INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

SECOND HAND BIKES[EDA]

Name : Mohammed Shahbaaz khan
Qualification : MCA
Course : Data Analysis
Batch no.:237

Content:

- Objective of the Project
- Summary of the Data
- Exploratory Data Analysis
- a. Data cleaning
- b. Data Manipulation
- c. Univariate Analysis steps

. d. Bivariate Analysis steps

- Conclusion

Objective of the Project:

- Collect data from user listings, user reviews, and any available third-party data on bike conditions.
- Provide clear documentation on best practices for accurate bike descriptions.
- Analyze the distribution of key variables such as bike conditions, prices, Brand, kilometers and model .
- Understanding the distribution helps in identifying patterns, outliers, and trends in the data.

Web Scraping:

- .Web scraping is a technique used to extract data from websites, and it can be a valuable tool for collecting information for an Exploratory Data Analysis (EDA) project on second-hand bikes.
- .Through “HTML code” we extract the features like Price, Kilometer, Model, Year ,Location and owner of the second hand bike by taking the div tag and class of “HTML code”.
- .Select a web scraping tool or library. Python offers several libraries for web scraping, such as BeautifulSoup and Scrapy. These tools allow you to navigate HTML and extract relevant information.

- Clean and preprocess the scraped data. Handle missing values, convert data types, and address any inconsistencies or errors in the extracted information.
- Save the cleaned data to a structured format such as CSV, Excel, or a database for further analysis in your EDA.

Summary of the Data:

	Brand	Year	Price	Model	Updated	Location	Kilometers	First_owner
0	Aprilia	2018	55000	2018	2023-08-31	Indore	14800	[1st Owner]
1	Bajaj	2018	55000	2018	2023-08-31	Gurgaon	54500	[1st Owner]
2	Suzuki	2022	175000	2022	2023-08-31	Pune	5300	[1st Owner]
3	Suzuki	2015	60000	2015	2023-08-31	Bhubaneswar	25000	[1st Owner]
4	Bajaj	2017	60000	2017	2023-08-31	Hyderabad	70000	[2nd Owner]
...	...	...	...	...	...	...	...	...
415	Royal	2015	130000	2015	2023-08-31	Mumbai	61000	[1st Owner]
416	KTM	2021	200000	2021	2023-08-31	Bangalore	15000	[1st Owner]
417	KTM	2021	150000	2021	2023-08-31	Navi Mumbai	60000	[1st Owner]
418	Royal	2015	75000	2015	2023-08-31	Delhi	9080	[1st Owner]
419	Royal	2023	225000	2023	2023-08-31	Bangalore	2250	[1st Owner]

420 rows × 8 columns

- The dataset was collected by scraping information from [List of Websites] that host second-hand bike listings.
- An EDA project on second-hand bikes aims to enhance the functionality and user experience of the platform, build trust among users, and create a more efficient and reliable marketplace for buying and selling second-hand bikes.
- Variables include Brand, model, year, price, kilometer , owner, location and update.
- Provide basic statistics such as mean, median, and standard deviation for relevant numerical variables (e.g., price, kilometer).
- Determine the most popular bike brands and models in the second-hand market based on the dataset.
- Highlight any trends or patterns related to brand preferences.

Exploratory Data Analysis

a.Data cleaning:

Columns:

check the null values in the DataFrame[No null values in the DataFrame].

Brand:info(),shape(),isnull(),unique(),nunique(),value_counts().Datatype is “obj”.

Year:Datatype is converted from “obj” to “int”.

Price:shape(),isnull(),value_counts(), unique(),nunique(),info().Datatype is converted from “obj” to “int”.Removed special character like comma(,).

- Model:info(),shape(),isnull(),value_counts(), unique(),nunique().Remove model from the column.Datatype is “obj”.
- Updated: info(),shape(),isnull(),unique(),nunique().Datatype is converted from “obj” to “int”.Converted month name to number.
- Location: info(),shape(),isnull(),unique(),nunique().Datatype is “obj”.
- Kilometer: info(),shape(),isnull(),value_counts(), unique(),nunique().Datatype is converted from “obj” to “int” and removed special character like comma(,) and kms.
- First_owner: isnull().Datatype is “obj”.

b.Data Manipulation:

- Import libraries like requests, pandas, matplotlib, seaborn, BeautifulSoup and re.
- Prettify is a parameter used to understand the complete “HTML Code”.
- Brand:info(),shape() unique(),nunique(),value_counts().
- Price:shape(),isnull(),value_counts(), unique(),nunique(),info().
- Year: isnull(),unique(),nunique(),value_counts().

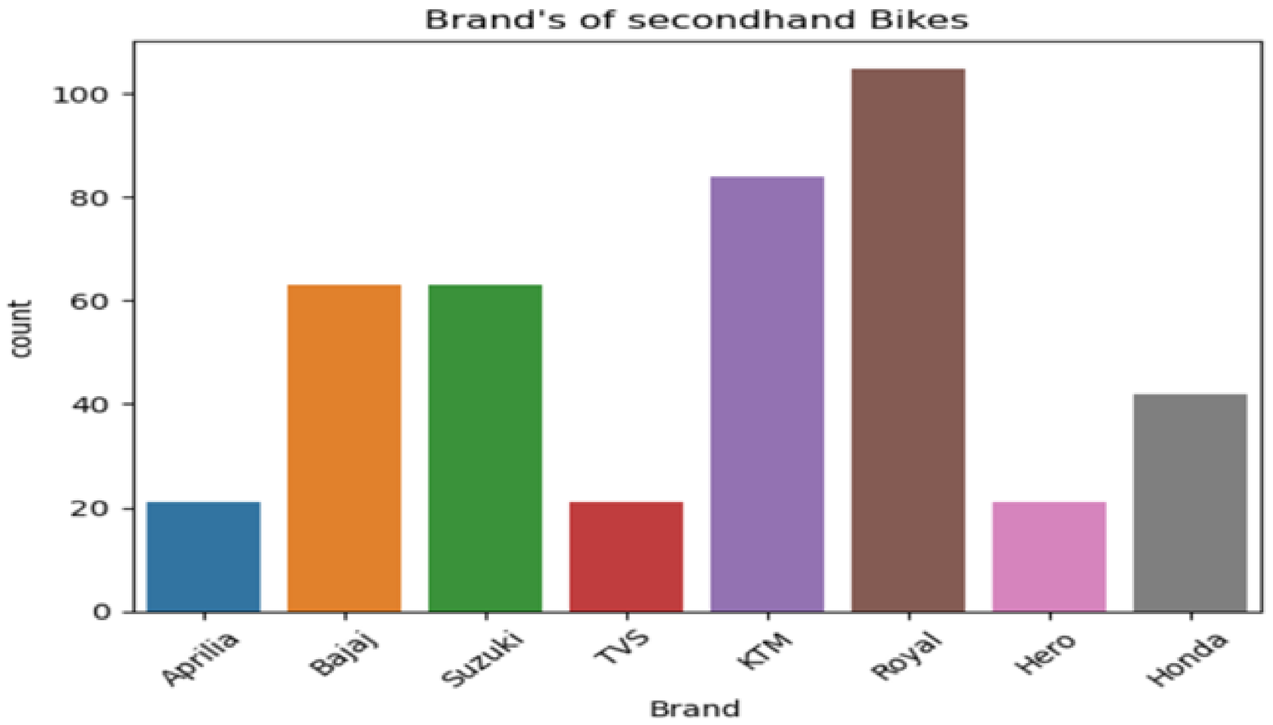
- Model:info(),shape(),isnull(),value_counts(), unique(),nunique().
- Updated: info(),shape(),isnull(),unique(),nunique().
- Location: info(),shape(),isnull(),unique(),nunique().
- Kilometer: info(),shape(),isnull(),value_counts(), unique(),nunique().
- First_owner: isnull().Datatype is “obj”.

C.Univariate Analysis steps:

- Gather the data for the variable of interest. Ensure that your dataset is complete, accurate, and in a format suitable for analysis.
- .Create visual representations of your variable's distribution. Choose appropriate plots or charts based on the type of variable (categorical or numerical). For example, bar plots, pie charts, histograms, or box plots can be useful.

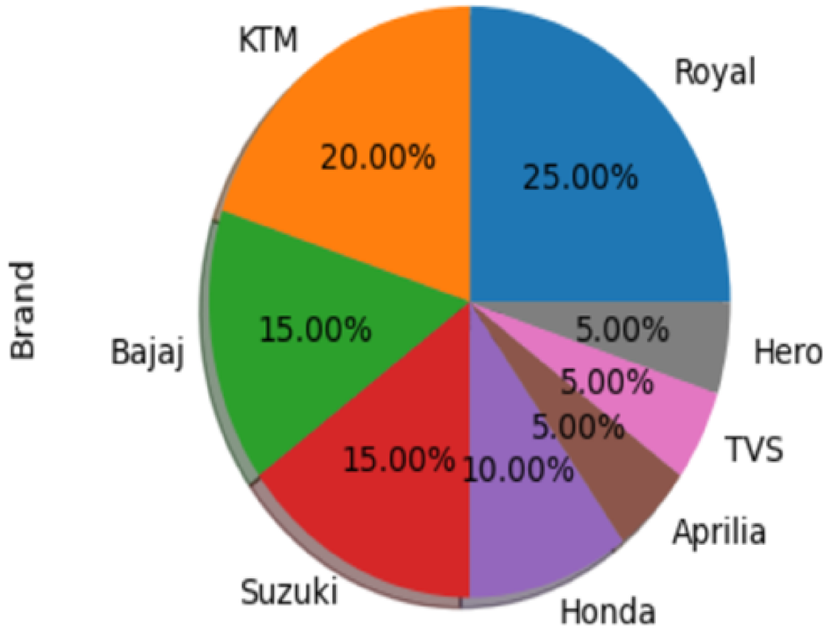
Countplot:

- X-Axis displays the Brand of second hand bikes.
- Y-Axis displays the count the Brand of second hand bikes.
- Maximum count of Brand of second hand bike is Royal and the minimum is count of Brand of second hand bike is Hero.



Pie chart:

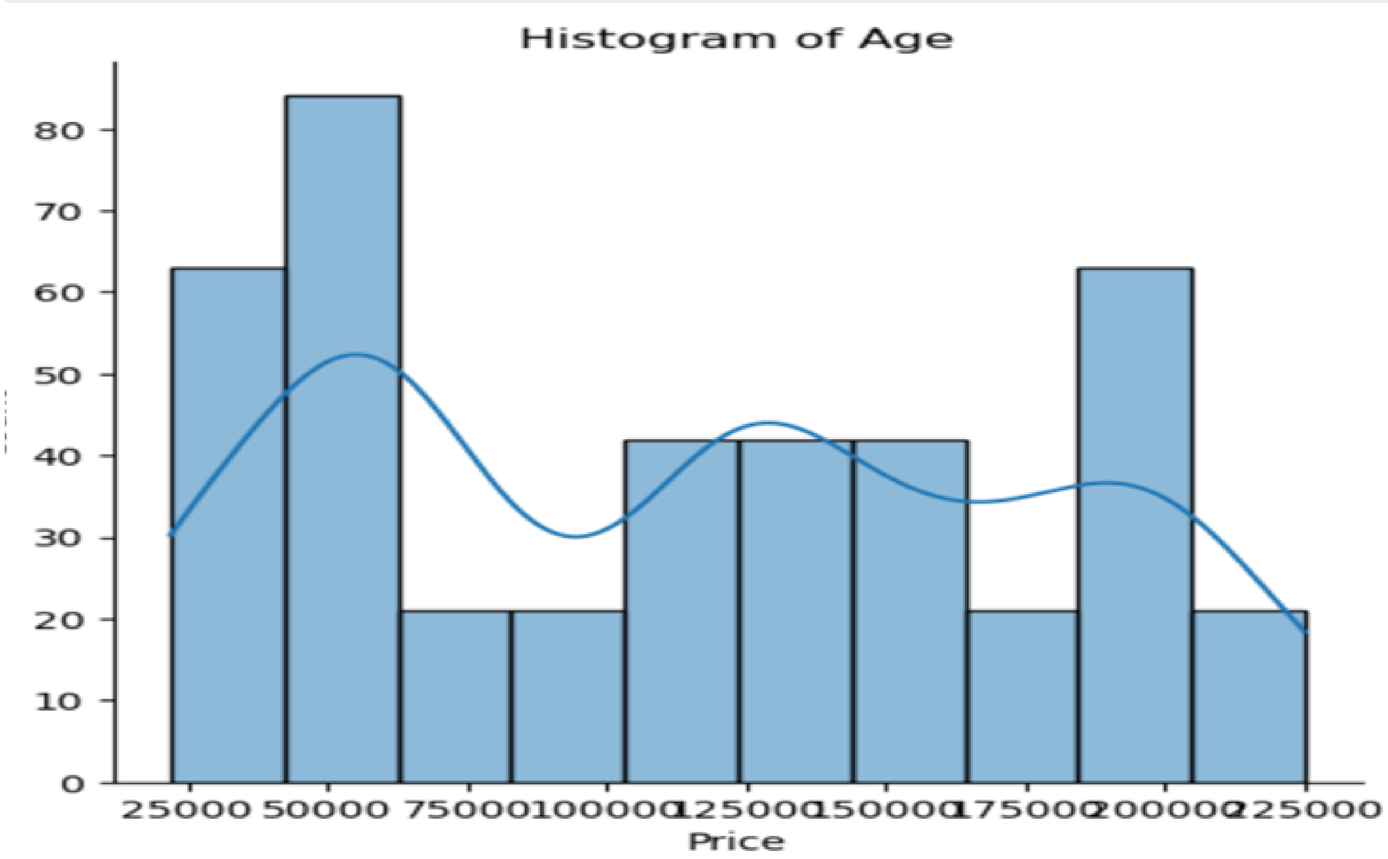
	Brand
Royal	105
KTM	84
Bajaj	63
Suzuki	63
Honda	42
Aprilia	21
TVS	21
Hero	21



- A pie chart is a circular statistical graphic that is divided into slices to illustrate numerical proportions.
- Royal Brand bike is having higher in percentage compare to other brand bikes i.e 25% and Aprilia , TVS and Hero is having least and same percentage compare to other brand bikes i.e 5%.

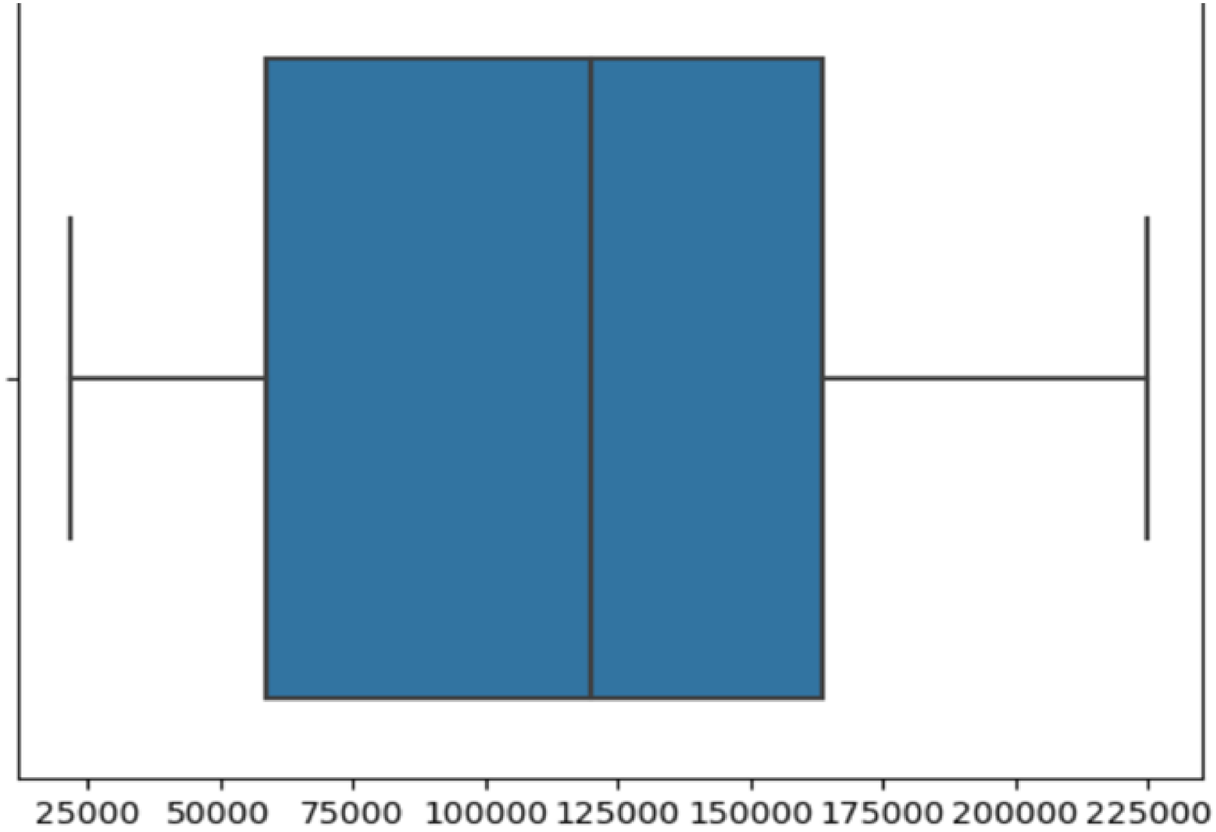
Histogram:

- .The x-axis (horizontal axis) typically represents the Price (variable) is being measured, while the y-axis (vertical axis) represents the frequency or count of observations in each bin.
- .Histograms are divided into bins, which are intervals or ranges of values. The data is grouped into these bins, and the height of each bar on the histogram represents the frequency of observations within that bin. It is the Bimodal Distribution because of two distinct peaks.



Boxplot:

- A boxplot, also known as a box-and-whisker plot, is a graphical representation that displays the distribution of a dataset based on a five-number summary: minimum, first quartile (Q1), median (Q2), third quartile (Q3), and maximum.
- Minimum is [Q1-1.5*IQR] -98750.0,Q1= 58750,Q3=163750,IQR=105000, Maximum is [Q3+1.5*IQR] 321250.0
- There is no outliers.



Bivariate Analysis steps:

- Bivariate analysis involves the simultaneous analysis of two variables to understand the relationship between them.
- Ensure that your dataset is clean and in a suitable format for analysis. Handle missing data, outliers, and transform variables if necessary.
- Identify the two variables of interest for your analysis. These variables can be numerical or categorical, depending on the type of analysis you want to perform.
- Explore the distribution of each variable individually. Use summary statistics, histograms, box plots, or bar charts to understand the characteristics of each variable.

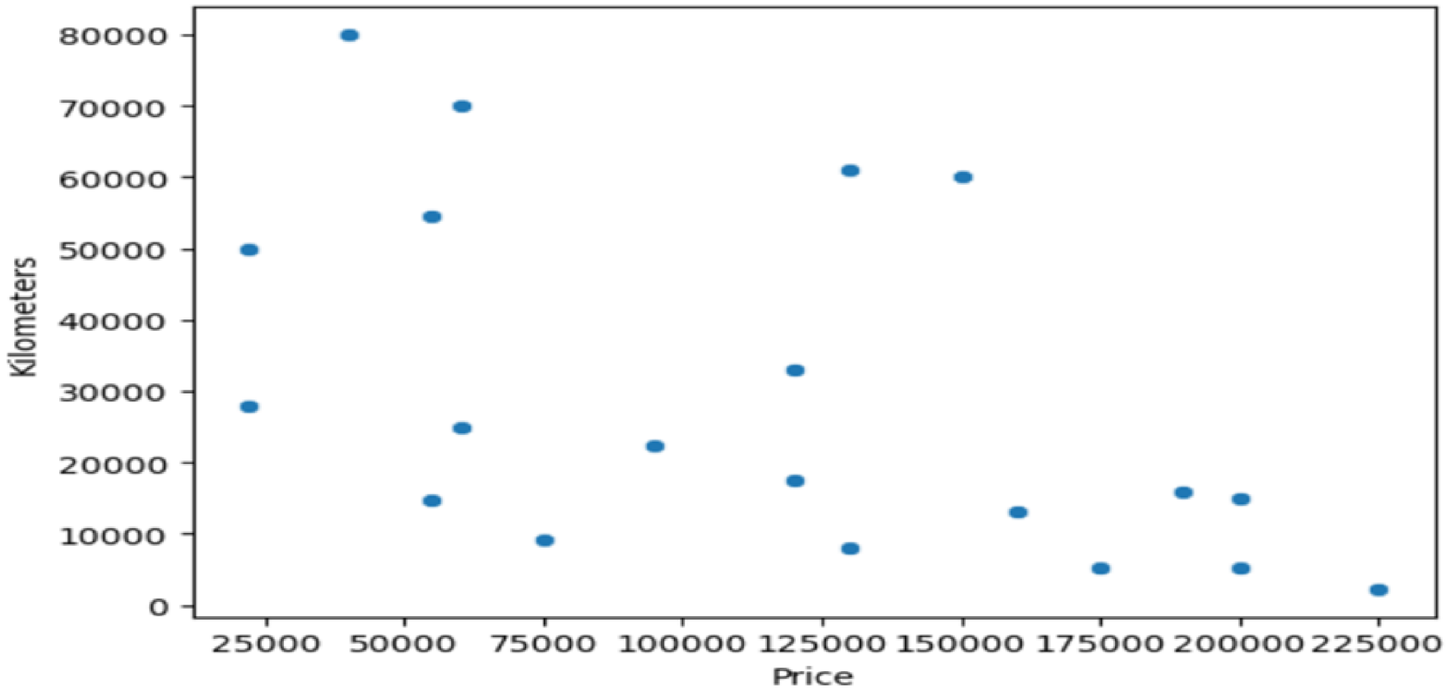
Numeric vs Numeric:

Scatter plot:

The x-axis represents the number of Price, and the y-axis represents the Kilometer. Each point on the scatter plot corresponds to a specific pair of kilometer and price values.

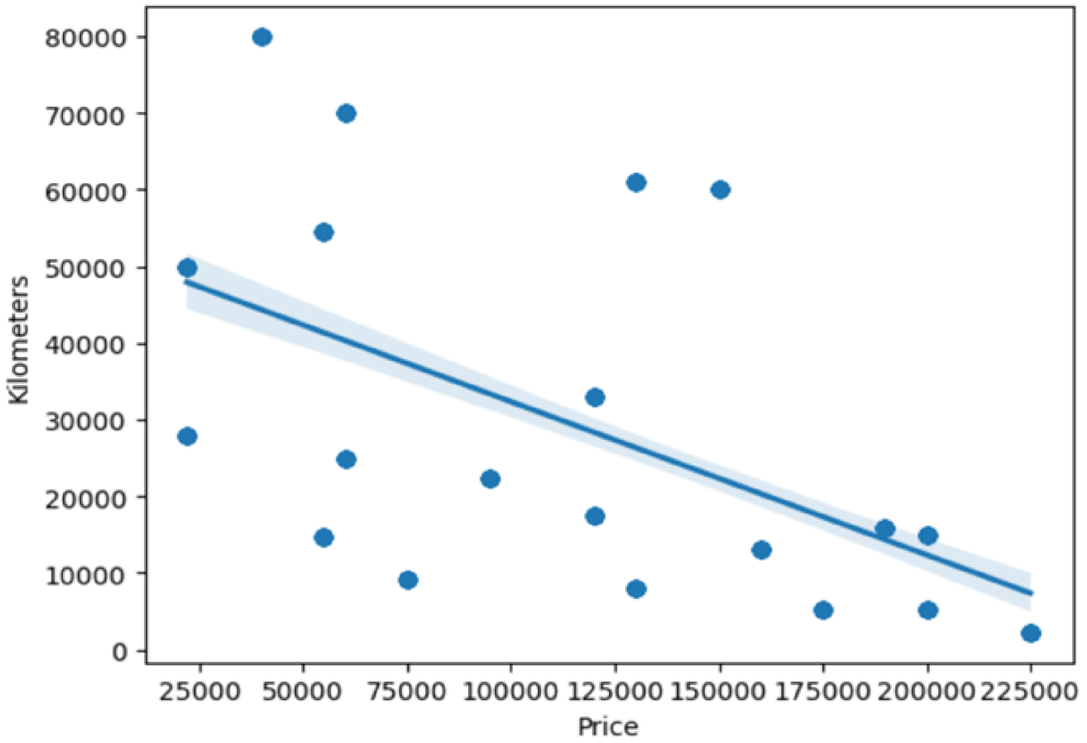
- .It displays individual data points on a two-dimensional plane, with one variable on the x-axis and the other on the y-axis. Each point on the plot represents a unique combination of values for the two variables.

.It shows each data point corresponds to a specific pair of values for the two variables being analyzed.

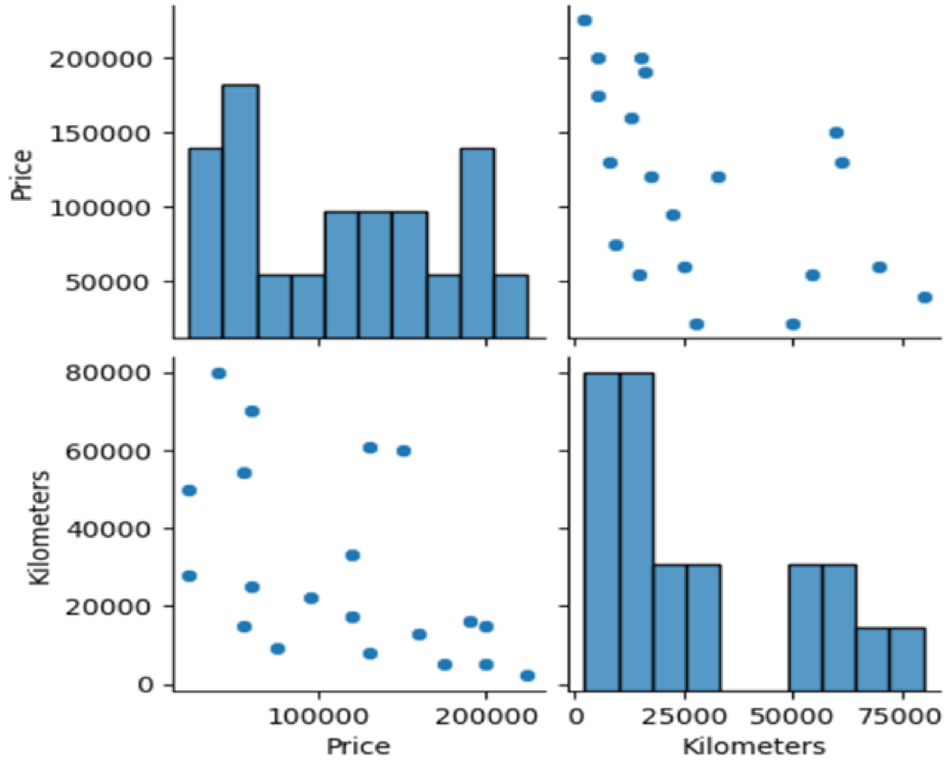


Regression Plot:

- visualization that shows the relationship between two numerical variables Price and Kilometers. It is a linear relationship between the two variables Price and Kilometers.



Pair plot:

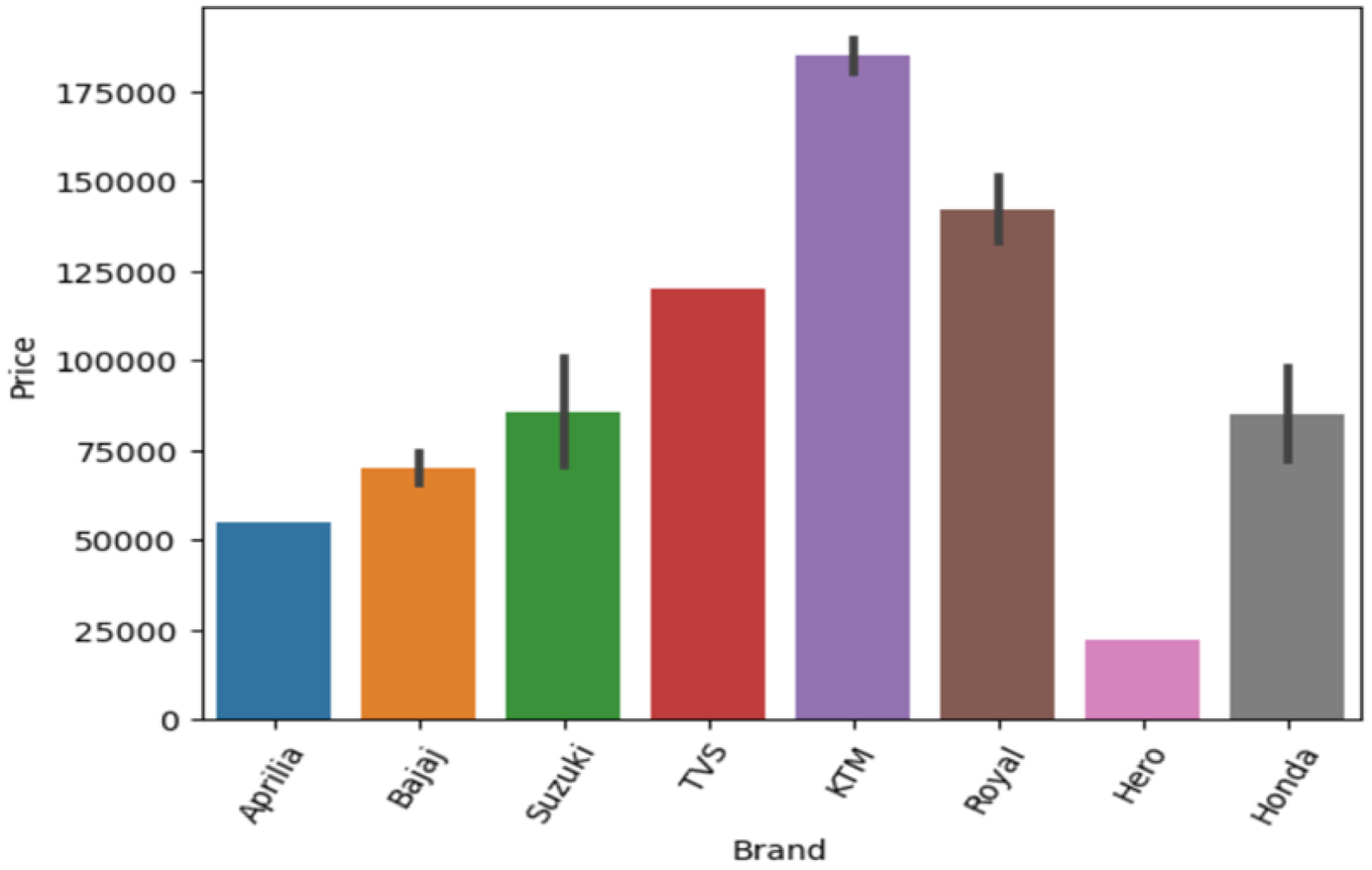


- The pair plot will display scatter plots for both "Price" and "Kilometer" individually (on the diagonal) and a scatter plot for the pair of variables . The upper triangle of the plot mirrors the lower triangle since the relationship between "Price" and "Kilometer" is symmetric.

Heat Map:

- The heatmap provides a visual representation of the correlation between pairs of variables like Price, Kilometer ,Model and Year, with warmer colors indicating stronger positive correlations and cooler colors indicating stronger negative correlations. Diagonal are equals to 1.

BarGraph:



- The x-axis represents the "Brand," and the y-axis represents the "Price" in currency units. .
- Each bar in the plot corresponds to a specific brand, and the height of the bar indicates the average or total price associated with that brand. Taller bars represent higher average prices(KTM), while shorter bars represent lower average prices(Hero).

Conclusion:

Opting for a second-hand bike can be a sensible choice for those looking to save money, access a wider range of models, and potentially avoid rapid depreciation. However, careful consideration of the bike's condition, maintenance history, and individual preferences is crucial to making a wise decision. It's also advisable to research the market thoroughly and, if possible, consult with a knowledgeable mechanic before making a purchase.