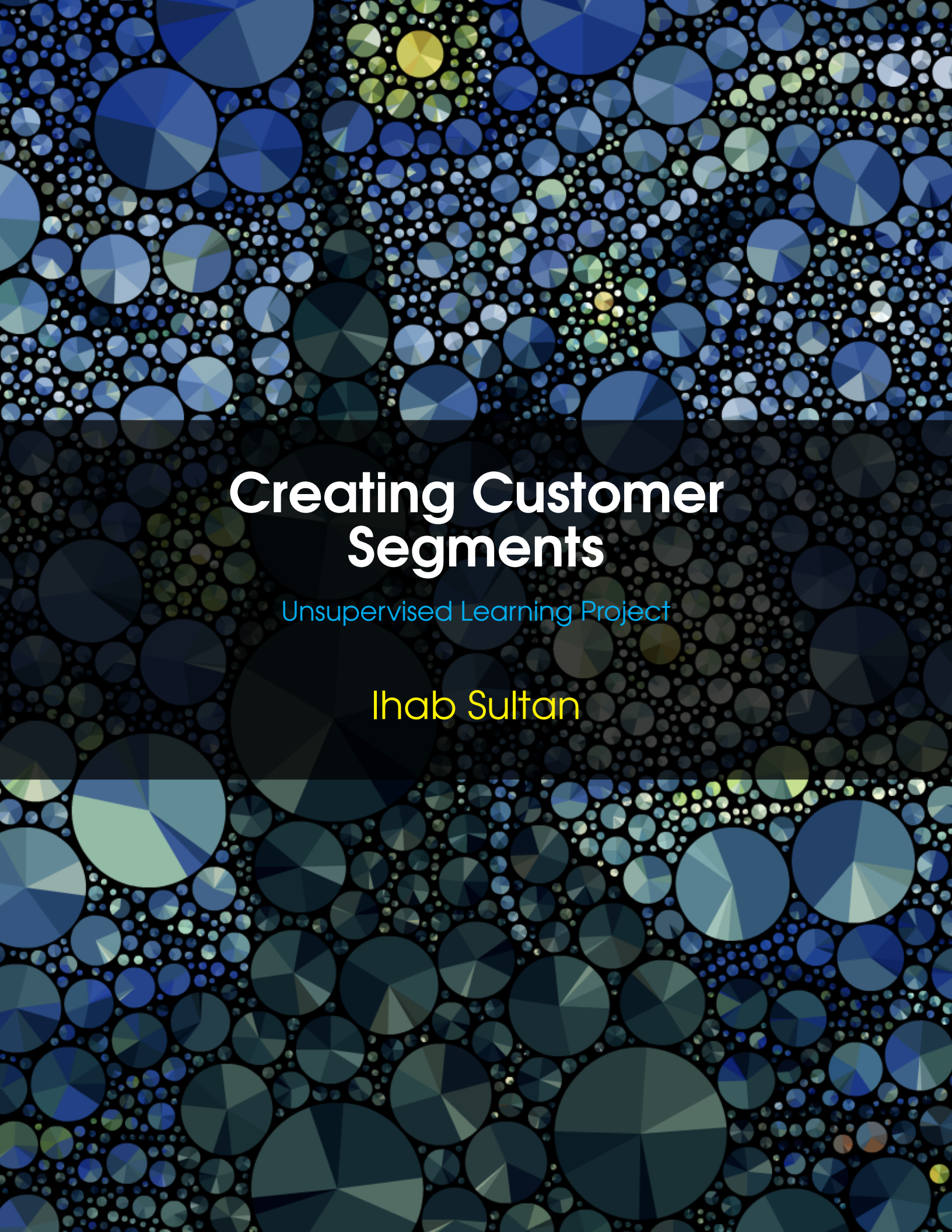


Creating Customer Segments

Unsupervised Learning Project

Ihab Sultan

Contents

1 Introduction 2
1.1 Project Overview 2
1.2 Software and Libraries 2

2 Customers Segmentation Data Set 3
2.1 Data Set Source 3
2.2 Available Attributes 3
2.3 Descriptive statistics 3

3 Project Report 4
3.1 Component analysis 4
3.2 Clustering 7
3.3 Conclusions 10

4 Further Explorations with R 11
4.1 Principal Component Analysis (PCA) 11
4.2 K-means 13

1. Introduction

1.1 Project Overview

The goal of this project is to use unsupervised techniques to see what sort of patterns exist among existing customers, and what exactly makes them different..

1.2 Software and Libraries

The following SW was used in the first part of the project:

- Python 2.7
- NumPy
- pandas
- matplotlib
- scikit-learn
- iPython Notebook

In the last part of this project, R was used as an EDA tool:

- R 3.2.3
- ggbiplot

The Legrand Orange Book by Mathias Legrand used under CC BY-NC-SA 3.0
The Starry Night Pie Packed by Mario Klingemann used under CC BY-ND 2.0
130824_gbg 008 by annakajsa used under CC BY-ND 2.0

2. Customers Segmentation Data Set

2.1 Data Set Source

The dataset refers to clients of a wholesale distributor. It includes the annual spending in monetary units (m.u.) on diverse product categories.

It is part of a larger database published with the following paper:

Abreu, N. (2011). Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional. Mestrado em Marketing, ISCTE-IUL, Lisbon.

2.2 Available Attributes

1. **Fresh**: annual spending (m.u.) on fresh products (Continuous)
2. **Milk**: annual spending (m.u.) on milk products (Continuous)
3. **Grocery**: annual spending (m.u.)on grocery products (Continuous)
4. **Frozen**: annual spending (m.u.)on frozen products (Continuous)
5. **Detergents_Paper**: annual spending (m.u.) on detergents and paper products (Continuous)
6. **Delicatessen**: annual spending (m.u.)on and delicatessen products (Continuous)

2.3 Descriptive statistics

Attribute: *(Minimum, Maximum, Mean, Std. Deviation)*

1. **Fresh**: (3, 112151, 12000.30, 12647.329)
2. **Milk**: (55, 73498, 5796.27, 7380.377)
3. **Grocery**: (3, 92780, 7951.28, 9503.163)
4. **Frozen**: (25, 60869, 3071.93, 4854.673)
5. **Detergents_Paper**: (3, 40827, 2881.49, 4767.854)
6. **Delicatessen**: (3, 47943, 1524.87, 2820.106)

3. Project Report

3.1 Component analysis

Reflection on PCA/ICA

- **What are likely candidates for early PCA dimensions?**
- **What might ICA dimensions look like?**

PCA will select orthogonal dimensions with highest variance. In terms of the data set at hand, this means the directions where customers had the maximum variation in annual spendings. The result will be a set of 6 orthogonal 6-D vectors with decreasing variance. For example, if we get the first principal component along [1/sqrt(2),1/sqrt(2),0,0,0,0] that would mean the sum of Fresh and Milk sales has the highest variance among all possible combinations of the 6 quantities.

In addition, some correlation relation exists among attributes that contribute to the same principal component. In the made-up example above, we can deduce that Fresh and Milk are usually bought together and hence they appear with similar lengths in the first principal component.

Going over the descriptive statistics of our data, we find that attributes with highest standard deviations are as follows:

- Fresh, SD = 12647
- Grocery, SD = 9503
- Milk, SD = 7380

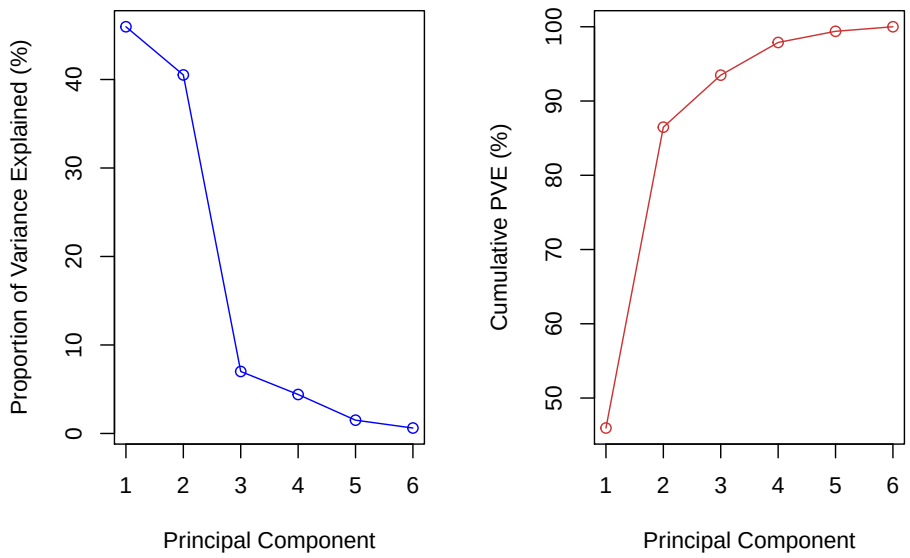
Therefore, we expect the first few principle components to lie along these dimensions (or a combination thereof).

By performing actual PCA, we find that first component is indeed close to the fresh direction, while second component is along the Grocery, Milk, and Detergents_Paper.

ICA well report a set of 6 6-D vectors along the directions of maximum statistical independence. The components are independent meaning that knowing the projection on any of the components does not give any information about the projection on other vectors, in addition, any customer sample can be explained by the sum of projections on these independent components.

The above means that the independent components reveal the underlaying latent sources driving all samples, and give an idea about the actual structure of the data. In light of the problem at hand, the independent components would reveal the different customers segments that we need to consider.

What proportion of variance is explained by each PCA dimension?



Variance along the first two principal components is high, it drops significantly along the third component. I would choose the first three components which would explain more than 93% of the variance in the original data.

PCA dimensions

- **What are the first few components? What might they represent?**
- **How can you use this information?**

First component is mainly along the 'Fresh' dimension, so customers do not pay similar amounts on fresh products(another way to say that variance along the Fresh dimension is the largest). Could be that some customers sell fresh products while some sell frozen products. We notice that statistics of Fresh products has the larges std. deviation (= 12647) among all products.

Second component is along 'Milk', 'Grocery', 'Detergents_Paper' dimensions. seems that a number of customers sell the three items together and some sell none of them.

Third component is mainly along 'Milk' and 'Frozen' dimensions, could be a difference between customers with fridges and others, but given that variance along this vector is very small, most customers agree on the total of these values.

This is a rough explanation as other factors exist along above principal components.

In general, the dimensions would tell us which vectors represent the differences between the customers, and this info is very useful in differentiating between different types of

customers. Basically, a direction where all customers seem to agree won't help much in differentiating between different customers types.

ICA

- **What are the components that arise?**
- **How could you use these components?**

First component: [0.24164952 -6.58859157 4.28385556 0.85143375 6.49078937 3.26190303] Customers who bought less than average Milk and lots of Grocery, Detergents, and Deli, and average of everything else. This could be indicative of large grocery stores with a deli counter.

Second component: [0.86341669 0.17817649 -0.86781197 -11.14739034 0.61841913 5.9517213] Customers who bought much less than average Frozen products and lots of Deli, and almost average of everything else. Could be smaller grocery stores with a deli counter.

Third component: [3.96270391 -0.94611005 -0.28411353 -0.71322329 1.30423217 -1.10023843] Customers who bought more Fresh products than average, and about average of everything else. This could be the group for small grocery stores or supermarkets which has no deli section.

Fourth component: [-0.33479516 -7.81234641 3.55607528 0.06090983 -5.18502639 5.30777666] Customers who brought less than average Milk and Detergents_Paper, and more than average Grocery and Deli. This could be the group of convenience stores.

Fifth component: [-0.31028354 1.07281059 13.77162459 -1.2758241 -27.2171808 -5.47709538] Customers who brought much more than average Grocery and much less than average Detergents_Paper. This is the group of large grocery stores which has no deli department.

Sixth component: [-0.38513386 -0.30973265 -0.58676198 -0.52955196 0.44955298 18.16781596] Customers who brought much more than average Deli, and about average of everything else. Customers in this group could be deli shops.

The above components describe the different customer segments that need to be considered, and each customer falling in one of the segments can be grouped with other customers of the same segment when studying or proposing changes affecting that group.

3.2 Clustering

Decide on K means clustering or Gaussian mixture methods

- **What are the advantages and disadvantages of each?**
- **How will you decide on the number of clusters?**

K-Means clustering clumps the customers in K groups based on a distance metric from cluster mean. This helps to identify the different types of customers and relate each customer to other customers within the same group.

Compared to hierarchical clustering, K-Means might give a center of cluster that is not itself a data point in the customers set, cluster region is isotropic around the cluster mean, and cluster sizes (in terms of number of data points) are in general more uniform.

K-means would fail exactly when its assumptions fail: when clusters shapes are irregular, and sizes of clusters are not uniform.

Gaussian mixture model is an EM algorithm that tries to fit Gaussian densities to the data to generate *soft* clusters.

Soft clustering is useful in that it gives the probability that each point is generated by any of the n Gaussian models.

Gaussian mixture model is similar to K-means in that it is an isotropic model and hence would have issues with irregular shaped clusters. In addition, if the clusters point densities are far from Gaussian, the model might fail to produce good clusters.

In terms of complexity, GMM is a more involved algorithm as it needs to calculate the probability that each point came from any of the clusters, in addition, the model requires maximum likelihood calculation of second order statistics, hence a GMM iterations is slower than K-means.

For the case at hand, the data set contains only 440 observations (customers), which makes building the Gaussian model, and specifically computing the second order statistic prone to error due to small number of samples, even if we assume the data was indeed generated from a Gaussian mixture.

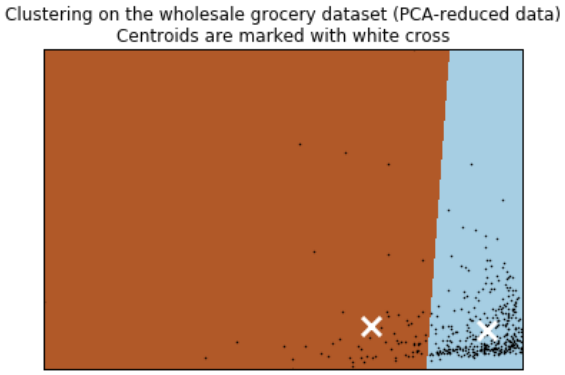
Implement clusters

- **Sample central points of the clusters**

Produce a graphic

- **Visualize important dimensions by reducing with PCA**
- **Are there clusters that aren't very well distinguished? How could you improve the visualization?**

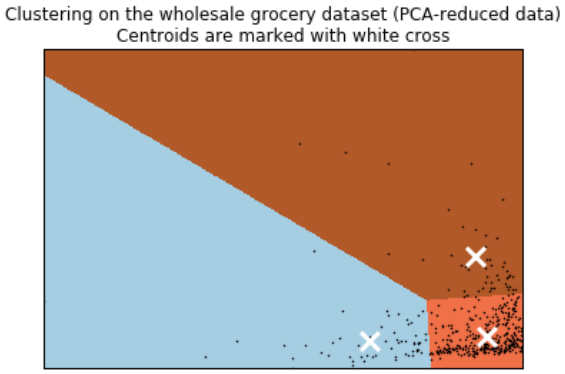
- **Test(1): 2 clusters**



First cluster: Customers who buy lots of "Fresh" products and about distributed on everything else. Second cluster: Customers who don't buy much of fresh products.
We can roughly divide the above as: First and second clusters present large customers in the dataset, whereas third cluster presents smaller customers.

Based on 3 clusters, the calculated silhouette score was 0.542

- **Test(2): 3 clusters**

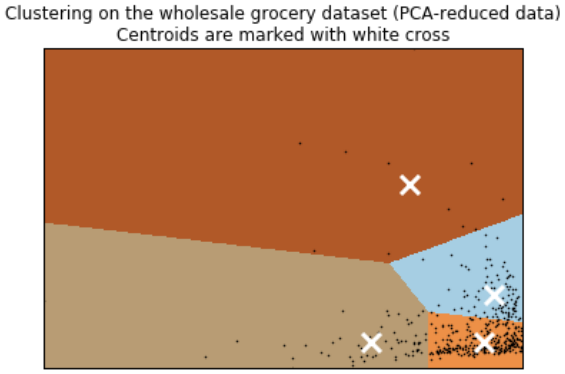


First cluster: Customers who buy lots of "Fresh" products and about average of everything else. Second cluster: Customers who buy lots of "Milk","Grocery", and "Detergent Paper"

and about average of everything else. Third cluster: Customer who buy about average of everything.
We can roughly divide the above as: First and second clusters present large customers in the dataset, whereas third cluster presents smaller customers.

Based on 3 clusters, the calculated silhouette score was 0.523

- **Test(3): 4 clusters**



Based on clusters centroids:

$$\begin{matrix} \mathbf{x} & \mathbf{y} \end{matrix}$$
$$\begin{bmatrix} 6166.2 & 11736.8 \\ -23984.6 & -4910.9 \\ 3496.8 & -5024.8 \\ -14526.9 & 50607.6 \end{bmatrix}$$

We can see that clusters are not as clearly separable as previous cases, but one can roughly identify clusters based on the two first PCA components like we did in the first two cases.

This clustering gave us the worst silhouette score among the three options we tried, which turned out to be 0.454

- **Choice of Clustering**

Based on the above trials, clustering with n=3 gave us sensible results that one could interpret intuitively, in addition, the silhouette score was not much worse than n=2 clusters, so we select n=3 as our clustering of choice for this project.

3.3 Conclusions

- **Which of these techniques felt like it fit naturally with the data?**

  The last, PCA + Clustering method, gave a useful summary of different customer types which can be easily explained visually.
  However, all the techniques gave some perspective into the data as explained above.

- **How would you use that technique to assist if the company conducted an experiment?**

  When a new experiment is to be designed, different clusters(segments) of customers need to be studied. It is not enough to conduct an experiment within customers with large daily Fresh products transactions, as these customers might have different perspective than customers from other segments.

  To succeed with multiple customer segments, one needs to cater to different segments needs separately.

  To have a concrete example of the above, lets assume we plan to perform an A/B test to check whether an environmental-friendly packaging of deli products shall have a positive or negative effect on customers.

  Following steps should be performed in order:

  1. Find the different customer segments that are affected by the change above. From the example above, this mounts to all customers buying deli products.
  2. **Within each of the segments**, pick a control group and a test group. Lets say we start with large grocery stores that have a deli counter, we pick some customers for the test group and keep the remaining customers within the segment as our control group (ie customers who won't observe the change we are testing). We do the same separately for other segments, such as deli shops.
  3. We quantify the results within each segment and conclude whether we should make the change to proposal B, or stay at our current version for that segment. In our example, a metric that can be used to measure our test success is the amount paid for deli products by group B against group A.

- **How would you use that data to predict future customer needs?**

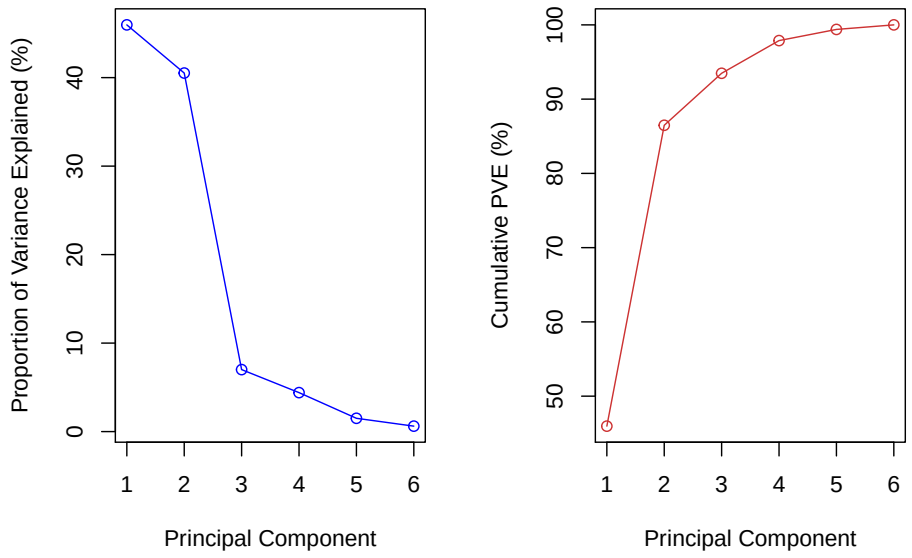
  Based on properties of the new customer (store size, location, popularity, types of sold products, etc..), one can map the customer to one of the clusters above, and predict the amount of different products to be requested by the customer based on other samples within the same cluster.

4. Further Explorations with R

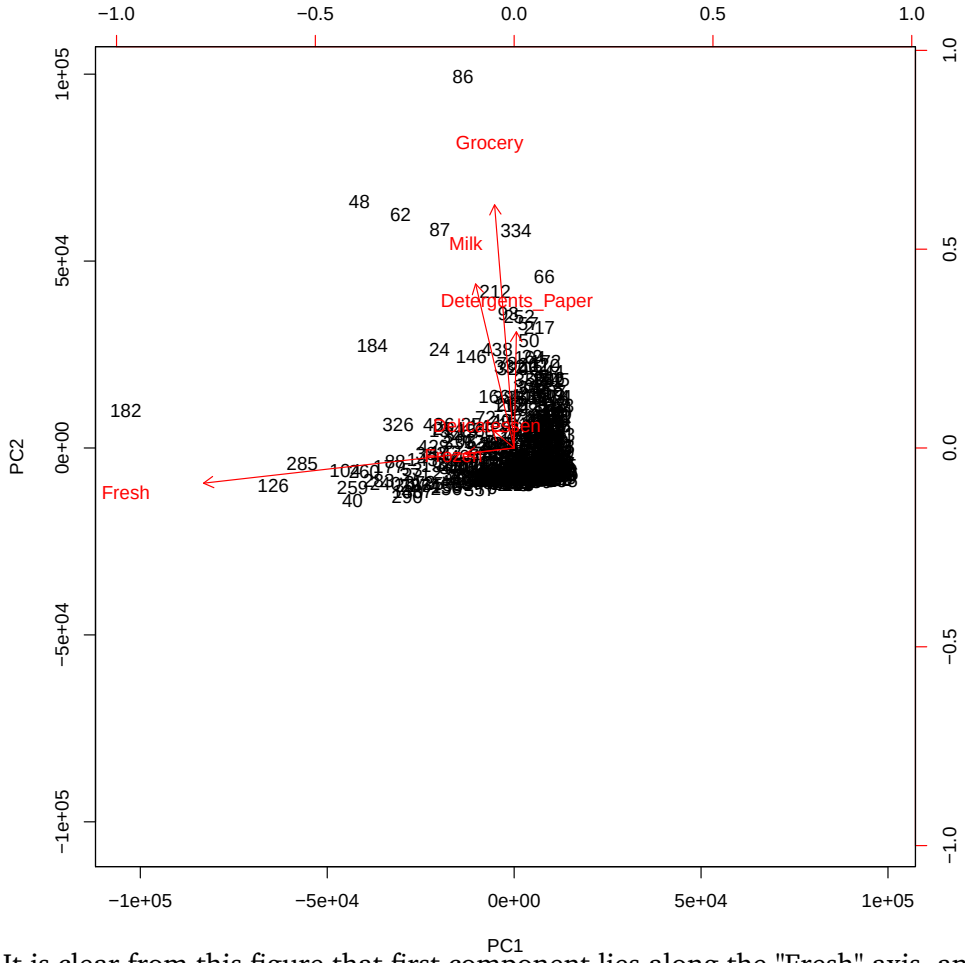
This is not part of the project, but an addendum for explorations performed using R statistical package on the customers dataset.

4.1 Principal Component Analysis (PCA)

PCA was applied to the dataset, and proportion of variance explained is shown in the plot below:



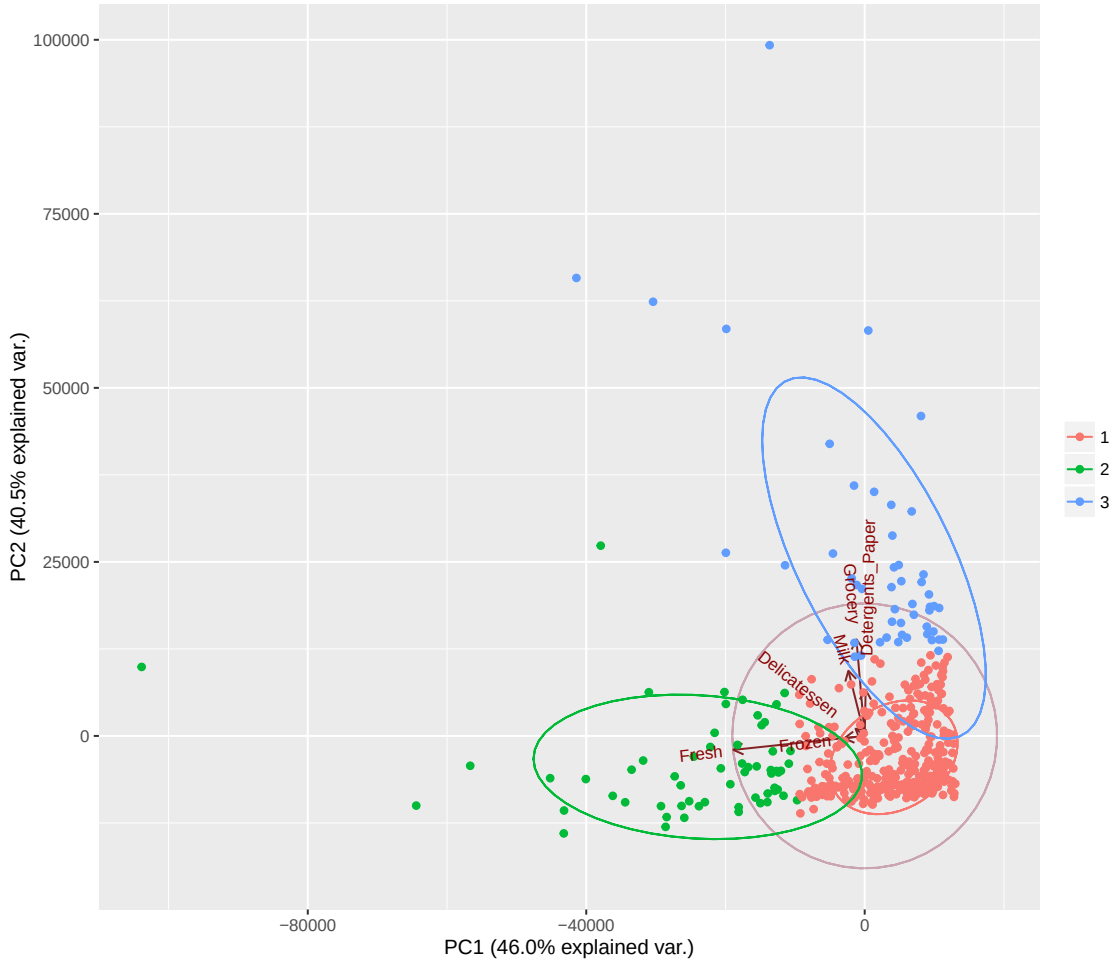
As described in the project report, variance explained drops significantly starting from third component onwards.
The following plot shows a *biplot* of the PCA using the first two principal components:



It is clear from this figure that first component lies along the "Fresh" axis, and the second lies along "Grocery", "Milk", and "Detergent_Paper".
by looking at variance of different features, we can see that indeed "Fresh" dimension has the largest variance, and that would explain that first principal component falls close the "Fresh" vector.

4.2 K-means

The last section studies the relation between K-means and principal components, and the figure below shows how the three clusters relate to the first 2 principal components:



The plot shows that the customers where clustered according to whether they buy lots of "Fresh" products, or lots of "Grocery" "Milk" "Detergent_Paper", more or less average of all the above items.