BIKE PRICES – FEATURE ENGINEERING & EXPLORATORY DATA ANALYSIS (EDA)

SUBTITLE:
A MACHINE LEARNING PROJECT

BY SWARAJ VERMA – DATA ANALYST

OBJECTIVE:
TO ANALYZE USED BIKE LISTINGS AND IDENTIFY KEY FACTORS AFFECTING SELLING PRICES USING FEATURE ENGINEERING AND EXPLORATORY DATA ANALYSIS.

- Understand the **impact of bike features** (e.g., power, mileage, age) on pricing.
- Extract hidden insights from **unstructured model data**.
- Engineer new variables like **bike age**, **brand**, and **owner category encoding**.
- Identify **data quality issues** (missing values, inconsistencies) and apply cleaning techniques.
- Perform **univariate and bivariate analyses** to uncover pricing trends.
- Provide a **data-driven foundation** for predictive modeling of used bike prices.

Dataset Overview

Dataset Columns:

model_name: Full name (includes model, year, engine info)
•model_year: Manufacturing year
•kms_driven: Total kilometers driven
•owner: Owner category (1st, 2nd, etc.)
•location: City/region of sale
•mileage: Fuel efficiency (kmpl)
•power: Engine power (BHP)
•price: Target variable (INR)
•brand: Brand name (to be extracted)
•cc: Engine capacity (if available)

STEP 1 – LOAD & INSPECT DATA
DATA COLLECTION & PREPARATION

- Checked **dataset shape** (rows × columns)
- Reviewed **data types** of each column
- Identified **missing values** using isnull().sum()
- Detected **duplicate records** using duplicated().sum()
- Generated **summary statistics** for numerical & categorical variables
- using describe(include='all')

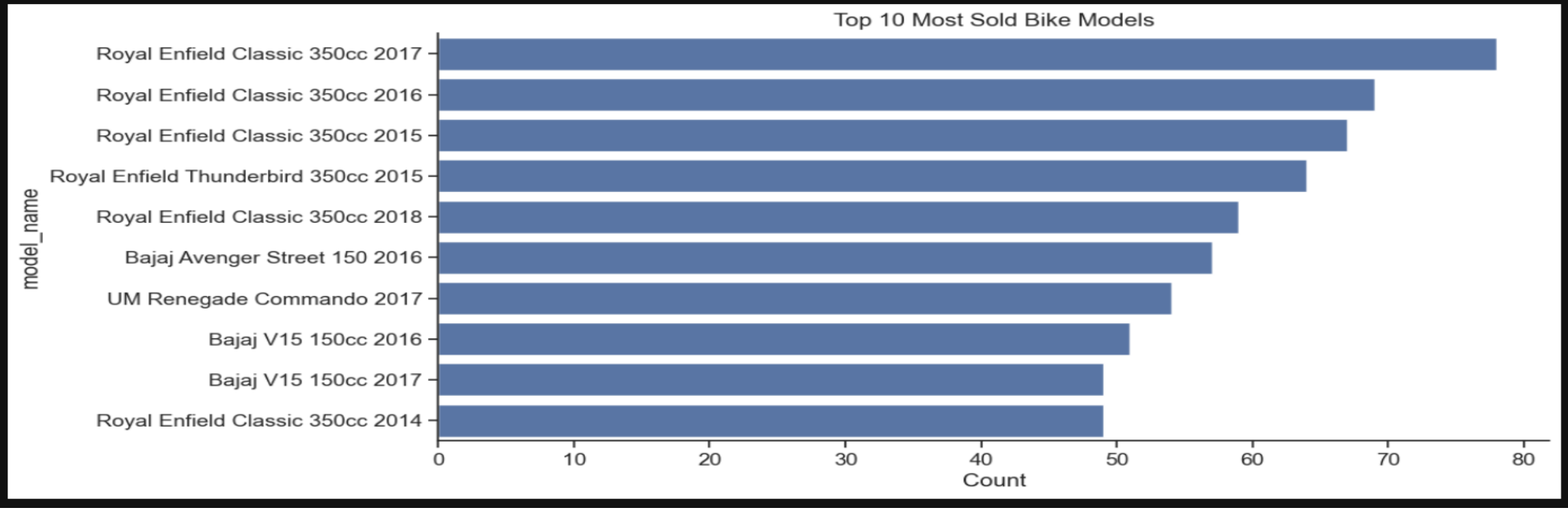
HEAT MAP SHOW


```
5]: plt.figure(figsize=(12,6))
```

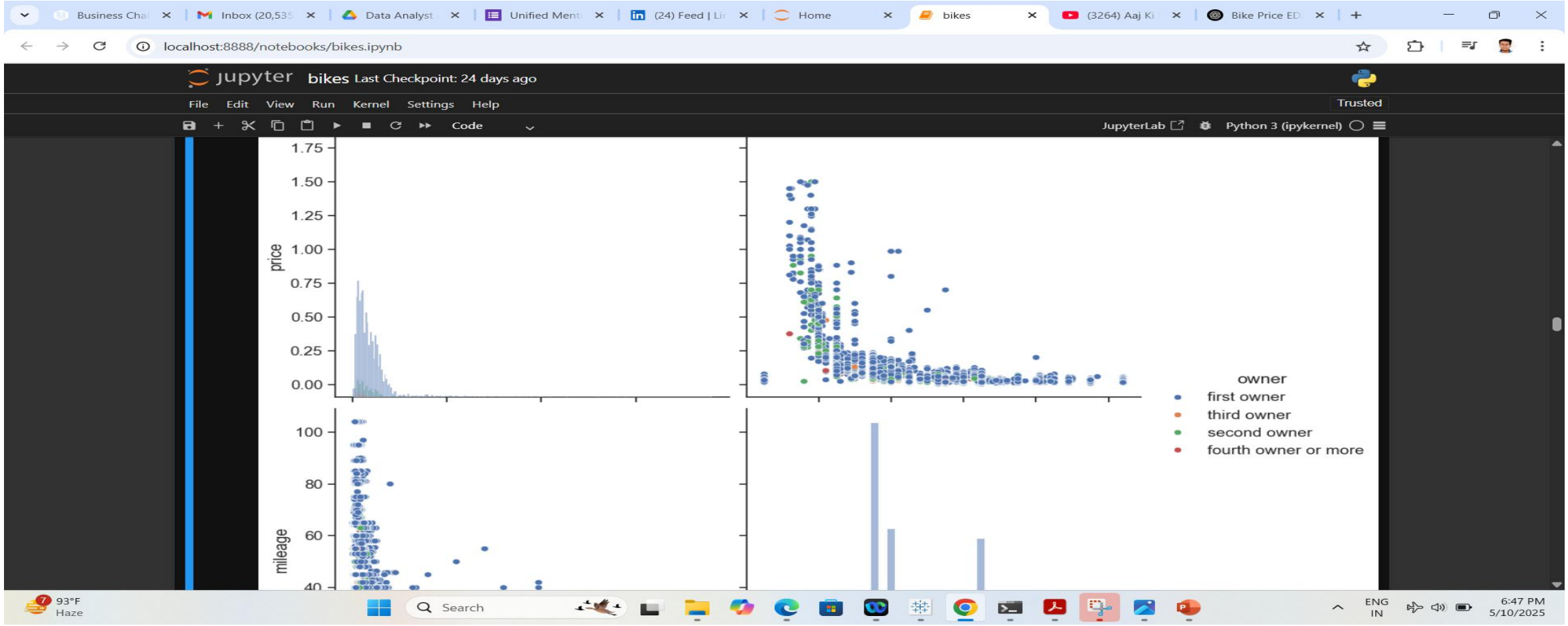
Search

TOP 10 MOST SOLD BIKE MODELS

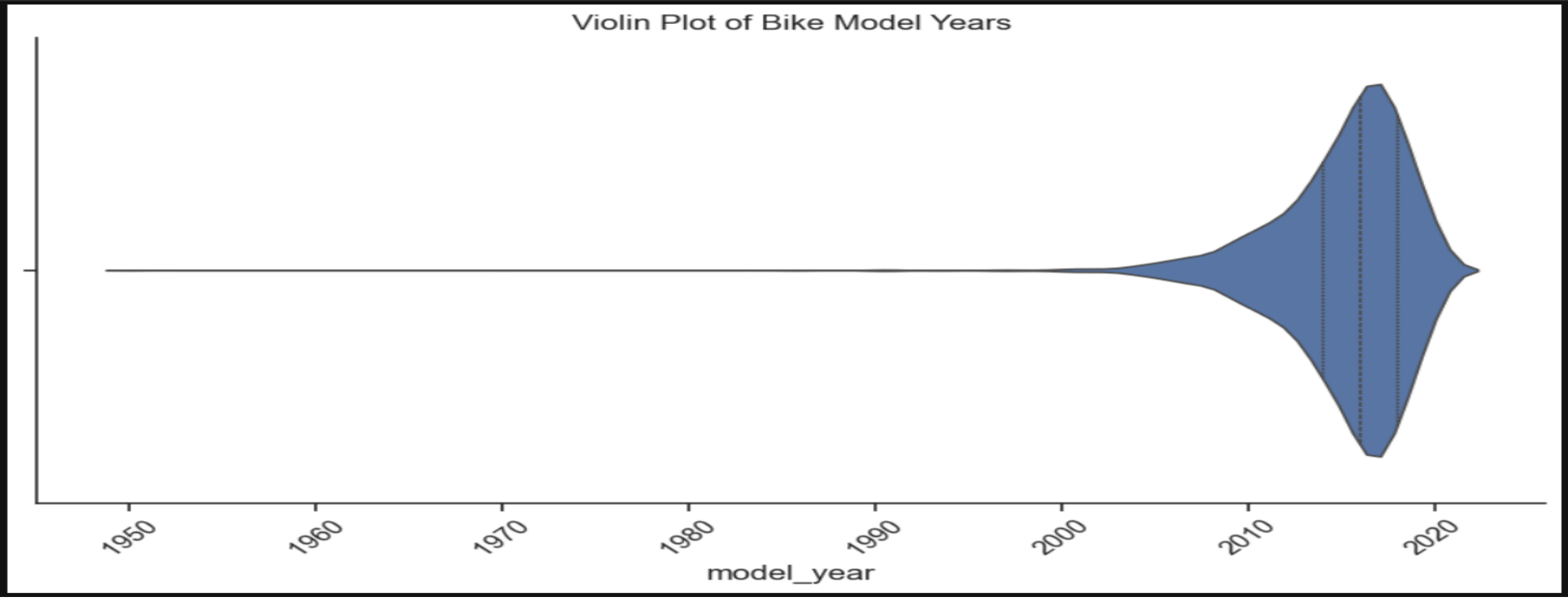
```
plt.title('Top 10 Most Sold Bike Models')
plt.xlabel('Count')
sns.despine()
plt.show()
```



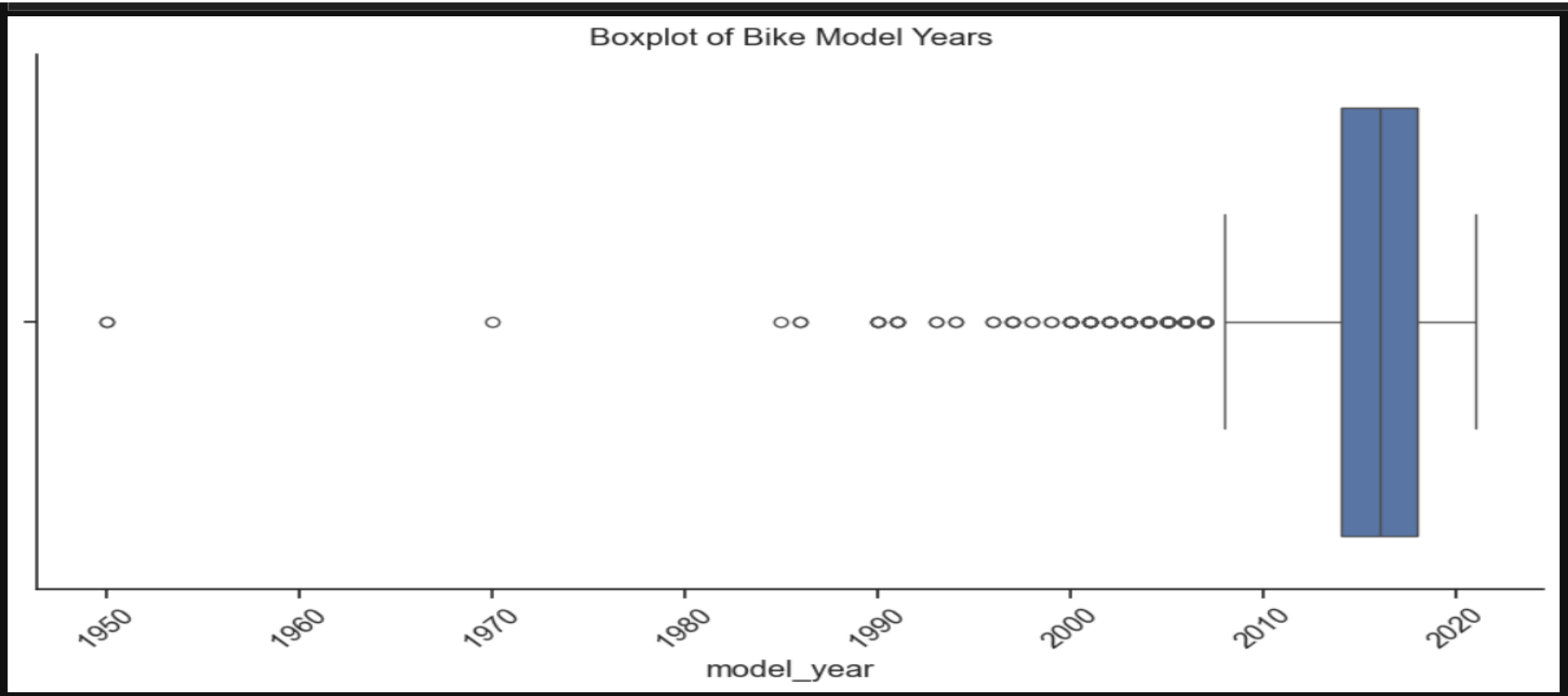
SCATTER PLOT



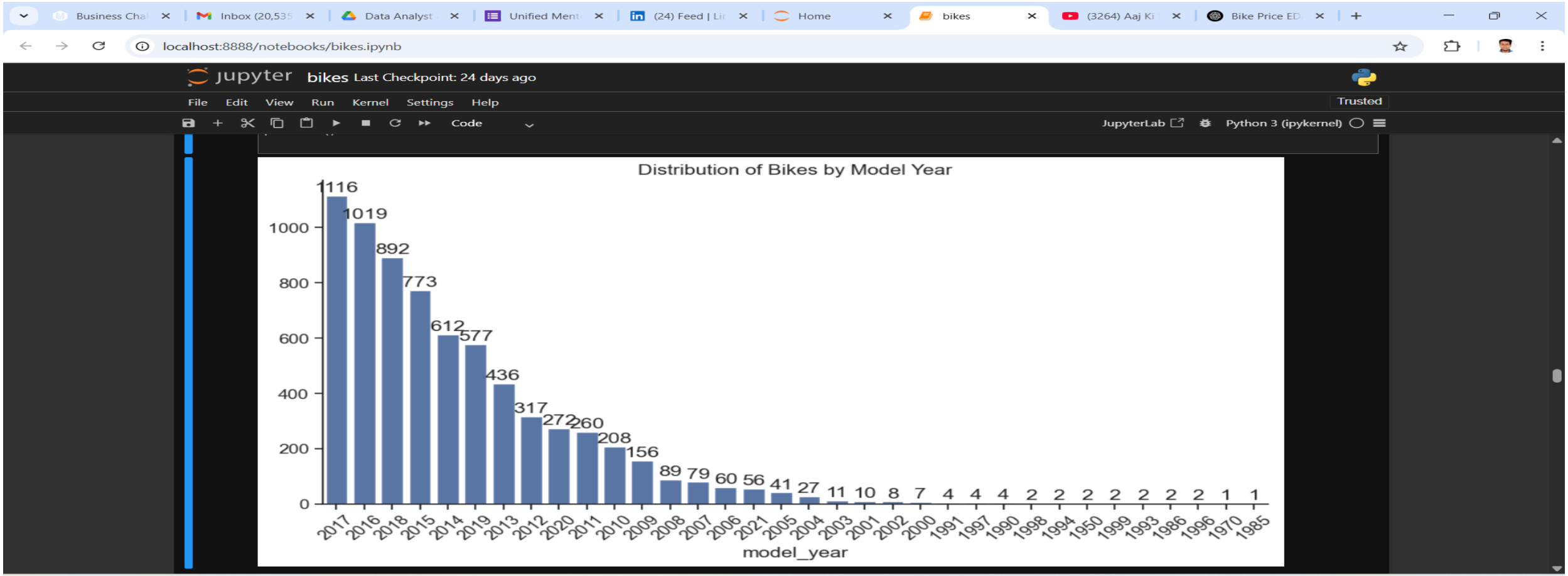
BIKE MODEL YEAR



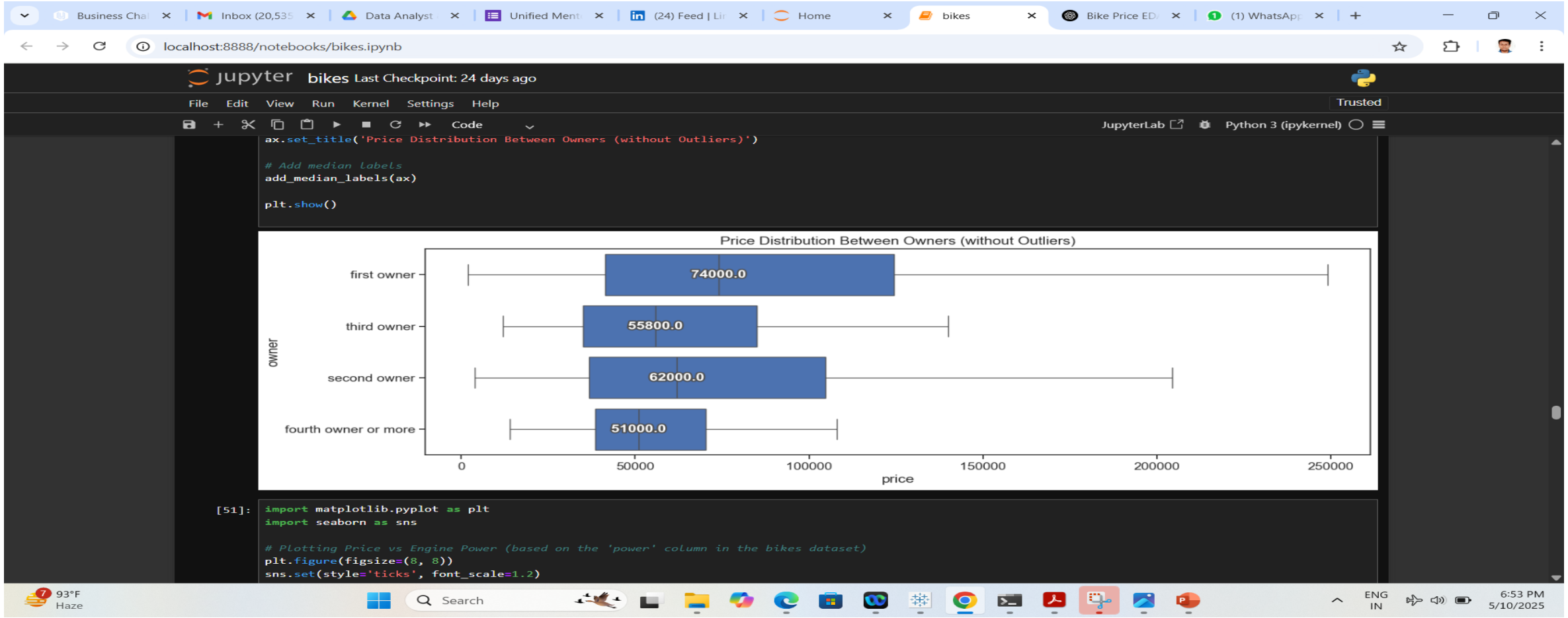
BOXPLOT OF BIKE MODEL YEARS



DISTRIBUTION OF BIKES BY MODEL YEAR

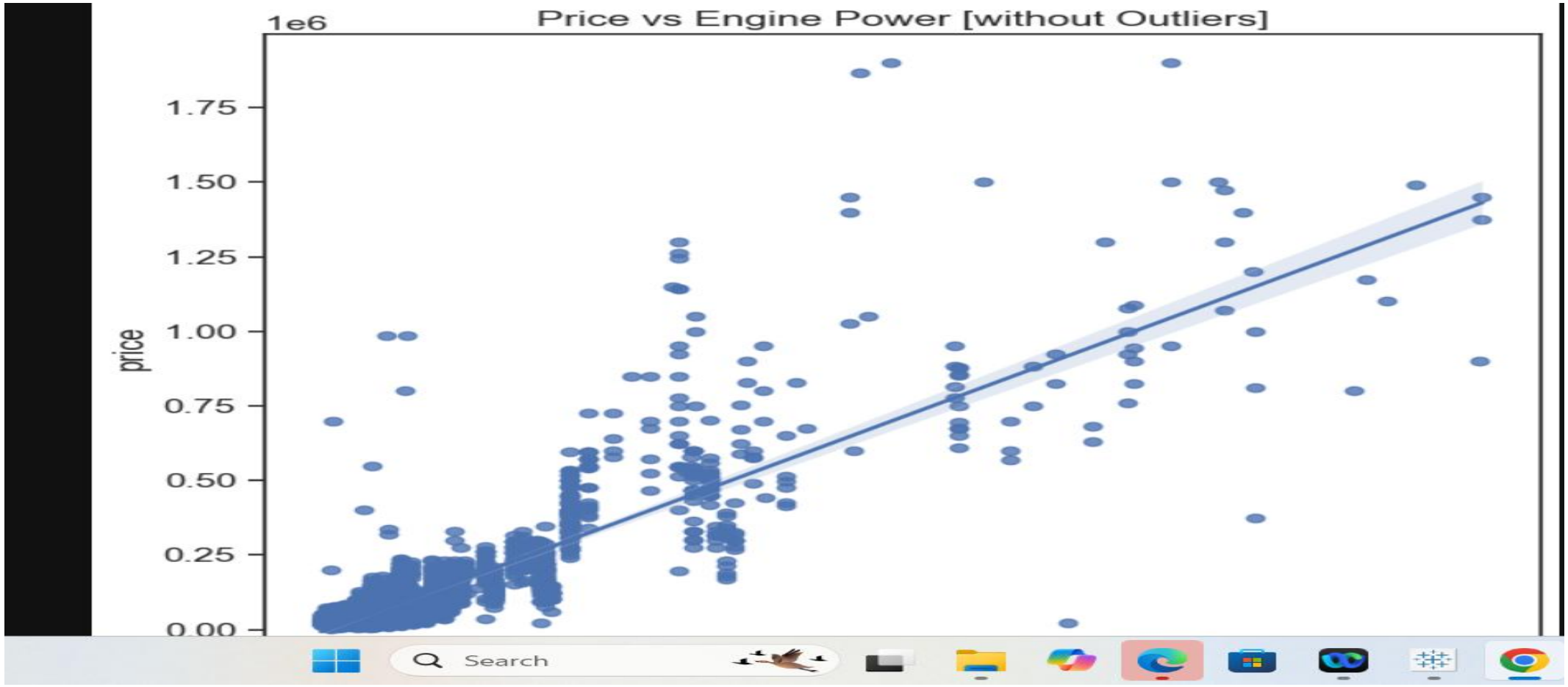


PRICE DISTRIBUTION OF BETWEEN OUTLIERS

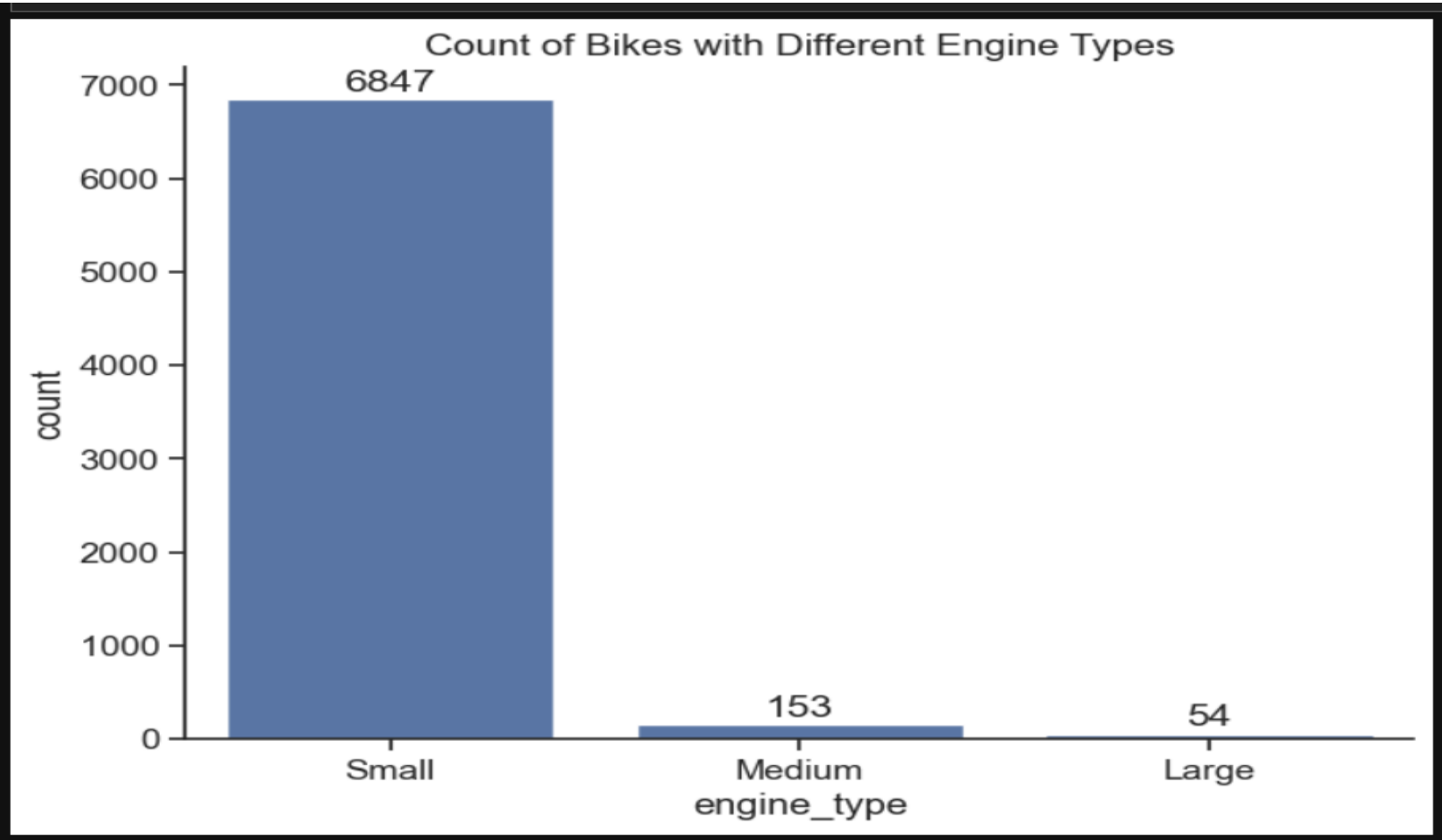


PRICE VS ENGINE POWER

The correlation between price and engine power is: 0.85

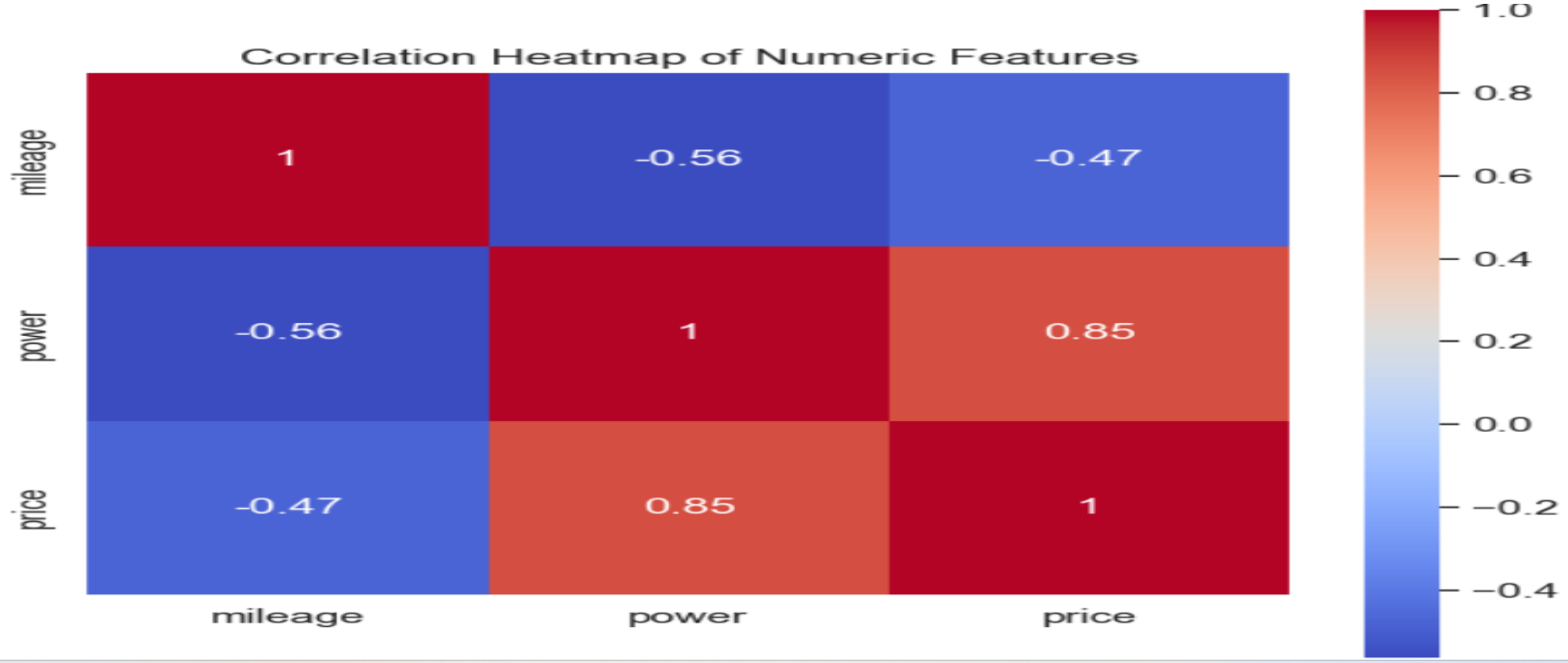


COUNT OF BIKES WITH DIFFERENT ENGINE TYPES

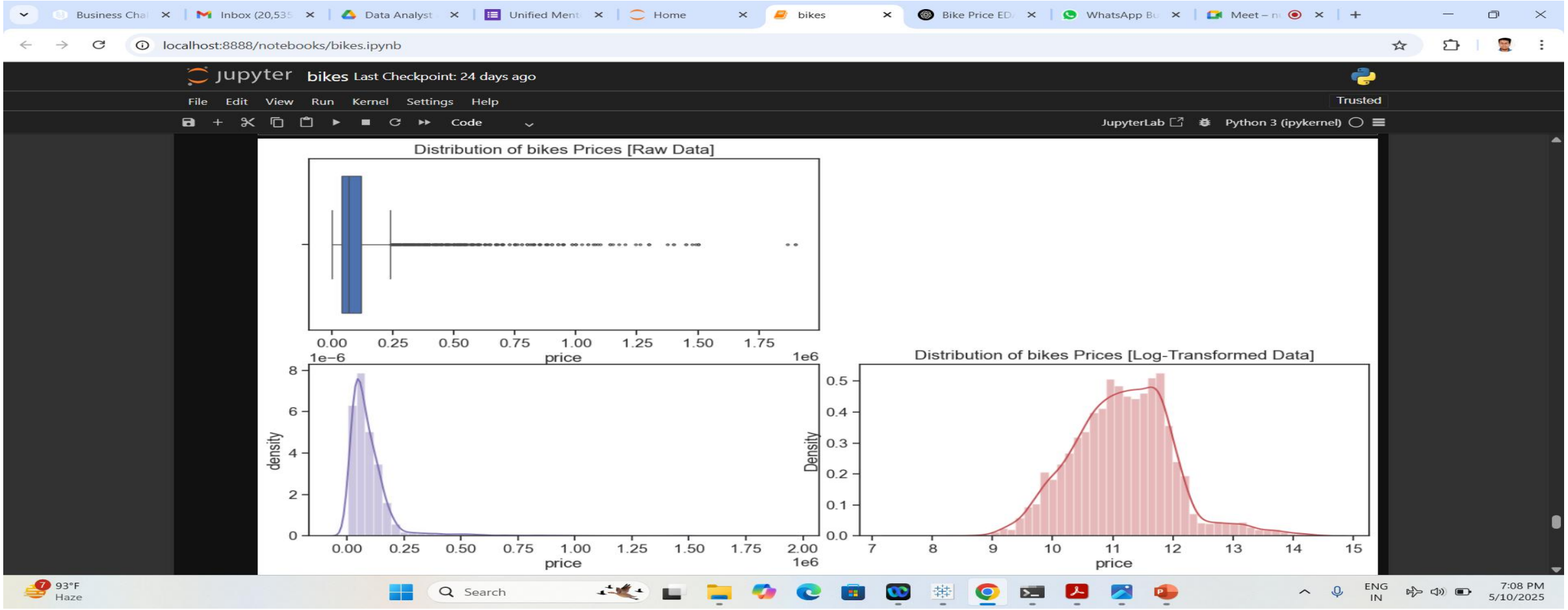



```
[56]: import seaborn as sns
```

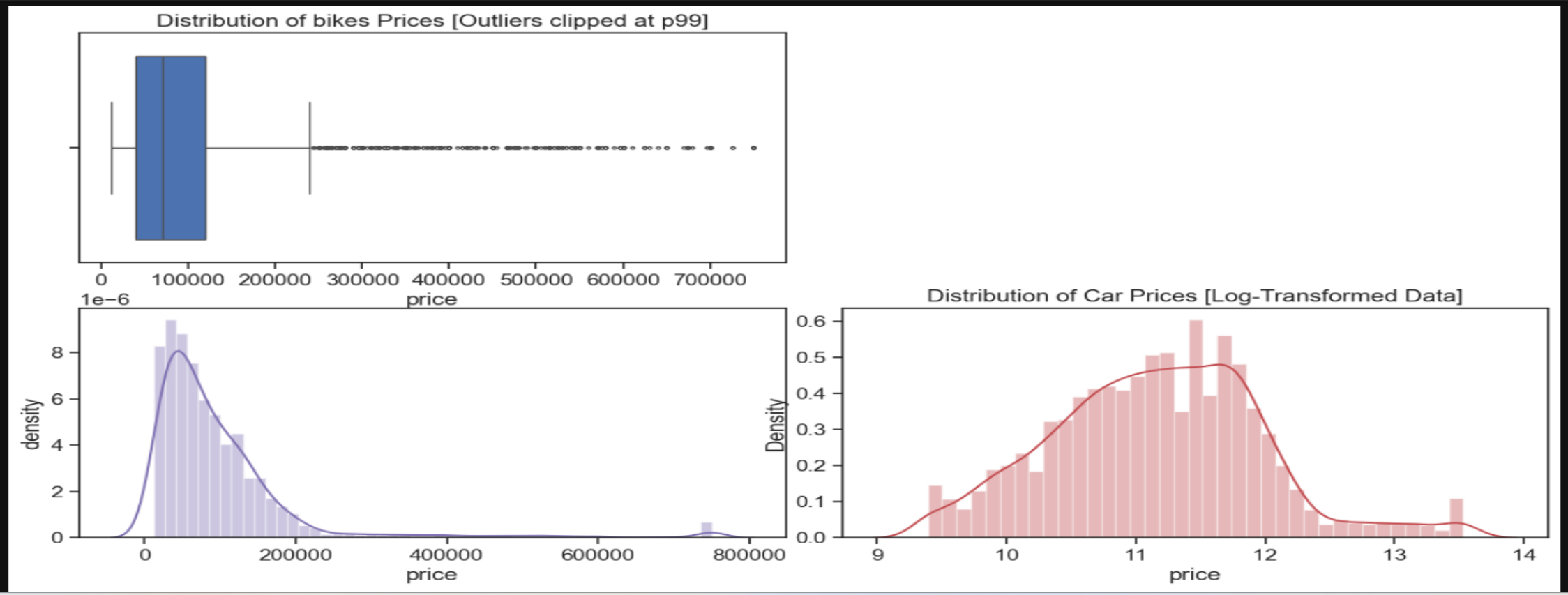
CORRELATION MATRIX



DISTRIBUTION OF BIKES PRICES



DISTRIBUTION OF BIKES PRICES



SUMMARY & INSIGHTS

- **Power, mileage, and bike age** emerged as strong predictors of price
- **Owner type** and **kilometers driven** show significant impact on pricing patterns
- **Feature engineering** (brand extraction, bike age, owner encoding) greatly enhanced data structure and model-readiness
- **Outliers and missing values** were identified and handled to ensure data integrity
- Insights provide a solid base for **building predictive pricing models**

- Majority of bikes are listed within the **₹30,000–₹1,00,000** price range
- **High power bikes** (above average BHP) tend to demand premium pricing
- **Older bikes** (age > 10 years) show a sharp depreciation in price
- **First-owner bikes** typically have higher resale value compared to second/third owners
- Popular brands and models (e.g., Royal Enfield, Honda, Bajaj) retain **better market value**
- **Missing values** in power, mileage, and cc needed careful imputation or exclusion
- **Text fields like model_name** required parsing to extract structured features (brand, engine)
- **Location-wise pricing patterns** suggest regional demand-supply influences
- Dataset was successfully transformed into a **clean, structured, and ML-ready format**