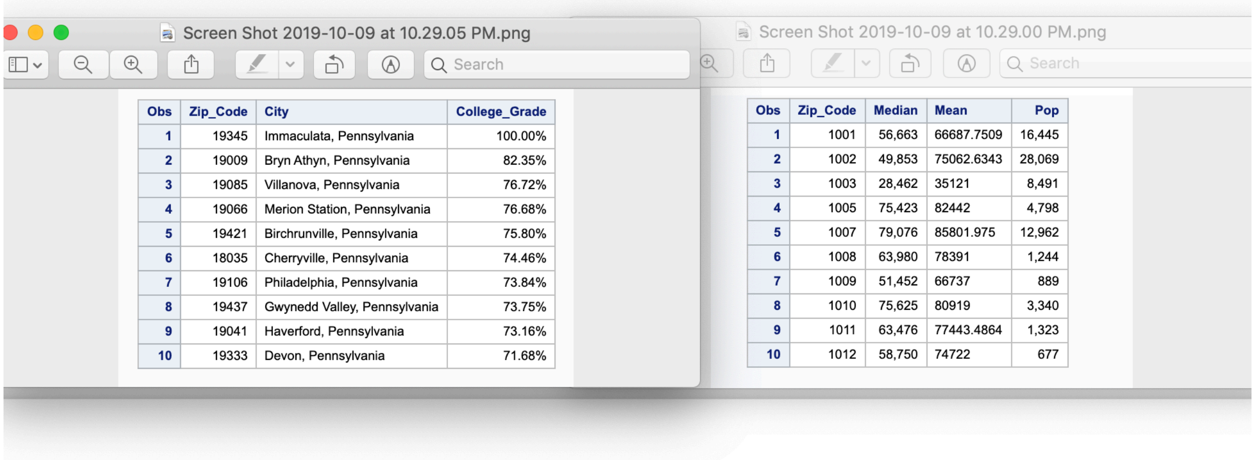
MAT7500
Dr. Posner
Haichuan Du
Oct 10, 2019

Project#1

Complete the following tasks:

1) Read in each file into SAS.

```
* 1) Read in each file into SAS and rename varibale in advance;
     proc import
          datafile='/folders/myfolders/project-1/Median Income by Zip
     Code in US.xlsx'
          dbms=xlsx replace out=data1(rename=(Zip=Zip_Code));
          sheet="nation"; getnames=yes;
     proc import
          datafile='/folders/myfolders/project-1/PA College Graduation
     By Zip Code.xlsx'
          dbms=xlsx replace
     out=data2(rename=(__College_Grad_=College_Grade));
          sheet="Sheet1"; getnames=yes;
     proc print data=data1 (obs=10);
     proc print data=data2 (obs=10);
     proc sort data=data1; by Zip_Code;
     proc sort data=data2; by Zip_Code;
```


Screen Shot 2019-10-09 at 10.29.05 PM.png
Search
Screen Shot 2019-10-09 at 10.29.00 PM.png
Search

Obs	Zip_Code	City	College_Grade
1	19345	Immaculata, Pennsylvania	100.00%
2	19009	Bryn Athyn, Pennsylvania	82.35%
3	19085	Villanova, Pennsylvania	76.72%
4	19066	Merion Station, Pennsylvania	76.68%
5	19421	Birchrunville, Pennsylvania	75.80%
6	18035	Cherryville, Pennsylvania	74.46%
7	19106	Philadelphia, Pennsylvania	73.84%
8	19437	Gwynedd Valley, Pennsylvania	73.75%
9	19041	Haverford, Pennsylvania	73.16%
10	19333	Devon, Pennsylvania	71.68%

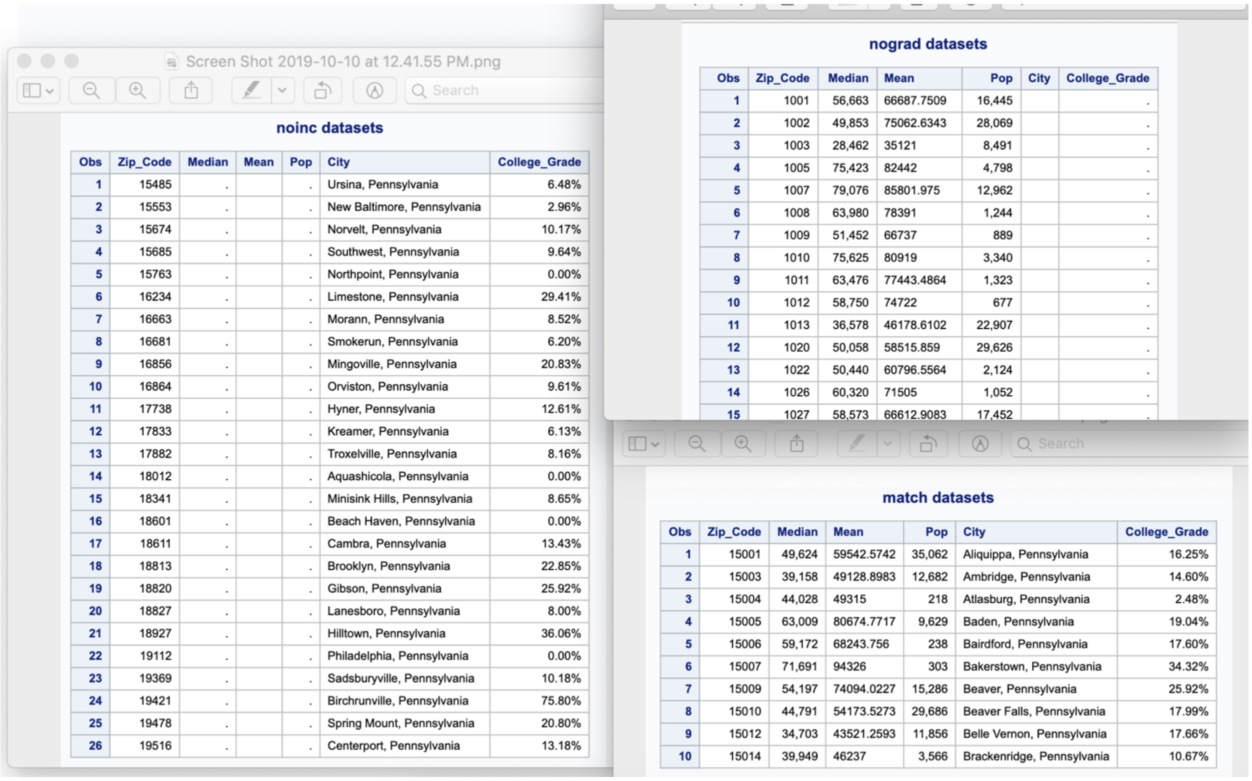
Obs	Zip_Code	Median	Mean	Pop
1	1001	56,663	66687.7509	16,445
2	1002	49,853	75062.6343	28,069
3	1003	28,462	35121	8,491
4	1005	75,423	82442	4,798
5	1007	79,076	85801.975	12,962
6	1008	63,980	78391	1,244
7	1009	51,452	66737	889
8	1010	75,625	80919	3,340
9	1011	63,476	77443.4864	1,323
10	1012	58,750	74722	677

2) Merge the two files together by zip code. When doing so, create three output datasets – one called *match* that contains only those zip codes that were in both files, a second one called *noinc* that includes zip codes that don't have income data, and a third one called *nograd* that includes zip codes that don't have college graduation data.

```
*2)Merge the two files together by zip code and create match datasets;
data MergeData match;
     merge data1(in=a) data2(in=b); by Zip_Code; output MergeData;
     if a and b then output match;
proc print data=match (obs=10);    /* match that contains only those 
zip codes that were in both files*/
proc print data=MergeData (obs=10);

* Create noinc datasets that includes zip codes that don’t have income 
data.;
data noinc;
     set MergeData;
     where Mean=' ';
proc print data=noinc;
run;

* Create nograd datasets that includes zip codes that don’t have 
college graduation data.;
data nograd;
     set MergeData;
     where College_Grade=.;
proc print data=nograd (obs=20);
run;
```



noinc datasets

Obs	Zip_Code	Median	Mean	Pop	City	College_Grade
1	15485	.		.	Ursina, Pennsylvania	6.48%
2	15553	.		.	New Baltimore, Pennsylvania	2.96%
3	15674	.		.	Norvelt, Pennsylvania	10.17%
4	15685	.		.	Southwest, Pennsylvania	9.64%
5	15763	.		.	Northpoint, Pennsylvania	0.00%
6	16234	.		.	Limestone, Pennsylvania	29.41%
7	16663	.		.	Morann, Pennsylvania	8.52%
8	16681	.		.	Smokerun, Pennsylvania	6.20%
9	16856	.		.	Mingoville, Pennsylvania	20.83%
10	16864	.		.	Orviston, Pennsylvania	9.61%
11	17738	.		.	Hyner, Pennsylvania	12.61%
12	17833	.		.	Kreamer, Pennsylvania	6.13%
13	17882	.		.	Troxelville, Pennsylvania	8.16%
14	18012	.		.	Aquashicola, Pennsylvania	0.00%
15	18341	.		.	Minisink Hills, Pennsylvania	8.65%
16	18601	.		.	Beach Haven, Pennsylvania	0.00%
17	18611	.		.	Cambra, Pennsylvania	13.43%
18	18813	.		.	Brooklyn, Pennsylvania	22.85%
19	18820	.		.	Gibson, Pennsylvania	25.92%
20	18827	.		.	Lanesboro, Pennsylvania	8.00%
21	18927	.		.	Hilltown, Pennsylvania	36.06%
22	19112	.		.	Philadelphia, Pennsylvania	0.00%
23	19369	.		.	Sadsburyville, Pennsylvania	10.18%
24	19421	.		.	Birchrunville, Pennsylvania	75.80%
25	19478	.		.	Spring Mount, Pennsylvania	20.80%
26	19516	.		.	Centerport, Pennsylvania	13.18%

nograd datasets

Obs	Zip_Code	Median	Mean	Pop	City	College_Grade
1	1001	56,663	66687.7509	16,445		.
2	1002	49,853	75062.6343	28,069		.
3	1003	28,462	35121	8,491		.
4	1005	75,423	82442	4,798		.
5	1007	79,076	85801.975	12,962		.
6	1008	63,980	78391	1,244		.
7	1009	51,452	66737	889		.
8	1010	75,625	80919	3,340		.
9	1011	63,476	77443.4864	1,323		.
10	1012	58,750	74722	677		.
11	1013	36,578	46178.6102	22,907		.
12	1020	50,058	58515.859	29,626		.
13	1022	50,440	60796.5564	2,124		.
14	1026	60,320	71505	1,052		.
15	1027	58,573	66612.9083	17,452		.

match datasets

Obs	Zip_Code	Median	Mean	Pop	City	College_Grade
1	15001	49,624	59542.5742	35,062	Aliquippa, Pennsylvania	16.25%
2	15003	39,158	49128.8983	12,682	Ambridge, Pennsylvania	14.60%
3	15004	44,028	49315	218	Atlasburg, Pennsylvania	2.48%
4	15005	63,009	80674.7717	9,629	Baden, Pennsylvania	19.04%
5	15006	59,172	68243.756	238	Bairdford, Pennsylvania	17.60%
6	15007	71,691	94326	303	Bakerstown, Pennsylvania	34.32%
7	15009	54,197	74094.0227	15,286	Beaver, Pennsylvania	25.92%
8	15010	44,791	54173.5273	29,686	Beaver Falls, Pennsylvania	17.99%
9	15012	34,703	43521.2593	11,856	Belle Vernon, Pennsylvania	17.66%
10	15014	39,949	46237	3,566	Brackenridge, Pennsylvania	10.67%

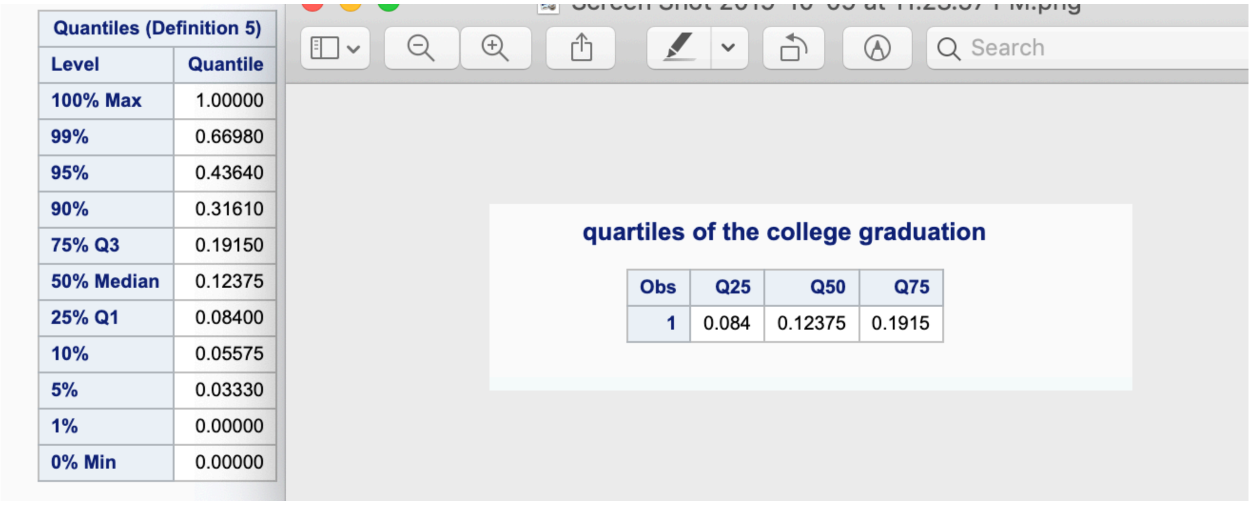
3) How many zip codes are in each of the three datasets that you created above? For the *noinc* and *nograd* datasets, speculate as to why these values did not match. You might need to look at the zip codes themselves to figure this out.

Common: match datasets have 1690 zip codes, noinc datasets have 26 zip codes, nongrad datasets have 30944 zip code. The reason why those Zip codes did not match because some Zip codes in the file of PA College Graduation were not included into the Median Income by Zip codes in US. For those that don't have income data, they also don't have population data. Because these places are remote and have very small populations. For large datasets with entire US, their effect is very small. But for a relatively small sample (PA), we have to be more precise and include these data. Therefore, that's the reason why these values did not match.

4) Using only the *match* dataset...

a) Calculate the quartiles of the college graduation rate variable.

```
*4) Using only the match dataset;
data UsingMatch;
*a) Calculate the quartiles of the college graduation rate
variable;
      set match;
proc univariate data=match noprint;
        var College_Grade;
        output out=quartiles pctlpre=Q pctlpts=25 50 75;
        data _null_;
        set quartiles;
        call symput('Q1',Q25);
        call symput('Q2',Q50);
        call symput('Q3',Q75);
        run;
proc print data=quartiles;
title 'quartiles of the college graduation';
run;
```



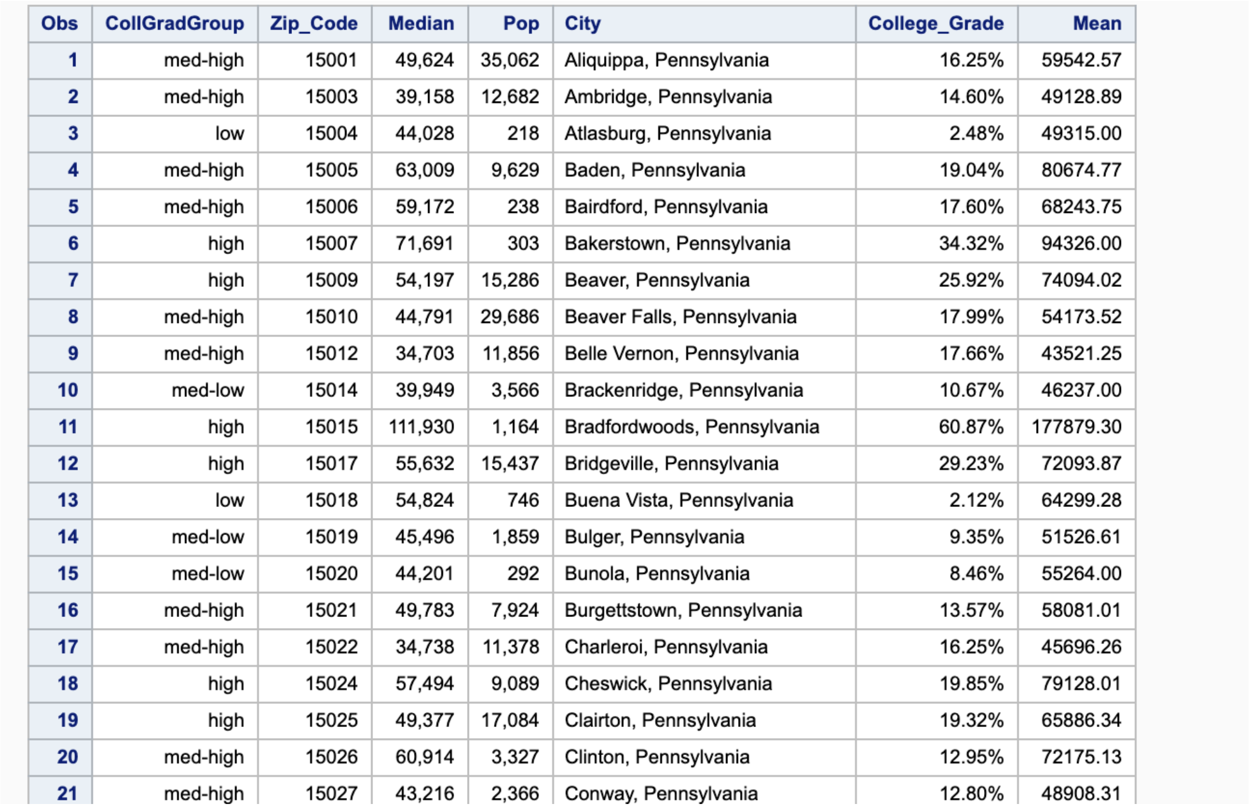
Quantiles (Definition 5)	
Level	Quantile
100% Max	1.00000
99%	0.66980
95%	0.43640
90%	0.31610
75% Q3	0.19150
50% Median	0.12375
25% Q1	0.08400
10%	0.05575
5%	0.03330
1%	0.00000
0% Min	0.00000

quartiles of the college graduation

Obs	Q25	Q50	Q75
1	0.084	0.12375	0.1915

b) Create a new variable called *CollGradGroup* which takes on the following values:

```
* b) Create a new variable called CollGradGroup and format the
CollGradGRoup variable to name the group “hight”, “med-high”,
“med-low”, and “low”, respectively.
proc format;
     value level 1='low' 2='med-low' 3='med-high' 4='high';
data mydata_b;
   label CollGradGroup = 'CollGradGroup';
   set match;
   new=input(Mean, 8.);     /* here we change Mean from Char to Num*/
   drop Mean;
   rename new=Mean;/*we drop the original char Mean and keep new one*/
     format CollGradGroup level.;
     if College_Grade > &Q3 then CollGradGroup=4;
     if &Q2 < College_Grade < &Q3 then CollGradGroup=3;
     if &Q1 < College_Grade < &Q2 then CollGradGroup=2;
     if College_Grade < &Q1 then CollGradGroup=1;
proc print data = mydata_b (obs=100);
```



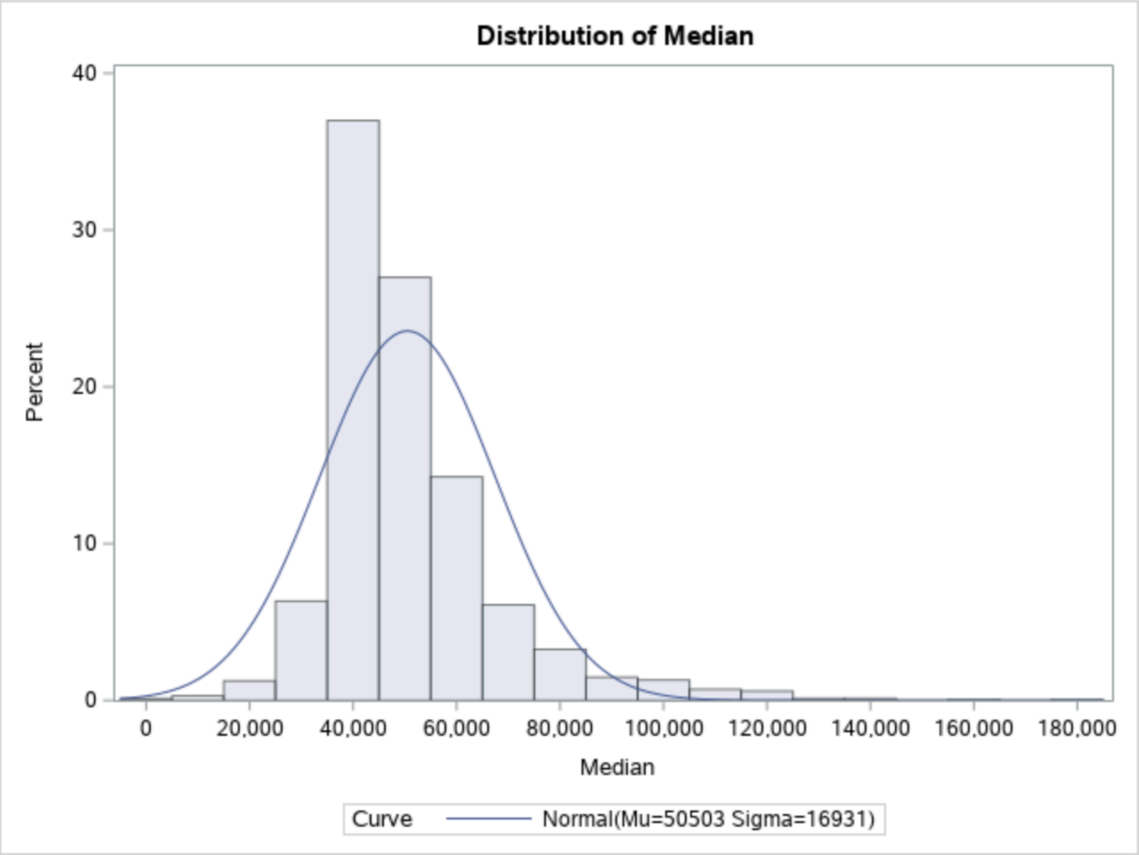
Obs	CollGradGroup	Zip_Code	Median	Pop	City	College_Grade	Mean
1	med-high	15001	49,624	35,062	Aliquippa, Pennsylvania	16.25%	59542.57
2	med-high	15003	39,158	12,682	Ambridge, Pennsylvania	14.60%	49128.89
3	low	15004	44,028	218	Atlasburg, Pennsylvania	2.48%	49315.00
4	med-high	15005	63,009	9,629	Baden, Pennsylvania	19.04%	80674.77
5	med-high	15006	59,172	238	Bairdford, Pennsylvania	17.60%	68243.75
6	high	15007	71,691	303	Bakerstown, Pennsylvania	34.32%	94326.00
7	high	15009	54,197	15,286	Beaver, Pennsylvania	25.92%	74094.02
8	med-high	15010	44,791	29,686	Beaver Falls, Pennsylvania	17.99%	54173.52
9	med-high	15012	34,703	11,856	Belle Vernon, Pennsylvania	17.66%	43521.25
10	med-low	15014	39,949	3,566	Brackenridge, Pennsylvania	10.67%	46237.00
11	high	15015	111,930	1,164	Bradfordwoods, Pennsylvania	60.87%	177879.30
12	high	15017	55,632	15,437	Bridgeville, Pennsylvania	29.23%	72093.87
13	low	15018	54,824	746	Buena Vista, Pennsylvania	2.12%	64299.28
14	med-low	15019	45,496	1,859	Bulger, Pennsylvania	9.35%	51526.61
15	med-low	15020	44,201	292	Bunola, Pennsylvania	8.46%	55264.00
16	med-high	15021	49,783	7,924	Burgettstown, Pennsylvania	13.57%	58081.01
17	med-high	15022	34,738	11,378	Charleroi, Pennsylvania	16.25%	45696.26
18	high	15024	57,494	9,089	Cheswick, Pennsylvania	19.85%	79128.01
19	high	15025	49,377	17,084	Clairton, Pennsylvania	19.32%	65886.34
20	med-high	15026	60,914	3,327	Clinton, Pennsylvania	12.95%	72175.13
21	med-high	15027	43,216	2,366	Conway, Pennsylvania	12.80%	48908.31

c) Create an appropriate graph of the median income variable. Comment on the graph.

```
*c) Create an appropriate graph of the median income variable. Comment
on the graph;
proc univariate data=mydata_b noprint;
      var Median;
      HISTOGRAM Median/normal(noprint);
      title 'The Median Income';
run;
```

The Median Income

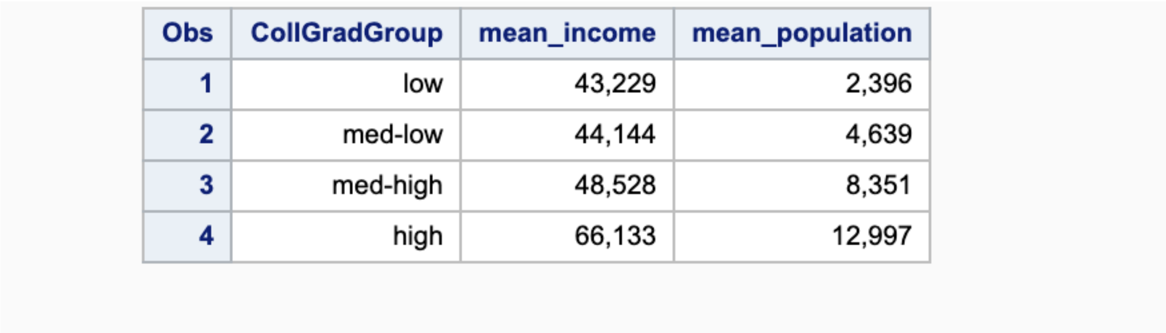
The UNIVARIATE Procedure



Comment: Pennsylvania Household Income According to the graph survey, the mean of the median household income for Pennsylvania was around $5000, and the mode is about $4200. which is less than the median annual income of $60,336 across the entire United States. Overall, the shape of this graph is a little skew right.

d) Create a SAS dataset that includes the mean value of the median income variable as well as the mean value of the population variable separately for each *CollGradGroup* value.

```
*d) Create a SAS dataset that includes the mean value of the median
income variable as well as;
*the mean value of the population variable separately for each
CollGradGroup value.;
data mydata_d;
     set mydata_b;
proc means data=mydata_b nway noprint;
     class CollGradGroup;
     var Median Pop;
     output out=abc(drop=_type_ _freq_) mean(Median)=mean_income
mean(Pop)=mean_population;
proc print data=abc;
run;
```



Obs	CollGradGroup	mean_income	mean_population
1	low	43,229	2,396
2	med-low	44,144	4,639
3	med-high	48,528	8,351
4	high	66,133	12,997

Comment: According to the table survey, the mean value of the median income is 43229$ for low college grade group, 44144$ for med-low college grade group, 48528$ for med-high college grade group, 48528$ for med-high college grade group, 66133$ for high college grade group. The higher college grade group has higher mean income. The college grade is proportional to the income. And the higher the score, the more people there are.

Appendix

```
/**************
 MAT 7500     *
 Dr. Posner   *
 Haichuan Du  *
 Oct.3 2019   *
 Project-1    *
 *************/

* 1) Read in each file into SAS.;
proc import
     datafile='/folders/myfolders/project-1/Median Income by Zip Code
in US.xlsx'
     dbms=xlsx replace out=data1(rename=(Zip=Zip_Code));
     sheet="nation"; getnames=yes;
proc import
     datafile='/folders/myfolders/project-1/PA College Graduation By
Zip Code.xlsx'
     dbms=xlsx replace
out=data2(rename=(__College_Grad_=College_Grade));
     sheet="Sheet1"; getnames=yes;
proc print data=data1 (obs=10);
proc print data=data2 (obs=10);
proc sort data=data1; by Zip_Code;
proc sort data=data2; by Zip_Code;

* 2) Merge the two files together by zip code.;
data MergeData match;
     merge data1(in=a) data2(in=b); by Zip_Code;  output MergeData;
     if a and b then output match;
proc print data=match ;    /* match that contains only those zip codes
that were in both files*/
title 'match datasets';
proc print data=MergeData ;
title 'MergeData datasets';

* a second one called noinc that includes zip codes that don’t have
income data.;
data noinc;
     set MergeData;
     where  Mean=' ';
```

```
proc print data=noinc;
title 'noinc datasets';
run;

* a third one called nograd that includes zip codes that don’t have
college graduation data.;
data nograd;
      set MergeData;
      where College_Grade=.;
proc print data=nograd ;
title 'nograd datasets';
run;



* 4) Using only the match dataset;
data UsingMatch;
*a) Calculate the quartiles of the college graduation rate variable;
      set match;
proc univariate data=match;
        var College_Grade;
        output out=quartiles pctlpre=Q pctlpts=25 50 75;
        data _null_;
        set quartiles;
        call symput('Q1',Q25);
        call symput('Q2',Q50);
        call symput('Q3',Q75);
        run;
proc print data=quartiles;
title 'quartiles of the college graduation';
run;

* b) Create a new variable called CollGradGroup which takes on the
following values;
proc format;
      value level 1='low' 2='med-low' 3='med-high' 4='high';
data mydata_b;
   label CollGradGroup = 'CollGradGroup';
   set match;
   new = input(Mean, 8.);     /* here we chang Mean from Char to Num
*/
   drop Mean;
   rename new=Mean;
      format CollGradGroup level.;
      if College_Grade > &Q3 then CollGradGroup=4;
      if &Q2 < College_Grade < &Q3 then CollGradGroup=3;
      if &Q1 < College_Grade < &Q2 then CollGradGroup=2;
      if College_Grade < &Q1 then CollGradGroup=1;
proc print data = mydata_b (obs=100);

*c) Create an appropriate graph of the median income variable. Comment
on the graph;
proc univariate data=mydata_b noprint;
      var Median;
      HISTOGRAM Median/normal(noprint);
      title 'The Median Income';
run;

*d) Create a SAS dataset that includes the mean value of the median
income variable as well as;
*the mean value of the population variable separately for each
CollGradGroup value.;
data mydata_d;
      set mydata_b;
proc means data=mydata_b nway noprint;
      class CollGradGroup;
      var Median Pop;
      output out=abc(drop=_type_ _freq_) mean(Median)=mean_income
mean(Pop)=mean_population;
proc print data=abc;
run;
```