# Longitudinal Analysis on Alternative Fueling Stations Trends and Anomalies

Christopher Ton

Department of Applied Data Science, San Jose State University

DATA 230: Data Visualization

Andrew Bond

December 6, 2023

## Abstract

As society makes incremental steps towards a cleaner and more sustainable future, the adoption of electric vehicles and alternative fueling resources are two such main considerations for achieving the ambitious initiative. To better anticipate the future landscape of EV momentum and the charging resources required to support it, a longitudinal analysis of past and existing trends or patterns in the adoption and growth of alternative charging stations is conducted. Important attributes to reasonably pinpoint potential areas for growth include capturing the momentum of EV expansion in the economic and consumer market or accessibility to resources for manufacturing, network placements and real estate. Holistic and segmented analyses were performed to extract insights into alternative fueling types and characteristics over a year to year basis. For interactivity purposes, a web dashboard is implemented to allow intended audiences to filter for specific periods of interest, assess the distribution of charging naming conventions, geographical presence of historical station placement types, and relationships among inherent attributes regarding operating hours, payment plans and utility. Finally, an unsupervised learning approach with Isolation Forest is implemented to classify historical anomalous points in the increase or decrease of station growth over time. Detection of abnormal growth trends can yield insights to uncover underlying factors contributing to the strongest indicators for expanding the EV network.

*Keywords: alternative, longitudinal, market, network, historical, unsupervised*

Comparatively, Consumer behavior is shifting and the International Energy Agency reports that “ Demand for electric cars is booming, with sales expected to leap 35% this year after a record-breaking 2022.” From a policy standpoint and with respect to the state of California for instance, all new cars that are sold from 2035 onwards must hold zero emissions status, including battery, PHEV, and fuel cell EV, according to the CA Air Resources Board. The demand in EV adoption rate must be met with a sufficient infrastructure of charging resources to supply its sustainability. If market needs are not properly addressed then constructing an already-overpopulated category of a particular charger type may likely lead to underutilization. Contrastingly, supplying too few resources would result in a shortage and inability to keep up with consumer needs. Exploratory analyses serve to provide a baseline expectation for the nationwide inventory of charging resources by utility and accessibility types. Business initiatives can be derived from informed decision-making based on the visual representations of longitudinal data.

### *1.2 Project Requirements and Deliverables*

The term project deliverables includes a fully functional and interactive dashboard based on geographical data that integrates at least 5 distinct visualizations including bar, line, scatter, geographical and at minimum, 2 variations of each type. A written report detailing the design, build approach and explanations for the insights generated from observational data are delivered alongside the dashboard.

### *1.3 Technology and Solution Survey*

Many different dashboarding tools exist for free and commercial use. The rationale for choosing to produce the delivered dashboard using Python programming language and web development frameworks such as Dash and Streamlit was based on the preference to straightforwardly define respective visual components and filters as needed. Streamlit is a popular package for deploying machine learning models and visualizing the outcomes and summary statistics associated with the models. Although Dash provides a greater range of customization options, Streamlit requires relatively less code to initialize user input functionalities and integrations with linking functions for chart building and predictive analytics outputs.

## 2. Data and Project Management Plan

### *2.1 Data Management Plan*

For mainly visualization purposes, the raw data files were stored locally in a personal harddrive for processing, transformations and analyses. A local directory was created specifically for the course and datasets were loaded into respective folders for efficient access.

### *2.2 Project Development Methodology and Plan*

The project follows a pre-defined schedule of deliverables including proposal submission, use case pitch and final report delivery. Initial research and exploration was the first step in defining the problem and business domain. Personal bias and expectations largely drove web queries for information regarding data and news sources related to the domain. Once a credible and a reliable data source was determined, a project hypothesis was formulated based on existing  and newfound impressions gained from research. Such platform sources include Alternative Fuels Data Center, Global EV Policy Explorer, and other mainstream media sources. Once sufficient information was gathered for the problem domain, the next step in the project development process involved crafting the use case pitch to the class. A proposal was drafted to address the intended objective, background context, tools and technologies utilized, course of action and planned deliverables. The condition for advancement was initial approval from the instructor to pursue the exploratory analysis. From that point onwards, project development took place within a preferred IDE and jupyter notebook for EDA. Upon completion of sufficient initial analyses, planning for the deployment phase of the project via an interactive dashboard was the next step. For quick prototyping, Tableau was explored as a medium for generating quick visuals in a no-code or low-code manner. However, it was quickly decided that frameworks such as Streamlit or Dash were better suited for the design and delivery of custom filters and/or functions to showcase intended visuals. Finally, with a dashboard constructed, a report to detail and document each chart deliverable was completed. All codes and notebooks are uploaded into GitHub, referenced with https://github.com/chriztopherton/data230_fall2023.

### *2.5 Project Schedule*

## 3. Data Engineering

### *3.1 Data Process and Collection*

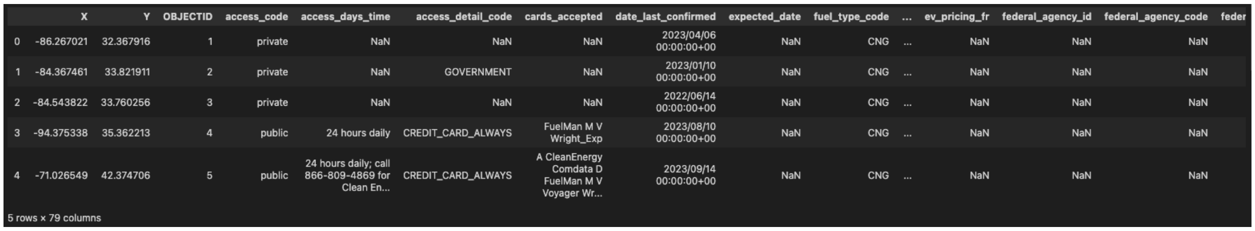
The data process and collection is straightforward, in that once the raw csv file was extracted and loaded into local storage, it was ready for processing. The Alternative Fuel Data center repository site ensures data is accessible through its platform and provides documentation for background collection information, a data dictionary, and a specialized dashboard for bespoke queries. For a holistic analysis, all 79,342 records covering all stations locations and types across the United States and Canada were exported for consideration. Columns that were not intended for use are dropped during the reading step (refer to lines 20-21 in app.py file).

### *3.3 Data Pre-processing*

Reading in the data was performed with the python library Pandas, a popular module for data wrangling and manipulation of various text file formats. By default, the module can recognize data types which should be of integer or float types but conventionally parses datetime objects as string. Converting date fields to their appropriate type using the function "pandas.to_datetime()" was the only pre-processing step required during the ingestion of stored datasets. Figure X shows the snapshot of the raw dataset that was read in before any pre-processing steps were performed.

**Figure X**

*Sample of raw dataset with 79 columns*



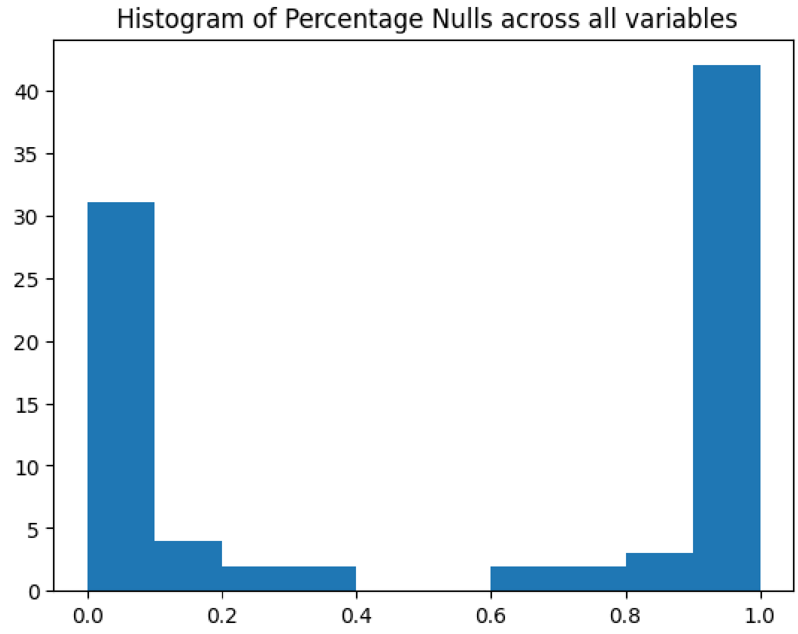
| | X | Y | OBJECTID | access_code | access_days_time | access_detail_code | cards_accepted | date_last_confirmed | expected_date | fuel_type_code | ... | ev_pricing_fr | federal_agency_id | federal_agency_code | feder |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -86.267021 | 32.367916 | 1 | private | NaN | NaN | NaN | 2023/04/06 00:00:00+00 | NaN | CNG | ... | NaN | NaN | NaN | |
| 1 | -84.367461 | 33.821911 | 2 | private | NaN | GOVERNMENT | NaN | 2023/01/10 00:00:00+00 | NaN | CNG | ... | NaN | NaN | NaN | |
| 2 | -84.543822 | 33.760256 | 3 | private | NaN | NaN | NaN | 2022/06/14 00:00:00+00 | NaN | CNG | ... | NaN | NaN | NaN | |
| 3 | -94.375338 | 35.362213 | 4 | public | 24 hours daily | CREDIT_CARD_ALWAYS | FuelMan M V Wright_Exp | 2023/08/10 00:00:00+00 | NaN | CNG | ... | NaN | NaN | NaN | |
| 4 | -71.026549 | 42.374706 | 5 | public | 24 hours daily; call 866-809-4869 for Clean En... | CREDIT_CARD_ALWAYS | A CleanEnergy Comdata D FuelMan M V Voyager Wr... | 2023/09/14 00:00:00+00 | NaN | CNG | ... | NaN | NaN | NaN | |

5 rows × 79 columns

Regarding data quality, many of the columns have entirely or majority null values. Given the lack of content provided by these scarce features, it may be reasonable to drop them from analysis.  Figure X shows the distribution of null values by proportion across all descriptive features. The bimodal distribution provides reasoning to drop the scarce fields from consideration.

**Figure X**

*Distribution of null values by feature*

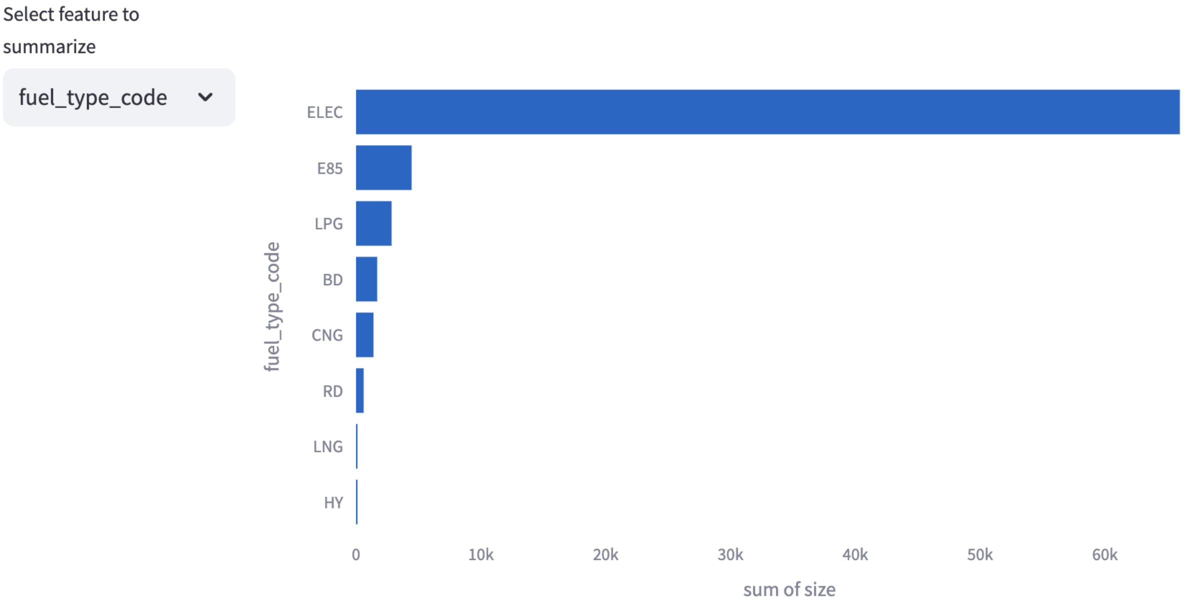


### *3.4 Data Transformation and Preparation*

To analyze the data on a categorical level by factors, a stratification approach was used with pandas' " groupby" function to partition by selecting variables of low cardinality. For example, by stratifying on the fuel type and measuring the size of each group, a distribution of the group sizes are obtained. Visually, figure X shows the distribution of counts of stations by fuel type code. The "ELEC" or electric charging group is outstandingly the largest group followed by "E85" or Ethanol85 then "LPG"  or propane. Understanding this distribution provides a basis for whether the counts by type are skewed or uniformly distributed. Alternatively,  a pie chart could present the proportions by total count size of each fuel type group. Figure X shows the pie chart and proportions for each factor level, "ELEC" shown in blue. By changing the input filter to "status_code" or " access_code", among other categorical variables, the distributions can be visualized for whether stations are operational, or public/private, respectively. Lines 42 - 49 in app.py shows an example of the function definition for the "groupby" statement based on the input data to find the median coordinate by city factor, and count the number of station instances. What materializes the display of the pie chart and overall geographical map are lines 140 to 149 in app.py, to format the charts in column format.

**Figure X**

*Bar chart for variables of relatively low cardinality*

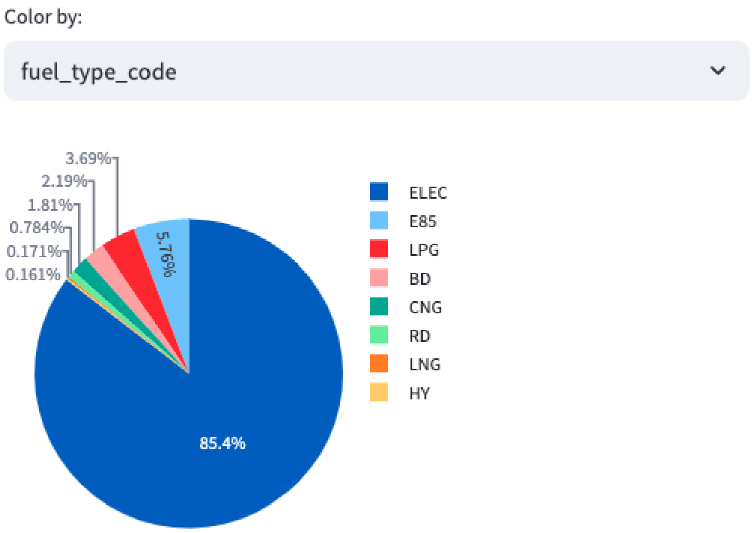


Select feature to summarize

fuel_type_code

**Figure X**

*Pie chart for fuel type attribute*



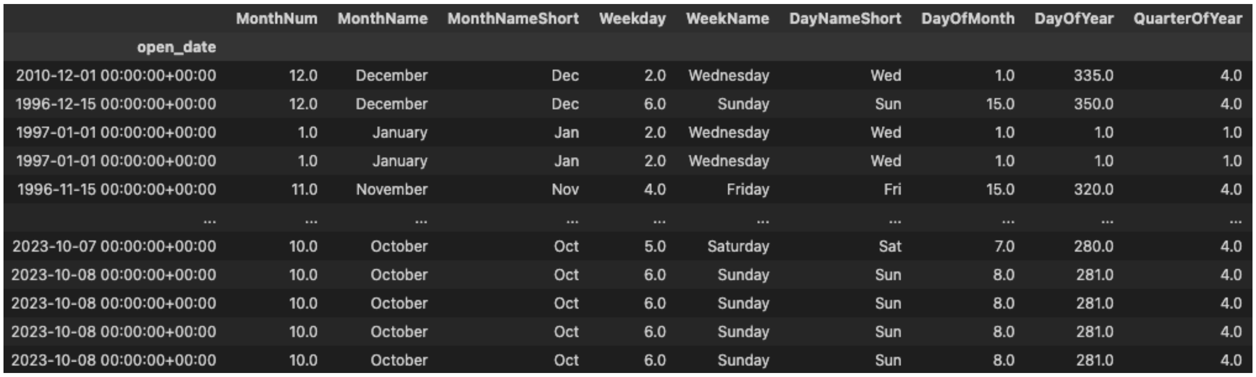
Color by:

fuel_type_code

From the pre-processing step that transforms date type fields into their intended types, the components such as weekday, months, or quarters can be extracted. The motivation for this step allows for stratification by date objects at a higher granularity level. Figure X displays a set of date components extracted from the “open_date” field during the .dt accessor from the pandas library. Insights or patterns that relate to specific months, days of the week, or quarters of the year can be derived from these new feature engineered variables. Lines 22-24 in app.py shows the datetime transformations and extraction of date components.

**Figure X**

*Data component attributes extracted from opening date variable*

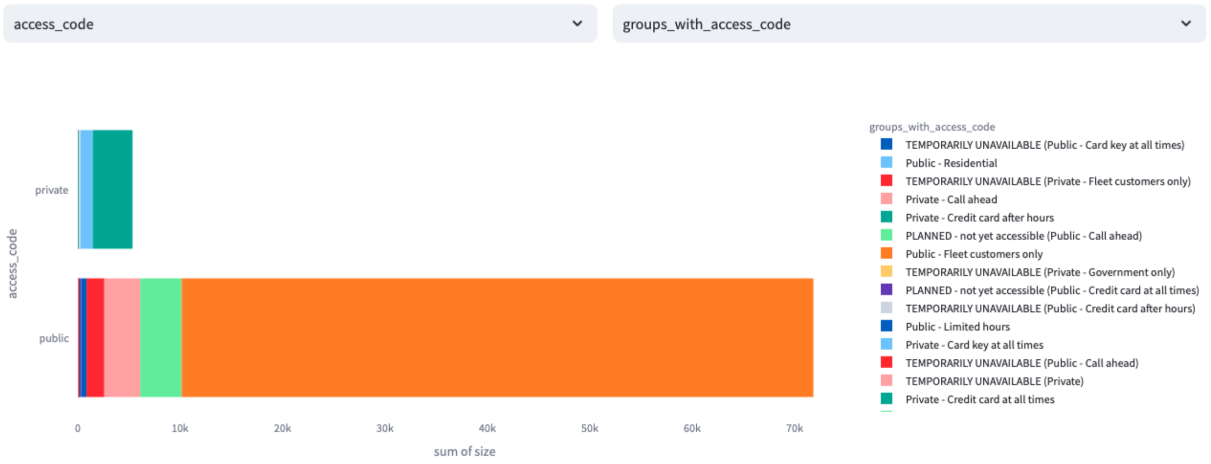


| | MonthNum | MonthName | MonthNameShort | Weekday | WeekName | DayNameShort | DayOfMonth | DayOfYear | QuarterOfYear |
|---|---|---|---|---|---|---|---|---|---|
| open_date | | | | | | | | | |
| 2010-12-01 00:00:00+00:00 | 12.0 | December | Dec | 2.0 | Wednesday | Wed | 1.0 | 335.0 | 4.0 |
| 1996-12-15 00:00:00+00:00 | 12.0 | December | Dec | 6.0 | Sunday | Sun | 15.0 | 350.0 | 4.0 |
| 1997-01-01 00:00:00+00:00 | 1.0 | January | Jan | 2.0 | Wednesday | Wed | 1.0 | 1.0 | 1.0 |
| 1997-01-01 00:00:00+00:00 | 1.0 | January | Jan | 2.0 | Wednesday | Wed | 1.0 | 1.0 | 1.0 |
| 1996-11-15 00:00:00+00:00 | 11.0 | November | Nov | 4.0 | Friday | Fri | 15.0 | 320.0 | 4.0 |
| ... | ... | ... | ... | ... | ... | ... | ... | ... | ... |
| 2023-10-07 00:00:00+00:00 | 10.0 | October | Oct | 5.0 | Saturday | Sat | 7.0 | 280.0 | 4.0 |
| 2023-10-08 00:00:00+00:00 | 10.0 | October | Oct | 6.0 | Sunday | Sun | 8.0 | 281.0 | 4.0 |
| 2023-10-08 00:00:00+00:00 | 10.0 | October | Oct | 6.0 | Sunday | Sun | 8.0 | 281.0 | 4.0 |
| 2023-10-08 00:00:00+00:00 | 10.0 | October | Oct | 6.0 | Sunday | Sun | 8.0 | 281.0 | 4.0 |
| 2023-10-08 00:00:00+00:00 | 10.0 | October | Oct | 6.0 | Sunday | Sun | 8.0 | 281.0 | 4.0 |

Data transformation of the textual fields such as “access_code” and “groups_with_access_code” can be performed to separate the nuances in the detailed specifications with respect to the private and public subgroups. Plotting the access code as the first level grouping and the detailed codes as the second-level grouping, it is observed that while public stations are generally accessible, a few minor proportions of them require credit card at all times, colored as pink/salmon, or advanced notice by call, colored as red (see Figure X). Single variable and grouped variable bar charts are defined in lines 197 to 222 in app.py.

**Figure X**

*Drilling down into usage specifications for public and private access stations*

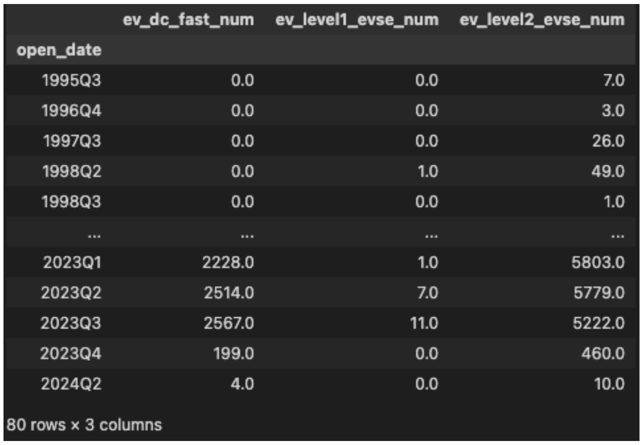

access_code
groups_with_access_code
private
public
access_code
0
10k
20k
30k
40k
50k
60k
70k
sum of size
groups_with_access_code
TEMPORARILY UNAVAILABLE (Public - Card key at all times)
Public - Residential
TEMPORARILY UNAVAILABLE (Private - Fleet customers only)
Private - Call ahead
Private - Credit card after hours
PLANNED - not yet accessible (Public - Call ahead)
Public - Fleet customers only
TEMPORARILY UNAVAILABLE (Private - Government only)
PLANNED - not yet accessible (Public - Credit card at all times)
TEMPORARILY UNAVAILABLE (Public - Credit card after hours)
Public - Limited hours
Private - Card key at all times
TEMPORARILY UNAVAILABLE (Public - Call ahead)
TEMPORARILY UNAVAILABLE (Private)
Private - Credit card at all times

### *3.5 Data Preparation*

Besides aforementioned transformations to visualize stratified variables, an additional step to prepare the training dataset for unsupervised classification was performed. The scope of the predictive modeling aspect was the subset of electric charging stations and the growth of level 1, 2 and dc fast chargers. To model growth over time, it meant structuring the data into a time series format and indexing on the "open_date" variable as the sequential date-based feature. To aggregate the data and reduce dimensionality, downsampling was applied to sum the number of opening stations per charger type on a quarterly basis. Figure X shows a snapshot of the prepared dataset after indexing on the opening date and selecting the three continuous descriptive features. From the third quarter of 1995 until the second quarter of 2024, there are 80 aggregated records. More visually, Figure X shows the corresponding time series graph and 3 individual series per charger type in different colors. The line plots are constructed in lines 228 to 239 in app.py, showing the yearly growth by states then station growth rates by type, respectively.

**Figure X**

*Prepared training set for electric stations including charger types*

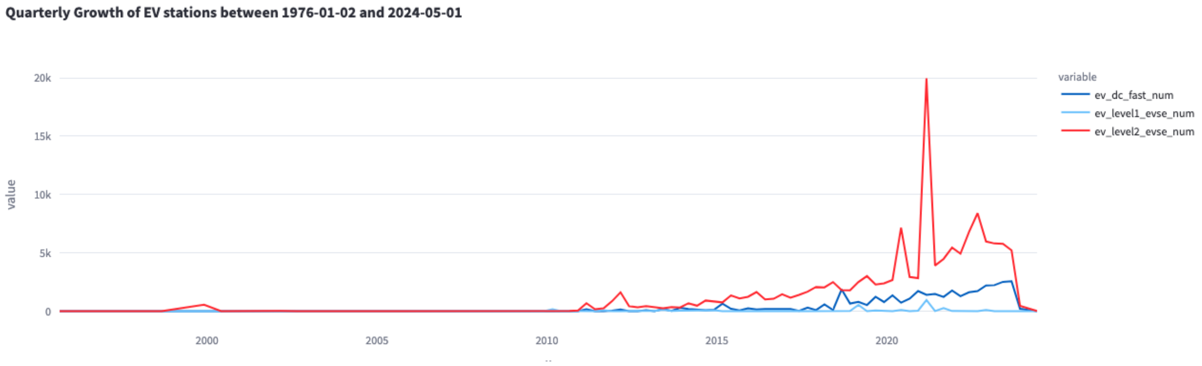


| open_date | ev_dc_fast_num | ev_level1_evse_num | ev_level2_evse_num |
|---|---|---|---|
| 1995Q3 | 0.0 | 0.0 | 7.0 |
| 1996Q4 | 0.0 | 0.0 | 3.0 |
| 1997Q3 | 0.0 | 0.0 | 26.0 |
| 1998Q2 | 0.0 | 1.0 | 49.0 |
| 1998Q3 | 0.0 | 0.0 | 1.0 |
| ... | ... | ... | ... |
| 2023Q1 | 2228.0 | 1.0 | 5803.0 |
| 2023Q2 | 2514.0 | 7.0 | 5779.0 |
| 2023Q3 | 2567.0 | 11.0 | 5222.0 |
| 2023Q4 | 199.0 | 0.0 | 460.0 |
| 2024Q2 | 4.0 | 0.0 | 10.0 |

80 rows × 3 columns

**Figure X**

*Time series plot for Electric charger types*

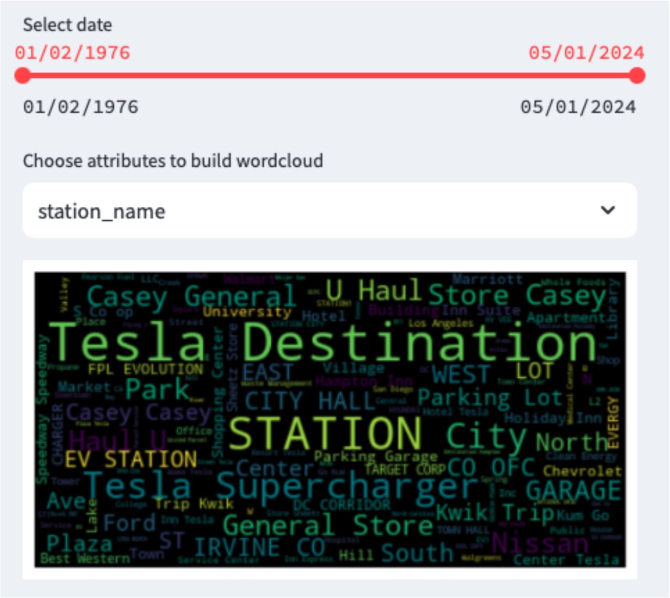


### *3.6 Data Statistics and Analytics Results*

The dashboard is a product of various visualizations ranging from bar and pie charts to geographical representations, scatter plots and line plots that can all react to date range inputs (see lines 112-117 in app.py for the initialization of the date range input). Analytical conclusions can be derived from visual observation of the graphs and different insights can be generated with the interactive filters for the entire dataset. For instance, a top-down approach can be taken by first plotting a word cloud of the most frequent texts given a time period (see lines 125 - 132 in app.py for the code that creates the input filter for textual columns for the word cloud). A use case could be the distribution of stations names across the entire longitudinal period from 1995 until 2024. A result of this is provided in figure X. The most prominent words that appear indicated by central placement tendencies and size, are Tesla superchargers, shopping centers or governmental sites such as city hall. These results are expected given that Tesla is leading in the EV charging provider industry (EV Magazine, 2023).

**Figure X**

Word map for station names from 1995 to 2024.



Select date

01/02/1976 05/01/2024

01/02/1976 05/01/2024

Choose attributes to build wordcloud

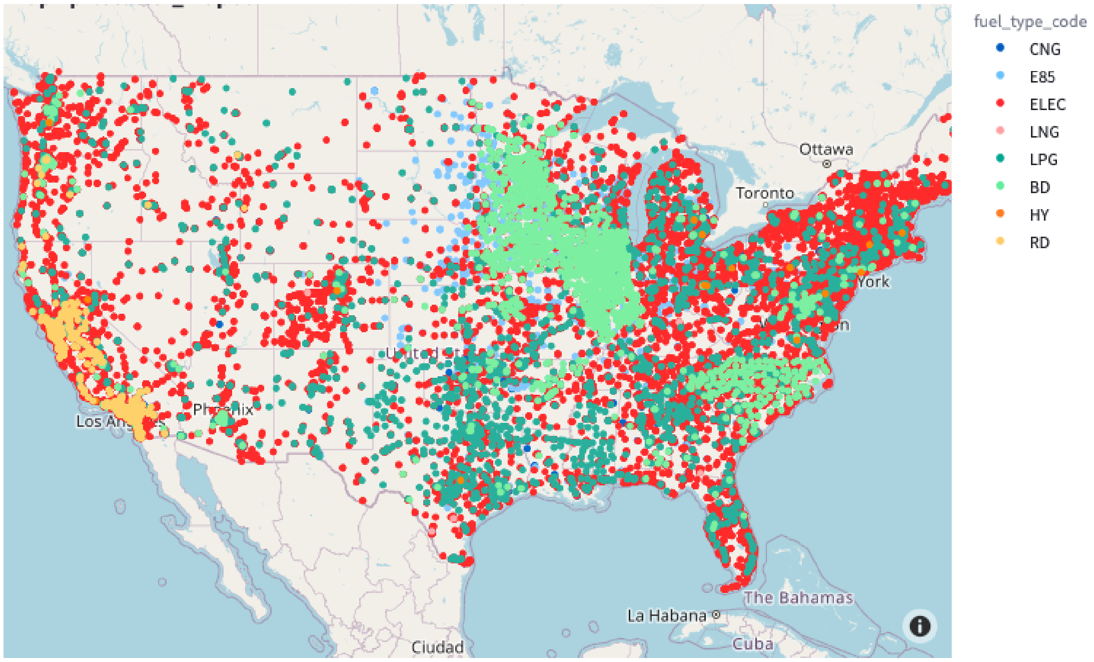
station_name

Casey General U Haul Store Casey
University Hotel
Tesla Destination
FPL EVOLUTION EAST Village WEST LOT
Market Park CITY HALL Parking Lot
Casey Casey Holiday Inn
Haul U STATION City North
EV STATION Center Parking Garage
TARGET CORP CO OFC Chevrolet
Tesla Supercharger GARAGE
Ave Trip Kwik DC CORRIDOR Kwik Trip
Ford General Store
Plaza ST IRVINE CO Hill South Nissan
Best Western Town Center Tesla
Speedway EVERGY Shopping Center Marriott

Geographically, it is also within expectations to hypothesize that superchargers are well distributed across the nation, with at least some presence in each state. Figure X shows all the historical stations plotted by latitude and longitude coordinates and their types colored by fuel type attribute. Lines 53-65 in app.py is the function for plotting the geographical map with inputs being the data to process and column for color definition. A strong concentration of Tesla superchargers, shown as red markers, populate the east and west coasts, with a minority presence in the central or midwest states. Opportunity for future expansion can likely be high given the right constraints for resources, real estate and policy allowances.

**Figure X**

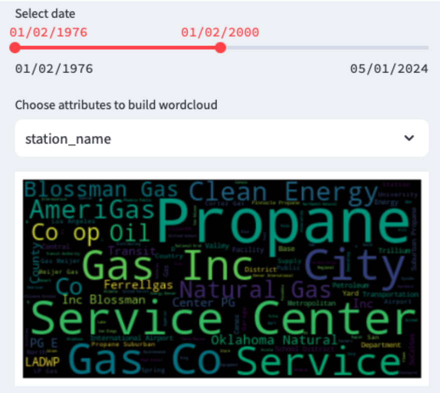
*Geographical map of all stations from 1995 to 2024, colored by fuel type*



However, adjusting the date range to the very first segment of the timeline and cutting all the years post January 2000, there is a drastically different story to be told with the distribution of stations by type and location (see Figure X). The LPG or propane stations dominate by presence across the United States and contrastingly, Electric stations are a minority during this earlier time period, with only a few seen in the state of California.

**Figure X**

*Word map and geographic map for stations between the years 1976 and 2000*



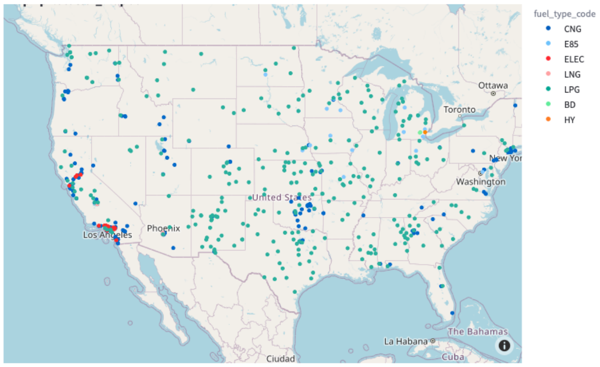
Select date

01/02/1976 01/02/2000

01/02/1976 05/01/2024

Choose attributes to build wordcloud

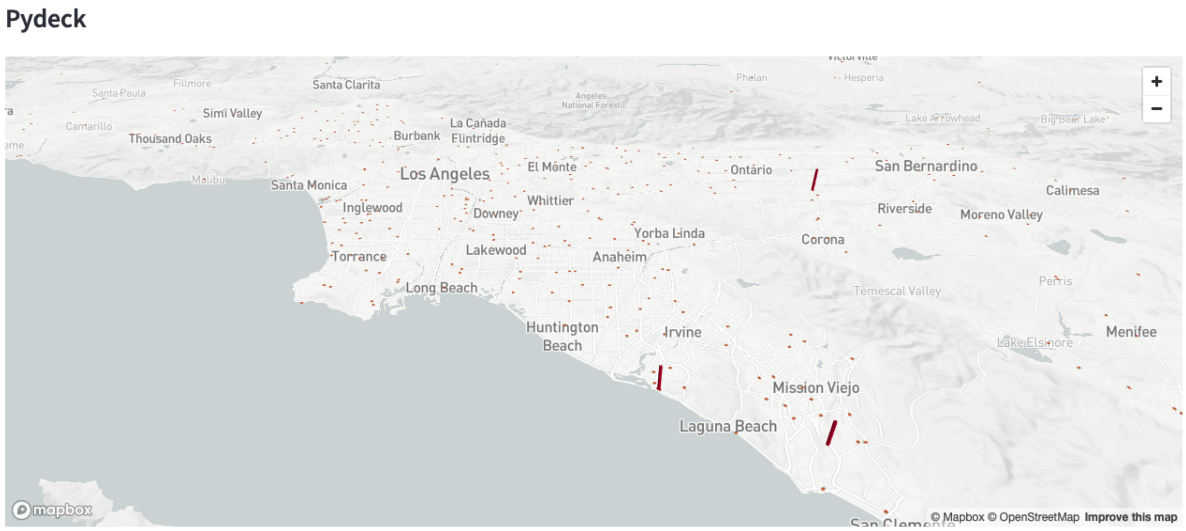
station_name



Geographic charts can be displayed more interestingly by directly plotting bar charts right where the latitude and longitude coordinates are physically. Using the pydeck library, Figure X shows this for the Los Angeles area and the count of stations indicated by the heights of the red bars. By overlaying the two charts and creating a hybrid map, an impression for where the highest frequency of stations are located can be made. The aerial view of the graph provides advantages over the flat representation by giving an accurate representation of the terrain or landscape and better spatial context. Refer to lines 160-192 in app.py for the function that builds the pydeck chart by first transforming the station counts into a standard scale, then specifying the chart parameters.

**Figure X**

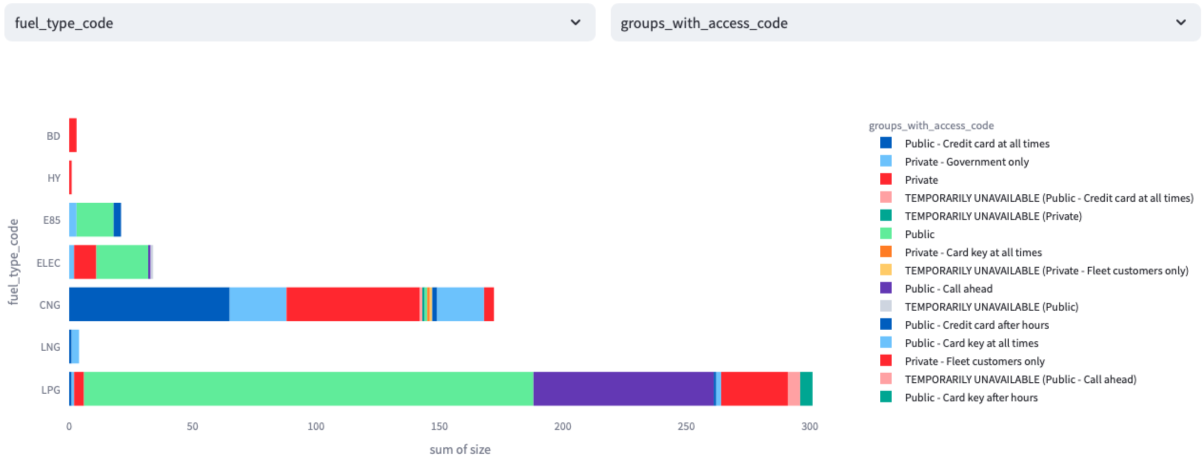
Hybrid visual combining geographic map and bar charts overlaid



The relationships or interactions between the categorical variables are important factors to consider when comparison between groups are studied. For instance, if the network is expanded and the question for where to place new stations arises, one would need to perform research into the existing market and consumer behaviors, for instance. For plotting a grouped bar chart and comparing pairwise categorical attributes such as “fuel_type_code” against “groups_with_access_code”, the actual public accessibility for electric charging stations is uncovered (see Figure X). Until January 2000, It is quite apparent that the LPG group dominated with the greatest offerings of public access codes, while electric offerings are more comparable to that of E85s.

**Figure X**

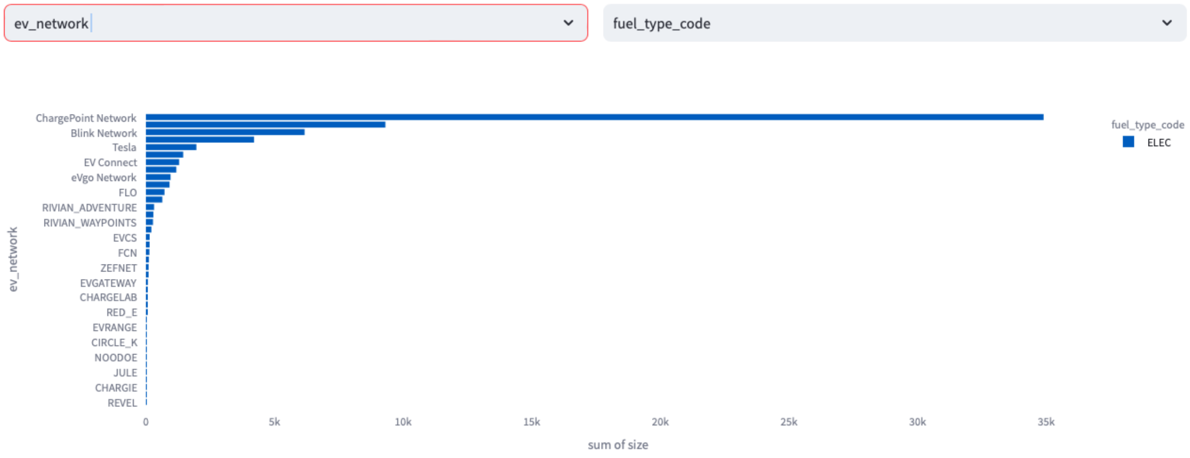
*Charting the fuel type variable against different access codes*



An example of market research is investigating the distribution of existing infrastructure of the top leaders in the industry. Figure X plots the EV charging companies and the count of stations currently operating in the present 2023 year. Top of the list is “ChargePoint” with nearly 35,000 stations followed by “Non-networke”  or independent entities, and Blink.

**Figure X**

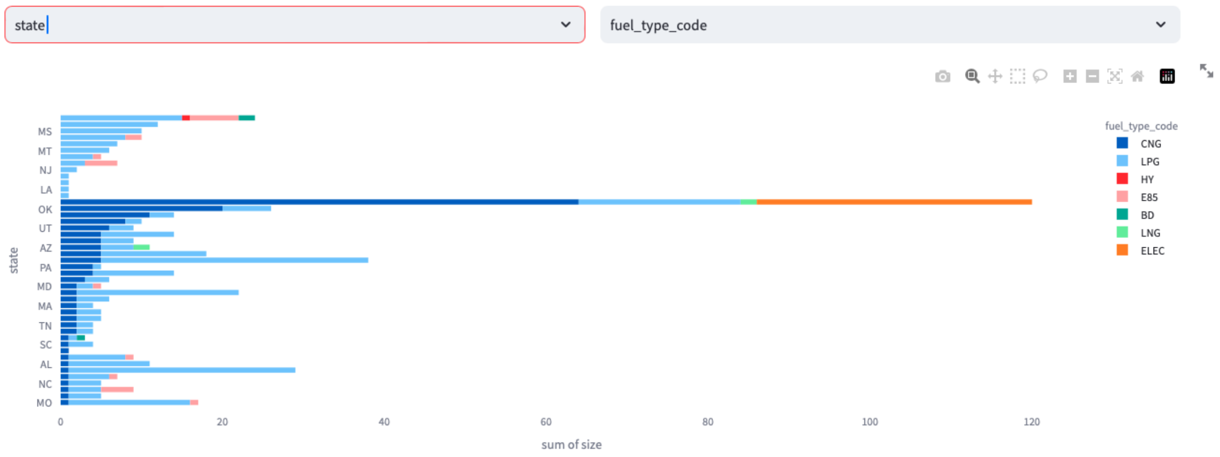
*Distribution of count of stations by company*



To complement the geographic map that depicts the frequency of points in each state, a grouped bar chart that involves state categorizations and fuel type can provide additional insight into the distribution of offerings. It is clear that California leads in the greatest number of stations but room for expansion can be found with the states with little to no orange bars labeled (see Figure X).

**Figure X**

*Plotting state levels against fuel type levels from 1976 until 2000*



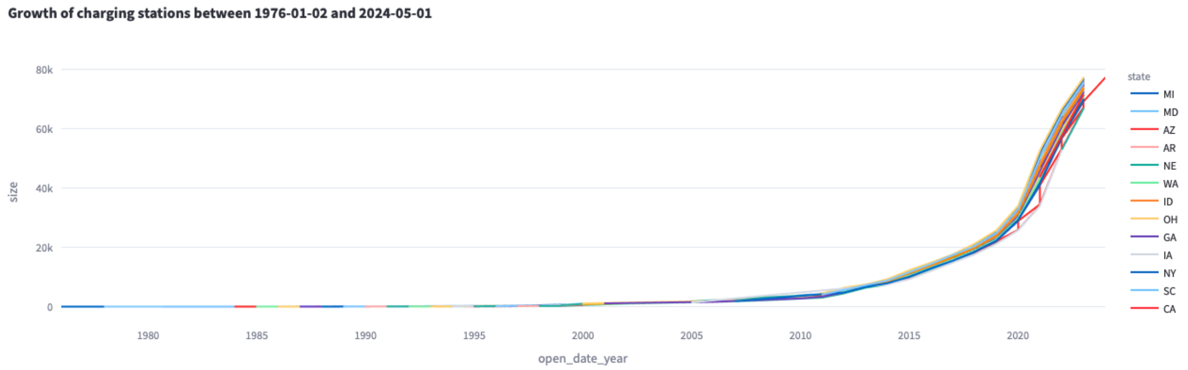
state

fuel_type_code

A year over year analysis for charging station growth rates, stratified by state, can provide insight for comparative adoption patterns. Figure X shows the line plot from the years 1976 until 2024 for the aggregated sum of stations that have opened in each state. Plotly's interactive function can allow users to zoom into specific windows of time to inspect differing rates.

**Figure X**

*Time series plot for charging station growth rates by state*



## 4. Model Development

### *4.1 Model Proposals*

Isolation Forest was selected as the classification model to implement and identify global anomalies from a longitudinal standpoint. To thoroughly assess historical performance and patterns or trends in the growth of charging stations by type, the model is fitted on the entire timeline from the first installed station until present day. Due to isolation forest's advantage for low computation, efficient scaling, and robustness to single outliers, the model is preferable for multivariate classification. Additionally, the model works well for high dimensions where there are potentially a high number of irrelevant features, and in cases where no anomalies actually exist (Liu et al., 2008). For simplicity, a proof of concept model is constructed based on 3 continuous variables being the sum of opened stations at each quarter of the year.

### *4.2 Model Supports*

**Hardware Requirements.** Model development took place locally on personal macbook machines. Given that the dataset was not relatively large, training computation time was almost instantaneous. For reference, all the local development took place on an M3 Macbook Pro with the Apple M3 Pro chip.

**Software Requirements.** Building the model required a development environment that can host python version 3, and software libraries such as pandas, numpy and scikit-learn to provide functions and APIs to execute analytical actions. To properly manage versioning and dependencies among these packages, and minimize version conflicts with external libraries, a virtual environment was necessary to isolate the workspace. Once downloaded and installed, the necessary packages can be utilized for experimentation, exploration and modeling. Refer to lines 1 -11 in app.py for all the packages utilized for the data exploration and dashboard development.

### *4.3 Model Comparison and Justification*

Only a single Isolation Forest model was implemented for the unsupervised anomaly detection objective. As an ensemble method, Isolation Forest combines the outputs from multiple outlier candidates. This helps to ensure that conclusions derived from  independent information-based learners are generated from majority vote.

### *4.4  Model Evaluation Methods*

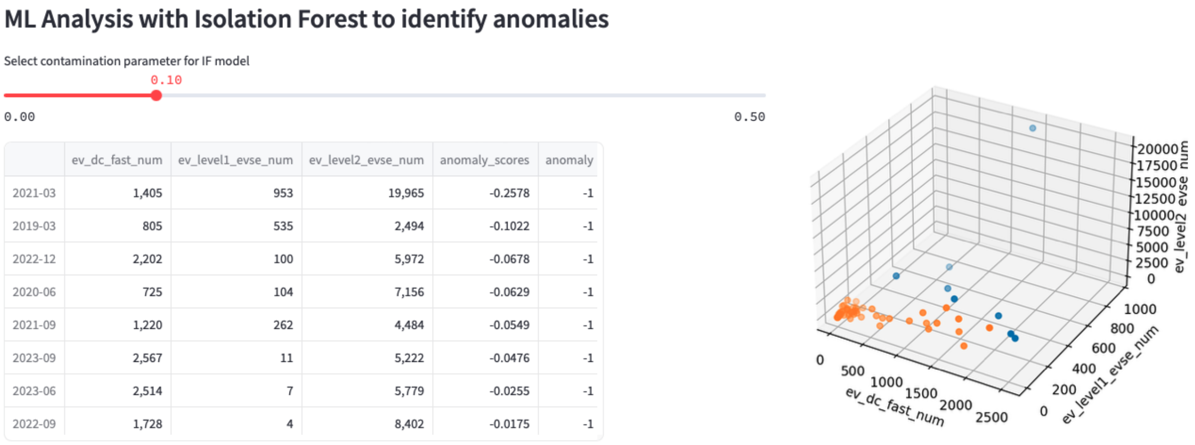
Classified timestamps for outliers are initially verified with a visual inspection of the time series plot. Points that are hypothesized to be anomalies are aligned with findings from the Isolation Forest model predictions. A classical approach to compare each point against stratified or group means was also considered but given the nature of year to year independencies, a multivariate classification approach was more suitable to address the case of overlapping anomalies by dates.

### *4.5 Model Validation and Evaluation Results*

Given the generative outputs of the unsupervised learning methods, there are limited ground truth for evaluation and one can subjectively grade the precision of anomalous dates or points. Particularly, recall from figure X that a few abnormal spikes in the level 2 chargers greatly deviate from the general trend in growth over the years. This is consistent with the Isolation Forest's model to distinguish one such anomaly during March of 2021 where nearly 20,000 level 2 superchargers were opened alongside 1400+ ds fast and 900+ level 1 chargers. Figure X shows the output table of classified dates quantified by measures of summed stations opened and a 3d scatter plot representation of the anomalous points. Refer to lines 242 - 272 for the code that initializes the model, fits on training data, displays output table and the 3d scatter plot.

**Figure X**

*Prediction table and 3d scatter plot for anomalous points*



ML Analysis with Isolation Forest to identify anomalies

Select contamination parameter for IF model

0.10

0.00 0.50

| | ev_dc_fast_num | ev_level1_evse_num | ev_level2_evse_num | anomaly_scores | anomaly |
|---|---|---|---|---|---|
| 2021-03 | 1,405 | 953 | 19,965 | -0.2578 | -1 |
| 2019-03 | 805 | 535 | 2,494 | -0.1022 | -1 |
| 2022-12 | 2,202 | 100 | 5,972 | -0.0678 | -1 |
| 2020-06 | 725 | 104 | 7,156 | -0.0629 | -1 |
| 2021-09 | 1,220 | 262 | 4,484 | -0.0549 | -1 |
| 2023-09 | 2,567 | 11 | 5,222 | -0.0476 | -1 |
| 2023-06 | 2,514 | 7 | 5,779 | -0.0255 | -1 |
| 2022-09 | 1,728 | 4 | 8,402 | -0.0175 | -1 |

# References

Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008, December). Isolation forest. In *2008 eighth ieee international conference on data mining* (pp. 413-422). IEEE.

Llanos, C. (2023, August 17). *What is the Electric Vehicle (EV) boom?: Chase*. What is the Electric Vehicle (EV) Boom? | Chase. https://www.jpmorgan.com/insights/investing/investment-strategy/what-is-the-electric-vehicle-boom#:~:text=International%20Energy%20Agency%2C%20“Demand%20for,Worldwide.”%20(2023).

Swallow, T. (2023, May 10). *Top 10 most successful EV charging businesses*. EV Magazine. https://evmagazine.com/top10/top-10-most-successful-ev-charging-businesses