Team Mini Project: Alternative Fueling Locations

Nicholas Romano, Michael Zelaya

DS450-01

Data Science Senior Capstone

Alternative Fueling Locations Report

Tools Used

We used the programming language Python and the visualization software Tableau for the Alternative Fueling Locations Project. Our team chose to use Python because the members of the team have had experience using Python in the past, and the fact that Python offers libraries that help in data cleaning and preparation. The library that we used primarily in Python was the Pandas library to help clean the data before moving on to creating visualizations. When creating the data visualizations, our team chose to use Tableau. Our team chose Tableau mainly because of its ease of use.

Cleaning and Preparing the Data

As mentioned previously, our team used the Pandas library in Python as the primary tool for cleaning and preparing the data. One of the first things our team did when we read in the data was to understand what data was being provided. The original dataset we downloaded had 56800 rows and 65 columns. Since much of the visualizations that we needed to create were based on the fuel type and locational data of the alternative fuel location, we chose to drop most information that was provided. In doing so, we choose to keep the columns on the fuel type code, station name, street address, city, state, zip code, EV network, EV network website, latitude, and longitude.

One of the first problems we ran into was the fact that there were many missing values within the EV Network and EV Network Web columns. Our intuition was that if an entry did not have an EV Network webpage associated with it then it was not going to have an EV Network to it and vice versa. Digging into the data we found that our intuition was right and that there was also a ‘Non-Networked’ EV Network value that does not have an EV Network Webpage associated with it as well. So, we filled in the missing EV Network value with ‘Non-Networked’ for all entries, even non-EV fuel types since we could filter these out in our EV Network visualizations. After replacing the missing EV Networked values, we no longer saw a need for the EV Network Web column and dropped it. Another missing value problem that we faced was that there were 2 entries with missing street addresses. Using Google Maps, we were able to obtain the street address by using station name and state.

Another problem we ran into was dealing with ZIP codes. When we read in the file, we noticed that there were a few ZIP codes with three or four digits instead of five. It turns out that the leading zero for some of the ZIP codes were being dropped. Our approach in fixing this was to check to see the number of ZIP codes that had less than 5 digits, which was 4302 out of the 56800 entries. For the ZIP code column, we formatted the column to have a datatype of string and format it to have a fixed length of five with any length shorter than five to have leading zeros placed in front of the current characters to make it five characters long.

Furthermore, we ran into a problem with concerning the accuracy of some of the locational data, mainly ZIP codes, latitude values, and longitude values. We created two functions to return index values and locational values for the highest and lowest values for each state for these values to help check for outliers – one for the ZIP codes and one for the longitude and latitude points. Reading through the rather long output for these functions, there were 51 entries of concern. Since these 51 entries make up a rather small portion of the 56800 total entries in the data, we choose to drop these entries due to concerns of the accuracy of the data.

The final problem that we faced with cleaning and preparing the data was preparing data for creating visualizations for the various regions and most populous states. For addressing the data on the regions, the dataset initially did not have a region column. So, we had to create a dictionary with the state abbreviations as the key and the region as the value, which allowed us to create a new column and extract the region value given the entry's state abbreviation. To address the data needed to create visualizations for the most populous states, we retrieved data from the US Census Bureau website on the 2020 census. We downloaded an Excel file, converted it into a CSV, cleaned it up, and sorted the population column to find the most populous states.

Visualization Process

Once we got the data cleaned and prepared to create the visualizations, we moved over to Tableau. For our visualizations, we chose to create maps, tree maps, bar charts, circle view graphs, and bubble graphs to represent the data.

The reasoning behind choosing to create maps was for the visualizations that included the locations of the alternative fuel sites. In doing so, we used the latitude and longitude points of the alternative fuel location to pinpoint its location and represented each alternative fuel location as a dot. To distinguish between alternative fuel locations, we included color in the maps where the colors were a feature of that alternative fuel location whether it be the fuel type or EV network that it belonged to. Furthermore, we also included locational details for each alternative fuel location along with the latitude and longitude points, including street name, city, and state. For the regional and state-specific maps, we placed a filter on the maps to include just the specified states or regions. Some maps we were able to compact into one worksheet using the page filter option in Tableau. For example, instead of having multiple worksheets, where each would represent one fuel type, we were able to create one worksheet that overlayed the maps for each fuel type. The filter allows the user to change the map by changing what value is being viewed.

The reasoning behind choosing tree maps, bar charts, stacked bar charts, circle view graphs, and bubble graphs were to help illustrate the breakdown between value counts for categorical data. All these graph types are useful in generating readable illustrations for value count data but are different ways of representing the data. We mainly used these graphs for the for both the state-specific and region-specific fuel type breakdown visualizations. For these graphs, we created them by dragging the data field we wanted to represent over into either the column or row and changing the illustration to display it as one of these graphs. Most of the data we were pulling over were the state abbreviations, fuel types, count of the fuel types, and EV networks. Furthermore, for state-specific and region-specific graphs we would place a filter on the data to include only the data pertaining to the state or region.

Works Cited

"2020 Census Apportionment Results". *United States Census Bureau*, 8 Oct. 2021, https://www.census.gov/data/tables/2020/dec/2020-apportionment-data.html. Accessed 13 Jan. 2024.

Map of Hawaii Transportation Department. *Google Maps*, 2024, https://www.google.com/maps/place/Hawaii+Transportation+Department/@21.3388412,-157.895119,19z/data=!4m10!1m2!2m1!1stransportation+department!3m6!1s0x7c006ef9b31239e1:0xdf3e0e891439cf37!8m2!3d21.3388706!4d-157.8942156!15sChl0cmFuc3BvcnRhdGlvbiBkZXBhcnRtZW50IgOIAQGSARxkZXBhcnRtZW50X29mX3RyYW5zcG9ydGF0aW9u4AEA!16s%2Fg%2F1w0r03j9?entry=ttu. Accessed 21 Jan. 2024.

Map of Piedmont Natural Gas CNG Station. *Google Maps*, 2024, https://www.google.com/maps/place/Piedmont+Natural+Gas+CNG+Station/@34.2822581,-77.9749824,16z/data=!4m10!1m2!2m1!1sPiedmont+Natural+Gas+-+Wilmington!3m6!1s0x89aa20311dd67f93:0x2f9718694029b861!8m2!3d34.2822584!4d-77.965455!15sCiFQaWVkbW9udCBOYXR1cmFsIEdhcyAtIFdpbG1pbmd0b24iA4gBAZIBC2NuZ19zdGF0aW9u4AEA!16s%2Fg%2F11svjrwqq2?entry=ttu. Accessed 21 Jan 2024.