Project Proposal

Milica Cvetkovic, Keyu Hu , Steven Min (nmin), Xintong Shi(xshi242)

Group 11

November 14, 2020

1 Introduction

For the project we chose the Open Election: USA dataset from Kaggle. The dataset contains the results from the previous US elections, with the last available data being from 2016 presidential elections. The data was collected by the OpenElections initiative whose goal is the create free collection of election data for the US. The dataset size is 28GB, and as of right now we are unsure of if the dataset needs to be preprocessed.

2 Statistical Questions

Since we will be working with a diverse dataset that will be giving us variety of options of which direction to take, our proposed statistical questions might require some pivoting, depending on the insights we gather from the preprocessing. Unfortunately, this specific dataset lack some demographics and population information, so we will try to experiment and bind another dataset to it that will provide this information. Note, this is very experimental, and we are unsure that we can do this. In this experimental attempt we will look into questions such as what does a typical millennial white voter lean towards. On the other hand, on a less experimental side, we will look into data by county and see how each county tends to vote. Another interesting thing we could investigate is the voter turnout for each district, and whether more conservative or democratic voters go out to vote.

3 Code

The following is a preview of loading a subset of data.

	state_legislative_district <int>	precinct <fctr>	office <fctr>	district <fctr>	party <fctr>	candidate <fctr>	votes <int>
1	1	01-446 Aurora	U.S. Senate		DEM	Begich, Mark	413
2	1	01-446 Aurora	U.S. Senate		LIB	Fish, Mark S.	30
3	1	01-446 Aurora	U.S. Senate		*NA*	Gianoutsos, Ted	22
4	1	01-446 Aurora	U.S. Senate		REP	Sullivan, Dan	434
5	1	01-446 Aurora	U.S. Senate		NP	Write-in 70	2
6	1	01-446 Aurora	U.S. House	1	DEM	Dunbar, Forrest	336

6 rows

Figure 1: Preview of the dataset

4 Variables

The data has 7 features: county, precinct, office, district, party, candidate and votes. The dataset has already been heavily preprocessed and we are hoping that this will make our job a little bit easier, to avoid heavy work in the beginning.

5 Statistical Methods

We plan to start with Regression models including Logistic Regression. Then we plan to try out Random Forest, Neural Networks and other advanced models.

6 Tools

For the processing data we will be using R, since that was our focus in STAT 605 so far, and we gained skills during this semester in R and bash. right now, we can't be more specific, except that we will be probably using tidyverse library to do our work. Since our data is not heavily loaded with information, it could be interesting to make some visual explorations of the data that we will be working on. Finally, we will be using CHTC resources in order to load and process such a large dataset.