

Configuration Manual

Racial disparities in Covid-19 incidence associated with Socio-Economic Factors in the United States

Ken Wheatley
Student ID: x16103785

School of Computing
National College of Ireland

Supervisor: Vikas Sahni

National College of Ireland
MSc Project Submission Sheet
School of Computing



Student Name:

Student ID:

Programme: **Year:**

Module:

Lecturer:

Submission Due Date:

Project Title:

Word Count: **Page Count:**

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:

Date:

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	<input type="checkbox"/>
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	<input type="checkbox"/>
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	<input type="checkbox"/>

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Ken Wheatley
Student ID: x16103785

1 Introduction

This document describes the hardware and software tools and processes used to implement the project “Racial disparities in Covid-19 incidence associated with Socio-Economic Factors in the United States”.

Each part of the development life cycle is described. Another purpose is that the document can be used to replicate all the technical work to complete the project.

2 System Configuration

2.1 Hardware

Dell Latitude 5410

Processor Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz

Installed RAM: 16.0 GB (15.6 GB usable)

System type: 64-bit operating system, x64-based processor

2.2 Software Environment

Windows 10 Operating System:

- OS build 19045.3324.
- Edition Windows 10 Pro
- Version 22H2

RStudio Version 1.4.1717R

R version 4.1.0 (2021-05-18):

- Platform: x86_64-w64-mingw32/x64 (64-bit)
- Running under: Windows 10 x64 (build 19045)

Microsoft Excel was the primary tool used to store external data, both raw input data and Generated output (persistent data) Both “.csv” and “.excel” formats used. The version of excel was “Microsoft 365 MSO (Version 2302 Build 16.0.16130.20684) 64-bit”.

Text files (.txt) were also used to store small datasets.

3 Data Processing and Exploration

This section describes how required raw data is obtained, loaded, explored, and pre-processed ready for model input.

All datasets are publicly available from following sources in the United States.

Centers for Disease Control and Prevention (CDC):

1. Monthly Covid-19 time-series dataset where each row represents a deidentified patient case. Broken down by county and race.
2. SVI Data broken down by county.
3. Population data by county and race.

University of Wisconsin Population Health Institute County Health Rankings:

4. Health rankings data broken down by county.

United States Census Bureau:

5. Community Resilience Estimates by state
6. Gini index of income inequality by state

3.1 Dataset Sources and Descriptions

The above referenced datasets are now discussed in further detail.

1. Monthly Covid-19 time-series dataset.

This dataset is provided by CDC and is entitled “COVID-19 Case Surveillance Public Use Data with Geography”. The data is a monthly time-series dating back to January 2020 of deidentified case data. The fields required are state, county, race, and month.

Races defined in the dataset are:

- American Indian/Alaska Native
- Asian
- Black
- Multiple/Other
- Native Hawaiian/Other Pacific Islander
- White

The data has an infection status field to indicate if probable or confirmed case. This is ignored as all entries are regarded as cases for modelling purposes.

The following link is the data contains all the filters required. Click on the link using Microsoft Edge browser (do not use Firefox as download does not work.)

The screen below (Fig 1) will open with all filters for the 16 states and exclusion of null or missing values already set. Click on “Export” and select “CSV for Excel” and click download. It takes several minutes as about 63 million rows are download.

Data Extract Link: [COVID-19 Case Surveillance Public Use Data with Geography | Data | Centers for Disease Control and Prevention \(cdc.gov\)](https://www.cdc.gov/data/2020/covid-19/covid-19-case-surveillance-public-use-data-with-geography/)

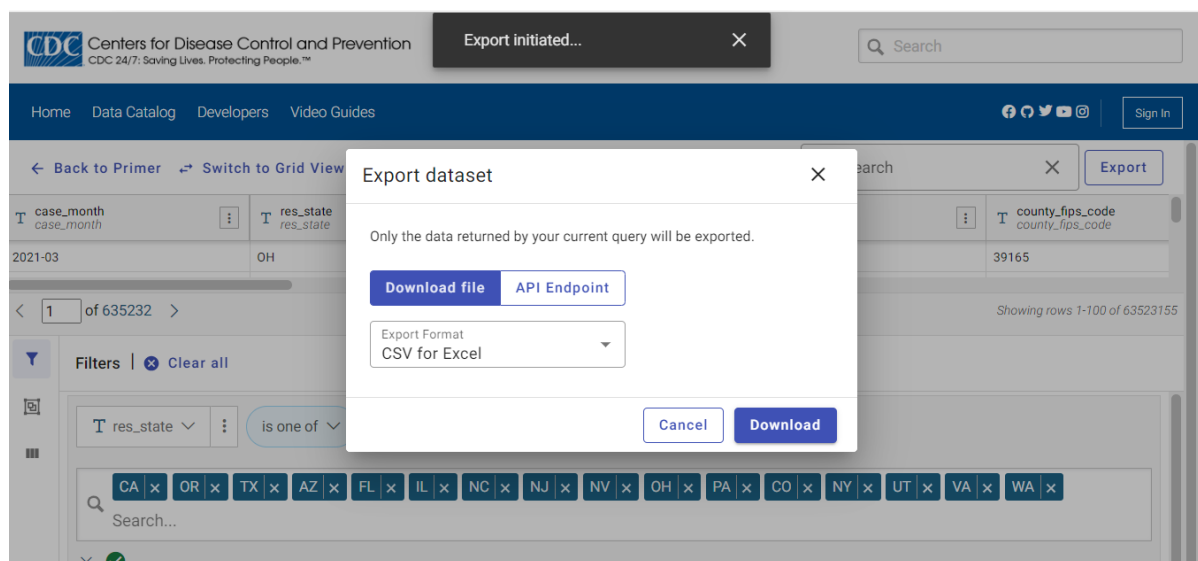


Fig 1. CDC download site for case data

2. SVI Data broken down by county.

The link to the SVI data below opens up the CDC/ATSDR SVI Data and Documentation Download page.

https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html

The top part of the SVI data download page is shown in Fig 2.a. Set year to 2020 (latest), Geography to “United States” and Geography Type to “Counties” and download as .csv file.

CDC/ATSDR SVI Data and Documentation Download

[Español \(Spanish\)](#) | [Print](#)

Year

2020

Geography

United States

If you choose “United States” as your option under Geography, all U.S. census tracts, or counties, are ranked against one another. Use “United States” for U.S.-wide or multi-state mapping and analysis.

If you choose an individual state, or “District of Columbia,” or “Puerto Rico,” tracts or counties are ranked against other tracts or counties in that state/district/territory.

Geography Type

Counties

File Type

- CSV File (table data)
- ESRI Geodatabase (map data)

Go

Fig 2.a SVI data download and documentation website page (top part)

File Type

- CSV File (table data)
- ESRI Geodatabase (map data)

Go

Our suggested citation for use of the database: Centers for Disease Control and Prevention/ Agency for Toxic Substances and Disease Registry/ Geospatial Research, Analysis, and Services Program. CDC/ATSDR Social Vulnerability Index [Insert 2020, 2018, 2016, 2014, 2010, or 2000] Database [Insert US or State]. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html. Accessed on [Insert date].

December 22, 2022: Data values for Housing Burden and Overall SVI in the 2020 SVI dataset have been updated. These variables were previously using E_HBURD instead of EP_HBURD in the calculation and thus the values after release on 10/27/22 and before 12/22/22 were incorrect.



Fig 2.b SVI data download and ddocumentation website page (bottom part)

Fig 2.b contains the link to the SVI documentaion which also contains the data dictionary. Click on link “CDC/ATSDR SVI Documentation 2020”

3. Population data by county and race.

The population data by state, county and race is obtained from the “CDC Wonder: Single-Race Population Estimates 2020-2021 Request” website (<https://wonder.cdc.gov/single-race-v2021.html>)

Population data for the following states is show in Tab 1 is obtained.

ANSI Code	State	US Census Region	ANSI Code	State	US Census Region
AZ	Arizona	West	NC	North Carolina	South
CA	California	West	OH	Ohio	Midwest
CO	Colorado	West	OR	Oregon	West
FL	Florida	South	PA	Pennsylvania	Northeast
IL	Illinois	Midwest	TX	Texas	South
NV	Nevada	West	UT	Utah	West
NJ	New Jersey	Northeast	VA	Virginia	South
NY	New York	Northeast	WA	Washington	West

Tab 1, US States selected for this project.

The data must be download piecwise one or two states at a time due to restriction of 75,000 rows enforced by website (Fig 3. The files must be downloaded with the following naming conventions where the state abbreviations Indicate file contents.

- Single-Race Population Estimates 2020-2021_AZ_IL.txt
- Single-Race Population Estimates 2020-2021_CA_FL.txt
- Single-Race Population Estimates 2020-2021_CO_OR_UT.txt
- Single-Race Population Estimates 2020-2021_NC.txt

- Single-Race Population Estimates 2020-2021_NJ_OH.txt
- Single-Race Population Estimates 2020-2021_NY.txt
- Single-Race Population Estimates 2020-2021_PA.txt
- Single-Race Population Estimates 2020-2021_States.txt
- Single-Race Population Estimates 2020-2021_TX_Hisp.txt
- Single-Race Population Estimates 2020-2021_TX_NotHisp.txt
- Single-Race Population Estimates 2020-2021_VA.txt
- Single-Race Population Estimates 2020-2021_WA_NV.txt

Single-Race Population Estimates 2020-2021 Request

Request Form | Results | Map | Chart | About

[Single-Race Population Estimates](#) | [Dataset Documentation](#) | [Other Data Access](#) | [Data Use Restrictions](#) | [How to Use WONDER](#)
Save | Reset

Messages:
 ▶ This request produces 375,726 rows, but 75,000 is the maximum allowed. Simplify this request, or send a series of smaller ones. For example, group results by race but limit each query to a single race. Send a request for each race group and then merge the results. Please contact user support for further assistance.

Make all desired selections and then click any **Send** button one time to send your request.

1. Organize table layout: Send | Help

Group Results By County ▾

And By Race ▾

And By Ethnicity ▾

And By Gender ▾

And By Age Group ▾

2. Select location: Send | Help

Click a button to choose locations by State or by Region.

States | **Regions**

Browse or search to find items in the States Finder Tool, then **highlight** the items to use for this request. (The *Currently selected* box displays all current request items.)

[Finder Tool Help](#) | [Advanced Finder Options](#)

Browse | Search | Details

States

- + 48 (Texas)
- + 49 (Utah)
- + 50 (Vermont)
- + 51 (Virginia)
- + 53 (Washington)
- + 54 (West Virginia)
- + 55 (Wisconsin)
- + 56 (Wyoming)

Open | Close | Close All

Currently selected:

- 04 (Arizona)
- 06 (California)
- 12 (Florida)
- 17 (Illinois)
- 36 (New York)
- 37 (North Carolina)
- 48 (Texas)
- 51 (Virginia)

Browse the list by opening and closing items.
Use Ctrl+Click to multiple select, Shift+Click for a range.

3. Select demographics and years: Send | Help

Hint: Use Ctrl + Click for multiple selections, or Shift + Click for a range.

Age Group

- All Ages
- 0-4 years
- 5-9 years
- 10-14 years
- 15-19 years
- 20-24 years
- 25-29 years
- 30-34 years
- 35-39 years
- 40-44 years
- 45-49 years
- 50-54 years

Race

- All Races
- American Indian or Alaska Native
- Asian
- Black or African American
- Native Hawaiian or Other Pacific Islander
- White
- More than one race

Ethnicity

- All Ethnicities
- Hispanic or Latino
- Not Hispanic or Latino

Yearly July 1st Estimates

- All Years
- 2020
- 2021

Gender

- All Genders
- Female
- Male

4. Other options: Send | Help

Export Results (Check box to download results to a file)

Show Totals

Show Zero Values

Precision decimal places

Data Access Timeout minutes

Fig 3. CDC Wonder Populations download page.

4. Population data by county and race.

Health rankings data is obtained from the University of Wisconsin Population Health Institute County Health Rankings website (<https://www.countyhealthrankings.org/explore-health-rankings/rankings-data-documentation>).

The following screen (Fig 4) appears. Select [2023 County Health Rankings National Data](#) to download the data as an excel file. The file also contain data dictionary information.

Rankings data & documentation

Find national statistics, state-level data and technical documentation including changes to our measures, guidelines for comparing data across states, information about data years and sources and more.

National data & documentation

2023 County Health Rankings

NATIONAL DATA	TREND DATA
2023 County Health Rankings National Data	2023 CHR CSV Trends Data
2023 CHR CSV Analytic Data	2023 CHR SAS Trends Data

Fig 4, Health Rankings data download page (University of Wisconsin Population Health Institute)

Two datasets are obtained from the United States Census Bureau:

5. Community Resilience Estimates by state

Community Resilience Estimates. By State or County as selected <https://www.census.gov/programs-surveys/community-resilience-estimates/data/datasets.html>

Click on the link and the page (Fig 5) below will appear. Select State file for download under 2021 Estimates. The page also contains guides and data dictionary information. 3 fields are used, the rates per population of individual with 0, 1 or 2, or 3 risk factors.

2021 Estimates

Data Files











-  [File Layout \[< 1.0 MB\]](#)
-  [All Geographic Levels \[< 1.0 MB\]](#)
-  [County \[< 1.0 MB\]](#)
-  [National \[< 1.0 MB\]](#)
-  [State \[< 1.0 MB\]](#)
-  [Tract \[< 1.0 MB\]](#)
-  [Top Vulnerable Counties and Tracts, All \[< 1.0 MB\]](#)
-  [Top Vulnerable Counties and Tracts, Hurricane Risk Areas \[< 1.0 MB\]](#)
-  [Quick Guide \[< 1.0 MB\]](#)
-  [Detailed Technical Documentation \[< 1.0 MB\]](#)

Fig 5. Community Resilience data

6. Gini index of income inequality by state

Gini Index by State data is available at the following link. Download dataset B19083 (Fig 6.). As a .csv file.

[https://data.census.gov/table?q=Gini&g=010XX00US,\\$0400000_040XX00US01,02,04,05,06,08,09,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,47,48,49,50,51,53,54,55,56,72&tid=ACSDT1Y2021.B19083&moe=false&tp=true](https://data.census.gov/table?q=Gini&g=010XX00US,$0400000_040XX00US01,02,04,05,06,08,09,10,11,12,13,15,16,17,18,19,20,21,22,23,24,25,26,27,28,29,30,31,32,33,34,35,36,37,38,39,40,41,42,44,45,46,47,48,49,50,51,53,54,55,56,72&tid=ACSDT1Y2021.B19083&moe=false&tp=true)

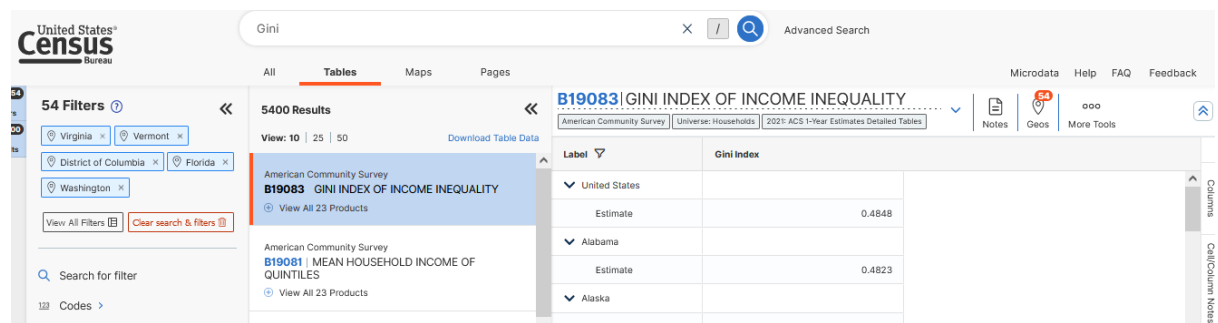


Fig 6. Gini Index download page

3.2 Data Pre-processing

All downloaded data is stored in the same project directory. It is loaded and pre-processed automatically using R code.

3.2.1 Monthly Cases

The largest dataset was the CDC cases monthly time-series data which contained over 63 million rows. This represented monthly data from Jan 2020 – June 2023 broken down by state, county, month and race. However around 27 million rows had missing key data and could not be used for the following reasons.

- No county description. Therefore, no population data exists. Approx 3 million rows
- No race description. Again, no population data could be determined as needed for modelling purposes. Approx 23.7 million rows

After processing, 36,868,293 rows of monthly case remained.

The times-series was cleaned using the `ts_clean()` function from forecast package. This function identifies and replaces outliers and missing values in a time series. Only data for two months was cleaned by the function. These months were Dec2020 and Jan 2021. These 2 months showed highly volatile data as shown in the graph of the raw data below (Fig 7). This spoke coincided with the emergence of the highly contagious Covid-19 omicron variant.

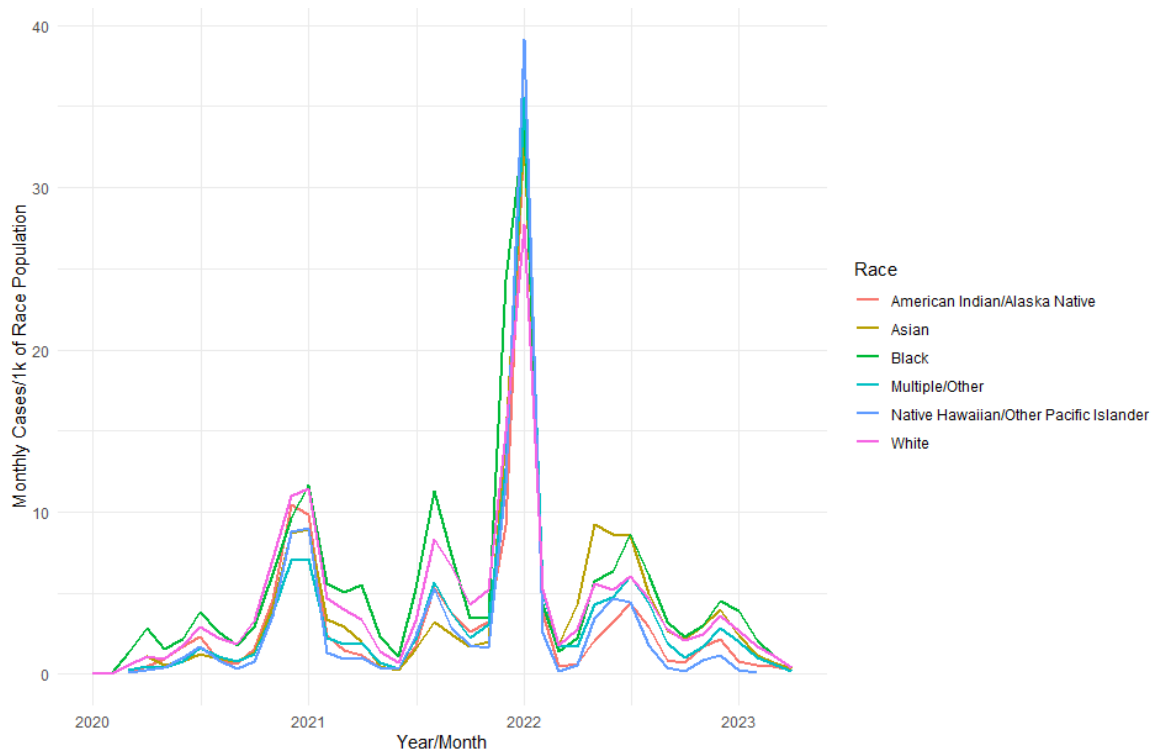


Fig 7. Monthly Cases/1k of population across selected States by Race

The next step was to aggregate the time-series data to create total cumulative case counts for each grouping of county and race. This created a cross-sectional snapshot of all cases as of June 2023. The snippet of code performing the aggregation is shown below in Fig 8.

```
# Create intermediate summary table for total cumulative cases by county and race. Both
# original time-series and corrected time-series are derived and stored for comparison.
#
G_R_M_Agg_tmp2 <-
  G_R_M_Agg_tmp1 %>%
  group_by(res_state, state_fips_code, res_county, county_fips_code, race) %>%
  mutate(CasesRace_Cnty = sum(CasesRace_Mnth),
         CasesRace_Cnty_raw = sum(CasesRace_Mnth_raw)) %>% ungroup() %>%
  select(res_state, state_fips_code, res_county, county_fips_code, race, CasesRace_Cnty, CasesRace_Cnty_raw) %>%
  unique()
```

Fig 8. R code for aggregation of time-series

3.2.2 Population

Population data from the inputs files was combined into a single file and rows with missing population or race data removed.

Population data was combined with aggregated cases data to derive rates of cases per population proportions for each race in every county. These proportions represent sample (maximum likelihood) estimates of infection probabilities and are assumed to follow a binomial distribution for the logistic regression models used in this project.

3.2.3 Health Rankings

Some county health rankings data was missing and was imputed using median values from all other counties within the same state.

3.2.4 Other Data

For the remaining SVI, CRE and Gini index data, no special processing other than formatting was required.

3.3 Data Exploration

Detailed data exploration was carried out using ggplot2 package for graph plotting. Fig 8 shows the distribution of cases per 1k of proportions broken down by race.

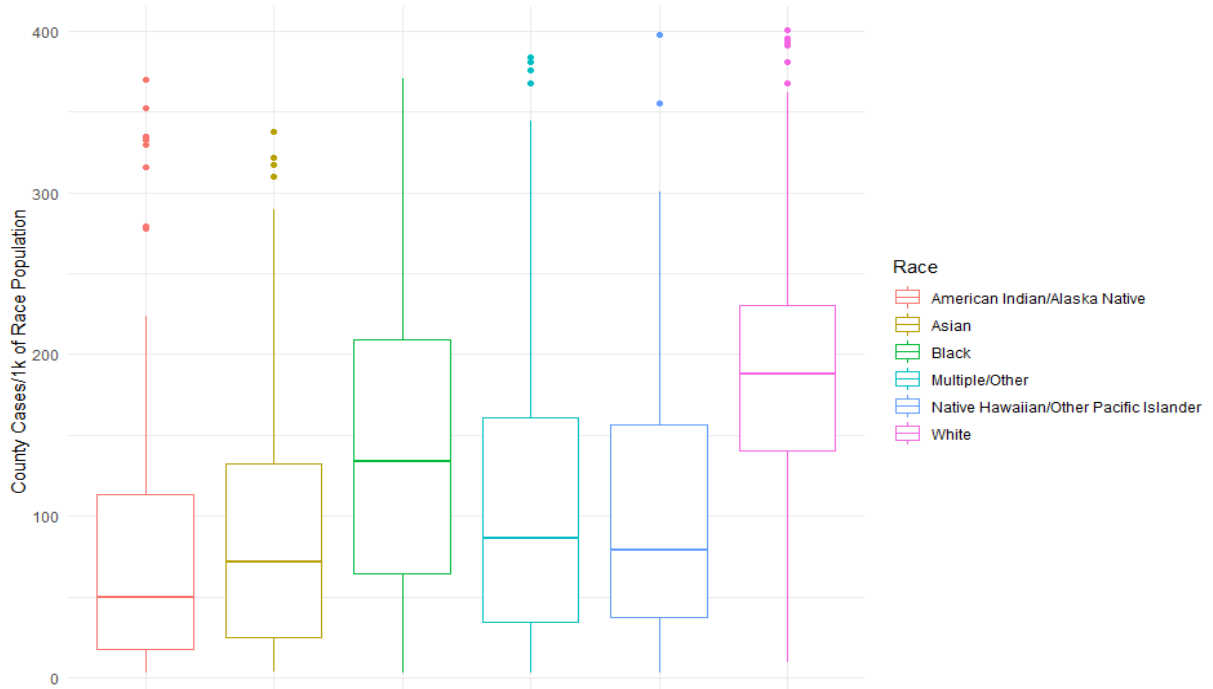


Fig 8. Cumulative county cases/per 1k of Population in period Jan 2020 - Apr 2023

Wide distributions are shown with a some extreme outliers. Whites are shown as having the highest infection rates per unit of 1k population. American Indian/Alaska Natives have the lowest.

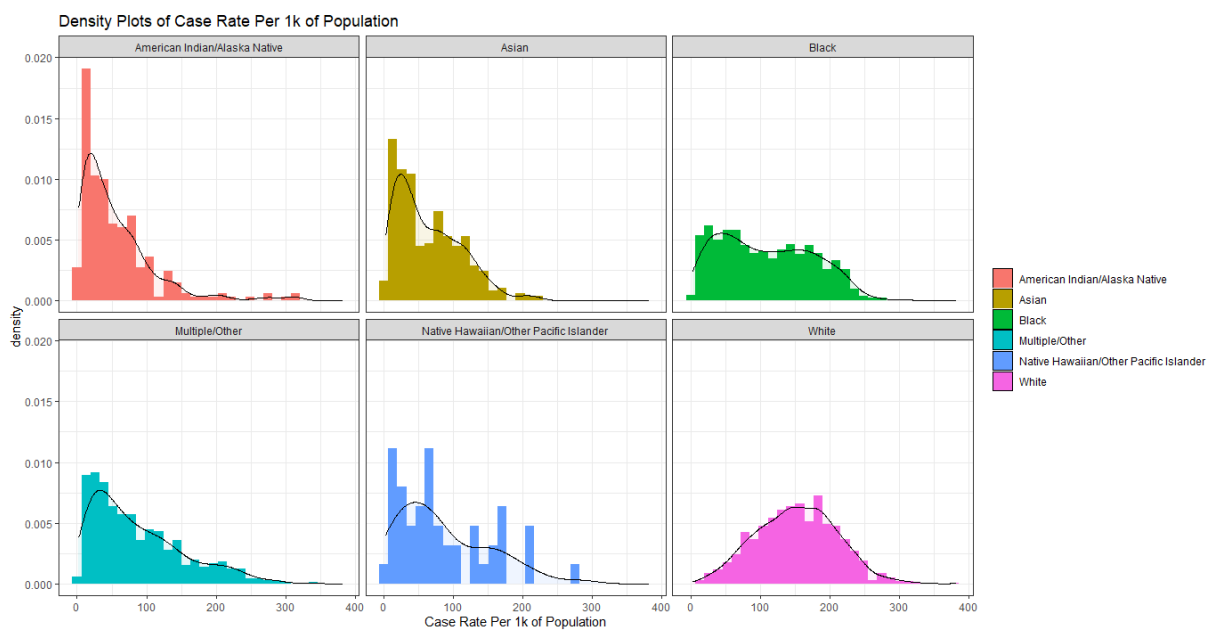


Fig 9. Density plots of Case rated per 1k of Population.

Fig 9. Shows the distribution of cases per 1k of population for each race. The proportion of cases/unit of population is assumed by this study to follow a binomial distribution. For whites, the distribution is symmetrical showing a binomial approximation to a normal distribution for large samples. Whites were both present in the most counties and typically had the largest populations, so had the largest sample size. The other distributions are somewhat right-skewed reflecting their smaller sample sizes.

Correlations plots were also examined, and a cross-tabulation was stored for future reference in feature selection.

4 Modelling

Logistic regression is used to model the probabilities of infection rates for each racial group. Infection rate proportion being defined as the number of cases per unit of population. The proportions are assumed to follow a binomial distribution. If p represents the estimated mean probability of infection for a given race and controlled for by a set of explanatory variables, then the variance should be expressed by:

Variance = $p(1-p)$. If the variance is larger than this, it is known as overdispersion.

In addition, the following assumptions must hold.

- The Response Variable is Binary or a Proportion
- The Observations are Independent.
- There is No Multicollinearity Among Explanatory Variables
- There are No Extreme Outliers
- There is a Linear Relationship Between Explanatory Variables and the Logit of the Response Variable
- The Sample Size is Sufficiently Large

Both fixed effects and random effects by state were modelled. The hypothesis regarding random effects is that socio-economic and health environmental factors differ between states due to differing policies, resources and many other factors which may not be known. Therefore, mean case rates for each population within each state cluster, may differ from one state to the next.

Fixed effects are modelled for 2 scenarios.

In one scenario, standard errors and overdispersion are accounted for, and in the other both are not.

5 Implementation

R scripts used to implement the model are described below. The “S” prefix represents the sequence model should be run. The sequence S01 and S04 must be run before S04-08. These details are also supplied in the README.txt document accompanying the project artifacts.

ProjectLaunch.Rproj	R project file to launch project
S01_ETL.R	Data loading and pre-processing
S02_EDA.R	Data exploration

S03_Corr_CountyFeatures.R	Feature correlations Analysis
S04_PrepFEModelData.R	Final preparations of data for modelling
S05_FEModel_StdErr.R	Fixed effect model using standard errors
S06_FEModel_RobustErr.R	Fixed effect model using robust errors and overdispersion correction.
S07_ModelOutput.R	Mixed effects and all model comparisons
S08_Report_Graohics.R	Graphics used in report

5.1.1 Fixed Effect Model

5.1.1.1 Accounting for clustered standard errors and overdispersion.

Clustering errors were assumed by the model due to between-state variances. Very high extra-binomial dispersion was also observed. In the fixed-effect scenario, these issues are considered a nuisance and accounted for. Overdispersion is often modelled using the quasi-binomial model.

The following approach was used, which had the benefit of facilitating feature selection, in conjunction with variance inflation (vif) and correlation analysis. Vif is a measure of linear relationships between explanatory features which can seriously distort model results.

1. Starting with saturated model (all candidate explanatory features), remove features successively with highest vif value. Use correlations, to help decide if highest vifs are very similar. Stop when all remaining features have $vifs < 10$.
2. With remaining features from 1. Perform backwards stepwise regression accounting for overdispersion to adjust standard errors and using AIC comparisons. Stop when all features are either significant (5% level) or AIC has reached minimum.
3. Using output from 2, perform forward stepwise regression with robust error corrections (using the Sandwich package). Repeat until all a feature set where all p-values are significant.

The above customised process reduced features from 18 to 6.

Sample sizes were decided based on recommendations in the following articles (Bujang, et al, Vittinghoff, E).

5.1.1.2 Ignoring clustered standard errors and overdispersion.

Models were ran using stepwise AIC regression and cross-validated regression (Caret package). Both produced the full list of original features all statistically significant. This was due to standard errors being massively underestimated.

Lasso regression was also tried, and although the feature list was reduced, standard errors remained small.

Because of the severe underestimation of standard errors, these models were no longer considered.

5.1.2 Random Effect Models

Models were run with state as random effect on the intercept and on the slope of the race category.

6 Evaluation

Correlations and variance inflation factors (multicollinearity test) were used to help select features. Using a combination of AIC model comparisons, significance testing using robust errors and overdispersion correction, a model of 6 features was obtained (from 18 candidates).

The table (Tab 2) shows the model output on the log(odds) scale and Whites are the reference group. The signs of the coefficients are important. Relative to Whites, Asian, Black and Native Hawaiian/Other Pacific Islander populations have lower odds of infection controlling for all variables.

However, the table also shows a negative coefficient for the percentage of persons aged 65 or over. This cohort is known to be especially vulnerable, so this result needs further investigation.

County Feature	Coefficient	Confidence Interval		Robust Standard Error
		2.5%	97.5%	
California	-0.604	-0.677	-0.531	0.108
Asian	-0.156	-0.339	0.027	0.096
Black	-0.156	0.029	0.027	0.096
Multiple/Other	0.116	-0.168	0.204	0.044
American Indian/Alaska Native	0.055	-0.749	0.278	0.108
Native Hawaiian/Other Pacific Islander	-0.377	-0.299	-0.005	0.201
% Households without internet connection	-0.001	0.076	0.298	0.153
Area in square miles	0.098	0.006	0.120	0.030
SVI Racial & Ethnic Minority Status Theme	0.014	-0.127	0.022	0.008
% of Hispanic or Latino persons	-0.059	-0.010	0.009	0.059
% uninsured in 90 th percentile	0.090	-0.148	0.190	0.046
% persons aged 65 or older	-0.094	-0.019	-0.039	0.035

Tab 2. Model summary output

References

Bujang, M. A., Sa'at, N., Bakar, T. M. I. T. A., & Joo, L. C. (2018). Sample size guidelines for logistic regression from observational studies with large population: emphasis on the accuracy between statistics and parameters based on real life clinical data. *The Malaysian journal of medical sciences: MJMS*, 25(4), 122.

Vittinghoff, E., & McCulloch, C. E. (2007). Relaxing the rule of ten events per variable in logistic and Cox regression. *American journal of epidemiology*, 165(6), 710-718.