HW2: Predicting 2022 citation numbers using the university rank and 2017-2021 citation numbers.

Dataset used: 41-50.csv

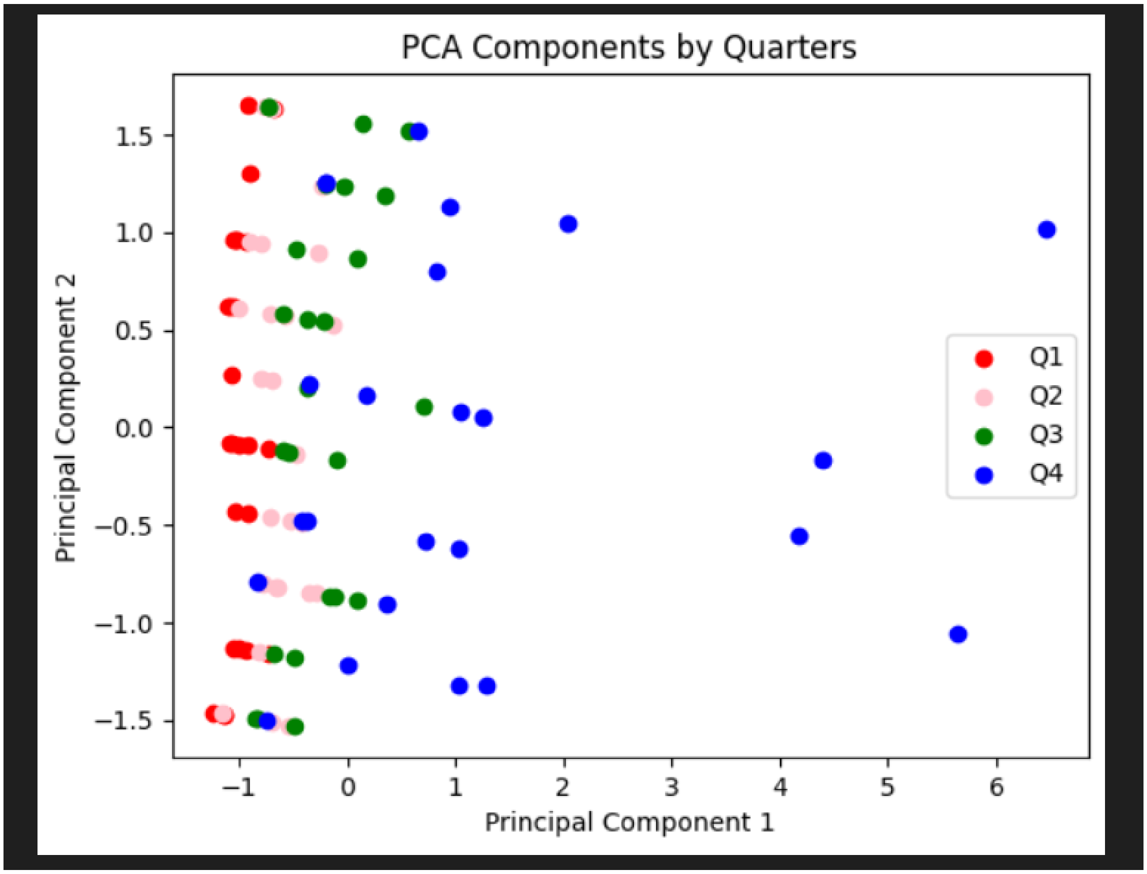
PCA for Dimension Reduction: PCA is a dimensionality reduction technique that helps simplify the dataset by transforming the original features into a new set of uncorrelated variables called principal components.

In the provided code, we used PCA to reduce the dimensionality of the dataset from six features (cit_2017, cit_2018, cit_2019, cit_2020, cit_2021) to two principal components (PC1 and PC2).This dimension reduction is valuable for data visualization and simplifying modeling, as it reduces the number of features and allows to work with a lower-dimensional dataset.

Scatter Plot Analysis: The scatter plot is a visual representation of data points in a two-dimensional space, with colors indicating the quarters (Q1, Q2, Q3, Q4) of 2022 citation numbers.

Observing the scatter plot, we can see that there is no clear separation or clustering of data points based on the quarters. Data points from all quarters appear overlapped and scattered throughout the plot. This suggests that the two principal components (PC1 and PC2) do not have a strong ability to discriminate or separate the quarters based on the 2022 citation numbers.

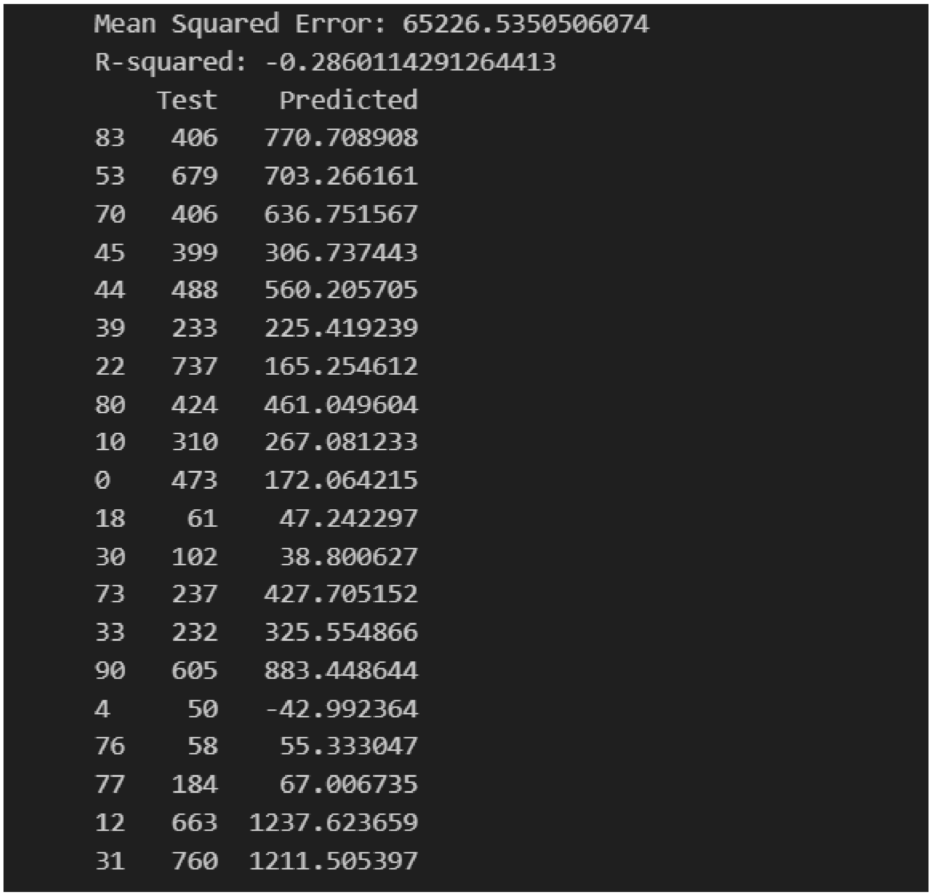
Output for the given data:



HW2: Predicting 2022 citation numbers using the university rank and 2017-2021 citation numbers.

Prediction of values using linear regression model:

Linear regression assumes a linear relationship between the independent variables (features) and the dependent variable (target). In this case, a more complex or non-linear model might be more appropriate for predicting 2022 citation numbers accurately.



```
Mean Squared Error: 65226.5350506074
R-squared: -0.2860114291264413
    Test    Predicted
83   406   770.708908
53   679   703.266161
70   406   636.751567
45   399   306.737443
44   488   560.205705
39   233   225.419239
22   737   165.254612
80   424   461.049604
10   310   267.081233
0    473   172.064215
18    61    47.242297
30   102    38.800627
73   237   427.705152
33   232   325.554866
90   605   883.448644
4     50   -42.992364
76    58    55.333047
77   184    67.006735
12   663  1237.623659
31   760  1211.505397
```

Mean Squared Error (MSE): The calculated MSE for the linear regression model is quite high (approximately 65,226.54), indicating a significant difference between the predicted and actual values. **R-squared (R2) Score:** The R-squared score is negative (-0.286), which indicates that the model does not fit the data well and may perform worse than a horizontal line.

In summary, while PCA is a valuable technique for dimensionality reduction and data visualization, its application to this specific prediction task did not lead to favorable results. The scatter plot confirmed that data points from different quarters were not clearly separable in the two-dimensional PCA space. To improve prediction accuracy, we should consider more advanced modeling techniques or feature engineering, as the relationship between features and the target variable seems to be non-linear and complex.

References: https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

https://www.youtube.com/watch?v=fkf4IBRSeEc