Questions and Report Structure

Component analysis

1. Reflection on PCA/ICA
    - What are likely candidates for early PCA dimensions?

      Based on a basic set of statistics, I would say that the first PCA dimensions will correspond to Fresh because it is the feature with the highest variance (in this case standard deviation). Grocery and Milk are up there next.

    - What might ICA dimensions look like?

      In the case of ICA we are going to find independent vector represented by a combination of the initial features. I would expect the independent components to represent the buying patterns of different types of customers composed by their spending in different product categories.

2. What proportion of variance is explained by each PCA dimension?

   The first two principal components make for about the 85% of the variance (45% and 40% respectively) and then drops to about 7%.

   Based on this, I would argue that we should choose two dimensions for PCA because they explain the vast majority of the variance. The next PCs seam not that relevant.

3. PCA dimensions
    - What are the first few components? What might they represent?

      In the case we use two dimensions, I would argue that the first Principal Component represent mainly Fresh due to the magnitude of the Fresh component being much higher than the rest. I would say that the second PC is a combination of Grocery, Milk and Detergents_Paper (in that order of importance) based on the magnitude of their components.

      *From searching the Discussion Board I found an implementation of a biplot (provided by other student, jjinking) that confirmed my findings based on the magnitude / direction of each variable arrow.

- How can you use this information?

  High dimensionality datasets require significantly more computer power than low dimensionality datasets. Most Machine Learning algorithms will struggle with this (in supervised or unsupervised learning). One solution is to find a solution through a sample from the dataset and another solution is to lower the dimensionality of the data set. In this sense, PCA can be used to identify the most important dimensions (or create new ones) in order to lower the dimensionality of the data.

  In this case, I would argue that using only 2 PCA dimensions to find customer segments would be enough to get a solution. This dimensions will probably represent Fresh and the combination of Grocery, Milk and Detergents_Paper.

4. ICA

   - What are the components that arise?

     I will explain this vectors in terms the different type of customers based on their expenditure behavior:

     - Vector 0: Customers that mainly consume Fresh and Detergents Paper in a correlated way.
     - Vector 1: Customers that mainly consume Grocery with a low negatively correlation with Milk.
     - Vector 2: Customers that consume a combination of Milk and Detergents Paper with a negative correlation with Grocery and Delicatessen.
     - Vector 3: Customer that almost exclusively consumes Delicatessen.
     - Vector 4: Customers that consume mainly Detergents Paper with some Delicatessen and a negative correlation with Grocery.
     - Vector 5: Consumer that has expenditures of Frozen with negative correlation with Delicatessen.

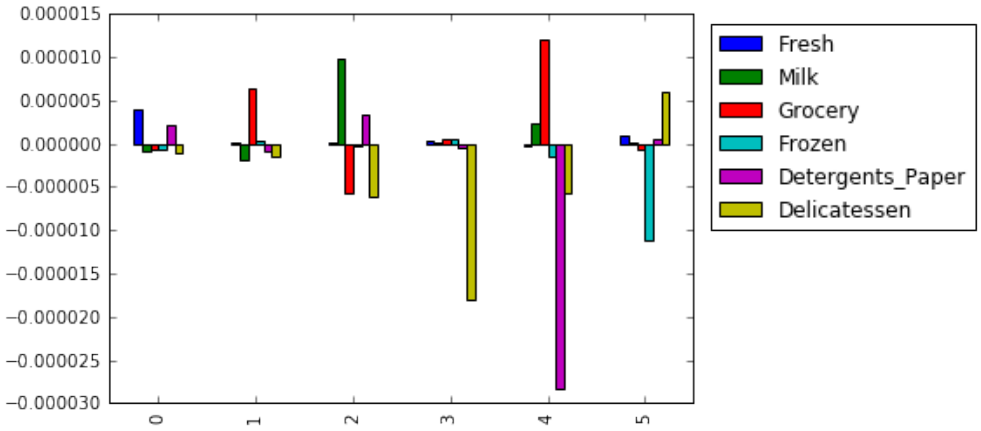

Figure N°1: Visual representation of the ICA components after fit

- How could you use these components?

  The components can be used to transform the original data set into a data set based on independent vectors which will help clustering and classification algorithms to understand trends in data more easily. This is specifically important for algorithm that require independence in between features.

  This vectors also give us insights related to have different expenditure patterns could show us how customers really are and help us understand/interpret the clustering results.

Clustering

2. Decide on K means clustering or Gaussian mixture methods
   - What are the advantages and disadvantages of each?
     - K-Means is a hard assignment algorithm in comparison to GMM that is a soft assignment algorithm. This difference relies in the fact that in K-Means assigns an element to a particular cluster (the element *only* belongs to that cluster) while in GMM an element *could* belong different clusters with different probabilities. The soft assignment can cause confusion interpreting the results.
     - Both Clustering algorithms need the number of clusters as an input (which could be a challenging task on its own).
     - GMM considers a higher number of parameters than K-Means which makes GMM would be take longer to tune (determine the optimal configuration).

- K-means is scalable on very large number of observations and on medium number of cluster while GMM is not scalable. In this sense, on big datasets K-means is recommended.
- In each iteration, K-means needs to calculate the mean of each cluster while GMM needs to calculate the maximum likelihood gaussian distribution (which involves the mean and the standard deviation of more observations). Considering this, we can say that GMM needs to perform more computations than K-means, therefore making it slower.**

Under this distinctions, I would rather work with K-Means than GMM mainly because I would like to work around an algorithm that is easier to interpret (each customer belongs to one cluster and not to many), K-means is easier to tune and the size of the dataset will make speed irrelevant.

** *Thanks to Mitchel and my code reviewer (sorry I didn't catch your name) for helping me address the speed comparison.*

- How will you decide on the number of clusters?

I will make a visual inspection of the reduced data using PCA on two PCs. Based on this inspection I will try to predict the number of clusters, and will start using both algorithms on this number of clusters. Results will be analyzed on several number of clusters to check their results.

My main goal here is to achieve clusters that are different from each other between clusters, and members of each clusters should be as similar as possible.

3. Implement clusters
    - Sample central points of the clusters

4. Produce a graphic
    - Visualize important dimensions by reducing with PCA

Based on Figure Nº2, we can see that most customers are placed on the bottom right of the graph. In this case the x axis (PC1) is represented mainly by Fresh and the y

axis (PC2) is represented by the combination of Grocery, Milk and Detergents_Paper based on the earlier conclusions.

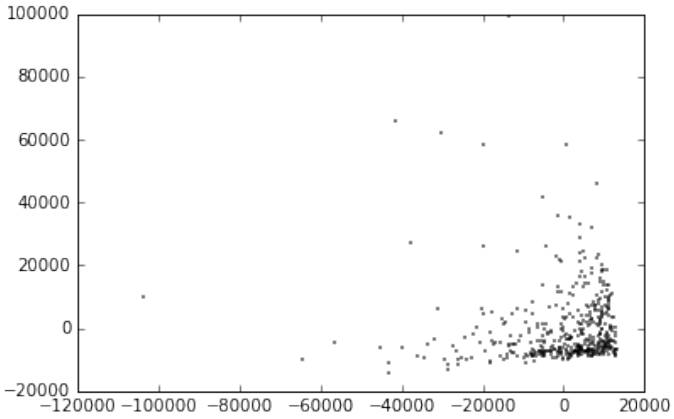

Figure Nº2: Visual representation on the reduced data set using PCA with two PCs

My first appreciation is that there are three clusters. One cluster representing the customers on the bottom right which mostly spend on Fresh, other cluster representing those who spend more on Fresh while spending the same on Grocery, Milk and Detergents_Paper and other cluster were customer spend more on Grocery, Milk and Detergents_Paper and the same on Fresh.

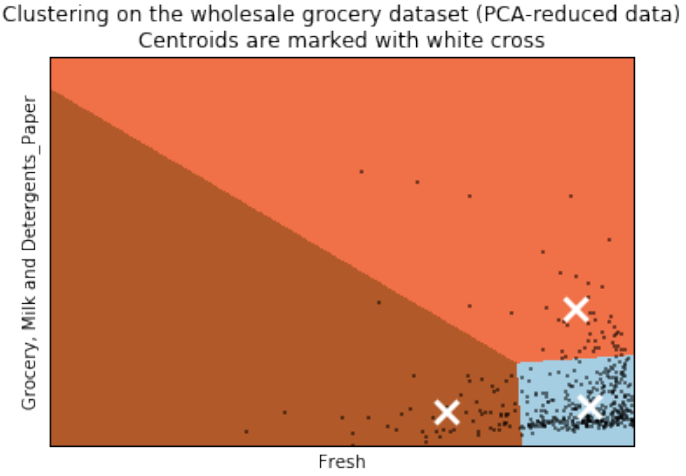

Figure Nº3: Visual representation of the K-Means results with k = 3.

To confirm this findings, I plotted the centroids with their original values (after an inverse transform from the PCA transformation) and made a plot with their spending (Figure Nº4). From this plot we can see that cluster 0 has spending's mainly on Fresh

and Grocery but has a lower spending in comparison with clusters 1 and 2. Cluster 1 has higher spending's on Grocery, Milk and Detergents Paper while having about the same spending on Fresh than cluster 0. Cluster 2 is focused almost entirely on Fresh while having relatively similar spending on the rest of the product categories than cluster 0. These findings confirm the first intuition rewarding how customer segments are.

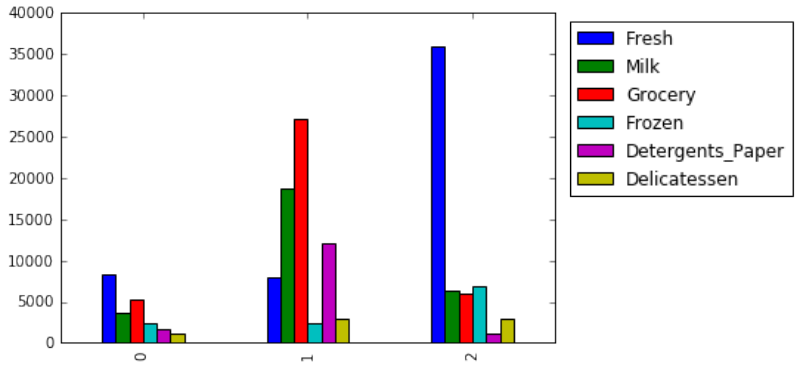

Figure N°4: Centroid spending for the results of the K-Means algorithm with k = 3.

After the visual inspection, I proceeded to test both clustering algorithms with k values between 2 and 4. The best outcome in my opinion came from the k-means clusters with k = 3.

After several iterations I chose this number of k based on the visual inspection of the results. The results associated with 4 clusters (Figure Nª5) showed one clusters with very few customers that were quite sparse (the ones at the top) while this same issues happened while increasing the number of clusters. 2 clusters (Figure Nª6) appeared to group customers that were not very similar between each other (cluster to the right, where some customers had a low spending on Grocery, Milk and Detergent Paper while others had a high spending).

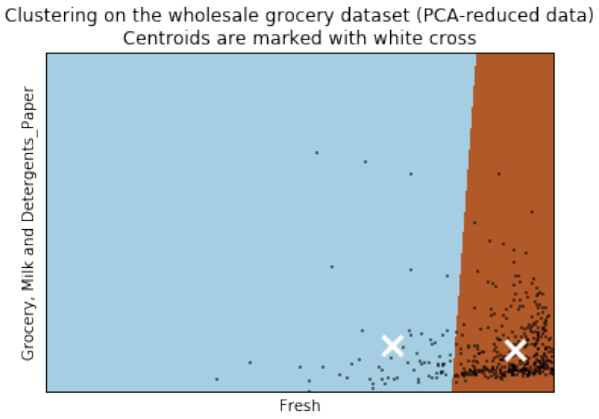

Figure N°5: Zoom over on cluster of the visual representation of the K-Means results with k = 2.

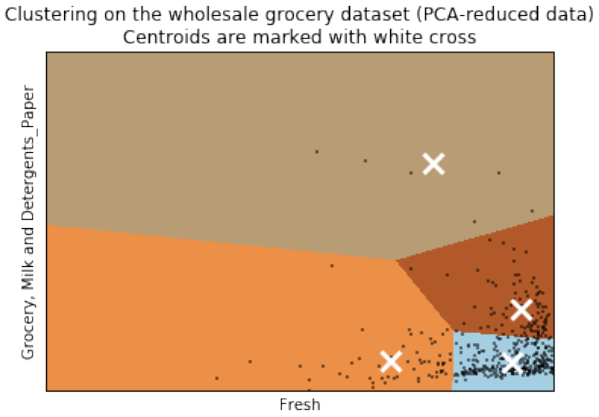

Figure N°6: Zoom over on cluster of the visual representation of the K-Means results with k = 4.

- Are there clusters that aren't very well distinguished? How could you improve the visualization?

  The bottom right cluster (light blue) has a high concentration of customers which are hard to visualize.

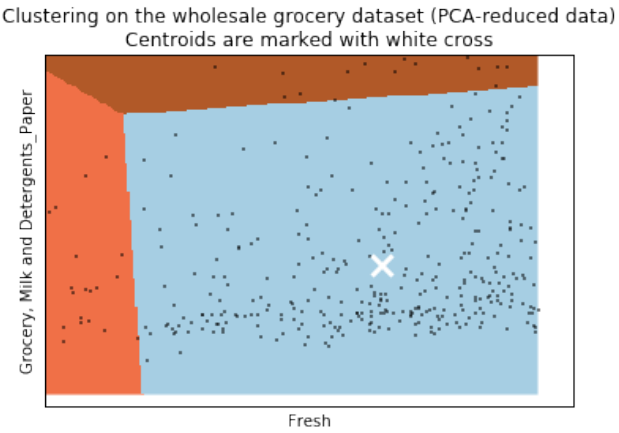

Figure N°7: Zoom over on cluster of the visual representation of the K-Means results with k = 3.

I think this cluster should be further analyzed based on the high density it has and to understand the differences between customers in this cluster. There seams to be a particular group with similar spending's on Grocery, Milk and Detergents_Paper but with different spending's on Fresh. This type of analyze would help us understand sub-groups on a particular type of customer and would be helpful for understanding their behavior in depth.

Conclusions

5. Which of these techniques felt like it fit naturally with the data?

The results for K-means show there are three customer segments, (1) one segment with high spending on Fresh and low spending on Grocery, Milk and Detergent (this segments has the highes amount of customers and should probably be analyzed more in depth), (2) one segment with the same levels of spending on Grocery, Milk and Detergent but with a lower amount of spending on Fresh and (3) a last segment with high spending on both Fresh and on Grocery, Milk and Detergent.

I would argue that segment (3) are customers of big supermarkets where they buy everything, (2) are segments of customers that go to local stores where they search for their day to day needs and (3) are small business customers (with low spendings).

After evaluating the results for GMM given 3 customer segments (cluster) I found that K-means gives more insights about this data mainly because the results on GMM showed clusters that overlapped a lot (mainly the one at the bottom right and top right) on GMM. This overlap resulted in some confusion about how the customers really are. Figure Nº8 shows the results for 3 clusters.

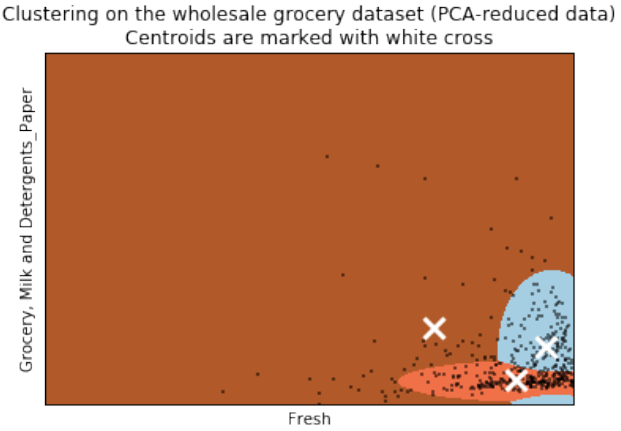

Figure Nº8: Visual representation of the GMM results with k = 3.

6. How would you use that technique to assist if the company conducted an experiment?

   After determining the number of customer segments, I would focus an experiment to a determined customer segment and not to all customers and check the performance of the experiment. This way we can control the effects of a certain experiment to a particular group were we could choose to expand the experiments to other segments after good results or we choose to stop the experiment after bad results.

   We could also design specific experiments customer segments, for example sending an email with a special offer associated with the spending to a subgroup and checking if their spending increased, remained the same o was lowered.

7. How would you use that data to predict future customer needs?

   Identifying the customer segments may not be enough for understanding customer needs. In this sense I would start conducting additional research on each customer segment separately. As mentioned before, I think the bottom right cluster needs additional research (based on the density and the amount of customers).

I would also try to find additional information (socio demographic and behavioral data) to analyze differences between customers and explore for further needs. Additional Supervised learning algorithms could be used to predict the probability of a customer to buy a certain product o to predict if the customer may churn the company.