Customer Segmentation In wholeSale Distribution Using PCA(Dimensionality Reductionality) and KMeans Clustering

UML Project
Name:Arisha Akhtar
Roll No. A041
SAP-id:86092300042

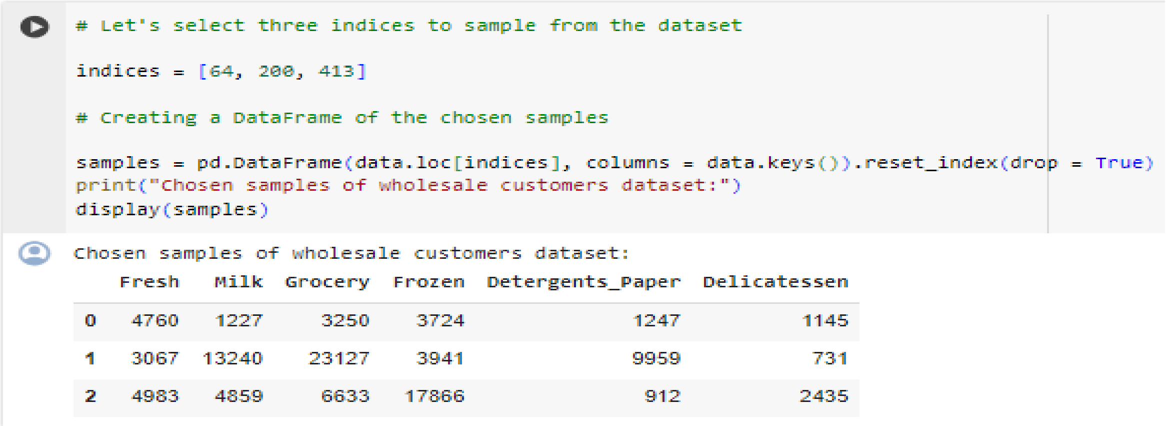
Problem Statement: Develop a customer segmentation strategy for a wholesale distributor to categorize clients based on the types of products they purchase, frequency of orders, and volume, in order to optimize inventory management and procurement strategies.

Goal: One goal of this project is to best describe the variation in the different types of customers that a wholesale distributor interacts with. Doing so would equip the distributor with insight into how to best structure their delivery service to meet the needs of each customer.

Dataset:A dataset containing data on various customers' annual spending amounts (reported in *monetary units*) of diverse product categories for internal structure.
The dataset for this project can be found on the UCI Machine Learning Repository. For the purposes of this project, the features `'Channel'` and `'Region'` is excluded in the analysis with focus instead on the six product categories recorded for customers.the dataset is composed of six important product categories: **'Fresh'**, **'Milk'**, **'Grocery'**, **'Frozen'**, **'Detergents_Paper'**, and **'Delicatessen'**.

Selecting Samples

To get a better understanding of the customers and how their data will transform through the analysis, let's select a few sample data points and explore them in more detail.



```
# Let's select three indices to sample from the dataset

indices = [64, 200, 413]

# Creating a DataFrame of the chosen samples

samples = pd.DataFrame(data.loc[indices], columns = data.keys()).reset_index(drop = True)
print("Chosen samples of wholesale customers dataset:")
display(samples)
```

```
Chosen samples of wholesale customers dataset:
```

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicatessen
0	4760	1227	3250	3724	1247	1145
1	3067	13240	23127	3941	9959	731
2	4983	4859	6633	17866	912	2435

Assumptions on Establishments

Considering the total purchase cost of each product category and by comparing the same with the mean values of each category, we can assume the following.

- The sample customer with index 0 looks like a Cafe as the values for all products (i.e. Fresh, Milk, Grocery and Frozen) are lower and almost equal to one another.
- The sample customer with index 1 may represent a Super-market as the values for product Milk, Grocery and Detergents_paper is high and only Super-market keep these items at high quantity.
- The sample customer with index 2 looks like a Restaurant as the values for Frozen product is comparatively higher than other products. A restaurant will surely try to serve fresh the food items to customer hence more quantity of frozen products.

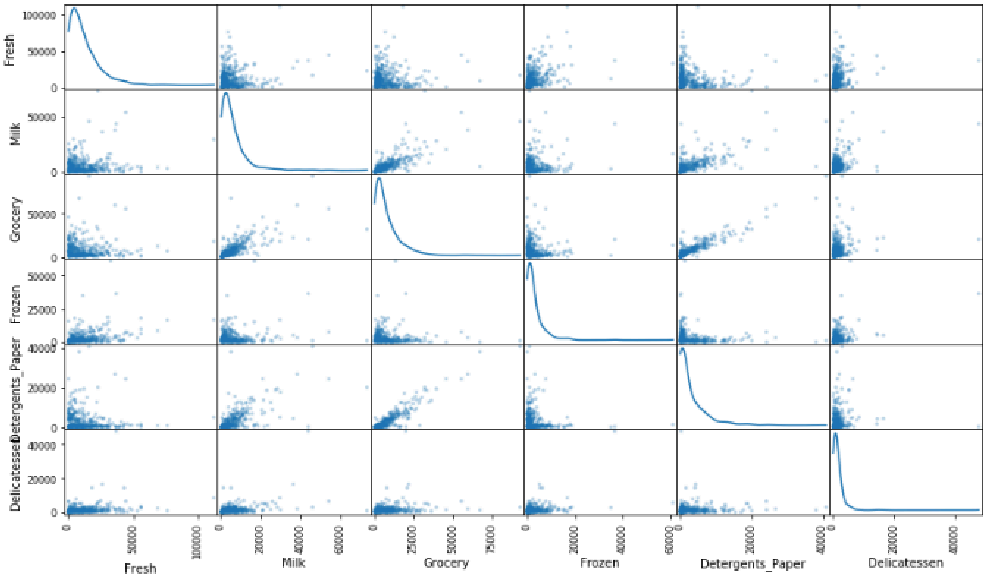
Visualizing Feature Distributions

To get a better understanding of the dataset, we can construct a scatter matrix of each of the six product features present in the data. If feature that we attempted to predict above is relevant for identifying a specific customer, then the scatter matrix below may not show any correlation between that feature and the others. Conversely, if we believe that feature is not relevant for identifying a specific customer, the scatter matrix might show a correlation between that feature and another feature in the data. Let's run the code block below to produce a scatter matrix.

```
# Producing a scatter matrix for each pair of features in the data

pd.scatter_matrix(data, alpha = 0.3, figsize = (14,8), diagonal = 'kde');
```

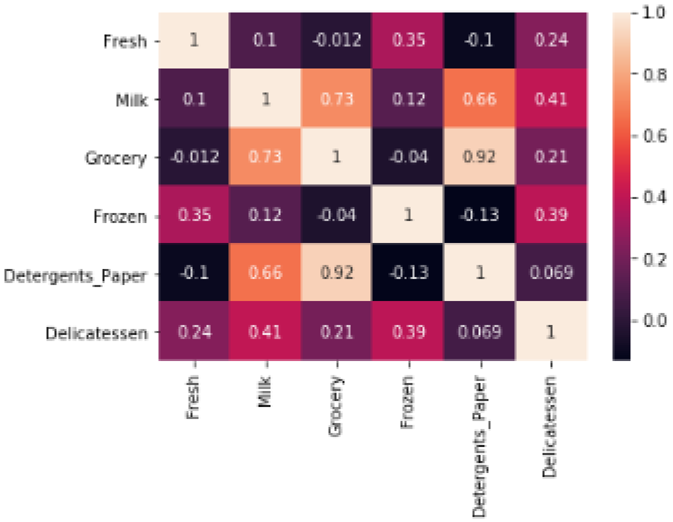
```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:3: FutureWarning: pandas.scatter_matrix is depr
  This is separate from the ipykernel package so we can avoid doing imports until
```



```
# Let's use seaborn library to plot the feature correlation

import seaborn as sns
sns.heatmap(data.corr(), annot=True)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x178f9d82438>
```



Observations

1. By looking at scatter matrix, we can say that the data is not normaly distributed due to many outliers within the dataset.

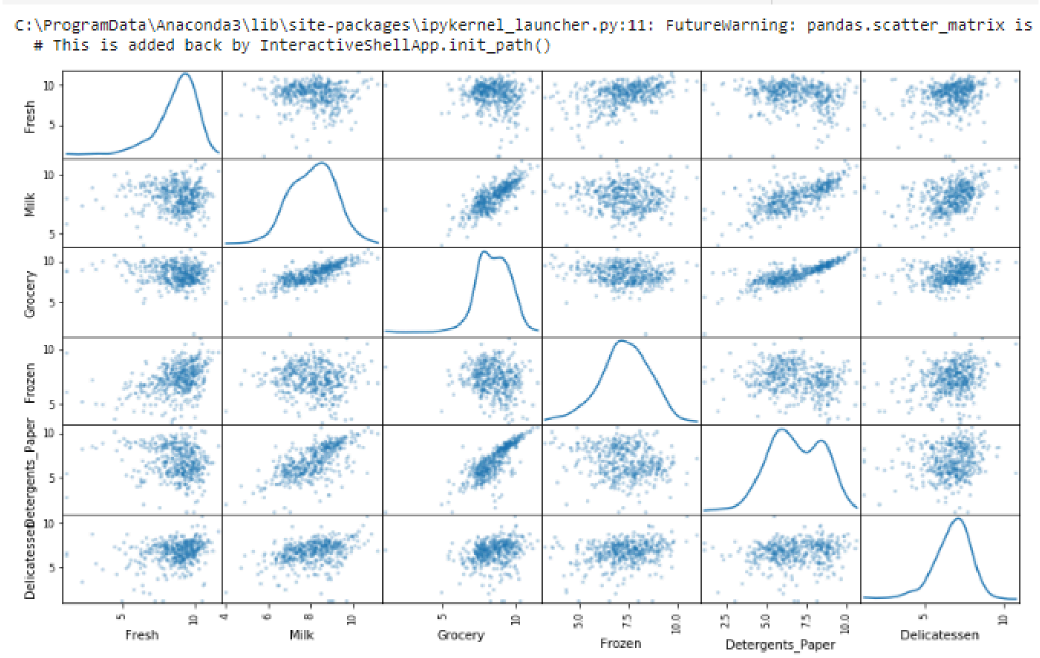
2. Large number of the data points are mostly skewed near to 0.

3. Milk, Detergent_Paper & Grocery features shows that there is high degree of correlation in relevance with each other. These features may not be important for identifying customers' spending habits.

4. This confirms our observation made about the relevance of the feature we predicted previously.

Feature Scaling

If data is not normally distributed, especially if the mean and median vary significantly (indicating a large skew), it is most often appropriate to apply a non-linear scaling — particularly for financial data. One way to achieve this scaling is by using a Box-Cox test, which calculates the best power transformation of the data that reduces skewness. A simpler approach which can work in most cases would be applying the natural logarithm.



```
C:\ProgramData\Anaconda3\lib\site-packages\ipykernel_launcher.py:11: FutureWarning: pandas.scatter_matrix is
  # This is added back by InteractiveShellApp.init_path()
```

Observations

After applying a natural logarithm scaling to the data, the distribution of each feature appeared much more normal.

Outlier Detection

Detecting outliers in the data is extremely important in the data preprocessing step of any analysis. The presence of outliers can often skew results which take into consideration these data points. There are many "rules of thumb" for what constitutes an outlier in a dataset. Here, we will use Tukey's Method for identfying outliers: An *outlier step* is calculated as 1.5 times the interquartile range (IQR). A data point with a feature that is beyond an outlier step (outside of the IQR) for that feature is considered abnormal.

Observations

, there are 5 data points considered as outliers for more than one feature. Hence, instead of removing all outliers we must remove the above 5 data points from the dataset

Feature Transformation:I have used principal component analysis (PCA) to draw conclusions about the underlying structure of the wholesale customer data. Since using PCA on a dataset calculates the dimensions which best maximize variance, we will find which compound combinations of features best describe customers.

Model Used:
PCA(Principal Component Analysis):PCA is used to discover which dimensions about the data best maximize the variance of features involved. In addition to finding these dimensions, PCA will also report the *explained variance ratio* of each dimension — how much variance within the data is explained by that dimension alone.

Dimensionality Reduction

When using principal coacmysis, one of the main goal is to reduce the dimensionality of the data — in effect, reponent analducing the complexity of the problem. Dimensionality reduction comes at a cost: Fewer dimensions used implies less of the total variance in the data is being explained. Because of this, the *cumulative explained variance ratio* is extremely important for knowing how many dimensions are necessary for the problem. Additionally, if a signifiant amount of variance is explained by only two or three dimensions, the reduced data can be visualized afterwards.

K-means Clustering

Advantages:

1. With more number of features, K-means can be computationally faster if K is small.

2. K-Means could result in tighter clusters than hierarchical clustering.

Gaussian Mixture Model Clustering

Advantages:

1. Works with different distributions of the data. It can fit more complex cluster shapes since each mixture component can freely fit its covariance matrix.

2. If you think that your model is having some hidden, not observable parameters, then you should use GMM. This is because, this algorithm is assigning a probability to each point to belong to certain cluster, instead of assigning a flag that the point belongs to certain cluster as in the classical k-Means.

3. Capable of "soft" classification i.e. each data point is assigned a probability for each cluster, indicating how likely it belongs to the cluster. K-means only provides hard assignments, i.e. it chooses a single cluster for each data point.

Choosing Algorithm

We will choose Gaussian Mixture Model because of its ability to apply "soft" classification.

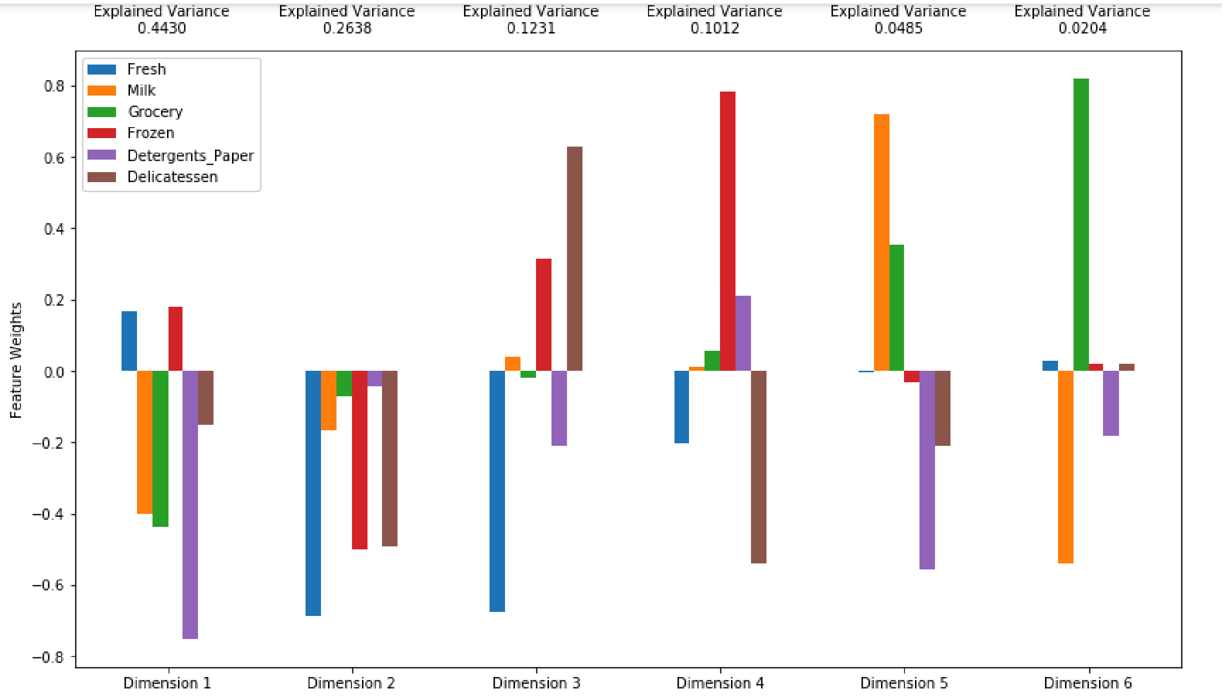
Why?

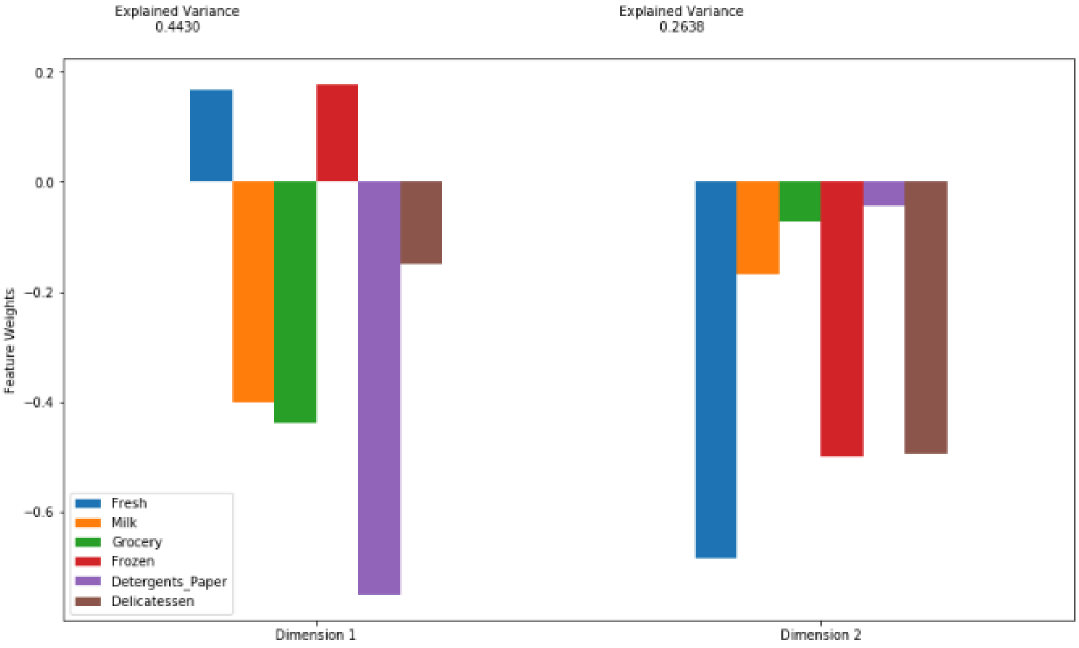
Soft Classification: GMM assigns probabilities to each data point belonging to each cluster rather than hard assigning it to just one cluster. This allows for more nuanced analysis, especially in scenarios where data points may belong to multiple clusters simultaneously.

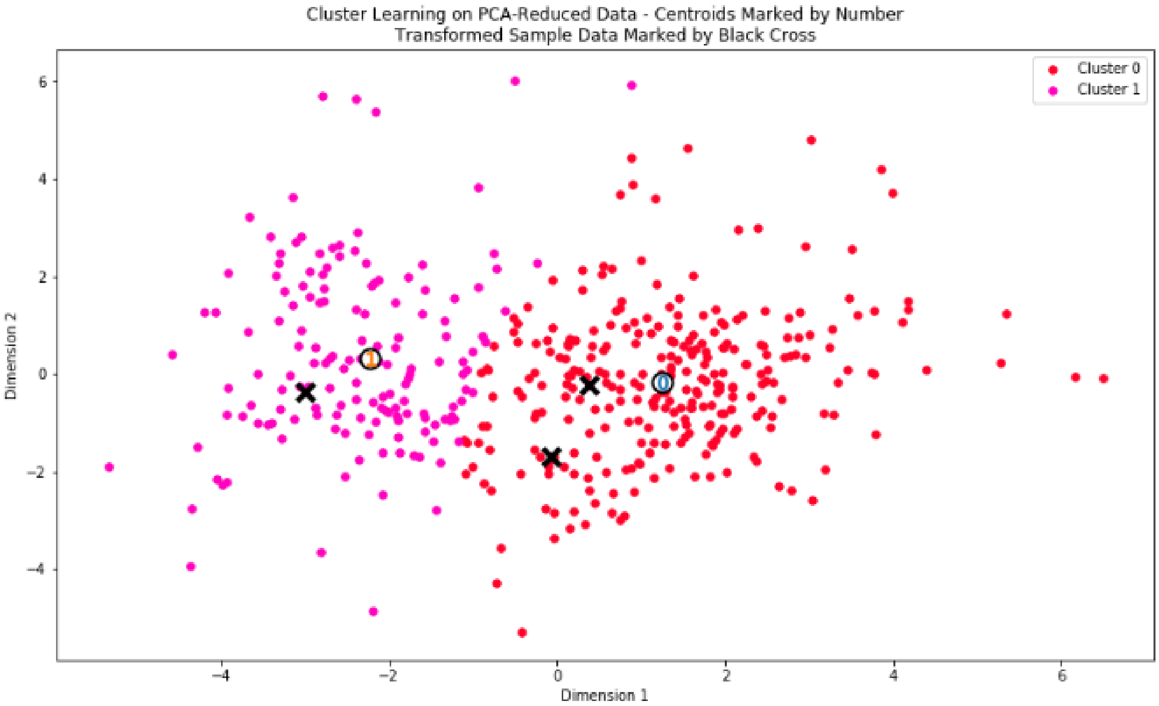
Flexibility in Cluster Shapes: Unlike some other clustering algorithms (e.g., K-means), GMM does not assume clusters to be of a specific shape. It models clusters as ellipsoids, which allows it to capture complex cluster structures in the data.

Visualizations:

PCA plot:

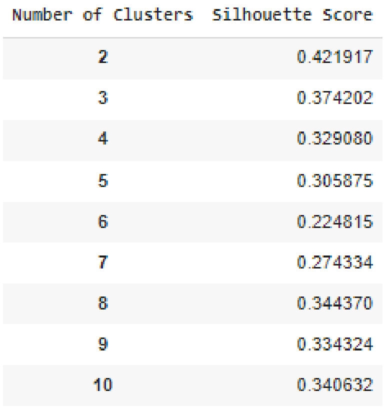






Performance Evaluation:

Silhouette score is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). A silhouette score ranges from -1 to 1, where a score closer to 1 indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters.



Number of Clusters	Silhouette Score
2	0.421917
3	0.374202
4	0.329080
5	0.305875
6	0.224815
7	0.274334
8	0.344370
9	0.334324
10	0.340632

Observations

1. Silhouette score for different number of clusters is as shown in above table.

2. Number of clusters = 2 has the best silhouette score which is 0.421917.

3.The optimal number of clusters is typically chosen based on the highest silhouette score. In this case I have observed that the number of clusters = 2 has the highest silhouette score of 0.421917. This suggests that dividing the data into 2 clusters provides the best balance between cohesion within clusters and separation between clusters.
4.Interpretation: The silhouette score of 0.421917 for 2 clusters indicates that the objects within each cluster are relatively well matched to their own clusters and poorly matched to other clusters. This suggests that the clustering algorithm has successfully separated the data into distinct groups.

Conclusion:
1.Optimal Cluster Number: The analysis suggests that dividing the data into 2 clusters yields the highest silhouette score of 0.421917, indicating that the data points are relatively well-clustered and distinct from each other. This suggests that the dataset exhibits strong patterns that can be effectively captured by dividing it into two distinct groups.

2.Cluster Interpretation: The two clusters identified likely represent meaningful segments within the data. Further analysis can be conducted to interpret the characteristics of each cluster and understand the distinguishing features that differentiate them. This could involve examining the cluster centroids, exploring the distribution of features within each cluster, and conducting statistical tests or visualizations to identify patterns and trends.

3.Sample Point 0 (Cafe):

- The features for sample point 0 show relatively low values for most categories, particularly for Grocery and Detergents_Paper.
- The higher values in categories like Fresh and Frozen suggest that this establishment likely focuses on serving fresh and/or frozen food items, which aligns with the characteristics of a cafe.
- The guess for sample point 0 being a cafe seems consistent with its feature values.

4.Sample Points 1 & 2 (Supermarket):

- The features for sample points 1 and 2 exhibit higher values across multiple categories, particularly in Grocery and Detergents_Paper.
- These higher values in Grocery and Detergents_Paper suggest that these establishments likely sell a variety of household and grocery items, which is characteristic of a supermarket.
- The guesses for sample points 1 and 2 being supermarkets seem reasonable given their feature values, which indicate a broader range of products compared to a cafe.

Future Recommendation:

Refine Clustering Model:

- Conduct further analysis to refine the clustering model by experimenting with different algorithms, hyperparameters, and preprocessing techniques.
- Consider incorporating additional features or feature engineering methods to better capture the underlying patterns in the data.

Collect Additional Data:

- Collecting additional data or expanding the dataset to include more samples from diverse establishments can improve the robustness and generalizability of the clustering model.
- Incorporate feedback and insights from domain experts to ensure that the dataset captures relevant information for the task at hand.

Enhance Feature Interpretation:

- Explore methods for enhancing the interpretation of features and their relevance to different types of establishments.
- Conduct feature importance analysis or visualization techniques to identify the most discriminative features for each cluster.

Colab File

LINK:https://colab.research.google.com/drive/19hhAdg3a6ZI9vfiOnFCitScch31JwzuO#scrollTo=WbTQs5E6-Qes