Project 1:

Deadline: November 10, 2024, 11:59pm Eastern Time

Group Members:

Dorothy (Gracie) Rehberg - drehberg1@student.gsu.edu
Lilly Parham - lparham2@student.gsu.edu
Pamela Alvarado-Zarate - palvaradozarate1@student.gsu.edu

The business problem centers around the EPA's goal of significantly increasing electric vehicle (EV) adoption in the United States to align with its Net-Zero Emissions targets for 2030-2032. Despite a growing push toward sustainability, only 6.9% of new cars sold in 2023 were electric, which is well below the EPA's target range of 35% to 56% by the early 2030s. The EPA suspects that a major barrier to higher EV adoption rates is the insufficient availability of charging stations, which discourages consumer demand for new electric vehicles. To address this, our solution involved developing a predictive model using demographic data to identify regions where additional charging infrastructure would most effectively drive EV purchases. By analyzing factors such as income levels, population density, and existing EV adoption patterns, this model will guide the EPA in strategically subsidizing new charging stations. The goal is to maximize the impact of infrastructure investments, thereby accelerating EV adoption, reducing emissions, and supporting the EPA's environmental targets.

The central question is as follows: within a given geography, what will be the marginal increase in the number of electric vehicles registered in that geography, under the assumption that the EPA adds one EV charging station to that geography by subsidizing a station's construction? The goal is to identify the geographies that will maximize the marginal increase in electric vehicle registrations. In order to predict the marginal increase in registrations, we must build a predictive model that predicts, as a target variable, the number of EV registrations, with predictive features including, but not limited to, the number (absolute and per capita) of fuel stations in that geography. Other predictive features that inform the model are the demographic features of the population of that geography, including income, population density, median age, and % completion of college. The premise is that geographies with certain kinds of demographics will have a higher marginal increase in registrations in response to marginal increase in fueling stations, even when the number of existing fuel stations and registrations are equal.

At this stage of the project, we have chosen U.S. states as the appropriate geography because the dataset of EV registrations from the Alternative Fuels Data Center (US DoE) reports the number of registrations by state and not a more specific geography. In the next stage of the project, we will use more fine-grained data to infer the number of EV registrations in a more specific geography, such as U.S. counties.

We built an analytic basetable (the ABT) consisting of the following features:

- A dataset of the number of EV registrations by state from the Alternative Fuels Data Center (the target feature)
- An exhaustive dataset of EV fuel stations in the United States from kaggle, aggregated to:
    - Number of accessible fuel stations in each state
    - Accessible fuel stations per capita in each state
- Demographic features:
    - State-level population data (by age and sex) from the U.S. Census (county-level population data is also available from the same source)
        - Population density (computed by dividing population feature by km^2 area of all the states)
        - Median age (computed by a cumulative sum on the population by age and sex by state table)
        - Sex ratio (by groupby on the population by age and sex by state table)

- Mean income by state from the Bureau of Economic Analysis (county-level mean income data is also available from the same source)
- College completion rate data from the Department of Agriculture

To ensure that all the datasets corresponded to the same set of states, the datasets that recorded data with state names encoded the state names as state abbreviations, and the datasets that recorded data with state FIPS codes encoded the state FIPS codes as state abbreviations.

At this stage of the project, it was not necessary to fill in missing data because all the state-level datasets were an exhaustive source of data on all 50 states. If, at the next stage of the project, data from smaller regions (such as counties) contains data with missing rows, we will handle the missing data of a column by replacing missing values with the median value of that column. It is recommended that we use the median for a robust measure of central tendency that is less sensitive to outliers compared to the mean. If, at the next stage of the project, we use a categorical data feature that includes missing values, we will replace the missing values with "Unknown." Assigning a placeholder like "Unknown" helps retain all rows in our dataset, allowing us to leverage as much information as possible. It is necessary that we fill in the missing values in our dataset because it is a crucial step in data preprocessing to ensure our analysis and models are accurate, reliable, and robust. Irrelevant or excessively incomplete data was and will be excluded ensuring that only high-quality data was used in subsequent analysis.

Other data preprocessing techniques used were feature selection and extraction. In the EV fuel stations dataset, we did encounter columns, like "RD Blends" or "Hydrogen Standards," that referred to alternative fuel stations (like renewable diesel or hydrogen) that are not relevant to this study. Other columns, with limited data, such as "Station Phone", "Funding Sources", or "Intersection Directions", provided detailed information about the stations but were also not relevant to the study's focus. Therefore, these features were removed to keep the analysis focused on electric fuel stations and to save space. The "fuel_station" dataset consists of features that describe the location, directory, accessibility, and additional information of each alternative fuel station, with the primary focus on electric vehicle stations for this study. Ultimately, the main feature extracted from the fuel station dataset was the raw number of fuel stations per state.

Feature extraction from the "population" dataset was required to make aggregations of median age, percentage of college completion (in decimal form), and the sex ratio of each state, to be used for further analysis. Additional features were created that described the statistical context of electric vehicle registrations or electric-fueled stations per person as "EV Registrations per capita" or "Fuel Stations per capita." (see figures 3-5).

Outliers in the dataset, specifically from "Fuel Stations per capita" and EV registrations per capita," will be included as part of the statistical analysis and visual representation. Figures 4 and 7 show the distributions of electric fuel stations per capita and electric vehicle registrations per capita. Both distributions skew right or positively. The most noticeable outliers, $x = >0.0005$ or $x = >0.30$, represent the state(s) with the highest number of fuel stations and electric vehicle registrations per capita. This is essential to the analysis and machine learning modeling to predict the target feature based on the change of electric fuel stations established in a state. This also infers that the represented state is inclining toward electricity as an alternative fuel source.

We constructed a new dataset, “States_ABT,” to consist of the selected and created features from the original dataset via joining features generated from the above-described dataset into a single table. The “States_ABT” is a table of 50 rows (data points), one for each state, with one target feature of the number of electric vehicle registrations, and several predictive variables about the geography and demographic features of the population in that geography.

On our States_ABT we built two regression models for comparison: (1) a random forest model to capture nonlinear relationships between predictive features and the target feature, and (2) a Ridge regularized linear regression model, which can mitigate the overfitting effect of large coefficients generated by an ordinary linear regression model. Both models used all the same predictive features predicting the target feature (number of EV registrations) using the remaining features as predictive variables. First we split the data into a train_data set and a test_data set, reserving 20% of the 50 rows for the test set. For each of the two types of models, we fit the best version of the model to the data by using hyperparameter tuning with cross validation to find the combination of hyperparameters that minimized the mean-square error of prediction values on the target variable in the test_data set.

Then, to use the model to solve the business question at hand, we iterated through the 50 states, and in each state we created a hypothetical data point corresponding to that same state with all the same predictive feature values, except with the “Fuel Station Count” feature incremented upward by +1. Then, we used the model to predict the new “EV Registration Count” value of this new hypothetical state, and subtracted out the actual “EV Registration Count” value for that corresponding state. The numerical result of this calculation is the expected marginal increase in electric vehicle registrations if a new EV fuel station is built.

The anticipated result of this process is that both models predict every state would respond to an additional EV fuel station with a modest positive number of additional vehicle registrations. However, the actual result of this process is that the random forest model predicted 0 change in EV registrations in most states, negative change in EV registrations in some states, and positive change in EV registrations in some states. This makes sense because the random forest model ultimately creates strict boundaries in the multidimensional feature-space, and an increment of +1 to one feature is not likely to cause a datapoint to cross a boundary in feature space. The actual result of this process for the ridge model is that many states had a positive response in EV registrations to an increment in EV fuel stations, but also many states had a negative response.

But, one positive indication that the two models are behaving similarly is that they both predicted that New York state would have the highest response in EV registrations to an increment in EV fuel stations. The Random Forest model predicted that 1 fuel station would cause 39,397 vehicle registrations in New York, and the Ridge model predicted that 1 fuel station would cause 125,365 more registrations in New York. These values are too large to correspond to an actual response because there are only 62 times as many electric vehicles registered in the USA as there are fuel stations in the USA. But, it indicates that New York may have among the largest actual response in registrations to an increment in fuel stations, even if the exact value of that response is not being correctly quantified.

Appendix

Descriptive Statistics	Registration Count	Fuel Station Count	Population	Median Age	Sex Ratio	Mean Income	Area_km2	Pop_per_km2	Fuel Stations per capita	EV Registrations per capita
Count	5.100e+01	51.000	5.10e+01	51.000	51.0	51.00	51.00	51.000	51.000000	51.00000
Mean	6.97e+04	1129.67	6.56e+06	38.46	.98	63862.08	1.93e+0	140.755	0.000164	0.007568
STD	1.78e+05	2226.50	7.45e+06	2.21	.03	9708.44	2.51e+05	531.473	0.000111	0.005840
Min	9.59e+02	60.00	5.81e+05	31.27	.90	46370.00	1.77e+02	0.413	0.000044	0.001226
25%	8.13e+03	261.50	1.87e+06	37.28	.05	57797.50	9.30e+04	19.141	0.000090	0.003832
50%	2.58e+04	500.00	4.51e+06	38.45	.98	62085.00	1.46e+0	37.545	0.000119	0.005661
75%	7.25e+04	1266.00	7.59e+06	39.57	1.00	68912.50	2.18e+05	80.663	0.000189	0.010922
Max	1.26e+06	15555.00	3.88e+07	43.79	1.06	95970.00	1.72e+06	3817.401	0.000572	0.032364

Figure 1. Descriptive Statistics for Stats_ABT dataset

Feature Name	Dataset	Type	Description
State	state, population, EV_registrations, fuel_stations	object	State name where the fueling station is located (e.g., “CA”)
Registration Count	EV_registration	int64	Electric vehicle registrations by state in 2023
Fuel Station Count	“extracted” from fuel_station	float64	Number of fuel stations (for electric vehicles) in each state
Population	population	float64	Number of people at a certain age in each state in 2023
Median Age	“extracted" from population	float64	Average age (in years) of the population in each state
Sex Ratio	“extracted” from population	float64	Ratio of the number of males & number of females in each state
Completing College	education	object	Percentage of adults 25 years of age and older (in decimal form)
Mean Income	“extracted” from income	float64	Average income (in dollars) in each state
Area_km2		float64	Area (in km^2) of state
Pop_per_km2	“extracted” from population	float64	Number of people per km^2 by state
Fuel Stations per Capita	“extracted” from population & fuel_stations	float64	Number of fuel stations per person
EV Registrations per capita	“extracted” from population & EV_registration	float64	Number of electric vehicles per person

The ABT table includes 11 descriptive features (describing 50 data points - one for each state).
Figure 2. Data Dictionary of States_ABT dataset.

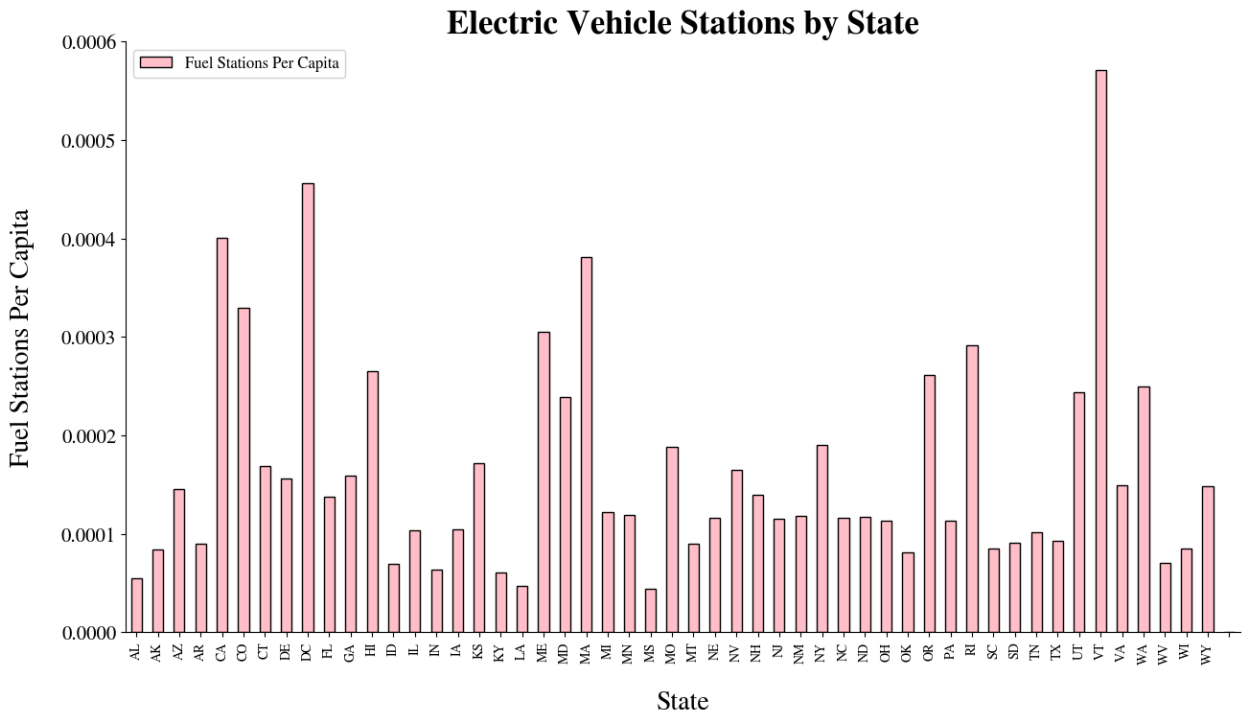

Figure 3. Bar Chart illustrating Fuel Stations Per Capita by Electric Vehicle Stations by State

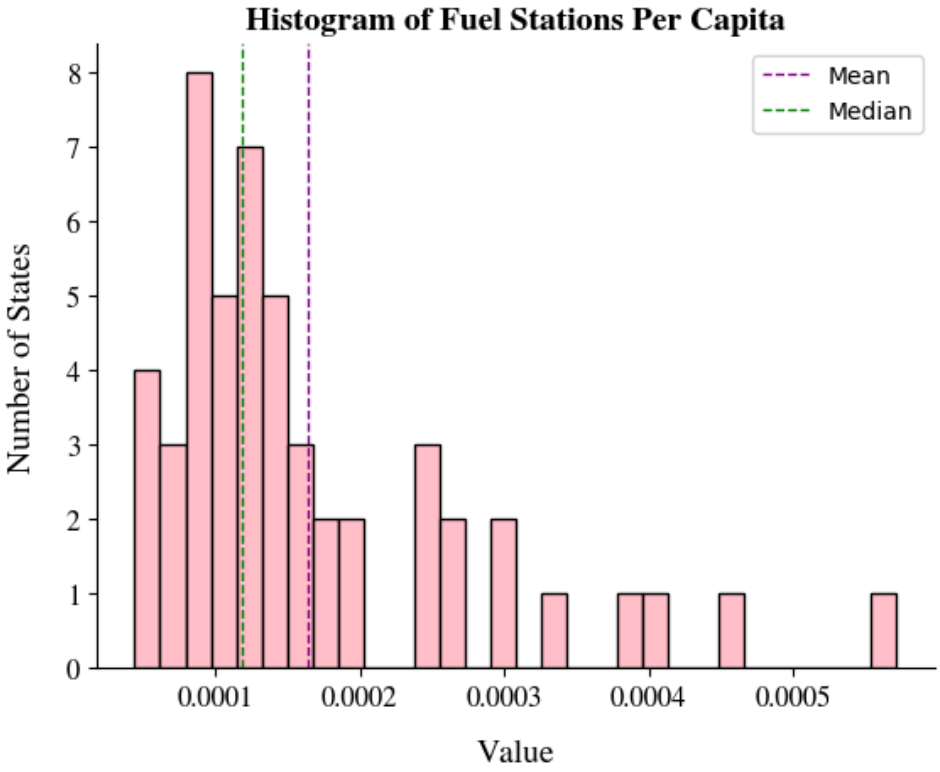

Figure 4. Fuel Stations per Capita Histogram

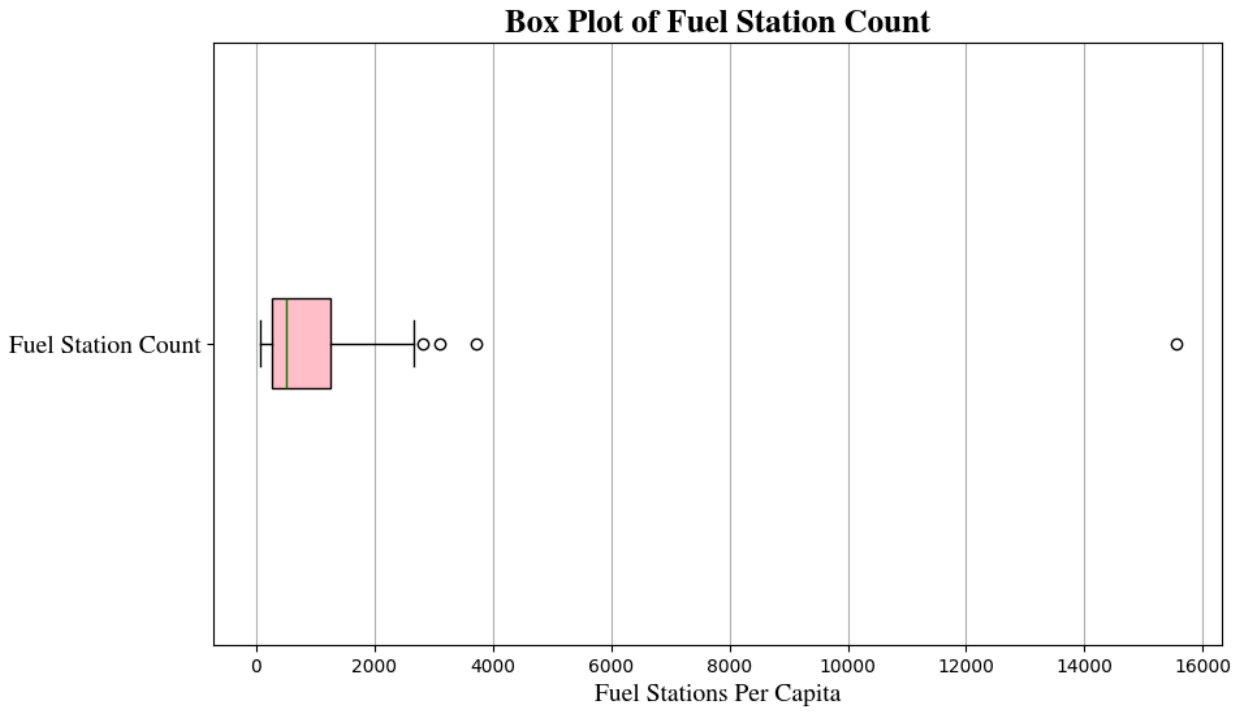

Figure 5. Box Plot of Fuel Station Count

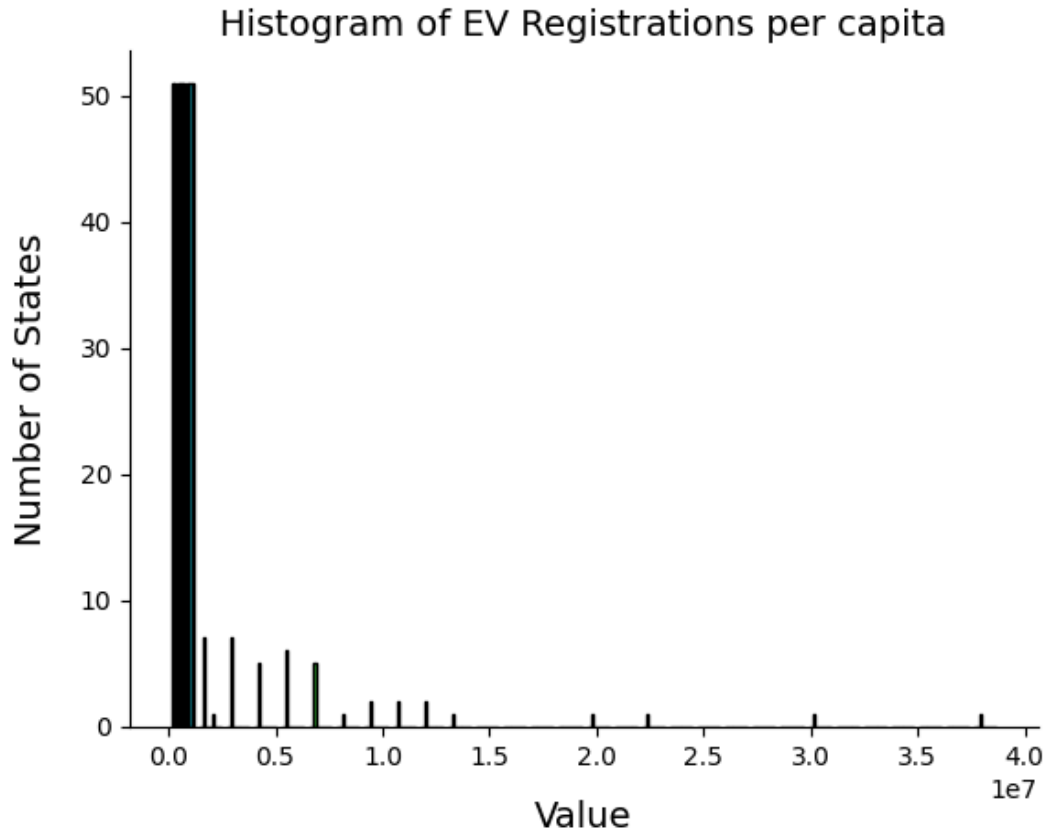

Figure 6. Histogram of EV Registrations per capita

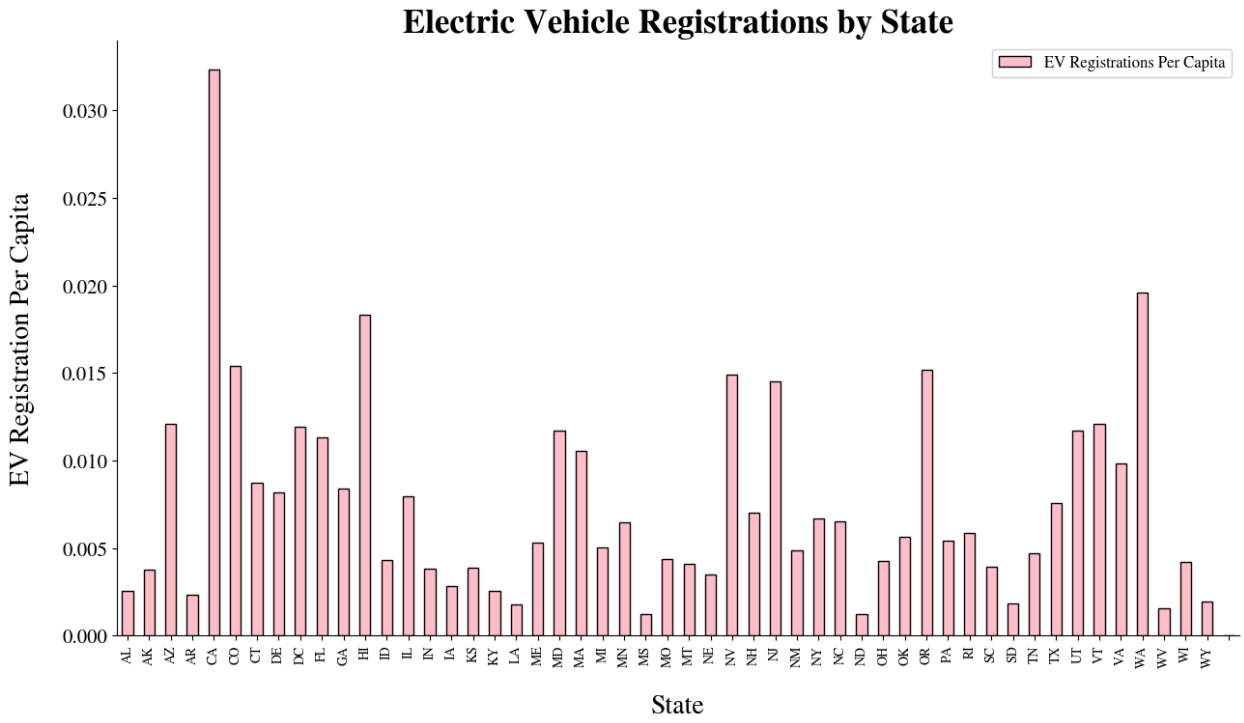

Figure 7. EV Registration by State Bar Graph

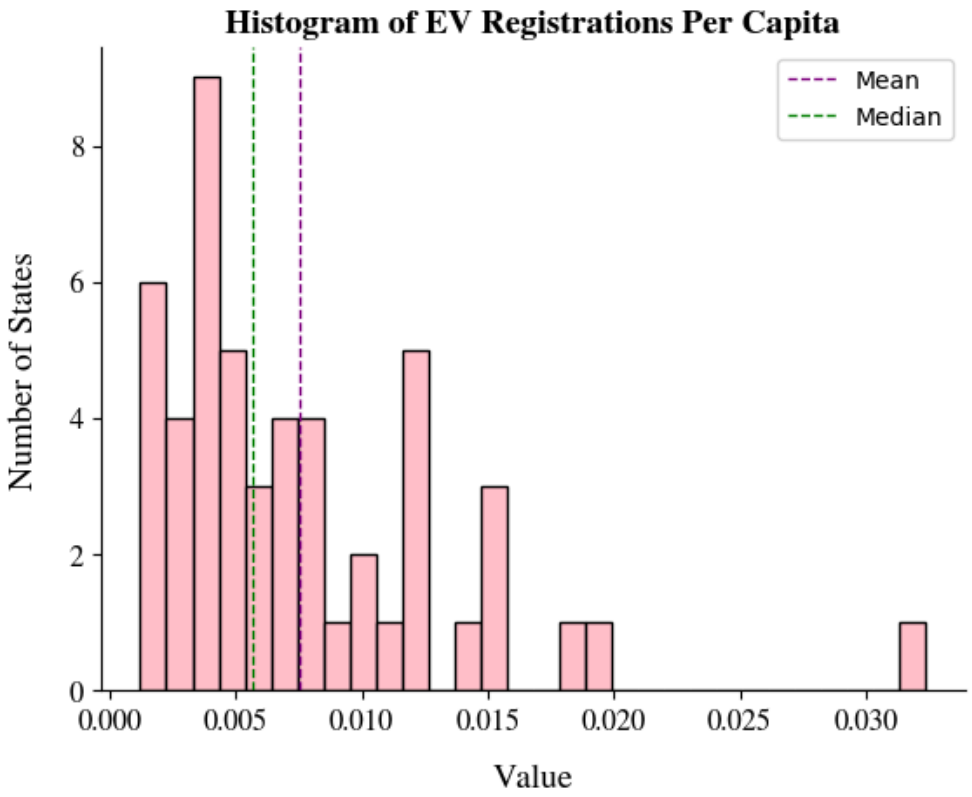

Figure 8. EV Registration per Capita by EV Registration by State Histogram

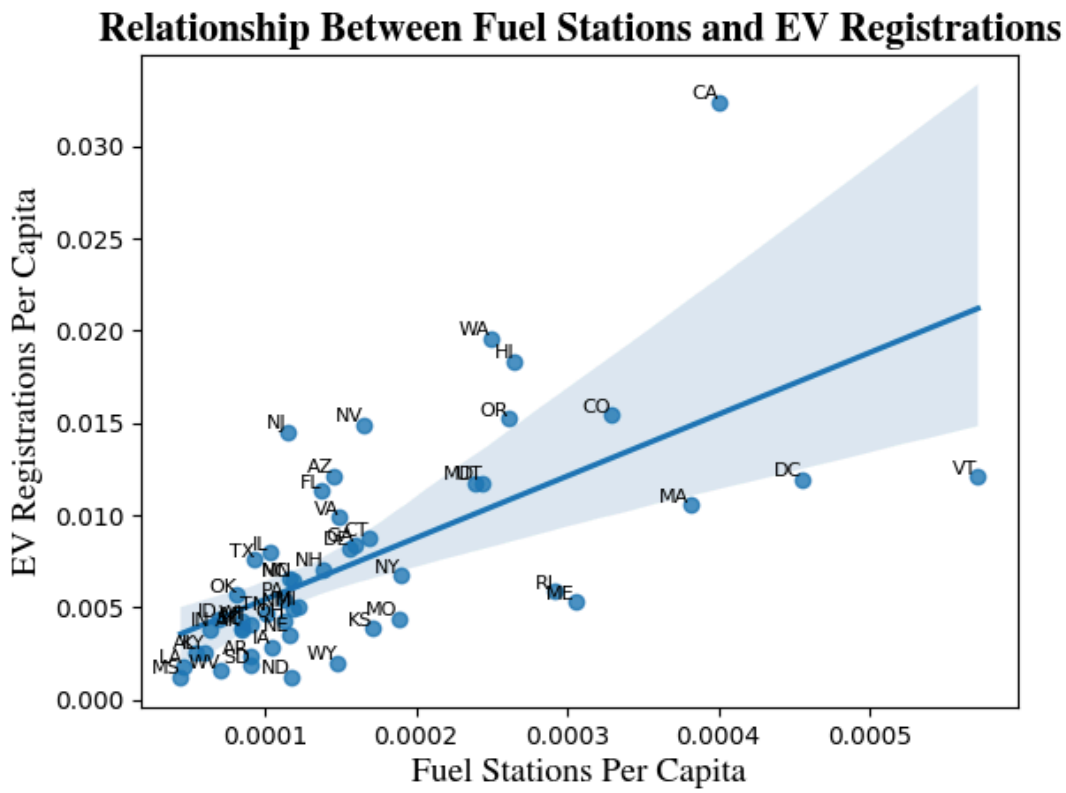

Figure 9. The single-variable relationship between EV Registrations Per Capita and Fuel Stations Per Capita

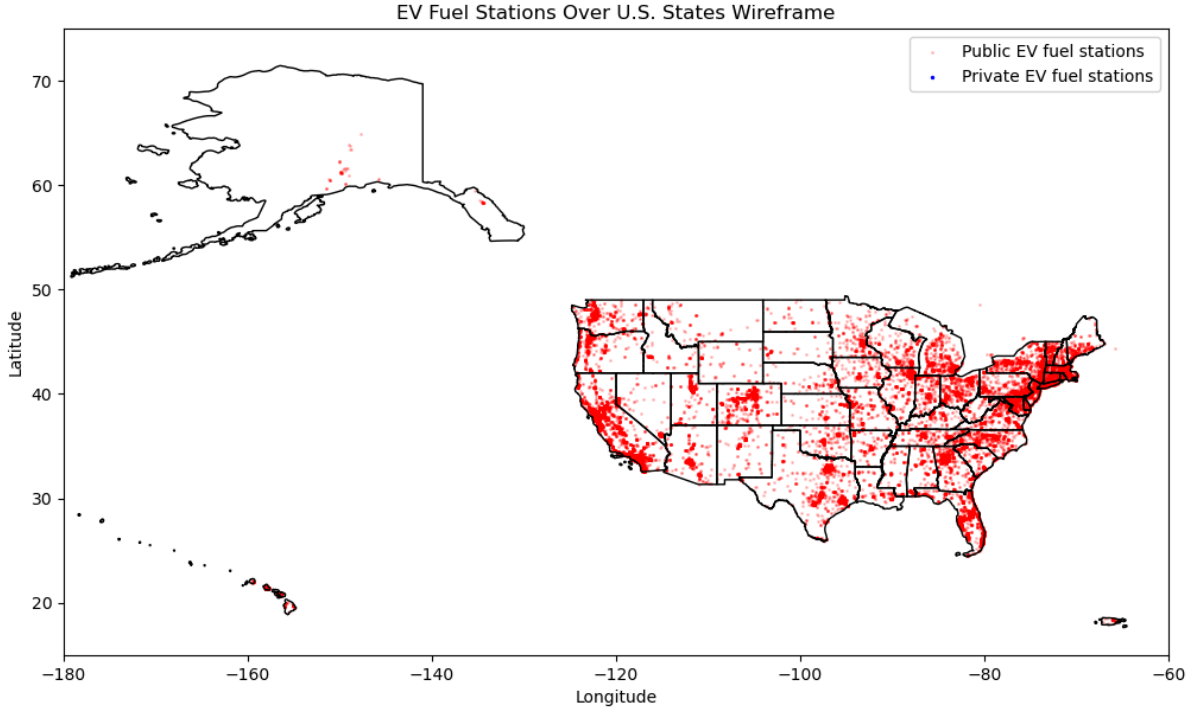

Figure 10. U.S. EV Fuel Station Density Map

Output of hyperparameter tuning for Random Forest Model

```
Fitting 5 folds for each of 96 candidates, totalling 480 fits
Best Parameters: {'bootstrap': False, 'max_depth': 20,
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators':
200}
Best Cross-Validation MSE: 22539447028.570824
Test MSE: 697285889.3740525
```

Output of hyperparameter tuning for Ridge Model

```
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Best Parameters: {'alpha': 100}
Best Cross-Validation MSE: 17456747747.90537
Test MSE: 2427465988.916198
```

Figure 11. Output of hyperparameter tuning for two regression models