IIT School of Applied Technology

ILLINOIS INSTITUTE OF TECHNOLOGY

information technology & management

527 Data Analytics

February 23,24 2016

Week 7 Presentation

Week 7 Topic: Agenda

- Revised Syllabus
- Wine 3 Cluster in SAS
- Tax Strategy Project Overview

Week 7 Topic: Revised Syllabus

Session	Date	Topic	Reading
1	January 12/14		Week 1 topics: Course Overview
2	January 19/21		Week 2 topics: Analysis in Excel/Basic Statistics
3	January 26/28		Week 3 topics: Analysis in Excel/Optimization Modeling I
4	February 2/4		Week 4 topics: Analysis in Excel/Optimization Modeling II
5	February 9/11		Week 5 topics: Cluster Analysis (in Excel) I
6	February 16/18		Week 6 topics: Cluster Analysis (in SAS) II
7	February 23/25		Week 7 topics: Cluster Analysis (Case Study in SAS) I
8	March 1/3		Week 8 topics: Cluster Analysis (Case Study in SAS) II
9	March 8/10		Week 9 topics: Midterm Quiz/Case Study Wrap up
10	March 15/17		No classes/Spring Break
11	March 22/24		Week 11 topics: Big Data Analytics I
12	March 29/31		No classes (Self Study Assignments)
13	April 5/7		Week 13 topics: Special Topic: Kx Systems
14	April 12/14		Week 14 topics: Big Data Analytics II
15	April 19/21		Week 15 topics: Big Data Analytics III
16	April 26/28		Week 16 topics: Final Reviews
Finals	May3 Week		Final Exam

Week 7 Topic: Revised Syllabus

The midterm grade for the class will be calculated as follows:

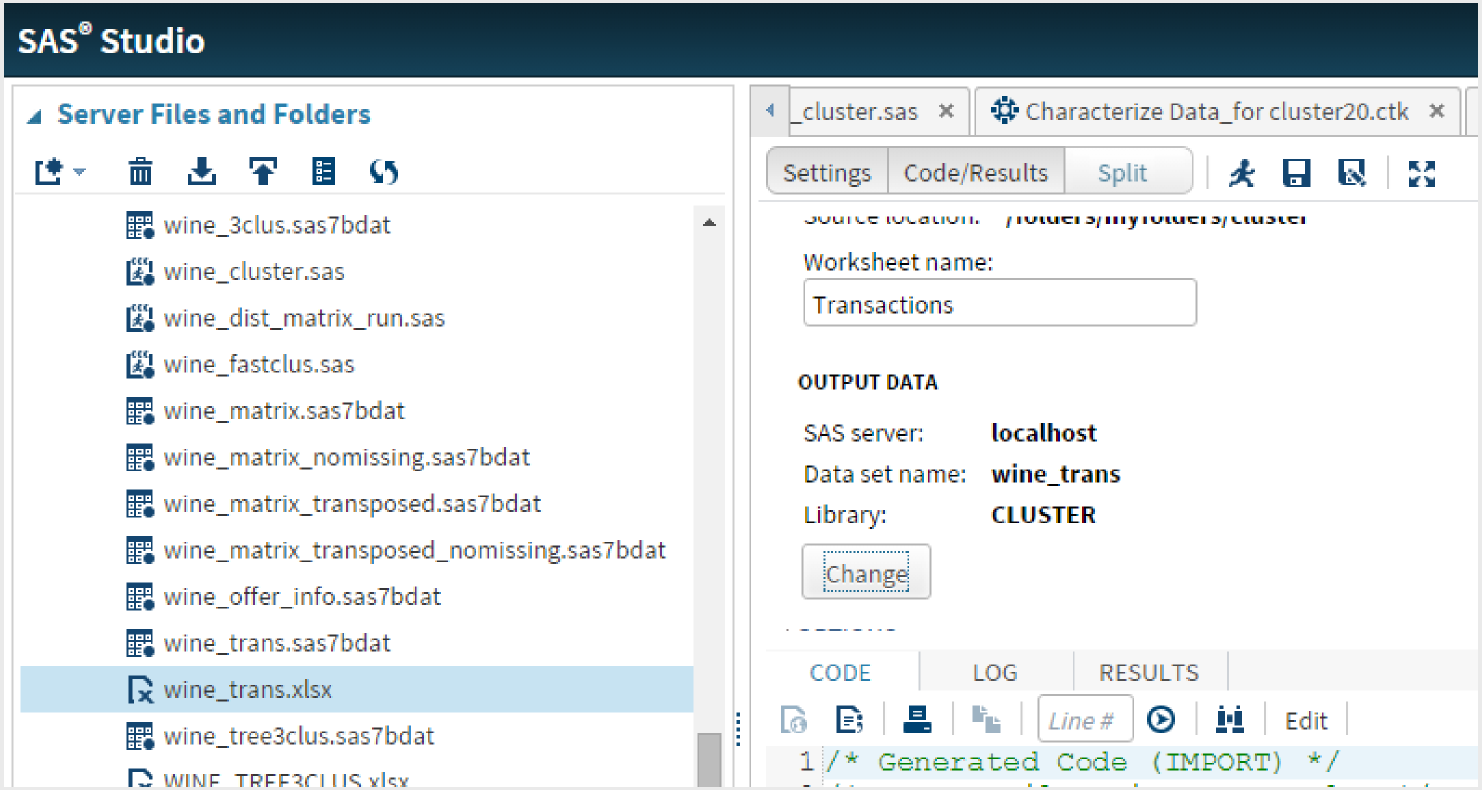
Midterm Quiz	**30%**
Class Exercises & Participation	**70%**

The final grade for the class will be calculated as follows:

Final Exam	**30%**
Midterm Quiz, Class Exercises & Participation	**70%**

Week 7 Topic:
Wine example in SAS: Part 1 – imports

1) Copy WINEKMC.xls into SAS folder
2) Follow import directions and create the following tables in SAS: wine_trans, wine_matrix, and wine_offer_info



- wine_3clus.sas7bdat
- wine_cluster.sas
- wine_dist_matrix_run.sas
- wine_fastclus.sas
- wine_matrix.sas7bdat
- wine_matrix_nomissing.sas7bdat
- wine_matrix_transposed.sas7bdat
- wine_matrix_transposed_nomissing.sas7bdat
- wine_offer_info.sas7bdat
- wine_trans.sas7bdat
- wine_trans.xlsx
- wine_tree3clus.sas7bdat
- WINE_TREE3CLUS.xlsx

Source location: /folders/myfolders/cluster

Worksheet name: Transactions

OUTPUT DATA

SAS server: localhost
Data set name: wine_trans
Library: CLUSTER

Change

```
/* Generated Code (IMPORT) */
```

Week 7 Topic:
Wine example in SAS: Part 2 - prep

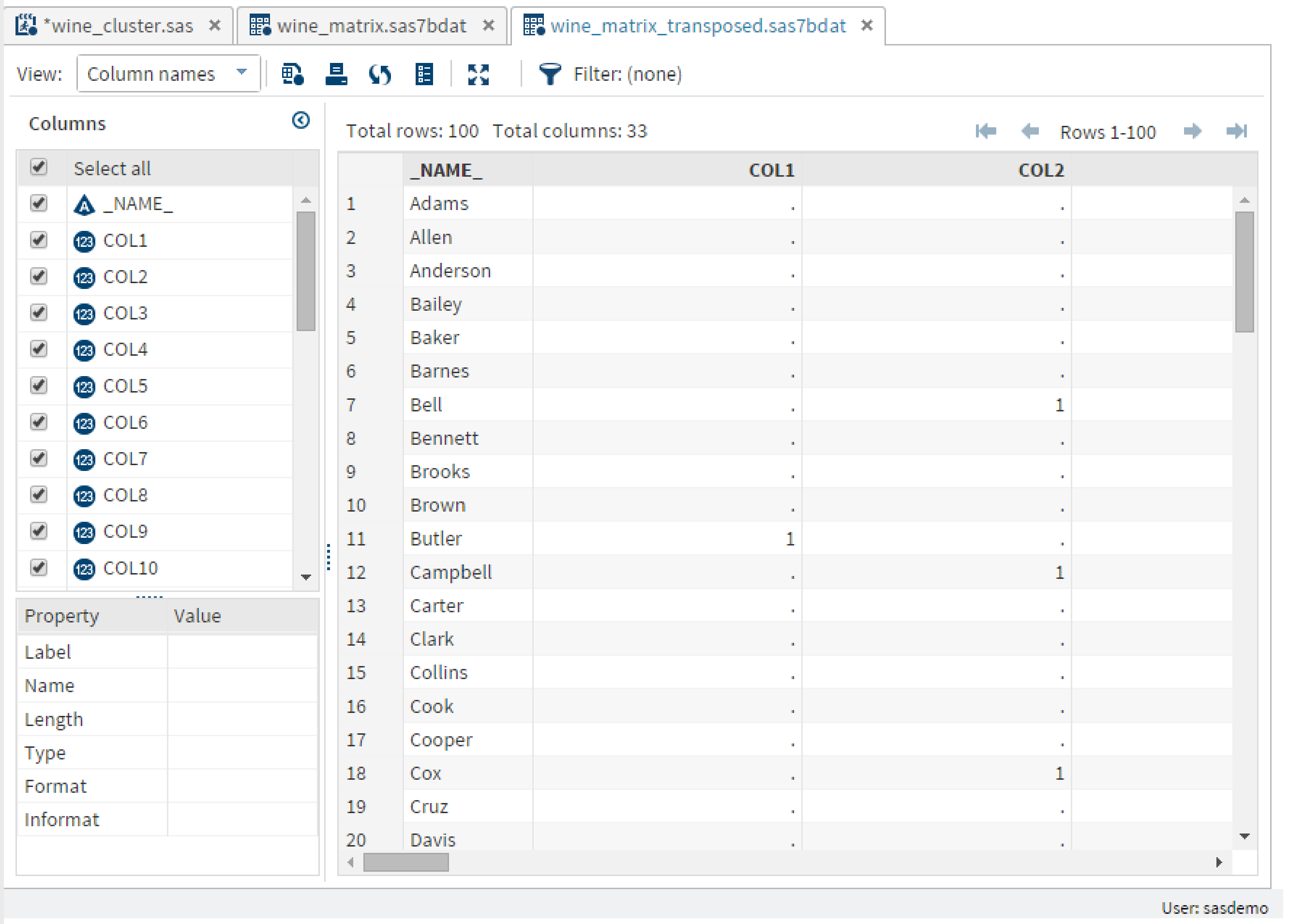
Create wine_transposed table (input data) for the PROC CLUSTER run by:

1) Transposing wine_trans
2) Rid of unwanted rows and columns

```
libname cluster "/folders/myfolders/cluster";

/*Start with imported wine_matrix from xls
 Transpose the matrix and save to a new file*/
proc transpose
data=cluster.wine_matrix
/*Delete the extra column created by the transpose step*/
out=cluster.wine_matrix_transposed(drop=_label_);
run;

/*Delete the extra row created by the transpose step*/
proc sql;
   delete
      from cluster.wine_matrix_transposed
      where _NAME_ like 'Offer%';
run;

proc print data=cluster.wine_matrix_transposed noobs;
run;
```

Week 7 Topic:
Wine example in SAS: Part 2 – outputs

wine_transposed:



Total rows: 100 Total columns: 33

Rows 1-100

☑ Select all
☑ _NAME_
☑ COL1
☑ COL2
☑ COL3
☑ COL4
☑ COL5
☑ COL6
☑ COL7
☑ COL8
☑ COL9
☑ COL10

Property	Value
Label	
Name	
Length	
Type	
Format	
Informat	

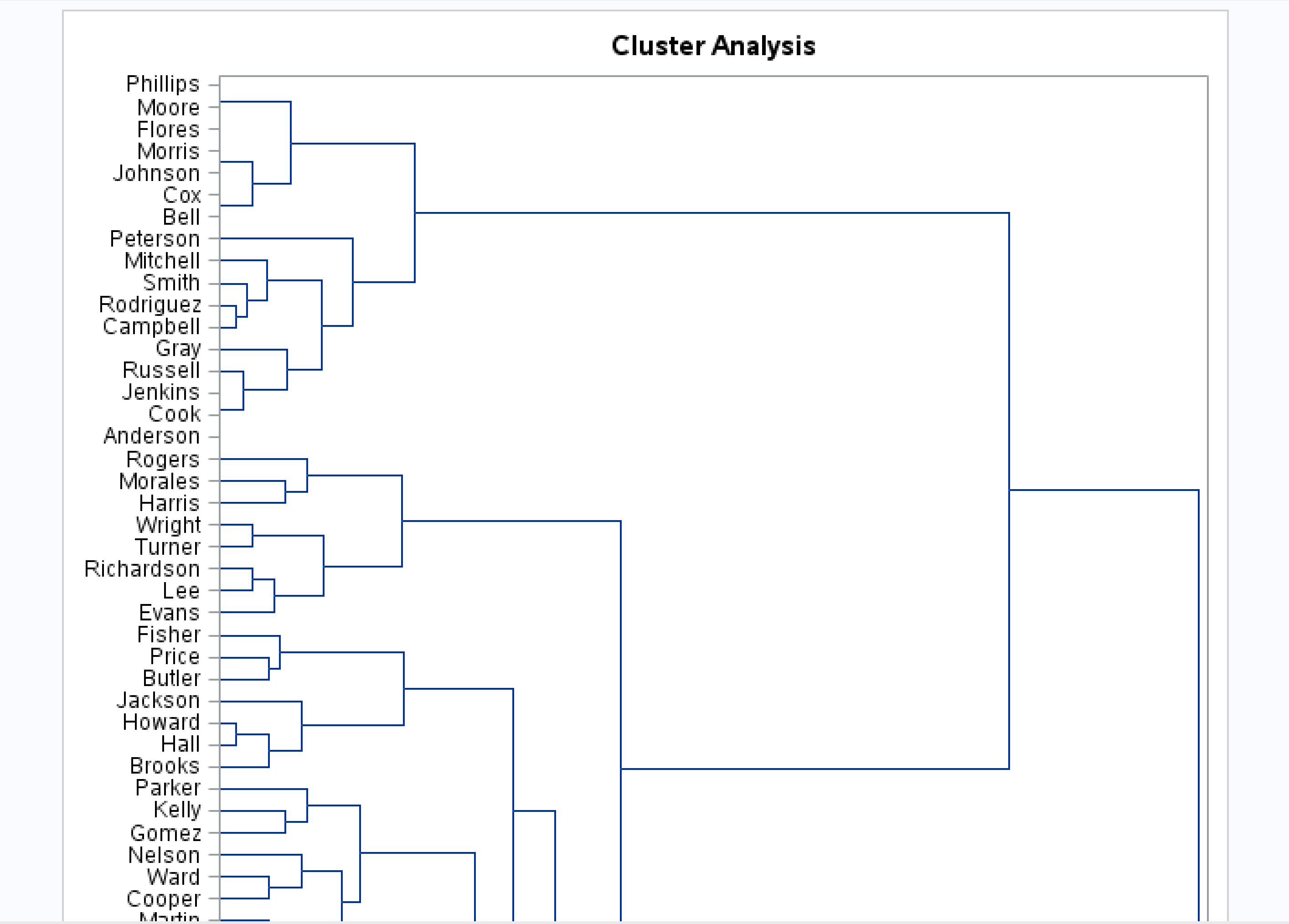
	NAME	COL1	COL2
1	Adams	.	.
2	Allen	.	.
3	Anderson	.	.
4	Bailey	.	.
5	Baker	.	.
6	Barnes	.	.
7	Bell	.	1
8	Bennett	.	.
9	Brooks	.	.
10	Brown	.	.
11	Butler	1	.
12	Campbell	.	1
13	Carter	.	.
14	Clark	.	.
15	Collins	.	.
16	Cook	.	.
17	Cooper	.	.
18	Cox	.	1
19	Cruz	.	.
20	Davis	.	.

Week 7 Topic:
Wine example in SAS: Part 3 - clustering

1) Run PROC CLUSTER: clustering is to be performed on the transposed matrix, it's a small data set and we already know the optimum k value of 3, which is small.
2) We transpose the data so that each row is an object to the cluster.
3) We fill missing cells with '0' to perform the distance calculations.

```
libname cluster "/folders/myfolders/cluster";

data cluster.wine_matrix_transposed_nomissing;
   set cluster.wine_matrix_transposed;
   array change _numeric_;
            do over change;
            if change=. then change=0;
            end;
run;

proc cluster data=cluster.wine_matrix_transposed_nomissing method=ward
outtree=cluster.wine_3clus noeigen simple;
   var _1 -- _32;
   id name;
run;

proc tree data=cluster.wine_3clus out=cluster.wine_tree3clus nclusters=3 noprint;
copy name;
run;
```

Week 7 Topic:
Wine example in SAS: Part 3 - output



Week 7 Topic:
Wine example in SAS: Part 4 - join

Add descriptive information (by joining wine_offer and wine_trans) to profile the cluster results:

```
/*Sort by cluster*/
proc sort; by cluster;

/*Join transaction information using name*/
proc sql;
create table cluster.wine_3clus_trans as

select *
from cluster.wine_tree3clus as a left join cluster.wine_trans as b
on a._NAME_=b.Customer_Last_Name;
run;

/*Join offer information using name*/
proc sql;
create table cluster.wine_3clus_results as

select *
from cluster.wine_3clus_trans as a left join cluster.wine_offer_info as b
on a.Offer__=b.Offer__;
run;
```

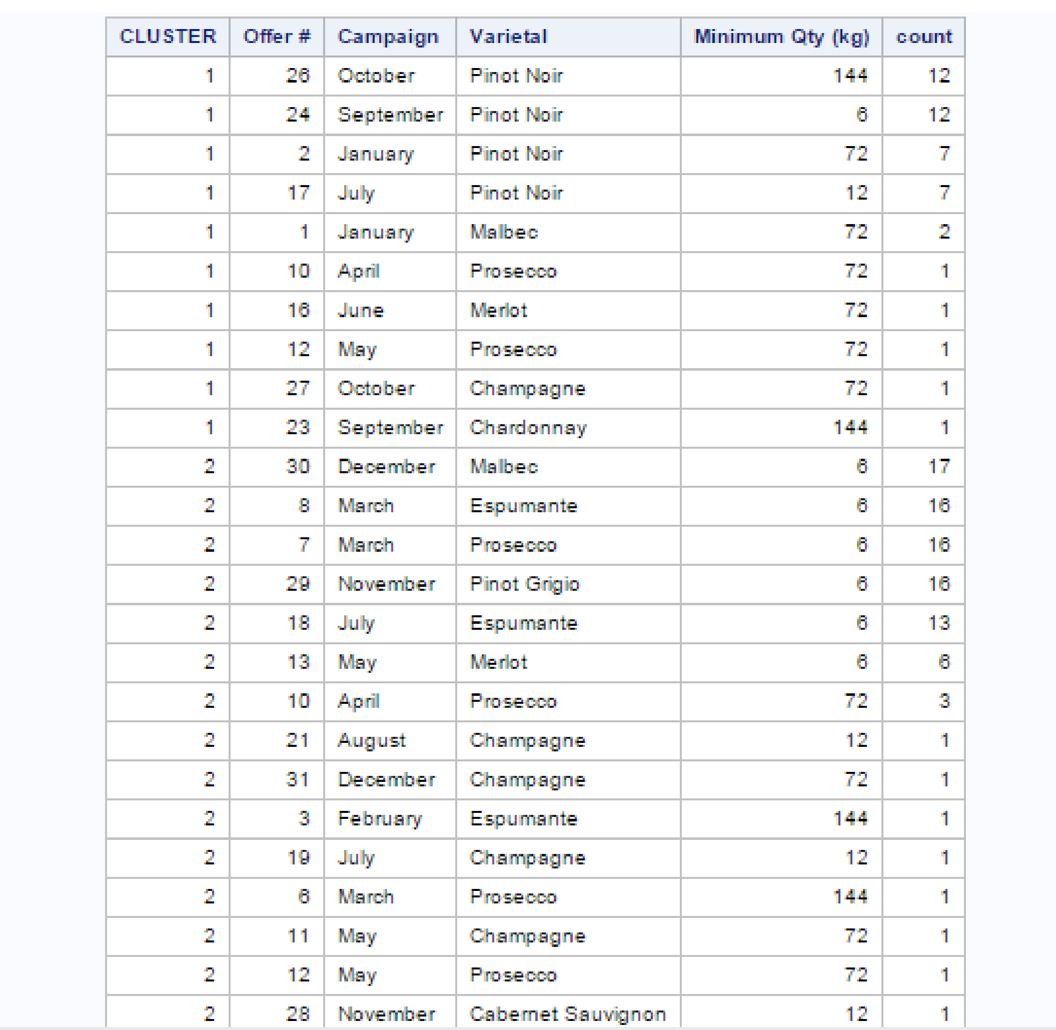
Week 7 Topic:
Wine example in SAS: Part 5 - counts

Get counts of offer by cluster:

```
proc sort ; by cluster _NAME_;

/*Get counts of cluster by offer*/
proc sql;
select distinct CLUSTER, Offer__, Campaign, Varietal, Minimum_Qty__kg_,
count(trans_num) as count
from cluster.wine_3clus_results
group by CLUSTER, Offer__
order by CLUSTER, count DESC;
run;
```

Week 7 Topic:
Wine example in SAS: profile results

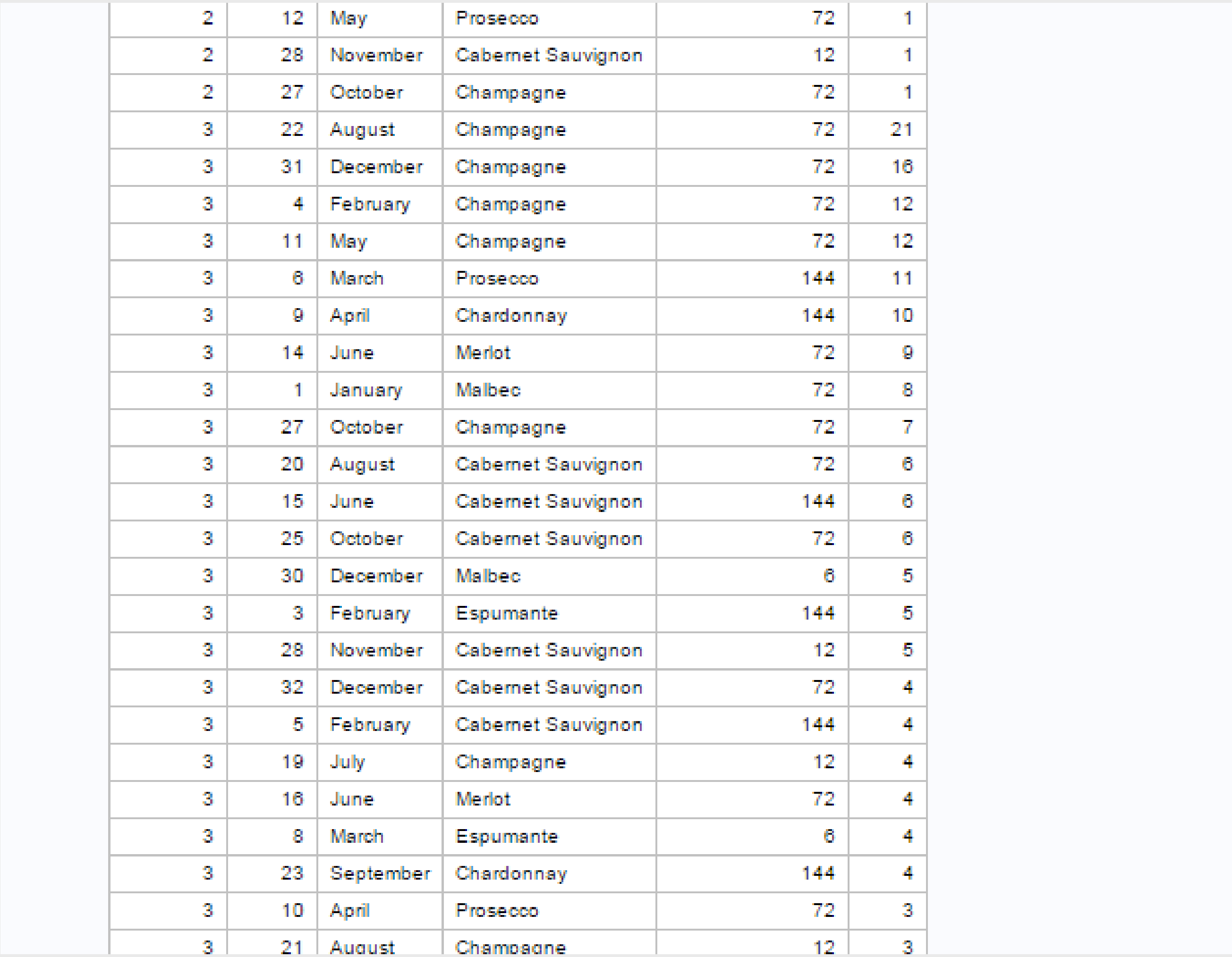
In SAS:



CLUSTER	Offer #	Campaign	Varietal	Minimum Qty (kg)	count
1	26	October	Pinot Noir	144	12
1	24	September	Pinot Noir	6	12
1	2	January	Pinot Noir	72	7
1	17	July	Pinot Noir	12	7
1	1	January	Malbec	72	2
1	10	April	Prosecco	72	1
1	16	June	Merlot	72	1
1	12	May	Prosecco	72	1
1	27	October	Champagne	72	1
1	23	September	Chardonnay	144	1
2	30	December	Malbec	6	17
2	8	March	Espumante	6	16
2	7	March	Prosecco	6	16
2	29	November	Pinot Grigio	6	16
2	18	July	Espumante	6	13
2	13	May	Merlot	6	6
2	10	April	Prosecco	72	3
2	21	August	Champagne	12	1
2	31	December	Champagne	72	1
2	3	February	Espumante	144	1
2	19	July	Champagne	12	1
2	6	March	Prosecco	144	1
2	11	May	Champagne	72	1
2	12	May	Prosecco	72	1
2	28	November	Cabernet Sauvignon	12	1

Week 7 Topic:
Wine example in SAS: profile results

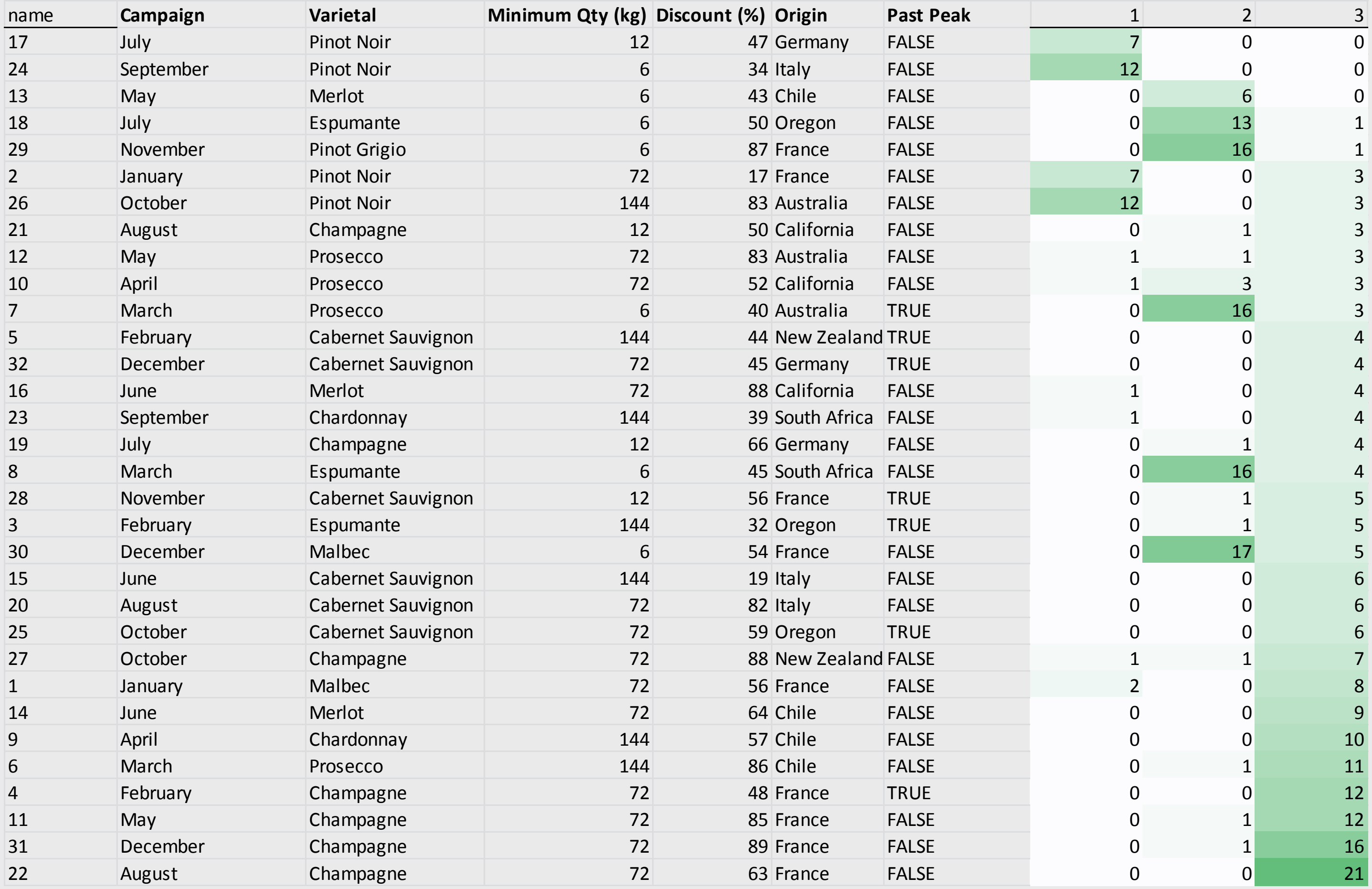
In SAS:



2	12	May	Prosecco	72	1
2	28	November	Cabernet Sauvignon	12	1
2	27	October	Champagne	72	1
3	22	August	Champagne	72	21
3	31	December	Champagne	72	16
3	4	February	Champagne	72	12
3	11	May	Champagne	72	12
3	6	March	Prosecco	144	11
3	9	April	Chardonnay	144	10
3	14	June	Merlot	72	9
3	1	January	Malbec	72	8
3	27	October	Champagne	72	7
3	20	August	Cabernet Sauvignon	72	6
3	15	June	Cabernet Sauvignon	144	6
3	25	October	Cabernet Sauvignon	72	6
3	30	December	Malbec	6	5
3	3	February	Espumante	144	5
3	28	November	Cabernet Sauvignon	12	5
3	32	December	Cabernet Sauvignon	72	4
3	5	February	Cabernet Sauvignon	144	4
3	19	July	Champagne	12	4
3	16	June	Merlot	72	4
3	8	March	Espumante	6	4
3	23	September	Chardonnay	144	4
3	10	April	Prosecco	72	3
3	21	August	Champagne	12	3

Week 7 Topic:
Wine example in SAS: profile results

We port the results into Excel and confirm that the cluster profiles match the Solver results for k = 3:



name	Campaign	Varietal	Minimum Qty (kg)	Discount (%)	Origin	Past Peak	1	2	3
17	July	Pinot Noir	12	47	Germany	FALSE	7	0	0
24	September	Pinot Noir	6	34	Italy	FALSE	12	0	0
13	May	Merlot	6	43	Chile	FALSE	0	6	0
18	July	Espumante	6	50	Oregon	FALSE	0	13	1
29	November	Pinot Grigio	6	87	France	FALSE	0	16	1
2	January	Pinot Noir	72	17	France	FALSE	7	0	3
26	October	Pinot Noir	144	83	Australia	FALSE	12	0	3
21	August	Champagne	12	50	California	FALSE	0	1	3
12	May	Prosecco	72	83	Australia	FALSE	1	1	3
10	April	Prosecco	72	52	California	FALSE	1	3	3
7	March	Prosecco	6	40	Australia	TRUE	0	16	3
5	February	Cabernet Sauvignon	144	44	New Zealand	TRUE	0	0	4
32	December	Cabernet Sauvignon	72	45	Germany	TRUE	0	0	4
16	June	Merlot	72	88	California	FALSE	1	0	4
23	September	Chardonnay	144	39	South Africa	FALSE	1	0	4
19	July	Champagne	12	66	Germany	FALSE	0	1	4
8	March	Espumante	6	45	South Africa	FALSE	0	16	4
28	November	Cabernet Sauvignon	12	56	France	TRUE	0	1	5
3	February	Espumante	144	32	Oregon	TRUE	0	1	5
30	December	Malbec	6	54	France	FALSE	0	17	5
15	June	Cabernet Sauvignon	144	19	Italy	FALSE	0	0	6
20	August	Cabernet Sauvignon	72	82	Italy	FALSE	0	0	6
25	October	Cabernet Sauvignon	72	59	Oregon	TRUE	0	0	6
27	October	Champagne	72	88	New Zealand	FALSE	1	1	7
1	January	Malbec	72	56	France	FALSE	2	0	8
14	June	Merlot	72	64	Chile	FALSE	0	0	9
9	April	Chardonnay	144	57	Chile	FALSE	0	0	10
6	March	Prosecco	144	86	Chile	FALSE	0	1	11
4	February	Champagne	72	48	France	TRUE	0	0	12
11	May	Champagne	72	85	France	FALSE	0	1	12
31	December	Champagne	72	89	France	FALSE	0	1	16
22	August	Champagne	72	63	France	FALSE	0	0	21

Week 7 Topic:
Wine example in SAS: FASTCLUS?

Running FASTCLUS may also seem valid but because we have such a small data set with binary information (customer either takes an offer or not), the results are inconclusive and highly tied to the order of the observations. You can try the following code for this. Because ordering matters greatly for this data set, if you run PROC CLUSTER on wine_trans instead of the matrix data, you will get similar results yielding inconclusive clusters :

```
libname cluster "/folders/myfolders/cluster";

proc fastclus data=cluster.wine_trans radius=0 replace=full maxclusters=3
converge=0 maxiter=100 OUTSTAT=cluster.wine_3clusters_stat OUT=cluster.wine_3clusters
list distance;
ID trans_num;
var artificial Offer__;
run;

proc sql;
title 'Cluster information for profiling';
create table cluster.wine_3clusters_results as
select *
from cluster.wine_3clusters as a left join cluster.wine_offer_info as b
on a.Offer__=b.Offer__;
run;

proc sql;
title '3 Cluster information for profiling';
select distinct CLUSTER, DISTANCE, Offer__, Campaign, Varietal, Minimum_Qty__kg_,
count(CLUSTER) as count
from cluster.wine_3clusters_results
group by Offer__
order by CLUSTER, count DESC;
run;
```

Week 7 Topic:
Tax Strategy: Data

AMERICAN COMMUNITY SURVEY 2013 ACS 1-YEAR PUMS FILES

Prepared by:

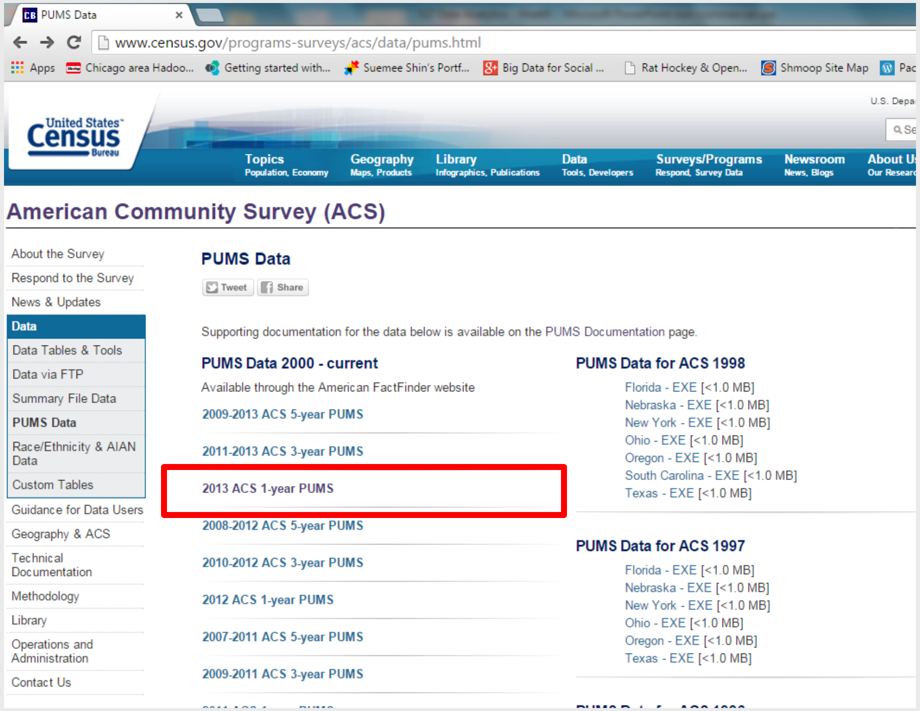
American Community Survey Office

U.S. Census Bureau

January 15, 2015

PUMS Data Link:

http://www.census.gov/programs-surveys/acs/data/pums.html



American Community Survey (ACS)

- About the Survey
- Respond to the Survey
- News & Updates
- Data
- Data Tables & Tools
- Data via FTP
- Summary File Data
- PUMS Data
- Race/Ethnicity & AIAN Data
- Custom Tables
- Guidance for Data Users
- Geography & ACS
- Technical Documentation
- Methodology
- Library
- Operations and Administration
- Contact Us

PUMS Data

Tweet Share

Supporting documentation for the data below is available on the PUMS Documentation page.

PUMS Data 2000 - current

Available through the American FactFinder website

- 2009-2013 ACS 5-year PUMS
- 2011-2013 ACS 3-year PUMS
- 2013 ACS 1-year PUMS
- 2008-2012 ACS 5-year PUMS
- 2010-2012 ACS 3-year PUMS
- 2012 ACS 1-year PUMS
- 2007-2011 ACS 5-year PUMS
- 2009-2011 ACS 3-year PUMS

PUMS Data for ACS 1998

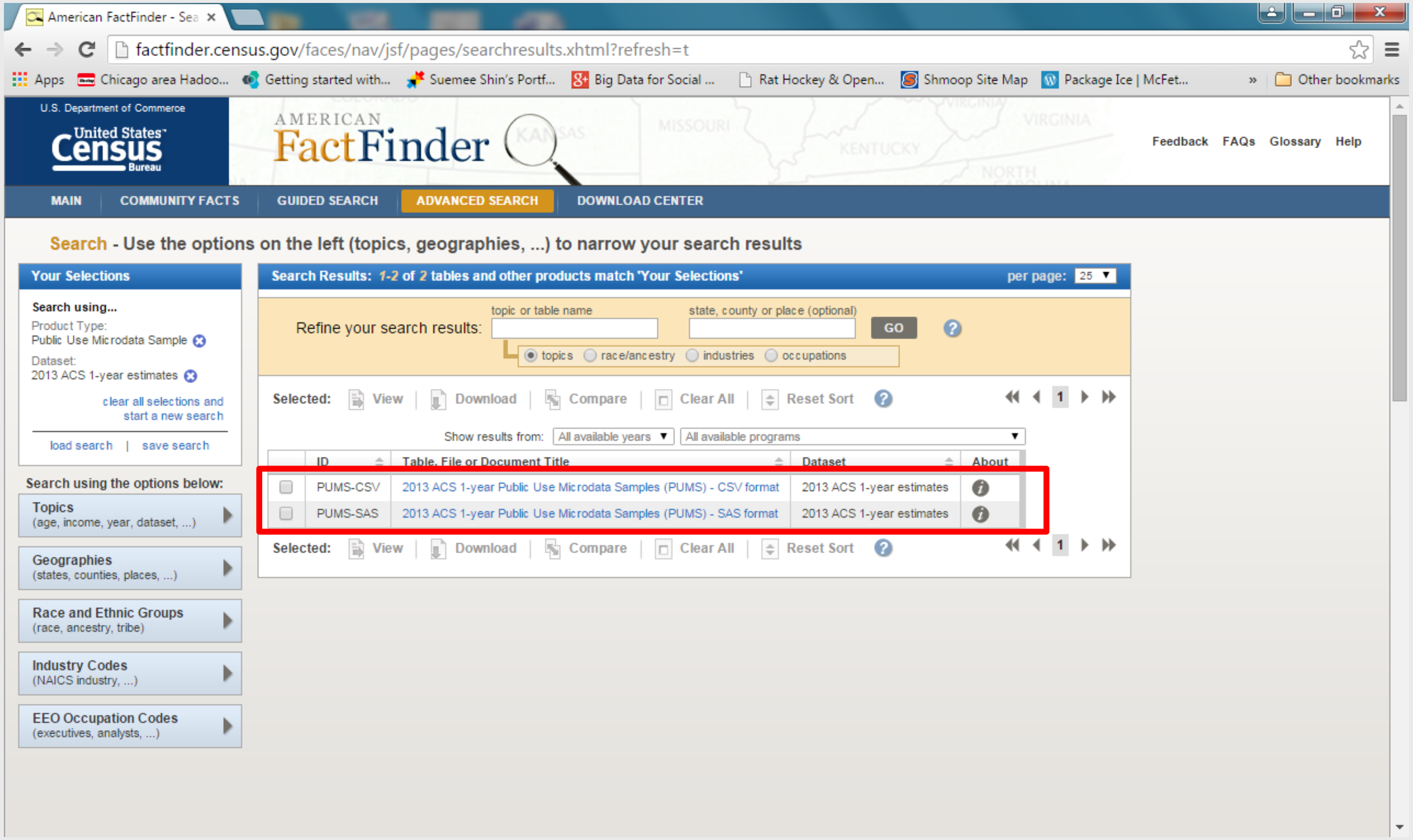
- Florida - EXE [<1.0 MB]
- Nebraska - EXE [<1.0 MB]
- New York - EXE [<1.0 MB]
- Ohio - EXE [<1.0 MB]
- Oregon - EXE [<1.0 MB]
- South Carolina - EXE [<1.0 MB]
- Texas - EXE [<1.0 MB]

PUMS Data for ACS 1997

- Florida - EXE [<1.0 MB]
- Nebraska - EXE [<1.0 MB]
- New York - EXE [<1.0 MB]
- Ohio - EXE [<1.0 MB]
- Oregon - EXE [<1.0 MB]
- Texas - EXE [<1.0 MB]

Week 7 Topic:
Tax Strategy: File Formats

- If using SAS, copy the SAS file into your SAS data folder to start
- If using other than SAS, upload the CSV file



American FactFinder - Sea
factfinder.census.gov/faces/nav/jsf/pages/searchresults.xhtml?refresh=t
Apps | Chicago area Hadoo... | Getting started with... | Suemee Shin's Portf... | Big Data for Social ... | Rat Hockey & Open... | Shmoop Site Map | Package Ice | McFet... | Other bookmarks
U.S. Department of Commerce
United States Census Bureau
AMERICAN FactFinder
Feedback FAQs Glossary Help
MAIN | COMMUNITY FACTS | GUIDED SEARCH | ADVANCED SEARCH | DOWNLOAD CENTER

Search - Use the options on the left (topics, geographies, ...) to narrow your search results

Your Selections

Search using...
Product Type:
Public Use Microdata Sample
Dataset:
2013 ACS 1-year estimates

clear all selections and start a new search

load search | save search

Search using the options below:

- Topics (age, income, year, dataset, ...)
- Geographies (states, counties, places, ...)
- Race and Ethnic Groups (race, ancestry, tribe)
- Industry Codes (NAICS industry, ...)
- EEO Occupation Codes (executives, analysts, ...)

Search Results: 1-2 of 2 tables and other products match 'Your Selections' per page: 25

Refine your search results: topic or table name ___ state, county or place (optional) ___ GO

● topics ○ race/ancestry ○ industries ○ occupations

Selected: View | Download | Compare | Clear All | Reset Sort

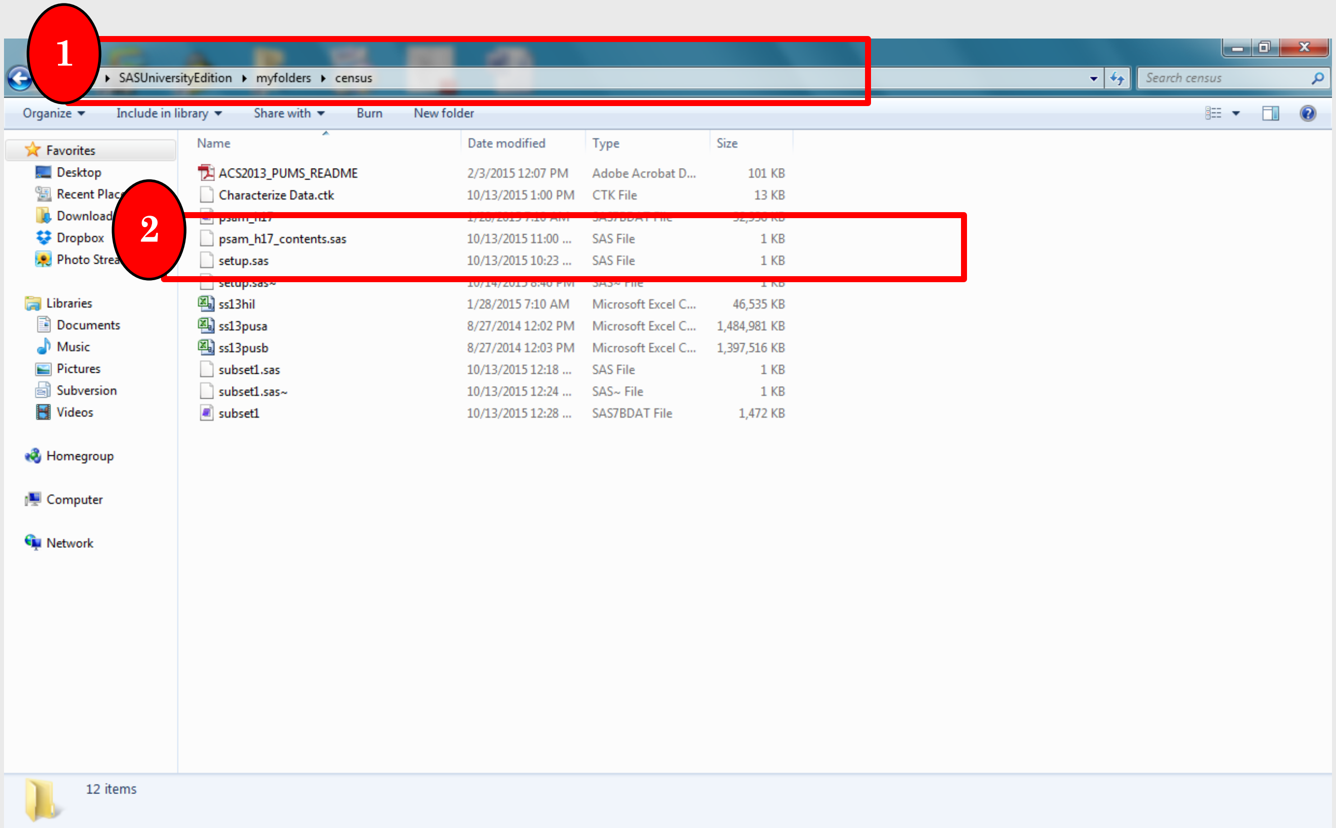
Show results from: All available years | All available programs

	ID	Table, File or Document Title	Dataset	About
□	PUMS-CSV	2013 ACS 1-year Public Use Microdata Samples (PUMS) - CSV format	2013 ACS 1-year estimates	i
□	PUMS-SAS	2013 ACS 1-year Public Use Microdata Samples (PUMS) - SAS format	2013 ACS 1-year estimates	i

Selected: View | Download | Compare | Clear All | Reset Sort

Week 7 Topic:
Tax Strategy: Folder set up

- Create a census folder in your SAS data folder
- Copy the SAS census download file into your SAS data folder



Name	Date modified	Type	Size
ACS2013_PUMS_README	2/3/2015 12:07 PM	Adobe Acrobat D...	101 KB
Characterize Data.ctk	10/13/2015 1:00 PM	CTK File	13 KB
psam_h17	1/28/2015 7:10 AM	SAS7BDAT File	[illegible]
psam_h17_contents.sas	10/13/2015 11:00 ...	SAS File	1 KB
setup.sas	10/13/2015 10:23 ...	SAS File	1 KB
setup.sas~	10/14/2015 8:46 PM	SAS~ File	1 KB
ss13hil	1/28/2015 7:10 AM	Microsoft Excel C...	46,535 KB
ss13pusa	8/27/2014 12:02 PM	Microsoft Excel C...	1,484,981 KB
ss13pusb	8/27/2014 12:03 PM	Microsoft Excel C...	1,397,516 KB
subset1.sas	10/13/2015 12:18 ...	SAS File	1 KB
subset1.sas~	10/13/2015 12:24 ...	SAS~ File	1 KB
subset1	10/13/2015 12:28 ...	SAS7BDAT File	1,472 KB

Week 7 Topic:
2013 IL ACS PUMS Metadata

Data Dictionary:
http://www2.census.gov/programs-surveys/acs/tech_docs/pums/data_dict/PUMSDataDict13.pdf

Focus on the following variables to start:

- **MRGP:** First mortgage payment (monthly amount)
- **HINCP:** Household income (past 12 months)
- **TAXP:** Property taxes (yearly amount)
- **VALP:** Property value

Week 7 Topic:
Tax Strategy Outline – PART 1

PART 1: Data work to process 2013 IL ACS PUMS data for clustering and analysis. This should include:

a) Treating for outliers
b) Treating for incomplete and/or inaccurate data
c) Selection of variables to be included for clustering – with reasons why
d) Summary of data work per variable used e.g., cleansing, trimming, metadata, etc.
e) Selection of variables for profiling and additional analysis – with reasons why

Week 7 Topic:
Tax Strategy Outline – PART 2

PART 2: Analytic work in SAS and/or Excel. This should include:

a) Histograms of variables and notation of key observations

b) Cluster analysis - to include:
- # of cluster chosen and why
- List any additional clusters from outliers
- Summarize results from cluster analysis run. For SAS, list options chosen and explain results. For Excel, describe run options chosen in Solver.

c) Profiling of clusters – to include:
- Use of additional descriptive variables to further profile cluster
- Tax bracket analysis of cluster
- Strategy for tax increase including optimum value of return per cluster

Week 7 Topic:
Tax Strategy Outline – PART 3

PART 3: Presentation of findings. This should include:

a) Overview/Background
b) Data processing summary
c) Analysis summary
d) Details of analysis:
   - Cluster analysis
   - Profiling of clusters
   - Tax increase strategy
   - Optimum return on strategy
e) Conclusion

Week 7 Topic:
Tax Strategy: Plan of analysis

Step 1: Select your variables for analysis – cluster and profile

- Consider select numeric base variables to start

Step 2: You can start with a small data set to explore clusters:

- 1,000 observations chosen at random from the main data set of observations

Step 3: Run a initial cluster analysis and try to understand the results from the analysis.

- Use 20 clusters to start

Step 4: Modify the analysis based on the initial cluster analysis results. Considerations include:

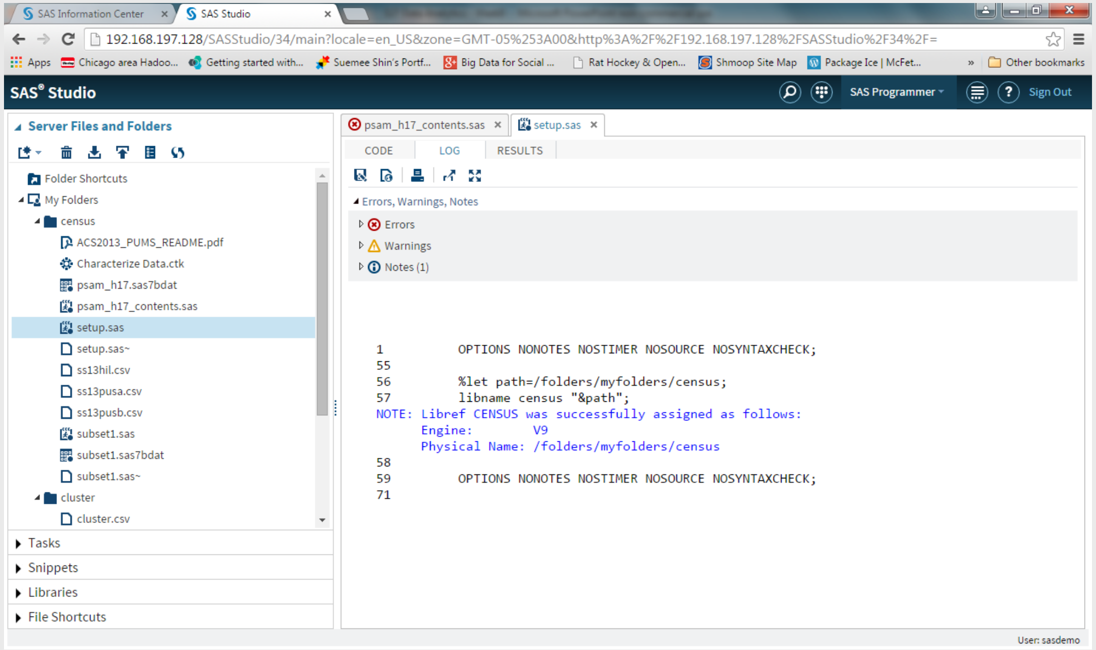
- Standardize variables
- Other k values
- Eliminating outliers
- Eliminating missing values

Week 7 Topic:
Tax Strategy: Library set up

Create a setup.sas file to reference your data folder as a library:

In **setup.sas:**

```
%let path=/folders/myfolders/census;
libname census "&path";
```



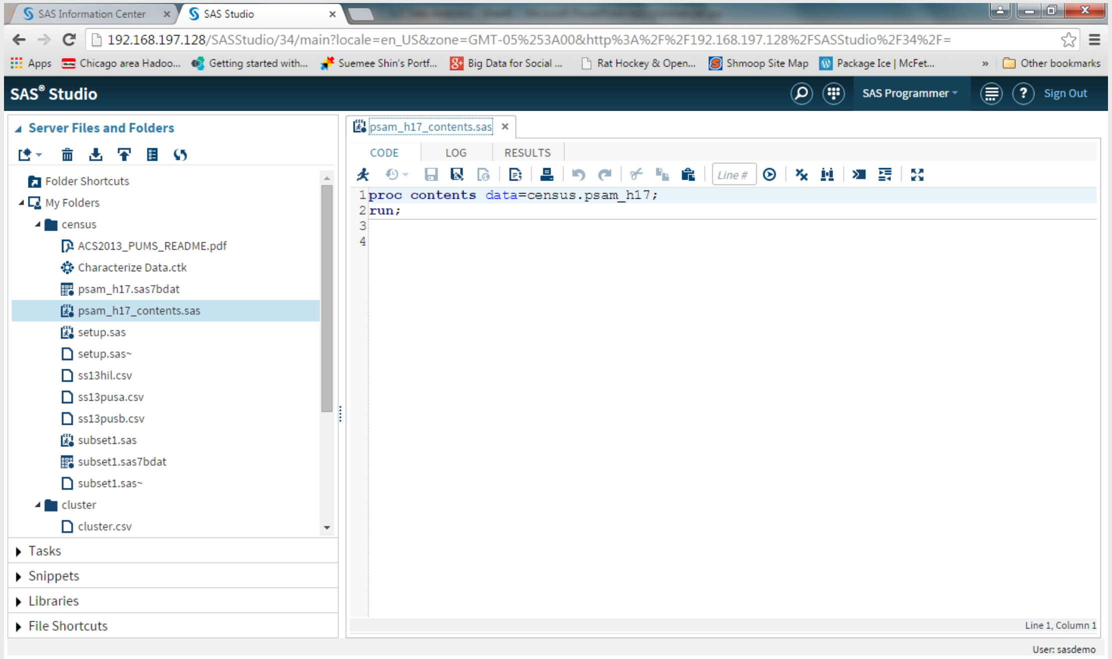
```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
55
56         %let path=/folders/myfolders/census;
57         libname census "&path";
NOTE: Libref CENSUS was successfully assigned as follows:
      Engine:        V9
      Physical Name: /folders/myfolders/census
58
59         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
71
```

Week 7 Topic:
Tax Strategy: PROC CONTENTS

Run PROC CONTENTS to explore dataset :

Create program **psam_h17_contents.sas:**

```
proc contents data=census.psam_h17;
run;
```



```
proc contents data=census.psam_h17;
run;
```

Week 7 Topic:
Tax Strategy: Subset of Variables

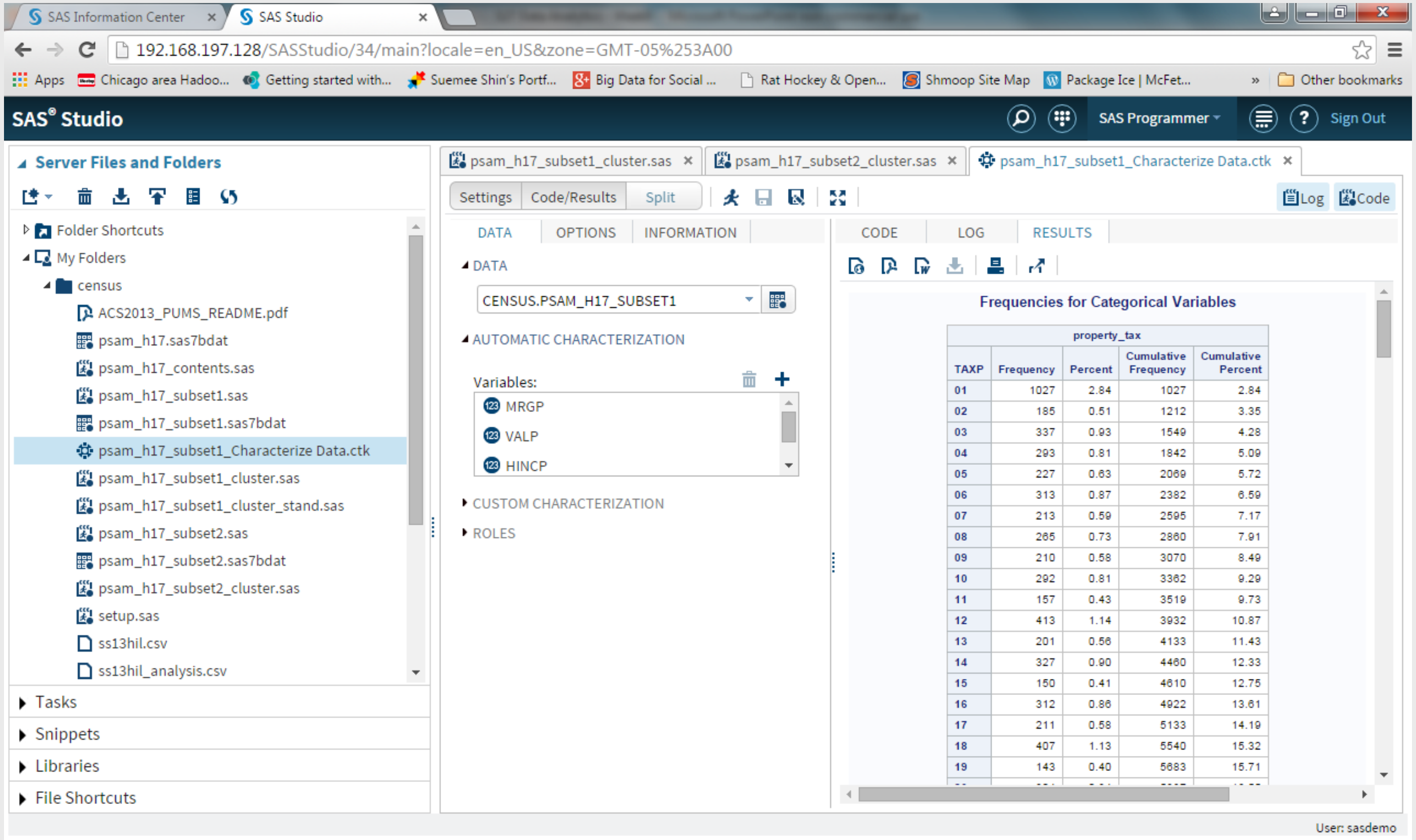
Run DATA step to select variables for analysis:

In **psam_h17_subset1.sas:**

```
data census.psam_h17_subset1;
                    set census.psam_h17;
                    keep SERIALNO MRGP HINCP TAXP VALP;
                    label SERIALNO='serial_id'
         MRGP='mortgage_payment'
         HINCP='household_income'
         TAXP='property_tax'
         VALP='property_value';
                    format SERIALNO $9.
                                          MRGP Z5.
                                          HINCP Z9.
          TAXP $2.
          VALP Z9.;
run;

proc contents data=census.psam_h17_subset1;
run;

proc print data=census.psam_h17_subset1(obs=20) label;
run;

proc sgplot;
scatter y=HINCP x=VALP;
run;
```

Week 7 Topic:
Tax Strategy: Characterize Data

Explore data using Characterize Data:



192.168.197.128/SASStudio/34/main?locale=en_US&zone=GMT-05%253A00

Server Files and Folders

- Folder Shortcuts
- My Folders
  - census
    - ACS2013_PUMS_README.pdf
    - psam_h17.sas7bdat
    - psam_h17_contents.sas
    - psam_h17_subset1.sas
    - psam_h17_subset1.sas7bdat
    - psam_h17_subset1_Characterize Data.ctk
    - psam_h17_subset1_cluster.sas
    - psam_h17_subset1_cluster_stand.sas
    - psam_h17_subset2.sas
    - psam_h17_subset2.sas7bdat
    - psam_h17_subset2_cluster.sas
    - setup.sas
    - ss13hil.csv
    - ss13hil_analysis.csv

psam_h17_subset1_cluster.sas | psam_h17_subset2_cluster.sas | psam_h17_subset1_Characterize Data.ctk

DATA

CENSUS.PSAM_H17_SUBSET1

AUTOMATIC CHARACTERIZATION

Variables:

- MRGP
- VALP
- HINCP

CUSTOM CHARACTERIZATION

ROLES

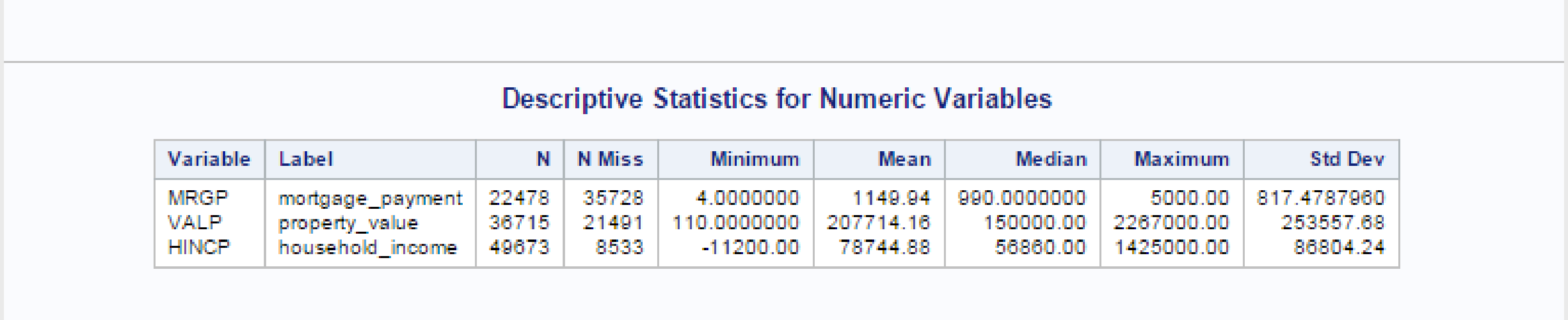
Frequencies for Categorical Variables

property_tax				
TAXP	Frequency	Percent	Cumulative Frequency	Cumulative Percent
01	1027	2.84	1027	2.84
02	185	0.51	1212	3.35
03	337	0.93	1549	4.28
04	293	0.81	1842	5.09
05	227	0.63	2069	5.72
06	313	0.87	2382	6.59
07	213	0.59	2595	7.17
08	265	0.73	2860	7.91
09	210	0.58	3070	8.49
10	292	0.81	3362	9.29
11	157	0.43	3519	9.73
12	413	1.14	3932	10.87
13	201	0.56	4133	11.43
14	327	0.90	4460	12.33
15	150	0.41	4610	12.75
16	312	0.86	4922	13.61
17	211	0.58	5133	14.19
18	407	1.13	5540	15.32
19	143	0.40	5683	15.71

Week 7 Topic:
Tax Strategy: Variables Analysis

For the numerical variables, note:

- **Negative values exist for HINCP. For our analysis, we may exclude negative values using the WHERE statement.**
- **Reviewing MRGP values versus using VALP values, we determine that VALP scales better with HINCP and more complete. Hence, we choose to use VALP instead of MRGP.**
- **Looking at N MISS, we see that many vales are missing in both VALP and HINCP. For our analysis, we will exclude missing data values using the WHERE statement.**



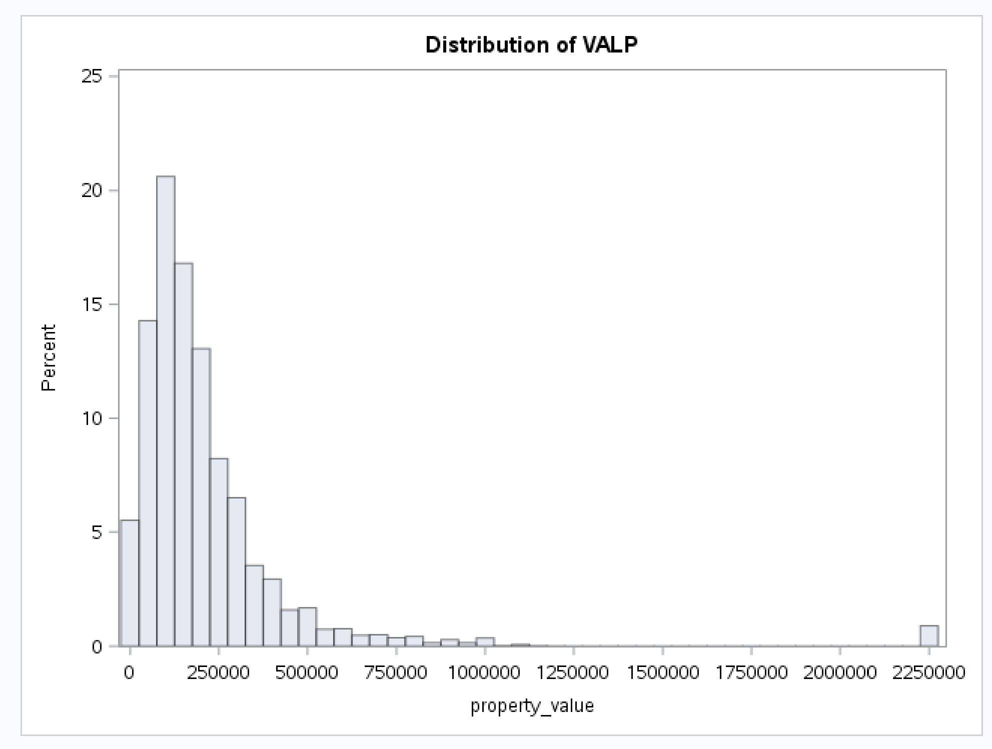
Descriptive Statistics for Numeric Variables

Variable	Label	N	N Miss	Minimum	Mean	Median	Maximum	Std Dev
MRGP	mortgage_payment	22478	35728	4.0000000	1149.94	990.0000000	5000.00	817.4787960
VALP	property_value	36715	21491	110.0000000	207714.16	150000.00	2267000.00	253557.68
HINCP	household_income	49673	8533	-11200.00	78744.88	56860.00	1425000.00	86804.24

Week 7 Topic:
Tax Strategy: VALP

For VALP:

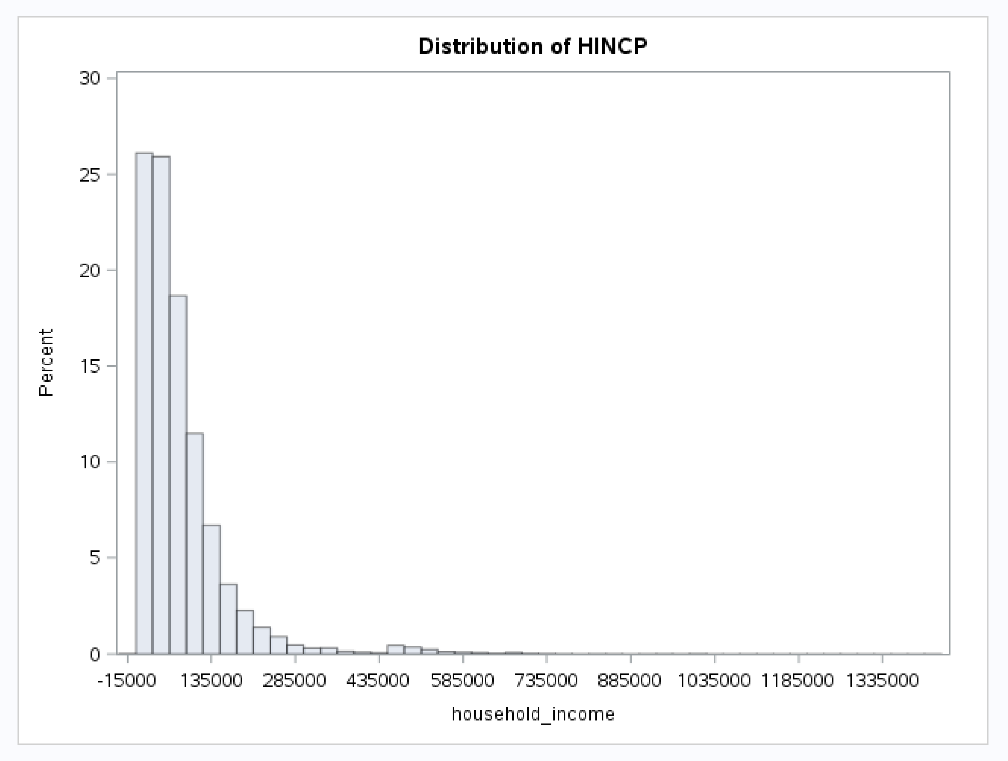
- **We see that the observations are highly skewed.**
- **We may subset to exclude high value properties and/or standardize dataset.**



Week 7 Topic: Tax Strategy: HINCP

For HINCP:

- ◆ **We see that the observations are highly skewed.**
- ◆ **We may subset to exclude high incomes and/or standardize dataset.**



Week 7 Topic:
Tax Strategy: SURVEYSELECT

Sometimes, it's easier to work with a smaller dataset to understand clustering options. To randomly choose a subset of data, you can use PROC SURVEYSELECT.

Subset of rows reference:

https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_surveyselect_sect003.htm

```
proc surveyselect data=census.subset1
          method=srs n=10000 out=census.subset2;
run;

proc print data=census.subset2(obs=20);
run;

proc contents data=census.subset2;
run;

proc sgplot;
scatter y=HINCP x=VALP;
run;
```

Week 7 Topic:
Tax Strategy: PROC FASTCLUS

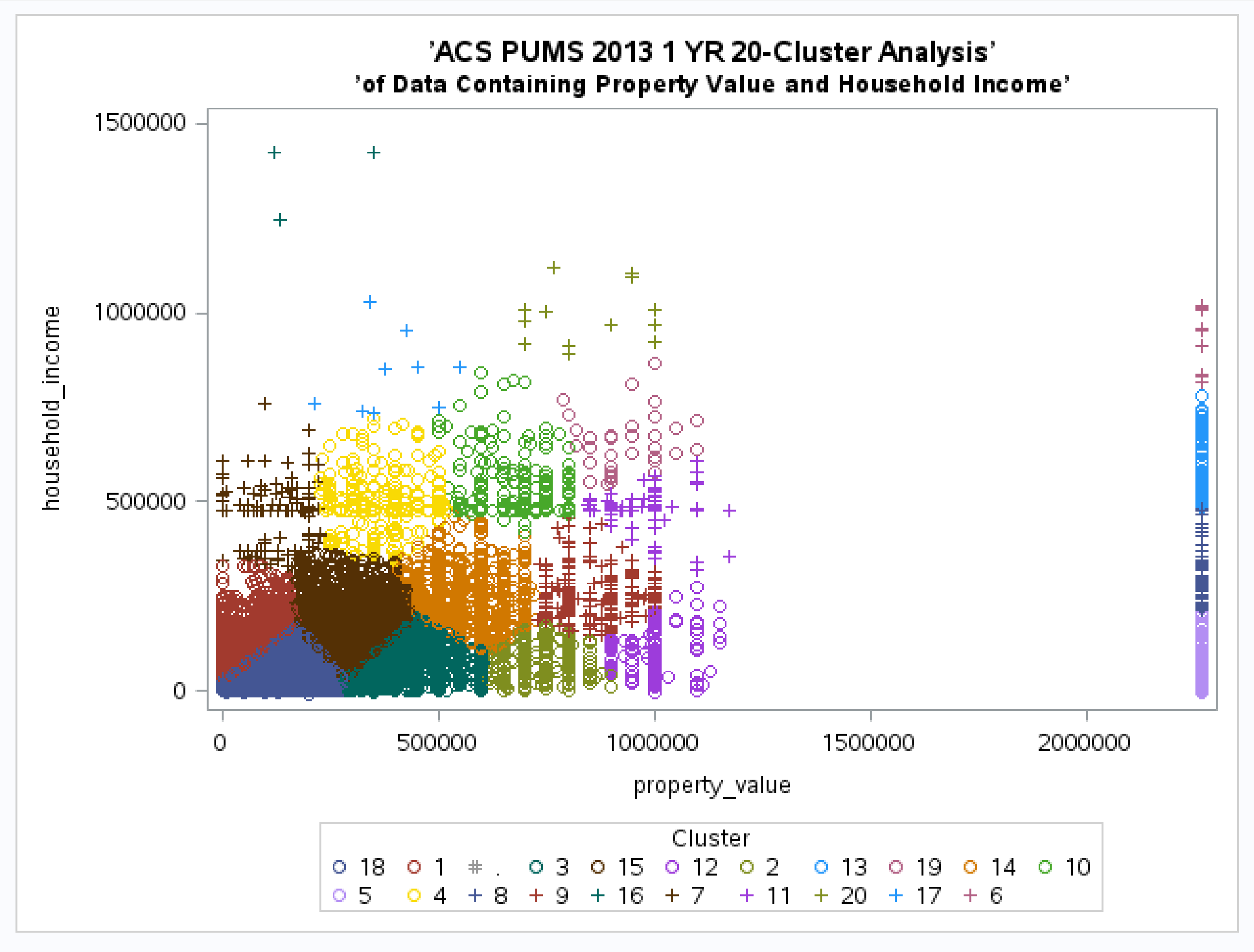
We choose to run with the following options. Note that we take out the LIST option as it will produce a larger than 4GB output of observations:

In **psam_h17_subset1_cluster.sas:**

```
proc fastclus data=census.psam_h17_subset1 radius=0 replace=full maxclusters=20
converge=0 OUTSTAT=subset1_20clusters_stat OUT=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 7 Topic:
Tax Strategy: Analysis of Clusters

We review the clusters using the scatterplot:



Week 7 Topic:
Tax Strategy: PROC FASTCLUS - stand

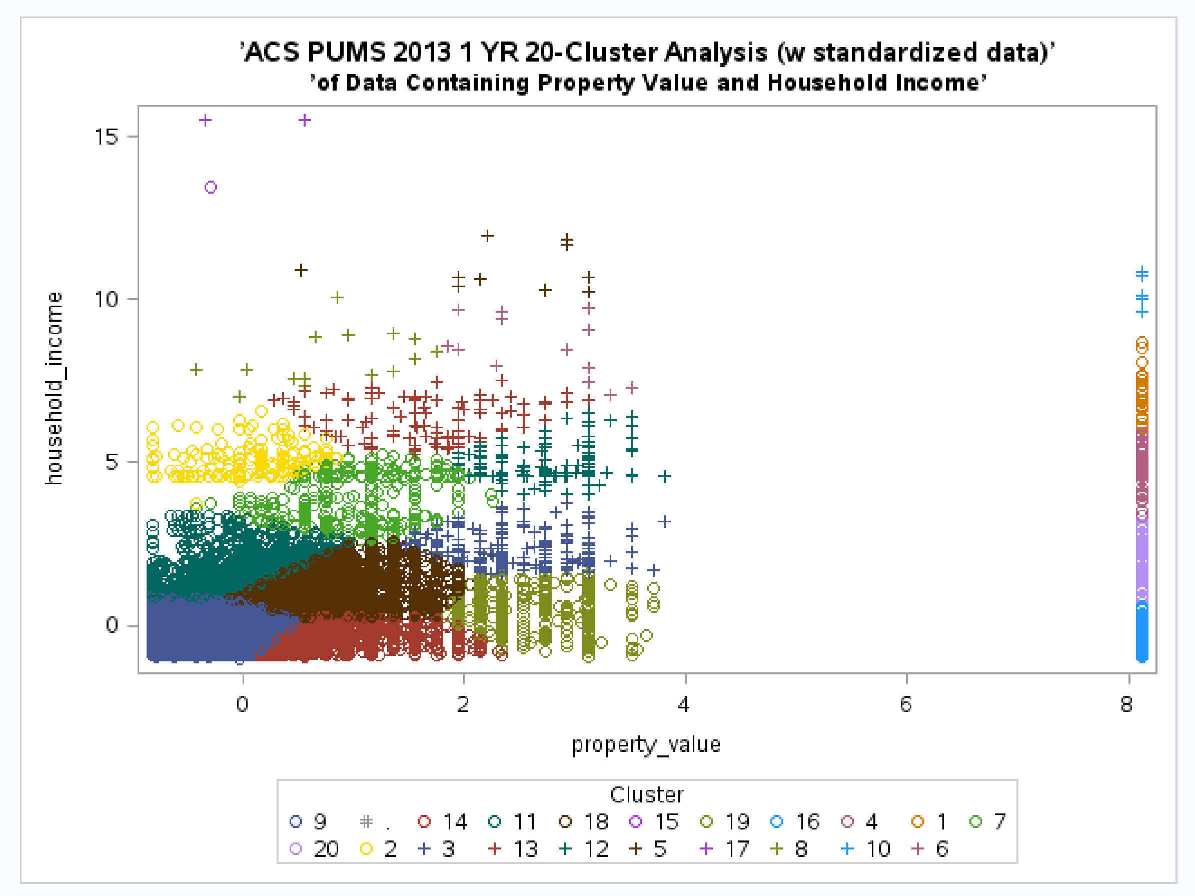
We run the previous example of the FASTCLUS run with 20 clusters with standardized variables:

In **psam_h17_subset1_cluster_stand.sas:**

```
proc standard data=census.psam_h17_subset1 out=Stand mean=0 std=1;
                 var VALP HINCP;
run;

proc fastclus data=Stand radius=0 replace=full maxclusters=20
converge=0 OUTSTAT=subset1_20clus_stat_stan OUT=subset1_20clus_stan
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 (w standardized data)’;
title2 ’of Data Containing Property Value and Household Income’;
run;
```

Week 7 Topic:
Tax Strategy: Analysis of Clusters - stand

We review the clusters using the scatterplot:



Week 7 Topic:
Tax Strategy: Analysis Considerations I

- 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.
- 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.
- 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 7 Topic:
Tax Strategy: Analysis Considerations II

- The pseudo ***F*** statistic, approximate expected overall R square, and cubic clustering criterion (CCC) are listed at the bottom of the figure. You can compare values of these statistics by running PROC FASTCLUS with different values for the MAXCLUSTERS= option. The R square and CCC values are not valid for correlated variables.
- Values of the cubic clustering criterion greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate potential clusters, but they should be taken with caution; large negative values can indicate outliers.
- 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 7 Topic: Tax Strategy: 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 7 Topic:
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.