Bike Price Prediction Project – Internship Report

Name: Monu Gupta
Role: Data Analyst Intern
Tools: Python, Pandas, NumPy, Scikit-learn, Jupyter Notebook, Power BI

Table of Contents

---

1. Introduction

The goal of this project is to predict the prices of used bikes in India based on various features such as model year, kilometers driven, engine power, mileage, owner type, and location. This helps buyers and sellers make informed decisions.

---

2. Problem Statement

Predicting accurate used bike prices is challenging due to varying factors like bike age, usage, engine power, and ownership history. The project aims to build a regression model to estimate the price and visualize insights through a dashboard.

---

3. Dataset Description

```
[38]: df.head()
```

[38]:

	model_name	model_year	kms_driven	owner	location	mileage	power	price
0	Bajaj Avenger Cruise 220 2017	2017	17000 Km	first owner	hyderabad	\n\n 35 kmpl	19 bhp	63500
1	Royal Enfield Classic 350cc 2016	2016	50000 Km	first owner	hyderabad	\n\n 35 kmpl	19.80 bhp	115000
2	Hyosung GT250R 2012	2012	14795 Km	first owner	hyderabad	\n\n 30 kmpl	28 bhp	300000
3	Bajaj Dominar 400 ABS 2017	2017	Mileage 28 Kms	first owner	pondicherry	\n\n 28 Kms	34.50 bhp	100000
4	Jawa Perak 330cc 2020	2020	2000 Km	first owner	bangalore	\n\n	30 bhp	197500

Dataset Size: 7857, 8

---

4. Data Cleaning & Preprocessing

Steps performed:
- Removed duplicates and inconsistent formatting
- Converted km, mileage, and power columns to numeric
- Handled missing values
- Standardized categorical columns (owner_clean, location_clean)
- Created numerical columns for modeling (kms_driven_num, mileage_num, power_num, price_num)
- Created bins for price, mileage, and kms_driven

[41]:
```
df.head()
```

[41]:

owner_clean	location_clean	mileage_clean	power_clean	kms_driven_num	mileage_num	power_num	price_num	price_range	kms_range	mileage_bin	brand	pow
first owner	hyderabad	35 kmpl	19 bhp	17000.0	35.0	19.0	63500	30K-70K	10k-20k	30-40	bajaj	
first owner	hyderabad	35 kmpl	19.80 bhp	50000.0	35.0	19.8	115000	70K-1.5L	40k-60k	30-40	royal	
first owner	hyderabad	30 kmpl	28 bhp	14795.0	30.0	28.0	300000	1.5L-3L	10k-20k	20-30	hyosung	
first owner	pondicherry	28 kms	34.50 bhp	28.0	28.0	34.5	100000	70K-1.5L	0-10k	20-30	bajaj	
first owner	bangalore	NaN	30 bhp	2000.0	40.0	30.0	197500	1.5L-3L	0-10k	30-40	jawa	

---

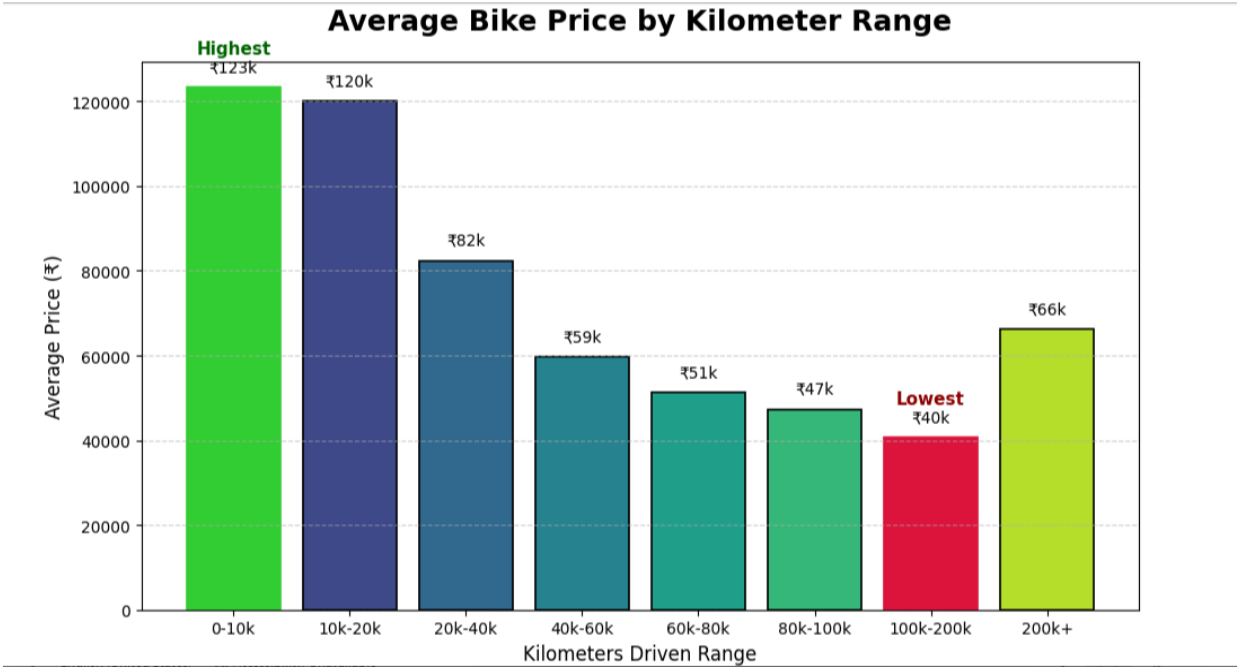
5. Exploratory Data Analysis (EDA)

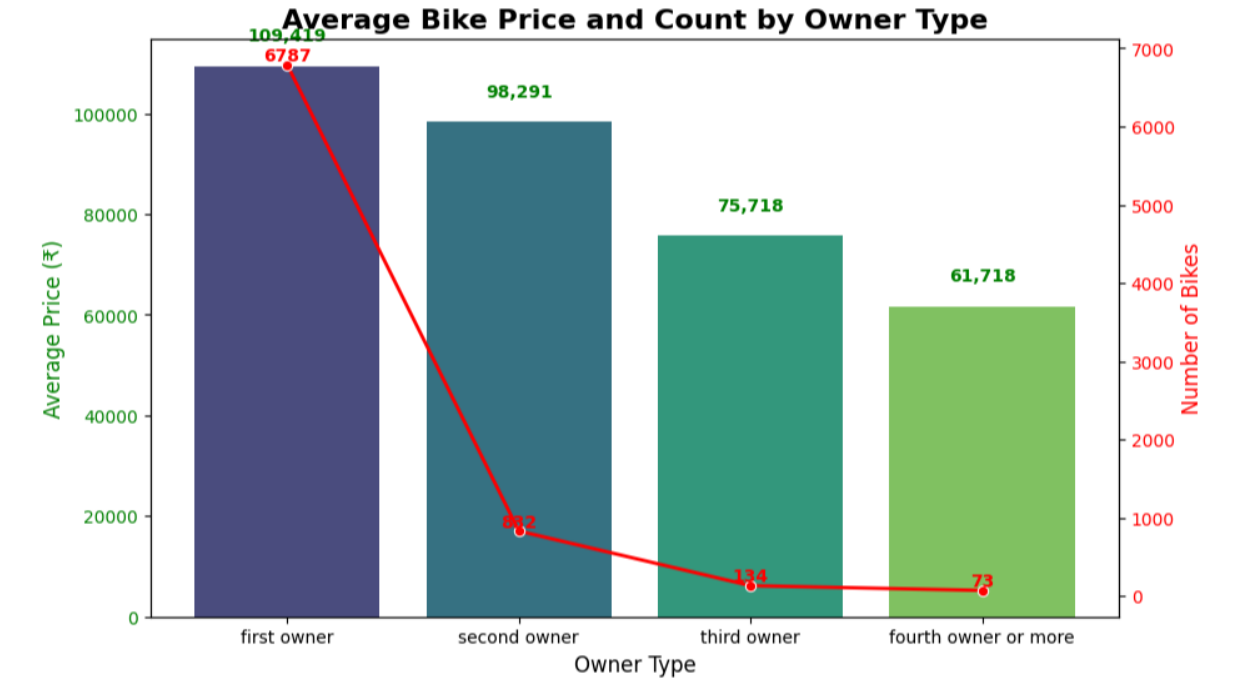
- Distribution of bikes by price, mileage, and owner

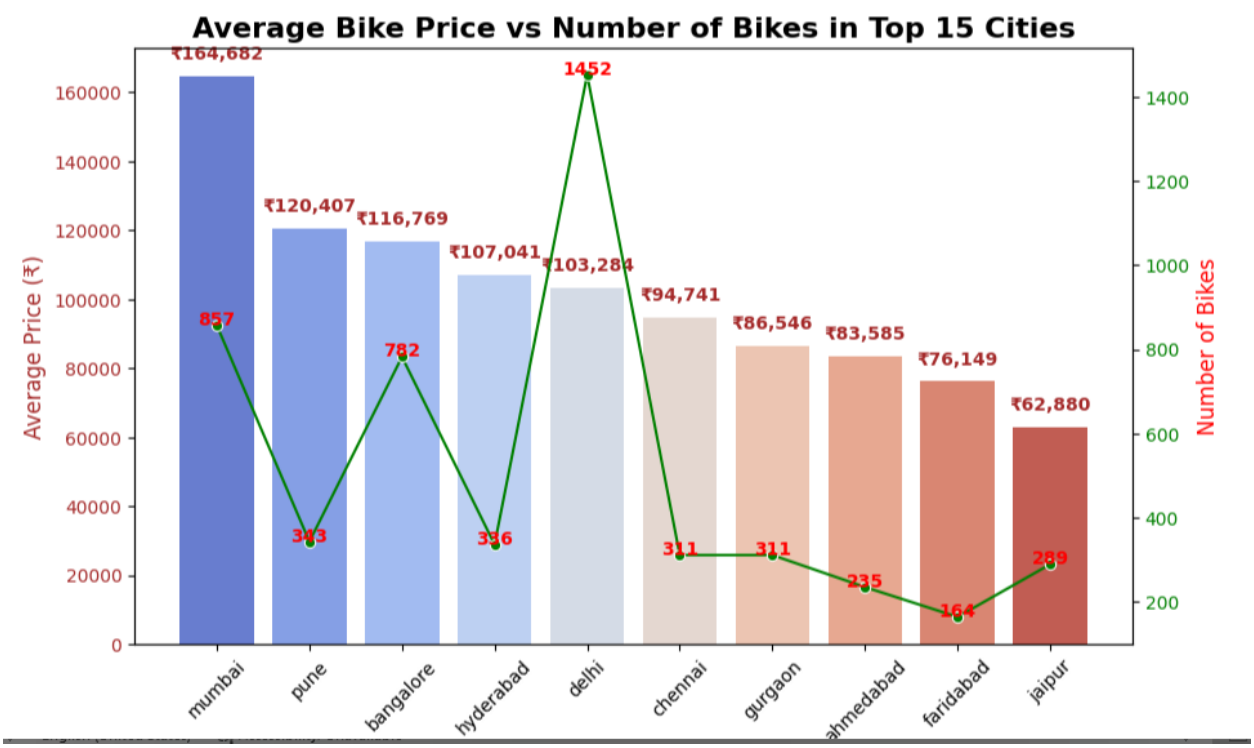
- Trends observed:
  - Older bikes cost less

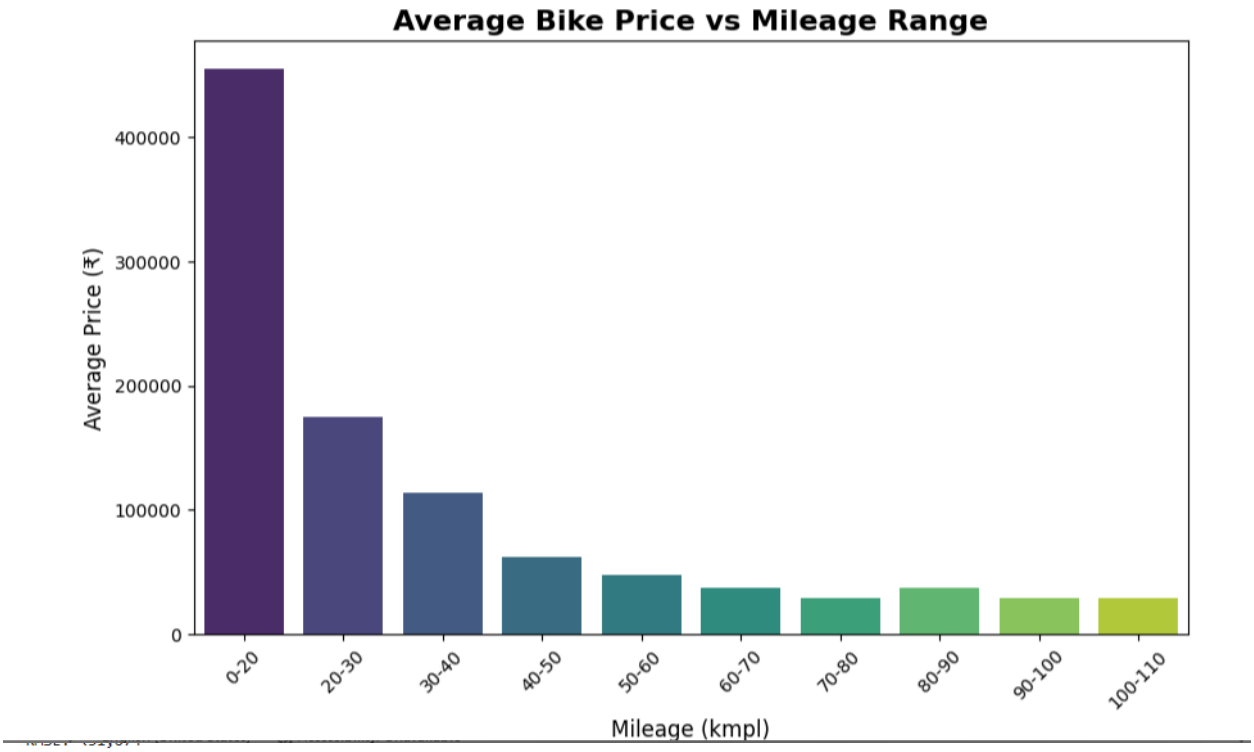
  - 1st owner bikes sell at higher prices

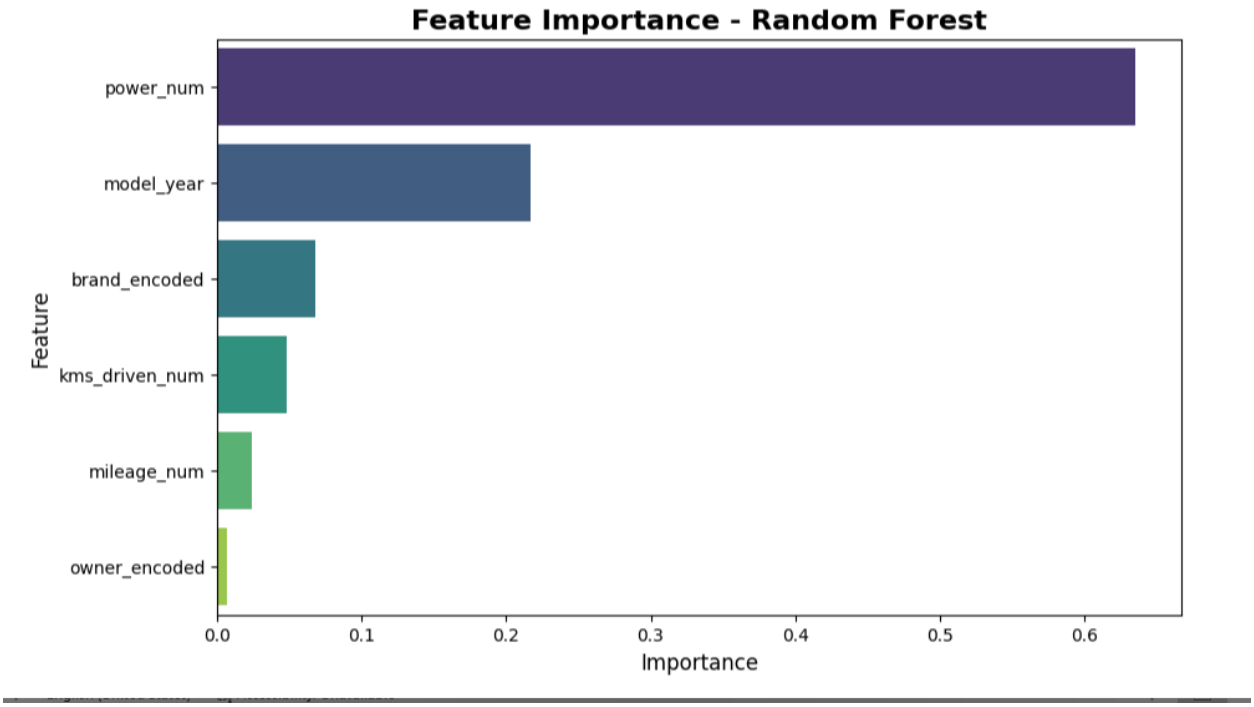
  Higher mileage correlates slightly with higher price




Average Bike Price and Count by Owner Type
109,419
98,291
75,718
61,718
6787
882
114
73
Average Price (₹)
Number of Bikes
Owner Type
first owner
second owner
third owner
fourth owner or more


Average Bike Price vs Number of Bikes in Top 15 Cities
₹164,682
₹120,407
₹116,769
₹107,041
₹103,284
₹94,741
₹86,546
₹83,585
₹76,149
₹62,880
857
343
782
336
1452
311
311
235
164
289
Average Price (₹)
Number of Bikes
mumbai
pune
bangalore
hyderabad
delhi
chennai
gurgaon
faridabad
jaipur





6. Feature Engineering

- Created derived columns: price_range, kms_range, mileage_bin
- Reason: To group values for better visualization and model performance

7. Machine Learning Model

Model Used: Random Forest Regressor

Input Features: model_year, kms_driven_num, mileage_num, power_num, owner_clean, location_clean
Target Variable: price_num

Why Random Forest: Handles non-linear relationships and performs well on tabular data

---

8. Model Evaluation

Metric	Log Scale	Original Scale
R^2 Score	0.91	-
MAE	0.14	₹11,758
RMSE	0.19	₹17,236

- **R^2 Score:** Measures model fit
- **MAE:** Average absolute error
- **RMSE:** Root mean squared error

```
Random Forest Performance (log scale):
R² Score: 0.90
MAE: 0.17
RMSE: 0.25

Random Forest Performance (original scale):
MAE: ₹14,397
RMSE: ₹31,074
```

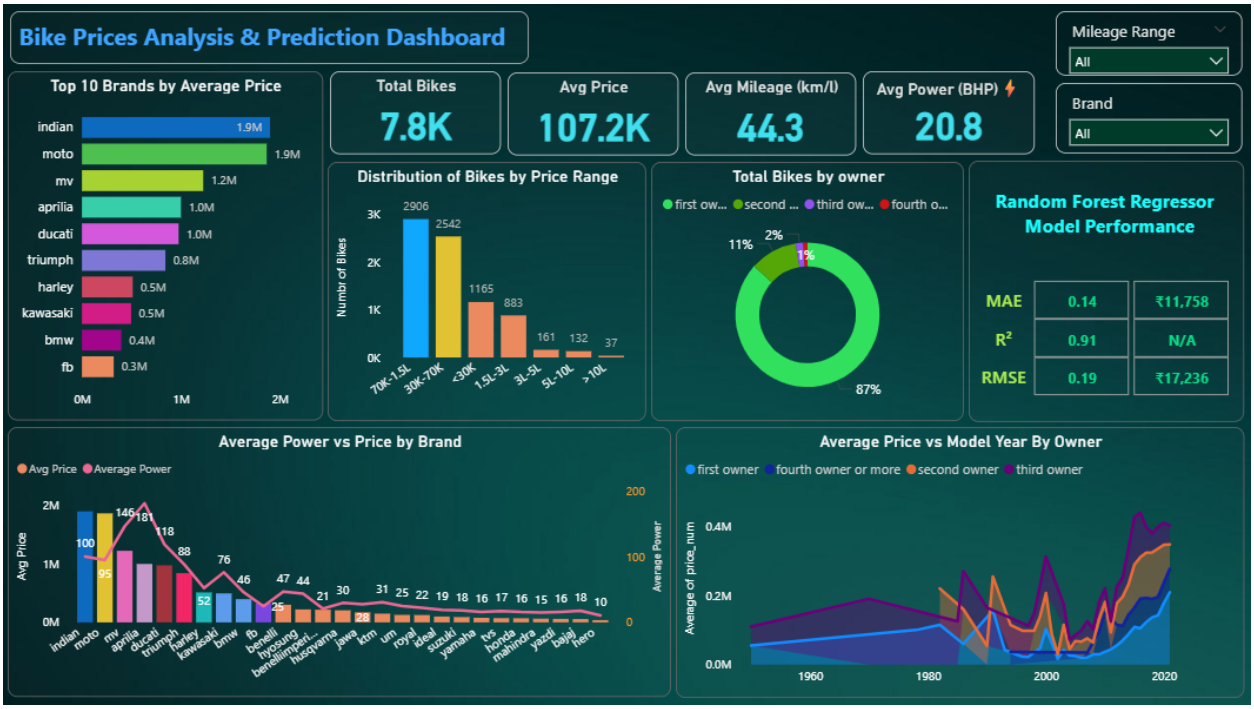
---

9. Dashboard & Visualization

- KPIs: Total bikes, Average price, Average mileage, KM driven
- Charts:
  - Price vs Model Year (Scatter)
  - Brand distribution (Bar)

- Mileage vs Price (Scatter)

- Cards showing model performance metrics



---

10. Insights & Conclusion

- 1st owner bikes cost more than 2nd or 3rd owner bikes
- Newer bikes have higher prices
- Mileage has a slight correlation with price
- Dashboard helps buyers and sellers make data-driven decisions

---

11. Tools & Technologies

- Python
- Pandas
- NumPy
- Scikit-learn

- Jupyter Notebook
- Power BI

---

12. References

- Google — *General research on used bike pricing factors and model evaluation techniques*
- YouTube — *Tutorials for Power BI dashboard design and Python data analysis*
- Scikit-learn Documentation — *Implementation details for Random Forest Regressor and performance metrics*
