Automotive Market Trend Prediction and Adoption of EV Technologies

MSc Research Project
Data Analytics

Arun Das Mohandas
Student ID: 23136766

School of Computing
National College of Ireland

Supervisor: Mr. Hicham Rifai

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Arun Das Mohandas
Student ID:	23136766
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Mr. Hicham Rifai
Submission Due Date:	12/12/2024
Project Title:	Automotive Market Trend Prediction and Adoption of EV Technologies
Word Count:	9500
Page Count:	27

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th December 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Automotive Market Trend Prediction and Adoption of EV Technologies

Arun Das Mohandas

23136766

Abstract

The research extensively studies the trends in the automobile industry in past decades and its shift toward electric vehicle technologies. It will unravel insights about how the automobile market is adapting to the change of the introduction of EVs. It will discuss Electric Vehicles, and plug-in Hybrids, and compare their sales growth with that of traditional non-electric vehicles. The research will gather sales figures for these automotive technologies from different automotive brands and will analyze the figures to build an understanding of where the market is heading. The research will also look into consumer acceptance of the new technologies, the environmental impact, and the market sustainability to understand the change in customer perception. To reach there we will delve into previous studies on the industries and extract, compare, and contrast on the topic to develop a better understanding of the subject. The key task that will be performed here will be employing methodologies of deep learning on time series datasets gathered from the automobile industries to understand the direction in which the market is heading, ie if the market is adopting Electric vehicles over plug-in hybrid vehicles and fuel engine vehicles. By doing this the automobile industry will have a viable long-term plan and course of action to follow and lead in the current competitive automotive market.

Keywords: EV Adoption, Electric Vehicles, Hybrids Electric Vehicles, Machine Learning, Time series Analysis Prediction, Deep Learning.

1. Introduction

Over the past decade, the automotive industry has drastically impacted the preferred mode of transportation with a noticeable shift towards electric vehicles.(Fernandez, 2021) Some leading companies like Tesla have made a mark for themselves as EV market pioneers and have shown other manufacturers the possibilities of EV technology. Tesla has launched revolutionary innovations to the market that include not just EVs, but also autopilot and self-drive capabilities that improve vehicle safety and security. The brand was founded in 2003 to make Electric cars the mainstream technology replacing the long-running trend of internal combustion engines. They chose a strategy based on a data-driven approach and they innovated based on the inference from this data. Their technology with adequate infrastructure and setup has been shown to help decrease emissions on a global scale. This has caused governments from various countries to back the change worldwide. This motivates governments to create EV-friendly policies, tax reductions, subsidies, and strong emissions-reduction goals which is essentially speeding up EV adoption.

(Adnan et al., 2017) Another key aspect of EV adoption is customer sentiments towards them. It is shaped by their perception, emotions, and awareness of environmental issues. It is seen that even though EVs do bring improvements in the emissions level across the world and reduce the use of fuel, The adoption is significantly less. Customer Behavior is what companies should be focusing on while setting up the framework as it will inspire people to adopt the new technologies. Furthermore, there exist issues like range anxiety, a lesser number of charging stations across the country, maintenance issues with EVs, and the costly unreliable batteries that these cars use. However recent events and advancements in battery technologies are improving the practicality of these batteries. This solid improvement pace may fuel customer interest in EVs and their adoption.

(Keohane et al., 2024)Some automakers haven't fully embraced the shift to vehicles like Toyota has done differently with their focus on hybrid electric vehicles (HEVs) and plug-in hybrid electric vehicles (PHEVs). Unlike EVs that run solely on electricity hybrids blend a conventional petrol engine with an electric motor to provide a dual power source with distinctive advantages. They utilize braking technology where the petrol engine aids, in recharging the electric battery that supplies energy to the electric motor. This system doesn't just enhance fuel efficiency; it also lowers emissions when compared to gasoline vehicles. Despite Toyota not embracing the EV technology and adopting the market trend it is still managing to make good sales figures. As of the past year 2024, Toyota has managed to be ranked second in the top-selling car brands in Europe. This raises the question that does Hybrids/Plugin-Hybrids have the potential to outpace EV sales figures. The success has shown the potential of hybrid technology and indicates how a quick EV adoption might not be the right way going forward. The brand strategy needs recognition for the same reason and this research will look at the global data to verify, analyze, and provide inference on how much of this is true.

The research will be mainly focus on the same, that is to explore the dynamics between the customer, the brand, and its technologies which are electric and hybrid engine technologies. The study will delve into the past sales figures of EVs vs Hybrids from different brands globally and compare both to derive a conclusion about the future of automobile technology. We will be adapting deep learning for the task and various methodologies from machine learning to visualize, analyze, cleanse, and develop metrics from the time series dataset opted for the research. Therefore, by analyzing the main factors in the scenario driving the market adoption, the study will seek to provide clarity on the future trajectory of automotive innovation and the roles these technologies are playing in it.

1.1 Research Question and Objectives

What is the future of EV adoption in the automobile market, and what EV technology will be adopted in the near future?

2. Literature Review of Related Work

2.1 Introduction to the research topic

The adoption of vehicles using alternative green energies has been a major concern for the consumer and government in recent years. A number of causes in recent years which include global warming, the depletion of fossil fuels, the massive urban population growth, and the increased pollution, have affected the shift in perception of individuals. These problems have led to the creation of engines that power from sustainable sources and are more efficient. EVs, hybrid EVs, fuel cell vehicles, plug-in hybrids, and other electric car architectures are being developed to compete with traditional combustion engines. EV and hybrid technology production and their acceptance have increased significantly in recent years. However, the proportion of EV sales is small in proportion compared to the overall automobile market. These cars must be more practical and interesting to buyers in order for them to be adopted widely. Power efficiency, dynamic reaction, cost-effective service, cooling system, improved range, and electrical accessibility are just a few of the features that the technology should be improving in the near future.

2.2 Automotive Companies Examples – Tesla:

(Fernandez, 2021) The study represents Telsa as the one company that revolutionized the EV market by building up a model of business that combines artificial intelligence, data, technology, and innovation driven by them. Tesla developed technology something called a modular EV architecture which comprises lithium-ion batteries, powerful motors, and its ecosystem. Tesla is overcoming its market troubles by pairing with brands like Panasonic and improving its battery technology. The AI based on cloud-based infrastructure continuously learns from situations and improves the ecosystem. Tesla's sales dropped only post-COVID by about 70%, which was recovered by introducing its online sales model. There are supply chain disruptions. However, the articles show no reduction in sales of Tesla even when people are not completely satisfied with the EV technology.

2.3 Automotive Companies Examples- Ford Motors:

(Boudette, 2024)The study talks about how Ford automobile industry entered the electric vehicle industry with its new EV vehicle Ford F150, which was praised for its many features and advancements. When there is a noticeable decline in range during cold weather, the consumer has had issues. This along with the fact that the real-world range is different from the claimed range gives the customer range anxiety and undermines customer confidence in the brand.

The battery-operated vehicle sales were taking off by the end of 2022, that is the sales rose to 46% in early 2023. This exceeded 1 million vehicles and out of this 7% was the range of new cars. The study points out how the sales declined in the late 2023 months and the automakers were cautious about it. California, which was one of the biggest markets for EV sales noticed a drop in EV sales at the end of the year. Altogether it demotivated Ford to slow down on EV investments.

2.4 Automotive Companies Examples - Toyota Company:

(Keohane et al., 2024)The study showed how the recent huge transformation in the automotive industry highlights the fact that the relationship between hybrids and battery-electric vehicles has been changing. Electric cars, thus, underperformed in comparison to hybrid cars in recent. Toyota was earlier strongly criticized for its hybrid and ICE engine investments and roadmap. The company received negative feedback from shareholders and environmentalists who think the company must have adopted BEVs and ditched their ICE and hybrid platforms. The firm conveyed that consumers find fully electric vehicles to be high on cost and that the lack of charging infrastructure throughout the country is concerning for the public.

The recent figures showed a declining trend of sales for brands that heavily invested in EVs. (Sigal, 2024) The data show the BEVs for the US and European markets are heading to a less glacial transition than Toyota predicted. Toyota's profitable hybrid vehicle sales have been so great that they have only been surpassed in recent years. They relied on Plug-in hybrids (PHEVs) and conventional hybrid cars which provide different types of electric driving cars with normal engines that cost less and most importantly the drivers are not burdened with monotonous practices. Both the convenience and cost-saving features have brought last-engine industrial strategies that have seen car manufacturers like General Motors re-introducing PHEVs.

EIA, Federal Energy Information Administration predicts that sales of PHEVs are expected to be exponentially increased by 2023. By that year Electric vehicles contributed 16% of the whole car sales in the United States last year and still were a point higher at 17.9 by the end of 2023. In Europe, Toyota is second to Europe's best-selling brands in the last year in terms of delivery of over eighty-two thousand eight hundred and forty-eight light vehicles. Some others who in the past correctly predicted the quick switchover to BEVs was Morgan Stanley's (GB) analyst Adam Jonas who is among those who see Toyota's way to be very practical. This said, though many experts still stress BEVs as the long-time growth-leading segment, one of them underscored the importance of hybrid technology in the near future for the transportation sector which wants to turn green.

2.5 Understanding customer perspective:

(Adnan et al., 2017)Based on the study conducted for the research 'Mainstream customers Driving BEVS and PHEVs cars: A qualitative analysis of responses and evaluation' consumers view EVs and Hybrids EVs differently. Factors such as price, range, and practicality drive the customer perspective. Customers compared EVs to internal combustion engines (ICE) because of the developing charging infrastructure, limited range, and perceived fun, functionality, and comfort. They pointed out how they have to sacrifice on features such as heating or entertainment to save energy in EVs. This makes the customers think unfavorably poorly of the EVs. What makes it worse is the design language and social identity associated with EV users.

PHEVs have the flexibility of switching to EVs or ICE engines. This helps with the practicality of the customers as they are no longer anxious about range or using features that need power. Hybrids are more cost-efficient, have more range and people opt for the EV mode for short distances.

2.6 The environmental impact of EVs and PHEVs comparison:

(Hawkins et al., 2012)The study investigates the environmental effects caused by the usage of Electric and Hybrid electric cars across all life cycles. This particular case is of great relevance because currently countries are aiming at cutting down emissions and thus the governments will be announcing policies in the near future supporting the more reliable technology. This also makes it clear on how critical it is to approve and conduct all the stages of the environmental impact assessments including all categories of vehicle life, types of emissions and the resources used for energy. It reviewed 51 studies on well-to-wheel analysis which in the latest of the tests focused on the effects of the production of these vehicles. Among the most important findings were the differences recorded in the kWh/km NAC use ranging from 0.10 – 0.24 and the estimated life span of the battery and vehicle as measuring from 150,000 up to 300,000 km. Most of the mentioned above are true regardless of the level of technologies engaged in EVs which GWP (Global warming potential) indicates as the most aggravated sources CO2 emissions. Even good ICEVs were outperformed by EVs if powered by low-carbon sources, but powered by coal electricity were found to have higher SOx emissions than any other type of energy. High-efficiency ICEVs and off-grid hybrid electric vehicles might turn out to be better performers than fuel-powered ICEs.

2.7 Growth in China’s plug-in Electric Vehicle market 2009-2015

(Ou et al., 2017)The study shows how China's plug-in EV market in 2015 witnessed a drastic growth in sales. The PEVs grew by 352% which is around 3.8 lakh units, which in turn made China the largest Plug-in EV market. The key factor that made PEVs prominent here was subsidies put on these vehicles by the Chinese government. It was noticed that the small versions of these vehicles called micro EVs, which were low on cost, dominated the market. This vehicle manufacturing was led by domestic young Automotive firms. The premium once was maintained by foreign automakers. The price-sensitive nature of the Chinese population focused on these mini EVs that drove the market moving into a sustainable future.

The government transitioned EV subsidies through three phases. First, they provided incentives to individual buyers. Secondly, they helped the customers with fraud concerns and thirdly they built robust charging infrastructure across the country which in turn accelerated the EV growth in cities.

2.8 Strategies for EV adoption to be made feasible: Evidence from Beijing

(Zhang and Bai, 2017)The shift to electric vehicles (EVs) is a decisive step in the path toward eco-friendly transportation, originating from various in-depth theoretical approaches. Many scholars have voiced the strategy of the Technology Acceptance Model (TAM) which focuses on functionality and simplicity in the use an EVs can lead to better customer acceptance. This along with better infrastructure for charging and cost-effectiveness motivates customers to opt for EVs(Wang et al., 2022). The Theory of Planned Behavior shows how attitudes, subjective norms, and perceived behavioral control are the factors that determine the intentions of EV adoption (Maichum et al., 2016). At the same time, the Diffusion of Innovations Theory deals with the fact of how the social influence and regulation of incentives guide adoption while things like driving experience, the image a brand creates, and financial subsidies are the dominant factors in this decision process (Bjerkan et al., 2016; Featherman et al., 2021). Likewise, the Value-Belief-Norm Theory showcases the way people perceive environmental values as the driving force for the adoption of electric vehicles given that they are the more environmentally friendly option (Yang et al., 2023). New research emphasizes a multi-dimensional approach that consists of economic, environmental, technological, and socio-cultural factors The urban development views also turn out to be significant, underscoring the roles of the environment and people's travel behaviors in the EV market adoption (Chen and Yang, 2022). This exhaustive methodology is filled with beneficial urban planning tips.

2.9 Example for Time Series Analysis:

("Study and analysis of SARIMA and LSTM in forecasting time series data - ScienceDirect," n.d.) The above literature conveys the advancements that happened in energy forecasting. Mathiyalagan et al.(2017) emphasize on the role of using smart meters for reducing electricity consumption. This enabled analytics to be done on the basis of usage based on time, demand, and tariff rates. This study relied upon time series models for time series analysis. As the dataset in the study involved patterns of trends, seasonality, and noise using time series analysis was crucial. Various models were implemented, some of them including ARIMA, SARIMA, which uses past values to make a forecast, and LSTM, which is a kind of neural network that is recurrent in nature. The LSTM uses memory of long-term data patterns to help with the forecast. LSTM was very important in the study because it could deal with scenarios of energy dynamically. The above models will be used to forecast, manage, and enhance grid availability.

2.10 Conclusion of Literature Review

From the literature review, it is evident that EV adoption does have challenges to tackle over the years. However, it is shown that the active involvement of the government in motivating the public to adopt EV do work. From customer perception what stops from adopting EVs are the issues of maintaining them and the initial cost of buying them. Some examples of successful adoption from China have shown how successful policies can drive EV adoption especially if it is accessible even for individuals on lower wages. We looked into time series models for performing prediction and some examples of how effectively they were implemented in the past. The research will aim at creating and leveraging similar models to keep track of where the automobile market is headed, which will help private companies and governments to make informed decisions.

3. Methodology

3.1 Introduction

The project will rely on some of the key Time series Models to investigate the adoption of patterns for three key categories of vehicles: 1. Battery operated Electric Vehicle (BEV), Plug-in Hybrid Electric Vehicle (PHEV), and Internal Combustion Engine Vehicle (ICE). We will be relying on several series of machine learning models namely ARIMA, SARIMA, ETS, Prophet, and, Long

Short-Term Memory Model networks. What we focus on here to achieve is building these models that could efficiently forecast the trends in the automobile market in adopting vehicle technology that could help brands recognize, plan, and make advancements in those technologies. For example, if the trend of Internal combustion engines is declining, the company will not need to plan ahead for a longer duration of 10 years but make a shorter plan. These models could be useful for different geographical locations across the world.

3.2 Architecture and Framework Overview

The framework, in Figure 3.1, involves absorbing the dataset for sales figures for EVs, Hybrids, and ICEs as well as preprocessing and categorizing the data so that it is clear enough to be trained for the models. There will be various steps involved in this starting with Data selection itself. The selected data is looked upon for data and format issues. The basic statistics of the data will be calculated, i.e. Mean, median, and Standard deviation to get an on how skewed, normalized or spread across is the data. The preprocessed data will be further used for Exploratory Data Analysis where different aspects of the data will be represented utilizing various graphs. This stage will give us more clarity on how the data exists and further clarify the observations from the basic statistics performed.

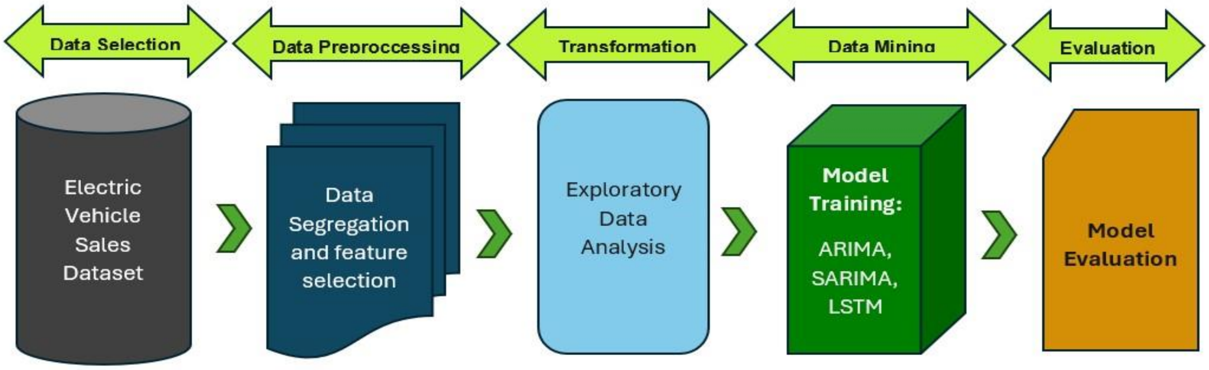


Fig 3.1 Electric Vehicle Sales, Adoption and Prediction framework

Once all the above steps are performed and we have a clarity on dataset we will move onto the next step of model creation where we ill split the data into training and testing data that will aid with the evaluation of these time series models. We ill be focusing on couple of efficient modeling techniques for time series analysis such as ARIMA, SARIMA, Prophet, and LSTM. The forecast result of these models will be thoroughly tested and evaluated on the basis of several metrics which will be discussed ahead.

3.3 Workflow for the Vehicular Technology Forecast

The workflow, in Figure 3.2, for the Electric Vehicle sales forecasting involves two tiers for handling the data. The first deals everything with the data itself and the second tier involves the visualization of the data and output. The first tier is the Business Logic tier which contains the core logic for the forecasting. The process starts with extracting the entire dataset from the source which is a .CSV input here. The data is cleansed for data issues and inconsistent values that may affect the accuracy of the forecasting is removed. Further processes include data transformation and feature selection for regression. The vehicle sales prediction for BEVs, HPEVS, and non-electric vehicle data is performed by separating time series by category and feeding the data to the different prediction models.

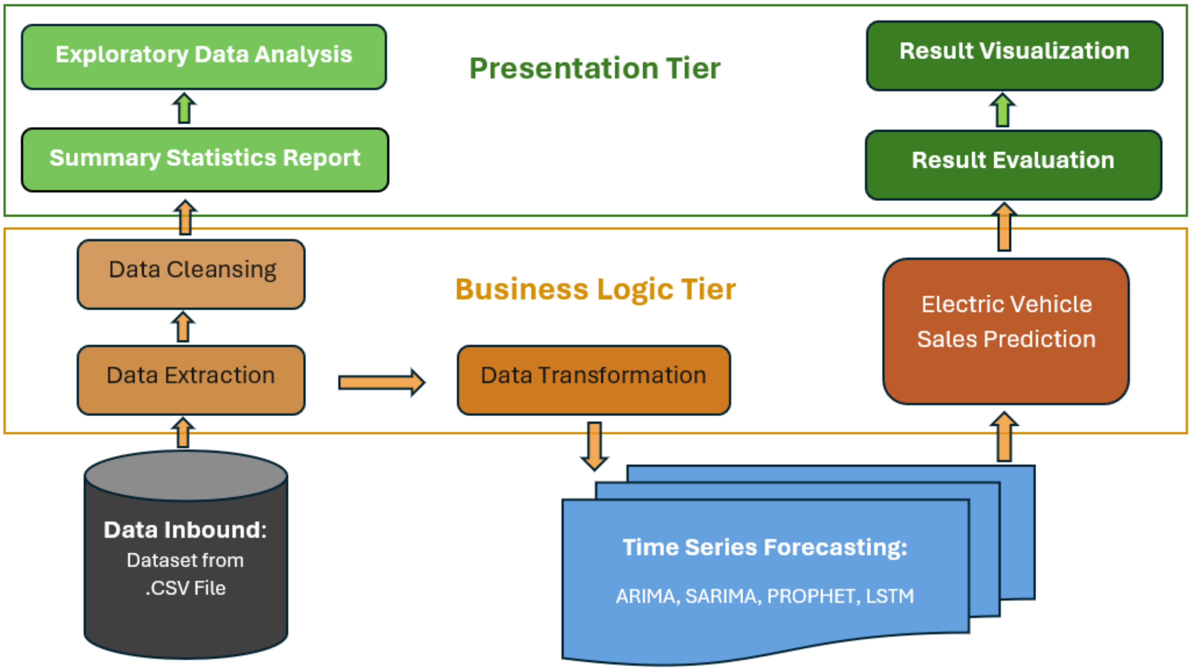


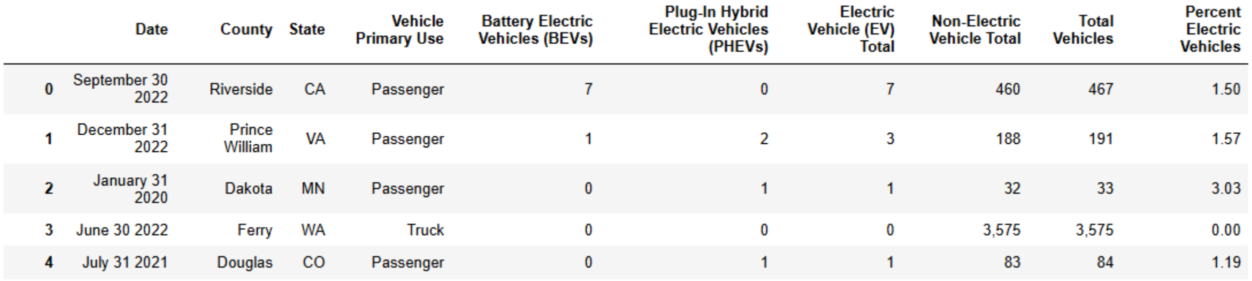
Fig 3.2 Electric Vehicle Sales, Adoption and Prediction Workflow

The second tier here is the presentation tier. This tier will be used specifically for visualization to get a better picture of the dataset or the output. There are statistical metrics used here to evaluate the source dataset. Also, different varieties of graphs are deployed using Python libraries like matplotlib and pandas. This enables one to identify any trends, seasonality, categorizations, intensities, outliers or any issues in the data. Once the data is visualized, cleaned and transformed, it is split into a training (70-80%) and a testing set (20-30%). The data is fetched to different models for prediction. The output of the model is verified and tested using two means: Metrices evaluation and Visualization Evaluation. For metrics evaluation, we will be depending on AIC i.e Akaike Information Criterion, RMSE i.e. Root Mean Squared Error, MSE i.e. Mean Squared Error, and MAPE i.e. Mean Absolute Percentage Error for comparison. For visualization, we will be opting for the actual vs forecasted values line graph. The various models will be executed, and results will be compared to reach a conclusion on which model to opt for the best fit.

There are key considerations for the above process, which involves: First the choice of time series models used based on the data, its pattern and its complexity. Secondly hyperparameter tuning for optimal model performance. This methodology should lead to a systematic approach and better performance with the models.

3.4 The Dataset Sourcing for Analysis

The dataset, in figure 3.3 chosen for this research is obtained from Kaggle. It shows the number of vehicles registered in the USA by the Washington State Department of Licensing. The data is separated by county and contains the passenger vehicles, as well as trucks, registered every month till early 2024. The source further shows the total percentage of vehicles registered and the percentage of EVs registered till the year 2024. The dataset has 20,820 total records in it. From the first observation of the database once can notice EV adoption in the dataset is still less compared to traditional ICE vehicles. Most of the places have a small vehicle population, but some of these regions have comparatively large vehicle counts, although this should not affect the analysis. However, there is a need to develop strategies to improve EV adoption in these regions. The factors that are making some regions adopt EV better should be looked upon. Overall, the dataset provides efficient details to be selected for Time series Analysis, even though it requires a certain level of data transformations.



	Date	County	State	Vehicle Primary Use	Battery Electric Vehicles (BEVs)	Plug-In Hybrid Electric Vehicles (PHEVs)	Electric Vehicle (EV) Total	Non-Electric Vehicle Total	Total Vehicles	Percent Electric Vehicles
0	September 30 2022	Riverside	CA	Passenger	7	0	7	460	467	1.50
1	December 31 2022	Prince William	VA	Passenger	1	2	3	188	191	1.57
2	January 31 2020	Dakota	MN	Passenger	0	1	1	32	33	3.03
3	June 30 2022	Ferry	WA	Truck	0	0	0	3,575	3,575	0.00
4	July 31 2021	Douglas	CO	Passenger	0	1	1	83	84	1.19

Fig 3.3 Electric Vehicle Sales, Adoption and Prediction Dataset

3.4 The Data Preprocessing

As in Figure 3.4, the process involves a series of transformations to the selected dataset to prepare them for Time Series Analysis (Bhagoji et al., 2018). These steps are as below:

Data Cleaning: We make sure there are no discrepancies in data that may hinder the result of the analysis. In the existing, there were null or Nan values in them. These records were dropped. There seem to be no further discrepancies in the data. Apart from the elimination of the null values, any record with a zero value or lesser in BEVS, PHEVS, or Non-Electric Vehicle columns has been removed.

Data Transformation: This step involves manipulating the dataset or structure of any of its dimensions to make it ready for further analysis. It is a prerequisite to enable the dataset for modeling. Here we have made a couple of transformations into columns. e.g. The date column details were not ideal for the analysis i.e. "DD <> Month_Name <> YYYY". We have split the column into three sub-columns namely Date, Month and Year. Further we have converted the date column to numerical value. The result is a column Recon date 'YYYY-MM-DD'.

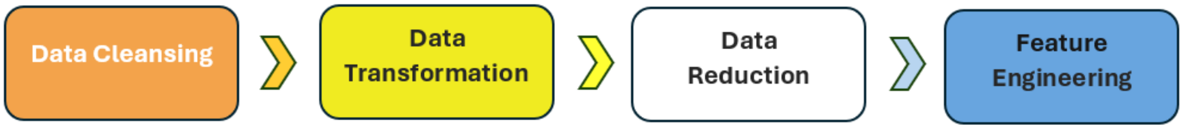


Fig 3.4 Electric Vehicle Sales, Adoption and Prediction Dataset

The second transformation made here is the columns 'BEV', HPEVs' and 'ICEs' are pivoted to transform them into a column 'Vehicle Count' and a supporting column 'Vehicle_Types' to represent the type of vehicle the 'Vehicle Count' represents.

Data Reduction: This process involves cutting down on the dataset to remove data that are not required for further modeling. Removing columns that are non-numerical or the ones that are not part of the analysis will be removed, e.g. we have columns such as County, State, Vehicle use, Total vehicles, and Percentage Electric vehicles that will be removed post Exploratory Data Analysis. Further, we would be sampling data for modeling only if required.

Feature Engineering: This methodology focuses on the creation, alteration or selection of features that are relevant to modeling. This can help improve the performance of the model. This involves resampling, scaling, encoding, or applying domain-specific information.

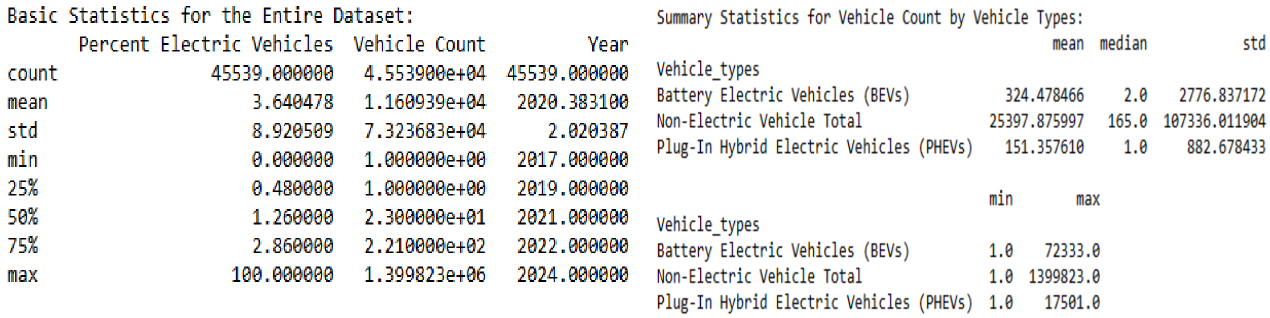
- **Resampling Data**: In the preprocessing stage the dataset was resampled according to the month column. This was done by aggregating the time series data to monthly frequency. This simplified the model and improved the accuracy of the model outputs.
- **Encoding Data**: It is the process of associating meaningful numerical values to non-numerical columns for the purpose of analysis. Here we have encoded the Month column to make it numerical as mentioned before. e.g. From "June" to "06".
- **Scaling Data**: This refers to adjusting the range of the distribution before modeling to enable the features to be compared effectively. The data is scaled using min-max (Normalization) scalar wherever necessary. We have, e.g. There is a scenario where we were required to scale the data for modeling for LTMC Modeling. LTMC is sensitive to magnitude of input. Here scaling will help by preventing large values from dominating the output.
- **Outlier Detection:** An outlier is any data point in the dataset that differs significantly from the average trend of the data. These values might be real high or real low compared to the rest of the data. The occurrence of these values are rare and most probably will not represent the overall pattern of the dataset even though this may hinder the result. Because of the nature of the outlier, it is to be often removed from the analysis. We can use statistical analysis or visual means to detect these. These can be removed by logarithmic

transformations, replacing them with average values from data or limiting the maximum and minimum range of the dataset.

3.5 Exploratory Data Analysis

(Chatfield, 1986)This step is one of the key approaches in the presentation tier for time series analysis. This step involves analyzing the data through statistical means and visual means to get clarity over the main characteristics of the data. The process should help us to uncover patterns or identify relations or correlations or verify our presumptions with the aid of graphs and statistical metrics. Generally, an EDA is done side by side with the data cleaning which ensures the data is presented in the appropriate format or as the requirement. We might be relying on Python libraries such as Pandas, Matplotlib, Seaborn, Plotly, or SciPy for the analysis. As part of our research, we have performed the following exploratory data analysis:

Statistical Data Analysis: This involves gathering, aggregating, calculating, and interpreting numerical values to understand data using statistical means. This helps to understand the characteristics of data through mathematical means such as mean, median, variance, and standard deviation summary. The inference provided by these metrics can provide valuable insights about the dataset.



```
Basic Statistics for the Entire Dataset:
      Percent Electric Vehicles  Vehicle Count          Year
count                45539.000000   4.553900e+04  45539.000000
mean                     3.640478   1.160939e+04   2020.383100
std                      8.920509   7.323683e+04      2.020387
min                      0.000000   1.000000e+00   2017.000000
25%                      0.480000   1.000000e+00   2019.000000
50%                      1.260000   2.300000e+01   2021.000000
75%                      2.860000   2.210000e+02   2022.000000
max                    100.000000   1.399823e+06   2024.000000
```

```
Summary Statistics for Vehicle Count by Vehicle Types:
                                               mean  median            std
Vehicle_types
Battery Electric Vehicles (BEVs)         324.478466     2.0    2776.837172
Non-Electric Vehicle Total             25397.875997   165.0  107336.011904
Plug-In Hybrid Electric Vehicles (PHEVs)   151.357610     1.0     882.678433

                                        min        max
Vehicle_types
Battery Electric Vehicles (BEVs)          1.0    72333.0
Non-Electric Vehicle Total                1.0  1399823.0
Plug-In Hybrid Electric Vehicles (PHEVs)  1.0    17501.0
```

Figure 3.5 Statistical Data Analysis of the Source Dataset

Based on the statistics from Figure 3.5, EV adoption seems to be at a very early stage with just *3.64%* of the total vehicles. However, the standard deviation of *8.92%* shows an uneven distribution signifying some states will be having more than or less than the average trend. The median here shows that the distribution is skewed, that means a few regions may have high number of vehicles registered.

When observing individual vehicle types, a mean count of *324.48* and a median for the **Battery Electric Vehicles (BEVs)** suggest its adoption is concentrated in certain places. The 75th percentile value of *8.0* confirms that the presence of EVs is limited in many regions. While **Plug-in Hybrid Electric (PHEVs)** statistical summary suggests an even lesser penetration into the market than BEVs with a mean value of *151.36* and a median value of just *1.0*. **Internal Combustion Engine (ICE)** vehicles on the other hand dominate the market with an average count of *25,397.88* and a median value of *165.0*. Their high standard deviation of *107,336.01* concludes that these vehicles are heavily spread across the regions. This must be due to technology's dominance in the market for decades.

Graphical Data Analysis: This analysis involves the depiction of the data through graphical and visual means. It uses tools for plotting various graphs namely Bar charts, Line graphs, Box plots, scatter plots, histograms, pie charts etc. For the current scenario, we have opted for the depiction below to analyze and understand our data.

- **Vehicle Registered Frequency vs State Bar Charts:** These graphs leverage rectangular bars to represent categorical values from a column or two on a graph. Each bar corresponds to each categorical value in the column. These are widely used to compare frequencies of different categorical values in the columns to provide a clear understanding. Here we have utilized the graph to plot BEVs, and PHEVs presence in each state of the USA.

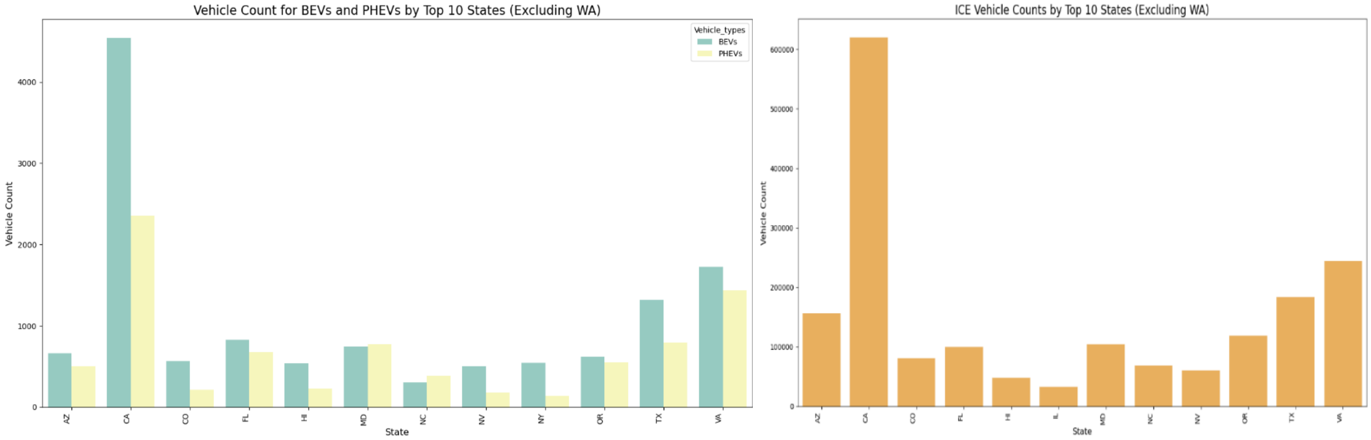


Figure 3.6 Bar Chart: a. BEVs & PHEVs Vehicle count vs State, b. ICE Vehicle Count vs States

Pertaining to EVs in Figure 3.6, Washington DC and California have the highest number of EVs and PHEVs vehicles. Other states that have high EV populations are Colorado, Florida, Texas, and New York. Whereas states like South Carolina, North Carolina, and Georgia seem to have the lowest registered EVs in the last decade. The high concentration of EV population in Washington DC and California might be due to pro-EV policies from the state governments like incentives or investments in EV infrastructure as noted from the literature review. Also, urban areas with higher populations and shorter commute distances might be another reason for favorable adoption in these regions. The counts of ICE vehicles being high in these regions suggest that these regions have sustainably more vehicles registered irrelevant of the vehicle technology.

- **Electric Vehicle percentage in each state Pie Charts:** A Pie chart represents the categories of data and their proportions in a circular graph divisioned into sectors including the numerical percentage of frequencies of each of these categories. The relative percentage of frequencies of these categories helps visualize, compare, and contrast the data. Here we compare the electric percentage of vehicles against each state/Counties in the USA.

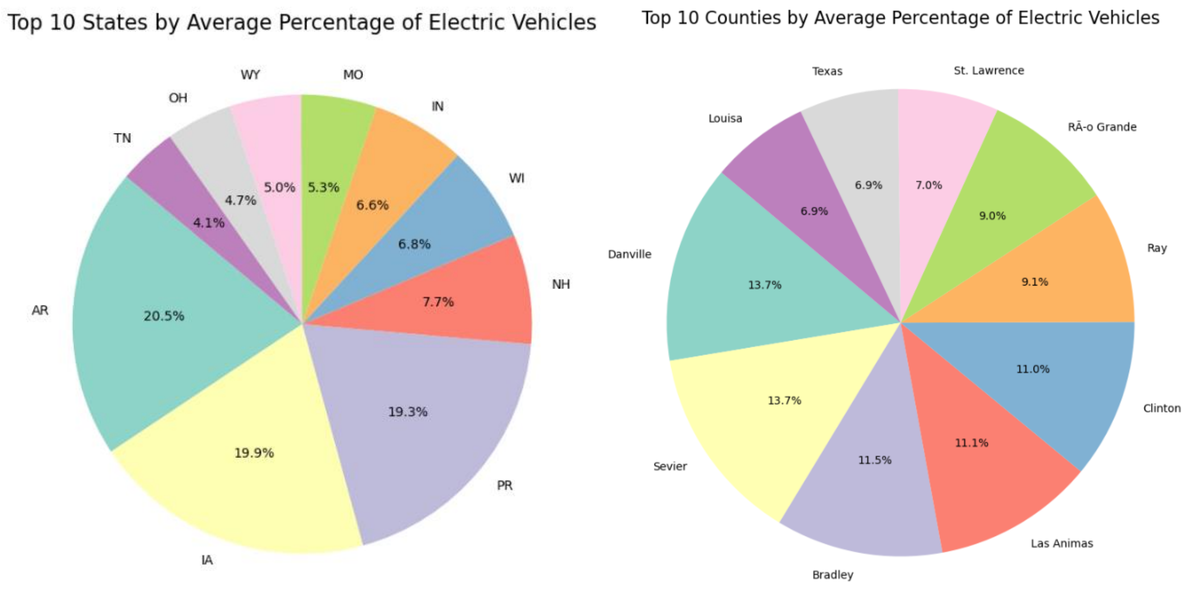


Figure 3.7 Pie Chart: a. vs State, b. Non-Electric Vehicle Count vs States

The Pie chart in Figure 3.7, depicts the state with the highest percentage of EVs namely California with *19.99%* EVs, New Hampshire with *7.7%* EVs, and Wisconsin with *6.8%* EVs respectively. The lower EV states include Missouri with *5.3%* EVs and Tennessee with *4.7%* EVs. When it comes to Counties, the county Danville in Virginia and the county Sevier Arkansas have the highest EV vehicle percentage with *13.7%* of EV vehicles. On the other hand, the county Las Animas in Colorado showed the lowest EV counts, with just 9.1% of the EV population. The above analysis further signifies the reading from the statistical analysis and Bar chart that there is a significant variation in EV adoption in different parts of the country.

- **EV Vehicle Count over the years using Line Graphs:** A line chart is utilized when we need to plot a continuous trend in data over a period. The data points are depicted on the graph joined by a line which represents the growing or diminishing trend. The graph could also show short-term or long-term fluctuations and correlations between two series of data. Here we have depicted the number of vehicles registered for each BEVs, PHEVs and ICE vehicle signifying their growth over the years.

  The graph in Figure 3.8, depicts the dominance of ICE vehicles in the automobile market over the period of 10 years. However, the dominance seems to have ceased post the year 2021, and since then a sharp decline has been noticed, that is post-2023. This might be due to 2 reasons: First, the total data for 2024 might have yet to be released, and second, the EV adoption and government policies pertaining to penalizing the use of traditional ICE vehicle technology like diesel engines. For the vehicles adopting EV technologies have displayed new details regarding their growth. It is observed that even though their population remains low, there has been a significant growth in the population of these vehicles over the years. The key reasons here might be due to improvements in battery technologies and improved range. Further growth will be forecasted using the time series analysis.

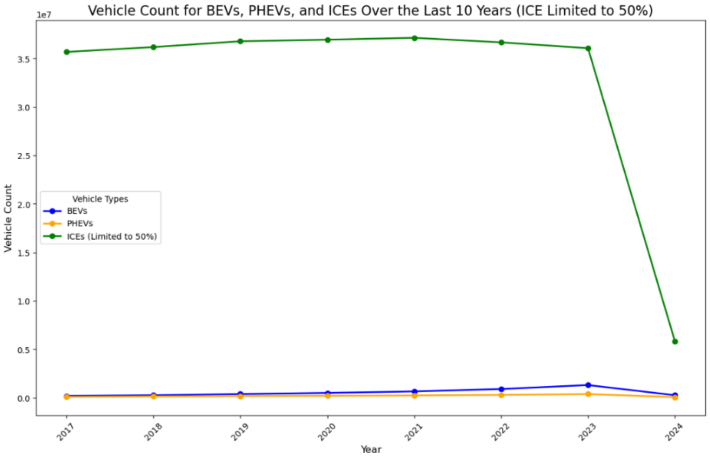


Figure 3.8 Pie Chart: a. BEVs/PHEVs vs State, b. ICE Vehicle Count vs States

3.6 Seasonal Decomposition

(“Fast RobustSTL: Efficient and Robust Seasonal-Trend Decomposition for Time Series with Complex Patterns | Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining,” n.d.) Seasonal decomposition in a time series analysis is a methodology in which the data's short-term patterns and long-term patterns are separated. Majorly time series data will have three components: First, the seasonality, which are period fluctuations, Second the Trend, which are long-term patterns, and finally residuals, which represent random noises in data. The method helps us understand all underlying patterns, making it easy to interpret the data.

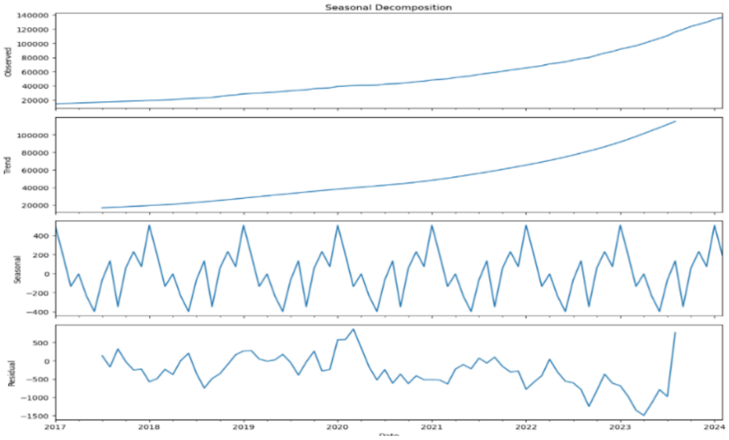


Figure 3.9 Seasonal decomposition of Battery Electric Vehicle (BEVs)

In Figure 3.9, The trend components of the BEV over the years show a growing pattern. This confirms the earlier observations made and shows how the overall trend is of growth. The seasonal component displays how there is an effect of yearly peaks and downs. These patterns might be an impact of period promotions, seasonal sales, or holidays. The last component residual, which is the left-over noise is minimal and suggests that the trend and seasonal components provide a good fit for the BEV data.

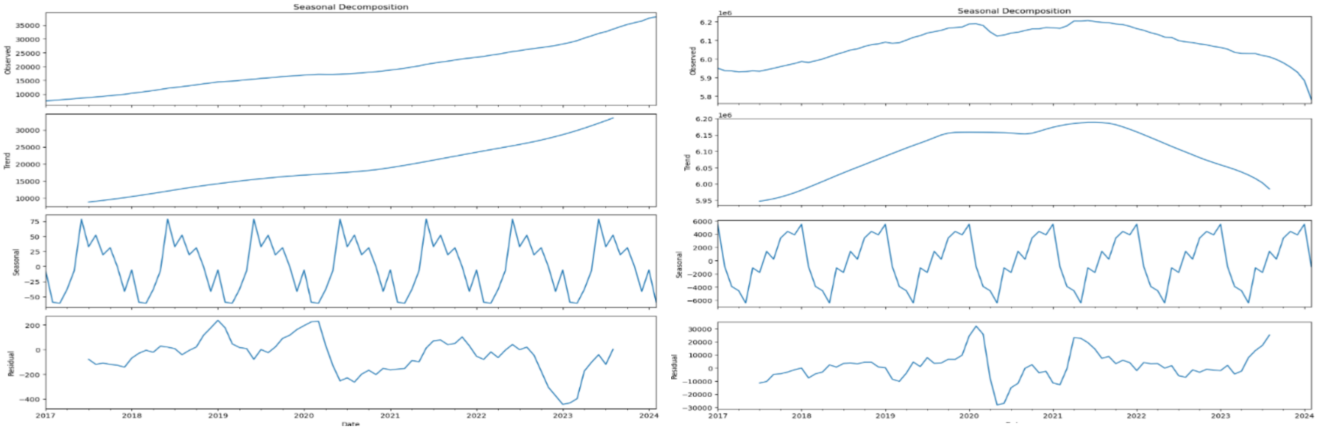


Figure 3.10 Seasonal decomposition of a. PHEVS, b. Non-electric Vehicle

In Figure 3.10 seasonal decomposition for PHEVs, the trend follows an upward trend similar to BEVs. However, for non-electrical vehicles, the trend displays a steady growth till 2021. Post 2021 the overall trend of non-electrical vehicles is declining. Both PHEVs and non-electric vehicles show the existence of seasonality and residuals.

For Time series models seasonality, trend, and residuals need to be handled differently. For example, an ARIMA model can handle trends by using differencing. However, it cannot handle seasonality. Therefore, a seasonal difference needs to be performed before ARIMA modeling. Other models such as SARIMA or LSTM can inherently handle trend and seasonality.

3.7 Stationarity Check and Differencing on Time Series

(Witt et al., 1998) A stationarity check is performed before a time series analysis which uses ARIMA or SARIMA models. It checks whether a time series data is stationary. Data is stationary, which means the statistical properties of it, i.e. mean, variance, and autocorrelation, remain constant over a period of time. This ensures a stable response from the designed time series model. Some of the key concepts in stationary check are:

ADF Test: The augmented Dickey-Fuller test is one of the techniques used to find out if a time series is stationary or a non-stationary one. It checks for a ‘unit root’ in series which in layman's terms, a sudden change or shock can have a lasting effect on the series, making it unpredictable and non-revertible. For ADF Test a null hypothesis assumes that the time series has a unit root. The ADF test statistics are calculated and compared with each of the critical values at 1%, 5%, and 10%. If the ADF statistics is greater than the critical values the hypothesis is rejected, and the time series is considered stationary.

KPSS TEST: The KPSS test is a second method used to evaluate if the series is stationary or a non-stationary one. In this process, the series is split into mean, trend, random walk and white noise. The test focuses on the random walk component which is essentially a drifting part of the series that can cause instability. If this part is stable the series is considered stationary.

First order Differencing: (Koreisha and Pukkila, 1993) Differencing is the process of subtracting the current value of the time series from its value before, over a period of time. By understanding this difference between the values over a period one can interpret the trend. These changes or differences can be analyzed to remove trends from the data.

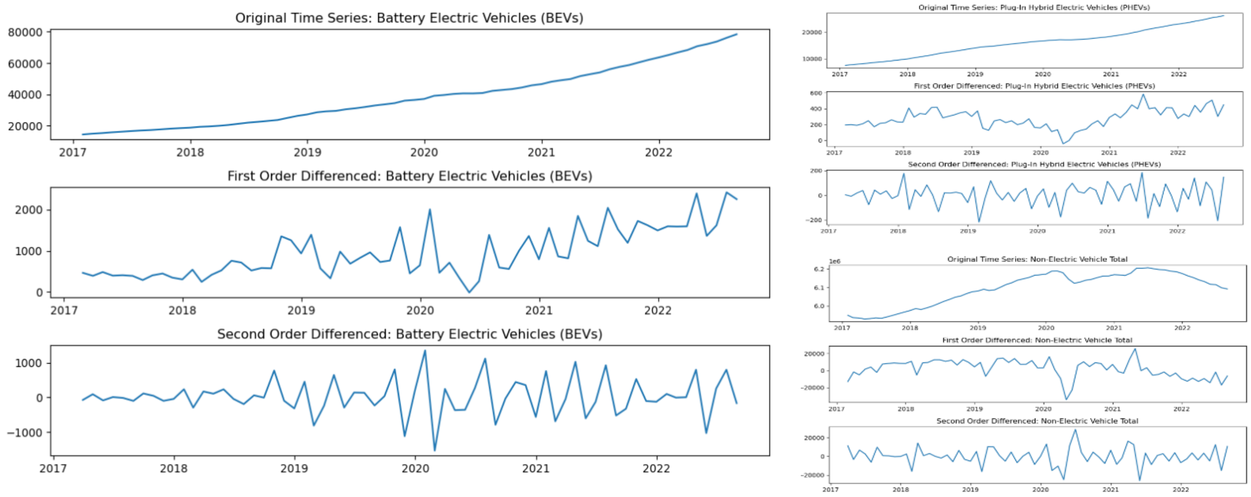


Figure 3.11 1st order and 2nd Order differencing for BEVs, HPEVs and ICEs

Second Order differencing: (Koreisha and Pukkila, 1993) This process is similar to first order differencing with the only difference being the process of differencing is done twice. This is used when the result data from the first is still not stationary.

Stationarity check and Differencing Technique: The stationary check is performed using the ADF test and KPSS tests here on the BEVs, PHEVs, and Non-Electric vehicle data. When the results of these tests results turn out to be non-stationary, a differencing is applied to the result and the tests are executed for a second time. This process continued until the second order differencing or until the data is stationary.

For the ADF test:

- Null Hypothesis (H_0): The time series is non-stationary (contains a unit root).
- p-value < 0.05 → Reject H_0, stationary.
- p-value > 0.05 → Fail to reject H_0, non-stationary.
- Also, ADF test statistics should be less than critical values at 1%, 5%, and 10 % significance level.

For a KPSS test:

- Null Hypothesis (H_0): The time series is stationary.
- p-value < 0.05 → Reject H_0, non-stationary.
- p-value > 0.05 → Fail to reject H_0, stationary.
- Also, KPSS test statistics should be less than the critical value.

Stationarity Check and Differencing Evaluation:

For BEVs in Figure 3.12, initial readings of ADF p-value of 0.999 fail to reject the null hypothesis. This is because the p-value is greater than 0.05 and therefore does not reject the null hypothesis of the existence of a unit root. The KPSS p-value of 0.01 is less than 0.05 which supports the above result. Post first-order differencing the series is still non-stationary with an ADF p-value of 0.907 and KPSS value of 0.01. However, after the second order differencing the series is stationary with ADF p-value 2.66 x 10^{-7} much less than 0.05, and KPSS p-value of 0.1 greater than 0.05. The critical value comparison suggests the same of ADF test statistics -5.909 against (**1%**: -3.542, **5%**: -2.910, **10%**: -2.593) and KPSS test statistics 0.173 against 12 suggests the same.

Vehicle Type	**Type of Data**	**ADF Test Statistic**	**ADF p-value**	**ADF Inference**	**KPSS Test Statistic**	**KPSS p-value**	**KPSS Inference**
BEVs	Original	2.705	0.999	Non-stationary	1.208	0.01	Non-stationary
BEVs	First-Differenced	-0.416	0.907	Non-stationary	1.127	0.01	Non-stationary
BEVs	Second-Differenced	-5.909	2.66 x10^{-7}	Stationary	0.173	0.1	Stationary

Figure 3.12 Stationarity Check and Differencing Results BEVs

For PHEVs in Figure 3.13, the initial reading ADF p-value of 0.989 is greater than 0.05 and therefore does not reject the null hypothesis of the existence of a unit root. The KPSS p-value of 0.01 is less than 0.05, which supports the above result. Post first-order differencing the series is still non-stationary with an ADF p-value of 0.305 but seems stationary with a KPSS p-value of 0.1. Since there is a conflict between the results, we consider the second-order differencing. After the second order differencing the series is stationary with ADF p-value $9.16\times10^{-239.1}$ much less than 0.05. The critical value comparison suggests the same of ADF test statistics -12.259 against (**1%**: -3.535, **5%**: -2.907, **10%**: -2.591) and KPSS test statistics 0.107 against 11 suggests the same.

Vehicle Type	**Data Type**	**ADF Test Statistic**	**ADF p-value**	**ADF Inference**	**KPSS Test Statistic**	**KPSS p-value**	**KPSS Inference**
PHEVs	Original	0.664	0.989	Non-stationary	1.216	0.01	Non-stationary
PHEVs	First-Differenced	-1.959	0.305	Non-stationary	0.335	0.1	Stationary
PHEVs	Second-Differenced	-12.259	$9.16\times10^{-239.1}$	Stationary	0.107	0.1	Stationary

Figure 3.13 Stationarity Check and Differencing Results PHEVs

For the Non-Electric vehicles in Figure 3.14, the initial reading ADF p-value of 0.243 is greater than 0.05 and therefore does not reject the null hypothesis of the existence of a unit root. The KPSS p-value of 0.01 is less than 0.05, which supports the above result. Post first-order differencing the series seems stationary with an ADF p-value of 0.0006 but is non-stationary with a KPSS p-value of 0.042. Since the KPSS test suggests non-stationary there might be an existence of trend stationarity, therefore we consider further differencing. After the second order differencing the series is stationary with a KPSS p-value 0.1 $^{\text{greater}}$ than 0.05. The critical value comparison suggests the same as ADF test statistics -6.151 against (**1%**: -3.540, **5%**: -2.909, **10%**: -2.59), and KPSS test statistics 0.175 against 9 suggests the same.

Vehicle Type	Data Type	ADF Test Statistic	ADF p-value	ADF Inference	KPSS Test Statistic	KPSS p-value	KPSS Inference
Non-Electric	Original	-2.103	0.243	Non-stationary	0.973	0.01	Non-stationary
Non-Electric	First-Differenced	-4.229	0.0006	Stationary	0.496	0.042	Non-stationary
Non-Electric	Second-Differenced	-6.151	7.56×10^{-87}	Stationary	0.175	0.1	Stationary

Figure 3.14 Stationarity Check and Differencing Results ICEs

3.8 Time Series Modeling Techniques

At the core, a time series analysis is a method derived from statistics that is leveraged to analyze data points spread across time. It deals with learning from the patterns, trends, and structures from the time series to understand the root behavior of the dataset to forecast the future of the data. Here we will be leveraging several time series model techniques which are as below:

ARIMA Model:

(“Study and analysis of SARIMA and LSTM in forecasting time series data - ScienceDirect,” n.d.) The method that stands for Autoregressive Integrated Moving Average is a popular modeling technique derived from statistics used for time series analysis. It leverages three components to capture the patterns of a series. The three components are:

- The relationship between one of the observations and its past values in time is called an AR (Autoregressive) component.
- Fluctuations in data over time which will be smoothened by the model by differencing called integrated component.
- Residual errors, observation, and the dependency between them form a moving Average component.

The model is defined as above and is denoted by (p, d, q). The model combines the above three components to forecast the future, especially when the past patterns could help predict future variables. ARIMA is capable of capturing temporal discrepancies and generating forecasts for future data.

SARIMA Model:

(“Study and analysis of SARIMA and LSTM in forecasting time series data - ScienceDirect,” n.d.) Similar to ARIMA a seasonal Autoregressive Integrated Moving average model extends the functionality of the ARIMA model by addition of the seasonal component. This makes SARIMA adapt a time series with seasonal patterns in it and makes predictions on models considering the series will have periodic patterns. A SARIMA is represented as SARIMA(p, d, q) (P, D, Q, m) where P, D, Q represents seasonal Autoregressive, Differencing, and Moving Average. The m represents the period e.g. m = 12 means represents 12 months of data with yearly seasonality.

ETS Model:

(Billah et al., 2006) An Exponential Smoothening State Space Model, known as ETS Model is a methodology that emphasizes trends and smoothness over the period. This kind of model is especially useful to interpret a series with seasonality, trends, and other factors that may arrive over time. The advantage of this is it doesn’t require the series to be stationary. It can inherently decompose a series into trends, seasonals, and errors and handle them dynamically. How it achieves it is through exponential smoothening. What it essentially does is it weighs the past observations exponentially with the current ones, where current observations will be receiving more weight.

Prophet Model:

(Ning et al., 2022) A forecasting modeling open-source tool developed from the developers of Facebook, the Prophet model can handle complex scenarios with missing data points and some strong seasonal patterns. It was majorly developed to analyze sales, website traffic, and activities of users. Similar to earlier models Prophet breaks down the series into Trends, seasonality, and Events such as Holidays. Unlike ARIMA, it can effectively handle non-stationary series. On top of that users can add custom seasonal patterns, handle outliers and missing points well, and need lesser parameter tuning.

LSTM Model:

(“Study and analysis of SARIMA and LSTM in forecasting time series data - ScienceDirect,” n.d.) Regular time series models tend to overlook older information when a series is too long, or the data is big. This is termed a vanishing or gradient exploding problem in time series analysis. While a Long short-term Memory Model collects important facts from each data point from the beginning that might be key in understanding a pattern later. The model leverages gates for this purpose. These gates essentially store details of what information to be added, what to delete, and what to be update. Some of these gates are the Update gate, forget gate, Input gate, and Output gate. Their functions are as their names suggest and are self-explanatory. Additionally, we have a candidate gate that represents a new candidate value to be added to the cell and a Hidden gate that provides the output of the LSTM at a particular time period.

3.9 Hyperparameter Tuning

(Liao et al., 2022) This methodology focuses on preparing the model before forecasting by finding the parameters for the configurations that are optimal. By doing this it aims at improving the accuracy, precision, and end result. It helps prevent the chances of overfitting or underfitting in forecasting models and can also aid with reducing computational resources. There are several kinds of hyperparameter tuning methods which include manual search, grid search, random search, Bayesian optimization, Hyperband optimization, and some automated libraries like Keras tuner, Optuna etc.

Here we are utilizing a Hyperband tuner from Keras Tuner Library. It is an automated tuner that is a combination of an early-stopping and a random search approach. The tuner is effective against long-running time series like Deep learning algorithms. Here we will be utilizing the setup for LSTM Modeling.

3.10 Evaluation Metrics

(Bergmeir and Benítez, 2012) For the evaluation of the time series models, we will be utilizing Six metrics that are discussed below:

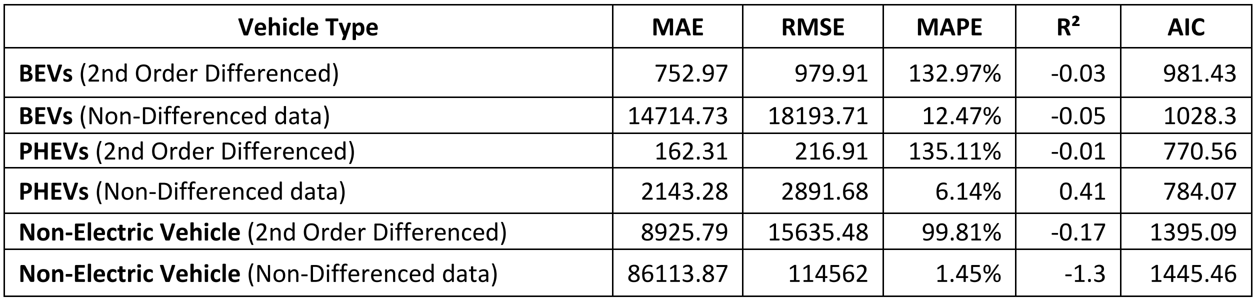
- **MAE** i.e. Mean Absolute Error is the average of the difference between original and forecasted values.
- **RMSE** i.e. Root of Mean Squared Error is similar to MSE except for the fact that the output it provides is in the same unit as inputs.
- **MAPE** i.e. Measure of the Average percentage between the original and forecasted values.
- **AIC** i.e. Akaike Information criterion is the relative quality of a model for data. It tries to balance the goodness of a fit to the model complexity. While AIC only provides a relative comparison between models, it can account for overfitting.
- **R Squared** value depicts how well the model explains the variability of actual data or the historical trends. R square cannot be relied upon for a time series model in several scenarios like for instance in the case of time series with autocorrelation. It can only explain the variability of data which uses prediction modeling.

4. Model Evaluation and Discussions

4.1 ARIMA Model evaluation:

For BEVs in Figure 4.2, the MAE & RMSE outputs were low, but the MAPE value, which is the average percentage of error, of 132.97%, is too high and shows how the prediction was unreliable. The R^2 value of -0.03 supports the MAPE readings. Therefore, a test with the original series was performed using ARIMA with an integrated component value ‘1’ to implicitly handle the stationarity. The output gave a much better performance in terms of MAPE with just a 12.47% average error percentage. However, there was only a minor improvement in R^2. Also, the MAE and RMSE errors increased substantially.

PHEVs performed similarly with high MAPE 135.11% average percent errors for the 2nd order differenced data. However, the ARIMA was executed with original data with an integrated component as ‘1’ to handle stationarity implicitly and the results improved. The MAPE error percentage was reduced to 6.14% and the model explained 41% of the data as per the R^2.



Vehicle Type	MAE	RMSE	MAPE	R²	AIC
BEVs (2nd Order Differenced)	752.97	979.91	132.97%	-0.03	981.43
BEVs (Non-Differenced data)	14714.73	18193.71	12.47%	-0.05	1028.3
PHEVs (2nd Order Differenced)	162.31	216.91	135.11%	-0.01	770.56
PHEVs (Non-Differenced data)	2143.28	2891.68	6.14%	0.41	784.07
Non-Electric Vehicle (2nd Order Differenced)	8925.79	15635.48	99.81%	-0.17	1395.09
Non-Electric Vehicle (Non-Differenced data)	86113.87	114562	1.45%	-1.3	1445.46

Figure 4.1 ARIMA Model Result Metrics

For the Non-Electric series, the model displayed high intensity of errors for both the 2nd order differenced and non-differenced series. The AIC reading shows how the results are not optimal.

ARIMA Model did not efficiently learn from the complexity of the series here. However, it was noticed that the differencing made the series unreliable in this case, and modeling using ARIMA on original data with integrated components produced better results. This might be due to reasons of over-differencing, differencing-induced noise, distortion of reliable information, or increased model complexity post-differencing.

4.2 SARIMA Model Evaluation:

The non-differenced series was fetched to the Seasonal ARIMA Model with a frequency configuration of 12 months with seasonal order (1, 1, 1, 12) and integrated component as ‘1’.

As in Figure 4.2, For BEVs, the SARIMA model has produced substantially better results. The MAE is relatively low and RMSE is higher than the MAE. This might be due to the existence of larger errors. This might not be significant compared to the overall trend. The MAPE is minimal with only 3.01% of errors. Based on the R^2 value the modeling could explain 94% of the actual values which is a huge improvement from earlier models. The AIC value of 608.86 is relatively low, suggesting a good fit.

For PHEVs, the MAE and RMSE produced low errors and the MAPE value is low at just 8.27%. In the Time series analysis, R^2(-0.05) can be negative and may only suggest the model is a poor fit. Further visualization will need to be considered to finalize the fitness.

Vehicle Type	MAE	RMSE	MAPE	R²	AIC
Battery Electric Vehicles (BEVs) (SARIMA)	3538.16	4402.05	3.01%	0.94	608.86
Plug-In Hybrid Electric Vehicles (PHEVs) (SARIMA)	2882.52	3851.32	8.27%	-0.05	488.12
Non-Electric Vehicle Total (SARIMA)	25103.75	44543.92	0.43%	0.65	881.96

Figure 4.2 SARIMA Model Result Metrics

For Non-Electric data, the MAE and RMSE are high and suggest the prediction might be less reliable than BEVs or PHEVs. However, the MAPE error percentage is very minimal 0.43% suggesting the absolute errors might be due to the larger volume and complexity of the non-electric data. The AIC value of 881.96 is moderately low, suggesting a good fit. Further visualization will need to be considered to finalize the fitness. The BEVs forecast vs actual graph in Figure 4.3 shows minimal variation in actual vs the predicted range of values and thus confirms the earlier readings. Whereas the graph for PHEVs depicts a huge variation in the prediction from the actual, suggesting that it is a poor fit after all.

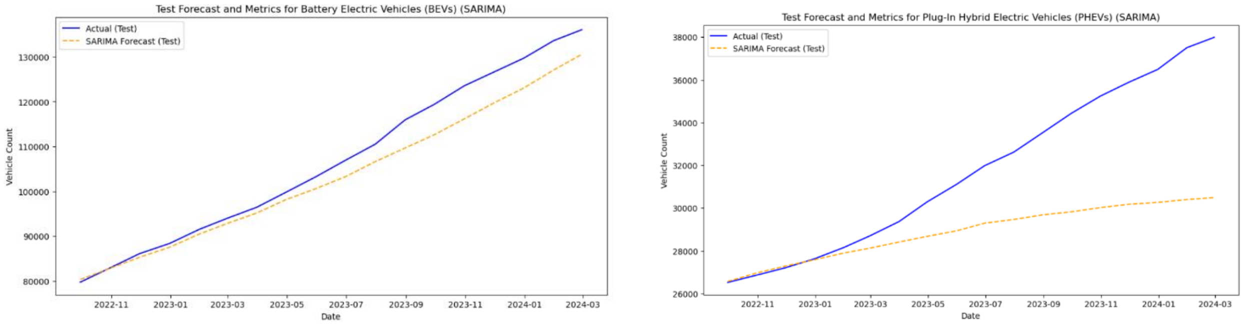


Figure 4.3 SARIMA Model Forecast vs Actual Plot for BEVs and PHEVs

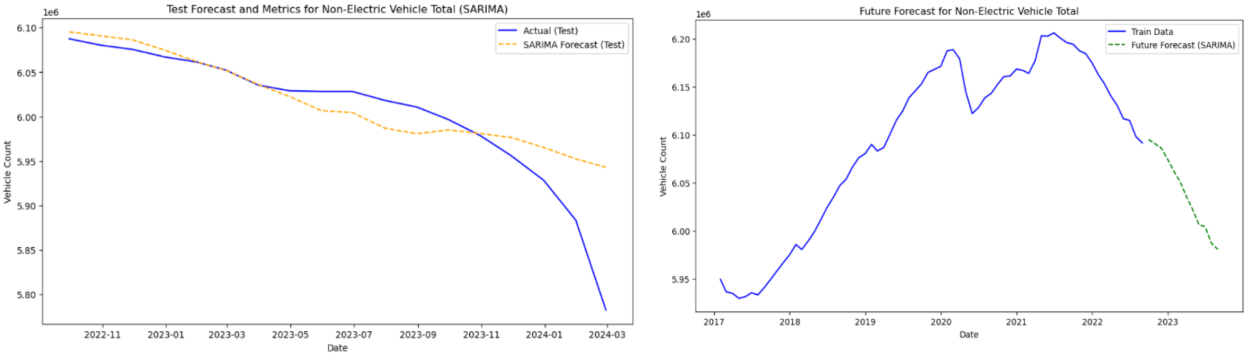


Figure 4.4 SARIMA Model Forecast vs Actual Plot for Non-Electric Vehicles.

The non-electric vehicle graph in Figure 4.4 suggests that there is very little variation in the predicted value from the actual, suggesting that it is a good fit.

4.3 Exponential Smoothening Model Evaluation:

Vehicle Type	MAE	RMSE	MAPE	R^2	AIC
Battery Electric Vehicles (BEVs) (ETS)	9871.11	12392.92	8.32%	0.51	838.19
Plug-In Hybrid Electric Vehicles (PHEVs) (ETS)	1898.01	2542.4	5.44%	0.54	653.77
Non-Electric Vehicle (ETS)	27353.67	39035.44	0.46%	0.73	1260.67

Figure 4.5 ETS Model Result Metrics

As shown in Figure 4.5, the ETS model displayed fairly good overall performance. For BEVs, the results showed no improvement in modeling performance even though the average errors were moderately low. The R^2 displays a relatively lower quality prediction compared to the SARIMA.

For PHEVS a much better performance was noticed compared with earlier models. The MAE, RMSE, and MAPE predicted very low errors with an average percentage rate of only 5.44%. The R^2 has been improved from the SARIMA model even though the forecast value only explains 54% of the data. The AIC results infer that the model is less complex and should be optimal.

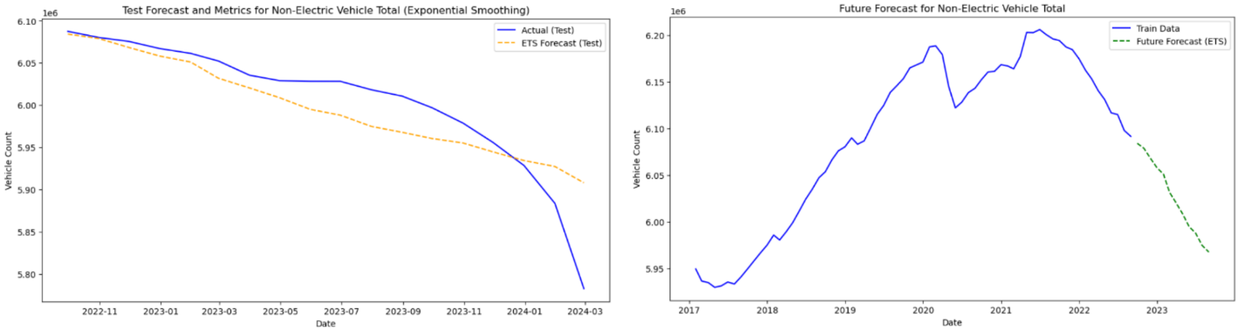


Figure 4.6 ETS Model Forecast vs Actual Plot for Non-Electric Vehicles.

Non-electric vehicles displayed better performance when compared to the SARIMA model results. The R^2 value improved, which explains 73% of the data, without any increase in error metrics. More visual representation can help us understand more on the distribution. The AIC value of 1260 displays a higher value which can be a reflection of complexity with a larger scale of data. Figure 4.6 shows better performance of the new model with minimal variation and negligible deviation in the trajectory of the forecast from the actual.

4.4 Prophet Model Evaluation:

The Prophet Model metric evaluation in Figure 4.7 suggests that the model couldn't effectively interpret the patterns in the series. For all three categories of vehicles, there seems to be no improvement in model performance. The MAPE values were at the lower end suggesting the presence of absolute errors are minimal.

Vehicle Type	MAE	RMSE	MAPE	R²	Cross-Validation RMSE	Cross-Validation MAPE
BEVs (Prophet)	13704.89	16403.65	11.76%	0.14	4205.98	8.58%
PHEVs (Prophet)	2009.36	2657.03	5.78%	0.5	2763.18	10.82%
Non-Electric Vehicle (Prophet)	26265.74	49493.67	0.45%	0.57	1159279.75	7.71%

Figure 4.7 Prophet Model Result Metrics

For BEVs, the cross-validation RMSE and MAPE produce better results than the original RMSE and MAPE suggesting when validated against multiple validation sets it produces better results. Altogether the model performed worse than earlier models.

4.5 LSTM Model Evaluation:

We have implemented a deep learning method LSTM, for our time series analysis with a Keras tuner to perform hyperparameter tuning for the model. The results, in Figure 4.8, show substantial improvement in performance especially for BEVs and PHEVs.

Vehicle Type	MAE	RMSE	MAPE	R²	Pseudo-AIC
BEVs (LSTM)	14697.34	17967.78	12.25%	-0.36	317.89
PHEVs (LSTM)	550.26	754.41	1.58%	0.95	222.78
Non-Electric Vehicle (LSTM)	86681.73	99477.23	1.46%	-0.81	369.23
BEVs (Hyperparameter tuned LSTM)	2981.95	3251.21	2.70%	0.96	77524.6
PHEVs (Hyperparameter tuned LSTM)	1440.12	1563.71	4.29%	0.78	77502.64
Non-Electric Vehicle (Hyperparameter tuned LSTM)	94587.63	110596.7	1.59%	-1.23	77630.41

Figure 4.8 Prophet Model Result Metrics

The BEVs LSTM model without hyperparameter tuning (Figure 4.9) produced high RMSE and MAE errors and a negative R^2 value of -0.36. While after the application of hyperparameter tuning, as shows in Figure 4.8, the performance improved drastically. The MAE and RMSE values (2981.95, 3241.21) were reduced considerably. The average percentage of errors i.e. MAPE reduced to 2.70% from 12.25%. The R^2 value explained 96% of the data.

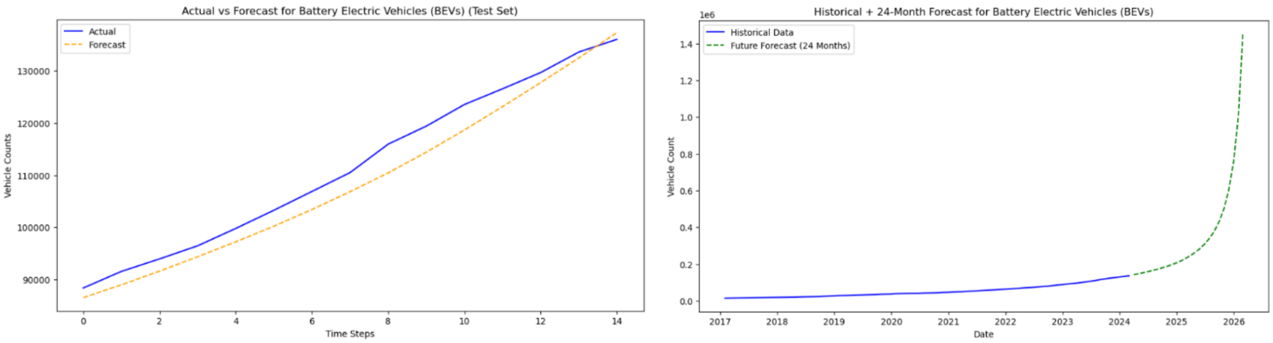


Figure 4.9 LSTM Model with hyperparameter tuning. Forecast vs Actual Plot for BEVs.

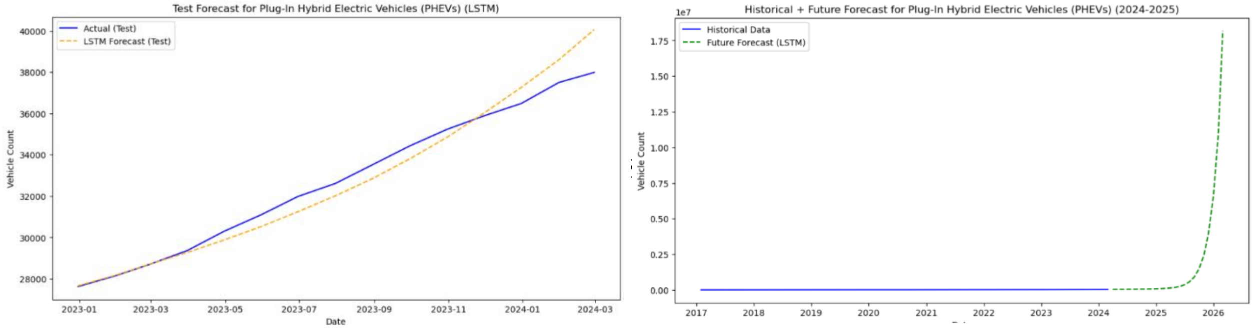


Figure 4.10 LSTM Model with hyperparameter tuning. Forecast vs Actual Plot for PHEVs.

While PHEVs LSTM model (Figure 4.10) produced greatly improved results LSTM without opting for hyperparameter tuning. The forecast produced had very few errors as clear from the MAE, RMSE, and MAPE scores of 550.26, 754.41, and 1.58%. The R^2 improved to a great extent from earlier models with the forecast explaining 95% of the data.

5. Conclusion and Future Work

I have extensively worked on analyzing the Vehicle Registration dataset to bring out patterns that affect the long-term trend and reliably forecast the future of the automobile market. To achieve this, I have implemented various time series machine learning and deep learning models. The above research was conducted on three categories of data i.e. **Battery Electric Vehicles, Plug-in Hybrid vehicles, and Non-Electric Vehicles**. Each of these categories of data displayed different long-term and short-term patterns. These patterns were analyzed and learned by models to forecast the future direction of the market.

Essentially for **Battery battery-operated electric** vehicles, an LSTM model using Hyperparameter tuning produced the best fit with minimal errors and the model explained 96% percent of the data. Similar models were executed for Plug-in Hybrid vehicle data and the LSTM Model without the need for hyperparameter tuning produced the best results with the model explaining 95% percent of the data. LSTM because of its deep learning characteristics of learning long-term sequential patterns could reliably forecast the data. For the large-scale data of non-electric vehicles, the SARIMA Model produced the best fit with a very minimal absolute error percentage of 0.46% and the model explained 73% percentage of the data. The AIC metrics of these models showed the balance between the complexity and performance of the models indicating their good fitness.

The research evidently displays a slow and gradual growth trend for EV models such as BEVs and PHEVs and the forecast shows that post 2025 there will be an exponential growth over the period. Also, the forecast shows that customers will prefer battery-operated Electric vehicles more than the hybrids in the near future even after the limitations of battery-operated vehicles. As for non-electric vehicles, the forecast suggests a steep downward trend. This is apparent with the fact that the government will focus on these fleets in the future as a move towards an eco-friendly future.

The research has instilled in me a curiosity to learn more about these methodologies and grow the research as there is room for improvement. So far, the research has focused on learning a particular dataset. The model should be exposed to more real-time data, adapting over the period and providing inference on the latest trends. Therefore, the research will be exposed to more datasets obtained from different countries essentially helping in unraveling new insights.

6. References

Adnan, N., Nordin, S.M., Rahman, I., Vasant, P.M., Noor, A., 2017. A comprehensive review on theoretical framework-based electric vehicle consumer adoption research. International Journal of Energy Research 41, 317–335. https://doi.org/10.1002/er.3640

Bergmeir, C., Benítez, J.M., 2012. On the use of cross-validation for time series predictor evaluation. Information Sciences, Data Mining for Software Trustworthiness 191, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Bhagoji, A.N., Cullina, D., Sitawarin, C., Mittal, P., 2018. Enhancing robustness of machine learning systems via data transformations, in: 2018 52nd Annual Conference on Information Sciences and Systems (CISS). Presented at the 2018 52nd Annual Conference on Information Sciences and Systems (CISS), pp. 1–5. https://doi.org/10.1109/CISS.2018.8362326

Billah, B., King, M.L., Snyder, R.D., Koehler, A.B., 2006. Exponential smoothing model selection for forecasting. International Journal of Forecasting 22, 239–247. https://doi.org/10.1016/j.ijforecast.2005.08.002

Boudette, N.E., 2024. How Ford's F-150 Lightning, Once in Hot Demand, Lost Its Luster. International New York Times NA-NA.

Chatfield, C., 1986. Exploratory data analysis. European Journal of Operational Research 23, 5–13. https://doi.org/10.1016/0377-2217(86)90209-2

Fast RobustSTL: Efficient and Robust Seasonal-Trend Decomposition for Time Series with Complex Patterns | Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining [WWW Document], n.d. URL https://dl.acm.org/doi/abs/10.1145/3394486.3403271 (accessed 12.12.24).

Fernandez, G., 2021. Tesla and the Stock Market. https://doi.org/10.2139/ssrn.3908812

Hawkins, T.R., Gausen, O.M., Strømman, A.H., 2012. Environmental impacts of hybrid and electric vehicles—a review. Int J Life Cycle Assess 17, 997–1014. https://doi.org/10.1007/s11367-012-0440-9

Keohane, D., Campbell, P., Bushey, C., 2024. "Told you so" moment for Toyota on hybrids. Carmaker enjoys measure of vindication after warning repeatedly that consumers would balk at going full electric. The Financial Times 9–9.

Koreisha, S.G., Pukkila, T.M., 1993. New approaches for determining the degree of differencing necessary to induce stationarity in ARIMA models. Journal of Statistical Planning and Inference 36, 399–412. https://doi.org/10.1016/0378-3758(93)90140-2

Liao, L., Li, H., Shang, W., Ma, L., 2022. An Empirical Study of the Impact of Hyperparameter Tuning and Model Optimization on the Performance Properties of Deep Neural Networks. ACM Trans. Softw. Eng. Methodol. 31, 53:1-53:40. https://doi.org/10.1145/3506695

Ning, Y., Kazemi, H., Tahmasebi, P., 2022. A comparative machine learning study for time series oil production forecasting: ARIMA, LSTM, and Prophet. Computers & Geosciences 164, 105126. https://doi.org/10.1016/j.cageo.2022.105126

Ou, S., Lin, Z., Wu, Z., Zheng, J., Lyu, R., Przesmitzki, S.V., He, X., 2017. A Study of China s Explosive Growth in the Plug-in Electric Vehicle Market (No. ORNL/TM-2016/750). Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States). National Transportation Research Center (NTRC). https://doi.org/10.2172/1341568

Sigal, P., 2024. Hybrids take Toyota to No. 2 in Europe; Automaker credits attention to popular segments, all markets. Automotive News 98, 0008–0008.

Study and analysis of SARIMA and LSTM in forecasting time series data - ScienceDirect [WWW Document], n.d. URL https://www.sciencedirect.com/science/article/abs/pii/S2213138821004847 (accessed 12.12.24).

Witt, A., Kurths, J., Pikovsky, A., 1998. Testing stationarity in time series. Phys. Rev. E 58, 1800–1810. https://doi.org/10.1103/PhysRevE.58.1800

Zhang, X., Bai, X., 2017. Incentive policies from 2006 to 2016 and new energy vehicle adoption in 2010–2020 in China. Renewable and Sustainable Energy Reviews 70, 24–43. https://doi.org/10.1016/j.rser.2016.11.211