Tarsus Lam, Jedidiah Tsang, Eugenia Tzeng
Data 100
May 13th, 2020
Final Project

The COVID-19 Crisis and Efforts to Combat It

Abstract

Undoubtedly one of the largest global events in recent history, the coronavirus pandemic has led to some of the strongest responses we've seen yet: the implementation of months-long shelter in place policies, stimulus checks sent directly to individuals, and other policies implemented in an effort to combat the varied impacts of the virus. In our research project, we explore crucial questions surrounding COVID-19, and come to the following conclusions. First, we used linear regression plots and bar plots to show that counties who lean Democrat typically have higher case counts of coronavirus, leading to an earlier implementation of policies such as shelter in place. Second, we used violin plots to show that the United States is struggling to contain community spread when compared to China. Last, we used a linear regression model and lasso regularization to find features such as number of hospitals and population density that would supplement existing knowledge to help predict the number of confirmed cases we'll see in the future. Although we have drawn significant conclusions in this project, we also acknowledge limitations given our current timeline, and offer some insight on how to best expand on our current work.

Introduction

As three college students of Asian-American descent, these identities led us to select the COVID-19 dataset as our topic of interest for our Data 100 final project. With so much uncertainty ahead of us amidst a global pandemic, we sought to answer questions not only about the virus itself and its spread, but also from a political and international angle. This interest, combined with our exploratory data analysis, helped us to form our guiding four questions surrounding COVID-19 case prediction, correlation between political leaning, confirmed cases, and ultimately policy such as shelter in place, and comparisons between the United States and China in effectiveness of response to the novel COVID-19.

Question Framing

Based on our EDA, our group chose to explore the following questions:

Question 1: Does the political leaning of a particular county have correlation with its cumulative number of cases, and does the cumulative number of cases have an impact on when emergency protocol such as shelter in place begins?
Question 2: How does the United States compare against China when it comes to preventing community spread?
Question 3: How effective are shelter in place policies when comparing counties that started earlier than others?
Question 4: What features would work best to predict the cumulative number of confirmed cases in a particular county within the United States?

The provided datasets consisted of four different .csv files: time_series_covid19_confirmed_US, time_series_covid19_deaths_US, 4.18states, and abridged_couties.

In order to determine what questions we would ask, we began with some EDA. We started by reading through a description of the columns provided on GitHub (Johns Hopkins), and began by taking a look at the first five rows of every dataset (Figure 1). From this, we summarized each dataset as follows:

1. time_series_covid_19_confirmed_US (confirmed_cleaned)

This dataset contains information about a particular county in the United States, which can be identified using the UID or FIPS (Federal Information Processing Number) and notes the cumulative number of confirmed cases from 1/22/20 to 4/18/20. The granularity of the dataframe goes down to each particular county, but also includes a 'Province_State' column and 'Country_Region' that can be used to increase the granularity if we choose to group by those columns.

2. time_series_covid19_deaths_US (death_cleaned)

Similar to covid_confirmed, covid_death contains the cumulative number of deaths for counties within the United States from 1/22/20 to 4/18/20. Unlike covid_confirmed, covid_death contains a column that contains the population ('Population') of that particular county.

3. Abridged_couties (county_cleaned)

This dataset also includes the Federal Information Processing Number for all counties in the US, labeled 'countyFIPS', and various information about that particular county, such as its population ('PopulationEstimate2018'), population density ('PopulationDensityperSqMile2010'), number of Hospitals ('#Hospitals), breakdown of sex and age, date in which certain policies such as shelter in place ('stay at home') orders were announced by Gregorian date format, and Democrat to Republican ratio (dem_to_rep_ratio).

4. 4.18states (countries_cleaned)

The main distinction between this dataset and the others is that the granularity of this dataframe goes only to the State level, but also includes data for countries beyond the U.S. such as Canada and China. This dataframe includes features found in other datasets, such as confirmed cases ('Confirmed'), number of deaths ('Deaths'), and other features that control for population such as Incident Rate ('Incident_Rate') and Testing Rate ('Testing_Rate'). We used

https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data to gather information on all the columns and their descriptions.

```
[45] covid_confirmed.head()
```

	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/
0	16	AS	ASM	16	60.0	NaN	American Samoa	US	-14.2710	-170.1320	American Samoa, US	0	0	0	0	0	
1	316	GU	GUM	316	66.0	NaN	Guam	US	13.4443	144.7937	Guam, US	0	0	0	0	0	
2	580	MP	MNP	580	69.0	NaN	Northern Mariana Islands	US	15.0979	145.6739	Northern Mariana Islands, US	0	0	0	0	0	
3	630	PR	PRI	630	72.0	NaN	Puerto Rico	US	18.2208	-66.5901	Puerto Rico, US	0	0	0	0	0	
4	850	VI	VIR	850	78.0	NaN	Virgin Islands	US	18.3358	-64.8963	Virgin Islands, US	0	0	0	0	0	

5 rows × 99 columns

This dataset contains information about a particular county in the United States, which can be identified using the UID or FIPS (Federal Information Processing Number) and notes the cumulative number of confirmed cases from 1/22/20 to 4/18/20. The granularity of the dataframe goes down to each particular county, but also includes a 'Province_State' column and 'Country_Region' that can be used to increase the granularity if we choose to group by those columns.

```
[46] county_info.head()
```

	countyFIPS	STATEFP	COUNTYFP	CountyName	StateName	State	lat	lon	POP_LATITUDE	POP_LONGITUDE	CensusRegionName	CensusDivisionName	U
0	01001	1.0	1.0	Autauga	AL	Alabama	32.540091	-86.645649	32.500389	-86.494165	South	East South Central	
1	01003	1.0	3.0	Baldwin	AL	Alabama	30.738314	-87.726272	30.548923	-87.762381	South	East South Central	
2	01005	1.0	5.0	Barbour	AL	Alabama	31.874030	-85.397327	31.844036	-85.310038	South	East South Central	
3	01007	1.0	7.0	Bibb	AL	Alabama	32.999024	-87.125260	33.030921	-87.127659	South	East South Central	
4	01009	1.0	9.0	Blount	AL	Alabama	33.990440	-86.562711	33.955243	-86.591491	South	East South Central	

5 rows × 87 columns

This dataset also includes the Federal Information Processing Number for all counties in the US, labeled 'countyFIPS', and various information about that particular county, such as its population ('PopulationEstimate2018'), population density ('PopulationDensityperSqMile2010'), number of Hospitals ('#Hospitals), breakdown of sex and age, date in which certain policies such as shelter in place ('stay at home') orders were announced by Gregorian date format, and Democrat to Republican ratio (dem_to_rep_ratio).

(Figure 1: First five rows of the provided datasets). Our start to EDA began with looking at the first five entries of each provided dataset and making remarks on what sort of information it could give us.

Data Cleaning

When cleaning our data to prep for visualizations, we wanted to keep two main goals in mind:

1. Limit the amount of data we filter out when merging different datasets
2. Ensure that we're filling in or removing data in a justifiable manner

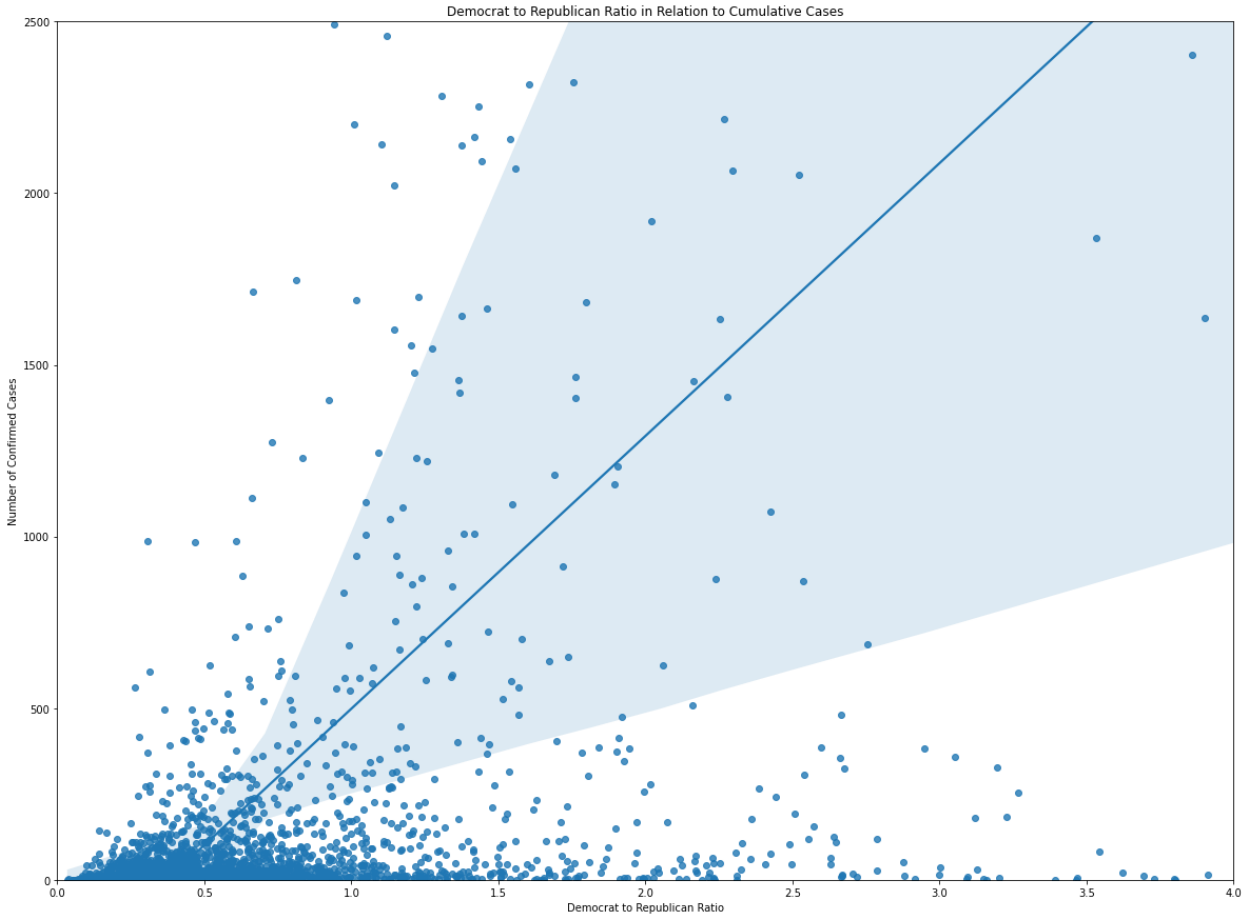
These two goals drove the majority of our EDA. We'll go into each question and note how we approached EDA for each one, starting with an overall clean. **Note: This chart is not comprehensive, but demonstrates how deeply we've thought about every modification.**

Question	Modification	Justification
Preliminary Clean	Removed two entries with FIPS 'City1' and 'City2'	Those entries had no data in them.
Preliminary Clean	Removed entries without an FIPS.	Only two entries had no FIPS number and would allow us to use FIPS while merging.
Preliminary Clean	Renamed column titled 'CountyFIPS' to 'FIPS'	Allows for grouping on 'FIPS'
Preliminary Clean	Converted values in FIPS column to integers	'FIPS' must be of the same type when merging on this primary key
Question 2	Removed all countries except US and China	For this question, we only care about data from the US and China.
Question 1	Filled in missing values for Alaska in dem_to_rep ratio with 0.71274948, then removed counties without the ratio.	Rather than removing all counties without information on dem_to_rep ratio, we found that almost all Alaskan counties were missing this data. Before filtering out counties missing this information, we first Googled the ratio of Democrats to Republicans in that state, filled in that information, then removed NA values.
Question 1, 3	Filled in missing values for stay at home with the Gregorian Ordinal 737533.0	Rather than filter out all the counties that had not instituted shelter in place orders by 4/18/20, we replaced them with the Gregorian Ordinal 737533 (corresponding to 4/18/20) to show that as of that date they had not yet instituted the policy.
Question 4	For selected features, before performing cross validation, removed counties with NAN values.	Of course, our biggest concern here was the removal of counties that don't have information about the features we're looking at. This actually helped guide our modification above with shelter in place, where we initially just removed those values. This modification removed a total of 131 counties from our dataset from 3242 to 3113 (4% of our total data).
Question 1, 3	Converted all Gregorian ordinals with Timestamps	Using Timestamp.fromordinal(), we converted all the Gregorian Ordinals to Timestamps for readability.
Question 1	Renamed dem_to_rep_ratio to Political Leaning, 4/18/20 to Cumulative Cases up to 4/18/20, PopulationDensityperSqMile2010 to Population Density	We did this to increase readability of the column names.

Data Visualization

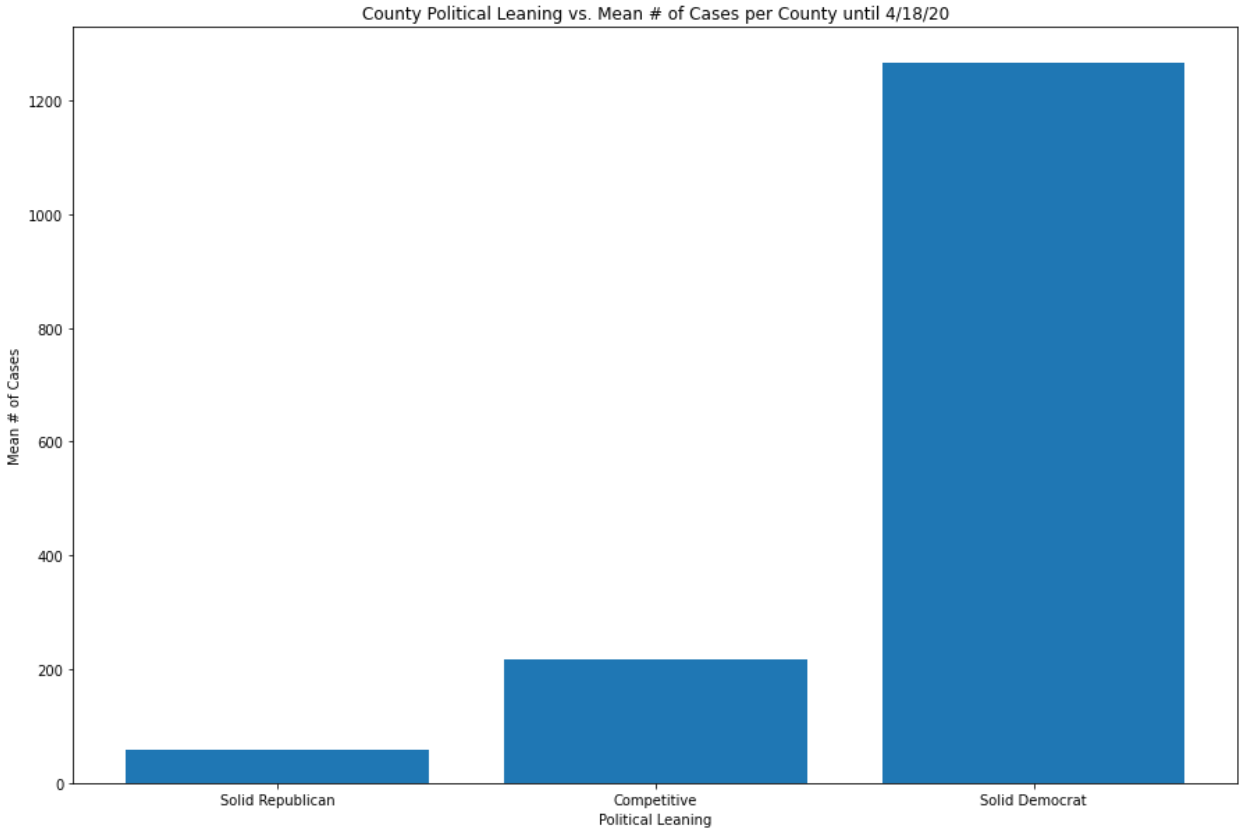
Question 1: Does the political leaning of a particular county have correlation with its cumulative number of cases, and does the cumulative number of cases have an impact on when emergency protocol such as shelter in place begins?

We first started by using sns.regplot to plot a linear regression model fit between the political leaning of a particular county using dem_to_rep_ratio from counties_cleaned and the cumulative number of confirmed cases using 4/18/20 from confirmed_cleaned (Figure 2). We chose a scatter plot as it could best visually represent each county (and also show that more counties lean Republican), but the addition of the linear regression line helps to demonstrate a clear correlation between political leaning and cumulative confirmed cases.



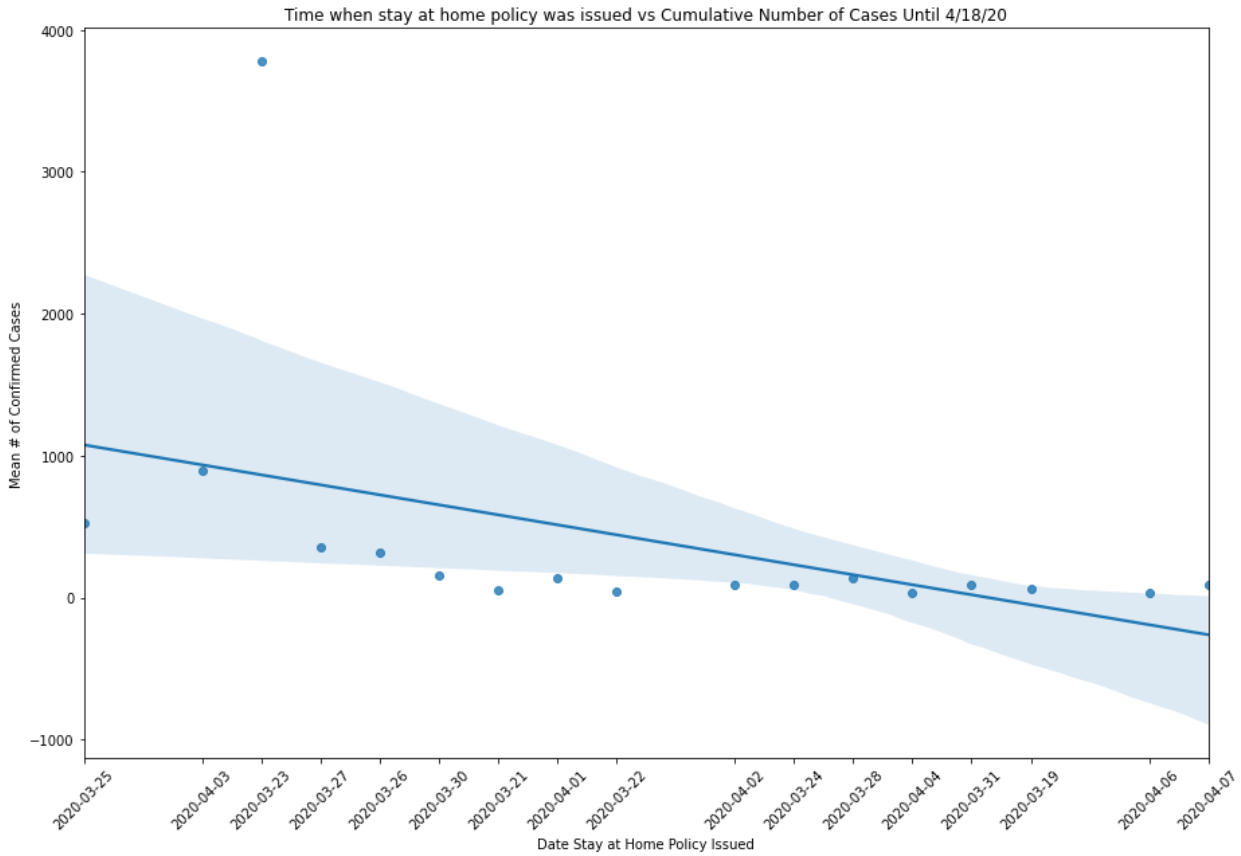
(Figure 2: Democrat to Republican Ratio in Relation to Cumulative Cases) The linear regression line shows that there are higher numbers of confirmed cases in counties that lean Democrat.

Realizing that the above graph may not be immediately intuitive, we did some research into how Gallup News classifies counties based on Democrat to Republican ratio using something called the Likert Scale. Now, being able to group a particular county as 'Strong Republican', 'Competitive', or 'Strong Democrat', we can visualize the relationship between political leaning and mean number of cases through a bar graph (Figure 3). This was done by classifying 'Strong Republican' as counties with a Democrat to Republican ratio of less than 0.9, 'Competitive' between 0.9 and 1.1, and 'Strong Democrat' greater than 1.1.



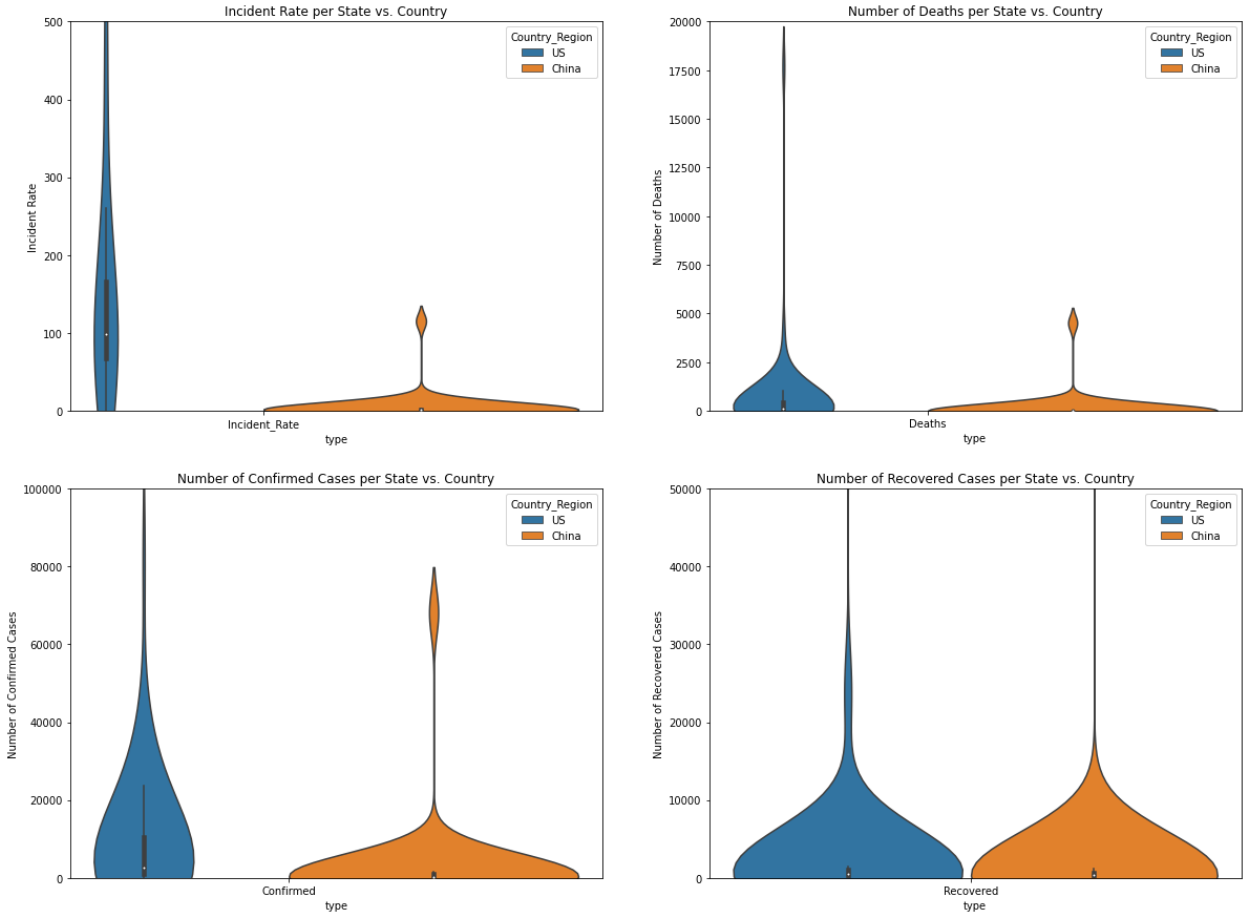
(Figure 3: County Political Leaning vs. Mean # of Cases per County until 4/18/20) This graph demonstrates a clear correlation between political leaning and the mean # of cases per county.

The last visualization we did for this question involved observing the correlation between a county's cumulative number of cases and when their shelter in place mandate started. To do this, we opted for another linear regression plot using sns.regplot, and grouped counties based on when their shelter in place date started (Figure 4).



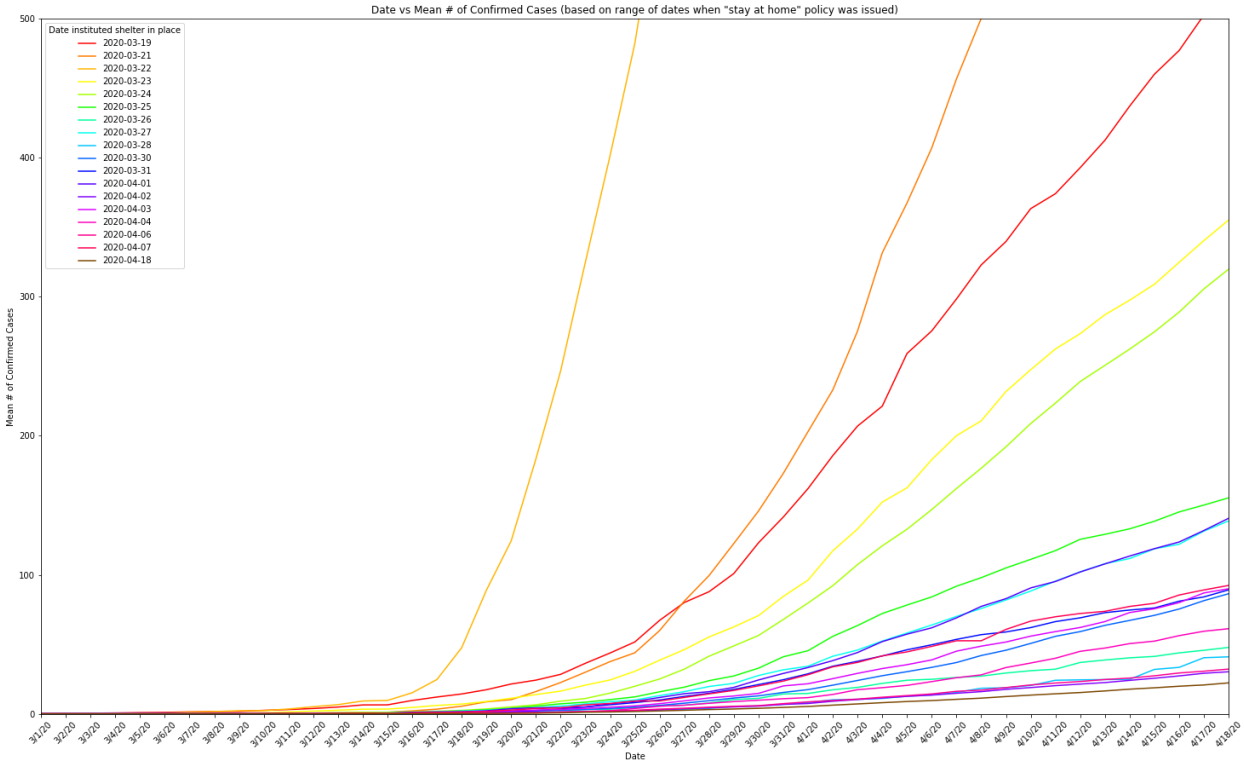
(Figure 4: Time when stay at home policy was issued vs. Cumulative Number of Cases until 4/18/20): This visualization shows a general negative correlation between # of confirmed cases and the date the stay at home policy was issued.

Question 2: How does the United States compare against China when it comes to preventing community spread?
We chose to do this through a violin plot, as outliers play a larger role in the visualization of the data (so areas with extremely high values such as New York and Hubei are not ignored), and the distribution of a violin plot gives us information about how each country may be performing in terms of restricting spread from area to area (Figure 5).

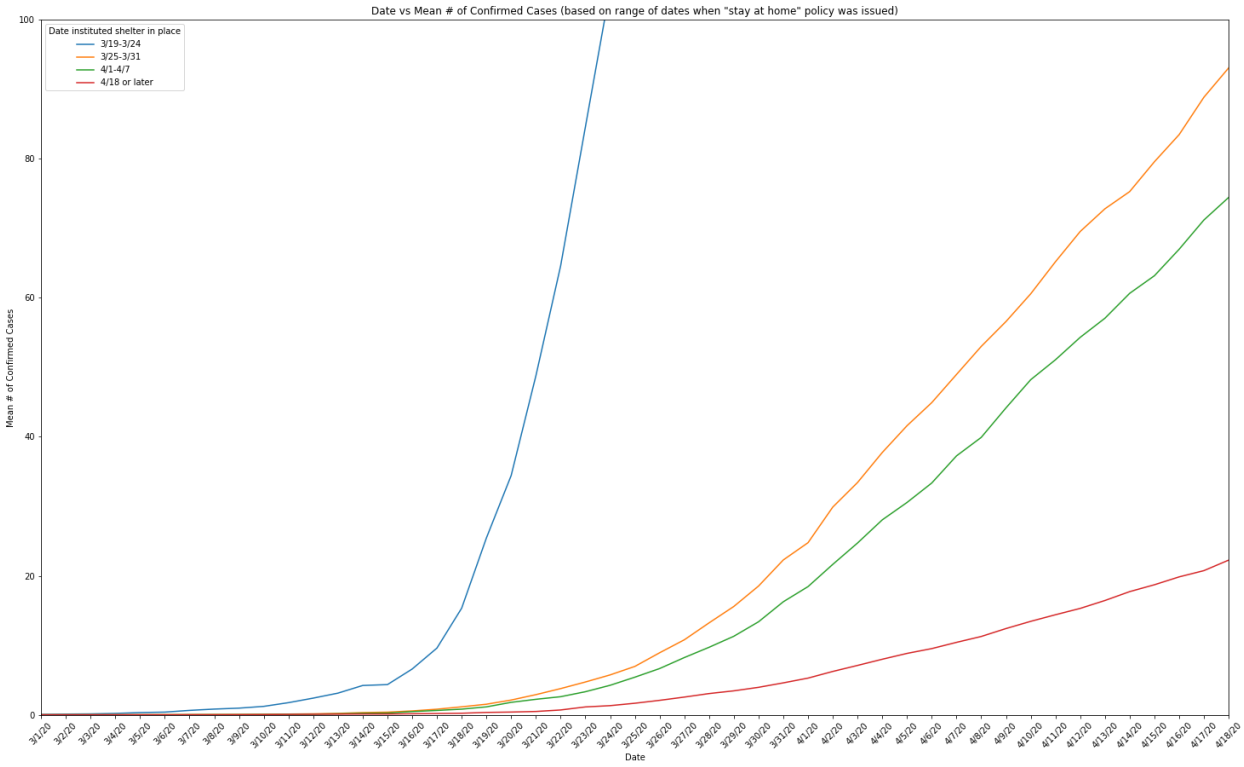


(Figure 5): Our most high level visualization, these violin plots compare the Incident Rate, Number of Deaths, Number of Confirmed Cases, and Number of Recovered Individuals between the USA (in blue) and China (in orange). As you can see, the US has more evenly spread-out distributions, suggesting that the US is not doing as well in terms of preventing spread from one location to another.

Question 3: How effective are shelter in place policies when comparing counties that started earlier than others?
For this question, we used a simple line plot that took in the start date of shelter in place as the x-axis, and plotted the mean # of confirmed cases on the y-axis. Since there were 18 different start dates (counting 4/18/20 for counties without shelter in place policies by that date), we used colors of the rainbow to distinguish each line (Figure 6). We also made a simpler visualization by grouping dates for readability (Figure 7).



(Figure 6: Date vs Mean # of Confirmed Cases (based on range of dates when 'stay at home' policy was instituted) This graph shows that the mean number of cases are higher for counties that began shelter in place at an earlier date.



(Figure 7: Date vs Mean # of Confirmed Cases (based on range of dates when 'stay at home' policy was issued)) Same concept as Figure 6, but groups certain dates that shelter in place started at. This graph also shows that the mean number of cases are higher for counties that began shelter in place at an earlier date.

Question 4: What features would work best to predict the cumulative number of confirmed cases in a particular county within the United States?

This will be further addressed in 'methods', but for this question we performed lasso regression and linear regression to help create our model, ultimately sticking with the model created by linear regression. Although we didn't use a formal visualization, we used print statements throughout our code (Figure 8) to demonstrate the features we were testing using cross validation , as well as print statements to effectively show our best cross validation error, worst cross validation error, and cross validation error (Figure 9) using only the number of confirmed cases two weeks prior to the test date. We finished by printing the Training Accuracy % and Test Error % of our linear model (Figure 10).

```
Trying feature(s): ['PopulationDensityperSqMile2010', 'stay at home', 'MedianAge2010', 'PopMale75-842010']
        RMSE: 558.8446431536714
Trying feature(s): ['PopulationDensityperSqMile2010', 'stay at home', '#Hospitals', 'PopFmle75-842010']
        RMSE: 553.9760288977716
Trying feature(s): ['PopulationDensityperSqMile2010', 'stay at home', '#Hospitals', 'PopMale75-842010']
        RMSE: 558.7163244077633
Trying feature(s): ['PopulationDensityperSqMile2010', 'stay at home', 'PopFmle75-842010', 'PopMale75-842010']
        RMSE: 553.5555199297379
Trying feature(s): ['PopulationDensityperSqMile2010', 'MedianAge2010', '#Hospitals', 'PopFmle75-842010']
        RMSE: 553.9903022877877
Trying feature(s): ['PopulationDensityperSqMile2010', 'MedianAge2010', '#Hospitals', 'PopMale75-842010']
        RMSE: 558.7831415254371
Trying feature(s): ['PopulationDensityperSqMile2010', 'MedianAge2010', 'PopFmle75-842010', 'PopMale75-842010']
        RMSE: 553.7119850736616
Trying feature(s): ['PopulationDensityperSqMile2010', '#Hospitals', 'PopFmle75-842010', 'PopMale75-842010']
        RMSE: 553.9159784150245
Trying feature(s): ['SVIPercentile', '4/4/20', 'dem_to_rep_ratio', 'stay at home']
        RMSE: 269.3858931260141
Trying feature(s): ['SVIPercentile', '4/4/20', 'dem_to_rep_ratio', 'MedianAge2010']
        RMSE: 269.3858931260141
Trying feature(s): ['SVIPercentile', '4/4/20', 'dem_to_rep_ratio', '#Hospitals']
        RMSE: 254.31149367397575
Trying feature(s): ['SVIPercentile', '4/4/20', 'dem_to_rep_ratio', 'PopFmle75-842010']
        RMSE: 257.524575441634
Trying feature(s): ['SVIPercentile', '4/4/20', 'dem_to_rep_ratio', 'PopMale75-842010']
        RMSE: 258.1459056459719
Trying feature(s): ['SVIPercentile', '4/4/20', 'stay at home', 'MedianAge2010']
        RMSE: 271.8449070885981
Trying feature(s): ['SVIPercentile', '4/4/20', 'stay at home', '#Hospitals']
        RMSE: 254.71451391886023
Trying feature(s): ['SVIPercentile', '4/4/20', 'stay at home', 'PopFmle75-842010']
        RMSE: 257.45777998163396
Trying feature(s): ['SVIPercentile', '4/4/20', 'stay at home', 'PopMale75-842010']
        RMSE: 258.29813769709324
Trying feature(s): ['SVIPercentile', '4/4/20', 'MedianAge2010', '#Hospitals']
        RMSE: 254.7144914801654
Trying feature(s): ['SVIPercentile', '4/4/20', 'MedianAge2010', 'PopFmle75-842010']
        RMSE: 257.45873143344727
Trying feature(s): ['SVIPercentile', '4/4/20', 'MedianAge2010', 'PopMale75-842010']
        RMSE: 258.2981243341935
Trying feature(s): ['SVIPercentile', '4/4/20', '#Hospitals', 'PopFmle75-842010']
        RMSE: 257.59646429937897
Trying feature(s): ['SVIPercentile', '4/4/20', '#Hospitals', 'PopMale75-842010']
        RMSE: 258.66978924369437
Trying feature(s): ['SVIPercentile', '4/4/20', 'PopFmle75-842010', 'PopMale75-842010']
        RMSE: 257.4577713825073
Trying feature(s): ['SVIPercentile', 'dem_to_rep_ratio', 'stay at home', 'MedianAge2010']
        RMSE: 617.4789903623412
Trying feature(s): ['SVIPercentile', 'dem_to_rep_ratio', 'stay at home', '#Hospitals']
        RMSE: 547.077312566849
Trying feature(s): ['SVIPercentile', 'dem_to_rep_ratio', 'stay at home', 'PopFmle75-842010']
        RMSE: 548.1822208401966
Trying feature(s): ['SVIPercentile', 'dem_to_rep_ratio', 'stay at home', 'PopMale75-842010']
        RMSE: 553.3321732858273
Trying feature(s): ['SVIPercentile', 'dem_to_rep_ratio', 'MedianAge2010', '#Hospitals']
        RMSE: 547.9113676452862
Trying feature(s): ['SVIPercentile', 'dem_to_rep_ratio', 'MedianAge2010', 'PopFmle75-842010']
```

(Figure 8: Various RMSE for different features during cross validation) This was our way of displaying the different features that our linear model was testing to come up with a model with the lowest RMSE.

```
The lowest CV error was 197.47373150633265%, which came from using these feature(s): ['PopulationDensityperSqMile2010', '4/4/20', 'dem_to_rep_ratio',
The highest CV error was 611.3359195562614% which came from using these feature(s): ['SVIPercentile'].
CV Error with just the date alone was 209.43701251209782%.
```

(Figure 9: Printed statements to show the percentage CV error of our best, worst, and simplest model). These statements tell us that using 'PopulationDensityperSqMile2010', '4/4/20', 'dem_to_rep_ratio', 'MedianAge2010', 'PopFmle 75-842010', and 'PopMale75-842010' yield the lowest CV error of 197.47%.

```
Training Accuracy: 92.06705675332464%
The test error is 1026.314388785715%.
```

(Figure 10: Printed statements to display what our total Training Accurate % and Test Error % is on average). This figure shows that our linear model's training accuracy was close to 92%, and that for the test set, we were off on average by about 1026% from the actual number, indicating overfitting.

Method and Experiments (and some analysis)

Question 1: Does the political leaning of a particular county have correlation with its cumulative number of cases, and does the cumulative number of cases have an impact on when emergency protocol such as shelter in place begins?

We started by using a linear regression plot to demonstrate a positive correlation between the ratio of Democrats to Republicans in a particular county and their cumulative number of cases up until 4/18/20. To further emphasize our point, we grouped counties based on a method that Gallup News does it (mentioned in Visualizations) and calculated the mean # of confirmed cases for the category to demonstrate that counties that run Democrat have much higher #s of confirmed cases than Republican counties using a bar plot. Lastly, we showed using a linear regression plot that counties that started shelter in place earlier tend to have higher #s of confirmed cases on average to start with. This last visualization helps to supplement our conclusion to Question #3, where we assert that the initial number of infected individuals at the start of shelter in place renders us unable (at least with the current data) to show the effectiveness of shelter in place.

Question 2: How does the United States compare against China when it comes to preventing community spread?

To tackle this question, we used country_cleaned to compare the Incident Rates, # of Deaths, # of Confirmed Cases, and # of Recovered Cases of the US and China using a violin plot. As you can see, the US has more evenly spread-out distributions for Incident Rate, Deaths, and Confirmed, suggesting that the US is not doing as well in terms of preventing spread from one location to another. The US and China have relatively identical distributions for the Recovered graph, showing that people are recovering at an equal rate as of 4/18/20.

Question 3: How effective are shelter in place policies when comparing counties that started earlier than others?

We attempted to solve this problem by grouping counties by the Timestamp by which they initiated shelter in place policies, then aggregated the number of cases by the mean, and then plotted them across time. Our method turned out to be ineffective to compare counties that started shelter in place earlier than those who started later or did not implement shelter in place policies at all. Intuitively, we thought our data should show a lower rate of change of confirmed cases, but we actually end up seeing that counties that start shelter in place earlier have exponentially growing mean # of confirmed cases, whereas counties without shelter in place policies seem to be doing much better (Figures 6 and 7). As a result, our approach didn't really show the effectiveness of the policy - we concluded that we would need more data to model the rate of change of mean # of confirmed cases in order to demonstrate that the policy would actually have an impact. A quick note is that we grouped based on timestamp because not all counties in a particular state started shelter in place at the same time.

Question 4: What features would work best to predict the cumulative number of confirmed cases in a particular county within the United States?

This question took the most amount of time to think about method and approach by far. The approach to this question went through several iterations. We initially started by using a linear model that would use 4-fold cross validation to take in n possible features, test the loss of each combination of the n features using RMSE, and eventually output the features that yielded the lowest cross validation error. We used a train-test split of 0.7-0.3, and normalized the data. Our selected potential features were Population Size, Density, Social Vulnerability Index, Cumulative Cases as of 4/4/20, Democrat to Republican Ratio, Stay at home policy date, Median Age, # of Hospitals, and the Population of Male and Female individuals between 75-84 years of age. We then found that our model was overfitting, as changes to the random state impacted the feature selection of our model heavily (which shouldn't happen if we're not overfitting). We then went on to apply Lasso regularisation. This proved to be more useful, as changes to alpha were much more stable with the CV error (only ranging from 227 to 234).

Analysis and Conclusion

We'll start this section with a direct response to the seven questions posed, then finish with summary of findings, analysis of approach, and future recommendations.

I. Some of the most interesting features that we came across for our questions were the usage of Democrat to Republican ratio when looking at cumulative number of cases (a member of our group is a Political Science major so this was particularly interesting to him), and learning about SVI percentile and how the CDC was able to quantify social vulnerability to things such as disease outbreak.

II. Although SVI percentile seemed to be a useful feature to use (we used the intuition that counties with a lower percentile would be more susceptible to COVID-19 spread), it turned out to not be one of the features ultimately used in our linear model.

III. One of our biggest challenges was realizing that we only had data up until 4/18/20, which really isn't a lot of data to work off of. For example, we were initially going to show how effective shelter in place orders were, but realized that 'flattening the curve' would be difficult to visualize given that COVID-19 cases will always have a lag time between the start of a policy and the manifestation of any impact it might have. Additionally, we were missing so many values that had no good replacement that we ended up not using some promising features.

IV. However, our greatest limitations came from a lack of domain knowledge (by the time we read about the SIR model, we had already completed the project and were in the midst of finals week) as well as simply the lack of data by being in the midst of the pandemic, rather than performing these experiments after the pandemic had finished. Additionally, some of our work was computationally expensive, and we didn't have enough storage in our kernel to test every feature provided.

V. One ethical dilemma we faced were the types of counties that we would not be including in our analysis when performing data cleaning. For example, when filtering out the two locations in confirmed_cleaned and death_cleaned,

those two locations were the Michigan Department of Corrections and the Federal Correction Institution. By filtering out those two completely, we neglect part of the incarcerated population which is an ethical concern. However, because we had to merge based on the Federal Information Processing Standard, we filtered that data out.

VI. If we had *accurate* data (the problem is that not enough testings are being administered) as well as performed these tests after COVID-19 has largely died down (rather than in the midst of the pandemic), we would be able to conduct tests for our question in response III (how shelter in place policies affected total number of confirmed cases).

VII. One ethical concern when looking at the effectiveness of shelter in place based on timestamp is the issue of aggregation. Effectively, when we are taking the mean, median, or mode of any dataset, we often neglect outliers, and for example, a smaller county's number of confirmed cases would be dwarfed by a larger county's number of confirmed cases when taking the mean.

Summary of Findings

First, we used linear regression plots and bar plots to show that counties who lean Democrat typically have higher case counts of coronavirus, leading to an earlier implementation of policies such as shelter in place. This was supplemented by our attempt to answer our third question, but we could not extrapolate any useful conclusions from our efforts to see the relationship between the date that shelter in place was instituted and # of confirmed cases. Second, we used violin plots to show that the United States is struggling to contain community spread when compared to China. Last, we used a linear regression model and lasso regularization to find features that would supplement existing knowledge to help predict the number of confirmed cases we'll see in the future.

Overall, the approaches for questions 1 and 2 were solid, and we were able to draw definitive conclusions from our subsequent visualizations and analysis. However, our greatest limitations for questions 3 and 4 came from a lack of domain knowledge (by the time we read about the SIR model, we had already completed the project and were in the midst of finals week) as well as simply the lack of data by being in the midst of the pandemic, rather than performing these experiments after the pandemic had finished. Additionally, some of our work was computationally expensive, and we didn't have enough storage in our kernel to test every feature provided. In terms of future work, we would recommend re-examining the same questions after the pandemic with updated data, soliciting more feedback from an expert with domain knowledge about pandemics, and running our code in a place where the kernel doesn't crash.

Works Cited

Agency for Toxic Substances and Disease Registery. “CDC SVI 2016 Documentation.” *CDC's Social Vulnerability Index (SVI)*, Center for Disease Control, 12 Sept. 2018, svi.cdc.gov/Documents/Data/2016_SVI_Data/SVI2016Documentation.pdf.

Alaska. “2016 General Election.” *GEMS ELECTION RESULTS*, 30 Nov. 2016, 12:39:48, www.elections.alaska.gov/results/16GENR/data/results.htm.

CSSEGISandData. “CSSEGISandData/COVID-19.” *GitHub*, Johns Hopkins University, 13 May 2020, github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data.

John, Arit. “These Eight States Haven't Issued Stay-at-Home Orders to Fight the Coronavirus Outbreak.” *Los Angeles Times*, Los Angeles Times, 22 Apr. 2020, www.latimes.com/politics/story/2020-04-22/states-without-coronavirus-stay-at-home-order.

Johns Hopkins University. “Hubei Timeline.” *Johns Hopkins Coronavirus Resource Center*, Johns Hopkins University & Medicine, 2020, coronavirus.jhu.edu/data/hubei-timeline.

Jones, Jeffrey M. “GOP Maintains Edge in State Party Affiliation in 2016.” *Gallup.com*, Gallup, 6 Nov. 2017, news.gallup.com/poll/203117/gop-maintains-edge-state-party-affiliation-2016.aspx.

Mervosh, Sarah, et al. “See Which States and Cities Have Told Residents to Stay at Home.” *The New York Times*, The New York Times, 24 Mar. 2020, www.nytimes.com/interactive/2020/us/coronavirus-stay-at-home-order.html.

Nelson, Cody. “Shelter-in-Place Orders: Looking at Other States to See What a MN Order Might Be.” *MPR News*, MPR News, 23 Mar. 2020, www.mprnews.org/story/2020/03/21/what-would-a-shelterinplace-order-look-like-in-minnesota.