Berkeley
UNIVERSITY OF CALIFORNIA

COVI TRAC

W200 - Summer 2021
Final Project

Prakash Krishnan, Eric Ellestad, Sharat Nellutla

Overview

Over a year ago, the World Health Organization (WHO) declared covid-19 a pandemic. Since the first case of infection with this new coronavirus was reported in China in December 2019, SARS-CoV-2 has killed over 2.5 million people and infected at least 116 million. Beginning as an unexplained, pneumonia-like illness, first detected in China's Wuhan province, it has since spread to almost every country, bringing life across most of the world to a near-standstill for the last year. World leaders became ill, entire countries were locked down to prevent the spread of infection and international travel ceased.

Given the tremendous amount of devastation created by one of the most impactful pandemics in our country's modern history, we set out to dig deeper into several aspects of this pandemic. This report will look at Covid-19 infections, deaths and vaccinations in the United States to assess if we are on a path to recovery and returning to normal life and where we stand in the battle against the Covid-19 pandemic.

Executive Summary of Key Insights

The first year of the Covid-19 pandemic in the United States was defined by three distinct surges of rising infections, each with a peak that surpassed that of the previous wave. The first peak was in April 2020 with over 35,000 daily cases and New York City as its epicenter. The second peaked in July 2020 with over 79,000 daily cases and more widespread infection. The third and deadliest wave occurred over December 2020 and January 2021 with a peak of over 392,000 daily cases and 4,500 daily deaths driven by winter weather and holiday gatherings.

In a groundbreaking feat of modern medicine and technology, three Covid vaccines were developed, tested, approved, and began to be distributed within less than 12 months of the first case on US soil. As of July 2021, almost 80% of Americans over the age of 65 have been fully vaccinated and over half of the entire US population has received their first dose. Increasing vaccination rates coincided with a sharp reduction in cases and deaths after the third wave, which led many to believe the pandemic was over.

Recent months have shown a concerning increase in Covid-19 infections. The CDC attributes the resurgence to the spread of the Delta Variant which has become the dominant strain globally and in the US. The Delta Variant is more virulent than previous strains and more than twice as infectious as the original virus due to mutations in its spike protein. These mutations also dampen the immune response in vaccinated individuals, which has led to increasing reports of "breakthrough" infections in those who are vaccinated and have already contracted the virus.

While this trend is highly concerning, the good news is that the vaccines have proven to be effective at preventing infections and even more effective at preventing serious symptoms, hospitalizations, and deaths. The path to defeating Covid-19 lies through an aggressive vaccination campaign that must include the vast majority of Amercians and states. Expect to see a return of mask mandates and social distancing requirements while infections continue to trend upward and until the nationwide vaccination rate reaches 70-80%. The Covid-19 pandemic is not over yet and continued vigilance is required.

A deeper county-level analysis of California data shows similar trends to the broader national trends. A review of the correlation coefficients for the different waves and also the most recent period after vaccine distribution began provides evidence that vaccination helps in reducing case counts, patients in the hospital and ultimately deaths. The other interesting information is the divide between metro and rural areas. Small and rural counties are significantly lagging behind in vaccination rates and a greater focus on these rural counties and the unvaccinated will be required to reach herd immunity targets.

Research Questions

National Covid-19 Cases and Deaths Over Time

1. What is the long-term case/death trend from the start of data collection and when did peaks and valleys occur? (Time Series) (2 years)
    a. Provide potential reasoning for the events
2. What is the short-term case/death trend in the recent X months and what does the trend tell us nationally? (last 3 months)?

National Covid-19 Vaccination Analysis

1. What percent of the US population has been partly or fully vaccinated against COVID-19?
2. Explore COVID-19 vaccinations by state and nationally (Number of doses administered, people with at least 1 dose, and people fully vaccinated)
    a. 5 states with highest and lowest vaccination rates and compare the number of vaccines distributed
    b. 5 states with highest and lowest with partly vaccinated population
    c. 5 states with highest and lowest with fully vaccinated population by manufacturer (Pfizer, Moderna, J&J)
3. What percent of the US population in each group are partially vaccinated against COVID-19?
4. What are the trends in vaccination rates nationally and by state and is it evenly administered by state?

Correlation of Cases, Deaths and Vaccinations

1. What is the trend for the same time slice as the vaccines (December 2020-Current)?
2. Understand the visual correlation between cases and vaccine rate trends. What is the data telling us?
3. Understand the visual correlation between cases and death trends. What is the data telling us?
4. Choose some counties in CA and investigate how vaccination affects these stats

Datasets

Primary Dataset

COVID-19 Vaccinations in the United States

a. Source: Centers for Disease Control & Prevention (CDC)
b. Link to dataset: COVID-19 Vaccinations in the United States,Jurisdiction | Data | Centers for Disease Control and Prevention
c. Size of Dataset: 69 Columns x 14.7k Rows
d. Variable names and descriptions listed here.

Secondary Datasets

United States COVID-19 Cases and Deaths by State over Time

e. Source: Centers for Disease Control & Prevention (CDC)
f. Link to dataset: United States COVID-19 Cases and Deaths by State over Time | Data | Centers for Disease Control and Prevention
g. Size of Dataset: 15 Columns x 33.3k Rows

California Datasets:

h. Team will analyze three CA Public Datasets on the following:
    i. Covid cases, deaths and tests (17 Columns x 33.2k Rows)
    ii. Hospitalization and ICU beds (9 Columns x 27.2k Rows)
    iii. Vaccines administered and provide summary assessment on where we are in the fight against Covid-19 and draw insights (17 Columns x 14.0k Rows)
i. Please see appendix for CA Dataset details.

Data Cleaning & Sanity Checks

National Cases & Deaths Data:

1. Data Cleaning:
    a. Converted "Date" variable from string to datetime to evaluate time series data.
    b. Reduced variables down to only what we will be analyzing: (submission_date, state, tot_cases, new_case, tot_death, new_death)
    c. Combined "NYC" with "NY" records since New York City was recorded separately from the rest of the New York State. We analyzed the combined data for the state.
    d. Filtered out jurisdictions that are not part of the 50 standard states and the District of Columbia (Guam, Virgin Islands, American Samoa, etc)
    e. Created a United States time series which is the sum of the 50 Official States plus DC. Filtered out percentage variables and recreated after summation.
    f. Added a new trailing 7-day average variable for "new_case" and "new_death" to avoid spikes in data from lack of weekend reporting and Monday backlog.

2. Sanity Checks:
    a. Confirmed number of variables and number of rows matched source data website.
    b. Used describe(), values_count(), unique(), and plots to evaluate min, max, and time series distribution of variable values.
    c. Source dataset is professionally maintained and did not require much cleaning.

National Vaccine Data:

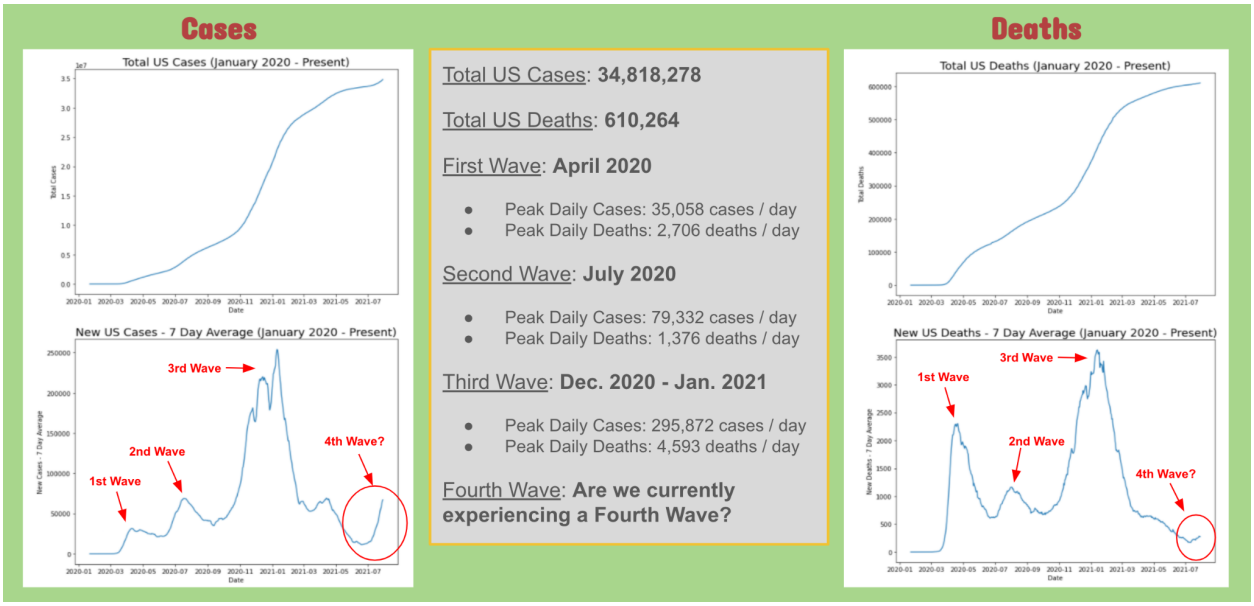
1. Data Cleaning:
    a. Converted “Date” variable from string to datetime to evaluate time series data.
    b. Filtered out jurisdictions that are not part of the 50 standard states and the District of Columbia (Guam, Virgin Islands, American Samoa, etc)
    c. Re-created a United States national time series which was the sum of the 50 Official States plus DC.
2. Sanity Checks:
    a. Confirmed number of variables and number of rows matched source data website.
    b. Used describe(), values_count(), unique(), and plots to evaluate min, max, and time series distribution of variable values.
    c. Source dataset is professionally maintained and did not require much cleaning.

California Data:

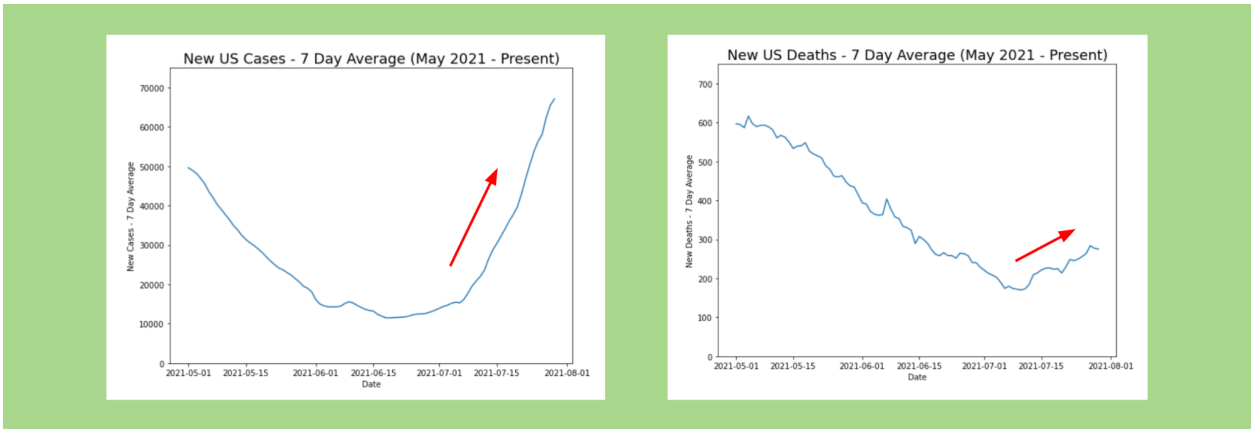
1. Data Cleaning:
    a. Filtered out jurisdictions not part of CA (outside CA, unknown)
    b. Filtered out summary CA data
    c. Standardized Column Names for Area and Date. The three datasets had different column names for county name and date
    d. Converted “Date” variable from string to datetime to evaluate time series data.
    e. Performed an outer join on three datasets with date and area (county name) as key
2. Sanity Checks:
    a. Performed a sanity check on population count, test count, death count and vaccine count against available and published CA data to ensure alignment

Covid-19 Cases and Deaths in the United States

The first Covid-19 infections were officially reported in the United States in January 2020. The first wave of infections (April 2020) hit New York City the hardest, but the second wave of infections (July 2020) was more widespread and the third wave was the deadliest and occurred during the winter months with spikes around the holidays.



The Delta Variant, which originated in India in late 2020, is more than 2x as transmissible as the original Covid-19 strain which originated in Wuhan, China. As the Delta Variant has spread, it outcompeted other less virulent strains and is now responsible for over 80% of all Covid-19 infections in the US. In addition to its increased virulence, there are increasing reports of "breakthrough infections" of the Delta Variant in fully vaccinated individuals due to the mutations to the spike protein which increases infectiousness and decreases the immune response in vaccinated individuals. The confluence of the spread of the Delta Variant with the reopening of businesses, restaurants, and offices has led to an increase in cases, which is apparent in the past 3 months of data:



The CDC and public health officials continue to warn of the risks of infection to both unvaccinated and vaccinated individuals. In order to combat rising cases, both Los Angeles County and San Francisco County have reinstated indoor mask mandates. In addition, vaccine requirements are increasingly being introduced by employers and federal agencies.

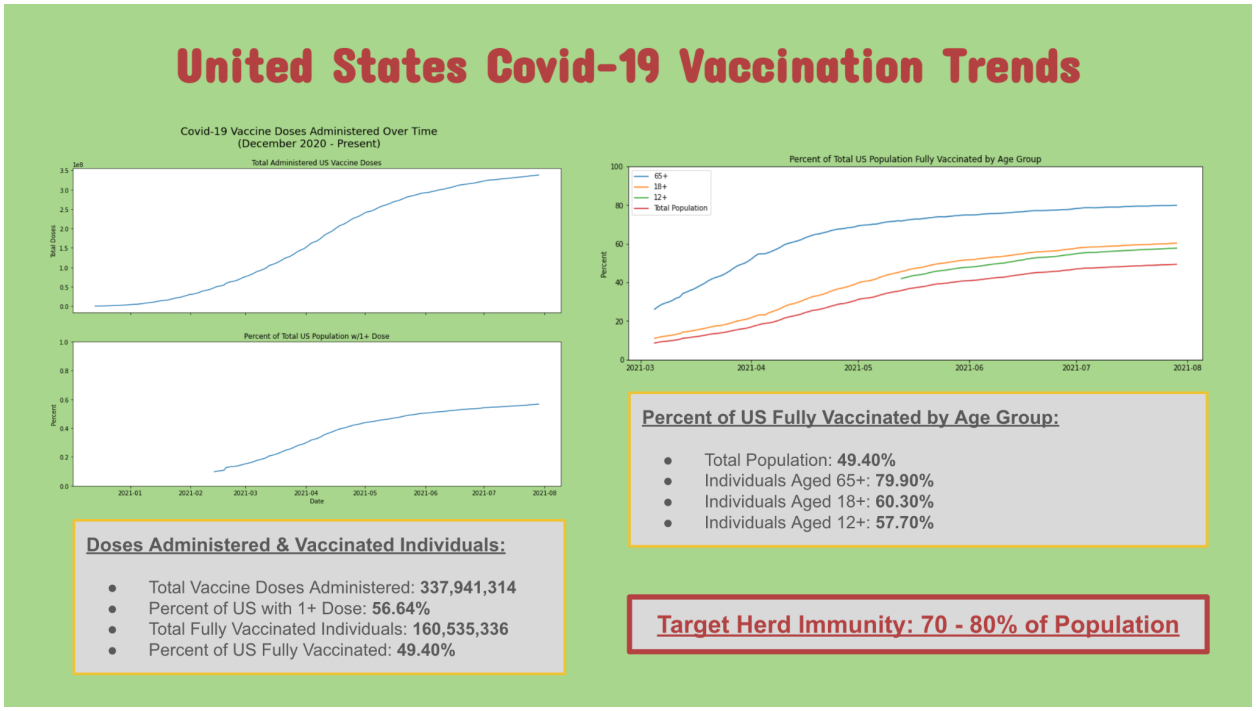
Correlating Daily Cases vs. Daily Deaths

	US_new_cases_avg	US_new_deaths_avg
US_new_cases_avg	1.000000	0.820818
US_new_deaths_avg	0.820818	1.000000

There is a visible and quantitative positive correlation between Covid-19 cases and deaths, which is what we would expect given infections lead to deaths. The first wave of infections had a much higher rate of deaths per infection than the second and third waves. The first wave was due to the original Wuhan strain of Covid-19 whereas the larger waves of infections were due to variants such as the Alpha Variant which originated in the UK and while it was more transmissible, it had a lower rate of hospitalizations and deaths.

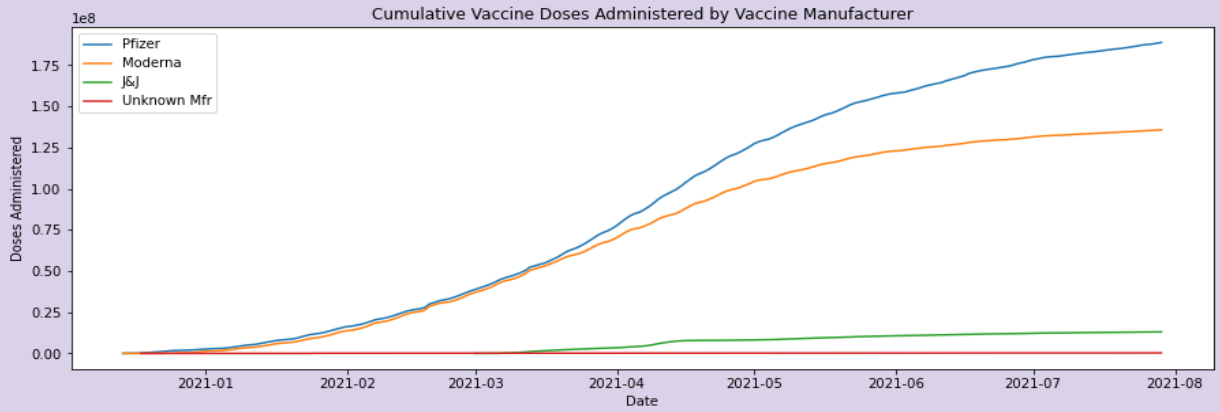
What's the nation's progress on vaccinations?

As of Friday 7/30, about 190.5 million or 58% people have received at least one dose of a Covid-19 vaccine, including about 164.2 million or 50% people who have been fully vaccinated by Johnson & Johnson's single-dose vaccine or the two-dose series made by Pfizer-BioNTech and Moderna.



The first vaccine dose administered in the United States was on December 13th, 2020. We can see by the summer of 2021, vaccine supply began to exceed demand, as most individuals who wanted the vaccine had already received it. Vaccinations rates began to drop as the remaining unvaccinated individuals were increasingly people who did not want the vaccine due to various concerns. Ongoing vaccination efforts are focused on educating people on vaccine benefits, vaccine safety, and fighting misinformation regarding Covid-19 and its vaccines.

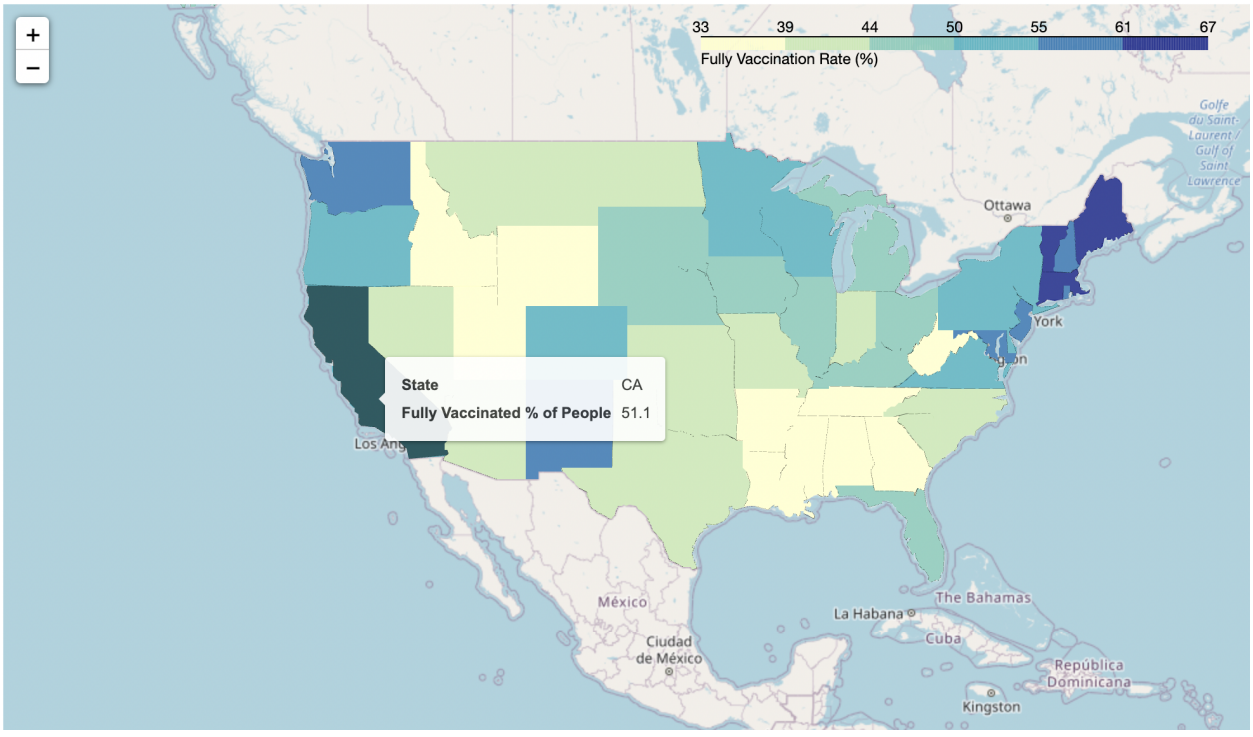
Cumulative Vaccine Doses Administered by Vaccine Manufacturer:



Cumulative Doses Administered by Vaccine Manufacturer:

- Pfizer: 188,819,532
- Moderna: 135,757,170
- Johnson & Johnson: 13,063,054
- Unknown Mfr: 301,558

Covid Vaccination Differences by State



Top 5 and Bottom 5 States by Per Capita Vaccination %

	State	Percent Fully Vaccinated	Vaccines Distributed per 100k
1	VT	67.4	152110
2	MA	63.7	144883
3	ME	63.4	142153
4	CT	63.1	137907
5	RI	61.3	147143

	State	Percent Fully Vaccinated	Vaccines Distributed per 100k
47	LA	36.8	92111
48	WY	36.6	91039
49	AR	36.2	98991
50	MS	34.4	93603
51	AL	34.2	103006

Top 5 vs Bottom 5 States by Vaccination Rate:

- There are substantial differences in vaccination rates by state:
  - **Vermont** has the **highest rate** of **67.4%**
  - **Alabama** has the **lowest rate** of **34.2%**
- All 5 of the states with the *highest* vaccination rates are from the *Northeast Region* of the United States
- All 5 of the states with the *lowest* vaccination rates are from the *Midwest and Southern Regions* of the United States
- In order to achieve "herd immunity" levels of national vaccination, vaccination rates must increase in the regions of the United States where *vaccine hesitancy* is strongest.

Correlation of Cases, Deaths and Vaccinations

	US New Cases 7-Day Avg	US New Deaths 7-Day Avg	US Percentage of Population Fully Vaccinated
US New Cases 7-Day Avg	1.000000	0.936966	-0.766211
US New Deaths 7-Day Avg	0.936966	1.000000	-0.866682
US Percentage of Population Fully Vaccinated	-0.766211	-0.866682	1.000000

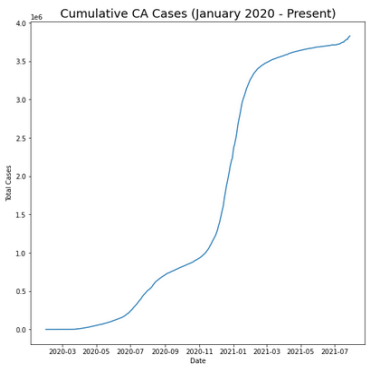
Cases, Deaths, and Vaccinations since January 2021:

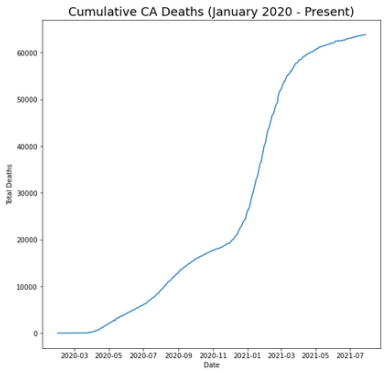
- Covid-19 Cases and Deaths are highly correlated (0.94).
- Cases and Deaths have decreased as Vaccinations have increased.
- Vaccinations are more strongly correlated with deaths (-0.87) than with cases (-0.77).

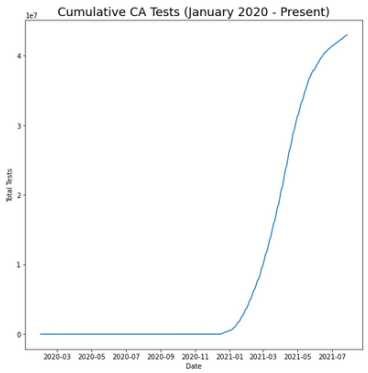
There is a visual and quantitative positive correction between cases and deaths, as well as negative correlation between vaccinations and cases as well as deaths. This supports the scientific research that vaccines help prevent infections and deaths from Covid-19 and widespread vaccination is the foundation of an effective strategy to finally defeat the Coronavirus.

Deep Dive into California Statistics

California has been at the forefront in the fight against Covid-19 as of July 31, 2021 there have been 3.84 million confirmed cases, 63,935 deaths and 73.56 million Covid tests.

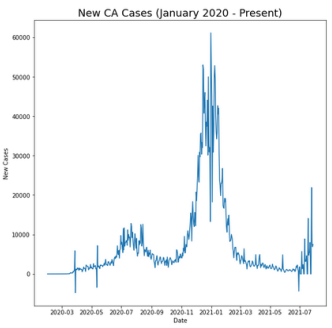


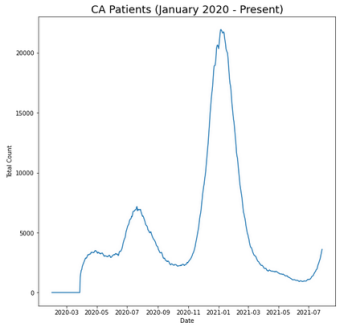


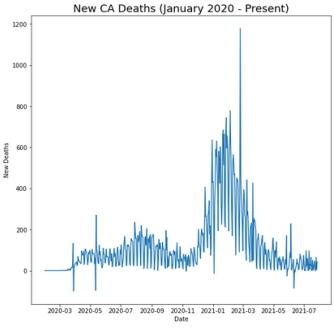


Long-Term Trends

Significant progress has been made in the fight against Covid-19 in terms of number of people vaccinated and a reduction in the number of cases, deaths and number of patients in the hospital. Covid-19 cases skyrocketed in the early part of the year with over 50,000 new cases every day, 800 daily deaths and over 20,000 patients in the hospital. This resulted in a severe strain on the hospital system resulting in significant reduction in ICU bed availability. This was about the time vaccines were made available for the high risk population and essential workers. Looking at the long terms charts for cases, patients in the hospital and deaths it is evident that the pandemic transpires in waves and very pronounced peaks and valleys. It is clear from the charts below the number of patients closely follows the cases as people get infected and the deaths happen 2-10 weeks later and this is also clear from the death trend chart.







Vaccination Statistics

Over 43 million doses have been administered with 51% of the population fully vaccinated and 58% of the population with at least one dose. California ranks 18th in the nation when measured by the percent of population fully vaccinated and that is not surprising given that California is the biggest state in the country.

CA Vaccination Summary

area	fully_vaccinated	at_least_one_dose	population	%_of_pop_fully_vaccinated	%_of_pop_atleast_one_dose
Total CA	20,847,606	23,908,552	40,129,160	51.95	59.58

But the progress on vaccination has not been evenly distributed across the counties.

The counties that are affluent and more progressively inclined (meaning Democratic) have a higher degree of population vaccinated while the rural counties and more conservative counties (meaning Republican) are lagging behind in vaccination.

See table showing the top five and bottom five counties with % fully vaccinated and partially vaccinated.

Top Five CA County Vaccination Summary

area	fully_vaccinated	at_least_one_dose	population	%_of_pop_fully_vaccinated	%_of_pop_atleast_one_dose
Marin	188,994	204,500	260,800	72.5	78.4
San Francisco	609,751	665,397	892,280	68.3	74.6
Santa Clara	1,324,402	1,441,527	1,967,585	67.3	73.3
San Mateo	517,568	568,621	778,001	66.5	73.1
Contra Costa	743,216	799,163	1,160,099	64.1	68.9

Bottom Five CA County Vaccination Summary

area	fully_vaccinated	at_least_one_dose	population	%_of_pop_fully_vaccinated	%_of_pop_atleast_one_dose
Modoc	2,954	3,190	9,475	31.2	33.7
Mariposa	5,275	8,061	17,795	29.6	45.3
Tehama	19,437	21,763	65,885	29.5	33.0
Kings	44,067	52,374	156,444	28.2	33.5
Lassen	6,060	6,716	30,065	20.2	22.3

<u>Correlation Between Cases, Deaths, Hospitalization and Vaccine Administration-</u>

The following four charts demonstrate how the pandemic operates and impacts society.

Cases are a leading indicator, followed by hospitalization and finally death. It is very clear visually that as the cases ramped at the start of 2021, Patient Trends in the hospital followed a similar pattern with a slight time lag and deaths occured, soon after. November-2020 to Mid-January 2021 was by far the worst period of the pandemic and then as vaccination began and social distancing was enforced, things started improving.

One other interesting observation from the case trend is the double wave of the Pandemic. After an initial high in August of 2020, number of cases declined to about end of October 2020. Then the second wave kicked in. This can be largely attributed to cold climate when people are more indoor and people possibly not adhering to social distancing.

Also when you closely look at the Case Trends, we are actually at the beginning of Wave 3 of the Pandemic and case trends are alarmingly increasing.

The data below provides correlation coefficients for the four key periods:

(1) Wave 1: Beginning of Pandemic to April, 2020
(2) Wave 2: April, 2020 - August 2020
(3) Wave 3: August 2020, February 2021
(4) Vaccination Impact: February 2021 - Current

```
Correlation Coefficient for Wave 1: [Beginning - April, 2020]
```

	CA_new_deaths	CA_new_cases	CA_new_patients
CA_new_deaths	1.000000	0.872982	0.754541
CA_new_cases	0.872982	1.000000	0.547829
CA_new_patients	0.754541	0.547829	1.000000

```
Correlation Coefficient for Wave 2: [April, 2020 - August, 2020]
```

	CA_new_deaths	CA_new_cases	CA_new_patients
CA_new_deaths	1.000000	0.506798	0.367611
CA_new_cases	0.506798	1.000000	0.862122
CA_new_patients	0.367611	0.862122	1.000000

```
Correlation Coefficient for Wave 3: [August, 2020 - February, 2021]
```

	CA_new_deaths	CA_new_cases	CA_new_patients
CA_new_deaths	1.000000	0.410367	0.576236
CA_new_cases	0.410367	1.000000	0.765986
CA_new_patients	0.576236	0.765986	1.000000

```
Correlation Coefficient for Impact of Vaccinations: [February, 2021 - Current]
```

	CA_new_doses	CA_new_deaths	CA_new_cases	CA_new_patients
CA_new_doses	1.000000	0.430582	-0.070192	0.101432
CA_new_deaths	0.430582	1.000000	0.396718	0.560133
CA_new_cases	-0.070192	0.396718	1.000000	0.653514
CA_new_patients	0.101432	0.560133	0.653514	1.000000

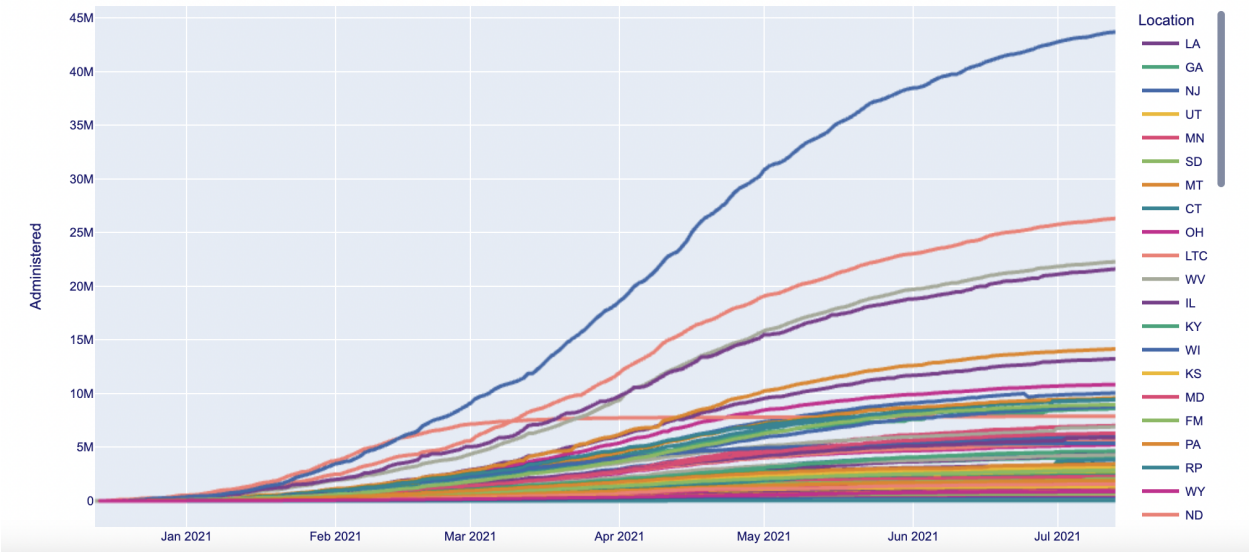
The above shows interesting observations:

(1) New cases to new patient correlation (meaning cases became hospitalizations) worsened with each wave and then as vaccines started rolling out early in the year the correlation coefficient is starting to decrease. This implies that vaccines have an impact in reducing severity of infection that requires hospital care.
(2) New cases to new death correlation also shows improvement over each of the waves and then as vaccines kicked in showing further improvement. This can attributed to several reasons. Early in the pandemic availability of respirators and ICU beds, caused many deaths for the elderly. As more ICU beds and respirators became available, with each wave the mortality count came down and finally as the vaccines rolled out, it further helped in preventing deaths.

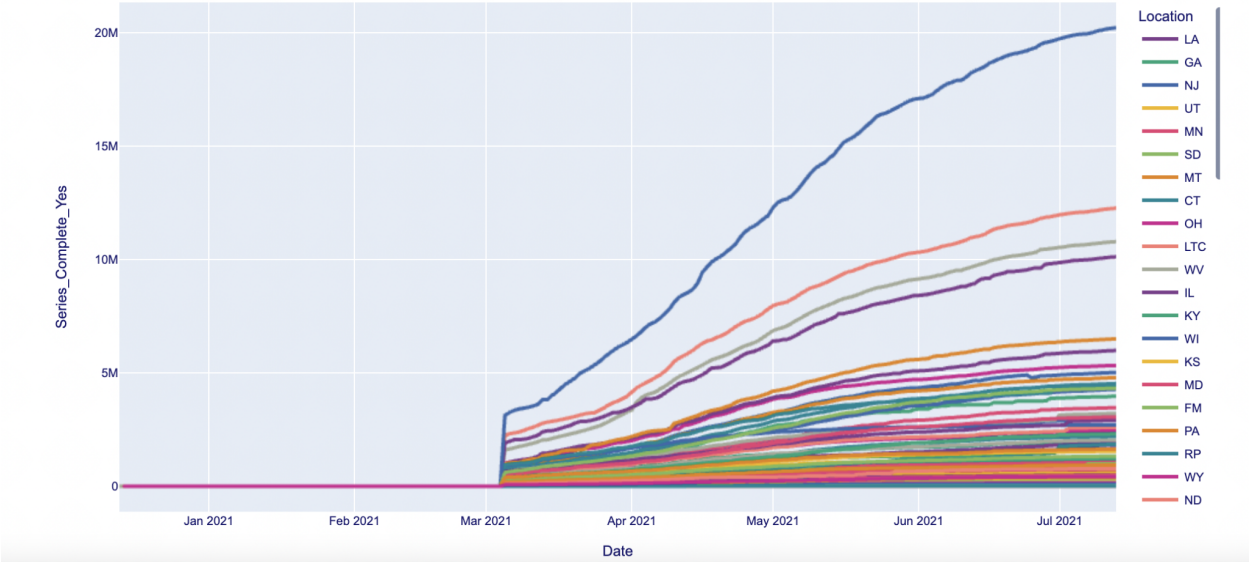
Appendix:

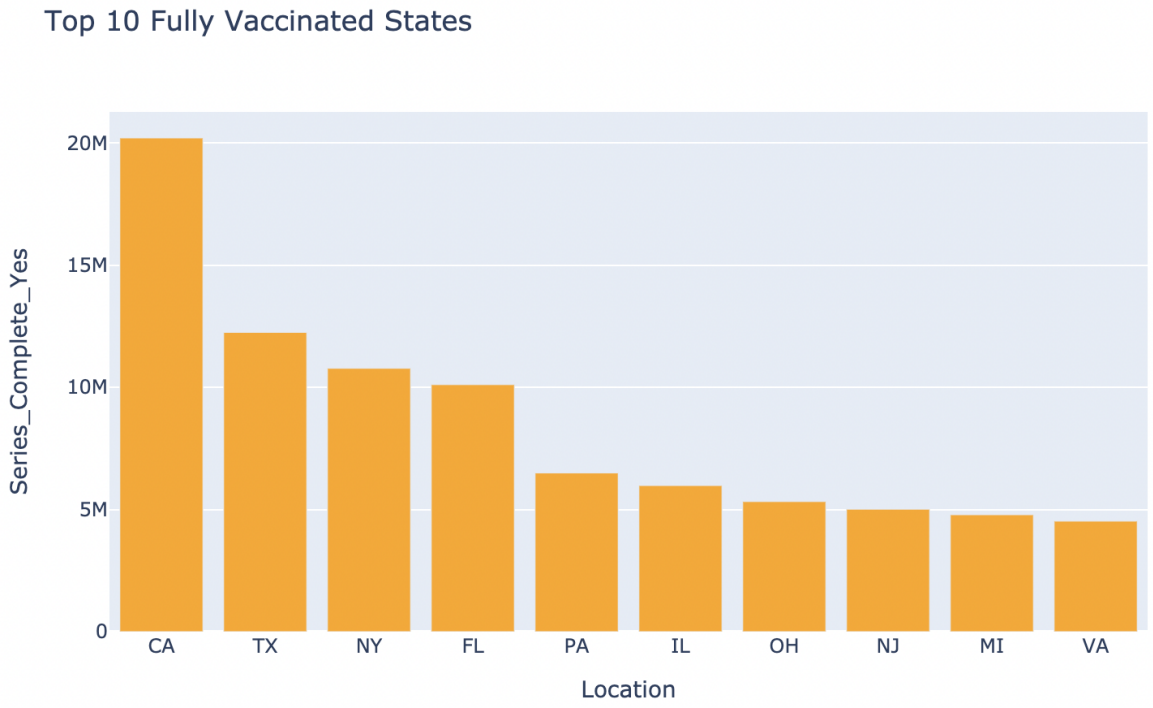
Supplemental National Charts:

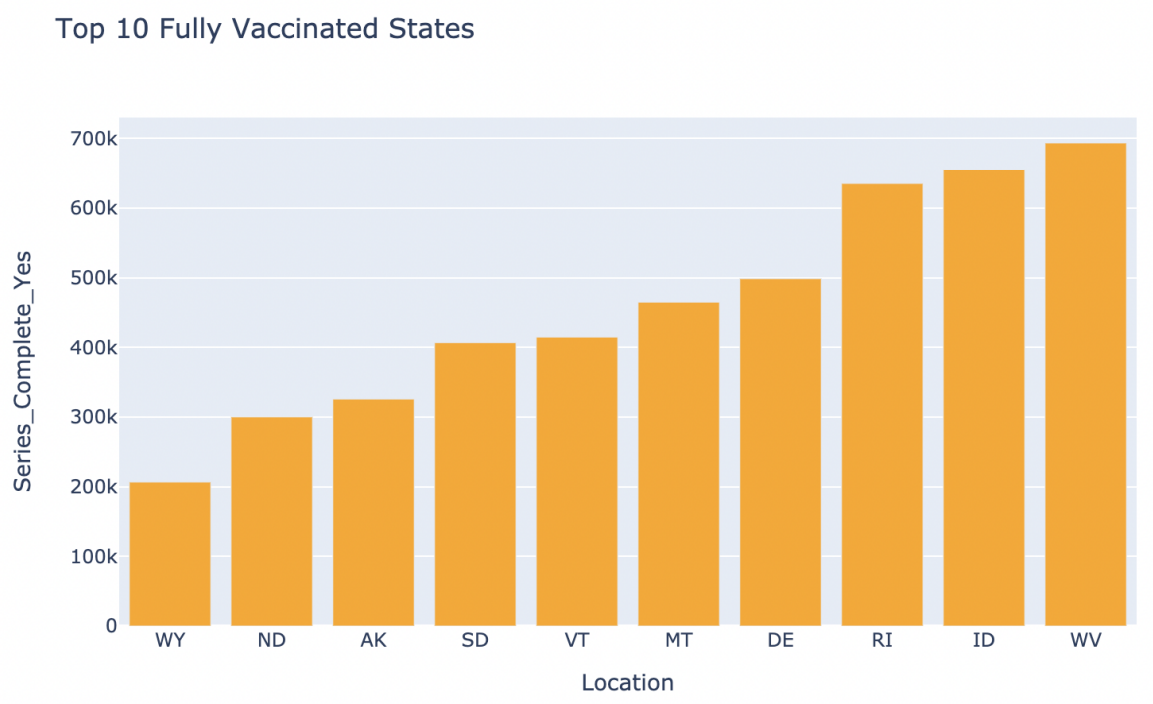
Number of Vaccinations Administered



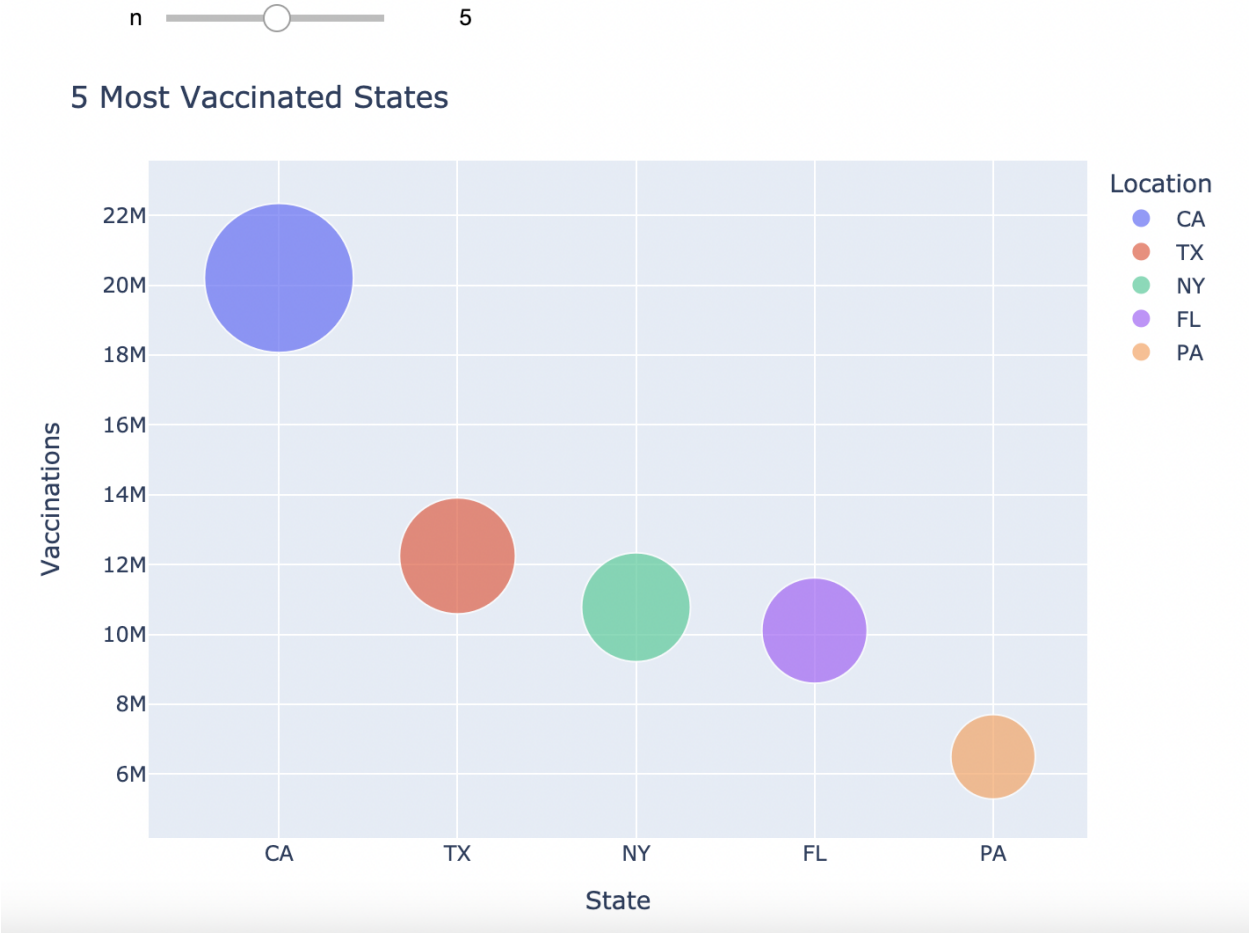
Fully Vaccinated States




Top 10 Fully Vaccinated States
Series_Complete_Yes
20M
15M
10M
5M
0
CA
TX
NY
FL
PA
IL
OH
NJ
MI
VA
Location


Top 10 Fully Vaccinated States
Series_Complete_Yes
700k
600k
500k
400k
300k
200k
100k
0
WY
ND
AK
SD
VT
MT
DE
RI
ID
WV
Location

Top 5 and Bottom 5 States by Per Capita Partial Vaccination %



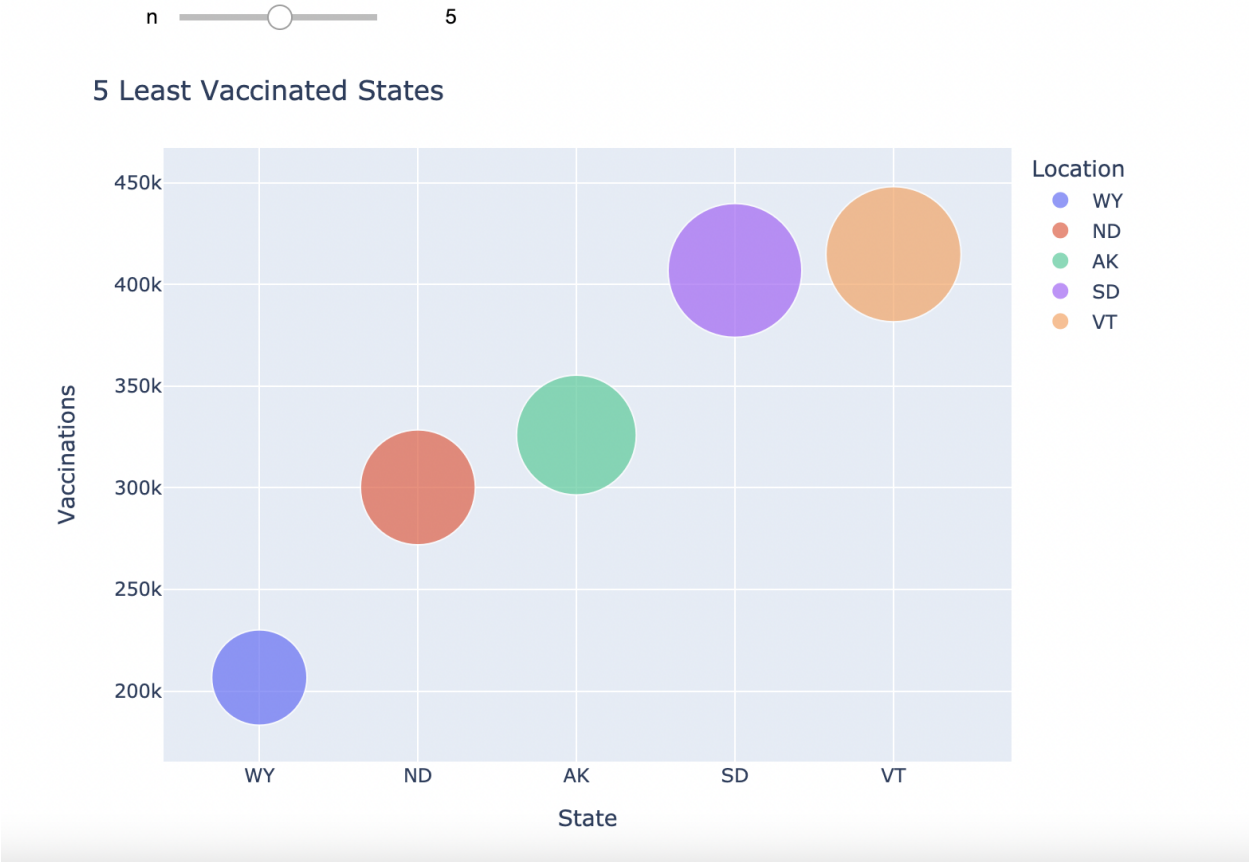
Top 5 Partially Vaccinated States

	State	Percent w/1+ Dose
1	VT	75.4
2	MA	72.4
3	HI	71.3
4	CT	69.6
5	ME	68.2

Bottom 5 Partially Vaccinated States

	State	Percent w/1+ Dose
47	AL	42.7
48	LA	41.8
49	WY	41.5
50	ID	41.0
51	MS	39.3

These results are fairly consistent with the Top 5 and Bottom 5 states by fully vaccination rates.



Top 5 and Bottom 5 States of Total Vaccinated Individuals by Vaccine Manufacturer:

	Pfizer_top5	Pfizer_bot5	Moderna_top5	Moderna_bot5	JnJ_top5	JnJ_bot5
1	CA	WY	CA	WY	CA	WY
2	TX	ND	TX	ND	FL	AK
3	NY	AK	NY	AK	TX	ND
4	FL	DC	FL	DC	NY	SD
5	PA	SD	PA	VT	PA	DC

Unsurprisingly, the most populous states have the highest number of vaccinated individuals for each vaccine, while the least populous states have the lowest number of vaccinated individuals per vaccine. A more interesting view would be the percentage of Vaccines administered by Vaccine manufacturers.

Top 5 and Bottom 5 States by Percentage of Vaccines Administered by Vaccine Manufacturer:

	Pfizer_top5_pct	Pfizer_bot5_pct	Moderna_top5_pct	Moderna_bot5_pct	JnJ_top5_pct	JnJ_bot5_pct
1	HI	WY	AR	HI	ME	HI
2	VA	AR	WY	VA	FL	WV
3	MO	ME	WV	MO	NV	GA
4	MD	AL	AL	MN	VT	MS
5	MA	MT	MS	MD	UT	OK

Here we see some interesting differences in state-by-state distribution of different vaccines. There is some overlap between the Top 5 Pfizer States and Bottom 5 Moderna States, as well as the Top 5 Moderna States and Bottom 5 Pfizer States. This makes sense as these are the two most common vaccines and states with a higher percentage of one would be lower in the other.

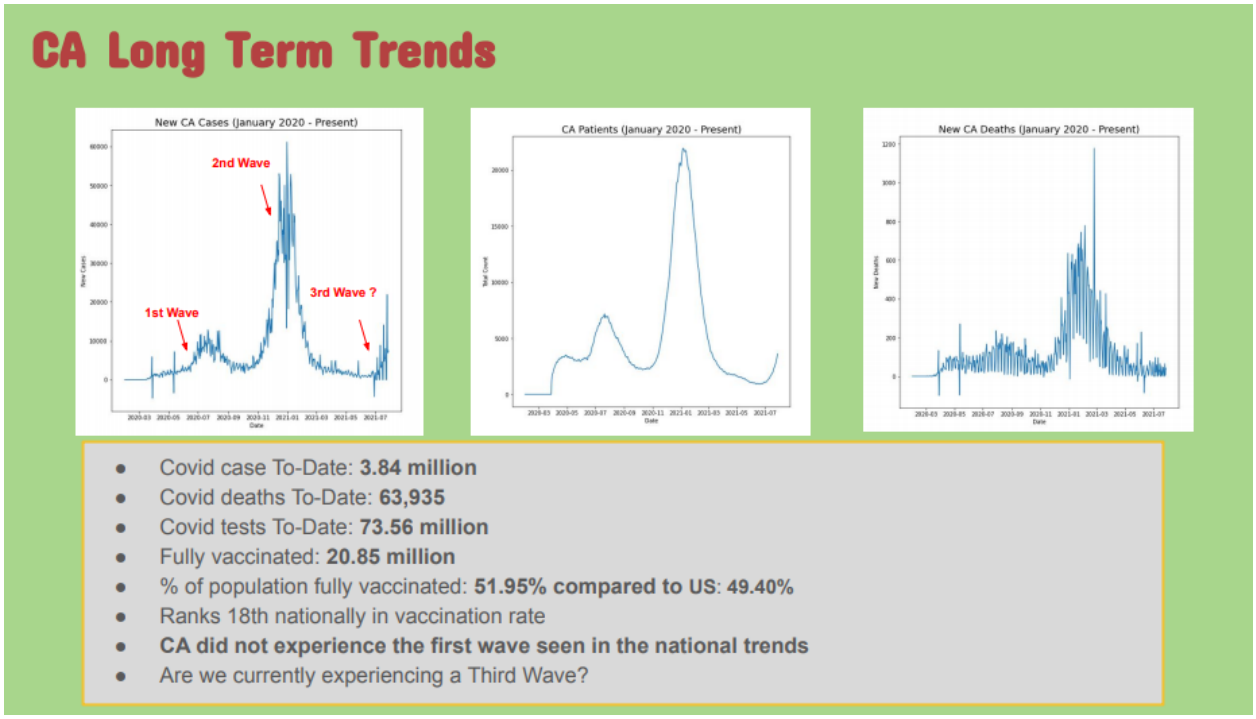
California Additional Details

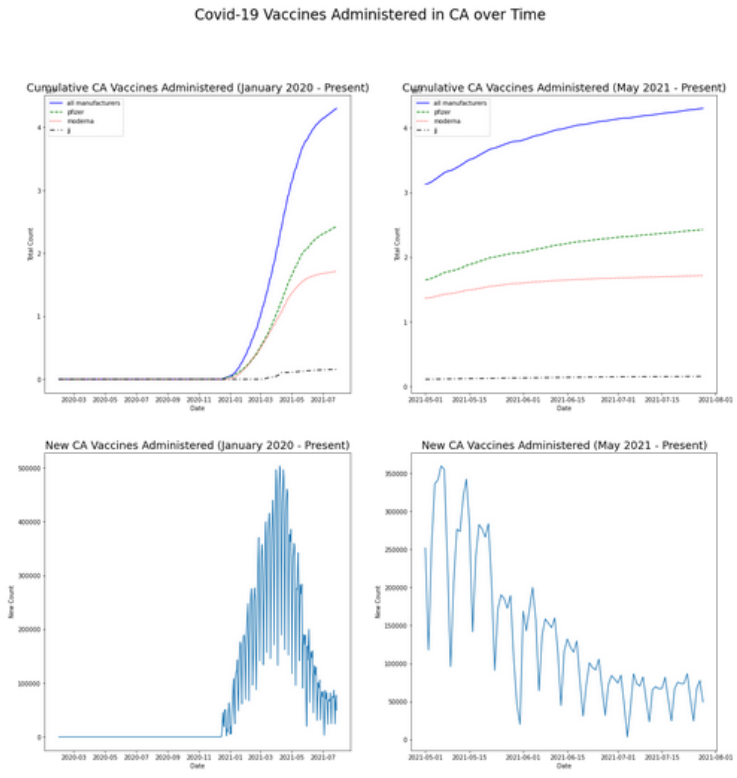
CA Facts and Figures:

California the biggest state in US in terms of population has been leading the fight against COVID-19 and the key takeaways from the analysis are:

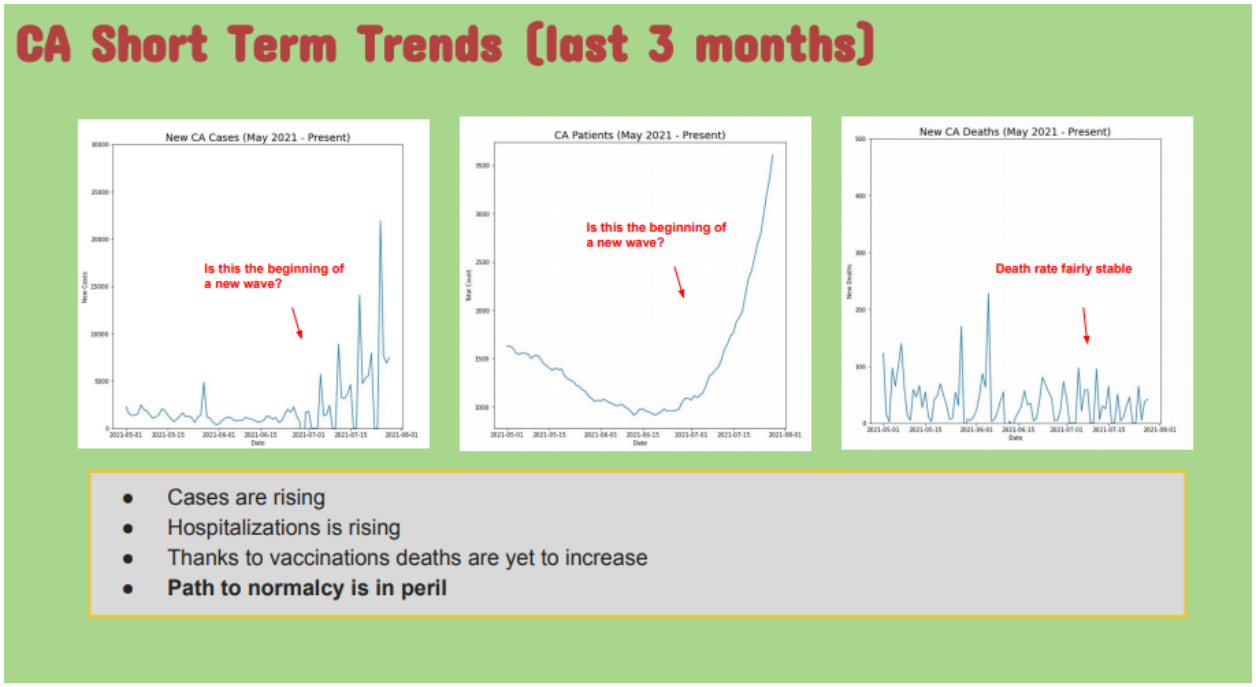
1. 59% of the Population are partially vaccinated with at least one dose of vaccine
2. 20.8 million people out of 40.1 million people are fully vaccinated
3. The vaccine administration is not uniform with the rural counties significantly lagging behind in vaccinations
4. On a per capita basis, CA ranks 11th nationally on vaccines administered per 100k population
5. Long term trends of cases, patients in hospitals and death counts have shown significant improvements since the peak of the pandemic in January-February of 2021
6. However short-term trends in the last few months are alarming. Daily count of cases, and patients in hospital are showing a steep uptick. Daily vaccines administered are not picking up indicating people are still reluctant to take the vaccines
7. This is not good news and the path normalcy is in peril. There needs to be concerted efforts to get the unvaccinated vaccinated and the political and religious leaders need to encourage people to get the vaccines. The rural and smaller counties have to increase the doses administered.
8. While progress has been made in the fight against COVID-19, the war is not won yet and we are seeing the beginnings of a new wave of pandemic.

Long-Term Trends





Recent Three Month Trends (May-2021 to Current)-



A view of the recent three months is alarming and concerning in the fight against the Pandemic. Cases are going up, the number of patients in the hospitals are rising and the rate of vaccination has tapered down. All of this points to a rough path ahead. Unless people socially distant, wear a mask and the unvaccinated get the dose, we are possibly seeing a next wave. All the hard work that was done so far may come to a screeching halt.

Mask mandate may be enforced, offices will delay workers coming back and school openings may be impacted. So the path to normalcy is in peril.

<u>County Drill Down-</u>

This analysis further drills down into three select counties. (1) Santa Clara County (2) Los Angeles County (3) Riverside County. The three counties were selected because they have different characteristics and are big populations centers.

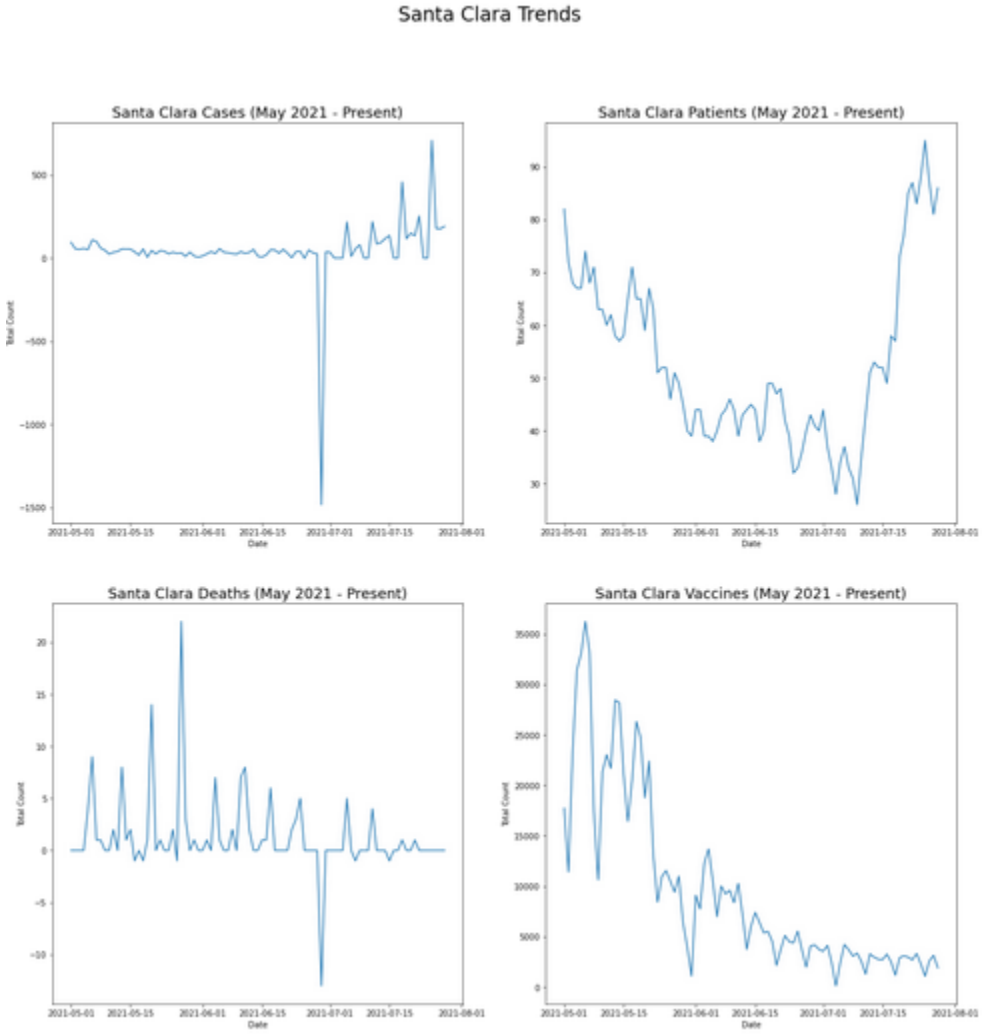
area	fully_vaccinated	at_least_one_dose	population	%_of_pop_fully_vaccinated	%_of_pop_atleast_one_dose
Santa Clara	1,310,903	1,429,678	1,967,585	66.6	72.7
Los Angeles	5,356,962	6,110,910	10,257,557	52.2	59.6
Riverside	1,027,295	1,187,517	2,468,145	41.6	48.1

Santa Clara County is California's 6th most populous county. Santa Clara is the most populous county in the San Francisco Bay Area and in Northern California. The county seat and largest city is San Jose, the 10th most populous city in the United States, California's 3rd most populous city and the most populous city in the San Francisco Bay Area.

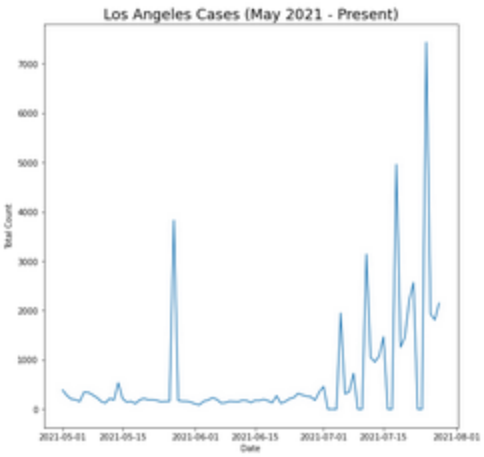
Home to Silicon Valley, Santa Clara County is an economic center for high technology and has the third highest GDP per capita in the world (after Zurich, Switzerland and Oslo, Norway), according to the Brookings Institution.

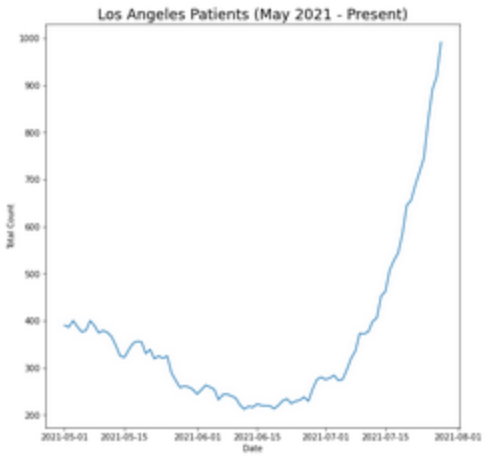
Los Angeles County, officially the County of Los Angeles, is the most populous county in the United States. It is the most populous non–state-level government entity in the United States. Its population is greater than that of 41 individual U.S. states. The county is home to more than one-quarter of California residents and is one of the most ethnically diverse counties in the United States. Its county seat, Los Angeles, is also California's most populous city and the second most populous city in the United States, with about four million residents.

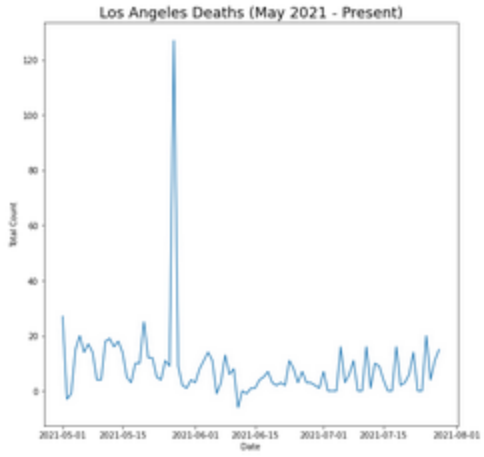
Riverside County is a county located in the southern portion of the U.S. state of California and the 10th-most populous in the United States.

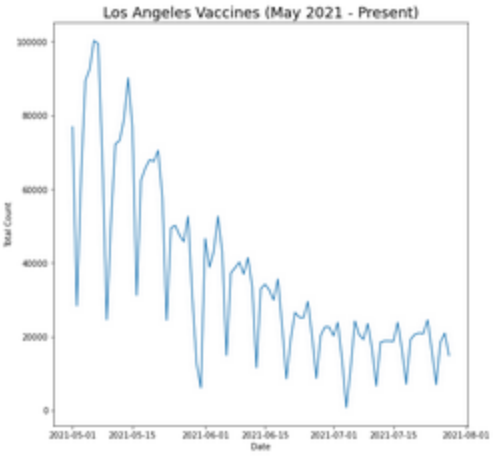


Los Angeles Trends









Riverside Trends