INNOMATICS RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Used Cars Analysis With Cars24.com

Web Scraping with EDA

Presented By : Raj Shivade

About Me:

Background: I have completed my B.Tech in Data Science and Engineering from G H Raisoni College of Engineering and Management.

Why I want to learn Data Science: I'm passionate about solving real-world problems using data-driven insights. With a strong foundation in Data Science, I find data science an exciting intersection of statistics, programming, and domain knowledge. It allows me to uncover patterns, predict outcomes, and support decision-making processes

LinkedIn Profile: www.linkedin.com/in/raj-shivade25
GitHub Profile: https://github.com/RajShivade

Problem Statement :

- The demand for used cars in India is steadily increasing due to affordability and rising car ownership aspirations. However, platforms like **Cars24.com**, while rich in vehicle listings, do not provide structured, downloadable data for market analysis. As a result, it becomes difficult for buyers, sellers, and dealerships to understand pricing trends, the impact of vehicle usage, and other key factors that influence resale value.

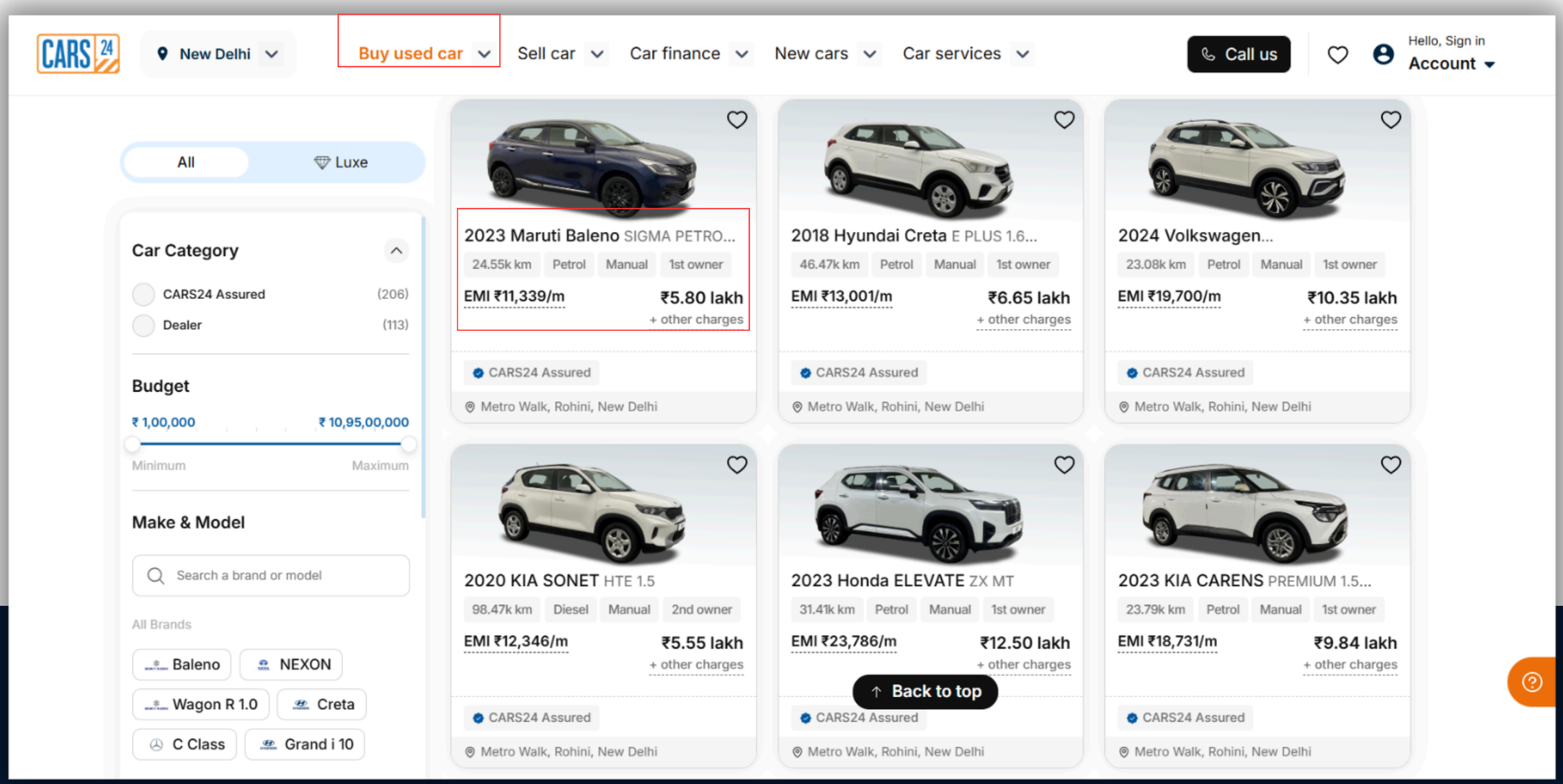
- This project solves this problem by scraping data from Cars24.com and performing detailed **Exploratory Data Analysis (EDA)**. The aim is to extract patterns from attributes like car model, ownership status, EMI options, kilometers driven, manufacturing year, fuel type, transmission type, and price to help users make better, data-driven decisions.

Objective of the Analysis:

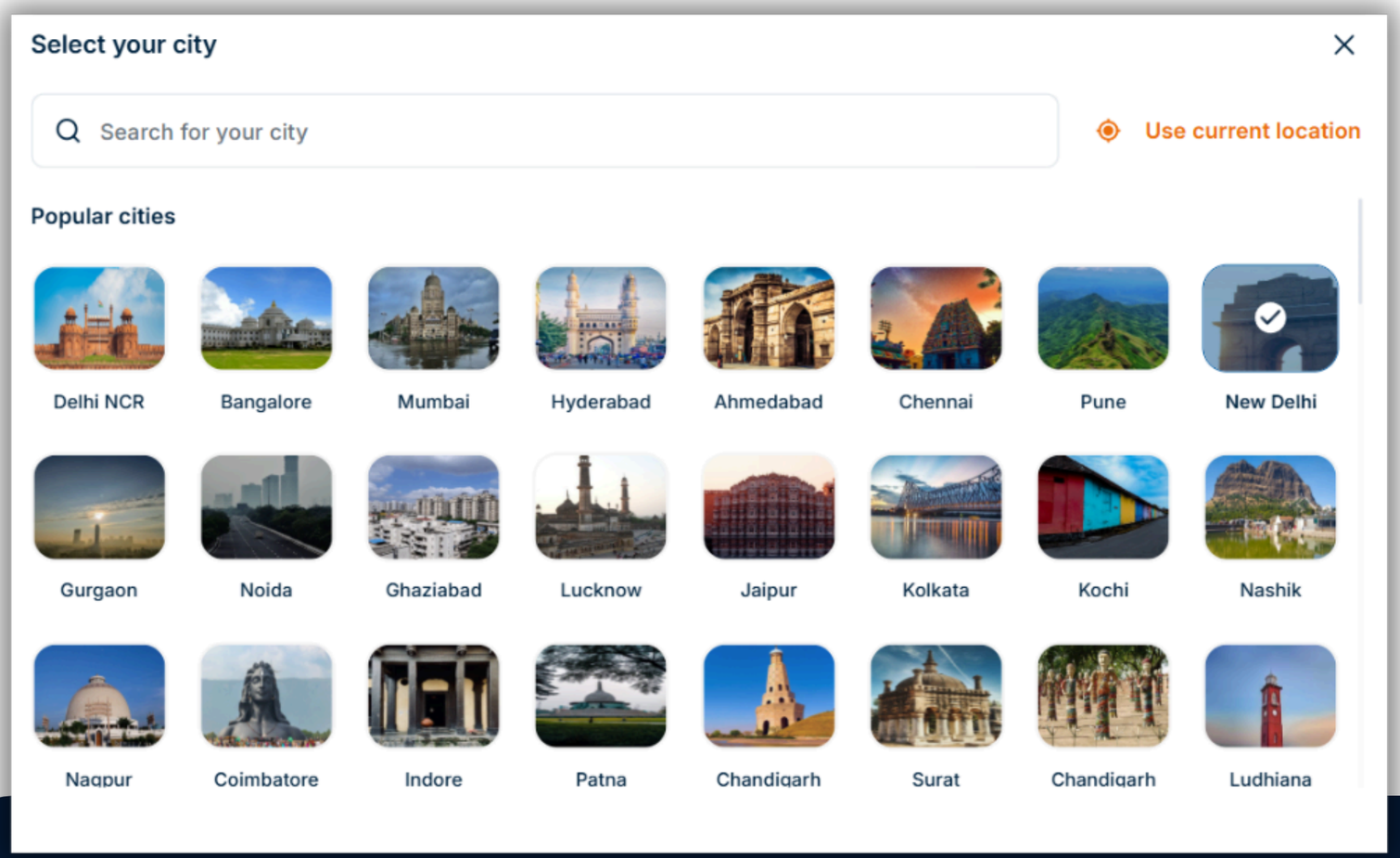
- **Web Scraping:**
  - Extract real-time used car listings from Cars24.com including:
    - Brand, Model, Year, Fuel Type, Transmission, Owner, KM Driven, EMI Price, Price.
- **Data Cleaning & Preparation:**
  - Remove special characters (e.g., “₹”, “km”, commas) and convert columns to proper formats (e.g., integers, floats).
  - Handle missing or inconsistent values.
  - Parse and standardize ownership data (e.g., First Owner, Second Owner).
- **Exploratory Data Analysis (EDA):**
  - Understand how each feature (like ownership, year, or transmission) impacts pricing.
  - Visualize relationships between:
    - Year & Price
    - KM Driven & Price
    - Fuel Type & Price
    - Owner Type & Price
    - Transmission & Price
    - EMI Price vs. Market Price
- **Insight Generation:**
  - Derive meaningful takeaways to inform buying strategies.

INNOMATICS RESEARCH LABS

DATA COLLECTION (SCRAPING PROCESS)



Web Scraping & EDA: From Raw Data to Insights



Extracted Data using Web Scraping & Regex: Collected data from the cars24 website using BeautifulSoup and cleaned it using regular expressions(regex) to fetch structured information.

Analyzed & Visualized Data: Performed EDA using pandas, matplotlib, and seaborn to uncover patterns and display insights through visualizations.

INNOMATICS RESEARCH LABS

Initial Scraped Dataset Overview :

Categorical Columns

	Brand	Model	Fuel_Type	Transmission	KM_Driven	Owner	EMI_Price	Price
0	Tata	Bolt	Petrol	Manual	99.79k km	1st owner	EMI ₹5,539/m	₹2.49 lakh
1	Nissan	MAGNITE	Petrol	Manual	22.49k km	1st owner	EMI ₹13,099/m	₹6.70 lakh
2	Maruti	Celerio	Petrol	Auto	1.2L km	1st owner	EMI ₹8,103/m	₹3.64 lakh
3	Hyundai	NEW	Petrol	Manual	44.81k km	1st owner	EMI ₹7,859/m	₹4.18L
4	Maruti	Celerio	Petrol	Auto	71.63k km	1st owner	EMI ₹8,579/m	₹3.26 lakh
5	Hyundai	i20	Petrol	Manual	72.54k km	1st owner	EMI ₹9,595/m	₹4.91 lakh
6	Honda	Amaze	Petrol	Manual	25.31k km	1st owner	EMI ₹12,094/m	₹6.19 lakh
7	Renault	Kwid	Petrol	Manual	020.62k km	2nd owner	EMI ₹7,661/m	₹3.92 lakh
8	Hyundai	Grand	Petrol	Auto	65.57k km	1st owner	EMI ₹9,037/m	₹4.06 lakh
9	MG	ASTOR	Petrol	Manual	12.92k km	1st owner	EMI ₹15,543/m	₹8.17 lakh
10	Maruti	Baleno	Petrol	Manual	213.05k km	1st owner	EMI ₹10,614/m	₹5.43 lakh
11	Maruti	Ertiga	Diesel	Manual	71.39k km	1st owner	EMI ₹17,630/m	₹6.69 lakh
12	Hyundai	Elite	Petrol	Manual	38.10k km	2nd owner	EMI ₹11,401/m	₹5.83 lakh
13	Maruti	Wagon	CNG	Manual	1.2L km	1st owner	EMI ₹11,346/m	₹3.42 lakh
14	Hyundai	Verna	Petrol	Manual	48.75k km	1st owner	EMI ₹12,828/m	₹6.56 lakh
15	Honda	WR	Petrol	Manual	40.07k km	1st owner	EMI ₹14,954/m	₹7.65 lakh
16	Honda	Amaze	Petrol	Manual	49.35k km	1st owner	EMI ₹9,801/m	₹5.01 lakh
17	Maruti	Celerio	Petrol	Auto	48.96k km	2nd owner	EMI ₹8,504/m	₹4.35 lakh

Numerical Columns

	Year
0	2015
1	2023
2	2017
3	2020
4	2014
5	2016
6	2019
7	2022
8	2015
9	2022
10	2017
11	2015
12	2018
13	2015
14	2017
15	2021
16	2018
17	2018
18	2019

Preprocessed Data Ready for Analysis :

Categorical Columns

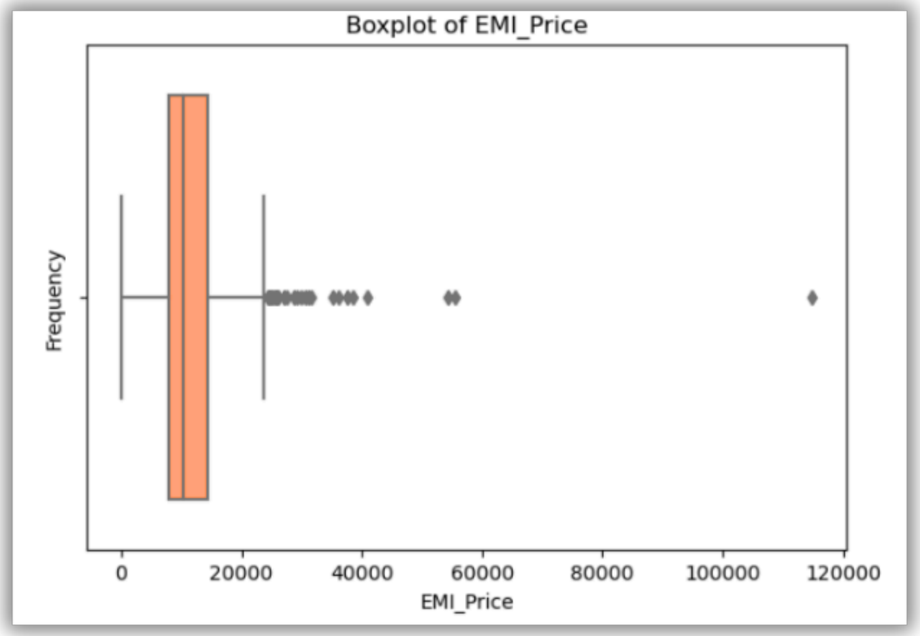
	Brand	Model	Fuel_Type	Transmission	Owner
0	tata	bolt	petrol	manual	1st owner
1	nissan	magnite	petrol	manual	1st owner
2	maruti	celerio	petrol	auto	1st owner
3	hyundai	new	petrol	manual	1st owner
4	maruti	celerio	petrol	auto	1st owner
5	hyundai	i20	petrol	manual	1st owner
6	honda	amaze	petrol	manual	1st owner
7	renault	kwid	petrol	manual	2nd owner
8	hyundai	grand	petrol	auto	1st owner
9	mg	astor	petrol	manual	1st owner
10	maruti	baleno	petrol	manual	1st owner
11	maruti	ertiga	diesel	manual	1st owner
12	hyundai	elite	petrol	manual	2nd owner
13	maruti	wagon	cng	manual	1st owner
14	hyundai	verna	petrol	manual	1st owner
15	honda	wr	petrol	manual	1st owner

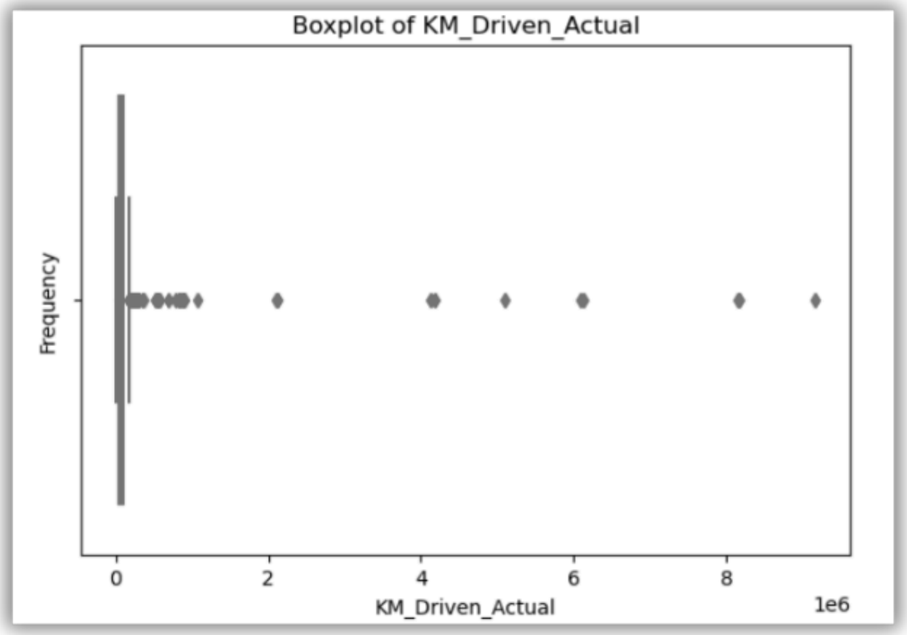
Numerical Columns

	Year	EMI_Price	Price	KM_Driven_Actual
0	2015	5539.0	249000	99790
1	2023	13099.0	670000	22490
2	2017	8103.0	364000	120000
3	2020	7859.0	418000	44810
4	2014	8579.0	326000	71630
5	2016	9595.0	491000	72540
6	2019	12094.0	619000	25310
7	2022	7661.0	392000	20620
8	2015	9037.0	405999	65570
9	2022	15543.0	817000	12920
10	2017	10614.0	543000	213050
11	2015	17630.0	669000	71390
12	2018	11401.0	583000	38100
13	2015	11346.0	342000	120000
14	2017	12828.0	656000	48750
15	2021	14954.0	765000	40070

Outlier Detection & Distribution Spread :

Before Outlier Handling:

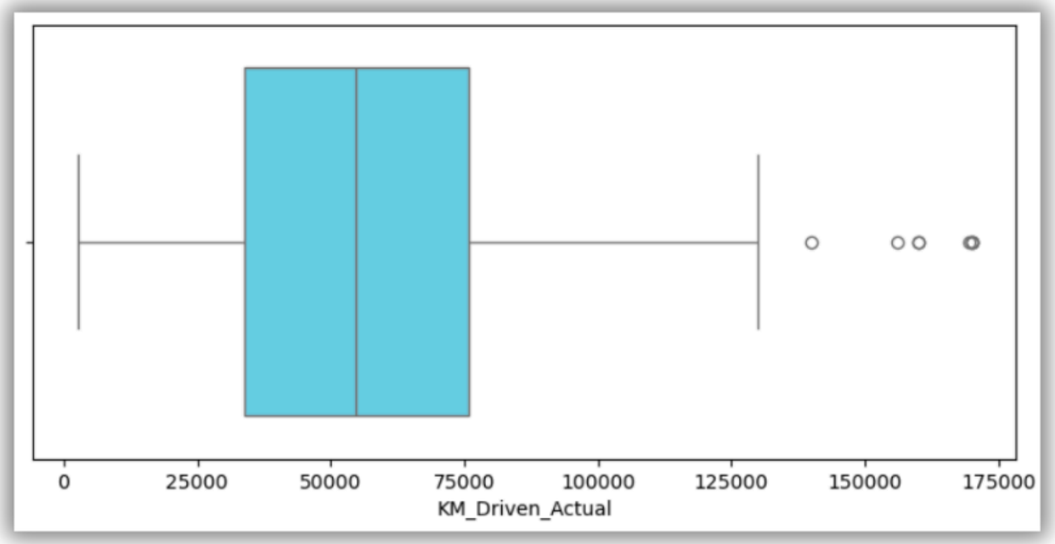


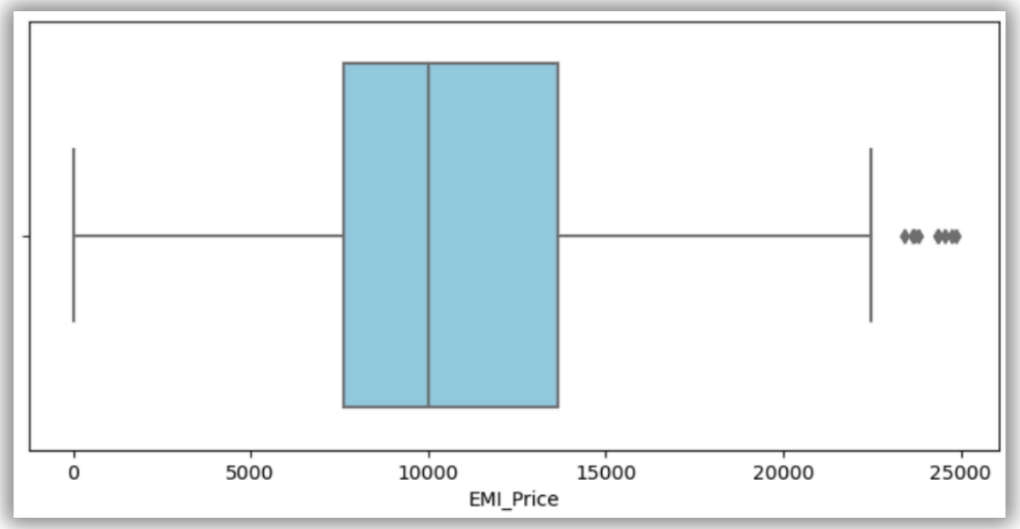


Outlier handling improves data quality and model accuracy.
Helps reduce skew, noise, and overfitting.
Indicates buyer preference for mid-range EMI and moderately used cars — useful for pricing, marketing, and loan structuring.

INNOMATICS RESEARCH LABS

After Handling Outliers :



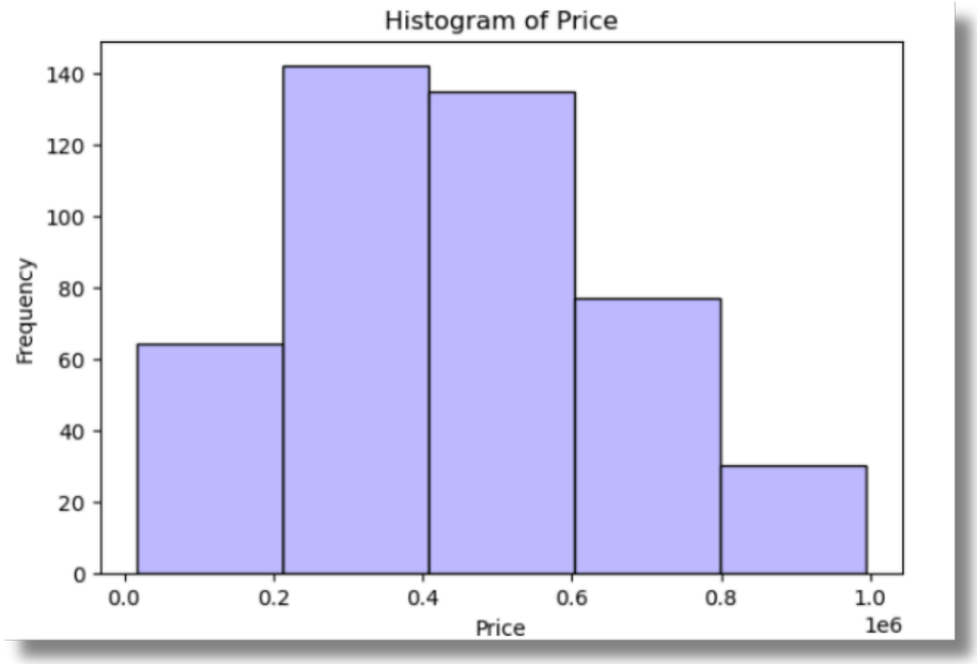


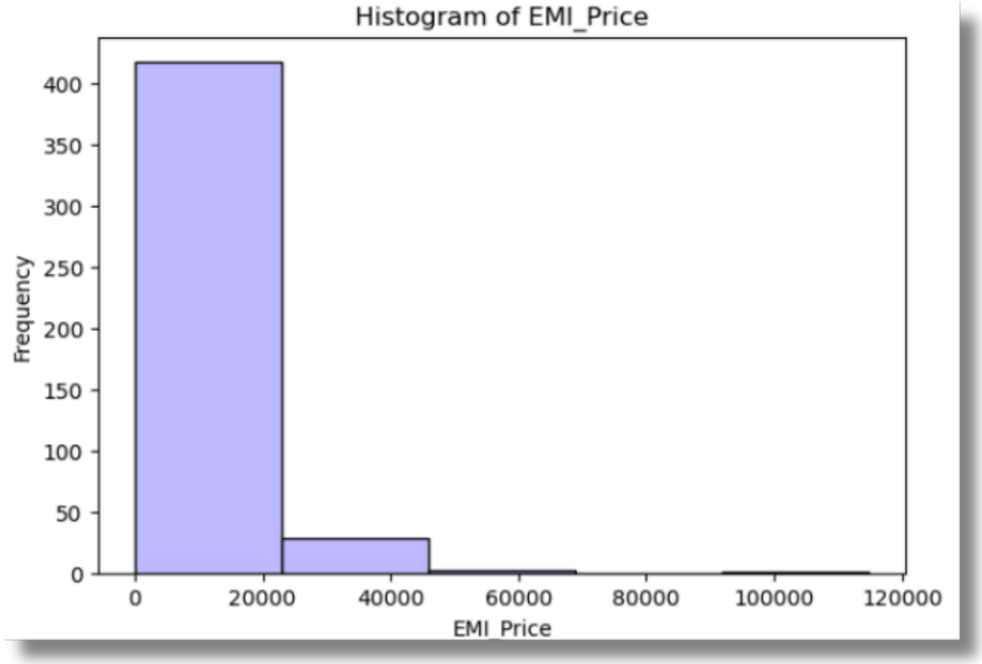
KM_Driven_Actual:
Most vehicles are driven between ~10,000 and ~125,000 km.
Median is slightly above 50,000 km.
Few outliers beyond 150,000 km (high usage cars)

EMI_Price:
Most EMI values lie between ₹2,500 and ₹20,000.
Median EMI is ~₹10,000.
Outliers above ₹20,000 (likely high-end cars).

INNOMATICS
RESEARCH LABS

Univariate Graphs On Numerical Columns :



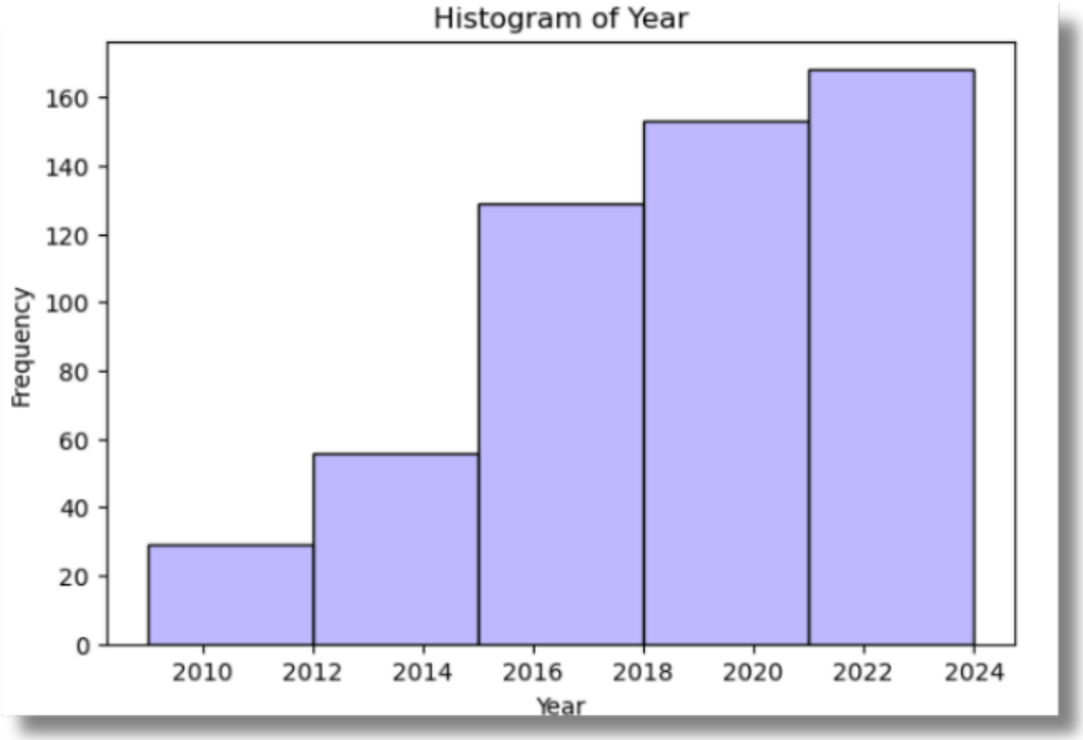


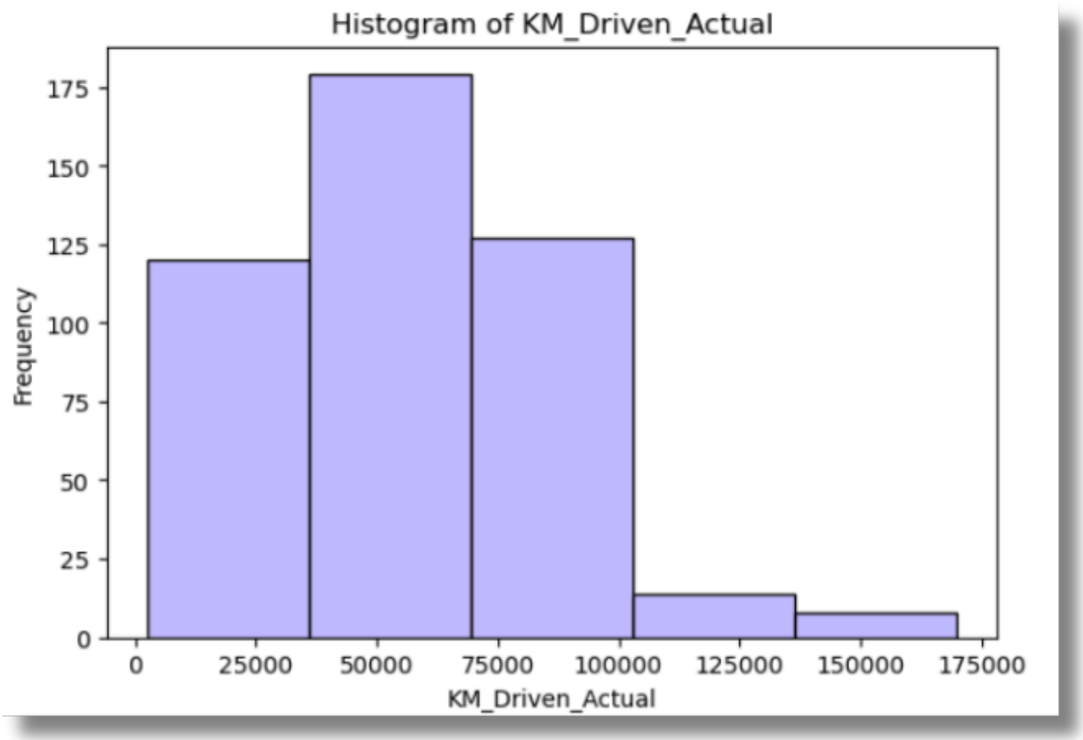
Price :
Most vehicles are priced between ₹2,00,000 and ₹6,00,000.
Fewer vehicles are in the ₹8,00,000 to ₹10,00,000 range.
Distribution is slightly left-skewed — high-priced vehicles are less common.

EMI_Price :
Majority of EMI values are below ₹20,000.
Very few vehicles have EMI over ₹40,000.
Distribution is highly right-skewed, indicating most buyers prefer lower EMI options.

INNOMATICS RESEARCH LABS

Univariate Graphs On Numerical Columns :



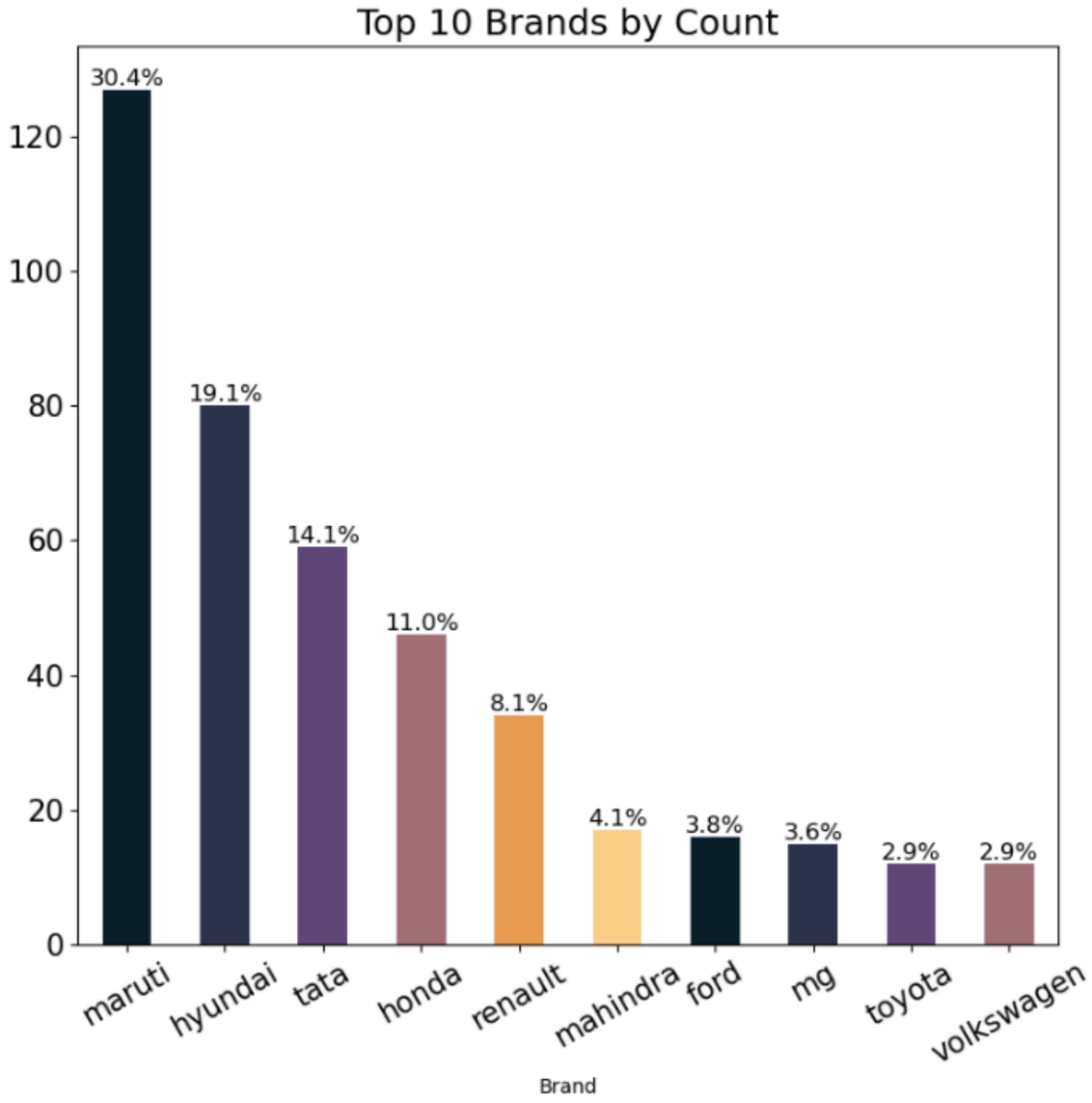


Year :
Majority of vehicles are from 2018–2023.
Very few vehicles older than 2014.
Dataset is dominated by newer vehicles.

KM_Driven_Actual:
Most vehicles have run 25,000–75,000 km.
Few vehicles exceed 100,000 km.
Distribution is right-skewed, showing more low-mileage cars.

INNOMATICS RESEARCH LABS

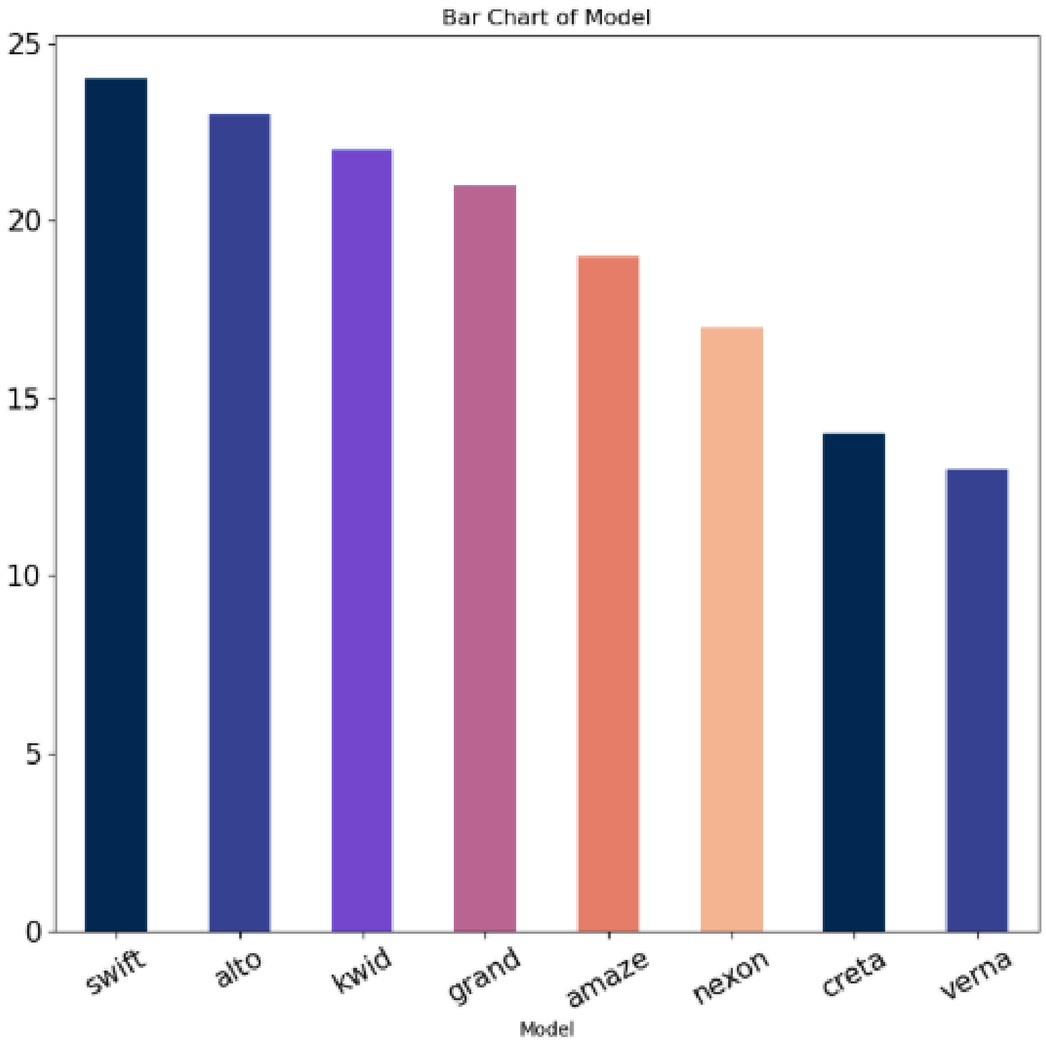
Univariate Analysis – Categorical Columns :



- **Maruti (30.4%), Hyundai (19.1%), Tata (14.1%), and Honda (11.0%) are the top 4 brands, making up ~75% of the dataset.**
- **Other brands like Renault, Mahindra, Ford, MG, Toyota, and Volkswagen have less representation (<10% each).**
- **The data is skewed toward popular, budget-friendly brands, which suggests:**
- **The model will perform best for these top brands.**
- **For rare brands, results might be less reliable due to fewer examples.**
- **Useful for targeting inventory and marketing toward high-volume brands.**

INNOMATICS RESEARCH LABS

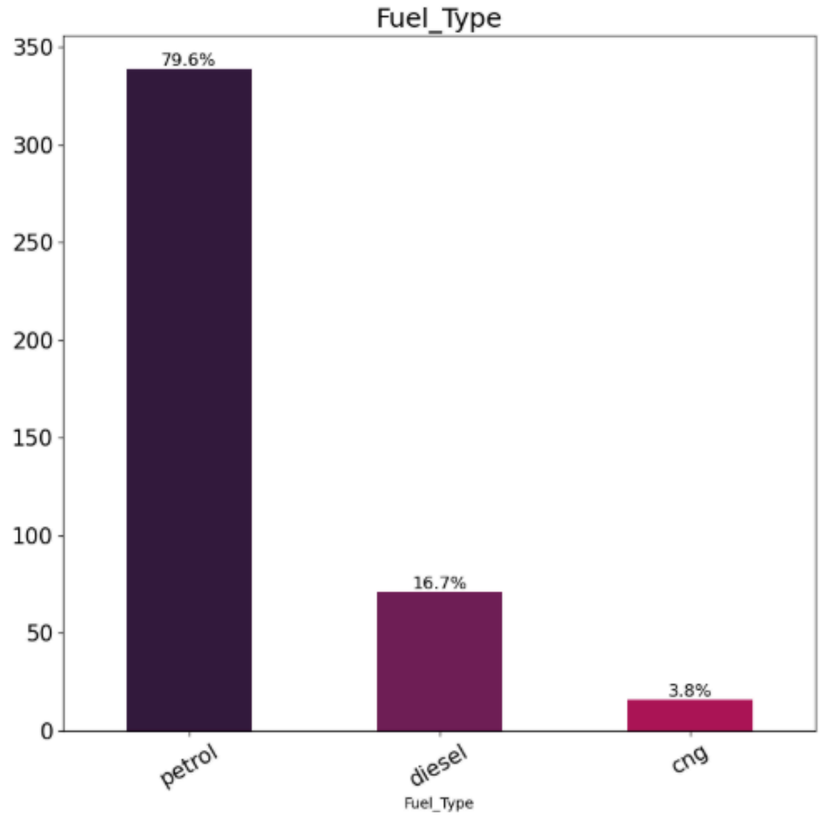
Categorical Univariate Analysis – Car Models:

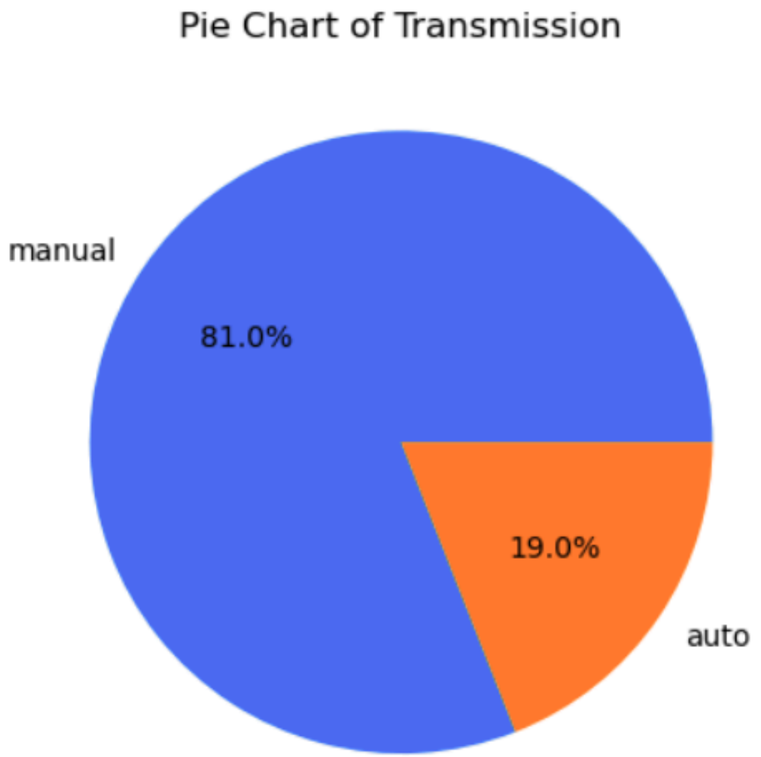


- **Top models in the dataset are Swift, Alto, and Kwid, each with over 20 entries.**
- **Other frequently listed models include Grand, Amaze, Nexon, Creta, and Verna.**
- **The dataset is dominated by entry-level and mid-range models, reflecting consumer preference for affordable and compact cars.**
- **This distribution helps:**
- **Build better predictive models for popular models like Swift and Alto.**
- **Focus marketing, pricing, and loan strategies on high-demand models.**
- **Understand that less data exists for premium or less common models, which may impact model generalizability for those.**

INNOMATICS RESEARCH LABS

Categorical Univariate Analysis:

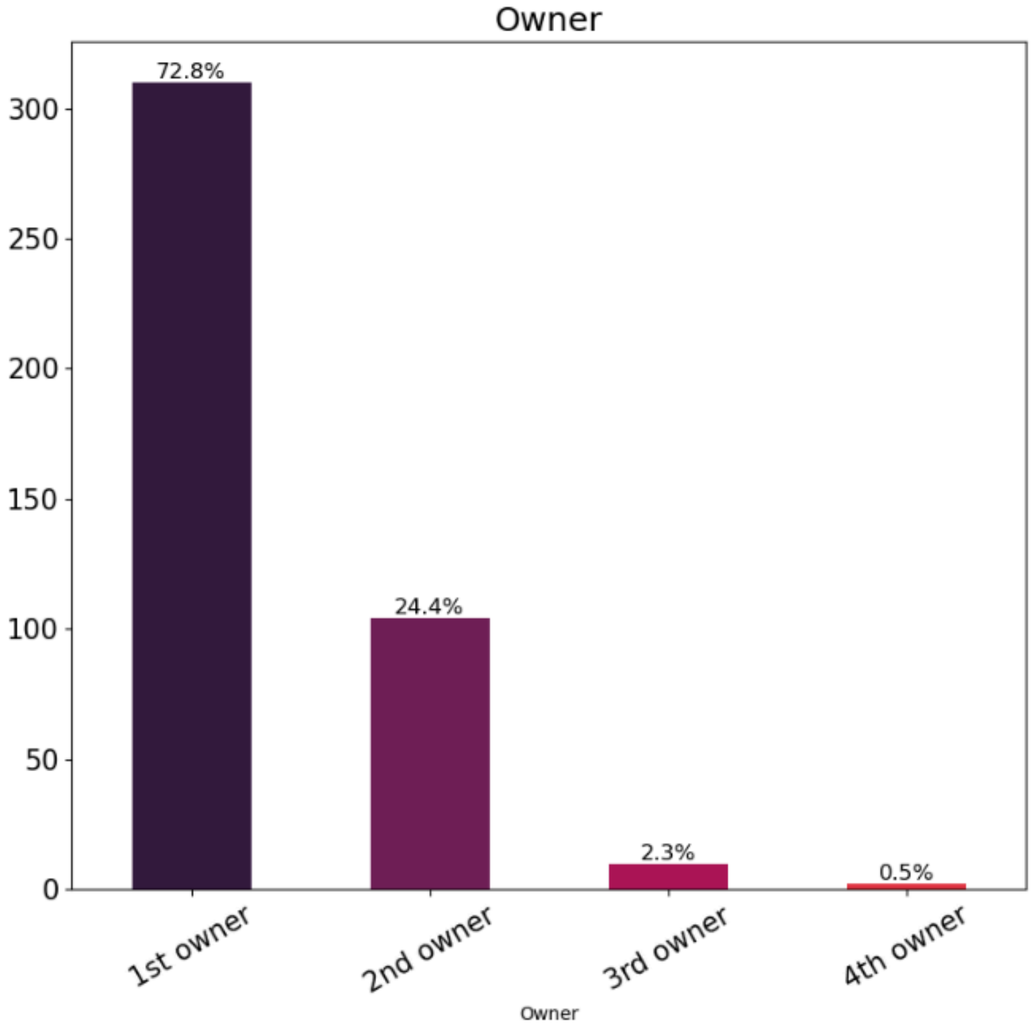




Fuel Type :
Petrol vehicles dominate the dataset at 79.6%.
Diesel makes up 16.7%, and CNG is just 3.8%.
Indicates a strong preference for petrol cars, which may reflect market trends or inventory focus.

Transmission Type:
Manual transmission is the most common at 81%.
Automatic vehicles make up only 19%.
Suggests that manual cars are more available or more in demand, likely due to affordability or customer preference.

Categorical Univariate Analysis :



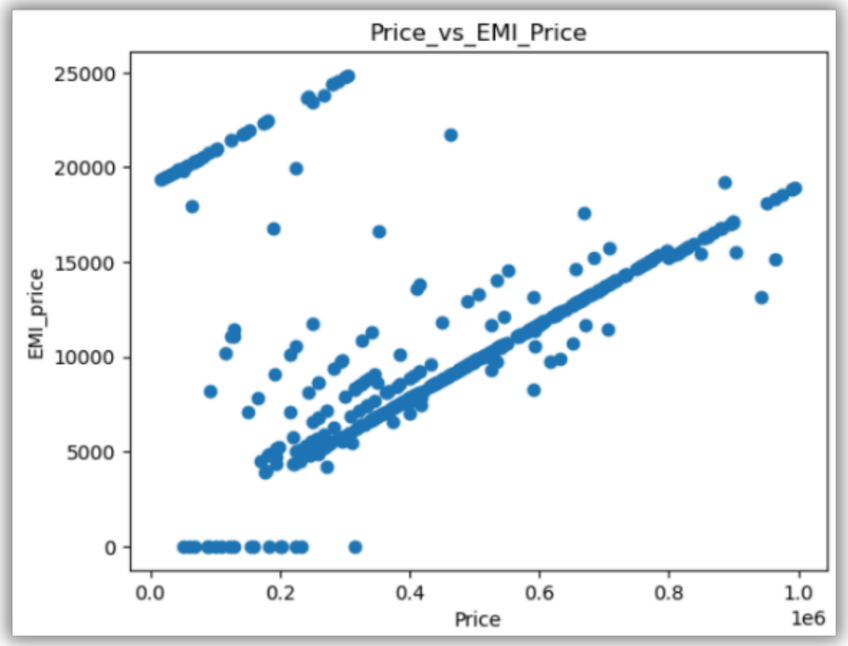
- **1st Owner cars dominate the dataset at 72.8%.**
- **2nd Owner (24.4%).**
- **Very few listings are for 3rd Owners (2.3%)**
- **4th Owners (0.5%).**

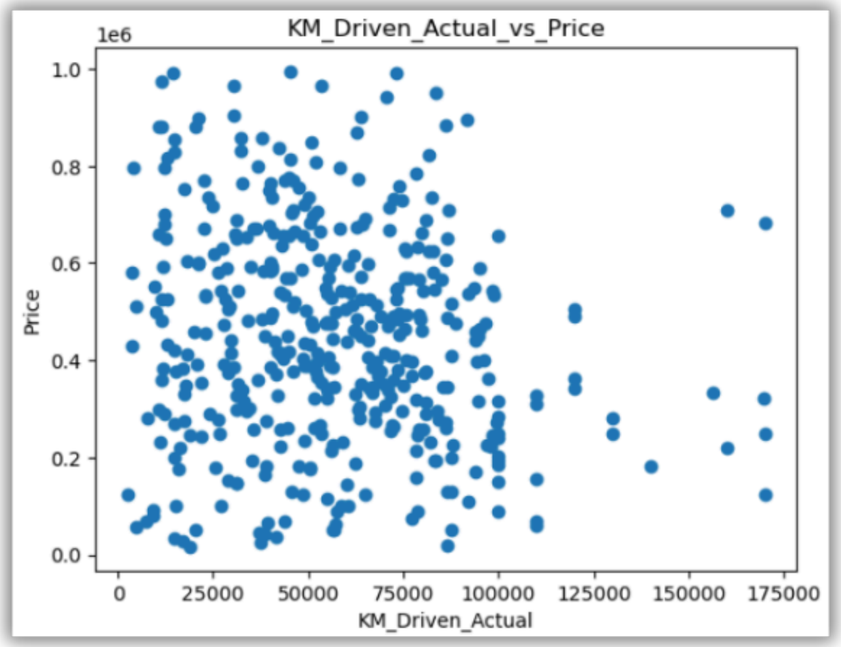
- **Majority of cars are lightly used (1st or 2nd owner), making the dataset ideal for predicting resale value or EMI for newer second-hand cars.**
- **Useful for targeting customers who prefer low-owner vehicles, a common buyer preference in used car markets.**

Bivariate Analysis :

- Here We Select 2 columns to perform our Analysis.
- Understand the type of relationship
- **Numerical - Numerical**
- a. You can plot graphs like scatterplot(regression plots), 2D histplot, 2D KDEplots
- b. Check the correlation coefficient to check the linear relationship
- **Numerical - Categorical**
- create visualizations that compare the distribution of the numerical data across different categories of the categorical data.
- a. You can plot graphs like barplot, boxplot, kdeplot,violinplot even scatterplots
- **Categorical - Categorical**
- a. You can create cross-tabulations or contingency tables that show the distribution of values in one categorical column, grouped by the values in the other categorical column.
- b. You can plots like heatmap, stacked barplots, treemaps

Numerical vs Numerical :



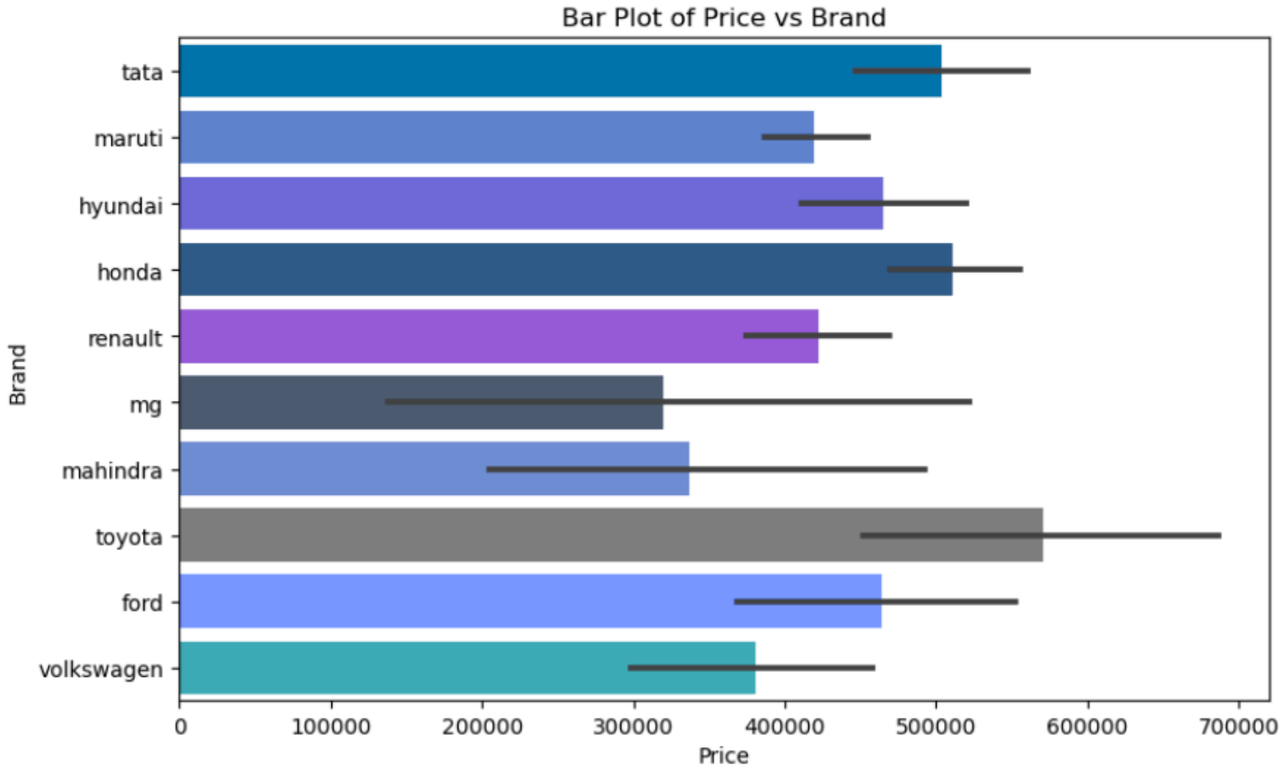


Price vs EMI_Price:
Shows a strong positive linear relationship.
As Price increases, EMI_Price also increases consistently.
Indicates that EMI is directly influenced by vehicle price, which is expected.
Useful for predicting EMI based on price with high confidence.

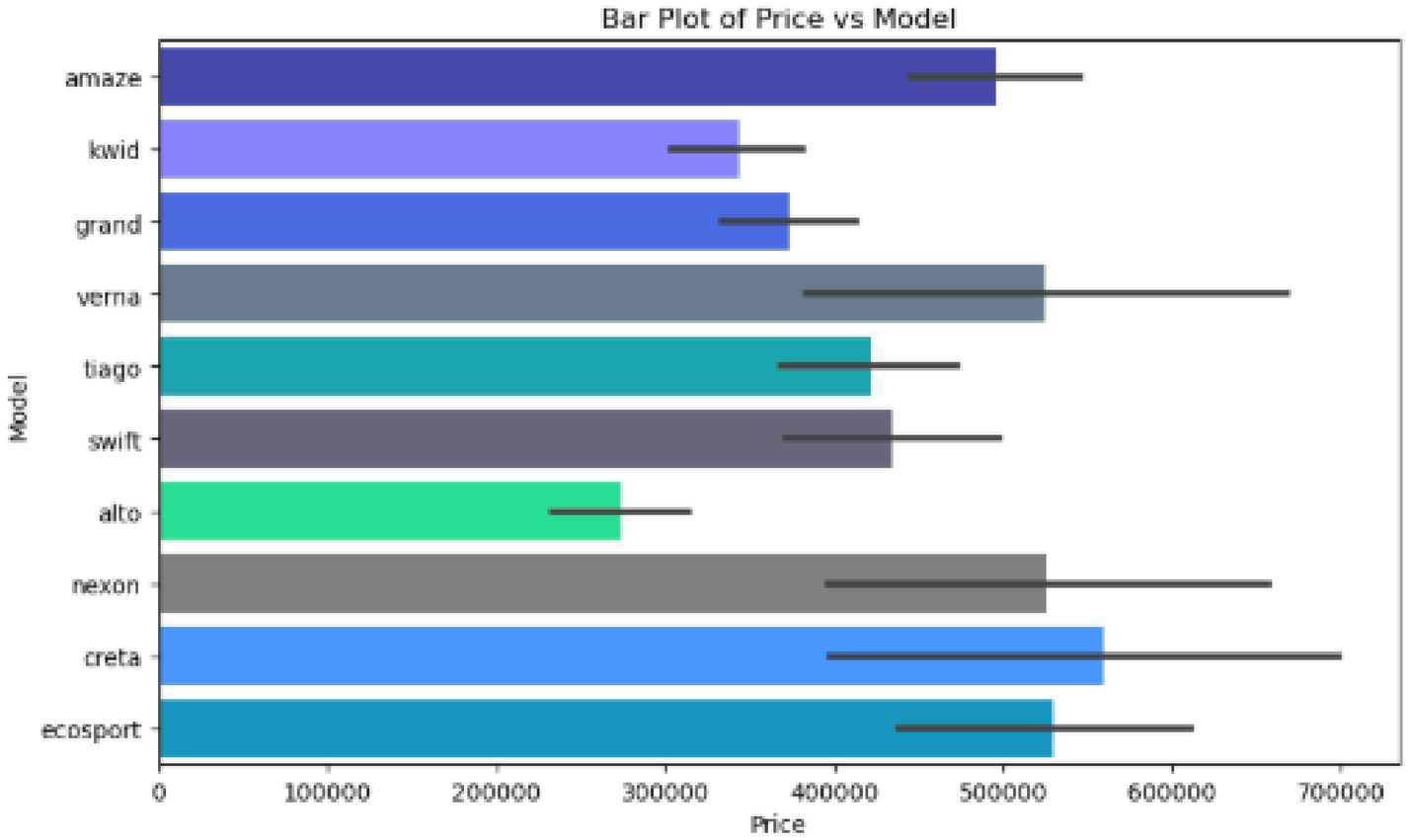
KM_Driven_Actual vs Price:
Shows a weak negative correlation.
As kilometers driven increases, price tends to decrease, but the relationship is scattered and not very strong.
Suggests other factors (like brand, year, or condition) also heavily influence price.

INNOMATICS RESEARCH LABS

Categorical vs Numerical :

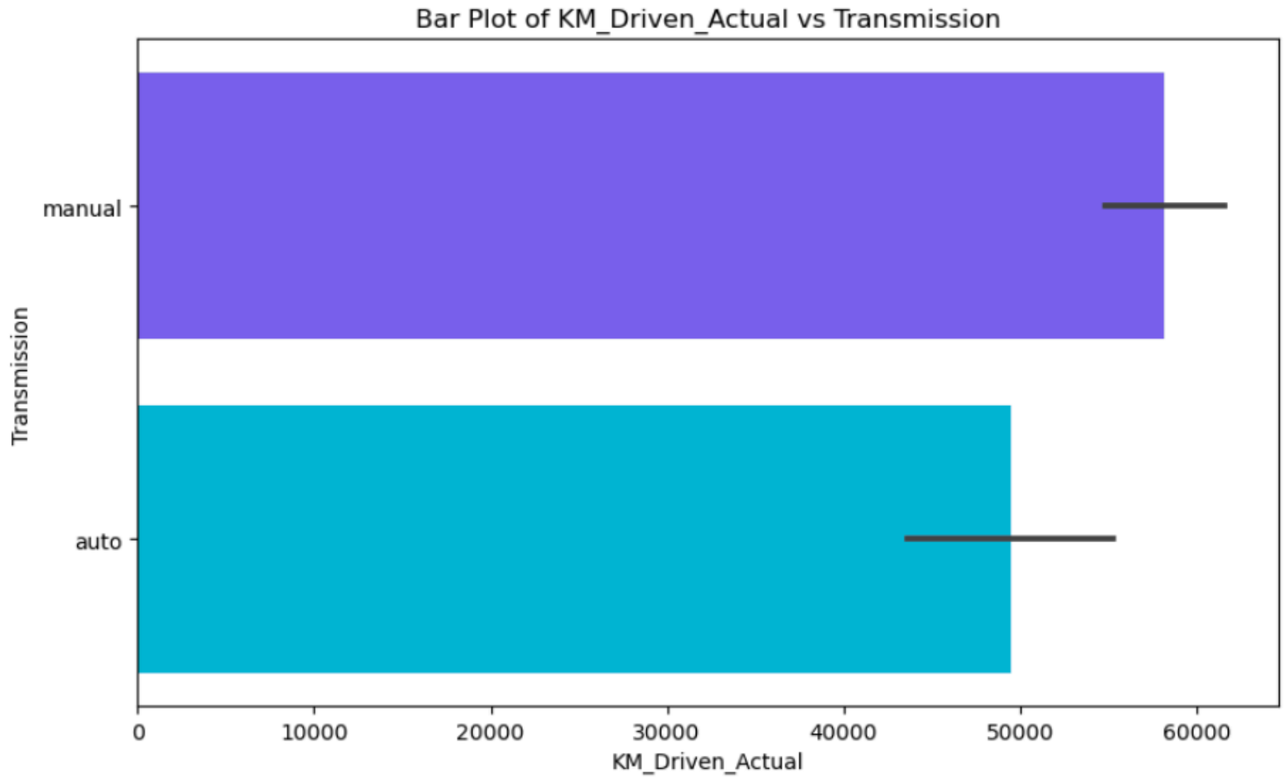


- **Toyota has the highest average price, followed by Mahindra and MG.**
- **Maruti, Renault, and Volkswagen have the lowest average prices, indicating they are more budget-friendly brands.**
- **Tata, Hyundai, and Honda fall in the mid-price range.**
- **This pricing trend helps segment the market by brand value — luxury (Toyota), mid-range (Hyundai, Tata), and economy (Maruti, Renault).**
- **Useful for setting brand-specific pricing strategies and predictive modeling.**
- **Helps align customer recommendations with budget preferences.**



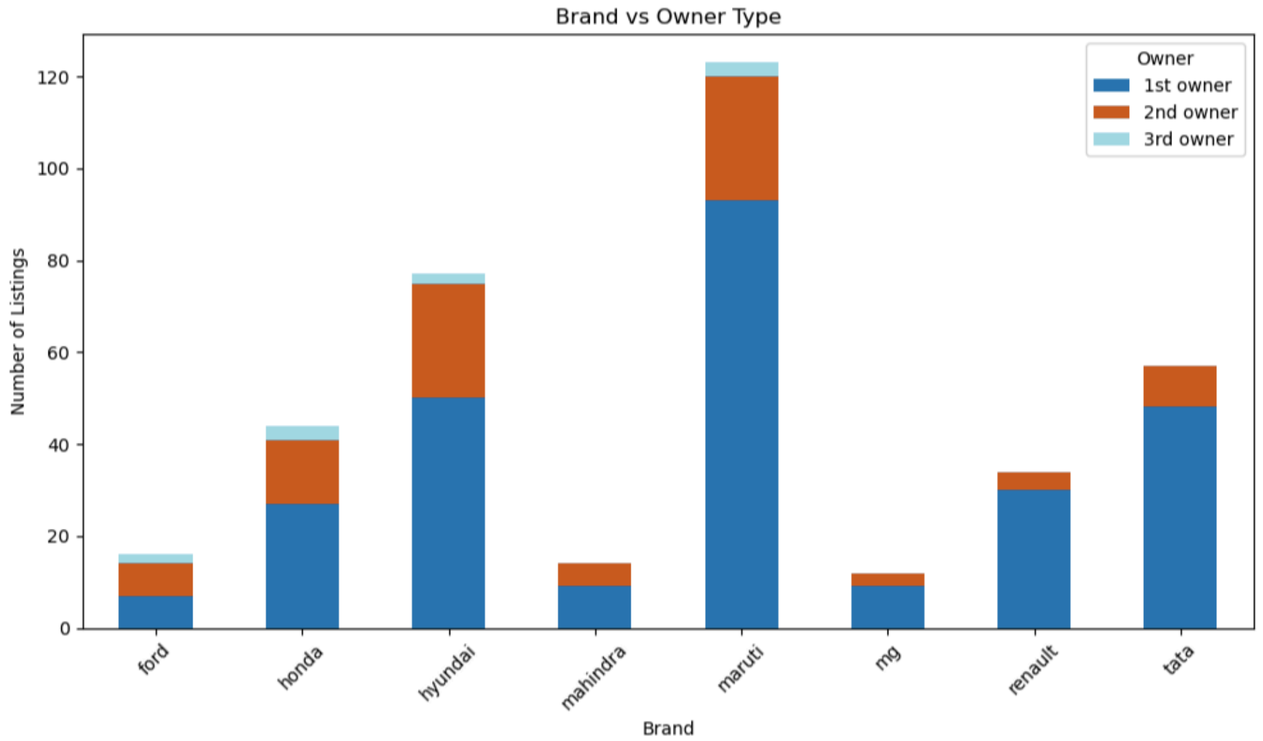
- **Verna and Creta are the most expensive models with high price variation, indicating premium variants.**
- **Alto and Kwid are the most affordable and price-stable, ideal for budget buyers.**
- **Nexon, Ecosport, and Grand fall in the mid-price range, offering a balance of cost and features.**
- **Large error bars in some models suggest wider price ranges due to different trims or conditions.**

KM Driven vs Transmission Type:



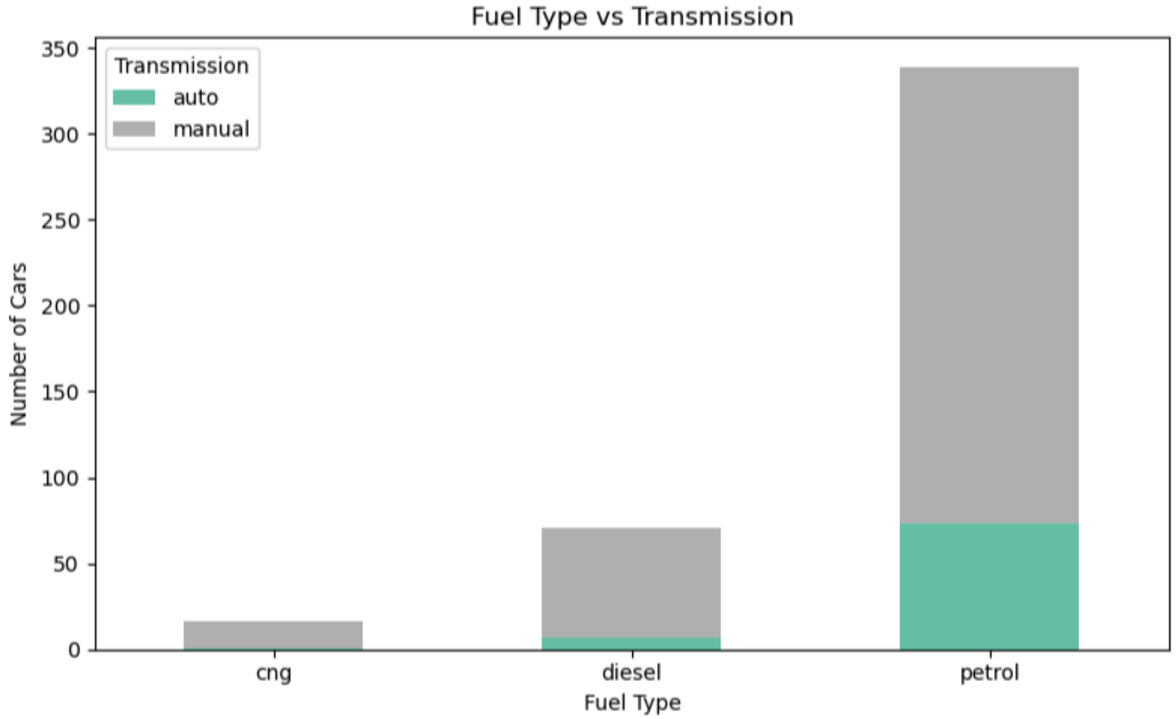
- **Manual transmission cars have a higher average kilometers driven than automatic cars.**
- **Auto cars are typically less driven, which may indicate:**
- **They are newer or used in urban settings.**
- **Manual cars are often owned longer or used for long-distance or commercial purposes.**

Categorical vs Categorical :



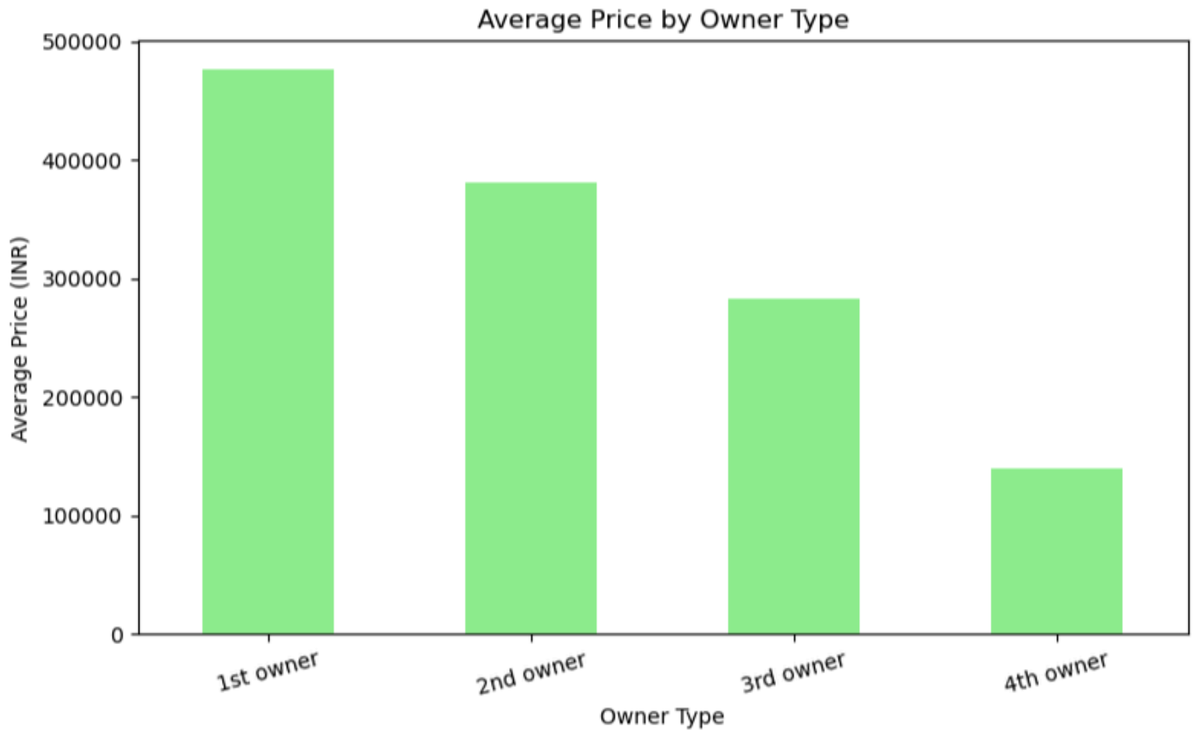
- **Maruti, Hyundai, and Tata have the highest number of listings, mostly from 1st owners.**
- **Other brands like Honda, Renault, and Ford show a higher proportion of 2nd and 3rd owner vehicles.**
- **Maruti shows strong presence across all owner types, especially 1st and 2nd owners.**
- **Mahindra and MG have fewer listings overall, mostly from 1st owners.**

Fuel Type vs Transmission:



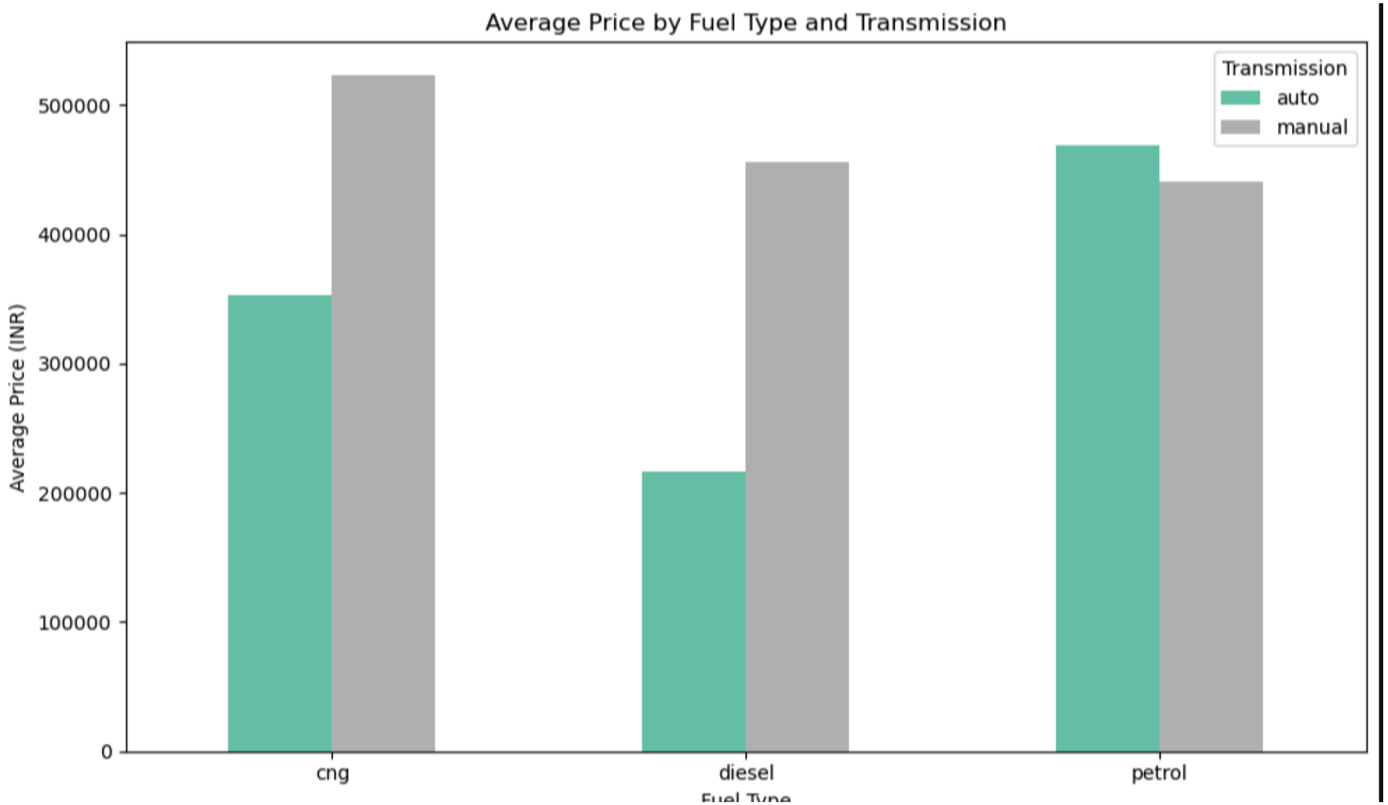
- **Petrol cars dominate the dataset and are available in both manual and automatic transmissions, though manual is more common.**
- **Diesel cars are mostly mannual, with very few automatic options.**
- **CNG cars are also predominantly manual, with almost no automatic variants.**
- **Petrol cars offer the widest choice in transmission types — useful for buyers with diverse preferences.**
- **Manual transmission dominates overall, especially in diesel and CNG segments.**

Average Price by Owner Type:



- **The average price decreases as the number of previous owners increases.**
- **1st owner cars have the highest average price, while 4th owner cars have the lowest.**
- **This trend reflects market perception: newer ownership = better condition and higher value.**
- **Ownership history is a strong factor in price prediction models.**
- **Helps in pricing strategy — vehicles with fewer owners can be sold at a premium.**
- **Useful for customer filtering — buyers seeking lower prices may consider 3rd or 4th owner cars, while others prefer 1st/2nd owner vehicles for quality and resale value.**

Fuel Type vs Transmission on Price:



- **For CNG and diesel cars, manual transmission models have a significantly higher average price than automatic ones.**
- **For petrol cars, automatic transmission models are priced slightly higher than manuals.**
- **This suggests:**
- **CNG and diesel manual cars are likely part of newer or more premium models.**
- **Petrol automatics are more in demand, pushing up their average price.**

INNOMATICS RESEARCH LABS

Overview of the Project :

- **What Each Plot Shows:**

- **Price Distribution**
- Most used car prices fall in the ₹2–7 lakh range.

- **KM Driven Distribution**
- The majority of cars have between 40,000 to 100,000 km driven.

- **Price vs EMI**
- Strong positive correlation — as price increases, EMI increases consistently.

- **Price by Owner Type**
- 1st owner cars command higher resale prices.

- Value declines for 2nd, 3rd, and 4th owners.

- **Top Brands**
- Maruti, Hyundai, Tata, and Honda dominate the dataset.

- Reflects market popularity.

- **Fuel Type vs Transmission**
- Petrol + Manual dominates.
- Diesel and CNG cars are almost entirely manual.

INNOMATICS RESEARCH LABS

Conclusion:

- This project focused on analyzing a used car dataset to understand key factors that influence the resale price of cars and overall market trends. The main aim was to extract insights that can help in building models for price prediction, improving customer recommendations, and supporting better decision-making for buyers and sellers.

- From the analysis, we found that most used cars fall within the affordable price range of ₹2 to ₹7 lakhs. Petrol cars are the most commonly available, followed by diesel and CNG. Manual transmission is dominant across all fuel types, showing that it remains the most popular choice in the market. Another important insight is that cars with fewer previous owners, especially first-owner cars, tend to have much higher resale value. As the number of owners increases, the price generally drops.

- Brands like Maruti, Hyundai, and Tata are the most listed and in-demand, indicating their strong presence and trust in the second-hand market. Cars with lower kilometers driven also tend to have better resale prices, as buyers prefer vehicles with less wear and tear. A strong correlation was also observed between car price and EMI, meaning higher-priced cars come with proportionately higher monthly payments.

- Overall, this dataset is valuable for building smart tools like price prediction models, car recommendation engines, and resale scoring systems. It clearly shows what factors customers care about and how sellers can improve their listings to attract more buyers. These insights provide a solid foundation for developing a more intelligent and user-friendly used car platform.

THANK YOU !!

INNOMATICS
RESEARCH LABS