Customer Segmentation

# Data Exploration

## What kind of establishment (customer) could each of the three samples chosen from the customer data represent?

## Which feature did you attempt to predict? What was the reported prediction score? Is this feature relevant for identifying a specific customer?

## 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?

# Data Pre-processing

## Are there any data points considered outliers for more than one feature? Should these data points be removed from the dataset? If any data points were added to the outliers list to be removed, explain why.

## 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 in the notebook, discuss what the first four dimensions best represent in terms of customer spending.

# Clustering

## 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?

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

## Consider the total purchase cost of each product category for the representative data points found from the cluster centers, and reference the statistical description of the dataset at the beginning of the notebook. What set of establishments could each of the customer segments represent?

## For each sample point, which customer segment from the question above best represents it? Are the predictions for each sample point consistent with this?

# Conclusion

## Companies often run **A/B tests** when making small changes to their products or services. If the wholesale distributor wanted to change its delivery service from 5 days a week to 3 days a week, how would you use the structure of the data to help them decide on a group of customers to test?

## Assume the wholesale distributor wanted to predict some other feature for each customer based on the purchasing information available. How could the wholesale distributor use the structure of the data to assist a supervised learning analysis?

## How well does the clustering algorithm and number of clusters you've chosen compare to the 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?