2/16/2025

# EXPLORATORY DATA ANALYSIS

WHICH CAR SHOULD YOU BUY?

Data Nerd
OKELLO RAYMOND

# EDA REPORT: CAR SALES ANALYSIS

## 1 INTRODUCTION

### 1.1 OBJECTIVE

The goal of this exploratory data analysis (EDA) is to analyze a car sales dataset to uncover patterns, trends, and insights related to car specifications, fuel efficiency, and pricing.

### 1.2 DATASET OVERVIEW

- **Source**: Click Here
- **Size**: 8286 rows and 16 columns
- **Features**:
    - **Make**: Manufacturer of the car (e.g., Toyota, BMW).
    - **Model**: Specific model of the car.
    - **Year**: Manufacturing year of the car.
    - **HP**: Horsepower of the car's engine.
    - **Cylinders**: Number of cylinders in the engine.
    - **Transmission**: Type of transmission (e.g., automatic, manual).
    - **Drive Mode**: Type of driven wheels (e.g., all-wheel drive, front-wheel drive).
    - **MPG-H**: Miles per gallon on the highway.
    - **MPG-C**: Miles per gallon in the city.
    - **Price**: Manufacturer's Suggested Retail Price (target variable).
    - **Car Age**: Age of the car since manufacturing
    - **Mpg**: Average miles per gallon in city and highway
    - **Price Segment**: Budget, mid-range, premium and luxury
    - **Age segment**: Old, Use, fairly new and new
    - **Hp segment**: Economy, standard, performance, sports
    - **Fuel segment**: gas guzzler, average, efficient
- **Tools and libraries:** Python, pandas, seaborn, matplotlib
- **Jupyter Notebook Code:** Click Here

## 2 DATA PREPROCESSING

### 2.1 DATA CLEANING

Data cleansing, also referred to as data cleaning or data scrubbing, is the process of fixing incorrect, incomplete, duplicate or otherwise erroneous data in a data set

- Created a function in python that reads and wrangles the dataset
- Dropped columns. Engine Fuel Type, Market Category, Vehicle style, Popularity, Number of doors, Vehicle Size doesn't make any sense to me so they were dropped.
- Renamed columns. To improve readability, I renamed the columns to a more human format for easy understanding of the dataset
- Since price column is read as object, converted the price columns to numbers.
- Transformed all the column names to small letters (lower cases).
- Dropped duplicates from the dataset.
- Dropped all the missing/null values.
- Calculated and dropped of the outliers from numerical columns.

### 2.2 FEATURE ENGINEERING

Feature engineering is the process of selecting, manipulating and transforming raw data into features that can be used in supervised learning

- Created a new feature **Car Age** column which is Current Year - Manufacturing Year.
- Since there are over 600 model types, calculated the most common 19 model types and the rest were labeled as 'Others' totaling to 20 model types.
- Same thing with make, reduced to 27 makes, applying others to model types which doesn't reach 100 in the dataset.
- Segmented based on age, price, hp and fuel to define simple categories that is understandable in car terms

# 3 EXPLORATORY DATA ANALYSIS (EDA)

In this section, we dive into the Exploratory Data Analysis (EDA) to uncover patterns, trends, and relationships within the car sales dataset. By visualizing key attributes such as price, age, horsepower, fuel efficiency, and drive modes, we gain insights into how these factors influence car values. The goal is to identify meaningful trends that can help both buyers and sellers make informed decisions.

## 3.1 UNIVARIATE ANALYSIS

### 3.1.1 Distribution of car price.

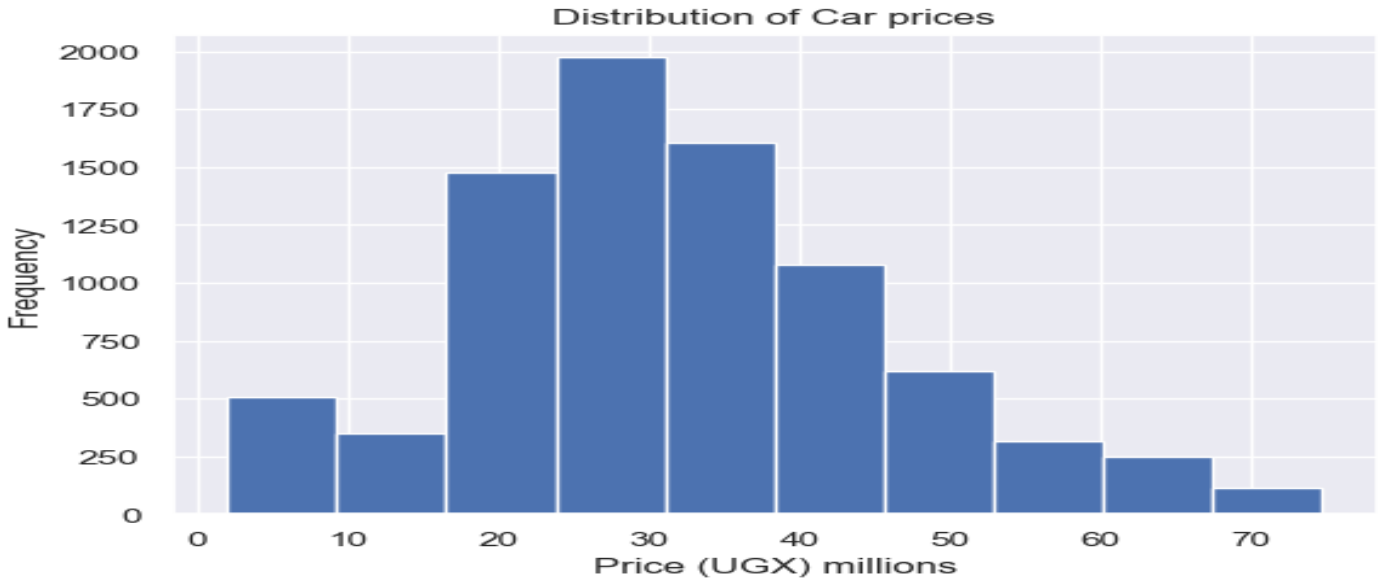



*Figure 1 The distribution of Car prices in the listing*

The histogram shows the distribution of car prices, revealing that most vehicles fall within the mid-range price category (UGX 20M-40M). The distribution appears right-skewed, indicating that while a few luxury cars exceed or equals to UGX 50M, the majority of listings are in the mid range to premium segment."

- Some of the budget cars in the listing includes **Ford, Chevrolet, Mazda, GMC and others**
- Options in the mid range and premium segment includes **Chevrolet, Ford, Toyota, Volkswagen, Nissan**
- In the luxury section, some of the classy cars are found in the **Cadillac, Mercedes-Benz, Chevrolet, BMW, Audi makes**

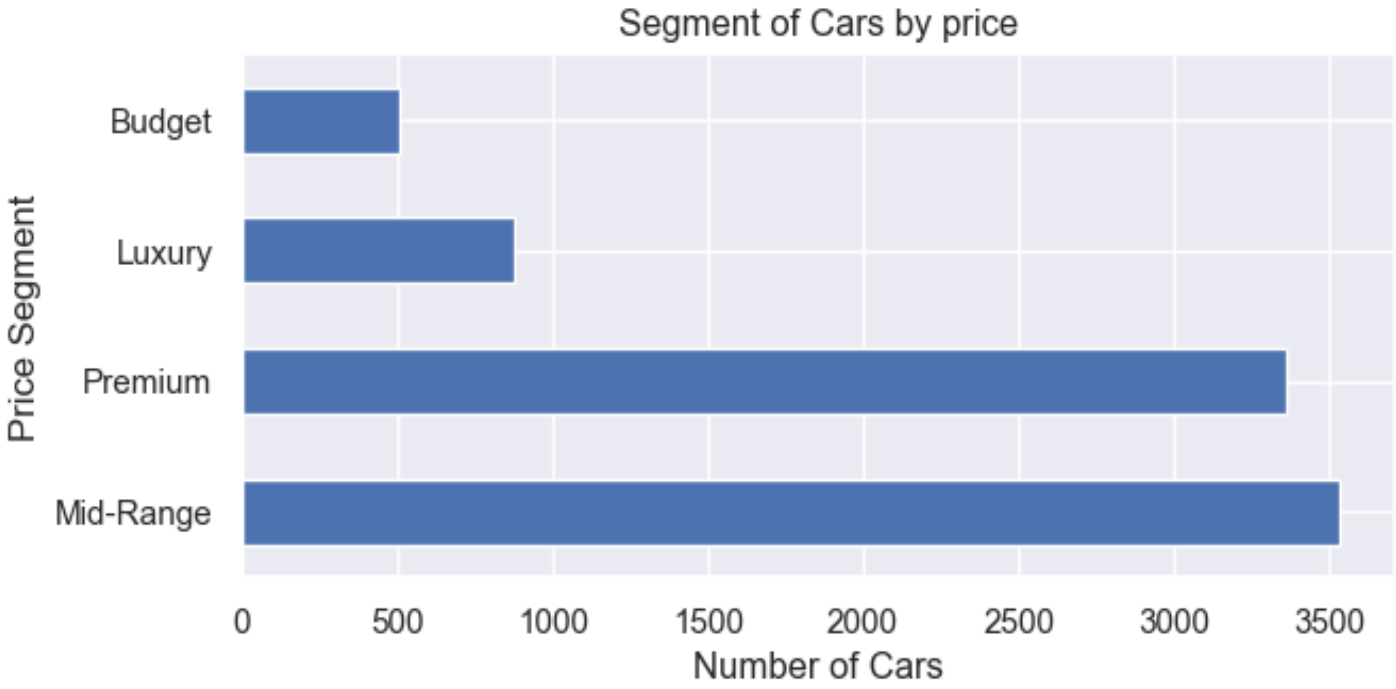



*Figure 2 The most popular price segments available in the dataset*

### 3.1.2 Distribution of car makes.

The distribution of car makes shows that a few brands dominate the dataset, with Chevrolet being the most common. This suggests that certain manufacturers have a larger market presence, likely due to affordability, reliability, or brand preference. Meanwhile, less frequent makes may represent niche or luxury brands with lower overall sales. The imbalance in distribution could also indicate market trends favoring specific manufacturers.

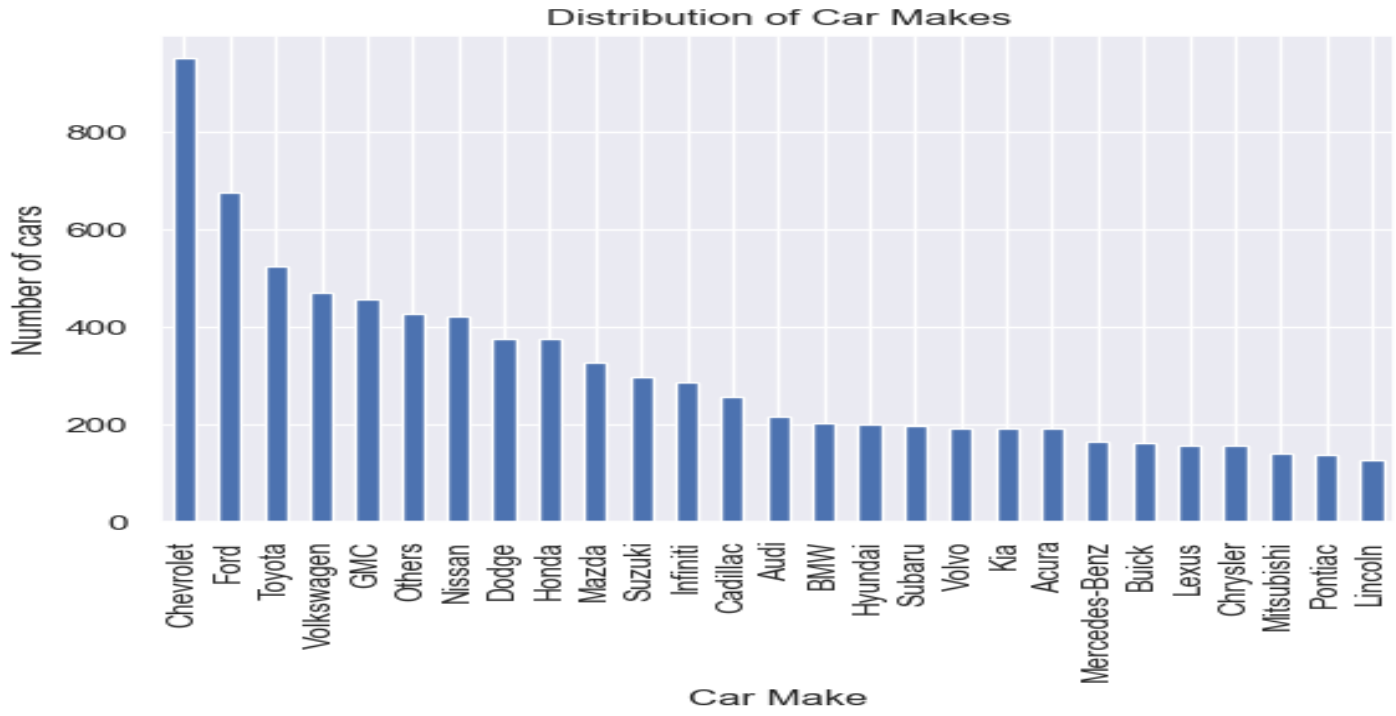



*Figure 3 Popular cars in the car listing*

## 3.2 BIVARIATE ANALYSIS

### 3.2.1 Car Price Vs Car Age:

The relationship between car price and age shows a clear depreciation trend—newer cars tend to have higher prices, while older vehicles lose value over time. Most cars above 25 years old rarely sell for more than UGX 10M, indicating a steep decline in value as they age. However, some exceptions exist, likely due to classic or high-performance models retaining their worth.

- If you are on a budget, some popular old cars include the following **Chevrolet, Ford, Mazda, GMC and others**
- Some of the popular used cars are in the listing are **Chevrolet, Dodge, Suzuki, Ford, Infiniti**
- The new and fairly new cars available are **Chevrolet, Toyota, Ford, Volkswagen, Nissan**


*Figure 4: Price decreases with the age of the car*

### 3.2.2 Car Price vs Horse Power

The analysis of horsepower (HP) and price reveals a positive correlation—cars with higher HP generally command higher prices. This trend suggests that performance-oriented vehicles, such as **Chevrolet, Cadillac, Infiniti, BMW, Mercedes-Benz,** are priced at a premium or above. In contrast, lower-HP vehicles, typically economy or standard models, tend to be more affordable. However, some variations exist, likely influenced by factors such as brand, model, and additional features.

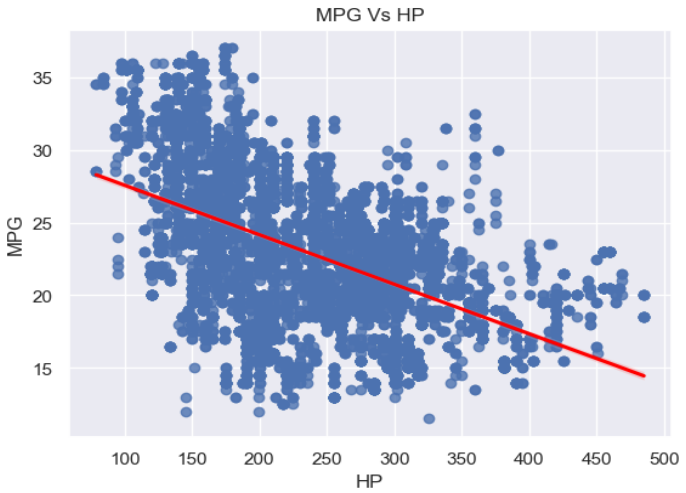
- In the performance section, Chevrolet, Ford, GMC, Toyota, Nissan takes up the chart while the classy sports make is available in **Chevrolet, Cadillac, Infiniti, BMW, Mercedes-Benz**
- For the best horse power and mid range price, customers can go for a **Toyota, Dodge, Chevrolet or a Ford**

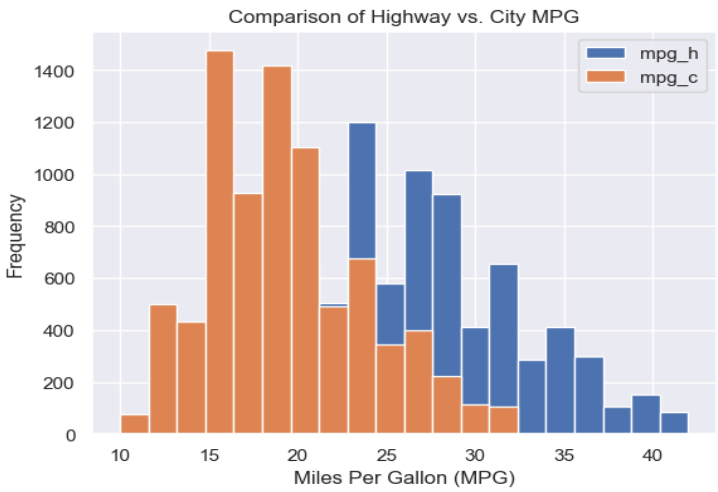

*Figure 5 Higher Horse Power (HP) means a higher price for the car*

### 3.2.3 Fuel or Power?

Not only are cars with high horse power (HP) expensive but they are also very expensive in fuel, cars with lower horse power tends to go longer miles per gallon compared to cars with higher horse power.

- However, there are a few cars with good horse power and some efficient fuel consumption. These cars include the **Infiniti, Mercedes-Benz and the Acura**
- If you're looking to save on fuel, keep in mind that city driving burns more gas than highway driving. The constant stopping, braking, and idling in traffic make your car work harder, meaning fewer miles per gallon.
- On the highway, it's a different story—steady speeds and fewer stops let your engine run more efficiently, stretching your fuel further. So, if most of your driving is in the city, a fuel-efficient car might save you a lot in the long run



### 3.2.4 Car Price vs Drive mode

FWD cars are the most affordable, with fewer high-end options. AWD and RWD tend to be pricier due to better performance and handling. 4WD vehicles also have a higher price range, built for off-road use. If you want affordability, go for FWD; for performance or traction, AWD, RWD, or 4WD will cost more.

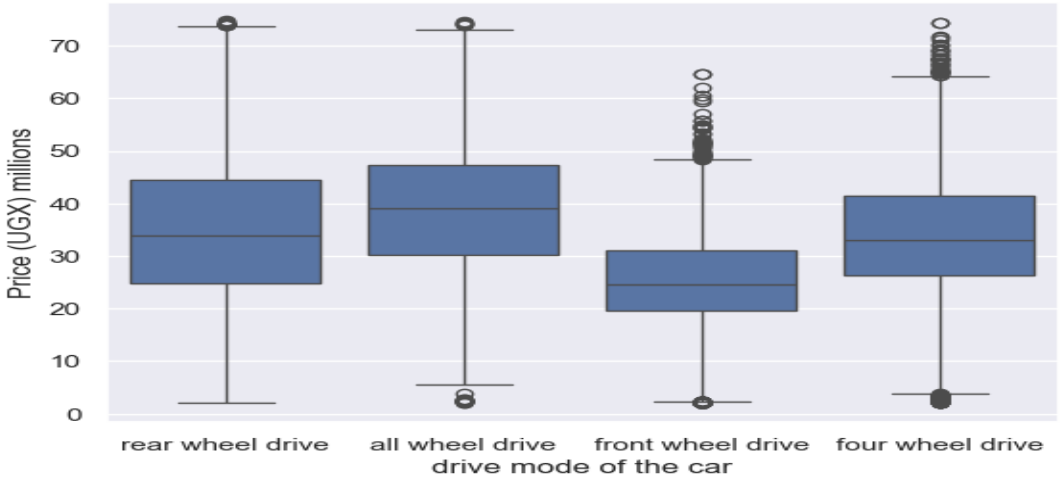


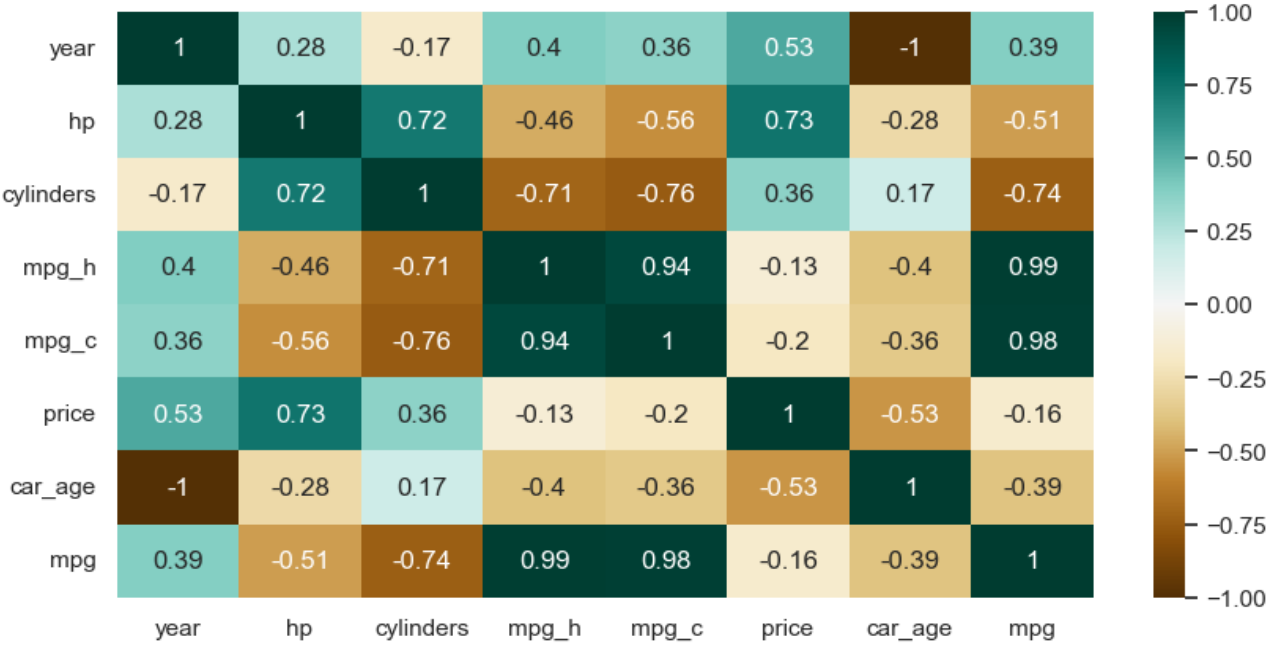

*Figure 6 Drive mode and price of cars*

## 3.3 MULTIVARIATE ANALYSIS

### 3.3.1 Heatmap

The heatmap helped in feature selection by revealing **correlations between numerical variables**. Key takeaways:

- **Car Age & Price** showed a strong negative correlation, confirming that older cars lose value.
- **HP & Price** had a positive correlation, making horsepower an important factor in pricing.
- **HP & MPG** showed an inverse relationship, highlighting the trade-off between performance and fuel efficiency.



## 4 LIMITATIONS

- The dataset may not include all relevant features, such as location or seller type.
- The dataset may be imbalanced, with certain car makes or models overrepresented.

## 5 CONCLUSION

From this analysis, it's clear that multiple factors influence car prices, including age, horsepower, fuel efficiency, and drive mode. For individuals likely to buy a car, these are some of the insights uncovered from the exploratory data analysis above

- Newer cars cost more, but older cars (25+ years) are more affordable. If you're on a budget, consider a slightly older model.
- Cars with higher MPG-H (highway) and MPG-C (city) save you money on gas. Look for cars with MPG-H above 20 and MPG-C above 25 for maximum savings.
- Higher HP means better performance but lower fuel efficiency. There is a trade off in higher or lower with fuel consumption. This must be highly considered
- AWD is great for rough terrain, FWD is fuel-efficient for city driving, and RWD is better for performance. Pick FWD for city driving or AWD if you need off-road capability.