COVID-19 Case Study

Graham Branscom Mahika Calyanakoti Andrew Raine

Contents

1 Background **1**

2 Data Summary **2**

3 EDA **2**
3.1 Understand the data . . . 3
3.2 COVID case trend . . . 3
3.3 COVID death trend . . . 4

4 COVID factor **5**

5 Executive summary **8**

6 Appendix 0: Intermediate Results **9**

7 Appendix 1: Data description **11**
7.1 Infection and fatality data . . . 12
7.2 Socioeconomic demographics . . . 12

8 Appendix 2: Data cleaning **18**
8.1 Clean NYC data . . . 19
8.2 Continental US cases . . . 20
8.3 COVID date to week . . . 20
8.4 COVID infection/mortality rates . . . 20
8.5 NA in COVID data . . . 21
8.6 Formatting date in `int_dates` . . . 21
8.7 Merge demographic data . . . 21

---

1 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 755 million cases have been confirmed worldwide, with nearly 6.8 million associated deaths by Feb of 2023. Within the US alone, there have been over 1.1 million deaths and upwards of 102 million cases reported by Feb of 2023. 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 assembled this dataset for our research with the goal of investigating the effectiveness of lockdowns in flattening the COVID curve. We are providing a portion of the cleaned dataset for this case study.

There are two main goals for this case study:

1.We aim to show the dynamic evolution of COVID cases and COVID-related deaths at the state level.

2. We aim to identify county-level demographic and policy interventions that are associated with mortality rates in the US. We will construct models to find possible factors related to county-level COVID-19 mortality rates.

3. This is a rather complex project, but with our team's help, we have made your job easier.

4. Please hide all unnecessary lengthy R-output and keep your write-up neat and 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` You may go through it if you are interested or would like to make any changes.

It may need more data wrangling.

First read in the data.

3.1 Understand the data

The detailed description of variables is in `Appendix 1: Data description`. Please get familiar with the variables. Summarize the two data briefly.

We first look at county_data.

There are 207 different variables that we are looking at for each of the 3278 different counties. Each column measures some sort of socioeconomic quality of that county. county_data combines 4 different data sets: Income, Jobs, People, and County Classifications.

```
## [1] 3279  208
```

We now look at covid_rate.

There are 1,008,984 rows for counties at different days (i.e. each county has many rows, one for each day). Each row has 7 different variables. The columns are listed below. It provides cumulative counts of cases (infections) and fatalities (deaths).

```
## [1] 1008984       8
```

```
## [1] "FIPS"            "date"            "County"          "State"
## [5] "cum_cases"       "cum_deaths"      "week"            "TotalPopEst2019"
```

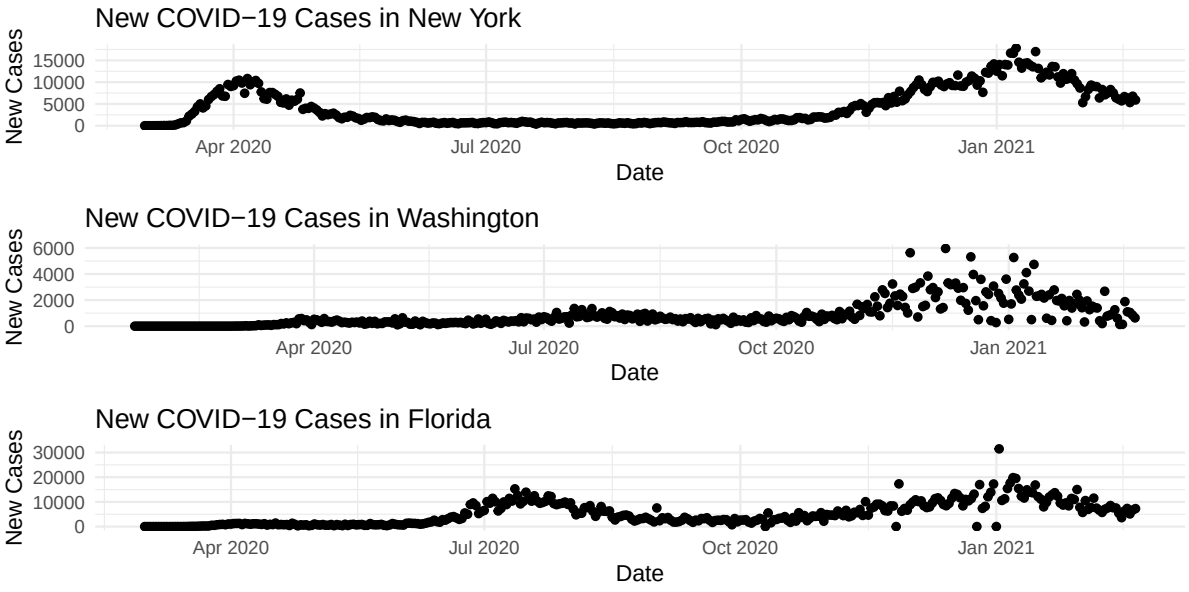
We can see that county and state are characters, but we should change them to factors

3.2 COVID case trend

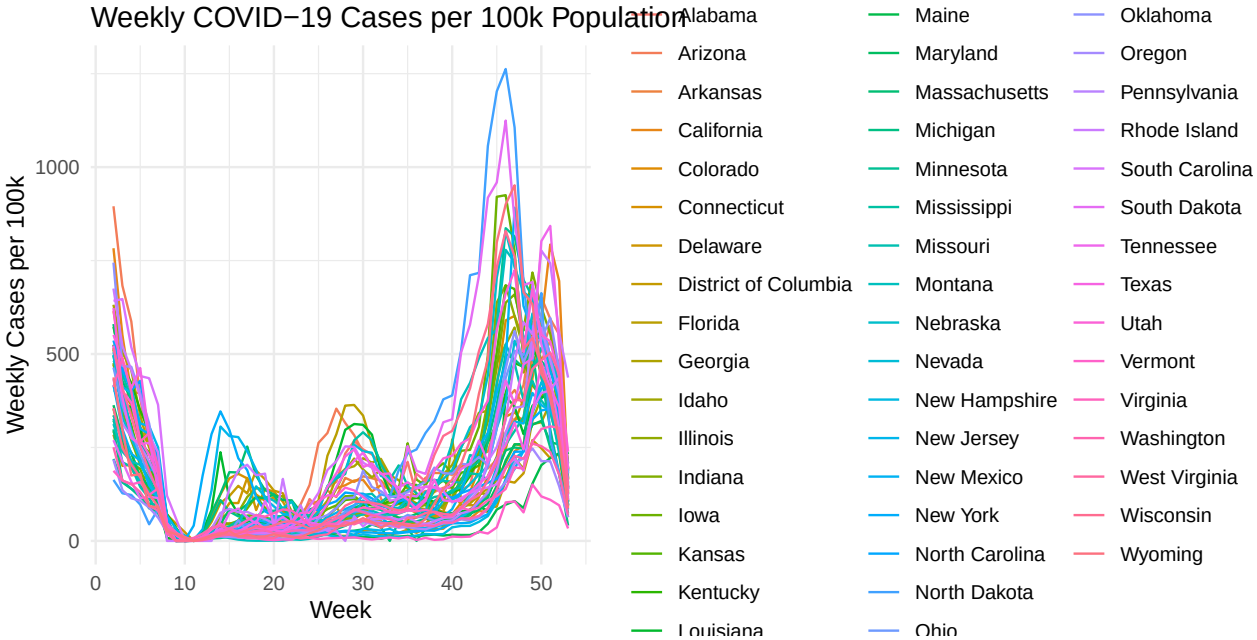
It is crucial to decide the right granularity for visualization and analysis. We will compare daily vs weekly total new cases by state and we will see it is hard to interpret daily report.

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?

We see a wave in NY in April 2020 not seen in the other states to the same degree. The issue with single day data is that it is not normalized to each state's beginning population, so a change in one state is not necessarily as drastic as a change in another. We should instead use proportions. Also, for instance, in Washington in January 2021 there is a large amount of variability because it varies day-to-day so much, so it is hard to interpret. Weekly data would probably be easier to interpret because it would be less varied.



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



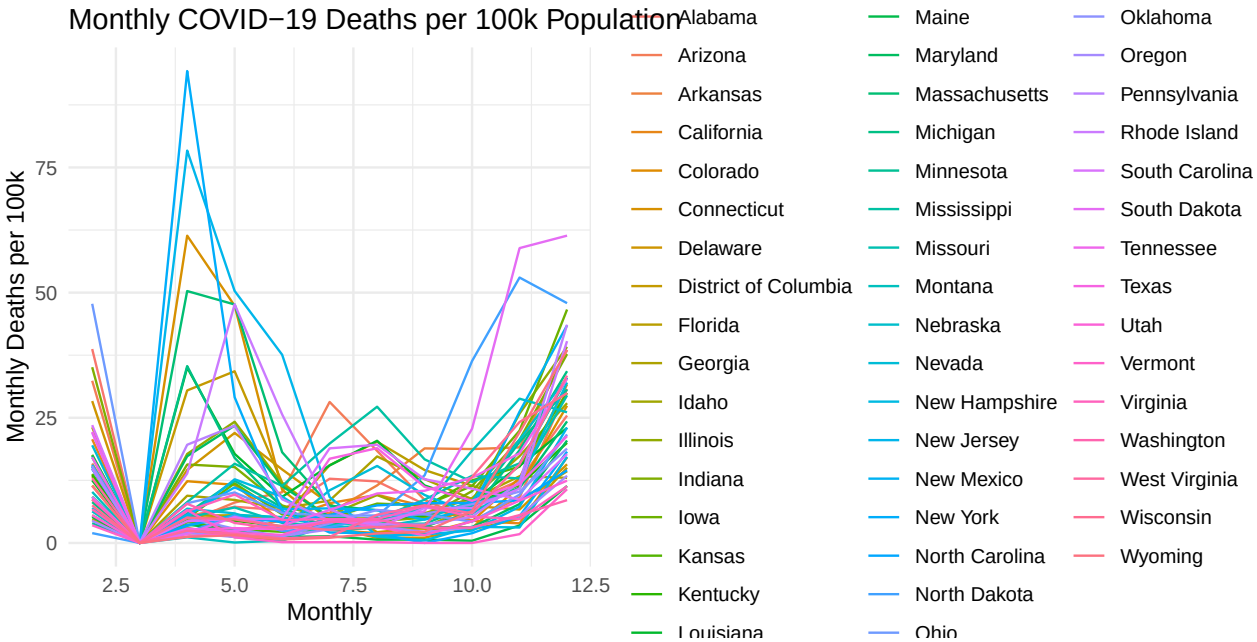
iii) Summarize the COVID case trend among states based on the plot in ii). What could be the possible reasons to explain the variabilities?

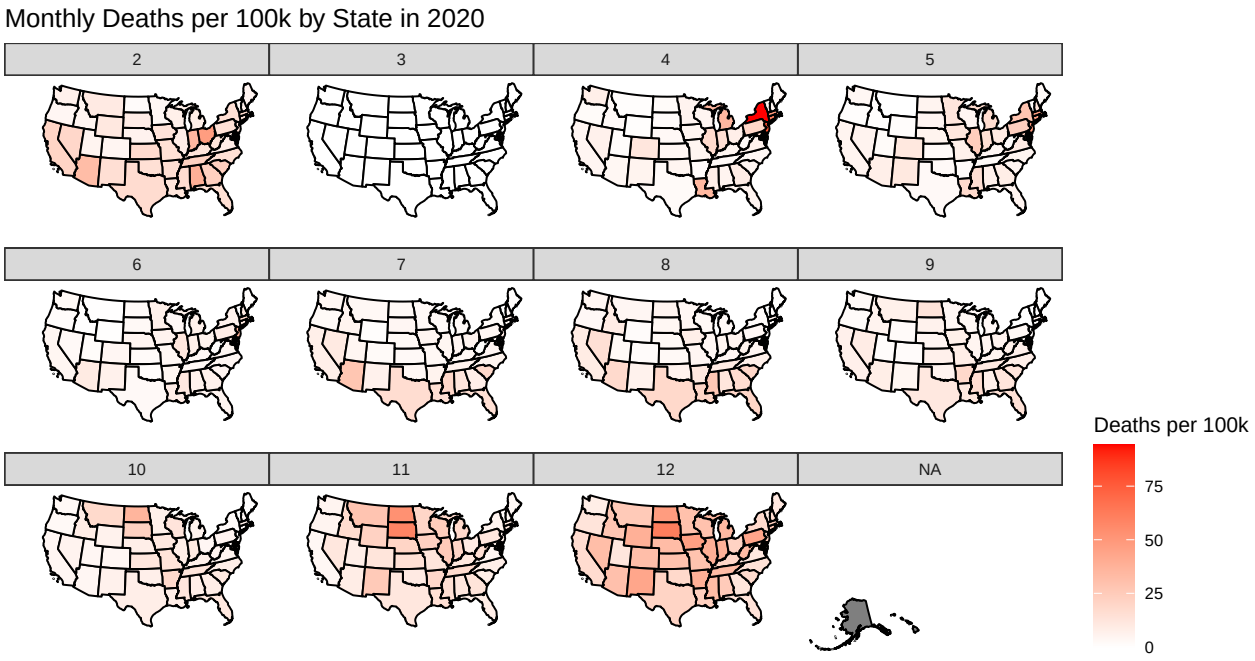
Based on the plot in part ii, we can see a COVID case trend of three waves, with the last wave having the highest peak. This is shown in the spaghetti plot for all of the states in terms of weekly new cases per 100k, which thus adjusts for the state populations and standardizes the metric across all states. These waves are on weeks 13, 28, and 46. The variabilities might be caused by spreads of the virus in specific locations, which might cause that state to be impacted more by the wave. Also varied regulations, such as masks and vaccine requirements, might cause certain states to have more cases per 100k than others.

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.)

After cleaning the data for NA's and extracting the monthly data for each state, we plotted the spaghetti plots for the states, as well as a heatmap for 2020. We see that the peaks in the spaghetti plot correspond to the months in the heatmap that have darker shades of red, specifically around month 4 (in the northeast), slightly in month 8, and increasing in months 10-12. The darker red shades correspond to more monthly deaths per 100k.





ii) (Optional) Use `plotly` to animate the monthly maps in i). Does it reveal any systematic way to capture the dynamic changes among states? (Hints: Follow *Appendix 3: Plotly heatmap::* in Module 6 regularization lecture to plot the heatmap using `plotly`. Use `frame` argument in `add_trace()` for animation. `plotly` only recognizes abbreviation of state names. Use `unique(us_map(regions = "states") %>% select(abbr, full))` to get the abbreviation and merge with the data to get state abbreviation.)

4 COVID factor

We now try to build a good parsimonious model to find possible factors related to death rate on 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.

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

We added the variables NonEnglishHHPct, FemaleHHNum, and AvgHHSize to the existing model variables. For example, we thought higher average household sizes might lead to more viral spread and therefore more deaths. So now, county_data_sub has 45 variables. We also included the response variable, log_total_death_per100k, for when we eventually will run a regression on the data.

`a) Report missing values in your final subset of variables.`

We found 174 rows with NA values.

```
## [1] 174  45
```

`b) In the following anaylsis, you may ignore the missing values.`

After omitting the NA rows, we have 3,105 rows left in county_data_sub.

```
## [1] 3279   45
```

```
## [1] 3105   45
```

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

We need to force in state so that all levels of this factor variable must be included, so no penalties on those betas. This way, none of the states get omitted from the LASSO model. We use penalty factor in the code for this. After we run the model using lambda.1se (for a smaller model), we see all of the states included in selected_vars.

iv) Use a quick backward elimination to fine tune the LASSO model from iii). Again **force in State** in the process.

We ran a backward elimination by removing the highest p-value variables in each iteration one by one until no p-value was greater than 0.05. We set this final model, fit6 to be our final.fit mode. We didn't remove any state from anything. In final.fit, all p-values are less than 0.05.

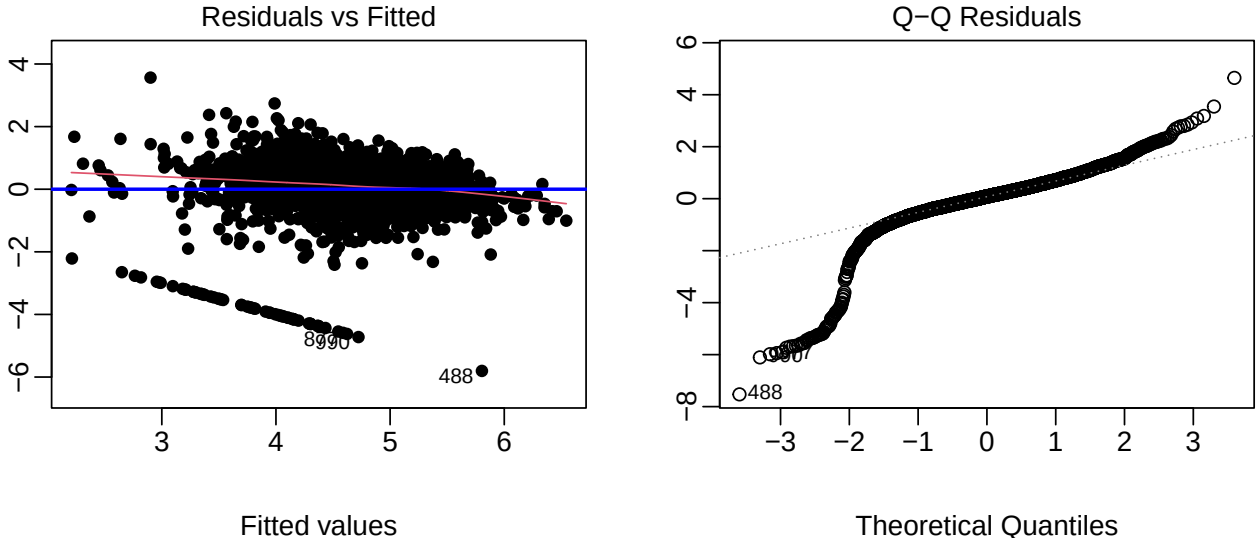
```
## Anova Table (Type II tests)
##
## Response: log_total_death_per100k
##                           Sum Sq   Df F value  Pr(>F)
## State.x                      520   48   17.90 < 2e-16 ***
## PctEmpConstruction            34    1   56.05 9.2e-14 ***
## PctEmpMining                  11    1   18.98 1.4e-05 ***
## PctEmpAgriculture             87    1  144.63 < 2e-16 ***
## PopDensity2010                 4    1    6.73  0.0095 **
## Age65AndOlderPct2010          12    1   19.16 1.2e-05 ***
## Under18Pct2010                36    1   59.57 1.6e-14 ***
## Ed3SomeCollegePct             12    1   19.65 9.7e-06 ***
## Ed5CollegePlusPct             30    1   49.76 2.1e-12 ***
## NetMigrationRate1019          16    1   26.17 3.3e-07 ***
## NaturalChangeRate1019         12    1   19.06 1.3e-05 ***
## WhiteNonHispanicPct2010        3    1    5.41  0.0201 *
## HispanicPct2010               11    1   17.55 2.9e-05 ***
## Type_2015_Update               2    1    4.00  0.0456 *
## Residuals                   1840 3043
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Finally fit our final model

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?

We have ensured that all variables have less than 0.05 as their p-value when doing the backward elimination step, and we forced in State (so even though some states have large p-values, we force them in, but all other non-state variables have low p-values). To test if the linear model assumptions are reasonably met, we run residuals vs fitted, as well as the Q-Q residuals test. We can see in residuals vs fitted that the values are spread out randomly, with some positives and some negatives, along with random variance. In the Q-Q residuals test, we see a generally positive linear relationship among the points. Thus, our model fits the linear model assumptions reasonable with linearity and homoscedasticity. We see a couple of abnormalities, however: in the residuals vs fitted plot, we see a negatively sloped line of a few points, but overall the points are random and evenly spread. In the Q-Q plot, there is a wave in the generally positively sloped line, but overall the Q-Q plot fits the linear model assumption expectations. We believe these abnormalities might be due to forcing State in.



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?

Looking at the summary of our final.fit LASSO model, we can see that the variable Age65AndOlderPct2010 has an estimate (beta value) 3.27e-02, with a p-value of 1.2e-05 *** (very small). This suggests that there is a positive correlation between percentage of elderly people and the number of (log) deaths per 100k, since we have a positive beta value. Additionally, since we have a very small p-value, this implies that Age65AndOlderPct2010 is a significant variable in our model. Likewise, our model includes the variable HispanicPct2010 with a positive beta estimate of 8.72e-03, and a p-value that is very small: 2.9e-05 ***. Again, this suggests that there is a positive correlation between percentage of Hispanic people and the number of (log) deaths per 100k, since we have a positive beta value. Additionally, since we have a very small p-value, this implies that HispanicPct2010 is a significant variable in our model. These two variables in our LASSO model support these arguments.

In terms of African Americans, that variable was not included in our LASSO model when we used lambda.1se. However, we have also run a separate linear model of log_total_death_per100k versus BlackNonHispanicPct2010. Our one-variable linear model with BlackNonHispanicPct2010 gave a positive beta estimate of 0.01305, with a p-value <2e-16 ***. This suggests that there is a positive correlation between percentage of African American people and the number of (log) deaths per 100k, since we have a positive beta value,

though it has a pretty small magnitude. Thus, our analysis more strongly supports the claims about COVID affecting the elderly and Hispanic populations, rather than the African American population.

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.

Based on the final model, we can see that the state effect, controlling for other variables, significantly influences the log death rate per 100k. Several states show significant associations with the death rate, like California and Washington, among other states. For instance, California has a negative coefficient, indicating that, on average, the death rate tends to be lower in California compared to the reference state. However, some states like South Dakota and Montana have notably large positive coefficients, suggesting relatively higher death rates, with very low p-values.

In addition to the state effect, several other non-state variables show significant impact on the log death rate per 100k in the final model. Variables such as PctEmpConstruction, PctEmpMining, and PctEmpAgriculture exhibit negative coefficients, indicating that higher employment rates in these sectors are associated with lower COVID-19 death rates. This suggests that individuals in these sectors might have lower exposure to the virus or better access to healthcare resources. Moreover, factors such as population density, age demographics (Age65AndOlderPct2010, Under18Pct2010), educational variables (Ed3SomeCollegePct, Ed5CollegePlusPct), migration rates (NetMigrationRate1019), and ethnic composition (WhiteNonHispanicPct2010, HispanicPct2010) also play significant roles in shaping the death rate. For instance, higher proportions of older adults and individuals with lower educational attainment are associated with higher death rates, highlighting the importance of targeted interventions to protect vulnerable populations. Overall, these findings show the complex interaction of various sociodemographic factors influencing the death rate and emphasize the need for various policy approaches to lower the death rate.

Policy interventions to reduce the COVID death rate should be tailored to address state-specific needs, with increased help for states like Montana and South Dakota, which have higher relative death rates. Strategies should focus on enhancing healthcare infrastructure, promoting vaccination campaigns, implementing and enforcing public health measures (like mask mandates and social distancing), improving access to healthcare services (especially in rural areas/states), and providing adequate support to vulnerable populations, such as more accessible healthcare for the elderly and Hispanic populations. Additionally, targeted efforts should be made to address socioeconomic disparities, since variables like population density, education levels, and employment also play significant roles in the COVID death rate. Collaboration between federal, state, and local governments, as well as community stakeholders, would help to successfully implement these interventions.

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

One variable that we definitely missed was the African American population percentage. The model could also be improved with more COVID-specific data, such as vaccination records based on the first dose, second dose, and booster, types of vaccines administered, data regarding masking and social distancing, and more. These variables could have been included in the model produced by the LASSO analysis. Our final model also could have had an income variable, or median household income, to analyze the economic impacts on the death rate. These other variables could be more informative for our LASSO model.

ix) (Optional) Would your findings be very different if you had refined the data in some way or imputed the missing values in part ii). Check PCA lecture, section 10 for imputations via `softImpute`.

5 Executive summary

- Goal of the study: The goal of this study was to investigate the effectiveness of lockdowns in flattening the COVID curve, and to evaluate which county-level socioeconomic/demographic variables impacted the log death rate per 100,000 population using LASSO models. Throughout the project, we also show the dynamic evolution of state-level COVID cases and deaths.
- Data: The data includes: covid_county.csv, which contains county-level socioeconomic classification

data from four tables (Income, Jobs, People, and County Classifications) from the USDA Economic Research Service. Covid_rates.csv contains daily cumulative numbers on infection and fatality for each county, and the data comes from the NY Times. Covid_intervention.csv contains county-level lockdown intervention data; this dataset is a manually compiled list of the dates that interventions/lockdown policies were implemented and lifted at the county level (source: JieYingWu's github). To assemble the data, counties were uniquely identified using FIPS, with some cleaning for FIPS in NYC and only including states in the continental US. Additionally, we set the week of Jan 21, 2020 as the first week, and the dates are formatted properly. For data assembly, we merge the `TotalPopEst2019` variable from the demographic data with `covid_rates` by FIPS. NA's are replaced with zeroes. Lastly, we merge the demographic data sets by FIPS and output as `covid_county.csv`.

- Analyses: We conducted analyses of time series data (weekly and monthly) for different states, using heatmaps, spaghetti plots, and scatterplots for data visualization. We also used LASSO while forcing the factorized State variable in to analyze what variables impacted the response variable, which we put in log scale (log_total_death_per100k). Using lambda.1se, Anova, and backward elimination, we minimized our model, checked the linear model assumptions, and analyzed our summary of the model.
- Methods used: The methods we use include LASSO model selection (forcing in certain variables, using penalty modification), k-fold cross validation, backward elimination, Anova analysis, summary statistics for linear and LASSO models, residual and Q-Q plots for linear model assumption analysis, and lubridate for time series analysis. We also use spaghetti plots, scatterplots, and heatmaps for data visualization. Packages: lm(), Anova, regsubsets(), glmnet(), cv.glmnet().
- Findings: Based on the final model, we can see that the state effect, controlling for other variables, significantly influences the log death rate per 100k. Several states show significant associations with the death rate, like California and Washington, among other states. For instance, California has a negative coefficient, with lower relative death rates, while some states like South Dakota and Montana have notably large positive coefficients, suggesting relatively higher death rates. Several other non-state variables show significant impact on the log death rate, such as PctEmpConstruction, PctEmpMining, and PctEmpAgriculture which exhibit negative coefficients, indicating that higher employment rates in these sectors are associated with lower COVID-19 death rates. Factors such as population density, age demographics (Age65AndOlderPct2010, Under18Pct2010), educational variables (Ed3SomeCollegePct, Ed5CollegePlusPct), migration rates (NetMigrationRate1019), and ethnic composition (WhiteNonHispanicPct2010, HispanicPct2010) also play significant roles in shaping the death rate. For instance, higher proportions of older adults and individuals with lower educational attainment are associated with higher death rates. Policy interventions to reduce the COVID death rate should be tailored to address state-specific needs, with increased help for states like Montana and South Dakota, which have higher relative death rates. Strategies should focus on enhancing healthcare infrastructure, promoting vaccination campaigns, implementing and enforcing public health measures (masks/social distancing), improving access to healthcare services (especially in rural areas/states), and providing adequate support to vulnerable populations, such as more accessible healthcare for the elderly and Hispanic populations.
- Limitations: To improve our model, we could have included some extra information, like the African American population percentage. The model could also be improved with more COVID-specific data, such as vaccination records based on the first dose, second dose, and booster, types of vaccines administered, data regarding masking and social distancing, and more. Our final model also could have had an income variable, or median household income, to analyze the economic impacts on the death rate. These other variables could be more informative for our LASSO model.

6 Appendix 0: Intermediate Results

```
summary(final.fit)
```

```
##
```

```
## Call:
## lm(formula = log_total_death_per100k ~ State.x + PctEmpConstruction +
##     PctEmpMining + PctEmpAgriculture + PopDensity2010 + Age65AndOlderPct2010 +
##     Under18Pct2010 + Ed3SomeCollegePct + Ed5CollegePlusPct +
##     NetMigrationRate1019 + NaturalChangeRate1019 + WhiteNonHispanicPct2010 +
##     HispanicPct2010 + Type_2015_Update, data = county_data_sub)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.804 -0.252  0.064  0.374  3.565
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              4.76e+00   2.79e-01   17.11  < 2e-16 ***
## State.xAR                5.10e-02   1.32e-01    0.39  0.70006
## State.xAZ                2.53e-01   2.33e-01    1.09  0.27604
## State.xCA               -8.33e-01   1.52e-01   -5.47  4.9e-08 ***
## State.xCO               -3.78e-01   1.49e-01   -2.54  0.01118 *
## State.xCT                1.06e-01   2.95e-01    0.36  0.72048
## State.xDC                4.16e-01   7.92e-01    0.53  0.59907
## State.xDE               -7.96e-02   4.60e-01   -0.17  0.86264
## State.xFL               -1.22e-02   1.40e-01   -0.09  0.93023
## State.xGA                5.36e-02   1.14e-01    0.47  0.63796
## State.xIA                2.55e-01   1.31e-01    1.95  0.05145 .
## State.xID               -3.27e-01   1.62e-01   -2.01  0.04420 *
## State.xIL                1.73e-01   1.29e-01    1.34  0.17943
## State.xIN               -8.41e-02   1.31e-01   -0.64  0.52126
## State.xKS               -6.48e-01   1.34e-01   -4.83  1.5e-06 ***
## State.xKY               -7.60e-01   1.24e-01   -6.12  1.1e-09 ***
## State.xLA                3.50e-01   1.39e-01    2.53  0.01149 *
## State.xMA                1.98e-01   2.34e-01    0.84  0.39911
## State.xMD               -1.40e-01   1.87e-01   -0.75  0.45584
## State.xME               -1.41e+00   2.20e-01   -6.40  1.8e-10 ***
## State.xMI               -1.56e-01   1.33e-01   -1.17  0.24069
## State.xMN               -1.98e-01   1.34e-01   -1.48  0.13977
## State.xMO               -3.92e-01   1.25e-01   -3.13  0.00174 **
## State.xMS                2.73e-01   1.29e-01    2.11  0.03502 *
## State.xMT                5.72e-01   1.53e-01    3.73  0.00019 ***
## State.xNC               -3.80e-01   1.24e-01   -3.06  0.00222 **
## State.xND                8.83e-01   1.59e-01    5.56  3.0e-08 ***
## State.xNE               -4.17e-01   1.39e-01   -3.00  0.00271 **
## State.xNH               -8.60e-01   2.67e-01   -3.22  0.00131 **
## State.xNJ                2.47e-01   2.00e-01    1.24  0.21681
## State.xNM               -6.82e-01   1.86e-01   -3.66  0.00026 ***
## State.xNV               -6.20e-01   2.22e-01   -2.79  0.00523 **
## State.xNY               -4.19e-01   1.45e-01   -2.90  0.00378 **
## State.xOH               -5.47e-01   1.32e-01   -4.15  3.4e-05 ***
## State.xOK               -3.84e-01   1.34e-01   -2.87  0.00418 **
## State.xOR               -8.87e-01   1.71e-01   -5.20  2.1e-07 ***
## State.xPA               -4.65e-03   1.41e-01   -0.03  0.97369
## State.xRI               -2.35e-01   3.65e-01   -0.64  0.52000
## State.xSC               -1.97e-02   1.50e-01   -0.13  0.89572
## State.xSD                8.47e-01   1.46e-01    5.79  7.6e-09 ***
## State.xTN                5.01e-02   1.28e-01    0.39  0.69469
```

```
## State.xTX                1.47e-01   1.24e-01    1.19  0.23557
## State.xUT               -1.08e+00   1.89e-01   -5.71  1.3e-08 ***
## State.xVA               -5.11e-01   1.19e-01   -4.31  1.7e-05 ***
## State.xVT               -2.17e+00   2.34e-01   -9.30  < 2e-16 ***
## State.xWA               -8.71e-01   1.65e-01   -5.28  1.4e-07 ***
## State.xWI               -1.62e-01   1.37e-01   -1.18  0.23723
## State.xWV               -7.12e-01   1.49e-01   -4.78  1.8e-06 ***
## State.xWY                1.67e-01   2.01e-01    0.83  0.40700
## PctEmpConstruction      -5.03e-02   6.72e-03   -7.49  9.2e-14 ***
## PctEmpMining            -2.24e-02   5.13e-03   -4.36  1.4e-05 ***
## PctEmpAgriculture       -3.97e-02   3.30e-03  -12.03  < 2e-16 ***
## PopDensity2010           2.32e-05   8.94e-06    2.59  0.00952 **
## Age65AndOlderPct2010     3.27e-02   7.46e-03    4.38  1.2e-05 ***
## Under18Pct2010           5.66e-02   7.33e-03    7.72  1.6e-14 ***
## Ed3SomeCollegePct       -2.26e-02   5.10e-03   -4.43  9.7e-06 ***
## Ed5CollegePlusPct       -1.59e-02   2.25e-03   -7.05  2.1e-12 ***
## NetMigrationRate1019    -1.27e-02   2.47e-03   -5.12  3.3e-07 ***
## NaturalChangeRate1019   -4.25e-02   9.74e-03   -4.37  1.3e-05 ***
## WhiteNonHispanicPct2010 -3.22e-03   1.38e-03   -2.32  0.02014 *
## HispanicPct2010          8.72e-03   2.08e-03    4.19  2.9e-05 ***
## Type_2015_Update        -1.68e-02   8.42e-03   -2.00  0.04560 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.778 on 3043 degrees of freedom
## Multiple R-squared:  0.365,  Adjusted R-squared:  0.352
## F-statistic: 28.7 on 61 and 3043 DF,  p-value: <2e-16
```

```
summary(lm_afr_american)
```

```
##
## Call:
## lm(formula = log_total_death_per100k ~ BlackNonHispanicPct2010,
##     data = county_data_sub)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.416 -0.342  0.145  0.534  2.087
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               4.63771    0.01992   232.8   <2e-16 ***
## BlackNonHispanicPct2010   0.01305    0.00117    11.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.947 on 3103 degrees of freedom
## Multiple R-squared:  0.0383, Adjusted R-squared:  0.038
## F-statistic:  124 on 1 and 3103 DF,  p-value: <2e-16
```

7 Appendix 1: Data description

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

7.1 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

7.2 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

8 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.

```r
# COVID case/mortality rate data
covid_rates <- fread("data/us_counties.csv", na.strings = c("NA", "", "."))
nyc <- fread("data/nycdata.csv", na.strings = c("NA", "", "."))

# Socioeconomic data
income <- fread("data/income.csv", na.strings = c("NA", "", "."))
jobs <- fread("data/jobs.csv", na.strings = c("NA", "", "."))
people <- fread("data/people.csv", na.strings = c("NA", "", "."))
county_class <- fread("data/county_classifications.csv", na.strings = c("NA", "", "."))

# Internvention policy data
int_dates <- fread("data/intervention_dates.csv", na.strings = c("NA", "", "."))
```

8.1 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.

```r
# NYC county fips matching table
nyc_fips <- data.table(FIPS = c('36005', '36047', '36061', '36081', '36085'),
                       County = c("BX", "BK", "MN", "QN", "SI"))

# nyc case
nyc_case <- nyc[,.(date = as.Date(date_of_interest, "%m/%d/%Y"),
                   BX = BX_CASE_COUNT,
                   BK = BK_CASE_COUNT,
                   MN = MN_CASE_COUNT,
                   QN = QN_CASE_COUNT,
                   SI = SI_CASE_COUNT)]

nyc_case %<>%
  pivot_longer(cols = BX:SI,
               names_to = "County",
               values_to = "cases") %>%
  arrange(date) %>%
  group_by(County) %>%
  mutate(cum_cases = cumsum(cases))

# nyc death
nyc_death <- nyc[,.(date = as.Date(date_of_interest, "%m/%d/%Y"),
                   BX = BX_DEATH_COUNT,
                   BK = BK_DEATH_COUNT,
                   MN = MN_DEATH_COUNT,
                   QN = QN_DEATH_COUNT,
                   SI = SI_DEATH_COUNT)]

nyc_death %<>%
  pivot_longer(cols = BX:SI,
               names_to = "County",
               values_to = "deaths") %>%
  arrange(date) %>%
  group_by(County) %>%
  mutate(cum_deaths = cumsum(deaths))
```

```
nyc_rates <- merge(nyc_case,
                   nyc_death,
                   by = c("date", "County"),
                   all.x= T)

nyc_rates <- merge(nyc_rates,
                   nyc_fips,
                   by = "County")

nyc_rates$State <- "New York"
nyc_rates %<>%
  select(date, FIPS, County, State, cum_cases, cum_deaths) %>%
  arrange(FIPS, date)
```

8.2 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`.

```
covid_rates <- covid_rates %>%
  arrange(fips, date) %>%
  filter(!(fips %/% 1000 %in% c(2, 15, 72))) %>%
  filter(county != "New York City") %>%
  filter(!(state %in% c("Virgin Islands", "Northern Mariana Islands"))) %>%
  rename(FIPS = "fips",
         County = "county",
         State = "state",
         cum_cases = "cases",
         cum_deaths = "deaths")


covid_rates$date <- as.Date(covid_rates$date)

covid_rates <- rbind(covid_rates,
                     nyc_rates)
```

8.3 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).

```
covid_rates[, week := (interval("2020-01-19", date) %/% weeks(1)) + 1]
```

8.4 COVID infection/mortality rates

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

```
covid_rates <- merge(covid_rates[!is.na(FIPS)],
                     people[,.(FIPS = as.character(FIPS),
                               TotalPopEst2019)],
                     by = "FIPS",
                     all.x = TRUE)
```

8.5 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`.

```
covid_rates$cum_cases[is.na(covid_rates$cum_cases)] <- 0
covid_rates$cum_deaths[is.na(covid_rates$cum_deaths)] <- 0
```

```
fwrite(covid_rates %>%
         filter(week < 58) %>%
         arrange(FIPS, date),
       "data/covid_rates.csv")
```

8.6 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`.

```
int_dates <- int_dates[-1,]
date_cols <- names(int_dates)[-(1:3)]
int_dates[, (date_cols) := lapply(.SD, as.Date, origin = "0001-01-01"),
          .SDcols = date_cols]

fwrite(int_dates, "data/covid_intervention.csv")
```

8.7 Merge demographic data

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

```
countydata <-
  merge(x = income,
        y = merge(
          x = people,
          y = jobs,
          by = c("FIPS", "State", "County")),
        by = c("FIPS", "State", "County"),
        all = TRUE)


countydata <-
  merge(
    x = countydata,
    y = county_class %>% rename(FIPS = FIPStxt),
    by = c("FIPS", "State", "County"),
    all = TRUE
  )

# Check dimensions
# They are now 3279 x 208
dim(countydata)
fwrite(countydata, "data/covid_county.csv")
```