***Question 1: Consider the total purchase cost of each product category and the statistical description of the dataset above for your sample customers. What kind of establishment (customer) could each of the three samples you've chosen represent? Hint: Examples of establishments include places like markets, cafes, and retailers, among many others. Avoid using names for establishments, such as saying "McDonalds" when describing a sample customer as a restaurant.***

Our first example looks like it might be a restaurant or cafe. There is high spend in Fresh (I'm assuming this is fresh produce) and Grocery. There is also higher spend in Detergents\_Paper as well as more Delicatessen spend than the other two samples. The second sample might be a cafeteria for a school or university. There is low spend in Fresh foods compared to the Milk, Grocery, and Delicatessen samples. Additionally, the Frozen and Detergents\_Paper categories have a higher spend than both the other samples. Lastly, the third would look something more of a Farmer's Market (although I'm not sure if food for famer’s markets run through a warehouse first). Fresh food represents almost half of the spend for all categories combined and Frozen foods only contribute a fraction. Milk and Grocery represent the next sizeable categories in spend, but they are not the prominent drivers.

***Question 2: Which feature did you attempt to predict? What was the reported prediction score? Is this feature necessary for identifying customers' spending habits? Hint: The coefficient of determination, `R^2`, is scored between 0 and 1, with 1 being a perfect fit. A negative `R^2` implies the model fails to fit the data.***

I attempted to predict Detergents\_Paper and the R^2 score was about 0.729. Without testing the other scores, I would infer that it is important enough to keep for us to correctly identify our customer’s spending habits.

***Question 3: Are there any pairs of features which exhibit some degree of correlation? Does this confirm or deny your suspicions about the relevance of the feature you attempted to predict? How is the data for those features distributed? Hint: Is the data normally distributed? Where do most of the data points lie?***

We can see that there about three pairs of features that correlate with one another; Grocery and Milk, Detergents\_Paper and Milk, and Grocery and Detergents\_Paper. This does confirm that Detergents\_Paper would be one of the important features to keep when trying to predict a customer’s spending. The data is skewed right, meaning the mean is greater than the median spend amount. It would be important to dive understand what is happening with our customers driving that skew first as they are the big spenders. Fixing the accounts that drive the most spend and applying the same methods to most other accounts can help us solve our monetary bleeding faster than focusing on the small spenders first.

***Question 4: Are there any data points considered outliers for more than one feature based on the definition above? Should these data points be removed from the dataset? If any data points were added to the `outliers` list to be removed, explain why.***

Points 66, 75, 142, 154, 289, and 338 are all outliers that occur in two or more of the categories. Removing the points who outliers in multiple categories might not be productive if we are trying to infer relations between customer spending and product categories. Especially the outliers that are on the upper bound because those are the ones with the most spend.

***Question 5: How much variance in the data is explained in total by the first and second principal component? What about the first four principal components? Using the visualization provided above, discuss what the first four dimensions best represent in terms of customer spending. Hint: A positive increase in a specific dimension corresponds with an increase of the positive-weighted features and a decrease of the negative-weighted features. The rate of increase or decrease is based on the individual feature weights.***

Components One and Two: `0.4470 + 0.2746 = 0.7216`<br>

Components One through Four: `0.4470 + 0.2746 + 0.1123 + 0.1004 = 0.9343`<br>

The first two components account for about 72% of the variance in our data and the sum of the first four represents about 93% of variance.<br>

Our first dimension shows a strong relation between Milk, Grocery, and Detergents\_Papers and a similar relation with Milk and Frozen. The second dimension is almost the inverse of the first where Milk, Frozen, and this time Delicatessen are the primary drivers of the variance with Milk, Grocery, and Detergents\_Paper are also related with smaller weights. These two dimensions explain more than half the variance.

***Question 6: What are the advantages to using a K-Means clustering algorithm? What are the advantages to using a Gaussian Mixture Model clustering algorithm? Given your observations about the wholesale customer data so far, which of the two algorithms will you use and why?***

<b>K-Means:</b> This clustering algorithm tries to put the data into K groups that equalizes the variance between each other and minimizes the sum-of-squares from the centroid. However, the K-Means algorithm does not accommodate non-uniform shaped clusters very well. (1)

<b>Gaussian Mixture Model:</b> This algorithm is an efficient way to group together data that are not uniform in shape or size. While the K-Means model tries to find an equal variance between groups, a Gaussian Mixture Model will try to capture the shape instead. When we look at the Biplot above we can see that our data will not be completely separable into clusters as the K-Means algorithm will force. "It can also draw confidence ellipsoids for multivariate models"(2). This would allow us to draw boarders with soft boundaries so that points can be represented into multiple groups rather than one or the other. For that reason, we will continue the project using the Gaussian Mixture Model.

(1): <href><http://scikit-learn.org/stable/modules/clustering.html#k-means></href>  
(2): <href><http://scikit-learn.org/stable/modules/mixture.html></href>

***Question 7: Report the silhouette score for several cluster numbers you tried. Of these, which number of clusters has the best silhouette score?***

2: GMM: 0.415, KMeans: 0.426

3: GMM: 0.401, KMeans: 0.400

4: GMM: 0.314, KMeans: 0.335

My initial choice of model was the Gaussian Mixed Model and it looks like two clusters was optimal for this set. After comparing the KMeans with it however, we can notice that the KMeans score of 0.426 (still two clusters) performs slightly better than the GMM.

***Question 8: Consider the total purchase cost of each product category for the representative data points above, and reference the statistical description of the dataset at the beginning of this project. What set of establishments could each of the customer segments represent? Hint: A customer who is assigned to `'Cluster X'` should best identify with the establishments represented by the feature set of `'Segment X'`.***

<b>Segment 0</b> looks like a market of sorts. The very high spend on Fresh foods, mid spend on Milk, Grocery, and Frozen, and low spend in Detergents\_Paper and Delicatessen make it look like the focus is on fresh foods and other complimentary foods.<br>

<b>Segment1</b> might be a restaurant/cafe. We can see high spending in Milk, Grocery, and Detergents\_Paper. There is lower spending in the Fresh category.

***Question 9: For each sample point, which customer segment from Question 8 best represents it? Are the predictions for each sample point consistent with this?***

No, my predictions did not match these results. My initial thoughts were that point 0 would be close to the middle of the two clusters, but more on the restaurant/cafe side. The second point would also be restaurant/cafe. The third point I had predicted to be a market. All three points however, were classified as cluster 1, which I had guessed to be a restaurant/cafe.

***Question 10: Companies will often run*** [***[A/B tests]***]((https://en.wikipedia.org/wiki/A/B_testing)) ***when making small changes to their products or services to determine whether making that change will affect its customers positively or negatively. The wholesale distributor is considering changing its delivery service from currently 5 days a week to 3 days a week. However, the distributor will only make this change in delivery service for customers that react positively. How can the wholesale distributor use the customer segments to determine which customers, if any, would react positively to the change in delivery service? Hint: Can we assume the change affects all customers equally? How can we determine which group of customers it affects the most?***

The distributor can take each segment and split it them into two groups. In Segment0 there would be group A and group B and the same for Segment1. Each group should have an even spread of total spend. Our A group can be our constant group with no changes while the B group tries the new 3-day delivery program. If the program is successful in the B group, then it can be applied to the A group. Having the two segments helps us analyze what sorts of customers this affects and whether this would be applicable to all customers, just the ones in Segment0 or just the ones in Segment1.

***Question 11: Additional structure is derived from originally unlabeled data when using clustering techniques. Since each customer has a customer segment it best identifies with (depending on the clustering algorithm applied), we can consider 'customer segment' as an engineered feature for the data. Assume the wholesale distributor recently acquired ten new customers and each provided estimate for anticipated annual spending of each product category. Knowing these estimates, the wholesale distributor wants to classify each new customer to a customer segment to determine the most appropriate delivery service. How can the wholesale distributor label the new customers using only their estimated product spending and the customer segment data? Hint: A supervised learner could be used to train on the original customers. What would be the target variable?***

We can follow the same process that we had for this project. Starting with a normalization of the spending categories across all the original customers, we can then reduce the current feature set using PCA. After compressing the feature set, we can search for the optimal number of clusters using trial and error and measuring the Silhouette Score or we could use BIC (Bayesian Information Criterion) to help us find the right number of clusters. Once we have our clusters and find the cluster centers, we can unpack the reduced dimensions and see what our centers would be as normalized spend. When we introduce our new customers, their spend will also be normalized. Rather than going through the whole PCA process, we can simply measure their distance from each cluster center and group them under the one with the shortest distance.

***Question 12: How well does the clustering algorithm and number of clusters you've chosen compare to this underlying distribution of Hotel/Restaurant/Cafe customers to Retailer customers? Are there customer segments that would be classified as purely 'Retailers' or 'Hotels/Restaurants/Cafes' by this distribution? Would you consider these classifications as consistent with your previous definition of the customer segments?***

The two clusters that we predicted back in Question 7 are consistent with the customer segments here. The split and number of clusters we predicted are like the visualization. Our predictions of the cluster labels were flipped from the actual names though. Segment0 (the red cluster) we predicted to be a Farmer’s Market but it was Restaurants and Cafés while Segment1 (green grouping predicted as Restaurants and Cafés) was Retail. For this split, there are some points whom are much more defined than others. A handful of the Retail and Restaurant/Cafés points are mixed in the opposite clusters. There is no perfect way to split this data.