Sentiment Analysis of COVID-19 Social Media Posts

By Paulene Barnes, Ari Kerdell, Matt Williams

Problem Statement

Problem

The COVID-19 response in the US has been largely regional and state-based in nature. Some states have enacted strictly enforced stay-at-home policies, while others have provided guidelines. Incidences of COVID-19 cases has varied from state-to-state as well.

The public's reaction to this information has changed over time. July, specifically, has had many new topics of conversation. Examples include:

- The CARES Act stimulus package expired on July 31st, bringing an end to the additional federal unemployment benefits for the many job-displaced Americans
- School systems developed plans for how to resume classes in the fall
- Many states began to re-evaluate COVID-19 response policies
- Many states started to experience a "second wave" of Coronavirus cases

We will investigate trends in language used and sentiment of social media posts for each state during July 2020. We will compare these to both the local policies on social distancing and the occurrences of the pandemic in those areas.

Gathering Data

Twitter - July Tweets Only

- Tweepy API - We hit a wall using this API to gather our data
- Rabindra Lamsal's Coronavirus GEO-tagged tweet dataset (https://ieee-dataport.org/open-access/coronavirus-covid-19-geo-tagged-tweets-dataset) was much better suited to our needs.
- DocNow's Hydrator (https://github.com/DocNow/hydrator) tool to repopulate tweet information from JSON Tweet IDs

Reddit - July Posts Only

- Pushshift API
- Pulled Reddit posts from each state’s coronavirus/COVID-19 related subreddit using a function
- Resulted in a combined data frame with 9839 rows
    - Note: Delaware and South Dakota did not have specific subreddits regarding coronavirus or COVID-19

Policy Data - Latest Updates for July

- Found a site that had the policies for each state exportable to a CSV
  - https://www.kff.org/coronavirus-covid-19/issue-brief/state-data-and-policy-actions-to-address-coronavirus/
  - Included policy on reopening, stay at home order, emergency declaration, large gatherings, and public mask requirements.

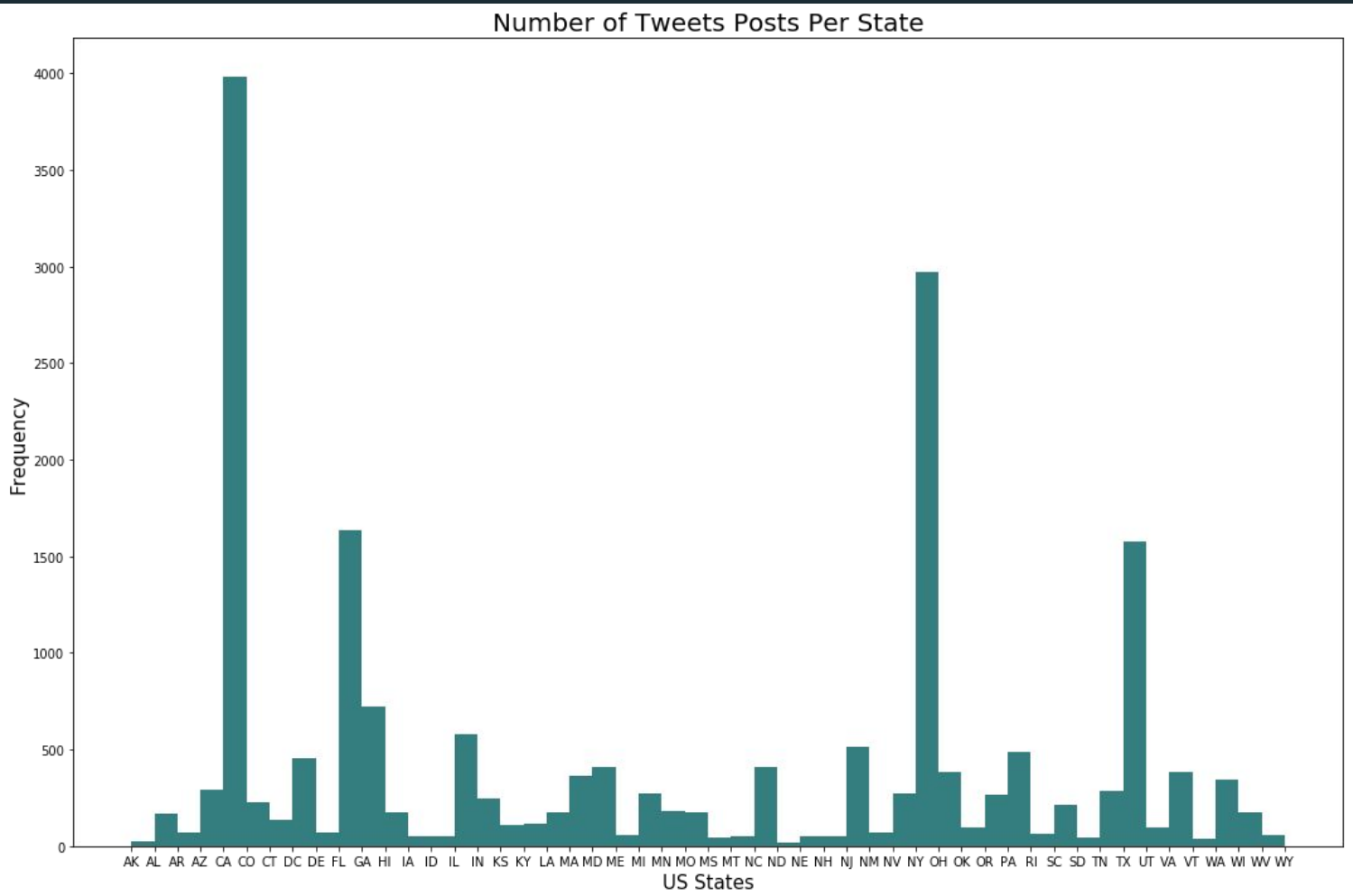
Health Data - Latest Updates for July

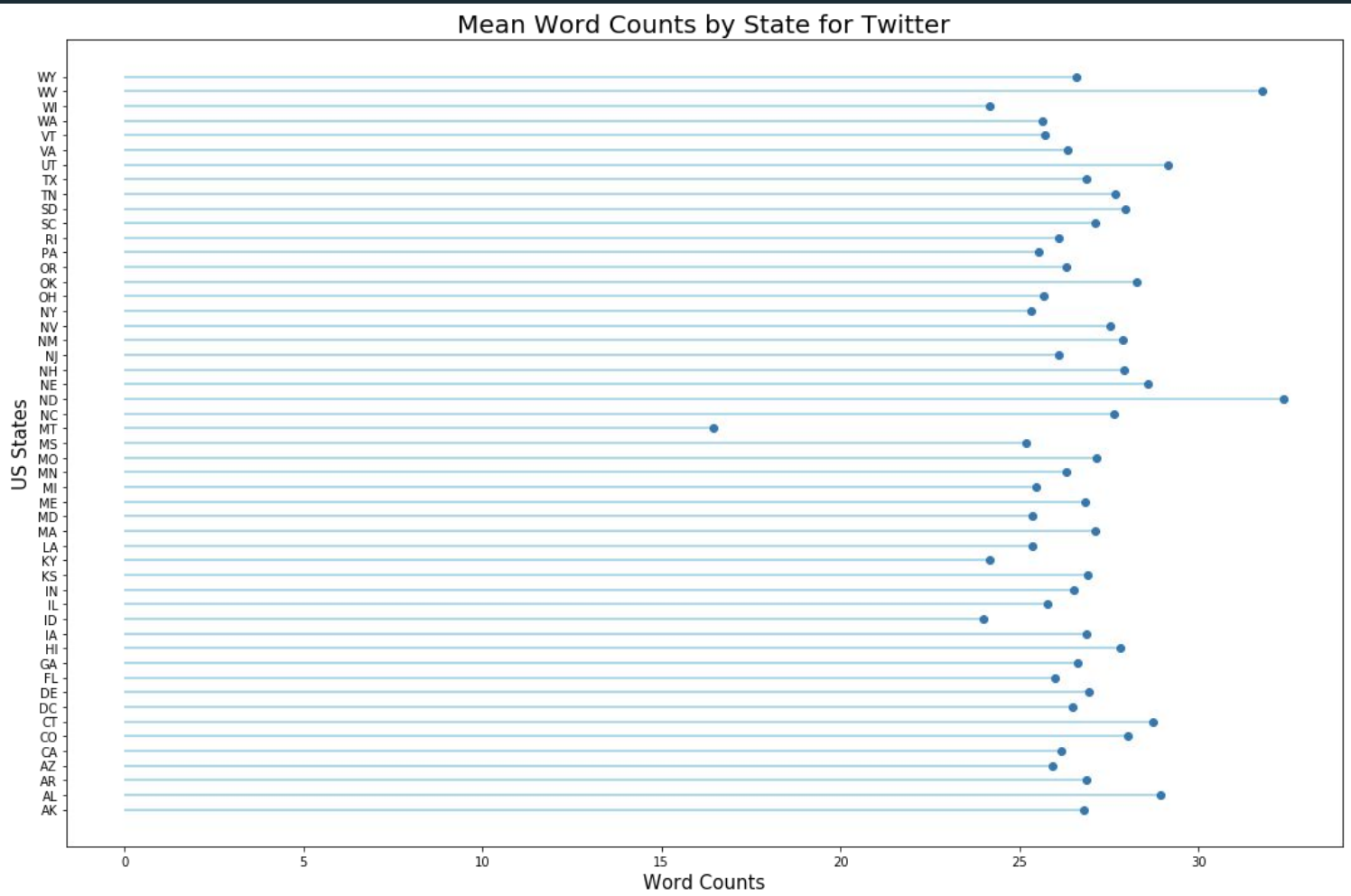
- Found a site that had the policies for each state exportable to a CSV
  - https://covidtracking.com/api/v1/states/daily.csv
  - Included total tests, positive rest rate, total hospitalized, total positive cases, and a lot of other important health data for each state.

Cleaning and EDA

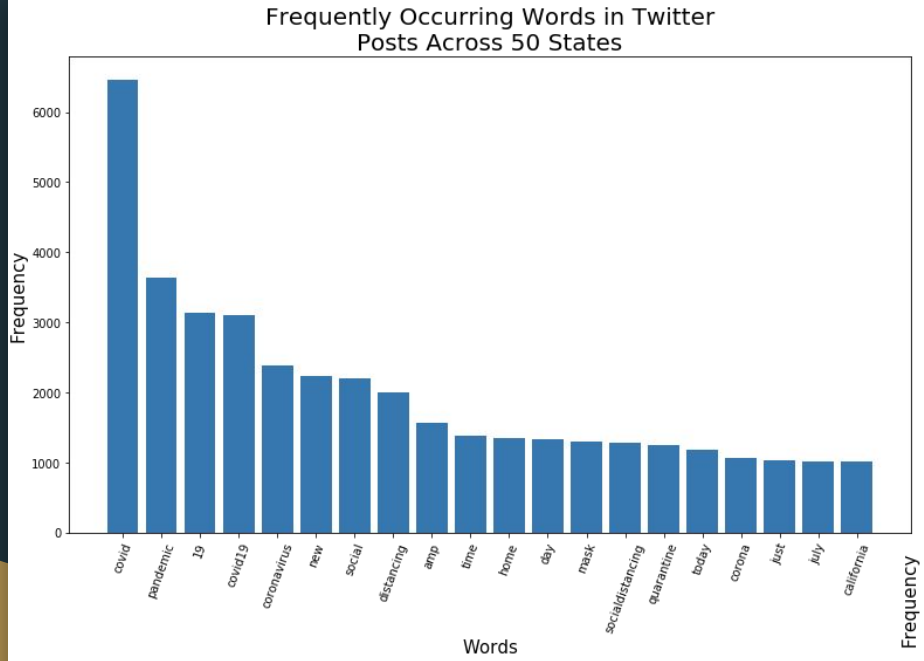
Twitter

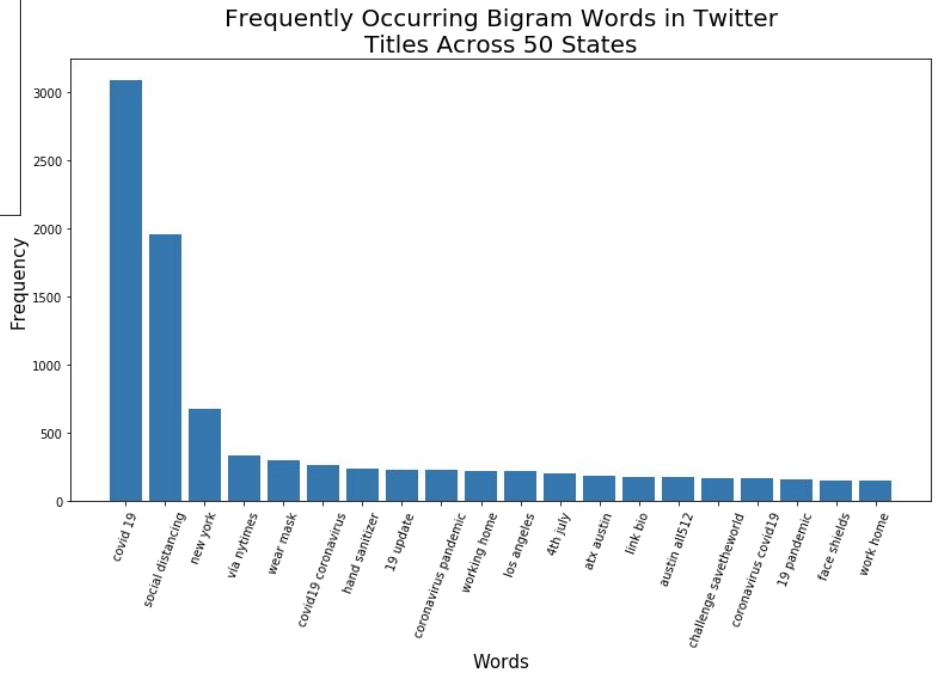
- We attempted to fill in missing place values by reverse geocoding using GeoPandas. The coordinates were unreliable and the missing places had to be dropped.
- We filtered out the tweets that came from outside the U.S.
- Mapped tweets that had oddly formated ‘place’ values to their proper state


Number of Tweets Posts Per State
Frequency
0
500
1000
1500
2000
2500
3000
3500
4000
AK AL AR AZ CA CO CT DC DE FL GA HI IA ID IL IN KS KY LA MA MD ME MI MN MO MS MT NC ND NE NH NJ NM NV NY OH OK OR PA RI SC SD TN TX UT VA VT WA WI WV WY
US States


Mean Word Counts by State for Twitter
US States
WY
WV
WI
WA
VT
VA
UT
TX
TN
SD
SC
RI
PA
OR
OK
OH
NY
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MO
MN
MI
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GA
FL
DE
DC
CT
CO
CA
AZ
AR
AL
AK
0
5
10
15
20
25
30
Word Counts

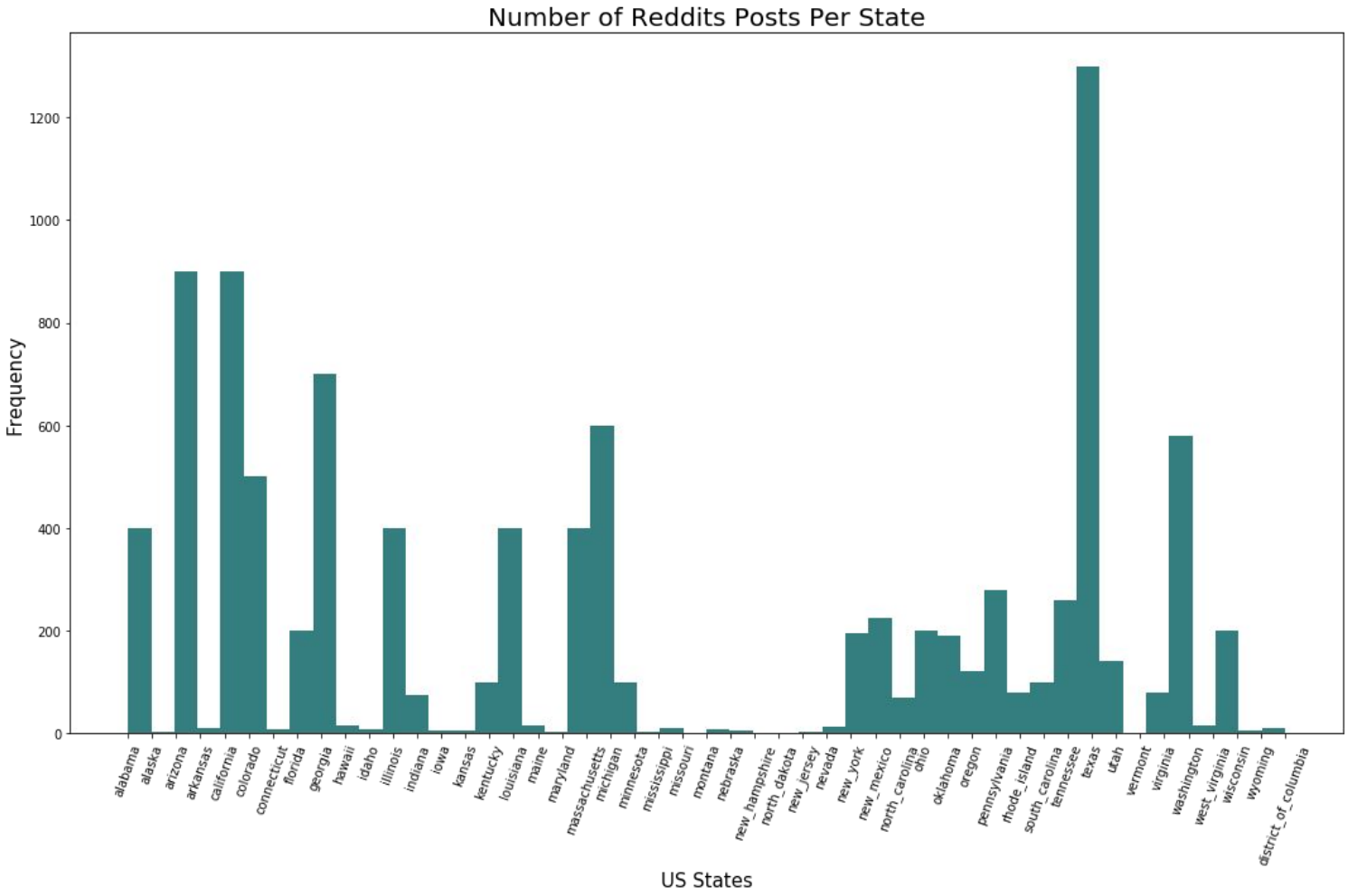
20 Most Frequently Occurring Words

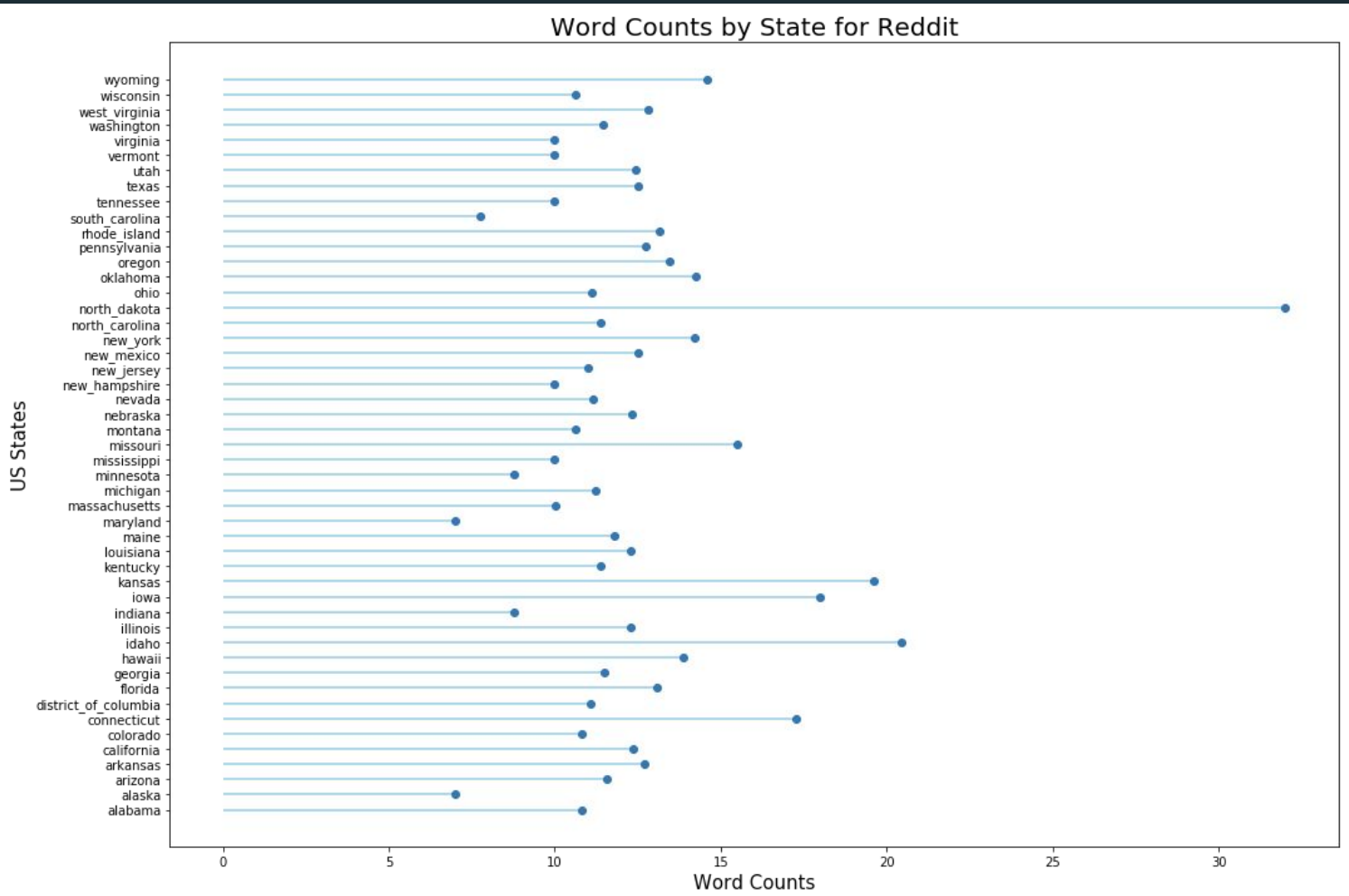


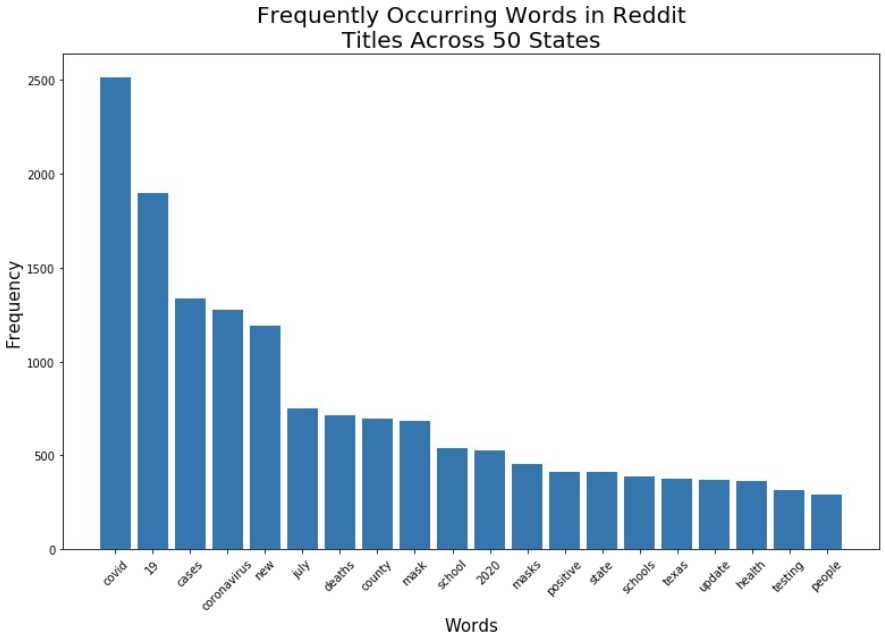


Reddit

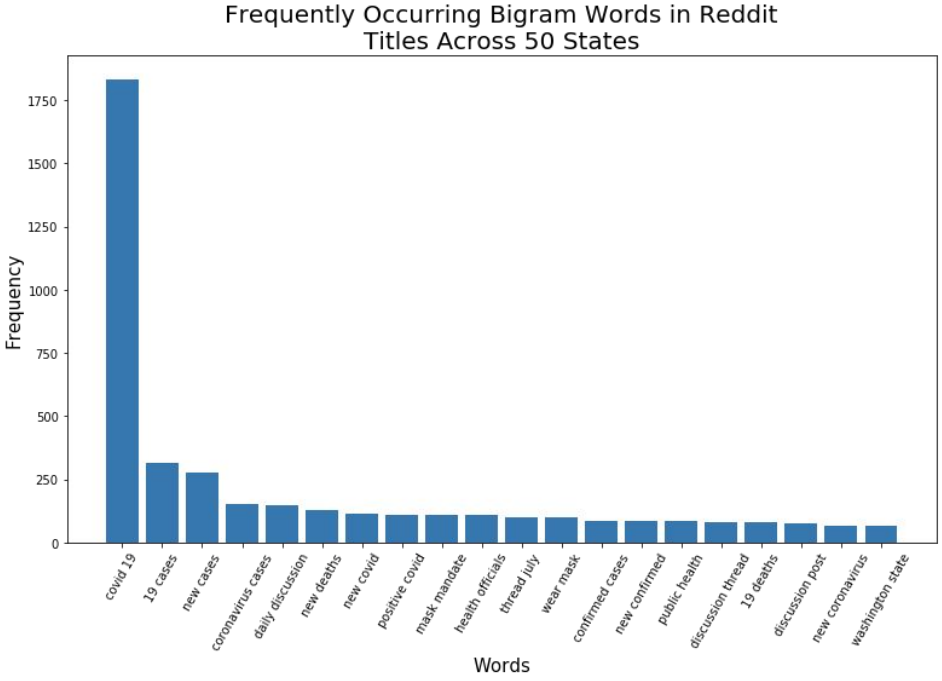
- Data frames for each of the 50 state plus Washington DC were concatenated together
- Check for null values in the subreddit titles
  - There were no null values


Number of Reddits Posts Per State
Frequency
0
200
400
600
800
1000
1200
alabama
alaska
arizona
arkansas
california
colorado
connecticut
florida
georgia
hawaii
idaho
illinois
indiana
iowa
kansas
kentucky
louisiana
maine
maryland
massachusetts
michigan
minnesota
mississippi
missouri
montana
nebraska
new_hampshire
north_dakota
new_jersey
nevada
new_york
new_mexico
north_carolina
ohio
oklahoma
oregon
pennsylvania
rhode_island
south_carolina
tennessee
texas
utah
vermont
virginia
washington
west_virginia
wisconsin
wyoming
district_of_columbia
US States


Word Counts by State for Reddit
US States
wyoming
wisconsin
west_virginia
washington
virginia
vermont
utah
texas
tennessee
south_carolina
rhode_island
pennsylvania
oregon
oklahoma
ohio
north_dakota
north_carolina
new_york
new_mexico
new_jersey
new_hampshire
nevada
nebraska
montana
missouri
mississippi
minnesota
michigan
massachusetts
maryland
maine
louisiana
kentucky
kansas
iowa
indiana
illinois
idaho
hawaii
georgia
florida
district_of_columbia
connecticut
colorado
california
arkansas
arizona
alaska
alabama
0
5
10
15
20
25
30
Word Counts



20 Most Frequently Occurring Words



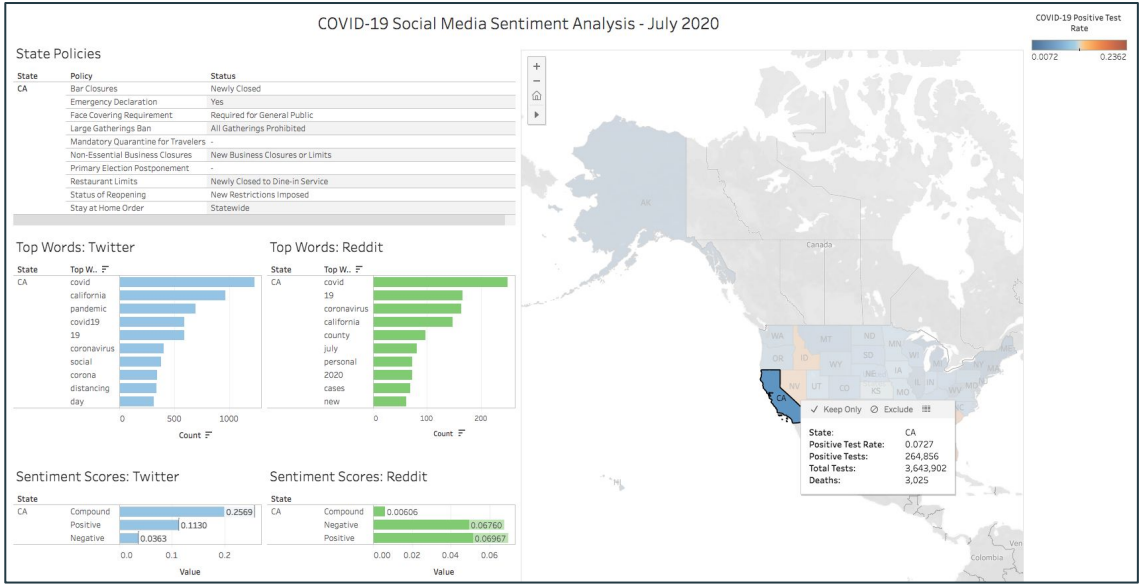
Policy Data

- Status of Reopening
- Stay at Home Order
- Large Gatherings Ban
- Restaurant Limits
- Bar Closures
- Face Covering Requirement
- Emergency Declaration

Tableau

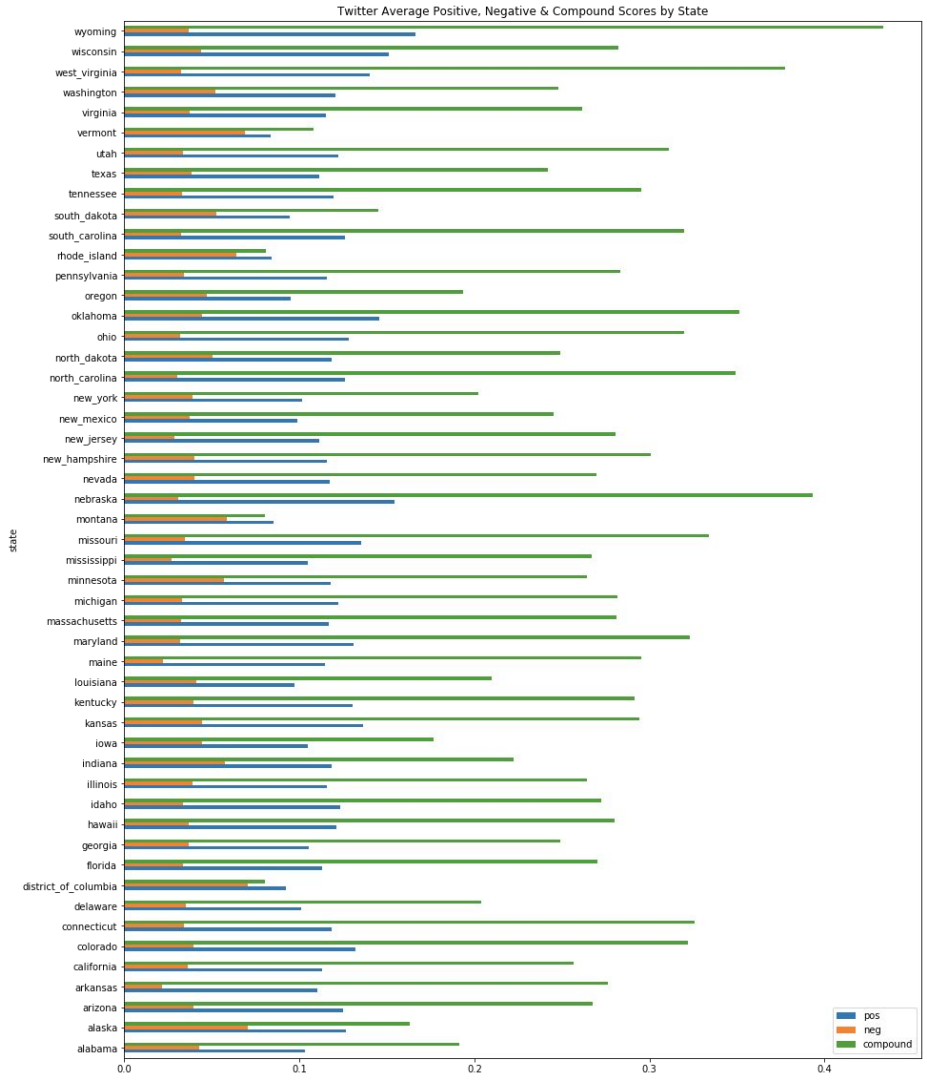
- Created dashboard to visualize state-specific health data, COVID-19 policy statuses, and social media data (sentiment scores and most frequently used words per state)

- Link: https://public.tableau.com/views/project_5_15971471559550/COVID-19SocialMediaSentimentAnalysis-July2020?:language=en&:display_count=y&:origin=viz_share_link



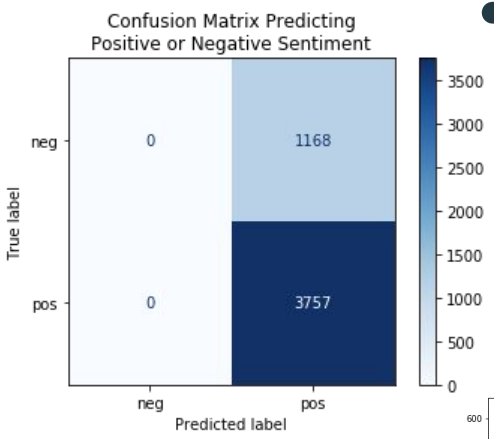
Modeling

Twitter

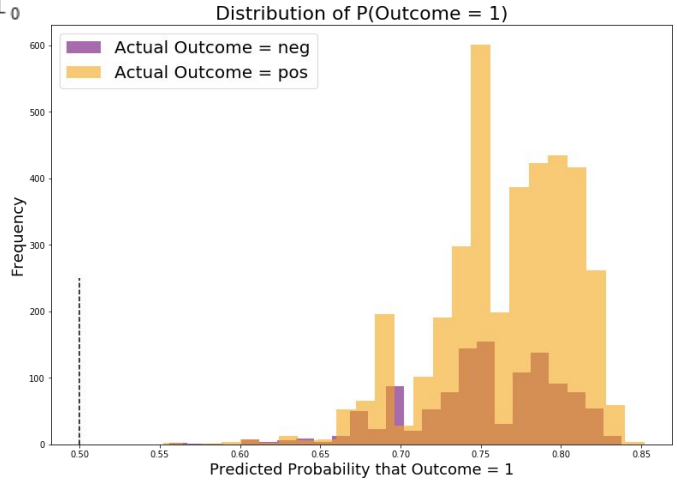

Twitter Average Positive, Negative & Compound Scores by State
state
wyoming
wisconsin
west_virginia
washington
virginia
vermont
utah
texas
tennessee
south_dakota
south_carolina
rhode_island
pennsylvania
oregon
oklahoma
ohio
north_dakota
north_carolina
new_york
new_mexico
new_jersey
new_hampshire
nevada
nebraska
montana
missouri
mississippi
minnesota
michigan
massachusetts
maryland
maine
louisiana
kentucky
kansas
iowa
indiana
illinois
idaho
hawaii
georgia
florida
district_of_columbia
delaware
connecticut
colorado
california
arkansas
arizona
alaska
alabama
pos
neg
compound
0.0
0.1
0.2
0.3
0.4

Twitter

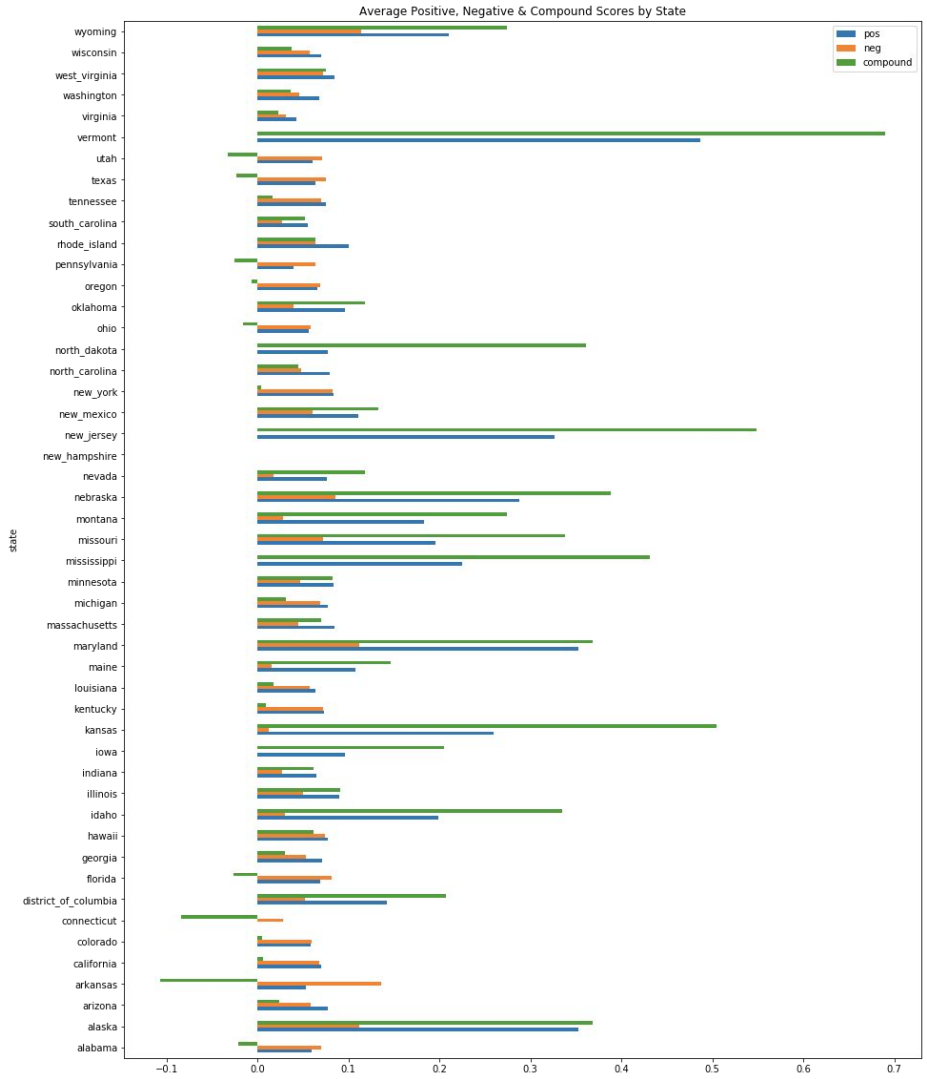
- Linear Regression
  - Predicting **total deaths** due to COVID-19
  - Train R^2: 0.98
  - Test R^2: 0.95



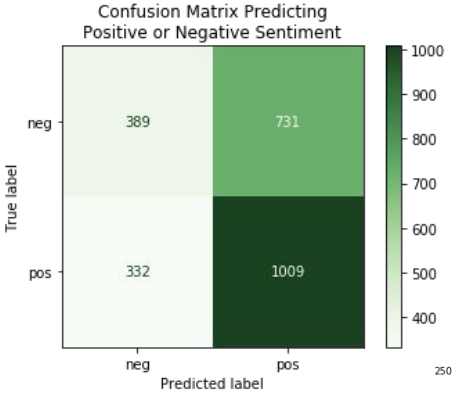
- Random Forest Classifier
  - Predicting **negative or positive sentiments**
  - Train accuracy: 0.76
  - Test accuracy: 0.76



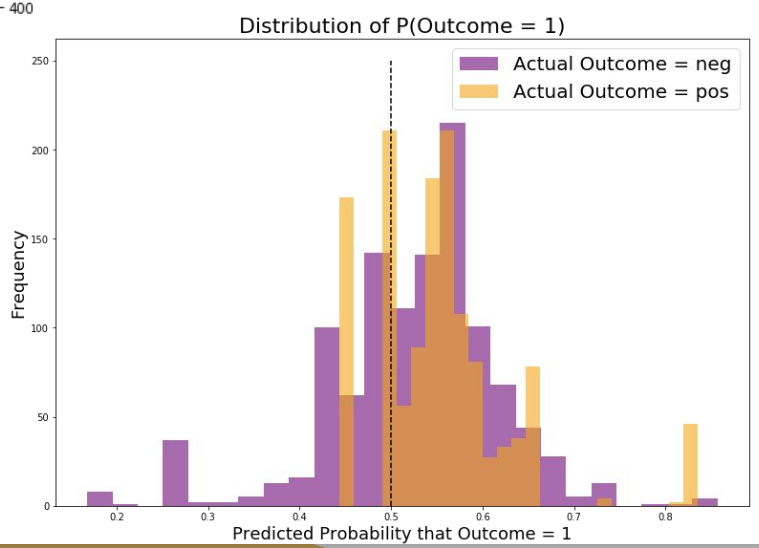
Reddit


Average Positive, Negative & Compound Scores by State
pos
neg
compound
state
wyoming
wisconsin
west_virginia
washington
virginia
vermont
utah
texas
tennessee
south_carolina
rhode_island
pennsylvania
oregon
oklahoma
ohio
north_dakota
north_carolina
new_york
new_mexico
new_jersey
new_hampshire
nevada
nebraska
montana
missouri
mississippi
minnesota
michigan
massachusetts
maryland
maine
louisiana
kentucky
kansas
iowa
indiana
illinois
idaho
hawaii
georgia
florida
district_of_columbia
connecticut
colorado
california
arkansas
arizona
alaska
alabama
−0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7

Reddit



- Random Forest Classifier
  - Predicting **negative or positive sentiments**
  - Train accuracy: 0.59
  - Test accuracy: 0.56



- Linear Regression
  - Predicting **total deaths** due to COVID-19
  - Train R^2: 0.98
  - Test R^2: 0.98

Conclusions and Recommendations

Conclusion and Next Steps

One next step would be to take a look at the months before and after July to see which state policies had the biggest impact on lowering positive test cases and deaths within each state. Creating a model based off of this could potentially highlight the best steps a government should take to combat a pandemic like COVID-19 in the future.

Another next step would be to take a look at sentiment of each state based on policy, but certain states were very difficult to gather data on, especially the lower population states. Learning how each state felt about their respective policies and potential emotional impact could prove very useful in the future.

Future projects like this one might want to take a closer look at other social media platforms as a way to gather data on public sentiment.

Questions?