In [ ]:
```
# libraries
import numpy as np
import pandas as pd
import altair as alt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import r2_score
from sklearn.preprocessing import add_dummy_feature
alt.data_transformers.disable_max_rows()
tidydata = pd.read_csv('data.csv').drop(columns = 'Unnamed: 0').iloc[186:, :].reset_index()
```

Vaccine Efficacy in California

Spencer Zeng and Shivani Kharva

Author contributions

Spencer contributed: "Visualizing deaths of each vaccination status in different months", "Exploring correlations within dataset", "Scattering boosted deaths against boosted cases by months and fitting a linear model", "Principal component analysis"

Shivani contributed: "Abstract", "Aims", "Methods", "Initial exploratory analysis", "Exploring the general relationship between death rate and cases (before and after Omicron spike)", overall formatting

Both contributed: "Background", "Datasets", "Discussion"

Abstract

Covid-19 has rampaged throughout the entire world ever since its spread began in 2019. However, there have been scientific advances made to combat the virus with vaccinations and boosters. This project was conducted to explore the effectiveness of vaccination status on cases and deaths in California. There is an emphasis on the effects of the booster vaccination on deaths since that is the most recent advancement. Overall, it was found that, as individuals have more vaccinations, the death rate appears to be lower. Moreover, although total deaths appear to have a negative relationship with the population of California that has been boosted, it is unlikely that a linear model would describe the relationship between the two variables the most effectively.

Introduction

Background

COVID-19 (coronavirus disease 2019) is a disease caused by a virus named SARS-CoV-2. It mainly attacks the respiratory system and spreads out rapidly. Since December 2020, COVID-19 vaccines have shown to be effective at keeping people from severe illness. During August 2021, vaccine boosters were introduced to further enhance protection. Several studies have been dedicated to determine vaccine efficacy; specifically, the California Department of Public Health provides updates on post-vaccination infection data in California weekly.

The data in this project describes the vaccination status of individuals who have contracted, died, or been hospitalized due to COVID-19 in California. The data was collected in order to track the spread of COVID-19, along with its variants, and how it affects different individuals, depending on their vaccination status. No vaccine can be 100% effective and it is important to know whether the different vaccine treatments (initial vaccine and booster) have made a significant difference compared to individuals being unvaccinated. From this data, we can learn how effective the vaccine has been over time in California and whether individuals should be pushed towards getting the vaccination. Also, since the data is over time, the findings can show us whether the different vaccinations remain effective, especially against the variants of Covid-19 over time.

Aims

This project aims to compare the effect of vaccination status on cases and deaths of California individuals over time. The average deaths rates over time by vaccination status imply that those who are unvaccinated have higher death rates than those who are vaccinated, who in turn have higher death rates than those who are boosted. Furthermore, the relationship between death rates and case rates, divided between time periods before and after Omicron, appears to be similar for all three vaccination status. Before Omicron, the relationship between death rates and case rates appears to be nonlinear at first, which then turns into a slightly nonlinear positive relationship; however, after Omicron, the two variables have a nonlinear, curved relationship, where the death rate increases at first and then decreases.

When focusing on the effect of the booster at death prevention after the Omicron spike, it was found that there was a nonlinear, curved, negative trend in the relationship between the total death rate and the population boosted. After fitting a linear model to the data, it was found that, though much of the total death rate can be explained by the explanatory variables, there is still some nonlinearity not explained by the fitted model. Principal component analysis (PCA) was then conducted in order to identify whether specific variables explained most of the variation in the death rate (testing for collinearity).

Materials and methods

Datasets

The data are counts of cases, hospitalizations, and deaths from COVID-19 for unvaccinated, vaccinated, and boosted individuals in California. The data is from the California Open Data Portal and was collected by the California Department of Public Health (CDPH). The CDPH tracks the spread and effect of COVID-19 depending on vaccination status among California residents in order to monitor the impact of immunization campaigns.

The data is publicly available:

> Citation: California Department of Public Health. (2022, May 16). COVID-19 Post-Vaccination Infection Data. Retrieved May 16, 2022, from California Open Data Portal.

> https://data.ca.gov/dataset/covid-19-post-vaccination-infection-data (https://data.ca.gov/dataset/covid-19-post-vaccination-infection-data)

The data is collected by the CDPH from California's state immunization registry and registry of confirmed COVID-19 cases. The counts are updated by the CDPH after receiving reports from California hospitals (in this case, we are using the term 'hospitals' to include clinics, school clinics, etc.).

As for sampling, the population is all California residents. Since the data is collected by the CDPH to gather more information about COVID-19 in California, this is administrative data. The sampling frame is the California residents who get tested/vaccinated/hospitalized/etc. at hospitals that report to the California registry (does not include individuals who do not report having COVID-19 hospitals, asymptomatic/untested individuals who are positive for COVID-19, etc.). Our sample is equal to our sampling frame in this case.

Since our sampling frame partly overlaps the population and our sampling mechanism is a census of the frame, we have no scope of inference. This may be a limitation to the particular topic we are investigating because we cannot make general conclusions about how vaccination status affects the entire population of California. Since the goal of the data is to provide more information about whether the vaccine is effective, having no scope of inference may hinder whether all California residents believe the findings in this data are relevant to them. However, these findings are still important because they can be generalized about individuals who do report their health status to some form of California hospitals (still a considerable proportion of the state).

The dataset contains 441 observations in total. The observation units are the dates of sample collection, and each observation is a daily record of vaccination status of COVID-19 cases, hospitalizations, and deaths in California from 02/01/2021 to 04/24/2022.

Below is the table of variable descriptions:

Name	Variable Description	Type	Units of Measurement
———	—	—	———
date	reporting time period	day	none
population_unvaccinated	number of persons age 12+ that have not received any does of COVID-19 vaccine	numeric	persons
population_vaccinated	number of persons age 12+ with a complete primary COVID-19 vaccine series	numeric	persons
population_boosted	number of persons age 12+ with a complete primary COVID-19 vaccine series and additonal booster dose	numeric	persons
unvaccinated_cases_per_100k	rates of laboratory-confirmed COVID-19 cases among the unvaccinated per 100k persons	numeric	persons
vaccinated_cases_per_100k	rates of laboratory-confirmed COVID-19 cases among the vaccinated per 100k persons	numeric	persons
boosted_cases_per_100k	rates of laboratory-confirmed COVID-19 cases among the boosted per 100k persons	numeric	persons
unvaccinated_hosp_per_100k	rates of hospitalized laboratory-confirmed COVID cases among the unvaccinated per 100k persons	numeric	persons
vaccinated_hosp_per_100k	rates of hospitalized laboratory-confirmed COVID cases among the vaccinated per 100k persons	numeric	persons
boosted_hosp_per_100k	rates of hospitalized laboratory-confirmed COVID cases among the boosted per 100k persons	numeric	persons
unvaccinated_deaths_per_100k	rates of laboratory-confirmed COVID-19 deaths among the unvaccinated per 100k persons	numeric	persons
vaccinated_deaths_per_100k	rates of laboratory-confirmed COVID-19 deaths among the vaccinated per 100k persons	numeric	persons
boosted_deaths_per_100k	rates of laboratory-confirmed COVID-19 deaths among the boosted per 100k persons	numeric	persons

Below is the first five rows of the tidied dataset:

In [ ]:
```
tidydata.head()
```

Out[2]:

	index	date	population_unvaccinated	population_vaccinated	population_boosted	unvaccinated_cases_per_100k	vaccinated_cases_per_100k	boost
0	186	2021-08-13	9767659	22033557	1099	85.180521	14.354728	
1	187	2021-08-14	9729344	22056444	1947	85.169448	14.291904	
2	188	2021-08-15	9708761	22066250	2613	84.940955	14.079031	
3	189	2021-08-16	9654877	22094991	3470	84.931170	13.959854	
4	190	2021-08-17	9595515	22128729	4395	84.718151	13.701000	

Methods

Exploratory analysis aimed at discovering whether specific vaccination status' appear to be more effective than others. This stage of the analysis identified that unvaccinated individuals have higher death rates compared to those who are vaccinated, who have higher death rates than boosted individuals. Furthermore, the relationship between deaths (per 100k) and cases (per 100k) by vaccination status was plotted and separated by time

periods before and after Omicron to observe and compare the pattern of death depending on vaccination status. Similar patterns were observed, and the aforementioned comparisons of death rates was further confirmed. In order to analyze the specific effect of the booster on the total death rate, an initial exploratory scatterplot of total deaths against population boosted was analyzed by months. A multiple linear model was fitted to this plot with month (qualitative with four levels) and population boosted (quanititative) as regressors on the log transformation of total deaths. The model generally fit the data well and revealed that each month appears to follow a general trend; however, the linear model did not fully encompass the nonlinear parts of the data and seemed to contain too many regressors (based on the model summary). Principal component analysis was then conducted to discover any collinearity to test whether we had any unnecessary variables.

Results

Initial exploratory analysis

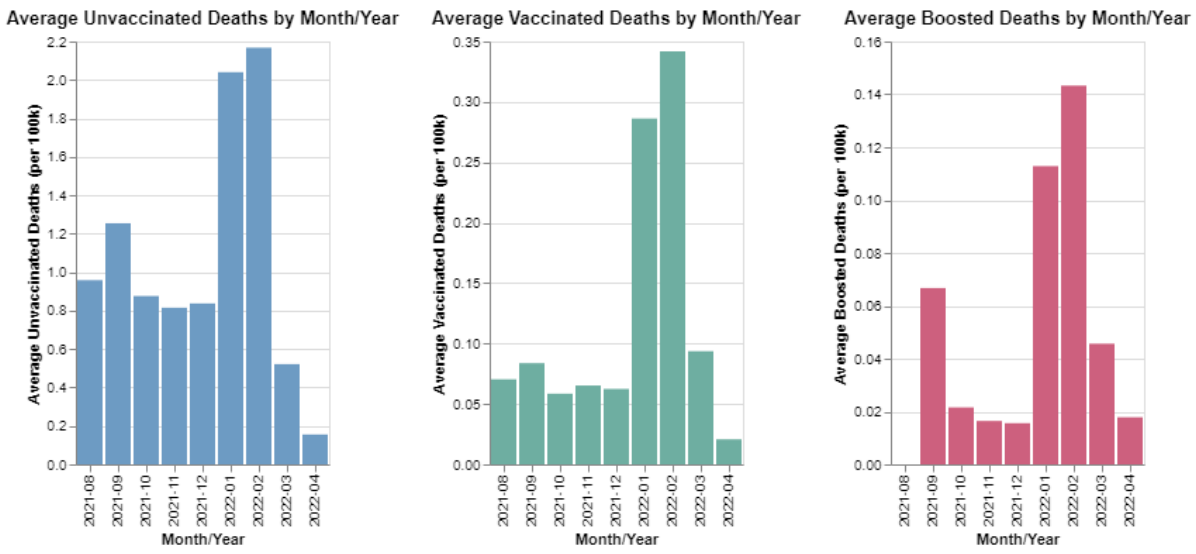

Figure 1: Bar graphs depicting average deaths (per 100k) per month by vaccination status (unvaccination, vaccinated, boosted)

As we can see in each of the graphs, there appears to be a spike in average deaths (per 100k) for all vaccination status in the month of January (which also appears). In particular, by viewing the raw data, it appears to shift on January 09, 2022. From this observation, we have decided to analyze the data in two subgroups: before and after January 09, 2022.

Exploring the general relationship between death rate and cases (before and after Omicron spike)

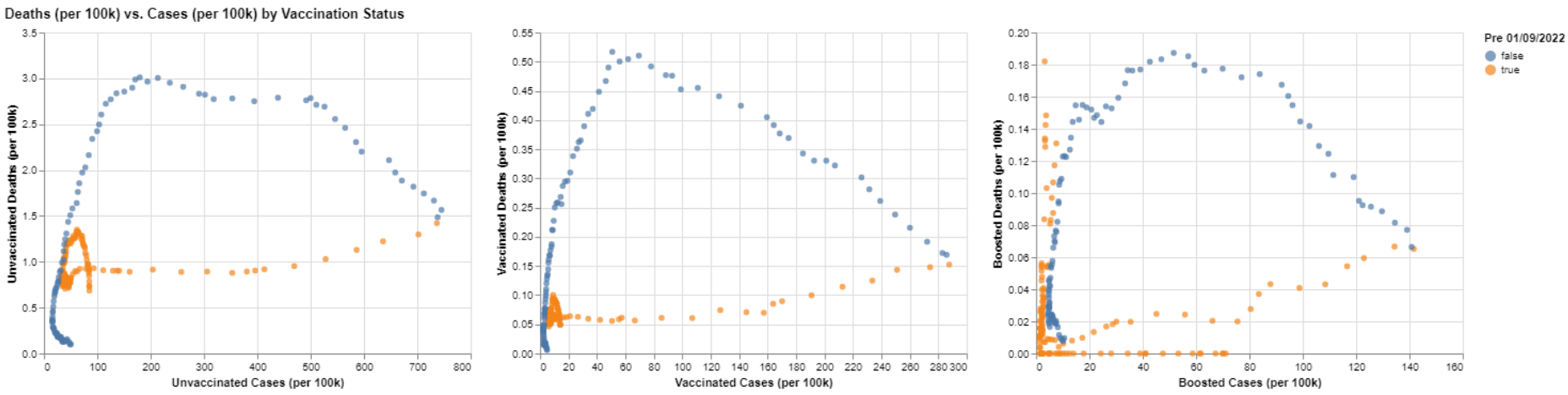

Figure 2: Scatterplots displaying the relationship between deaths (per 100k) and cases (per 100k) by vaccination status (unvaccination, vaccinated, boosted), divided by the periods before (yellow) and after (blue) the Omicron spike on January 09, 2022

Each of the graphs show that there were generally fewer deaths before January 2022 than after. Another trend that is apparent in each of the graphs is that, before January 2022, there was a positive pattern that was relatively close to linear between deaths and cases for each vaccination status; however, after January 2022, it is shown that there is a curved pattern, with deaths increasing and then decreasing with the increase of cases. Finally, by paying attention to the y-axis, we can see that, though the graphs tend to follow similar patterns, relatively, the unvaccinated cases are associated with higher deaths (peak around 3 deaths per 100k), vaccinated cases have the second highest deaths (peak around 0.52 deaths per 100k), and boosted cases have the fewest deaths (peak around 0.19 deaths per 100k).

Visualizing deaths of each vaccination status in different months

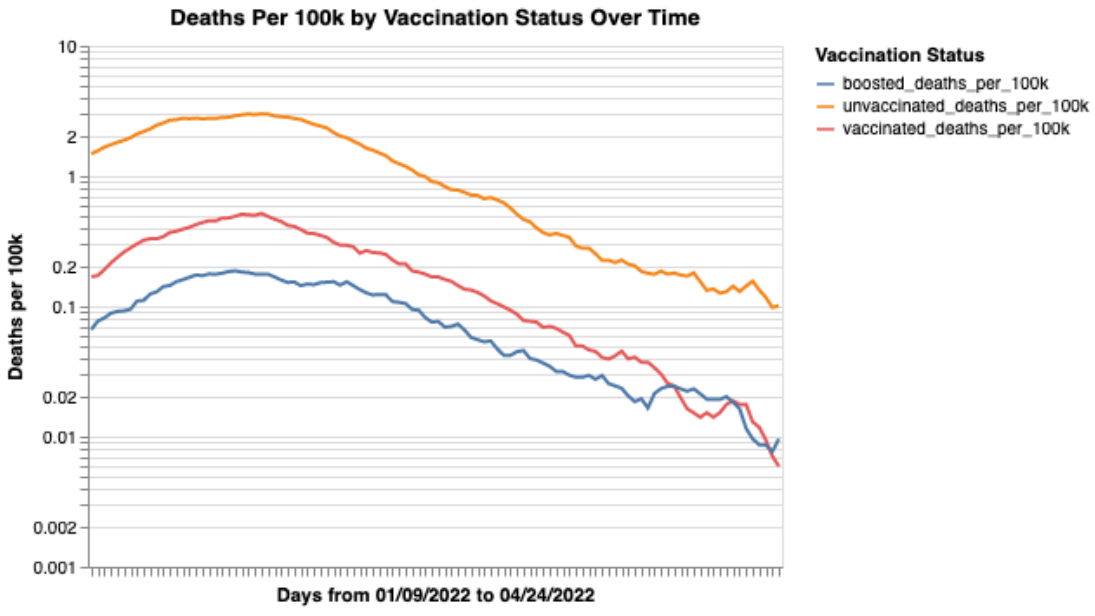

Figure 3: Point-and-line plots of positive cases/deaths/hospitalizations per 100k against months on each vaccination status (unvaccinated,vaccinated,and boosted)

In Figure 3, Death rates and hospitalization rates are decreasing over time, while lines in positive cases per 100k show a slight increase from March to April at each vaccination status. Boosted cases in general are observed to have the lowest positive/deaths/hospitalizations rates, although in April, there's a crossing over between lines of vaccinated hospitalizations and lines of boosted hospitalizations

Exploring correlations within dataset

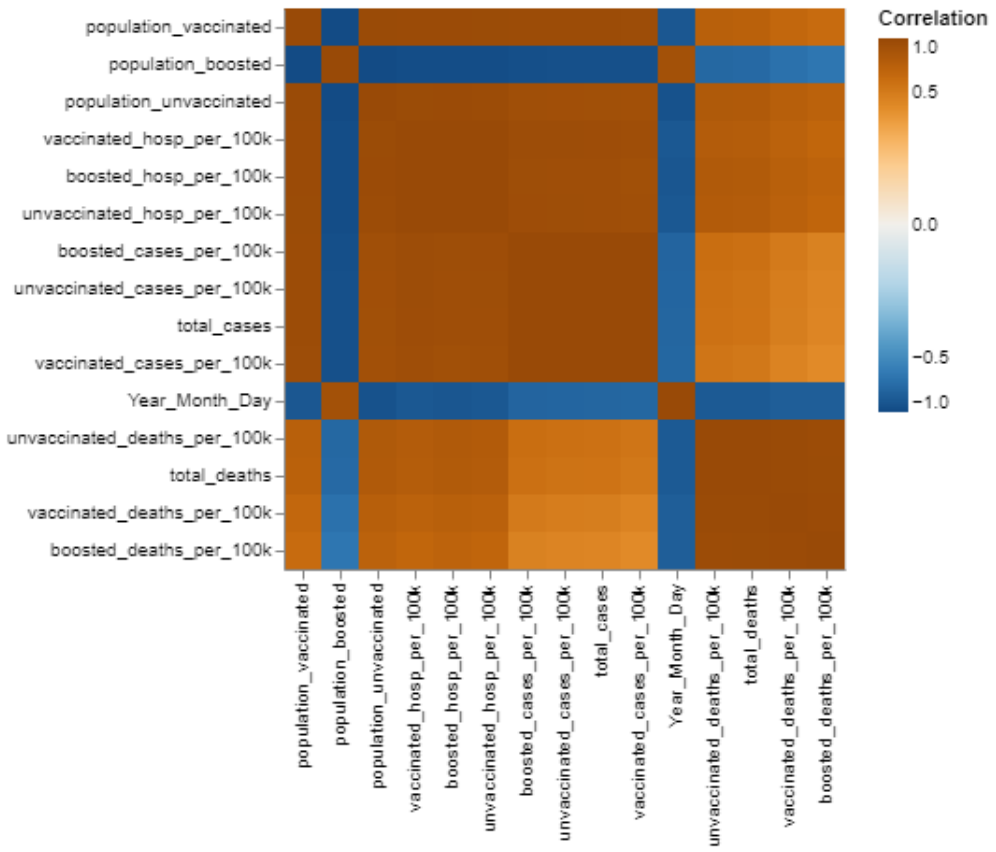

Figure 4: Heat map for correlatoins of variables in the data

When looking at boosted deaths, boosted cases, and date, there is a moderate positive correlation between boosted deaths and boosted cases, a strong negative correlation between date and boosted deaths, and a moderate negative correlation between boosted cases and date. Between population boosted and date, the association is strong and positive

Scattering boosted deaths against boosted cases by months and fitting a linear model

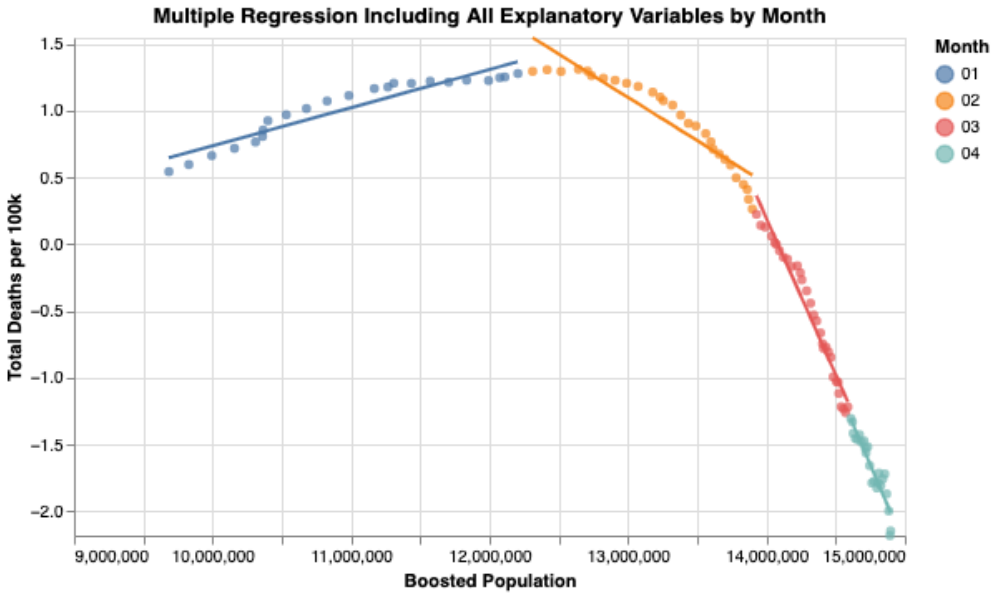

Figure 5: Scatterplot of log transformation of total deaths per 100k against the number of boosted persons by month with fitted multiple linear regression lines

Variable	Exponentiate coefficient estimates	Standard Error
—	—	———
boosted_cases	1.000000026820428	2.6820427842103426e-8
months_Feburary	1.821386	0.5995977306259569
months_March	3.741629	1.3195210630687657
monhts_April	27.652413	3.3197129783978796
Feb x boosted	1.0	4.757395046481492e-8
Mar x boosted	1.0	9.380510899654589e-8
April x boosted	1.0	2.2567797909232606e-7

Name	Value
—	—
Residual sum of squares	0.992917791097417
Estimated error variance	9.74879055e-03

Table 1: Summary of Figure 5 plot with variable names and their coefficients, R^2, and estimated variance

In Figure 5, different trends are shown at each month. In January, there is a positive correlation; negative associations are observed in Feburary, March, and April. The trend is relatively steady each month, and the R^2 value is quite high (0.993), meaning that about 99.3% of the variation in total deaths per 100k is explained by the predictors. However, by looking at the fit, we can see that some of the nonlinearity is not captured in the fitted lines.

Principal components analysis

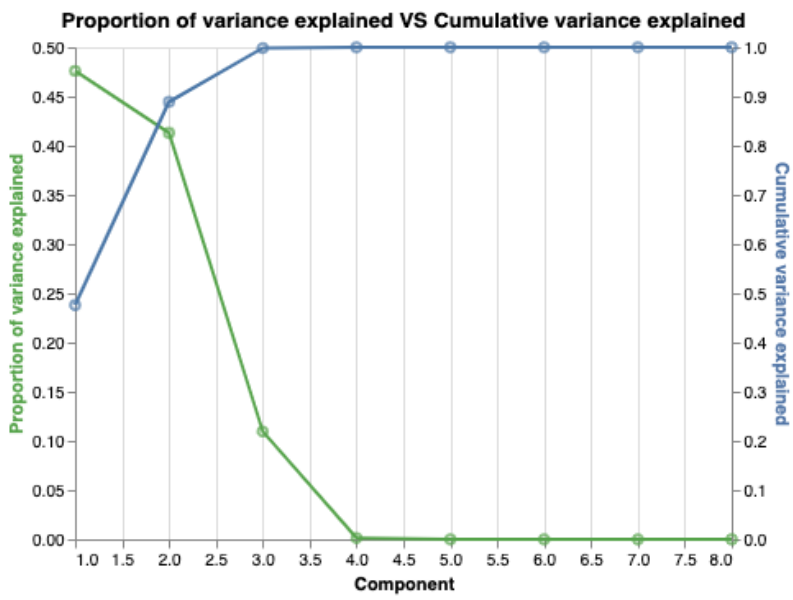

Figure 6: dual-axis plot showing the proportion of variance explained (y) as a function of component (x) in green on the left side and the cumulative variance explained (y) also as a function of component (x) in blue on the right side.

In Figure 6, we see that thwo components out of seven explain more than 80 percent of variations and covariations. Components after the third one do not contribute to the total variance explained. Thus, the principal components selected are PC1 and PC2

Discussion

In Figure 1, we see that unvaccinated individuals have higher deaths than vaccinated individuals, who have higher deaths than boosted individuals, and this observation is similar in terms of hospitalizations and cases. Also, a peak was observed in January 2022. With further investigation, it was found that this peak in deaths and overall illness was due to the increasing spread of the Omicron variant of Covid-19. Therefore, it is sensical that the data was split into time periods before and after the peak of Omicron. Furthermore, in Figure 2, we can see that the points for deaths after January 2022 are relatively higher than the points for deaths before 2022. This further confirms the differential effect of Omicron. An interesting observation from both time periods in all three graphs is that there is nonlinearity. Through further analysis of the raw data, it becomes apparent that a few weeks after spikes in Covid cases, the number of cases tends to decrease; however, though the case counts decrease, the number of deaths tends to be higher in the following weeks because individuals from the spike in cases may die in later weeks.

The aforementioned analyses led us to question how well month and population boosted together could estimate the total number of deaths per 100k in California. Based on our multiple linear regression model of months and population boosted against deaths per 100k, the model generally fits the data well. However, some non-linear trends were not captured by the fitted lines, and, based on our model summary, it seemed to contain too many regressors. From our PCA, we learned that three out of seven components explain all variation and correlation in the original data, indicating that the set of full regressors is collinear. This means that there were variables in our analysis possibly explaining the same variation in total deaths, as we suspected from the model summary. In terms of further analysis, we might question whether we should fit variables in the

dataset other than the ones we chose (since collinearity exists within our model) and what type of model might fit the data better. Also, it may be useful to explore how the different variants of Covid-19 have affected California and whether the changes in regulations over time along with medical breakthroughs have lessened the effect of the variants.