Data Selection and Exploration

We chose to use data from the Election Administration and Voting Survey (EAVS). We are using two datasets from them; one from 2016, and one from 2020. This survey contains data pertaining to the administration of voting, rather than results themselves; Examples of this include the number of ballots cast, the number of ballots returned, the number of mail-in ballots received, and how many ballots were cast on any given polling machine make and model. This survey is filled out by election officials in each county or county equivalent, and each row is identified by the FIPS Code for the county or county equivalent the vote was cast in.

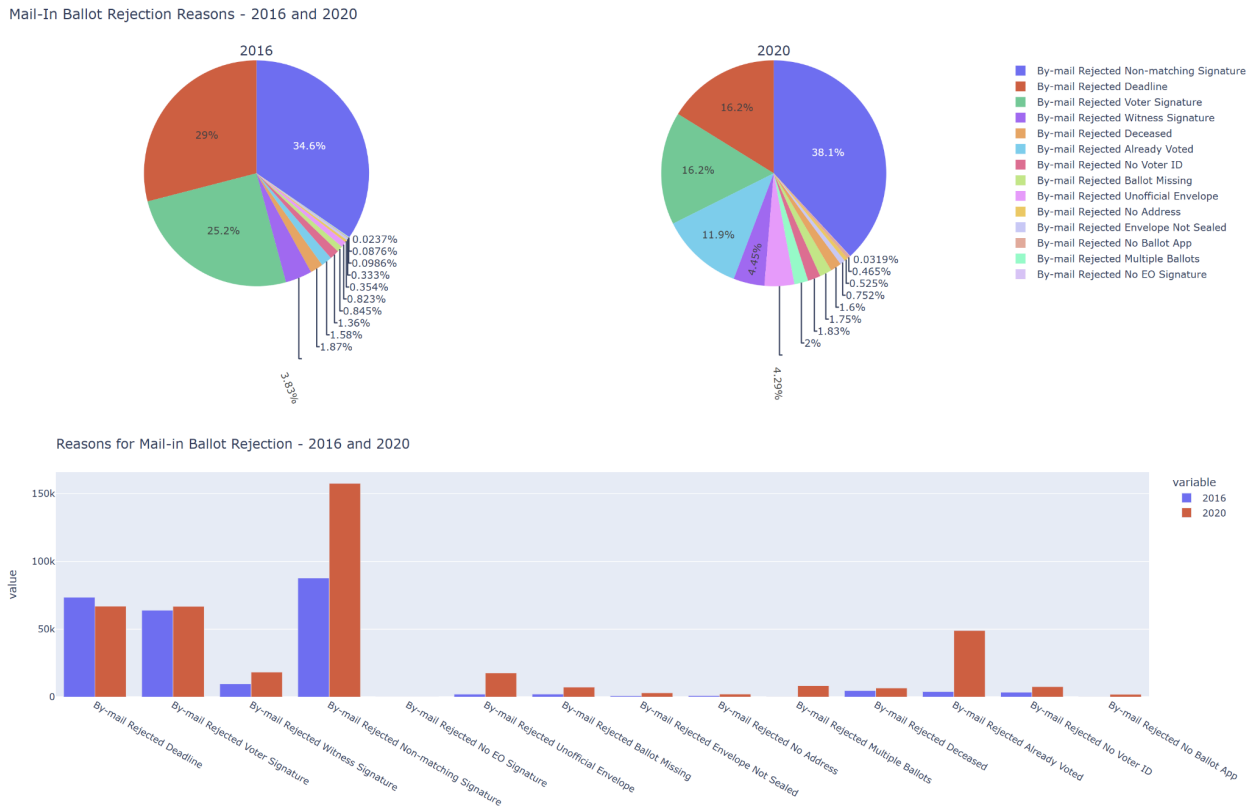
To explain what we specifically are interested in observing here, we want to take a data-driven approach to observing the change in mail-in voting. Section C of the EAVS has questions pertaining to how many mail in votes were sent out, received, rejected, voided and more. The idea behind using both 2016 and 2020 data is to see how these amounts changed in response to the Covid-19 pandemic. While it is pretty obvious that the absolute number of mail-in votes increased, we also want to look at whether things like the rate of voided ballots changed, or how many people requested absentee ballots but still ended up casting provisional ballots at polling places on election day.

There are an exceptional amount of data points in each dataframe, as there are just under 6500 observations, along with over 400 columns. It is worth noting that the 2016 and 2020 dataframes have their differences, particularly in their columns; Some questions were added, removed, or changed in the interim between when the surveys were conducted. An exploration of the data led us to these specific notes to consider in our cleaning and analysis processes:

- Missing data vs. Intentionally blank data: Data that was coded -888888: Not Applicable and -999999: Data Not Available needed to be removed. Additionally, codes were slightly different between 2016 and 2020.
- Comments: There are columns that contain the responses of those filling out the survey, who can leave comments on each question. These responses, while anecdotally interesting, require a level of analysis beyond our ability (language processing) to use in any statistically significant way, and as such are being excluded.
- The survey had changed, and a lot of effort was put into ensuring that the questions were lined up (such as question C3a in 2016 being question C2a in 2020). A table was made to track questions that were the same and comparable and is attached at the end of the document.

Data Analysis

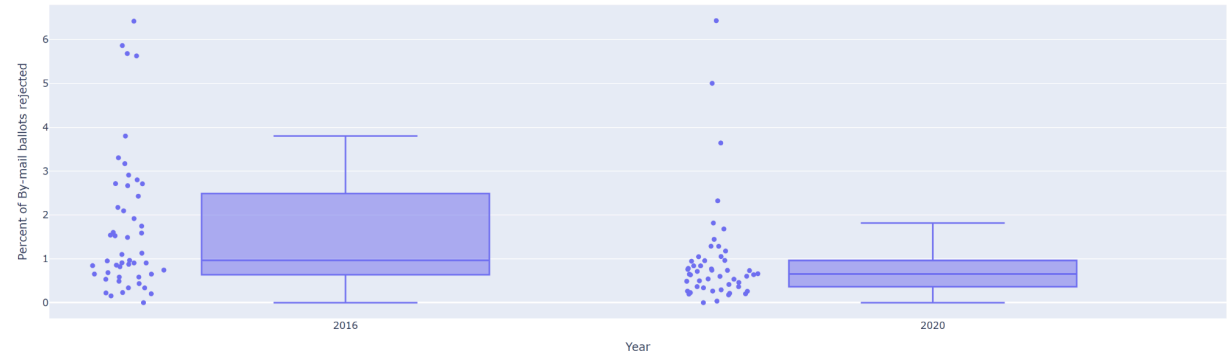
The first part of the dataset we looked at were the reasons for mail-in ballot rejection across the United States. Graphing the sums of this data give the following representations:



These charts show the reasons for rejections of mail-in ballots in 2016 and 2020, both in real and relative terms. The two representations are both important because they demonstrate different results. The relative pie chart shows two major findings. First, many of the less common reasons for mail-in ballot rejection (on the bottom-right of the charts) became relatively more common in 2020, and the number of voters that missed the deadline decreased. In 2020, many more people voted by mail due to the COVID-19 pandemic. This makes sense that voters missing the deadline decreased, since there was a lot more awareness on deadlines (such as when ballots had to be postmarked or when they had to arrive, depending on state).

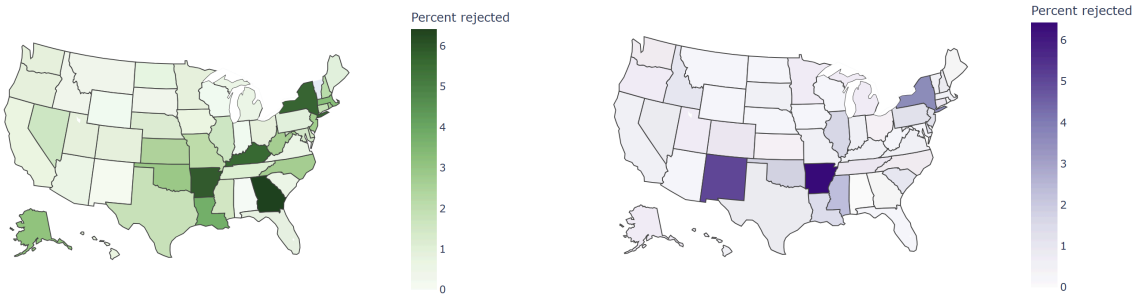
The bar chart, which shows the numbers in absolute terms, gives other information. First, it shows the large increase in rejected ballots for “non-matching signatures”. This error is given when the election official believes that the signature on the ballot does not match the voter’s signature on the record. Though this did not greatly increase in relative terms (only a 3.5% increase, this number almost doubled. Additionally, the increase in the number of ballots rejected for people that have already voted, though also visible on the pie chart, is even more jarring.

The second part of our data analysis was looking at the rate of mail-in ballot rejection by state. We wanted to answer the question: “Are the rates of mail-in ballot rejection different between 2016 and 2020?” To answer this, we ran a one-way ANOVA test with year as our independent variable. For n=49 (Alabama’s data is incomplete), F=6.90, p=0.010, which indicates that the sample means are different. Visualizations support this finding:



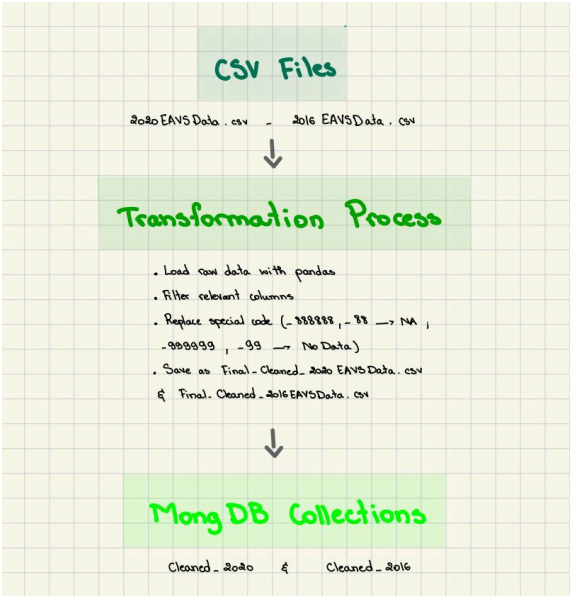
Percent of Mail-In Ballots Rejected in 2016

Percent of Mail-In Ballots Rejected in 2020



ETL Pipeline and Cloud Storage

We chose MongoDB as the storage solution for this project because it aligns well with the nature of the data and my existing knowledge. MongoDB provides flexibility in handling the semi-structured data from the Election Administration and Voting Survey (EAVS). Additionally, I was already familiar with MongoDB, which made it the most comfortable and efficient choice for storing and managing the cleaned datasets. The flowchart for the ETL is shown below:



Appendix: Questions Asked in 2016 and 2020 Survey

2016 Question	2020 Question	Question Description
C1a	C1a	TOTAL domestic by-mail ballots transmitted
C1b	C1b	Returned by Voters
C1c	C1c	Returned as Undeliverable
C1d	C1d	Surrendered, Spoiled or Replaced Ballots (Voided)
-	C1e	By-mail voters who voted Provisionally
C1e	C1f	Status Unknown
C2	-	Does the state have a permanent absentee voter list
C3a	C2a	Total permanent absentee voter list ballot recipients
C4a	C3a	Total by-mail ballots counted
C4b	C4a	Total by-mail ballots rejected
C5a	C4b	Ballots missing deadline
C5b	C4c	No voter signature
C5c	C4d	No witness signature
C5d	C4e	Non-matching signature
C5e	C4f	No election official signature
C5f	C4g	Ballot in unofficial envelope
C5g	C4h	Ballot missing from envelope
C5h	C4i	Envelope not sealed
C5i	C4j	No resident address on envelope
C5j	C4k	Multiple ballots in one envelope
C5k	C4l	Voter deceased
C5l	C4m	Voter already voted in person
C5m	C4n	First-time voter without proper identification
C5n	C4o	No ballot application on record