COVID-19 ANALYSIS

TEAM MEMBERS:

ARADHYA AGRAWAL
aagrawal4497@sdsu.edu
MS in Big Data Analytics

RYAN LAFLER
rlafler@sdsu.edu
MS in Big Data Analytics

BHAGYASHRI PATIL
bpatil0085@sdsu.edu
MS in Big Data Analytics

AISHWARYA BHANGARSHETTRA
abhangarshettr6376@sdsu.edu
MS in Big Data Analytics

Website Link: https://sites.google.com/sdsu.edu/covid19/main

Video Link: https://youtu.be/HYXaKkpt72M

Table of Content:

1. INTRODUCTION

California is a diverse and expansive state containing 58 counties that are divided demographically, economically, and politically into three main regions: “NorCal”, “SoCal”, and central California. Counties line the Golden state’s coast, inland valleys, mountainous regions, and desert communities, each exhibiting characteristics similar and dissimilar from one another.

There is a fundamental distinction between a county and a city. Counties lack broad powers of self-government that California cities have (e.g., cities have broad revenue generating authority and counties do not). In addition, legislative control over counties is more complete than it is over cities. Unless restricted by a specific provision of the state Constitution, the Legislature may delegate to the counties any of the functions which belong to the state itself. Conversely, the state may take back to itself and resume the functions which it has delegated to counties (e.g., state funding of trial courts).

The California Constitution recognizes two types of counties: general law counties and charter counties. General Law counties adhere to state law as to the number and duties of county elected officials. Charter counties, on the other hand, have a limited degree of “home rule” authority that may provide for the election, compensation, terms, removal, and salary of the governing board; for the election or appointment (except the sheriff, district attorney, and assessor who must be elected), compensation, terms, and removal of all county officers; for the powers and duties of all officers; and for consolidation and segregation of county offices. A charter does not give county officials extra authority over local regulations, revenue-raising abilities, budgetary decisions, or intergovernmental relations.

A county may adopt, amend, or repeal a charter with majority vote approval. A new charter or the amendment or repeal of an existing charter may be proposed by the Board of Supervisors, a charter commission, or an initiative petition. The provisions of a charter are the law of the state and have the force and effect of legislative enactments. There are currently 44 general law counties and 14 charter counties. They are as follows:

General Law Counties: Alpine, Amador, Calaveras, Colusa, Contra Costa, Del Norte, Glenn, Humboldt, Imperial, Inyo, Kern, Kings, Lake, Lassen, Madera, Marin, Mariposa, Mendocino, Merced, Modoc, Mono, Monterey, Napa, Nevada, Plumas, Riverside, San Benito, San Joaquin, San Luis Obispo, Santa Barbara, Santa Cruz, Shasta, Sierra, Siskiyou, Solano, Sonoma, Stanislaus, Sutter, Trinity, Tulare, Tuolumne, Ventura, Yolo, Yuba

Charter Counties: Alameda, Butte, El Dorado, Fresno, Los Angeles, Orange, Placer, Sacramento, San Bernardino, San Diego, San Francisco, San Mateo, Santa Clara, Tehama.[7]

1.1. SCOPE

With the pandemic persisting into 2021, our topic of interest focused on the percentage of each county's population that was fully vaccinated from any of the FDA-approved vaccines. More specifically, our analysis sought to characterize which variables were of greatest significance explaining the variation ("spread") between vaccination rates in different counties.

1.2. ABOUT DATASET

Four datasets were used in our analysis: (1) estimates of county demographics collected by the U.S. Census Bureau, (2) daily COVID-19 infection and mortality frequencies from the New York Times, (3) cumulative county vaccination rates published by the U.S. Centers for Disease Control, and (4) the 2020 presidential election results by U.S. counties. By first cleaning these datasets, visualizing their variables, and examining whether spatial dependencies exist between neighboring counties, we constructed a final model to best explain the variation in counties' vaccination rates.[3]

2. EXPLORATORY ANALYSIS

The four datasets were imported into R, Python and subsequently filtered for counties located *only* within the state of California. This was completed by filtering for FIPS codes between 6000 and 7000. Following the filtering process, the datasets merging occurred using intersecting joins between tables. In the effort to reduce anomalies and redundancy, only the most up-to-date cumulative vaccine and COVID-case data (September 24th, 2021) were merged with the American Community Survey estimates and 2020 presidential election results. Merging was completed on the FIPS variable, serving as the primary key and foreign keys linking tables together.

2.1. DEMOGRAPHICS

County demographics estimates were extracted from the U.S. Census Bureau's five-year American Community Survey. Gathered from 2014 to 2019, the five-year survey included the most recent estimates of various socio-economic characteristics from all counties in the United States regardless of their population size/density. The R `tidycensus` package contains an API for downloading Census Bureau datasets directly to R, which can then be exported as a CSV file. Each county is assigned a unique FIPS code doubling as the primary key. Variables of interest included a county's median age, median household income, levels of educational attainment, ethnic and racial compositions. Avoiding the effect of larger counties (i.e., Los Angeles County) exerting overwhelming influence over smaller counties (i.e., Alpine County), all frequencies were converted to proportions, locking values between 0 and 1. Converting from frequencies to proportions also standardized our variables, greatly assisting in the model-building and interpretation process.[5]

The 2020 presidential election[7] results were obtained from **GitHub** where the original data aggregated from several sources, including (as the author describes), "The Guardian, townhall.com, Fox News, Politico, and the New York Times". Using R, the data were downloaded directly from GitHub and contained columns for the total numbers of Democratic and Republican

votes and their proportions of the vote. Counties are again identified by their FIPS codes, permitting a relational database to exist between the four tables [4].

Fig.1 shows the proportion of participation of the counties in 2020 elections.

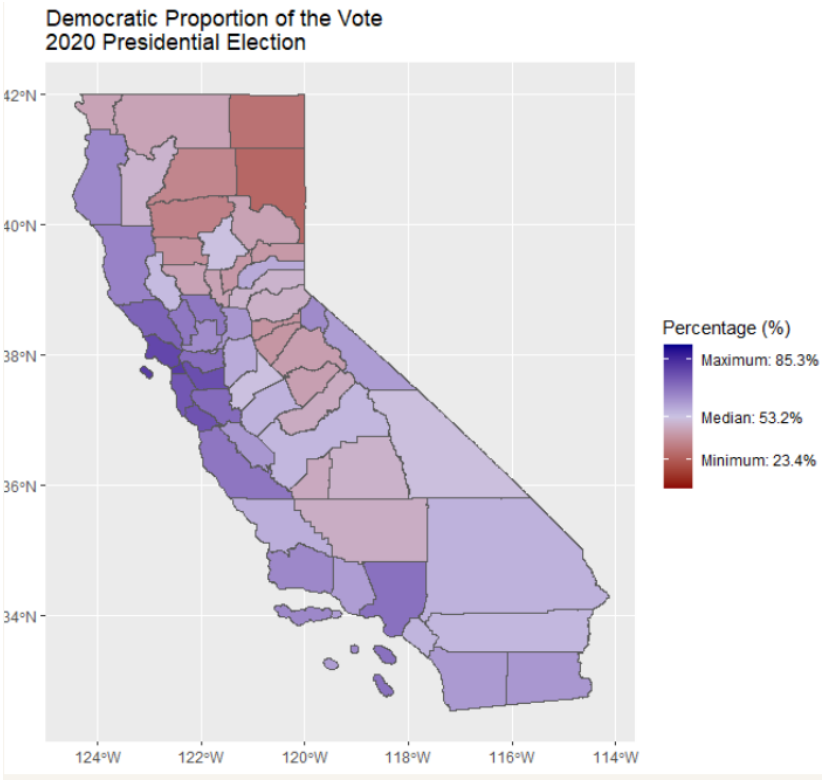
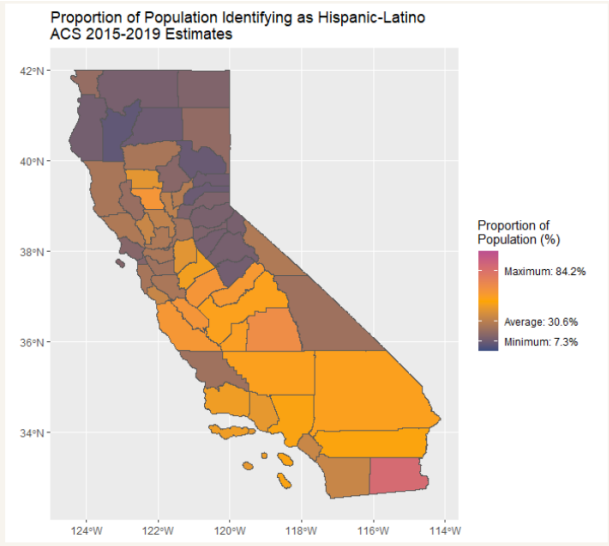


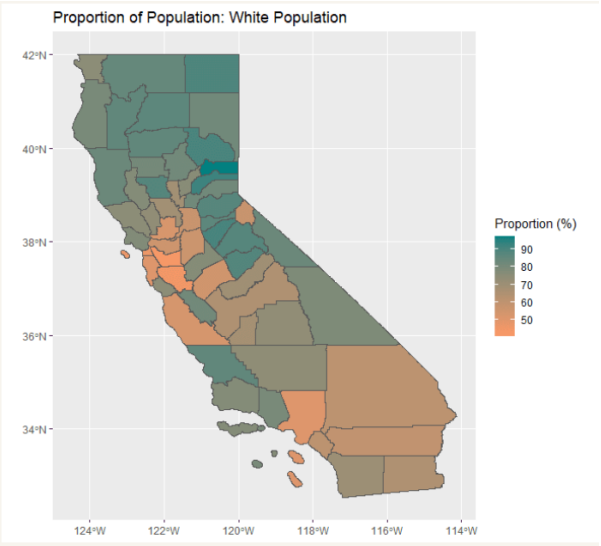
Fig 1

The below figures show each county's proportion of the population based on race and ethnicity—variables included the proportion identifying as Hispanic-Latino, white-alone, black/African American, and Asian. The following two figures examine counties' proportions of those holding a bachelor's degree or higher and median household incomes.



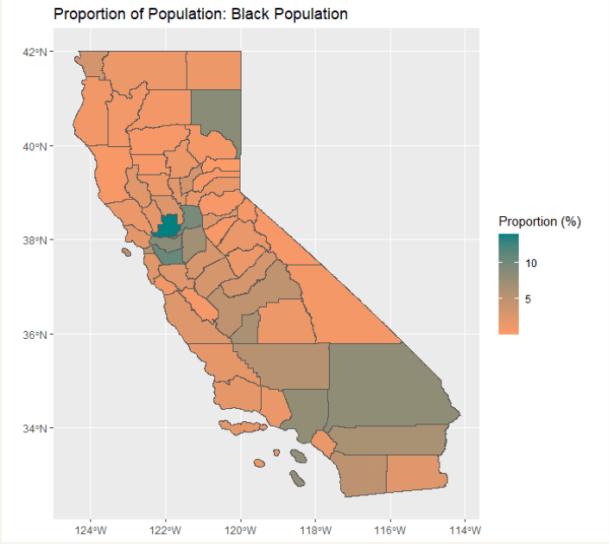
ACS 2015-2019 estimates of the proportion of the population self-identifying as Hispanic-Latino.

Fig 2

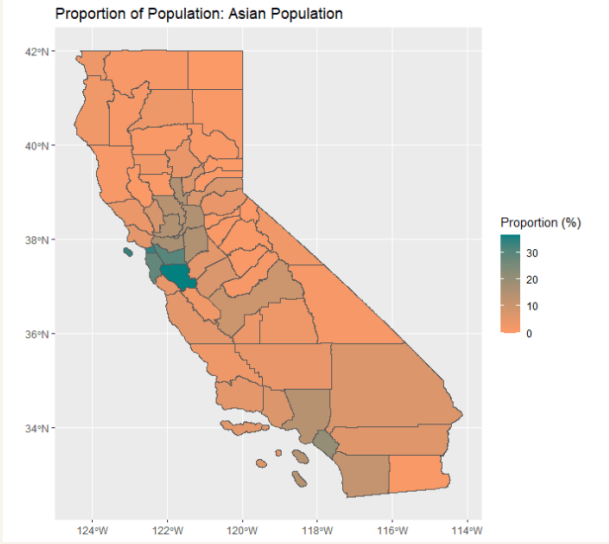


ACS 2015-2019 estimates of the proportion of the population self-identifying as white / Caucasian.

Fig 3

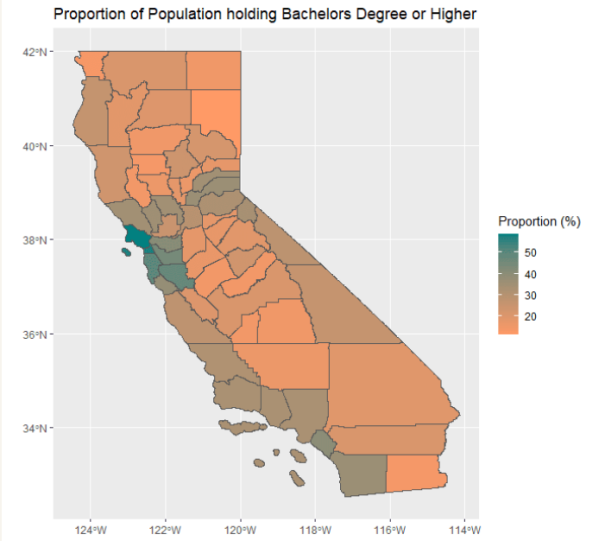


ACS 2015-2019 estimates of the proportion of the population self-identifying as black / African American.



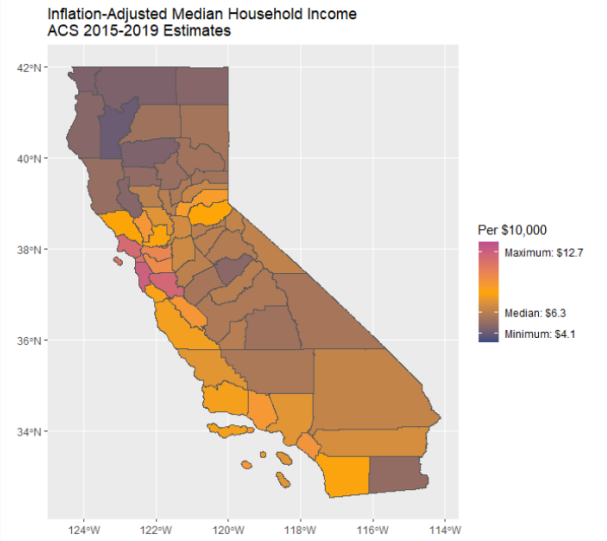
ACS 2015-2019 estimates of the proportion of the population self-identifying as Asian.

Fig 4



ACS 2015-2019 estimates of the proportion of the population holding bachelors degrees or higher.

Fig 5



ACS 2015-2019 estimates of median household income (adjusted for inflation).

Fig 6

Fig 7

2.2. COVID-19 SPREAD

With the first case being recorded on January 21st, 2020, the COVID-19 infection and mortality dataset contains the daily number of confirmed COVID-positive cases and fatalities for all counties in the United States up to the date of retrieval on September 24th, 2021. The dataset is a time series showing the cumulative frequencies (thereby including frequencies prior to and on that day) of confirmed cases and fatalities caused from COVID-19. Counties are uniquely identified by their Census Bureau FIPS codes, serving as the primary key for the dataset.

The vaccination dataset contains the cumulative frequencies and proportions of a county's population that are either partially or fully vaccinated. Updated daily, the dataset is a time series with the first observation recorded on December 13th, 2020, up to September 24th, 2021—the date of retrieval.

This line represents the total number of active cases and the total death recorded from Jan 25, 2020 to Sept 23, 2021. The graph is built using the extensive python libraries such as Numpy, pandas, folium, matplotlib, plotly. From the graph we can infer that there has been a significant raise in the count from the number of people who has been exposed to Covid 19 and the deaths which took people in that time frame. While there has been tremendous increase in cases as shown in Fig. 8, the death rate seems to be significantly low(Fig. 9) when compared with the active cases. However, according to the World Health Organization(WHO), though the death rate(787 K(latest)) is much less than the active cases(4.68 million and more)), COVID-19 has recorded the highest number of deaths in The United States after the 1918 influenza flu(675 K)[1].

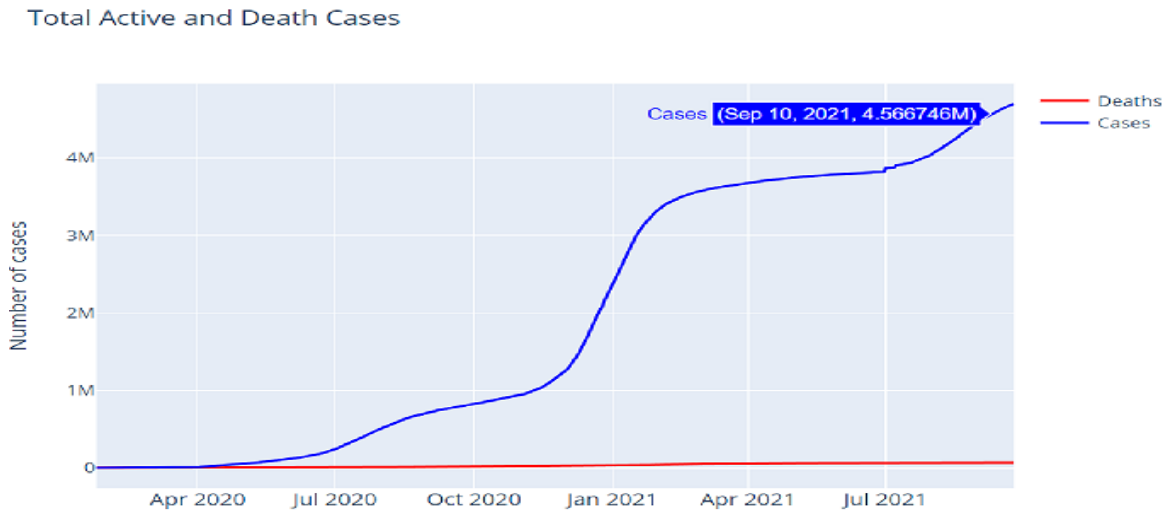


Fig 8

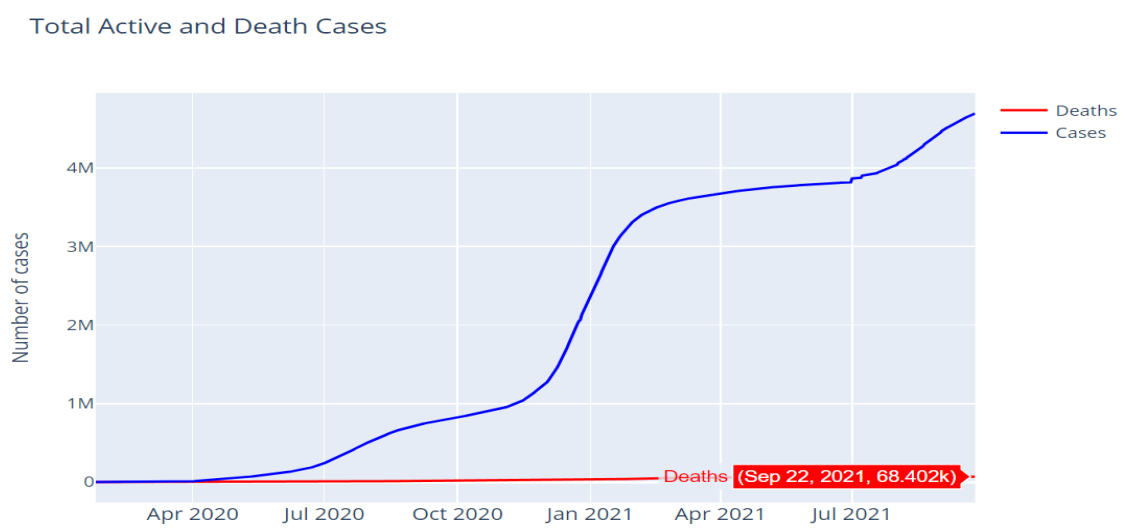


Fig 9

2.3. VACCINATION RATES

The below fig. 10 shows the vaccination rates in all the counties in California.
Counties that has the highest vaccination rate are green in color, while the counties that are color orange is least vaccinated. Various colors in the graph represent various Vaccination Status in that county.

The fig. 11 examines distribution for the proportion of fully vaccinated individuals by county over time. The diagram shows in majority of the counties, the vaccination rate is in between 30% to 60%.

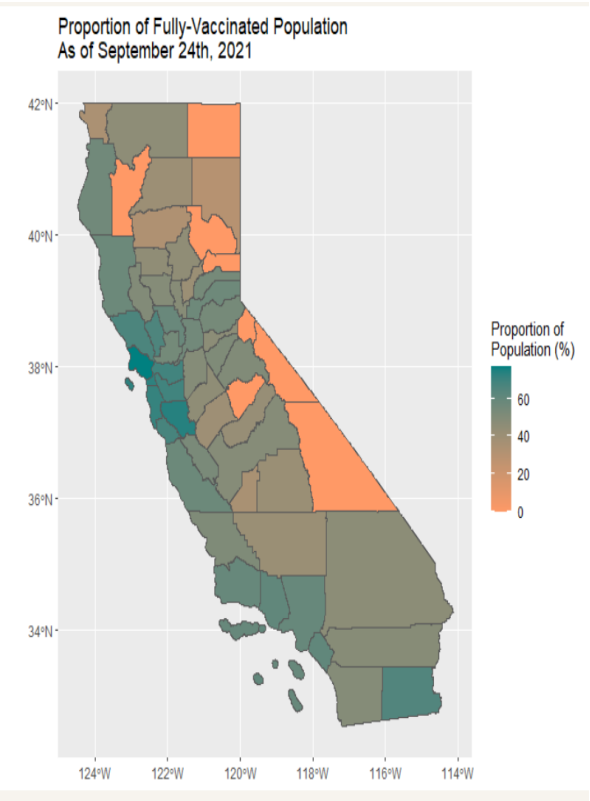


Fig 10

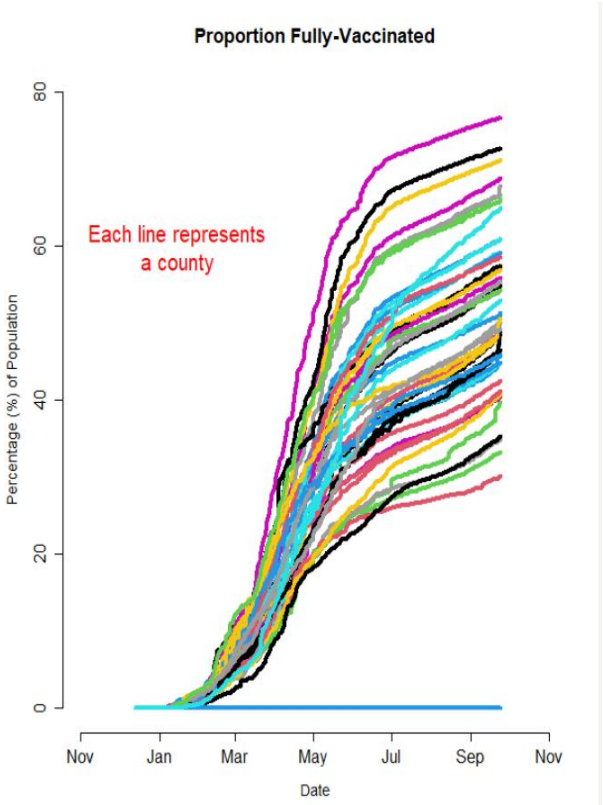


Fig 11

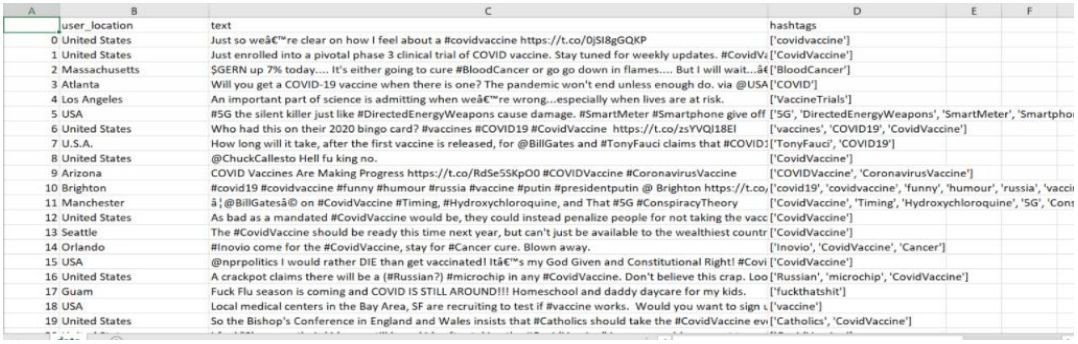
2.4. TWITTER

The Twitter dataset was downloaded from Kaggle. It contained various features like date, tweets, user, retweet, hashtags, etc. The problem with the dataset was that it contained user location from all over the world but we required data only related to the USA. So by using geonamecache library we were able to sort out the data for the USA. Initially contained 300,000 rows but after sorting out data contains only 14,000 rows.[6]

DATASET CLEANING, SOURCING AND PREPARATION

It contained various features like date, tweets, user, retweet, hashtags, etc. The problem with the dataset was that it contained user location from all over the world but we required data only related to the USA. So by using geonamecache library we were able to sort out the data for the USA. Initially contained 300,000 rows but after sorting out data contains only 14,000 rows.

The Fig. 13 is of the cleaned data now the data contains factors that are important for our tweets analysis. The user_location now contains places in the United States only. Working on this particular column was the most difficult task because there are the same places in the United Kingdom and Australia having the same names as those places in the United States. To remove those duplicates and consider only places in the United States, considering states was also important, So we added states to a list and compared that list with the after comma words which are basically abbreviations of states. And the other two columns contain tweets and hashtags.



	user_location	text	hashtags
0	United States	Just so weâ€™re clear on how I feel about a #covidvaccine https://t.co/0jSI8gGQKP	['covidvaccine']
1	United States	Just enrolled into a pivotal phase 3 clinical trial of COVID vaccine. Stay tuned for weekly updates. #CovidVa	['CovidVaccine']
2	Massachusetts	$GERN up 7% today.... It's either going to cure #BloodCancer or go go down in flames.... But I will wait...â€	['BloodCancer']
3	Atlanta	Will you get a COVID-19 vaccine when there is one? The pandemic won't end unless enough do. via @USA	['COVID']
4	Los Angeles	An important part of science is admitting when weâ€™re wrong...especially when lives are at risk.	['VaccineTrials']
5	USA	#5G the silent killer just like #DirectedEnergyWeapons cause damage. #SmartMeter #Smartphone give off	['5G', 'DirectedEnergyWeapons', 'SmartMeter', 'Smartpho
6	United States	Who had this on their 2020 bingo card? #vaccines #COVID19 #CovidVaccine https://t.co/zsYVQI18EI	['vaccines', 'COVID19', 'CovidVaccine']
7	U.S.A.	How long will it take, after the first vaccine is released, for @BillGates and #TonyFauci claims that #COVID1	['TonyFauci', 'COVID19']
8	United States	@ChuckCallesto Hell fu king no.	['CovidVaccine']
9	Arizona	COVID Vaccines Are Making Progress https://t.co/RdSeSSKpO0 #COVIDVaccine #CoronavirusVaccine	['COVIDVaccine', 'CoronavirusVaccine']
10	Brighton	#covid19 #covidvaccine #funny #humour #russia #vaccine #putin #presidentputin @ Brighton https://t.co/	['covid19', 'covidvaccine', 'funny', 'humour', 'russia', 'vacci
11	Manchester	â¦@BillGatesâ© on #CovidVaccine #Timing, #Hydroxychloroquine, and That #5G #ConspiracyTheory	['CovidVaccine', 'Timing', 'Hydroxychloroquine', '5G', 'Cons
12	United States	As bad as a mandated #CovidVaccine would be, they could instead penalize people for not taking the vacc	['CovidVaccine']
13	Seattle	The #CovidVaccine should be ready this time next year, but can't just be available to the wealthiest countr	['CovidVaccine']
14	Orlando	#Inovio come for the #CovidVaccine, stay for #Cancer cure. Blown away.	['Inovio', 'CovidVaccine', 'Cancer']
15	USA	@nprpolitics I would rather DIE than get vaccinated! Itâ€™s my God Given and Constitutional Right! #Covi	['CovidVaccine']
16	United States	A crackpot claims there will be a (#Russian?) #microchip in any #CovidVaccine. Don't believe this crap. Loo	['Russian', 'microchip', 'CovidVaccine']
17	Guam	Fuck Flu season is coming and COVID IS STILL AROUND!!! Homeschool and daddy daycare for my kids.	['fuckthatshit']
18	USA	Local medical centers in the Bay Area, SF are recruiting to test if #vaccine works. Would you want to sign	['vaccine']
19	United States	So the Bishop's Conference in England and Wales insists that #Catholics should take the #CovidVaccine evi	['Catholics', 'CovidVaccine']

Fig. 13

2.4.1. HASHTAG BAR CHARTS

This bar plot was made using the matplotlib library in python. For each tweet there were many hashtags used so every hashtag was appended to a list and using NLTK library frequency distribution was done. This frequency distribution was then converted to a dictionary. And using this dictionary the bar plot was formed.

Fig. 14 bar plot for hashtags used. As it was clear that ***#covidvaccine*** was the most used tag but it can be seen that #tokyoolympics, #parisolympics were also commonly used as olympics was held at during that period. The most common vaccine hashtag was for ***#pfizer*** than ***#moderna***. These were the top 50 hashtags used. It can be seen that people are aware of Olympics spreading covid-19 so ***2024 year*** and ***2028year*** also were the most common ones.

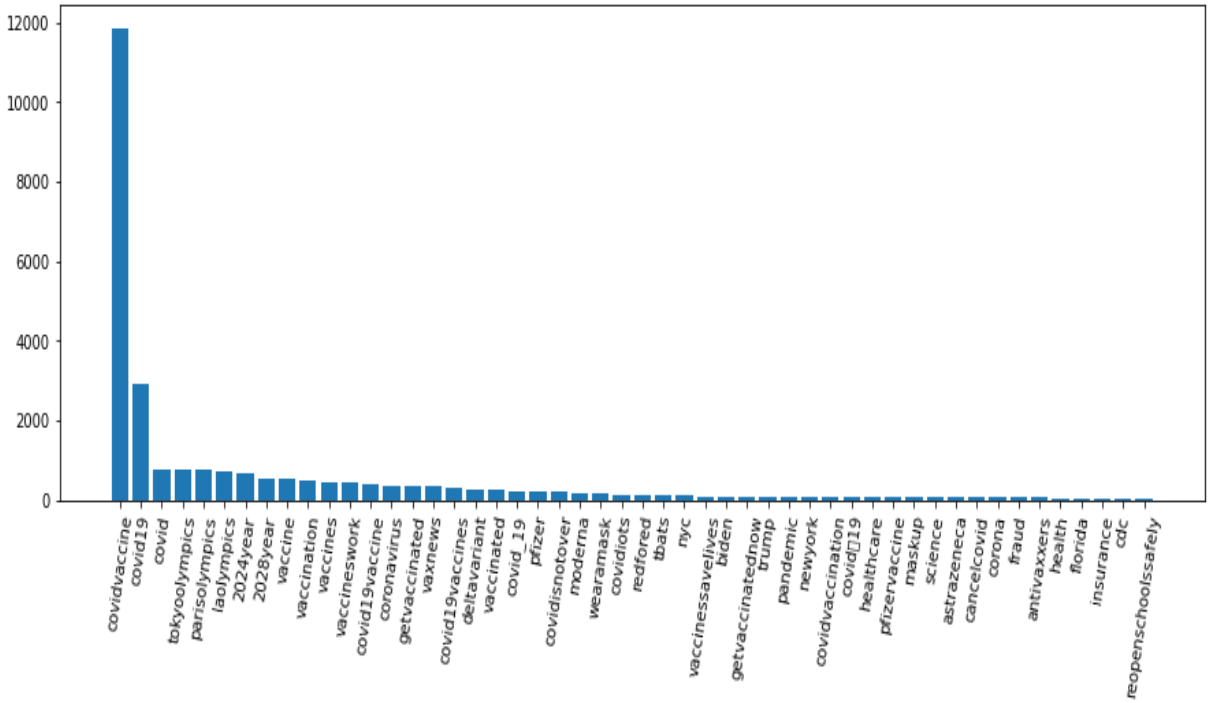


Fig 14

2.4.2. WORD CLOUD

Word Cloud was created using the Word Cloud library in the word cloud package. All the tweets were combined and then the word Cloud () function was used. Then matplotlib was used to plot the graph.

#REPLACE THIS WITH WORD CLOUD PARA 1

Before plotting the graph, the stop words are removed and stop words are the words like articles used in the English language. These removal of stop words are important because they add unnecessary space or words in the word cloud.

#REPLACE THIS WITH WORD CLOUD PARA 2

Fig. 15 is a word cloud of all the tweets of Covid Vaccines. Here it can be clearly seen that there are many links attached in the tweets as https and co are biggest. It could be because there were various news articles, blogs and other websites sharing updates about the covid. But the point to be noticed is people are opposing Olympics that is why it is written postpone tokyo olympics and please postpone. It can be seen that users want Olympics to be conducted in 2024year and continue Olympics at laolympics and words related to 2028year and parisolympics is also used often and is more common. Many tweets also include words related to education and trust on vaccines

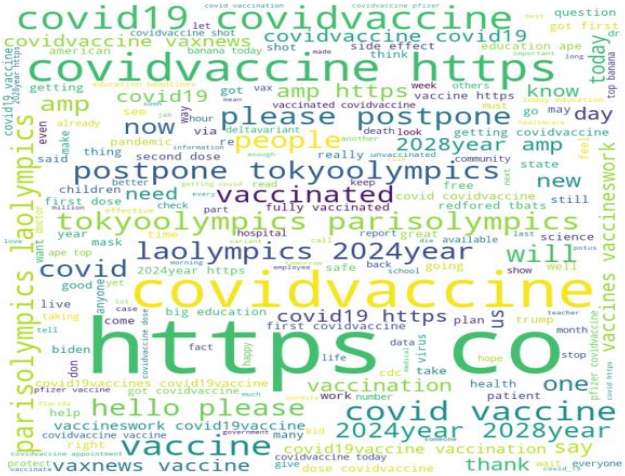


Fig 15

2.4.3. SENTIMENTAL ANALYSIS

Fig. 16 is sentiment analysis bar plot was built using the R programming language. Libraries like syuzhet, lubridate, ggplot2, scales, reshape2, and dplyr. And after this get_nrc_sentiment() function was used to perform sentiment analysis. And finally, the use of barplot() function was done to plot the bar chart.

We even performed sentiment analysis of the tweets. It can be seen that there is a mistake in the sentiment analysis because of the positive keyword. Positive is always used covid positive and here in sentiment analysis positive is referred to something good. Ignoring positive and negative we see people are using words related to trust for covid vaccine. And might be anticipating or predicting about the covid vaccine. Even words related to fear are more often used. It can be seen that people are less likely to disapprove or disgust things related to covid vaccine.

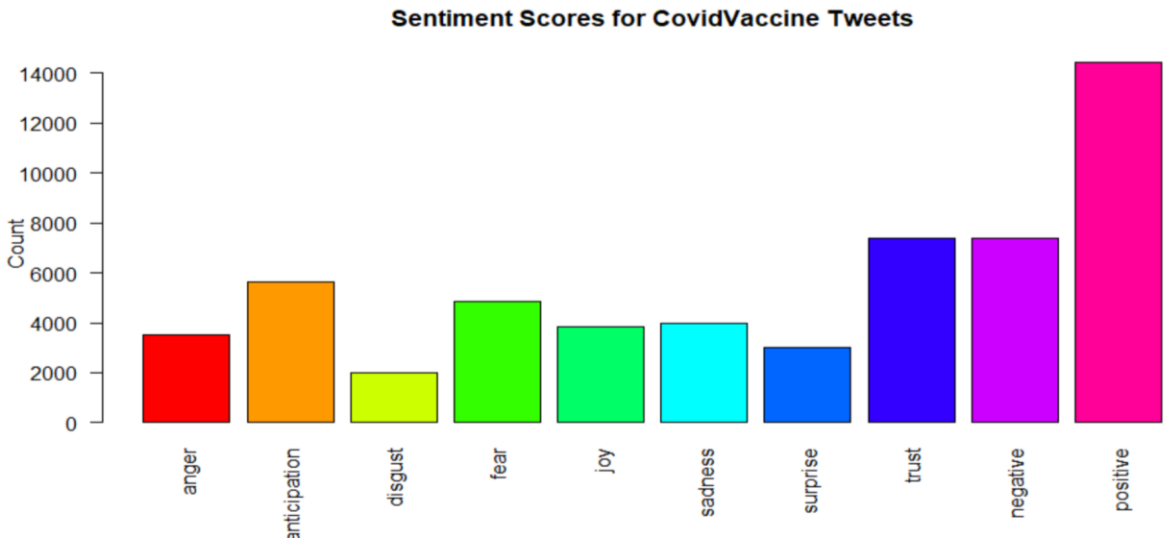


Fig 16

3. DASHBOARD

Fig. 17 is an interactive Dashboard represents the insights on the analysis of our Covid-19 Datasets. This Dashboard is built using Tableau data analysis tool. We have used the Dashboard Actions feature for the visuals to interact with each other using a single-click event[2].

Graph 1: Trends of Covid Cases vs. Deaths: This visual represents the Covid positive cases versus Death's trend graph for the range of Jan 2020 to September 2021.

Graph 2 : **Series Completion by County:** This visual represents the California State counties and uses the symbol map for representing the counties who have completed their covid-19 vaccination series. The map represents the color shading from red to green where red represents the lowest numbers and the green color represents the highest numbers, as shown in the legend.

Graph 3 : **Age Groups wise Series Completion by County:** This visual represents the Age group bins for each county in the California state, which have finished taking their covid-19 vaccination series. The data labels shows the actual count for the vaccination series taken by residents of those counties.

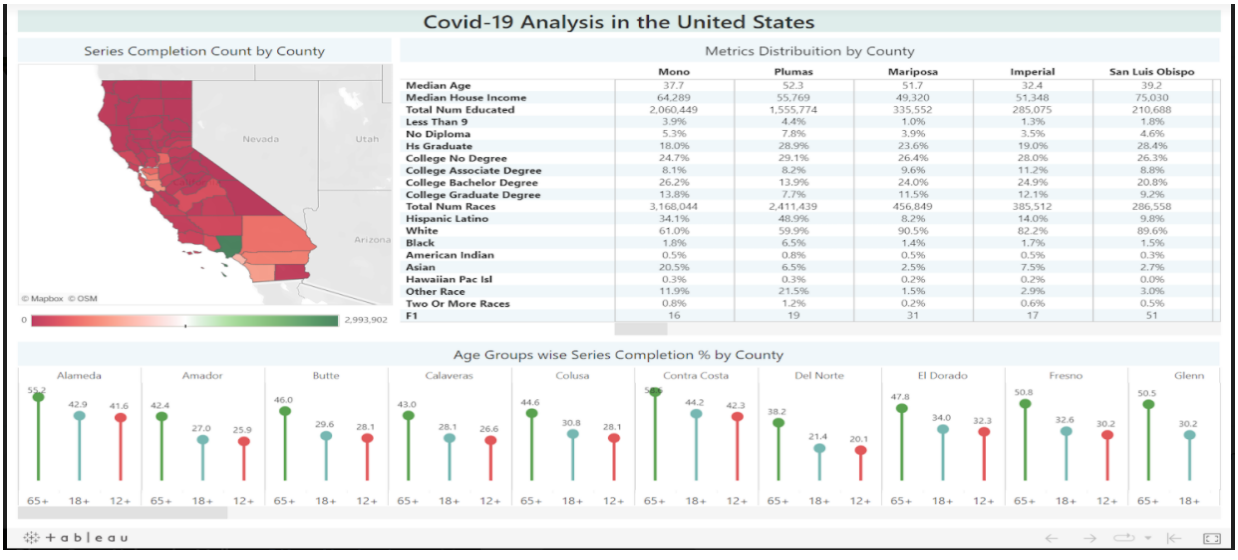


Fig 17

4. RELATIONSHIPS AND FINDINGS

Using the datasets, several variables emerged as potential candidates for explaining the variation in vaccination rates between California counties. Our goal was to include variables relevant to the question in the models—pulling them out from the unexplained error term to examine their effects on vaccination rates.

Variables included demographic and political insights including how Democratic a county voted in the 2020 Presidential Election, median household income, median age, the proportion of citizens holding bachelor's degrees or higher, and estimates for racial/ethnic proportions of the population. Also included were counties' fatality rates and confirmed positive COVID-19 case rates. Rather than regressing on a set of frequency variables, predictors were transformed to proportions, locking their range of values between 0 and 1 to prevent large counties like Los Angeles from exerting overwhelming influence on the models.

The first constructed model does not factor in spatial dependencies, Table 1, and as such, is represented by an ordinary least-squares multiple regression model. Using R, predictors' individual p-values, assuming the null hypothesis excludes them from the model (setting their partial slopes equal to zero), are given below:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0624	0.5170	-0.1207	0.9045
prop_cases	-1.9283	4.7972	-0.4020	0.6895
prop_deaths	1.8661	3.9210	0.4759	0.6363
white	0.5728	0.3843	1.4906	0.1427
asian	0.4027	0.6074	0.6631	0.5105
black	0.4112	1.1156	0.3686	0.7141
hispanic_latino	-0.1273	0.2797	-0.4552	0.6510
bachelors_or_higher	-0.6686	0.6721	-0.9948	0.3249
median_income	0.0000	0.0000	2.3227	0.0246
medianAge	-0.0134	0.0060	-2.2227	0.0311
per_dem	0.6842	0.3214	2.1288	0.0385

Table 1

In predicting a county's proportion of fully vaccinated individuals, several variables were found to be statistically significant. Median household income, median age, and the percentage a county voted for Biden in the 2020 election proved significant at the 5% level, and the proportion of a county's white population was significant at the 15% level, holding all other factors constant.

The residuals plot from the OLS multiple regression model is shown in the fig. 18:

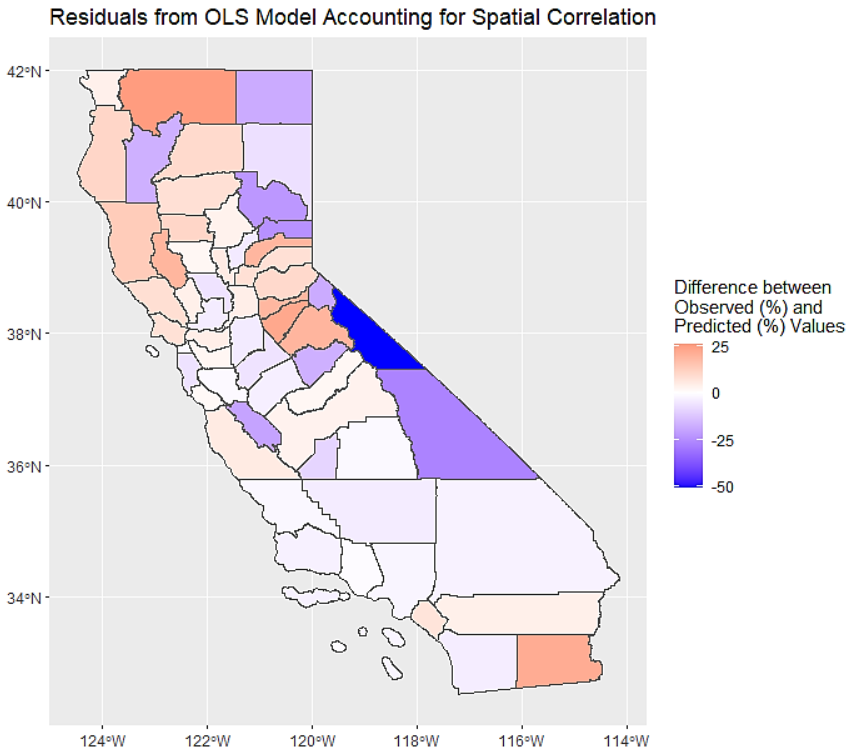


Fig 18

Ultimately however, two models were constructed: (1) a multiple linear regression model showing predictors' partial effects on vaccination rates and (2) a simultaneous autoregressive (SAR) model factoring in the effects of nearby counties. Built from the same set of predictors, the SAR model accounted for relationships between nearby counties using an adjacency structure. The adjacency structure applied weights to nearby counties bordering each other— "neighboring" counties were defined as being horizontal, vertical, or diagonal to each other. The following "queens" approach is shown in the fig. 19:

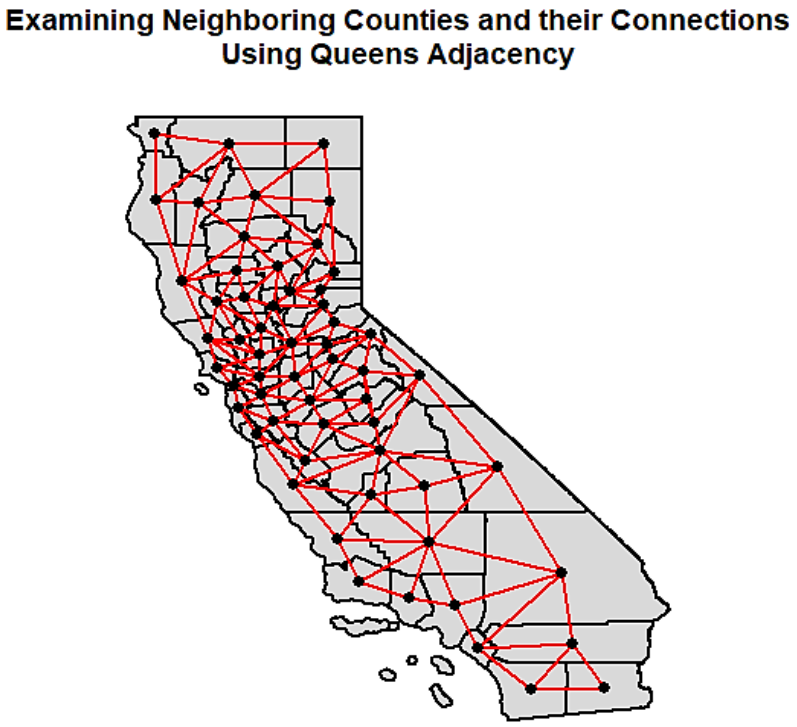


Fig. 19

For instance, San Diego County shares neighbors with Imperial, Orange, and San Bernardino Counties. Setting up the neighboring structure, a Moran's test of spatial correlation was conducted to determine whether there was statistically significant evidence suggesting nearby counties are spatially dependent on one another. A test statistic and Monte Carlo simulated sampling distribution for the test statistic were computed, resulting in a highly significant p-value ($\Pr(I > 0.349) \approx 0.0001$). This provided overwhelming evidence against the claim (initial assumption) of spatial independence. The Moran test statistic and its simulated sampling distribution (assuming spatial independence) are produced in the following fig. 20:

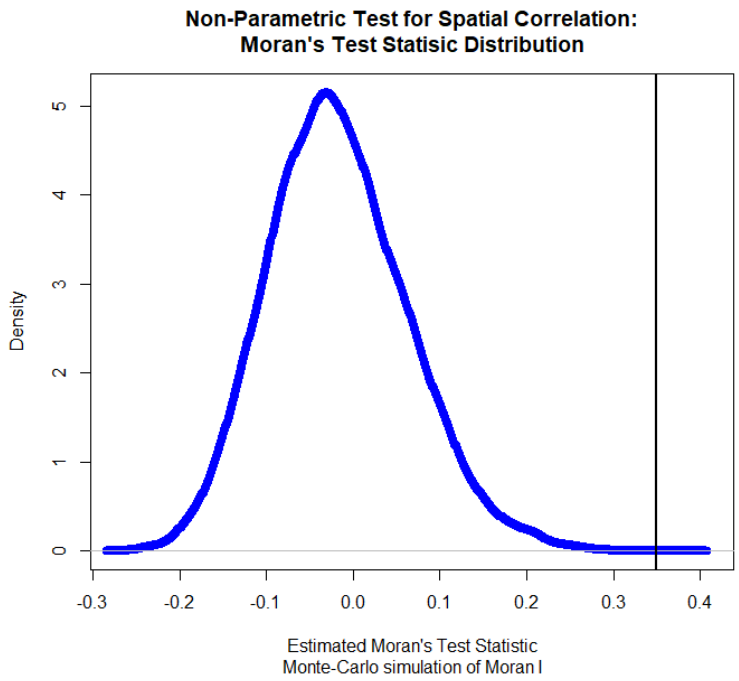


Fig 20

Confirming initial suspicions that nearby counties were influence each other, an autoregressive model incorporating spatial dependencies was constructed. A simultaneous autoregressive model was chosen, which, when using the same set of predictors yielded slightly more significant (smaller) p-values for each predictor. One reason for this increase in significance was due to a source of unexplainable error (spatial variation) now being accounted for in the model. The coefficient estimates, standard errors, and p-values are shown below for each predictor in Table .2

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.0416	0.4593	-0.0906	0.9278
prop_cases	-2.7299	4.2876	-0.6367	0.5243
prop_deaths	2.4939	3.5023	0.7121	0.4764
white	0.5608	0.3416	1.6418	0.1006
asian	0.3818	0.5451	0.7004	0.4837
black	0.4941	1.0152	0.4867	0.6265
hispanic_latino	-0.0887	0.2508	-0.3536	0.7236
bachelors_or_higher	-0.5553	0.6086	-0.9125	0.3615
median_income	0.0000	0.0000	2.5258	0.0115
medianAge	-0.0135	0.0053	-2.5340	0.0113
per_dem	0.6137	0.2954	2.0778	0.0377

Table 2

Individually significant predictors at the 5% significance level include the partisan lean of a county, with counties voting in larger proportions for Biden exhibiting, on average and holding all other factors constant, higher vaccination rates compared to counties that voted more for Trump. The median age of a county also proved highly significant, in which counties with older populations exerted, on average and holding all other factors constant, a negative effect on that county's vaccination rate. Median income similarly proved significant, suggesting that counties with higher household incomes yield, on average, a positive effect on vaccination rates. Partisanship, household income, and age best explained a county's vaccination rate. Examining the difference between the observed and fitted (predicted) values from the model, the following residuals mapping was produced.

5. CONCLUSION

Blue-shaded residuals show where the model underestimated the actual proportion of people vaccinated, red residuals show model overestimation, while white-shaded residuals show the model's prediction closely aligning with what was observed. The model typically did well along heavily populated coastal regions and performed poorly in rural, inland areas—with the issue of incorrect predictions persisting despite accounting for spatial variation. This suggests that more powerful explanatory variables are unidentified and excluded from the model, permitting further investigation beyond this project to uncover what these predictors in Fig. 21.

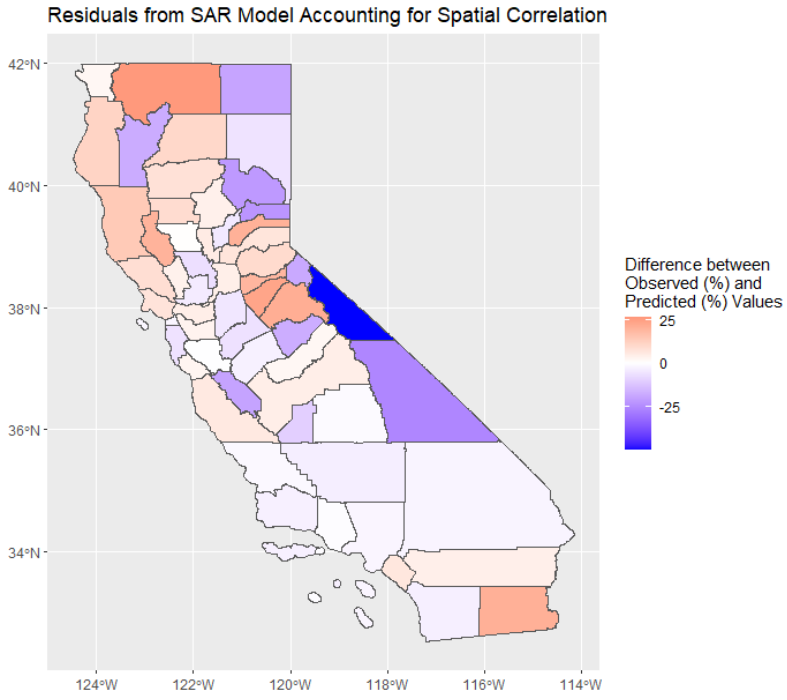


Fig 21

There are many insights from the twitter dataset but the major three of them were:

1. Pfizer is known by more people than any other vaccine. This can be seen from the Hashtag bar plot.
2. People in the United States really wanted to postpone the Olympics due to fear of covid-19. This can be clearly seen in the bar plot of hashtags and can be majorly addressed in the Word Cloud.
3. People in the United States had more hopes and trust on Vaccines because the bar plot of the sentiment analysis shows us the values of trust and anticipation were the highest.

6. LINKS

[1] https://jovian.ai/bhangarshettra-aishwarya/bda-project

[2] https://public.tableau.com/app/profile/bhagyashri.patil/viz/Covid-19Analysis_16387675599910/Covid-19Analysis

[3] https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh

[4] https://github.com/nytimes/covid-19-data

[5] https://www.census.gov/programs-surveys/acs

[6] https://drive.google.com/file/d/1mcCC-N07TK0YJkLGmbMDPrplrnXvcA_L/view?usp=sharing

[7] https://www.counties.org/general-information/county-structure-0