Analysis

Our project has two main sections of data sets in it. The first and primary one containing candidate information and previous presidential election data and the second containing twitter data. We believed the second needed more time because the first dataset contained straightforward facts while the twitter dataset is more susceptible to different perspectives and viewpoints.

TWITTER:

In this project we are analyzing twitter data of key presidential candidates and other key supporters.We gathered the data from kaggle datasets. Data is mainly divided into 3 sub categories. Tweets made by key personnel.Twitter profiles of the two candidates(all info including followers, following, # of tweets,etc.).The final category involves graphs for visualization purposes.A problem with twitter data is the fact that it is huge. We are using google drive and with it comes the problem of storage. To combat this we are only using 4 twitter data sets. The datasets of Donald J. Trump, Joe Biden, Kamala Harris, and Mike Pence. We also downloaded these data sets to use them locally.There are mainly 3 types of formats used in everyday twitter use: images, text, and video. Only text will be used in this project because the others are not useful for the purpose of the experiment.

Our project mostly uses twitter data as support to the primary dataset.It is there to strengthen the already predicted result. The reason why we cannot make the twitter data set our primary data set is because the data(tweets) are mostly opinion based with only some exceptions. So we cannot predict with the twitter data,however, it can be used to show public support which will be vital in supporting the prediction derived from the primary data set.So, we found many twitter data sets on keggle and used certain parts from each to make our final four. The difference between the background sets and our final four datasets is the fact that their primary dataset was the twitter data while we used twitter data as our secondary dataset. We realized that twitter data is best used as secondary data that supports the primary dataset, which is more fact based.We can use three of the four OSoMe tools available: trends, networking, and maps. Trends and networking can be combined to find a group that involves every user that is taking part in the elections in some way. Mapping can show these users and their location. Giving us the area based result that we seek. However this method is already a part of our project .Due to the fact that all this data is in keggle in a wider array. Which gave us the option to condense into four large data sets.

Candidate Info and Previous Data:

The intent of this section is to look at candidate information from previous years and compare with present candidate information. This part was pretty straightforward when it came to planning and execution. We ran into no problems and it was easier for us to create visual representations like graphs.