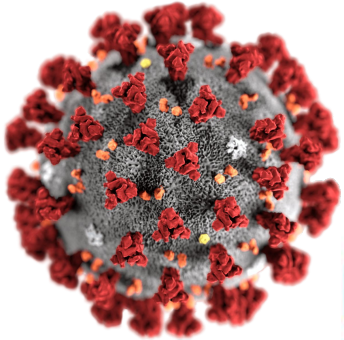
#Covid19



twitter

Investigating the relationship between social media activity and the coronavirus

Group 6: Andrew Cummings,
Oliver Posewitz, Ian Ustanik

Agenda

- Data Acquisition
  - Twitter
  - Covid-19
- Data Preprocessing
  - Compilation
  - Cleansing
  - Aggregation
- Methods of Analysis
  - Statistical
  - NLP - Sentiment
  - Regression/Location

```python
def tweet_search(hashtag):
    query = hashtag
    count = 1000
    searched_tweets = [status for status in tweepy.Cursor(api.search, q=query).items(count)]

    tweet_list = list(searched_tweets)
    tweet_json = [tweet._json for tweet in tweet_list]
    return tweet_json

def tweets_to_df(tweet_json):
    ID = []
    tweets = []
    location = []
    time = []

    for tweet in tweet_json:
        ID.append(tweet['id'])
        tweets.append(tweet['text'])
        location.append(tweet['user']['location'])
        time.append(tweet['created_at'])

    data_tuples = list(zip(ID, tweets,location,time))
    tweets_df = pd.DataFrame(data_tuples, columns=['ID','tweets','location','time'])
    return tweets_df
```

Data Acquisition

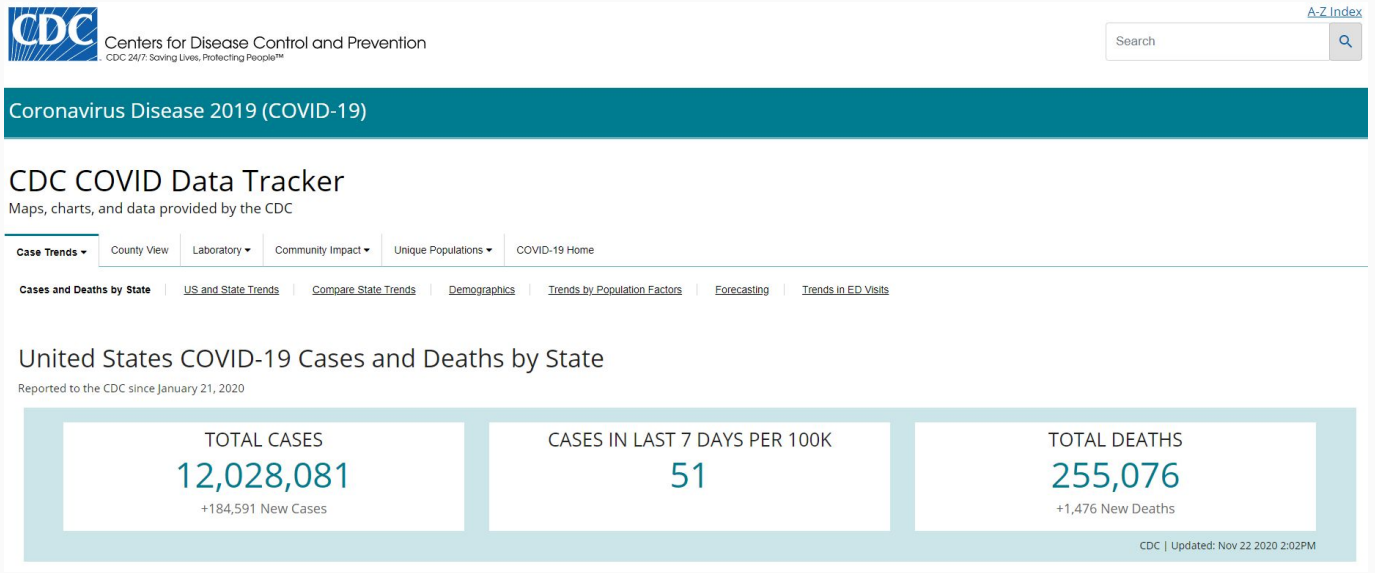
Twitter Data

- Tweepy request to Twitter API
- Use three ‘bins’ of keywords:
  - #covid19
  - #maskup
  - #travel
- Request limitations → Pagination
- Retrieved ID, tweets, location, and time
- 21,000 tweets total over 7-day period

	ID	tweets	location	time
0	1328125627812114433	RT @ArTallks: GodSpeed 👴🚀👦🚀🌑🚀👴🚀@SpaceX ...	Los Angeles Universe Omniverse	Sun Nov 15 23:59:59 +0000 2020
1	1328125567271538688	As if we needed evidence that the Pandemic is ...	#notinabubble	Sun Nov 15 23:59:45 +0000 2020
2	1328125516222754818	RT @nonsequiteuse: Hey @HiltonHouston & @H...	Houston, Texas	Sun Nov 15 23:59:33 +0000 2020
3	1328125492545798144	RT @cartoonnetwork: Rep your fav #CartoonNetwo...	someone's mind	Sun Nov 15 23:59:27 +0000 2020
4	1328125490255835136	Waiting to take a pandemic seriously until you...	Norman, OK	Sun Nov 15 23:59:26 +0000 2020
...	...	...	...	...
995	1328063679926038531	RT @Alt_ReddTruq: #MaskUp	Florida, USA	Sun Nov 15 19:53:50 +0000 2020
996	1328063582974734339	Do you think racist folks will finally get #BL...	Texas, USA	Sun Nov 15 19:53:27 +0000 2020
997	1328063420739031043	RT @cartoonnetwork: Rep your fav #CartoonNetwo...	Naahhhh, have a cawffee -w-	Sun Nov 15 19:52:48 +0000 2020
998	1328063393438322688	For those of you that need a reminder. #maskup...	New York, NY	Sun Nov 15 19:52:41 +0000 2020
999	1328063376891801600	RT @cartoonnetwork: Rep your fav #CartoonNetwo...		Sun Nov 15 19:52:37 +0000 2020

Covid-19 Data

- Center for Disease Control website



Data Preprocessing

Cleansing and Purging

- Used conditionals to tag records with location data like state name in Twitter data, formatted time attribute
- Concatenate each bin of tweets into one dataframe

location
Universe Omniverse
#notinabubble
Houston, Texas
someone's mind
Norman, OK

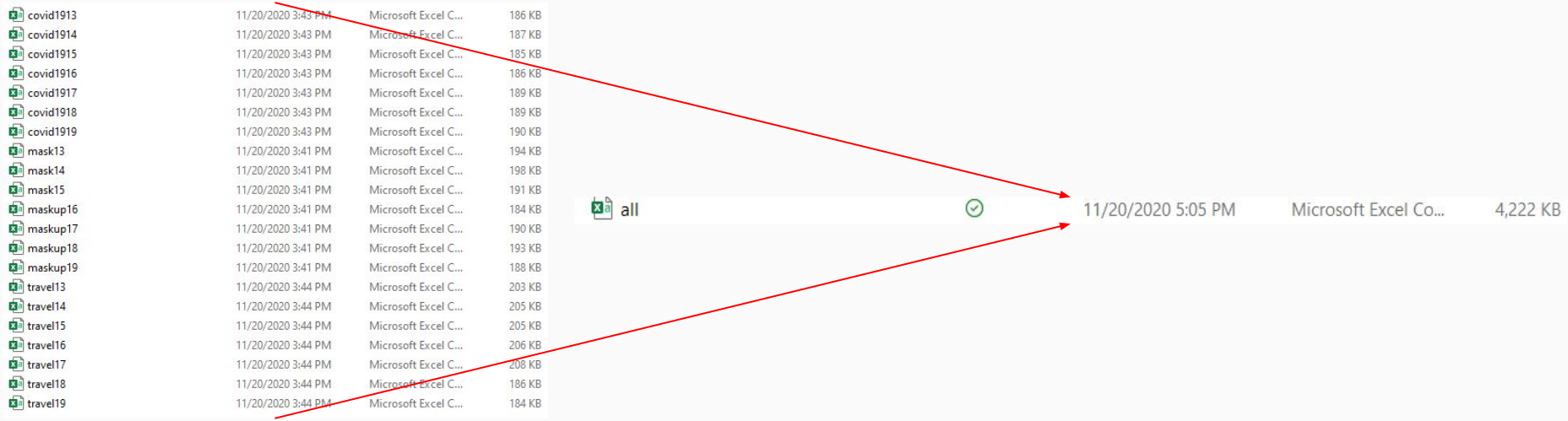
location
Miami, FL
Houston, TX
Montville, NJ
MD
The Woodlands, TX

location
FL
TX
NJ
MD
TX

```
#stitch csv's together

import os

directory = 'travel'

li = []

for csv_file in os.listdir(directory):
    mask_df = pd.read_csv(directory+'/'+csv_file, index_col=None, header = 0)
    li.append(mask_df)

travel_df = pd.concat(li, axis=0, ignore_index=True)

# travel_df.to_csv('travel.csv')
```

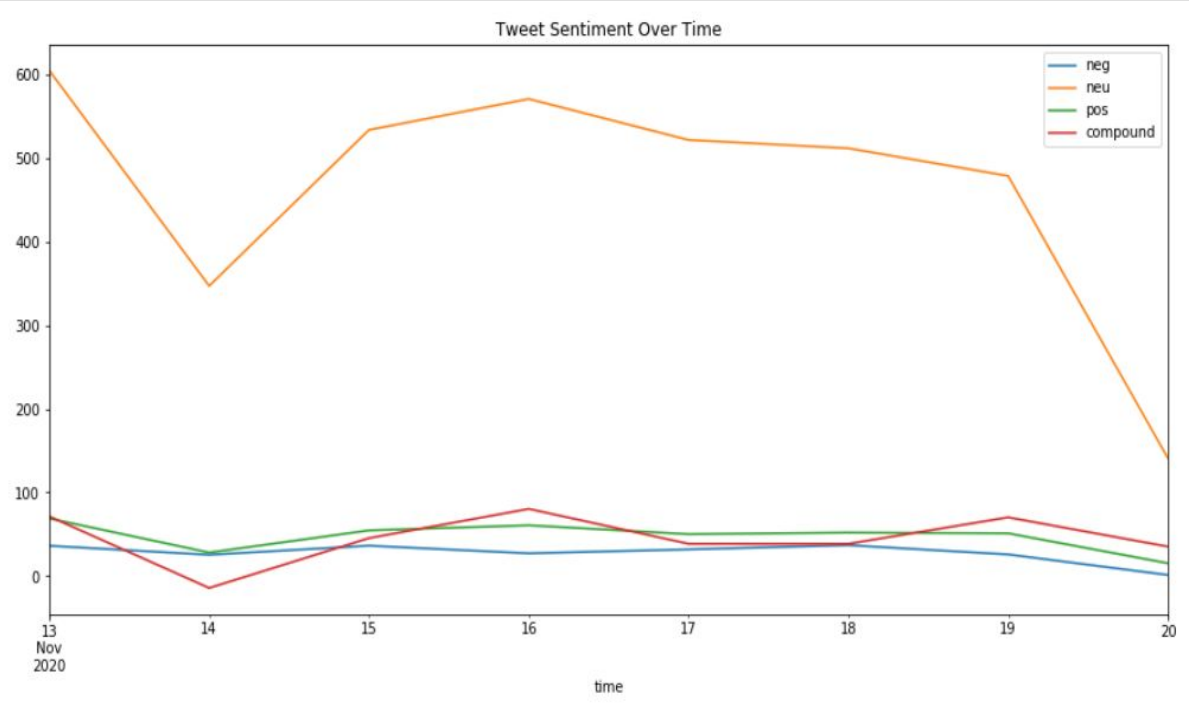
Combining Different Data Sets

- Data was initially stored in many separate csv's
- Iterate over csv's using os package to create one file for each bin



Aggregation and Summarization

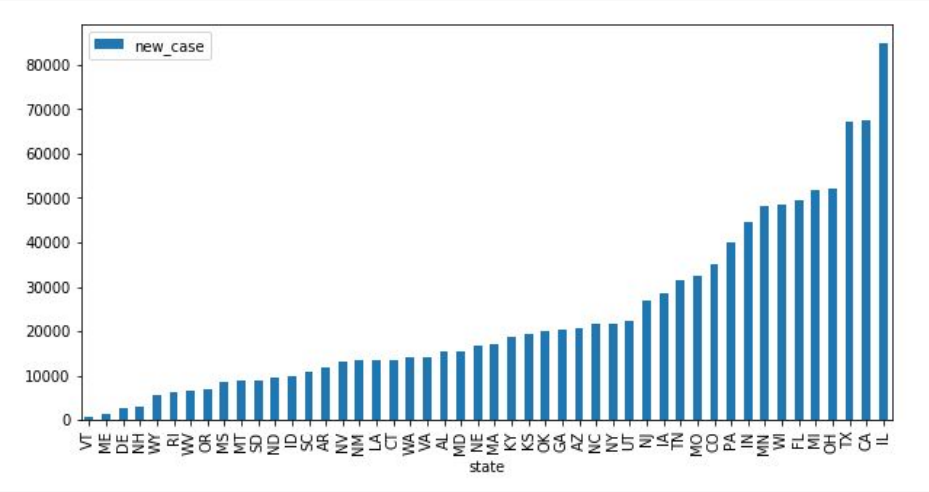
- Dimension:
  - 21,000 rows
  - 9 columns
- Continental US
  - Lower 48
- 7-day period
  - (11/13-11/19)
- 4,314 tweets
  - Location formatted



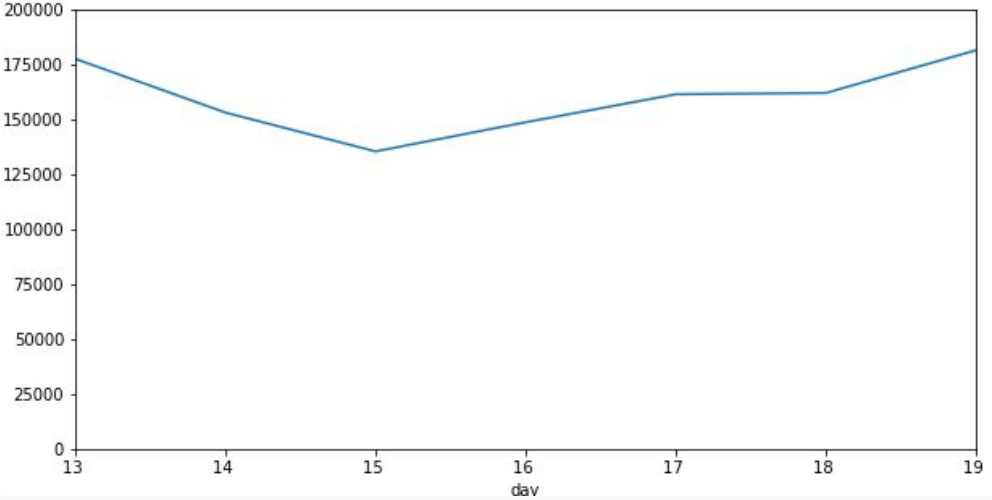
Methods of Analysis

What is the current state of Covid-19?

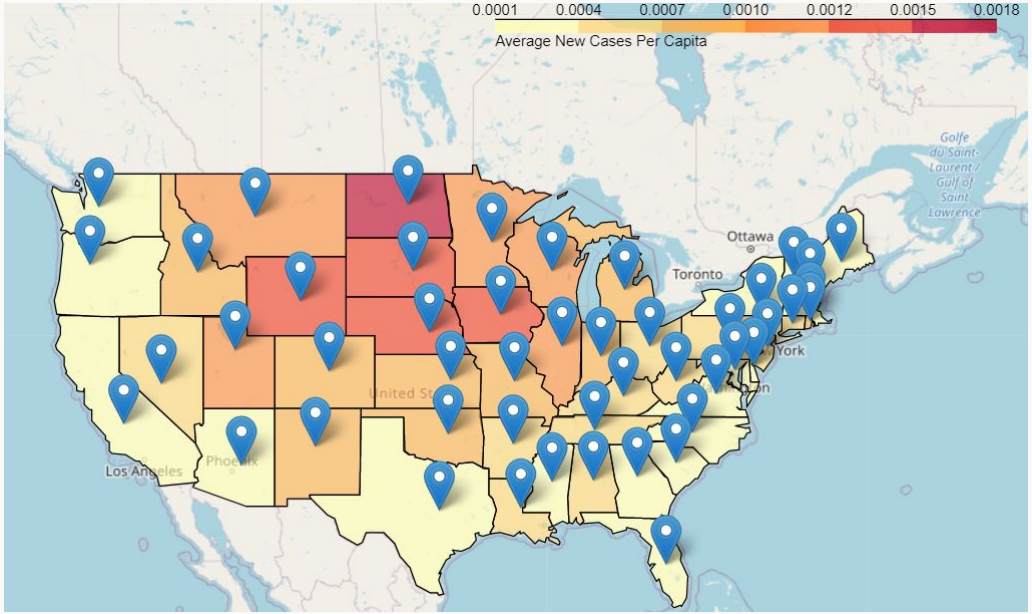
New Cases By State



New Cases By Day



Supporting Visualization



Conclusion

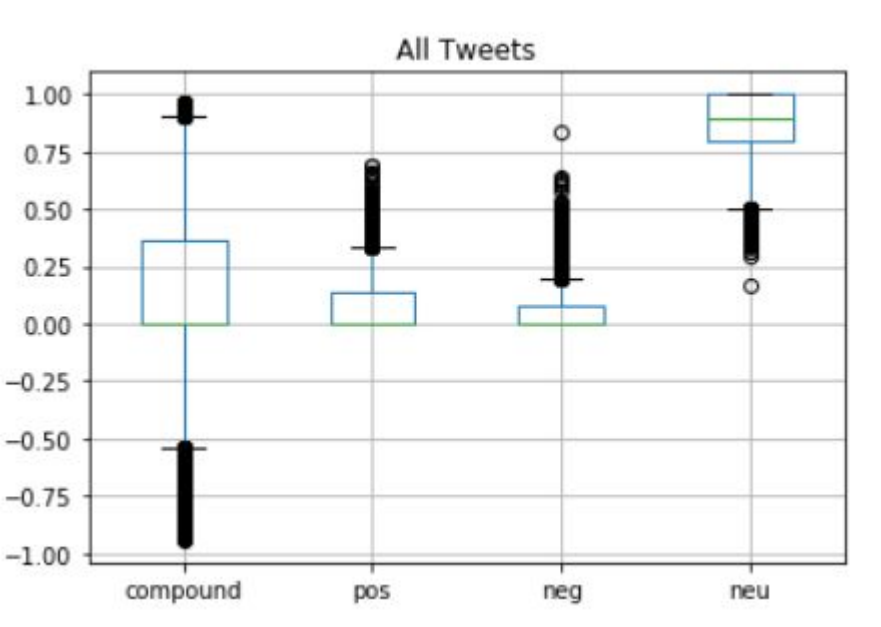
- Top 3 states for new cases: Illinois, California, Texas

state	tot_cases	conf_cases	new_case	tot_death	conf_death	prob_death	new_death
IL	4099809	4099809.0	84841	78778	75762.0	3016.0	802
CA	7200493	0.0	67658	127996	0.0	0.0	358
TX	7216532	0.0	67042	137620	0.0	0.0	966

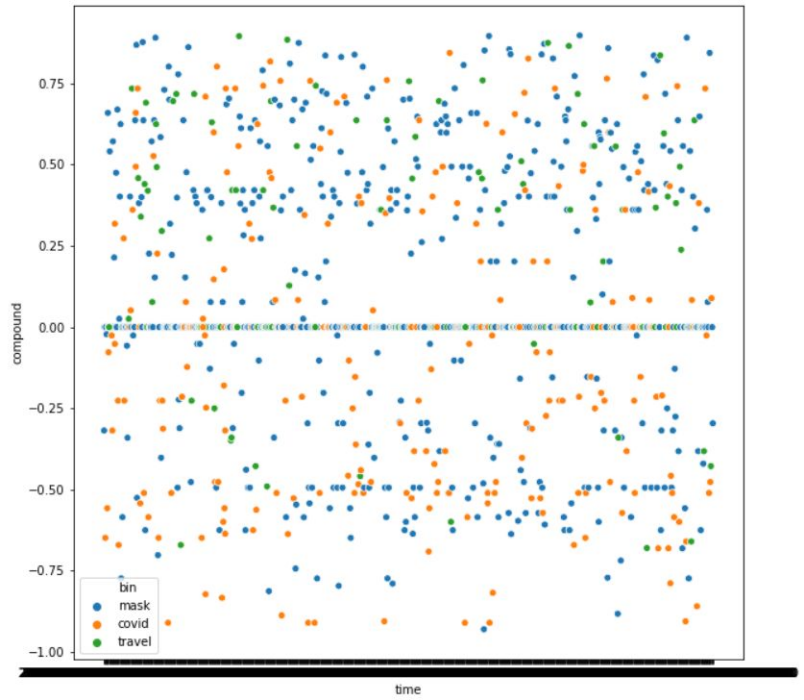
Focusing on new cases, we can look at distribution by state and by day in order to identify trends that may correlate with parallel trends with social media.

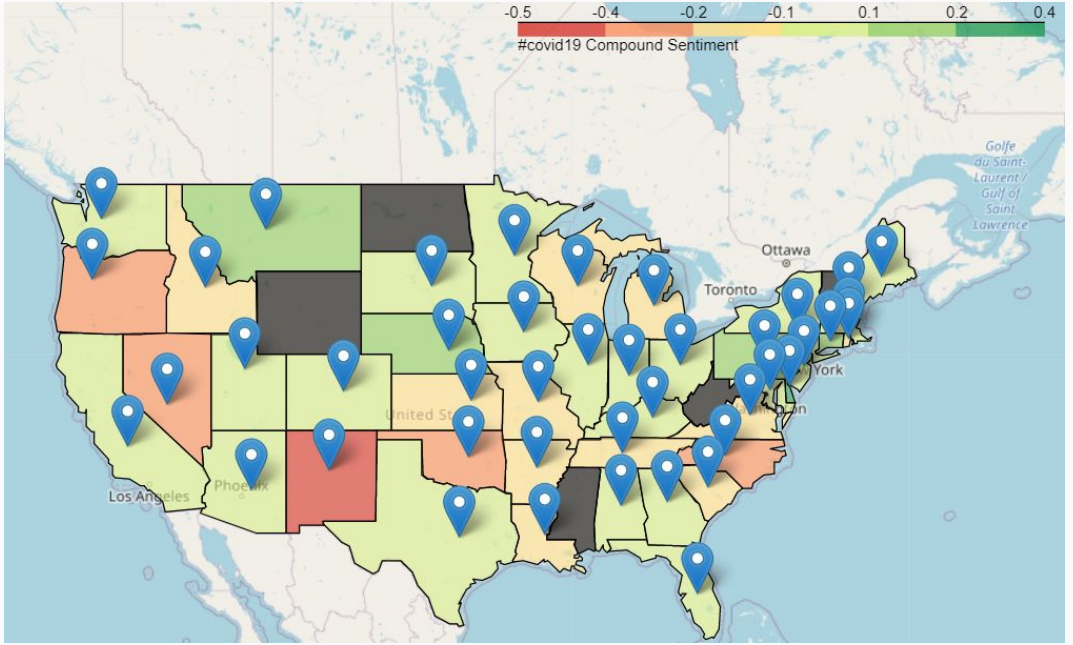
What is the sentiment of coronavirus-related tweets?

- Used VADER model for sentiment analysis - NLTK
- Returned positive, neutral, negative, and compound scores
- Applied on all hashtags and for all locations



Supporting Visualization





Conclusion

- While #travel had the highest compound score:
  - #covid19 → Highest negative
  - #maskup → Highest positive
- New Mexico and North Carolina
  - High % change of new cases
  - High negative score for #covid19
- Sentiment scores somewhat consistent over 7-day period
- Correlation warrants further investigation

	bin	neg	neu	pos	compound
2	travel	0.020455	0.893188	0.086358	0.170609
1	mask	0.049105	0.854347	0.096553	0.105901
0	covid	0.081858	0.845110	0.073033	-0.029543

Does a correlation exist between social media sentiment and Covid-19 in the United States?

- Carried out regression and location analysis

```
#define response variable
y = data['compound']

#define explanatory variable
x = data[['time']]

#add constant to predictor variables
x = sm.add_constant(x)

#fit linear regression model
model = sm.OLS(y, x).fit()

#view model summary
print(model.summary())
```

OLS Regression Results

Dep. Variable:	compound	R-squared:	0.003
Model:	OLS	Adj. R-squared:	0.003
Method:	Least Squares	F-statistic:	957.7
Date:	Mon, 23 Nov 2020	Prob (F-statistic):	6.04e-210
Time:	14:15:08	Log-Likelihood:	-1.6245e+05
No. Observations:	303000	AIC:	3.249e+05
Df Residuals:	302998	BIC:	3.249e+05
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-8408.2516	271.705	-30.946	0.000	-8940.786	-7875.718
time	0.0114	0.000	30.946	0.000	0.011	0.012

Omnibus:	7583.644	Durbin-Watson:	0.006
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3814.192
Skew:	-0.005	Prob(JB):	0.00
Kurtosis:	2.450	Cond. No.	2.67e+11

Regression Examples

OLS Regression Results

Dep. Variable:	new_case	R-squared:	0
Model:	OLS	Adj. R-squared:	0
Method:	Least Squares	F-statistic:	4
Date:	Mon, 23 Nov 2020	Prob (F-statistic):	8.34
Time:	14:26:13	Log-Likelihood:	-1.5837
No. Observations:	173619	AIC:	3.167
Df Residuals:	173616	BIC:	3.167
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.
const	1.773e+07	1.89e+06	9.360	0.000	1.4e+07	2.14
time	-24.0374	2.568	-9.359	0.000	-29.071	-19
compound	14.4045	12.976	1.110	0.267	-11.027	39

Omnibus:	112920.449	Durbin-Watson:	0
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1818704
Skew:	2.919	Prob(JB):	
Kurtosis:	17.742	Cond. No.	2.63

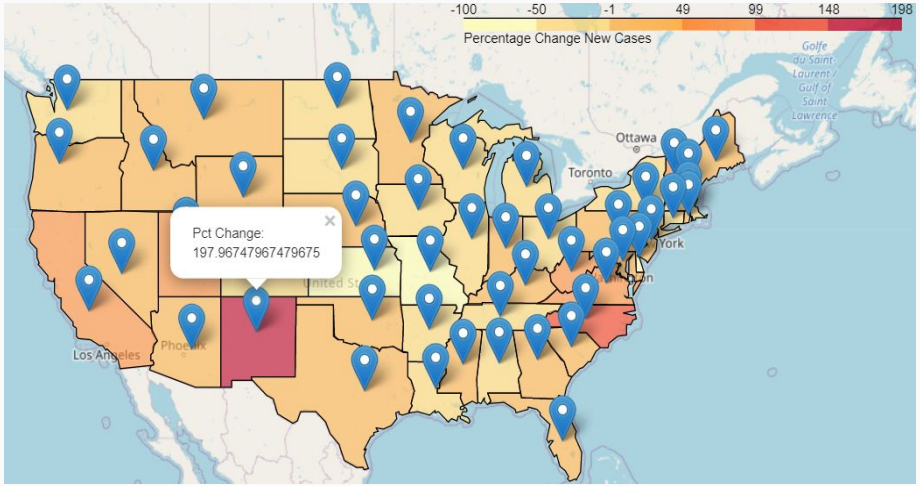
OLS Regression Results

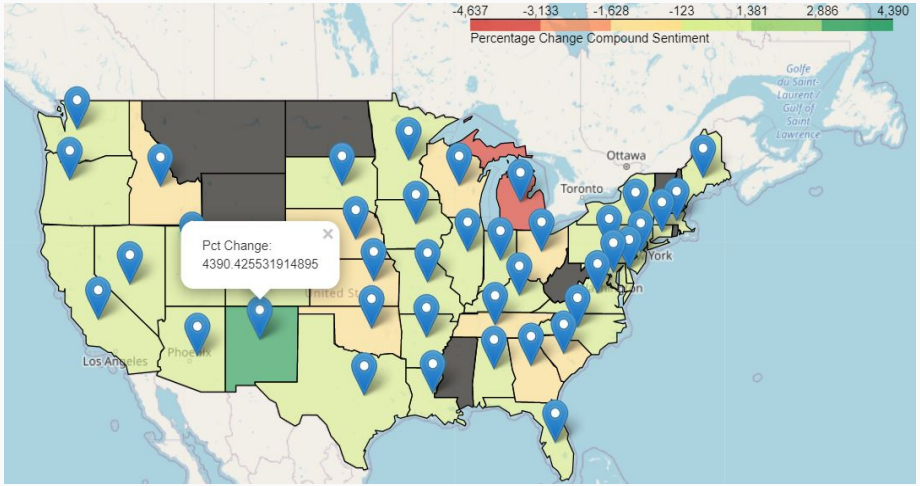
Dep. Variable:	compound	R-squared:	
Model:	OLS	Adj. R-squared:	
Method:	Least Squares	F-statistic:	
Date:	Mon, 23 Nov 2020	Prob (F-statistic):	4.46
Time:	14:30:41	Log-Likelihood:	-1.624
No. Observations:	303000	AIC:	3.24
Df Residuals:	302997	BIC:	3.24
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0
const	-8402.0063	271.711	-30.923	0.000	-8934.552	-786
time	0.0114	0.000	30.923	0.000	0.011	
new_case	-8.856e-07	3.15e-07	-2.815	0.005	-1.5e-06	-2.6

Omnibus:	7574.269	Durbin-Watson:	
Prob(Omnibus):	0.000	Jarque-Bera (JB):	381
Skew:	-0.005	Prob(JB):	
Kurtosis:	2.451	Cond. No.	2.6

Supporting Visualization





Conclusion

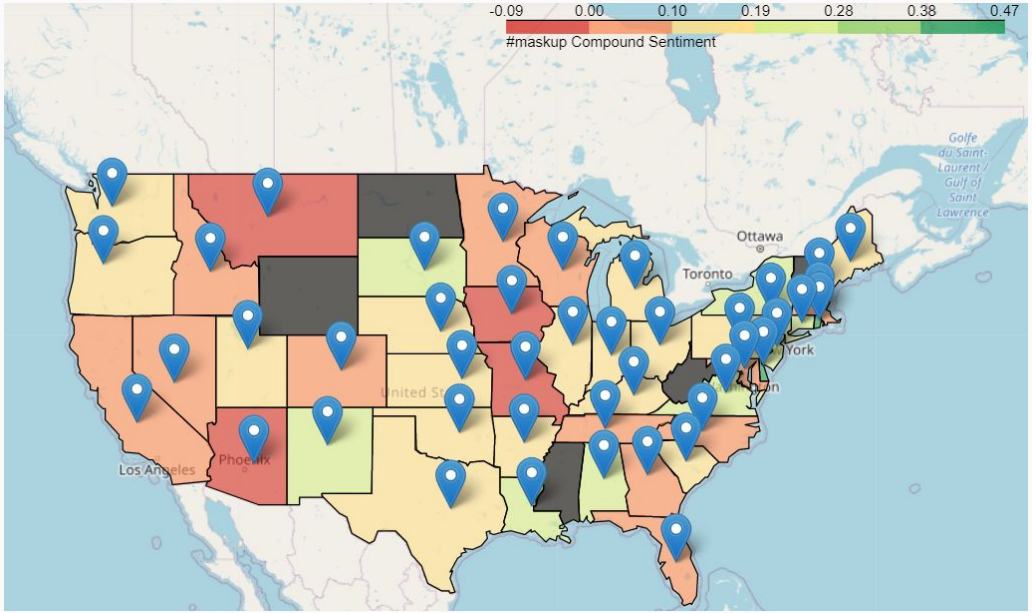
-Limited by time frame of publicly accessible tweets

-Data thoroughly tested for any correlation

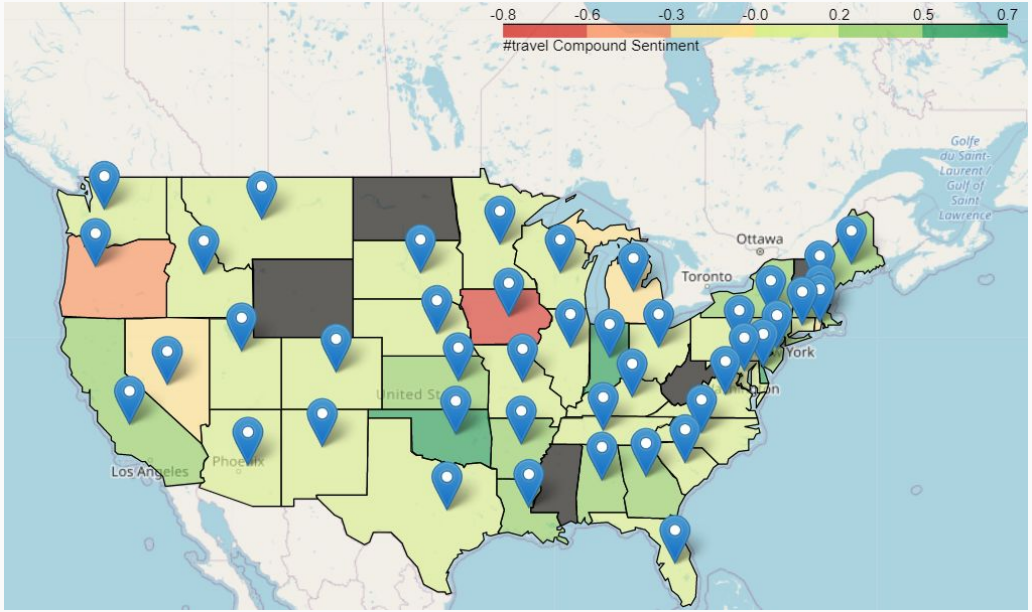
-Process projected on a larger scale would yield more statistically significant results

Thank You!

Appendix I


-0.09
0.00
0.10
0.19
0.28
0.38
0.47
#maskup Compound Sentiment
Golfe du Saint-Laurent / Gulf of Saint Lawrence
Ottawa
Toronto
York
Los Angeles
Phoenix

Appendix II


-0.8
-0.6
-0.3
-0.0
0.2
0.5
0.7
#travel Compound Sentiment
Golfe du Saint-Laurent / Gulf of Saint Lawrence
Ottawa
Toronto
York
Los Angeles
Phoenix