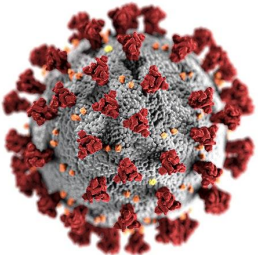

Coronavirus

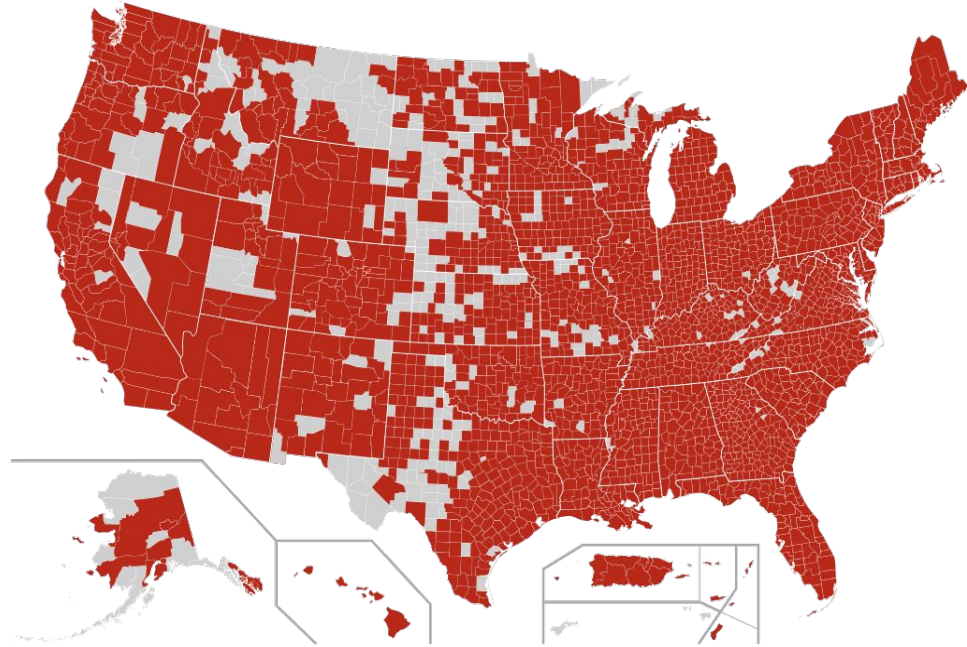
Team R²

Tariq Attarwala, Preston Hinkel, Sai Reddy,
Addison Sams, and Veronica Valencia



Introduction

Given the recent pandemic, our team's goal is to evaluate the relationship between income per capita, social distancing efforts, number of administered test, and population density on the number of confirmed cases by US counties and states.



—

Hypothesis 1a:

Income per capita does not impact the number of confirmed cases by US county.

Data

	date	county	state	fips	cases	deaths
0	2020-01-21	Snohomish	Washington	53061.0	1	0
1	2020-01-22	Snohomish	Washington	53061.0	1	0
2	2020-01-23	Snohomish	Washington	53061.0	1	0
3	2020-01-24	Cook	Illinois	17031.0	1	0
4	2020-01-24	Snohomish	Washington	53061.0	1	0

Covid-19 data for cases by county pulled from *New York Times* COVID-19 Tracker

	county	state	2016	2017	2018
0	United States	NaN	49870.0	51885.0	54446.0
1	Alabama	Alabama	39224.0	40467.0	42238.0
2	Autauga	Alabama	39561.0	40450.0	41618.0
3	Baldwin	Alabama	42907.0	43989.0	45596.0
4	Barbour	Alabama	31595.0	33048.0	35199.0

Income per capita data pulled from the Bureau of Economic Analysis based on Census data for 2016, 2017, and 2018

3 of the top income
counties for 2018 are
located in

CALIFORNIA

3 of the bottom income
counties for 2018 are
located in **Georgia**



Counties

- Santa Clara
- San Francisco
- San Mateo

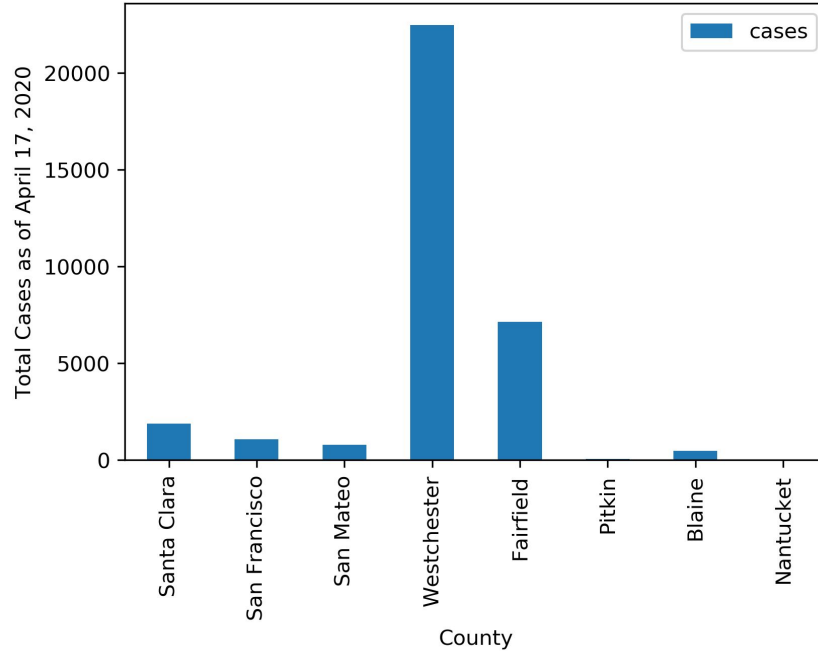


Counties

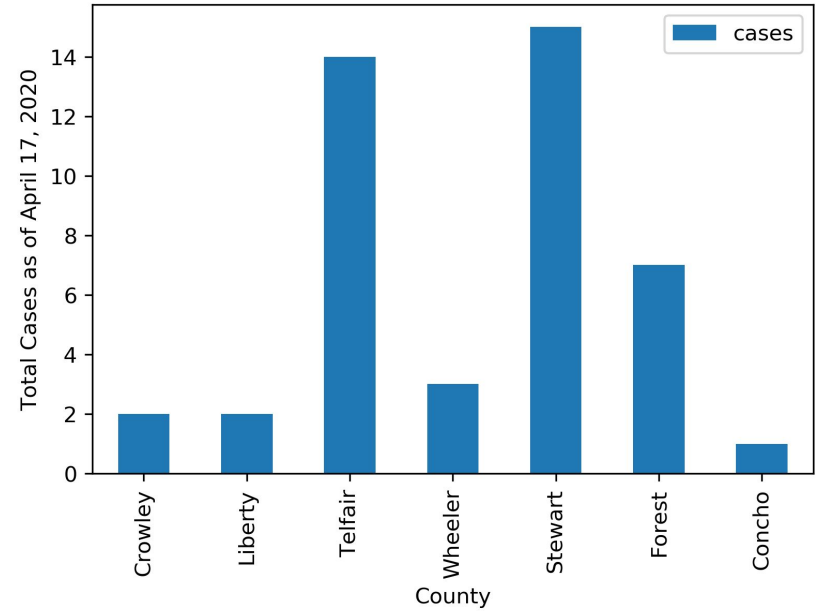
- Stewart
- Telfair
- Wheeler

Graphs

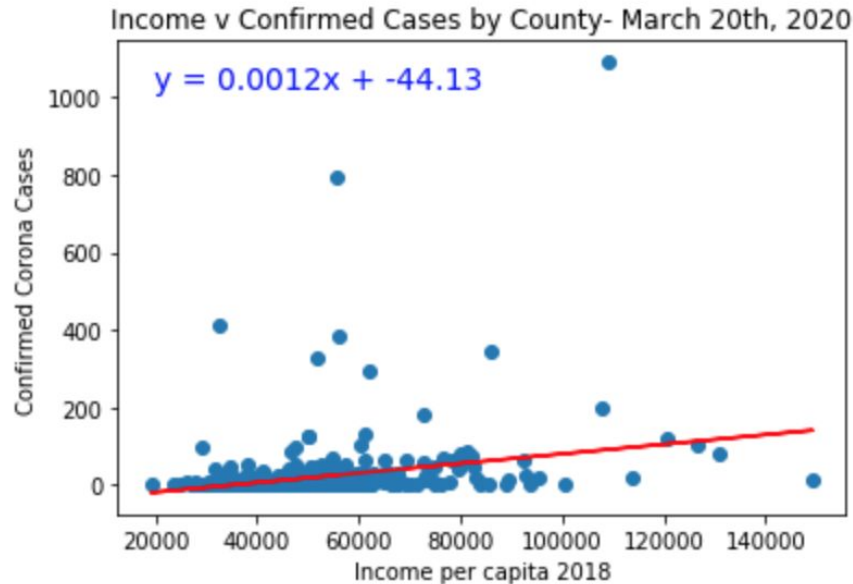
Confirmed Cases for the Top 8 - 2018 Income Counties



Confirmed Cases for the Bottom 7 - 2018 Income Counties



Graph 1

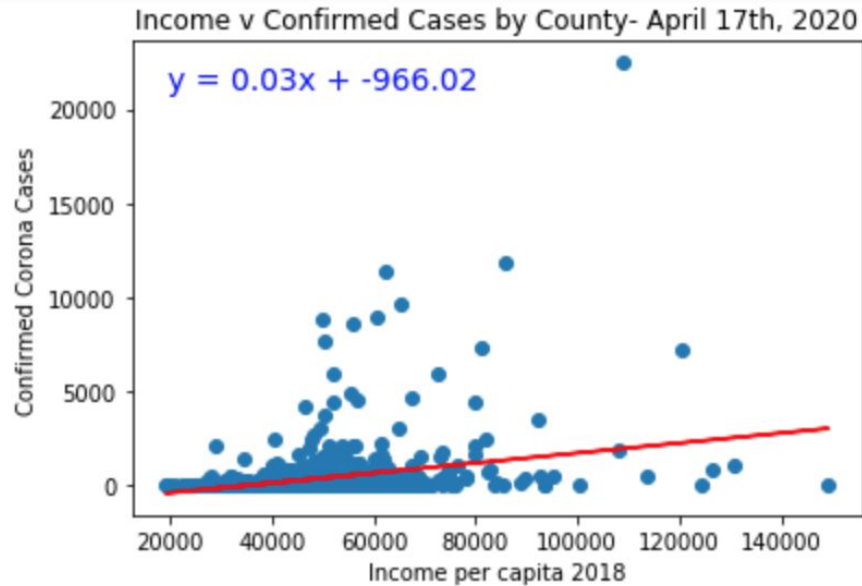


	coef	std err	t	P> t	[0.025	0.975]
2018	0.0004	4.86e-05	8.047	0.000	0.000	0.000

R squared: 0.07991299959205628

The correlation between income per capita and confirmed corona cases is 0.28

Graph 2



	coef	std err	t	P> t	[0.025	0.975]
2018	0.0062	0.001	11.361	0.000	0.005	0.007

R squared: 0.10118349600866594

The correlation between income per capita and confirmed corona cases is 0.32

Limitations of Analysis

- Income calculated based on Census data, this data often limited to people who complete survey
- Data not updated to reflect incomes for 2019 or Q1 of 2020
- External factors may contribute to the overall confirmed cases across counties regardless of income

Conclusion

- **Reject null hypothesis- $p \text{ value} < 0.05$**
- **Income does impact the number of confirmed corona cases by county.**

—
Hypothesis 1b:

**Average Household income
does not impact the number
of deaths caused by
COVID-19 by state**

Data

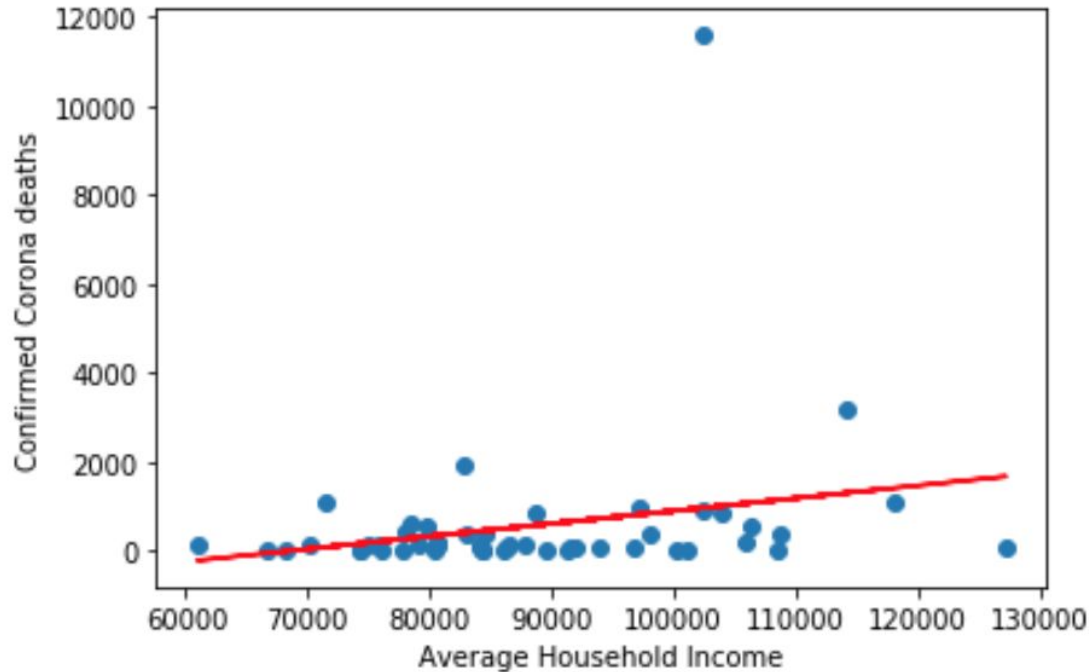
	date	state	fips	cases	deaths
1	2020-01-21	Washington	53	1	0
2	2020-01-22	Washington	53	1	0
3	2020-01-23	Washington	53	1	0
4	2020-01-24	Illinois	17	1	0
5	2020-01-24	Washington	53	1	0
6	2020-01-25	California	06	1	0
7	2020-01-25	Illinois	17	1	0
8	2020-01-25	Washington	53	1	0
9	2020-01-26	Arizona	04	1	0
10	2020-01-26	California	06	2	0
11	2020-01-26	Illinois	17	1	0
12	2020-01-26	Washington	53	1	0
13	2020-01-27	Arizona	04	1	0
14	2020-01-27	California	06	2	0
15	2020-01-27	Illinois	17	1	0

**Covid-19 data for cases by county pulled
from *New York Times* COVID-19 Tracker**

	state	Income
1	Alabama	70090.35
2	Alaska	91268.57
3	Arizona	87823.18
4	Arkansas	68079.14
5	California	102375.47
6	Colorado	98129.07
7	Connecticut	103865.55
8	Delaware	93957.11
9	District of Columbia	127263.58
10	Florida	78433.46

**Average Household Income data is
based on Census data for 2019**

Analysis



The r^2 value is 0.06, which indicates that Average Household income only explains 6% of the variation in confirmed corona deaths.

Conclusion

	coef	std err	t	P> t	[0.025	0.975]
Income	0.0069	0.003	2.654	0.011	0.002	0.012

Since our p value is 0.011, which is less than 0.05, we can reject the null hypothesis.

Omnibus:	102.976	Durbin-Watson:	2.029
Prob(Omnibus):	0.000	Jarque-Bera (JB):	3106.345
Skew:	5.926	Prob(JB):	0.00
Kurtosis:	39.350	Cond. No.	1.00

Therefore, we can safely reject the null hypothesis, and say that Average Household income by state impacts the number of coronavirus deaths in each state.

Limitations

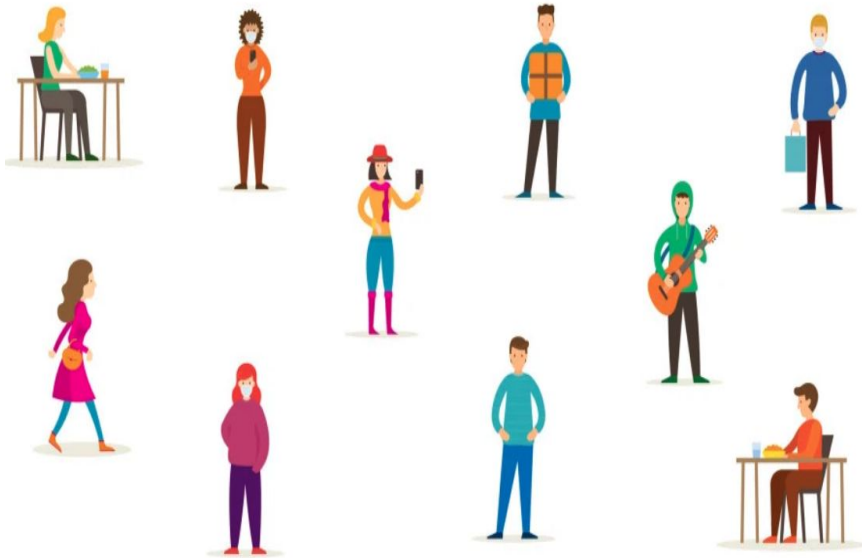
1. As the p value indicates, there are several other factors that impact the death rate
2. Average Household income tracks families and not individuals
3. All states in the US do not have equal levels of testing

—

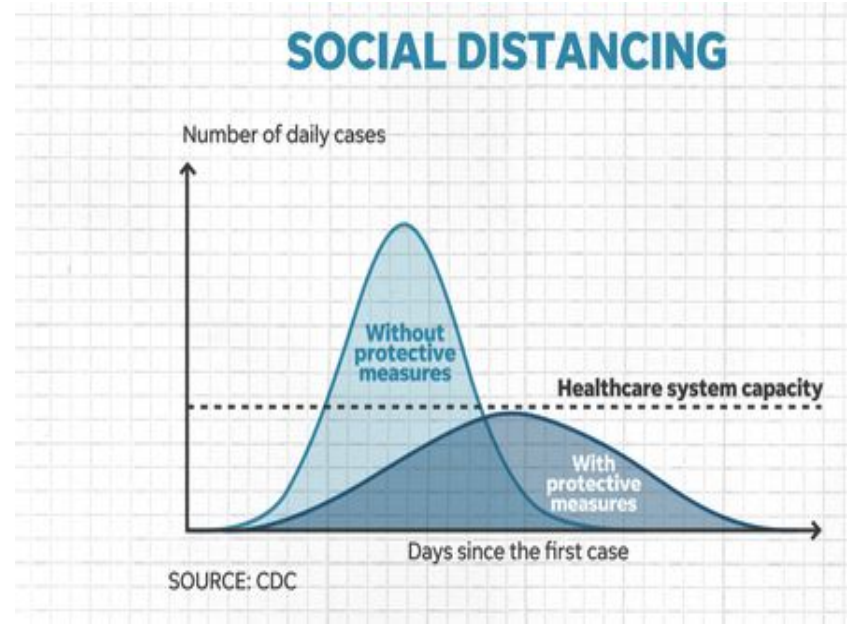
Hypothesis 2:

The rate of cases spreading around the US is not impacted by the social distancing measures and lockdown status enforced by US states/counties.

Coronavirus: Is social distancing working?



SOCIAL DISTANCING



Identifying data sources


- ❖ How do you measure social distancing?

Google COVID-19 Community Mobility Reports



See how your community is moving around differently due to COVID-19

- ❖ How do you quantify, If it is working?

HomeAPIsCovid-19 DataGet StartedSign In

Covid-19 Data

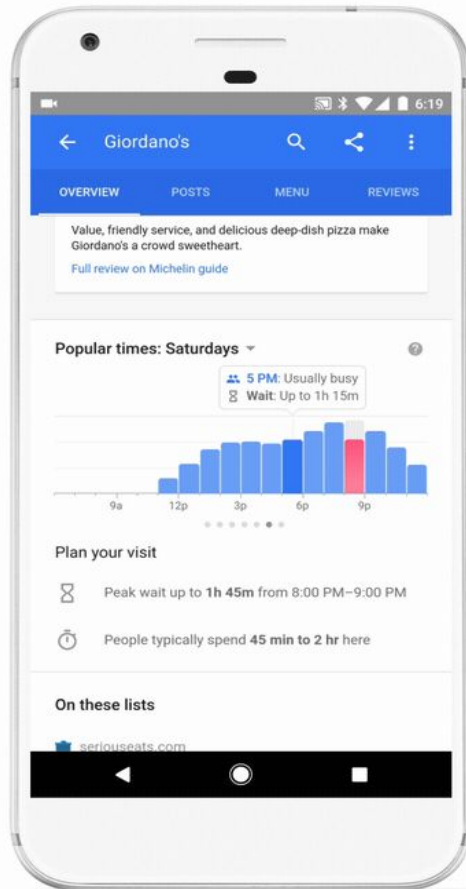
Coronavirus (Covid-19) Data

The New York Times is releasing a series of data files with cumulative counts of coronavirus cases in the United States, at the state and county level, over time. We are compiling this time series data from state and local governments and health departments in an attempt to provide a complete record of the ongoing outbreak.

For more information see:

- <https://github.com/nytimes/covid-19-data>
- <https://www.nytimes.com/article/coronavirus-county-data-us.html>

Build Data Retrieval plan



README.md

Covid-19 Mobility Tracker

last synced from source **April 22th** last update from google **April 16th**

Note

Note: Google recently started publishing raw CSV data for this, and following it is recommended for all future updates.

This project would still continue update sources, but it might not be as realtime.

[Google Mobility Reports](#) show aggregate activity in each country, and how it changes in response to policies aimed at combating COVID-19. However, it is only published as a PDF and the data isn't available in a machine-readable format that could enable more richer analysis.

This is an effort to reverse-engineer the PDFs into vectors and ultimately into time-series data available as a JSON Rest API.

Assemble and Clean Data

	date	cases	deaths
0	2020-01-21	1	0
1	2020-01-22	1	0
2	2020-01-23	1	0
3	2020-01-24	2	0
4	2020-01-25	3	0

Covid 19 Cases

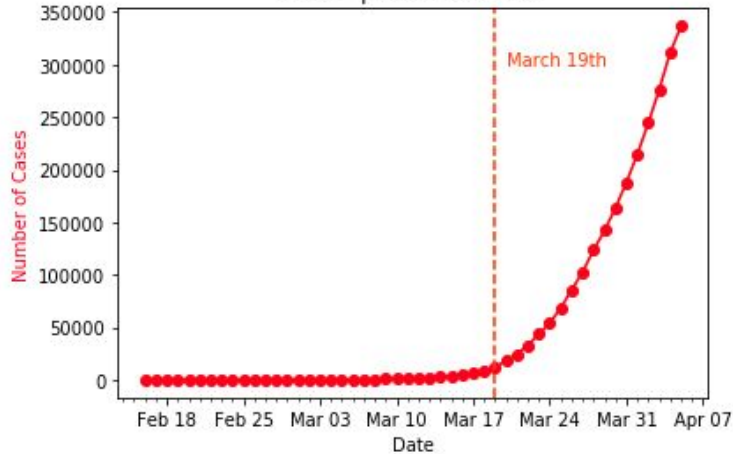
	date	Per_decrease_ret_rec	Per_decrease_parks	Per_decrease_grocery	Per_decrease_transit	Per_decrease_workplace	Per_decrease_residential
49	2020-02-16	6	28	0	-9	-23	5
48	2020-02-17	1	7	0	1	-2	1
47	2020-02-18	2	8	0	1	1	0
46	2020-02-19	1	5	0	0	0	1
45	2020-02-20	2	4	-2	1	0	0

Mobility Data

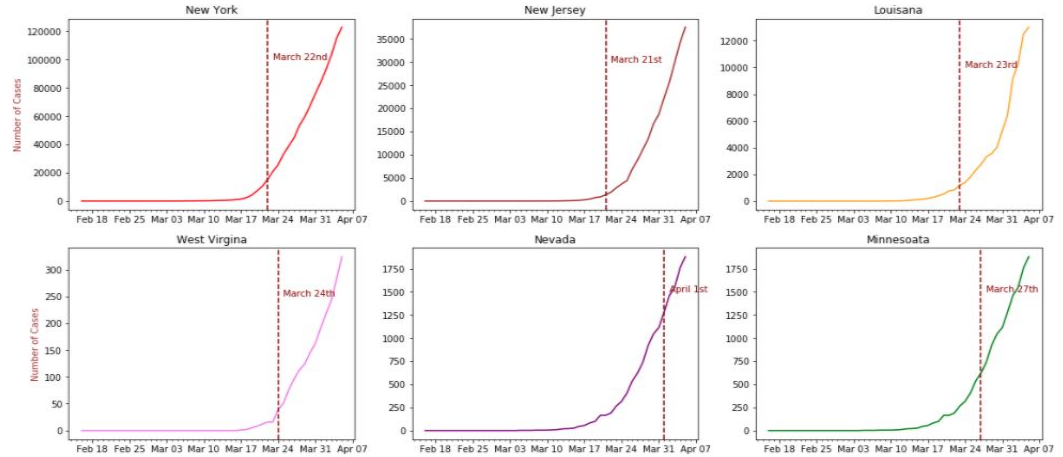
- The mobility data for 5 major U.S territories was not published by google, as a result they were excluded from the analysis.

Analyze for Trends

Daily Case count
Feb - April 2020 for USA

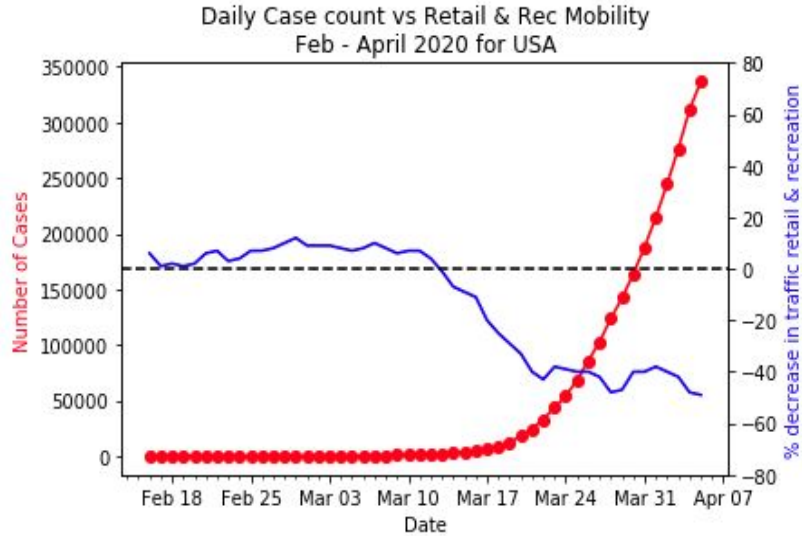


Top 3 Worst & Best performing States
Feb - April 2020



- As of April 7th, we still see an increasing trend in covid-19 new case count for USA.
- The rate of increase of Covid-19 new cases has decreased in the first week of April
- The case count for Louisiana has reached its peak and new case count is trending downwards.
- The timing of the lockdown coincides with the period of exponential growth.
- Nevada and Minnesota are the best performing states, In Spite of significant delay in lock down.

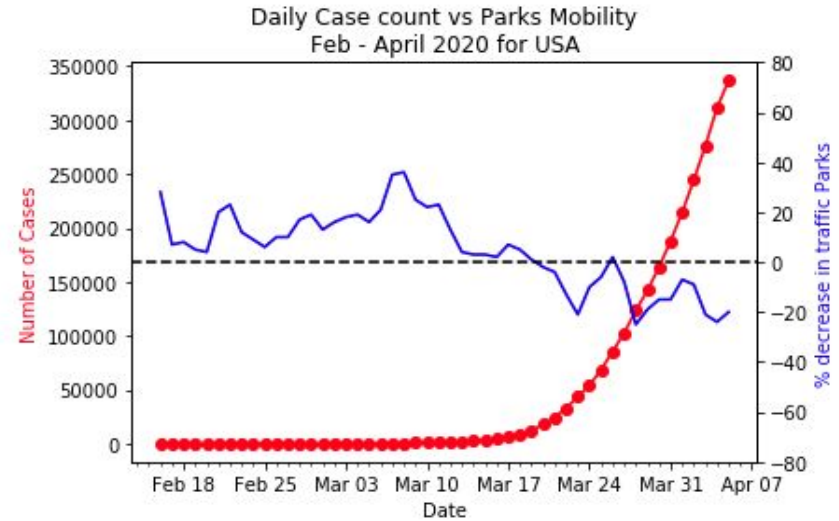
Analyze for Trends



Retail & Recreation

-49%

compared to baseline



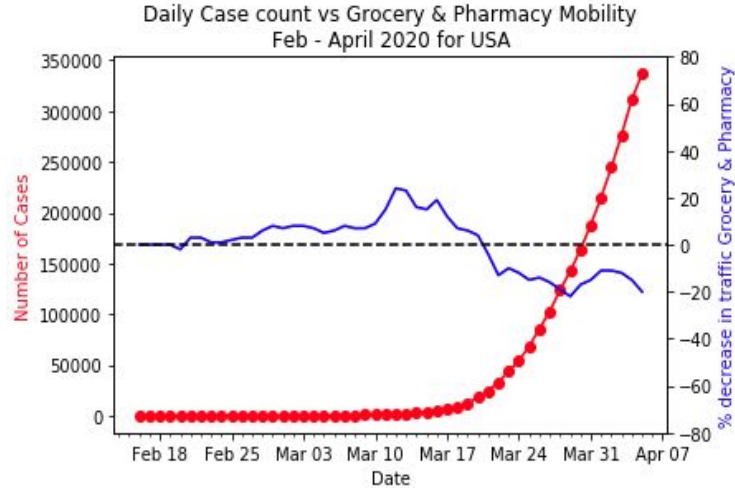
Parks

-20%

compared to baseline

- There is a 2-3 week lag before we can actually see the impact of social distancing.

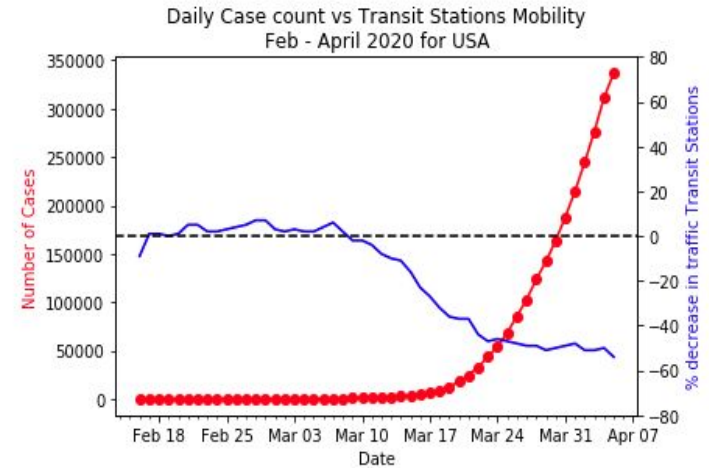
Analyze for Trends



Grocery & Pharmacy

-20%

compared to baseline



Transit Stations

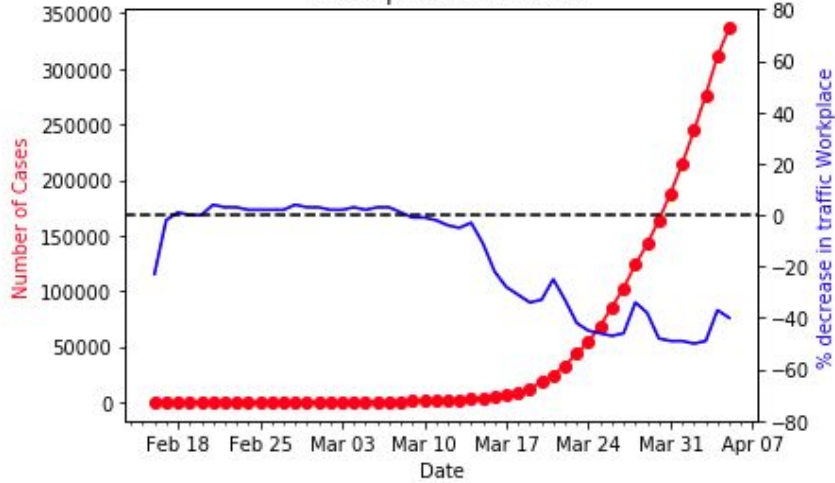
-54%

compared to baseline

- The exponential growth in new cases is due to people that were infected before the social distancing guidelines came into effect.

Analyze for Trends

Daily Case count vs Workplace Mobility
Feb - April 2020 for USA

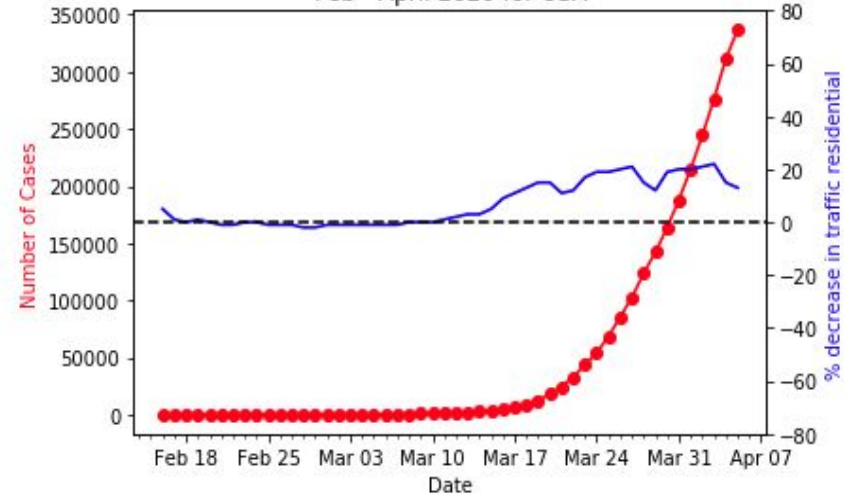


Workplaces

-40%

compared to baseline

Daily Case count vs residential Mobility
Feb - April 2020 for USA



Residential

13%

compared to baseline

“ At least **Several Weeks** needed to measure coronavirus restriction impact.” - Dr Fauci, March 23 rd

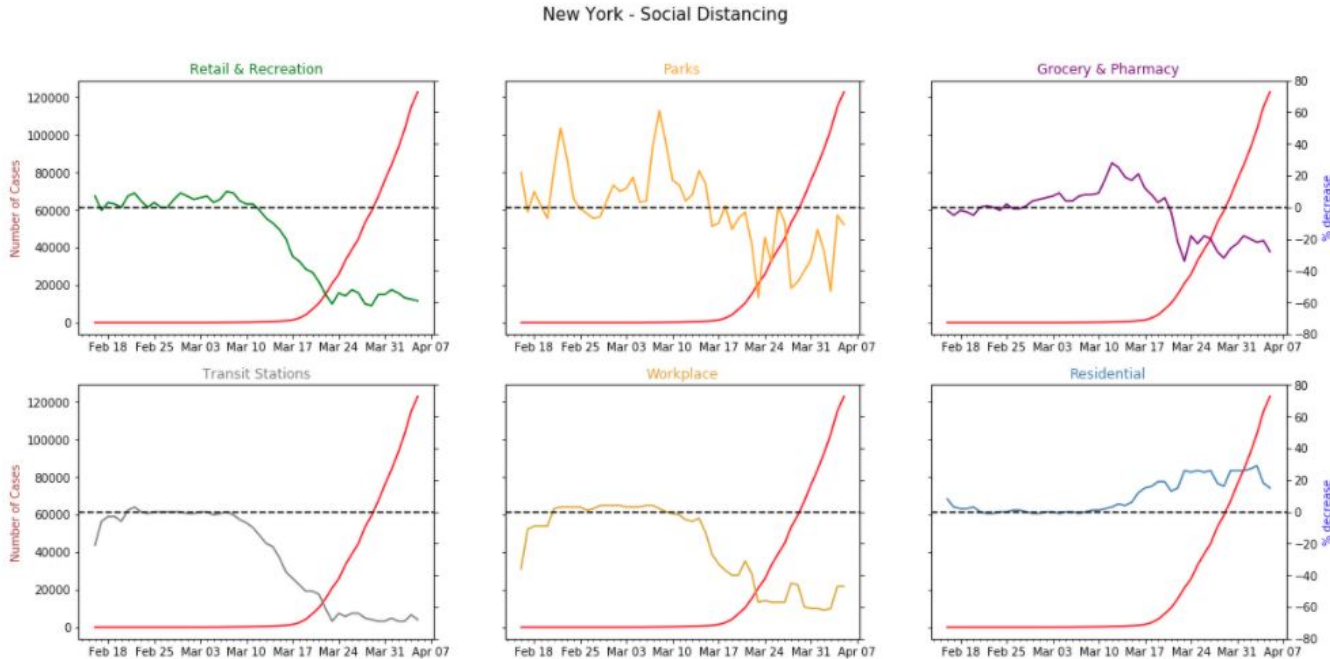
Analyze for Trends

	Cases	Death
Percent decrease Retail	-0.743	-0.655
Percentage decrease Parks	-0.722	-0.652
Percentage decrease Grocery	-0.714	-0.626
Percentage decrease Transit	-0.750	-0.661
Percentage decrease Workplace	-0.724	-0.631
Percentage decrease Residential	0.702	0.606

Pearson Correlation coefficients

Analyze for Trends

Is New York Social Distancing?



- New Yorkers are doing a **significantly better** job at social distancing compared to the national average.

Retail & Recreation

-59%

compared to baseline

Grocery & Pharmacy

-28%

compared to baseline

Transit Stations

-68%

compared to baseline

Workplaces

-47%

compared to baseline

Limitations of Analysis

- We need to expand the time period to completely quantify the impact of social distancing.
- The increase in number of new cases could just be due to increase in the number of tests that being performed.

Conclusion

- From our analysis, we can conclude that social distancing has not yet reduced the number of new covid-19 cases in USA.
- Social distancing has reduced the rate of increase in new cases, thereby helping flatten the curve.

—

Hypothesis 3:

The number of confirmed cases is not directly dependent on the number of tests performed

Hypothesis 3: Data

- The COVID Tracking Project API was used to pull state level testing data
- This data contains daily testing information for each state
- In testing the null hypothesis is utilized on:
 - The full dataset
 - The dataset without clear outliers
 - The dataset narrowed down to the 8 states with the highest positivity rates

<https://covidtracking.com/>

Data Retrieval and Assembly

```
response = requests.get('https://covidtracking.com/api/v1/states/daily.json').json()
results = response
```

```
field_list = ['date', 'state', 'positive', 'negative', 'pending', 'hospitalizedCurrent']
results_dict = dict.fromkeys(field_list)
for field in field_list:
    results_dict.update({field : []})
```

```
for result in results:
    for field in field_list:
        try:
            results_dict[field].append(result[field])
        except:
            results_dict[field].append(0)
```

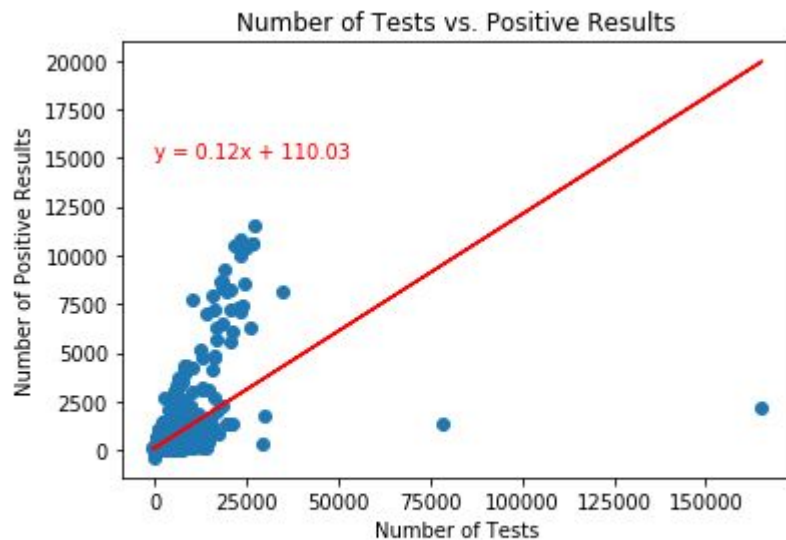
```
covid_states_df = pd.DataFrame.from_dict(results_dict)
covid_states_df.info()
```

Cleaning the Data

	date	state	positive	negative	pending	hospitalizedCurrently	hospitalizedCumulative	inlcuCurrently	inlcuCumulative	onVentilatorCurrently	...	hospitalized	total
0	20200424	AK	339.0	11942.0	NaN	36.0	NaN	NaN	NaN	NaN	...	NaN	12281
1	20200424	AL	5832.0	46863.0	NaN	NaN	768.0	NaN	288.0	NaN	...	768.0	52695
2	20200424	AR	2741.0	32837.0	NaN	101.0	291.0	NaN	NaN	24.0	...	291.0	35578
3	20200424	AS	0.0	3.0	17.0	0.0	0.0	0.0	0.0	0.0	...	0.0	20
4	20200424	AZ	6045.0	54669.0	NaN	639.0	984.0	332.0	NaN	186.0	...	984.0	60714

```
covid_states_df['positiveIncrease'] = covid_states_df['positiveIncrease'].replace("None", 0)
covid_states_df['positiveIncrease'] = covid_states_df['positiveIncrease'].replace("", 0)
covid_states_df = covid_states_df.replace([np.inf, -np.inf], np.nan)
covid_states_df = covid_states_df.dropna(subset=['positiveIncrease'])
covid_states_df.info()
```

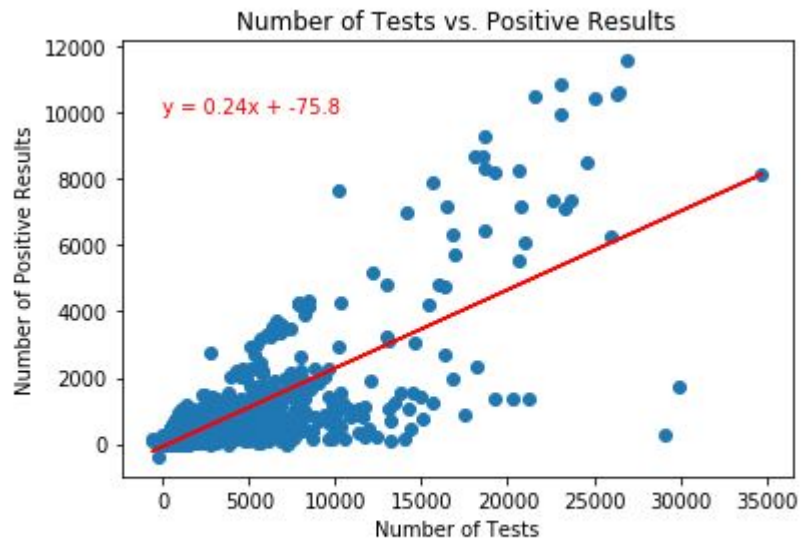

Graphs



The R-squared between both factors is 0.34

The p-value is $2.1095264583594818 \times 10^{-251}$

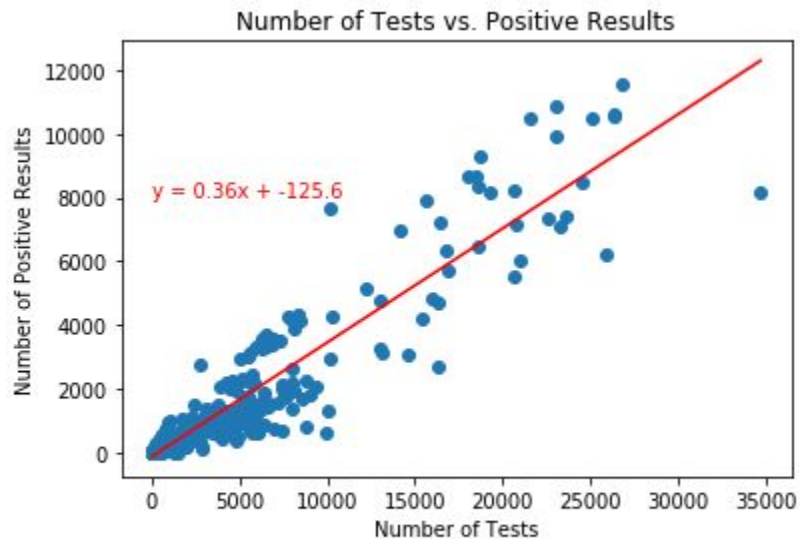
Graphs



The R-squared between both factors is 0.63663

The p-value is 0.0

Graphs



The R-squared between both factors is 0.87

The p-value is 7.498650691558495e-181

Limitations of Analysis

- There are some obvious reasons that the number of confirmed cases will at least be partially tied to the number of tests
- Testing has not continued to grow in the past few weeks
- This analysis focuses solely on the United States and then hones in on some of the worst hit states

Conclusion

- The number of confirmed cases is tied directly to the number of tests that are performed in a given area
- As this currently stands true this means that case growth has likely outgrown test growth
- Until testing outgrows case growth we will not have enough knowledge about the spread of the virus

—

Hypothesis 4:

The number of confirmed cases is not directly influenced by the population density of an area.

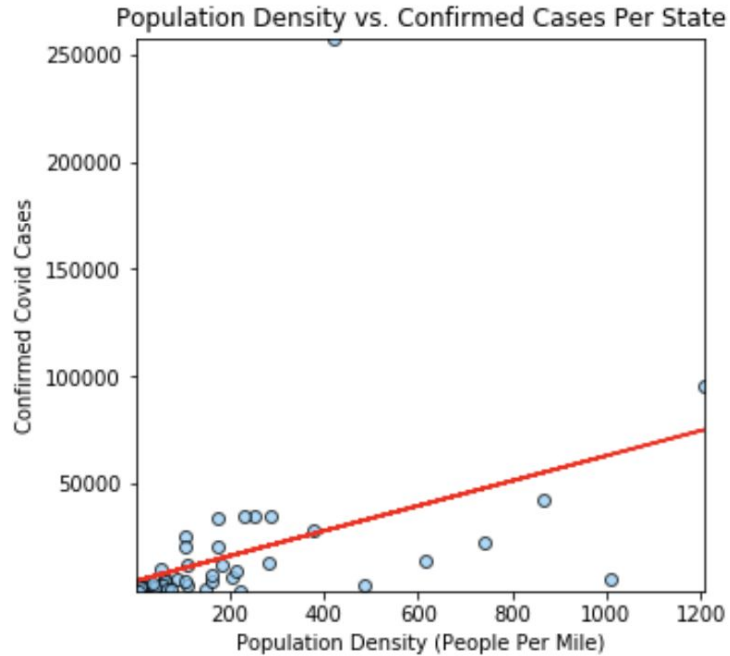
Hypothesis 4: Data

- New York Times Covid API was used to track confirmed cases at both the county and state level.
- The US Census Bureau website was used to acquire the following:
 - Population density for each state.
 - Net population for each state.
 - Population density for each county (number of people per square mile of land).
 - Net population of each county.

Relationships tested

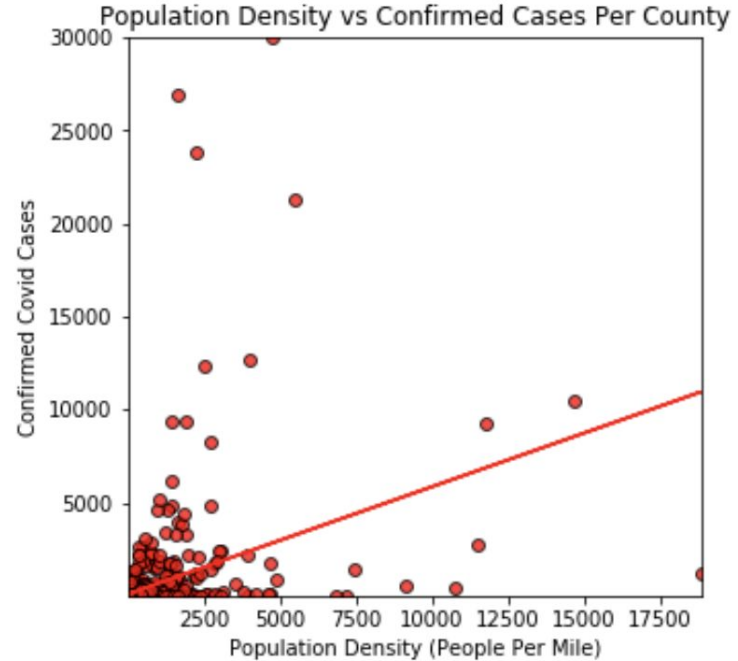
- State level analysis of population density vs. number of confirmed cases
- County level analysis of population density vs. number of confirmed cases.
- County level analysis of population density vs. confirmed cases per capita.
 - Confirmed cases per capita was calculated by dividing the number of cases per county by the population of that county.
 - Counties without confirmed cases were not included in the analysis data set.

State Level Analysis



The correlation coefficient is 0.4
The P Value is 0.004464940094622526

County Level Analysis

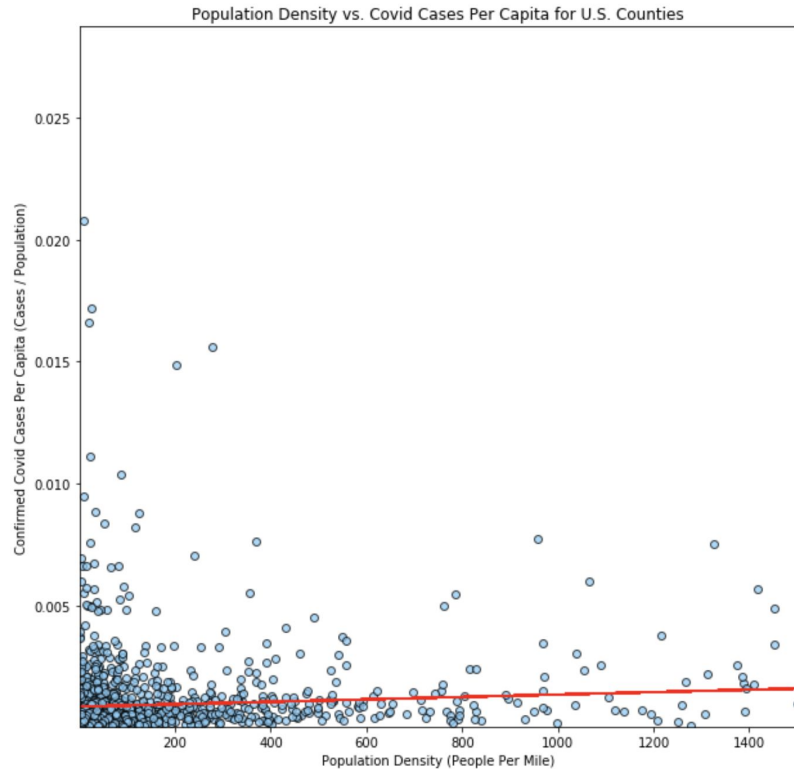


The correlation coefficient is 0.38
The P Value is 4.173629213245211e-59

```
agg_df.sort_values(['casesPerCapita'], ascending = False).head()
```

	State	Area	Units	2019	date	county	state	cases	deaths
36087	New York	Rockland County	Persons per square mile	1877.85401	2020-04-19	Rockland	New York	9364	276
36119	New York	Westchester County	Persons per square mile	2247.17713	2020-04-19	Westchester	New York	23803	831
36059	New York	Nassau County	Persons per square mile	4772.31142	2020-04-19	Nassau	New York	30013	1577
16013	Idaho	Blaine County	Persons per square mile	8.56832	2020-04-19	Blaine	Idaho	470	5
36103	New York	Suffolk County	Persons per square mile	1624.34389	2020-04-19	Suffolk	New York	26888	845

Sorting by cases per capita shows cases per capita is a better indication of the effect of population density than strictly plotting number of cases.



The correlation coefficient is 0.25
The P Value is 2.618398591890597e-25

Conclusions

- The data shows that there is a slight positive correlation between population density and confirmed cases per capita, so population density may be a factor in accelerating the spread of the disease.

Limitations

- Not all areas of the U.S. have equal rates of testing.
- Outliers in the datasets may be skewing results.
- The confounding variable of population alone may be driving up the number of confirmed cases in areas with high population density.



Thank you

