**A COMPREHENSIVE ANALYSIS OF ELECTRIC VEHICLES TRENDS**

**Introduction:**

The automotive landscape is undergoing a revolutionary transformation with the increasing adoption of electric vehicles (EVs). With the commercialization of new Tesla cars and newfound startups like Zoox who are on a boom in the EV industry, more and more of the general population is starting to switch to EVs. This brings to question, will there be a point in time where every vehicle will be an EV? The present scenario prompted us to do an analysis on the current market of EV in the US. With the data found from the vehicles registered under the Department of Liscencing of Washington State, we want to present an analysis to get answers to some questions such as:

1. What is the distribution of electric vehicles across different counties and cities and are there any notable patterns or concentrations?

2. What is the relationship between the Electric Vehicle type, CAFV eligibility and the electric utility?

3. What insights can we gain about the future of electric mobility by examining the data on newer models and their technological features? For instance, how are factors like electric range and Electric vehicle type evolving with advancements in electric vehicle technology?

Team Members

Sajan Kumar Kar (G40140248) – Topic idea, Data Cleaning, Data Analysis(Q2)

Pratiksha Dange (G24444604) – Coming up with data questions, Data Analysis(Q1, Q3)

**Description of Data:**

The dataset we used here is found from the vehicles registered under the Department of Licensing of Washington State. We have accessed the data by visiting the link: <https://catalog.data.gov/dataset/electric-vehicle-population-data>

The dataset is a comprehensive record of the electric vehicle (EV) population, including data points such as Vehicle Identification Number (VIN), county, city, state, postal code, model year, make, model, electric vehicle type, Clean Alternative Fuel Vehicle (CAFV) eligibility, electric range, base Manufacturer's Suggested Retail Price (MSRP), legislative district, Department of Licensing (DOL) vehicle ID, vehicle location, electric utility, and 2020 Census Tract.

The dataset consists of 177,866 entries and 17 columns of data. Each record presumably represents an individual EV, and the VINs have been anonymized for privacy. The information spans a range of model years and includes various manufacturers and models, showcasing the diversity in the EV market. This dataset provides a snapshot of the geographic distribution of EVs, their economic attributes (such as base MSRP), and technical specifications like electric range, which can be pivotal for analyzing the EV adoption rate, market penetration, and infrastructure requirements.

**Exploratory Data Analysis:**

**1. Analysis of Data Quality**

For the data quality analysis, we will delve into each variable. The primary goal is to understand the dataset's completeness, accuracy, consistency, and uniformity.

|  |  |  |
| --- | --- | --- |
| **Feature** | **Data Type** | **Definition** |
| **VIN (1-10)** | Text | The 1st 10 characters of each vehicle's Vehicle Identification Number (VIN) serving as a unique identifier for each vehicle. |
| **County** | Text | This is the geographic region of a state that a vehicle's owner is listed to reside within. Vehicles registered in Washington state may be located in other state. |
| **City** | Text | The city in which the registered owner resides. |
| **State** | Text | This is the geographic region of the country associated with the record. These addresses may be located in other states. |
| **Postal Code** | Float | The 5 digit zip code in which the registered owner resides. |
| **Model Year** | Integer | The model year of the vehicle, determined by decoding the Vehicle Identification Number (VIN). |
| **Make** | Text | The manufacturer of the vehicle, determined by decoding the Vehicle Identification Number (VIN). |
| **Model** | Text | The specific model of the vehicle, determined by decoding the Vehicle Identification Number (VIN). |
| **Electric Vehicle Type** | Text | This distinguishes the vehicle as Battery Electric Vehicle (BEV) or a plug-in hybrid. |
| **CAFV Eligibility** | Text | This categorizes vehicle as Clean Alternative Fuel Vehicles (CAFVs) based on the fuel requirement and electric-only range requirement in House Bill 2042 as passed in the 2019 legislative session |
| **Electric Range** | Integer | Describes how far a vehicle can travel purely on its electric charge. |
| **Base MSRP** | Integer | This is the lowest Manufacturer's Suggested Retail Price (MSRP) for any trim level of the model in question. |
| **Legislative District** | Float | The specific section of Washington State that the vehicle's owner resides in, as represented in the state legislature. |
| **DOL Vehicle ID** | Integer | Unique number assigned to each vehicle by Department of Licensing for identification purposes. |
| **Vehicle Location** | Point | The center of the ZIP Code for the registered vehicle. |
| **Electric Utility** | Text | This is the electric power retail service territories serving the address of the registered vehicle. All ownership types for areas in Washington are included: federal, investor owned, municipal, political subdivision, and cooperative. If the address for the registered vehicle falls into an area with overlapping electric power retail service territories then a single pipe | delimits utilities of same TYPE and a double pipe || delimits utilities of different types. Vehicle address and Homeland Infrastructure Foundation Level Database (HIFLD) Retail\_Service\_Territories feature layer has been combined using a geographic information system to assign values for this field. Blanks occur for vehicles with addresses outside of Washington or for addresses falling into areas in Washington not containing a mapped electric power retail service territory in the source data. |
| **2020 Census Tract** | Float | The census tract identifier is a combination of the state, county, and census tract codes as assigned by the United States Census Bureau in the 2020 census, also known as Geographic Identifier (GEOID) |

The dataset is substantial in size, which provides a rich foundation for analysis, but it does exhibit some missing values and potential inaccuracies (such as zero values for range and MSRP) that will need to be addressed during the data cleaning process.

# Data Import

df = pd.read\_csv('/Users/pratikshadange/Desktop/My Docs/MSDS Courses/DataViz/Prof.Sarah/Proj\_Final/Electric\_Vehicle\_Population\_Data (1).csv')

df.head()

Output:

A screenshot of a computer

Description automatically generated

#Dimension of the dataframe

df.shape

Output:

(177866, 17)

# descriptive statistics of the data

df.describe()

Output:

A black screen with numbers

Description automatically generated

For Electric Range, median is 0 and the 3rd quantile is 84, mean is 61.5. This indicates a heavy right skew and a possibility of outliers.

df.describe(include='O')

Output:

A screenshot of a computer

Description automatically generated

# Summary of dataframe

df.info()

Output:

A screenshot of a computer program

Description automatically generated

2. **Data Cleaning -** The initial step involved thorough data cleaning to ensure the integrity and quality of the dataset. Here's a detailed outline of the data cleaning process:

* + 1. Identifying the missing values- The dataset presented with missing values across several crucial attributes, notably 'County', 'City', 'Postal Code', 'Legislative District', 'Vehicle Location', 'Electric Utility', and '2020 Census Tract'. Notably, 5 records exhibited missing values across 'County', 'City', and 'Postal Code', suggesting a potential correlation among these omissions.

# Detecting Missing Values

df.isna().sum()

A screenshot of a computer

Description automatically generatedOutput:

df[df['County'].isna()]

A black and white table with white text

Description automatically generatedOutput:

* + 1. Handling Missing Records- To maintain data consistency, the decision was made to eliminate the 5 records with missing values across 'County', 'City', 'Postal Code', 'Vehicle Location', 'Electric Utility', and '2020 Census Tract'. Additionally, another 5 records lacking 'Vehicle Location' values were also removed.

df\_cleaned = df.dropna(subset=['County'])

df\_cleaned = df.dropna(subset=['Vehicle Location'])

df\_cleaned.isna().sum()

A screenshot of a computer

Description automatically generatedOutput:

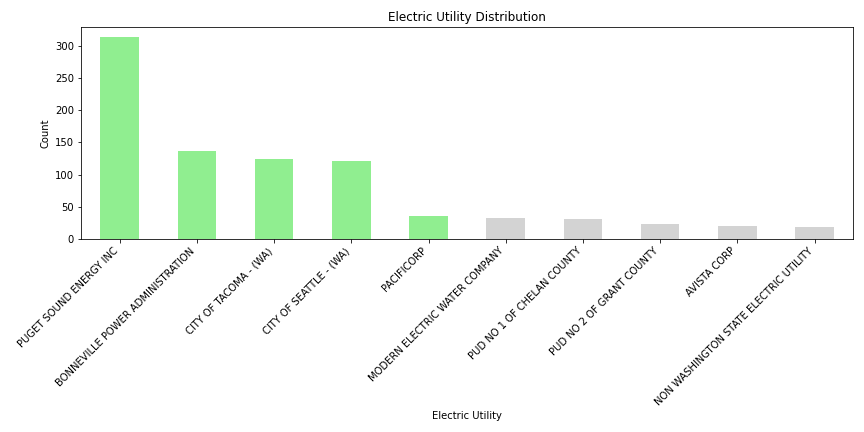
* + 1. Addressing 'Electric Utility' Values- The 'Electric Utility' attribute displayed a varied range of values, including single values, pairs separated by '|' or '||', and even multiple values separated by both '|' and '||'. To standardize this attribute, a systematic approach was adopted:
       - 1. Single Value: No alteration required.
         2. Two Values Joined by '|': One value was randomly chosen due to their similar nature.
         3. Two Values Joined by '||' : The first value was retained as the primary territory.
         4. Three Values Joined by '||' and '|or||': The middle value was selected to ensure comprehensive territory coverage.
         5. Multiple Values Joined by '||' and '|': A random selection between the second and third values was made to encompass all territories in between.
    2. Reducing 'Electric Utility' Categories- Subsequently, only the top subset of 'Electric Utility' categories, with higher counts, was retained to streamline the analysis, reducing the total from 34 unique categories.
    3. Feature Name and Values Mapping- 'Clean Alternative Fuel Vehicle (CAFV) Eligibility' was renamed to 'CAFV Eligibility' for clarity. Mapping was applied to standardize 'CAFV Eligibility' values into three categories. 'Electric Vehicle Type' values underwent mapping to enhance clarity and consistency.

df\_cleaned['Electric Utility'].unique()

df\_cleaned['Make'].nunique()

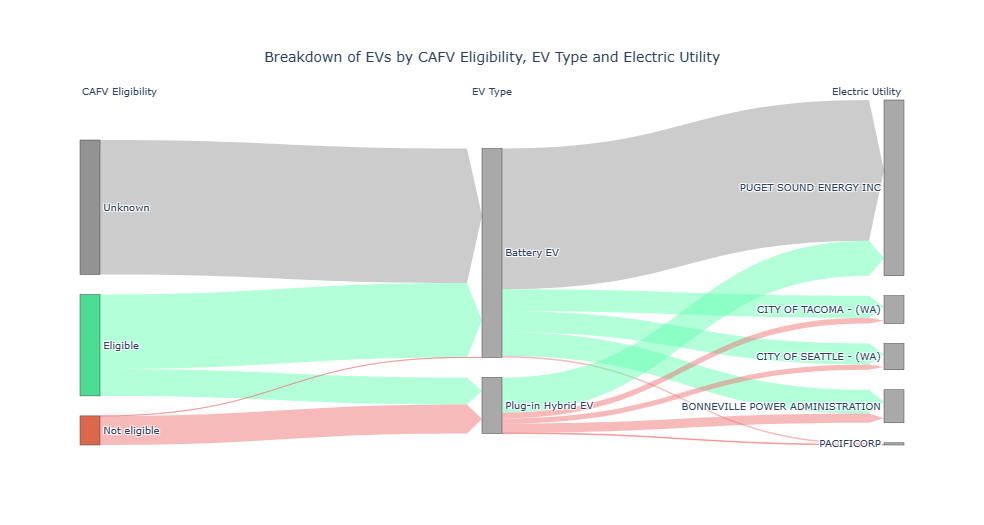
Output:

A screenshot of a computer

Description automatically generated**Main Analysis :**

1. To explore the relationship between Electric Vehicle (EV) types, Clean Alternative Fuel Vehicle (CAFV) eligibility, and Electric Utility, a Sankey plot was constructed. First, the top 10 Electric Utility categories were visualized, with the top 5 selected for further analysis. The top 5 Electric Utilities are as highlighted in the below image:

Subsequently, the Sankey plot was generated using the top 5 Electric Utilities. The plot illustrates the distribution of EVs based on CAFV eligibility, EV type, and Electric Utility.

Key observations include:

* The majority Battery EV vehicles have an unknown CAFV eligibility and are associated with PUGET SOUND ENERGY INC Electric Utility.
* Nearly all CAFV eligible Battery EVs are evenly distributed among CITY OF TACOMA, CITY OF SEATTLE, and BONNEVILLE POWER ADMINISTRATION Electric Utilities.
* CAFV eligible Plug-in Hybrid EVs predominantly belong to PUGET SOUND ENERGY INC.
* Non-eligible CAFV EVs are primarily Plug-in Hybrid EVs, with exceptions being Battery EVs, all associated with PACIFICORP Electric Utility.

Creating the Sankey plot posed challenges in customizing colors to represent CAFV eligibility, EV types, and Electric Utility. Balancing clarity and coherence in color interpretation and setting category names for interpretability required meticulous coding and manual iterative refinement to ensure a visually compelling and informative visualization.

2. What is the distribution of electric vehicles across different counties and cities and are there any notable patterns or concentrations?

This is our second business question and to resolve this question we have created 2 bar charts and one tree map as below. The aim is to identify patterns of EV adoption and the geographical concentration of EVs, which could be indicative of infrastructural and policy impacts on EV distribution. Prior to analysis, the dataset underwent a cleaning process to ensure accuracy in the findings (‘df\_cleaned’ suggests a cleaned DataFrame).

1. **Histogram: Distribution of Electric Vehicle by City**

The data was visualized using Tableau, he underlying dataset included fields for the city name and the count of EVs in each city. This data was likely aggregated during the visualization process in Tableau.

* The **x-axis** was assigned to the city names. These would be categorical variables.
* The **y-axis** represents the count of EVs, which is a quantitative variable.
* Color: Count of EV measure is used to color the histogram and green color pallet is used.
* Sorting: The data is sorted by the count of EVs in descending order, bringing the city with the highest count to immediate attention.
* Filters: We exclude cities with very low counts of EVs for clarity, and applied condition to visualize the cities having greater than or equal to 1000 EV’s.

A graph of distribution of ev by city

Description automatically generated

**Key Finding:**

• This visualization indicates a significant variance in EV distribution across cities.

• Seattle city demonstrates an exceptionally high concentration of EVs, which may suggest a strong local EV ecosystem or incentives. Bellevue, Redmond, Vancouver, Sammamish, Kirkland, Renton, Bothell, Olympia are ranging the EV count from 8000 to 2000.

1. **Histogram: Distribution of EV by County**

* Data Preparation: The data needed to be grouped by county with a sum of EVs in each to create this visualization.
* Variable Assignment:
  + The **x-axis** was used for county names, another categorical variable.
  + The **y-axis** represents the total count of EVs in each county, a quantitative variable.
* Color: Just like the first chart, the color here is consistent across all bars, Count of EV measure is used to color the histogram and green color pallet is used.
* Sorting: The bars are sorted, likely in descending order, to emphasize the counties with the most significant number of EVs.
* Filters: Filters have been applied to include only certain counties based on criteria like the counties having greater than or equal to 500 EV’s.

A graph of a number of people

Description automatically generated

**Key Finding:**

The histogram reveals a similar pattern at the county level, with King County displaying a notably higher number of EVs, and highlighted in dark green followed by Snohomish, Pierce, and Clark.

**3. Treemap: Electric Vehicle by County and City**

The treemap displays a nested distribution of electric vehicles, organized by county and further by city within each county. The tree map is a powerful visualization tool for revealing the distribution of a variable across a nested categorical hierarchy. In this case, it elucidates the spread of EVs in a region, highlighting areas of high and low adoption, which can be pivotal information for policymakers, urban planners, and businesses involved in the EV market.

* The treemap, generated using Plotly Express, provides a hierarchical representation of EV distribution, with size indicating the total number of EVs in counties and color saturation representing the count within cities.
* This visualization was designed to intuitively convey relative densities and distributions, using annotations to clarify the represented dimensions.

#Treemap of Electric Vehicle by county and city

county\_city\_counts = df\_cleaned.groupby(['County', 'City']).size().reset\_index(name='Count')

county\_counts = county\_city\_counts.groupby('County')['Count'].sum().reset\_index()

combined\_counts = pd.merge(county\_city\_counts, county\_counts, on='County', suffixes=('\_city', '\_county'))

fig = px.treemap(combined\_counts, path=['County', 'City'], values='Count\_city',

color='Count\_city', color\_continuous\_scale='Viridis',

title='Treemap of Electric Vehicles by County and City',

labels={'Count\_city': 'Count of EV <br>(City)', 'Count\_county': 'Count of EV (County)'})

fig.update\_traces(root\_color='lightgrey', hovertemplate='<b>%{label}</b><br>Count: %{value}<extra></extra>')

fig.add\_annotation(text="Color represents the count of electric vehicles within each city",

xref="paper", yref="paper",

x=0.5, y=1.15, showarrow=False,

font=dict(size=12, color="black"))

fig.add\_annotation(text="Size represents the count of electric vehicles within each county",

xref="paper", yref="paper",

x=0.5, y=1.1, showarrow=False,

font=dict(size=12, color="black"))

fig.update\_layout(

title={

'text': 'Treemap of EV by County and City',

'y': 0.97,

'x': 0.5,

'xanchor': 'center',

'yanchor': 'top',

'font': {'size': 24, 'color': 'Black'},

},

width=1000,

height=500

)

fig.show()

**Technical Approach and Challenges**

* The code used for the treemap creation involves grouping data by county and city and summing up the counts. This hierarchical data was then merged to present a combined view of city-level data within the context of their respective counties.
* One technical challenge was ensuring the accuracy of merging operations, where data alignment and integrity are crucial.
* Another consideration was the selection of an appropriate color scale (**Viridis**) to ensure the visualization is understandable and accessible.

**Approaches That Didn't Work**

While not explicitly stated, the analysis likely underwent several iterations, including:

* **Trial and Error with Visualizations**: Finding the most informative and interpretable type of visualization can require experimenting with multiple forms, such as bar charts, scatter plots, or heatmaps, before settling on histograms and treemaps.
* **Data Cleaning Iterations**: Initial passes at cleaning the data might not have captured all inconsistencies or missing values, requiring multiple iterations.

A purple and blue map

Description automatically generated with medium confidence

**Key Findings:**

1. **County-Level Distribution:**

* The size of the county boxes indicates that there's significant variance in the number of EVs by county. A few counties such as King, Snohomish, Pierce, Clark, Spokane, Thurston have much larger boxes, showing a higher total count of EVs.

1. **City-Level Concentration:**
   * Within these counties, certain cities like Seattle, Bellevue, Redmond, Vancouver, Sammamish, Kirkland, Renton, Bothell, Olympia stand out due to their box size, indicating a concentration of EVs. Notably, the largest boxes are not always in the largest counties, suggesting that certain cities have high adoption rates, potentially due to localized incentives, charging infrastructure, or socioeconomic factors.
2. **Comparative Analysis:**
   * Comparing the colors and sizes of the city boxes within a single county indicate the relative distribution of EVs. For instance, Seattle city's box is much darker and larger than Auburn city within same county, it suggests a particularly high concentration of EVs in that city.
3. **Data Density and Urban Centers:**
   * Larger and darker-colored boxes often correspond to urban centers or cities with better EV infrastructure. These cities likely have more resources dedicated to supporting EVs, such as public charging stations.
4. **Rural vs. Urban Spread:**
   * Smaller and lighter-colored boxes might represent rural areas or smaller towns where EV adoption is lower. This could be due to a variety of factors, including lower availability of charging infrastructure, higher reliance on long-range transportation where EVs are less practical, or socioeconomic barriers.
5. **Policy Implications:**
   * Areas with low EV concentration could be targeted for policy interventions to improve adoption rates, suggesting that state or local governments could invest in infrastructure improvements or provide incentives in these areas.

Top of Form

Bottom of Form

**Conclusion:**

Throughout the analysis, it became evident that the EV market is experiencing a substantial surge in popularity. The dataset revealed a plethora of new features and advancements being introduced within the EV landscape, indicative of the industry's rapid evolution and innovation.

One notable limitation of the dataset was the prevalence of 'Base MSRP' values set at 0 for the majority of entries. This anomaly suggests potential data inaccuracies, as it would imply that vehicles were available for purchase at no cost, which is unrealistic. This limitation hindered the ability to conduct comprehensive analyses relating to pricing trends, model features, or price fluctuations over time. Moreover, the scope of the analysis was restricted to the state of Washington, limiting the generalizability of the findings.

Navigating the dataset revealed the importance of data accuracy and pre-processing for meaningful analysis. Despite limitations such as faulty 'Base MSRP' values, insights into the booming EV market and its evolving features were gained. Future endeavours’ should prioritize access to comprehensive, reliable datasets beyond state borders, enabling nationwide analyses. Enhanced data quality will facilitate in-depth examinations of pricing dynamics, model features, and market trends. By addressing these challenges and leveraging improved data sources, future research can unlock deeper insights into the dynamic landscape of electric vehicles, paving the way for informed decision-making and strategic advancements within the industry.