Prediction of Covid-19 Deaths and Cases by Method of Regression Modeling

Data 100. Spring 2020 Final Project

Names: Emmy Yu, Ruqian Wang, Shaye Hong

Abstract & Introduction

This Notebook is aimed at trying to answer the questions relating to the spread of Covid-19 in the United States. In particular, we will be applying regression models such as Linear, Ridge, and Lasso Regression to predict Covid-19 deaths and cases for counties in the United States. Some questions we intend to answer are: What factors/features are most important in the spread of Covid-19? Can we use modeling techniques to predict the number of Covid-19 deaths and cases in the next few months?

Modeling and analysis will be carried out based on datasets detailing confirmed deaths and confirmed cases of Covid-19 as of April 18, 2020, as well as statistics of individual counties and states. Of course, these datasets will be cleaned to be made compatible with our modeling; erroneous data values will be dropped or imputed. A section on Exploratory Data Analysis is also carried out, showing interesting findings from these statistics and mapping the spread of Covid-19 across the U.S.

This analysis comes at an appropriate time as the Covid-19 pandemic continues to spread globally. Creating data-driven predictive models has the potential to provide estimates on how many people will be directly affected by this disease. Early prediction of Covid-19 could allow the government to effectively allocate funding, and to allow local hospitals to adequately prepare for the potential influx of new patients.

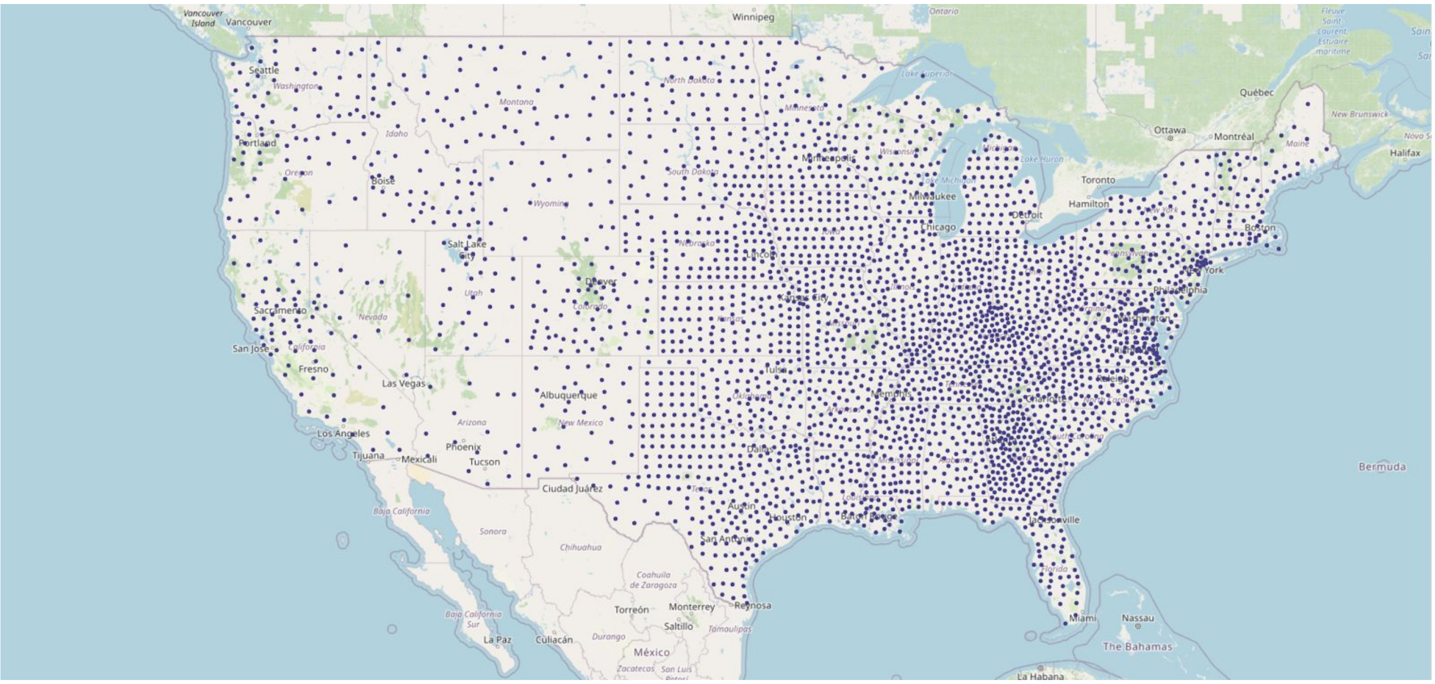
About the Datasets: Covid-19

The datasets used in this Notebook are publicly-available. Links to original sources are detailed below:

- **Confirmed Cases:** (time_series_covid19_confirmed_US.csv)
- **Confirmed Deaths:** (time_series_covid19_deaths_US.csv)
- **Counties Statistics:** (abridged_couties.csv, Column Descriptions)
- **States Statistics:** (4.18states.csv)

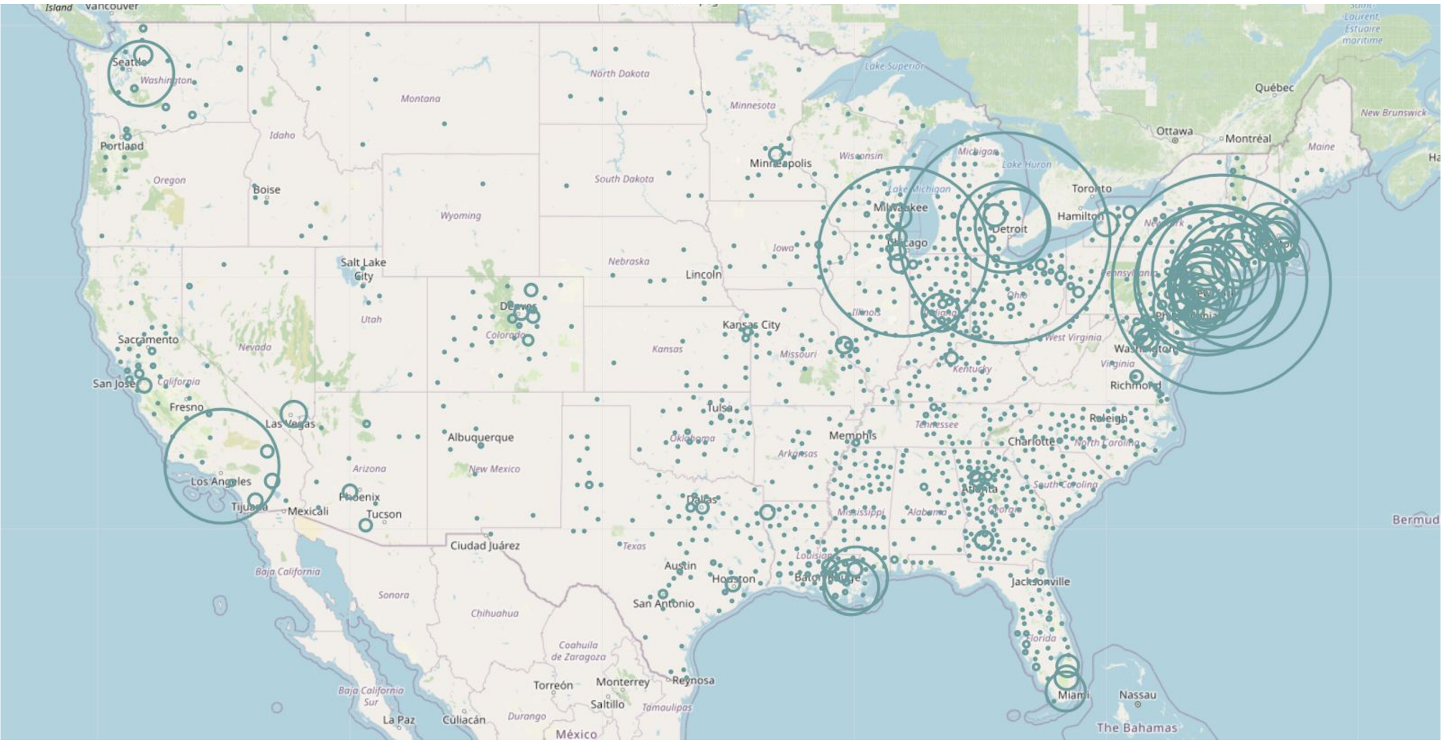
Exploratory Data Analysis

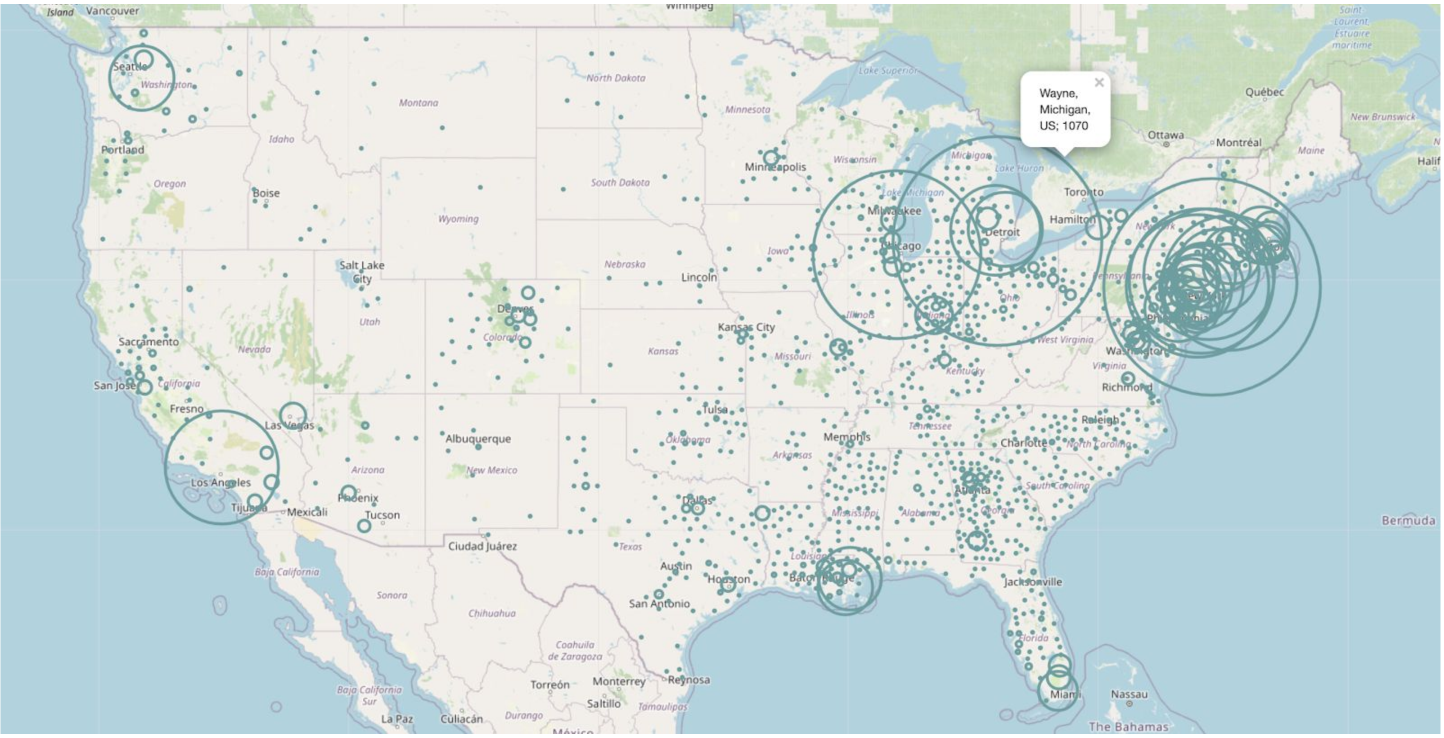
The following plot shows the geographical location of each of the counties given in the dataset of confirmed deaths. Note the geographical density of the counties. In the next few maps, the number of confirmed deaths and cases due to Covid-19 will be displayed.



Mapping confirmed deaths in the U.S. as of 4/18/20.

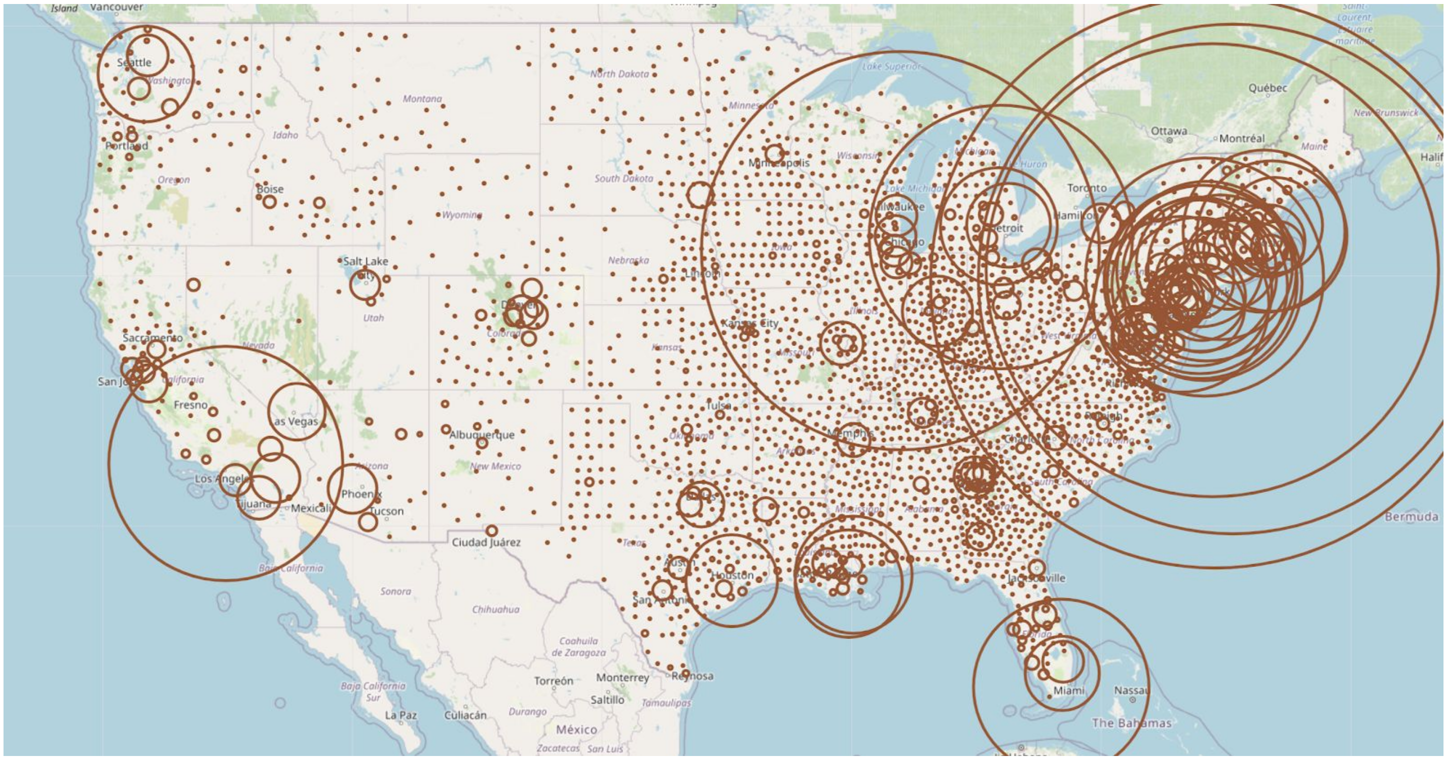
The size of each of the markers corresponds to the number of confirmed deaths due to Covid-19. As an interactive feature in this Notebook, clicking on each marker will display the location name and the actual number of deaths in the county as of 4/18/20. Note that the counties with 0 confirmed deaths are left out of the map. Non-contiguous U.S. states and territories are left out of the initial zoomed frame. Note also that misreported longitude and latitude coordinates show a hotspot off the coast of Africa (these points can be attributed to cruise ships and correctional facilities which did not have a reported geographical location in our datasets; these points are removed from our features training sets).

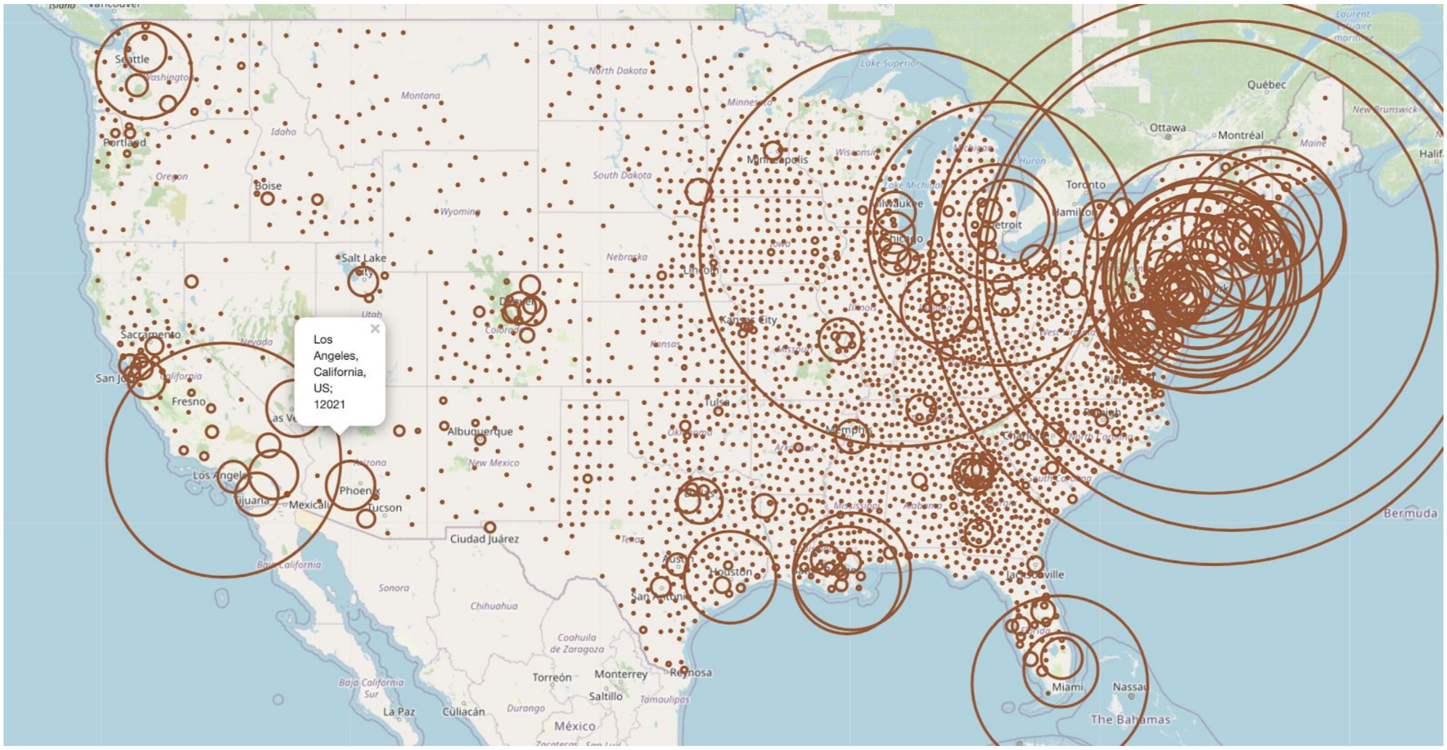




Mapping confirmed cases in the U.S. as of 4/18/20.

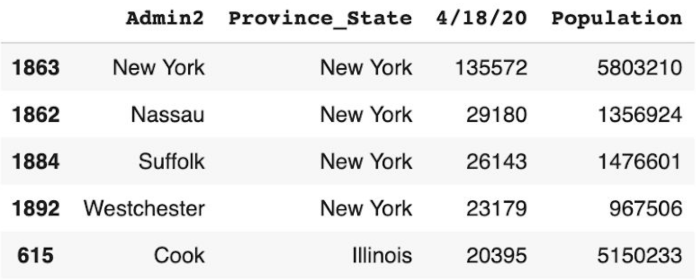
The size of each of the markers corresponds to the number of confirmed cases due to Covid-19. As an interactive feature, clicking on each marker will display the location name and the actual number of cases in the county as of 4/18/20. Note that the counties with 0 confirmed cases are left out of the map. For the ease of visualization, the scale of the marker sizes has increased (10 times) from the map of confirmed deaths (there are much more confirmed cases than deaths).





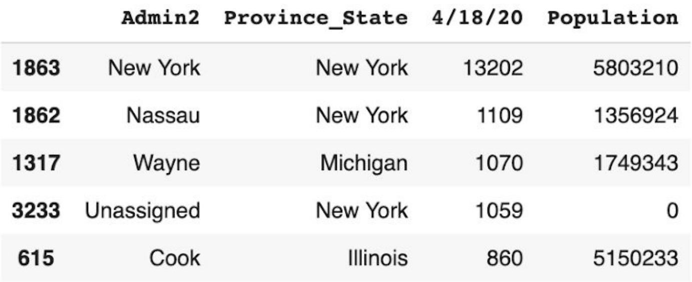
By looking at the marker sizes in the graphs above, it is obvious that the greatest number of confirmed cases come from New York. After sorting the cumulative cases counts, the top 4 counties with the highest confirmed cases and the top 2 counties with the highest deaths come from New York. The actual values on 4/18/20 are displayed below alongside their respective population size. Unassigned data values will be dropped in the data cleaning portion of this notebook.

Counties with the highest confirmed cases of Covid-19



	Admin2	Province_State	4/18/20	Population
1863	New York	New York	135572	5803210
1862	Nassau	New York	29180	1356924
1884	Suffolk	New York	26143	1476601
1892	Westchester	New York	23179	967506
615	Cook	Illinois	20395	5150233

Counties with the highest deaths due to Covid-19

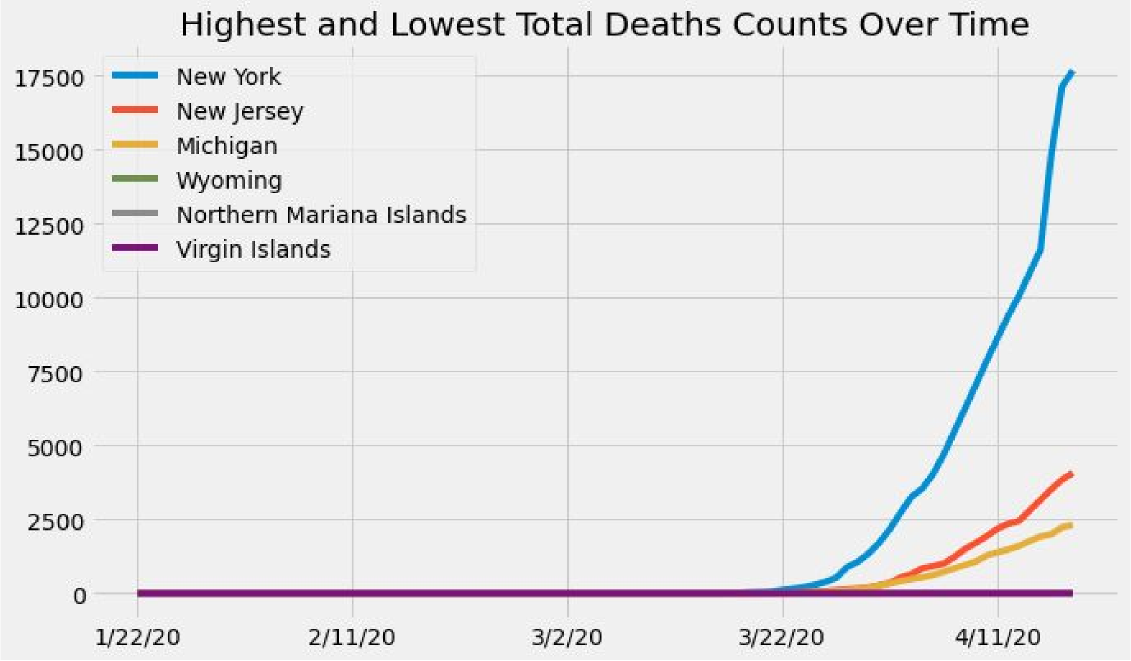


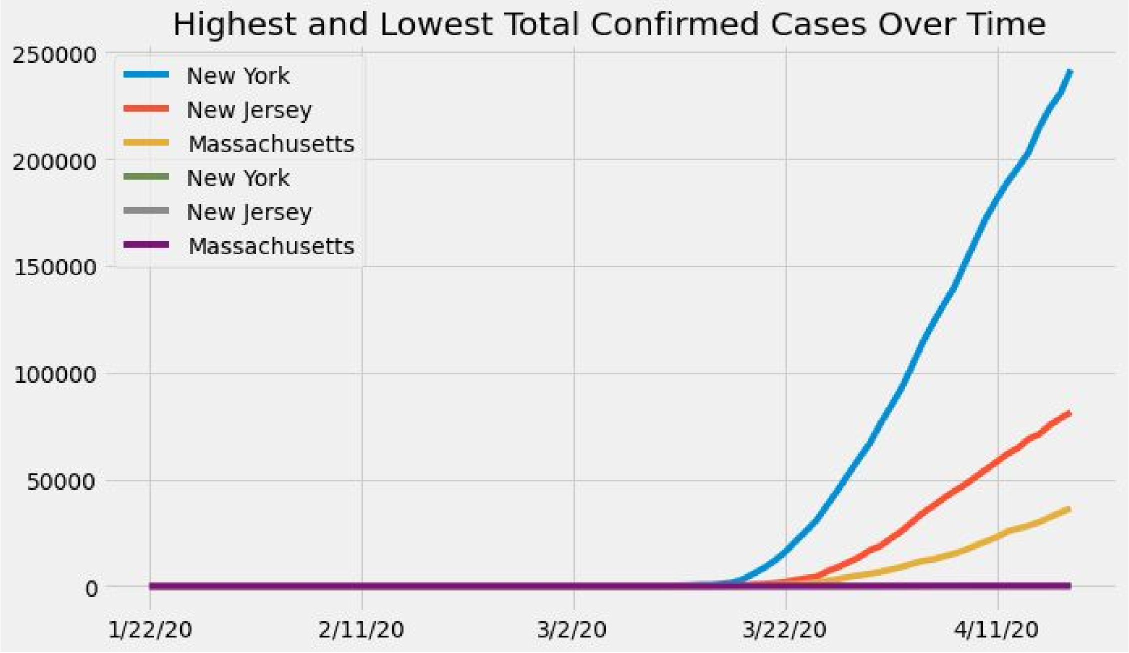
	Admin2	Province_State	4/18/20	Population
1863	New York	New York	13202	5803210
1862	Nassau	New York	1109	1356924
1317	Wayne	Michigan	1070	1749343
3233	Unassigned	New York	1059	0
615	Cook	Illinois	860	5150233

To help determine which features and geographic locations might be important for understanding the spread of COVID-19, we examined the number of deaths and infections over time for the areas with the highest and lowest infection rates.

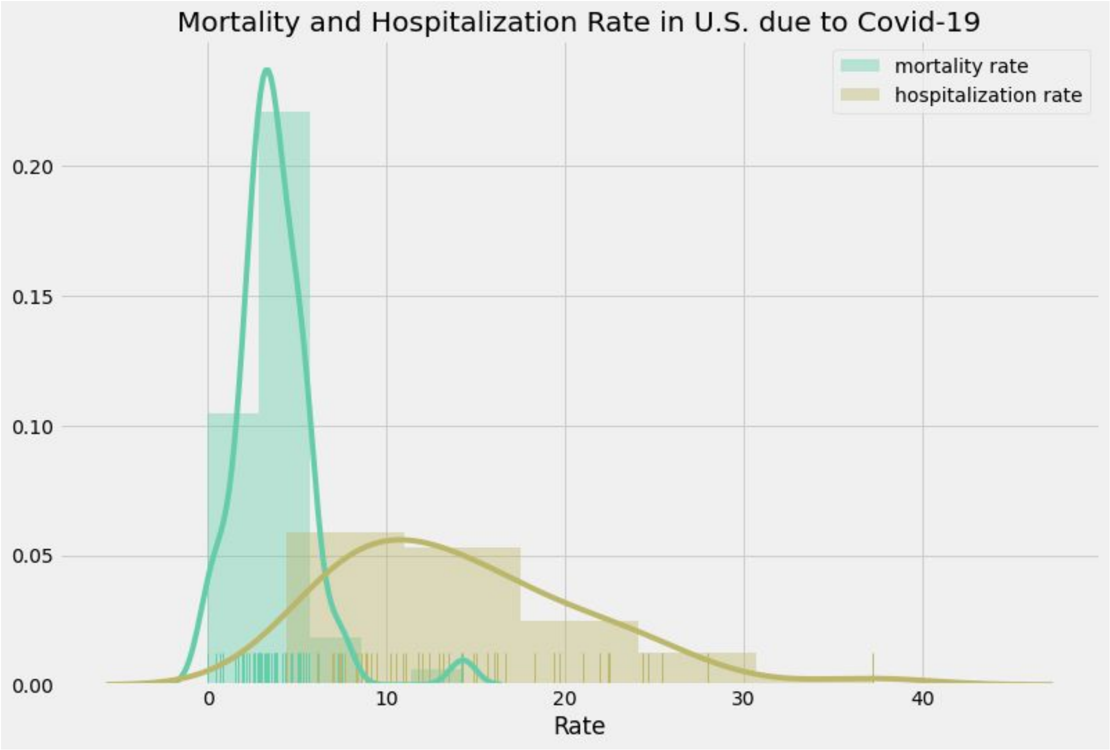
- Highest number of deaths found in: ['New York', 'New Jersey', 'Michigan']

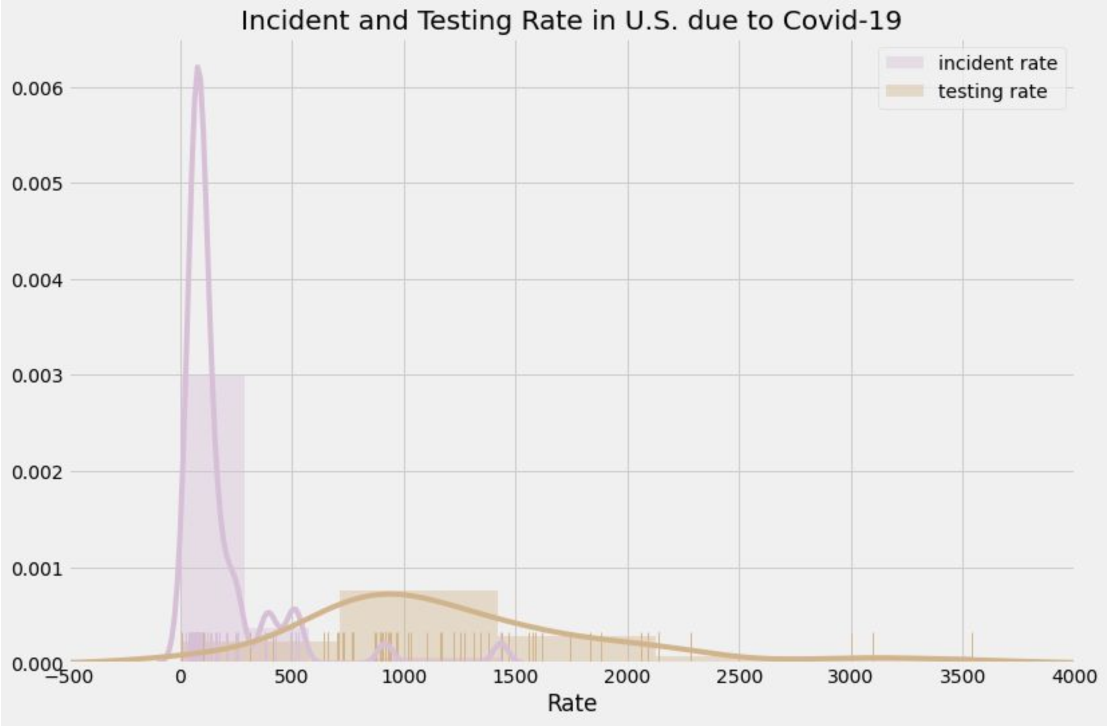
- Lowest number of deaths found in: ['Wyoming', 'Northern Mariana Islands', 'Virgin Islands']
- Highest number of cases found in: ['New York', 'New Jersey', 'Massachusetts']
- Lowest number of cases found in: ['Northern Mariana Islands', 'Virgin Islands', 'Guam']





To help visualize the mortality rate, hospitalization rate, incident rate, and testing rate, we chose to use a distplot to depict their distributions. The kde and rug distributions are overlaid in the visualization.





Data Cleaning and Standardization

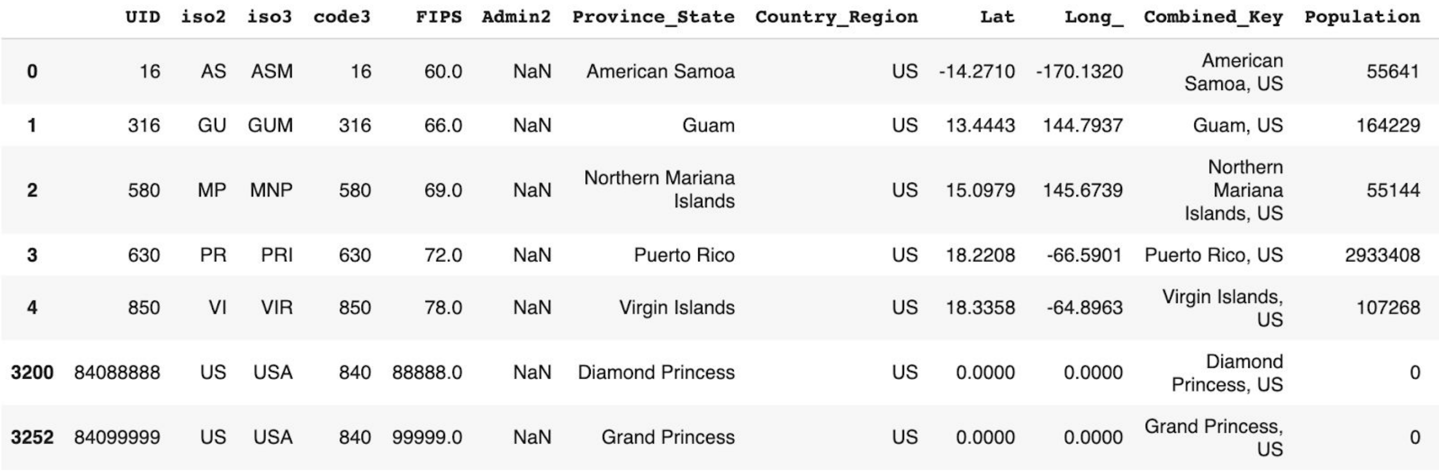
The goal of our analysis is to use the confirmed deaths and cases data for each county to predict future estimates over time. To promote fair comparisons, the following considerations were taken into account during standardization/cleaning of the data:

- **Only data from counties are used** (cruise ships, correctional facilities, unassigned values, etc are dropped)
- **Adding Population** (simply using counts of cases is misleading since population totals change from county to county)

- **Counties without any confirmed cases** (counties without any confirmed cases are dropped from the cases and deaths datasets to prevent imbalanced datasets. Note that there are counties with confirmed cases but no confirmed deaths)
- **Erroneous populations** (Example: in one case, two counties (Dukes and Nantucket, Massachusetts) were lumped together in deaths and cases dataset, and population was erroneously given a 0 population count)
- **Cases dataset given population totals** (the cases dataset was given a column for county populations as a modeling feature)

The dataset of confirmed deaths has rows for each county in the United States. Taking a look at the null values for the **Admin2** column for county names, the null values are cruise ships (Grand Princess, Diamond Princess) and U.S. territories (Guam, Virgin Islands, etc).

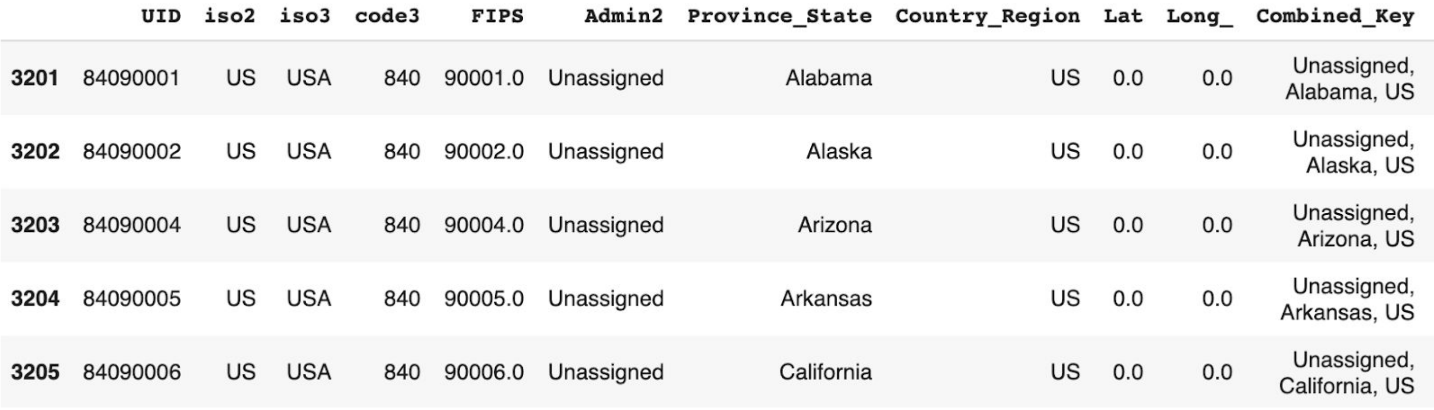
Given that different people regularly embark and disembark cruise ships, this data would not suit our regression models that aim to predict cumulative estimates of Covid-19 cases and deaths. Therefore, the analysis will be limited to only counties. While other counties provide death/case data that is cumulative over time, the Grand Princess cruise ship for example abruptly dropped their death count back to 0 on April 12, 2020 when their passengers disembarked. **Therefore, the cruise ships will be removed from our both the deaths and cases datasets.**



	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	Population
0	16	AS	ASM	16	60.0	NaN	American Samoa	US	-14.2710	-170.1320	American Samoa, US	55641
1	316	GU	GUM	316	66.0	NaN	Guam	US	13.4443	144.7937	Guam, US	164229
2	580	MP	MNP	580	69.0	NaN	Northern Mariana Islands	US	15.0979	145.6739	Northern Mariana Islands, US	55144
3	630	PR	PRI	630	72.0	NaN	Puerto Rico	US	18.2208	-66.5901	Puerto Rico, US	2933408
4	850	VI	VIR	850	78.0	NaN	Virgin Islands	US	18.3358	-64.8963	Virgin Islands, US	107268
3200	84088888	US	USA	840	88888.0	NaN	Diamond Princess	US	0.0000	0.0000	Diamond Princess, US	0
3252	84099999	US	USA	840	99999.0	NaN	Grand Princess	US	0.0000	0.0000	Grand Princess, US	0

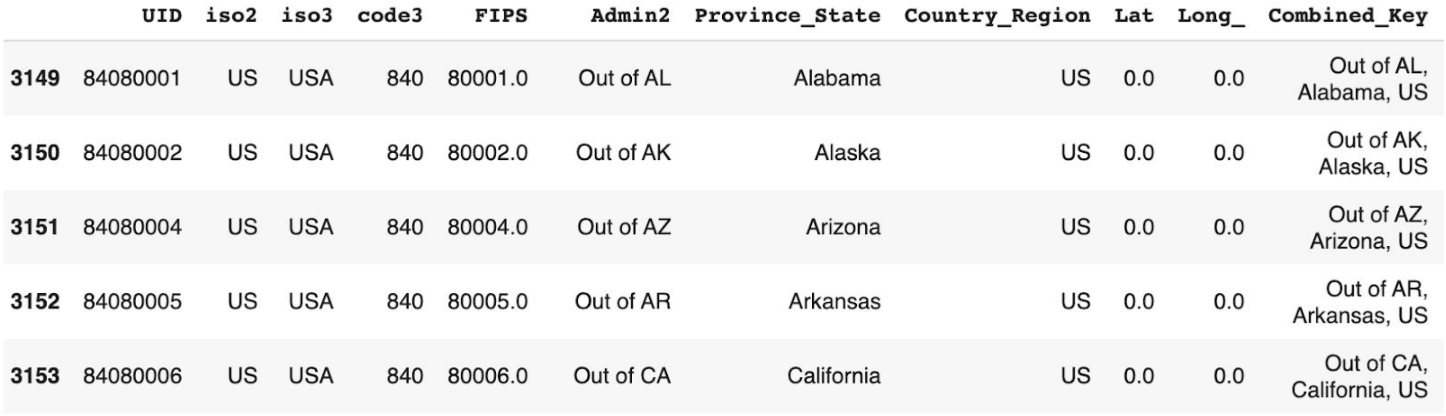
To aid in future merging with other datasets, the null values in both deaths and cases datasets for the U.S. territories in **Admin2** will be replaced with their **Province_State** name. The province names for these locations will effectively act as county names. Now, there are 0 null values for country names in **Admin2**.

In addition to null values representing county names, there are also 51 data entries that label the county to be "Unassigned". Taking a closer look at the death counts, the values are not all cumulative over time, which suggests that these values may be misreported or may represent non-cumulative information. **Since the goal of our analysis is to potentially predict death/cases rates for different counties, these unassigned rows will be dropped from our dataset.**



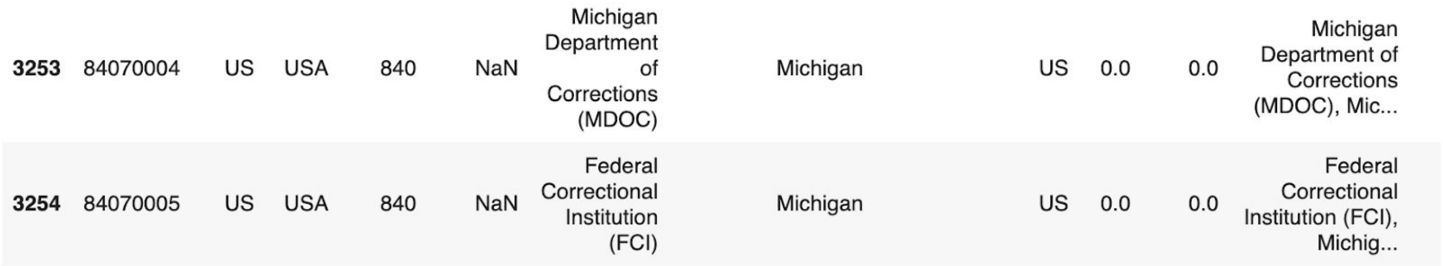
	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key
3201	84090001	US	USA	840	90001.0	Unassigned	Alabama	US	0.0	0.0	Unassigned, Alabama, US
3202	84090002	US	USA	840	90002.0	Unassigned	Alaska	US	0.0	0.0	Unassigned, Alaska, US
3203	84090004	US	USA	840	90004.0	Unassigned	Arizona	US	0.0	0.0	Unassigned, Arizona, US
3204	84090005	US	USA	840	90005.0	Unassigned	Arkansas	US	0.0	0.0	Unassigned, Arkansas, US
3205	84090006	US	USA	840	90006.0	Unassigned	California	US	0.0	0.0	Unassigned, California, US

A similar situation is seen in with **Admin2** names that label the counties as "Out of AL" or "Out of CA" for different states. **Since these data entries do not correspond to any particular county, these values will also be dropped from our datasets.**



	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key
3149	84080001	US	USA	840	80001.0	Out of AL	Alabama	US	0.0	0.0	Out of AL, Alabama, US
3150	84080002	US	USA	840	80002.0	Out of AK	Alaska	US	0.0	0.0	Out of AK, Alaska, US
3151	84080004	US	USA	840	80004.0	Out of AZ	Arizona	US	0.0	0.0	Out of AZ, Arizona, US
3152	84080005	US	USA	840	80005.0	Out of AR	Arkansas	US	0.0	0.0	Out of AR, Arkansas, US
3153	84080006	US	USA	840	80006.0	Out of CA	California	US	0.0	0.0	Out of CA, California, US

In addition, it is found that two entries correspond to the Michigan Department of Corrections (MDOC) facilities. **Since these locations are not counties as well, these two correctional facilities will also be dropped from the dataset.**

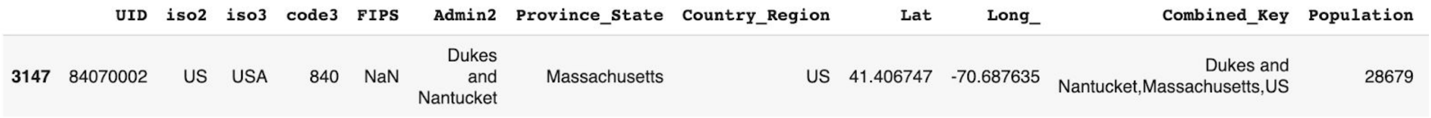


3253	84070004	US	USA	840	NaN	Michigan Department of Corrections (MDOC)	Michigan	US	0.0	0.0	Michigan Department of Corrections (MDOC), Mic...
3254	84070005	US	USA	840	NaN	Federal Correctional Institution (FCI)	Michigan	US	0.0	0.0	Federal Correctional Institution (FCI), Michig...

As seen in the Exploratory Data Analysis section above, it can be seen that there is a substantial proportion (approximately 14% of the dataset) of counties without a single confirmed case of Covid-19 as of 4/18/20. Since the goal of our models is to predict future estimates of cases and deaths, it would be pointless to train on nonexistent data. This high prevalence of 0 values would also lead to an imbalanced dataset, where the model may overfit and tend to predict exclusively 0 values.

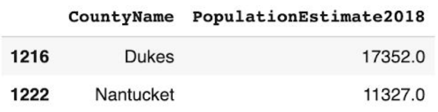
Therefore, the counties without a single confirmed case will be dropped from the dataset. These same counties will be dropped from the deaths dataset. Note that there are counties with confirmed cases, but no deaths as of 4/18/20.
Taking a look at the **Population** column provided in the deaths dataset, there is a single entry for the Dukes and Nantucket counties with a population of 0. It appears that these two counties were lumped together in this dataset, and a population count wasn't inputted properly. **Therefore, using the counties dataset, the 2018 population estimate for both counties were summed and imputed to replace the erroneous 0 value.**

As it appears in the deaths dataset:



	UID	iso2	iso3	code3	FIPS	Admin2	Province_State	Country_Region	Lat	Long_	Combined_Key	Population
3147	84070002	US	USA	840	NaN	Dukes and Nantucket	Massachusetts	US	41.406747	-70.687635	Dukes and Nantucket,Massachusetts,US	28679

As it appears in the counties dataset:



	CountyName	PopulationEstimate2018
1216	Dukes	17352.0
1222	Nantucket	11327.0

Since the cases dataset doesn't have a column for population, the population column from the deaths dataset will also be copied over.

Modeling and Experiments

Due to the current nature of Covid-19, any type of regression modeling to predict future cases and deaths is considered unsupervised learning. However, in order to use metrics that test the accuracy of our models, we will use the available data from the dates 1/22/20 - 4/11/18 as features to predict death/cases counts for 4/18/20. Early predictions of Covid-19 in modeling would be more beneficial since this gives counties time to prepare. We do not include the week of 4/11/20 - 4/17/20 because we want to make early predictions and not overtrain the model. In this way, the available data 4/18/20 will be used as true values that can be compared to the predicted values. Regression will be performed on the cases and deaths datasets separately.

Goals:

- Use OLS, RIDGE, LASSO models to predict the spread of COVID in the next few months
- Determining which features are most important using cross validation
- Assess the accuracy of each model

To perform modeling and analysis on our dataset, the dataset will be split into training and validation sets:

- Training set 70%
- Validation set 15%
- Testing set 15%

Least Squares Regression Model

As a baseline model, the analysis will start with the Least Squares Linear Regression model.

Why Least Squares?

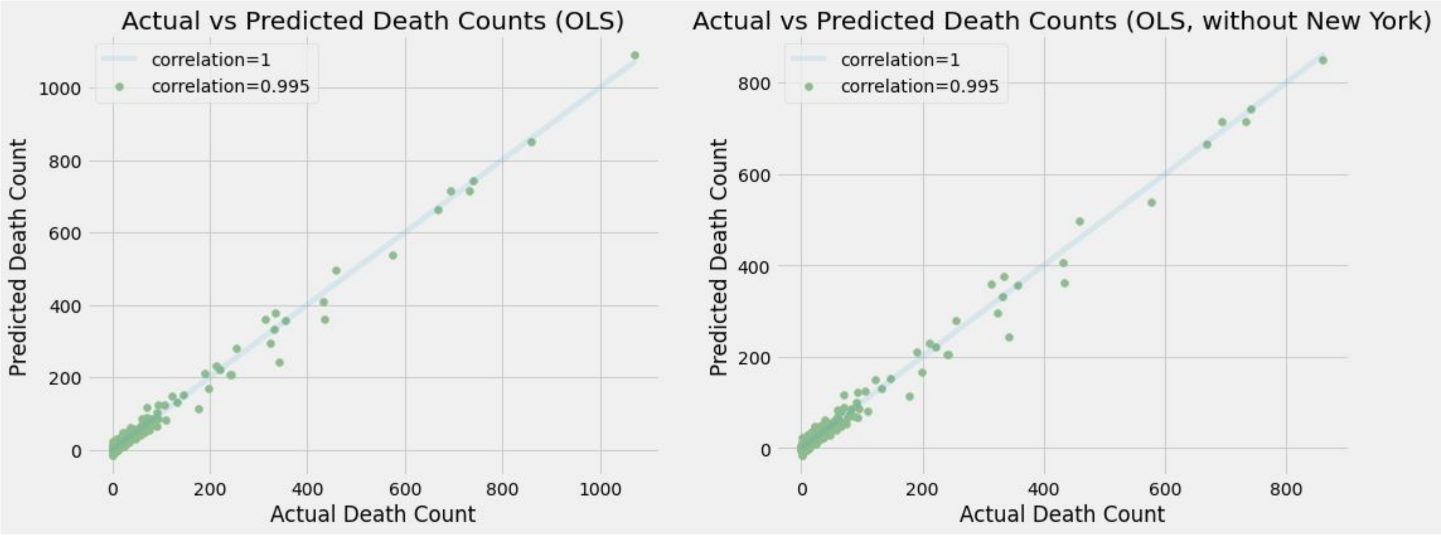
OLS is an optimization method that minimizes the sum of the squared residuals. For example, this method will draw a linear line through data points that will minimize the distance between those points and the predicted values on that line. In this case, the function is parametric since there is a linear relationship, $y = \beta_0+\beta_1 x$, in which the parameters β_0, β_1 are known. OLS is best used when the underlying data itself has a linear association, and a line of best fit can be drawn through the data. This makes it easy and efficient to implement either by hand or through computational linear algebra. OLS is easily interpretable and understandable, in which the best prediction for an underlying linear distribution should be the line of best fit.

However, the least squares method is very susceptible to outliers because of the squaring effect of the calculation. In addition, OLS (not surprisingly) does poorly when used to model nonlinear distributions. For example, a sinusoidal or parabolic curve cannot be well modeled by a single straight line through the data.

Deaths

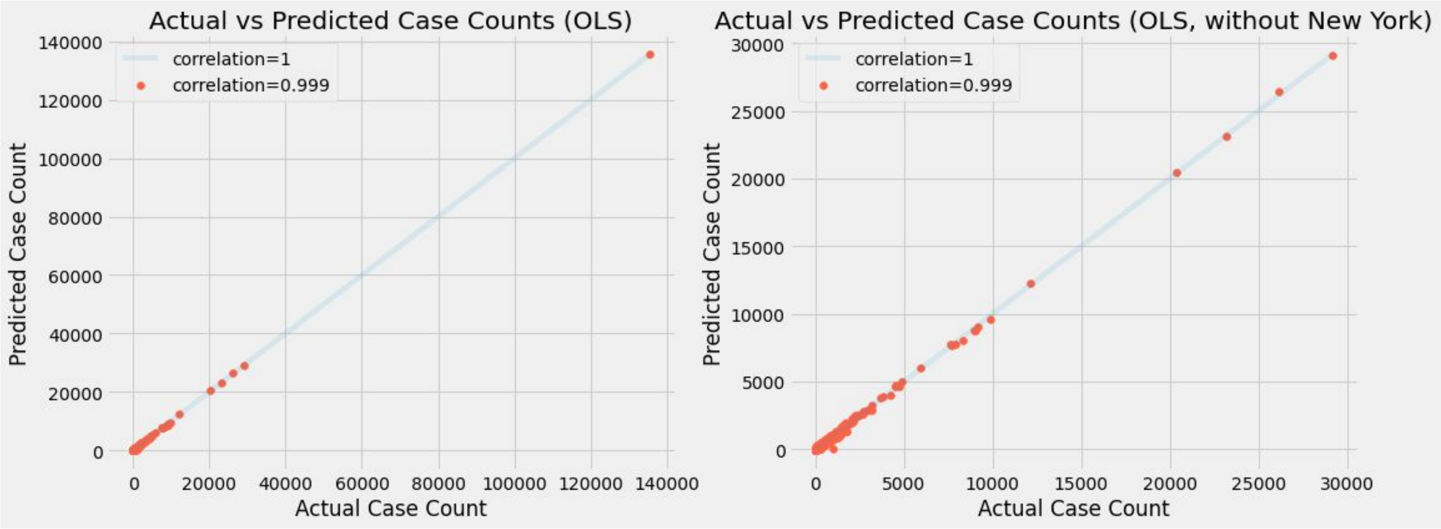
- The accuracy is: 0.9899189800137906
- The root mean squared error is: 5.6412767356529425
- The correlation between actual and predicted values is: 0.9949467221986263

In the plots below, the actual and predicted death count values are plotted. The maximum value corresponding to predictions for New York are left out in the plot on the right to see the underlying distributions for the majority of the points. The differences in these two plots are more apparent when plotting the cases counts in the next parts.



Cases

- The accuracy is: 0.9997677209700798
- The root mean squared error is: 51.71940799568102
- The correlation between actual and predicted values is: 0.9998838758179113



Ridge Regression Model

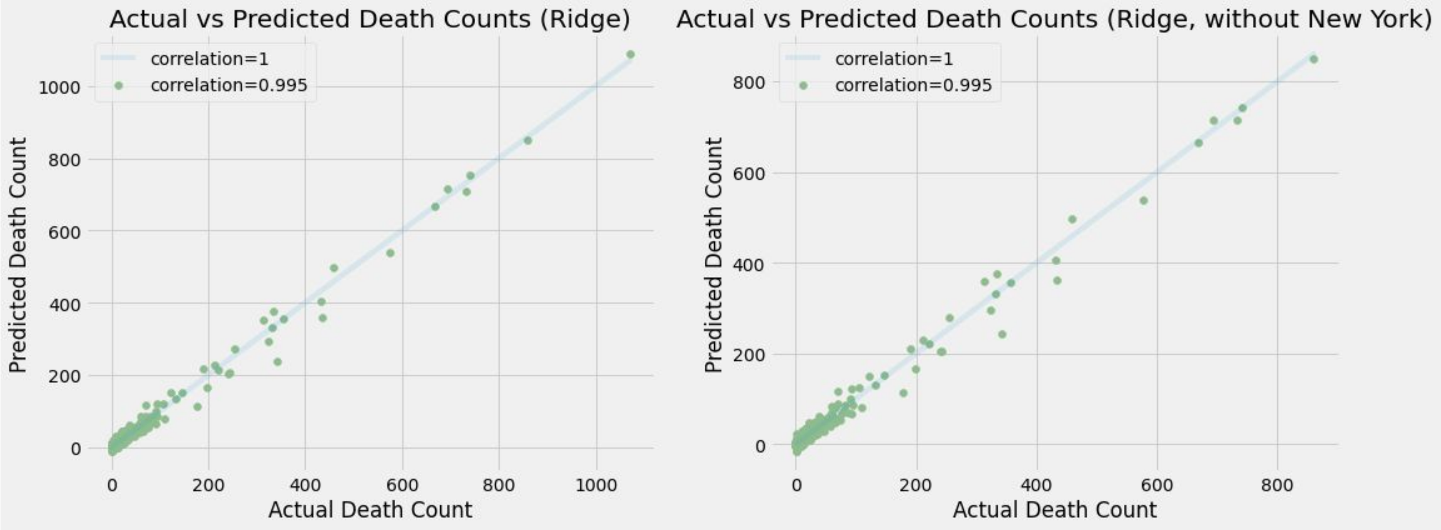
Why Ridge Regression?

Ridge regression is a type of squared loss regression used when the number of predictors exceeds the number of observations (eg. p>n) and when the model experiences multicollinearity. Since both p>n and multicollinearity are issues when using linear least squares regression, ridge regression would be used instead. Ridge regression works by using a shrinkage estimator that essentially would produce new estimates that are "shrunk" to the population's true parameters. It is a L2 "squared" regularization that adds a penalty equal to the squared magnitude of the coefficients. A tuning parameter λ would determine the strength of this penalty.

However, a disadvantage to ridge regression is the tradeoff of bias for variance. Unlike least squares, the model is biased since the coefficients are given different weights, ie. biased estimators. The constraints put on each of the estimators helps to shrink extreme variance and fluctuations; this sacrifices training accuracy for a model that is likely to generalize better. In other words, ridge regression strives to introduce enough bias that shrinks variance to make estimates closer to the true population values.

Deaths

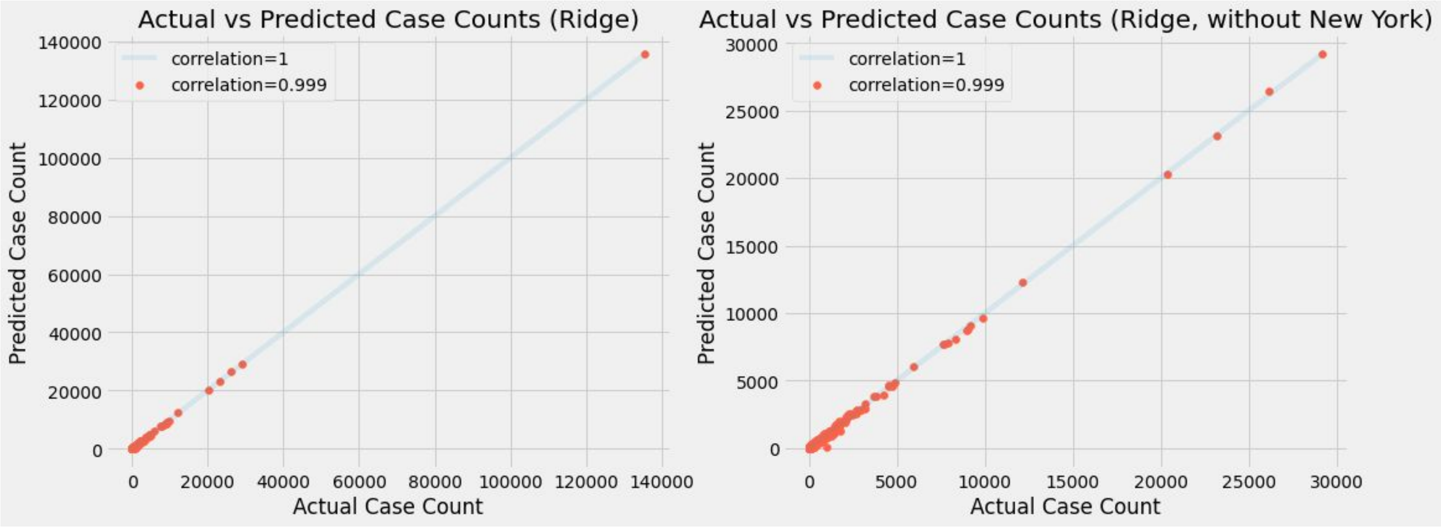
- The accuracy is: 0.9897601965203211
- The root mean squared error is: 5.685530292752235
- The correlation between actual and predicted values is: 0.9948669422511904



The results from Ridge Regression remain very similar to the Linear Regression model. As part of feature engineering, it would be beneficial to take a look at multicollinearity.

Cases

- The accuracy is: 0.9997650153455985
- The root mean squared error is: 52.0197540085709
- The correlation between actual and predicted values is: 0.9998825007725552



The Problem with Multicollinearity

There is a red line for y = -x because values should be correlated with themselves. However, any red or blue columns show there's a strong correlation/anticorrelation that requires more investigation. Since most of the values in the months of January and February show 0 confirmed deaths, there is whitespace in the features correlation heatmap.

Regression assumes that the feature parameters used are independent from one another.

LASSO Regression Model

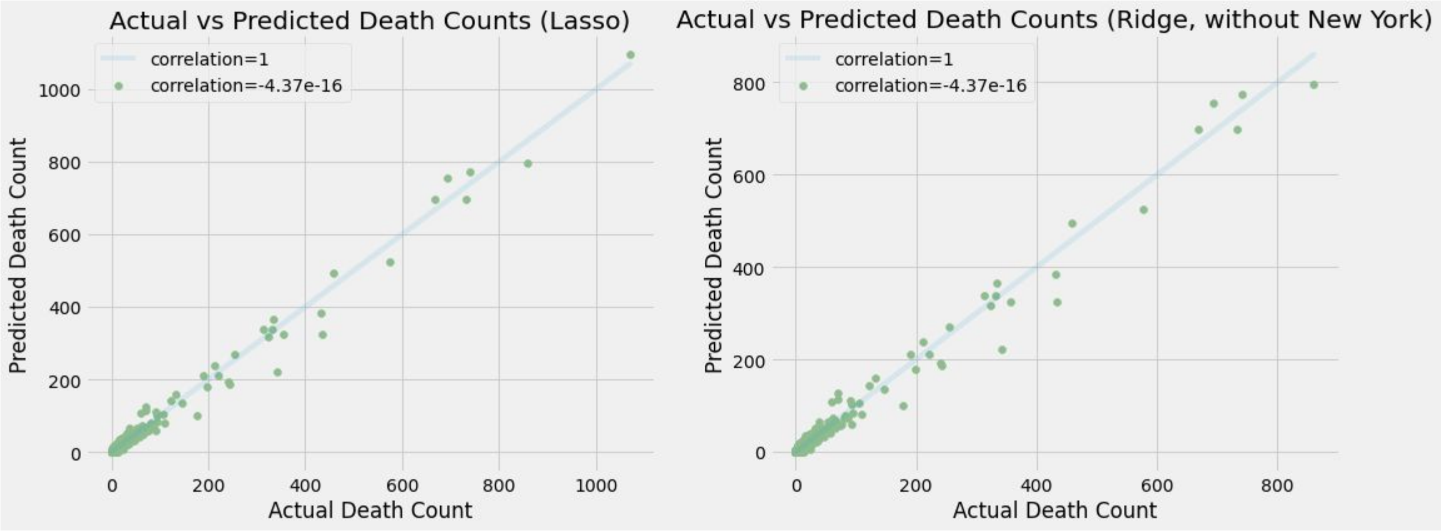
Why LASSO?

Lasso Regression (Least Absolute Shrinkage and Selection Operator) is similar to Ridge Regression in which it uses a shrinkage penalty as well. Instead of the L2 squared regularization, it uses the L1 regularization that takes the absolute value of the coefficient magnitudes. In this way, the model can be parsed by giving some coefficients penalties of 0, effectively removing this estimator from the model. Lasso regression would potentially include fewer features while still solving the least squares issue of multicollinearity. Lasso regression is a parametric method. The same disadvantage of biasness mentioned above for Ridge regression also applies to the Lasso method as well.

In this iteration of linear regression using Lasso, a for loop has been set up to find the best alpha value. By looping through values of alpha from 0-1, it was found using sklearn **.cross_val_score()** function that alpha=1 produced the highest score.

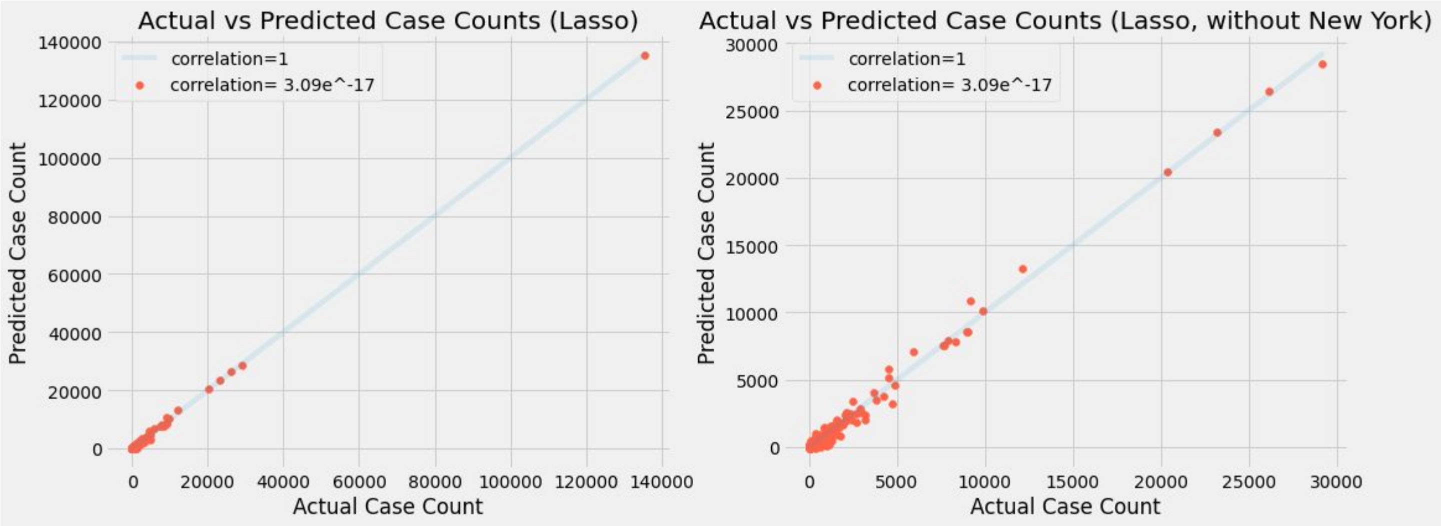
Deaths

- The accuracy is: 0.9856726585766261
- The root mean squared error is: 79.10518710431734
- The correlation between actual and predicted values is: -4.37165021751137e-16



Cases

- The accuracy is: 0.9987522881922346
- The root mean squared error is: 4798.780617376308
- The correlation between actual and predicted values is: 3.091760321740942e-17



Modeling Conclusions

Evaluation of Models Used and Numerical Results

Overall, each model (Linear, Ridge, Lasso) produced relatively accurate 1-week predictions using only population totals and cumulative deaths and cases data for each county. In particular, the baseline Linear model showed high accuracy scores of 98.9% for deaths predictions and 99.9% for cases predictions. The Ridge and Lasso models show similar results. Even without training on the most immediate data (since we excluded data one week before 4/18/20), these predictions still scored high accuracy and correlation marks. The root mean squared error for the Linear model on deaths predictions is about 5.64, a relatively small distance between actual and predicted death values.

Challenges to the Data

As noted before in the section on Data Cleaning and Standardization of this notebook, there was significant data preparation involved to create training, validation, and testing sets that would fit our analysis. Since our goal is to get early predictions for the number of Covid-19 cases and deaths in each county, it was imperative to only select the rows corresponding to counties. The datasets were littered with extraneous data from unassigned counties, cruise ships, and even correctional facilities. In addition to this, some data entries were also misreported and did not follow the cumulative deaths and cases trends. In one case, two counties (Dukes and Nantucket, Massachusetts) were even lumped together in the datasets. This caused the corresponding population count to be erroneously given a 0 population value. These considerations were taken in the data cleaning process.

Reflecting on Limitations

There are a couple limitations to these models. First, we are limited by the data we were given. Specifically, data collection plays a huge role in how our models will predict the spread of Covid-19. For example, testing in the United States is very limited. There are likely many more unconfirmed cases than those shown in our datasets. Also, some of the information provided such as the number of hospital beds and ventilators are very dynamic and will change everyday based on the number of cases and deaths. Therefore, as the number of deaths and cases increase, the number of beds and equipment do not remain constant. Some surprising discoveries we made were during data cleaning; there were some cruise ships and correctional facilities listed in the datasets while the majority of entries were counties. From the news, we expected that men are more likely to die from Covid-19, however, from examining the data we were not able to draw those same conclusions.

Potential Improvements to the Model

The models can be improved if we had a larger data set to train with. The current model is trained on data up to 4/18. However, data 4/18 after would be important to determining the effectiveness of sheltering in place and the use of face masks. It is also likely that our model is overtrained, perhaps in the future we can cross validate on a greater fold. Additionally, if the data set were larger it would be computationally expensive to cross-validate on both L1 and L2 regularization penalties. The model could also be improved by eliminating outliers that might be disproportionately skewing the data. Since we are limited by the amount of data we have, we could bootstrap the data to add more variation.

Additional Data

Of course, as new information is made available in the coming months, the datasets used in modeling should be updated. Since the spread of Covid-19 is very much still on-going as this Notebook is created, more data through the month of May would greatly aid in our models' predictive capabilities. In addition, to strengthen our analysis, more data on the demographics of deaths and cases would be another interesting question to look into. In particular, the distribution of age and gender could be another area that predictive models can be applied. Are men more likely to die from Covid-19 than women? What percentage of people in each age group are contracting the disease? These questions could be answered with additional information since our datasets only provide a total number of deaths and cases in each county.

Ethical Concerns

When using data-driven predictive models, researchers often look to increase the amount of data to train their models on to increase accuracy of their models. Even in our analysis, it is obvious that there was an imbalanced dataset, where there were only a handful of counties with high (>1000) deaths and high (>10,000) confirmed cases. Imbalanced datasets potentially lead to overfitting of the model, in which the model trains more closely to high prevalence cases. However, it is unethical for researchers to push for more data on the spread of a potentially fatal disease. Solutions to overcome the limited data could be to oversample or bootstrap to create more variability.

In addition, another ethical concern to these predictions arises from disproportionate funding of counties. If these predictive models are used to allocate resources for each county (eg. government aid, sending doctors, opening hospital space), this may create scenarios in which this model chooses which counties to save. For example, if the model predicts that the most cases will overwhelmingly be in New York, most of the government funding would be given to New York, effectively ignoring individuals in other counties.