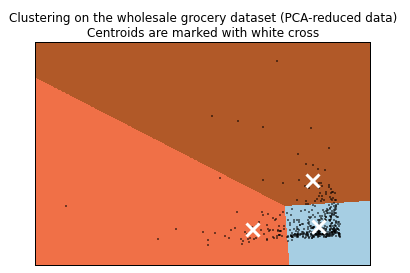
**Questions and Report Answers**

**Component analysis**

1. Reflection on PCA/ICA
   * What are likely candidates for early PCA dimensions?
   * What might ICA dimensions look like?
   * I think that PCA will find that some of the categories like Fresh and Milk are similar and will find a way to group them together in one of its principle components; this might be a surrogate for freshness. ICA is harder to predict because it is looking to create statistically independent features from the current features. Maybe ICA might find features that are more like the cuisine/type of food?
2. What proportion of variance is explained by each PCA dimension?
   * The output above shows how we can create the new PCA feature/dimensions. If we multiply the rows by the original data (as columns) we'll get our transformed principle components. The pca.explained\_variance\_ratio (which is the vector of the eigenvalues for this analysis) shows that the first 2 components explain a lot of the variance (0.4596 and 0.40517). After that the components don't seem to describe much of the variance at all (<= 0.07004). So, I would probably only use the first 2 components in my analysis.
3. PCA dimensions
   * What are the first few components? What might they represent?
   * How can you use this information?
   * Each row in the pca.components output represents one of the principle components and each column represents one of the original features input into PCA. This table seems to suggest that Fresh is one strong (negative) component (-0.9765), and the 2nd component seems to be a pretty strong mix of Milk, Grocery, and a little Detergent-Paper (0.515, 0.764, and 0.365 respectively). I'll call the first component Fresh Food and the 2nd component Regular Groceries, since this seems to convey the meaning that these new components would have in reality. We can now use these 2 dimensions (instead of 6) to describe the whole data set while maintaining 86.5% (46% + 40.5%) of the original data’s variance. Using these 2 dimensions can help us avoid the “curse of dimensionality”. We can also use these 2 dimensions for future work like clustering, future supervised learning model fitting, etc.
4. ICA
   * What are the components that arise?
   * How could you use these components?
     1. This vector takes a lot of negative effect from Grocery (-1.116) and a lot of positive effect from Detergents\_Paper (1.151). Maybe this vector would be good for describing how much non-food items a store buys.
     2. This vector takes a lot of its effect from Delicatessen (0.867). This one just seems to be the Delicatessen effect with a little bit from others to make it statistically independent.
     3. This vector is positively influenced by Grocery (0.599) and negatively a little by Milk (-0.138). This one might represent how much the store buys of non-dairy Grocery items.
     4. This vector is negatively influenced by Frozen (-0.678) and slightly positively influenced by Delicatessen (0.285). The negative signal from this vector could describe how much the store buys frozen foods, compared to refrigerated/ready-made foods in a Deli.
     5. This vector is negatively influenced by Milk (-0.723) and positively by Grocery (0.539). This negative signal of this vector could describe how much Milk is bought compared to groceries. That might be an important ratio...
     6. This vector is positively influenced by Fresh (0.446), with traces of the others...
   * These components could be used to cluster the stores by the type of items they buy, perhaps regardless of the quantity. They could also be used to describe different (statistically independent) types of store purchasing behaviors.

**Clustering**

1. 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?
   * Gaussian Mixture Models find clusters centers much like K-means, but Gaussian Mixture Models have the advantage of being able to change the size/shape of their probability density functions to better match the cluster shape. For example, if we had a cluster that was shaped like a horizontal oval, Gaussian Mixture models would be able to change the variance of the gaussian distribution to more accurately include the long/wide shape of that cluster. A K-means cluster just looks for the distance between the cluster center and the closests points. The k-means algorithm would be faster, but might not cluster the data as well. The other big advantage of the K-means algorithm is that the cluster boundaries will often be linear and therefore potentially easier for a human to understand. Since I'm doing this analysis for a business client, I'll try the simpler K-means first to see if the result looks reasonable, and then try GMM as a comparison (you can see commented code below for the GMM solution as well).
2. Implement clusters
   * Sample central points of the clusters
     1. Looking for Elbow in K-Means score
     2. k=2; Score: -610.4
     3. k=3; Score: -380.91
     4. k=4; Score: -277.41
     5. k=5; Score: -207.38
     6. k=6; Score: -169.16
     7. k=7; Score: -136.67
     8. KMeans(copy\_x=True, init='k-means++', max\_iter=300, n\_clusters=3, n\_init=10,
     9. n\_jobs=1, precompute\_distances='auto', random\_state=None, tol=0.0001,
     10. verbose=0)
     11. #centroid outputs
     12. [[ 0.32398252 -0.25421161]
     13. [-1.86890029 -0.36902956]
     14. [ 0.10439573 2.12063212]]
3. 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?
  + There are 3 cluster centers:
    1. The first center is at (0.3239, -0.2542) which is the white X in the blue region of the plot. Since in the first principle component represents Fresh Food and the 2nd represents groceries, this would be a cluster of stores that buy more fresh food and not as many groceries. Perhaps these are fruit stands, or health food stores...
    2. The 2nd cluster center is at (-1.8689, -0.36) which is the white X in the orange region of the plot. These stores are don't really buy fresh foods, and don't buy as many groceries either. Maybe these are convenience stores?
    3. The 3rd cluster center is at (0.1044, 2.121) which is the white X in the brown region of the plot. These stores buy fresh foods, but they are mostly huge purchasers of groceries. I would describe these stores as being larger grocery stores like Wal-mart.

**Conclusions**

1. Which of these techniques felt like it fit naturally with the data?
   * I thought that I got the most insight into the data with PCA and K-means. PCA did a great job of quickly telling me what the most important (combined) factors were in the feature set. K-means output was much more human understandable (3 separate regions). Those 3 regions helped me think about the customers and what was different about them. I could intuitively tell what types of businesses they were based on the clustering. ICA and GMM were new concepts to me and I was hoping that they would give new and interesting insights, but the information/output they gave seemed hard to interpret. ICA didn't seem to add much value to the interpretation of the data, and didn't reduce the dimensionality of the problem either. If I had just taken 2 of the features like I did with PCA, I would have lost a lot of the information in the feature set. GMM's output plot, while you can't see it in this final version, was weird to look at. The clusters were roundish and one of the clusters was divided by another cluster running right through the middle of it. Luckily, PCA allowed me to get the feature space down to 2 dimensions so I could see this with my eye and decide which method I liked better for this problem.
2. How would you use that technique to assist if the company conducted an experiment?
   * Using the cluster assignment in the experiment would allow the company to determine if there were different responses from different customer segments, instead of averaging results over the entire customer population. They might run an A/B test in just one customer segment at a time in the future to get a clearer understanding of each customer segments' needs. For example, they could retry the experiment on just one segment they ran on the whole customer base, “chang[ing] their delivery method, from a regular morning delivery to a cheaper, bulk evening delivery”. Using A/B testing, different customers within the new customer segment would be randomly selected to either get the original service or the trial bulk delivery. Effects from this experiment (perhaps revenue or customer complaints) could be analyzed to determine if there was any statistical significance between the 2 randomly selected sub-groups within the customer segment being tested.
3. How would you use that data to predict future customer needs?
   * If we ran small A/B experiments as suggested above in targeted segments, then the business would be able to extend the improved results to the rest of the segment, without fear of hurting results in the other segments (like what happened to them in the past). Perhaps they might try to optimize profit for the biggest segment, or they might try to grow their business in a smaller customer segment through this type of experimentation. Over time, this might allow the company to find that these different segments prefer different types of service/products thus helping them to better serve their customers, improve customer loyalty, and increase the company’s profits.

If by future customer needs, the question is referring to new customers, then the business would be able to assign the new customer to a cluster pretty quickly and then be able to serve that customer like other similar customers. This would help to retain those new customers.