A COMPREHENSIVE ANALYSIS OF FACTORS INFLUENCING THE PATH TO ZERO-EMISSION VEHICLES SALES TARGET OF CALIFORNIA

University of Exeter
BEMM466 – Business Project

Trang Quynh Nguyen
730017134
MSc Business Analytics

Project Advidor: Dr. Stuart So

September 2024

Table of Contents

Github link: https://github.com/trangng99/businessproject/tree/main

Glossary

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BEVs	Battery Electric Vehicles
CVRP	Clean Vehicle Rebate Project
EVs	Electric Vehicles
EDA	Exploratory Data Analysis
k-NN Imputation	k-Nearest Neighbor Imputation
MAE	Mean Absolute Error
OLS	Ordinary Least Squares
PHEVs	Plug-in hybrid Electric Vehicles
RMSE	Root Mean Squared Error
The U.S.	The United States
VIF	Variance Inflation Factor
ZEVs	Zero-emission Vehicles

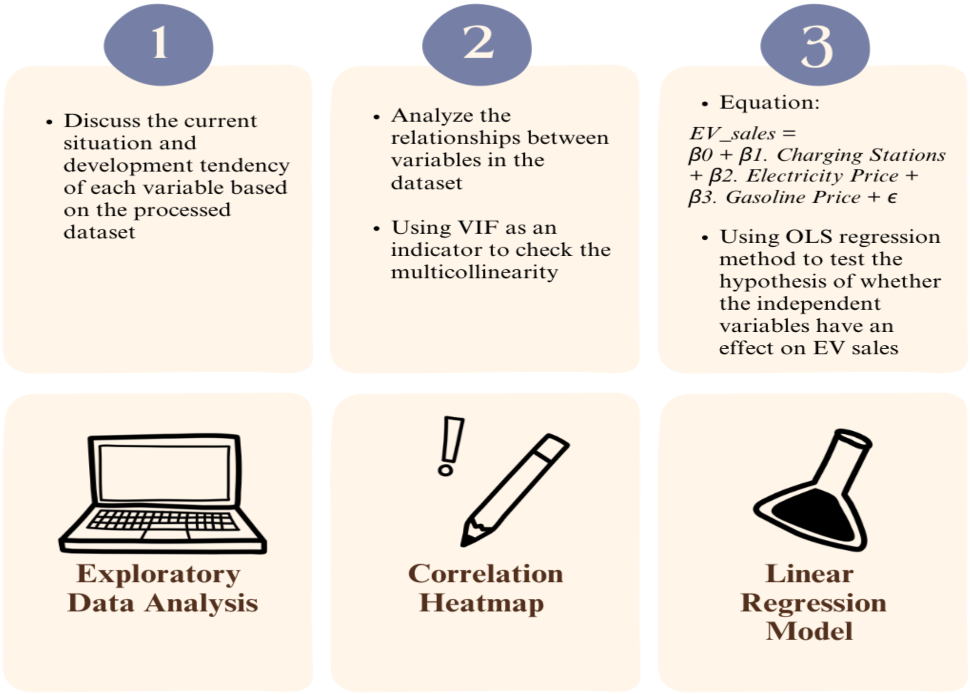
Executive Summary

Purpose

In order to achieve the Paris Agreement to address the global climate-warming emergency, the U.S. government is making great efforts to implement commitments and policies for greenhouse gas emissions reduction. The goal of reducing emissions by 50-52% by 2030 compared to 2005 levels is proposed and the option of switching to BEVs is a feasible option for this goal (United Nations Climate Change).

California, the city leading the development of the BEV market and investing in the number of charging stations in the U.S., has set an ambitious goal of accelerating to 100% new zero-emission vehicle sales by 2035. Currently, they have successfully achieved 1.5 million EVs prior to targeted 2025, but recently, the BEV sales experienced a decline while its growth is critical in determining whether the state can accomplish its 2035 goal of prohibiting carbon-emitting new automobile sales. This project aims at providing a comprehensive understanding of the future trend of BEVs in California while analyzing external factors including *charging station counts, gasoline and electricity price*, affecting their EV adoption progress.

Methodology



1

- Discuss the current situation and development tendency of each variable based on the processed dataset

Exploratory Data Analysis

2

- Analyze the relationships between variables in the dataset
- Using VIF as an indicator to check the multicollinearity

Correlation Heatmap

3

- Equation:

$EV\_sales = \beta 0 + \beta 1.\ Charging\ Stations + \beta 2.\ Electricity\ Price + \beta 3.\ Gasoline\ Price + \epsilon$

- Using OLS regression method to test the hypothesis of whether the independent variables have an effect on EV sales

Linear Regression Model

To conduct research, the author has collected four different secondary data from government websites and filled in missing data values using the k-NN imputation method using Python. Following imputation, the data is distributed more clearly by quarter. The final merged table includes data from Q1 2010 to Q4 2023 for EV sales and external factors with 56 records each. Exploratory Data Analysis, Correlation Heatmap and Linear Regression Model will be the key approach to answer the research questions with qualitative insights policymakers.

Techniques based on data type and objective suggested by Komorowski, M. et al., (2016) will be the foundation for developing appropriate EDA contents. The Correlation Heatmap explores the potential relationships between the variables and detects multicollinearity guiding further analysis. Linear regression analysis, by fitting a line to data points makes better understanding how EV sales would vary if there are some changes to external factors. Hypothesis was tested to check the assumed relationships between independent variables and EV_sales using OLS regression method. Statistical indicators are key elements to evaluate the efficiency level of the forecasting model.

This project strictly complies with the Data Ethics Framework and General Data Protection Regulation (2018) for appropriate and responsible data use purposes.

Key findings

Based on previous empirical research, it is expected that all independent variables considered have a correlation with the EV sales. This applied to several studies in international studies as well. Inspired by them, the project has revealed some significant findings:

- Exploratory Data Analysis:

Through analyzing processed dataset, it indicates that BEVs accounted for the majority of total vehicle sales in California during the studied period and generally trended upward year-over-year. However, if we look closely at each quarter, EV sales showed signs of slowing down during the COVID-19 pandemic and have trended downward in recent quarters.

The number of public charging stations in California ranks first among the country with uneven distribution, in fact, mostly concentrated in large counties and scattered in rural areas. Gasoline and electricity prices are considered remarkably high due to the impact of the energy crisis.

- Correlation Coefficient Matrix

The Heatmap describes the relationships between the variables in the dataset. The results show that the variables mostly have positive correlations with each other, with only Gasoline Price and Charging Stations Counts having negative but insignificant relationships. In general, this unclear strong inverse relationship indicates that high gasoline prices might not necessarily increase as a result of a substantial rise in EV sales.

Outstandingly, Electricity Price has an extremely strong positive correlation with the Charging Stations Count, which might not cause multicollinearity based on all tested VIF values being below 5. However, some outliers may indicate significant urban centers or areas with high EV adoption rates.

Table 1: Relationship between each variable

	Number of Vehicles	**Charging Stations Count**	**Gasoline Price**	**Electricity Price**
Number of Vehicles	1	Moderately Strong	Moderately Strong	Extremely Strong
Charging Stations Count	Moderately Strong	1	Insignificantly Negative	Moderately Strong
Gasoline Price	Moderately Strong	Insignificantly Negative	1	Moderately Strong
Electricity Price	Extremely Strong	Moderately Strong	Moderately Strong	1

- Linear Regression Model:

Linear Regression Model, compared to other Ridge and Lasso regularization model, is the most optimal one when criticizing the R^2, RMSE and MAE.

When running OLS Regression with 56 observations, the R^2 was 0.77, however other indicators, especially P-value, did not meet the standard to reject the null hypothesis. After processing the outliers based on the analysis of the Actual vs Predicted Number of Vehicles simulation graph, to improve the model efficiency, 4 observations were removed and the results had a positive change. The statistical indicators were all analyzed to meet the requirements to reject the null hypothesis with an accuracy rate of $R^2 = 0.903$. This proves that the independent variables have a positive impact on EV sales in California.

The final regression model is:

$$\textbf{EV\_sales} = \mathbf{19540 + 5062.0151 * Charging\ Stations + 23880 * Electricity\ Price}$$
$$\mathbf{+ 2951.8923 * Gasoline\ Price} + \boldsymbol{\varepsilon}$$

Although the residuals do contain some outliers, overall they have a relatively symmetric distribution around zero and a right-skewed distribution; the model remains applicable to forecast the future EV sales.

Conclusion and Recommendations for Future Research

In conclusion, EV sales could possibly experience the upward trend if there is a proper adjustment from the authorities on charging stations counts, electricity and gasoline price due to their positive impact analyzed in the linear regression model. Although not considered in this project due to data limitation, financial incentives are certainly an important factor influencing EV sales, which has been confirmed in previous studies on the same topic.

Any changes in the price along with strong developments in technology, EVs will become more attractive to potential buyers to seek for alternatives. The study will also discuss in detail the limitations in the implementation process.

I. Introduction

The Paris Agreement, a legally binding international treaty on climate change that adopted by 196 parties in France, came into force in 2016. Its overarching goal is to hold "the increase in the global average temperature to below 2°C above pre-industrial levels" and pursue efforts "to limit the temperature increase to 1.5°C above pre-industrial levels (United Nations Climate Change).

Given the climate crisis and the need to restrict warming stated in The Paris Agreement, the United States (U.S.) aims to achieve net-zero greenhouse gas emissions by 2050 at the latest, with a commitment to significantly reduce net GHG emissions by 50-52%. To achieve the target, reducing emissions and fossil fuel use by hurriedly zero-emission vehicle (ZEV) deployment is a great solution and thanks to technology innovations, the historic investments in the Inflation Reduction Act, and additional investments made by manufacturers and throughout the battery supply chain, the U.S. transportation sector is rapidly shifting towards zero emission vehicles, which include battery electric vehicles (BEVs).

Electric vehicles (EVs) have the ability to reduce CO_2 emissions as electricity can be produced from renewable energy sources (Mersky et al., 2016). In this sense, EVs have become an alternative in the transport sector. Increasing the use of EVs not only helps reduce greenhouse gas emissions but also contributes to improving air quality, reducing dependence on fossil fuels and promoting green economic development. In addition, transportation electrification provides a major opportunity for economic prosperity and job creation as well as international competitiveness.

California, renowned for its progressive environmental laws in the United States, is actively pushing for the increased sales of EVs as a crucial part of its strategy. The EV industry is swiftly turning California into a frontrunner in the global exploration, development, production, and export of transportation-related goods, services, and technologies, shifting the industry's focus from Michigan to California. According to the Los Angeles County Economic Development Corporation (LAEDC), California has an unprecedented ecosystem of EV industry assets that has led the U.S. in the number of EVs and charging locations every year since 2016, accounting for 37% of registered light-duty EVs and 27% of EV charging locations by the end of 2022, according to new State Energy Data System (SEDS) estimates.

California continues to see record-breaking sales of ZEVs, driven by the state government's ambitious target: all new passenger cars, trucks, and SUVs sold in California will need to be zero-emission by 2035 under the 2022 Advanced Clean Cars II legislation. With numerous policies and incentives for consumers, Californians purchased approximately 450,000 new ZEVs in 2023, a 30% increase from 2022. ZEVs accounted for 25% of new vehicle sales, up from 20% in 2022, and California sold 1.5 million EVs two years ahead of its targeted sales milestone in 2025, according to the California Energy Commission. However, the market proportion of EV is basically flat, reported by the California New Car Dealers Association. The statewide climate plan claims that market share must reach 35% by 2026, which requires an annual sales growth rate of approximately 20%.

Despite leading in the U.S. market, EV sales in California have recently declined while its growth is critical in determining whether the state can accomplish its goal of prohibiting carbon-emitting new automobile sales by 2035. Various barriers have hindered the journey of selling EVs contributing to a sustainable economy. A better knowledge of the factors influencing EV adoption in California is crucial and required to meet the zero-emissions objective and remain a leader in EV uptake among U.S. states.

Therefore, effective models to understand and predict EV sales dynamics are required, which can assist California policymakers and government in developing better options and supporting people as they implement the plan to transition from selling only traditional gasoline cars to EVs by 2035. Significantly, the government must offer adequate charging stations to alleviate fear and raise awareness among automobile owners, therefore accurate sales volume forecasts for EVs is critical to supporting infrastructure development (Kumar et al., 2024) and energy efficiency.

As a result, the research questions are:

- ***Could this sales growth slowdown turn into a longer-term trend, reducing the overall EV adoption rate of not only California but also the U.S.?***
- ***Has the growth of EV sales in California been challenged by any external factors, including charging infrastructure, gasoline prices and electricity prices*?**

Particularly, the main purpose of this project is to provide (1) a comprehensive Exploratory Data Analysis to explore quarterly data from 2010 to 2023, identify trends in order to better understand its structure and qualities; and (2) the influence of external factors on EV sales in California through the application of linear regression model - a fundamental machine learning algorithm in the process of EV adoption to reach the 2035 target.

II. Literature Review

In this section, the overview of the studies of the effect of external factors to EV adoption over the past few years is presented. The chosen common factors are classified as *context factors*, comprising relatively Charging Stations Counts, Electricity Price and Gasoline Price, which have been applied in many studies.

The comprehensive incentives have been extensively studied, with particular focus on the Clean Vehicle Rebate Project (CVRP) in California, which was first launched in 2010 and is considered the most important initiative. Through buyer rebates and reduced greenhouse gas emissions, CVRP has played a significant role in the increased adoption of EVs in California; however, it is criticized for not providing equitable benefits to lower-income groups. An income cap was imposed, eliminating consumers with higher earnings, and consumers in families below a certain threshold became eligible for a larger rebate amount. Although this rebate has been improved by implementing specific income-based eligibility requirements from March 2016, CVRP remains perceived to deliver better benefits among high-income individuals (Arthur L. Ku and John D. Graham, 2022). Following the study of Soltani-Sobh et al., (2017), due to the lack of quarterly data as well as there is no available electric vehicles price data over quarters in monetary value, therefore this project will not consider the impact of incentive to EV sales in California.

There is mixed evidence within the literature in terms of the effect of gasoline prices on EVs adoption. Gasoline prices, affecting the comparative cost advantage of EVs over internal combustion engine vehicles. The studies from Gallagher and Muehlegger (2011) or Qiu et al. (2019) illustrates that there is a positive correlation between gasoline prices and EV adoption in China. There is a likelihood that consumers are more inclined to transition to EVs to minimize the higher operational costs of gasoline-powered automobiles. This aligns with other previous studies from Wee et al. (2018), Javid and Nejat (2017) and many more. On the other hand, in the study of Adepetu et al., (2016), it is claimed that there is no clear evidence for the positively correlated result of increasing PEV sales from the growth of gasoline prices.

Soltani-Sobh et al., (2016) had found out using a cross-sectional/time-series (panel) analysis that there is a proper sign that lower utilization of EVs in areas with higher electricity prices. Javid & Nejat, (2017) concluded that electricity price does not have any effect towards the EV adoption as there is no considerably diverse electricity price compared to gas price within California in 2012. Based on the research of Zhenzhen Jiang & Xinwei Gao in 2023, changes in gasoline and electricity prices will have a significant impact on electric car sales in low- and middle-income areas, whereas electric vehicle sales in high-income cities will be unaffected. Both influencing the consumer preference, gasoline prices have a more significant impact on consumer decisions to buy EVs than electricity prices, according to a study by Bushnell et al., (2022), which concluded that a change in gasoline prices has four to six times the effect on EV demand as a similar change in electricity prices.

Several studies have highlighted the heavy reliance of EVs on the development and availability of charging infrastructures, which is a crucial factor in consumer acceptance of alternative fuel vehicles. Ghamami et al., (2014), Cherchi (2017), Coffman et al., (2017) and Tang et al., (2019) stress that this factor could enhance purchase intention and support the spread adoption of EVs, while Yang et al. (2016) also states that the low availability of charging stations creates such a barrier. In addition, Sierzchula et al., (2014) find out that along with financial incentives, the provision of a robust and accessible charging network increases the possibility that consumers would choose EVs. The importance of investments in charging infrastructure aiming at fostering EV market growth is implied by this relationship. According to Mersky, Sprei, Samaras, and Qian (2016), EV charging infrastructure is the primary indicator of EV adoption at the regional and municipal levels in Norway. However, in California, where the number of charging stations is comparatively higher than rest of the United States, public EV charging stations are not equitably distributed with respect to race and income levels (Hsu & Fingerman, 2021) thereby, widening the gap in access and adding to the sociodemographic and economic disparities in the region.

The reviewed literature underscores the significant role of charging infrastructure, gasoline prices, and electricity prices in promoting the adoption of EVs in California, which is the inspiration for this project. While previous studies have focused on data from across the U.S. or worldwide, this project will focus on California, using the latest real-world data with the longest time frame to most accurately reflect EV sales during the period that is considered experiencing the most active market in the state.

The project methodology allows model extension to any other region having a homological dataset. It also intends to contribute to a more comprehensive understanding of how forecasting models deliver actionable insights for policymakers and authorities so as to optimize their strategies in greater EV usage stimulation.

III. Data and Methodology

1. Data Collection and Preprocessing

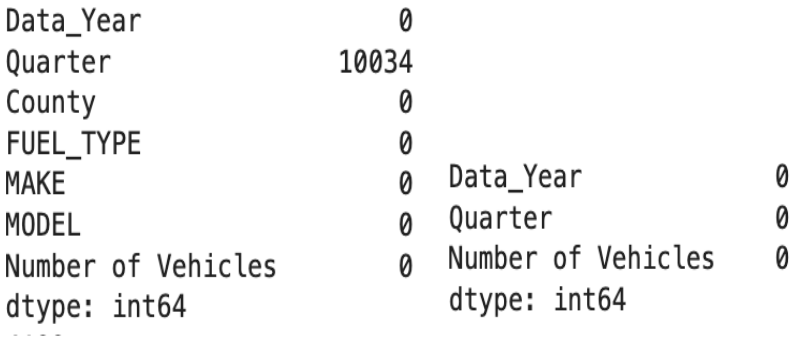
For this project, the secondary data has been gathered from four different public sources from U.S. government websites and eventually merged the selected columns into one usable table from Quarter 1 2010 to Quarter 4 2023. Collecting original data on EVs from government websites can be transparent and up-to-date. It is preferable to conduct market research on EVs using secondary data to analyze consumer behavior, environmental impact, and economic matters. On the other hand, due to the geographical limitations and jurisdiction, the author was unable to directly collect primary data on EVs in California for this research.

All original datasets contain introductory information that is unrelated to the project, these will be eliminated to avoid confusion. The files uploaded on *Github* have been processed and specifically explained in the README section. The final merged table *(Link)* contains six columns with total 56 rows as mentioned below:

(1) *Year*: Year ranging from 2010 to 2023
(2) *Quarter:* There will be four quarters for each year
(3) *Number of Electric Vehicles sold*: gathered from California Energy Commission website, offering total number of BEVs, Hydrogen Vehicles and Plug-in hybrid Electric Vehicles (PHEVs) sales across the period. As California is targeting at promoting faster EV adoption

and minimize the emission, this project will filter only battery EV sales and exclude Hydrogen Vehicles and PHEVs as they are still having some tailpipe emissions. And in the process of data collection, BEVs accounted for the majority of vehicle sales compared to the other two types throughout the 13-year time period, which will be discussed more detailed in the Exploratory Data Analysis section). In the final table, only Data_Year, Quarter and Number of Vehicles will be used for the merging step at the end.

It is outstanding that data from Quarter 1 2010 until Quarter 4 2020 have 10,034 missing values for the "Quarter" column, which will be solved by **k-Nearest Neighbor (kNN) Imputation** method to transform data. This supervised-learning method estimates missing values in a dataset by considering the values of the closest data points, determined by a distance. n_neighbors = 5 is used to implement based on the existed values (following the pattern from available data from year 2021 to 2023) from the original dataset to forecast the missing values. After implementation, there would be no missing values for the "Quarter" column (as shown in Figure 1).



```
Data_Year                 0
Quarter               10034
County                    0
FUEL_TYPE                 0
MAKE                      0   Data_Year              0
MODEL                     0   Quarter                0
Number of Vehicles        0   Number of Vehicles     0
dtype: int64                  dtype: int64
```

Figure 1: "Quarter" column before and after processing

After imputation, the data points in Figure 2 are distributed across all four quarters, however there is a noticeable horizontal trend, with several points clustered around quarters 2 and 3. This allows the original data distribution quite accurately after the process of filling in the missing values using the imputation procedure. The imputed data seemed to fit the entire dataset, however there were still some different patterns or outliers detected in Quarter 4. Quarter 2 to 3 had a higher density of points, suggesting that generally the imputed values were reliable and consistent with the overall data pattern.

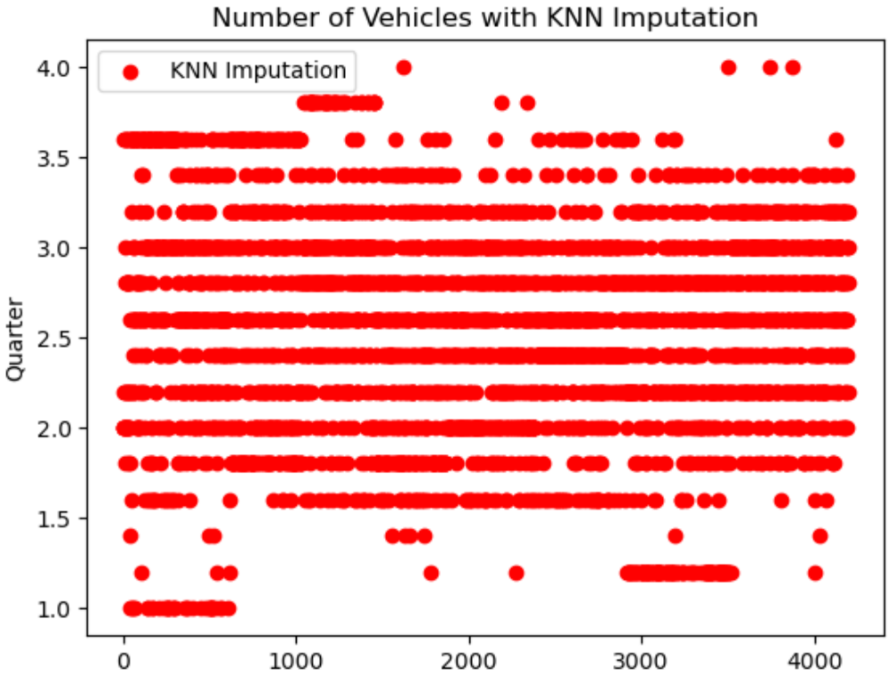
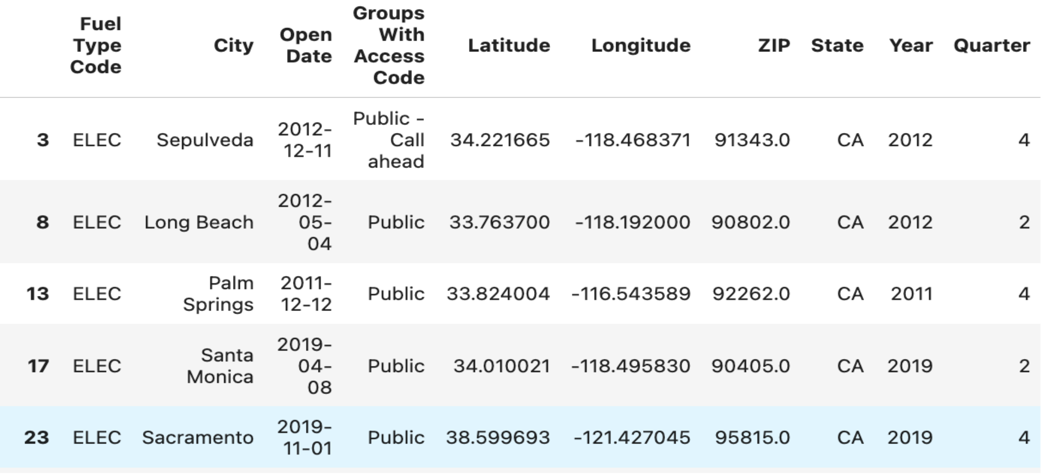


Figure 2: Distribution of Number of EVs with k-NN Imputation

(4) *Gasoline Prices*: sourced from U.S. Energy Information Administration website. This dataset containing U.S. all grades all formulations retail gasoline prices (dollars per gallon) is collected in a monthly frequency. After filtering only values from 2010 to 2023, the average gasoline price will be calculated using values of 4 months per quarter (e.g. January until April will be Quarter 1). Eventually, 56 values of gasoline price will be qualified to be used in the final merged table.

(5) *Electricity Prices:* sourced from Federal Reserve Bank of St. Louis website. This dataset containing average price of electricity per kilowatt-hour in U.S. city average (USD) is also collected in a monthly frequency. The data preparation process for electricity will be similar with the gasoline price process above, containing 56 average gasoline price from the specific time period for the final table.

(6) *Number of Alternative Fueling Stations in California*: gathered from Alternative Fueling Data Center of U.S. Department of Energy. This dataset obtains the locations of nearly 18,000 electric charging stations (both private and public stations). This project, which aims to encourage sustainable development and accelerate the transition from traditional vehicles to EVs in California, will merely include data from public charging stations that were open from January 1, 2010 to December 31, 2023 and remained available. Based on the "Open Date" column, the quarter for each station will be classified and values will be displayed in an additional column named "Quarter" (shown in Figure 3).

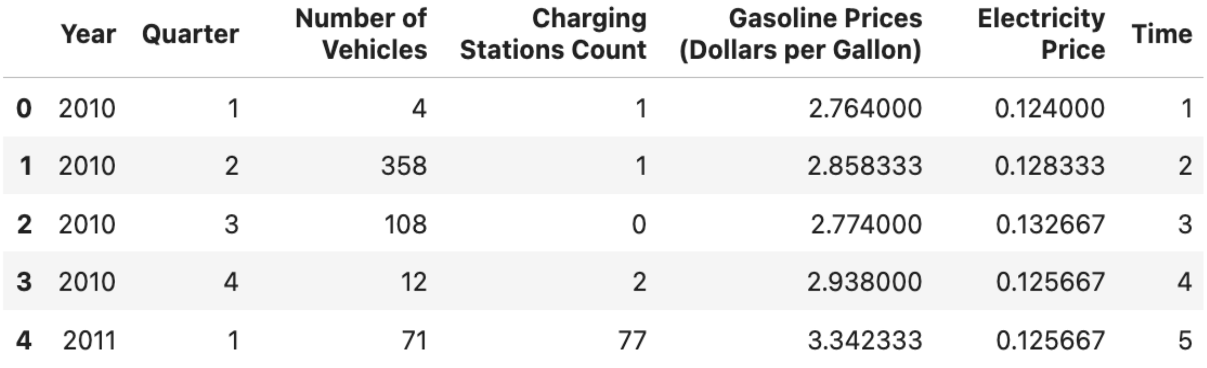


	Fuel Type Code	City	Open Date	Groups With Access Code	Latitude	Longitude	ZIP	State	Year	Quarter
3	ELEC	Sepulveda	2012-12-11	Public - Call ahead	34.221665	-118.468371	91343.0	CA	2012	4
8	ELEC	Long Beach	2012-05-04	Public	33.763700	-118.192000	90802.0	CA	2012	2
13	ELEC	Palm Springs	2011-12-12	Public	33.824004	-116.543589	92262.0	CA	2011	4
17	ELEC	Santa Monica	2019-04-08	Public	34.010021	-118.495830	90405.0	CA	2019	2
23	ELEC	Sacramento	2019-11-01	Public	38.599693	-121.427045	95815.0	CA	2019	4

Figure 3: *The first five lines of Number of Alternative Fueling Stations in California shown after filtering*

The total number of public charging stations newly opened every quarter will be calculated and finalize in the merged table. These would be more accessible to every EV consumer when targeting to sell 100% ZEV in 2035. The research findings assist policymakers in developing programs to stimulate the usage of EVs and regulate charging station standards, thereby facilitating the transition to EVs. In addition, for the purpose of analyzing the current state of charging station density in California in the Exploratory Data Analysis section, the city names, latitude and longitude data from this dataset will be used to create a map representing charging station density.

Below are the first five lines using the .head() method containing 6 columns with one additional "Time" column and contains no duplicates to clarify the time by quarter, in which "Number of Vehicles" is the dependent variable; "Charging Stations Count", "Gasoline Price" and "Electricity Price" are three independent variables studied in the linear regression model:



	Year	Quarter	Number of Vehicles	Charging Stations Count	Gasoline Prices (Dollars per Gallon)	Electricity Price	Time
0	2010	1	4	1	2.764000	0.124000	1
1	2010	2	358	1	2.858333	0.128333	2
2	2010	3	108	0	2.774000	0.132667	3
3	2010	4	12	2	2.938000	0.125667	4
4	2011	1	71	77	3.342333	0.125667	5

Figure 4: *The first five lines of the final merge table shown after processing*

2. Ethics Consideration

This project strictly complies with the Data Ethics Framework and General Data Protection Regulation (2018) for appropriate and responsible data use purpose.

- *Data Privacy*

  Data anonymization techniques will be thoroughly used to protect personally identifiable information. The datasets used in this study are publicly available on U.S. government websites (Federal Reserve Bank, Energy California Commission,…) and does not contain any individually identifiable information to ensure no privacy and confidentiality leakage of individuals. Furthermore, this project will collect merely the data necessary for the analysis to avoid unnecessary or excessive data collection.

- *Data Security*

  Strict access controls would be implemented to limit data access to authorized personnel only to protect against unauthorized access or breaches.

- *Fairness*

  The dataset includes the number of several counties of California that reflects the diversity of California's population to avoid bias. Besides, choice of machine learning models or analytical methods based on their suitability for analyzing EV adoption trends or optimizing charging infrastructure will be clarified.

  When using government, policy, or consultancy reports, it is important to ask questions of the materials. All questions for the wider context (e.g. political, cultural, economic, legal) of the report – for what purpose the report was produced? when it was produced - woudl be carefully considered.

- *Transparency*

  Data from across different datasets will be combined to satisfy the factors affecting EV adoption by cataloging all data sources and data types. All analytical methods and the rationales are clearly described, including handling missing data, normalizing data formats, and performing temporal alignment.

3. *Methodology*

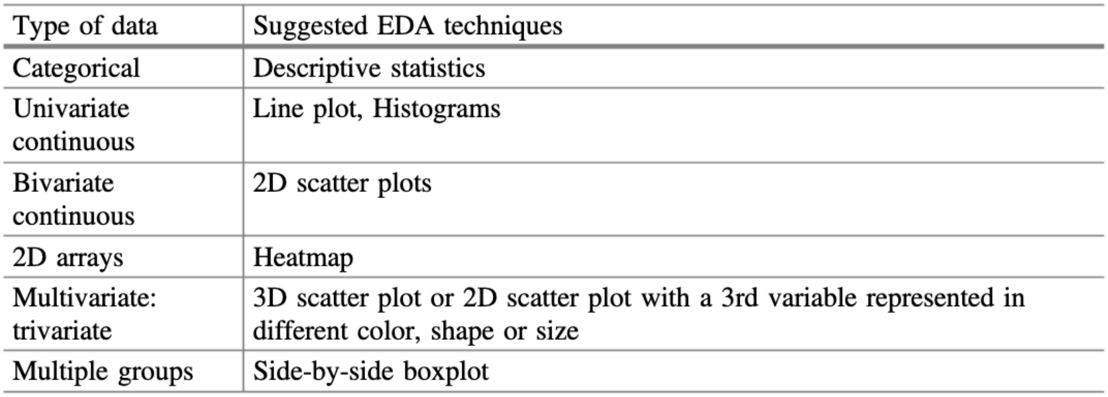
The methodology for this project is based on the development of Exploratory Data Analysis, Correlation Analysis and Linear Regression Methods using Python – a powerful programming language that provides the standard libraries and is capable of handling complex datasets efficiently.

3.1 Exploratory Data Analysis (EDA)

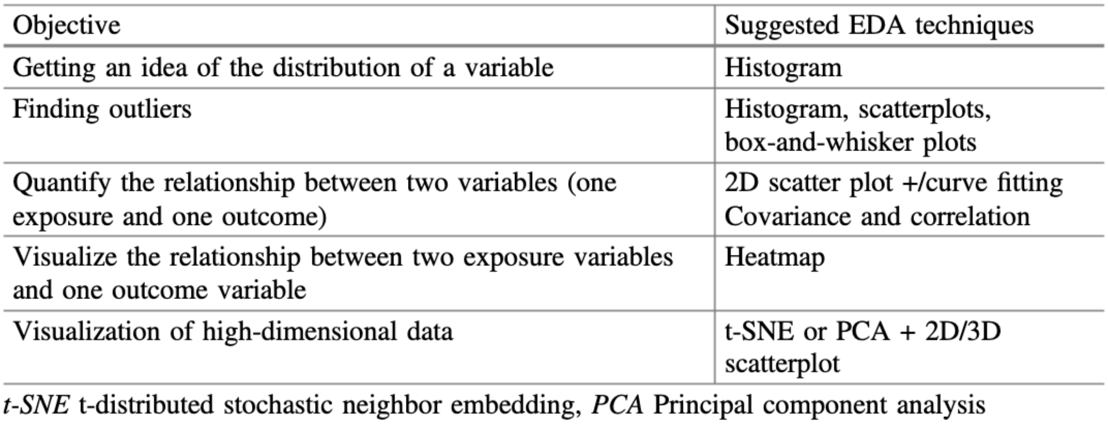
EDA is an important preliminary step in a business project involving analyzing and visualizing data to understand its key characteristics, uncover patterns, and identify relationships between variables. This method taking advantage of using graphics would be helpful in:

- Selecting relevant analysis or prediction techniques requires an understanding of data structures.
- Visualizations and statistical summaries in EDA show hidden patterns and correlations between variables. These findings can help guide future research and enable more effective feature engineering and model development.
- Delving into the nature of the case through descriptive statistics such as the median, mode, maximum, range, and through sorting out abnormalities and outliers in data, analytical results can possibly be improved.
- Correlation heatmaps will be the method to identify significant correlations between variables, which helpes in understanding the interdependencies among features and their potential impact on the target variable.

In the study of Komorowski, M. et al., (2016), there will be two classification of EDA: Graphical or non-graphical methods and Univariate (only one variable, exposure or outcome) or multivariate (several exposure variables alone or with an outcome variable) methods based on the type of data and the objective of the analysis, which would be extremely helpful in order to provide insight into the characteristics and variable(s) distribution:



Type of data	Suggested EDA techniques
Categorical	Descriptive statistics
Univariate continuous	Line plot, Histograms
Bivariate continuous	2D scatter plots
2D arrays	Heatmap
Multivariate: trivariate	3D scatter plot or 2D scatter plot with a 3rd variable represented in different color, shape or size
Multiple groups	Side-by-side boxplot



Objective	Suggested EDA techniques
Getting an idea of the distribution of a variable	Histogram
Finding outliers	Histogram, scatterplots, box-and-whisker plots
Quantify the relationship between two variables (one exposure and one outcome)	2D scatter plot +/curve fitting Covariance and correlation
Visualize the relationship between two exposure variables and one outcome variable	Heatmap
Visualization of high-dimensional data	t-SNE or PCA + 2D/3D scatterplot

t-SNE t-distributed stochastic neighbor embedding, *PCA* Principal component analysis

Figure 5: Recommended EDA techniques based on data type and objective
by Komorowski, M. et al., (2016)

In most cases, conducting EDA using Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn is a thorough approach to uncovering significant further insights from processed data, which forms a strong basis for all subsequent data analysis efforts. This project will make use of EDA to picture the current situation of EV sales in California by assessing historical data and examining the distribution of other external factors, followed by calculating the correlation coefficient between variables using appropriate visual representations.

3.2 Correlation Coefficient Matrix

The Pearson Correlation Coefficient Matrix is the key approach to visualize correlations between each pair of variables in a data collection. This is a statistical measure that assesses the strength and direction of the relationship between two continuous variables. This coefficient not only indicates the magnitude of the correlation but also shows its direction. It is simulated using the Correlation Heatmap, as proposed by Komorowski, M. et al. (2016) (Figure 5). The direction and strength of the association between two variables is reflected by correlation, involving a two-way relationship where a change in variable A results in a change in variable B and vice versa.

This critical step determines which variables to be used as input features for the regression model while checking for multicollinearity. This occurs when two independent variables are substantially associated in an ordinary least squares (OLS) regression model, indicating that collinear independent variables are not genuinely independent, potentially leading to a weak regression model. Therefore, a statistical indicator called the variance inflation factor (VIF) can detect and measure the amount of collinearity in a linear regression model. VIF measures how much the variance of a regression coefficient is inflated due to multicollinearity among the predictor variables in a model.

The VIF for the j-th predictor variable is calculated as: $VIF_j = \frac{1}{1-R_j^2}$

where R_j^2 is the coefficient of determination (R-squared) obtained by regressing the j-th predictor on all the other predictors.

A VIF of 1 will mean that the variables are not correlated; a VIF between 1 and 5 shows that variables are moderately correlated, while values between 5 and 10 will mean that variables are highly correlated, which can be problematic for regression analysis.

Remarkedly, the degree and direction of a linear relationship between two variables will be demonstrated by the value of the correlation coefficient ranging from -1 to 1. A correlation value of one indicates that the two variables have a complete positive correlation. A correlation value of -1 indicates a perfect negative correlation (one variable increases while the other decreases). Adversely, 0 indicates no linear correlation between the two variables.

3.3 Linear Regression Model

This is a common predictive modelling technique that investigates the relationship between dependent and independent variables, creating an equation. In contrast to Correlation, Regression is a one-way relationship from the independent variables to the dependent variable.

This technique uses Python's modules to forecast outcomes based on input factors and identify the causal effect relationship. The variables will be incorporated into an OLS regression, a technique creating the best-fitting straight line to minimize the squared errors (the distance between the line and each observation or differences between the predicted and actual values). Later on, estimate the coefficients that minimize the sum of squared residuals to avoid any OLS assumption violations.

There are three important issues that need to focus when analyzing OLS: (1) *whether the regression coefficient is statistically significant*, (2) *whether the model is meaningful*, and (3) *the level of model explanation.*

(1) The hypothesis would be $\beta_0 = 0$, aiming at rejecting this hypothesis. That means the beta coefficient is different from 0, the estimated beta coefficient can be used to interpret the influence of the independent variables X_1, X_2, ..., X_n on the fluctuation of the dependent variable Y. To validate a hypothesis against observed data, t-test or a p-value, which equals to 0.05 or lower is generally considered statistically significant, will be analyzed in this project to perform this test.

(2) Model testing, known as F-statistic, has the hypothesis for this test that all regression coefficients are simultaneously equal to 0. If the hypothesis is not rejected, this means that the model is not statistically significant.

(3) The level of model explanation will be tested through the value of R-Squared (R^2).

Some other metrics that will also be considered for model selection to balance bias and variance or model accuracy and complexity on the available data are Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Using the estimated log-likelihood and number of parameters as well as sample size, the smaller the AIC or BIC the better. BIC is more useful approach in selecting a correct model while the AIC is more appropriate in finding the best model for predicting future observations (Chakrabarti et al., 2011).

The basic equation is:

$$\boldsymbol{Y = \beta_0 + \beta_1 .\ x_1 + \beta_2 .\ x_2 + \ldots + \beta_n .\ X_n + \epsilon}$$

where Y is the dependent variable; $x_1, x_2, \ldots, x_n$ *are the independent variables;*
β_0 *is the y-intercept;* $\beta_1, \beta_2, \beta_3$ *is the slope or coefficients to be estimated;*
and ε *is an error term.*

The final equation is:

$$\boldsymbol{EV\_sales = \beta_0 + \beta_1 .\ \text{Charging Stations} + \beta_2 .\ \text{Electricity Price} + \beta_3 .\ \text{Gasoline Price} + \epsilon}$$

where

EV_sales	*prediction of wholesale total EV (BEV in specific) in the future in California;*
Charging Stations	*total number of electric charging stations;*
Electricity Price	*the price of Electricity Price in United States*
Gasoline Price	*the price of Gasoline Price in United States*

Assumption of Regression Model:

- Linearity: Number of EV sales in California are linearly correlated with Number of Charging Stations, Electricity Price and Gasoline Price.
- Homoscedasticity: Constant variance of the errors should be maintained.
- Multivariate normality: The model assumes that the residuals are normally distributed.
- Lack of Multicollinearity: It is assumed to have little or no multicollinearity in the data.

Hypothesis testing:

Null Hypothesis: H_0: $\beta_i = 0$ (assuming independent variables have no effect on EV_sales)
Alternative Hypothesis: H_a: $\beta_i \neq 0$ *(*assuming independent variables have effect on EV_sales)

Significantly, it is crucial to evaluate the regression model effectiveness using key metrics namely:

- *R-Squared (R^2):* measures the proportion of variance in the dependent variable explained by the model. It ranges from 0 to 1, where higher value generally indicates a better fit.
- *Mean Absolute Error (MAE):* is the mean of the absolute value of the difference between actual and predicted values. MAE is also referred to as a loss function since the goal is to minimize this loss function (Belyadi et al., 2021): $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$

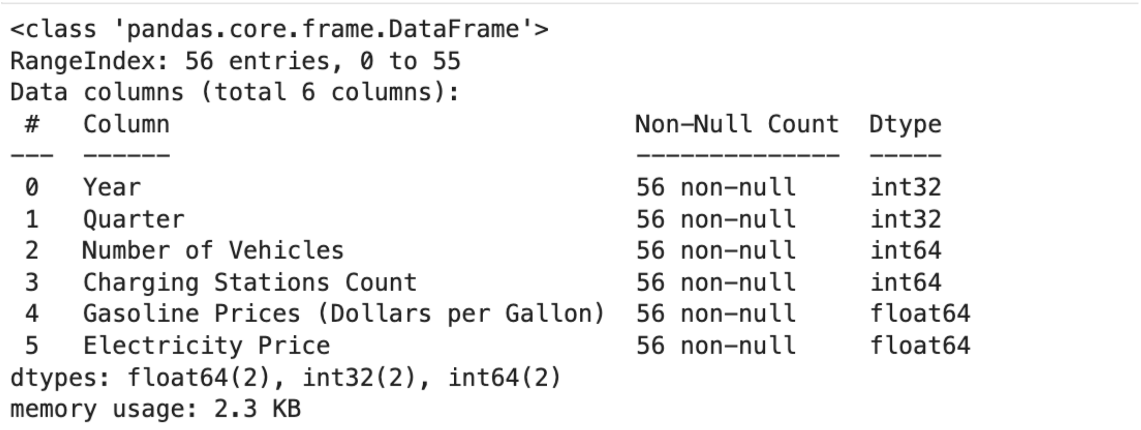
  where y_i and $\hat{y}_i$ are relatively the observed and predicted value of the dependent variable for the i^{th} observation
- *Root Mean Squared Error (RMSE):* is the square root of mean squared error (MSE), which is the same as MAE but it happens before taking the sum of all values: $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$.

  This loss function is commonly used due to its interpretive capability and gives more weight to larger errors.

IV. Results and Discussion

1. Exploratory Data Analysis

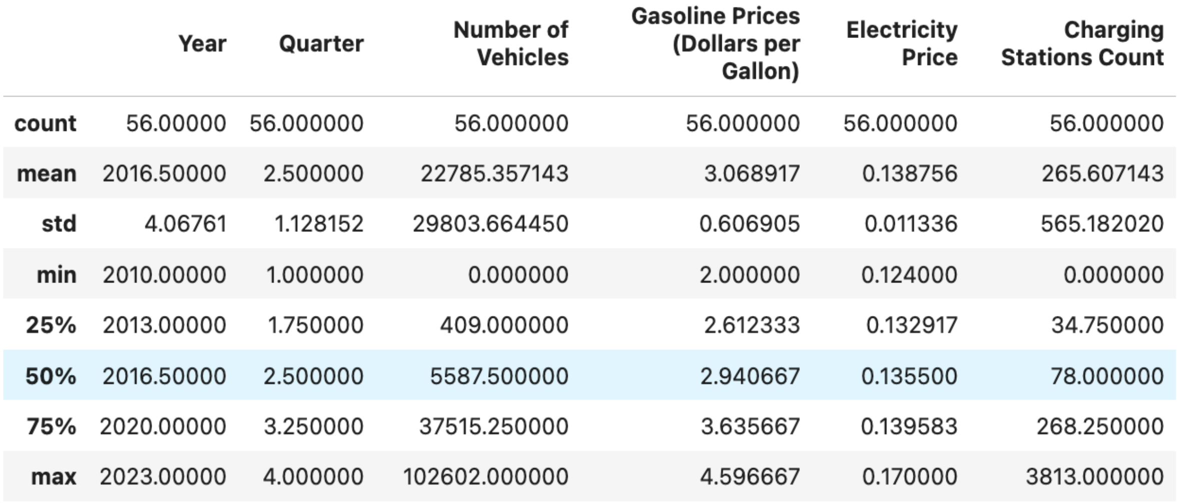
After the merging process mentioned in the previous part, a concise summary of the Dataframe will be detailedly provided using the *.info()* to understand the dataset. Based on the given information, there is no missing value and columns “Year”, “Quarter”, “Number of Vehicles” and “Charging Stations Count” are in integer form, meanwhile the rest has the float data type.



```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 56 entries, 0 to 55
Data columns (total 6 columns):
 #   Column                             Non-Null Count  Dtype  
---  ------                             --------------  -----  
 0   Year                               56 non-null     int32  
 1   Quarter                            56 non-null     int32  
 2   Number of Vehicles                 56 non-null     int64  
 3   Charging Stations Count            56 non-null     int64  
 4   Gasoline Prices (Dollars per Gallon)  56 non-null     float64
 5   Electricity Price                  56 non-null     float64
dtypes: float64(2), int32(2), int64(2)
memory usage: 2.3 KB
```

Figure 6: Data Summary

The *.describe()* function shows basic statistical characteristics of each numerical feature (int64 and float64 types): number of non-missing values, mean, standard deviation, range, median, the first and third quartiles. Overall, each variable has 56 records and the mean number of vehicles is approximately 22,785, but this value has high variability (standard deviation of 29,803) while approximately 266 charging stations were newly established in California each quarter, which implies building one new charging station can support the electricity supply for approximately 86 EVs. The evolution of charging stations counts over the time period range from zero to 3,813 stations. The minimum number of vehicles is 0, and the maximum is 102,602, based on the k-NN Imputation, a significant growth or fluctuations in the number of vehicles over time has been witnessed. There was a moderate fluctuation in gasoline prices, in contrast, electricity prices experienced a stability, remaining approximately $0.14 per unit with extremely low variability (standard deviation of $0.01).



	Year	Quarter	Number of Vehicles	Gasoline Prices (Dollars per Gallon)	Electricity Price	Charging Stations Count
count	56.00000	56.000000	56.000000	56.000000	56.000000	56.000000
mean	2016.50000	2.500000	22785.357143	3.068917	0.138756	265.607143
std	4.06761	1.128152	29803.664450	0.606905	0.011336	565.182020
min	2010.00000	1.000000	0.000000	2.000000	0.124000	0.000000
25%	2013.00000	1.750000	409.000000	2.612333	0.132917	34.750000
50%	2016.50000	2.500000	5587.500000	2.940667	0.135500	78.000000
75%	2020.00000	3.250000	37515.250000	3.635667	0.139583	268.250000
max	2023.00000	4.000000	102602.000000	4.596667	0.170000	3813.000000

Figure 7: Variable Descriptive Statítics

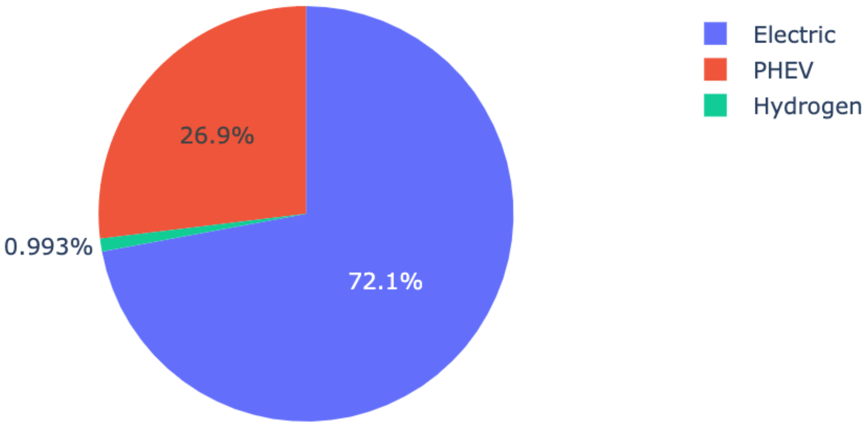


Figure 8: Percentage of three vehicle types in 2010 - 2023

According to the original sales dataset from the California Energy Commission website, being the most prevalent among three car types, traditional BEVs accounted for 72% of all vehicles sold statewide, while PHEV and hydrogen vehicles accounted for nearly 27% and 1%, respectively. Comprising the majority of sales in California, BEVs has grasped the growing preference and infrastructural support from numerous users, this also visualizes the still-niche market for hydrogen-powered vehicles. The fact that BEVs account for the bulk of vehicle sales is a good sign for customer preference and California's effort in transition to zero-emission vehicles as BEVs are more environmentally-friendly. Based on *"Electric vehicles and charging infrastructure in California*" reported by Statista, BEVs are also the vehicle category that has received the most CVRP rebate in California since 2010, with about $101.3 million granted across 38,307 applications by 2022. As stated, this study will only analyze the BEV sales sector as one variable in order to encourage further adoption of these vehicles in order to fulfill the 2030 emissions reduction target set by President Biden's Historic Climate Agenda in 2021.

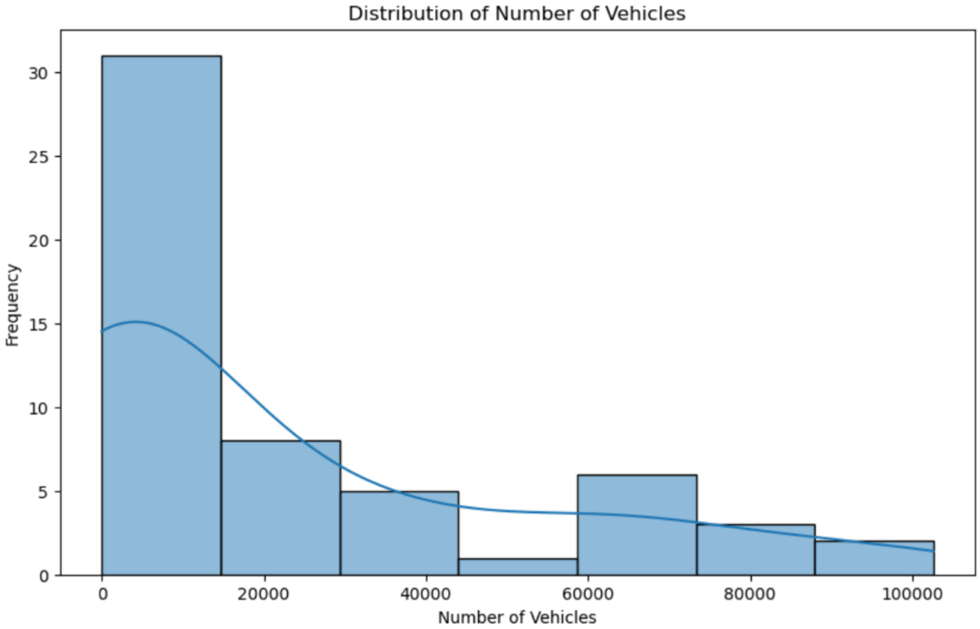


Figure 9: Distribution of Number of Vehicles

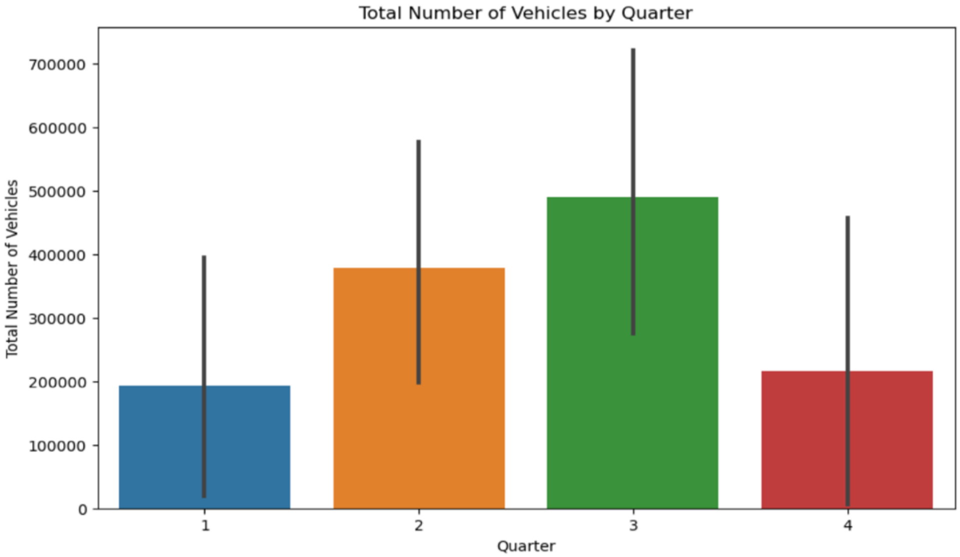


Figure 10: Total Number of Vehicles by Quarter

The first variable to be explored is the EV sales with values ranging from 0 to over 100,000 EVs and Quarter 3 being the month with the highest volume. However, the application of the kNN Imputation method will possibly affect this distribution due to the shortage of Quarter data for the time period 2010 to 2020. Although the result is structured and consistent for the distribution model, in reality there will be differences compared to this data.

The two charts below depict the rise of EV sales over time and quarters. The growth tendency has been substantial, particularly since 2020. Changes to an income cap of CVRP funding, which went into effect in 2016, have enabled people to get better access to the EV market. Since 2018, the growth rate has slowed because to the COVID-19 epidemic, and consumers demand for automobile sales has decreased. However, demand has generally now improved, even reaching its previously set EV sales target of 1.5 million EVs two years earlier than its projected 2025 goal.

The second graph shows that the number of automobiles remained relatively low and constant between 2010 and 2016. During this early age of EVs growth, minor changes were made in EV sales, no substantial growing pattern was seen. Frequent spikes and decreases in the EV sales can only be seen after the year of 2016.

The swings indicate seasonal or cyclical trends in vehicle adoption, or even responses to external variables such as economic conditions or legislative changes. Since 2020, the graph indicates a clear exponential increase in the number of automobiles. The high surge implies the car adoption or manufacturing has accelerated significantly in the latest years. The peaks rise dramatically, reaching over 100,000 vehicles by the end of the observation period resulting from technological advancements. However, during the last three quarters in 2024, there has been a marginal decrease in the sales, which creates debates among the policymakers.

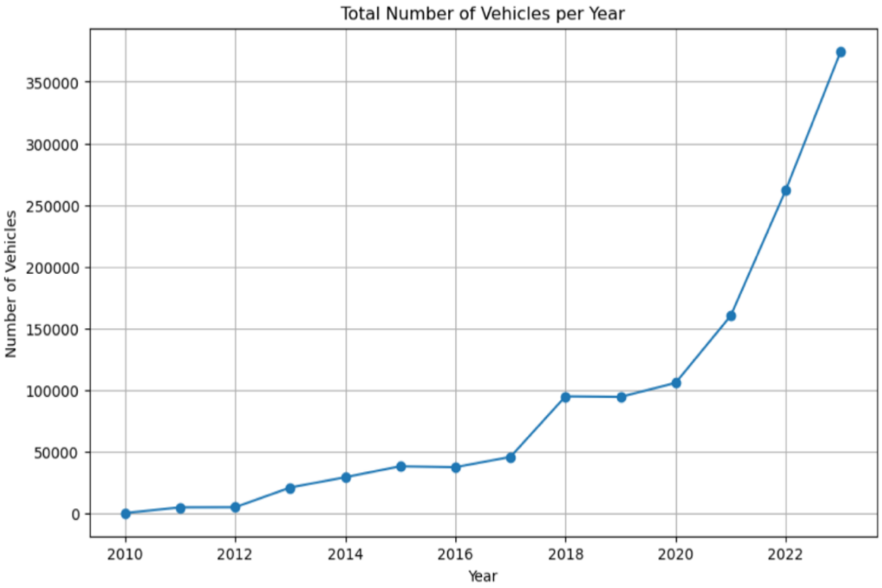


Figure 9: Total Number of Vehicles by Year

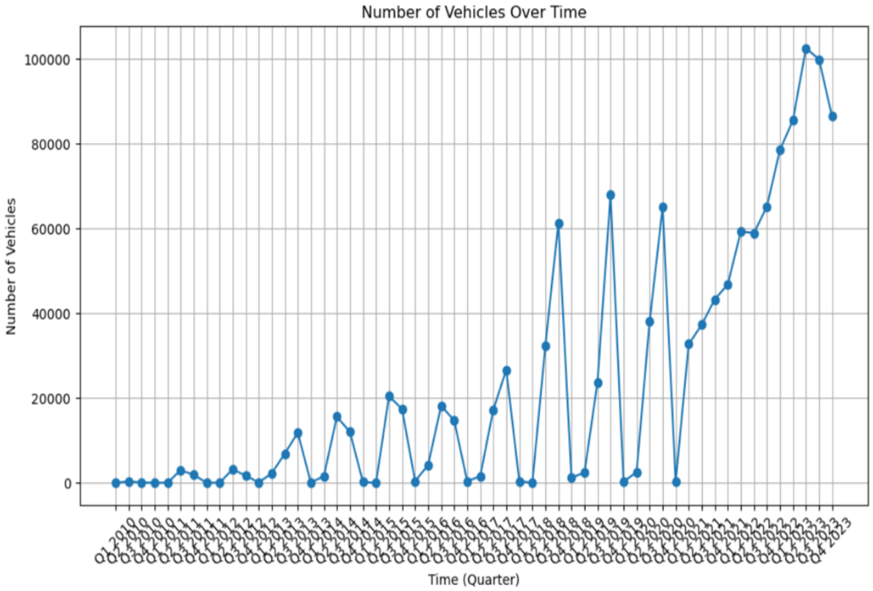


Figure 10: *Total Number of Vehicles over quarters*

In terms of alternative fuel stations in California, longitude and latitude values from the dataset sourced from Alternative Fueling Data Center of U.S. Department of Energy had been used to provide insights about density and distribution of all active charging stations. Utilized data from Figure 13, the scatter plot and mapbox of location of charging stations were created. First of all, the top 10 cities in California with the most charging stations are concentrated in large urban areas with the top highest population in this state, particularly in which Los Angeles has over 1,400 charging stations, nearly 2.5 times more than the two cities ranked 2nd and 3rd.

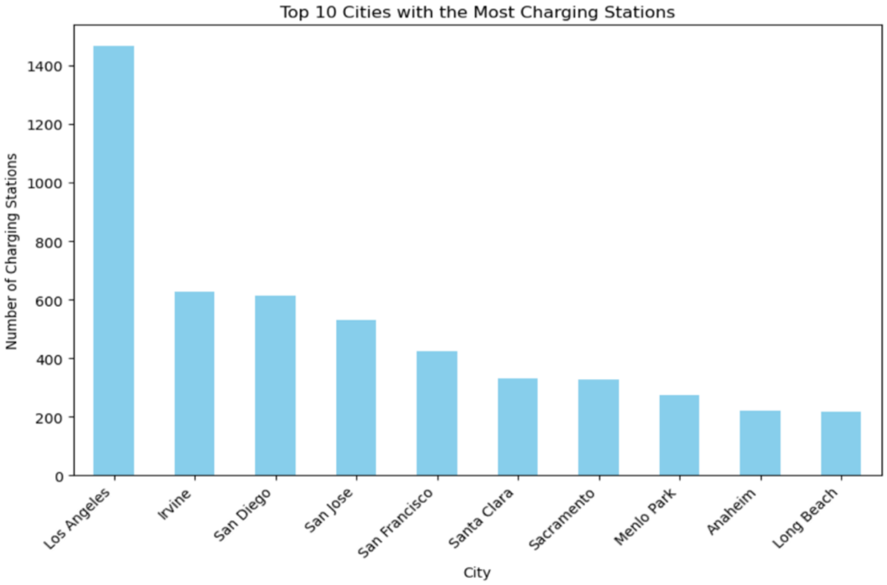
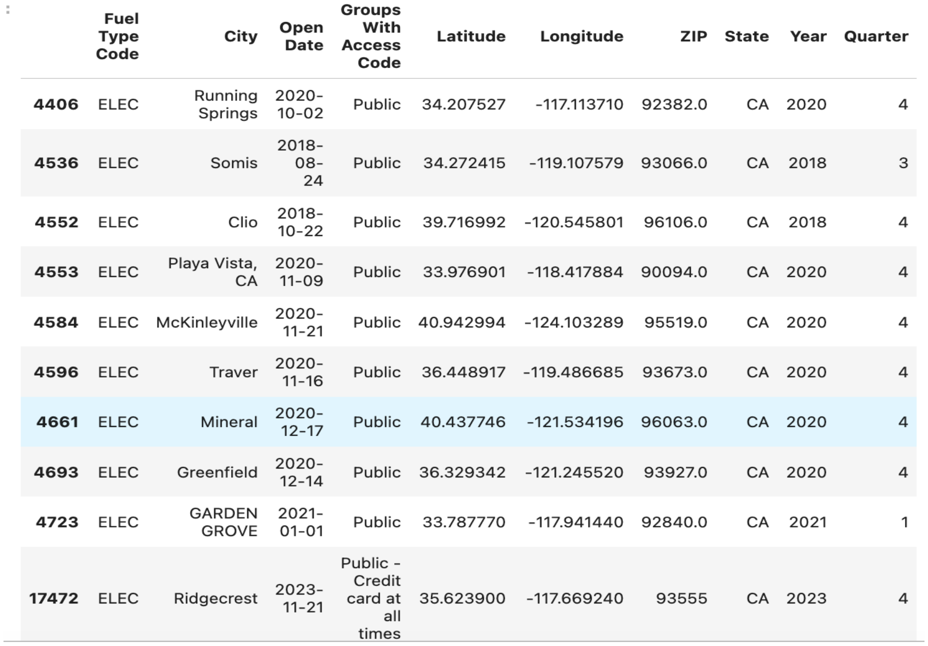


Figure 11: *Top 10 Cities with the most charging stations*

On the contrary, in some smaller cities like Garden Grove or Ridgecrest,… up to now, there is only one charging station, creating such a huge difference and not truly effective in increasing the recognition of charging stations as well as EVs and educating local people in the area about the application of EVs.

Based on the charts, the following recommendations can be made for charging infrastructure planning in California. Since the majority of charging stations are located in urban areas, policymakers and California authorities should focus on promoting the use of electric cars in particularly small cities. In addition, while Los Angeles counties have the highest number of electric cars, there is still potential to expand the use of electric cars in other counties in the state of California. There may be specific cities or neighborhoods with lower adoption rates and poor charging station counts in Figure 12. Targeting these rural areas could be a strategic plan to promote the greater use of EVs through customised marketing with outreach efforts. Increasing the number of public chargers by 2035 when selling only EV cars is also a long-term plan to better meet the needs of the entire state.



	Fuel Type Code	City	Open Date	Groups With Access Code	Latitude	Longitude	ZIP	State	Year	Quarter
4406	ELEC	Running Springs	2020-10-02	Public	34.207527	-117.113710	92382.0	CA	2020	4
4536	ELEC	Somis	2018-08-24	Public	34.272415	-119.107579	93066.0	CA	2018	3
4552	ELEC	Clio	2018-10-22	Public	39.716992	-120.545801	96106.0	CA	2018	4
4553	ELEC	Playa Vista, CA	2020-11-09	Public	33.976901	-118.417884	90094.0	CA	2020	4
4584	ELEC	McKinleyville	2020-11-21	Public	40.942994	-124.103289	95519.0	CA	2020	4
4596	ELEC	Traver	2020-11-16	Public	36.448917	-119.486685	93673.0	CA	2020	4
4661	ELEC	Mineral	2020-12-17	Public	40.437746	-121.534196	96063.0	CA	2020	4
4693	ELEC	Greenfield	2020-12-14	Public	36.329342	-121.245520	93927.0	CA	2020	4
4723	ELEC	GARDEN GROVE	2021-01-01	Public	33.787770	-117.941440	92840.0	CA	2021	1
17472	ELEC	Ridgecrest	2023-11-21	Public - Credit card at all times	35.623900	-117.669240	93555	CA	2023	4

Figure 12: *Top 10 Cities with the lowest number of charging stations*

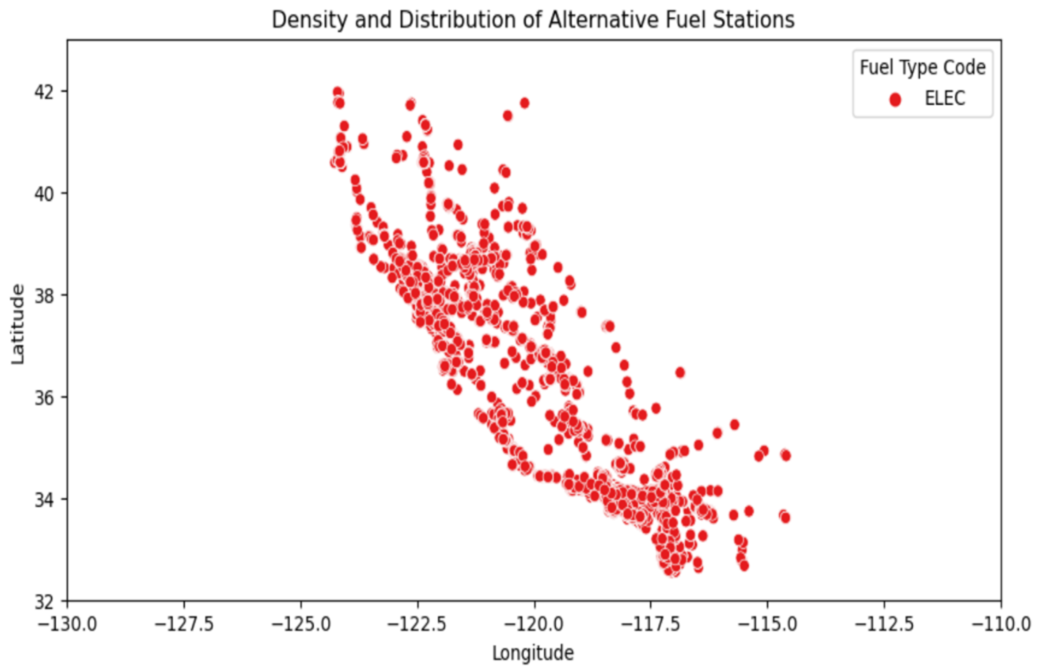


Figure 13: Charging stations density by latitude and longitude

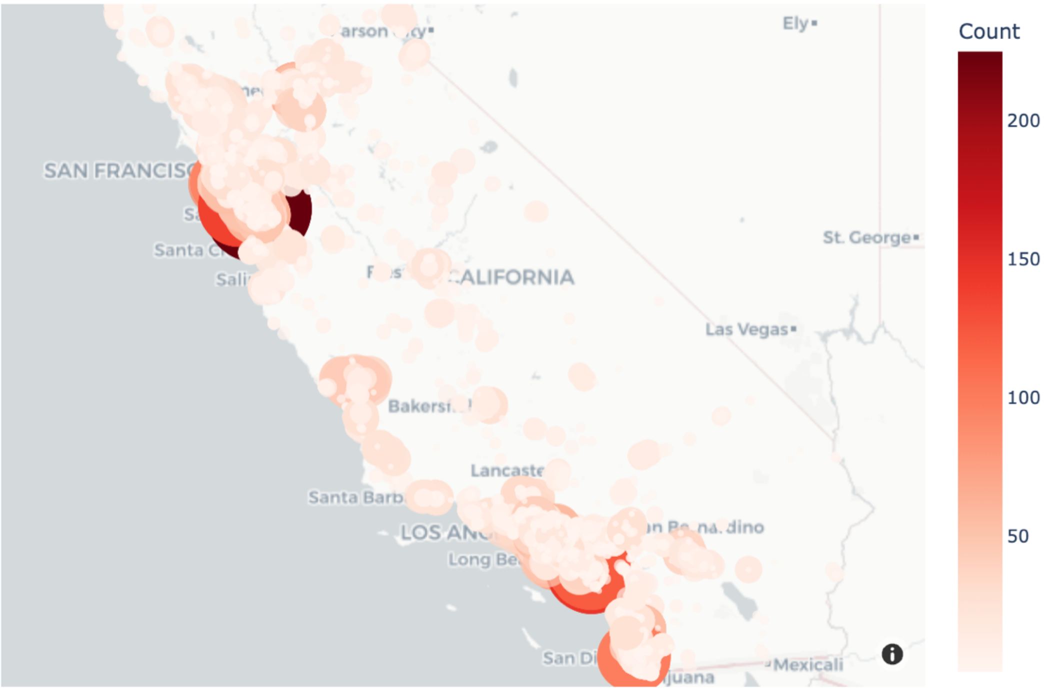


Figure 14: Charging stations spread map

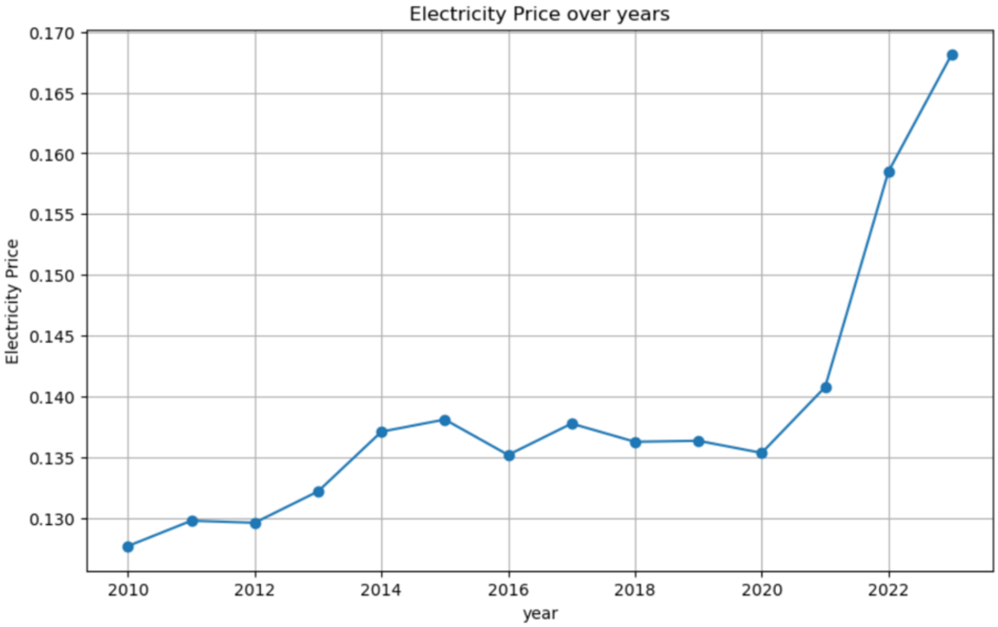


Figure 15: Changes in Electricity Price over years

In Figure 15, there was an upward trend in Electricity prices witnessed from 2010 to 2022, with particularly strong increases in these recent years. Prior to 2014, electricity prices had some minor fluctuations, then 7 years later, electricity prices remained relatively stable, fluctuating slightly below 0.14 dollars, which was followed by a dramatic surge to 0.17.

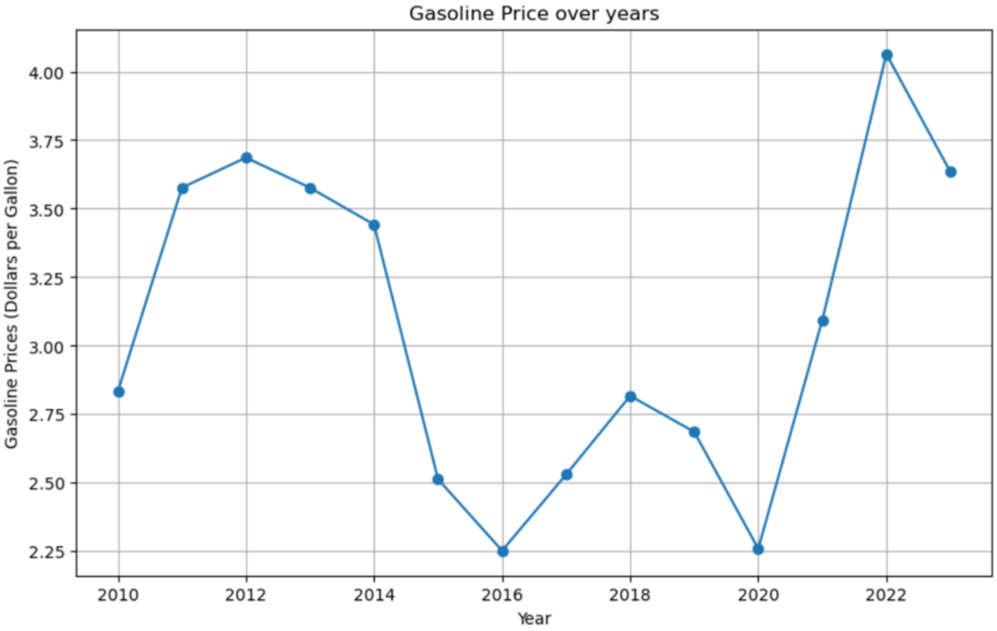


Figure 16: Changes in Gasoline Price over years

Figure 16 above depicts the varying trajectory of gasoline prices over time, reaching peaks around 2012-2014 and 2022, as well as steep drops in 2016 and 2020. In 2016, the lowest price was approximately \$2.25 per gallon. From 2020 to 2022, gasoline prices rose rapidly from the lowest point to highest at over \$4, despite being quite steady than in previous years. Gasoline prices fell slightly to below \$3.75 after peaking in 2022. This volatility demonstrates that gasoline prices are highly varied due to changes in economic, political, and environmental issues.

Lately, the global energy crises, as well as rising inflation and demand, have contributed to the sharper changes in power and gasoline costs in recent years. According to energy experts, every cent increase in gasoline prices will cost Americans \$4 million more per day in 2022. So, will fluctuations in energy prices affect the desire in purchasing EVs? The Correlation Heatmap below will show the correlation between independent variables and the dependent variables, as well as checking for multicollinearity before performing a regression.

2. Correlation Heatmap

In Figure 17, the scatter plots and histograms had been generated by Python and specifically analyzed so as to gain a better understanding of the distribution and correlation between variables. It is noticeable that the distribution is right-skewed and there exists some extremely high outliers due to EV sales fluctuations over years. The density of charging stations in all cities, in general, is low, with an average growth of less than 1,000 stations. Some of the outliers may indicate significant urban centers or areas with high EV adoption rates. Another factor to consider is that Gasoline Price and Electricity Price have a symmetrical distribution around the mean value.

Based on the direction of the quantity of EVs and the number of charging stations, the two variables would increase simultaneously illustrated by the consistent increase in the correlation. This trend can be practically understandable in the light of booming EVs, the greater charging demand will require extended charging station availability. Although there are a few data points indicating a slight upward trend between number of vehicles and electricity price, numerous scattered data points shows a weak or non-existent relationship between these variables. It also seems that there might not be a clear relationship between EV sales and gasoline price. Additionally, there is no distinct linear relationship between the other pairs of variables.

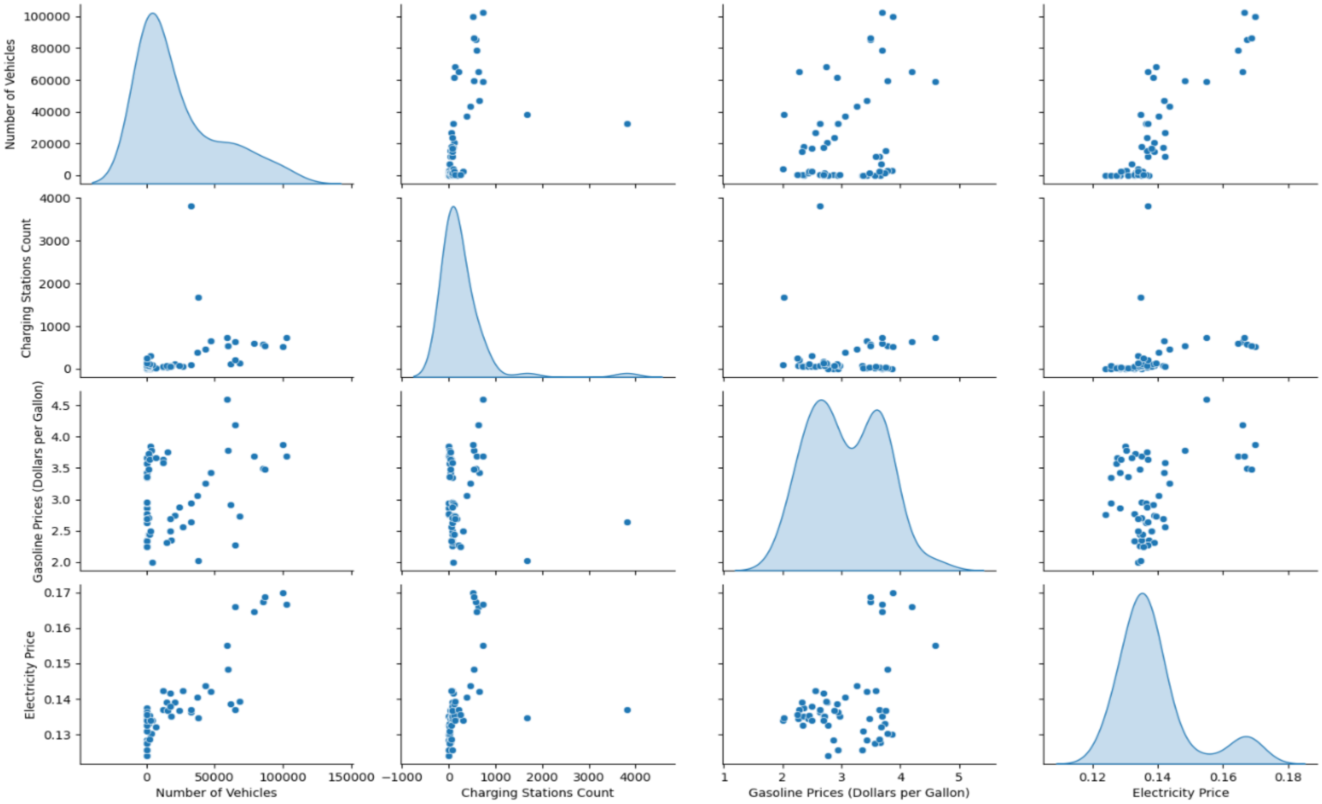


Figure 17: The distribution and correlation between variables

The Correlation Heatmap consolidates and quantifies the observed relationships in the Figure 18, representing the value of the correlation coefficient in color and a specific number from -1 to 1. The sign of the Pearson correlation coefficient determined the direction of the relationship. If the coefficient is positive, an increase in one variable is associated with an increase in the other variable, graphed in an upward slope on a scatterplot. Conversely, negative coefficients signify that an increase in one variable is linked to a decrease in the other variable, producing a downward slope.

A significant relationship between "Number of Vehicles" and "Charging Stations Counts" with the correlation coefficient of 0.37 aligns with the growing charging demand. In contrast, a weak relationship between variables "Gasoline Price" and "Charging Stations Counts" has a negative correlation of -0.019. In general, it is not necessary for high gasoline prices to increase as a result of a substantial rise in EV sales as there is no clear strong inverse relationship. A weak positive correlation between "Charging Stations Counts" and "Electricity Price" is shown by a moderate coefficient of 0.29.

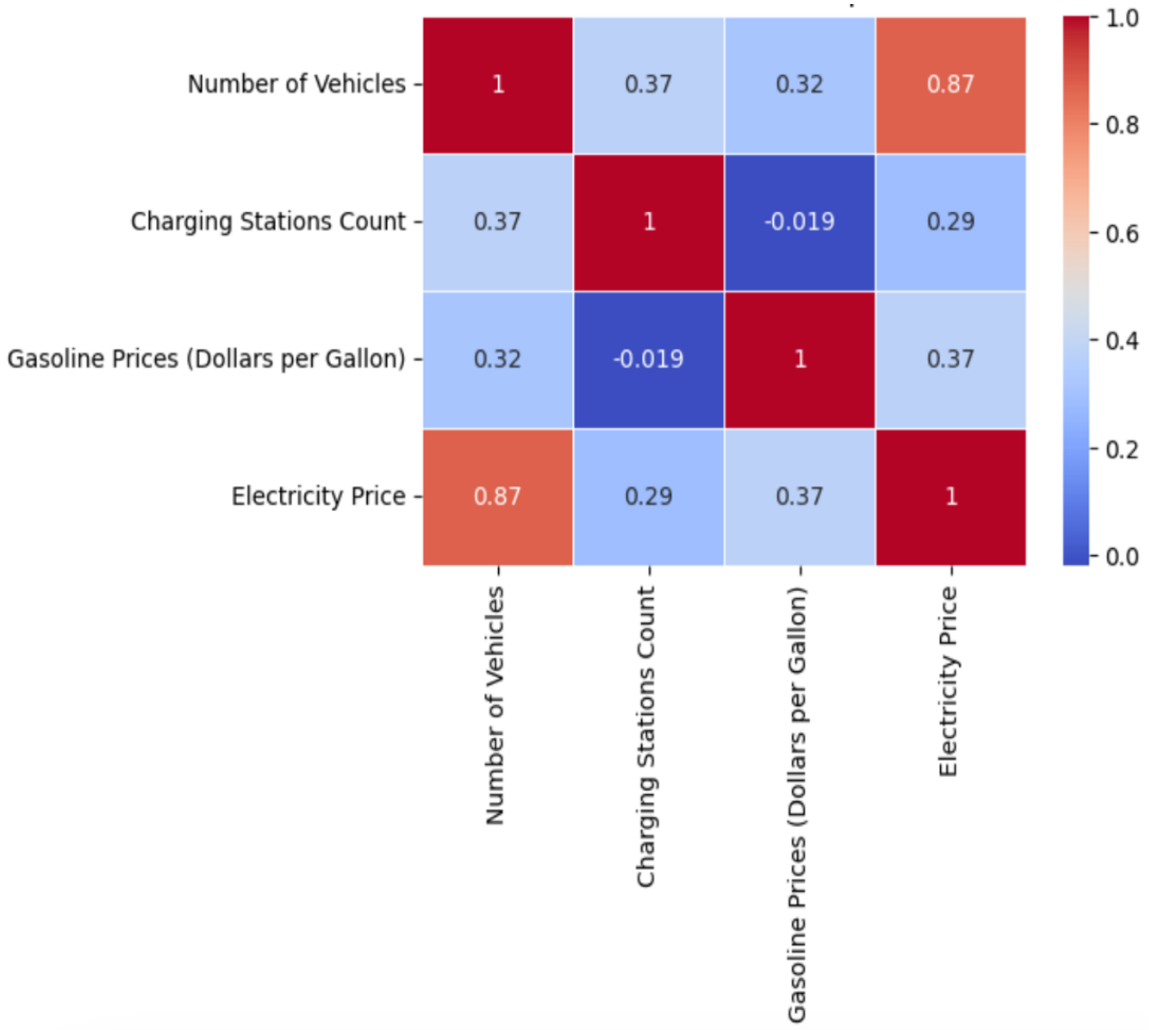
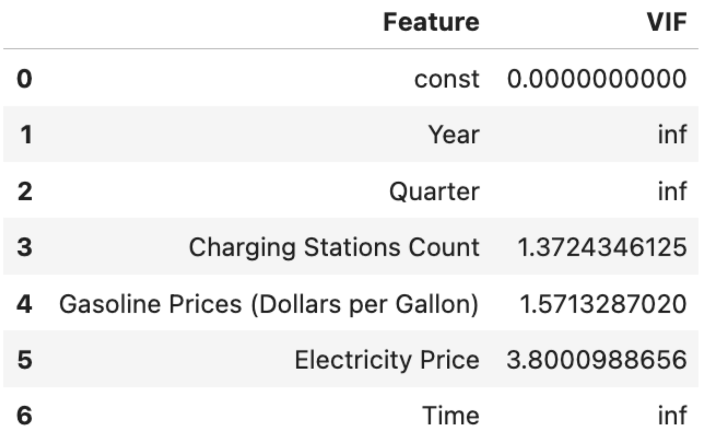


Figure 18: Correlation Heatmap

Upon reviewing the graph, it is evident that there is a strong correlation of 0.87 between "Number of Vehicles" and "Electricity Price". In this case, severe multicollinearity might be a consequence of high correlations among these two variables. Accordingly, VIF would be the assessing indicator in order to avoid this statistical concept leading to a less reliable model prediction and remove violating variables. Computing VIF algorithm using variance_inflation_factor() functions from Python, values of all three independent variables are below 5, which warrants the satisfactory variance of the estimated coefficients. As a rule of thumb, VIF below 5 indicates an unproblematic amount of collinearity and does not cause concern to the predicting model. Eventually, the regression model will remain unchanged with three independent variables.



	Feature	VIF
0	const	0.0000000000
1	Year	inf
2	Quarter	inf
3	Charging Stations Count	1.3724346125
4	Gasoline Prices (Dollars per Gallon)	1.5713287020
5	Electricity Price	3.8000988656
6	Time	inf

Figure 19: VIF result

To sum up, the Correlation Heatmap demonstrates the complex relationship between factors related to EV sales, including Charging Stations Counts, Gasoline and Electricity Price. The strongest correlation is between vehicle count and electricity price. However, after checking the VIF indicator, no multicollinearity exists, implying that the variables and regression equations remain unchanged. On the other hand, external factors such as Charging Stations Counts and Gasoline Price have a moderate or weaker influence.

3. ***Regression Model***

3.1. Model Comparison

In the light of selecting the most optimal model for the project based on the input variables, it is necessary to take the first step comparing the performance of the basic regression model including OLS and regularized methods Ridge and Lasso Regression. OLS can easily overfit if the data has multicollinearity, while Ridge and Lasso apply regularization to deal with this problem by reducing the variance at the expense of adding bias. The three statistical metrics R^2, RMSE and MAE are considered to be criteria for this comparison.

OLS regression coefficients produces unbiased estimators of the population values. The other two techniques use L2 regularization, adding penalties in order to eliminate overfitting from the model. The purpose of Ridge regression is to reduce model complexity and multicollinearity while Lasso regression aims at simplifying the model.

In fact, regression models are often sensitive to the scale of the input variables, so StandardScaler() will generalize them to avoid the situation where some variables create a significant influence. K-Fold Cross Validation method is now combined to increase the generalizability and overall model performance on unseen data. Hyperparameter tuning can lead to much better performance on test sets. However, optimizing parameters to the test set can lead to information leakage causing the model to perform worse on unseen data. The training data used in the model is split into k number of smaller sets, to be used to validate the model. The model is then trained on (k – 1) folds of training set. The remaining fold is then used as a validation set to evaluate the model.

Hyperparameters 'alpha' are parameters controlling the strength of regularization to be identified before training the model. In this case, GridSearchCV will be used to find the optimal hyperparameter. Each alpha value (0.1, 1.0, 10.0, 100.0) will be subjected to 5-fold cross-validation (cv=5) to choose the alpha that gives the best result on the validation data.

	RMSE	MAE	R^2
Linear Regression	13945.2477	10187.3548	0.7771
Ridge Regression Best Ridge Params: {'alpha': 1.0}	13953.0277	10075.8898	0.7768
Lasso Regression Best Lasso Params: {'alpha': 100.0}	13946.0268	10192.5926	0.7771
Best Model	**Linear Regression (OLS)**		

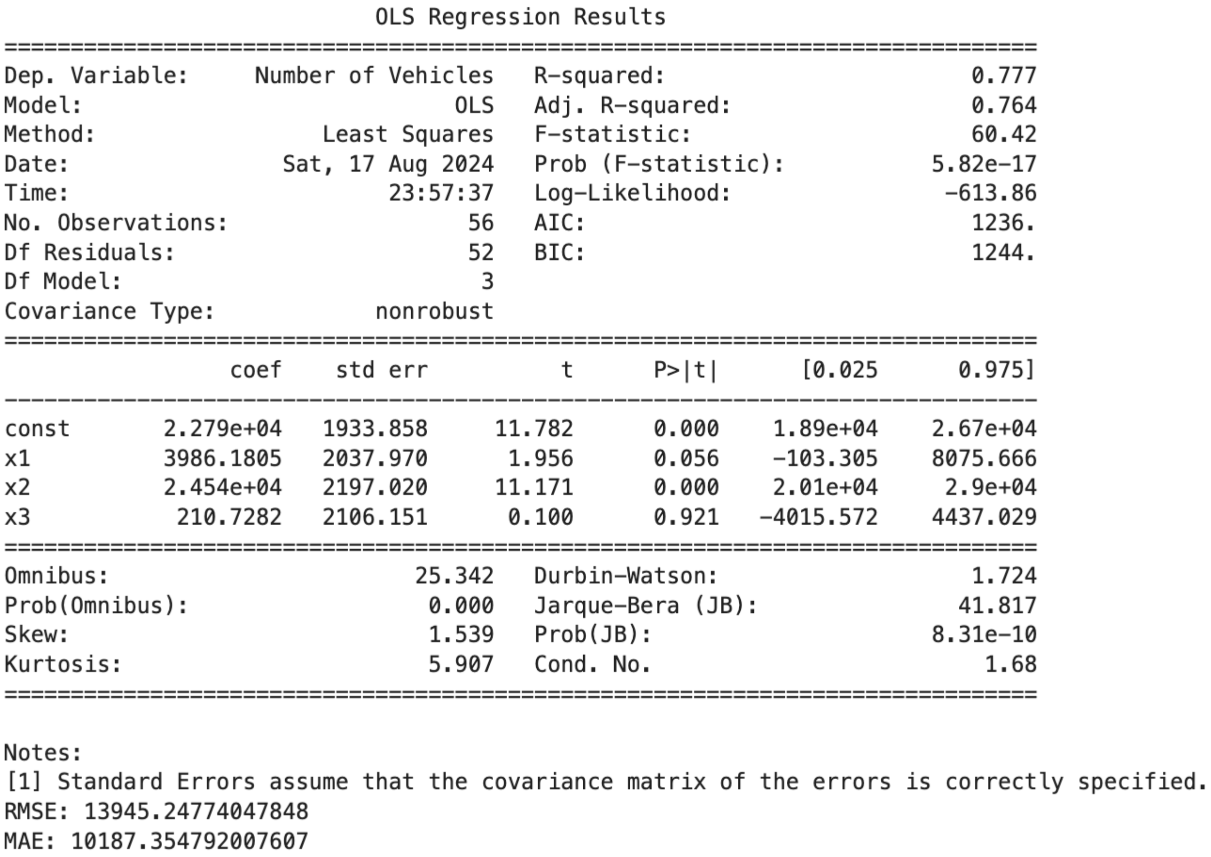
Figure 20: Comparison results of three model performance based on RMSE and MAE

Based on Figure 20, Linear Regression (OLS) model has provided the best performance on the dataset compared to other models with the lowest RMSE. This best performing model has no regularisation and will be selected to analyze the influence of independent variables on the dependent variable.

3.2. OLS Regression Model

Once having checked multicollinearity, training OLS regression model would be the next step using the *statmodels* library in Python. This process is to estimate the coefficients of the independent variables and calculate other statistical measures associated with the regression model. Normally, in order to detect a machine learning model behavior, the dataset observations will be divided into Train Set and Test Set. If the dataset is relatively small (n < 10,000), 70:30 split would be considered a suitable choice. However, due to the limitation of data (n < 1,000), for this project, each valuable observation would be utilized. In this case, k-fold cross-validation is a better model evaluation choice than splitting the observations, which will be analyzed later.

Figure 21 presents the result summary of the OLS regression model after fitting to the preprocessed dataset using .fit() function.



```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:     Number of Vehicles   R-squared:                       0.777
Model:                            OLS   Adj. R-squared:                  0.764
Method:                 Least Squares   F-statistic:                     60.42
Date:                Sat, 17 Aug 2024   Prob (F-statistic):           5.82e-17
Time:                        23:57:37   Log-Likelihood:                -613.86
No. Observations:                  56   AIC:                             1236.
Df Residuals:                      52   BIC:                             1244.
Df Model:                           3                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       2.279e+04   1933.858     11.782      0.000    1.89e+04    2.67e+04
x1          3986.1805   2037.970      1.956      0.056    -103.305    8075.666
x2          2.454e+04   2197.020     11.171      0.000    2.01e+04     2.9e+04
x3           210.7282   2106.151      0.100      0.921   -4015.572    4437.029
==============================================================================
Omnibus:                       25.342   Durbin-Watson:                   1.724
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               41.817
Skew:                           1.539   Prob(JB):                     8.31e-10
Kurtosis:                       5.907   Cond. No.                         1.68
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
RMSE: 13945.24774047848
MAE: 10187.354792007607
```

Figure 21: OLS Regression Results

With 56 original records, the OLS result shows that R^2 is 0.77, meaning 77.7% of the variance has been explained, which is not a low result. F-statistic equals to 60.42 with the corresponding p-value equals to 5.82e-17, a small value to assure a statistically significant model. What is more, the Omnibus test, which evaluates the normality assumption of the residuals, is reported at 25.342 and the associated p-value (Prob(Omnibus)) is 0.000. In addition, with a value of Jarque-Bera statistics is 41.817 and Prob(JB) is 8.31e-10. A p-value less than 0.05, there is significant evidence to reject the null hypothesis of normality, indicating the residuals may depart significantly from a normal distribution. Later on, the Durbin-Watson statistic tests for autocorrelation in the residuals is 1.724, which is between 1 and 3 assuming that there is no correlation between the residuals.

Additionally, a right-skew value of 4.65 is the result of the asymmetry of the residuals' distribution. Values greater than -2 or +2 suggest substantial non-normality (Hair et al., 2022, p. 66). The kurtosis equals to 34.321, this means there is a high peak in the center of the residual distribution as well as heavy tails in comparison with a normal distribution. Potential multicollinearity can be the repercussion of the sensitivity of the regression coefficients to data changes, which is measured by a higher condition number. Not to mention, the condition number is only 1.68, a low value which might not lead to severe multicollinearity.

In terms of regression coefficients, x_1, x_2 and x_3 are Charging Stations Count, Electricity Price, Gasoline Prices respectively. Of which, x_1 and x_3 have P>|t| of 0.056 and 0.921, which do not meet the criterion of $p < 0.05$. That means these two independent variables have almost no impact on the dependent variable, the null hypothesis can not be rejected. Based on previous research, this result is unexpected so that the model will be reassessed with further data improvements.

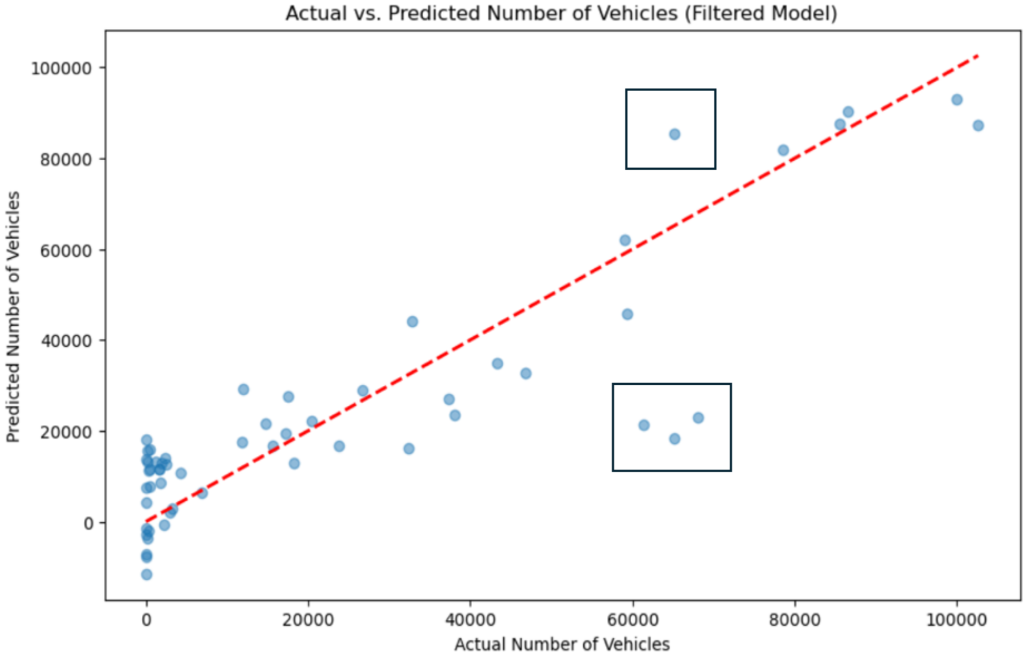
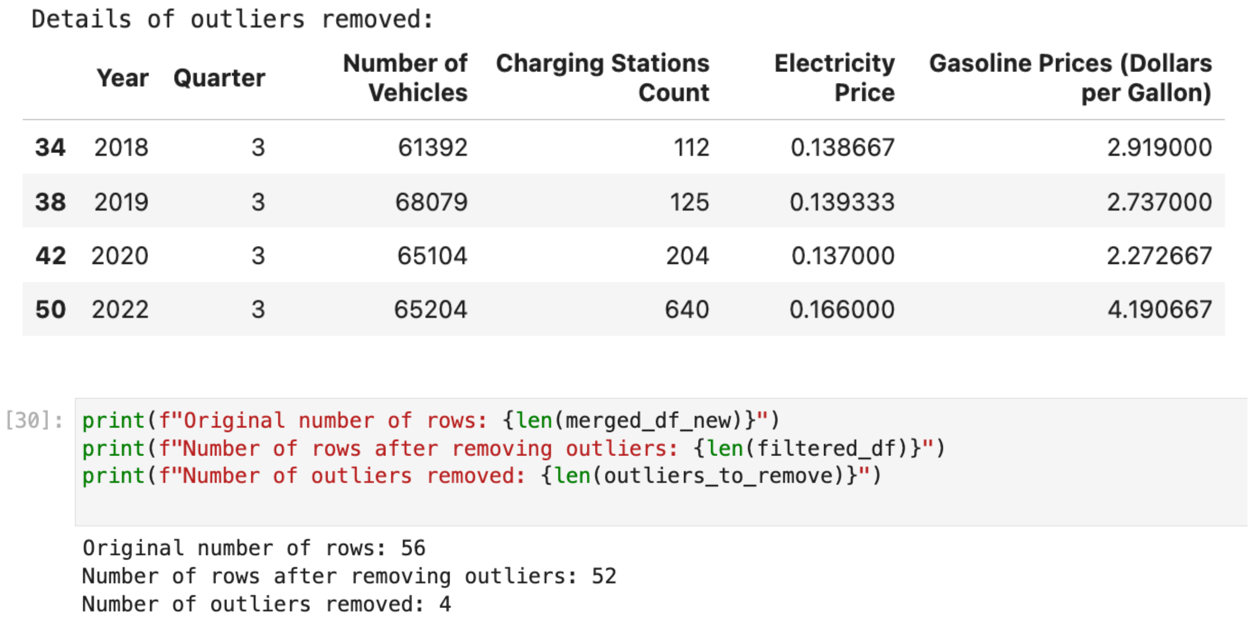


Figure 22: Graph for Actual and Predicted Number of EVs

As can be seen in the Figure 21, there are four highlighted points that lie far from the fitting line, values ranging from 60,000 to 80,000 EVs. This could possibly be the reason leading to high RMSE and MAE and other extreme indicators as mentioned. Therefore, the solution to figure out the cause and improve the model performance is to test removing these four outliers. Firstly, they will be determined based on their value range shown in the graph:



```
Details of outliers removed:
```

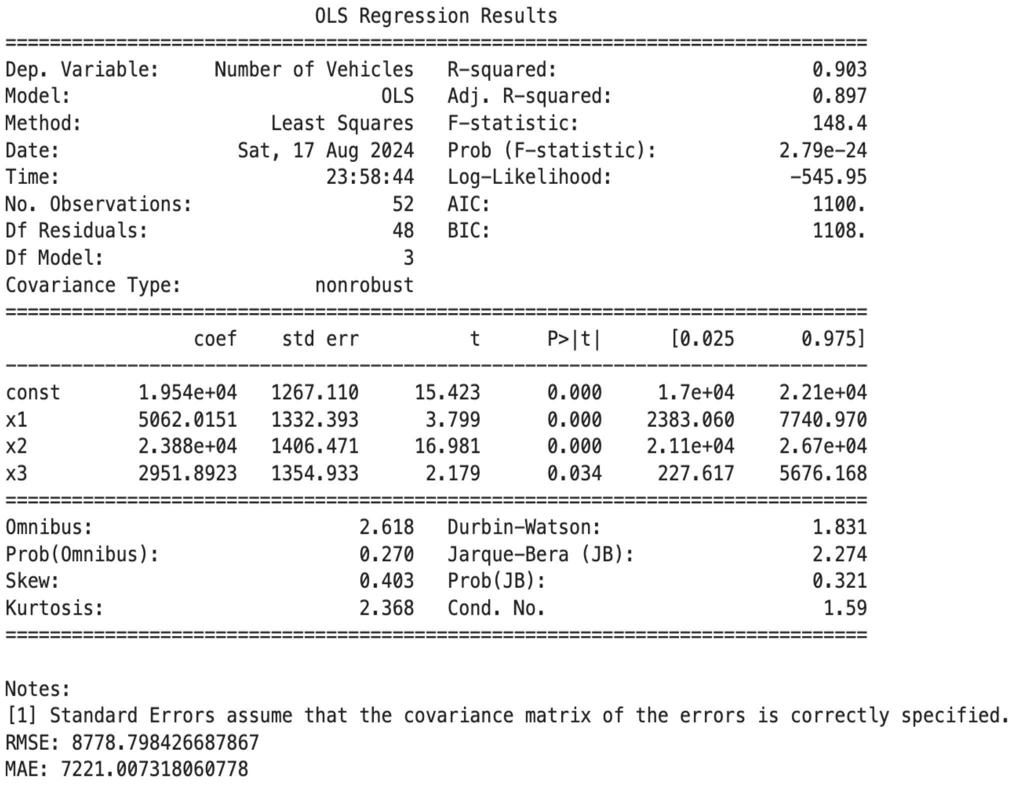
	Year	Quarter	Number of Vehicles	Charging Stations Count	Electricity Price	Gasoline Prices (Dollars per Gallon)
34	2018	3	61392	112	0.138667	2.919000
38	2019	3	68079	125	0.139333	2.737000
42	2020	3	65104	204	0.137000	2.272667
50	2022	3	65204	640	0.166000	4.190667

```
[30]: print(f"Original number of rows: {len(merged_df_new)}")
      print(f"Number of rows after removing outliers: {len(filtered_df)}")
      print(f"Number of outliers removed: {len(outliers_to_remove)}")
```

```
Original number of rows: 56
Number of rows after removing outliers: 52
Number of outliers removed: 4
```

Figure 23: Detail of Outliers

Shown in Figure 22, only 52 records will be kept for further analysis. After removing the outliers, the revised OLS regression result will be shown in Figure 23:



```
                            OLS Regression Results
==============================================================================
Dep. Variable:     Number of Vehicles   R-squared:                       0.903
Model:                            OLS   Adj. R-squared:                  0.897
Method:                 Least Squares   F-statistic:                     148.4
Date:                Sat, 17 Aug 2024   Prob (F-statistic):           2.79e-24
Time:                        23:58:44   Log-Likelihood:                -545.95
No. Observations:                  52   AIC:                             1100.
Df Residuals:                      48   BIC:                             1108.
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       1.954e+04   1267.110     15.423      0.000     1.7e+04    2.21e+04
x1          5062.0151   1332.393      3.799      0.000    2383.060    7740.970
x2          2.388e+04   1406.471     16.981      0.000    2.11e+04    2.67e+04
x3          2951.8923   1354.933      2.179      0.034     227.617    5676.168
==============================================================================
Omnibus:                        2.618   Durbin-Watson:                   1.831
Prob(Omnibus):                  0.270   Jarque-Bera (JB):                2.274
Skew:                           0.403   Prob(JB):                        0.321
Kurtosis:                       2.368   Cond. No.                         1.59
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
RMSE: 8778.798426687867
MAE: 7221.007318060778
```

Figure 24: OLS Regression Results (After removing outliers)

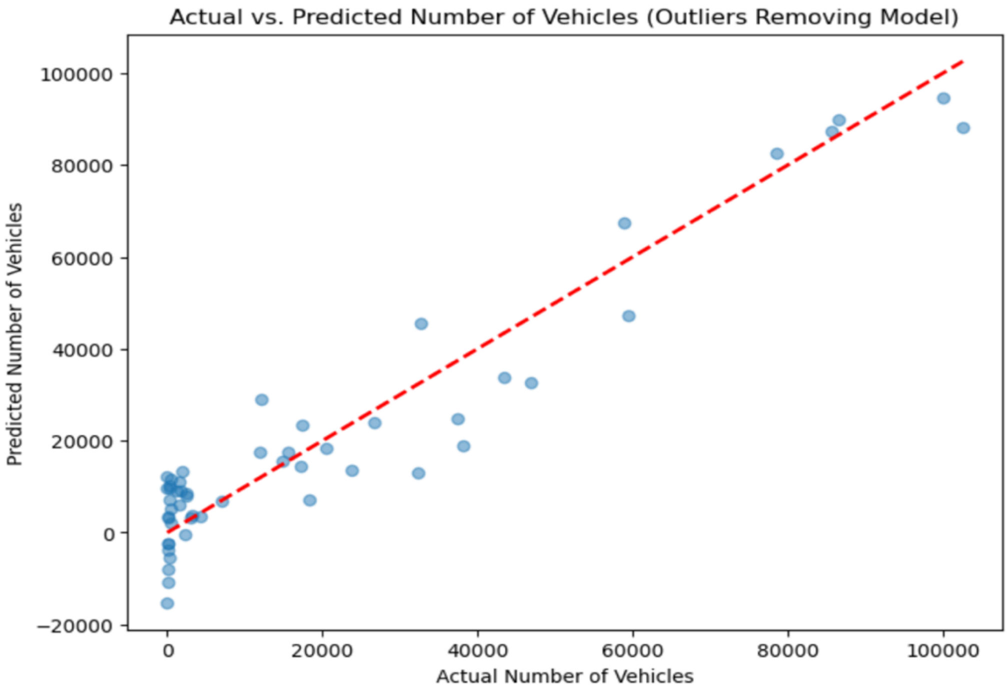


Figure 25: Graph for Actual and Predicted Number of EVs (After removing outliers)

Figure 25 shows that there are several differences compared to the original OLS regression result after fitting the model by removing specific outliers. First of all, value of R^2 has been much improved, reaching at a good fit value of 0.903. This can be interpreted that 90.3% of the data for the variance of dependent variable can be explained by this regression model, most of the data points are now close to the regression line. Technically, this could be an optimal statistical model to predict the outcome of EV sales in California.

Analyzing the model summary, with the increase in Log-likelihood, and lower AIC (1100.) and BIC (1108.) indicative indicators, the model has been much improved. The model is considered highly statistically significant when F-statistic is now 148.4 with a small Prob (F-statistic) = 2.79e-24. Therefore, the null hypothesis can be rejected, all three independent variables are significantly contributing to explain the variance in the dependent variable.

Another point worth noting is regression coefficients, which have also been improved due to the outliers removal. A common threshold of the P-value is 0.05. If p-value (P>|t|) associated with the t-test is less than 0.05, the null hypothesis is not rejected and accepts the alternative hypothesis H_a. All variables are statistically significant at the 5% level and there exists a relationship between the variables. Moreover, [0.025 and 0.975] are both measurements of values of our coefficients within 95% of our data, or within two standard deviations. Outside of these values can generally be considered outliers, which did not happen in this case.

	Indicators	**Before removing outliers**	**After removing outliers**
Model Summary	Number of Observations	56	52
	R^2	0.777	0.903
	Adjusted R^2	0.764	0.897
	F-statistic	60.42	148.4
	Prob (F-statistic)	5.82e-17	2.79e-24
	Log-Likelihood	-613.86	-545.95
	AIC	1236.	1100.
	BIC	1244.	1108.
Regression Coefficients	Const (intercept)	*Coefficient: 2.279e+04* *P>|t|: 0.000*	*Coefficient: 1.954e+04* *P>|t|: 0.000*
	x1	*Coefficient: 3986.1805* *P>|t|: 0.056*	*Coefficient: 5062.0151* *P>|t|: 0.000*
	x2	*Coefficient: 2.454e+04* *P>|t|: 0.000*	*Coefficient: 2.388e+04* *P>|t|: 0.000*
	x3	*Coefficient: 210.7282* *P>|t|: 0.921*	*Coefficient: 2951.8923* *P>|t|: 0.034*

Diagnostic Statistics	Omnibus	25.342	2.618
	Prob (Omnibus)	0.000	0.270
	Durbin-Watson	1.724	1.831
	Jarque-Bera (JB)	41.817	2.274
	Prob (JB)	8.31e-10	0.321
	Skew	1.539	0.403
	Kurtosis	5.907	2.368
	Cond. No.	1.68	1.59
Model Fit Indicators	RMSE	13945.2477	8778.7984
	MAE	10187.3548	7221.0073

Figure 26: Result difference prior to and after outliers removal

On top of that, Diagnostic Statistics has also been changed in a more positive direction. The Omnibus index (2.618) with Prob(Omnibus) = 0.270 is at the non-significant level, although it was previously at 0.000 in the previous model, closer to 1 to reach residual normality. There is no evidence to reject H_0, as well as to conclude that the residuals are not normally distributed. Thus, residuals might have a normal distribution.

The Durbin-Watson statistical test for autocorrelation on the residuals has changed from 1.724 to 1.831, a value that is closer to 2. There is not much evidence of autocorrelation in the residuals due to the outliers removal. Furthermore, a large value of Jarque-Bera test shows that the errors are not normally distributed. Moving from 8.31e-10 to 0.321, the p-value is currently greater than 0.05. Using both methods, it can be confirmed that there is no significant evidence against the null hypothesis of normality. There is a huge change for the Skew (from 1.539 to 0.403) and Kurtosis (from 5.907 to 2.368). Skew value close to zero is more preferable to have normal residual distribution. Kurtosis is lower than 3, residuals are distributed normally with slightly heavier tails. Lastly, the condition number of 1.59 is a bit lower but still remains a good result for strengthening a non-multicollinearity model. Based on the criteria that the smaller the RMSE and MAE, the more accurate the model is. It appears that these two forecast values have reduced the difference between the actual model and the predicted model.

Applying the coefficients into the regression model:

$$\mathbf{EV\_sales = 19540 + 5062.0151 * Charging\ Stations + 23880 * Electricity\ Price}$$
$$\mathbf{+\ 2951.8923 * Gasoline\ Price} + \boldsymbol{\varepsilon}$$

This regression equation shows that: with each coefficient having a statistical significance p-value < 0.05, each independent variable has a significant effect on the dependent variable EV_sales. As all coefficients are positive, this prediction model predicts that, assuming other factors remain constant, when an additional charging station is newly built, the number of EVs sold is expected to increase by about 5062 units. Likewise, for every $1 increase in electricity prices, the number of EVs sold in California is expected to increase by about 23880 units. Similarly, for every $1 increase in gasoline prices, 2952 units of EVs are expected to be sold in California.

To validate the regression forecasting model, the residual is calculated by subtracting the actual value of the data point from the predicted value of that data point. If the error term ε_i is smaller, the relationship between the independent and the dependent variables is larger and vice versa.

$$residuals = \ actual\ y(y_i) - \ predicted\ y(\hat{y}_i)$$

Overall, there is a normal distribution of residual data points since they are concentrated mainly near the red reference line and are relatively symmetric from the origin data points. However, the residuals might have heavier tails or be skewed. The histogram in Figure 28 also shows a relatively symmetric distribution around zero and a right-skewed distribution, with some outliers having large deviations due to the impact of generating data through k-NN imputation. Although there are some points at the tail that are slightly off, the general regression model is still considered appropriate to forecast the future EV sales. More attention should be paid to the outliers in the model.

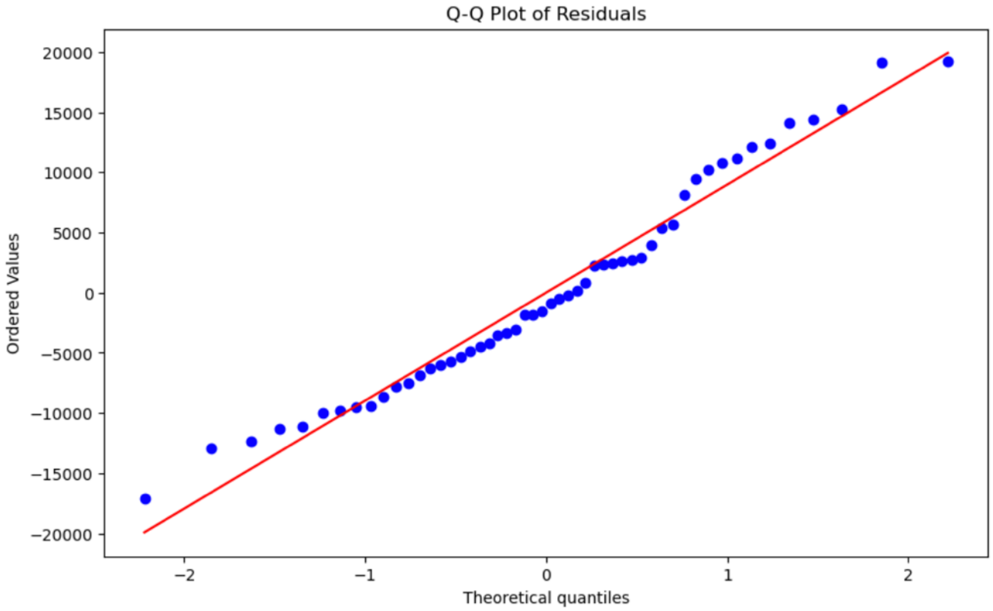


Figure 27: Q-Q Plot for Residuals

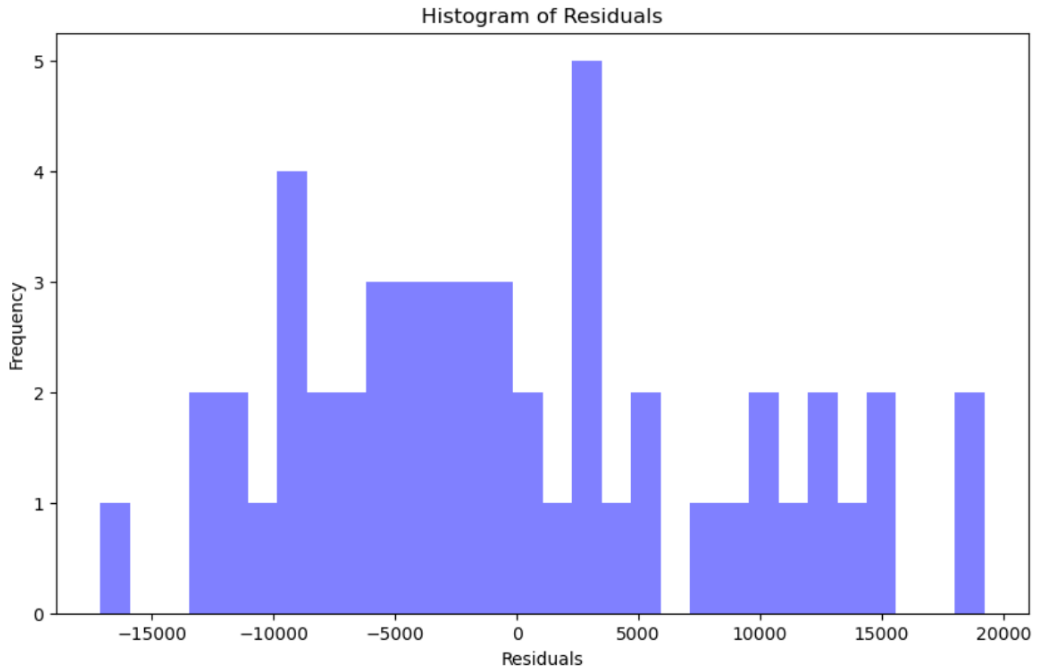


Figure 28: Histogram of Residuals

After conducting the OLS regression analysis, what is worth mentioning is that after handling specific outliers, the important coefficients to test the reliability of the model have been improved and have the ability to predict the number of cars sold well based on the changes of Gasoline Price, Electricity Price and Charging Stations Counts. Most importantly, the model does not seriously violate the basic assumptions of linear regression even though the residuals are slightly right-skewed.

4. Discussion

There are some lingering questions about the decline in EV sales in recent years. To answer this question, the average retail price of an EV is considered to be significantly higher from that of gasoline- and diesel-powered vehicles. The Los Angeles Times notes that high interest rates have also created a systematic barrier for buyers to switch to EVs. Public charging stations are few and far between, and charging station electricity prices are soaring, making EVs less attractive to consumers. EV owners also have to queue to charge their vehicles at public charging stations.

Results from the final regression model are consistent with those from previous studies and have confirmed that the number of charging stations and energy prices have an impact on EV sales. Based on this model: if the California state government better regulates financial incentives while adjusting external factors more efficiently, creating the premise for more buyers to access EVs, then the trend of EV ownership would continue to keep the upward trend in the long term. The idea of raising the pricing incentives has been recommended in other international studies: Malvik et al., (2013); Sierzchula et al., (2014)…

The positive correlation between EV sales and charging stations counts suggests that the installation of additional charging stations is beneficial for promoting EV sales and increasing user access. Currently, there are three types of electric charging stations: Level 1 and Level 2 are typically installed at home or work, whereas DC fast charging stations are intended for long-distance travel. Although Level 2 charging stations are popular in California, the growing number of EVs requires the reduction of charging time, which the installation of DC fast charging stations could be a possible solution.

In particular, after 2035, if California successfully implements the selling 100% ZEV market, the growth of the number of stations is required to meet the dynamic charging demand of citizens. "Public chargers must be built at an unprecedented pace to meet the target in less than 7 years, and then doubled to 2 million in 2035. The high cost - $120,000 or more for one fast charger - is just one obstacle" by CalMatters (2024). The Biden administration is pouring $168.5 million into charging projects in California. The government should continue implementing greater support programs and make stations available in rural areas, to alleviate consumers' concerns about the cost and convenience of charging their vehicles. This has been similarly concluded in research for developed countries such as China (Wang et al., 2017) or Sweden (Egnér and Trosvik, 2018).

Both electricity and gasoline prices are positively correlated with EV sales. Rising gas prices are prominently displayed along roadways. Any changes in the price could stimulate consumers to seek alternatives to reduce their dependence on gasoline. Along with strong developments in technology, EVs will become more attractive to potential buyers.

On the contrary, electricity prices will be less attentive. If ZEV sales reach 100% by 2035, electricity prices will no longer be a major concern for consumers when they are forced to use it. However, to encourage those who are still using conventional vehicles to switch to EVs, the government should consider adjusting electricity prices to stabilize them to reduce consumers' concerns about costs.

V. Conclusion

The purpose of this project is to explore the long-term trend of EV sales and its relationship with other factors across California counties. EDA and Correlation analysis has been utilized to generally gain insights of the extended dataset compared with previous empirical studies prior to implying linear regression analysis.

With an accuracy rate of 90.3%, the linear regression model predicts that the number of EVs sold would suffer a positive direct impact from charging station counts, gasoline and electricity prices. The empirical results have several important implications for the strategies to support California to gradually reach its ambitious goal by 2035. Consequently, a failure to significantly expand the network of stations for alternative fuels and energy price adjustment would significantly hamper the EV adoption of California in coming years. While this project did not directly investigate financial purchase incentives as an effect, previous studies state that this is the most effective solution in stimulating individual adoption of EVs, as consumers prefer to have a lower initial purchasing price (Hardman et al., 2017), (Gong et al., 2020). California should focus more on distributing incentives more equitably to lower-income people to encourage EV adoption. This will also be the premise for developing and expanding this in-depth research project.

VI. Limitations and Recommendations for Future Research

Despite its significant contributions, this project presents significant limitations that necessitate solutions. Collected records on EVs are generally limited since this market has only truly taken off in recent years. Furthermore, the project includes k-NN Imputation to optimize filling in missing values due to the lack of data in the original dataset. The assumption that close observations in the feature space have similar missing values can lead to inaccurate estimation of missing values, thereby reducing the accuracy of the analysis results compared to the actual data. In addition, the data used include year 2019-2021 when the COVID-19 pandemic was a serious factor that impacted EV sales. Its influence has not been assessed thoroughly to determine whether it is a temporary or long-lasting impact.

A potential future research would consider incorporating other socio-economic factors such as financial incentives or income, which might have undisclosed effects in this project. With the aim of generalizing the regression model for other cities, the future solution would be collecting a qualified dataset with diverse variables and covering extra testing on significant social impacts such as the COVID-19 pandemic.

References

Datasets:

(1) Total Number of Electric Vehicles:
Commission, C. E. (n.d.). *ZEV and Infrastructure Stats Data*. California Energy Commission. https://www.energy.ca.gov/files/zev-and-infrastructure-stats-data

(2) Number of Alternative Fueling Stations in California:
Alternative Fuels Data Center: Alternative Fueling Station Locator. (2024). Energy.gov. https://afdc.energy.gov/stations#/analyze?country=US®ion=US-CA&tab=location

(3) Gasoline Price in U.S.:
U.S. Energy Information Administration. (2023, October 23). *U.S. All Grades All Formulations Retail Gasoline Prices (Dollars per Gallon)*.
https://www.eia.gov/dnav/pet/hist/LeafHandler.ashx?n=pet&s=emm_epm0_pte_nus_dpg&f=m

(4) Electricity Price in U.S.:
U.S. Bureau of Labor Statistics. (1978, November 1). *Electricity Per KWH in U.S. City Average*. FRED, Federal Reserve Bank of St. Louis.
https://fred.stlouisfed.org/series/APU000072610

1. California Air Resources Board. (n.d.). *Cars and light trucks are going zero: Frequently asked questions*. Retrieved from https://ww2.arb.ca.gov/resources/documents/cars-and-light-trucks-are-going-zero-frequently-asked-questions

2. Chakrabarti, A., & Ghosh, J. K. (n.d.). AIC, BIC and recent advances in model selection. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

3. Chantal Larose and Daniel Larose, *Data Science Using Python and R*, Wiley, 2019.

4. Coffman, M., Bernstein, P., & Wee, S. (2016). Electric vehicles revisited: A review of factors that affect adoption. *Transportation, 43*(1), 79-93. https://doi.org/10.1080/01441647.2016.1217282

5. *Detecting multicollinearity using variance inflation factors* | STAT 462. (n.d.). https://online.stat.psu.edu/stat462/node/180/

6. Egnér, F., & Trosvik, L. (2018). Electric vehicle adoption in Sweden and the impact of local policy instruments. *Energy Policy*, *121*, 584–596. https://doi.org/10.1016/j.enpol.2018.06.040

7. Federal Reserve Bank of St. Louis. (n.d.). *Gasoline, all types, per gallon/3.785 liters in U.S. city average*. Retrieved from https://fred.stlouisfed.org/series/APU000072610

8. *First EV sales decline in a decade; hiccup or lasting trend?* (2024, February 15). Los Angeles Times. https://www.latimes.com/california/newsletter/2024-02-15/essential-california-ev-sales-essential-california

9. Frost, J. (2018). *Interpreting Correlation Coefficients*. Statistics by Jim. https://statisticsbyjim.com/basics/correlations/

10. Gallagher, K. S., & Muehlegger, E. (2011). Giving green to get green? Incentives and consumer adoption of hybrid vehicle technology. *Journal of Environmental Economics and Management*, *61*(1), 1–15. https://doi.org/10.1016/j.jeem.2010.05.004

11. GDPR. (2018). *General data protection regulation (GDPR)*. General Data Protection Regulation (GDPR). https://gdpr-info.eu/

12. GeeksforGeeks. (2021, July 22). *What is Exploratory Data Analysis ?* GeeksforGeeks. https://www.geeksforgeeks.org/what-is-exploratory-data-analysis/

13. GOV.UK. (2020). *Data Ethics Framework*. https://assets.publishing.service.gov.uk/media/5f74a4958fa8f5188dad0e99/Data_Ethics_Framework_2020.pdf

14. Hair, J. F., Hult, G. T. M., Ringle, C. M., & Sarstedt, M. (2022). A primer on partial least squares structural equation modeling (PLS-SEM) (3rd ed.). Sage.

15. Hannisdahl, O. H., Malvik, H. V., & Wensaas, G. B. (2013). The future is electric! The EV revolution in Norway — Explanations and lessons learned. In 2013 World Electric Vehicle Symposium and Exhibition (EVS27) (pp. 1-13). IEEE. https://doi.org/10.1109/EVS.2013.6914921

16. HAYES, A. (2022, January 4). *Multicollinearity*. Investopedia. https://www.investopedia.com/terms/m/multicollinearity.asp#:~:text=Multicollinearity%20is%20a%20statistical%20concept

17. Higueras-Castillo, E., Guillén, A., Herrera, L. J., & Liébana-Cabanillas, F. (2020). Adoption of electric vehicles: Which factors are really important? *International Journal of Sustainable Transportation, 15*(10), 799–813. https://doi.org/10.1080/15568318.2020.1818330

18. Javadnejad, F., Jahanbakhsh, M., Pinto, C. A., & et al. (2023). Analyzing incentives and barriers to electric vehicle adoption in the United States. *Environment Systems and Decisions*. https://doi.org/10.1007/s10669-023-09958-3

19. Javid, R. J., & Nejat, A. (n.d.). A comprehensive model of regional electric vehicle adoption and penetration. https://doi.org/10.1016/j.tranpol.2016.11.003

20. Jiang, L., & Gao, X. (2024). Will changes in charging and gasoline prices affect electric vehicle sales? Evidence from China. *Environmental Science and Pollution Research, 31*, 3123–3133. https://doi.org/10.1007/s11356-023-31389-5

21. Komorowski, M., Marshall, D. C., Salciccioli, J. D., & Crutain, Y. (2016). Exploratory Data Analysis. *Secondary Analysis of Electronic Health Records*, 185–203. springer. https://doi.org/10.1007/978-3-319-43742-2_15

22. Ku, A. L., & Graham, J. D. (2022). Is California's Electric Vehicle Rebate Regressive? A Distributional Analysis. https://doi.org/10.1017/bca.2022.2

23. Kumar, S., Singh, V., & Goel, R. (2024). Strategic forecasting for electric vehicle sales: A cutting edge holistic model leveraging key factors and machine learning technique. Journal of The Institution of Engineers (India): Series C. https://doi.org/10.1007/s40890-024-00213-1

24. Lazo, A. (2024, July 16). *California needs a million EV charging stations — but that's "unlikely" and "unrealistic."* CalMatters; CalMatters. https://calmatters.org/environment/climate-change/2024/07/california-electric-car-chargers-unrealistic-goals/

25. *Machine Learning Guide for Oil and Gas Using Python | ScienceDirect*. (n.d.). Www.sciencedirect.com. https://www.sciencedirect.com/book/9780128219294/machine-learning-guide-for-oil-and-gas-using-python

26. Mays, K. (2024, July 18). California EV sales decline again, after years of rapid expansion. *Los Angeles Times*. Retrieved from https://www.latimes.com/environment/story/2024-07-18/california-ev-sales-decline-again

27. Newcastle University. (2023). *Numeracy, Maths and Statistics - Academic Skills Kit*. Www.ncl.ac.uk. https://www.ncl.ac.uk/webtemplate/ask-assets/external/maths-resources/statistics/regression-and-correlation/residuals.html

28. Perraillon, M. (n.d.). *Week 14: Choosing models*. Retrieved September 1, 2024, from https://clas.ucdenver.edu/marcelo-perraillon/sites/default/files/attached-files/week_14_selection_v08_0.pdf

29. Qiu, Y. Q., Zhou, P., & Sun, H. C. (2019). Assessing the effectiveness of city-level electric vehicle policies in China. *Energy Policy, 130*, 22-31. https://doi.org/10.1016/j.enpol.2019.03.052

30. *Rebate Statistics | Clean Vehicle Rebate Project*. (n.d.). Cleanvehiclerebate.org. https://cleanvehiclerebate.org/en/rebate-statistics

31. *Regression - Sociology 3112 - Department of Sociology - The University of utah*. (n.d.). Soc.utah.edu. https://soc.utah.edu/sociology3112/regression.php#:~:text=Ordinary%20least%20squares%20(OLS)%20regression%3A%20a%20technique%20in%20which

32. Roy, A., & Law, M. (2022). Examining spatial disparities in electric vehicle charging station placements using machine learning. *Sustainable Cities and Society, 86*, 104123. https://doi.org/10.1016/j.scs.2022.103978

33. Sierzchula, W., Bakker, S., Maat, K., & van Wee, B. (2017). Analysis of the electric vehicles adoption over the United States. *Sustainable Cities and Society, 28*, 177-185. https://doi.org/10.1016/j.trpro.2017.03.027

34. Sierzchula, W., Bakker, S., Maat, K., & van Wee, B. (2017). The influence of financial incentives and other socio-economic factors on electric vehicle adoption. *Transportation Research Part A: Policy and Practice, 106*, 234-250. https://doi.org/10.1016/j.enpol.2014.01.043

35. Tang, L., & Sun, J. (2019). Predict the sales of New-energy Vehicle using linear regression analysis. *E3S Web of Conferences*, *118*, 02076. https://doi.org/10.1051/e3sconf/201911802076

36. The White House. (2021, January 27). *National Climate Task Force*. The White House. https://www.whitehouse.gov/climate/

37. U.S. Energy Information Administration. (2023). Retrieved from https://www.eia.gov/todayinenergy/detail.php?id=61082

38. UNFCCC. (2015). *The Paris Agreement*. United Nations Climate Change; United Nations. https://unfccc.int/process-and-meetings/the-paris-agreement

39. Wee, S., Coffman, M., & La Croix, S. (2018). Do electric vehicle incentives matter? Evidence from the 50 U.S. states. *Research Policy, 47*(9), 1601-1610. https://doi.org/10.1016/j.respol.2018.05.003