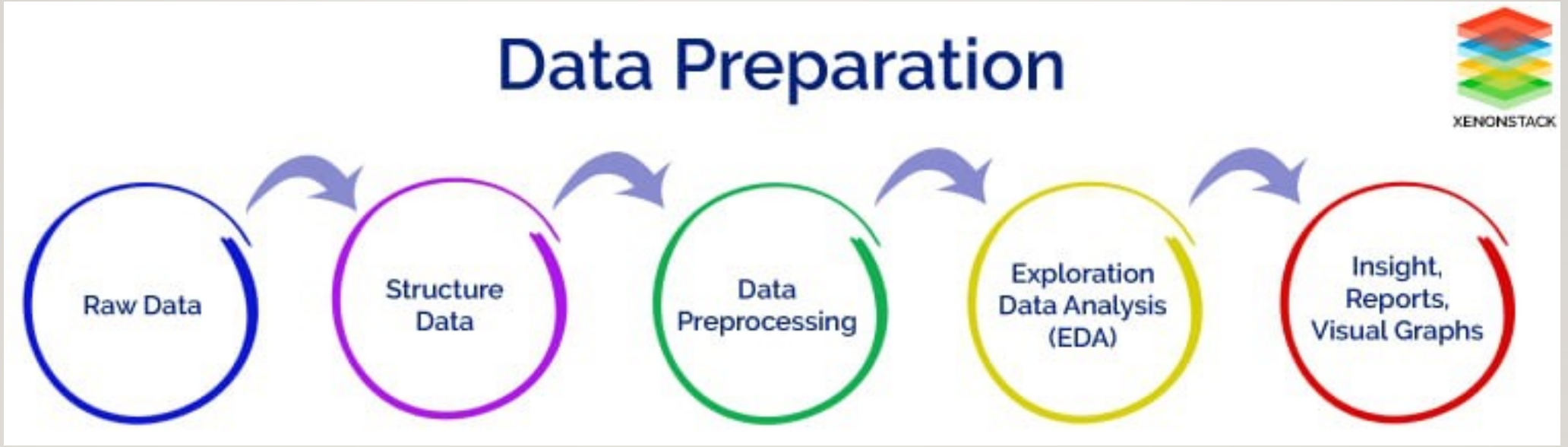
EXPLORATORY DATA ANALYSIS (EDA) USING PYTHON

JOKO TRILOKA, PH.D

INTRODUCTION TO EDA

The main objective of this course is to cover the steps involved in Data pre-processing, Feature Engineering, and different stages of Exploratory Data Analysis, which is an essential step in any research analysis. Data pre-processing, Feature Engineering, and EDA are fundamental early steps after data collection. Still, they are not limited to where the data is simply visualized, plotted, and manipulated, without any assumptions, to assess the quality of the data and building models. This course will guide you through data pre-processing, feature engineering, and EDA using Python.

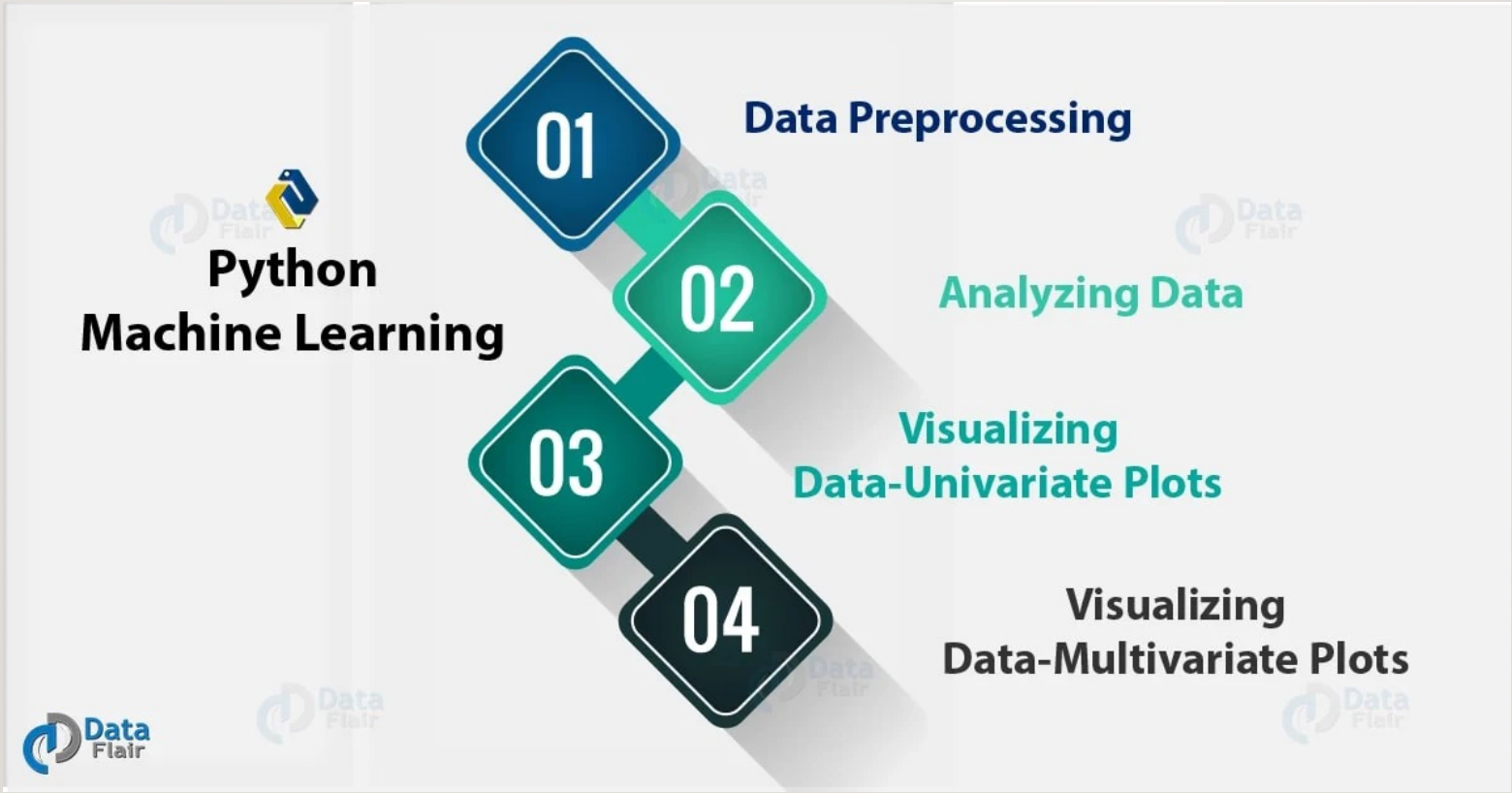
Data Preparation



WHAT IS DATA PRE-PROCESSING AND FEATURE ENGINEERING?

- In our data-driven processes, we consider refining our raw data. Both data pre-processing and feature engineering play pivotal roles in this endeavour. Data pre-processing encompasses a range of activities, including data integration, analysis, cleaning, transformation, and dimension reduction.
- Data pre-processing involves cleaning and preparing raw data to facilitate feature engineering. Meanwhile, feature engineering entails employing various techniques to manipulate the data. This may include adding or removing relevant features, handling missing data, encoding variables, and dealing with categorical variables, among other tasks.
- Undoubtedly, feature engineering is a critical task that significantly influences the outcome of a model. It involves crafting new features based on existing data while pre-processing primarily focuses on cleaning and organizing the data.

WHAT IS DATA PRE-PROCESSING AND FEATURE ENGINEERING?



STEP 1: IMPORT PYTHON LIBRARIES

- The first step involved in ML using python is understanding and playing around with our data using libraries. Here is the link to the dataset.
- Import all libraries which are required for our analysis, such as Data Loading, Statistical analysis, Visualizations, Data Transformations, Merge and Joins, etc.
- **Pandas and Numpy have been used for Data Manipulation and numerical Calculations**
- **Matplotlib and Seaborn have been used for Data visualizations.**

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
#to ignore warnings
import warnings
warnings.filterwarnings('ignore')
```

STEP 2: READING DATASET

- The Pandas library offers a wide range of possibilities for loading data into the pandas DataFrame from files like JSON, .csv, .xlsx, .sql, .pickle, .html, .txt, images etc.
- Most of the data are available in a tabular format of CSV files. It is trendy and easy to access. Using the **read_csv()** function, data can be converted to a pandas DataFrame.
- In this article, the data to predict **Used car price** is being used as an example. In this dataset, we are trying to analyze the used car's price and how EDA focuses on identifying the factors influencing the car price. We have stored the data in the DataFrame **data.**

```
data = pd.read_csv("used_cars.csv")
```

ANALYZING THE DATA

- Before we make any inferences, we listen to our data by examining all variables in the data.
- The main goal of data understanding is to gain general insights about the data, which covers the number of rows and columns, values in the data, datatypes, and Missing values in the dataset.
- **shape** – **shape** will display the number of observations(rows) and features(columns) in the dataset
- There are 7253 observations and 14 variables in our dataset
- **head()** will display the top 5 observations of the dataset

HEAD() WILL DISPLAY THE TOP 5 OBSERVATIONS OF THE DATASET

```
data.head()
```

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_price	Price
0	0	Maruti Wagon R LXI CNG	Mumbai	2010	72000	CNG	Manual	First	26.60	998.0	58.16	5.0	NaN	1.75
1	1	Hyundai Creta 1.6 CRDi SX Option	Pune	2015	41000	Diesel	Manual	First	19.67	1582.0	126.20	5.0	NaN	12.50
2	2	Honda Jazz V	Chennai	2011	46000	Petrol	Manual	First	18.20	1199.0	88.70	5.0	8.61	4.50
3	3	Maruti Ertiga VDI	Chennai	2012	87000	Diesel	Manual	First	20.77	1248.0	88.76	7.0	NaN	6.00
4	4	Audi A4 New 2.0 TDI Multitronic	Coimbatore	2013	40670	Diesel	Automatic	Second	15.20	1968.0	140.80	5.0	NaN	17.74

TAIL() WILL DISPLAY THE LAST 5 OBSERVATIONS OF THE DATASET

- data.tail()

:

	S.No.	Name	Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_price	Price
7248	7248	Volkswagen Vento Diesel Trendline	Hyderabad	2011	89411	Diesel	Manual	First	20.54	1598.0	103.6	5.0	NaN	NaN
7249	7249	Volkswagen Polo GT TSI	Mumbai	2015	59000	Petrol	Automatic	First	17.21	1197.0	103.6	5.0	NaN	NaN
7250	7250	Nissan Micra Diesel XV	Kolkata	2012	28000	Diesel	Manual	First	23.08	1461.0	63.1	5.0	NaN	NaN
7251	7251	Volkswagen Polo GT TSI	Pune	2013	52262	Petrol	Automatic	Third	17.20	1197.0	103.6	5.0	NaN	NaN
7252	7252	Mercedes-Benz E-Class 2009-2013 E 220 CDI Avan...	Kochi	2014	72443	Diesel	Automatic	First	10.00	2148.0	170.0	5.0	NaN	NaN

INFO() HELPS TO UNDERSTAND THE DATA TYPE AND INFORMATION ABOUT DATA, INCLUDING THE NUMBER OF RECORDS IN EACH COLUMN, DATA HAVING NULL OR NOT NULL, DATA TYPE, THE MEMORY USAGE OF THE DATASET

- data.info()

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 7253 entries, 0 to 7252
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   S.No.              7253 non-null   int64
 1   Name               7253 non-null   object
 2   Location           7253 non-null   object
 3   Year               7253 non-null   int64
 4   Kilometers_Driven  7253 non-null   int64
 5   Fuel_Type          7253 non-null   object
 6   Transmission       7253 non-null   object
 7   Owner_Type         7253 non-null   object
 8   Mileage            7251 non-null   float64
 9   Engine             7207 non-null   float64
 10  Power              7078 non-null   float64
 11  Seats              7200 non-null   float64
 12  New_price          1006 non-null   float64
 13  Price              6019 non-null   float64
dtypes: float64(6), int64(3), object(5)
memory usage: 793.4+ KB
```

data.info() shows the variables Mileage, Engine, Power, Seats, New_Price, and Price have missing values. Numeric variables like Mileage, Power are of datatype as float64 and int64. Categorical variables like Location, Fuel_Type, Transmission, and Owner Type are of object data type

MISSING VALUES CALCULATION

- **isnull()** is widely been in all pre-processing steps to identify null values in the data
- In our example, **data.isnull().sum()** is used to get the number of missing records in each column
- The below code helps to calculate the percentage of missing values in each column
- (data.isnull().sum()/(len(data)))*100

STEP 3: DATA REDUCTION

- Some columns or variables can be dropped if they do not add value to our analysis.
- In our dataset, the column S.No have only ID values, assuming they don't have any predictive power to predict the dependent variable.
- # Remove S.No. column from data data = data.drop(['S.No.'], axis = 1) data.info()

STEP 4: FEATURE ENGINEERING

- Feature engineering refers to the process of using domain knowledge to select and transform the most relevant variables from raw data when creating a predictive model using machine learning or statistical modeling. The main goal of Feature engineering is to create meaningful data from raw data.

STEP 5: CREATING FEATURES

- We will play around with the variables Year and Name in our dataset. If we see the sample data, the column “Year” shows the manufacturing year of the car.
- **It would be difficult to find the car’s age if it is in year format as the Age of the car is a contributing factor to Car Price.**
- Introducing a new column, “Car_Age” to know the age of the car
- from datetime import date date.today().year data['Car_Age']=date.today().year-data['Year'] data.head()

- Since car names will not be great predictors of the price in our current data. But we can process this column to extract important information using brand and Model names. **Let's split the name and introduce new variables "Brand" and "Model"**

- ```
  data['Brand'] = data.Name.str.split().str.get(0)data['Model'] =
  data.Name.str.split().str.get(1) +
  data.Name.str.split().str.get(2)data[['Name','Brand','Model']]
  ```
  ```

STEP 6: DATA CLEANING/WRANGLING

- Some names of the variables are not relevant and not easy to understand. Some data may have data entry errors, and some variables may need data type conversion. We need to fix this issue in the data.
- In the example, **The brand name 'Isuzu' 'ISUZU' and 'Mini' and 'Land' looks incorrect. This needs to be corrected**
- print(data.Brand.unique()) print(data.Brand.nunique())

```
['Maruti' 'Hyundai' 'Honda' 'Audi' 'Nissan' 'Toyota' 'Volkswagen' 'Tata'
 'Land' 'Mitsubishi' 'Renault' 'Mercedes-Benz' 'BMW' 'Mahindra' 'Ford'
 'Porsche' 'Datsun' 'Jaguar' 'Volvo' 'Chevrolet' 'Skoda' 'Mini' 'Fiat'
 'Jeep' 'Smart' 'Ambassador' 'Isuzu' 'ISUZU' 'Force' 'Bentley'
 'Lamborghini' 'Hindustan' 'OpelCorsa']
33
```

- searchfor = ['Isuzu' ,'ISUZU','Mini','Land']
  data[data.Brand.str.contains('|'.join(searchfor))].head(5)

Location	Year	Kilometers_Driven	Fuel_Type	Transmission	Owner_Type	Mileage	Engine	Power	Seats	New_price	Price	Car_Age	Brand	Model
Delhi	2014	72000	Diesel	Automatic	First	12.70	2179.0	187.70	5.0	NaN	27.00	8	Land	RoverRange
Pune	2012	85000	Diesel	Automatic	Second	0.00	2179.0	115.00	5.0	NaN	17.50	10	Land	RoverFreelander
Jaipur	2017	8525	Diesel	Automatic	Second	16.60	1998.0	112.00	5.0	NaN	23.00	5	Mini	CountrymanCooper
oimbatore	2018	36091	Diesel	Automatic	First	12.70	2179.0	187.70	5.0	NaN	55.76	4	Land	RoverRange
Kochi	2017	26327	Petrol	Automatic	First	16.82	1998.0	189.08	4.0	44.28	35.67	5	Mini	CooperConvertible

```
data["Brand"].replace({"ISUZU": "Isuzu", "Mini": "Mini Cooper","Land":"Land Rover"}, inplace=True)
```

STEP 7: **EDA EXPLORATORY DATA ANALYSIS**

- Exploratory Data Analysis refers to the crucial process of performing initial investigations on data to discover patterns to check assumptions with the help of summary statistics and graphical representations.
- EDA can be leveraged to check for outliers, patterns, and trends in the given data.
- EDA helps to find meaningful patterns in data.
- EDA provides in-depth insights into the data sets to solve our business problems.
- EDA gives a clue to impute missing values in the dataset

STEP 8: STATISTICS SUMMARY

- The information gives a quick and simple description of the data.
- Can include Count, Mean, Standard Deviation, median, mode, minimum value, maximum value, range, standard deviation, etc.
- Statistics summary gives a high-level idea to identify whether the data has any outliers, data entry error, distribution of data such as the data is normally distributed or left/right skewed
- In python, this can be achieved using describe()
- describe() function gives all statistics summary of data
- **describe()**– Provide a statistics summary of data belonging to numerical datatype such as int, float

DATA.DESCRIBE().T

Out[12]:

	count	mean	std	min	25%	50%	75%	max
S.No.	7253.0	3626.000000	2093.905084	0.00	1813.000	3626.00	5439.0000	7252.00
Year	7253.0	2013.365366	3.254421	1996.00	2011.000	2014.00	2016.0000	2019.00
Kilometers_Driven	7253.0	58699.063146	84427.720583	171.00	34000.000	53416.00	73000.0000	6500000.00
Mileage	7251.0	18.141580	4.562197	0.00	15.170	18.16	21.1000	33.54
Engine	7207.0	1616.573470	595.285137	72.00	1198.000	1493.00	1968.0000	5998.00
Power	7078.0	112.765214	53.493553	34.20	75.000	94.00	138.1000	616.00
Seats	7200.0	5.280417	0.809277	2.00	5.000	5.00	5.0000	10.00
New_price	1006.0	22.779692	27.759344	3.91	7.885	11.57	26.0425	375.00
Price	6019.0	9.479468	11.187917	0.44	3.500	5.64	9.9500	160.00

From the statistics summary, we can infer the below findings :

- Years range from 1996- 2019 and has a high in a range which shows used cars contain both latest models and old model cars.
- On average of Kilometers-driven in Used cars are ~58k KM. The range shows a huge difference between min and max as max values show 650000 KM shows the evidence of an outlier. This record can be removed.
- Min value of Mileage shows 0 cars won't be sold with 0 mileage. This sounds like a data entry issue.
- It looks like Engine and Power have outliers, and the data is right-skewed.
- The average number of seats in a car is 5. car seat is an important feature in price contribution.
- The max price of a used car is 160k which is quite weird, such a high price for used cars. There may be an outlier or data entry issue.

DESCRIBE(INCLUDE='ALL') PROVIDES A STATISTICS SUMMARY OF ALL DATA, INCLUDE OBJECT, CATEGORY ETC

- `data.describe(include='all').T`

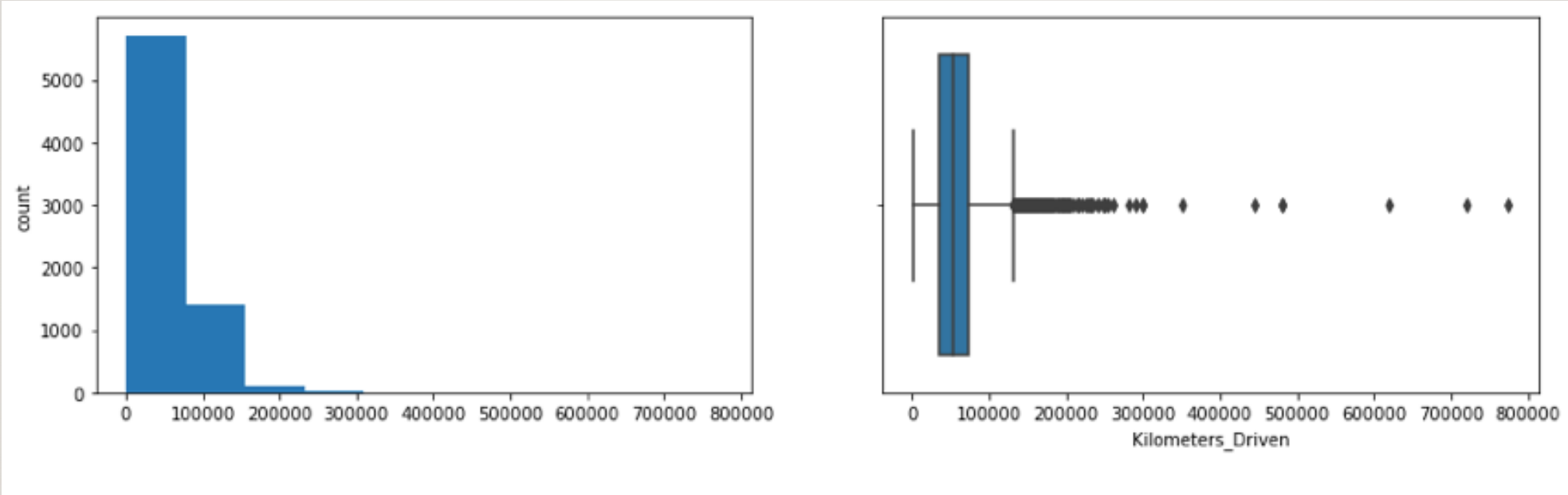
Out[13]:

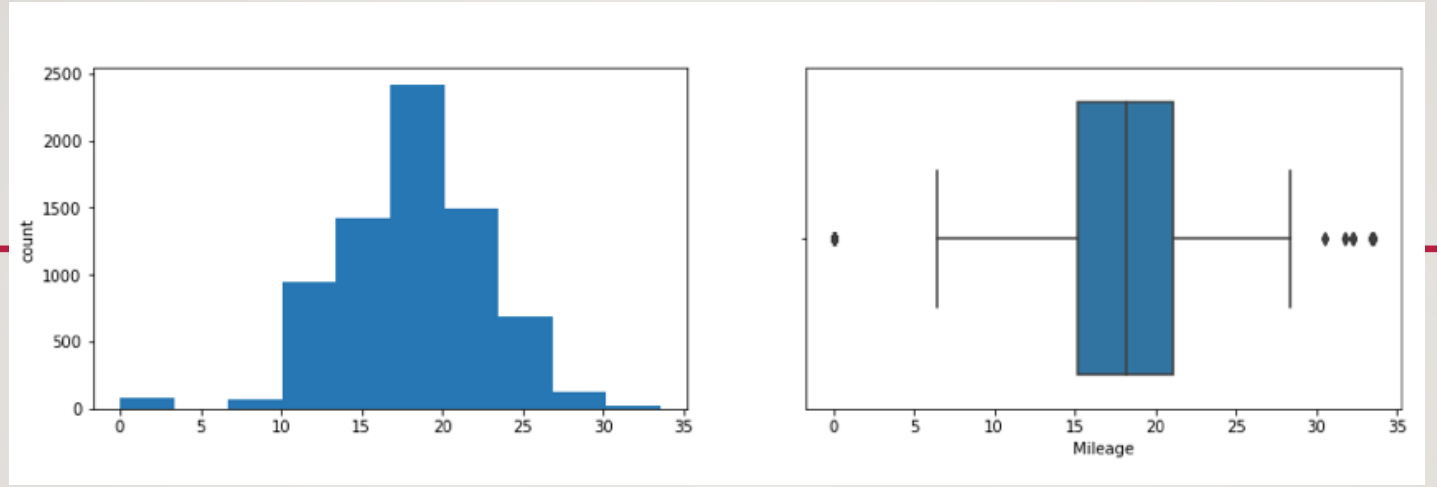
	count	unique	top	freq	mean	std	min	25%	50%	75%	max
S.No.	7253.0	NaN	NaN	NaN	3626.0	2093.905084	0.0	1813.0	3626.0	5439.0	7252.0
Name	7253	2041	Mahindra XUV500 W8 2WD	55	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Location	7253	11	Mumbai	949	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Year	7253.0	NaN	NaN	NaN	2013.365366	3.254421	1996.0	2011.0	2014.0	2016.0	2019.0
Kilometers_Driven	7253.0	NaN	NaN	NaN	58699.063146	84427.720583	171.0	34000.0	53416.0	73000.0	6500000.0
Fuel_Type	7253	5	Diesel	3852	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Transmission	7253	2	Manual	5204	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Owner_Type	7253	4	First	5952	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Mileage	7251.0	NaN	NaN	NaN	18.14158	4.562197	0.0	15.17	18.16	21.1	33.54
Engine	7207.0	NaN	NaN	NaN	1616.57347	595.285137	72.0	1198.0	1493.0	1968.0	5998.0
Power	7078.0	NaN	NaN	NaN	112.765214	53.493553	34.2	75.0	94.0	138.1	616.0
Seats	7200.0	NaN	NaN	NaN	5.280417	0.809277	2.0	5.0	5.0	5.0	10.0
New_price	1006.0	NaN	NaN	NaN	22.779692	27.759344	3.91	7.885	11.57	26.0425	375.0
Price	6019.0	NaN	NaN	NaN	9.479468	11.187917	0.44	3.5	5.64	9.95	160.0

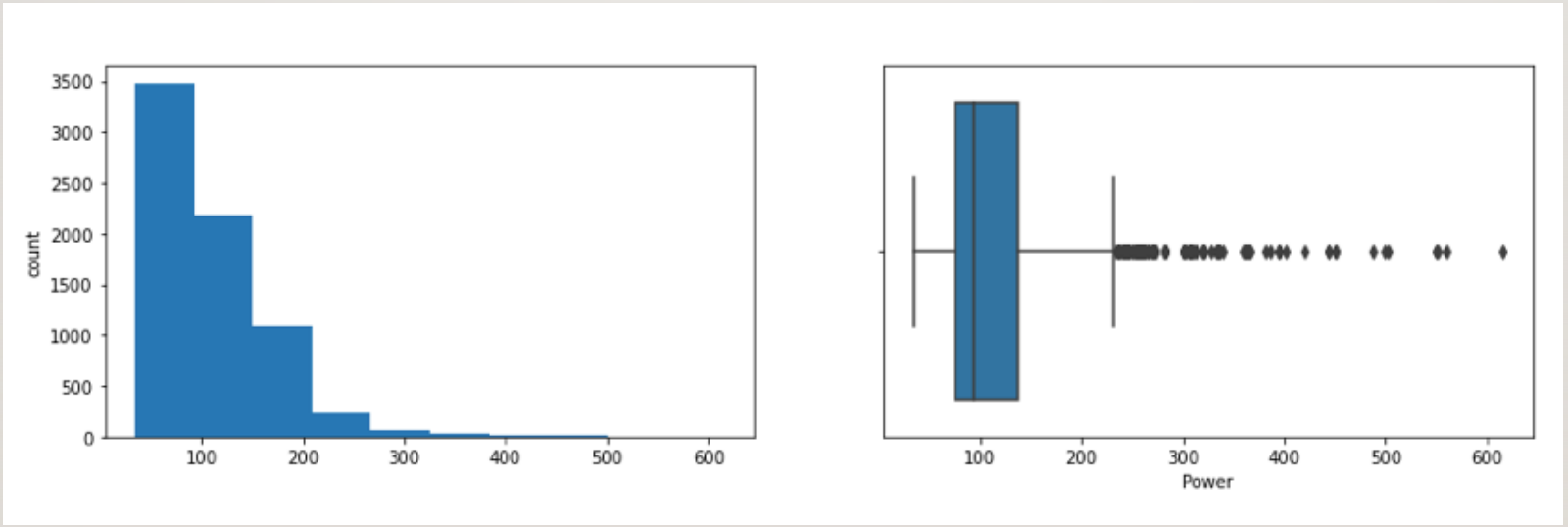
STEP 9: **EDA UNIVARIATE ANALYSIS**

- Analyzing/visualizing the dataset by taking one variable at a time:
- Data visualization is essential; we must decide what charts to plot to better understand the data. In this article, we visualize our data using Matplotlib and Seaborn libraries.
- Matplotlib is a Python 2D plotting library used to draw basic charts we use Matplotlib.
- Seaborn is also a python library built on top of Matplotlib that uses short lines of code to create and style statistical plots from Pandas and Numpy
- Univariate analysis can be done for both Categorical and Numerical variables.
- Categorical variables can be visualized using a Count plot, Bar Chart, Pie Plot, etc.
- Numerical Variables can be visualized using Histogram, Box Plot, Density Plot, etc.
- In our example, we have done a Univariate analysis using Histogram and Box Plot for continuous Variables.
- In the below fig, a histogram and box plot is used to show the pattern of the variables, as some variables have skewness and outliers.

```
FOR COL IN NUM_COLS:
PRINT(COL)
PRINT('SKEW :', ROUND(DATA[COL].SKEW(), 2))
PLT.FIGURE(FIGSIZE = (15, 4))
PLT.SUBPLOT(1, 2, 1)
DATA[COL].HIST(GRID=FALSE)
PLT.YLABEL('COUNT')
PLT.SUBPLOT(1, 2, 2) SNS.BOXPLOT(X=DATA[COL])
PLT.SHOW()
```



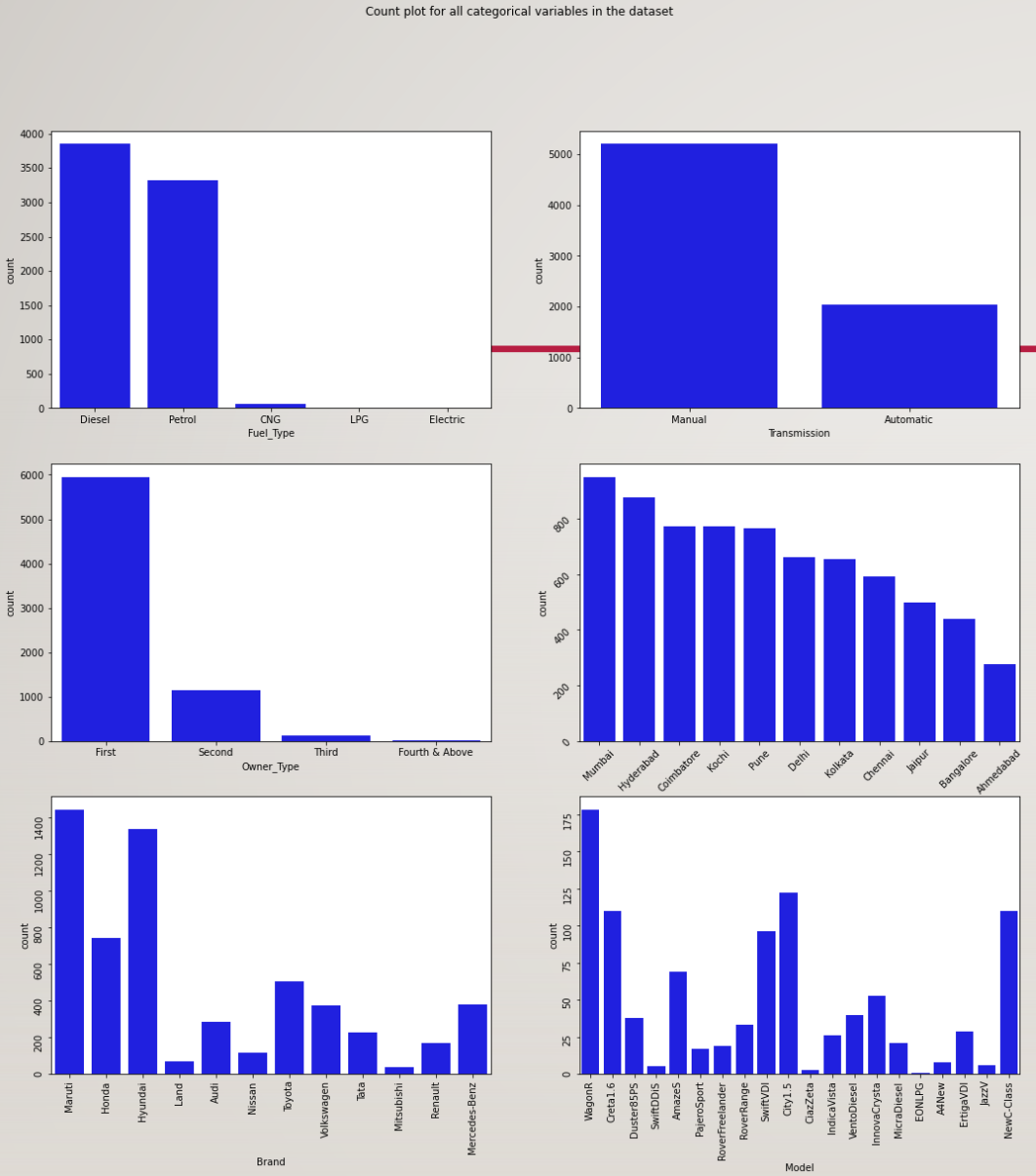

count
2500
2000
1500
1000
500
0
0
5
10
15
20
25
30
35
Mileage


count
3500
3000
2500
2000
1500
1000
500
0
100
200
300
400
500
600
Power

PRICE AND KILOMETERS DRIVEN ARE RIGHT SKEWED FOR THIS DATA TO BE TRANSFORMED, AND ALL OUTLIERS WILL BE HANDLED DURING IMPUTATION CATEGORICAL VARIABLES ARE BEING VISUALIZED USING A COUNT PLOT. CATEGORICAL VARIABLES PROVIDE THE PATTERN OF FACTORS INFLUENCING CAR PRICE

```
fig, axes = plt.subplots(3, 2, figsize = (18, 18))
fig.suptitle('Bar plot for all categorical variables in the dataset')
sns.countplot(ax = axes[0, 0], x = 'Fuel_Type', data = data, color = 'blue',
              order = data['Fuel_Type'].value_counts().index);
sns.countplot(ax = axes[0, 1], x = 'Transmission', data = data, color = 'blue',
              order = data['Transmission'].value_counts().index);
sns.countplot(ax = axes[1, 0], x = 'Owner_Type', data = data, color = 'blue',
              order = data['Owner_Type'].value_counts().index);
sns.countplot(ax = axes[1, 1], x = 'Location', data = data, color = 'blue',
              order = data['Location'].value_counts().index);
sns.countplot(ax = axes[2, 0], x = 'Brand', data = data, color = 'blue',
              order = data['Brand'].head(20).value_counts().index);
sns.countplot(ax = axes[2, 1], x = 'Model', data = data, color = 'blue',
              order = data['Model'].head(20).value_counts().index);
axes[1][1].tick_params(labelrotation=45);
axes[2][0].tick_params(labelrotation=90);
axes[2][1].tick_params(labelrotation=90);
```

Count plot for all categorical variables in the dataset



From the count plot, we can have below observations

- Mumbai has the highest number of cars available for purchase, followed by Hyderabad and Coimbatore
- ~53% of cars have fuel type as Diesel this shows diesel cars provide higher performance
- ~72% of cars have manual transmission
- ~82 % of cars are First owned cars. This shows most of the buyers prefer to purchase first-owner cars
- ~20% of cars belong to the brand Maruti followed by 19% of cars belonging to Hyundai
- WagonR ranks first among all models which are available for purchase

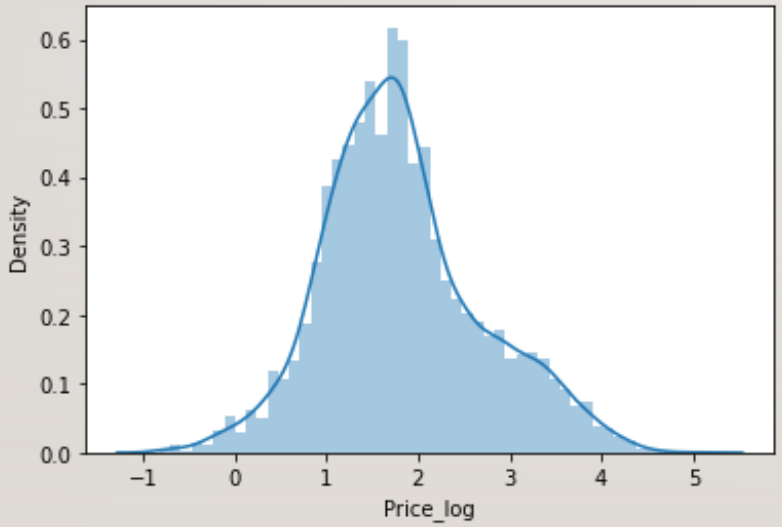
STEP 10: DATA TRANSFORMATION

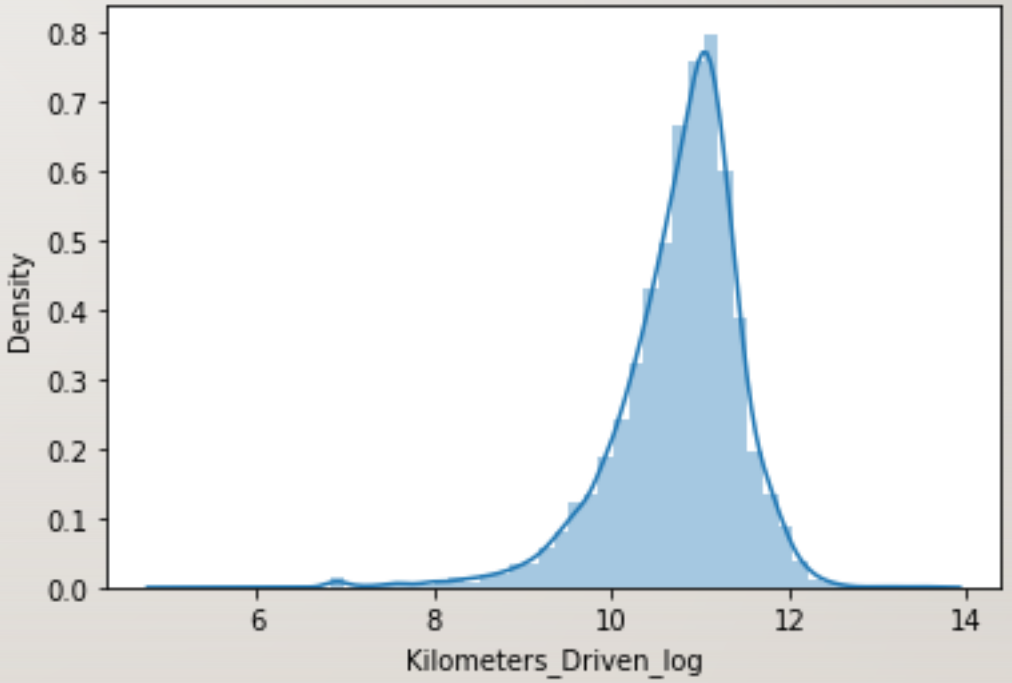
- Univariate analysis demonstrated the data pattern as some variables to be transformed.
- Price and Kilometer-Driven variables are highly skewed and on a larger scale. Let's do log transformation.
- Log transformation can help in normalization, so this variable can maintain standard scale with other variables:

```python
# Function for log transformation of the column
def log_transform(data,col):
    for colname in col:
        if (data[colname] == 1.0).all():
            data[colname + '_log'] = np.log(data[colname]+1)
        else:
            data[colname + '_log'] = np.log(data[colname])
    data.info()
```

```python
log_transform(data,['Kilometers_Driven','Price'])
```

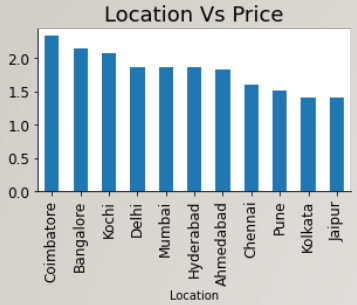
```python
#Log transformation of the feature 'Kilometers_Driven'
sns.distplot(data["Kilometers_Driven_log"], axlabel="Kilometers_Driven_log");
```

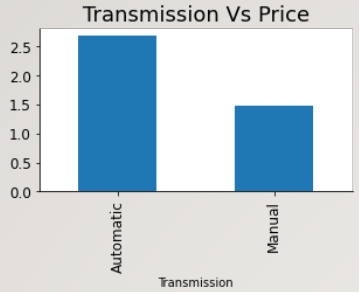


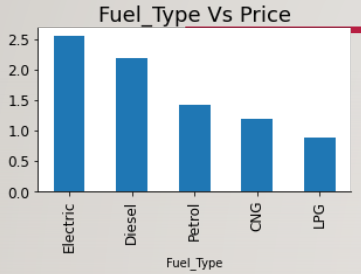


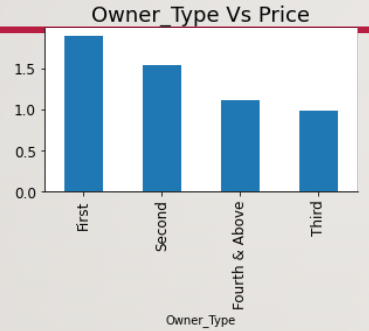
A BAR PLOT CAN BE USED TO SHOW THE RELATIONSHIP BETWEEN CATEGORICAL VARIABLES AND CONTINUOUS VARIABLES

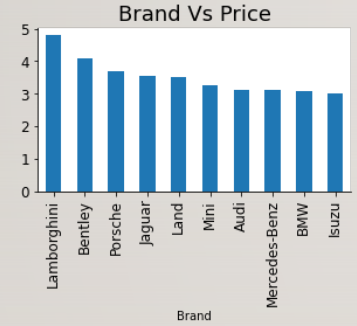
```
fig, axarr = plt.subplots(4, 2, figsize=(12, 18))
data.groupby('Location')['Price_log'].mean().sort_values(ascending=False).plot.bar(ax=a
axarr[0][0].set_title("Location Vs Price", fontsize=18)
data.groupby('Transmission')['Price_log'].mean().sort_values(ascending=False).plot.bar(
axarr[0][1].set_title("Transmission Vs Price", fontsize=18)
data.groupby('Fuel_Type')['Price_log'].mean().sort_values(ascending=False).plot.bar(ax=
axarr[1][0].set_title("Fuel_Type Vs Price", fontsize=18)
data.groupby('Owner_Type')['Price_log'].mean().sort_values(ascending=False).plot.bar(ax
axarr[1][1].set_title("Owner_Type Vs Price", fontsize=18)
data.groupby('Brand')['Price_log'].mean().sort_values(ascending=False).head(10).plot.ba
axarr[2][0].set_title("Brand Vs Price", fontsize=18)
data.groupby('Model')['Price_log'].mean().sort_values(ascending=False).head(10).plot.ba
axarr[2][1].set_title("Model Vs Price", fontsize=18)
data.groupby('Seats')['Price_log'].mean().sort_values(ascending=False).plot.bar(ax=axar
axarr[3][0].set_title("Seats Vs Price", fontsize=18)
data.groupby('Car_Age')['Price_log'].mean().sort_values(ascending=False).plot.bar(ax=ax
axarr[3][1].set_title("Car_Age Vs Price", fontsize=18)
plt.subplots_adjust(hspace=1.0)
plt.subplots_adjust(wspace=.5)
sns.despine()
```

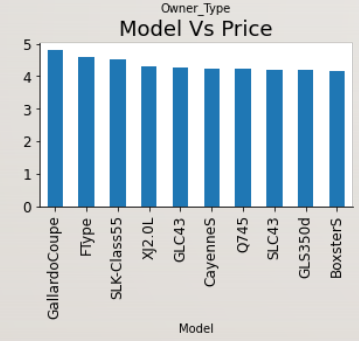


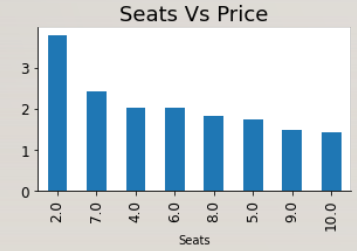


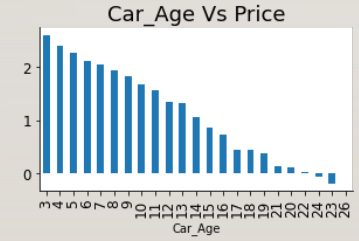












Observations:

- The price of cars is high in Coimbatore and less price in Kolkata and Jaipur
- Automatic cars have more price than manual cars.
- Diesel and Electric cars have almost the same price, which is maximum, and LPG cars have the lowest price
- First-owner cars are higher in price, followed by a second
- The third owner's price is lesser than the Fourth and above
- Lamborghini brand is the highest in price
- Gallardocoupe Model is the highest in price
- 2 Seater has the highest price followed by 7 Seater
- The latest model cars are high in price

CONCLUSION

In this course, we tried to analyse the factors influencing the used car's price.

- Data Analysis helps to find the basic structure of the dataset.
- Dropped columns that are not adding value to our analysis.
- Performed Feature Engineering by adding some columns which contribute to our analysis.
- Data Transformations have been used to normalise the columns.
- We used different visualisations for EDA like Univariate Analysis.

CONCLUSION

Through EDA, we got useful insights, and below are the factors influencing the price of the car and a few takeaways:

- Most of the customers prefer 2 Seat cars hence the price of the 2-seat cars is higher than other cars.
- The price of the car decreases as the Age of the car increases.
- Customers prefer to purchase the First owner rather than the Second or Third.
- Due to increased Fuel price, the customer prefers to purchase an Electric vehicle.
- Automatic Transmission is easier than Manual.

This way, we perform EDA on the datasets to explore the data and extract all possible insights, which can help in model building and better decision-making. However, this was only an overview of how EDA works; you can go deeper into it and attempt the stages on larger datasets. If the EDA process is clear and precise, our model will work better and give higher accuracy.

THANK YOU