**EXPLORATORY DATA ANALYSIS**

**ON THE**

**AUTOMOBILE DATASET**

INTRODUCTION

The dataset consists of a wide range of 26 data points.

The 26 column headings in the dataset are:

“***symboling, normalized-losses, make, fuel-type, aspiration, num-of-doors, body-style, drive-wheels, engine-location, wheel-base, length, width, height, curb-weight, engine-type, num-of-cylinders, engine-size, fuel-system, bore, stroke, compression-ratio, horsepower, peak-rpm, city-mpg, highway-mpg, price***”.

DATA CLEANING

A series of functions were used to investigate the dataset to identify data-points requiring cleaning.

The primary functions were:-

* .describe()
* .info()
* .nunique()
* .head()

MISSING DATA

Using the functions mentioned above missing values were identified as being filled in with a question-mark (‘?’) character.

These were the values identified as missing and filled-in with a ‘?’ character.

- normalized-losses  (164 non-null)    = 41

- num-of-doors       (203 non-null)    = 2

- bore               (201 non-null)    = 4

- stroke             (201 non-null)    = 4

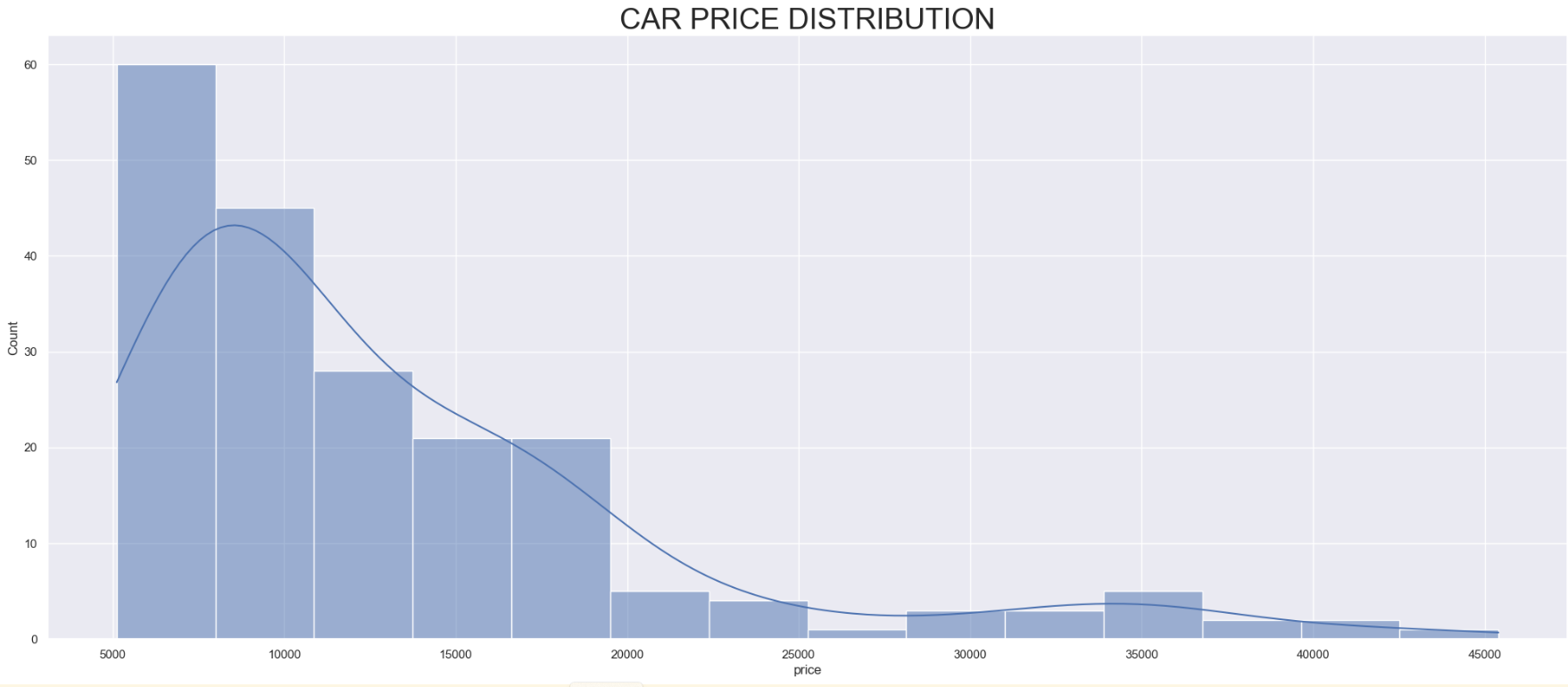
- horsepower         (203 non-null)    = 2

- peak-rpm           (203 non-null)    = 2

- price              (201 non-null)    = 4

The missing/(’?’) values were replaced with NaN values when reading the file into the dataframe using an (na\_values='?') argument in the pd.read\_csv() function.

DATA STORIES AND VISUALISATIONS



PRICE

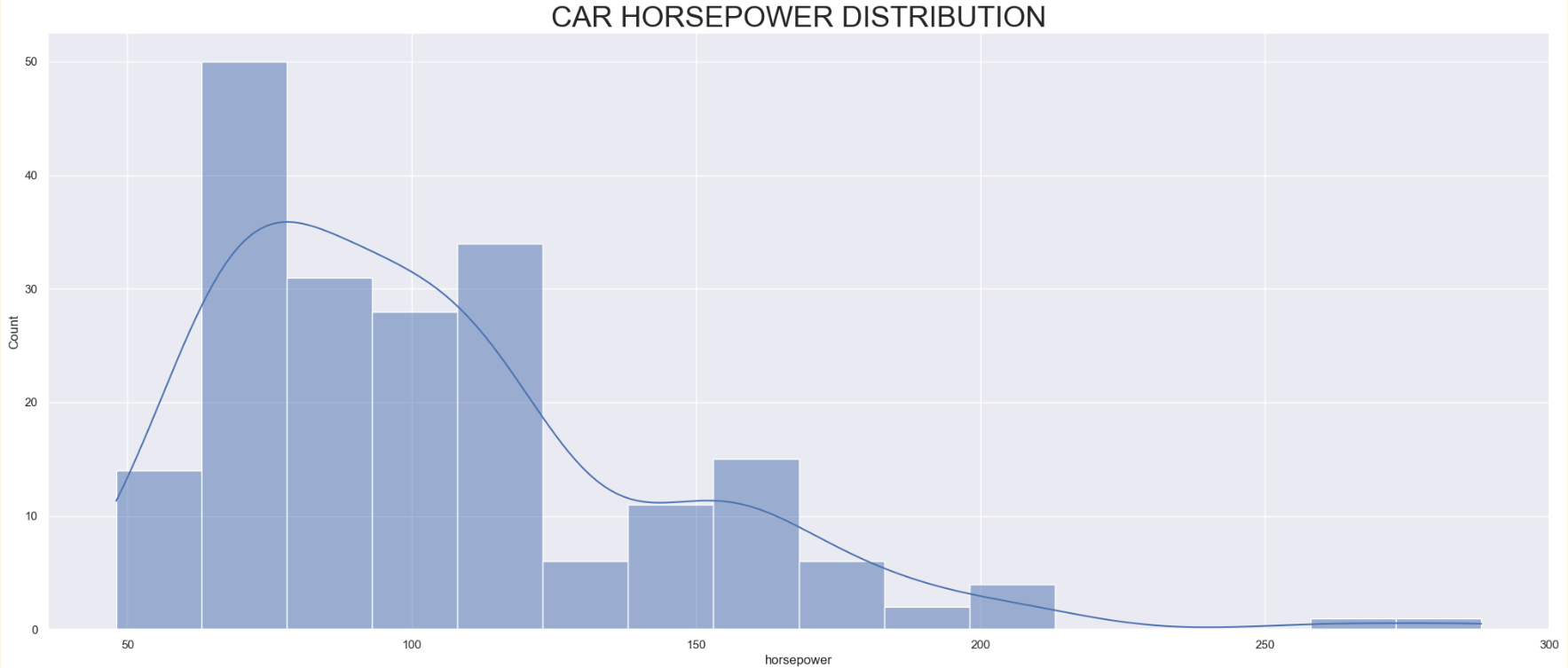
A large distribution of cars is in the price categories at the lower end of the scale.

- Most car prices are located below the 20,000 value.

- It is unsurprising that lower value cars are more popular because they are more affordable for customers.

- It is surprising that the lowest priced cars make up so much of the total because they are often made to a price and can be unreliable or underpowered.

- It would be interesting to look at the horsepower of the cars in the dataset.



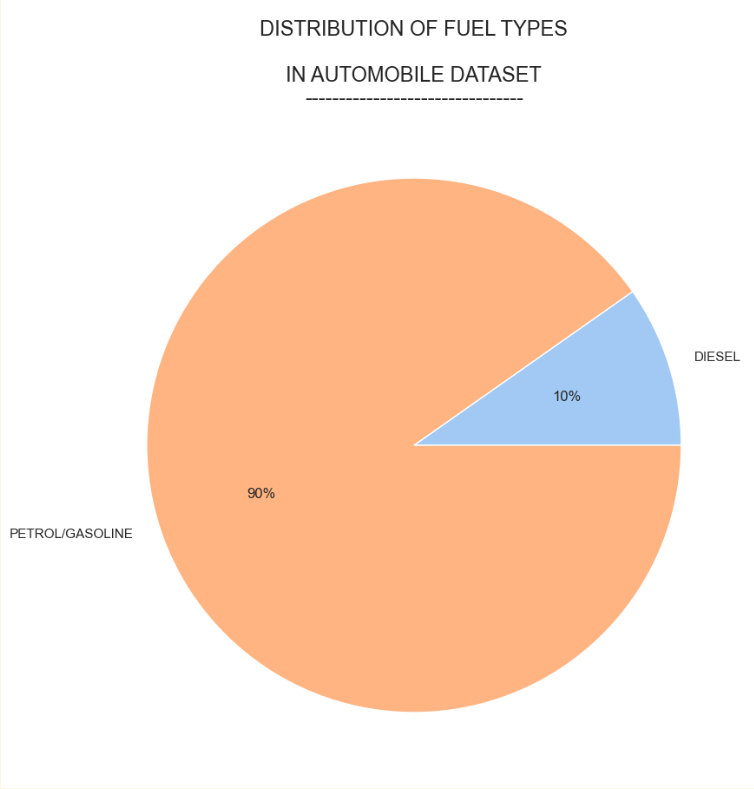
HORSEPOWER

As anticipated a large number of the cars in the dataset are at the lower end to the middle of the horsepower scale.

- As most cars sold are lower to mid-priced it makes sense a lower powered engine is a feature of lower to mid-priced cars.

- The correlation is not a direct 1:1 correspondence because looking at the lowest powered vehicles in the above graph does not correspond exactly with the previous ‘price’ graph in which the largest number is in the lowest priced bracket of cars.

- This suggests that not all lower-priced cars have the weakest powered engines and that manufacturers provide customers choice in there consumer offering of lower-priced cars.



FUEL-TYPES

The distribution of fuel-types is surprising because in this dataset ‘Diesel’ fuel has only a 10% market share. ‘Petrol’ has a 90% market share.

- The tiny market share of Petrol is surprising because for many years Diesel has been a popular fossil fuel for automobiles. This dataset indicates a skewed preference towards Petrol.

- Diesel became a more popular fuel in the 1990s to the mid-2000s because it was considered more environmentally-friendly as it was more fuel-efficient and reduced CO2 emissions compared to petrol.

- It would be interesting to compare the popularity of fuel-types for automobiles over time and to identify the factors influencing fashions in fuel-types.

This report was written by: Prabhjit Gill (01/02/2023)