THE LEGION

# PROJECT PROPOSAL

## ELECTION POLLS

## Phase 1: EDA Analysis

Section-002

Jayraj Radadiya-Team leader

Harshil Patel

Rajvi Mehta

Surbhi Patel

Krim Patel

Karan Patel

# INDEX

# 1.INTRODUCTION

Decisions in the United States are held for government authorities at the administrative, state, and neighborhood levels. At the government level, the country's head of express, the president, is chosen by implication by individuals of each state, through an Electoral College. Today, these voters quite often vote with the well-known vote of their state.

The most well-known technique utilized in U.S. decisions is the first-past-the-post (Plurality voting system) framework, where the most noteworthy surveying applicant wins the election. Under this system, a candidate needs only a majority of the votes to win, instead of a direct majority. Some may use a two-round system, where if no one gets the required votes then there is competition between the two who gets the most votes.

Citizens rank the competitors arranged by inclination as opposed to deciding in favor of a solitary applicant. Assuming a competitor gets the greater part of the votes cast, that up-and-comer wins. In any case, the competitor with the least votes is wiped out. Voting forms appointed to the disposed of the competitor are related and allocated to those of the excess applicants who rank next arranged by inclination on each voting form. All individuals from the government council, the Congress, are straightforwardly chosen by individuals of each state.

# 2. PROBLEM STATEMENT

Pollster has created vivid type of polls to collect vote through different methodology namely, Online, IVR/Online, Live Phone, IVR/Online/Text, Live Phone/Online.

Usage of different method to collect vote is convenience for Candidate to use preferable polls to give vote. This makes more votes for Politician to elect.

There is a unique id for certain variable to keep data secure like Question id, Polls id, Pollster id, Pollster rating id, Race id, Candidate id and Politician id.

Many sponsors participate to assist pollster to conduct election polls and each sponsor has unique id. list of sponsors participated such as Economist, Politico, Winning the Issues, John Bolton Super PAC, and Reuters. Each sponsor selects their own state to help the candidate to have their chosen politician get elected.

- Candidate-specific information, such as State, methodology used, and so on, will be included in the polls.

- To avoid identity theft and duplication, the first upload must be done under the supervision of a trustworthy pollster.

- The votes recorded on the identity details MUST be secret, and the counting of votes is done from the polls itself once the votes have been read.

## Target Audience

Mostly government political parties (both ruling and opposition) will make use of historical data to forecast the future poll and media will telecast by analyzing the election poll.

**Political campaigns** have information about Americans and how they utilize it to design their plans. In the United States, political campaigns employ data on more than 200 million eligible voters to influence their strategy and tactics.

The two major political parties in the United States strive to utilize the most precise statistics to target voters in several ways. Republicans and Democrats collaborate with data businesses to develop national voter databases, gathering data from a variety of sources to produce complete voter profiles with hundreds of data points and algorithms that predict people's attitudes toward topics and candidates.

This information may be used by political campaigns to assist them decide who to reach out to, how to contact them, and how they might react to various messages.

### Political campaigns collect data in four diverse ways mentioned below:

### National database:

The data utilized by the campaigns comes from a variety of public voter files, which are stacked with hundreds of data points purchased from commercial vendors and updated on a regular basis by organizations like TargetSmart, which works for Democrats, and Data Trust, which works for Republicans.

### Layering data:

To develop detailed voter profiles, firms layer data from a variety of sources onto the national database.

### Predictive models:

Models that anticipate people's attitudes on a candidate or subject are built using voter data and opinion polls.

### Data-informed campaigns:

Campaigns get access to the voter database and models, which they use to help decide whom to target in their outreach efforts and how to reach them.

## Motivation

As Americans, a person must cultivate patience, accept the uncertainties of the present, and wait for election results and reliable polling data to reveal what voters decided and why. Pre-election polls are an effective attempt to get into people's brains. They try to figure out why Americans have the attitudes, beliefs, and worries that they have.

In the past, polls have been quite accurate in predicting popular sentiments, but not so much in predicting public conduct. It is easy to dismiss the decision to vote or not vote in an election as insignificant. It is a minor decision that we make occasionally and that has little or no societal or personal consequences. Because the likelihood of a single vote deciding an election's outcome is negligible, whether one votes will determine which party will win the election. However, the decision to vote or abstain is not one that should be taken lightly. With the help of the data, we can gauge the method highly preferred by people.

# 3. PROPOSAL

Opinion polls are often aimed to depict a population's views by asking a series of questions and then extrapolating generalities in ratios or within confidence intervals. The term "pollster" refers to someone who conducts polls. To educate people about the voting process. To inform residents about the relevance of Electors Photo Identity Cards (EPIC) and how to utilize them in various government initiatives, such as passport processing and bank account opening.

Encourage citizens to participate in democracy by enrolling in electoral rolls and voting at election time. Examine how the United States' election administration contributes to the functioning of democracy. Define the role of polling stations and local precincts in the electoral process. Describe the numerous methods for voters to cast ballots. Describe the role of voting machines in the electoral process.

## Data Description

This is the dataset of election polls which contains 12 CSV files. Each CSV files provide information about various polls such as generic ballot polls, governor polls, house polls, president polls, senate polls, president approval polls, and president primary polls.

The data can be compared easily because this dataset contains both current polls and historical polls files except president approval polls and president primary polls as it has only current polls files. There is numeric as well as categorical data in all the files.

Current polls files include data from the most recent election whereas historical files contain data prior to the most recent election.

President approval polls contain 28 variables, generic ballot polls contain 32 variables, president primary polls contain 34 variables, and governor polls, house polls, president polls and senate polls contain 38 variables in both current and historical files.

The data is imported from Kaggle website.

Reference link: www.kaggle.com/gmkeshav/election-polls-datasets

Data Source: www.usa.gov/government-works/

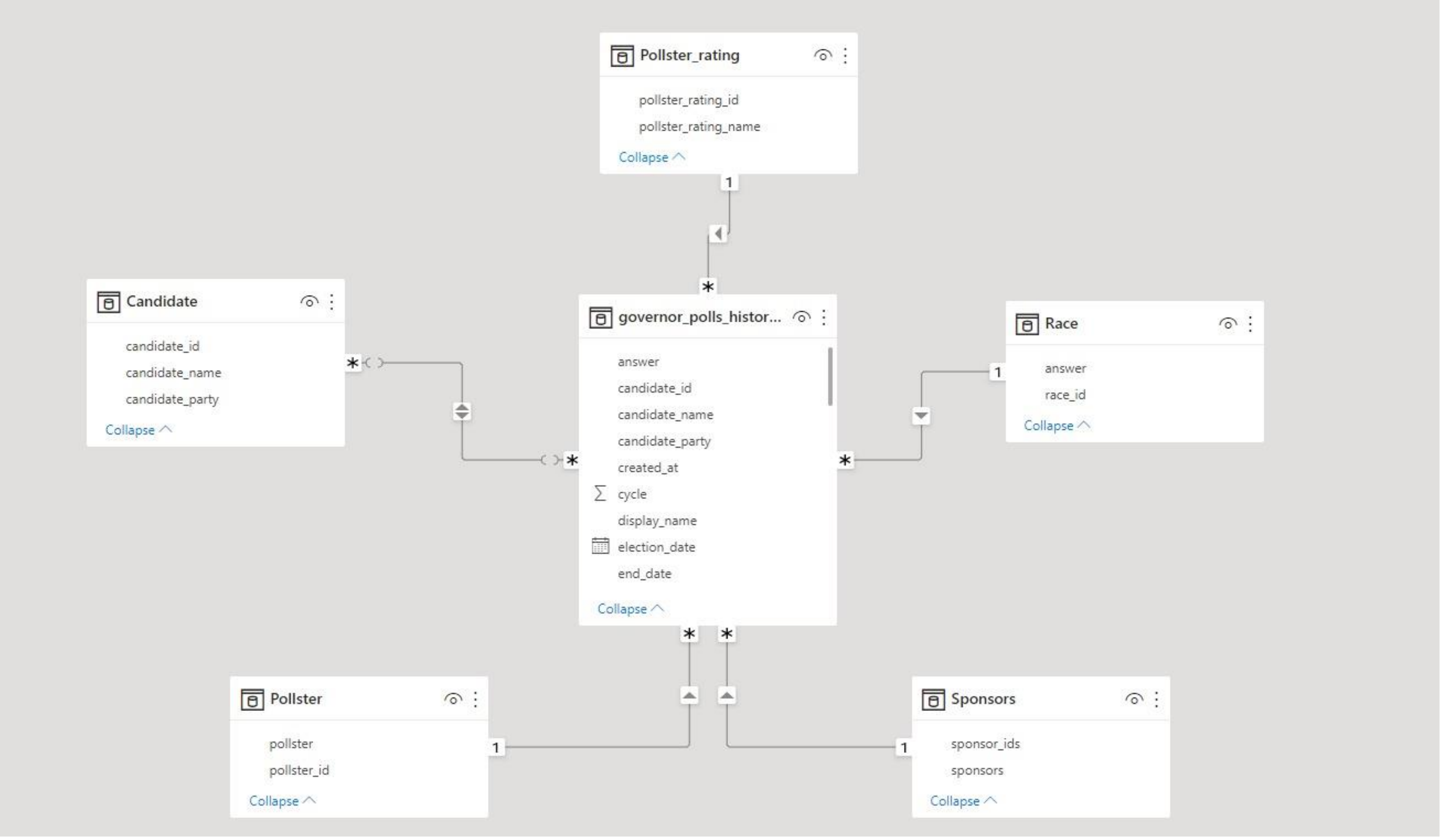
| Variables | Data Type |
|---|---|
| State | Categorical |
| Pollster | Categorical |
| Sponsors | Categorical |
| Display name | Categorical |
| Pollster rating name | Categorical |
| Fte grade | Categorical |
| Population | Categorical |
| Population full | Categorical |
| Methodology | Categorical |
| Office type | Categorical |
| Seat name | Categorical |
| Sponsor candidate | Categorical |
| Partisan | Categorical |
| url | Categorical |
| Stage | Categorical |
| Answer | Categorical |
| Candidate name | Categorical |
| Candidate party | Categorical |
| Party | Categorical |
| Cycle | Numerical |
| Sample size | Numerical |
| Seat number | Numerical |
| Pct | Numerical |
| Dem | Numerical |
| Rep | Numerical |
| Source | Numerical |
| Yes | Numerical |
| No | Numerical |
| Alternative answer | Numerical |
| Sponsor IDs | Unique |
| Question ID | Unique |
| Polls ID | Unique |
| Pollster ID | Unique |
| Pollster rating ID | Unique |
| Race Id | Unique |
| Candidate Id | Unique |
| Politician ID | Unique |
| Internal | Logical |
| Nationwide batch | Logical |
| Ranked choice relocated | Logical |
| Created at | Logical |
| Tracking | Logical |
| Start date | Date |

| | |
|---|---|
| **End date** | Date |
| **Election date** | Date |

# 4. EDA ANALYSIS

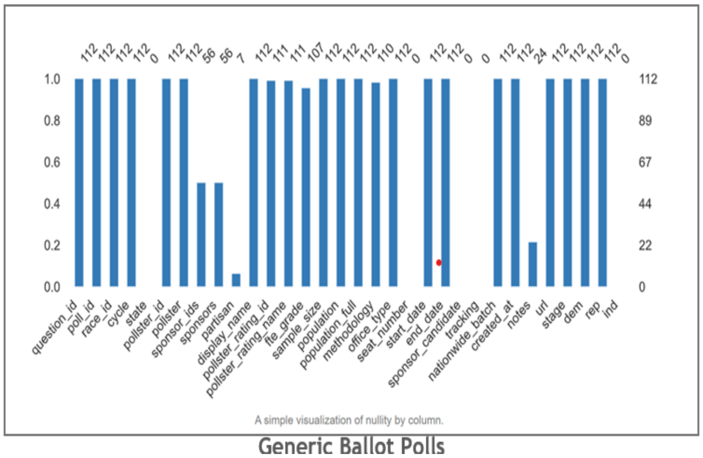
From this given dataset we can be able to analyze the data regarding the Recent Election of US. There are many different groped data such as data regarding the pollsters, Sponsors, and candidate.

In this data, we are going to analyze what was the flow of election. which party won the elections. We can also determine which candidates have won how many votes and through which method and from where.

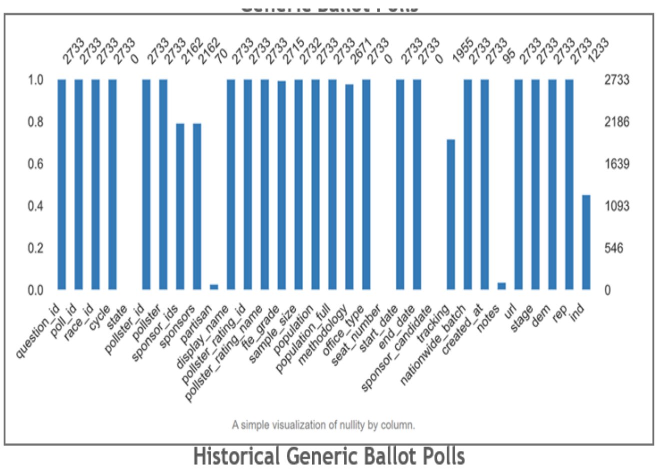


## Invalid and missing values

The csv files do contain missing values in a few fields. Most of the fields have 100% valid records, but there are some fields from sponsor id, sponsor, fte_grade or notes that have some missing values, and state, seat number, sponsor candidate, tracking with 100% missing values.



Generic Ballot Polls

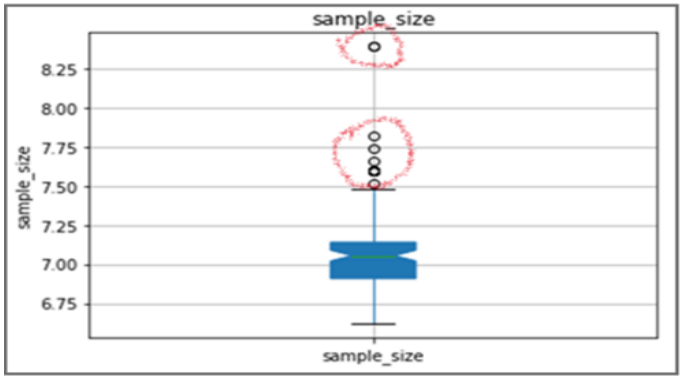


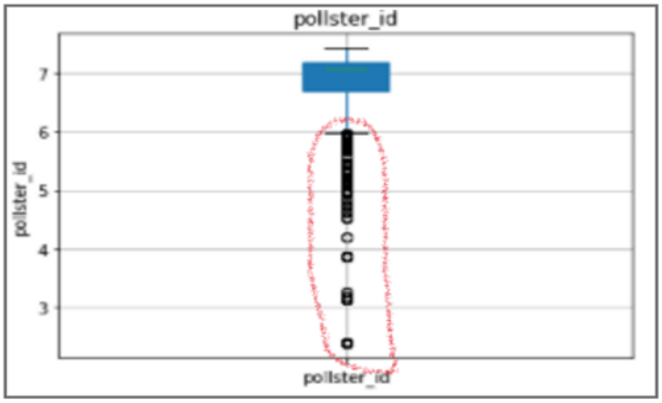
Historical Generic Ballot Polls

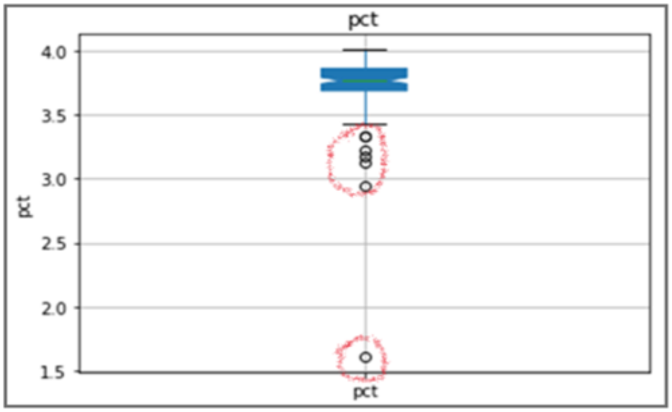
## Outliers

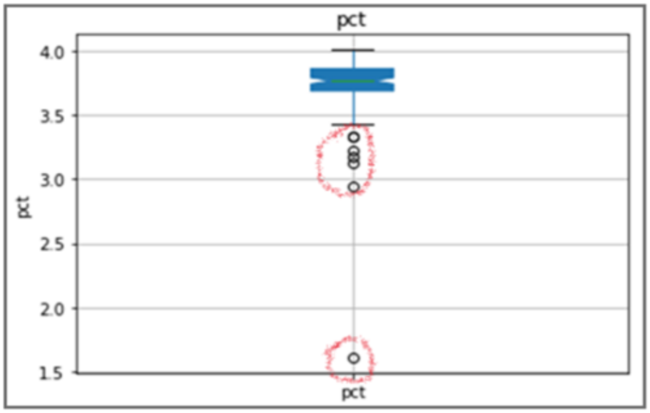
Identification of Outlier values has been done on all the column of all csv files. Mostly, fields do not have any outlier but some of them such as **sample_size, dem, pct, candidate_id, pollster_id, pollster_rating_id** and few more showed an outlier.

Below are some boxplot graphs which were produced during the Analysis of dataset, in these graphs' dots are located at outside of the lower bound and upper bound of boxplot those are the outliers of that fields.









## Segmentation

The data is segmented into the group in this election poll dataset. In some fields such as pollster, state, sponsors data divided in groups.

Let’s take pollster for an example, pollster column has same name of pollster like Rasmussen (pulse opinion research), YouGov, Morning Consult, IBD/TIPP, Ipsos, Harris Poll, NBC, RMG Research, AP-NORC, Saint Leo U, which is grouped with pollster_id. Because of the same pollsters are making some polls in year

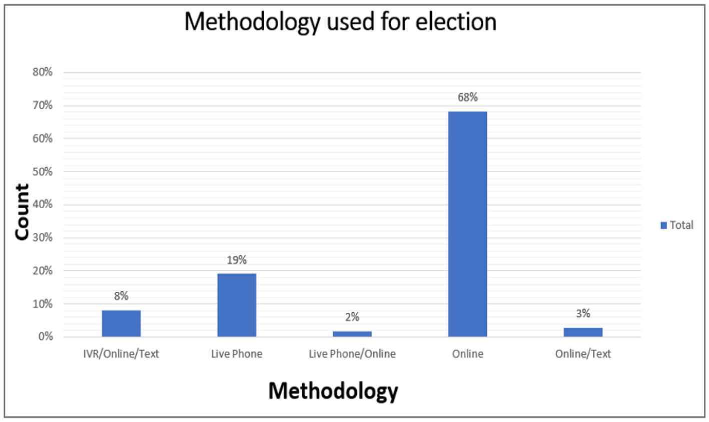
In sponsors column also sponsor’s names are same likewise Wayne Allyn Root, Economist, Reuters, Harvard, The Canadian Press, Politico, The ANTIFA, Yahoo News, CNN, CNBC-All America Economic Survey. This filed is in group with sponsor_id, due to these all sponsors are giving a sponsorship for polls.

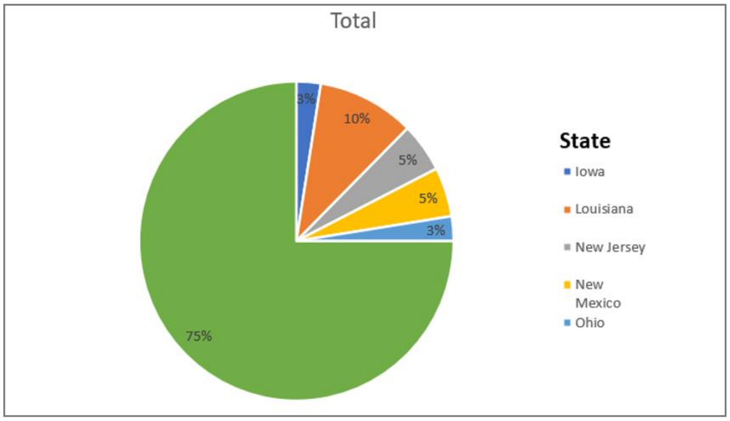
Lastly, state field contains all US states is also in group with question_id, because some polls has taken from same state.

## Data Imbalance

Most of the data in data elements are not imbalance in our dataset, but some missing values still present in few columns like methodology, state, partisan, etc.

We use a bar graph and pie chart to display the imbalance of data. From the bar graph it’s clear there are different types of method used for election, but online method is most common as compared to others methodology. The pie chart depicts the total number of states participate in election. As per the graph, we observed that there are 75% of missing data present in state column which is a good example of imbalance data.





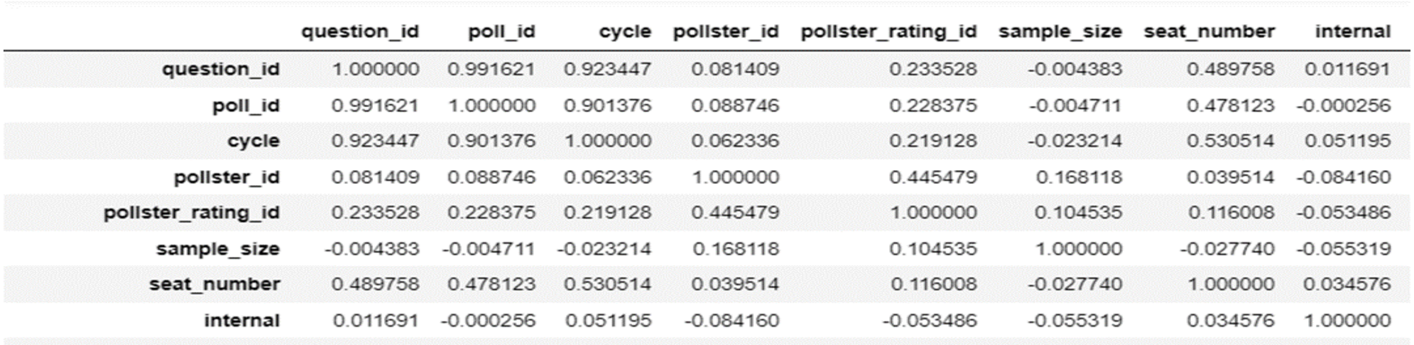
## Correlation

Many variables are highly correlated with each other by finding a correlation of senate_polls_historical dataset variables it concludes that most of categorical values like states, URL and notes are correlated with unique values such as Question ID, Polls ID, Pollster ID, Pollster rating ID, Candidate Id, Politician ID, Race Id. All these ID variables are also highly correlated with other because of unique values.

By analyzing it can be visible that unique values are mostly correlated with each other's than numerical variable. There are very few numerical variables like cycle, sample size, pct and seat number are present but there is minimal correlation found between them.

**Table of senate_polls_historical shows correlation:**

We can discover statistical links between variables by using the correlation function. The correlation of one variable to another is shown numerically in the table below.

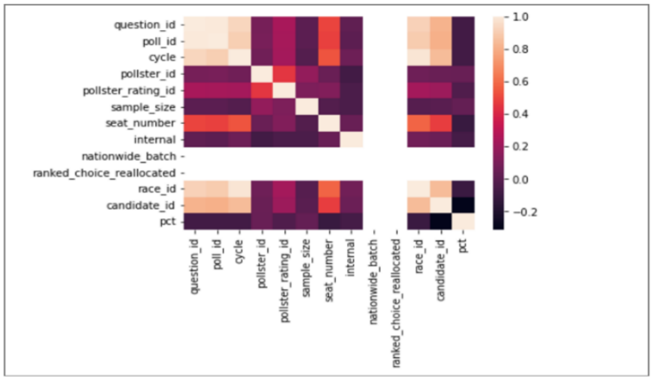


| | question_id | poll_id | cycle | pollster_id | pollster_rating_id | sample_size | seat_number | internal |
|---|---|---|---|---|---|---|---|---|
| **question_id** | 1.000000 | 0.991621 | 0.923447 | 0.081409 | 0.233528 | -0.004383 | 0.489758 | 0.011691 |
| **poll_id** | 0.991621 | 1.000000 | 0.901376 | 0.088746 | 0.228375 | -0.004711 | 0.478123 | -0.000256 |
| **cycle** | 0.923447 | 0.901376 | 1.000000 | 0.062336 | 0.219128 | -0.023214 | 0.530514 | 0.051195 |
| **pollster_id** | 0.081409 | 0.088746 | 0.062336 | 1.000000 | 0.445479 | 0.168118 | 0.039514 | -0.084160 |
| **pollster_rating_id** | 0.233528 | 0.228375 | 0.219128 | 0.445479 | 1.000000 | 0.104535 | 0.116008 | -0.053486 |
| **sample_size** | -0.004383 | -0.004711 | -0.023214 | 0.168118 | 0.104535 | 1.000000 | -0.027740 | -0.055319 |
| **seat_number** | 0.489758 | 0.478123 | 0.530514 | 0.039514 | 0.116008 | -0.027740 | 1.000000 | 0.034576 |
| **internal** | 0.011691 | -0.000256 | 0.051195 | -0.084160 | -0.053486 | -0.055319 | 0.034576 | 1.000000 |

**Heatmap visualization of correlation.**

The degree of relatedness between variables is measured by correlation and for the accurate result we find correlation through Pearson's coefficient.

Below mentioned graph indicate Positive correlation (+1) to Negative correlation (-1) and in between Pearson's coefficient correlation "r" (0). If r is near to -1 it denotes there is inverse relation between two variables if it is +1 or near to +1 it indicates perfect positive correlation. If else "r" coefficient correlation is 0 it shows no relation between data points between data points.



*Note: Blank white color shows missing values. (ranked_choice_reallocated and nationwide batch)

## Preliminary Visualization

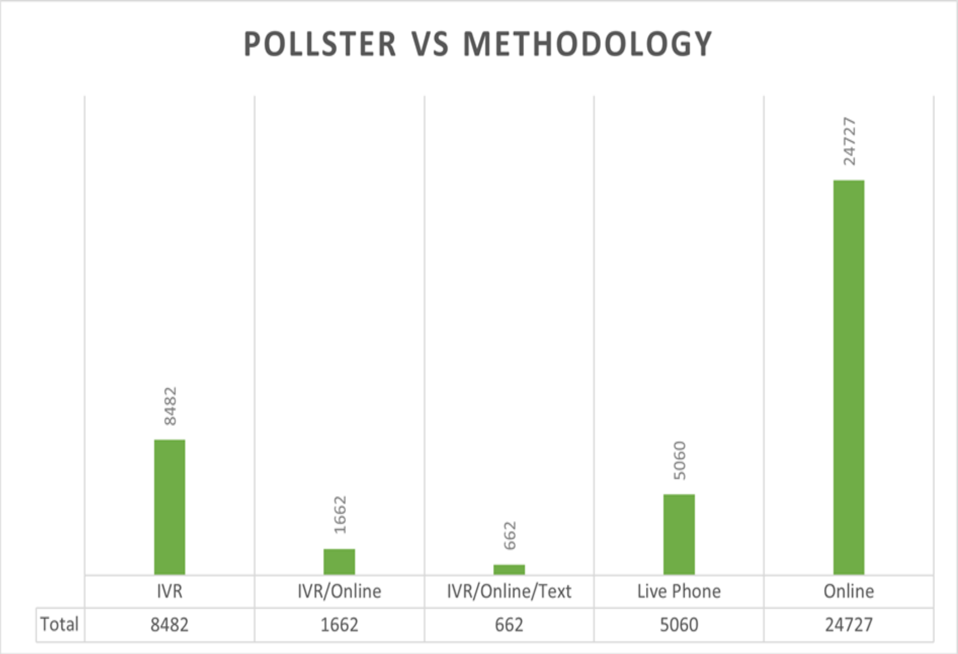
Below are some preliminary visualizations from our data.

1)Pollster VS Methodology
Question. - Which is the most preferred and least preferred Voting method across the people?

2)Candidate Party vs Vote Percentage
Question - Which party is winning?

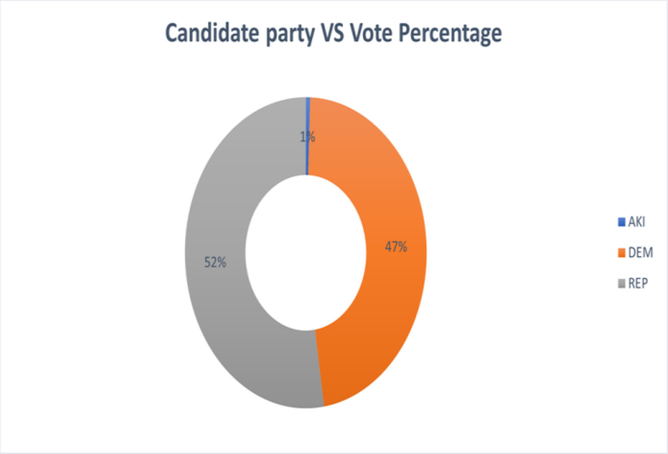
3)Pollster count.
Question - Which banner has the greatest number of pollsters under it?

4)Distribution of pollsters across the USA.
Question - Are the pollster Divided equally within each state?

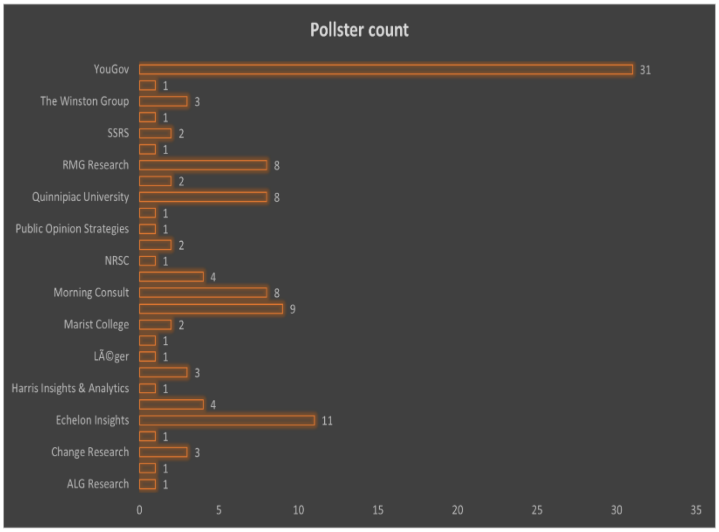


| | IVR | IVR/Online | IVR/Online/Text | Live Phone | Online |
|---|---|---|---|---|---|
| Total | 8482 | 1662 | 662 | 5060 | 24727 |

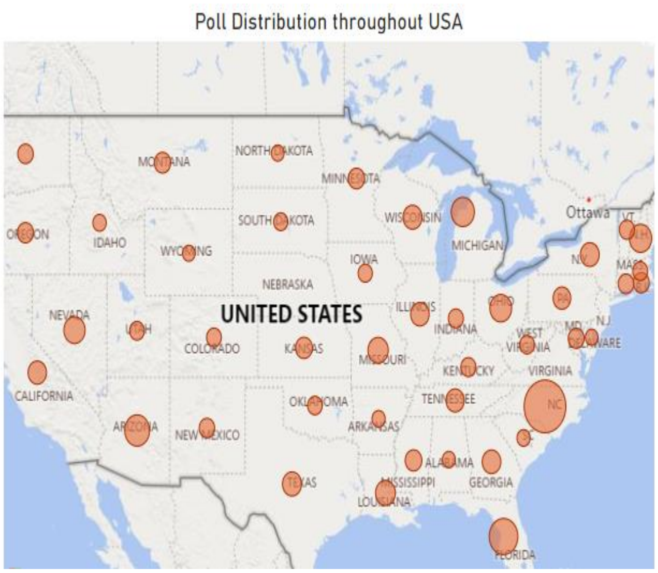
The above graph is about the Method used to vote. As now a days many different methods to vote are available such as Online, IVR, Live Phone,Text etc.



The above plot represents the Distribution of votes among the Parties. With this representation we can clearly identify the winning party.



This graph shows the number of Polling booths under the particular pollster.We can see the Goverment body YouGov is the one with maximum amount of pollster under it.



This graph depicts the distribution of pollsters throughout the USA. We can identify the locations of pollsters statewise here.

## Conclusion

We can conclude that from the above visualization we can clearly see that the most used method to vote is online and least one is IVR/Online/Text. Also the Republic party (REP) wins it with 52% of votes while the Democratic party (DEM) is at 2nd position with 47 % votes. We can even analyze that the Goverment body YouGov is the leader in having the pollster and the 4th graph depicts the distribution of pollsters throughout the USA. We can identify the locations of pollsters statewise here.

## References

https://en.wikipedia.org/wiki/Elections_in_the_United_States#Election_information_on_the_web

https://www.pewresearch.org/fact-tank/2020/10/29/what-we-can-trust-2020-election-polls-to-tell-us/

https://www.ubcpress.ca/asse