Analysis of Electric Vehicles and Charging Infrastructure

Kanchan Naik Mrunali Katta Prasad Shimpatwar Yashasvi Kanchugantla
Dept. of Applied Data Science
San Jose State University

***Abstract*—Integrating the available diverse electric vehicle data in the market we provide an analysis of the electric vehicles' growth, Existing charging infrastructure, customer demographics, energy consumption, and also if the current EV infrastructure can support Electric Vehicle needs, etc. We aim to help stakeholders, including government organizations, business owners, and consumers to get insights into market growth, infrastructure adequacy, and future needs helping them make strategic decisions toward sustainable mobility.**

Key Terms - ***Database Management, EVs, Database and query design, ETL, Data Warehouses, Neo4j, Kuzu, BigQuery, Tableau, Python, Jupyter Notebook, MySQL, Google Trifacta Dataprep (ETL), Data visualization, NoSQL, RDBMS, web crawling***

***Index Terms*—data analytics, project report**

I. Introduction

A. Motivation

According to an article published in the scientific journal Nature titled "Oil reserves will be depleted by 2043" [1], In the past few years, we have always encountered discussions about climate change and increasing global warming due to pollution. Not only pollution but also shrinking sources of fossil fuel can result in drastic effects on energy production across the globe. Electric vehicles have almost zero direct pollutant and greenhouse gas emissions in the environment; hence, EVs contribute significantly to improving air quality and reducing carbon footprint. In addition to this, we recently came across an article "California seeks EPA approval to ban sales of new gasoline-only vehicles by 2035". Recently many countries have set a goal to completely decarbonize the transport sector by 2050[1]. To achieve this goal huge growth in charging infrastructure is required. To explore this further, we researched the growth of EVs and found the following graph which shows there has been a tremendous growth in the EV population in the past few years[6]. This motivated us to explore the topic of analyzing existing EV charging infrastructure and energy consumption during the fleet.

B. Problem Statement

Analyze the growth trends to assess EV ecosystem and its readiness, spatially and temporally, through the following:

- Electric Vehicles
- Charging Stations
- Energy Supply and Demand

C. Scope and Limitations

Literature review of existing electric vehicles, customer demographics, EV energy consumption, and charging infrastructure. Perform ETL on CSV, Web pages, JSON, pdf data marts available. Stores into BigQuery and perform visualizations. Evaluate visualizations of existing trends, and patterns and come up with recommendations for government organizations, transportation business owners, states, and existing trends and patterns.

We assumed median energy consumption values for different vehicle models and charging equipment. Limitations on free/ open source technologies and the availability of public datasets restricted our analysis to go granular. Forecasting has not been done. We considered only historic data.

II. Sources-DataMarts

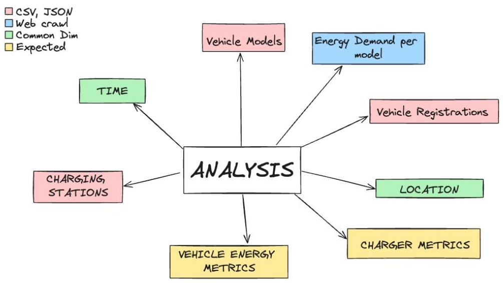


Fig. 1. Schema used for Data Warehousing

The project talks about growth of electric vehicles and supporting infrastructure. Datasets are surveyed from government websites, EV analytics companies - OpenChargeMap, public datasets and experimental results from EV companies etc., and available web pages for vehicle models and their specifications. From plethora of available data, we have picked 4 datasets:

The Alternative Fuels Data Center (AFDC) provides information, data, and tools to help fleets, fuel providers, policymakers, cities, states, Clean Cities and Communities coalitions, and other transportation decision makers find ways to reach their energy, environmental, and economic goals through the use of alternative and renewable fuels, advanced vehicles, and other fuel-saving strategies.

1) Vehicle Models:

- **Source** AFDC [2]
- **Format** CSV
- **Description** Vehicle Model, Model year, Fuel Type, Manufactures data from 2016-2024.
- **Ingested into** BigQuery

2) Vehicle Registrations:

- **Source** AFDC [3]
- **Format** CSV
- **Description** Year, State, Fuel Type, Number of vehicles registered from 2016-2022.
- **Ingested into** BigQuery

- **Source** Data Atlas [3]
- **Format** CSV
- **Description** Year, State, Vehicle Make, Vehicle Model, Zip Code, Day, Month
- **Ingested into** BigQuery

3) Customer Demoraphics:

- **Source** Nyserda [4]
- **Format** CSV
- **Description** Questions, and answer choice
- **Ingested into** BigQuery

4) Charging Stations:

- **Source** AFDC and OpenChargeMap's public Dataset [5]
- **Format** JSON, CSV
- **Description** Year, State, Access Code, Number of Level1 Charging, Number of Level 2 Charging, Charging sockets type..
- **Ingested into** BigQuery and Neo4j

5) Energy Demand Estimation:

- **Source** AFDC [2]
- **Format** Web crawling with a python script
- **Description** Year, State, Access Code, Number of Level1 Charging, Number of Level 2 Charging, Charging sockets type..
- **Ingested into** MySQL and BigQuery

III. Methodology

A. Data Collection

- Vehicle Models data from 2016 - 2024 for all categories [1]
- Vehicle Registration data from 2016-2024 [1]
- Existing fuel station data from 2014-2024 [6]
- Energy Demand Estimation by Vehicles

B. Workflow

2The data collection involved gathering information from various sources, including Excel spreadsheets, JSON files, PDF, and web crawling techniques. The raw data was imported into the cloud using Google Cloud Platform (GCP) for further processing and cleaning. To organize the data efficiently We created multiple buckets based on the type of data we are using and imported data from the above-mentioned sources. Each bucket was used as a repository for a specific category of data making it easier to manage and access during subsequent data processing stages. Handling data from PDF format was specifically challenging due to sheer scale of data. In order to solve this problem we leveraged the 'Get Data' feature of Excel for extracting data from various file types. We cross-validated data with the original source so that no data is overlooked and lost during transformation. For certain data we stored in RDBMS (MySQL) after extracting it from web using web scraping.

The Extract, Transform, Load (ETL) process plays crucial role in preparing data for analysis. ETL was performed on the cloud using the Google Trifacta DataPrep tool, In DataPrep the data was imported from Google Cloud, and cleaned using different recipes. Recipes encompasses of tasks such as dropping NaNs or null values, converting columns to appropriate datatypes(integers, state codes, dates..), combining various datasets with a unified Time dimension, replacing certain values to handle mismatch and inconsistent values, and other relevant columns.

After cleaning and preparing the data we loaded it into BigQuery using DataFlow, BigQuery was used as a central repository for cleaned and transformed data. We read data from BigQuery using Python and Tableau to create interactive dashboards. The combination of cloud-based data preparation, cleaning, and storage, along with the use of APIs and Jupyter Notebook for data processing and visualization, demonstrates a comprehensive approach to data management and analysis.

C. Analysis

Adoption of electric vehicles: As shown in all the Fig. 3(a), 3(b), 3(c), Registrations started increasing from 2019, so are the hybrid and electric models offered by vehicle manufacturers. Growth of electric vehicles can be emphasize by the decrease in models released in other fuel type categories.

Fig. 3(d) From the figure we can see that the number of public charging points is way more compared to private charging stations also growth in public and private charging stations over the years also has significant disparity. From this trend, we can say that not many people or offices installed charging private stations. One of the reasons could be expenses for private charging stations. This gives us all the more necessity to study this problem. From 2023 there are not many stations opened. But new vehicles are still coming into the market from Fig. 3(c) on registrations. Which makes states that our infrastructure is not adequate.

Fig. 3(a) From Fig we can observe that companies are starting to launch more and more Electric or Hybrid Electric models into the market. Companies have decreased investing in other fuel models. A few years back diesel and gasoline vehicles were prominent in the market now this trend has shifted towards electric and hybrid electric models. Hybrid

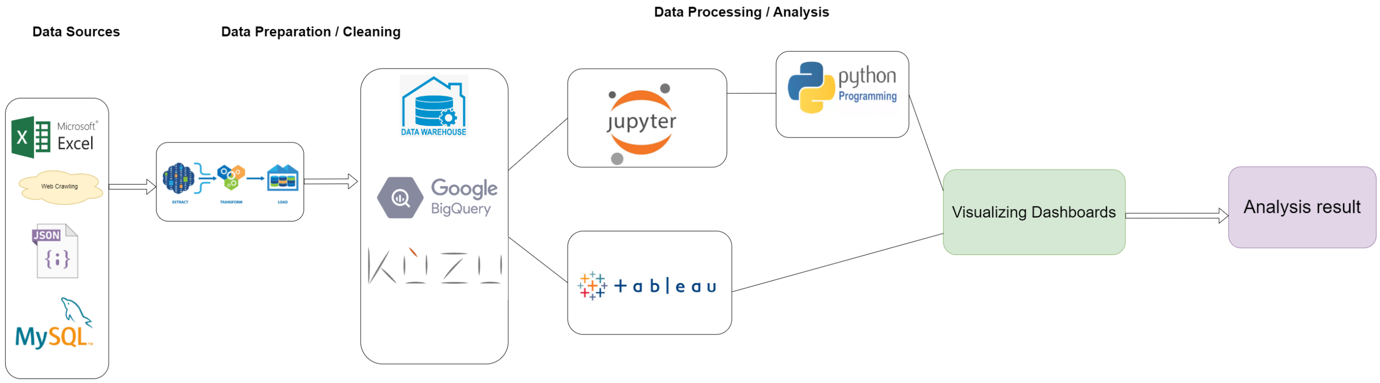


Fig. 2. Workflow for the data systems used

electrics have increased more than others. This can be read a people wanting to rely on both fuel types, Which might be telling us that there are not enough charging stations in the market or in remote areas.

EV registrations are rapidly increasing, but a fall observed in charging stations opening. This creates a lot of gaps to fill the energy demand that is being created.

From the above two visualizations, we can see that California has the highest number of electric vehicles compared to the rest of the US. Hence there is still scope for EV adoption in other states. The population-dense states have adopted EVs more compared to lightly populated states. In Fig. 4(e) we can see that even though California has many charging stations these are very few compared to the number of electric vehicles on road. Also, from the Fig. 4(f). we can see that charging station availability is good only in major cities compared to remote areas. Even in California which has the highest number of electric vehicles has sparse charging station availability in remote areas.

IV. Evaluation

A. Vehicle Load on a Charging station:

To understand the sufficiency of charging stations for electric vehicles coming into the market for a state, we came up with a metric - Vehicle Load. Vehicle Load statistic for a state is the average number of vehicles every charging station is catering to meet the energy demand. This metric is used to understand if there are more electric vehicles needing energy than an charging station is capable to handle. In simple terms, it is number of patients per doctor as a metric for healthcare in countries or number of students per professor as a metric for universities.

$$\text{Vehicle load} = \frac{\text{Total vehicles in a state}}{\text{Total charging stations in a state}} \tag{1}$$

Fig. 5(b) shows the decreasing load on charging stations over the years. 2021 is the year where a number of vehicles per charging station dipped for states with high EV adoption. NewYork's load is the least with only 319 vehicles per charging station in 2021. Idaho's load stands highest at 690 EVs per charging station. This means that in Idaho, there is the highest stress compared to other stress.

B. Energy Supply = Energy Demand ?

Now with the number of charging stations per vehicle, we can only get limited insights. Charging equipment for EVs is classified by the rate at which the batteries are charged. Charging times vary based on how depleted the battery is (i.e., state-of-charge), how much energy it holds (i.e., capacity), the type of battery, the vehicle's internal charger capacity, and the type of charging equipment (e.g., charging level, charger power output, and electrical service specifications). The charging time can range from less than 20 minutes using DC fast chargers to 20 hours or more using Level 1 chargers, depending on these and other factors [7].

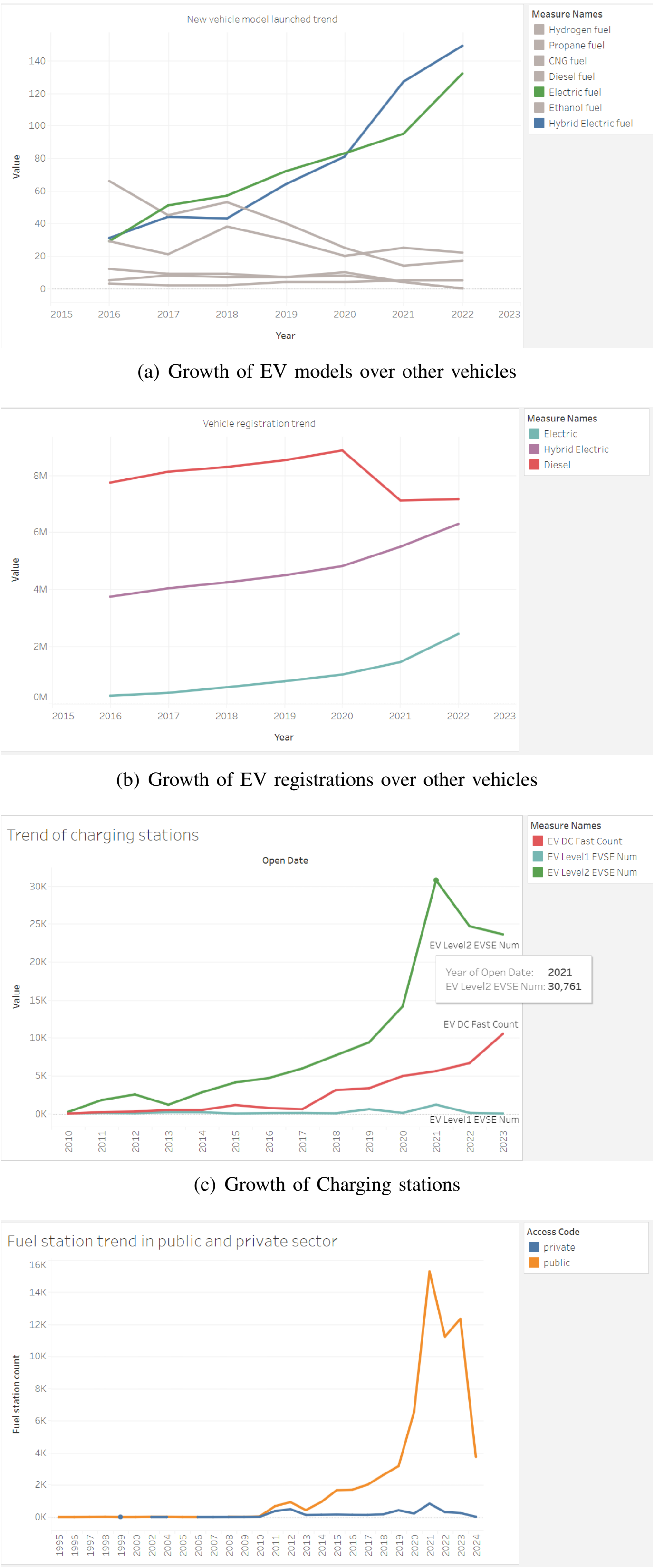
The next level of granularity is the capacity of energy each charging station can offer to the vehicles. Our dataset is limited, we could not fetch energy contributed by a charging station specifically. Taking the number of miles one hour of charging at a charger can boost from Fig. 7, which is mentioned in [7] and an average utilization time per day, we have come up with a metric, Number of miles supplied by a charging station.

$$\text{Miles in a week per station} = 126 \times \sum_{i=1}^{3} \text{Level}_i \times \text{miles/hour}_i \tag{2}$$

where:

- Level_i is the charging level (1, 2, or 3) for the i-th type of charger
- The constant 18 represents the average number of hours charged in a charging station per day
- The constant 7 represents the number of days in a week
- The summation is taken over all types of chargers ($i = 1$ to 3)

Miles of demand in the market. [8] shows us that an average vehicle in US is driven for 300 miles every week. So, we have taken that figures as the amount of demand created for every new electric vehicles. Matching is done for energy balance



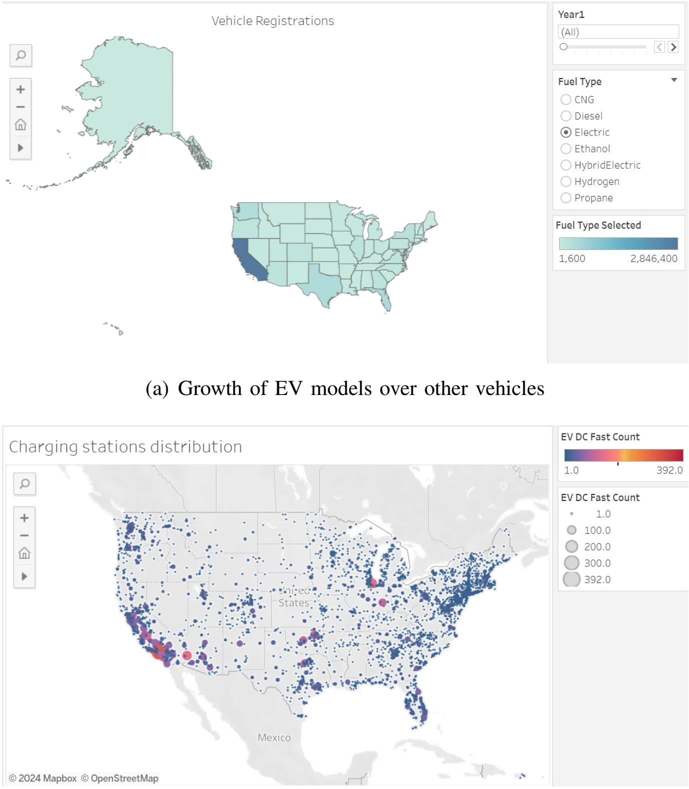
(a) Growth of EV models over other vehicles

(b) Growth of EV registrations over other vehicles

(c) Growth of Charging stations

(d) Growth of private and public Charging stations

Fig. 3. Growth trends of Electrics vehicles and ecosystem



(a) Growth of EV models over other vehicles

(b) Growth of EV registrations over other vehicles

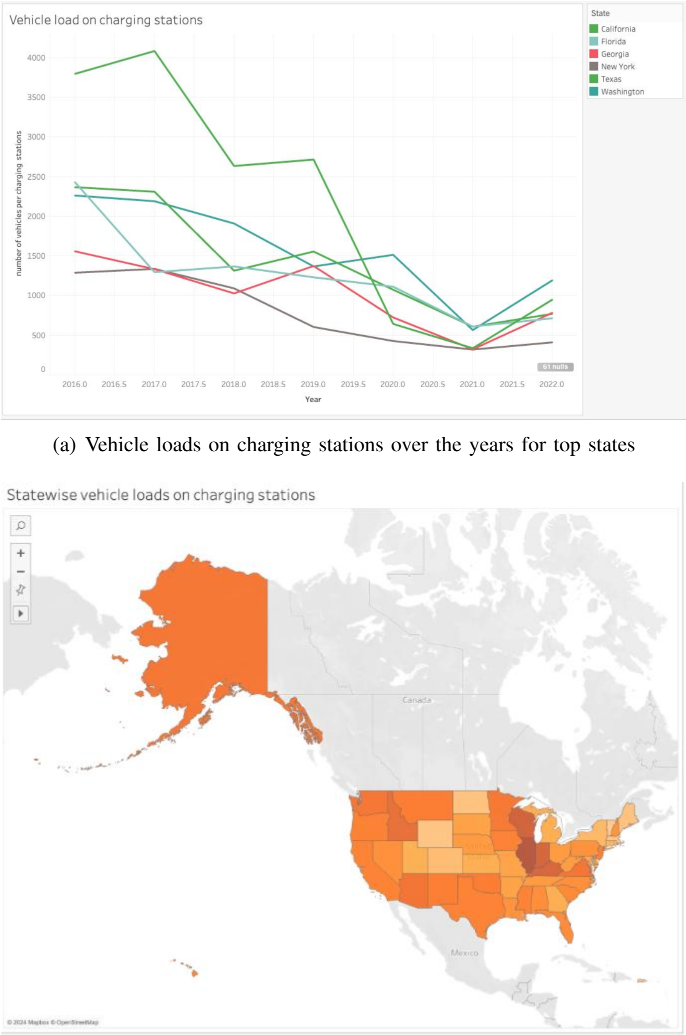
Fig. 4. statewise distribution of Electrics vehicles and ecosystem

based on the miles demanded by vehicles and supplied by charging stations.

Fig. 6(a) Represents a scatter plot for years with miles supplied by charging stations and miles demanded by electric vehicles. If you observe, even in the last 3 years, demand is greater than the supply. Then how is the energy demand is being met? Probably because hybrid vehicles switch to fuel resources when they are not bale to get charging stations. From here, we can say that there is still work that is needed to meet the demand created by electric vehicles. Policy makers, Manufacturers and Business men can leverage these metrics statewise and county wise to cut the underutilized areas and balance by making new stations at overutilized areas.

V. Tools and Technologies

- **Google Cloud Platform** GCP provides huge range of cloud computing solutions which includes storage, computing, data cleaning tools and data analytics tools. We used various tools provided by GCP for various purposes but mainly we used GCP as a Data Warehouse. After extracting data from various sources we imported this data into buckets. Buckets were categorized based on the type of data eg. vehicle registration data, energy consumption data, charging infrastructure d,ata, and vehicle model and manufacturing data. Since GCP provides free tier tools upto only certain extent we couldn't process



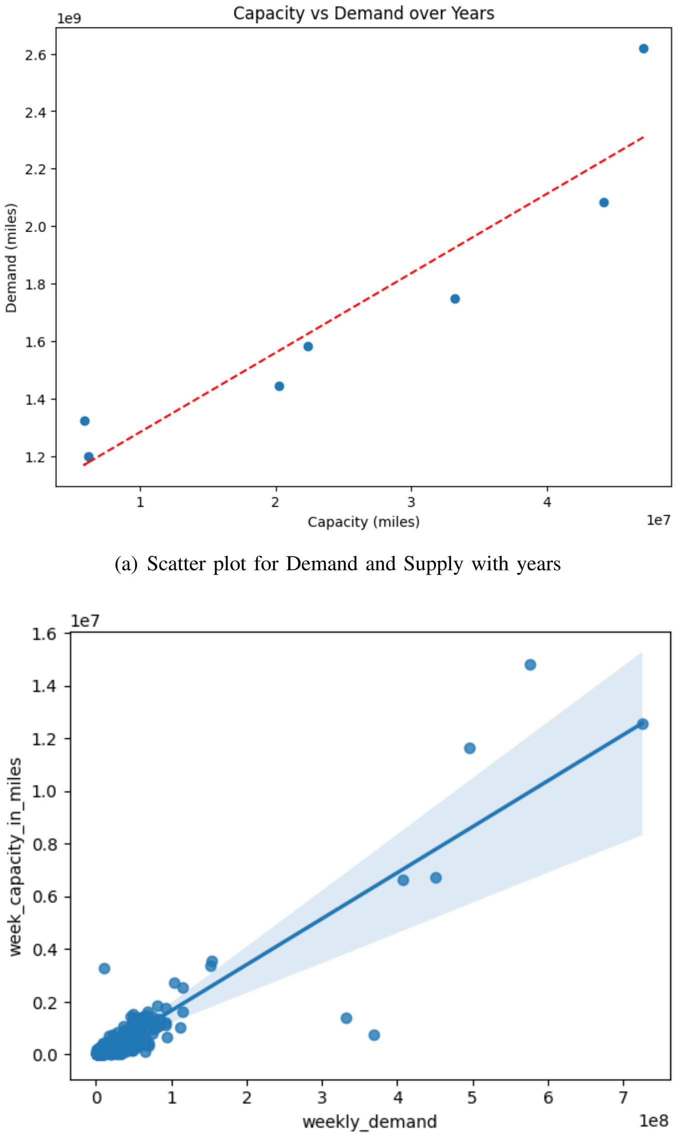
(a) Vehicle loads on charging stations over the years for top states

(b) Overall Statewise vehicle load on charging stations

Fig. 5. Variation of Vehicle Loads on Charging Stations

large volumes of data and sometimes we had to migrate data from one account to another account.

- **Excel** Excel not only helped us understand our data but it also helped us for extracting data from various sources like PDFs or normal text files. We used the GetData feature of Excel to extract relevant data from PDF files and other file formats. The limitation with Excel was that we faced many challenges opening Excel files if the size of file was huge.
- **Google Trifacta Data** This tool helped us for data preparation. Trifacta provides GUI tools for exploring, cleaning, and transforming data which omits the need for depending on programming languages such as Python for data preparation purposes. It provides a wide range of pre-built recipes for data cleaning and allows us to write out our own custom recipes. For cleaning the data some of the recipes we used are as follows - We used the UNION recipe to combine multiple sheets on a unified time dimension. For certain mismatches and inconsistent



(a) Scatter plot for Demand and Supply with years

(b) Regression plot of Demand and Supply over the years and over states

Fig. 6. Comparison of Energy Demand and Energy Supply

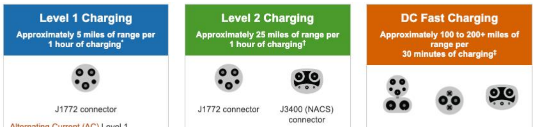

Level 1 Charging
Approximately 5 miles of range per 1 hour of charging*
J1772 connector
Alternating Current (AC) Level 1
Level 2 Charging
Approximately 25 miles of range per 1 hour of charging†
J1772 connector
J3400 (NACS) connector
DC Fast Charging
Approximately 100 to 200+ miles of range per 30 minutes of charging‡

Fig. 7. Charger Levels and their capacities

values, we replace text or units. For a few data columns, we converted data types into appropriate data types. We removed some columns from the original data as these data points were not relevant to our analysis. Handled null values by removing a few rows or by adding some default values.

- **Python** Python is widely used for extraction and data processing. In our project, we used Python for data cleaning, extraction and visualizations. We used libraries such as Pandas, NumPy, seaborn, etc. We used Python mainly for web scraping. We needed data related to vehicle models available in the market and their respective energy consumption statistics. This data was not available in any of the ready-to-use datasets hence we had to scrape the data using Python. For web scraping, we used the BeautifulSoup library. Apart from web scraping, we used Python to connect with BigQuery and run queries from Python to get data to create visualizations. Also, we used Python to add data from Excel and other sources to graph databases KUZU and MySQL.
- **Tableau** We used Tableau to provide interactive visualizations/dashboards. These dashboards allow users to explore the data visually, uncover patterns, and understand how various factors of electric vehicle infrastructures are distributed. Using Tableau we connected to BigQuery and extracted the data. The main challenge we faced was connecting the data to use in the same Tableau book. Since most of our data was not directly related to establishing a connection among it (foreign key) hence, for some visualizations we had to rely on Python for data frame manipulations. We built two dashboards, one explaining how a vehicle registration and models changed over the period and another one explaining how charging infrastructure is changing over time. In addition to doing analysis over time dimension, we also did analysis on space/location. We explored how various states in USA adopted electric vehicles and whether these states have adequate resources to support this adoption.
- **Jupyter Notebook** Jupyter Notebook allowed us to write code for data extraction, cleaning, and visualization.
- **APIs and Drivers** The data we scraped from websites is stored in MySQL database. To connect to MySQL Database we used mysql.connector driver. This driver helped us create a table and insert values directly in the database from Python code. For connecting with BigQuery we used google.cloud.bigquery driver. This driver helped us establish a connection between Python and BigQuery and allowed us to query BigQuery.
- **MySQL** MySQL is widely used where relational databases are required. We used MySQL to store web scraped data. This data is then uploaded on the common storage of BigQuery.
- **Kuzu** Kuzu is an embedded graph database technology. We used Kuzu to explore charging station data. Charging station data was available in JSON format. Kuzu can handle huge amount of data compared to Neo4j
- **Neo4j** We initialy used Neo4j for storing and queriying using Neo4j but we encountered a problem due to huge data size. Neo4j has limit of 400K relation and 250 K nodes. More than these many nodes and relationships we can not create or add.
- **Git** We used git for version control where we uploaded all our source code.
- **Jira** We used JIRA by Atlassian for maintaining task list and dividing tasks among team members.

A. Cost Analysis

Existing analytics companies track data more precisely. Datarade provides realtime charging station data. These insights will be helpful to dig deeper into the analyses. Free resources of Google Cloud Platform were used. We had to cut some of the data on Neo4j while doing analysis because there was support only for 400k relationships. Cost would be at two points - Data procurement and Data Storage Time is a cost. Given time to pursue it as a Term project, there are many more facets which can be digged into for deeper insights.

B. Recommendations

- Certain areas of states that have high Electric vehicles still do not have proper charging infrastructure. There is a huge gap in demand of charging station and availability of charging stations which needs to be met.
- Remote areas and many states still do not have adequate charging stations, which must be addressed.
- Allocate resources towards sustainable transport solutions based on data-driven insights, promoting EV adoption and infrastructure development.
- Foster collaboration among stakeholders to expand charging infrastructure, addressing gaps and promoting widespread adoption of electric mobility.
- Underutilized charging stations can be made available for public usage.
- New openings for public charging stations are depleted after 2023 which needs to be addressed.

VI. Conclusion

In conclusion, our project supports the development of sustainable transport solutions by offering vital data insights to stakeholders in the electric vehicle sector. It helps businesses align with environmental goals by addressing issues including EV adoption rates, the requirement for charging infrastructure, and environmental implications. By making these kinds of efforts, the industry can support a robust and environmentally responsible transportation ecosystem, paving the way for a cleaner future and a strong future.

VII. Progress

A. Task Distribution

- Data collection - Gathering data and finding various data sources -
    - Vehicle Models data from 2016 - 2024 for all categories - Mrunali Katta
    - Vehicle Registration data from 2016-2024 - Kanchan Naik
    - Existing fuel station data from 2014-2024 - Yashasvi Kanchugantla
    - Energy Demand Estimation by Vehicle - Prasad Simpatvar
- Data Cleaning and Transformation -
    - Vehicle Models data from 2016 - 2024 for all categories - Mrunali Katta
    - Vehicle Registration data from 2016-2024 - Kanchan Naik
    - Existing fuel station data from 2014-2024 - Yashasvi Kanchugantla
    - Energy Demand Estimation by Vehicle - Prasad Simpatvar
- Data modeling - Prasad Simpatwar
- Data warehouse design and implementation - Mrunali Katta, Kanchan Naik, Yashasvi Kanchugantla
- Data visualization

    Growth of Electric Vehicles over time - Mrunali Katta Analysis of existing charging infrastructure - Prasad Pramod Shimpatwar Analysis of energy consumption by Electric Vehicles - Kanchan Ashok Naik Understanding if the current infrastructure can support the growing need for electric vehicles - Yashasvi Kanchugantla

  Project Report and slides - Mrunali Katta Kanchan Ashok Naik Prasad Pramod Shimpatwar Yashasvi Kanchugantla

VIII. Acknowledgment

We would like to express our gratitude to all those who helped us in the successful completion of this project titled 'Analysis of Electric Vehicle Infrastructure'. Our thanks go to: Our professor Dr. Vishnu Pendyala, for his guidance and support throughout the project and semester. His lectures and guidance have been instrumental in the completion of this project.

Our classmates and team members, for their dedication and efforts. Each member's unique skills and contributions played a crucial role in achieving our project goals. Our TAs for this course whose guidance and assistance in clearing our doubts and questions helped us complete this project. Lastly, we acknowledge all individuals, organizations, and sources referenced in our project work. Their insights, data sources, and available resources helped us strengthen our project outcomes.

IX. Appendix A

A-1: Presentation Skills.Includes time management:
"To be done at the time of presentation"

A-2: Code Walkthrough:
"To be done at the time of Presentation"

A-3: Discussion / QA:
"To be done at the time of Presentation"

A-4: Demo:
"To be done at the time of Presentation"

A-5:Version Control Use of Git / GitHub or equivalent; must be publicly accessible:
Link(Google Drive): [9]

Link(Github): [10]

A-6:Significance to the real world:

"Significance found for Policymakers and contractors to bring up more charging stations in overutilized areas and areas which have growing EV registrations. Can be very helpful for planning for infrastructure in states. This data helps fleet managers make informed decisions when transitioning to EVs."

A-7: Lessons learned. Included in the report and presentation? How substantial and unique are they? "We could explore many database technologies. We understood the importance of data cleaning and its effect on the productivity of developing algorithms and also the accuracy of the results"

A-8: Innovation: "We explored a unique embedded Graph Database, Kuzu. Also came with some metrics to make meaningful information from the data we have."

A-9: Teamwork We collaborated to find various data sources and cleaned data using gmeet. We connected over a meeting every week to work on various topics like data collection, cleaning, and processing. We connected to derive formulas for calculations.

A-10: Technical difficulty: Data Integration: Managing and merging data from diverse sources and datasets for charging stations, energy consumption, etc, and challenging due to variations in data formats, standards, and accessibility. EV charging station data is in JSON format, it will viewed and processed in NoSQL format. Other datasets are tabular, hence it is easy to use RDBMS. Integrating all these datasets and modeling data is another Data Quality Assurance: Ensuring the precision and reliability of collected data, encompassing details on charging station locations, availability, power ratings, and usage patterns, is essential for informed analysis and decision-making. Handling missing values/null values

A-11: Practiced pair programming? "Yes We have practiced pair programming : 1) At the time of ETL half sheets were done by one person and then paired and explained and done by other 2)Regular meetings were done and discussed. 3) Planning and organizing were done combined. " A-12: Practiced agile / scrum (1-week sprints)? Submit evidence on Canvas - meeting minutes, other artifacts

A-13:Used Grammarly / other tools for language? Grammarly free version is sufficient; can use other tools as well. Submit a report screenshot on Canvas.

A-14: Slides
"To be presented at the time of Presentation"

A-15: Report Format, completeness, language, plagiarism, whether turnItIn could process it (no unnecessary screenshots), etc:// Yes we have taken care of the screenshots which are required and necessary are only uploaded.

A-16: Used unique tools
E.g.: LaTeX for writing a report (submit .tex that is not generated from another format such as .docx; generating from .lyx and similar LaTeX editor outputs is fine. Checkout [11]

A-17: Performed substantial analysis using database techniques The project must include an analytics component://Yes we have provided visualization which can be used to do analysis.

A-18: Used a new database or data warehouse tool not covered in the HW or class://Worked on Google data prep where 'ETL can be performed. Use of Bigquery and using it to connect in tableau and create Visualization. We have also used "KuzuDB" and also used "NEO4j""

A-19: Used appropriate data modeling techniques://Yes

A-20: Used ETL tool//Yes We have used. Worked on Google data prep where 'ETL can be performed We used DataPrep by trifact google

A-21: Demonstrated how Analytics support business decisions:// "We have given the analysis how stakeholders can make decisions using the "Electric Vehicle analysis" and

A-22: Used RDBMS:// Web pages were scraped to fetch some useful information and written to MySQL

A-23: Used Datawarehouse We used BigQuery, Neo4j and Kuzu. Datamarts were the datasets available and we assimilated on these tools

A-24: Includes DB Connectivity / API calls Possibly using Python - Yes, for interactions with BigQuery, Neo4j, MySQL. we used mysql.connector

A-25: Used NOSQL: Used Graph database -Neo4j, Kuzu for JSON data

References

[1] "For a livable climate: Net-zero commitments must be backed by credible action." https://www.un.org/en/climatechange/net-zero-coalition, 2022.

[2] "Vehiche models." https://ev-database.org/.

[3] "Vehicle registration counts by state." https://afdc.energy.gov/vehicle-registration, 2022.

[4] "Nyserda." An Electric-Vehicle Consumer Segmentation Roadmap: Strategically Amplifying Participation in the New York Drive Clean Rebate Program, 2024.

[5] "Poi." https://github.com/openchargemap/ocm-data, 2022.

[6] "An associated dataset is also found with databook." https://tedb.ornl.gov/data/, updated 2022.

[7] "Electric vehicle charging stations." https://afdc.energy.gov/fuels/electricity-stations, 2022.

[8] "The average annual miles driven in each state." https://www.agilerates.com/advice/auto/average-miles-driven-per-year/, 2022.

[9] "Drive link." https://drive.google.com/drive/folders/1-sL4i0od-QU9siBqDCCfo9_elJAezRl7, 2024.

[10] "Github link for the repository." https://github.com/Yashasvi1225/electric_vehicles_analytics, 2024.

[11] "Overleaf link." https://www.overleaf.com/read/vqpwhhnwbrgv#e0c667, 2024.