CS F320 Foundations of Data Science

Assignment – 2A

Implementing PCA from Scratch and Applying it to Car Data

Group Details

	Name	ID
1	Manthan Patel	2021A7PS2691H
2	Teerth Patel	2021A7PS2090H
3	Shrey Paunwala	2021A7PS2808H

Principal Component Analysis:

Principal Component Analysis (PCA) is a dimensionality reduction technique commonly used in data analysis and machine learning. It works by transforming the original variables into a new set of uncorrelated variables called principal components. The first few principal components capture the maximum variance in the data, allowing you to represent the data in a lower-dimensional space.

Importing and cleaning data:

- The file used for the PCA Analysis is **“audi.csv”.** The file was imported using pandas library of python.
- After importing, we remove non-numeric data columns from the dataset like **“model”,” transmission” and “fuel-type”.**
- The column “price” is also removed from the dataset because it the target feature.
- Lastly, we normalise the data in each column by using the following:

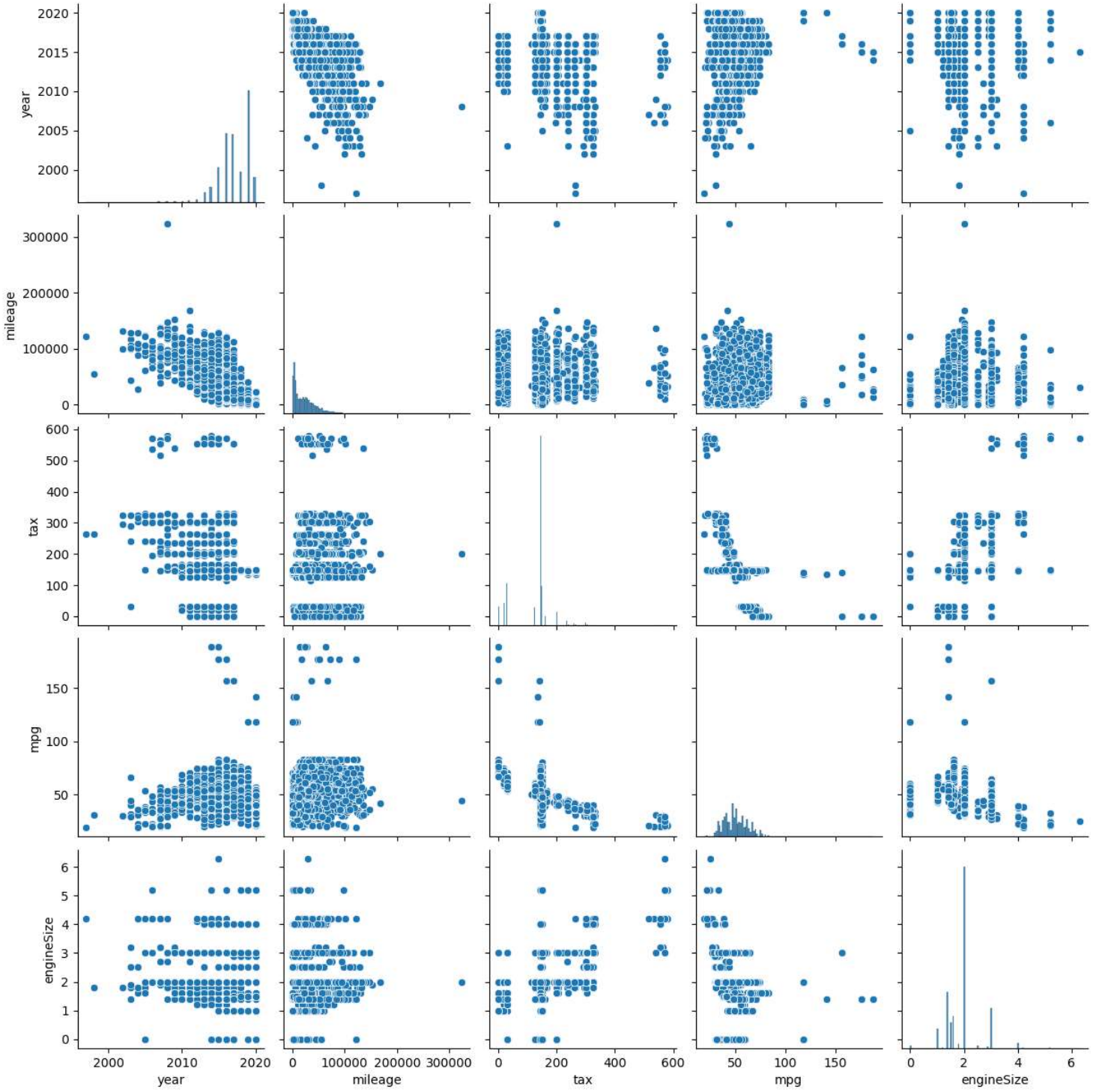
$$X' = \frac{X - \mu}{\sigma}$$

Where μ is the mean of values in a given feature and σ is the standard deviation of values of a given feature.

Interpreting and visualizing relationships between different features:

- Used pair-plots from seaborn library to visualize the data using the following:

```
sns.pairplot(numeric_data)
plt.show()
```



Interpreting Results from PCA Analysis:

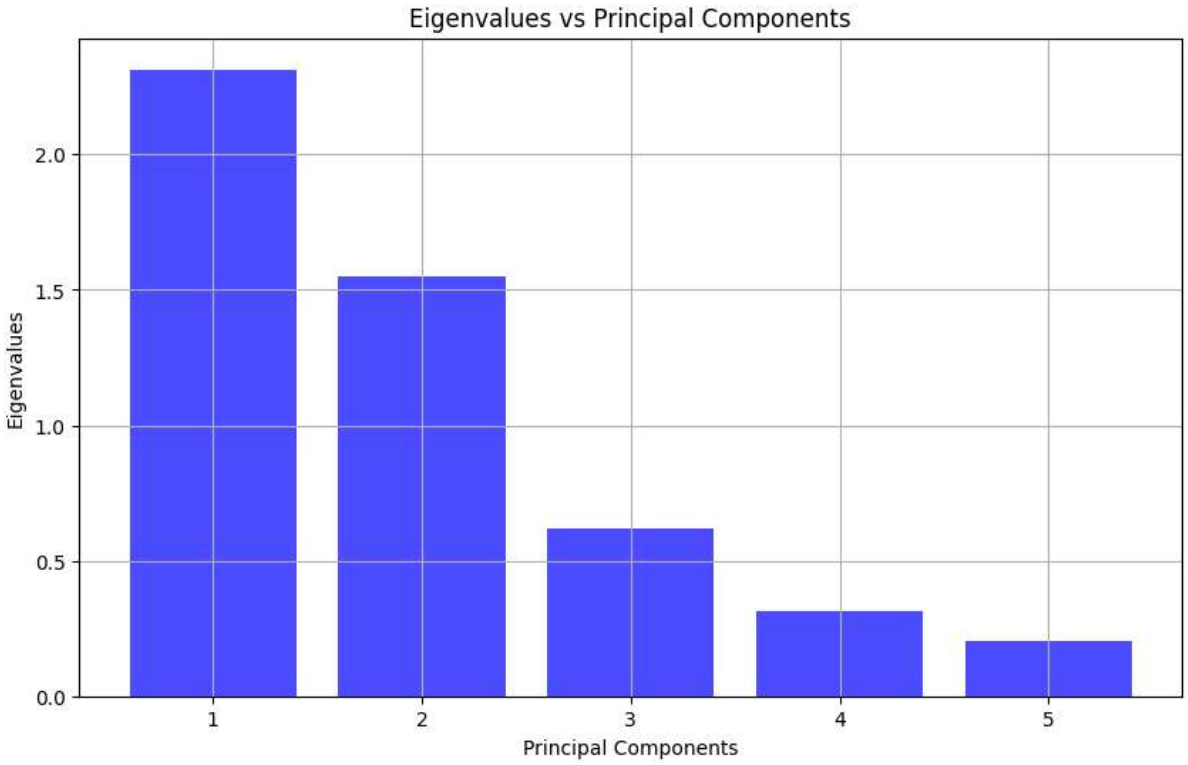
- **Calculating Eigenvalue and Eigenvectors from the co-variance matrix:**

  Co-variance matrix calculated using function: **np.cov()**

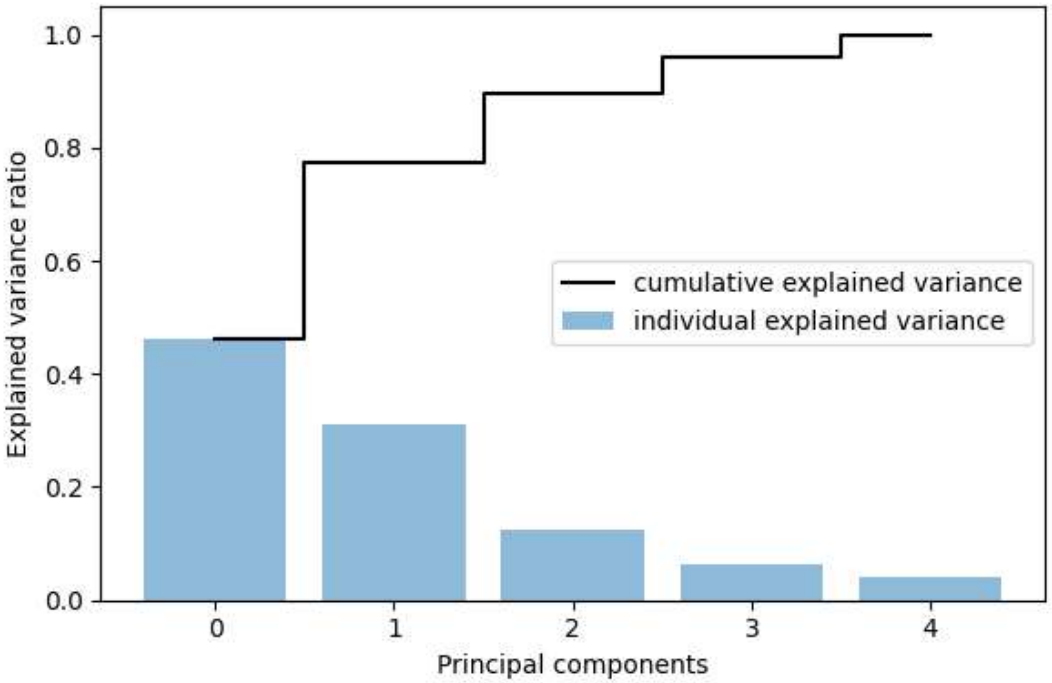
```
          0         1         2         3         4
0  1.000094 -0.789741  0.093075 -0.351314 -0.031585
1 -0.789741  1.000094 -0.166563  0.395140  0.070717
2  0.093075 -0.166563  1.000094 -0.635968  0.393112
3 -0.351314  0.395140 -0.635968  1.000094 -0.365655
4 -0.031585  0.070717  0.393112 -0.365655  1.000094
```

  Eigenvalues and Eigenvectors are computed using python function:

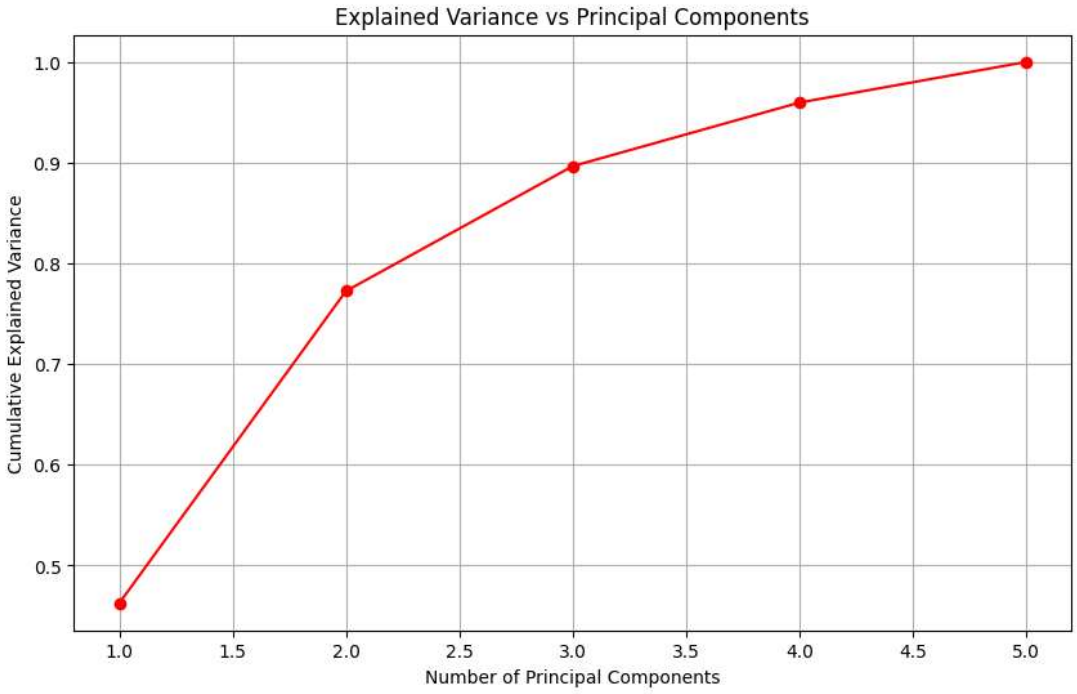
```
# Formulate and solve the eigenvalue-eigenvector equation
eigenvalues, eigenvectors = np.linalg.eig(cov_matrix)
```



- We get eigenvalues associated with each principal component. These represent the amount of variance captured by each component.
- The sum of all eigenvalues is equal to the total variance in the original data. You can calculate the **proportion of variance explained by each principal component** by dividing its eigenvalue by the total sum of eigenvalues.

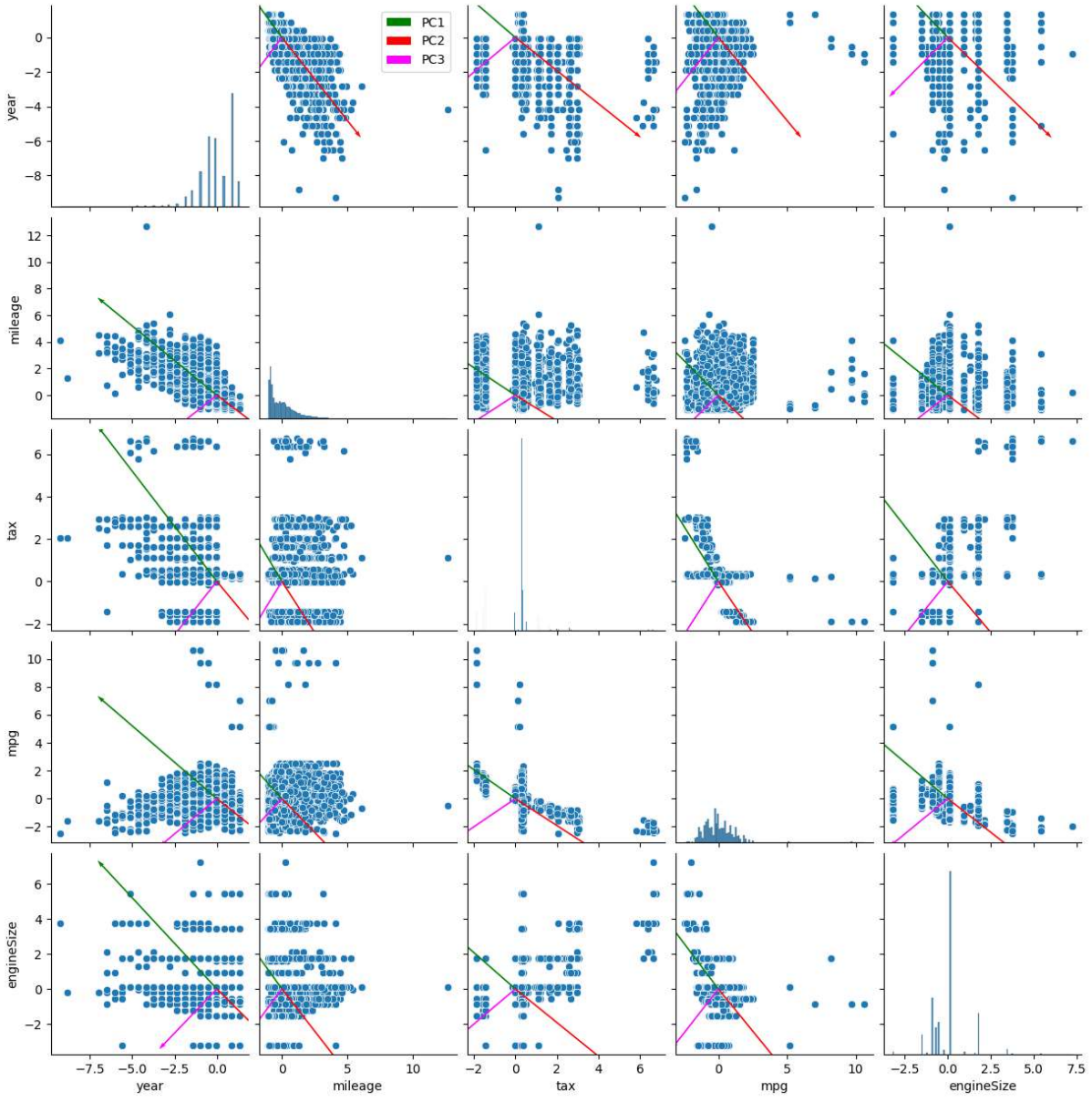


- Graph indicates the sequential increase in the explained variance. Using this result, we can now compute the value of k (**Top k principal components**). The value of k depends on the percentage of cumulative variance to be captured.
  The value of K is computed by taking **85% of the cumulative variance:**
  `k=np.argmax(explained_variance >= 0.85) + 1`



Number of Components	Cumulative explained variance	Individual explained variance
1	0.46259565	0.46259565
2	0.77253170	0.30993605
3	0.89653682	0.12400512
4	0.95964136	0.06310454
5	1.00000000	0.04035864

Projecting the principal components onto the Pair Plots:



- When eigenvectors associated with principal components are plotted on a pair plot, insights into variable contributions and relationships emerge. Longer vectors indicate greater variable influence, with similar directions signifying positive correlations. Clusters of similarly directed vectors suggest variable groups, while orthogonality implies component independence. Alignments with axes and data scatter reveal interpretability and dimensionality reduction effectiveness. This visual analysis enhances understanding of variable impact on principal components in multivariate datasets.

Conclusion:

- In conclusion, dimensionality reduction, particularly through approaches such as PCA, can be extremely useful in boosting computational efficiency, improving model performance, and revealing significant insights via visualisations. However, the trade-offs and problems involved with information loss, as well as the assumptions of the chosen dimensionality reduction method, must be carefully considered.

.