Data Analytics Bootcamp by EDEX.org

Data Analysis of Electric Vehicles Registered in Washington State

Project 1

Lashway, Hannah
Montemayor, Andrew
Parker, Will
Rodriguez, Fatima
Rosensweig, Ben

DATA-PT-EAST-APRIL-041524-MTTH

Instructor: Booth, Alexander

June 17th, 2024

Table of contents

Introduction

Electric Vehicle is a vehicle that uses one or more electric motors for propulsion. It can be powered by a collector system, with electricity from extravehicular sources, or it can be powered autonomously by a battery (sometimes charged by solar panels, or by converting fuel to electricity using fuel cells or a generator) EVs include, but are not limited to, road and rail vehicles, surface and underwater vessels, electric aircraft, and electric spacecraft. (Wikipedia)

Electric vehicles have been around for a while, the very first one invented was in the late 1820. in the early 20th century, the high cost, low top speed and short-range of battery electric vehicles, compared to internal combustion engine vehicles, led to a worldwide decline in their use as private motor vehicles. For this reason, electric vehicles didn't seem like a good option compared to the internal combustion engine vehicles.

In the 21st century Electric vehicles starting to be a more attractive option due to growing concern about the environment and all the problems associated with combustion. It wasn't until 2010 the electric vehicles started to be more popular and sales for those increased considerably and have been the trend since then.

We were inspired by the cultural relevance and growing popularity of electric vehicles to select our dataset. As a group, we understand on a surface level the many appeals of electric and hybrid vehicles such as the tax incentives and their cost effectiveness. Through this project we aim to gain a deeper understanding into why consumers are choosing electric vehicles.

Data Set

The data set was retrieved on May 28th, 2024, from Kaggle.com, the name was: Electric Vehicle Data. The Data set was formatted as a csv file. It provides a comprehensive snapshot of the current landscape of Battery Electric Vehicles (BEVs) and Plug-in Hybrid Electric Vehicles (PHEVs) registered through the Washington State Department of Licensing (DOL) starting in 1997 until May 21st, 2024. This data base contains 181,458 rows and the columns are listed as follows:

1. VIN (1-10)- Vehicle Identification Number of the vehicle mentioned in the dataset.
2. County- Name from where the data is gathered.
3. City- Cities Name from where the data is gathered.
4. State- State Name from where the data is gathered.
5. Postal Code- The postal code from where the data is Present.
6. Model Year- Manufacturing year of the model mentioned in the data set.
7. Make- Manufacturer of the vehicle.
8. Model- Model Name of the mentioned vehicle.
9. Electric Vehicle Type- Type of the vehicle present in the dataset.
10. Clean Alternative Fuel Vehicle (CAFV) Eligibility- Clean Alternative for the data present in this dataset.
11. Electric Range- The range the car provides.
12. Base MSRP- Base cost without any accessories.
13. Legislative District- Legislative District it falls under.
14. DOL Vehicle ID- Department of Licensing issues Vehicle ID.
15. Vehicle Location- Vehicle coordinates.

16. Electric Utility- Electricity provider.

17. 2020 Census Tract- Census area or district defined for the purpose of taking a census.

By harnessing a wealth of information including geographic data such as zip codes, consumer preferences regarding model popularity, and technical specifications such as battery range, our objective is to delve deep into the motivations behind consumers' choices to embrace electric vehicles. Through meticulous analysis, we aim to uncover intricate patterns and trends that shed light on why certain regions exhibit higher adoption rates, what factors drive individuals to opt for specific electric vehicle models, and how these preferences evolve over time.

Data Cleaning

After selecting the data set, we proceed to download the csv file and use jupyter notebooks to proceed with the data cleaning. The first step was to import the csv file and explore the shape of the data looking for null values and select and evaluate the columns that we will keep for this analysis.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 181458 entries, 0 to 181457
Data columns (total 17 columns):
 #   Column                                             Non-Null Count   Dtype  
---  ------                                             --------------   -----  
 0   VIN (1-10)                                         181458 non-null  object 
 1   County                                             181455 non-null  object 
 2   City                                               181455 non-null  object 
 3   State                                              181458 non-null  object 
 4   Postal Code                                        181455 non-null  float64
 5   Model Year                                         181458 non-null  int64  
 6   Make                                               181458 non-null  object 
 7   Model                                              181458 non-null  object 
 8   Electric Vehicle Type                              181458 non-null  object 
 9   Clean Alternative Fuel Vehicle (CAFV) Eligibility  181458 non-null  object 
 10  Electric Range                                     181458 non-null  int64  
 11  Base MSRP                                          181458 non-null  int64  
 12  Legislative District                               181060 non-null  float64
 13  DOL Vehicle ID                                     181458 non-null  int64  
 14  Vehicle Location                                   181450 non-null  object 
 15  Electric Utility                                   181455 non-null  object 
 16  2020 Census Tract                                  181455 non-null  float64
dtypes: float64(3), int64(4), object(10)
memory usage: 23.5+ MB
None
```

Since we anticipated to use vehicle location in our visualizations, we proceed to drop the rows that were missing this value using the dropna() function.

```
# Drop rows with missing location data
df = df.dropna(subset=["Vehicle Location"]).reset_index(drop=True)
```

We found that there were registrations for other states other than Washington, then we applied a mask to keep values that matched Washington as state.

```
# Keep only the vehicles in Washington
df = df[df["State"] == "WA"]
```

The values for the column Vehicle location were formatted as follows: POINT (-122.374105 47.54468). We needed to first strip the string around the latitude and longitude and separate these two values into two different columns for later use. To remove the ¨POINT¨ characters we used the strip() function and to separate the latitude and longitude we used the split() function.

```
# Remove the "POINT(" and ")" from the "Vehicle Location" column
df["Vehicle Location"] = df["Vehicle Location"].str.strip("POINT( )")
```

```
# Split the "Vehicle Location" column into two separate columns: "Longitude" and "Latitude"
df["Longitude"] = df["Vehicle Location"].str.split(" ", expand=True)[0]
df["Latitude"] = df["Vehicle Location"].str.split(" ", expand=True)[1]
```

The new two columns latitude and longitude had string data type, we needed to change it to a float to be able to use it, the same as the zip code column.

```
# Convert the "Longitude" and "Latitude" columns to floats and "Postal Code" to an integer
df["Longitude"] = df.loc[:,"Longitude"].astype(float)
df["Latitude"] = df.loc[:,"Latitude"].astype(float)
df["Postal Code"] = df.loc[:,"Postal Code"].astype(int)
```

We also checked for duplicate values and dropped them.

```
# Drop duplicate rows based on the "VIN (1-10)" column
df.drop_duplicates(subset="VIN (1-10)", keep="first")

df = df.set_index("VIN (1-10)")
```

We created a new Data Frame with only the columns that were going to be relevant to our analysis.

```
<class 'pandas.core.frame.DataFrame'>
Index: 181055 entries, 0 to 181449
Data columns (total 12 columns):
 #   Column                 Non-Null Count   Dtype
---  ------                 --------------   -----
 0   VIN (1-10)             181055 non-null  object
 1   County                 181055 non-null  object
 2   City                   181055 non-null  object
 3   State                  181055 non-null  object
 4   Postal Code            181055 non-null  int64
 5   Model Year             181055 non-null  int64
 6   Make                   181055 non-null  object
 7   Model                  181055 non-null  object
 8   Electric Vehicle Type  181055 non-null  object
 9   Electric Range         181055 non-null  int64
 10  Latitude               181055 non-null  float64
 11  Longitude              181055 non-null  float64
dtypes: float64(2), int64(3), object(7)
memory usage: 18.0+ MB
None
```

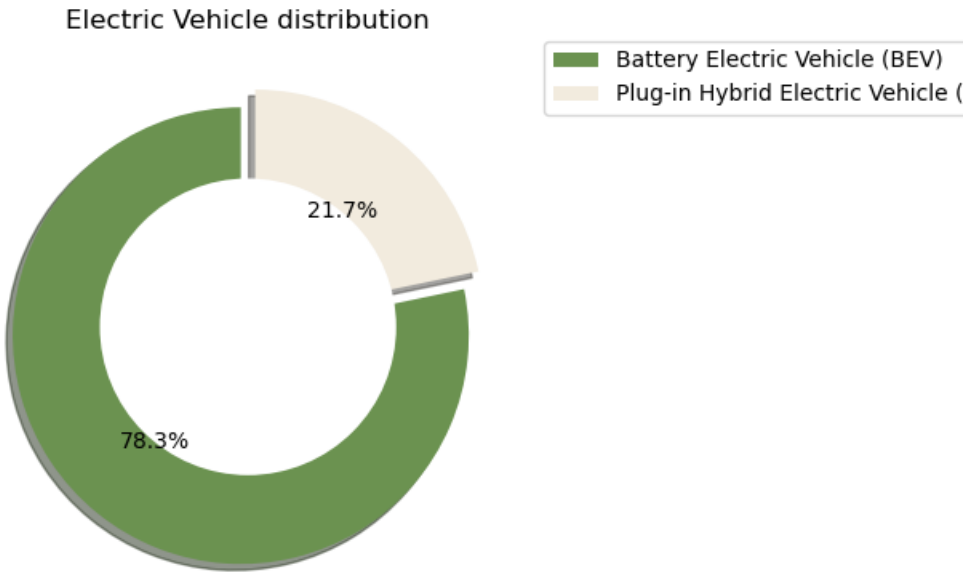
As an extra step to create our linear regressions we dropped the values with Range values equal to 0.

```
# Drop vehicles with electric range of 0
df = df[df["Electric Range"] > 0]
```

And finally the new data set was saved as a csv file using the to_csv() function.

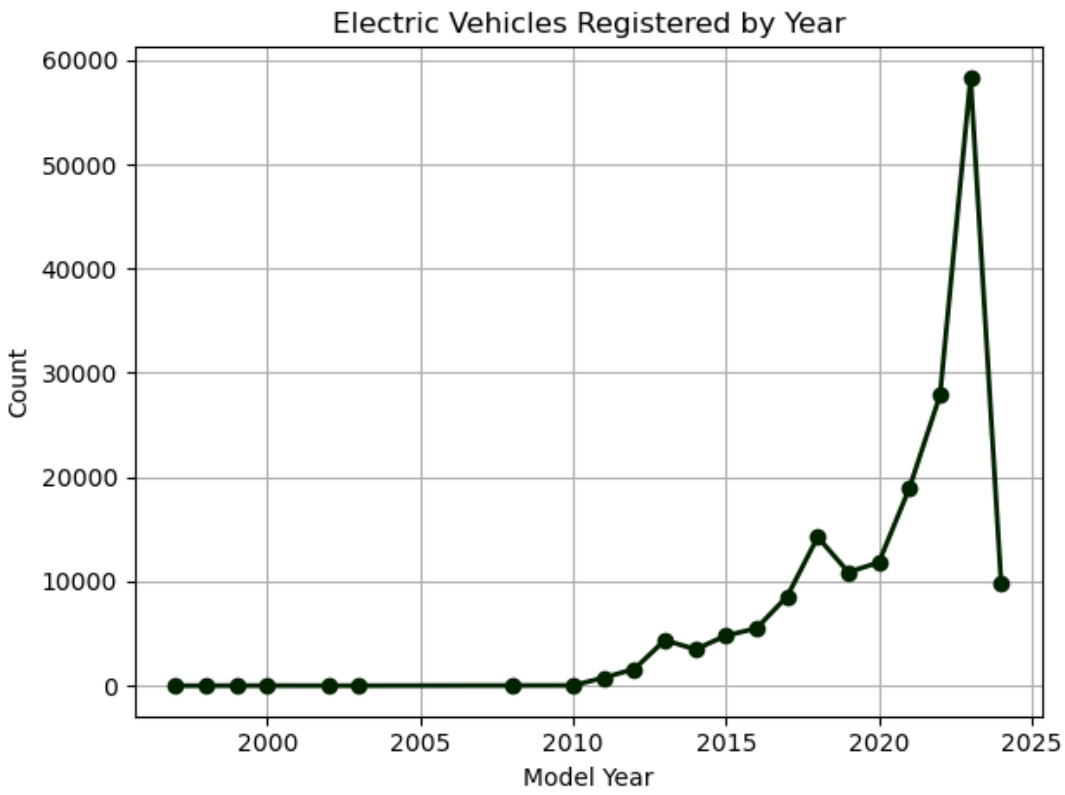
Research Question 1: What are the most and least popular car make and models for each year based of annual registrations?

After cleaning our data, we were left with 181,055 vehicle registrations. There are two main types of electric vehicles, Battery Electric Vehicle (BEV) and Plug-in Hybrid Vehicle (PHEV). The vehicle population was distributed as follows:



78.3% were Battery Electric Vehicles (BEV) and 21.7% Plug-in Hybrid Vehicle (PHEV). A majority of registered vehicles are Battery electric vehicles, having a significant difference against the plug-in hybrid.

As the years go by, EV vehicles have become more popular, with more affordable prices among other incentives and technological improvements. The next graph shows the Electric Vehicle registered by year.



Registration for Electric Vehicles started to increase in 2010, probably incentivize by The federal government that has been offering tax credits for electric vehicles (EVs) and plug-in hybrid electric vehicles (PHEVs) since 2008.

There was a particular increase in EV sales in the United States in 2018, driven by the market launch of the standard version of the Tesla Model 3. In 2019 the sales decrease due to limited supply, loss of popularity of other models different than Tesla Model 3, No New High-Volume EVs Were Introduced among others reasons. As of June 2024, the credit for new EVs is up to $7,500, while the credit for used EVs is up to $4,000.

It's important to note that in this graph the year 2024 looks like a decrease in registrations, nevertheless, this is data is only a reflection of the first few months of the year. We would need to wait for more updated data in 2025 to account for all registrations in 2024.

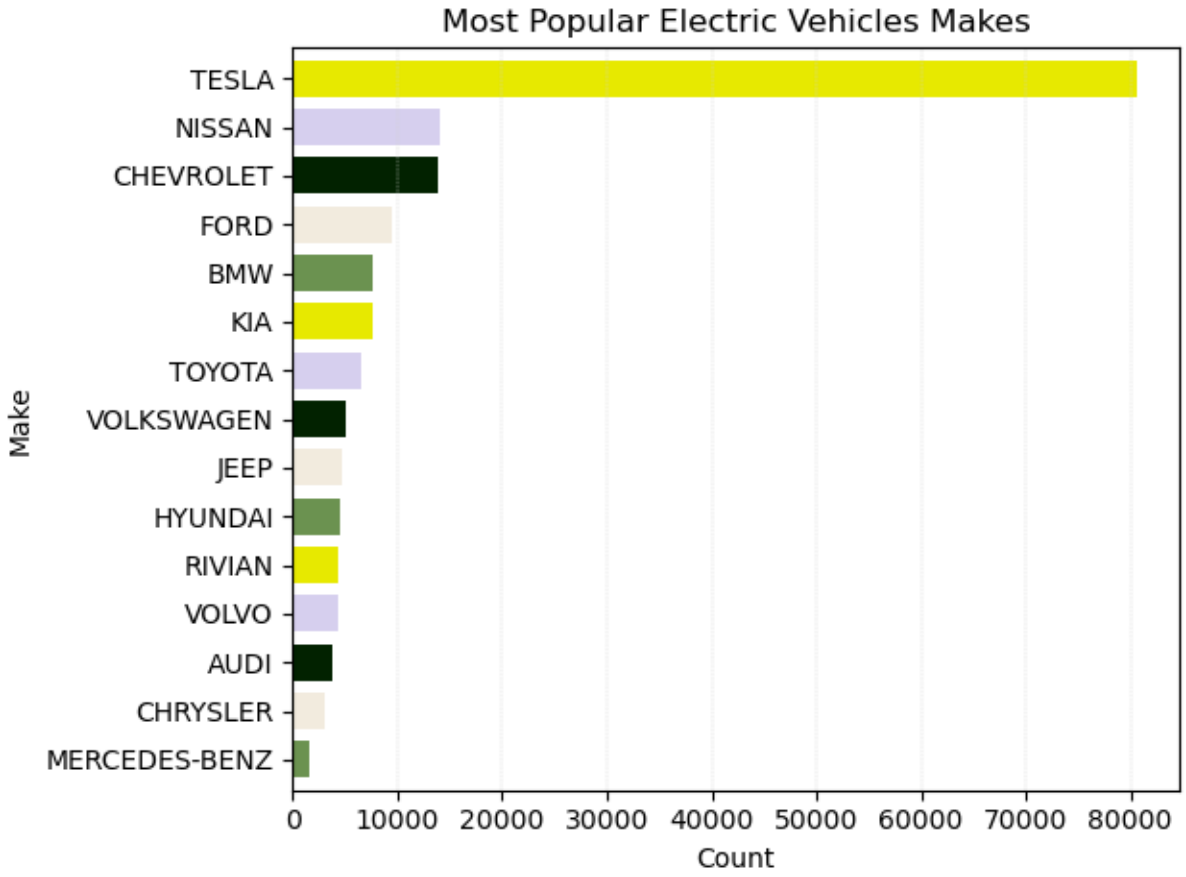
Now that we understand the distribution and trends for the EV vehicles, we can move to answer our questions.

The number of registrations by make goes like this:

Make	
TESLA	80627
NISSAN	14024
CHEVROLET	13838
FORD	9503
BMW	7666
KIA	7632
TOYOTA	6485
VOLKSWAGEN	5153
JEEP	4678
HYUNDAI	4552
RIVIAN	4419
VOLVO	4278
AUDI	3729
CHRYSLER	3039
MERCEDES-BENZ	1646
PORSCHE	1157
MITSUBISHI	979
MINI	925
POLESTAR	894
SUBARU	837
HONDA	834
FIAT	783
DODGE	607
MAZDA	506
CADILLAC	432
LEXUS	398
SMART	269
LINCOLN	269
LUCID	238
JAGUAR	236
GENESIS	189
FISKER	111
LAND ROVER	57
ALFA ROMEO	42
AZURE DYNAMICS	8
TH!NK	5
GMC	3
BENTLEY	3
WHEEGO ELECTRIC CARS	3
ROLLS ROYCE	1

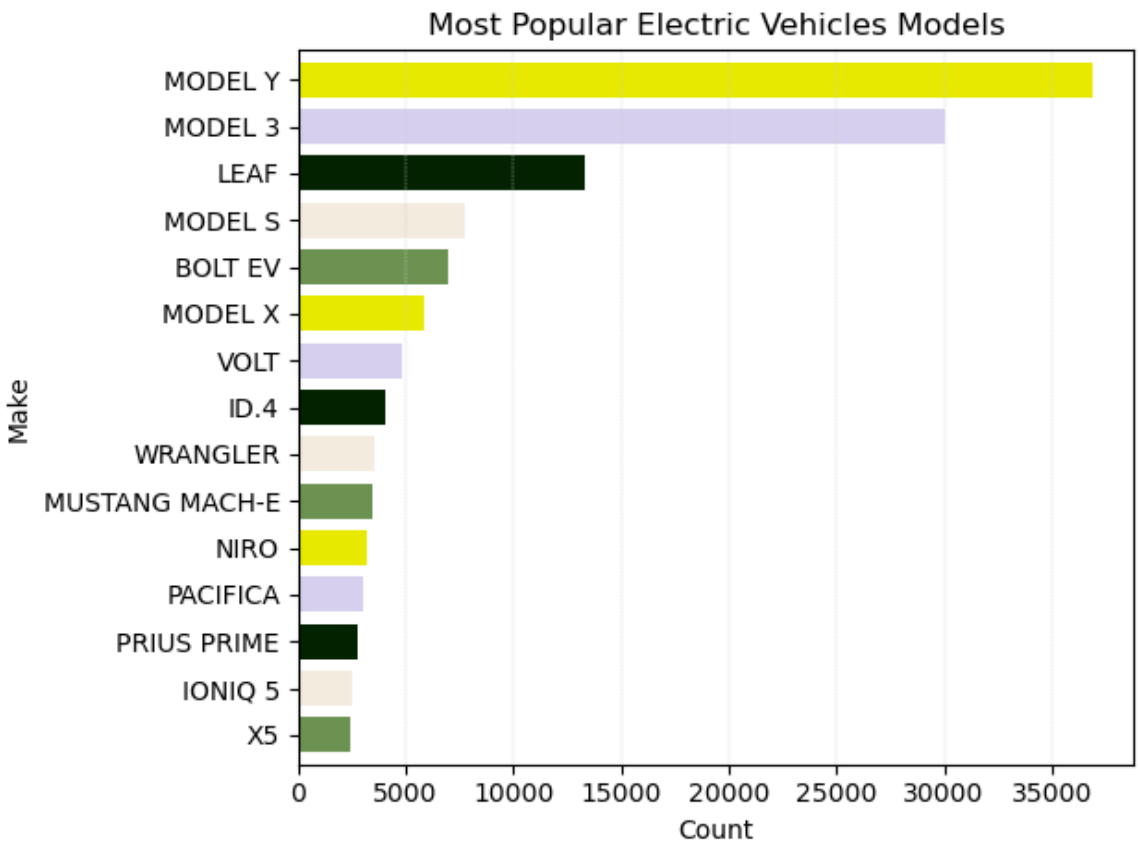
dtype: int64

Leaderboard of the top 15 makes by the number of registrations in the State of Washington.



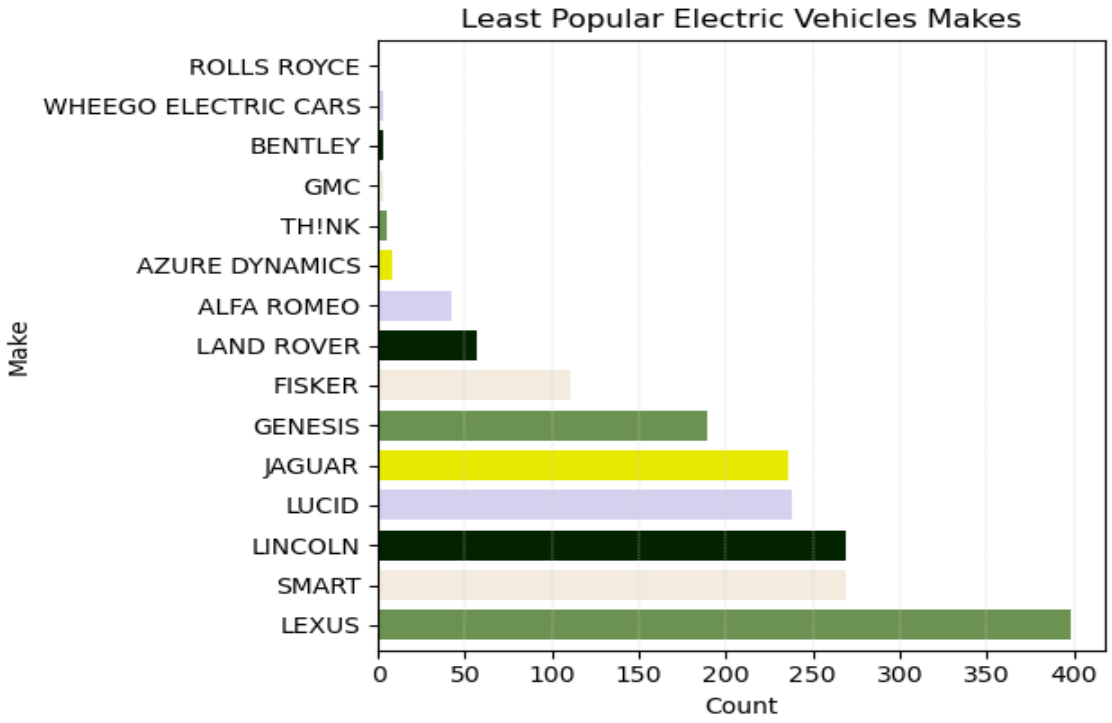
The most popular make is Tesla with 80,627 registrations, a lot more than its closest competitor, Nissan with 14,024 and Chevrolet with 13,838. Tesla started as Tesla motors in 2003 focusing on electric mobility, its vehicles are known for their cutting-edge technology, long range and high performance. Tesla also has a very robust network of fast charging stations, making long distance travel convenient and increasing their appeal among other makes. For these reasons we infer that people are more inclined to buy Tesla overall.

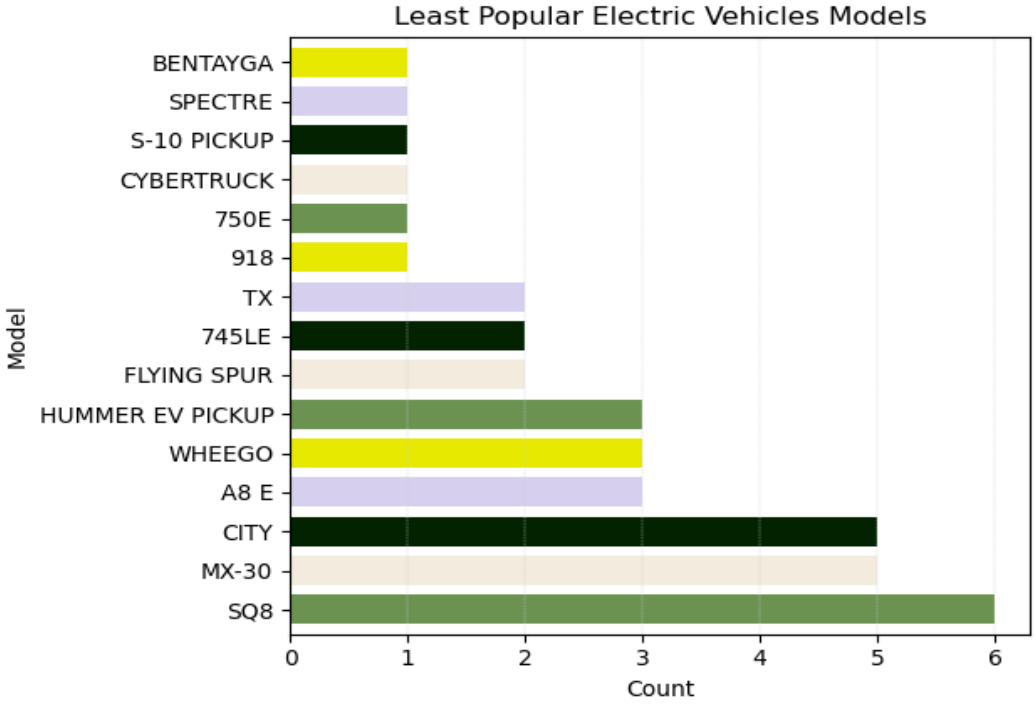
When it comes to the most popular models, we observed the top 15 models registered as follows:



Tesla being the most popular make, it was expected that its models would be in the most preferred models as well. The Tesla model Y and 3 are particularly popular due to their practicality, affordability and stylish design. Tesla model Y is the vehicle with more registrations, this is particularly interesting since is also the most recent model for Tesla, its production started in January 2020 and rapidly became popular in just few years, surpassing the Tesla model 3 that was the number before this. The Nissan Leaf was first introduced in 2010, new models have a range of 212 miles and it is very affordable in the EV market. The Chevrolet Bolt EV was commercialized from 2017-2023, there is not a 2024 Bolt EV. Although, this vehicle was always among the cheapest EVs sales were affected by a battery recall in 2021, followed by a MRSP drop in 2023 that was their best-selling year ever.

Looking at the least popular makes and models for our database, the leaderboard looks like this.

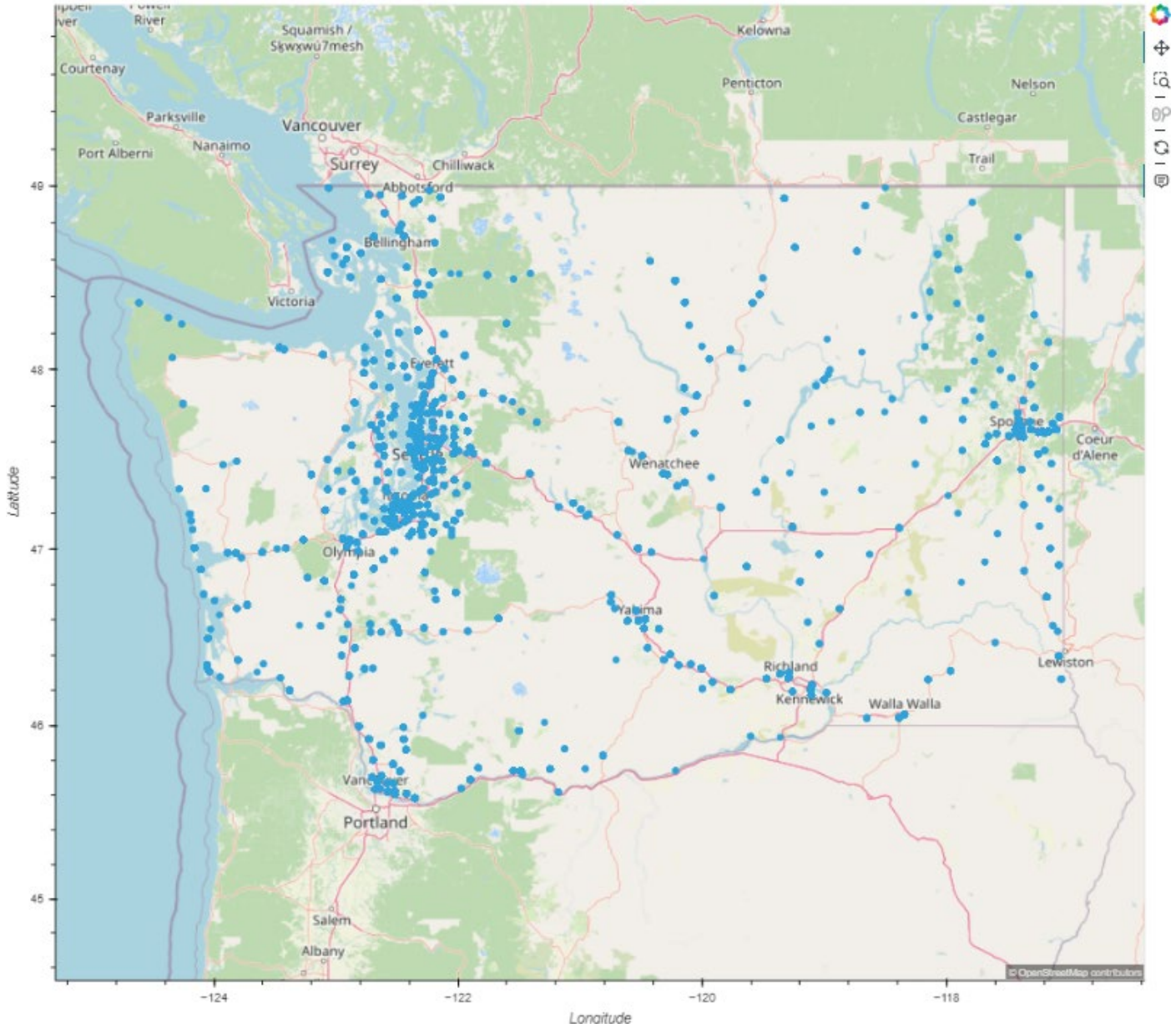


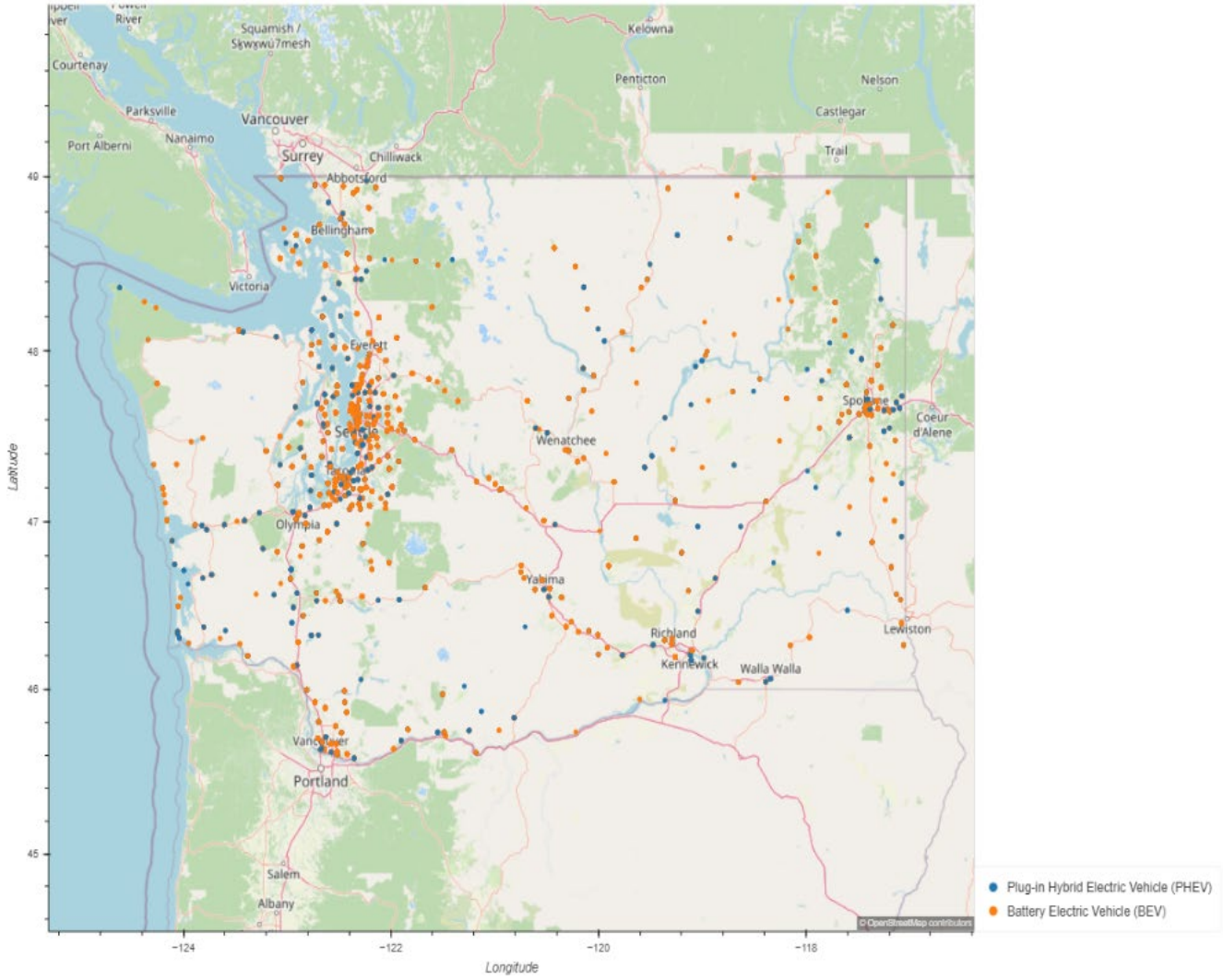


Rolls Royce and Bentley are makes that cater to a very specific population, usually affluent people who value exclusivity. Their models the Bentayga and the Spectre only had one registration in our database, being the least popular makes and models in the distribution. Their models are also very expensive, where just very few individuals will be able to afford these particular EVs. Wheego Electric cars retired the market of EVs in 2013, and turned its attention to developing tolls and systems for autonomous vehicles. The cybertruck is the newest Tesla model that was first introduced as a concept in 2019 but production didn't start until 2021. The first delivered was in November 2023, as the year 2024 progress we start to see more of these trucks on the streets, so even when at the time this data base was updated, it only showed one registration, more registrations are expected in the following years.

Research Question 2: Is there a relationship between geographic location and type of Electric Vehicle (i.e. BEV vs. Hybrid EV)?

To help provide results, for the second hypothesis, we created two visualizations. These visualizations consisted of two geographical maps of the state of Washington. The first one detailing the geolocation of all electric vehicle owner registrations in Washington; and the second one detailing those geolocations broken down by Hybrid and Full Battery electric vehicles.





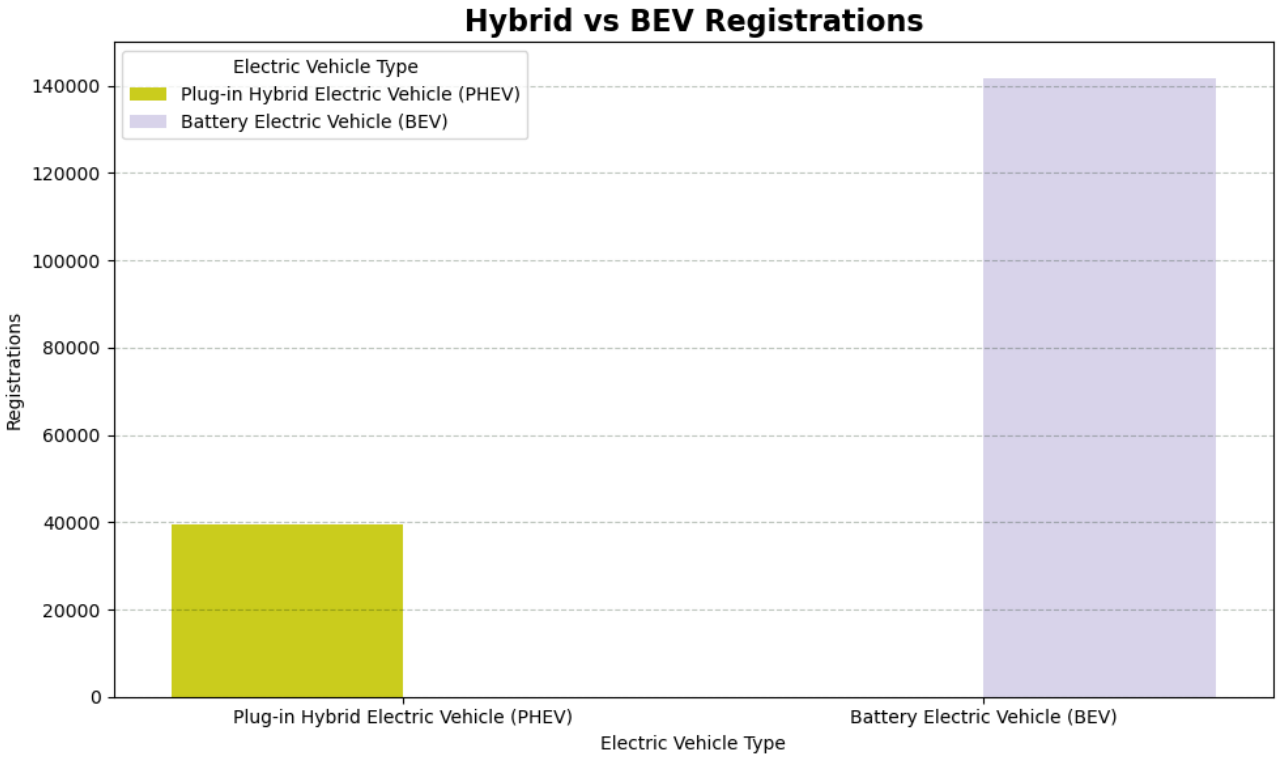
Though we found our dataset to lack pertinent information to provide conclusive results regarding the second hypothesis, it can be observed that there seems to be a “relatively” even distribution of both vehicle types, Hybrids and Full Battery electric vehicles, in respective geolocational “clusters” throughout the entire state of Washington.

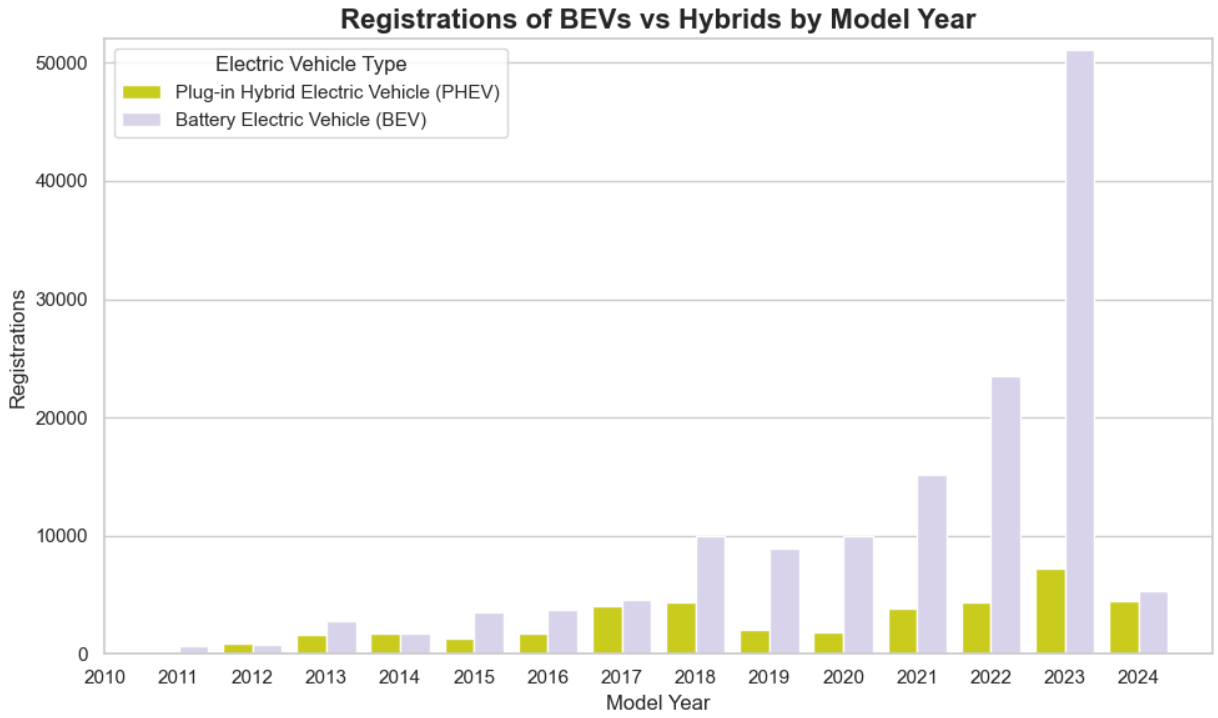
To address the lack of evidence/information to adequately test our hypothesis, we researched additional sources of relevant metadata that we could analyze and apply to our project in the future. We found that the integration of API sources, such as “Open Charge Map API”, that

address the geolocations of charging stations could help us test our hypothesis by comparing vehicle registration locations to their proximity of the nearest charging station. The integration of vehicle value data, such as Vehicle Market Value API, could help us test the hypothesis by comparing vehicle registration locations to specific vehicle type/model MSRP. Additionally, we speculate that the integration of census API sources, such as Economic Census API, Poverty Statistics API, and Quarterly Workforce Indicators API could help test our hypothesis by comparing vehicle registration locations to economy-based data such as household income and county workforce financial statistics.

Research Question 3: is there a correlation between model years and annual car registrations, and what are the comparisons between BEVs and PHEVs. Is there a correlation between the increase in battery range and annual car registrations?

We used several types of plots to visualize registrations by Model Year. Using a count plot and a donut chart, we were able to see that for BEVs there is a strong correlation between model year and registrations. Of all the 141,705 registrations in fiscal year 2023, starting from model year 1997 to 2023, 92% fall within the model years 2015-2023. Model Years 2022 and 2023 take up over 52% of registrations alone. Starting with Model Year 1999, it almost seems to be an exponential increase. There is a preference for newer models. Why is that?

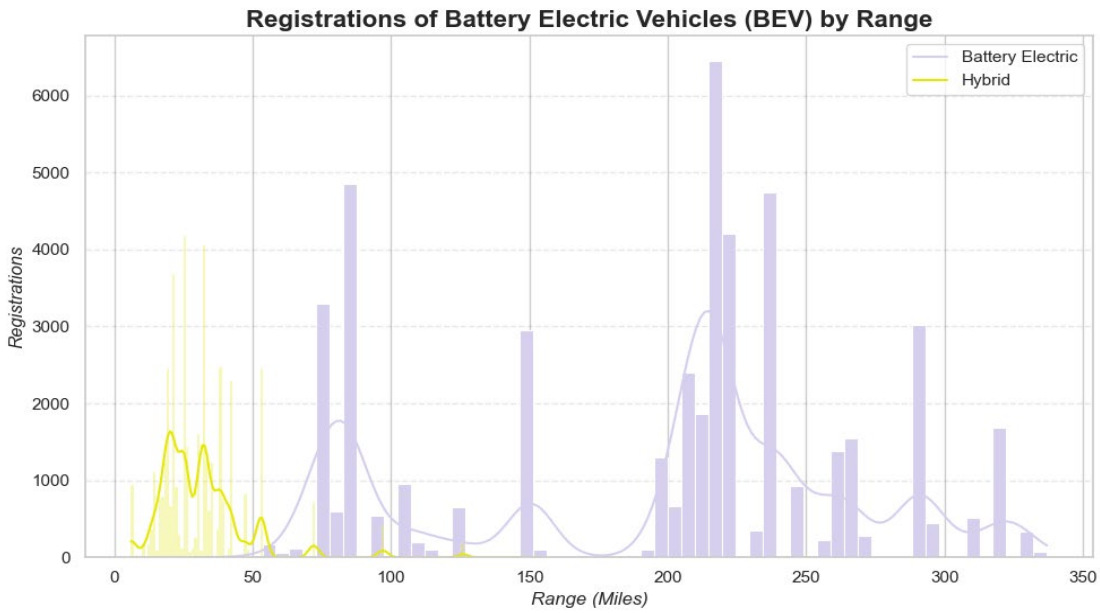


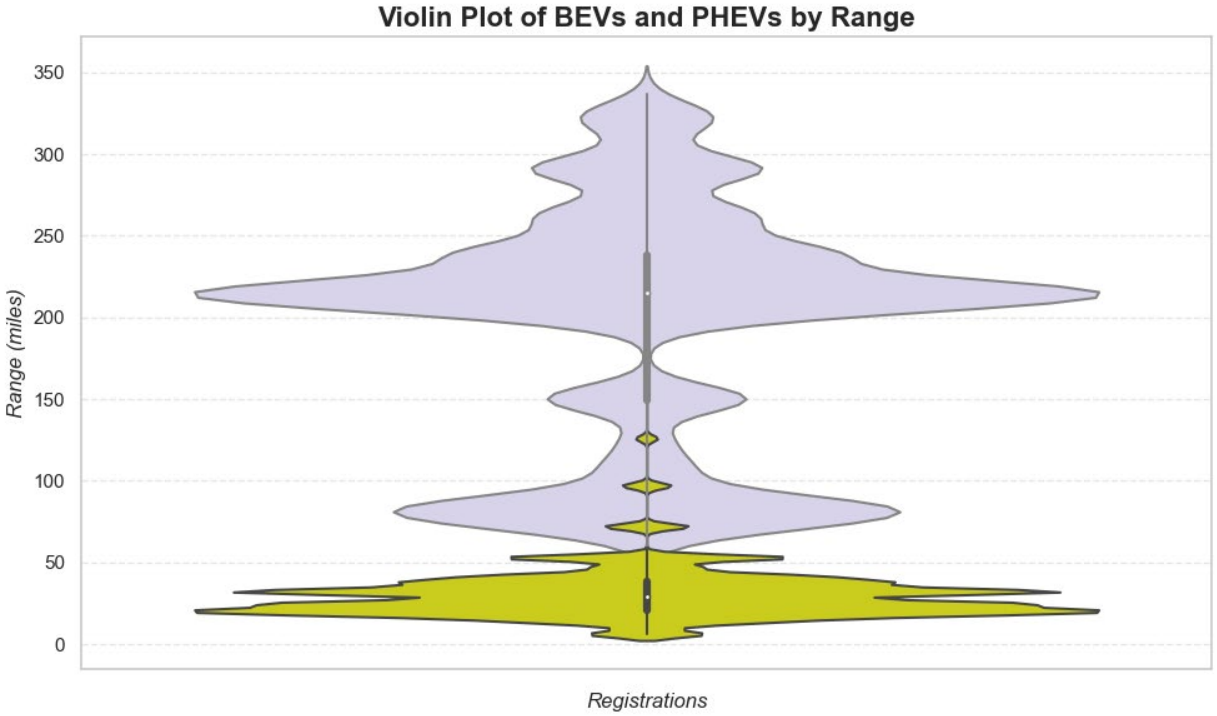


For hybrid registrations in 2023, although there doesn't seem to be as much of a preference for newer models, there does seem to be some. There is even some bimodality. Further work would be needed to study exactly why there are spikes for 2017, 2018, and 2023 model years.

At this point we realized that many BEVs had a range of 0 and had to clean the dataset even further. We went from around 181,000 rows to 86,500 - dropping approximately 94,500 rows, or over 52% of the data, all of which were BEVs from 2013-2024.

We were still able to get valuable insights from this new dataset that focused on range. After plotting a few histogram violin, and scatter plots, we were able to see that hybrids have much lower range than BEVs. Most hybrids fell between a range of 20-30 miles, while the majority of BEVs had ranges over 200 miles. For BEVs there is a clear correlation of range increase with newer models, but a very weak correlation for newer hybrids.

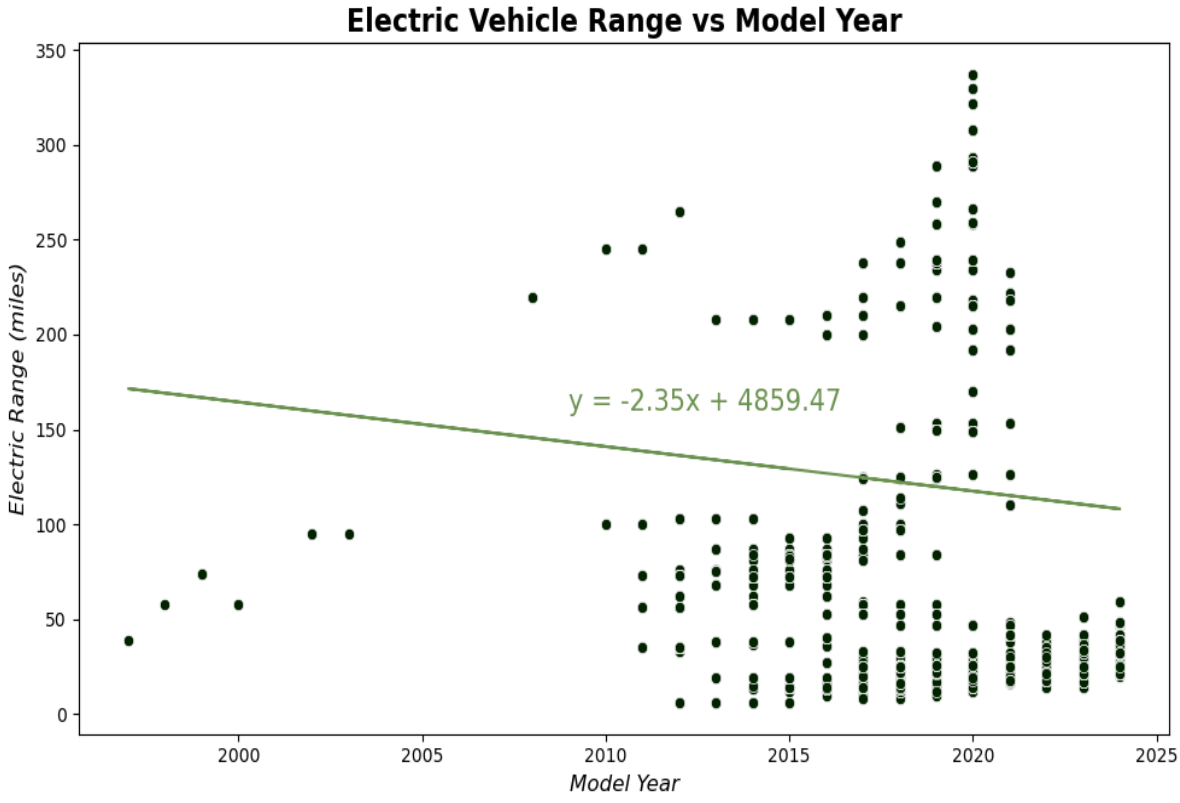


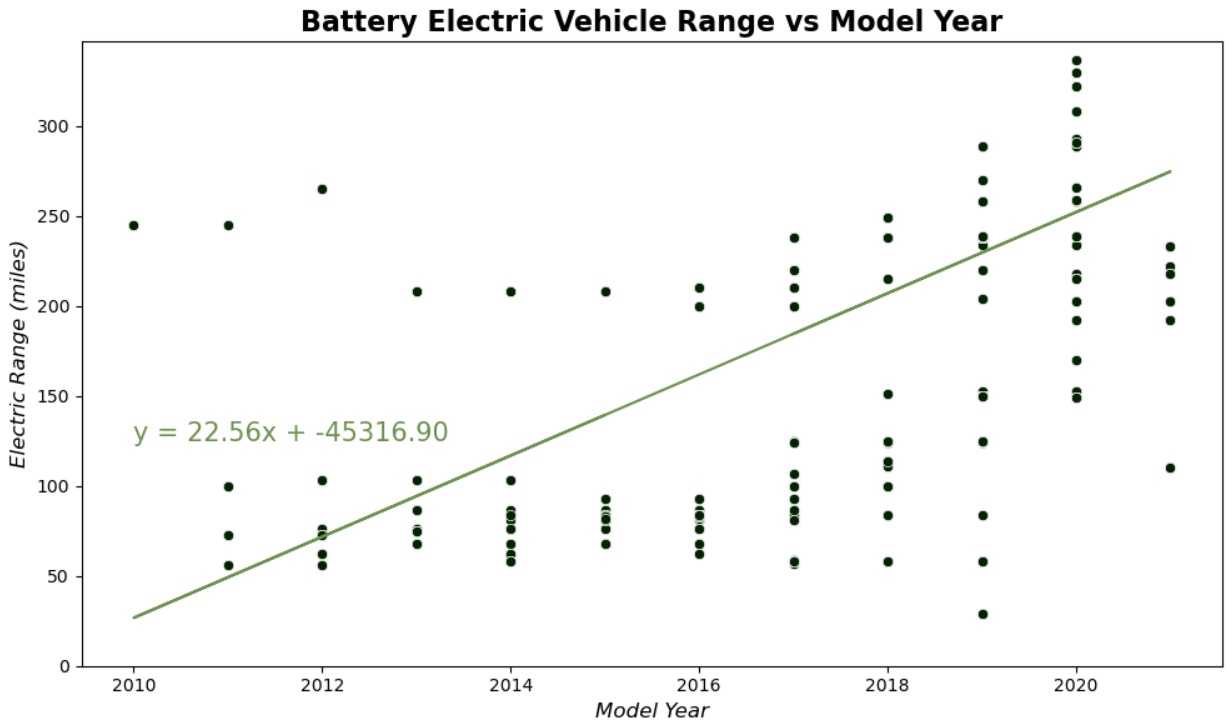


We can safely conclude that people prefer BEVs over hybrids, as the ranges for newer model BEV have continued to grow and have far outpaced the ranges of newer model hybrids. In the future, it will be interesting to see whether hybrids are either completely phased out, or if manufacturers increase hybrid range.

Linear Modelling and/or Statistical Modelling and T-test

To test the third hypothesis and the overall data story, regressions were run based on model year and electric range. With the hybrid and full electric vehicle data combined, there was no correlation found. After narrowing the data to only full electric vehicles and removing outliers from before 2010 in the electric vehicle data, an r-squared value of 0.51 was obtained. This gave the confidence to test for statistical significance in the form of a one sample t-test. The yearly mean of full electric vehicle ranges was tested against the overall mean range of the dataset. A p-value of 0.02 was obtained.





There are limitations to what these values can communicate because of the original dataset. Future work should include registration data from multiple years to measure year-over-year changes and from a wider geographic area to increase the population size. It would also be interesting to compare this data to non-electric vehicles to see if electric vehicles, in general, are overtaking petroleum powered vehicles in popularity.

Call to Action

With sufficient data, we can develop predictive models to forecast future trends in electric vehicle adoption based on socio-economic indicators. This can be valuable for urban planners, policymakers, and businesses looking to anticipate demand and tailor their strategies accordingly.

The insights derived from analyzing the correlation between location and electric vehicle preferences can inform policy decisions aimed at promoting sustainable transportation solutions. For instance, policymakers may use this information to target subsidies or infrastructure investments in areas where electric vehicle adoption is lagging behind due to socio-economic barriers.

Bias and Limitations

No Population Data

We could sort our data by zip code, county or city but did not have information such as the population or average income for those cities, counties and zip codes. This made it difficult to understand if people were selecting EV's based off of their income or location.

Incomplete Columns

Our MSRP column had some values but for the majority of our rows, we only had '0' values as a result, we had to drop this column during our data cleaning and revise one of our thesis questions. Without this data we were unable to address or incorporate any financial questions.

Limited BEV Data

During data cleaning, we removed all rows that had a battery range of '0' to have a more accurate reflection of the actual battery ranges. Our cleaned dataset did not include any ranges for BEV's from 2022 to 2024. As a result, our data for BEV's is very limited. Challenging to gauge the actual improvement in battery range for BEVs Limited data may explain the bias that Hybrids hold in our linear regression resulting in a negative graph.

Incomplete Registration

Our Registration Year column was only partially completed As a result we had to remove rows that didn't include a registration year. We had registration data for 2023 for all vehicles but not for 2022 or 2024. This meant we couldn't compare year over year change

Future Work

Our next steps would be to incorporate APIs with socio-economic data Incorporating APIs (Application Programming Interfaces) with socio-economic data can significantly enhance our understanding of the relationship between location and the type of electric vehicles (EVs) people are selecting. Here's how:

- **Accessing Socio-Economic Data**: APIs provide a bridge to access vast amounts of socio-economic data, including demographics, income levels, education, employment rates, and more. By integrating these APIs into our analysis, we gain insights into the social and economic characteristics of different geographical areas.
- **Correlation Analysis**: Once we have access to socio-economic data, we can conduct correlation analysis to determine the relationship between various socio-economic factors and the choice of electric vehicles. For example, we can explore whether areas with higher income levels tend to prefer luxury electric vehicles, or if regions with better access to public transportation opt for more affordable electric models as secondary vehicles.
- **Geospatial Mapping**: By combining socio-economic data with geographic information, such as zip codes or GPS coordinates, we can create geospatial maps to visualize patterns and trends. These maps can highlight clusters of specific types of electric vehicles in certain areas and identify potential influencing factors, such as proximity to charging stations, tax incentives, or local environmental policies.

Overall, by leveraging APIs to integrate socio-economic data into our analysis, we can gain a more nuanced understanding of the complex factors influencing electric vehicle adoption patterns,

ultimately facilitating more effective decision-making and interventions to accelerate the transition to sustainable transportation.

Works Cited

Kaggle Dataset:

Jainaru. "Electric Vehicle Population." *Kaggle*, URL: https://www.kaggle.com/datasets/jainaru/electric-vehicle-population.

Stack Overflow:

Stack Overflow. URL: https://stackoverflow.com/.

Pandas Documentation:

"Pandas Documentation." *Pandas*, URL: https://pandas.pydata.org/pandas-docs/stable/index.html.

EV Statistic Sources:

"How Does the Electric Car Tax Credit Work?" *U.S. News & World Report*, URL: https://cars.usnews.com/cars-trucks/advice/how-does-the-electric-car-tax-credit-work#:~:text=Since%202008%2C%20the%20federal%20government,hybrid%20electric%20vehicle%20(PHEV).

"2019 US EV Sales Decreased: An Estimated 7% to 9.6% Reasons Why." *EV Adoption*, URL: https://evadoption.com/2019-us-ev-sales-decreased-an-estimated-7-to-9-6-reasons-why/.

"Government Incentives for Plug-in Electric Vehicles." *Wikipedia*, URL: https://en.wikipedia.org/wiki/Government_incentives_for_plug-in_electric_vehicles.