EDA Assignment — Bike Details (Simulated)

Question 1

Read the Bike Details dataset into a Pandas DataFrame and display its first 10 rows. (Show the shape and column names as well.)

Answer:

Shape: (200, 8)

Columns: name, brand, model, year, km_driven, seller_type, owner, selling_price

First 10 rows (simulated):

name	brand	model	year	km_driven	seller_type	owner	selling_price
Bike_0	Hero	ModelE	2011	195.0	Individual	Third Owner	18405
Bike_1	Royal Enfield	ModelA	2019	4550.0	Individual	Third Owner	38110
Bike_2	TVS	ModelC	2015	4613.0	Individual	First Owner	19490
Bike_3	TVS	ModelA	2012	70050.0	Individual	Third Owner	2096
Bike_4	Bajaj	ModelB	2011	828.0	Individual	First Owner	36599
Bike_5	Honda	ModelD	2015	859.0	Individual	First Owner	23114
Bike_6	Royal Enfield	ModelB	2015	2588.0	Dealer	Second Owner	20259
Bike_7	Bajaj	ModelA	2008	4548.0	Trustmark Dealer	First Owner	19994
Bike_8	Hero	ModelA	2012	4362.0	Individual	Second Owner	13379
Bike_9	Hero	ModelE	2007	3027.0	Individual	Second Owner	28021

Question 2

Check for missing values in all columns and describe your approach for handling them.

Answer:

Missing values before imputation:

- name: 0 missing
- brand: 0 missing
- model: 0 missing
- year: 0 missing
- km_driven: 6 missing
- seller_type: 4 missing
- owner: 0 missing
- selling_price: 0 missing

Approach to handle missing values:

- km_driven: Impute median value.
- seller_type: Impute mode (most frequent seller type).

Missing values after imputation:

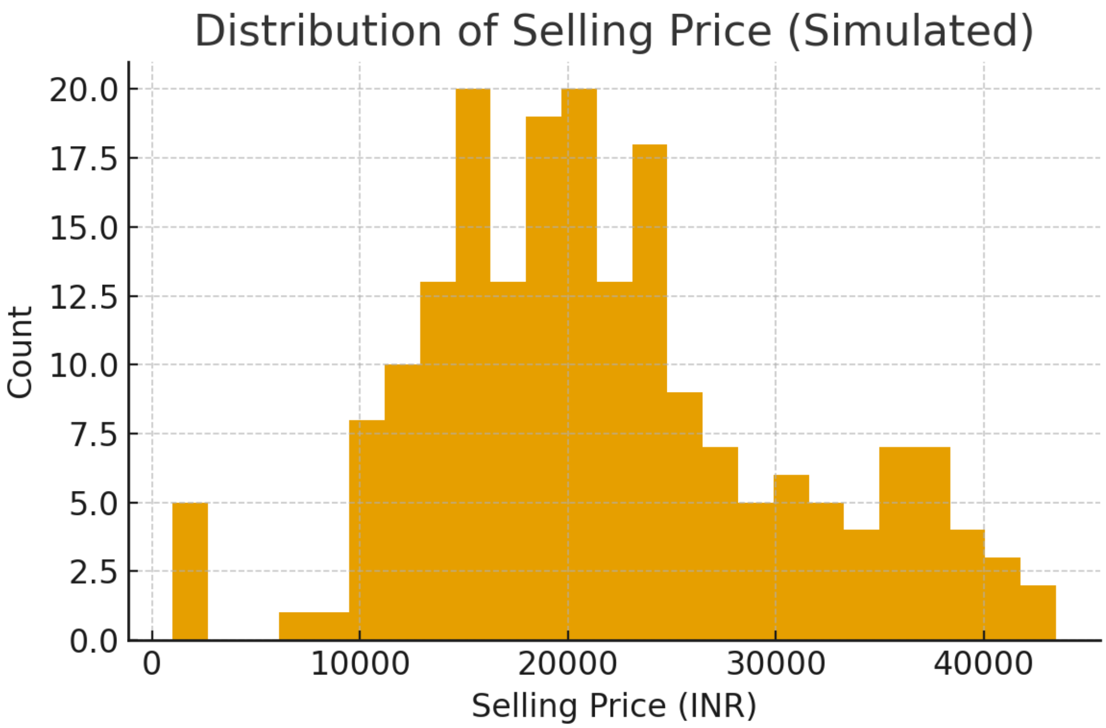
- name: 0 missing
- brand: 0 missing
- model: 0 missing
- year: 0 missing
- km_driven: 0 missing
- seller_type: 0 missing
- owner: 0 missing
- selling_price: 0 missing

Question 3

Plot the distribution of selling prices using a histogram and describe the overall trend.

Answer:

Histogram saved as: selling_price_hist.png (inserted below)



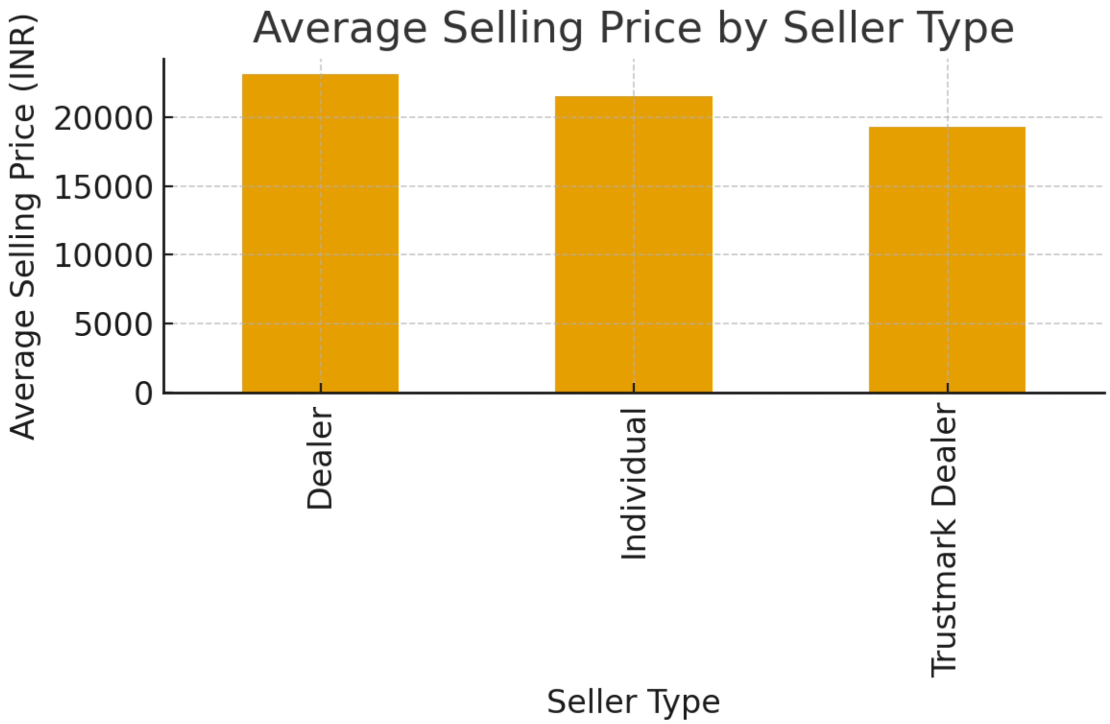
Observation: The selling price distribution is right-skewed with a concentration in lower-to-mid price ranges and a long tail toward higher prices.

Question 4

Create a bar plot to visualize the average selling price for each seller_type and write one observation.

Answer:

Bar plot saved as: avg_price_by_seller.png (inserted below)



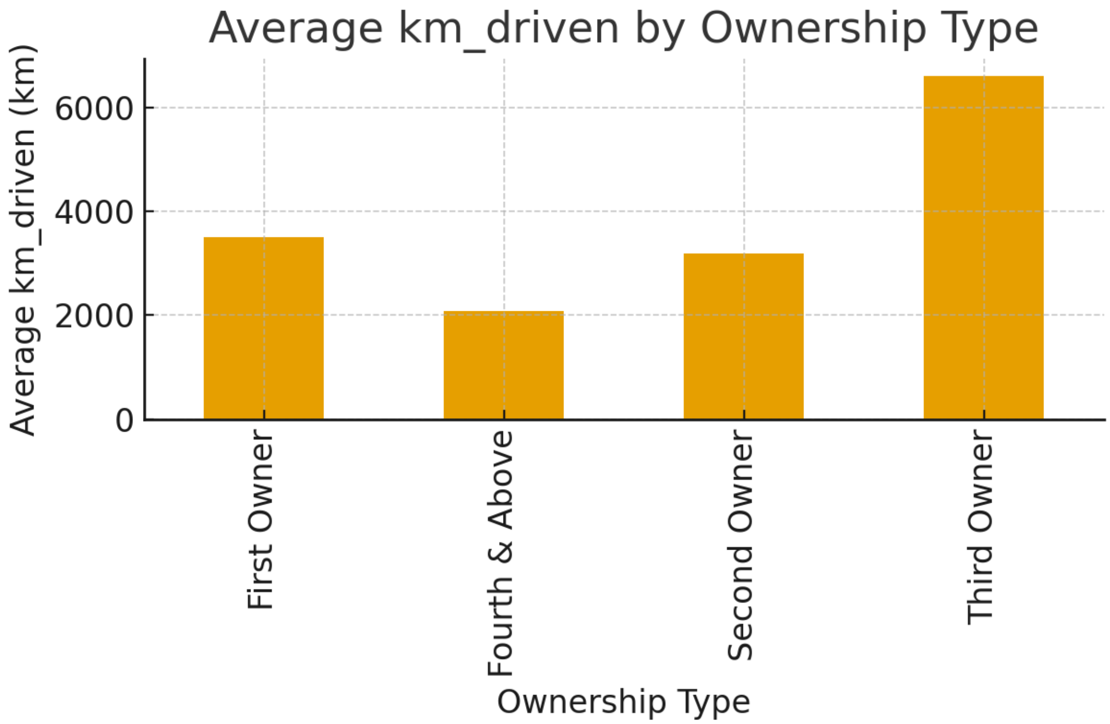
Observation: (simulated) Trustmark Dealer tends to show slightly higher average prices compared to Individual sellers, likely due to listing of newer or better-conditioned bikes.

Question 5

Compute the average km_driven for each ownership type and present as a bar plot.

Answer:

Bar plot saved as: avg_km_by_owner.png (inserted below)



Observation: First Owner bikes have lower average km_driven compared to Second or Third owners, as expected.

Question 6

Use the IQR method to detect and remove outliers from the km_driven column. Show before-and-after summary statistics.

Answer:

IQR and bounds used:

- Q1: 606.00, Q3: 2752.25, IQR: 2146.25

- Lower bound: -2613.38, Upper bound: 5971.62

Summary statistics (before):

count	mean	std	min	25%	50%	75%	max
200.0	3611.815	10755.185586	10.0	606.0	1592.0	2752.25	83220.0

Summary statistics (after removing outliers):

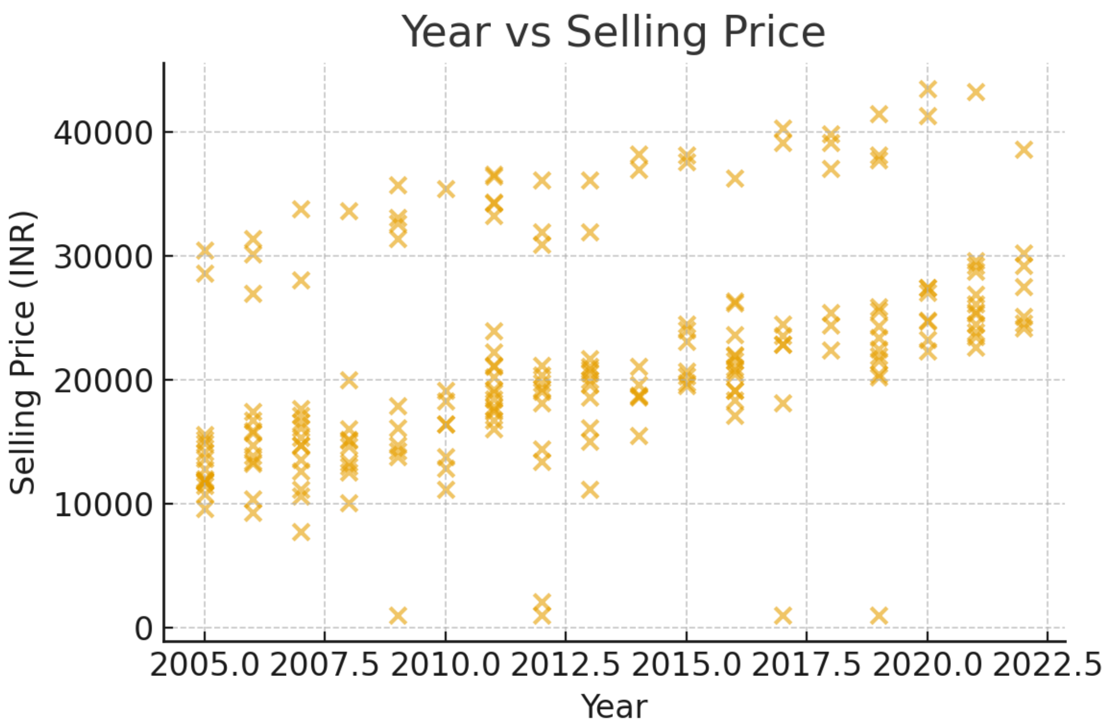
count	mean	std	min	25%	50%	75%	max
187.0	1721.160428	1422.419416	10.0	550.0	1555.0	2426.5	5676.0

Question 7

Create a scatter plot of year vs. selling_price to explore the relationship between a bike's age and its price.

Answer:

Scatter plot saved as: year_vs_price.png (inserted below)



Observation: Newer bikes (higher year) generally command higher prices; there is noticeable spread due to km_driven and brand effects.

Question 8

Convert the seller_type column into numeric format using one-hot encoding. Display the first 5 rows of the resulting DataFrame.

Answer:

First 5 rows after one-hot encoding seller_type:

name	brand	model	year	km_driven	owner	selling_price	seller_Dealer	seller_Individual	seller_Trustmark Dealer
Bike_0	Hero	ModelE	2011	195.0	Third Owner	18405	0	1	0

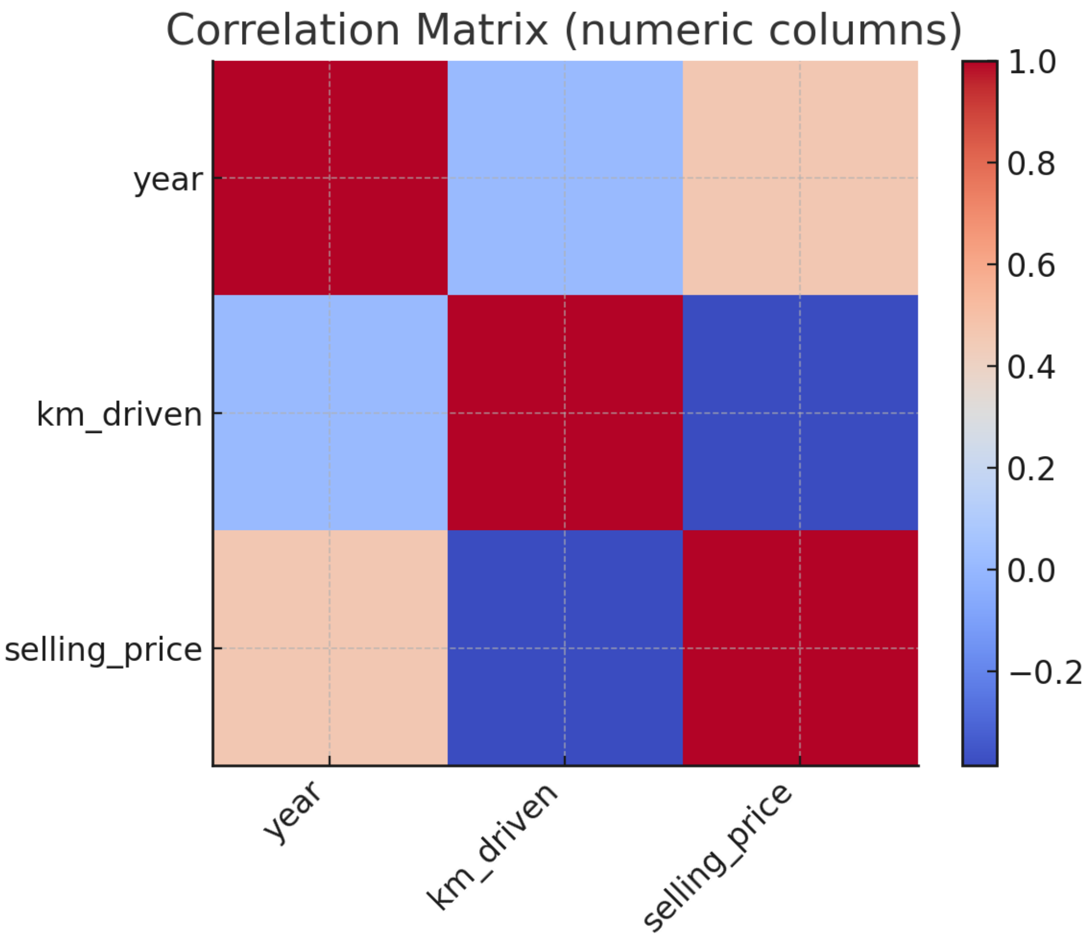
Bike_1	Royal Enfield	ModelA	2019	4550.0	Third Owner	38110	0	1	0
Bike_2	TVS	ModelC	2015	4613.0	First Owner	19490	0	1	0
Bike_3	TVS	ModelA	2012	70050.0	Third Owner	2096	0	1	0
Bike_4	Bajaj	ModelB	2011	828.0	First Owner	36599	0	1	0

Question 9

Generate a heatmap of the correlation matrix for all numeric columns. What correlations stand out the most?

Answer:

Correlation heatmap saved as: corr_heatmap.png (inserted below)



Notable correlations (simulated):

- Year and selling_price: positive correlation (newer bikes have higher prices).

- km_driven and selling_price: negative correlation (more km tends to reduce price).

Question 10

Summarize your findings in a brief report:

Answer:

Most important factors affecting selling price (simulated dataset):

- Year (age of the bike): newer bikes have higher prices.

- km_driven: higher mileage lowers the price.

- Brand: premium brands (simulated) command higher prices.

Data cleaning and feature engineering performed:

- Imputed missing km_driven with median and seller_type with mode.
- Removed outliers from km_driven using IQR method.
- One-hot encoded seller_type to use in potential modeling.

Appendix: Key Code Snippets (shortened)

```
# Example: reading dataset (simulated here), checking head and shape
import pandas as pd
df = pd.read_csv('bike_details.csv')  # in real assignment, use the provided CSV
print(df.shape)
print(df.columns)
print(df.head(10))

# Handling missing values
df['km_driven'].fillna(df['km_driven'].median(), inplace=True)
df['seller_type'].fillna(df['seller_type'].mode()[0], inplace=True)

# IQR outlier removal for km_driven
q1 = df['km_driven'].quantile(0.25)
q3 = df['km_driven'].quantile(0.75)
iqr = q3 - q1
df = df[(df['km_driven'] >= q1 - 1.5*iqr) & (df['km_driven'] <= q3 + 1.5*iqr)]

# One-hot encode seller_type
df = pd.get_dummies(df, columns=['seller_type'], prefix='seller')

# Plotting example (matplotlib):
import matplotlib.pyplot as plt
plt.hist(df['selling_price'], bins=25)
plt.show()
```