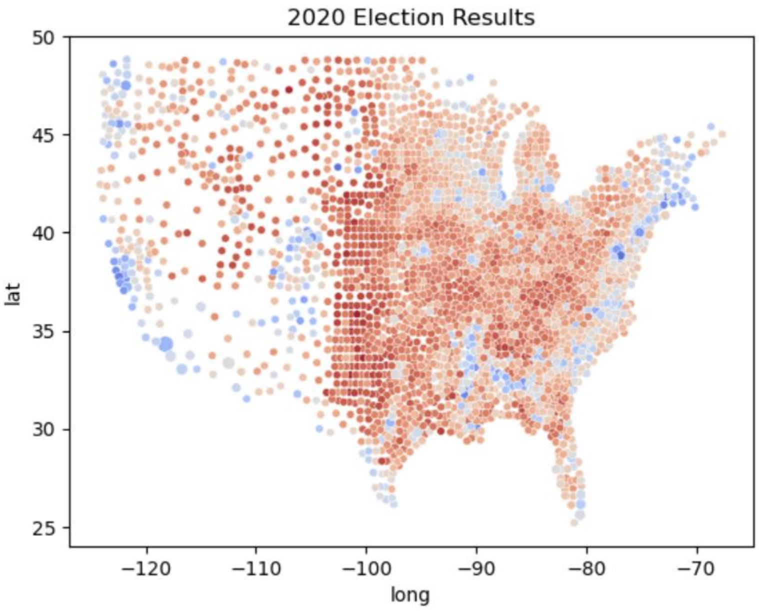
ADSP 31014 Statistical Models for Data Science

Course Project Part 1 - Hira Stanley

Overview of Data Set

This data set contains 4,954 rows of data, each pertaining to a county in the United States. With 3,143 counties in the U.S., all counties were included at least once. Each county is supposed to have records of the 2016 and 2020 election results, combined with demographic, economic, and COVID-related figures. The demographic and economic data is from 2017, so we do not have the ability to detect shifts in these metrics from 2016 to 2020. The COVID data is from November 1, 2020, which is after the early voting period had begun in 2020, but is still a good representative of the pervasiveness of COVID in that county in 2020.



Election, COVID, and Demographic Data by County

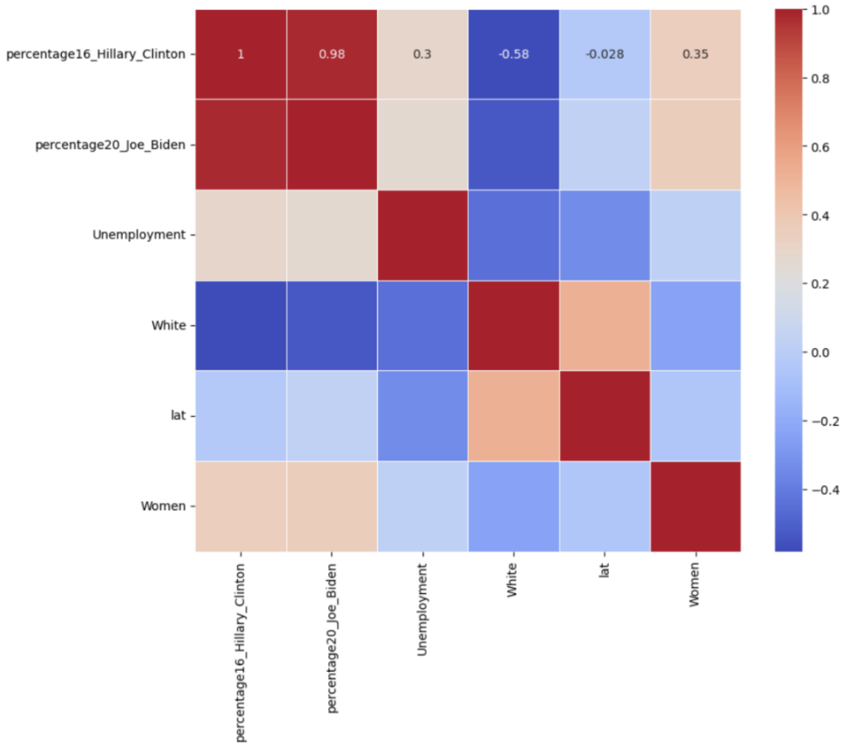
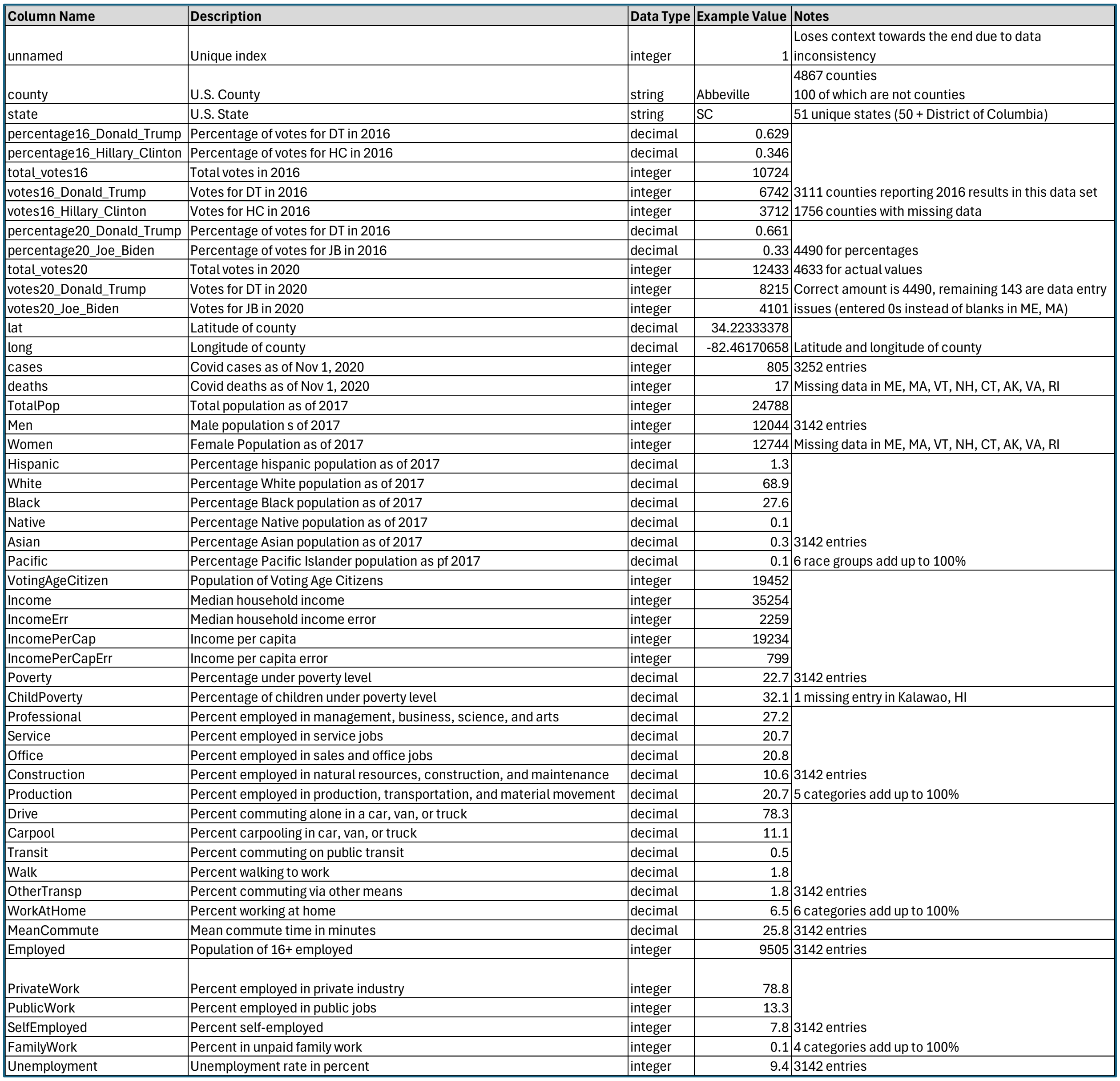


Figure 1 (above): Percentage of votes won by Donald Trump (red) in 2020, sized by total votes in 2020.

Figure 2 (left): Some preliminary analysis on the dataset shows some interesting correlations and lack of correlations between fields. Fields include election result data, an economic indicator, demographic and geographic data.

Data Table Schema

The data set consists of 4,867 rows and 51 columns. Each row represents a county in the United States.



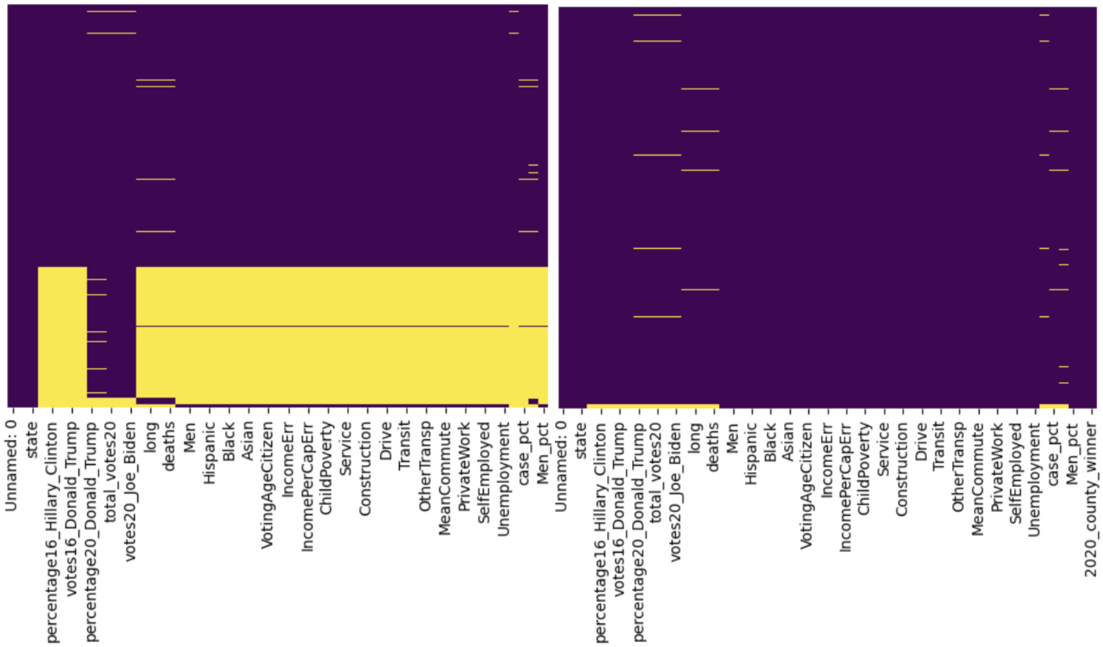
Column Name	Description	Data Type	Example Value	Notes
unnamed	Unique index	integer	1	Loses context towards the end due to data inconsistency
county	U.S. County	string	Abbeville	4867 counties 100 of which are not counties
state	U.S. State	string	SC	51 unique states (50 + District of Columbia)
percentage16_Donald_Trump	Percentage of votes for DT in 2016	decimal	0.629	3111 counties reporting 2016 results in this data set 1756 counties with missing data
percentage16_Hillary_Clinton	Percentage of votes for HC in 2016	decimal	0.346	
total_votes16	Total votes in 2016	integer	10724	
votes16_Donald_Trump	Votes for DT in 2016	integer	6742	
votes16_Hillary_Clinton	Votes for HC in 2016	integer	3712	
percentage20_Donald_Trump	Percentage of votes for DT in 2016	decimal	0.661	4490 for percentages 4633 for actual values Correct amount is 4490, remaining 143 are data entry issues (entered 0s instead of blanks in ME, MA)
percentage20_Joe_Biden	Percentage of votes for JB in 2016	decimal	0.33	
total_votes20	Total votes in 2020	integer	12433	
votes20_Donald_Trump	Votes for DT in 2020	integer	8215	
votes20_Joe_Biden	Votes for JB in 2020	integer	4101	
lat	Latitude of county	decimal	34.22333378	Latitude and longitude of county
long	Longitude of county	decimal	-82.46170658	
cases	Covid cases as of Nov 1, 2020	integer	805	3252 entries Missing data in ME, MA, VT, NH, CT, AK, VA, RI
deaths	Covid deaths as of Nov 1, 2020	integer	17	
TotalPop	Total population as of 2017	integer	24788	3142 entries Missing data in ME, MA, VT, NH, CT, AK, VA, RI
Men	Male population s of 2017	integer	12044	
Women	Female Population as of 2017	integer	12744	
Hispanic	Percentage hispanic population as of 2017	decimal	1.3	3142 entries 6 race groups add up to 100%
White	Percentage White population as of 2017	decimal	68.9	
Black	Percentage Black population as of 2017	decimal	27.6	
Native	Percentage Native population as of 2017	decimal	0.1	
Asian	Percentage Asian population as of 2017	decimal	0.3	
Pacific	Percentage Pacific Islander population as pf 2017	decimal	0.1	
VotingAgeCitizen	Population of Voting Age Citizens	integer	19452	3142 entries
Income	Median household income	integer	35254	
IncomeErr	Median household income error	integer	2259	
IncomePerCap	Income per capita	integer	19234	
IncomePerCapErr	Income per capita error	integer	799	
Poverty	Percentage under poverty level	decimal	22.7	
ChildPoverty	Percentage of children under poverty level	decimal	32.1	1 missing entry in Kalawao, HI
Professional	Percent employed in management, business, science, and arts	decimal	27.2	3142 entries 5 categories add up to 100%
Service	Percent employed in service jobs	decimal	20.7	
Office	Percent employed in sales and office jobs	decimal	20.8	
Construction	Percent employed in natural resources, construction, and maintenance	decimal	10.6	
Production	Percent employed in production, transportation, and material movement	decimal	20.7	
Drive	Percent commuting alone in a car, van, or truck	decimal	78.3	3142 entries 6 categories add up to 100%
Carpool	Percent carpooling in car, van, or truck	decimal	11.1	
Transit	Percent commuting on public transit	decimal	0.5	
Walk	Percent walking to work	decimal	1.8	
OtherTransp	Percent commuting via other means	decimal	1.8	
WorkAtHome	Percent working at home	decimal	6.5	
MeanCommute	Mean commute time in minutes	decimal	25.8	3142 entries
Employed	Population of 16+ employed	integer	9505	3142 entries
PrivateWork	Percent employed in private industry	integer	78.8	3142 entries 4 categories add up to 100%
PublicWork	Percent employed in public jobs	integer	13.3	
SelfEmployed	Percent self-employed	integer	7.8	
FamilyWork	Percent in unpaid family work	integer	0.1	
Unemployment	Unemployment rate in percent	integer	9.4	3142 entries

Data Cleaning and Processing

The main data quality issues in this data set are inconsistency in the county data. Because the data set is combing data from multiple sources like the U.S. Census, election results, and COVID results, there is a mismatch in the county names.

The data set has 4,867 entries for counties, corresponding with a specific "county". However, further analysis shows 100 of these entries are "Unassigned" and begin with "Out of [state name]", which are catch-all buckets for additional COVID cases and deaths by state that are not assigned to a particular county. These entries are removed as our analysis will be at the county-level.

The states showing inconsistent data are ME, MA, VT, NH, CT, AK, VA, and RI. External research helps us understand what is going on. For example, Maine (ME) has 16 counties officially, but in this data set, it has 506 entries. 16 of these 506 entries correspond to real county data, with 6 rows (out of 16) missing values in the 2020 election results. The remaining 490 entries are these 16 counties broken down further with town-specific 2020 election data. Because the actual county election data corresponds to 16 real counties, and is matched with 16 entries of U.S. Census and COVID data, it is sufficient for analysis. This means, we can remove the town-level detail. Doing so for the 8 states with this data inconsistency issue results in 3,142 entries/counties in Figure 3 (below), which aligns closely with the 3,143 counties in the United States.



One of the main factors to look at when analyzing the drivers of the 2020 election result is the 2016 election result. There are only 2 entries where 2016 results are missing values, but their corresponding 2020 results are available. This appears to be at random. Alternatively, there are 22 entries where 2016 results are available, but not 2020 results. These 22 entries all belong to states with the earlier discussed data inconsistency issue, meaning that some of these 22 results may be present at the town-level that we separated from this data set.

Other columns added to this data frame were:

- Turnout_difference to show difference in total votes between 2016 and 2020, as additional or fewer voters may have an impact
- Case_pct to show COVID cases as a percentage of the total population of that county
- Death_pct to show COVID deaths as a percentage of COVID cases in that county
- Men_pct to show men as a percentage of total population in that county
- 2016_county_winner to add a 'D' or 'R' label as the winner should we run any classification techniques in the future on this data set
- 2020_county_winner to add a 'D' or 'R' label as the winner should we run any classification techniques in the future on this data set

The resulting data set contains 3,142 rows and 57 columns.