PROJECT PROPOSAL

ELECTION POLLS

Phase 1: EDA Analysis

Phase 2: Data Cleaning & Visualization

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.

• 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 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, The method that is highly preferred by people can be identified as well.

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/

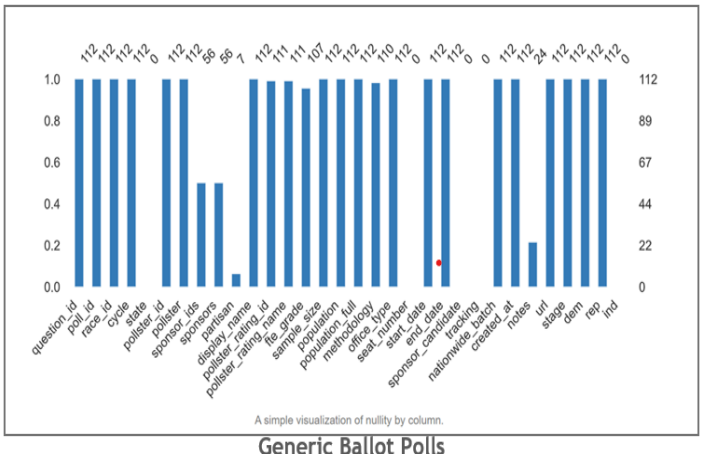
Variables	Data Type
State	Categorical
Pollster	Categorical
Sponsors	Categorical
Display name	Categorical
Pollster rating name	Categorical
Fte grade	Categorical
Population	Categorical
Population full	Categorical
Methodology	Categorical
Office type	Categorical
Seat name	Categorical
Sponsor candidate	Categorical
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

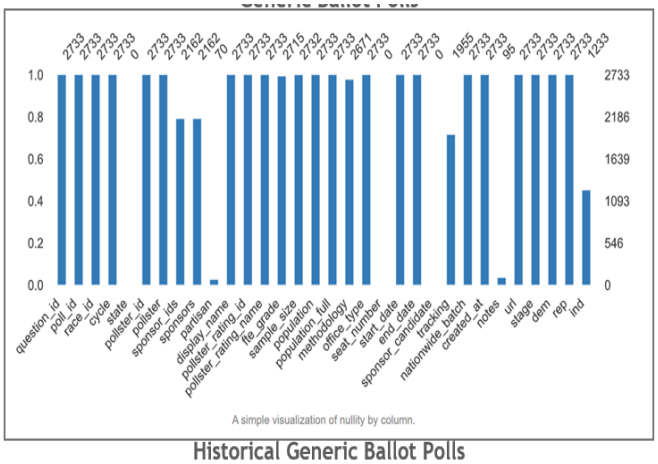
4. EDA ANALYSIS

From this dataset analysis can be done regarding the Recent Election of US. There are many different groped data such as data regarding the pollsters, Sponsors, and candidate.

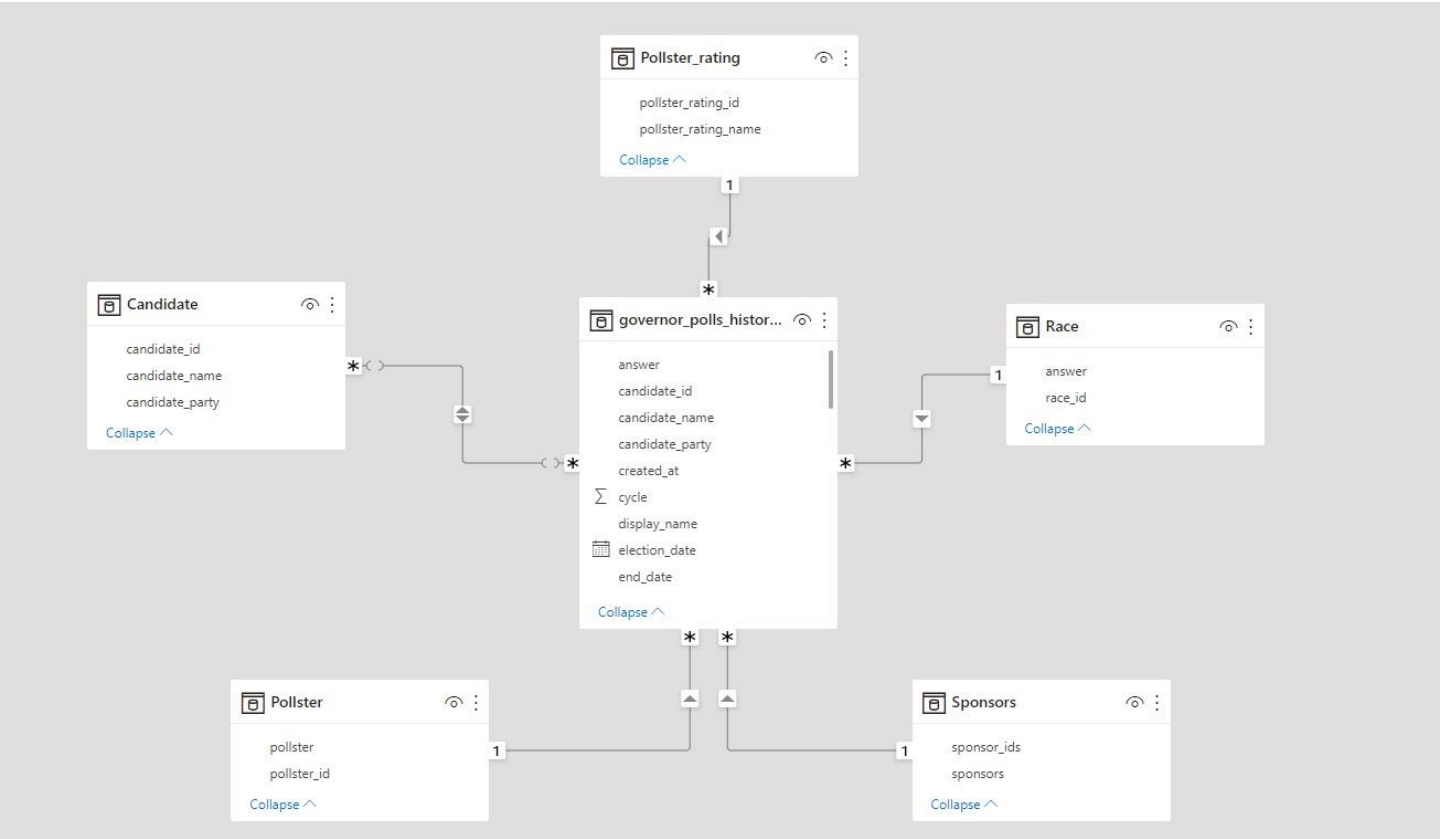
In this data, main analysis is going to be the flow of election. which party won the elections. It can also determine which candidates have won how many votes and through which method and from where.



Generic Ballot Polls



Historical Generic Ballot Polls



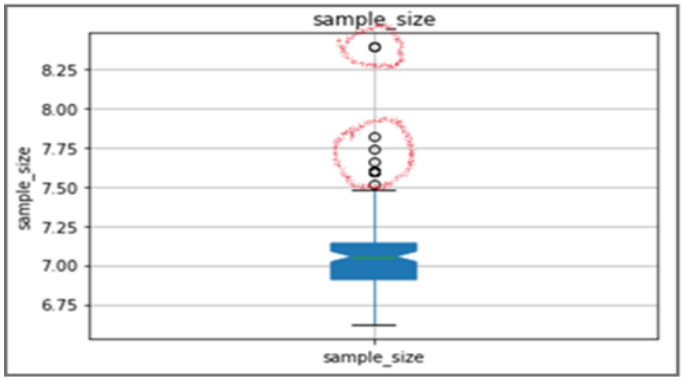
Invalid and missing values

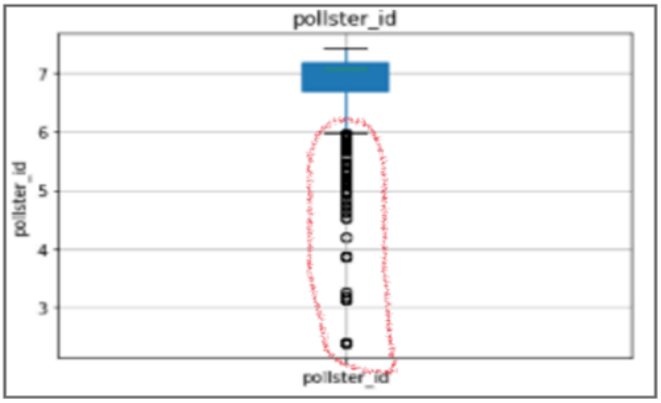
The csv files do contain missing values in a few fields. Most of the fields have 100% valid records, but there are some fields from sponsor id, sponsor, fte_grade or notes that have some missing values, and state, seat number, sponsor candidate, tracking with 100% missing values.

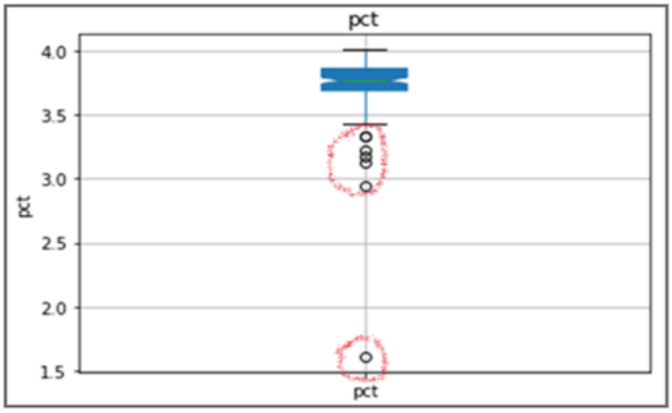
Outliers

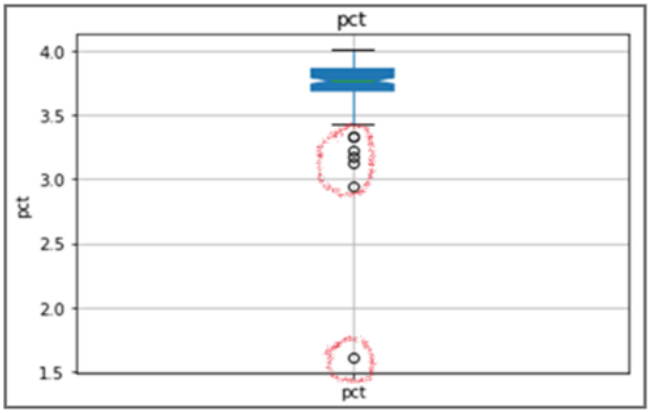
Identification of Outlier values has been done on all the column of all csv files. Mostly, fields do not have any outlier but some of them such as **sample_size, dem, pct, candidate_id, pollster_id, pollster_rating_id** and few more showed an outlier.

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

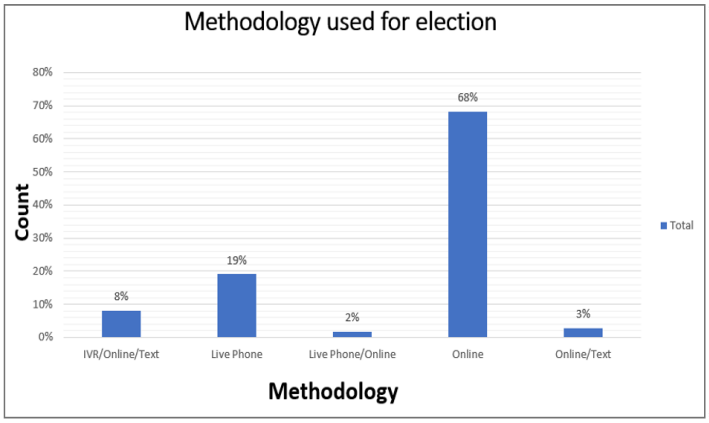
In sponsors column also sponsor's names are same likewise Wayne Allyn Root, Economist, Reuters, Harvard, The Canadian Press, Politico, The ANTIFA, Yahoo News, CNN, CNBC-All America Economic Survey. This filed is in group with sponsor_id, due to these all sponsors are giving a sponsorship for polls.

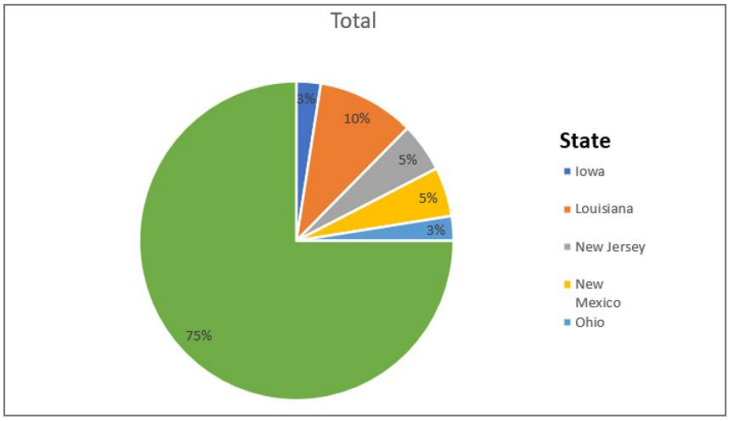
Lastly, state field contains all US states is also in group with question_id, because some polls has taken from same state.

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.

bar graph and pie chart are used 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, it is 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.

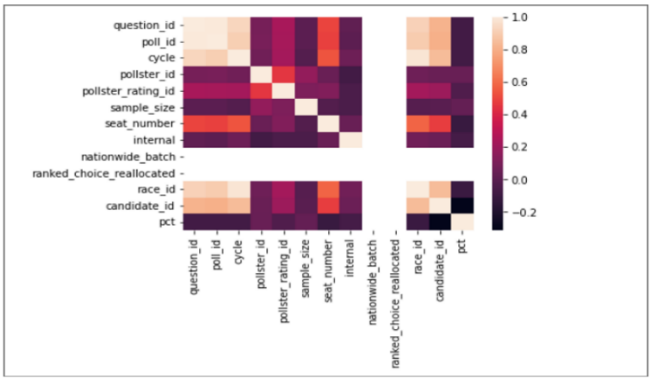
Table of senate_polls_historical shows correlation:

The statistical links between variables by using the correlation function are discoverable. 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 the correlation through Pearson's coefficient will help.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

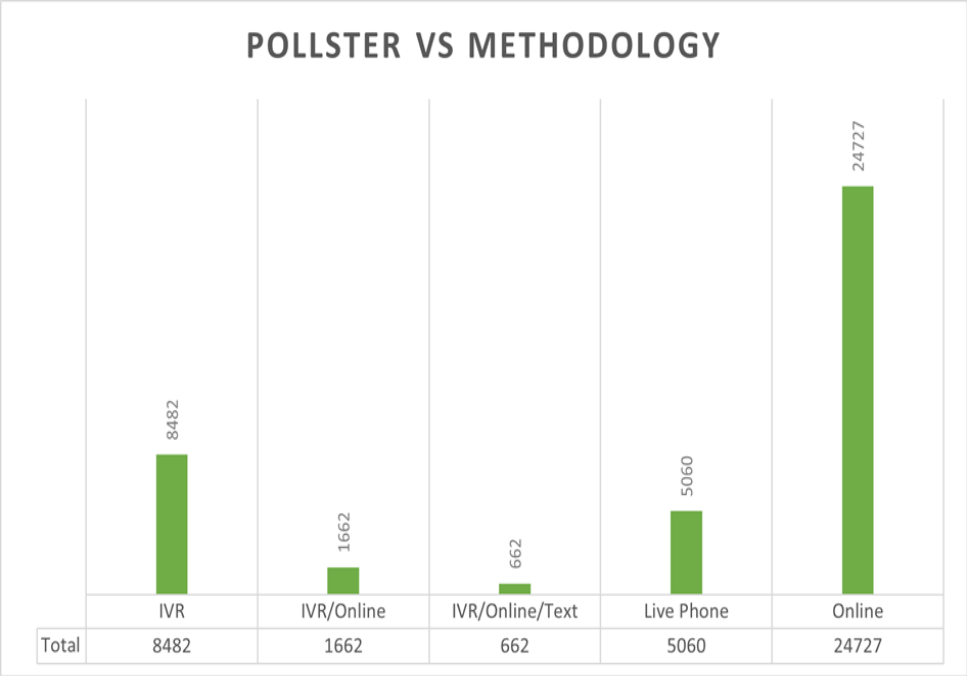
Below are some preliminary visualizations from our data.

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

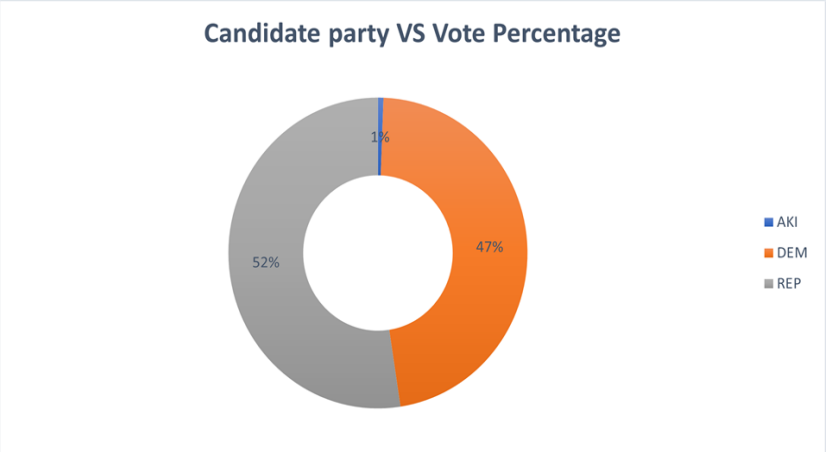
2)Candidate Party vs Vote Percentage
Question - Which party is winning?

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

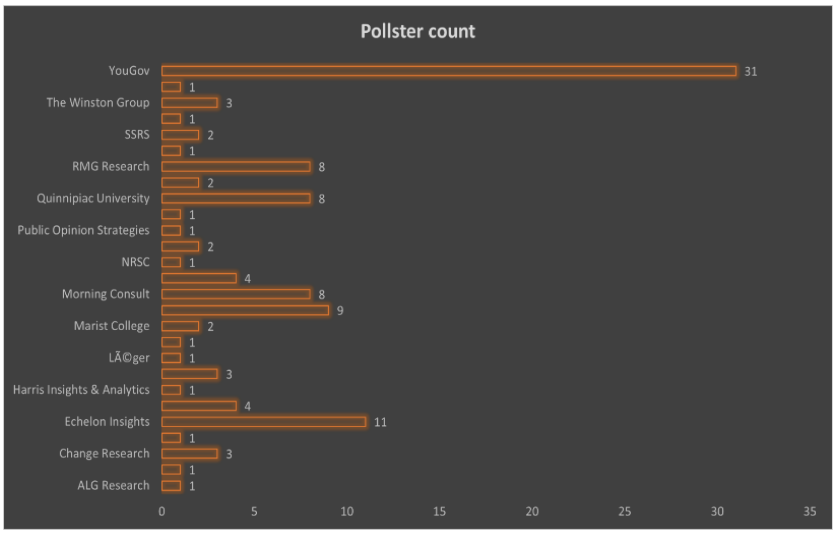
4)Distribution of pollsters across the USA.
Question - Are the pollster Divided equally within each state?



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.

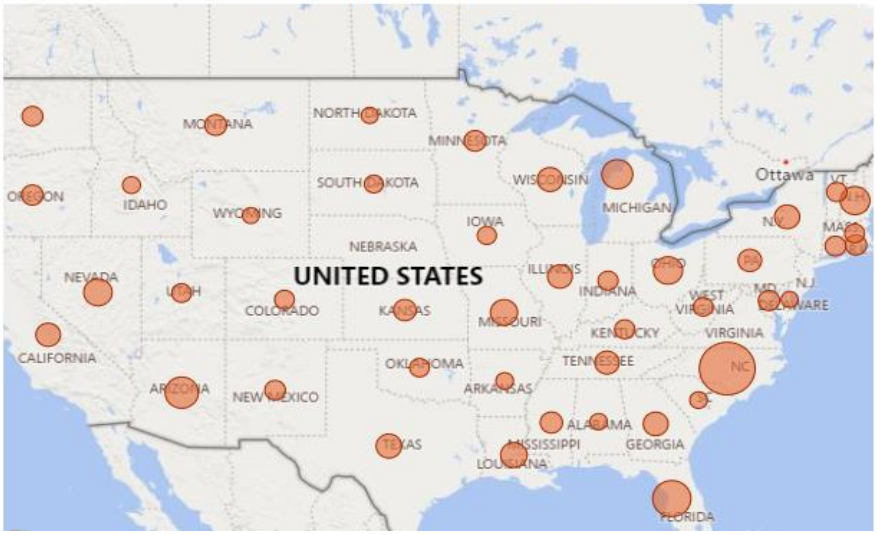


The above plot represents the Distribution of votes among the Parties. With this representation from which it is clear identify the winning party.



This graph shows the number of Polling booths under the particular pollster.It is seen that the Goverment body YouGov is the one with maximum amount of pollster under it.

Poll Distribution throughout USA



This graph depicts the distribution of pollsters throughout the USA. The Location can be identified of the pollsters statewise here.

It is to be concluded from the above visualization 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. It is to be 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.

5. DATA CLEANING

Data cleaning of 12 CSV files includes following steps:

- Missing Values
    - Most of the variable has missing values (NaN), So it has been replaced based on variable data type.
    - Categorical variable is replaced with Unknown values.
    - Replacing NaN values for Numerical column with 0.
- Dropping Irrelevant Column
    - Column with 100% values or more than 50% are dropped.
    - Column which are dropped listed below:
    - Partisan, seat number, sponsor candidate, tracking, notes, url, dem, rep, ind, Party, Politician I'd, Politician, Source, yes, No, Alternate answer.
    - After dropping useless column now total there are 33 columns with 32602 entries.
- Converting Datatype for Date containing columns
    - Three variables with Dates are: Start Date, End Date, Created at are converted from datatype object to datetime.
- Creating new CSV
    - After cleaning each CSV files using python on jupyter notebook and generating New CSV files by converting data frame to csv format.
    - Technology used : - Microsoft Excel , Power Query , Tableau.

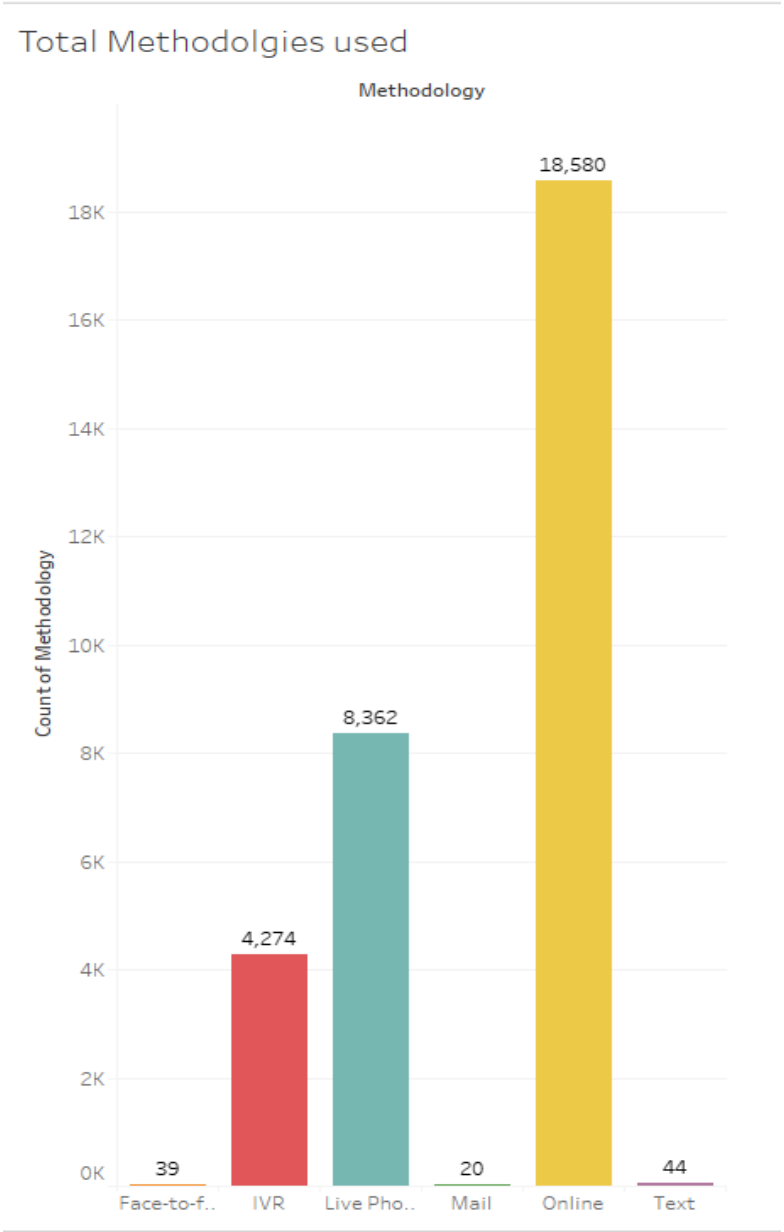
6. DATA TRANSFORMATION

- Merging data into single DataFrame
    - After Cleaning each csv files and creating new one will get merged into single dataframe for further analysis.
- Need to do Data cleaning Again
    - After Merging all files data cleaning phase should be repeated for better visualization.
    - Repeated Steps:
    - Replacing NaN values with 0 for Numeric variable.
    - Filling Categorical variable with Unknown.
    - Converting Data Type of Date variable with datetime.
- Renaming all columns
    - All columns are renamed with valid names. Total 33 columns with dtypes: datetime64[ns](3), float64(2), int64(4), object(24)

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Sr No.                    32602 non-null  int64
 1   Question ID               32602 non-null  int64
 2   Poll ID                   32602 non-null  int64
 3   Race ID                   32602 non-null  object
 4   Cycle                     32602 non-null  object
 5   Pollster ID               32602 non-null  int64
 6   Pollster                  32602 non-null  object
 7   Sponsor IDs               32602 non-null  object
 8   Sponsors                  32602 non-null  object
 9   Display Name              32602 non-null  object
 10  Pollster Rating ID        32602 non-null  float64
 11  Pollster Rating Name      32602 non-null  object
 12  FTE Grade                 32602 non-null  object
 13  Sample Size               32602 non-null  float64
 14  Population                32602 non-null  object
 15  Population Full           32602 non-null  object
 16  Methodology               32602 non-null  object
 17  Office Type               32602 non-null  object
 18  Start Date                32602 non-null  datetime64[ns]
 19  End Date                  32602 non-null  datetime64[ns]
 20  Nationwide Batch          32602 non-null  object
 21  Created At                32602 non-null  datetime64[ns]
 22  Stage                     32602 non-null  object
 23  State                     32602 non-null  object
 24  Seat Number               32602 non-null  object
 25  Election Date             32602 non-null  object
 26  Internal                  32602 non-null  object
 27  Ranked Choice Reallocated 32602 non-null  object
 28  Answer                    32602 non-null  object
 29  Candidate ID              32602 non-null  object
 30  Candidate Name            32602 non-null  object
 31  Candidate Party           32602 non-null  object
 32  Pct                       32602 non-null  object
dtypes: datetime64[ns](3), float64(2), int64(4), object(24)
```

7. DATA ANALYSIS & VISUALIZATION

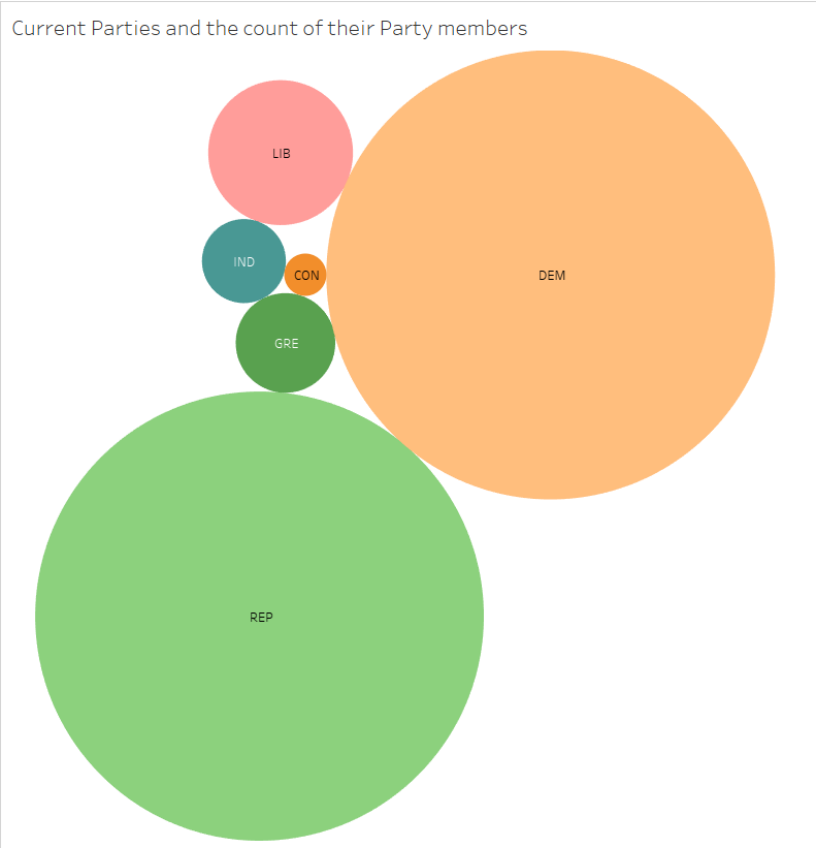
What medium of voting is preferred the most by the voters?



A. Today, there are a plethora of voting technologies available, but there is a notable absence of remote voting options. But with the increasing technology, there is an increase in options open for the voters. There are six options available which are face to face, IVR (Interactive Voice Response), live phone, Mail, online and text. With the data we have when we perform analysis the highest method which is preferred by the voters is the online method. Secure, simple, and

easy online voting to modernize and simplify elections. Voting online has its own benefits like you cannot vote from your phone or online if you are in the military or a citizen of a foreign country registered in specific US states. During this time online voting comes as a solution. With this analysis as we know that online voting is highly preferred, we can make it easier, safer, and secure for the future elections.

Which political party rules the election and has the highest number of members?

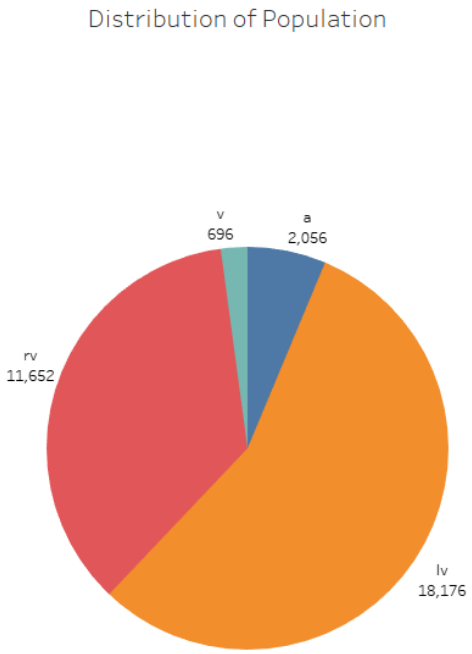


The electoral system in the U.S. is called a two-party system. There are more than 69 registered political parties, each with its own set of political beliefs. The Republicans and Democrats, on the other hand, are the only two national political parties.

There are other minor parties in the United States the Libertarian Party, the Green Party, the Constitution Party, and other minor parties have less influence than the main parties. With the graph above it is certain that both parties' rule but there is a minor difference between them with minimal numbers. Lower taxes, deregulation, more military expenditures, abortion restrictions, immigration limitations, gun rights restrictions, and labor union restrictions are all priorities for the Republic party. Americans who share the same political views as candidates from one of the third parties may instead vote for either the Democratic or Republican parties.

They do this to guarantee that their vote goes to the candidate with the best possibility of winning. This maintains the two-party system. A two-party system has the advantage of ensuring that the two major parties in power have a broad platform that represents the overall public. Because the two parties are so vast, each can accommodate a wide spectrum of political positions. This means that within any party, there may be minor differences in political perspectives on certain issues.

How many people come out to vote?



The right to vote is one of the most essential rights that American citizens have. Citizens over the age of 18 cannot be refused the right to vote based on their race, religion, gender, disability, or sexual orientation. The types of voters are divided into four major categories according to the data used, that is LV (likely voters), RV (registered voters), A(adults), V(voters). Likely voters own most of the number and voters have the least. Likely voters (LV) are those who have indicated to polling companies that they have a strong will to vote on election day.

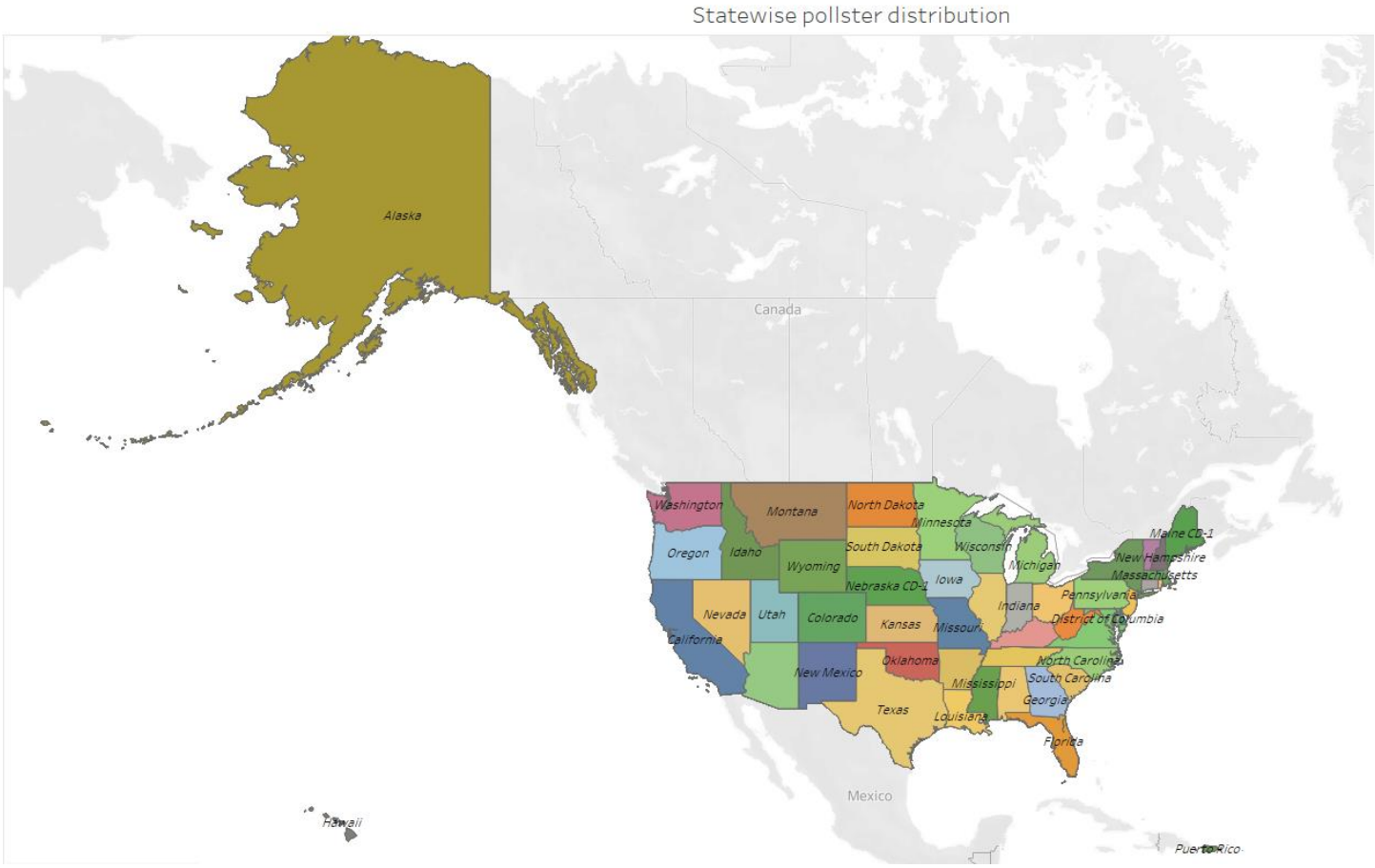
Some characteristics of likely voters are higher levels of income, education, and occupational status associated with a keen sense of party identity and the belief that voting is a critical act.

Registered voters (RV) are individuals who declare they are "registered to vote in their precinct or election district" in response to a typical poll question. This is the group whose data Gallup

most frequently publishes since it represents an estimate of Americans who are eligible to vote and who might vote if they so choose. Not all the people who have registered to vote will cast a ballot. As a result, over the years, Gallup has developed algorithms to identify likely voters, or those who the business believes are most likely to vote.

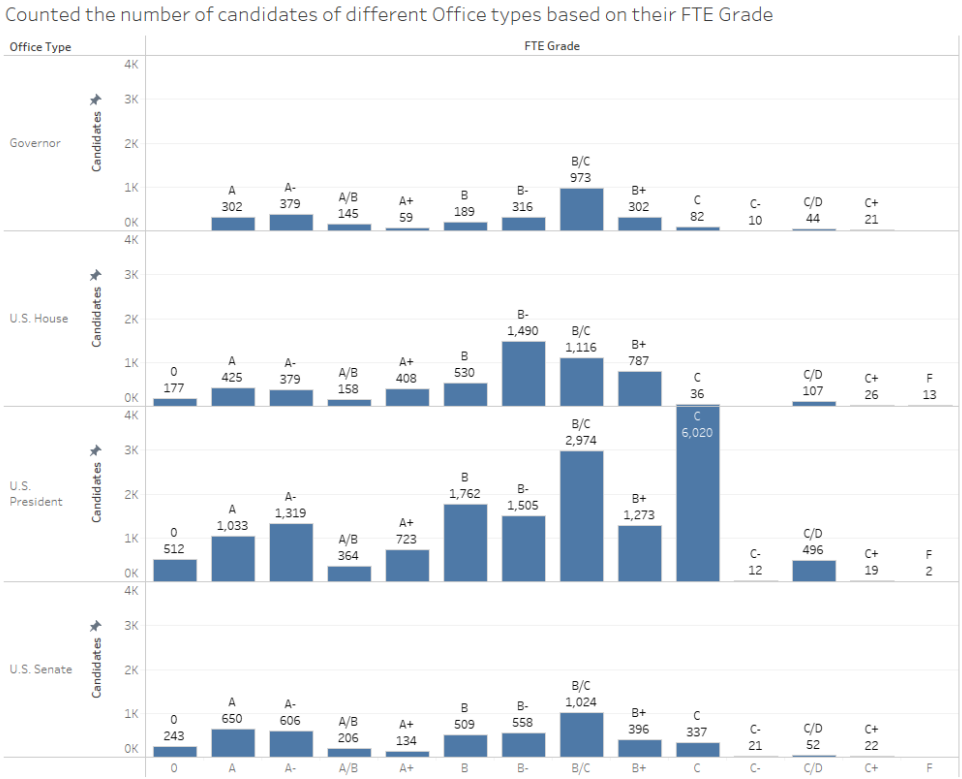
All adult citizens(A), except for a few minor exclusions, have the right to vote regardless of their money, income, gender, social standing, race, ethnicity, political attitude, or any other restriction. Voters(V) are the citizens who have the right to vote.

The different pollster assigned to each state?



Different type of pollster is allocated to a specific state to conduct an election poll in United State of America.

According to the different office types, which candidate has the highest/lowest fte grades?

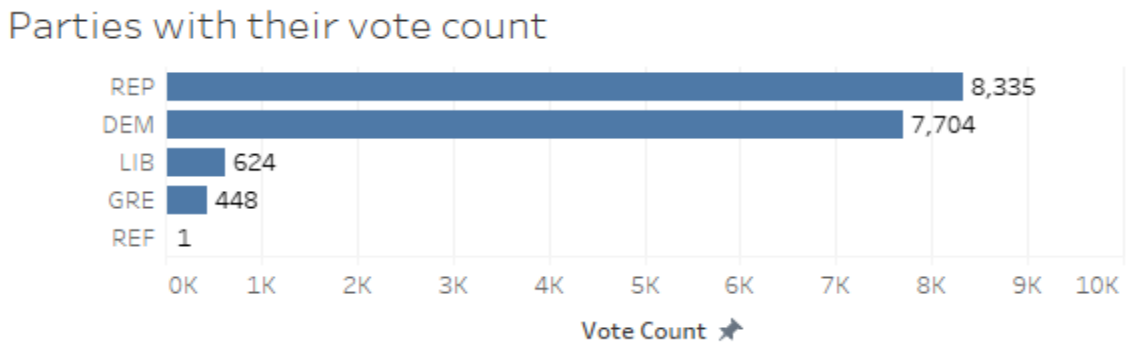


By analyzing, There are total of four office types:

- Governor
- U.S.House
- U.S.President
- U.S.Senate

U.S.President and U.S.House's office candidate receives more grades than others. There are a total of 6020 candidate of U.S.President office who achieves c grades, however very few candidates from U.S.Senate and Governor's office has c grades, which is 337 and 82 respectively. On the other hand, there are many candidates that receive c, a-,b/c, and b+ grades according to their office types. Moreover, none of the candidates receives c- and f grades of a different office. Only 21 candidates have c- grades from U.S.Senate and merely 13 candidates have f grades from the U.S.House office.

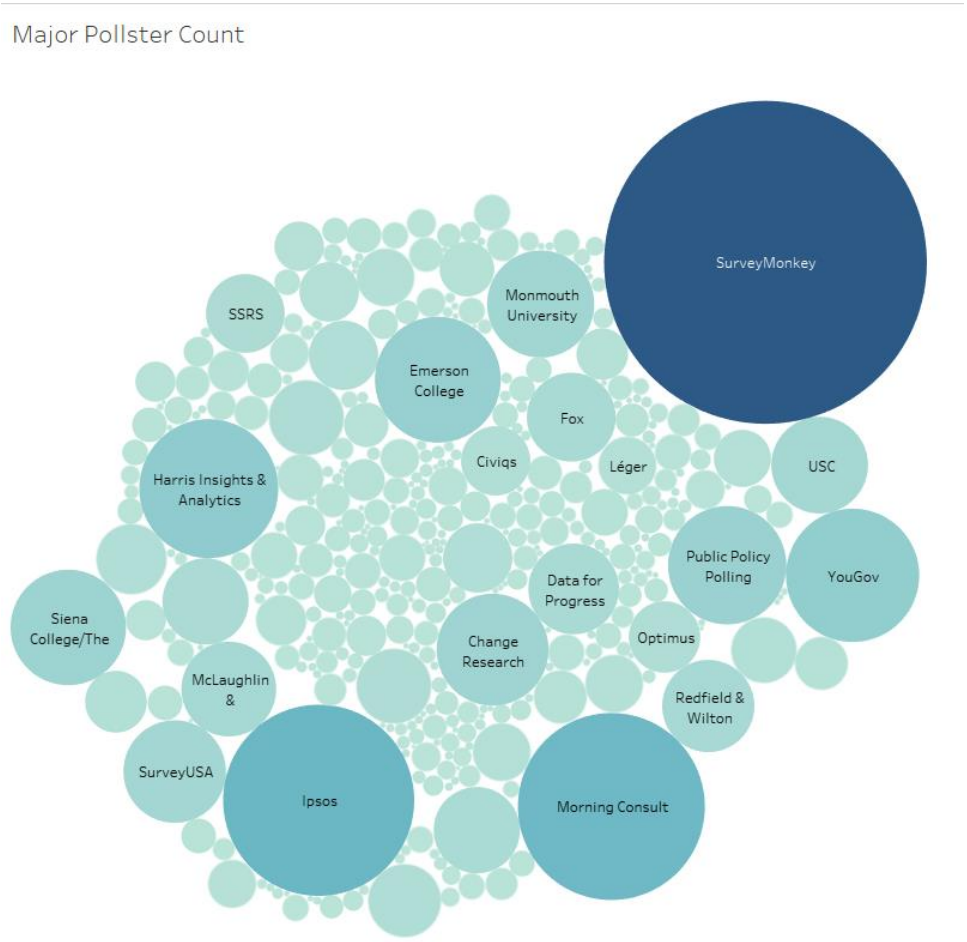
Which party won the most vote?



It is obvious that those parties which work hard get highest vote. But as a future perspective it will not be identified which party covers the most vote as every year, parties work very hard to win the election.

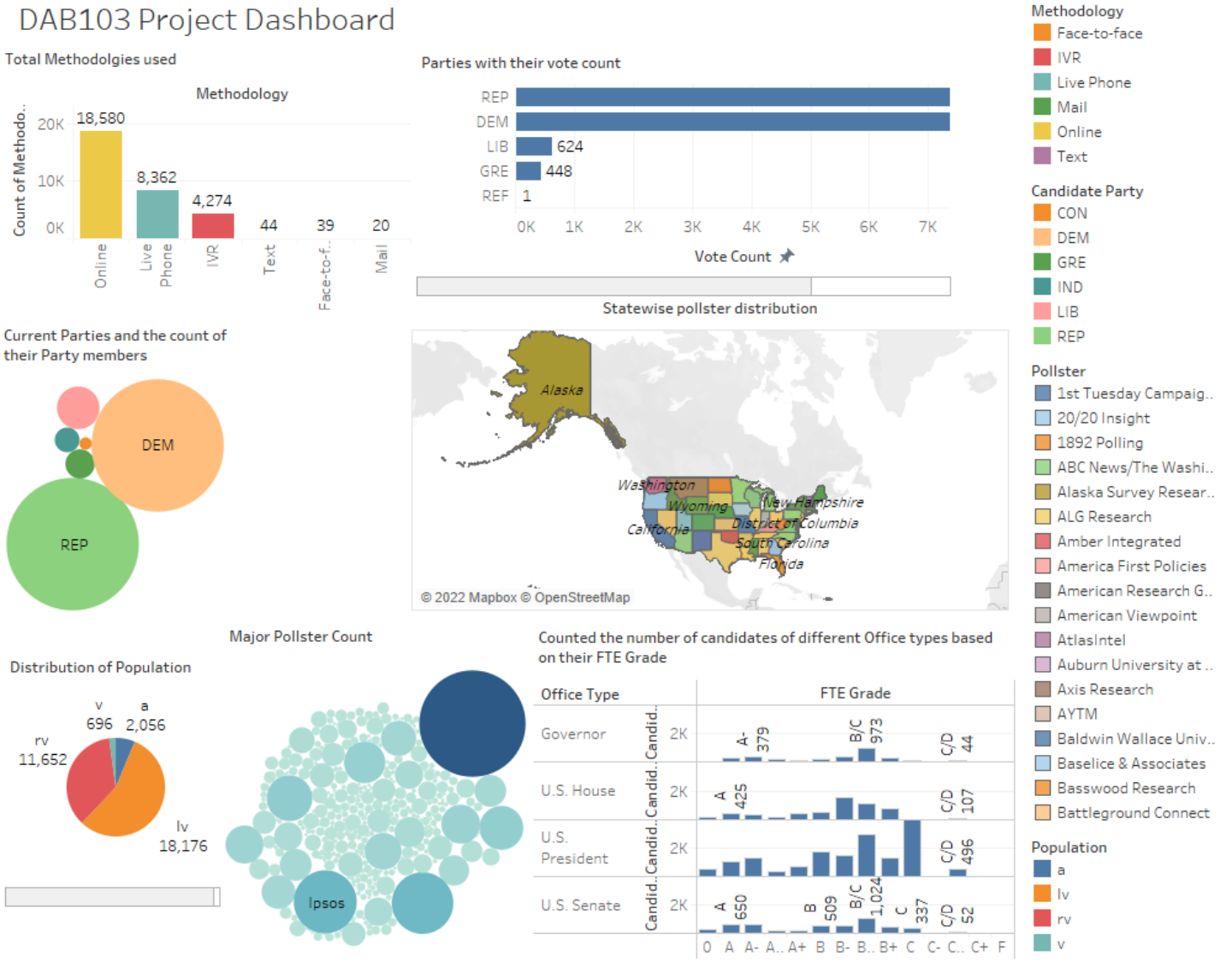
There is total 5 parties involved in election. DEM and REP has highest number of candidates. As the data describes the candidates from REP has won the highest number of seats by beating all other parties which were fighting election against them.

Which pollster counts more winning probability for election in USA?



Pollster is a person who conducts an election survey to gather information based on the opinio ns of a sample of people. By analyzing the above graph, SurveyMonkey has the highest number of polls counting as compared to others, whereas civics countless number of the pollster. However, there are many others who conduct a survey for elections like Harris insights & analysis, public policy, YouGov, SSRS,etc.

Dashboard



The above is the dashboard of our project which included the data visualizations obtained from the data.

8.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

https://news.gallup.com/poll/110287/what-difference-between-registered-voters-likely-voters.aspx

https://dk.usembassy.gov/da/youth-education-da/the-american-political-system/the-democrats-and-the-republicans/