Customers Segmentation in the Insurance Company (TIC) Dataset

Ningyuan Zhang

1. The Proposed Model

1.1. Elbow Method to Choose Number of Clusters(K)

One method to validate the number of clusters is the elbow method. The idea of the elbow method is to run K-means clustering on the dataset for a range of values of K (K from 2 to 11 in the following experiments), for each value of K calculate the sum of squared errors (SSE), and plot a line chart of the SSE for each value of K.

$$\text{SSE} = \sum_{i=1}^{k} \sum_{p \in C_i} ||p - m_i||^2$$

If the line chart looks like an arm, then the "elbow" on the arm is the value of K that is the best. It makes sense because adding another cluster doesn't give much better modeling of the data.

The best K is 5 for the first experiment while the best K is 6 for the second one.

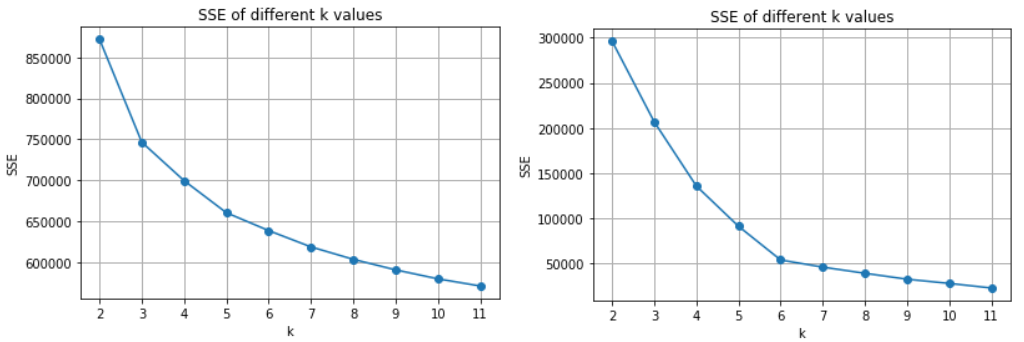


Figure 1 (a) Elbow Method Plot of the First Experiment;(b) Elbow Method Plot of the Second Experiment

1.2. K-Means Algorithm

K-means clustering is one of the simplest and popular unsupervised machine learning algorithms. To process the learning data, the K-means algorithm starts with K randomly defined initial centroids and performs iterative calculations to find the closest centroid for each data point. The process halts when the centroids have stabilized, or the maximum number of iterations has been achieved.

Algorithm 1 K-means clustering algorithm

1: Compute the intensity distribution /*the histogram of intensities*/.
2: Initialize the centroids with k random intensities /*the number of clusters to be found*/.
3: Initialize $\{u_i\}i^k = 1$
4: FOR: Each cluster C_j
5: REPEAT:
6: Cluster the points based on distance of their intensities from the centroid intensities.

$$c^{(i)} := arg\min_j \|x^{(i)} - \mu_i\|^2 \tag{2}$$

7: Compute the new centroid for each of the clusters.

$$\mu_i := \frac{\sum_{i=1}^{m} 1\{c_{(i)=j}\}x^{(i)}}{\sum_{i=1}^{m} 1\{c_{(i)=j}\}} \tag{3}$$

where i iterates over the all intensities, j iterates over all the centroids, and μ_i is the centroid intensity.
8: UNTIL: cluster labels of the image does not change anymore.
9: ENDFOR

Figure 2 K-means Algorithm

1.3. Self-Organized Maps (SOM)

Visualization is difficult in high dimensional clustering. In order to reduce the dimensions of the data, I use the Self-Organized Maps (SOM).

SOM is a type of artificial neural network which is trained using unsupervised learning. It consists of two layers of units: a one-dimensional input layer and a two-dimensional competitive layer, organized as a 2D grid of units. This layer can also be called output layer or computational layer. Each node on the competitive layer is a neuron (for example, a 7*7 layer has 7*7 neurons), and each neuron is associated with a "weight" vector. The “weight” vector has the same dimension as each input vector.

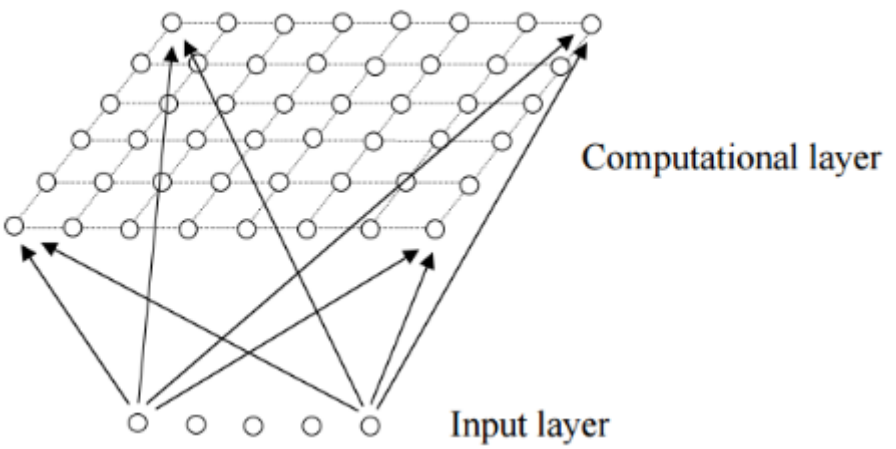


Figure 3 SOM

The learning process is as follows. The weights of the neurons are initialized to small random values, when an input vector is fed to the network, its similarity to all weight vectors is computed. The neuron whose weight vector is most similar to the input is the winner and the winner is considered to represent the input vector and other similar ones. The weights of the winner neuron and its neighbor neurons would be adjusted towards the input vector. This process is repeated for each input vector for a large number of times.

After training, every input vector could be associated to a single optimal neuron and similar input patterns tend to produce a response in units that are close to each other in the competitive layer.

2. Experiment 1: K-means Clustering

2.1. Execution

I had applied the K-means algorithm for two times. The first time I directly used the most 21 informative attributes which are mentioned in the paper, while I picked the most 21 informative attributes baed on their Shannon entropy for the experiment.

The centroids tables are as follows:

	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
1-MOSTYPE	34.165	8.466	23.403	7.081	35.704
5-MOSHOOFD	7.933	2.440	5.192	2.022	8.413
16-MOPLHOOG	0.539	1.792	1.449	2.931	0.976
19-MBERHOOG	0.911	2.014	1.451	3.551	1.607
25-MSKA	0.639	1.648	1.226	3.158	1.425
30-MHHUUR	7.153	6.323	6.955	1.163	2.249
31-MHKOOP	1.857	2.689	2.055	7.846	6.756
32-MAUT1	5.693	6.318	5.284	6.776	6.013
34-MAUT0	2.547	1.940	3.277	0.988	1.624
35-MZFONDS	7.483	6.129	6.934	4.634	6.320
37-MINKM30	3.942	2.689	3.972	1.106	1.946
39-MINK4575	1.586	2.682	1.974	4.114	2.947
42-MINKGEM	2.921	3.696	3.078	4.935	3.955
43-MKOOPKLA	3.067	5.540	2.112	6.777	3.766
44-PWAPART	0.711	0.682	0.951	0.855	0.710
47-PPERSAUT	3.083	2.922	2.651	3.118	2.956
59-PBRAND	1.546	1.553	1.288	2.210	2.116
61-PPLEZIER	0.012	0.019	0.015	0.027	0.020
65-AWAPART	0.382	0.355	0.495	0.439	0.371
68-APERSAUT	0.577	0.558	0.492	0.592	0.565
82-APLEZIER	0.004	0.004	0.005	0.010	0.006

	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4
1-MOSTYPE	7.193	34.292	8.208	23.422	35.668
5-MOSHOOFD	2.050	7.960	2.375	5.207	8.407
10-MRELGE	7.188	5.895	6.146	4.052	6.720
16-MOPLHOOG	2.923	0.560	1.819	1.446	0.970
18-MOPLLAAG	2.492	6.365	3.523	4.415	5.214
25-MSKA	3.127	0.687	1.722	1.233	1.403
30-MHHUUR	1.112	7.118	6.312	6.922	2.200
31-MHKOOP	7.899	1.893	2.699	2.088	6.805
32-MAUT1	6.802	5.657	6.300	5.279	6.046
34-MAUT0	0.970	2.564	1.936	3.279	1.593
37-MINKM30	1.068	3.929	2.718	3.983	1.912
39-MINK4575	4.140	1.599	2.666	1.976	2.961
42-MINKGEM	4.941	2.922	3.717	3.067	3.976
43-MKOOPKLA	6.799	3.107	5.560	2.119	3.748
44-PWAPART	0.851	0.710	0.691	0.946	0.712
47-PPERSAUT	3.136	3.047	2.917	2.641	2.980
59-PBRAND	2.185	1.552	1.609	1.291	2.122
61-PPLEZIER	0.027	0.012	0.019	0.014	0.020
65-AWAPART	0.437	0.381	0.360	0.492	0.372
68-APERSAUT	0.596	0.571	0.555	0.490	0.569
80-ABRAND	0.620	0.523	0.520	0.550	0.600

Figure 4 (a) Centroids Table of experiment a;(b) Centroids Table of experiment b

2.2. Evaluation

Here I used the Calinski Harabasz score to evaluate the clustering model. It's a popular evaluation measure.

$$\mathrm{CH(k)} = \frac{B(k)}{W(k)} \frac{n-k}{k-1}$$

where n is the number of data points, k is the number of clusters,W(k) = within cluster variation and B(k) = between cluster variation. The model having larger Calinski Harabasz score is better.

In this model, the Calinski Harabasz score is 5109.

2.3. Analysis

I analyzed the most 10 interesting attributes mentioned in the paper. For each attribute, I analyzed the distribution of it and the relation between each cluster and each attribute. All the scatter plots are shown in the appendix section.

Cluster	86-CARAVAN	47-PPERSAUT	59-PBRAND	68-APERSAUT	5-MOSHOOFD	43-MKOOPKLA	61-PPLEZIER	42-MINKGEM	82-APLEZIER	1-MOSTYPE	44-PWAPART
0.0	0.0	3.008389	1.526007	0.561242	7.931208	3.070470	0.012584	2.915268	0.004195	34.166107	0.694631
	1.0	4.740741	1.981481	0.925926	7.962963	2.981481	0.000000	3.055556	0.000000	34.148148	1.074074
1.0	0.0	2.817647	1.500000	0.536765	2.457353	5.526471	0.020588	3.688235	0.004412	8.514706	0.664706
	1.0	4.658537	2.439024	0.902439	2.146341	5.756098	0.000000	3.829268	0.000000	7.658537	0.975610
2.0	0.0	2.580000	1.271250	0.480000	5.191250	2.115000	0.005000	3.068750	0.001250	23.405000	0.951250
	1.0	5.631579	2.000000	1.000000	5.210526	2.000000	0.421053	3.473684	0.157895	23.315789	0.947368
3.0	0.0	2.909341	2.122711	0.552198	2.020147	6.763736	0.023810	4.915751	0.008242	7.054945	0.798535
	1.0	4.848485	2.931818	0.924242	2.037879	6.886364	0.053030	5.098485	0.022727	7.295455	1.318182
4.0	0.0	2.870760	2.097076	0.547368	8.417544	3.759064	0.005848	3.945029	0.001754	35.701170	0.690643
	1.0	4.392157	2.441176	0.862745	8.333333	3.882353	0.254902	4.117647	0.078431	35.745098	1.039216

Figure 5 Attribute Analysis Table Grouping by 'Cluster' and '86-CARAVAN'

A. Cluster 0: This cluster represents customers whose Feature 1 value (1 MOSTYPE: Customer Subtype) is from 29 to 41, consisting of porchless seniors, village families and large traditional families. The average purchasing power class and average income are pretty low, which could explain why most of them do not own a caravan policy. They do not spend much on boat policy since low-income people living in villages and elderly tend to in no need for boats. However, the average car policy contribution is quite high no matter whether or not they own a caravan. It's because cars are indispensable for village families and large family farms.
B. Cluster 1: This cluster represents customers whose Feature 1 value is from 1 to 15. Most of them have comparatively high income and even higher purchasing power, so they are more likely to have a caravan policy than people in Cluster 0. People who have a caravan policy tend to spend more on car policies, which makes sense since the caravan needs to be pulled by the car.
C. Cluster 2: This cluster represents customers whose Feature 1 value is from 16 to 30. Customers in Cluster 2 have high contribution in boat policies. Now that they have already had boats and the average purchasing power is limited, few of them would choose to buy a caravan.
D. Cluster 3: Subtypes of customers in this cluster are pretty similar to the ones in Cluster 1, but they have even higher income and purchasing power. They spend more on fire policies, boat policies and private third-party insurance than any other cluster. Unsurprisingly, they are most likely to buy caravan policies.
E. Cluster 4: Subtypes of customers in this cluster are pretty similar to the ones in Cluster 0, but they have higher income and purchasing power. It also explains why they are more likely to buy caravan policy and boat policy.

3. Experiment 2: Self Organized Map (SOM) and K-means

3.1. Execution

I still selected the most 10 interesting attributes mentioned in the paper and trained them in a SOM model. After training, the 10-dimension input data could be mapped onto a 2-dimention grid. Next, I ran K-means algorithm again and visualized it.

The centroids table is as follows:

	Cluster_0	Cluster_1	Cluster_2	Cluster_3	Cluster_4	Cluster_5
som1	23.022	3.776	11.663	21.964	1.937	13.159
som2	11.544	0.447	12.393	0.834	12.711	1.611

Figure 6 Centroids Table

The following figure is a scatter plot of the clustering result in 2D feature space.

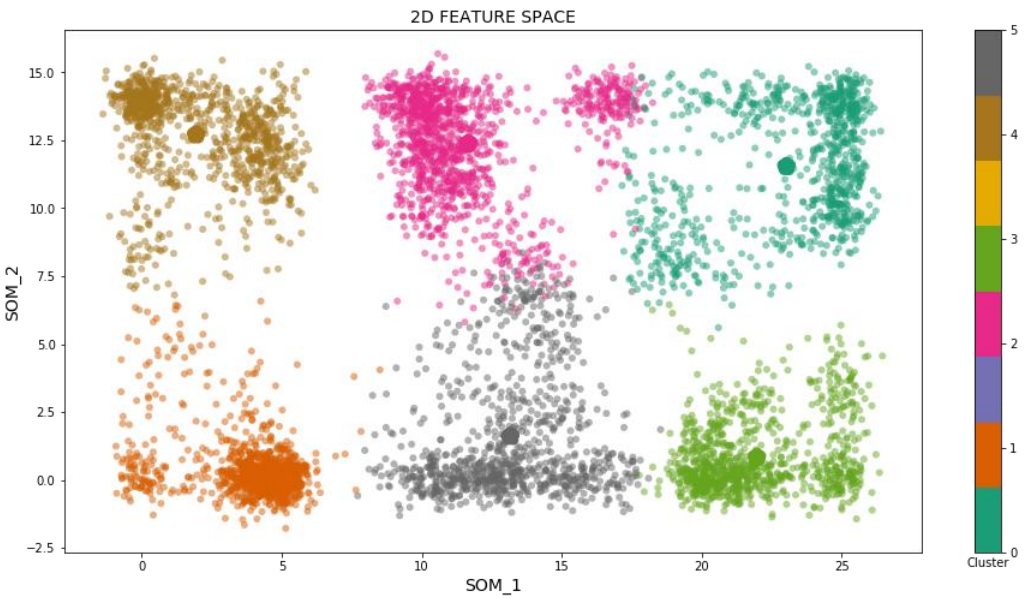


Figure 7 2D Feature Space after Clustring

The following figure shows the relation between every clusters. The darker area represent similarity (low average distance of the weights) while the brighter area represents dissimilarity. It's clear that the comparatively white bands separate the cluster.

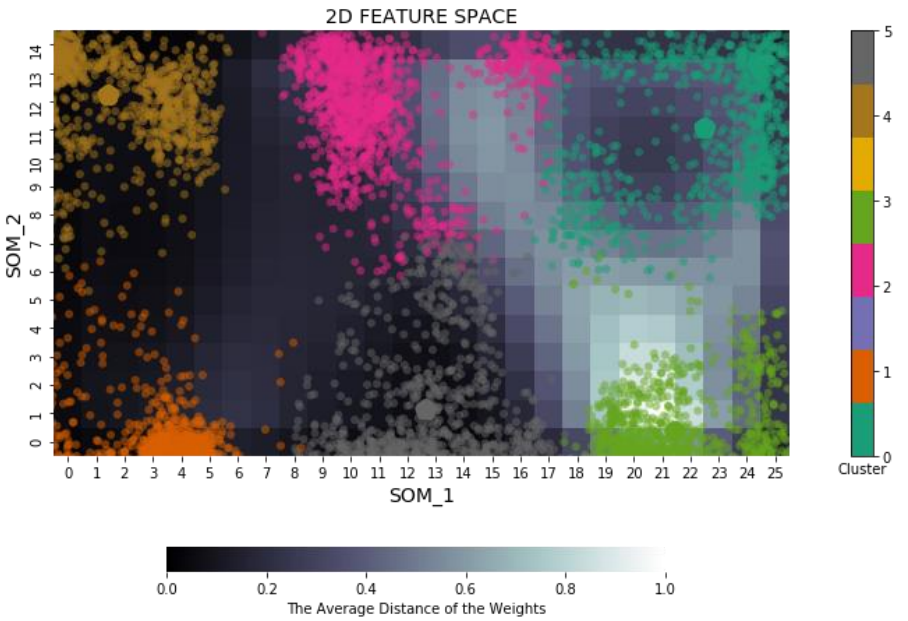


Figure 8 Average Distance of Weights

3.2. Evaluation

The Calinski Harabasz score of the model is 14285, which is much better than the first model.

3.3. Analysis

I still analyze the 10 attributes to study and compare each cluster.

A. Cluster 0: Most of the customers in this cluster are from high-income families. They have the strongest purchasing power and live an affluent life. They are willing to pay more for anything, including the caravan policy.
B. Cluster 1: Cluster 1 consists of traditional families, retired and religions and most of them do not spend money on car policies. It explains why they do not choose to own caravan.
C. Cluster 2: Customers in this clusters have medium income and purchasing power and have cars. But their likelihood of owning a caravan policy is low. I noticed that their cost on boat policies is close to people in Cluster 0. The only one reasonable is that they love boats and spend the constrained budgets on this hobby.
D. Cluster 3: Customers in Cluster 3 is almost as rich as the ones in Cluster 0, but they don't have cars, which explains why they do not own caravan policies neither. Instead of camping out, they could choose to have a sea trip, so they spend more on boats.
E. Cluster 4: This cluster is made up of low-income customers, most of them could not afford car policies and many other policies, let alone the caravan.
F. Cluster 5: Cluster 5 is a union of all types of customers who spend much on car policies and they also tend to own a caravan although they are not rich. It shows the strong correlation between car policies and caravan policies again.

4. Conclusion

In the first experiment, the most 10 informative attributes I picked by Shannon entropy are slightly different from those mentioned in the paper, but the centroids are very similar. In my opinion, it is because the we used different methods to measure information gain.

It's clear that SOM successfully reduces the high dimension of data to realize visualization and improves the performance of clustering. In the second experiment, I directly used the same features in the first experiment because keeping attributes same is better for the comparison model performance. I also tried to calculate the correlation between every two attributes, I found that some attributes are highly correlated(for example, attribute 1 and attribute 5), and I filtered out them whose correlation is higher than 0.8, and trained SOM model using the left 63 features, but the clustering result is not good. So, I still kept using the 10 attributes I mentioned before. As a future plan, extra efforts should be spent on finding features suitable for training and clustering since some attributes in the 10 attributes mentioned in the paper are associated and some important attributes, like the average size of household and marital status, should be taken into consideration. I also tried to write the SOM model from scratch, but the time complexity is too high, hence I used the MiniSom library directly.

Appendix

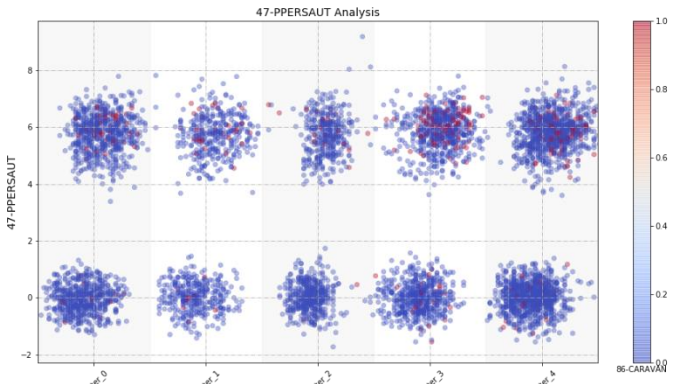


Figure 9 Feature 47

Customers owning 4-7 car policy contributions, especially in Cluster 3 are more likely to buy a caravan policy.

Customers with no car policy contribution have less chance to own a caravan policy.

Customers with 4 contribution in fire policies, especially in Cluster 3 and Cluster 4 are more likely to buy a caravan policy.

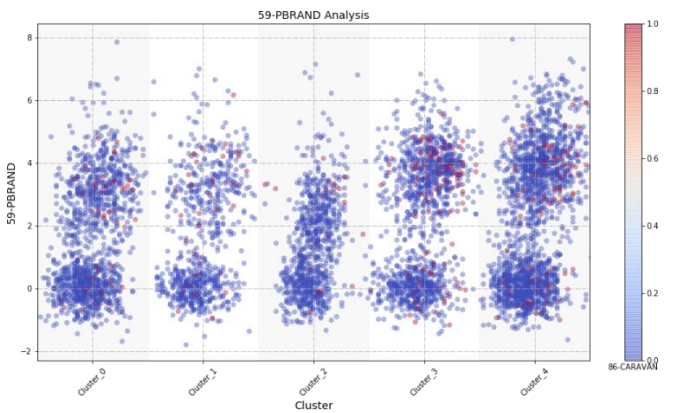


Figure 10 Feature 59

Customers with 3-4 contribution in fire policies are more likely to own a caravan policy.

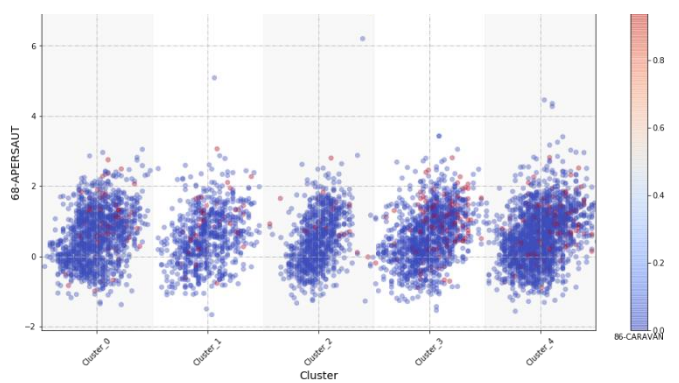
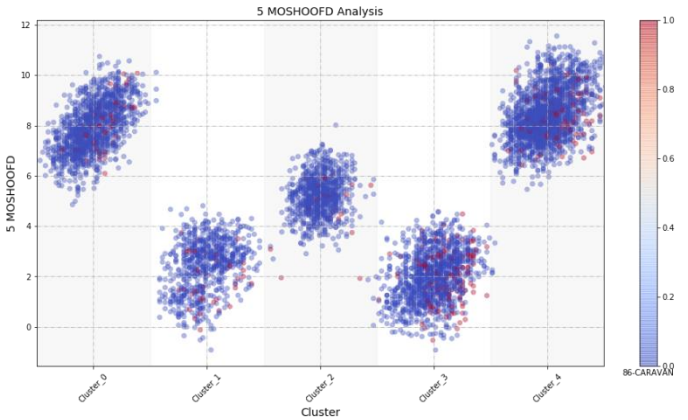


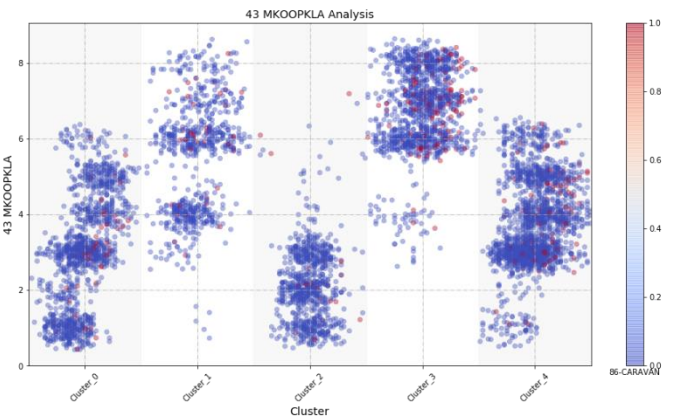
Figure 11 Feature 68

Most customers have less than 2 car policies.



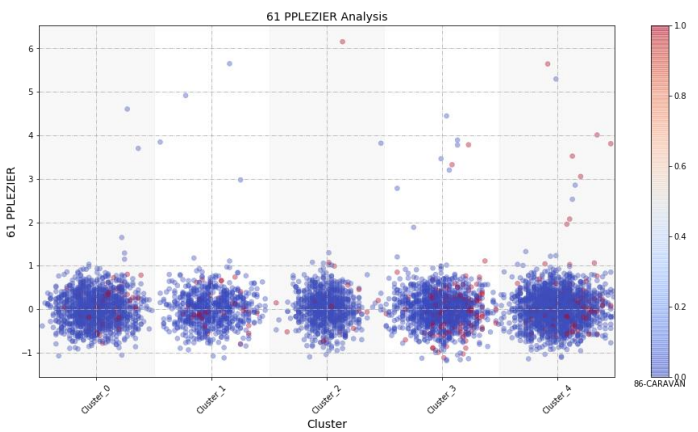
Customers in Cluster 0 are from type 6 -10, customers in Cluster 1 are from type 1- 4, customers in Cluster 2 are from type 4 – 6, customers in Cluster 3 are from type 1 – 4, customers in Cluster 4 are from type 6- 10.

Figure 12 Feature 5



The purchasing power order is Cluster 3, Cluster 1, Cluster 4, Cluster 0 and Cluster 2 in a decreasing order.

Figure 13 Feature 43



Most customers have 0 boat policy contribution.

Figure 14 Feature 61

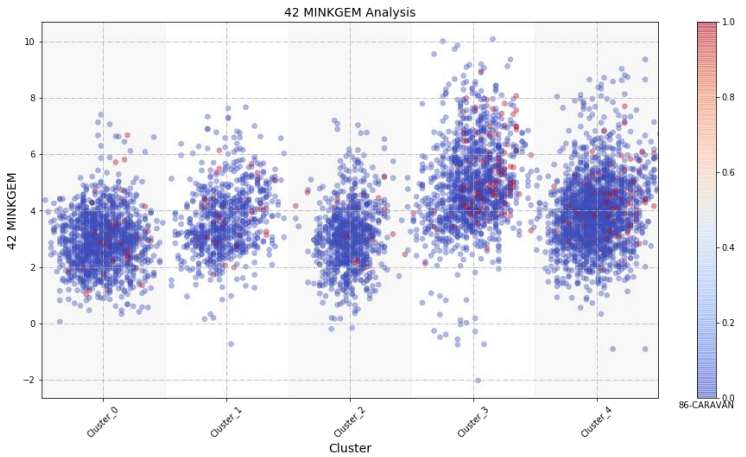


Figure 15 Feature 42

Cluster 3 and Cluster 4 have the comparatively high income and they are more likely to own a caravan policy.

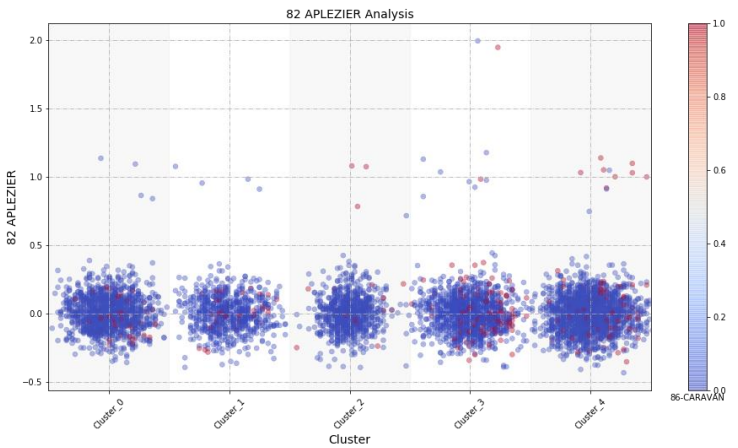


Figure 16 Feature 82

Most customers have no boat policy. Customers owning boat policies do not buy caravan policies in Cluster 0 and Cluster 1.

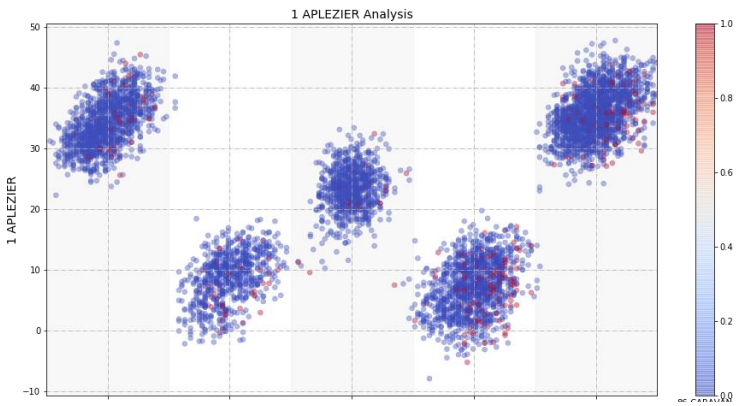


Figure 17 Feature 1

Customers in Cluster 0 are from subtype 26-41, customers in Cluster 1 are from subtype 1-16, customers in Cluster 2 are from subtype 16-30, customers in Cluster 3 are from subtype 1-16, customers in Cluster 4 are from subtype 26-41

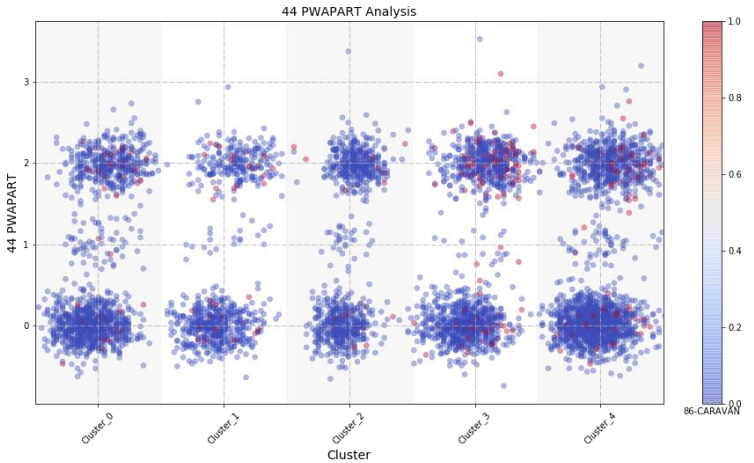


Figure 18 Feature 44

Customers with 2 contributions are more likely to own a caravan policy.