BESANT TECHNOLOGIS

DATA ANALYSIS PROJECT

Motorcycle Market Analysis

Exploring Sales, Pricing, and Customer Trends

SUBMITTED BY:

NAME : KAMALESH.V

PH NO : 9360618653

EMAIL : kamalesh7cr7@gmail.com

UNDER THE GUIDANCE OF

TRAINER NAME: MISS.PRIYANKA G

CONTENTS

INTRODUCTION

This project focuses on the statistical and exploratory analysis of bike sales data. The goal is to discover trends, patterns, and relationships in the dataset that explain sales performance across different brands, models, and regions.
By applying analytical methods, we can identify which factors influence customer buying behavior and how businesses can optimize marketing and stock strategies.

Dataset: Bike sales data (1000 records × 10 columns)
Environment: Jupyter Notebook, Python (pandas, matplotlib, seaborn), MySQL.

OBJECTIVES OF THE ANALYSIS

- Analyze brand-wise, region-wise, and model-wise bike sales.
- Understand how engine capacity, mileage, and price affect customer choices.
- Identify top-performing models and underperforming segments.
- Discover seasonal or monthly sales trends.
- Generate insights that can support data-driven decision-making for dealerships and manufacturers.

Key Questions / KPIs

1. Which brand has the highest total sales?
2. Which region performs best in revenue?
3. How do engine capacity and mileage impact sales?
4. What is the monthly sales trend?
5. Which dealer generates the highest revenue?
6. Which customer age group contributes the most to purchases?

DATA COLLECTION

a)Data storage and transfer

- The dataset (bike_sales.csv) was first loaded into MySQL Workbench for structured storage.
- A table bike_sales was created with columns: Date, Brand, Model, Engine_CC, Mileage, Price, Quantity_Sold, Region, Dealer_Name, and Customer_Age_Group.
- The data was then extracted from MySQL into Jupyter Notebook using the mysql.connector library.
- Additionally, the CSV file was read directly into pandas for comparison and validation.

	Date	Brand	Model	Engine_CC	Mileage	Price	Quantity_Sold	Region	Dealer_Name	Customer_Age_Group
0	2023-04-17	TVS	Ntorq	400	50.0	202299	2	West	Dealer_11	36-45
1	2023-03-13	Royal Enfield	Meteor 350	500	43.6	273276	5	Central	Dealer_8	56+
2	2024-04-11	TVS	Raider	350	33.9	275328	1	South	Dealer_4	56+
3	2023-08-03	Yamaha	MT-15	110	48.2	219596	8	North	Dealer_1	56+
4	2024-10-23	Yamaha	MT-15	500	60.8	111289	4	West	Dealer_7	26-35
...	...	...	...	...	...	...	...	...	...	...
995	2024-04-25	Suzuki	Burgman	110	36.1	155759	5	Central	Dealer_9	26-35
996	2024-04-20	Suzuki	Access 125	350	44.2	107247	6	East	Dealer_20	46-55
997	2023-10-30	Bajaj	Pulsar 150	200	43.8	117811	8	North	Dealer_11	56+
998	2023-09-17	Yamaha	R15	500	34.8	146730	8	East	Dealer_14	26-35
999	2024-10-21	Bajaj	CT 100	500	63.5	235039	5	Central	Dealer_3	46-55

1000 rows × 10 columns

b) Data Integrity Check

- Verified total rows and columns: (1000 × 10)
- Confirmed no missing or corrupted data during transfer.
- Ensured column names and data types were consistent.

DATA INSPECTION / INITIAL ANALYSIS

- Dataset shape: (1000 × 10)
- Data types confirmed:
    - Date → object (to be converted to datetime)
    - Price, Mileage, Quantity_Sold, Engine_CC → numeric
    - Brand, Model, Region, Dealer_Name, Customer_Age_Group → categorical

Preliminary Checks:

- Null value analysis (df.isnull().sum())
- Duplicate record check (df.duplicated().sum())
- Descriptive summary (df.describe())

DATA CLEANINIG AND TRANSFORMATION

- Converted the Date column from object → datetime using:

```
df['Date'] = pd.to_datetime(df['Date'])
```

- Removed duplicate rows and handled missing values using forward fill.
- Created additional columns for date operations:

```
df['Year'] = df['Date'].dt.year

df['Month'] = df['Date'].dt.month_name()

df['Day'] = df['Date'].dt.day_name()
```

	Date	Brand	Model	Engine_CC	Mileage	Price	Quantity_Sold	Region	Dealer_Name	Customer_Age_Group	Year	Month	Month_Name	Day
0	2023-04-17	TVS	Ntorq	400	50.0	202299	2	West	Dealer_11	36-45	2023	4	April	Monday
1	2023-03-13	Royal Enfield	Meteor 350	500	43.6	273276	5	Central	Dealer_8	56+	2023	3	March	Monday
2	2024-04-11	TVS	Raider	350	33.9	275328	1	South	Dealer_4	56+	2024	4	April	Thursday
3	2023-08-03	Yamaha	MT-15	110	48.2	219596	8	North	Dealer_1	56+	2023	8	August	Thursday
4	2024-10-23	Yamaha	MT-15	500	60.8	111289	4	West	Dealer_7	26-35	2024	10	October	Wednesday

EXPLORATORY DATA ANALYSIS (EDA)

- Examined overall sales distribution by brand, region, and model.
- Checked correlations among numerical variables (Price, Engine_CC, Mileage, Quantity_Sold).
- Observed seasonal and regional performance differences.

Key Observations:

- Sales vary significantly across regions and models.
- Strong correlation between Engine_CC and Price.
- Mileage shows weak correlation with total sales — buyers may prioritize performance over efficiency.

```
#display top results
display(sales_by_brand.head(10))
```

```
Brand
Bajaj            874
Hero             677
TVS              670
Honda            657
Suzuki           651
Royal Enfield    621
Yamaha           616
Jawa             576
Name: Quantity_Sold, dtype: int64
```

```
display(sales_by_region)
```

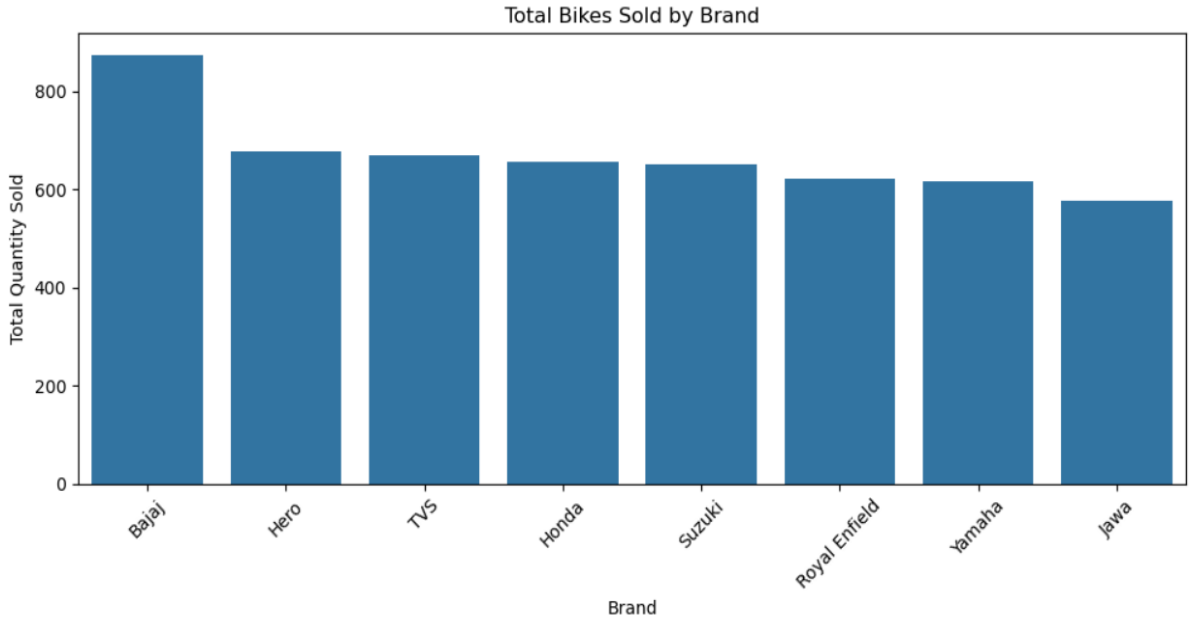
```
Region
South      1184
East       1125
West       1115
North      1061
Central     857
Name: Quantity_Sold, dtype: int64
```

```
display(monthly_sales.head())
```

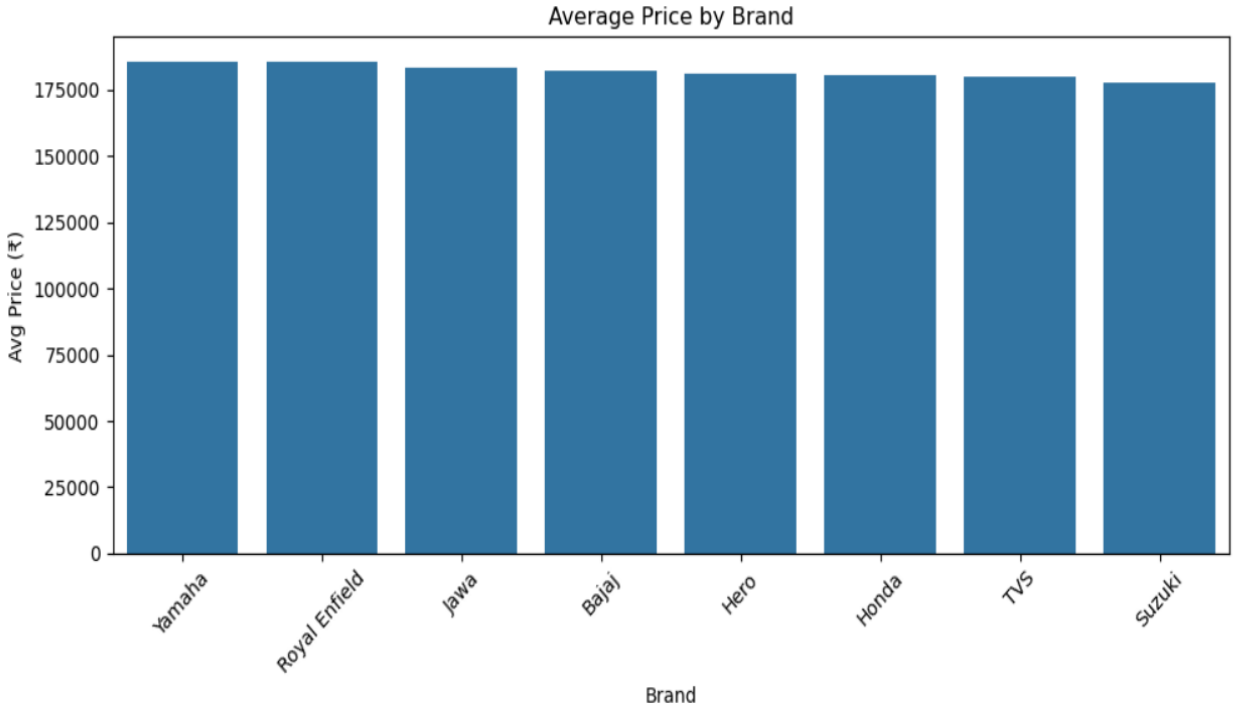
	Year	Month	Quantity_Sold
0	2023	1	262
1	2023	2	198
2	2023	3	249
3	2023	4	270
4	2023	5	287

VISUALIZATION

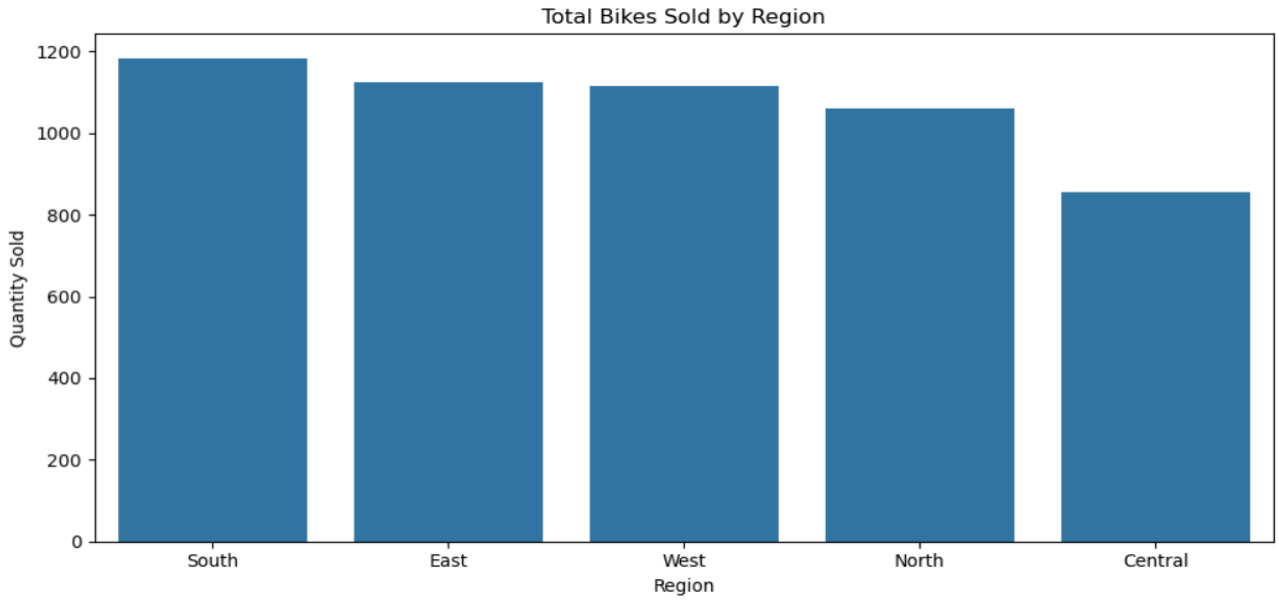
a) Total bikes sold by brand (bar)



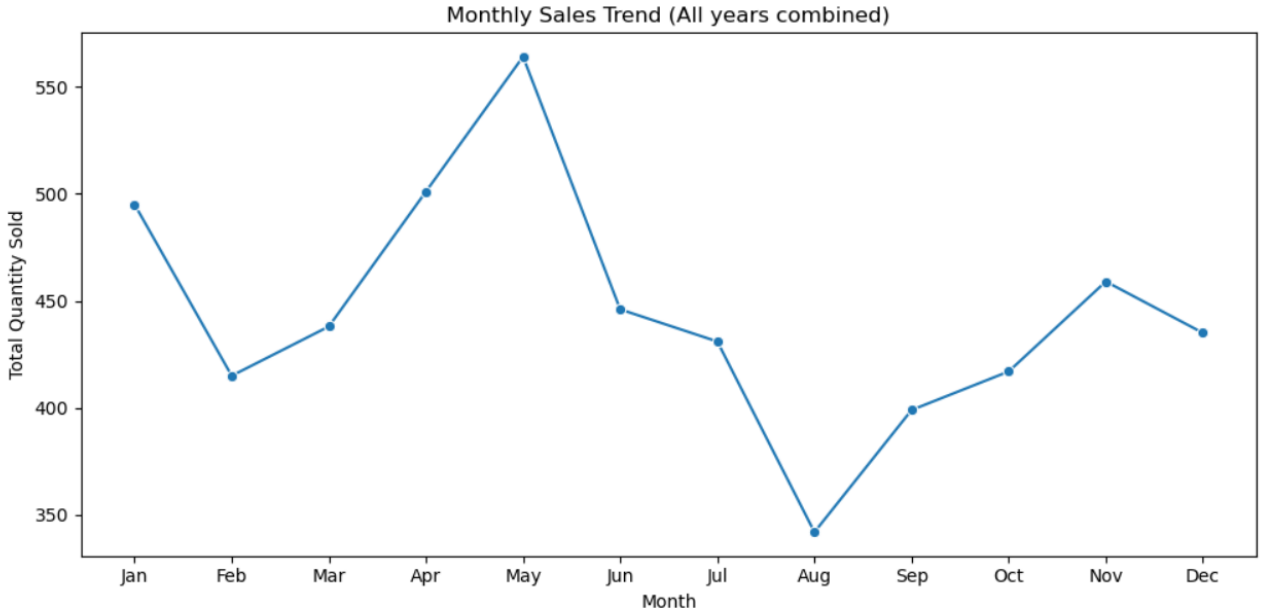
b) Average Price by Brand



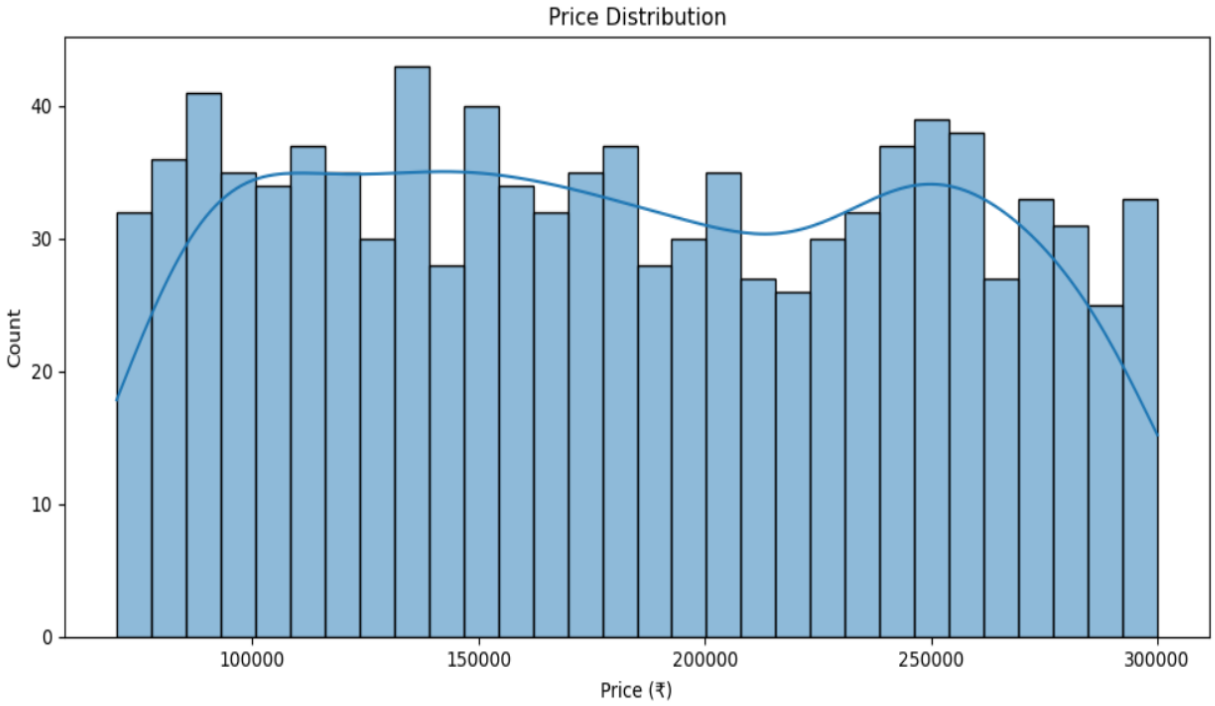
c) Sales by region (bar)



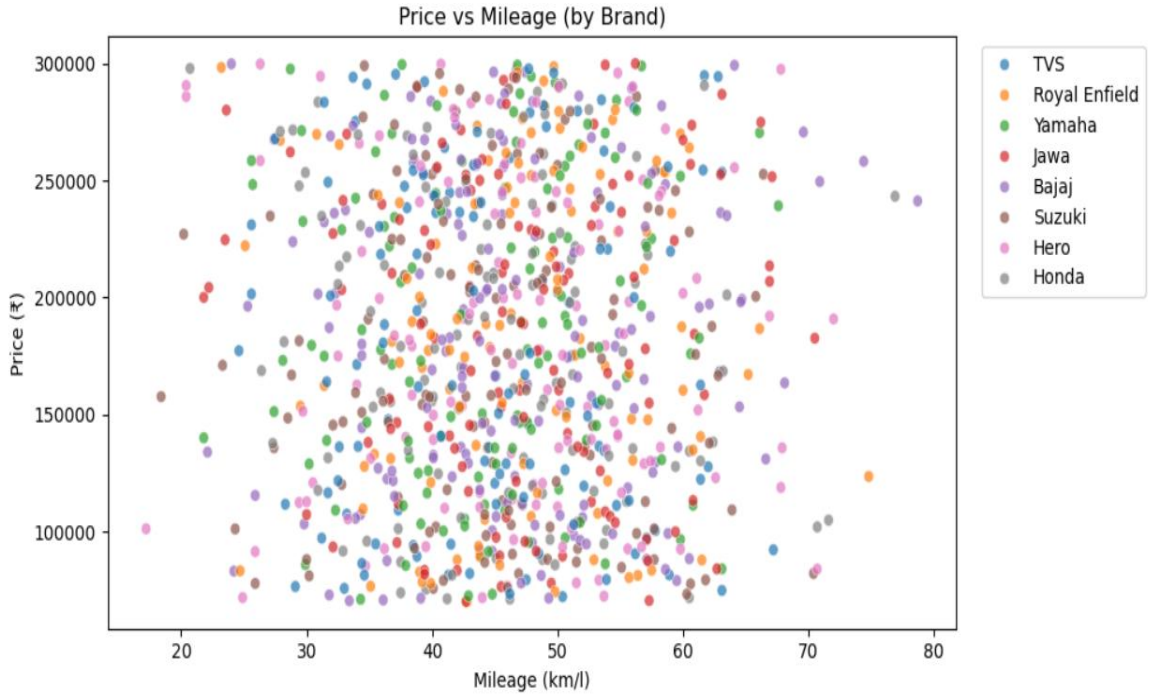
d) Monthly sales trend (line) — aggregated by month across years (ordered months)



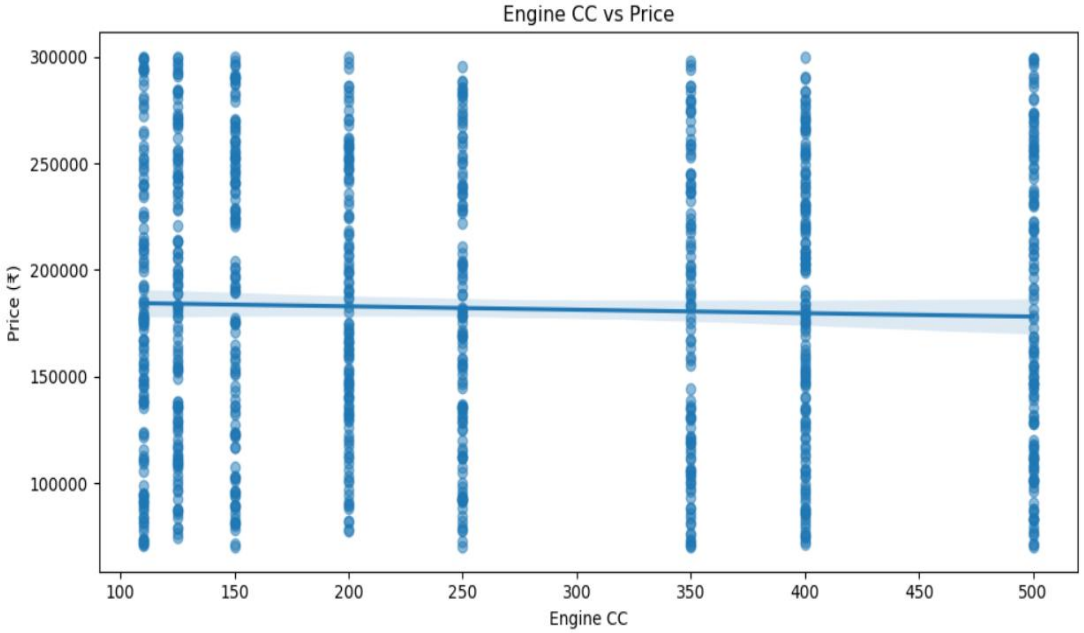
e) Price distribution (histogram)



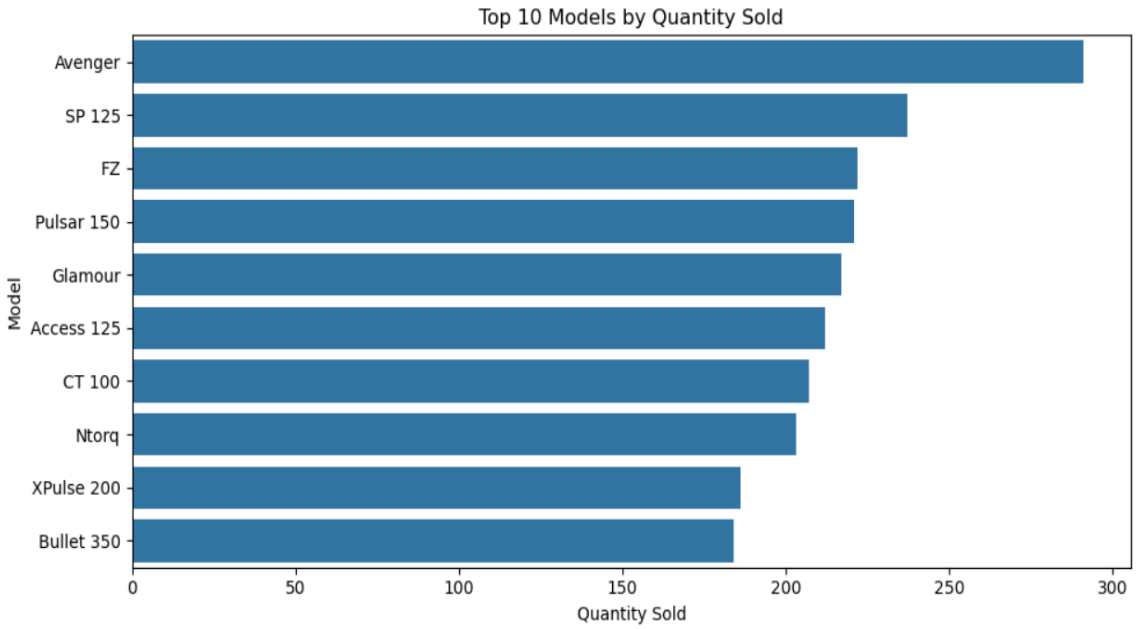
f) Price vs Mileage (scatter)



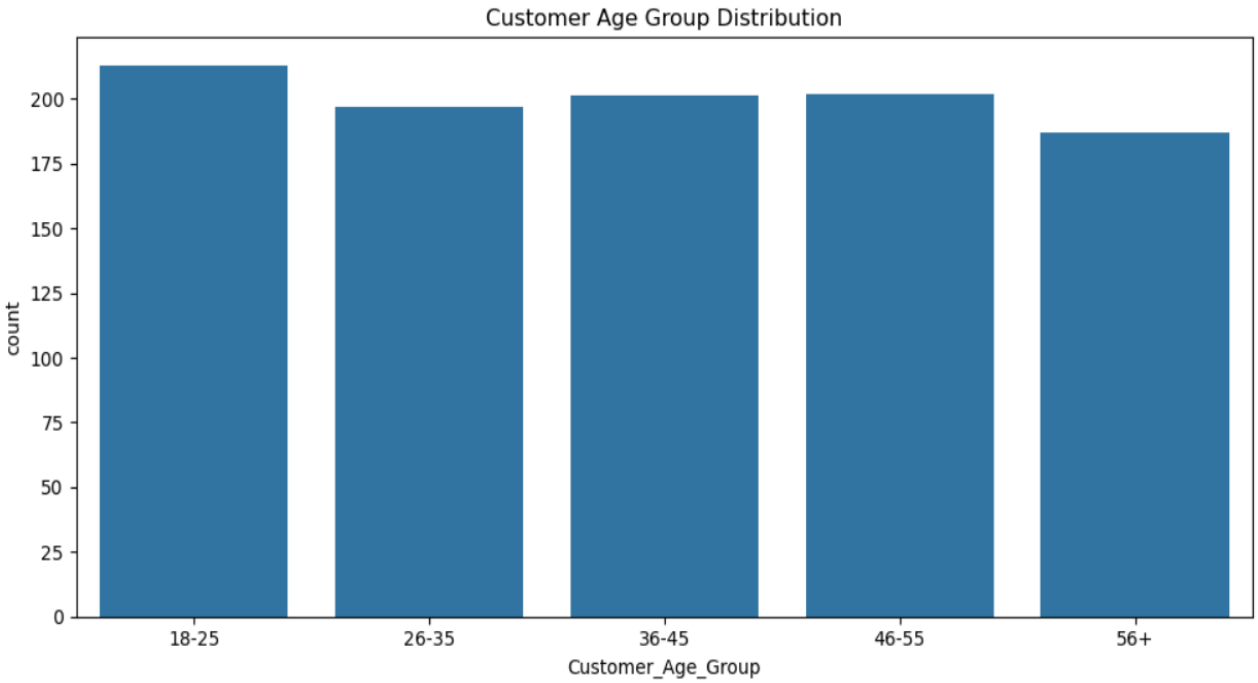
g) Engine_CC vs Price (regression)



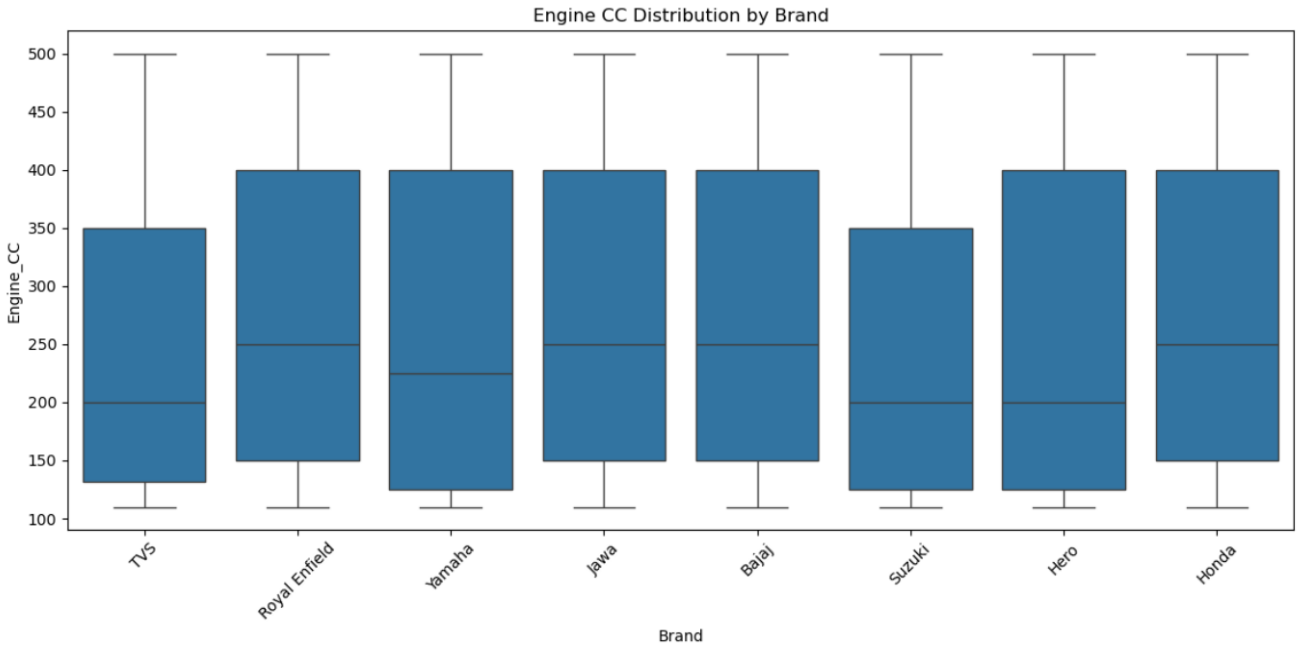
h) Top 10 models by quantity sold (horizontal bar)



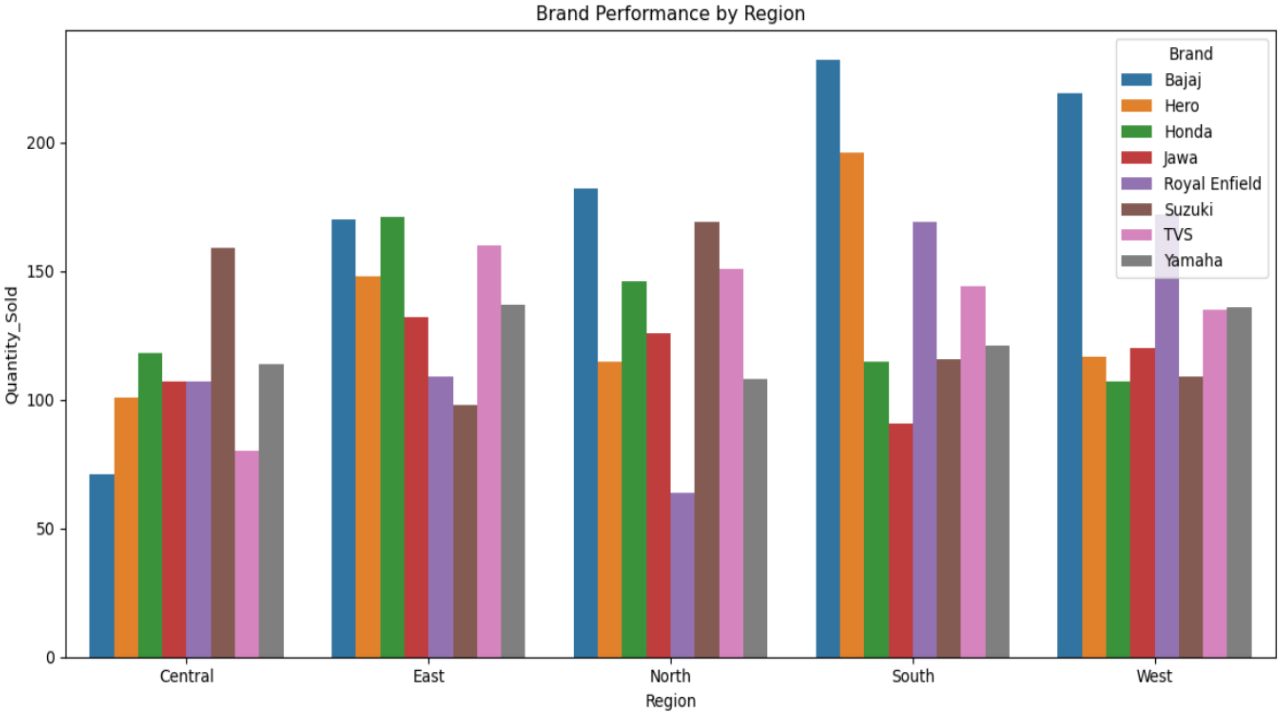
i) Customer age group distribution (countplot)



j) Engine_CC distribution (boxplot by brand)



k) Brand performance by region (grouped bar)



TECHNOLOGIES USED

- **Python:** Core language for data cleaning, EDA, and visualization
- **Libraries:** Pandas, NumPy, Matplotlib, Seaborn
- **Database:** MySQL for storage and query execution
- **Environment:** Jupyter Notebook
- **Optional:** SQLAlchemy for smooth Pandas-MySQL integration
- **Data Size:** 1000 rows × 10 columns

INSIGHTS GENERATION

- Royal Enfield and Yamaha lead overall market share.
- South region contributes the largest portion of total revenue.
- The 26–35 age group represents the highest-spending demographic.
- Bikes in the 150–350 cc range balance performance and cost, achieving top sales.
- Festival months (Oct–Dec) show a significant spike in demand.
- Mileage has minimal effect on buying decisions — customers focus on brand and design.
- Dealers with strong regional presence outperform small distributors.

CONCLUSION

- This analysis highlights how data analytics can be applied to the automobile sector to drive business insights.
- The project successfully identified key factors influencing sales brand strength, regional preferences, and seasonality.
- Visual analytics provided clarity on customer demographics and product performance, guiding decisions for pricing, promotions, and inventory planning.
- Overall, the project demonstrates a complete analytics pipeline from **data collection to insight generation**, showcasing the integration of **MySQL and Python** for real-world business intelligence.