Joseph Cardenas
IS 497
April 10th, 2022
Final Project Proposal

Idea: Clean some data with open refine and write a python script that will read in the json file and add those changes to a new version of the data.

Dataset selection

My data that I am going to be working on is the electric vehicle population. This data is going to be for the state of Washington having all the raw data of all the electric cars registered to the state. Furthermore, it's going to have anything from VIN number to the type of car and location.

Problem formulation

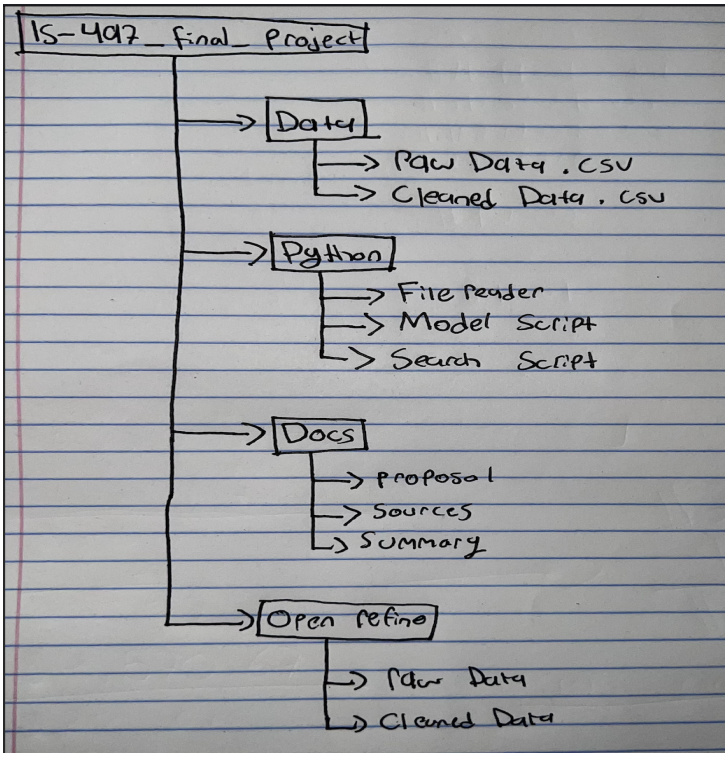
The problem with this data is tons of data. Having just about 10000 cells of data. This data is also all over the place having some cells with unorganized data and it is not nice and neat. Therefore, I plan to use open refinement and cleaning my data so that it's nice and neat. This will allow me to see my data much easier. I will also delete or add to cells that need to be changed to keep the same format overall keeping everything in the same flow. Next I plan on using the csv file in python so that it can read the changes and be able to view the data. Finally, I plan on writing a script so you will be able to search any type of model and all the cleaned data for that model car will appear in the python script.

Dataset description

The data set was found from “https://catalog.data.gov/dataset”. This is a government website in which you are able to search any type of dataset in any format csv, json, pdf, and etc. This website is great because they have a huge catalog of data in which you are able to find all the information for and download however you want. Furthermore, while looking for dataset I have always been curious about electric cars as they are coming up in our new technology being more and more popular over the years. As I searched for something similar I found this data set for the state of Washington. This data set was the population of electric vehicles in the state of Washington having as much detail as possible. The dataset was created on November 10, 2020 and last updated on March 30, 2022. Therefore, the data set is two years old which can be a problem with the huge increase in electric cars over the two years but has been updated in

the last few weeks having more data for those two gap years. Nevertheless, the data set is huge having just a little less than 10000 registered electric cars in the data. It also has tons of data like VIN numbers, county, city, model year. Model, location, type of electric car, and etc. This data set is very broad in the point that if you want to know about a specific electric vehicle in the state of washington you are able to find out everything about it with this data set. Therefore, because of how interesting and huge the data set is, it's what got my attention and wanting to be able to work with it. As big as this data is it's also all over the place being difficult to read and not find what you are looking for if you don't know how. Furthermore, this is also a key thing that draws my attention to this data. The need for it to be organized and cleaned is what wanted me to take on the challenge and be able to run scripts in which I can search what I want to find from it. Also, electric cars are something I've always been interested in. It's something I hope I own in the future and will like to know what electric cars are being bought. Therefore, with all these things I list is really why I decided to work with this data and allow me to apply what I learned from this class into this data set.

Project organization



Project repository

Create the repository for your project in github (or similar alternative). You can make it private if you need, but you will need to add the instructional team as collaborators so we can see it.

LINK: https://github.com/josephc90/IS-497_FInal_Project

Source:

https://catalog.data.gov/dataset/electric-vehicle-population-data