**DATA7201 Data Analytics at Scale**

**Project Report**

**Jaskeerat Singh**

**47610039**

**Abstract**

The project aims to implement big data techniques to analyze Facebook Ad sponsored datasets. This dataset is a collection of Ad posts over 23 months. As the quantity of data available is substantial, we will use PySpark for data exploration and implementing preprocessing techniques to uncover trends and patterns in the data. A high-quality subset of data for the primary analysis will be produced through this. Analysing this dataset will make it easier to gauge how the sponsored Ads influenced voter participation in the US presidential election. The report will look at Ad volume, target population distribution, expenditure and funding entities of different ad campaigns. This way, we can tell where and how the political posts targeted the general audience.

**Table of content**

1. **Introduction**
   1. General area of big data analytics ……………………………………… 3
   2. Motivation for distributed system solutions ……………………………. 3
2. **Dataset Analytics**
   1. Pre-Processing …………………………………………………………. 4
   2. Main Analysis …………………………………………………….……. 6
3. **Summary and conclusion**
   1. Summary Discussion ………………………………………...…………. 11
   2. Conclusion …………………………………………………………….... 11
4. **Reference**
5. **Appendix**
6. **Introduction**
   1. **General area of big data analytics**

Big data analytics is defined as the process of collection, examination and analysis of large and complex data using a group of resource sharing nodes based on commodity hardware. It focuses on discovering patterns and that can help in better decision-making and improve our services' quality. Big Data Analytics have a major impact on all aspects of our life be it marketing, finance, healthcare, IT/ITES and sports and entertainment.

Big data involves the ingestion of data from different sources, transforming and pushing it forward for further analysis. This data from different sources can be structured, unstructured or streaming in nature. The data received in real-time, e.g., from IoT devices, can become redundant by the time it is processed and stored. The amount of data processed and stored surpasses the hardware limitation day by day. This is evident from the increasing problem of handling multimedia data in the form of video, images, signals etc. The data from various systems can be of varying quality and requires the usage of analytical techniques to help improve the data quality and integrity. After processing comes the main task of deriving values from the data collected. This value can help in decision-making, improve product quality and help in generating value for the business.

This whole process can be applied and leveraged for tasks involving Predictive Analytics, Streaming Analytics, sentiment Analysis and much more.

* 1. **Motivation for distributed system solutions**

A distributed system is a collection of multiple interconnected nodes that work together to achieve a shared task. These nodes use the network to communicate and coordinate actions to achieve the shared task.

When we are working on extensive data and the processing power of our system is insufficient to process our dataset, we can move forward with vertical or horizontal scaling. Vertical scaling means improving the hardware of our current system, but the system acts like a single point of failure, and the new, improved hardware is expensive. However, with horizontal scaling, we can add nodes to the chain. Here we can use the processing power and memory of the new nodes to distribute and process the data more optimally. An excellent example is the cloud infrastructure like AWS, where on-demand or dynamic allocation of resources takes place.

Distributed systems help improve our system's availability and fault tolerance by incorporating multiple nodes. This improves the system's performance by removing single point of failure. For example, if a website is hosted on a single node, then the availability of that website to the user depends upon the availability of that node.

A distributed system is heterogeneous. The system should be able to operate on commodity hardware by pooling its available resources. This helps a distributed system run parallel processes on different systems. For example, docker containers which applies the same environment on any underlying hardware.

In our system, the dataset is hosted on Hadoop Distributed File System. The files related to the project are distributed and replicated on the storage of different nodes. This makes it easier for the user to access and ensure the files are always available. Hence making the entire system fault tolerant. The files are partitioned on the storage for faster access and to ensure the network traffic is not clogged at a singular point.

SPARK, which encompasses multiple data processing frameworks like MLlib, Graphx, Spark Streaming and Spark SQL, helps user to process different types of data. MLlib gives users the ability to perform machine learning tasks on the data. Spark Streaming is developed to handle real-time or near real-time time data handling capabilities. Spark SQL is available to work with large amounts of structured data. Graphx is available to users for handling graph-based data like networks.

1. **Dataset analytics**

The dataset used in this project for big data analysis is a collection of Facebook Ad JSON files collected over 23 months from 03/2020 - 01/2022. This Ad data can be used to gain insight into measuring the influence of political parties on the different demographic of users. For data exploration and configuring pre-processing techniques, all the dataset columns will be considered.

For our Main analysis, we will look into the data from pages of US presidential candidates - Joe Biden and Donald Trump, during 2020 (the presidential election year). I would particularly be interested in the data of the months leading up to the election. Hence a subset of the total data is filtered and stored as a spark SQL view. Here we will be visualizing information like target demographic like age range, gender and regions in the US. We will also look at the themes and the views shared in the Ads posted on different pages.

Columns used in Analysis:

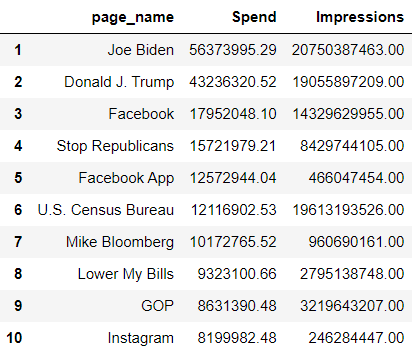
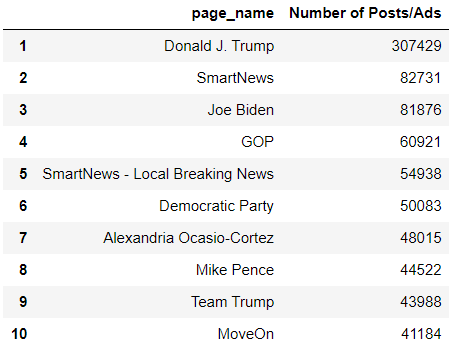
|  |  |
| --- | --- |
| **Column** | **Description** |
| ad\_creative\_body | Contains the body or content of the ad. |
| ad\_delivery\_start\_time | The start time when the ad campaign begins. |
| demographic\_distribution | Includes details such as age, gender, and the percentage distribution. |
| funding\_entity | Entity or organization responsible for funding the ad campaign. |
| id | Unique identifier for each ad. |
| impressions | The range of impressions generated by the ad campaign. |
| page\_name | Name of the page belonging to the ad campaign. |
| region\_distribution | Percentage of ad reach in a specific region. |
| spend | Lower and upper bounds of the spending range of the Ad campaign. |
|  |  |

* 1. **Pre-Processing**

The data available to us is stored in multiple files present inside HDFS. The data is present in JSON files and is structured in nature. For this purpose, PySpark and Spark SQL will be used to manipulate the data.

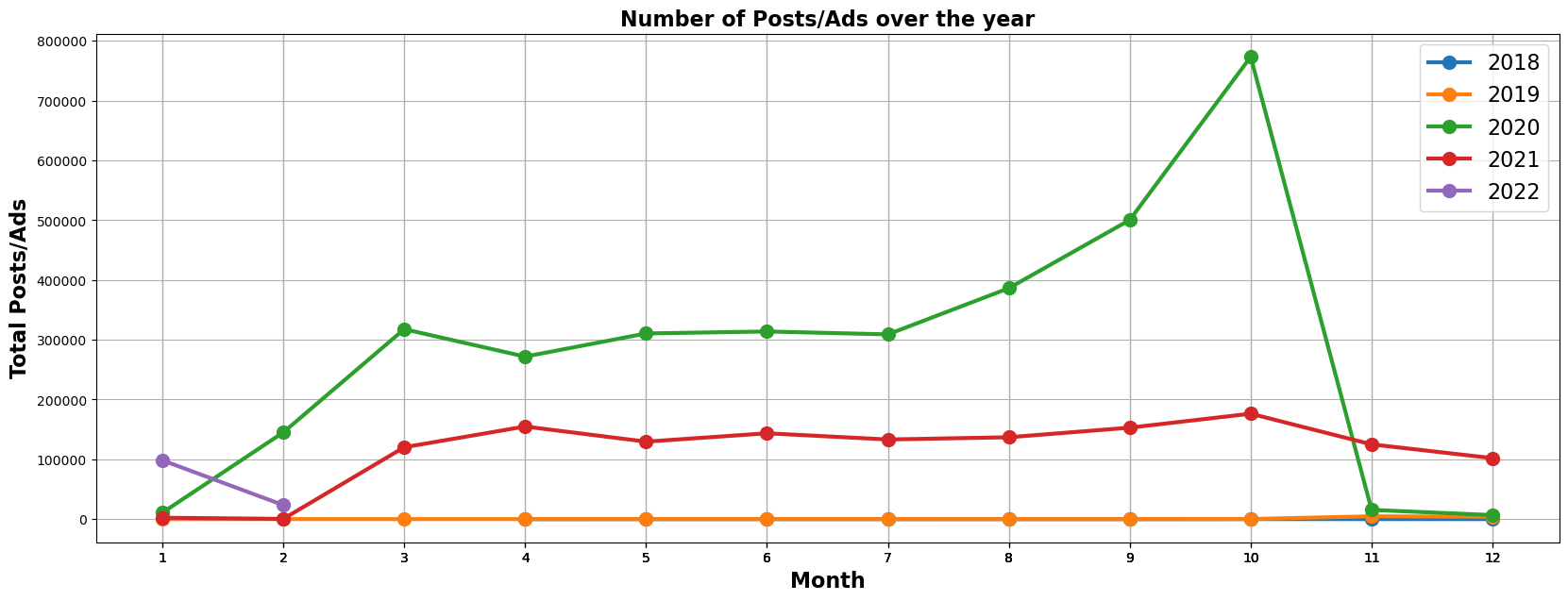
We will first calculate and drop columns with more than 50% of nulls for data cleaning. We will now remove any duplicate rows in our dataset. Post this feature engineering is implemented to produce columns like average spend, average impressions, file date, age, region and gender.

The Primary Exploration and analysis will aim to extract information about different pages and funding organisations. The study will show how much the page spends and how many impressions they got during the 23 months.



**Figure 1:** Top 10 Pages with highest Spend, Impressions and Number of Posts/Ads

From the above rankings, we can see that pages with higher spending and impressions use more ads. Except for the News Handles having the highest number of posts, Pages related to politicians and their supporting pages got the highest impressions. Pages of Joe Biden and Donald Trump spent the most by showing more ads and hence got many impressions.



**Figure 2:** Total Posts/Ads over the year

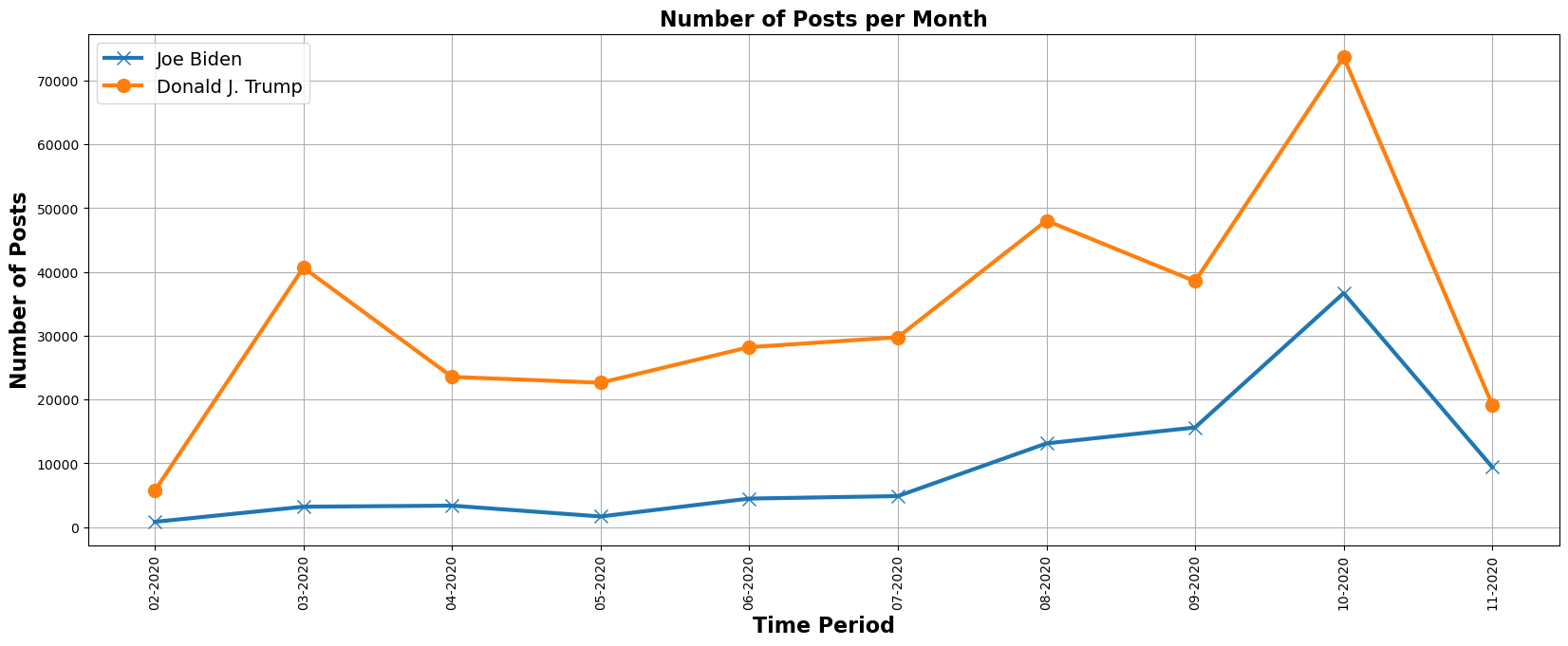
Here we can see that the Posts/Ads every Month over the year. We can see a large uptick in the posts of October 2020. This coincides with the election period in US. We will be taking a deeper look at the data in this period for our main analysis.

* 1. **Main Analysis**

After primary exploration and analysis, we can gain some insight into the trends of the major political parties and organisations. We will now take a deeper view by closely analysing the pages of Donald Trump and Joe Biden during the US election year of 2020.

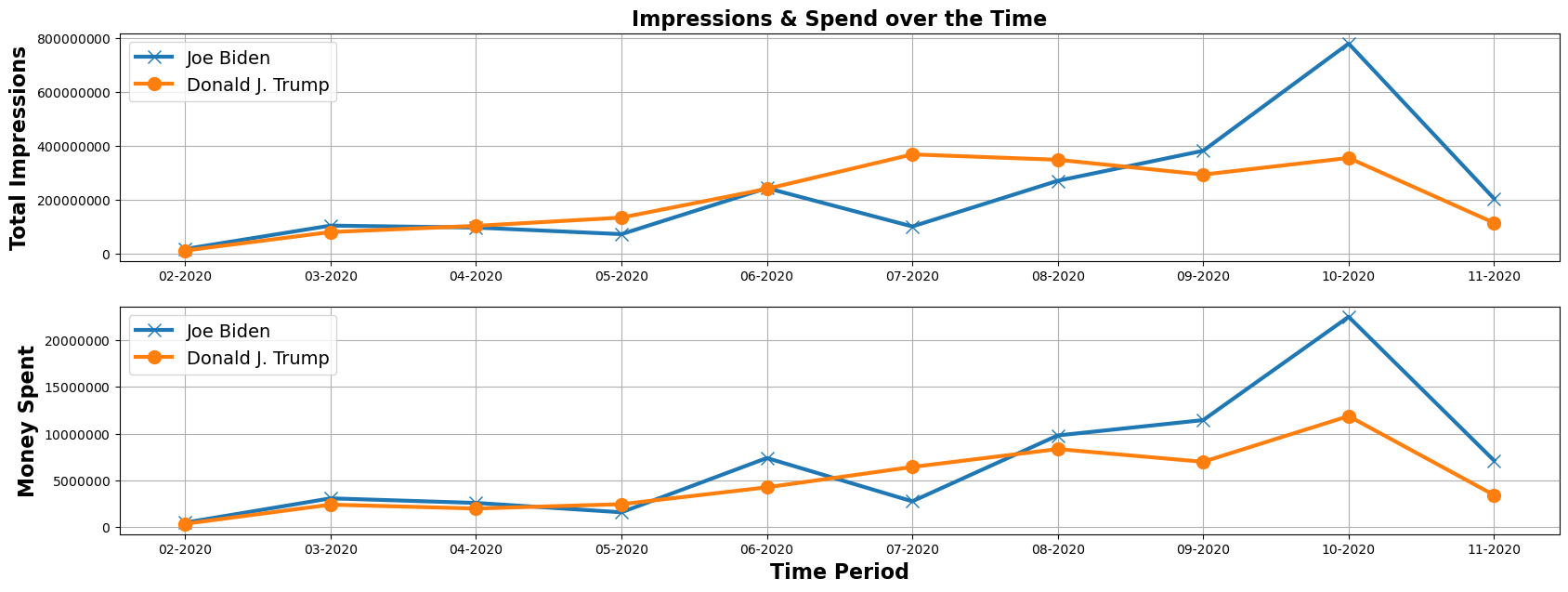
Steps taken to achieve this analysis –

* Read the available data
* Filter the data for the pages Donald Trump and Joe Biden
* Filter the data for the year 2020
* Group by the selected fields
* Save the results in a pandas dataframe
* Pass the dataframe into the matplotlib or seaborn python module



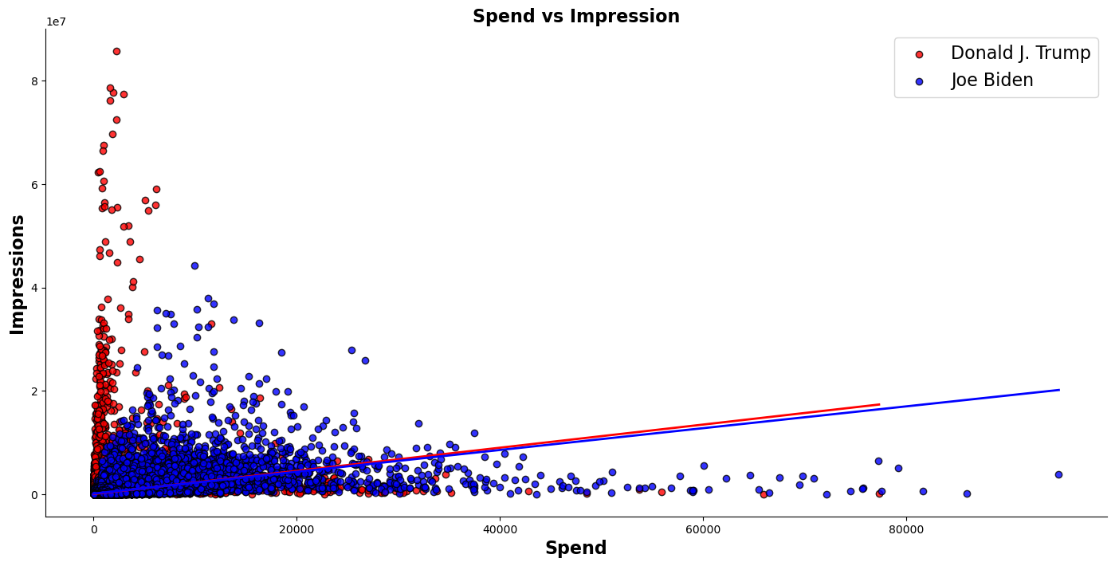
**Figure 4:** Number of Posts/Ads supporting Donald Trump and Joe Biden

In the plotted graph, we can see over the months the number of posts posted by each candidate in 2020. We can see a significant uptick in the number of posts for each candidate in August 2020 when both candidates got assurance of representing their respective political parties. We can again see a significant uptick in October 2020 when the election would take place in the US.



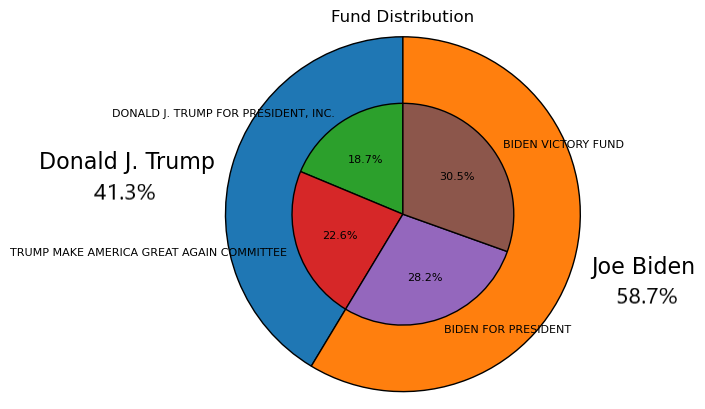
**Figure 5:** Money Spent and Impressions received by Donald Trump and Joe Biden

In the dual chart, we can see the spend and impressions of each candidate over the months. The general trend of increasing the number of Impressions and spending over the last 6 months of the election is true for both candidates. In August, we can see that both the candidates were neck in neck regarding money spent and interactions received. Joe Biden's money spent on Ad campaigns increased after August, leading to more interactions with the general audience.



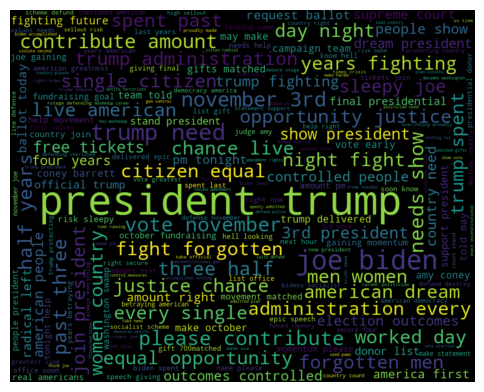
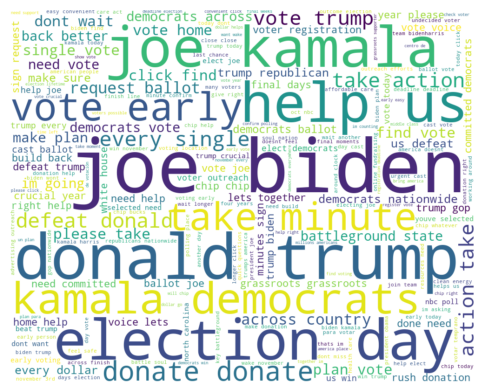
**Figure 6:** Trend Line plotted for Spend vs Impressions

The above figure shows that most of the sponsored posts for the 2 candidates were clustered in the low-spend range. And such posts managed to get a large number of interactions. The trend line of spend vs impression for both candidates is quite similar. Most of the posts spent fell between the 0 to 20000 USD range.



**Figure 7:** Nested Pie chart for Funding Entities

Here we can see the spending distribution of the funding entities for both candidates. Donald Trump spent less in total on Ad campaigns than joe Biden. The funding entities spending money on the campaign of Donald Trump spent around 40% of the total spend by both the candidates while it was 60% for Joe Biden.

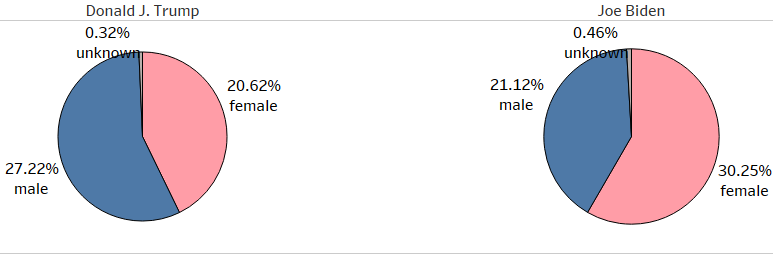


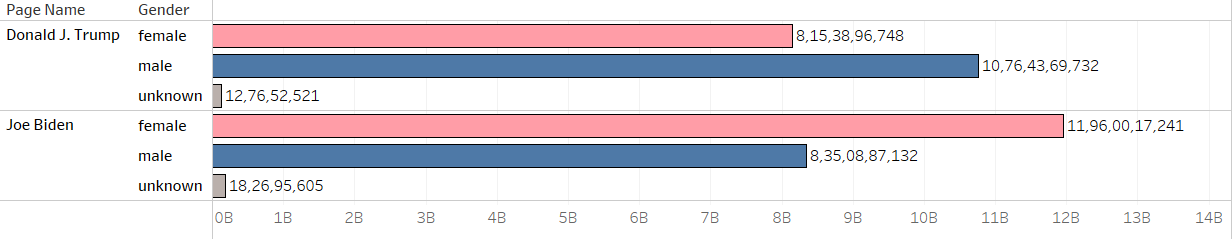
**Figure 8:** Word Cloud for most occurring words

Text Analysis was done to extract the frequently occurring words in both candidates' text bodies of Ads. We can see some common words occurring in the text body of both candidates. Keywords like donate, contribute, and help were seen in the word cloud of the Ad body. Both the candidates urged the audience to vote on 3rd November. Donald Trump talked more about opportunities, Equal Rights, radical left, fake news and American life, while joe Biden talked more about taking action, defeating Trump and the nation.

Steps taken to achieve this analysis –

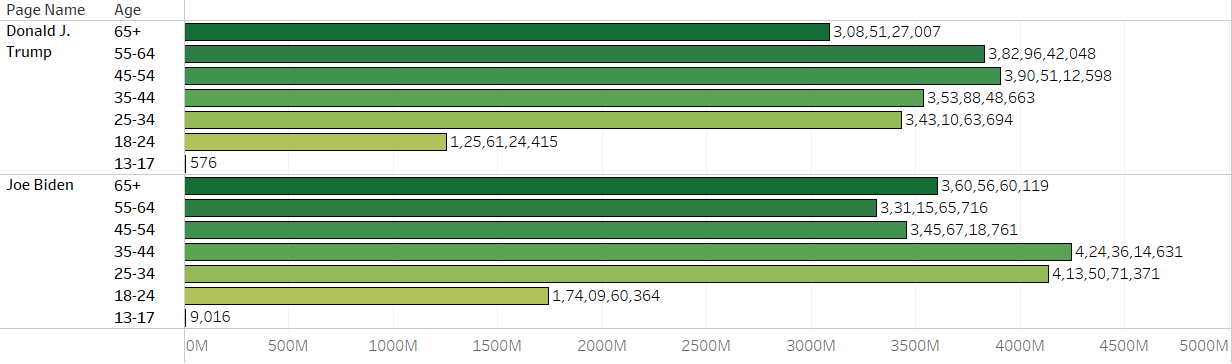
* Calculate the distribution of Age and Gender by multiplying the percentage with impressions
* Group the data by the page name and id
* Save results as pandas dataframe
* Export the dataframe as a CSV file
* load the CSV file into Tableau





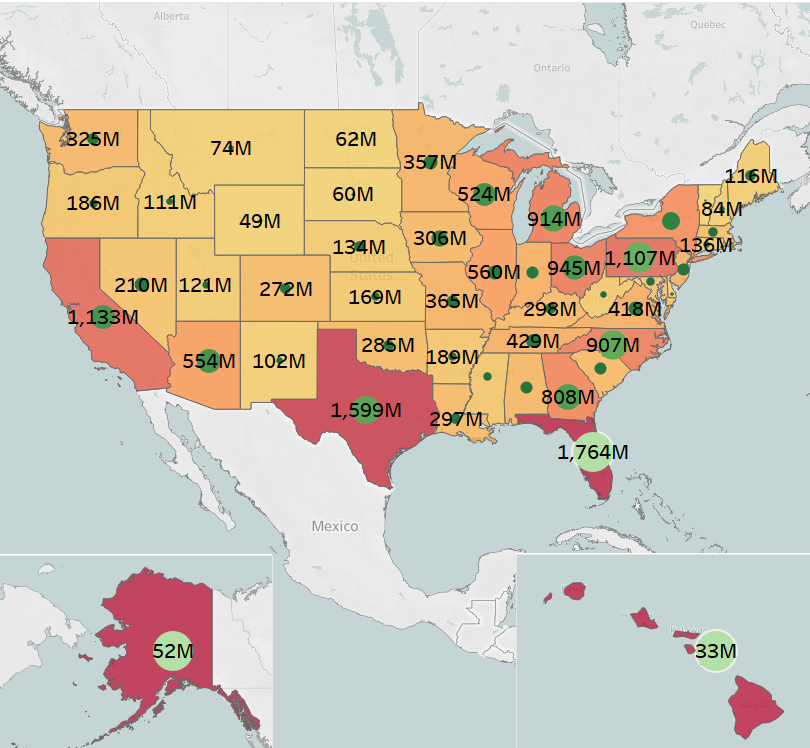
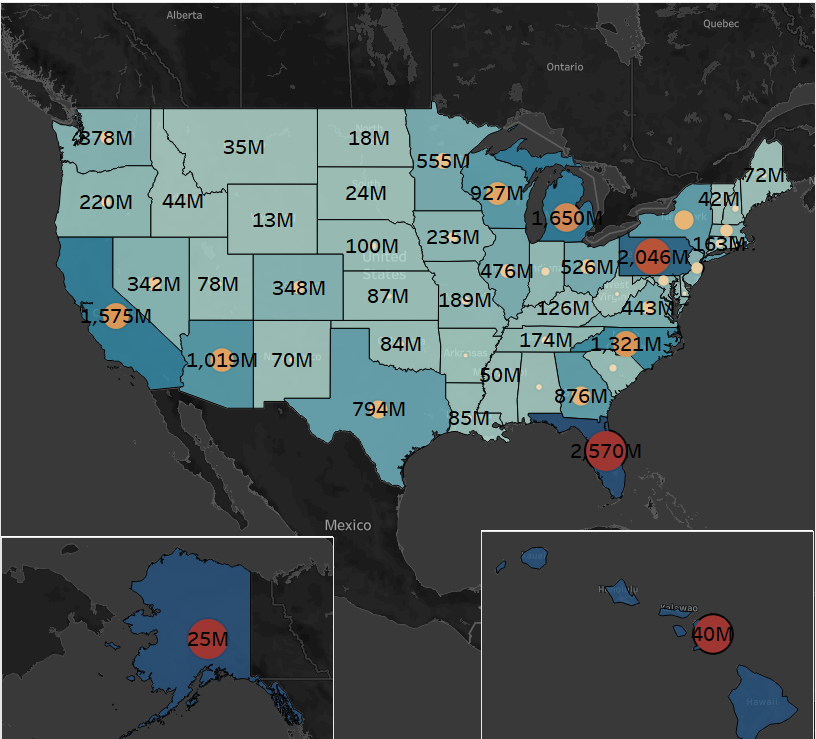
**Figure 9:** Gender Distribution of the Audience

In this chart, we can see that Donald Trump targeted and got more impressions from the Male Audience, while Joe Biden catered more to the Female Population. Overall, Joe Biden got a higher percentage of impressions on his Political Ads.



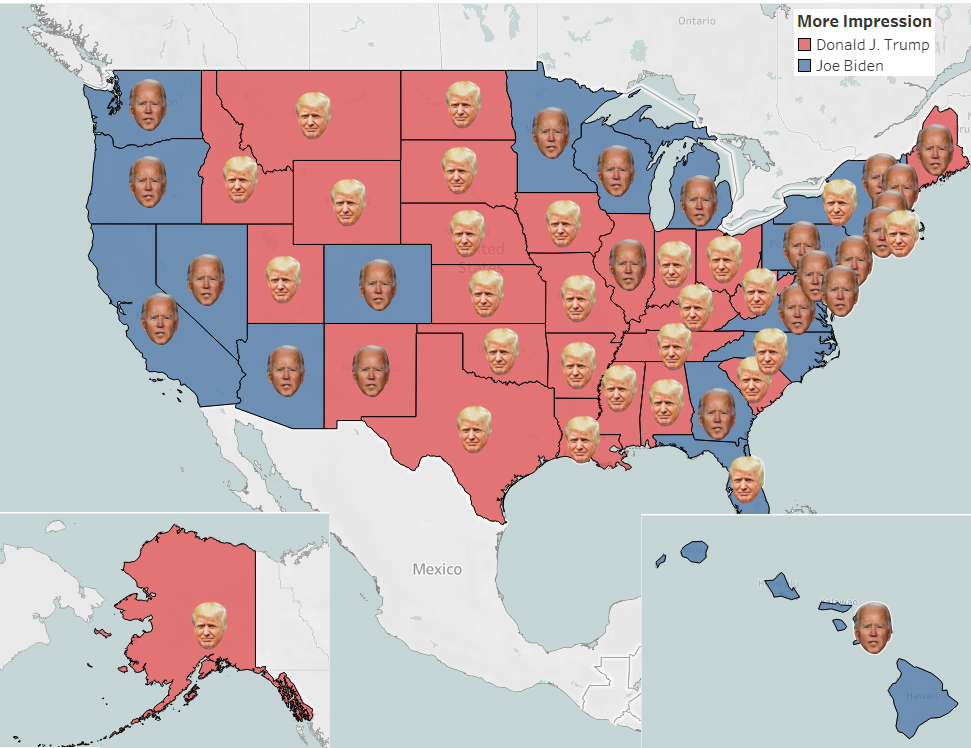
**Figure 10:** Age Distribution of the Audience

In the above chart, we can see that Donald Trump received more impressions and had a more prominent target population belonging to the well-settled adult section of the age range of 45 – 65 years, while Joe Biden got more impressions and support from the younger demographic of the population belonging to 18 to 45 years of age range.



**Figure 11:** Joe Biden and Donald Trump Impressions and Sales distributed by State

From the plotted above maps, we can see the average spending and average impressions of each state in the US leading up to the elections. The number indicating the labels and a darker colour palette signifies an increase in the impressions, while the size of the circle signifies the spending distributed by the region.



**Figure 12:** Greater Impressions Received in State vs Actual Winner in 2020 Election

The Final map signifies the election winner in each state, with the red and blue background signifying the higher number of impressions received in that particular state for that candidate. Red indicates Donald Trump having more impressions in the state, while blue signifies the same for Joe Biden. The face logo signifies the winner of the election in that state. From this, we can infer those higher impressions received in the state typically meant the state's winner. Some exceptions exist, like Joe Biden winning in Utah and Donald Trump winning in North Carolina.

1. **Summary and conclusion**
   1. **Summary Discussion**

The project gives an overview of the insights derived from the Ads of the political campaigners who participated in the 2020 US Election. First, we saw that Joe Biden had more impressions and spent more on posts/Ads for these impressions during the election period. Both the candidates preferred to make posts/Ads in the 0 to 20000 USD range. Both the candidates had 2 major funding entities, with Donald Trump making 40% and joe Biden making 60% of the total funds spent on ads. Through the text analysis of these funded ads, we can see the themes and topics under which the posts were made on Facebook. Both the candidates asked for support and donations and asked citizens to participate in the upcoming elections. Donald Trump's posts talked more about the life of American citizens and the administration. In contrast, Joe Biden talked about Trump destroying America and taking steps to bring America in the right direction. Posts related to Donald Trump gained impressions majorly from the Male gender while it was the woman for Joe Biden. The working sector in the age group of 45 to 65 was the primary audience of Trump, while the younger population was following Joe Biden. In the end, we plotted spend according to the states for both the candidates and which candidate managed to win in what state even after having a large following in that region. An excellent example is Joe Biden winning in New Mexico while Donald Trump winning in North Carolina. Thus, winning the popular vote does not necessarily mean winning the presidency.

* 1. **Conclusion**

This project gives great insight into how the Ads/Posts are used to modify the audience's perception. It also tells us how the political parties use these ads and modify them to target specific sections of society. By leveraging big data analytics, the stakeholders like campaign managers, organisations and funding entities can derive and develop good intuition to enhance their upcoming Ads/Posts, which can help boost and expand their audience base. They use the metrics of these Ad posts to improve user engagement. Using the latest trends and changes, the political parties will target specific sections of society to try and convert them into their loyal voter base.

**WORD COUNT - 2300**

**4 References**

[1] The 2020 election by the numbers. (2020, December 15). Council on Foreign Relations. <https://www.cfr.org/blog/2020-election-numbers>

[2] Presidential election results: Biden wins (Published 2020). (2021, June 1). The New York Times - Breaking News, US News, World News and Videos. <https://www.nytimes.com/interactive/2020/11/03/us/elections/results-president.html>

[3] Ad library API. (n.d.). Facebook - log in or sign up. <https://www.facebook.com/ads/library/api/>