REPORT

Question 1¶

What kind of establishment (customer) could each of the three samples you've chosen represent?

Chosen samples of wholesale customers dataset:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	9413	8259	5126	666	1795	1451
1	4591	15729	16709	33	6956	433
2	9198	27472	32034	3232	18906	5130

Customer 0 (index 5): Café (has above average milk)

Customer 1 (index 38): Grocery store (has above average milk, grocery, Detergents_paper)

Customer 2 (index 92): Restaurant (has above average milk, grocery, frozen)

Question 2

Which feature did you attempt to predict? What was the reported prediction score? Is this feature is necessary for identifying customers' spending habits?

I attempted to predict Milk. The reported prediction score was: 0.207516

Milk is a relevant feature as it can't be predicted by other features. To identify a customer habit, we would need a feature that provides more information about the customer to give us a better understanding and be able to differentiate a customer from another.

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?

(Detergents_paper, grocery), (Milk, grocery), (Milk, Detergents_paper) exhibit correlations (positive).

This confirms the suspicion that milk does seem to be a relatively important feature as it can be used to predict customer spending habits for grocery and Detergents_paper.

Most of the data points of each feature lie closer to the origin. Data is not normally distributed (looks exponential).

Question 4

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.*

Yes, the indices 65, 66, 75, 128, appeared in 2 features and 154 appeared in 3 features.

Since they are outliers in multiple features, the clusters would get skewed. I considered them as extreme outliers and removed them to reduce their effect on the data.

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.*

First component variance: 0.4470, Second component variance: 0.2746; Total: 0.7216

Total variance of first 4 components: 0.9343 (nearly all the variance is explained in the first 4 components)

The dimensions have features with highest variance are considered to be dominant:

Dimension No.	Interpretation of category spending from dimensions	Categorisation by spending habits
1	Large positive weight on Detergents_paper, relatively insignificant negative weights on Fresh and Frozen	Retail goods
2	The largest positive weight is on Fresh, there are smaller positive weights on Frozen and Delicatessen	Fast food store
3	Fresh has a high negative weight and Delicatessen has almost just as much positive weight	Market
4	Significant amount of positive weight on Frozen and a lesser magnitude of negative weight on Delicatessen	Ice cream shop

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?

K-means clustering:

- Fastest algorithm to perform clustering on large data sets

- Relatively efficient when data is well separated

GMM:

- Fastest algorithm for mixture models
- Assumes the data is normally distributed so automatically makes the clusters elliptical

KMeans performs hard clustering and thus it is faster.

I would like to use GMM for the following reasons:

- The wholesale customer data features have been transformed into normal distributions, which is ideal for GMM algorithm.
- Also, using PCA, we have maximised the variances explained by the components, and GMM takes into account variance and covariance of the features in the dataset.

Question 7¶

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

Sr. No.	n_components	Silhouette_score	aic	bic
1	2	0.41509095452	3594.3662	3631.106
2	3	0.400760073738	3562.8961	3620.0471
3	4	0.313945934136	3562.8961	3620.0471

The best score is recorded when we have 2 clusters. The aic and bic score is marginally lower for n=3, but I would like to choose a less complex model

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?*

Cluster 0 could represent a restaurant (fast food) as it has nearly average frozen items.

Cluster 1 could represent a grocery store as it has higher than average milk, grocery and Detergents_paper.

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?*

All customers have their money spent on milk, grocery, Detergents_paper, Delicatessen. I think all 3 customers belong to cluster 1 as their spending habits are consistent with cluster 1.

After running the code, all 3 samples are predicted to be in cluster 1.

Question 10

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?

Testing will be done in 2 batches. Each cluster will be divided into 2 groups: control, experiment. The control group will have delivery 5 days a week and experiment group will have to 3 days a week; 1 cluster will be tested at a time. Sample size will be 10% of the entire cluster chosen randomly, without repetition.

The A/B test would negatively affect some customers in cluster 1 which would lead to reduced buying patterns. These few customers can be identified by outlier analysis.

Question 11

Assume the wholesale distributor wanted to predict a new 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?

After performing the clustering algorithm, we get to know there are mainly 2 types of customers, those that are satisfied by deliveries 3 times a week and those who don't. Thus the new feature gained as a product of clustering would be a label for each customer depending on their choice. This has now become a supervised learning problem. So when we have a new unlabelled customer, the wholesaler can simply compare the spending habits (the 6 features) of the new customer with the existing clusters, predict what kind of delivery schedule would be suitable (which is the 7^{th} feature) and finally label them.

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 clustering done earlier was more 'pure' than this distribution.

According to this distribution, there can’t be customer segments which are purely retailers. We need a ‘fuzzy’ allocation of data points.

Yes, this classification is consistent with the segments I described earlier.