Name: Arlene Antony D'costa
Seating Pin: 79
SUID: 594303899

CIS 662 HW2 Report

The task I have selected is **h-index prediction** and the dataset is **71-80.csv**.

The colums considered for the same are: *univ_rank , cit_2017, cit_2018 ,cit_2019 ,cit_2020 cit_2021,cit_2022*

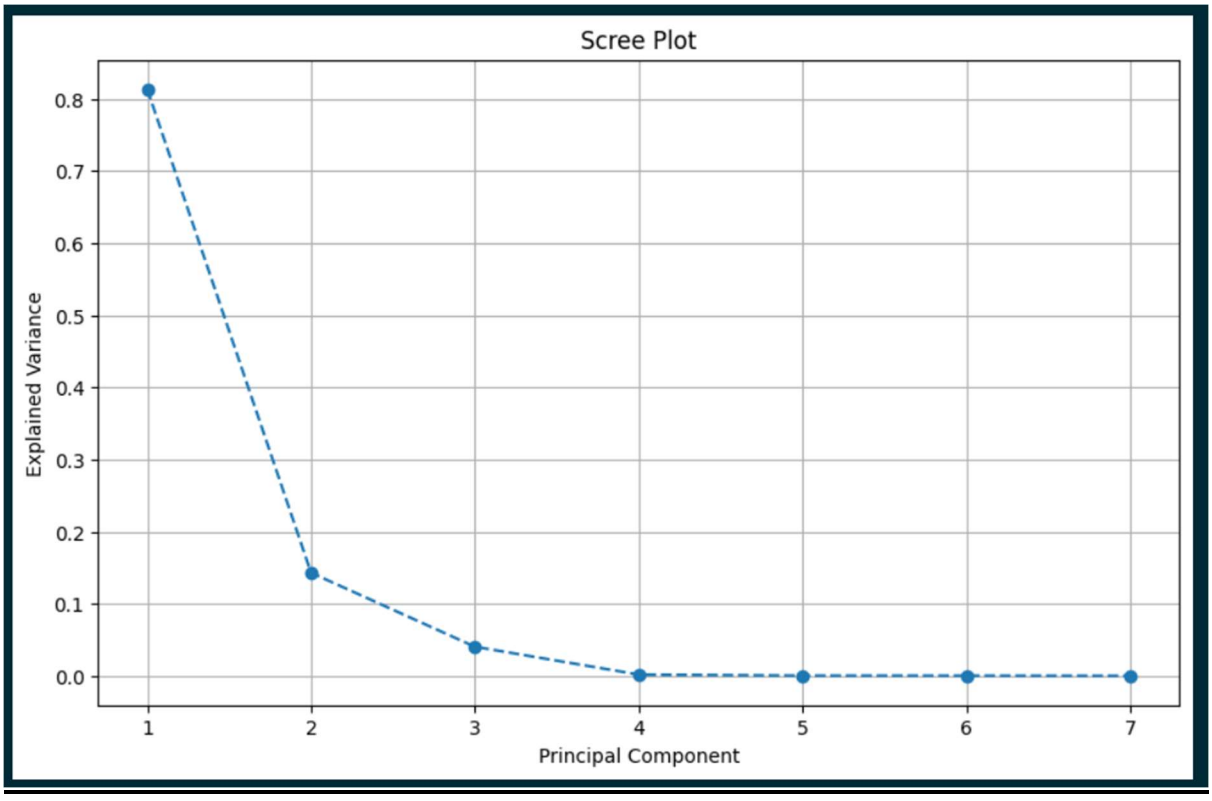
Data is normalised using **Standard Scaler**.

Inorder to determine the number of correct components for PCA that will capture the maximum variance for prediction tasks we use **Scree Plot**.

A scree plot is a valuable tool in factor analysis and PCA to help researchers decide how many factors or principal components to include in their analysis, striking a balance between model complexity and the ability to capture meaningful variation in the data.

It provides a visual representation of the eigenvalues, which assists in identifying an appropriate cutoff point for dimensionality reduction.

OUTPUT:



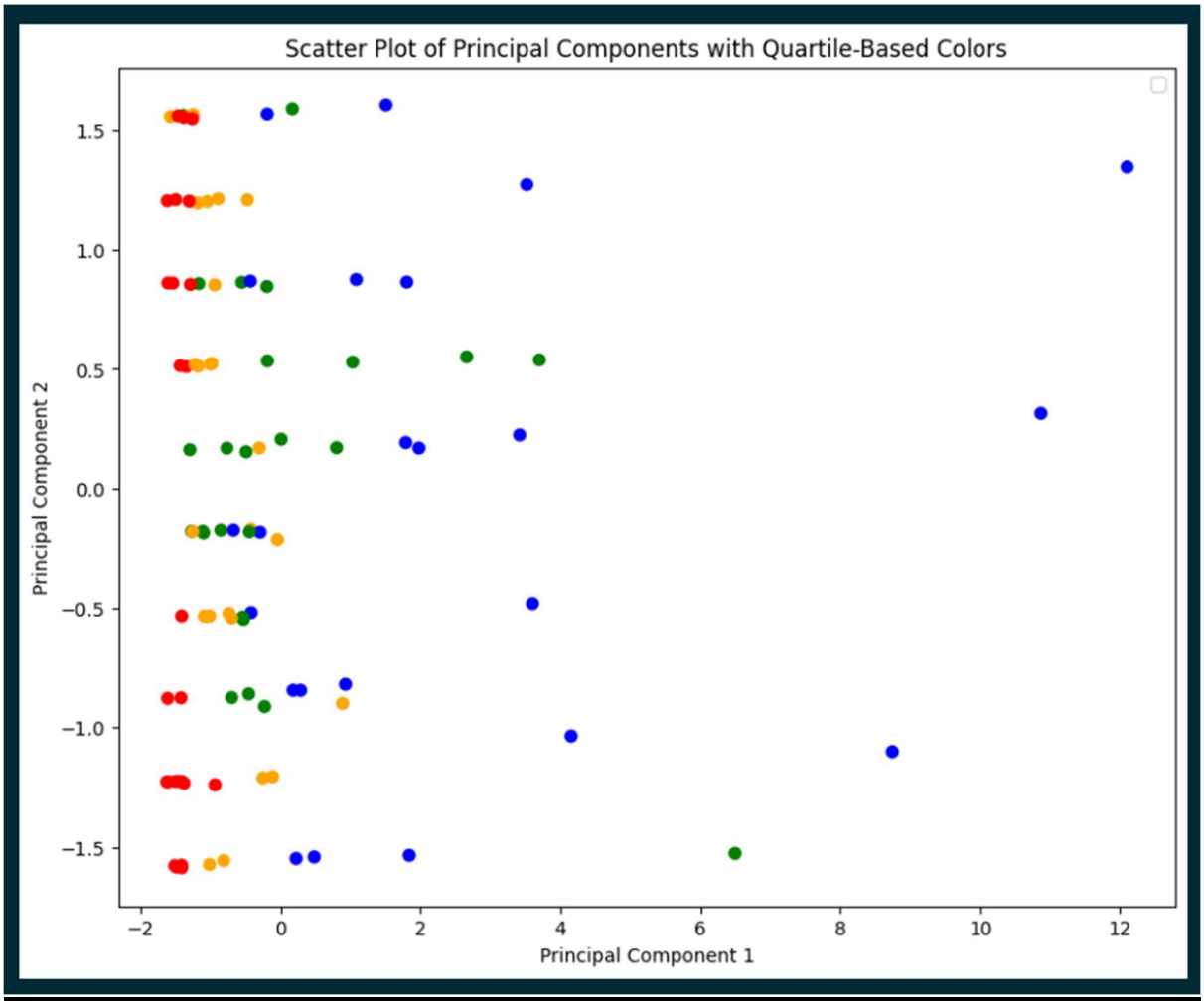
RESULTS/CONCLUSION:

Point of inflexion can be referred as the point that captures the maximum slope. Here no components that should be provided for Dimensionality Reduction with PCA is 2, since the inflection point is at x = 2.

Name: Arlene Antony D'costa
Seating Pin: 79
SUID: 594303899

Analysing Scatter Plot for PCA Output:

OUTPUT:



RESULTS:

In the above output, we can analyse that PCA Component 1 captures the maximum variation in the h-index target data which is explained below.

There is no pattern observed for PCA2 as the data points of different quartiles are distributed along the Y-axis. Hence it can be concluded that PCA2, does not have much impact in capturing variation essential for target data.

The quartile distribution is based on the frequency of h-index and helps to categorize the data based on the frequency as well as identify potential outliers.

Lower quartile data which is Red in colour is clustered around lower values of PCA1. As the PCA1 value increases, higher quartile-valued data points are mapped in accordance with the PCA1 values.

A clear separation does not exist among the PCA data points from different quartiles as they are not forming distinct groups or clusters along PCA1. Since PCA1 is not effectively separating data points into distinct quartiles, it suggests that the most important source of variation for the captured data points may not be strongly associated with quartile membership.

PCA1 explains the variation in the initial quartiles well, and it helps distinguish faculty members with different h-index quartiles in this region.

The scattering of quartile data as PCA1 values increase could be due to various factors:

- Noise in the data: Higher PCA1 values may introduce more variability that is unrelated to quartile membership, leading to increased scatter.
- Heterogeneity: The data points in the higher PCA1 range may be more diverse in terms of h-index quartile membership, making it challenging for PCA1 to separate them effectively.
- Non-linear relationships: PCA captures linear relationships between variables. If the relationship between PCA1 and quartile membership becomes non-linear at higher values, this can lead to increased scatter.

Understanding which factors drive the initial correlation can provide valuable insights into the factors that distinguish quartiles. This can be done using a **correlation matrix** or different algorithms such as **Mutual Inclusion**.

CONCLUSION:

The features mapped along PCA1 as opposed to PCA2 can also be regarded as dominant features essential in capturing explained variance for target variable essential for h-index prediction. Data points for the initial h-index quartiles are positively correlated with the PCA1 values, however, as the PCA1 component value increases the quartile data tends to be scattered. It implies that PCA1 is weak positively correlated with higher quartiles.

Thus, PCA1 effectively captures some patterns or relationships within the initial quartiles but loses its ability to explain or discriminate higher quartile values. PCA2 is ineffective in capturing variance essential to the h-index.