



DATA SCIENCE TOOLBOX: PYTHON PROGRAMMING

PROJECT REPORT

(Project Semester January-April 2025)

**BEYOND THE BOUNDARY:
EXPLORATORY DATA ANALYSIS OF IPL USING PYTHON**

Submitted by

Himanshu Yadav

Registration No. 12309897

Programme and Section. K23DP

Course Code: INT375

Under the Guidance of

Dr. Dheeraj Kapila

UID. 23509

Discipline of CSE/IT

Lovely School of Computer Science and Engineering

CERTIFICATE

This is to certify that Himanshu Yadav, bearing Registration no. 12309897 has completed INT375 project titled, “**Electric Vehicle Population Data Analysis in Washington State using Python**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

Dr. Dheeraj Kapila

Professor

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab.

Date: 10-04-25

DECLARATION

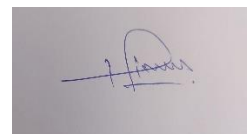
I, Himanshu Yadav, student of Computer Science and Engineering under CSE/IT Discipline at Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 10-04-25

Registration No. 12309897

Signature:

Himanshu Yadav

A rectangular box containing a handwritten signature in blue ink. The signature is stylized and appears to read 'Himanshu'.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Dr. Dheeraj Kapila for their valuable guidance and support throughout the course of this project.

TABLE OF CONTENT

1. Introduction
2. Source
3. EDA Process
4. Correlation
5. Analysis
 - i. Proportion of Battery Electric Vehicles (BEVs) vs Plug-in Hybrid Electric Vehicles (PHEVs)
 - ii. How EV registrations have changed over the year
 - iii. Identify the top 10 EV manufacturers based on registration counts
 - iv. Determine the top 10 most registered EV models using a combination of Make and Model fields.
 - v. Discover the top 10 counties and cities with the highest EV registration numbers.
 - vi. Display the number of EVs registered per model year using a line plot.
 - vii. Show the distribution of BEVs and PHEVs for each year using a count plot (grouped bar chart).
 - viii. Estimate the total annual CO₂ emissions saved by EVs.
 - ix. Compute and visualize correlation between numerical variables using a heatmap.
 - x. Use a T-test to compare the average model year of BEVs vs PHEVs to see if there's a significant difference.
 - xi. Identify vehicle makes with significantly high or low registration counts using **Z**-score analysis.
6. Conclusion
7. Future Scope
8. References

INTRODUCTION:

As the global shift toward sustainable transportation accelerates, electric vehicles (EVs) have become a focal point of environmental and technological advancement. This analysis explores a real-world dataset detailing electric vehicle registrations across different regions, years, makes, and models. By leveraging Python libraries such as Pandas, Matplotlib, Seaborn, and SciPy, the study conducts a comprehensive data-driven exploration to uncover trends, patterns, and insights into EV adoption. The data is first cleaned and prepared to ensure reliability, followed by initial exploratory steps to understand the structure and quality of the dataset.

The core of the analysis investigates several key aspects, including the distribution of EV types, leading manufacturers and models, regional adoption trends, and the evolution of EV usage over time. Visualizations such as bar charts, pie charts, and line graphs are used to make the insights more intuitive and impactful. Additionally, statistical techniques such as hypothesis testing, correlation analysis, and normality checks are applied to derive deeper meaning from the data. A notable inclusion is the estimation of CO₂ emissions potentially saved by EV usage, offering an environmental perspective to complement the market analysis. Together, these insights contribute to a clearer understanding of the current state and growth trajectory of electric vehicle adoption.

.

SOURCE:

The data used in this project was obtained from the https://catalog.data.gov/dataset/?res_format=CSV , Categories covered by this include:

- **VIN (1-10)** – Vehicle Identification Number (partial)
- **County** – The county where the vehicle is registered
- **City** – The city where the vehicle is registered
- **State** – The state of registration (likely all WA)
- **Postal Code** – ZIP code of the registered address
- **Model Year** – Year the EV model was manufactured
- **Make** – Vehicle manufacturer (e.g., Tesla, Nissan)
- **Model** – Specific model name (e.g., Model 3, Leaf)
- **Electric Vehicle Type** – BEV (Battery Electric) or PHEV (Plug-in Hybrid)
- **Clean Alternative Fuel Vehicle (CAFV) Eligibility** – Whether the vehicle qualifies for Washington state incentives
- **Electric Range** – Estimated electric-only driving range
- **Base MSRP** – Manufacturer's suggested retail price
- **Legislative District** – Political district of registration
- **DOL Vehicle ID** – Department of Licensing ID
- **Vehicle Location** – Latitude and longitude of the vehicle's registered location
- **Electric Utility** – The utility company servicing the address
- **2020 Census Tract** – Census data tract code for demographic reference

The raw CSV files were processed using Python to extract relevant attributes for deriving vehicle data, registrations and match environmental impact.

LINK: <https://catalog.data.gov/dataset/electric-vehicle-population-data/resource/fa51be35-691f-45d2-9f3e-535877965e69>

EXPLANATORY DATA ANALYSIS (EDA):

The EDA process was carried out in a structured and categorized manner to reveal information in a systematic order. All steps followed in achieving a comprehensive EDA are listed below coupled with my personal understanding of what the data indicates as deemed necessary.

1. Data Cleaning

The first and most important step in EDA is ensuring the quality of data. In this process:

- The dataset is loaded from a CSV file.
- Rows with missing or incomplete key information—like vehicle type, make, model, location, and model year—are removed. This prevents unreliable results during analysis and ensures that only meaningful and valid data is retained.

2. Initial Data Exploration

After cleaning, we need to understand the structure and content of the dataset:

- Viewing the first and last few records helps check if the data was read correctly.

- Summary functions like `info()` and `describe()` give an overview of data types, value ranges, and statistics (like mean, median, mode, etc.).
 - Checking the shape tells us the number of rows and columns.
- Counting missing values in each column helps assess data completeness.

This stage builds familiarity with the dataset before deeper analysis begins.

3. Univariate Analysis (Single-Variable Trends)

This step involves analysing one variable at a time:

- The proportion of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) is shown using a pie chart. This reveals which type of EV is more common.
 - Distribution of EV types over model years is visualized with a stacked bar chart, showing how each type's popularity has changed over time.
-

4. Brand and Model Popularity

Identifies the top 10 manufacturers by number of EVs registered.

- Constructs a combined "Make + Model" name and finds the top 10 EV models. These bar plots highlight the dominant players in the EV market.
-

5. Geographic Trends

This part focuses on locations where EVs are popular:

- The top 10 counties and cities with the highest number of EV registrations are visualized. This helps identify geographic hotspots of EV adoption.
-

6. Trend Analysis Over Time

To understand the evolution of EV adoption:

- A line chart plots the number of vehicles registered over the years.

- A grouped bar chart compares the number of BEVs and PHEVs registered year by year.

This gives insight into growth trends and shifts in preferences between types.

7. Environmental Impact Estimation

A simple estimation is made of how much CO₂ emissions have been reduced due to the use of EVs:

- It assumes a fixed average mileage and CO₂ emission rate for internal combustion vehicles.
- The total emissions avoided is calculated and presented in millions of kilograms.

This connects the dataset to broader environmental implications.

8. Correlation Analysis

The code then explores relationships between numeric variables:

- A correlation heatmap shows how different numeric features are related (e.g., do newer cars cost more or travel farther?). This is useful to identify possible patterns or trends among numeric data.
-

9. Statistical Analysis

Finally, several statistical tests are applied:

- A t-test compares the average model years of BEVs and PHEVs to see if one type tends to be newer.
- A z-score analysis identifies manufacturers whose EV registration numbers are significantly higher than others (i.e., statistical outliers).

****1. Data Cleaning & Preparation:**** The first objective of the code is to ensure that the dataset is cleaned and organized, making it ready for analysis. Missing values in critical columns such as `Electric Vehicle Type`, `Make`, `Model`, `County`, `City`, and `Model Year` are dropped to avoid inaccuracies or biases in the results. Data cleaning is an essential step, as incomplete or irrelevant data can lead to misleading visualizations or incorrect statistical conclusions. By removing rows with missing values, the dataset becomes consistent and reliable, laying a strong foundation for further analysis.

****2. Initial Data Exploration:**** This objective focuses on understanding the dataset's structure and contents. The `.head()` and `.tail()` methods provide snapshots of the beginning and end of the dataset, helping to grasp its layout and range of entries. The `.info()` method sheds light on data types, null values, and memory usage, giving insight into the dataset's technical composition. Descriptive statistics are produced by `.describe(include='all')`, summarizing measures such as central tendency, dispersion, and frequency for numerical and categorical columns. Additionally, `.isnull().sum()` identifies the exact number of missing values across columns, ensuring no overlooked gaps in data integrity.

```
=== HEAD ===
  VIN (1-10)  County  City  ...  Vehicle Location  Electric Utility  2020 Census Tract
0  5YJ3E1EBXK   King  Seattle  ...  POINT (-122.23825 47.49461)  CITY OF SEATTLE - (WA)  CITY OF TACOMA - (WA)  5.303301e+10
1  5YJYGDEE3L  Kitsap  Poulsbo  ...  POINT (-122.64681 47.73689)  PUGET SOUND ENERGY INC  5.303509e+10
2  KMBKRDAF5P  Kitsap  Olalla  ...  POINT (-122.54729 47.42602)  PUGET SOUND ENERGY INC  5.303509e+10
3  5UXTA6C0XM  Kitsap  Seabeck  ...  POINT (-122.81585 47.64509)  PUGET SOUND ENERGY INC  5.303509e+10
4  JTMAB3FV7P  Thurston  Rainier  ...  POINT (-122.68993 46.88897)  PUGET SOUND ENERGY INC  5.306701e+10
```

```

=== INFO ===
<class 'pandas.core.frame.DataFrame'>
Index: 235689 entries, 0 to 235691
Data columns (total 17 columns):
#   Column                                Non-Null Count  Dtype
---  -
0   VIN (1-10)                            235689 non-null object
1   County                                235689 non-null object
2   City                                  235689 non-null object
3   State                                235689 non-null object
4   Postal Code                           235689 non-null float64
5   Model Year                            235689 non-null int64
6   Make                                  235689 non-null object
7   Model                                 235689 non-null object
8   Electric Vehicle Type                 235689 non-null object
9   Clean Alternative Fuel Vehicle (CAFV) Eligibility 235689 non-null object
10  Electric Range                        235653 non-null float64
11  Base MSRP                            235653 non-null float64
12  Legislative District                  235198 non-null float64
13  DOL Vehicle ID                       235689 non-null int64
14  Vehicle Location                      235682 non-null object
15  Electric Utility                      235689 non-null object
16  2020 Census Tract                    235689 non-null float64
dtypes: float64(5), int64(2), object(10)
memory usage: 32.4+ MB
None

```

```

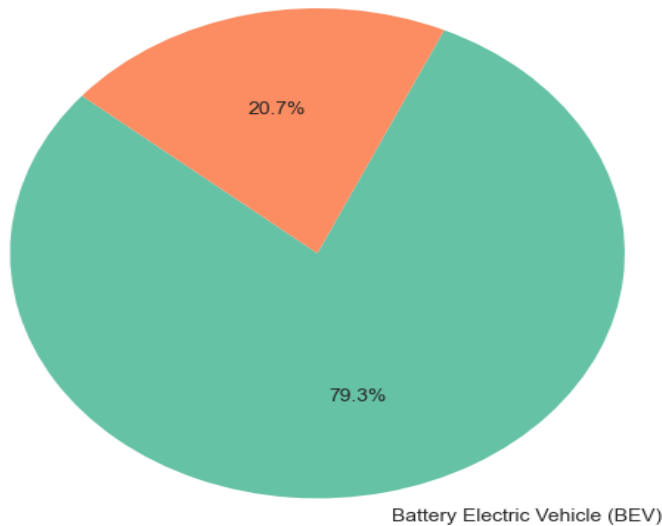
=== DESCRIBE ===
VIN (1-10)  County  City  ...  Vehicle Location  Electric Utility  2020 Census Tract
count      235689  235689  235689  ...  235682  235689  2.356890e+05
unique     13763   212    788    ...  957      76      NaN
top        7SAYGDDE6P  King  Seattle  ...  POINT (-122.13158 47.67858)  PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA)  NaN
freq       1190   118711  37410  ...  5824    85298  NaN
mean       NaN    NaN    NaN    ...  NaN     NaN    5.298066e+10
std        NaN    NaN    NaN    ...  NaN     NaN    1.521066e+09
min        NaN    NaN    NaN    ...  NaN     NaN    1.001020e+09
25%        NaN    NaN    NaN    ...  NaN     NaN    5.303301e+10
50%        NaN    NaN    NaN    ...  NaN     NaN    5.303303e+10
75%        NaN    NaN    NaN    ...  NaN     NaN    5.305307e+10
max        NaN    NaN    NaN    ...  NaN     NaN    5.602100e+10

```

3. Distribution of EV Types: To visualize the distribution between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), this objective utilizes a pie chart. The chart provides a clear visual representation of the proportion of the two EV types within the dataset. This segmentation highlights consumer preferences and the relative popularity of fully electric vs. hybrid technology vehicles, which can influence business strategies or government policies supporting specific EV types.

Proportion of EV Types (BEV vs PHEV)

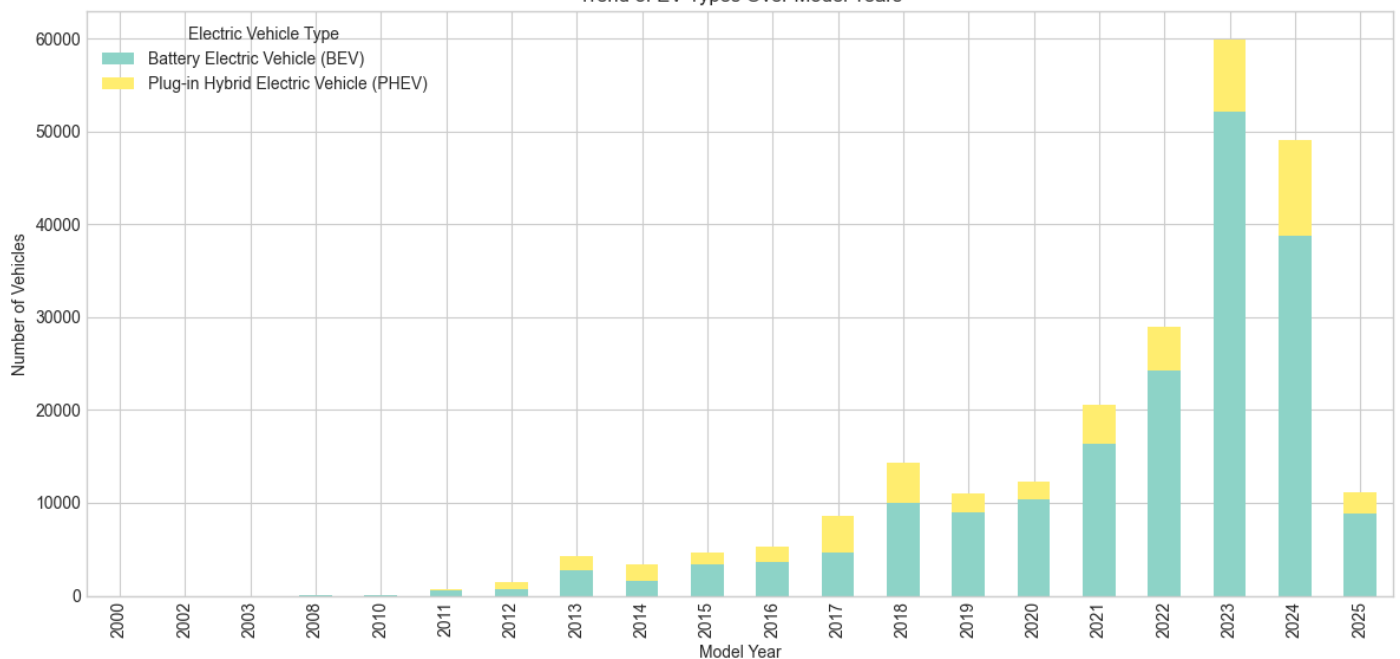
Plug-in Hybrid Electric Vehicle (PHEV)



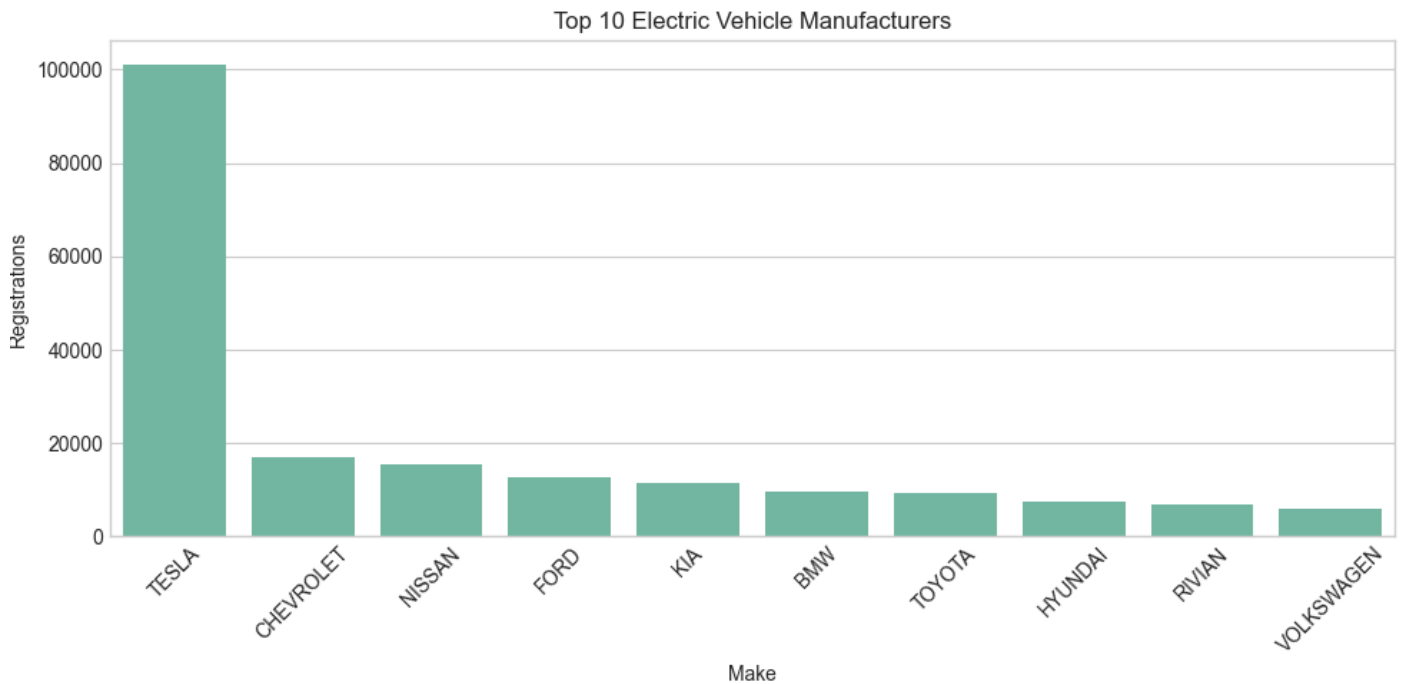
Battery Electric Vehicle (BEV)

****4. Trends Over Time:**** This analysis aims to observe how electric vehicle registrations have evolved over time, particularly focusing on the distinction between BEVs and PHEVs. Using a stacked bar chart, the data is grouped by model year and categorized into the two EV types. This visualization helps identify year-on-year growth or decline, industry milestones, or the impact of external factors such as policy changes or technological advancements.

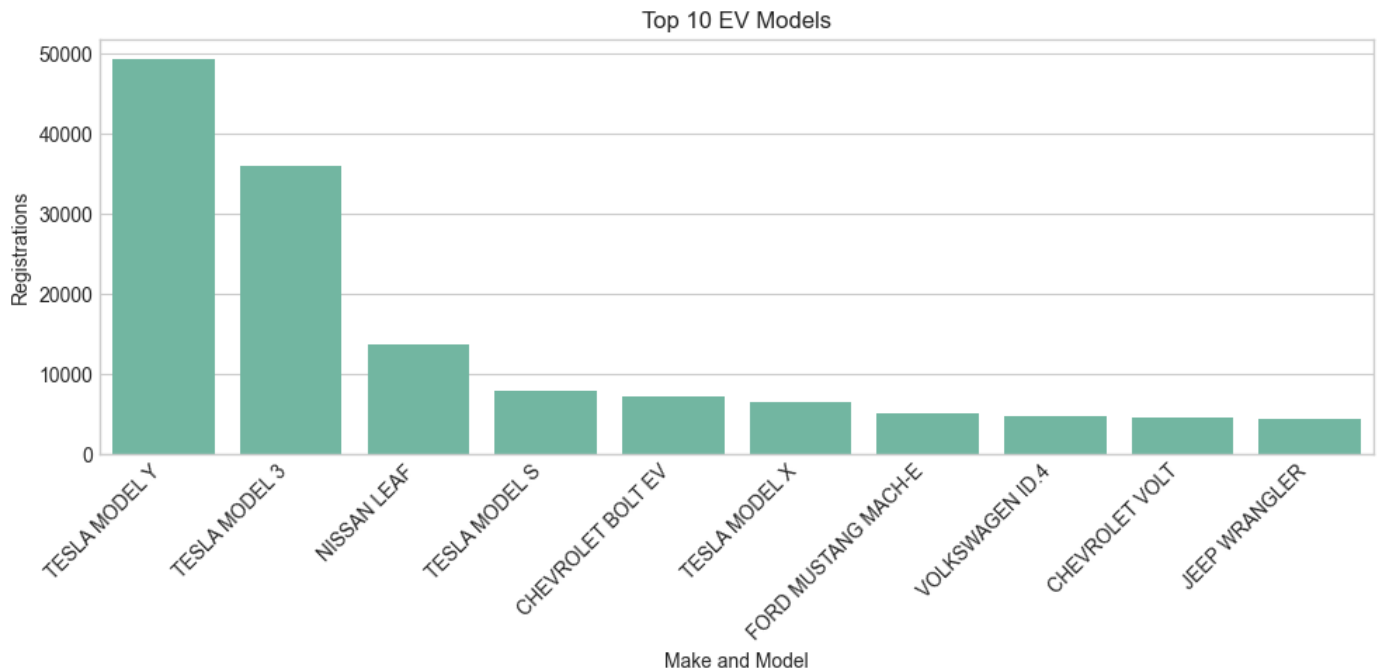
Trend of EV Types Over Model Years



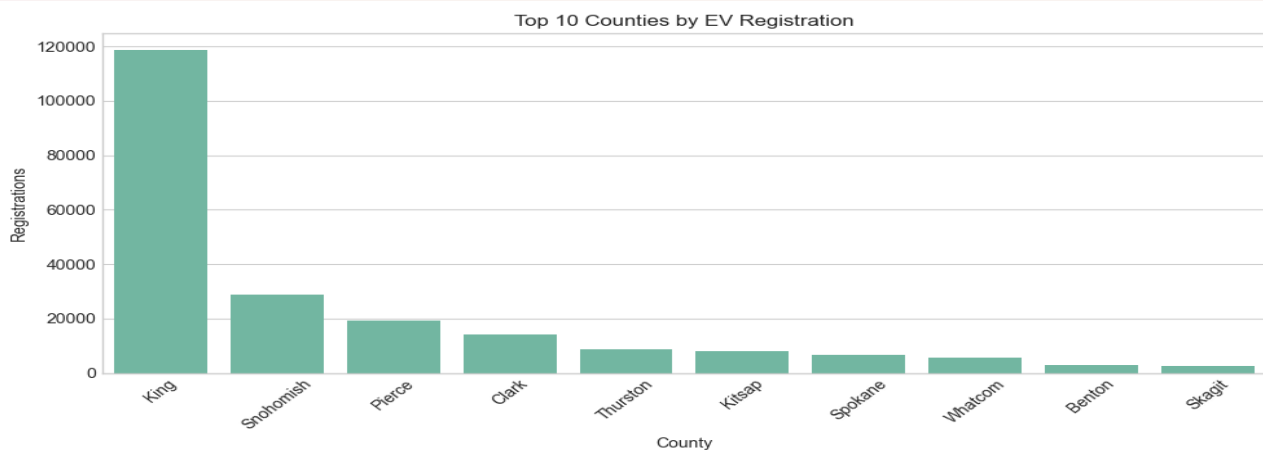
****5. Top Manufacturers:**** Understanding market leaders in the EV space is achieved by identifying the top 10 manufacturers with the highest registrations. A bar plot provides a clear ranking of the most successful companies, showcasing how consumer preferences are distributed among the brands. Insights gained from this analysis can be valuable for competitors, marketers, or policymakers focusing on EV industry growth.



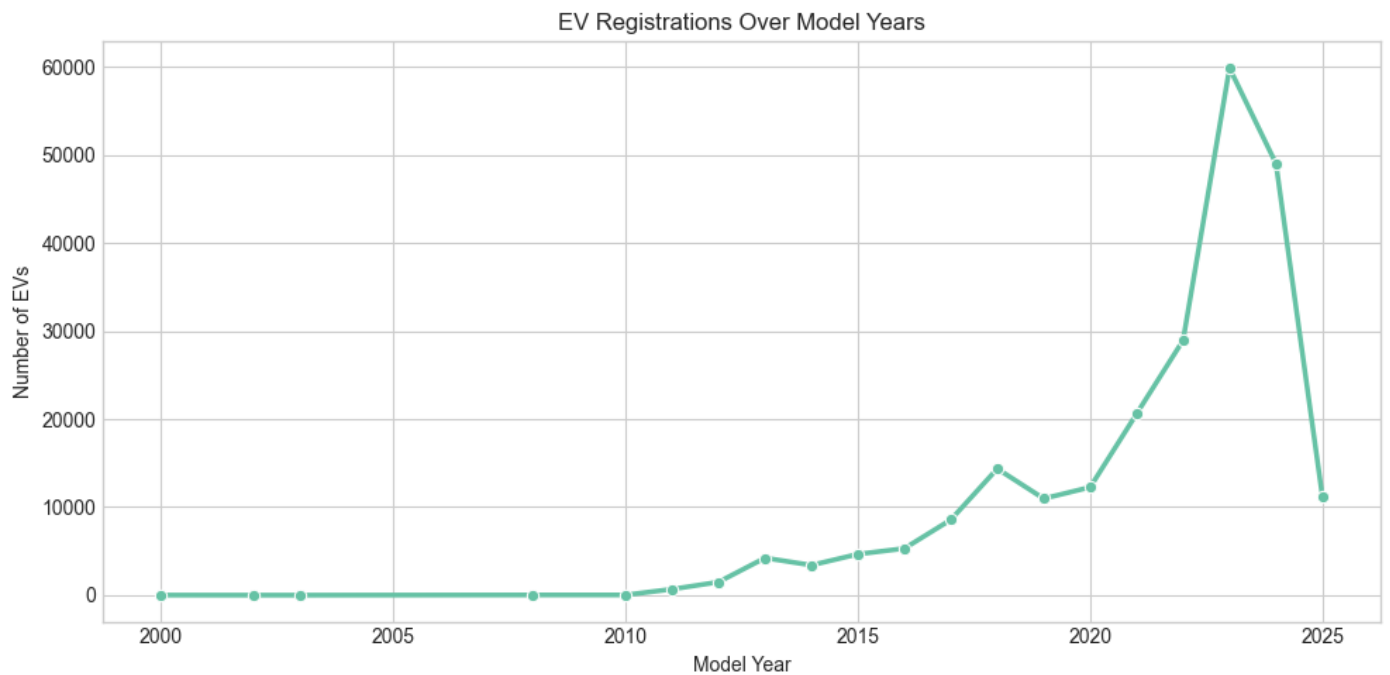
****6. Popular EV Models:**** To gain a more granular understanding of EV preferences, this objective identifies the top 10 registered vehicle models by combining the `Make` and `Model` fields. The resulting bar plot reveals specific configurations that resonate most with customers. By highlighting the most popular EV models, this analysis can guide manufacturers in understanding successful design elements or features.



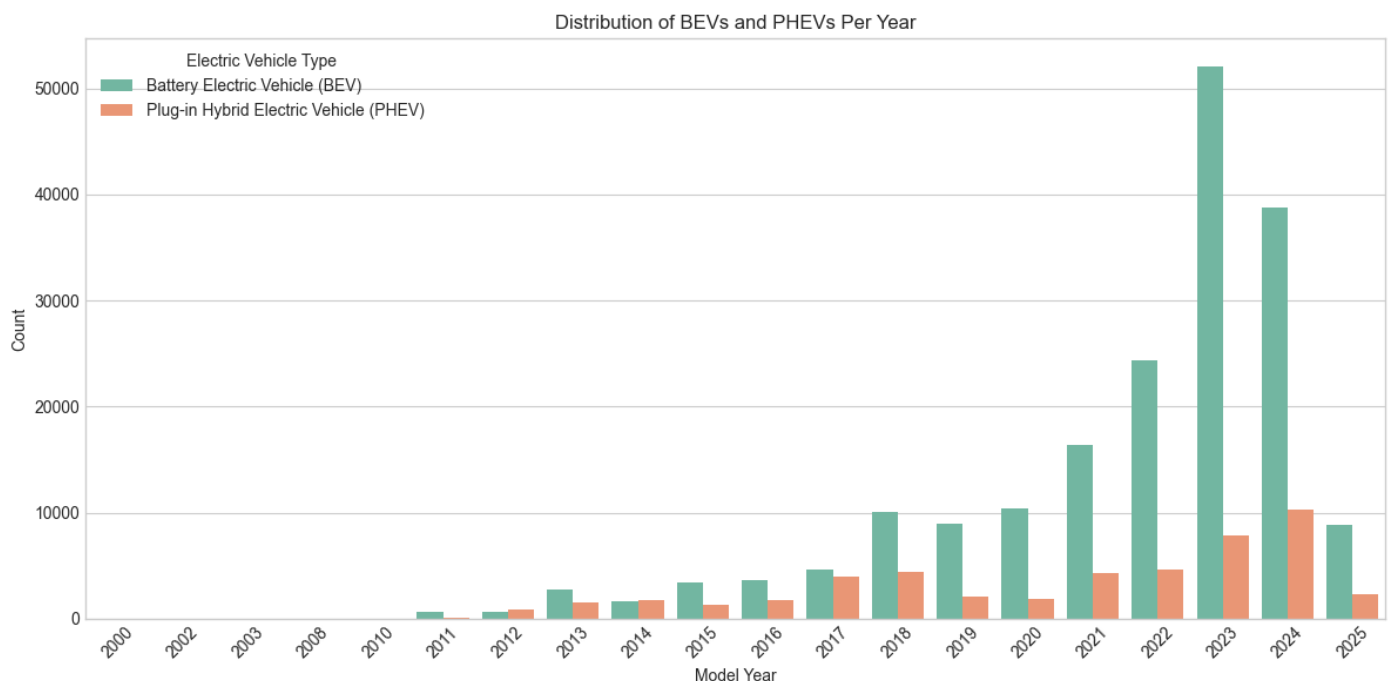
****7. Geographical Analysis:**** This part of the analysis explores the geographic distribution of EV registrations, focusing on top-performing counties and cities. Separate bar plots for each region provide a ranked list of areas with the highest EV adoption rates. These findings can reveal trends, such as urban areas leading adoption efforts, or specific regions responding well to subsidies or charging infrastructure.



****8. Yearly Registration Trend:**** Using a line plot, this analysis examines the overall trend of EV registrations over different model years. This objective highlights whether registrations are increasing, plateauing, or decreasing over time. Peaks or troughs in the trend can point to historical events, technological breakthroughs, or shifts in consumer interest in electric vehicles.

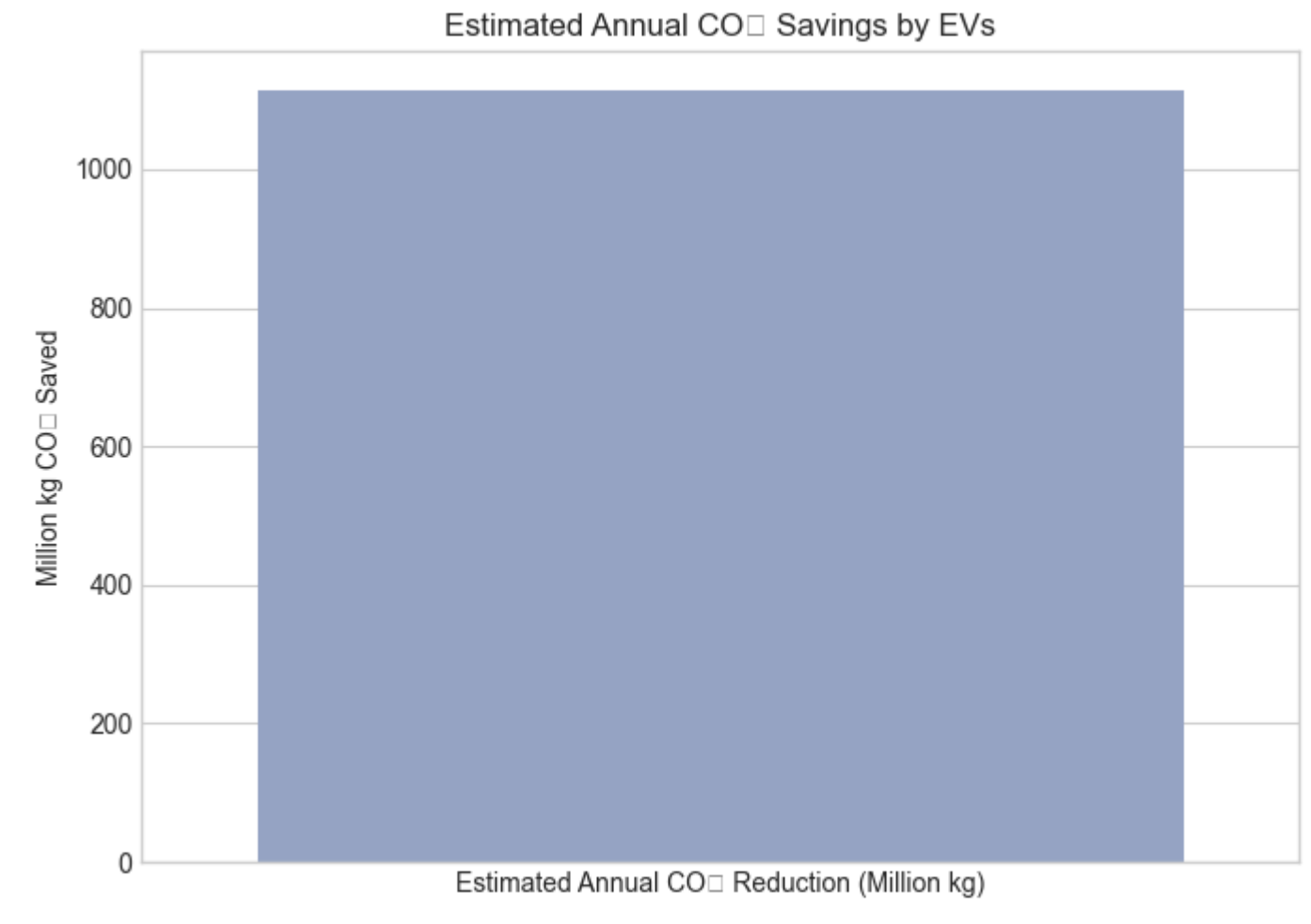


****9. EV Type Distribution by Year:**** The distribution of BEVs and PHEVs over different model years is analysed using a count plot, or grouped bar chart. By splitting the data by year and categorizing by EV type, this visualization shows how the preference for one type of EV has evolved relative to the other. This analysis can provide valuable insights into whether certain policies or advancements have led to shifts in consumer behaviour.

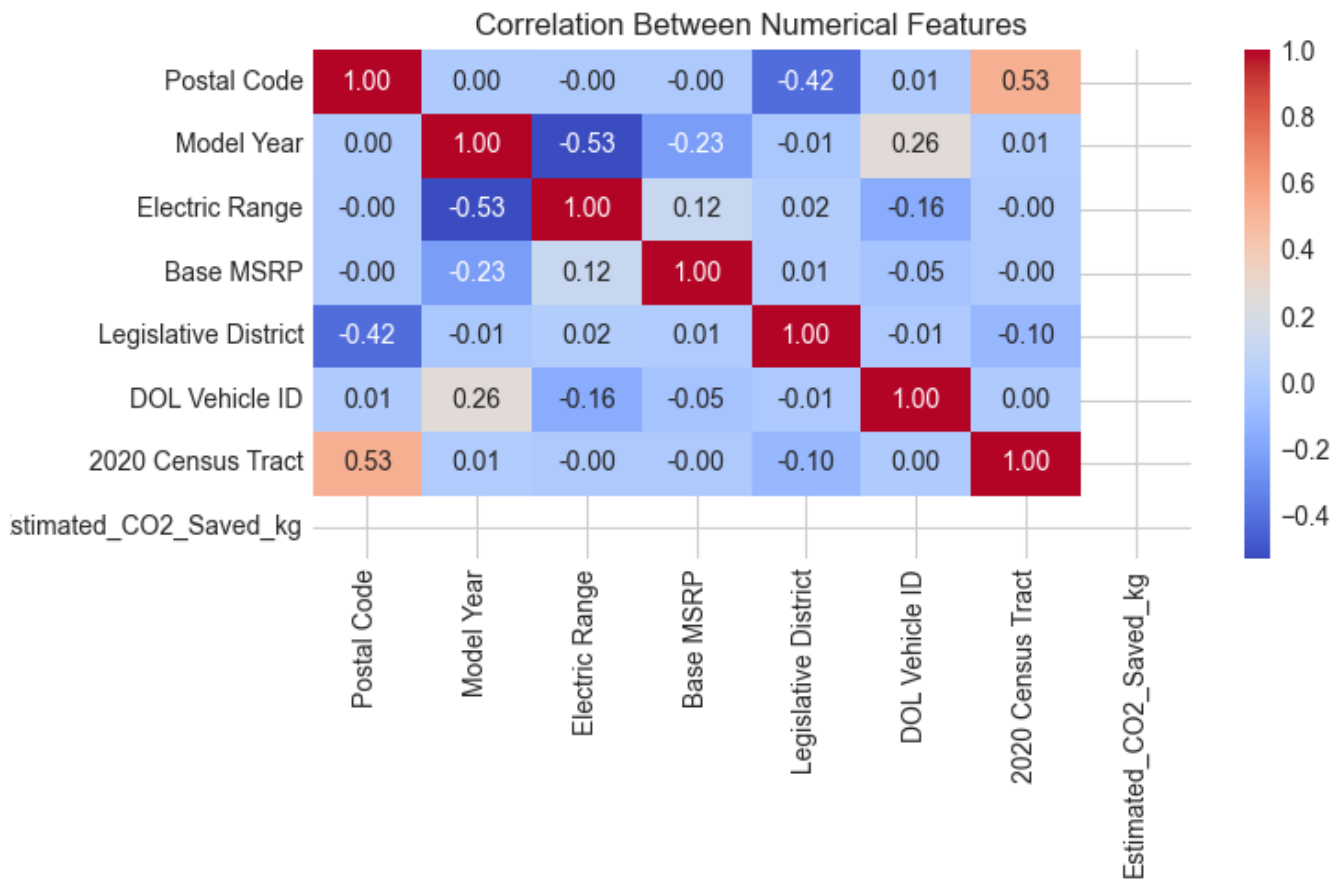


****10. Environmental Impact Estimation:**** To estimate the environmental benefits of electric vehicles, the code calculates total annual CO₂ emissions saved by using EVs instead of internal combustion engine (ICE) vehicles. Assuming average miles driven and ICE emission rates, a new

column computes the CO₂ savings for each vehicle. Summarizing the total CO₂ saved, a bar plot visualizes the positive impact of EVs in reducing emissions, highlighting their role in addressing climate change.



****11. Correlation Analysis:**** The relationship between numerical variables in the dataset is explored using a correlation matrix visualized through a heatmap. By examining pairwise correlation coefficients, this analysis uncovers dependencies or patterns among variables such as model year, vehicle count, or estimated CO₂ savings. Strong correlations can provide actionable insights or guide further hypothesis testing.



****12. Statistical Comparison of Model Years:**** To identify significant differences between BEV and PHEV model years, the analysis performs a T-test. This statistical test compares the average model years of the two EV types under the assumption of unequal variances. The test results, represented by the T-statistic and P-value, help determine whether the difference in model years is statistically significant or likely due to random chance.

****13. Outlier Detection:**** The analysis aims to identify vehicle makes with unusually high or low registration counts using Z-score analysis. By calculating how far each make deviates from the mean registration count in terms of standard deviations, it identifies brands that are statistical outliers. These outliers may represent exceptional performers or anomalies worth investigating further.

```
T-Test Between Model Years of BEVs and PHEVs:
T-statistic: 64.02623610947947
P-value: 0.0

Outlier Vehicle Makes Based on Z-Score Analysis:
Make
TESLA    101078
Name: count, dtype: int64
```

CONCLUSION

In conclusion, this analysis provides a comprehensive exploration of the electric vehicle (EV) population dataset, offering valuable insights into various aspects of EV adoption. By ensuring the data is clean and reliable, the study lays a strong foundation for meaningful examination. Key findings include the dominance of certain EV manufacturers and models, the regional preferences for EV registrations, and clear trends in the shift toward electric mobility over the years. The visualizations, ranging from pie charts to bar plots and heatmaps, effectively present the data, making complex information easy to interpret. The environmental impact analysis, which estimates annual CO₂ savings from EVs, highlights the significant role they play in reducing carbon emissions and combating climate change. Moreover, the statistical analysis, including T-tests and normality checks, adds depth by testing hypotheses and identifying outliers in the data.

Overall, the analysis achieves its objectives by providing a comprehensive, data-driven understanding of EV adoption trends and their implications. It not only uncovers patterns and outliers but also contextualizes the findings within broader environmental and societal impacts. From the growth of battery-electric vehicles (BEVs) to the geographic spread of registrations, the results reflect the rising popularity of EVs and their potential to revolutionize transportation. The inclusion of environmental impact metrics and statistical tests enriches the analysis, offering insights that can inform manufacturers, policymakers, and consumers alike. This study underscores the importance of embracing data analytics in understanding and fostering the adoption of sustainable technologies like electric vehicles.

FUTURE SCOPE

The future scope of this analysis lies in expanding the dataset, integrating additional features, and refining methodologies to gain deeper insights into electric vehicle adoption and its impacts. Incorporating data on charging infrastructure, consumer demographics, government incentives, and technological advancements can enrich the analysis and uncover more nuanced trends. Predictive modeling and machine learning techniques can be applied to forecast EV adoption rates and assess the impact of emerging technologies like solid-state batteries or autonomous driving. Additionally, extending the environmental impact assessment to include metrics like lifecycle emissions or renewable energy usage could provide a broader picture of EV sustainability. This analysis can also serve as a foundation for policymakers and researchers to design strategies that accelerate EV adoption, improve infrastructure, and achieve climate goals. By continuously updating the dataset and

expanding the analytical scope, this study can evolve into a comprehensive tool for shaping the future of sustainable transportation.

REFERENCES

1. Dataset - <https://catalog.data.gov/dataset/electric-vehicle-population-data>.
2. Python Data Analysis Library (Pandas) – <https://pandas.pydata.org>
3. NumPy Documentation – <https://numpy.org/doc>
4. Matplotlib Documentation – <https://matplotlib.org/stable/contents.html>
5. Seaborn Documentation – <https://seaborn.pydata.org>
6. Scipy.stats (for chi-square test) – <https://docs.scipy.org/doc/scipy/reference/stats.html>