Insights and Analysis on COVID-19

Jerry Sun

11 December 2020

1 Introduction

For the purposes of this project, I have decided to collect and analyze data on Covid-19 to try and find some insightful correlations and trends present within the data. This also includes examining trends that may validate already existing literature on the spread, virality, mortality, and knowledge about the US healthcare system. This data primarily comes from Amazon Web Services and their public access data, the NSDA, and a dataset from another class (CS 4780) with useful features about counties in the US, all of which can be found in my references. Also included in the references is the code that created the figures, plots, and analyses present in this document. The language used is Python through Jupyter Notebook as well as the packages Pandas and matplotlib to do much of the plotting and dataframe manipulation for the datasets.

2 Description of Data

The data obtained from Amazon, AWSDF, contains data on the precise counties and their number of cases and deaths that occured from their first case up until the day of Thanksgiving. The data includes the date, county, state, fips, cases and deaths for each row. The data from the NSDA contained yearly Unemployment Rates data by state and county from the years 2011 up to 2019 as well as Median Household Income data for 2018. Rather than analyzing every single state, I decided to analyze a few select states that represent key demographic areas within the US with respect to industry, size, and unemployment rate categories. I collected this NSDA data from the states Ohio, Virginia, Washington, and Wyoming. Each state by the NSDA's own records falls into a separate category for Unemployment rates and have different majority industry sectors accounting for their state GDP's: Finance/Insurance/Real-Estate, Professional and Business services, Information Technologies, and Mining/Quarrying/Oil Extraction respectively. They are also drastically varying in population sizes and degree of rural to urban divide. None of them are also outliers with respect to the pandemic, resulting in more equal and less biased comparisons between them. Lastly, the data incorporated from another class surrounds the topic of election data from 2016 so the data is not consistent with the years from the other datasets but I consider them valid since they shouldn't

change drastically in between the difference of a few years. The dataset has the columns FIPS, County/State, DEM, GOP, MedianIncome, MigraRate, BirthRate, DeathRate, and UnemploymentRate.

3 Analysis

We start with doing a bit of broad analysis over the entire US where the data encompasses all 50 states and 5 territories as well. When we conduct some summary analysis of the rate of cases per day within the last 5 days, we find that the mean rate is 2331 cases/day with a standard deviation of 2439 cases/day. The quartiles follow as such: min - 0; 25% - 746; 50% - 1630; 75% - 3101; max - 11497 cases/day. From the violin plot in Figure 1 we can determine that this distribution is right-skewed with a mean greater than the median and that there appears to be 3 outliers with significant rates of infection in the boxplot.

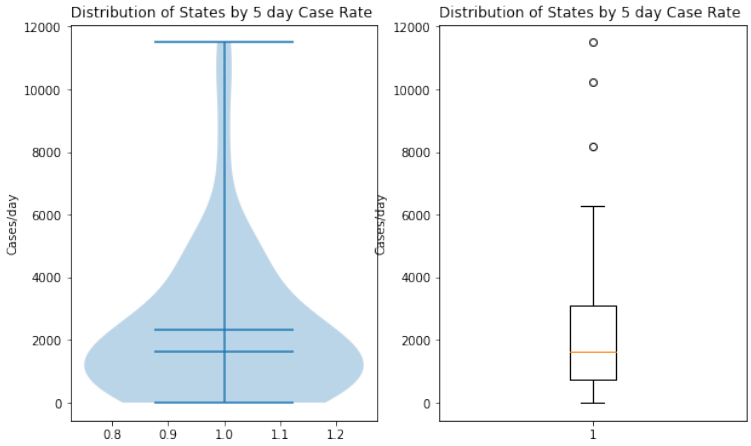

Figure 1: Distribution

We can also estimate what kind of distribution the states form. The most obvious choice would appear to be an exponential distribution and so we will appropriately determine the theoretical values of lambda that would approximate it best. From class, we stated the variance of a random variable could be equal to 1 over the lambda squared and thus solving for lambda we obtain a value of: 0.000409. Thus we make our QQ plot and obtain Figure 2. The dots form a fairly accurate and straight line with the points actually hovering back and forth between above and below the $y = x$ line. This shows that an exponential distribution indeed describes the states' and their infection rates for the past 5 days. From this we can see that the majority of the states have similar degrees of success for their containment policies with a few spiking states that are doing exceptionally poor as we head into winter.

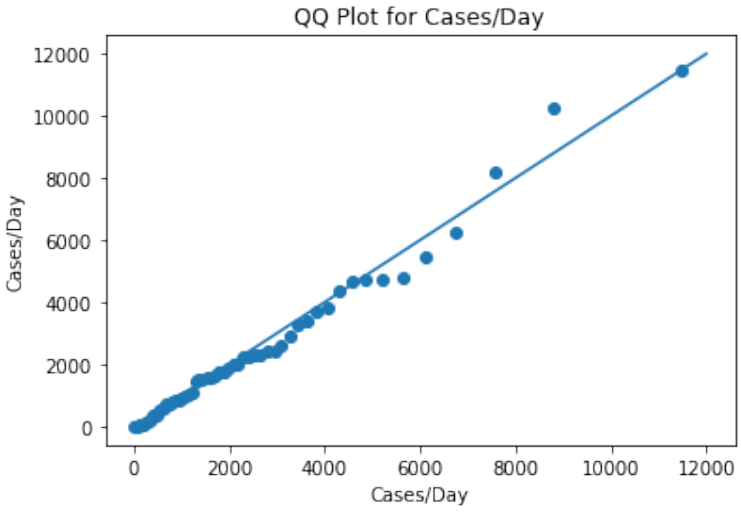

Figure 2: QQ Plot

Here in Figure 3, we will examine the relationship between Median Household Income and Cases.

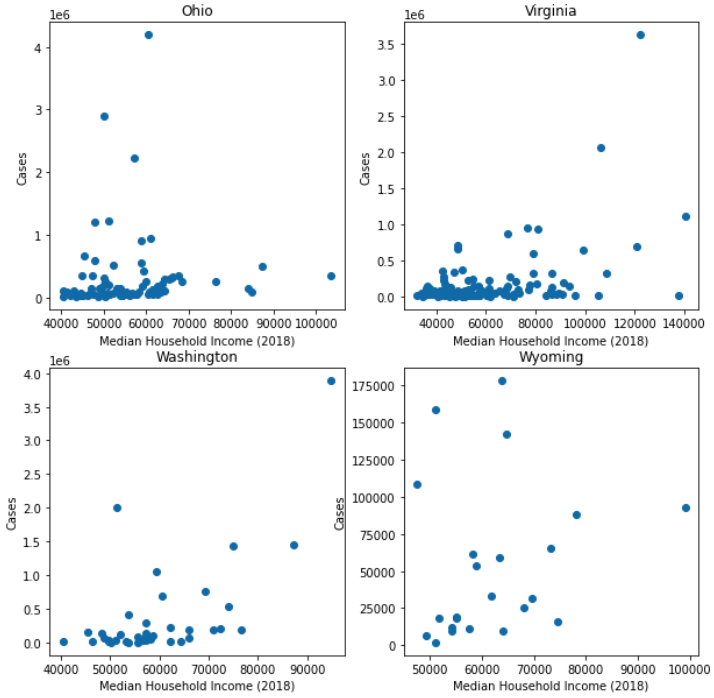

Figure 3: Income Case Plot

We see here that there exists a relationship between poorer and richer counties that has been relatively well-documented in existing literature. In three of the four states (Ohio, Washington, Wyoming), the infection rates amongst the poorer communities was significantly larger than their richer counterparts (although Wyoming had the lowest statewide infection rates out of the four). This can be explained due to lack of testing resources, access to healthcare, and other resources poor communities lack

in order to combat teh virus (Finch Finch, 2020). However we also see a few well-off communities with high levels of infection as well which can be explained by a recent Cornell study. Zitek and Schlund in their study found that "entitled individuals were less likely to report that they were following, or would follow, the health guidelines of the COVID-19 pandemic, and they were more likely to report that they had contracted COVID-19" (Zitek Schlund, 2020). This seems consistent with Figure 3 where a sizeable portion of higher income communities have Covid infection rates higher than their poorer counterparts in all states but Ohio. The narrative this data presents is further solidified and compounded when you take into account the mortality rates as shown in Figure 4.

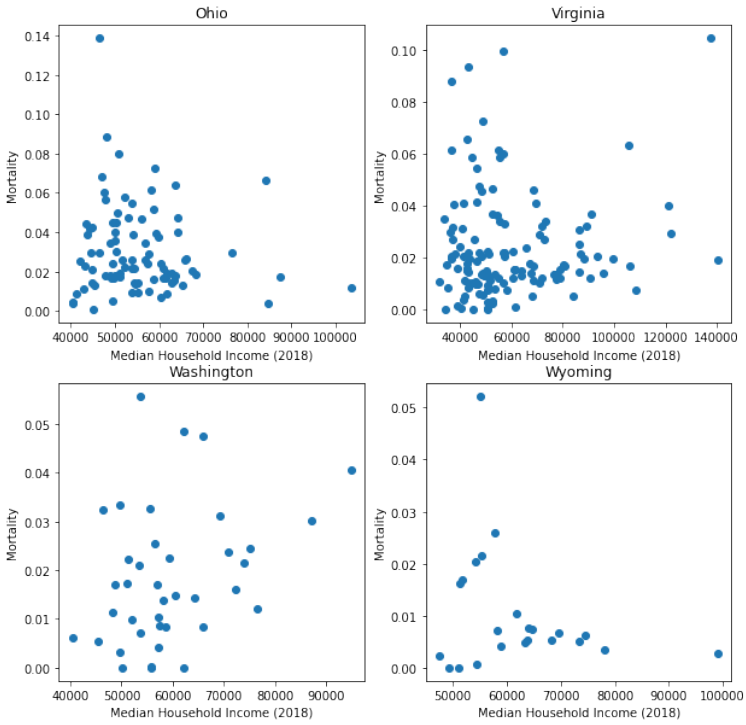

Figure 4: Income Mort Plot

From Figure 4, we can see that in all four states, the mortality rate is high for lower income communities regardless of their infection rates in Figure 3, showing that these communities simply lack the resources wealthier ones have, not to mention the pricing of such options. An interesting note is the high mortality in some of the wealthier communities, presumably because of the virus' natural deadliness and novel appearance, but also due to those counties having high infection rates as well if you compare with Figure 3.

Moving on to the cumulative case and death plots for each of our states, we can see that the line plots actually reflect a variety of events that we've observed so far.

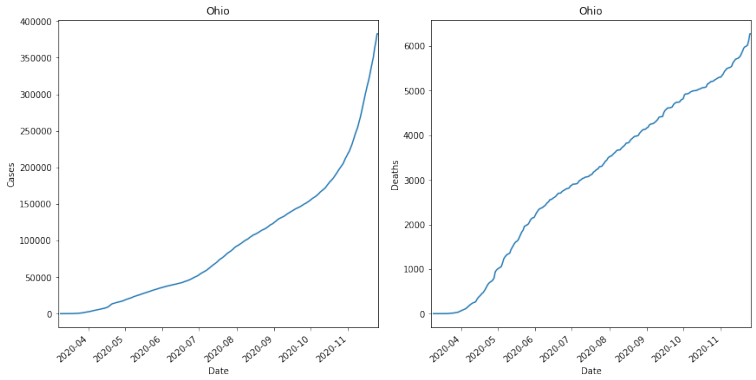

Figure 5: Ohio Case Plot

We can see how the death rate was extremely steep in the beginning as hospitals were overwhelmed and little was known about the disease and how it spread and how deadly it was. Even now as Ohio's cases have spiked recently, the deaths have not spiked as we saw in the beginning, meaning that the rate of survival has increased with time as one would expect.

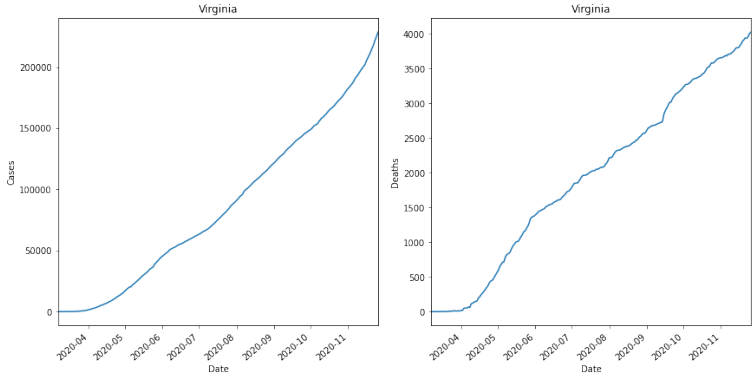

Figure 6: Virginia Case Plot

What we can see in Virginia is actually pretty interesting as it pertains to recent events. Most news coverage will lead you to believe that winter is causing some major spikes in Covid cases but in Virginia it appears as if the rate of infection has remained pretty constant and stable throughout, even dipping during the summer when many states were spiking. This could also be due to the fact that Virginia has wealthier counties overall which on average appear to have lower infection rates and death rate than their counterparts as mentioned before.

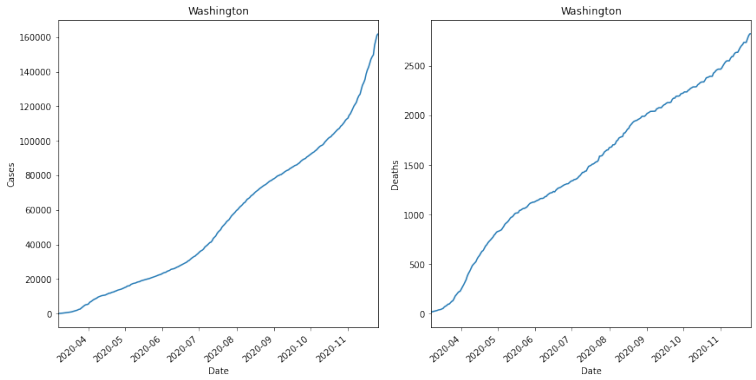

Figure 7: Washington Case Plot

From Washington, we see many of the trends present in Ohio and Virginia but it seems like their cases are fairly lower than either with major spikes in cases being the past month and towards the end of summer. With respect to protests during the summer, the data shows mixed results as the beginning of protests started end of May and continued throughout the summer while the spike in Covid cases only began in the middle and end of July. It's unclear what major event if any contributed to this major spike.

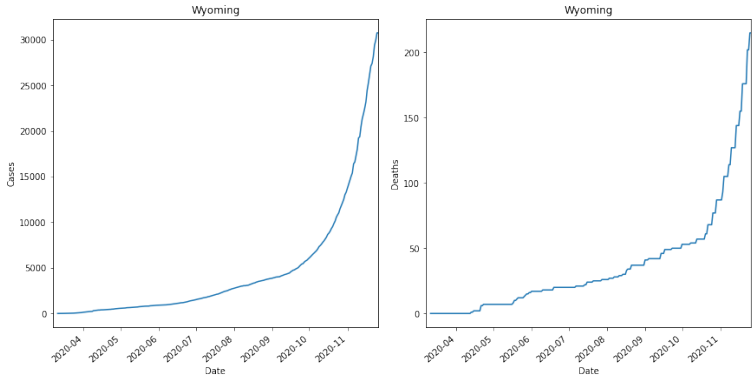

Figure 8: Wyoming Case Plot

What we see from Wyoming is an astonishing explosion of cases recently. In fact the very sparse and lightly populated state had only 5,000 cumulative cases until the past month skyrocketed them towards 30,000 which is a 600% increase. Seeing as they never had to deal with the severity of the Covid outbreak as other states had to, their mortality rate has also skyrocketed along with the case growth. Furthermore public policy has also shaped the spread of the virus with the lack of strong imposed health restrictions on the general public (Gruver 2020).

Something to realize for all of these states is that with the exception of Wyoming, over a third of all of their deaths occurred within the first month of their first case of coronavirus. In the case of Wyoming, three quarters of their deaths have resulted in the past month. But furthermore, the current rate of infection for the past few weeks has been the highest that every state has seen, making the concerns and

panic that was induced earlier in the year seem a bit out of touch as society has progressed to become more weary of this new reality. Less coverage and a much too early relaxation of state restrictions have coincided with winter, a season known to already be harsh on living conditions, to induce Covid spikes in every state, including many other states in the US outside of this sample size of four. This goes to show that if restrictions are lifted or when there are no restrictions, deaths and cases exponentially increase. These regulations and safety protocols are the only reason why the pandemic hasn't decimated the US by now. We can easily project that if the initial rate of mortality was consistent with the initial months, we could see double the number of deaths that we see currently. It is crucial that we as a society stay vigilant as we approach the heart of winter.

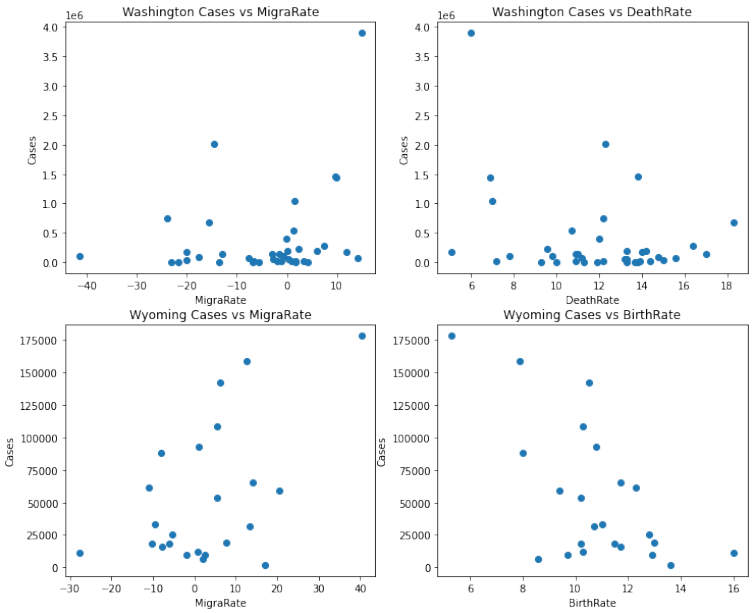

Figure 9: Correlation Plots

The examined variables for correlations with number of cases in each county and state included the Migration Rate, Birth Rate, Death Rate, and Unemployment Rate. Plotted above in separate scatterplots are the four highest correlated features out of all of the states. We can see the top two plots are weakly correlated with correlation coefficient values of 0.26 for Migration Rate and -0.32 for Death Rate. From what we can see it appears counties in Washington with higher rates of migration into the state have higher Covid cases, albeit only slightly. A slightly stronger correlation exists in Washington's Death Rate where surprisingly, the lower the previous non-Covid year Death Rate the higher the number of cases. This is only a slight correlation however and the few wealthy counties with higher number of cases maybe the reason for cause as wealthier counties also have lower Death Rates simply due to better medical access. With respect to the bottom two plots, they have much higher correlation coefficient values; coming in at 0.51 is the Migration Rate whereas the Birth Rate is -0.62. These correlations also appear much more strongly correlated, although they in actuality aren't strong enough to really suspect any major causations or potential

strong trends/rules of thumb. Similarly with before, we can gander a guess at the rough trends behind these correlations where counties with higher level of influx of residents generally result in more populous and urban areas and potentially more traveler's, meaning more people coming in and out of the state to transmit and spread Covid. Furthermore, Birth Rate is known to be lower among higher income households, reflecting the same principles as wealthy vs poorer nations. However when comparing these ties to income in Figure 3, we realize that indeed these guesses are incorrect since the top 3 counties in cases are amongst the lower income counties. Thus, the most we can conclude is that these trends are simply rough correlations that are loosely rooted in some connection to these features.

4 Conclusion and Future Work

The bulk of this paper's focus centers around correlations and trends rather than any concrete causal inference which is entirely the point as there exists circulating literature that has either already documented systemic issues in the US's health care system or analyzed the reasons behind the viral spread of Covid in the US. The aim of this paper was simply to display these trends and dive deeper into some comparisons between different states in the US with varying populations, demographics, prevalent industries, and so forth. The main takeaways of the analysis is that population wealth, public health policies, and knowledge about the disease are the main contributing factors to the spread, mortality, and response to the pandemic. Future extensions of this analysis would include further investigations into the causal nature of certain factors as well as the direct comparison between case numbers and enactment of public policy or occurrences of major events to determine the degree of influence of those factors.

5 Resources

Economic Research Service. "County-Level Data Sets." USDA ERS - County-Level Data Sets, United States Department of Agriculture, 13 May 2020, www.ers.usda.gov/data-products/county-level-data-sets/.

"Open-Access Data and Computational Resources to Address COVID-19." National Institutes of Health, U.S. Department of Health and Human Services, 2020, datascience.nih.gov/covid-19-open-access-resources.

Finch, W. Holmes, and Maria E. Hernández Finch. Poverty and Covid-19: Rates of Incidence and Deaths in the United States During the First 10 Weeks of the Pandemic, Frontiers in Sociology, 29 May 2020, www.frontiersin.org/articles/10.3389/fsoc.2020.00047/full.

Zitek, Emily M., and Rachel J. Schlund. Psychological Entitlement Predicts Non compliance with the Health Guidelines of the COVID-19 Pandemic. U.S. National Library of Medicine, 29 Oct. 2020, www.ncbi.nlm.nih.gov/pmc/articles/PMC7598540/.

Gruver, Mead. Wyoming among Top Five States for COVID-19 Infection Rate, Associated Press, 27 Oct. 2020, apnews.com/article/virus-outbreak-wyoming-cheyenne-4994b1e60aa0ac40297a064c0584d45e.

6 Code

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib.dates as mdates
from scipy.stats import norm
import datetime
df = pd.read_csv('us-counties.csv')
df2 = pd.read_csv('train_2016.csv')
df3 = pd.read_csv('test_2016_no_label.csv')
df5 = pd.concat([df2.drop(['DEM', 'GOP'], axis=1), df3])
WY = pd.read_csv('WyomingUnemployment.csv')
WA = pd.read_csv('WashUnemployment.csv')
VA = pd.read_csv('VirgUnemployment.csv')
OH = pd.read_csv('OhioUnemployment.csv')

def preprocess(state, us):
    us = us[us['state'] == state['name'][0]]
    state['fips1'] = state['fips1'].astype(float)
    fips = state['fips1']
    for i, code in enumerate(fips):
        fips[i] = sum(us[us['fips'] == code]['cases'])
    state['Cases'] = np.array(fips)
    state = state.drop(['year', 'year.1', 'year.2', 'year.3', 'year.4', 'year.5', 'year.6',
                        'year.8', 'UR', 'UR.1', 'UR.2', 'UR.3', 'UR.4', 'UR.5', 'UR.6', 'UR.8'],axis=1)
    money = state['Median Household Income 2018']
    incomes = []
    for income in money:
        incomes.append(float(income.split(',')[0][1:] + income.split(',')[1]))
    state['Median Household Income 2018'] = incomes
    return state

joined = [preprocess(WY.copy(), df.copy()), preprocess(WA.copy(), df.copy())
          preprocess(VA.copy(), df.copy()), preprocess(OH.copy(), df.copy())]
fig, axs = plt.subplots(2,2, figsize=(10,10))
order = ["Ohio", "Virginia", "Washington", "Wyoming"]
for i, ax in enumerate(axs.flat):
    ax.scatter(joined[i]['Median Household Income 2018'][1:], joined[i]['Cases'][1:])
    ax.set(xlabel="Median Household Income (2018)", ylabel="Cases")
    ax.set_titel(order[i])

def preprocess2(state, us):
    us = us[us['state'] == state['name'][0]]
    state['fips1'] = state['fips1'].astype(float)
    fips = state['fips1']
    for i, code in enumerate(fips):
        fips[i] = sum(us[us['fips'] == code]['deaths'])/
sum(us[us['fips'] == code]['cases']) if sum(us[us['fips'] == code]['cases']) != 0 else 0
    state['Mortality'] = np.array(fips)
    state = state.drop(['year', 'year.1', 'year.2', 'year.3', 'year.4', 'year.5', 'year.6',
                        'year.8', 'UR', 'UR.1', 'UR.2', 'UR.3', 'UR.4', 'UR.5', 'UR.6', 'UR.8'],axis=1)
    money = state['Median Household Income 2018']
    incomes = []
    for income in money:
        incomes.append(float(income.split(',')[0][1:] + income.split(',')[1]))
    state['Median Household Income 2018'] = incomes
```

Figure 10: Code

```
    return state

joined2 = [preprocess2(WY.copy(), df.copy()), preprocess2(WA.copy(), df.copy())
           preprocess2(VA.copy(), df.copy()), preprocess2(OH.copy(), df.copy())]
fig, axs = plt.subplots(2,2, figsize=(10,10))
for i, ax in enumerate(axs.flat):
    ax.scatter(joined2[i]['Median Household Income 2018'][1:], joined2[i]['Mortality'][1:])
    ax.set(xlabel="Median Household Income (2018)", ylabel="Mortality")
    ax.set_title(order[i])

def trends(us, state):
    us = us[us['state'] == state].drop(['county', 'fips'], axis=1)
    dates = us['date'].unique()
    cases, deaths = [], []
    for date in dates:
        temp = us[us['date'] == date]
        cases.append(sum(temp['cases']))
        deaths.append(sum(temp['deaths']))
    us = pd.DataFrame()
    base = datetime.datetime(2020, int(dates[0][-5:-3]), int(dates[0][-2:]))
    dates = np.array([base + datetime.timedelta(hours=(24 * i)) for i in range(len(dates))])
    us['dates'], us['cases'], us['deaths'] = dates, np.array(cases), np.array(deaths)
    return us

trendsWY = trends(df.copy(), 'Wyoming')
fig, axs = plt.subplots(1, 2, constrained_layout=True, figsize=(12, 6))
order = ['cases', 'deaths']
for i, ax in enumerate(axs.flat):
    ax.plot(trendsWY['dates'], trendsWY[order[i]])
    ax.set(xlabel="Date", ylabel=order[i])
    ax.set_title("Wyoming")
    ax.set_xlim((np.datetime64('2020-03-07'), np.datetime64('2020-11-26')))
    for label in ax.get_xticklabels():
            label.set_rotation(40)
            label.set_horizontalalignment('right')

def summary(us):
    states, points = us['state'].unique(), []
    for state in states:
        trend = trends(us.copy(), state)
        rate = trend['cases'][len(trend)-1] - trend['cases'][len(trend)-5]
        points.append(np.sum(rate/5))
    return sorted(points)

data = summary(df.copy())
fig, axs = plt.subplots(1, 2, figsize=(10, 6))
axs[0].violinplot(data, showmedians=True, showmeans = True)
axs[1].boxplot(data)
for i, ax in enumerate(axs.flat):
    ax.set(ylabel="Cases/Day")
    axs.set_title("Distribution of States by 5 day Case Rate ")
qq, res, i = [], [], 1
for point in data:
    qi = (i - 0.5)/55
    expo = np.log(1 - qi)/-0.000409
```

Figure 11: Code

```
    qq.append(expo)
    i += 1
plt.scatter(qq, data)
plt.plot([0,12000],[0,12000])
plt.xlabel("Cases/Day")
plt.ylabel("Cases/Day")
plt.title("QQ Plot for Cases/Day")
pd.DataFrame(data).describe()

def ml_preprocess(dframe, us, code, state):
    name, county= [], []
    for i in range(1555):
        county.append(dframe['County'][i].iloc[0][0:-11])
        county.append(dframe['County'][i].iloc[1][0:-11])
        name.append(dframe['County'][i].iloc[0][-2:])
        name.append(dframe['County'][i].iloc[1][-2:])
    dframe['code'], dframe['County'] = np.array(name), np.array(county)
    dframe, cases, us = dframe[dframe['code'] == code], [], us[us['state'] == state]
    for county in range(len(dframe.index)):
        fips = dframe['County'].iloc[county]
        cases.append(sum(us[us['county'] == fips]['cases']))
    dframe['cases'] = np.array(cases)
    return dframe

oh = ml_preprocess(df5.copy(), df.copy(), 'OH', 'Ohio')
va = ml_preprocess(df5.copy(), df.copy(), 'VA', 'Virginia')
wa = ml_preprocess(df5.copy(), df.copy(), 'WA', 'Washington')
wy = ml_preprocess(df5.copy(), df.copy(), 'WY', 'Wyoming')
fig, axs = plt.subplots(2,2, figsize = (12,10))
lst = ["MigraRate", "DeathRate", "MigraRate", "BirthRate"]
lst2 = ["Washington Cases vs ", "Washington Cases vs ", "Wyoming Cases vs ", "Wyoming Cases vs "]
lst3 = [[wa, "MigraRate"], [wa, "DeathRate"], [wy, "MigraRate"], [wy, "BirthRate"],]
for i, ax in enumerate(axs.flat):
    ax.scatter(lst3[i][0][lst3[i][1]], lst3[i][0]['cases'])
    ax.set(xlabel=lst[i], ylabel="Cases")
    ax.set_title(lst2[i] + lst[i])
wa.corr()
wy.corr()
```

Figure 12: Code