Data Description

This is the guidance to help you to understand how the data look like. “Data\_Wrangling” folder consists of two parts: “Raw\_data” and “Final\_data”. “Raw\_data” contains all sources for cleaned and validated datasets in “Final\_data.” There are 54 sources files in “Raw\_data” and 8 files in “Final\_data.” Each file and directories were grouped by *number* by topic (1 ~ 4):

Topic 1: Early Vote data in Texas by County in 2020 Presidential Election,

Topic 2: Presidential Election data by State from 1976 to 2016,

Topic 3: Demographic data in Texas by County from 2010 to 2019

Topic 4: General Election (Governor and Senator) data in Texas in 2018

*\* Topic 5: Voter Registration data Texas from 1996 to 2020 (it was combined with   
Topic 1. Please see “1.Early\_Vote\_data.csv” below for detail).*

***Final\_data***

*1. Early\_Vote\_data.csv (+* Voter Registration data in Texas by County from 1996 to 2020)

\* This dataset is combined with Voter Registration.

This dataset shows the number of early votes by 1) in\_person and 2) mail-in by year from 1996 to 2020. Additionally, it shows the number of precincts, detail about the registered voters with the number suspense voters and non-suspense voters.

Notice that original source only reports top 15 county by Registered\_Voters from 1996 to 2016. In 2020, we have data for all 254 counties. This dataset contains 338 rows and 12 columns.

* County: name of the county
* Precints: number of precincts in the county
* Registered\_Voters: total number of registered voters   
   (Suspense\_Voters + Non\_Suspense\_Voters)
* Suspense\_Voters: number of voters who are suspended
* Non\_Suspense\_Voters: number of voters who can actually vote.
* Cumulative\_In\_person: cumulative number of in-person voter in *Year*.
* Cumulative\_In\_Person\_Percent:

Cumulative\_In\_person / Non\_Suspense\_Voters \* 100 in *Year*

* Cumulative\_By\_Mail: cumulative number of in-person voter in *Year*.
* Cumulative\_By\_Mail\_Percent:

Cumulative\_By\_Mail / Non\_Suspense\_Voters \* 100 in *Year*

* Cumulative\_In\_Person\_By\_Mail: Cumulative\_In\_person + Cumulative\_By\_Mail
* Cumulative\_Percent\_Early\_Vote:

Cumulative\_In\_Person\_By\_Mail / Non\_Suspense\_Voters \* 100 in *Year*

* Year: year of the data election

*Sources: Texas Secretary of State, https://earlyvoting.texas-election.com/Elections/getElectionEVDates.do for early vote data,* https://www.sos.texas.gov/elections/historical/vrfig.shtml for registered voters data

*1. Early\_Vote\_data\_By\_Day.csv*

This dataset shows the number of early votes by 1) in\_person and 2) mail-in from Oct/04/2020 to Oct/30/2020 by County. It contains 6858 rows and 7 columns. I removed the Registered\_Voters column from original sources and import the Non\_Suspense\_Voters from the registered voters data for accurate and consistent analysis.

* Date: date of the data collected
* County: name of the county
* Non\_Suspense\_Voters: number of voters who can actually vote.
* In\_person: number of in-person voter on *Date*.
* In\_Person\_Percentage: In\_person / Non\_Suspense\_Voters \* 100 on *Date*
* Mail: number of mail-in voter on *Date*.
* Mail\_Percentage: Mail / Non\_Suspense\_Voters \* 100 on *Date*

*Sources: Texas Secretary of State, https://earlyvoting.texas-election.com/Elections/getElectionEVDates.do*

*2.General\_Election\_Data.csv*

This dataset shows the summary of the presidential election in Texas from 1976 to 2016. It contains 22 rows and 6 columns. Each year has two observations that indicate winner and runners-up for the election.

* Year: year of the election
* Candidate: name of the candidate
* Party: candidate’s party (democrat or republican)
* Candidatevotes: number of votes the candidate received
* Totalvotes: total number of registered voters
* Result: dummary variable (either winner or runners-up)

*Sources: Harvard Dataverse, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/42MVDX*

*3.2018\_Demographic\_Data\_By\_County.csv / 2020\_Demographic\_Data\_By\_County.csv*

This dataset shows the demographic information in the Texas by county by age group in 2018/2020. Each dataset contains 3556 rows and 21 columns. For detail description of the original data, you can check Data\_Wrangling/Raw\_data/3.Demographic/cc-est2019-alldata.pdf

* CTYNAME: name of the county
* AGEGRP: categorical variable indicating the age:
  + 4 = Age 15 to 19 years
  + 5 = Age 20 to 24 years
  + 6 = Age 25 to 29 years
  + 7 = Age 30 to 34 years
  + 8 = Age 35 to 39 years
  + 9 = Age 40 to 44 years
  + 10 = Age 45 to 49 years
  + 11 = Age 50 to 54 years
  + 12 = Age 55 to 59 years
  + 13 = Age 60 to 64 years
  + 14 = Age 65 to 69 years
  + 15 = Age 70 to 74 years
  + 16 = Age 75 to 79 years
  + 17 = Age 80 to 84 years
  + 18 = Age 85 years or older

Notice that AGEGRP == 4 is dropped because it contains the people who are not eligible for voting (age < 18).

*(Continued on the next page)*

* TOT\_POP: Total population (TOT\_MALE + TOT\_FEMALE **OR** WA\_MALE + WA\_FEMALE + BA\_MALE + BA\_FEMALE + IA\_MALE + IA\_FEMALE + AA\_MALE + AA\_FEMALE + NA\_MALE + NA\_FEMALE + TOM\_MALE + TOM\_FEMALE **OR** NH\_MALE + NH\_FEMALE + H\_MALE + NH\_FEMALE)
* TOT\_MALE: Total male population
* TOT\_FEALE: Total female population
* WA\_MALE: White male population
* WA\_FEMALE: White Female population
* BA\_MALE: Black or African American male population
* BA\_FEMALE: Black or African American female population
* IA\_MALE: American Indian and Alaska Native alone male population
* IA\_FEMALE: American Indian and Alaska Native alone female population
* AA\_MALE: Asian male population
* AA\_FEMALE: Asian female population
* NA\_MALE: American Indian and Alaska Native alone male population
* NA\_FEMALE: American Indian and Alaska Native alone female population
* TOM\_MALE: Two or More Races male population
* TOM\_FEMALE: Two or More Races female population
* NH\_MALE: Not Hispanic male population
* NH\_FEMALE: Not Hispanic female population
* H\_MALE: Hispanic male population
* H\_FEMALE: Hispanic female population

*3.2018\_Demographic\_Data\_By\_County\_Total.csv / 2020\_Demographic\_Data\_By\_County\_Total.csv*

This dataset is the **aggregated** version of *2018\_Demographic\_Data\_By\_County.csv* and *2020\_Demographic\_Data\_By\_County.csv.* For example, in *2018\_Demographic\_Data\_By\_County.csv* file, if you add each column by County (i.e., we do not separate the data by AGEGRP), you will obtain this dataset. Notice that I do not include AGEGRP = 1, 2, 3, 4 from original dataset as I described earlier. Each dataset contains 250 rows and 20 columns.

NOTE: Since the original dataset does not contain 2020 data, I estimate the population using the percent change between 2018 and 2019. In other words, I impose the assumption that the percentage of increasing or decreasing of population is the same between 2018-2019 and 2019-2020.

*Sources: CENSUS, https://www.census.gov/data/tables/time-series/demo/popest/2010s-counties-detail.html*

*4.2018\_Election\_Data.csv*

This dataset shows the result of the 2018 general election in Texas. It contains 508 (254 \* 2) rows and 8 columns. First 254 observations are for Governor election and next 254 observations are for Senator election.

*(Continued on the next page)*

* County: name of the county
* REP: total number of votes for Democratic Party
* DEM: total number of votes for Republican Party
* LIB: total number of votes for Libertarian Party
* Total\_Voters: total number of votes (REP + DEM + LIB)
* Registered\_Voters: total number of registered voters.
* Turn\_out: Turn out rate (Total\_Voters / Registered\_Voters \* 100)
* Race: indicator for type of race (either Governor or Senator).

*Sources: Texas Secretary of State, https://earlyvoting.texas-election.com/Elections/getElectionEVDates.do*