ETL Project

By Zach Spahr, Manuel Jean, Mario Iglesias

Extraction

We used three data sets for this project. The first one was a dataset from Data World that displayed the percent of people who have preexisting health conditions by congressional district. The second data set , which came from Kaggle, displays political fundraising data, election results, and demographic characteristics of each congressional. The third data set is Cook Political Report’s Dave Wasserman spreadsheet with the results of every congressional election in the 2018 midterms. (Note I have worked with this data for other projects and had made some edits it to before inputting into python myself.

<https://data.world/carlvlewis/pre-exisiting-conditions-by-state-congressional-district/workspace/file?filename=pre-existing-conditions-by-congressional-district.xlsx+-+number+with+pre-ex+by+district.csv>

<https://www.kaggle.com/landonwall/aggregate-congressional-district-data>

<https://docs.google.com/spreadsheets/d/1WxDaxD5az6kdOjJncmGph37z0BPNhV1fNAH_g7IkpC0/edit>

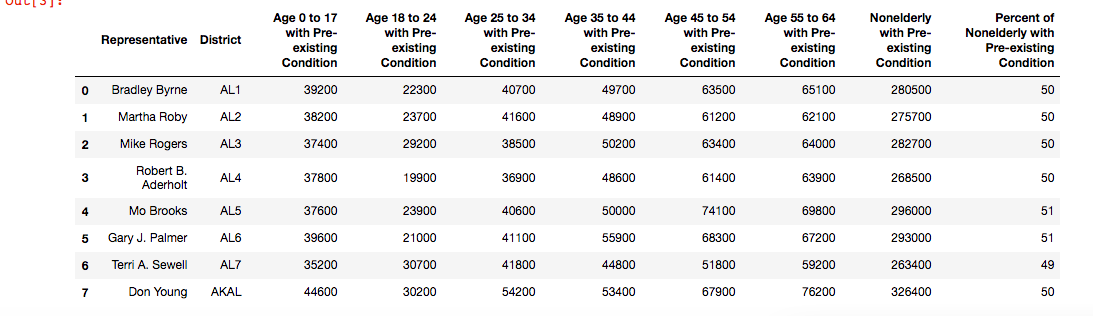
Transformation

Initial Data Set:

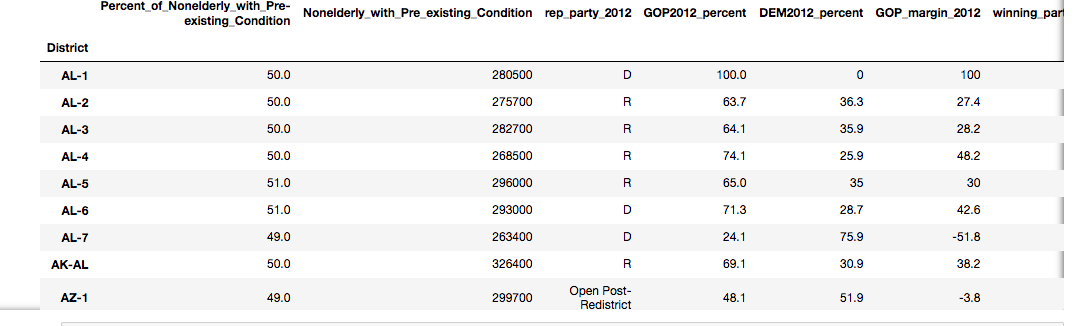


The first step in cleaning this data set was to eliminate the commas from each number on the table so that we could cast each one of the main columns as an integer or float rather than a string. We also eliminated the hyphen and the parenthesis in the “District” column to set that column up to merge with the larger data set containing election results and demographic information from congressional districts. Lastly, we dropped the “Representative (District)” column because it became redundant at this point.

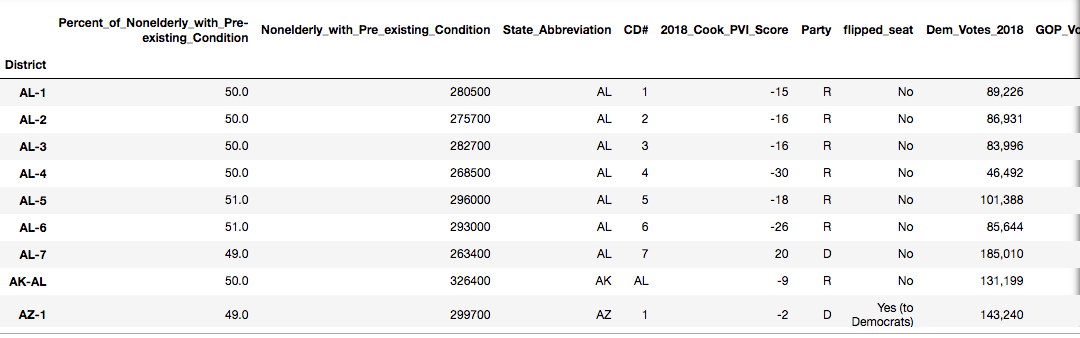
Pre Existing Condition Data After Initial Transformation:



After this, the next step was to merge this data set with the other two data sets because the other data sets already had the same values in the District column as this data set does.



This is the merge of the pre-existing condition data set with the congressional districts data set from Kaggle. we dropped the fundraising data from the congressional districts because those numbers took up a lot of columns and were not being analyzed for the purpose of this project. Also, after the merge was completed, we reset the district column to have a hyphen between the state abbreviation and the number of the congressional district because this is normally how congressional districts are displayed when they are referred to in the media graphics. The data frame below is the merge with the Cook Political Report data merged with the pre-existing conditions data.

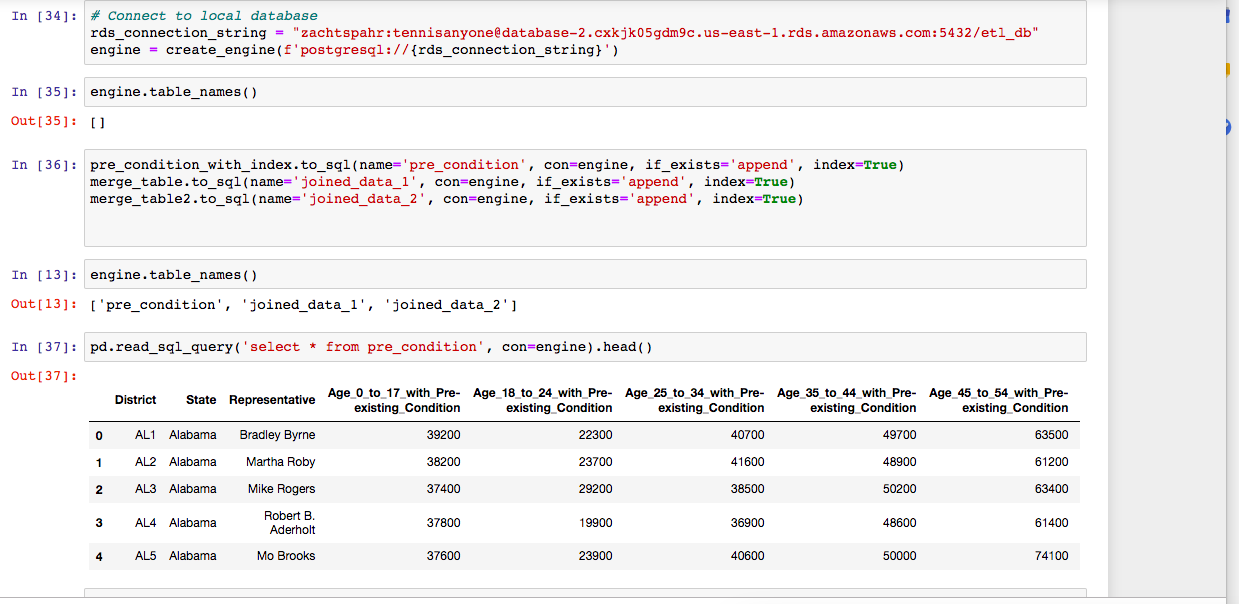


Note: We used the .Iloc function to show a section of this. The much larger data set from these merges are stored in the merged data csv’s in the GitHub repository.

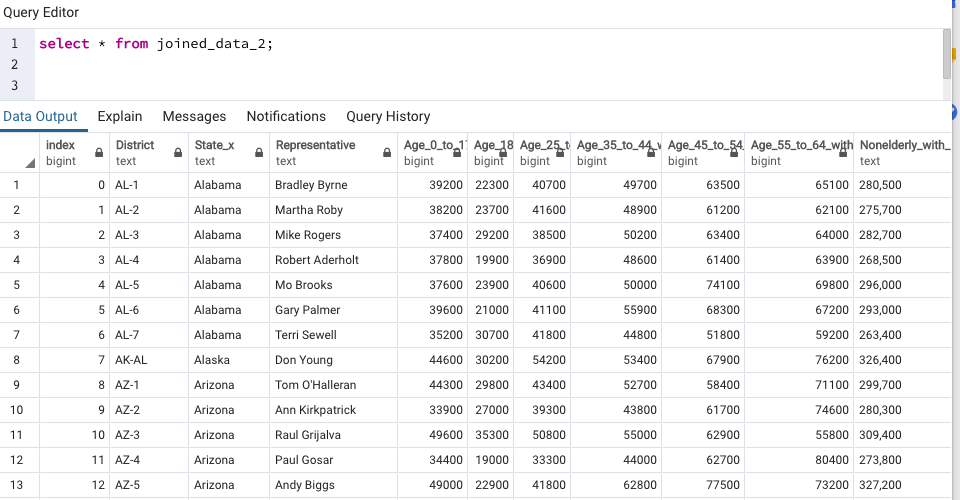
**Load:**

The last step of this process was to put the data from pandas into a database that was stored in a cloud. Using SQL Alchemy and pandas functions, we were able to convert the tables into a database in Postgres. This Postgres database is connected to amazon webservices.

Here is how this process went in python :



Here is that database queried into Postgres using PG-admin:



Concluding Thoughts:

By bringing these three data sets together, we could examine several relationships. First, it would be interesting to see if the variation in the rates of pre-existing conditions has any noticeable relationship with how people have voted. One might expect that districts with more pre-existing conditions would be more supportive of the affordable act, and; therefore, possibly more supportive of Democrats; however, it is unclear to me if there is enough variation in the different districts to know if that is the case. It would also be interesting to continue to update this dataset with future election results to see how demographic trends will continue to shape political behavior in the near future.