Kristine Pau, Selena Weng

Looking at Covid-19 Death Trends in the U.S. as a Whole and a Special Feature at New York

Project Introduction

When we started the project, we were interested in how the mortality rate looked across the U.S. by states, regions, and provinces after adjusting for population size. The question that we were trying to answer, looking at data for all the U.S. states and provinces, was whether there was a relationship between states and provinces when it came to the number of cases and deaths, and how the trends looked once we adjusted percentage of deaths with the state's population size. We also created a linear regression model to predict the number of deaths in a state or province at a given time. After creating a few visualizations, we realized that compared to other states and provinces, New York was a covid-19 hotspot with over 200,000 confirmed cases and over 17,500 deaths by April 18th, 2020, so we decided to look into the relationships and patterns of NY counties with high numbers of confirmed COVID cases. The main questions we wanted to address with the NY counties were whether we could predict the total number of cases per country a week in advance. Along the same thread, we wanted to see underlying relationships that result in high or low numbers of total cases in these counties. We had 2 sets of prediction models, one that used numerical features and information about cases/day in each county and another that did not include case/day information. Each set used a regular linear, lasso, ridge and elastic net regression model. We concluded that the set of models that had case/day information did better. Within that set, the ridge model did best if we compare mean absolute error.

Data Summary and Exploratory Data Analysis

U.S. states: We started with creating a bar plot (Figure Ai) of the total confirmed cases ordered from lowest to greatest number of cases across all the states and provinces in the U.S. on April 18th, 2020. Then we did the same for the total number of deaths (Figure Aii). Next, we created another bar plot (Figure B) that looked at the percentage of deaths (deaths/population) in each state to adjust for a state's population. We then wanted to look at the distribution of deaths in U.S. states and provinces from January to mid-April (Figure C) with a line plot. From the plot, it is difficult to understand what is happening. Since multiple states shared the same color, distinguishing which line belonged to which state or province was almost impossible. As a result, we divided the data into 6 groups - regions and provinces and created line plots for the distribution of deaths and percentage of deaths for each group (Figure D). Lastly, we created a comparison model comparing the rmse of a range features using linear regression on our training data and when we cross validate (Figure E).

NY counties: We created a line graph that looked at the number of cases per NY county over the span of several weeks from Jan - Apr 18 (Figure F). We also created a heatmap that looked at the correlation between the total number of cases and all of the feature columns (Figure I). This also helped us look to see which columns had values that were nominal data, which would later prompt us to not use these columns for our prediction models. For example, we knew that county FIPS and federal guidelines values were not numerical data. We then created a correlation graph that looked at the different feature columns' correlation to the total number of cases in the NY (target) dataset (Figure H). Number of doctors per county corresponded to the highest correlation (Figure G). We also made a scatter plot looking at the counties with most, least (with at least 1 confirmed case) and no confirmed covid cases. We plotted the number of cases to doctors per county for these 3 respective groups of NY counties (Figure K). We then plotted a similar scatter plot that was based on density per square mile by the number of cases of the groups (Figure J).

Figure Ai and Aii

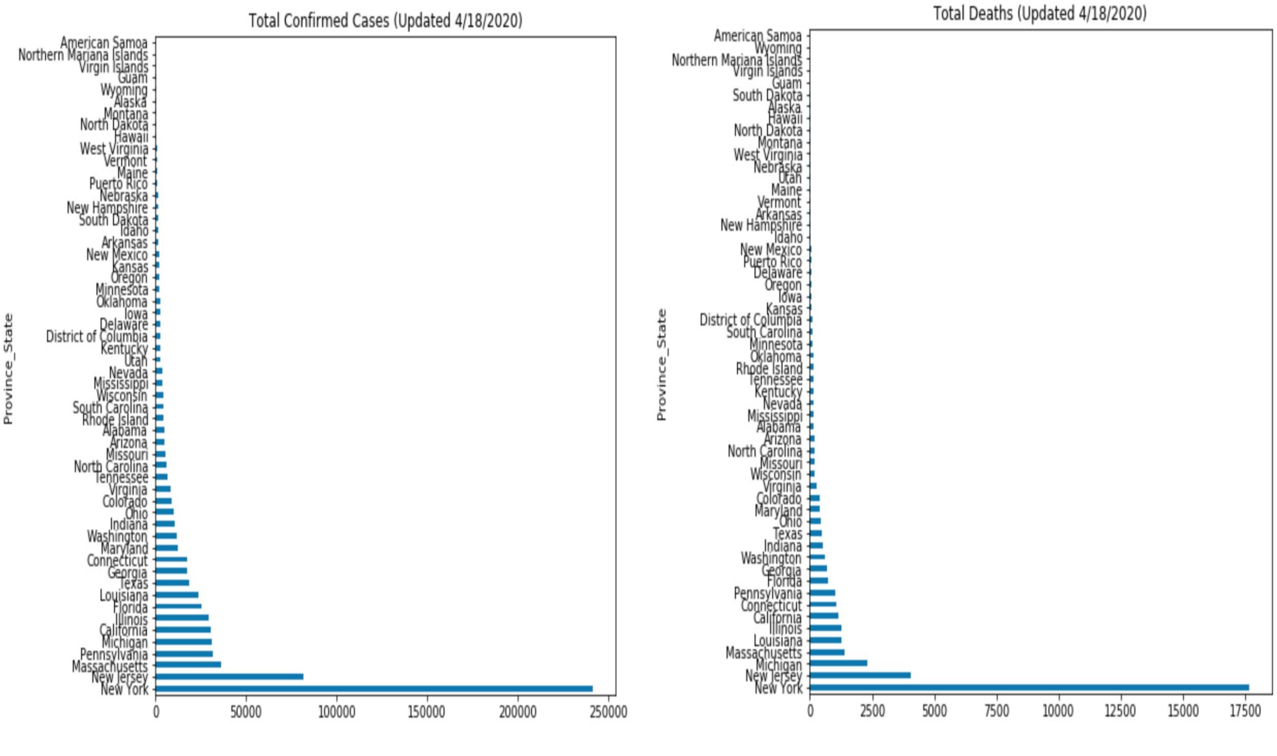


Figure B

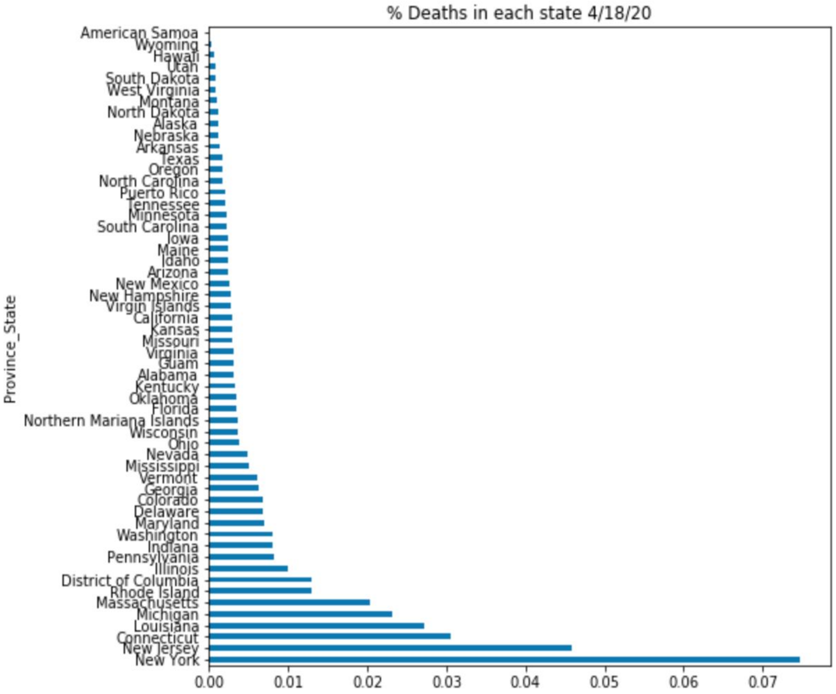


Figure C

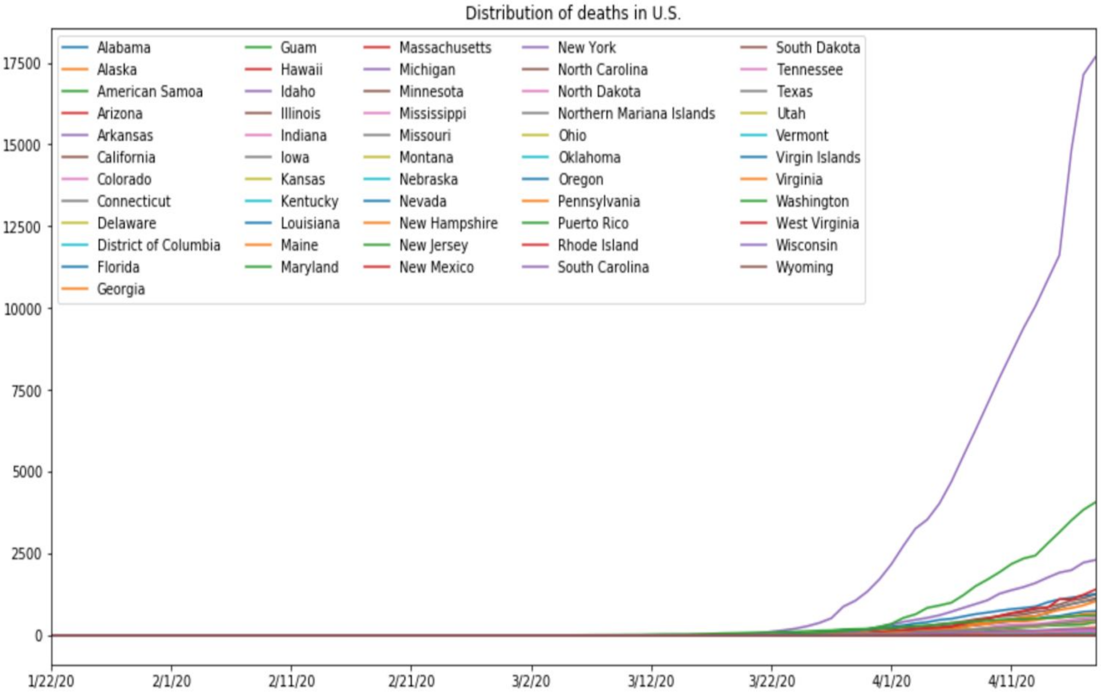


Figure D

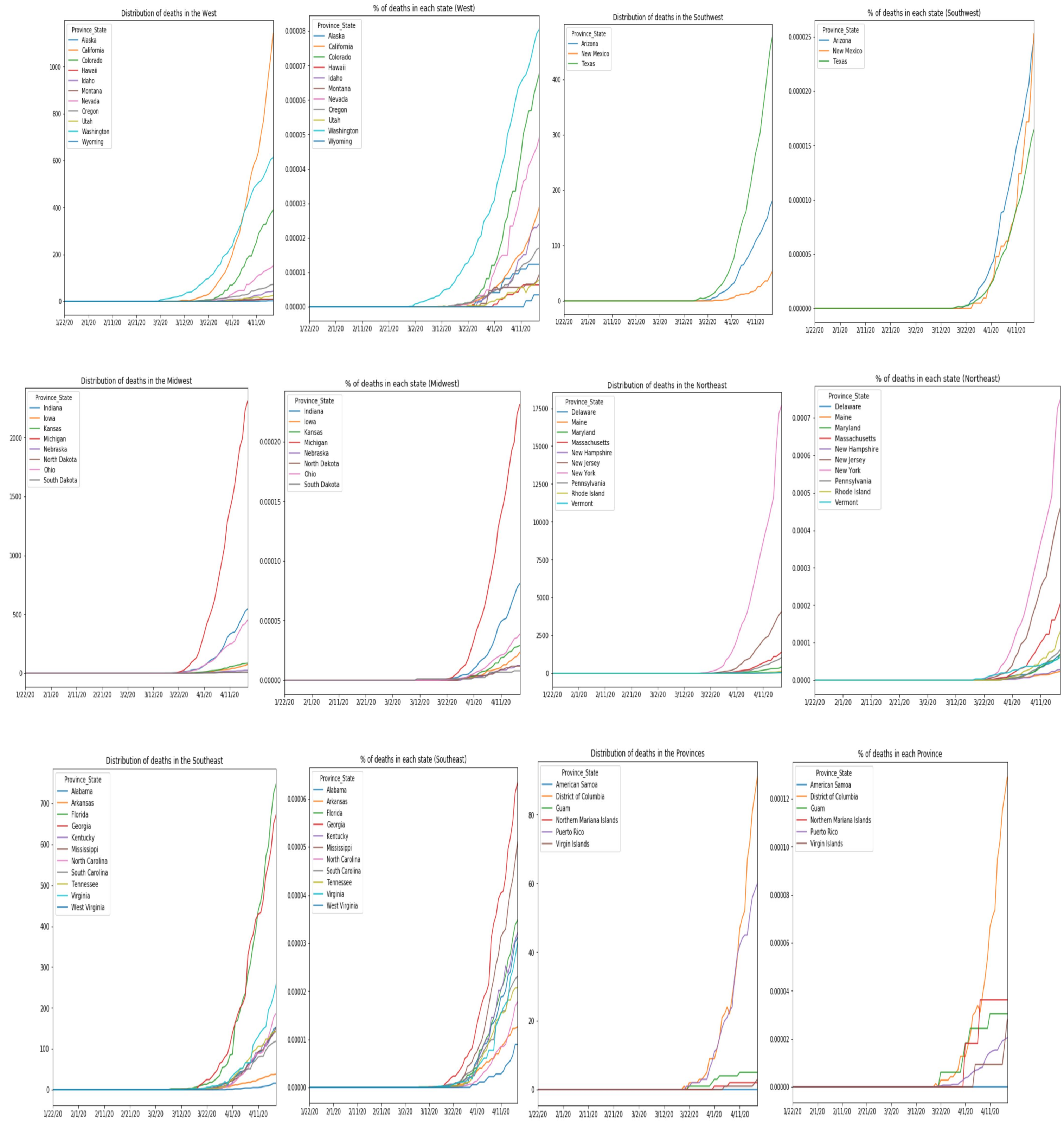

Distribution of deaths in the West
Province_State
Alaska
California
Colorado
Hawaii
Idaho
Montana
Nevada
Oregon
Utah
Washington
Wyoming
0
200
400
600
800
1000
1/22/20
2/1/20
2/11/20
2/21/20
3/2/20
3/12/20
3/22/20
4/1/20
4/11/20
% of deaths in each state (West)
Province_State
Alaska
California
Colorado
Hawaii
Idaho
Montana
Nevada
Oregon
Utah
Washington
Wyoming
0.00000
0.00001
0.00002
0.00003
0.00004
0.00005
0.00006
0.00007
0.00008
Distribution of deaths in the Southwest
Province_State
Arizona
New Mexico
Texas
0
100
200
300
400
% of deaths in each state (Southwest)
Province_State
Arizona
New Mexico
Texas
0.000000
0.000005
0.000010
0.000015
0.000020
0.000025
Distribution of deaths in the Midwest
Province_State
Indiana
Iowa
Kansas
Michigan
Nebraska
North Dakota
Ohio
South Dakota
0
500
1000
1500
2000
% of deaths in each state (Midwest)
Province_State
Indiana
Iowa
Kansas
Michigan
Nebraska
North Dakota
Ohio
South Dakota
0.00000
0.00005
0.00010
0.00015
0.00020
Distribution of deaths in the Northeast
Province_State
Delaware
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
0
2500
5000
7500
10000
12500
15000
17500
% of deaths in each state (Northeast)
Province_State
Delaware
Maine
Maryland
Massachusetts
New Hampshire
New Jersey
New York
Pennsylvania
Rhode Island
Vermont
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.0007
Distribution of deaths in the Southeast
Province_State
Alabama
Arkansas
Florida
Georgia
Kentucky
Mississippi
North Carolina
South Carolina
Tennessee
Virginia
West Virginia
0
100
200
300
400
500
600
700
% of deaths in each state (Southeast)
Province_State
Alabama
Arkansas
Florida
Georgia
Kentucky
Mississippi
North Carolina
South Carolina
Tennessee
Virginia
West Virginia
0.00000
0.00001
0.00002
0.00003
0.00004
0.00005
0.00006
Distribution of deaths in the Provinces
Province_State
American Samoa
District of Columbia
Guam
Northern Mariana Islands
Puerto Rico
Virgin Islands
0
20
40
60
80
% of deaths in each Province
Province_State
American Samoa
District of Columbia
Guam
Northern Mariana Islands
Puerto Rico
Virgin Islands
0.00000
0.00002
0.00004
0.00006
0.00008
0.00010
0.00012

Figure E

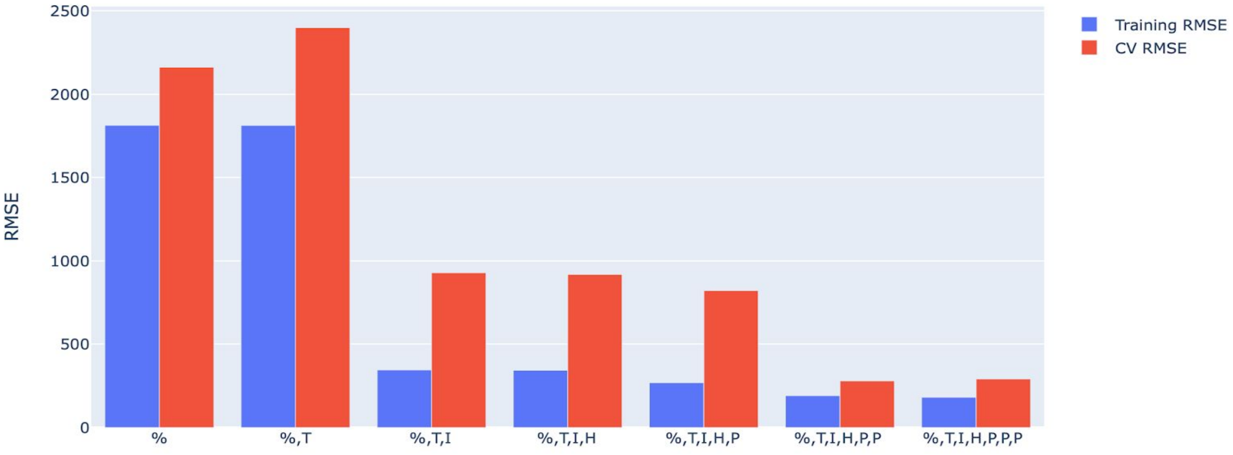


Figure F

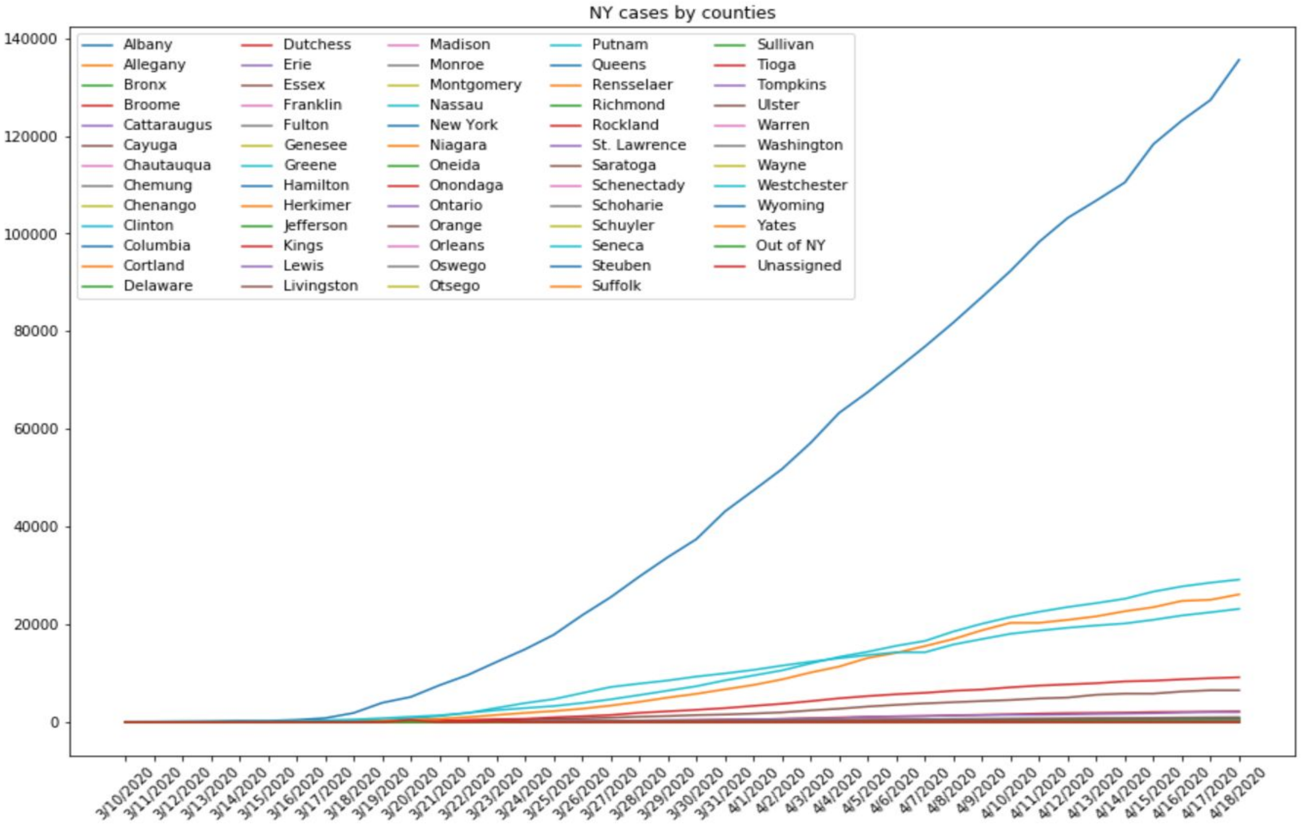
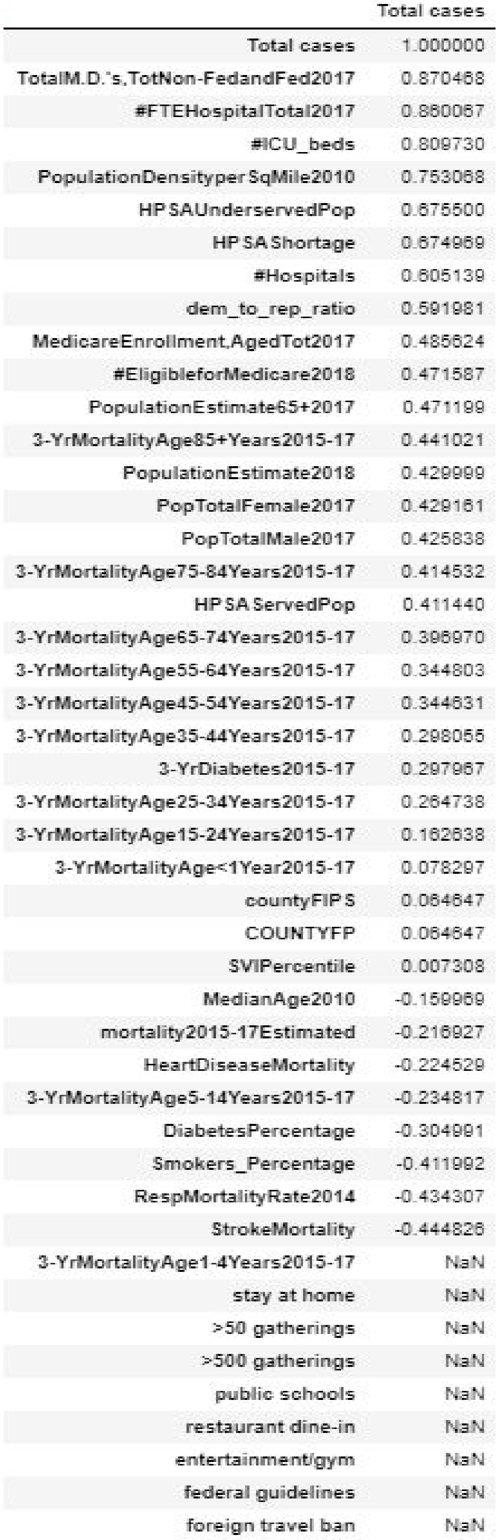


Figure G



	Total cases
Total cases	1.000000
TotalM.D.'s,TotNon-FedandFed2017	0.870468
#FTEHospitalTotal2017	0.860067
#ICU_beds	0.809730
PopulationDensityper SqMile2010	0.753068
HP SAUnderservedPop	0.675500
HP SA Shortage	0.674969
#Hospitals	0.605139
dem_to_rep_ratio	0.591981
MedicareEnrollment,AgedTot2017	0.485624
#EligibleforMedicare2018	0.471587
PopulationEstimate65+2017	0.471199
3-YrMortalityAge85+Years2015-17	0.441021
PopulationEstimate2018	0.429999
PopTotalFemale2017	0.429161
PopTotalMale2017	0.425838
3-YrMortalityAge75-84Years2015-17	0.414532
HP SA ServedPop	0.411440
3-YrMortalityAge65-74Years2015-17	0.396970
3-YrMortalityAge55-64Years2015-17	0.344803
3-YrMortalityAge45-54Years2015-17	0.344631
3-YrMortalityAge35-44Years2015-17	0.298055
3-YrDiabetes2015-17	0.297967
3-YrMortalityAge25-34Years2015-17	0.264738
3-YrMortalityAge15-24Years2015-17	0.162638
3-YrMortalityAge<1Year2015-17	0.078297
countyFIP S	0.064647
COUNTYFP	0.064647
SVIPercentile	0.007308
MedianAge2010	-0.159969
mortality2015-17Estimated	-0.216927
HeartDiseaseMortality	-0.224529
3-YrMortalityAge5-14Years2015-17	-0.234817
DiabetesPercentage	-0.304991
Smokers_Percentage	-0.411992
RespMortalityRate2014	-0.434307
StrokeMortality	-0.444826
3-YrMortalityAge1-4Years2015-17	NaN
stay at home	NaN
>50 gatherings	NaN
>500 gatherings	NaN
public schools	NaN
restaurant dine-in	NaN
entertainment/gym	NaN
federal guidelines	NaN
foreign travel ban	NaN

Figure H

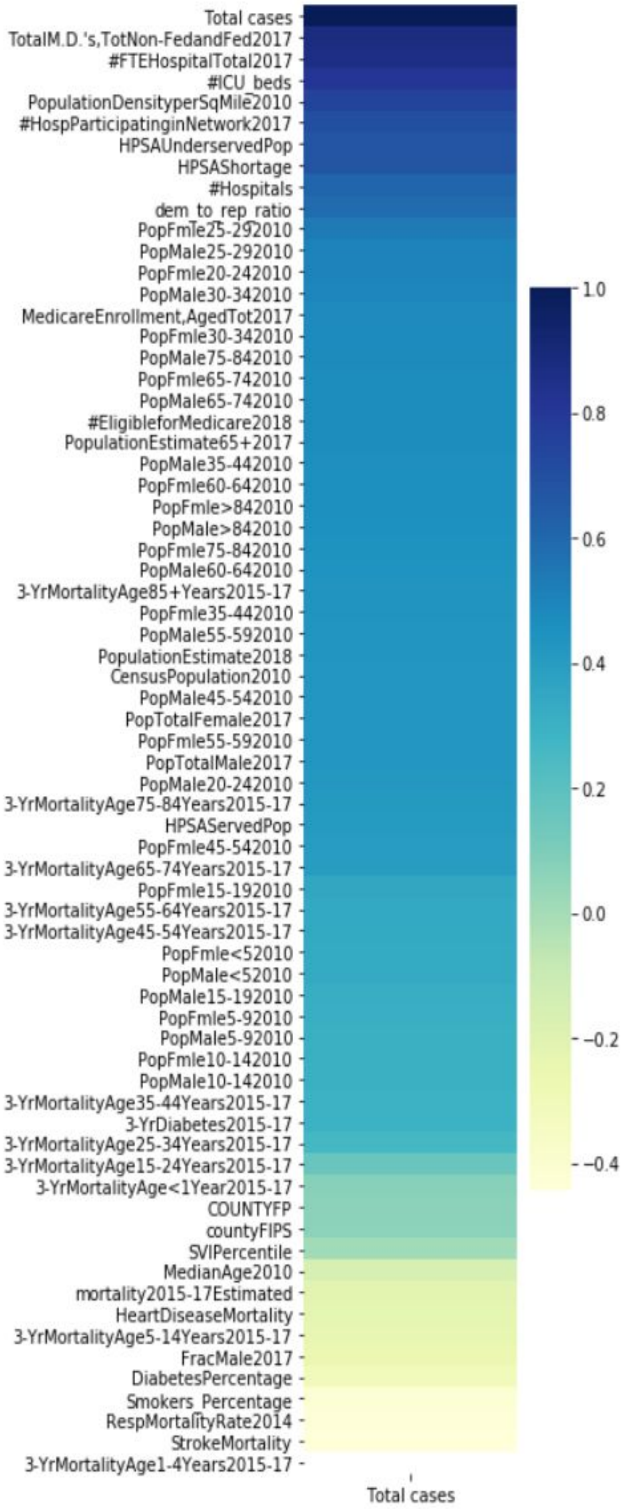


Figure I

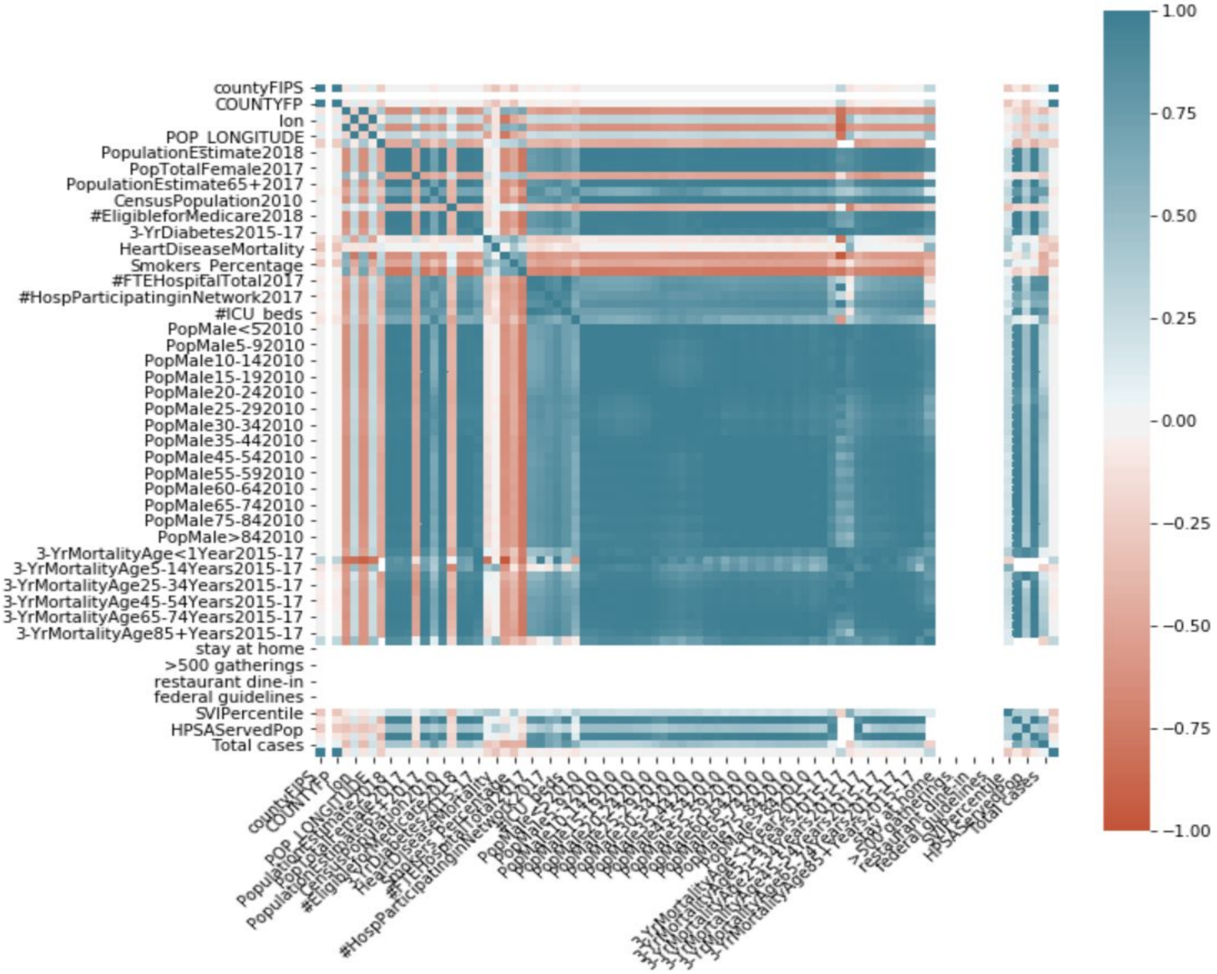


Figure J

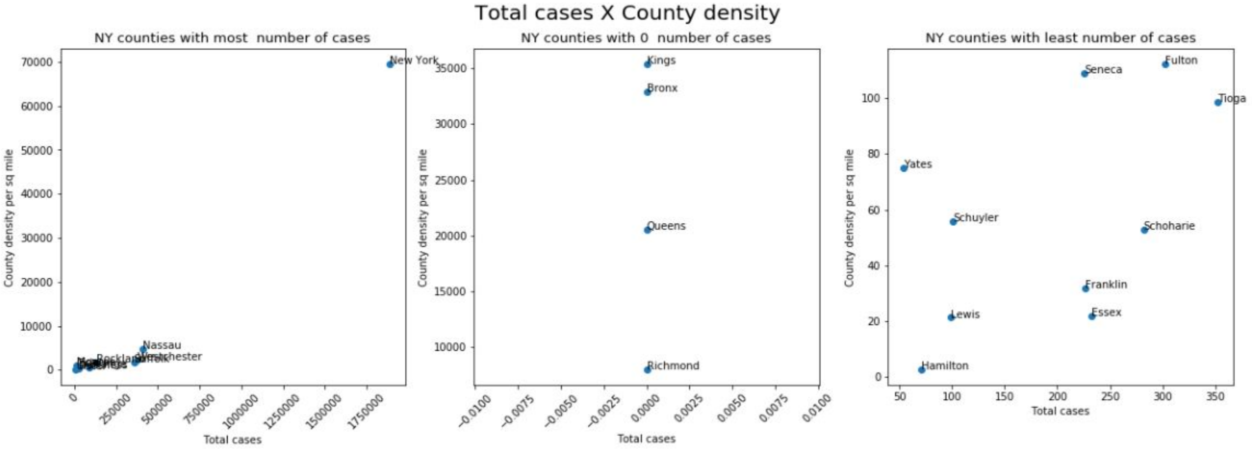


Figure K

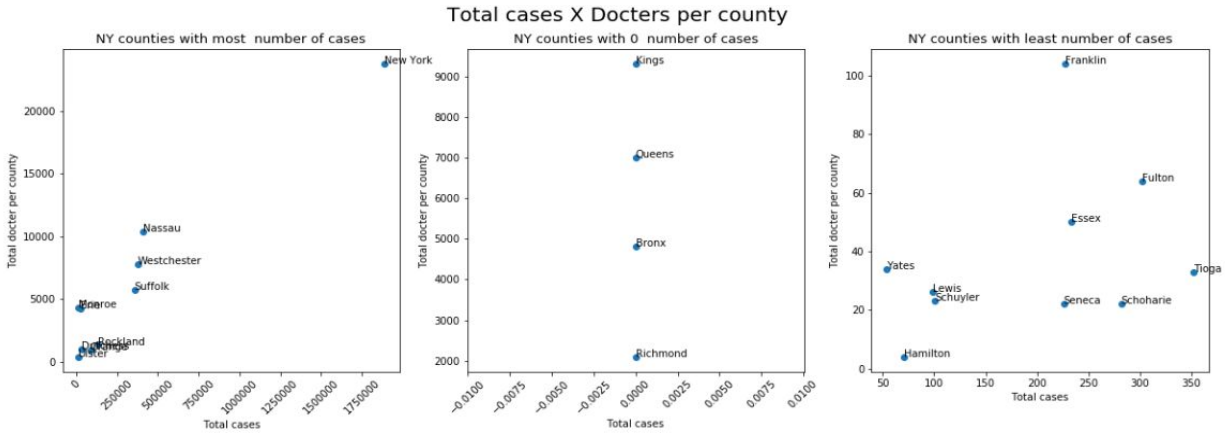
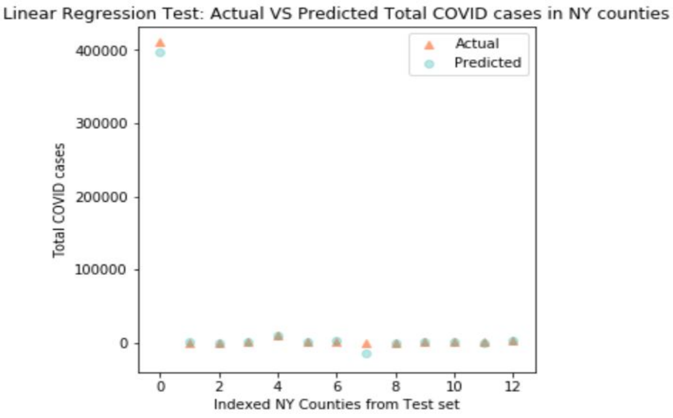
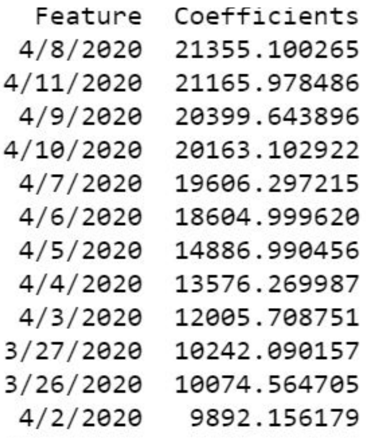
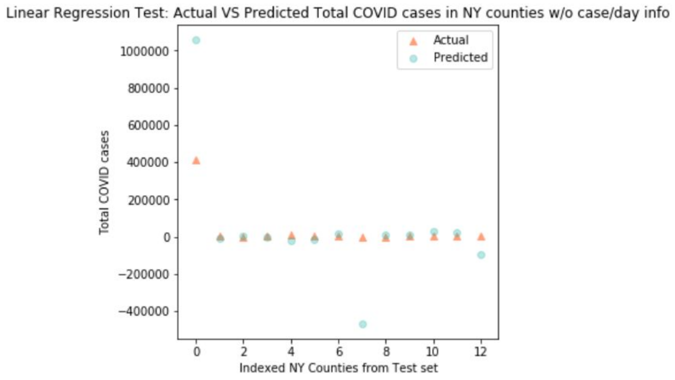


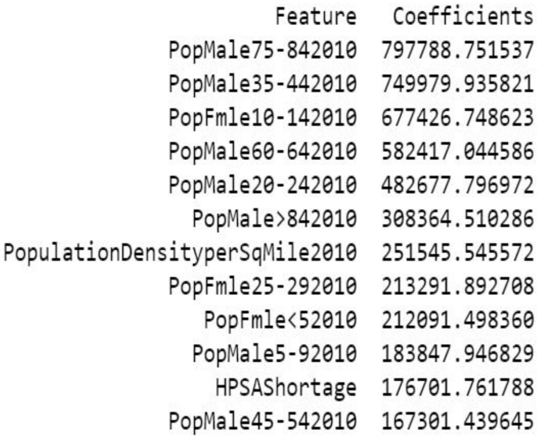
Figure L





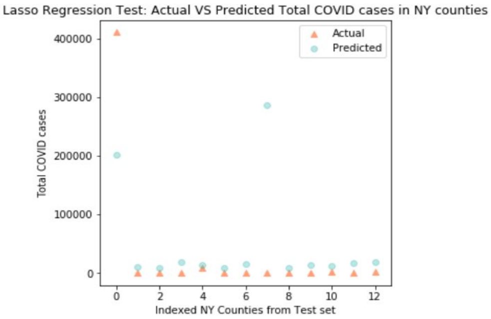
```
     Feature   Coefficients
   4/8/2020   21355.100265
  4/11/2020   21165.978486
   4/9/2020   20399.643896
  4/10/2020   20163.102922
   4/7/2020   19606.297215
   4/6/2020   18604.999620
   4/5/2020   14886.990456
   4/4/2020   13576.269987
   4/3/2020   12005.708751
  3/27/2020   10242.090157
  3/26/2020   10074.564705
   4/2/2020    9892.156179
```

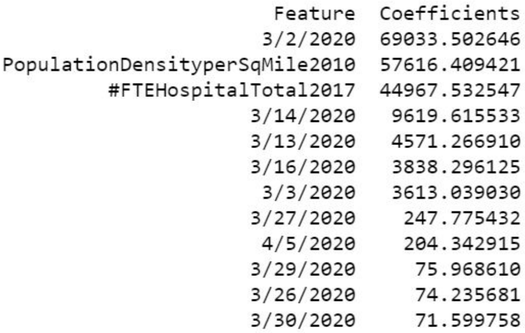


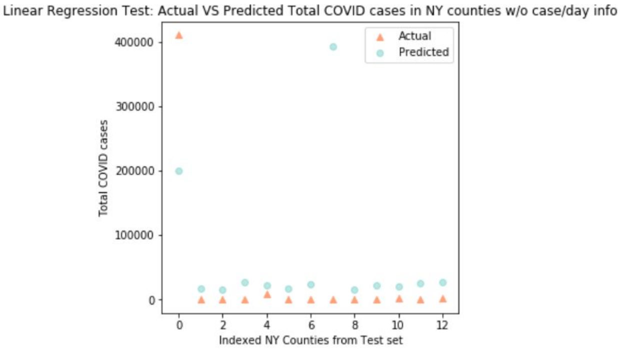


```
                           Feature    Coefficients
               PopMale75-842010   797788.751537
               PopMale35-442010   749979.935821
               PopFmle10-142010   677426.748623
               PopMale60-642010   582417.044586
               PopMale20-242010   482677.796972
                 PopMale>842010   308364.510286
PopulationDensityperSqMile2010   251545.545572
               PopFmle25-292010   213291.892708
                  PopFmle<52010   212091.498360
                PopMale5-92010   183847.946829
                    HPSAShortage   176701.761788
               PopMale45-542010   167301.439645
```

Figure M


Lasso Regression Test: Actual VS Predicted Total COVID cases in NY counties


```
                         Feature   Coefficients
                        3/2/2020   69033.502646
 PopulationDensityperSqMile2010    57616.409421
          #FTEHospitalTotal2017    44967.532547
                       3/14/2020    9619.615533
                       3/13/2020    4571.266910
                       3/16/2020    3838.296125
                        3/3/2020    3613.039030
                       3/27/2020     247.775432
                        4/5/2020     204.342915
                       3/29/2020      75.968610
                       3/26/2020      74.235681
                       3/30/2020      71.599758
```


Linear Regression Test: Actual VS Predicted Total COVID cases in NY counties w/o case/day info

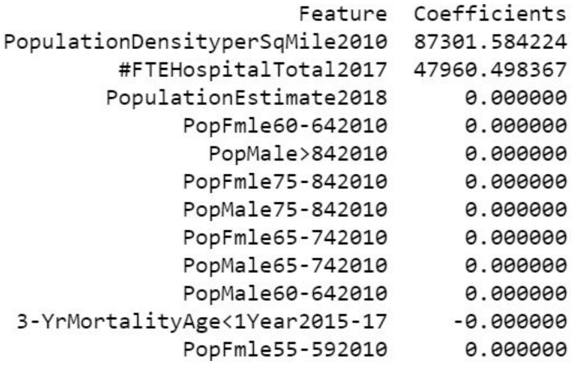
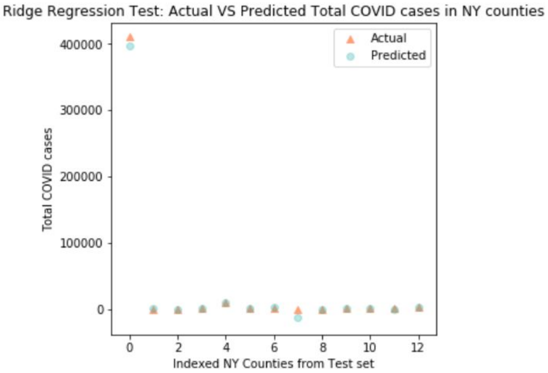
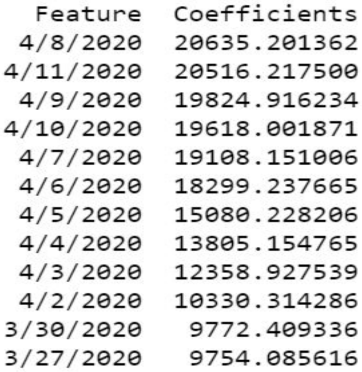
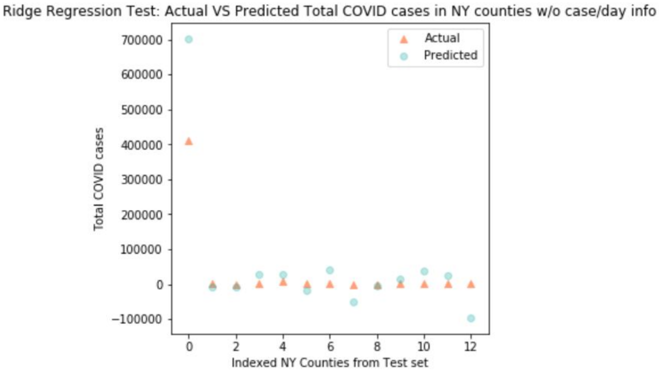

```
                         Feature   Coefficients
 PopulationDensityperSqMile2010    87301.584224
          #FTEHospitalTotal2017    47960.498367
       PopulationEstimate2018          0.000000
                PopFmle60-642010       0.000000
                 PopMale>842010        0.000000
                PopFmle75-842010       0.000000
                PopMale75-842010       0.000000
                PopFmle65-742010       0.000000
                PopMale65-742010       0.000000
                PopMale60-642010       0.000000
  3-YrMortalityAge<1Year2015-17       -0.000000
                PopFmle55-592010       0.000000
```

Figure N


Ridge Regression Test: Actual VS Predicted Total COVID cases in NY counties


```
  Feature   Coefficients
 4/8/2020   20635.201362
4/11/2020   20516.217500
 4/9/2020   19824.916234
4/10/2020   19618.001871
 4/7/2020   19108.151006
 4/6/2020   18299.237665
 4/5/2020   15080.228206
 4/4/2020   13805.154765
 4/3/2020   12358.927539
 4/2/2020   10330.314286
3/30/2020    9772.409336
3/27/2020    9754.085616
```


Ridge Regression Test: Actual VS Predicted Total COVID cases in NY counties w/o case/day info

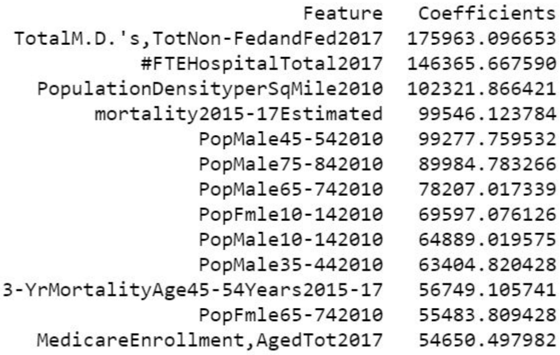
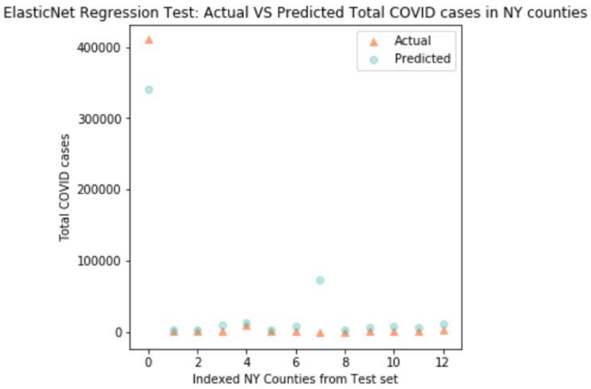
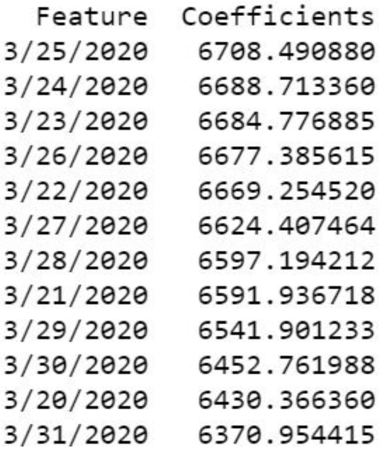
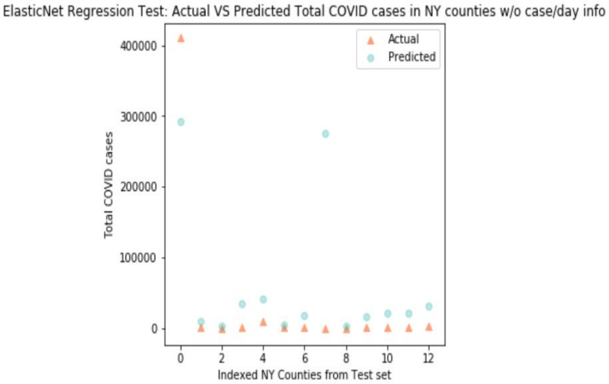
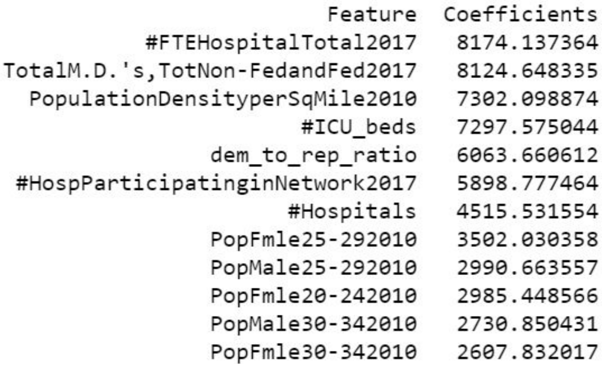

```
                              Feature   Coefficients
 TotalM.D.'s,TotNon-FedandFed2017     175963.096653
              #FTEHospitalTotal2017   146365.667590
      PopulationDensityperSqMile2010  102321.866421
          mortality2015-17Estimated    99546.123784
                    PopMale45-542010    99277.759532
                    PopMale75-842010    89984.783266
                    PopMale65-742010    78207.017339
                    PopFmle10-142010    69597.076126
                    PopMale10-142010    64889.019575
                    PopMale35-442010    63404.820428
  3-YrMortalityAge45-54Years2015-17    56749.105741
                    PopFmle65-742010    55483.809428
        MedicareEnrollment,AgedTot2017    54650.497982
```

Figure O


ElasticNet Regression Test: Actual VS Predicted Total COVID cases in NY counties


```
 Feature   Coefficients
3/25/2020   6708.490880
3/24/2020   6688.713360
3/23/2020   6684.776885
3/26/2020   6677.385615
3/22/2020   6669.254520
3/27/2020   6624.407464
3/28/2020   6597.194212
3/21/2020   6591.936718
3/29/2020   6541.901233
3/30/2020   6452.761988
3/20/2020   6430.366360
3/31/2020   6370.954415
```


ElasticNet Regression Test: Actual VS Predicted Total COVID cases in NY counties w/o case/day info


```
                          Feature   Coefficients
            #FTEHospitalTotal2017    8174.137364
 TotalM.D.'s,TotNon-FedandFed2017    8124.648335
   PopulationDensityperSqMile2010    7302.098874
                       #ICU_beds     7297.575044
                dem_to_rep_ratio     6063.660612
#HospParticipatinginNetwork2017      5898.777464
                      #Hospitals     4515.531554
                PopFmle25-292010     3502.030358
                PopMale25-292010     2990.663557
                PopFmle20-242010     2985.448566
                PopMale30-342010     2730.850431
                PopFmle30-342010     2607.832017
```

Data Cleaning

U.S. states: We first checked to see which rows or columns had null values and made sure that all the states and provinces belonged in the U.S by checking the rows in the country region column. We then dropped columns that were redundant or unnecessary like latitude, longitude, country region, uid, code3, and admin2 and rows that contained data for cruise ships. We also merged two of the datasets together so that the new dataset would contain columns for total deaths and population, before creating a new column where we divided the deaths column by the population column for Figures C, D, and E. Since some states and provinces had 0 deaths, we filled the null values in this new column with 0s. For the line plots (Figure D), since multiple states shared the same color, distinguishing which line belonged to which state or province was almost impossible, so we divided the data into 6 groups - regions and provinces and created line plots for the distribution of deaths and percentage of deaths for each group.

NY counties: We made sure to clean the datasets to only keep rows that were a part of the US and were part of NY state. When we initially made the correlation heatmap, we saw that there were some columns that had 0 correlation, and upon further review, noticed that the columns had nominal values. For certain parts of the process, such as when we made the predictive model, we had to drop columns that were not numerical. Other times, we would keep these nominal columns as primary keys to help with the data merging process. For the first visualization, (Figure F), I had to transpose the data so that I had the columns as NY counties and the rows as the different dates showing the number of confirmed cases as opposed to how it originally had the dates as columns and counties as rows. We also noticed that there were some columns that had null values. For most of the census related data about the population percentage of certain health groups, we filled in the null values with the mean of the column. For the model, we also cleaned the data by standardizing all the feature columns before training out models.

Prediction Modeling + Methods

We decided to create two models - one for predicting deaths in a country at any given time and one for predicting the total number of cases in New York counties one week into the future.

U.S. states: We tried to create a model that would predict the total number of deaths in a state at a given time based on data from the feature columns First, we split the data into a test and train set and used a comparison model to see which distribution of features would give the lowest training rmse and also for the cross validation rmse, ensuring that we did not overfit the model and had the best combination of features. After trying many different combinations of features, the final features that we decided to use for this model were percent confirmed, testing rate, incident rate, hospitalization rate, people tested, and people hospitalized because the cross validation rmse had the smallest error using these features.

NY counties: We tried to create a model that would predict the total number of cases in the coming week based on feature columns and cases per day data collected since Jan 23 2020 (the start of the recorded data given). In other words, we would have the features such as population, population density, number of ICU's as well as number of cases from 1/23/2020-4/11/2020 to predict total number of cases by 4/18/2020. We created a test/train split and first created a regular linear regression model (Figure L). We then tried to improve it by creating regularized linear models :lasso, ridge and elastic net (Figures M,N,O respectively). We believe that it was important to also regularize the models given the number of columns we were using. We wanted to add a penalty term to tackle the issue of potentially overfitting. The 3 types were all tested because they all have its strengths when it comes to how the penalty term is used. Prior to making these models, we did a 5-fold CV to find the optimal alpha values for each respective model. We had to play around with different sets with varying ranges of alpha values to find the optimal alphas. We then fit our new models with the optimal alpha values and then finally looked at the mean absolute error (MAE) of the testing set.

Interpretation and Conclusions

U.S. states: From the bar plots (Figure Ai and Aii), we observed that states with more confirmed cases also had a higher number of deaths, so deaths and confirmed cases most likely have a positive correlation. Out of all the states, New York had the most confirmed cases and number of deaths - almost triple the number of cases and deaths of New Jersey - the second state with the most cases and deaths. Even after accounting for the state's population size, New York's percentage of deaths was still higher than all the other states and provinces in the U.S. One of the challenges that we faced was how to use visualizations to show the distribution of deaths in the U.S. from January to April. We decided to separate the data by regions and provinces, and use line plots to show the distributions. We observed that in the West coast, although California has the highest number of deaths, after adjusting for the state population size, Washington and Colorado's percentage of deaths was much higher than California's. The Northeast region had the most deaths overall. For all the regions and provinces, there is an exponential increase in deaths around mid-March and as of April 18th, most of the states have not seen a decrease in the number of deaths. What was surprising was that even though the number of deaths for many countries were well above 1,000, this only accounted for less than a fraction of a percent of the state's population. Even though this is a low number, we need to take into account that not the whole population has been infected or developed immunity for covid-19 and that some states went into lockdown a lot earlier than others, preventing the spread of covid-19 in the population. Our analysis was limited because we didn't have information on the demographics of the people who died from covid-19, which could further suggest whether a particular demographic was more likely to die from the virus, and help to explain why certain states or provinces have more deaths than others.

NY counties: The main assumption we made was that the dataset accurately and consistently showed the number of new cases per day in each county; we assumed there would be no random backlogs in cases per day when recording. When looking at the NY counties dataset, we first wanted to see if it would be possible for us to project the total number of cases in the NY counties a week in advance. When we did the initial EDA of the number of total cases (by 4/18/2020) in NY counties, we definitely saw a skew in the number of cases in NYC. Knowing that NYC has a large population density we wanted to see if there was indeed a high correlation with the number of cases. Interestingly, of all the features (not including data about cases/days), we found that the number of doctors/county had the highest correlation with total number of cases. This is interesting because we could interpret this as: confirmed cases locations would be based on which county hospital the confirmed person was at/tested. This is a slight nuance from our initial logic that people in NYC have higher rates of contracting the virus because they are living/from NYC.

We made **2 sets** of models: one that had cases/day as its features (top row of Figures L,M,N,O) and one that didn't (bottom row of Figures L,M,N,O). The former model was used mainly to address the issue of predicting a week in advance. The latter model was more to see if it was possible to predict total cases by a certain date without any other previous information about the number of cases whatsoever. From our predictions models that used case/day information, we noticed that those columns would have the highest coefficients regardless of whether we used a simple linear regression or regularized regression model. The model performed relatively well in our small data set of just NY counties. For this set of models, the ridge regression model just beat out the linear regression model if we compare the MAE. The worst performer was the lasso regression. In our 2nd set of models that did not use case/day features, the MAE was significantly higher, although from the scatter plots of actual VS predicted total cases, it didn't look too bad which was pretty surprising. A big concern when creating our models initially was we weren't sure if using cases/day data (from 1/22-4/11) used to predict the total number of cases by 4/18 would be giving away too much information to actually call it a prediction model. The second set of models that didn't rely on case/day features had a lot more emphasis put on total hospitals/doctors per county and population density rather than # cases/day like in the other models. Surprisingly, in both sets of models, health factors such as percentage smokers/diabetics had little to no major effects on the model. However, for the 2nd set of prediction models, the order from best to worst was Elastic Net>Ridge>Lasso>Normal Linear if we were to compare the lowest MAE. I think the reason why the regularized models did better for models that didn't have case/day information was that there was more of a need to penalize features that didn't help the model.

Again, it's interesting that the amount of medical capabilities AND population density (in certain models) dictated the number of cases in our prediction. We had always assumed that just density of population would be the main attributing factor in seeing which areas had most cases. Our EDA and models draw us to a different potential story/explanation--where cases are largely confirmed (generally areas with strong medical capabilities) may not represent where the patients are from.

However, it is important to note that there could be ethical dilemmas we face with this possibile narrative. This model shouldn't be used solely to make policy on addressing this pandemic. We need to look at a multitude of lenses when making decisions and not focus purely at the numbers. There may very well be many underlying inequalities and disparities in the number of confirmed cases that are not addressed in our simple prediction models that need to be considered when making decisions. For example, a certain county could have very few confirmed cases, but we can't boldly assume that people there wouldn't equally be at risk. Maybe people in those counties don't have access to medical care and aren't getting tested to confirm whether or not they have contracted the virus. Simply looking at our model, that county would look to be in good shape when it's not.

Lastly, we definitely had limitations ranging from the level of detail we had from the datasets to the amount of null values we had in columns. Additionally, there are aspects that could be better improved such as further fine tuning the hyperparameters, adding more features and dealing with more precise and up-to date data. Some other steps that could be implemented in the future would be to use logistic regression as well.