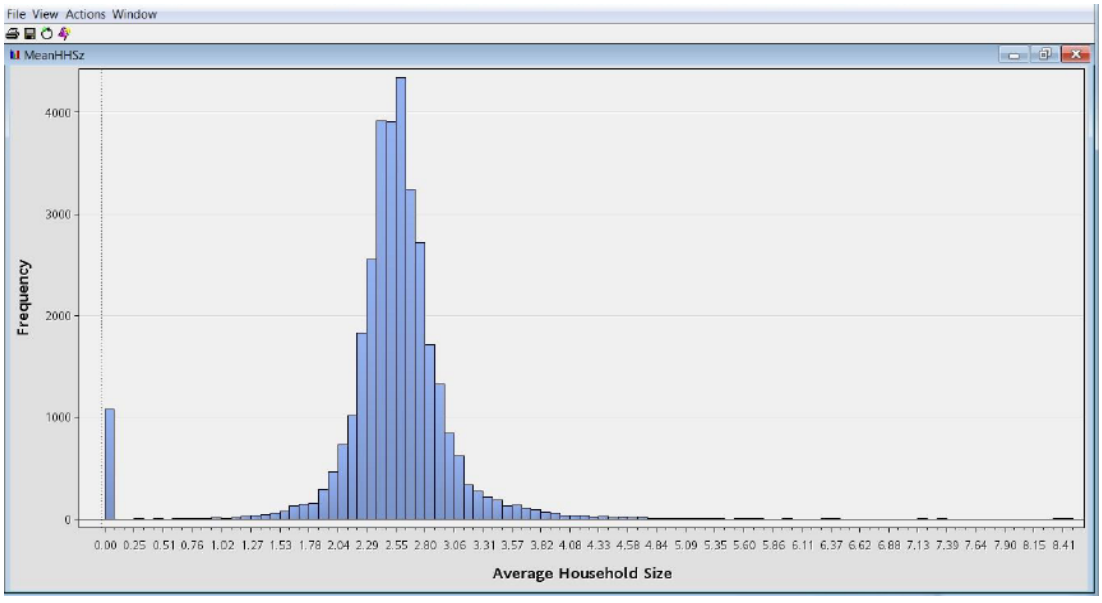
Clustering Zip Codes

What is the business purpose of this exercise?
To group geographic areas in the US into distinct subsets based on urbanization, household size, and income factors. The cluster analysis will then be used to identify candidate areas for each of the company's three grocery store types: low-end budget groceries, small urban boutique groceries, and large full-service supermarkets.

Exploring the Data Source

1. Maximize the SASEM.CENSUS2000 window.
    a. How many variables are in the data set? 7
    b. How many records (rows/observations) are in the data set? 33,178
    c. What does each record/row represent? A zip code

2. Maximize the SAMPLE STATISTICS window.
    a. What does **MeanHHSz** stand for? The average household size in the region
        i. What is the maximum (mean) number of individuals in a household? 8.49
        ii. What is the mean (of the mean) HHsz for all records? 2.50071
    b. What does a RegDens of 100 mean? The region population density percentile is at its highest density
    c. Why is Median Income a more appropriate measure than Mean Income for each region? The distribution will be highly skewed since income levels vary by zip code. Each zip will consist of people with low income and people with high income. The median is a better representation of the center than the mean when skew is high.
3. Why does each of the RegionID numbers have a frequency of 1 in the bar chart of ID?
   ID is the primary key and must be unique.
4. What is the purpose of a histogram of a variable? It shows the distribution of a given variable for all records.
    a. In the histogram of HHSize, describe the tallest bar – what does it indicate?
       15,647 records have a mean household size between 2.55 and 3.40.
    b. **Copy/Paste** a screen shot of the Histogram of Avg HouseHold Size **after** re-binning to 100.



i. State what the tallest bar indicates. 4,348 records have mean household sizes between 2.55 and 2.63
ii. State what the bar near 0 indicates. The spike in the histogram near 0 indicates a data quality problem within the records – there are other missing attributes. It doesn't make sense for a household to have a 0 average household size.
iii. What is causing the bar near zero in the histogram of Avg HH size? Most of the cases have with a zero average household size have 0 or missing on the remaining nongeographic attributes. Don't want to form clusters on bad pieces of data.
iv. 'Restore' the histogram of HHsz.
    1. Select the bar near zero in the HHsz histogram.
    2. Where do these values appear in the Regional Density Percentile histogram? Between 0-1 on the Regional Density percentile – which indicates missing records.

B. Case Filtering

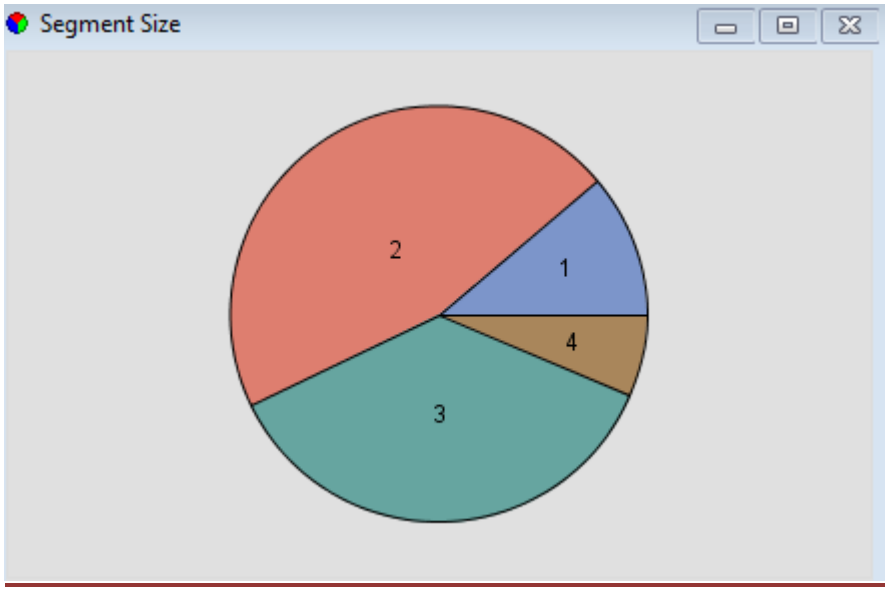
1. What is the purpose of the filtering node? It enables you to remove unwanted records from an analysis.
    a. What is the default filtering method for "class" (qualitative) variables? Rare values (percentage)
    b. What is the default filtering method for "interval" (quantitative) variables? +/- 3 standard deviations from the mean
    c. What are you filtering out? Records with mean household sizes of .1 or less
    d. Why is this data being filtered out? We want to improve the integrity of our model. The poor data quality caused by the missing records skew the cluster.
2. Right click on the filter node and select RUN.
    a. How many records were removed? 1,081

C. Clustering

1. What variables are used to cluster the records? MedHHInc (median house hold income), MeanHHSz (average household size), and RegDens (regional density)
2. What are the units of measure of the 3 input variables used to cluster the records?
    1. Average Household Size uses percentages
    2. Median Household Income uses dollars
    3. Region Population Density Percentile uses percentages
3. Why should the input variables be 'standardized' for this cluster method? We standardize units of measure to get meaningful clustering.
    a. What is the process of standardization in the Cluster Node? Subtract the mean and divide by the standard deviation of the input values.

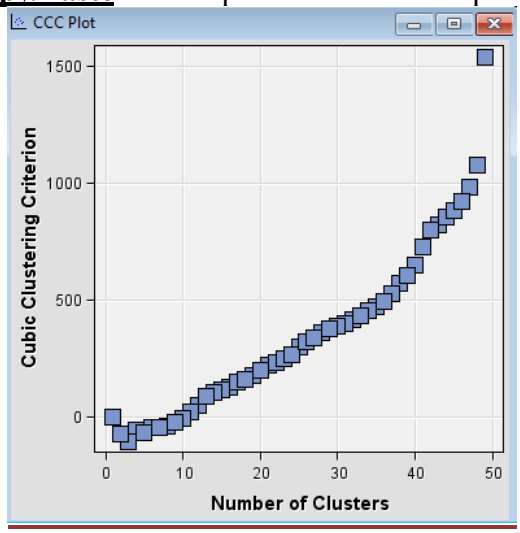
4. Steps 1-3 in the Default Cluster Method
    a. How many clusters are initially created in **Step 1**? 50
    b. What statistic is used to reduce the # of clusters in **Step 2**? The cubic clustering criterion (CCC)

5. Run the cluster node - Results
    a. **Copy/Paste** the Segment Size Pie Chart



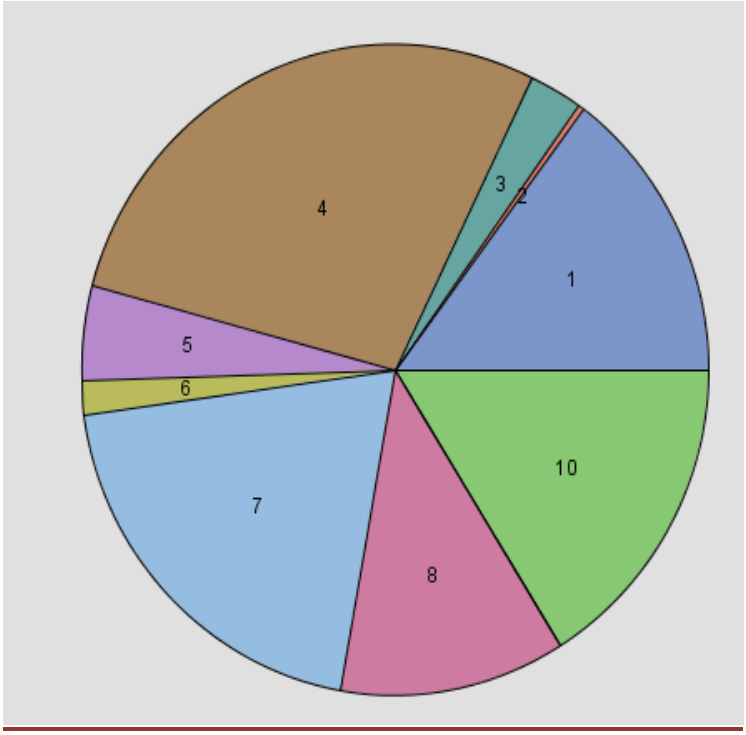
b. What value was used for *k* in this *k-means* clustering analysis? 4
c. How many records (rows) are in the largest cluster? 14729
    i. What is the median HH income of this cluster? (Mean Statistics window) $32,920.12
d. The largest Avg HHSize is in what cluster (segment id)? 1
e. Which cluster (segment id) has the lowest Regional Density? 2
    i. Does this suggest a rural or urban region? a rural region
f. Why are four clusters convenient for this case related to grocery store types? Having four clusters increases the likelihood of mapping onto the demographic that is likely to shop at one of three grocery stores.

6. **Copy/Paste** a small picture of the CCC plot.



a. What is the CCC value for 4 clusters? -60.2249
b. What is the problem with this value? The CCC plot showed no clear peak, and the four-cluster solution corresponded to a negative CCC. A negative CCC is not meaningful.

7. Change the Number of Clusters
    a. **Copy/Paste** the Segment Size Pie Chart

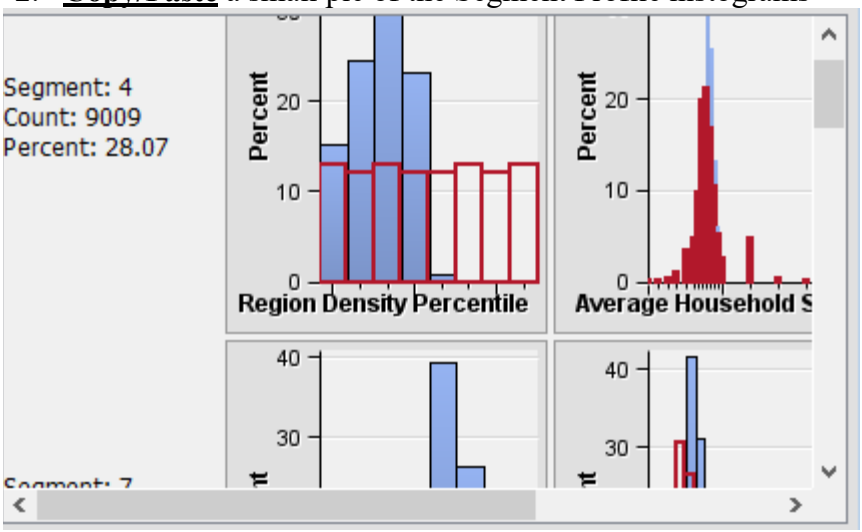


b. How many data records are in cluster 9 (Segment id 9)? 10
   i. Why is this problematic? There are too few records for effective grouping.

c. Why are 10 clusters probably too many for our Business Case? The objective of the cluster analysis is to find the best candidate zip codes for only 3 grocery stores. An effective analysis should include about 5 clusters with the most records.

D. Profiling Clusters

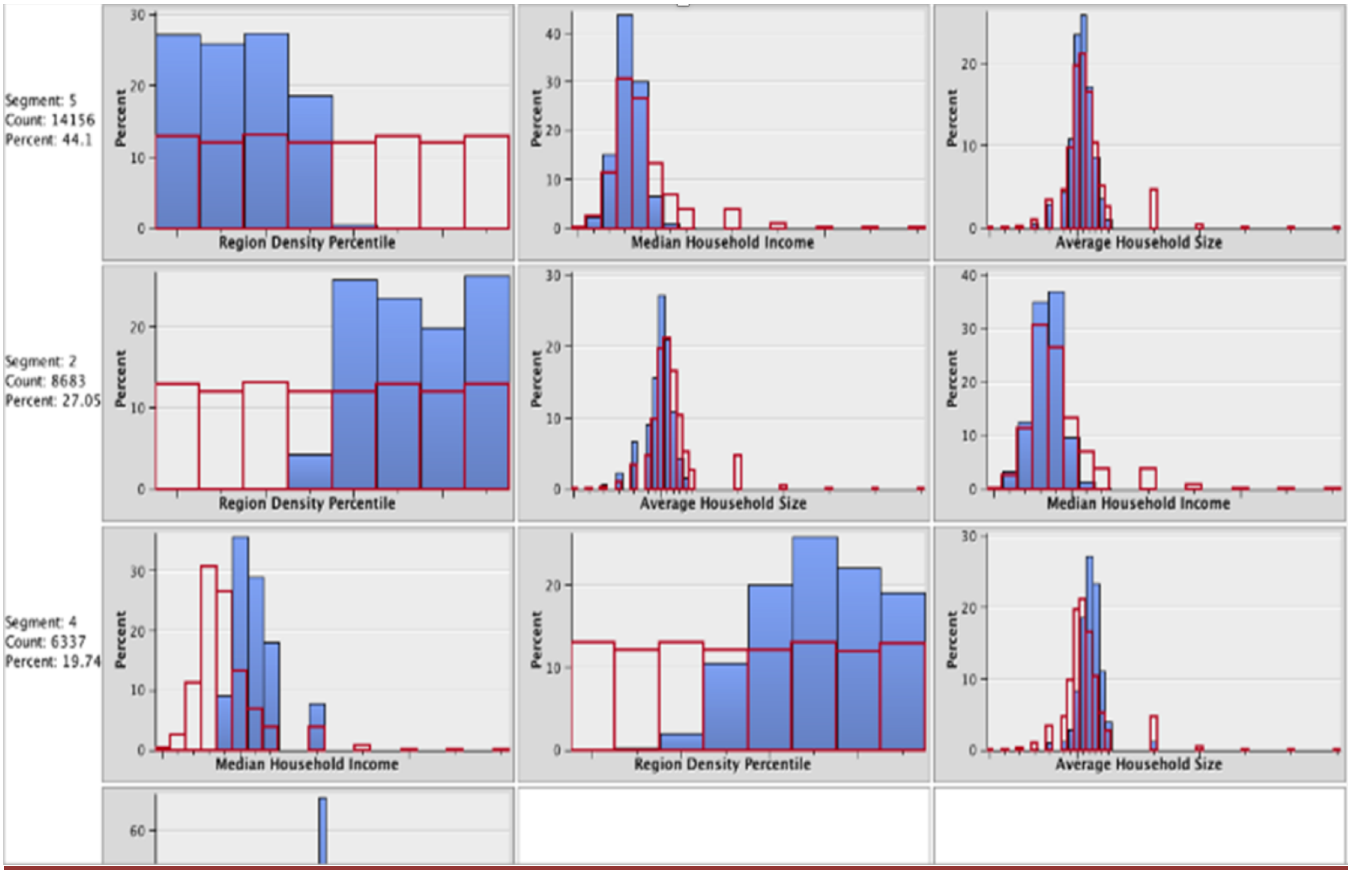
1. What is the purpose of the Segment Profile Node? It enables you to compare the distribution of a variable in an individual segment to the distribution of the variable overall.
2. **Copy/Paste** a small pic of the Segment Profile histograms



a. How many clusters were profiled? 7
b. Why is Cluster #4 the first profiled? The output is arranged by cluster size and by variable importance. It has the highest frequency count and is the largest cluster.
c. How many data records are in the largest cluster? 9009

d. For Segment 10 (or cluster 10), why is Regional Density Profile the first variable profiled? Since Regional Density Profile is the most important variable for segment 10.
e. What do the red outlines in the histograms represent? The red outlines represent the population distribution (overall distributions).
f. Describe the attributes of the households clustered in Segment 7. How are the records in this cluster of zip codes similar? In comparison to the overall distributions, Segment 7 has a higher Region Density Percentile, a higher average household size, and a more central median household income.

3. Select the Cluster Node and change the "Maximum number of clusters" in the properties panel to 5.
   a. Run the Profile Node
   b. **Copy/Paste** the Profile: Histograms of the clusters.



4. Using the Profiles for the 3 largest Clusters, match each cluster of zip codes to **one** of the three types of grocery stores that might be best for those locations. **Explain why** you think that cluster is best for the type of store you selected. (Use the store types and clusters one time only).

   a. Budget grocery: segment 5
   The region density percentile is less dense than the population's. Opening a budget grocery store in a less densely populated area may be cheaper in terms of real estate and operational costs to run the grocery store. Several budget grocery stores can therefore be built in less densely populated regions. The less dense a region is, the lower its wages and income levels are going to be. Families in this cluster have a below average household size, which may be due to their lower income levels. A budget grocery store selling goods at affordable prices will maximize sales within this cluster.

b. Small boutique grocery: segment 4
The distribution of mean household income is greater than the population's, consumers in this area would be able to afford products sold at a boutique store that are more specialized and expensive. A region's household income is more important than its population density since for determining the location of a boutique store. This cluster has a higher region density and average household size than the population's.

c. Large full-service grocery: segment 2
This cluster has a higher region density percentile than the average. A dense area will increase consumer traffic into the grocery store. Supermarkets should be placed in a central area, within a dense population, to be easily accessible to consumers. The average income is less than most (not as low as more poor areas, but low enough to be your middle class average incomes)