IST652 - Scripting for Data Analysis M002: Spring 2023

DATA DRIVEN CAMPAIGNS

A Look at the 2020 Presidential Election

YASHANJALI CHAVAN

TABLE OF CONTENTS

INTRODUCTION

The 2020 United States presidential election was the 59th quadrennial presidential election, held on Tuesday, November 3, 2020. Incumbent President Donald Trump faced off against former Vice President Joe Biden in a contentious and polarizing campaign that highlighted deep political divisions within the country.The election took place against the backdrop of the global COVID-19 pandemic and related recession. It was the first election since 1992 in which the incumbent president failed to win a second term. The election saw the highest voter turnout by percentage since 1952.

Campaigns used data to target their messages to specific audiences, track the effectiveness of their messaging, and develop strategies that could help them win the election. Data also played a role in informing and managing the campaigns' resources, such as identifying areas where the campaign needed to focus its efforts and resources. By looking at the net worth data, political campaigns can get a better sense of the economic concerns of high net worth individuals and tailor their messaging to address those concerns.The expenditure data, makes it possible to see which channels were the most successful in terms of driving donations and raising awareness. This data can then be used to create more targeted outreach strategies in the future, as well as inform decisions on where to allocate resources in order to maximize impact. This provides a comprehensive and unbiased view of the election process. It can help to identify any irregularities. Furthermore, it enables citizens to make more informed decisions when voting and can also help to identify any potential areas of improvement in the electoral system.

Throughout this project I will take a look at the different strategies campaigns used, their fundraising efforts, and the differences in their expenditure. I will also look into the effectiveness of their approaches and the implications of the election results.

UNIT OF ANALYSIS

Considering the fact that it was one of the most highly anticipated and closely watched elections I was particularly curious to know the answers to a few questions I had in my mind.

- What was the total voter turnout and the demographic breakdown of the voters?
- What was the sentiment on Twitter regarding Joe Biden and Donald Trump during the election period, and how did it correlate with the final election results?
- Did a candidate's net worth and sources of income have any correlation with their campaign's success in the 2020 presidential election?
- How much money did each presidential candidate raise during the 2020 election campaign?
- Key areas of expenditure for each candidate and their recipients?

For the above questions to be answered I look into the data related to the following units of analysis:

- **Funds Raised**: Details of the funds raised by a candidate's campaign committees and outside groups.
- **Twitter Sentiment**: Based on tweets collected using #joebiden, #donaldtrump, and #election2020 from 15-10-2020 up until 08-11-2020.
- **Voter Demographics**: Based on the demographics of voters, such as gender, race, etc. It helps to understand the voter base in the USA.
- **Net Worth**: Details of a candidate's financial holdings, debt, and sources of income so that the public can identify any conflicts of interest they may have.
- **Expenditure**: Details on the expenditure made by candidates and their top recipients
- **Election Results**: The results of the presidential elections in each state.

DATA AND SOURCES

For the project I will be using multiple sources of the data. These sources provide valuable information on the election process, campaign strategies, voter behavior, and the impact of money in politics.

- **OpenSecrets.org**: OpenSecrets.org maintains a comprehensive database of political contributions, campaign finance records, lobbying data, and other related information. The website offers a range of tools and resources to help users explore this data and understand the impact of money on politics. For the data related to expenditure for each candidate, their net worth and the funds raised I have used urls from this website to access information which was available in HTML format.
- **Census.gov**: Census.gov is the official website of the United States Census Bureau, which is responsible for conducting the decennial census of the US population. The Census Bureau collects a wide range of data on the US population, including demographic, social, economic, and housing data. For the data related to the demographics of the total population including voters I have used an Excel file available on this website.
- **FoxNews.com**: Foxnews.com is the official website of Fox News, a major American news organization that provides news and analysis on a wide range of topics, including politics, business, sports, and entertainment. For the data related to the election results I have used the url from their website to access the information which was available in HTML format.
- **Kaggle** : Due to recent restrictions on the Twitter API, I came across CSV structured files of tweets collected using #joebiden, #donaldtrump, and #election2020 from 15-10-2020 up until 08-11-2020 on Kaggle. It contains around 970917 tweets for Donald Trump and 776885 for Joe Biden

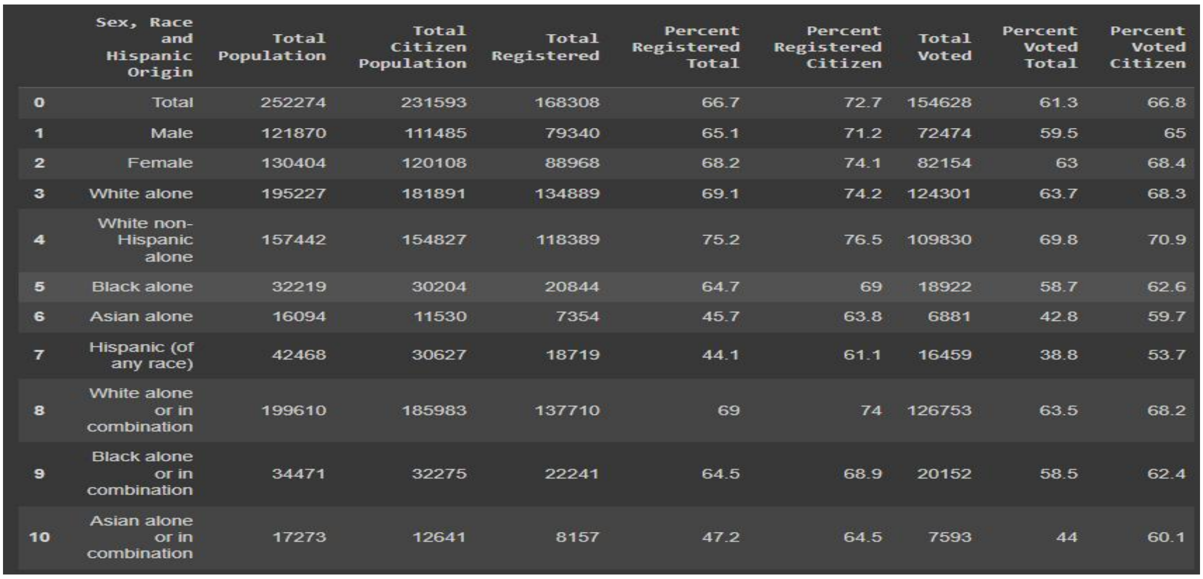
PACKAGES USED

The following packages were used for the analysis:

- Pandas for data cleaning, manipulation and analysis.
- urllib for opening the urls to get html data and handling errors that may occur when making the http requests.
- BeautifulSoup for extracting the data from a HTML file.
- re for manipulating text patterns.
- Numpy for array manipulation
- Plotly for creating interactive and visually appealing data plots.

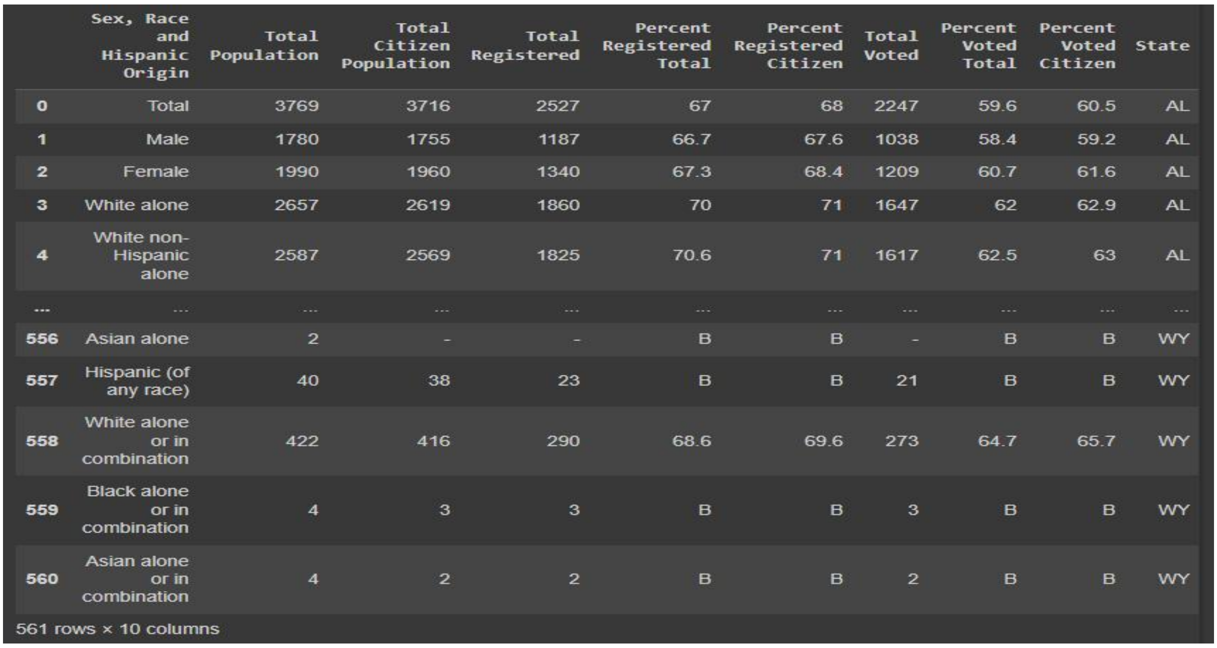
DATA PREPROCESSING

I begin by importing all the necessary libraries and mounting my drive in my Google Colab notebook. The first data file that I look into is the Excel file which contains the information of the demographics across the United States downloaded from the Census.gov website. The data in the original excel file is divided as the demographic distribution throughout the United States as a whole and the state wise demographic distribution of the population. As I read the file in using read_excel() I realize that the data table has a lot of empty spaces filled with NaN values which makes it difficult to understand. I decided to split the data into two completely different data frames that will contain the demographic distribution of the United States as a whole and the demographic distribution for each state. I locate and drop all the unnecessary rows and columns. I rename the columns as per the original file and assign them and the first 11 rows to the us_demo data frame containing the data for the entire country. Below is the image of the cleaned data frame.



	Sex, Race and Hispanic Origin	Total Population	Total Citizen Population	Total Registered	Percent Registered Total	Percent Registered Citizen	Total Voted	Percent Voted Total	Percent Voted Citizen
0	Total	252274	231593	168308	66.7	72.7	154628	61.3	66.8
1	Male	121870	111485	79340	65.1	71.2	72474	59.5	65
2	Female	130404	120108	88968	68.2	74.1	82154	63	68.4
3	White alone	195227	181891	134889	69.1	74.2	124301	63.7	68.3
4	White non-Hispanic alone	157442	154827	118389	75.2	76.5	109830	69.8	70.9
5	Black alone	32219	30204	20844	64.7	69	18922	58.7	62.6
6	Asian alone	16094	11530	7354	45.7	63.8	6881	42.8	59.7
7	Hispanic (of any race)	42468	30627	18719	44.1	61.1	16459	38.8	53.7
8	White alone or in combination	199610	185983	137710	69	74	126753	63.5	68.2
9	Black alone or in combination	34471	32275	22241	64.5	68.9	20152	58.5	62.4
10	Asian alone or in combination	17273	12641	8157	47.2	64.5	7593	44	60.1

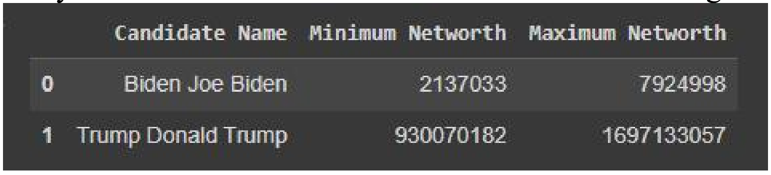
For the state wise distribution data frame I first create a list of abbreviated state names in the order as mentioned in the original data file. The reason I decided to use abbreviations was the fact that plotly only identifies abbreviated state names necessary for creating map plots. I then assign these states names 11 times each over a loop as there are 11 demographic categories to a column called State. Below is the image of the cleaned data frame. It contains around 561 rows and 10 columns in total.



	Sex, Race and Hispanic Origin	Total Population	Total Citizen Population	Total Registered	Percent Registered Total	Percent Registered Citizen	Total Voted	Percent Voted Total	Percent Voted Citizen	State
0	Total	3769	3716	2527	67	68	2247	59.6	60.5	AL
1	Male	1780	1755	1187	66.7	67.6	1038	58.4	59.2	AL
2	Female	1990	1960	1340	67.3	68.4	1209	60.7	61.6	AL
3	White alone	2657	2619	1860	70	71	1647	62	62.9	AL
4	White non-Hispanic alone	2587	2569	1825	70.6	71	1617	62.5	63	AL
...	...	...	...	...	...	...	...	...	...	...
556	Asian alone	2	-	-	B	B	-	B	B	WY
557	Hispanic (of any race)	40	38	23	B	B	21	B	B	WY
558	White alone or in combination	422	416	290	68.6	69.6	273	64.7	65.7	WY
559	Black alone or in combination	4	3	3	B	B	3	B	B	WY
560	Asian alone or in combination	4	2	2	B	B	2	B	B	WY

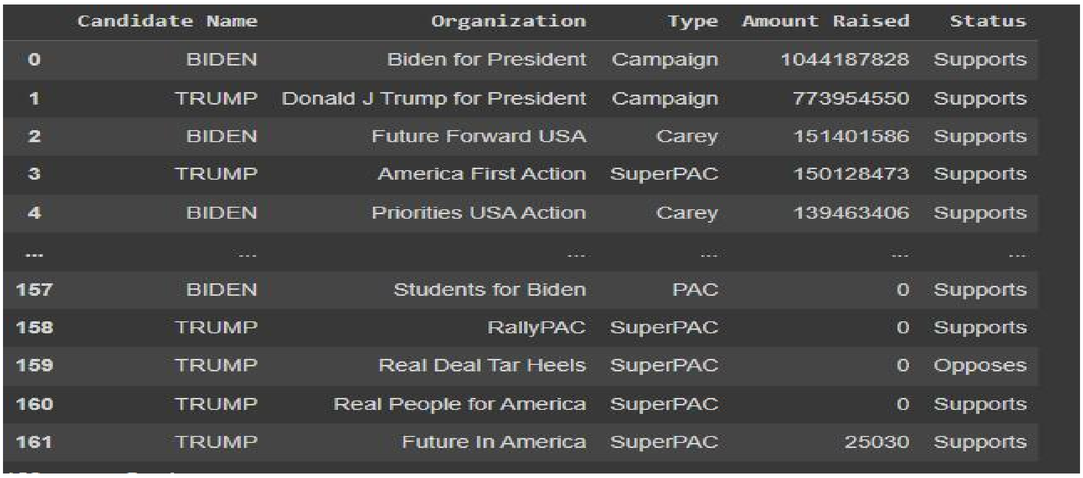
561 rows × 10 columns

Next step would be getting the HTML data from the urls. For that purpose I initially define a function called html_parse() which takes in any url and returns a pandas data frame of the data present in the html file on the website. Inside the function I start off by assigning a user agent so that my http requests are not blocked by the server. I then send a request to open the url and if any occurs during the process it will be printed else it will proceed to the next step which is decoding the data from bytes to strings. Using BeautifulSoup the html string is then parsed and all the data is extracted. I then specifically look for data present in tables in the html file and append all of it to and empty list which will then be finally coverted to pandas data frame. I then get the data for the different units of analysis and put them in different data frames. As all these data frames do not have column names to them I decide to give every data frame column names as per the names available on the web site. I also drop the columns that I will not be requiring for my analysis. I define another set of functions called clean_string, clean_int and clean which will all take in values and remove any non alphabetic, no numeric, $#,- signs present in the data as we are dealing with values in dollars. I apply this function across all data frames in order to remove the $ and , from the values and select only the rows of candidates that will be necessary for my analysis. For the net worth data frame below is the image of the cleaned data frame.



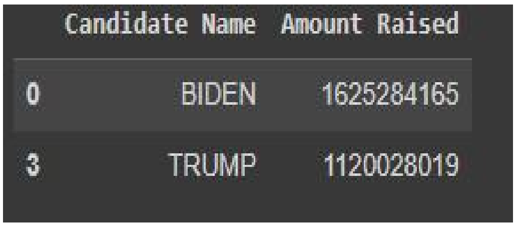
	Candidate Name	Minimum Networth	Maximum Networth
0	Biden Joe Biden	2137033	7924998
1	Trump Donald Trump	930070182	1697133057

For the funds raised data frame after performing the above steps this is what the data frame looks like.



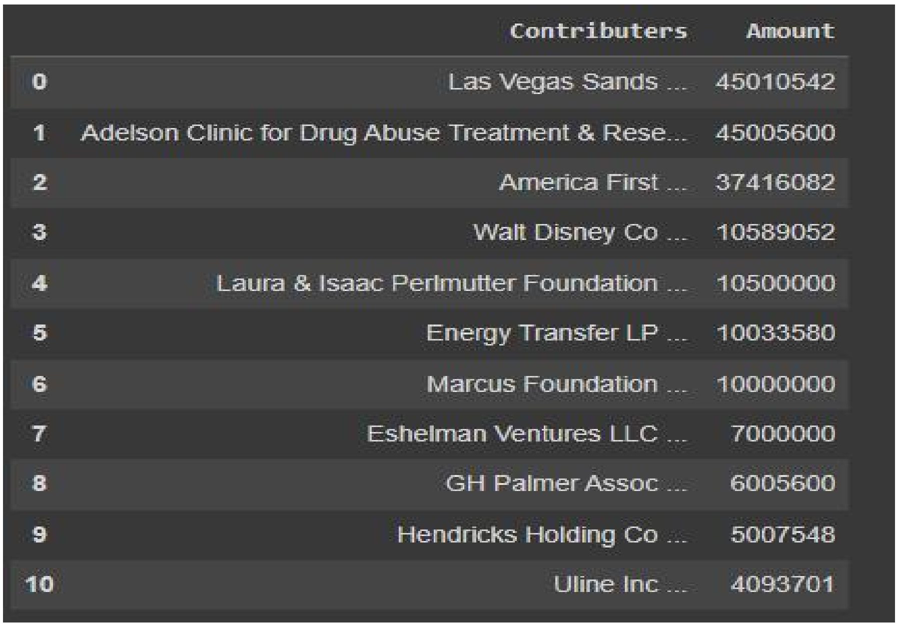
	Candidate Name	Organization	Type	Amount Raised	Status
0	BIDEN	Biden for President	Campaign	1044187828	Supports
1	TRUMP	Donald J Trump for President	Campaign	773954550	Supports
2	BIDEN	Future Forward USA	Carey	151401586	Supports
3	TRUMP	America First Action	SuperPAC	150128473	Supports
4	BIDEN	Priorities USA Action	Carey	139463406	Supports
...	...	...	...	...	...
157	BIDEN	Students for Biden	PAC	0	Supports
158	TRUMP	RallyPAC	SuperPAC	0	Supports
159	TRUMP	Real Deal Tar Heels	SuperPAC	0	Opposes
160	TRUMP	Real People for America	SuperPAC	0	Supports
161	TRUMP	Future In America	SuperPAC	25030	Supports

I then divide the data frame into two separate ones for each candidate as I wanted to get the different committees which helped raise the funds. And finally to get the total amount raised I group by the candidates name and sum the total amount. The resultant data frame looks like



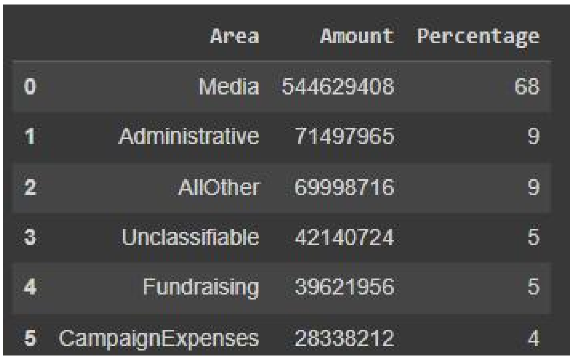
	Candidate Name	Amount Raised
0	BIDEN	1625284165
3	TRUMP	1120028019

For the list of contributors to the fundraising amount for both the candidates I again repeat the same procedure of cleaning values and renaming columns.

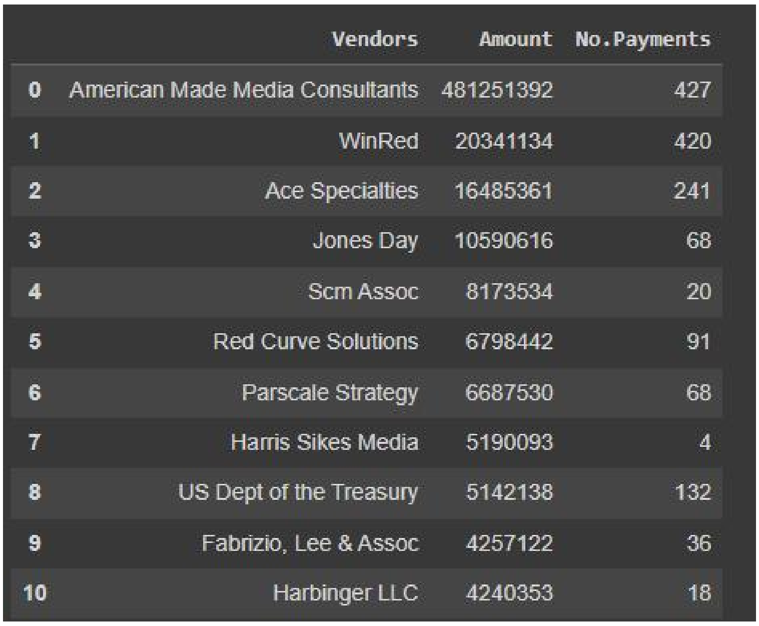


	Contributers	Amount
0	Las Vegas Sands ...	45010542
1	Adelson Clinic for Drug Abuse Treatment & Rese...	45005600
2	America First ...	37416082
3	Walt Disney Co ...	10589052
4	Laura & Isaac Perlmutter Foundation ...	10500000
5	Energy Transfer LP ...	10033580
6	Marcus Foundation ...	10000000
7	Eshelman Ventures LLC ...	7000000
8	GH Palmer Assoc ...	6005600
9	Hendricks Holding Co ...	5007548
10	Uline Inc ...	4093701

For the expenditure data after I repeated the above process I then divided the data for each candidate as the key areas of expenditure and the recipients of the money. Below are the two data frames.

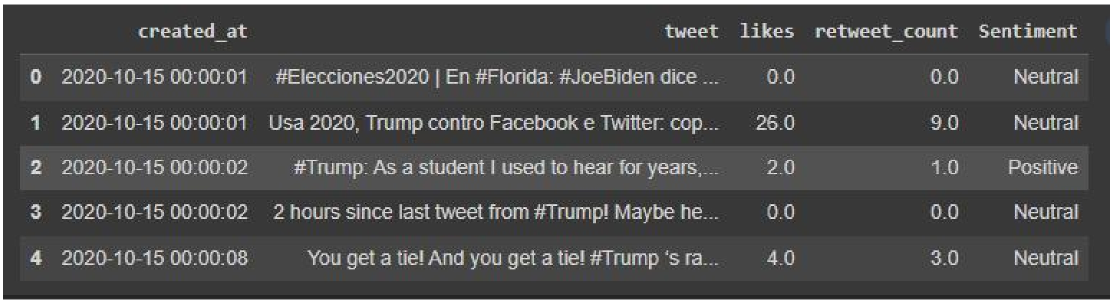


	Area	Amount	Percentage
0	Media	544629408	68
1	Administrative	71497965	9
2	AllOther	69998716	9
3	Unclassifiable	42140724	5
4	Fundraising	39621956	5
5	CampaignExpenses	28338212	4



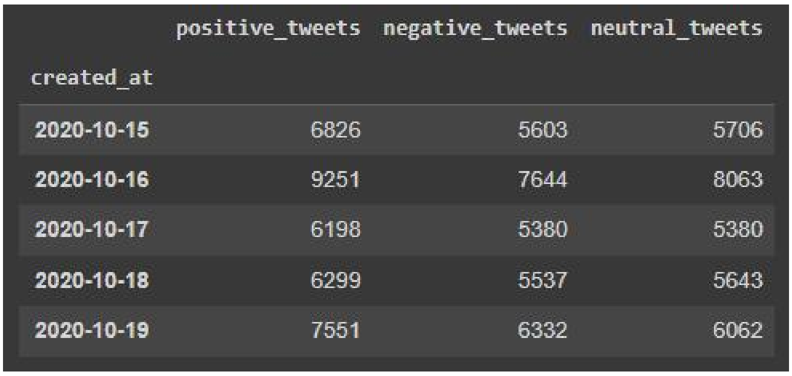
	Vendors	Amount	No.Payments
0	American Made Media Consultants	481251392	427
1	WinRed	20341134	420
2	Ace Specialties	16485361	241
3	Jones Day	10590616	68
4	Scm Assoc	8173534	20
5	Red Curve Solutions	6798442	91
6	Parscale Strategy	6687530	68
7	Harris Sikes Media	5190093	4
8	US Dept of the Treasury	5142138	132
9	Fabrizio, Lee & Assoc	4257122	36
10	Harbinger LLC	4240353	18

Then I move on to the next segment of analysis where I look into the Twitter sentiment for each candidate around the time of the elections. I have used csv files which were available on Kaggle to get tweets for each candidate. After reading in the files I decide to just select only the columns that are going to be useful out of all the 21 columns that will be the date, tweet text, likes and retweet count. I then look for any null values in the data frames and remove all no alphabetic characters from the tweets texts and put them in two separate lists for each candidates. For assigning the sentiment score to the tweet I will be using a pre-trained Sentiment Analyszer tool called Vader Sentiment which has a lexicon of words and phrases with a score assigned to them and on the basis of that score we get the overall calculated score for any text. I create a list called sent list which will contain the sentiment score i.e positive, negative or neutral for each tweet in the tweet list created above. This list is then assigned to a new column called Sentiment in each candidate's data frame. This entire process took around 6 hours so I decided that I will save and download this csv file for future ease of use.



	created_at	tweet	likes	retweet_count	Sentiment
0	2020-10-15 00:00:01	#Elecciones2020 | En #Florida: #JoeBiden dice ...	0.0	0.0	Neutral
1	2020-10-15 00:00:01	Usa 2020, Trump contro Facebook e Twitter: cop...	26.0	9.0	Neutral
2	2020-10-15 00:00:02	#Trump: As a student I used to hear for years,...	2.0	1.0	Positive
3	2020-10-15 00:00:02	2 hours since last tweet from #Trump! Maybe he...	0.0	0.0	Neutral
4	2020-10-15 00:00:08	You get a tie! And you get a tie! #Trump 's ra...	4.0	3.0	Neutral

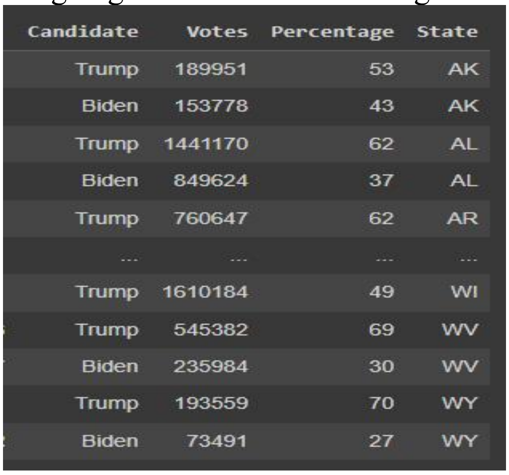
Above is the glimpse of what the new cleaned files look like. I again read these files in look for null values and drop them and decide to group the data frame to get a total count of the positive, negative and neutral tweets for each given day for both the candidates.



	positive_tweets	negative_tweets	neutral_tweets
created_at			
2020-10-15	6826	5603	5706
2020-10-16	9251	7644	8063
2020-10-17	6198	5380	5380
2020-10-18	6299	5537	5643
2020-10-19	7551	6332	6062

I then merge the original data frame and the grouped data frame together and proceed to create a new column which will display the adjusted sentiment score for each day by subtracting the total negative tweets from the positive tweets and dividing them by the overall total tweets for both the candidates.

And for the final part where I want to display the election results that I get from the foxnews website I repeat the entire procedure of getting the data using the html_parse function defined, cleaning the data to remove all non alphabetic characters and assigning states for each row to get the final cleaned data frame.

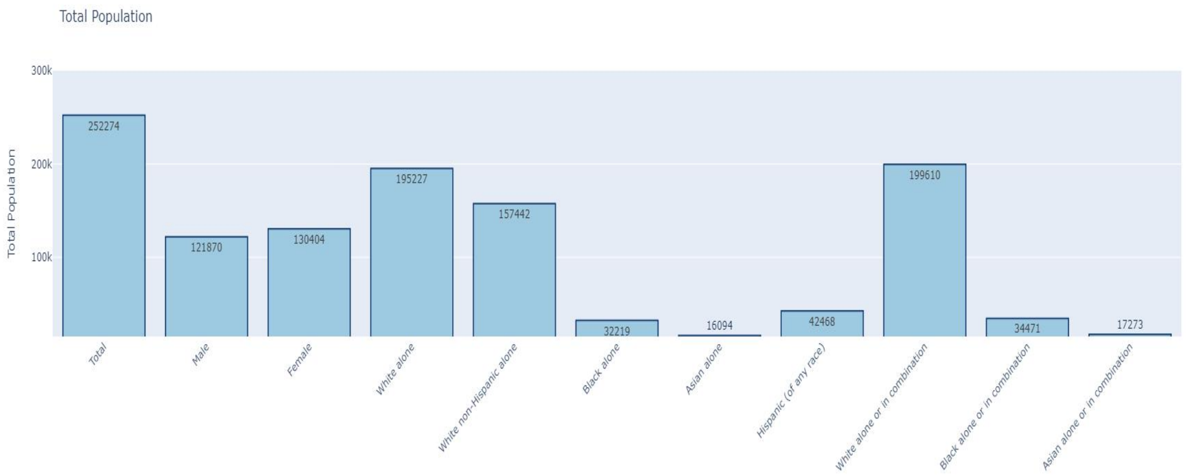


Candidate	Votes	Percentage	State
Trump	189951	53	AK
Biden	153778	43	AK
Trump	1441170	62	AL
Biden	849624	37	AL
Trump	760647	62	AR
...	...	...	...
Trump	1610184	49	WI
Trump	545382	69	WV
Biden	235984	30	WV
Trump	193559	70	WY
Biden	73491	27	WY

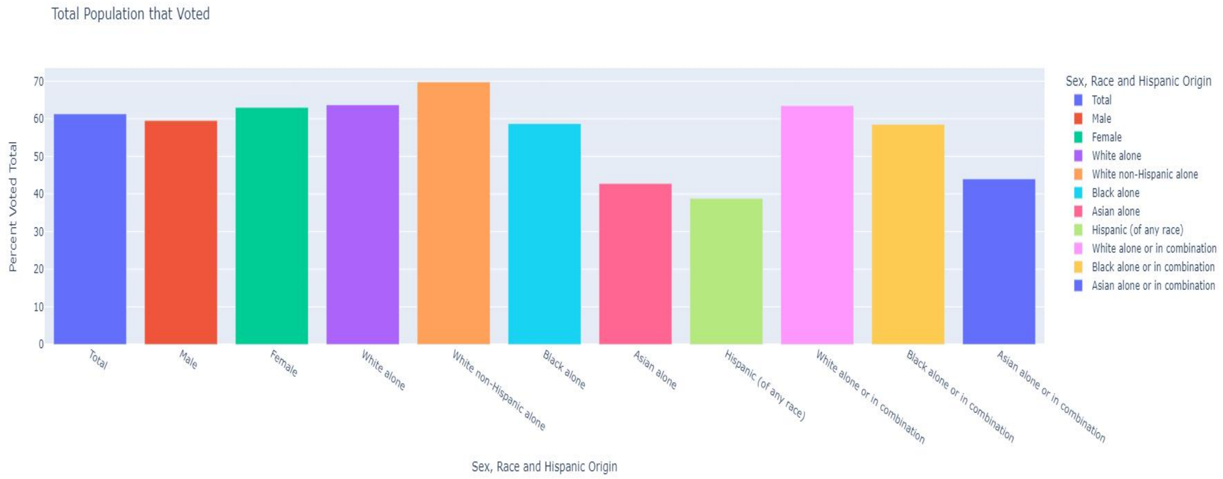
DATA VISUALIZATION AND INTERPRETATION

For the visualization of plots I will use Plotly as the plots created are visually appealing and interactive. We now go through all of the plots created using the above data frame.

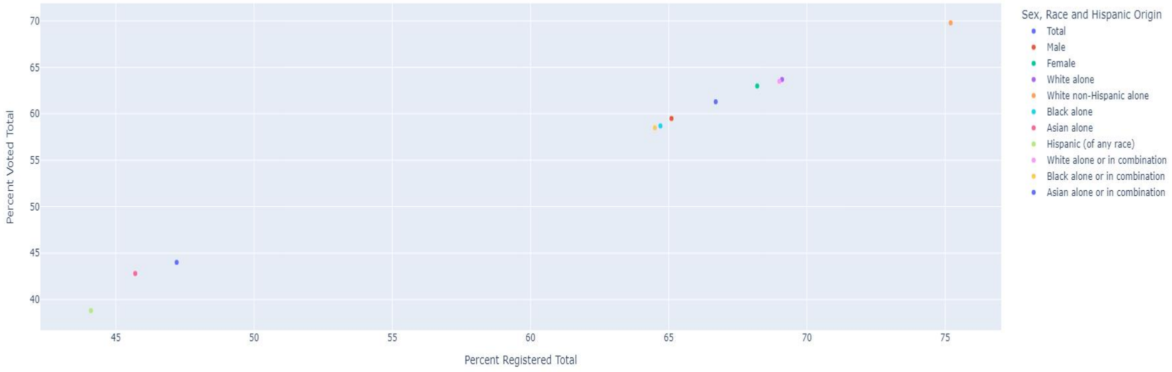
DEMOGRAPHICS:



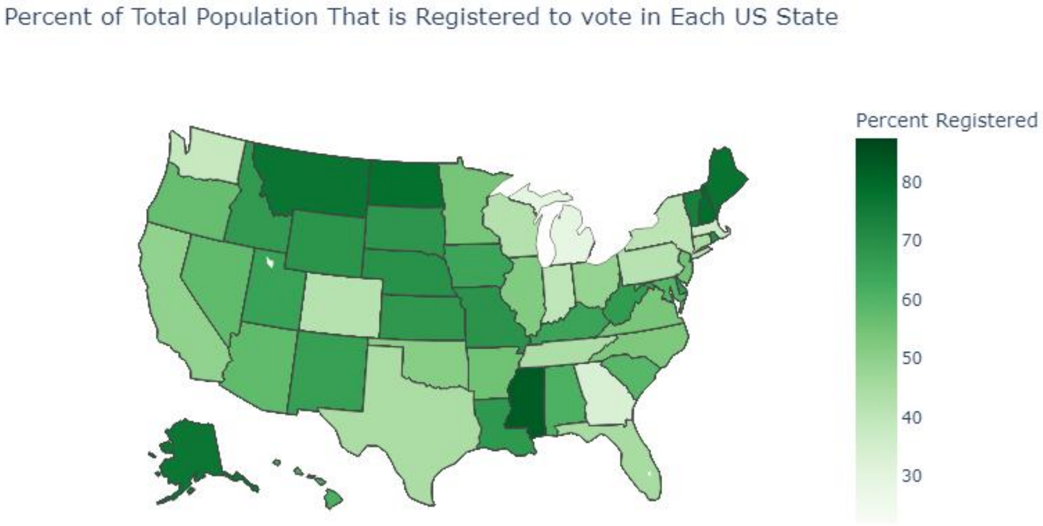
As we can see from the plot above the total female population exceeds the male population and White alone seems to be the category with the highest population while Asian alone is the lowest.



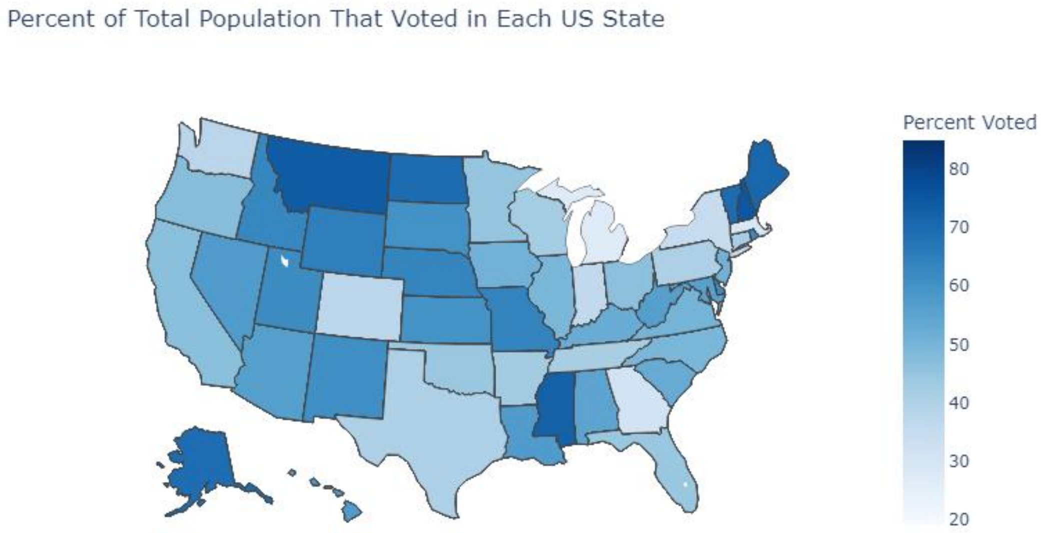
Looking at the total population that voted females voters seem to be more in number than the male voters. Hispanic category seems to have the least voting population.



Looking at the scatter plot for the relation between the category of demographic that is registered to vote and the ones that actually voted we observe that Hispanic seems to have the least registered voters and also the actual voters while the highest will be White non-Hispanic, even exceeding the total population.

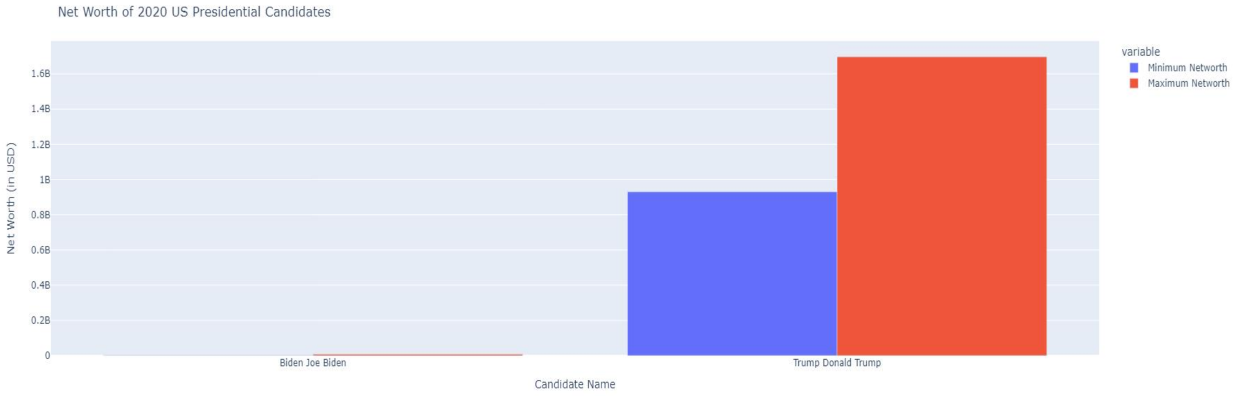

Percent of Total Population That is Registered to vote in Each US State

When we look at the state wise distribution of the registered population we can observe that Montana and North Dakota have the highest percentage of registered voter population.



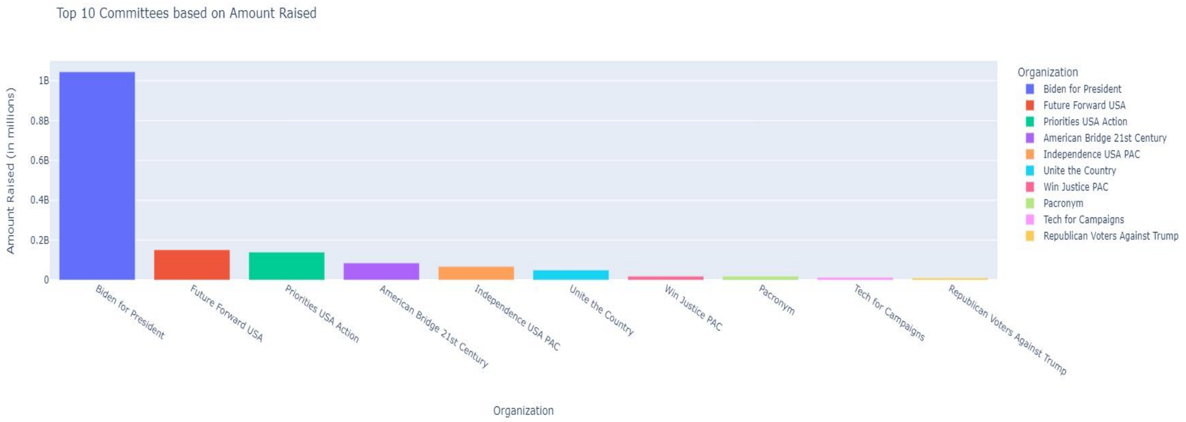
Where as here for the actual voter population we can see that Montana and Mississippi seem to have the highest turnout.

NET WORTH:

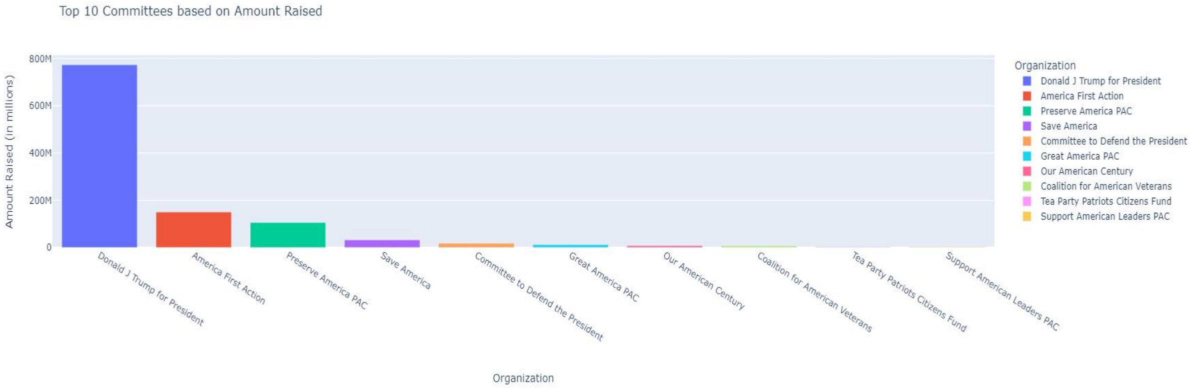


Clearly Donald Trump has a net worth much higher than Joe Biden due to his several business ventures.

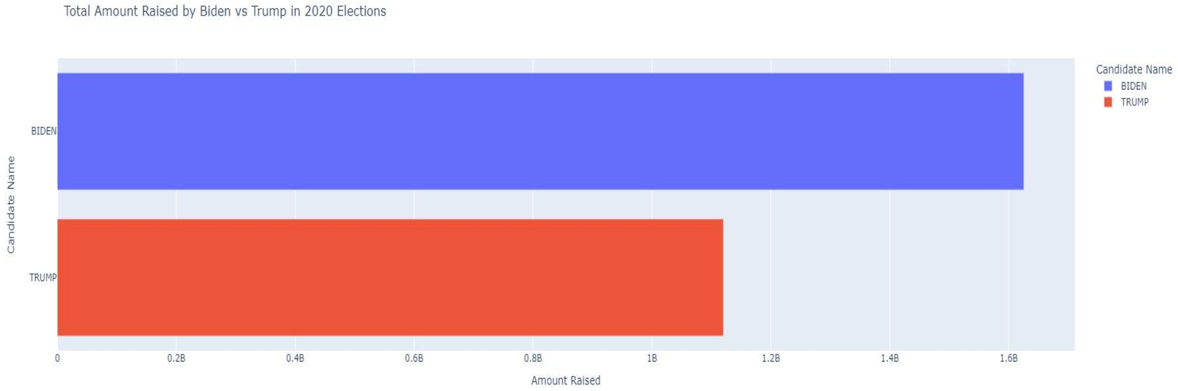
FUNDS RAISED:



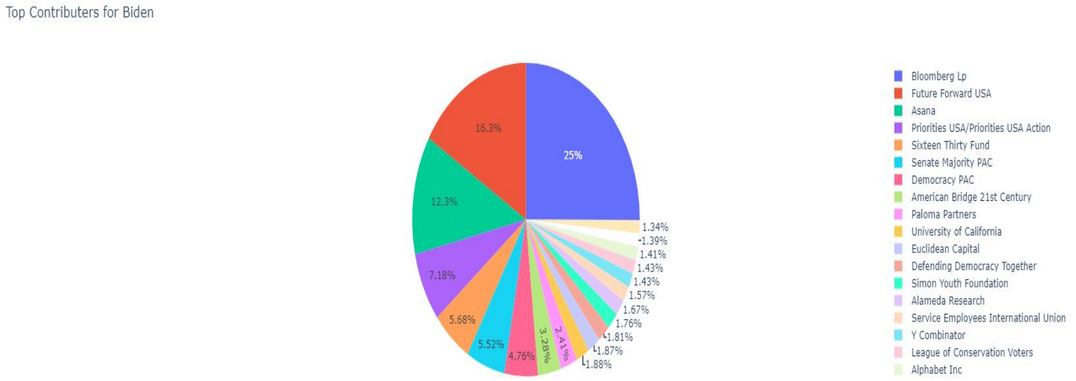
Looking into the top 10 committees that raised the highest funds for Biden we can see that Biden for President has the highest amount raised and managed to raise more than a Billion dollars



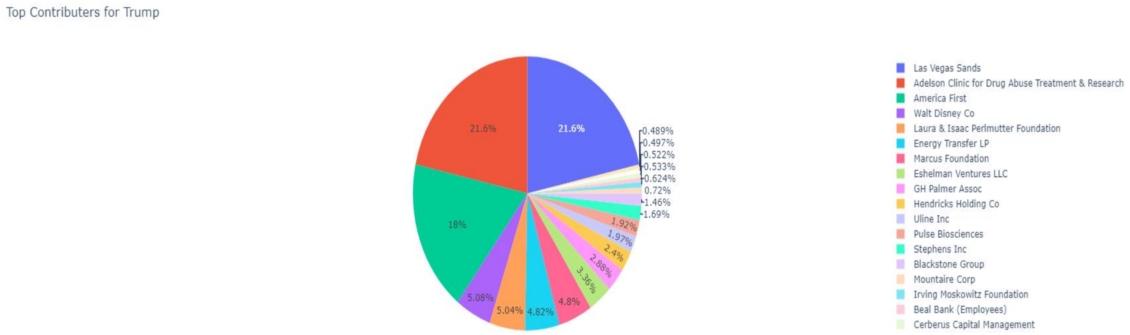
Looking into the top 10 committees that raised the highest funds for Trump we can see that Donald J Trump for President has the highest amount raised and managed to raise close to 800 million dollars



Comparing the overall amount raised by both the candidates Joe Biden manages to raise more amount than Donald Trump.

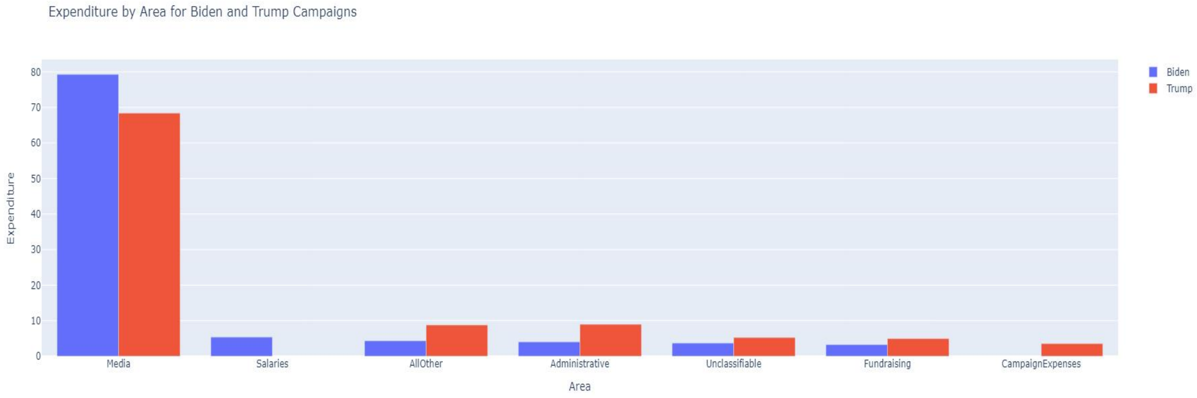


Above we look at the top 20 contributors to the fundraising committee for Joe Biden.

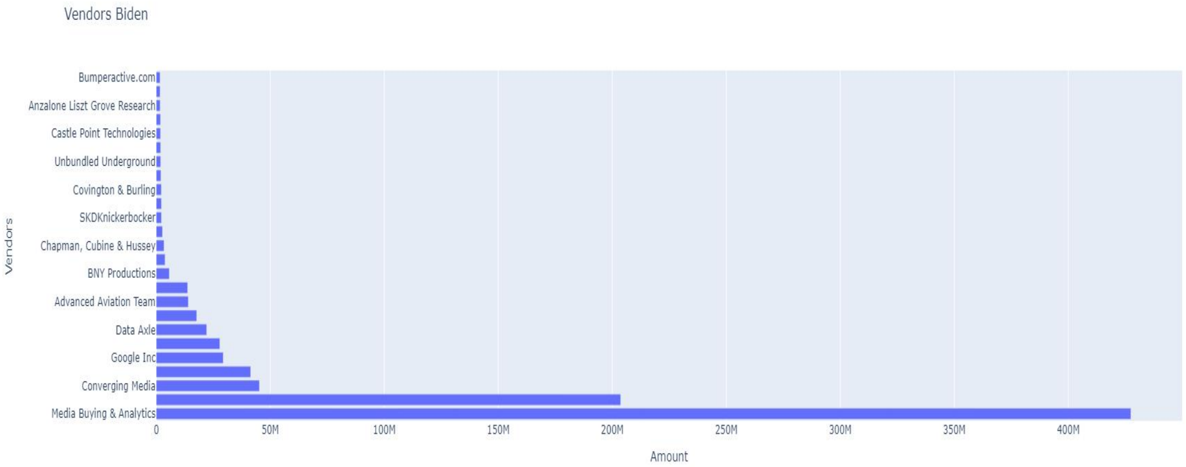


Above we look at the top 20 contributors to the fundraising committee for Donald Trump

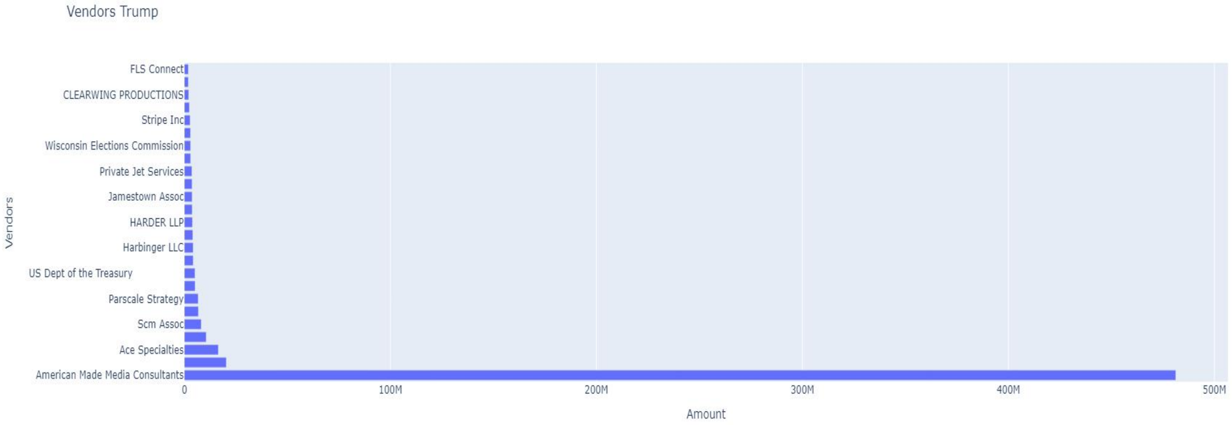
EXPENDITURE:



The key area of expenditure for both the candidates seems to be Media

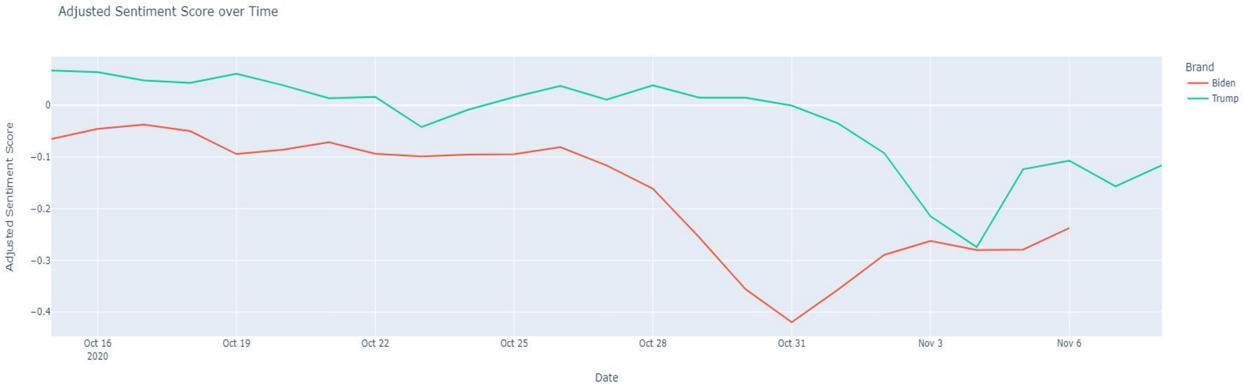


Above are the top recipients of the amount for Joe Biden and we can see that majority of it has been spent on media and advertising agencies.



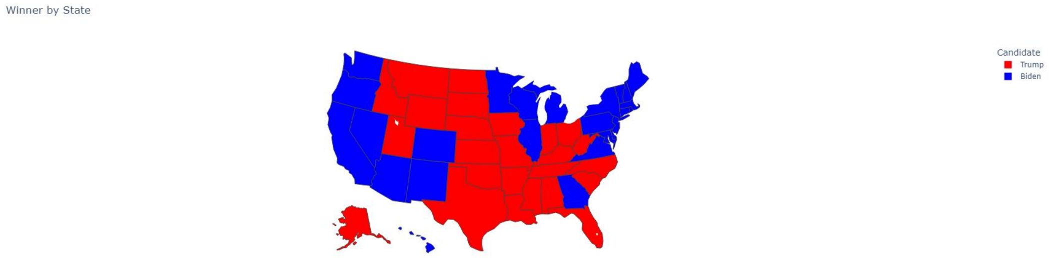
Above are the top recipients of the amount for Donald Trump and we can see that majority of it has been spent on media and advertising agencies.

TWITTER SENTIMENT:



The overall sentiment on Twitter during the election period seems to more negative for Joe Biden than Donald Trump.

ELECTION RESULTS:



Above plot represents the state wise winners of the 2020 presidential elections. Joe Biden was declared winner over Donald Trump.

CONCLUSION

- We can see that Biden's campaign and supporting committees raised more money than Trump's by comparing the amounts raised by each candidate and their supporting committees. This indicates that he had access to more financial resources to support his campaign efforts.
- According to our analysis in 2020 voter turnout was 68.4% for women and 65.0% for men.
- Bloomberg Lp and Las Vegas Sands were the highest contributors for Biden and Trump respectively.
- Majority of the money spent by both candidates was on media and advertising, which includes TV and digital ads. Campaigns that invest more in media and advertising can more effectively communicate their message to voters, which can sway their opinion and ultimately determine the outcome of the election.
- From our Twitter sentiment analysis, we can see that even though Biden had a lot more negative sentiment than Trump he still won the elections. This could possibly be because of the media covering the negative aspects of the campaign, people critical of his policies were more active on Twitter. This certainly tells us that sentiment analysis may have limitations and might not necessarily be accurate.
- Having a higher net worth might not guarantee victory in the elections.
- Every upcoming candidate needs to understand the importance of the role that media plays during elections and thus plan on spending more in that particular area as it may lead to possible victory.