TITLE:

ELECTRIC VEHICLE POPULATION DATA

NAME:

SANTHOSH KUMAR B

DATE:

29/3/2025

COURSE:

DATA ANALYSIS

INTRODUCTION:

With the increase in realization world wide of the fact that conventional vehicles are soon going to be unsustainable due to depletion of fuel reserves, increased fuel costs and above all their devastating impact on the environment. Hence, many responsible countries have started adopting Electric Vehicle (EV) policies that will help them phase out conventional vehicles and adopt electric vehicles. This shift has necessitated to review the performance of electric vehicles. The major constraint in adoption of Electric Vehicles (EV) is range per charge and battery life of these vehicles. We are thankful to the government of State of Washington for releasing Electrical Vehicle Population dataset in the public domain for data analysts and data scientists along with other stakeholders to evaluate their dataset, present analysis and share insights.

Objective: The objective of this analysis is to present insights on adoption of the Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) that are currently registered through Washington State Department of Licensing (DOL).

DATA CLEANING QUESTIONS :

1.How many missing values exist in the dataset, and in which columns?

Base MSRP Column Analysis

Description: The "base msrp" column represents the Manufacturer's Suggested Retail Price for each product in the dataset. This price is recommended by the manufacturer and serves as a standard for retail pricing.

Data Type: Numeric

Count: 147027

Purpose and Importance:

The "base msrp" column is crucial for understanding the market value of products. It aids retailers in pricing strategies, allows for market comparisons, and helps evaluate sales performance against the suggested retail price.

Missing Values: The analysis revealed that there were too many missing values in the "base msrp" column, which were addressed by replacing them with the mean value of the column.

2.How should missing or zero values in the base MSRP and electric range columns be handled?

Zeros: A electric range total of 69698 entries had a value of zero, indicating potential missing data. These were replaced with the median value to maintain data integrity.

147027 have a total number of zeroes in the base msrp column

3.Are there duplicate records in dataset? If so how should they be managed?

YES, Handling duplicates in a large dataset requires a systematic approach to ensure data integrity while maintaining efficiency. Here's a step-by-step guide on how to identify, analyze, and manage duplicates in a large dataset, using tools like Excel, Python (with pandas), or SQL.

Analyze Duplicates

Understand the Context: Determine why duplicates exist. Are they due to data entry errors, merging datasets, or other reasons?

After handling duplicates, validate the dataset to ensure that the changes made are correct and that no important data has been lost.

Re-run your duplicate identification process to confirm that all duplicates have been addressed.

Keep a record of how duplicates were identified and handled. This documentation is crucial for reproducibility and for understanding the data cleaning process.

4. How can VIN be anonymized while maintaining uniqueness?

VIN number first three characters identifies the manufacturer and country of origin

And(4-9)characters provides information about the vehicle's model, body style, engine type,etc.

Data Integrity:

Unique identifiers ensure that each vehicle can be distinctly recognized, which is crucial for tracking ownership, service history, and recalls.

Regulatory Compliance:

Many regulations require that data remains identifiable for auditing and compliance purposes, even when anonymized.

Data Utility:

Maintaining uniqueness allows for effective data analysis and reporting, enabling organizations to derive insights without compromising individual privacy.

Fraud Prevention:

Unique identifiers help in preventing fraud by ensuring that each vehicle's history can be accurately traced and verified.

Operational Efficiency:

Unique VINs streamline processes in various sectors, including insurance, law enforcement, and automotive services, by providing a reliable reference point for each vehicle.

5.How can vehicle location(GPS coordinates) be cleaned or converted for better readability?

Use Descriptive Labels: Clearly label the coordinates as "Latitude" and "Longitude" to avoid confusion.

Use Decimal Degrees: Present coordinates in decimal degrees (e.g., 37.7749, -122.4194).

Label Clearly: Clearly label coordinates as "Latitude" and "Longitude".

Example: Latitude: 37.7749, Longitude: -122.4194

Add Location Names: Include the name of the location for context.

Example: San Francisco, CA

Format Consistently: Use a consistent format with clear separators.

Example: (37.7749, -122.4194)

Use Tables: Present multiple locations in a table for better organization.

Reverse Geocode: Convert GPS coordinates to human-readable addresses for clarity.

DATA EXPLORATION QUESTIONS:

1.what are the top 5 most common EV makes and models in the dataset?

The top most common **EV makes in BMW, Chevrolet, Ford, Nissan, Tesla**

Ev models are Tesla model Y, tesla model 3, Ford mustang mach-E, Chevrolet bolt, Tesla model S

These models reflect the current trends in Ev sales and popularity. The top 5 most common electric vehicle EV makes and models. These models highlight leading choices among consumers in EV market today.

2.What is distribution of EV’s by country? Which country has most registrations?

King -79075

Snohomish- 17307

Thurston -5403

Spokane -3690

Pierce -11542

3.How has EV adoption changed over different model years?

EV model year 2020 -11294 ,2021-18684, 2022-2779, 2023-37079 EV

4.What is average electric range of EV in dataset?

67.83KM is average electric range of EV in dataset.

5.what percentage of EV are eligible for CAFV?

41.81% percentage of EV are eligible for CAFV

6.What is average base MSRP for each EV model?

Manufacturer's Suggested Retail Price is average base E-tron ,Model -3,Rav 4

7.How does electric range vary across different makes and models?

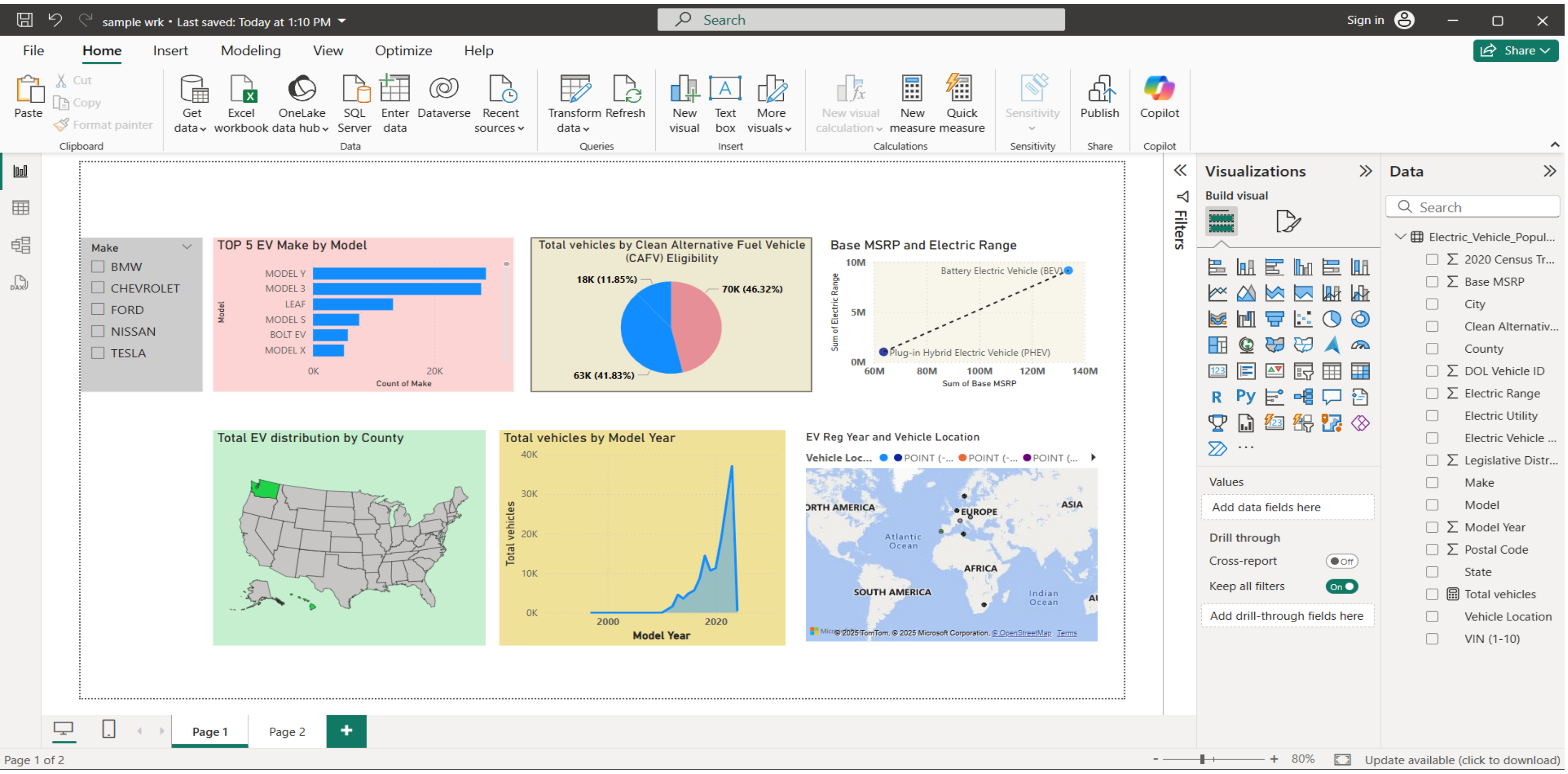
The most common EV makes in BMW, Chevrolet, Ford, Nissan, Tesla

Ev models are Tesla model Y, tesla model 3, Ford mustang mach-E, Chevrolet bolt, Tesla model S

Electric range measured in kilowatt-hours .the energy consumption rate km per kwh.

Many modern EV offers ranges from approx. 150km to 500km depending on make and model.

3.DATA VISUALISATION QUESTIONS:



1.**TOP 5 EV makes and models by count**

Using bar chart to find **top 5 makes and models is BMW, Chevrolet, Ford, Nissan, Tesla**

2.Use a filled map (I don't have heat map in visualization pane) to find EV distribution by country

3.A line graph to show **the trend of model year and total vehicles for each model .**

4.A Scatter plot comparing difference between **base MSRP vs electric range to see the pricing trends**

5.A pie chart showing variations **percent of CAFV eligible vs non-eligible EV**

6.use map visualization to display EV registrations based on vehicle location

4.LINEAR REGRESSION

1. How can we use Linear Regression to predict the Electric Range of a vehicle?

Linear Regression is a statistical method that models the relationship between a dependent variable (in this case, Electric Range) and one or more independent variables (features). By fitting a linear equation to the observed data, we can predict the Electric Range based on the values of the independent variables. The model will learn the coefficients for each feature, which represent the expected change in Electric Range for a one-unit change in that feature.

2. What independent variables (features) can be used to predict Electric Range?

Several independent variables can be used to predict Electric Range, including:

Model Year: Newer models may have better technology and efficiency.

Base MSRP: The price of the vehicle may correlate with its features and performance.

Make: Different manufacturers may have varying technologies and efficiencies.

Battery Capacity: Larger batteries typically provide a longer range.

Weight of the Vehicle: Heavier vehicles may have a shorter range due to increased energy consumption.

Tire Type and Size: These can influence rolling resistance and efficiency.

Motor Efficiency: Different motors may have different efficiencies.

3.How do we handle categorical variables like Make and Model in regression analysis?

Categorical variables can be handled using techniques such as:

One-Hot Encoding: This involves creating binary (0/1) columns for each category. For example, if "Make" has three categories (A, B, C), we create three new columns: Make_ A, Make _B, and Make_ C.

Label Encoding: Assigning a unique integer to each category. However, this method is less preferred for nominal categories as it may imply an ordinal relationship.

Dummy Variables: Similar to one-hot encoding, but one category is omitted to avoid multicollinearity.

4. What is the R² score of the model, and what does it indicate about prediction accuracy?

The R² score (coefficient of determination) measures the proportion of variance in the dependent variable (Electric Range) that can be explained by the independent variables in the model. It ranges from 0 to 1:

An R² score of 0 indicates that the model does not explain any of the variance.

An R² score of 1 indicates that the model explains all the variance.

A higher R² score generally indicates a better fit, but it is important to consider other metrics and potential overfitting.

5. How does the Base MSRP influence the Electric Range according to the regression model?

The influence of Base MSRP on Electric Range can be determined by examining the coefficient associated with the Base MSRP variable in the regression model. If the coefficient is positive, it suggests that as the Base MSRP increases, the Electric Range is expected to increase, indicating that more expensive vehicles may have better technology or larger batteries. Conversely, a negative coefficient would suggest that higher prices are associated with lower ranges, which may be counterintuitive.

6. What steps are needed to improve the accuracy of the Linear Regression model?

To improve the accuracy of the model, consider the following steps:

Feature Engineering: Create new features or transform existing ones to better capture relationships.

Data Cleaning: Remove outliers and handle missing values appropriately.

Feature Selection: Use techniques to select the most relevant features and reduce dimensionality.

Regularization: Apply techniques like Lasso or Ridge regression to prevent overfitting.

Cross-Validation: Use cross-validation to ensure the model generalizes well to unseen data.

Model Tuning: Adjust hyperparameters and consider using polynomial regression if relationships are non-linear.

7. Can we use this model to predict the range of new EV models based on their specifications?

Yes, once the Linear Regression model is trained and validated, it can be used to predict the Electric Range of new EV models based on their specifications, provided that the new data includes the same features used in training. However, the accuracy of predictions for new models will depend on how similar the new models are to the training data and whether the relationships learned by the model still hold true for the new data.