SIT720 Machine Learning

Task 4

Michael Rideout
Student Id: 225065259

Item 1.1 Main Points Summary

Week 3 Main Points Summary - Clustering Concepts

Clustering is the method of grouping data points based on a similarity or distance metric / measure.

Distance Metrics

A distance metric measures the similarity or distance between data points. Some distance metrics include:

Manhattan distance: $$d_{Cityblock}(x_i, x_j) = \sum_{k=1}^{D} |x_{i,k} - x_{j,k}|$$

Euclidean distance: $$d_{Euclidean}(x_i, x_j) = \sqrt{\sum_{k=1}^{D} (x_{i,k} - x_{j,k})^2}$$

Chebyshev distance: $$d_{\text{Chebyshev}}(x_i, x_j) = \max(|x_{i,1} - x_{j,1}|, |x_{i,2} - x_{j,2}|, \ldots, |x_{i,D} - x_{j,D}|)]$$

Minkowski distance: $$d(x, y) = (\sum_{i=0}^{n-1} |x_i - y_i|^p)^{1/p}$$

Cosine distance: $$d_{Cosine}(x_i, x_j) = 1 - \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2}$$

Mahalanobis distance: $$d_{Mahalanobis}(x_i, x_j) = \sqrt{(x_i - x_j)^T M^{-1} (x_i - x_j)}$$

Jaccard distance: $$d_{Jaccard}(x_i, x_j) = 1 - \frac{|x_i \cap x_j|_1}{|x_i \cup x_j|_1}$$

Clustering Algorithms

Clustering algorithms use distance metrics to group together similar datapoints into potentially interesting clusters.
Some clustering algorithms include:

K-means:
Is a clustering algorithm which performs the following steps:

1. Initialise k centroids
2. Assign each datapoint to the closest centroid
3. Recalculate the centroids to be the mean of all data points in each cluster
4. Repeat 2 and 3 until cluster assignments do not change substantially

Cluster Evaluation

- Two categories exist to evaluate clusters:

External Assessment - Compares the assigned cluster to a known ground truth cluster.
Internal Assessment - Evaluates the quality of the clustering based on intrinsic properties of the clusters themselves.

Rand Index - measures how similar two clusters are.
Purity - is a measure of how well a cluster algorithm's output matches some ground truth
Mutual Information - measures the consensus of two clustering assignments
Silhouette Coefficient - Measure the degree of similarity an instance is to its cluster as opposed to other clusters

Limitation of K-means

K-means has several limitations:

1. Results can vary due to random initialisation
2. Number of clusters needs to be specified
3. Has difficulty with arbitrary shapes
4. Sensitive to noisy data

Week 4 Main Points Summary

Eigenvalues and Eigenvectors

Are tools that aid in the investigation of linear transforms. Attributes of them are:

- Defined as pairs (λ, u) satisfying $Au = \lambda u$ for a square matrix A
- A d x d matrix has d eigenvalue/eigenvector pairs
- The number of non-zero eigenvalues equals the matrix rank
- Eigenvectors form an orthogonal matrix U.
- Finding eigenvalues involves solving the characteristic polynomial $\det(A - \lambda I) = 0$
- Eigenvectors are found by solving $(A - \lambda I)u = 0$ for each eigenvalue.

Singular Value Decomposition

A way to break down a matrix into three matrices
Attributes of the method are:

- Decomposes a matrix X into $X = USV^T$, where U and V are orthogonal matrices and S is a diagonal matrix of singular values.
- Singular values (σ_i) are the square roots of eigenvalues from XX^T or X^TX.
- Eigenvectors of XX^T form U, and eigenvectors of X^TX form V.
- SVD represents data in a coordinate system where the covariance matrix is diagonal

Curse of Dimensionality

Highly dimensional data is common in areas like text, image and genomic data. Increased dimensionality can cause exponential increase in the size of data. In high dimensions, more data points reside near the surface of a hyperplane. Distances between points become less distinct making clustering less effective. Dimensionality reduction aims to mitigate these effects whilst preserving information

Principal Component Analysis (PCA)

The goal of PCA is to summarise correlated high-dimensional data using a smaller set of uncorrelated variables called principal components. These components are linear combinations of the original dimensions, sorted by the amount of variance they capture.

Formulation (Maximising Error)

- Find the direction (eigenvector) that maximises the variance of the project data
- Leads to an eigenvalue $Cu_i = \lambda_i u_i$,
- Subsequent principal components are found similar to the above, maximising variance while being orthogonal to the previous principal components

Formulation (Minimising Error)

An alternate method which minimises the reconstruction error when projecting data onto a lower k-dimensional subspace.

1.2 Summary of Reading List Items

Week 3 Readings

k-means++: The Advantages of Careful Seeding by Arther and Vassilviskii.

Article - k-means++: The Advantages of Careful Seeding

This article discussed improvements to the k-means algorithm. It proposed an enhanced seeding technique that involved choosing initial cluster centers sequentially with specific probabilities related to the points' distances from existing centers.

Video - How Does The DBSCAN Algorithm Work

Video is no longer publicly accessible

Video - Spectral Clustering and How It Works
An introduction to spectral clustering. It is a clustering technique that doesn't assume specific cluster shapes, handles intertwined data well and avoids the iterative process and sensitivity to initialisation.

Week 4 Readings

Video - Eigenvectors and eigenvalues
Video explaining how to find principal components in PCA using linear algebra. This is achieved by finding eigenvalues and eigenvectors of the covariance matrix.

Video - Lecture: The Singular Value Decomposition (SVD)
A lecture on Singular Value Decomposition. It explains the matrix multiplications that fundamentally involve the rotation and stretching of vectors. The method to compute the SVD using eigenvalue decomposition is explained.

Video - StatQuest: Principal Component Analysis (PCA)
A step by step guide to principal component analysis using singular value decomposition. Demonstrates how PCA can reduce the dimensionality of data while retaining important information.

Video - PCA 3: direction of greatest variance
This video explains the significations of the components in PCA, that being the first component is the direction of the greatest variance, the second orthogonal to the first and has the next greatest variance and so on.

Video - PCA 4 : principal components = eigenvectors
Video about PCA explaining that principal components are eigenvectors of the covariance matrix.

Video - finding eigenvalues and eigenvectors
A video about eigenvectors and eigenvalues. The core idea is that eigenvectors are special vectors which, when transformed by a matrix, only scale and don't rotate. It explains how to compute eigenvectors and eigenvalues.

Python Libraries

Python libraries utilised in this task include:

- **Pandas:** Data manipulation and analysis library
- **Numpy:** Library for scientific and numerical computing
- **Mathplotlib:** Visualisation library
- **Seaborn:** Visualisation library
- **Sklearn:** Machine learning toolkit
- **Mplot3d:** Generates 3D plots

- **Yellowbrick:** ML visualisation library
- **SciPy:** Scientific library
- **Kneed:** Library to detect elbow points in a curve

1.3 Learning Reflection

Weeks 3 and 4 have provided a foundational basis for the key machine learning concepts of clustering and dimensionality reduction.

I learnt that clustering the process of grouping similar data points together. From that the focus was on what is ‘similarity’ and that was described by the concept of distance metrics. The k-means clustering algorithm was explained in detail as it is currently the most popular clustering technique. Cluster evaluation was also described in detail as it is important to not only generate clusters, but to know how to compare them to ground truths or determine inherent characteristics of clusters.

The lecture on dimensionality reduction introduced the eigenvalues and eigenvectors and the mathematical underpinnings for these, that being linear algebra. Singular Value Decomposition was also described as a powerful matrix factorisation technique that is useful in uncovering underlying structures in data. Principal Component Analysis was presented as a means to reduce dimensionality by finding new, uncorrelated variables that capture the maximum variance in the data.

1.4 Quiz Results

Week 3 Quiz:

Week 3 quiz SIT 720

Your work has been saved and submitted

Written 18 April, 2025 8:36 AM - 18 April, 2025 8:39 AM • Attempt 2 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

Attempt Score	**9 / 10 - 90 %**
Overall Grade (Highest Attempt)	**9 / 10 - 90 %**

Week 4 Quiz:

Week 4 Quiz for SIT720

Your work has been saved and submitted

Written 18 April, 2025 2:47 PM - 18 April, 2025 2:58 PM • Attempt 1 of unlimited

Your quiz has been submitted successfully, the answer(s) for the following question(s) are incorrect.

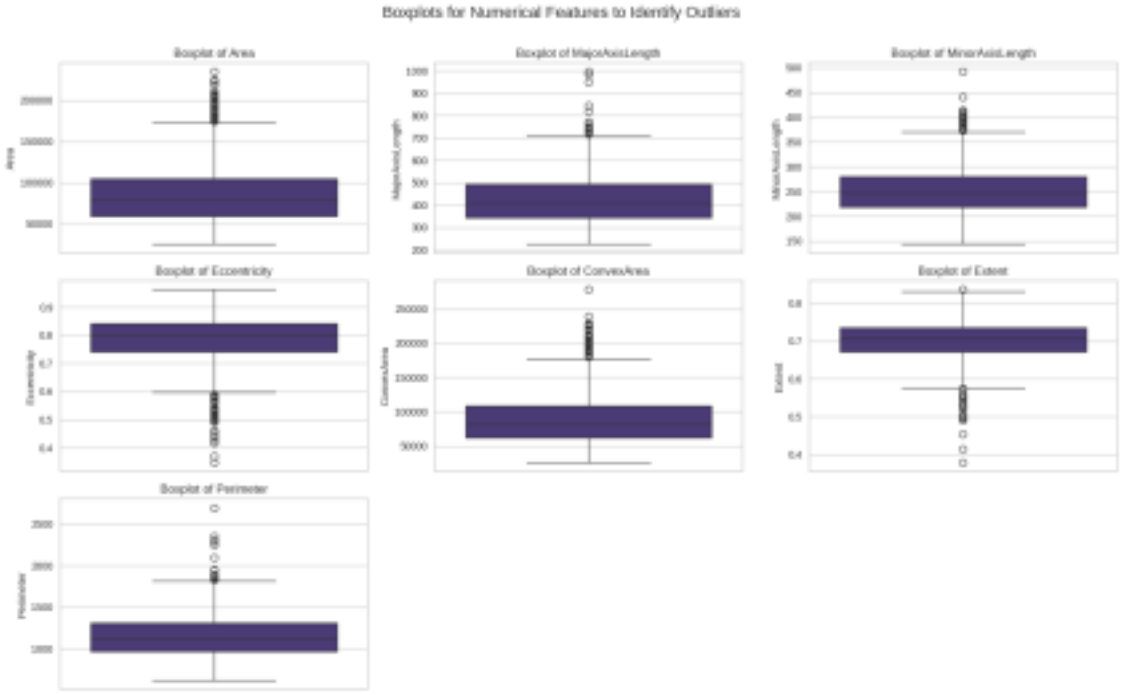
Attempt Score	**9 / 10 - 90 %**
Overall Grade (Highest Attempt)	**9 / 10 - 90 %**

Task 4 Jupyter Notebook Graph Outputs

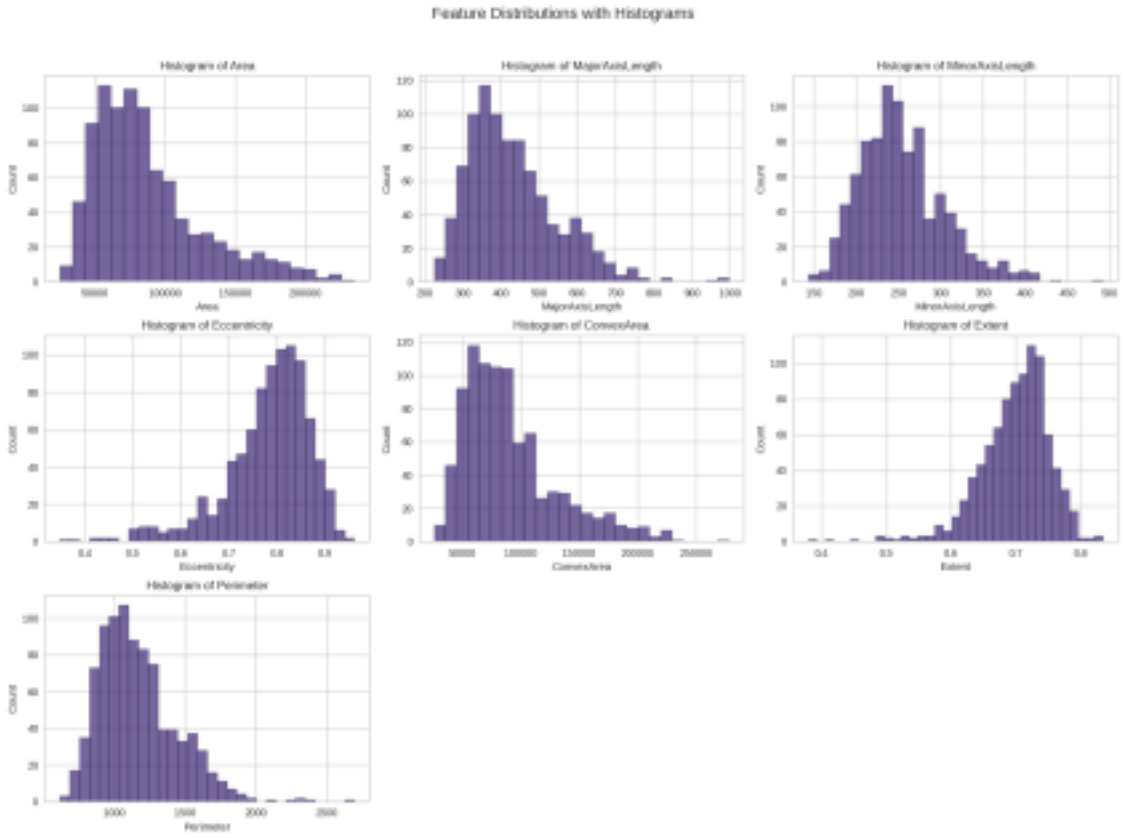
April 23, 2025

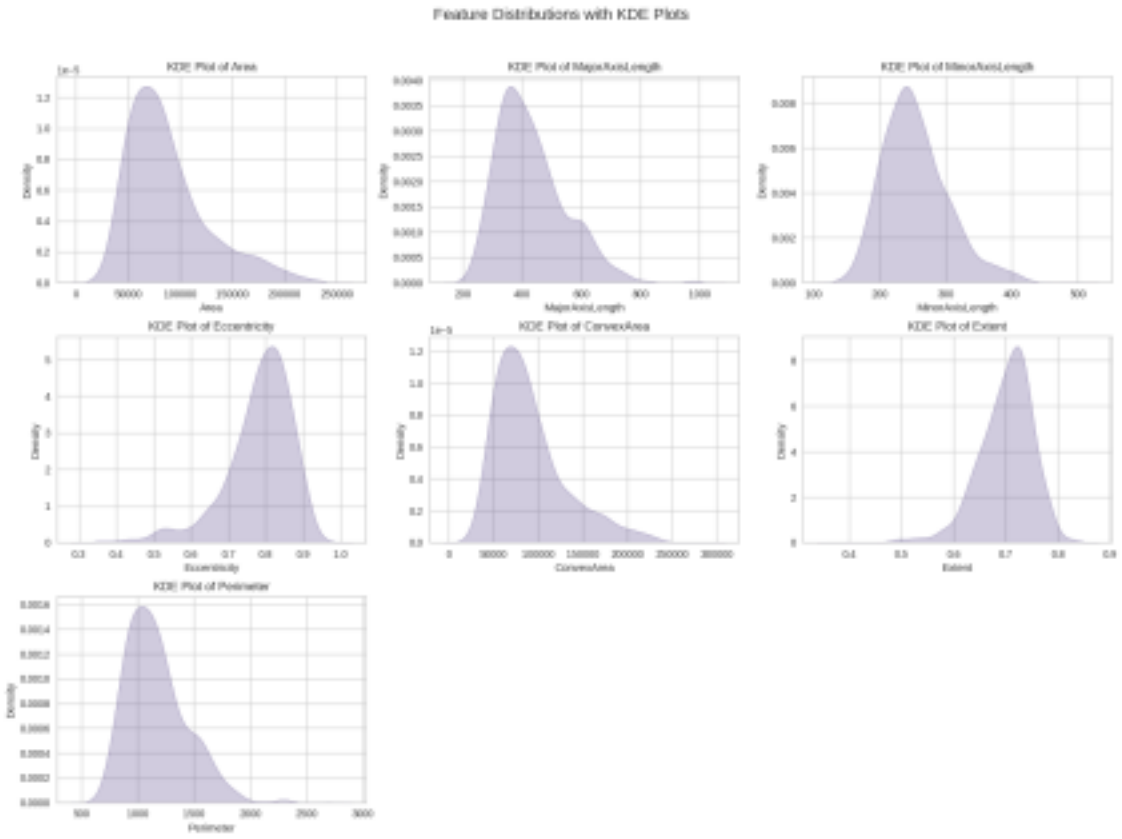
1 Task 1 Graphs

Outlier Visualisation



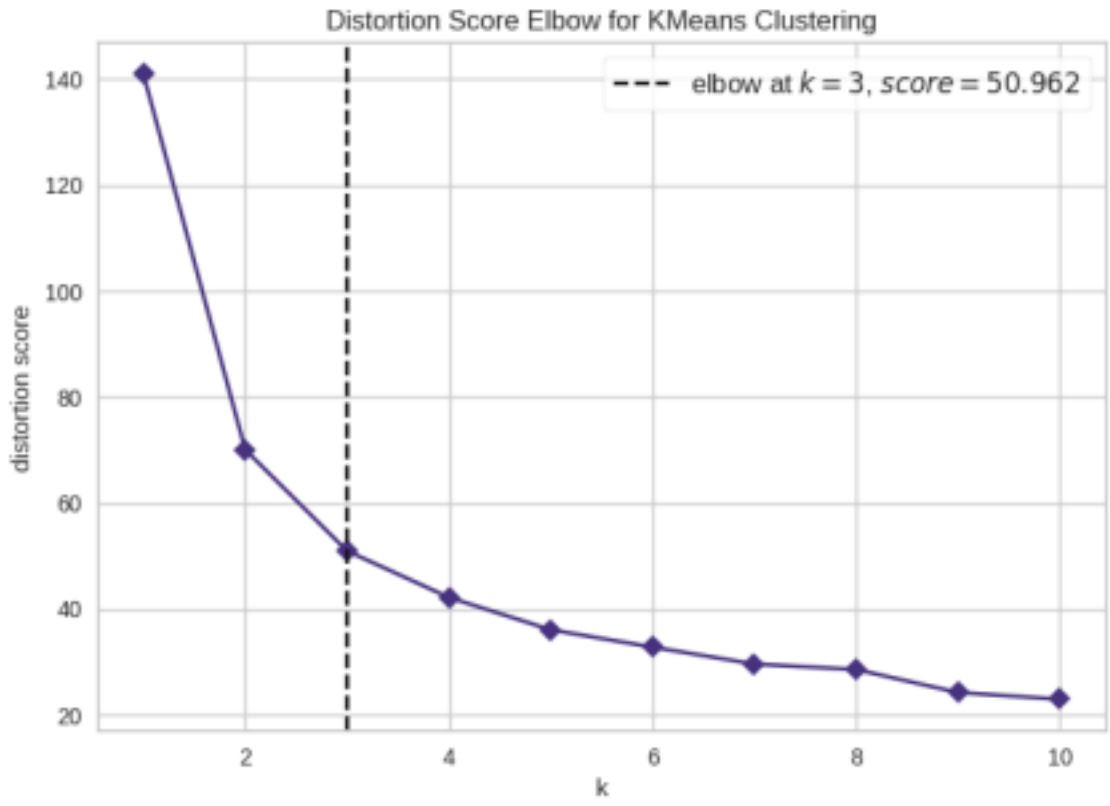
Distribution Visualisations



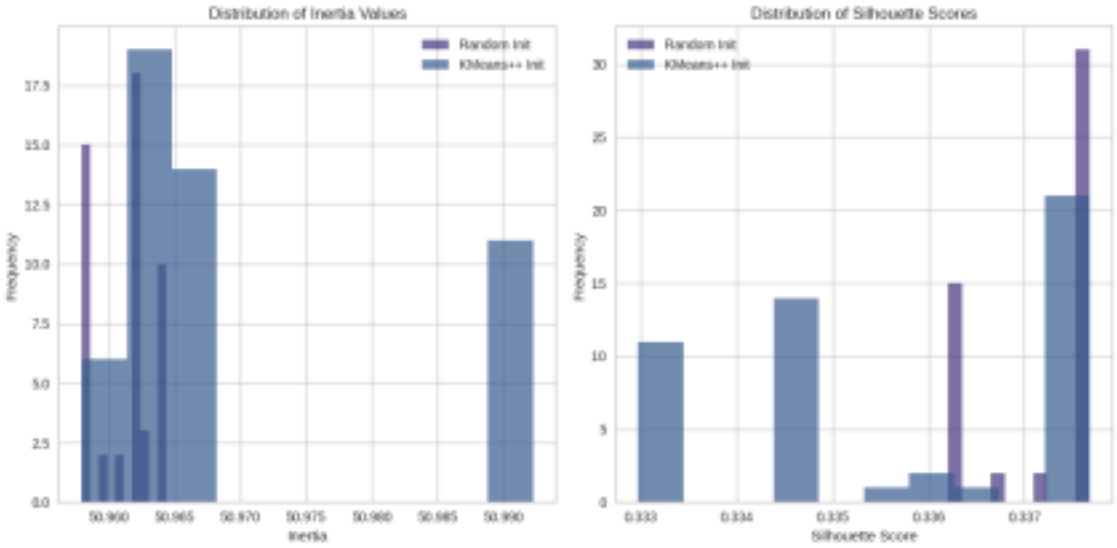

Feature Distributions with KDE Plots
KDE Plot of Area
KDE Plot of MajorAxisLength
KDE Plot of MinorAxisLength
KDE Plot of Eccentricity
KDE Plot of ConvexArea
KDE Plot of Extent
KDE Plot of Perimeter
Density
Area
MajorAxisLength
MinorAxisLength
Eccentricity
ConvexArea
Extent
Perimeter

Task 2 Plots

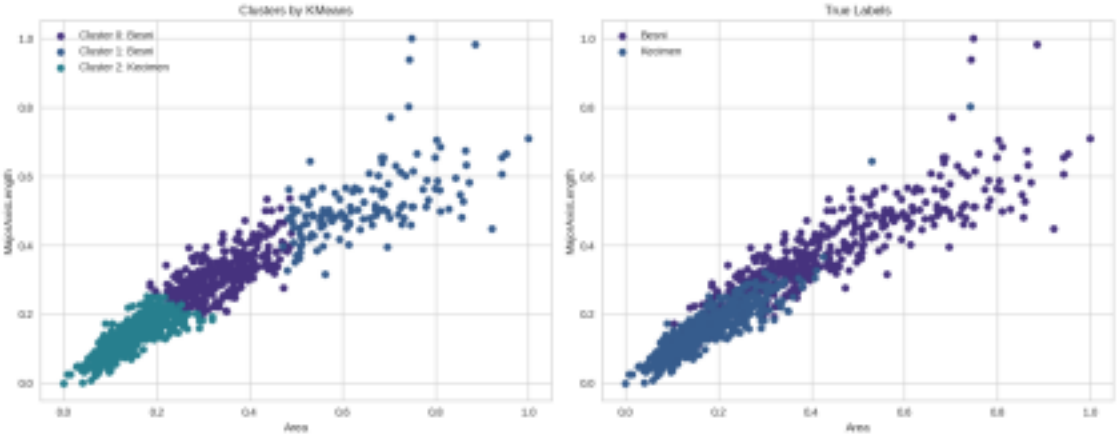
Kmeans Elbow Plot



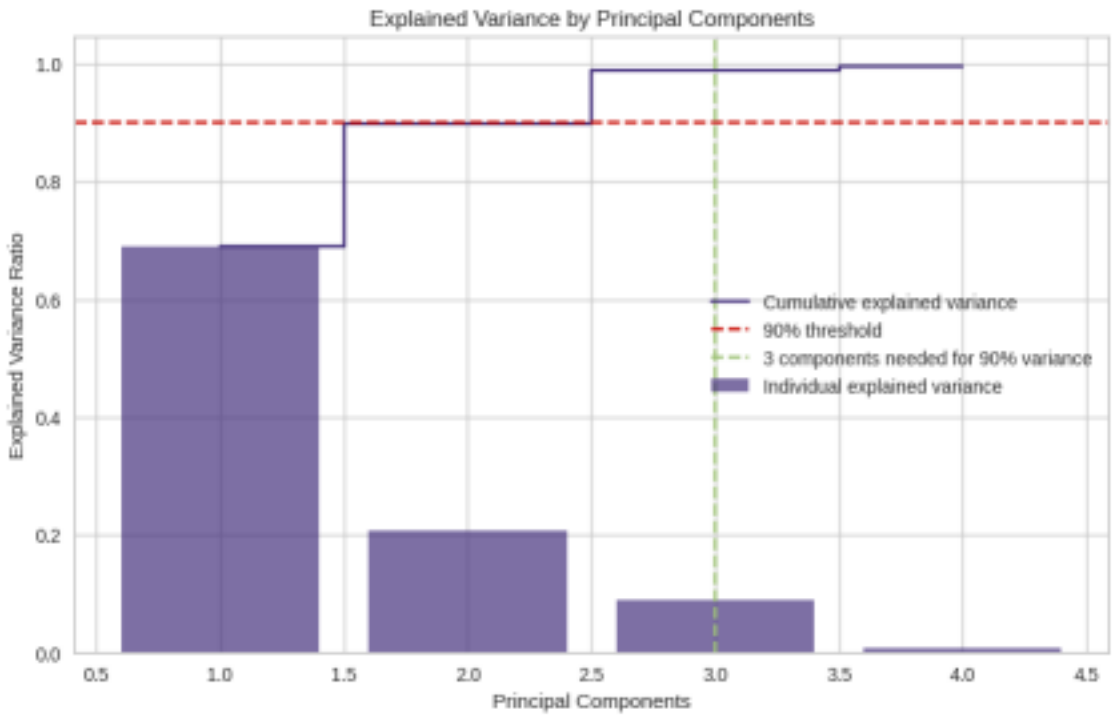
3 Task 3 Metric Distributions Plot

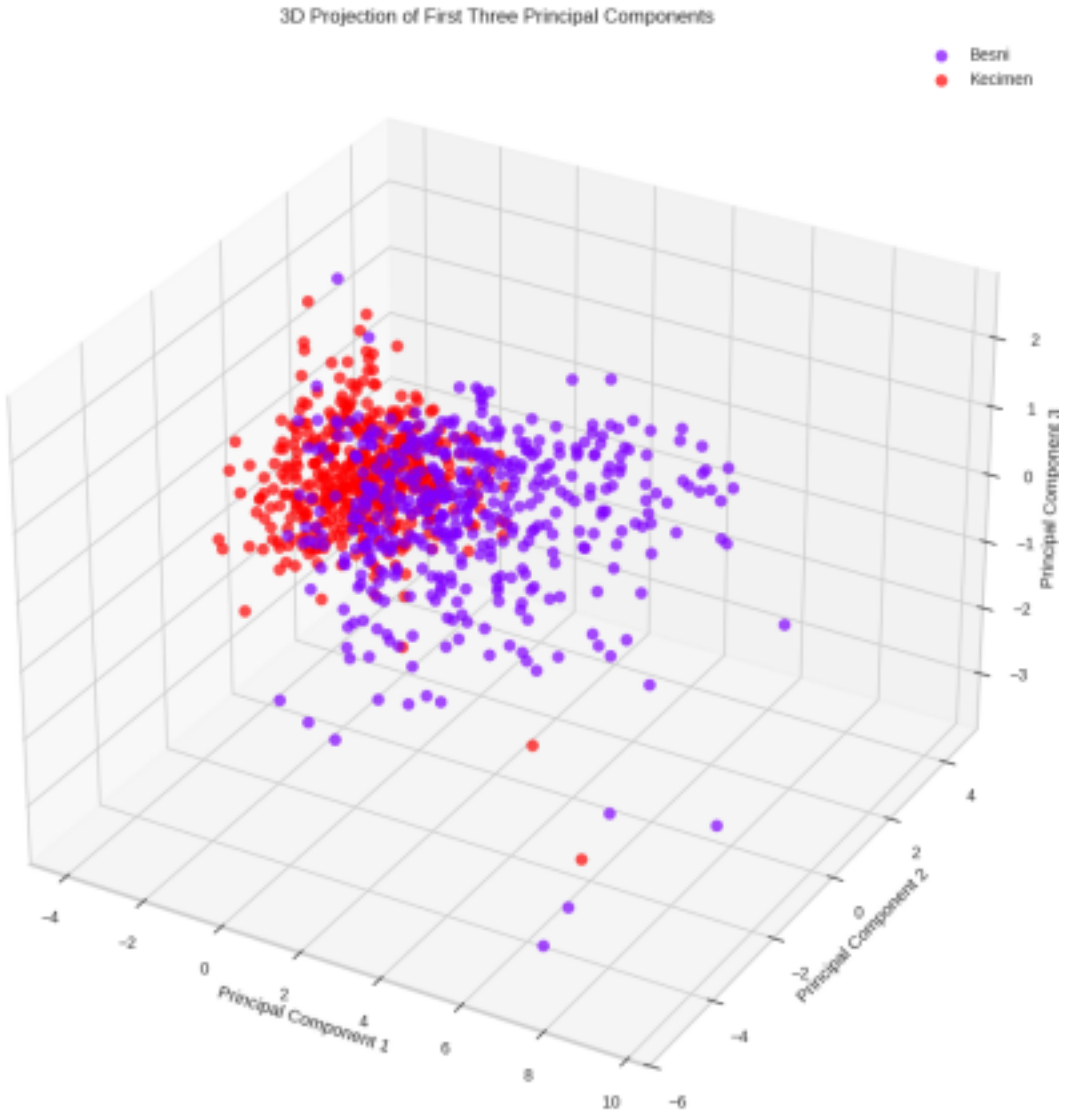


4 Task 4 Cluster Evaluation Plots

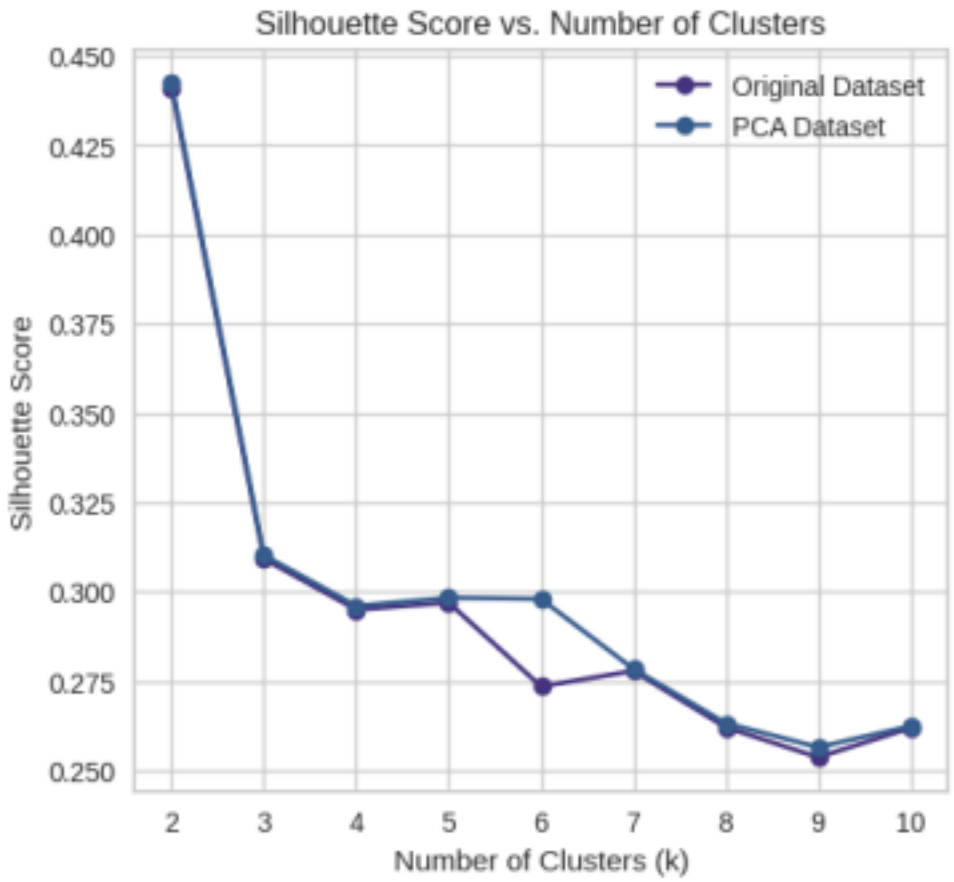


Task 5 PCA Plots




3D Projection of First Three Principal Components
Besni
Kecimen
Principal Component 1
Principal Component 2
Principal Component 3

Task 7 Cluster vs Silhoutte Score Plot



t-SNE Visualisation