# Used car Dataset ReadMe

## by Dike Ujunwa

## Dataset

> Provide basic information about your dataset in this section. If you selected your own dataset, make sure you note the source of your data and summarize any data wrangling steps that you performed before you started your exploration.

The dataset has 7907 records of cars and 19 columns showing the features, attributes, characteristics, sales status and mechanical properties of each car in the dataset. The dataset is also publicly available and obtainable from Kaggle at (https://www.kaggle.com/datasets/shubham1kumar/usedcar-data?select=UsedCarData.csv). The main feature of interest in this dataset are year of purchase, selling price or car worth and whether the cars have been sold or not. The key features that can support this investigation would be the non-mechanical properties of cars such as seats number, region, name, and mechanical features like mileage, engine, max_power, torque, max_rpm etc.

The wrangling process included assessment and cleaning stages of the dataset:

The dataset was assessed visually for some potential data issues. Some of the issues assessed include mission values, presence of '0s', presence of abbreviated entries etc. In the programmatic assessment stage, the dataset was manipulated using code to spot potential and cleanable data issues. Some of the codes included a. head() b. query() c. info() d. sum() e. duplicated() f. describe() g. value_counts() etc. In addition, some visualizations where used to further assess and demonstrated some of these assessments. Major issues assessed were the incorrect datatypes and the presence of abbreviated strings.

## Summary of Findings

The dataset revealed some major findings.

1. 25% of the cars in the dataset has been sold while the remaining 75% are still unsold. Interestingly, the mean price for both types of cars are almost equivalent at $676.60bn and $642.6bn respectively

2.The total worth of all cars at their purchase are$5.137bn.

3. The minimum and maximum price for cars in the dataset are $29.999k and $10bn respectively.

4. Purchase of cars were majorly done between 2010 and 2020.

5. The selling price of cars also increased as the engine size increases

6. Over the course of time, the features of cars purchased increased, this included engine sizes, car power, mileage and selling price

7. The Test_drive_cars also showed to be most expensive reaching up to $6m

8. There's a negative correlation between the car power and the fuel per distance consumed by a car. This explains that cars of lower horsepower would use high amount fuel per distance during movement, this in turn would increase demand of fuel on the car user.

9. Most cars are first-owner type cars.

10. Four-seater unsold vehicles (probable for family use) were cheapest in the West.

11. Majority of the cars have low prices below 4million have equivalently low engine sizes and most of them are five-seater cars.

12. None of the test_drive_cars (which are the most expensive type of cars in relation to car ownership) have been sold out. This makes sense as they were still test driven.

13. Majority of the cars in the data set are first_owned car and more than 80% of these cars are five-seater cars.

14. Cars with fuel not petrol or diesel are generally cheap

## Key Insights for Presentation

The used_car dataset contain a lot of explorable features and points. The steps taken during the exploration were aimed at comparing the distribution of features of interest and other(supplementary) features. First, the distribution of the different data types; continuous, discrete and categorical data. Then, the bivariate exploration examined the nature of correlation between the mechanical features of the car as well as the distribution of these features over the categorical data. Finally, at the multivariate exploration stage, the insights gotten from the bivariate exploration were used to further explore the addition of a third variable(both categorical and discrete).

Key insights from the dataset include

1. The prices of purchased cars increased over time.
2. The engine sizes determine the car prices, in other words, the higher the engine sizes the higher the prices
3. The increase in engine sizes over time could have resulted to increase in car prices, this is because of the observed strong correlation between engine sizes and selling_price
4. Seven-seater cars have higher engine sizes and five-seater have lower engine sizes
5. Cars sold by car dealers are generally higher in prices than the rest types of car by seller-type
6. Generally, the car power/horsepower and engine size majorly determined the prices of car in the dataset.