Project 2

Deadline: December 01, 2024, 11:59pm ET.

Team Name: Eco-Warriors

Group Members:

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

The transition to electric vehicles (EVs) is essential for addressing the global climate crisis and reducing greenhouse gas emissions. The U.S. Environmental Protection Agency (EPA) has set ambitious targets to achieve Net-Zero Emissions for 2030-2032, requiring a substantial increase in EV adoption. However, despite growing interest in sustainability, only 6.9% of new car sales in 2023 were electric - far short of the EPA's goal of 35% to 56%. A critical obstacle to broader EV adoption is the insufficient availability of charging stations, which discourages potential buyers who fear limited access to recharging infrastructure. This lack of infrastructure creates a gap that must be addressed to make EV ownership viable for a larger portion of the population.

Solving this problem is crucial for several reasons. First, it directly impacts environmental goals by facilitating a faster transition away from fossil-fuel-powered vehicles. Second, strategic investment in EV infrastructure ensures that resources are allocated efficiently, avoiding wasteful spending on stations in areas where they will have little impact. Third, overcoming this barrier will encourage consumers to adopt EVs, driving economic growth in the renewable energy and EV manufacturing sectors. Government agencies, private investors, and environmental organizations all stand to benefit from data-driven strategies to prioritize charging station deployment. The solution to this problem not only supports national climate objectives but also empowers policymakers and stakeholders with actionable insights to optimize their investments.

The problem can be clearly stated as follows: **Within a given geography, what will be the marginal increase in the number of electric vehicles registered if the EPA subsidizes the addition of one EV charging station?** To answer this, we need to develop a predictive model that identifies areas where adding a charging station will have the greatest impact on EV adoption. By analyzing factors like income, population density, and existing EV infrastructure, the model will guide strategic investments in charging stations to maximize their impact. This approach ensures that every new charging station contributes meaningfully to the growth of EV adoption, paving the way for a cleaner, more sustainable future.

To address the challenge of increasing EV adoption by strategically expanding changing infrastructure, our solution involves developing a data-driven predictive model. This model will analyze key demographic, geographic, and infrastructure-related variables to identify regions where subsidizing the construction of new EV charging stations will yield the greatest marginal increase in EV registrations. By leveraging state-level data on EV registrations, charging stations, and demographic characteristics, the model will guide the EPA in prioritizing investments for maximum impact.

The proposed solution is built on a foundation of rigorous statistical analysis and machine learning. Specifically, we employ regression models - including random forest and Ridge regression - to predict the expected change in EV registrations when a new charging station

is added to a given state. These models are trained on an analytical base table (ABT) constructed from comprehensive datasets, including the number of existing EV registrations, charging stations, and demographic indicators such as population density, income levels, and college completion rates. This data-driven approach ensures that predictions are grounded in real-world evidence and tailored to the unique characteristics of each state.

This solution is the right approach because it balances practicality and precision. The use of predictive modeling allows us to account for complex, non-linear relationships between variables, ensuring that the recommendations are nuanced and actionable. By focusing on marginal increases in EV registrations, the solution directly aligns with the EPA's goals of maximizing the impact of infrastructure investments while staying cost-effective. Moreover, this method is scalable; as more granular data (e.g., at the county level) becomes available, this model can be refined to provide even more recommendations. This adaptability makes it an ideal tool for long-term strategic planning to accelerate EV adoption and achieve sustainability targets.

Our approach to addressing this problem of insufficient EV charging infrastructure is built on a predictive framework that identifies regions where additional charging stations would have the highest impact on increasing EV registrations. The solution architecture is composed of three key parts: data preprocessing, predictive modeling, and impact analysis. These components are designed to work together to produce actionable insights for guiding the strategic placement of EV charging stations.

The first component, data preprocessing, ensured that the datasets are clean, consistent, and ready for analysis. We compiled data from various sources including EV registrations from the Alternative Fuels Data Center, charging station data from Kaggle, and demographic information from the U.S. Census Bureau and other government repositories. All data was standardized to the state level for consistency and combined into a unified analytical base table (ABT). Key steps included selecting the relevant features from the datasets, including population density, income levels, and college completion rates. As well, excluding irrelevant and incomplete data like the information on hydrogen fuel stations. We built derived features like "EV registrations per capita" and "fuel stations per capita" to better capture the relationship between demographics and infrastructure. All of these steps ensured the dataset was both comprehensive and focused on the variables directly related to EV adoption.

The second component, predictive modeling, is the core component of the solution. We used two regression models - random forest and Ridge regression - to predict the marginal increase in EV registrations resulting from the addition of a charging station in a specific state. The random forest model is useful because it handles complex relationships in the data, making it effective for analyzing how factors like income, population density and current

infrastructure influence EV adoption. The Ridge regression model provides a simpler way to see how each factor contributes to the prediction. Both models were trained on the ABT, with the number of EV registrations as the target variable. To ensure both models worked accurately, we optimized them using methods like hyperparameter tuning and tested them on a separate dataset to check their accuracy. These models give us a reliable way to estimate the impact of adding charging stations across different states.

The final component, impact analysis, involves using the predictive models to simulate the effect of adding one charging station in each state. For each state, the "fuel station count" feature was incremented by one, and the models predicted the resulting change in EV registrations.

The model fitting occurred in two stages: First with an ABT of 50 states (one row per state) in the first iteration of model-fitting, and then with an ABT of 1166 counties (one row per county) in the second iteration of model-fitting.

Both iterations of models were built to predict the single target variable: "Count of Electric Vehicle Registrations" based on the following 8 target variables of demographic information of the population of that state (or county):

- Count of Electric Vehicle charging stations (EV Fuel Stations)
- Population
- Median Age
- (Logarithmic) Sex ratio
- Rate of completing college
- Mean Income (USD)
- Population per square kilometer
- Fuel Stations per capita

The fitting occurred via a standard search of candidate hyperparameter values chosen to minimize the average mean-squared error in validation sets in a 5-fold cross-validation procedure.

The only parameter that needed to be searched in the ridge regression model fitting was the regularization parameter, "alpha."

The parameters that needed to be searched in the random forest model fitting were "n_estimators" (integer), "max_depth," (integer), "min_samples_split" (integer), "min_samples_leaf" (integer), and "bootstrap" (True or False).

To predict the impact on the marginal number of EV registrations which would be caused by the additional construction of +1 EV charging stations in the county, the following procedure

was used:
Each of the 1166 rows of the county ABT were individually input back into the regression model to predict the number of EV registrations in the county. Then, the row was modified by incrementing the "Count of Electric Vehicle charging stations" variable by 1, and again input into the regression model to predict a hypothetical number of EV registrations. The difference between the first predicted number of vehicles and the second, hypothetical, predicted number of vehicles was recorded for each of the 1166 counties. Then, all counties were compared to find the county with the highest predicted marginal change in electric vehicle registrations.

A random forest model can plausibly capture interactions between predictive feature variables better than a regularized linear regression model. However, an 18% increase in electric vehicle registrations in response to a 0.2% increase in charging stations is not likely to occur, and probably represents an abrupt change in target variable predicted value regions in feature space of the random forest model. So the random forest model may not be well suited to the problem at hand.

By combining the insights from both models, we ensured our recommendations were appropriate and actionable. We chose these approaches because they balance complexity and practicality. Together, these models provide a comprehensive view of the factors influencing EV adoption. To improve the models, we performed rigorous hyperparameter tuning, derived normalized features like "stations per capita," and conducted outlier analysis to retain meaningful data points. Our methodology ensures that our solution can adapt to more granular data in future iterations. Our solution provides a powerful tool to guide strategic EV infrastructure investments, accelerating the transition to sustainable transportation.

Based on the results of our analysis and predictive modeling, **we recommend building an EV charging station in two distinct counties in the US: Los Angeles County, California and New York County, New York**. Los Angeles County was identified as one of the most impactful among all regions analyzed, with an expected increase of **81 additional EV registrations** resulting from the addition of one singular charging station, according to the Ridge Regression Model fitted to the county-level ABT of 1166 counties. Los Angeles County currently has 320,110 electric vehicles registered and 3,738 EV charging stations. New York County was identified as one of the most impactful among all regions analyzed, with an expected increase of **1140 additional EV registrations** resulting from the addition of one singular charging station, according to the random forest regression model fitted to the county-level ABT of 1166 counties. New York County currently has only 6,220 electric vehicles registered and 398 EV charging stations.

This recommendation aligns with our findings that regions with high population density, strong income levels, and existing but limited EV infrastructure tend to show the greatest responsiveness to additional charging stations. Los Angeles and New York, are both major

metropolitan counties, fitting the profile and offering a high return on investment for the EPA's infrastructure subsidies. By prioritizing these locations, and other locations that "agree" based on the predictive models, policymakers and stakeholders can maximize the environmental and economic benefits of their investment, accelerating EV adoption and advancing sustainability goals.

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.

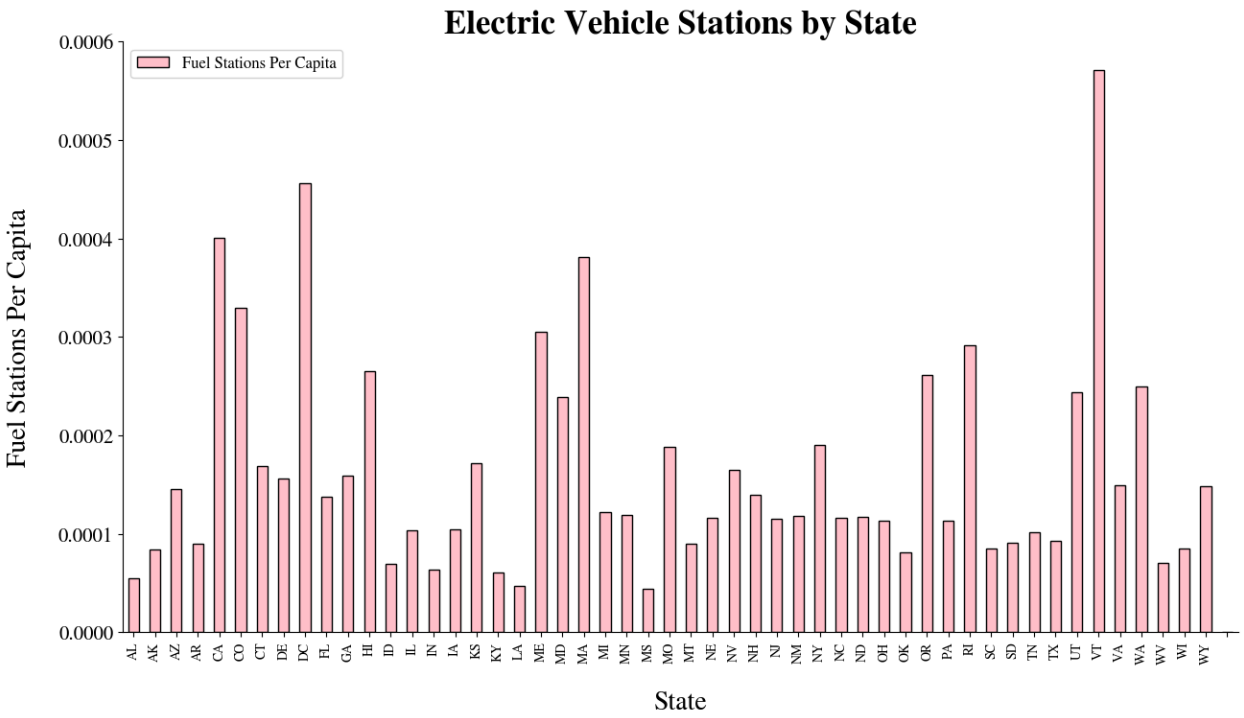

Figure 3. Bar Chart illustrating Fuel Stations Per Capita by Electric Vehicle Stations by State

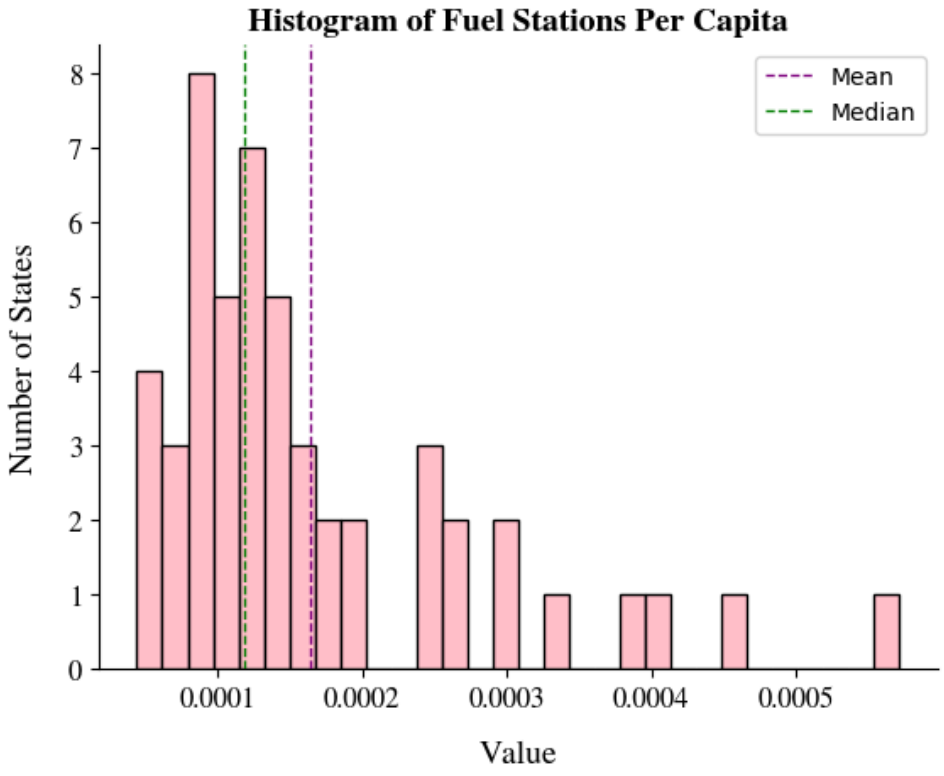

Figure 4. Fuel Stations per Capita Histogram

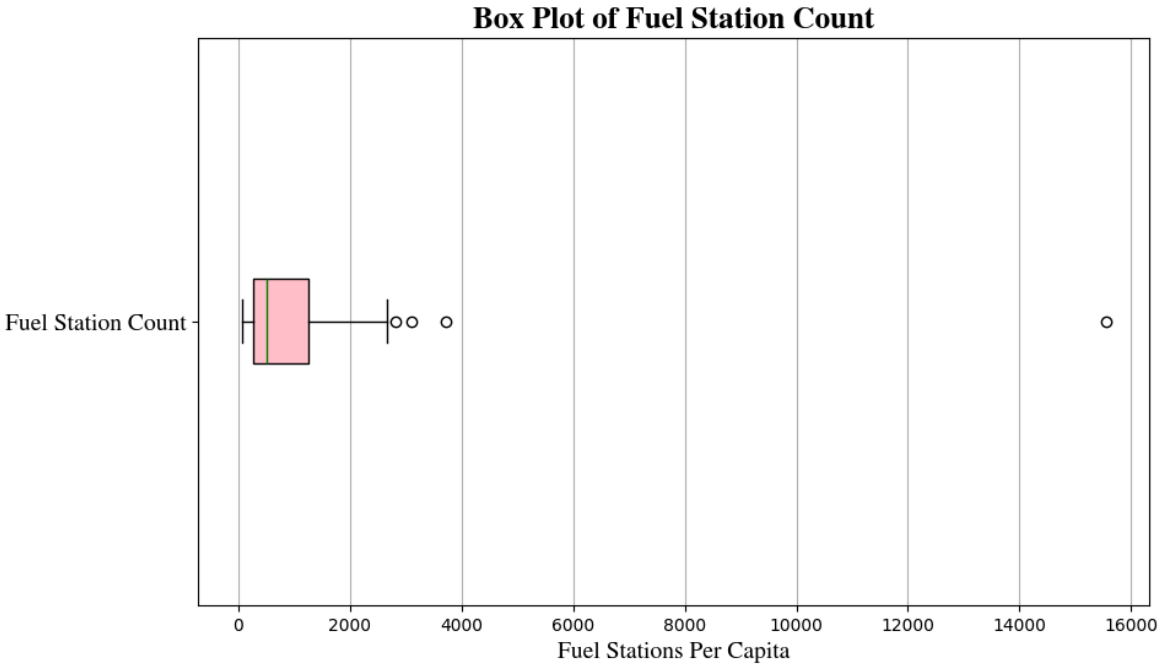

Figure 5. Box Plot of Fuel Station Count

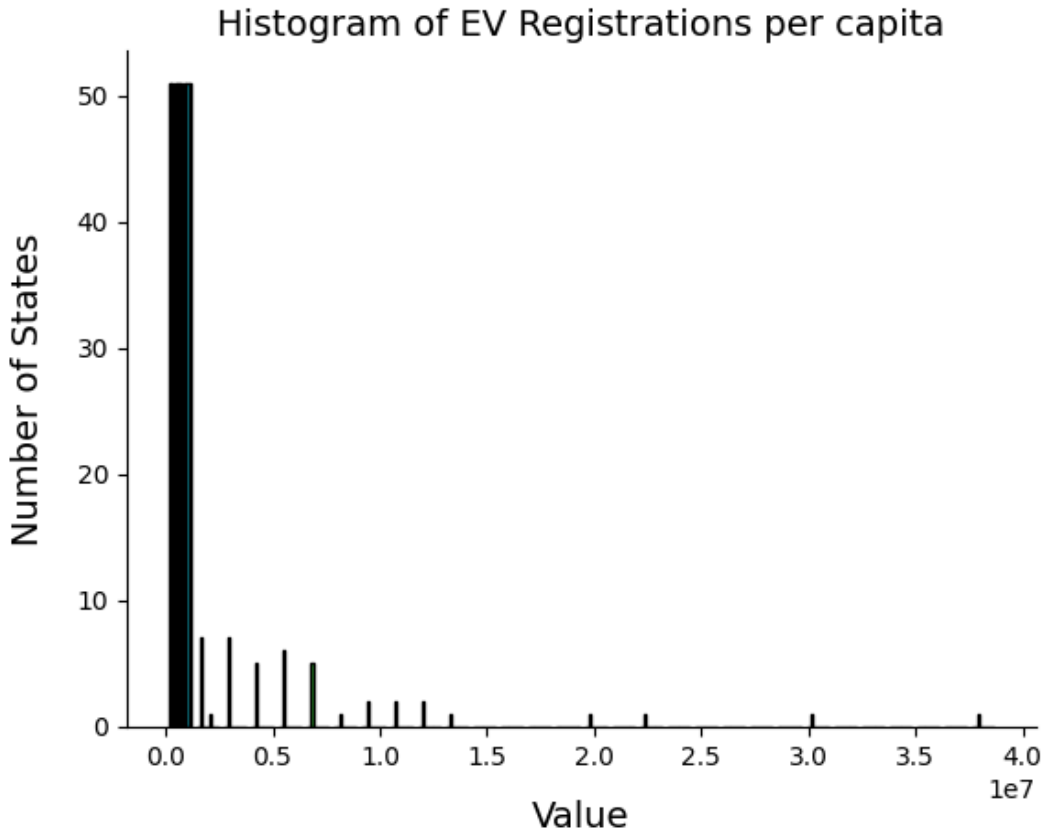

Figure 6. Histogram of EV Registrations per capita

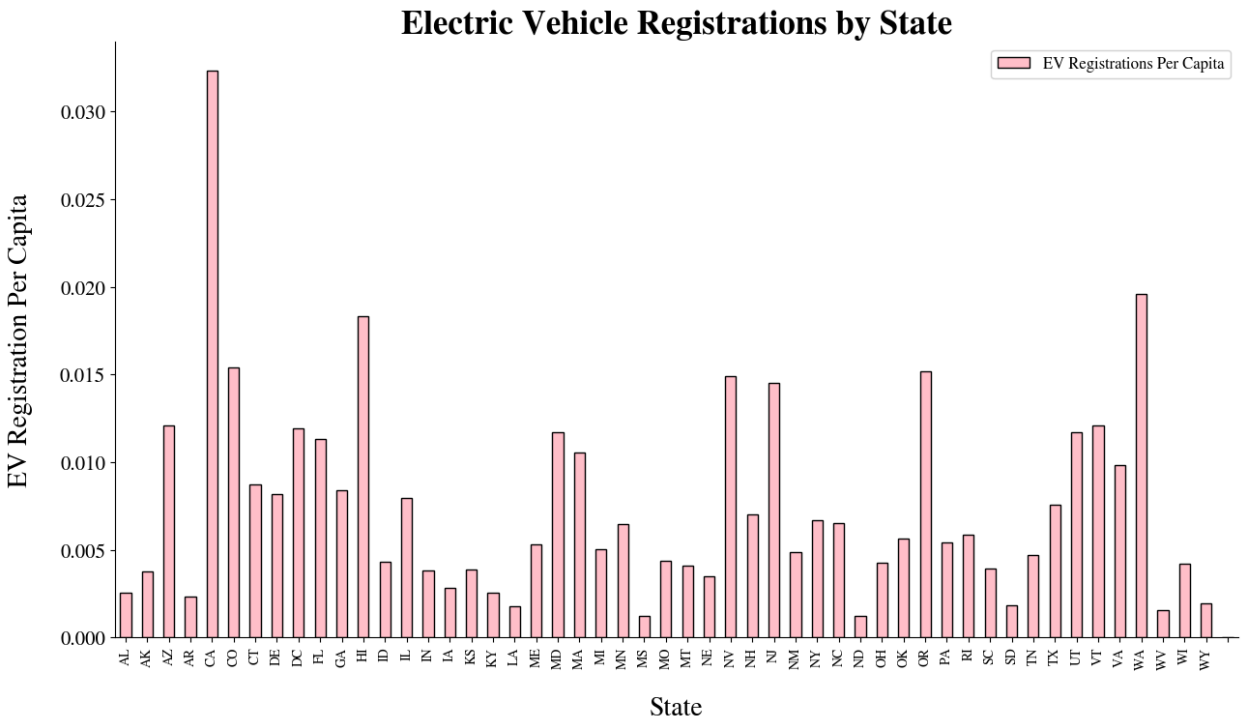

Figure 7. EV Registration by State Bar Graph

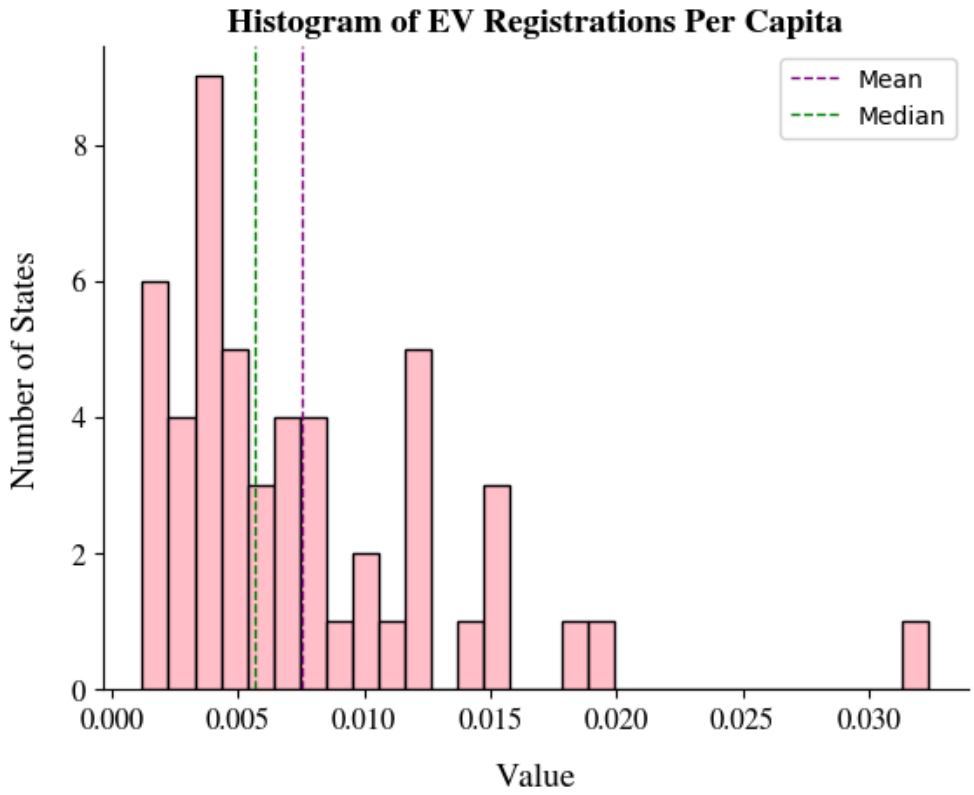

Figure 8. EV Registration per Capita by EV Registration by State Histogram

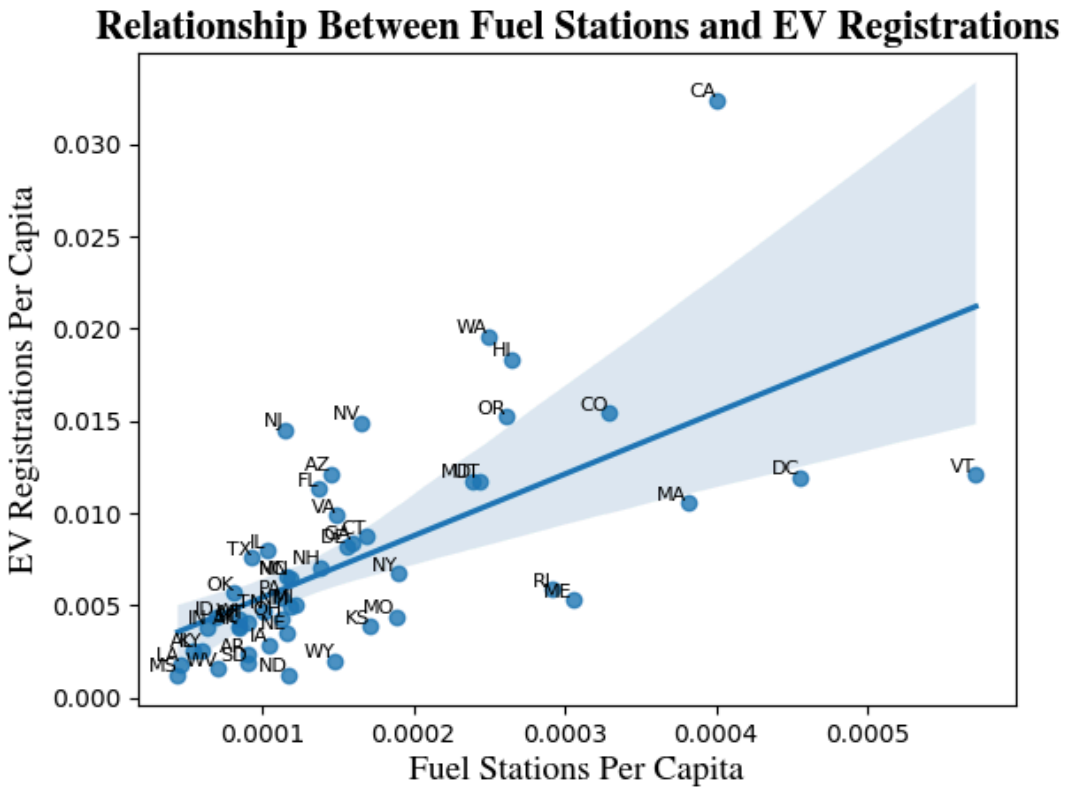

Figure 9. The single-variable relationship between EV Registrations Per Capita and Fuel Stations Per Capita

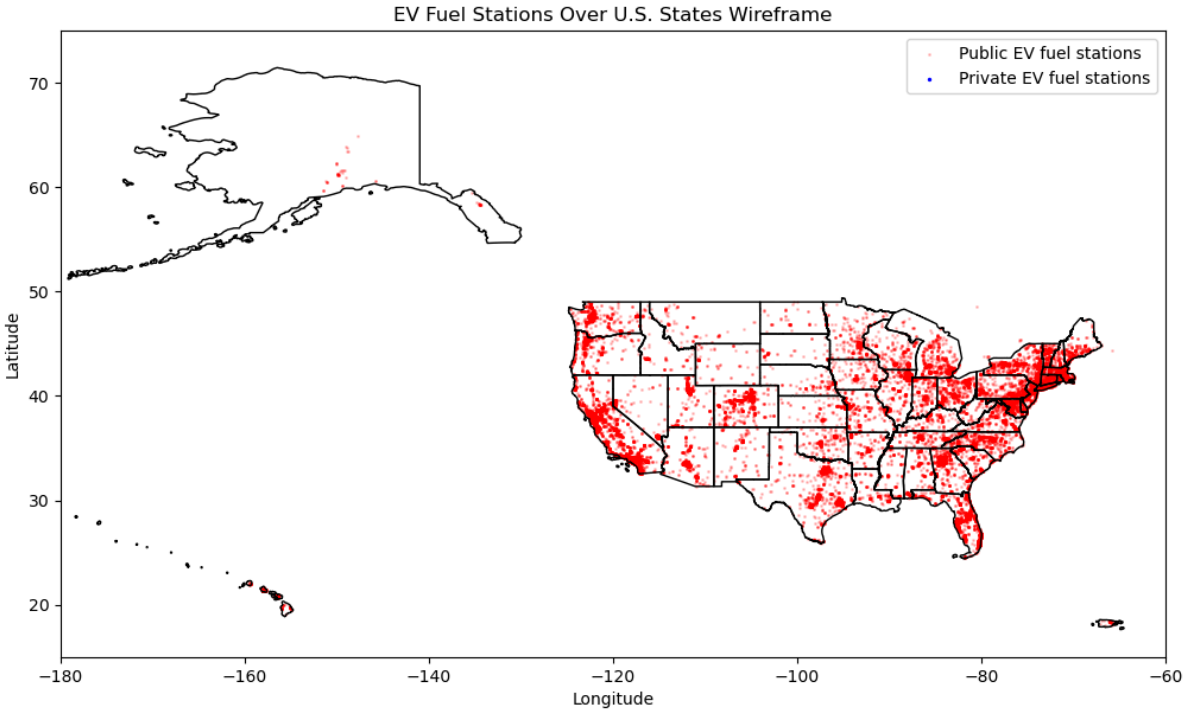

Figure 10. U.S. EV Fuel Station Density Map

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 (Project 1)

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': 30,
'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators':
100}
Best Cross-Validation MSE: 61010638.146991864
Test MSE: 4868617.38851825
```

Output of hyperparameter tuning for Ridge Model

```
Fitting 5 folds for each of 5 candidates, totalling 25 fits
Best Parameters: {'alpha': 0.01}
Best Cross-Validation MSE: 10645149.745769624
Test MSE: 6752570.461839935
```

Figure 12. Output of updated hyperparameter tuning for two regression models (Project 2)

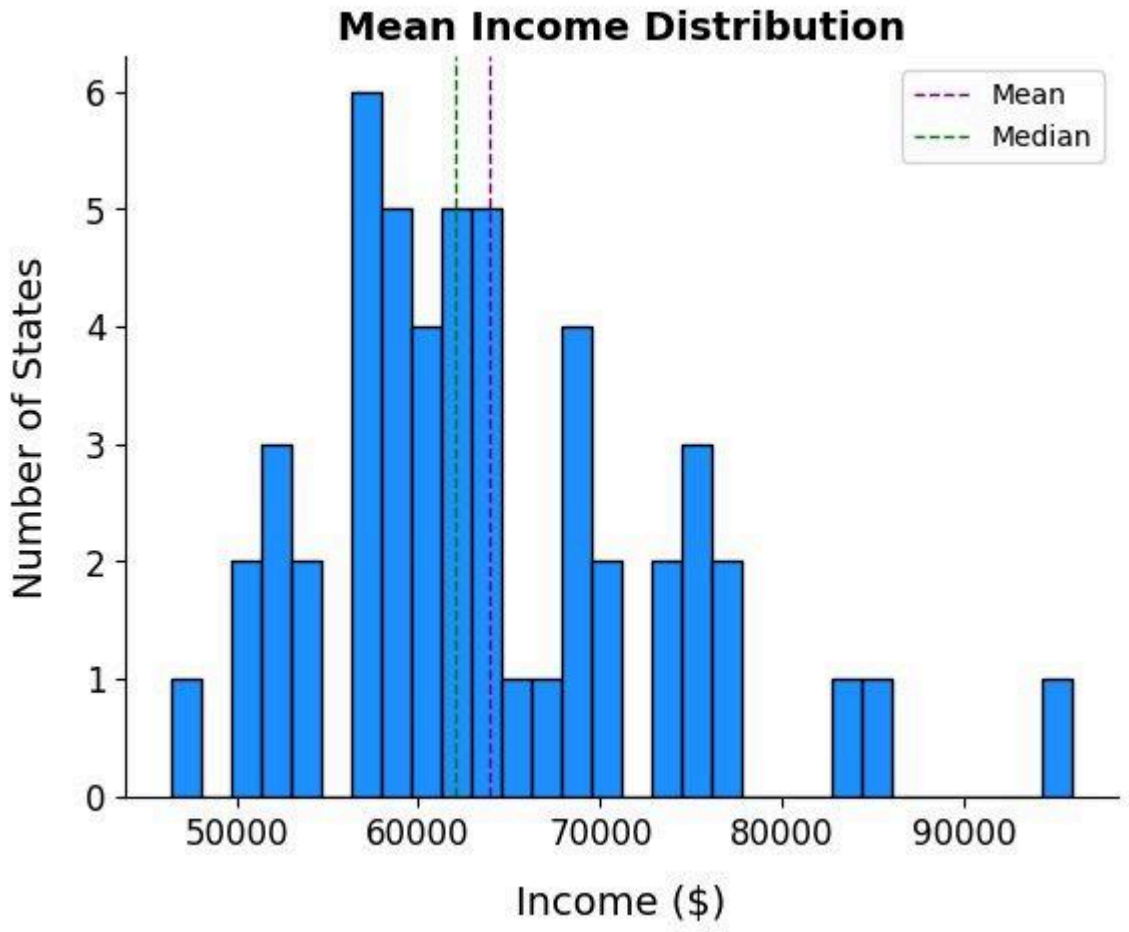

Figure 13. Mean Income Distribution by Number of States