COVID-19 Case Study

Jose Cervantez Bethany Hsaio Rob Kuan

Contents

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.

```
# county-level socialeconomic information
county_data <- fread("data/covid_county.csv")
# county-level COVID case and death
covid_rate <- fread("data/covid_rates.csv")
# county-level lockdown dates
covid_intervention <- fread("data/covid_intervention.csv")
```

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.

'covid_rate' contains the following columns:

- 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

'county_data' contains the 208 columns. The data covers a wide range of socioeconomic and demographic indicators. At a broad stroke, the data can broken up into three categories:

- Socioeconomic Demographics: Details on income levels, poverty rates (including deep poverty), median household income, and per capita income.
- Jobs: Information on employment rates, changes in employment over time, sectors of employment, and the civilian labor force.
- People: Population size and density, household characteristics, migration rates, education levels, race and ethnicity percentages, and foreign-born populations.

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?

Answer:

- See plot below.
- There seems to be two irregular patterns in the data. First, there are days where the data is very noisy, with the number of new cases spiking and dropping dramatically. Second, there are days where the number of new cases is zero. This is likely due to the fact that the data is being reported by day, and some counties may not report new cases every day.
- This leads to a problem with reporting daily case information because each county may operate under different reporting schedules. For example, some counties may not be equipped to support daily reporting so it may artificially inflate the day counts for some days (and deflate the counts for other days). In addition, the noise in the data makes it difficult to discern any meaningful trends. This would also make it problematic discerning across counties.

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

Answer:

- See plot below.

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

Answer:

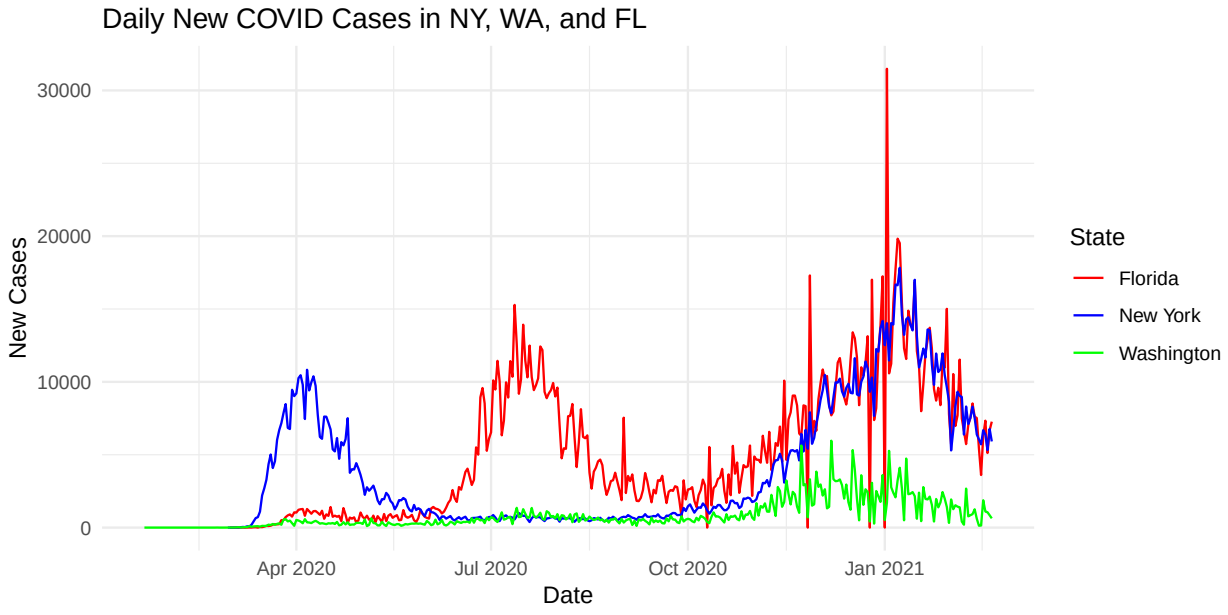
- The plot shows that the number of new cases per 100,000 people varies widely across states. For example, New York and Washington have experienced much lower rates of weekly new cases per 100,000 people than Florida. This could be due to a number of factors, including differences in population density, differences in the timing and stringency of lockdown measures, differences in the prevalence of new COVID-19 variants, and differences in the availability and distribution of vaccines.

```
###############
### part i) ###
###############

# Plot **new** COVID cases in NY, WA and FL by state and by day. Any irregular pattern? What is the big


covid_rate_filtered <- covid_rate %>%
  filter(State %in% c("New York", "Washington", "Florida"))

# Calculate daily new cases
```

```r
covid_rate_filtered <- covid_rate_filtered %>%
  arrange(State, County, date) %>%
  group_by(State, County) %>%
  mutate(new_cases = cum_cases - lag(cum_cases, default = 0))

# Aggregate daily new cases by state and date
daily_new_cases_by_state <- covid_rate_filtered %>%
  group_by(State, date) %>%
  summarise(daily_new_cases = sum(new_cases))
```

```
## `summarise()` has grouped output by 'State'. You can override using the
## `.groups` argument.
```

```r
# Plotting
ggplot(daily_new_cases_by_state, aes(x = date, y = daily_new_cases, color = State)) +
  geom_line() +
  labs(title = "Daily New COVID Cases in NY, WA, and FL",
       x = "Date",
       y = "New Cases") +
  theme_minimal() +
  scale_color_manual(values = c("New York" = "blue", "Washington" = "green", "Florida" = "red"))
```

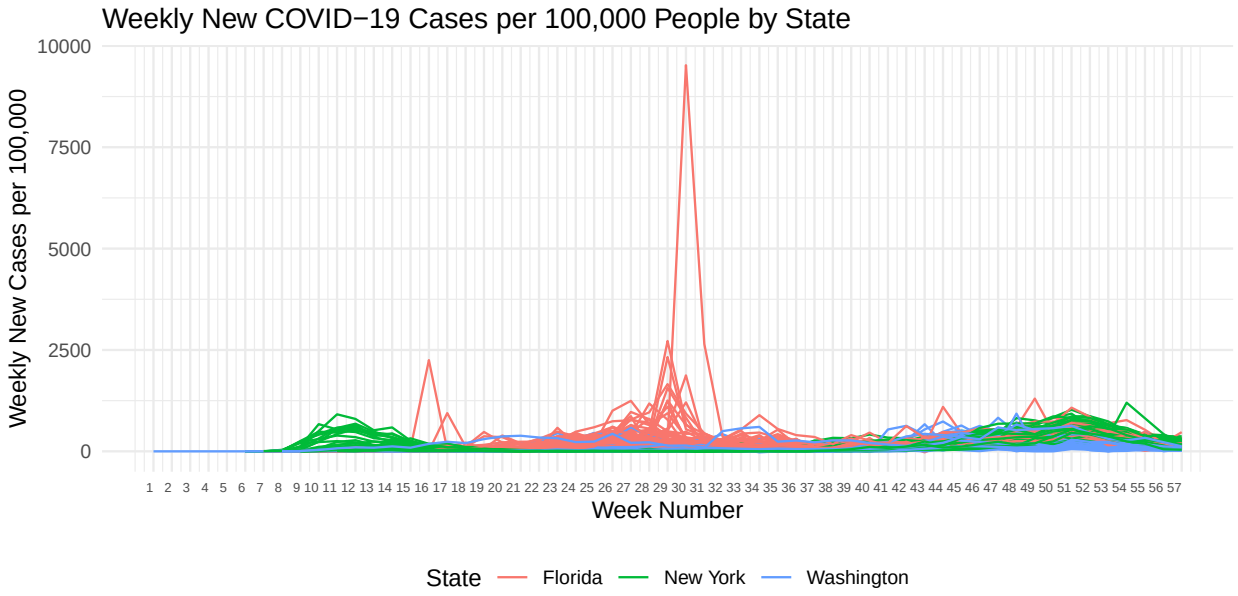



```r
##################
#### part ii) ###
##################


#Create **weekly new** cases per 100k `weekly_case_per100k`. Plot the spaghetti plots of `weekly_case_p

covid_weekly = covid_rate_filtered %>%
  group_by(State, County, week) %>%
  summarise(
```

```
    weekly_new_cases = sum(new_cases),
    TotalPop = first(TotalPopEst2019) # Assuming TotalPopEst2019 is constant for each County
  ) %>%
  ungroup()
```

```
## `summarise()` has grouped output by 'State', 'County'. You can override using
## the `.groups` argument.
```

```
# Calculate weekly new cases per 100,000 people
covid_weekly <- covid_weekly %>%
  mutate(weekly_case_per100k = (weekly_new_cases / TotalPop) * 100000)

# Plotting
ggplot(covid_weekly, aes(x = week, y = weekly_case_per100k, color = State, group = interaction(State, C
  geom_line() +
  labs(title = "Weekly New COVID-19 Cases per 100,000 People by State",
       x = "Week Number",
       y = "Weekly New Cases per 100,000") +
  scale_x_continuous(breaks = seq(1, 57, by = 1), # Change 'by = 1' to a higher number for fewer labels
                     labels = seq(1, 57, by = 1)) + # Adjust accordingly
  theme_minimal() +
  theme(legend.position = "bottom",
        axis.text.x = element_text(angle = 0, hjust = 1, size = 6)) + # Rotate x-axis labels for better
  guides(colour = guide_legend(title = "State"))
```



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.)

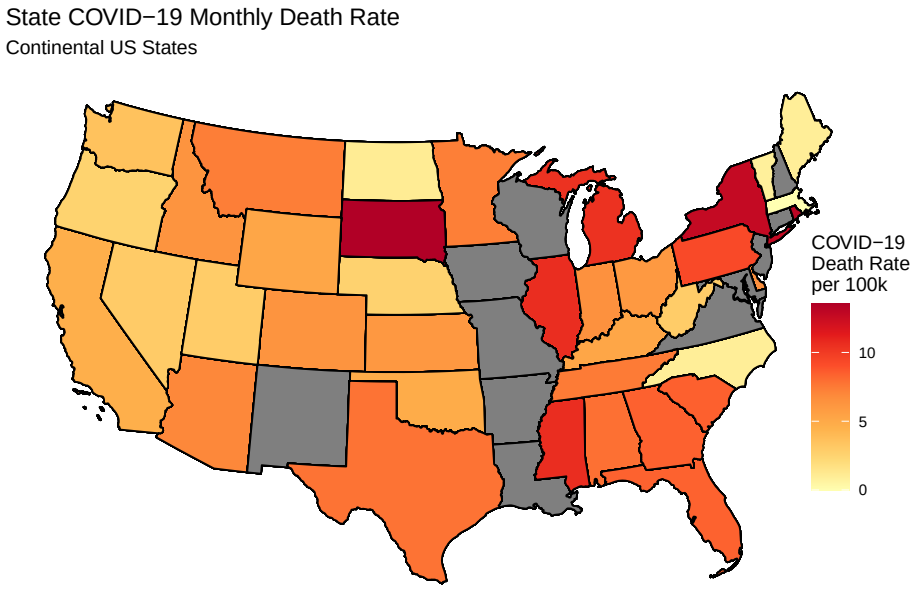
Answer:

- See plot below.

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.)

```
###############
### part i) ###
###############



# For each month in 2020, plot the monthly deaths per 100k heatmap by state on US map. Use the same col

# Sum total state population by adding up all the counties' populations
state_pop <- covid_rate %>%
  select(State, County, TotalPopEst2019) %>%
  group_by(State) %>%
  summarize(TotalStatePop = sum(TotalPopEst2019)) %>%
  ungroup()

# Calculate monthly deaths at the county level and then aggregate to state level
covid_rate_filtered_all <- covid_rate %>%
  mutate(month = month(date), year = year(date)) %>%
  filter(year == 2020) %>%
  group_by(State, month) %>%
  summarize(monthly_deaths = sum(cum_deaths), .groups = "drop") %>%
  ungroup()

# Join with state-level population estimates to calculate deaths per 100k
covid_rate_filtered_all <- covid_rate_filtered_all %>%
  left_join(state_pop, by = "State") %>%
  mutate(deaths_per100k = (monthly_deaths / TotalStatePop) * 100000)

# Use complete() to ensure all months are represented for each state, filling missing values as needed
covid_complete <- covid_rate_filtered_all %>%
  complete(State, month, fill = list(deaths_per100k = NA)) %>%
  rename(state = State)

# Calculate the color scale limits based on the 0th and 98th percentiles
max_covid_col <- quantile(covid_complete$deaths_per100k, .98, na.rm = TRUE)
min_covid_col <- quantile(covid_complete$deaths_per100k, 0, na.rm = TRUE)

plot_usmap(regions = "state",
                            data = covid_complete,
                            values = "deaths_per100k", exclude = c("Hawaii", "Alaska"), color = "black"
  scale_fill_distiller(palette = "YlOrRd", direction = 1,
                       name = "COVID-19\nDeath Rate\nper 100k",
```

```
                    limits = c(min_covid_col, max_covid_col)) +
  labs(title = "State COVID-19 Monthly Death Rate", subtitle = "Continental US States") +
  theme(legend.position = c(.92,.2))
```



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.

```
## Create crosswalk for State names and abbreviations
st_crosswalk <- tibble(State = state.name) %>%
   bind_cols(tibble(abb = state.abb)) %>%
   bind_rows(tibble(State = "District of Columbia", abb = "DC"))

## Calculate the total deaths per 100k for each county

covid_data_filtered <- covid_rate %>%
  left_join(., st_crosswalk, by = "State") %>%
  filter(date <= as.Date("2021-02-01")) %>%
  group_by(abb, County) %>%
  summarize(total_deaths = max(cum_deaths), .groups = "drop") %>%
  rename(State = abb)


# Calculate total deaths per 100k for each county
```

```r
covid_data_filtered <- covid_data_filtered %>%
  left_join(county_data %>% select(County, State, TotalPopEst2019), by = c("County", "State")) %>%
  mutate(total_death_per100k = (total_deaths / TotalPopEst2019) * 100000)

# Apply log transformation
covid_data_filtered <- covid_data_filtered %>%
  mutate(log_total_death_per100k = log(total_death_per100k + 1))

# Ensure county_data has unique identifiers for merging
county_data_unique <- county_data %>% distinct(County, State, .keep_all = TRUE)

# Merge with county_data
county_data <- county_data_unique %>%
  left_join(covid_data_filtered %>% select(County, State, total_death_per100k, log_total_death_per100k)

# Handle duplicate counties by creating the variable county_state
county_data <- county_data %>%
  mutate(county_state = paste(County, State, sep = "-"))


## STEPS TO INCLUDE STATE LEVEL COUNTS IN COUNTY DATA

# state_level_data <- county_data %>%
#   filter(!is.na(total_death_per100k)) %>% # Exclude rows with NA in total_death_per100k
#   left_join(st_crosswalk, by = c("State" = "abb")) %>%
#   group_by(State) %>%
#   summarise(
#     total_death_per100k = sum(total_death_per100k, na.rm = TRUE),
#     log_total_death_per100k = log(sum(total_death_per100k * TotalPopEst2019 / 100000, na.rm = TRUE) +
#   )
#
# # Join the state-level aggregated data back to the original dataset
# county_data <- county_data %>%
#   left_join(state_level_data, by = "State", suffix = c("", "_state")) %>%
#   mutate(
#     total_death_per100k = if_else(is.na(total_death_per100k), total_death_per100k_state, total_death_
#     log_total_death_per100k = if_else(is.na(log_total_death_per100k), log_total_death_per100k_state,
#   ) %>%
#   select(-total_death_per100k_state, -log_total_death_per100k_state) # Remove the temporary state-lev
```

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.

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

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

```r
predictive_variables <- c(
  "PovertyUnder18Pct", "Deep_Pov_All", "Deep_Pov_Children", "PovertyAllAgesPct",
  "MedHHInc", "PerCapitaInc", "PovertyAllAgesNum", "PovertyUnder18Num",
  "PctEmpChange1019", "PctEmpChange1819", "PctEmpChange0719", "PctEmpChange0710",
  "NumCivEmployed",
  "PctEmpFIRE", "PctEmpConstruction", "PctEmpTrans", "PctEmpMining", "PctEmpTrade",
  "PctEmpInformation", "PctEmpAgriculture", "PctEmpManufacturing", "PctEmpServices",
```

```
  "PctEmpGovt",
  "PopDensity2010", "LandAreaSQMiles2010", "TotalHH", "TotalOccHU", "AvgHHSize",
  "OwnHomeNum", "OwnHomePct", "NonEnglishHHPct", "HH65PlusAlonePct", "FemaleHHPct",
  "FemaleHHNum", "NonEnglishHHNum", "HH65PlusAloneNum",
  "Age65AndOlderPct2010", "Age65AndOlderNum2010", "TotalPop25Plus", "Under18Pct2010",
  "Under18Num2010", "Ed1LessThanHSPct", "Ed2HSDiplomaOnlyPct", "Ed3SomeCollegePct",
  "Ed4AssocDegreePct", "Ed5CollegePlusPct", "Ed1LessThanHSNum", "Ed2HSDiplomaOnlyNum",
  "Ed3SomeCollegeNum", "Ed4AssocDegreeNum", "Ed5CollegePlusNum", "ForeignBornPct",
  "ForeignBornEuropePct", "ForeignBornMexPct", "ForeignBornCentralSouthAmPct",
  "ForeignBornAsiaPct", "ForeignBornCaribPct", "ForeignBornAfricaPct", "ForeignBornNum",
  "ForeignBornCentralSouthAmNum", "ForeignBornEuropeNum", "ForeignBornMexNum",
  "ForeignBornAfricaNum", "ForeignBornAsiaNum", "ForeignBornCaribNum",
  "Net_International_Migration_Rate_2010_2019", "Net_International_Migration_2010_2019",
  "Net_International_Migration_2000_2010", "Immigration_Rate_2000_2010",
  "NetMigrationRate0010", "NetMigrationRate1019", "NetMigrationNum0010", "NetMigration1019",
  "NaturalChangeRate1019", "NaturalChangeRate0010", "NaturalChangeNum0010", "NaturalChange1019",
  "TotalPop2010", "TotalPopEst2010", "TotalPopEst2011", "TotalPopEst2012", "TotalPopEst2013",
  "TotalPopEst2014", "TotalPopEst2015", "TotalPopEst2016", "TotalPopEst2017", "TotalPopEst2018",
  "TotalPopEst2019", "TotalPopACS", "TotalPopEstBase2010",
  "NonHispanicAsianPopChangeRate0010", "PopChangeRate1819", "PopChangeRate1019",
  "PopChangeRate0010", "NonHispanicNativeAmericanPopChangeRate0010",
  "HispanicPopChangeRate0010", "MultipleRacePopChangeRate0010",
  "NonHispanicWhitePopChangeRate0010", "NonHispanicBlackPopChangeRate0010",
  "MultipleRacePct2010", "WhiteNonHispanicPct2010", "NativeAmericanNonHispanicPct2010",
  "BlackNonHispanicPct2010", "AsianNonHispanicPct2010", "HispanicPct2010",
  "MultipleRaceNum2010", "WhiteNonHispanicNum2010", "BlackNonHispanicNum2010",
  "NativeAmericanNonHispanicNum2010", "AsianNonHispanicNum2010", "HispanicNum2010",
  "Type_2015_Recreation_NO", "Type_2015_Farming_NO", "Type_2015_Mining_NO",
  "Type_2015_Government_NO", "Type_2015_Update", "Type_2015_Manufacturing_NO",
  "Type_2015_Nonspecialized_NO", "RecreationDependent2000", "ManufacturingDependent2000",
  "FarmDependent2003", "EconomicDependence2000", "RuralUrbanContinuumCode2003"
)

orig_vars <- c("county_state", "State", "FIPS", "Deep_Pov_All", "PovertyAllAgesPct",
                      "PerCapitaInc", "UnempRate2019", "PctEmpFIRE", "PctEmpChange1819",
                      "PctEmpConstruction", "PctEmpTrans", "PctEmpMining", "PctEmpTrade",
                      "PctEmpInformation", "PctEmpAgriculture", "PctEmpManufacturing",
                      "PctEmpServices", "PopDensity2010", "TotalOccHU", "LandAreaSQMiles2010",
                      "OwnHomePct", "Age65AndOlderPct2010", "TotalPop25Plus", "Under18Pct2010",
                      "Ed2HSDiplomaOnlyPct", "Ed3SomeCollegePct", "Ed4AssocDegreePct",
                      "Ed5CollegePlusPct", "ForeignBornPct",
                      "Net_International_Migration_Rate_2010_2019", "NetMigrationRate1019",
                      "NaturalChangeRate1019", "Metro2013", "TotalPopEst2019",
                      "WhiteNonHispanicPct2010", "NativeAmericanNonHispanicPct2010",
                      "BlackNonHispanicPct2010", "AsianNonHispanicPct2010", "HispanicPct2010",
                      "Type_2015_Update", "Low_Education_2015_update",
                      "RuralUrbanContinuumCode2013", "UrbanInfluenceCode2013",
                      "Perpov_1980_0711", "HiCreativeClass2000", "HiAmenity",
                      "Retirement_Destination_2015_Update", "RecreationDependent2000",
                      "log_total_death_per100k")

all_vars <- union(predictive_variables, orig_vars)
```

```r
county_data_sub <- county_data %>%
  select(all_of(all_vars))

colSums(is.na(county_data_sub))
```

```
##                          PovertyUnder18Pct
##                                         84
##                               Deep_Pov_All
##                                          5
##                          Deep_Pov_Children
##                                          6
##                          PovertyAllAgesPct
##                                         84
##                                  MedHHInc
##                                         84
##                               PerCapitaInc
##                                          5
##                          PovertyAllAgesNum
##                                         84
##                          PovertyUnder18Num
##                                         84
##                           PctEmpChange1019
##                                          6
##                           PctEmpChange1819
##                                          6
##                           PctEmpChange0719
##                                         11
##                           PctEmpChange0710
##                                         11
##                             NumCivEmployed
##                                          6
##                                  PctEmpFIRE
##                                          6
##                         PctEmpConstruction
##                                          6
##                                 PctEmpTrans
##                                          6
##                                PctEmpMining
##                                          6
##                                 PctEmpTrade
##                                          6
##                           PctEmpInformation
##                                          6
##                           PctEmpAgriculture
##                                          6
##                         PctEmpManufacturing
##                                          6
##                              PctEmpServices
##                                          6
##                                 PctEmpGovt
##                                          6
##                              PopDensity2010
##                                          5
```

```
##                       LandAreaSQMiles2010
##                                         5
##                                   TotalHH
##                                         5
##                                TotalOccHU
##                                         5
##                                 AvgHHSize
##                                         5
##                                OwnHomeNum
##                                         5
##                                OwnHomePct
##                                         5
##                            NonEnglishHHPct
##                                         5
##                           HH65PlusAlonePct
##                                         5
##                                FemaleHHPct
##                                         5
##                                FemaleHHNum
##                                         5
##                             NonEnglishHHNum
##                                         5
##                           HH65PlusAloneNum
##                                         5
##                         Age65AndOlderPct2010
##                                         5
##                         Age65AndOlderNum2010
##                                         5
##                               TotalPop25Plus
##                                         5
##                               Under18Pct2010
##                                         5
##                               Under18Num2010
##                                         5
##                            Ed1LessThanHSPct
##                                         5
##                          Ed2HSDiplomaOnlyPct
##                                         5
##                            Ed3SomeCollegePct
##                                         5
##                           Ed4AssocDegreePct
##                                         5
##                           Ed5CollegePlusPct
##                                         5
##                             Ed1LessThanHSNum
##                                         5
##                          Ed2HSDiplomaOnlyNum
##                                         5
##                            Ed3SomeCollegeNum
##                                         5
##                           Ed4AssocDegreeNum
##                                         5
##                           Ed5CollegePlusNum
##                                         5
```

```
##                             ForeignBornPct
##                                         83
##                         ForeignBornEuropePct
##                                         83
##                            ForeignBornMexPct
##                                         83
##                  ForeignBornCentralSouthAmPct
##                                         83
##                           ForeignBornAsiaPct
##                                         83
##                          ForeignBornCaribPct
##                                         83
##                         ForeignBornAfricaPct
##                                         83
##                              ForeignBornNum
##                                         83
##                  ForeignBornCentralSouthAmNum
##                                         83
##                           ForeignBornEuropeNum
##                                         83
##                            ForeignBornMexNum
##                                         83
##                         ForeignBornAfricaNum
##                                         83
##                           ForeignBornAsiaNum
##                                         83
##                          ForeignBornCaribNum
##                                         83
## Net_International_Migration_Rate_2010_2019
##                                         83
##      Net_International_Migration_2010_2019
##                                         83
##      Net_International_Migration_2000_2010
##                                         93
##                  Immigration_Rate_2000_2010
##                                         93
##                          NetMigrationRate0010
##                                         93
##                          NetMigrationRate1019
##                                         83
##                           NetMigrationNum0010
##                                         93
##                              NetMigration1019
##                                         83
##                         NaturalChangeRate1019
##                                         83
##                         NaturalChangeRate0010
##                                         93
##                          NaturalChangeNum0010
##                                         93
##                             NaturalChange1019
##                                         83
##                                TotalPop2010
##                                          5
```

```
##                             TotalPopEst2010
##                                           5
##                             TotalPopEst2011
##                                           5
##                             TotalPopEst2012
##                                           5
##                             TotalPopEst2013
##                                           5
##                             TotalPopEst2014
##                                           5
##                             TotalPopEst2015
##                                           5
##                             TotalPopEst2016
##                                           5
##                             TotalPopEst2017
##                                           5
##                             TotalPopEst2018
##                                           5
##                             TotalPopEst2019
##                                           5
##                                TotalPopACS
##                                           5
##                          TotalPopEstBase2010
##                                           5
##          NonHispanicAsianPopChangeRate0010
##                                          50
##                            PopChangeRate1819
##                                          83
##                            PopChangeRate1019
##                                           5
##                            PopChangeRate0010
##                                          11
## NonHispanicNativeAmericanPopChangeRate0010
##                                          34
##                   HispanicPopChangeRate0010
##                                          11
##               MultipleRacePopChangeRate0010
##                                          13
##          NonHispanicWhitePopChangeRate0010
##                                          11
##          NonHispanicBlackPopChangeRate0010
##                                          72
##                          MultipleRacePct2010
##                                           5
##                     WhiteNonHispanicPct2010
##                                           5
##           NativeAmericanNonHispanicPct2010
##                                           5
##                     BlackNonHispanicPct2010
##                                           5
##                     AsianNonHispanicPct2010
##                                           5
##                              HispanicPct2010
##                                           5
```

```
##                       MultipleRaceNum2010
##                                         5
##                    WhiteNonHispanicNum2010
##                                         5
##                    BlackNonHispanicNum2010
##                                         5
##             NativeAmericanNonHispanicNum2010
##                                         5
##                    AsianNonHispanicNum2010
##                                         5
##                            HispanicNum2010
##                                         5
##                    Type_2015_Recreation_NO
##                                       141
##                       Type_2015_Farming_NO
##                                       141
##                       Type_2015_Mining_NO
##                                       141
##                    Type_2015_Government_NO
##                                       141
##                          Type_2015_Update
##                                       141
##                  Type_2015_Manufacturing_NO
##                                       141
##                 Type_2015_Nonspecialized_NO
##                                       141
##                    RecreationDependent2000
##                                       143
##                  ManufacturingDependent2000
##                                       143
##                            FarmDependent2003
##                                       143
##                       EconomicDependence2000
##                                       143
##                 RuralUrbanContinuumCode2003
##                                        60
##                               county_state
##                                         0
##                                      State
##                                         0
##                                       FIPS
##                                         0
##                              UnempRate2019
##                                         6
##                                 Metro2013
##                                        63
##                   Low_Education_2015_update
##                                       141
##                 RuralUrbanContinuumCode2013
##                                        63
##                      UrbanInfluenceCode2013
##                                        63
##                            Perpov_1980_0711
##                                       141
```

```
##                         HiCreativeClass2000
##                                         145
##                                    HiAmenity
##                                         176
##          Retirement_Destination_2015_Update
##                                         141
##                      log_total_death_per100k
##                                         203
```

```
county_data_sub <- na.omit(county_data_sub)

county_data_sub <- as.data.frame(county_data_sub)

y_col <- grep("log_total_death_per100k", colnames(county_data_sub))
```

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.

```
set.seed(42)
y_col <- grep("log_total_death_per100k", colnames(county_data_sub))
Y <- county_data_sub[, y_col] # extract Y
X <- model.matrix(log_total_death_per100k~., data=county_data_sub)[, -1]

state_cols <- grep("State", colnames(X))
penalty_vec <- rep(1, ncol(X))
penalty_vec[state_cols] <- 0

fit.fl.force.cv <- cv.glmnet(X, Y, alpha=1, nfolds=10, intercept = T,
                      penalty.factor =penalty_vec)  # force the states into the model
coef.force.min <- coef(fit.fl.force.cv, s="lambda.1se")
var.force.min <- coef.force.min[which(coef.force.min !=0),]
rownames(as.matrix(var.force.min))[-1]
```

```
##  [1] "Ed5CollegePlusPct"    "NetMigrationRate1019" "PopChangeRate1019"
##  [4] "StateAR"              "StateAZ"              "StateCA"
##  [7] "StateCO"              "StateCT"              "StateDE"
## [10] "StateFL"              "StateGA"              "StateIA"
## [13] "StateID"              "StateIL"              "StateIN"
## [16] "StateKS"              "StateKY"              "StateLA"
## [19] "StateMA"              "StateMD"              "StateME"
## [22] "StateMI"              "StateMN"              "StateMO"
## [25] "StateMS"              "StateMT"              "StateNC"
## [28] "StateND"              "StateNE"              "StateNH"
## [31] "StateNJ"              "StateNM"              "StateNV"
## [34] "StateNY"              "StateOH"              "StateOK"
## [37] "StateOR"              "StatePA"              "StateRI"
## [40] "StateSC"              "StateSD"              "StateTN"
## [43] "StateTX"              "StateUT"              "StateVA"
## [46] "StateVT"              "StateWA"              "StateWI"
## [49] "StateWV"              "StateWY"
```

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

```r
coef.min <- coef(fit.fl.force.cv, s="lambda.1se")
coef.min <- coef.min[which(coef.min !=0),]   # get the non=zero coefficients
var.min <- rownames(as.matrix(coef.min))[-1] # output the names without intercept
lm.input <- as.formula(paste("log_total_death_per100k", "~", paste(paste0("`", var.min, "`"), collapse =
# prepare for lm fomulae
lm.input
```

```
## log_total_death_per100k ~ Ed5CollegePlusPct + NetMigrationRate1019 +
##     PopChangeRate1019 + StateAR + StateAZ + StateCA + StateCO +
##     StateCT + StateDE + StateFL + StateGA + StateIA + StateID +
##     StateIL + StateIN + StateKS + StateKY + StateLA + StateMA +
##     StateMD + StateME + StateMI + StateMN + StateMO + StateMS +
##     StateMT + StateNC + StateND + StateNE + StateNH + StateNJ +
##     StateNM + StateNV + StateNY + StateOH + StateOK + StateOR +
##     StatePA + StateRI + StateSC + StateSD + StateTN + StateTX +
##     StateUT + StateVA + StateVT + StateWA + StateWI + StateWV +
##     StateWY
```

```r
data_lm = as.data.frame(X)
data_lm$log_total_death_per100k <- Y
grep("log_total_death_per100k", colnames(data_lm))
```

```
## [1] 3143
```

```r
fit.min.lm <- lm(lm.input, data=data_lm)  # debiased or relaxed LASSO
summary(fit.min.lm)
```

```
##
## Call:
## lm(formula = lm.input, data = data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.357 -0.248  0.069  0.368  4.183
##
## Coefficients:
##                       Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           5.414766   0.096576   56.07 < 0.0000000000000002 ***
## Ed5CollegePlusPct    -0.013576   0.001902   -7.14  0.00000000000119940 ***
## NetMigrationRate1019 -0.014626   0.004781   -3.06               0.0022 **
## PopChangeRate1019    -0.000241   0.004369   -0.06               0.9561
## StateAR              -0.053692   0.123928   -0.43               0.6649
## StateAZ               0.285025   0.210513    1.35               0.1759
## StateCA              -1.017111   0.133006   -7.65  0.00000000000002772 ***
## StateCO              -0.715703   0.133609   -5.36  0.00000009135112437 ***
## StateCT               0.150293   0.277181    0.54               0.5877
## StateDE              -0.113051   0.434220   -0.26               0.7946
## StateFL              -0.072733   0.129212   -0.56               0.5735
## StateGA              -0.044809   0.107797   -0.42               0.6777
## StateIA               0.061631   0.116741    0.53               0.5976
## StateID              -0.847484   0.145769   -5.81  0.00000000676586805 ***
```

```
## StateIL               0.014634   0.116533    0.13               0.9001
## StateIN              -0.168066   0.118186   -1.42               0.1551
## StateKS              -1.012061   0.116398   -8.69 < 0.0000000000000002 ***
## StateKY              -0.895773   0.112537   -7.96  0.00000000000000244 ***
## StateLA               0.117294   0.129539    0.91               0.3653
## StateMA               0.102044   0.219558    0.46               0.6421
## StateMD              -0.259652   0.176448   -1.47               0.1412
## StateME              -1.676534   0.205370   -8.16  0.00000000000000048 ***
## StateMI              -0.300165   0.121007   -2.48               0.0132 *
## StateMN              -0.500057   0.119940   -4.17  0.00003145304302617 ***
## StateMO              -0.551843   0.113129   -4.88  0.00000112913476408 ***
## StateMS               0.219213   0.121630    1.80               0.0716 .
## StateMT              -0.313294   0.139384   -2.25               0.0247 *
## StateNC              -0.489511   0.116380   -4.21  0.00002675950822453 ***
## StateND               0.293580   0.144361    2.03               0.0421 *
## StateNE              -0.572224   0.124726   -4.59  0.00000466702572401 ***
## StateNH              -1.054982   0.250660   -4.21  0.00002644719296680 ***
## StateNJ               0.418308   0.187066    2.24               0.0254 *
## StateNM              -0.405873   0.161870   -2.51               0.0122 *
## StateNV              -0.903275   0.210076   -4.30  0.00001765990054290 ***
## StateNY              -0.540315   0.133792   -4.04  0.00005518941448218 ***
## StateOH              -0.573246   0.119236   -4.81  0.00000160444966380 ***
## StateOK              -0.603446   0.123179   -4.90  0.00000101615017655 ***
## StateOR              -1.174480   0.154025   -7.63  0.00000000000003273 ***
## StatePA               0.017914   0.127435    0.14               0.8882
## StateRI              -0.200606   0.343196   -0.58               0.5589
## StateSC               0.023906   0.140972    0.17               0.8653
## StateSD               0.352583   0.135298    2.61               0.0092 **
## StateTN               0.015022   0.118117    0.13               0.8988
## StateTX               0.094222   0.102483    0.92               0.3580
## StateUT              -1.318602   0.170165   -7.75  0.00000000000001269 ***
## StateVA              -0.712320   0.116944   -6.09  0.00000000126833425 ***
## StateVT              -2.497953   0.217785  -11.47 < 0.0000000000000002 ***
## StateWA              -1.305311   0.150484   -8.67 < 0.0000000000000002 ***
## StateWI              -0.387371   0.125452   -3.09               0.0020 **
## StateWV              -0.846208   0.133891   -6.32  0.00000000030124072 ***
## StateWY              -0.579734   0.178414   -3.25               0.0012 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.735 on 2918 degrees of freedom
## Multiple R-squared:  0.328,  Adjusted R-squared:  0.317
## F-statistic: 28.5 on 50 and 2918 DF,  p-value: <0.0000000000000002
```

```
Anova(fit.min.lm)
```

```
## Anova Table (Type II tests)
##
## Response: log_total_death_per100k
##                       Sum Sq   Df F value               Pr(>F)
## Ed5CollegePlusPct         27    1   50.94  0.00000000000119940 ***
## NetMigrationRate1019       5    1    9.36               0.0022 **
## PopChangeRate1019          0    1    0.00               0.9561
## StateAR                    0    1    0.19               0.6649
```

```
## StateAZ                    1    1    1.83              0.1759
## StateCA                   32    1   58.48  0.00000000000002772 ***
## StateCO                   15    1   28.69  0.00000009135112437 ***
## StateCT                    0    1    0.29              0.5877
## StateDE                    0    1    0.07              0.7946
## StateFL                    0    1    0.32              0.5735
## StateGA                    0    1    0.17              0.6777
## StateIA                    0    1    0.28              0.5976
## StateID                   18    1   33.80  0.00000000676586805 ***
## StateIL                    0    1    0.02              0.9001
## StateIN                    1    1    2.02              0.1551
## StateKS                   41    1   75.60 < 0.0000000000000002 ***
## StateKY                   34    1   63.36  0.00000000000000244 ***
## StateLA                    0    1    0.82              0.3653
## StateMA                    0    1    0.22              0.6421
## StateMD                    1    1    2.17              0.1412
## StateME                   36    1   66.64  0.00000000000000048 ***
## StateMI                    3    1    6.15              0.0132 *
## StateMN                    9    1   17.38  0.00003145304302617 ***
## StateMO                   13    1   23.79  0.00000112913476408 ***
## StateMS                    2    1    3.25              0.0716 .
## StateMT                    3    1    5.05              0.0247 *
## StateNC                   10    1   17.69  0.00002675950822452 ***
## StateND                    2    1    4.14              0.0421 *
## StateNE                   11    1   21.05  0.00000466702572401 ***
## StateNH                   10    1   17.71  0.00002644719296680 ***
## StateNJ                    3    1    5.00              0.0254 *
## StateNM                    3    1    6.29              0.0122 *
## StateNV                   10    1   18.49  0.00001765990054290 ***
## StateNY                    9    1   16.31  0.00005518941448217 ***
## StateOH                   12    1   23.11  0.00000160444966380 ***
## StateOK                   13    1   24.00  0.00000101615017655 ***
## StateOR                   31    1   58.14  0.00000000000003273 ***
## StatePA                    0    1    0.02              0.8882
## StateRI                    0    1    0.34              0.5589
## StateSC                    0    1    0.03              0.8653
## StateSD                    4    1    6.79              0.0092 **
## StateTN                    0    1    0.02              0.8988
## StateTX                    0    1    0.85              0.3580
## StateUT                   32    1   60.05  0.00000000000001269 ***
## StateVA                   20    1   37.10  0.00000000126833425 ***
## StateVT                   71    1  131.56 < 0.0000000000000002 ***
## StateWA                   41    1   75.24 < 0.0000000000000002 ***
## StateWI                    5    1    9.53              0.0020 **
## StateWV                   22    1   39.94  0.00000000030124072 ***
## StateWY                    6    1   10.56              0.0012 **
## Residuals               1575 2918
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
fit2 <- update(fit.min.lm, .~. -PopChangeRate1019, data_lm)
Anova(fit2)
```

```
## Anova Table (Type II tests)
```

```
##
## Response: log_total_death_per100k
##                       Sum Sq   Df F value               Pr(>F)
## Ed5CollegePlusPct        31    1   57.39  0.00000000000004768 ***
## NetMigrationRate1019     24    1   44.72  0.00000000002712146 ***
## StateAR                   0    1    0.19               0.6647
## StateAZ                   1    1    1.83               0.1760
## StateCA                  32    1   58.87  0.00000000000002282 ***
## StateCO                  16    1   28.73  0.00000008971418559 ***
## StateCT                   0    1    0.30               0.5864
## StateDE                   0    1    0.07               0.7941
## StateFL                   0    1    0.32               0.5739
## StateGA                   0    1    0.18               0.6734
## StateIA                   0    1    0.28               0.5976
## StateID                  19    1   34.34  0.00000000513276263 ***
## StateIL                   0    1    0.02               0.8993
## StateIN                   1    1    2.04               0.1537
## StateKS                  41    1   75.63 < 0.0000000000000002 ***
## StateKY                  34    1   63.48  0.00000000000000230 ***
## StateLA                   0    1    0.82               0.3659
## StateMA                   0    1    0.22               0.6404
## StateMD                   1    1    2.17               0.1412
## StateME                  36    1   66.85  0.00000000000000043 ***
## StateMI                   3    1    6.16               0.0131 *
## StateMN                   9    1   17.44  0.00003057819542012 ***
## StateMO                  13    1   23.82  0.00001111302612092 ***
## StateMS                   2    1    3.25               0.0716 .
## StateMT                   3    1    5.06               0.0246 *
## StateNC                  10    1   17.70  0.00002667467670399 ***
## StateND                   2    1    4.14               0.0420 *
## StateNE                  11    1   21.07  0.00000461911522730 ***
## StateNH                  10    1   17.74  0.00002611034093284 ***
## StateNJ                   3    1    5.00               0.0254 *
## StateNM                   3    1    6.30               0.0121 *
## StateNV                  10    1   18.52  0.00001732834135524 ***
## StateNY                   9    1   16.31  0.00005508780400380 ***
## StateOH                  12    1   23.15  0.00001157653263733 ***
## StateOK                  13    1   24.03  0.00000999860842684 ***
## StateOR                  31    1   58.16  0.00000000000003245 ***
## StatePA                   0    1    0.02               0.8860
## StateRI                   0    1    0.34               0.5601
## StateSC                   0    1    0.03               0.8658
## StateSD                   4    1    6.82               0.0091 **
## StateTN                   0    1    0.02               0.8985
## StateTX                   0    1    0.84               0.3581
## StateUT                  33    1   61.79  0.00000000000000532 ***
## StateVA                  20    1   37.17  0.00000000122736564 ***
## StateVT                  71    1  131.80 < 0.0000000000000002 ***
## StateWA                  41    1   75.40 < 0.0000000000000002 ***
## StateWI                   5    1    9.54               0.0020 **
## StateWV                  22    1   40.03  0.00000000028845672 ***
## StateWY                   6    1   10.61               0.0011 **
## Residuals              1575 2919
## ---
```

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

```
summary(fit2)
```

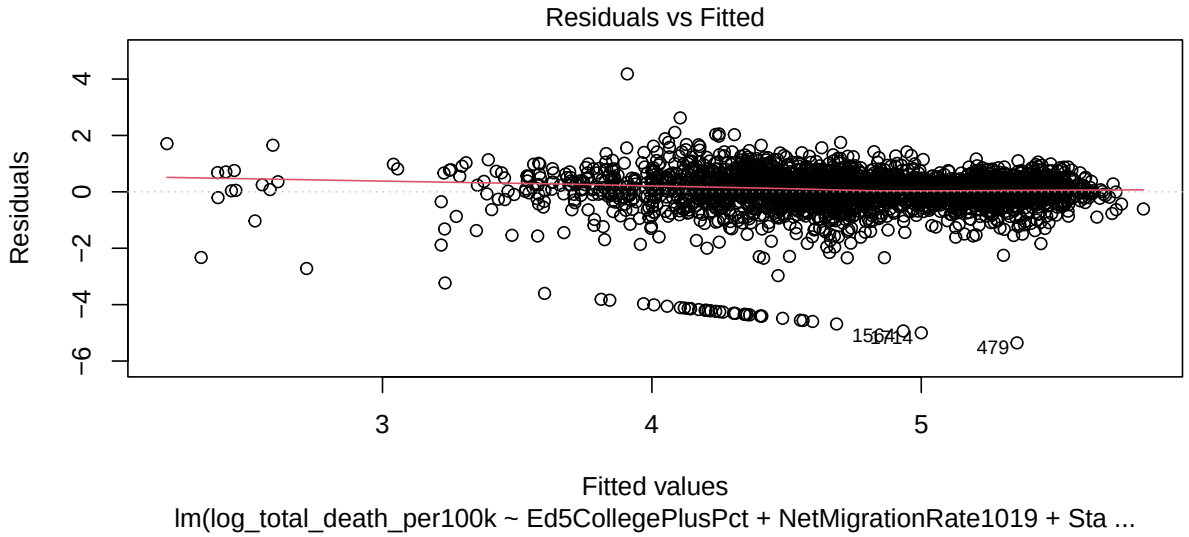
```
##
## Call:
## lm(formula = log_total_death_per100k ~ Ed5CollegePlusPct + NetMigrationRate1019 +
##     StateAR + StateAZ + StateCA + StateCO + StateCT + StateDE +
##     StateFL + StateGA + StateIA + StateID + StateIL + StateIN +
##     StateKS + StateKY + StateLA + StateMA + StateMD + StateME +
##     StateMI + StateMN + StateMO + StateMS + StateMT + StateNC +
##     StateND + StateNE + StateNH + StateNJ + StateNM + StateNV +
##     StateNY + StateOH + StateOK + StateOR + StatePA + StateRI +
##     StateSC + StateSD + StateTN + StateTX + StateUT + StateVA +
##     StateVT + StateWA + StateWI + StateWV + StateWY, data = data_lm)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.356 -0.248  0.069  0.368  4.182
##
## Coefficients:
##                      Estimate Std. Error t value             Pr(>|t|)
## (Intercept)           5.41544    0.09578   56.54 < 0.0000000000000002 ***
## Ed5CollegePlusPct    -0.01361    0.00180   -7.58  0.00000000000004768 ***
## NetMigrationRate1019 -0.01486    0.00222   -6.69  0.00000000002712146 ***
## StateAR              -0.05371    0.12391   -0.43                0.6647
## StateAZ               0.28432    0.21009    1.35                0.1760
## StateCA              -1.01764    0.13263   -7.67  0.00000000000002282 ***
## StateCO              -0.71586    0.13356   -5.36  0.00000008971418559 ***
## StateCT               0.15074    0.27702    0.54                0.5864
## StateDE              -0.11331    0.43412   -0.26                0.7941
## StateFL              -0.07264    0.12918   -0.56                0.5739
## StateGA              -0.04528    0.10743   -0.42                0.6734
## StateIA               0.06163    0.11672    0.53                0.5976
## StateID              -0.84841    0.14477   -5.86  0.00000000513276263 ***
## StateIL               0.01474    0.11650    0.13                0.8993
## StateIN              -0.16840    0.11801   -1.43                0.1537
## StateKS              -1.01207    0.11638   -8.70 < 0.0000000000000002 ***
## StateKY              -0.89598    0.11245   -7.97  0.00000000000000230 ***
## StateLA               0.11665    0.12898    0.90                0.3659
## StateMA               0.10249    0.21937    0.47                0.6404
## StateMD              -0.25963    0.17642   -1.47                0.1412
## StateME              -1.67587    0.20498   -8.18  0.00000000000000043 ***
## StateMI              -0.29984    0.12084   -2.48                0.0131 *
## StateMN              -0.50033    0.11982   -4.18  0.00003057819542012 ***
## StateMO              -0.55197    0.11309   -4.88  0.00000111302612092 ***
## StateMS               0.21890    0.12147    1.80                0.0716 .
## StateMT              -0.31335    0.13936   -2.25                0.0246 *
## StateNC              -0.48951    0.11636   -4.21  0.00002667467670399 ***
## StateND               0.29305    0.14401    2.03                0.0420 *
## StateNE              -0.57233    0.12469   -4.59  0.00000461911522730 ***
## StateNH              -1.05432    0.25033   -4.21  0.00002611034093284 ***
## StateNJ               0.41839    0.18703    2.24                0.0254 *
## StateNM              -0.40615    0.16176   -2.51                0.0121 *
```

```
## StateNV              -0.90362    0.20995   -4.30  0.00001732834135524 ***
## StateNY              -0.54016    0.13374   -4.04  0.00005508780400380 ***
## StateOH              -0.57340    0.11918   -4.81  0.00001157653263733 ***
## StateOK              -0.60360    0.12312   -4.90  0.00000999860842684 ***
## StateOR              -1.17443    0.15400   -7.63  0.00000000000003245 ***
## StatePA               0.01824    0.12727    0.14                0.8860
## StateRI              -0.19973    0.34277   -0.58                0.5601
## StateSC               0.02383    0.14094    0.17                0.8658
## StateSD               0.35193    0.13476    2.61                0.0091 **
## StateTN               0.01507    0.11809    0.13                0.8985
## StateTX               0.09360    0.10184    0.92                0.3581
## StateUT              -1.32010    0.16793   -7.86  0.00000000000000532 ***
## StateVA              -0.71201    0.11679   -6.10  0.00000000122736564 ***
## StateVT              -2.49743    0.21754  -11.48 < 0.0000000000000002 ***
## StateWA              -1.30561    0.15036   -8.68 < 0.0000000000000002 ***
## StateWI              -0.38736    0.12543   -3.09                0.0020 **
## StateWV              -0.84582    0.13369   -6.33  0.00000000028845672 ***
## StateWY              -0.58022    0.17816   -3.26                0.0011 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.735 on 2919 degrees of freedom
## Multiple R-squared:  0.328,  Adjusted R-squared:  0.317
## F-statistic: 29.1 on 49 and 2919 DF,  p-value: <0.0000000000000002
```

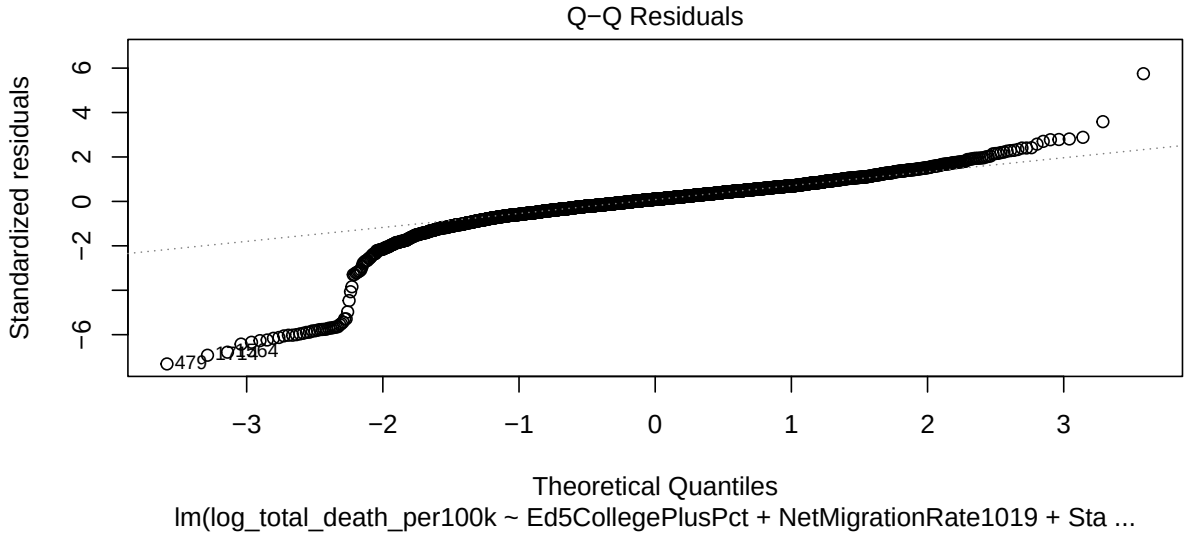
v) If necessary, reduce the model from iv) to a final model with all variables being significant at 0.05 level. Are the linear model assumptions all reasonably met?

Answer: All variables (other than states, which we have forced in) are significant at a 0.05 level. The linear model assumptions are reasonably met. Linearity appears to be satisfied as the residuals seem to follow a symmetric pattern. Homoscedasticity appears to be satisfied since the residuals are mostly evenly distributed within a band. Finally, normality appears to be reasonably satisfied as the QQ plot shows a relatively linear trend.

```
# residual plot
plot(fit2, 1)
```




```
# QQ plot
plot(fit2, 2)
```



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?

Answer: My analysis does not support these arguments. I included variables that included the proportion of the population that was above the age of 65 as well as the proportion of the population that was African American or Latino. However, LASSO did not select these variables, indicating that these variables do not affect the repsonse variable, which is the log death rate. Thus, based on my analysis, age and race do not seem to impact the COVID death rate.

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.

Answer: Based on my final model, it appears that Ed5CollegePlusPct (Percent of persons with a 4-year college degree or more, adults 25 and over, 2014-18) and NetMigrationRate1019 (the net migration rate from 2010 to 2019) are the most important non-state factors that affect the COVID death rate. Both variables have a small negative coefficient, indicating that higher levels of education and higher levels of net miration rates lead to lower COVID death rates. Controlling for other variables, the state effect sizes tend to be larger in magnitude, indicating that the state of the county has a larger impact on the COVID death rate than do Ed5CollegePlusPct and NetMigrationRate1019. The vast majority of states with significant effects have negative effects.

To reduce the COVID death rates, policymakers can consider creating interventions that target populations with a lower proportion of individuals with 4-year college degrees. This may include educational campaigns that inform citizens of how COVID spreads, how the virus spreads, and the efficacy of the vaccine. Another policy recommendation could be how to place vaccine location sites; to reach more of the population without a college education, more vaccine location sites could be placed near work locations of blue-collar workers. Since NetMigrationRate1019 is also a factor, policy makers could investigate what factors lead more people to leave a state since on average, when holding other variables constant, the more people who leave a county, the higher the COVID death rate. By understanding what factors lead to higher exit rates, policy makers can identify similarities between such counties and then create policies that address those factors.

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

Answer: To improve our model, we could incorporate interactions. It is unlikely that all of these variables are independent of each other, and by not including interactions, we may miss information on how the interaction between factors, such as age and race, affect the COVID death rate. Another piece of information omitted from our model is the political affiliation of each county. In general, states' and counties' responses to COVID were largely influenced by their political associations. Including this information would allow us to see how the death rate was impacted by a county being Democratic versus Republican.

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

Please summarize this project as follows (no more than one page):

- Goal of the study
- Data
  - Source and a brief description of the data
  - How do you assemble them together (mostly done by our team but you may present them as if you have done so)
- Analyses
- Methods used
- Findings
- Limitations

Executive Summary:

Goal of the Study:

The goal of this study is to identify factors at the county level that have the most impact on COVID death rates. By identifying such factors, we can use this information to guide what interventions policy makers create to reduce COVID death rates.

Data:

The dataset used for this analysis is assembled from several sources:

- County-level COVID-19 infection and fatality data from the New York Times, which provides daily cumulative numbers on infections and fatalities for each county in the United States. Additional data for New York City is obtained from the NYC Department of Health and Mental Hygiene.
- County-level socioeconomic data from the USDA Economic Research Service's Atlas of Rural and Small-Town America. This includes data on 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), and county classifications (rural or urban on a rural-urban continuum scale).
- Intervention policy data manually compiled by the research team, which includes the dates that interventions and lockdown policies were implemented and lifted at the county level.

The data was assembled using the following steps:

- Extracting relevant information from the NYC data and formatting it to match the structure of the county-level data.
- Filtering the data to include only counties in the continental United States and combining the NYC data with the county-level data.
- Converting the date information to a standardized format and creating a 'week' variable based on the date of the first COVID-19 case in the US.
- Merging the COVID-19 data with the county-level socioeconomic data using the FIPS (Federal Information Processing Standards) code as a unique identifier for each county.
- Handling missing data by replacing NA values in the cumulative cases and deaths columns with zeros and dropping counties with missing FIPS codes.
- Converting date columns in the intervention policy data to a standardized format.
- Merging the various socioeconomic datasets and the intervention policy data to create a comprehensive dataset for analysis.

Analyses and Findings:

To conduct our analysis, we used linear regression to understand the effect sizes of each covariate and LASSO to select the variables with the highest impact on COVID death rates. We also used backwards elimination to retain the set of variables with significant effects on COVID death rates. Through this analysis, we identified the percentage of adults age 25 and over with at least a 4-year college degree and the net migration rate from 2010-2019 as the variables with the highest impact on COVID death rates at the county level. Both of these variables had small negative effects on COVID death rates, indicating that higher proportions of adults with at least a college education and higher net migration rates are associated with lower COVID death rates. State effect sizes had larger magnitudes, indicating that the state a county is in has a larger effect on the COVID death rate than both of the aforementioned variables.

Limitations:

This analysis has several limitations. First, we did not include any interaction effects, so our analysis only investigates the effect of independent variables. This means that our analysis does not consider how interactions between variables, such as age and race, impact COVID death rates. Second, our analysis is limited by the variables available to us. One important variable omitted from the data is the political orientation of each county. Policymakers' reactions to COVID were highly polarized between political parties, and without this variable, we are unable to quantify the effect size of political orientation on COVID death rates. Finally, all of these analyses are merely correlational and we cannot conclude any causal effects. That is, although we can quantify effect sizes, we cannot conclude that any covariate causes an increase or decrease in COVID death rates.

6 Potential final projects

6.1 Does the govenment lockdown policy work in reducing covid death rates?

Please talk to Linda if interested in this.

You may use the available data together with `covid_intervention` as well as more potential data not included here.

6.2 Do vaccinations help reducing death rates from `COVID`?

Please talk to Linda if interested in this.

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.

```
# Indicator to check whether the data files exist
data_cleaned <- !(file.exists("data/covid_county.csv") &
                    file.exists("data/covid_rates.csv") &
                    file.exists("data/covid_intervention.csv"))
```

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

```
# 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.

```
# 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))
```

```r
# 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`.

```r
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")
```