

Postgraduate coursework  
DATA7201 Data Analytics at Scale (2022)

## **Project Report – Report on Dataset Analytics**

**Student Name: Cheng-Yu Wu**

**Student ID: S4623099**

## Structured abstract

**Analytics the relationship between US FB political ads and its region distribution.**

### Summary

**Background** Here is the dataset a collection of sponsored political posts on Facebook targeted at US users during 23 months (03/2020-01/2022). This includes the period preceding the latest US Presidential election in November 2020. I would use it do some dataset analytics.

**AIM** I would like to figure out what makes an impact on the region distribution of US Facebook advertisements in 2021, trying to find out hidden patterns behind them.

**Method** I used Python and Pyspark through Jupyter to do dataset analytics. Analyse and sort ad data by ads' title, body, funding entity, link caption and region distribution. Then combine some additional data and information to get the results.

**Findings** I found that at least in US, the region distribution of Facebook political advertisements was effluent by regional population. If the population in a state is large, the percentage of ads posted tends to be high. Apart from it, in some particular periods, data might have obvious differences course of regional political events like elections.

## **Table of contents**

**Introduction----- p.3**

**Dataset Analytics----- P.4 ~ P.16**

**Discussion and conclusions of the analysis----- P.17**

**Appendix ----- P.17**

## Introduction

Big data analytics is the less-cost way to deal with a large amount of data and uncover meaning, patterns and other messages hidden. It can be used in lots areas, for example, business. If the company wants to gauge customer needs and customer satisfaction through analysing all of customers' purchase records in every store and other channels in a long period, it will need big data analytics with a distribution system, a computing environment supporting calculating tasks across multiple computing devices which can work parallely, and offering an efficient way to deal with complex, massive and even unstructured datasets.

Of course, big data analytics and distribution systems are suit for analysing Facebook political advertisements for finding their region distribution. Look at FB political dataset file, I noticed that ads were not posted averagely in every region. As a result, I would like to analyse the advertisements' region distribution in each US state in 2021, and trying to find reasons.

To figure out what caused the region distribution of ads, I would deal with their title, body, funding entity and link caption. Besides, I compared region distributions between two related link captions, "winred.com" and "secure.actblue.com", trying to find the differences between them.

## Dataset Analytics

### Tool:

I would use python and Pyspark, which is the Python API for Apache Spark through Jupyter to do this project.

### Data Used:

In order to analyse region distribution of FB political advertisements monthly by United States in 2021, I loaded all the files recording 2021 ad data. In this project, I mainly focused on regions in US. In addition, I also needed to analyse related titles, bodies, link captions and funding entities to figure out possible reasons why ads were posted in this distribution. As a result, those attributes, "ad\_creative\_body", "ad\_creative\_link\_caption", "ad\_creative\_link\_title", "funding\_entity" and "region\_distribution" were needed.

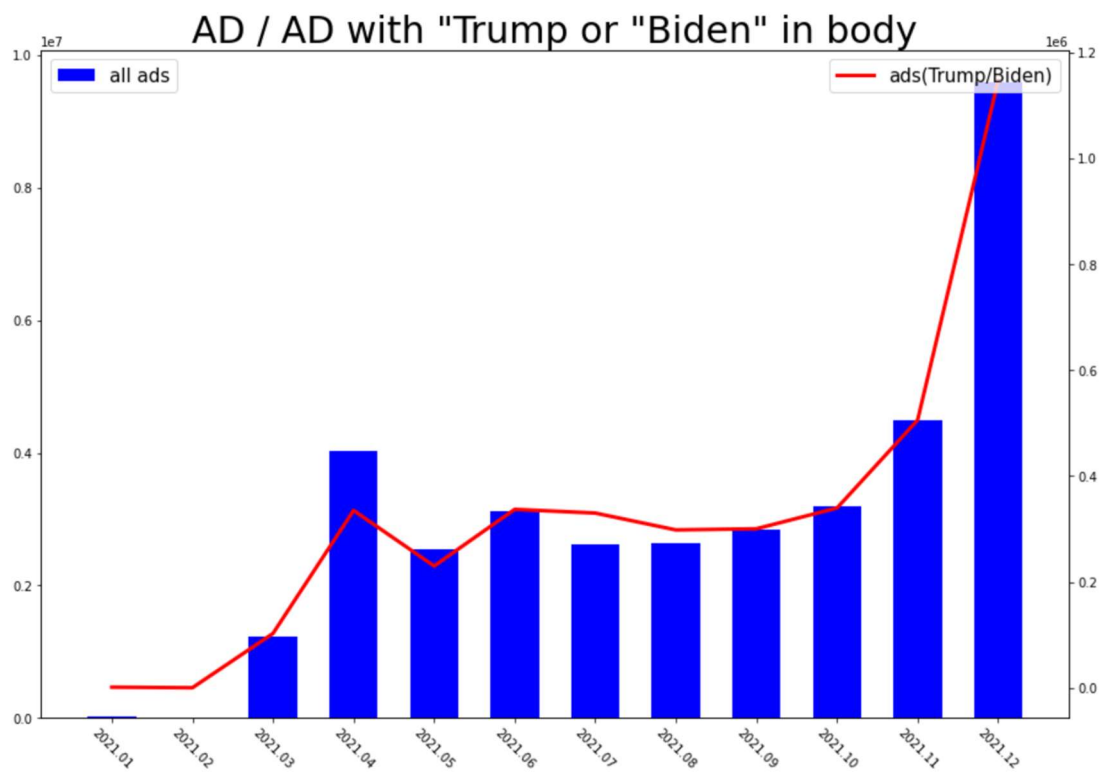
### Data Loading:

At first, I loaded files from HDFS and transferred them as 12 data frames by month. During loading all the 2021 files, I found there were several files missing.

```
/data/ProjectDatasetFacebook/FBads-US-2021121-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021130-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021201-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021205-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021209-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021214-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021216-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021226-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021229-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021230-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021518-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021603-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021625-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021708-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021715-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021724-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021725-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021804-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021824-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021830-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021901-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021907-05_00_01
/data/ProjectDatasetFacebook/FBads-US-2021908-17_00_01
/data/ProjectDatasetFacebook/FBads-US-2021925-17_00_01
/data/ProjectDatasetFacebook/FBads-US-20211004-17_00_01
/data/ProjectDatasetFacebook/FBads-US-20211015-17_00_01
/data/ProjectDatasetFacebook/FBads-US-20211018-17_00_01
/data/ProjectDatasetFacebook/FBads-US-20211021-05_00_01
/data/ProjectDatasetFacebook/FBads-US-20211024-17_00_01
/data/ProjectDatasetFacebook/FBads-US-20211028-05_00_01
/data/ProjectDatasetFacebook/FBads-US-20211114-05_00_01
/data/ProjectDatasetFacebook/FBads-US-20211204-05_00_01
/data/ProjectDatasetFacebook/FBads-US-20211212-05_00_01
```

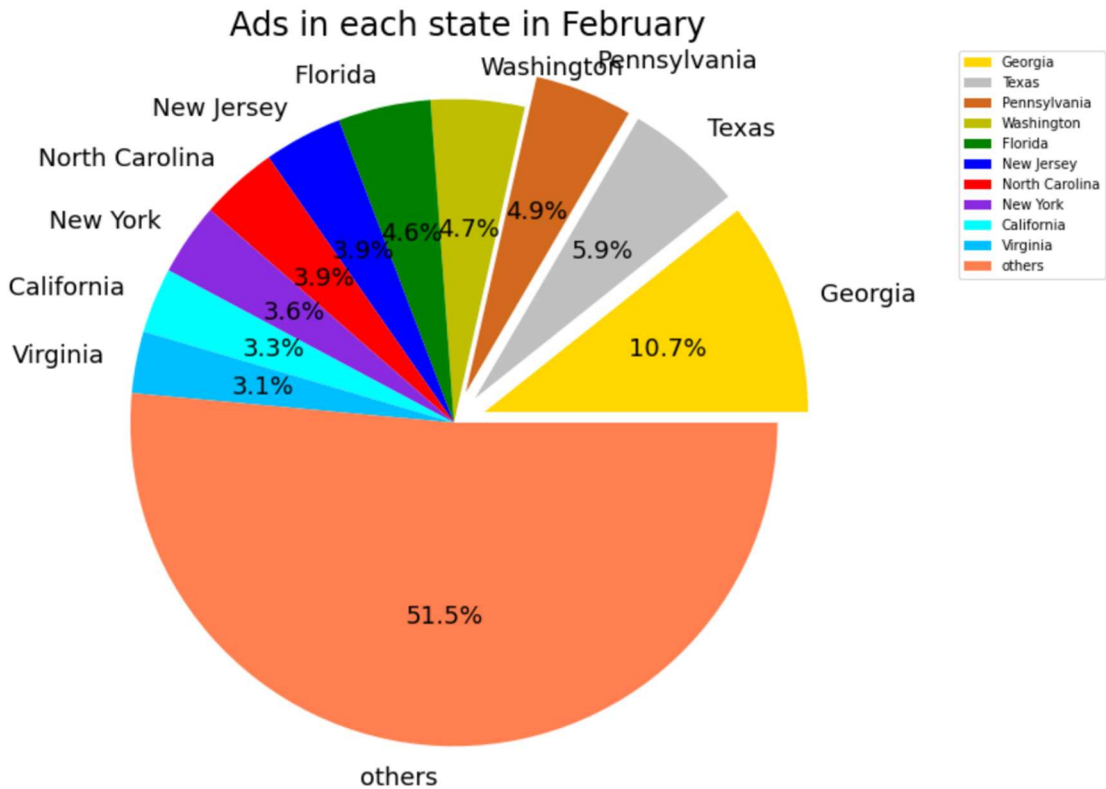
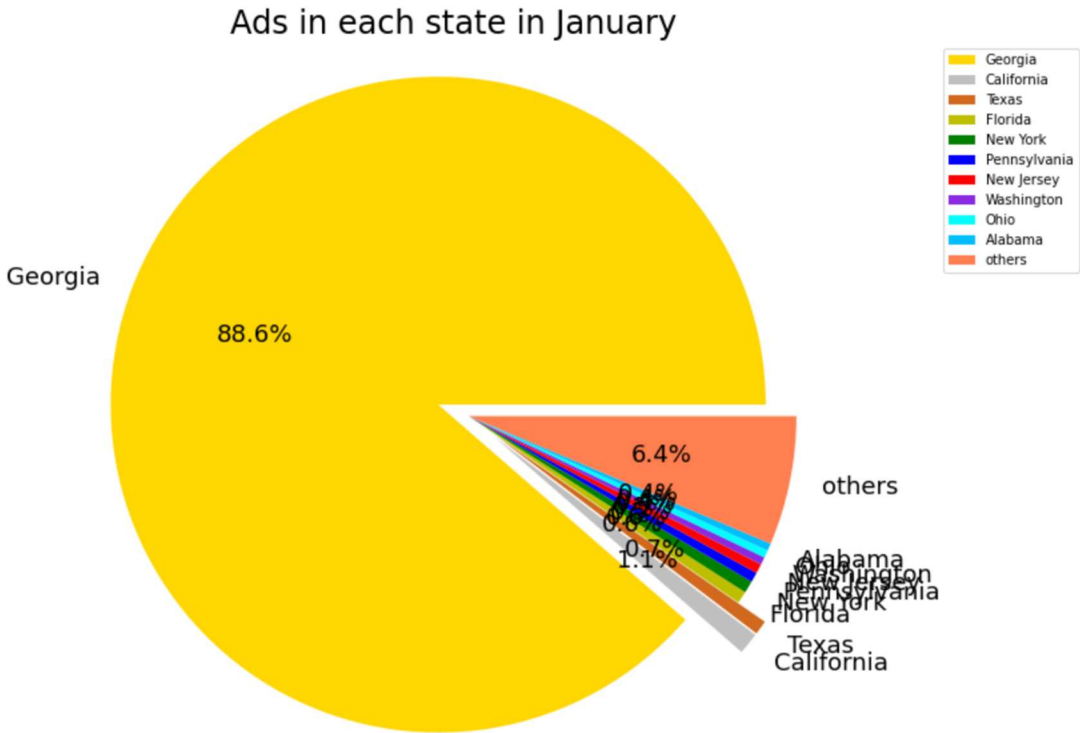
## Data Counting:

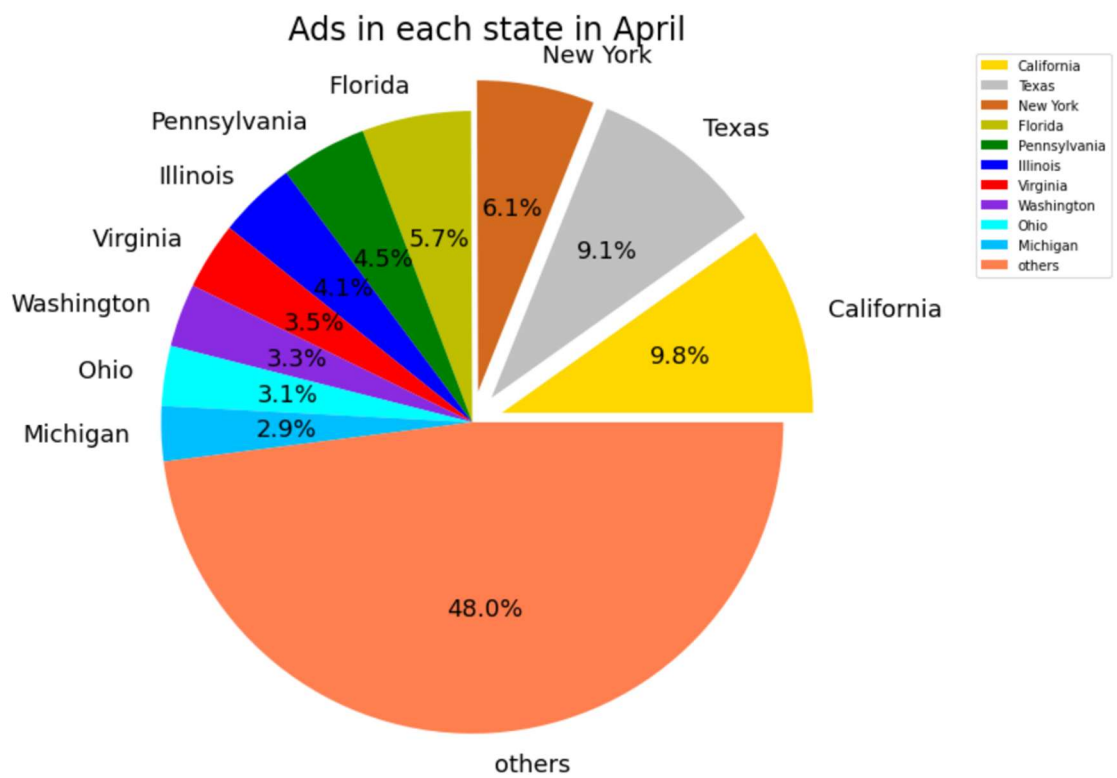
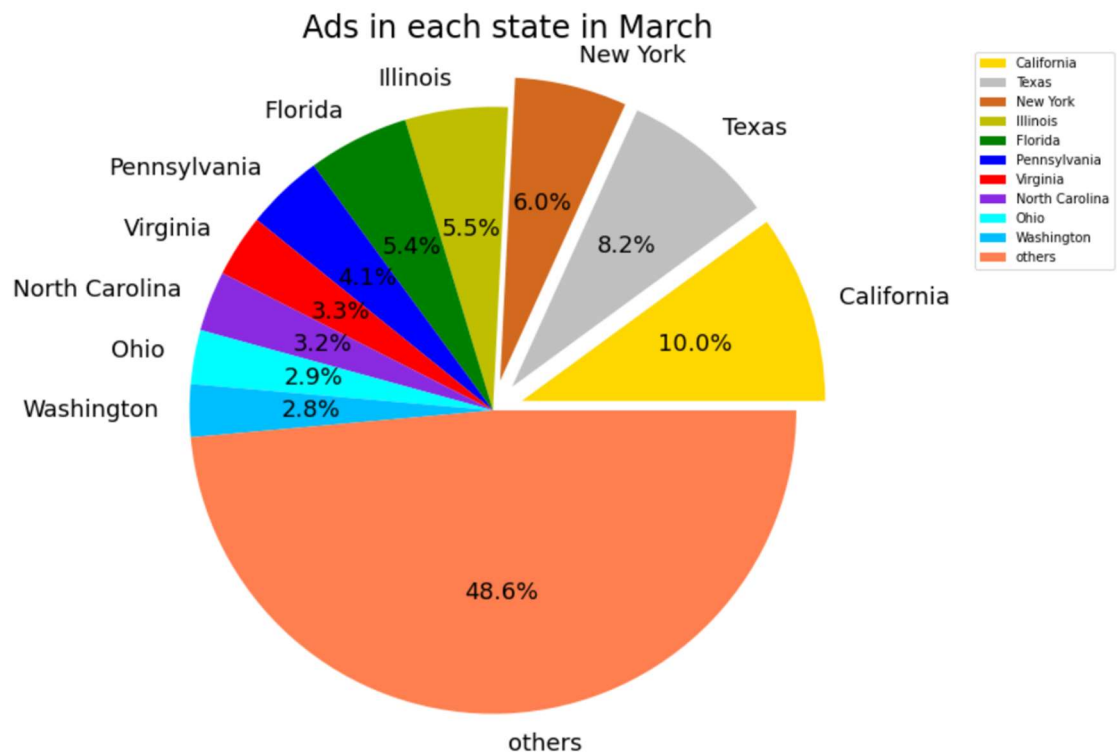
Here I count all the data I used to make sure how many rows of data I used. And when it came to US politics in 2021, Trump and Biden were definitely big wheels. As a result, I also counted ads that mentioned their names in their body. After both of results were presented monthly, I found that they were basically proportional. Besides, both of them were increasing basically.



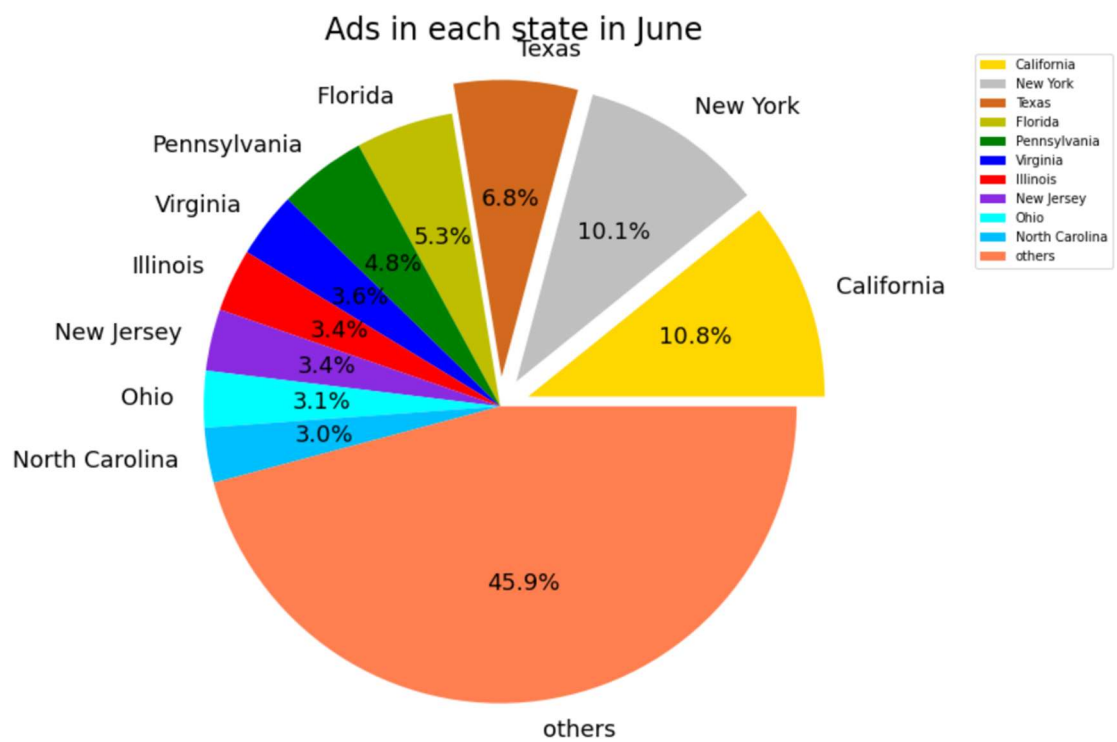
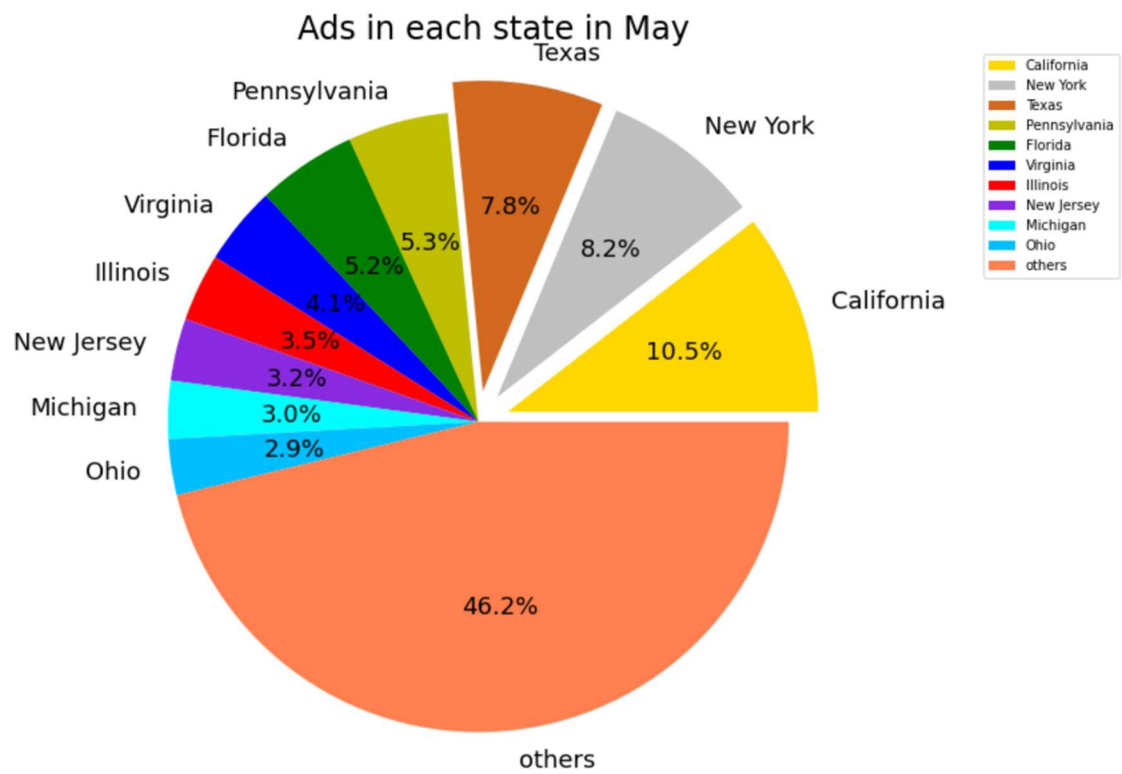
Region distribution (Attributes: "region\_distribution"):

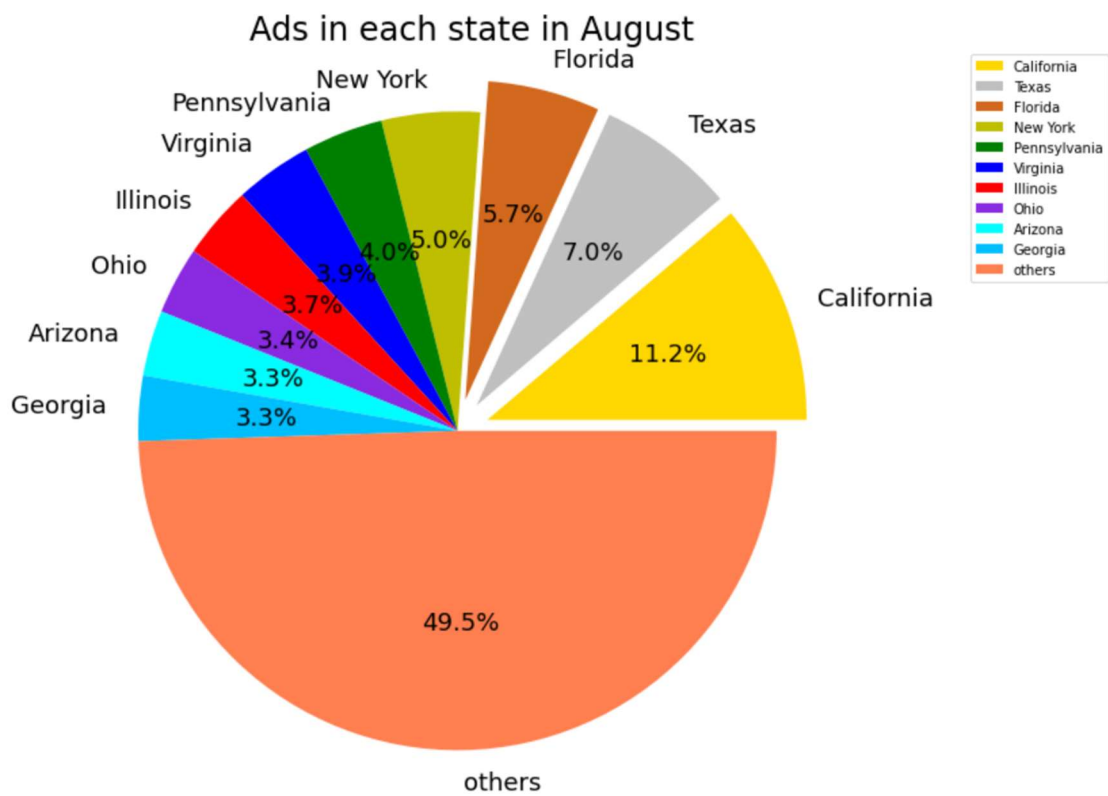
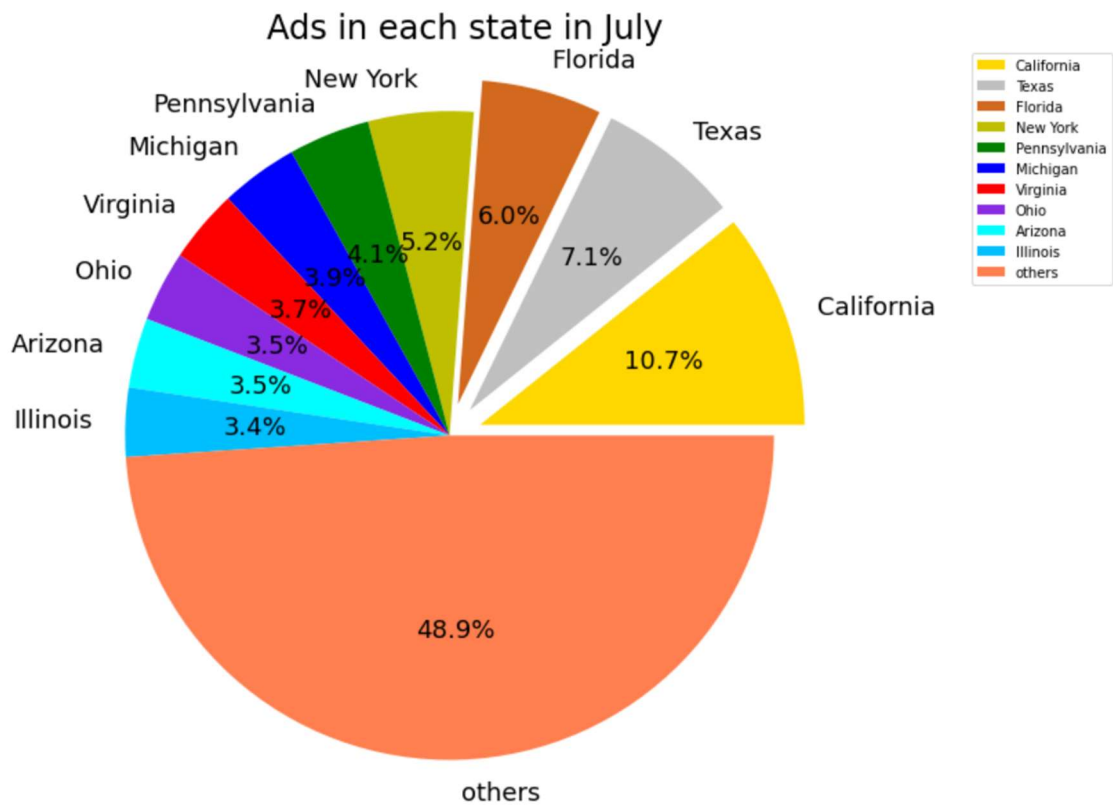
Then I selected all the regions which were US states and their own percentage by month and state separately. Here I attached attention on top 10 states which occupied higher percentage.

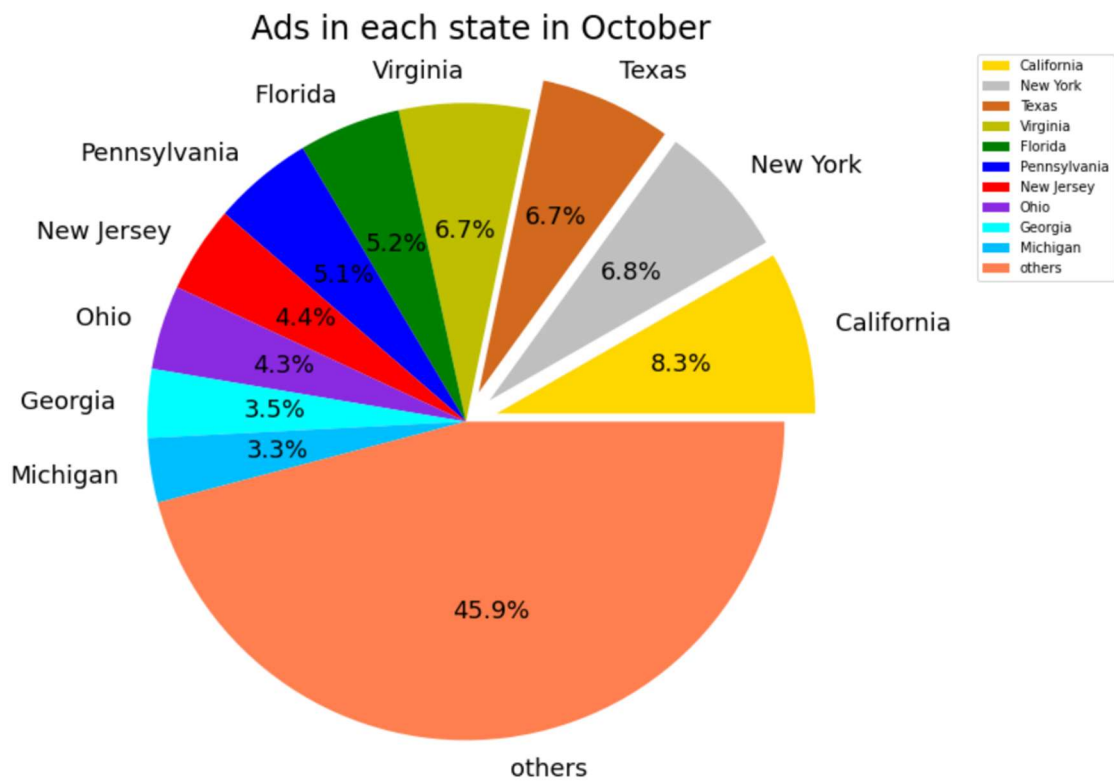
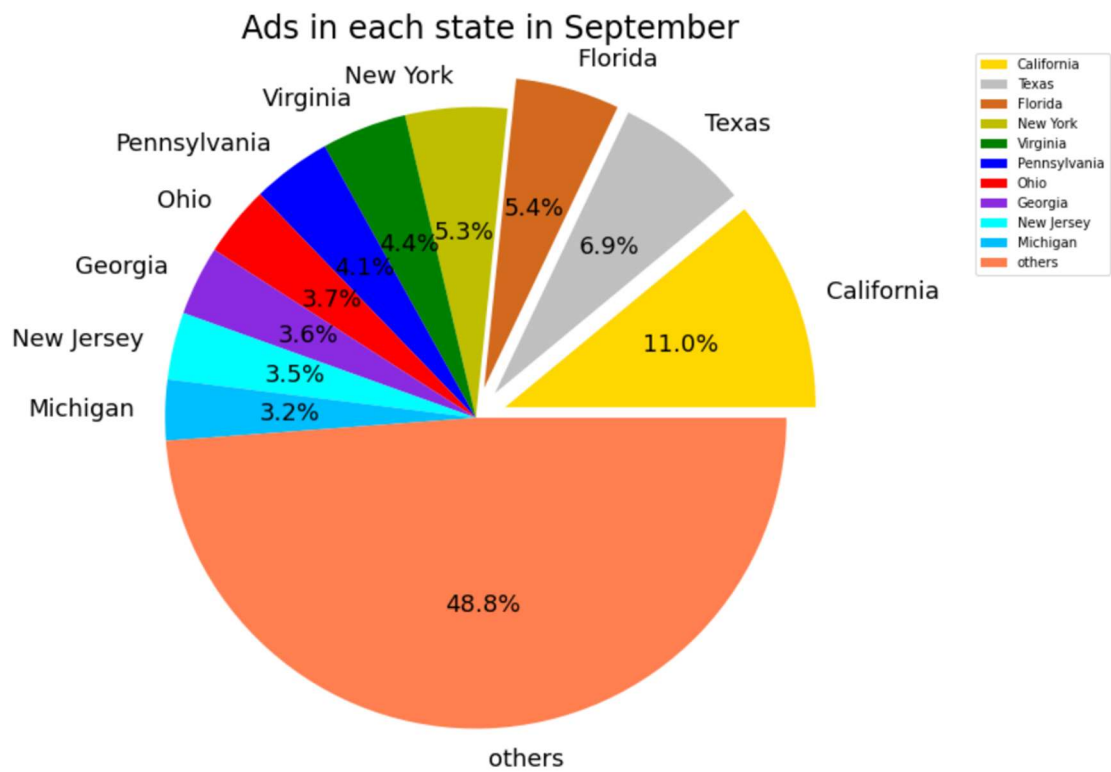


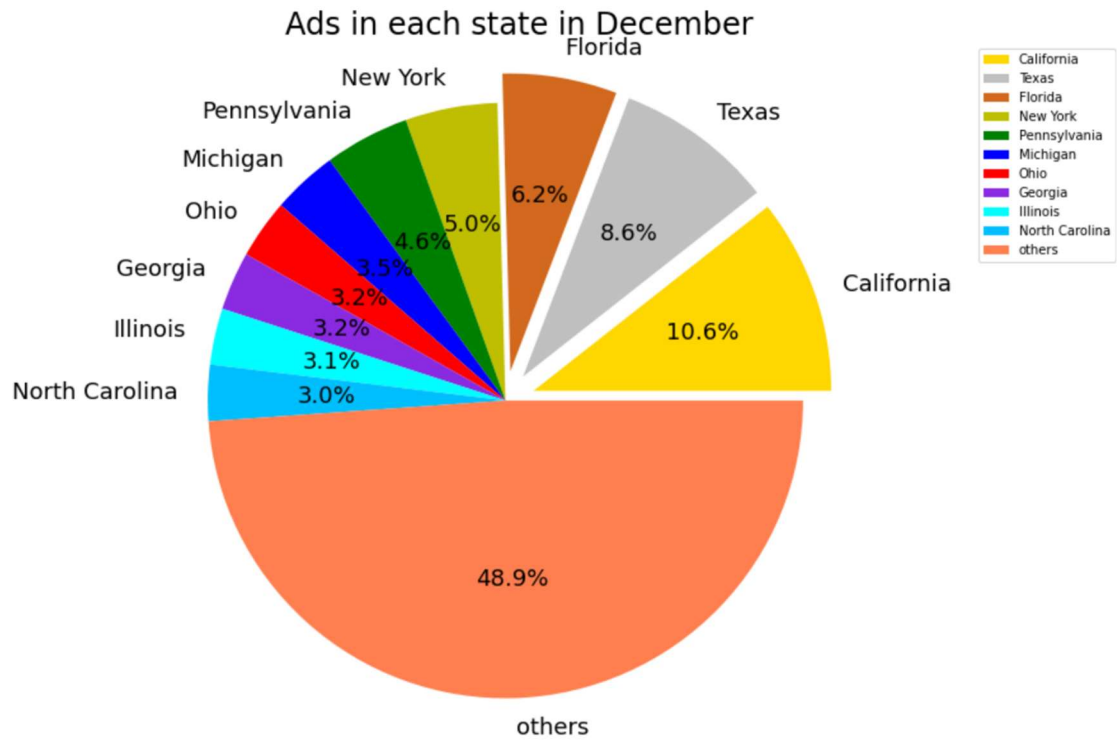
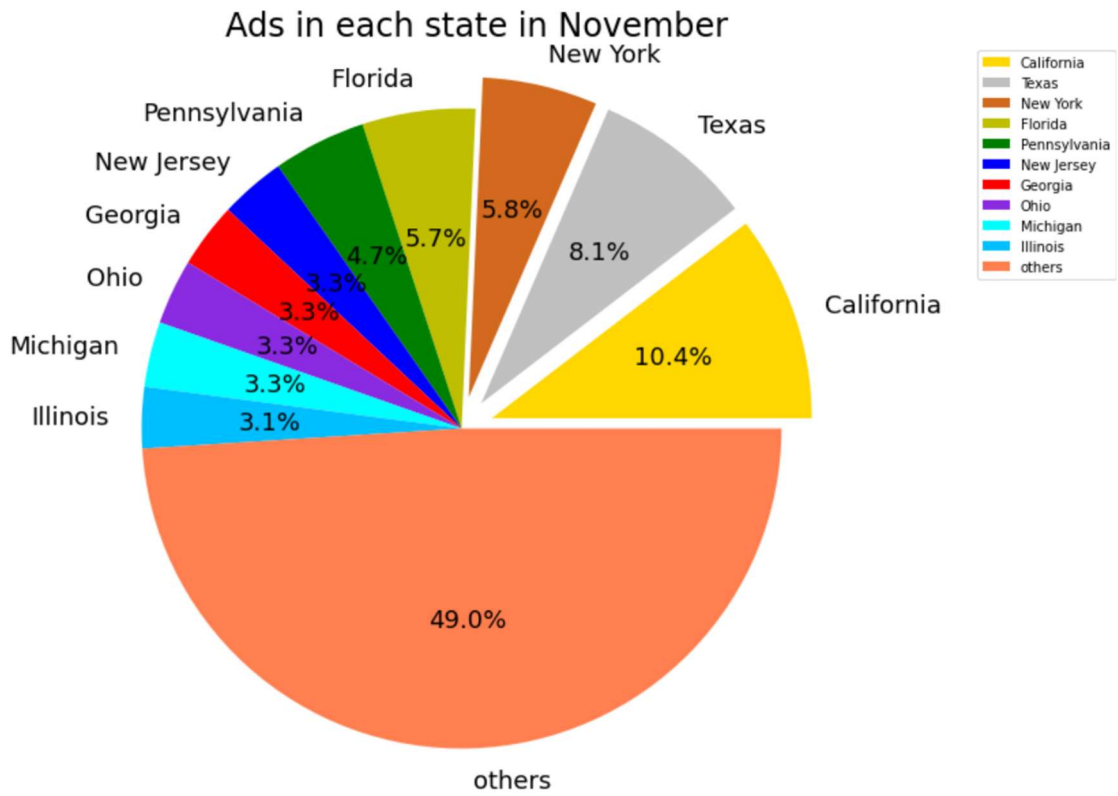






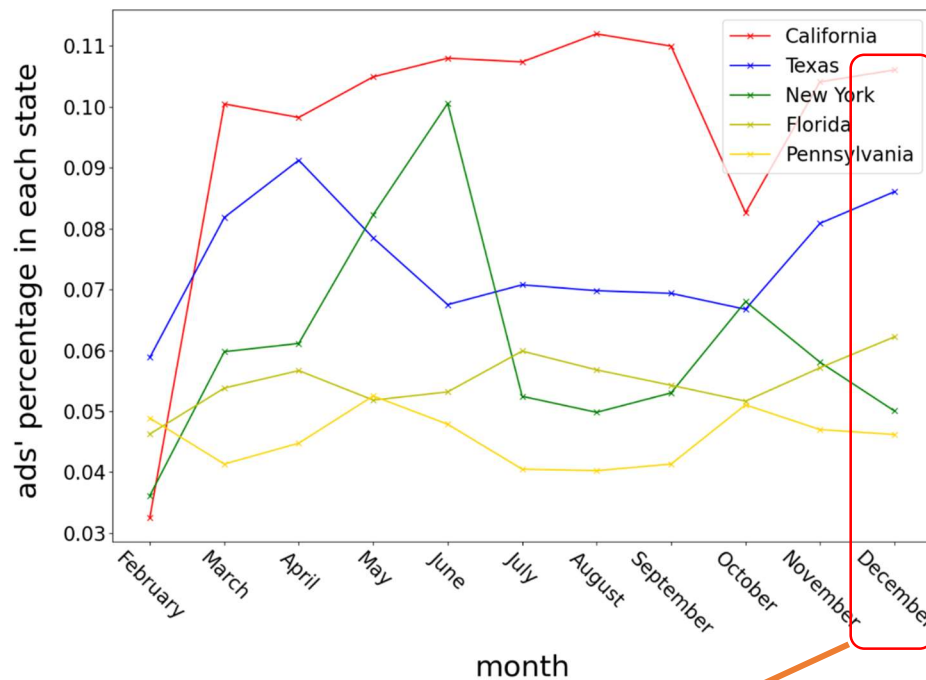






From the pie charts above I found the result in January was really different than others. As a result, I set it as an outlier. I collected other 11 results and summed all the percentage by US states to find the top 5 regions that ads posted in. It illustrated that the proportion of ads in California was the highest after February. More than that, by comparing with Top 10 most populous states in 2021, the top5 were the same. Specially in December, even the order was the had no difference. As a result, I figured out that the percentage of ads in different states depended on population.

### Ads in top 5 US states.



Top 10 Most Populous States: 2021

Rank	Geographic Area	April 1, 2020 (Estimates Base)	July 1, 2020	July 1, 2021
1	California	39,538,223	39,499,738	39,237,836
2	Texas	29,145,505	29,217,653	29,527,941
3	Florida	21,538,187	21,569,932	21,781,128
4	New York	20,201,249	20,154,933	19,835,913
5	Pennsylvania	13,002,700	12,989,625	12,964,056
6	Illinois	12,812,508	12,785,245	12,671,469
7	Ohio	11,799,448	11,790,587	11,780,017
8	Georgia	10,711,908	10,725,800	10,799,566
9	North Carolina	10,439,388	10,457,177	10,551,162
10	Michigan	10,077,331	10,067,664	10,050,811

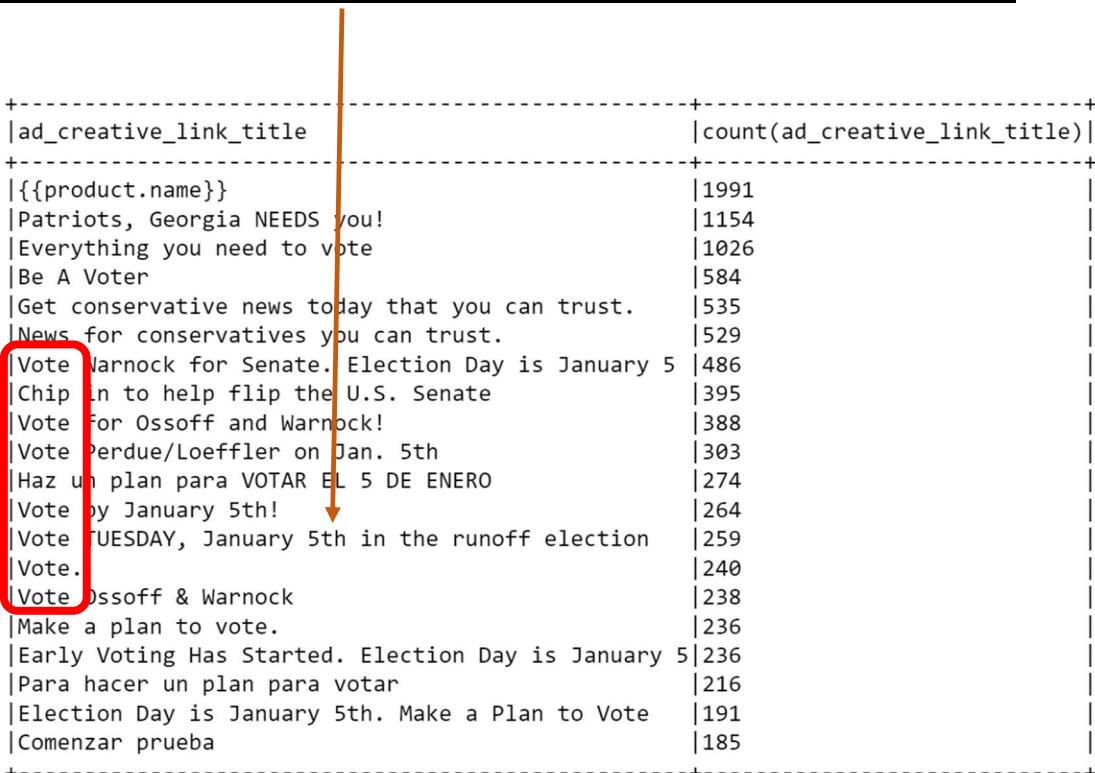
(See Appendix A1)

Contents of January (Attributes: “ad\_creative\_link\_title”, “ad\_creative\_body”, “funding\_entity”, “region\_distribution”):

As I said, I labelled the result in January was an outlier. For discovering the reason, I had to uncover hidden patterns in their contents. First, I aimed at ads in “Georgia”, which percentage occupied more than 88% and selected their title, body and funding entity.

After checking titles and bodies and counting each of them to make some analytics, I found lots of them included “vote” in the strings. As a result, I counted all the titles and bodies with “vote” as key point. The data frame showed that approximately 37% of titles and 44% bodies met the condition. It is obvious that something grand related to politics happened in Georgia in January.

In one of the bodies, I noticed there was a line about an election in January 5<sup>th</sup>.



ad_creative_link_title	count(ad_creative_link_title)
{{product.name}}	1991
Patriots, Georgia NEEDS you!	1154
Everything you need to vote	1026
Be A Voter	584
Get conservative news today that you can trust.	535
News for conservatives you can trust.	529
Vote Warnock for Senate. Election Day is January 5	486
Chip in to help flip the U.S. Senate	395
Vote For Ossoff and Warnock!	388
Vote Perdue/Loeffler on Jan. 5th	303
Haz un plan para VOTAR EL 5 DE ENERO	274
Vote by January 5th!	264
Vote TUESDAY, January 5th in the runoff election	259
Vote.	240
Vote Ossoff & Warnock	238
Make a plan to vote.	236
Early Voting Has Started. Election Day is January 5	236
Para hacer un plan para votar	216
Election Day is January 5th. Make a Plan to Vote	191
Comenzar prueba	185

number of January ads in Georgia	number of titles with 'vote'	number of bodies with 'vote'
32407	11945	14389
% of January ads in Georgia	% of titles with 'vote'	% of bodies with 'vote'
1.0	0.368593	0.444009



To confirm this detection, I turned to analyse bodies and funding entities. Therefore, I compared funding entities with bodies, and checked the top 5 funding.

There were two steps:

- (1) Found the top 5 funding entities.
- (2) Checked their ad bodies.

Then I found those entities were organization about election or political parties.

Check

ad_creative_body	funding_entity
Hey Georgia! Do ...	MAJORITY FORWARD
"Me and Trump Fis...	Our Friendly Forest
Hey Georgia! Show...	MAJORITY FORWARD
Make your plan to...	MAJORITY FORWARD
Haga un plan para...	MAJORITY FORWARD
Your vote is bigg...	Fight For The Base
Georgia, we've co...	Fight For The Base
Georgia, we've co...	Fight For The Base
Haz tu plan para ...	MAJORITY FORWARD
Healthcare is on ...	Fight For The Base
Haga su plan para...	MAJORITY FORWARD
"I hit the button...	Our Friendly Forest
¡Hola Georgia! H...	MAJORITY FORWARD
"I'm on the road ...	Our Friendly Forest
Make a plan to vo...	MAJORITY FORWARD
{{product.brand}}	MAJORITY FORWARD
We owe it ourselv...	Fight For The Base
Who you vote for ...	MAJORITY FORWARD

funding_entity	count(funding_entity)
Fight For The Base	2491
Our Friendly Forest	2137
Black Voters Matter Fund Inc	1789
REPUBLICAN NATIONAL COMMITTEE	1671
MAJORITY FORWARD	1544
MASTV EL PLANETA LLC	1385
WARNOCK FOR GEORGIA	1363
Real Voices Media	1350
WorkMoney	1254
BizPac Review	1064
JON OSSOFF FOR SENATE	1032
Democratic Party of Georgia	950
AB PAC	877
Asian American Advocacy Fund PAC	863
TED CRUZ FOR SENATE	767
RESTORATION PAC	743
Asian American Advocacy Fund	667
Future Coalition	623
Progress Georgia, Inc.	620
Feel Good Action	556

Organizations encouraged people to vote and uphold their own right.

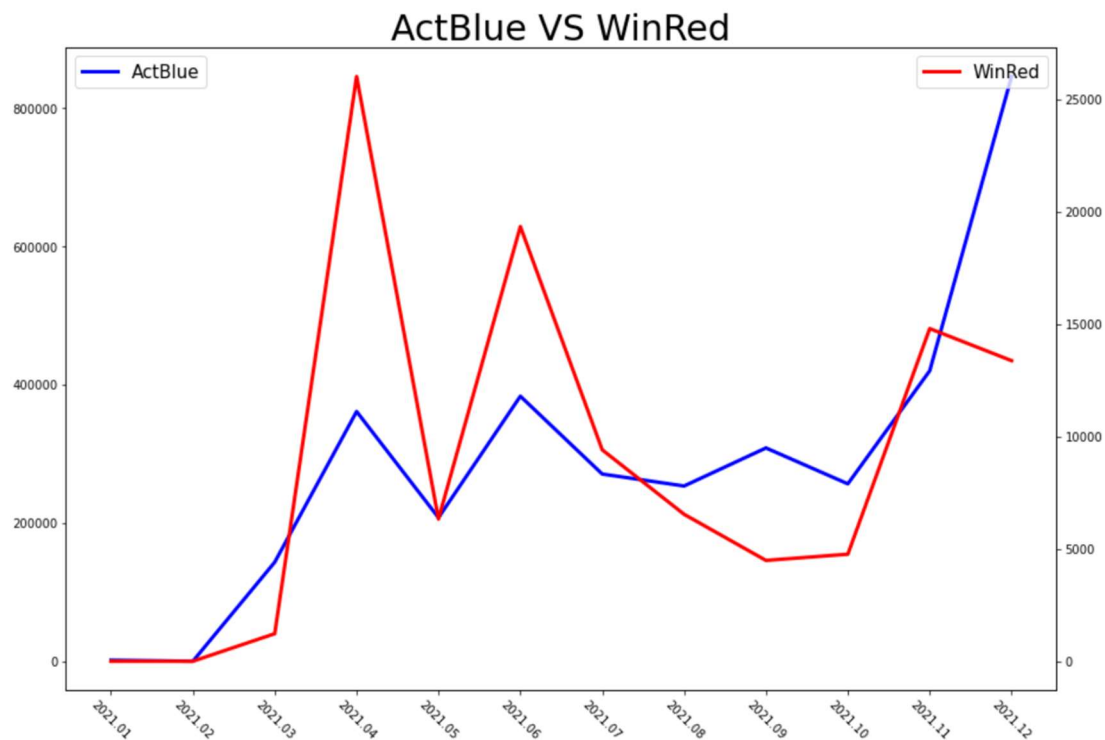
Organizations related to political parties.

## Link Captions Comparison (Attributes: "ad\_creative\_link\_caption", "region\_distribution"):

When checking link captions and counting their times appearing in the whole ad dataset, I noticed that there were two link captions which numbers in top 5 and seems to be related to each other. (See Appendix A3)

ad_creative_link_caption	count
secure.actblue.com	3378263
null	3185105
fb.me	1473974
facebook.com	1161745
secure.winred.com	395407
p2a.co	379224
democrats.org	304071
secure.ngpvan.com	267692
aclu.org	260462
about.fb.com/regulations	246830

As a result, I would like to compare them with region distribution, and limited the regions as US states. At first, I compared them with numbers by month. ActBlue shows that its trend was like the trend of all the ads. On the other hand, WinRed seemed to be different, especially a dramatically drop in December.





Because they were some different trend performances, especially after August 2021, I decided to analyse their percentage of region distribution. Additionally, due to the number of WinRed is zero in not only January but February, I summed the percentage starting from March.

WinRed			ActBlue		
	State	Percentage		State	Percentage
0	Texas	2.036594	0	California	1.445977
1	Pennsylvania	1.497595	1	Texas	0.680528
2	California	0.549940	2	New York	0.632121
3	Florida	0.538509	3	Florida	0.562601
4	Indiana	0.492619	4	Pennsylvania	0.433749
5	Ohio	0.307608	5	Washington	0.408437
6	New York	0.287132	6	Illinois	0.397129
7	Michigan	0.278199	7	Virginia	0.387629
8	Arizona	0.249845	8	North Carolina	0.373262
9	Illinois	0.220219	9	Arizona	0.348786
10	Georgia	0.218479	10	Michigan	0.299620
11	North Carolina	0.210814	11	Ohio	0.288747
12	Tennessee	0.178976	12	Massachusetts	0.279374
13	Missouri	0.176298	13	Colorado	0.262233
14	Washington	0.166532	14	New Jersey	0.236601
15	Virginia	0.166481	15	Oregon	0.236239
16	New Jersey	0.155557	16	Maryland	0.218524
17	Alabama	0.143329	17	Georgia	0.210039
18	South Carolina	0.131783	18	Minnesota	0.195739
19	Kentucky	0.116184	19	Wisconsin	0.182748

It could be observed that the order of region percentage of ads which link caption was RedWin was not really like the order of population of US states from the data frames. On the other hand, the order of ActBlue was similar to the population order. It might be the reason that the line trend of WinRed was not like ActBlue and the drop in December 2021.

## Discussion and conclusions of the analysis:

In conclusion, the region distribution of US Facebook advertisements in 2021 followed some regular patterns on the whole. First, the number of ads was increasing month and month, consistent and proportional with those with “Trump” or “Biden”, names of two of American Presidents.

When it comes to the relationship between the region distribution of ads and American states, I found that the order of percentage ads occupied by state was similar to the order of population of states after analysing both data. In other words, how to allocate the percentage of ads posted in that state must depend on how large the population there to a large extent. For example, in December 2021, the top 5 states most ads posted in were literally same as the top 5 most populous states.

However, the region distribution of ads in January 2021 was really an outlier, in which more than 88% of ads were posted in Georgia. To figure out the reason, I put an emphasis on analysing ads in Georgia in January 2021. After checking and getting the results from analytics, I found it was because an election - United States Senate special election in Georgia. (See Appendix A2)

In addition, the region distribution and number by month of link captions were based on ads they belonged to. It might not be similar to the distribution and order of population.

With these analytics, Facebook operators can predict region distributions of ads by populations and some considerable regional political events, like elections to distribute their resources reasonably. Apart from it, they are able to design different plans based on those factors, and get more benefits from fund entities. For who wants to post political ads, he/she can decide the region distribution by population as reference. Last but not least, if users want to get more ads information or escape from harassments brought by political ads, he/she might can refer to those analytics to change his/her internet domain.

## Appendix

A1 Picture is from <https://www.census.gov/newsroom/press-releases/2021/2021-population-estimates.html>

A2 [United States Senate special election in Georgia](#)

A3 [ActBlue VS WinRed](#)

A4 [Jupyter Code](#)