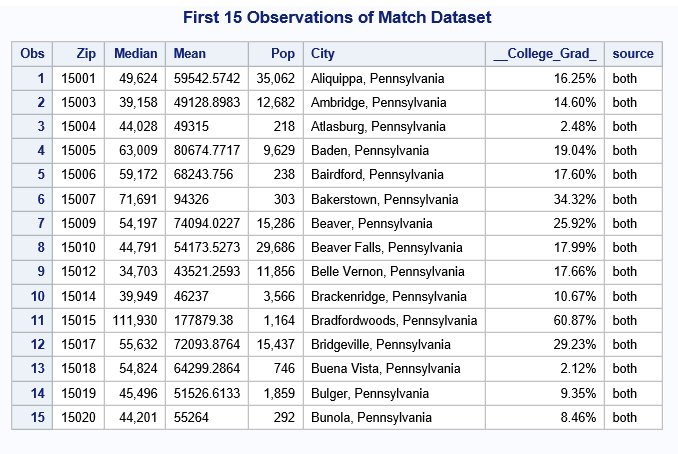
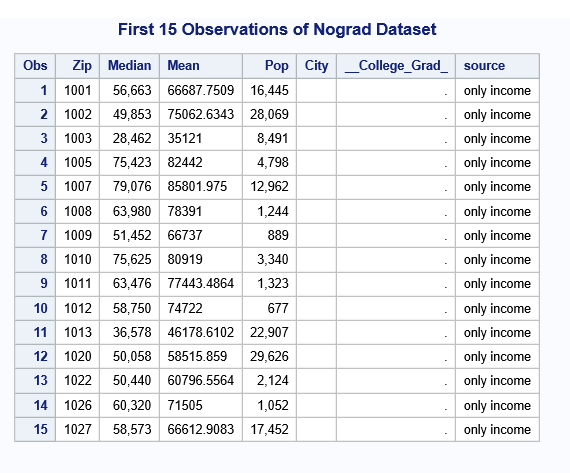
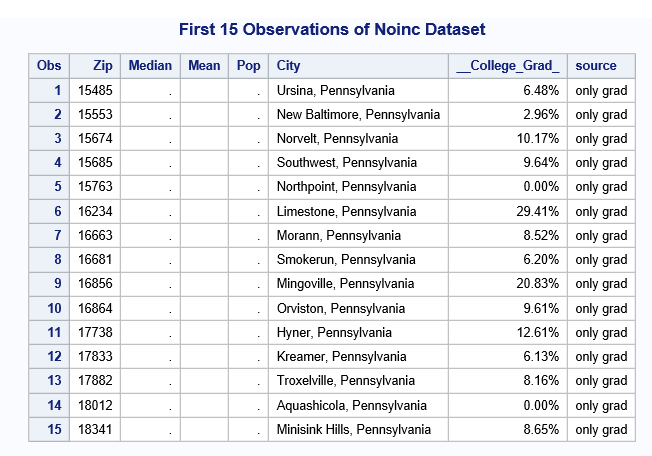
1. I read the files in using proc import statements. See the appendix for details.

2. Here are the first 15 observations of each dataset.



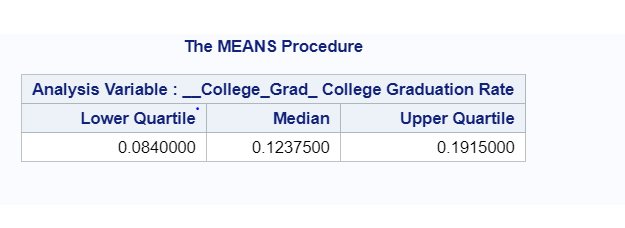




3. The *match* dataset has 1,690 zip codes; the *no grad* dataset has 30,944 zip codes; and the *noinc* dataset has 26 zipcodes. Most of the zipcodes in the *noinc* dataset have very small populations or even a population of zero, so the Census Bureau—or the source of our Median Income by Zip dataset—probably didn’t calculate income information for these sparsely populated areas. For instance, Mingoville (zip 16856) has a population of zero. Birchrunville (zip 19421) is a small collection of farms and large homes along with a high-end restaurant. The *noinc* dataset does include 19112, the Philadelphia zipcode for the Navy Yard, however, an industrial area with 8 full-time residents and 107 businesses, according to zip-codes.com. Likewise, the *nograd* dataset is primarily zipcodes not in Pennsylvania. The vast majority of the *nograd* dataset are zipcodes outside of Pennsylvania , so they didn’t match with our data on Pennsylvania college graduation rates. Yet I found 99 Pennsylvania zipcodes that did not have college graduation data. Again, most of the were sparsely populated, but they did include 17015, which covers Carlisle (population 19,037) , home to Dickinson College, the U.S. Army War College, and the Penn State Dickinson School of Law. Other zipcodes on the *nograd* included 17050 which covers Mechanicsburg, PA, and has a population of 27,008 and 17202 which includes Chambersburg and covers a population 25,368. It might be worth investigating the data sources to see why these communities were not included.

4.

a. Here are the quartiles of the college graduation rate variable.



b. See the appendix for the creation of a new variable called *CollGradGroup*.

4.c. The graphs of the median income variable reveal how the mean of median income distribution increases as we move from zipcodes with low college graduation rate to those with high college graduation rates. In the “low” group, the mean appears to be roughly centered at $49,000 and in the “high” group, it appears to be centered around $66,000. Also, the distribution is much more spread out in the “high” group, although the other distributions don’t seem to increase in spread.

|  |
| --- |
| ***Median Income Distribution in PA Zipcodes by College Graduation Rate***  ***Distribution of Median Income***  ***The Univariate Procedure*** |



d. Here is the SAS dataset with the mean values of the median income variable and the population variable separately for each *CollGradGroup*. These values confirm what we first observed in the above graph: the range for the high college graduation rate zipcodes is much bigger (173,425.) than for the other groups (98,382 for low; 93,5878 for med-low; and 80,369 for med-high). Indeed, the standard deviation for high college graduation rate group is more than twice the standard deviation of the other groups.

| **College Graduate Rate** | **N Obs** | **Variable** | **Label** | **N** | **Mean** | **Std Dev** | **Minimum** | **Maximum** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| low | 423 | Median Pop | Median Pop | 423 423 | 43197.75 2534.21 | 10240.15 7241.14 | 13763.61 14.0000000 | 112146.00 62905.00 |
| med-low | 422 | Median Pop | Median Pop | 422 422 | 44143.88 4639.35 | 9922.46 7508.97 | 743.8831000 23.0000000 | 94331.74 68584.00 |
| med-high | 424 | Median Pop | Median Pop | 424 424 | 48601.99 8327.76 | 9735.89 10283.87 | 14341.14 3.0000000 | 94710.00 70634.00 |
| high | 421 | Median Pop | Median Pop | 421 421 | 66133.43 12996.99 | 22591.16 12478.85 | 8049.44 29.0000000 | 181474.82 58524.00 |

Appendix: SAS Code

\*\*1. Reading in the first file;

proc import out=medianzip datafile='/folders/myfolders/Stat Programming/Median Income by Zip Code in US.xlsx'

dbms=xlsx replace;

getnames=yes;

run;

\*\*Importing the graduation rate data, renaming zipcode as zip;

proc import out=gradzip datafile='/folders/myfolders/Stat Programming/PA College Graduation By Zip Code.xlsx'

dbms=xlsx replace;

getnames=yes;

run;

\*\*2. Merging the files together by zipcode;

\*Now I rename the zipcode labels so they are the same;

data gradzip2;

set gradzip (rename=(zip\_code=Zip));

run;

\*\*Now we merge, creating variable called source;

proc sort data=medianzip; by zip;

proc sort data=gradzip2; by zip;

data merged;

merge medianzip (in=income) gradzip2 (in=grad); by zip;

if income and grad then source='both ';

else if income then source='only income ';

else if grad then source='only grad ';

run;

\*\*2 and 3. Splitting the merged dataset into three output datasets, match that containes only those zipcodes

that were in both files, noinc that has zipcodes but no income data, and a third one called nograd that

includes zip codes that don't have grad data;

data match noinc nograd;

set merged;

if source='both ' then output match;

if source='only income ' then output nograd;

if source='only grad ' then output noinc;

run;

\*\*Printing a selection of the three datasets for the assignment;

proc print data=match (obs=15); title "First 15 Observations of Match Dataset"; run;

proc print data=nograd (obs=15); title "First 15 Observations of Nograd Dataset"; run;

proc print data=noinc (obs=15); title "First 15 Observations of Noinc Dataset"; run;

\*\*Calculating the quartiles of the college graduation rate variable in the match dataset;

proc print match; run;

proc sort data=match; by \_\_College\_Grad\_; run;

proc print match; run;

proc means data=match q1 median q3;

var \_\_College\_Grad\_;

run;

\*\*\*4b. Creating the CollGradGroups;

proc sort data=match; by \_\_College\_Grad\_; run;

proc format; value CollGradGroupf 1="low" 2="med-low" 3="med-high" 4="high";

data match2;

set match;

format CollGradGroup CollGradGroupf.;

length CollGradGroup 8.;

if (0.0000 <= \_\_College\_Grad\_) and (\_\_College\_Grad\_ <= 0.084) then CollGradGroup=1;

else if (0.084 < \_\_College\_Grad\_) and (\_\_College\_Grad\_ <= 0.12375) then CollGradGroup=2;

else if (0.12375 < \_\_College\_Grad\_) and (\_\_College\_Grad\_ <= 0.191500) then CollGradGroup=3;

else if (0.191500 < \_\_College\_Grad\_) and (\_\_College\_Grad\_ <= 1.000) then CollGradGroup=4;

label CollGradGroup="College Graduate Rate";

run;

\*\*\*4c. Creating a graph of the median income variable;

title 'Median Income Distribution in PA Zipcodes by College Graduation Rate';

proc univariate data=work.match2 noprint; class CollGradGroup;

VAR median; histogram / normal (COLOR=RED W=5) nrows=2 ncols=2 cfill=blue;

INSET N='N:' (4.0) MIN='MIN:' (4.0) MAX='MAX:' (4.0) MEAN='MEAN:' (4.1) / NOFRAME POSITION=NE HEIGHT=2;

RUN;

label median="Median Income Distribution";

RUN;

\*\*\*4d. Creating a SAS dataset with the mean values of the median income variable;

proc means data=match2;

var median pop;

class CollGradGroup;

output out=new(drop=\_type\_ \_freq\_) / autoname;

label median=median income pop=population;

run;