COVID-19 Case Study

DATASCI 347

Group Member 1
Group Member 2
Group Member 3

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 112 million cases have been confirmed worldwide, with nearly 2.5 million associated deaths. Within the US alone, there have been over 500,000 deaths and upwards of 28 million cases reported. Governments around the world have implemented and suggested a number of policies to lessen the spread of the pandemic, including mask-wearing requirements, travel restrictions, business and school closures, and even stay-at-home orders. The global pandemic has impacted the lives of individuals in countless ways, and though many countries have begun vaccinating individuals, the long-term impact of the virus remains unclear.

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

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

There are two main goals for this case study:

1. We show the dynamic evolvement of COVID cases and COVID-related death at state level.

2. We try to figure out what county-level demographic and policy interventions are associated with mortality rate in the US. We try to construct models to find possible factors related to county-level COVID-19 mortality rates.

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

2 Data Summary

The data comes from several different sources:

- **County-level infection and fatality data** - This dataset gives daily cumulative numbers on infection and fatality for each county.

- **NYC data**

- **County-level socioeconomic data** - The following are the four relevant datasets from this site:

  - 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 - Type of county (rural or urban on a rural-urban continuum scale).

- **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 Exploratory Data Analysis

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

- `covid_county.csv`: County-level socioeconomic information that combines the above-mentioned 4 datasets: Income, Jobs, People, and 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.

First read in the data using `pandas.read_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.

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?

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?

iv) (Optional) Use `covid_intervention` to see whether the effectiveness of lockdown in flattening the curve.

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 in* `matplotlib.pyplot.clim()` *or use* `seaborn.FacetGrid()`*; use* `pandas.to_datetime()` *with* `.dt.month` *and* `.dt.year` *to extract month and year from date; use* `pandas.DataFrame.merge()` *with appropriate parameters to complete the missing months with no cases.*)

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: Use* `plotly.express.choropleth()` *with* `animation_frame` *argument for animation. Plotly recognizes state abbreviations for mapping.*)

4 COVID Factors

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 = np.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 variables from `county_data`, for you to get started. Please add any potential variables as you wish.

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

In the following analysis, you may ignore the missing values.

iv) Use LASSO (from `sklearn.linear_model.Lasso` or `sklearn.linear_model.LassoCV`) 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?

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

vi) 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?

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

viii) 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.

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

A Data Description

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

Infection and fatality data

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

Socioeconomic demographics

Income: Poverty level and household income

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

Jobs: Employment type, rate, and change

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

- `NumCivLaborforce2007-2019`: Civilian labor force, 2007-2019
- `PctEmpFIRE`: Percent of the civilian labor force 16 and over employed in finance and insurance, and real estate and rental and leasing, 2014-18
- `PctEmpConstruction`: Percent of the civilian labor force 16 and over employed in construction, 2014-18
- `PctEmpTrans`: Percent of the civilian labor force 16 and over employed in transportation, warehousing and utilities, 2014-18
- `PctEmpMining`: Percent of the civilian labor force 16 and over employed in mining, quarrying, oil and gas extraction, 2014-18
- `PctEmpTrade`: Percent of the civilian labor force 16 and over employed in wholesale and retail trade, 2014-18
- `PctEmpInformation`: Percent of the civilian labor force 16 and over employed in information services, 2014-18
- `PctEmpAgriculture`: Percent of the civilian labor force 16 and over employed in agriculture, forestry, fishing, and hunting, 2014-18
- `PctEmpManufacturing`: Percent of the civilian labor force 16 and over employed in manufacturing, 2014-18
- `PctEmpServices`: Percent of the civilian labor force 16 and over employed in services, 2014-18
- `PctEmpGovt`: Percent of the civilian labor force 16 and over employed in public administration, 2014-18

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

- `PopDensity2010`: Population density, 2010
- `LandAreaSQMiles2010`: Land area in square miles, 2010
- `TotalHH`: Total number of households, 2014-18
- `TotalOccHU`: Total number of occupied housing units, 2014-18
- `AvgHHSize`: Average household size, 2014-18
- `OwnHomeNum`: Number of owner occupied housing units, 2014-18
- `OwnHomePct`: Percent of owner occupied housing units, 2014-18
- `NonEnglishHHPct`: Percent of non-English speaking households of total households, 2014-18
- `HH65PlusAlonePct`: Percent of persons 65 or older living alone, 2014-18
- `FemaleHHPct`: Percent of female headed family households of total households, 2014-18
- `FemaleHHNum`: Number of female headed family households, 2014-18

- `NonEnglishHHNum`: Number of non-English speaking households, 2014-18
- `HH65PlusAloneNum`: Number of persons 65 years or older living alone, 2014-18
- `Age65AndOlderPct2010`: Percent of population 65 or older, 2010
- `Age65AndOlderNum2010`: Population 65 years or older, 2010
- `TotalPop25Plus`: Total population 25 and older, 2014-18 - 5-year average
- `Under18Pct2010`: Percent of population under age 18, 2010
- `Under18Num2010`: Population under age 18, 2010
- `Ed1LessThanHSPct`: Percent of persons with no high school diploma or GED, adults 25 and over, 2014-18
- `Ed2HSDiplomaOnlyPct`: Percent of persons with a high school diploma or GED only, adults 25 and over, 2014-18
- `Ed3SomeCollegePct`: Percent of persons with some college experience, adults 25 and over, 2014-18
- `Ed4AssocDegreePct`: Percent of persons with an associate's degree, adults 25 and over, 2014-18
- `Ed5CollegePlusPct`: Percent of persons with a 4-year college degree or more, adults 25 and over, 2014-18
- `Ed1LessThanHSNum`: No high school, adults 25 and over, 2014-18
- `Ed2HSDiplomaOnlyNum`: High school only, adults 25 and over, 2014-18
- `Ed3SomeCollegeNum`: Some college experience, adults 25 and over, 2014-18
- `Ed4AssocDegreeNum`: Number of persons with an associate's degree, adults 25 and over, 2014-18
- `Ed5CollegePlusNum`: College degree 4-years or more, adults 25 and over, 2014-18
- `ForeignBornPct`: Percent of total population foreign born, 2014-18
- `ForeignBornEuropePct`: Percent of persons born in Europe, 2014-18
- `ForeignBornMexPct`: Percent of persons born in Mexico, 2014-18
- `ForeignBornCentralSouthAmPct`: Percent of persons born in Central or South America, 2014-18
- `ForeignBornAsiaPct`: Percent of persons born in Asia, 2014-18
- `ForeignBornCaribPct`: Percent of persons born in the Caribbean, 2014-18
- `ForeignBornAfricaPct`: Percent of persons born in Africa, 2014-18
- `ForeignBornNum`: Number of people foreign born, 2014-18

- `ForeignBornCentralSouthAmNum`: Number of persons born in Central or South America, 2014-18
- `ForeignBornEuropeNum`: Number of persons born in Europe, 2014-18
- `ForeignBornMexNum`: Number of persons born in Mexico, 2014-18
- `ForeignBornAfricaNum`: Number of persons born in Africa, 2014-18
- `ForeignBornAsiaNum`: Number of persons born in Asia, 2014-18
- `ForeignBornCaribNum`: Number of persons born in the Caribbean, 2014-18
- `Net_International_Migration_Rate_2010_2019`: Net international migration rate, 2010-19
- `Net_International_Migration_2010_2019`: Net international migration, 2010-19
- `Net_International_Migration_2000_2010`: Net international migration, 2000-10
- `Immigration_Rate_2000_2010`: Net international migration rate, 2000-10
- `NetMigrationRate0010`: Net migration rate, 2000-10
- `NetMigrationRate1019`: Net migration rate, 2010-19
- `NetMigrationNum0010`: Net migration, 2000-10
- `NetMigration1019`: Net Migration, 2010-19
- `NaturalChangeRate1019`: Natural population change rate, 2010-19
- `NaturalChangeRate0010`: Natural population change rate, 2000-10
- `NaturalChangeNum0010`: Natural change, 2000-10
- `NaturalChange1019`: Natural population change, 2010-19
- `TotalPop2010`: Population size 4/1/2010 Census
- `TotalPopEst2010`: Population size 7/1/2010
- `TotalPopEst2011`: Population size 7/1/2011
- `TotalPopEst2012`: Population size 7/1/2012
- `TotalPopEst2013`: Population size 7/1/2013
- `TotalPopEst2014`: Population size 7/1/2014
- `TotalPopEst2015`: Population size 7/1/2015
- `TotalPopEst2016`: Population size 7/1/2016
- `TotalPopEst2017`: Population size 7/1/2017
- `TotalPopEst2018`: Population size 7/1/2018
- `TotalPopEst2019`: Population size 7/1/2019

- `TotalPopACS`: Total population, 2014-18 - 5-year average
- `TotalPopEstBase2010`: County Population estimate base 4/1/2010
- `NonHispanicAsianPopChangeRate0010`: Population change rate Non-Hispanic Asian, 2000-10
- `PopChangeRate1819`: Population change rate, 2018-19
- `PopChangeRate1019`: Population change rate, 2010-19
- `PopChangeRate0010`: Population change rate, 2000-10
- `NonHispanicNativeAmericanPopChangeRate0010`: Population change rate Non-Hispanic Native American, 2000-10
- `HispanicPopChangeRate0010`: Population change rate Hispanic, 2000-10
- `MultipleRacePopChangeRate0010`: Population change rate multiple race, 2000-10
- `NonHispanicWhitePopChangeRate0010`: Population change rate Non-Hispanic White, 2000-10
- `NonHispanicBlackPopChangeRate0010`: Population change rate Non-Hispanic African American, 2000-10
- `MultipleRacePct2010`: Percent multiple race, 2010
- `WhiteNonHispanicPct2010`: Percent Non-Hispanic White, 2010
- `NativeAmericanNonHispanicPct2010`: Percent Non-Hispanic Native American, 2010
- `BlackNonHispanicPct2010`: Percent Non-Hispanic African American, 2010
- `AsianNonHispanicPct2010`: Percent Non-Hispanic Asian, 2010
- `HispanicPct2010`: Percent Hispanic, 2010
- `MultipleRaceNum2010`: Population size multiple race, 2010
- `WhiteNonHispanicNum2010`: Population size Non-Hispanic White, 2010
- `BlackNonHispanicNum2010`: Population size Non-Hispanic African American, 2010
- `NativeAmericanNonHispanicNum2010`: Population size Non-Hispanic Native American, 2010
- `AsianNonHispanicNum2010`: Population size Non-Hispanic Asian, 2010
- `HispanicNum2010`: Population size Hispanic, 2010

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

- `Type_2015_Recreation_NO`: Recreation counties, 2015 edition
- `Type_2015_Farming_NO`: Farming-dependent counties, 2015 edition

- `Type_2015_Mining_NO`: Mining-dependent counties, 2015 edition
- `Type_2015_Government_NO`: Federal/State government-dependent counties, 2015 edition
- `Type_2015_Update`: County typology economic types, 2015 edition
- `Type_2015_Manufacturing_NO`: Manufacturing-dependent counties, 2015 edition
- `Type_2015_Nonspecialized_NO`: Nonspecialized counties, 2015 edition
- `RecreationDependent2000`: Nonmetro recreation-dependent, 1997-00
- `ManufacturingDependent2000`: Manufacturing-dependent, 1998-00
- `FarmDependent2003`: Farm-dependent, 1998-00
- `EconomicDependence2000`: Economic dependence, 1998-00
- `RuralUrbanContinuumCode2003`: Rural-urban continuum code, 2003
- `UrbanInfluenceCode2003`: Urban influence code, 2003
- `RuralUrbanContinuumCode2013`: Rural-urban continuum code, 2013
- `UrbanInfluenceCode2013`: Urban influence code, 2013
- `Noncore2013`: Nonmetro noncore, outside Micropolitan and Metropolitan, 2013
- `Micropolitan2013`: Micropolitan, 2013
- `Nonmetro2013`: Nonmetro, 2013
- `Metro2013`: Metro, 2013
- `Metro_Adjacent2013`: Nonmetro, adjacent to metro area, 2013
- `Noncore2003`: Nonmetro noncore, outside Micropolitan and Metropolitan, 2003
- `Micropolitan2003`: Micropolitan, 2003
- `Metro2003`: Metro, 2003
- `Nonmetro2003`: Nonmetro, 2003
- `NonmetroNotAdj2003`: Nonmetro, nonadjacent to metro area, 2003
- `NonmetroAdj2003`: Nonmetro, adjacent to metro area, 2003
- `Oil_Gas_Change`: Change in the value of onshore oil and natural gas production, 2000-11
- `Gas_Change`: Change in the value of onshore natural gas production, 2000-11
- `Oil_Change`: Change in the value of onshore oil production, 2000-11
- `Hipov`: High poverty counties, 2014-18
- `Perpov_1980_0711`: Persistent poverty counties, 2015 edition

- `PersistentChildPoverty_1980_2011`: Persistent child poverty counties, 2015 edition
- `PersistentChildPoverty2004`: Persistent child poverty counties, 2004
- `PersistentPoverty2000`: Persistent poverty counties, 2004
- `Low_Education_2015_update`: Low education counties, 2015 edition
- `LowEducation2000`: Low education, 2000
- `HiCreativeClass2000`: Creative class, 2000
- `HiAmenity`: High natural amenities
- `RetirementDestination2000`: Retirement destination, 1990-00
- `Low_Employment_2015_update`: Low employment counties, 2015 edition
- `Population_loss_2015_update`: Population loss counties, 2015 edition
- `Retirement_Destination_2015_Update`: Retirement destination counties, 2015 edition

B 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 first read in the table using `pandas.read_csv()`.

B.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.

B.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 floor 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`.

B.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).

B.4 COVID infection/mortality rates

Merge the `TotalPopEst2019` variable from the demographic data with `covid_rates` by FIPS using `pandas.merge()`.

B.5 NA in COVID data

`NaN` values in the `covid_rates` data set correspond to a county not having confirmed cases/deaths. We replace the `NaN` values in these columns with zeros using `fillna(0)`. FIPS for Kansas city, Missouri, Rhode Island and some others are missing. We drop them for the moment using `dropna()` and output the data up to week 57 as `covid_rates.csv`.

B.6 Formatting date in int_dates

We convert the columns representing dates in `int_dates` to Python datetime types using `pandas.to_datetime()`. We output the data as `covid_intervention.csv`.

B.7 Merge demographic data

Merge the demographic data sets by FIPS using `pandas.merge()` and output as `covid_county.csv`.