A computer screen shot of a computer

Description automatically generated

**CST3344 Business Intelligence Course Work 2**

A car charging at a charging station

Description automatically generated

Analysis of Electric Vehicle Population Data

By: Rian Qadir M00975827

Contents

[ Introduction 2](#_Toc183811567)

[ Data Cleaning: 2](#_Toc183811568)

[ Data Analysis & Visualization 3](#_Toc183811569)

[ Selection of Data Mining Algorithm and Data Pre-processing: 13](#_Toc183811570)

[ Data Mining 15](#_Toc183811571)

[1. Clustering-Simple K means: 15](#_Toc183811572)

[ Data Ethics 18](#_Toc183811573)

[ Conclusion: 20](#_Toc183811574)

## Introduction

The dataset was sourced from Kaggle, which is a well-known platform for data analysis and competitions that hosts a variety of public datasets. This data set primarily represents the population of electric vehicles registered with the Washington State Department of Licensing (DOL) and other states. It includes both Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), offering valuable insights into the adoption and distribution of electric vehicles across the state. The dataset includes vehicle details such as make, model, model year, electric vehicle type, electric range, base MSRP, and geographic and registration details like city, county, and legislative district. The dataset provides a picture of Washington State's transition to sustainable transportation.

The dataset includes several types of variables, such as vehicle-specific information (e.g., make, model, electric range), geographic details (e.g., city, county), and registration data (e.g., VIN, vehicle ID). Together, these variables offer a comprehensive view of the electric vehicle ecosystem in the states.

### Data Cleaning:

While the dataset was relatively well structured and clean, a few preprocessing steps were taken to ensure the dataset was ready for analysis in Tableau. These preprocessing steps include:

1. **Duplicate records**, particularly in the VIN field, were identified and removed to prevent redundancy in vehicle counts.
2. **Missing values** in fields like Base MSRP and Electric Range were addressed.
3. **Text fields**, such as Make and Model, were standardized to correct inconsistencies in spelling or formatting.

When the data set had been entirely cleaned, potential cleaning steps were still applied to ensure that the dataset was in its best possible form for analysis. For example, Data transformation by creating calculated fields such as Total Vehicles, Total BEV Vehicles, Total PHEV Vehicles, AVG Electric Range % BEV Vehicles, and % PHEV Vehicles. Additionally, the data was also cross validated against external sources from the internet, to ensure its accuracy and reliability.

## Data Analysis & Visualization

The dataset on electric vehicles was analyzed using Tableau to uncover trends, patterns, and actionable insights. Below are the key visualizations and their interpretations.

1. **Total Vehicles by State:**

**A map of the united states

Description automatically generated**

This map visualization shows the distribution of electric vehicles across the United States, with a specific emphasis on Washington State. Each state is represented by a mark whose size and color intensity indicate the total number of vehicles registered. Most vehicles, totaling 150,141, are registered in Washington State, far surpassing the numbers in other states. California and a few neighboring states show moderate EV registrations, while other regions, such as Wyoming and Mississippi, have notably lower figures.

**Key Insight:** The concentration of vehicles in Washington highlights the state's leadership in EV adoption, supported by progressive environmental policies and urban infrastructure. States with minimal registrations indicate untapped markets or challenges such as limited charging stations or lower awareness.

1. **Top 10 Vehicles by Make:**

**A screenshot of a computer

Description automatically generated**

This horizontal bar chart shows the top 10 manufacturers of electric vehicles based on total registrations. Tesla dominates the chart, accounting for 52.71% of all registered vehicles, followed by Nissan (10.31%) and Chevrolet (9.19%). Ford, BMW, and Kia make up smaller, yet notable, portions of the market. The chart labels include both the total number of vehicles for each manufacturer and their respective market share percentages.

**Key Insight**: Tesla’s overwhelming market share underscores its position as the leader in electric vehicle innovation and adoption. Traditional automakers like Nissan and Chevrolet continue to hold competitive spots, especially in the hybrid and entry-level EV markets.

1. **Popular Vehicle Models:**

**A screenshot of a graph

Description automatically generated**

This bar chart ranks the top 10 most popular electric vehicle models by the total number of registrations. Tesla models dominate the list, with the **Model Y** and **Model 3** accounting for **28.15%** and **27.36%** of total registrations, respectively. Other notable models include the Nissan Leaf (**13.02%**), Tesla Model S (**7.52%**), and Chevrolet Bolt EV (**5.66%**). Models like the Volkswagen ID.4 and Chrysler Pacifica have smaller shares but still feature in the top 10.

**Key Insight**: Tesla Model Y and Model 3's dominance highlights the company’s strong appeal across different segments, offering both affordability and premium features. The presence of models like the Nissan Leaf reflects the importance of budget-friendly EV options in the market.

1. **Total Vehicles by Model Year:**

**A screenshot of a graph

Description automatically generated**

This chart depicts the trend in electric vehicle registrations over time, categorized by model year. Registrations have steadily increased since 2011, with a dramatic surge between 2021 and 2023. The peak year was 2023, with 37,079 vehicles registered, highlighting the recent acceleration in EV adoption. Registrations for 2023 and 2024 are notably lower, likely due to incomplete data at the time of reporting.

**Key Insight:** The sharp growth in registrations reflects advancements in EV technology, greater affordability, and increased government incentives during recent years.

1. **Total Vehicles by Clean Alternative Fuel Vehicle (CAFV) Eligibility:**

**A chart with a colorful circle

Description automatically generated with medium confidence**

This pie chart breaks down vehicles by their eligibility for Clean Alternative Fuel Vehicle (CAFV) incentives. The data reveals that **41.83%** of vehicles are eligible, while **46.32%** have an unknown eligibility status. The remaining **11.85%** of vehicles are not eligible, primarily due to limited electric range in certain PHEVs.

**Key Insight**: The large percentage of unknown eligibility suggests the need for better data collection and transparency. Expanding eligibility criteria, especially for PHEVs, could boost adoption rates among a wider audience.

1. **Total Vehicles by Model:**

**A screenshot of a computer

Description automatically generated**

**Total Vehicles by Model** visualization provides a comprehensive breakdown of electric vehicle registrations by model, make, and type (BEV or PHEV), along with their total count and percentage of total registrations. Tesla dominates the landscape, with the **Model Y** and **Model 3** accounting for **28.15%** and **27.36%** of total registrations, respectively, together comprising over half of all-electric vehicles in the dataset. Other popular models include the **Nissan LEAF** (13.02%), **Chevrolet Bolt EV** (5.66%), and **Chevrolet Volt** (4.83%), indicating that affordability and established brands play an important role in market share.

While BEVs represent the majority of the top-ranking models, PHEVs like the Chevrolet Volt and Chrysler Pacifica contribute to the dataset, though at smaller scales.

**Key Insight:** The visualization highlights Tesla’s market leadership, driven by its ability to offer desirable features and performance across price points, while affordable BEVs like the Nissan LEAF demonstrate the significance of cost-effective options in promoting EV adoption. PHEVs, while present, show limited market penetration compared to fully electric vehicles.

1. **Average Electric Range by Make:**

**A screenshot of a computer

Description automatically generated**

This bubble chart displays the average electric range for vehicles from different manufacturers. Jaguar leads with an average range of 204.2 miles, followed by Tesla, Nissan, and Chevrolet. Other manufacturers, such as Smart and Fiat, have shorter ranges, likely to target urban commuters with affordable options. The size of each bubble represents the average range, and the colors differentiate between the manufacturers.

**Key Insight:** Jaguar's superior electric range reaffirms that even with the technological leadership that Tesla comprises there is still a lot to do in order to improve the vehicle range so that it can appeal to consumers seeking long-distance travel options. However, shorter-range vehicles remain relevant for specific urban markets.

1. **Electric Vehicle Registrations Over Time:**

**A graph showing a green line

Description automatically generated**

This line chart tracks the total number of vehicles registered over time, segmented by model year which can also be filtered by the Electric Vehicle Type. Similar to the earlier analysis, it highlights consistent growth, with significant spikes in registrations from 2021 to 2023. The data suggests that 2023 saw the highest registrations, coinciding with increased public interest and support for EVs.

**Key Insight:** The continued growth over time indicates that EVs are becoming mainstream, driven by improvements in affordability, performance, and public infrastructure.

1. **Average Base MSRP by Make**

**A graph with numbers and a bar chart

Description automatically generated with medium confidence**

This bar chart shows the average base Manufacturer's Suggested Retail Price (MSRP) for different car manufacturers. Fisker stands out with an extraordinarily high average MSRP of $1,734,000, significantly surpassing all other brands. Wheego follows as the second-highest with an average MSRP of $98,985.

1. **Total Vehicles by Model Year**

**A graph showing the number of vehicles

Description automatically generated**

The area chart illustrates the total number of vehicles manufactured over time, based on model year. The trend shows a significant increase in production volume in recent years, peaking at 37,079 vehicles in 2022. Earlier years (1996–2010) have significantly lower production numbers, indicating a focus on more modern models in the dataset.

**Key Performance Indicators (KPIs):**

1. **Total Vehicles:**

**A screenshot of a computer

Description automatically generated**

The dataset includes a total of 150,482 vehicles, reflecting the overall scale of EV adoption.

1. **Average Electric Range:**

**A screenshot of a computer

Description automatically generated**

The average range for all vehicles is 67.88 miles, with variations based on vehicle type and manufacturer.

1. **Total BEV Vehicles and % BEV:**

**A screenshot of a computer

Description automatically generated**

BEVs account for 116,807 vehicles, representing 77.6% of the total.

1. **Total PHEV Vehicles and % PHEV:**

**A screenshot of a computer

Description automatically generated**

PHEVs account for 33,675 vehicles, representing 22.4% of the total.

## Selection of Data Mining Algorithm and Data Pre-processing:

The **k-means Clustering** data mining algorithm was chosen for further analysis and implemented using the Weka SimpleKMeans tool. The algorithm was selected to segment the data into distinct groups of electric vehicles based on shared characteristics.

K-means is an unsupervised learning algorithm that partitions data into k clusters; each cluster contains instances with similar attributes. The algorithm works by randomly selecting k initial centroids, which are the starting points for clusters. It then assigns each data point to the nearest centroid based on a distance metric, which is Euclidean distance. The algorithm then recalculates the centroids by averaging the positions of all points in a cluster. It then repeats the assignment and recalculates the centroids until centroids stabilize or a maximum number of iterations is reached. The final output is a set of clusters where intra-cluster variance is minimized, and inter-cluster variance is maximized.

The visualizations in the previous section highlighted distinct trends and patterns such as the dominance of Battery Electric Vehicles (BEVs) over Plug-in Hybrid Electric Vehicles (PHEVs), Variation in electric range, and geographic disparities in adoption.

K-means clustering is ideal for exploring the differences as it covers the segmentation of vehicles based on their attributes which then reveal insights into market trends and customer preferences by grouping vehicles by their type, range, and MSRP to identify the clusters that represent specific market segments.

**Variable Used:**

* Electric Vehicle Type (BEV or PHEV)
* Electric Range (maximum distance covered by vehicle on electric power)
* Base MSRP (Price set by the manufacturer)
* Model Year (Year of manufacture)

**Data Pre-Processing Steps:**

|  |
| --- |
| 1. **Handling Missing Values:** Missing values in numeric variables were replaced with the mean, and categorical variables were replaced with the mode to ensure consistency. |
| 1. **Outlier Detection:** Outliers in numeric variables, such as exceptionally low or high electric ranges, were analyzed but no extreme anomalies were removed, instead their influence was mitigated by standardizing the data**.** |
| 1. **Standardization:** Numeric variables were normalized to ensure equal weighting during distance calculations in K-means. |
| 1. Categorical variables like Electric Vehicle Type were converted into numeric representations to make them compatible with the algorithm |

By applying the pre-processing steps, the dataset was prepared for clustering by ensuring the accuracy of the resulting insights. K-means enables segmentation that aligns with the visual pattern observed, providing an in-depth understanding of market dynamics.

## Data Mining

The data set containing almost 150,000 rows was then analyzed in Weka using the Clustering- Simple K mean data mining algorithm through several iterations. This algorithm was applied to uncover patterns, relationships, and classifications in electric vehicle population data.

### Clustering-Simple K means:

A screenshot of a computer

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

**Implementation:**

The dataset, which consisted of 150,476 instances and 17 attributes, was preprocessed to handle missing values, which were replaced with the mean for numeric attributes and the mode for categorical attributes. K-means was run with the following configuration:

* **Number of Clusters (k):** 2
* **Distance Metric:** Euclidean Distance
* **Initialization Method:** Random starting points.

**Clustering Details:**

* **Cluster 0**: Represents **78%** of the dataset, comprising 116,964 instances.
* **Cluster 1**: Represents **22%** of the dataset, comprising 33,512 instances.
* The clustering process was completed in six iterations, with a **within-cluster sum of squared errors (SSE)** of **875,752.91**, indicating the variability within each cluster.

The clustering process ran 6 times, with a within-cluster sum of squared errors of 875,52.91, which indicates the variability within each cluster.

|  |  |
| --- | --- |
| Cluster 0 | Cluster 1 |
| Dominated by Battery Electric Vehicles (BEVs). | Composed mainly of Plug-in Hybrid Electric Vehicles (PHEVs). |
| Higher average electric range (78.54 miles) and lower average base MSRP ($1,146.06). | Significantly lower average electric range (30.65 miles) and higher base MSRP ($1,894.29). |
| The most common make and model is the Tesla Model Y. | The most common make and model is the Toyota Volt. |
| Vehicles in this cluster tend to be newer, with an average model year of 2020. | Older vehicles on average, with an average model year of 2018. |

**Geographical Analysis:**

Both clusters are concentrated in Washington State, while the same geographic patterns were observed in their vehicle locations and electric utility providers. However, Cluster 0 vehicles are slightly more prevalent in King County which is an urban region that has a slight upper hand in urban regions.

**CAFV Eligibility:**

Vehicles in Cluster 0 often have unknown eligibility for Clean Alternative Fuel Vehicle (CAFV) incentives. While, Cluster 1 vehicles are frequently not eligible due to their lower battery ranges, which are very common in PHEVs.

**Evaluation:**

The clustering results highlight a clear distinction between BEVs and PHEVs, reflecting their technical and economic differences. The clustering model fits the data well, as indicated by the low within-cluster SSE.

1. The low SSE value (875,752.91) refers to relatively tight clusters which means that vehicles within each group have very similar characteristics.
2. The split between clusters (78% BEVs and 22% PHEVs) aligns with broader trends in the dataset, confirming the algorithm's effectiveness in identifying natural groupings.

**Insights:**

1. Cluster 0: BEV vehicles dominate the market, driven by their higher range and affordability. Policymakers and manufacturers should focus on incentivizing BEV purchases and expanding charging infrastructure in urban regions.
2. Cluster 1: PHEVs, while less prevalent, cater to a niche market. Improving battery ranges could make these vehicles more attractive, potentially increasing their market share.

## Data Ethics

**Ethical Considerations:** Ethical data analysis prioritizes fairness and the responsible use of information. It must be ensured that the data is not misrepresented or manipulated to support biased conclusions.

The privacy of individuals associated with the data is another critical aspect. Ethical principles require ensuring data anonymization and avoiding unnecessary aggregation that could inadvertently identify individuals.

Additionally, the potential misuse of findings must be considered, such as avoiding reports that could discriminate against certain regions or demographics in infrastructure planning or policy decisions.

**Legal Considerations:** Legal compliance is paramount in data analysis. Adherence to regulations such as the General Data Protection Regulation (GDPR) ensures data is handled appropriately.

It is also essential to verify that the data was obtained legally and that any insights generated do not violate copyright or intellectual property laws.

Furthermore, if the data includes any inferred attributes, such as geographic patterns or preferences, these insights must be handled within the boundaries of applicable legal frameworks to avoid breaches of confidentiality or discriminatory practices.

**Professional Consideration:** Professional conduct in data analysis requires transparency, accuracy, and accountability. Analysts must document every step of the analysis, including data cleaning, transformations, predictions, and interpretations, to ensure the results can be reproduced and verified.

Visualizations and reports must be presented honestly, avoiding any misleading claims or representations.

Professionals must also acknowledge the limitations of the analysis and highlight areas where data gaps could influence results.

Communication with stakeholders should focus on the responsible use of findings and insights, to ensure that data-driven decisions align with ethical and legal norms.

## Conclusion:

The electric vehicle population dataset analysis gave insights into important trends and insights through visualization and data mining. The visualizations highlighted key patterns in vehicle registrations, manufacturer dominance, and consumer preferences. Tesla emerged as the leading manufacturer, with its **Model Y** and **Model 3** accounting for over half of the total registrations, highlighting its dominance in the electric vehicle market. Additionally, the geographic distribution of vehicles showed a significant number in urban areas like Washington, reflecting the influence of infrastructure and government policies on EV adoption. The data also revealed a difference in electric range and MSRP between Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), further highlighting the importance of affordability and performance in consumer choice.

Using the K-means clustering algorithm, the data mining process successfully segmented the dataset into two distinct groups: Cluster 0, dominated by BEVs, and Cluster 1, comprising mostly PHEVs. Cluster 0 represented newer, more affordable vehicles with higher electric ranges, while Cluster 1 contained older, more expensive models with lower ranges. This segmentation provides a clear understanding of market dynamics and vehicle attributes.

The findings have significant business intelligence applications. For manufacturers, insights into consumer preferences and market segments can inform production strategies and targeted marketing campaigns. Tesla and other BEV manufacturers can focus on expanding infrastructure to attract urban buyers, while PHEV makers might improve battery performance to increase their market share. Policymakers can use geographic insights to allocate resources for charging stations in high-demand areas. Dealerships can leverage the identified clusters to tailor promotions and improve customer engagement. These insights enable data-driven decision-making, promoting sustainable growth in the electric vehicle industry.