**Project on Geo Consumption data by Zip Code**

**Source:**

Geo Consumption data by Zip Code (From Blackboard)

**Data description**

The Geo Consumption dataset contains Demographic, socio-economic, and environmental variable details at almost 29,791 Zip codes. It has covered the facts such as Households, Population Average Household Size, Average Household Income, Population Ages groups, Race, Owner-Occupied Dwellings, Median Cash Rent (2000), Education level, Housing Units Built details at Zip code level. It also covers Average Household spending on Apparel, Education, Entertainment, Food, Gifts, Health Care, Furnishing, Miscellaneous, Tobacco, Utilities, and Transportation. It also captures environmental details such as Snow Fall, Rain fall, Temperature levels at different months, Hurricane, Tornado, Wind index etc. Using this dataset exploratory research could be done to understand the spending behavior at Zip code level. Which helps us to realize the consumer spending behaving and to identify potential zip codes. which indeed helps us taking further strategic steps.

**Key areas of marketing:**

This research could help us to take further steps on,

* Who is our target: Where to invest (at Zip code level)?
* Identify Opportunities
* What kind of product needs more advertising?
* Realizing the market potential for particular category of products.
* Understanding the consumer spending patters
* Targeting the right product with right consumer segment
* Providing discounts on slow moving product, creating the need, which leads demand for the that product in the future
* Coming up with Pricing Strategy etc.

**Nature/Characteristics of variable:**

All the variable details are continuous in nature

**Key variables**

Average Household expenditure on Apparel, Education, Entertainment, Food, Gifts, Health Care, Furnishing, Miscellaneous, Tobacco, Utilities, Transportation at the Zip Code level are the mail variable helps us to understand the consumer behavior.

**Observation on these variables**

Looking at the expenditure variables in detail, it gives a rough idea about the nature of spends at zip code level, like are they more focused on food, entertainment, Fashion, education, health etc.

**Advantages of using this dataset:**

1. Gives the broader picture consumer behavior at the Zip Code level
2. Products and price offerings could be varied base on consumer needs, like by increasing the prices on fast moving products, by proving more discounts on least selling item
3. Could be used to increase the product portfolios by analyzing the consumer needs & buying tendency.

**Disadvantages of using this dataset**

1. Details are not covered at more granular level (at each customer level)
2. We cannot customize the offer for individuals
3. Hard to analyze the individual need and demands

**Problem formulation**

**Problems and/or opportunities addressed by the project**

1. It gives the opportunity to analyze the consumer behavior at the Zip Code level
2. Helps to customize the products and price offerings at identified potential zip codes. Identifying zip codes could also be considered as our main Research objectives for the project
3. Problem could be sometimes it is hard to pin point the customer requirements at zip code level

**Target audience**

The target audience could be at strategic level audience; like marketing directors, Marketing VP’s, CMO’s, Channel partners like Ad agencies, Media partners etc.

**Model building:**

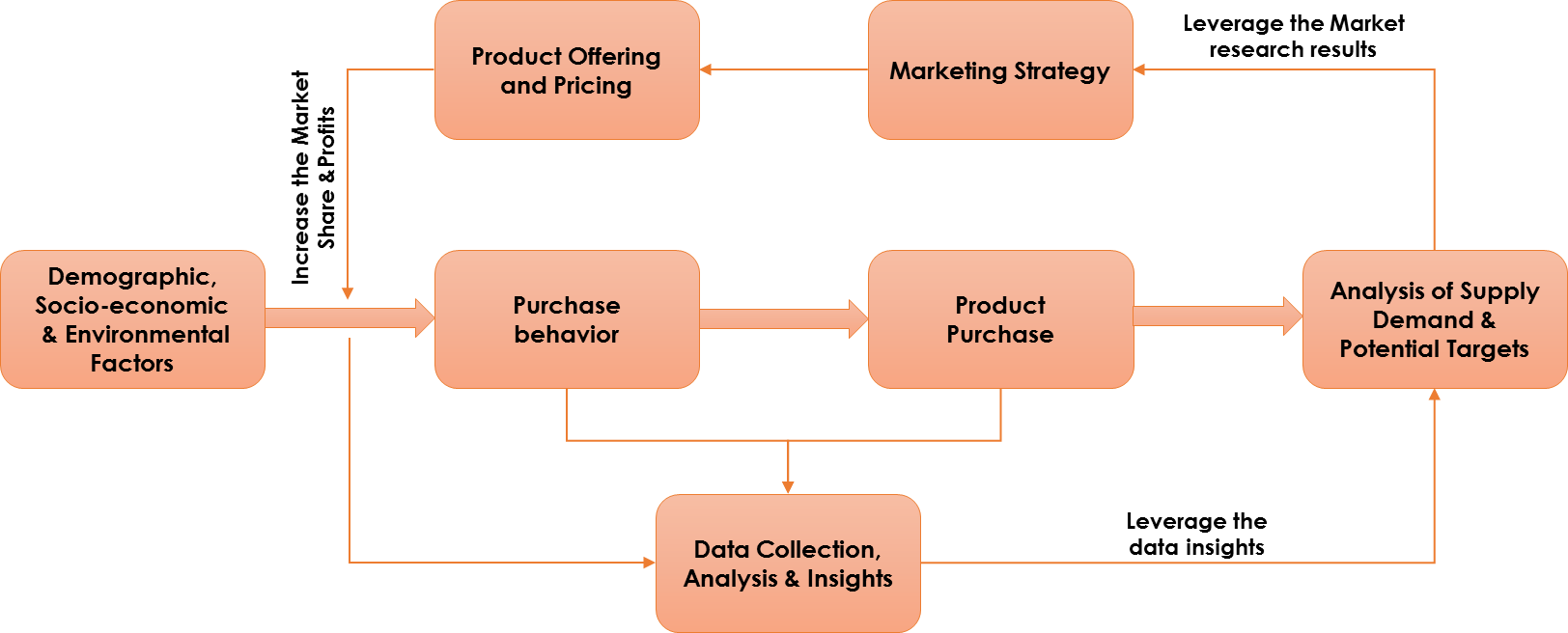
**Associations between variables:**

The dataset containing, Demographic, socio-economic, and environmental variables will have a direct impact on the purchase behavior, which indeed affect the type of expenditure made by the consumers. Based on these parameters the basic needs and demands will be prioritized in the consumer perception.

**Reasoning behind the models:**

Using these independent and dependent factors, we can analyze the business opportunities like potential areas or zip codes to invests, pushing new product etc. Which in turn helps in building marketing strategies and models to improve the business. Below model give the overall picture of how each parameter are interrelated, how these data could be leveraged in improving the business.

**Marketing model:**



**Outcomes of understanding the marketing model:**

1. Gives the broader picture consumer behavior at the Zip Code level
2. Products and price offerings could be varied base on consumer needs, like by increasing the prices on fast moving products, by proving more discounts on least selling item
3. Helps in building the Marketing Strategy & Product Pricing

**Benefits from the marketing models**

1. Could Increase the product portfolio
2. Creating needs by providing discounts, could turned into demands in the future
3. Helps in increasing the Revenue and Profits buckets

**Analysis methods**

**Available methods to analyze the models:**

1. Factor analysis
2. Regressing Analysis
3. Correlation
4. Clustering

**Strengths and weaknesses of each models:**

**Factor Analysis:**

**Advantages:** Reduction of number of variables, by combining two or more variables into a single factor. Factors are selected by grouping related items

**Disadvantages:** The factors that appear can only come from the answers to the questions you as. Selecting a good set of questions is complicated, and different researchers will choose different sets of questions.

**Regression Analysis:**

**Advantages:** The estimates of the unknown parameters obtained from linear least squares regression are the optimal. It uses data very eﬃciently. Good results can be obtained with relatively small data sets.

**Disadvantages:** It is very sensitive to outliers. It often gives optimal estimates of the unknown parameters. Assumptions needs to be checked (Assumptions: Errors from the model are normally distributed; that the errors have constant variance; that the mean of the errors is zero; and that the errors are independent)

**Cluster Analysis:**

**Advantages:** CA may reveal associations and structure in data which was not previously evident. The results of cluster analysis may contribute to the definition of a formal classification scheme or taxonomy. Cluster analysis is relatively simple, and can use a variety of input data

**Disadvantages:** Cluster analysis methods are not clearly established consequently most of the guidelines for using cluster analysis are rules of thumb. There are no completely satisfactory methods for determining the appropriate number of clusters

**Choices of the methods:**

1. **Factor analysis** could help us in reducing the no of independent variable and capturing the most import factors from the dataset, which helps in explaining the impact of group of independent identical variables on outcomes.
2. Applying the **regressing models** on expenditure variable using new independent variable (factors) to analyze the impact of Demographics, Social-economic and environmental variables
3. Identifying the **correlation** among the dependent variables
4. **Clustering** the top zip codes and identifying the characteristics of each cluster

**Results and discussion**

***Data considered:***

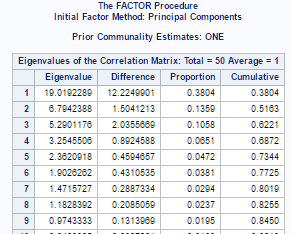
1. *Used all the Zips codes data for analysis, the predictions were not really significant, as the dataset is huge (Over 29,000 zip code data)*
2. *Tried using Top 200 zip codes by population, the results were not significant*
3. *Finally, Considered* ***Top 200 Zip codes by Average Household Size (AHS2004)*** *as a Subset for the Analysis*

**Methodology 1: Factor Analysis**

**Factor analysis** has been done on the Demographic, socio-economic, and environmental variables (SED) to generate an interpretable variable list (factors) as the new set of independent variables.

**Identification of NO of Factors**

* **Eigenvalues -** There are 8 Factors with Eigenvalues > 1, it implies factor with an eigenvalue of 1 accounts for as much variance as a single variable. 8 Factors could be considered for our analysis.



* **Scree Plot -** A scree plot shows the eigenvalues on the y-axis and the number of factors on the x-axis. It always displays a downward curve. The point where the slope of the curve is clearly leveling off **(the “elbow)** indicates the number of factors that should be generated by the analysis. According to the observation of “the elbow”, we can consider 8 Factors for our analysis, which almost explains 82.5% variation in the data.



(As the output contain large no. of variable, I couldn’t take the screen shot. Please refer the attached PDF: **“Factor\_Output1”** for detailed results).

**Factor Interpretation:**

**Top Loads by Factor 1:**  It represents demographics variable like households, Population, Age group 15 to 55 and above, Owner & Renter occupied dwellings, housing unites build, Education details from Schooling to professional degree.



**Top load by Factor2:** It represent household food spends and use of coupons



**Top load by Factor3:** It represents environmental variables like, Temperature range and Snowfall measures.



**Top load by Factor4:** It represents demographic variable like, Household income, Graduates or Professional degree, Bachelor’s degrees, with White & Asian/pacific Islander details



**Top load by Factor5:** It represents environmental variable like, Hail, Tornado Index and Non-Hispanic population details



**Top load by Factor6:** It represents environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature



**Top load by Factor7:** It represents environmental variable like July temperature & Annual Rainfall



**Top load by Factor8:** It represents household size & Housing Units build



**Summary of Variables:**

**Factor 1:**  It represents demographics variable like households, Population, Age group 15 to 55 and above, Owner & Renter occupied dwellings, housing unites build, Education details from Schooling to professional degree.

**Factor2:** It represent household food spends and use of coupons

**Factor3:** It represents environmental variables like, Temperature range and Snowfall measures.

**Factor4:** It represents demographic variable like, Household income, Graduates or Professional degree, Bachelor’s degrees, with White & Asian/pacific Islander details.

**Factor5:** It represents environmental variable like, Hail, Tornado Index and Non-Hispanic population details

**Factor6:** It represents environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

**Factor7:** It represents environmental variable like July temperature & Annual Rainfall

**Factor8:** It represents household size & Housing Units build



**Methodology 2: Regression Analysis**

**Regression Analysis** has been done using new independent variable (Factors) as Input and Expenditure metrics as output variables.

## Regression analysis for **Ave Household Expenditure on Apparel** using all the Independent variable (Factor1 to Factor 8)

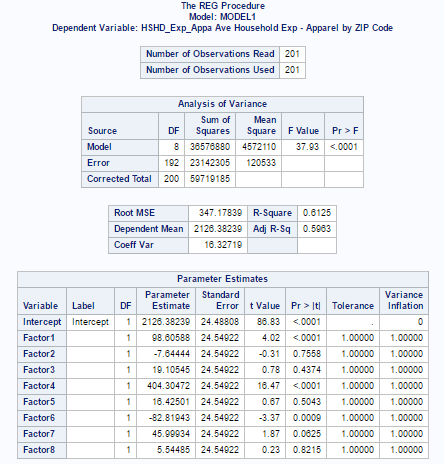
**Proc Reg data = Independent;**

**model HSHD\_Exp\_Appa = Factor1- Factor8**

**/ tol vif collin;**

**run;**

From the below parameter estimates, it shows, Factors 2 and Factor6 are negatively correlated with **Ave Household Expenditure on Apparel.** The parameter estimates measure the rate of change corresponding to a one‑unit change in the predictor variable. P Value shows Factor2, 3, 5,7 and 8 are Not significant in explaining the LHS (**Ave Household Expenditure on Apparel**) at the 95% significance level.



We can also observe that, (Adjusted R Square) 59.63% variation in the output (**Ave Household Expenditure on Apparel)** can be explained by the new independent variables (except Factor 2, 3, 5, 7 & 8)

From + Ve correlation and significance, it clearly shows, Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Household expenditure on Apparel** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Similarly, Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Household expenditure on Apparel** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.

Where as in –Ve Correlation with Factor 6 shows. Top 5 Zip codes of **Factor 6** 23521 32215 33039 78574 78589, in these zip codes the **Household expenditure on Apparel** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

## Regression analysis for **Ave Household Expenditure on Contributions** using all the Independent variable (Factor1 to Factor 8)

**Proc Reg data = Independent;**

**model HSHD\_Exp\_Contr = Factor1- Factor8**

**/ tol vif collin;**

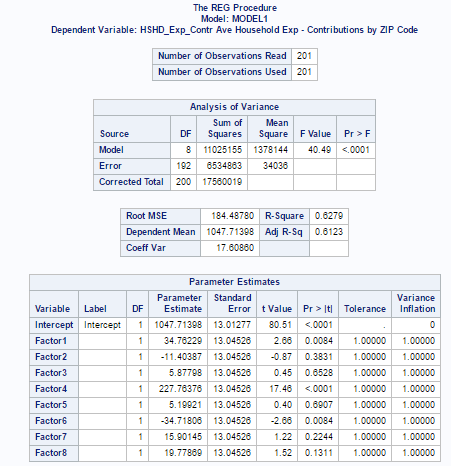
**run;**

From the below parameter estimates, it shows, Factors 2 and Factor6 are negatively correlated with **Ave Household Expenditure on Contributions.** The parameter estimates measure the rate of change corresponding to a one‑unit change in the predictor variable. P Value shows Factor2, 3, 5,7 and 8 are Not significant in explaining the LHS (**Ave Household Expenditure on Contributions**) at the 95% significance level.

We can also observe that, (Adjusted R Square) 61.23% variation in the output (**Ave Household Expenditure on Contributions)** can be explained by the new independent variables (except Factor 2, 3, 5, 7 & 8)

From + Ve correlation and significance, it clearly shows, Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Ave Household Expenditure on Contributions** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Similarly, Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Ave Household Expenditure on Contributions** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.



Where as in –Ve Correlation with Factor 6 shows. Top 5 Zip codes of Factor 6 23521 32215 33039 78574 78589, in these zip codes the **Ave Household Expenditure on Contributions** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

## Regression analysis for **Ave Household Expenditure on Education** using all the Independent variable (Factor1 to Factor 8)

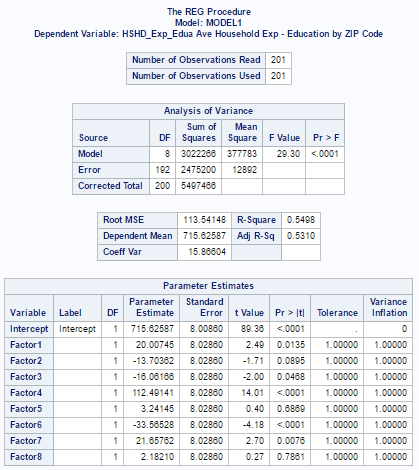
**Proc Reg data = Independent;**

**model HSHD\_Exp\_Edua = Factor1- Factor8**

**/ tol vif collin;**

**run;**

From the below parameter estimates, it shows, Factors 2 ,3 and Factor6 are negatively correlated with **Ave Household Expenditure on Education.** The parameter estimates measure the rate of change corresponding to a one‑unit change in the predictor variable. P Value shows Factor2, 5 and 8 are Not significant in explaining the LHS (**Ave Household Expenditure on Education**) at the 95% significance level.



We can also observe that, (Adjusted R Square) 53.1% variation in the output (**Ave Household Expenditure on Education)** can be explained by the new independent variables (except Factor 2, 5 & 8)

From + Ve correlation and significance, it clearly shows, Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Ave Household Expenditure on Education** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Similarly, Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Ave Household Expenditure on Education** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.

Where as in –Ve Correlation with Factor3 shows, Top 5 Zip Codes of **Factor 3** 78548, 79849, 96135, 96762, 96819 in these zip codes the **Ave Household Expenditure on Education** significantly decreases as one unite increase in Temperature range and Snowfall measures.

Similarly, –Ve Correlation with Factor 6 shows. Top 5 Zip codes of **Factor 6** 23521 32215 33039 78574 78589, in these zip codes the **Ave Household Expenditure on Education** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

We can summarize saying that, better the demographics & Socio-economic conditions higher the education spends, as the climatic conditions are unfavorable the spending on education decreases. We can conclude that **during summer the people likes to spend more on education**.

## Regression analysis for **Ave Household Expenditure on Entertainment** using all the Independent variable (Factor1 to Factor 8)

**Proc Reg data = Independent;**

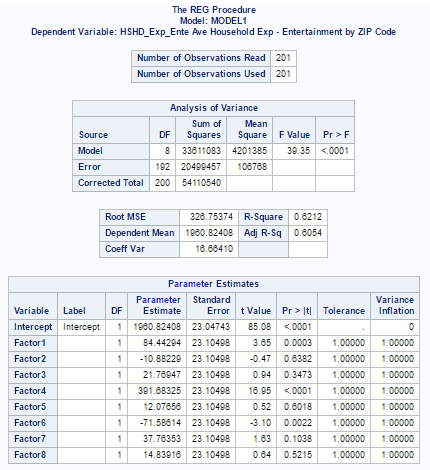
**model HSHD\_Exp\_Ente = Factor1- Factor8**

**/ tol vif collin;**

**run;**

From the below parameter estimates, it shows, Factors 2 and Factor6 are negatively correlated with **Ave Household Expenditure on Entertainment.** The parameter estimates measure the rate of change corresponding to a one‑unit change in the predictor variable. P Value shows Factor2, 3, 5, 7 and 8 are Not significant in explaining the LHS (**Ave Household Expenditure on Entertainment**) at the 95% significance level.

We can also observe that, (Adjusted R Square) 60.54% variation in the output (**Ave Household Expenditure on Entertainment)** can be explained by the new independent variables (except Factor 2, 3, 5, 7 & 8)



From + Ve correlation and significance, it clearly shows, Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Ave Household Expenditure on Entertainment** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Similarly, Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Ave Household Expenditure on Entertainment** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.

Whereas –Ve Correlation with Factor 6 shows. Top 5 Zip codes of **Factor 6** 23521 32215 33039 78574 78589, in these zip codes the **Ave Household Expenditure on Entertainment** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

We can summarize saying that, better the demographics & Socio-economic conditions higher the **Entertainment** spends, as the climatic conditions are unfavorable the spending on **Entertainment** decreases.

## Regression analysis for **Ave Household Expenditure on Food/Beverages** using all the Independent variable (Factor1 to Factor 8)

**Proc Reg data = Independent;**

**model HSHD\_Exp\_Food = Factor1- Factor8**

**/ tol vif collin;**

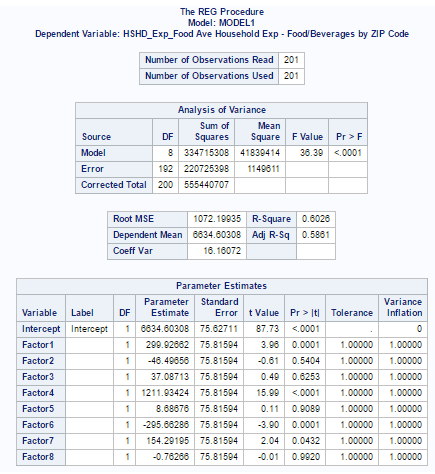
**run;**

From the below parameter estimates, it shows, Factors 2, 6 and Factor8 are negatively correlated with **Ave Household Expenditure on Food/Beverages.** The parameter estimates measure the rate of change corresponding to a one‑unit change in the predictor variable. P Value shows Factor2, 3, 5 and 8 are Not significant in explaining the LHS (**Ave Household Expenditure on Food/Beverages)** at the 95% significance level.

From + Ve correlation and significance, it clearly shows, Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Ave Household Expenditure on Food/Beverages** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Similarly, Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Ave Household Expenditure on Food/Beverages** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.

Similarly, Top 5 Zip Codes of **Factor 7** 99727, 99736, 99749, 99751, 99763 in these zip codes the **Ave Household Expenditure on Food/Beverages** significantly increases as one unite increase in July high and Average temperature & Annual Rainfall.



Whereas –Ve Correlation with Factor 6 shows. Top 5 Zip codes of **Factor 6** 23521 32215 33039 78574 78589, in these zip codes the **Ave Household Expenditure on Food/Beverages** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

We can summarize saying that, better the demographics & Socio-economic conditions higher the **Food/Beverages** spends, as the climatic conditions are unfavorable the spending on **Food/Beverages** decreases.

## Regression analysis for **Ave Household Expenditure on Gifts** using all the Independent variable (Factor1 to Factor 8)

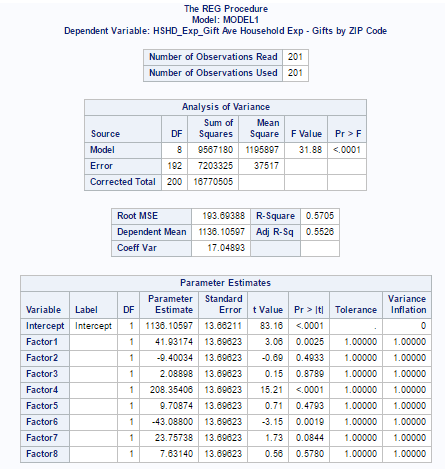
**Proc Reg data = Independent;**

**model HSHD\_Exp\_Gift = Factor1- Factor8**

**/ tol vif collin;**

**run;**

From the below parameter estimates, it shows, Factors 2 and Factor 6 are negatively correlated with **Ave Household Expenditure on Gifts.** The parameter estimates measure the rate of change corresponding to a one‑unit change in the predictor variable. P Value shows Factor2, 3, 5, 7 and 8 are Not significant in explaining the LHS (**Ave Household Expenditure on Gifts)** at the 95% significance level.



From + Ve correlation and significance, it clearly shows, Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Ave Household Expenditure on Gifts** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Similarly, Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Ave Household Expenditure on Gifts** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.

Whereas –Ve Correlation with Factor 6 shows. Top 5 Zip codes of **Factor 6** 23521 32215 33039 78574 78589, in these zip codes the **Ave Household Expenditure on Gifts** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

We can summarize saying that, better the demographics & Socio-economic conditions higher the **Gifts** spends, as the climatic conditions are unfavorable the spending on **Gifts** decreases.

Similarly, Expenditure on **Health Care**, **Furnishing, Household Operations, Personal Care, Spending on Reading, on Tobacco, on Transportation and Utilities**; shows better the demographics & Socio-economic conditions higher the spends, as the climatic conditions are unfavorable the spending on **most of these categories** decreases.

(Attached PDF: **“2. Regression\_Output2”** with detailed regression results for all the expenditure variables).

**Methodology 3: Correlation of All expenditure variable**

(As the output contain large no. of variable, I couldn’t take the screen shot. Please refer the attached PDF: **“3. Correlation\_Output3”** for detailed results)

The correlation results clearly show that all the expenditure variables are related to each other, hence in most of the regression outputs, the same the same socio-economic factors and environmental factor are becoming more significant in determining the spending nature.

**Methodology 4:**

**Cluster Analysis for Top 10 Zip Code by each Factor – “Ward’s Method”**

**proc aceclus data=Topzip out=Ace p=.03 noprint;**

**var HSHD\_Exp\_Appa HSHD\_Exp\_Contr HSHD\_Exp\_Edua**

**HSHD\_Exp\_Ente HSHD\_Exp\_Food HSHD\_Exp\_Gift HSHD\_Exp\_Heal**

**HSHD\_Exp\_Furn HSHD\_Exp\_Hhop HSHD\_Exp\_Misc HSHD\_Exp\_Pers HSHD\_Exp\_Read HSHD\_Exp\_Toba HSHD\_Exp\_Tran HSHD\_Exp\_Util;**

**run;**

**/\*---------------------- WARD'S --------------------\*/**

**/\* title2 'Using METHOD=WARD'S'; \*/**

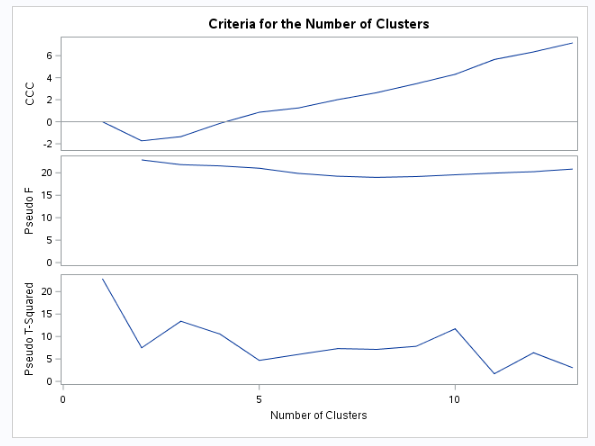
**ods graphics on;**

**proc cluster data=Ace method=ward ccc pseudo print=80 outtree=Tree;**

**var can1 can2 can3 Can4 Can5 Can6 Can7 Can8 Can9 Can10 Can11 Can12 Can13 Can14 Can15;**

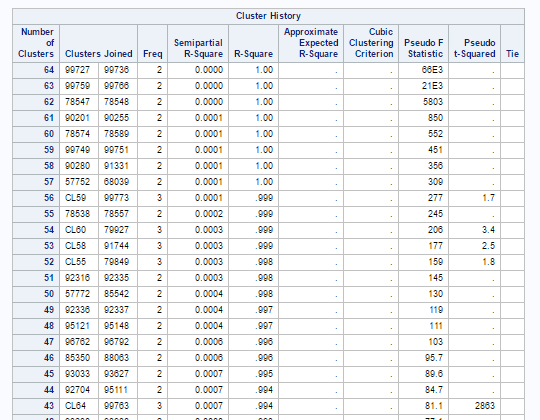
**id Zipcode;**

**run;**



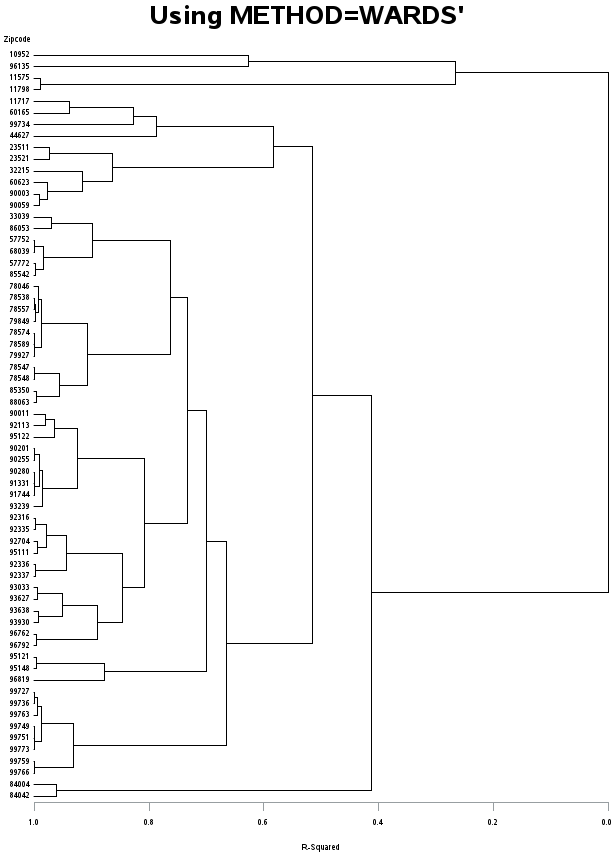
From the above image **“Criteria for the No of Clusters”** - Using CCC, Pseudo-F & Pseudo T-Squared images we can group them to 3 clusters.

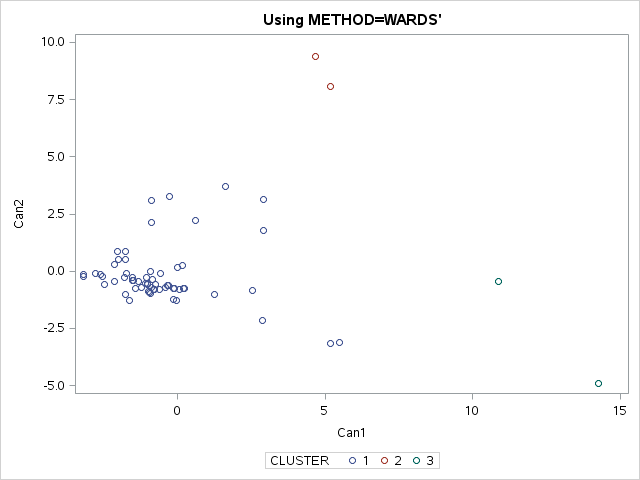
**Cluster history** below show how the clusters are formed. Overall 64 no’s of clusters are formed in the ward’s method. For instance, 59th cluster is formed by grouping 99749 & 99751 zip codes, like that 56th cluster is formed by grouping with 59th cluster & 99773 zip codes and so on. In this way totally 64 clusters are formed. Frequency shows the no. of zip codes in each cluster & semi partial R-Squared value is used as X axis in plotting the dendrogram in the further steps.



(As the output contain large no. of variable and Clusters, I couldn’t take the screen shot. Please refer the attached PDF: **“Ward’s\_Cluster\_Output4”** for detailed results)

Below dendrogram shows the how the Zip codes are grouped together to form a cluster.

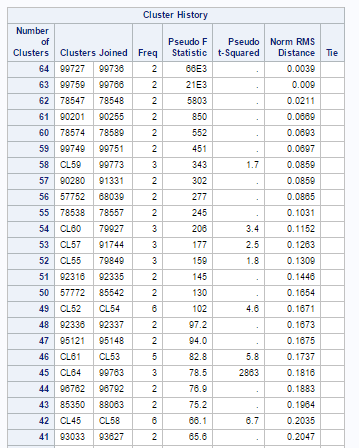


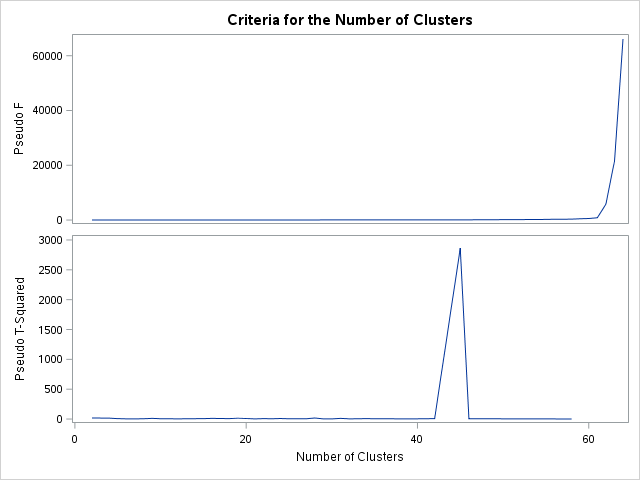
****

**Cluster Analysis for Top 10 Zip Code by each Factor - “Average Linkage Method”**

**Cluster history** below show how the clusters are formed. Overall 64 no’s of clusters are formed in the ward’s method. For instance, 59th cluster is formed by grouping 99749 & 99751 zip codes, like that 58th cluster is formed by grouping with 59th cluster & 99773 zip codes and so on. In this way totally 64 clusters are formed. Frequency shows the no. of zip codes in each cluster & semi partial R-Squared value is used as X axis in plotting the dendrogram in the further steps.

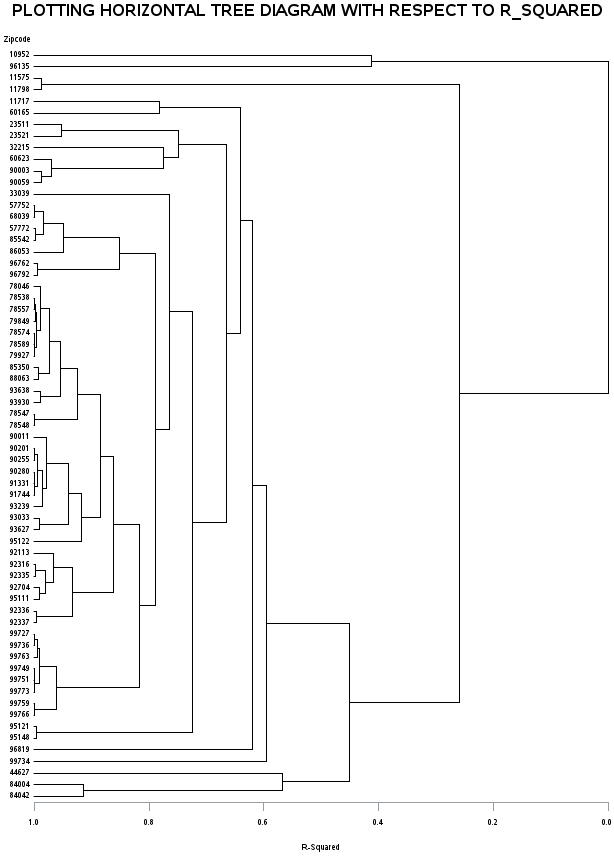
From the below image “Criteria for the No of Clusters” - Using CCC, Pseudo-F & Pseudo T-Squared images, nothing much information about grouping the cluster, I am restricting them to form to 3 clusters manually.

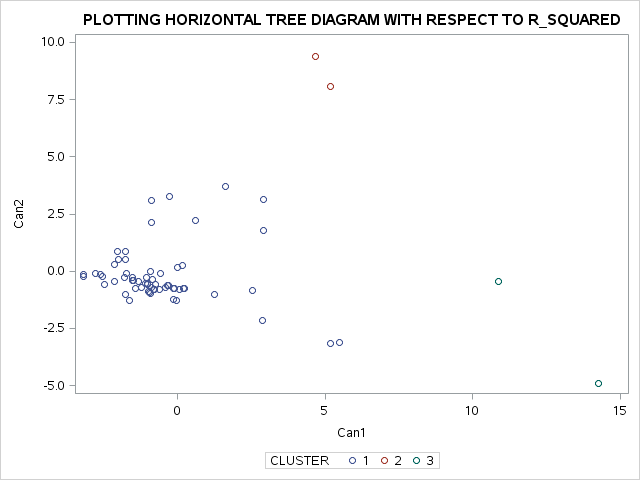


****

(As the output contain large no. of variable and Clusters, I couldn’t take the screen shot. Please refer the attached PDF: **“Average\_Cluster\_Output5”** for detailed results)

Below dendrogram shows the how the Zip codes are grouped together to form a cluster.

****

****

**Cluster Analysis for Top 10 Zip Code by each Factor – “Centroid Method”**

**/\*---------------------- Centroid method --------------------\*/**

**title2 'Using METHOD=CENTROID';**

**ods graphics on;**

**proc cluster data=Ace method=centroid pseudo print=80 outtree=Tree;**

**var can1 can2 can3 Can4 Can5 Can6 Can7 Can8 Can9 Can10 Can11 Can12 Can13 Can14 Can15;**

**id Zipcode;**

**run;**

**\*\*\*\* \*\*\*\* \* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\* \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;**

**title 'PLOTTING HORIZONTAL TREE DIAGRAM WITH RESPECT TO R\_SQUARED';**

**goptions vsize=9in hsize=6.4in htext=.9pct htitle=3pct;**

**axis1 order=(0 to 1 by 0.2);**

**proc tree data=Tree out=New nclusters=3**

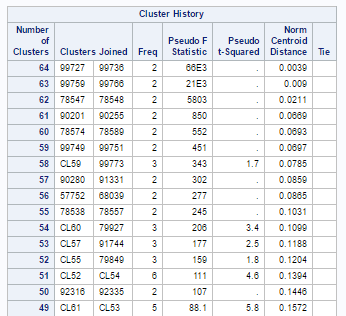
**haxis=axis1 horizontal;**

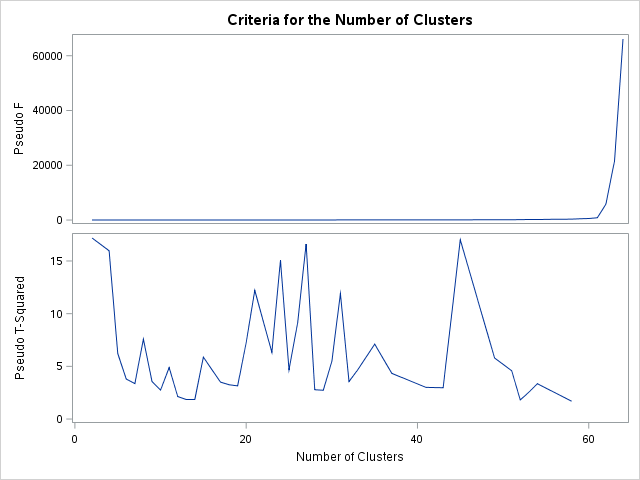
**height \_rsq\_;**

**copy can1 can2 can3;**

**id Zipcode; run;**

**Cluster history** below show how the clusters are formed. Overall 64 no’s of clusters are formed in the ward’s method. For instance, 59th cluster is formed by grouping 99749 & 99751 zip codes, like that 58th cluster is formed by grouping with 59th cluster & 99773 zip codes and so on. In this way totally 64 clusters are formed. Frequency shows the no. of zip codes in each cluster & semi partial R-Squared value is used as X axis in plotting the dendrogram in the further steps.

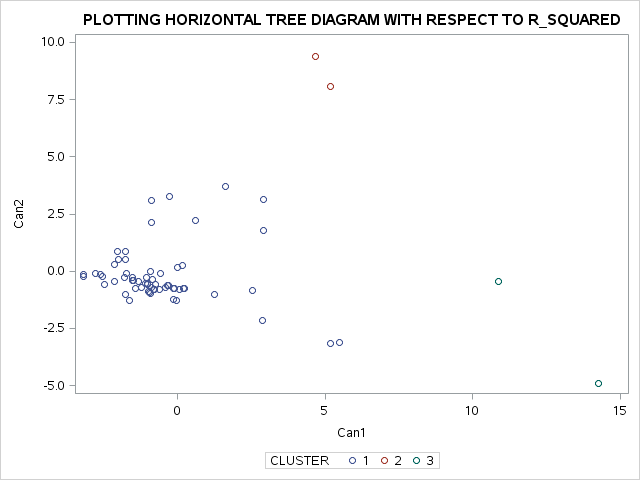


****

From the above image “Criteria for the No of Clusters” - Using Pseudo-F & Pseudo T-Squared images, we can restrict them to 3 clusters.

(As the output contain large no. of variable and Clusters, I couldn’t take the screen shot. Please refer the attached PDF: **“Centroid\_Cluster\_Output5”** for detailed results)

Below scatter plot shows the how the Zip codes are grouped together to form clusters.

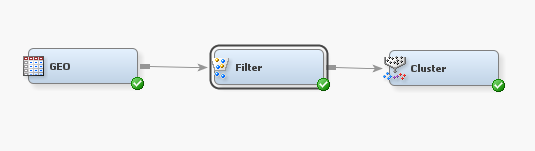
****

**Methodology 5:**

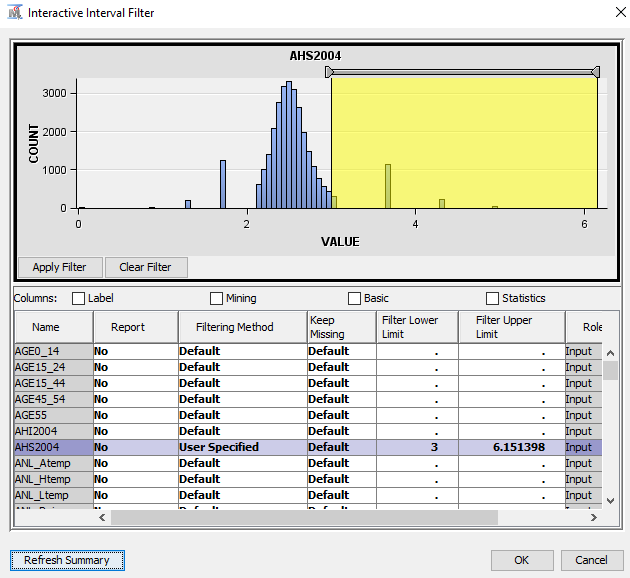
**Cluster Analysis Using “enterprise Miner” – “Ward’s Method”**

Filter node – Used to filter the data

Cluster Node – Used to generate the cluster using filter test samples

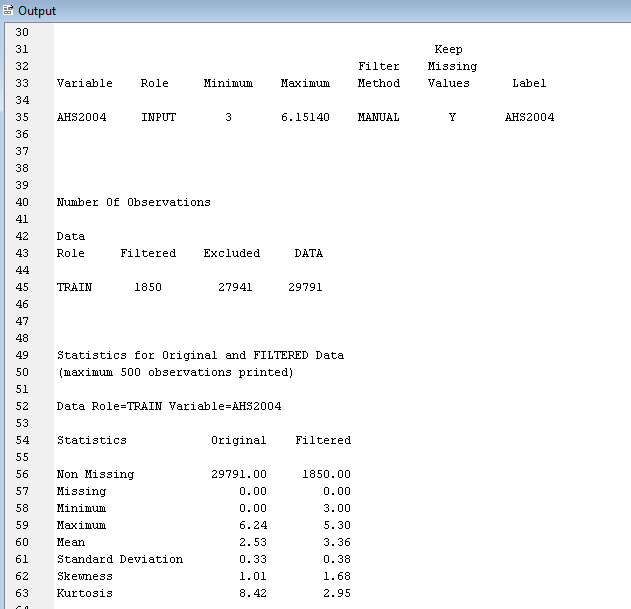


Using the interactive filter (shown below) from Filter node, Average Household Size lower limit is set to 3, hence the zip codes with 3 more Average Household size will be filtered out as a test sample for our analysis.



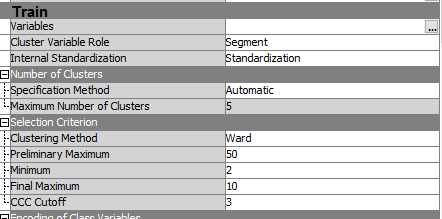
Below diagram shows the output after running the filter node. After limiting the AHS2004(average household size) to Range (Min-3 to Max-6.151), No. of observation shows the summary of Training data set. It says out of 29791 samples (original dataset), 27941 are exclude (means AHS for these samples must be lesser that 3). 1850 sample are considered as training dataset for our analysis.

Further below, it shows some statistics figures about original and test datasets.

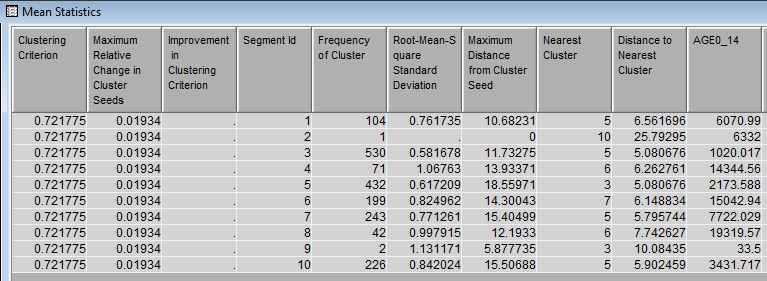


Cluster Node:

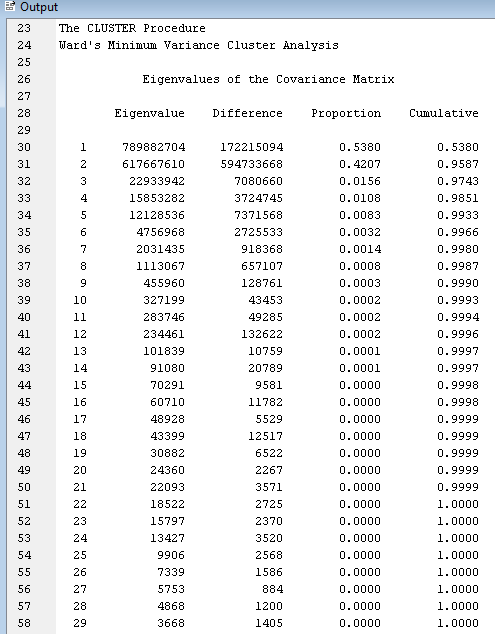
Data is standardized, Used Ward method & Number of Cluster set to automatic as shown below.



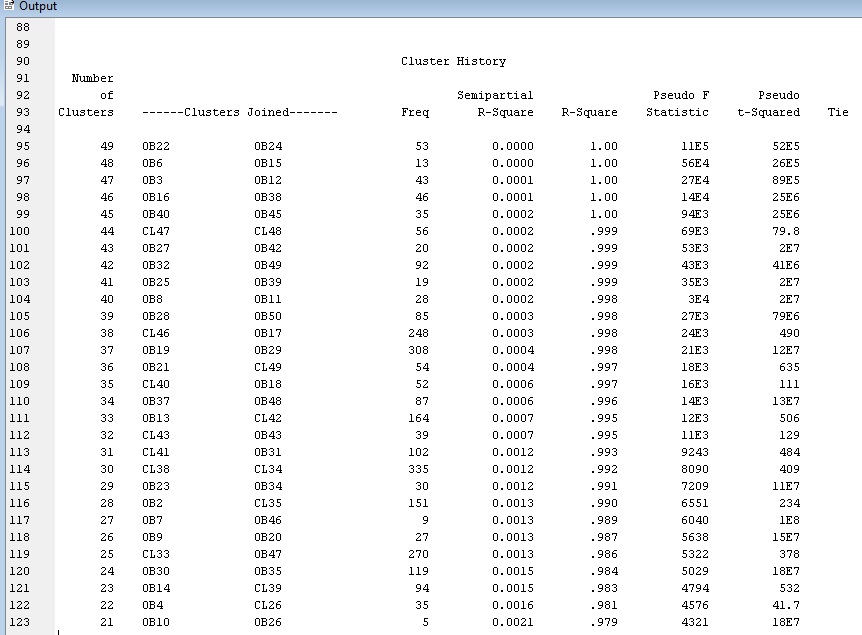
Mean Statistics: After running the cluster node, 10 clusters are formed as show below with frequencies highlighted. As seen in the Mean Statistics window, segment frequency counts vary from 1 cases to more than 500 cases.



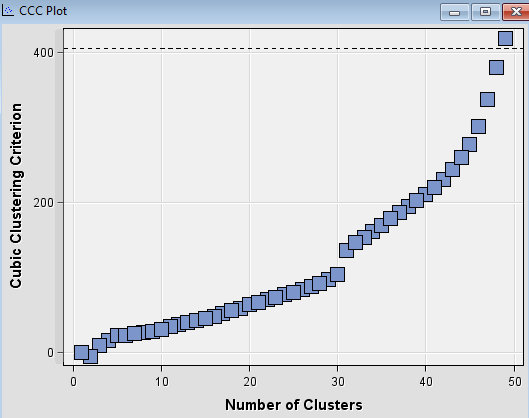
Below **output shows** the details such as Cluster Procedure Used: Ward’s Method, Eigen Value of Covariance Matrix



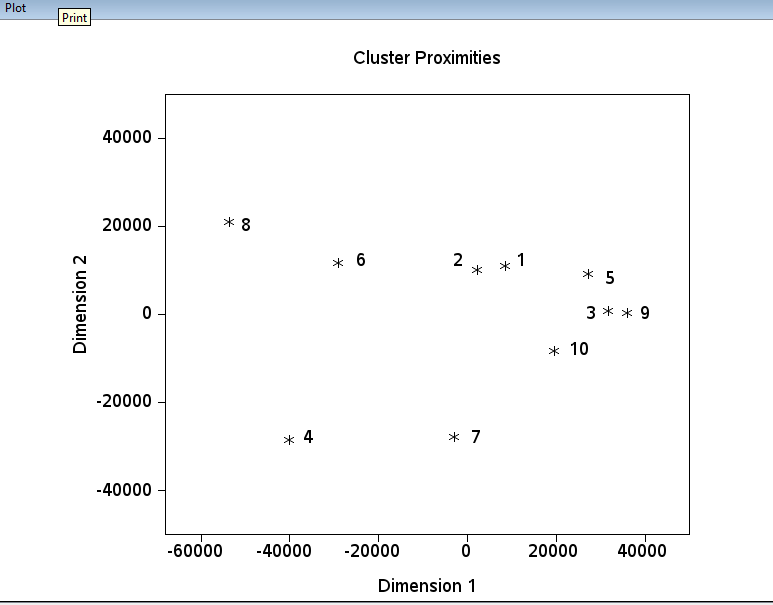
**Cluster history** below show how the clusters are formed. Overall 49 no’s of clusters are formed in the ward’s method. For instance, 49th cluster is formed by grouping 22nd observation & 24th observation zip codes, like that 44th cluster is formed by grouping with 47th cluster & 48th cluster of zip codes and so on. In this way totally 49 clusters are formed. Frequency shows the no. of zip codes in each cluster.



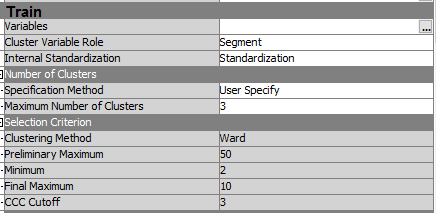
CCC plot could be used to decide the no of cluster outcomes.



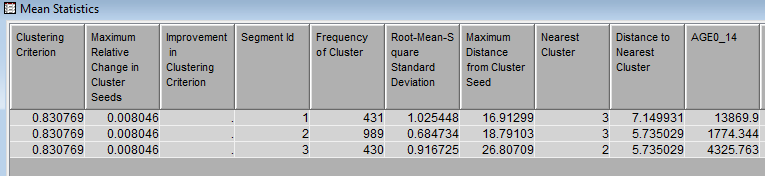
Plot below show the all the 10 cluster proximities in the Two dimensional space



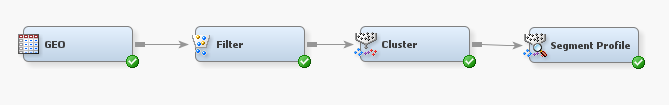
**User Specific Clusters:** Changed the number of clusters to user specific and restrict the maximum cluster to 3



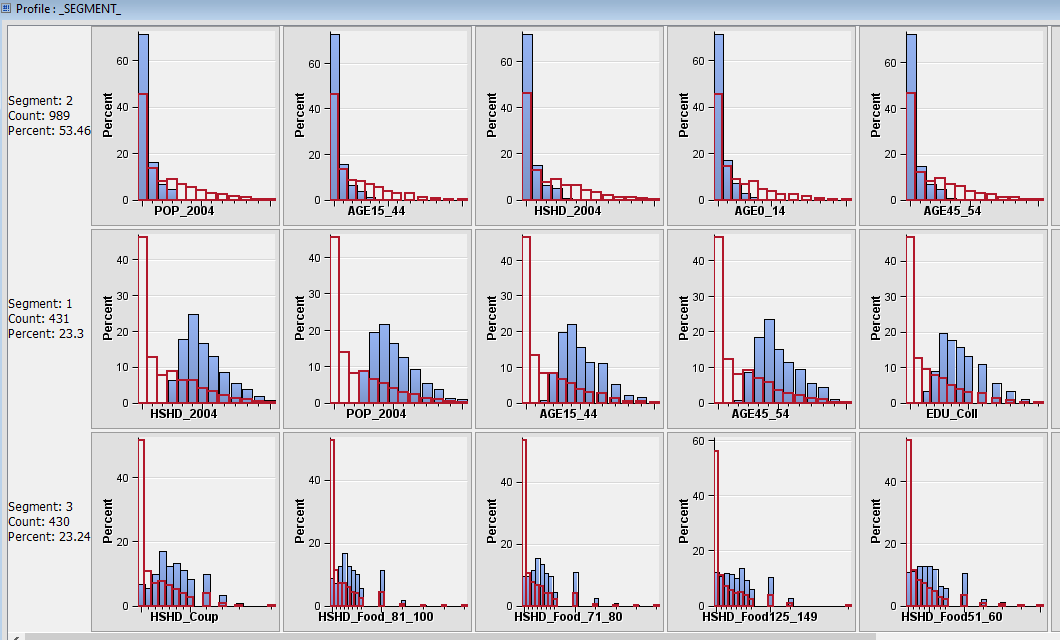
Mean Statistics: After running the cluster node, 10 clusters are formed as show below with frequencies highlighted. As seen in the Mean Statistics window, segment frequency counts vary from 430 cases to more than 980 cases.

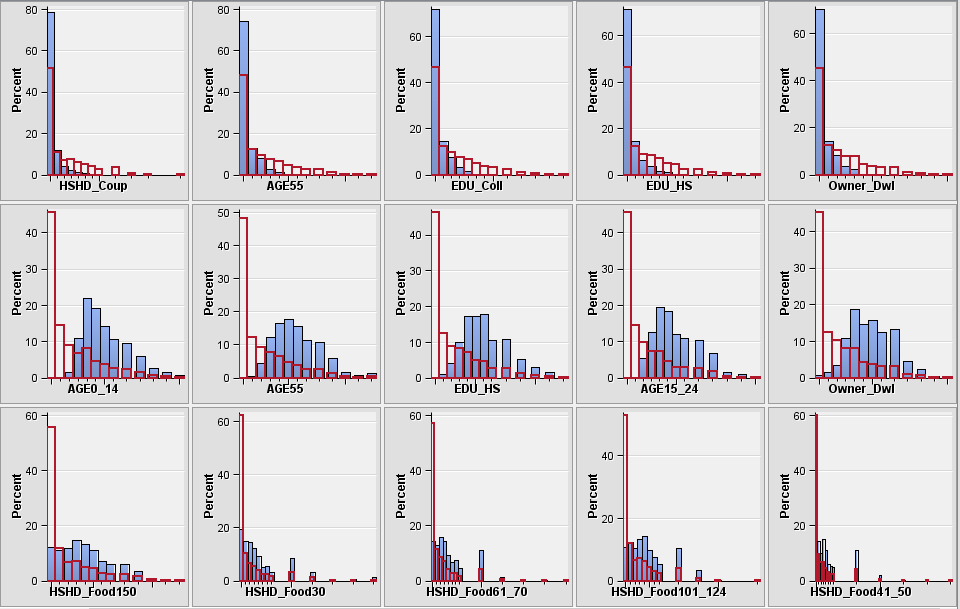


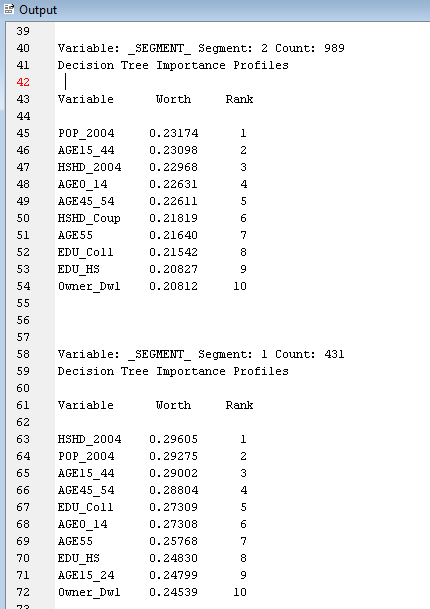
Included the **segment profiling** node to analyze the clusters



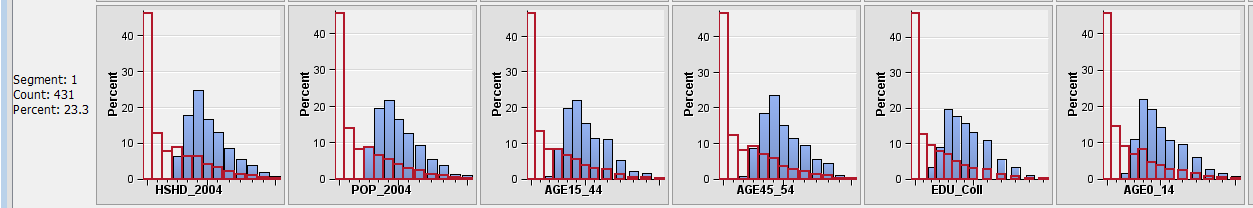
**Segment 2:** Below table shows the list of top10 most important independent variables, helps to explain the characteristics of this cluster. This 2nd cluster consist of 989 zip code, have almost 53% of the samples from test dataset.

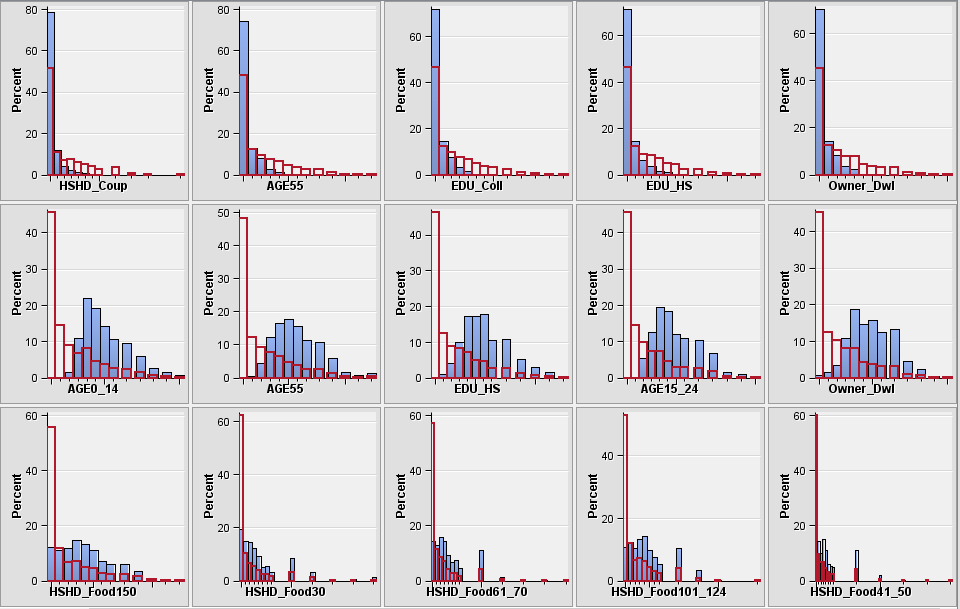


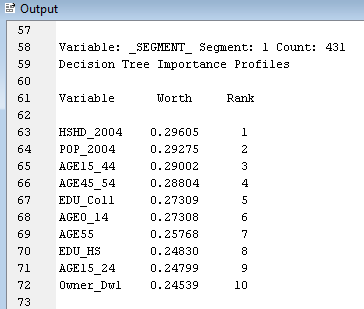




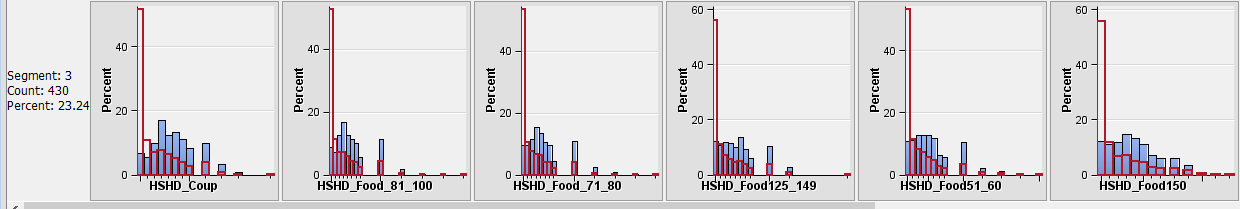
**Segment 1:** Below table shows the list of top10 most important independent variables, helps to explain the characteristics of 1st cluster. This cluster consists of 431 zip code, accounts for almost 23% of the samples from test dataset.

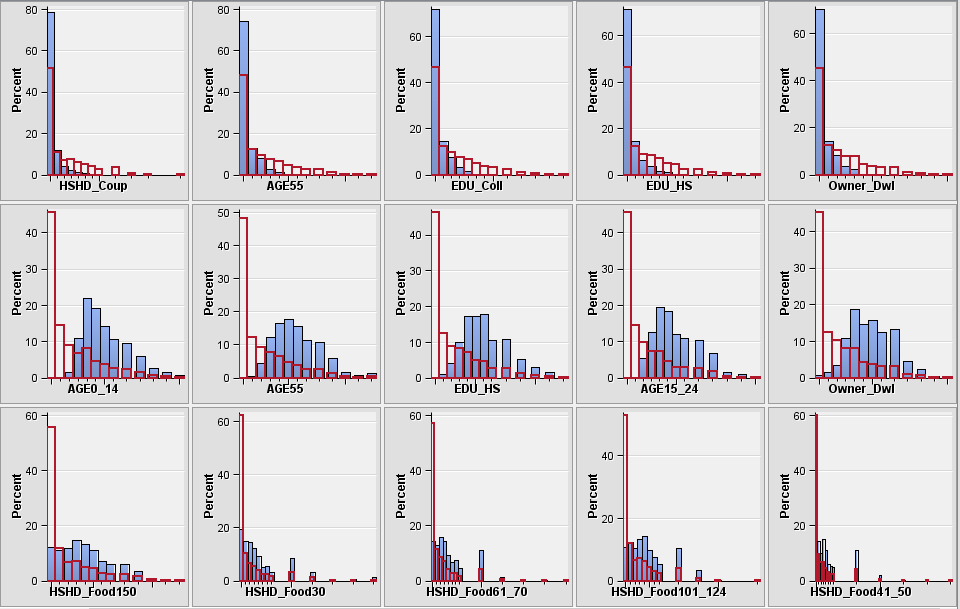


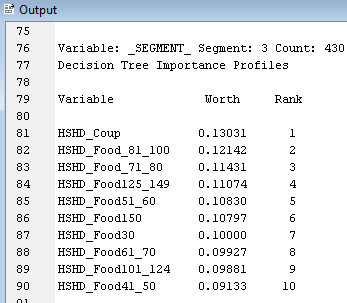




**Segment 3:** Below table shows the list of top10 most important independent variables, helps to explain the characteristics of 3rd cluster. This cluster consists of 430 zip code, accounts for almost 23% of the samples from test dataset.







**Expected findings**

As expected the Demographic, Socio-economic & variable directly affect the consumer behavior and their spending patterns. We saw that better the environmental condition, people tend to spend more. We also realized that higher the socio-economic status, higher the consumer spends.

**Discussion of the analysis process, take-away insights:**

**Factor Analysis**

It Helped to reduce the no of independent variables, we could reduce to only 8 independent factors, which characterizing demographic, socio-economic and environmental variable.

**Regression Analysis**

Using this we could able to prove that as expected Demographic, Socio-economic & variable directly affect the consumer behavior and their spending patterns. In most of the regression analysis of expenditure variable we saw, higher the socio-economic status, higher the consumer spends & better the environmental condition, higher the consumer spends.

**Correlation Analysis**

There was not much variation in the output of regression analysis, it may be because as seen by correlation analysis results, all the expenditure variables are correlated to each other.

Cluster Analysis: Top 10 Zip Codes for each factor has been considered to perform the cluster analysis. In forming the clusters, only the expenditure variables are taken into consideration.

Below mentioned 3 methods have been used for forming the cluster:

* Ward’s
* Average Linkage
* Centroid Method

The result generated we not that significantly different from each other, managed to create 3 clusters from each method.

**Correlation Analysis (Using Enterpriser Miner):**

Filter node – Used to filter the data

Cluster Node – Used to generate the cluster using filter test samples

Segment profile Node – For analyzing and interpreting the cluster

Filter Criteria: Sample with Average Household Size (AHS2004) > 3 has been selected for analysis

Cluster criteria: Data is standardized, Used Ward method & Number of Cluster set to automatic. 10 Cluster has been generated from the test dataset (1850 samples) frequency counts vary from 1 to more than 530 cases.

**Recommendations**

**Relate the results to marketing decisions**

Expenditure on **Apparel, Education, Entertainment, Food/Beverage, Gift, Health Care**, **Furnishing, Household Operations, Personal Care, spending on Reading, on Tobacco, on Transportation and Utilities**; shows better the demographics & Socio-economic conditions higher the spends, as the climatic conditions are unfavorable the spending on **most of these categories** decreases.

**Example:**

Top 5 Zip Codes of **Factor1** - 60623, 90201, 90280 91331, 92335, in these zip codes the **Ave Household Expenditure on Entertainment** significantly increases as one unite increase in No. of households, Population, Age group 15 to 55 and above, No of Owner & Renter occupied dwellings or increase in the Schooling to professional degree holders.

Top 5 Zip Codes of **Factor 4** 10952, 84004, 95121, 95148, 96819 in these zip codes the **Ave Household Expenditure on Entertainment** significantly increases as one unite increase in Household income, No of Graduates or Professional degree & Bachelor’s degree holders, or with White & Asian/pacific population.

Top 5 Zip codes of **Factor 6** 23521 32215 33039 78574 78589, in these zip codes the **Ave Household Expenditure on Entertainment** significantly decrease as one unite increase in environmental variables like Hurricane, Tornado Index, Weather risk and July Low temperature

We could generalize these rules & build the marketing strategy across all the zip codes as stated below.

**Recommendations for answering the research questions:**

**Top Zip Code to be Focus based on factors influence (independent variable) :**

Identify which factor are the most significant in explaining the output, identify what Demographic, Socio-economic and Environmental variable takes the major load, simultaneously identify which are the top Zip Codes need to be focused and targeted from the table below. Search for those top Zip Codes from the generated Clusters and apply the same marketing strategy across all the Zip codes



**Limitations**

**Insights about the limitations of the current dataset**

1. Details are not covered at more granular level (at each customer level)
2. We cannot customize the offer for individuals
3. Hard to analyze the individual needs and demands

**Insights about the limitations of the current model and analysis methods**

1. Regression analysis results for all the expenditure variable are almost same, maybe we are not considering the time series factor into consideration, for the single zip code (data over the period of time). That could help to sense more meaning out of it.
2. Supporting to that the correlation between the expenditure variables are same and all are highly correlated to each other.
3. Cluster analysis could help classifying the zip codes into different groups, but in identifying the characters of each cluster are always to opaque. Most of the guidelines for using cluster analysis are rules of thumb. There are no completely satisfactory methods for determining the appropriate number of clusters

**Future plan for pursuing in the same directions**

1. I Would consider the starting step as analyzing at single zip code using historical data
2. In regression analysis, we are not really checking the assumptions of regression analysis. I would rather consider checking all the assumption before proceed it further.