Big Data Processing and Applications

Electric Vehicle Population Data Analysis

Name	Email	University
Muhammad Ahmed	Muhammad.ahmed@student.oulu.fi	University of Oulu (Finland)

Project Code

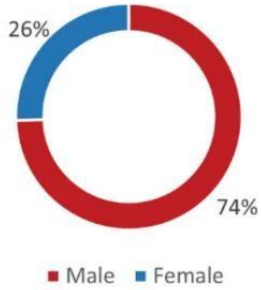
Project description

The automotive industry plays a vital role in transportation, but it demands continuous improvement to enhance energy efficiency. Electric cars offer numerous benefits, aiming to create a better world with their low operating costs, eco-friendliness, and energy efficiency. Despite industry advancements, achieving certain goals requires consideration of government policies and marketing strategies. Through data analysis, we aimed to understand why some areas embrace electric vehicles more readily. Factors such as charging infrastructure, battery technology advancements, and environmental awareness may influence people's decisions. We will be using Apache Spark and some libraries of data visualization to analyze the data trends after processing the data and later we will aim to project the electric vehicles sales in the coming years.

Related work

The increasing appeal of electric vehicles (EVs) is credited to their environmental benefits and cost-saving potential. However, selecting the most appropriate EV models for a particular location remains challenging, primarily due to diverse consumer preferences and inherent limitations of EV technology. With the growth of the EVs Industry, researchers are proposing different algorithms to analyze the data and extract information from it that can be used further to enhance the usage and production. In a standard growing mode, EV is expected to dominant the new sales market by 2030 in Shanghai, and to completely replace ICEV before 2050 [1]. In the paper [2], different analysis were performed regarding the distribution of gender for the usage of EVs and results were found out as 74 percent males and 26 percent females depicted in 300,000 posts.

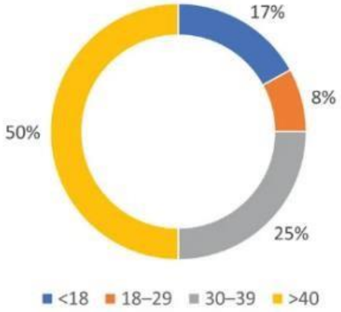
Fig. 1
Gender distribution



Note: https://www.mdpi.com/2071-1050/16/1/ 305#fig body display sustainability-16-00305-f001

Further analysis were carried out to find the age groups using EV’s and people above age of 40 were found to be using around 50 percent of the total EV’s found in the dataset as depicted in the graph below.

Fig. 2.
Age distribution



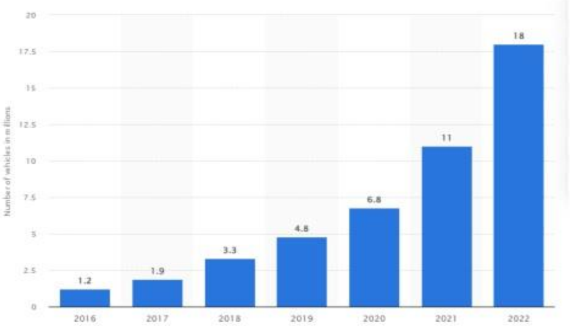
Note: https://www.mdpi.com/2071-1050/16/1/305# fig body display sustainability-16-00305-f001

Young generation from the age group of 18 to 29 were found to be the least in the analysis to be using EV's. Through careful analysis, researchers have uncovered important topics such as the need for better charging stations and tax incentives that appeal to people from all walks of life.

These insights come from our methodical approach, which looks at people's feelings and uses computer programs to understand what they care about the most. Our study's focus on charging station issues demonstrates how important it is to address these issues for all people, regardless of identity. This relationship between our research methodology and findings indicates that our approach is effective in capturing the emotional states of individuals. With the information available, a number of studies are examining the global popularity of electric vehicles. The global number of battery-electric vehicles in use from 2016 to 2022 is shown in the graph below. The graphs' units are expressed in million.

Fig. 3.

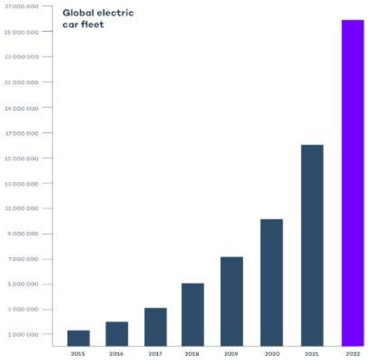
Worldwide number of battery electric vehicles in use from 2016 to 2022



Note: https://www.statista.com/statistics/270603/ worldwide-number-of-hybrid-and-electric-vehicles-since-2009/

From Fig 4, sudden increase in the number of EV's can be observed in 2022 just after the years of pandemic. According to a report [3] in 2020, there was a notable lack of significant growth in overall new car registrations worldwide. The global automotive market, encompassing all vehicle types, experienced a downturn due to the COVID-19 pandemic and ensuing economic challenges. The outlook for global electric vehicle (EV) sales was initially uncertain amidst the pandemic. However, 2020 ultimately defied expectations, witnessing a remarkable surge. EV sales globally soared by 43 percent compared to 2019, with the electric car industry capturing a record market share of 4.6 percent. The automotive market is experiencing rapid expansion, with growth evident across regions. Fueled by the urgent need for decarbonization, which is now being prioritized by many leading nations, and bolstered by a range of policies and incentives, global electric vehicle (EV) sales continue to surge in 2023. The first quarter of the year saw an impressive 2.3 million EVs sold, marking a 25 percent increase compared to the same period in 2022. Projections indicate that by the end of 2023, we could witness a staggering 14 million EV sales, potentially constituting 18 percent of total car sales worldwide whereas another report [4] states that in 2022, there were remarkable strides in the adoption of electric cars and vans within the European Union, with electric vehicles representing 30.24% of new car registrations. This translated to nearly 2.8 million electric car registrations within the year, marking a substantial rise from 1.22 million in 2021. Additionally, the prevalence of electric vans on European roads continued to ascend, accounting for 7.7% of new registrations in 2022. Over the past year, there was a 35% surge in the number of newly registered battery electric vehicles, while the count of plug-in hybrid cars remained constant. Notably, battery electric vehicles dominated the registrations of electric vans in 2022.

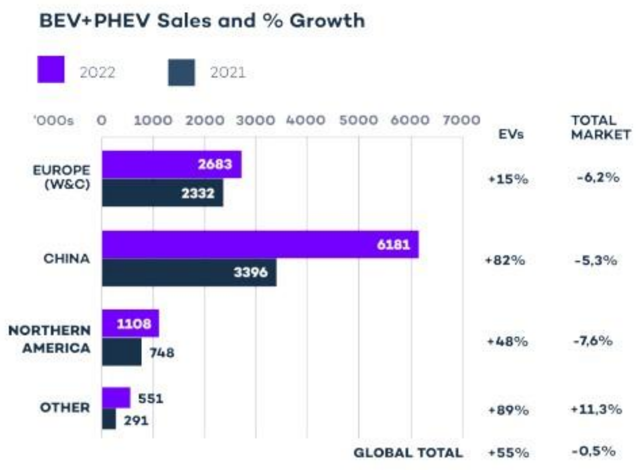
Fig. 4.
Worldwide number of battery electric vehicles in use from 2016 to 2022



Note: **https://www.virta.global/en/global-electric-vehicle-market**

Plug-in hybrid cars remained constant. Notably, battery electric vehicles dominated the registrations of electric vans in 2022 Breaking down the analysis by each country further, the report [5] States that the top 5 countries with the highest share of EV sales are Norway (all-electric vehicles made up 80%of passenger vehicle sales in 2022), Iceland (41%), Sweden (32%), the Netherlands (24%) and China (22%) Another study [6], highlights the impact of tax exemption policies on the proliferation of electric cars in Norway. According to analysis by the World Economic Forum, these policies have significantly contributed to the surge in electric vehicle adoption in the country. As a result, Norway has achieved a high ratio of electric vehicles per capita compared.

Fig. 5.
Global Electric car fleet

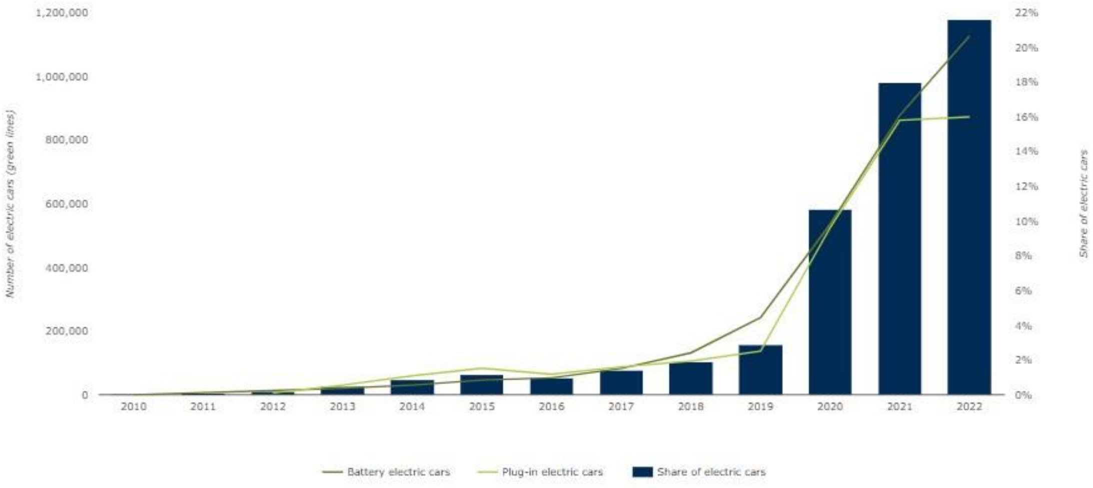


Note: https://www.statista.com/statistics/ 270603/worldwide-number-of-hybrid-and-electric-vehicles-since-2009/

Furthermore, the Norwegian government has set an ambitious goal to phase out the use of oilfueled cars by 2023, aiming to transition entirely to electric vehicles. Following Norway, Iceland emerges as the second country with the highest electric car population. With 36.8 percent of electric vehicles per 1,000 population, Iceland has experienced a rapid increase in electric vehicle adoption since 2017, when the share of electric vehicles stood at a mere 8.7 percent.

Fig. 6.

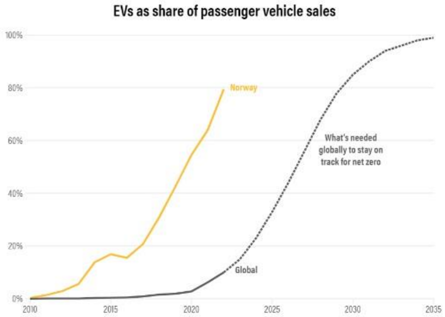
New registrations of electric cars, EU-27



Note: https://www.eea. europa.eu/en/analysis/indicators/new-registrations-of-electric-vehicles

Fig. 7.

New registrations of electric cars, EU-27



Note: https://www.eea. europa.eu/en/analysis/indicators/new-registrations-of-electric-vehicles

Data description

The data [7] used in this project for analysis is sourced from the Data.gov USA website, providing open data pertaining to battery electric vehicles (BEVs) and plug-in hybrid electric vehicles (PHEVs) registered through the Washington State Department of Licensing (DOL). The dataset contains various attributes including Vehicle Identification Number (VIN), county, city, model year, manufacturer, electric vehicle type, Clean Alternative Fuel Vehicle (CAFV), electric range, base MSRP, legislative district, DOL vehicle ID, vehicle location, electric utility, and 2020 Census Tract. The data encompasses all ownership types and areas in Washington, covering federal, investor-owned, municipal, political subdivision, and cooperative territories.

✞ Data Volume and License

The dataset consists of 177866 records. As it is sourced from the Data.gov USA website, the data is openly available for use. However, it's crucial to review the specific license agreement associated with the dataset for any restrictions or requirements.

✞ Data Temporality and Meaningful Statistics

The dataset contains temporal elements such as model year and census tract identifiers from the 2020 census. Mean, median, minimum, maximum, mode, and other quantiles can be calculated for numerical attributes such as electric range. Statistical analysis can provide insights into the distribution and variability of these attributes, aiding in understanding trends and patterns within the data.

✞ Structural Details and Data Pre-processing

The dataset is organized and predominantly comprises both text and numeric data fields. During our analysis, we identified certain columns that were irrelevant and consequently excluded them from our study. For instance, the "state" column was redundant as the dataset exclusively pertained to Washington, thus we opted to discard it. Similarly, the "Manufacturer's Suggested Retail Price

(MSRP)" column lacked pricing information for the vehicles, prompting its removal from consideration. Missing values in the electric range column were filled by computing the mean value of the entire column and fill in those missing values. Below is a detailed description of the fields:

Table 1. Data description example

Field Name	Description	Type	Example
VIN	Vehicle Identification Number	Text	1HGCM82AN
County	Geographic region of vehicle owner's	Text	King
City	City of vehicle owner's residence	Text	Seattle
Postal Code	Postal code of the location where the vehicle is registered	Numeric	98597
Model	Model year of the vehicle	Numeric	2019
Make	Manufacturer of the vehicle	Text	Tesla
Electric Vehicle Type	Battery Electric Vehicle (BEV) or Plug-in Hybrid Electric Vehicle (PHEV)	Text	BEV
Clean Alternative Fuel Vehicle (CAFV) Eligibility	Categorization of vehicle as Clean Alternative Fuel Vehicle or Electric only range requirement	Text	Eligible
Electric Range	Distance vehicle can travel purely on electric	Numeric	250 miles
Legislative District	Specific section of Washington State for vehicle owner	Numeric	37
DOL Vehicle ID	Unique number assigned by Department of	Numeric	12345678
Vehicle Location	Center of ZIP code for registered vehicle	Text	98101
Electric Utility	Electric power retail service territory	Text	Puget Sound Energy
2020 Census Tract	Combination of the state, county, and census tract codes.	Numeric	53033000100

Methods and tools

A. Environment Setup

Our research relied on a strong setup designed for handling lots of data and doing complicated analyses. We used Python and Apache Spark because they’re good at dealing with big datasets and tough jobs.

- Programming Environment:

  At the core of our environment is **Python 3.11**, chosen for its extensive library support and ease of use in scientific computing. Our project relied on key Python libraries, each serving a specific role in our data analysis pipeline:

  - **Matplotlib** and **Seaborn** for generating visualizations, crucial for data exploration and communication of findings.
  - **Apache PySpark MLLib** for developing and evaluating machine learning models, particularly focusing on regression analyses.

- **Distributed Computing Environment**:

  Our analytical capabilities were significantly enhanced by leveraging Apache Spark, enabling scalable data processing across multiple nodes. Within this environment, we utilized:

  - **Sark SQL** and **DataFrames** for structured data processing, applying complex transformations and aggregations with ease:
  - **Window Functions** and **Regular Expression Functions** for sophisticated data analysis and preprocessing techniques
  - A variety of data processing functions, such as col, **concat_ws**, and **count**, further extended our ability to manipulate and summarize datasets. Our setup was super important for our research. It made it easy to do data analysis and play around with different computer models. This helped us understand our data better and discover new things about it.

B. Spark Cluster Overview

Our data processing pipeline utilizes Apache Spark for efficient distributed computing. The Spark cluster comprises: ✞ Master Node

Coordinates tasks and manages worker nodes.

- Worker Nodes

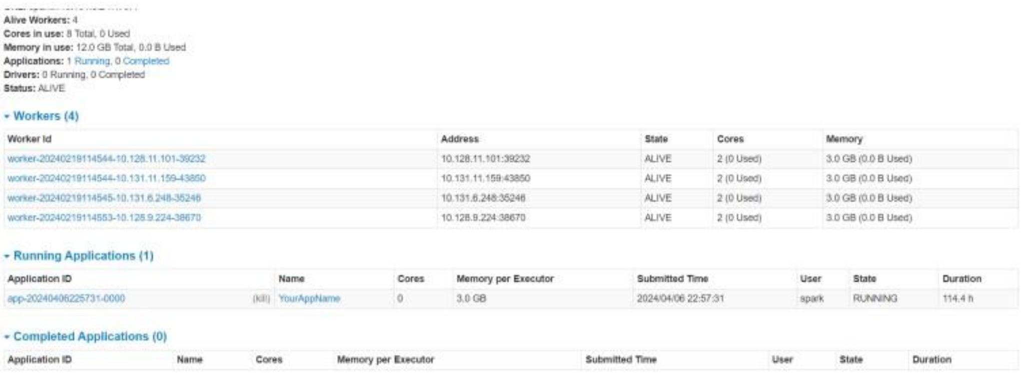
  Execute tasks distributed by the master node.

- Driver Node

  Acts as the application entry point, coordinating with the master node.

Fig. 8.

Spark Cluster Topology



Alive Workers: 4
Cores in use: 8 Total, 0 Used
Memory in use: 12.0 GB Total, 0.0 B Used
Applications: 1 Running, 0 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (4)

Worker Id	Address	State	Cores	Memory
worker-20240219114544-10.128.11.101-39232	10.128.11.101:39232	ALIVE	2 (0 Used)	3.0 GB (0.0 B Used)
worker-20240219114544-10.131.11.159-43850	10.131.11.159:43850	ALIVE	2 (0 Used)	3.0 GB (0.0 B Used)
worker-20240219114545-10.131.6.248-35246	10.131.6.248:35246	ALIVE	2 (0 Used)	3.0 GB (0.0 B Used)
worker-20240219114553-10.128.9.224-38670	10.128.9.224:38670	ALIVE	2 (0 Used)	3.0 GB (0.0 B Used)

Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration
app-20240406225731-0000 (kill)	YourAppName	0	3.0 GB	2024/04/06 22:57:31	spark	RUNNING	114.4 h

Completed Applications (0)

Application ID	Name	Cores	Memory per Executor	Submitted Time	User	State	Duration

C. Tools and Libraries Used

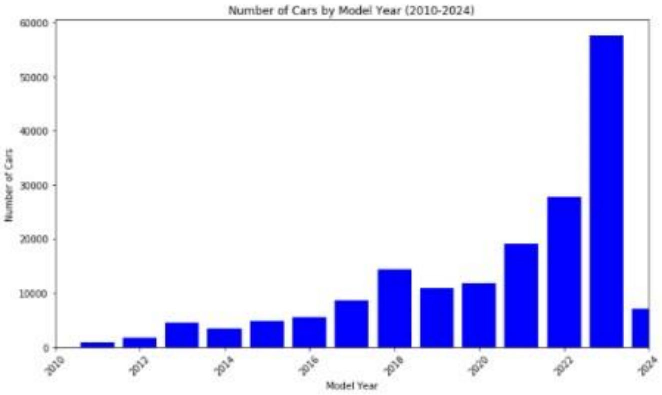
Our project utilized a suite of computational tools and libraries, pivotal for analyzing data and deriving meaningful insights. These tools enabled us to handle, visualize, and interpret complex datasets effectively. Key tools and libraries employed in our research include.

- **Python 3.11**
  Served as our primary programming language, known for its efficiency and versatility in scientific computing.
- **Matplotlib and Seaborn**
  Used for crafting visual representations of our data, enabling intuitive analysis and insights.
- **Apache PySpark MLLib**
  Powered our machine learning models, particularly for implementing regression analyses, aiding in predictive analysis. Apache Spark: A critical component for our distributed computing needs, allowing us to process large datasets efficiently

Data analysis and Results

After data refinement, we have found out the general trend of cars across each year from 2010 to 2024 whereas the records of 2024 were very few as latest. By the graph a rapid increase in the usage of EVs could be observed in the year of 2022 and 2023. Since data has very few records of 2024, so graph bar of 2024 is small. From the years 2010 to 2016 trend seems to be very minimum and in the era of 2019 to 2020 due to pandemic, trends show a significant consistency as of 2018.

Fig. 9.
Bar chart illustrating the annual number of cars registered in Washington, USA, from 2010 to 2024



Further, we have analyzed the data and figured out annual car counts in the top 5 counties from 2010 to 2024. In the graph below, each color denotes a different county. The distribution highlights the leading counties by car presence each year, illustrating the changing automotive landscape over the 15 years. We can see that King County has the highest number of electric cars, exhibiting steady growth from 2010 to 2022. It has outperformed many other counties to become a leader in electric car usage.

Below graph shows the number of cars across each county in the dataset. Bar chart showcasing the disparity in car counts among the top 20 counties. The data reveals a steep decline from the county with the highest number of cars to subsequent counties, indicating a concentration of vehicles in the leading county far exceeding that of its counterparts.

Fig. 10.

Scatter plot representing the annual car counts in the top 5 counties from 2010 to 2024

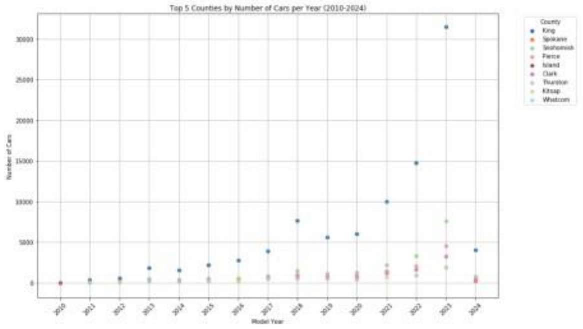
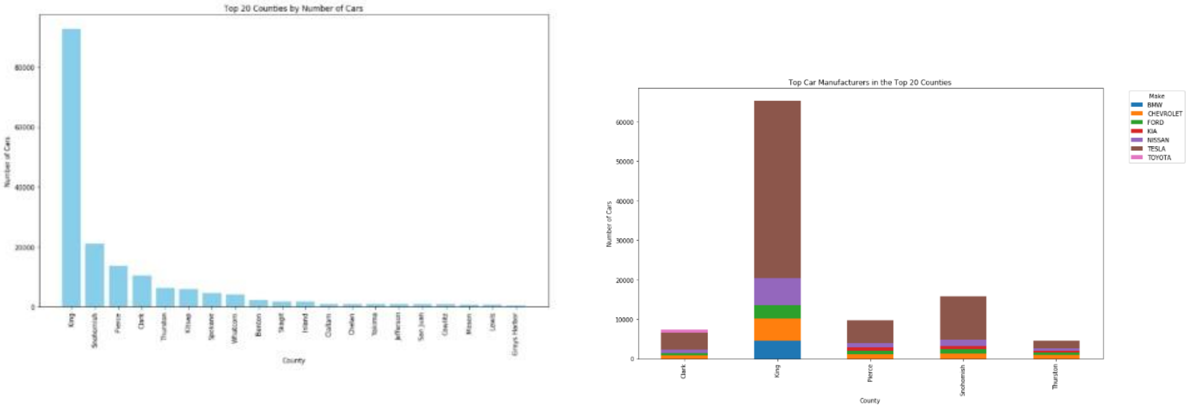


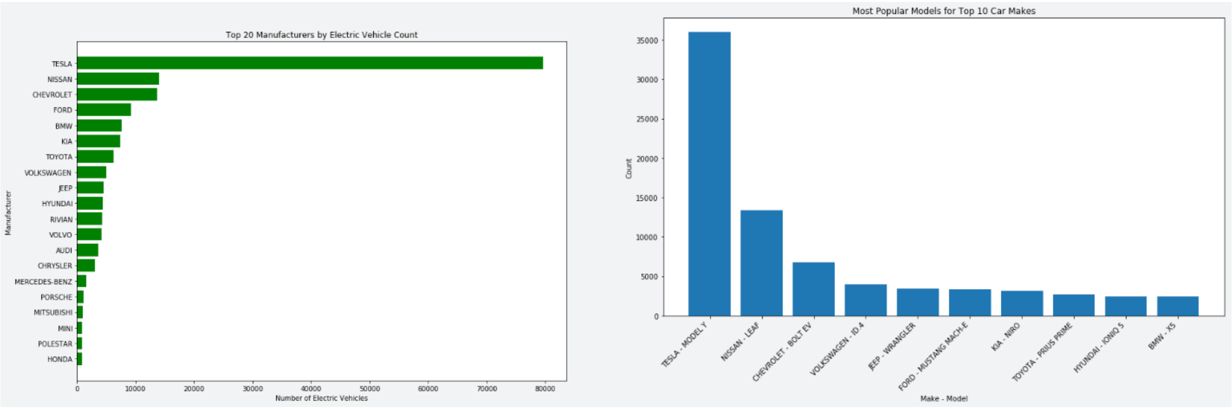
Fig. 11.

Bar charts showcasing the disparity in car counts among the top 20 counties on the left and top 5 countries car manufacturers on the right



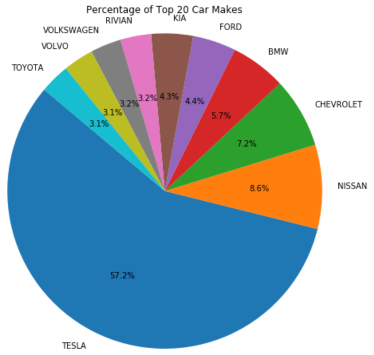
On the right side of figure, Stacked bar chart showing the comparative presence of car manufacturers across the top 20 counties. Each bar represents a county, segmented into colors that correspond to different car manufacturers. The chart highlights the dominance of specific manufacturers in certain counties and the overall diversity of vehicles within the regions analyzed.

Fig. 12.
The company makes most of the electric vehicles on the left and most popular 10 models in those countries on the right.



A horizontal bar chart displaying the top 20 electric vehicle manufacturers ranked by the number of vehicles produced. The lengths of the bars reflect the quantity of electric vehicles for each manufacturer, highlighting the leading contributors to the electric vehicle market and on the right-side bar chart presenting the most popular models for the top 10 car makes. The makes and models are arranged on the x-axis with corresponding model counts on the y-axis, demonstrating a clear decline in frequency from the most to the least common model within the dataset.

Fig. 13.
Country and card manufacturer ratio on the right and the maximum electric range of the top 20 car manufacturers



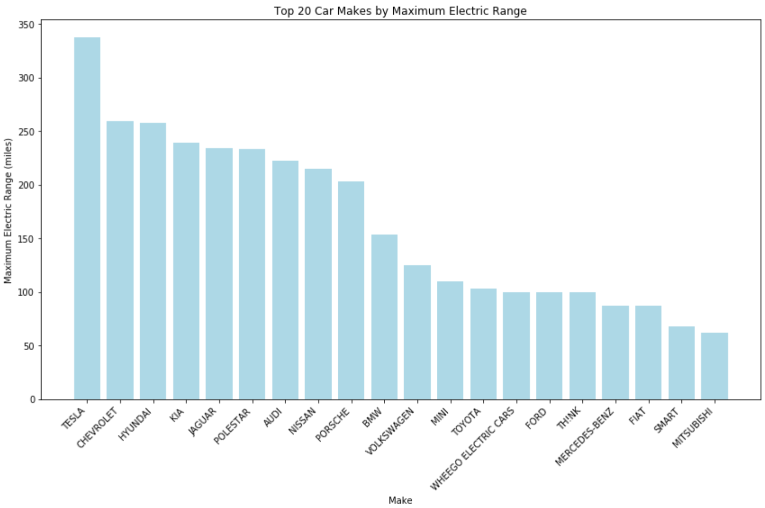
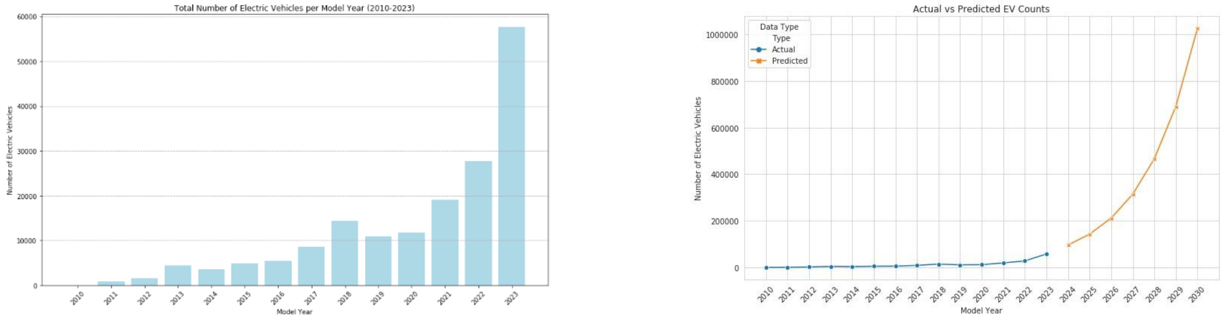


Fig. 14.

Total number of electric vehicles on the left until 2023 and our prediction to 2030 on the right



The graph displays the actual and predicted electric vehicle (EV) counts from 2010 to 2030. The blue line represents actual historical EV data up to the year 2023, showing a gradual increase over time. The orange line illustrates the predicted counts from 2024 onwards, indicating an exponential growth pattern, with the number of EVs sharply rising each year. The plot includes a legend distinguishing between actual data and predictions, with the 'Model Year' on the x-axis and the Number of Electric Vehicles' on the y-axis

Conclusion

As we come to the end of our project, it's been quite a journey. We've been focused on reaching our main goal of finding data trends from the data and analyze them based on the literature review. We've tried out different ways and tools to help us, but it hasn't all been easy. One of the hardest parts was dealing with the setup of Apache Spark but documentation helped us a lot which made it tough to use some of the ways and tools we wanted to.

But it's not all been bad news. We've had some real successes, especially in finding data trends from the data after processing it. Those wins have taught us a lot about what works and what doesn't in this field. Looking ahead, we will try to find some stats and possible outcomes of government laws for electrical vehicles and based on that we will try to estimate the projection for coming years.

Contribution report

This chapter explains how each student contributed to the project, in detail.

- Muhammad Talha Arshad
    - R & D on literature
    - Report Compilation
- Muhammad Ahmed
    - Coding
    - Environment Setup
- Aamir Sohail
    - R & D on proposed work
    - Report Compilation

References

[1] Shi, Y., Feng, D., Yu, S., Fang, C., Li, H., & Zhou, Y. (2022). The projection of electric vehicle population growth considering scrappage and technology competition: A case study in Shanghai. Journal of Cleaner Production, 365, 132673.

[2] Priyam, T.; Ruan, T.; Lv, Q. Demographic-Based Public Perception Analysis of Electric Vehicles on Online Social Networks. Sustainability 2024, 16, 305. https://doi.org/10.3390/su16010305

[3] The Global Electric Vehicle Market In 2024." Virta Global. Retrieved April 30, 2024, from https://www.virta.global/en/global-electric-vehicle-market (Published on March 04, 2024)

[4] New registrations of electric vehicles in Europe." European Environment Agency's home page. Retrieved April 30, 2024, from https://www.eea.europa.eu/en/analysis/indicators/new-registrations-of-electric-vehicles

[5] World Resources Institute. (2023, September 14). These Countries Are Adopting Electric Vehicles the Fastest. World Resources Institute. Retrieved April 7, 2024, from https://www.wri.org/insights/ countries-adopting-electricvehicles-fastest

[6] OI. (2022, October 12). List 4 Countries with The World's Most Electric Car Population. VOI. Retrieved April 7, 2024, from https: //voi.id/en/economy/215422

[7] Catalog. (2024, April 19). State of Washington - Electric Vehicle Population Data. Retrieved April 30, 2024, from https://catalog.data.gov/dataset/electric-vehicle-population-data