Team Mini Project: Alternative Fuel Locations

Eric Schneider

Jalin Roberts

DS450-01

Data Science Senior Capstone

The initial stage of the project tasked us with using either Tableau or PowerBI to create a visual representation alternative fuel locations across the entirety of the United States. We chose to use Tableau over PowerBI due to the increased usability and likelihood of use in future job positions. In other words, because both members were more likely to use tableau over PowerBI after graduation, it made sense to learn and utilize it. Additionally, the visual representations we were tasked with creating were best handled by Tableau since its system could easily generate geographic representations based on the longitude and latitude of the locations listed in the dataset. Because of both of these factors, we chose to primarily use Tableau for this project, relying on python, specifically the pandas package, for cleaning and encoding of the region data.

Because all of the entries were given a valid latitude and longitude, very little cleaning was necessary in the initial steps of the project. The only issue which existed were three datapoints whose coordinated did not match the states which they were said to exist in. Specifically, two points were stated to be located in New York, but had coordinates slated for the United Kingdom, and a third point was said to be located in Florida but had coordinated which mapped the case to the Dominican Republic. Since there were very few values located outside of the feasible range, we decided to utilize tableau’s internal exclusion function rather than hardcoding to fully clean the data. From there we mapped the longitude and latitude of each case onto a geographic map, and sorted the fuel type codes into different colored dots. In order to achieve a specific state only, we utilized the filter to exclude any cases which did not reside in Alaska, and later Hawaii. Finally, when it came to differentiating each fuel type into different maps, we utilized the pages function. This separated the cases for each fuel type into separate geographic representations in order to better see where each of the fuel types were located. This completed the initial stage of the project, and we moved on to creating visual models for each individual region.

The main problem with creating visual representations for each region was the fact that the initial dataset lacked a column listing the region that each case resided in. This meant that we had to manually encode the dataset with the regions. This process was done through Python, specifically with the pandas package. A method was created named Region Encode, which took a single case from the dataset and returned the region that case was located in. This was done by matching the state codes with lists containing the same codes, differentiated by region. A for loop was used to match the state codes, with a value counter which would count the index values of the regions list. When a match was found, the method returned the item in the region names list with the matching index of the counter established before the for loop. The fact that the counter was listed before the for loop ensures that the value will not be overwritten every time the loop continues, as well as overwriting it each time a new case is taken after the region for the previous case is returned. From there, an external for loop was initiated which took the individual cases from the state column of the dataset and appended the returned region to the previously initialized list named us\_region. Finally, after the list containing the regions for each case was created, the list was assigned to the original dataset to map each case with their respective region internally. From there, the newly encoded fuel data was exported to a csv text file for future use in tableau.

The creation of the visuals for each region had the same process as the nationwide data, with one specific addition. Each region required a part to a whole graph showing the percentage of each fuel type as it made up the entire region. The way we achieved this was through a bar graph, as it was the best for showing data split up into a relatively small number of groups. Each state data was mapped to a different representation through the pages function in Tableau. The third task for the project was to focus on the specific ev networks in each region. The same process as other tasks was followed, except with the addition of using the filter to only count the cases with the ELEC fuel type code. For the final task, we filtered out the top ten most populous states by filtering the state codes and manually selecting the top ten most populous states. Then, to separate each visualization, we placed the state variable in the pages section. From this, we noticed that there were a few cases which had been mapped to the wrong state than they had been listed as. However, as this amount was less than .1% of all of the data, we concluded that it would not alter any analysis in a significant way. Instead, like with the three cases located out of the united states, an effective visualization would’ve simply used the exclusion function located inside of Tableau rather focusing on cleaning the data.