COVID-19 Case Study

Alyssa Frantz Angela Chen Lori Sun

Due before midnight, Feb 27

1 COVID-19 Case Study Background

The outbreak of the novel Corona virus disease 2019 (COVID-19) was declared a public health emergency of international concern by the World Health Organization (WHO) on January 30, 2020. Upwards of 112 million cases have been confirmed worldwide, with nearly 2.5 million associated deaths. Within the US alone, there have been over 500,000 deaths and upwards of 28 million cases reported. Governments around the world have implemented and suggested a number of policies to lessen the spread of the pandemic, including mask-wearing requirements, travel restrictions, business and school closures, and even stay-at-home orders. The global pandemic has impacted the lives of individuals in countless ways, and though many countries have begun vaccinating individuals, the long-term impact of the virus remains unclear.

The impact of COVID-19 on a given segment of the population appears to vary drastically based on the socioeconomic characteristics of the segment. In particular, differing rates of infection and fatalities have been reported among different racial groups, age groups, and socioeconomic groups. One of the most important metrics for determining the impact of the pandemic is the death rate, which is the proportion of people within the total population that die due to the the disease.

We assemble this dataset for our research with the goal to investigate the effectiveness of lockdown on flattening the COVID curve. We provide a portion of the cleaned dataset for this case study.

There are two main goals for this case study.

1. We show the dynamic evolvement of COVID cases and COVID-related death at state level.
2. We try to figure out what county-level demographic and policy interventions are associated with mortality rate in the US. We try to construct models to find possible factors related to county-level COVID-19 mortality rates.
3. This is a rather complex project. With our team's help we have made your job easier.
4. Hide all unnecessary lengthy R-output. Keep your write up neat, readable.

Remark1: The data and the statistics reported here were collected before February of 2021.

Remark 2: A group of RAs spent tremendous amount of time working together to assemble the data. It requires data wrangling skills.

Remark 3: Please keep track with the most updated version of this write-up.

2 Data Summary

The data comes from several different sources:

1. County-level infection and fatality data - This dataset gives daily cumulative numbers on infection and fatality for each county.
   - NYC data
2. County-level socioeconomic data - The following are the four relevant datasets from this site.
   i. Income - Poverty level and household income.
   ii. Jobs - Employment type, rate, and change.
   iii. People - Population size, density, education level, race, age, household size, and migration rates.
   iv. County Classifications - Type of county (rural or urban on a rural-urban continuum scale).
3. Intervention Policy Data - This dataset is a manually compiled list of the dates that interventions/lockdown policies were implemented and lifted at the county level.

3 EDA

In this case study, we use the following three nearly cleaned data:

- **covid_county.csv**: County-level socialeconomic information that combines the above-mentioned 4 datasets: Income (Poverty level and household income), Jobs (Employment type, rate, and change), People (Population size, density, education level, race, age, household size, and migration rates), County Classifications
- **covid_rates.csv**: Daily cumulative numbers on infection and fatality for each county
- **covid_intervention.csv**: County-level lockdown intervention.

Among all data, the unique identifier of county is `FIPS`.

The cleaning procedure is attached in `Appendix 2: Data cleaning`.

3.1 Understand the data

The detailed description of variables is in `Appendix 1: Data description`. Brief summaries of the `covid_rates` and `county_data` datasets are below.

Infection and fatality data

- 3108 unique county FIPS codes
- Dates ranging from 1/21/2020 to 2/20/2021
- Data from 48 continental states and Washington DC (Alaska and Hawaii not included)
- Cumulative COVID-19 infections range from 0 to 1.2M (Los Angeles County on 2/20/2021)
  - Median: 366 infections
  - Mean: 2933 infections
- Cumulative COVID-19 deaths range from 0 to 19,793 (Los Angeles County on 2/20/2021)
  - Median: 7 deaths
  - Mean: 66 deaths

Socioeconomic demographics (please see Appendix for detailed data description)

- 3278 unique county FIPS codes

Income

- Per capita income in the past 12 months ranged from 5,974 to 72,832, with mean 26,720
- Median household income ranged from 25,385 to 140,382, with mean 52,945.
- Poverty rate for children ages 0-17 in 2018 ranged from 2.5% to 68.3%, with mean 21.0%

Jobs

- Unemployment rates in 2019 ranged from 0.70 to 19.30, with a mean of 4.14.
- The percent of the civilian labor force 16 and over employed in services ranged from 8.3% to 81.6%, with a mean of 43.1%
- Other industries included in the data:
  - FIRE (finance and insurance, real estate and rental and leasing)
  - Construction

  - Transportation, warehousing and utilities
  - Mining, quarrying, oil and gas extraction
  - Wholesale and retail trade
  - Information services
  - Agriculture, forestry, fishing and hunting
  - Manufacturing
  - Public administration

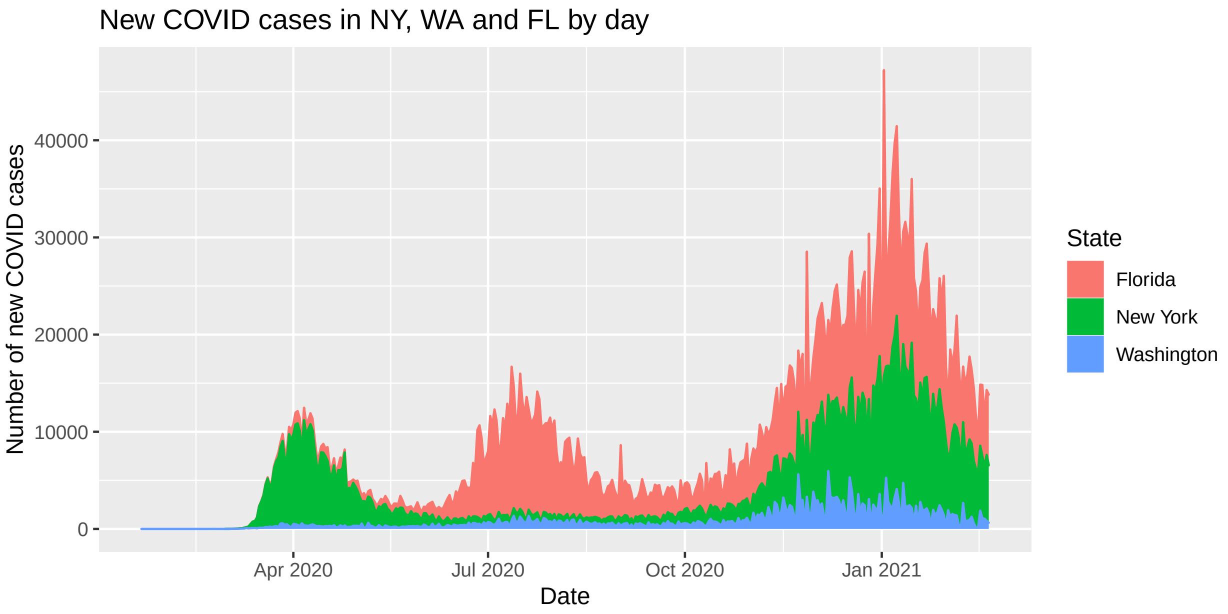
People

- Average household size ranged from 1.34 to 4.97, with a mean of 2.53
- Percent of non-English speaking households of total households ranged from 0 to 89.5%, with a mean of 3.6%
- The percent of the population 65 or older in 2010 ranged from 3.5% to 43.4%, with a mean of 15.8%

3.2 COVID case trend

It is crucial to decide the right granularity for visualization and analysis. We will compare daily vs weekly total new cases by state and we will see it is hard to interpret daily report.

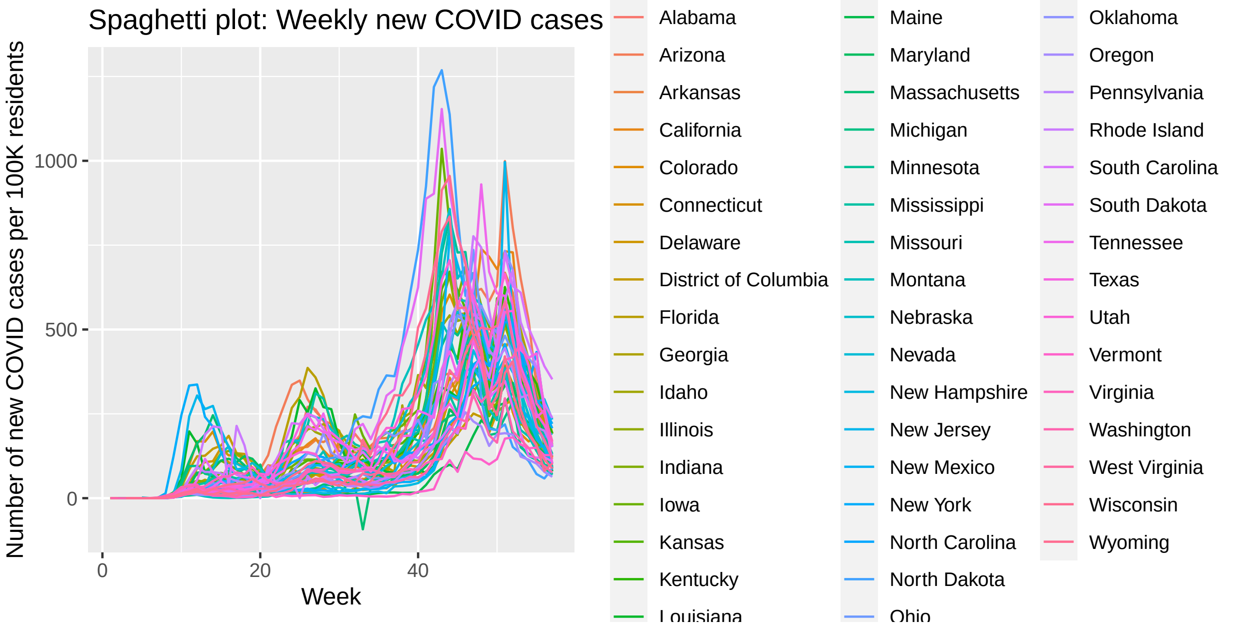
i) Plot **new** COVID cases in NY, WA and FL by state and by day. Any irregular pattern? What is the biggest problem of using single day data?



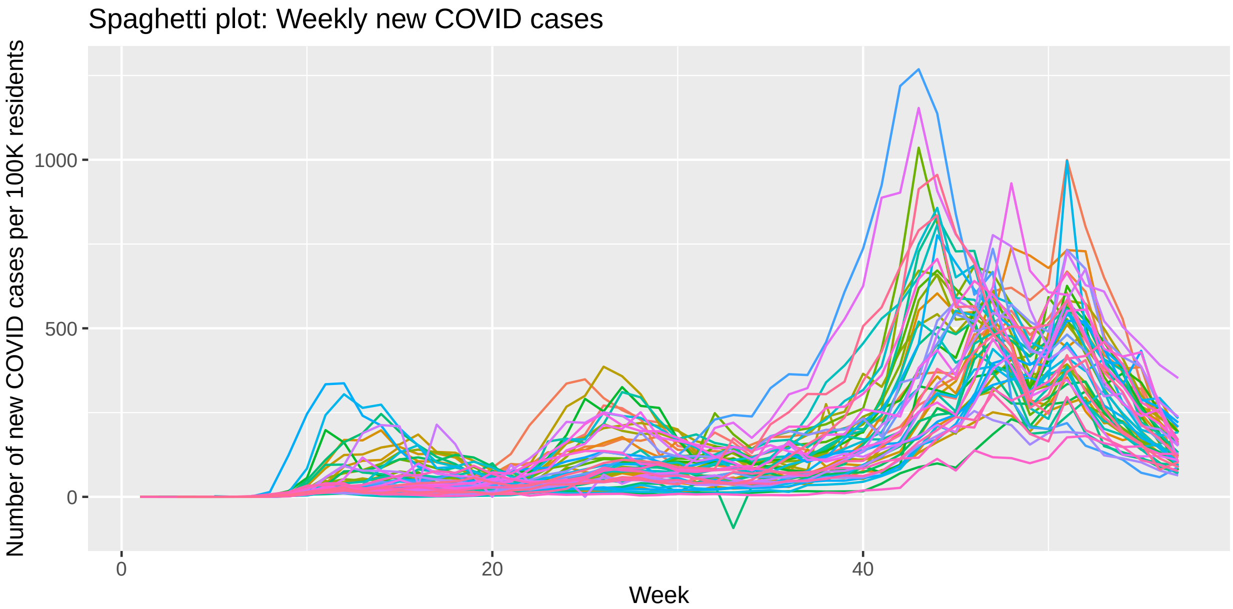
When plotting new case counts by day, we do not notice major irregularities or unusual patterns in the data. Of note, there are two small peaks before a large peak in cases around Jan 2021; in the second of these two peaks, FL experiences many more cases than either NY or WA.

Problems with using single day data may be twofold: 1) the daily cases result in more noise on the visual representations of the data, and 2) mostly likely, not all counties report on a daily basis. For example, if a number of counties report their case counts on a weekly basis, the daily representation of data may not be accurate. This can be seen in the graphic above, in which there appear to be regular, small peaks over time, most visible in the Florida data.

ii) Create **weekly new** cases per 100k `weekly_case_per100k`. Plot the spaghetti plots of `weekly_case_per100k` by state. Use `TotalPopEst2019` as population.



We also present this same image without the legend for greater clarity:



iii) Summarize the COVID case trend among states based on the plot in ii). What could be the possible reasons to explain the variabilities?

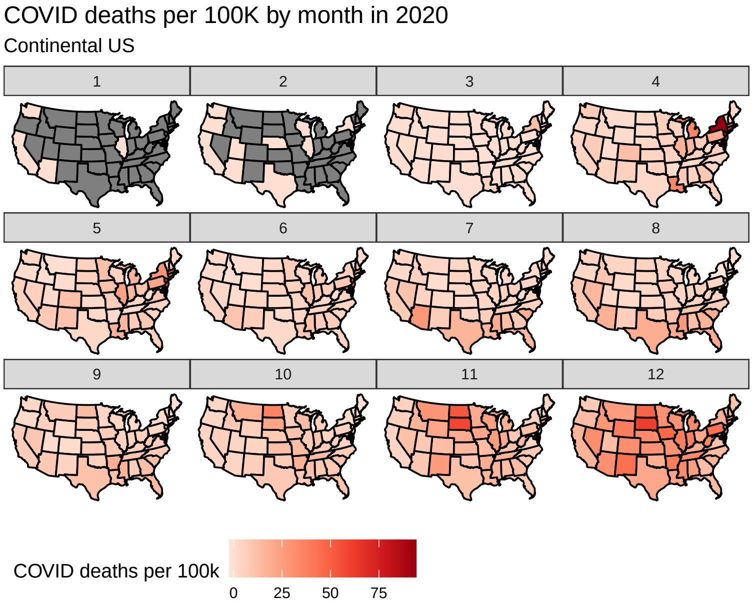
With 50 states included in the spaghetti plot, we can observe overall trends across states but identifying individual states can be challenging. Generally, we see two smaller COVID case peaks: one that begins after week 10 and another that takes place around week 25. These peaks in cases per 100K residents are small comared with the largest peak that occurs after week 40.

Clearly, there is also variation from state to state. Any number of reasons may contribute to this variation. Given that the measure of interest is weekly new cases per 100K people by state, factors that might affect variability in the numerator include public health measures taken by state and local officials that affect spread, such as the degree of lockdown, presence and prevalence of mask-wearing mandates, and vaccination rates. Geography and population density can also affect variability, with more urban environments experiencing faster spread. Since this measure is also scaled in terms of population (i.e., 100K people in 2019) variability in this graph may also occur due to differences in local population. For example, states with lower populations but massive spread might reveal higher rates than more populous states with a greater number of absolute cases.

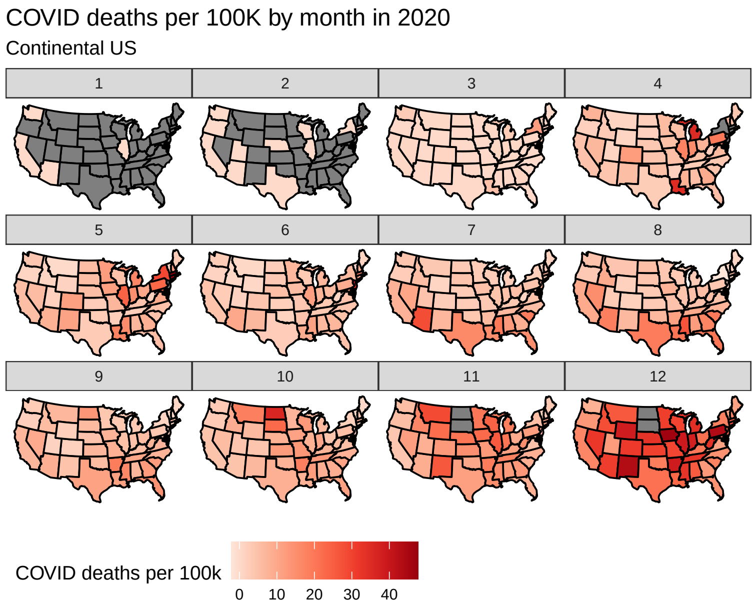
Finally, we note that around week 33, one state has a “negative” number of weekly new COVID cases. Since this is not possible, the value is most likely erroneous. We leave this as is for the purposes of this question, but this small blip in the data as evidenced by the spaghetti plot may warrant closer examination and future cleaning.

3.3 COVID death trend

i) For each month in 2020, plot the monthly deaths per 100k heatmap by state on US map. Use the same color range across months. (Hints: Set `limits` argument in `scale_fill_gradient()` or use `facet_wrap()`; use `lubridate::month()` and `lubridate::year()` to extract month and year from date; use `tidyr::complete(state, month, fill = list(new_case_per100k = NA))` to complete the missing months with no cases.)



In the figure above, each panel number represents a month in 2020 (1 = January, 2 = February, etc.). Gray states in months 1 and 2 indicate that no COVID deaths had yet occurred. We notice that the COVID death rate per 100K in New York in month 4 dominates the rest of the maps (i.e., the contrast in the remainder of the heatmap is somewhat washed out). To improve contrast, we redo the image above, limiting the range of coloring so that the top 2% of COVID death rates per 100K are not included:


COVID deaths per 100K by month in 2020
Continental US
1
2
3
4
5
6
7
8
9
10
11
12
COVID deaths per 100k
0
10
20
30
40

4 COVID factor

We now try to build a good parsimonious model to find possible factors related to death rate on the county level. Let us not take time series into account for the moment and use the total number as of *Feb 1, 2021.*

i) Create the response variable `total_death_per100k` as the total of number of COVID deaths per 100k by *Feb 1, 2021.* We suggest to take log transformation as `log_total_death_per100k = log(total_death_per100k + 1)`. Merge `total_death_per100k` to `county_data` for the following analysis.

We created the variable `log_total_death_per100k` and merged the `covid_rate` data witht `county_data` by `FIPS`.

ii) Select possible variables in `county_data` as covariates. We provide `county_data_sub`, a subset variables from `county_data`, for you to get started. Please add any potential variables as you wish.

We used the suggested variables and also added `MedHHInc`: Median household income in 2018 and `BlackNonHispanicNum2010`. Across these 44 total variables, we found that a number of them (such as `Deep_Pov_All`, `PerCapitaInc` and `PctEmpFIRE`) were missing one value; at most, `HiAmenity` was missing two values. When filtering for missing values, we found that 3 counties - Broomfield, CO; Champaign, IL; and Rio Arriba, NM - accounted for all the missing values. Because this dataset had 3108 observations, we addressed the missing values by dropping these three counties.

iii) Use LASSO to choose a parsimonious model with all available sensible county-level information. **Force in State** in the process. Why we need to force in State? You may use `lambda.1se` to choose a smaller model.

We used LASSO to choose a parsimonious model from the covariates selected in part (ii). In doing so, we forced in `state` so that the dummy variables for state are locked into the model, and no penalty is imposed on these coefficients. We do this because, as we showed in Part 3 (the EDA), COVID severity varies by state. As such, we want to make sure we control for any state effects (i.e., hold states constant) to better separate state effects from the county-level information that we are interested in. Furthermore, in deciding on specific model to use, we used `lambda.1se` as recommended in order to limit the number of coefficients included, within reason.

Below are the non-zero coefficients included in our model after LASSO, including all the state dummy variables:

```
##               (Intercept)              stateArizona             stateArkansas
##                  5.22e+00                  2.68e-01                  2.12e-02
##           stateCalifornia             stateColorado          stateConnecticut
##                 -9.13e-01                 -6.78e-01                  1.51e-01
##             stateDelaware stateDistrict of Columbia              stateFlorida
##                 -1.10e-01                  3.16e-01                 -2.32e-02
##              stateGeorgia                stateIdaho             stateIllinois
##                 -6.78e-03                 -5.94e-01                  1.28e-01
##              stateIndiana                 stateIowa               stateKansas
##                 -1.19e-01                  2.02e-01                 -8.04e-01
##             stateKentucky            stateLouisiana                stateMaine
##                 -8.08e-01                  2.28e-01                 -1.52e+00
##             stateMaryland        stateMassachusetts             stateMichigan
##                 -1.92e-01                 -2.79e-02                 -2.30e-01
##            stateMinnesota          stateMississippi             stateMissouri
##                 -3.23e-01                  2.25e-01                 -4.39e-01
##              stateMontana             stateNebraska               stateNevada
##                  8.94e-02                 -6.41e-01                 -8.94e-01
##        stateNew Hampshire           stateNew Jersey           stateNew Mexico
##                 -9.56e-01                  3.89e-01                 -5.50e-01
##             stateNew York       stateNorth Carolina         stateNorth Dakota
##                 -3.62e-01                 -4.58e-01                  4.79e-01
##                 stateOhio             stateOklahoma               stateOregon
##                 -5.32e-01                 -5.04e-01                 -1.11e+00
##         statePennsylvania          stateRhode Island       stateSouth Carolina
##                  6.51e-02                 -1.99e-01                 -3.19e-02
##         stateSouth Dakota            stateTennessee                stateTexas
##                  5.43e-01                  4.32e-02                  9.98e-02
##                 stateUtah              stateVermont             stateVirginia
##                 -1.28e+00                 -2.33e+00                 -5.79e-01
##           stateWashington        stateWest Virginia            stateWisconsin
##                 -1.11e+00                 -7.24e-01                 -2.67e-01
##              stateWyoming         PovertyAllAgesPct                  MedHHInc
##                 -2.60e-01                  8.41e-04                 -9.36e-07
##              PerCapitaInc        PctEmpConstruction              PctEmpMining
##                 -8.94e-07                 -2.42e-02                 -1.88e-03
##               PctEmpTrade         PctEmpAgriculture       PctEmpManufacturing
##                  1.56e-04                 -1.76e-02                  4.60e-03
##            PopDensity2010      Age65AndOlderPct2010            Under18Pct2010
##                  2.73e-06                  8.05e-03                  2.07e-02
##         Ed3SomeCollegePct         Ed5CollegePlusPct      NetMigrationRate1019
##                 -5.27e-03                 -9.93e-03                 -9.76e-03
##     NaturalChangeRate1019   WhiteNonHispanicPct2010           HispanicPct2010
##                 -1.74e-02                 -2.74e-03                  2.54e-03
##          Type_2015_Update
##                 -8.67e-03
```

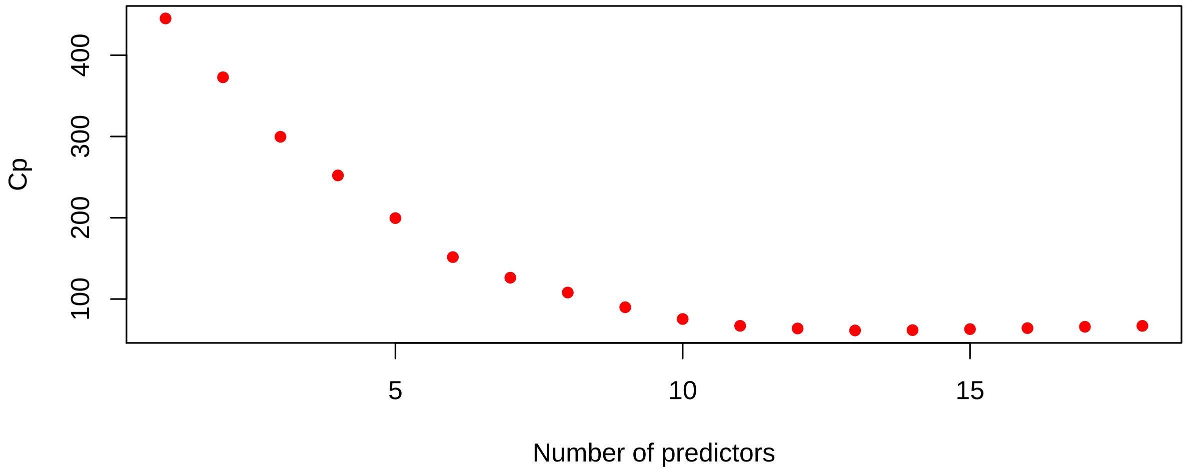
Using LASSO, we have narrowed down the initial set of 41 covariates (in addition to `state`) down to 18. We now fit all 18 of these covariates to a linear model as follows, assuming all the linear model assumptions to be true.

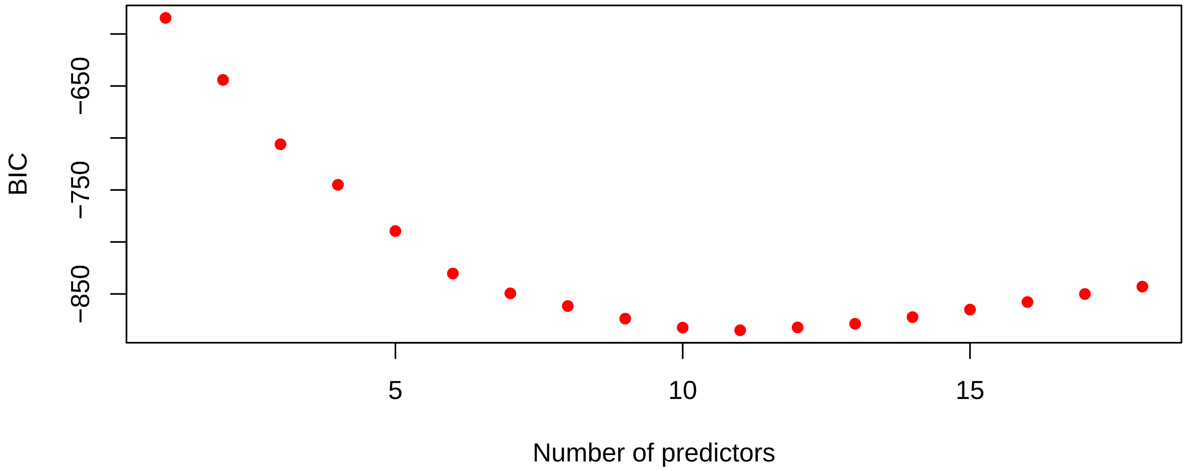
```
##
## Call:
## lm(formula = log_total_death_per100k ~ ., data = data.model)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.784 -0.251  0.067  0.372  3.511
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.28e+00   4.61e-01    9.27  < 2e-16 ***
## stateArizona               2.57e-01   2.39e-01    1.07  0.28249
## stateArkansas              5.32e-02   1.35e-01    0.39  0.69361
## stateCalifornia           -7.89e-01   1.61e-01   -4.89  1.0e-06 ***
## stateColorado             -3.56e-01   1.53e-01   -2.33  0.01998 *
## stateConnecticut           1.72e-01   3.03e-01    0.57  0.57078
## stateDelaware             -2.22e-02   4.69e-01   -0.05  0.96217
## stateDistrict of Columbia  5.71e-01   8.09e-01    0.71  0.48057
## stateFlorida               2.93e-02   1.47e-01    0.20  0.84223
## stateGeorgia               6.35e-02   1.17e-01    0.54  0.58593
## stateIdaho                -3.20e-01   1.67e-01   -1.92  0.05510 .
## stateIllinois              2.01e-01   1.33e-01    1.51  0.13231
## stateIndiana              -7.20e-02   1.34e-01   -0.54  0.59138
## stateIowa                  2.77e-01   1.34e-01    2.06  0.03926 *
## stateKansas               -6.28e-01   1.38e-01   -4.55  5.5e-06 ***
## stateKentucky             -7.53e-01   1.29e-01   -5.86  5.2e-09 ***
## stateLouisiana             3.79e-01   1.43e-01    2.65  0.00811 **
## stateMaine                -1.37e+00   2.25e-01   -6.07  1.5e-09 ***
## stateMaryland             -6.69e-02   1.98e-01   -0.34  0.73525
## stateMassachusetts         7.37e-02   2.41e-01    0.31  0.75950
## stateMichigan             -1.51e-01   1.36e-01   -1.11  0.26682
## stateMinnesota            -1.66e-01   1.38e-01   -1.20  0.23028
## stateMississippi           2.65e-01   1.32e-01    2.01  0.04428 *
## stateMissouri             -3.78e-01   1.29e-01   -2.94  0.00332 **
## stateMontana               5.61e-01   1.58e-01    3.55  0.00039 ***
## stateNebraska             -3.93e-01   1.43e-01   -2.75  0.00595 **
## stateNevada               -5.79e-01   2.28e-01   -2.54  0.01119 *
## stateNew Hampshire        -8.13e-01   2.73e-01   -2.97  0.00297 **
## stateNew Jersey            3.03e-01   2.11e-01    1.44  0.15000
## stateNew Mexico           -6.50e-01   1.93e-01   -3.37  0.00075 ***
## stateNew York             -3.70e-01   1.51e-01   -2.45  0.01451 *
## stateNorth Carolina       -3.66e-01   1.27e-01   -2.89  0.00390 **
## stateNorth Dakota          9.53e-01   1.66e-01    5.74  1.1e-08 ***
## stateOhio                 -5.21e-01   1.35e-01   -3.87  0.00011 ***
## stateOklahoma             -3.70e-01   1.38e-01   -2.69  0.00724 **
## stateOregon               -8.72e-01   1.75e-01   -4.99  6.5e-07 ***
## statePennsylvania          3.95e-02   1.46e-01    0.27  0.78736
## stateRhode Island         -1.82e-01   3.72e-01   -0.49  0.62492
## stateSouth Carolina       -1.01e-02   1.53e-01   -0.07  0.94738
## stateSouth Dakota          8.10e-01   1.51e-01    5.37  8.5e-08 ***
## stateTennessee             6.47e-02   1.30e-01    0.50  0.61958
## stateTexas                 1.67e-01   1.28e-01    1.31  0.18962
## stateUtah                 -1.09e+00   1.97e-01   -5.56  2.9e-08 ***
## stateVermont              -2.13e+00   2.39e-01   -8.93  < 2e-16 ***
## stateVirginia             -4.83e-01   1.24e-01   -3.90  9.9e-05 ***
## stateWashington           -8.64e-01   1.69e-01   -5.10  3.6e-07 ***
## stateWest Virginia        -6.70e-01   1.56e-01   -4.31  1.7e-05 ***
## stateWisconsin            -1.37e-01   1.40e-01   -0.97  0.33067
## stateWyoming               2.16e-01   2.06e-01    1.05  0.29486
## PovertyAllAgesPct          5.45e-03   5.82e-03    0.94  0.34943
## MedHHInc                   3.17e-06   3.39e-06    0.93  0.35057
## PerCapitaInc              -9.06e-06   6.98e-06   -1.30  0.19471
## PctEmpConstruction        -4.72e-02   7.42e-03   -6.37  2.2e-10 ***
## PctEmpMining              -1.81e-02   5.81e-03   -3.12  0.00181 **
## PctEmpTrade                6.07e-03   6.13e-03    0.99  0.32178
## PctEmpAgriculture         -3.85e-02   3.70e-03  -10.40  < 2e-16 ***
## PctEmpManufacturing        4.21e-03   3.47e-03    1.21  0.22536
## PopDensity2010             2.65e-05   9.36e-06    2.83  0.00462 **
## Age65AndOlderPct2010       3.65e-02   8.33e-03    4.37  1.3e-05 ***
## Under18Pct2010             5.85e-02   8.00e-03    7.32  3.3e-13 ***
## Ed3SomeCollegePct         -2.02e-02   5.54e-03   -3.65  0.00027 ***
## Ed5CollegePlusPct         -1.10e-02   3.88e-03   -2.84  0.00458 **
## NetMigrationRate1019      -1.24e-02   2.64e-03   -4.70  2.7e-06 ***
## NaturalChangeRate1019     -4.36e-02   1.00e-02   -4.34  1.5e-05 ***
## WhiteNonHispanicPct2010   -2.89e-03   1.67e-03   -1.73  0.08377 .
## HispanicPct2010            8.82e-03   2.18e-03    4.04  5.5e-05 ***
## Type_2015_Update          -1.86e-02   8.66e-03   -2.15  0.03178 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.79 on 3038 degrees of freedom
## Multiple R-squared:  0.359,  Adjusted R-squared:  0.345
## F-statistic: 25.8 on 66 and 3038 DF,  p-value: <2e-16
```

We see that not all the predictors shown in the linear model above are significant at the 0.05 level. Thus, we go one step further and use `regsubsets` and `Cp` to eliminate some insignificant predictors, as shown in the next step.

iv) Use `Cp` or BIC to fine tune the LASSO model from iii). Again **force in State** in the process.

We use `Cp` to fine tune the LASSO model, and check these findings against BIC. In this step, we again force in `state`, for the reasons described above. `Cp` and `BIC` plots are as shown below.



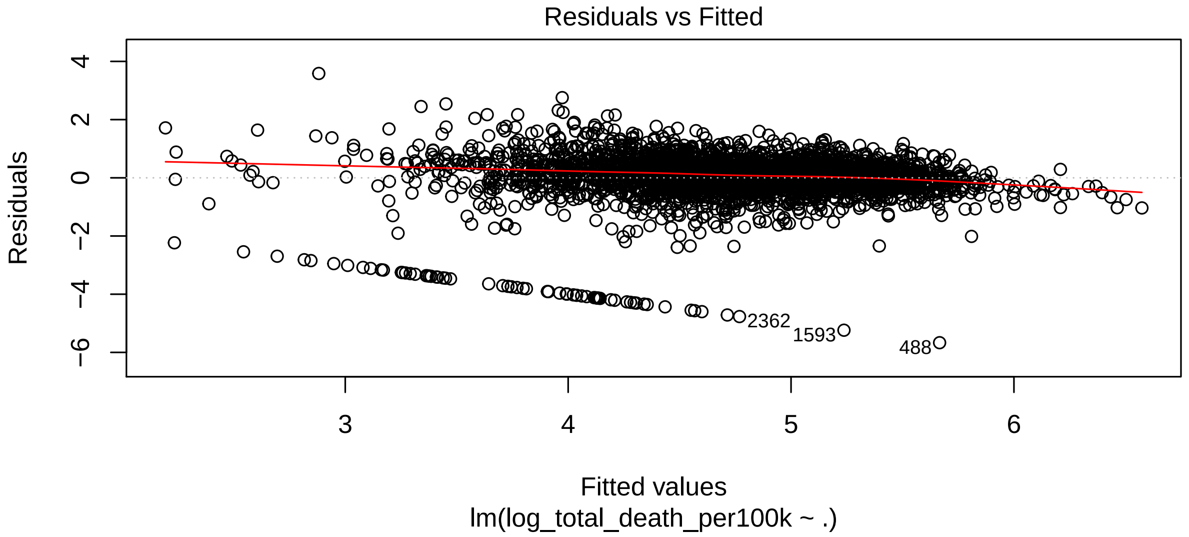


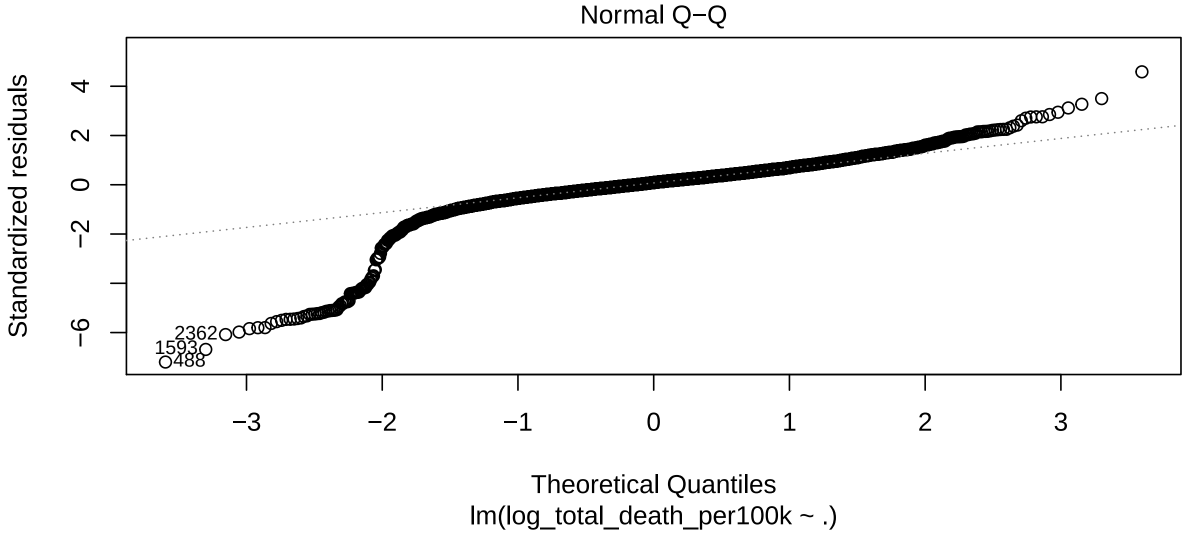
Given both the `Cp` and `BIC` plots, on top of the `state` dummies forced in, it seems sensible to incorporate 10 additional variables into our model. The summary of our final model after `Cp` can be found below:

```
##
## Call:
## lm(formula = log_total_death_per100k ~ ., data = data.model[,
##     c("log_total_death_per100k", "state", final.var)])
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.667 -0.259  0.070  0.377  3.585
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                4.52273    0.26709   16.93  < 2e-16 ***
## stateArizona               0.20775    0.23668    0.88  0.38012
## stateArkansas              0.02140    0.13393    0.16  0.87306
## stateCalifornia           -0.86340    0.15481   -5.58  2.7e-08 ***
## stateColorado             -0.43379    0.15037   -2.88  0.00394 **
## stateConnecticut           0.06973    0.29926    0.23  0.81576
## stateDelaware             -0.09926    0.46873   -0.21  0.83231
## stateDistrict of Columbia  0.65712    0.80381    0.82  0.41370
## stateFlorida              -0.01647    0.14232   -0.12  0.90786
## stateGeorgia               0.05272    0.11610    0.45  0.64981
## stateIdaho                -0.39909    0.16257   -2.45  0.01415 *
## stateIllinois              0.10669    0.12702    0.84  0.40104
## stateIndiana              -0.18257    0.12775   -1.43  0.15306
## stateIowa                  0.17764    0.12851    1.38  0.16697
## stateKansas               -0.70585    0.13325   -5.30  1.3e-07 ***
## stateKentucky             -0.84118    0.12140   -6.93  5.2e-12 ***
## stateLouisiana             0.37069    0.14113    2.63  0.00867 **
## stateMaine                -1.49302    0.22194   -6.73  2.1e-11 ***
## stateMaryland             -0.15727    0.19066   -0.82  0.40952
## stateMassachusetts        -0.00519    0.23772   -0.02  0.98259
## stateMichigan            -0.23709    0.13219   -1.79  0.07298 .
## stateMinnesota           -0.26774    0.13309   -2.01  0.04434 *
## stateMississippi          0.30881    0.13124    2.35  0.01868 *
## stateMissouri            -0.45915    0.12306   -3.73  0.00019 ***
## stateMontana              0.48472    0.15512    3.12  0.00180 **
## stateNebraska            -0.48242    0.13797   -3.50  0.00048 ***
## stateNevada              -0.64674    0.22587   -2.86  0.00422 **
## stateNew Hampshire       -0.94950    0.27058   -3.51  0.00046 ***
## stateNew Jersey           0.25412    0.20293    1.25  0.21056
## stateNew Mexico          -0.72505    0.18966   -3.82  0.00013 ***
## stateNew York            -0.42884    0.14231   -3.01  0.00261 **
## stateNorth Carolina      -0.38692    0.12646   -3.06  0.00224 **
## stateNorth Dakota         0.84816    0.16071    5.28  1.4e-07 ***
## stateOhio                -0.63227    0.12904   -4.90  1.0e-06 ***
## stateOklahoma            -0.40009    0.13630   -2.94  0.00336 **
## stateOregon              -0.93173    0.17248   -5.40  7.1e-08 ***
## statePennsylvania        -0.06826    0.14018   -0.49  0.62634
## stateRhode Island        -0.28687    0.37035   -0.77  0.43864
## stateSouth Carolina      -0.00123    0.15223   -0.01  0.99357
## stateSouth Dakota         0.73795    0.14796    4.99  6.5e-07 ***
## stateTennessee           -0.00340    0.12761   -0.03  0.97875
## stateTexas                0.11936    0.12573    0.95  0.34251
## stateUtah                -1.20145    0.18805   -6.39  1.9e-10 ***
## stateVermont             -2.27975    0.23566   -9.67  < 2e-16 ***
## stateVirginia            -0.51620    0.12062   -4.28  1.9e-05 ***
## stateWashington          -0.92342    0.16722   -5.52  3.6e-08 ***
## stateWest Virginia       -0.79220    0.14770   -5.36  8.8e-08 ***
## stateWisconsin           -0.25159    0.13589   -1.85  0.06421 .
## stateWyoming              0.10609    0.20264    0.52  0.60064
## PctEmpConstruction       -0.05708    0.00667   -8.55  < 2e-16 ***
## PctEmpMining             -0.02403    0.00518   -4.64  3.7e-06 ***
## PctEmpAgriculture        -0.04144    0.00334  -12.42  < 2e-16 ***
## Age65AndOlderPct2010      0.03201    0.00757    4.23  2.4e-05 ***
## Under18Pct2010            0.06021    0.00717    8.40  < 2e-16 ***
## Ed3SomeCollegePct        -0.02414    0.00519   -4.65  3.4e-06 ***
## Ed5CollegePlusPct        -0.01578    0.00227   -6.94  4.6e-12 ***
## NetMigrationRate1019     -0.01388    0.00244   -5.68  1.5e-08 ***
## NaturalChangeRate1019    -0.03966    0.00981   -4.04  5.4e-05 ***
## HispanicPct2010           0.01120    0.00189    5.93  3.4e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.792 on 3046 degrees of freedom
## Multiple R-squared:  0.354,  Adjusted R-squared:  0.342
## F-statistic: 28.8 on 58 and 3046 DF,  p-value: <2e-16
```

v) If necessary, reduce the model from iv) to a final model with all variables being significant at 0.05 level. Are the linear model assumptions all reasonably met?

From the summary of our final model fit above, we see that all 10 variables in addition to the state dummy variables are significant at the 0.05 level. Below, we run some model diagnostics.





From the model diagnostics above, we see that the linear model assumptions of this model are a bit tenuous. Linearity and homoscedasticity may be reasonably met; however, there appears to be a heavy left skew to the data. In practice, this means that many counties have fewer COVID deaths per 100K people than expected. This might make sense, since while COVID has seriously affected the US on the whole, there may be pockets of the country that have been largely unaffected. Regardless, this apparent violation of normality may warrant future study.

Based on the model, That is, the spread is not constant, which provides evidence for heteroskedasticity as discussed above. An important conclusion from the model is also the heavy left skew.

vi) It has been shown that COVID affects elderly the most. It is also claimed that the COVID death rate among African Americans and Latinos is higher. Does your analysis support these arguments?

Our model does support the argument that COVID affects the elderly population the most. Specifically, `Age65AndOlderPct2010` was included in our final model, with a positive coefficient of 0.03201. This means that, for a one percent increase in the percent of population age 65 or older in 2010, the COVID death rate per 100K people is expected to increase by about 3.3% on average, holding other variables in the model fixed. This finding was significant at the 0.05 level.

Similarly, we also find support for the argument that the COVID death rate among Latino communities may be higher. The variable `HispanicPct2010` was also included in our model, and had a positive coefficient of 0.01120. This means our model predicts that, for a one percent increase of the Hispanic population in the county, holding other variables in the model fixed, the COVID death rate per 100K individuals among these communities was expected to increase by 1.1%, on average.

Surprisingly, our model did not find support for the claim that the COVID death rate among African Americans is higher. We included two relevant variables - `BlackNonHispanicPct2010` and `BlackNonHispanicNum2010` - in our initial pool of covariates, but neither remained after we used LASSO. In fact, the finding described above regarding Hispanic communities was our only race-related covariate included in our final model.

vii) Based on your final model, summarize your findings. In particular, summarize the state effect controlling for others. Provide intervention recommendations to policy makers to reduce COVID death rate.

We discussed some of our model inplications above, regarding the elderly population and the Latino population. For other non-state effects, we see that controlling for other variables in the model, higher levels of education (e.g., some college or more) had a negative relationship with predicted COVID mortality, and counties with larger percents of individuals under age 18 had increased rates of COVID-19 mortality. Interestingly, larger percentages of individuals in three industries of employment - Construction, Mining, and Agriculture - were associated with lower rates of COVID mortality, controlling for the other variables. This may be because places where these occupations prevail are less urban in nature, and thus, have lower rates of COVID spread.

In building this model, we also forced in state dummy variables because during EDA, we noted that there is clear variation of COVID infection and mortality rates by state. So, forcing in state dummies allows us to examine the state effect, or the effect that any individual state may have on COVID-19 mortality per 100K individuals. Thus, to highlight some interesting state effects, we can look at the statistically significant coefficients for individual states at the 0.05 level. For example, controlling for all else, Kansas, Kentucky, Oregon, Ohio, and New Mexico all have a negative estimate, suggesting that counties in those states had lower rates of COVID-19 mortality relative to Alabama. On the other hand, some states that had positive statistically significant coefficients (i.e., have higher deaths per 100k), include Montana, North Dakota, Mississippi, and Louisiana. Thus, our final model clearly suggests significant geographic variation in COVID-19 mortality. This may be for various reasons, but at baseline, these results may provide evidence for policymakers to direct current or future infectious disease prevention efforts in a more targeted manner, perhaps providing certaing states, or counties with greater proportions of Latino and older adults, with more resources.

Additionally, combining our results with some EDA of the `covid_intervention` dataframe, we saw that the `stay.at.home` lockdown and rollback did not vary dramatically between states with higher and lower death rates. This may indicate that the stay.at.home lockdown was not enforced strictly enough and that the intervention was not successful in combating covid death rates. While this is speculative, how and when policymakers choose to implement interventions is critical for consideration; hopefully, the experience and data generated in the past couple years' experience with the ongoing COVID-19 provides data and lessons for evidence-based approaches to future policy interventions.

viii) What else can we do to improve our model? What other important information we may have missed?

In a health crisis, having clinical or health data could help improve the predictive ability of a model. For example, something important to consider with COVID deaths is whether or not an individual has pre-existing conditions, which may have been worsened significantly by contracting COVID or comprised the immune system in battling the illness. In this example, example, counties with higher percentages of individuals with comorbidities may experience higher COVID death rates in their communities. Thus, in this case, there could be another variable added to the data which identifies what percent of individuals with, for example, 3 or more comorbidities as a predictor. In this way, adding clinical data to the predictive variables in the model may improve the model.

As we also discussed, there is also a lot of variability within states in terms of population density and demographics, which likely affects rates of COVID infection and mortality. These may be a result of the difference between urban and rural areas within states. A dummy variable for rural/urban (or a similar approach to specifying population density) could be an informative addition to the model. As with the state variation described above, if these predictors were significant, they could give public health officials better insight into how to tackle COVID in areas based on the population density - which may directly impact transmission risk - in a given area.

Appendix 1: Data description

A detailed summary of the variables in each data set follows:

Infection and fatality data

- date: Date
- county: County name
- state: State name
- fips: County code that uniquely identifies a county
- cases: Number of cumulative COVID-19 infections
- deaths: Number of cumulative COVID-19 deaths

Socioeconomic demographics

Income: Poverty level and household income

- PovertyUnder18Pct: Poverty rate for children age 0-17, 2018
- Deep_Pov_All: Deep poverty, 2014-18
- Deep_Pov_Children: Deep poverty for children, 2014-18
- PovertyAllAgesPct: Poverty rate, 2018
- MedHHInc: Median household income, 2018 (In 2018 dollars)
- PerCapitaInc: Per capita income in the past 12 months (In 2018 inflation adjusted dollars), 2014-18
- PovertyAllAgesNum: Number of people of all ages in poverty, 2018
- PovertyUnder18Num: Number of people age 0-17 in poverty, 2018

Jobs: Employment type, rate, and change

- UnempRate2007-2019: Unemployment rate, 2007-2019
- NumEmployed2007-2019: Employed, 2007-2019
- NumUnemployed2007-2019: Unemployed, 2007-2019
- PctEmpChange1019: Percent employment change, 2010-19
- PctEmpChange1819: Percent employment change, 2018-19
- PctEmpChange0719: Percent employment change, 2007-19
- PctEmpChange0710: Percent employment change, 2007-10
- NumCivEmployed: Civilian employed population 16 years and over, 2014-18
- NumCivLaborforce2007-2019: Civilian labor force, 2007-2019

- PctEmpFIRE: Percent of the civilian labor force 16 and over employed in finance and insurance, and real estate and rental and leasing, 2014-18

- PctEmpConstruction: Percent of the civilian labor force 16 and over employed in construction, 2014-18

- PctEmpTrans: Percent of the civilian labor force 16 and over employed in transportation, warehousing and utilities, 2014-18
- PctEmpMining: Percent of the civilian labor force 16 and over employed in mining, quarrying, oil and gas extraction, 2014-18

- PctEmpTrade: Percent of the civilian labor force 16 and over employed in wholesale and retail trade, 2014-18

- PctEmpInformation: Percent of the civilian labor force 16 and over employed in information services, 2014-18

- PctEmpAgriculture: Percent of the civilian labor force 16 and over employed in agriculture, forestry, fishing, and hunting, 2014-18

- PctEmpManufacturing: Percent of the civilian labor force 16 and over employed in manufacturing, 2014-18

- PctEmpServices: Percent of the civilian labor force 16 and over employed in services, 2014-18

- PctEmpGovt: Percent of the civilian labor force 16 and over employed in public administration, 2014-18

People: Population size, density, education level, race, age, household size, and migration rates

- PopDensity2010: Population density, 2010

- LandAreaSQMiles2010: Land area in square miles, 2010
- TotalHH: Total number of households, 2014-18

- TotalOccHU: Total number of occupied housing units, 2014-18

- AvgHHSize: Average household size, 2014-18

- OwnHomeNum: Number of owner occupied housing units, 2014-18

- OwnHomePct: Percent of owner occupied housing units, 2014-18
- NonEnglishHHPct: Percent of non-English speaking households of total households, 2014-18

- HH65PlusAlonePct: Percent of persons 65 or older living alone, 2014-18

- FemaleHHPct: Percent of female headed family households of total households, 2014-18

- FemaleHHNum: Number of female headed family households, 2014-18

- NonEnglishHHNum: Number of non-English speaking households, 2014-18

- HH65PlusAloneNum: Number of persons 65 years or older living alone, 2014-18
- Age65AndOlderPct2010: Percent of population 65 or older, 2010
- Age65AndOlderNum2010: Population 65 years or older, 2010

- TotalPop25Plus: Total population 25 and older, 2014-18 - 5-year average

- Under18Pct2010: Percent of population under age 18, 2010

- Under18Num2010: Population under age 18, 2010
- Ed1LessThanHSPct: Percent of persons with no high school diploma or GED, adults 25 and over, 2014-18

- Ed2HSDiplomaOnlyPct: Percent of persons with a high school diploma or GED only, adults 25 and over, 2014-18

- Ed3SomeCollegePct: Percent of persons with some college experience, adults 25 and over, 2014-18

- Ed4AssocDegreePct: Percent of persons with an associate's degree, adults 25 and over, 2014-18

- Ed5CollegePlusPct: Percent of persons with a 4-year college degree or more, adults 25 and over, 2014-18

- Ed1LessThanHSNum: No high school, adults 25 and over, 2014-18
- Ed2HSDiplomaOnlyNum: High school only, adults 25 and over, 2014-18

- Ed3SomeCollegeNum: Some college experience, adults 25 and over, 2014-18

- Ed4AssocDegreeNum: Number of persons with an associate's degree, adults 25 and over, 2014-18

- Ed5CollegePlusNum: College degree 4-years or more, adults 25 and over, 2014-18
- ForeignBornPct: Percent of total population foreign born, 2014-18

- ForeignBornEuropePct: Percent of persons born in Europe, 2014-18

- ForeignBornMexPct: Percent of persons born in Mexico, 2014-18
- ForeignBornCentralSouthAmPct: Percent of persons born in Central or South America, 2014-18
- ForeignBornAsiaPct: Percent of persons born in Asia, 2014-18
- ForeignBornCaribPct: Percent of persons born in the Caribbean, 2014-18
- ForeignBornAfricaPct: Percent of persons born in Africa, 2014-18
- ForeignBornNum: Number of people foreign born, 2014-18
- ForeignBornCentralSouthAmNum: Number of persons born in Central or South America, 2014-18
- ForeignBornEuropeNum: Number of persons born in Europe, 2014-18
- ForeignBornMexNum: Number of persons born in Mexico, 2014-18
- ForeignBornAfricaNum: Number of persons born in Africa, 2014-18
- ForeignBornAsiaNum: Number of persons born in Asia, 2014-18
- ForeignBornCaribNum: Number of persons born in the Caribbean, 2014-18
- Net_International_Migration_Rate_2010_2019: Net international migration rate, 2010-19
- Net_International_Migration_2010_2019: Net international migration, 2010-19
- Net_International_Migration_2000_2010: Net international migration, 2000-10
- Immigration_Rate_2000_2010: Net international migration rate, 2000-10
- NetMigrationRate0010: Net migration rate, 2000-10
- NetMigrationRate1019: Net migration rate, 2010-19
- NetMigrationNum0010: Net migration, 2000-10
- NetMigration1019: Net Migration, 2010-19
- NaturalChangeRate1019: Natural population change rate, 2010-19
- NaturalChangeRate0010: Natural population change rate, 2000-10
- NaturalChangeNum0010: Natural change, 2000-10

- NaturalChange1019: Natural population change, 2010-19
- TotalPop2010: Population size 4/1/2010 Census
- TotalPopEst2010: Population size 7/1/2010
- TotalPopEst2011: Population size 7/1/2011
- TotalPopEst2012: Population size 7/1/2012
- TotalPopEst2013: Population size 7/1/2013
- TotalPopEst2014: Population size 7/1/2014
- TotalPopEst2015: Population size 7/1/2015
- TotalPopEst2016: Population size 7/1/2016
- TotalPopEst2017: Population size 7/1/2017
- TotalPopEst2018: Population size 7/1/2018
- TotalPopEst2019: Population size 7/1/2019
- TotalPopACS: Total population, 2014-18 - 5-year average
- TotalPopEstBase2010: County Population estimate base 4/1/2010
- NonHispanicAsianPopChangeRate0010: Population change rate Non-Hispanic Asian, 2000-10
- PopChangeRate1819: Population change rate, 2018-19
- PopChangeRate1019: Population change rate, 2010-19
- PopChangeRate0010: Population change rate, 2000-10
- NonHispanicNativeAmericanPopChangeRate0010: Population change rate Non-Hispanic Native American, 2000-10
- HispanicPopChangeRate0010: Population change rate Hispanic, 2000-10
- MultipleRacePopChangeRate0010: Population change rate multiple race, 2000-10
- NonHispanicWhitePopChangeRate0010: Population change rate Non-Hispanic White, 2000-10
- NonHispanicBlackPopChangeRate0010: Population change rate Non-Hispanic African American, 2000-10
- MultipleRacePct2010: Percent multiple race, 2010
- WhiteNonHispanicPct2010: Percent Non-Hispanic White, 2010

- NativeAmericanNonHispanicPct2010: Percent Non-Hispanic Native American, 2010
- BlackNonHispanicPct2010: Percent Non-Hispanic African American, 2010
- AsianNonHispanicPct2010: Percent Non-Hispanic Asian, 2010
- HispanicPct2010: Percent Hispanic, 2010
- MultipleRaceNum2010: Population size multiple race, 2010
- WhiteNonHispanicNum2010: Population size Non-Hispanic White, 2010
- BlackNonHispanicNum2010: Population size Non-Hispanic African American, 2010
- NativeAmericanNonHispanicNum2010: Population size Non-Hispanic Native American, 2010
- AsianNonHispanicNum2010: Population size Non-Hispanic Asian, 2010
- HispanicNum2010: Population size Hispanic, 2010

County classifications: Type of county (rural or urban on a rural-urban continuum scale)

- Type_2015_Recreation_NO: Recreation counties, 2015 edition
- Type_2015_Farming_NO: Farming-dependent counties, 2015 edition
- Type_2015_Mining_NO: Mining-dependent counties, 2015 edition
- Type_2015_Government_NO: Federal/State government-dependent counties, 2015 edition
- Type_2015_Update: County typology economic types, 2015 edition
- Type_2015_Manufacturing_NO: Manufacturing-dependent counties, 2015 edition
- Type_2015_Nonspecialized_NO: Nonspecialized counties, 2015 edition
- RecreationDependent2000: Nonmetro recreation-dependent, 1997-00
- ManufacturingDependent2000: Manufacturing-dependent, 1998-00
- FarmDependent2003: Farm-dependent, 1998-00
- EconomicDependence2000: Economic dependence, 1998-00

- RuralUrbanContinuumCode2003: Rural-urban continuum code, 2003
- UrbanInfluenceCode2003: Urban influence code, 2003
- RuralUrbanContinuumCode2013: Rural-urban continuum code, 2013
- UrbanInfluenceCode2013: Urban influence code, 2013
- Noncore2013: Nonmetro noncore, outside Micropolitan and Metropolitan, 2013
- Micropolitan2013: Micropolitan, 2013
- Nonmetro2013: Nonmetro, 2013
- Metro2013: Metro, 2013
- Metro_Adjacent2013: Nonmetro, adjacent to metro area, 2013
- Noncore2003: Nonmetro noncore, outside Micropolitan and Metropolitan, 2003
- Micropolitan2003: Micropolitan, 2003
- Metro2003: Metro, 2003
- Nonmetro2003: Nonmetro, 2003
- NonmetroNotAdj2003: Nonmetro, nonadjacent to metro area, 2003
- NonmetroAdj2003: Nonmetro, adjacent to metro area, 2003
- Oil_Gas_Change: Change in the value of onshore oil and natural gas production, 2000-11
- Gas_Change: Change in the value of onshore natural gas production, 2000-11
- Oil_Change: Change in the value of onshore oil production, 2000-11
- Hipov: High poverty counties, 2014-18
- Perpov_1980_0711: Persistent poverty counties, 2015 edition
- PersistentChildPoverty_1980_2011: Persistent child poverty counties, 2015 edition
- PersistentChildPoverty2004: Persistent child poverty counties, 2004
- PersistentPoverty2000: Persistent poverty counties, 2004
- Low_Education_2015_update: Low education counties, 2015 edition

- LowEducation2000: Low education, 2000
- HiCreativeClass2000: Creative class, 2000

- HiAmenity: High natural amenities

- RetirementDestination2000: Retirement destination, 1990-00

- Low_Employment_2015_update: Low employment counties, 2015 edition
- Population_loss_2015_update: Population loss counties, 2015 edition
- Retirement_Destination_2015_Update: Retirement destination counties, 2015 edition

Appendix 2: Data cleaning

The raw data sets are dirty and need transforming before we can do our EDA. It takes time and efforts to clean and merge different data sources so we provide the final output of the cleaned and merged data. The cleaning procedure is as follows. Please read through to understand what is in the cleaned data. We set `eval = data_cleaned` in the following cleaning chunks so that these cleaning chunks will only run if any of `data/covid_county.csv`, `data/covid_rates.csv` or `data/covid_intervention.csv` does not exist.

We first read in the table using `data.table::fread()`, as we did last time.

4.1 Clean NYC data

The original NYC data contains more information than we need. We extract only the number of cases and deaths and format it the same as the `covid_rates` data.

4.2 Continental US cases

We only consider cases and death in continental US. Alaska, Hawaii, and Puerto Rico have 02, 15, and 72 as their respective first 2 digits of their FIPS. We use the `%/%` operator for integer division to get the first 2 digits of FIPS. We also remove Virgin Islands and Northern Mariana Islands. All data of counties in NYC are aggregated as `County == "New York City"` in `covid_rates` with no FIPS, so we combine the NYC data into `covid_rate`.

4.3 COVID date to week

We set the week of Jan 21, 2020 (the first case of COVID case in US) as the first week (2020-01-19 to 2020-01-25).

4.4 COVID infection/mortality rates

Merge the `TotalPopEst2019` variable from the demographic data with `covid_rates` by FIPS.

4.5 NA in COVID data

NA values in the `covid_rates` data set correspond to a county not having confirmed cases/deaths. We replace the NA values in these columns with zeros. FIPS for Kansas city, Missouri, Rhode Island and some others are missing. We drop them for the moment and output the data up to week 57 as `covid_rates.csv`.

4.6 Formatting date in `int_dates`

We convert the columns representing dates in `int_dates` to R Date types using `as.Date()`. We will need to specify that the `origin` parameter is `"0001-01-01"`. We output the data as `covid_intervention.csv`.

4.7 Merge demographic data

Merge the demographic data sets by FIPS and output as `covid_county.csv`.