

Data Analysis for Cars Dataset

Implementation of Statistical Analysis, ACP, and Predictive Analysis

M1 in Big Data and Business intelligence

**Authors: Superviso**r:

Abdelouhab Bella Pr. Abdelghani GHAZDALI

Oussama Qouti

Nizar Bennani

2023-2024

# Abstract

The analysis aims to explore the relationship between various car attributes and their performance. The dataset includes information such as miles per gallon (mpg), number of cylinders, engine displacement, horsepower, vehicle weight, time to accelerate from 0 to 60 mph, manufacturing year, and brand. The report utilizes exploratory data analysis and principal component analysis to uncover patterns and insights within the data.

# Table of content

[Abstract 2](#_Toc157904821)

[Table of content 3](#_Toc157904822)

[Introduction 4](#_Toc157904823)

[Chapter 1: Data Preprocessing 5](#_Toc157904824)

[1. Introduction 5](#_Toc157904825)

[2. Analyzing Extracted Data 5](#_Toc157904826)

[2.1 Manual Data Analysis 5](#_Toc157904827)

[2.2 Performing Data Cleaning 6](#_Toc157904828)

[2.3 Statistical Analysis to Detect Data Anomalies 6](#_Toc157904829)

[2.4 Normalizing the Data 7](#_Toc157904830)

[2.5 The Correlation 8](#_Toc157904831)

[Chapter 2: Reducing Data Dimensionality with PCA 9](#_Toc157904832)

[PCA Analysis 9](#_Toc157904833)

[Initializing PCA Analysis: 9](#_Toc157904834)

[Modeling with PCA Analysis 9](#_Toc157904835)

[Checking Variance and Singular Values 9](#_Toc157904836)

[Creating and Displaying DataFrame 9](#_Toc157904837)

[Creating Subset DataFrame 9](#_Toc157904838)

[Visualising PCA 10](#_Toc157904839)

[Executing Correlation Circle Plot: 10](#_Toc157904840)

[Scree Plot for Cumulative Explained Variance: 11](#_Toc157904841)

[Chapter 3: Cluster Analysis with KMeans 12](#_Toc157904842)

[Executing KMeans Analysis 12](#_Toc157904843)

[3D Plotting for Visualization: 12](#_Toc157904844)

[Chapter 4: Cluster Analysis with KMeans 14](#_Toc157904845)

[1. Dataset Comparison: 14](#_Toc157904846)

[2. Decision Tree Analysis: 14](#_Toc157904847)

[3. Logistic Regression Analysis: 15](#_Toc157904848)

[4. Comparative Analysis: 16](#_Toc157904849)

[5. Insights and Conclusions : 16](#_Toc157904850)

[Conclusion 17](#_Toc157904851)

# Table of figures

[Figure 1 : The boxplots 8](#_Toc157904892)

[Figure 2 : the correlation matrix for our variables 9](#_Toc157904893)

[Figure 3 : Correlation circle 11](#_Toc157904894)

[Figure 4 : Scree Plot for PCA 12](#_Toc157904895)

[Figure 5 : the clusters in 3d space 13](#_Toc157904896)

[Figure 6 : Decision tree results 16](#_Toc157904897)

[Figure 7 : Logistic regression results 16](#_Toc157904898)

# 

# Introduction

The analysis presented in this report aims to explore the relationship between various car attributes and their performance. The dataset used in the analysis includes information such as miles per gallon (mpg), number of cylinders, engine displacement, horsepower, vehicle weight, time to accelerate from 0 to 60 mph, manufacturing year, and brand. The report utilizes exploratory data analysis and principal component analysis to uncover patterns and insights within the data. The report is structured as follows: first, we provide an overview of the data and the analysis methodology. Then, we present the findings and insights gained from the analysis. Finally, we conclude with a summary o

# Chapter 1: Data Preprocessing

## Introduction

This chapter focuses on preparing the data for analysis. Various tasks were performed, including checking for duplicate and null values, changing categorical data to numeric values, handling missing data, and exploring data statistics.

## Analyzing Extracted Data

### Manual Data Analysis

In this section, we perform a manual analysis of the dataset to gain insights into its structure and contents.

| **mpg** | **cylinders** | **cubicinches** | **hp** | **weightlbs** | **time-to-60** | **year** | **brand** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 14.0 | 8 | 350 | 165 | 4209 | 12 | 1972 | US. |
| 31.9 | 4 | 89 | 71 | 1925 | 14 | 1980 | Europe. |
| 17.0 | 8 | 302 | 140 | 3449 | 11 | 1971 | US. |
| 15.0 | 8 | 400 | 150 | 3761 | 10 | 1971 | US. |
| 30.5 | 4 | 98 | 63 | 2051 | 17 | 1978 | US. |

Table 1 : Data Head

DataFrame Information: We start by examining the basic information about the dataset using cars.info().

Total Cells Count: Checking the total count of cells to identify any missing values with cars.count().

Data Statistics: Utilizing cars.describe() to gather statistical information about the dataset.

Duplicate Data: Identifying and counting any duplicated rows with cars.duplicated().sum().

Null Data: Checking for the presence of null values in the dataset using cars.isnull().any().

Brand Distribution: Investigating the distribution of car brands in the dataset with cars.brand.value\_counts().

Cars per Year: Analyzing the distribution of cars across different years using cars.year.value\_counts().

Cylinders Classification: Exploring the classification of the total number of cylinders in the engine with cars.cylinders.value\_counts().

### Performing Data Cleaning

In this section, we focus on preparing the data for further analysis by addressing any issues or inconsistencies.

Creating a Copy for Modification: Creating a copy of the original dataframe (cars2) to avoid altering the original data.

Changing Regions to Categorical Values: Converting the regions of the car models into categorical values for machine learning analysis.

Converting Volume and Weight to Numeric: Changing the data types of the volume (cubicinches) and weight (weightlbs) columns from objects to numeric values.

Handling NA Values: Dealing with missing values by replacing them with the mean of the respective columns to ensure data integrity for machine learning models.

### Statistical Analysis to Detect Data Anomalies

In this section, we perform statistical analysis to identify potential anomalies or outliers in the dataset.

Boxplot Analysis for Outliers: Utilizing boxplot analysis on various features like mileage (mpg), cylinders, volume (cubicinches), power (hp), weight (weightlbs), and time-to-60mph (time-to-60) to identify potential outliers.

Spread of Data in Boxplot: Examining the spread of data for features like year and brand to gain insights into the distribution and identify any anomalies.

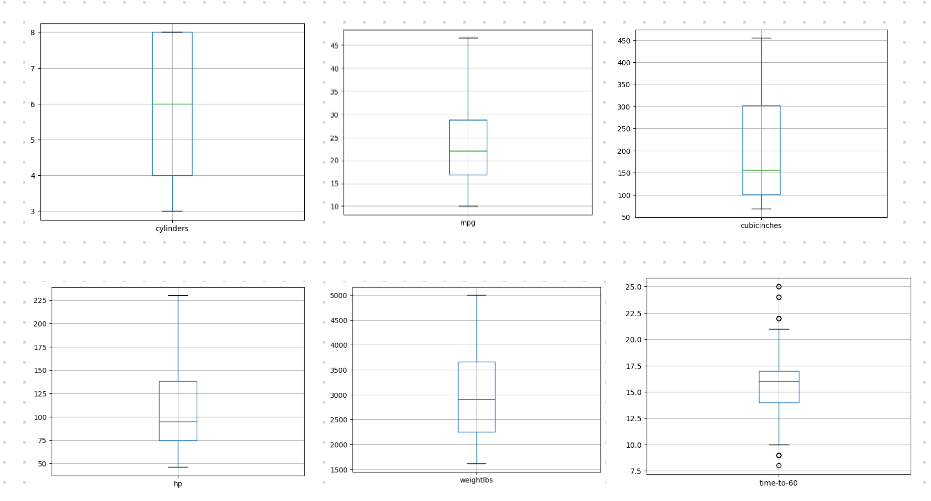


Figure 1 : The boxplots

We can see from the boxplots above that the their isn’t many outliers (except for the variable time-to-60) but will conduct the analysis using them as they don’t have a big impact

### Normalizing the Data

In this section, we focus on normalizing the data to ensure consistent scales for machine learning analysis.

Creating a Copy DataFrame for Normalization: Creating a copy of the cars2 dataframe (dftest) to normalize the data for machine learning analysis.

Data Preprocessing with StandardScaler: Using the StandardScaler() function from the sklearn library to normalize the selected features.

Verifying Data Normalization: Checking the mean and standard deviation of the normalized data to ensure that the values are centered around zero with a standard deviation of the findings and their implications.

### The Correlation

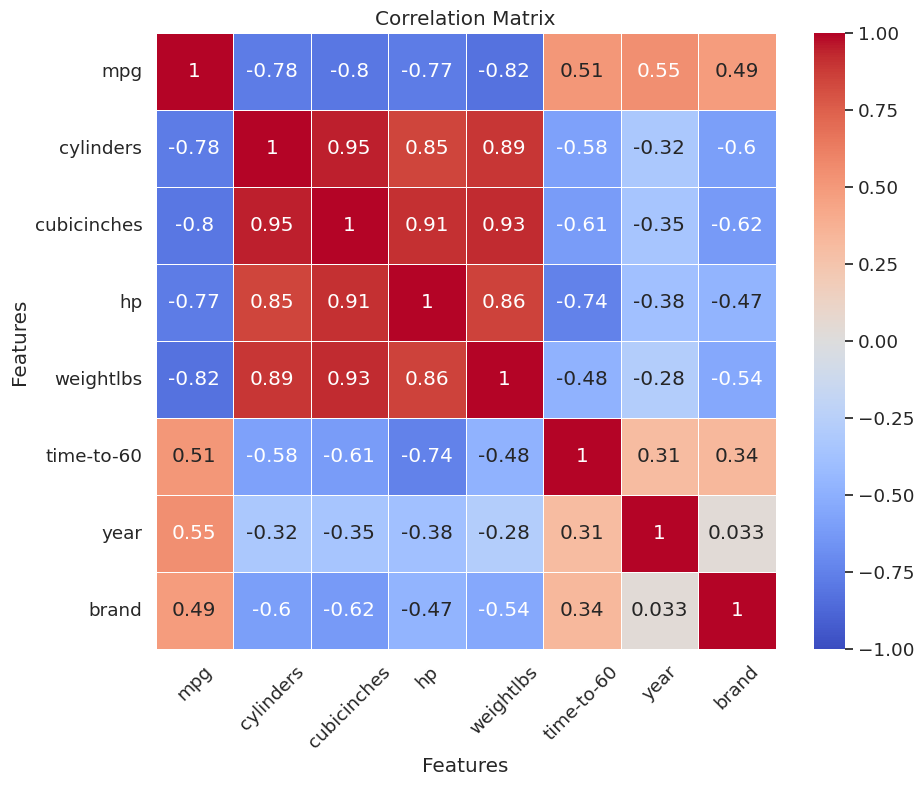


Figure 2 : the correlation matrix for our variables

Upon reviewing the correlation matrix, it is possible to identify a high correlation (>0.7) between mileage(mpg), total cylinders (cylinders), volume (cubicinches), power (hp), and weight (weightlbs).

# Chapter 2: Reducing Data Dimensionality with PCA

## PCA Analysis

### Initializing PCA Analysis:

- The journey begins with the initiation of Principal Component Analysis (PCA) on the dataset (`dftest`). PCA is configured with `n\_components = 7`, indicating the desire to retain seven principal components.

### Modeling with PCA Analysis

- The data is modeled using the PCA analysis, resulting in transformed principal components stored in `principalcomponents`.

### Checking Variance and Singular Values

- Insights into the variance captured by each principal component are unveiled through `pca.explained\_variance\_ratio\_`. Additionally, singular values obtained during the PCA analysis are printed using `pca.singular\_values\_`.

### Creating and Displaying DataFrame

- A DataFrame (`principalDf`) is crafted to showcase the data based on the PCA analysis, providing a clear view of the distribution across the seven principal components.

### Creating Subset DataFrame

- A new DataFrame (`principalDf2`) is extracted, containing only the first three principal components, setting the stage for subsequent visualizations.

Displaying Subset DataFrame:

- The values of the subset DataFrame (`principalDf2`) are presented, offering a focused perspective on the reduced dimensionality.

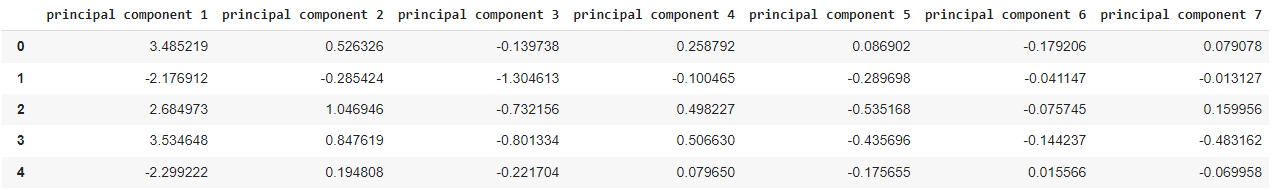


Table 2 : principalDF2

## Visualising PCA

### Executing Correlation Circle Plot:

- The function is applied to the `cars2` dataset, offering a visual representation of variable relationships within the PCA space. The plot showcases arrows representing variables, where length indicates importance, and angles depict relationships.

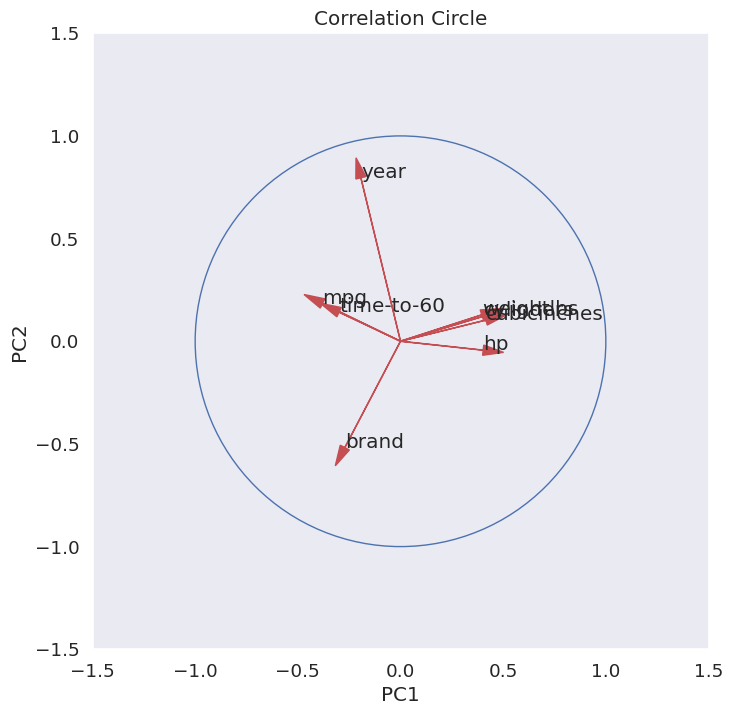


Figure 3 : Correlation circle

The correlation circle plot visualizes how variables contribute to principal components in PCA. Longer arrows indicate greater importance, while angles between arrows show relationships between variables. Closer arrows suggest similar patterns, while farther ones indicate less correlation. Understanding these relationships helps in dimension reduction and interpretation of data patterns.

### Scree Plot for Cumulative Explained Variance:

- A Scree Plot is generated to display the cumulative explained variance ratio. This plot aids in determining the optimal number of components to retain, a crucial step in the dimensionality reduction process.

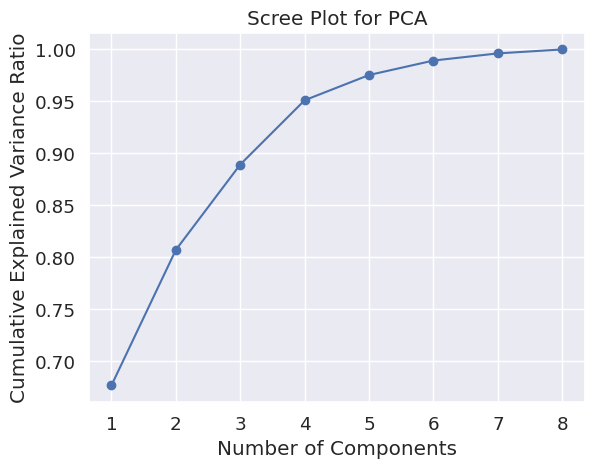


Figure 4 : Scree Plot for PCA

# Chapter 3: Cluster Analysis with KMeans

# Executing KMeans Analysis

KMeans clustering is applied to the reduced principal components (principalDf2) with three clusters, providing an initial insight into the dataset's inherent structure.

Checking Results and Reporting Centroids:

The results of the KMeans analysis are examined by printing predicted cluster values (pred\_y) and reporting the centroids of the identified clusters. This information serves as a foundation for understanding cluster characteristics.

Checking Variance (Inertia):

The variance (inertia) of the KMeans analysis is assessed using kmeans.inertia\_, providing a quantitative measure of the compactness of the identified clusters.

## 3D Plotting for Visualization:

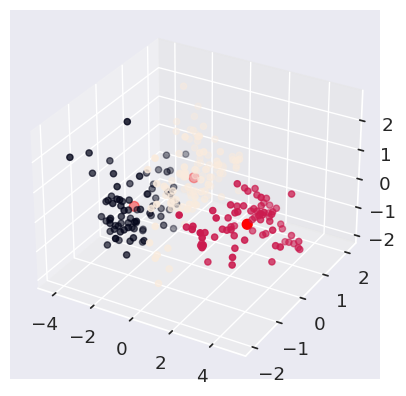


Figure 5 : the clusters in 3d space

A 3D plot is generated to visually represent the clustered data in the principal components space. Data points are colored according to predicted clusters, with KMeans centers highlighted in red for enhanced comprehension. This immersive visualization facilitates a holistic understanding of the clustering results.

# Chapter 4: Cluster Analysis with KMeans

*Introduction:* This report delves into the realm of predictive analysis, leveraging the power of supervised learning algorithms—Decision Tree and Logistic Regression. The objective is to predict the efficiency of cars based on crucial parameters, offering a comparative evaluation of the two models.

## Dataset Comparison:

* + The exploration begins with a comparative overview of the **cars2** and **dftest** dataframes, setting the stage for subsequent predictive analyses. Understanding the dataset nuances is fundamental to drawing meaningful insights.

## Decision Tree Analysis:

* + *Model Initialization:*
    - Parameters such as the total number of cylinders, engine volume, power, weight, and time to reach 60mph are chosen as inputs for the Decision Tree model.
  + *Training and Testing:*
    - The dataset is partitioned into training and testing sets, ensuring robust model evaluation.
  + *Model Evaluation:*
    - The Decision Tree model is trained and evaluated for accuracy, revealing a high accuracy score.
  + *Performance Metrics:*
    - In-depth analysis includes a confusion matrix and a detailed classification report, emphasizing the f1-score as a key metric for model quality.

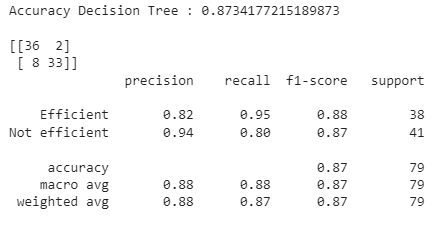


Figure 6 : Decision tree results

## Logistic Regression Analysis:

* + *Model Initialization:*
    - Linear Regression is employed to predict car efficiency, mirroring the Decision Tree setup.
  + *Model Evaluation :*
    - The Logistic Regression model demonstrates strong predictive capabilities, with an accuracy score matching the efficacy of the Decision Tree.
  + *Performance Metrics:*
    - Similar to the Decision Tree, a comprehensive examination includes a confusion matrix and a detailed classification report.

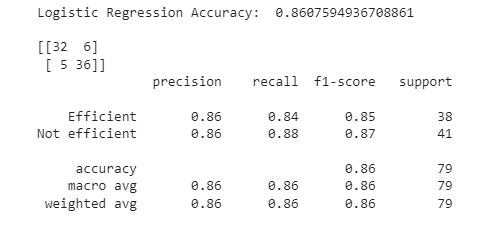


Figure 7 : Logistic regression results

## Comparative Analysis:

* + *Model Effectiveness:*
    - Both Decision Tree and Logistic Regression models prove effective in predicting car efficiency, showcasing similar metric values.
  + *Insights from Metrics:*
    - The high f1-score across both models signifies robustness in classification quality.

## Insights and Conclusions :

* + *Comparative Evaluation :*
    - The report concludes with a thoughtful comparison of the Decision Tree and Logistic Regression models. Despite their distinct methodologies, both models offer reliable predictions.
  + *Informed Decision-Making:*
    - The comparative analysis equips decision-makers with insights to choose the most suitable predictive algorithm based on specific requirements.
  + *Path Forward:*
    - The success of both models opens avenues for further exploration and fine-tuning to enhance predictive accuracy.

# Conclusion

To sum it up, this report delved into car data, from preprocessing and statistical analysis to dimensionality reduction using PCA and clustering with KMeans. The predictive analysis using Decision Tree and Logistic Regression models showed high accuracy, making both models reliable for predicting car efficiency. The comparative analysis highlighted their effectiveness, offering insights for decision-makers. In conclusion, the report provides valuable information for big data and business intelligence, guiding decision-making and suggesting avenues for future exploration.