IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

March 1,3 2016

Week 8 Presentation

Week 8 Topic: Agenda

- Questions on Clustering
- Profiling
- Tax Strategy

Week 8 Topic:
Any Questions on Week 7 Assignment?

- Perform clustering for k=20 as in class.
- Perform clustering for k=15.
- Then, in the submission text:
  - State top 3 differences between the k=20 and k=15 runs.
  - State your next steps in the clustering analysis (not profiling steps). These will be steps that address what you will do next with the 15 clusters:
    - Will you proceed with k=15 or k=20? Will you run for a different k value? State reasons why?
    - Will you perform sub-clustering on the clustered results? State reasons why? If yes, which clusters will you additionally segment? Which clusters will you keep as is?
- What variables will you consider in profiling the clusters? Just list the variables chosen.

Week 8 Topic: FASTCLUS vs CLUSTER

	# of Clusters	# of Observations	Takes Matrix as input	Model	Sensitivity
PROC CLUSTER	*small k*	*small data set*	Yes	Hierarchical Clustering	# of Observations, # of Variables
PROC FASTCLUS	20~100	>100	No	K-means Clustering	Outliers, Order of Observations

Notes:

- The time required by PROC FASTCLUS is roughly proportional to the number of observations, whereas the time required by PROC CLUSTER with most methods varies with the square or cube of the number of observations. Therefore, you can use PROC FASTCLUS with much larger data sets than PROC CLUSTER.
- If you want to hierarchically cluster a data set that is too large to use with PROC CLUSTER directly, you can have PROC FASTCLUS produce, for example, 50 clusters, and let PROC CLUSTER analyze these 50 clusters instead of the entire data set.

Week 8 Topic: FASTCLUS vs CLUSTER (cont.)

From http://analytics.ncsu.edu/sesug/2010/SDA10.Reiss.pdf

Table 2 summarizes some of the advantages, disadvantages, and overall differences of both clustering methods as applied to this project.

Table 2: PROC FASTCLUS vs. PROC CLUSTER

	PROC FASTCLUS	PROC CLUSTER
Method	distance-based disjoint	hierarchical
Steps	• Arbitrarily choose *k* observations as the "seeds" • Assign all other observations to their closest "seed • Update the cluster mean and re-define the center	• Each observation begins in a cluster by itself • The two closest clusters are merged to form a new cluster • Merging continues until only one cluster is left
Advantages	• Faster • Can handle large datasets	• Produces a tree visualization • Provides more process details
Disadvantages	• Must specify number of clusters • Different seeds = different results	• Slower • Not suitable for larger datasets • Missing value imputation necessary

Week 8 Topic: Analysis Considerations I

- **# of Observations:** The FASTCLUS procedure is intended for use with large data sets, with 100 or more observations. With small data sets, the results can be highly sensitive to the order of the observations in the data set.

- **Outliers:** Most cluster solutions are affected heavily by presence of outliers and/or observations that are just too different from the others. These observations can also indicate potential business opportunities. You could subset the data using the WHERE statement in your DATA step. The initialization method used by the FASTCLUS procedure makes it sensitive to outliers. PROC FASTCLUS can be an effective procedure for detecting outliers because outliers often appear as clusters with only one member.

- **Standardization:** Before using PROC FASTCLUS, decide whether your variables should be standardized in some way, since variables with large variances tend to have more effect on the resulting clusters than those with small variances. If all variables are measured in the same units, standardization might not be necessary. Otherwise, some form of standardization is strongly recommended. The STANDARD procedure can standardize all variables to mean zero and variance one. PROC FASTCLUS uses algorithms that place a larger influence on variables with larger variance, so it might be necessary to standardize the variables before performing the cluster analysis.

Week 8 Topic: Analysis Considerations II

- **Convergence:** If PROC FASTCLUS runs to complete convergence, the final cluster seeds will equal the cluster means or cluster centers. If PROC FASTCLUS terminates before complete convergence, which often happens with the default settings, the final cluster seeds might not equal the cluster means or cluster centers. If you want complete convergence, specify CONVERGE=0 and a large value for the MAXITER= option.

- PROC FASTCLUS always selects the first complete (no missing values) observation as the first seed. The next complete observation that is separated from the first seed by at least the distance specified in the RADIUS= option becomes the second seed. Later observations are selected as new seeds if they are separated from all previous seeds by at least the radius, as long as the maximum number of seeds is not exceeded.

Week 8 Topic: Analysis Considerations III - CCC

- **CCC:** The best way to use the CCC is to plot its value against the number of clusters, ranging from one cluster up to about one-tenth the number of observations. The CCC may not behave well if the average number of observations per cluster is less than ten. The following guidelines should be used for interpreting the CCC:
  - Peaks on the plot with the CCC greater than 2 or 3 indicate good clusterings.
  - Peaks with the CCC between 0 and 2 indicate possible clusters but should be interpreted cautiously.
  - There may be several peaks if the data has a hierarchical structure.
  - Very distinct nonhierarchical spherical clusters usually show a sharp rise before the peak followed by a gradual decline.
  - Very distinct nonhierarchical elliptical clusters often show a sharp rise to the correct number of clusters followed by a further gradual increase and eventually a gradual decline.
  - If all values of the CCC are negative and decreasing for two or more clusters, the distribution is probably unimodal or long-tailed.
  - Very negative values of the CCC, say, -30, may be due to outliers. Outliers generally should be removed before clustering and their removal documented.
- If the CCC increases continually as the number of clusters increases, the distribution may be grainy or the data may have been excessively rounded or recorded with just a few digits. A final and very important warning: neither the CCC nor R2 is an appropriate criterion for clusters that are highly elongated or irregularly shaped. If you do not have prior substantive reasons for expecting compact clusters, use a nonparametric clustering method such as Wong and Lane's (1983) rather than Ward's method or k-means clustering.

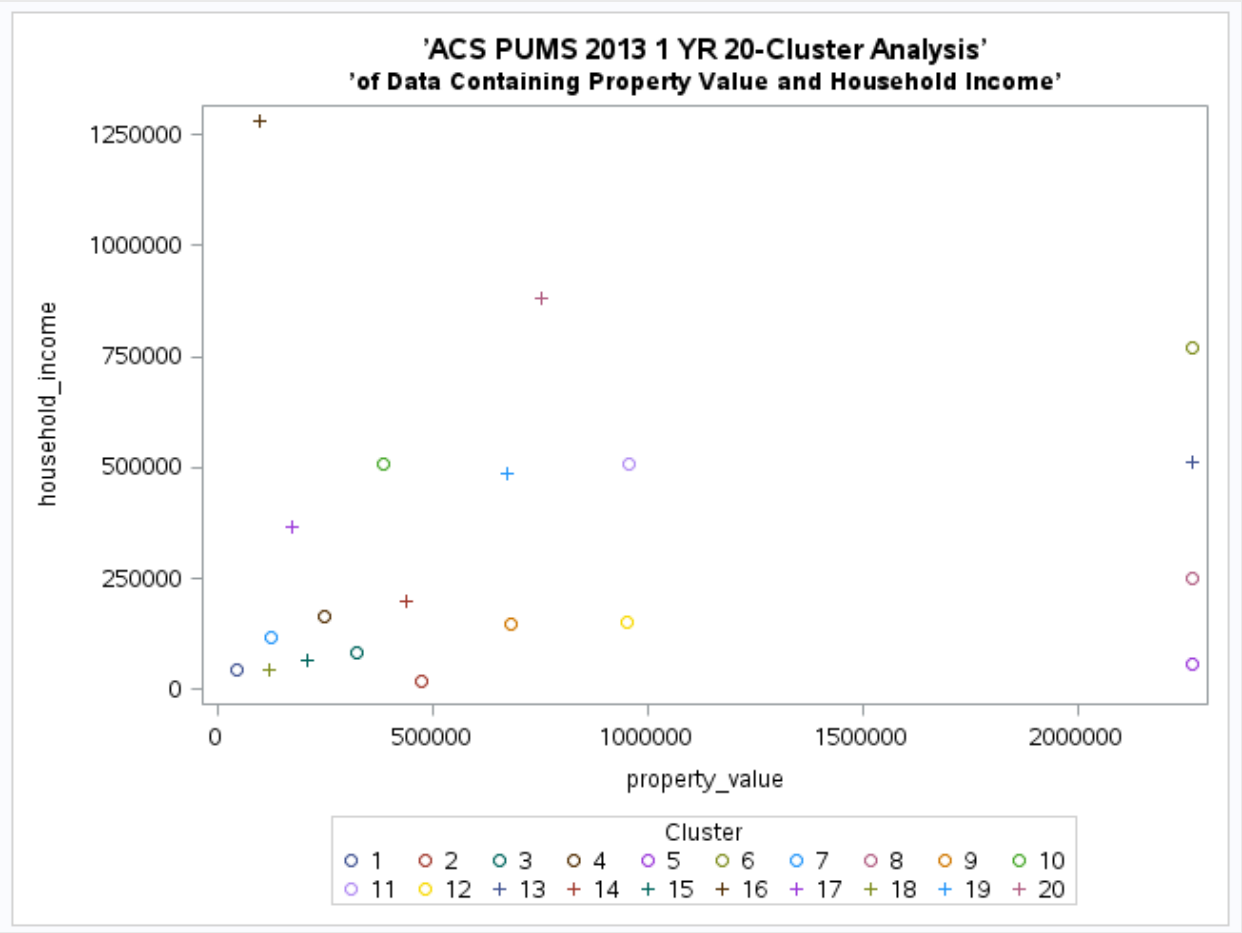
Week 8 Topic:
20 Clusters: PROC FASTCLUS

We choose to run with the following options using HINCP and VALP as variables. Note that we take out the LIST option as it will produce a larger than 4GB output of observations:

```
/*20 Cluster run of FASTCLUS*/

libname census "/folders/myfolders/census";

proc fastclus data=census.psam_h17_subset1
radius=0 replace=full
converge=0 maxiter=200
maxclusters=20
OUTSTAT=census.psam_h17_subset1_20clusters_stat
OUT=census.psam_h17_subset1_20clusters
distance;
id SERIALNO;
var VALP HINCP;
run;

proc sgplot;
scatter y=HINCP x=VALP  / group=cluster;
title 'ACS PUMS 2013 1 YR 20-Cluster Analysis';
title2 'of Data Containing Property Value and Household Income';
run;
```

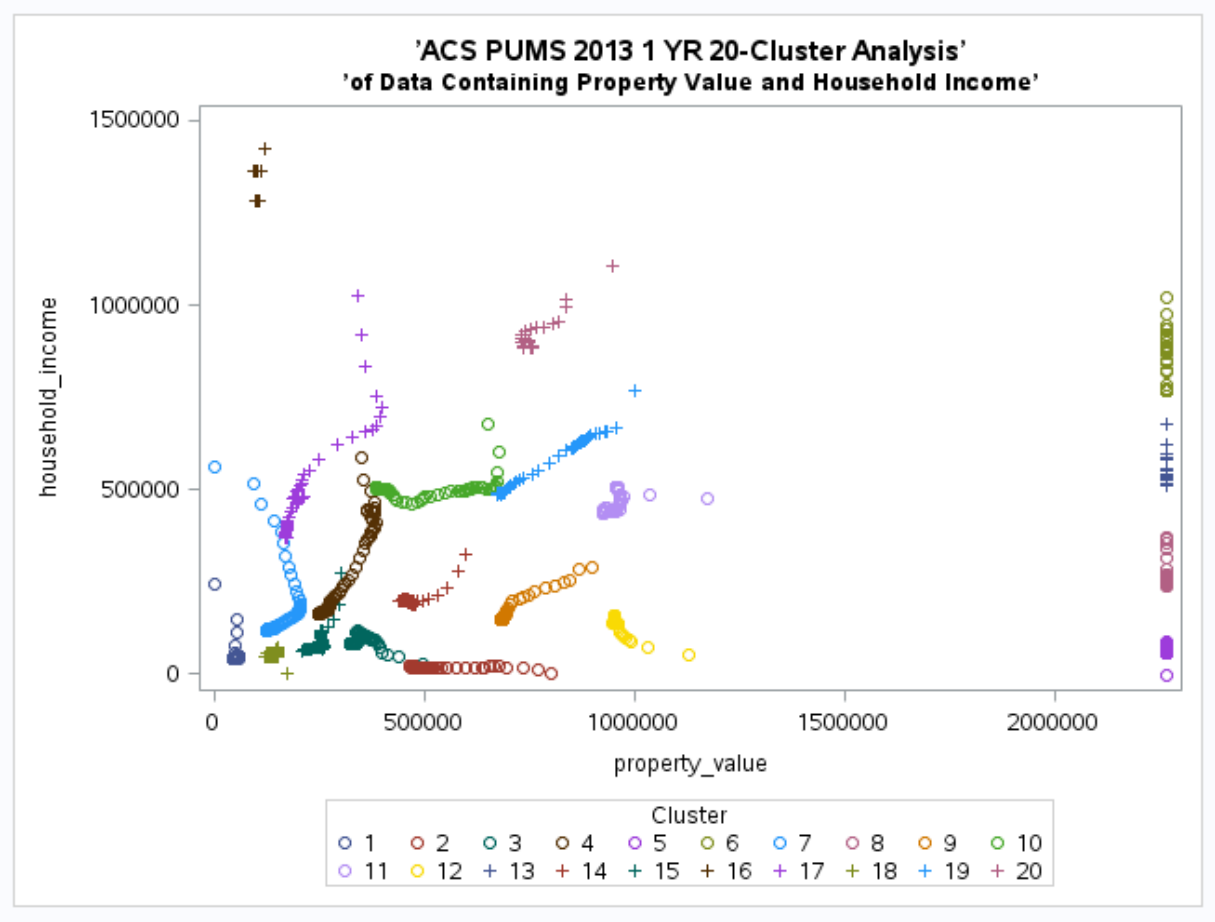
Week 8 Topic:
20 Clusters: Initial Seeds

Add `OUTSEED=census.psam_h18_subset1_20clusters_seed` to see initial seeding:



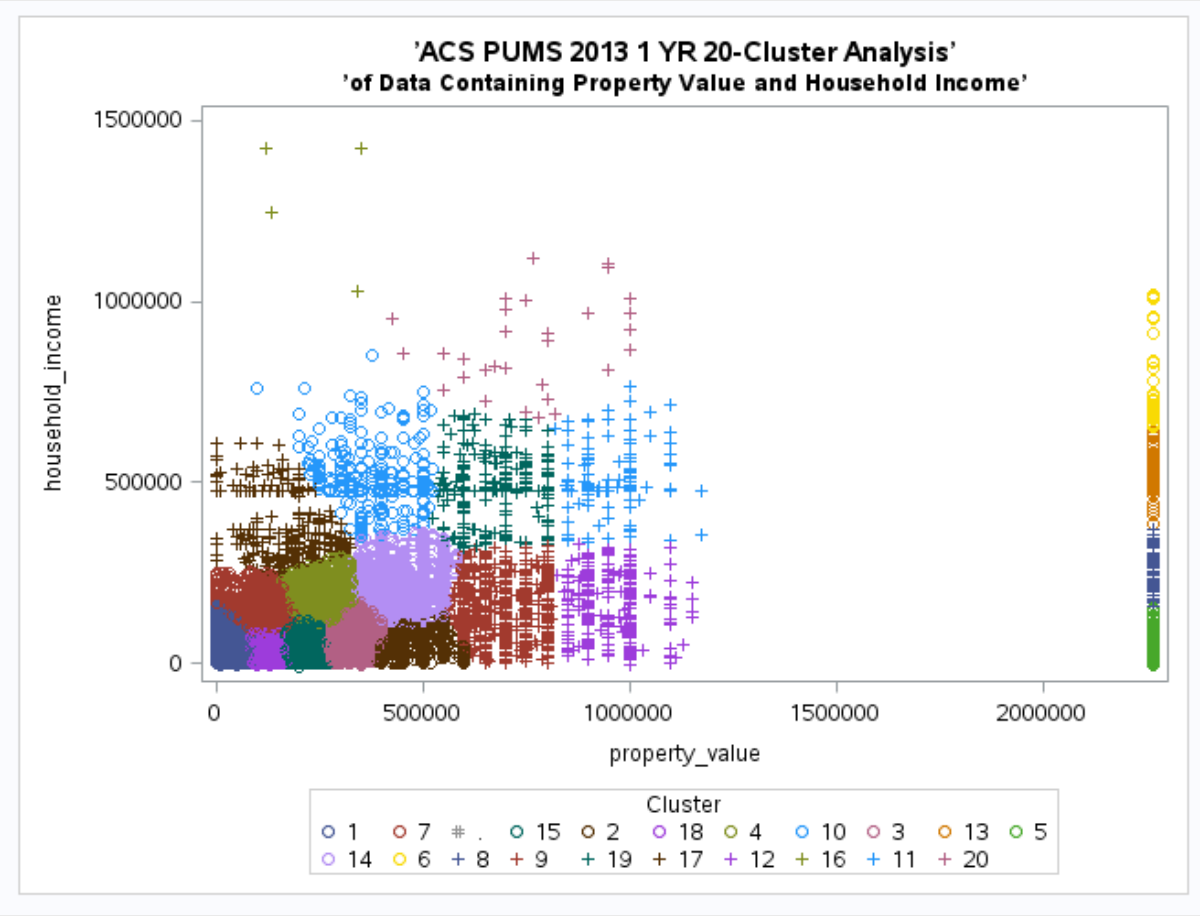
Week 8 Topic:
20 Clusters: Seed Iteration

Add `OUTITER OUTSEED=census.psam_h18_subset1_20clusters_seed` to see initial seeding and seed iteration to convergence:



Week 8 Topic:
20 Clusters: Final

Final 20 clusters, observe the grid like clusters along the axis:



Week 8 Topic:
20 Clusters: Analysis Factors

- **Model Selection:** Due to the data set size and number of clusters anticipated, we choose FASTCLUS for the cluster analysis

- **Variables Selection:** We choose HINCP and VALP numeric variables with similar scale. Since both variable scale is similar, no standardization is required. As discussed last week, we ignore MRGP (the other numeric variable contender) due to its incompleteness and scale differences compared to VALP.

- **Trimming of data:** Looking at the initial run of clustering, we see that the clusters follow a grid like pattern along the x and y axis. If we were to perform segmentation without using models, we may segment similarly. Hence, we accept the results. Also, since we used FASTCLUS:
  - Outliers are grouped into clusters of its own
  - Negative HINCP values can be treated at the cluster level
  - Missing values are ignored by the program run or treated at the cluster level
  - Alternative would be to trim outliers, missing, and negative variable values before the cluster run. However, we go in with the assumption that these did not impact the cluster analysis.

Week 8 Topic:
20 Clusters: TAXP Frequency Distribution

We examine TAXP in more detail, the main variable we will use to develop the tax increase strategy for the City. We use PROC FREQ to understand how tax groupings distribute amongst the various clusters:

```
/*Plot TAXP distribution across the clusters*/

proc freq data=census.psam_h17_subset1_20clusters;
tables TAXP*Cluster;
run;
```

Table of TAXP by CLUSTER

Frequency
Percent
Row Pct
Col Pct

TAXP(property_tax)	CLUSTER(Cluster) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
01	844 2.33 82.18 10.07	4 0.01 0.39 0.55	12 0.03 1.17 0.41	8 0.02 0.78 0.30	0 0.00 0.00 0.00	0 0.00 0.00 0.00	56 0.15 5.45 1.27	0 0.00 0.00 0.00	2 0.01 0.19 0.30	1 0.00 0.10 0.37	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 0.01 0.29 0.20	23 0.06 2.24 0.39	0 0.00 0.00 0.00	10 0.03 0.97 3.69	64 0.18 6.23 0.85	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1027 2.84
02	167 0.46 90.27 1.99	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 0.01 1.62 0.07	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.00 0.54 0.39	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 0.01 2.16 0.07	0 0.00 0.00 0.00	1 0.00 0.54 0.37	9 0.02 4.86 0.12	0 0.00 0.00 0.00	0 0.00 0.00 0.00	185 0.51
03	307 0.85 91.10 3.66	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	7 0.02 2.08 0.16	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	7 0.02 2.08 0.12	0 0.00 0.00 0.00	0 0.00 0.00 0.00	16 0.04 4.75 0.21	0 0.00 0.00 0.00	0 0.00 0.00 0.00	337 0.93
04	251 0.69 85.67 2.99	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	9 0.02 3.07 0.20	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	1 0.00 0.34 0.07	5 0.01 1.71 0.09	0 0.00 0.00 0.00	0 0.00 0.00 0.00	27 0.07 9.22 0.36	0 0.00 0.00 0.00	0 0.00 0.00 0.00	293 0.81
05	202 0.56 88.99 2.41	0 0.00 0.00 0.00	1 0.00 0.44 0.03	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	7 0.02 3.08 0.16	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	4 0.01 1.76 0.07	0 0.00 0.00 0.00	0 0.00 0.00 0.00	13 0.04 5.73 0.17	0 0.00 0.00 0.00	0 0.00 0.00 0.00	227 0.63
06	265 0.73 84.66 3.16	1 0.00 0.32 0.14	2 0.01 0.64 0.07	1 0.00 0.32 0.04	0 0.00 0.00 0.00	0 0.00 0.00 0.00	12 0.03 3.83 0.27	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	0 0.00 0.00 0.00	3 0.01 0.96 0.05	0 0.00 0.00 0.00	4 0.01 1.28 1.48	25 0.07 7.99 0.33	0 0.00 0.00 0.00	0 0.00 0.00 0.00	313 0.87

Week 8 Topic: Profiling Considerations

- Two types of questions are often asked in profiling:
  - How do the members in one cluster differ from the members in another cluster with regard to the base variables?
  - How do the members in a particular cluster differ from all the members in the data with regard to the base variables?
- Profiling involves examining the distinguishing characteristics of each cluster's profile and identifying substantial differences between clusters.
  - For numeric variables, this involves
    - comparing mean of each variable across clusters
    - comparing mean of each variable in a cluster with the mean for the same variable for the entire data (population)
    - comparing distribution of each variable in a cluster with the distribution of the same variable for the entire data (population).
- Cluster solutions failing to show substantial variation indicates that other cluster solutions need to be examined.

Week 8 Topic: Additional Variables for Profiling

- TAXP: Property taxes (yearly amount)
  - bb .N/A (GQ/vacant/not owned or being bought)
  - 01 .None
  - 02 .$ 1 - $ 49
  - 03 .$ 50 - $ 99
  - 04 .$ 100 - $ 149
  - …
- WIF: Workers in family during the past 12 months
  - b .N/A (GQ/vacant/non-family household)
  - 0 .No workers
  - 1 .1 worker
  - 2 .2 workers
  - 3 .3 or more workers in family
- SMX: Second or junior mortgage or home equity loan status
  - b .N/A (GQ/vacant/not owned or being bought)
  - 1 .Yes, a second mortgage
  - 2 .Yes, a home equity loan
  - 3 .No
  - 4 .Both a second mortgage and a home equity loan

Week 8 Topic: Additional Variables for Profiling (cont.)

- MV: When moved into this house or apartment
  - b .N/A (GQ/vacant)
  - 1 .12 months or less
  - 2 .13 to 23 months
  - 3 .2 to 4 years
  - 4 .5 to 9 years
  - 5 .10 to 19 years
  - 6 .20 to 29 years
  - 7 .30 years or more
- WORKSTAT: Work status of householder or spouse in family households
  - bb .N/A (GQ/not a family household)
  - 01 .Husband and wife both in labor force, both employed or in
  - .Armed Forces
  - 02 .Husband and wife both in labor force, husband employed or in
  - .Armed Forces, wife unemployed
  - 03 .Husband in labor force and wife not in labor force, husband
  - .employed or in Armed Forces
  - 04 .Husband and wife both in labor force, husband unemployed, wife
  - .employed or in Armed Forces
  - …

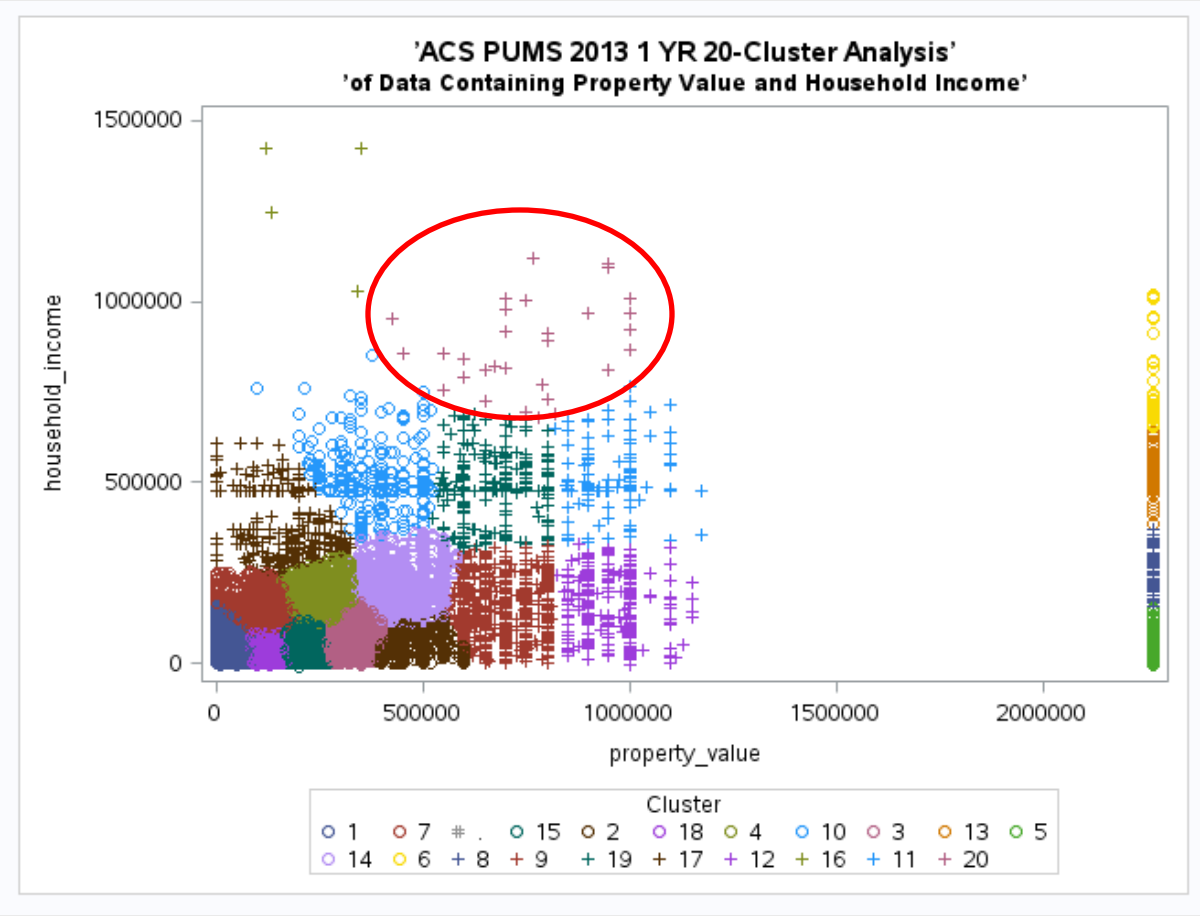
Week 8 Topic:
JOIN tables to get additional variables

We use PROC SQL to join the clustered results table with the original data file to get the additional variables for profiling:

```
/*Join additional variables from psam_h17 table for profiling*/

proc sql;
title '20 Cluster aggregated information for profiling';
create table census.psam_h17_subset1_20clusters_agg as

select a.SERIALNO, a.CLUSTER, a.DISTANCE, a.HINCP, a.VALP, a.TAXP,
 b.MV, b.SMX, b.WORKSTAT, b.WIF
from census.psam_h17_subset1_20clusters as a left join census.psam_h17 as b
on a.SERIALNO=b.SERIALNO;
run;
```

Week 8 Topic: Cluster #20

We examine and profile cluster #20 in more detail



Week 8 Topic:
Cluster #20: Create table

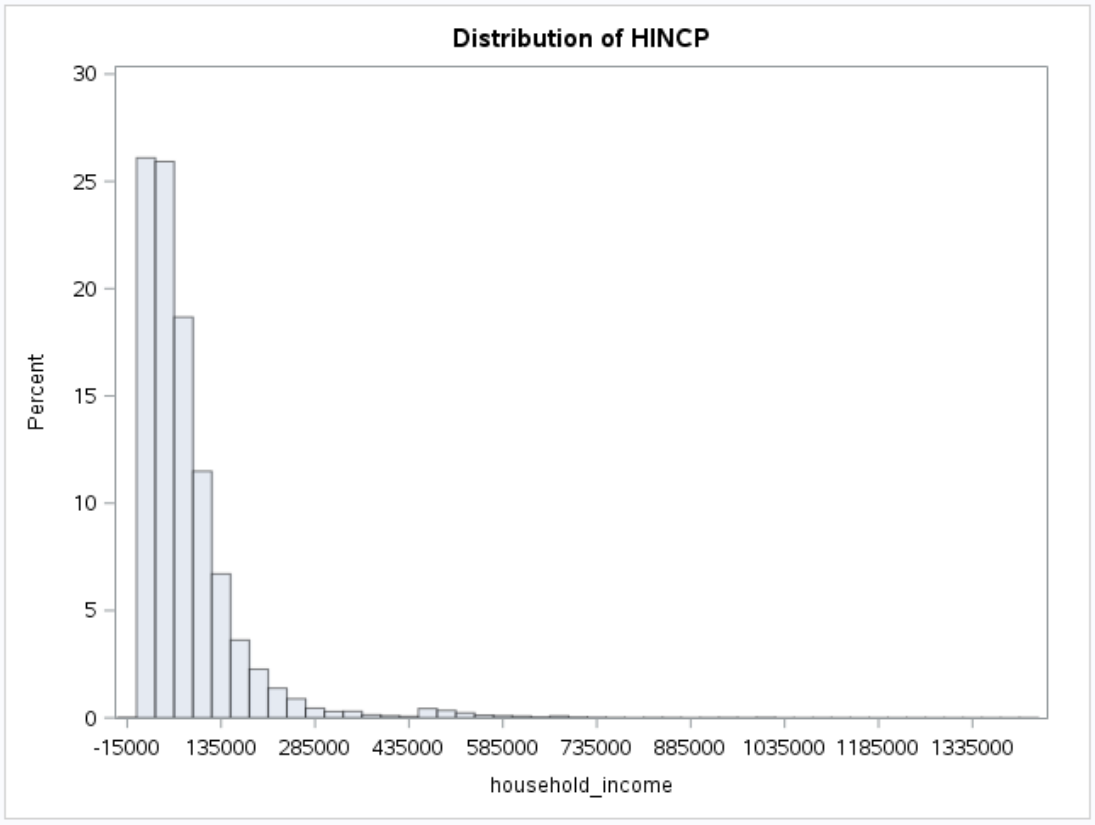
Create a table for cluster #20 for ease of variables analysis:

```
/*Create #20 cluster table for analysis*/

proc sql;
title 'Cluster #20 aggregated information for profiling';
create table census.psam_h17_subset1_cluster_20_agg as

select *
from census.psam_h17_subset1_20clusters_agg
where CLUSTER=20;
run;
```

Week 8 Topic:
Cluster #20: Understand HINCP

Descriptive Statistics for Numeric Variables

Analysis Variable : HINCP household_income						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
49673	8533	-11200.00	78744.88	56860.00	1425000.00	86804.24

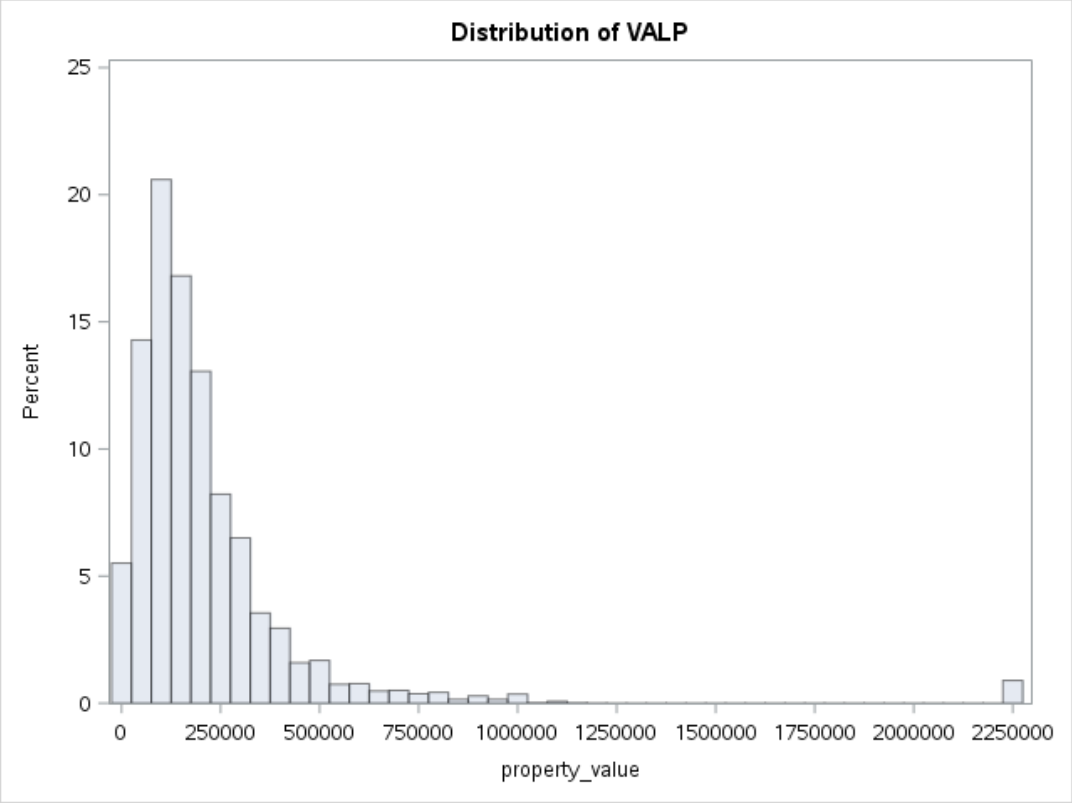


- To understand HINCP for cluster #20, we examine HINCP for the population in more detail.
- On slide 22, we validate the mean and median values of the data set to be in line with the population
- In keeping with the segmentation/grouping of household income on Slide 23, we determine that **Cluster #20 is in the middle to high income range**

Week 8 Topic:
Cluster #20: Understand VALP

Descriptive Statistics for Numeric Variables

Analysis Variable : VALP property_value						
N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
36715	21491	110.0000000	207714.16	150000.00	2267000.00	253557.68



- With mean, median, and skewed distribution, close in shape with HINCP, we use the same segmentation/grouping as seen on Slide 23
- Hence, from a VALP perspective, we deem **Cluster #20 to have property valued in the medium to high range**

Week 8 Topic:
Cluster #20: Additional Stats

ITM - 527

United States

2014 Population Estimate
318,857,056
Source: Vintage 2014 Population Estimates: US, State, and PR Total Population and Components of Change

Median Housing Value
$ 176,700
Source: 2009-2013 American Community Survey 5-Year Estimates

Individuals below poverty level
15.4 %
Source: 2009-2013 American Community Survey 5-Year Profiles

Educational Attainment: Percent high school graduate or higher
86.0 %
Source: 2009-2013 American Community Survey 5-Year Profiles

Health Insurance Coverage: Percent uninsured
14.9 %
Source: 2009-2013 American Community Survey 5-Year Profiles

Total Housing Units
132,057,804
Source: 2009-2013 American Community Survey 5-Year Estimates

Number of Companies
27,092,908
Source: 2007 Survey of Business Owners

From the census.gov website:

- We see that the median housing value for the US for 2014, $176,700, is close to the median property value in our data set – VALP of $150,000
- We see that the median household income for the US, $53,046, is close to the median household income in our data set – HINCP of $56,860

United States | *Median Household Income*
$ 53,046
Source: 2009-2013 American Community Survey 5-Year Estimates

Week 8 Topic:
Cluster #20: Additional Stats (cont.)

- In addition, we look at the census analysis on household income in more detail: https://www.census.gov/content/dam/Census/library/publications/2014/demo/p60-249.pdf
- Notice the mean HINCP value is higher than the median as expected and close to our data set value of $78,744
- We use the distribution groupings as a input for profiling our clusters

U.S. Census Bureau

Table A-1.

Households by Total Money Income, Race, and Hispanic Origin of Householder: 1967 to 2013

(Income in 2013 CPI-U-RS adjusted dollars. Hou... Before 2010, standard errors were calculated using the generalized variance function. For information ...veys/cps/techdocs/cpsmar14.pdf)

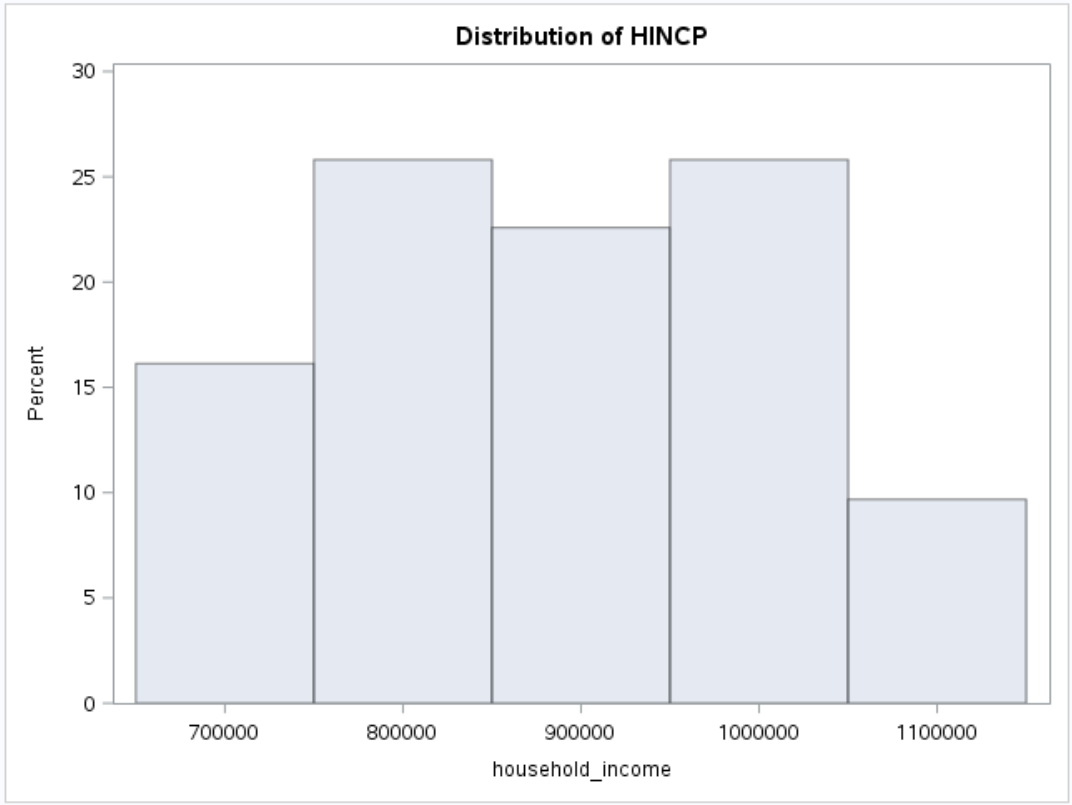
Race and Hispanic origin of householder and year	Number (thousands)	Percentage distribution: Total	Low: Under $15,000	Low: $15,000 to $24,999	Mid Low: $25,000 to $34,999	Mid Low: $35,000 to $49,999	Mid-High: $50,000 to $74,999	Mid-High: $75,000 to $99,999	High: $100,000 to $149,999	High: $150,000 to $199,999	High: $200,000 and over	Median income (dollars): Value	Median income (dollars): Standard error	Mean income (dollars): Value	Mean income (dollars): Standard error
ALL RACES															
2013[1]	122,952	100.0	12.7	11.3	10.4	13.6	17.6	11.9	12.4	5.3	4.8	51,939	276	72,641	499
2012	122,459	100.0	12.8	11.6	10.6	13.5	17.4	11.7	12.6	5.1	4.6	51,759	212	72,310	427
2011	121,084	100.0	13.1	11.3	10.6	13.9	17.5	11.5	12.3	5.3	4.6	51,842	260	72,166	381
2010[2]	119,927	100.0	12.8	11.5	10.5	13.3	17.6	11.8	12.9	5.1	4.6	52,646	347	72,001	385
2009[3]	117,538	100.0	11.8	11.2	10.6	13.8	17.4	12.3	13.1	5.3	4.7	54,059	231	73,824	264
2008	117,181	100.0	11.8	11.1	10.5	13.6	17.2	12.6	13.3	5.3	4.7	54,423	148	74,029	262
2007	116,783	100.0	11.5	10.6	10.2	13.0	17.9	12.3	14.1	5.5	4.9	56,436	157	75,957	265
2006	116,011	100.0	11.1	10.6	10.5	13.7	17.5	12.5	13.6	5.5	5.1	55,689	239	76,912	297
2005	114,384	100.0	11.6	10.8	9.8	13.7	18.0	12.5	13.5	5.2	4.9	55,278	185	75,584	285
2004[4]	112,242	100.0	11.9	10.8	10.2	13.9	17.5	12.6	13.2	5.2	4.5	54,674	242	74,569	291

ITM - 527

Week 8 Topic:
Cluster #20: Analyze HINCP distribution

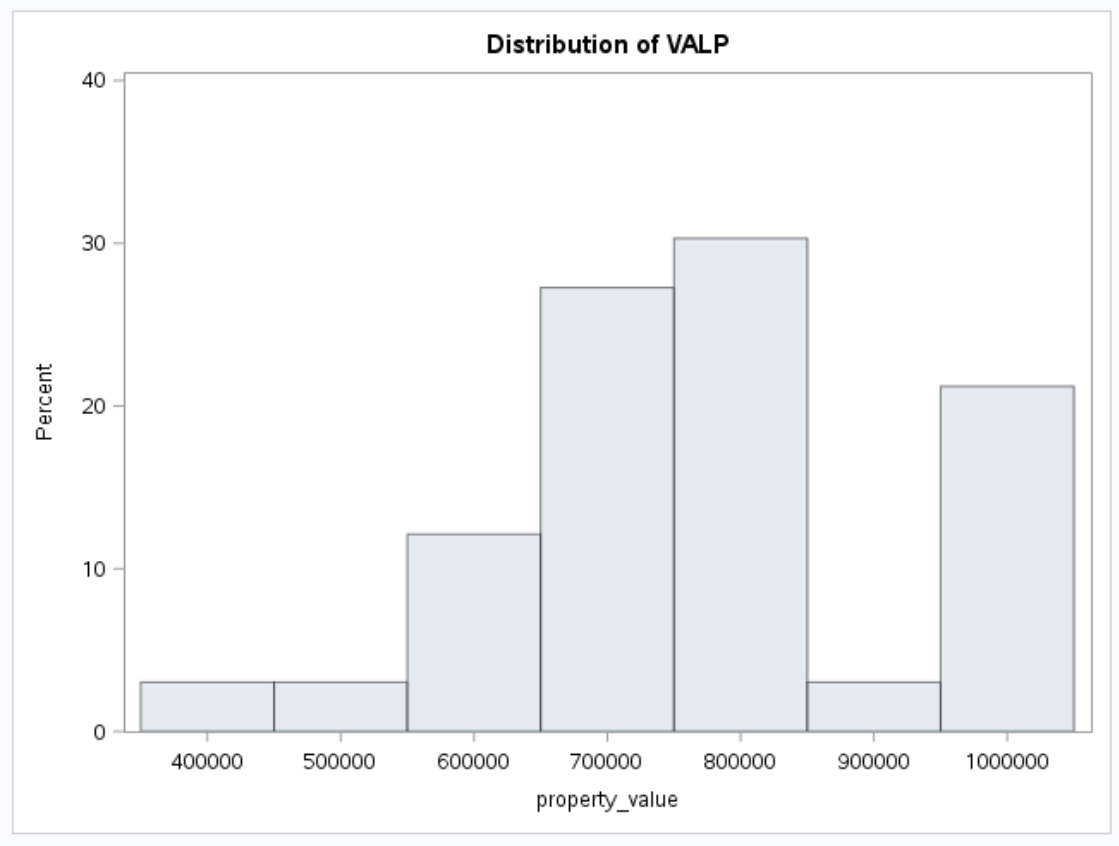
Descriptive Statistics for Numeric Variables

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
HINCP	household_income	31	2	678000.00	882724.52	866000.00	1117700.00	124659.12
VALP	property_value	33	0	425000.00	755272.73	750000.00	1000000.00	154745.61



- As discussed previously, we deem Cluster #20 in the **Medium-High** income category

Week 8 Topic:
Cluster #20: Analyze VALP distribution



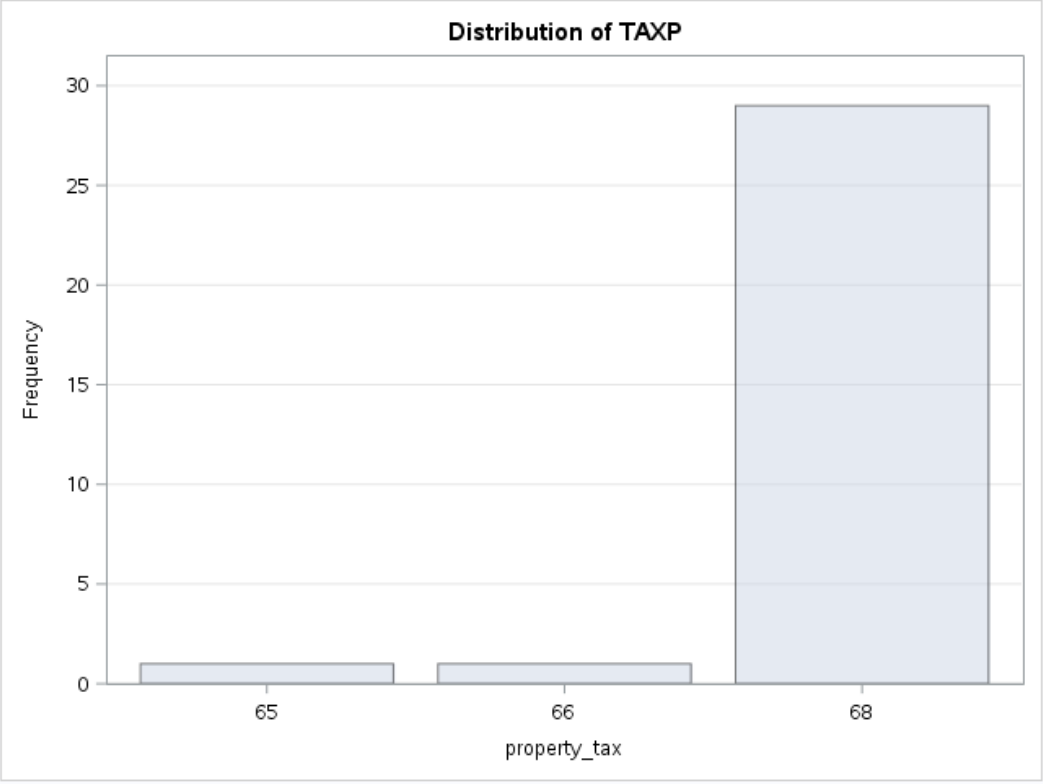
- As discussed previously, we deem Cluster #20 in the **Medium-High** property value category
- The mean and median VALP values for this cluster seems low if the average ratio (income to housing) is expected to be about 2+, or 3+ using recent rates
- Next step would be to correlate the two values, HINCP and VALP, to validate the data set, get a percentage of household over/under extended
- We may also look for armed forces employment for this cluster

Week 8 Topic:
Cluster #20: Analyze TAXP distribution

property_tax

TAXP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
65	1	3.23	1	3.23
66	1	3.23	2	6.45
68	29	93.55	31	100.00

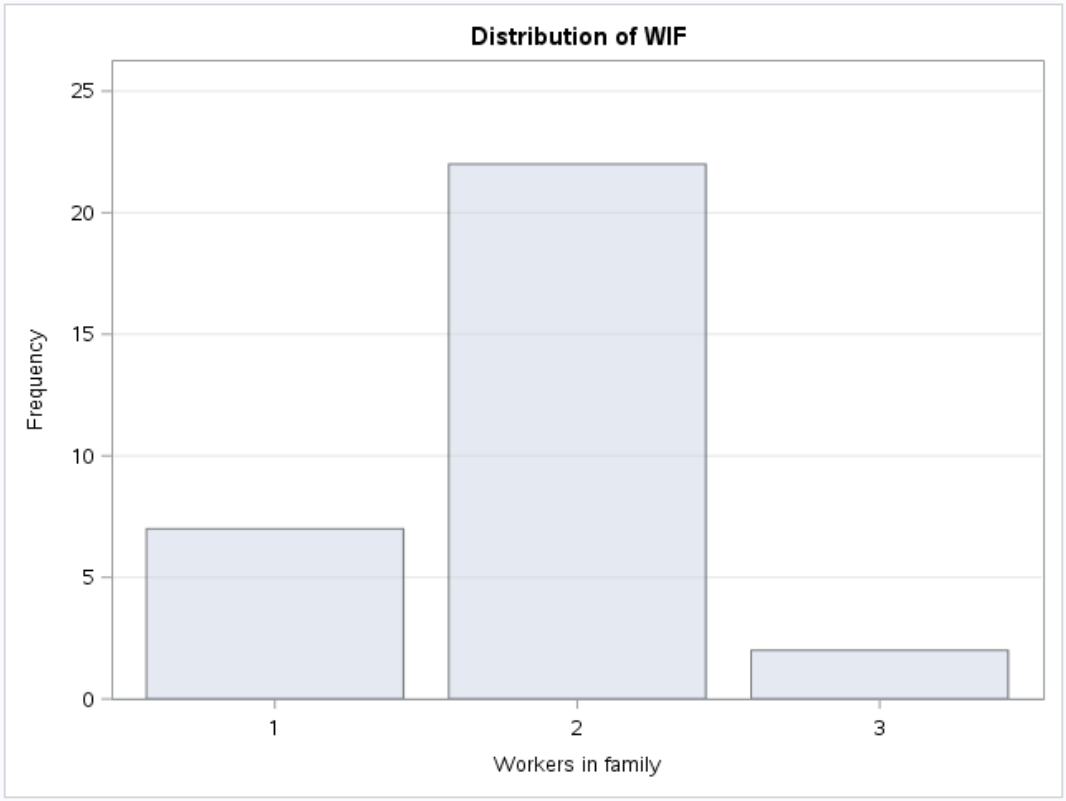
Frequency Missing = 2



- We see that majority of households in Cluster #20 pay over $10,000 in property taxes
- This seems high for the median home valued at $75,000.
- Next step would be to gather mortgage & tax payment (MRGP and TAXP) information as a percentage of the household income.

Week 8 Topic:
Cluster #20: Analyze WIF distribution

Workers in family				
WIF	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	7	22.58	7	22.58
2	22	70.97	29	93.55
3	2	6.45	31	100.00
Frequency Missing = 2				



- We see that majority, 70%, of Cluster #20 population has two workers in the family which is a good indicator of the household withstanding a tax hike
- Next steps would be to compare this against the other clusters using PROC FREQ

ITM - 527

Week 8 Topic:
Cluster #20: WIF distribution in clusters

- Interesting to note that Cluster #20 has the highest percentage of 2 workers in family.
- Clusters 4, 7, 8, 10, and 14 seem to have similar proportions to Cluster #20 whereas other clusters have a more even distribution between 1 or 2 workers in family. These seem to be neighborhood clusters with similar characteristics
- Next step would be to analyze these 'like' clusters in tandem for the tax hike

Table of WIF by CLUSTER

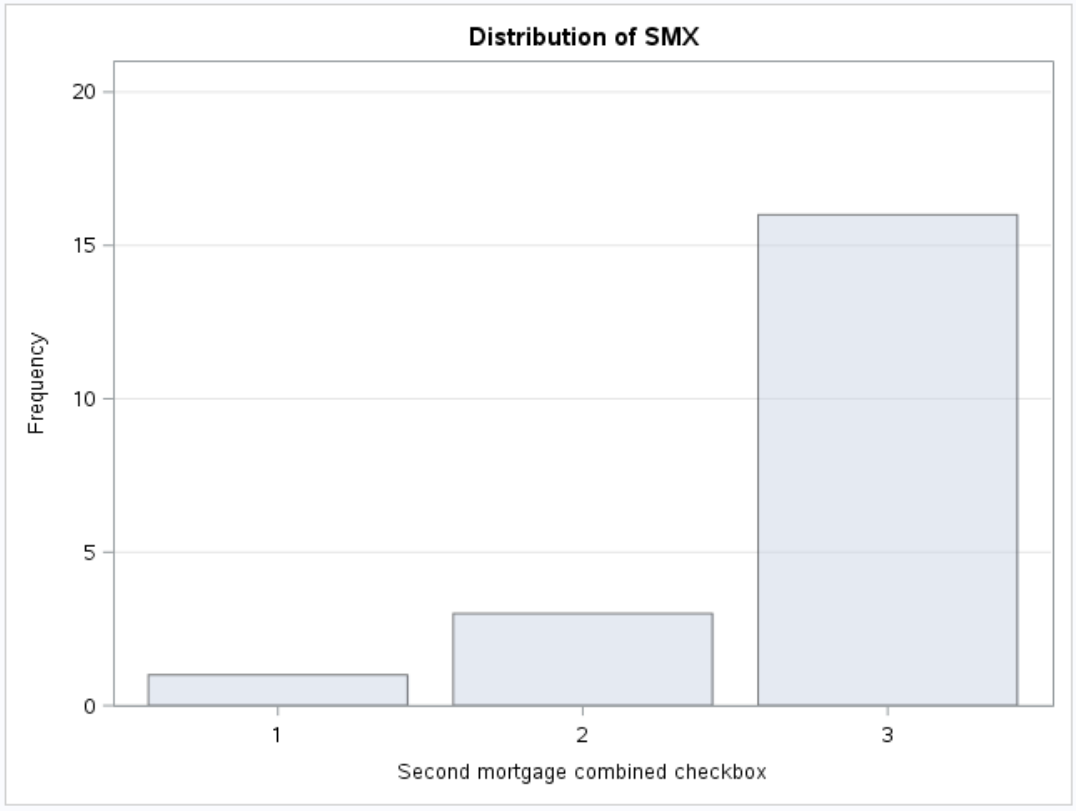
	CLUSTER(Cluster)																			
WIF(Workers in family)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	Total
0	1202 3.62 24.78 19.47	872 2.63 17.98 27.30	380 1.14 7.83 12.76	67 0.20 1.38 2.73	34 0.10 0.70 5.70	0 0.00 0.00 0.00	160 0.48 3.30 3.76	15 0.05 0.31 10.49	58 0.17 1.20 8.95	2 0.01 0.04 0.84	3 0.01 0.06 2.46	27 0.08 0.56 10.55	4 0.01 0.08 4.04	45 0.14 0.93 3.17	832 2.51 17.15 17.48	0 0.00 0.00 0.00	12 0.04 0.25 4.76	1136 3.42 23.42 21.36	2 0.01 0.04 0.87	0 0.00 0.00 0.00	4851 14.61
1	2256 6.79 21.99 36.54	1651 4.97 16.09 51.69	899 2.71 8.76 30.19	338 1.02 3.29 13.77	225 0.68 2.19 37.75	19 0.06 0.19 43.18	694 2.09 6.76 16.33	35 0.11 0.34 24.48	197 0.59 1.92 30.40	67 0.20 0.65 28.03	54 0.16 0.53 44.26	77 0.23 0.75 30.08	44 0.13 0.43 44.44	251 0.76 2.45 17.70	1435 4.32 13.99 30.15	1 0.00 0.01 25.00	70 0.21 0.68 27.78	1864 5.61 18.17 35.05	75 0.23 0.73 32.47	7 0.02 0.07 22.58	10259 30.89
2	2217 6.68 15.89 35.91	579 1.74 4.15 18.13	1284 3.87 9.20 43.12	1455 4.38 10.43 59.29	277 0.83 1.99 46.48	24 0.07 0.17 54.55	2378 7.16 17.04 55.95	76 0.23 0.54 53.15	290 0.87 2.08 44.75	129 0.39 0.92 53.97	56 0.17 0.40 45.90	107 0.32 0.77 41.80	44 0.13 0.32 44.44	834 2.51 5.98 58.82	1959 5.90 14.04 41.16	1 0.00 0.01 25.00	124 0.37 0.89 49.21	1972 5.94 14.13 37.08	124 0.37 0.89 53.68	22 0.07 0.16 70.97	13952 42.01
3	499 1.50 12.03 8.08	92 0.28 2.22 2.88	415 1.25 10.00 13.94	594 1.79 14.32 24.21	60 0.18 1.45 10.07	1 0.00 0.02 2.27	1018 3.07 24.54 23.95	17 0.05 0.41 11.89	103 0.31 2.48 15.90	41 0.12 0.99 17.15	9 0.03 0.22 7.38	45 0.14 1.08 17.58	7 0.02 0.17 7.07	288 0.87 6.94 20.31	534 1.61 12.87 11.22	2 0.01 0.05 50.00	46 0.14 1.11 18.25	346 1.04 8.34 6.51	30 0.09 0.72 12.99	2 0.01 0.05 6.45	4149 12.49
Total	6174 18.59	3194 9.62	2978 8.97	2454 7.39	596 1.79	44 0.13	4250 12.80	143 0.43	648 1.95	239 0.72	122 0.37	256 0.77	99 0.30	1418 4.27	4760 14.33	4 0.01	252 0.76	5318 16.01	231 0.70	31 0.09	33211 100.00

Frequency Missing = 24995

ITM - 527

Week 8 Topic:
Cluster #20: Analyze SMX distribution

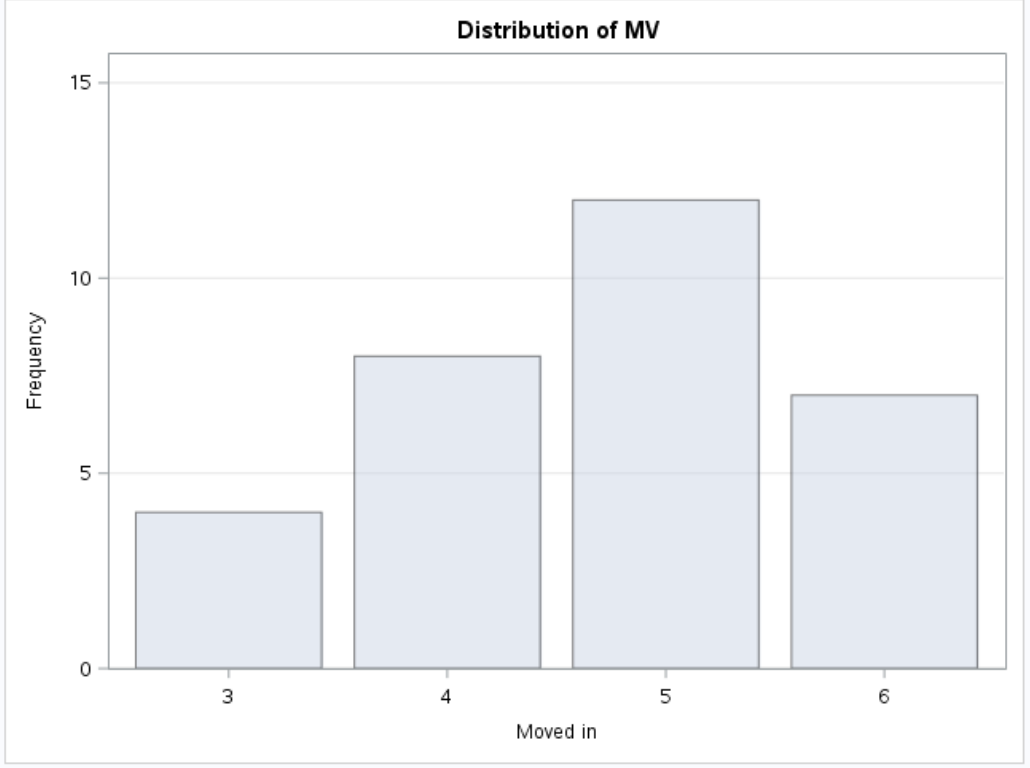
Second mortgage combined checkbox				
SMX	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	1	5.00	1	5.00
2	3	15.00	4	20.00
3	16	80.00	20	100.00
Frequency Missing = 13				



- Majority do not have a second mortgage or home equity loan in Cluster #20 which is a good indicator of household withstanding a tax hike

Week 8 Topic:
Cluster #20: Analyze MV distribution

Moved in				
MV	Frequency	Percent	Cumulative Frequency	Cumulative Percent
3	4	12.90	4	12.90
4	8	25.81	12	38.71
5	12	38.71	24	77.42
6	7	22.58	31	100.00
Frequency Missing = 2				

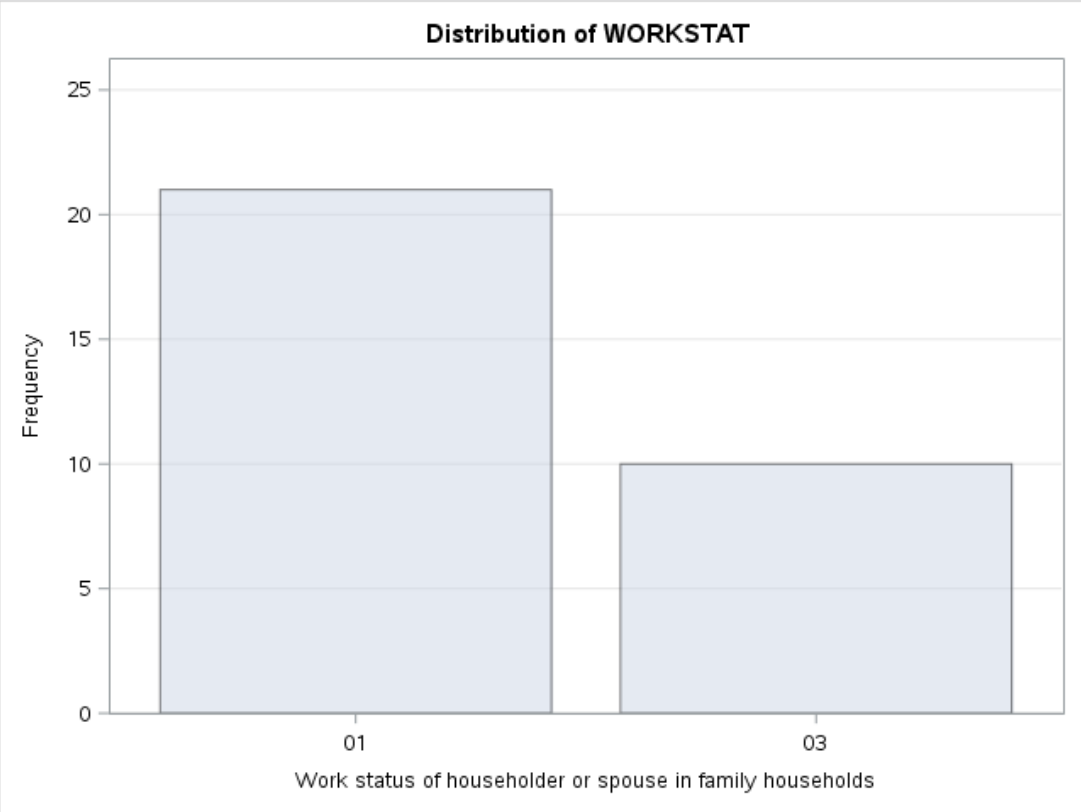


- Majority have owned the property for more than 5 years with more than half over 10 years.
- Next steps would be understand the loan terms of the property ownership.
- Either way, this is a good indicator for the cluster to withstand a tax hike

ITM - 527

Week 8 Topic:
Cluster #20: Analyze WORKSTAT dist.

Work status of householder or spouse in family households				
WORKSTAT	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01	21	67.74	21	67.74
03	10	32.26	31	100.00
Frequency Missing = 2				



- This percentage break out of husband and wife work status seems to be in line with the WIF data
- As all the cluster members are in the workforce, this also indicates that the cluster is in good standing to withstand a tax hike

Week 8 Topic:
Cluster Analysis Sample Code

Sort the data to get a better understanding of the clusters:

```
libname census "/folders/myfolders/census";

/*sort the 20 clusters by population*/
proc sort data=census.psam_h17_subset1_20clusters_stat;
by descending OVER_ALL;
run;

/*print the sorted data by cluster population*/
proc print;
var  CLUSTER VALP HINCP OVER_ALL;
WHERE _TYPE_ = "FREQ";
title2 '20-cluster solution sorted by population';
run;

/*plot the seed values per cluster*/
proc sgplot;
scatter y=HINCP x=VALP  / group=cluster;
WHERE _TYPE_ = "SEED";
title 'ACS PUMS 2013 1 YR 20-Cluster Analysis - Means';
title2 'of Data Containing Property Value and Household Income';
run;
```

Week 8 Topic: Cluster Analysis - Grouping

We may consider applying a similar tax strategy for these groups of clusters.

Group A:

- These 8 out of 20 clusters (1, 18, 2, 15, 7, 3, 4, and 14) have over 1000 members which is high compared to the other 12 clusters with under 1000 members.
- Cluster 1, with 10,460 members, is especially over populated.
- This was expected since we left the variables skewed. We may want to consider additionally clustering the larger clusters.

Group B:

- These 5 out of 20 clusters are less populated and high valued which indicate outlier clusters. We may consider a singular tax strategy for these after investigating why the high property owns have low household income (Clusters 6, 8, and 13).
- We have Cluster 5 with close to 1000 members that we expected to be less populated as an outlier cluster. More investigation is needed to determine whether to include in Group B.

20-cluster solution sorted by population

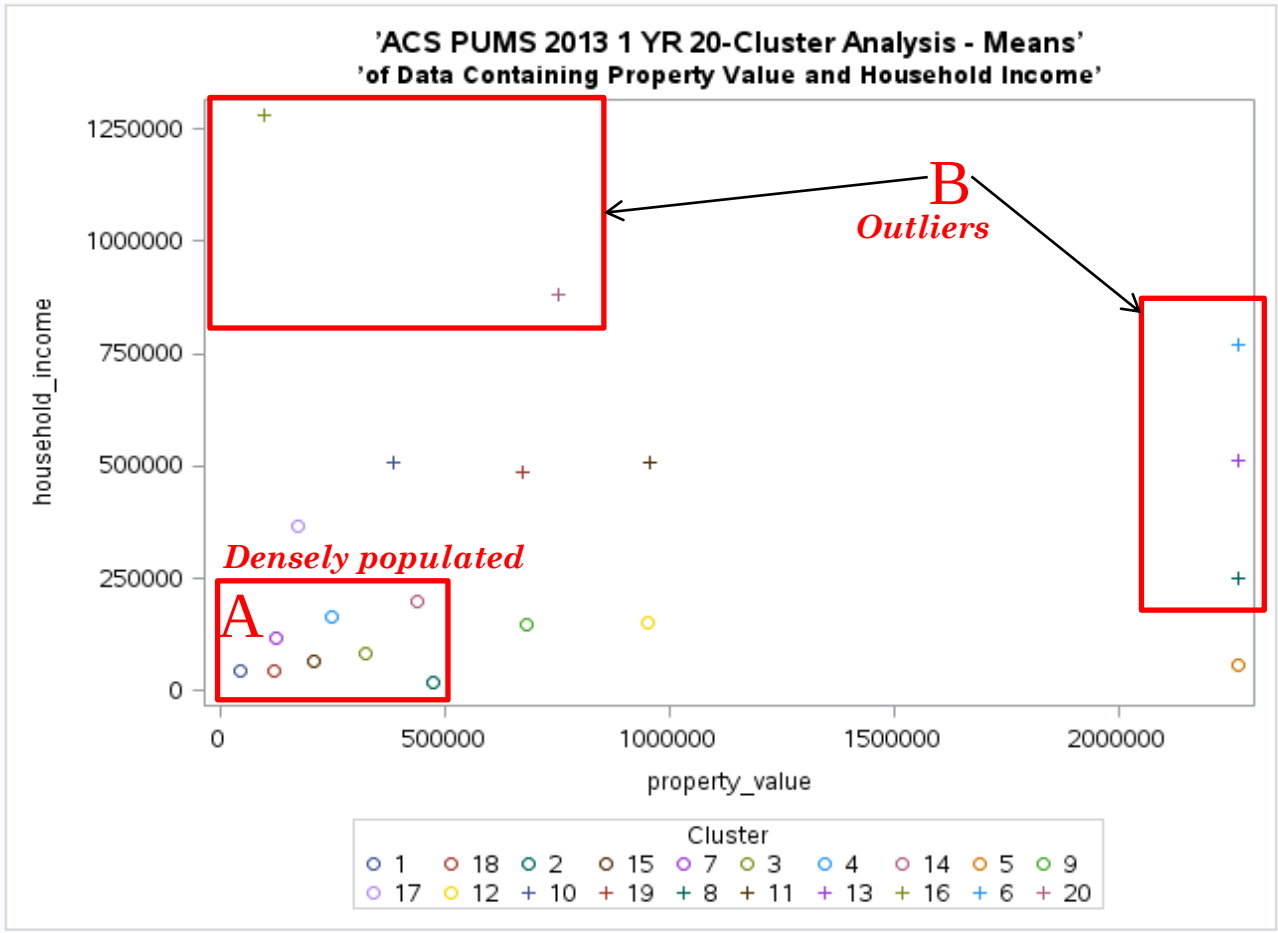
Obs	CLUSTER	VALP	HINCP	OVER_ALL
25	1	8565	10278	10460
28	18	7538	8430	8454
30	2	737	7377	7381
32	15	5891	6696	6734
34	7	4463	5028	5070
36	3	2974	4114	4138
38	4	2663	2741	2770
40	14	1506	1603	1605
42	5	115	987	993
44	9	657	793	794
46	17	348	293	370
48	12	258	294	294
50	10	283	271	284
52	19	248	267	271
54	8	73	175	175
56	11	124	130	131
58	13	99	116	116
60	16	99	4	99
62	6	41	45	45
64	20	33	31	33

A

B

Week 8 Topic:
Cluster Analysis - Grouping

Using the scatterplot below, we review the mean values and location of the clusters to confirm the groupings:



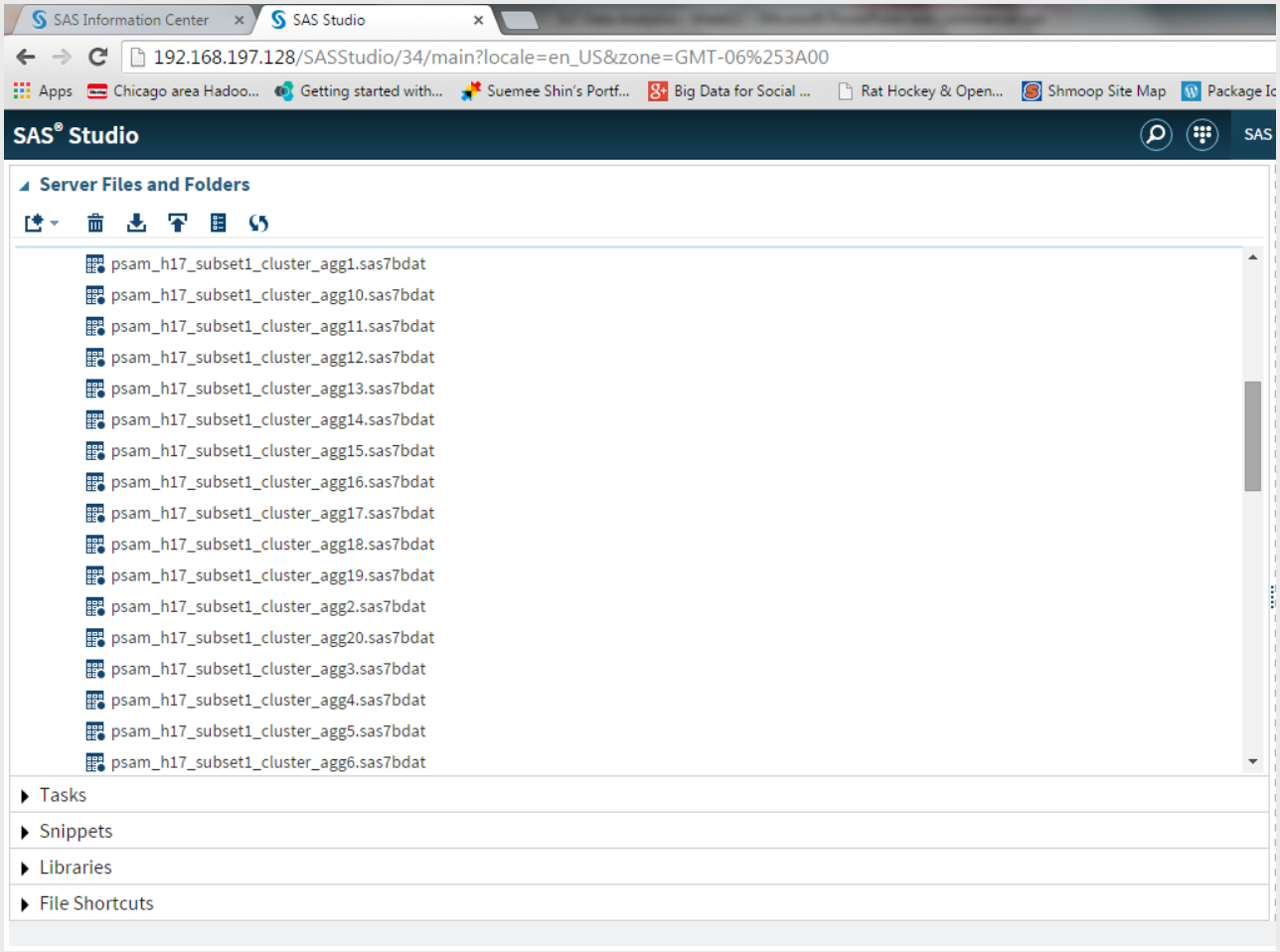
Week 8 Topic: Creating tables for each cluster

- This is so that you can use the Characterize Data function for each cluster to review the profiling variable values and its distribution.
- We gather the most frequent values for TAXP, SMX, WIF, MV, and WORKSTAT and create a matrix in Excel.

```
%macro sqlloop;
proc sql;

/*Loop through 20 times to create a table for each cluster*/
%DO k=1 %TO 20;
create table census.psam_h17_subset1_cluster_agg&k. as
select *
from census.psam_h17_subset1_20clusters_agg
where CLUSTER=&k.;
%END;
QUIT;

%mend;
%sqlloop;
```

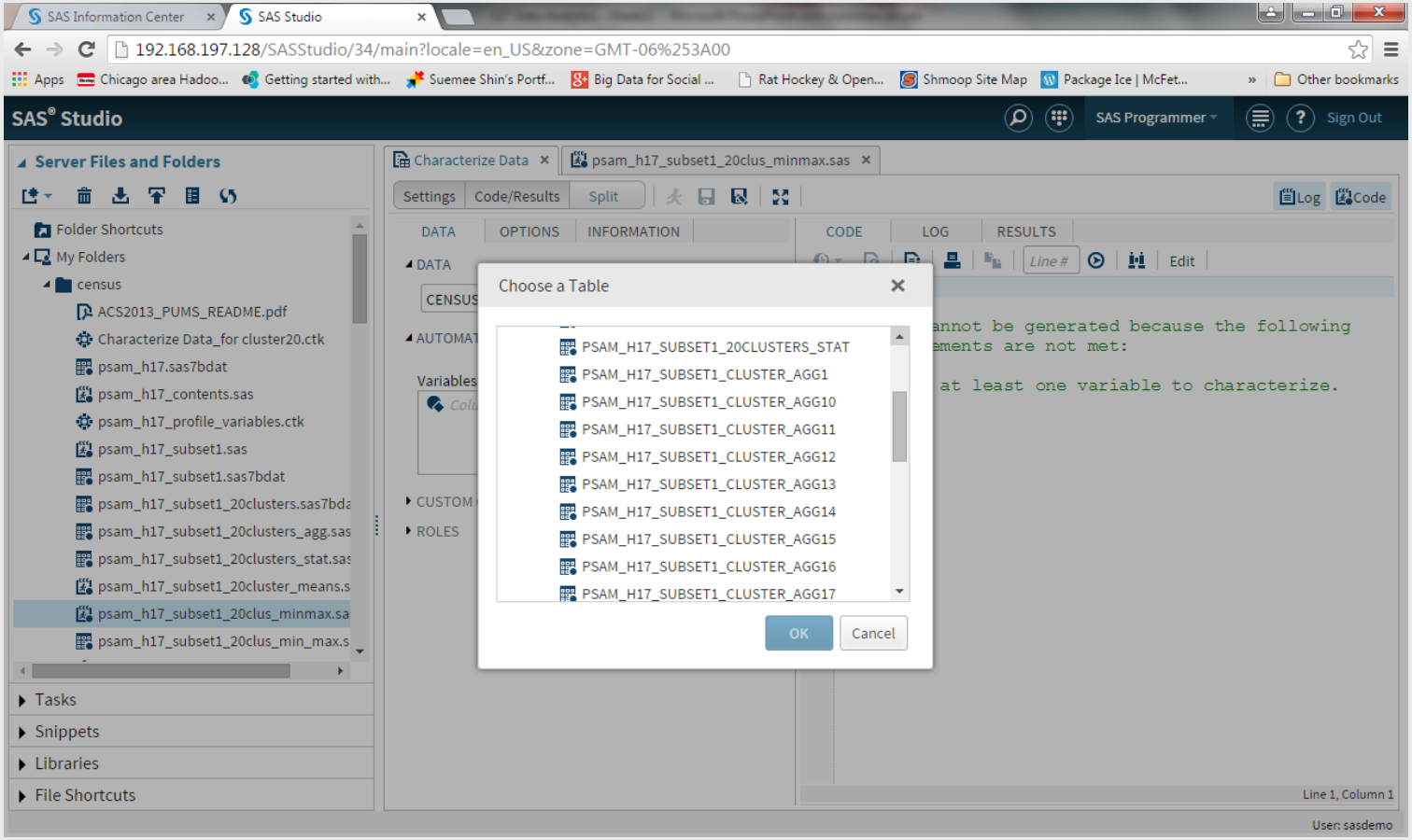
Week 8 Topic:
SAS view of each SAS cluster table

- This saves the results of the clustering for future use.



Week 8 Topic:
Choosing each cluster table for review

- Using Characterize Data, review the profiling variables for each cluster.
- Log top values (most frequent) of the profiling variables in the matrix.



Week 8 Topic: Getting Mean, Min, & Max

- This is to compare the VALP and HINCP values for each cluster. We will merge this information and the other profiling variables' most frequent values into a matrix.

```
/*Get Min and Max VALP HINCP values by cluster*/

proc sql;
title 'Mean, Min and Max VALP HINCP values by cluster';
create table census.psam_h17_subset1_20clus_min_max as

select CLUSTER,
Mean(HINCP) as HINCP_Mean, Min(HINCP) as HINCP_Min, Max(HINCP) as
HINCP_Max,
Mean(VALP) as VALP_Mean, Min(VALP) as VALP_Min, Max (VALP) as
VALP_Max
from census.psam_h17_subset1_20clusters_agg group by CLUSTER;

run;
```

Week 8 Topic:
Cluster matrix for profiling analysis

- In Excel, we merge the information and review the variables together for profiling:

Example of Analysis

CLUSTER	Count	HINCP_Mean	HINCP_Min	HINCP_Max	VALP_Mean	VALP_Min	VALP_Max	MF TAXP	MF SMX	MF WIF	MF MV	MF WORKSTAT	Tax Group	Income Cat	Property Cat	Notes
20	33	$882,725	$678,000	$1,117,700	$755,273	$425,000	$1,000,000		68	3	2	5	1	A	high	high	
6	45	$770,242	$645,000	$1,019,000	$2,267,000	$2,267,000	$2,267,000		68	3	2,1	3,4,5	1,3	A	high	max	
13	116	$510,772	$385,000	$636,000	$2,267,000	$2,267,000	$2,267,000		68	3	1,2	4,3,5	3,1	A	high	max	
11	131	$506,125	$334,200	$766,000	$958,242	$820,000	$1,175,000		68	3	2,1	5,4,3	1,3	A	high	high	
8	175	$251,049	$155,000	$370,800	$2,267,000	$2,267,000	$2,267,000		68	3	2,1	3,1,5	1,3	A	high	max	
19	271	$487,479	$321,000	$688,500	$671,310	$525,000	$800,000		68	3	2,1	5,4,3	1,3	A	high	mid-high	
12	294	$152,966	$3,300	$329,000	$951,426	$825,000	$1,150,000		68	3	2,1	5,4,3	1,3	A	mid-high	high	
9	794	$148,692	$0	$322,400	$683,473	$555,000	$810,000		68	3	2,1	5,4,3	1,3	A	mid-high	mid-high	
10	284	$508,055	$350,500	$848,800	$384,134	$100,000	$525,000	68, 64, 65, 67, 66		3	2,1	5,4,6	1,3	B	high	mid	
17	370	$367,531	$249,000	$609,000	$171,630	$110	$325,000	64,65,63,66		3	2,1	5,4,7	1,3	B	high	mid	
14	1605	$197,773	$104,200	$361,380	$437,610	$329,000	$580,000	68,66,67		3		2 5,4	1	B	mid-high	mid	
4	2770	$164,348	$101,000	$299,800	$248,184	$160,000	$340,000	64,65,62		3		2 5,4,6	1	B	mid-high	mid	
3	4138	$81,571	$5,100	$170,000	$322,122	$264,000	$415,000	64,65,66		3	2,1	5,4	1,3	C	mid	mid-low	
7	5070	$118,756	$80,700	$252,000	$123,634	$170	$190,000	42,62,52		3		2 5,4	1	C	mid-high	mid-low	
2	7381	$18,972	$5,100	$126,400	$472,739	$385,000	$600,000	68,66,64		3	1,0	1,3,4	13,15	D	low	mid-low	woman head of household, 15-unemployed
15	6734	$63,518	$11,200	$128,060	$206,932	$160,000	$275,000	64,62,63		3	2,1,0	5,4,7	1,9,3	D	mid	mid-low	9-unemployed
18	8454	$45,487	$5,100	$83,630	$120,010	$82,000	$170,000	42,52,62		3	2,1,0	5,4,7	1,9,3	D	mid	mid-low	9-unemployed
1	10460	$41,993	$5,100	$165,000	$46,484	$110	$85,000	1,24		3	1,2,0	5,7,4	1,9,13,3	D	mid	low	9-unemployed, woman head of household
16	99	$1,280,750	$1,027,000	$1,425,000	$96,222	$72,000	$350,000										too many missing HINCP - ignore cluster.
5	993	$57,161	$5,100	$151,000	$2,267,000	$2,267,000	$2,267,000										too many missing VALP - ignore cluster.

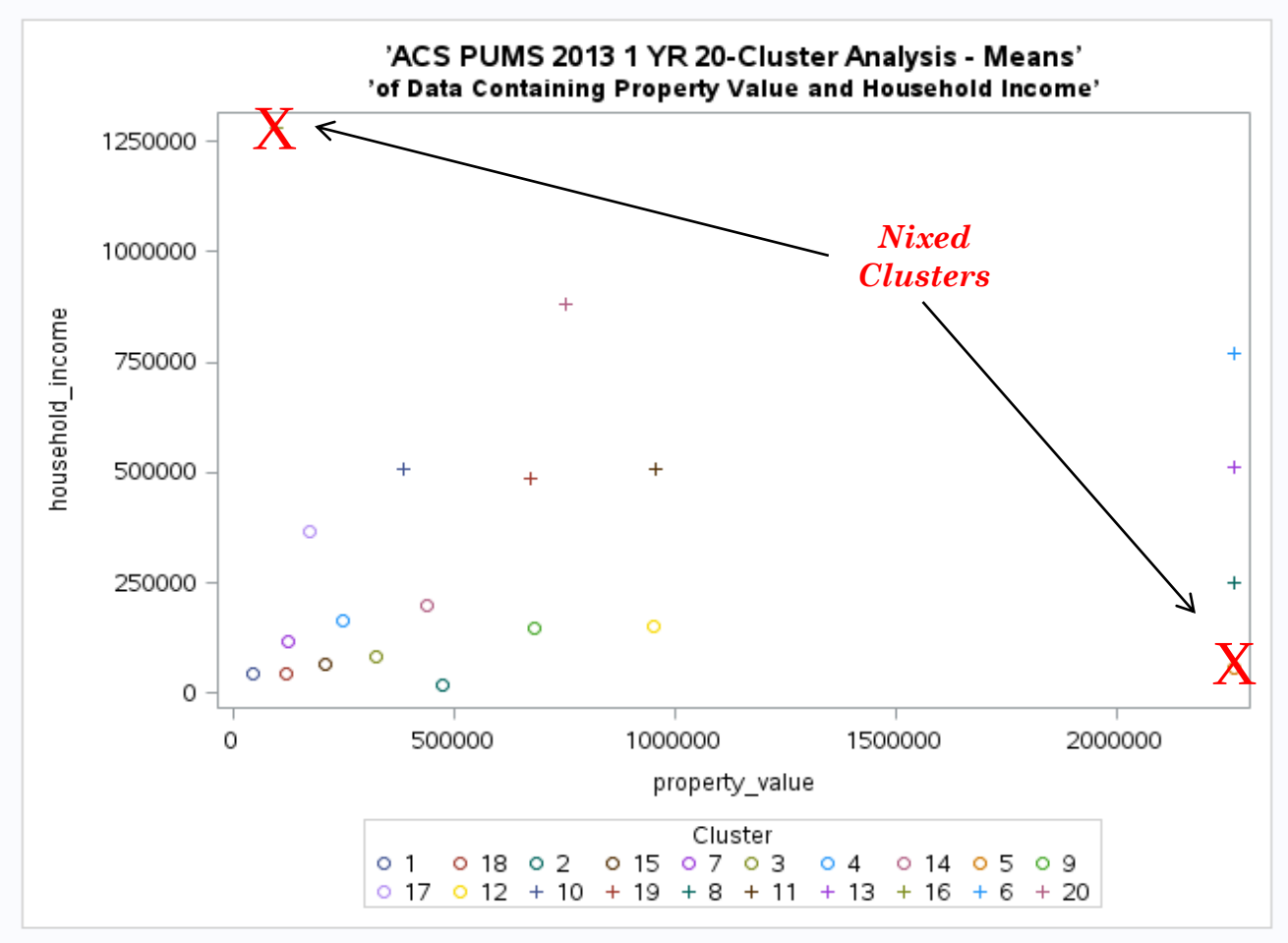
Week 8 Topic:
Tax Groups of Clusters

- Group A:
  - Clusters: 6,8,9,11,12,13,19,20
  - High Income, High Property Value Owners
  - Paying above $10k in property taxes already (TAXP=68)
- Group B:
  - Clusters: 4,10,14,17
  - High Income, Mid Property Value Owners
  - Paying above ~$5k in property taxes already
- Group C:
  - Clusters: 3,7
  - Mid to Mid-Low levels
  - Paying above ~$3k in property taxes already
- Group D:
  - Clusters: 1,2,15,18
  - Unemployed, woman household, or single worker household
  - Paying low range property taxes

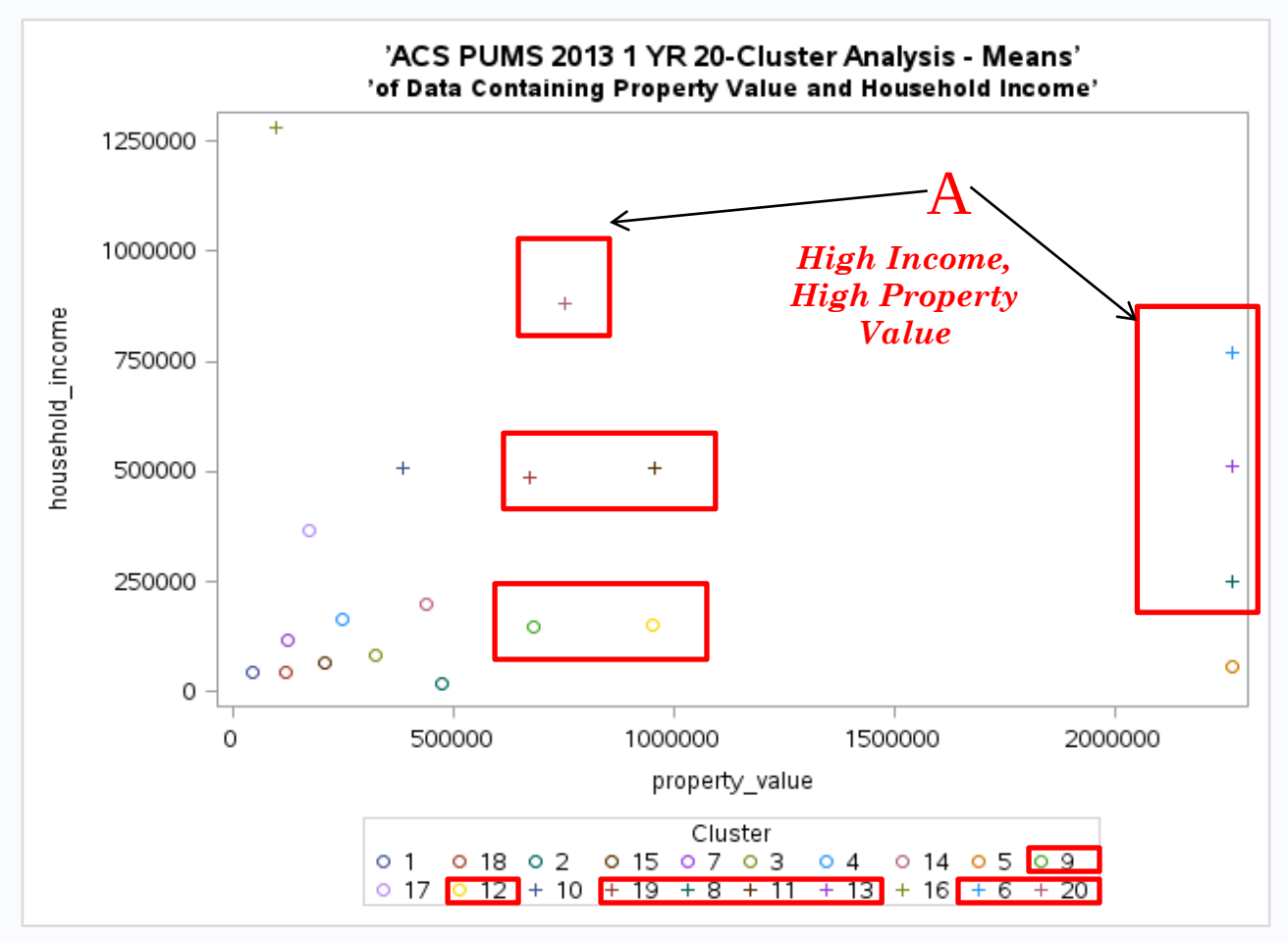
Income Category:	
<35k	low
35k-100k	mid
100k-200k	mid-high
>200k	high

Property Category:	
>2M	max
>700k	high
500k-700k	mid-high
250k-500k	mid
100k-250k	mid-low
<100k	low

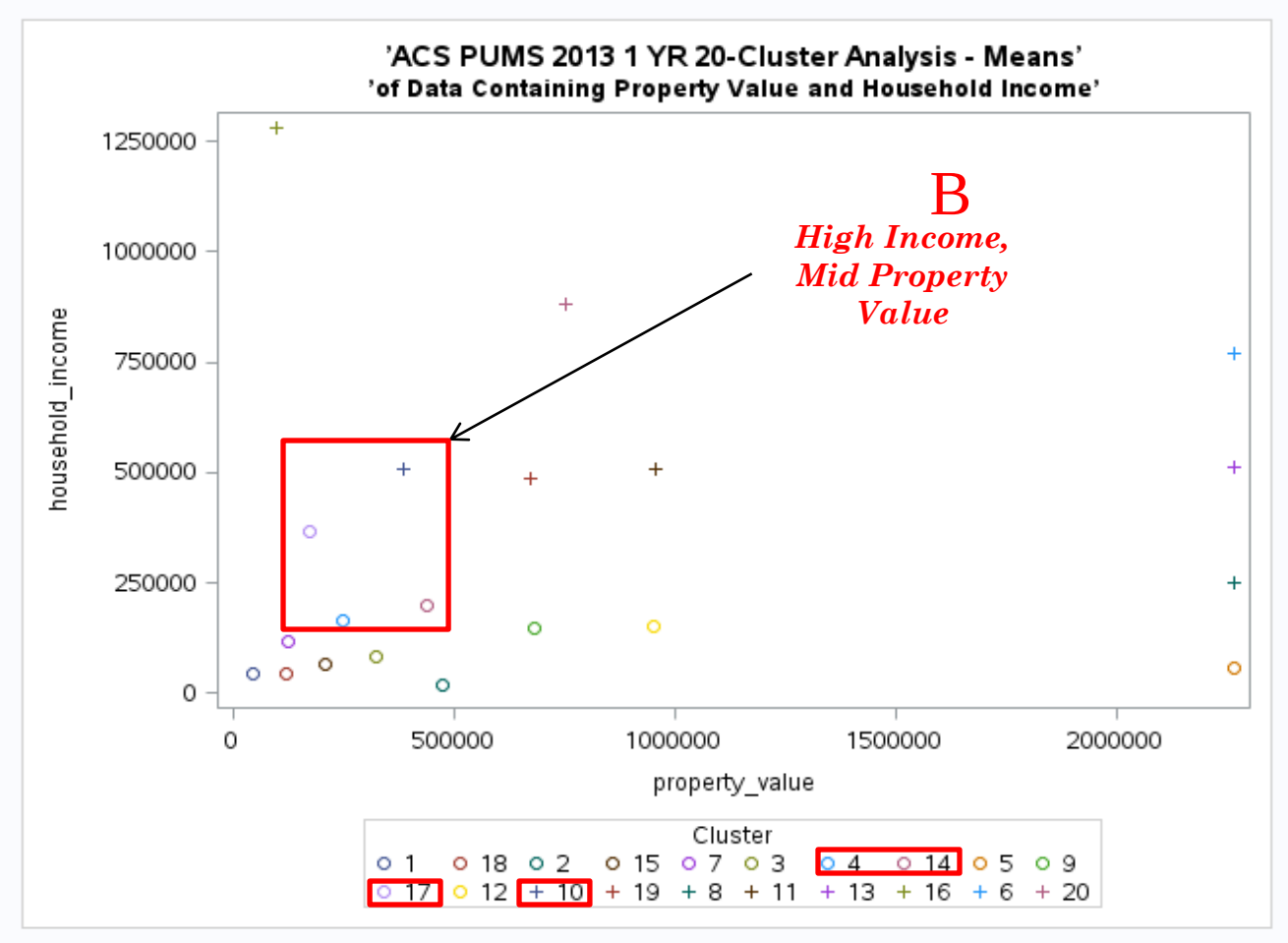
Week 8 Topic:
Tax Group 'A' on the plot



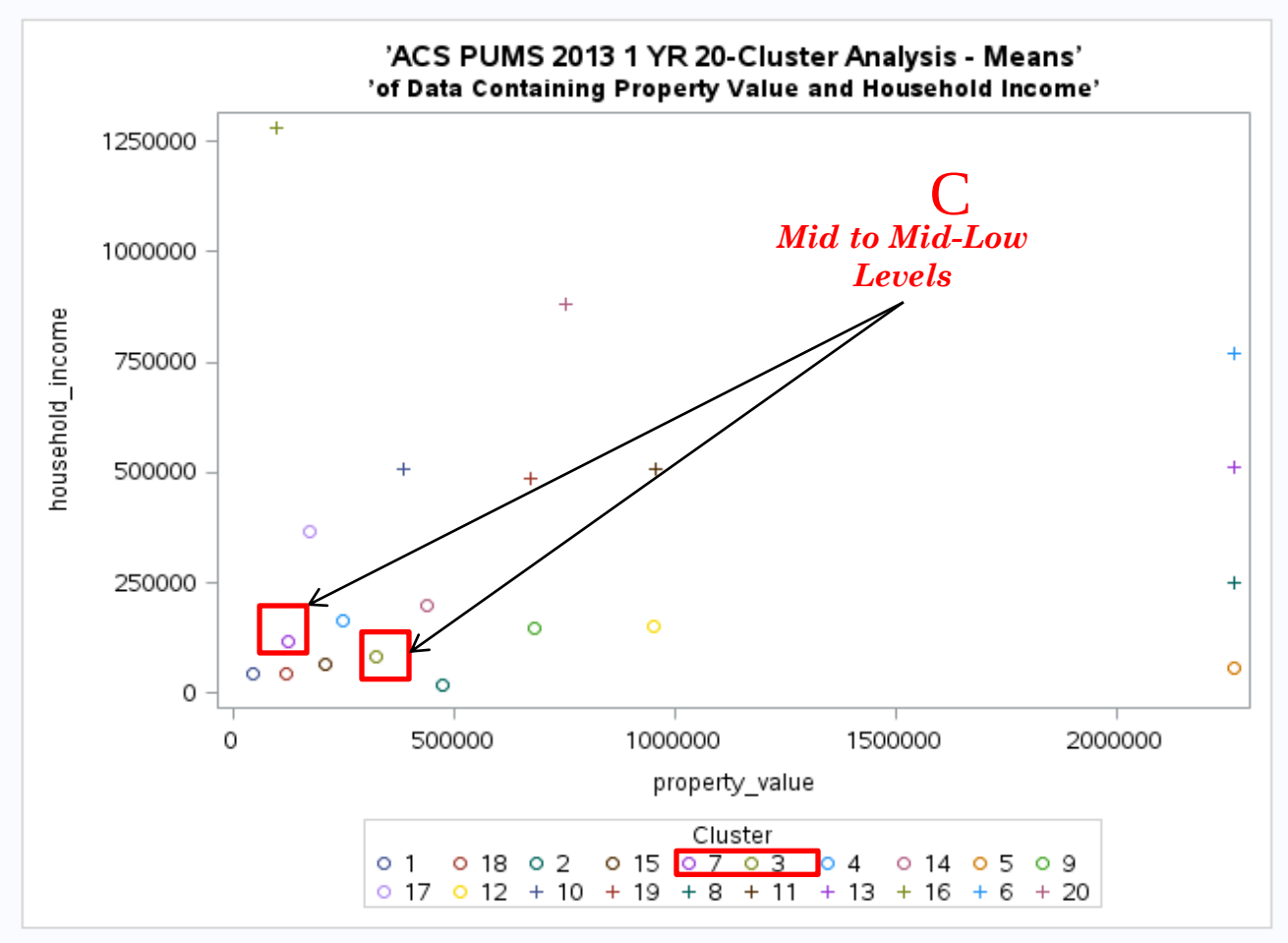
Week 8 Topic:
Tax Group 'A' on the plot



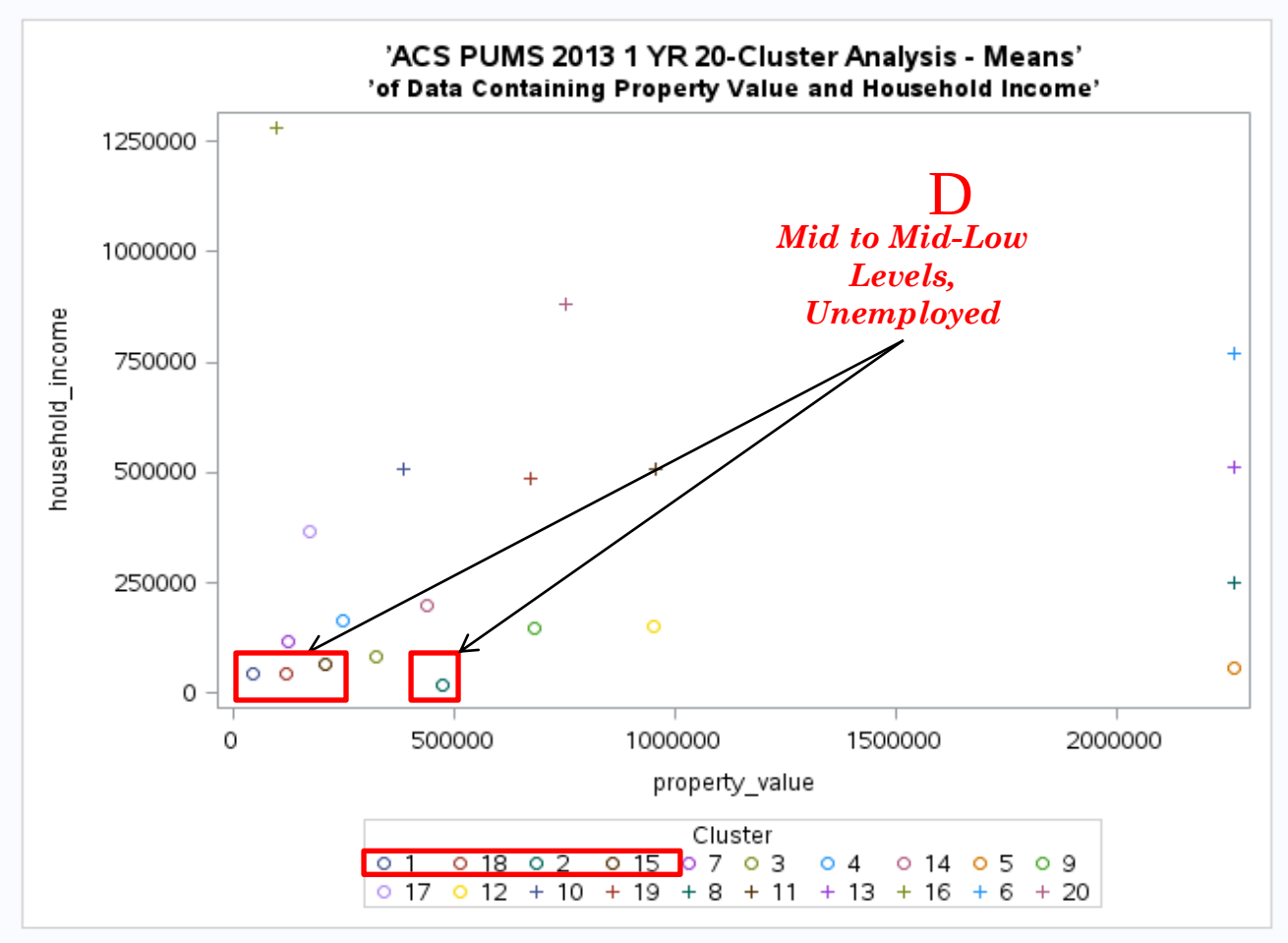
Week 8 Topic:
Tax Group 'B' on the plot



Week 8 Topic:
Tax Group 'C' on the plot



Week 8 Topic:
Tax Group ‘D’ on the plot



Week 8 Topic:
Tax Increase Example - Group 'C'

For the tax increase example, we take Group C from last class:

- Consists of Clusters: 3,7
- Mid to Mid-Low HINCP & VALP values
- Paying above ~$3k in property taxes already

Income Category:	
<35k	low
35k-100k	mid
100k-200k	mid-high
>200k	high

Property Category:	
>2M	max
>700k	high
500k-700k	mid-high
250k-500k	mid
100k-250k	mid-low
<100k	low

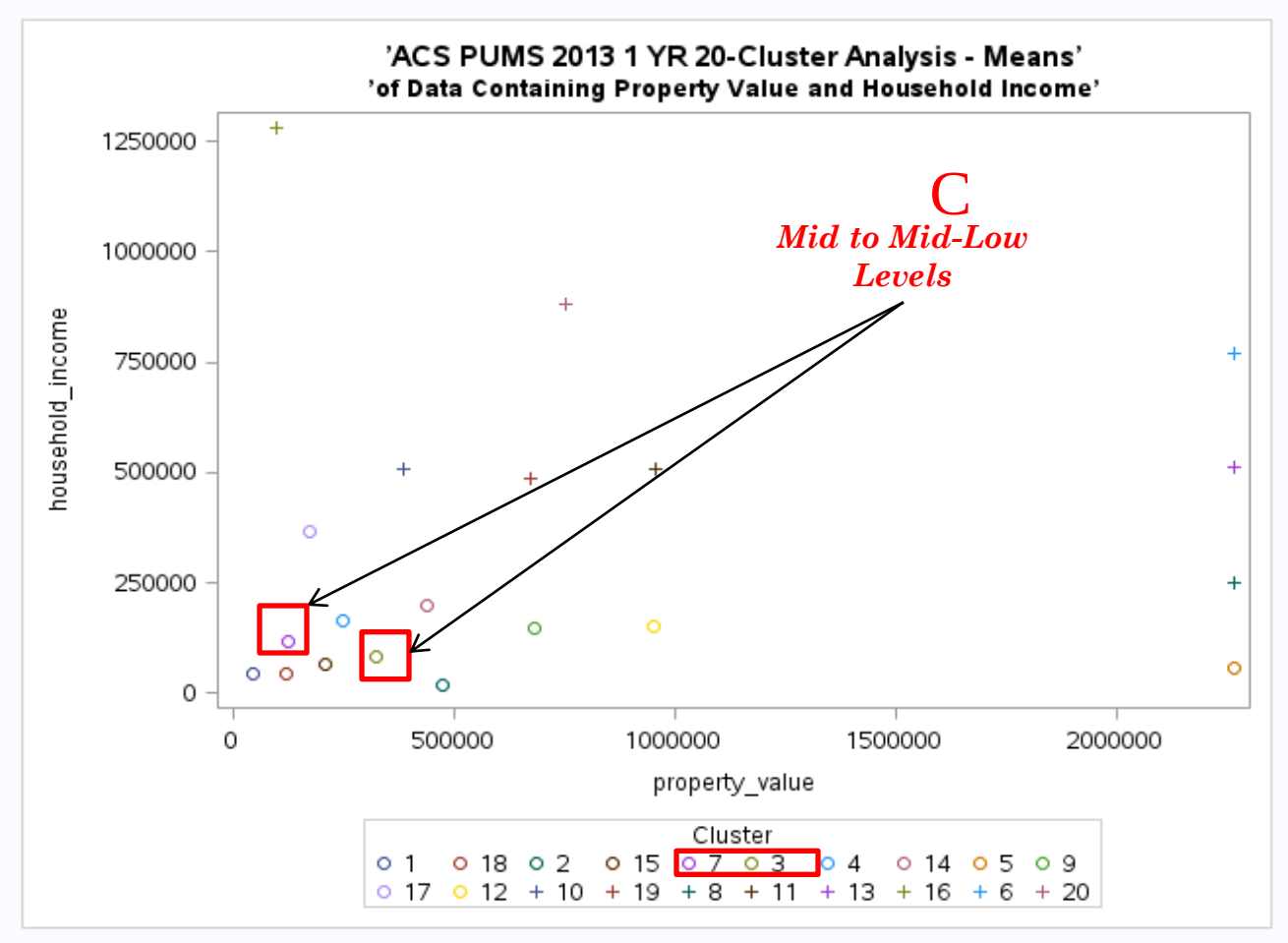
HINCP and VALP values for Group C:

CLUSTER	Count	HINCP_Mean	HINCP_Min	HINCP_Max	VALP_Mean	VALP_Min	VALP_Max
3	4138	$81,571	$5,100	$170,000	$322,122	$264,000	$415,000
7	5070	$118,756	$80,700	$252,000	$123,634	$170	$190,000

Tax Bracket values for Group C:

CLUSTER	MF TAXP	Income Cat	Property Cat
3	64,65,66	mid	mid-low
7	42,62,52	mid-high	mid-low

Week 8 Topic:
Tax Increase Example - Group 'C' (cont.)



Week 8 Topic:
Tax Increase Example - Tax Scheme

General Scheme for Tax Increases:

- Households currently making less than 100k will incur no tax increases
- Households currently paying taxes in the 22 (>$1000) ~ 31 (<$2000) range, looking at TAXP values, will pay an additional ***10%*** if making >100k per household
- Households currently paying taxes in the 32 (>$2000) ~ 42 (<$3000) range, looking at TAXP values, will pay an additional ***15%*** if making >100k per household
- Households currently paying taxes in the 42 (>$3000) ~ 61 (<$5000) range, looking at TAXP values, will pay an additional ***20%*** if making >100k per household
- Households currently paying taxes above 61 (>$5000), looking at TAXP values, will pay an additional ***25%*** if making >100k per household

Week 8 Topic:
Tax Increase Example – Tax Agg Table

```
libname census "/folders/myfolders/census";

/*Add TAXP_INCREASE (average of bracket range) column
by using lookup table created from Excel*/

proc sql;

create table census.psam_h17_subset1_cluster_agg3t as

select a.*, b.TAXP_AVG, (b.TAXP_AVG*0.1) as
TAXP_INCREASE
from census.psam_h17_subset1_cluster_agg3 as a left join
census.taxp_lookup as b
on a.TAXP=b.TAXP
WHERE a.HINCP > 100000
and a.TAXP not in ('01','02','03','04','05','06','07','08','09','10'
,'11','12','13','14','15','16','17','18','19','20','21')
and VALP is not null;

run;
```

1) Using the aggregate table for cluster 3 and TAXP lookup table, generate a tax aggregate table with TAXP_AVG and TAXP_INCREASE
   - TAXP_AVG is an average of the TAXP bracket min/max values
   - TAXP_INCREASE is a % of the TAXP_AVG. For the initial field population, we use 10%
2) Filter the new tax aggregate table for households with >100k in earnings, TAXP values above 21, and valid VALP values.

* Note that we use a series of TAXP values to filter as the variable is a text field ($2.)

Week 8 Topic:
Tax Increase Example – Tax Increase


```
/*15% Increase level*/
proc sql;
update census.psam_h17_subset1_cluster_agg3t
set TAXP_INCREASE=(TAXP_AVG*0.15)
where TAXP in ('32','33','34','35','36','37','38','39','40','41');
run;

/*20% Increase level*/
proc sql;
update census.psam_h17_subset1_cluster_agg3t
set TAXP_INCREASE=(TAXP_AVG*0.2)
where TAXP in ('42','43','44','45','46','47','48','49','50',
'51','52','53','54','55','56','57','58','59','60','61');
run;

/*25% Increase level*/
proc sql;
update census.psam_h17_subset1_cluster_agg3t
set TAXP_INCREASE=(TAXP_AVG*0.25)
where TAXP in ('62','63','64','65','66','67','68');
run;

/*Total Tax Increase*/
proc sql;
select sum(TAXP_INCREASE)
from census.psam_h17_subset1_cluster_agg3t;
run;
```

1) Using a series of PROC SQL statements, apply the percent increases according to the tax scheme described in Slide 15
2) Then, SUM the anticipated tax increases for Group C (Clusters 3 & 7) as: ***$1,694,829***

Week 8 Topic:
Week 8 Assignment – Profiling & Tax

(Worth 20 points)

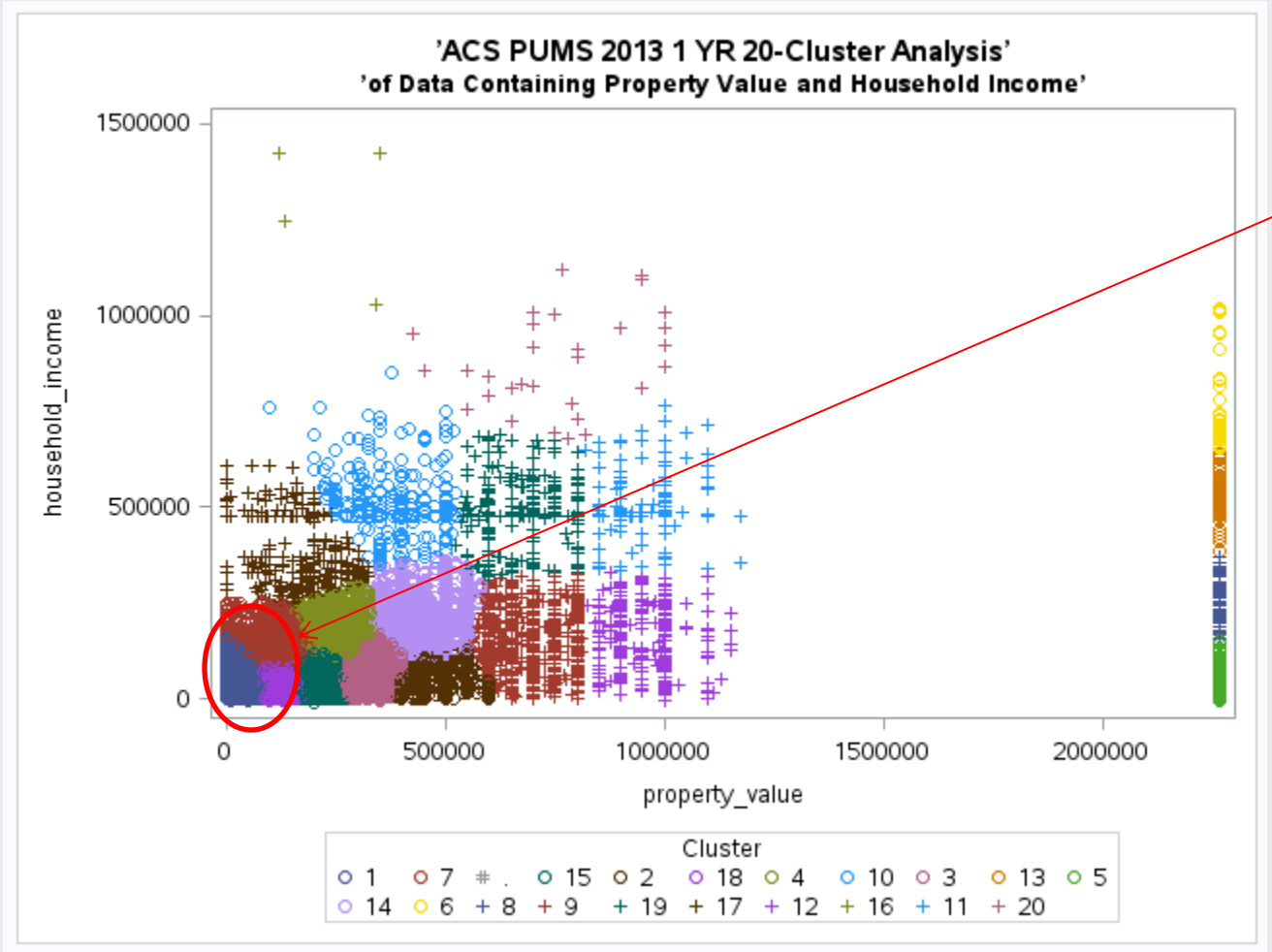
- Fill in the Profile Matrix Spreadsheet following the example in class.
- Determine your own Tax Groups. Similar to the example in class. Provide answers in the same Matrix spreadsheet.
- Determine the tax strategy per Tax Group(s) and provide the overall tax increase amount. Provide answers in the same Matrix workbook.

Week 8 Topic:
Week 8 Assignment - Bonus

(Worth 1.5 points to be added to your previous lowest score, not to exceed 10 once added to that score)

- Run for k=100
- Compare the results with your k value clusters e.g., k=20
- Highlight the differences e.g., compare means, value ranges, and other characteristics
- Determine which set of clusters is better to work with for the tax strategy. State reasons why.

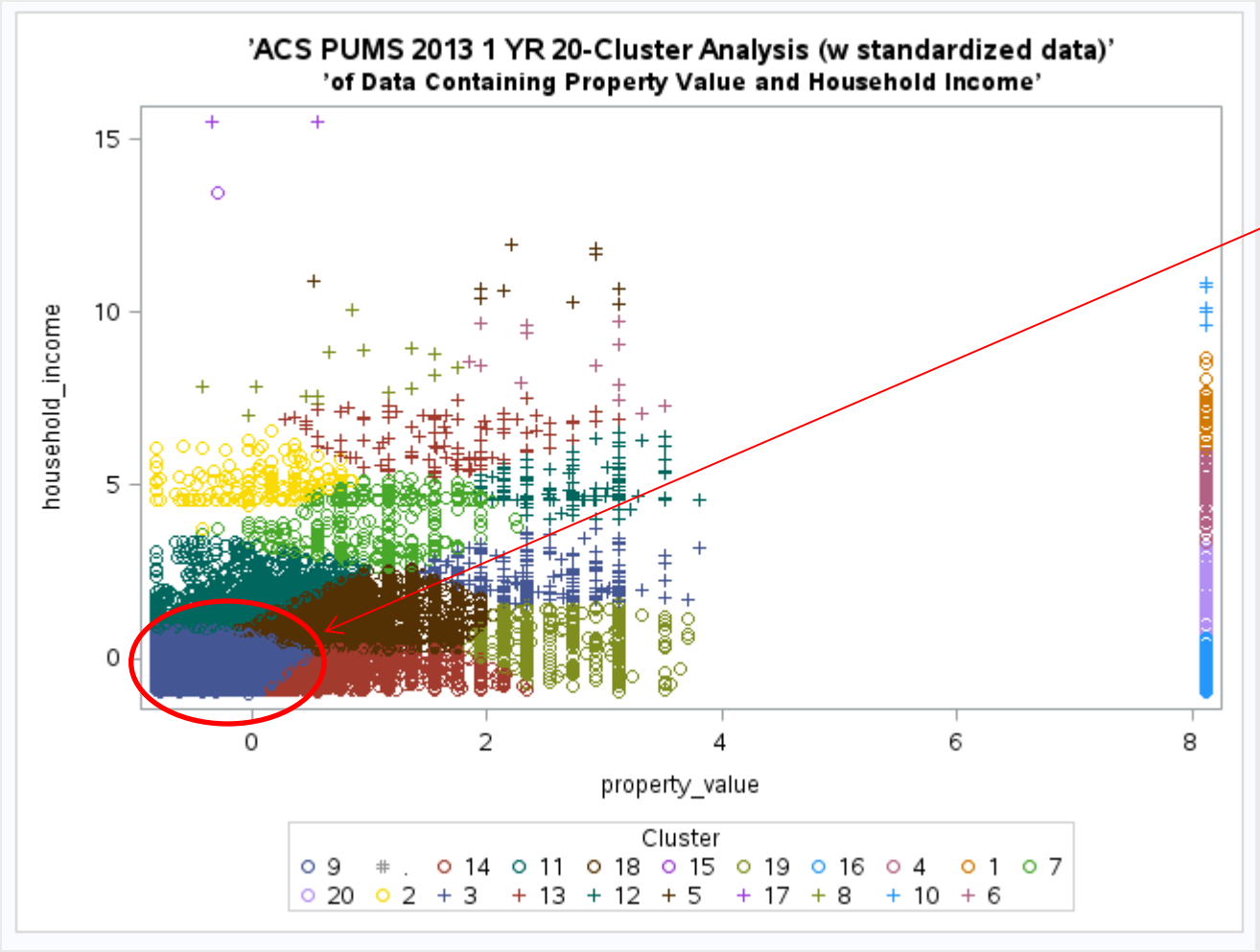
Week 8 Topic: Scatterplot – Cluster 1



Comparison Cluster: 1 from a k=20 PROC CLUSTER run without standardized variables.

ITM - 527

Week 8 Topic: Scatterplot – Cluster 9



Comparison Cluster: 9 from a k=20 PROC CLUSTER run with standardized variables.

ITM - 527

Week 8 Topic: HINCP & VALP comparison

- Note that Cluster 9 from the standardized variable run has almost twice as many members in its cluster. This will additional dilute the cluster to have a broader range of values hence making the cluster hard to profile. The VALP range for this cluster is too broad.

- From Cluster 1 w non-standardized variables:

Descriptive Statistics for Numeric Variables

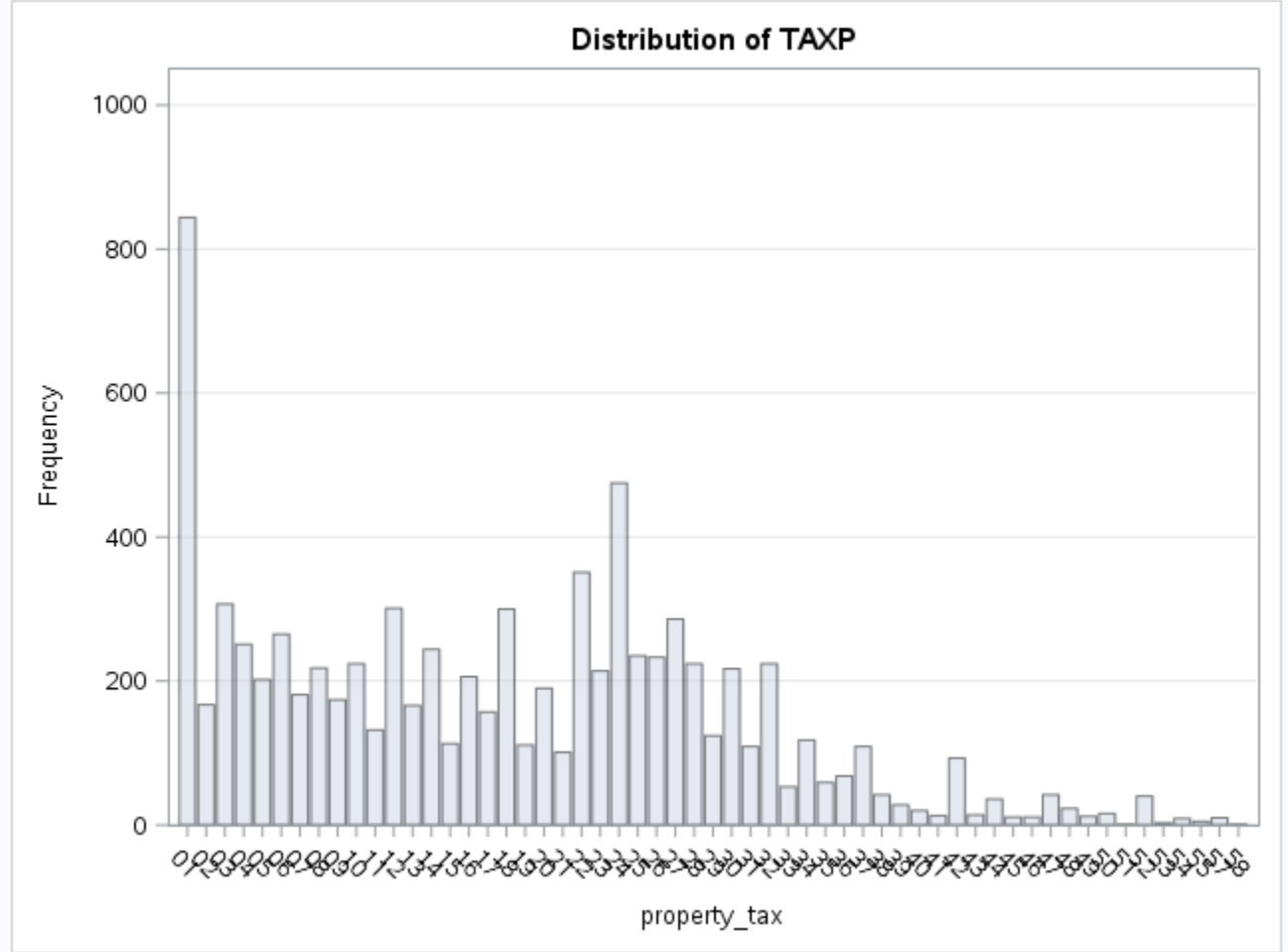
Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
HINCP	household_income	10278	182	-5100.00	41993.46	37800.00	165000.00	24694.98
VALP	property_value	8565	1895	110.0000000	46483.70	50000.00	85000.00	25161.08

- From Cluster 9 w standardized variables:

Descriptive Statistics for Numeric Variables

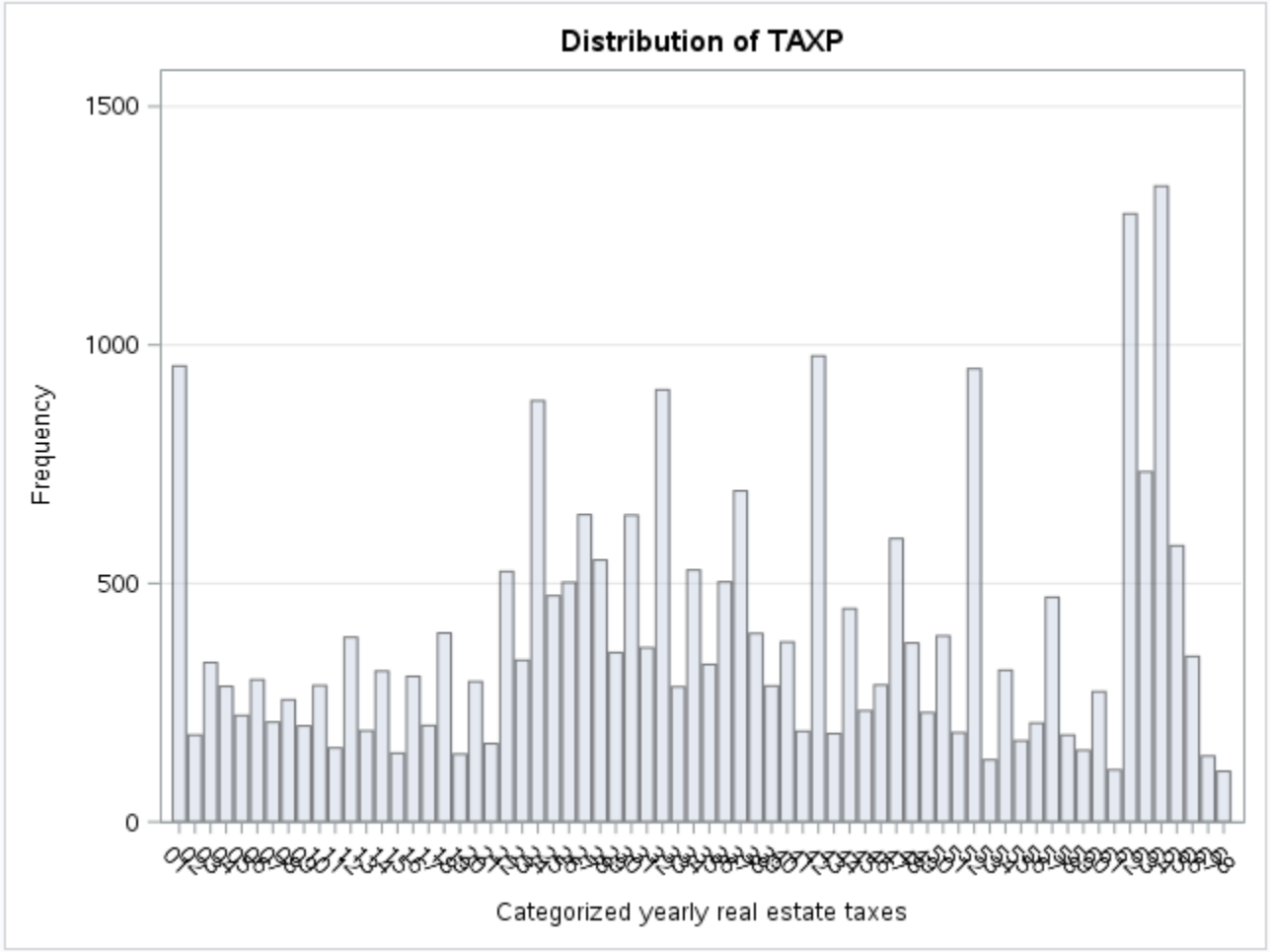
Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
HINCP	PUMS Household income	29109	318	-11200.00	61404.84	60000.00	152000.00	33522.10
VALP	PUMS Property value	27389	2038	110.0000000	124710.97	120000.00	340000.00	72543.29

Week 8 Topic:
TAXP Distribution for Cluster 1



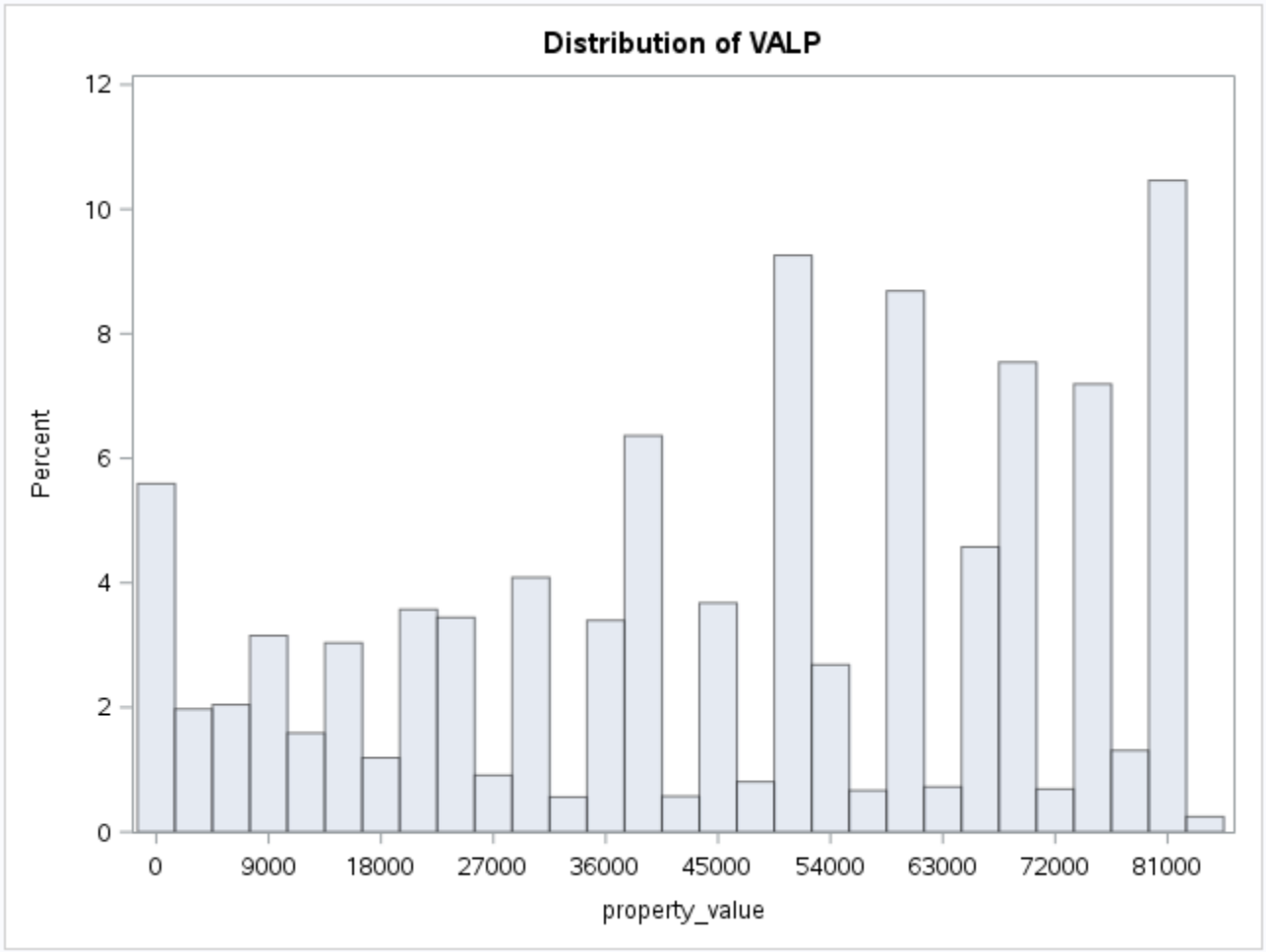
The concentration of TAXP values in the lower range makes this cluster profile-able

Week 8 Topic:
TAXP Distribution for Cluster 9



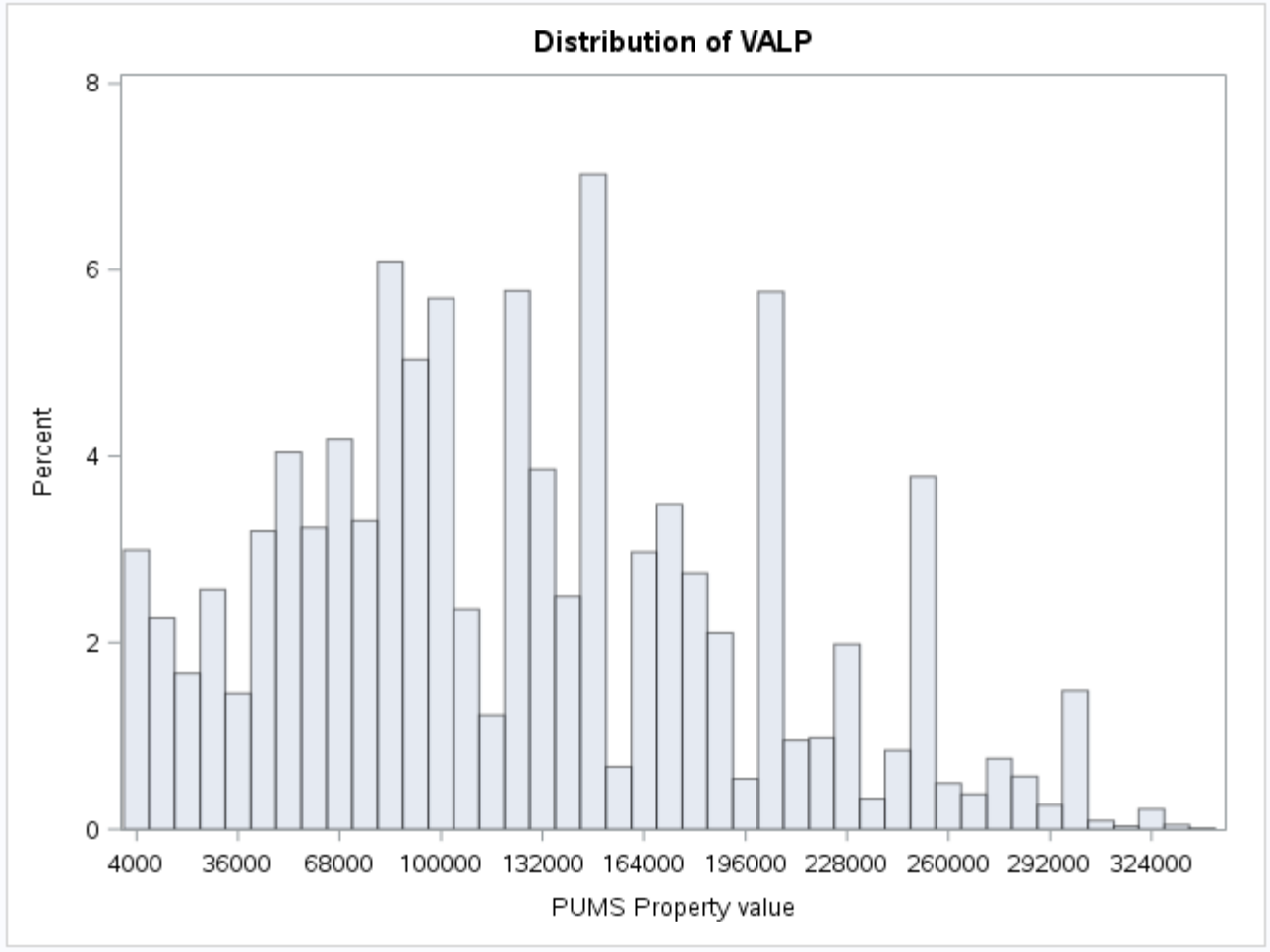
The broad range of TAXP values spanning all levels with multiple peaks and valleys make this cluster hard to profile

Week 8 Topic:
VALP Distribution for Cluster 1



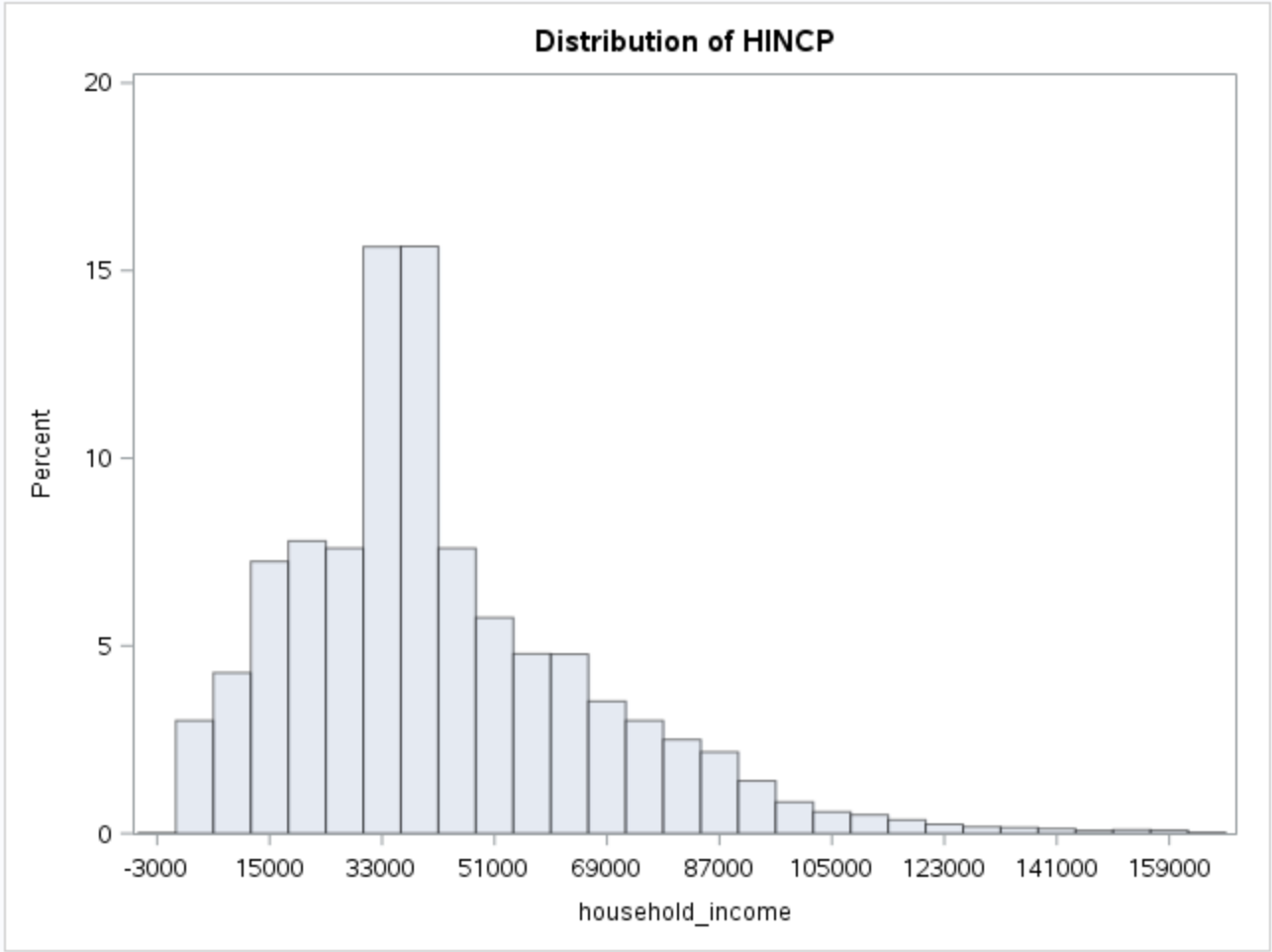
Property values are ~<100k which can be use to profile this cluster as 'Low' for VALP

Week 8 Topic:
VALP Distribution for Cluster 9



Property value range is too broad spanning multiple category levels for VALP, making it difficult to profile

Week 8 Topic:
HINCP Distribution for Cluster 1



Most the household income stays below <100k which puts this cluster in the 'Low' to 'Mid' category for HINCP

ITM - 527

Week 8 Topic:
Cluster 1 vs Cluster 9 - conclusion

- Even though Cluster 1 is dense with >10k in members, it is profile-able as 'Low' to 'Mid' in both VALP and HINCP categories.
- Cluster 9, with >20k members, is difficult to profile with the HINCP and VALP values spanning too broad a range.
- If using clusters with standardized variables, additional clustering is needed for Cluster 9.