Documentation: CS F320 - FODS - Assignment 2

Names of students (with ID no):

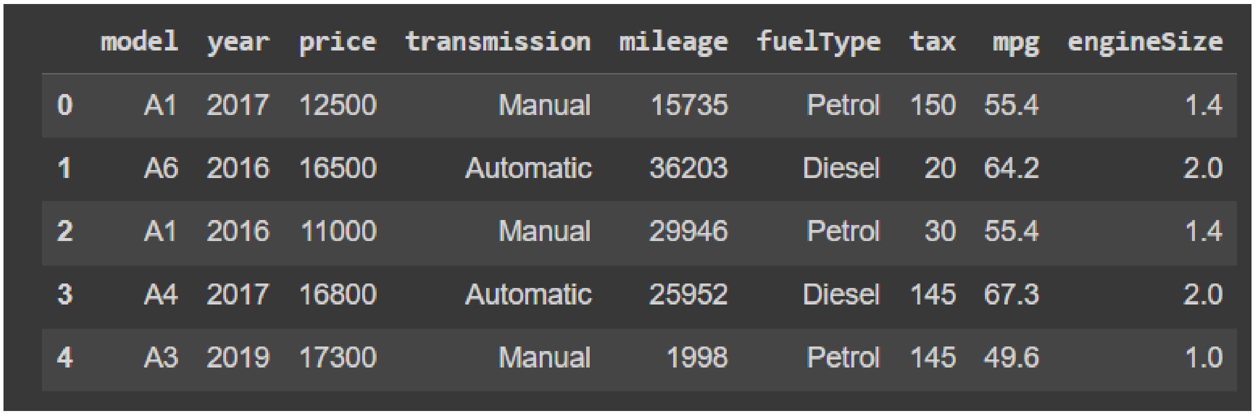
1. Aditya Kumar Sharma - 2021A7PS3112H
2. Ninad Agrawal - 2021A7PS2948H
3. Prachi Shah - 2021A7PS2589H

Part 2A

This assignment was focussed on application of the Principal Component Analysis (PCA). We go over each step to understand, how we did it and what was the outcome of it.

STEP 1: Data Understanding and Representation:

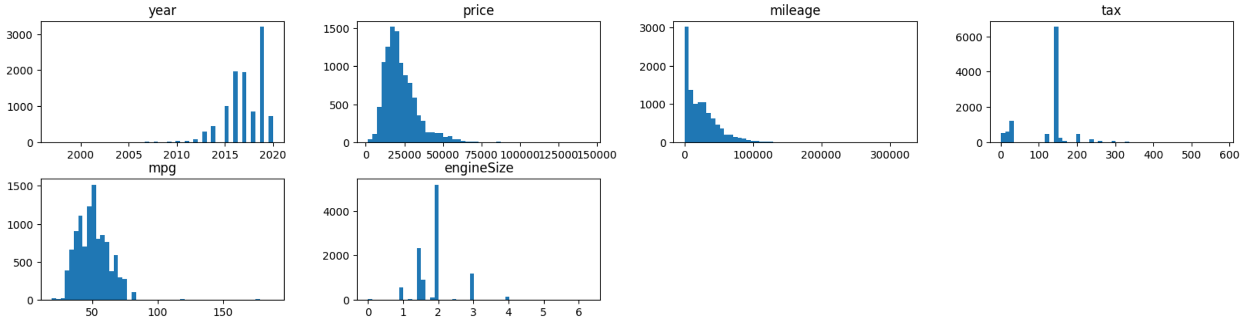
1. To understand the data we first imported it, and it looked like this:



	model	year	price	transmission	mileage	fuelType	tax	mpg	engineSize
0	A1	2017	12500	Manual	15735	Petrol	150	55.4	1.4
1	A6	2016	16500	Automatic	36203	Diesel	20	64.2	2.0
2	A1	2016	11000	Manual	29946	Petrol	30	55.4	1.4
3	A4	2017	16800	Automatic	25952	Diesel	145	67.3	2.0
4	A3	2019	17300	Manual	1998	Petrol	145	49.6	1.0

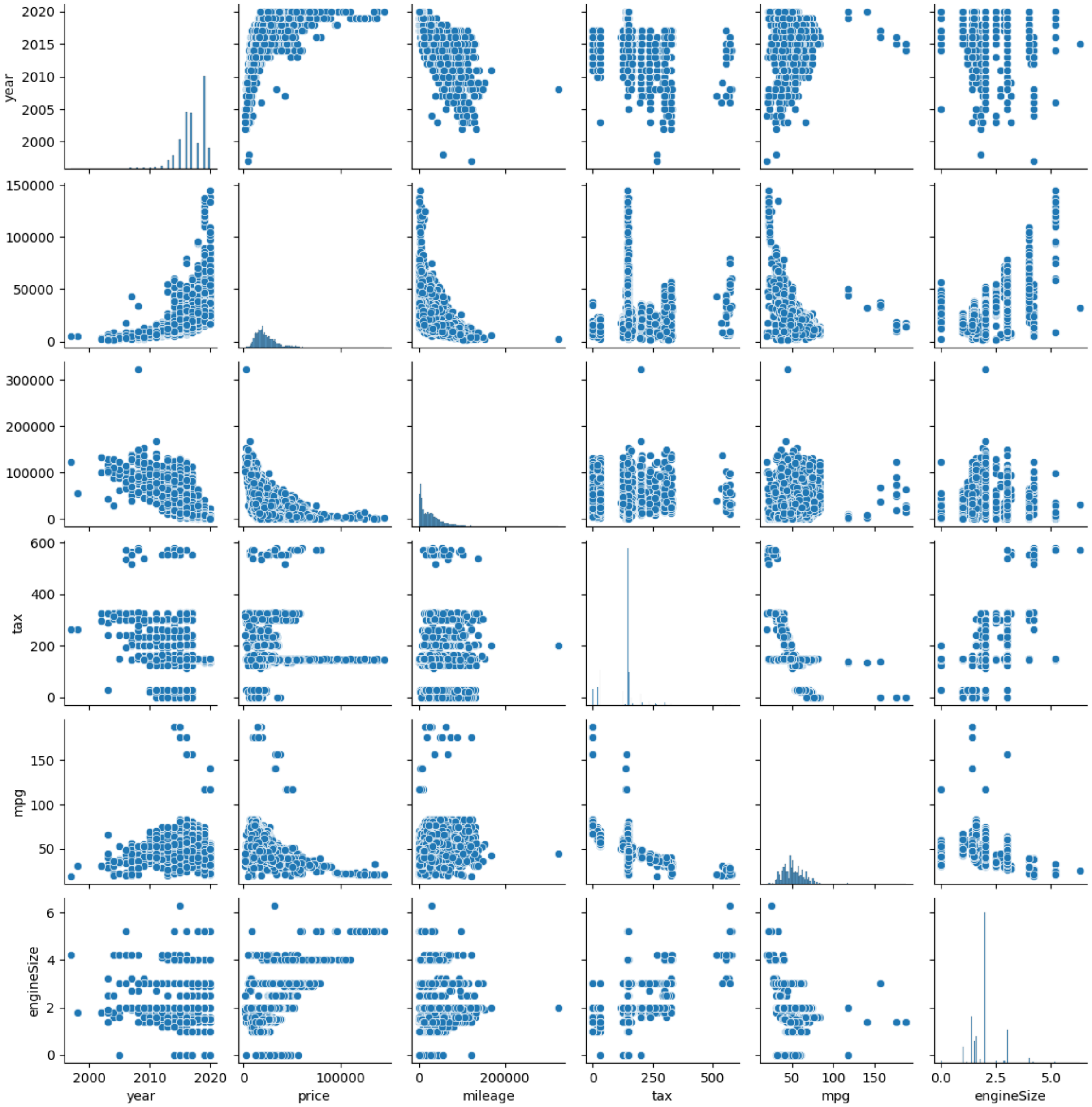
1. To understand data better we drew histograms.

```
Each feature is provided here:

array([[<Axes: title={'center': 'year'}>,
        <Axes: title={'center': 'price'}>,
        <Axes: title={'center': 'mileage'}>,
        <Axes: title={'center': 'tax'}>],
       [<Axes: title={'center': 'mpg'}>,
        <Axes: title={'center': 'engineSize'}>, <Axes: >, <Axes: >],
       [<Axes: >, <Axes: >, <Axes: >, <Axes: >],
       [<Axes: >, <Axes: >, <Axes: >, <Axes: >]], dtype=object)
```



1. Then we go on to understand the relation between each feature. we draw pairplots for that.

```
Each feature is plotted against another, to show their correlation:
```



1. Then we divide data into X and y: where X is the matrix of features and y is the matrix of target values.

Here, 'mileage', 'tax', 'mpg', 'engineSize' constitute X, while 'price' constitutes y.

Step 2: Finding Covariance Matrix:

1. First we find out the covariance matrix

```
This represents the covariance matrix:

array([[ 1.00009375, -0.16656277,  0.39514041,  0.0707168 ],
       [-0.16656277,  1.00009375, -0.63596815,  0.39311229],
       [ 0.39514041, -0.63596815,  1.00009375, -0.36565501],
       [ 0.0707168 ,  0.39311229, -0.36565501,  1.00009375]])
```

Step 3: Finding out the eigen values and the eigrn vectors

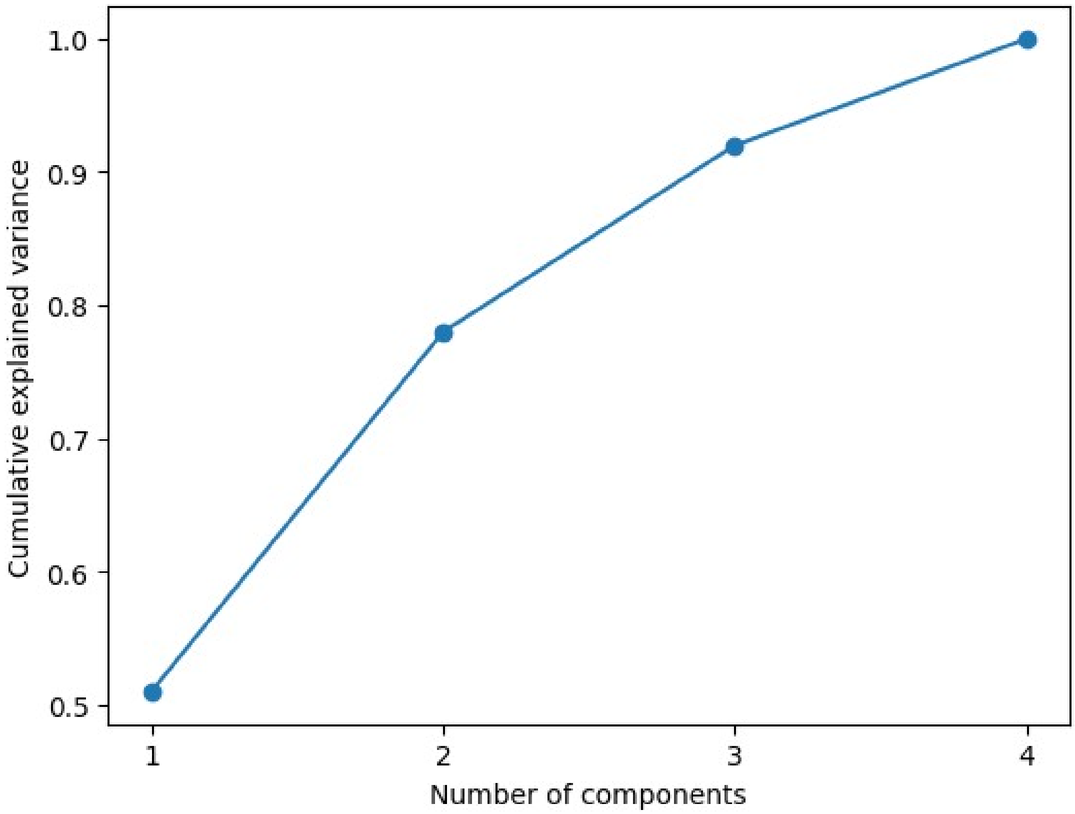
1. We follow the process asked in the assignment and arrive at the following result:

```
Following is the list of eigen values and eigen vectors:

Eigenvalues
 [2.03992434 1.09512926 0.55289517 0.31242622]
Eigenvectors
 [[ 0.3018911  -0.78612981 -0.43981596 -0.31212755]
 [-0.58766187 -0.12952322 -0.56935379  0.56010135]
 [ 0.62195239 -0.13838822  0.13586185  0.75866032]
 [-0.42034572 -0.58827914  0.68113122  0.11531428]]
```

Step 4: Solving for the principal components

1. Now finally we apply the PCA to find the principal components
1. We need to find top k features, so that our variance is around 75 % of the original variance.

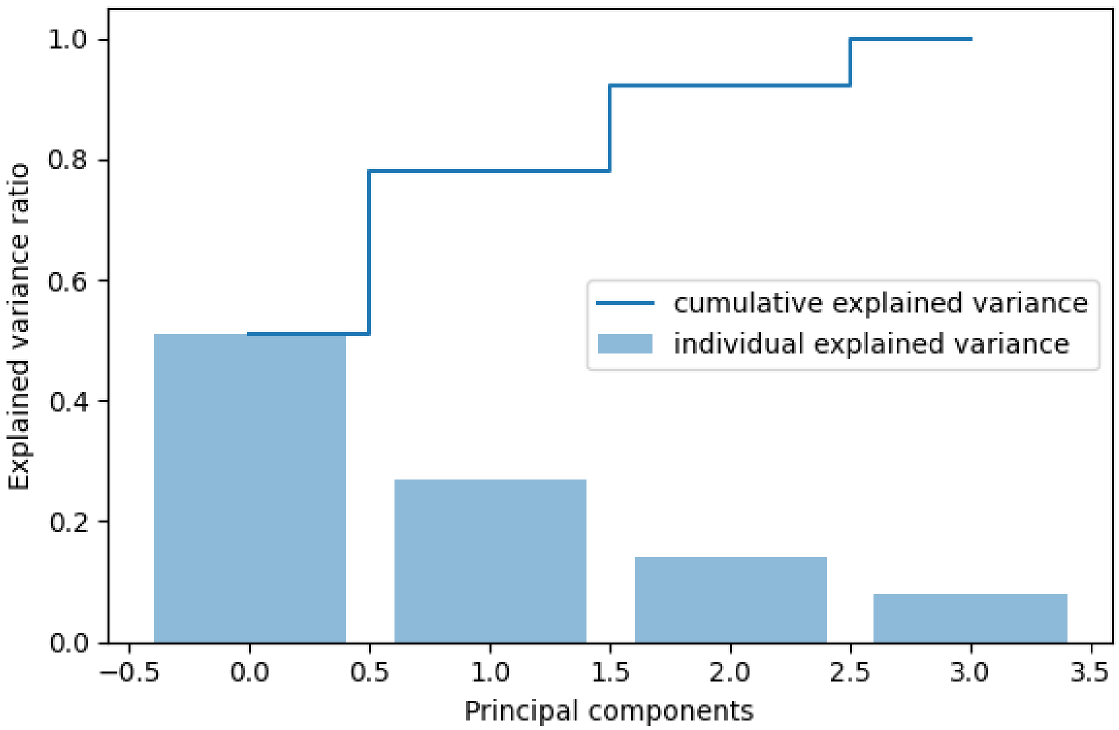
```
Here we have shown the variance for each of the features:

Explained variance: [0.51 0.27 0.14 0.08]
Cumulative explained variance: [0.51 0.78 0.92 1.  ]
```



1. As we can see, with 2 features we can reach enough variance, so we go forward with 2 features only.

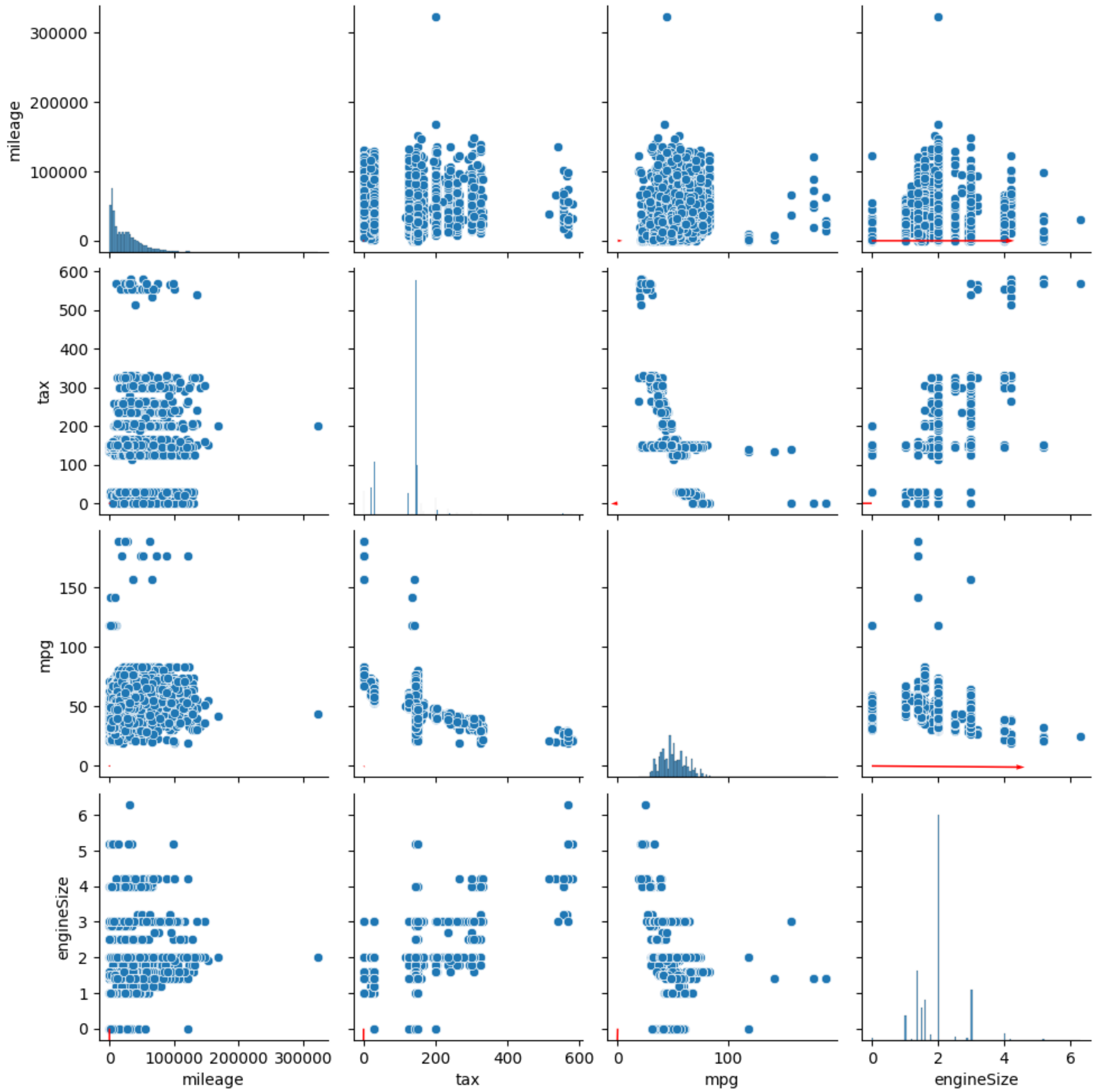
Step 5: Showing sequential variance increase:

```
1. This the the variance as previously shown, now we plot it further.

Components:
 [[ 0.3018911  -0.78612981 -0.43981596 -0.31212755]
 [-0.58766187 -0.12952322 -0.56935379  0.56010135]]
Explained variance scratch:
 [0.51 0.27 0.14 0.08]
Cumulative explained variance from scratch:
 [0.51 0.78 0.92 1.  ]

2. Following is the graph of individual and cumulative variance
```



Step 6: Pair Plots:

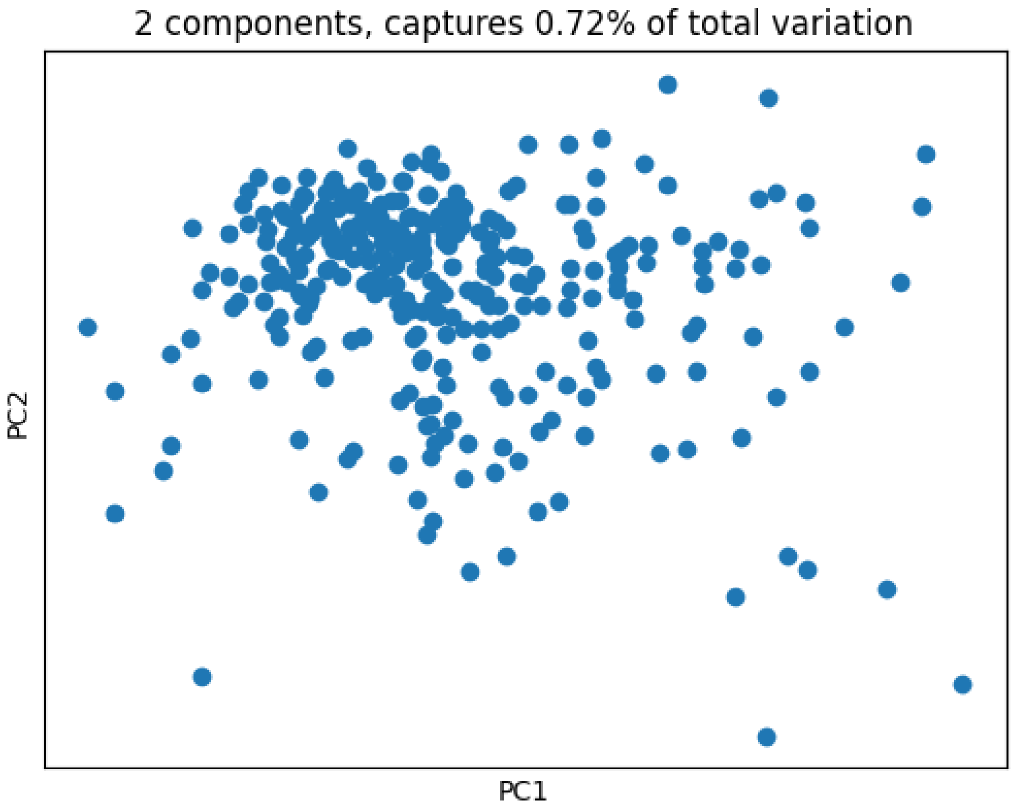
```
These are the pair plots, which are also visualised as vectors
```



Step 7: Conclusion and Interpretation

Principal Component Analysis (PCA) was applied to a dataset initially comprised of four components. Following the analysis, two principal components were identified that collectively explain 78% of the original dataset's variance.

```
This shows the plot of the 2 components:
```



7.1 Interpretation of results

1. Dimensionality Reduction: The reduction from four to two components demonstrates effective dimensionality reduction, simplifying the dataset while retaining key information.

2. Variance Retention: The selected two principal components capture a substantial 78% of the original variance, indicating their efficacy in summarizing the major patterns in the data.

3. Interpretability: Examination of variable loadings on these components will provide insights into the features contributing most to the captured variance, aiding in the interpretation of the principal components.

4. Application: The reduced dataset, now represented by two principal components, is suitable for subsequent analyses, offering a streamlined yet informative input for modeling.

5. Validation: Consideration of the cumulative explained variance reinforces the appropriateness of choosing these two components.

7.2 Significance of Principal Components in Capturing Variance:

1. Efficient Representation: Principal components are ordered by the amount of variance they explain. The first few principal components often capture the majority

of the dataset's variability. By focusing on these components, one can efficiently summarize the essential information contained in the data.

2. Reduction of Redundancy: Principal components are orthogonal to each other, meaning they are uncorrelated. This orthogonality ensures that each component represents a unique source of variability, reducing redundancy. This property is particularly valuable when dealing with multicollinearity in high-dimensional datasets.

7.3 Significance of Principal Components in Reducing Dimensionality:

1. Simplification of Models: High-dimensional datasets often lead to overfitting in models and increased computational complexity. Principal components allow for the creation of a reduced-dimensional representation of the data, facilitating the development of simpler and more interpretable models.

2. Computational Efficiency: Working with a reduced set of principal components can significantly speed up computational processes. This is particularly important in machine learning and statistical modeling, where computational efficiency is often a critical factor.

3. Enhanced Interpretability: A lower-dimensional representation of the data is inherently more interpretable than the original high-dimensional space. Principal components, being combinations of the original variables, can often be associated with underlying patterns or features in the data, aiding in a more intuitive understanding.

7.4 Effectiveness of Dimensionality Reduction:

1. Simplified Representation: The reduction from four to two principal components using PCA effectively simplifies the original dataset. By focusing on these two components, we capture a significant portion of the data's variability while reducing complexity.

2. Computational Efficiency: The streamlined dataset with two principal components contributes to computational efficiency. This is particularly advantageous in scenarios where quicker analyses and model training are essential.

3. Model Performance: Models trained on this reduced-dimensional representation are expected to perform well, as the chosen components collectively explain 78% of the original variance. This suggests that the key patterns and information in the data are retained in the reduced space.

Step 8: Conclusion

In summary, the application of Principal Component Analysis effectively reduced the dataset from four to two components, retaining 78% of the original variance. This streamlined representation enhances interpretability, facilitates computational efficiency, and offers valuable insights through visualizations. The reduced dimensions present a promising foundation for subsequent analyses and modeling, showcasing the effectiveness of dimensionality reduction in capturing essential patterns in the data.