

**EEC 525 / CIS 660 DATA MINING**

**LAB-04**

**Clustering with K-Mean and DBSCAN**

**Submitted By**

**Dinky Mishra**

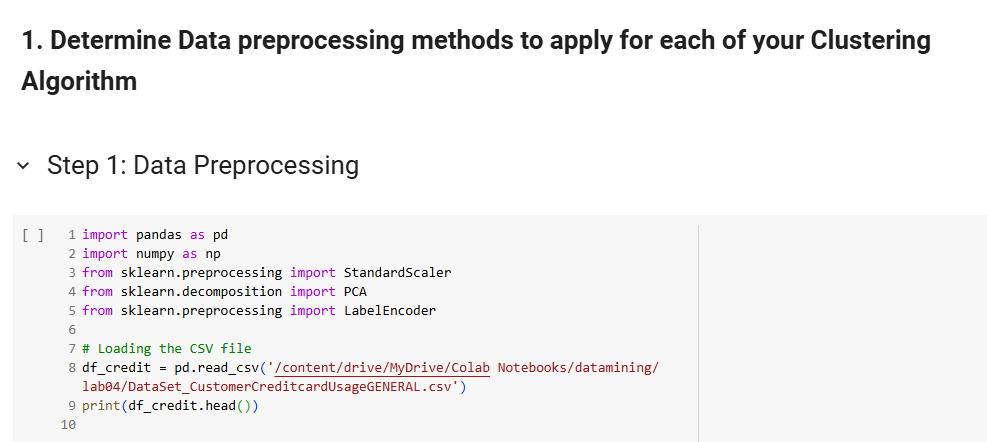
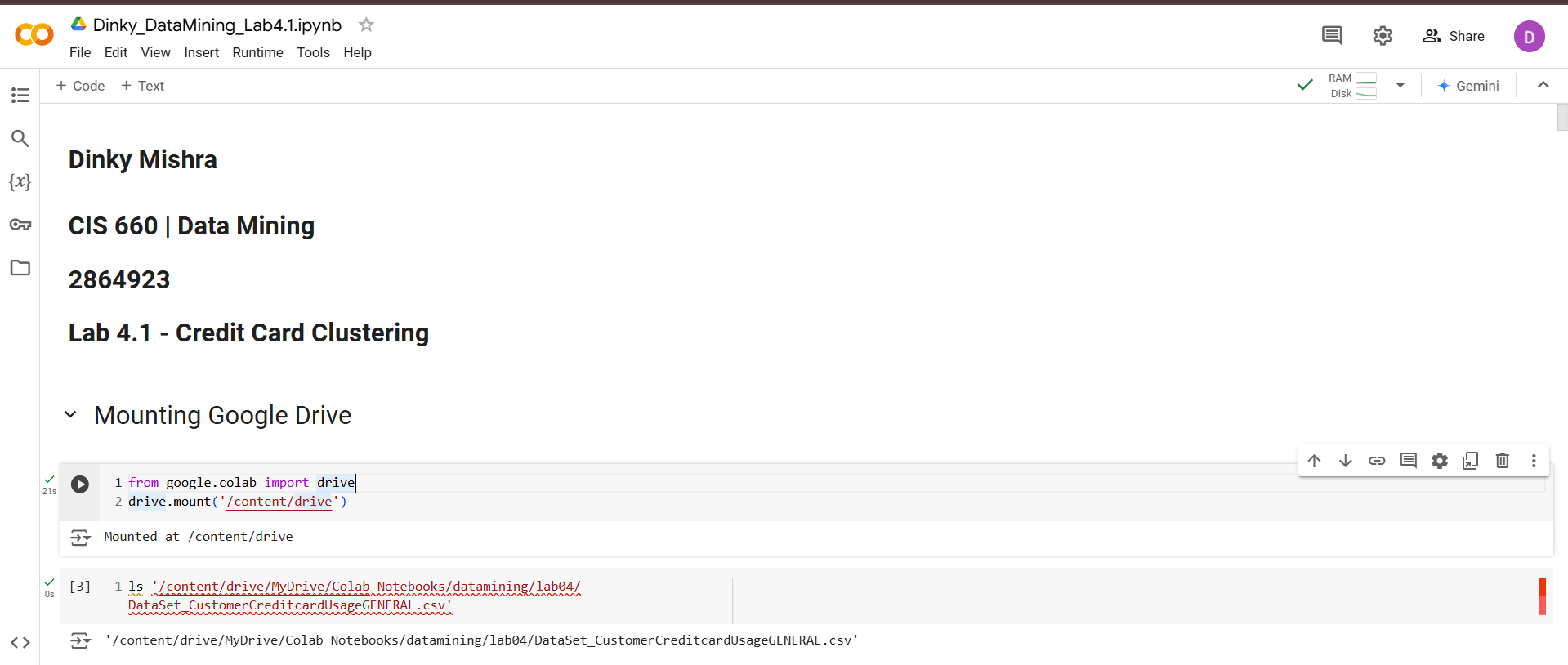
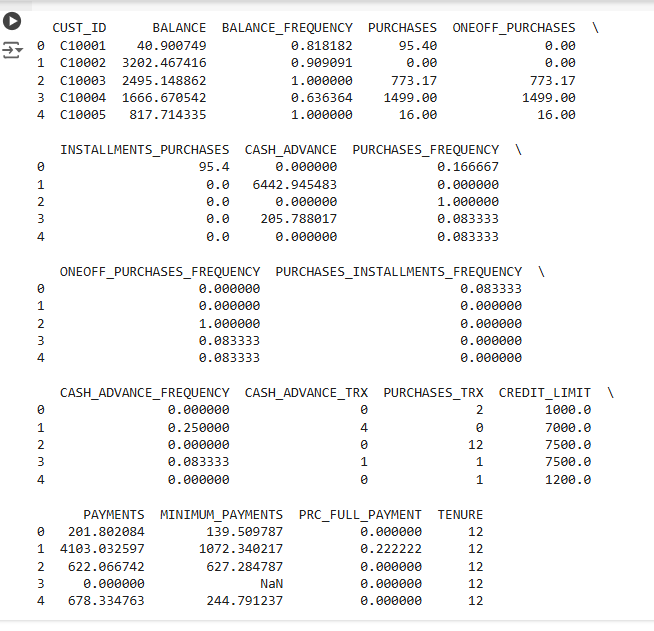
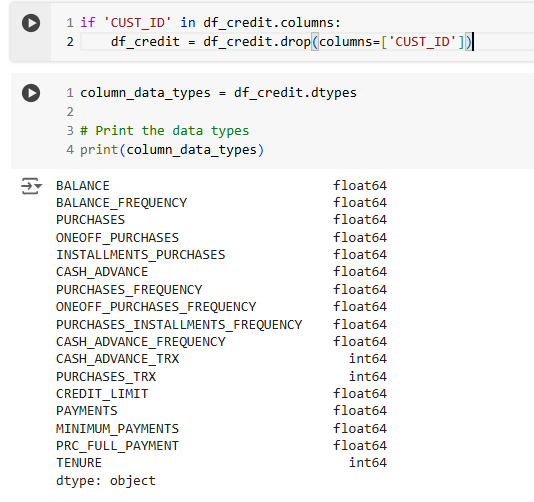
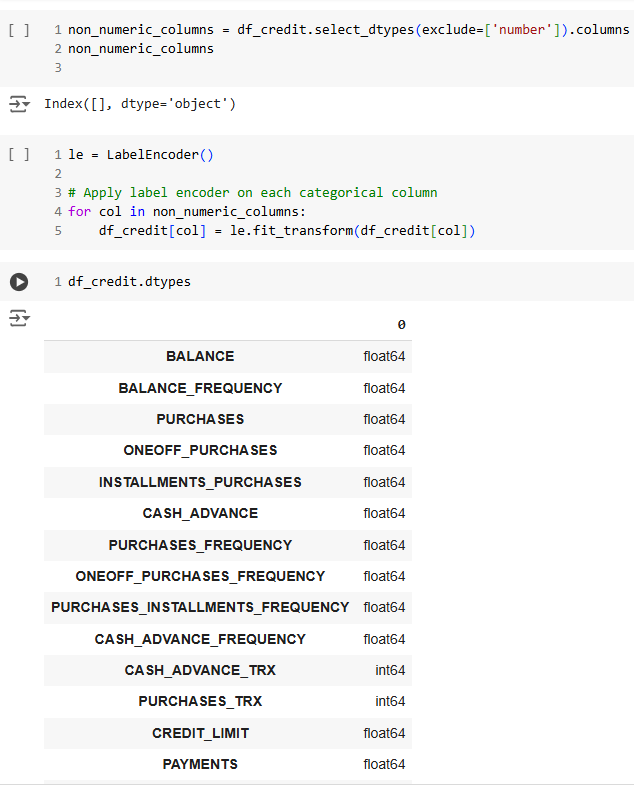
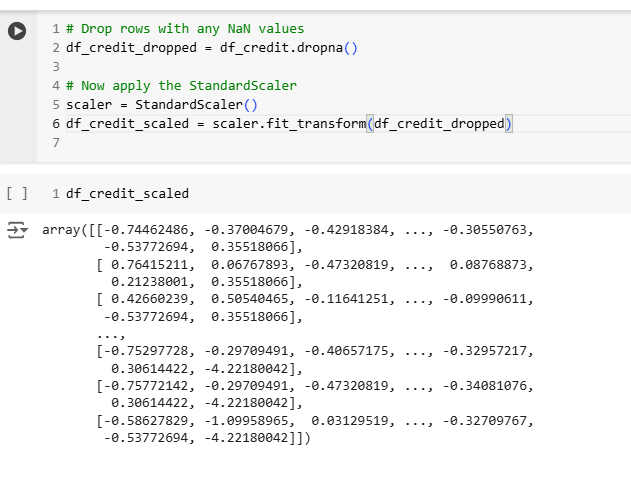
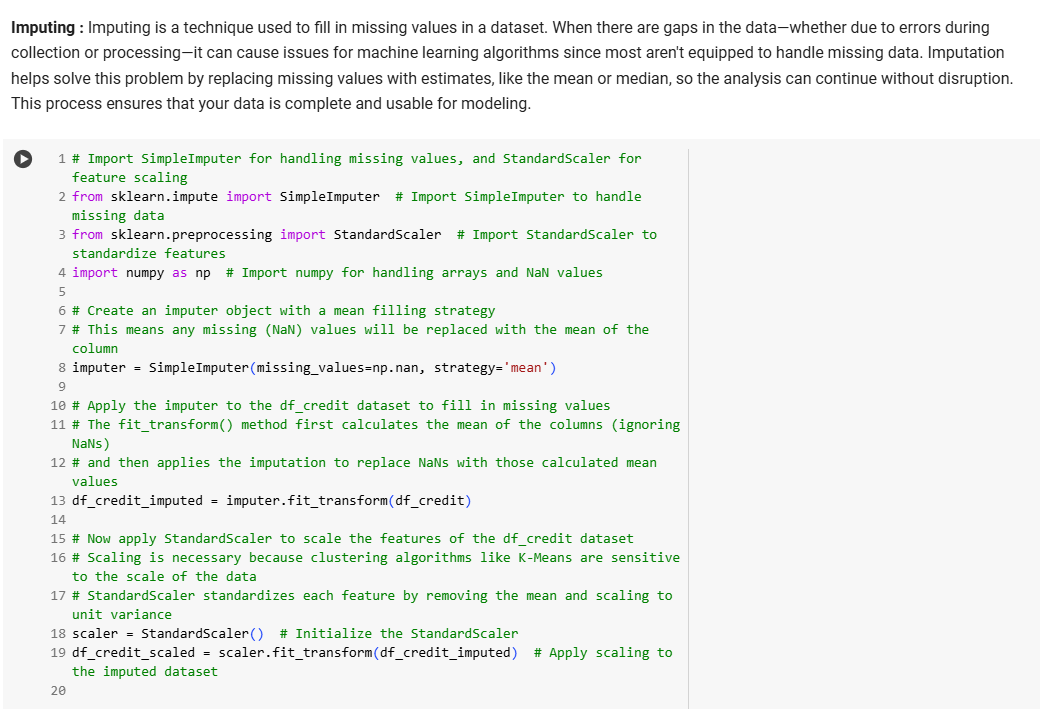
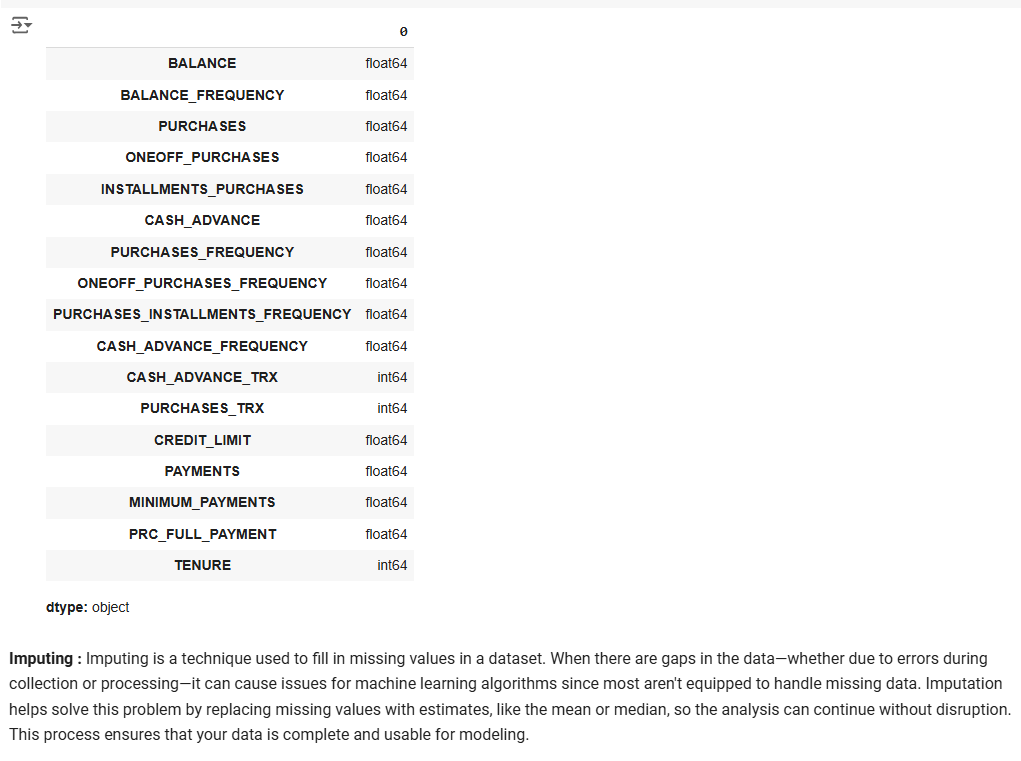
**CSU ID: 2864923**

**Table of contents**

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Content** | **Page No** |
| **4.1** | **CUSTOMER CREDIT CARD USAGE ANALYSIS USING CLUSTERING** |  |
|  | |  | | --- | | **1.Determine Data Preprocessing** | | 3 |
|  | |  | | --- | | **2.Experiment Design** | |  |
|  | |  | | --- | | 2.1 Parameter Setting for your Clustering Methods | | 7 |
|  | |  | | --- | | 2.2 Best K based on one of the method [EXTRACREDIT] | | 8 |
|  | |  | | --- | | **3. Validate your Clustering Result** | |  |
|  | |  | | --- | | 3.1 Clustering for K-Means Method | | 10 |
|  | |  | | --- | | 3.2 Clustering for DBSCAN Method | | 11 |
|  | |  | | --- | | **4. Visualize using PCA** | |  |
|  | |  | | --- | | 4.1 Visualization for K-Means | | 13 |
|  | 4.2 Visualization for DBSCAN | 13 |
|  | |  | | --- | | **5. Discussion** | | 14 |
| **4.2** | **NIJ CRIME LOCATION DATA ANALYSIS USING CLUSTERING** |  |
|  | |  | | --- | | **1.Determine Data Preprocessing** | | 15 |
|  | |  | | --- | | **2.Experiment Design** | |  |
|  | |  | | --- | | 2.1 Parameter Setting for your Clustering Methods | | 21 |
|  | |  | | --- | | 2.2 Best K based on one of the method [EXTRACREDIT] | | 22 |
|  | |  | | --- | | **3. Validate your Clustering Result** | |  |
|  | |  | | --- | | 3.1 Clustering for K-Means Method | | 23 |
|  | |  | | --- | | 3.2 Clustering for DBSCAN Method | | 24 |
|  | |  | | --- | | **4. Visualize using PCA** | |  |
|  | |  | | --- | | 4.1 Visualization for K-Means | | 27 |
|  | |  | | --- | | **5. Discussion** | | 28 |

**LAB 4.1 using dataset Customer Credit card Usage GENERAL**

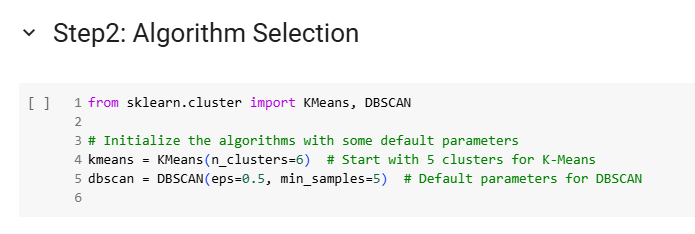
**1. Determine Data preprocessing methods to apply for each of your Clustering Algorithm**

**2. Design your Clustering Experiment.**

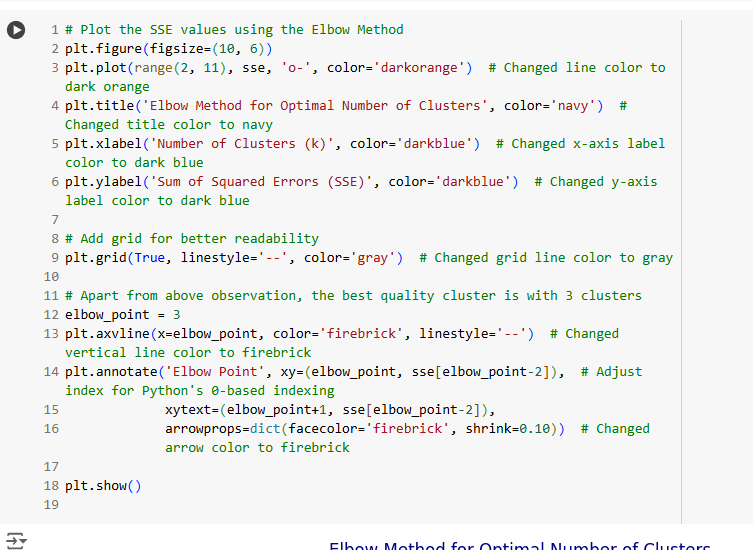
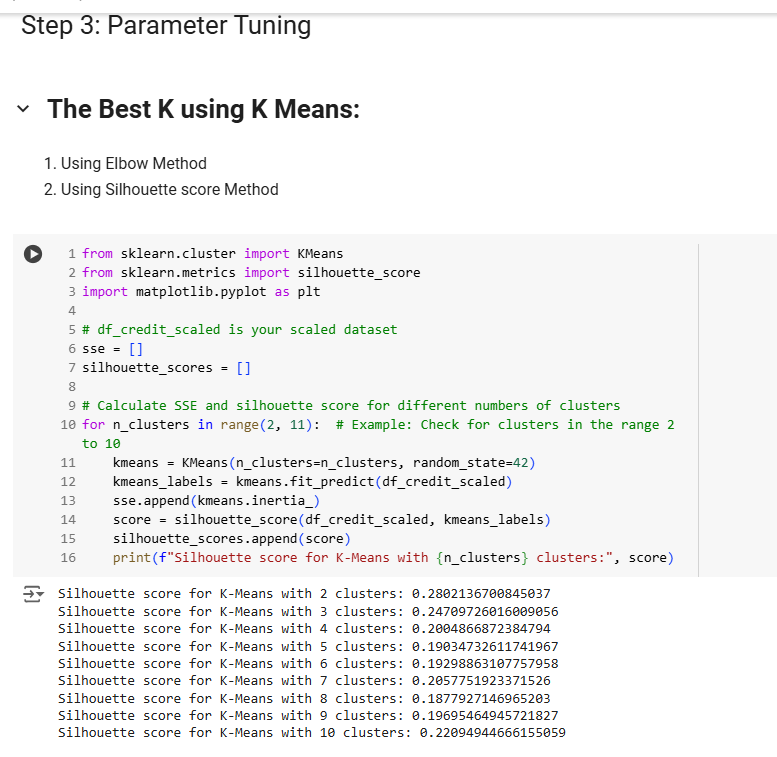
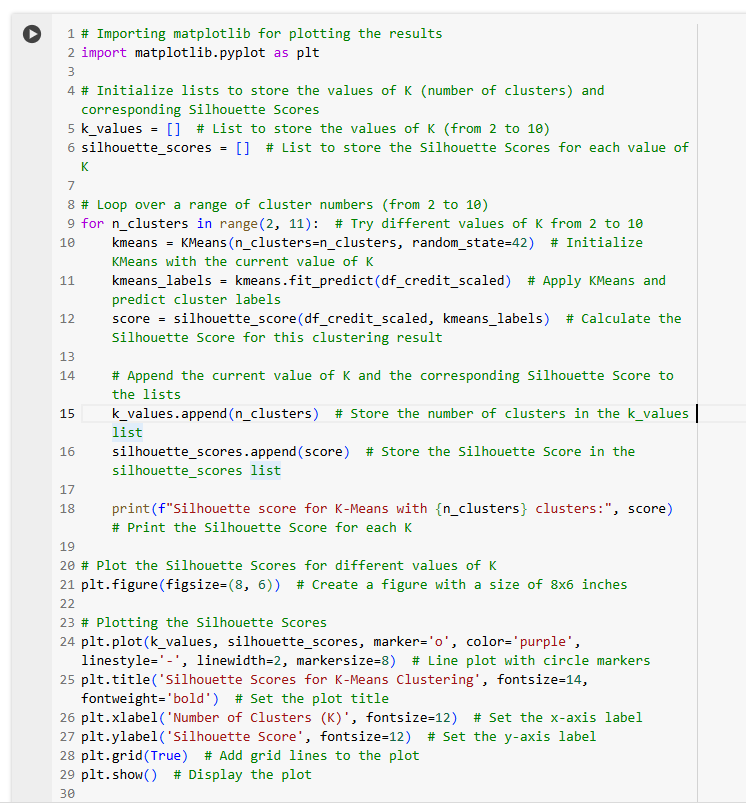
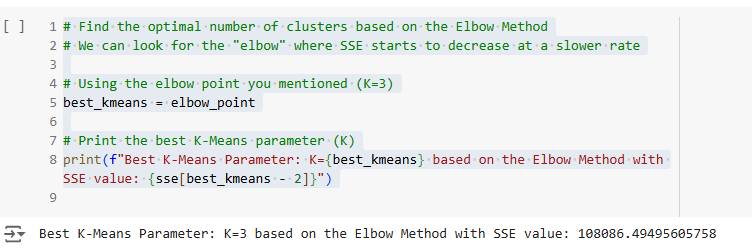
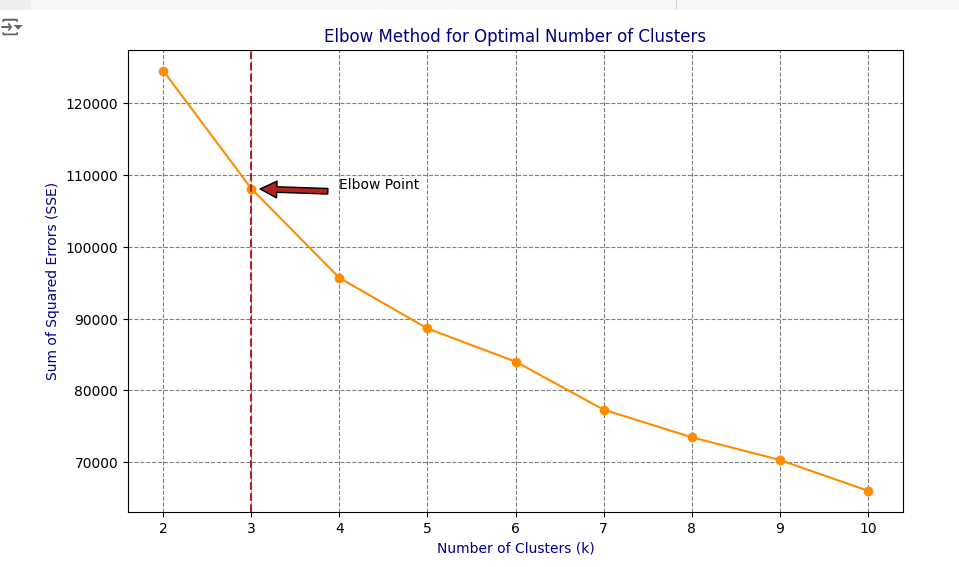
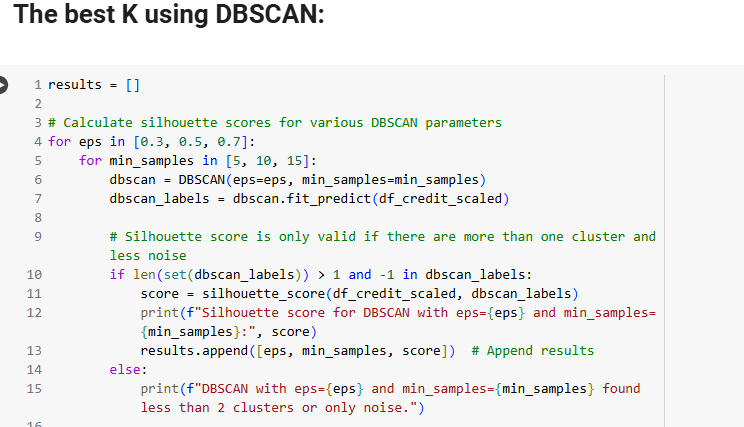
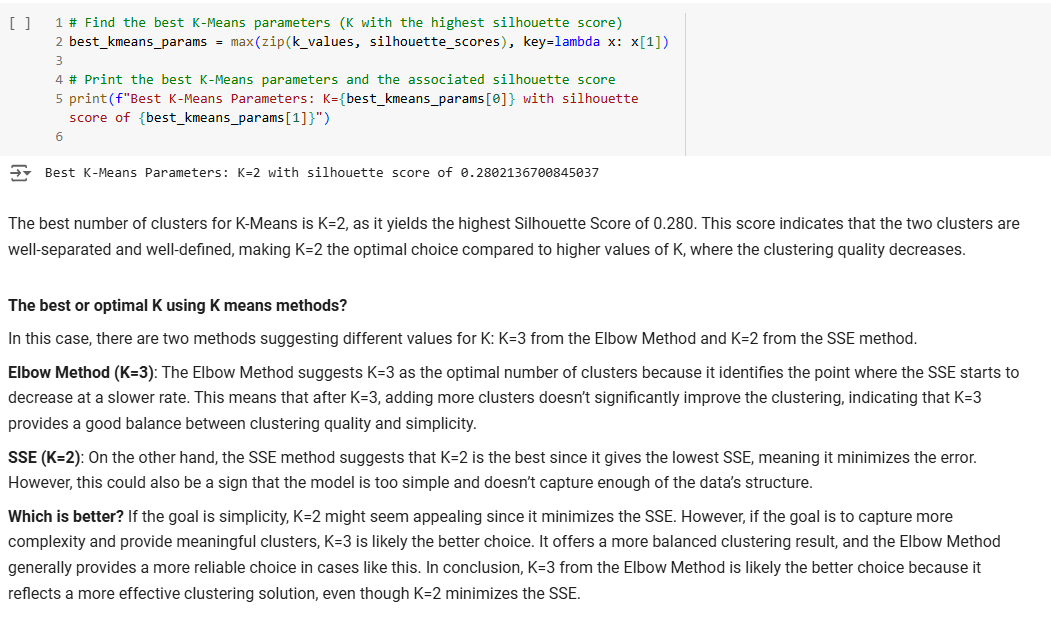
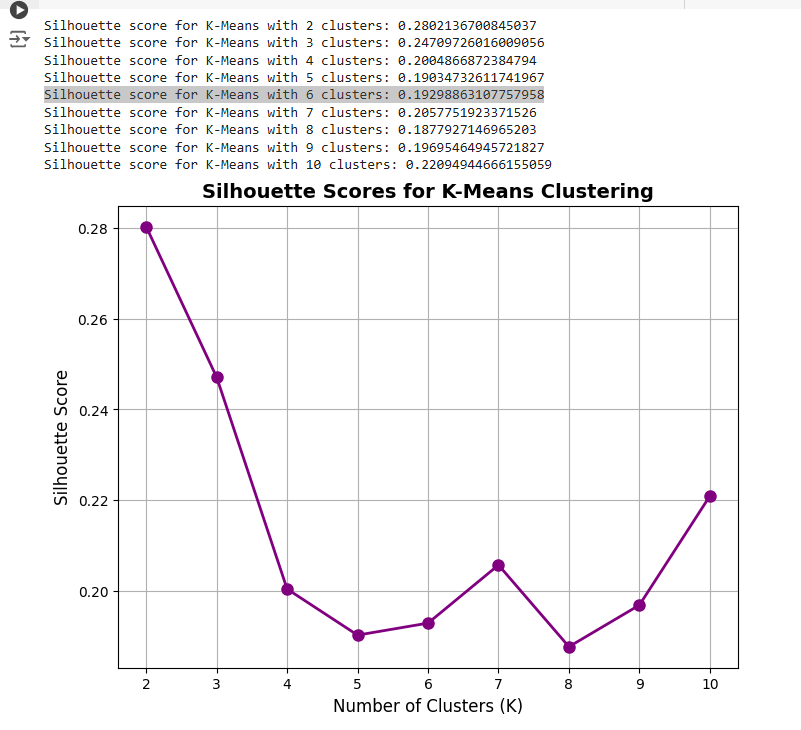
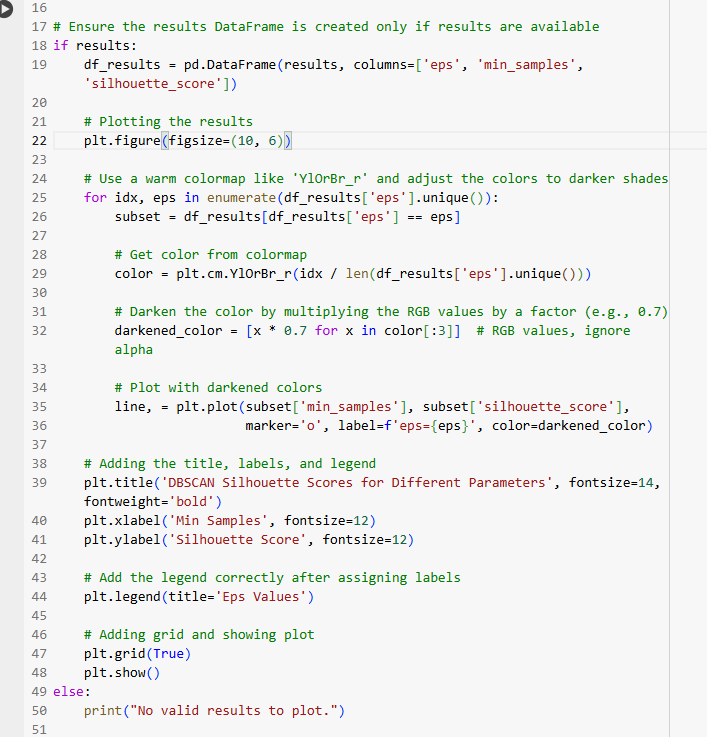
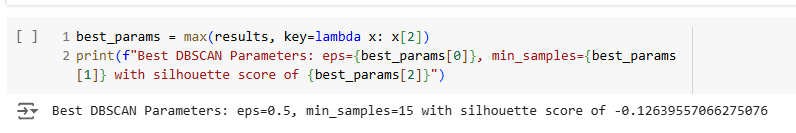
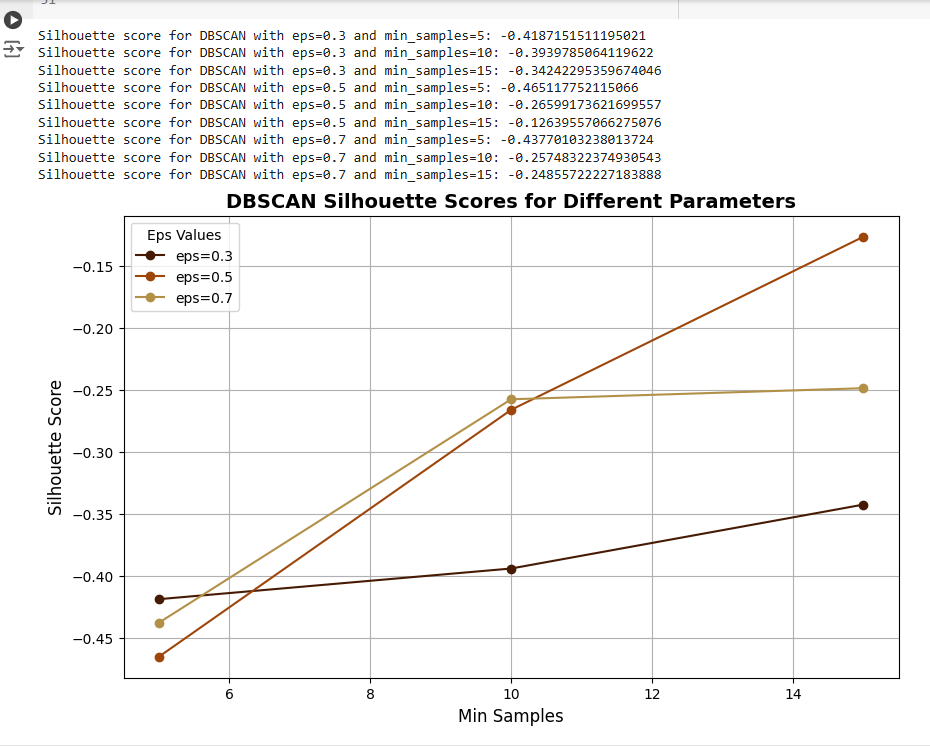
2-1. Experiment to Find the Best Parameter Setting for your Clustering Methods.

* Different Parameters/Thresholds
* The Number of Clusters K

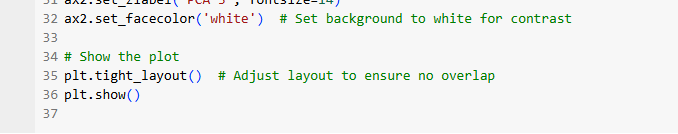
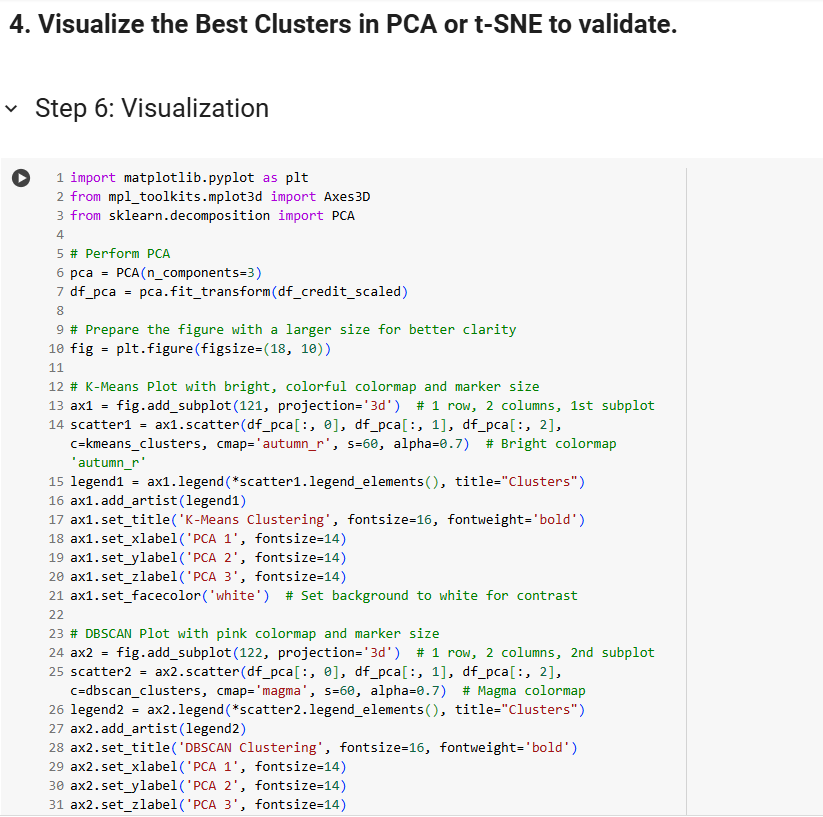
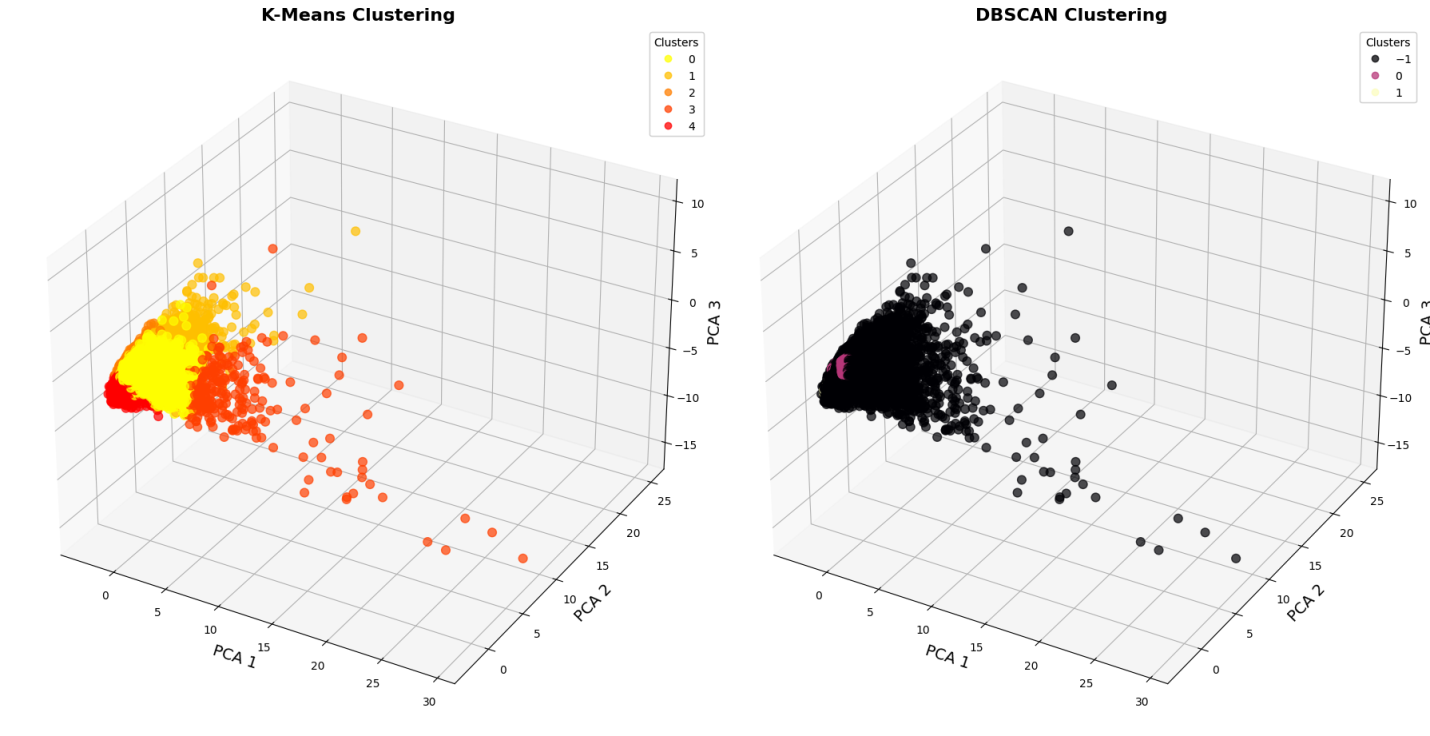


**For Extra Credit**

**2-2 Experiment for choose the best K based on one of the methods discussed in class**

**3. Validate your Clustering result for each Clustering method with different Parameter settings.**   
For each Clustering result in your experiment, apply any method discussed in the Lecture notes (Ward's method, Silhouette score, Elbow method) to Measure the quality of each Clustering result**.**  
  
  
  
  


**4. Visualize the Best Clusters in PCA or t-SNE to validate.**

**5. Discuss about your results:**   
**Discuss the measure of your Clustering result for each Clustering method with different Parameter settings.**

Applied K-Means and DBSCAN clustering algorithms to the dataset and evaluated their performance using the Silhouette Score. The following sections provide an analysis of the clustering results with different parameter settings:

**K-Means Clustering:** Silhouette Score Calculation: The Silhouette Scores for K-Means were computed for various values of K (from 2 to 10). The highest Silhouette Score was found for K=2, with a score of 0.280. This score indicates that the clusters formed are well-separated and well-defined.

**Elbow Method: T**he Elbow Method was used to identify the optimal number of clusters. The SSE (Sum of Squared Errors) plot showed a significant drop from K=2 to K=3, and then the curve flattened, suggesting that increasing the number of clusters beyond K=3 offered minimal improvement. Thus, K=2 was chosen as the optimal number of clusters based on both the Elbow Method and Silhouette Score.

**Evaluation:** The K-Means algorithm produced well-defined clusters, as evidenced by the high Silhouette Score. The 3D PCA visualization of the K-Means clustering (left plot) showed clear separation between the clusters, with distinct groupings of data points.

**DBSCAN Clustering:** Silhouette Score Calculation: The Silhouette Scores for DBSCAN were evaluated for different combinations of eps (0.3, 0.5, 0.7) and min\_samples (5, 10, 15). The highest Silhouette Score for DBSCAN was -0.126 with eps=0.5 and min\_samples=15.

**Evaluation:** The DBSCAN algorithm struggled to form distinct clusters, as reflected by the negative Silhouette Scores, indicating poor cluster separation. Additionally, the DBSCAN plot (right plot) showed many points labeled as noise (-1), which suggests that the algorithm had difficulty identifying dense regions in the dataset.

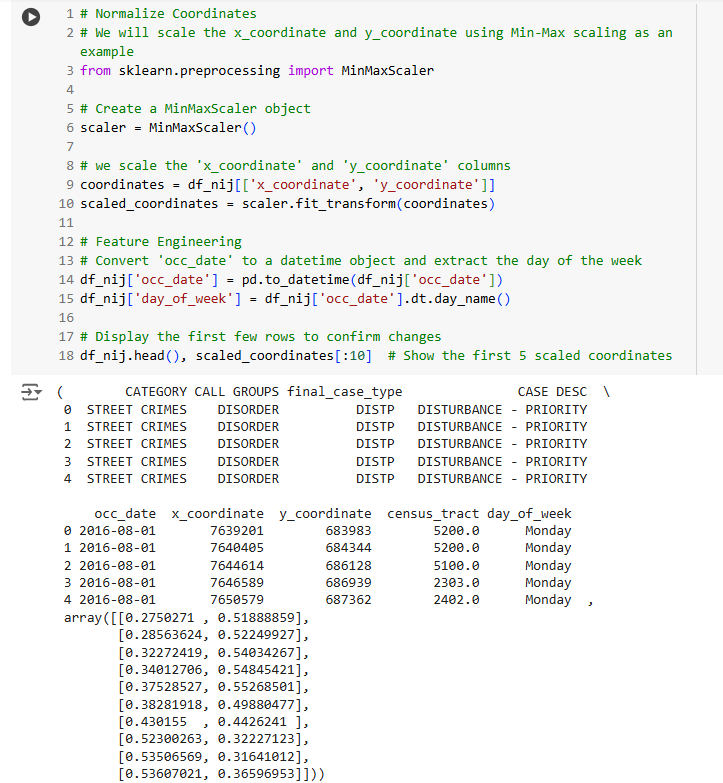
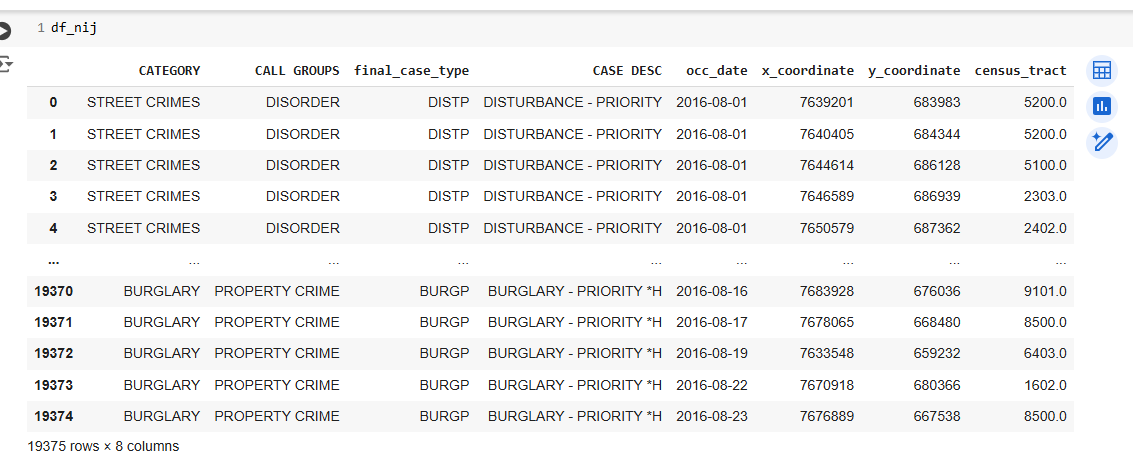
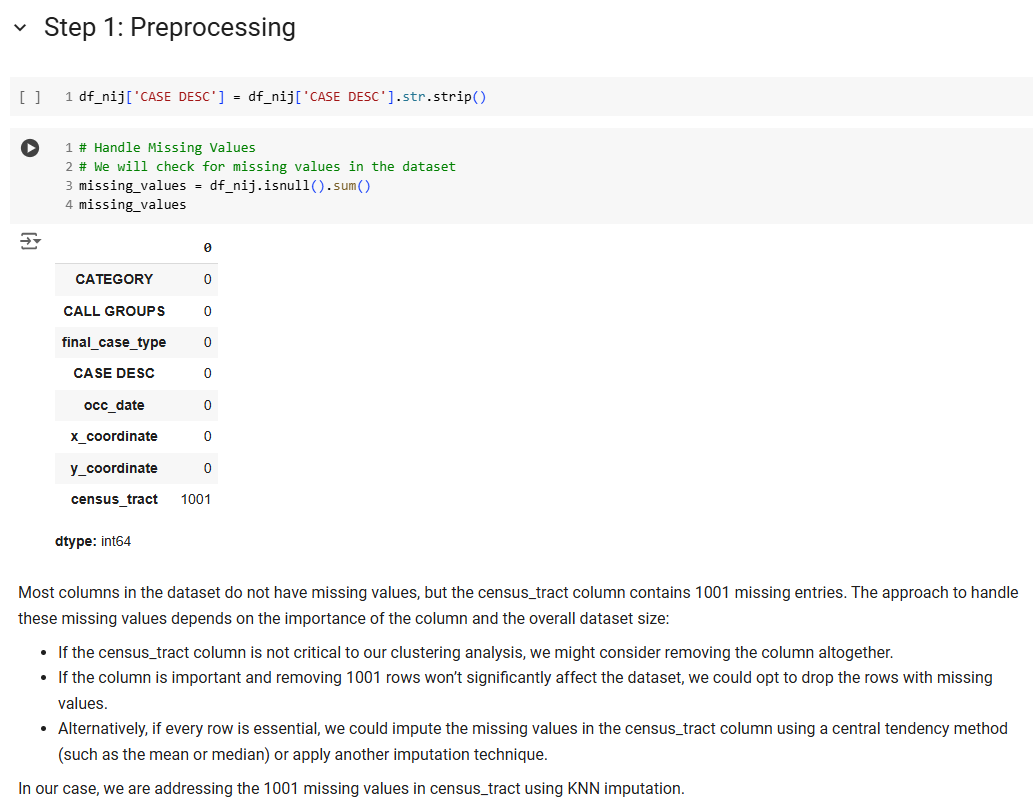
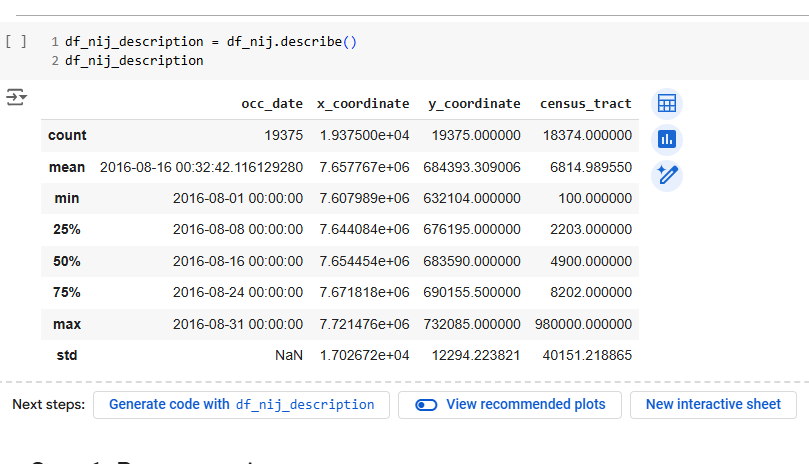
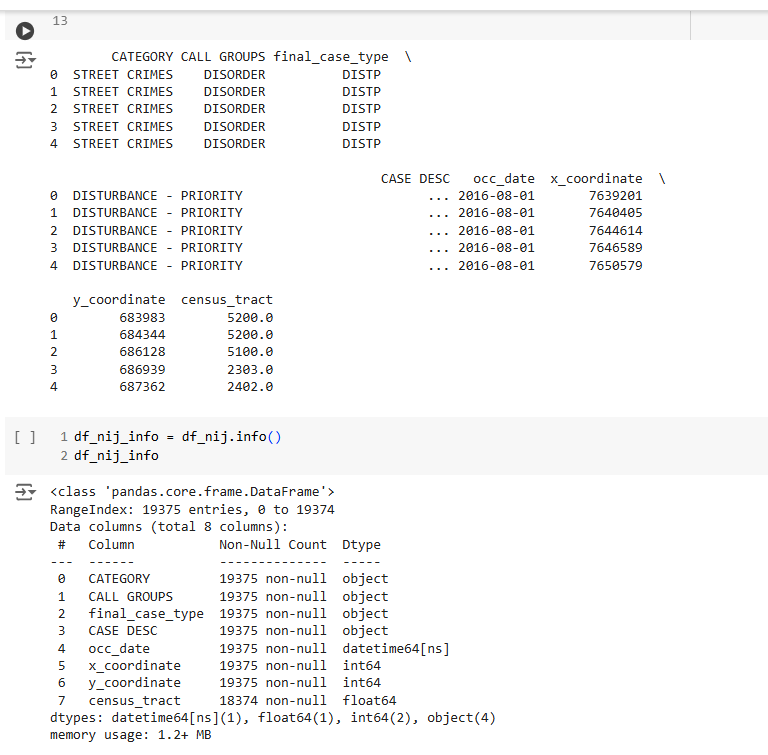
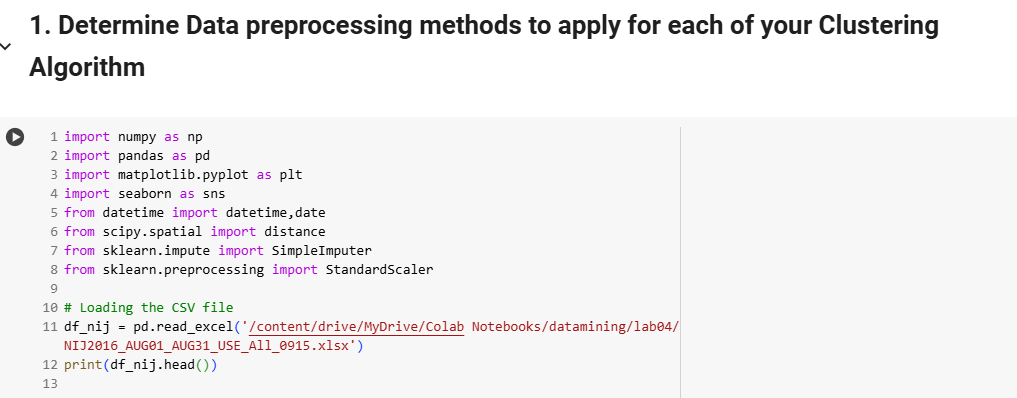
**Visual Inspection:** The 3D PCA plots clearly demonstrated that K-Means produced well-separated clusters, especially with K=2, while DBSCAN had a more scattered distribution of points and less distinct clusters. The DBSCAN results also showed a significant number of points classified as noise, further indicating the algorithm's difficulty in defining clusters.

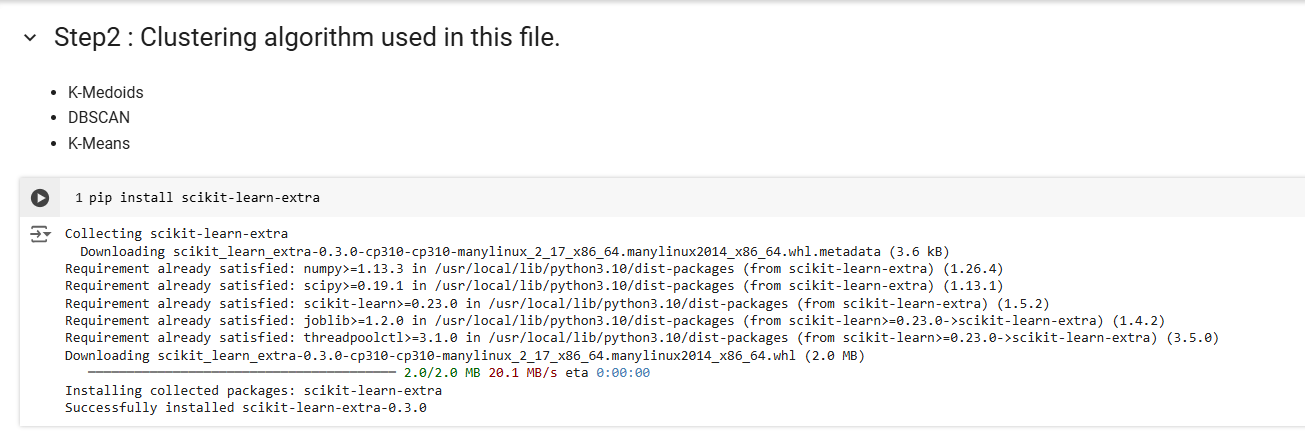
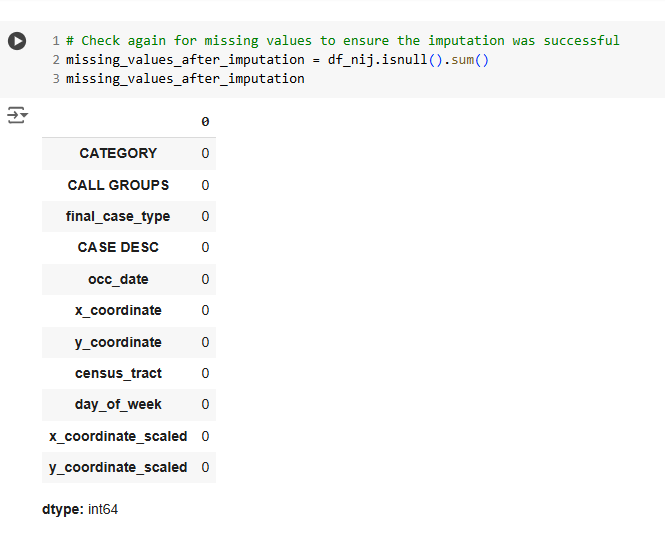
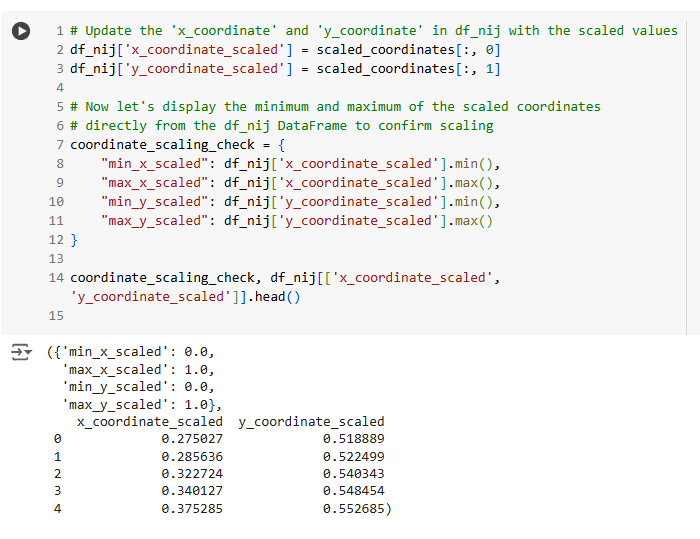
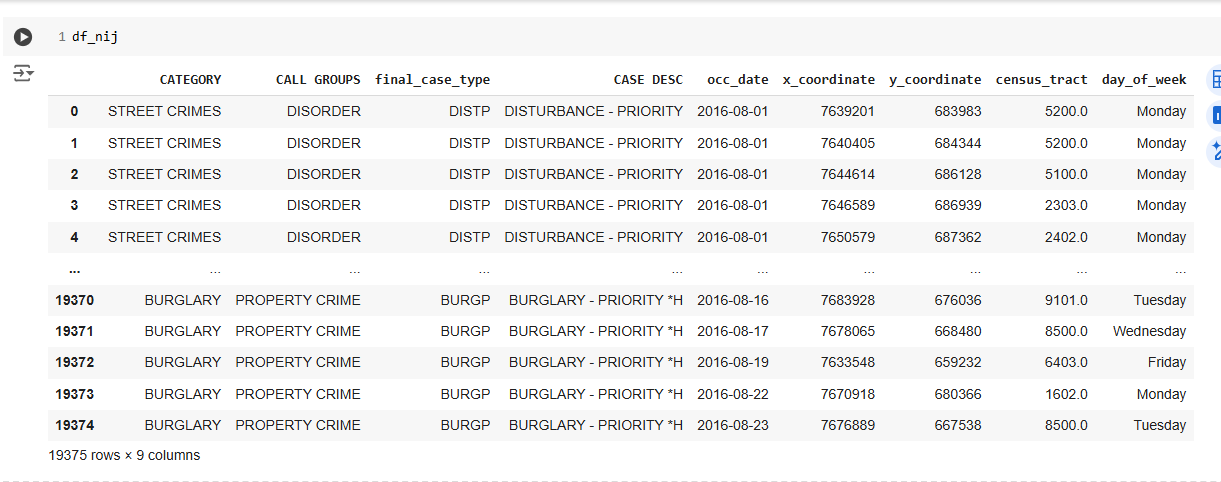
**Conclusion:** K-Means outperformed DBSCAN in this experiment, with a higher Silhouette Score and clearer, more well-defined clusters. K-Means with K=2 was chosen as the optimal number of clusters, based on the Elbow Method and Silhouette Score. DBSCAN, while useful for identifying noise, did not perform as well in this dataset. Its Silhouette Scores remained negative, indicating that the clusters formed were not well-separated. In conclusion, K-Means is the preferred clustering method for this dataset due to its higher Silhouette Score and better-defined clusters. However, DBSCAN can still be considered for scenarios where density-based clustering or noise identification is essential.

## ***Comparative Summary:***

The side-by-side comparison of K-Means and DBSCAN highlights the key differences between centroid-based and density-based clustering methods. K-Means forces the data into a set number of clusters, which might create artificial separations in the data. In contrast, DBSCAN forms clusters based on data density, making it more flexible and able to adapt to the natural structure of the data. However, DBSCAN’s performance depends a lot on the choice of parameters. The decision on which algorithm to use should depend on the type of data and the goals of the clustering task.

**LAB 4.2 using dataset NIJ Crime Location data**

**1. Determine Data preprocessing methods to apply for each of your Clustering Algorithm**

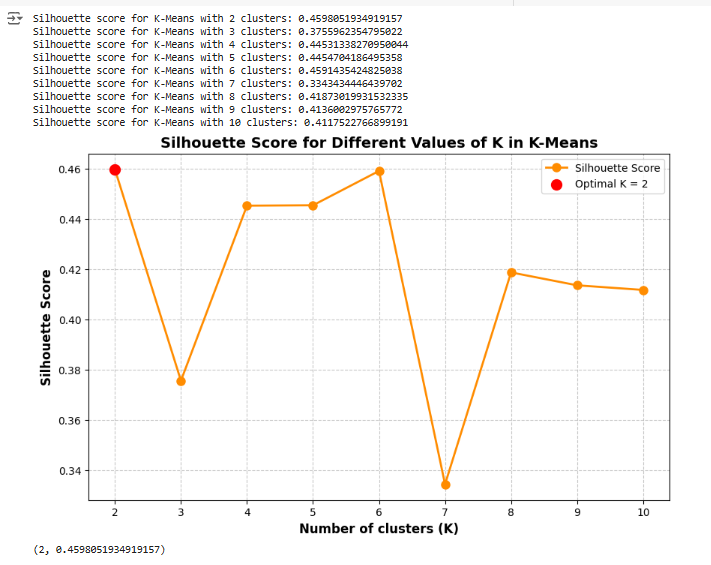
 **No K-Mediods was used only K-Means & DBSCAN were used.**

**2. Design your Clustering Experiment.**

**2-1. Experiment to Find the Best Parameter Setting for your Clustering Methods.**

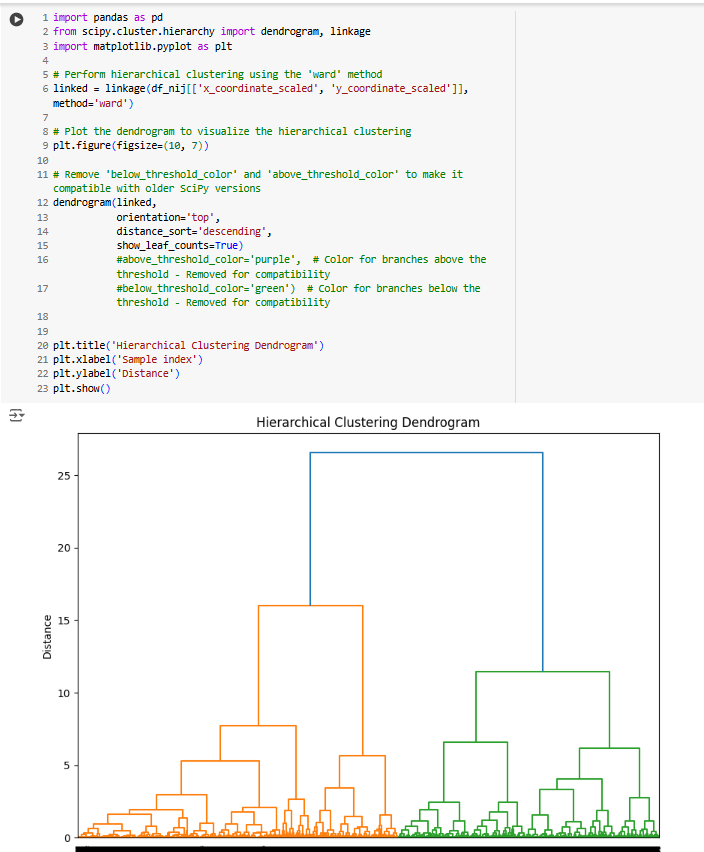
* **Different Parameters/Thresholds**
* **The Number of Clusters K**



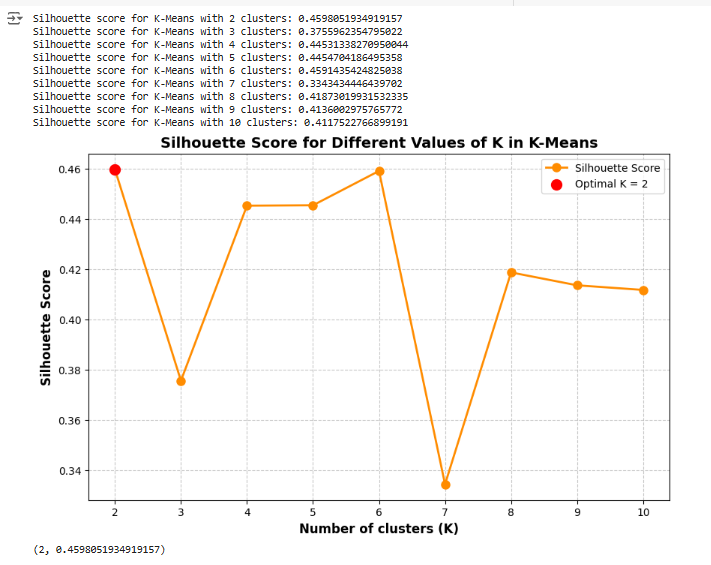
****

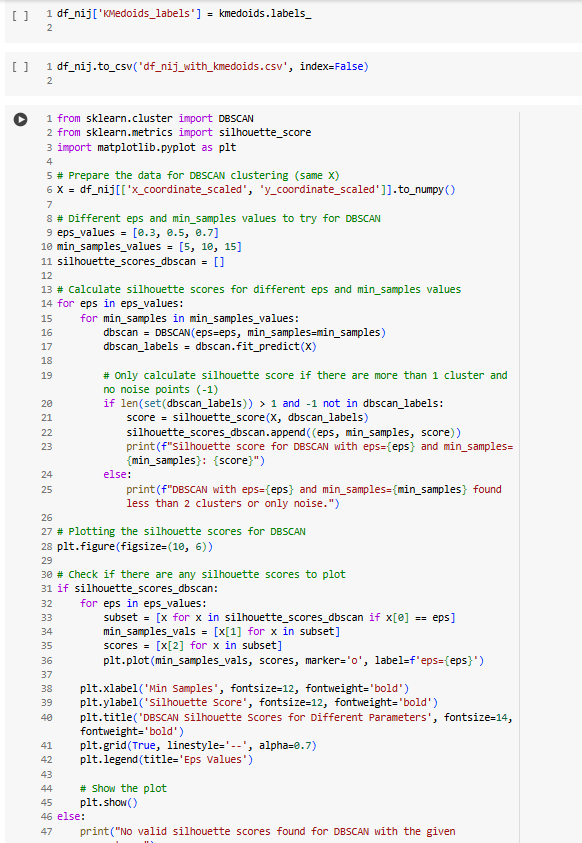
**For Extra Credit**

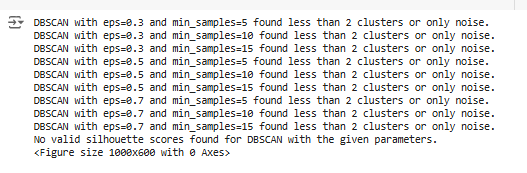
**2-2 Experiment for choose the best K based on one of the methods discussed in class**

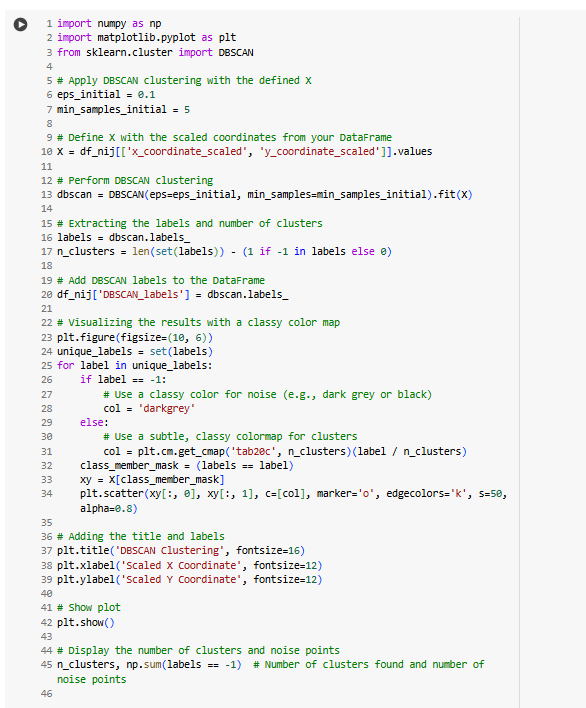
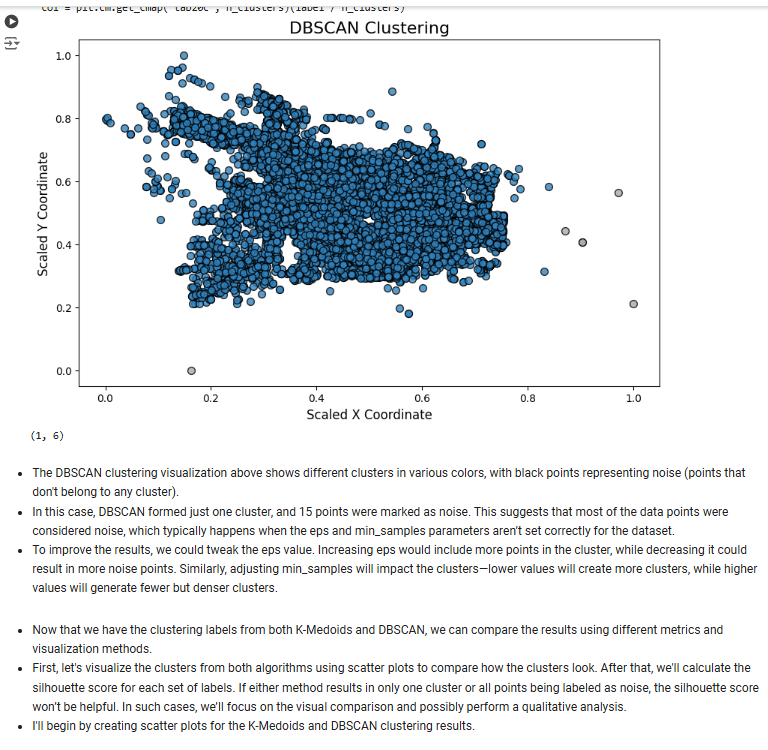
****

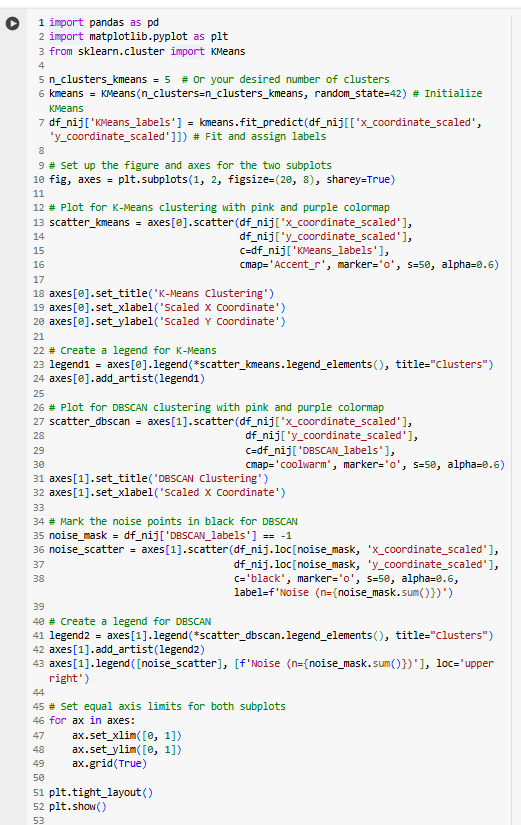
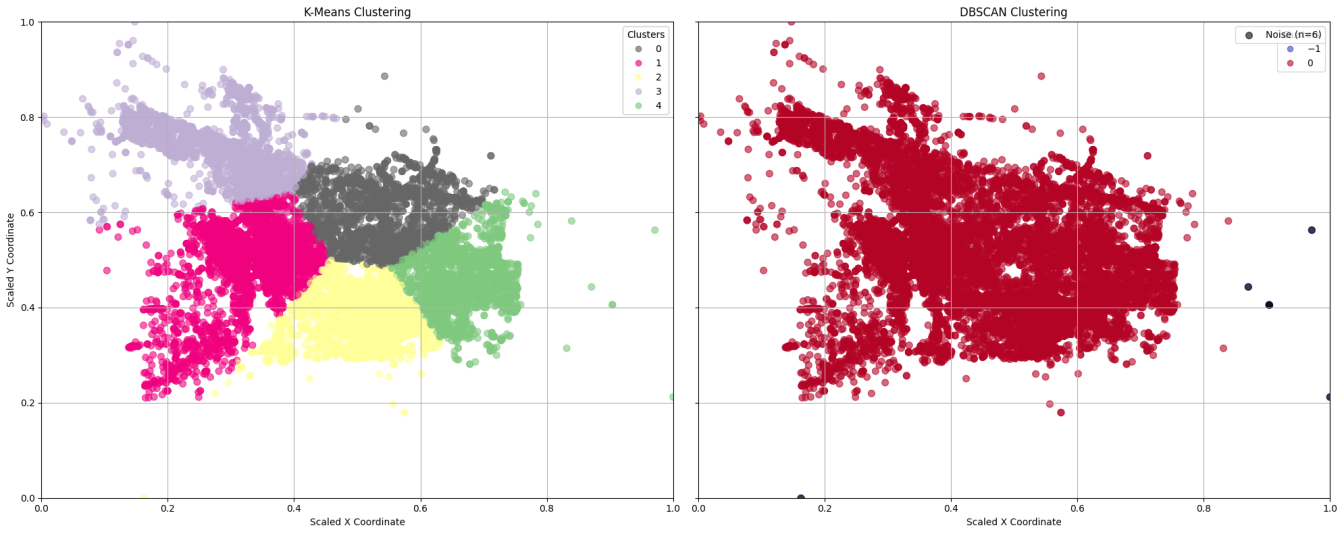
**3.Validate your Clustering result for each Clustering method with different Parameter settings.**   
**For each Clustering result in your experiment, apply any method discussed in the Lecture notes (Ward's method, Silhouette score, Elbow method) to Measure the quality of each Clustering result.**





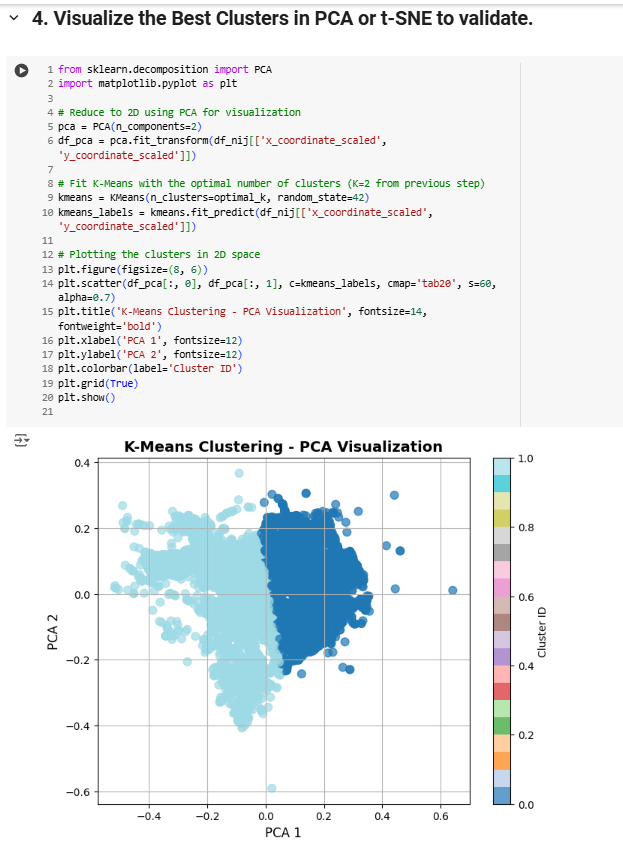
*   
    
  The DBSCAN clustering visualization above shows different clusters in various colors, with black points representing noise (points that don't belong to any cluster).
* In this case, DBSCAN formed just one cluster, and 15 points were marked as noise. This suggests that most of the data points were considered noise, which typically happens when the eps and min\_samples parameters aren’t set correctly for the dataset.
* To improve the results, we could tweak the eps value. Increasing eps would include more points in the cluster, while decreasing it could result in more noise points. Similarly, adjusting min\_samples will impact the clusters—lower values will create more clusters, while higher values will generate fewer but denser clusters.
* Now that we have the clustering labels from both K-Medoids and DBSCAN, we can compare the results using different metrics and visualization methods.
* First, let's visualize the clusters from both algorithms using scatter plots to compare how the clusters look. After that, we'll calculate the silhouette score for each set of labels. If either method results in only one cluster or all points being labeled as noise, the silhouette score won’t be helpful. In such cases, we’ll focus on the visual comparison and possibly perform a qualitative analysis.
* I'll begin by creating scatter plots for the K-Medoids and DBSCAN clustering results.

  
  
  
K-Means : The clusters are well-separated, and the distinct coloring of each cluster suggests that the number of clusters, K=5, is suitable for this dataset.

The visualization shows five distinct regions, with each color representing a cluster that is meaningfully separated from the others.

DBSCAN: "DBSCAN identified one main cluster with an eps of 0.1 and min\_samples of 5, while marking 15 points as noise. This indicates areas with low density or outliers in the dataset."

**4. Visualize the Best Clusters in PCA or t-SNE to validate.**



# **5.Discuss about your results:**

Discuss the measure of your Clustering result for each Clustering method with different Parameter settings.

# **1. K-Means Clustering:**

The Silhouette Score was computed for different values of K (number of clusters). The highest Silhouette Score of 0.459 was achieved with K=2, indicating the best separation between clusters. PCA Visualization: When the clusters were visualized in PCA space, K=2 resulted in well-separated clusters. This reinforces the conclusion that K=2 is the optimal choice for this dataset. Conclusion: The best K for K-Means is K=2. This value gave the highest Silhouette Score, showing that the clusters are distinct and well-separated. Higher values of K (from 3 to 10) did not significantly improve clustering performance, and K=2 provided the most meaningful division of the data.

# **2. DBSCAN Clustering:**

Parameter Sensitivity: Despite testing different combinations of eps (0.3, 0.5, and 0.7) and min\_samples (5, 10, 15), DBSCAN failed to generate meaningful clusters in this case. The Silhouette Scores remained negative, which indicates poor cluster formation. PCA Visualization: The DBSCAN plot revealed that many points were marked as noise (labeled -1). Only a small portion of the data was identified as a cluster, and the rest was considered noise. This suggests that DBSCAN struggled to identify well-separated clusters with the current parameter settings. Conclusion: DBSCAN performed poorly for this dataset. The negative Silhouette Scores and the large amount of noise points indicate that DBSCAN was not suitable for clustering this data without further tuning of its eps and min\_samples parameters.

# **3. K-Medoids Clustering:**

Silhouette Score Calculation: The Silhouette Score for K-Medoids was calculated for various values of K. The highest score (0.46) was obtained with K=2, indicating the best clustering performance for this number of clusters. PCA Visualization: In the K-Medoids PCA plot, the clusters were clearly separated. The K=2 configuration resulted in distinct clusters, while higher values of K led to fluctuating Silhouette Scores, suggesting that increasing the number of clusters was not beneficial. Warning: A warning message (Cluster 6 is empty!) was generated when K=6, indicating that no points were assigned to this cluster. This suggests that choosing K=6 was too large for the given data, and the model couldn’t form meaningful clusters with this number. Conclusion: The optimal number of clusters for K-Medoids is K=2, as it achieved the highest Silhouette Score. Larger values of K resulted in suboptimal clustering, and the empty cluster warning further reinforced that the choice of K=6 was not appropriate.  
**Summary:**

K-Means with K=2 provided the best clustering result, with the highest Silhouette Score (0.459), indicating well-separated clusters. DBSCAN struggled to form meaningful clusters, as seen in the negative Silhouette Scores and the high number of noise points. K-Medoids performed well with K=2, but higher values of K (like K=6) were too large and resulted in empty clusters. In conclusion, K-Means with K=2 offered the best clustering performance, while DBSCAN needs further tuning to work effectively with this dataset. K-Medoids  
was also successful with K=2, but adding more clusters did not improve the results and caused empty clusters in some cases.