Documentation 2A:

1. Data Understanding and Representation:

Displayed the head (starting) rows in matrix form. We understand the form of data provided to us by doing this.

Here each row represents an observation (car) and each column represents a feature.

Here is the screenshot of the data's head:

	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

2. Implementing PCA using Covariance Matrices:

- Calculated the mean and standard deviation of each feature in the dataset.
- The dataset has been standardized by subtracting mean of corresponding features from the respective data points' corresponding features.
- The eigenvalues and eigenvectors were calculated and correspondingly sorted.
- Below attached are the eigenvectors and eigenvalues:

```
sorted_eigenvalues
```

```
array([3.00600026, 1.55286623, 0.78562111, 0.31859008, 0.20305466,
       0.13386766])
```

```
sorted_eigenvectors
```

```
array([[-0.40282543, -0.50443824, -0.10088103,  0.37645692,  0.63995831,
         0.1477508 ],
       [-0.51016611,  0.03785255, -0.41901248, -0.73835396,  0.0976721 ,
        -0.08955372],
       [ 0.4047773 ,  0.49359229, -0.08336572, -0.08457989,  0.72984732,
        -0.21387209],
       [-0.28190526,  0.54769809, -0.55984939,  0.49654138, -0.16767081,
         0.18016117],
       [ 0.46372679, -0.19415189, -0.372847  , -0.20625118,  0.02788077,
         0.75161875],
       [-0.34511511,  0.40349442,  0.59569749, -0.1298729 ,  0.13908837,
         0.57185643]])
```

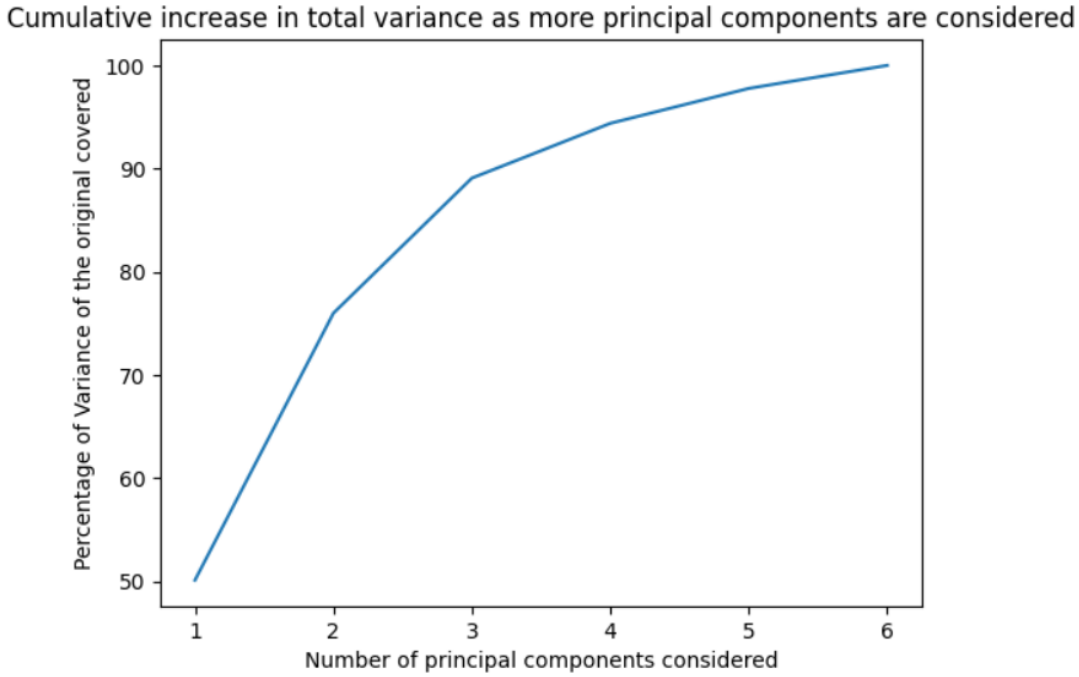
- They were calculated by solving the characteristic equation ($C - \lambda I = 0$)
- The λ values give the eigenvalues and C values are the eigenvectors
- Cumulative increase of variance has been attached below:

```
Total Variance covered by all principal components: 6.000000000000003
Total variance when top 1 principal components are considered: 3.006000257279837
Total variance when top 2 principal components are considered: 4.55886649155676
Total variance when top 3 principal components are considered: 5.344487601473161
Total variance when top 4 principal components are considered: 5.66307768260635
Total variance when top 5 principal components are considered: 5.866132341418811
Total variance when top 6 principal components are considered: 6.000000000000003
```

- In tabular form:

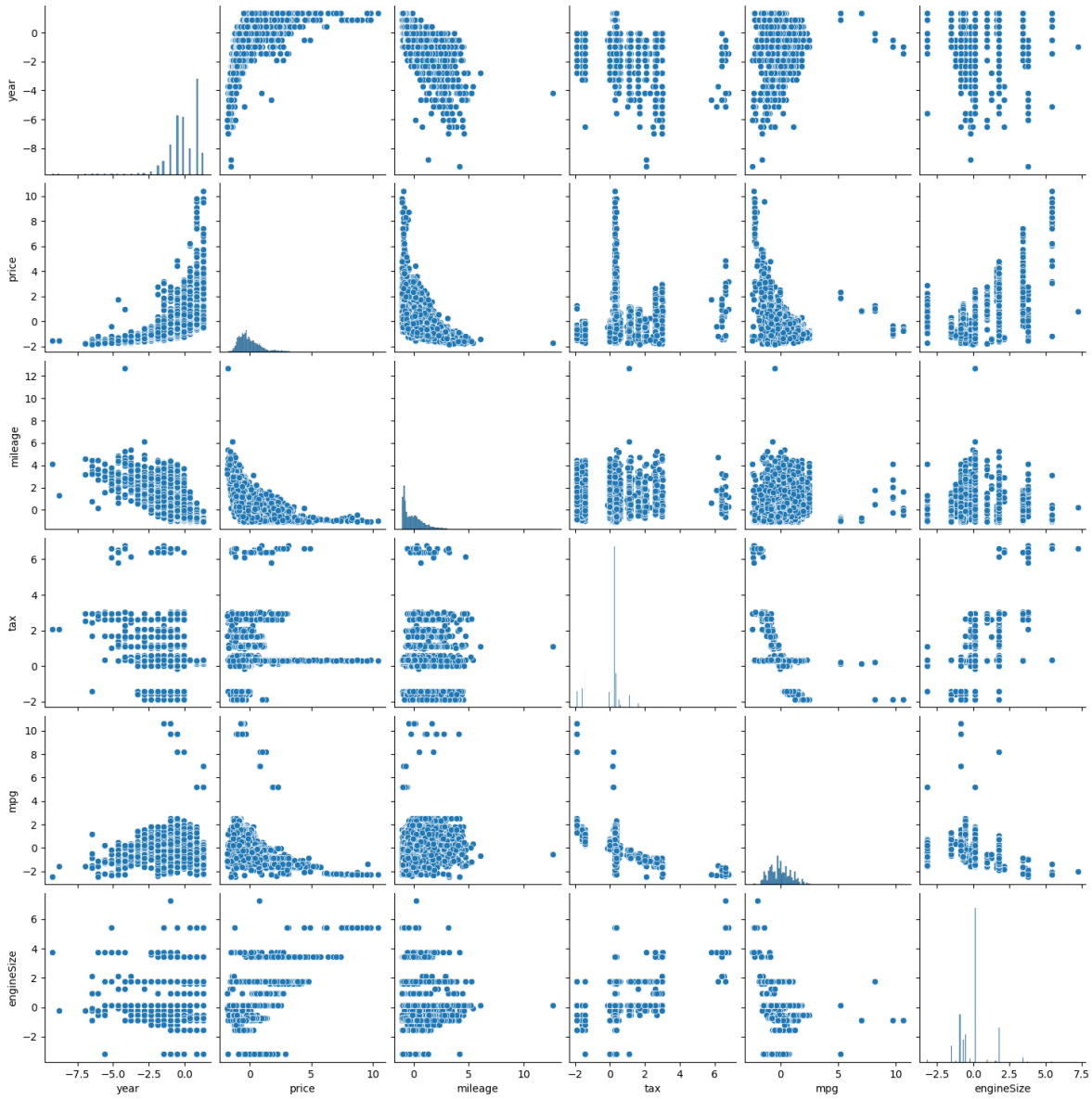
	Num of Principal Components	Percentage Variance covered
0	1.0	50.100004
1	2.0	75.981108
2	3.0	89.074793
3	4.0	94.384628
4	5.0	97.768872
5	6.0	100.000000

- We see that if we use just 5 principal components, we can cover more than 95% of variance of the data.
- In graphical form:



3. Implementing PCA using Covariance Matrices:

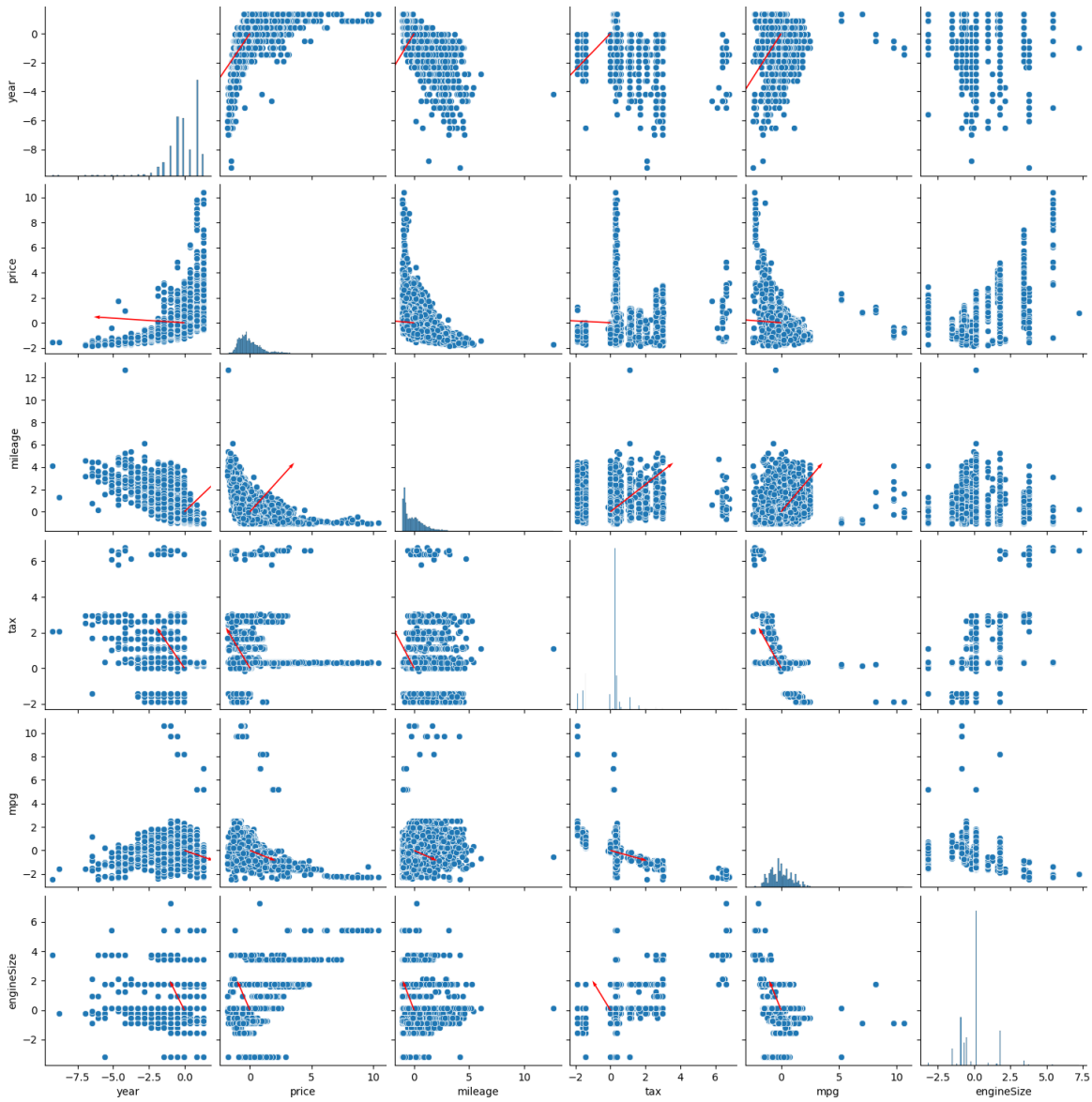
The pair plots of all the pairs of features are given below:



- The diagonal graphs are just the probability distributions of the features.

4. Projecting Principal Components into Pair Plots:

The principal components projected onto the pair plots are displayed below:



5. Results & Conclusions:

Interpreting the results from Principal Component Analysis (PCA) involves understanding the role of principal components, the explained variance, and the implications of dimensionality reduction. Let's break down each aspect:

1. Principal Components (PCs):
   - Principal components are linear combinations of the original variables in your dataset.

- The first principal component (PC1) explains the most variance in the data, followed by PC2, PC3, and so on.
- Each principal component is orthogonal to the others, meaning they are uncorrelated.
- The eigenvectors associated with each principal component indicate the direction in which the data varies the most.

2. Explained Variance:
- The eigenvalues associated with each principal component represent the amount of variance explained by that component.
- The total variance in the data is the sum of the eigenvalues.
- The proportion of explained variance for each principal component is calculated by dividing its eigenvalue by the total variance.
- High eigenvalues indicate that the corresponding principal component captures a large amount of variability in the data.

3. Significance of Principal Components:
- Principal components allow you to prioritize and focus on the most critical patterns and structures in your data.
- They provide a way to reduce the dimensionality of your dataset while retaining as much of the original variability as possible.
- The significance of each principal component is related to its ability to capture and represent patterns or trends in the data.

4. Reducing Dimensionality:
- Dimensionality reduction is a critical aspect of PCA. It involves selecting a subset of the most important principal components.
- By retaining the top principal components, you can achieve a lower-dimensional representation of your data while preserving most of its variability.
- Reduced dimensionality can lead to more efficient modeling, faster computation, and improved interpretability.

5. Effectiveness of Dimensionality Reduction:
- Evaluate the cumulative explained variance across the selected principal components. A common practice is to choose the number of components that explain a sufficiently high percentage of the total variance (e.g., 90%).
- Visualize the data in the reduced-dimensional space to gain insights into the structure and relationships among observations.
- Compare the results of models trained on the original and reduced-dimensional datasets to assess the impact on predictive performance.

6. Insights from Visualizations:
   - Visualizations, such as biplots or scatter plots of selected principal components, help interpret the relationships among variables and observations.
   - Clustering or grouping patterns in reduced-dimensional space may reveal hidden structures or similarities in the data.
   - Interpret the direction and magnitude of eigenvectors in biplots to understand which variables contribute the most to the principal components.

In summary, PCA provides a powerful tool for data exploration, dimensionality reduction, and visualization. The interpretation of results involves understanding the contribution of each principal component to the overall variance and assessing the implications of dimensionality reduction on model performance and data representation. Visualizations play a crucial role in gaining insights from the reduced-dimensional space.