Data Noobs

Ritesh Suhag, Tushar Pandey, Sambandh Dhal, Swarnabha Roy

Executive Summary

Website: https://share.streamlit.io/ritesh-suhag/us_elections_analysis/us_elections_app.py

This project focuses on analyzing the role of money in the US Presidential Elections (2012 to 2020). The Presidential elections play a huge part in determining a lot of policies in the United States owing to the amount of power the position entails. Many companies donate money in an attempt to help the party they support and the amount donated has increased significantly over the years. In 2020, the parties received a total contribution of $2.7 billion as compared to $681 million in 2012. These donations include the money raised by the Political parties in the years leading up to the elections.

The role of money can also be seen through the fact that in 2020, the total expenditure spent in the elections was $6.2 billion as compared to $1.4 billion in 2012. This expenditure included all the elections (Presidential, State, and House). In this project, we focus only on the Presidential Elections. The expenditure done by the parties was divided into the following categories using Natural Language Processing (NLP) - Advertisement, Logistics, Communications, and Others. The expenses spent on advertisement were the highest at $5.8 billion in 2020 as compared to $32 million spent on logistics in the same year.

Along the same line, we analyzed how the expenditures varied according to different parties and states. This was analyzed in conjunction with the demographics data of the state to get a better idea. It was seen that in 2020, Democrats were confident in some states and knew about the states they couldn’t win, thereby decreasing their investment in those states. In 2012, an interesting correlation was found between states with low expenditure and election results. All the states where Republicans have spent around less than $ 3 million are red states crushing the other parties.

To further analyze the expenditure, we formulated an index named ROI (Return on Investment) to account for the percent vote difference to the amount of money spent in the state. In various states, the Republicans lost votes despite an increase in expenditure of millions of dollars. If we look at the Democrats, only five states witnessed a surge in blue votes whereas, in 2020, all the states had an increase in the number of votes for the Democratic party. Most of the states with high ROI went Blue, except Utah in 2020.

Analysis was carried out between population data, expenditure data, and the voting percentage for three consecutive Presidential elections. There was a correlation between certain populations in some states and how they affect the electoral votes. States with a higher population of people who are neither African American nor White tend to favor the democrats.

The analysis was ended by carrying out a Network analysis on the data. There are some great insights from this model, specifically on the clustering part. There is one cluster with states where Democrats won and their expense is around the same as the Republicans. Similarly, there is another one where Republicans won, which gives us some information on the states where parties spend a similar amount of money and people tend to vote for the same party.

Problem Statement

Money has a huge impact on US presidential elections. The aim is to look into the depth of how political parties spend and get money as donations during the Presidential elections. As an analyst, we have to observe patterns, check the effectiveness of the expenditures and provide some inferences and recommendations as to where to invest the money for the next Presidential elections.

Data Collection and Preprocessing

The analysis has been made on two groups of data: Donations and Expenditures. We created a dataset to analyze the Donations received by the two major political parties (i.e. Democrats and Republicans). The data from the major corporate donations have been taken into account. The corporate donation data has been segregated based on the industry sector i.e. Finance, Healthcare, Defence, law, Energy, etc. The following links have been used to collect the donation data during various Presidential Elections over the years:

- https://www.opensecrets.org/orgs/top-donors?topdonorcycle=All
- https://www.opensecrets.org/industries/indus.php?ind=w04&cycle=2016
- https://www.opensecrets.org/industries/indus.php?ind=W05&cycle=2016
- https://www.opensecrets.org/industries/indus.php?ind=Q
- https://www.opensecrets.org/industries/indus.php?ind=B
- https://www.opensecrets.org/industries/indus.php?ind=P
- https://www.opensecrets.org/industries/indus.php?ind=F&cycle=2020
- https://www.opensecrets.org/industries/indus.php?ind=M
- https://www.opensecrets.org/industries/indus.php?ind=N
- https://www.opensecrets.org/industries/indus.php?ind=D
- https://www.opensecrets.org/industries/indus.php?ind=E
- https://www.opensecrets.org/industries/indus.php?ind=A
- https://www.opensecrets.org/industries/indus.php?ind=K

Donation data over the years have been collected and consolidated into a single datasheet. Additionally, a state column has been created based on the location of the Corporate Headquarters.

Dataset related to individual party expenditure and county elections.

- https://www.fec.gov/data/independent-expenditures/?most_recent=true&data_type=processed&is_notice=true
- https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ
- https://data.world/bdill/county-level-population-by-race-ethnicity-2010-2019

Data associated with different political parties for each election year were collected and analyzed. At first, the features were cleaned and null entries corresponding to the state were removed. Further, the data were merged to create a dataset, where for each state, the expenditure for political parties as well as the share of the vote they received for each election year is reported. We also worked on the expenditure data set, categorized the purpose of different investments made by each political party for further analysis.

Once it was done, we looked into census data and pruned it to make it mergeable with the previously constructed data set. Since there are multiple races, we cut it down to three, namely: White population, Black population, and Other population. In this way, data cleaning and feature engineering were done.

Data Exploration

For each of the companies, in either of the categories, the headquarters were appended as a column to the dataset. Based on the state where their Headquarters were located, corporate fundings were analyzed as to how the donation made by these companies impacted the winners of the Presidential race in each of these states. The last three presidential campaigns have been observed and the donations made by the companies belonging to each of these sectors have been scrutinized and the pattern of these corporate fundings has been analyzed in detail. A detailed analysis with interactive visualizations has been shown on the website.

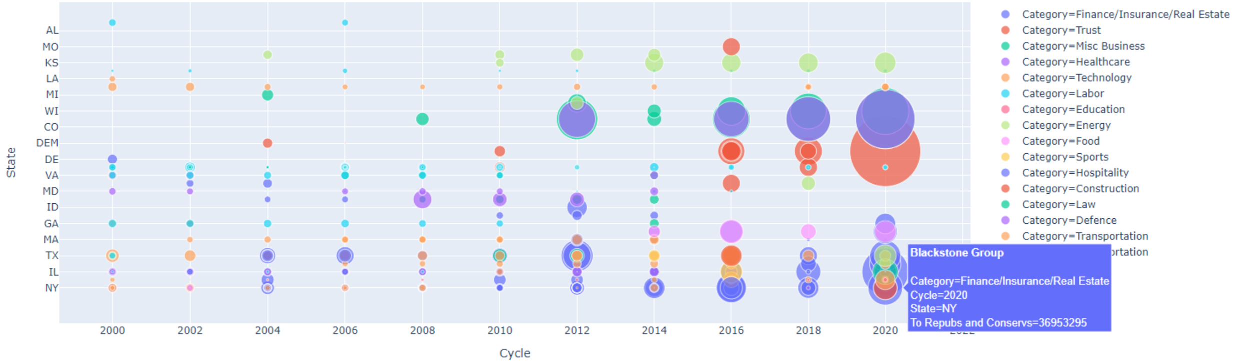


Figure 1: State-wise Year-wise Corporate donations to the Republicans

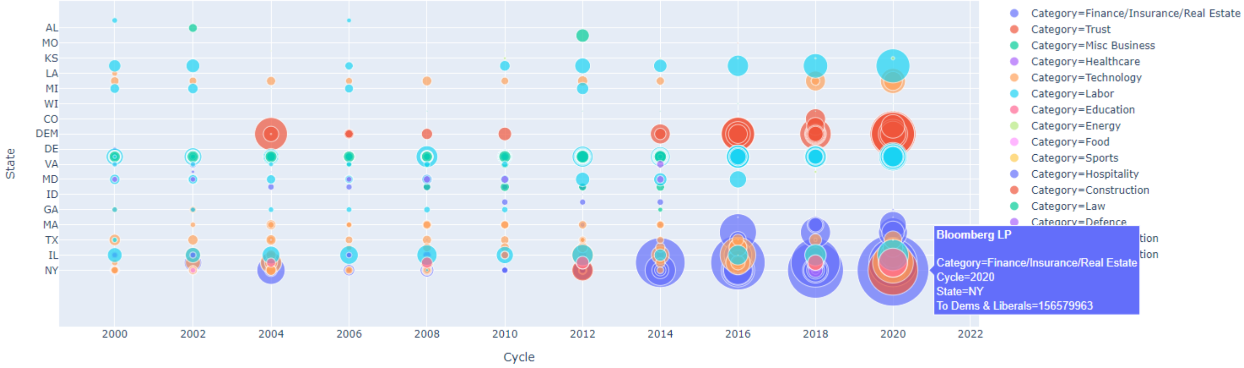


Figure 2: State-wise Year-wise Corporate donations to the Democrats

The donations received by the Political parties have drastically increased over the years as shown in the snapshots Fig.1 and Fig.2. The radius of the circle is proportional to the amount donated by the companies. Companies have been divided into various sectors which are denoted by various colors. Some interesting observations have been made based on the collected data. For instance, companies headquartered in New York have made significant donations to the Democrats while companies headquartered in Wisconsin have made heavy donations to the Republicans in the past Presidential elections.

Methodology

- **Feature Engineering**: After cleaning the dataset, some more features were constructed which includes the population of several races as well as changing the total number of votes to the percentage of votes. Further, we changed the expenditure amount in millions of dollars and population in 100,000. This helped in creating better data which will be further used in unsupervised learning.
- **Unsupervised Learning**: Once suitable features were engineered, we moved to some clustering methods. We approached the data in several ways using KNN, AgglomerativeClustering, Dendrograms, and Topological Data Analysis. Since the data points are not large enough, we didn't keep the number of clusters as a hard number. Rather, with the mapper package, we see a better picture of our clusters which brings us to the visualizations.
- **Dynamic Visualization**: The graph which contains clusters is colored in a way that represents much better graphics along with information for every node. We host a website to represent different findings we have with an option to choose the party the analyst would like to know about. For each year, anyone on the website can look into how money or people of different races are related to vote shares of different parties for the past three elections as well as which institution prefers which party. Therefore, the website we have designed is very dynamic and suitable for everyone.

Modeling, Analysis, and Interpretation

Expenditure Analysis

The key idea of democracy is that every citizen should have input, via a vote, as to who is elected. Enormous amounts of money are expended by political campaigns to engage with voters, and so fundraising has become a major activity by candidates and other actors. The money is spent for various purposes.

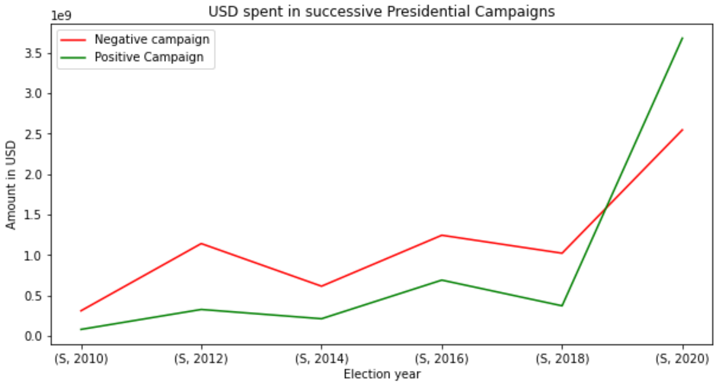


Figure 3: Amount spent over the years in Election campaigns

In Figure 3, two kinds of election campaigns have been addressed: Positive Campaign and Negative Campaign. By positive campaign, it refers to a wholesome release of a manifesto and developmental objectives by a Political Campaign. By negative campaign, it refers to finding out the faults and shaming the other party viciously to gain political mileage. The amount spent in USD in the last three consecutive election campaigns has been addressed. We can go on to see that, despite the negativity that clouded the recently concluded Presidential elections, the amount spent in positive campaigns by each of the political parties outcast the past presidential elections.

Fig. 4 demonstrates the trend in amount expenditure of each party for the last 10 years. From the graph, the first obvious conclusion is: The campaigning expenditure for each party increased five-fold for the last presidential year. This increased the total voting numbers. Another subtle inference is the change and the difference in the expenditure of Republicans and Democrats from the last presidential election to the 2020 presidential election.

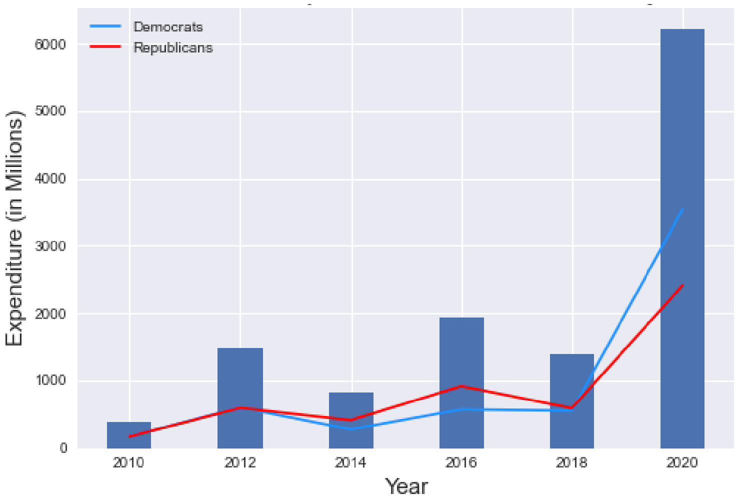
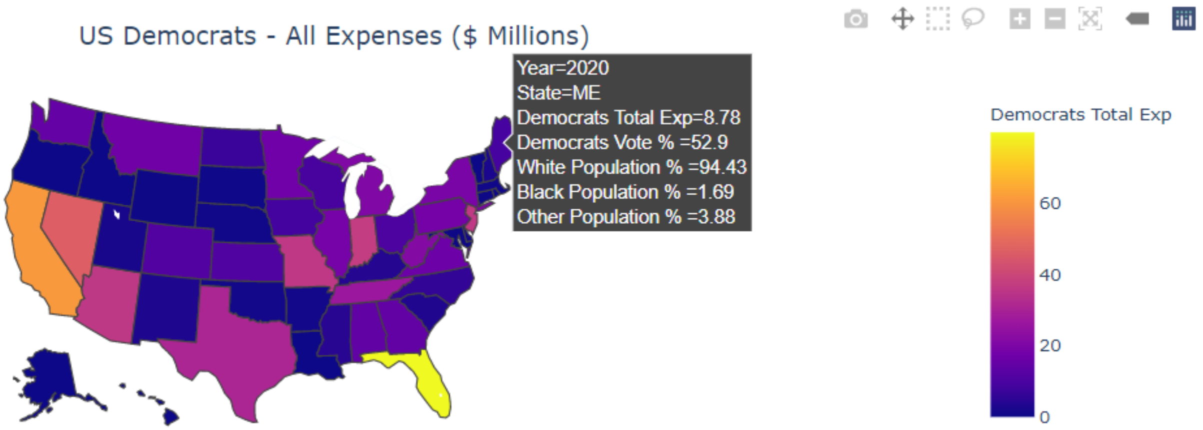


Figure 4: US Presidential Election expenditures (in USD) over the last 10 years

After analyzing the bigger picture, the next thing needed is “micro-analysis”. From national scope, we switch to states and start looking at different expenses. The dataset which comprised various purposes of expenditure was categorized using NLP into categories: Advertisement, Communications, Logistics, and others.



Year=2020
State=ME
Democrats Total Exp=8.78
Democrats Vote % =52.9
White Population % =94.43
Black Population % =1.69
Other Population % =3.88

Figure 5: Democrat expenses in 2020

An outlier corresponding to the state DC was removed because of the unexpectedly high investment from both parties even though it's a predominantly Democratic stronghold for ages. For different parties and categories of expenditure, graphs are displayed.

Plot Overview: Democrats

In the year 2012, interestingly, the Democrats invested a lesser amount in the states where Republicans are in a huge majority. At the same time, Democrats spent a huge amount of money to get the clear majority, in states like Illinois, California, Nevada. However in 2016, all the states where Republicans have a huge margin in their victory witnessed a huge drop in expenditure from the Democrats. Since Democrats were in power for the last two sessions, they decided on cutting down the expenses. Last year, after some political events, Democrats were confident in some states and knew about the states they couldn't win, thereby decreasing their investment in those states. In some states, Democrats kept investing money even though they were far away from winning, planning for long-term success as shown in Fig. 5.

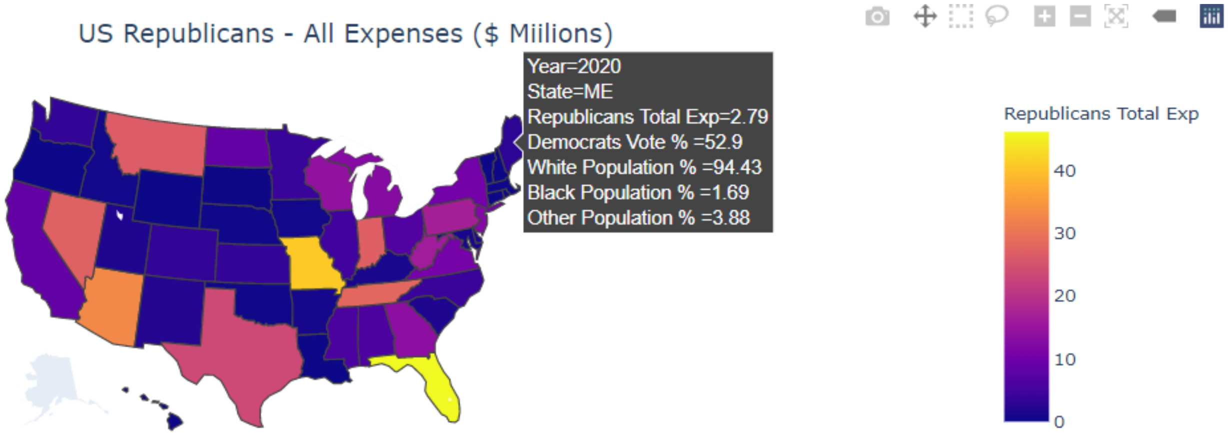


Figure 6: Republican expenses in 2020

Plot Overview: Republicans

An interesting correlation was found between states with low expenditure and election results for the year 2012. All the states where Republicans have spent around less than $ 3,000,000 are red states crushing the other parties. For 2016, the previous states where Republicans didn't feel the need to spend money and still get a victory witnessed a decrease in expenditure along with the other red states. Since Democrats were in power for the last two sessions, Republicans strategically cut down their expenses. For the current one, this year's data is rather interesting because only the aforementioned states (in 2012) had less expenditure. Perhaps, even in some red states, Republicans have invested a large amount of money because of the increase in Democratic votes. There are a few states where Democrats and Republicans were on an investing spree to secure at least the votes they had in 2016.

From the previous graph, we can see that each party wanted to go all in this time and win the elections. But does investment always lead to an increase in the voting percentage? To answer this question, we defined a **Return of Investment (ROI)** index for different states based on how much campaigning influences the results for both parties.

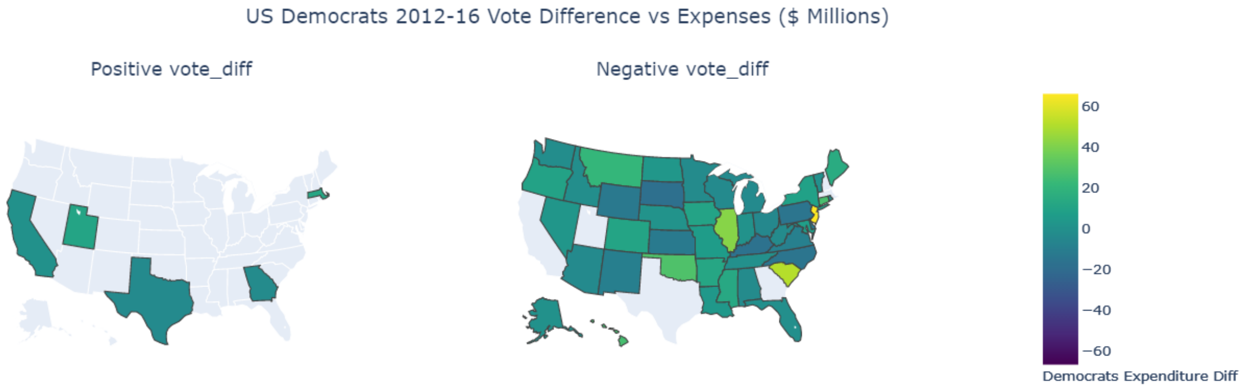

US Democrats 2012-16 Vote Difference vs Expenses ($ Millions)
Positive vote_diff
Negative vote_diff
60
40
20
0
−20
−40
−60
Democrats Expenditure Diff

Figure 7: Democrat Vote Differences vs. Expenses (2012-16)

The graphs are divided into two subgraphs where one corresponds to the states where the percentage of votes received by Republicans/Democrats increased and the other subgraph where we see the states when the percentage of votes received by Republicans/Democrats decreased between two presidential elections.

The first important question, whether investing always helps is answered through all the graphs. In various states, the Republicans lost votes despite an increase in expenditure of millions of dollars. The ROI for Republicans is positive (significantly larger than 0) for the states (except Utah) where they lost, which means they reduced their expenditures there. In 2016 in Utah, they reduced their investment and therefore, lost a lot of votes (more than 25%), but the state remained red. In the next presidential election, they invested more and were able to get the votes in their favor. If we look at the data where the Democrats won, only five states witnessed a surge in blue votes, whereas in 2020, all the states had an increase in the number of votes for Democratic party. Most of the states with high ROI went Blue, except Utah. Because of a decrease in Republican votes in 2016, the Democrats invested just a little in 2020 and were able to secure 10% more votes than in 2016, thereby giving a huge ROI for the same. This way, the graph helps us understand the trends and the power of campaigning on the minds of different people beautifully.

People, Money, and Voting!

A scatter plot (Fig 8) with population data, expenditure data, the voting percentage for three presidential election years tells us about some correlation between the certain population in some states and how they affect the electoral votes in their respective states.

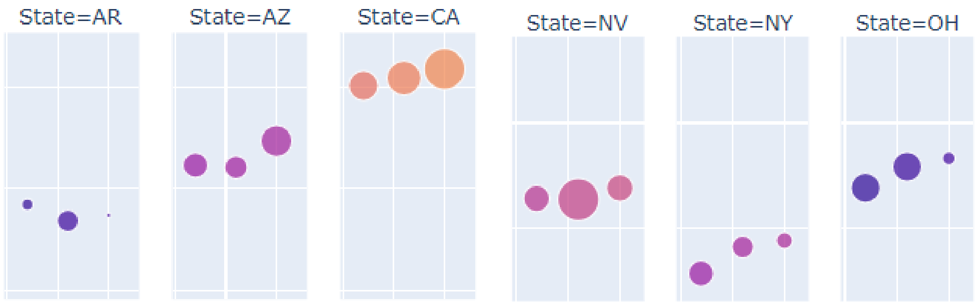


Figure 8: Percentage of votes state-wise (2012,2016 and 2020)

For Democrats, states with a higher population of people who are neither African American nor White tend to favor the Democrats. Another riveting find is that the expenditure is surging for the states in which the Democrats already have a huge advantage. Meanwhile, where there is an equal chance of both Democrats and Republicans or Republicans winning, the expenditure has decreased for 2016 which is concurrent with the US politics trend. States with a higher population of people who are neither African American nor White tend not to favor the Republicans. In contrast to the conclusions drawn from the graph for Democrats, the trend in change of expenditure throughout the election years is a little different. Some states saw an increase in the expenditure throughout the last three presidential elections, while a general trend is an increase in investment in 2016 because it was a year marked by the victory of the Republicans.

Correlation Heat Map

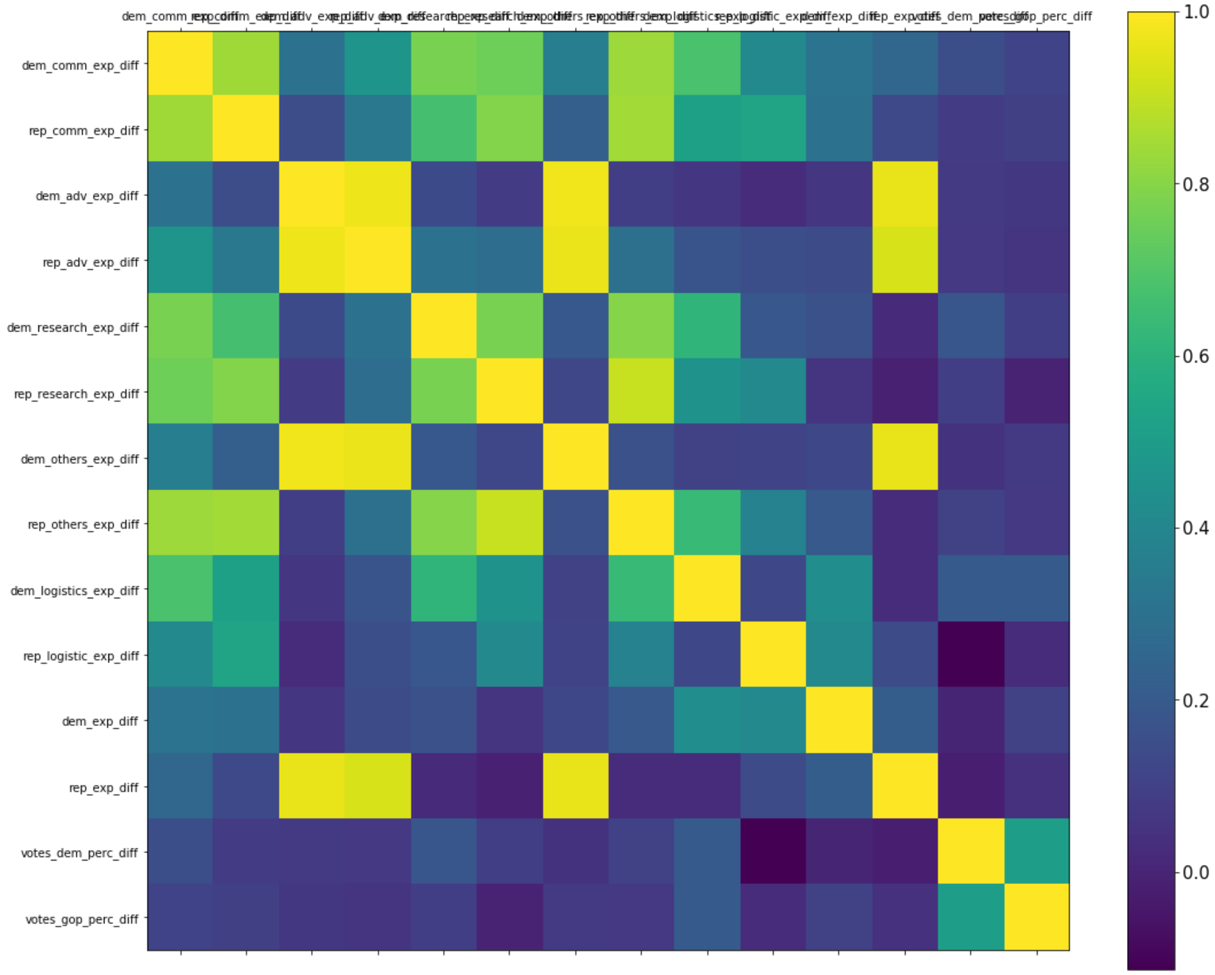


Figure 9: Finding feature correlation

A basic heat map (Fig. 9) is plotted to look into correlation between different features of our clean dataset. There are some columns with highly negative correlations. This is also evident from the previous graphs and inferences.

Unsupervised Learning:

Since the dataset is small, the number of clusters found will not be optimal. Therefore multiple approaches were taken to get an idea about different clusters and later about whether or not these clusters make sense. The first approach is the Elbow method using KNN for different values of k and identifying the elbow.

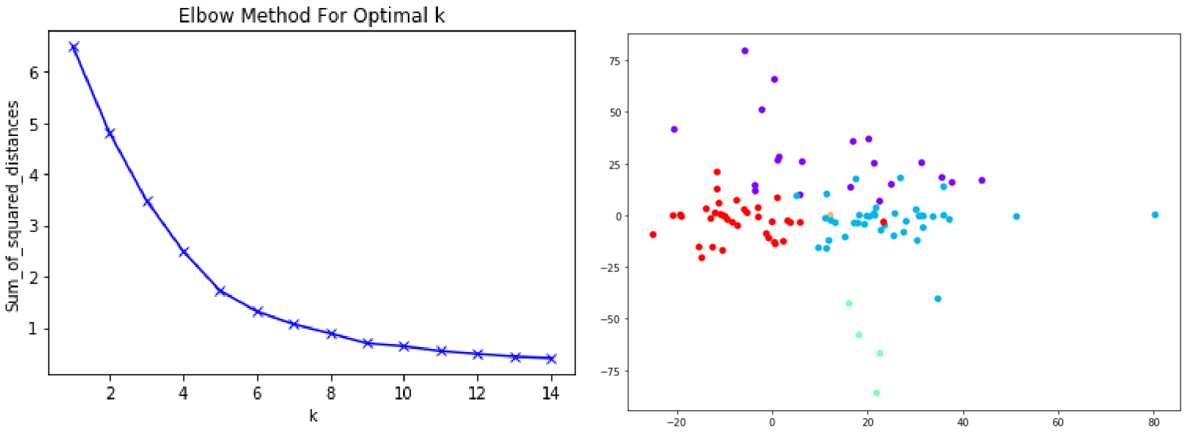
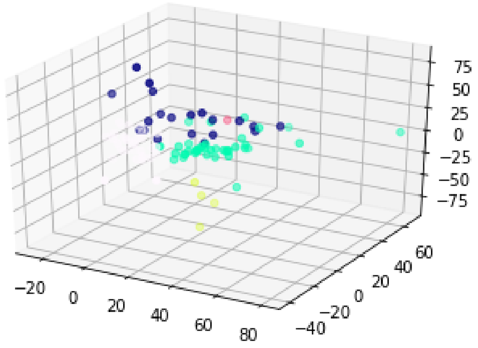
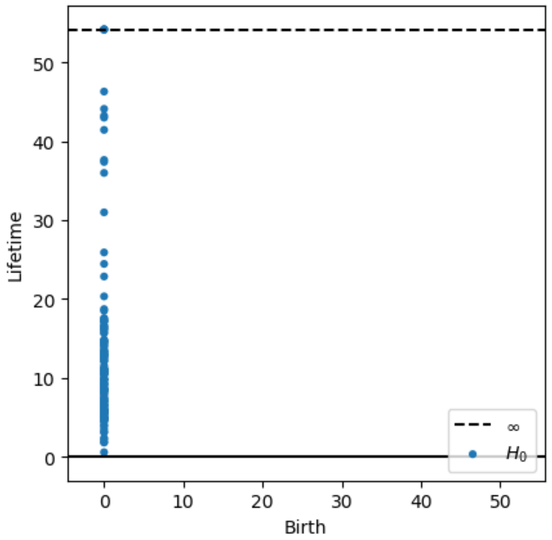


Figure 10: Knn optimal clustering (top left); AgglomerativeClustering projection 2D (top right); AgglomerativeClustering projection 3D (center); Persistent Homology (bottom)





Here, 5 or 6 seems like a good choice for the cluster number. Later, using AgglomerativeClustering, each data point was categorized and a 2-D projection is plotted. It does not seem very clear if one needs to add or remove labels, thus, another 3D projection is analyzed. It gives a slightly better identification of the clusters. Because of relatively low sample points and higher dimension, we use Topological Data Analysis.

Topological Data Analysis is a clustering algorithm that relies on topology and creates a Cech complex of a point-cloud. Using persistent diagrams, the number of clusters is calculated. From there, using t-distributed stochastic neighbor embedding, the data is projected in a manner such that some properties are preserved. After the projection, a clustering algorithm along with covering balls is applied to the dataset to obtain a 3-dimensional graph. The hyperparameters are autotuned based on the variance and mean valency of the graph. Lastly, the graph is colored according to different attributes associated with the original dataset to provide better visualization.

There are some great insights from this model, specifically on the clustering part. There is one cluster with states where DEMOCRATS won and their expense is around the same as republicans. Similarly, there is another one where REPUBLICANS won, which gives us some information on the states where parties spend a similar amount of money and the people tend to vote for the same party. Another cluster has states where democrats did not spend more than the Republicans, but still, they won, which says that the state is blue. These insights were helpful in understanding the relationship between expenditure and voting in different states.

Conclusions and Recommendations

The trivial binary metric from the data is not sufficient to hand in any recommendations, therefore we created our fluid one, called Return on Investment(ROI). ROI index can be used to determine the strategy of expenditure for future presidential elections. If the ROI for a particular state is positive and high, this means the party can remove some funds and invest in the states with lower or negative ROI. The advantage of using ROI is its historical value. The ROI changes after every Presidential election which makes the investment strategy dynamic. For the states where a party has a lower negative ROI, then the party needs to look into minute expenditures and political rallies. For higher negative ROI, the party can spend a little more and increase their vote share in that state.

Some other analyses involving demographics can also be looked into to determine voting correlations. In such states lowering the expenditure will not harm the party too much.

Possible areas of improvement: The demographics data for most of the MNCs are not publicly available. If we could have had a hold of the demographics of the companies, we could have linked it to the donations that these corporations have made to the parties depending on the diversity of these companies. If we can get more historical data for expenditure and sponsors, our dataset can be modeled into some AR(1) and ML models to know more about the effects of another (possibly new) feature.