

Faculty of Engineering and Technology

Electrical and Computer Engineering Department

Machine Learning and Data Science ENCS5341

**Assignment 1**

Exploratory Data Analysis of Electric Vehicle Registrations in Washington State: Trends, Insights, and Spatial Distributions.

**Prepared by:** Adam Nassan, 1202076

**Instructor:** Dr. Ismail Khater

**Section:** 3

**Date:** 27-10-2024

# Introduction

**Dataset Description**: This dataset, titled Electric Vehicle Population Data, is sourced from the State of Washington’s public records on electric vehicles (EVs), available on Data.gov. It captures a range of details regarding electric vehicle registrations, including vehicle make, model, electric range, MSRP, and geographic location (e.g., county). This information provides insights into EV adoption rates and helps analyze trends in the EV market, supporting environmental and infrastructural planning, policy-making, and targeted EV incentives.

The analysis aims to:-

* Clean and preprocess the dataset by handling missing values and encoding categorical features.
* Perform exploratory data analysis (EDA) to uncover key descriptive statistics, trends, and spatial distribution patterns within the EV population.
* Explore relationships between features (e.g., electric range, base MSRP) to identify correlations and consumer preferences.
* Highlight the popularity of specific models, spatial concentration, and growth trends, enabling a comprehensive understanding of the electric vehicle landscape in Washington State.

# Data cleaning and feature engineering

**Missing Values**: Upon analysis, the dataset exhibited missing values across various features, with each feature's missing count and percentage calculated to gauge impact. Missing values in numerical fields were handled using:

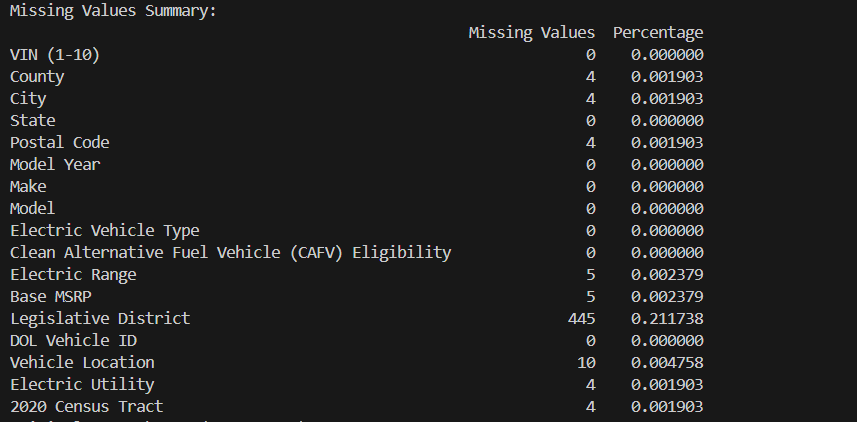


Figure 1: Missing values percentage

* **Mean Imputation**: This strategy replaced missing numerical values with the mean of each feature, effectively filling gaps while preserving the dataset's overall shape.
* **Median Imputation**: Replaced missing values with the median, potentially better suited for skewed data by avoiding the influence of outliers.
* **Row Dropping**: Rows with any missing values were removed, providing a complete but reduced dataset.

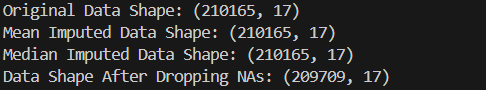


Figure 2: Handle missing values strategies

Each method was tested to observe how the changes impacted data shape and integrity. Mean and median imputation retained the dataset's dimensions, while row dropping resulted in data loss, which could limit insights for certain analyses.

**Feature Encoding**: To handle categorical data effectively, one-hot encoding was applied to features like Make and Model, creating binary columns for each unique category. This process:

* Added columns corresponding to each make and model, which increased the dataset’s dimensionality.

The resulting dataset structure improved compatibility for numerical analysis, enabling comparisons across EV makes and models.

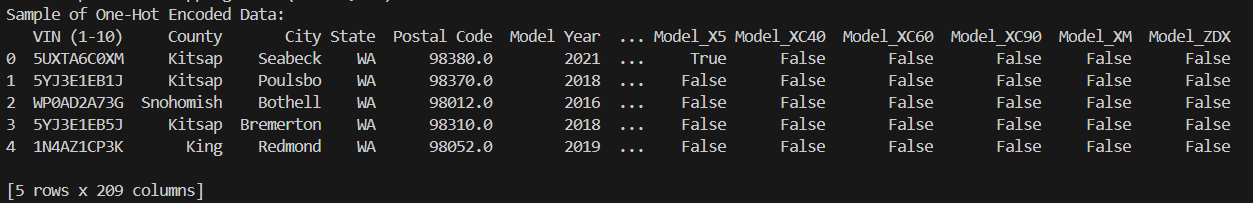


Figure 3: Features Encoding

**Normalization**: Normalization was applied to numerical features (e.g., Electric Range, Base MSRP, Model Year) using MinMax scaling, transforming values to a range between 0 and 1. This scaling:

* Allowed consistent feature ranges, preventing features with larger scales from dominating analyses.
* Enhanced interpretability; for instance, an Electric Range value close to 1 represented the maximum observed range, while values near 0 denoted the lowest.

This process standardized feature comparison and was essential for fair analysis in relationship explorations.

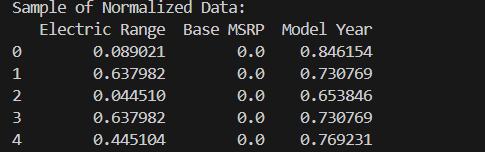


Figure 4: Normalized data

# Exploratory Data Analysis (EDA)

**Descriptive Statistics**

In examining the key descriptive statistics (mean, median, and standard deviation) of numerical features like Electric Range and Base MSRP, we gain insights into the central tendencies and variability in the EV dataset.

* **Electric Range**:
  + Mean and Median: These values provide an indication of the typical electric range for vehicles in the dataset. If the mean and median are close, the distribution is likely symmetric. A lower median compared to the mean would suggest a right-skewed distribution, potentially due to a few vehicles with very high ranges.
  + Standard Deviation: This metric reflects the spread in electric ranges, indicating whether most vehicles have similar ranges or if there is a wide variation.
* **Base MSRP**:
  + Mean and Median: Similar to Electric Range, the mean and median values for Base MSRP reveal the typical price point of vehicles. A significant difference between the mean and median could suggest a skewed price distribution, likely influenced by high-end EV models.
  + Standard Deviation: A large standard deviation here would indicate that the dataset includes vehicles with a wide range of prices, from more economical models to premium ones.

These statistics help identify the common ranges and prices that may influence consumer choices. For instance, a relatively high standard deviation in Base MSRP and Electric Range can highlight diverse consumer preferences in terms of budget and required driving range.

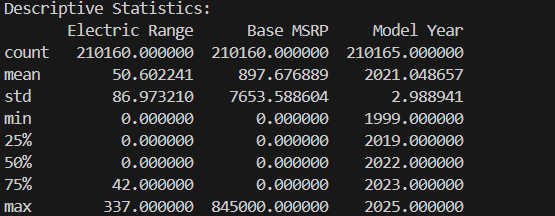


Figure 5: Descriptive statistics

**Spatial Distribution**

Analyzing the geographic distribution of EVs across various counties reveals how EV adoption varies by location. Using bar charts or a map visualization, we can observe which counties have higher concentrations of EVs. Urban counties with more developed infrastructure and greater access to charging facilities often have higher EV counts.

**Trends and Influencing Factors**:

* **Urban Areas**: Higher EV density in urban areas can indicate greater adoption rates due to infrastructure readiness, environmental policies, or urban incentives.
* **Rural Areas**: Lower EV counts in rural counties might reflect reduced access to charging stations or different lifestyle needs (e.g., higher preference for gasoline vehicles for longer distances).
* **Policy Influence**: Certain counties with local incentives may show a larger number of EVs, highlighting the impact of targeted government support in driving adoption.

This analysis informs policymakers and stakeholders about the regions where EV adoption is high and the areas where increased infrastructure could support EV growth.

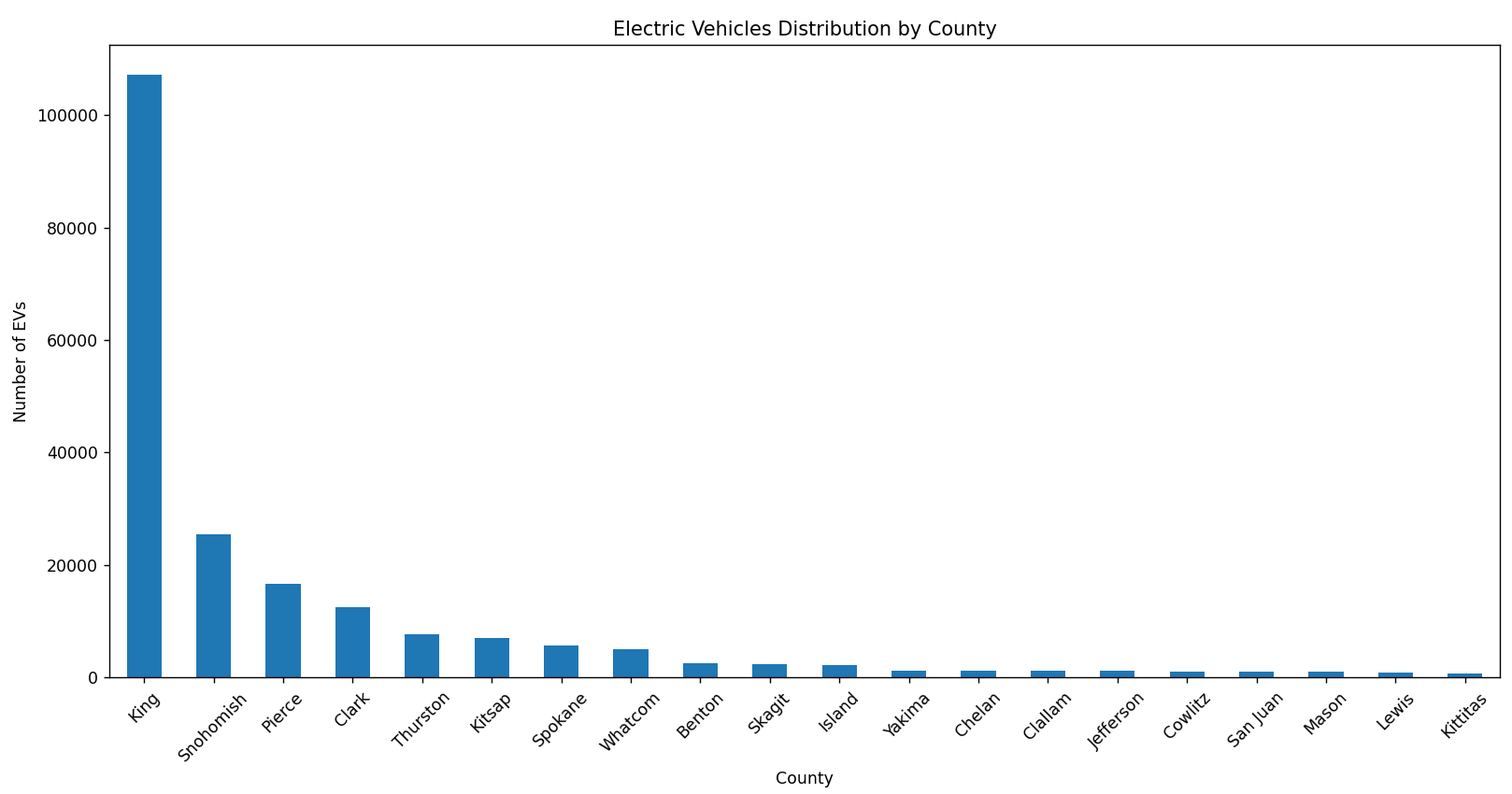


Figure 6: Vehicle distribution by county

**Model Popularity**

The popularity of EV models, visualized through a bar chart, shows which models are most commonly registered. This analysis reveals:

* **Brand Dominance**: Certain brands may dominate the market, indicating strong consumer preference for specific manufacturers. Popular brands often have a reputation for reliability, affordability, or range, which could influence buyers.
* **Type Preference**: A trend toward specific EV types (e.g., SUVs vs. sedans) may reflect consumer needs based on family size, budget, or driving habits.

From a market perspective, high popularity for particular brands and models suggests where the highest demand lies, allowing manufacturers to target features or services that appeal to this market segment.

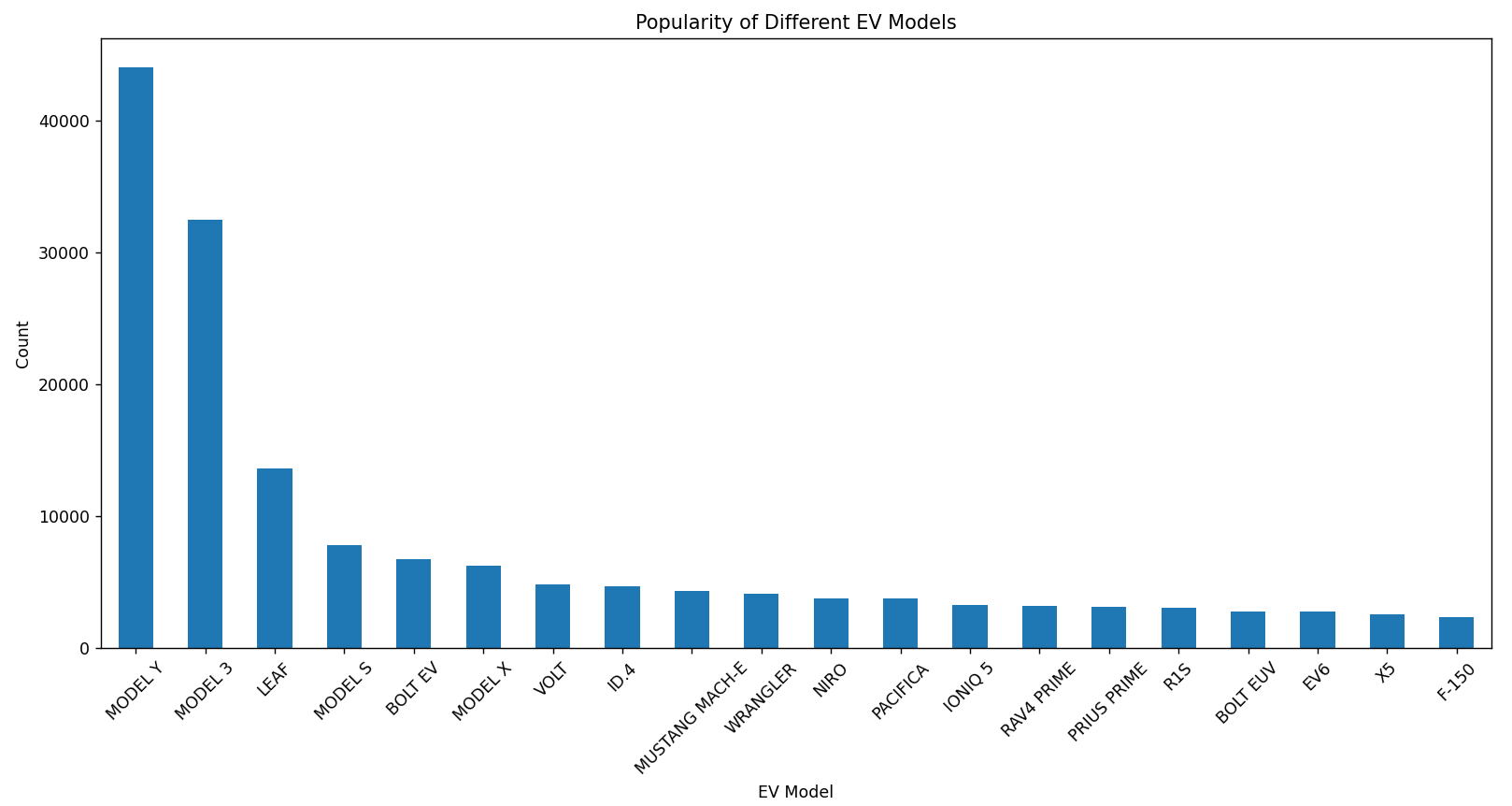


Figure 7: Models popularity

**Pairwise Feature Relationship Analysis**

**Correlation Analysis**

The correlation matrix and heatmap provided insights into relationships between numerical features, especially Electric Range, Base MSRP, and Model Year.

* **Electric Range and Base MSRP**: A positive correlation between Electric Range and Base MSRP suggests that vehicles with longer ranges tend to have higher base prices, likely due to more advanced battery technology in high-range models. This trend may indicate that consumers willing to pay more prefer vehicles capable of longer trips without charging.
* **Model Year and Electric Range**: A weaker correlation between Model Year and Electric Range implies that newer models may not always have extended ranges. This could reflect a focus on optimizing other features, such as efficiency or affordability, rather than purely increasing range in newer models.

The heatmap visually emphasizes these correlations, allowing us to easily identify stronger and weaker relationships.

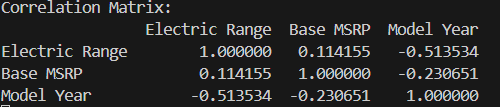


Figure 8: Correlation analysis

# **Visualization Analysis**

**Data Exploration Visualizations**

Exploratory data visualizations, including histograms, scatter plots, and box plots, are instrumental in uncovering the characteristics and variability within the dataset.

* **Histogram of Electric Range**:
  + The histogram for Electric Range reveals the distribution of EV ranges in the dataset. A peak in the histogram indicates the most common range value, reflecting the preferences or availability of EVs with that range. This visualization shows the concentration of ranges around certain values, helping to identify whether most vehicles are intended for shorter or longer trips.

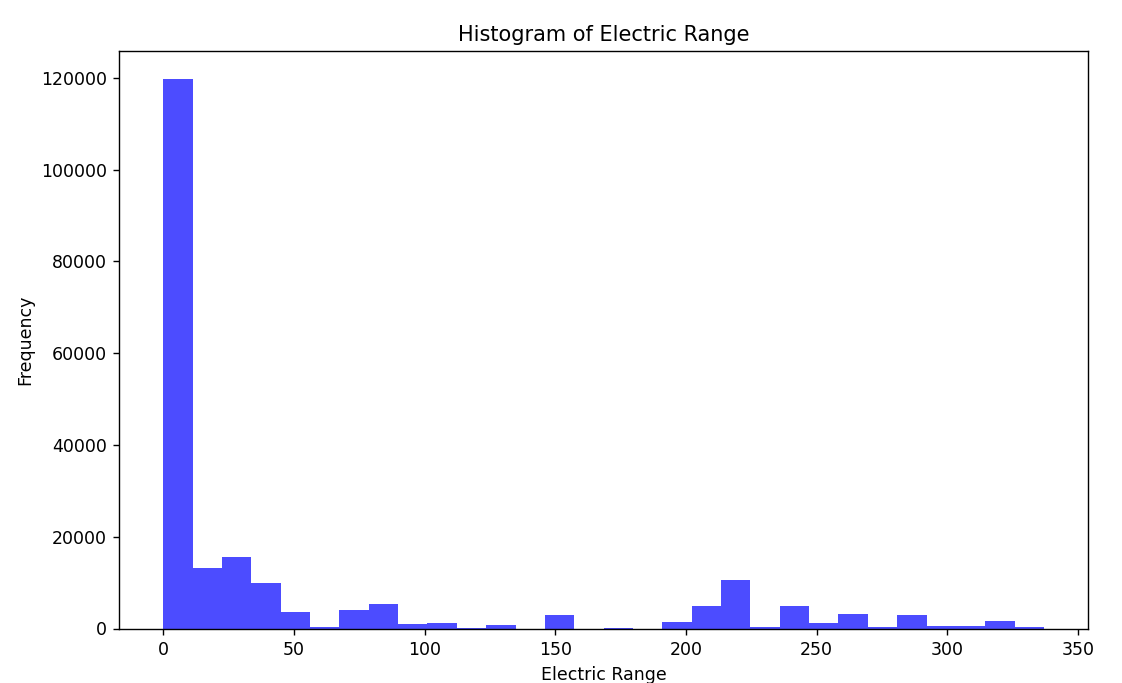


Figure 9: Electric Range frequency Histogram

* **Box Plot by Model Year**:
  + **Range Variation by Model Year**: From around 2015 to 2019, the spread (interquartile range) of electric range values is wider, indicating a greater variability in EV ranges available in those years. This may reflect a variety of models catering to different needs, from short-range to long-range options.
  + **Recent Decline in Range**: After 2019, the range seems to drop, with more recent models (2021-2025) showing lower median values and less variation. This could imply that the dataset includes many low-range EVs from these years or that there is less data for newer models.
  + **Outliers**: The dots outside the whiskers represent outliers, showing some models with significantly higher or lower ranges than the typical values for each year.
  + **Mature Market Trend**: From around 2018 onward, the electric range does not increase as significantly as in previous years, which could imply that EV range improvements are stabilizing.

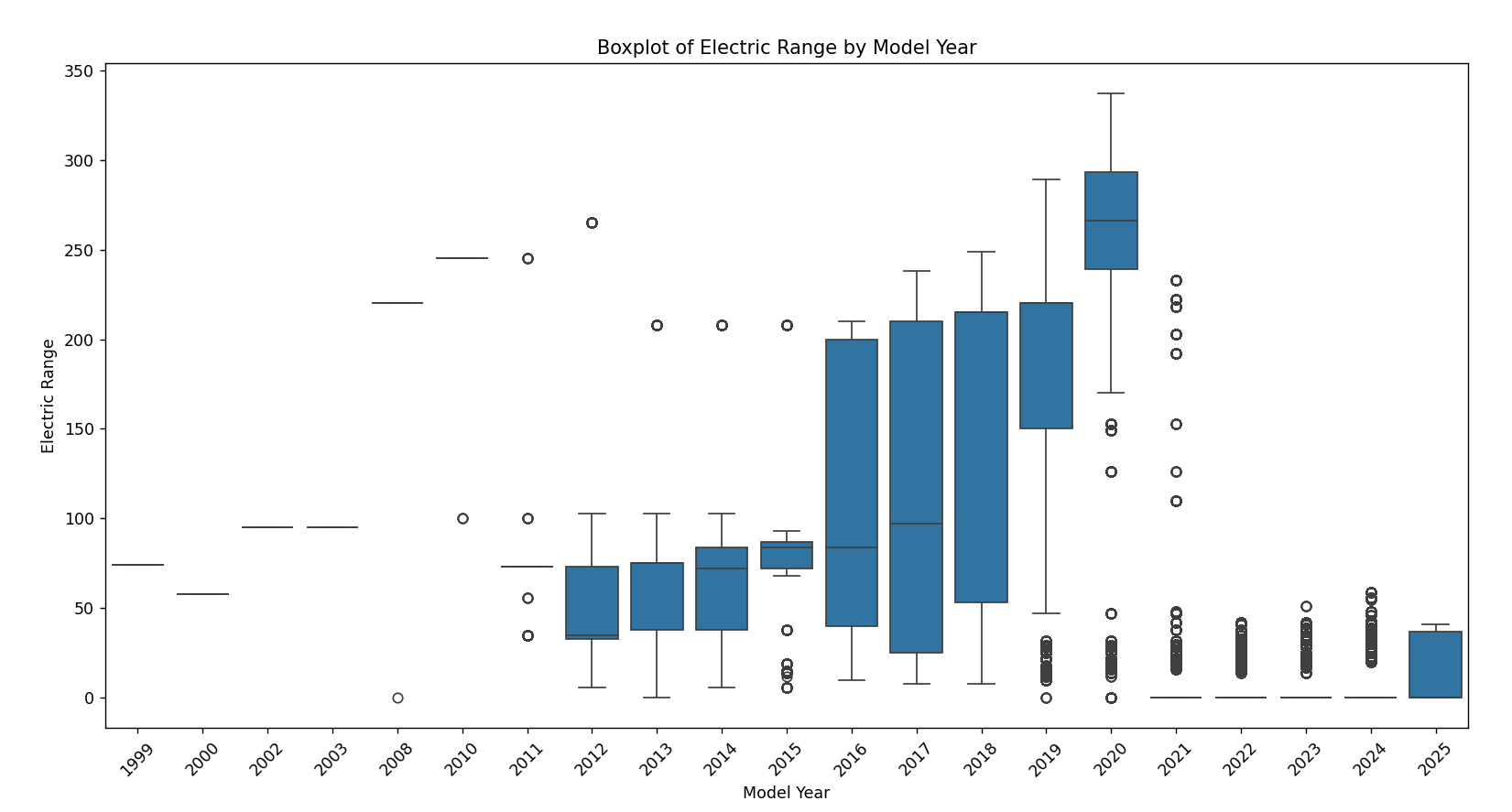


Figure 10: Electric Range vs Model Year BoxPlot

* **Scatter Plot of Electric Range vs. Base MSRP**:
  + The scatter plot comparing Electric Range and Base MSRP displays how price correlates with range. Clusters within this plot reveal typical price-range combinations, and a visible upward trend indicates that longer-range vehicles generally come with a higher price tag. This visualization helps identify market segments, like mid-range EVs at moderate prices, and provides insights into pricing based on consumer demand for longer ranges.

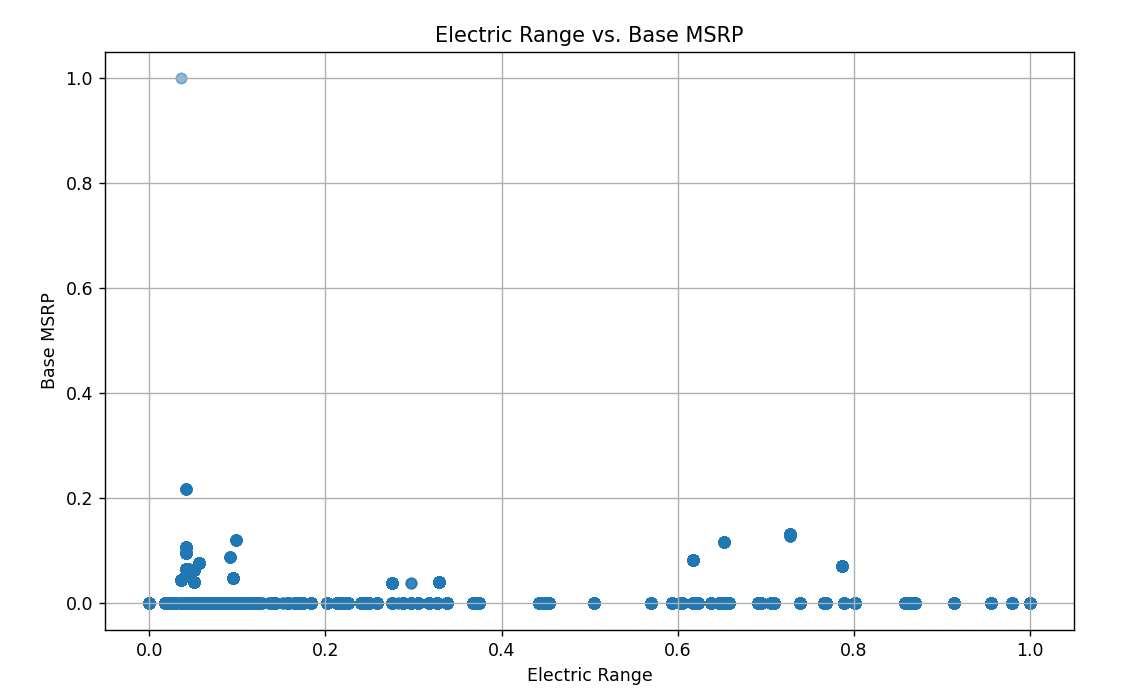


Figure 11: Base MSRP vs Electric Range ScatterPlot

These visualizations clarify the distribution and variability of key features, helping us understand common vehicle characteristics, the influence of technological advancement, and price relationships.

**Comparative Visualization**

To analyze the geographic distribution of EV registrations, bar charts and stacked bar charts were used to compare EV counts across different locations.

* **Bar Chart of EV Registrations by County**:
  + A bar chart of EV registrations across counties shows the concentration of EVs, typically highlighting urban counties with high EV counts. This is likely due to better charging infrastructure, environmental incentives, and greater public awareness. Counties with lower counts may indicate regions where EV adoption is still growing or where infrastructure needs improvement.

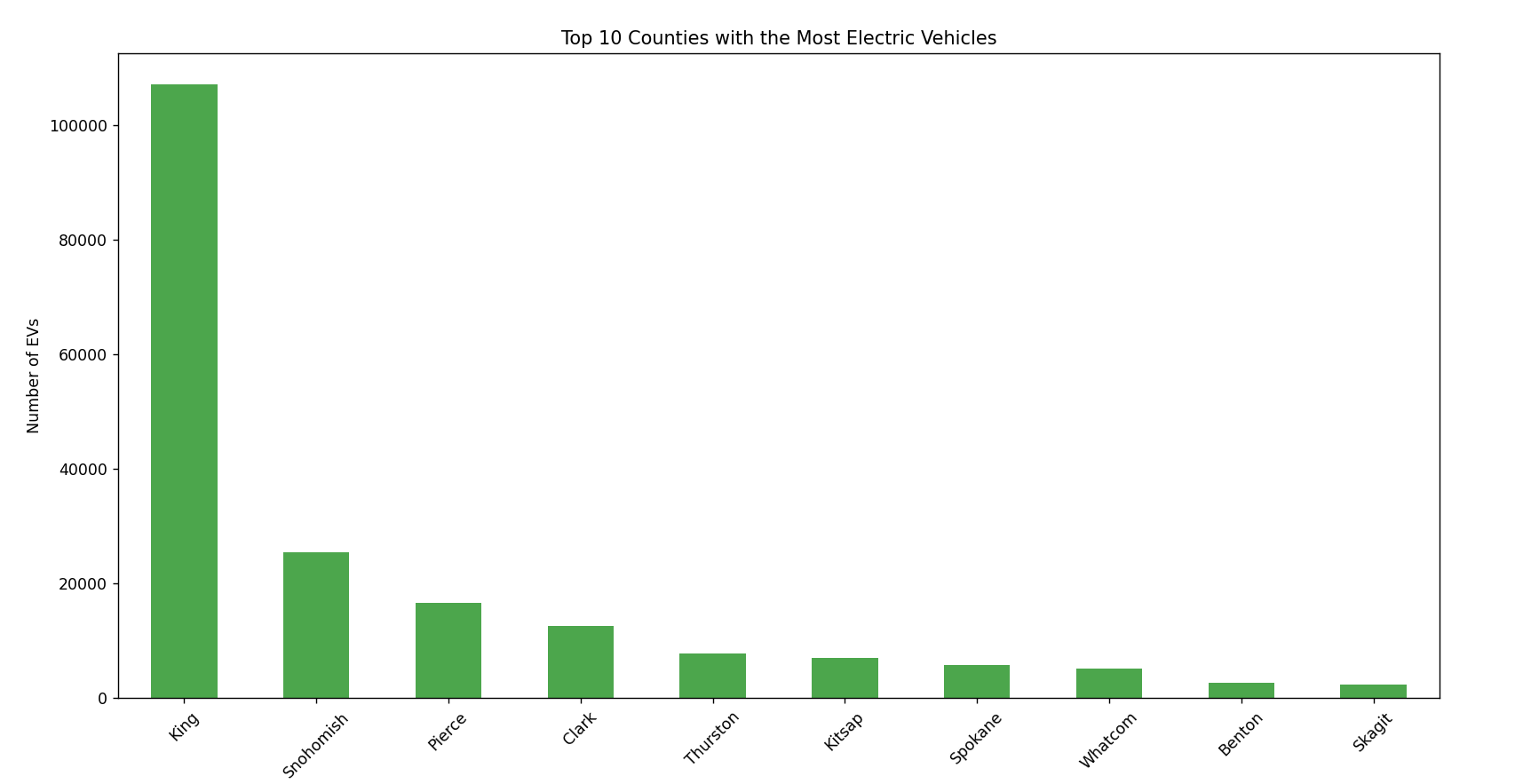


Figure 12: Top 10 Counties with most EVs

* **Stacked Bar Chart for EV Types by County**:
  + Stacked bar charts illustrate the composition of different EV types (e.g., plug-in hybrid vs. all-electric) across various counties. This comparison provides insights into the popularity of EV types within urban vs. rural areas. Urban areas may have a higher proportion of all-electric vehicles due to proximity to charging stations, while rural areas might show a stronger preference for hybrids.

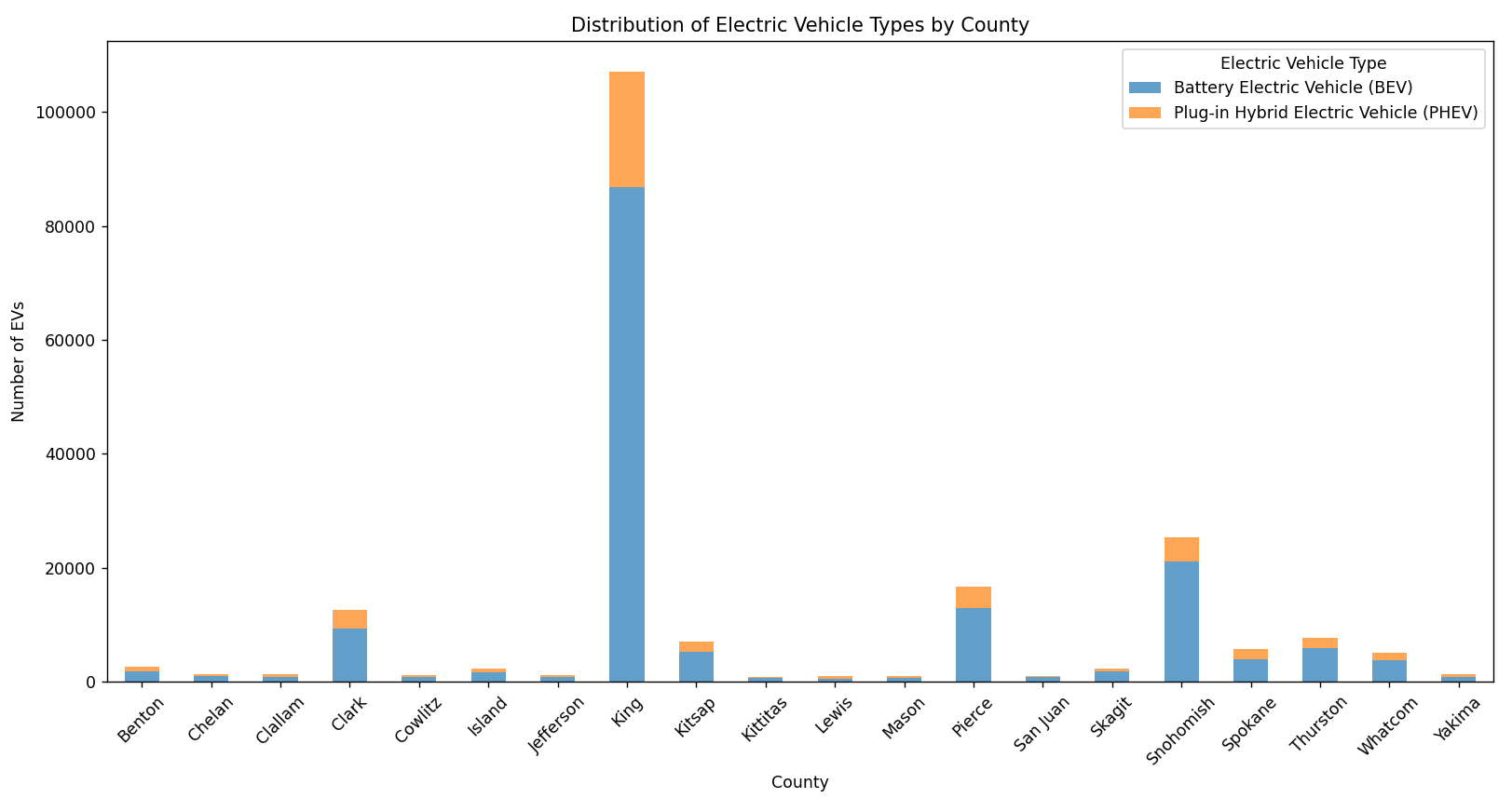


Figure 13: County vs plug-in / all electric EV

These visual comparisons emphasize urban-rural patterns in EV distribution, helping to pinpoint where infrastructure improvements or policy initiatives could accelerate adoption.

**Additional Analysis**

**Temporal Analysis**

For datasets covering multiple years, temporal analysis reveals trends in EV adoption rates and model popularity, using line charts to highlight these trends over time.

* **EV Adoption Trends by Model Year**:
  + The line chart of EV registrations over the years provides a visual of adoption rates, often showing a general increase over time. This trend reflects the growing popularity and acceptance of EVs, driven by improved technology, reduced costs, and environmental awareness. Sharp increases in specific years might coincide with the release of popular new models or government incentives.

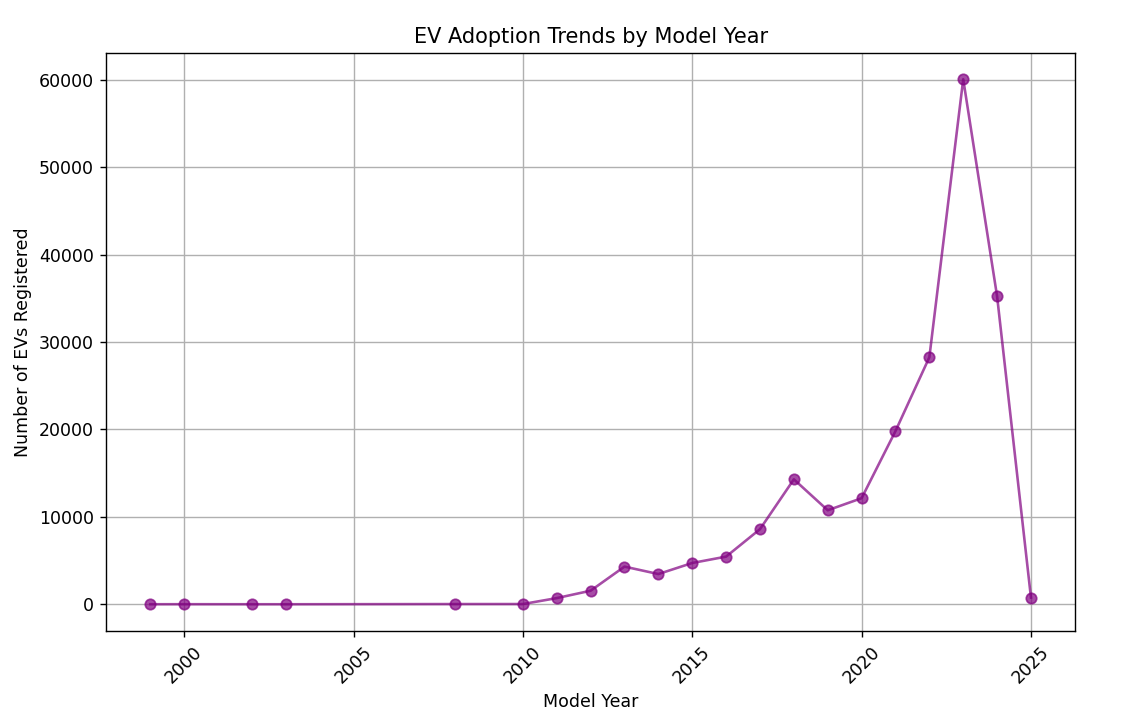


Figure 14: Trends by model year

* **Model Popularity Over Time**:
  + Tracking individual model popularity across years reveals shifts in consumer preferences. Certain models might see spikes in popularity when new versions or features are introduced. Models that consistently remain popular over the years can indicate strong brand loyalty or the appeal of specific features.

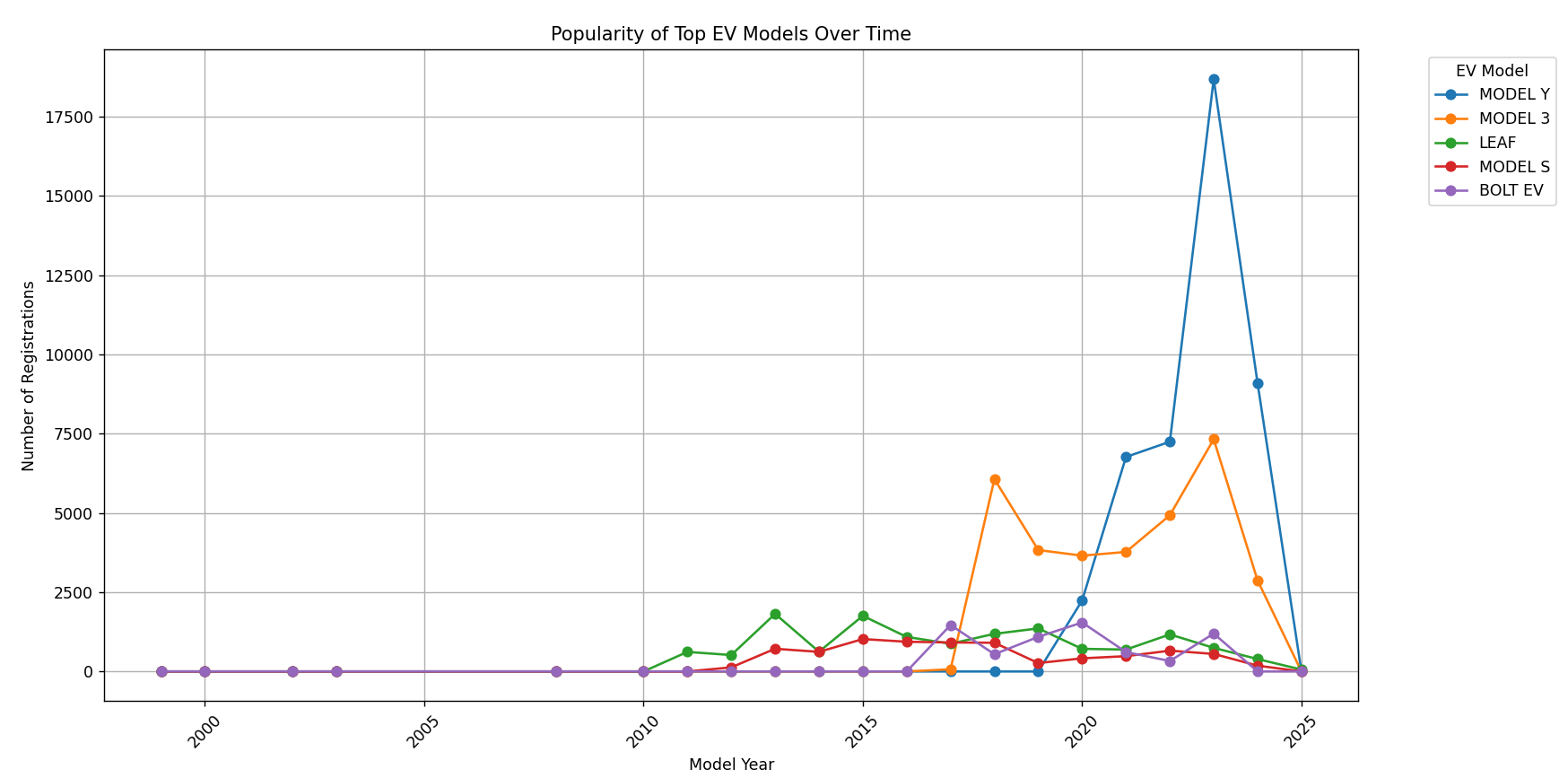


Figure 15: Top EV models over time

These temporal analyses highlight key moments in the industry’s development, allowing us to understand market evolution, predict future growth, and identify the impact of new model releases and policy changes on EV adoption.

# **Conclusion**

**Summary of Findings**

The analysis revealed that electric vehicle (EV) registrations are highly concentrated in urban counties, with a trend favoring all-electric models in areas with better infrastructure. Popular models dominated registrations, with certain brands consistently leading in adoption. Significant correlations were found between Electric Range and Base MSRP, indicating that higher ranges typically command a premium. Temporal trends showed a steady increase in EV adoption, reflecting both technological advancement and greater consumer acceptance.

**Implications**

These findings suggest that **policymakers** could focus on expanding charging infrastructure in less-served regions to support wider adoption. **Automotive manufacturers** may benefit from targeting mid-range EV options for price-sensitive regions, while **environmental planners** can leverage these insights to prioritize areas for green infrastructure development.

**Future Work**

Further analysis could include a more detailed temporal study to track seasonal trends or policy impacts, regional comparisons to understand adoption in different demographic areas, and predictive modeling to forecast EV growth under various scenarios. This would enhance strategic planning and decision-making across stakeholders.