INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Title: EDA Web Scrapping on “Used cars analysis on Area wise and all over India"

By:
CH BHAVANI CHARY
D AKSHAY KUMAR REDDY
B UMESH CHANDRA

ABOUT US:

CH BHAVANI CHARY
MBA Graduate / Data Analytics and Visualization.
LinkedIn : http://www.linkedin.com/in/bhavani-chary-chityala/
GitHub : https://github.com/chityala-bhavani-chary

D AKSHAY KUMAR REDDY
MCA Graduate / Data Analytics and Visualization.
LinkedIn :
GitHub :

B UMESH CHANDRA
B.Tech Graduate / Data Analytics and Visualization.
LinkedIn :
GitHub :

INNOMATICS RESEARCH LABS

Agenda :

Introduction

Business Problem

Objective of the Project

Tools used

Web Scraping – Details (Websites, Processor followed by us)

Summary of the Data

Exploratory Data Analysis:

Data Cleaning Steps

Data Manipulation Steps

Univariate Analysis Steps

Bivariate Analysis Steps

Multi Variate Analysis steps

Conclusion (Key finding overall)

Your Experience/Challenges working on Web Scraping – Data Analysis Project.

INNOMATICS RESEARCH LABS

Introduction

Introduction :

- The Indian used car market is expanding rapidly due to rising vehicle costs, urban mobility needs, and digital resale platforms.
- Cars24 operates as a large-scale online marketplace enabling transparent buying and selling of used vehicles across India.
- Used car pricing is influenced by multiple factors including fuel type, vehicle age, mileage, and regional demand.
- Understanding these factors is essential for accurate pricing, better inventory planning, and informed decision-making.
- Raw marketplace data often contains variability, outliers, and hidden patterns that require systematic analysis.
- Exploratory Data Analysis (EDA) helps uncover relationships, trends, and distributions within used car data.
- This project applies EDA techniques on Cars24 data to extract insights into resale value drivers and market behavior.

Business Problem :

- Used car prices vary significantly due to multiple factors such as **fuel type, vehicle age, mileage, and regional demand**, making pricing complex and inconsistent.
- Absence of clear, data-driven pricing insights can result in:
- **Over pricing** leading to slower sales and inventory backlog, **Under pricing** causing revenue loss
- Without proper analytical insights, platforms may face **inefficient inventory allocation** across regions and fuel types.
- Regional and fuel-based demand variations are often overlooked, resulting in **suboptimal market and pricing strategies**.
- The Cars24 dataset reflects a **real-world used car marketplace**, capturing diverse vehicle attributes and regional behavior.
- Analyzing this data enables platforms like Cars24 to:
  - **Understand regional demand patterns**,
  - **Identify high-value and high-liquidity vehicle segments**,
  - **Optimize pricing and inventory planning decisions**

Objective :

- To analyze **area-wise and state-wise distribution of used car prices across India**.
- To identify **regional variations in resale value** and pricing behavior.
- To study the impact of **fuel type, kilometres travelled, and manufacturing year** on used car prices.
- To examine **state-wise fuel preferences** and their influence on resale pricing.
- To compare **high-demand and low-demand regions** in the used car market.
- To apply **univariate, bivariate, and multivariate exploratory analysis** techniques.
- To generate **data-driven insights** supporting regional pricing, inventory, and market strategies.

Tools Used :

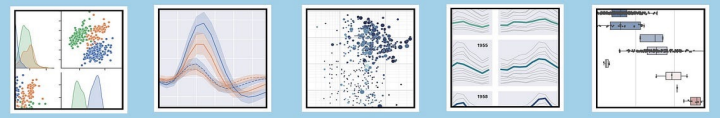
NumPy

Requests
http for humans

Selenium
Python

pandas

seaborn



.[RegEx]*

matplotlib

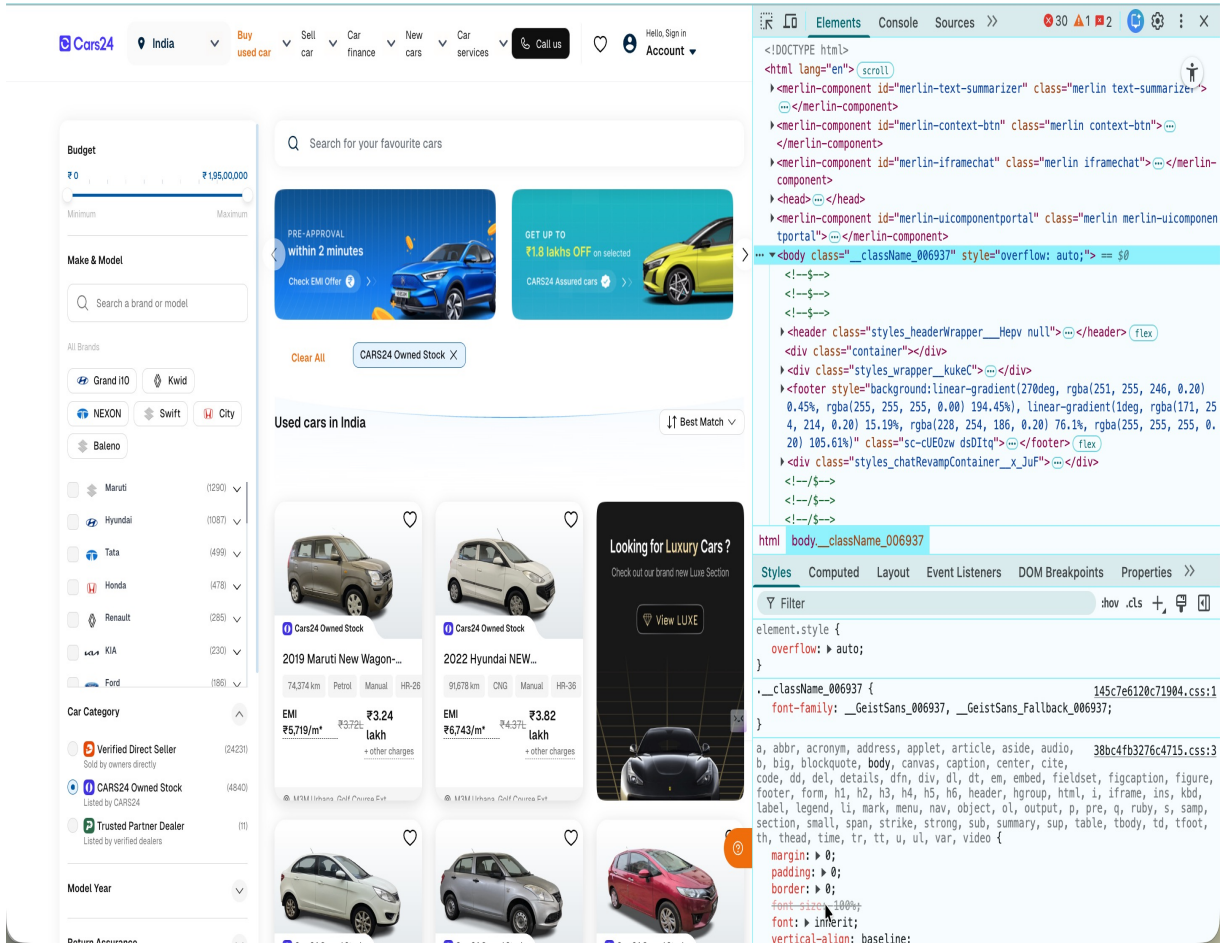
BeautifulSoup

INNOMATICS
RESEARCH LABS

Web Scraping

Web Scraping:

- Web scraping is the process of **automatically extracting structured data from websites** for analysis and decision-making.
- Online platforms like Cars24 generate large volumes of **real-time market data** that are not always available in downloadable formats.
- Manual data collection from such platforms is **time-consuming, error-prone, and inefficient**.
- Web scraping enables analysts to **collect large datasets quickly and consistently** from online marketplaces.
- Scraped data can include key attributes such as **price, fuel type, location, mileage, and manufacturing year**.
- This collected data forms the foundation for **exploratory data analysis (EDA)** and statistical evaluation.
- In this project, web-scraped data supports **area-wise and India-wide analysis of used car pricing trends**.



INNOMATICS RESEARCH LABS

Dataset:

	year_manufactured	brand	model	km_travel	fuel	transmission	state	emi_per_month	price_of_car
0	2019	Maruti	Celerio	85350	CNG	Manual	Haryana	6432.0	329000.0
1	2019	Maruti	New Wagon-R	74374	Petrol	Manual	Haryana	5719.0	324000.0
2	2022	Hyundai	NEW SANTRO	91678	CNG	Manual	Haryana	6743.0	382000.0
3	2022	Tata	NEXON	21703	Petrol	Manual	Karnataka	12233.0	693000.0
4	2020	KIA	SONET	38236	Petrol	Manual	Uttar Pradesh	10203.0	578000.0
...	...	...	...	...	...	...	...	...	...
3175	2022	Nissan	MAGNITE	75684	Petrol	Manual	Karnataka	10463.0	593000.0
3176	2019	Hyundai	VENUE	63224	Petrol	Manual	Telangana	11383.0	645000.0
3177	2018	Honda	WR-V	76894	Diesel	Manual	Telangana	11671.0	597000.0
3178	2019	MG	HECTOR	83300	Petrol	Auto	Haryana	15376.0	898000.0
3179	2018	Maruti	Alto 800	95111	Petrol	Manual	Rajasthan	4590.0	260000.0

3180 rows × 9 columns

```
<class 'pandas.DataFrame'>
RangeIndex: 3180 entries, 0 to 3179
Data columns (total 9 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   year_manufactured  3180 non-null   str    
 1   brand              3180 non-null   str    
 2   model              3180 non-null   str    
 3   km_travel          3180 non-null   int64  
 4   fuel               3180 non-null   str    
 5   transmission       3180 non-null   str    
 6   state              3180 non-null   str    
 7   emi_per_month      3180 non-null   float64
 8   price_of_car       3180 non-null   float64
dtypes: int64(4), str(5)
memory usage: 325.3 KB
```

Data Cleaning Steps:

1.Loading and Inspecting Data

- Loaded the Cars24 dataset and examined its structure, column names, data types, and basic statistics to understand the data quality.

2.Deleting Irrelevant Columns

- Removed unnecessary or non-informative columns that did not contribute to price analysis or regional insights.

3.Handling Duplicate Records

- Checked for duplicate car listings and ensured that duplicate entries did not affect the analysis results.

4.Data Type Conversion

- Converted columns such as price, kilometres travelled, EMI, and year of manufacture into appropriate numerical formats for accurate analysis.

5.Handling Missing Data

- Identified missing values and handled them appropriately to avoid distortion in statistical calculations and visualizations.

6.Dealing with Outliers

- Detected extreme price values and high-mileage vehicles; used median-based interpretation and visual analysis to minimize outlier impact.

INNOMATICS RESEARCH LABS

Cleaned Dataset & checking data structure:

```
df.info()
```

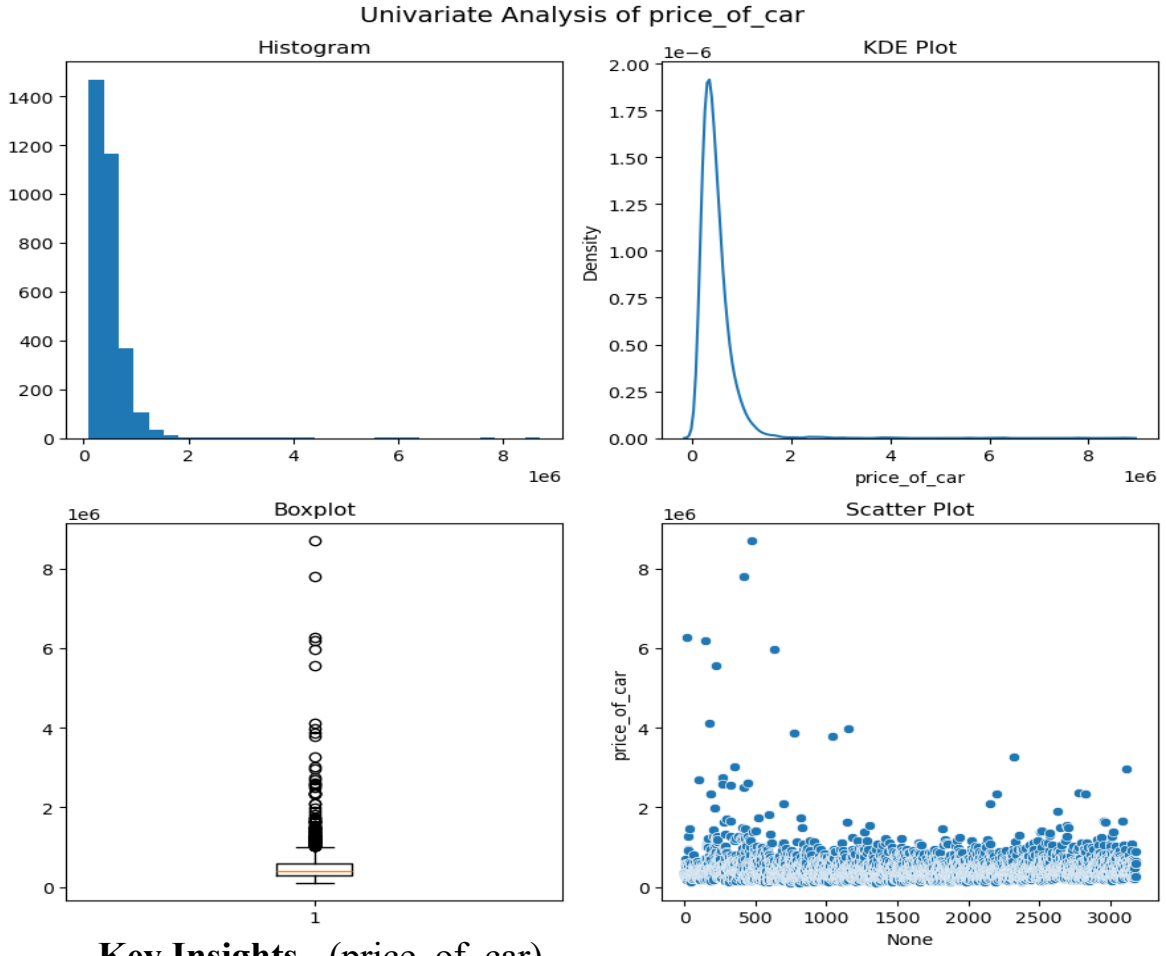
```
<class 'pandas.DataFrame'>
RangeIndex: 3180 entries, 0 to 3179
Data columns (total 9 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   year_manufactured   3180 non-null   int64 
 1   brand               3180 non-null   str   
 2   model               3180 non-null   str   
 3   km_travel           3180 non-null   int64 
 4   fuel                3180 non-null   str   
 5   transmission        3180 non-null   str   
 6   state               3180 non-null   str   
 7   emi_per_month       3180 non-null   int64 
 8   price_of_car        3180 non-null   int64 
dtypes: int64(4), str(5)
memory usage: 325.3 KB
```

```
df
```

	year_manufactured	brand	model	km_travel	fuel	transmission	state	emi_per_month	price_of_car
0	2019	Maruti	Celerio	85350	CNG	Manual	Haryana	6432	329000
1	2019	Maruti	New Wagon-R	74374	Petrol	Manual	Haryana	5719	324000
2	2022	Hyundai	NEW SANTRO	91678	CNG	Manual	Haryana	6743	382000
3	2022	Tata	NEXON	21703	Petrol	Manual	Karnataka	12233	693000
4	2020	KIA	SONET	38236	Petrol	Manual	Uttar Pradesh	10203	578000
...	...	...	...	...	...	...	...	...	...
3175	2022	Nissan	MAGNITE	75684	Petrol	Manual	Karnataka	10463	593000
3176	2019	Hyundai	VENUE	63224	Petrol	Manual	Telangana	11383	645000
3177	2018	Honda	WR-V	76894	Diesel	Manual	Telangana	11671	597000
3178	2019	MG	HECTOR	83300	Petrol	Auto	Haryana	15376	898000
3179	2018	Maruti	Alto 800	95111	Petrol	Manual	Rajasthan	4590	260000

Data Analysis

1. UNIVARIATE ANALYSIS:

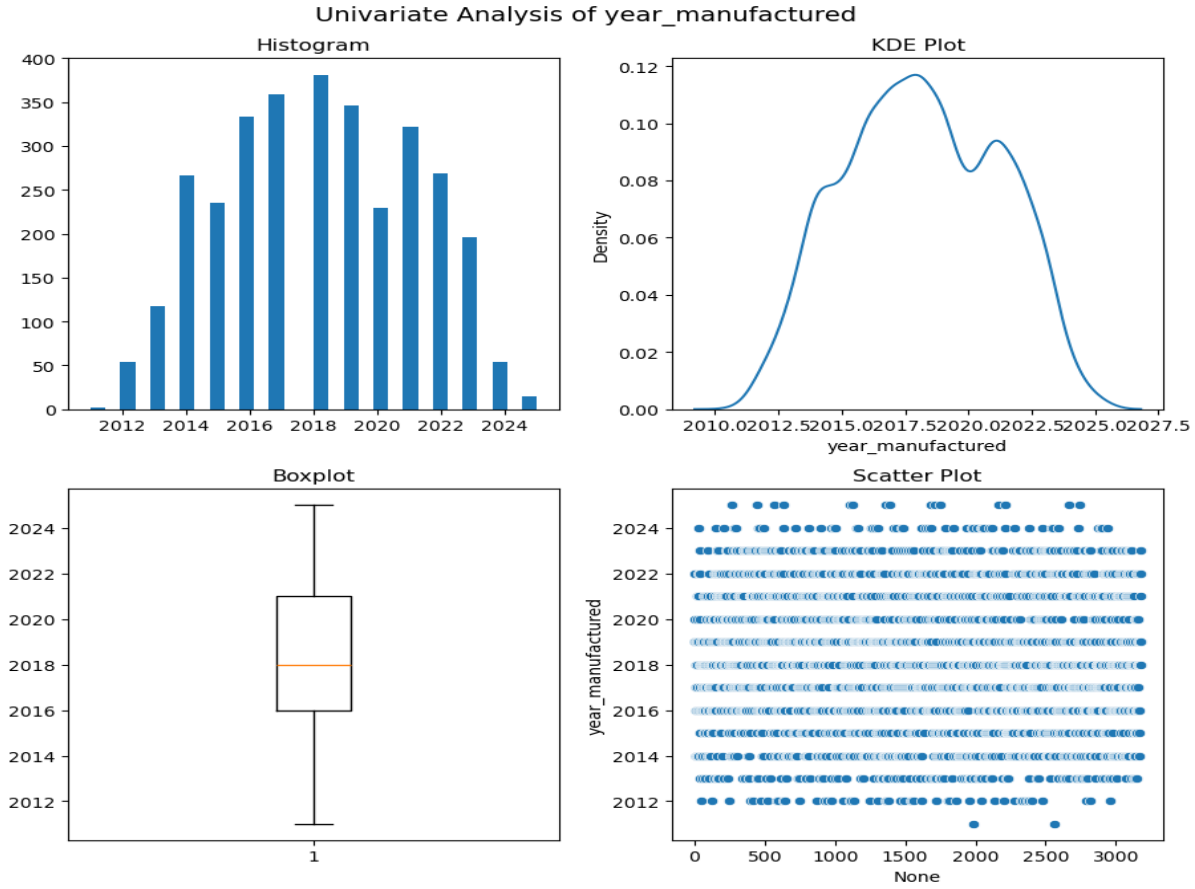


Key Insights - (price_of_car)

•**Affordable segment dominates:** Majority of cars are priced in the lower to mid-range, with most values below ₹10 lakh.

•**Presence of luxury vehicles:** A long right tail and several high-end outliers (up to ₹80+ lakh) indicate a small luxury segment.

•**Highly skewed distribution:** Very high skewness and kurtosis confirm prices are non-uniform and outlier-driven, not normally distributed.



Key Insights - (year_manufacture)

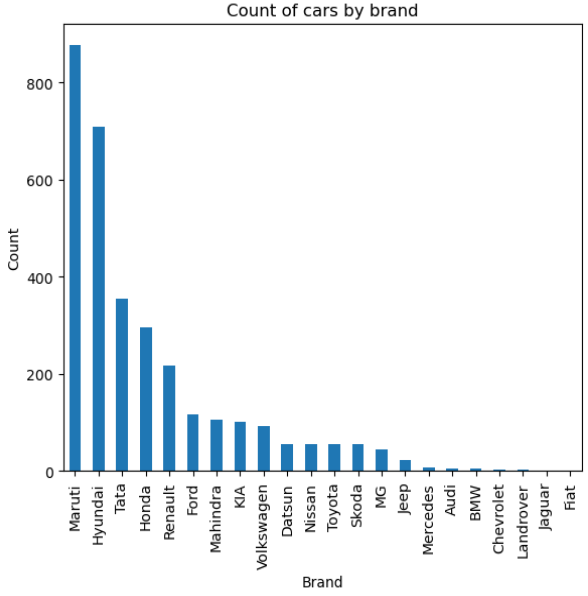
•**Used-car dominant inventory:** Most cars are 4–10 years old, with the highest concentration around 6–8 years, indicating a mature and stable used-car market.

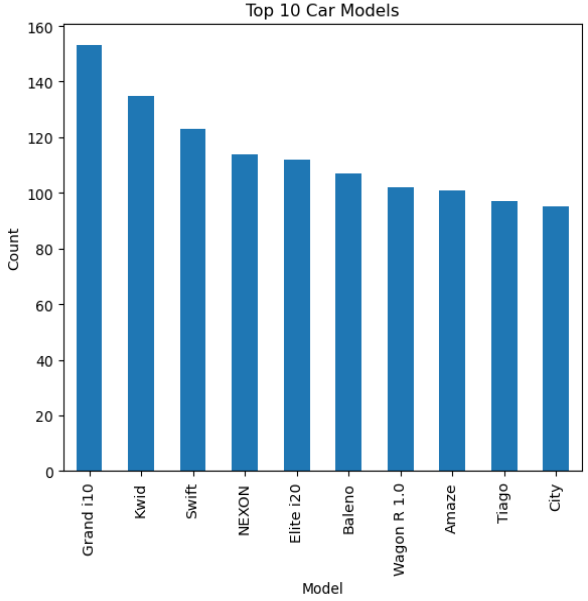
•**Well-balanced age distribution:** Nearly symmetric distribution with very low skewness, showing no bias toward extremely new or very old vehicles.

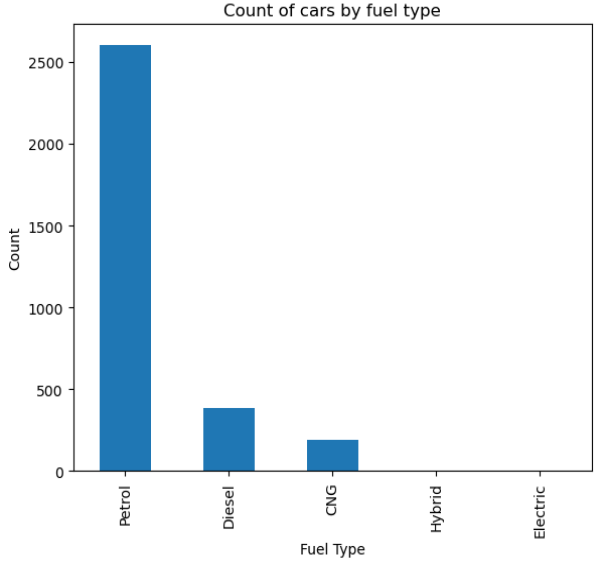
•**Clean and reliable data:** Age ranges from 0 to 14 years with no extreme outliers, reflecting controlled data quality.

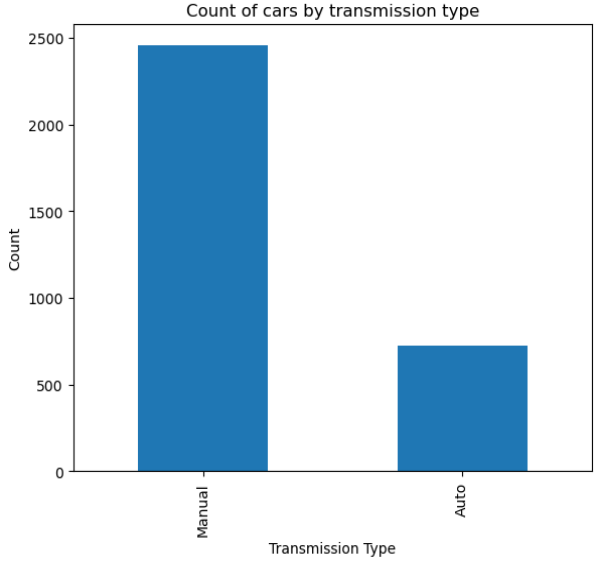
INNOMATICS RESEARCH LABS

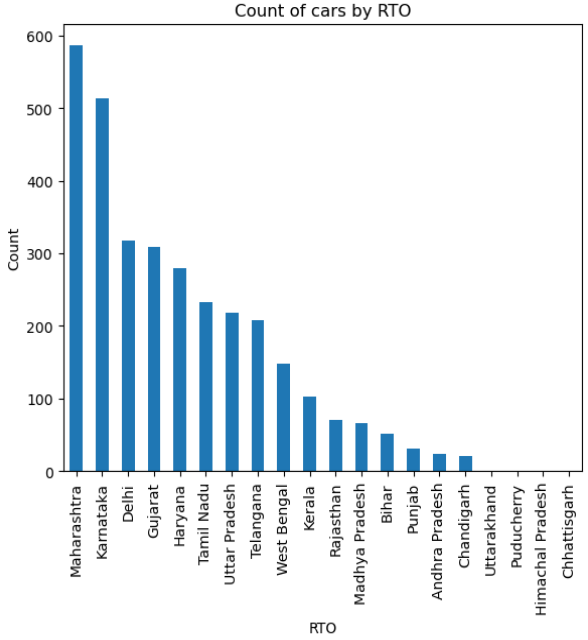
1. UNIVARIATE ANALYSIS:

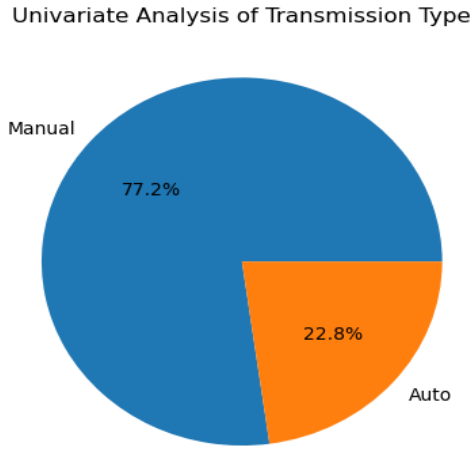












Key Insights:

•**Maruti and Hyundai** clearly dominate the dataset with the highest car counts.

•**No single car model** overwhelmingly dominates among the top 10 models.

•**Petrol** is the most common fuel type, with Diesel far behind and others negligible.

•**Manual transmission** cars clearly dominate the dataset, accounting for about 77% of all vehicles, while **automatic cars** make up only around 23%.

•Car registrations are concentrated in a few RTOs, especially **Maharashtra and Karnataka**.

INNOMATICS RESEARCH LABS

2. BIVARIATE ANALYSIS:

5.1 Numerical vs Numerical

A. Non-Visualization (Statistical Measures)

[28]:

```
# 1) Correlation Coefficient - [KM traveled - sale price, emi]
df[['km_travel', 'price_of_car', 'emi_per_month']].corr()
```

[28]:

	km_travel	price_of_car	emi_per_month
km_travel	1.000000	-0.14098	-0.031756
price_of_car	-0.140980	1.00000	0.728370
emi_per_month	-0.031756	0.72837	1.000000

5.2 Numerical vs Categorical

A. Non-Visualization (Statistical Measures)

[35]:

```
# 1) GroupBy - [state & price of cars]
group_by = df.groupby('state')['price_of_car'].median().round(0)
group_by
```

[35]:

```
state
Andhra Pradesh       257500.0
Bihar                445000.0
Chandigarh           445000.0
Chhattisgarh         245000.0
Delhi                322000.0
Gujarat              400000.0
Haryana              391000.0
Himachal Pradesh    2350000.0
Karnataka            496000.0
Kerala               310000.0
Madhya Pradesh       426000.0
Maharashtra          380000.0
Puducherry          5975000.0
Punjab               499000.0
Rajasthan            437500.0
Tamil Nadu           386000.0
Telangana            428000.0
Uttar Pradesh        461500.0
Uttarakhand          340000.0
West Bengal          400000.0
Name: price_of_car, dtype: float64
```

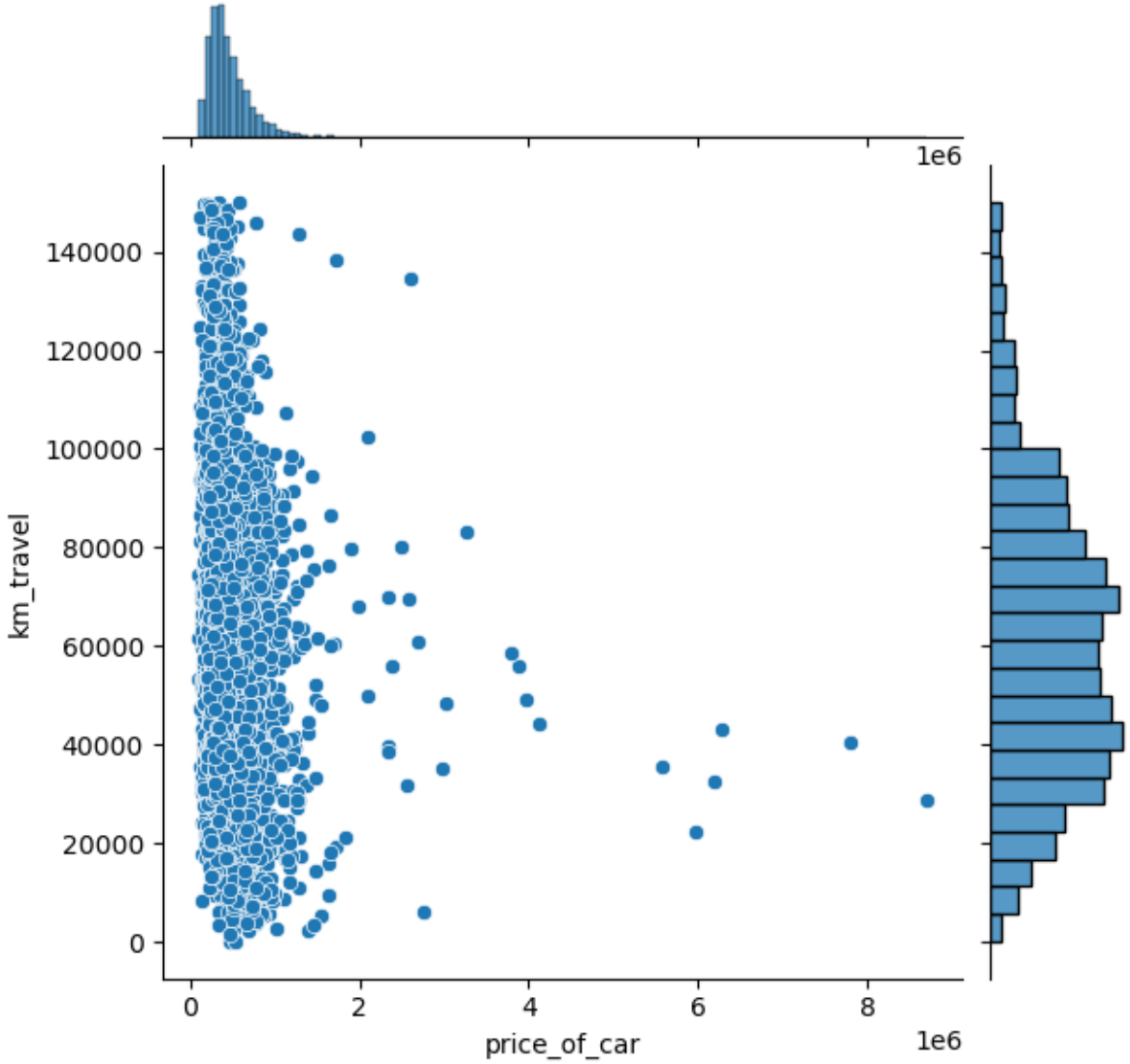
5.3 Categorical vs Categorical

A. Non-Visualization (Statistical Measures)

```
# 1) Crosstab - [state Vs  fuel]
pd.crosstab(df['state'], df['fuel'])
```

fuel	CNG	Diesel	Electric	Hybrid	Petrol
state					
Andhra Pradesh	0	4	0	0	20
Bihar	2	9	0	0	40
Chandigarh	1	3	0	0	17
Chhattisgarh	0	0	0	0	1
Delhi	25	4	0	0	288
Gujarat	33	41	0	0	234
Haryana	29	24	0	0	226
Himachal Pradesh	0	1	0	0	0
Karnataka	5	85	0	1	422
Kerala	3	10	0	0	90
Madhya Pradesh	8	7	0	0	51
Maharashtra	65	63	1	2	455
Puducherry	0	0	0	0	1
Punjab	1	12	0	0	18
Rajasthan	1	10	0	0	59
Tamil Nadu	2	28	0	0	203
Telangana	4	40	0	0	164
Uttar Pradesh	12	32	0	0	174
Uttarakhand	0	0	0	0	1
West Bengal	0	12	0	0	136

2. BIVARIATE ANALYSIS:



1) Num vs Num - Visualization

Jointplot - [km_travel vs price_of_car]

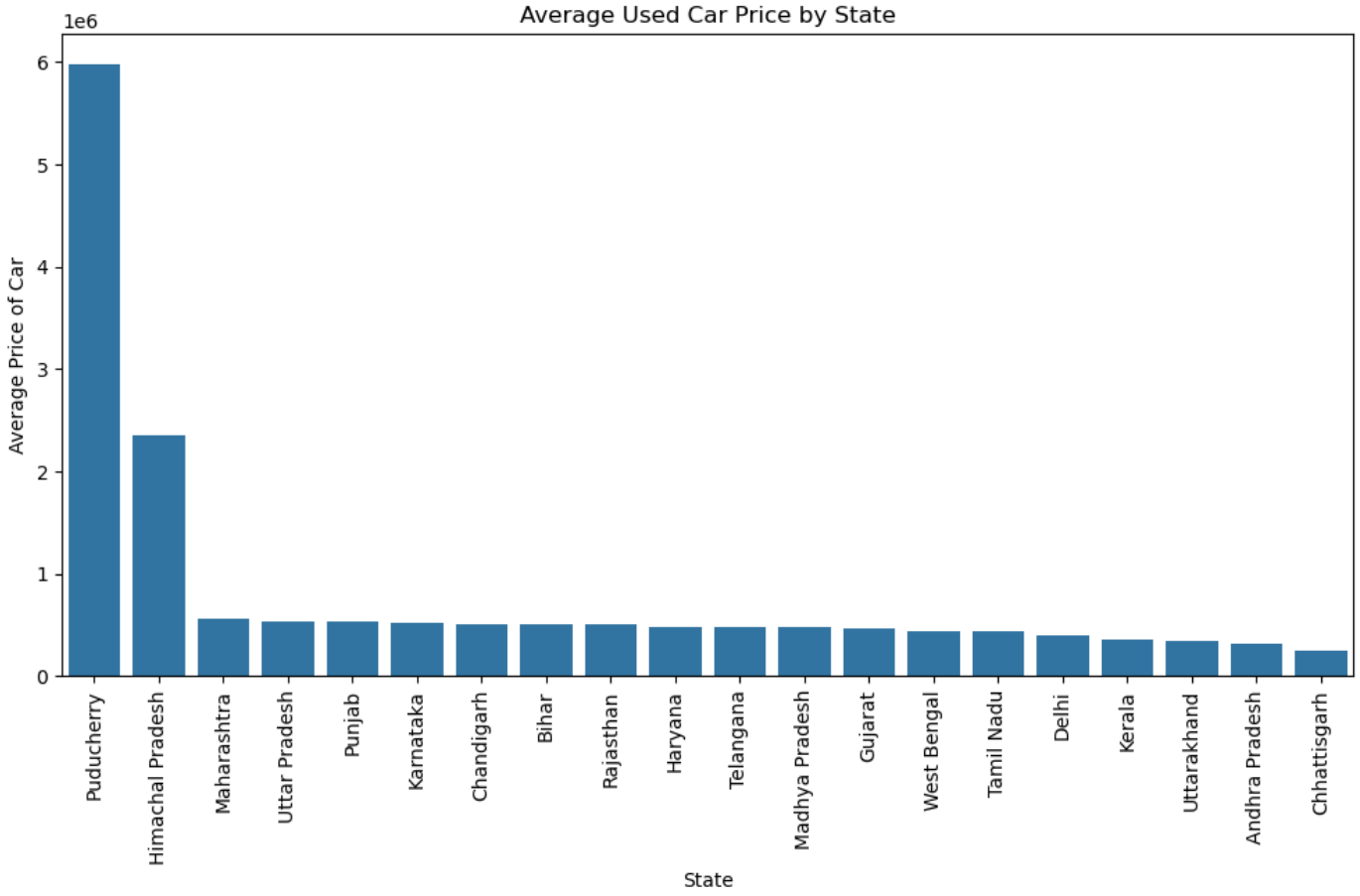
Insights

•**Car price and kilometers travelled** show a weak inverse relationship, confirming mileage-related depreciation.

•**High-priced vehicles are concentrated at lower mileage levels**, indicating premium cars retain value when usage is limited.

INNOMATICS RESEARCH LABS

2. BIVARIATE ANALYSIS:



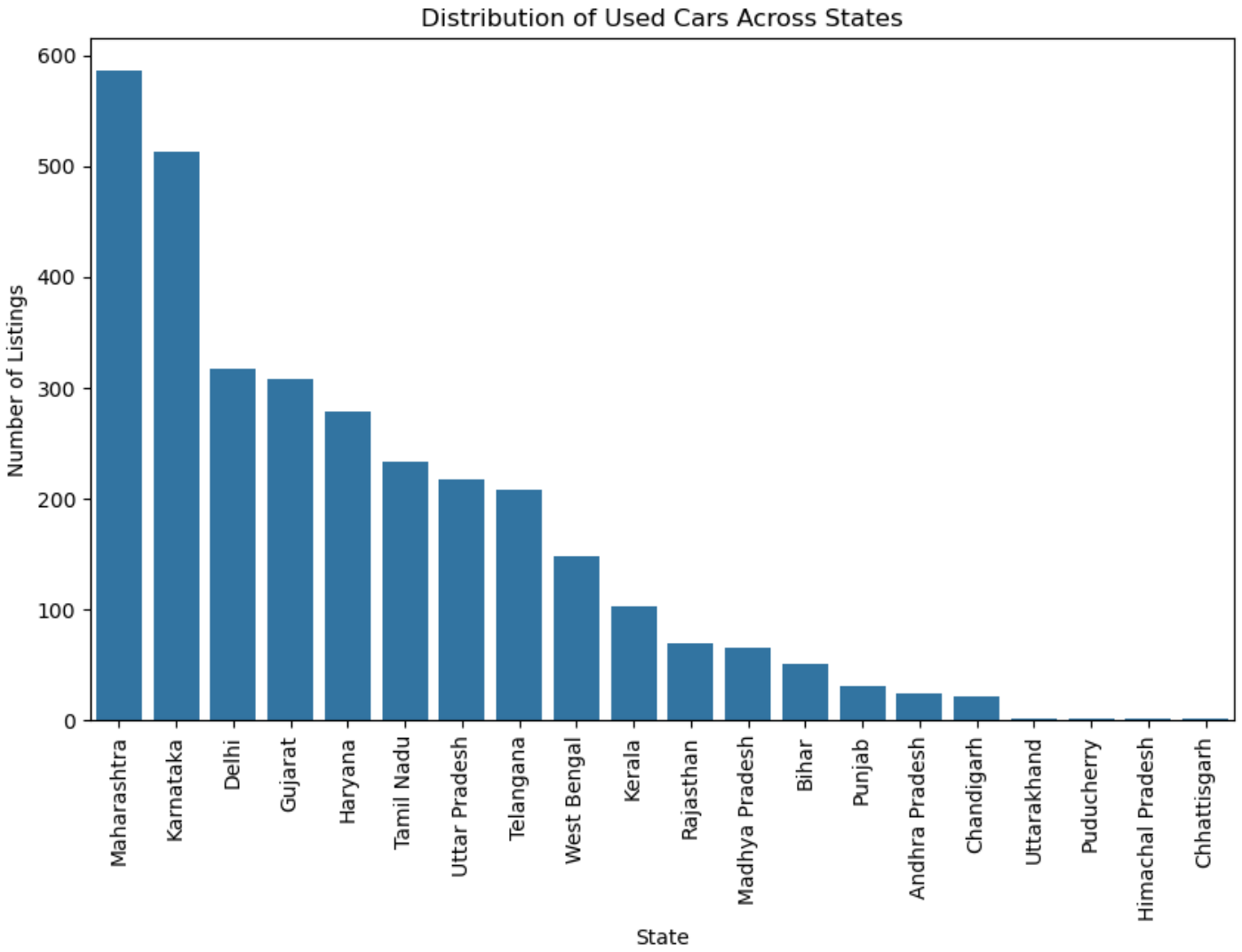
2) Num vs Cat - Visualization

Bar Plot - [state vs average price_of_car]

Insights

•**Average used car prices vary substantially across states**, indicating strong regional differences in resale market dynamics.
•**States with higher average prices** (such as Maharashtra, Karnataka, Punjab, and Uttar Pradesh) likely reflect greater demand for newer or premium vehicles, higher income levels, and stronger urban presence.
•**States with lower average prices** suggest a higher concentration of older or budget vehicles, possibly driven by lower purchasing power or rural dominance.
•**Extremely high average prices in certain states** may be influenced by a small number of high-value listings, highlighting the impact of outliers on mean-based analysis.
•Overall, the plot indicates that **state-level factors play a crucial role in determining used car pricing**, emphasizing the need for region-specific pricing and inventory strategies.

2. BIVARIATE ANALYSIS:



2) Cat vs Cat - Visualization

Count Plot - [state vs count]

Insights

•**Used car listings are highly concentrated in a few states**, indicating regional dominance in the resale market.
•**States with higher listing counts** likely represent larger urban populations, stronger vehicle turnover, and higher digital marketplace adoption.
•**States with fewer listings** may reflect lower market activity, limited platform penetration, or smaller population bases.
•Overall, the distribution highlights **significant regional imbalance** in used car availability across the country.

INNOMATICS RESEARCH LABS

3. MULTIVARIATE ANALYSIS:

Unique Insights from Pair Plot Analysis:

•**Price vs EMI per Month**: Car price increases sharply with EMI per month, showing a strong positive relationship.

•**Price vs Kilometers Travelled**: Car price generally decreases as kilometers travelled increase, though the relationship is weak.

•**Price vs Year Manufactured**: Newer manufactured cars tend to have higher prices, indicating a positive trend.

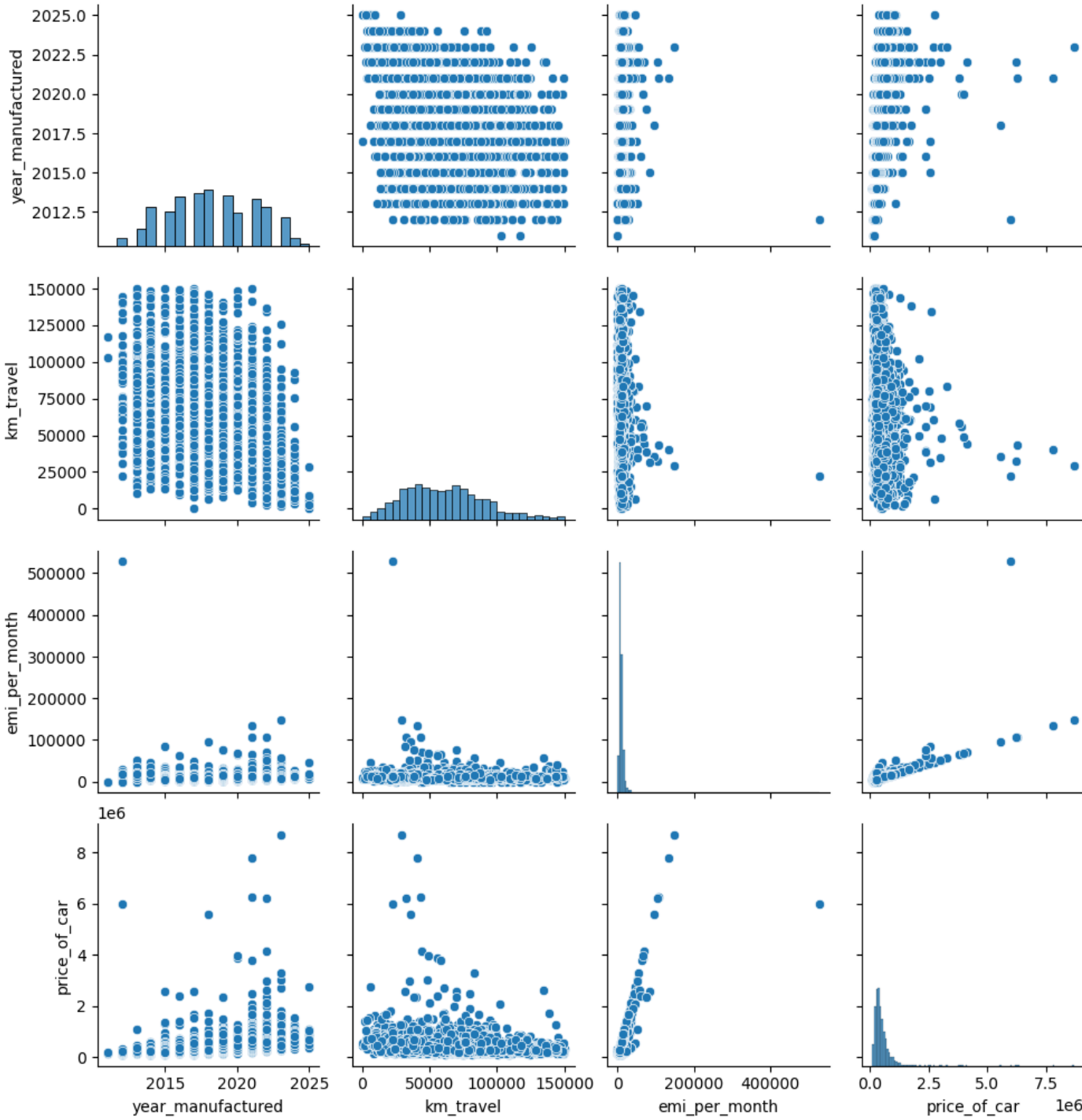
•**EMI vs Kilometers Travelled**: EMI values are scattered across all kilometer ranges, showing no clear relationship.

•**EMI vs Year Manufactured**: EMI shows a very weak upward trend with newer manufacturing years.

•**Kilometers Travelled vs Year Manufactured**: Older manufactured cars tend to have higher kilometers travelled, showing a clear negative relationship.

•**Kilometers Travelled Distribution**: Most cars fall within a mid-range of usage, with fewer extreme values.

•**Year Manufactured Distribution**: The dataset is dominated by newer cars, indicating more recent manufacturing years.



3. MULTIVARIATE ANALYSIS:

Unique Insights from Correlation Analysis:

•**Price vs EMI per Month**: As car price increases, EMI per month also increases strongly, indicating a high positive relationship.

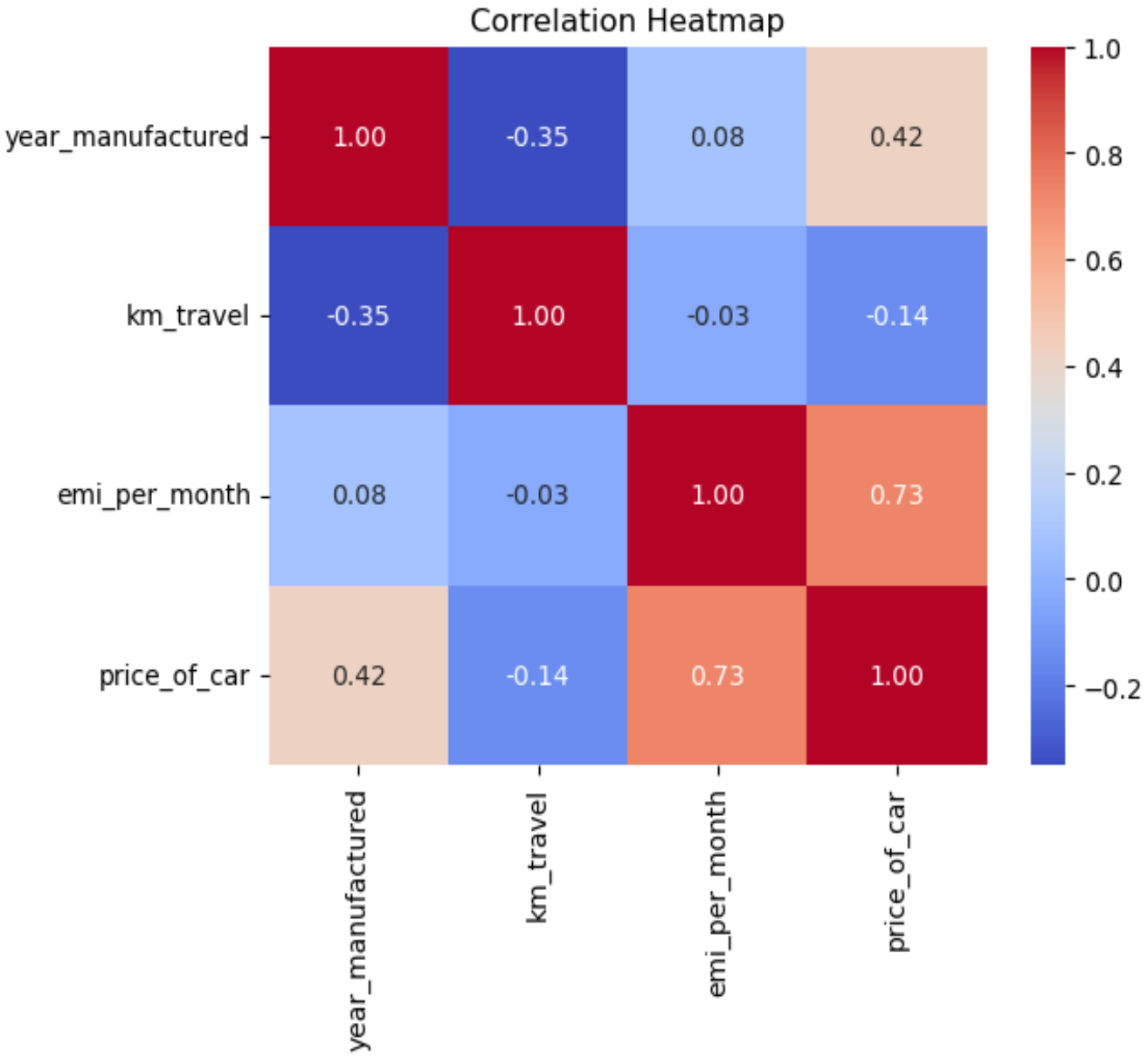
•**Price vs Kilometers Travelled**: As kilometers travelled increase, car price slightly decreases, showing a weak negative relationship.

•**Price vs Year Manufactured**: Newer manufacturing years are associated with higher car prices, indicating a moderate positive relationship.

•**EMI vs Kilometers Travelled**: EMI has almost no relationship with kilometers travelled, meaning usage does not influence EMI values.

•**EMI vs Year Manufactured**: Manufacturing year has a very weak influence on EMI, showing minimal impact.

•**Kilometers Travelled vs Year Manufactured**: Older cars generally have higher kilometers travelled, indicating a moderate negative relationship.



INNOMATICS RESEARCH LABS

Conclusion (Key finding overall) :

- Used car prices in India show **significant area-wise and state-wise variation**, confirming that regional demand strongly influences resale value.
- **Fuel type is a major determinant of price**:
  Petrol vehicles dominate listings, while Diesel vehicles generally retain higher resale value across many states.
- **Manufacturing year has a strong positive relationship with price**, with newer vehicles consistently commanding higher resale prices.
- **Kilometres travelled has a weak negative impact on price**, indicating mileage alone is not a strong predictor of resale value.
- Used car listings are **highly concentrated in a few major states**, such as Maharashtra, Karnataka, and Delhi, reflecting higher market activity in urban regions.
- **CNG vehicles are positioned in the lower price segment**, while Electric and Hybrid vehicles have very limited presence due to early-stage adoption.
- Presence of **price outliers and premium vehicles** in certain states affects average prices, highlighting the importance of median-based analysis.
- Overall, **vehicle characteristics, fuel preference, and regional demand together drive used car pricing trends across India**.

Our Experience/Challenges working on
Web Scraping – Data Analysis Project:

- Scraping data from the Cars24 website was challenging due to **dynamic content loading**, where listings appeared only after scrolling.
- Handling **lazy loading and pagination** required repeated page interactions to ensure complete data extraction.
- Variations in webpage structure caused **inconsistent or missing fields** during the scraping process.
- Raw web-scraped data contained **missing values, mixed data types, and formatting issues**, requiring careful cleaning.
- The dataset exhibited **highly skewed price distributions and extreme outliers**, complicating statistical analysis.
- Limited data for **Electric and Hybrid vehicles** restricted comparative analysis across fuel types.
- Converting complex analytical results into **clear, business-friendly insights** required thoughtful interpretation.

THANK YOU

By:
CH BHAVANI CHARY
D AKSHAY KUMAR REDDY
B UMESH CHANDRA

INNOMATICS RESEARCH LABS