Cluster this!

June 2011

Agenda

On the agenda today:

- SAS Enterprise Miner (some of the pros and cons of using)
- How multivariate statistics can be applied to a business problem using clustering
- Some cool variable reduction methods
- Type of modelling techniques possible and scenarios where each is applicable
- How to evaluate the cluster models once built

Power of applied MV statistics

- Applied statistical methods can be a very powerful tool in answering some difficult business problems.

- Often it has the stigma of being difficult to understand since some methods are very complex such as multivariate analysis (MV)! But using tools such as SAS Enterprise Miner and Enterprise Guide can assist you in helping explain some of the more complex methods through graphs, visualizations and other diagnostics.

- In banking where on occasion the business problems can be complex multivariate statistics can come in handy since it helps uncover patterns not easily revealed by simple statistics.

- Today we will discuss in more detail the specific MV method of clustering..

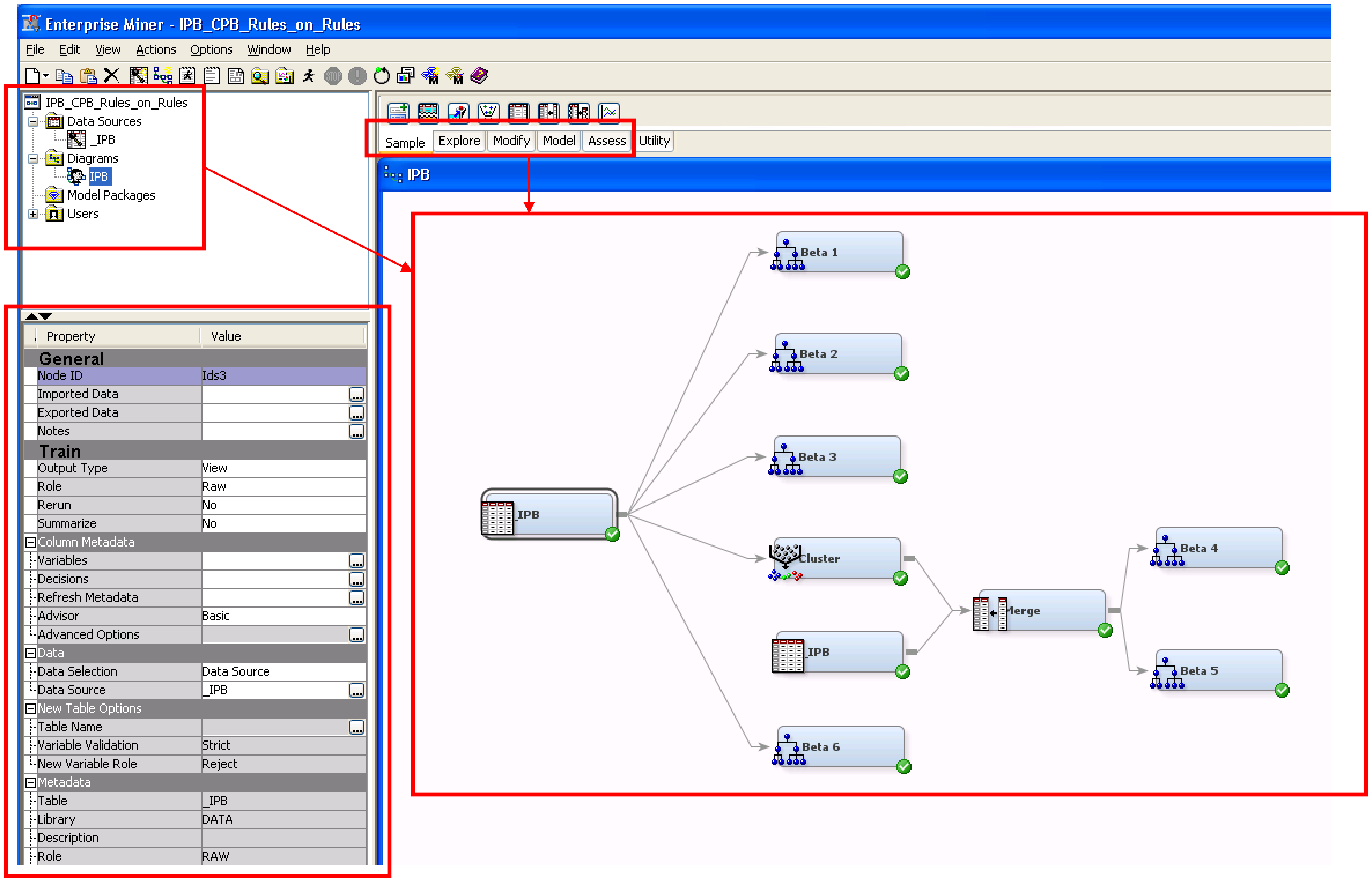
What is clustering?

By examining more than one characteristic, similar cases can be group together into a ‘cluster’. Clusters are distinguished from each other based on the differences.

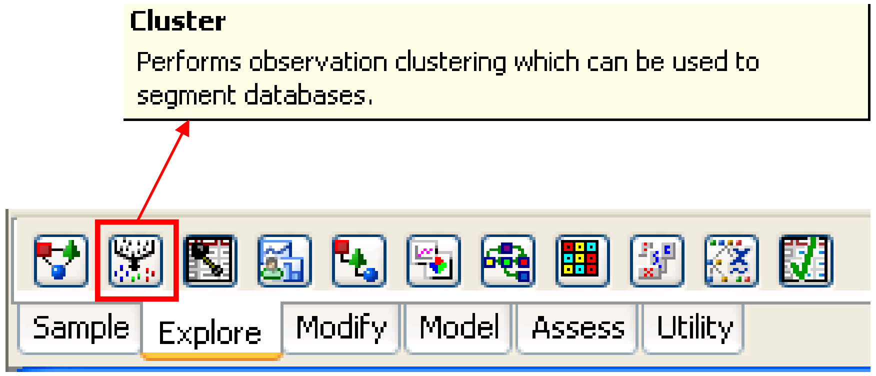
- Although you may need a PhD develop the statistics to create new clustering techniques you certainly do not need a PhD to understand when and how to apply it!
- Can solve difficult questions or business problems with clustering
- As an example think about the classification of cars.. Different types of vehicles on the surface you can say all have 4 doors and engines, but by digging a little deeper and looking at it on a number of variables such as engine type, fuel efficiency, luxury options, you can start to create a very distinct taxonomy: sports cars, SUVs, etc.

Using Enterprise Guide to cluster

We all know that Enterprise Miner can be used to do modelling! It uses the SEMMA -> **S**ample; **E**xplore; **M**odify; **M**odel; and **A**ssess framework to build and deploy models quickly!!



Nodal structure



Sample:

- Functions related to data sampling such as appending, partitions, file import, merging etc.

Explore:

- Functions related to finding relationships in your data, such as Multi-plots, associations, clustering, self organizing maps etc.

Modify:

- Transform your data, but imputing, creating new variables, consolidating (PCA) etc.

Model:

- Run various modeling frameworks, neural networks, decision trees, regression etc.

Assess:

- Evaluate and measure model performance

Utility:

- Run custom SAS code, edit metadata, control points etc.

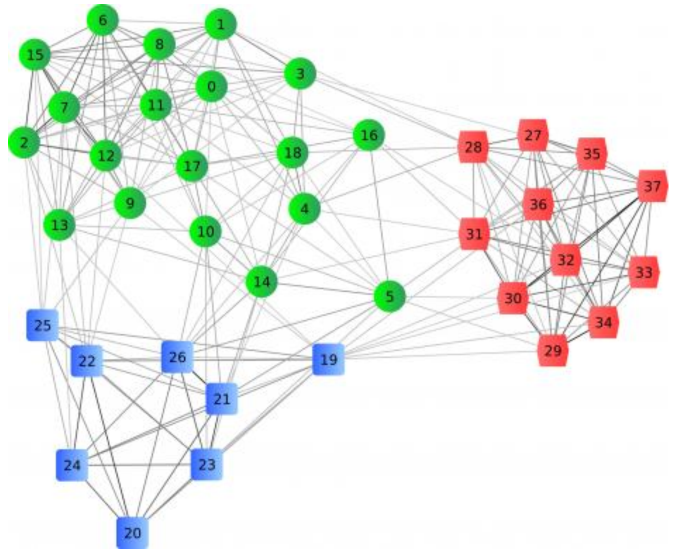
SEMMA outside of Enterprise Miner

SEMMA is a way to organize your models and has nothing to do with the modelling itself.

- You can implement SEMMA without using EM
- Limited *customizable* visualization techniques with EM
- Limited options to customize (not all modeling options are listed, for example in clustering limited to average, centriod, and Wards), great for black box environments!!
- Not so good for audit when you have to explain yourself
- Already do the Sampling, Exploring, Modifying, and Assessing outside of Enterprise Miner, only bring it in to do some Modelling!

Business Problem: UTR reporting from Branches

From a UTR reporting perspective many of the best leads of suspicious activity occur at the Branch level as these cases have already been looked at by a human being.



From a regulatory perspective, these cases must be reviewed upon receipt.

- Accurate reporting is important, but more importantly want to ensure each branch is reporting what they should be.
- How to compare whether a branch is reporting accurately? This can be difficult since branches can vary widely from area to area, differences in size, type of business they conduct, area, and general client demographic and greatly effect the amount of UTR reporting that a branch does.

How clustering can assist in your analysis

Defined our business problem: identify branches that reporting zero UTRs, or under reporting the number of UTRs

- Using a cluster model will assist in determining similar branches and group them together.

- Once this task is complete, the analysis can be continued by examining branches within a cluster with each other to determine who appears to be conducting normal vs. non-normal activity.

- A very powerful tool to profile and group data together. Very good method to find similarities between branches by digging deeper and finding connections that are not apparent.

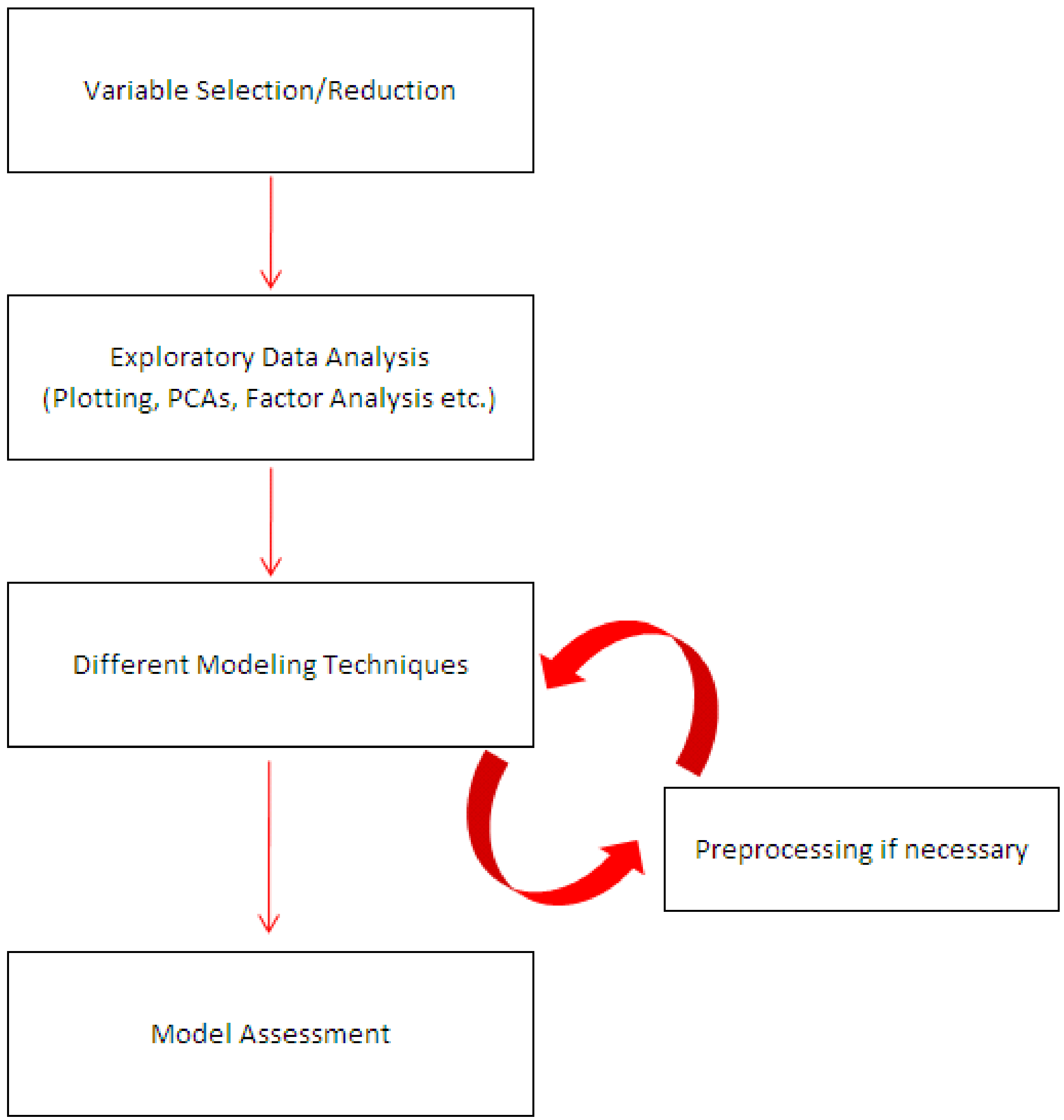
How do you build a cluster model

Several steps that you need to do when building out a cluster model:

- Data Gathering
- Data Cleansing
- Perform the Clustering
- Evaluate the Clusters

} 85% of your time will be spent on these steps, as they are the most time intensive, and likely to skew results if not done properly, GIGO!

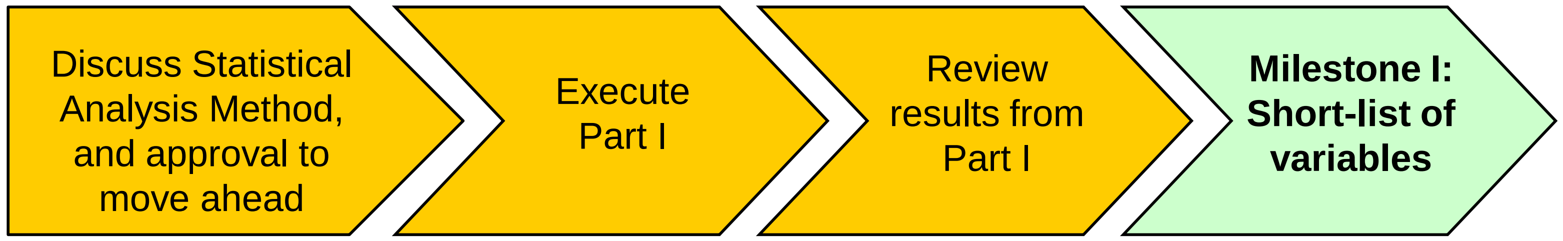
Overall Process Methodology



Variable Selection

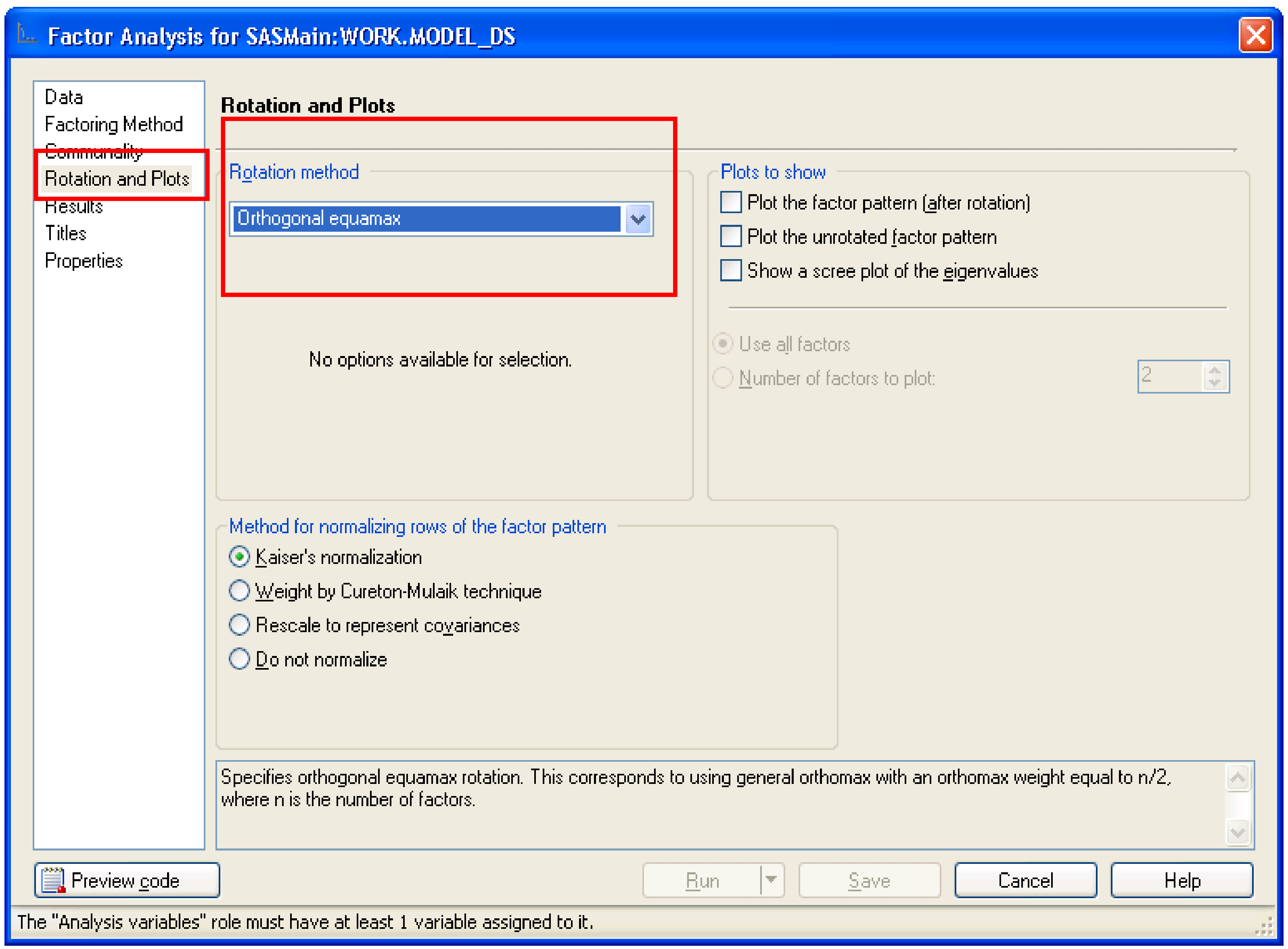
Variable selection can be done using a variety of analysis techniques such as principal components, factor analysis, SAS process such as varclus, or straight correlations. In our exercise this portion of the analysis also included:

- Ensuring the data files are accurate
- Looking for Outliers in the data
- Checking if there are restricted ranges in the continuous variables
- Checking if there are unequal cell sizes in categorical variables
- Distributions of the variables
- Co-linearity between variables
- Covariance matrices are homogeneous
- Extent and nature of missing data



Factor Analysis

Factor analysis and PCAs are very similar in outcome but the roads and reasons for using either are very different, as the assumptions when using either also differ greatly.



Variable Selection and Exploratory Data Analysis

Proc factor was run to reduce the number of redundant variables down to 35:

Type	Rotated Factor Pattern						
			Factor1	Factor2	Factor3	Factor4	Factor5
Population data (Stats Canada)	E09TENOWN	Owned	0.94536	0.22234	0.04879	0.15446	-0.03158
	E09FSFAMT	Total Census Families	0.94168	0.19639	0.18281	0.16141	-0.06284
	E09CHTOT	Total Number of Children at Home	0.9405	0.13655	0.04479	0.13254	-0.12657
	E09TYHOUSE	Houses	0.93315	0.12103	-0.23162	0.10705	-0.02141
	WSMORTI	Mortgage - Incidence	0.92036	0.24026	-0.00089	0.12336	-0.08451
	WSNETREI	Net Real Estate Incidence	0.91491	0.28241	0.03539	0.15554	-0.01483
	E09HHPOPIM	Household Population For Immigration	0.89965	0.24276	0.28411	0.19343	-0.08768
	E09HHPOP	Number of Persons In Private Households	0.89953	0.24263	0.28412	0.19321	-0.09046
	E09PFTOT	Females	0.89903	0.24349	0.28483	0.19359	-0.0776
	E09TOTPOP	Total Population	0.896	0.24627	0.29172	0.19594	-0.08809
	E09PMTOT	Males	0.89127	0.24864	0.29818	0.19797	-0.09859
	WSSECLOCI	Secured Lines of Credit - Incidence	0.88696	0.30464	0.06879	0.14666	-0.054
	E09HHPOP15	Household Population 15 Years or Over	0.87889	0.26329	0.32049	0.21313	-0.07515
	E09POP15P	Total Population 15 Years or Over	0.87431	0.26706	0.32863	0.21586	-0.07252
Investments and Financial data	WSINVESTB	Total Investments-Balance	0.2373	0.93933	0.15167	0.16727	-0.03467
	WSSTOKORB	Stocks outside of RRSP-Balance	0.19483	0.93265	0.12609	0.16341	-0.04447
	WSLIQASTV	WealthScapes Liquid Assets - Value	0.28314	0.92858	0.16264	0.16094	-0.02566
	WSRRSPB	Total RRSPs-Balance	0.31415	0.91886	0.16478	0.13788	-0.02121
	WSFUNDORB	Mutual Funds outside of RRSP-Balance	0.24173	0.90808	0.15633	0.20351	-0.0263
	WSCHQSAVB	Chequing & Savings Accounts - Balance	0.30565	0.89372	0.20526	0.14945	-0.03944
	WSBONDIRB	Bonds in RRSP-Balance	0.0454	0.89103	0.19733	0.05054	-0.01834
	WSFUNDIRB	Mutual Funds in RRSP-Balance	0.38856	0.87961	0.11572	0.15938	-0.02652
	WSSTOKIRB	Stocks in RRSP-Balance	0.16144	0.85949	0.2315	0.12804	-0.04277
	WSSAVNGB	Total Savings - Balance	0.40911	0.84288	0.18745	0.13224	0.00366
Housing Types	E09TYAPT	Apartment, Building Low and High Rise	0.27085	0.35854	0.84433	0.23383	-0.05944
	E09TYAPT5P	Apartment, Building that has Five or more Storeys	0.0963	0.3929	0.72014	0.21904	-0.07034
	E09IMNPERM	Non-Permanent Residents	0.1607	0.41106	0.71665	0.25138	-0.17363
	E09TYAPT_5	Apartment, Building that has fewer than Five	0.34691	0.19416	0.66205	0.16377	-0.02699
Branch details	tot_fte_pte_cnt	Total count of FTE and PTE per transit	0.3451	0.29149	0.26943	0.80265	-0.07188
	Personal	Total number of personal clients	0.39611	0.25968	0.31906	0.78136	-0.02832
	branch_resize	Branch size	0.40265	0.20774	0.13433	0.76942	-0.13153
	Business	Total number of business clients	0.28452	0.39453	0.27341	0.72144	-0.06551
Age	tot_pop_age_med	Median Total Population Age	-0.11474	-0.035	-0.06711	-0.05975	0.98576
	female_age_med	Median Female Age	-0.10788	-0.046	-0.06622	-0.06013	0.96943
	male_age_med	Median Male Age	-0.12488	-0.0226	-0.06735	-0.06116	0.96616

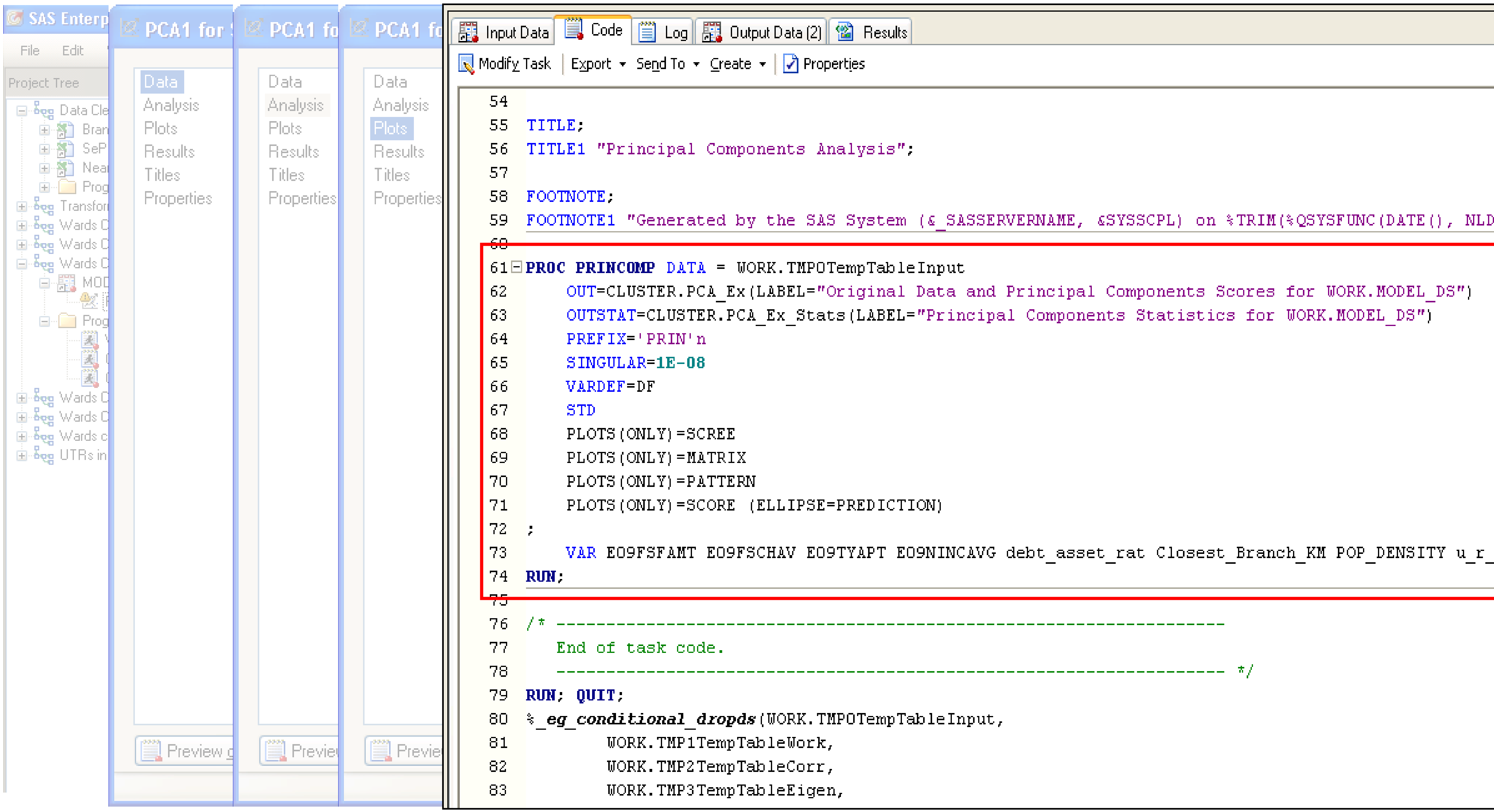
Primary use of either Principal components or factor analysis is both data reduction and summarization. Getting the most bang for your buck, that is, less is more! With PCA you are accounting for the maximum variance in a minimal number of variables ('super-variables').

Its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data.

The results of a PCA are usually discussed in terms of component scores (the transformed variable values corresponding to a particular case in the data) and loadings (the weight by which each standardized original variable should be multiplied to get the component score)[1].

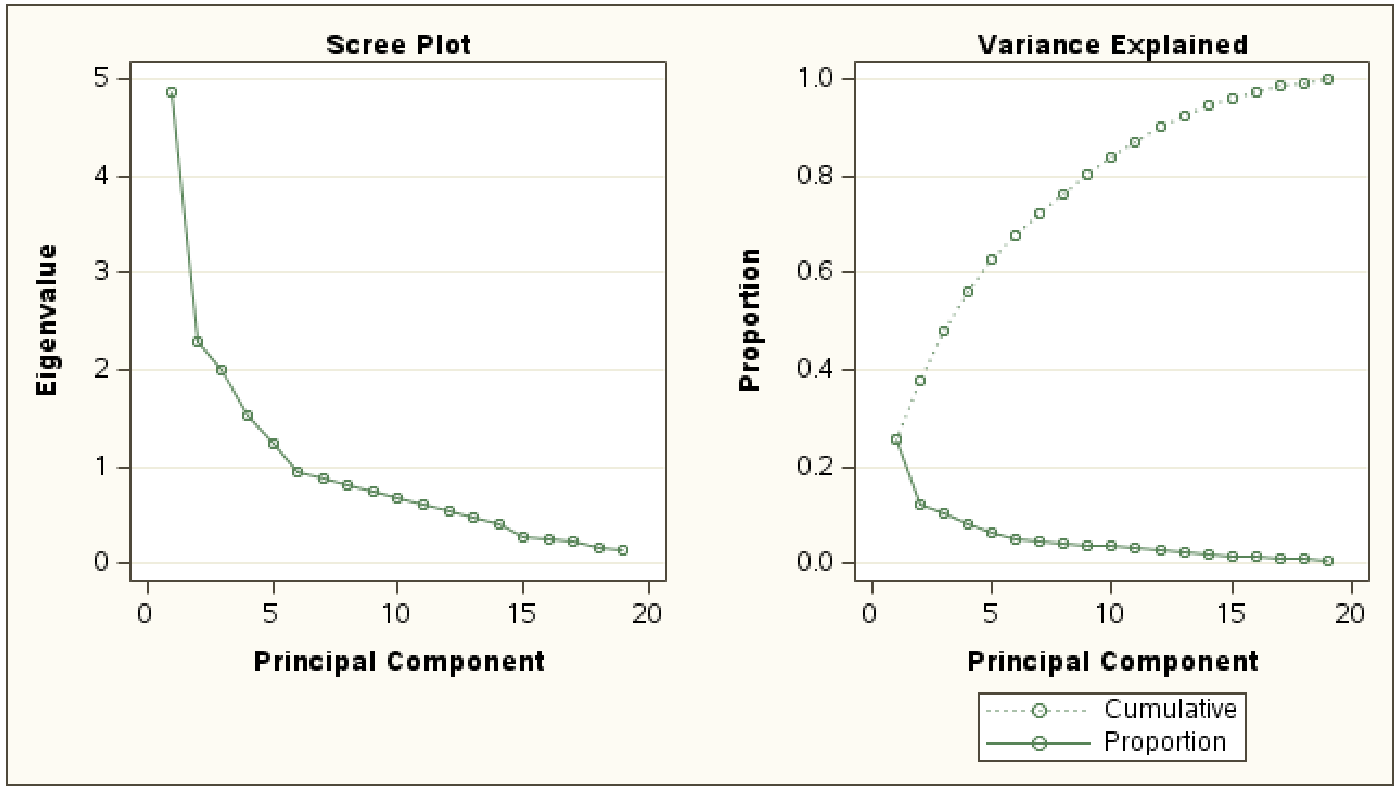
[1] Shaw PJA, Multivariate statistics for the Environmental Sciences, (2003) Hodder-Arnold

PCA in Enterprise Guide



Variance Explained by the PCA Factors

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.86041198	2.56953591	0.2558	0.2558
2	2.29087607	0.28314898	0.1206	0.3764
3	2.00772709	0.48349777	0.1057	0.4821
4	1.52422932	0.29499951	0.0802	0.5623
5	1.22922981	0.29044368	0.0647	0.627
6	0.93878613	0.07012215	0.0494	0.6764
7	0.86866398	0.06566382	0.0457	0.7221
8	0.80300016	0.06886492	0.0423	0.7644
9	0.73413524	0.06251745	0.0386	0.803
10	0.67161779	0.05258319	0.0353	0.8384
11	0.61903459	0.08102724	0.0326	0.8709
12	0.53800736	0.06471474	0.0283	0.8992
13	0.47329262	0.06678178	0.0249	0.9242
14	0.40651084	0.12376538	0.0214	0.9456
15	0.28274545	0.03647968	0.0149	0.9604
16	0.24626578	0.02706868	0.013	0.9734
17	0.2191971	0.06585729	0.0115	0.9849
18	0.15333981	0.0204109	0.0081	0.993
19	0.13292891		0.007	1



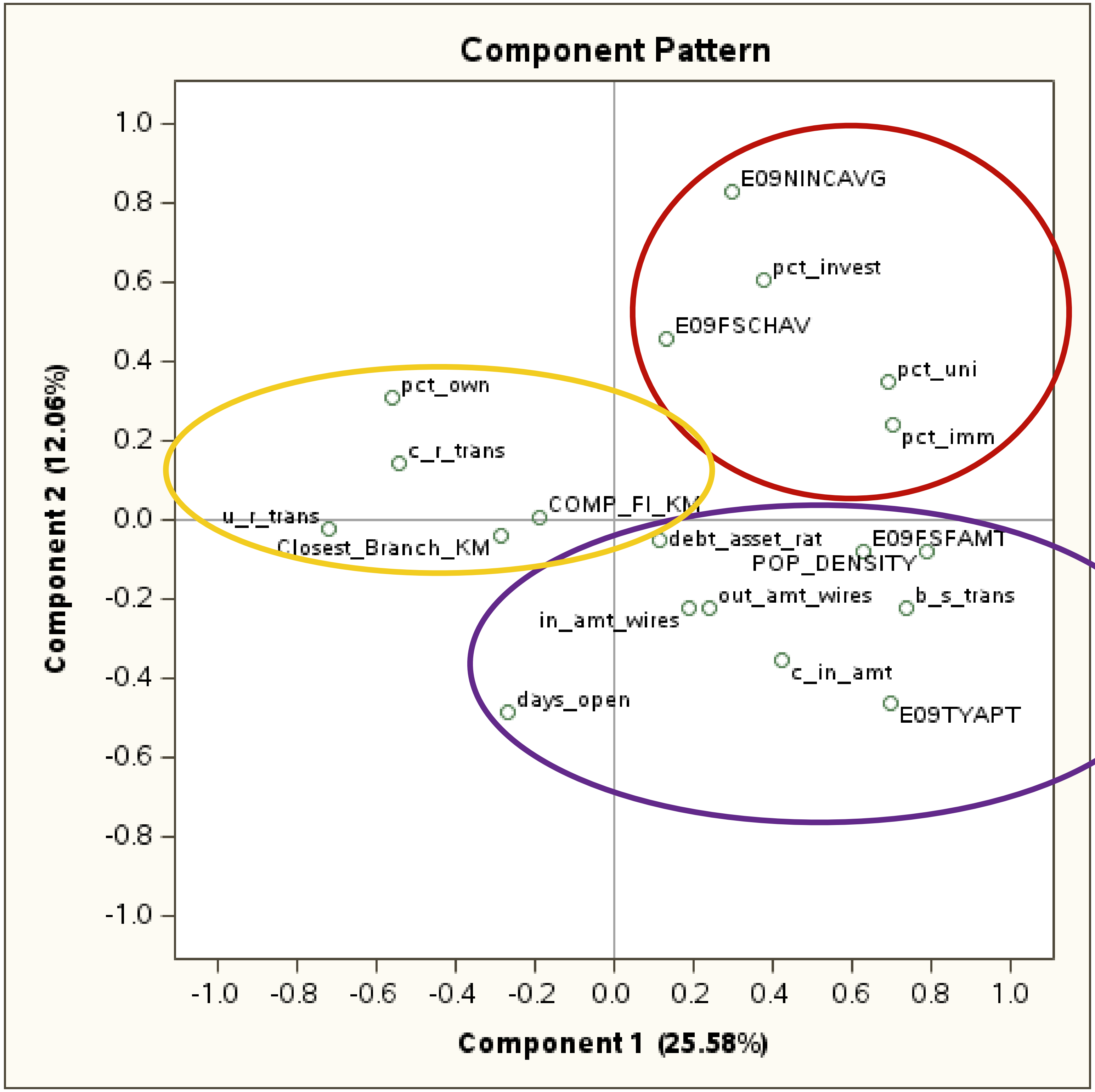
PCA Driving Factors

Driving Factors	Principal Components			
Primary Cluster Variables	1	2	3	4
% of Homeowners	-0.25	0.20	-0.07	-0.29
% of Population that Immigrated	0.32	0.16	0.11	0.03
% of University Graduates	0.31	0.23	-0.29	0.07
Amount Invested per Average Income	0.17	0.40	-0.35	0.06
Amount of Incoming Wires by Transit	0.09	-0.15	-0.12	0.08
Amount of Outgoing Wires by Transit	0.11	-0.15	-0.11	0.07
Average Income	0.13	0.55	-0.06	0.05
Average Number of Children per family	0.06	0.30	0.46	0.15
Branch Size	0.33	-0.15	0.11	-0.05
Cash in Amount	0.19	-0.23	0.15	0.05
Closest RBC Branch in KM	-0.13	-0.03	0.08	0.63
Crime Rank	-0.25	0.09	0.08	-0.04
Debt to Asset Ratio	0.05	-0.03	0.53	-0.10
Nearest Competitor in KM	-0.08	0.00	0.05	0.66
Number of Apartment Dwellers (Renters)	0.31	-0.31	-0.16	0.08
Number of Days Transit is open	-0.12	-0.32	-0.26	-0.02
Number of Families	0.29	-0.05	0.28	-0.08
Population Density	0.36	-0.05	-0.15	0.08
Urban or Rural area	-0.33	-0.02	-0.11	0.13

Strong Porportion

Positive Negative

PCA Plots show evidence of Clusters



What method to use?

Depends on what you want to do!

Objective

PCA
- Variables are relatively error free
- Small variances between variables
- Going from many (∞) to few

*** Account for maximum amount of variance in a minimum number of variables**

Common Factor Analysis
- Find underlying groups of variables
- Why things are related to each other
- Helps identify relationships to group data together (that which is not obvious)

*** What are your drivers? Attempts to identify latent constructs.**

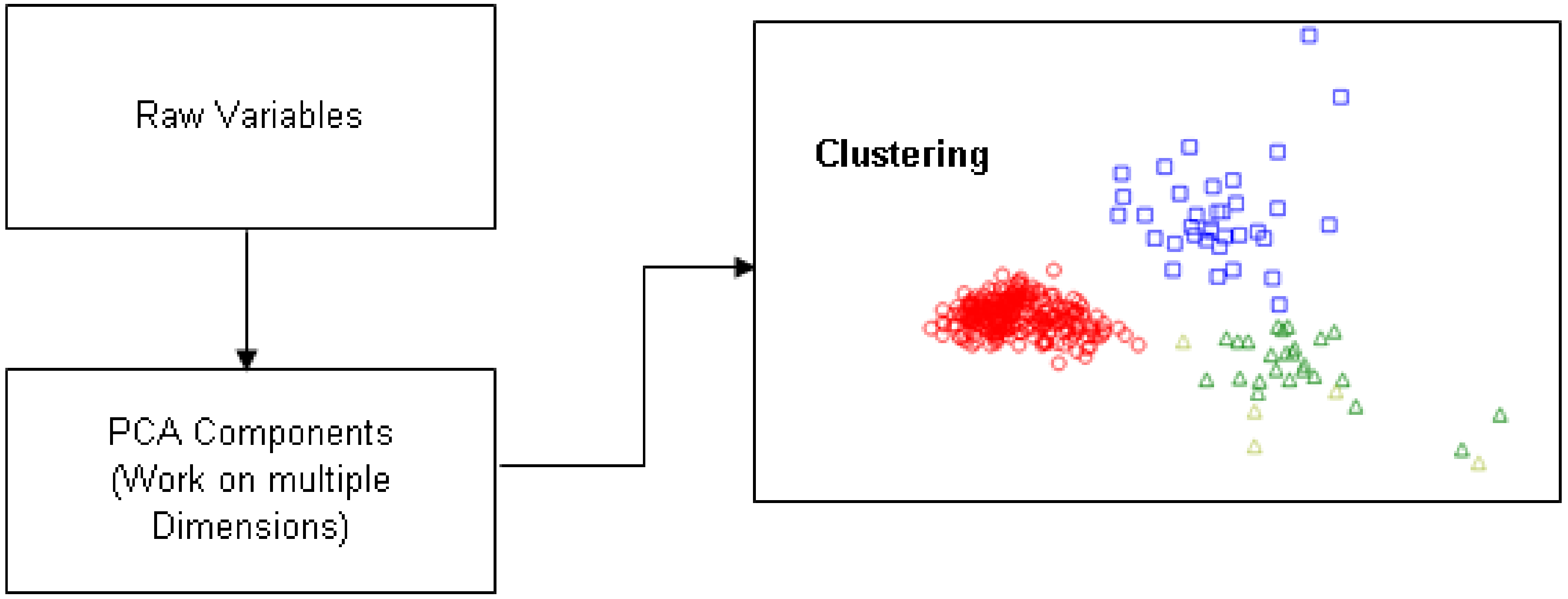
What Variables are most Important

Key Drivers	Description
Population related	Average Income, Population age, Population type by gender, and family unit (single, families), university graduates, homeowners, renters, children per family etc.
Geographic	Urban or Rural area, Population density etc.
Branch Related	Number of days branch open, number of competitors near by, closest branches, branch size etc.
Internal RBC Variable	Data warehouse related variables such as transit postal code, city, number of business and personal clients etc.

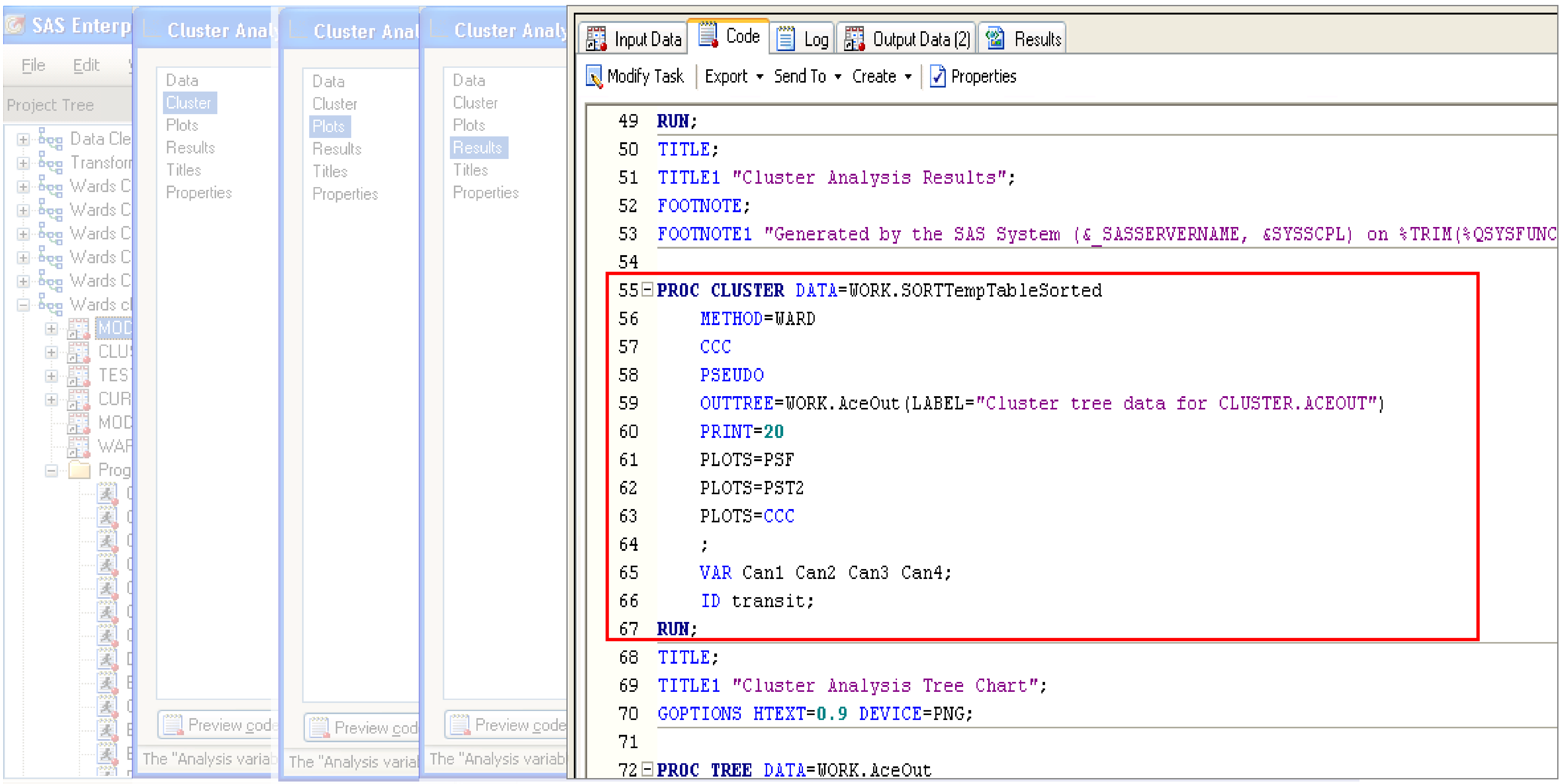
Eventually using a combination of techniques 300+ variables were reduced to about 35. Then PCA analysis was performing to get the most out of the 35 remaining variables.

Cluster Methodology

A short list of variables have already been identified through previous analysis, the following methodology was used to find meaningful groups of homogeneous branches:



Wards Clustering in Enterprise Guide



Semi-Partial R^2, R^2, and CCC

What statistics you can look at and why?

Like many SAS outputs, cluster output gives you a number of different statistics to look at to help evaluate, first if the clustering worked, secondly how many clusters are optimal for the solution. Referring to the output of Wards clustering, the following selected statistics are helpful:

1. SpRSq (semipartial R-squared) is a measure of the homogeneity of merged clusters, in other words how similar the cluster elements are to each other. Thus, the SPRSQ value should be small to imply that we are merging two homogeneous groups.

2. RSq (R-squared) measures the extent to which groups or clusters are different from each other (so, when you have just one cluster RSQ value is, intuitively, zero). RSQ value should high, or as close to 1 as possible as it explains the proportion of variance accounted for by the clusters.

3. CCC (Cubic Clustering Criterion), rule of thumb using this statistic is that a value of greater than 2 indicates good clusters, values between 0 and 2 indicate potential clusters [and used with caution]; large negative values could indicate outliers in the data.

Cluster Performance

Before pre-processing

Cluster History										
NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	BSS
20	CL31	CL27	290	0.008	0.768	0.644	42.9	197	73.1	55.26
19	CL33	CL63	140	0.0086	0.76	0.637	41.7	199	54.5	59.165
18	CL22	CL111	35	0.0093	0.75	0.629	40.4	201	18	63.921
17	CL51	CL28	8	0.01	0.74	0.622	39.1	202	7.2	69.158
16	CL32	CL29	105	0.0103	0.73	0.613	38	205	36.7	70.901
15	CL26	CL23	164	0.0113	0.719	0.604	36.7	208	58.9	78.083
14	CL25	CL19	261	0.0126	0.706	0.594	35.4	210	66.1	86.979
13	CL20	CL53	386	0.0171	0.689	0.583	32.8	210	137	117.89

After pre-processing:

Cluster History										
NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	BSS
17	CL33	CL35	253	0.0049	0.878	0.847	13.4	511	97.6	167.68
16	CL20	CL30	283	0.0065	0.872	0.842	12.3	515	89.6	220.98
15	CL23	CL28	191	0.0068	0.865	0.836	11.4	520	72.1	231.97
14	CL19	CL72	46	0.0068	0.858	0.83	10.7	529	24	232
13	CL25	CL36	152	0.0069	0.851	0.824	10.3	543	82.7	234.11
12	CL22	CL27	115	0.0102	0.841	0.816	8.91	548	71	347.56
11	CL38	CL12	179	0.0132	0.828	0.807	6.88	548	73.9	449.94
10	CL15	CL17	444	0.017	0.811	0.798	3.62	543	178	578.11
9	CL13	CL10	596	0.018	0.793	0.786	1.74	547	136	610.65
8	CL16	CL18	315	0.0213	0.771	0.772	-0.2	552	175	723.81
7	CL45	CL34	7	0.0222	0.749	0.755	-1.4	570	33.2	755.07

How to evaluate the cluster models

We already looked at a series of statistics that can help us decide but what else??

- Graph PCAs to check for evidence of clusters
- Box plots
- Testing!! Go out and validate and see if the groupings make sense

There is no one right answer!! It's a model, and will never be 100% correct all the time.

Example: Cluster 7, Urban Competition

Cluster 7 contains **<u>61 transits</u>**.

Main drivers in this cluster:

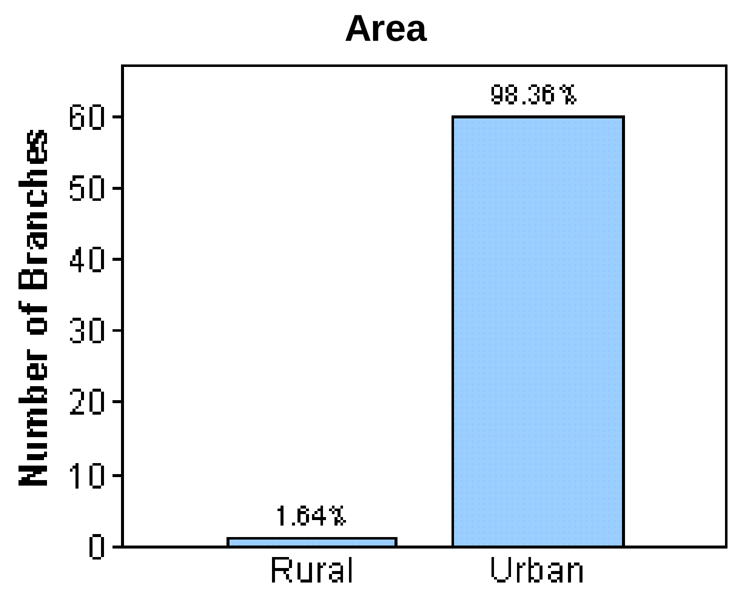
•High Amount of Investors

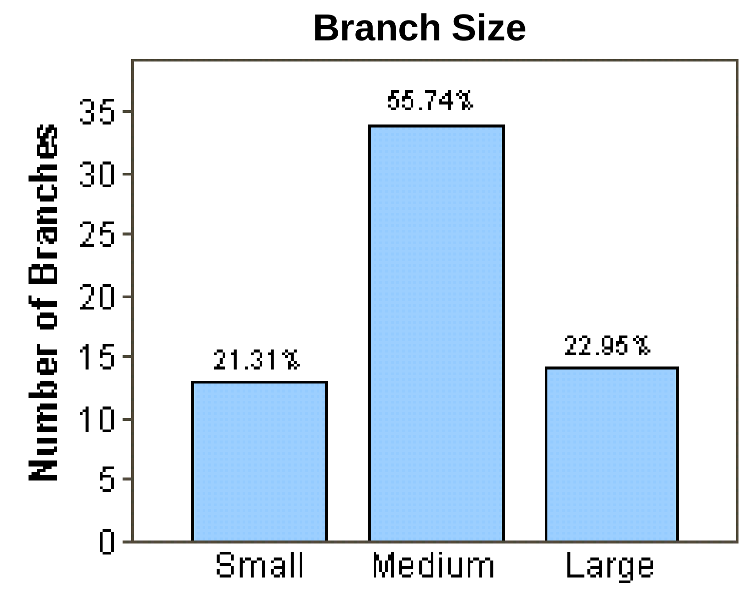
•High Number of Renters

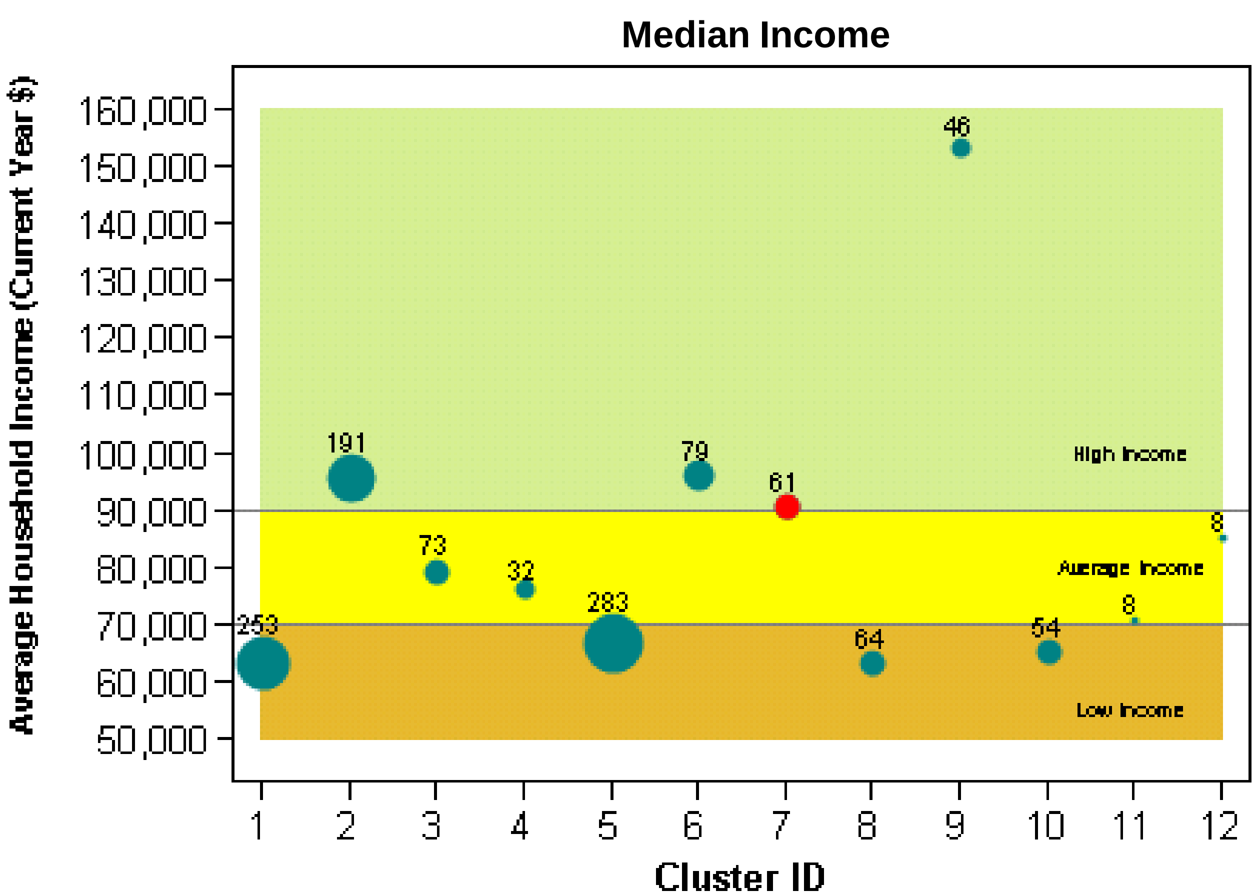
•Lots of Competition nearby

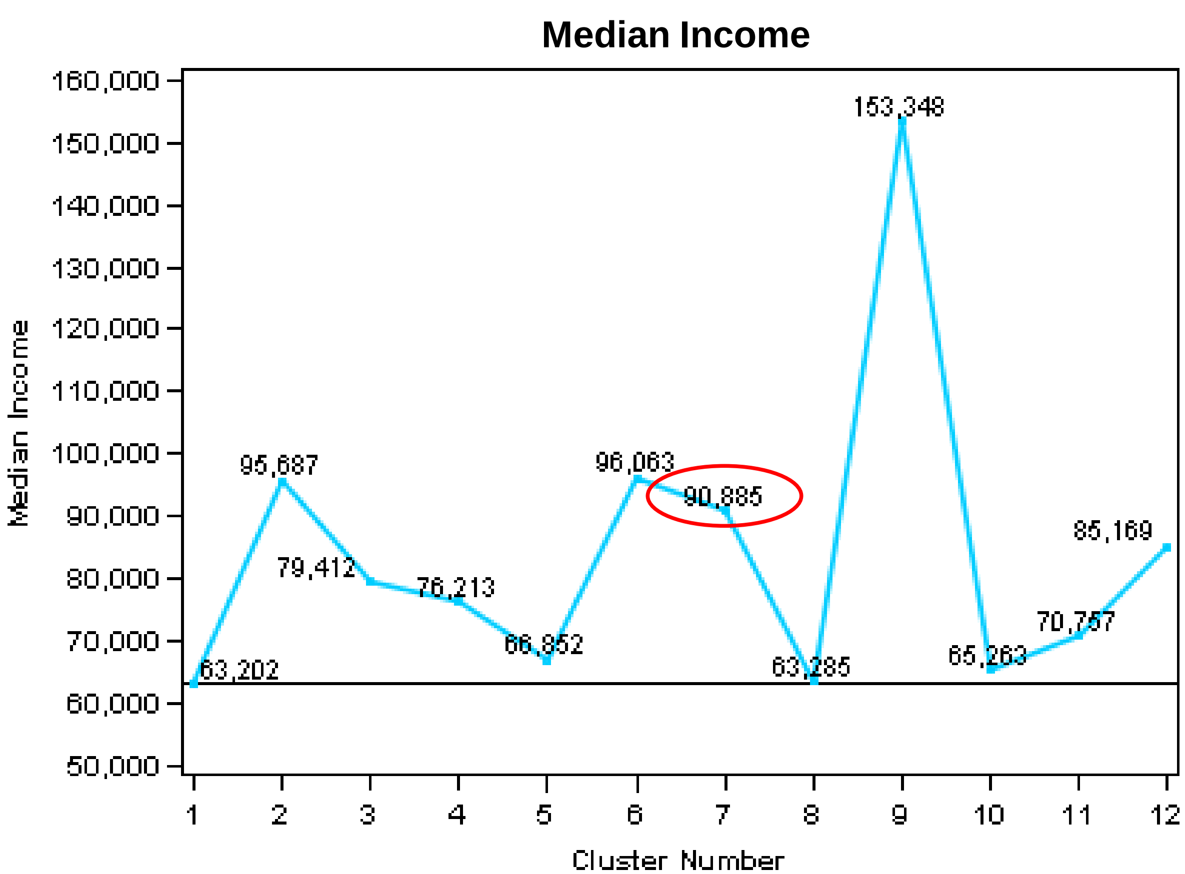
•Higher Crime areas

•High number of anonymous sessions per branch

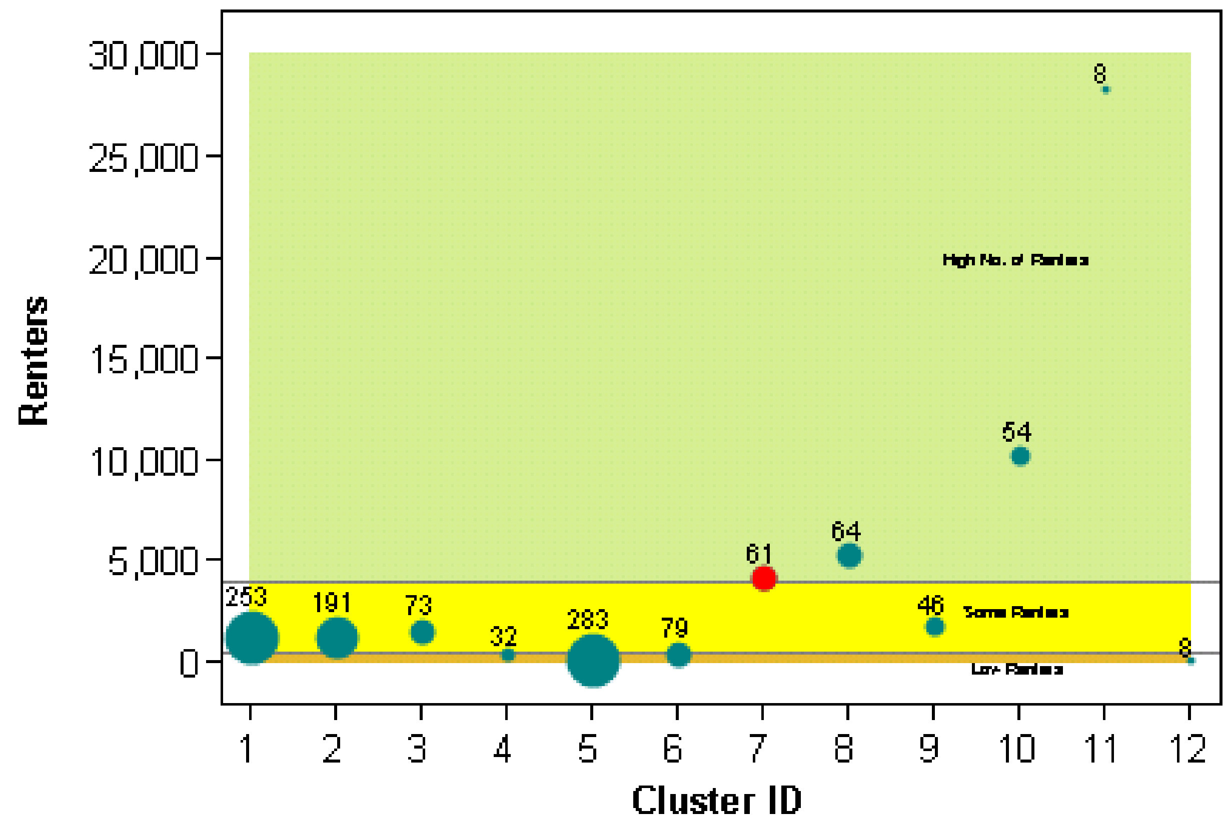


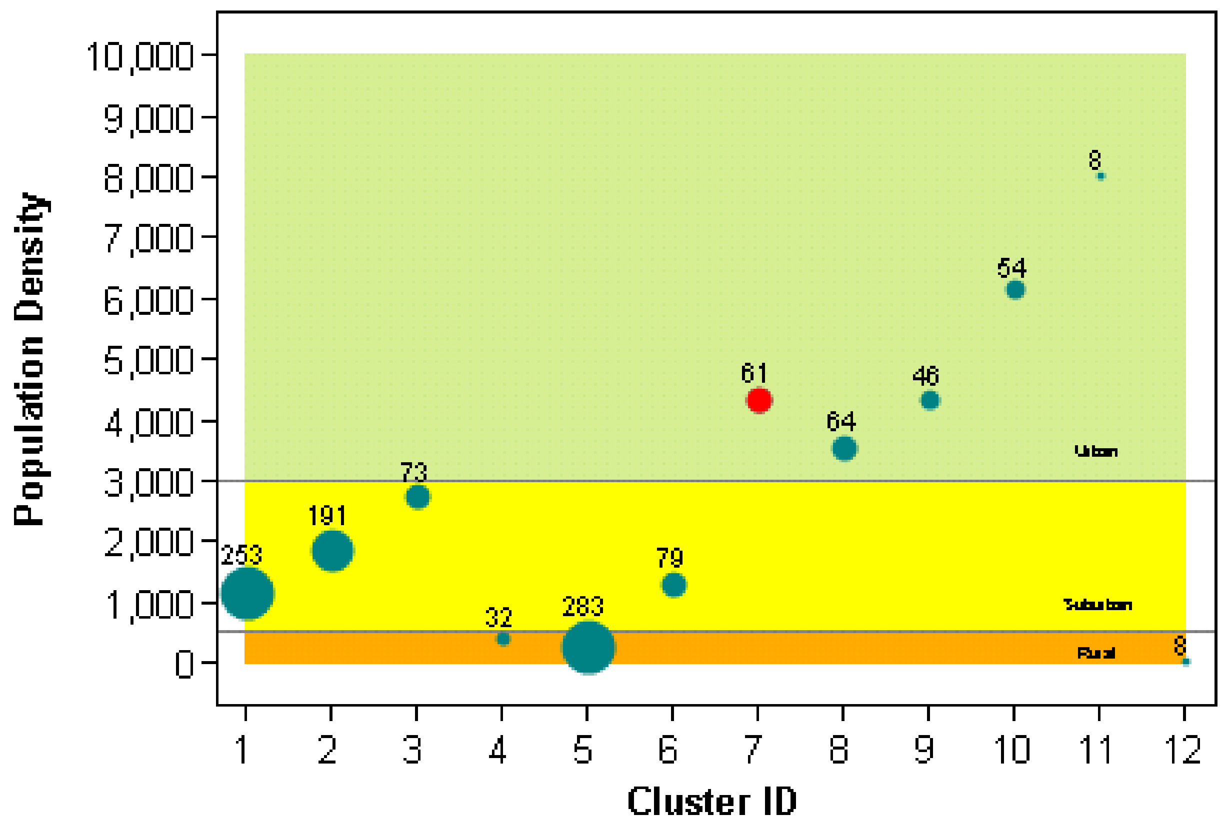


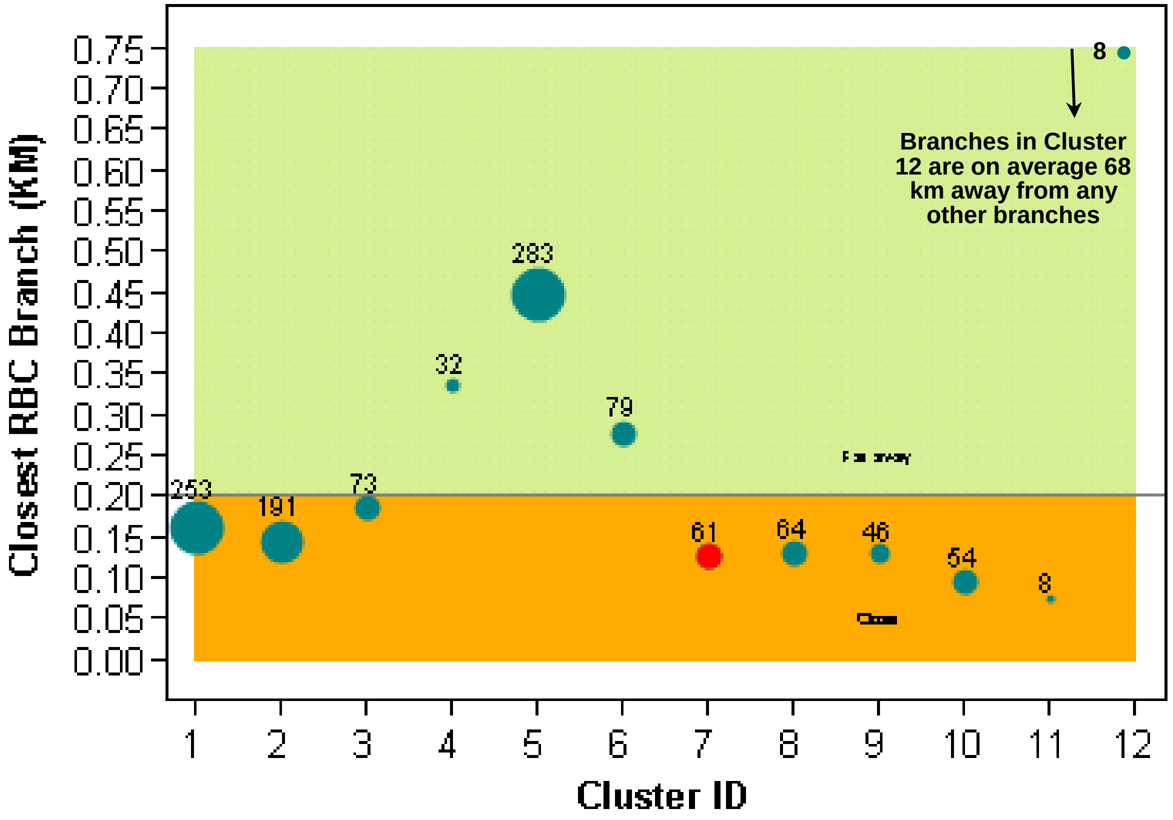


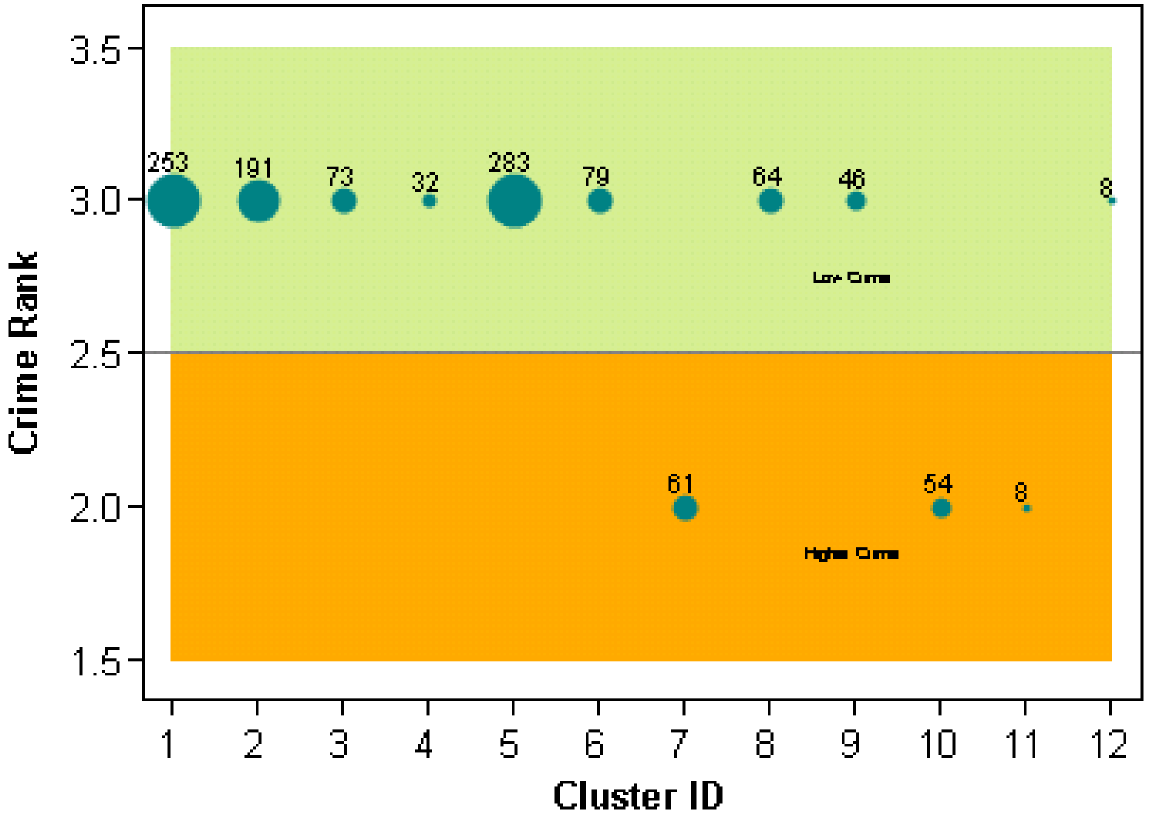


Cluster 7, Urban Competition

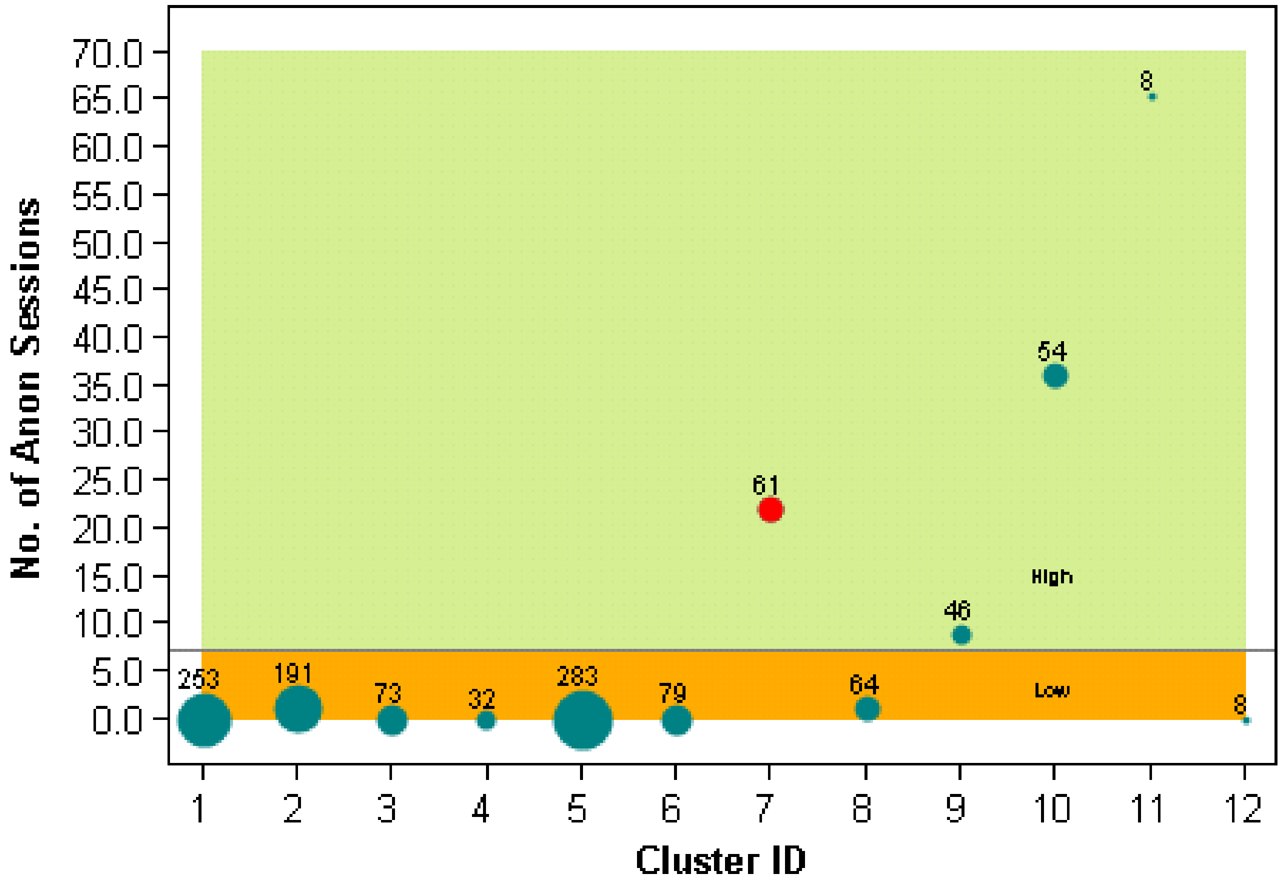


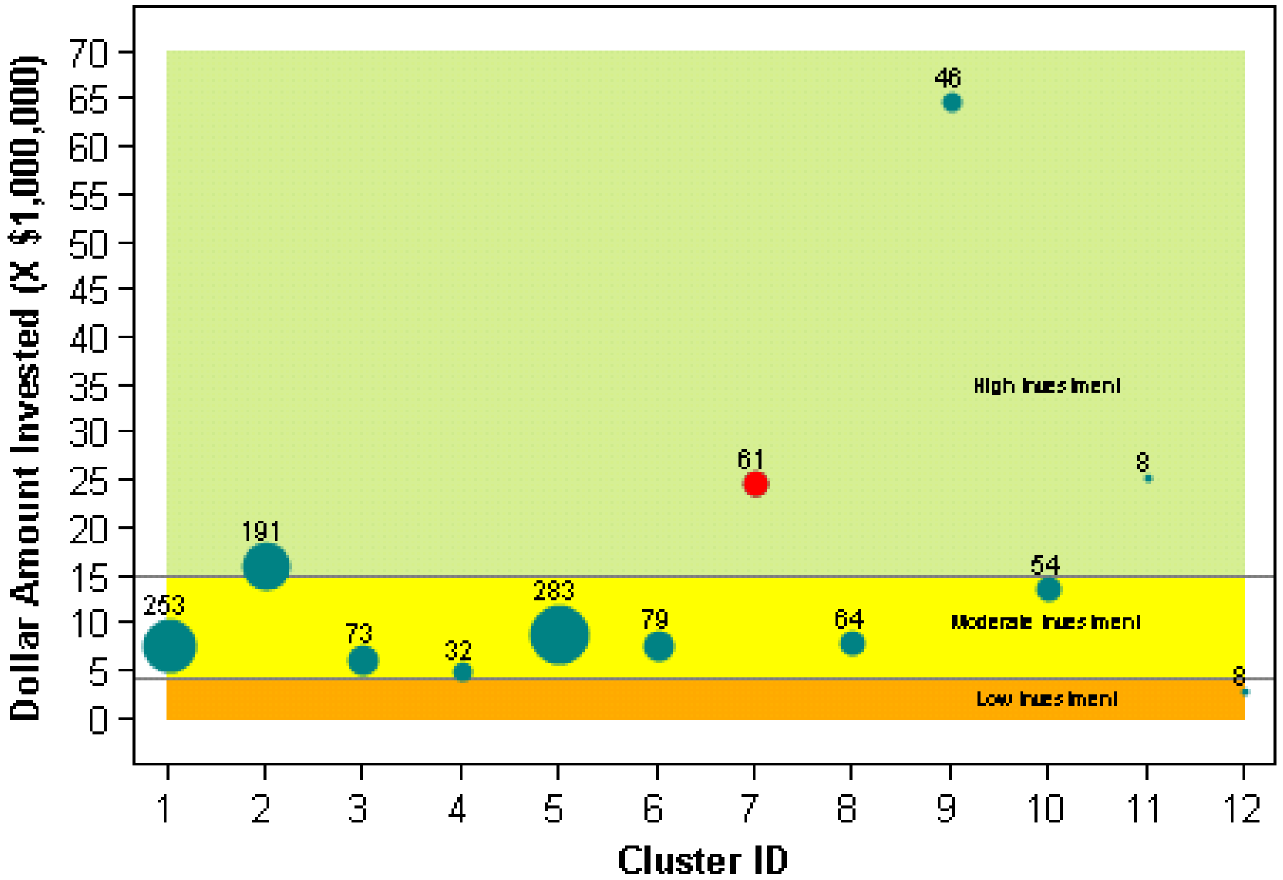




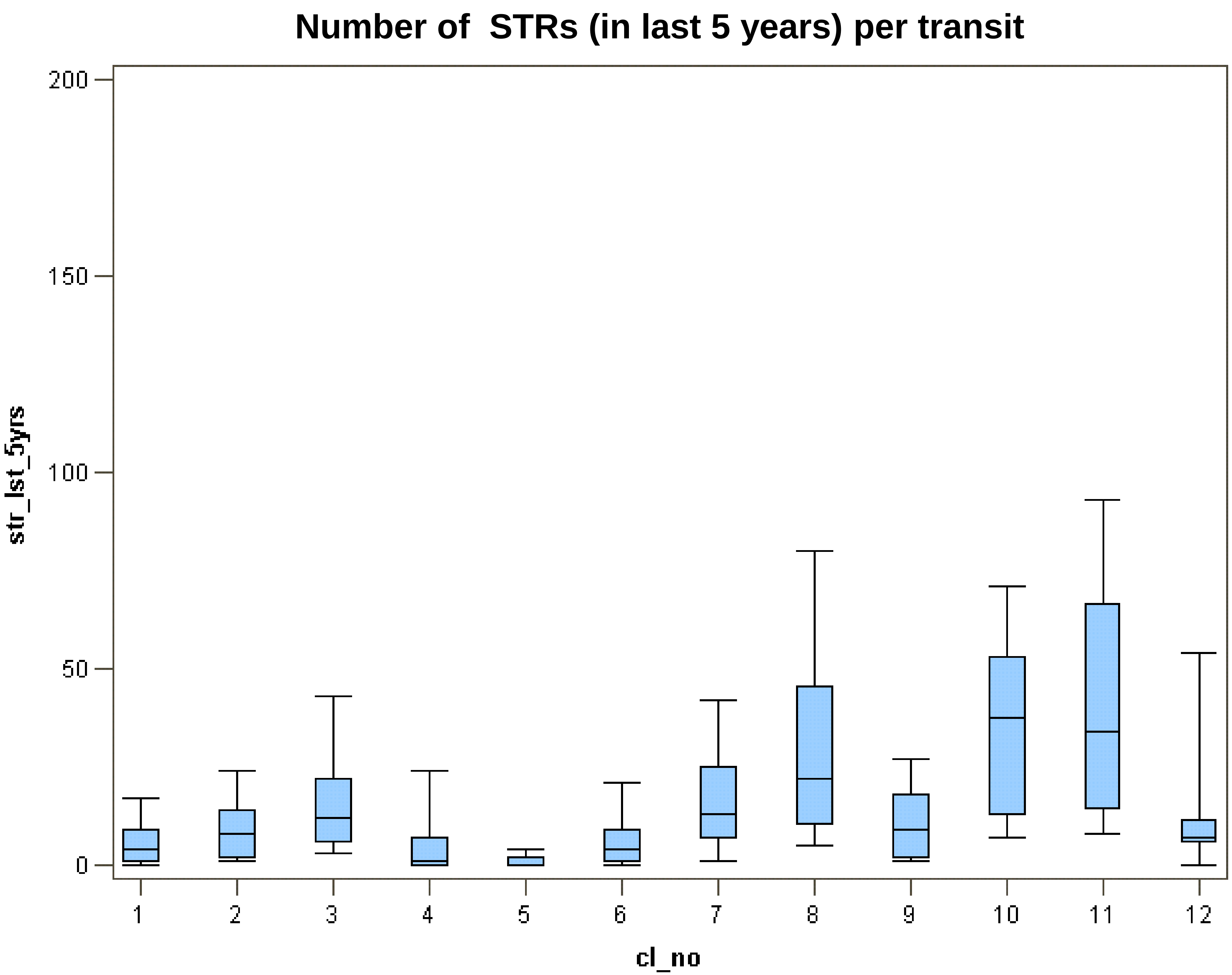


Cluster 7, Urban Competition





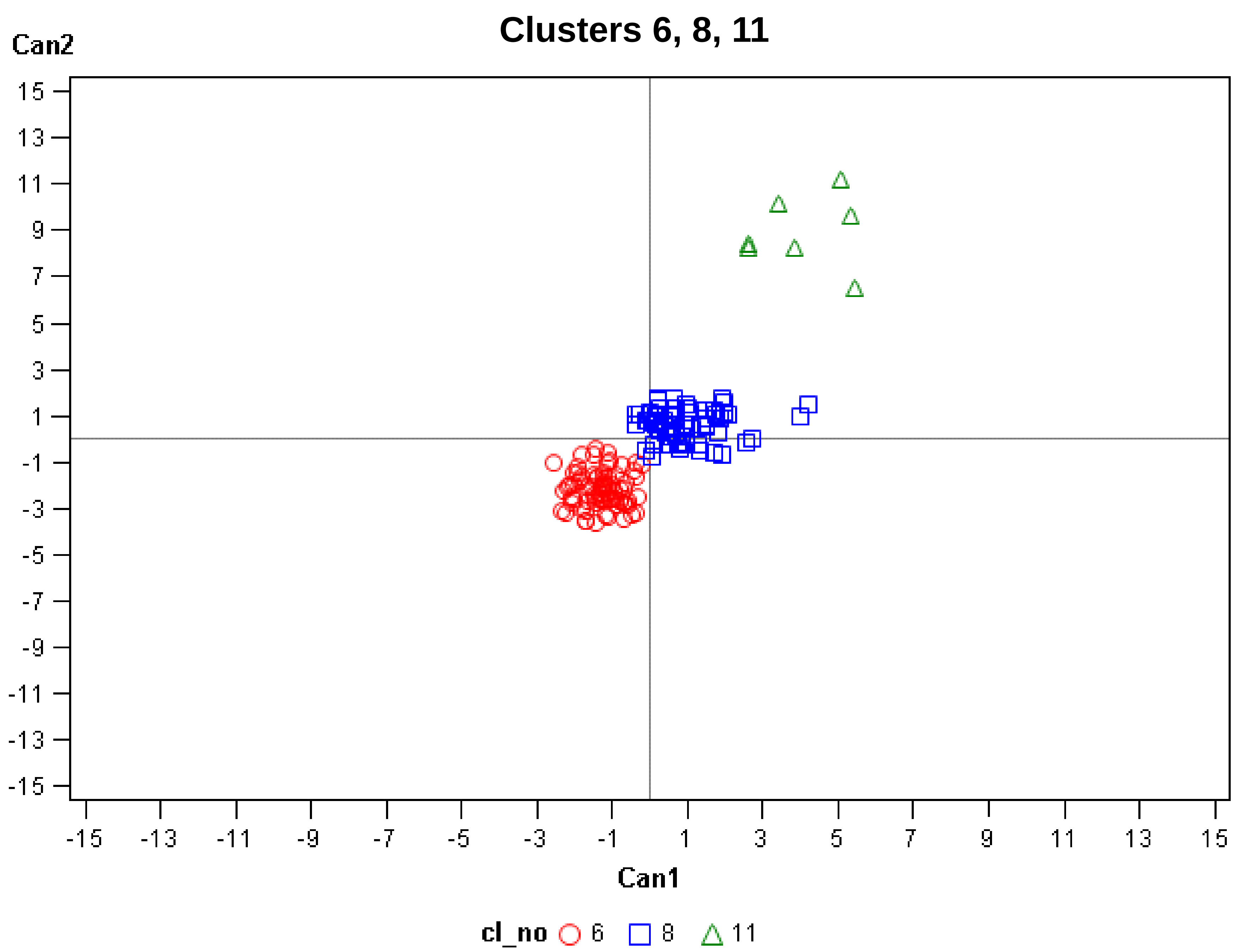
Box Plots



Cluster	Count
1	253
2	191
3	73
4	32
5	283
6	79
7	61
8	64
9	46
10	54
11	8
12	8

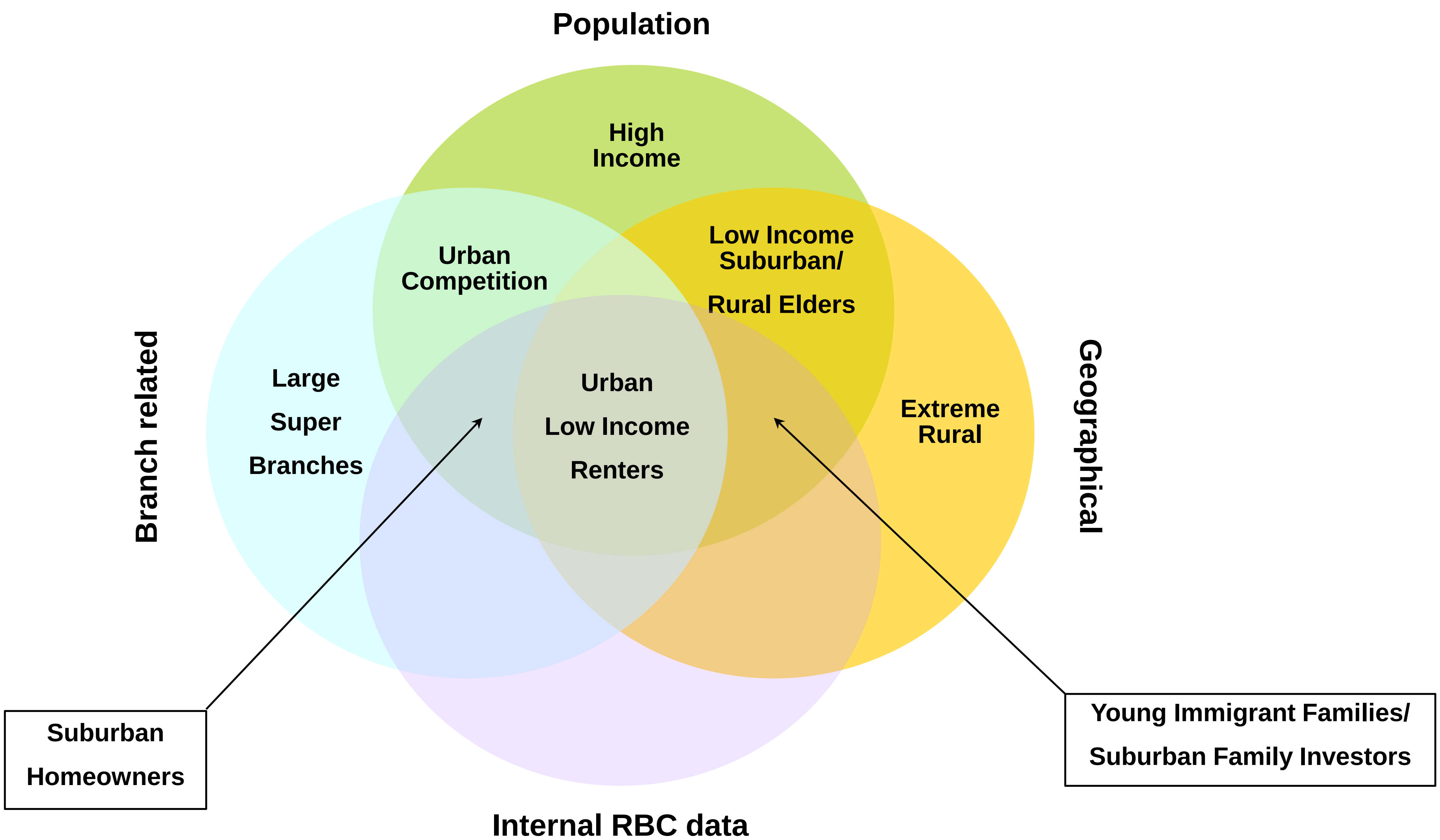
*** Note that Box plots are based on 10th-90th percentiles for each cluster.**

Plot of PCA



Cluster	Count	%
1	253	21.96
2	191	16.58
3	73	6.34
4	32	2.78
5	283	24.57
6	79	6.86
7	61	5.3
8	64	5.56
9	46	3.99
10	54	4.69
11	8	0.69
12	8	0.69

How everything fits together!



Findings

- Key groupings include:
  - Small branches / Rural areas / well established
  - Mid size branches in Urban / young / educated / immigrants
  - Large branches in Lower income / higher crime areas
  - Suburban mid-size branches / families / new branches
- Solution effectively identifies groups based on size and potential to better measure AML risk.

In Summary..

- Solution effectively identifies groups based on size and potential to better measure AML risk.

- EG played a key role

- Apply statistics without deriving them (just know how and when to use them)!

Questions?

Contact: meera.das@rbc.com