BIOS 611 Project Organization

Ning Zhang

2024-10-10

Everything has been pushed to GitHub: https://github.com/nzhang09/bios611. Since canvas doesn't support plain text submission, I copied my README file, Docker file, and Makefile in this report_canvas.Rmd for grading purpose. Thanks for understanding!

README file:

```
# # BIOS 611 Project -- COVID-19 Mortality in the U.S.
#
# Ning Zhang
#
# ning.zhang@unc.edu
#
# GitHub: https://github.com/nzhang09/bios611
#
#
# ### Dockerfile:
# The Dockerfile contains the instructions to build a Docker image for this project,
# including the base image used (i.e., rocker/verse), and instructions for installing
# additional packages or software.
#
# ### Makefile:
# The Makefile creates the datasets used in the analysis and builds the report from R Markdown.
# It simplifies the workflow by defining targets and dependencies.
#
# ### report.Rmd:
# The report.Rmd is an R Markdown file that generates a pdf report,
# which contains description of the dataset, column names, and potential project ideas.
```

Dockerfile

```
#Docker file
# FROM rocker/verse
#
# # Update package list, install man-db, and clean up
# RUN apt-get update && \
#     apt-get install -y man-db && \
#     apt-get clean && \
#     rm -rf /var/lib/apt/lists/*
```

```
#
# # Set the default command
# CMD ["/init"]
```

Makefile

```
#Makefile
# .PHONY: clean
# .PHONY: init
#
# init:
#   mkdir -p derv_data
#   mkdir -p figures
#   mkdir -p logs
#
# clean:
#   rm -rf derv_data
#   rm -rf figures
#   mkdir -p derv_data
#   mkdir -p figures
#   mkdir -p logs
#
# #This creates a complete dataset by combining COVID-19 mortality data, age distribution,
# and county-level covariates.
# #Incomplete cases are removed.
# derv_data/covid_data_all.csv: orgdata/usa_population_age_county.csv orgdata/county_dat.csv
# orgdata/nyt_reported_covid19_june23.parquet covid_data.R
#   Rscript covid_data.R
#
# report.pdf: report.Rmd
#   Rscript -e "rmarkdown::render('report.Rmd')"
```

Project

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.5.0     v purrr   1.0.2
## v tibble  3.2.1     v dplyr   1.1.4
## v tidyr   1.3.1     v stringr 1.5.1
## v readr   2.1.2     v forcats 1.0.0
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
covid_dat_all<-read_csv("derv_data/covid_data_all.csv")
```

```
## Rows: 2978 Columns: 71
```

```
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (6): fips, county, state, abbreviation, NAME, Description
## dbl (65): population, deaths_per_100k, tpopE, twhiteE, tblackE, tasianE, thl...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
colnames(covid_dat_all)
```

```
##  [1] "fips"              "county"            "state"
##  [4] "population"        "abbreviation"      "deaths_per_100k"
##  [7] "NAME"              "tpopE"             "twhiteE"
## [10] "tblackE"           "tasianE"           "thlE"
## [13] "tforeignbornE"     "tpovertyE"         "tnoinsurance19E"
## [16] "tnoinsurance1934E" "tnoinsurance3564E" "tnoinsurance65E"
## [19] "temploypopE"       "temployedE"        "tworkerE"
## [22] "twfhE"             "tedupopE"          "tnohsE"
## [25] "thhE"              "tmarriedhhE"       "tfemalehhE"
## [28] "tmalehhE"          "tmalealonehhE"     "tfemalealonehhE"
## [31] "tpubassishhE"      "tbbinternethhE"    "medhhincomeE"
## [34] "giniE"             "white"             "black"
## [37] "asian"             "hispaniclatino"    "foreignborn"
## [40] "poverty"           "noinsurance"       "highschool"
## [43] "employed"          "workfromhome"      "marriedhh"
## [46] "femalehh"          "malehh"            "alonehh"
## [49] "pubassishh"        "bbinthh"           "popdensity"
## [52] "RUCC_2013"         "Description"       "crudemortality"
## [55] "0_4"               "10_14"             "15_19"
## [58] "20_24"             "25_29"             "30_34"
## [61] "35_39"             "40_44"             "45_49"
## [64] "5_9"               "50_54"             "55_59"
## [67] "60_64"             "65_69"             "70_74"
## [70] "75_79"             "80_110"
```

Dataset

This project will use publicly available county-level COVID-19 data to investigate the relationship between population factors (e.g., age, race and ethnicity, population density), socioeconomic status (e.g., education, poverty, GINI index, employment), health care access (e.g., insurance, urbanicitiy, crude mortality), and COVID-19 mortality.

Weekly data on per capita COVID-19 mortality in each county was downloaded from the New York Times. Data on county-level population age distribution was pulled from CDC Vintage 2020 Bridged-Race Postcensal Population Estimates. Data on county level factors was obtained from US Census 2017-2021 American Community Survey, USDA, and National Center for Health Statistics Mortality Data on CDC WONDER (2019 crude mortality rates).

References:

https://www.nytimes.com/interactive/2023/us/covid-cases.html

https://covid19.census.gov/datasets/21843f238cbb46b08615fc53e19e0daf/explore

https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/

https://wonder.cdc.gov/mcd.html

Project Ideas

1. Predictors of county-level COVID-19 mortality in the U.S. The outcome of interest is the county-level COVID-19 mortality rate per 100k population (death_per_100k), and predictors include all county-level factors in the dataset, including demographics, socioeconomic status, health care access, and crude death rate before the pandemic. Traditional regression models and dimension reduction methods will be applied in variable selection.
2. Disparities of COVID-19 mortality using SuperLearner. Variables will be the same as #1, but this idea will specifically focus on comparing SuperLearner variable selection algorithms, such as generalized linear model, LASSO, and random forest. Differences in selected variables and model performance will be reported.
3. Interactive visualization of COVID-19 mortality trends. Visualize the trend of COVID-19 mortality rate per 100k population (death_per_100k) across key factors, e.g., geographical region, COVID-19 pandemic waves (Delta, Omicron etc.), policy changes (pre-vaccination, post-booster etc.). Iteractive plots will be the major deliverables.
4. Relationship between access to care and COVID-19 mortality. Regression analysis will be performed to evaluate whether rural counties, or counties with higher rates of uninsured individuals have higher COVID-19 mortality rates during the pandemic. Potential confounders such as demographics (e.g., population density, age distribution) will be adjusted.
5. Socioeconomic inequalities and COVID-19 mortality. This idea will focus on the correlation between county-level income inequalities (e.g., GINI index, median household income) and COVID-19 mortality, controlling for age and urbanicity.