Sample 1

| Fresh | Milk | Grocery | Frozen | Detergents\_Paper | Delicatessen |

|--------------------------------------------------------------------|

| 11818 | 1648 | 1694 | 2276 | 169 | 1647 |

From the above descriptive statistics of wholesale data, the sample 1 looks like more of Delicatessen or Wholesale outlet for Fresh foods, since the sample has Delicatessen value greater than the mean and lies between 50 to 75% maxima and also the sample value for Fresh food is near to mean value.

Sample 2

| Fresh | Milk | Grocery | Frozen | Detergents\_Paper | Delicatessen |

|--------------------------------------------------------------------|

| 14755 | 899 | 1382 | 1765 | 56 | 749 |

From the above descriptive statistics, the sample 2 looks like more of fresh and frozen food establishment, since the sample value for fresh and frozen category is larger than the mean and falls in 50-75% maxima distribution.

Sample 3

| Fresh | Milk | Grocery | Frozen | Detergents\_Paper | Delicatessen |

|--------------------------------------------------------------------|

|9790|1786| 5109 |3570| 182| 1043

Sample 3 looks like more of supermarket where you can find fresh, grocery, frozen foods along with Deli products. Since the sample value has fresh, grocery, frozen and deli more than mean value.

I tried to predict Grocery feature, the reported prediction score for the same was (R^2 = 0.681884008544).

Since the score is positive and close to one, thus the data fits well to predict grocery feature.

But since we want to separate customers spending habits, this feature seems highly unlikely to decide the customer’s spending habits.

So we need feature with say negative score, which help me in clear segmentation.

Thus, grocery feature is not necessary for identifying customers spending habits.

From the above scatter matrix, the distribution mostly looks like it is positively skewed distribution, whereas some pair of features show some degree of correlation (e.g. Grocery vs Milk or Grocery vs Detergents\_Paper).

The distribution of data is positively skewed, whereas it’s clear from the scatter matrix that there are few outliers to look after.

Whereas large number of data points are scattered near zero, that’s also the reason why distribution looks like positively skewed distribution.

Below plot shows how large number of data points are nearby zero.

\* Are there any pairs of features which exhibit some degree of correlation?

\*\* Answer \*\*

Yes, Grocery vs Milk

Grocery vs Detergents\_Paper

Detergents\_Paper vs milk

Shows some degree of correlation and below plots confirm the argument.

\* Does this confirm or deny your suspicions about the relevance of the feature you attempted to predict?

\*\* Answer \*\*

This plots and scatter matrix above thus confirms that the grocery feature is not relevant for identifying specific customers.

\* How is the data for those features distributed?

The distribution of data for grocery feature is somewhat correlated with milk, Detergents\_Paper and for other features the distribution is positively skewed.

Yes, following are the indices of data points that were considered as outliers for more than one feature.

65 - An outlier for Frozen and Fresh

66 - An outlier for Delicatessen and Fresh

75 - An outlier for Detergents\_Paper and Grocery.

128 - An outlier for Delicatessen and Fresh.

154 - An outlier for Delicatessen, Milk and Grocery.

Data points that are outliers in multiple categories warranty removal, since they are much below the tukey's IQR range and have very small magnitude and they skew the distribution.

Since the k-means algorithm has a loss function that is calculated based on distance of data points to the cluster centroid, these outliers can pull the cluster centers towards them in order to minimize the loss function.

Thus, since these outliers fall below tukeys IQR range and since k means algorithm can be affected by this outliers, we add them to the outliers list, so that they can be removed.