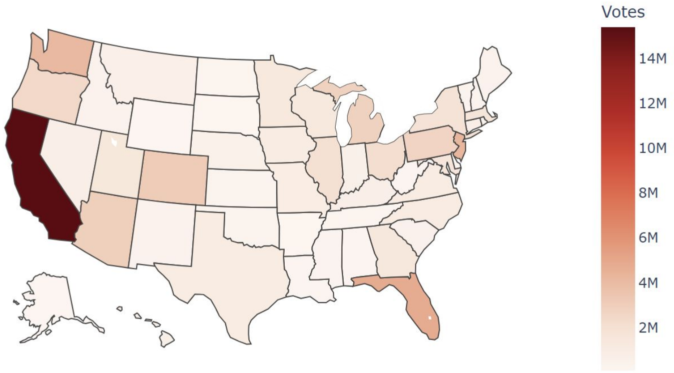
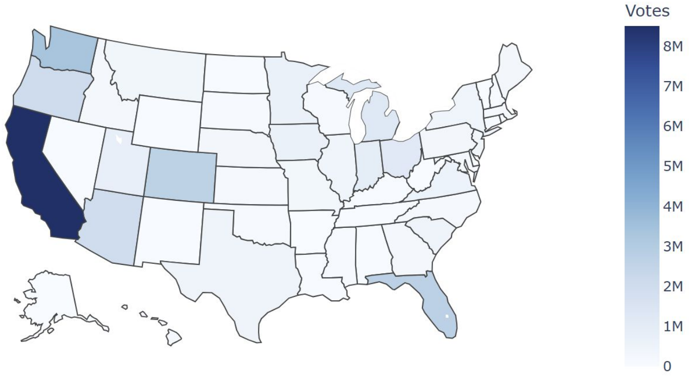
DS 2002 Final Project: Mail-In Ballot Data Analysis

Balkees Rekik, Connor Rose, Alexander Church

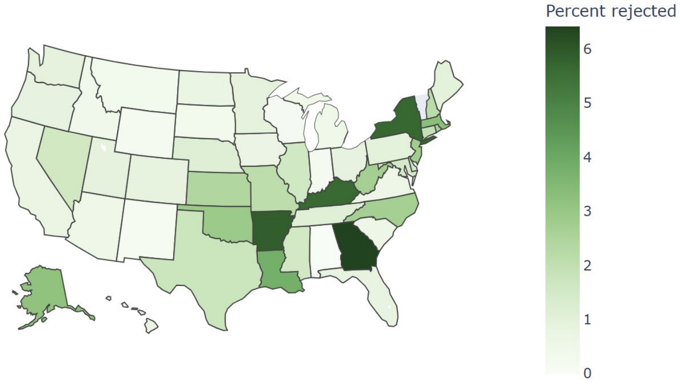
Mail-In Ballots Submitted by Voters in 2020



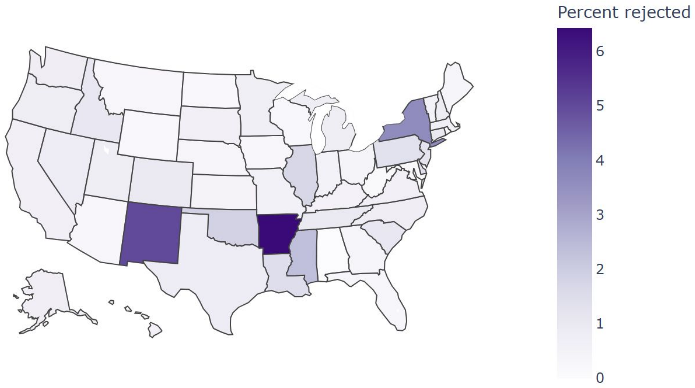
Mail-In Ballots Submitted by Voters in 2016

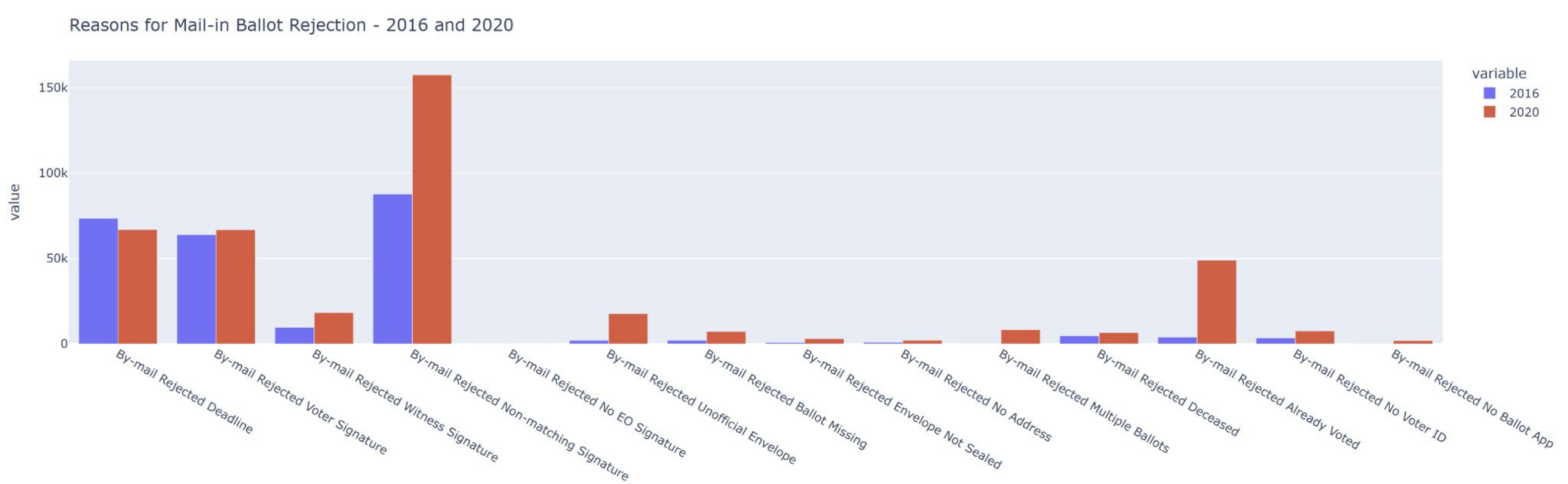


Percent of Mail-In Ballots Rejected in 2016

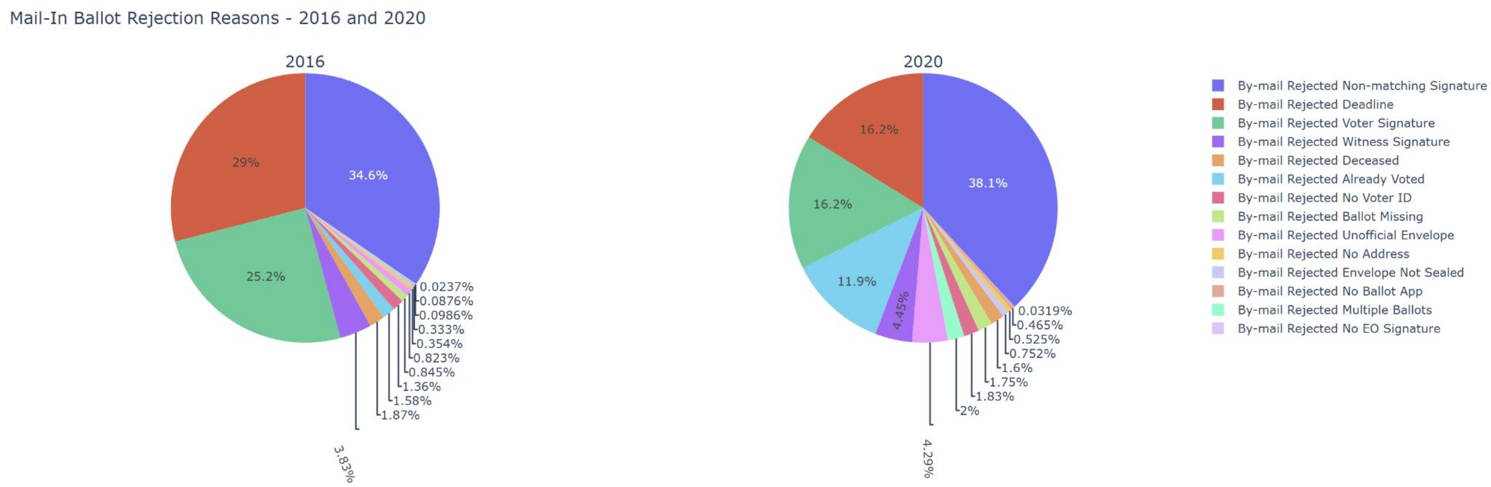


Percent of Mail-In Ballots Rejected in 2020





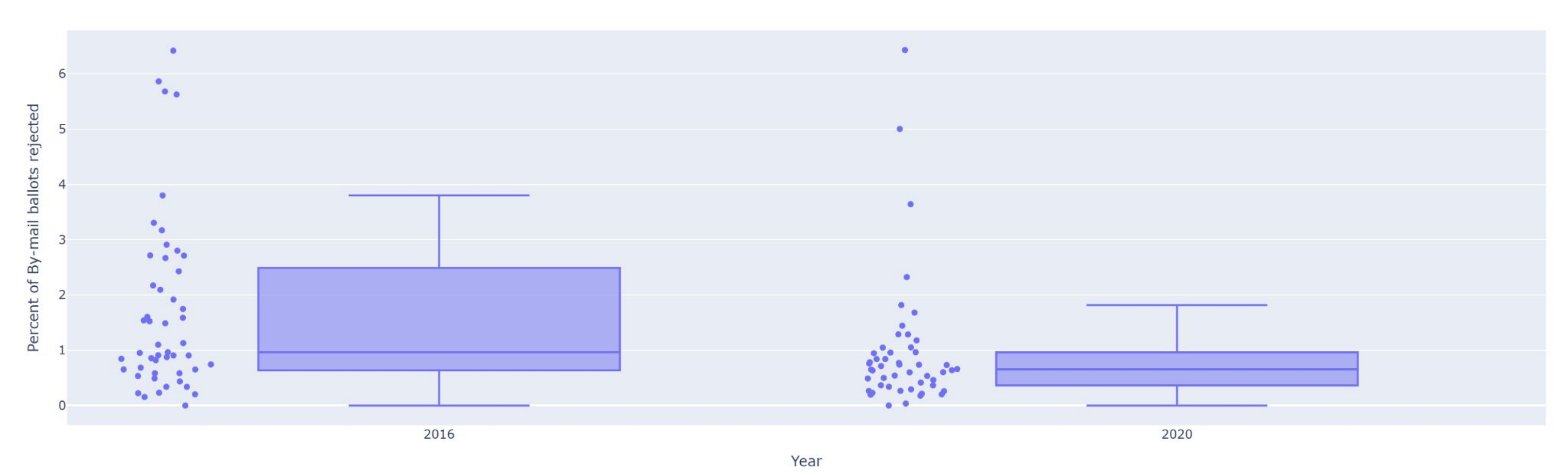
Reasons for Mail-in Ballot Rejection - 2016 and 2020



Mail-In Ballot Rejection Reasons - 2016 and 2020

Are the rates of mail-in ballot rejection different between 2016 and 2020?

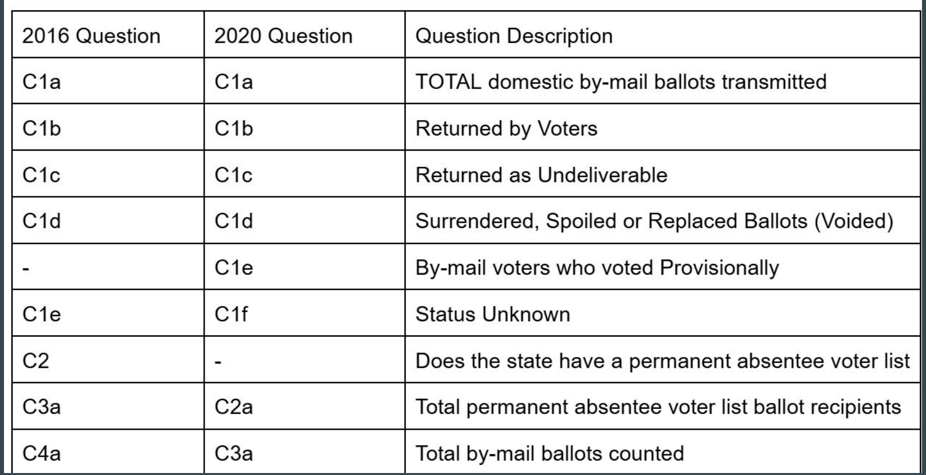
To answer this, we run a one-way ANOVA test with year as our independent variable. For n=49 (Alabama's data is incomplete), F=6.90, p=0.010, which indicates that the sample means are different. Visualizations support this finding.



Main Challenges and Solutions - Part 1

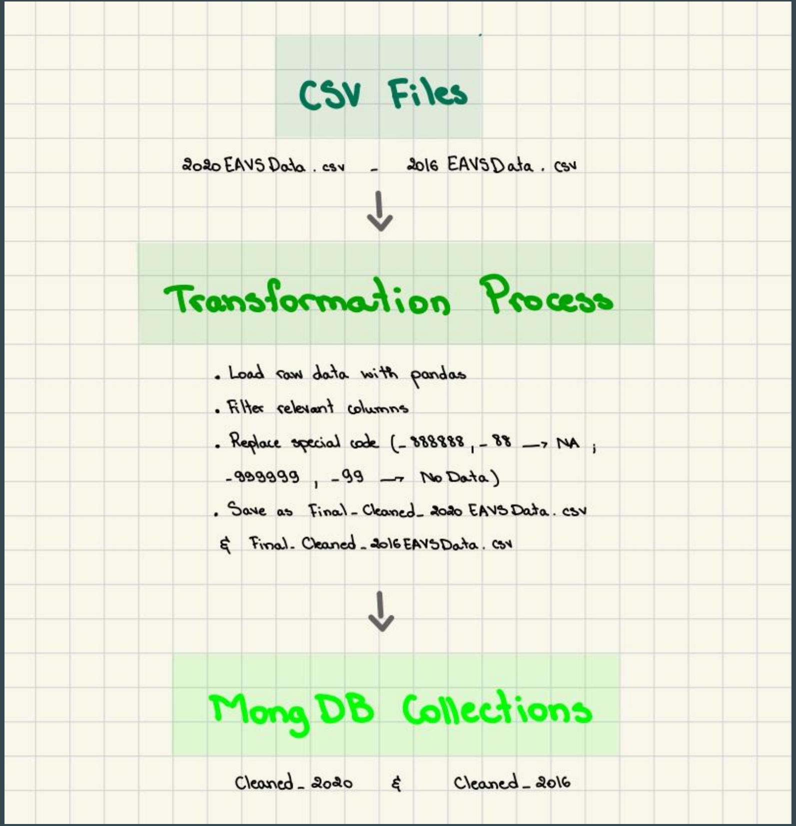
Selection

We knew we wanted to work with the EAVS, and look at the trends of mail-in voting as it changed in the election post-Covid. However, 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.



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

Main Challenges and Solutions - Part 2


CSV Files
2020 EAVS Data.csv - 2016 EAVS Data.csv
Transformation Process
• Load raw data with pandas
• Filter relevant columns
• Replace special code (-888888, -88 → NA ;
-999999 , -99 → No Data)
• Save as Final_Cleaned_2020 EAVS Data.csv
& Final_Cleaned_2016 EAVS Data.csv
Mong DB Collections
Cleaned_2020 & Cleaned_2016

ETL

The ETL process was essential for preparing the EAVS datasets for analysis. Data was first extracted from raw CSV files for 2016 and 2020. During the transformation step, unnecessary columns were removed, special codes were replaced with meaningful values, and the data was cleaned to ensure consistency. Finally, the cleaned datasets were loaded into MongoDB to enable seamless storage and retrieval for analysis.

Main Challenges and Solutions - Part 3

Cloud Storage

Using Google BigQuery's native support in python was difficult to find documentation for. Instead, this was resolved by finding the to_gbq method for pandas_gbq, which was a much more straightforward.

Analysis

Often times, it is difficult to decide what to look for in order to analyze trends in data. EDA was very helpful for this, as it showed the very large variations of rejection rates between states. It also showed that the difference in rejection reasons was not particularly significant year-over-year. Additionally, some states (particularly Alabama) do not have complete data, which requires it to be neglected.