**Milestone 3**

**EDA and Revised Project Statement**

Voter Turnout Project Overview

Our goal in this project is to build a classification model to predict voter turnout in the general election using variables available in public voting records. Specifically, we want draw conclusions about and better understand the voting behavior of minorities.

Data Description

Given that Florida is a swing state, with large minority populations and differences in voting behavior between counties, it is the ideal state to examine in this project. The Florida dataset we are utilizing contains information of over 8 million registered voters and consists of over 250 million data points.

The variable we are trying to predict is participation in the general election. The primary variable we are looking at is voter ethnicity. Because ethnicity is not available in public datasets, we will first need to classify ethnicities by using only the names of voters, before looking at what other factors impact voter turnout.

Other available and possibly relevant predictors we will look at include age, gender, party affiliation, absentee vs. in person ballots, years since receiving a voter ID, and previous voting behavior. We can also see how these variables differ between ethnicities.

Before EDA, we needed to perform initial processing for the data to be usable. This consisted of importing all text files as Pandas dataframes, cleaning the data, and merging the dataframes by voter ID, so that voter registration information corresponded with voting history.

EDA Visualizations

The first step of our data exploration was to create a baseline model for the ethnicity classifier. An open source classifier (https://github.com/kitofans/ethnicityguesser/blob/master/NLTKMaxentEthnicityClassifier.py), yielded these results:



Bar chart of age vs. voting in the general election

Voting behavior by gender

Revised Project Statement

This project will consist of multiple parts. Firstly, we aim to create an effective ethnicity classifier using voters’ names. Though the results of our baseline model are reasonably accurate, especially given how many ethnic groups are being used, there is room for improvement. We hope to tune the model to give stronger results, so for instance we might exclude names that the model does not strongly predict.

After creating a reliable ethnicity classifier, we will compare and contrast voting behavior between ethnic groups. We will then add in additional variables into the prediction, which could be factors such as age, gender, party affiliation, absentee vs. in person ballots, years since receiving a voter ID, and previous voting behavior.

From this, we can determine which variables are most useful in predicting voter turnout, as well as how these factors differ between ethnicities. After this is finished, we will evaluate our prediction accuracy on our test set.

These are our primary goals, however we may discover additional directions for analysis and they may change slightly as we continue our project.