**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, delis, wholesale retailers, among many others. Avoid using names for establishments, such as saying *"McDonalds"* when describing a sample customer as a restaurant. You can use the mean values for reference to compare your samples with. The mean values are as follows:

* Fresh: 12000.2977
* Milk: 5796.2
* Grocery: 7951.277273
* Frozen: 3071.931818
* Detergents\_paper: 2881.4
* Delicatessen: 1524.8

Knowing this, how do your samples compare? Does that help in driving your insight into what kind of establishments they might be?

|  | **Fresh** | **Milk** | **Grocery** | **Frozen** | **Detergents\_Paper** | **Delicatessen** |
| --- | --- | --- | --- | --- | --- | --- |
| **0** | 3366 | 5403 | 12974 | 4400 | 5977 | 1744 |
| **1** | 6269 | 1095 | 1980 | 3860 | 609 | 2162 |
| **2** | 11594 | 7779 | 12144 | 3252 | 8035 | 3029 |

Answers:

The purchase cost of each product category for each sample customer is compared to the corresponding mean value from the whole dataset.

The first customer purchases much fewer Fresh, on average Milk, much more Grocery, more Frozen, much more Detergents\_Paper, and on average Delicatessen. The establishment of this sample represents fast food restaurant because of the low level of Fresh and the high level of Grocery, Frozen, and Detergents\_Paper.

The second customer purchases fewer Fresh, much fewer Milk, much fewer Grocery, on average Frozen, much fewer Detergents\_Paper, and much more Delicatessen. The establishment of this sample represents to-go deli because of the high level of Delicatessen and low level of other categories.

The third customer purchases on average Fresh, more Milk, much more Grocery, on avarage Frozen, much more Detergents\_Paper, and much Delicatessen. The establishment of this sample represents dine-in restaurant serving sandwiches and desserts because of the high level of Milk, Grocery, Detergents\_Paper, and Delicatessen.

**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. If you get a low score for a particular feature, that lends us to beleive that that feature point is hard to predict using the other features, thereby making it an important feature to consider when considering relevance.

**Answer:**

I attempt to predict the feature \*\*Grocery\*\*. The prediction score is R^2= 0.656050008527. This is a low score, lending me to believe that \*\*Grocery\*\* is hard to predict using the other features and is an important feature to consider when considering relevance.

Question 3

Using the scatter matrix as a reference, discuss the distribution of the dataset, specifically talk about the normality, outliers, large number of data points near 0 among others. If you need to sepearate out some of the plots individually to further accentuate your point, you may do so as well.

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? You can use corr() to get the feature correlations and then visualize them using a heatmap(the data that would be fed into the heatmap would be the correlation values, for eg: data.corr()) to gain further insight.

Answer:

\* Six features are all approximately in truncated normal distribution (truncated at x=0), all have a few outliners (Detergents\_Paper only has one outliner), and all have large number of data points near 0 among others.

\* Heatmap shows that Grocery-Milk (0.73), Grocery-Detergents\_Paper (0.92), Milk-Detergents\_Paper (0.66) exhibit high degree of correlation, and other pairs exhibit limited degree of correlation.

\* The correlation values confirm my prediction that features \*\* Delicatessen Fresh, Frozen, and Milk\*\* are relevant for specifying a customer.

\* Data of Delicatessen Fresh, Frozen, and Milk are in truncated normal distribution, have a few outliner, and have large number of data points near 0 among others.