جامعة بيرزيت

BIRZEIT UNIVERSITY

ENCS5341, MACHINE LEARNING AND DATA SCIENCE

Assignment 1

Students Names:

Yafa Naji	**1200708**
Noor Shamali	**1200016**

Instructor: Dr. Yazan Abu Farha

Section: #3

Date: 30/10/2024

Table of Contents

Table of Figures:

1. Data Cleaning and Feature Engineering

1.1 Document Missing Values:

1_We identified the distribution of missing values across different columns in our data analysis, with the following findings:
1_The columns "**County**," "**City**," "**Postal Code**," and "**Electric Utility**" each have four missing values. Given the low percentage of missing values, we can handle them with simple techniques:

- ✓ Replace missing values with the most frequent value (mode).
- ✓ Alternatively, if these columns contain categorical data, substitute a placeholder like "Unknown" for missing values.

2_The columns "**Electric Range"** and "**Base MSRP**" each have five missing values. Since they contain numerical data, we can use imputation techniques such as:

- ✓ Replacing missing values with the mean or median to maintain the original data distribution.
- ✓ Alternatively, if the percentage of missing values is low, we may consider removing the rows with missing data.

3_We found a large number of missing values (445) in the "**Legislative District**" column, suggesting limited data availability in certain areas. Potential strategies include:

- ✓ Replacing missing values with a default value like zero (0) if district data is unknown.
- ✓ Considering removal of this column if it is not critical to our analysis.

4_For the "**Vehicle Location**" column, with only 10 missing values, we can handle it by:

- ✓ Using "Unknown" if location data is less critical.
- ✓ Replacing missing values with the most frequent location (mode).

5_The "2020 Census Tract" column has only four missing values, making their impact minimal. We can either:

- ✓ Remove these rows if they're not crucial to the analysis.
- ✓ Replace the missing values with the most frequent value.

Our analysis revealed that most columns have very few missing values, allowing us to use imputation techniques (mean or most frequent values) without affecting accuracy. For columns with a high number of missing entries (e.g., "Legislative District"), we considered deletion or using default values. After initial data cleaning, the dataset is now in good shape for further analysis, including statistical analysis, EDA, and advanced processing.

1.2 Missing Value Strategies: Analysis and Comparison:

Objective:

To evaluate the impact of different approaches for handling missing values in the dataset.

Strategies Applied:

➢ Dropping Rows with Missing Values:

- **Method**: Used df.dropna() to remove rows with missing values.
- **Impact**: Ensures complete data but may lead to data loss, impacting analysis quality if many rows are affected.

➢ Replacing Missing Values with the Mean for Numeric Columns:

- **Method**: Used df.fillna(df.mean()) to replace missing values in numeric columns.
- **Impact**: Maintains dataset size without data loss but may skew data if missing values deviate from the mean.

➢ Replacing Missing Values with the Most Frequent Value for Categorical Columns:

- **Method**: Used df.fillna(df.mode().iloc[0]) to fill missing categorical values.
- **Impact**: Preserves data integrity but can distort distribution by overrepresenting certain categories.

Comparison of Strategy Impact:

- ❖ **Dropping Rows**: Effective for minimal, random missing values but causes data loss.
- ❖ **Replacing with Mean**: Retains all rows but risks skewing statistics if missing values are far from the mean.
- ❖ **Replacing with Mode**: Preserves categorical data but can overrepresent some categories, impacting distribution.

1.3 Feature Encoding:

Feature Encoding: One-Hot Encoding Analysis:

Objective:

To simplify analysis and model building, we applied one-hot encoding to categorical columns "Make" and "Model."

Implementation:

Resulting dataset dimensions: (210,165 rows, 209 columns), reflecting:

- 210,165 rows representing electric vehicles.
- Increase to 209 columns due to new binary columns from one-hot encoding.

Results Analysis:

- **Dimension Increase**: One-hot encoding expanded column count, enabling the model to better differentiate vehicle types but also increasing model complexity and risk of overfitting.
- **Sparsity**: Many columns contain numerous zeros due to unique categories, leading to storage inefficiency and potential issues for algorithms requiring dense data.
- **Considerations for Further Analysis:** Dimensionality reduction or merging rare categories may be necessary to manage increased complexity.

1.4 Normalization:

Analysis of Normalization Results

After applying Min-Max normalization to numerical features, we observed:

- **Rescaling**: Features like "Postal Code," "Model Year," and "Electric Range" were transformed to a 0-1 range, removing scale differences.
- **Value Interpretation**: Normalized values now represent feature ranges as percentages, aiding interpretability (e.g., "Model Year" of 0.846 indicates proximity to the max value).
- **Consistency**: Ensures uniform feature scales, essential for distance-based models, preventing any feature from dominating due to scale.
- **Algorithm Performance**: Normalization aids in faster convergence and potentially better accuracy for machine learning models sensitive to feature scaling.
- **Prepared for Further Analysis**: Normalization establishes a solid foundation for additional analyses and algorithms requiring standardized data.

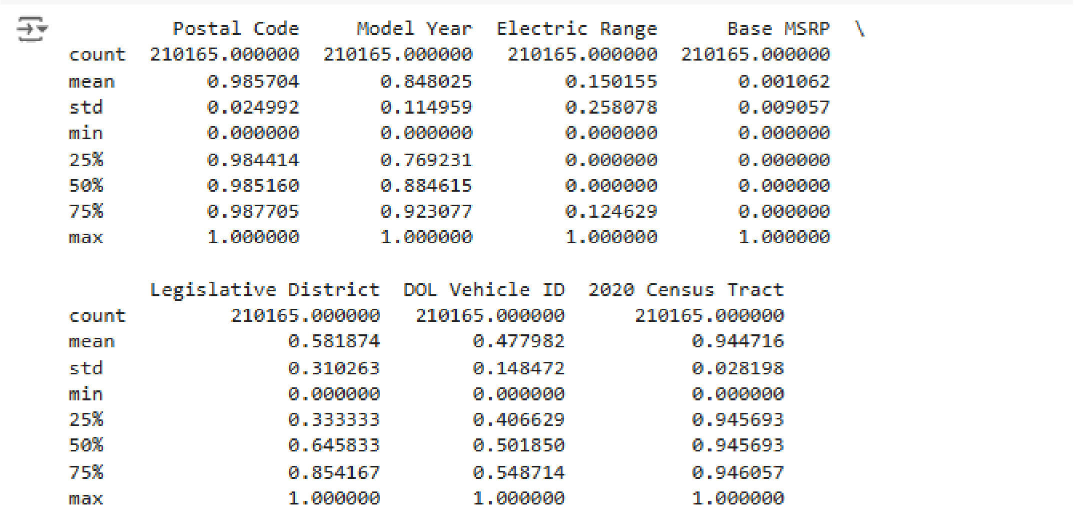
2. Exploratory Data Analysis:

2.1 Descriptive Statistics:

Exploratory Data Analysis: Descriptive Statistics Summary

- **Count**: Each numerical feature has 210,165 entries, confirming no missing values.
- **Mean Values**: Most electric vehicles are in the upper ranges of features, such as "2020 Census Tract" and "Postal Code." "Base MSRP" shows a low mean, warranting further exploration on pricing factors.
- **Standard Deviation**: Higher variability in "Legislative District" and "Electric Range" suggests distinctions that may need further analysis.
- **Min/Max Values:** Features span from 0 to 1, indicating normalization.

- **Quartiles**: "Model Year" quartiles show a concentration in recent years, highlighting the predominance of newer vehicles in the dataset.



```
        Postal Code    Model Year  Electric Range     Base MSRP  \
count  210165.000000  210165.000000  210165.000000  210165.000000
mean        0.985704       0.848025        0.150155       0.001062
std         0.024992       0.114959        0.258078       0.009057
min         0.000000       0.000000        0.000000       0.000000
25%         0.984414       0.769231        0.000000       0.000000
50%         0.985160       0.884615        0.000000       0.000000
75%         0.987705       0.923077        0.124629       0.000000
max         1.000000       1.000000        1.000000       1.000000

       Legislative District  DOL Vehicle ID  2020 Census Tract
count         210165.000000   210165.000000      210165.000000
mean               0.581874        0.477982           0.944716
std                0.310263        0.148472           0.028198
min                0.000000        0.000000           0.000000
25%                0.333333        0.406629           0.945693
50%                0.645833        0.501850           0.945693
75%                0.854167        0.548714           0.946057
max                1.000000        1.000000           1.000000
```

2.2 Spatial Distribution:

Objective: The goal is to analyze the geographic distribution of electric vehicles (EVs) using a heatmap, revealing patterns in EV concentration across various regions.

The whole map:

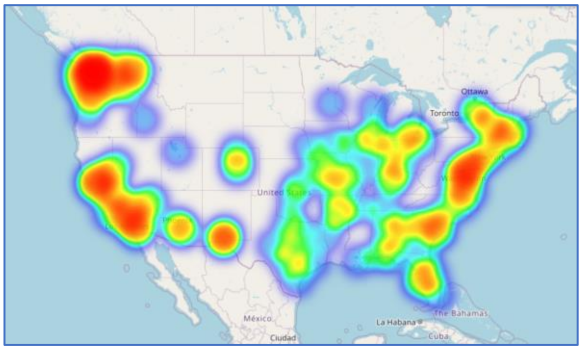


Figure 1 The complete map

Findings from the Map:

West Coast Dominance:

- California: High EV density around San Francisco Bay Area, Los Angeles, and San Diego, driven by environmental awareness, strong infrastructure, and government incentives.

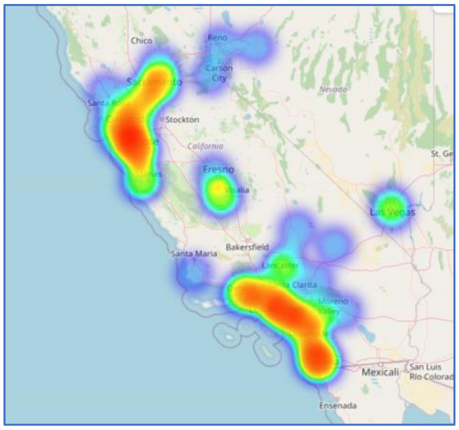


Figure 2 West Coast Dominance 1

- Pacific Northwest: Seattle and Portland also show significant concentrations, likely due to similar green policies.

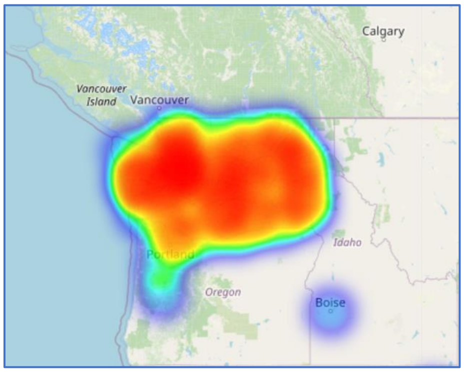


Figure 3 West Coast Dominance 2

East Coast Clusters:

- **Northeast Corridor:** High adoption from Boston through New York, Philadelphia, and down to Washington, D.C., indicating strong interest in sustainable transport.

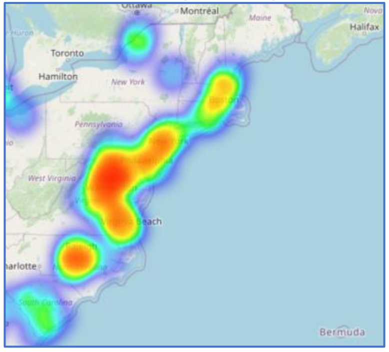


Figure 4 East Coast Clusters 1

- **Southeast Growth:** Areas like Raleigh, NC, show increasing EV activity, though not as dense as the Northeast

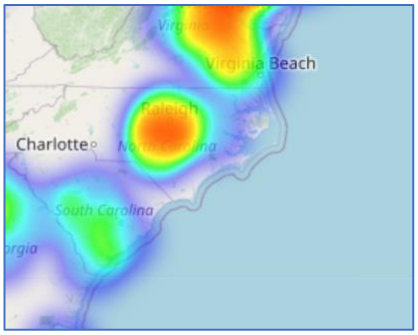


Figure 5 East Coast Clusters 2

Midwest and Mountain West:

- Urban clusters in cities like Denver, CO, and Salt Lake City, UT, show localized adoption in these regions.

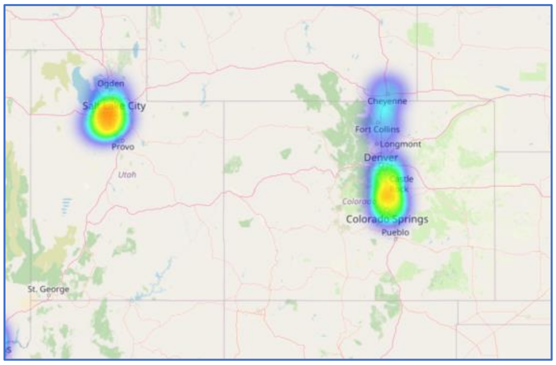


Figure 6 Midwest and Mountain West

Southern United States:

- Moderate activity in Texas (Austin, San Antonio, Houston) and Florida (Miami, Orlando), showing gradual growth in EV interest.

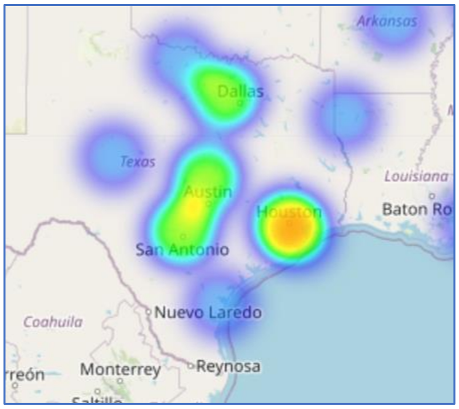


Figure 7 Southern United States 1

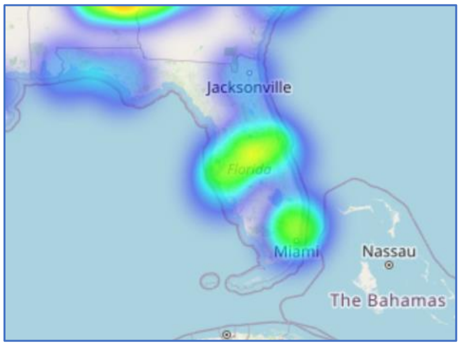


Figure 8 Southern United States 2

Findings and Implications

- **Urban vs. Rural Divide:** Higher adoption in urban/suburban areas due to population density, infrastructure, and incentives.
- **Coastal vs. Inland Adoption:** Coastal areas lead in EV adoption, while inland regions lag due to infrastructure gaps and different priorities.

2.3 Model Popularity:

Objective:

To analyze the popularity of EV models based on registration frequency, providing insights into market leaders and consumer trends.

The output:

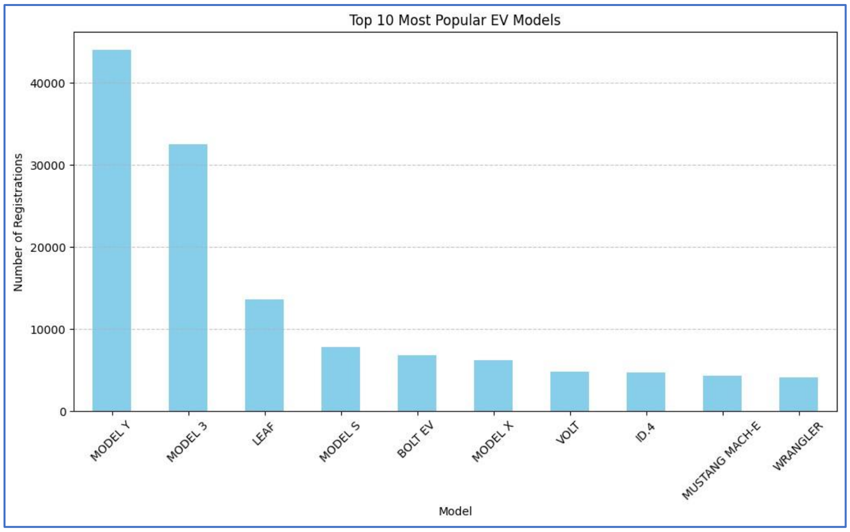


Figure 9 Analyze the popularity of different EV models

Findings and Analysis from the Output:

1. **Tesla Dominance**
    - **Model Y** is the most popular, followed by **Model 3** with over 40,000 and 30,000 registrations, respectively.
    - Tesla's success is attributed to its charging infrastructure, vehicle performance, and strong brand trust.
2. **Model Diversity**
    - **Nissan Leaf** ranks third, highlighting demand for affordable EVs.
    - Top 10 includes both budget-friendly (e.g., Chevrolet Bolt EV) and luxury models (e.g., Model S, Model X), showing varied consumer preferences.
3. **Market Trends**
    - Affordable models like the Leaf and Bolt EV are essential, while luxury models retain a loyal segment.
    - New entries like the **Mustang Mach-E** signal growing consumer openness to non-Tesla EVs, hinting at an increasingly competitive market.

Conclusion

Tesla leads with the Model Y and Model 3, but traditional automakers are gaining ground. This trend reflects a diversifying EV market with options for both budget-conscious and premium-seeking consumers.

2.4 Correlation Analysis Between Numeric Features

Objective:

To examine relationships between numeric features in the dataset, using a correlation matrix to identify positive or negative associations and gain insights into their interactions.

Correlation Matrix:

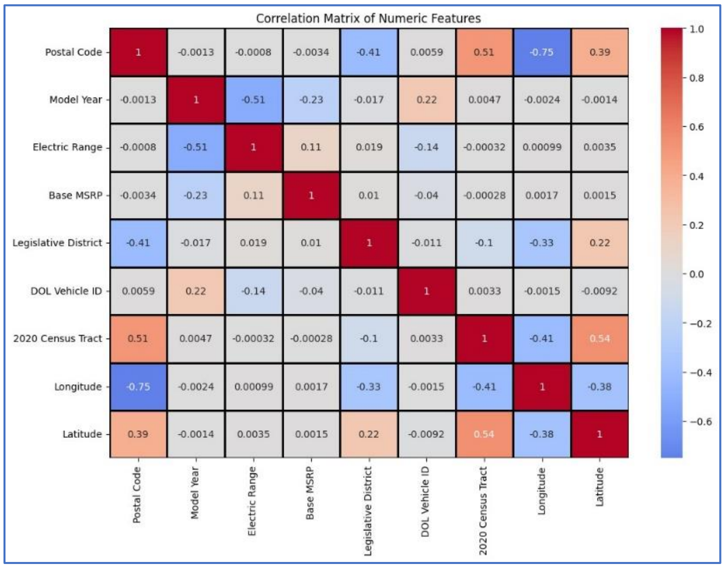


Figure 10 Correlation Matrix

Key Findings and Insights from the Correlation Matrix:

- **Model Year and Electric Range** (-0.51)
  - Newer models show shorter ranges, potentially indicating a trend towards more affordable, city-oriented EVs over long-range options.
- **Longitude and Postal Code** (-0.75)
  - Reflects geographic logic: moving west (decreasing longitude) corresponds to lower postal codes, consistent with U.S. postal code assignments.
- **Latitude and 2020 Census Tract** (0.54)
  - Moderate positive relationship, aligning population patterns with northern geographic regions.
- **Electric Range and Base MSRP (Price)** (0.11)

    - Weak positive correlation; higher-range EVs tend to cost more, though other factors like brand and features influence pricing.

- **Model Year and DOL Vehicle ID** (0.22)
    - Sequential ID assignment is evident, with newer vehicles having higher Department of Licensing (DOL) IDs.

- **Longitude and Latitude** (-0.38)
    - Moderate negative relationship, expected given that moving west correlates with a slight decrease in latitude.

- **Legislative District**
    - Shows minimal correlation with other features, indicating political boundaries do not significantly affect EV characteristics like price or range.

3. Visualization:

3.1 Data Exploration Visualizations:

1. Histogram: Distribution of Electric Range

- **Description**:

  This histogram shows the distribution of electric range across all EVs in the dataset. The range is divided into 20 bins.

- **Key Insights**:

○ Most EVs have a low range (0-50 miles), likely due to the inclusion of hybrids or budget models. Higher ranges (150+ miles) are less common, possibly indicating premium models.

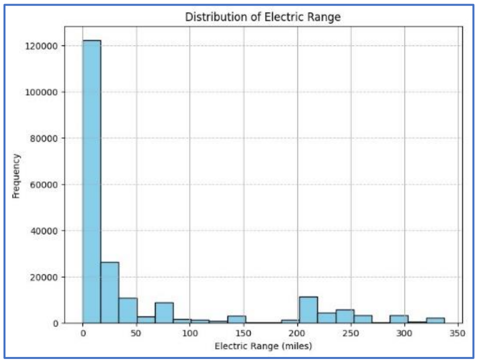


Figure 11 Distribution of Electric Range

2. Boxplot: Electric Range by Model Year

- **Description**:

  This boxplot visualizes how the electric range varies over different model years.

- **Key Insights**:

○ Significant range improvement is evident from 2015 onwards, likely due to advances in battery technology. Some recent years show outliers with high ranges (300+ miles), while recent declines may indicate more urban-focused or hybrid models.

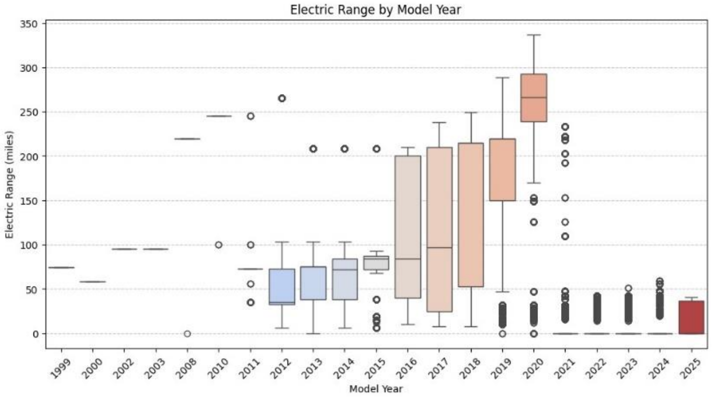


Figure 12 Electric Range by Model Year

3. Scatter Plot: Electric Range vs. Base MSRP

- **Description**:

  This scatter plot shows the relationship between electric range and base price (MSRP) of the vehicles.

- **Key Insights**:

○ No strong correlation between price and range. Affordable EVs (<$50,000) still offer reasonable range, while outliers with high MSRPs represent luxury models.

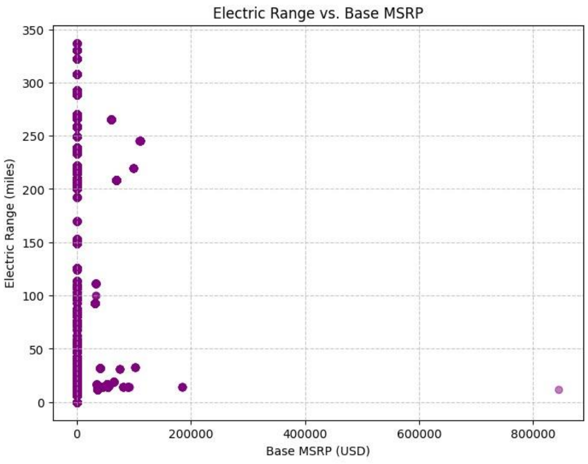


Figure 13 Electric Range vs. Base MSRP

4. Scatter Plot: Electric Range vs. Model Year

- **Description**:
  This plot explores how electric range varies with the model year.
- **Key Insights**:

○ Clear increase in range for newer models, with older models (pre-2010) showing limited range (<50 miles). Range diversity from 2018 onwards indicates a broader selection of long- and short-range EVs.

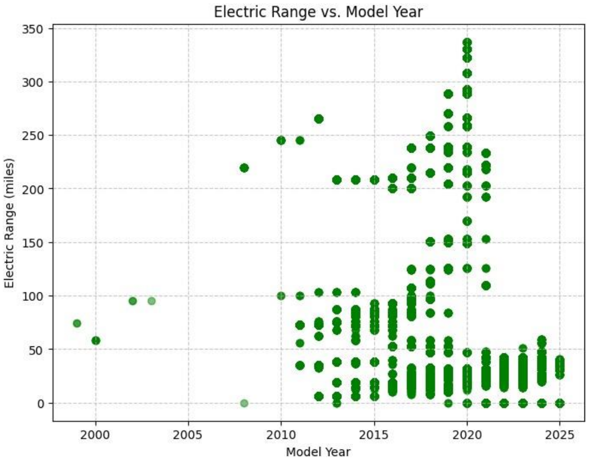


Figure 14 Electric Range vs. Model Year

5. Histogram: Distribution of Base MSRP

- **Description**:

  This histogram shows the distribution of base prices (MSRP) for EVs.
- **Key Insights**:
  - Most vehicles are priced below $50,000, reflecting the market's shift towards affordability.

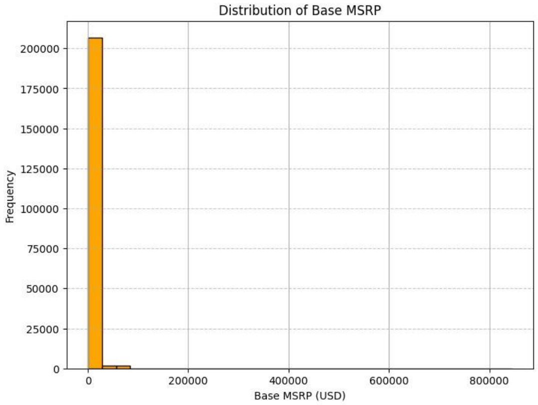


Figure 15 Distribution of Base MSRP

6. Boxplot: Base MSRP by Model Year

- **Description**:

  This boxplot shows how the base MSRP of vehicles varies by model year.
- **Key Insights**:
  - Consistent pricing observed from 2010 to 2020, with recent years showing a trend towards more affordable models. Outliers with high MSRPs suggest the presence of luxury options.

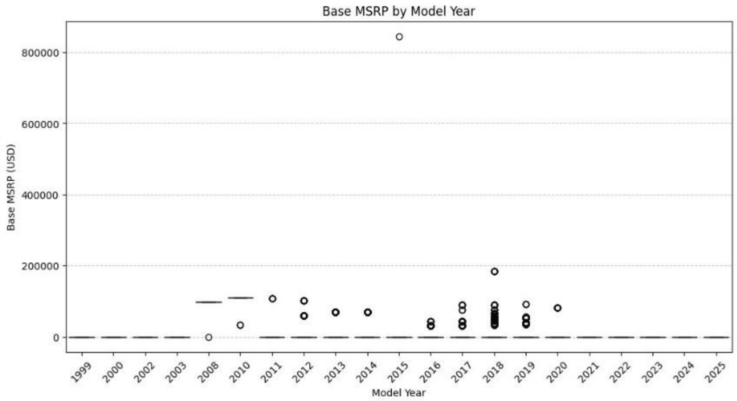


Figure 16 Base MSRP by Model Year

Overall Findings

- **Technological Progress:** Noticeable improvements in EV range, especially post-2015, highlight advancements in battery technology.
- **Affordability Focus:** The majority of EVs are priced below $50,000, aligning with a push towards mass-market accessibility.
- **Diverse Options:** The EV market now includes both short- and long-range models, catering to various consumer needs.
- **Price-Range Dynamics:** Weak correlation between price and range indicates that improved battery technology allows for reasonable ranges at accessible prices.

3.2 Comparative Visualization - Analysis of EV Distribution Across Locations

Objective: Analyze the distribution of EV registrations across different cities and counties, revealing adoption patterns and regional trends.

1. Top 10 Cities by Number of EV Registrations

- Seattle leads significantly with over 30,000 EV registrations, indicating it as a central hub for EV adoption.
- Bellevue and Vancouver show considerable EV usage, but at much lower numbers than Seattle.
- Cities like Redmond, Bothell, and Renton have similar registration levels, indicating consistent adoption.
- Insight: Seattle’s strong EV presence may reflect greater environmental consciousness, strong infrastructure, or local policies favoring EVs.

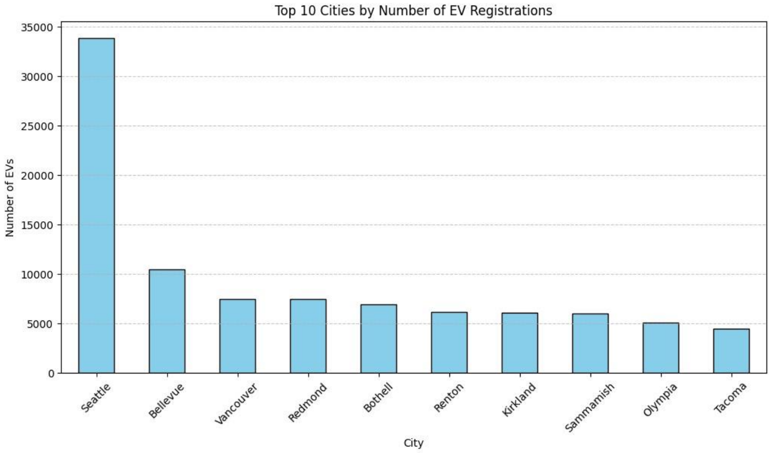


Figure 17 Cities by Number of EV Registrations

2. Top 10 Counties by Number of EV Registrations

- King County dominates with over 100,000 registrations, reinforcing Seattle’s position as an EV hotspot within this county.
- Neighboring counties, Snohomish and Pierce, have over 20,000 registrations each, suggesting broader regional adoption.
- Clark, Thurston, and Kitsap counties show moderate EV adoption, which may reflect differences in infrastructure or policy support.
- Insight: King County’s high EV registration numbers likely result from favorable policies, infrastructure, and socio-economic factors promoting EV adoption.

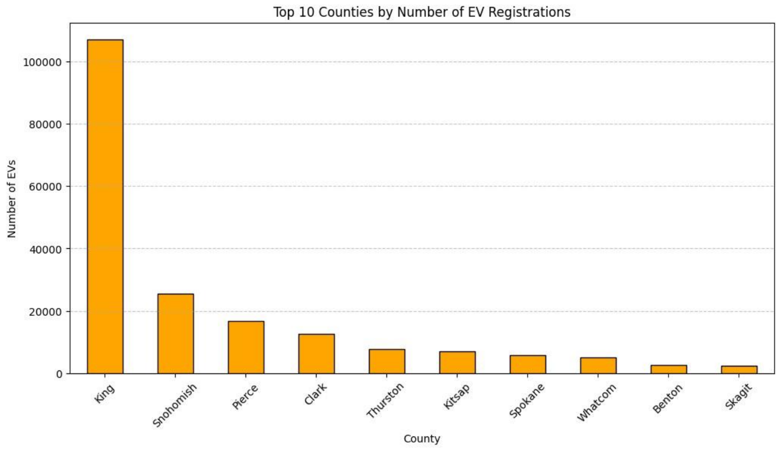


Figure 18 Top 10 Counties by Number of EV Registrations

3. Stacked Bar Chart - Distribution of EV Types Across Top 5 Counties

- King County has the largest share of both Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs), showing a diverse adoption pattern.
- In Snohomish County, BEVs lead but PHEVs also represent a significant share.
- Pierce, Clark, and Thurston counties follow this trend, with BEVs dominating but PHEVs maintaining a visible presence.
- Insight: BEVs are more popular across top counties, indicating a shift towards fully electric options, while PHEVs still attract users who prefer hybrid flexibility.

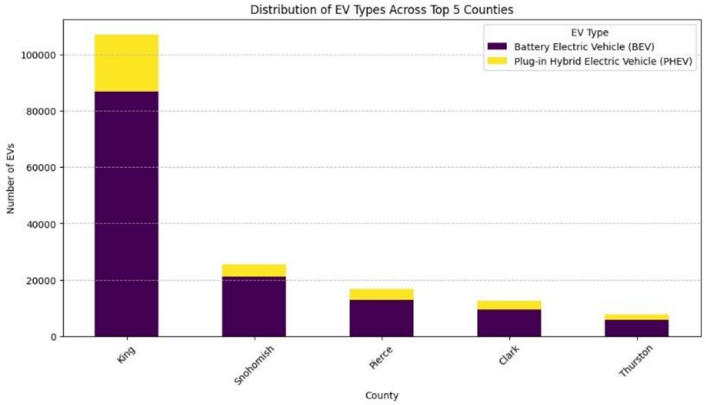


Figure 19 Distribution of EV Types Across Top 5 Counties

4. Additional Analysis:

4.1 Temporal Analysis of EV Adoption Rates and Model Popularity

1. EV Registrations Over Time (by Model Year)

- Trend: EV registrations remained low from 2000 to 2010, began rising around 2015, and sharply increased post-2018.
- Peak: Registrations peaked in 2023, reflecting a surge in EV adoption, followed by a sharp decline, potentially due to incomplete data for 2025 or temporary market factors.
- Insight: The upward trend highlights growing consumer interest, spurred by technological advancements, incentives, and environmental awareness. The recent drop may suggest market saturation or challenges like supply chain issues.

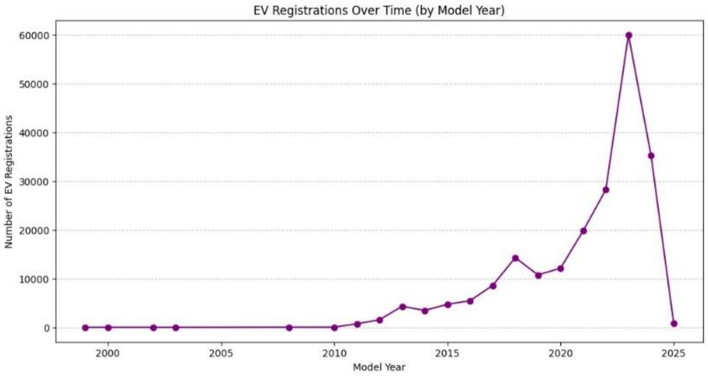


Figure 20 EV Registrations Over Time (by Model Year)

2. Popularity Trends of Top 5 Models Over Time

- Model Trends: MODEL Y leads with significant growth, especially in 2023, while MODEL 3 also shows steady popularity. Other models, such as the LEAF, BOLT EV, and MODEL S, display modest growth.
- Insight: Tesla’s dominance is clear, driven by strategic expansion, strong infrastructure, and consumer preference.

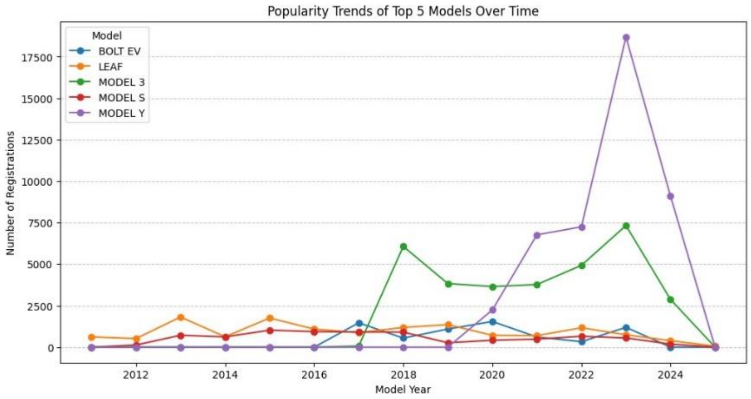


Figure 21 Popularity Trends of Top 5 Models Over Time

3. Adoption Trends of EV Types (BEV vs. PHEV) Over Time

- BEV Dominance: BEVs have grown exponentially since 2015, while PHEVs grew slower and occupy a smaller market share.
- Insight: The shift towards BEVs suggests consumer preference for fully electric vehicles, likely due to better battery tech, reduced maintenance, and stronger incentives for zero-emission options.

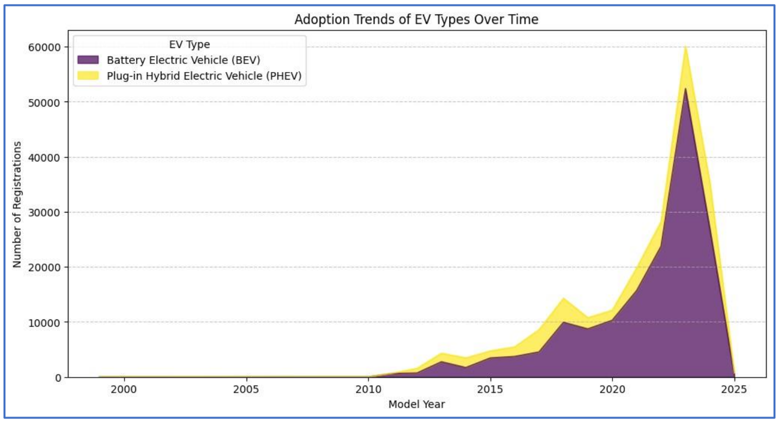


Figure 22 BEV vs. PHEV