September 21 2021

Division of Computing Science and Mathematics
Faculty of Natural Sciences
University of Stirling

COVID-19 Cases and Deaths and Demographic Features of California Counties

Sarah Good

August 25, 2023

Dissertation submitted in partial fulfillment for the degree of
Master of Science in Mathematics and Data Science

September 2021

Abstract

- Problem: Various demographic features are known to effect the spread of diseases such as COVID-19. However, it is unclear which features have had the greatest impact at the population level in California specifically.
- Objectives: The goal of this study is to determine correlations between cumulative case and deaths per 100,000 and demographic data marking features such as socioeconomic status, race and ethnic background, healthcare access, household size and overcrowding, and other features, as well as the statistical significance of these correlations.
- Methodology: The procedure of the study was as follows:
  - Merge demographic data of Californian counties to COVID-19 cumulative cases and deaths per 100,000.
  - Determine the top 5 features which correlate best with cumulative cases per 100,000 population and cumulative deaths per 100,000 population relative to each county.
  - Determine the top 20 features which correlate best with cumulative cases per 100,000 and cumulative deaths per 100,000 population, and subtract features until no strong correlations within the feature set remain.
  - Generate linear models that relate to cumulative cases per 100,000 and cumulative deaths per 100,000 from these feature sets to determine both the significance of a model as a whole and its individual features.
  - Determine correlations between demographic features included in the models and without the models to extrapolate potential causes behind higher or lower COVID-19 cumulative case and death rates per 100,000.
- Achievements: It was found that the top 5 correlations for cases per 100,000 were as follows: “Average Household Size,” “Hispanic ethnicity” (% of population), “Ages 0-19” (% population), education level of “9th to 12th with no high school diploma” (% of population older than 25 years), and incidence rates of “Long-term Diabetes Complications” (per 100,000 population). For deaths per 100,000, the top 5 correlations were the same except Overcrowding replaced Long-term Diabetes Complications. The feature set calculated for both cases and deaths from the top 20 correlations without strong correlations were identical to each other and are as follows: Overcrowding (% of households), “Wholesale trade” (% of workforce employed in), “Transportation, warehousing, and utilities” (% of workforce employed in), and “Graduate or professional degree” (% of population older than 25 years). All linear models were found to be significant overall. Examples of features which correlated strongly to these input sets include Long-term Diabetes Complications, 9th to 12th grade, no diploma, Ages 0-19, health insurance status, “Agriculture, forestry, fishing and hunting, and mining” (% of workforce), and Hispanic ethnicity. This indicates populations with high concentrations of essential workers, low income, and subject to racial/ethnic inequality are more vulnerable to COVID-19 cases and deaths.

Attestation

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this dissertation reports original work by me during my University project.

Signature: Sarah Good **Date: August 25, 2023**

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background and Context

COVID-19 has had a massive impact on the world, including the United States. However, some communities and individuals are at higher risk of the severe outcomes and death from the disease [1]. These risks are compounded for people of lower socioeconomic status and those with limited access to health care. This is due to both greater barriers to obtaining care [2], [3], [4] as well as the higher rates of chronic diseases, such as hypertension, diabetes (type-2), obesity, and other illnesses in these communities. This is partially due to the chronic stress of such living conditions [5], [6]. Essential workers in particular are also at a higher risk of COVID due to their higher rates of exposure to other people [7], [8].

1.2 Scope and Objectives

The goal of this study was to determine correlations between demographic features and the magnitude of COVID-19 outcomes in terms of cumulative cases and deaths per 100,000 of population since 1 February 2020 to 5 July 2021. This was done by calculating the Pearson correlation coefficients (r values) between the population features and cumulative cases per 100,000 and cumulative deaths per 100,000. The most highly correlated features were to be determined based on the absolute value of their Pearson Correlation Coefficients in relation to cases per 100,000 or deaths per 100,000.

Linear regression models were created from these top correlated features in two ways: The first feature set uses the top 5 correlations. The second uses the top 20 correlations with features highly correlated with each other filtered out to limit interactions between features. Linear Models 1 and 3 use the top 5 correlations for cases and deaths respectively while Models 2 and 4 do the same but with the filtered feature set. These models can be used to determine a) the most significant features out of these subsets and b) features that approximate different potential forces on COVID cases and deaths (especially in the case of the latter set of features).

Additionally, co-correlations, defined as demographic features *not* within a given input feature set for the regression models but which are strongly correlated with the features included within, were calculated for all features. By determining the most highly correlated features to COVID-19 outcomes and which sets of features correlate with each other, the underlying factors behind these outcomes can be hypothesized.

1.3 Achievements

The five features which had the highest correlations to cumulative cases per 100,000 were found to be the following: "Average Household Size," "Hispanic ethnicity" (% of population), "Ages 0-19" (% population), education level of "9th to 12th with no high school diploma" (% of population older than 25 years), and incidence rates of "Long-term Diabetes Complications" (per 100,000 population). For deaths per 100,000, the feature set was the same except "Overcrowding" (% of households) replaced Long-term Diabetes Complications. The feature set for uncorrelated features was the same for both cases and deaths. This set was comprised of Overcrowding (% of households), "Wholesale trade" (% of workforce employed in), "Transportation, warehousing, and utilities" (% of workforce employed in), and "Graduate or professional degree" (% of population older than 25 years).

All linear regression models were found to be significant overall. Model 1's (for cases per 100,000, top 5 correlations) only significant feature for predicting cases per 100,000 is Hispanic ethnicity in predicting cumulative cases per 100,000. Model 2's (for cases per 100,000, top 20 filtered features) significant features were Transportation, warehousing, and utilities and Graduate or professional degree. In Model 3 (for deaths per 100,000, top 5 correlations), the significant features were Hispanic ethnicity and Average Family Size, whereas in Model 4 (for deaths per 100,000, top 20 filtered features), these features were Transportation and warehousing, and utilities and Overcrowding.

Co-correlations for all features within Models 1 and 3 include "Ages 60-69" (% population), "Ages 80+" (% population), "Commuting - Worked from home" (% workers), "health insurance status" (% noninstitutionalized civilians), and "Less than 9th grade education" (% population $\geq$ 25 years). For Models 2 and 4, the features co-correlated with at least 3 of the input features include Long-term Diabetes Complications (incidence per 100,000), 9th to 12th grade, no diploma (% population $\geq$ 25 years), and Ages 0-19 (% population).

This evidence suggests a strong correlation between cumulative cases and deaths per 100,000 and the conditions of poverty and low socioeconomic status, high levels of employment in essential work, and racial inequalities. Due to the weaker correlation between the features in Model 2 and 4's feature set, it is possible that essential work, housing overcrowding, and certain levels of graduate education are all relatively independent of each other and point to different underlying causes of county-wide COVID outcomes.

1.4 Overview of Dissertation

First, a literature review of COVID-19 and how it relates to demographic features in general and specific to the United States are outlined, followed by this study's goals as they pertain to existing research. Second, the methods used to determine the correlations between features and cumulative cases per 100,000 and cumulative deaths per 100,000 respectively are outlined. Next, methods used to find correlations between demographic features are described. Lastly, the results of the study are presented, followed by an evaluation of the study and suggestions for future work and policy suggestions.

2 Background

2.1 About COVID-19

Coronavirus 2019 (COVID-19) is the respiratory illness caused by the virus SARS-CoV-2. It was discovered in Wuhan China in late 2019 and usually causes mild respiratory and flu-like symptoms, including (but not limited to) fever, dry cough, and fatigue [9]. It is mainly spread by aerosolized droplets when infected people exhale, such as through breathing, talking, coughing, and sneezing [10], [11]. Certain groups, including those who are older, those who have pre-existing health conditions, or both, are more likely to suffer severe illness, become hospitalized, and die should they become infected [1], [12].

The impact of COVID has been immense worldwide. As of July 1, 2021, World Health Organization (WHO) estimated that over 181 million cases and almost 4 million COVID-related deaths had been confirmed since the beginning of the pandemic [13]. On the same date, the Center for Disease Control and Prevention (CDC) in the United States reported over 33 million cases and over 600 thousand COVID-related deaths had been recorded so far [14]. It is expected that that COVID-19 will reduce the average life expectancy of Americans by more than a year, with this effect being three to four times as large for Black and Latinx populations compared to Whites [15].

2.2 Social Inequities and Health

In the United States in particular, health care access is limited by insurance status as well as the availability of local hospitals and other health services in general. Those living in poverty are more likely to lack health insurance. For example, 11.0% of all Californians between the ages 19 to 64 are uninsured compared to 17.8% of Californians living below the Federal poverty line in the same age group [16], [17]. In a nationwide poll conducted by the Kaiser Family Foundation, 73.7% of uninsured Americans cite the cost of insurance as the reason for their uninsured status, even after the implementation of the Affordable Care Act (ACA) [18]. Uninsured Americans have limited access to healthcare, especially preventative care services and treating chronic conditions compared to those who have health insurance [2], [3], [4].

This is especially worrying since people in poverty are more vulnerable to severe outcomes of COVID-19. Lower social status is associated with poorer health outcomes than those of higher status, including greater levels of stress and greater risk of chronic conditions such as obesity, hypertension, and diabetes [5], [6]. Moreover, experiencing periods of unemployment increases the risk of metabolic syndrome [19]. Because such conditions are known risk factors for severe disease, low income populations are more susceptible to severe disease and death when they are infected with COVID-19 [1].

Inadequate housing is also a significant risk factor for poorer health outcomes in general and COVID-19 in particular. It has been well established that overcrowded and substandard housing conditions are correlated with higher rates of diseases, including respiratory diseases and worse mental health outcomes (the latter of which increases physiological stress, increasing risks of chronic disease) [20], [21], [22], [23]. Additionally, poor housing conditions have been linked to higher rates of COVID-19 case rates and mortality [24]. Furthermore, overcrowded households are more vulnerable to COVID as the virus spreads more easily indoors and especially when the infected person is in proximity to others [10], [11]. All of these factors raise the risk of COVID transmission and severe disease for populations in poor living conditions.

Education, both in terms of work opportunities and health literacy, also effects one's health outcomes. Lower levels of education limit one's opportunities in obtaining higher paying jobs [25], and are associated with lower levels of health literacy, or the skills needed to understand and assess health-related information [26]. Low health literacy has also been connected to increased risk of hospital admission for preventable conditions [27]. Those with poor health literacy skills may also be more vulnerable to false information about the pandemic, as they lack the skills necessary to accurately determine which information is true [28].

Racial inequalities have been highlighted by the pandemic. Americans of Hispanic and Latinx decent were found to represent 28.7% of the cases nationwide while accounting for 18.45% of the total U.S. population. Hispanic/Latnix and Black non-Hispanic people also represent more COVID-related deaths nationwide than their share of the population would suggest (Hispanic/Latinx: 18.70% deaths vs. 18.45% of population, Black: 13.70% vs. 12.54%) [14]. Additionally, there is evidence that Black people in the US are more likely to be infected, hospitalized, and die from the illness compared to their White counterparts [29], [30].

Essential workers are especially at risk of contracting COVID as well, given that they cannot self isolate like other workers. Essential workers comprise about 40% of the US workforce, and include industries such as healthcare, social services, food services, cleaning, transit, essential manufacturing, and agriculture. Studies suggest that essential workers are at higher risk of being infected with COVID-19 compared to the general population due to increased proximity to others [7], [8]. A UK study found that essential workers were at higher risk of severe disease and mortality due to COVID-19 compared to non-essential workers [31]. Moreover, 25% of US essential workers were estimated to be low income, and 11% were estimated to lack health insurance [32], putting them at even greater risk of worse COVID outcomes.

2.3 Significance of this Study

The uniqueness of this study in terms of COVID-19 impact lies in its focus on Californian counties and their demographic features. Studies which are similar in design tend to focus on all US counties, such as a study linking overcrowding and COVID incidence and mortality rates [33], rather than a single state.

Furthermore, many COVID studies have linked environmental and demographic factors to increased risk of transmission and mortality at the individual level, but not as they pertain to the population level. The goal of this study, however, is to measure the correlation of the demographic features in terms of how they effect population at a large scale rather than the individual. Features that represent a marginal set of the population may not have a significant correlation to the outcomes compared to other features which describe or effect larger sections of the population.

It should be noted that general measures of a community's vulnerability to disease outbreaks such as COVID exist. The California Healthy Places Index (HPI) describes the health of Californian communities based on social determinants of health [34]. The Social Vulnerability Index (SVI) by the CDC and Agency of Toxic Substances and Disease Registry (ATSDR) is a more general measure that determines a community's vulnerability to emergencies, such as natural disasters and disease outbreaks, based on its socioeconomic status, housing and transport conditions, racial and ethnic background, and household composition [35]. Therefore, this study aims to compare some of the features also included in these vulnerability measures together to see which are the most impactful for COVID-19 specifically, not in terms of general health or disaster preparedness.

More specifically, this study seeks to find which of these features are most correlated with COVID cases and deaths in California specifically. While social inequities apply generally across the entire United States, this study may be used as a point of comparison for similar studies done either nationwide or for other states. The results may also be followed up with future studies focusing on specific features or subsets of features and how they correlate with COVID-19 outcomes at the individual level.

3 Methods

3.1 Data Sources

COVID-19 data on daily cases and deaths by county were sourced from the COVID-19 Time Series Metrics dataset by county was sourced from the Vaccine Progress Dashboard data on 27 July 2021 through the links to the .csv files available on the California Open Data Portal [36]. Cases per 100,000 and deaths per 100,00 were calculated using the raw case and death count values per county divided by county population and multiplied by 100,000. The analysis was restricted to COVID data on and before 5 July 2021 in order to account for cases and deaths counted retroactively and to account for potential delays in reporting.

Most demographic data were sourced from the United States American Community Survey (ACS) Data and specifically from the 2019 1-Year Estimates. While 2020 data were available at time of analysis, 2019 demographic data were used for this analysis since it is more representative of what communities were like approaching the pandemic. Age group data and race/ethnicity data were downloaded as .csv files from the US Census website. Age group data was split into bins of 5 years in length (i.e. "0-4", "5-9", etc.) cutting off at the category for "85+" years.

All other demographic characteristics, namely population employed in various industries, education attainment, disabled population (civilian, non-institutionalized), health insurance status (private, public, or uninsured), language spoken at home, English ability, median income, and average family size, were downloaded through the use of the US ACS API for 2019 1 Year data [37]. The reference populations for each group of variables are listed in Table 3.1.

Table 3.1: Reference Populations for Data

Feature	Sampled Population
Race/Ethnicity	Total Population
Sex	Total Population
Age Categories	Total Population
Registered Party Affiliation	Registered Voters
Education Level	Population $\geq$ 25 years
Employment Status	Population $\geq$ 16 years
Health Insurance Status	Civilian Noninstitutionalized Population
Industry	Full-time, year-round civilian employed population $\geq$ 16 years*
Commuting	Workers $\geq$ 16 years
Language Skills	Population $\geq$ 5 years
Residence Continuity	Total Population
Disability Status	Civilian Noninstitutionalized Population
Preventable Illness (Incidence per 100,000)	Incidence per 100,000
Overcrowding	All Households

*Estimate includes the industry category of those who have been employed within the last 5 years.

Race and ethnicity data contains the following groupings: Asian, American Indian and Alaska Native, Black, White, Native Hawaiian and Other Pacific Islander, Hispanic, and

Two or More Races. In this case, “Hispanic” refers to people of Spanish or Latin-American origin aside from Brazilian. This is not to be confused with the terms “Latino,” “Latina,” and “Latinx,” which refer to people of Latin-American origin, including Brazil. The Hispanic designation is treated as a binary (Hispanic and non-Hispanic) distinct from race in the data. All racial categories retained for analysis except Two or More Races refer groupings of either 1) alone or 2) alone and in combination. “Alone” (such as White only) groupings refer to race groups identifying *only* as that race whereas “Alone and in Combination” (such as White Alone and in Combination with) refers to people in the “Alone” grouping *and* anyone who also identifies with at least one other race.

Industry values are determined based on the job full-time, year-round employed civilians have worked at the most in the past week relative to when the survey was conducted. For unemployed civilians who have worked in the past 5 years, this refers to the industry of the job previously held [37].

Employment status categories sample the total population of those 16 years and older. The list of features here includes enlistment in the armed forces as a percent of this population in addition to civilian employment, which may appear as lower civilian employment rates in areas with high percentages of military personnel [37].

Some data groupings were merged. Age data were regrouped from groups of 5 years to the following groups: 0-19, 20-39, 40-49, 50-59, 60-69, 70-79, 80+. Race/ethnicity and industry data were originally subdivided by gender, but were summed together to create a total value. In the case of age and race/ethnicity, percentages of total population were attained for both the age and race/ethnicity variables by dividing each variable by the total population for each county. This was not needed for the industry data since the percentage values could be summed from the percentages sourced from the API.

Housing overcrowding data by county were also downloaded as a .csv file from the California Open Data Portal. In this dataset, overcrowding is defined as a housing density greater than 1.0 persons per room and severe overcrowding is defined as housing density greater than 1.5 persons per room. Data from 2019 for county-wide populations were used [38]. Data constrained to 2019 were extracted and only county, overcrowding, and severe overcrowding estimates (as percentage of households) were retained.

Preventable hospitalization data were downloaded from the California Open Data Portal. This data used from this data set were the county-wide 2019 observed rates (per 100,00 of population) of preventable hospitalizations related to Diabetes (Short-term complications, Long-term complications, uncontrolled, and lower-extremity amputations among patients with diabetes), Asthma in Younger Adults (Ages 18-36), COPD or Asthma in Older Adults (Ages 40+), Hypertension, Community-Acquired Pneumonia, and Urinary Tract Infections [39].

Voter registration data were sourced from Statement of Vote: General Election November 3, 2020. Voter registration numbers were used from the “Report of Registration as of October 19, 2020” and were copied into a .csv file manually [40]. Republican and Democrat party estimates were acquired for each county by dividing total people registered in each party for that county by the total people registered to vote in that county.

3.2 Data Processing

All data above was processed using Python within the Jupyter Notebook environment [41], [42]. An abridged version of the Python code used can be found in Appendix A.

Demographic data groupings for ages were redistributed. Age data was regrouped from groups of 5 years to the following groups: 0-19, 20-39, 40-49, 50-59, 60-69, 70-79, 80+. Race/ethnicity and industry data were originally subdivided by gender, but were summed together to create a total value. In the case of age and race/ethnicity, percentages of total population were attained for both the age and race/ethnicity variables by dividing each variable by the total population for each county. This was not needed for the industry data since the percentage values could be summed from the percentages sourced from the API.

Demographic data were imported as data frames and merged together by county, then merged with COVID-19 data restricted to the most recent date at the time of analysis (5 July 2021). Variables of cumulative cases per 100,000 and cumulative deaths per 100,000 were created using the population and cumulative case count and cumulative deaths count features within the newly merged data frames, and will be the variable of interest for this analysis.

Counties missing census data due to low population figures (< 65,000 people) were removed for two key reasons: 1) they were missing a substantial amount of demographic data and 2) low population counties are more likely to be subject to high variability in COVID spread compared to more populous counties. Counties removed were: Del Norte, Siskiyou, Modoc, Trinity, Lassen, Plumas, Glenn, Sierra, Colusa, Alpine, Amador, Calaveras, Tuolumne, Mono, Mariposa, San Benito, and Inyo counties. This leaves 41 counties in the analysis.

To reduce the dataset further, features that fell under the following categories were removed from the dataset: a) features missing data, b) racial features not defined as “percent of people of (a race) alone or combination with another race”, c) daily data that changes over time (such as COVID-19-related daily figures), d) all other COVID data, e) raw number features already represented by a percent of the population, and f) all other redundant or obsolete features (such as working population over 16 years).

3.3 Description of Cumulative Cases and Deaths per 100,000

A summary of cases and deaths per 100,000 were generated, including mean, quartile values, minimum, and maximum values. A histogram of these features were plotted using Matplotlib [43]. A heatmap of cases and deaths per 100,000 were generated by way of using the GeoPandas package to merge the relevant data to the shapefile of Californian counties [44], [45].

3.4 Calculating Correlations

Pearson correlation coefficients (r values) as well as r^2 values were calculated between all population features, cumulative cases per 100,000, and cumulative deaths per 100,000 as of 5 July, 2020. Feature pairs and their corresponding correlation values were exported into a Microsoft Excel spreadsheet which was filtered to find the highest and lowest correlation values for each feature, by sorting by r^2 in the appropriate direction.

For the purposes of describing correlations, a pair of features were considered "strongly" correlated if the absolute value of their correlation coefficient is greater than 0.5 ($|r| \geq 0.5$), moderately correlated if the absolute value is between 0.5 and 0.2 ($0.5 \geq |r| \geq 0.2$), and weakly correlated if the absolute value is between 0.2 and 0.1 ($0.2 \geq |r| \geq 0.1$) Any correlation whose absolute value is less than 0.1 ($|r| \leq 0.1$) is considered to be almost not-existent. Note that while it is fact that the closer the absolute value of r is to 1, the closer the relationship between the features resembles a linear relationship.

3.5 Linear Regression

Linear models were created to further quantify the strength of the relationship between the most highly correlated features and our target features—cumulative cases per 100,000 and cumulative deaths per 100,000—especially when multiple input features are used within the same model. This approach specifically was used due to the ease of implementation compared to other models. The purpose of these models is to prove the correlations between the chosen features and cases and deaths per 100,000 through tests for significance for the model overall and each feature specifically. Testing whether these models fail the assumptions of a linear regression model also determines if a (simple) linear regression model is adequate to explain such a relationship.

Using these features, two linear regression models were created using the Statsmodel library for Python [46]. The target variable and feature sets were assigned to each model as shown in Table 3.2. Model 1 contains the top 5 most significant features to cumulative cases per 100,000. Model 2 uses the same target variable, but attempts to minimize colinearity by selecting from a pool of the top 20 features by F regression test which are the not likely to be colinear with each other. To meet this requirement, all input features in Model 2 must have a Pearson correlation of less than 0.5 with all other input features. These features were selected manually according to this rule. Models 3 and 4 follow the same procedures, but for the target variable of cumulative deaths per 100,000 instead.

All models will be evaluated on their significance based on the F-test for overall significance and the t-test for the significance of each input variable. The former is to measure if the models performs statistically significantly better than a model with no input variables ($H_0 : \beta_1 = \beta_2 = \ldots = \beta_i$ where β signifies a coefficient in the linear model and i number of input variables vs. $H_1 : \beta_j \neq 0$ for any j integer in $[0, i]$). The latter test is used to determine the significance of a given input variable in the model ($H_0 : \beta = 0$,

Table 3.2: Feature Selection for Linear Regression Models

	Model 1	Model 2	Model 3	Model 4
Target Feature	Cases per 100,000	Cases per 100,000	Deaths per 100,000	Deaths per 100,000
Input Features	Top 5 Correlations to Target	Top 20 Correlations to Target; eliminate co-correlaitons	Top 5 Correlations to Target	Top 20 Correlations to Target; eliminate co-correlations

$H_1 : \beta \neq 0$ for β coefficient of a given input variable in the linear model).

In order for a linear model to be viable with no caveats, it must follow the following assumptions: 1) that the relationships between the target variable and each input variable are linear in nature, 2) that residuals are independent of each other, 3) that residuals are homoscedastic, or that they have a constant variance for all values of the target variable, and 4) that the residuals follow a normal distribution [47].

Assumption 1) was checked by performing the Linear Rainbow test for linearity, which tests if the "middle" 50% of residuals fit significantly worse than the whole distribution. (H_0 : The middle 50% of residuals do not differ significantly in fit compared to the entire set of residuals; H_1 : The middle 50% of residuals overall fit is significantly worse than for the entire set of residuals.) This test assumes homoscedasticity from the residuals, and may reject an otherwise linear model if assumption 3) is violated [48].

Assumption 2) was assessed by calculating the Variable Inflation Factors (VIF) values for the input variables of each model. A rule of thumb is that if any of these values exceeds 5, then there is likely a linear relationship between at least two of the input variables, which would violate this assumption [49].

Assumption 3) was checked by performing the Breusch-Pagan test for heteroscedasticity on the residuals of a linear regression model. (H_0 : The variance of errors from the model are homoscedastic, H_1: The variance of errors from the model are *not* homoscedastic, but heteroscedastic.) [50]. The Koenker variant of this test was used as it does not assume the residuals follow a normal distribution [51].

Lastly, assumption 4) was checked by performing the Jarque-Bera test for normality of residuals. The test detects if the residuals of the linear model have the skewness and kurtosis that match a normal distribution, with the null hypothesis stating that the distribution is normal. (H_0 : The distribution of the residuals follows a normal distribution, H_1 : The distribution of the residuals does *not* follow a normal distribution) [52].

All tests were performed with the Statsmodels package [46]. As these are all hypothesis tests, the significance level these tests were evaluated against will be set at 0.95 (or $\alpha = 0.05$), just like with the significance of each input variable and the significance of the model in general.

3.6 Calculating Co-Correlations

Input features from all models were evaluated for other demographic features which they were strongly correlated with (defined as $|r| \geq 0.50$ between the feature pair). These relationships will be defined as "co-correlations" or "co-correlated features." The procedure for this was to create a function that iterates through the list of each models features and counts the number of times each feature outside of that set is strongly correlated with each item in that list. The resulting data of co-correlations and number of model features which they were correlated with were merged into a single data frame and exported to a .csv file.

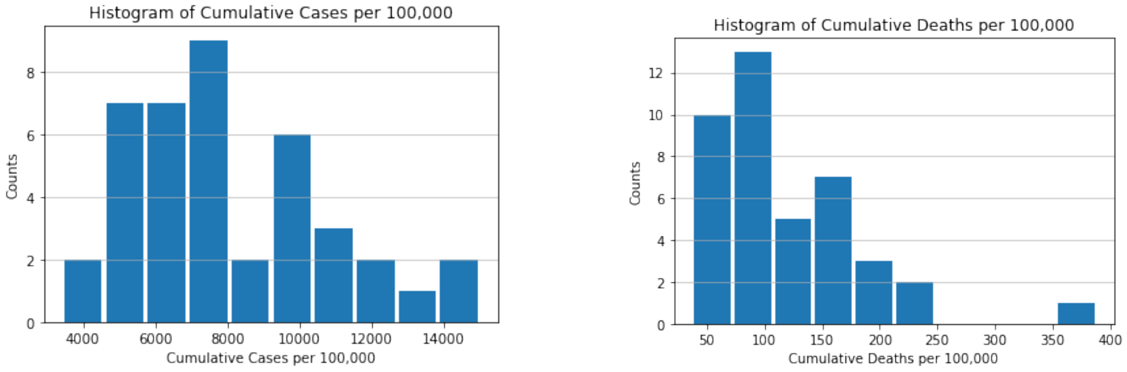
4 Results

4.1 Distribution of Target Features

After filtering counties for which demographic data was missing, 41 counties remained. Between the remaining counties, cumulative cases per 100,000 and cumulative deaths per 100,000 had means of 8068 and 119 (rounded to the nearest whole number), standard deviations of 2788 and 66, and IQR ranges of 5947 to 9957 and 73 to 156, respectively. Minimum and maximum values for cumulative cases per 100,000 were 3418 and 14,999. For cumulative deaths per 100,000, minimum and maximum values were 37 and 388.

Plotting cumulative cases (Figure 4.1a) per 100,000 shows an approximately normal distribution skewed slightly right. Cumulative deaths per 100,000 (Figure 4.1b) is skewed strongly right with one clear outlier (Imperial county). According to the county maps for cases and deaths (represented in Figures 4.2 and 4.3 respectively), cases and deaths in southern and inland counties were usually higher compared to northern and coastal counties.

The counties with the top 5 cumulative cases per 100,000 values were (in descending order) Imperial, Kings, San Bernardino, Riverside, and Los Angeles counties, and the bottom 5 (in ascending order) were Humboldt, San Francisco, Mendocino, Nevada, and Alameda counties. For cumulative deaths per 100,000, the counties with the top 5 highest values, in descending order, were Imperial, Los Angeles, San Bernardino, Inyo, and San Joaquin counities, with the bottom 5, in ascending order, being Mariposa, Mono, Trinity, Del Norte, and Plumas counties.



(a) Cases per 100k by County

(b) Deaths per 100k by County

Figure 4.1: Histograms of Cases and Deaths per 100,000 by California Counties

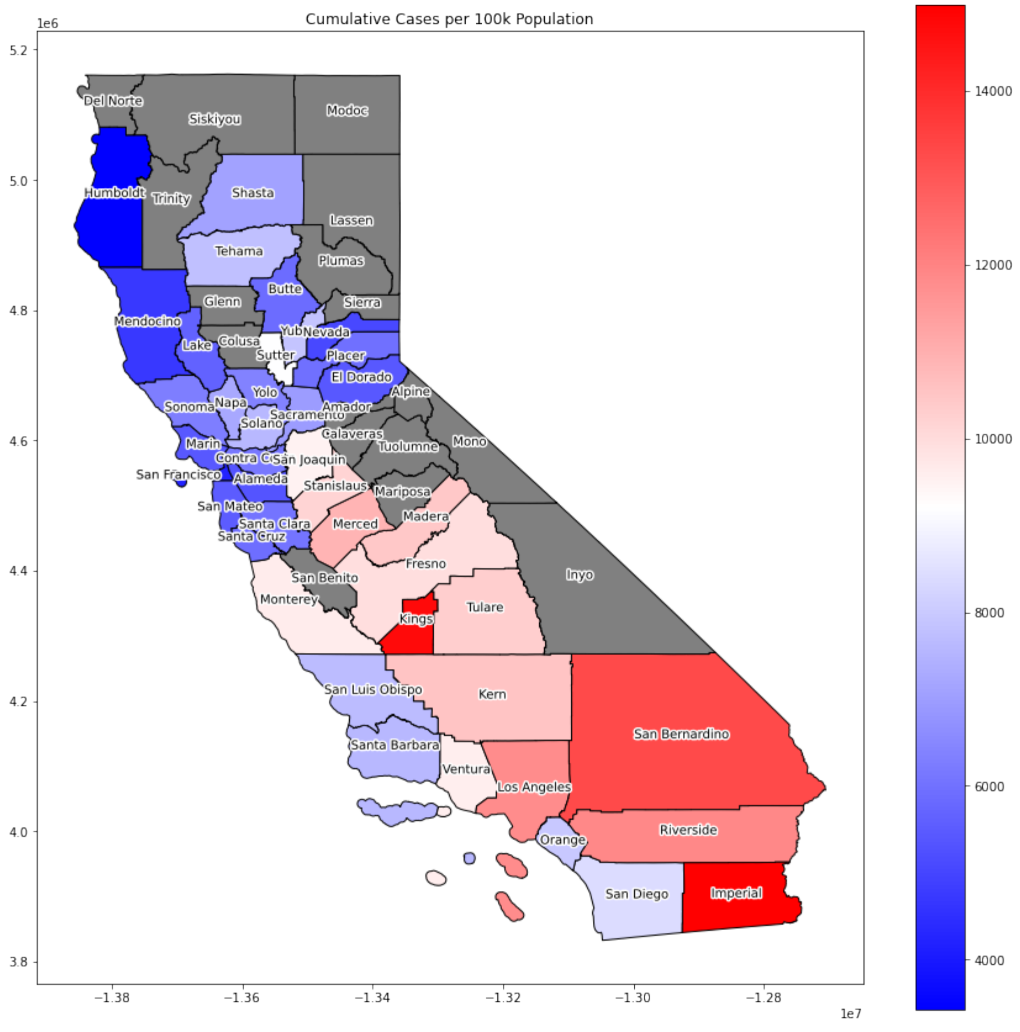


Figure 4.2: County Map by Cases per 100,000

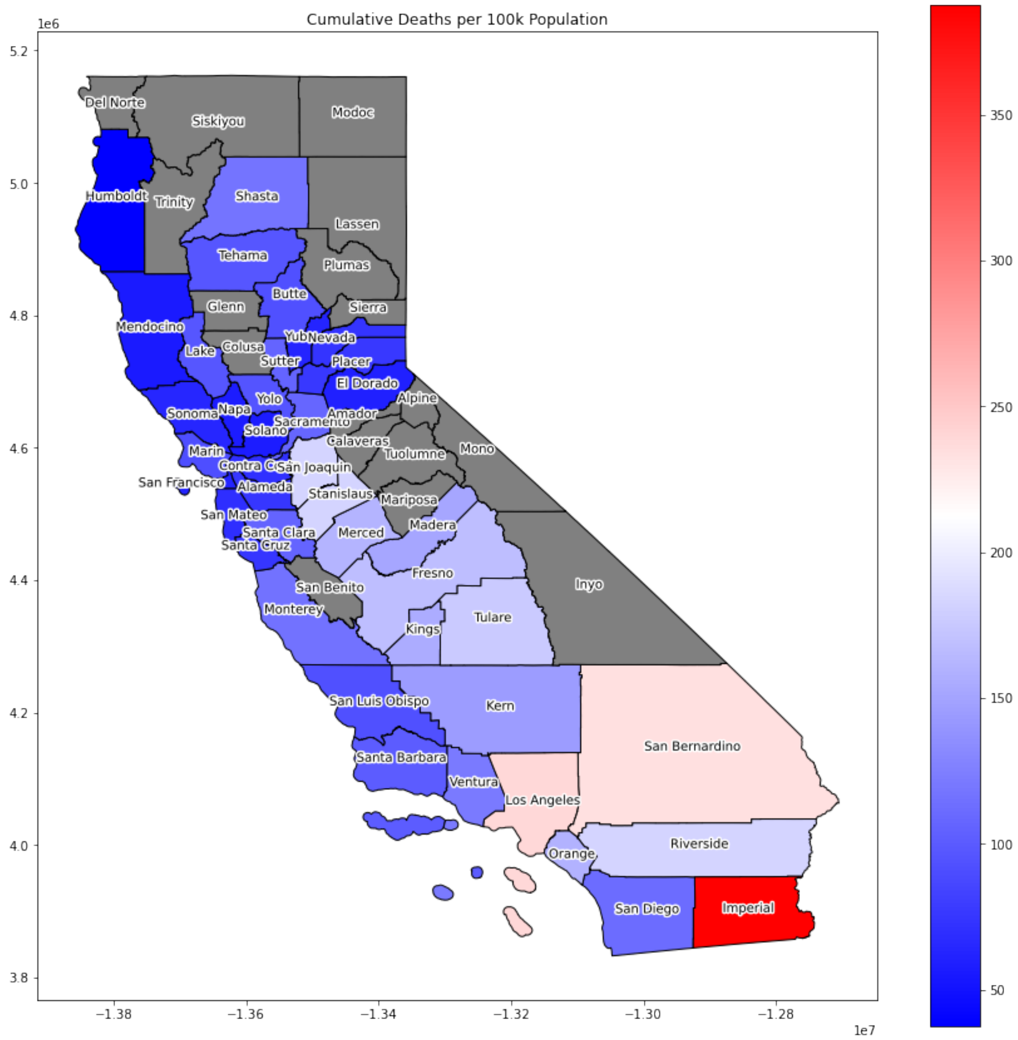


Figure 4.3: County Map by Deaths per 100,000

4.2 Correlations with Cases per 100,000 and Deaths per 100,000

4.2.1 Race, Ethnicity and Sex

Percentage of female population correlates moderately with lower values of cumulative cases per 100,000 (-0.480) and fewer cumulative deaths per 100,000 (-0.242). Correlations for male population are directly opposite to female population correlations (0.480, 0.242).

Most racial categories had moderate to weak correlations with cases per 100,000 and deaths per 100,000 with exception to percentages of individuals of two or more races and Hispanic ethnicity. Hispanic ethnicity strongly, positively correlates with cases and deaths per 100,000 (0.875, 0.795). Two or more races correlates negatively and moderately for cases but strongly and negatively for deaths (-0.494, -0.557). Asian (-0.267, -0.167), Native Hawaiian or Pacific Islander (-0.186, -0.228), and American Indian or Alaska Native (-0.092, -0.129) populations had moderate and weak correlations with cases and deaths at best. Black (0.257, 0.135) populations had a moderate correlation with cases and weak correlation with deaths while White (0.129, 0.078) populations had weak positive correlations with cases but essentially no correlation with deaths.

4.2.2 Age

Younger populations are linked with increases in cumulative cases and deaths per 100,000 while older populations are linked with a decrease in these outcomes. The 0-19 group is strongly correlated with both of these variables (0.752, 0.604) and a substantially weaker correlation exists for the 20-39 group (0.270, 0.210). On the other hand, age groups of 50-59 or greater are strongly associated with fewer cases and deaths per 100,000.

4.2.3 Education Level

As a general rule, higher levels of education correlate with fewer cumulative cases and deaths per 100,000 while the opposite is true for lower levels of education. Categories of “less than 9th grade,” “9th to 12th grade, no diploma”, “High school graduate (or equivalent)”, and “Some College, no degree” all correlate positively with cases (0.683, 0.747, 0.443, 0.101 respectively). The first three of these features correlate positively with cumulative deaths per 100,000 (0.528, 0.632, 0.346 respectively) with “Some College” having essentially no correlation with this variable (0.015). “Associate’s degree”, “Bachelor’s degree”, and “Graduate or professional degree” all correlate to fewer cases per 100,000, with the relationship being strongest for the last two education levels (-0.256, -0.591, -0.580). Similar relationships can be found between the above features and cumulative deaths per 100,000 (-0.326, -0.418, -0.444).

4.2.4 Health Insurance Status

Greater proportions of people possessing health insurance correlates negatively with COVID-19 cases and deaths per 100,000 (-0.647, -0.517) as does the proportion of people possessing private health insurance (-0.577, -0.570). However, the correlation between cases and deaths becomes moderately positive for the proportion of those with public health insurance (0.326, 0.341).

4.2.5 Disability Status

The proportion of non-institutionalized civilians with a disability correlates positively with COVID-19 cases and deaths per 100,000 (0.513, 0.399).

4.2.6 Preventable Illness Incidence Rates

Most preventable illnesses were found to have a slight positive correlation with COVID cases. Long-term diabetes complications were found to have the highest correlation with COVID incidence, followed by instances of lower-extremity amputations (0.691, 0.490). Short-term diabetes complications, uncontrolled diabetes, community-acquired pneumonia, hypertension, and urinary tract infections had less strong correlations to COVID cases (0.223, 0.299, 0.209, 0.257, 0.346). Heart failure, asthma in younger adults (ages 18-39), and COPD or Asthma in Older Adults (ages 40+) had very weak to no correlation with COVID cases (0.188, -0.056, 0.021).

4.2.7 Industry Employment

While most sectors had weak or moderate correlations, a few stand out as having stronger correlations to COVID cases. “Agriculture, forestry, fishing and hunting, and mining” (0.536, 0.306), “Transportation and warehousing, and utilities” (0.631, 0.509), and “Wholesale trade” (0.541, 0.481) had the strongest positive correlations with cumulative cases and deaths per 100,000 whereas “Finance and insurance, and real estate and rental and leasing” (-0.500, -0.344) and “Professional, scientific, and management, and administrative and waste management” (-0.488, -0.348) had the strongest negative correlations. Industries with moderate-to-weak correlations include “Public administration” (0.314, 0.187), which had positive correlations with cases and deaths. “Arts, entertainment, and recreation, and accommodation and food services” (-0.308, -0.175), “Educational services, and health care and social assistance” (-0.202, -0.123), “Information” (-0.352, -0.171), and “Other services, except public administration” (-0.206, -0.127) were negatively correlated to cases and deaths. “Construction” (-0.013, 0.045), “Manufacturing” (-0.012, -0.066), and “Retail trade” (0.031, 0.060) had virtually no correlation with cases or deaths.

4.2.8 Commute Methods and Duration

Commuting by car by carpooling (0.354, 0.209) and alone (0.486, 0.339), was found to be positively correlated with cumulative cases and deaths per 100,000. Working from home (-0.601, -0.434) was found to have the strongest negative correlation with cases and deaths out of all features describing commute methods. However, walking (-0.385, -0.242), public transportation (-0.358, -0.227), and “other means” (-0.134, -0.120) were all found to have moderate-to-weak negative correlations with cases and deaths. This is a surprising result that will be explored in a later section. Mean travel time to work had no correlation with either cases or deaths (0.019, 0.034).

4.2.9 Language Proficiency

Language proficiency (or lack thereof) and whether or not English was the only language spoken at home had almost no correlation with COVID incidence. Proportion of those who only speak English “less than ’very well’” (0.016, 0.053) as well as those who speak only English at home (0.016, 0.053) both had no correlation to cases or deaths

4.2.10 Party Affiliation

Correlations with party affiliations appear moderate to almost non-existent. Percentages of registered Democrats (-0.201, -0.090) and Republicans (0.264, 0.097) correlated moderately negatively and moderately positively in relation to cumulative cases per 100,000, but had no significant correlation with cumulative deaths per 100,000.

4.2.11 Overcrowding and Family Size

Average family size (0.816, 0.777) and overcrowding (>1.0 people per room) (0.681, 0.593) correlate strongly with both cases and deaths. There is a moderate correlation between severe overcrowding (>2.0 people per room) (0.361, 0.347) and cases and deaths, respectively.

4.2.12 Residence Continuity

The proportion of people who lived in the same residence a year ago (0.353, 0.387) and the proportion of those who lived in a different residence in the same county a year ago (-0.164, -0.057) weakly correlate to cumulative cases per 100,000 negatively and positively respectively. However, the correlation with deaths per 100,000 only exists for the former with no correlation with the latter variable.

Table 4.1: Feature Correlations to Cases per 100k and Deaths per 100k

Feature correlations		
Race/Ethnicity (% of population)	**cases**	**deaths**
American Indian or Alaska Native	-0.092	-0.129
Asian	-0.267	-0.167
Black	0.257	0.135
Hispanic	0.875	0.795
Native Hawaiian or Pacific Islander	-0.186	-0.228
Two or more races	-0.494	-0.557
White	0.129	0.078
Sex	**cases**	**deaths**
Female	-0.480	-0.242
Registered Party Affiliation (% of registered voters)	**cases**	**deaths**
Democrat	-0.201	-0.090
Republican	0.264	0.097
Education Level (% population $\geq$ 25 yrs)	**cases**	**deaths**
Less than 9th grade	0.683	0.528
9th to 12th grade, no diploma	0.747	0.632
High school graduate (or equivalent)	0.443	0.346
Some college, no degree	0.101	0.015
Associate's degree	-0.256	-0.326
Bachelor's degree	-0.591	-0.418
Graduate or professional degree	-0.580	-0.444
Employment Status (% population)	**cases**	**deaths**
Civilian labor force	-0.389	-0.286
Not in labor force	0.313	0.288
In labor force	-0.313	-0.288
Employed (Civilian)	-0.439	-0.360
Unemployed (Civilian)	0.449	0.501
Armed Forces	0.434	0.059
Health Insurance Status (% civilian population)	**cases**	**deaths**
No Health Insurance	0.647	0.517
Has Health Insurance	-0.647	-0.517
Private Health Insurance	-0.577	-0.570
Public Health Insurance	0.326	0.341
Industry (% employed population)	**cases**	**deaths**
Agriculture, forestry, fishing and hunting, and mining	0.536	0.309
Arts, entertainment, and recreation, and accommodation and food services	-0.308	-0.175
Construction	-0.013	0.045
Educational services, and health care and social assistance	-0.202	-0.123

Finance and insurance, and real estate and rental and leasing	-0.500	-0.344
Information	-0.352	-0.171
Manufacturing	-0.012	-0.066
Other services, except public administration	-0.206	-0.127
Professional, scientific, and management, and administrative and waste management services	-0.488	-0.348
Public administration	0.314	0.187
Retail trade	0.031	0.060
Transportation and warehousing, and utilities	0.631	0.509
Wholesale trade	0.541	0.481
Commuting (% employed population)	**cases**	**deaths**
Car, truck, van, carpooled	0.354	0.209
Car, truck, van, drove alone	0.489	0.339
Other means	-0.134	-0.120
Public Transportation	-0.358	-0.227
Walked	-0.385	-0.242
Worked from home	-0.601	-0.434
Mean travel time to work (min)	0.019	0.034
Language Skills (% of population $\geq$ 5yrs)	**cases**	**deaths**
Speak English less than “very well”	0.016	0.053
English only	0.016	0.053
Residence Continuity (% population)	**cases**	**deaths**
Lived in same house 1 year ago	0.353	0.387
Lived in different house, same county 1 year ago	-0.164	-0.057
Disability Status (% of population)	**cases**	**deaths**
Noninstutitionalized Civilian Population with a Disability	0.513	0.399
Preventable Illness (Incidence per 100,000)	**cases**	**deaths**
Asthma in Younger Adults (Ages 18-39)	-0.056	0.072
Community-Acquired Pneumonia	0.209	0.236
COPD or Asthma in Older Adults (Age 40+)	0.021	0.013
Diabetes Long-term Complications	0.690	0.578
Diabetes Short-term Complications	0.223	0.072
Heart Failure	0.188	0.143
Hypertension	0.257	0.252
Lower-Extremity Amputation (Diabetes)	0.490	0.267
Uncontrolled Diabetes	0.299	0.190
Urinary Tract Infection	0.346	0.351
Housing Density	**cases**	**deaths**
Average family size	0.816	0.777
Overcrowding	0.681	0.593
Severe Overcrowding	0.361	0.347

Age Categories	cases	deaths
0-19	0.752	0.604
20-39	0.270	0.210
40-49	-0.092	-0.058
50-59	-0.492	-0.390
60-69	-0.659	-0.540
70-79	-0.622	-0.500
80+	-0.623	-0.464

4.3 Linear Models

4.3.1 Cases per 100,000

Model 1

Model 1 used the top 5 correlated features to cases per 100,000, which consists of the following features: percent of population identifying as "Hispanic ethnicity" (%) ($r = 0.875$), "average family size" (0.816), "Age 0-19" (%) (0.752), education level of "9th to 12th grade, no diploma" (% $\geq$ 25 years) (0.747), and incidences of "Long-term Diabetes Complications" (per 100,000 of population) (0.690). The overall model was found to be statistically significant by F-test of overall significance ($F = 29.12, p > 0.001$). Out of the selected input variables, only Hispanic ethnicity was found to be significant to the model at significance level of $\alpha = 0.05$ ($t = 2.473$, $p = 0.018$).

This model violates the assumption that the above input variables are not colinear with each other. Calculating the Variance Inflation Factor (VIF) of each variable show that Hispanic ethnicity and average family size have VIF values equal to 7.595 and 5.269. Since a recommendation is that VIF values greater than 5 suggests high collinearity with other input variables [49], we can assume that this model does a poor job of establishing correlations between each separate input variable to the target variable of cumulative cases per 100,000.

However, Model 1 did exceed the level of significance required to accept the corresponding alternative hypotheses for Jarque-Bera test for normality, the Breusch-Pagan test for heteroscedacity, or the Linear Rainbow test for linearity at our chosen significance level. ($p = 0.470, 0.114, 0.474$ respectively at $\alpha = 0.05$)

Model 2

Model 2 is composed of the following features: "overcrowding" (percent of households) ($r = 0.681$), "Wholesale trade" (% workers) (0.541), "Transportation, warehousing, and utilities" (% workers) (0.631), and education level of "Graduate or professional degree" (% population $\geq$ 25 years) (-0.580). The overall model was found to be significant by F-test of overall significance ($F = 35.32, p < 0.001$). The variables of "overcrowding," "Transportation, warehousing, and utilities," and "Graduate or professional degree" were all found to be significant at significance level of $\alpha = 0.05$ ($t = 5.288$, $p < 0.001$, $t = 4.384$, $p < 0.001$, and $t = -2.452$, $p < 0.019$ respectively). There was not significant evidence to suggest that Model 2 violated any assumptions of the linear model. It did not fail the Jarque-Bera test for normality, the Breusch-Pagan test for heteroscedacity, or the Linear Rainbow test for linearity at our chosen significance level. ($p = 0.352, 0.276, 0.547$ respectively). The VIF values for each input feature for Model 2 never exceed 5, but this assumption is violated as seen in Table 4.2. Therefore, it is reasonable to assume that multicollinearity is minimal in our model, which is to be expected given the way our input features were selected.

4.3.2 Deaths per 100,000

Model 3 is composed of the following features:overcrowding (percent of households) ($r = 0.593$), Hispanic ethnicity (0.795), average family size (0.777), Age 0-19 (as percent of population) (0.604), and education level of 9th to 12th grade, no diploma (Percent >= 25 years) (0.632). The model was found to be significant overall by F-test of overall significance ($F = 17.05, p < 0.001$). Only Hispanic ethnicity and average family size were found to be significant features according to t-test ($t = 2.713, p = 0.010$ and, $t = 2.086, p = 0.044$ for each feature respectively).

Model 3 mostly appears to follow the assumptions of the linear model except for the assumption of independence between input variables. It did not fail the Jarque-Bera test for normality, the Breusch-Pagan test for heteroscedacity, or the Linear Rainbow test for linearity at our chosen significiance level. ($p = 0.600, 0.216, 0.269$ respectively). However, the VIF values for Ages 0-19 and average family size are 9.568 and 4.755 respectively. This suggests the input variables (as with Model 1) are highly collinear with each other.

Model 4 is composed of the same features as Model 2: "overcrowding" (percent of households) ($r = 0.593$), "Wholesale trade" (% workers) (0.481), "Transportation, warehousing, and utilities" (% workers) (0.509), and education level of "Graduate or professional degree" (% population $\geq$ 25 years) (-0.444). The overall model was found to be significant by F-test of overall significance ($F = 11.53, p > 0.001$). By t-test, only "overcrowding" and "Transportation, warehousing, and utilities" were found to be significant features ($t = 3.244, p = 0.003$ and $t = 2.511, p = 0.017$ respectively).

However, while the model was found to be significant, it fails two of the four assumptions of a linear regression model. It fails the Jarque-Bera test for normality of residuals ($p < 0.001$) as well as the Linear Rainbow test ($p = 0.011$). The Breusch-Pagan test failed to reject the assumption that the data were homoscedastic ($p = 0.471$). None of the VIF values for these variables exceeded 2, indicating low multicollinearity in this model.

Table 4.2: Goodness of Fit Test Results for Linear Models

Model 1 - Cases		
Jarque-Bera Test	Test Statistic	P-value
	1.510	0.470
Breusch-Pagan test	Test Statistic	P-value
	8.718	0.114
Linear Rainbow	F-Statistic	P-value
	1.050	0.474
VIF values		
Feature	VIF	
constant	665.682	
Age 0-19	4.104	
Hispanic Ethnicity (%)	7.595	
Long Term Diabetes Complications (incidence per 100,000)	3.880	
Education: 9th to 12th grade, no diploma (%)	3.492	
Average family size	5.270	
Model 2 - Cases		
Jarque-Bera Test	Test Statistic	P-value
	2.088	0.352
Breusch-Pagan test	Test Statistic	P-value
	5.116	0.276
Linear Rainbow	F-Statistic	P-value
	0.957	0.547
VIF		
Feature	VIF	
constant	40.823	
Overcrowding (% households)	1.267	
Industry: Wholesale trade (% employed)	1.316	
Industry: Transportation and warehousing, and utilities (% employed)	1.334	
Education: Graduate or professional degree (% $\geq$ 25 yrs)	1.414	
Model 1 - Deaths		
Jarque-Bera Test	Test Statistic	P-value
	1.021	0.600
Breusch-Pagan Test	Test Statistic	P-value
	7.069	0.216
Linear Rainbow	F-Statistic	P-value
	1.385	0.269
VIF		
Feature	VIF	
constant	568.208	
Overcrowding (% households)	3.604	

Age 0-19	9.568	
Hispanic Ethnicity (%)	2.414	
Education: 9th to 12th grade, no diploma (%)	2.414	
Average family size	4.755	
Model 2 - Deaths		
Jarque-Bera Test	Test Statistic	P-value
	137.839	0.000
Breusch-Pagan Test	Test Statistic	P-value
	3.549	0.471
Linear Rainbow	F-Statistic	P-value
	3.310	0.011
VIF		
Feature	VIF	
constant	40.833	
Overcrowding (% households)	1.267	
Industry: Wholesale trade (% employed)	1.316	
Industry: Transportation and warehousing, and utilities (% employed)	1.334	
Education: Graduate or professional degree (% $\geq$ 25 yrs)	1.414	

4.4 Co-Correlated Features

Correlations between the Model 1's features overlap substantially. In addition to features all being correlated to each other, the following variables were co-correlations to all input features: overcrowding, Ages 60-69, Ages 80+, health insurance status (none or any), private health insurance, percent of employed working from home, and education level less than 9th grade. Co-correlations with at least 4 of the input features include the age groups of 50-59 and 70-79, the "Agriculture, forestry, fishing and hunting, and mining" industry group, Bachelor's degrees, the "Finance and insurance, and real estate and rental and leasing" industry group, and unemployment rates. Other correlated features can be found in Appendix C.

Co-correlations between Model 2 and 4's features and the main feature set have less overlap with each other than Model 1. Features co-correlated with 3 input features are 9th to 12th grade education, no diploma, long-term diabetes complications, and Ages 0-19. Features strongly correlated with 2 input features are proportion of those who work from home as their commute, health insurance status (any and none), the "Agriculture, forestry, fishing and hunting, and mining" industry group, the high school graduate (or equivalent) education, proportion of disabled population, average family size, and proportion of population identifying as Hispanic.

For Model 3, the features have many co-correlations in common with Model 1 given that both have a similar feature set. The features which co-correlate with all input features for Model 3 were long-term diabetes complications, working from home, health insurance status, age groups of 60-69 and 80+, and less than 9th grade education level. Co-correlated features with at least 4 of the input features were private health insurance, ages 50-59, ages 70-79, and “Agriculture, forestry, fishing and hunting, and mining” industry group. Other correlated features can be found in Appendix C.

5 Conclusion

5.1 Summary

The correlations suggest a strong link between COVID spread and the following characteristics: housing size and density, low education attainment, low access to health services, employment in specific industries, and minority status. This is consistent with existing literature on health outcomes as it pertains to social status. Interestingly, these correlations were stronger for cumulative cases per 100,000 compared to deaths per 100,000.

It is known that COVID spreads best indoors and where the infected are in close proximity to others [10], [11]. Those who live in especially crowded conditions may lack the ability to self-isolate from family and housemates. Low levels of education attainment correlates with fewer and more lower paying work opportunities [25]. Lacking health insurance is both an indicator of a lack of means to afford it and may lead to situations where care is accessed at later stages of illness, such as for diseases like diabetes. It is also well known that higher income and status leads to better health outcomes overtime, due to greater ability to meet one's own needs and decrease or eliminate poverty-related stressors [5].

Many of the industries that positively correlated with COVID spread were related to sectors considered "essential," such as agriculture, wholesale trade, transportation and warehousing, utilities, and similar industries. On the other hand, industries that had strong negative correlations with COVID spread included finance, insurance, real estate, information, professional, scientific, and management jobs. These industries are *not* considered essential and many of these positions would have most likely been reconfigured into remote positions. This suggests essential workers in essential industries are more likely to be at risk of contracting COVID since they would not have been able to do their work remotely.

Lastly, the limited correlation within the feature sets for Models 2 and 4 implies a limited dependence of each component feature on each other. In other words, the levels of graduate level education (or lack thereof), prevalence of "Wholesale trade" and "Transportation and warehousing, and utilities," and housing overcrowding may point to differing forces that led to higher rates of transmission and mortality. However, given that these features point to varying levels of socioeconomic status, this inference should be taken with caution.

5.2 Evaluation

The main strength of this study is its observation of a variety of demographic features all at once. While this study cannot achieve the same depth of a specific subset features that a more tailored study can, this study may provide a jumping off point for future work in tying those features to COVID incidence and mortality in a study where interacting features are easier to control for.

The use of the Pearson Correlation as the main metric for this study implicitly assumes any correlation will follow a linear relationship with few, if any, outliers. This is was not always the case for some variables. For example, the r value for commuting by public transportation suggests there is a moderate negative correlation (-0.358). However, not only does this not line up with research [10], [53], but it is apparent in Figure 5.1 that the correlation is being driven by a few outliers in the Bay Area (especially San Francisco). When eliminating these outlier counties, there appears to be almost no correlation between the use of public transportation and COVID outcomes.

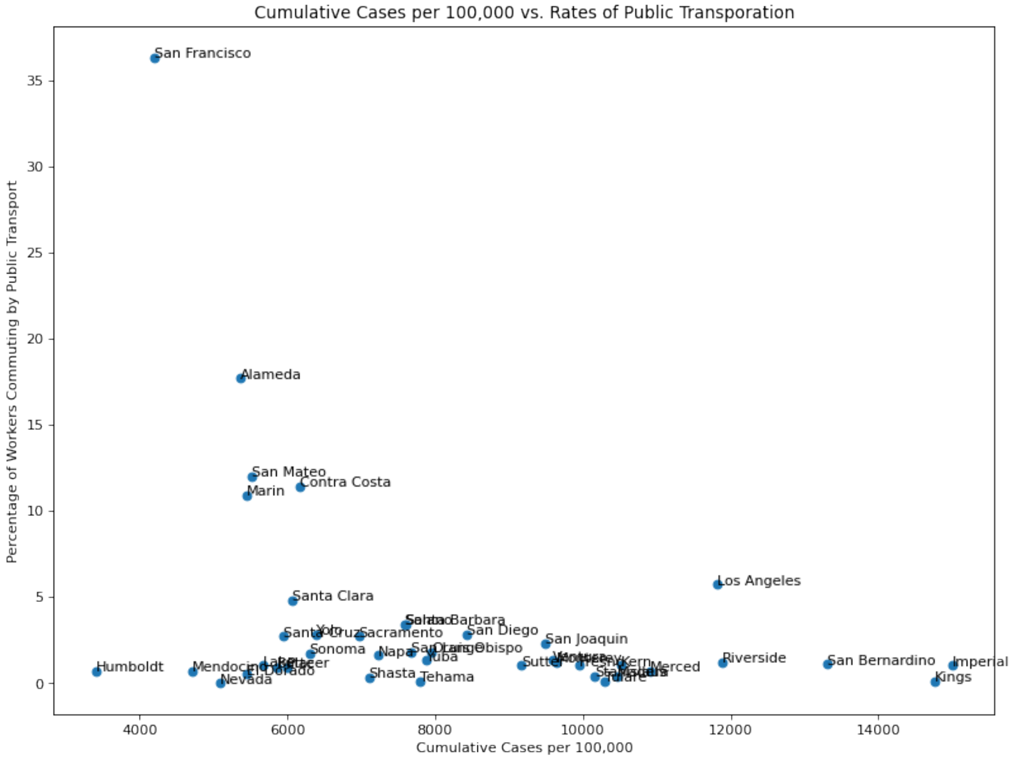


Figure 5.1: Graph of Cumulative Cases per 100,000 vs. Percentage of Workers Commuting by Public Transportation

Secondly, this study is limited in its ability to isolate the effects of every variable from each other. The correlation between higher proportions of people age 0-19 and higher incidences of cases and deaths seems like a contradiction considering that older individuals are typically at greater risk of severe illness and mortality [1]. However, this age group correlates positive with larger average family sizes and the prevalence of overcrowded households. Therefore, it is likely that the more people ages 0-19 there are in a population (relative to other age groups), the larger households are on average and more crowded homes there are on average. This means that children do not cause worse COVID outcomes per se, but that the proportion of children in a population correlates with other features that are more inherently harmful from a public health standpoint.

The correlation between certain groups within the population and COVID outcomes become less reliable as said groups make up smaller portions of the population. For example, the interquartile range for percent of White and Hispanic populations in a given county are 76.9% to 90.0% and 22.9% to 48.6% respectively. Meanwhile, Black, Native American/Alaska Native, and Native Hawaiian/Pacific Islander populations have much narrower IQRs in addition to making up a smaller share of county populations overall as can be see in Table 4.1. This means that conclusions based on correlations from these population percentages are limited to this narrow range of percentages. Furthermore, these correlations may be more descriptive of counties where people of these groups tend to make up more of the relative population rather than describe the actual correlation of racial populations to COVID cases, especially if said populations represent a fraction of a relatively small percent of most counties.

Furthermore, demographic data for each county is sourced from 2019, meaning that these data do not capture the effects the pandemic had on the population after the pandemic took hold in the United States. Some demographic data, such as unemployment levels, would have changed substantially throughout the pandemic. However, this was a deliberate choice for most of the data since the goal of the study was to see how these features would effect COVID outcomes dependent on how counties appeared before the onset of the pandemic. Additionally, attempting to account for features, such as unemployment, which were highly variable during this period, is outside the scope of this study.

5.3 Policy Suggestions

The evidence suggests that vaccination strategies must prioritize the most disadvantaged and essential workers if the disease is to be brought under control. At the time of this writing (8 September, 2021), only 67.4% of all Californians are fully vaccinated, with 22.5% of Californians completely unvaccinated [54]. Medi-Cal (Californian public insurance for low-income, non-elderly individuals) recipients are to receive monetary incentives for receiving their vaccinations to boost vaccination rates among low-income populations (though this program does not account for the uninsured) [55]. However, a substantial number of low income, uninsured individuals do not currently qualify for these incentives [16].

The vaccination program should also include efforts to educate everyone on the safety of the vaccine in an way that is inclusive to these groups. In a US nationwide poll by the Kaiser Family Foundation, about 21% and 16% of respondents who reported not getting the COVID-19 vaccine reported side effects and the newness of the vaccine as the main reason for their decision, respectively. Furthermore, many of these respondents were not resistant to getting a COVID vaccine in the future [56]. This suggests a need to communicate to the public that the vaccines are safe and that the risk of side effects outweigh the risk of the virus. Considering that many people in poverty may not be able to miss a day of work or access alternative childcare in order to acquire the vaccine *or* to recover from its side effects, efforts must also be sensitive to these concerns.

With the increased transmissiblity of the Delta variant, including in breakthrough cases of the vaccinated, continued measures, such as testing and social distancing should be practiced for the foreseeable future. The Delta variant is about 40 to 60% more transmissible than the original Alpha variant [57] and cases recently peaked at 33.2 thousand cases per 100,000 people statewide on 13 August, 2021 despite vaccination efforts [58].

Measures to assist those in overcrowded housing to self-isolate may also be beneficial. An Australian study suggested that the strategy of isolating COVID-19 patients in hotel rooms rather than their homes becomes more cost effective to the state as the patients household size increases and/or if the household has older individuals [59]. While this cost effectiveness analysis is not applicable to the United States, which ostensibly lacks a universal healthcare system, a similar strategy may still be effective in reducing the burden on hospitals, protecting larger and more vulnerable households, and reducing the number of people who will bear the burden of medical costs due to their COVID-19 related hospital stays.

5.4 Future Work

5.4.1 Research on Individual Factors

This study may serve as a jumping off point for future, more specific studies of these demographic features in relation to COVID outcomes. While the analysis found correlations at the population level, it does not directly isolate these correlations from each other.

An example study may involve an observational study aiming to determine if the size of a household or family increases risk of COVID-19 illness and death. This study would ideally control for socioeconomic status and overcrowding within the household. A similar study could be done to determine the effect of COVID-19 on racial and ethnic minorities to determine if simply being a specific minority puts one at increased risk of COVID illness and mortality even independent of racial and ethnic disparities in wealth.

5.4.2 Imperial County

Further investigation of Imperial County COVID-19 mortality rates may be beneficial in determining why this county is such an outlier in this regard, even when accounting for demographic features. A hypothesized cause for this is that a substantial number of people (Mexican and American citizens alike) that commute between the county and nearby Mexicali in Mexico [60]. This would increase the risk of transmission in times when Mexicali was experiencing their own COVID spike. Moreover, it may be that a substantial number of American citizens living in Mexicali (14 such cases were cited during a hospital surge at El Centro Hospital in May 2020 [61]) sought medical care in Imperial County, placing hospitals under even more pressure relative to population. Further research would be needed to clarify which factors drove COVID mortality and to what extent.

Bibliography

[1] CDC. (Feb. 11, 2020). “COVID-19 and Your Health”, Centers for Disease Control and Prevention, [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/people-with-medical-conditions.html (visited on 06/28/2021).

[2] H. Liang, M. A. Beydoun, and S. M. Eid, “Health needs, utilization of services and access to care among Medicaid and uninsured patients with chronic disease in health centres”, *Journal of Health Services Research and Policy*, vol. 24, no. 3, pp. 172–181, Jul. 1, 2019, ISSN: 1355-8196. DOI: 10.1177/1355819619836130. pmid: 31291765. [Online]. Available: https://jhu.pure.elsevier.com/en/publications/health-needs-utilization-of-services-and-access-to-care-among-med (visited on 09/09/2021).

[3] M. B. Cole, A. N. Trivedi, B. Wright, and K. Carey, “Health Insurance Coverage and Access to Care for Community Health Center Patients: Evidence Following the Affordable Care Act”, *Journal of General Internal Medicine*, vol. 33, no. 9, pp. 1444–1446, Sep. 2018, ISSN: 0884-8734. DOI: 10.1007/s11606-018-4499-2. pmid: 29845464. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6108997/ (visited on 09/09/2021).

[4] S. McMorrow, G. M. Kenney, and D. Goin, “Determinants of Receipt of Recommended Preventive Services: Implications for the Affordable Care Act”, *American Journal of Public Health*, vol. 104, no. 12, pp. 2392–2399, Dec. 2014, ISSN: 0090-0036. DOI: 10.2105/AJPH.2013.301569. pmid: 24432932. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4232157/ (visited on 09/09/2021).

[5] M. G. Marmot, G. D. Smith, S. Stansfeld, C. Patel, F. North, J. Head, I. White, E. Brunner, and A. Feeney, “Health inequalities among British civil servants: The Whitehall II study”, *Lancet (London, England)*, vol. 337, no. 8754, pp. 1387–1393, Jun. 8, 1991, ISSN: 0140-6736. DOI: 10.1016/0140-6736(91)93068-k. pmid: 1674771.

[6] M. H. Algren, O. Ekholm, L. Nielsen, A. K. Ersbøll, C. K. Bak, and P. T. Andersen, “Associations between perceived stress, socioeconomic status, and health-risk behaviour in deprived neighbourhoods in Denmark: A cross-sectional study | BMC Public Health | Full Text”, *BMC Public Health*, Feb. 13, 2018. [Online]. Available: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-018-5170-x (visited on 07/02/2021).

[7] W. R. Milligan, Z. L. Fuller, I. Agarwal, M. B. Eisen, M. Przeworski, and G. Sella, “Impact of essential workers in the context of social distancing for epidemic control”, *PLOS ONE*, vol. 16, no. 8, e0255680, Aug. 4, 2021, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0255680. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0255680 (visited on 09/09/2021).

[8] J. M. Cox-Ganser and P. K. Henneberger, “Occupations by Proximity and Indoor/Outdoor Work: Relevance to COVID-19 in All Workers and Black/Hispanic Workers”, *American Journal of Preventive Medicine*, vol. 60, no. 5, pp. 621–628, May 1, 2021, ISSN: 0749-3797. DOI: 10.1016/j.amepre.2020.12.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0749379721000696 (visited on 09/08/2021).

[9] CDC. (Jun. 11, 2021). “COVID-19 and Your Health”, Centers for Disease Control and Prevention, [Online]. Available: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html (visited on 07/02/2021).

[10] C. C. Wang, K. A. Prather, J. Sznitman, J. L. Jimenez, S. S. Lakdawala, Z. Tufekci, and L. C. Marr, “Airborne transmission of respiratory viruses”, *Science*, Aug. 27, 2021. DOI: 10.1126/science.abd9149. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.abd9149 (visited on 09/08/2021).

[11] T. Greenhalgh, J. L. Jimenez, K. A. Prather, Z. Tufekci, D. Fisman, and R. Schooley, “Ten scientific reasons in support of airborne transmission of SARS-CoV-2”, *The Lancet*, vol. 397, no. 10285, pp. 1603–1605, May 1, 2021, ISSN: 0140-6736. DOI: 10.1016/S0140-6736(21)00869-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673621008692 (visited on 09/08/2021).

[12] P. Halvatsiotis, A. Kotanidou, K. Tzannis, E. Jahaj, E. Magira, M. Theodorakopoulou, G. Konstandopoulou, E. Gkeka, C. Pourzitaki, N. Kapravelos, S. Papoti, M. Sileli, C. Gogos, D. Velissaris, N. Markou, E. Stefanatou, G. Vlachogianni, E. Aimoniotou, A. Komnos, T. Zafeiridis, P. Koulouvaris, A. Armaganidis, A. Bamias, and G. Dimopoulos, “Demographic and clinical features of critically ill patients with COVID-19 in Greece: The burden of diabetes and obesity”, *Diabetes Research and Clinical Practice*, vol. 166, p. 108 331, Aug. 2020, ISSN: 01688227. DOI: 10.1016/j.diabres.2020.108331. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0168822720305830 (visited on 04/24/2021).

[13] World Health Organization. (Jul. 1, 2021). “WHO Coronavirus (COVID-19) Dashboard”, World Health Organization, [Online]. Available: https://covid19.who.int (visited on 07/02/2021).

[14] CDC. (Mar. 28, 2020). “COVID Data Tracker”, Centers for Disease Control and Prevention, [Online]. Available: https://covid.cdc.gov/covid-data-tracker (visited on 07/02/2021).

[15] T. Andrasfay and N. Goldman, “Reductions in 2020 US life expectancy due to COVID-19 and the disproportionate impact on the Black and Latino populations”, *Proceedings of the National Academy of Sciences*, vol. 118, no. 5, Feb. 2, 2021, ISSN: 0027-8424, 1091-6490. DOI: 10.1073/pnas.2014746118. pmid: 33446511. [Online]. Available: https://www.pnas.org/content/118/5/e2014746118 (visited on 09/10/2021).

[16] (Oct. 23, 2020). “Health Insurance Coverage of Adults 19-64 Living in Poverty (under 100% FPL)”, KFF, [Online]. Available: https://www.kff.org/other/state-indicator/poor-adults/ (visited on 09/09/2021).

[17] (Oct. 23, 2020). “Health Insurance Coverage of Adults 19-64”, KFF, [Online]. Available: https://www.kff.org/other/state-indicator/adults-19-64/ (visited on 09/09/2021).

[18] J. Tolbert, A. D. P. Nov 06, and 2020. (Nov. 6, 2020). “Key Facts about the Uninsured Population”, KFF, [Online]. Available: https://www.kff.org/uninsured/issue-brief/key-facts-about-the-uninsured-population/ (visited on 09/09/2021).

[19] E. Brunner and M. Marmot, “Social organization, stress, and health”, in *Social Determinants of Health*, M. Marmot and R. G. Wilkinson, Eds., 2nd ed, Oxford ; New York: Oxford University Press, 2006, pp. 6–30.

[20] L. Stein, “A Study of Respiratory Tuberculosis in Relation to Housing Conditions in Edinburgh”, *British Journal of Social Medicine*, vol. 4, no. 3, pp. 143–169, Jul. 1950, ISSN: 0366-0842. pmid: 14777885. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1037252/ (visited on 09/09/2021).

[21] H. Thomson, M. Petticrew, and D. Morrison, “Health effects of housing improvement: Systematic review of intervention studies”, *BMJ (Clinical research ed.)*, vol. 323, no. 7306, pp. 187–190, Jul. 28, 2001, ISSN: 0959-8138. DOI: 10.1136/bmj.323.7306.187. pmid: 11473906.

[22] J. Sharfstein, M. Sandel, R. Kahn, and H. Bauchner, “Is Child Health at Risk While Families Wait for Housing Vouchers?”, *American Journal of Public Health*, vol. 91, no. 8, pp. 1191–1192, Aug. 2001, ISSN: 0090-0036. pmid: 11499101. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1446743/ (visited on 09/09/2021).

[23] D. J. Pevalin, A. Reeves, E. Baker, and R. Bentley, “The impact of persistent poor housing conditions on mental health: A longitudinal population-based study”, *Preventive Medicine*, vol. 105, pp. 304–310, Dec. 2017, ISSN: 00917435. DOI: 10.1016/j.ypmed.2017.09.020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0091743517303419 (visited on 09/09/2021).

[24] K. Ahmad, S. Erqou, N. Shah, U. Nazir, A. R. Morrison, G. Choudhary, and W.-C. Wu, “Association of poor housing conditions with COVID-19 incidence and mortality across US counties”, *PLOS ONE*, vol. 15, no. 11, e0241327, Nov. 2, 2020, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0241327. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0241327 (visited on 08/31/2021).

[25] OECD, “Education at a Glance 2014: Highlights”, OECD, Text, Nov. 2014. [Online]. Available: http://dx.doi.org/10.1787/eag_highlights-2014-en (visited on 09/08/2021).

[26] “The Health Literacy of America’s Adults: Results From the 2003 National Assessment of Adult Literacy”, p. 76, 2003.

[27] A. M. Arozullah, S. Lee, T. Khan, and S. Kurup, “Low health literacy increases the risk of preventable hospital admission.”, *Journal of General Internal Medicine*, vol. 18, pp. 221–221, Apr. 2003, ISSN: 0884-8734. [Online]. Available: http://www.webofscience.com/wos/woscc/full-record/WOS:000182564300882?SID=C5MBLg6Hu5Qf3FAiP79 (visited on 08/02/2021).

[28] L. Paakkari and O. Okan, “COVID-19: Health literacy is an underestimated problem”, *The Lancet. Public Health*, vol. 5, no. 5, e249–e250, May 2020, ISSN: 2468-2667. DOI: 10.1016/S2468-2667(20)30086-4. pmid: 32302535. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7156243/ (visited on 08/02/2021).

[29] K. Mackey, C. Ayers, K. Kondo, S. Saha, S. Advani, S. Young, H. Spencer, M. Rusek, J. Anderson, S. Veazie, M. Smith, and D. Kansagara, “Racial and Ethnic Disparities in COVID-19-Related Infections, Hospitalizations, and Deaths : A Systematic Review”, *Annals of internal medicine*, vol. 174, no. 3, pp. 362–373, 2021. DOI: 10.7326/M20-6306.

[30] E. Price-Haywood, E. Price-Haywood, J. Burton, D. Fort, and L. Seoane, “Hospitalization and mortality among black patients and white patients with Covid-19”, *New England Journal of Medicine*, vol. 382, no. 26, pp. 2534–2543, 2020. DOI: 10.1056/NEJMsa2011686.

[31] M. Mutambudzi, C. Niedzwiedz, E. B. Macdonald, A. Leyland, F. Mair, J. Anderson, C. Celis-Morales, J. Cleland, J. Forbes, J. Gill, C. Hastie, F. Ho, B. Jani, D. F. Mackay, B. Nicholl, C. O’Donnell, N. Sattar, P. Welsh, J. P. Pell, S. V. Katikireddi, and E. Demou, “Occupation and risk of severe COVID-19: Prospective cohort study of 120 075 UK Biobank participants”, *Occupational and Environmental Medicine*, vol. 78, no. 5, pp. 307–314, May 1, 2021, ISSN: 1351-0711, 1470-7926. DOI: 10.1136/oemed-2020-106731. pmid: 33298533. [Online]. Available: https://oem.bmj.com/content/78/5/307 (visited on 09/09/2021).

[32] G. McCormack, C. Avery, A. K.-L. Spitzer, and A. Chandra, “Economic Vulnerability of Households With Essential Workers”, *JAMA*, vol. 324, no. 4, p. 388, Jul. 28, 2020, ISSN: 0098-7484. DOI: 10.1001/jama.2020.11366. [Online]. Available: https://jamanetwork.com/journals/jama/fullarticle/2767630 (visited on 07/02/2021).

[33] C. Kamis, A. Stolte, J. S. West, S. H. Fishman, T. Brown, T. Brown, and H. R. Farmer, “Overcrowding and COVID-19 mortality across U.S. counties: Are disparities growing over time?”, *SSM - Population Health*, vol. 15, p. 100 845, Jun. 12, 2021, ISSN: 2352-8273. DOI: 10.1016/j.ssmph.2021.100845. pmid: 34189244. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8219888/ (visited on 09/09/2021).

[34] Public Health Alliance of Southern California. (). “California Healthy Places Index Map”, The California Health Places Index (HPI), [Online]. Available: https://map.healthyplacesindex.org/ (visited on 09/16/2021).

[35] ATSDR (Agency of Toxic Substances and Disease Registry). (Aug. 30, 2021). "CDC/ATSDR SVI Fact Sheet | Place and Health | ATSDR", ATSDR (Agency of Toxic Substances and Disease Registry), [Online]. Available: https://www.atsdr.cdc.gov/placeandhealth/svi/fact_sheet/fact_sheet.html (visited on 09/16/2021).

[36] California Department of Public Health, *COVID-19 Time-Series Metrics by County and State - California Open Data*, Jun. 30, 2021. [Online]. Available: https://data.ca.gov/dataset/covid-19-time-series-metrics-by-county-and-state1 (visited on 07/01/2021).

[37] US Census Bureau, *American Community Survey (ACS)*. [Online]. Available: https://www.census.gov/programs-surveys/acs (visited on 07/09/2021).

[38] California Department of Public Health, *Percent of Household Overcrowding ($>$ 1.0 persons per room) and Severe Overcrowding ($>$ 1.5 persons per room) - California Open Data*, Jun. 30, 2021. [Online]. Available: https://data.ca.gov/dataset/percent-of-household-overcrowding-1-0-persons-per-room-and-severe-overcrowding-1-5-persons-per- (visited on 07/01/2021).

[39] California Office of Statewide Health Planning & Development, *Rates of Preventable Hospitalizations for Selected Medical Conditions by County (LGHC Indicator) - California Open Data*. [Online]. Available: https://data.ca.gov/dataset/rates-of-preventable-hospitalizations-for-selected-medical-conditions-by-county-lghc-indicator (visited on 07/08/2021).

[40] A. Padilla, *Statement of Vote: General Election November 3, 2020*, Nov. 3, 2020. [Online]. Available: https://elections.cdn.sos.ca.gov/sov/2020-general/sov/complete-sov.pdf (visited on 06/30/2021).

[41] J. Reback, W. McKinney, jbrockmendel, J. V. den Bossche, T. Augspurger, P. Cloud, gfyoung, Sinhrks, A. Klein, M. Roeschke, S. Hawkins, J. Tratner, C. She, W. Ayd, T. Petersen, M. Garcia, J. Schendel, A. Hayden, MomIsBestFriend, V. Jancauskas, P. Battiston, S. Seabold, chris-b1, h-vetinari, S. Hoyer, W. Overmeire, alimcmaster1, K. Dong, C. Whelan, and M. Mehyar, *Pandas-dev/pandas: Pandas 1.0.3*, Zenodo, Mar. 18, 2020. DOI: 10.5281/zenodo.3715232. [Online]. Available: https://zenodo.org/record/3715232 (visited on 08/10/2021).

[42] T. Kluyver, B. Ragan-Kelley, F. Pérez, M. Bussonnier, J. Frederic, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, S. Abdalla, and C. Willing, "Jupyter Notebooks—a publishing format for reproducible computational workflows", p. 4,

[43] J. D. Hunter, "Matplotlib: A 2D Graphics Environment", *Computing in Science Engineering*, vol. 9, no. 3, pp. 90–95, May 2007, ISSN: 1558-366X. DOI: 10.1109/MCSE.2007.55.

[44] K. Jordahl, J. V. den Bossche, M. Fleischmann, J. McBride, J. Wasserman, J. Gerard, A. G. Badaracco, A. D. Snow, J. Tratner, M. Perry, C. Farmer, G. A. Hjelle, M. Cochran, S. Gillies, L. Culbertson, M. Bartos, G. Caria, N. Eubank, sangarshanan, S. Rey, maxalbert, A. Bilogur, B. Ward, C. Ren, D. Arribas-Bel, Flavin, L. Wasser, L. J. Wolf, M. Journois, and abonte, *Geopandas/geopandas: V0.9.0*, Zenodo, Feb. 28,

2021. DOI: 10.5281/zenodo.4569086. [Online]. Available: https://zenodo.org/record/4569086 (visited on 09/14/2021).

[45] California Department of Technology. (Oct. 23, 2019). "CA Geographic Boundaries - California Open Data", California Open Data Portal, [Online]. Available: https://data.ca.gov/dataset/ca-geographic-boundaries (visited on 09/14/2021).

[46] S. Seabold and J. Perktold, "Statsmodels: Econometric and Statistical Modeling with Python", presented at the Python in Science Conference, Austin, Texas, 2010, pp. 92–96. DOI: 10.25080/Majora-92bf1922-011. [Online]. Available: https://conference.scipy.org/proceedings/scipy2010/seabold.html (visited on 09/02/2021).

[47] M. A. Poole and P. N. O'Farrell, "The Assumptions of the Linear Regression Model", *Transactions of the Institute of British Geographers*, no. 52, p. 145, Mar. 1971, ISSN: 00202754. DOI: 10.2307/621706. JSTOR: 621706.

[48] J. Utts, "The Rainbow Test for Lack of Fit in Regression", *Communications in Statistics - Theory and Methods*, vol. 11, pp. 2801–2815, Jan. 1, 1982. DOI: 10.1080/03610928208828423.

[49] R. M. O'brien, "A Caution Regarding Rules of Thumb for Variance Inflation Factors", *Quality & Quantity*, vol. 41, no. 5, pp. 673–690, Oct. 1, 2007, ISSN: 1573-7845. DOI: 10.1007/s11135-006-9018-6. [Online]. Available: https://doi.org/10.1007/s11135-006-9018-6 (visited on 09/10/2021).

[50] T. S. Breusch and A. R. Pagan, "A Simple Test for Heteroscedasticity and Random Coefficient Variation", *Econometrica*, vol. 47, no. 5, pp. 1287–1294, 1979, ISSN: 0012-9682. DOI: 10.2307/1911963. JSTOR: 1911963.

[51] R. Koenker, "A note on studentizing a test for heteroscedasticity", *Journal of Econometrics*, vol. 17, no. 1, pp. 107–112, Sep. 1, 1981, ISSN: 0304-4076. DOI: 10.1016/0304-4076(81)90062-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0304407681900622 (visited on 09/03/2021).

[52] C. M. Jarque and A. K. Bera, "Efficient tests for normality, homoscedasticity and serial independence of regression residuals", *Economics Letters*, vol. 6, no. 3, pp. 255–259, Jan. 1, 1980, ISSN: 0165-1765. DOI: 10.1016/0165-1765(80)90024-5. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0165176580900245 (visited on 09/03/2021).

[53] A. Heinzerling, M. J. Stuckey, T. Scheuer, K. Xu, K. M. Perkins, H. Resseger, S. Magill, J. R. Verani, S. Jain, M. Acosta, and E. Epson, "Transmission of COVID-19 to Health Care Personnel During Exposures to a Hospitalized Patient - Solano County, California, February 2020", *MMWR. Morbidity and Mortality Weekly Report*, Apr. 17, 2020. DOI: 10.15585/mmwr.mm6915e5. [Online]. Available: https://www.meta.org/papers/transmission-of-covid-19-to-health-care-personnel/32298249 (visited on 06/29/2021).

[54] California Department of Public Health, *COVID-19 Vaccine Progress Dashboard Data - California Open Data*, Jun. 30, 2021. [Online]. Available: https://data.ca.gov/dataset/covid-19-vaccine-progress-dashboard-data (visited on 07/01/2021).

[55] Office of Governor. (Aug. 11, 2021). "California Implements First-in-the-Nation Measure to Encourage Teachers and School Staff to Get Vaccinated", California Governor, [Online]. Available: https://www.gov.ca.gov/2021/08/11/california-implements-first-in-the-nation-measure-to-encourage-teachers-and-school-staff-to-get-vaccinated/ (visited on 09/09/2021).

[56] A. Kirzinger, G. Sparks, and M. Brodie, *KFF COVID-19 Vaccine Monitor: In Their Own Words, Six Months Later*, Jul. 13, 2021. [Online]. Available: https://www.kff.org/coronavirus-covid-19/poll-finding/kff-covid-19-vaccine-monitor-in-their-own-words-six-months-later/ (visited on 07/30/2021).

[57] *SPI-M-O: Consensus Statement on COVID-19*, Jun. 2, 2021. [Online]. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/993321/S1267_SPI-M-O_Consensus_Statement.pdf (visited on 09/08/2021).

[58] State of California. (Sep. 8, 2021). "Tracking COVID-19 in California", [Online]. Available: https://covid19.ca.gov/state-dashboard/ (visited on 09/09/2021).

[59] A. Melia, D. Lee, N. Mahmoudi, Y. Li, and F. Paolucci, "Cost-Effectiveness Analysis of COVID-19 Case Quarantine Strategies in Two Australian States: New South Wales and Western Australia", *Journal of Risk and Financial Management*, vol. 14, no. 7, p. 305, Jul. 4, 2021, ISSN: 1911-8074. DOI: 10.3390/jrfm14070305. [Online]. Available: https://www.mdpi.com/1911-8074/14/7/305 (visited on 09/09/2021).

[60] J. Small. (Jun. 17, 2020). "What's Behind a COVID-19 Spike in Imperial County", KQED, [Online]. Available: https://www.kqed.org/news/11824749/whats-behind-a-covid-19-spike-in-imperial-county (visited on 06/23/2021).

[61] J. Bowman. (May 19, 2020). "Citing Overnight Rises In Cases, Imperial County Hospitals Turn Away COVID-19 Patients", KPBS Public Media, [Online]. Available: https://www.kpbs.org/news/2020/may/19/citing-overnight-rises-cases-imperial-county-hospi/ (visited on 06/23/2021).

A Appendix – Source Code

A.1 Import Census Data from API call

```
# Import packages
import requests
import pandas as pd
import numpy as np
import json
import math

# Assign api key; omitted for security
api_key ='XXXXX'

# Import table including all data
table =pd.read_csv('codes_employment.csv')

# Set up string to call all variables from outside list
call_string =''
labels =[]
ind_mapper ={}

for i in range(len(table['Code'])):
    call_string += str(table["Code"][i]) + 'E&' + str(table["Code"][i]) + 'PE&'
    labels.append(table['Employment Status'][i])
    labels.append(str(table['Employment Status'][i]) + ' (Percent)')
    ind_mapper[str(table["Code"][i]) + 'E'] =table['Employment Status'][i]
    ind_mapper[str(table["Code"][i]) + 'PE'] =str(table['Employment Status'][i])
        ↪ + ' (Percent)'

# Takes json object from api call and applies labels through a mapper
def create_table(dump,mapper,labels_apply=False):
    columns =dump[0]
    df =pd.DataFrame(dump[1:], columns=columns)
    df =df.rename(columns=mapper)
    if labels_apply:
        df[labels] =df[labels].apply(pd.to_numeric)
    return df

# Sends call to retrieve data for all Californian counties
response3 =requests.get('https://api.census.gov/data/2019/acs/acs1/profile?get=
    ↪ NAME,'
                    +call_string+'for=county:*&in=state:06&key='+api_key)
json_dict =response3.json()

# Convert response into dataframe
df =create_table(json_dict,ind_mapper)
df.set_index('NAME',inplace=True)

# Save to dataframe
df.to_csv('emp_industry_commute.csv')
```

A.2 Merge Demographic Data

```
# Import Packages
import numpy as np
import pandas as pd

# Import age data
popage =pd.read_csv('C:/Users/shoyr/courses_stir/dissertation/datasets/custom/
    ↪ pop_by_county_agegroup.csv')

# Create copy of dataframe with renamed columns
agecats =[str(float(i)) for i in range(1,19)]
popagecopy =popage.copy()
for cat in agecats:
    popagecopy[cat + '_percentage'] =popagecopy[cat] / popagecopy['0.0']
    popagecopy.drop(cat, inplace=True,axis=1)
popagecopy.rename({'0.0':'Total_pop'},inplace=True,axis=1)

# Import overcrowding data, remove index column
housing =pd.read_csv('C:/Users/shoyr/courses_stir/dissertation/datasets/custom/
    ↪ housing_overcrowding_county.csv')
housing =housing.iloc[:,1:]

# Filter overcrowding data to filter out race/ethnicity subgroups,
housing =housing[housing['race_eth_name'] =='Total']
housing.drop('race_eth_name',axis=1,inplace=True)

# Rename counties to match otehr data
housing['county_name'] =housing['county_name'] + ' County'

# Merge housing and population data
concatted =popagecopy.merge(housing[['county_name','estimate','estimate_severe'
    ↪ ]], how='outer',
                        left_on='Geographic Area',right_on='county_name')
concatted.rename({'estimate': 'overcrowding','estimate_severe':'
    ↪ overcrowding_severe'}, inplace=True,axis=1)
concatted.drop('county_name',inplace=True,axis=1)

# Regroup age data into smaller groupings
concatted['Age_0-19'] =concatted['1.0_percentage'] + concatted['2.0_percentage']
    ↪ + concatted['3.0_percentage'] + concatted['4.0_percentage']
concatted['Age_20-39'] =concatted['5.0_percentage'] + concatted['6.0_percentage']
    ↪  + concatted['7.0_percentage'] + concatted['8.0_percentage']
concatted['Age_40-49'] =concatted['9.0_percentage'] + concatted['10.0_percentage'
    ↪ ]
concatted['Age_50-59'] =concatted['11.0_percentage'] + concatted['12.0_percentage
    ↪ ']
concatted['Age_60-69'] =concatted['13.0_percentage'] + concatted['14.0_percentage
    ↪ ']
concatted['Age_70-79'] =concatted['15.0_percentage'] + concatted['16.0_percentage
    ↪ ']
concatted['Age_80+'] =concatted['17.0_percentage'] + concatted['18.0_percentage']
```

```
# Drop old age groups
agecats =[str(float(i))+'_percentage' for i in range(1,19)]
for item in agecats:
    concatted.drop(item,axis=1,inplace=True)

# Import race/ethnic and sex data
popc =pd.read_csv('C:/Users/shoyr/courses_stir/dissertation/datasets/pop-county-
    ↪ char/cc-est2019-alldata-06.csv')

# Drop redundant columns
popc.drop(['SUMLEV','STATE','COUNTY','STNAME'],axis=1,inplace=True)

# Filter to total population race/ethnic groups and most recent year estimates
popc =popc[popc.AGEGRP ==0]
popc =popc[popc.YEAR ==12]

# Calculate percentages within population

# Sex
popc['F_percent'] =popc['TOT_FEMALE'] / popc['TOT_POP']
popc['M_percent'] =popc['TOT_MALE'] / popc['TOT_POP']

# Race/ethnic dems
# init list
dem_vars =['WA','BA','IA','AA','NA','TOM','WAC','BAC','IAC','AAC','NAC','NH','
    ↪ NHWA','NHBA',
          'NHIA','NHAA','NHNA','NHTOM','NHWAC','NHBAC','NHIAC','NHAAC','NHNAC','H
              ↪ ','HWA',
          'HBA','HIA','HAA','HNA','HTOM','HWAC','HBAC','HIAC','HAAC','HNAC']

# loop the list calculate percents for the above vars
for item in dem_vars:
    popc[item + '_percent'] =(popc[item + '_MALE']+popc[item + '_FEMALE'])/popc['
        ↪ TOT_POP']

# Create new dataframe with only percentages of selected groups
dem_cols =[item + '_percent' for item in dem_vars]
pop =popc.copy()
pop =pop[['CTYNAME','F_percent','M_percent'] + dem_cols]

# Merge overcrowding/age data with race/ethnic and sex data
concatted =concatted.merge(pop,how='outer',left_on='Geographic Area',right_on='
    ↪ CTYNAME')
concatted =concatted.drop('CTYNAME',axis=1)

# Import preventable hospitalization data
prev =pd.read_csv('C:/Users/shoyr/courses_stir/dissertation/datasets/preventable-
    ↪ hospitalizations/rates-of-preventable-hospitalizations-for-selected-
    ↪ medical-conditions-by-county-.csv')

# Filter to most recent data, avoid nans, filter out extra data
prev =prev[prev.Year ==2019]
prev =prev[prev.PQI.isin([1,3,4,5,6,7,8,10,11,12,13,14,15,16,17,18])]
prev =prev[['County','PQIDescription','ObsRate_ICD10','RiskAdjRate_ICD10',
              'AnnotationCode','AnnotationDesc']]

# Pivot labels for observed rates of hospitalizations
prev =prev.pivot(index='County',columns='PQIDescription', values=['ObsRate_ICD10'
    ↪ ])

# Merge previous data with preventable hospitalization data
cc =concatted.copy()
pp =prev.copy()
pp.index =pp.index + ' County'
cc =cc.merge(pp, left_on='Geographic Area',right_on='County')

# Merge previous data with api census data
d =pd.read_csv('emp_industry_commute.csv').iloc[:,1:]
d['NAME'] =d['NAME'].map(lambda x: x.rstrip(', California'))
new =cc.merge(d, left_on=['Geographic Area'], right_on=['NAME'])
new =new.drop(columns=['NAME'])

# Merge previous data with voter data
party =pd.read_csv(r'C:/Users/shoyr/courses_stir/dissertation/datasets/party_aff/
    ↪ party_aff.csv')
party =party[['County','Dem_percent','Rep_percent']]
party['County'] =party['County'] + ' County'

new =new.merge(party, left_on=['Geographic Area'],right_on=['County'])
new =new.drop(columns=['County'])

# Save result to .csv file
new.to_csv('all_statics.csv')
```

A.3 Merge COVID and Demographic Data

```
# Import packages

import numpy as np
import pandas as pd
from datetime import date

# Get today's date (MM-DD_YYYY)
today =date.today()

# Download COVID data
df =pd.read_csv('https://data.chhs.ca.gov/dataset/f333528b-4d38-4814-bebb-12
    ↪ db1f10f535/resource/046cdd2b-31e5-4d34-9ed3-b48cdbc4be7a/download/
    ↪ covid19cases_test.csv')
df =df.merge(df1, left_on=['date','area'], right_on=['todays_date','county'])
df =df.drop(columns=['todays_date','county'])[df['date'] <= '2021-07-05']
df['weekday'] =pd.to_datetime(df['date']).dt.weekday

# Import demographic data
st =pd.read_csv('all_statics.csv').iloc[:,1:]
```

```
# Modify county column to match COVID data
st['area'] =st['area'].str.replace(' County','')

# Merge data together
df =df.merge(st, how='left', left_on=['area'], right_on=['area'])

# Save whole dataframe
df.to_csv('all_time_{}.csv'.format(str(today)),index=False)
```

A.4 Plotting Data and Linear Regression

```
# Import packages
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import matplotlib as mpl
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf

import geopandas as gpd
import matplotlib.patheffects as pe

# Import Data
df =pd.read_csv('all_time_2021-07-27.csv')

# New Vars
df['cases_100k'] =df['cases'] / df['population'] * 100000
df['deaths_100k'] =df['deaths'] / df['population'] * 100000
df['percent_vaxxed'] =df['fully_vaccinated'] / df['population']
df['cum_cases_100k'] =df['cumulative_cases'] / (df['population'] / 100000)
df['cum_deaths_100k'] =df['cumulative_deaths'] / (df['population'] / 100000)

# Flatten data to 5 July 2021 Cumulatives
dfc =df.copy()
dfc =dfc[dfc['date'] =='2021-07-05']

# Filter counties missing substantial amount of data (as in smallest ones)
dfc =dfc.dropna(subset=['area','overcrowding','No Health Insurance (Percent)','
    ↪ H_percent'])

# Create and export data for cases and deaths per 100k
sorted_cases =df[df['date'] =='2021-07-05'].dropna(subset=['area','overcrowding',
    ↪ 'No Health Insurance (Percent)','H_percent'])[['area','cum_cases_100k','
    ↪ cum_deaths_100k']]
sorted_cases.to_csv('raw_cumulatives.csv', index=False)

# Export histograms of cases and deaths per 100k
plt.hist(sorted_cases['cum_cases_100k'], bins=10, rwidth=0.9)
plt.title('Histogram of Cumulative Cases per 100,000')
plt.xlabel('Cumulative Cases per 100,000')
plt.ylabel('Counts')
plt.grid(axis='y', alpha=0.75)
plt.show()
plt.savefig('cum_cases_hist.png')

plt.hist(sorted_cases['cum_deaths_100k'], bins=10, rwidth=0.9)
plt.title('Histogram of Cumulative Deaths per 100,000')
plt.xlabel('Cumulative Deaths per 100,000')
plt.ylabel('Counts')
plt.grid(axis='y', alpha=0.75)
plt.show()
plt.savefig('cum_deaths_hist.png')

# Import shapefile and merge with cases and deaths per 100k
test =gpd.read_file('county_shpfiles/CA_Counties_TIGER2016.shp')
test =sorted_cases.merge(test,how='right', left_on=['area'],right_on=['NAME'])
test['coords'] =test['geometry'].apply(lambda x: x.representative_point().coords
    ↪ [:])
test['coords'] =[coords[0] for coords in test['coords']]
test =gpd.GeoDataFrame(test)

# Create and export heatmaps
for item in [['cum_cases_100k', 'Cases'],['cum_deaths_100k', 'Deaths']]:
    fig, ax =plt.subplots(figsize =(15,15))
    test.plot(ax=ax, column=item[0], categorical=False, cmap='bwr',legend=True,
        ↪ edgecolor="black",
           missing_kwds={'color': 'grey'},)
    plt.title('Cumulative {} per 100k Population'.format(item[1]))
    for idx, row in test.iterrows():
        plt.annotate(text=row['NAME'], xy=row['coords'], horizontalalignment='
            ↪ center',
                path_effects=[pe.withStroke(linewidth=3, foreground="white")])
    plt.show()
    plt.savefig('{}_map.png'.format(item[1]))

# Rename some hospitalization names
dfc =dfc.rename(columns={"('ObsRate_ICD10', 'Urinary Tract Infection')": 'uti',
                     "('ObsRate_ICD10', 'Diabetes Long-term Complications')": '
                         ↪ long_term_diabetes'})

# Function reads a text file as a list, delimitted by return carriage
def file_to_list(filename):
    f =open(filename, 'r')
    flist =f.readlines()
    flist =[x[:-1] for x in flist]
    return flist

# Import list from text file
drop_me =file_to_list('drop_me.txt')

# Drops irrelevant variables through "drop_me" and exports exog and endog
def scale_values(df, target, drop_list =drop_me):
    # Make copy of original to prevent modification of global variable
    drop =drop_list.copy()

    # Prepare X and y for linear regression
    df_copy =df.copy()
    X =df_copy.drop(columns=[x for x in drop if x in df_copy.columns])
    if target in drop:
        drop.remove(target) # Remove target variable from column filter
    y =df_copy.drop(columns=[x for x in drop if x in df_copy.columns])[target]

    return X, y

# Create dataframe of all correlations as four-column frame
def get_corr_matrix(dfc):
    dfc_corrs =dfc.corr().stack().reset_index()
    dfc_corrs.columns =['f1','f2','r']
    dfc_corrs =dfc_corrs[(dfc_corrs.f1 != dfc_corrs.f2)]
    dfc_corrs['r2'] =dfc_corrs['r']**2
    dfc_corrs =dfc_corrs.sort_values(by=['r2'], ascending=False)
    return dfc_corrs

dfc_corrs =get_corr_matrix(df[df['date'] =='2021-07-05'].drop(columns=drop_me))

# Export dataframe of correlations to .csv
dfc_corrs.to_csv('all_corrs.csv', index=False)

# Function used to get dataframe of co-correlations for a given feature set
def get_model_corrs(df, feats, model_name):
    df_copy =df.copy()
    df_copy =df_copy[df_copy['f1'].isin(feats)]
    df_copy =df_copy[-df_copy['f2'].isin(feats)]
    items ={}
    for index, row in df_copy.iterrows():
        if row['r2'] >= 0.25 and row['f2'] in items.keys():
            items[row['f2']] += 1
        elif row['r2'] >= 0.25:
            items[row['f2']] =1
    newdf =pd.DataFrame.from_dict(items, orient='index', columns=[model_name])
    return newdf

# Takes above function and iterates over multiple feature lists
def all_model_corrs(df, feats, model_names):
    main_df =get_model_corrs(df, feats[0], model_names[0])
    for i in range(1, len(model_names)):
        new_df =get_model_corrs(df, feats[i], model_names[i])
        main_df =main_df.merge(new_df, left_index=True, right_index=True, how='
            ↪ outer')
    return main_df

# Define feature lists for each model
m1_feats =['H_percent', 'Average family size', 'Age_0-19', '9th to 12th grade, no
    ↪  diploma (Percent)',
         "('ObsRate_ICD10', 'Diabetes Long-term Complications')"]
m2_feats =['overcrowding','Wholesale trade (Percent)',
          'Transportation and warehousing, and utilities (Percent)',
           'Graduate or professional degree (Percent)']
m3_feats =['H_percent', 'Average family size', 'Age_0-19', '9th to 12th grade, no
    ↪  diploma (Percent)',
         "overcrowding"]

# Get correlation lists for all models and export to .csv
all_model_corrs(dfc_corrs, [m1_feats, m2_feats,m3_feats], ['m1','m2','m3']).
    ↪ to_csv('co_corrs.csv')

# Get X, y for linear regression
remove_list =['cum_cases_100k']
X, y =scale_values(dfc, 'cum_cases_100k', drop_list=set(remove_list+drop_me))

# Model 1
# Note: Code is almost identical for all other models and is thus omitted

# Exog selects only selected features; endog is y
exog =sm.add_constant(X[m1_feats])
endog =y.copy()

# Fit to model, select features recursively
mod =sm.OLS(endog, exog)
res =mod.fit()
print(res.summary())

# Test for heterostatisicity
sm.stats.diagnostic.het_breuschpagan(res.resid, exog)

# Test for non-linearity
sm.stats.diagnostic.linear_rainbow(res)

from statsmodels.stats.outliers_influence import variance_inflation_factor

# VIF dataframe
vif =pd.DataFrame()
vif["feature"] =exog.columns

# calculating VIF for each feature
vif["VIF"] =[variance_inflation_factor(exog.values, i)
                       for i in range(len(exog.columns))]

print(vif)

# Make influence plot
plt.rc("figure", figsize=(7,5))
fig =sm.graphics.influence_plot(res, criterion="cooks")
fig.tight_layout(pad=1.0)

# Make partial regression plots
plt.rc("figure", figsize=(15,10))
fig =sm.graphics.plot_partregress_grid(res)
fig.tight_layout(pad=1.0)
```

B Appendix – Output for Linear Models

B.1 Model 1

```
                        OLS Regression Results
==============================================================================
Dep. Variable: cum_cases_100k R-squared: 0.806
Model: OLS Adj. R-squared: 0.779
Method: Least Squares F-statistic: 29.12
Date: Fri, 20 Aug 2021 Prob (F-statistic): 1.52e-11
Time: 14:23:26 Log-Likelihood: -349.29
No. Observations: 41 AIC: 710.6
Df Residuals: 35 BIC: 720.9
Df Model: 5
Covariance Type: nonrobust
=================================================================================
    ↪ ==========================
                                       coef std err t P>|t| [0.025 0.975]
---------------------------------------------------------------------------------
    ↪ --------------------------
const -3305.6703 5287.235 -0.625 0.536 -1.4e+04 7427.988
Age_0-19 -292.8862 1.06e+04 -0.028 0.978 -2.17e+04 2.12e+04
H_percent 7971.6264 3223.780 2.473 0.018 1427.005 1.45e+04
long_term_diabetes 1.3844 13.488 0.103 0.919 -25.998 28.767
9th to 12th grade, no diploma (Percent) 212.5274 118.959 1.787 0.083 -28.973
    ↪ 454.028
Average family size 2030.6911 1670.421 1.216 0.232 -1360.443 5421.825
==============================================================================
Omnibus: 2.174 Durbin-Watson: 2.094
Prob(Omnibus): 0.337 Jarque-Bera (JB): 1.510
Skew: 0.467 Prob(JB): 0.470
Kurtosis: 3.105 Cond. No. 5.00e+03
==============================================================================
```

B.2 Model 2

```
                        OLS Regression Results
==============================================================================
Dep. Variable: cum_cases_100k R-squared: 0.797
Model: OLS Adj. R-squared: 0.774
Method: Least Squares F-statistic: 35.32
Date: Mon, 30 Aug 2021 Prob (F-statistic): 5.30e-12
Time: 12:55:11 Log-Likelihood: -350.25
No. Observations: 41 AIC: 710.5
Df Residuals: 36 BIC: 719.1
Df Model: 4
Covariance Type: nonrobust
=================================================================================
    ↪ ==========================================
                                                       coef std err t P>|t| [0.025
    ↪ 0.975]
----------------------------------------------------------------------------------
    ↪ ------------------------------------------
const 1269.3932 1321.701 0.960 0.343 -1411.142 3949.928
overcrowding 465.4333 86.027 5.410 0.000 290.962 639.905
Wholesale trade (Percent) 744.9200 313.434 2.377 0.023 109.247 1380.593
Transportation and warehousing, and utilities (Percent) 574.3395 129.262 4.443
    ↪ 0.000 312.183 836.496
Graduate or professional degree (Percent) -103.7377 40.892 -2.537 0.016 -186.670
    ↪ -20.806
==============================================================================
Omnibus: 2.985 Durbin-Watson: 2.335
Prob(Omnibus): 0.225 Jarque-Bera (JB): 2.088
Skew: 0.540 Prob(JB): 0.352
Kurtosis: 3.238 Cond. No. 98.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
    ↪ specified.
```

B.3 Model 3

```
                         OLS Regression Results
==============================================================================
Dep. Variable: cum_deaths_100k R-squared: 0.709
Model: OLS Adj. R-squared: 0.667
Method: Least Squares F-statistic: 17.05
Date: Thu, 02 Sep 2021 Prob (F-statistic): 1.56e-08
Time: 13:34:14 Log-Likelihood: -203.97
No. Observations: 41 AIC: 419.9
Df Residuals: 35 BIC: 430.2
Df Model: 5
Covariance Type: nonrobust
==================================================================================
    ↪ ==========================
                                        coef std err t P>|t| [0.025 0.975]
----------------------------------------------------------------------------------
    ↪ --------------------------
const -175.5297 141.092 -1.244 0.222 -461.963 110.903
overcrowding -7.5399 4.151 -1.816 0.078 -15.968 0.888
Age_0-19 -434.9085 281.417 -1.545 0.131 -1006.215 136.398
H_percent 283.4923 104.513 2.713 0.010 71.319 495.665
9th to 12th grade, no diploma (Percent) 3.9634 2.857 1.387 0.174 -1.836 9.763
Average family size 95.6078 45.833 2.086 0.044 2.562 188.653
==============================================================================
Omnibus: 1.829 Durbin-Watson: 2.242
Prob(Omnibus): 0.401 Jarque-Bera (JB): 1.021
Skew: 0.357 Prob(JB): 0.600
Kurtosis: 3.296 Cond. No. 550.
```

```
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
    ↪ specified.
```

B.4 Model 4

```
                            OLS Regression Results
==============================================================================
Dep. Variable: cum_deaths_100k R-squared: 0.562
Model: OLS Adj. R-squared: 0.513
Method: Least Squares F-statistic: 11.53
Date: Thu, 02 Sep 2021 Prob (F-statistic): 3.98e-06
Time: 13:35:39 Log-Likelihood: -212.37
No. Observations: 41 AIC: 434.7
Df Residuals: 36 BIC: 443.3
Df Model: 4
Covariance Type: nonrobust
=================================================================================
    ↪ ==========================================
                                                     coef std err t P>|t| [0.025
                                                        ↪ 0.975]
---------------------------------------------------------------------------------
    ↪ ------------------------------------------
const -32.7124 45.776 -0.715 0.479 -125.550 60.125
overcrowding 9.6641 2.979 3.244 0.003 3.622 15.707
Wholesale trade (Percent) 17.9259 10.855 1.651 0.107 -4.090 39.942
Transportation and warehousing, and utilities (Percent) 11.2413 4.477 2.511 0.017
    ↪  2.162 20.321
Graduate or professional degree (Percent) -1.4362 1.416 -1.014 0.317 -4.308 1.436
==============================================================================
Omnibus: 35.893 Durbin-Watson: 2.117
Prob(Omnibus): 0.000 Jarque-Bera (JB): 137.839
Skew: 1.968 Prob(JB): 1.17e-30
Kurtosis: 11.074 Cond. No. 98.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
    ↪ specified.
```

C Appendix – Co-correlation Lists

C.1 Model 1

C.1.1 Co-correlated to 5 features

- Ages 60-69 (% population)
- Ages 60-69 (% population)
- Ages 80+ (% population)
- Commuting - Worked from home (% workers)
- Health Insurance (% noninstitutionalized civilians)
- Less than 9th grade education (% population $\geq$ 25 years)
- No Health Insurance (% noninstitutionalized civilians)
- Private Health Insurance (% noninstitutionalized civilians)
- Overcrowding (% households)

C.1.2 4 Features

- Ages 50-59 (% population)
- Ages 70-79 (% population)
- Agriculture, forestry, fishing and hunting, and mining (%)
- Bachelor’s degree (% $\geq$ 25 yrs)
- Finance and insurance, and real estate and rental and leasing (%)
- Unemployed (% population $\geq$ 16 years)

C.1.3 3 Features

- Lower Extremity Diabetes Amputations (incidence per 100,000)
- Employed, civilians (% $\geq$ 16 yrs)
- Graduate or professional degree (% $\geq$ 25 yrs)
- High school graduate (or equivalent) (% $\geq$ 25 yrs)
- Information (%)
- Median Income (USD)
- Professional, scientific, and management, and administrative and waste management services (%)
- Total Civilian Noninstitutionalized Pop with disability (%)
- Transportation and warehousing, and utilities (%)

C.1.4 2 Features

- Community-Acquired Pneumonia (incidence per 100,000)
- Diabetes - Short-term Complications (incidence per 100,000)
- Urinary Tract Infection (incidence per 100,000)
- Commuting - Car, truck, van, drove alone (% workers)
- Public Health Insurance (% noninstitutionalized civilians)
- Wholesale trade (%)

C.2 Models 2 and 4

C.2.1 Co-correlated to 3 features

- Diabetes Long-term Complications (incidence per 100,000)
- 9th to 12th grade, no diploma (% population $\geq$ 25 years)
- Ages 0-19 (%)

C.2.2 2 features

- Commuting - Worked from home (% workers)
- Health Insurance (% noninstitutionalized civilians)
- No Health Insurance (% noninstitutionalized civilians)
- Agriculture, forestry, fishing and hunting, and mining (%)
- High school graduate (or equivalent) (% population $\geq$ 25 years)
- Total Civilian Noninstitutionalized Pop with disability (%)
- Average family size
- Hispanic (%)

C.3 Model 3

C.3.1 Co-correlated to 5 features

- Ages 60-69 (% population)
- Ages 80+ (% population)
- Commuting - Worked from home (% workers)
- Health Insurance (% noninstitutionalized civilians)
- Less than 9th grade education (% population $\geq$ 25 years)
- No Health Insurance (% noninstitutionalized civilians)
- Private Health Insurance (% noninstitutionalized civilians)
- Overcrowding (% households)
- Long-Term Diabetes Complications (incidence per 100,000)

C.3.2 4 Features

- Private Health Insurance (% noninstitutionalized civilians)
- Ages 50-59 (% population)
- Ages 70-79 (% population)
- Agriculture, forestry, fishing and hunting, and mining (%)

C.3.3 3 Features

- Bachelor's degree (% population $\geq$ 25 years)
- Finance and insurance, and real estate and rental and leasing (%)
- Unemployed, civilians (% population $\geq$ 16 years)

C.3.4 2 Features

- Lower Extremity Diabetes Amputations (incidence per 100,000)
- Employed, civilians (% $\geq$ 16 yrs)
- Graduate or professional degree (% $\geq$ 25 yrs)
- High school graduate (or equivalent) (% $\geq$ 25 yrs)
- Information (%)
- Median Income (USD)
- Professional, scientific, and management, and administrative and waste management services (%)
- Total Civilian Noninstitutionalized Pop with disability (%)
- Transportation and warehousing, and utilities (%)
- Wholesale trade (%)
- Severe Overcrowding (% households)