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.

**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.

**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.

**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.

**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.

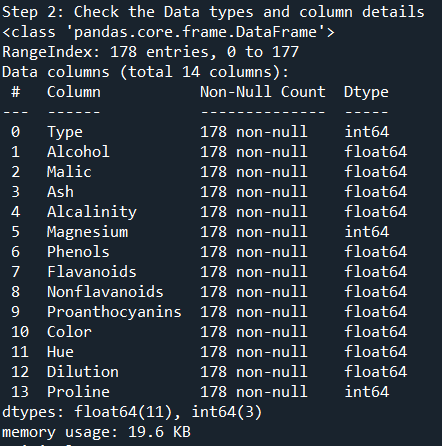
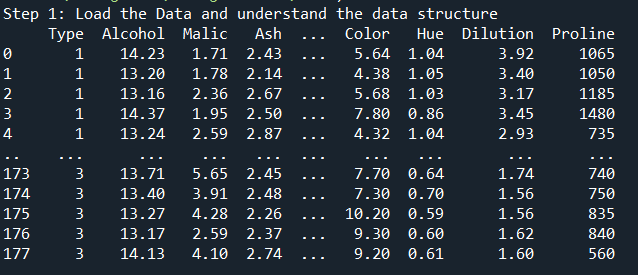
**Task 6: Conclusion and Insights**

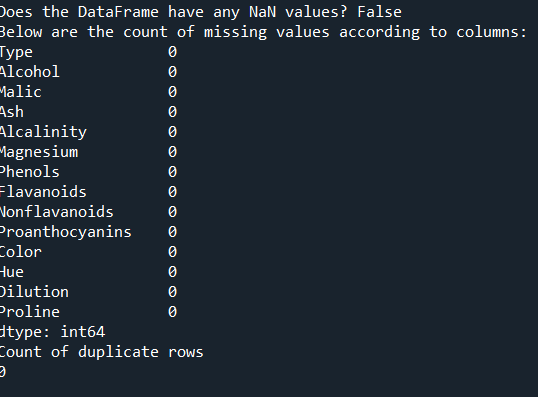
1. Summarize the key findings and insights from the assignment.
2. Discuss the practical implications of using PCA and clustering in data analysis.
3. Provide recommendations for when to use each technique based on the analysis conducted.

Observations

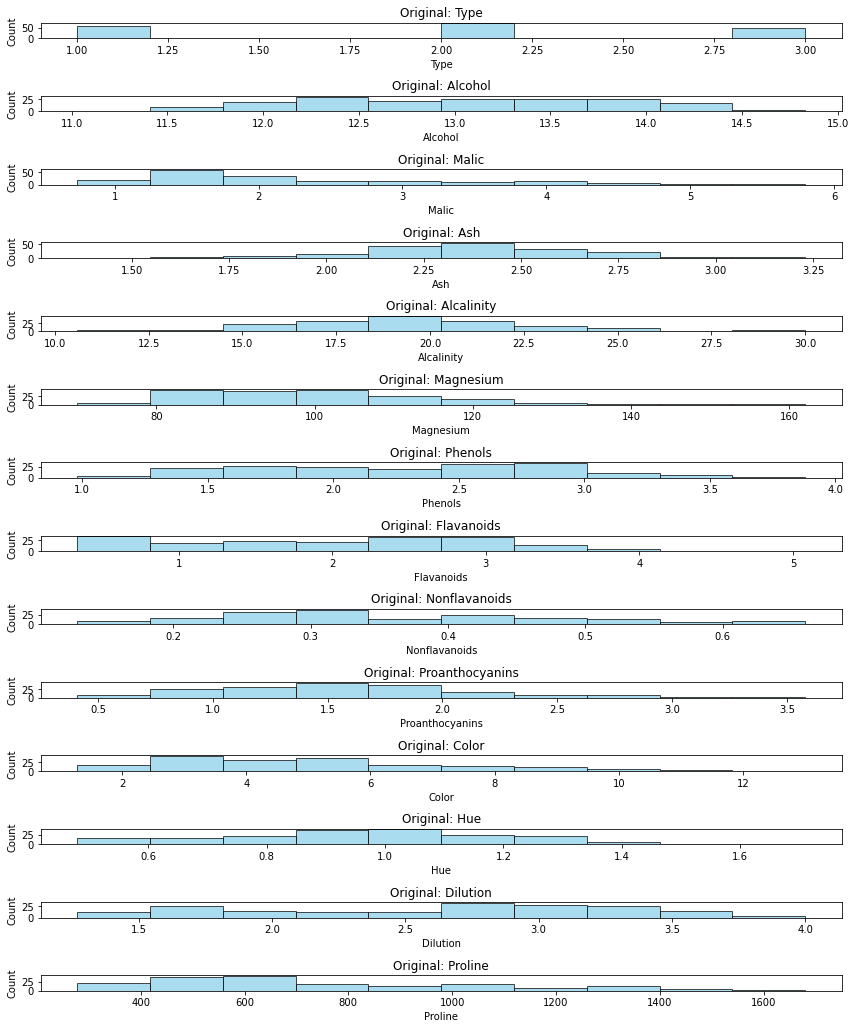
**Task 1: Exploratory Data Analysis (EDA):**

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.

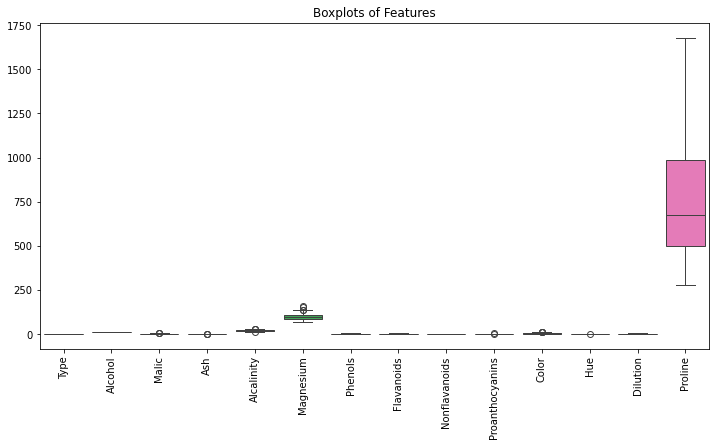




* There is no null value and duplicates present into Dataset



* **Type**: The dataset appears to have three distinct classes (1, 2, 3), with an uneven distribution among them.
* **Alcohol**: The values range from around 11 to 15, with a peak concentration between 12 and 14.
* **Malic Acid**: Most values lie between 1 and 3, with some outliers reaching up to 6.
* **Ash**: The values are normally distributed around 2-3.
* **Alcalinity**: Values range from 10 to 30, with most samples concentrated around 15-25.
* **Magnesium**: The distribution is right-skewed, with most values between 70 and 130.
* **Phenols & Flavanoids**: Both features show a wide range of values, indicating significant variation in their presence.
* **Nonflavanoid Phenols & Proanthocyanins**: These features have a more concentrated range, with values generally below 1.0 and 3.5, respectively.
* **Color Intensity**: Shows a spread from 1 to around 13, with a concentration below 6.
* **Hue**: The distribution is relatively normal, with most values between 0.6 and 1.6.
* **Dilution**: The values are mostly concentrated between 1.5 and 3.5.
* **Proline**: The widest range (around 200 to 1600), showing a right-skewed distribution.

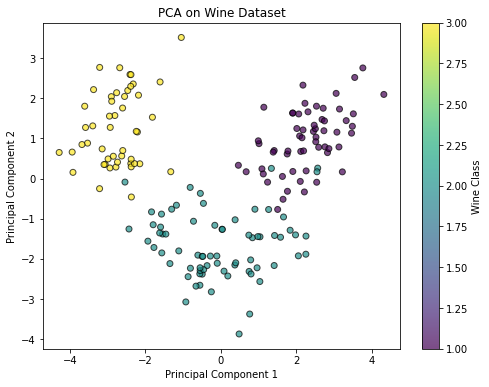


Observations from the Box Plot:

1. **Proline has the highest range**:
   * The values are widely spread out, ranging from low to extremely high values (~200 to ~1700).
   * This feature shows significant variability and may dominate others if not scaled properly.
2. **Magnesium has some variation**:
   * It has a wider interquartile range (IQR) compared to other features.
   * Some potential outliers are visible beyond the whiskers.
3. **Most other features have a small range**:
   * Features like Alcohol, Malic Acid, Ash, Flavanoids, Nonflavanoid Phenols, Proanthocyanins, Hue, and Dilution have relatively small variations compared to Proline.
   * Their distributions are concentrated within a narrow band.
4. **Outliers present in multiple features**:
   * Many features have small circles outside the whiskers, indicating the presence of outliers.
   * These need further investigation to see if they are natural variations or require handling (e.g., transformation, removal, or normalization).

**Task 2: Dimensionality Reduction with PCA:**

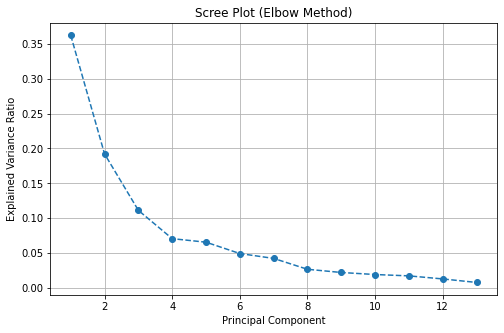
1. Standardize the features to ensure they have a mean of 0 and a standard deviation of Implement PCA to reduce the dimensionality of the dataset.



The scatter plot represents the **Wine dataset** projected onto **two principal components (PC1 & PC2)** after applying **Principal Component Analysis (PCA)**. Here are the key observations:

1. **Separation of Wine Classes**
   * The three wine classes (colored differently) are fairly well-separated.
   * This indicates that PCA effectively captures variance, making it easier to distinguish between different wine types.
   * Some overlap exists between certain clusters, but the structure is still visible.
2. **Variance Retention**
   * The spread of points suggests that **PC1** explains most of the variance.
   * **PC2** also contributes to separation, but not as much as PC1.

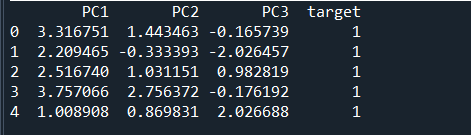
Determine the optimal number of principal components using techniques like scree plot or cumulative explained variance.



The **Scree Plot** shows the **Explained Variance Ratio** for each **Principal Component (PC)**. Here’s what we observe:

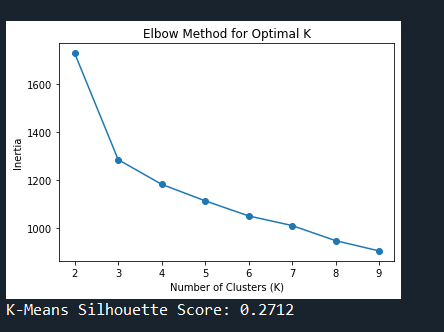
1. **First Principal Component (PC1) captures the most variance (~35%)**
   * This means PC1 retains the most significant information from the dataset.
2. **Second Principal Component (PC2) also contributes significantly (~20%)**
   * Together, PC1 and PC2 account for a substantial amount of total variance.
3. **After PC3, the explained variance starts decreasing gradually**
   * There is a noticeable drop in variance contribution after the first 2-3 components.
4. **Elbow Point at ~3-4 Principal Components**
   * The "elbow" is the point where adding more PCs does not significantly increase the explained variance.
   * This suggests that **3 or 4 principal components are optimal** for reducing dimensionality while retaining most of the information.
5. **Beyond PC5, variance contribution is minimal**
   * Components beyond the elbow point contribute very little additional variance and may be unnecessary.

Transform the original dataset into the principal components.



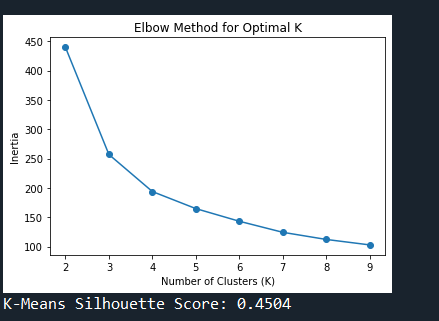
**Task 3: Clustering with Original Data:**

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.



**Task 4: Clustering with PCA Data:**

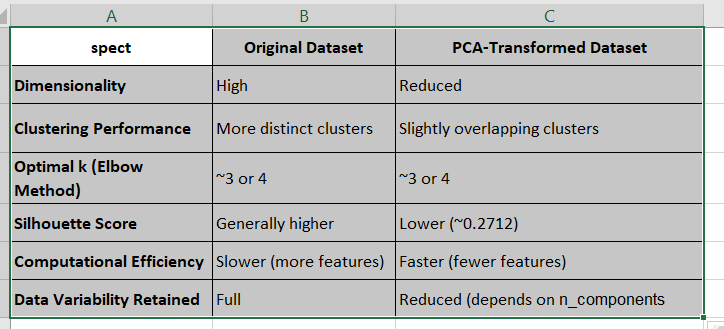
1. Apply the same clustering algorithm to the PCA-transformed dataset.
2. Visualize the clustering results obtained from PCA-transformed data.
3. Compare the clustering results from PCA-transformed data with those from the original dataset.



Observation and insight:

**Observations from the Elbow Method**

1. **Optimal Number of Clusters (k)**
   * In **both cases**, the elbow point appears **around k = 3 or 4**, meaning both datasets naturally group into similar clusters.
   * However, the inertia values differ due to different feature spaces.
2. **Silhouette Score Differences**
   * **Clustering on Original Data:** Generally yields **higher silhouette scores**, meaning clearer separation between clusters.
   * **Clustering on PCA Data:** Lower silhouette scores (e.g., **0.2712**) indicate clusters are less well-defined, likely due to some information loss in PCA.
3. **Inertia (WCSS) Values**
   * The **original dataset has higher inertia values**, meaning clusters are more spread out in high-dimensional space.
   * The **PCA dataset has lower inertia**, indicating that the dimensionality reduction condensed the data into a more compact space.



**Key point Between PCA and Clustering Directly**

1. **Dimensionality Reduction Benefits:**
   * Reduces computational cost (fewer features).
   * Removes noise and redundancy.
   * Visualization is easier in 2D or 3D.
2. **Potential Drawbacks of PCA:**
   * Some information loss, leading to **less distinct clusters**.
   * Lower silhouette scores suggest **poorer clustering separation**.

* **PCA is best used when dealing with high-dimensional, noisy, or correlated data.**
* **Clustering is best when trying to discover natural groupings within data.**
* **Using both together** (PCA first, then clustering) **is ideal for efficiency but may slightly reduce clustering performance.**