TASK

Exploratory Data Analysis on the Automobile Data Set

INTRODUCTION

In this exploratory data analysis (EDA) report, we will be analyzing the Automobile Data Set. The aim is to extract stories and assumptions based on the visualizations of the data.

The Automobile Data Set:

Contains information on various attributes of different cars. This report is focused on conducting exploratory data analysis (EDA) to gain insights into the data and understand the relationships between the variables.

Summary of the data set:

The dataset contains information about different types of cars, including make, body style, number of doors, drive wheels, engine size, horsepower, curb weight, and price. The dataset has 205 rows and 26 columns.

DATA CLEANING

Data cleaning:

During data cleaning, I checked for missing data and handled it appropriately. I also conducted some visualizations to better understand the data.

Summary of the methods and visualizations done during data cleaning:

During data cleaning, I used various methods to check for missing data, including using the isnull() and sum() methods.
I also used visualizations such as Scatter Plot and Bar charts to help me identify missing data.

MISSING DATA

I found that there were some missing data in the data set, in the variables such as;
normalized-losses, num-of-doors, bore, stroke, horsepower, peak-rpm, and price, etc.

Missing data can be handled in various ways;

- Dropping the rows or
- Filling the missing values with appropriate data based on the data type (Imputation).

DATA STORIES AND VISUALISATIONS

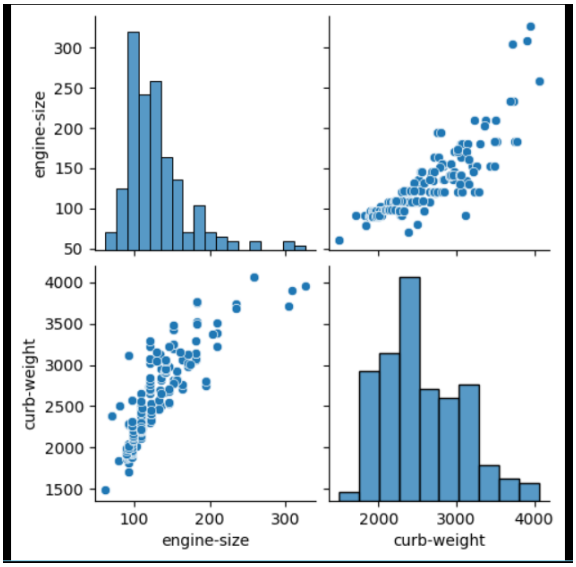
Code for the visualizations:

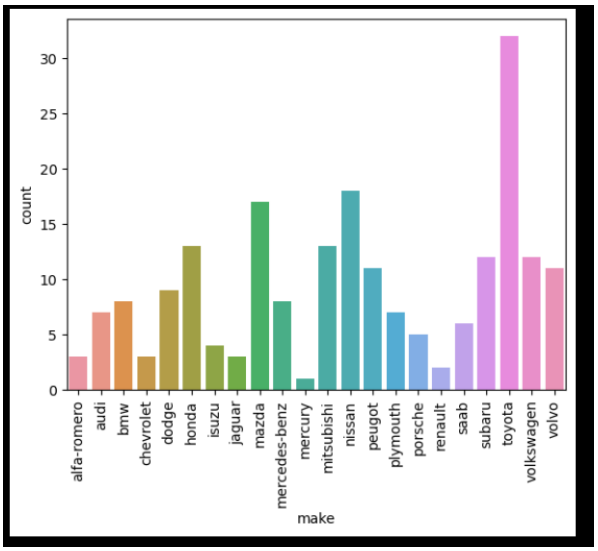
To visualize the data, we imported the necessary libraries such as:

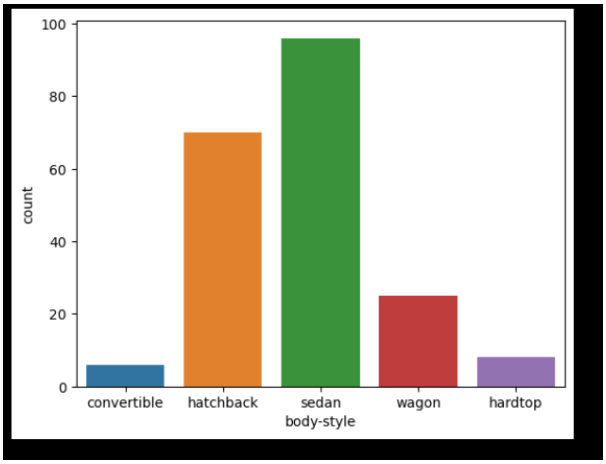
Numpy | pandas | seaborn | matplotlib.pyplot.

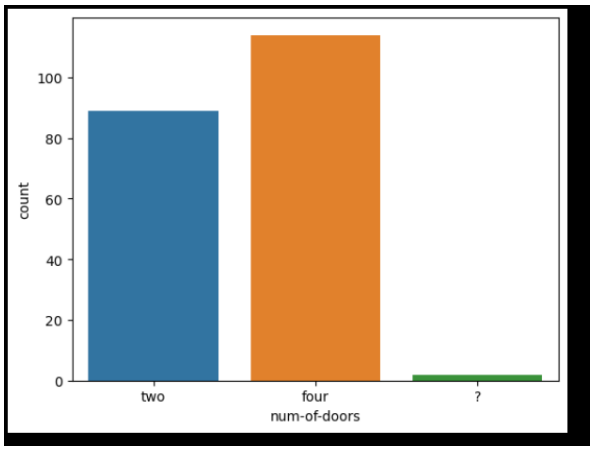
We then loaded the data set using pd.read_csv('automobile.txt') and

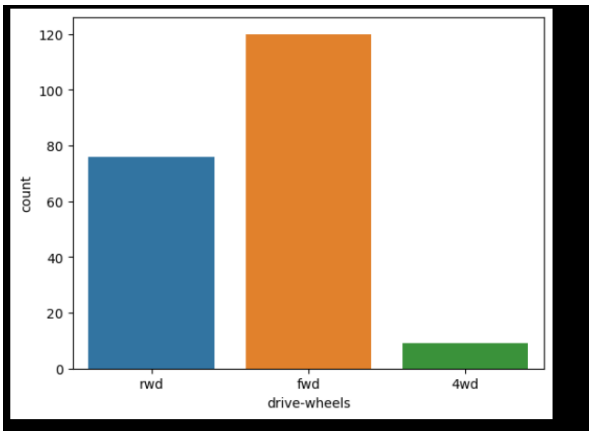
conducted various visualizations using *sns.pairplot*, *sns.countplot*, and *plt.show*.

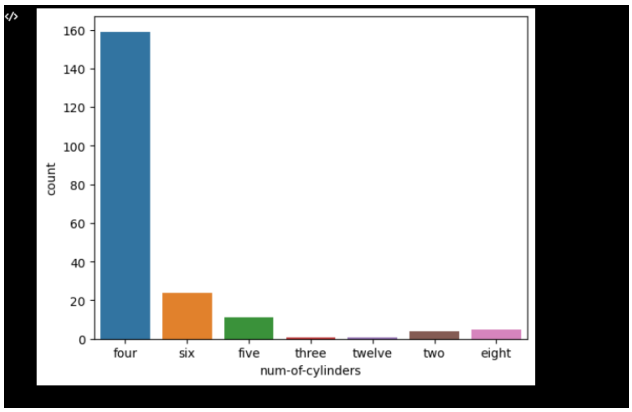












Data stories and visualizations:

During EDA, I conducted various visualizations to gain insights into the data.

Some of the visualizations included:

(a) Pair-Plot:

The pair-plot visualization was used to show the relationship between the variables: **price, engine-size, horsepower, and curb-weight**.

It is observed that there is a positive linear relationship between **price and engine-size, horsepower, and curb-weight**.

This means that as these variables increase, the price of the car also tends to increase.

(b) Count-Plot:

I used count-plot visualization to show the frequency distribution of categorical variables in the dataset. I looked at the distribution of the following variables:

Make:

I observed that Toyota and Nissan were the most common car makes in the dataset.

Body Style:

The most common body style was sedan, followed by hatchback.

Number of Doors:

The majority of cars in the dataset had four doors.

Drive Wheels:

Most cars in the dataset had a front-wheel drive.

Number of Cylinders:

Four-cylinder cars were the most common in the dataset.

CONCLUSION

Overall, This EDA report provides insights into the Automobile Data Set and helps us understand the relationships between the variables.

I conducted data cleaning, checked for missing data, and performed six visualizations to extract stories and assumptions based on the data.

Through the visualizations, I observed that there is a positive linear relationship between the price of a car and its engine size, horsepower, and curb weight.

I also observed that Toyota and Nissan were the most common car-make, while sedan was the most common body-style.

Additionally, the majority of cars in the dataset had four doors and a front-wheel drive.

THIS REPORT WAS WRITTEN BY:

Sameera Mohamed