<https://github.com/Hep7vv/DS3001-Final-Project>

1. What is in your data?

County Data contains many important demographic measures such as general population numbers broken down by sex, age, and race. There's also more specific categories such as educational attainment >25yrs, household income, industry employed, home value, and rent all generally broken down into the same categories. When joined with voting\_VA trends can be observed between voting habits and household/individual traits. Voting data is provided for every national election since 2000 and demographic data since 2008with votes for individual candidates and parties (listed as Republican, Democrat, Other).

**Key Variables:**

**voting\_VA**

* year
* state
* state\_po
* County\_name
* county\_fips (county id)
* office (all president)
* Candidate
* Party
* Candidatevotes
* Totalvotes
* Version
* mode (all are total)

**county\_data**

* Sex by Age
* Median Age by Sex
* Total Population
* Race
* Hispanic or Latino Origin by Race
* Sex by Educational Attainment for the Population 25 Years and Over
* Ratio of Income to Poverty Level in the Past 12 Months
* Household Income in the Past 12 Months
* Per Capita Income in the Past 12 Months
* Aggregate Income in the Past 12 Months
* Sex by Industry for the Civilian Employed Population 16 Years and Over
* Housing Units
* Occupancy Status
* Tenure (occupied housing Units)
* Units in Structure
* Year Structure Built
* Median Gross Rent
* Median Value (of housing in dollars)

2. How will these data be useful for studying the phenomenon you're interested in?

These variables break down voting habits by county within Virginia. By joining this dataset with another that contains more specific demographic information we hope to predict future voter tallies and election results. In doing so, we can study voter turnout (how it has changed over time) and candidate performance (compare the vote share of specific candidates across a myriad of counties to identify where they are stronger or weaker) to reveal how each county may vote.

3. What are the challenges you've resolved or expect to face in using them?

We encountered a few initial challenges when merging the two datasets, **voting\_VA** and **county\_data**. The first issue was that the 2020 **voting\_VA** data included excess information detailing different voting methods (e.g., absentee and ballot voting). This initially skewed the data by nearly tripling the overall vote count per county in Virginia for 2020. We have some ideas on how to address this issue to further clean our data, which will be our next step.The second issue arose after merging the two datasets. Every fourth row contained the relevant information we needed. While this was a minor issue, we resolved it by excluding all irrelevant data and keeping only every fourth row.

In the end, we don’t expect any further challenges, but it was important to clean and refine the data early on to avoid any major setbacks.