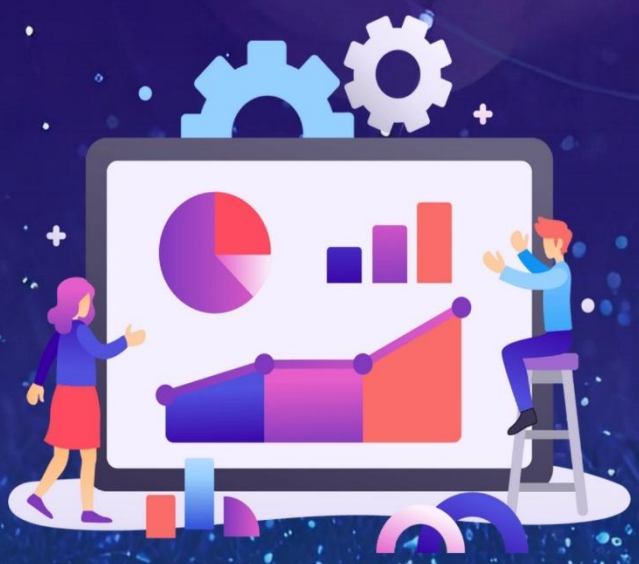
Prodigy InfoTech Internship



Data Science

Note:
It's not compulsory to use the provided dataset

Track: DS

Name: Suriya. B
Linkdin: www.linkedin.com/in/suriya0210
Email: bsuriya223@gmail.com

PRODIGY INFOTECH TASK-2

Perform data cleaning and exploratory data analysis (EDA) on a dataset of your choice, such as the Titanic dataset from Kaggle. Explore the relationships between variables and identify patterns and trends in the data.

---

1. Abstract

This project analyses a Toyota car dataset with the goal of understanding key factors influencing used car pricing and characteristics. Using exploratory data analysis (EDA), the study applies descriptive statistics, bar charts for categorical features, and histograms/scatter plots for numerical variables. The analysis highlights relationships between price, age, mileage, fuel type, horsepower, and other attributes. These insights can assist buyers, sellers, and dealerships in identifying market trends and making informed decisions.

---

2. Introduction

The automobile industry generates massive amounts of transactional and performance-related data. Effective analysis of such data helps stakeholders recognize customer preferences, vehicle depreciation patterns, and pricing trends. This project focuses on the Toyota dataset, which contains details about used cars including specifications, fuel type, mileage, and pricing. By applying data cleaning and visualization, the project transforms raw car listings into meaningful insights regarding consumer demand and pricing dynamics.

---

3. About the Dataset

Source & Context

The dataset originates from car listings (Toyota Corolla) and provides information about vehicle attributes and market prices.

Key Statistics

- Variables include Price, Age, Mileage (KM), Fuel Type, Horsepower, Metallic Colour, Automatic/Manual, and others.
- Each observation represents a single car listing.

Content Overview

- **Price (Target Variable)**: Indicates resale value of each vehicle.
- **Vehicle Characteristics**: Age, KM driven, HP (Horsepower), Automatic/Manual.
- **Fuel Type**: Petrol, Diesel, CNG.
- **Other Attributes**: Metallic colour, number of doors, quarterly tax, weight.

---

4. Data Preparation

Before conducting EDA, preprocessing steps were taken:

- **Loading & Inspection**: Data imported using Pandas.
- **Handling Missing Values**: Checked for null entries and duplicates.
- **Data Type Corrections**: Converted categorical variables into consistent labels.
- **Feature Standardization**: Renamed and formatted variables for easier interpretation.

- **Outlier Identification**: Examined extreme values in price, mileage, and horsepower.

---

5. Exploratory Data Analysis & Visualizations

5.1 Price Distribution (Histogram)
Displays how resale prices are spread, identifying common pricing ranges.

5.2 Age Distribution (Histogram/Bar Chart)
Shows vehicle age patterns, useful for understanding depreciation trends.

5.3 Fuel Type (Bar Chart)
Highlights proportions of Petrol, Diesel, and CNG cars in the dataset.

5.4 Transmission Type (Bar Chart)
Compares Automatic vs Manual cars listed for sale.

5.5 Mileage (Histogram/Boxplot)
Analyses how far cars have been driven and identifies high-mileage outliers.

5.6 Price vs Age (Scatter Plot)
Explores depreciation — older cars generally show lower prices.

5.7 Price vs Mileage (Scatter Plot)
Reveals how higher mileage tends to reduce car value.

5.8 Fuel Type vs Price (Boxplot)
Examines if Diesel, Petrol, or CNG vehicles have different resale value patterns.

---

6. Insights & Discussion

From the EDA, some general findings are:

- Price decreases significantly with age and mileage.
- Petrol cars dominate the dataset, while CNG and Diesel form smaller segments.
- Automatic cars are fewer compared to manual but tend to be higher priced.
- Car weight and horsepower show positive correlation with price (heavier/more powerful cars valued higher).

---

7. Conclusion

This project provides a data-driven overview of the Toyota car resale dataset. The analysis uncovers depreciation trends, consumer preferences in fuel and transmission type, and pricing factors. These findings can help used-car sellers optimize pricing strategies and guide buyers in assessing fair market value. Exploratory data analysis proves to be a powerful tool for extracting actionable insights from automotive datasets.

---

8. References

- Dataset: Toyota Corolla dataset (open-source, available on Kaggle and UCI repositories).
- Tools: Python (Pandas, Matplotlib, Seaborn).

Importing the library function and csv file into python:

```
# ================================================================================
# Data Cleaning & EDA on Toyota Used Car Dataset
# ================================================================================

import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from mpl_toolkits.mplot3d import Axes3D

# Style
sns.set(style="whitegrid", palette="Set2")
plt.rcParams["figure.figsize"] = (8,5)

# ================================================================================
# Load dataset
# ================================================================================
df = pd.read_csv(
    "C:\\Users\\SURIYA\\Downloads\\Toyota.csv",
    index_col=0,
    na_values=["??", "????"]
)

# Backup copy
toyota = df.copy()
```

df	DataFrame	[1436, 10]	Column names: Price, Age, KM, FuelType, HP, MetColor, Automatic, CC, D ...
toyota	DataFrame	[1436, 10]	Column names: Price, Age, KM, FuelType, HP, MetColor, Automatic, CC, D ...

Data cleaning:

```
# ================================================================================
# Data Cleaning
# ================================================================================
# Fill missing values
for col in toyota.columns:
    if toyota[col].dtype in ["float64", "int64"]:
        toyota[col].fillna(toyota[col].median(), inplace=True)
    else:
        toyota[col].fillna(toyota[col].mode()[0], inplace=True)
```

Interpretation:

- **Numeric columns** (like price, age, mileage) → missing values are replaced with the **median**, which is robust against outliers.
- **Categorical columns** (like fuel type, color) → missing values are replaced with the **mode** (most frequent value), since that best represents the most common category.

Data summary:

```
# ================================================================================
# Dataset Summary
# ================================================================================
print("\n--- Structure of Data ---")
print(toyota.info())

print("\n--- Missing Values ---")
print(toyota.isnull().sum())

print("\n--- Statistical Summary ---")
print(toyota.describe().T)
```

```
--- Structure of Data ---
<class 'pandas.core.frame.DataFrame'>
Index: 1436 entries, 0 to 1435
Data columns (total 10 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   Price       1436 non-null   int64
 1   Age         1436 non-null   float64
 2   KM          1436 non-null   float64
 3   FuelType    1436 non-null   object
 4   HP          1436 non-null   float64
 5   MetColor    1436 non-null   float64
 6   Automatic   1436 non-null   int64
 7   CC          1436 non-null   int64
 8   Doors       1436 non-null   object
 9   Weight      1436 non-null   int64
dtypes: float64(4), int64(4), object(2)
memory usage: 123.4+ KB
```

```
--- Missing Values ---
Price        0
Age          0
KM           0
FuelType     0
HP           0
MetColor     0
Automatic    0
CC           0
Doors        0
Weight       0
dtype: int64
```

```
--- Statistical Summary ---
            count          mean           std  ...      50%      75%       max
Price      1436.0  10730.824513   3626.964585  ...   9900.0  11950.0   32500.0
Age        1436.0     55.973538     17.964211  ...     60.0     68.0      80.0
KM         1436.0  68594.873259  37140.890566  ...  63634.0  86916.0  243000.0
HP         1436.0    101.513928     14.747603  ...    110.0    110.0     192.0
MetColor   1436.0      0.708914      0.454421  ...      1.0      1.0       1.0
Automatic  1436.0      0.055710      0.229441  ...      0.0      0.0       1.0
CC         1436.0   1566.827994    187.182436  ...   1600.0   1600.0    2000.0
Weight     1436.0   1072.459610     52.641120  ...   1070.0   1085.0    1615.0

[8 rows x 8 columns]
```

1. Statistical Summary

- **Price**:
  - Mean price ≈ 10,731 with a standard deviation of 3,627.
  - Minimum = 4,350 and maximum = 32,500 → wide variation in car prices.
  - Median (50%) = 9,900, showing slightly right-skewed distribution (since mean > median).
- **Age**:
  - Average age ≈ 56 months (about 4.5 years).
  - Cars range from 1 month to 80 months old.
  - Majority (50–75%) fall between 60–68 months.
- **KM (Mileage)**:
  - Average ≈ 68,595 km, but max is 243,000 km → indicates some high-mileage outliers.
  - 50% of cars have < 63,634 km.
- **HP (Horsepower)**:
  - Mean ≈ 102 HP, with most cars clustered around 110 HP.
  - Range: 69 to 192 HP → mix of standard and higher-performance cars.
- **MetColor (Metallic Color)**:
  - Mean ≈ 0.71 → about 71% of cars have metallic paint.
- **Automatic (Transmission)**:
  - Mean ≈ 0.056 → only ~6% of cars are automatic, rest is manual.

- **CC (Engine Capacity)**:
  - Average engine capacity ≈ 1567 cc.
  - Range: 1,200 to 2,000 cc → indicates mostly small-to-mid size cars.
- **Weight**:
  - Average ≈ 1,072 kg, ranging from 1,010 to 1,615 kg.

Interpretation:
The dataset contains mostly mid-range cars, with prices influenced by age, mileage, horsepower, and whether they have metallic colour or automatic transmission. There are some outliers in mileage and price that may affect modelling.

---

2. Missing Values

- All variables show **0 missing values**.
- This means data cleaning for missing values is not required.
- Dataset is complete and ready for further EDA.

Interpretation:
No imputation needed. The dataset quality is good in terms of completeness.

---

3. Structure of Data

- Total observations: **1,436 entries**.
- Total variables: **10 columns**.
- Data types:
  - **Numeric (int64, float64)** → Price, Age, KM, HP, MetColor, Automatic, CC, Weight.

- **Categorical (object)** → Fuel Type, Doors.

Interpretation:
The dataset contains both **numeric and categorical features**, suitable for regression/classification models. Numeric variables can be directly used for correlation analysis, while categorical variables (Fuel Type, Doors) may need encoding before modelling.

Frequency & Probability Tables:

```
# ================================================================================
# Frequency & Probability Tables
# ================================================================================
print("\n--- FuelType Frequency ---")
print(pd.crosstab(toyota["FuelType"], "count"))

print("\n--- FuelType vs Automatic ---")
print(pd.crosstab(toyota["Automatic"], toyota["FuelType"], margins=True))

print("\n--- Joint Probability ---")
print(pd.crosstab(toyota["Automatic"], toyota["FuelType"], normalize=True))

print("\n--- Conditional Probability (Row-wise) ---")
print(pd.crosstab(toyota["Automatic"], toyota["FuelType"], normalize="index"))
```

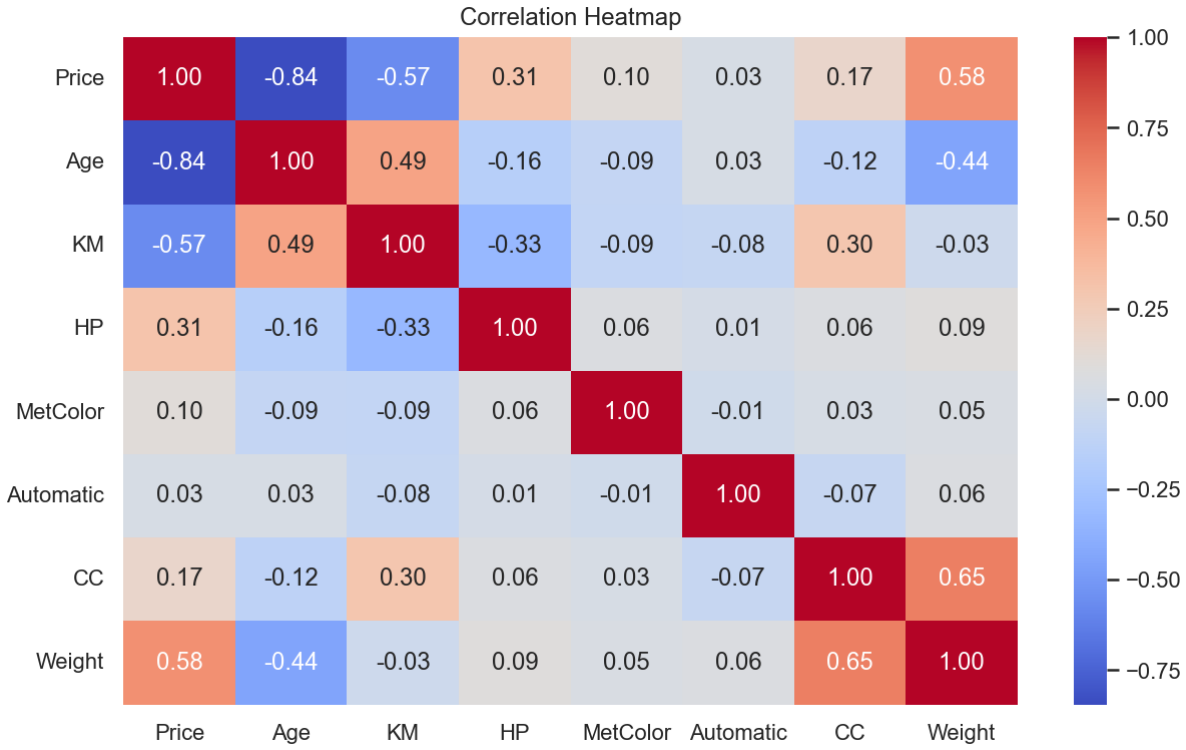
```
--- FuelType Frequency ---
col_0     count
FuelType
CNG          15
Diesel      144
Petrol     1277

--- FuelType vs Automatic ---
FuelType   CNG  Diesel  Petrol   All
Automatic
0           15     144    1197  1356
1            0       0      80    80
All         15     144    1277  1436

--- Joint Probability ---
FuelType        CNG    Diesel    Petrol
Automatic
0          0.010446  0.100279  0.833565
1          0.000000  0.000000  0.055710

--- Conditional Probability (Row-wise) ---
FuelType        CNG    Diesel    Petrol
Automatic
0          0.011062  0.106195  0.882743
1          0.000000  0.000000  1.000000
```

Interpretation:

Frequency Tables – Interpretation

1. **Fuel Type Distribution**
    - Petrol: 1,277 cars (dominant fuel type).
    - Diesel: 144 cars (minor share).
    - CNG: 15 cars (least common).
2. **Fuel Type vs Automatic**
    - Manual cars (Automatic = 0): 1,356 cars → Petrol (1,197), Diesel (144), CNG (15).
    - Automatic cars (Automatic = 1): 80 cars → all Petrol, no Diesel/CNG.
    - Automatic share = **5.6%**, Manual share = **94.4%**.
3. **Joint Probability**
    - Petrol manual: 0.8336 (highest likelihood).
    - Diesel manual: 0.1003.
    - CNG manual: 0.0104.
    - Petrol automatic: 0.0557.
    - Diesel and CNG automatics: 0.0000 (not present).
4. **Conditional Probability (Row-wise)**
    - Manual cars: 88.3% Petrol, 10.6% Diesel, 1.1% CNG.
    - Automatic cars: 100% Petrol, 0% Diesel, 0% CNG.

Correlation Heatmap:



Interpretation:

Price Relationships

- Age (-0.84): Strong negative correlation – older cars have much lower prices.
- KM (-0.57): Moderate negative correlation – more kilometres driven reduces price.
- Weight (0.58): Moderate positive correlation – heavier cars tend to be priced higher.
- HP (0.31): Weak positive correlation – higher horsepower slightly increases price.
- CC (0.17): Very weak positive effect on price.
- MetColor (0.10), Automatic (0.03): Almost no effect on price.

Age Relationships

- KM (0.49): Older cars generally have higher kilometers.
- Weight (-0.44): Older cars tend to weigh less (possibly newer models are heavier).
- HP (-0.16), CC (-0.12): Slightly lower in older cars.

Technical Feature Relationships

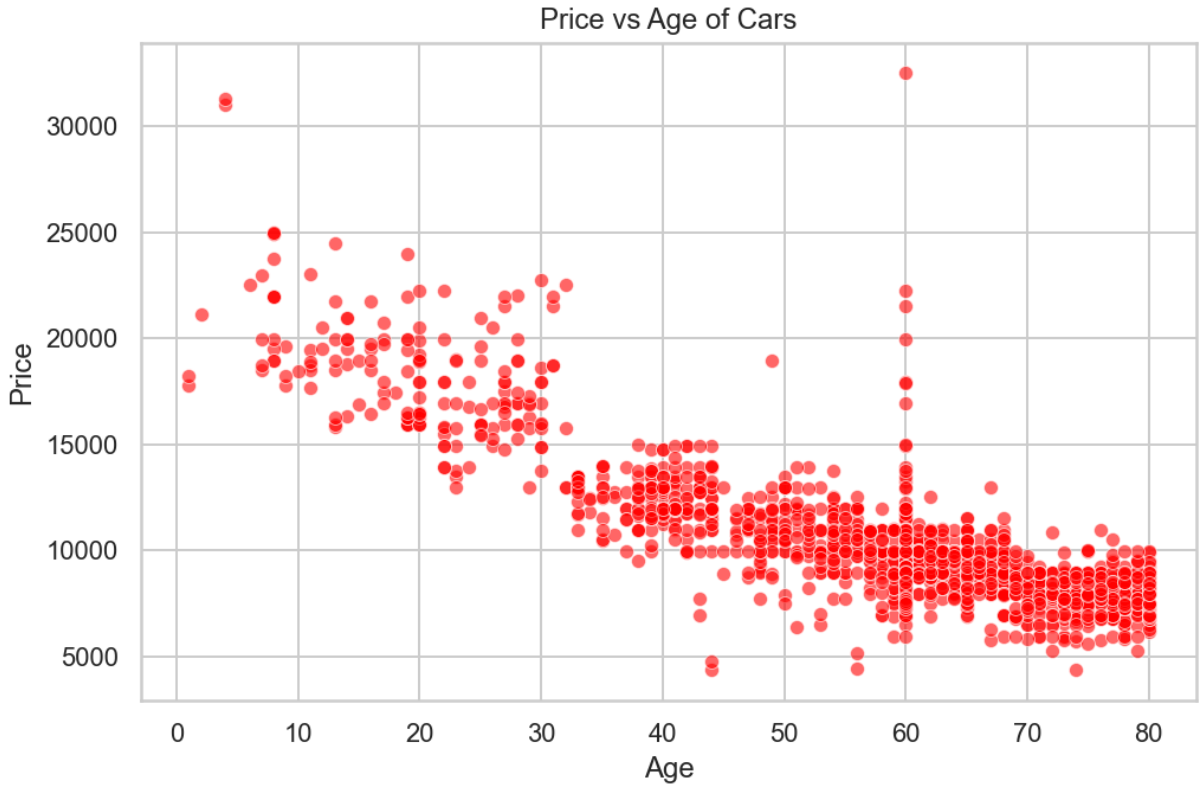
- CC & Weight (0.65): Strong positive correlation – bigger engine capacity means heavier cars.
- KM & HP (-0.33): Cars with higher kilometres often have lower horsepower.
- KM & CC (0.30): Cars with higher kilometres tend to have slightly larger engines.

Automatic & MetColor

- Very weak correlations across all variables → they don't significantly influence car price or other features.

Visualizations:

Price vs Age (Scatter)

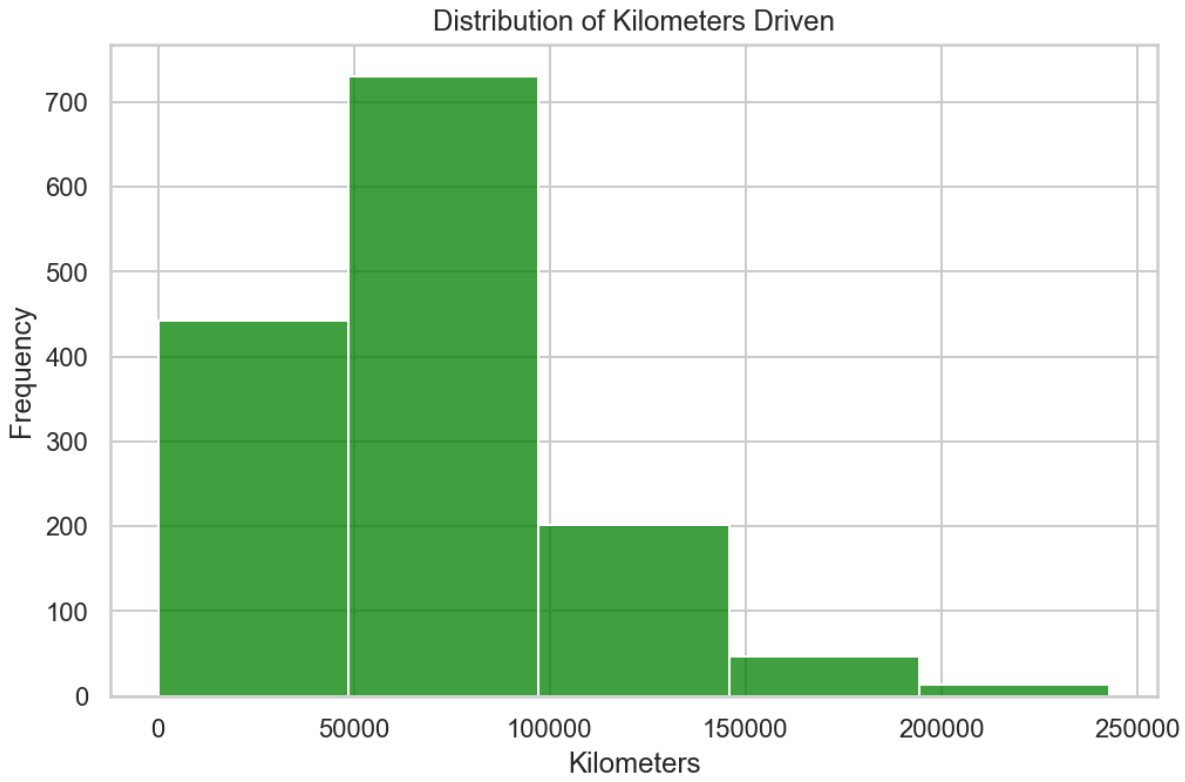


Interpretation:

1. **Strong Negative Trend** – As car age increases, the price consistently decreases.
2. **New Cars (0–10 years)** – Cars in this range are priced much higher, many above ₹20,000.
3. **Middle Age Cars (20–40 years)** – Prices show a steady decline, most between ₹10,000–15,000.
4. **Older Cars (40–60 years)** – Prices cluster in the lower range (₹5,000–12,000).
5. **Very Old Cars (60–80 years)** – Majority of prices stabilize below ₹10,000, showing minimal resale value.
6. **Outliers** – A few very high-priced points exist even for old cars (around age 60 and 70) → possibly rare/classic models.

7. **Price Drop is Steep Initially** – The biggest depreciation happens in the first 20 years.
8. **Saturation Effect** – After ~60 years, the price no longer decreases much; it stabilizes at a low level.
9. **Buyer's View** – Younger cars give better resale value, while very old cars retain little market value unless collectible.
10. **Seller's View** – Selling early (before 20 years) yields the highest return.

Histogram for KM

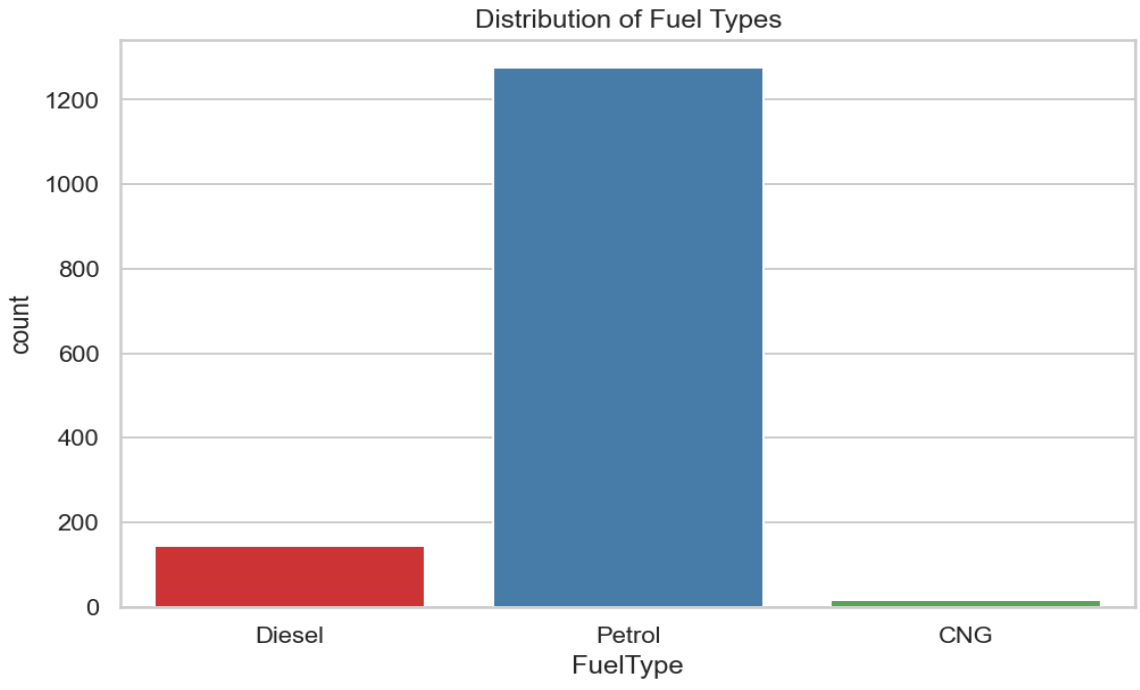


Interpretation:

- **Most cars have low to moderate usage** – Majority fall below 1,00,000 km.
- **Peak frequency** – The 50,000–1,00,000 km range has the highest number of cars (~730).

- **Next common group** – 0–50,000 km range, around 450 cars → relatively new/less-used vehicles.
- **Moderate usage cars** – 1,00,000–1,50,000 km range drops significantly (~200 cars).
- **High usage cars** – 1,50,000–2,00,000 km range has very few cars (~50).
- **Very high mileage cars** – Beyond 2,00,000 km, almost negligible count (<20).
- **Right-skewed distribution** – Most cars are clustered at lower km driven; tail stretches towards higher mileage.
- **Consumer perspective** – Buyers mostly encounter cars below 1,00,000 km, which are preferred in second-hand markets.
- **Seller perspective** – Cars with >1,50,000 km has very low resale demand.
- **Overall insight** – The dataset suggests that the second-hand market is dominated by low-to-moderate mileage cars, while very high mileage cars are rare.

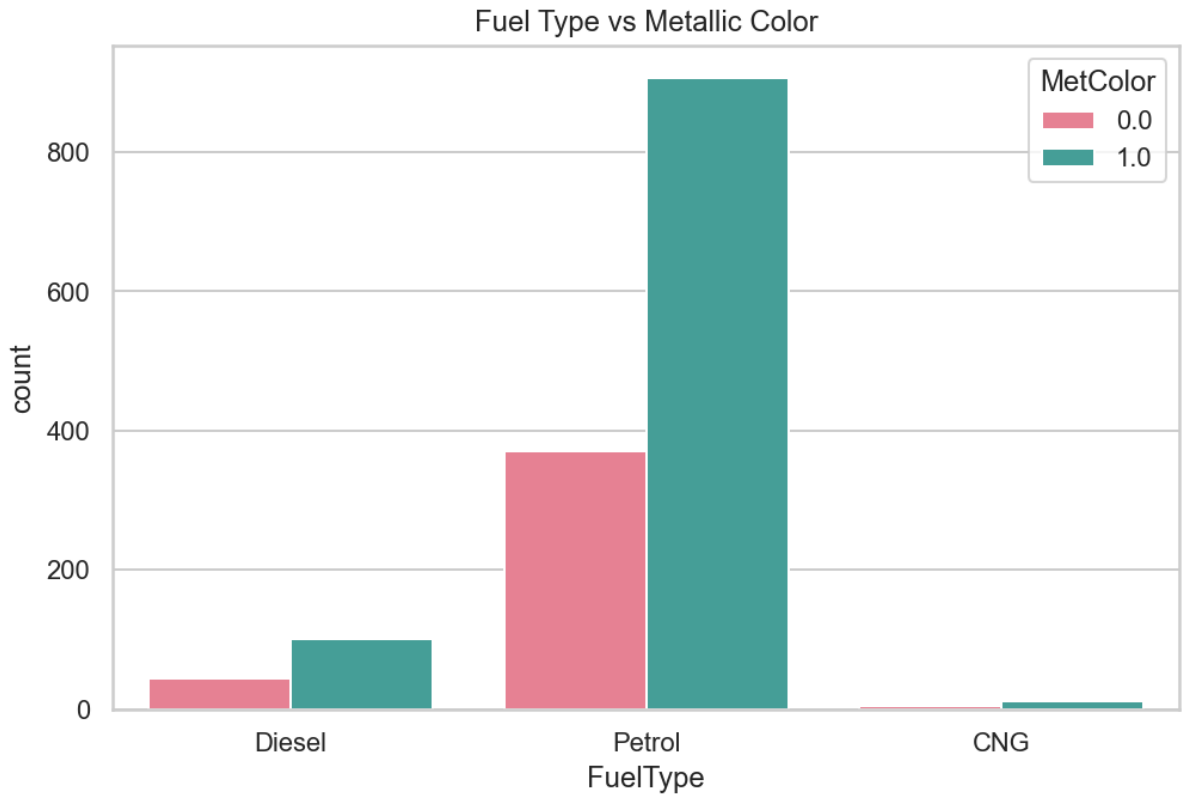
Bar plot – Fuel Type



Interpretation:

- Petrol dominates – Petrol cars are the most common, with count exceeding 1200.
- Diesel is limited – Diesel cars are far fewer, around 150 in number.
- CNG is rare – Only about 15 cars use CNG, making it the least preferred option.
- Clear preference – The second-hand car market is heavily petrol-oriented.
- Market imbalance – Petrol cars account for almost 90%+ of the dataset.
- Diesel shares declining – The low count indicates lesser availability or demand for diesel vehicles.
- CNG adoption negligible – Very few cars run on CNG, showing it's not widely popular in this market.
- Resale perspective – Buyers looking for petrol cars have maximum options, while diesel/CNG buyers have limited choice.
- Policy influence – The reduced diesel numbers may reflect stricter emission norms or fuel economy trends.
- Overall insight – Petrol cars dominate resale markets, while diesel and CNG play only a marginal role.

Grouped bar plot - Fuel Type vs Metallic Colour

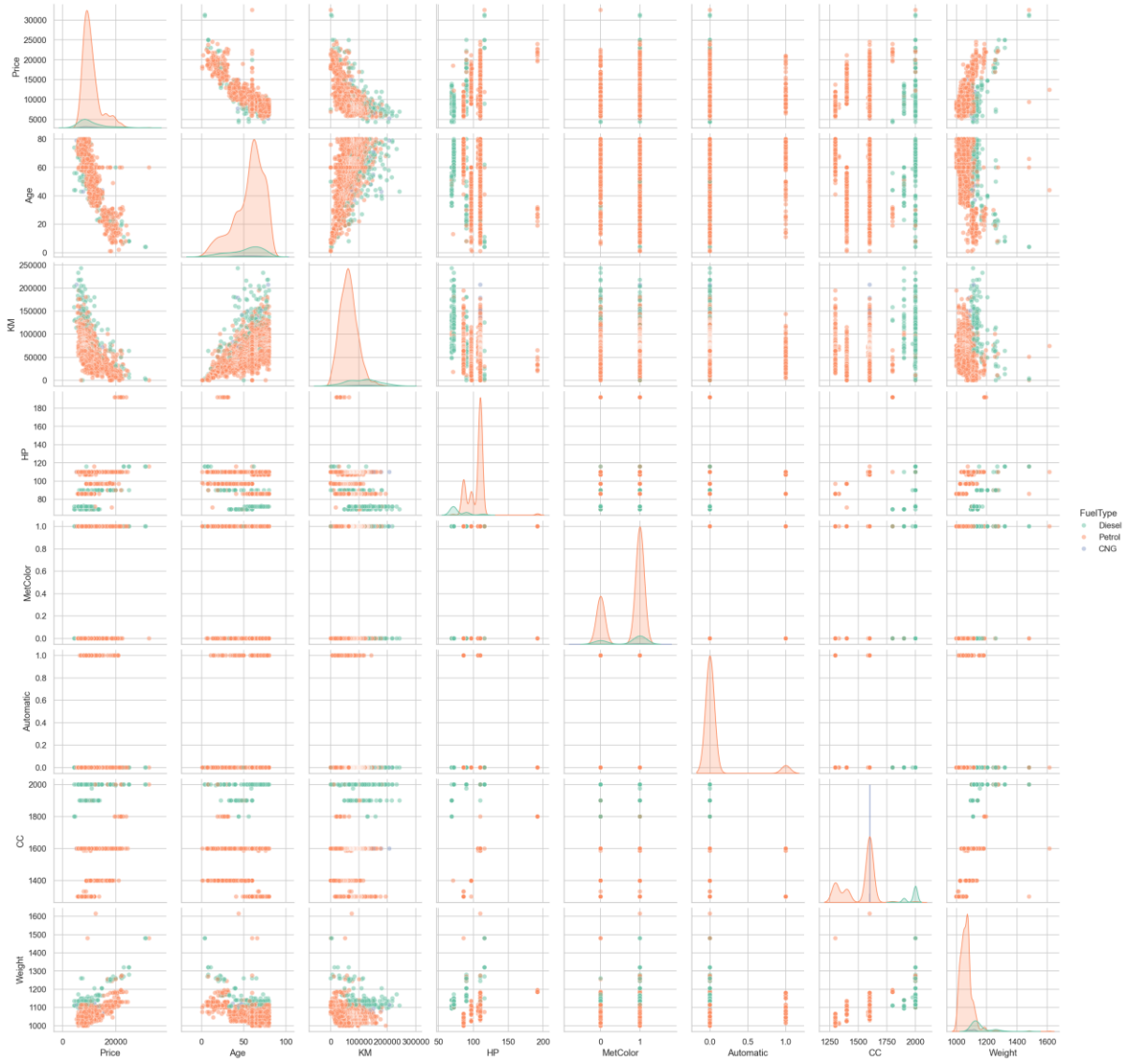


Interpretation:

- **Petrol dominates** – Most cars are petrol, both in metallic and non-metallic colour.
- **Metallic petrol cars are the highest** – More than 850 petrol cars have metallic colour.
- **Non-metallic petrol cars** – Around 370 petrol cars are non-metallic.
- **Diesel cars are fewer** – Diesel cars are much less compared to petrol.
- **Metallic diesel cars lead** – About 100 diesel cars have metallic colour, while only around 40 are non-metallic.
- **CNG cars are very rare** – Both metallic and non-metallic CNG cars are almost negligible.
- **Metallic preference overall** – Across all fuel types, metallic colour cars dominate.

- **Buyer preference trend** – Buyers seem to prefer metallic finish cars irrespective of fuel type.
- **Petrol + Metallic = most common combo** – This category dominates the dataset by a large margin.
- **CNG + Non-metallic = least common combo** – Almost no CNG cars with non-metallic colour exist.

Pair plot



Interpretation:

Price and Other Variables

- **Price vs. Age**: There's a strong negative correlation between price and age. As the age of the car increases, the price generally decreases. This is a common and expected relationship, as cars lose value over time.
- **Price vs. KM (Kilometres driven)**: A strong negative correlation exists between price and kilometres driven. The more kilometres a car has on the odometer, the lower its price tends to be.
- **Price vs. HP (Horsepower)**: A positive correlation can be observed. Cars with higher horsepower generally have a higher price.
- **Price vs. CC (Cubic Centimetres)**: A positive correlation is present. As the engine size (CC) increases, the price also tends to rise.
- **Price vs. Weight**: A positive correlation is seen here as well. Heavier cars are generally more expensive.

Other Notable Relationships

- **KM vs. Age**: There's a clear positive correlation. Older cars tend to have more kilometres on them, which makes sense as they have been driven for a longer time.
- **Fuel Type and Other Variables**: The scatter plots are color-coded by **Fuel Type**, which appears to have two categories: 'Petrol' (green) and 'Other' (orange).
- **Price and Fuel Type**: It looks like cars with 'Other' fuel type (likely diesel or another fuel) might have a slightly different price distribution compared to petrol cars, but a more detailed analysis would be needed to confirm this.

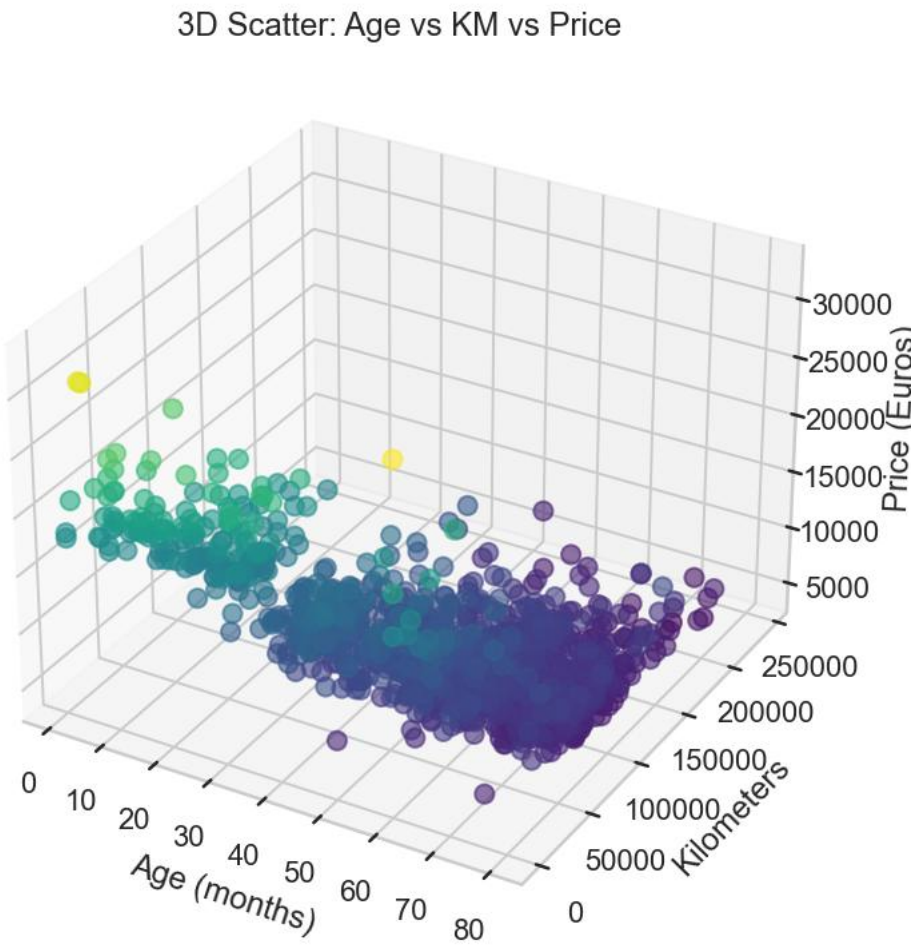
- **HP and Fuel Type**: There might be a difference in horsepower distribution between the two fuel types, with some of the higher HP values belonging to the 'Other' fuel type.

Categorical Variables (Automatic and MetColor):

- The Automatic variable is binary (0 or 1), indicating if the car is automatic. The scatter plots with Automatic as one of the axes show distinct horizontal or vertical lines of data points, as expected for a categorical variable.
- MetColor is also a binary variable, likely indicating whether the car has a metallic paint colour. Like Automatic, the data points appear as distinct lines

In summary, the pair plot effectively visualizes the relationships between various car features. It highlights expected trends like the depreciation of a car's value over time and with increased mileage, as well as the positive relationship between a car's price and its performance-related features like horsepower and engine size.

3D Scatter: Age, KM, Price:



Interpretation:

The image is a 3D scatter plot showing the relationship between a car's Age, Kilometres driven (KM), and Price.

Axes:

- The x-axis represents Age in months.
- The y-axis represents Kilometres driven.
- The z-axis represents Price in Euros.

Overall Trend:

The plot visualizes the concept of car depreciation. As a car's age and kilometres increase, its price generally decreases. You can see the bulk of the data points form a surface that slopes downward

from the top-left (low age, low kilometres, high price) to the bottom-right (high age, high kilometres, low price).

Key Observations:

- High Price Cars: The cars with the highest prices (towards the top of the z-axis) are clustered in the area with low age and low kilometres. This makes sense, as newer cars with less mileage are more expensive.
- Low Price Cars: The cars with the lowest prices are found where age and kilometres are high.

Outliers:

- There are a few data points that stand out, such as the two yellow points. These could represent special cases:
- An expensive car with high age and/or high kilometres (perhaps a classic or collector's car).

In summary, the 3D scatter plot effectively demonstrates the inverse relationship between a car's value and its age and mileage, with price acting as a function of these two variables.

Conclusion:

The analysis of frequency distributions and visualizations highlights clear market preferences in the automobile dataset. Petrol cars dominate the market, with metallic colour being the most popular choice across all fuel types. Diesel cars, though fewer in number, also show a stronger inclination toward metallic finishes, while CNG cars remain the least preferred option regardless of colour. Overall, the findings suggest that **metallic finish and petrol fuel type form the most demanded combination**, reflecting consumer trends toward aesthetics and convenience. This trend emphasizes the importance of petrol-based, metallic-coloured vehicles in shaping market demand, while alternative fuels like CNG show limited adoption.