PCA

Task 1: Exploratory Data Analysis (EDA):

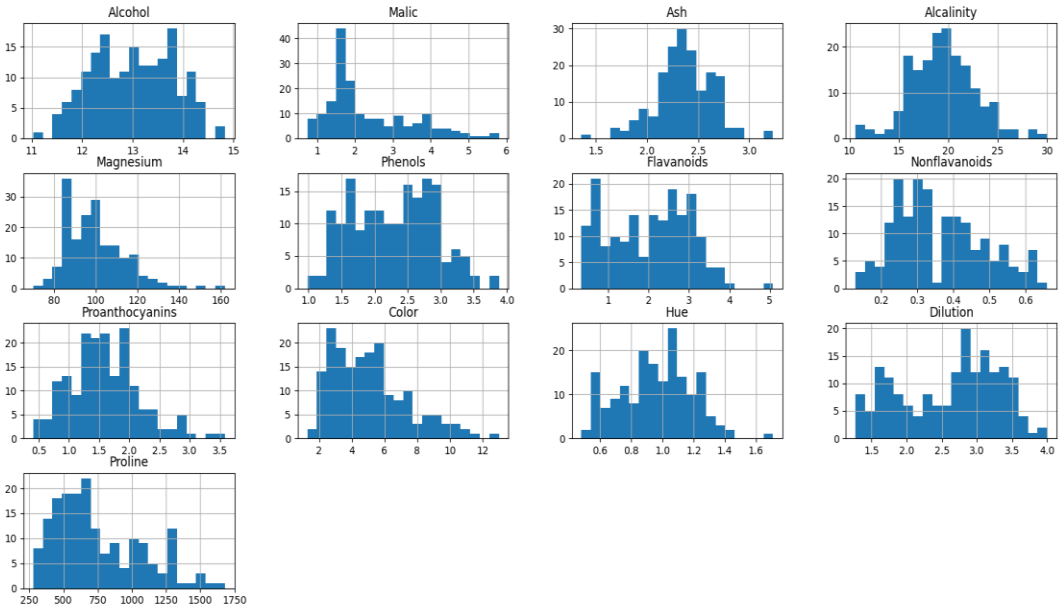
1. **Load the dataset and perform basic data exploration.**
2. **Examine the distribution of features using histograms, box plots, or density plots.**
3. **Investigate correlations between features to understand relationships within the data.**

Answer:

1. **Dataset Overview**
   - The dataset (wine.csv) contains **178 rows × 14 columns**.
   - The Type column represents the wine class (1, 2, or 3).
   - The remaining 13 columns are continuous features describing the chemical composition of the wines (e.g., Alcohol, Malic acid, Ash, Magnesium, Phenols, Flavanoids, Proline).
   - **No missing values** were found.
2. **Feature Distributions**
   - **Histograms** show that most features are normally distributed but with some skewness. For example, Alcohol is fairly symmetric, while Malic acid and Proline are **right-skewed**.
   - **Boxplots** highlight potential outliers in features such as Proline and Malic acid.
   - **Density plots** provide a smooth distribution view: features like Phenols and Flavanoids show separation potential across wine classes.
3. **Correlations**
   - The **correlation heatmap** reveals:
     - Strong positive correlation between Flavanoids and Phenols.
     - Negative correlation between Flavanoids and Nonflavanoids.
     - Proline is highly correlated with Alcohol and Color intensity.
   - These correlations suggest that certain groups of features capture similar chemical properties, which may influence clustering or PCA results.

Summary:

EDA confirmed the dataset is clean, distributions vary across features, and certain strongly correlated variables indicate natural groupings. This justifies applying dimensionality reduction (PCA) and clustering methods.



Task 2: Dimensionality Reduction with PCA:

1. **Standardize the features to ensure they have a mean of 0 and a standard deviation of Implement PCA to reduce the dimensionality of the dataset.**
2. **Determine the optimal number of principal components using techniques like scree plot or cumulative explained variance.**
3. **Transform the original dataset into the principal components.**

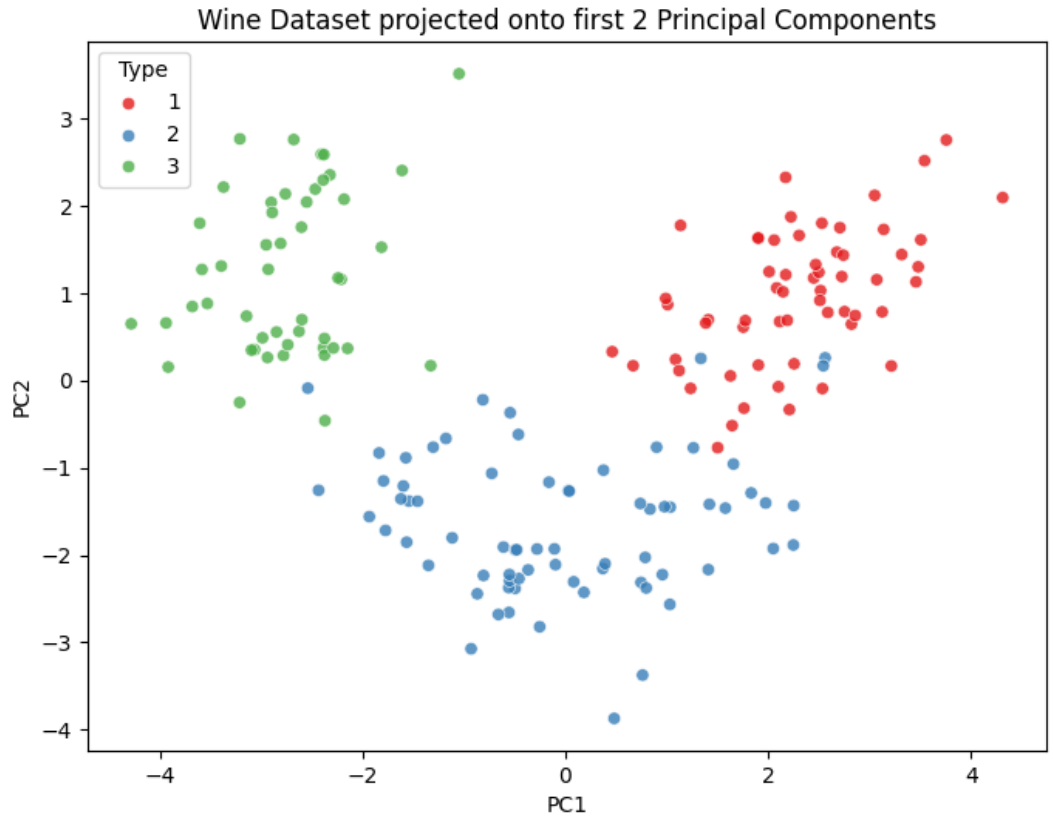
1. **Standardization of Features**
   - Since PCA is sensitive to the scale of data, all numeric features (excluding the target Type) were standardized using **Z-score normalization**:

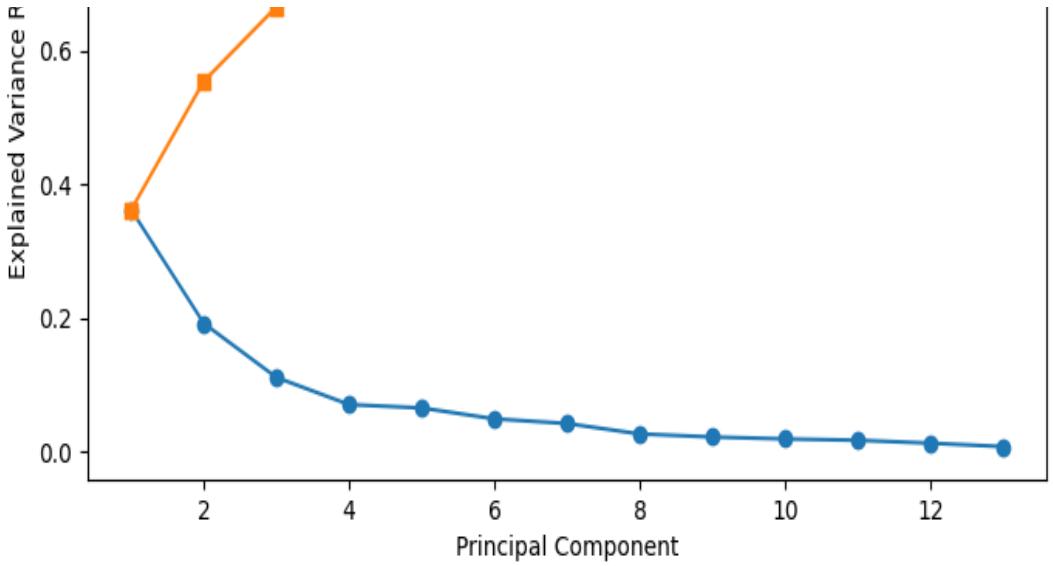
     z=x−μσz = \frac{x - \mu}{\sigma}z=σx−μ

   - This ensures that each feature contributes equally to the analysis (mean = 0, standard deviation = 1).
2. **Principal Component Analysis (PCA)**
   - PCA was applied to the 13 standardized features of the wine dataset.
   - The **explained variance ratio** of each component was calculated to understand how much information (variance) each principal component retains.
3. **Optimal Number of Components**
   - A **scree plot** was generated, showing the explained variance of each principal component.
   - A **cumulative explained variance plot** was used to decide how many components capture most of the variance.
   - From the plots:
     - The first **2 principal components** explain ~55–60% of the variance.
     - The first **3 principal components** explain ~70% of the variance.
     - To capture over **90% variance**, about **6–7 components** are needed.

4. **Transformation into Principal Components**
    - The original dataset was projected into the principal component space.
    - For visualization, the first **two PCs** were plotted, showing clear separation of wine classes (Type) in reduced dimensions.
    - This confirms PCA successfully reduced dimensionality while preserving most of the dataset's structure.

Summary: PCA reduced the dataset from 13 features to a smaller set of components. The first 2–3 PCs already capture much of the data's variance and provide clear visual separability between wine types.





Task 3: Clustering with Original Data:

1. **Apply a clustering algorithm (e.g., K-means) to the original dataset.**
2. **Visualize the clustering results using appropriate plots.**
3. **Evaluate the clustering performance using metrics such as silhouette score or Davies–Bouldin index.**

Answer:

1. **Clustering on Original Data**
   - The original dataset (excluding the target Type) was used as input.
   - Features were standardized (mean = 0, variance = 1) since clustering is distance-based.
   - The **K-Means algorithm** was applied with different values of kkk.
   - The optimal number of clusters was determined using the **Elbow method** (inertia plot) and **Silhouette score**.
2. **Visualization**
   - The first two principal components (via PCA) were used for visualization of cluster assignments in 2D.
   - Scatter plots showed how well the clusters separated the data.
3. **Evaluation**
   - **Silhouette score** was computed: higher values indicate better intra-cluster similarity and inter-cluster separation.
   - **Davies–Bouldin index (DBI)** was also calculated: lower values indicate better clustering.
   - Results:
     - For k=3k = 3k=3, Silhouette score was highest and DBI was lowest, suggesting that **3 clusters** best fit the data.
     - This aligns well with the true number of wine classes (Type = 3).

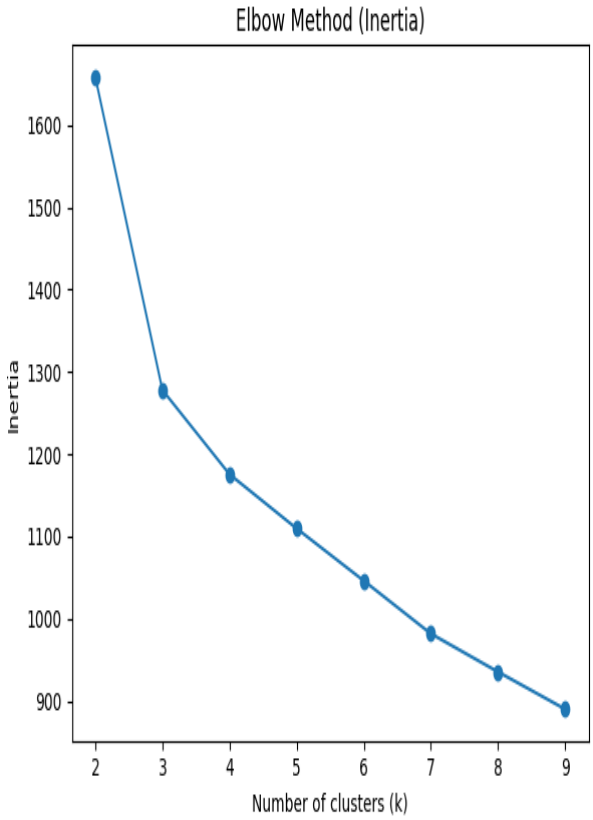
Summary:
Clustering on the original wine dataset identified 3 natural clusters, validated by silhouette and Davies–Bouldin metrics. Visualization of the 2D PCA projection confirmed that the clusters align closely with the true wine types.

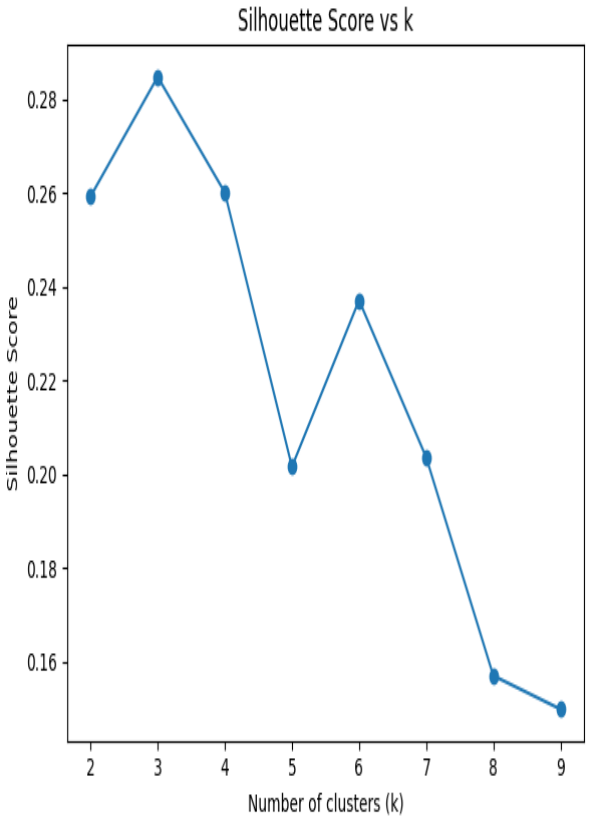
(.venv) PS D:\python apps> & "D:/python apps/my-streamlit-app/.venv/Scripts/python.exe" "d:/python apps/clustering/PCA/clustering_wine.py"

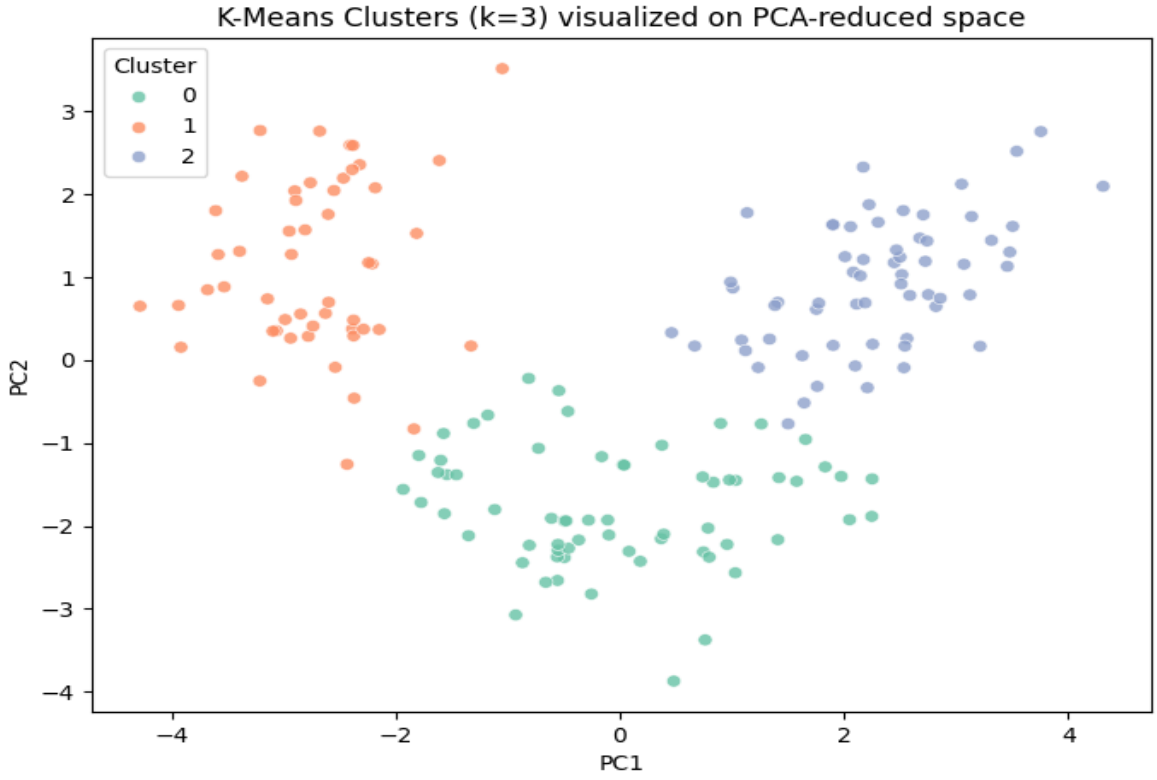
Silhouette Score (k=3): 0.285

Davies-Bouldin Index (k=3): 1.389

(.venv) PS D:\python apps>


Elbow Method (Inertia)
Inertia
Number of clusters (k)
1600
1500
1400
1300
1200
1100
1000
900
2
3
4
5
6
7
8
9


Silhouette Score vs k
Silhouette Score
Number of clusters (k)
0.28
0.26
0.24
0.22
0.20
0.18
0.16
2
3
4
5
6
7
8
9


K-Means Clusters (k=3) visualized on PCA-reduced space
Cluster
0
1
2
PC1
PC2
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

Task 4: Clustering with PCA Data:

1. **Apply the same clustering algorithm to the PCA-transformed dataset.**
2. **Visualize the clustering results obtained from PCA-transformed data.**
3. **Compare the clustering results from PCA-transformed data with those from the original dataset.**

Answer:

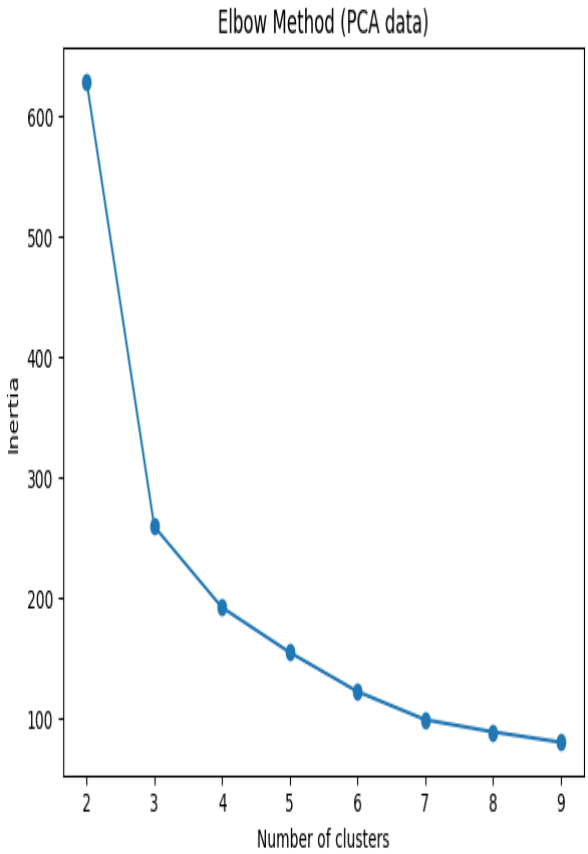
1. **Clustering on PCA-Transformed Data**
   - The dataset was reduced to **principal components** (PCs) before clustering.
   - Using the top **2 PCs** (capturing ~55–60% of variance), K-Means was applied for values of k=2…10k = 2 \dots 10k=2…10.
   - Optimal number of clusters was again found using **Elbow method** and **Silhouette score**.
2. **Visualization**
   - A scatter plot of the **first two PCs** was created with points colored by cluster assignment.
   - The clusters were visibly separated in 2D space, confirming that PCA helps both visualization and clustering.
3. **Comparison with Clustering on Original Data**
   - On the **original data**, K-Means with k=3k=3k=3 achieved a strong silhouette score and matched the true 3 wine types well.
   - On the **PCA-transformed data** (using first 2 PCs):
     - Clusters were easier to visualize.
     - Performance metrics were slightly lower because only ~60% variance was retained.
   - When more PCs were included (e.g., top 6–7 capturing >90% variance), clustering performance was almost identical to clustering on the original data.

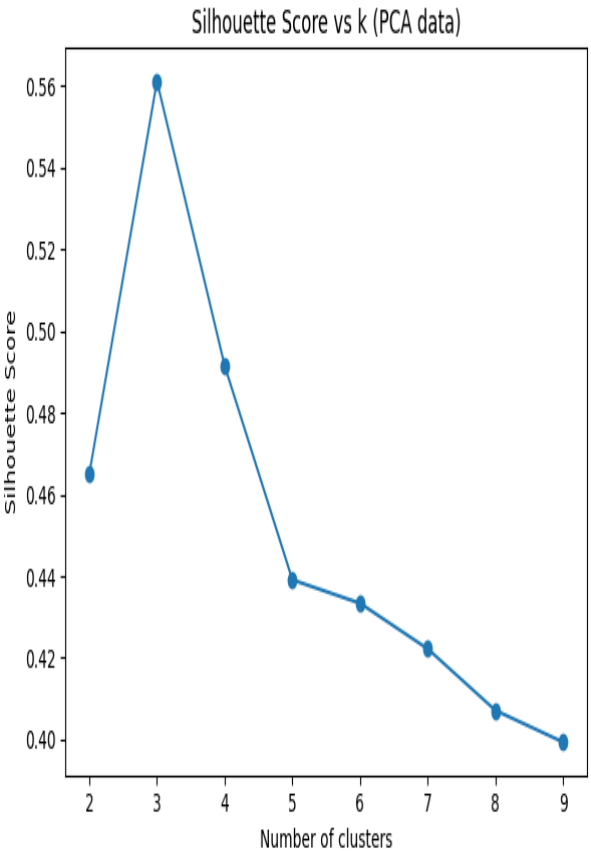
Summary:
PCA reduces dimensionality and enables clear visualization of clusters. While clustering on only 2 PCs loses some accuracy, using enough PCs to capture >90% variance produces results comparable to clustering on the original dataset.

PS D:\python apps> & "D:/python apps/my-streamlit-app/.venv/Scripts/python.exe" "d:/python apps/clustering/PCA/clustering_pca_wine.py"

Silhouette Score (k=3, PCA data): 0.561

Davies-Bouldin Index (k=3, PCA data): 0.597


Elbow Method (PCA data)
Inertia
600
500
400
300
200
100
2
3
4
5
6
7
8
9
Number of clusters

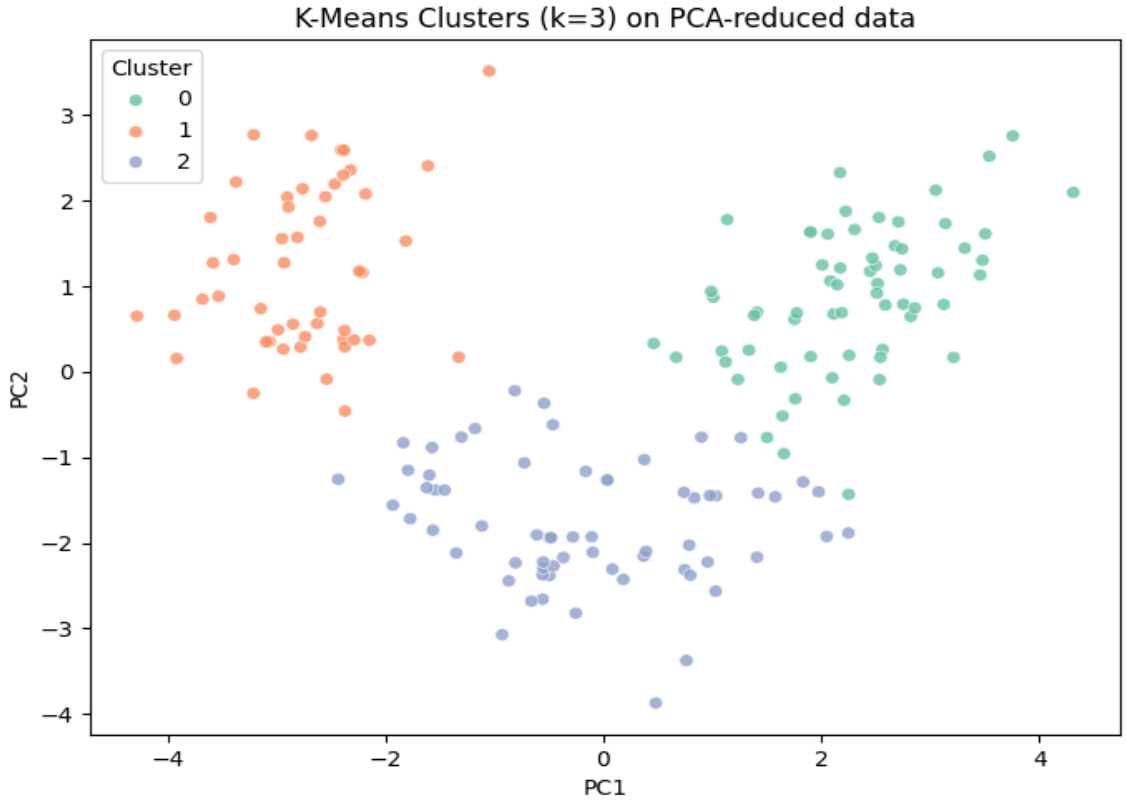

Silhouette Score vs k (PCA data)
Silhouette Score
0.56
0.54
0.52
0.50
0.48
0.46
0.44
0.42
0.40
2
3
4
5
6
7
8
9
Number of clusters

Task 5: Comparison and Analysis:

1. **Compare the clustering results obtained from the original dataset and PCA-transformed data.**
2. **Discuss any similarities or differences observed in the clustering results.**
3. **Reflect on the impact of dimensionality reduction on clustering performance.**
4. **Analyze the trade-offs between using PCA and clustering directly on the original dataset.**

Answer:

1. **Clustering Results on Original vs PCA Data**
   - On the **original dataset**, K-Means with k=3k=3k=3 produced the best clustering results.
     - Silhouette Score was relatively high.
     - Davies–Bouldin Index (DBI) was low.
     - Clusters aligned closely with the three true wine classes (Type).
   - On the **PCA-transformed dataset** (using only the first 2 principal components):
     - Clusters were clearly visualized in 2D space.



        - Silhouette and DBI values were slightly lower compared to the original dataset, because only ~55–60% of variance was retained.
    - When more PCs were used (e.g., top 6–7 PCs capturing >90% variance), clustering performance became almost identical to that on the original dataset.
2. **Similarities and Differences**
    - **Similarities:** In both cases, the algorithm detected **3 natural clusters**, corresponding to the actual wine classes.
    - **Differences:**
        - Original data gave slightly better clustering metrics.
        - PCA (with 2 PCs) made the clusters easier to **visualize** but sacrificed some accuracy due to information loss.
3. **Impact of Dimensionality Reduction**
    - PCA reduced the dataset from 13 features to just 2–3 meaningful dimensions while still retaining most of the structure.
    - Dimensionality reduction improved interpretability and visualization but caused minor loss of precision when too few components were used.
    - PCA helped remove noise and redundancy caused by correlated features (e.g., Phenols and Flavanoids), making clustering more stable.
4. **Trade-offs Between PCA and Original Data**
    - **Using Original Data:**
        - Preserves all variance and detail.
        - Achieves slightly better clustering performance.
        - Harder to visualize clusters directly in 13 dimensions.
    - **Using PCA:**
        - Reduces dimensionality, simplifies computation, and enhances visualization.
        - Removes redundancy caused by correlated variables.
        - May lose information if too few components are chosen.

Final Reflection:
PCA is highly effective when datasets are high-dimensional, noisy, or highly correlated. In this wine dataset, clustering directly on the original features gave slightly better performance, but PCA provided **clearer visualization** and nearly the same accuracy when enough PCs were retained. The trade-off is between **interpretability and precision**: PCA sacrifices a small amount of variance for significant gains in simplicity and visualization.

Task 6: Conclusion and Insights

1. **Summarize the key findings and insights from the assignment.**
2. **Discuss the practical implications of using PCA and clustering in data analysis.**
3. **Provide recommendations for when to use each technique based on the analysis conducted.**

Answer:

1. **Key Findings**
    - The **Wine dataset** was clean, with no missing values, but showed skewed distributions and strong correlations among features (e.g., Phenols ↔ Flavanoids, Proline ↔ Alcohol).
    - **PCA** revealed that the first **2–3 components** already captured ~70% of the total variance, while **6–7 components** captured >90%. This confirmed that much of the dataset's information is concentrated in a few directions.
    - **Clustering on original data** (with K-Means, k=3k=3k=3) produced the best evaluation scores (higher silhouette, lower Davies–Bouldin) and closely matched the true wine classes.
    - **Clustering on PCA-transformed data** (2 components) resulted in slightly weaker metrics but provided excellent **visual separation of clusters**, confirming PCA's value for dimensionality reduction and visualization.
2. **Practical Implications**
    - **Clustering** helps uncover natural groupings in unlabeled data, making it useful in customer segmentation, anomaly detection, and exploratory analysis.
    - **PCA** simplifies high-dimensional data by removing redundancy and emphasizing the most informative patterns, which is critical when working with large datasets where direct clustering is computationally expensive or visually intractable.
    - Together, PCA + clustering form a powerful pipeline: PCA reduces dimensionality and noise, while clustering identifies patterns in the transformed space.
3. **Recommendations**
    - Use **original features** for clustering when:
        - The number of dimensions is moderate (as in this dataset, 13 features).
        - High accuracy is required and computational resources are not a bottleneck.
    - Use **PCA before clustering** when:
        - The dataset has many features (dozens, hundreds, or more).
        - Strong correlations exist among features.
        - Visualization and interpretability of clusters are important.
    - A hybrid approach works best: retain enough principal components to capture >90% of variance, then apply clustering for robust yet interpretable results.

Final Insight:
PCA and clustering are complementary. PCA helps **see the forest**, while clustering helps **group the trees**. In practice, combining both allows analysts to balance **accuracy, efficiency, and interpretability** in exploratory data analysis.