentage of the population that is age 65 and older. Lastly, SHLH is the percentage of employed people in the Leisure and Hospitality sector. SH65UP can be used as a proxy for migrant retirees and SHLH is a proxy for the importance of tourism in the local economy.

Table 1: Summary Statistics

summarize

Variable	Obs	Mean	Std. Dev.	Min	Max
countyno	67	44	19.48504	11	77
county	0				
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

The data for COUNTYNO and COUNTY are not relevant because they are simply a number indicator for the county and the name of the county itself. With a mean below 1, RW shows that some counties in the state artificially inflate the average wage. PCI further reinforces the wage disparity hypothesis with a minimum wage of under $20k and a maximum of over $65k and a mean wage of $35k. As with any state, there is a large difference in the population of counties with large cities and without. POP illustrates this well with a minimum population of 8.5k residents and a maximum

of 2.5m residents. The weighted density, WDEN, has a large range between a minimum value of 12.48 and maximum of 9075.18. SH65UP is less varied with a mean of 18.68% and a maximum of 45.40%. The share of leisure workers, SHLH is distributed similarly with a mean of 10.08% and a maximum of 26.39%.

Analysis and Discussion

In Table 1, WDEN shows that while there is a large disparity in population by county, there is an even larger disparity in weighted density by county. The largest population is approximately 295 times larger than the smallest while the most dense county is approximately 730 times more dense than the least dense county. No other variables have such stark and significant differences in minimum and maximum.

While the initial summary statistics are important and useful, some of the numbers are heavily impacted by county population. By weighting the variables by county population, the data becomes much more usable.

Table 2: Summary Statistics Weighted by POP

```
summarize [aw=pop]
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
countyno	67	19074434	41.90642	19.7	11	77
county	0	0				
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

In Table 1, the mean value for RW is .94 which does not make sense because each datapoint is supposed to be representative of the wage ratio which means that the average should be equal to one. However, once RW is weighted by population, the data is corrected and the mean becomes equal to 1. PCI and WDEN see significant changes from the unweighted mean as well, but the new values themselves are not significant. SH65UP and SHLH see very little change when weighted because the initial values are the share of the population.

To make the scaling of any graphs easier to read, the natural log of RW, PCI, POP, and WDEN were found. The four new natural log variables as well as SH65UP and SHLH were then put into a correlation matrix. The correlation matrix allows us to see relationships between variables that might not otherwise be apparent just by looking at raw numbers or graphs.

Table 3: Correlation Matrix of All Variables

```
. cor lnRW lnPCI lnPOP lnWDEN sh65up shlh
(obs=67)
```

	lnRW	lnPCI	lnPOP	lnWDEN	sh65up	shlh
lnRW	1.0000					
lnPCI	0.8223	1.0000				
lnPOP	0.8162	0.6962	1.0000			
lnWDEN	0.8312	0.7534	0.9419	1.0000		
sh65up	0.1836	0.3069	0.1500	0.1694	1.0000	
shlh	0.4346	0.5486	0.3843	0.4725	0.0999	1.0000

The correlation between lnPOP and lnWDEN is the highest with a score of .9419. We can then visually verify the similarities in distributions by comparing the histograms (Figures 1 and 2) and graphing them in a graphical version of the correlation matrix (Figure 3). The high correlation between the two makes sense because as the population of an area increases, it will become more dense unless the area is able to expand at the same rate that people enter. In a well-established region like Florida, this is unlikely to happen.

Figure 1: Histogram of POP

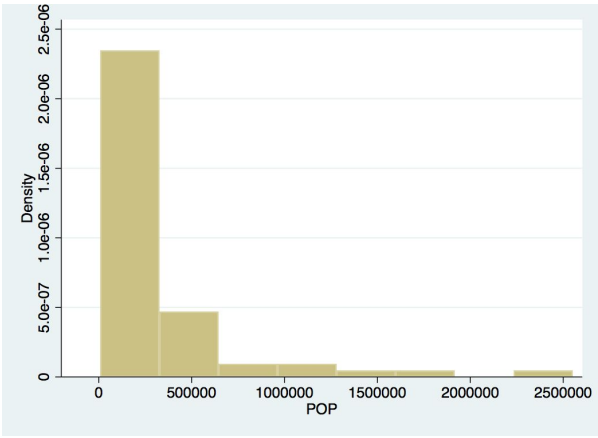


Figure 2: Histogram of WDEN

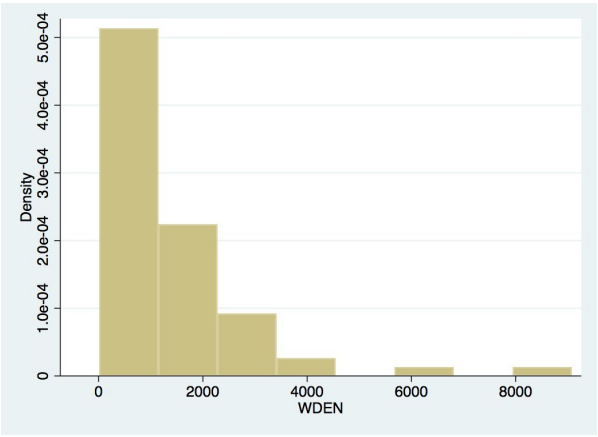
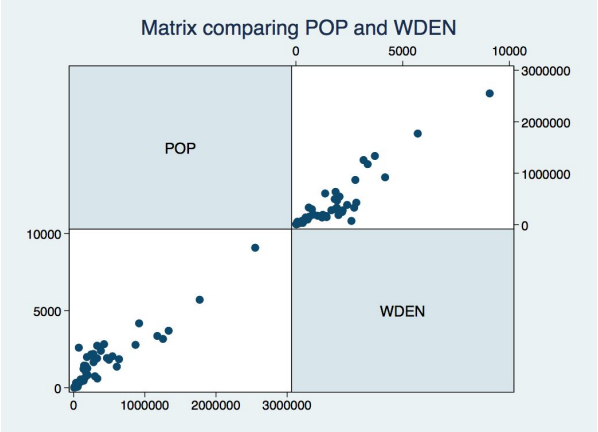
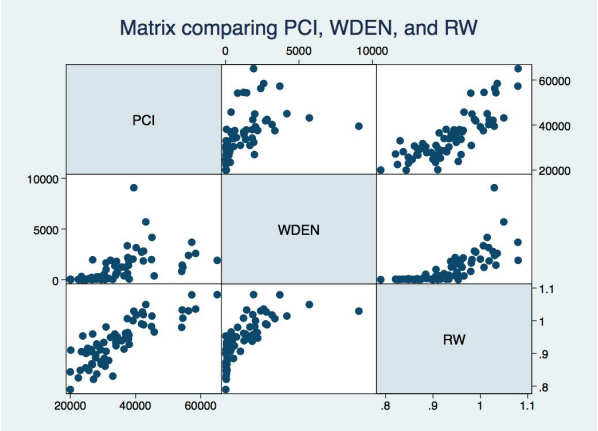


Figure 3: Graphical Matrix of POP and WDEN



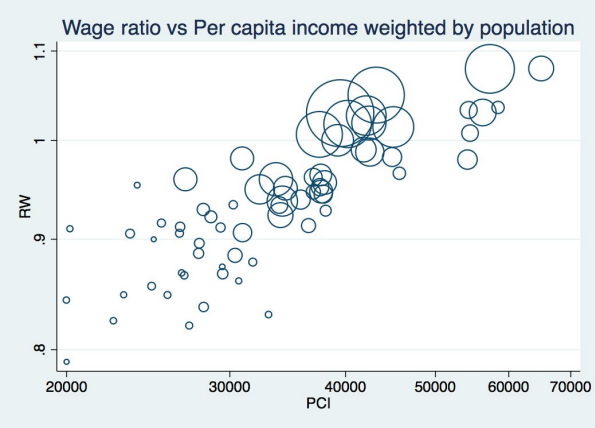
If we further examine Table 3, we see that lnRW is highly correlated with lnPCI and lnWDEN with correlation scores of .8223 and .8312 respectively. These relationships are then graphed and shown in Figure 4.

Figure 4: Graphical Matrix of PCI, WDEN, and RW



It makes sense that PCI and WDEN would be correlated as wages generally are higher in cities than in rural areas and cities are more densely populated than rural areas. Thus, as population density increases, so do wages. Because per capita income and the wage ratio are so closely related, it also makes sense that RW and PCI would be correlated. Figure 4 reinforces this by showing a graph of the wage ratio set against the per capita income with data points scaled by population.

Figure 4: Scatterplot of RW vs PCI weighted by POP



As the population of a county increases, and therefore becomes more urban, wages rise and present higher in the wage ratio. The most populous counties in Florida tend to have the highest per capita income while also scoring above a 1 on the wage ratio.

Conclusion

As an exploration into a new piece of software goes, this first look at using Stata went well. The graphs produced are neat, easy to read, and clearly show the data and any correlations that there may be such as the one between population and weighted population density or the one between wage ratio and per capita income. As expected, there is a clear correlation between the wage ratio, per capita income, and weighted population density because they all rely on population as a factor of their calculation. Without delving too deep into research about the primary jobs in each county or how much different cities pay compared to average wages, it is difficult to draw any meaningful conclusions. Said research would be the next step if we wanted to take this further and learn anything useful.

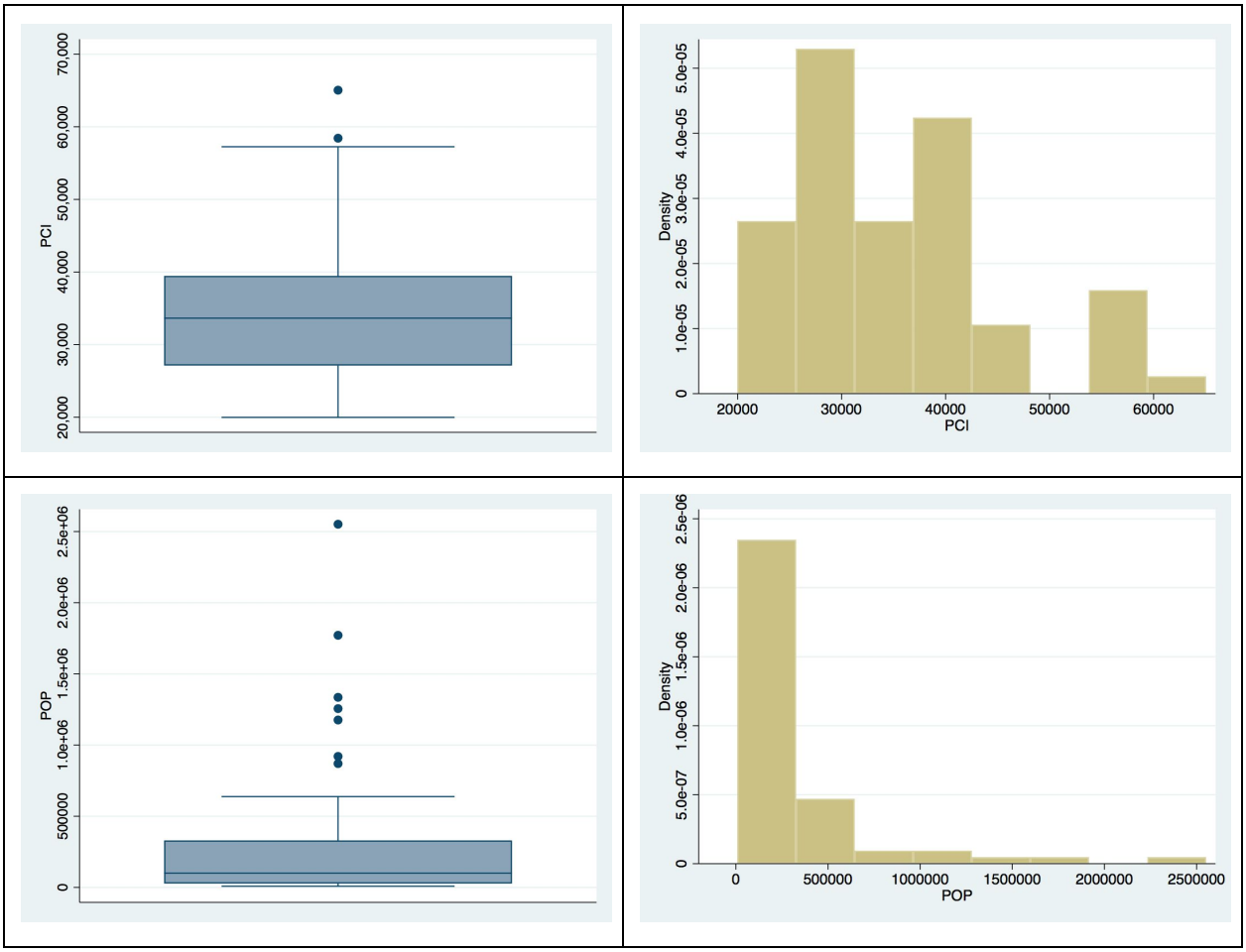
Appendix

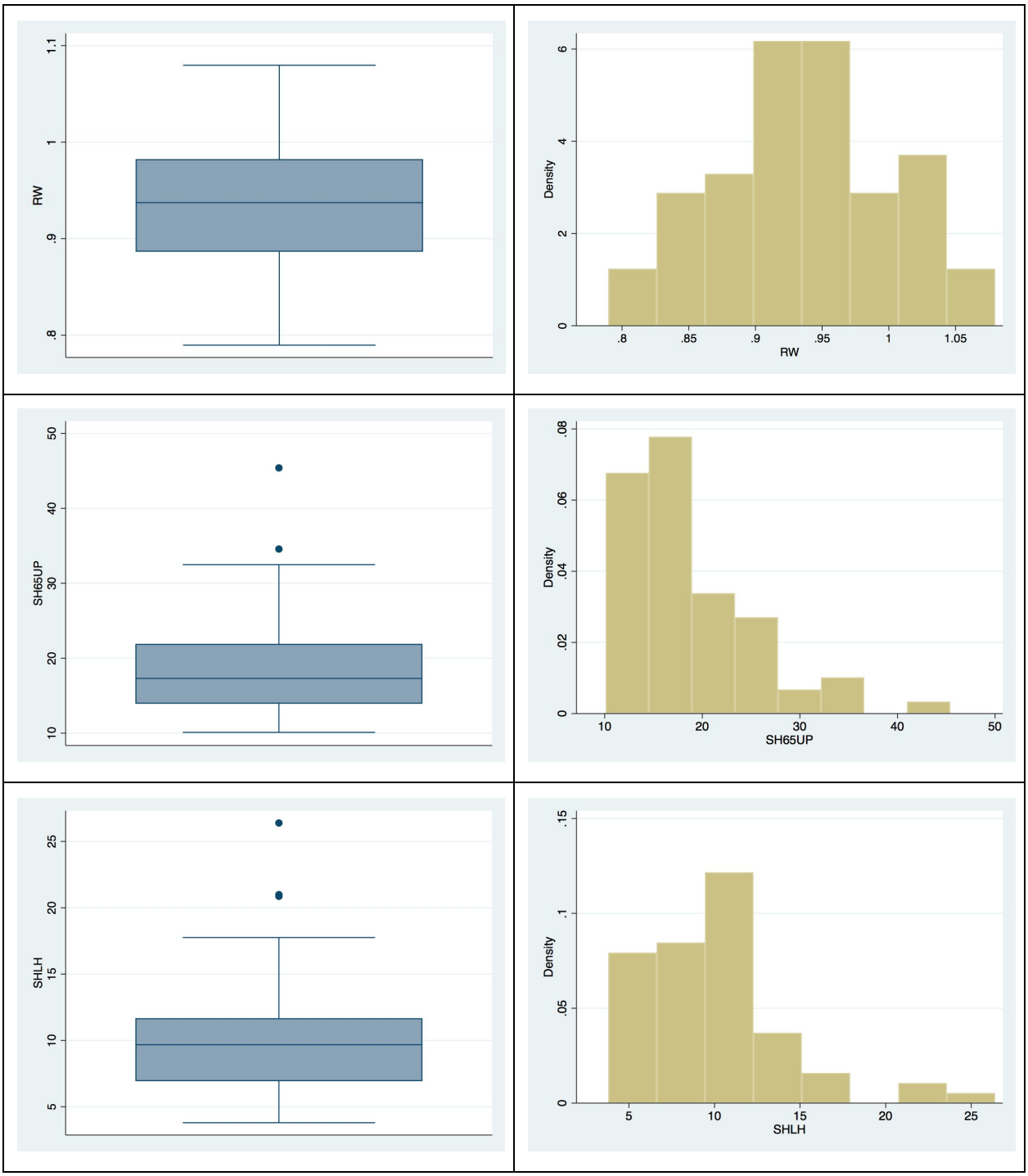
1 Calculate the summary statistics for each variable

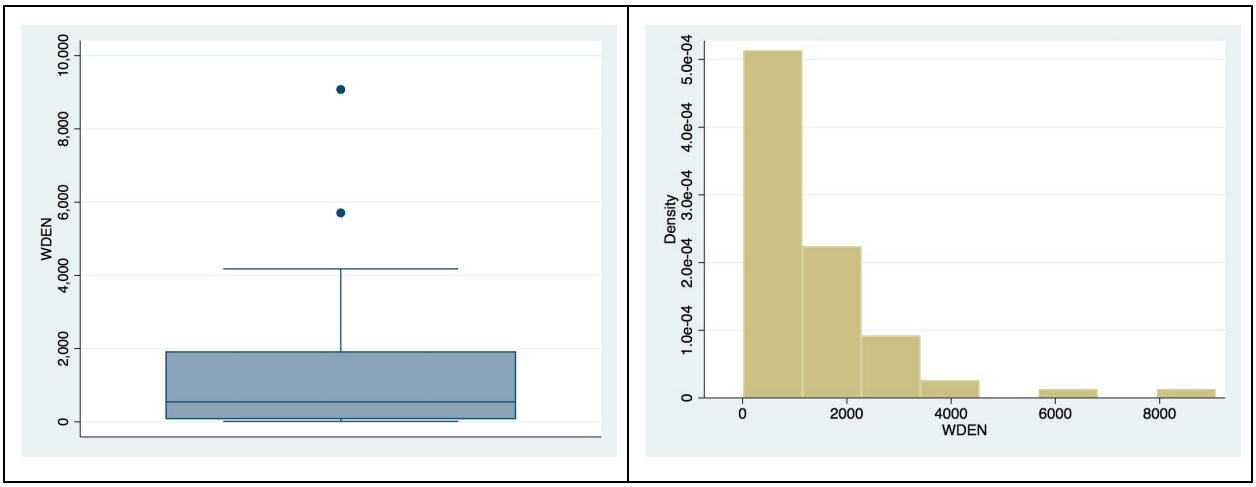
```
summarize
```

Variable	Obs	Mean	Std. Dev.	Min	Max
countyno	67	44	19.48504	11	77
county	0				
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 Produce box and whisker and histograms for each variable




RW
.8
.9
1
1.1
Density
0
2
4
6
.8
.85
.9
.95
1
1.05
RW
SH65UP
10
20
30
40
50
Density
0
.02
.04
.06
.08
10
20
30
40
50
SH65UP
SHLH
5
10
15
20
25
Density
0
.05
.1
.15
5
10
15
20
25
SHLH



3 Calculate summary statistics for each variable, weighted by county population

```
summarize [aw=pop]
```

Variable	Obs	Weight	Mean	Std. Dev.	Min	Max
countyno	67	19074434	41.90642	19.7	11	77
county	0	0				
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

4 Create new variables that equal the natural log of…

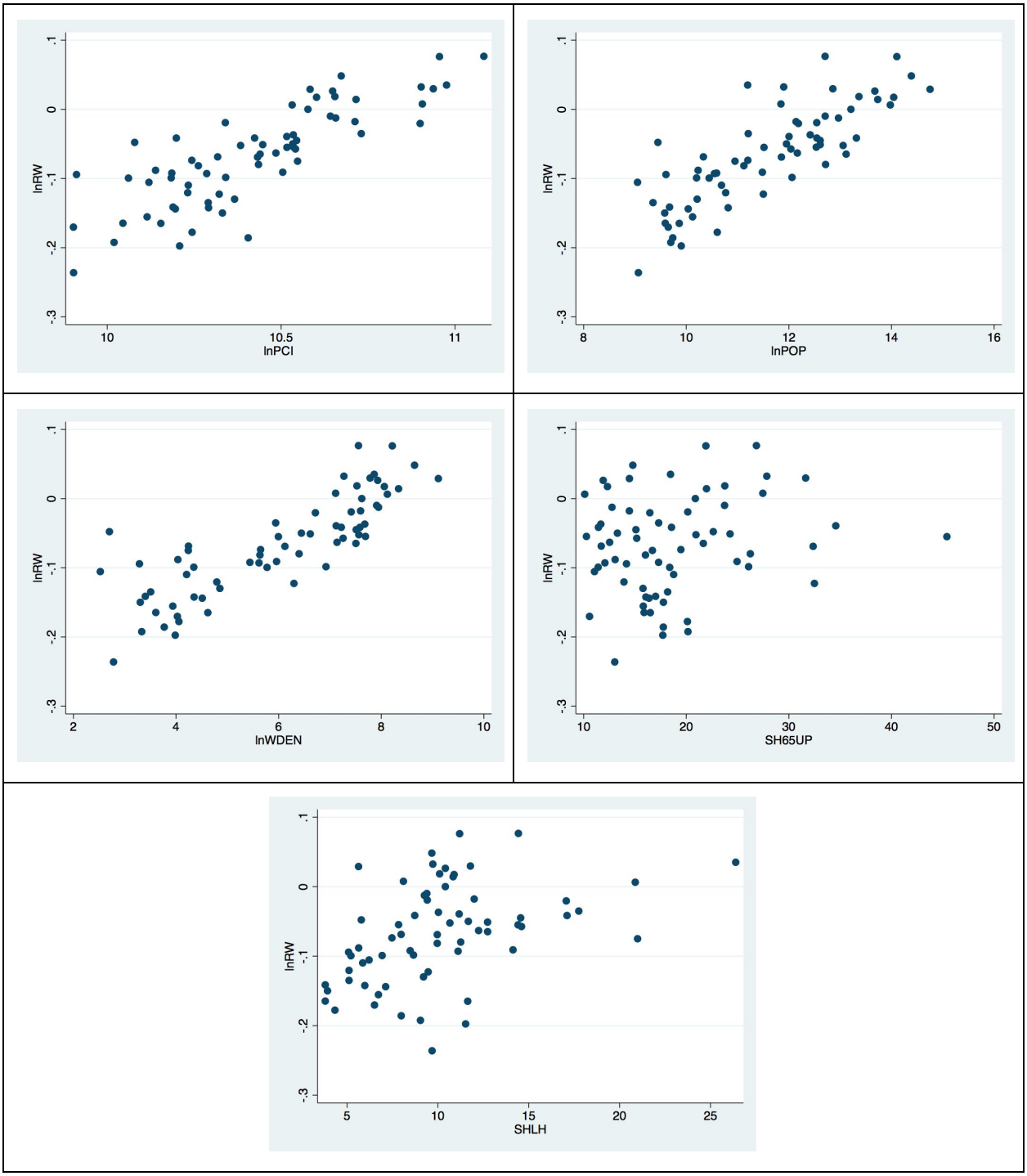
This is not shown graphically. The code for it is included in the 'Stata do-file' section.

5 Produce the correlation matrix for…

```
. cor lnRW lnPCI lnPOP lnWDEN sh65up shlh
(obs=67)
```

	lnRW	lnPCI	lnPOP	lnWDEN	sh65up	shlh
lnRW	1.0000					
lnPCI	0.8223	1.0000				
lnPOP	0.8162	0.6962	1.0000			
lnWDEN	0.8312	0.7534	0.9419	1.0000		
sh65up	0.1836	0.3069	0.1500	0.1694	1.0000	
shlh	0.4346	0.5486	0.3843	0.4725	0.0999	1.0000

6 Produce scatterplots for lnRW against…



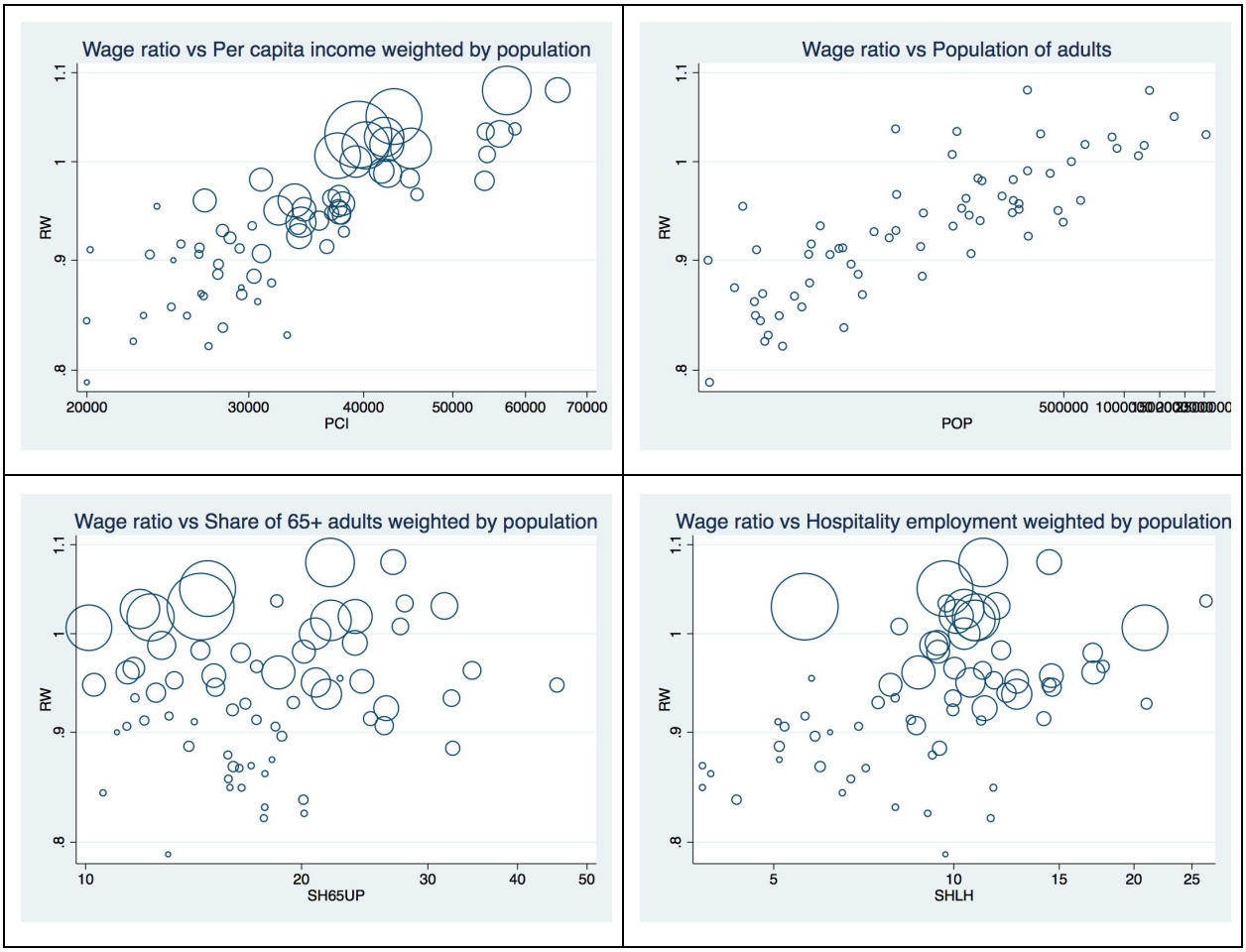
a Change the markers to hollow circles and use weighting…

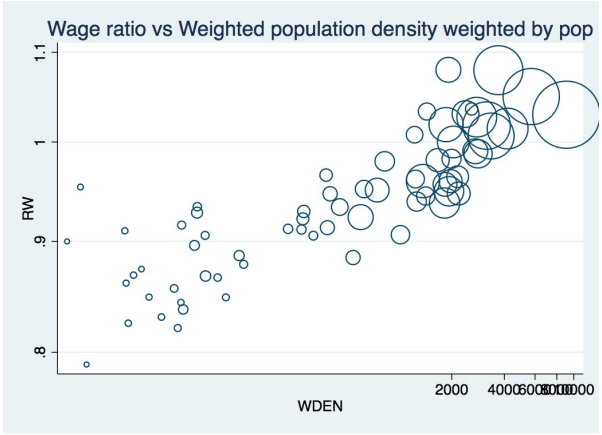
Answers for extra credit a, b, and c are aggregated in section ‘c Try changing the scatterplots to use a log scale…’

b Change the labels on the titles on the scatterplots to be self-explanatory

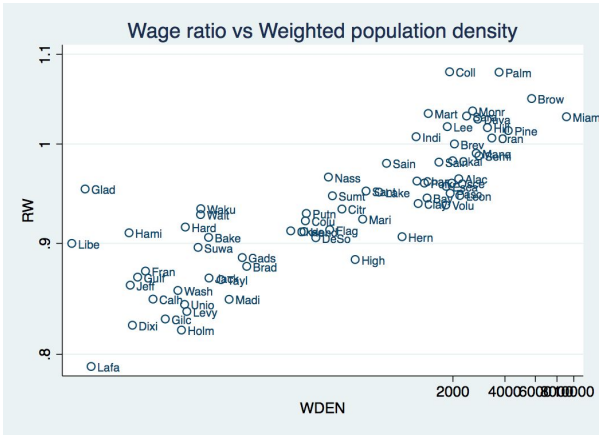
Answers for extra credit a, b, and c are aggregated in section ‘c Try changing the scatterplots to use a log scale…’

c Try changing the scatterplots to use a log scale…

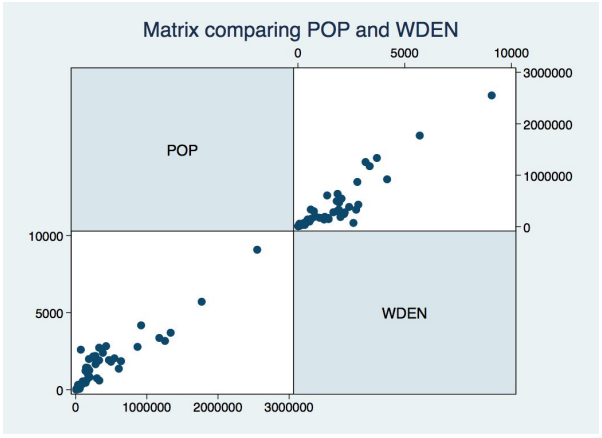


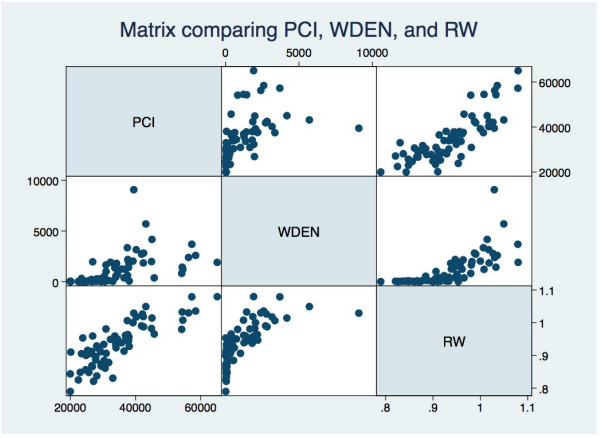


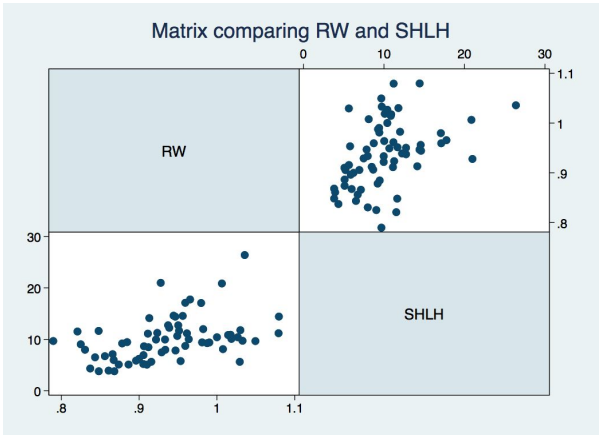
d See if you can figure out how to create a variable…



e Use “graph matrix” to produce a matrix of scatter plots…







Stata do-file

```
capture log close
log using gus_lipkin_Assignment_3, replace
import delimited "/Users/guslipkin/Documents/Fall2020/QMB 3200 ~
Advanced Quantitative Methods/FloridaCountyData.csv"
clear all

* Main assignment

* Problem 1
summarize

* Problem 2
graph box rw
graph box pci
graph box pop
graph box wden
graph box sh65up
graph box shlh

hist rw
hist pci
hist pop
hist wden
hist sh65up
hist shlh

* Problem 3
summarize [aw=pop]

* Problem 4
gen lnPOP = ln(pop)
gen lnWDEN = ln(wden)
gen lnPCI = ln(pci)
gen lnRW = ln(rw)

* Problem 5
cor lnRW lnPCI lnPOP lnWDEN sh65up shlh
```

```
* Problem 6
scatter lnRW lnPCI
scatter lnRW lnPOP
scatter lnRW lnWDEN
scatter lnRW sh65up
scatter lnRW shlh

* Extra Credit a
scatter lnRW lnPCI [w=pop], msymbol(circle_hollow)
scatter lnRW lnPOP
scatter lnRW lnWDEN [w=pop], msymbol(circle_hollow)
scatter lnRW sh65up [w=pop], msymbol(circle_hollow)
scatter lnRW shlh [w=pop], msymbol(circle_hollow)

* Extra Credit b
scatter lnRW lnPCI [w=pop], msymbol(circle_hollow) title("Wage ratio
vs Per capita income weighted by population")
scatter lnRW lnPOP, title("Wage ratio vs Population of adults")
scatter lnRW lnWDEN [w=pop], msymbol(circle_hollow) title("Wage ratio
vs Weighted population density weighted by pop")
scatter lnRW sh65up [w=pop], msymbol(circle_hollow) title("Wage ratio
vs Share of 65+ adults weighted by population")
scatter lnRW shlh [w=pop], msymbol(circle_hollow) title("Wage ratio
vs Hospitality employment weighted by population")

* Extra Credit c
scatter rw pci [w=pop], msymbol(circle_hollow) title("Wage ratio vs
Per capita income weighted by population") xscale(log) yscale(log)
scatter rw pop, msymbol(circle_hollow) title("Wage ratio vs
Population of adults") xscale(log) yscale(log)
scatter rw wden [w=pop], msymbol(circle_hollow) title("Wage ratio vs
Weighted population density weighted by pop") xscale(log) yscale(log)
scatter rw sh65up [w=pop], msymbol(circle_hollow) title("Wage ratio
vs Share of 65+ adults weighted by population") xscale(log)
yscale(log)
scatter rw shlh [w=pop], msymbol(circle_hollow) title("Wage ratio vs
Hospitality employment weighted by population") xscale(log)
yscale(log)
```

```
* Extra Credit d
gen shortCounty = substr(county, 1, 4)
scatter rw pci [w=pop], msymbol(circle_hollow) title("Wage ratio vs
Per capita income") xscale(log) yscale(log) mlabel(shortCounty)
scatter rw pop, title("Wage ratio vs Population of adults")
xscale(log) yscale(log) mlabel(shortCounty)
scatter rw wden [w=pop], msymbol(circle_hollow) title("Wage ratio vs
Weighted population density") xscale(log) yscale(log)
mlabel(shortCounty)
scatter rw sh65up [w=pop], msymbol(circle_hollow) title("Wage ratio
vs Share of 65+ adults") xscale(log) yscale(log) mlabel(shortCounty)
scatter rw shlh [w=pop], msymbol(circle_hollow) title("Wage ratio vs
Hospitality employment") xscale(log) yscale(log) mlabel(shortCounty)

* Extra Credit e
graph matrix rw shlh, title("Matrix comparing RW and SHLH")
graph matrix pci wden rw, title("Matrix comparing PCI, WDEN, and RW")

log close
```