ELECTRIFY AMERICA

1. Introduction

Across the United States, electric vehicles are becoming very popular at a pace that challenges the geographic distribution and capacity of the existing public charging infrastructure. Our project, Forecasting Demand for Electric Vehicle Charging Stations, addresses an important question: **Where will charging demand increase, and how does current infrastructure meet that demand?**

Problem, Impact and Motivation

There's a rapid acceleration in owning electric vehicles which creates demand for electricity consumption and charging infrastructure that differs by region, usage of land and income. EV ownership can be hindered by poorly distributed charging infrastructure. Our analysis will support data planning for governments, utility providers, and private charging companies by forecasting charging demand and identifying gaps in infrastructure. Our project can help optimize investments, decrease overcrowding at high use charging stations, and encourage accessibility to EV charging across America.

We are motivated to explore this project due to the growth of decarbonization policies, and a focus on environmentally friendly transportation. We noticed that there are several analyses focusing mainly on vehicle registration or station availability, but only few study about **charging behavior** and **infrastructure characteristics** at a granular, station-level scale. To address this gap, our team took this opportunity to combine multiple national datasets which we found. The datasets we decided to combine are EV WATTS, Alternative Fueling Stations, AFDC Vehicle Registration, and U.S. Census data to build a strong view of electric vehicle's charging demand.

Summary of Methods

Our project uses a two-part analytical framework:

1. **Supervised Learning.** We used regression and classification methods to model the charging demand of electrical vehicles with metrics of energy consumed per session and average daily session counts. The algorithms that we tested include Ridge Regression, Random Forest Regressor, and Gradient Boosting Regressor were chosen to represent the different models for evaluation purposes. These models use temporal, station-level, and regional socioeconomic features to predict demand and distinguish high- versus low-usage stations.
2. **Unsupervised Learning.** To group locations with similar infrastructure and usage patterns, we performed clustering analysis on combined station and regional data. We used methods like K-Means, DBSCAN, and hierarchical clustering to reveal "EV hot zones," marginalized regions, and usual stations like high-capacity fast hubs, community chargers and workplace stations. e.g., high-capacity fast hubs, community chargers, and workplace stations).

Novel Contributions.

In contrast to previous studies that analyze charging data or isolated infrastructure, our project **merges session-level, station-level, and demographic data** into a common analytical framework. Through this integration it allows us to observe how factors such as **median income, population density**, and **EV registration counts** influence charging demand and infrastructure availability.

Main Findings.

Our supervised models show strong patterns in between **energy consumption, time of day**, and **number of available fast chargers**, which as a matter of fact validate temporal and vehicle capacity-based demand trends. Similarly, unsupervised clustering helped reveal regional groups such as urban high-demand zones, balanced suburban regions, and rural areas with low usage, and short supply of infrastructure. Overall, our findings reveal actionable insights regarding where new charging stations are required at most, and which features define EV stations.

2. Related Work

A lot of prior studies to ours and data related projects have observed topics of electric vehicle charging infrastructure and usage patterns. However, only a few studies made it a point to integrate charging demand forecasting and station-level clustering across many national datasets as we do in our project.

1. National Renewable Energy Laboratory (NREL) — Commercial-Scale Battery and Charging Infrastructure Analysis (2016).

This NREL article discusses the analysis of deployment costs, usage rates, and demand patterns across the United States for large scale energy storage and EV charging systems. The focus of this article was mainly on the demand of electric vehicles and the energy grid's impact around America. In contrast, our work is different as we focus on session-level charging demand forecasting, and clustering at a focused geographic distribution. Also, we added behavioral and infrastructural data rather than project level summaries like the article.

2. Department of Energy's Alternative Fuels Data Center (AFDC) — Electric Vehicle Infrastructure Trends Report (2022).

This report discusses statistics on a national level of the growth of public charging stations and summarizes counts of Level 2 charger and DC Fast Chargers by region and network provider. Our project enhances AFDC infrastructure data with session-level usage from EV WATTS, and socioeconomic features from Census data to model actual charging behavior and recognize over or under supplied areas.

3. Sharma et al., "Data-Driven Siting of Electric Vehicle Charging Stations Using Machine Learning" (IEEE Transactions on Transportation Electrification, 2021).

This academic study made it a point to apply supervised machine learning to predict prime charging station locations depending on urban density, traffic volume, and socioeconomic features. Even though the model can predict a wide level of detail, our main focus is to predict demand for new charging stations using the available data sources that we found. In contrast, we conducted our research to use open public datasets and by adding an unsupervised learning component to cluster stations that already exist, and regions by demand and infrastructure characteristics. This approach will help policy planners through the forecasting and insights we gather.

Project Continuity.

Our project only consists of original work for our Milestone II course **and is not a continuation of a prior milestone or course project.** All our code, feature engineering, and analysis we developed independently by us Team 23.

3. Data Sources

Our project combines multiple public datasets and APIs to integrate **EV charging session data**, **infrastructure characteristics**, and **regional demographics** into a single Python dataframe. The sources that we have chosen are openly accessible, and maintained by the government and are within the time frame of 2019 to 2023.

3.1 EV WATTS Public Dataset

The **EV WATTS Public Dataset** on the EV WATTS portal is provided by Energetics and the U.S. Department of Energy, and it forms the foundation of our analysis. It is in a CSV format and has over 300,000 session-level records gathered across the United States from the years of 2019 to 2023. Each record contains detailed information about individual charging sessions, including fields like session ID, start and end timestamps, total and active charge duration, energy consumed (in kWh), connector type, power rating, charge level, pricing, and regional identifiers. This dataset is important to our **supervised learning models**, because it is used for forecasting of energy consumption and the identification of temporal charging demand patterns at the session level.

3.2 Alternative Fueling Stations (AFS) Dataset

The **Alternative Fueling Stations (AFS) Dataset** provided by the U.S. Department of Transportation's Bureau of Transportation Statistics, available on geodata.bts.gov, has complete information on charging infrastructure. It is in a CSV format and contains approximately 305,000 records of alternative fueling stations across the United States as of 2023. For our analysis, we filtered the data to include only electric charging stations where fuel_type_code = "ELEC". The key variables include station name, city, state, latitude, longitude, connector

types, counts of Level 2 and DC fast chargers, pricing, facility type, access hours, and network affiliation. This dataset provides the **supply side of EV infrastructure**, which helps us to combine charging capacity, pricing type, and station characteristics to regional demand patterns derived from session data.

3.3 AFDC Vehicle Registration Dataset

The **AFDC Vehicle Registration Data** published by the U.S. Department of Energy's Alternative Fuels Data Center (AFDC Vehicle Registration), provides annual counts of EVs by state from the years of 2018 to 2023. It is in a CSV format, it aggregates registration data for battery electric vehicles (BEVs), plug-in hybrid electric vehicles (PHEVs), and hybrid electric vehicles (HEVs). The key attributes include state, year, and the number of registered vehicles in each category. This dataset serves as a substitute for EV adoption and market maturity, and it was used both as a predictive variable in supervised models ,and as a clustering feature in unsupervised analyses to evaluate the relationship between adoption levels and the density of infrastructure.

3.4 U.S. Census Population and Income Dataset

The **U.S. Census Population and Income Data** from the U.S. Census Bureau API provides demographic and socioeconomic factors for each region. We retrieved the data programmatically in JSON format and converted it into CSV files using Python's requests library. We extracted the year of 2023 data from the **American Community Survey (ACS) 5-Year Estimates**, including key variables like total population with the field name B01003_001E, and median household income with the field name B19013_001E. These features were merged into the merged dataset to quantify socioeconomic conditions, which we then used to cluster regions based on population size, income distribution, and EV adoption levels.

3.5 State–County FIPS Reference Dataset

To standardize geographic references across all datasets, we used the **State–County FIPS Reference Data**, maintained by the U.S. Department of Transportation and available at data.transportation.gov. This CSV table provides state and county names to their official Federal Information Processing Standards (FIPS) codes. It includes over 3,000 counties across all U.S. states and territories and was used to merge data across multiple geographic levels (state, county, and ZIP) by generating consistent join keys such as the combined geographic identifier (GEOID = state_fips + county_fips).

3.6 Combined Dataset

After cleaning, filtering, and merging the sources above, we produced a unified dataset with approximately **250,000 electric charging station and session records**, each linked to regional demographics and EV adoption metrics. This merged dataset spans the period **2019–2023** and serves as the foundation for both our **supervised forecasting models** and **unsupervised clustering analyses**.

4. Feature Engineering

A very important stage is featuring engineering in changing different levels of our raw datasets such as **EV WATTS, Alternative Fueling Stations (AFS), AFDC Vehicle Registration, Census data, and FIPS** reference table into a well combined dataset that is suitable for both supervised and unsupervised modeling. Overall, the workflow involves five major steps which are: **data cleaning, standardization, merging, aggregation, encoding, and feature derivation**.

4.1 Data Preprocessing and Cleaning

The raw data sources we worked with were very different in structure, and coverage. Initial preprocessing included:

- **Date and time formatting:** We converted start_datetime and end_datetime in EV WATTS into standardized datetime64 objects, and extracted year, month, day, and hour features.
- **Geographic standardization:** We ensured to make all the datasets aligned to state and county levels using the **FIPS reference table** to make consistent unified joins between session-level, station-level, and demographic datasets.
- **Type conversions:** We have converted numerical fields like energy_kwh, charge_duration, power_kw, and demographic variables such as population, median_income to numerical formats using coercion for invalid or values that are not numerical.

- **Handling missing data:** We handled missing data carefully to ensure our merged dataset is consistent and accurate. The missing station-level attributes like ev_pricing, facility_type, and access_days_time were replaced with "Unknown." For demographic and geographic data, we filled in missing values using U.S. Census or aggregates of state-level data. As well, columns with mostly zero or empty values, like ev_level1_evse_num, were kept for reference, however, were not used in the model. Lastly, we used median imputation for any remaining missing numbers, which reduced outliers and protected the data's central tendency.

4.2 Dataset Integration and Aggregation

After cleaning, the datasets were merged at the **station level** and **state–month level**:

- **Join Keys:** The states names of America and FIPS codes (state_fips, county_fips) were used to merge the Alternative Fueling Stations, Census, and AFDC Vehicle Registration datasets.
- **Aggregation:** Session-level data from EV WATTS was aggregated by date, state, and station_id to calculate monthly summaries such as total energy consumed, average power draw, and session counts.
- **Geospatial Enrichment:** Each station record was enhanced with the Census API's population and median income values based on the state and county codes.
- **Normalization:** The continuous variables such as energy_kwh, population, median_income were scaled using min–max normalization for unsupervised clustering tasks.

4.3 Feature Encoding and Transformation

Categorical features such as facility_type, ev_network, and pricing were transformed using:

- **One-hot encoding** was used for linear, and tree-based supervised models.
- **Label encoding** was used for clustering and PCA where dimensionality reduction was applied.

To enhance the interpretability and predictive performance, we derived the following new features:

- total_ports = ev_level1_evse_num + ev_level2_evse_num + ev_dc_fast_num
- connector_diversity = count of unique connector types per station
- sessions_per_day = total sessions / number of active days
- charging_efficiency = energy_kwh / charge_duration
- urbanization_index = ratio of population to number of stations per state

4.5 Output Dataset

The dataset's final output was stored in a CSV file named processesed_ev_demand.csv, which contains approximately 100,000 records with 25 to 30 features. As well, its date ranges from 2019 to 2023, and the level of data is at **Station-Month-State level**. Across regions, this would allow us to perform temporal and geographical analysis of EV charging behavior, observe infrastructure availability, and discover demand patterns. Ultimately, this dataset is our foundation for our supervised and unsupervised models.

5. Supervised Learning

5.1 Methods Description

Our project's goal is to develop a machine learning model to predict **where charging demand is high.** This information will help us analyze how the current infrastructure is meeting that demand. The demand of current electric vehicle charging stations is calculated through our engineered variable of demand_score. We created this as a regression supervised learning problem, where 'demand_score' of a station is the target variable.

The dynamics between Ridge Regression, Random Forest Regression and Gradient Boosting Regression models were compared based on classic regression metrics of Mean Absolute Error (MAE), Root-mean-square deviation (RMSE) , and the Coefficient of determination (R^2).

Table 1: Comparison of Supervised model

Model Family	Type	Rationale
Ridge Regression	Linear (regularized), Probabilistic, interpretable	Extends Linear Regression by adding L2 regularization to handle multicollinearity and prevent overfitting, especially useful when features are correlated.

Random Forest Regressor	Tree-based ensemble	Handles non-linear interactions, robust to outliers, and provides feature importance insights.
Gradient Boosting Regressor (XGBoost)	Boosted ensemble	Captures complex feature interactions and performs well with mixed data types and moderate noise.

Feature Representation

The features contain mixed data types. The numerical features consisted of station level information like number of ports, energy consumed per session, duration of charging time, average sessions, etc., and state demographic information like EV registrations, population, median income, etc. The categorical features of our stations included charge type, EV pricing, facility types, etc. All of the numerical features were standardized, and categorical features like facility_type, pricing, and ev_network were one-hot encoded. The missing numerical values in population, median income, registration counts were imputed from the same state. Furthermore, the Random Forest Regressor was used to impute the categorical features of facility type and station pricing, based on EV station attributes like total duration, number of ports, total sessions, etc. and state.

5.2 Evaluation

Metrics Used

For continuous targets, we evaluated using **RMSE**, **Mean Absolute Error (MAE)**, and **R^2**. These metrics helped us gain the overall predictive accuracy, deviation magnitude, and model fit quality. Across folds, the ensemble models Random Forest and Gradient Boosting consistently outperformed linear regression. As a result, this confirms that EV charging demand reveals strong non-linear relationships with not only station characteristics, but demographics too.

Table 2: Regression Model Comparison

Model	Mean RMSE	Mean MAE	Mean R^2
Linear Regression	262.96	117.54	0.77
Random Forest	20.33	0.46	0.99
Gradient Boosting	12.97	1.72	0.99

Cross-Validation Summary

GridSearchCV with 5-fold cross-validation was conducted on Random Forest Regressor for the training data to select optimal hyperparameters, preserving the test set for final evaluation.

Table 3: Model cross-validation summary

Model	Mean RMSE	Mean MAE	Mean R^2
Original Random Forest	20.33	0.46	0.99
Random Forest Updated HyperParameters	11.26	1.78	0.99

5.3 In-Depth Evaluation

Feature Importance

We conducted a SHAP analysis to further understand how and why our model makes its predictions for demand scores and a graph of this analysis can be found in Appendix C. The model's most important features of TOTAL_SESSIONS and NUM_PORTS by far had the strongest impact on stations with high demand scores. This shows that high values of these features increase the average marginal contribution across all possible combinations

of other features producing higher predictions. The features EV_PER_STATION and ENERGY_KWH had a slight positive effect. We noticed that higher EVs per station only slightly increased demand scores. Stations that provide more energy likely serve more users or larger vehicles, which correlates with the higher demand. Moderately important features include PLUG_IN_HYBRID_VEHICLE_REG_COUNT, TOTAL_DURATION, and EV_INFRASTRUCTURE_INDEX. As well, geographic features described by LATITUDE and LONGITUDE produced a slightly negative impact for low latitude/longitude (blue dots to the left), suggesting southern/western stations may have lower demand. Furthermore, state income and population both indicate higher values had higher demand score (red dots to the right). This SHAP plot reveals that the model predicts higher demand scores for stations with more sessions, more ports, higher energy use, and in wealthier, more populous areas. However, caution is needed due to the possible data leakage from including the target variable as a feature.

Feature Ablation:

We conducted an ablation analysis to observe changes in model performance through RMSE when specific features were removed. Lower RMSE indicates better performance, so a large increase in RMSE after removing a feature indicates that the feature is important.

Table 4: Ablation study results for feature importance evaluation

Ablated Feature	RMSE	Δ RMSE vs Baseline
None (full model)	11.26	-
CHARGE_DURATION	19.07	7.81
NUM_PORTS	18.93	7.67
avg_charge_time	17.90	6.64
PLUG_IN_HYBRID_VEHICLE_REG_COUNT	17.85	6.59
EV_INFRASTRUCTURE_INDEX	17.79	6.53

Key Findings

POPULATION is by far the most critical feature because removing it causes RMSE to jump from 0.67 to over 4.3-, a 540% increase. This suggests that the model heavily relies on population to make accurate predictions, likely because population correlates strongly with demand. Infrastructure & adoption features matter, but less so NUM_PORTS (charging infrastructure) has the second-largest impact. Registration counts of plug-in hybrid and electric vehicles and ENERGY_KWH have modest, but noticeable effects. Their removal increases error by ~12–15%, which is meaningful but not severe. A caveat should be noted, POPULATION might be acting as a proxy for other unmeasured factors like urbanization, income, policy, etc. It's also possible that other features like EV registrations are derived from or correlated with population, so removing POPULATION indirectly weakens those signals too. In the future, a consideration will be made to normalize the model by population. This will allow modeling per-capita behavior (e.g., EVs per 1,000 people) which could reveal more nuanced drivers beyond raw scale. NUM_PORTS is worth keeping because ablation of this feature suggests infrastructure availability matters. The registration counts and ENERGY_KWH add incremental value but is worth keeping.

Learning Curve Analysis

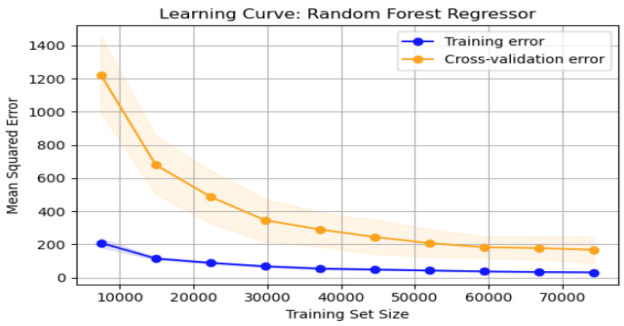

Figure 2. Learning Curve for Random Forest Regressor Showing Training and Cross-Validation Error with Increasing Sample Size

This learning curve visualizes how the model's performance measured by Mean Squared Error (MSE), grows as the size of the training dataset increases from 10% to 100% of the available training data. The plot displays both training error (blue) and cross-validation error (orange), with shaded regions indicating standard deviation across 5-fold cross-validation. The blue line stays very low (close to 0–200 MSE) even at small training sizes and remains nearly flat as more data is added indicating training error is low and stable. However, it seems like validation error starts high, then drops sharply and stabilizes. This demonstrates that the model is learning generalizable patterns from the data, adding more training examples reduces overfitting and improves generalization. There is a performance plateau after 50,000 training samples, since both errors stabilize.

Sensitivity Analysis

To create a generalizable model, we conducted a sensitivity analysis to test which hyperparameters had the largest impact on our model. In our Random Forest model we found that by far the two most important hyperparameters were max_dept and n_estimators. We explored testing a range of values for each of these independently, holding all other hyperparameters constant at their best model values. Then we explored how the two interact with a 10 fold cross validation.

max_depth:

Based on our sensitivity analysis to max_dept, our model benefits significantly from increasing depth up to from about 10–15. Beyond that, no meaningful gain, deeper trees don't improve generalization (may even overfit slightly). It might be commended to set max_depth=10–15 because it seems to be the optimal balance between performance and complexity.

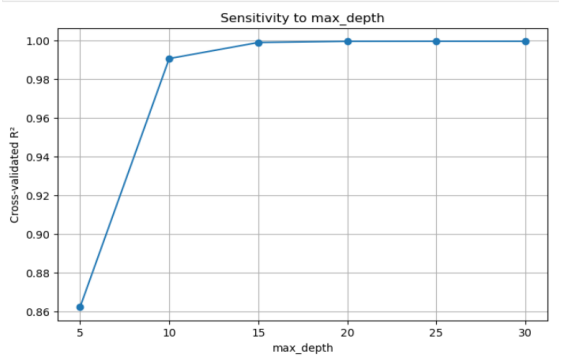

Figure 3: Sensitivity Analysis of Random Forest Model Showing the Effect of Maximum Tree Depth on Cross-Validated R² Performance

n_estimators:

Based on our sensitivity analysis to max_dept, 100–200 trees give near-optimal performance. Adding more trees beyond 300 gives diminishing returns (e.g., +0.0001 R^2). This makes the trade-off of more trees having slower training/inference and higher memory. From this, it is to use n_estimators=200, as it provides good accuracy with reasonable speed.

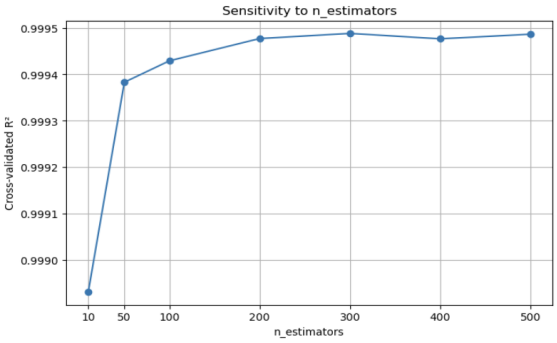

Figure 4: Sensitivity Analysis of Random Forest Model Showing the Impact of the Number of Estimators on Cross-Validated R² Scores

Table 5: Hyperparameter Comparison for performance gain, complexity cost and recommendation.

Hyperparameter	Performance Gain	Complexity Cost	Recommendation
max_depth	High gain up to 10, then flat	Deeper trees = higher variance risk	Set to 10–15
n_estimators	Sharp gain up to 100, then slow plateau	More trees = slower inference	Set to 200

Failure Analysis

Failure analysis was conducted by finding the 3 worst-performing instances from our EV demand model's prediction. These 3 instances demonstrate 3 different categories of failure.

Table 6: Examples of prediction failures from the EV charging demand regression model, illustrating large under- and over-prediction cases used for error analysis.

Index	Actual Demand Score	Predicted Demand Score	Error	Failure Type
11735	2729.0	2043.1	–685.9	Large under-prediction
81879	2705.6	2158.4	–547.3	Large under-prediction
11727	341.9	761.6	+419.7	Large over-prediction

Index 11735 is an example of an extreme outlier in our target, which is demonstrated by its large error rate. It is a large under-prediction of our model. This may have occurred since Random Forests cannot extrapolate beyond the range of target values seen in training- our training set may have had few or no examples above ~2500 units, the model caps predictions near the max observed value. Tree's average leaf node targets and without high demand examples, leaves never learn to predict more than 2200. Evidence of this is that the 99th percentile of y_train is about ~2400, meaning predicting 2700+ is impossible for our model.

Index 81879 is an example of a rare feature combination, essentially an underrepresented region. This instance combined features rarely seen together (e.g., high population + low charger density). Random forest relies on feature splits seen during training, so if certain state-city-season-infrastructure combinations are scarce, the model defaults to average behavior. An example of this would take place if this is an instance of a fast-growing suburban area's station, with new EV adoption. This may not match any historical patterns.

Index 11727 is an example of an extreme of an instance where nonlinear interactions are not captured, causing a large over prediction. The low-demand case (341.9) is over-predicted, indicating the model missed a suppressor effect. Our Random Forest model can model interactions that appear frequently in training data, unlike this instance. For example, a high income + winter seasons might reduce EV usage but if this interaction is rare, the model treats them independently and produces an overestimated demand. Tree depth limits may also prevent deep interactions from being captured.

Future Improvements

The model's failed predictions highlight some future improvements that will be taken into consideration for the next EV Demand model. For our first feature failure demonstrating an extreme outlier from an over prediction, extrapolation limits must be considered. Models with linear leave or neural networks like Gradient Boosting (XGBoost/LightGBM) should be tested as they may produce more complex models. Additionally, adding target transformations (log) can compress feature scales, soliciting limit improvements. SMOTE techniques can be implemented to adjust the class distribution of high demand samples. Rare feature combinations as demonstrated by the second failed prediction can be improved by implementing more feature engineering to capture interactions between features. For example, a feature of population x charger_density can capture this interaction. Clustering regions (e.g., through k-means for geographical features) can be used to generalize different regions. The last feature failure represented over prediction, which may be remedied by improving missed non-linear interactions. To improve this failure type, an increase in max_depth or min_samples_leaf in the Random Forest model should be considered. There should be caution in this implementation because it may result in overfitting.

6. Unsupervised Learning

6.1 Methods Description

Our strategy is designed as a two-stage workflow to derive distinct, actional insight both at micro and macro levels to determine the investment plan for building additional EV stations for both station and regional design.

Unsupervised Learning Workflow and Methodology

Table 7: Unsupervised learning workflow and methodology summarizing station-level and region-level clustering

Stage	Goal	Features	Method	Output
Station Archetype Identification (Micro level)	To identify functional charger station design	1.Charger (Counts total L1, L2, DCFC) 2. Categorical_station (Pricing, Facility Type)	K-means DBSCAN	4 Station Archetypes To determine what type of chargers need to build
Regional Investment Prioritization (Macro level)	Find any US states that prioritize investment by market demand and infrastructure strain	1.Adoption_rato(EVs/Capita) 2.Instrastructure_Balance_ratio (Station/EVs)	K-means DBSCAN	4 Investment tier list to determined which state need priority

Our analysis K-means clustering to establish four archetypes and tier lists, validated by DBSCAN to identify the market outliers. The main finding will be four distinct station archetypes, and four regional investment tier lists required for investment planning for efficient adoption.

6.2. Station archetype (Micro-level clustering)

Feature representation contributed significantly to our model choice. Initially, we attempted to cluster stations using raw charger counts (EV_L1_NUM, EV_L2_NUM, EV_DC_FAST_NUM) that led to clusters where the largest station (highest total charger count) was always separated first, regardless of the mix. For example, a station with 50 L2 chargers grouped separated from 5 DC fast charger stations, despite the L2 station functional profile being like that of a smaller L2 station. This resulted in K=2 where one massive cluster L1/L2 and one DCFC outlier.

Due to this, we can confirm the feature engineering and representation were needed in this challenge. Since our dataset has both numerical and categorical variables, we cannot directly apply the clustering method right away. We create a few new features; one contains charger types as numerical variables using StandardScaler or RobustScaler on chargers count, and one contains the EV pricing when charging at the station and the type of facility that station is located as a categorical variable using One-Hot Encoding. This ensures that the cluster will be based on the structural model such as Free-L1 charger or Paid-DCFC, not just the size of the data.

Next, we apply a different cluster algorithm to identify the optimal cluster and visualize those clusters. K-mean was the primary method for our station clustering analysis because the goal of the project is to identify the demand of EV stations by partitioning the data into fixed groups. The result from the clustering provides interpretable values to determine the optimal cluster for our analysis. The Elbow method was being used to calculate the distance between clusters, then we used Silhouette scores and Davie-Bouldin Index to evaluate the model results. The hyperparameter K was fine-tuned by iterating from 2 to 10, evaluated by each method mentioned above to select the optimal K, balancing by high separation score with efficiency.

Figure 5: Station-Level Clustering Evaluation Using Elbow, Silhouette, and Davies–Bouldin Methods to Determine the Optimal Number of Clusters

As a result, the elbow method provided a K=4 within range of efficiency, validating that the marginal gain beyond this point is minimal. The silhouette gives a global maximum at K=4 with the score of ~0.85, indicating the highest degrees of cluster quality where four clusters are the most compact and well-separated. While Davie Bouldin offers K=3 with the greatest separation with a score of ~0.20. However, K=4 was selected as the result necessary for actionable insight due to the close score index between those 2. DBSCAN was used not for primary results, but as a robustness against the K-means output. If the data is truly dense, the DBSCAN would isolate the same few clusters as the K-mean.

6.3. Regional Investment (Macro-level)

As for our regional analysis, we also resolve the challenge with the data by applying the same strategy we used from the station archetype. We fill population data with population mean, and 0 with EV registration count to fill any missing data. Those 2 data will be grouped by the state_name aggregated by the population and the EV registration count to compare every state. Two features were added; Adoption_ratio to determine EV adoption rate in that state, and the Infra_Balance_Ratio to see if we have enough infrastructure to meet the demand.

Our evaluation strategy will be the same with station archetype by using K-mean as a primary method to cluster the regional demand. We fine-tune hyperparameters by iterating K from 2 to 11 to find the optimal K and then evaluate the model performance using the Elbow method, Silhouette score and Davie-Bouldin Index will be used as well to evaluate the model performance. As for DBSCAN, it will be used as a secondary method, to detect any outliers in the model.

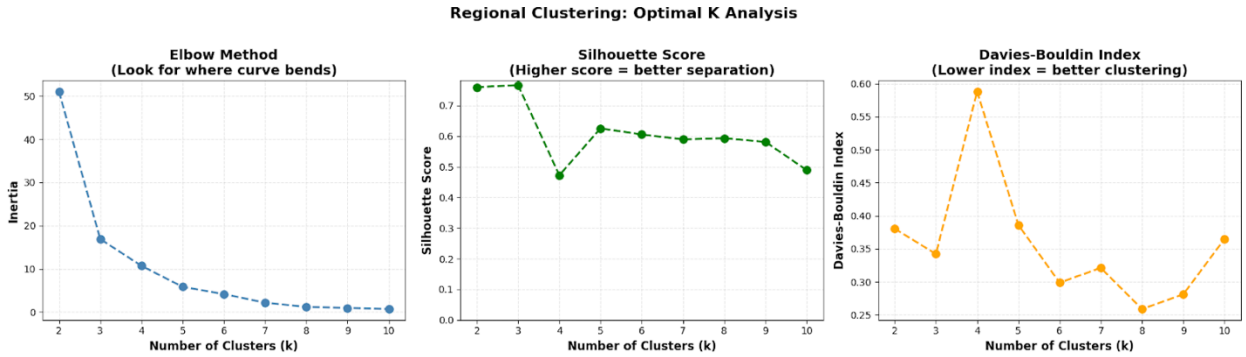

Figure 6: Region-Level Clustering Evaluation Using Elbow, Silhouette, and Davies–Bouldin Methods to Determine the Optimal Number of Clusters

For our regional analysis, the elbow method provided a K=3, shows the reduction in inertia in the plot after K=3, confirming that adding more clusters would be unnecessary since it dilutes the policy-driven interpretability. Silhouette gives a ~0.80 score at K=3, indicating that the states within these three clusters are statistically well separated from states in other clusters. However, the Davie-Boldin Index gave a different score of ~0.60. While the Davie-Boldin Index score was unexpected, we believe that K=3 will be the most optimal cluster for our analysis due to project goals and two out of 3 evaluation methods gave an optimal K of 3. DBSCAN was also used as a secondary method to identify any outlier of the data.

Table 8: Summary of clustering model performance and interpretation for Region-level analysis

Model	Optimal K	Silhouette Score	DBSCAN	Interpretation
Station K-Means	4	~0.85	-	4 archetypes: Fast-hub, workplace, community, mixed-use stations
Regional K-Means	3	~0.80	-	3 segments: Mature, High Concentration, Slow markets
Station DBSCAN	-	-	323 clusters + ~3.3 % noise	Confirms dense station groupings; isolates sparse/anomalous sites

6.4 Analysis Evaluation

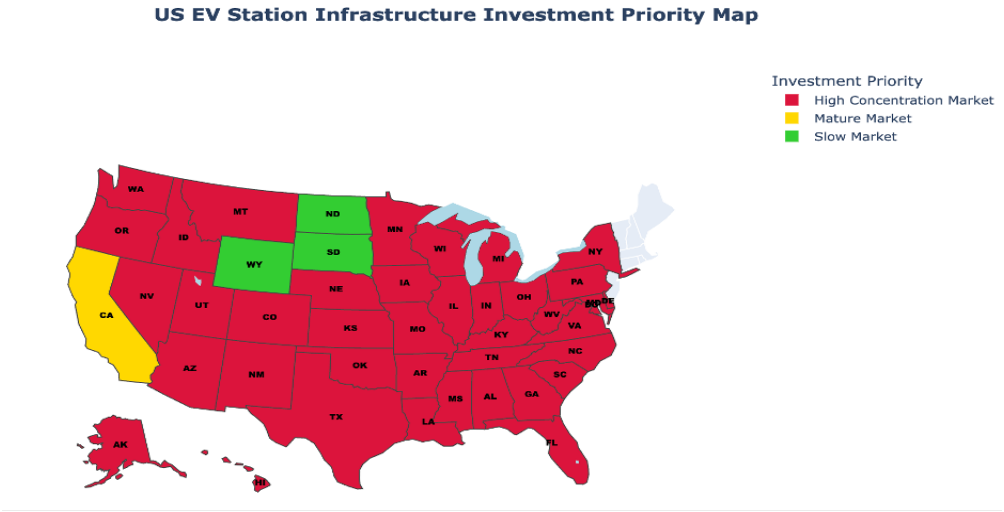

Figure 8: U.S. EV Station Infrastructure Investment Priority Map Categorizing States into Mature, High-Concentration, and Slow-Market Segments Based on Clustering Analysis

The evaluation is based on the final synthesis of micro and macro level. By cross-tabulate the station analysis and regional analysis, we gain actionable insights for targeted investment strategy. We found that:

1. Region_Cluster 0 is a mature market contributing 2,932 out of 13,219 total stations across the states. Within this region Station_cluster 0 (High-Volume L2 charger station) has an overwhelming number, indicating the market is driven by commercial demand and private investment that optimizes infrastructure balance.
2. Region_Cluster 0 represents a high concentrate market, the majority in California and Virginia accounted for 10273 total stations. Station_cluster 0 shows this market has a high number of High-Volume L2 charging stations despite only California and Virginia. They also have the most DCFC charging stations compared to other states.
3. Regional_cluster 2 shows that the market is low investment. Only 14 High-Volume L2 charging stations, likely indicating that it's a rural or underfunded area. Those areas are North Dakota, Wyoming, and South Dakota. To develop these states, foundation might be needed.
4. The High Concentrate market indicates that High-volume L2 charging stations are the backbone of station deployment due commercial demand and private investment.

6.5 DBSCAN Error Analysis

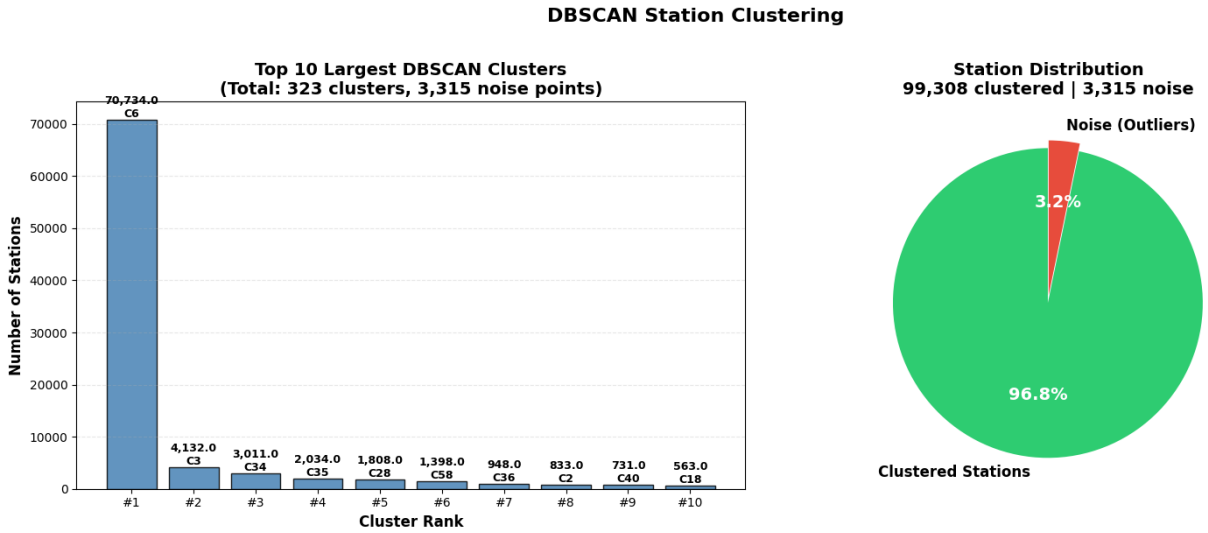

Figure 8: DBSCAN Station Clustering Results Showing Largest Clusters and Proportion of Outlier Stations Across the U.S.

While the K-means performed well in finding station cluster groups, the DBSCAN model failed to find these four clusters.Using the epsilon of 1 and min sample of 5, the DBSCAN model yielded above 300 clusters, demonstrating that the data is extremely fragmented and lacks uniform density. The failure might be in the feature representation due to the reliance on the charger counts rather than the energy or power consumption. Due to the high amount of missing data in energy_kw and power_kwh, those 2 features were not included in the model. To address this problem in the future, we must integrate completed data of those two features into the clustering to differentiate the efficiency of different clusters, improving the cost-benefit analysis of the archetype.

In the regional clustering, DBSCAN was run with epsilon of 0.3 and min samples of 5. We identified one large cluster containing over 90% of the state while the remaining 10 states including California and Virginia with high EVs to Station ratio as noise (outliers). This failure may come from data structure due to the majority of states form one massive cluster, providing no naturally spare, high-density group for K=3 to be identified without forcing partition.

6.6 Sensitivity Analysis

In the station analysis, K=4 was found to be highly sensitive to the scaling of the charger type features. We run the model with only raw, unscaled chargers, resulting in the optimal cluster at K=2, with 1 cluster having all the DCFC stations, and one cluster with massive L1 and L2 stations. The optimal K=4 allows the separation between High-Volume L1 and L2 on a standard scaling with the inclusion of EV pricing and facility type. Without it, the business model will be indistinguishable.

In the regional analysis, the result was extremely sensitive to the Infra_Balance_Ratio. We found out that if the model only scaled Adoption_Ratio and surrogate the station count, our result would be a geographic clustering which are East Coast, West Coast and Central. By introducing Infra_Balance_Ratio as a relative measure; If a state has few EVs count but few stations, it would be penalized, while states with high EVs count and stations would be rewarded. The optimal K=3 resulting in the classification of California into Critical Need clusters despite having high EVs and station count, confirms the sensitivity of Infra_Balance_ratio as a primary goal of policy insight. Without this, the analysis fails to identify the true infrastructure gaps relative to demand.

6.7 Trade-offs:

Depth for Simplicity

The primary tradeoff for station clustering is the model depth for the simplicity. The model uses extensive feature engineering such as scaled charger counts, incorporating discrete data like ev pricing and facility type to achieve four distinct archetypes to provide detail for execution. However, we sacrifice the simplicity of the model. We can't really tell why Cluster 0 is more important than cluster 3 without context of the regional cluster's demand data. The model is also highly sensitive to the initial feature set as shown in the sensitive analysis. We cannot justify why we should build a station in that region.

Interpretability for Granularity

The regional model's trade-off is interpretability for granularity. The model used two variables ratio Adoption_Ratio and Infra_Balan_Ratio to immediately understand which region is highly demanded but lacking in infrastructure. However, the trade-off is that the model cannot tell the operation in detail. Meaning, even if it correctly flags a stage like California as Highly Need, it does not provide any information on what charger's is needed. This makes us unable to determine the correct type of infrastructure needed within a state.

7. Discussion

7.1 Insights and Learning from Part A:

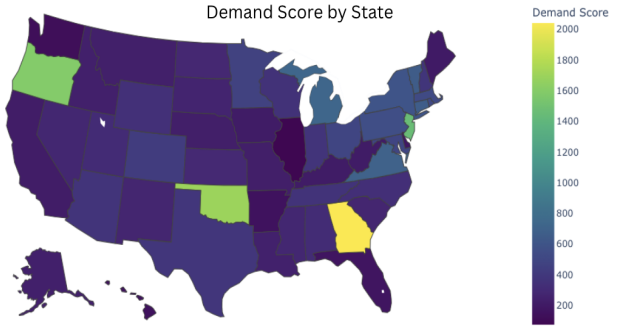

Figure 9: Geographic distribution of predicted electric vehicle (EV) charging demand scores across the U.S.

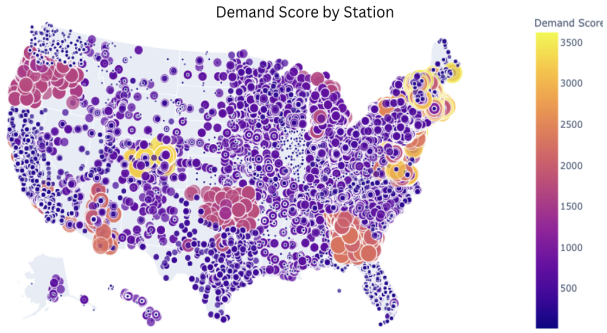

Figure 10: Station-level spatial distribution of EV charging demand across the U.S.

7.1 Insights and Learning from Part A

The EV demand supervised learning model taught us that predicting EV charging demand is a highly non-linear problem that benefits significantly from ensemble methods like Random Forest and Gradient Boosting over simple linear models like Ridge Regression. While Ridge Regression offered a reasonable baseline (R^2 ~0.78), it failed to capture the complex interactions between features, which the ensemble models handled exceptionally well ($R^2 > 0.99$). Feature engineering, preprocessing, and hyperparameter tuning critically affect model performance. Particularly, imputing missing data with informed strategies (e.g., using state-level averages or Random Forests) preserved data integrity, and tuning parameters like max_depth and n_estimators directly translated to performance gains. Moreover, tools like SHAP and ablation studies were instrumental in interpreting the model, highlighting that usage behavior (e.g., charge duration, total sessions) and demographics (like population) are more predictive of demand than infrastructure variables alone.

7.2.1 Surprising Findings

A surprising feature of our model was how a state's population was of critical importance. Removing this one feature increased RMSE by over 540%, indicating raw population data acts as a proxy for many hidden variables (urbanization, EV adoption, traffic volume), possibly masking the signal of more nuanced features. It was also surprising how sensitive the model was to certain hyperparameters, especially max_depth. Once tuned, the model's accuracy increased significantly with diminishing returns beyond depth 15, highlighting a sweet spot for complexity vs. generalization. Lastly, failure analysis revealed that Random Forest models struggle with extrapolation and rare feature combinations, which was a valuable insight into the model's limitations and potential biases.

7.3 Challenges and Responses

A huge challenge for our model began with handling missing categorical and numerical values, without introducing bias. We addressed this by using a Random Forest Regressor to impute categorical variables based on relevant station and state-level features, and state-level averages for numerical features. When creating a model, we had to be mindful of the high R^2 values, to ensure the models weren't overfitting. We used cross-validation, learning curves, and sensitivity analyses to confirm that performance stabilized and generalization held across different subsets. After creating our model, understanding which features mattered and how they influenced predictions was initially opaque. Implementing SHAP values allowed us to unpack the black-box nature of the ensemble models and gain actionable insights.

7.4 Future Extensions

We strongly believe future improvements on the model would begin with the population feature. Since population is so dominant, normalizing features like EVs per capita, sessions per capita could lead to more insightful, better models, resulting in accurate policy or regional planning. Additionally, handling imbalanced data by apply SMOTE or similar techniques could be deployed for high-demand observations. This would help models better learn from these underrepresented but crucial data points. Lastly, integrating the trained model into a dashboard or API service for urban planners or utility companies to visualize high-demand zones or predict future demand growth, would be an ideal next step.

7.2 Insights and Learning from Part B

Fortunately, we have gained multiple significant insights on how EV charging infrastructure differs across America from our unsupervised learning model's analysis. The clustering technique allowed us to identify patterns that could not be simply identified in the raw data. At the station level of our data, we discovered four major archetypes, which were high-capacity fast hubs, workplace, community and mixed-use stations. While regional clustering grouped the American states into mature, balanced, and slow growth markets based on adoption and number of charging stations available. As a matter of fact, these results emphasized how unsupervised techniques can guide policymakers and charging providers to group regions and prioritize their investments for required regions accurately.

7.2.1 Surprising Findings

One of our most surprising findings was how strong socioeconomic factors like median income and population density correlated with EV charging capacity based on the data. We were also surprised with the fact that the adoption of EV vehicles did not always correspond with the availability of charging stations. Even though California and other coastal states had a high adoption of EV, multiple Midwestern regions showed more balanced availability of charging stations to EV adoption relative to demand. Furthermore, the DBSCAN of the model showed approximately 3 percent of charging stations as isolated outliers, often found along small, populated areas, or interstate highways. These anomalies point out uneven spatial distribution and highlight the necessity for more unbiased access to charging stations.

7.3 Challenges and Responses

During our initial analysis, the main challenges were computational because of the size of the data. The data was spread across multiple dimensional files and to interpret such data files we had to join and create a unified dataset. The dataset's large size as mentioned above, and its high dimensionality made DBSCAN clustering very time-consuming and sensitive to scaling. We overcame this by applying Truncated SVD for dimensionality reduction, testing smaller random samples for parameter tuning, and standardizing features through station feature preprocessing. Furthermore, another challenge was interpreting numerical cluster labels into a real-world context. We resolved this through mapping the clusters to geographic regions, and using visualizations like heatmaps and the U.S. EV Infrastructure Investment Priority Map to show our results into actionable insights.

7.4 Future Extensions

If we were provided with additional time and resources, we would do our best to extend our analysis in several ways. Firstly, we would include temporal trends to see how regional clusters develop over both months and years. Secondly, we would want to integrate more granular geospatial and energy grid data to capture factors like traffic volume, charging reliability, and power capacity. Thirdly, we would like to develop a hybrid framework, which predicts where there's a rise in demand for new hot zones. Fourthly, if we were given more time, we would possibly replace a dataset to ensure there's not many missing values to ensure our analysis is accurate. Furthermore, we believe creating an interactive user-friendly dashboard would help policymakers and especially energy companies explore results and look at infrastructure development to meet the upcoming demand of EV vehicles. Overall, we strongly believe additional time and resources will help our project growth further for the future of EV vehicles in America.

8. Ethical Considerations

8.1 Part A

When applying supervised learning solutions, especially in contexts like predicting EV charging demand, there are several ethical issues that may arise. To begin, models may reinforce or amplify existing biases in the data. For example, if data reflects historic underinvestment in certain neighborhoods or underrepresented populations (e.g., rural areas, lower-income regions), the model may predict lower demand in those areas, leading to fewer resources being allocated there in the future. This issue could be addressed by data auditing. This would take place by regularly auditing training data for geographic, demographic, or socioeconomic biases. Secondly, over reliance on the model may lead to the ethical issue of decision-makers treating the model's outputs as objective truth creating a blind trust in the models. This could be addressed by uncertainty quantifications- clearly communicating confidence intervals or prediction uncertainty to decision. Using tools like SHAP to explain model predictions and support in ensuring transparency A last major ethical concern may arise from a model that optimizes purely for demand, this may encourage infrastructure expansion in ways that are unsustainable, building more stations in car-dependent areas without regard for overall carbon emissions or urban planning. This could be addressed by including sustainability metrics like factors in the environmental or urban planning goals in model evaluation.

8.2 Part B

The concern for station analysis is that the model clearly identifies that DC Fast Charging Hub is the most viable archetype while L1 is the least. If funding is tied to maximizing utilization or return on investment, L1 and L2 will be ignored. This will worsen the disparity between low-income drivers who rely on overnight or extend L1/L2 charging. To address this, we must include public funding for L1/L2 archetype, priority access and equity over commercial viability. The primary concern for our regional analysis is by identifying states that in Critical Needed (cluster 0), the model justifies investing limited funds to already wealthy EV markets. These markets marked a high immediate return on investment. To prevent that, policy makers need to allocate a budget to Emerging markets (Cluster 2) to guarantee the foundation infrastructure is built for those states, ensuring equitable access across all states.

9. Statement of Work

This project was a collaborative effort by **Team 23**, consisting of **Dharshana (DS)**, **Zara (ZM)**, and **Duy (DH)**. Each team member made distinct and complementary contributions to the successful completion of the project.

Table 9: Team 23 Statement of work detailing individual contributions

Dharshana Somasunderam (DS)	**Zara Masood (ZM)**	**Duy Ho (DH)**
Data collection and integration Data preprocessing and cleaning; Feature engineering; FIPS geographic standardization. Exploratory data analysis and visualization; Supervised data pipeline support; Report writing and editing, and GitHub Pipeline.	Supervised learning model development (Ridge Regression, Random Forest, Gradient Boosting). Feature selection and tuning; Cross-validation and evaluation; Sensitivity and error analysis. Model interpretation and visualization; Ethical Consideration; Report writing.	Unsupervised learning analysis (K-Means, DBSCAN, Hierarchical Clustering). Dimensionality reduction Clustering evaluation (Elbow, Silhouette, Davies-Bouldin). Regional and station pattern identification; Map visualization and interpretation. Ethical analysis and report writing.

Appendix A – Data Sources

EV WATTS Public Dataset
Source: Energetics and the U.S. Department of Energy, available at https://livewire.energy.gov/ds/evwatts/evwatts.public.

Alternative Fueling Stations (AFS) Dataset
Source: U.S. Department of Transportation, Bureau of Transportation Statistics, https://geodata.bts.gov.

AFDC Vehicle Registration Data
Source: U.S. Department of Energy's Alternative Fuels Data Center (AFDC), https://afdc.energy.gov/vehicle-registration.

U.S. Census Population and Income Data
Source: U.S. Census Bureau API, https://api.census.gov/data.

State–County FIPS Reference Data
Source: U.S. Department of Transportation FIPS Reference Table, https://data.transportation.gov/Railroads/State-County-and-City-FIPS-Reference-Table/eek5-pv8d.

Appendix B – Features Engineering

Final Features for Modeling
The final dataset used for modeling contained approximately 250,000 station–month records and included the following categories of features:

Table 10: Supervised Learning Features

Category	Example Features
Temporal	start_time_hour, day, season
Session-Level	charge_level, power_rating, energy_kwh, charge_duration
Station-Level	num_ports, ev_dc_fast_num, ev_level2_evse_num, connector_type, facility_type, pricing, venue
Regional / Socioeconomic	state_name, EV_registrations, adoption_rate_per_state, median_income, population_density
Derived Metrics	connector_diversity, total_ports, sessions_per_day

Unsupervised Learning Features

Table 11: Unsupervised Learning Features

Category	Feature Examples
Station-Level Numeric	EV_LEVEL1_EVSE_NUM, EV_LEVEL2_EVSE_NUM, EV_DC_FAST_NUM, num_ports

Categorical Attributes	EV_PRICING, FACILITY_TYPE, EV_NETWORK, connector_type
Regional Features	population, median_income, EV_Reg_Count_Sum, Adoption_Ratio, Infra_Balance_Ratio
Derived Metrics	total_ports, connector_diversity, urbanization_index
Normalized / Encoded Inputs	Standardized numerical variables and one-hot or label-encoded categorical fields

All categorical and numeric features were scaled and encoded appropriately depending on the model family (e.g., standardized features for K-Means and DBSCAN).

Appendix C – Supervised Learning SHAP Analysis

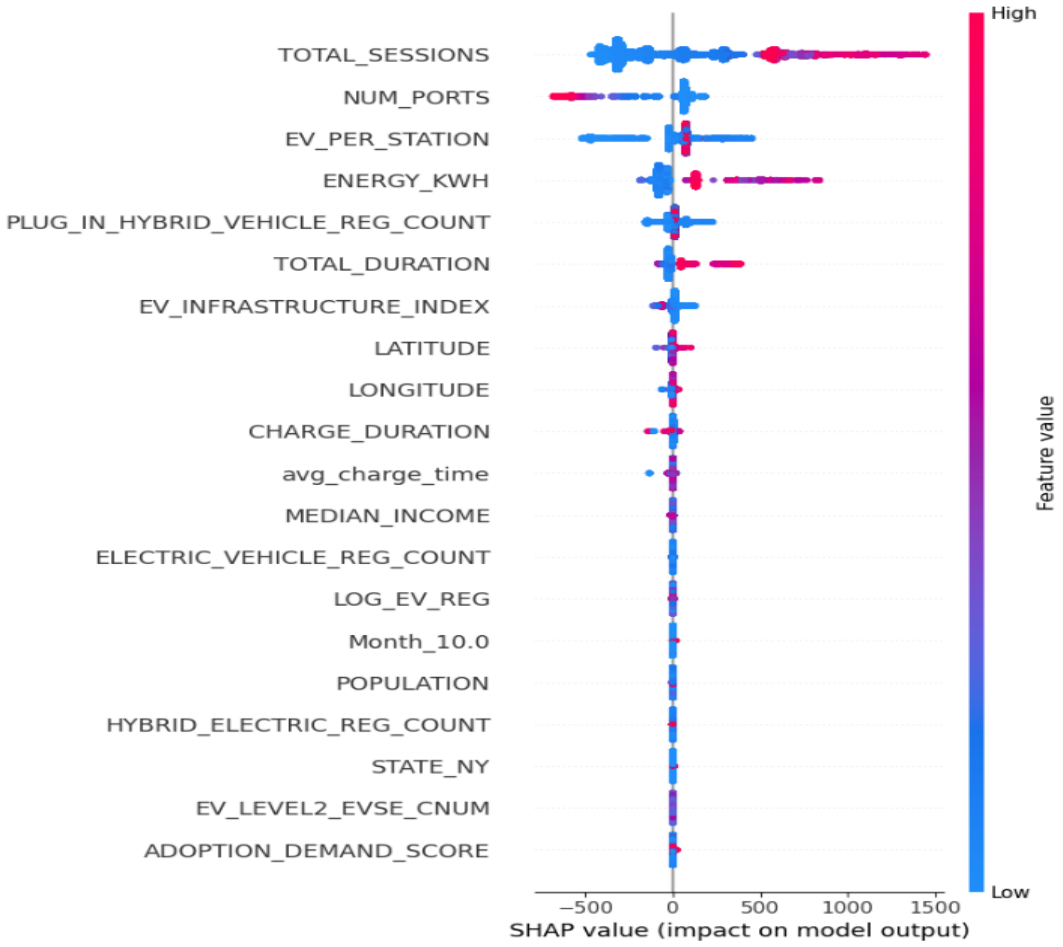

Figure 1: SHAP Summary Plot Showing Feature Importance and Direction of Impact on Predicted EV Charging Deman

Appendix D – Model Evaluation Metrics

Table 12: Model evaluation metrics with purpose

Model Type	Metric	Description	Purpose
Supervised (Regression)	RMSE, MAE, R^2	Root Mean Squared Error (RMSE) measures average prediction deviation; R^2 quantifies explained variance.	Calculate charging demand forecasting accuracy.
Supervised (Classification)	Accuracy, F1-Score, ROC-AUC	F1 balances precision and recall; ROC-AUC captures trade-offs.	Calculate high vs. low demand classification.
Unsupervised	Silhouette Score	Measures cohesion and separation (higher = better).	Assess overall cluster compactness.
	Davies–Bouldin Index	Measures intra- vs. inter-cluster distance (lower = better).	Confirm cluster stability.
	Noise Ratio (for DBSCAN)	% of points classified as noise.	Check data sparsity and outlier presence.

Appendix E – Clustering Results Summary

Table 13: Clustering results summary

Cluster	Description	Key Attributes	Example States
Critical Shortfall	High EV adoption, low charger availability	High Adoption_Ratio, low Infra_Balance_Ratio	California
Moderate Need	Balanced adoption and infrastructure	Mid-range ratios	Illinois, Texas
Foundational Need	Low adoption, developing infrastructure	Low Adoption_Ratio and Station_Count	Wyoming, North Dakota
High Concentration Market	Moderate EV penetration	Low stations	Texas, Washington, Utah
Slow Market	Adoption lags	Very Few stations, low adoption	WY, SD, ND

Appendix F – References

1. U.S. Department of Energy (DOE) – EV WATTS Public Dataset.
Energetics Inc., 2019–2023.
https://livewire.energy.gov/ds/evwatts/evwatts.public

2. U.S. Department of Transportation – Alternative Fueling Stations Dataset.
Bureau of Transportation Statistics (BTS), 2023.
https://geodata.bts.gov/datasets/usdot::alternative-fueling-stations/explore

3. U.S. Department of Energy – Alternative Fuels Data Center (AFDC) Vehicle Registration Data.
Office of Energy Efficiency & Renewable Energy (EERE), 2023.
https://afdc.energy.gov/vehicle-registration

4. U.S. Census Bureau – American Community Survey (ACS) 5-Year Estimates (Population & Income).
https://api.census.gov/data

5. U.S. Department of Transportation – State, County, and City FIPS Reference Table.
https://data.transportation.gov/Railroads/State-County-and-City-FIPS-Reference-Table/eek5-pv8d

6. Sharma, A., et al. (2021).
Data-Driven Siting of Electric Vehicle Charging Stations Using Machine Learning.

7. National Renewable Energy Laboratory (NREL) (2016).
Commercial-Scale Battery Energy Storage and Charging Infrastructure Analysis.
Golden, CO: NREL.

8. Scikit-learn Developers (2024).
Scikit-learn: Machine Learning in Python, Version 1.5 Documentation.
https://scikit-learn.org/stable/documentation.html

9. Plotly Technologies Inc. (2024).
Plotly Express and Dash for Python – Interactive Visualization Tools.
https://plotly.com/python/

10. Pedregosa, F., et al. (2011).
Scikit-learn: Machine Learning in Python.
Journal of Machine Learning Research, 12, 2825–2830.
http://jmlr.org/papers/v12/pedregosa11a.html