Name : - Sonal Kumari

Registration N0. :- 12107709

Roll No. :- RK21URA14

Section :- K21UR [Group – 1]

Course Code :- INT 353

Course Name:- EDA Project

Project on Electric vehicle Population dataset

Program Name:- Computer Science and Engineering

(Specialization in Data Science, AI and ML)

School of Computer Science and Engineering

Lovely Professional University, Phagwara

Acknowledgement

“GOD HELPS THOSE WHO HELP THEMSELVES.”

“ARISE! AWAKE! AND STOP NOT UNTIL THE GOAL IS REACHED.”

Success often requires preparation, hard work, and perspiration. The path to success is a long journey that calls for tremendous effort with many bitter and sweet experiences. This can only be achieved by the Graceful Blessing from the Almighty on everybody. I want to submit everything beneath the feet of God.

I want to acknowledge my regards to my teacher, Ms. Shivangini Gupta, for her constant support and guidance throughout my training. I would also like to thank HOD Ms. Harjeet Kaur, School of Computer Science and Engineering for introducing such a great program.

To my friends, whose companionship has been a source of joy, laughter, and strength: your camaraderie has been a constant source of inspiration. Thank you for being my pillars of support through thick and thin.

To my parents, the foundation of my world, whose unconditional love, sacrifices, and unwavering belief in me have been my driving force: thank you for being my greatest advocates and for instilling in me the values that shape who I am today.

Table Of Contents

Sr.No.	Contents	Page No.
01	Acknowledgement	02
02	Introduction	04
03	Domain Knowledge	04 - 09
04	Reason for this dataset	09 - 11
05	Library Used and Approached	11 - 12
06	Data Description	12-14
07	Data Cleaning	15
08	Data Exploration	15-17
09	Univariate Analysis	17-18
10	Bivariate Analysis	18-19
11	Multivariate Analysis	19
12	Distributions	20-21
13	Hypothesis Testing	21-24
14	Findings and Insights	25-29
15	Limitations	30
16	Recommendations	30
17	Conclusion	31
18	References	31
19	Links	31

INTRODUCTION

I have chosen a dataset from Kaggle. It is about Electric Vehicle Population and this data collected in the year from 1997to 2023.

This dataset contains over 124716 EVs from the USA, along with information on their regional origin, where the EVs are mostly used, and which model year it was made.

It's focused on Electrical Vehicle Type with an aim of appreciating zero emission vehicles, and which state in the USA is using it more, and Clean Alternative Fuel Vehicle (CAFV) Eligibility.

Domain Knowledge

"Amidst the humming electric currents and the whispers of a cleaner, more sustainable future, lies a treasure trove of data—an invaluable insight into the dynamic world of electric vehicle population. Beyond the surface of charging stations and zero-emission promises, this dataset holds the pulse of a rapidly evolving industry."

Electric Vehicle:

Electric vehicles emerge as the possible strategy for decarbonization and green transportation due to social demand. Researchers have made multiple efforts and initiatives as the demand surge for sustainable development in the electric vehicle industry. This study analyzes the relevant research of the industry, thereby explores electric vehicle industry development trends with a scientometrics-based data evaluation system, where three key topics are detected: "Vehicle Exhaust Emissions", “Climate Change”, and “Integration”. The results are visualized in the view of clusters, timeline, and time zone to explore the dynamic direction and future trends of the industry. Further trend detection and forward-looking analysis suggest the significance of stakeholders and their interconnection. In response to the significant challenges in sustainable development, this study proposes the stakeholder engagement system in a comprehensive perspective. The system firstly reveals the driving forces behind the industry development, that is how different motivation and strategies orientate the industry towards sustainability. Then, it furtherly analyzes the commitments and efforts needed from multiple stakeholders, through policy implications, leading factors on demand side, and technology innovation requirements on supply side. The stakeholder engagement system may contribute to stakeholder synergy and resource optimization hence for electric vehicle industry sustainable development and novel ideas to other relevant research fields.

Electric vehicles (EVs) have gained significant popularity in recent years as the world seeks sustainable and environmentally friendly transportation options. The shift towards EVs is driven by concerns about climate change, air pollution, and the finite nature of fossil fuels. This document delves into a dataset containing information about various electric vehicles and their eligibility for Clean Alternative Fuel Vehicle (CAFV) programs. We will explore the dataset's structure, the concept of CAFV eligibility, and potential insights that can be extracted from this data.

1. Global EV Adoption:

- Electric vehicles have been experiencing significant growth in adoption worldwide due to concerns about climate change and a shift towards cleaner transportation options. Governments in many countries have also been offering incentives to promote EV adoption.

2. Types of EVs:

- There are different types of EVs, including Battery Electric Vehicles (BEVs), Plug-in Hybrid Electric Vehicles (PHEVs), and Hybrid Electric Vehicles (HEVs). BEVs run entirely on electricity, while PHEVs can run on both electricity and gasoline.

3. EV Market Leaders:

- Tesla has been a major player in the EV market, with its Model 3 being one of the most popular electric cars globally. Other automakers, such as Nissan (Nissan Leaf), Chevrolet (Chevy Bolt), and BMW (i3, i4), have also contributed to the growing EV market.

4. Charging Infrastructure:

- The availability of charging infrastructure is a crucial factor for EV adoption. Governments and private companies have been investing in expanding charging networks to make EVs more accessible. Charging options include home chargers, public chargers, and fast-charging stations.

5. Environmental Benefits:

- One of the primary drivers of EV adoption is their environmental benefits. EVs produce zero tailpipe emissions, which helps reduce air pollution and greenhouse gas emissions compared to traditional internal combustion engine vehicles.

6. Range and Battery Technology:

- Improvements in battery technology have led to increased EV range and reduced charging times. Newer EV models often feature longer ranges, making them more practical for daily use.

7. Cost of Ownership:

- While the upfront cost of EVs can be higher than that of traditional vehicles, lower operating and maintenance costs can make them more cost-effective over time. This includes lower fueling costs and reduced maintenance needs due to fewer moving parts.

8. Government Incentives:

- Many governments around the world offer incentives to encourage EV adoption. These incentives can include tax credits, rebates, reduced registration fees, and access to carpool lanes.

9. Challenges:

- Despite the growth of the EV market, challenges remain, including concerns about the environmental impact of battery production, limited charging infrastructure in some areas, and the need for continued improvements in battery technology.

10. Future Outlook:

- The EV market is expected to continue growing, with more automakers introducing new electric models and advancements in battery technology. Some experts predict that EVs will eventually become the dominant form of personal transportation as the industry matures.

Significance of Domain Knowledge:

Automakers: Manufacturers can use this data to assess market penetration, consumer preferences, and regional trends. It helps them tailor their product offerings and marketing strategies.

- Market Assessment: Manufacturers can use this dataset to assess the market's current state and predict future trends. They can identify regions with high EV adoption rates and tailor their production and marketing strategies accordingly.
- Product Development: Insights into the electric range and pricing can help manufacturers develop vehicles that align with consumer expectations and budget constraints.

Government and Policy Makers: Government agencies can leverage this data to make informed decisions regarding incentives, infrastructure development, and emissions reduction policies.

- Incentive Programs: Governments can use this data to evaluate the effectiveness of existing EV incentive programs and make adjustments as needed. They can target incentives to regions with lower adoption rates to encourage more EV purchases.
- Infrastructure Planning: Data on vehicle locations can inform decisions about where to deploy charging infrastructure, ensuring convenient access for EV owners.
- Emissions Reduction: Policymakers can assess the environmental impact of EV adoption in different areas and develop policies to reduce greenhouse gas emissions further.
- Legislative Impact: Legislative districts data can help policymakers understand how EV adoption varies across different political regions, influencing policy decisions.

Utilities: Electric utility companies can analyze this data to plan for increased electricity demand due to EV charging and explore opportunities for clean energy integration.

- Electric Grid Planning: Utilities can anticipate increased electricity demand due to EV charging by examining the dataset. This data can guide investments in grid infrastructure and the integration of renewable energy sources to meet this growing demand.

Consumers: Potential EV buyers can use this data to compare different models, assess their eligibility for incentives, and make informed purchase decisions.

- Electric Grid Planning: Utilities can anticipate increased electricity demand due to EV charging by examining the dataset. This data can guide investments in grid infrastructure and the integration of renewable energy sources to meet this growing demand.
- Service Reliability: Understanding the distribution of EVs across their service areas can help utilities ensure reliable electricity supply to meet charging needs.

Researchers: Academics and researchers can use this dataset to conduct studies on EV adoption, environmental impact, and energy consumption patterns.

- Academic Studies: Researchers can use this dataset for various academic studies, including analyzing the impact of EV adoption on air quality, conducting energy consumption studies, and investigating the correlation between demographic factors and EV ownership.

Reasons To choose this dataset

Choosing an electric vehicle (EV) population dataset for a real-life project can offer significant value for consumers in future, here are some reasons why it is important :

1. Environmental Impact Assessment: EV population data can be used to assess the environmental impact of electric vehicles in a specific region or country. This information is crucial for policymakers, environmental organizations, and researchers studying climate change and air quality.

2. Energy Planning: Governments and utility companies can use EV population data to plan for the increased electricity demand resulting from the widespread adoption of electric vehicles. This data helps in upgrading the electric grid infrastructure to support charging needs efficiently.

3. Transportation Planning: City planners and transportation authorities can utilize EV population data to make informed decisions about the placement of charging infrastructure, such as public charging stations, to support the growing number of electric vehicles.

4. Market Research: Businesses in the automotive and energy sectors can analyze EV population data to understand market trends, consumer preferences, and the demand for specific types of electric vehicles. This helps in product development and marketing strategies.

5. Incentive Programs: Governments and local authorities often offer incentives to promote EV adoption. Access to EV population data helps in targeting these incentives effectively and ensuring they reach the intended audience.

6. Grid Management: Electric utilities can use EV population data for load forecasting and demand management. This information helps in optimizing energy generation and distribution to meet the charging needs of EV owners efficiently.

7. Emissions Reduction: Monitoring the growth of electric vehicle populations allows governments to track progress toward reducing greenhouse gas emissions and air pollution, which is essential for achieving environmental goals.

8. Urban Planning: City planners can use EV data to design and implement urban mobility solutions that integrate electric vehicles into public transportation systems and reduce traffic congestion and emissions.

9. Research and Innovation: Researchers and academics can use EV population data to conduct studies on various aspects of electric vehicles, from battery technology to consumer behavior, leading to advancements in the EV industry.

10. Economic Impact: Understanding the growth of the electric vehicle market and its associated industries (e.g., battery manufacturing, charging infrastructure) can provide insights into the economic impact of this sector on a local, regional, or national level.

Libraries used and approched

1.**Warnings: -** This library is used to manage warnings in Python. In the code, warnings.filterwarnings('ignore') is used to ignore warning messages.

2.**Pandas: -** It is a Python library used for working with data sets. It has functions for analyzing, cleaning, exploring, and manipulating data, like pd.read_csv (for reading csv files), cleaning the empty rows and columns.

3. **Numpy** **:-** NumPy is a fundamental library for numerical and mathematical operations in Python. It provides support for arrays and

matrices. While it's not explicitly used in the code of dataset analysis, it is often used in conjunction with pandas for numerical computations.

4. **Matplotlib. Pyplot: -** It is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc. Like as plt.plot(), plt.show(), plt.figure().

5.**Seaborn: -** It is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics, like sns.pairplot(for bivariate analysis), sns.boxplot(for boxplot), sns.countplot(), sns.subplot(), etc.

6.**SciPY: -** SciPy is a scientific computation library that uses NumPy underneath. SciPy stands for Scientific Python.It provides more utility functions for optimization, stats and signal processing.

These are the libraries used in my datasheet which are used to provide insights of dataset Electrical Vehicle Population.

Data Description

Electric Vehicle Data Overview:

The dataset under consideration contains information related to various aspects of electric vehicles, primarily in the states of Washington (WA) and California (CA). Here are the key components and their significance:

- **Vehicle Identification Number (VIN)**: The first 10 characters of the VIN serve as a unique identifier for each vehicle. VINs are critical for tracking and managing vehicle data, including manufacturing details, recalls, and ownership history.
- **County, City, State, and Postal Code**: These fields provide geographical information about the location of electric vehicles. Analyzing the distribution of EVs across counties and cities can help identify regions with higher adoption rates and potential areas for infrastructure development.
- **Model Year, Make, and Model**: These attributes specify details about the electric vehicle, including its year of manufacture and the manufacturer's brand and model name. This information is essential for categorizing and comparing different EVs.
- **Electric Vehicle Type**: This field distinguishes between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs). BEVs rely solely on electric power, while PHEVs have both electric and gasoline power sources.
- **Clean Alternative Fuel Vehicle (CAFV) Eligibility**: CAFV eligibility indicates whether a vehicle qualifies for incentives and benefits associated with clean alternative fuels. Some vehicles may not be eligible due to factors like low battery range.
- **Electric Range**: The electric range represents the maximum distance an EV can travel on a single electric charge. It is a critical factor for consumers, as it determines the practicality and usability of the vehicle for their needs.
- **Base MSRP**: The Manufacturer's Suggested Retail Price (MSRP) reflects the vehicle's cost before additional fees and taxes. Analyzing the MSRP helps in understanding the affordability of electric vehicles.

- **Legislative District**: This field provides information about the legislative district associated with the vehicle's location. Legislative districts play a role in policymaking and infrastructure development.
- **DOL Vehicle ID:** This is a unique identifier for each vehicle in the dataset, facilitating data management and retrieval.
- **Vehicle Location**: The geographical coordinates (latitude and longitude) pinpoint the precise location of each electric vehicle. This information can be used for mapping and spatial analysis.
- **Electric Utility**: This field identifies the electric utility company serving the vehicle's location. It is important for understanding the sources of electricity powering EVs.
- **2020 Census Tract**: Census tracts are geographic areas used for demographic analysis. This information can be valuable for studying EV adoption patterns in different demographic contexts.

General Information :

This dataset contains 124716 entries.

Each entry has 17 columns.

There are 988 missing values in the dataset.

This dataset covers the electric vehicle population from 1997 to 2023.

Data Types:

- There are 7 columns: “Postal_Code”, “Model_Year”, “Electric_Range”, “Base MSRP”, “Legislative_District”, “DOL Vehicle_ID”, “2020 Census_Tract”.
- There are 10 columns: “VIN (1-10)”, “County”, “City”, “State”, “Make”, “Model”, “Electric_Vehicle_Type”,

“Clean_Alternative_Fuel_Vehicle (CAFV) Eligibility”, “Vehicle_Location”, “Electric_Utility”.

Data Cleaning

- ♦ **Handling Missing Values: -**This includes handling missing values in specific columns such as 'Model', ‘Electric_Utility’,‘Legislative_District’ and 'Vehicle_Location'.
- ♦ The missing values are handled in a way that is Missing values are either filled with appropriate values (e.g., mode) or, in the case of 'Electric_Utility', rows with missing values are dropped from the dataset.
- ♦ **Exploring Unique Value: -** The code explores unique values in specific columns using nunique() and value_counts() functions. For example, it checks the number of unique values in the 'Electric_Utility' column and counts the occurrences of each 'Electric_Vehicle_Type'.
- ♦ **Handling Duplicates: -** The code handles duplicates in the dataset by duplicating the first row and appending it to the data, creating a dataset with duplicated rows.
- ♦ **Outlier Detection and Handling: -**
- Boxplot Visualization: The code creates boxplots to visualize the distribution of the "Model_Year" feature and identify potential outliers using quartiles and the Interquartile Range (IQR).
- Outlier Removal: It filters out outliers by defining upper and lower bounds based on the IQR and removes data points outside these bounds.

Data Exploration

❑Summary Statistics:

Model_Year = This dataset contains models of EVs from the year 1997 to 2023.

Electric_Range:

Minimum = 0.000000

Maximum = 337.000000

Std = 100.331969

Mean = 79.471936

25% Percentile = 0.000000

50% Percentile = 25.000000

75% Percentile = 200.000000

Base _ MSRP:

Minimum = 0.000000

Std = 10053.289929

Mean = 1556.068906

25% Percentile = 0.000000

50% Percentile = 0.000000

75% Percentile = 0.000000

Maximum = 845000.000000

❑ Data Visualization:

♦ **Count of Vehicles by Make:** The count plot of vehicles by make provides a quick overview of the distribution of different car brands in the dataset. The rotation of x-axis labels enhances readability.

- ♦ **Scatter Plot of Electric Range vs. Base MSRP:** The scatter plot effectively visualizes the relationship between electric range and base MSRP, with color-coded points for different electric vehicle types. This helps in understanding the distribution and potential trends.
- ❖ **Pie Chart of Electric Vehicle Type Distribution:** The pie chart illustrates the distribution of electric vehicle types, providing a clear percentage breakdown.
- ❖ **Pie Chart of Clean Alternative Fuel Vehicle (CAFV) Eligibility:** The pie chart visualizes the distribution of CAFV eligibility.
- ❖ **Count Plot of TESLA Cars by Model:** This count plot specifically focuses on TESLA cars, providing insights into the distribution of TESLA models.
- ❖ **Count Plot of HYUNDAI Cars by Model:** Like the TESLA count plot, this visualizes the distribution of HYUNDAI car models. The rotation of x-axis labels improves readability.
- ❖ **Count Plot of Hybrid Cars by Make:** This count plot filters data for hybrid cars and displays their distribution by make.
- ❖ **Top 10 Counties by Car Sales:** The bar plot effectively shows the top 10 counties with the highest car sales.
- ❖ **Bar graph of Legislative District Distribution:** The bar graph visualizes the distribution of vehicles based on legislative districts.

Univariate Analysis

Univariate Analysis involves examining the distribution and characteristics of a single variable at a time.

Approaches: -

- Histogram (sns.histplot()): Histograms are created to visualize the distribution of the "Electric_Range" feature. It provides insights into the frequency distribution of electric vehicle ranges.

- Kernel Density Plot (KDE Plot) (sns.kdeplot()): Kernel Density Plots are used to visualize the probability density of "Electric_Range" by "Make." This helps understand the density of electric vehicle ranges for different car manufacturers.
- Rug Plot (sns.rugplot()): Rug plots are used in combination with histograms to visualize the distribution of "Model_Year." Rug plots show individual data points as ticks along the axis.
- Box Plots (sns.rugplot()): Box plots are generated for multiple numeric columns to visualize their distributions and identify potential outliers.
- Violin Plots (sns.violinplot()): Violin plots are used to visualize the distribution of numeric features, providing more information about the probability density at different values.
- Strip Plots (sns.stripplot()): Strip plots are created to visualize the distribution of numeric features by displaying individual data points as strips along the axis.

→

Categorical Features Analysis (COUNT PLOT and PIE CHART):

- Count Plot (sns.countplot()): Count plots are used to visualize the distribution of categorical features with fewer unique values. They count the occurrences of each category and display the frequency.
- Pie Chart (sns.pieplot()): A pie chart is generated to display the distribution of "Model_Year" categories as a percentage of the whole.

Bivariate Analysis

Bivariate analysis involves exploring the relationship between two variables.

Approaches:-

- Scatter Plots (plot.scatter()): Used to visualize the relationship between two numeric variables. It helps identify correlations or patterns between "Postal_Code" and "Model_Year," as well as "Electric_Range" and "Postal_Code."
- Line Chart (plt.plot()): Shows how one numeric variable ("Electric_Range") changes concerning another ("Postal_Code").
- Heatmap (sns.heatmap()): Visualizes the correlation matrix of all numeric variables, indicating the degree of correlation between pairs of variables.
- Hexbin Plot: Used to visualize the density of data points with "Electric_Range" less than 100 against "Postal_Code."
- Box Plot (sns.boxplot()): Visualizes the distribution of "Model_Year" across categories of "Clean_Alternative_Fuel_Vehicle (CAFV) Eligibility."
- Violin Plot (sns.violinplot()): Shows the distribution of "Model_Year" across different "Legislative_Districts."

Multivariate Analysis

Multivariate analysis involves the simultaneous analysis of two or more variables to understand their relationships and interactions.

- Pairplot (sns.pairplot()): Visualizes pairwise relationships between multiple numeric variables ("Postal_Code," "Model_Year," "Electric_Range," "Base_MSRP," "Legislative_District," and "DOL Vehicle_ID"). It helps identify correlations and patterns among these variables.

Distributions

Power Law Distribution :

- The power-law distribution is used to model heavy-tailed distributions. My code generates random samples and plots a histogram along with the power-law probability density function.
- Interpretation: The histogram and the power-law curve suggest whether your data follows a power-law distribution.

Binomial Distribution:

- Binomial distributions model the number of successes in a fixed number of independent Bernoulli trials. My code generates random samples and creates a bar plot.
- Interpretation: Check if the observed distribution aligns with the expected binomial distribution.

Uniform Distribution :

- Uniform distributions model events where all outcomes are equally likely. My code generates random samples and visualizes the uniform distribution.
- Interpretation: Ensure the data spreads uniformly across the defined interval.

Chi-Square Distribution:

- The chi-square distribution is commonly used in hypothesis testing. My code generates random samples and compares the histogram with the theoretical chi-square distribution.
- Interpretation: Assess how well the observed data fits the chi-square distribution.

F-Distribution:

- The F-distribution is often used in ANOVA. My code generates random samples and compares the histogram with the theoretical F-distribution.
- Interpretation: Evaluate how well the data matches the F-distribution.

Hypothesis Testing

Normality Testing using Shapiro-Wilk Test:

- The Shapiro-Wilk test assesses whether a sample comes from a normal distribution.
- Interpretation: My data is identified as not following a normal distribution.

Correlation Tests - Pearson and Spearman’s Rank Correlation:

- We use Spearman and Pearson correlation tests to assess the relationship between two variables.
- Interpretation: The tests suggest independence between the selected samples.

Contingency Table Creation:

- Use pd.crosstab to create a contingency table (contingency_data) that represents the frequency distribution of combinations of the two categorical variables: 'Clean_Alternative_Fuel_Vehicle (CAFV) Eligibility' and 'Make'.

Chi-square Test:

- Use chi2_contingency to perform the chi-square test on the created contingency table.Retrieve the chi-square statistic (stat), p-value (p), degrees of freedom (dof), and expected frequencies (expected).
- Interpretation: Print the chi-square statistic and p-value. If the p-value (p) is greater than 0.05, print 'independent categories', indicating that there is no significant association between the two categorical variables.
- If the p-value is 0.05 or less, print 'dependent categories', suggesting a significant association between the two categorical variables.

T-test:

One-Sample-Test:

- Use ttest_1samp to perform a one-sample t-test on the electric range data with the hypothesized mean of 79.47.
- Retrieve the t-statistic (tset) and the p-value (pval).
- Interpretation:Print the sample mean. Print the p-value.
  - If the p-value is less than the significance level (0.05), reject the null hypothesis.
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis.

Independent sample test:

- Use ttest_ind to perform an independent sample t-test between Electric_Range and Model_Year.
- Retrieve the t-statistic (ttest) and the p-value (pval).
- Interpretation - Independent Sample T-test:
  - Print the mean values and standard deviations.
  - Print the p-value.

  - If the p-value is less than the significance level (0.05), reject the null hypothesis, indicating that there is no significant association between Electric_Range and Model_Year.
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis, suggesting a significant association.

Paired sample Test:

- Use ttest_rel to perform a paired sample t-test between Electric_Range and Base_MSRP.
- Retrieve the p-value (pval).
- Interpretation - Paired Sample T-test:
  - Print the p-value.
  - If the p-value is less than the significance level (0.05), reject the null hypothesis, indicating a significant mean difference between Electric_Range and Base_MSRP.
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis, suggesting no significant mean difference.

ANOVA:

- Use scipy.stats.f_oneway to perform the ANOVA test on Electric_Range, Model_Year, and Base_MSRP.
- Retrieve the F-statistic (tstat) and the p-value (p).
- Interpretation - ANOVA Test:
  - Print the F-statistic and p-value.
  - If the p-value is less than the significance level (0.05), reject the null hypothesis, indicating that there is a significant difference in the means of at least one of the variables (Electric_Range, Model_Year, Base_MSRP).
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis, suggesting that the means of Electric_Range, Model_Year, and Base_MSRP are not significantly different.

NON-Parametric Test:

Mann-Whitney U Test:

- Use scipy.stats.mannwhitneyu to perform the Mann-Whitney U test on Electric_Range and Base_MSRP.
- Retrieve the U-statistic (tstat) and the p-value (p).
- Interpretation - Mann-Whitney U Test:
  - Print the U-statistic and p-value.
  - If the p-value is less than the significance level (0.05), reject the null hypothesis, indicating that the distributions of Electric_Range and Base_MSRP are significantly different.
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis, suggesting that the distributions are not significantly different.

Friedman Test:

- Use scipy.stats.friedmanchisquare to perform the Friedman test on Electric_Range, Base_MSRP, and Model_Year.
- Retrieve the test statistic (stat) and the p-value (p).
- Interpretation - Friedman Test:
  - Print the test statistic and p-value.
  - If the p-value is less than the significance level (0.05), reject the null hypothesis, indicating that the distributions of Electric_Range, Base_MSRP, and Model_Year are significantly different.
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis, suggesting that the distributions are not significantly different.

Test of Stationarity - very Important for time series analysis:

Augmented Dickey-Fuller Test:

- Use statsmodels.tsa.stattools.adfuller to perform the Augmented Dickey-Fuller test on the Electric_Range data.

- Retrieve the test statistic (stat), p-value (p), number of lags used (lags), number of observations used for the ADF regression (obs), critical values (crit), and t-statistic (t).
- Interpretation - Augmented Dickey-Fuller Test:
  - Print the test statistic and p-value.
  - If the p-value is less than the significance level (0.05), reject the null hypothesis, indicating that the series is stationary.
  - If the p-value is greater than or equal to 0.05, accept the null hypothesis, suggesting that the series is not stationary.

Finding and Insights

Q1. What is the most common electric vehicle type in the dataset?

Answer: - The most common electric vehicle type is 'Battery Electric Vehicle (BEV)' with 95953 occurrences.

Q2. Which are the Top 10 Electric Vehicle Model?

Answer: - Top 10 Electric Vehicle Models:

1. MODEL 3: 24481 vehicles
2. MODEL Y: 20609 vehicles
3. LEAF: 12897 vehicles
4. MODEL S: 7314 vehicles
5. BOLT EV: 4976 vehicles
6. VOLT: 4865 vehicles
7. MODEL X: 4635 vehicles
8. PRIUS PRIME: 2421 vehicles

9. ID.4: 2318 vehicles
10. NIRO: 2286 vehicles

Model 3 of Tesla has almost 30% of EV vehicles followed by Tesla's Model Y. In total almost 50% of the market is captured by Tesla in Washington, USA.Nissan's Leaf is the 3rd vehicle that has captured almost 16% of the market.

Q3. What is the average electric range of battery electric vehicles (BEVs) in the dataset?

Answer: - The average electric range of Battery Electric Vehicles (BEVs) is nan miles.

Q4. How many plug-in hybrid electric vehicles (PHEVs) are eligible for clean alternative fuel vehicle (CAFV) incentives?

Answer: - Number of eligible PHEVs for CAFV incentives: 0

There are not any PHEVs vehicles who are eligible for CAFV incentives.

Q5. Provide the visualization of the distribution of Electric Vehicle Models?

86K vehicles are Battery Electric Vehicles whereas only 27K vehicles are Plug-in Hybrid Vehicle that means they use some form of fossil fuels for internal combustion.

Q6. Can you provide the details of the electric vehicle with the highest legislative district?

Answer: - Electric Vehicle with the Highest Legislative District:

```
VIN (1-10)
5YJYGDEEXM
County                                                    Clark
City                                                  Vancouver
State                                                        WA
Postal_Code                                             98665.0
```

```
Model_Year                                                  2021
Make                                                       TESLA
Model                                                    MODEL Y
Electric_Vehicle_Type                             Battery Electric
Vehicle (BEV)
Clean_Alternative_Fuel_Vehicle (CAFV) Eligibility    Eligibility
unknown as battery range has not b...
Electric_Range                                                 0
Base_MSRP                                                      0
Legislative_District                                        49.0
DOL Vehicle_ID
168722382
Vehicle_Location                                  POINT (-
122.64443 45.67871)
Electric_Utility                    BONNEVILLE POWER
ADMINISTRATION||PUD NO 1 OF C...
2020 Census_Tract
53011041008.0
```

Tesla Model Y has the highest Legislative District with 49.0.

Q7. Plot the correlation map for all the numerical columns of the dataset?

Answer: - Electric Range has the strongest positive relation with Base MSRP and has a negative relation with Model Year.

Q8. How many electric vehicles have a base MSRP above $50,000?

Answer: - Number of electric vehicles with a base MSRP above $50,000: 2229

Q9. Which electric utility company serves the most electric vehicles in the dataset?

Answer: - The electric utility company serving the most electric vehicles is PUGET SOUND ENERGY INC||CITY OF TACOMA - (WA) with 44995 vehicles.

Q10. Which year contains more manufacturing of electric vehicles?

Answer: - Most of the data is concentrated on the 2019- 2020 Model Years, no data prior to 2010 has been shown.

Q11.For which electric range electrical vehicle has cleaner fuel (CAFV)?

Answer: -

It is noted that for electric ranges greater than 50 are eligible vehicles with clean fuel. The uncertainty lies in those ranges below 50 as they are either not eligible or there is no clear information on them.

Q12. How many electric vehicles are eligible for CAFV incentives in Washington state?

Answer: - The number of electric vehicles eligible for CAFV incentives in Washington state is 0.

There is not any electric vehicle which is eligible for CAFV incentives in Washington, US.

Q13. Can you provide the graph for distributions of CAFV Eligibility?

Answer: - In Clean Alternative Fuel Vehicle Eligibility (CAFV) the maximum EV cars from the dataset are Clean Alternative Fuel Vehicle Eligible.

Q14. Can you provide details on electric vehicles with a legislative district of 1?

Answer: - There are 5345 electric vehicles have legislative district of 1.

Q15. How many electric vehicles have a 0 electric range?

Answer: - Number of electric vehicles with a 0 electric range: 49537
There are a total of 49537 vehicles having an electric range of 0.

Q16. Plot the distribution of Cars in Various Years in the dataset?

Answer: - Most EV vehicles were manufactured in 2022 followed by 2021, year 2023 is low because almost half of the year is still left.

Q17. Which are the top 15 Electric Vehicle Models makers in the dataset?

Answer: - Tesla captures almost 52% of the market whereas 9 companies together capture 50%.

Q18. Visualize the number of electric vehicles (EVs) in each county?

Answer: - King has the most no. of EV Cars followed by Snohomish then Pierce. The reason could be there might be more no. of charging stations in these areas.

Q19. Visualize the distribution of electric vehicle types for the top N companies with the most electric vehicles?

Answer: - We can see that Tesla is all about BEV vehicles, it does not use PHEV vehicles at all Nissan as well follows the same as Tesla Toyota uses only PHEV vehicles which is why Toyota is famous all over Chevrolet uses almost 60-40 of both types.

Q20. Which company has the most distribution of cars?

Answer: -Tesla has the most EV vehicles followed by Nissan, Chevrolet, Ford, BMW, etc.

Insights

Model 3 of Tesla has almost 30% of EV vehicles followed by Tesla's Model Y. In total almost 50% of the market is captured by Tesla in

Washington, USA. Nissan's Leaf is the 3rd vehicle that has captured almost 16% of the market. Tesla captures almost 52% of the market whereas 9 companies together capture 50%. King has the most no. of EV Cars followed by Snohomish then Pierce. The reason could be there might be more no. of charging stations in these areas. Most EV vehicles were manufactured in 2022 followed by 2021, year 2023 is low because almost half of the year is still left. There is not any electric vehicle which is eligible for CAFV incentives in Washington, US. It is noted that for electric ranges greater than 50 are eligible vehicles with clean fuel. The uncertainty lies in those ranges below 50 as they are either not eligible or there is no clear information on them.

Limitations

- **Missing Data:** The dataset has missing values in certain columns, which can affect the completeness of the analysis.
- **Sample Bias:** The dataset is not fully representative of the entire population of electric vehicles. It focusses on specific regions, manufacturers, or time periods, leading to a biased view.
- **Limited Variables:** The available variables do not cover all aspects relevant to the analysis. Some important features or factors influencing electric vehicles' performance or adoption are missing.
- **Temporal Limitations:** The dataset is limited to a specific time, and trends or patterns over time could be missed.

- **Scope of CAFV Eligibility:** The criteria for Clean Alternative Fuel Vehicle (CAFV) eligibility may not be explicitly defined in the dataset, leading to ambiguity in understanding the eligibility status.
- **Limited Context:** The dataset may lack contextual information about external factors influencing electric vehicle ownership or usage, such as government policies, charging infrastructure, or public perception.

Recommendations

As We have seen EVs used are limited to some cities only and many of them has high numbers of them while some has few. It may be because of less electric vehicle charge station. There should be installation of many charging stations and people need to know more about clean fuel and zero emission fuels.

Conclusion

After all the analysis of the dataset, we can conclude that Model 3 has been dominating till now and Tesla is the company which is making most of the electrical vehicles. There should be a greater number of charging stations, so that the number of electric vehicles can be used more.

References

https://drive.google.com/file/d/1MY0KSzcIQAAyrrkjvdl9vlltUMpBA36R/view?usp=sharing (kaggle)

https://en.wikipedia.org/wiki/Statistical_hypothesis_testing

https://www.analyticsvidhya.com/blog/2021/06/exploratory-data-analysis-using-data-visualization-techniques/

Links

https://docs.google.com/presentation/d/1VDCUr35lJjQ00Ug1LgRQJhyqphwaOuf6/edit?usp=sharing&ouid=114143997331721506711&rtpof=true&sd=true

https://github.com/sonal2909/Electric-Vehicle-Population-Analysis