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_g7lkp C0/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:

	Representative	District	Age 0 to 17 with Pre- existing Condition	Age 18 to 24 with Pre- existing Condition	Age 25 to 34 with Pre- existing Condition	Age 35 to 44 with Pre- existing Condition	Age 45 to 54 with Pre- existing Condition	Age 55 to 64 with Pre- existing Condition	Nonelderly with Pre- existing Condition	Percent of Nonelderly with Pre-existing Condition
0	Bradley Byrne	AL1	39200	22300	40700	49700	63500	65100	280500	50
1	Martha Roby	AL2	38200	23700	41600	48900	61200	62100	275700	50
2	Mike Rogers	AL3	37400	29200	38500	50200	63400	64000	282700	50
3	Robert B. Aderholt	AL4	37800	19900	36900	48600	61400	63900	268500	50
4	Mo Brooks	AL5	37600	23900	40600	50000	74100	69800	296000	51
5	Gary J. Palmer	AL6	39600	21000	41100	55900	68300	67200	293000	51
6	Terri A. Sewell	AL7	35200	30700	41800	44800	51800	59200	263400	49
7	Don Young	AKAL	44600	30200	54200	53400	67900	76200	326400	50

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.

	Percent_of_Nonelderly_with_Pre- existing_Condition	Nonelderly_with_Pre_existing_Condition	rep_party_2012	GOP2012_percent	DEM2012_percent	GOP_margin_2012 w	vinning_pa
District							
AL-1	50.0	280500	D	100.0	0	100	
AL-2	50.0	275700	R	63.7	36.3	27.4	
AL-3	50.0	282700	R	64.1	35.9	28.2	
AL-4	50.0	268500	R	74.1	25.9	48.2	
AL-5	51.0	296000	R	65.0	35	30	
AL-6	51.0	293000	D	71.3	28.7	42.6	
AL-7	49.0	263400	D	24.1	75.9	-51.8	
AK-AL	50.0	326400	R	69.1	30.9	38.2	
AZ-1	49.0	299700	Open Post- Redistrict	48.1	51.9	-3.8	

This is the merge of the pre-exisiting 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.

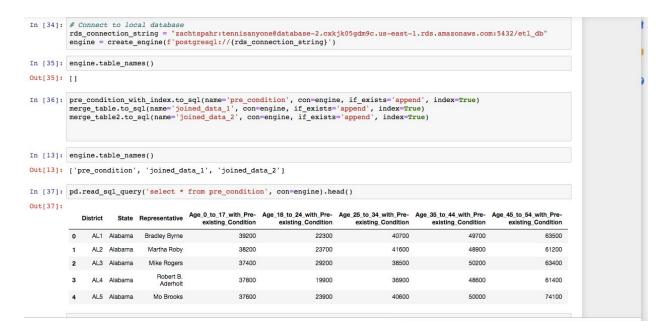
	Percent_of_Nonelderly_with_Pre- existing_Condition	$Nonelderly_with_Pre_existing_Condition$	State_Abbreviation	CD#	2018_Cook_PVI_Score	Party	flipped_seat	Dem_Votes_2018	GOP_V
District									
AL-1	50.0	280500	AL	1	-15	R	No	89,226	
AL-2	50.0	275700	AL	2	-16	R	No	86,931	
AL-3	50.0	282700	AL	3	-16	R	No	83,996	
AL-4	50.0	268500	AL	4	-30	R	No	46,492	
AL-5	51.0	296000	AL	5	-18	R	No	101,388	
AL-6	51.0	293000	AL	6	-26	R	No	85,644	
AL-7	49.0	263400	AL	7	20	D	No	185,010	
AK-AL	50.0	326400	AK	AL	-9	R	No	131,199	
AZ-1	49.0	299700	AZ	1	-2	D	Yes (to Democrats)	143,240	

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.

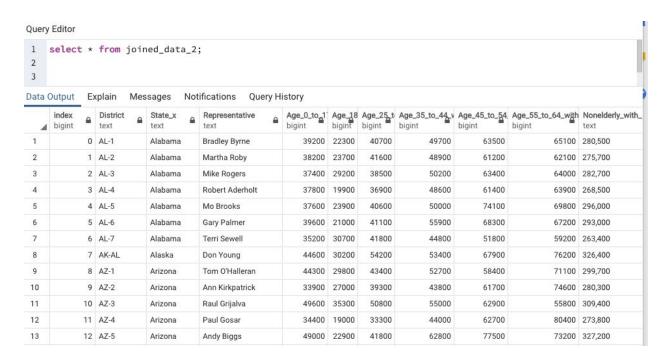
Load:

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 pgadmin:



Concluding Thoughts:

Manipulating this data and loading into a database would allow for many different types of data analysis to be covered. First, it would be interesting to see if the occurrence of pre-existing

conditions has any noticeable correlation 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, 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 are shaping political behavior in the near future.