INNOMATICS®
RESEARCH LABS

INNOVATION. AUTOMATION. ANALYTICS

PROJECT ON

Title: Web Scraping & Analysis of Electric Vehicle (EV) Data

INNOMATICS
RESEARCH LABS

About Us:

- We are **Mrutyunjaya Debata** and **Narottam Kar**, B.Tech graduates passionate about **Data Analytics and Visualization**.
- Skilled in **Python**, **Pandas**, **BeautifulSoup**, and **Power BI** for data-driven problem solving.
- Enthusiastic about **real-world data projects** — from web scraping to dashboard creation.
- Interested in **data extraction**, **business insights**, and **market trend analysis** in emerging domains like **Electric Vehicles**.

🔗 **LinkedIn Profiles:**

- https://www.linkedin.com/in/mrutyunjaya3806debata/
- https://www.linkedin.com/in/narottam-kar/

💻 **GitHub profiles:**

- https://github.com/Mrutyunjaya-1
- https://github.com/narottam2003

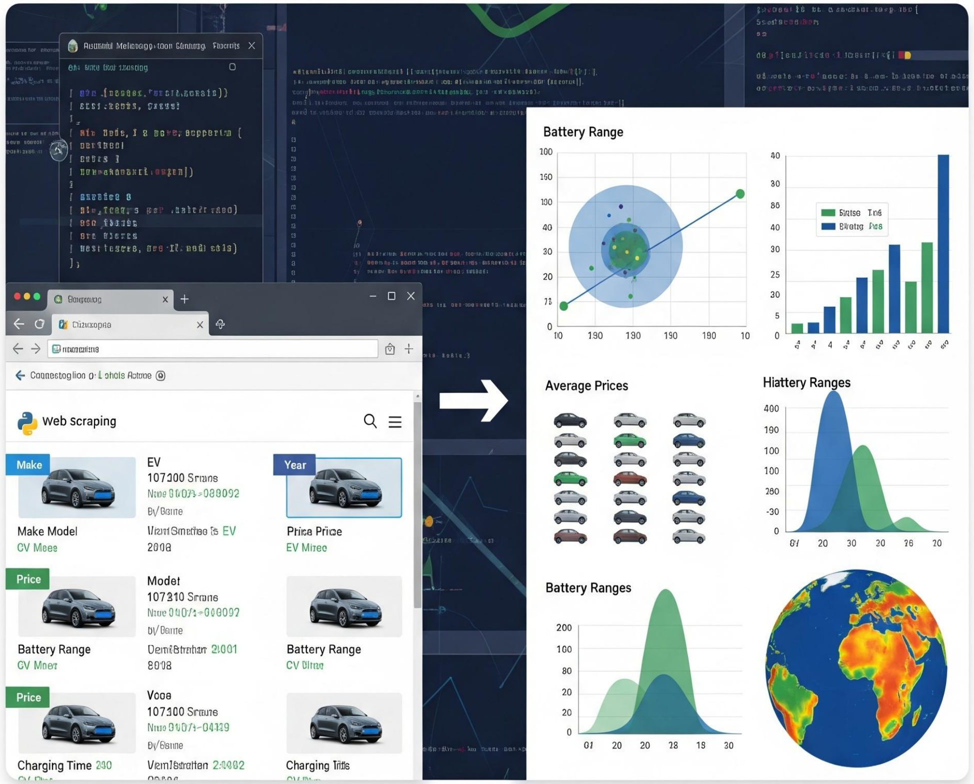
INNOMATICS RESEARCH LABS

Introduction:

- The EV market is rapidly growing, but data is scattered across multiple websites.
- Manual data collection is time-consuming and error-prone.
- **Goal:** Automate data extraction to create a clean, structured EV dataset for analysis.

Web scraping

EDA
Exploratory data Analysis



Data Flow

INNOMATICS
RESEARCH LABS

Business Problem:

1. **Fragmented Data Sources**
   - EV information (range, price, battery, charging time, etc.) is spread across multiple websites, making it difficult for analysts and consumers to compare models.
2. **Lack of Centralized, Updated Dataset**
   - There’s no unified or real-time database for EV specifications and performance metrics, causing decision delays for businesses and consumers.
3. **Inefficient Manual Data Collection**
   - Companies spend time and resources gathering data manually, which is prone to errors and becomes outdated quickly.
4. **Difficulty in Market Comparison & Insights**
   - Auto companies and consumers can’t easily analyze competitors’ EV models — e.g., how price relates to battery capacity or efficiency.
5. **Missed Opportunities for Data-Driven Decisions**
   - Without structured EV data, stakeholders (dealers, manufacturers, policymakers) can’t identify trends or optimize product strategies effectively.

Objectives:

1. **To Collect Reliable EV Data**
    - Automatically extract detailed information about electric vehicles (range, price, battery, efficiency, etc.) from multiple web sources.
2. **To Build a Structured Dataset**
    - Convert unorganized web data into a clean, consistent, and analyzable format (CSV/XLSX).
3. **To Enable Data-Driven Insights**
    - Use exploratory data analysis (EDA) to identify trends, comparisons, and correlations between key EV features.
4. **To Support Business & Consumer Decisions**
    - Provide insights for automakers, analysts, and customers to evaluate EV performance and value.
5. **To Automate and Streamline Data Collection**
    - Minimize manual effort and ensure that EV data can be updated efficiently and accurately.

Web Scraping:

- The official **EV car listing website** was selected as the main data source.
- Used **browser developer tools (Inspect Element)** to identify relevant HTML tags containing EV details like model, range, battery capacity, and price.
- Utilized **BeautifulSoup** and **Requests** libraries in Python to extract and parse structured EV data from multiple pages.
- Automated data retrieval by sending **HTTP requests** to fetch all car details efficiently.
- **Cleaned and consolidated** the scraped data for further analysis and visualization using Pandas.

```
[1]: import json
     import numpy as np
     import pandas as pd
     from pandas import json_normalize
     import requests
     import re
     from bs4 import BeautifulSoup
     import matplotlib.pyplot as plt
     import warnings
     warnings.filterwarnings("ignore")
     import seaborn as sns

[2]:  url = "https://ev-database.org/#group=vehicle-group&rs-pr=10000_100000&rs-er=0_1000&rs-ld=0_1000&rs-a

[3]: response  = requests.get(url)
     response

[3]: <Response [200]>

[4]: response.text[1:100]

[4]: '<!doctype html>\n\n<html lang="en" data-locale="en-IE" data-region="si">\n\t<!-- head -->\n\t<head>\t
     \t\n\t<m'

[5]: pc = response.text

[6]: soup = BeautifulSoup(pc)

[7]: headers = {"User-Agent": "Mozilla/5.0"}

     # Key labels and regex patterns
     FIELDS = {
         'Range': r'(\d+\.?\d*)\s*km',
         'Efficiency': r'(\d+\.?\d*)\s*Wh/km',
         'Weight': r'(\d+\.?\d*)\s*kg',
         '0-100': r'(\d+\.?\d*)\s*s',
         '1-Stop Range': r'(\d+\.?\d*)\s*km',
         'Battery': r'(\d+\.?\d*)\s*kWh',
         'Fastcharge': r'(\d+\.?\d*)\s*kW',
         'Towing': r'(\d+\.?\d*)\s*kg',
         'Cargo Vol.': r'(\d+\.?\d*)\s*L',
         'Price/range': r'€?(\d+\.?\d*)'
     }

     cars = []

     for page in range(1, 51):
         print(f"Scraping page {page}...")
         url = f'https://ev-database.org/?page={page}'
         res = requests.get(url, headers=headers)
         soup = BeautifulSoup(res.text, 'html.parser')

         for item in soup.select('.list-item'):
             model = item.select_one('.title')
             model = model.get_text(strip=True).split('(')[0] if model else 'N/A'

             # Collect all text in item-data block
             specs_text = item.select_one('.item-data')
             specs_text = specs_text.get_text(separator=' ', strip=True) if specs_text else ''

             data = {'Model': model}
             for label, pattern in FIELDS.items():
                 match = re.search(pattern, specs_text)
                 data[label] = match.group(1) if match else None
```

```
             data = {'Model': model}
             for label, pattern in FIELDS.items():
                 match = re.search(pattern, specs_text)
                 data[label] = match.group(1) if match else None

             cars.append(data)

     # Create DataFrame
     df = pd.DataFrame(cars)
     df.to_csv('clean_ev.csv', index=False)

     print(f"\n Scraped {len(df)} EVs successfully!")
     Scraping page 1...
     Scraping page 2...
     Scraping page 3...
     Scraping page 4...
     Scraping page 5...
     Scraping page 6...
     Scraping page 7...
     Scraping page 8...
     Scraping page 9...
     Scraping page 10...
```

INNOMATICS RESEARCH LABS

Tools Used:

BeautifulSoup

pandas

.[RegEx]*

matplotlib

NumPy

seaborn

Data Cleaning Steps:

- Removed unwanted symbols and text (e.g., “km”, “kWh”).
- Standardized units for Battery, Range, and Efficiency.
- Filled missing brand or numeric fields with “Unknown” or median values.
- Converted all numeric fields to float/int types.
- Removed duplicates and irrelevant entries.
- Ensured consistent column naming conventions.

```
dtype: int64
```

```
[15… df["Brand"].fillna(df["Brand"].mode()[0], inplace=True)
```

```
[16… df["Model Name"].fillna(df["Model Name"].mode()[0], inplace=True)
```

```
[17… df.isnull().sum()
```

```
[17… Model           0
     Range           0
     Efficiency      0
     Weight          0
     0-100           0
     1-Stop Range    0
     Battery         0
     Fastcharge      0
     Towing          0
     Cargo Vol.      0
     Price/range     0
     Brand           0
     Model Name      0
     dtype: int64
```

```
[18… df.duplicated().sum()
```

```
[18… np.int64(56196)
```

```
[19… df.drop_duplicates(subset=["Model"], inplace=True)
     df.reset_index(drop=True, inplace=True)
```

```
[20… df.duplicated().sum()
```

```
[20… np.int64(0)
```

```
[21… df.head()
```

[21…

	Model	Range	Efficiency	Weight	0-100	1-Stop Range	Battery	Fastcharge	Towing	Cargo Vol.	Price/range
0	BMWiX3 50 xDrive	610	178	2360	4.9	610	108.7	108.7	2360	2	3.[illegible]
1	MGMG4 Electric 64 kWhMG MG4	360	171	1726	7.9	360	64.0	64.0	1726	2	4.[illegible]

Data Visualization:

Key Insights:

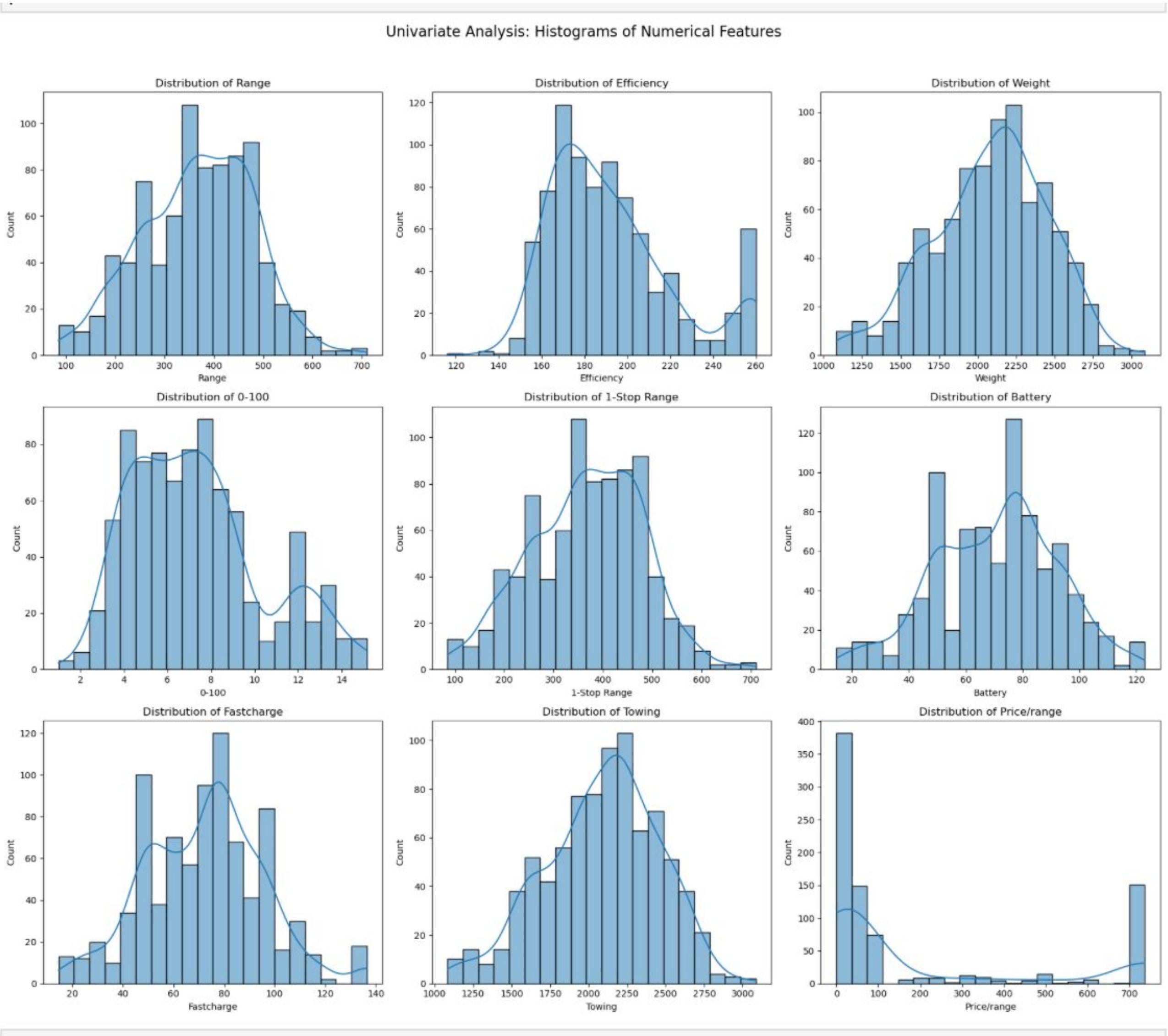
- Positive correlation between **Battery Capacity** and **Range**.
- Brands like **Tesla, BMW, BYD** show better efficiency per battery size.
- Outliers indicate optimization differences among brands.

Observations:

- Heavier EVs generally have **lower range**.
- Compact EVs (e.g., Mini, Fiat) balance efficiency better despite smaller batteries.
- Some luxury EVs offset weight with high battery capacity.

Distribution Insights:

- Most EVs fall within 300–500 km range.
- Efficiency clustered around 15–20 kWh/100 km.
- Few outliers (premium EVs) achieve higher range and efficiency.



INNOMATICS RESEARCH LABS

Data Visualization:

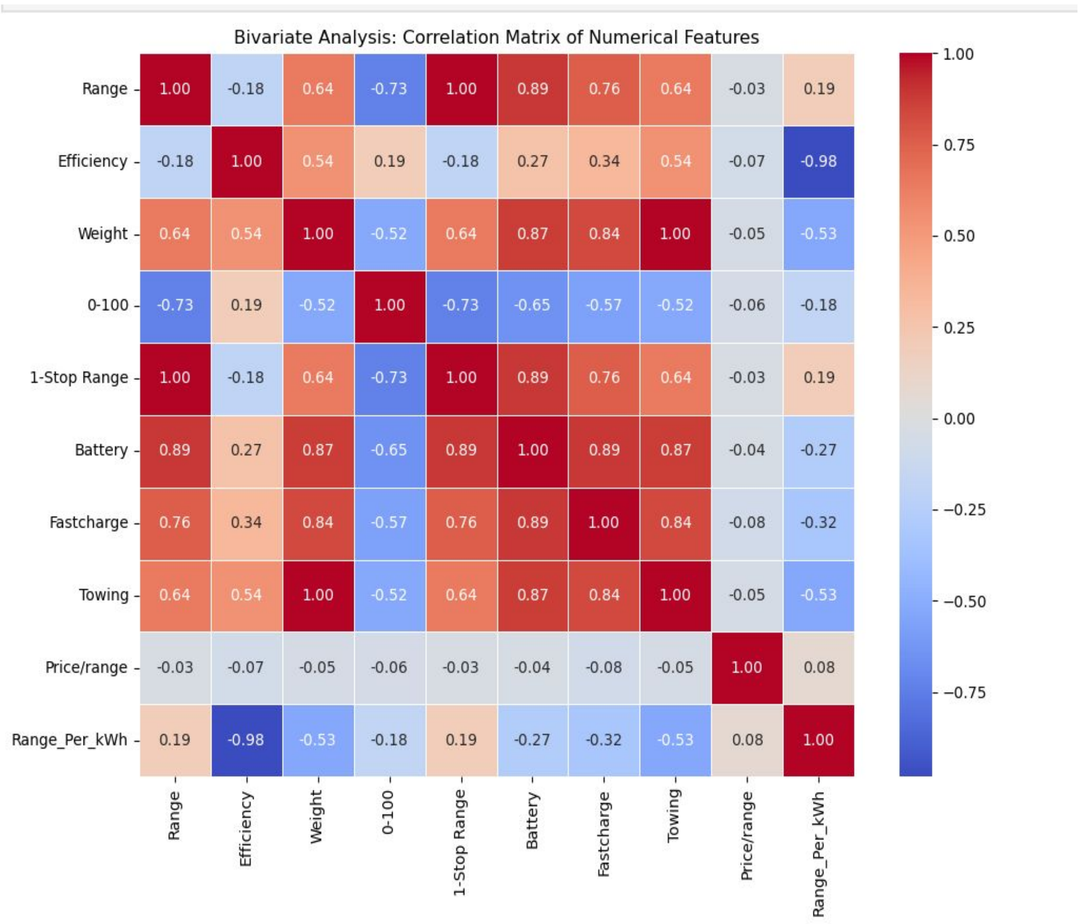
Highlights:

- Tesla and BYD lead in both range and efficiency.
- European brands show stable but moderate range.
- Some new Chinese brands show high battery-to-range ratios.
- **Key Findings**
- EV range grows almost linearly with battery capacity.
- Heavier cars reduce overall efficiency.
- Brands differ in design efficiency (battery utilization).
- Clean data can support research and buying decisions.

Data Visualization:

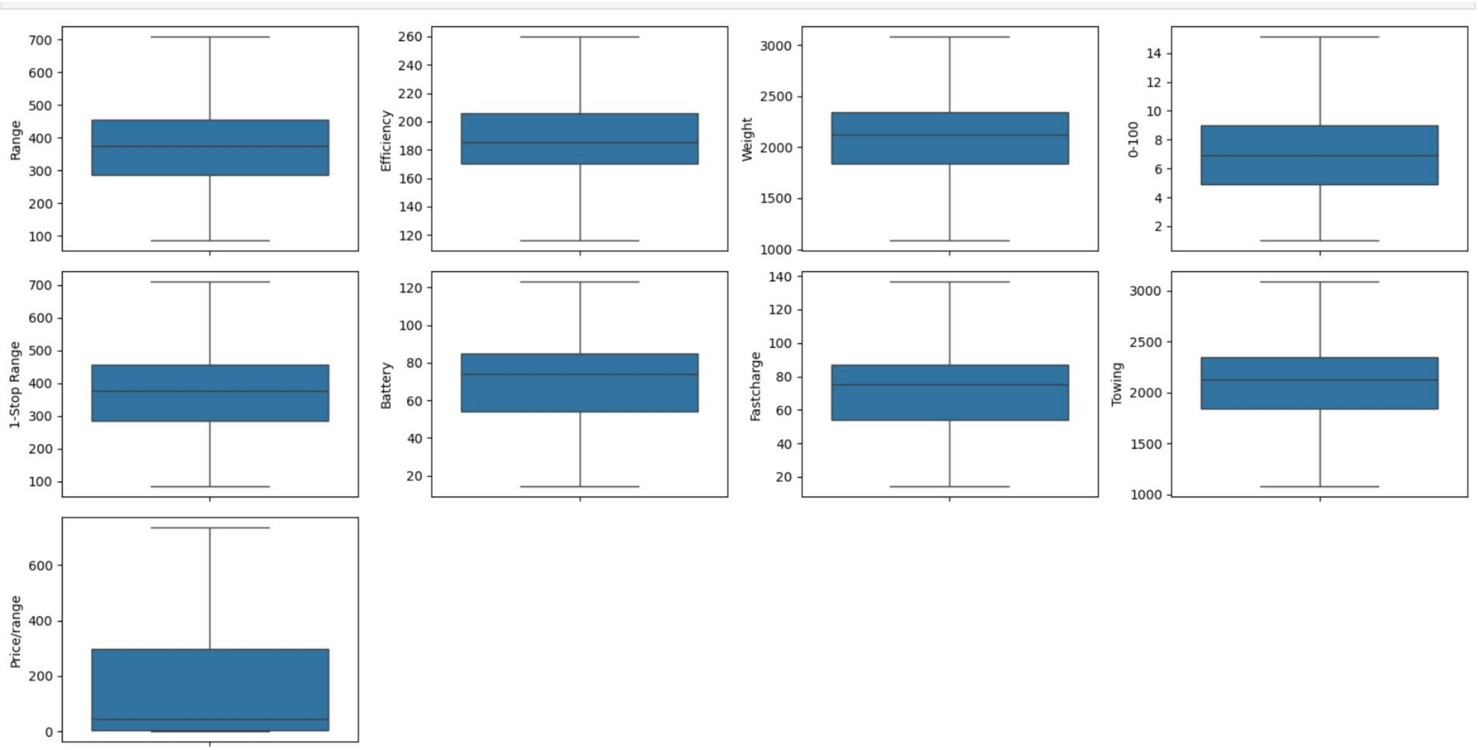
- **Correlation Heatmap**
- **Purpose:** Show how numerical features (Battery, Range, Efficiency, Weight) are related.

  What it shows: Battery and Range have the strongest correlation.



INNOMATICS RESEARCH LABS

Data Visualization:

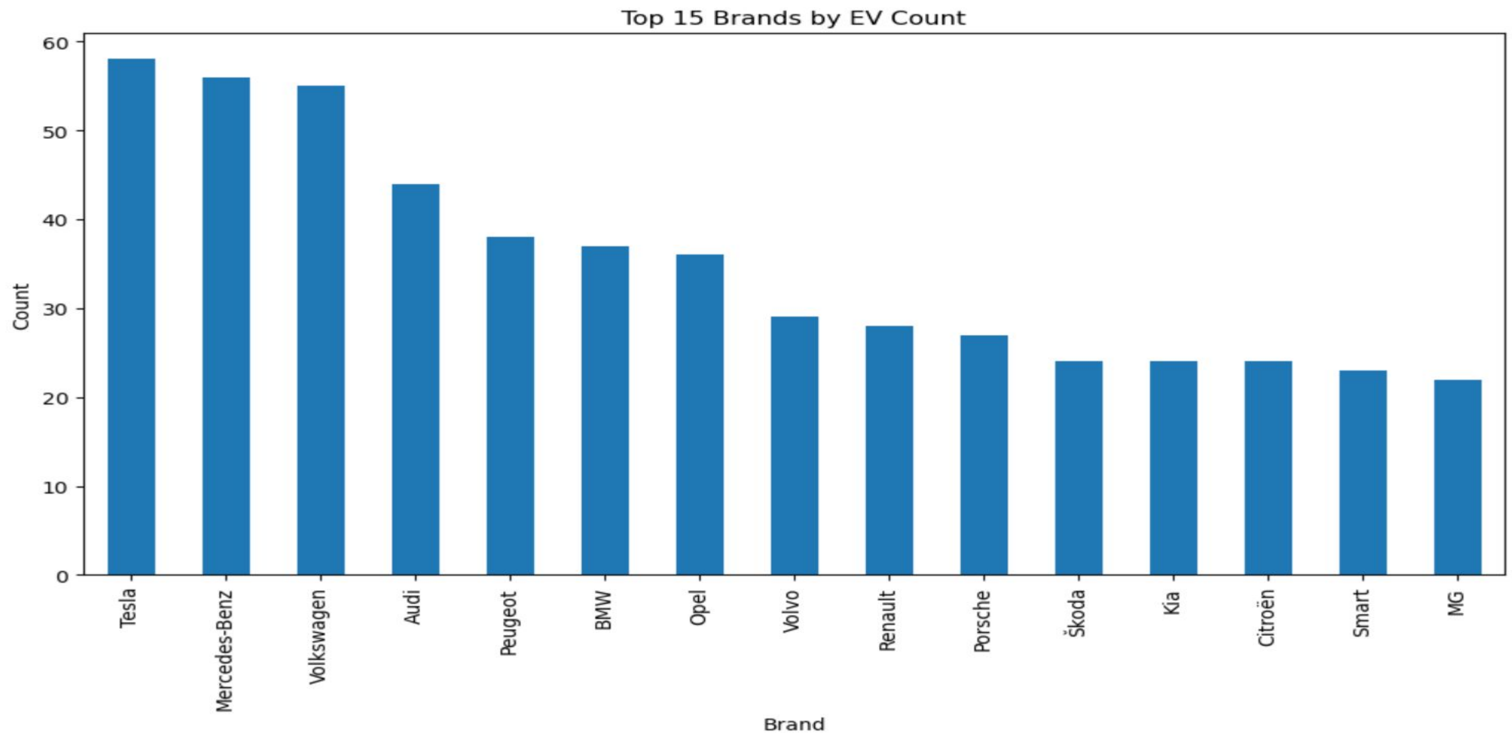


Bi-Variate Analysis — Battery Capacity vs Range

Insights:

- Strong positive correlation (more battery = higher range).
- Premium brands show better range efficiency for same capacity.
- Outliers show design inefficiencies or added weight penalties.

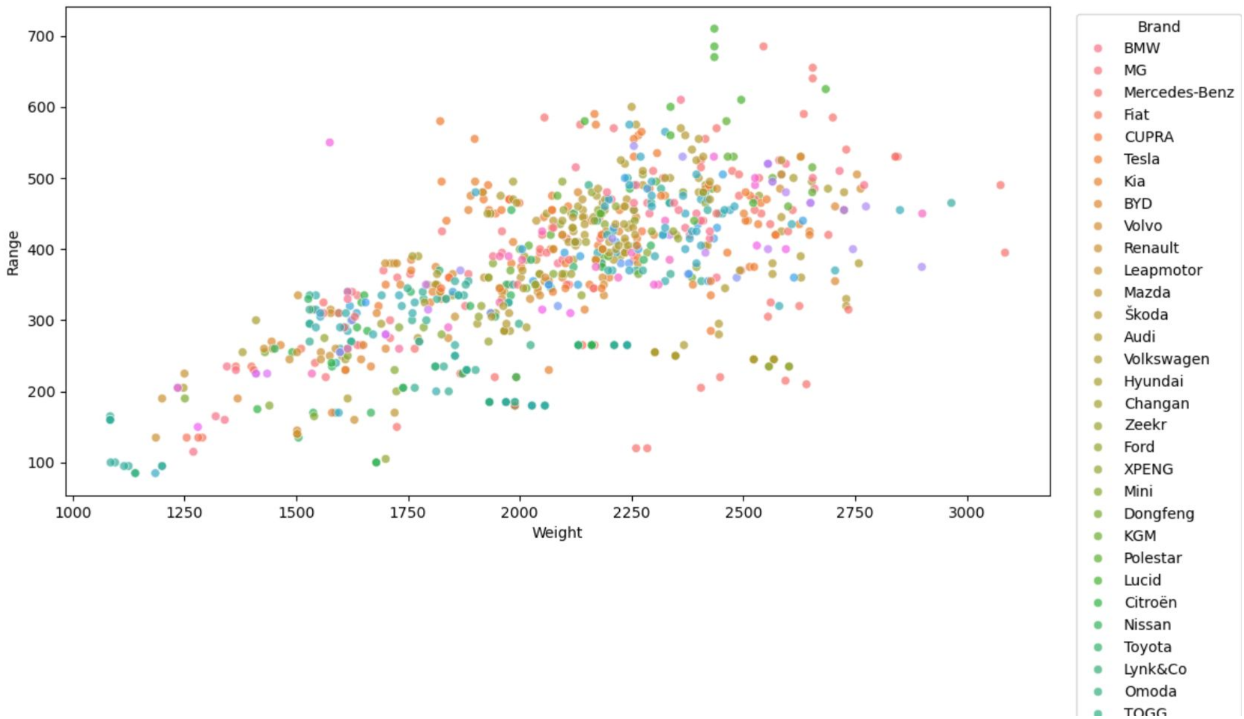
Data Visualization:



1. **Top 15 Brands by EV Count** (Bar chart)
2. **Battery Capacity vs Range** (Scatterplot)

Let’s now expand on that and add **more visualization slides + content descriptions** for your presentation, exactly like your “Data Visualization” template (bullets on the left, charts on the right).

Data Visualization:



Bi-Variate Analysis

- The dense upward trend shows a **strong positive correlation** between *Battery Capacity* and *Range*.
- Brands like **Tesla, BYD, and BMW** consistently appear in the higher range for similar battery sizes, highlighting their efficiency optimization.
- A few outlier points represent EVs that underperform in range, possibly due to heavier body weight or less efficient motor systems.

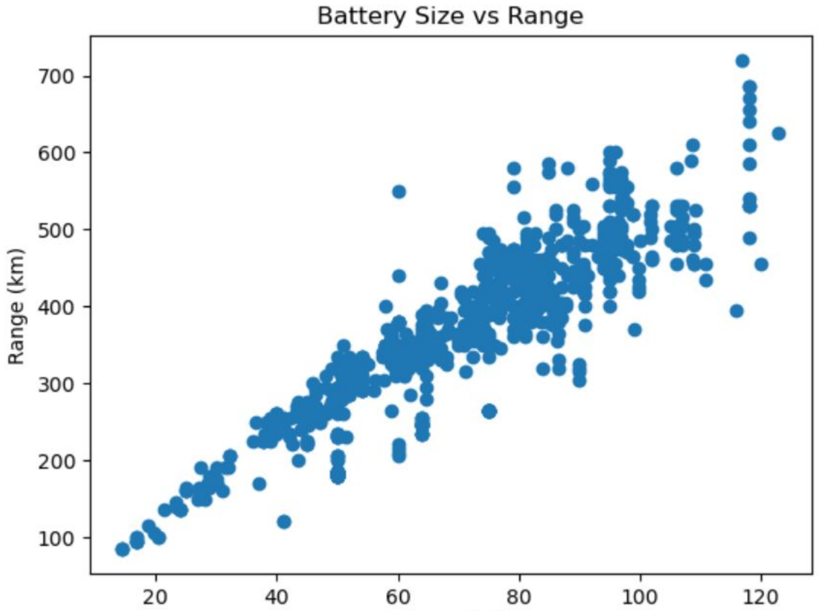
Data Visualization:

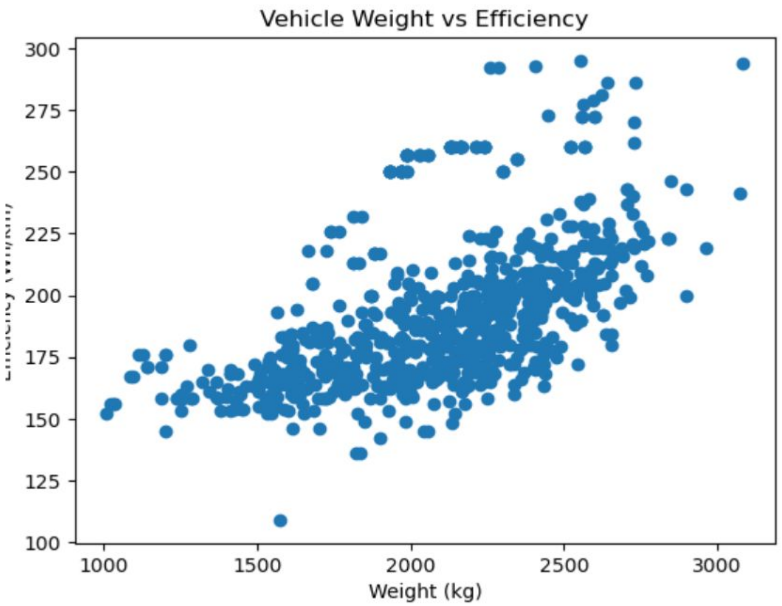
Bi-Variate Analysis

- A clear **positive correlation** is observed between *Battery Size* and *Range*, confirming that larger batteries enable longer driving distances.
- The dense cluster between **40–80 kWh** and **250–500 km** indicates the common capacity range for mid-segment EVs.
- A few outliers with high range per kWh suggest **better energy optimization** and **aerodynamic efficiency** in brands like Tesla and BYD.

Trade-Off Analysis

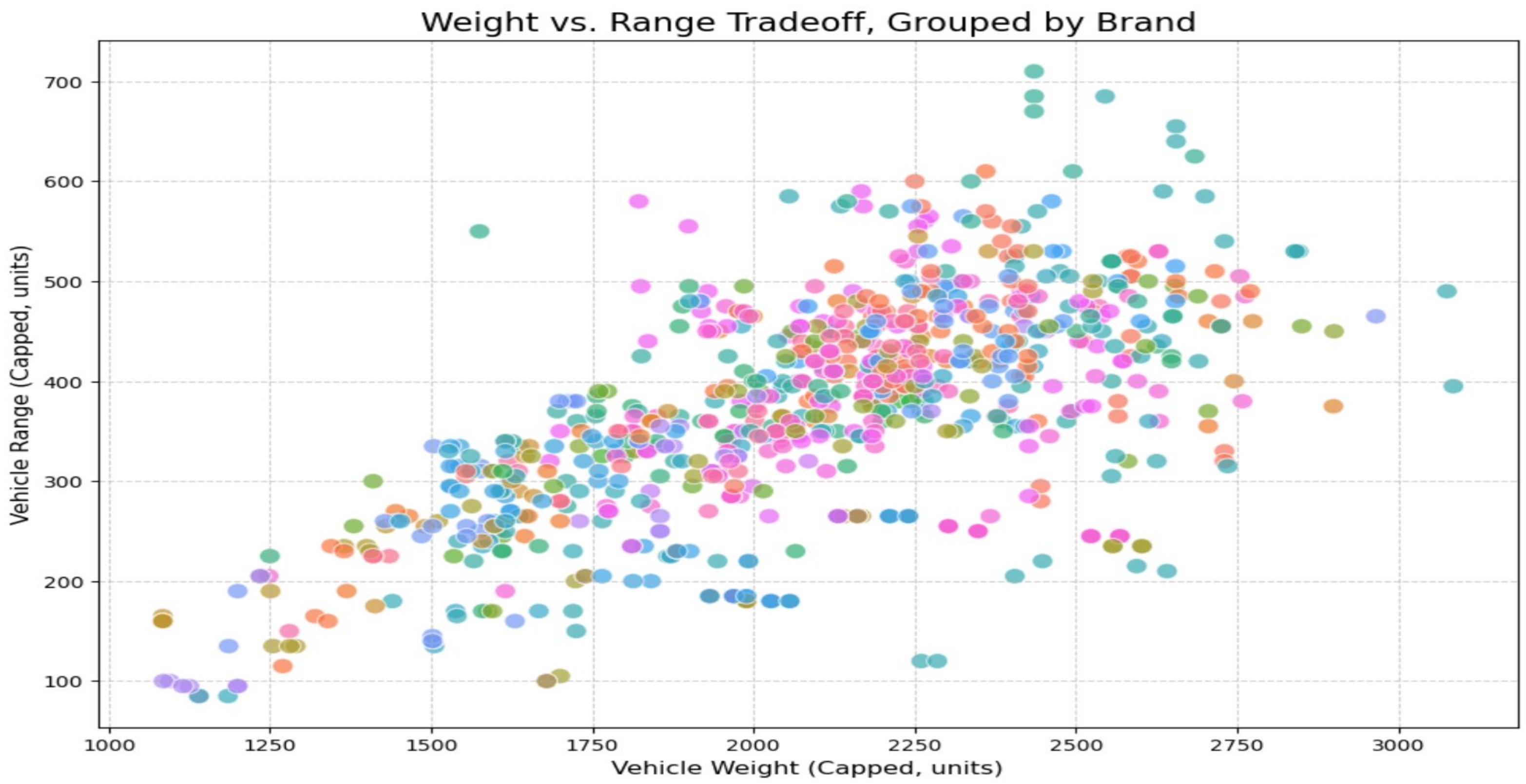
- The upward trend shows that as **vehicle weight increases**, energy consumption (**Efficiency in Wh/km**) also rises, confirming expected physical constraints.
- Lightweight EVs achieve **superior energy efficiency**, consuming less energy per km even with smaller batteries.
- Heavier luxury models exhibit higher Wh/km due to additional features, safety structures, and dual-motor configurations.





INNOMATICS RESEARCH LABS

Data Visualization:



Bi-Variate Analysis – Battery Capacity vs Range

Key Points:

- Dense upward trend → **Strong positive correlation**.
- **Tesla, BYD, BMW** stand out for efficiency with similar battery sizes.
- **Outliers** may represent heavier builds or less efficient motors.
-

INNOMATICS RESEARCH LABS

Key Business Questions

Which EV brands dominate the market in terms of available models and range performance?
→ Helps identify market leaders and emerging competitors.
How does battery capacity impact driving range?
→ Determines whether larger batteries always lead to proportionally higher range or if efficiency plays a key role.
What is the trade-off between vehicle weight and efficiency?
→ Reveals how design and build impact energy consumption and performance.
Which brands offer the most energy-efficient EVs?
→ Useful for comparing technological advancement and optimization among manufacturers.
What are the common battery and range segments in the EV market?
→ Helps manufacturers and consumers understand standard benchmarks for EV performance.
How do fast-charging capabilities vary across brands and battery sizes?
→ Assesses which brands are leading in charging technology and infrastructure readiness.
Are there any outlier EVs that deliver exceptional performance or poor efficiency?
→ Identifies top-tier innovations and underperforming models for further analysis.

Conclusion:

- Successfully **scraped and consolidated Electric Vehicle (EV) data** from multiple online sources into a structured dataset.
- Performed **data cleaning and preprocessing** to ensure consistency across key parameters such as *battery capacity, range, efficiency, and weight*.
- Conducted **exploratory data analysis (EDA)** revealing strong correlations — higher battery capacities generally yield longer ranges, while heavier vehicles reduce efficiency.
- Identified **top-performing brands** (Tesla, BYD, BMW) that consistently deliver optimal balance between battery size, range, and efficiency.
- The project demonstrates how **web scraping and data analytics** can generate valuable insights into **EV market trends, technology performance, and competitive benchmarking**.
- This analysis provides a data-driven foundation for consumers, researchers, and manufacturers to make informed decisions in the growing EV industry.

Q&A

Experience– Web Scraping & Data Analysis:

- **Extracted real-world EV data** from multiple web sources using **Python, Requests, and BeautifulSoup** for automated data collection.
- Applied **Regex and Pandas** for data cleaning — removing unwanted text, converting units, and handling missing or inconsistent values.
- Designed a **structured dataset** containing key EV specifications such as *Battery Capacity, Range, Efficiency, Weight,* and *Fastcharge*.
- Conducted **Exploratory Data Analysis (EDA)** using **Matplotlib** and **Seaborn** to uncover insights and visualize trends in EV performance.
- Identified **key market patterns**, such as correlations between battery capacity and range, and the impact of weight on energy efficiency.
- Delivered findings through **visual storytelling and dashboards**, enabling better understanding of EV market trends and brand competitiveness.

Challenges – Web Scraping & Data Analysis:

- **Inconsistent Website Structure:**
  Different EV listing pages used varied HTML tags and formats, making data extraction rules harder to standardize.
- **Dynamic Web Content:**
  Some sites loaded data dynamically using JavaScript, requiring additional handling or alternative scraping approaches.
- **Data Cleaning Complexity:**
  Extracted values contained mixed units (e.g., *km*, *kWh*, *Wh/km*) and symbols that needed Regex-based cleaning and conversions.
- **Missing & Duplicate Records:**
  Several entries lacked fields like *Efficiency* or *Weight*, demanding manual verification and data imputation.
- **Visualization Challenges:**
  Large number of brands and overlapping data points caused cluttered graphs that required layout tuning and filtering.

INNOMATICS RESEARCH LABS

INNOMATICS RESEARCH LABS