Hyperiondev

TASK

Exploratory Data Analysis on the Automobile Data Set

Visit our website

Introduction

The automobile data set is data collected on motor vehicle manufacturers that looks at various factors.

These factors can be described as:

- Motor vehicle specifications that look at various areas such as body style, engine type, number of doors etc
- Normalised losses when compared to other vehicle models

DATA CLEANING

The first step when looking at this data like any data set is to load the required packages you'll be using such as numpy, pandas and matplotlib among others.
I then loaded the dataset by using the pd.read function and when the dataset was loaded showing the 1st 5 rows & 26 columns, we could see that there were special characters in the dataset which wouldn't be able to be read into the data by the pandas library.

To clean the data I used the following (as can be seen in the automobile.ipnyb notebook):

- df.drop: to drop any columns I wouldn't be using in the data analysis
- df.drop_duplicates: to drop any duplicate rows
- temp.df: after cleaning the data & taking care of missing data, we then created a temp dataset from which we can carry out our exploratory data analysis

After creating the temp.df, I then did some data cleaning such as:

- Converting missing values in the price data into integers so that it can make more sense in our data analysis
- I also converted the data in the normalised losses into integers as this is data I wanted to use

MISSING DATA

With missing data I used the df.replace & df.isnull functions to see where there was missing data & decide on whether I could draw inferences from the data despite the missing data.

Missing data was found:

- normalized-losses: 41
- num-of-doors: 2
- bore: 4
- stroke: 4
- horsepower: 2
- peak-rpm: 2
- price: 4

DATA STORIES AND VISUALIZATIONS

Code for the visualizations can be referred to: automobile.ipynb jupyter notebook

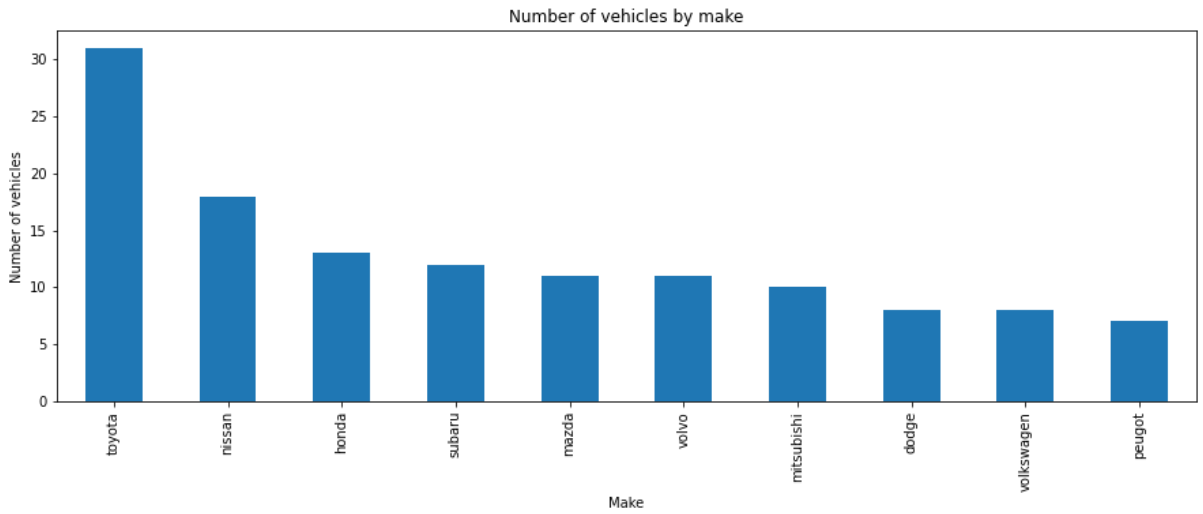

Figure 1: Bar graph showing Number of vehicles produced by each car maker

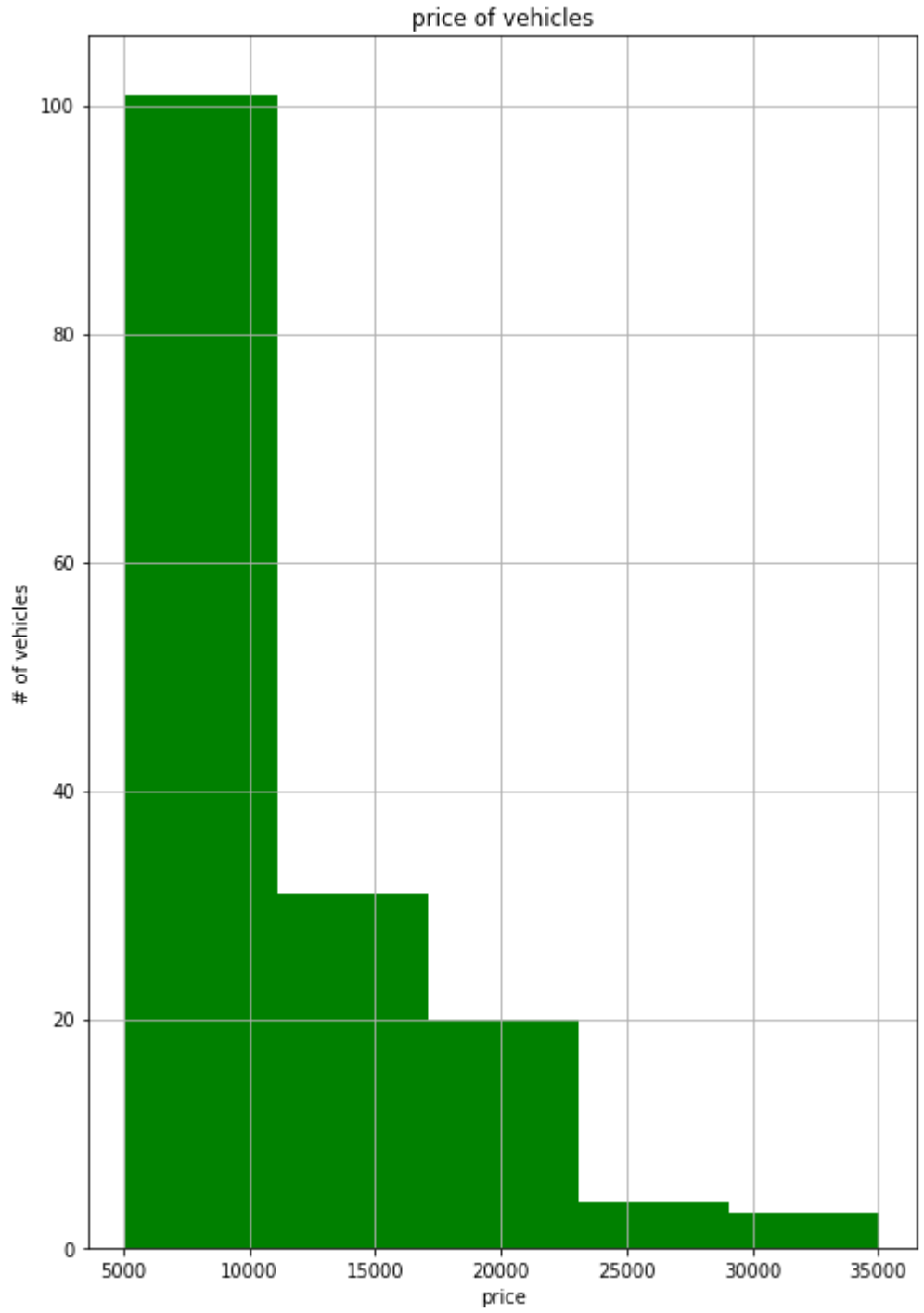

Figure 2: Histogram showing prices

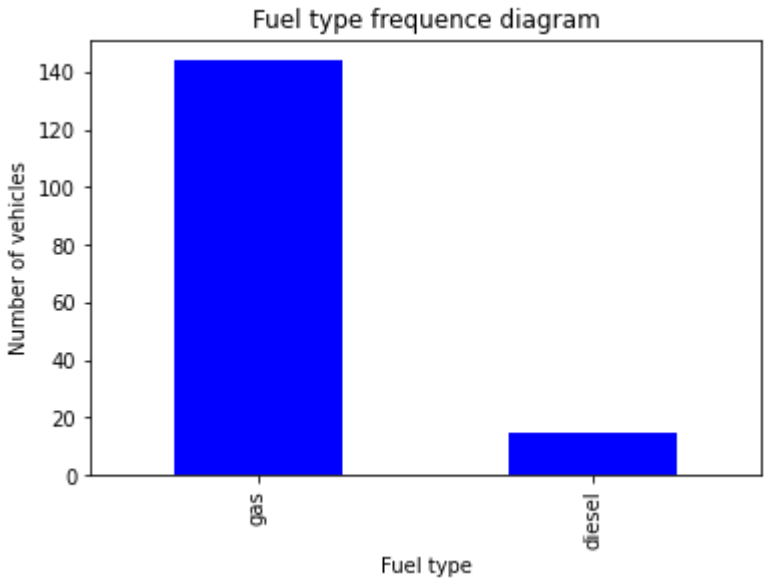

Figure 3: Bar graph showing fuel type

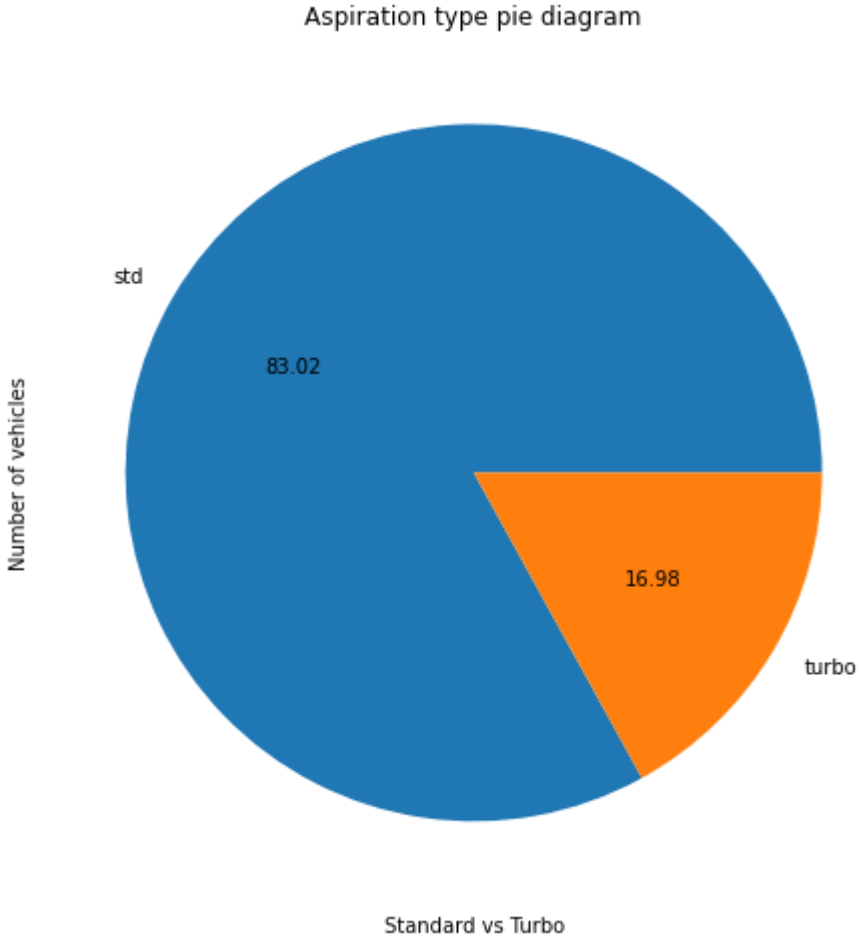

Figure 4: Pie chart showing Standard engine vs Turbo engine

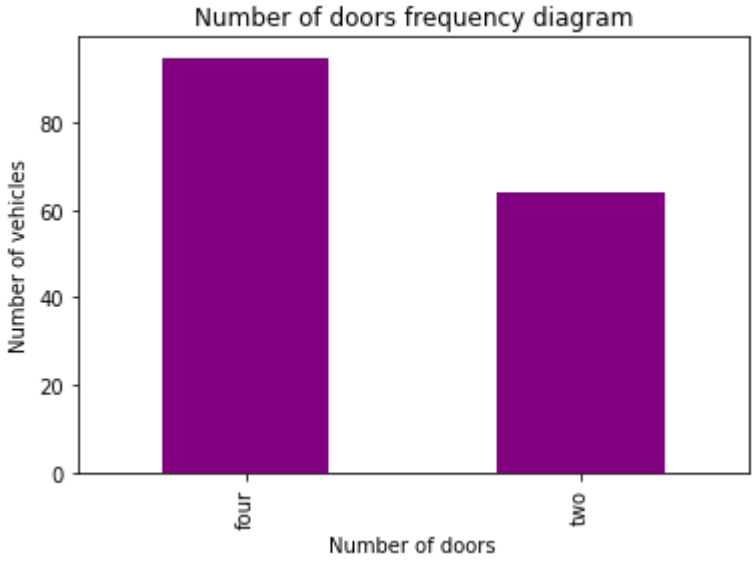

Figure 5: bar chart of 2 door vs 4 door vehicles

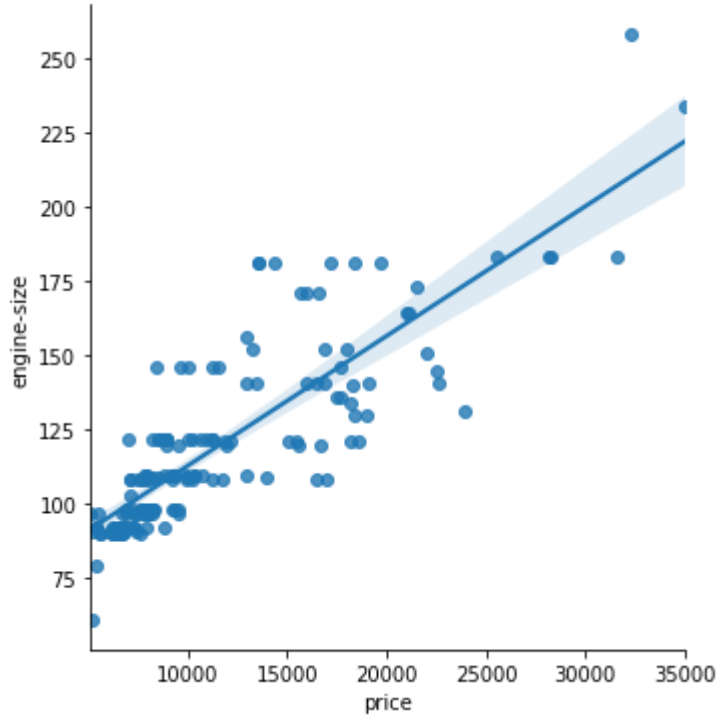

Figure 6: Scatter plot showing correlation between engine size & price

FINDINGS:

- **The top 3 carmakers in our dataset are Japanese: Toyota, Nissan & Honda with Toyota being the biggest car maker**
- **Within our dataset we can see that most of these cars are entry level cars with majority of cars sitting between the 5000 - 20000 price range**
- **We don't have many cars in the top end of the market**
- **Majority of our dataset has gas cars more than diesel cars**

- The data has more standard engine cars than there are turbo cars
- There's more 4 door cars than 2 door cars in the dataset
- Small engine cars cost less than big size engines showing us that the bigger the engine, the more it costs

CONCLUSIONS:

1. There's factors such as engine size that can affect the price of a car
2. Japanese carmakers mass produce vehicles
3. Smaller engines can point towards the dataset having cheaper cars
4. There's different factors which can affect the price of a car
5. Data set distribution will affect how we draw conclusions

THIS REPORT WAS WRITTEN BY : Matete Nchabeleng