Written Solutions for Project 3

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.**

Answer:

From the above results we can see that:

Before we begin one can explain how to approach this analysis We can analyse the customers standout purchased segment as it relates to the mean and the 25% and 75% interquartile range (IQ). For example, if the customer purchased a segment that was above the mean purchases of segment in the dataset it is considered to be an outlier because it is above the 75% IQ range. We will analyze the three customers below.

From the above table We can analyze all three customers:

Firstly, Analyzing Customer 0. From the table we can see that this customer stands out purchasing the segment Fresh. If we analyze a bit more in detail and make a comparison with the Dataset we can see that this customer purchases Fresh above the average , as well as above the percentage quartiles 25, 50 and 75 respectively. So we can successfully say that this consumer greatly prefers Fresh items so we can assume that this customer is a Vegetarian and Shops at an Organic Market.

Secondly, Analyzing Customer 1. From the table we can see that this customer stands out purchasing the segment Grocery. If we analyze a bit more in detail and make a comparison with the Dataset we can see that this customer purchases Grocery above the average, as well as above the Percentage Quartiles of 25, 50 and 75 respectively. So we can successfully say that this consumer greatly prefers Grocery items. So we can assume that this customer shops at a supermarket.

Third and Final, Analyzing Customer 2. From the table we can see that this customer stands out purchasing the Segment Grocery. Similar to Customer 1 However this customer purchases far more than the Average and Percentage Quartiles of 25, 50, and 75 of Grocery proporionately greater than Customer 1.

Question 2

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

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.

Answer:

The feature we attempted to predict was "Milk". The reported prediction score is an average score of 0.1734

From the analysis one can say that Milk is an important segment. Why so? Because when we drop the Milk Segment the average score dropped tremendously. One can confidentially say Milk is a relevant Customer Segment. From the model we ran a small experimental test to derive the importance of milk. How did we find this out? We did this by dropping another segment such as Grocery. When we ran this we got an average score of 0.963. What this showed us is that Grocery was not as relevant in comparison to the relevance of the segment Milk. Where as when dropping Milk we got an average of 0.1734 So one can clearly see the relevance of MIlk in the Dataset and the outcome of it, it were to be dropped it affects the average value greatly.

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?
Answer:

From the above diagram there are few pairs of features in which that we can state exhibit some degree of correlation but only will be specified. The first is the correlation between Milk and Fresh. From assessing the diagram we can determine hypothetically that there is a negative correlation between the two.

The second is the correlation between Grocery and Milk. From assessing the diagram we can determine hypothetically that there is a positive correlation between the two.

However much we analyze the graphs one can confidentially state that there are abnormalities with the data diagrams. We can look and see that the data is skewed which means in simple terms that the data is biased. We can also see that the data is not Normally distributed.

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.

Answer:

One can say, yes there are data points considered to be ouliers for more than one feature, we can see confidentially say this by looking at the results generated in the table above

If we normalize the data, the outliers won't necessarily need to be removed, given that all the datapoints will be centered with a mean of 0 and will all fall within a set range with a normal distribution.

Data points were added to the outlier list to be removed because points which are outliers can skew the data by not creating Normal Distribution. In simple terms in Normal Distribution everything should be on the same scale if not this causes the data to be Biased. In conclusion one or more outliers create distortion in the data so hence we remove them because generally it will give a better understanding of the customer base.

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.

Answer:

The total explained variance for the first and second principal components is by finding the sum of the two explained variance values which are:(0.4424 + 0.2766).

The sum of these two numbers gives us the number (0.719) which is the total of the first and second principal component.

The same method how we found out the total variance of the first and second principal components is also the same method that we will calculate the first four principal components. Which is simply finding the sum of:(0.4424 + 0.2766 + 0.1162 +0.0962). Which amounts too the value (0.9314).

From the vizualization above one can analyze and discuss about the First Four Dimensions determining customer spending:

From the First Graphical Dimension, Dimension 1. We can see a correlation that there is a major increase in spending on Detergents_Paper which significantly stands out over the others as the most positive increase, followed by a fairly positive increase in Grocery, closely followed by Milk and then somewhat of a satisfactory increase in Delicatessen. However we can see negative decreases in spending on segments such as Milk and the most negative decrease on Frozen.

From the Second Graphical Dimension, Dimension 2. From this correlation we do not see a significant standout 'towering' as we saw in Dimension 1. However the most positive increase in spending on a segment in this case is Frozen. This is followed by Fresh which can be concerned second which isn't too far and Delicatessen in third which has a fairly positive increase, Milk and Grocery have positive but satisfactory increases. There is only negative decrease in this Dimension and that is the only decrease of the segment Detergents_Paper which is not a major decrease in comparison to other segments like we saw in Dimension1.

Thirdly the Third Graphical Dimension, Dimension 3. From this correlation we can clearly see two "towering" opposites of one being positive and negative. The positive increase in spending in Fresh which uniquely stands out over the fairly positive increase in Detergents_Paper and Grocery just barely being a positive increase. However Unlike Dimension 1 and Dimension 2 which stood out proportionately greater on the positive side, We can see that Dimension 3 stands out proportionately greater on the negative sides with a large decrease in Milk followed by larger Decrease in Frozen and the largest Decrease in Delicatessen. But we can make another analysis and see that the positive increase in Fresh is proportionately greater than the negative decrease in Delicatessen.

Fourth and Final the Fourth Graphical Dimension, Dimension 4. From this correlation, we can see quite a bit of negative decreases even more than Dimension 3. We can first by looking at the only 2 positives. The major increase we can see "towering" is the Frozen segment and then further down not competing with Frozen is Detergents_Paper which stands fairly positive. But as we said earlier our main center of attraction to focus on are the negative decreases starting with the Delicatessen which stands out as the most negative decrease followed by Fresh, Milk and Grocery in this order. We can definitely see that this Dimension is far more proportionately negative then it is positive in comparison to the other 3 Dimensions.

In general we get an idea how customer spending correlates on one segment and greatly affects the other segment within the same Dimension.

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?

Answer:

The Advantages of Using K-means algorithm are:

-It is very easy to implement

-It is computationally very efficient compared to other cluster algorithms

-It is very effective in identifying clusters of spherical shape

The Advantages of using Gaussian Mixture Model algorithm:

-The Gaussian Process relys more on the family of functions rather than parameters. -(Speed) It is the fastest algorithm for learning mixture models

-(Agnostic) AS the algorithm maximizes only the likelihood, it will not bias the means towards zero, or bias the cluster sizes to have specific structures that might or might not apply.

The main difference between the two is that K means is a form of hard clustering and Gaussian is a form of soft clustering. We can remind ourselves that soft clustering is where one sample is assigned to one or more clusters and that hard clustering is where one dataset set is assigned to exactly one cluster

We can use both algorithims, However one would prefer K-Means because of its simplicity.

Question 7

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

In this case the algorithm was run several times using the numbers 1 through to 10 in the n_clusterers.

When n_clusterer = 2, The silhouette_score = 0.3715

When n_clusterer = 3, The silhouette_score = 0.3703

When n_clusterer = 4, The silhouette_score = 0.3783

When n_clusterer = 5, The silhouette_score = 0.3560

When n_clusterer = 6, The silhouette_score = 0.3582

When n_clusterer = 7, The silhouette_score = 0.3600

When n_clusterer = 8, The silhouette_score = 0.3733

When n_clusterer = 9, The silhouette_score = 0.3629

When n_clusterer = 10,The silhouette_score = 0.3591

From the several outcomes generated one can confidentially say that the clusters are most accurately defined when the n_clusteres = 4 which generates the value of 0.3783

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'.

Answer:

Each cluster represents a certain customer on average. Basically the cluster algorithm shows a good representation of each customer and what category they are primarily spending on the most. To conclude we can succesfully say, Based on the customers spending we can categorize and determine what type of customers they really are.

Such as, Segment 0 which shows us that there is a higher preference towards the customer spending on Fresh in comparison to the other categories. However we can make another small analyses that customers who purchased Fresh also purchased quite a bit of Grocery and minimal amounts of Milk, Frozen and Delicatessen .We can also compare this data to the customer dataset on how much they spend on the categories. Customer(s) spends above the 1st Quartiles but still consumes Fresh far below the 2nd and 3rd Quartile. Although he/she spends quite a bit on Grocery customer(s) still purchases below the average in comparison to the mean purchase of Fresh. Basically, customer(s) purchases under the average of all the categories within this Segment.

From the Analyses of Segment 1 we can see that there is a higher preference towards the customer spending on Grocery in comparison to the other categories. We can also see that customers also spend fairly on Milk, Fresh and Detergents_Paper in that order but not as much on Frozen and least on Delicatessen. We can compare this data to the customer dataset on how much they spend on all categories. Customer(s) spends far above average on Grocery and Milk and Delicatessen. But Spends below average on Fresh, Frozen 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?

Run the code block below to find which cluster each sample point is predicted to be.

Answer:
Sample 0,1 and 2 were predicted to fall in cluster 1. Fresh and Frozen were below average while milk and groceries were above average.

Yes, one can successfully say that the predictions for each sample point are consistent with this. We can say this because earlier we found out that sample_preds prints an array (1, 1, 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?
Hint: Would such a change in the delivery service affect all customers equally? How could the distributor identify who it affects the most?

Answer:
To perform A/B on the data, test the change on delivery to one group (experimental group) and see how they respond to it. If they respond well, apply the delivery service to the 'control group'. For instance, if the distributor wants to roll out a bulk delivery service to customers that largely purchases a customer segment, then run the service on an experimental group, or a subset of the dataset or a sample of customers from a particular cluster. Then if the experiment goes well, test the service on the control group or the entire cluster. For example if we plan to change the delivery service for customers who purchase Grocery, Wholesale Distributors main clientel are the SuperMarkets. The Distributor can try it on a few supermarkets first and analyze the results to determine whether the service receives positve feedback or negative feedback. If the results are positve and the customer which is the Supermarket is satsified then the Distributor can go ahead and implement this with all of its customers.

Question 11

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?

Answer:
Given that we classified the customers into clusters or segments, we now have a new feature to add to the dataset. So adding to the previous known features, we have added the cluster label classifications as an additional feature. This updated dataset can be used towards learning some new feature using supervised learning. For example, we could train a supervised learning model to predict whether a particular segment responds positively or negatively to a new bulk delivery service given the additional cluster information.

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?

Answer:

From the graph we cannot correctly classify samples within the dataset because we see samples overlapping, For example we see some red data points in areas alongside with green points and vice-versa. Unlike the PCA

Reduced_Data which can be classified this cannot be classified. One can only see one pure cluster which is Data point 2. Apart from that all data clusters look mixed up or chaotic.

Finally, we can say that these classifications are not consistent because the clusters cannot be classified.