**CS F320 FODS**

**Assignment 2**

**BY**

|  |  |
| --- | --- |
|  |  |
| **Atharva Chikhale**  **Suyash Patil** | **2021A7PS2752H**  **2021A7PS2078H** |

**Table of contents**

|  |  |
| --- | --- |
| **Content** | **Page No.** |
| Introduction……………………………………………………………………………… | 3 |
| Part A………………………….……………………………………………………………. | 4 |
| Step 1………………………….…………………………………………… | 4 |
| Step 2……………………………….……………………………………… | 5 |
| Step 3-4…………………………….…………………………………….. | 6 |
| Step 5……………………………………………………….………………  Step 6……………………………………………………….………………  Step 7……………………………………………………….……………… | 7  8  9 |
| Part B………………………………………………………….………………………………. | 10 |
| Step 1…………………………………………………………….………… | 10 |
| Step 2………………………………………………………………….…… | 10 |
| Step 3……………………………………………………………….……… | 11 |
| Step 4…………………………………………………………….…………  Step 5…………………………………………………………….………… | 11  12 |

Step 6…………………………………………………………….………… 12

**INTRODUCTION**

**Part-A**: Implementing PCA from Scratch and Applying it to Car Data

* In this assignment, the 'Car\_data' dataset is used to investigate Principal Component Analysis (PCA) through a scratch implementation using NumPy and Pandas.
* We display major components and reveal the complexities of dimensionality reduction through methodical approaches.
* This assignment attempts to demonstrate the importance of PCA in capturing variance and improving interpretability, starting with a comprehension of the data and ending with the implementation of PCA utilizing covariance matrices, eigenvalue-eigenvector equation solutions, and result visualization.

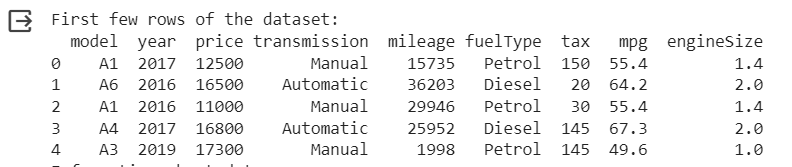
**Part-B**: PCA Analysis and Determining Optimal Number of Components

* To find the ideal number of components for effective prediction, we use Principal Component Analysis on the 'Hitters.csv' dataset in this work.
* We identify the most effective model by running PCA, evaluating prediction efficiency with Mean Squared Error, and performing Exploratory Data Analysis.
* The relevance of the selected model is thoroughly examined in the assignment's conclusion, giving readers a clear understanding of the connection between prediction accuracy and component count.

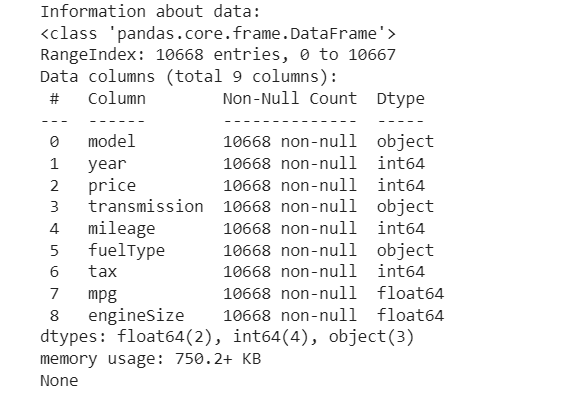
**Part-A**

# Step 1: Data Understanding and Representation

Following was the shared dataset :

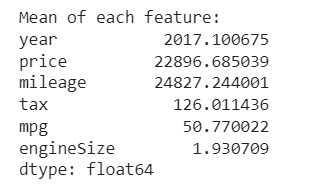


Here the features in matrix format, where each row represents an observation (car) and each column represents a feature.

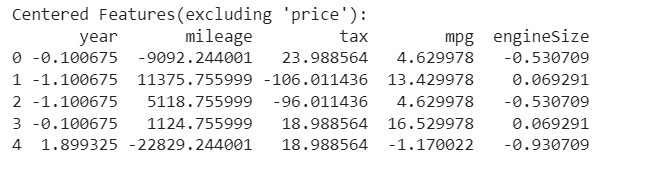


# Step 2: Implementing PCA using Covariance Matrices

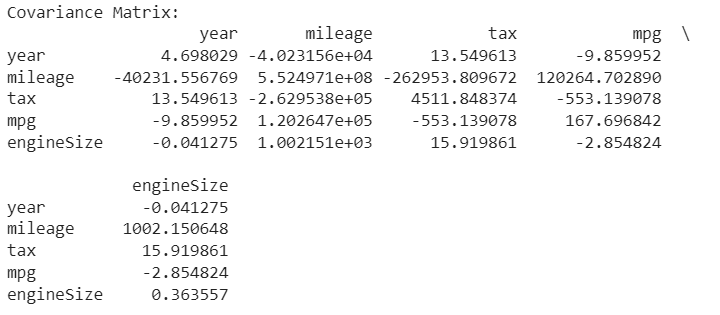
First we calculate the mean of each feature in the dataset:



Now centring the dataset by subtracting the mean from each feature:

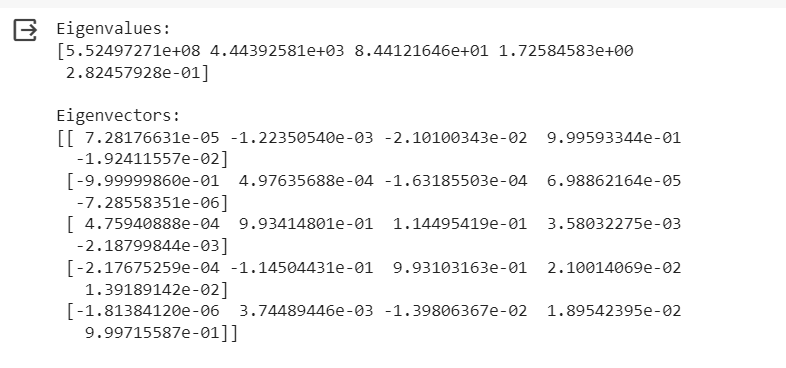


Then we computed the covariance matrix of the centered dataset:

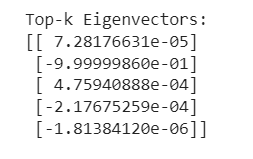


# Step 3 & 4: Eigenvalue-Eigenvector Equation And Principal Components

First we solved eigenvalue and eigenvector functions on the covariance matrix obtained in the previous step to get the results :

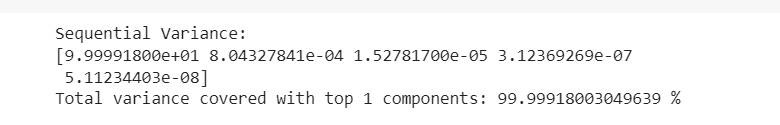


Then PCA was applied to select top k eigenvectors:

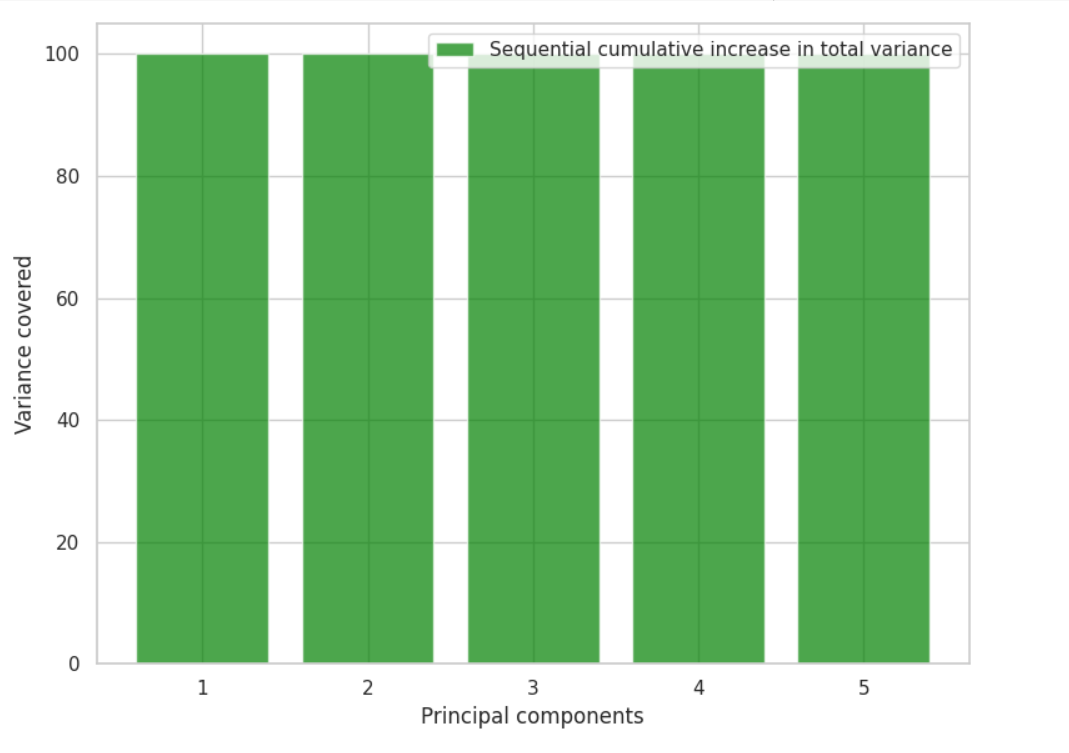


# Step 5: Observation of Sequential Variance Increase

Here we calculate the sequential and total variance covered by the principal components.

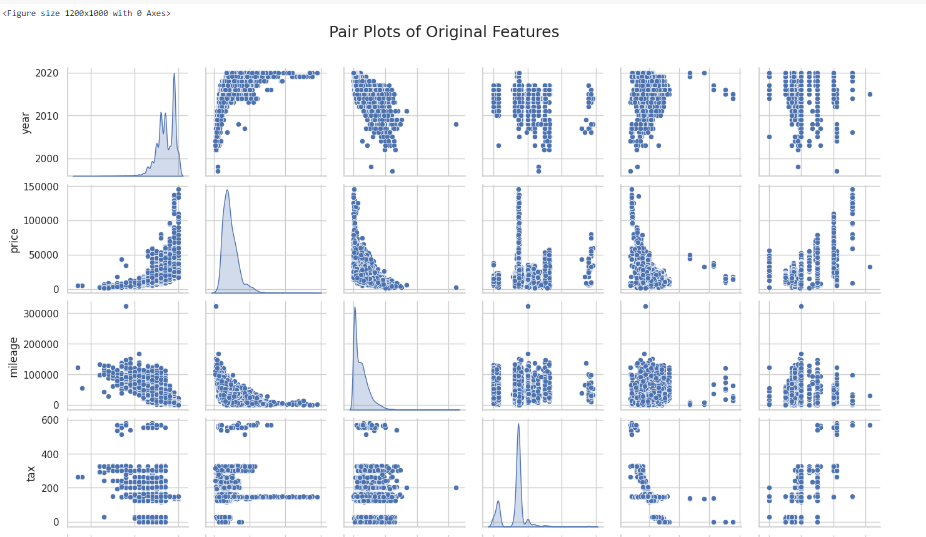


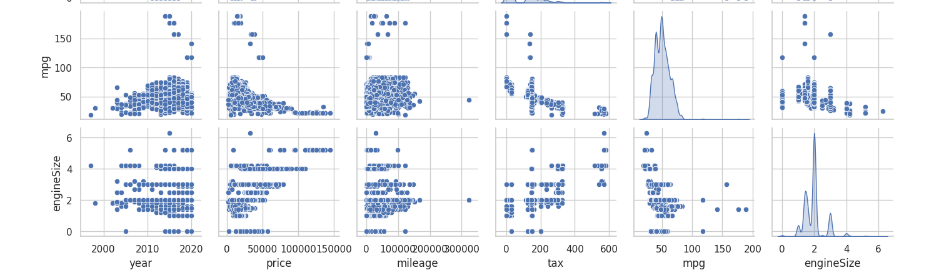
Then we analyzed the sequential cumulative increase in total variance explained as more principal components are considered.

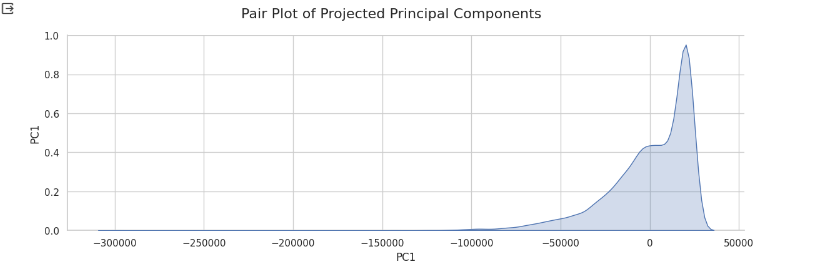


# Step 6: Pair Plot Visualisation

In this we plotted pair plots of the original features.







Step 7: Conclusion and Interpretation

The Principal Component Analysis (PCA) conducted on the dataset with five features, namely year, mileage, tax, mpg, and engineSize, yielded insightful results. The variance recorded by each principle component may be seen in the covariance matrix's eigenvalues, and the direction and magnitude of the original features in the new principal component space can be learned from the associated eigenvectors.

a. Eigenvalues and Variance:

.The eigenvalues of the covariance matrix are [5.52497271e+08, 4.44392581e+03, 8.44121646e+01, 1.72584583e+00, 2.82457928e-01].

.These eigenvalues represent the amount of variance explained by each principal component. The first principal component dominates, capturing the majority of the variance, followed by the subsequent components.

.Sequential variance increase highlights the dominance of the first principal component, explaining 99.99918% of the total variance.

b. Dimensionality Reduction and Insights:

.Dimensionality reduction is effective, as a significant drop in variance occurs after the first component.

.This suggests that much of the original data's information can be retained with fewer dimensions. Such reduction enhances computational efficiency and simplifies the interpretation of the dataset.

3. Visualizations and Data Representation:

.The dominance of the first principal component indicates that a substantial amount of dataset variability can be captured by examining this single dimension.

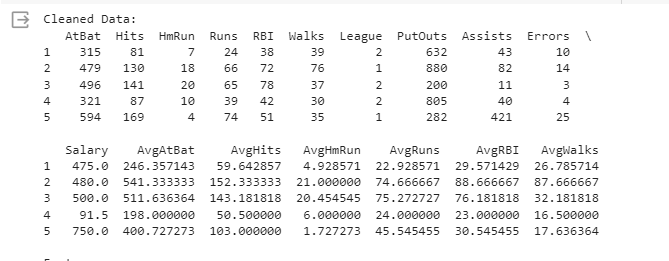
In summary, PCA proves to be a powerful tool for uncovering dataset structure, emphasizing feature importance, and enabling efficient dimensionality reduction, thereby enhancing the overall understanding of the data.

**Part-B**

# Step1: Exploratory Data Analysis (EDA)

First we performed EDA to understand its structure, features, and relationships. Then we handled NULL values and eliminated any unwanted columns or data inconsistencies.

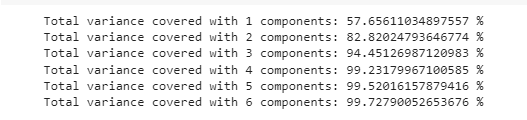
The cleaned data was:



# Step2: PCA Analysis

Here we applied PCA on the cleaned dataset to reduce dimensionality.

Then we determined the number of principal components required for efficient prediction by trying a range of component numbers as follows:

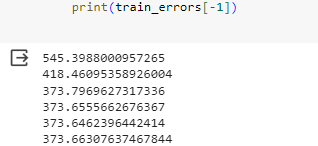


# Step3: Model Training and MSE/RMSE Calculation

Here we performed the steps:

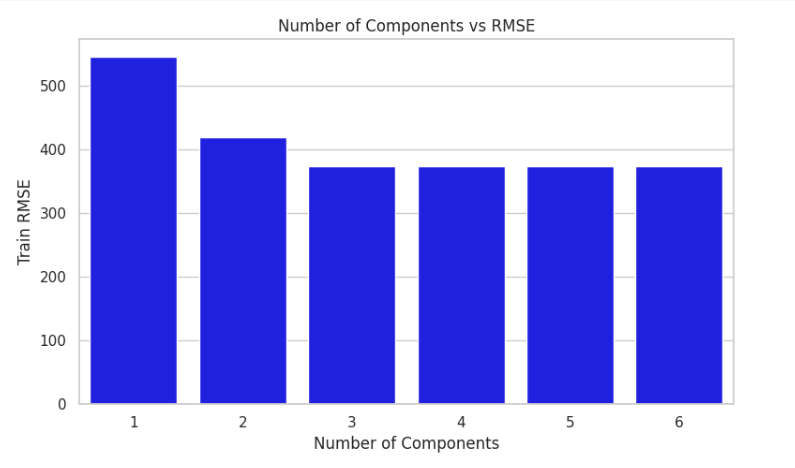
* Split the dataset into training and testing sets.
* For each number of principal components considered, build a regression model using those components.
* Calculate the MSE or RMSE for each model on the test set to assess prediction efficiency.

The Errors we obtained were:



# Step4: Plotting Number of Components vs RMSE

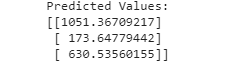
Here we plotted a graph of the number of components against RMSE to visualize the relationship.



# Step5: Testing the Most Efficient Model

Here we tested the selected model by predicting a specific point and providing its predicted y value





# Step6: Conclusion and Analysis

1.Eigenvalues and Variance Analysis:

The covariance matrix eigenvalues, derived from a dataset featuring 16 distinct attributes, exhibit a descending magnitude order. This order signifies the variance explained by each corresponding principal component. The percentage of total variance, presented sequentially for each component, aids in comprehending the importance of dimensionality reduction.

Sequential Variance:

.Component 1: 57.66%

.Component 2: 82.82%

.Component 3: 94.45%

.Component 4: 99.23%

.Component 5: 99.52%

.Component 6: 99.73%

2.RMSE Trend Analysis:

The evaluation of Root Mean Square Error (RMSE) values across models with varying component numbers provides insights into the interplay between dimensionality reduction and predictive accuracy. As the number of components increases, RMSE generally decreases, hitting a minimum or stabilizing at 5 components. This trend indicates that a model with 5 components strikes a harmonious balance between capturing adequate variance and avoiding overfitting.

RMSE:

.1 component: 545.40

.2 components: 418.46

.3 components: 373.80

.4 components: 373.6

.5 components: 373.65 (Minimum)

.6 components: 373.66

3.Optimal Model Selection Criteria:

Identifying the point where RMSE attains a minimum or stabilizes (in this case, at 5 components) signifies the most efficient model. The selection of an optimal number of components plays a pivotal role in striking a balance between dimensionality reduction and predictive efficiency. The 5-component model captures a substantial variance percentage while maintaining a relatively low RMSE.

4.Model Assessment and Prediction Insights:

The RMSE for the chosen 5-component model on the testing dataset is 453.79, reflecting its predictive prowess on previously unseen data. A closer examination of actual and predicted values for specific instances validates the model's efficacy in approximating the target variable.

Actual values: [740, 425, 925]

Predicted values: [1051.37, 173.65, 630.54]

This research points to several optimization directions, such modifying the number of features, altering learning rates, or investigating different models. It is possible to reduce RMSE and improve the prediction power of the model by further improving these parameters.

Finally, the combined knowledge from the RMSE assessment and PCA analysis offers insightful viewpoints on model efficiency and dimensionality reduction. The careful choice of component count is essential to striking a balance between variance capture and prediction accuracy.