Covid-19 - The Story of the States

Presented by:
Srihari Shekhar
Khushboo Zope
Aakash Shah

Key Objectives

- Understand the story of Covid-19 - for each state
  - How the infection/death count moved over time for each state?
  - When did each state Peak?

- See the vaccination progress for each state
  - Compare the state to the National Average

- Create a rough guideline for Covid Data Analysis
  - Multiple data points need cleaning and overall a widely available and used dataset!

- Fill PDFs using Template file & Python Code

Data Sources

Covid-19 Infection data:

https://www.kaggle.com/fireballbyedimyrnmom/us-counties-covid-19-dataset

Which derives its data from the New York Times GITHUB source:

https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv

(County-Wise - Daily Data since January 2020 when the first case was found in Washington State)

Covid-19 Vaccination data:

https://www.kaggle.com/paultimothymooney/usa-covid19-vaccinations

Which derives its data from https://ourworldindata.org/us-states-vaccinations & which in turn derives its data from the CDC Website

(State Wise Daily Data starting from the point when CDC began to publish this starting - January 2021)

The Common Potholes

- 65 distinct locations exist for Vaccination Data, 55 "States" for Infection data! *(an example being Bureau of Prisons)*

- Some values of infection data (cases/deaths) was negative!

- Vaccinations actually began before being recorded/reported on a daily basis by the CDC.

- Multiple days had complete vaccination data missing!

Let us dive into the data!

- Cumulative data does not easily show the rise and fall of the infections/vaccinations (we have seen this in class already!)

- Daily Data - has a lot of noise/variations

- 7 Day Moving Average - smoothes the curve and is also widely used/accepted

Total Daily Infections Reported in the U.S. as of 2021-05-27
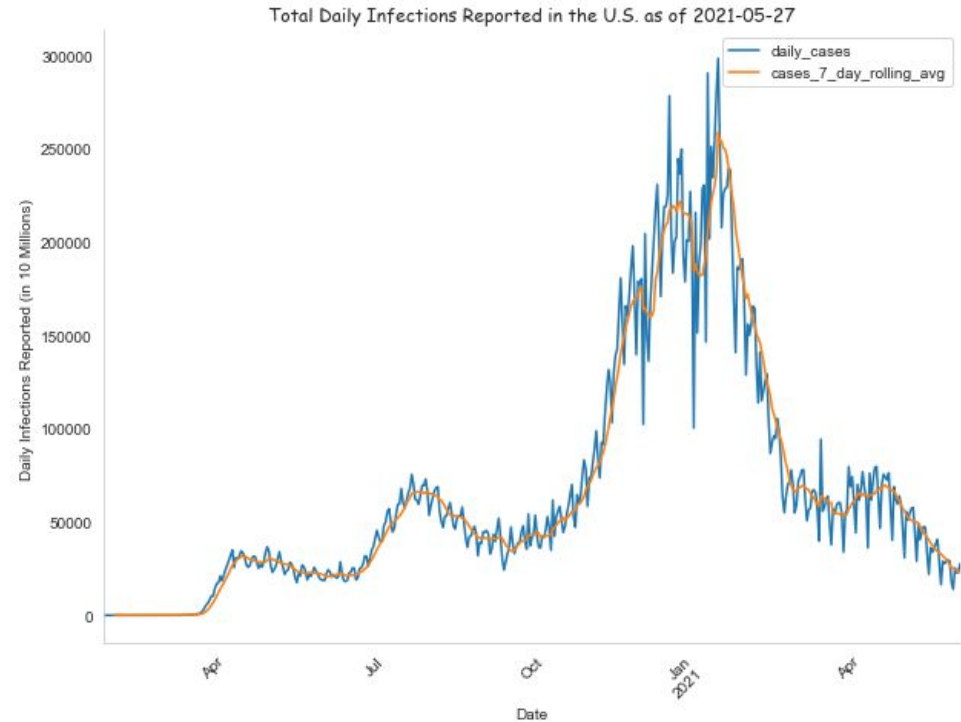
The Rise & Fall of the Infections

- First Infection - January 2020 - Washington State

- We have had multiple waves of the infection - the worst being the one of December - January

- Since then, there was a sharp decline in the cases

- We are still seeing a national average of 22-25 thousand cases every day

Rolling Average of Infections Reported in the United States as of 2021-05-27 (with markers showing the peaks) and the Max Peak was recorded as 2021-01-08
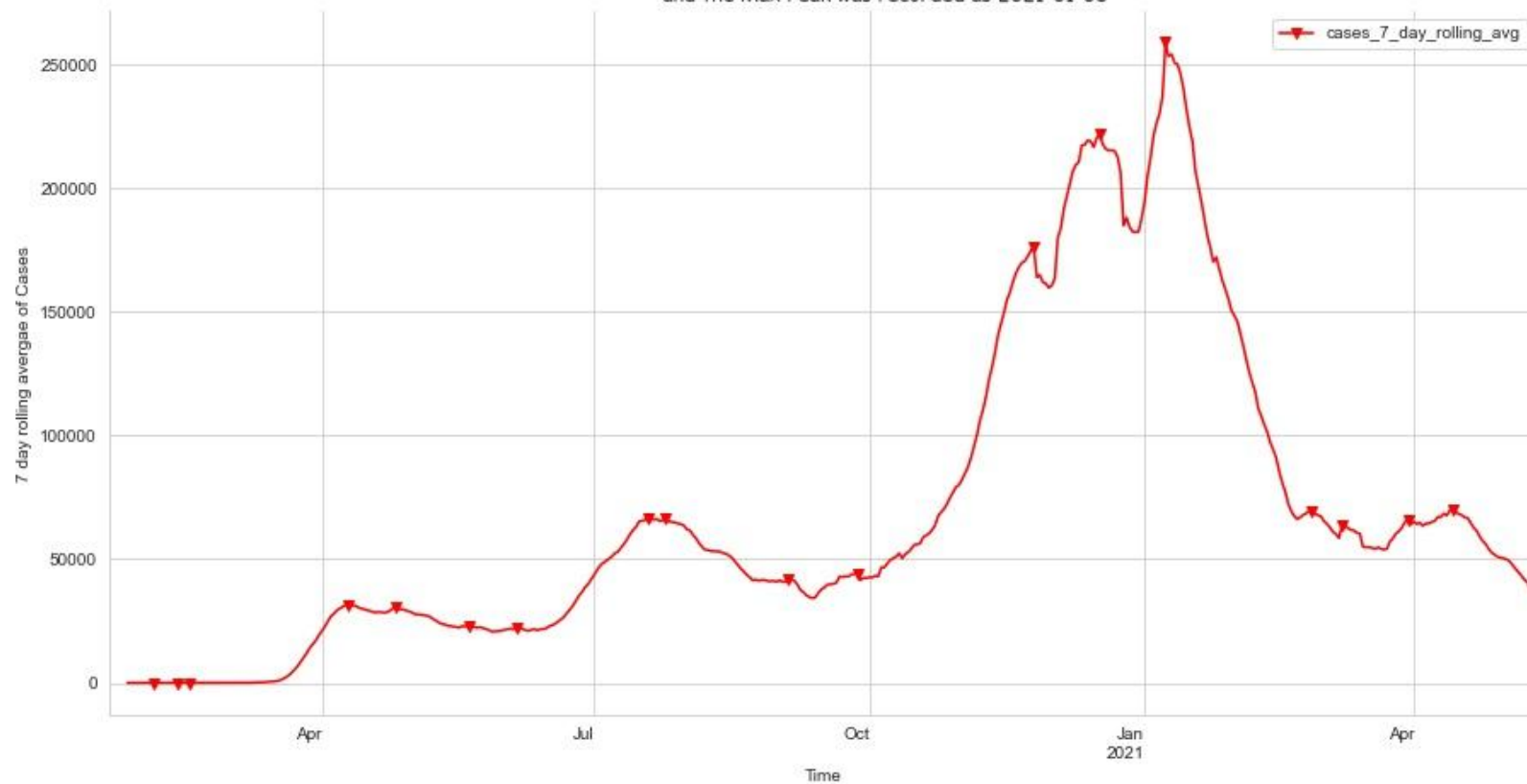
Rolling Average of Deaths Reported in the United States as of 2021-05-27 (with markers showing the peaks)
and the Max Peak was recorded as 2021-01-12
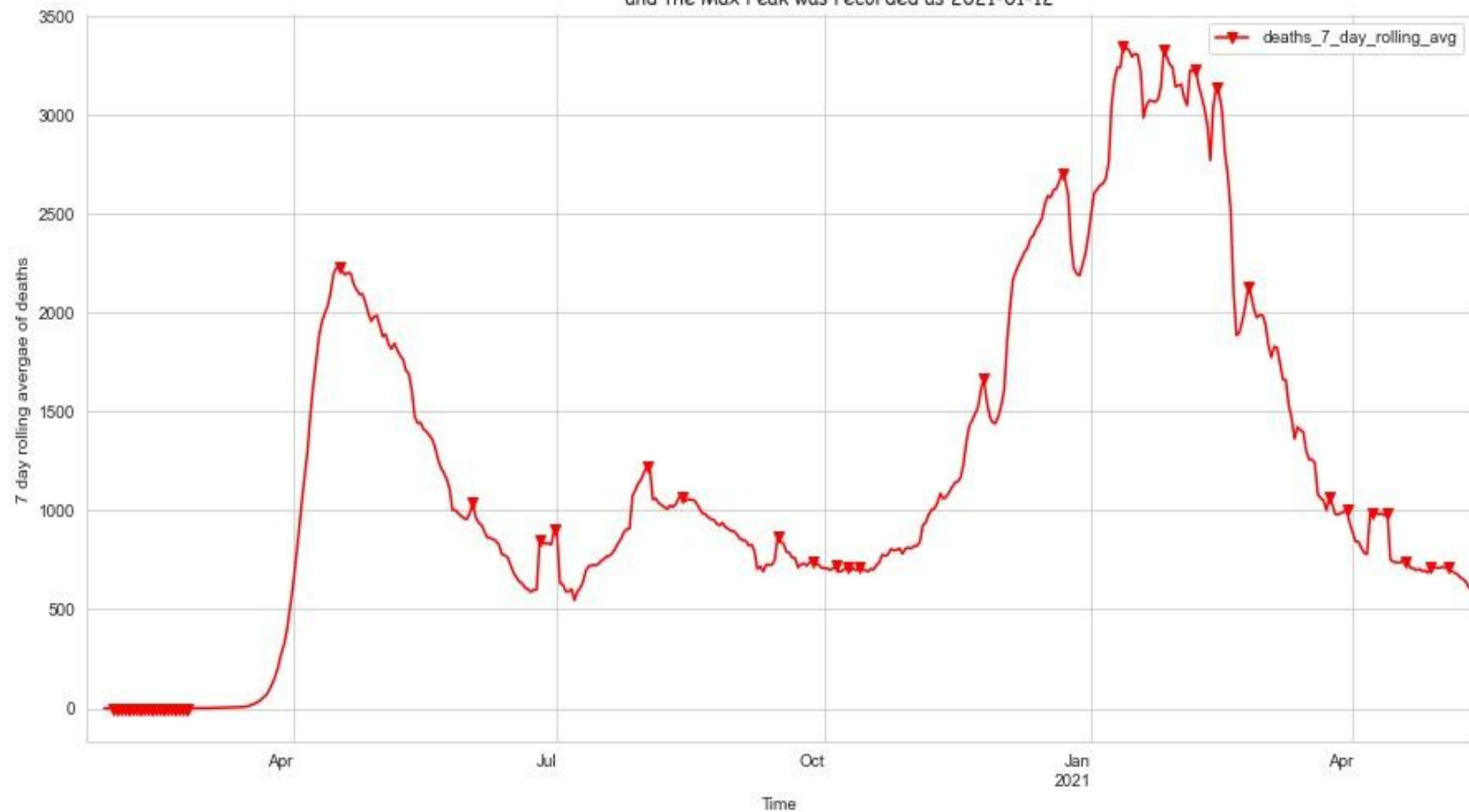
The Vaccination Drive

- First State Wise Vaccination data was recorded on 13th January, 2021 (actually started sometime in Dec 2020)

- We have seen a considerable decrease in the infections since then.

- More than 50% of the total U.S population has received at least 1 dose of the vaccine.

Percentage Covid-19 Vaccinated in USA as of 2021-05-30
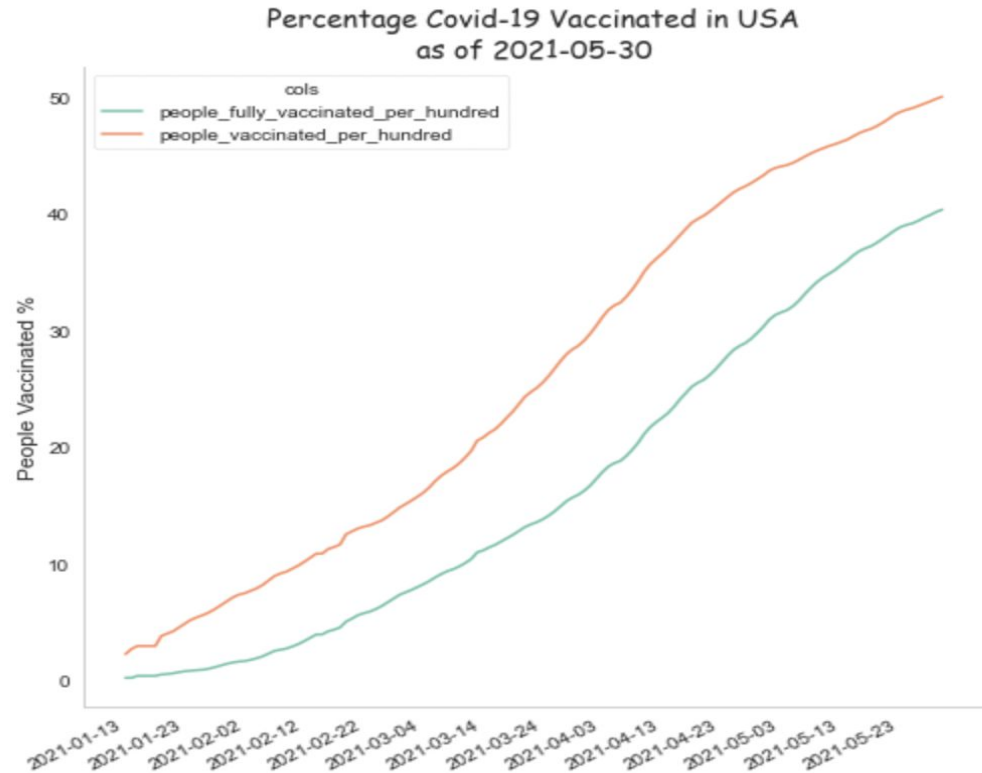
PDF Template Fill

- The aim is to fill in Templated PDFs using Python code

- The same is achieved using:
  - PDFTK (https://www.pdflabs.com/tools/pdftk-the-pdf-toolkit/)
  - pypdftk (python package, which is a wrapper around the PDFTK binary)
  - A few lines of code

- Process:
  - Install the PDFTK-Server (Command line tool, works on installed machine)
  - Ensure PDFTK is available on the PATH variable!
  - pip install pypdftk
  - Simple 2 lines of code will ensure PDF is filled in!

PDF Template before and after

Alabama

COVID-19 INFECTION DATA

TOTAL NUMBER OF PEOPLE INFECTED BY COVID-19:	
PERCENTAGE OF PEOPLE INFECTED BY COVID-19:	
TOTAL NUMBER OF DEATHS DUE TO COVID-19:	
PERCENTAGE OF DEATHS DUE TO COVID-19:	

COVID-19 VACCINATION DATA

TOTAL NUMBER OF PEOPLE VACCINATED TILL DATE:	
PERCENTAGE OF TOTAL POPULATION VACCINATED:	
TOTAL NUMBER OF PEOPLE FULLY VACCINATED TILL DATE:	
PERCENTAGE OF TOTAL FULLY POPULATION VACCINATED:	

COVID-19 INFECTION DATA

TOTAL NUMBER OF PEOPLE INFECTED BY COVID-19:	543177
PERCENTAGE OF PEOPLE INFECTED BY COVID-19:	11.08%
TOTAL NUMBER OF DEATHS DUE TO COVID-19:	11140.0
PERCENTAGE OF DEATHS DUE TO COVID-19:	0.23%

COVID-19 VACCINATION DATA

TOTAL NUMBER OF PEOPLE VACCINATED TILL DATE:	3027567.0
PERCENTAGE OF TOTAL POPULATION VACCINATED:	36.02%
TOTAL NUMBER OF PEOPLE FULLY VACCINATED TILL DATE:	1432446.0
PERCENTAGE OF TOTAL FULLY POPULATION VACCINATED:	29.21%

Covid-19 cases in Alabama

Rolling Average of Infections Reported in the Alabama as of 2021-05-27 (with markers showing the peaks)
and the Max Peak was recorded as 2021-01-10

Deaths in Alabama Due to Covid-19

Rolling Average of Deaths Reported in the Alabama as of 2021-05-27 (with markers showing the peaks)
and the Max Peak was recorded as 2021-01-29
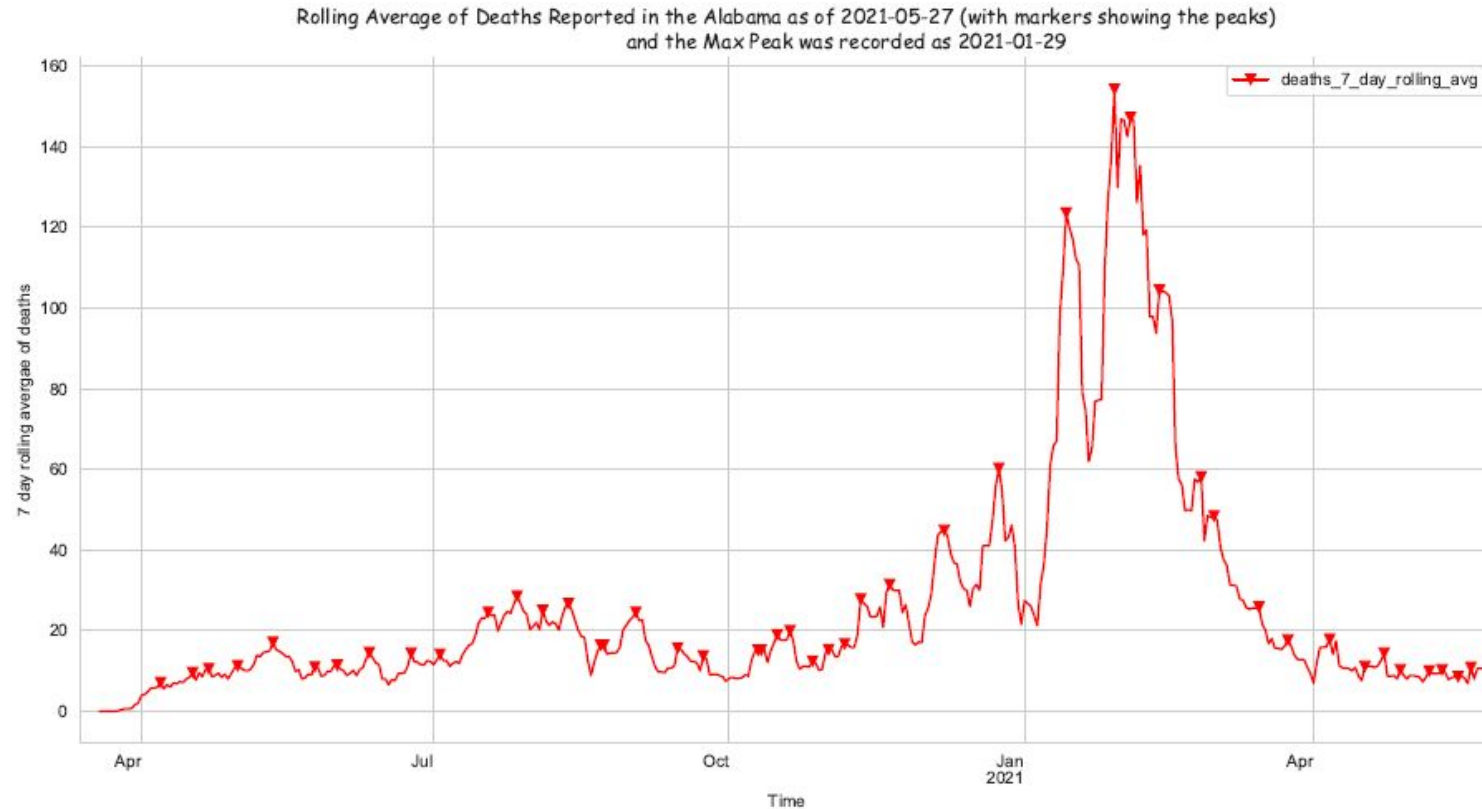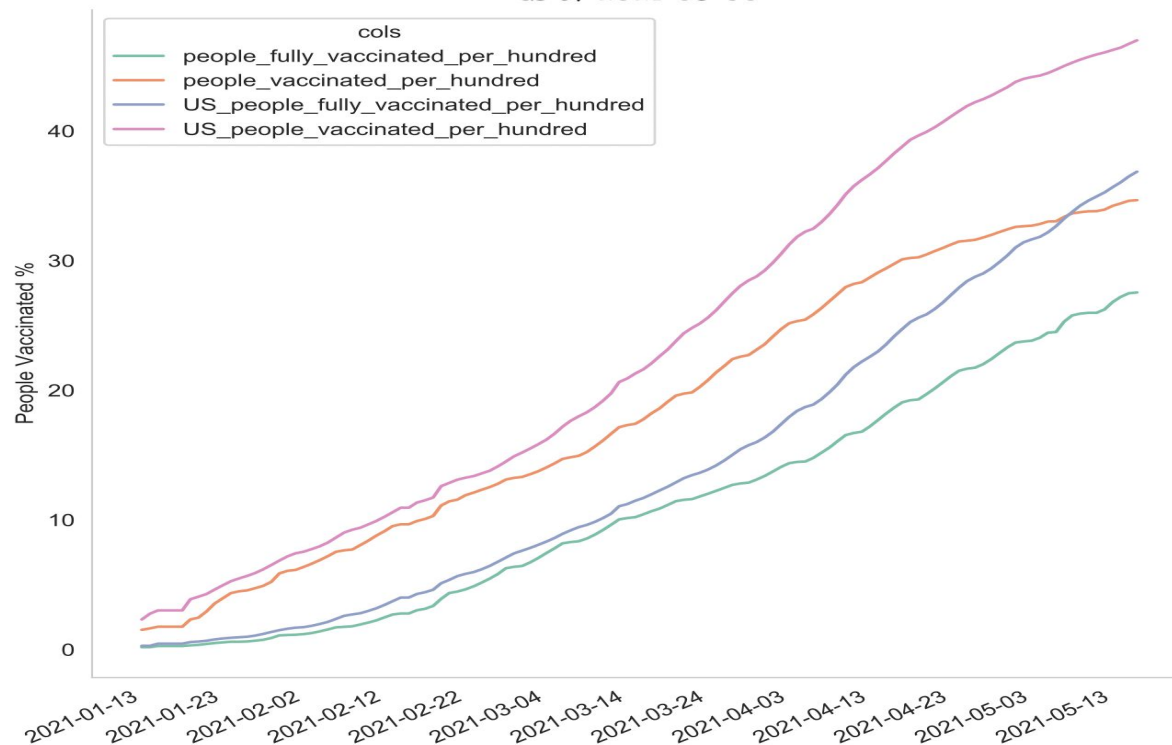
Vaccination progress in Alabama

Percentage Covid-19 Vaccinated in Alabama
v/s Vaccinations across all of USA
as of 2021-05-30

cols
— people_fully_vaccinated_per_hundred
— people_vaccinated_per_hundred
— US_people_fully_vaccinated_per_hundred
— US_people_vaccinated_per_hundred

Which States are performing the Best/Worst?

Top 5 and Worst 5 performing states
WRT Vaccinations as of 2021-05-30
With USA National Numbers to compare
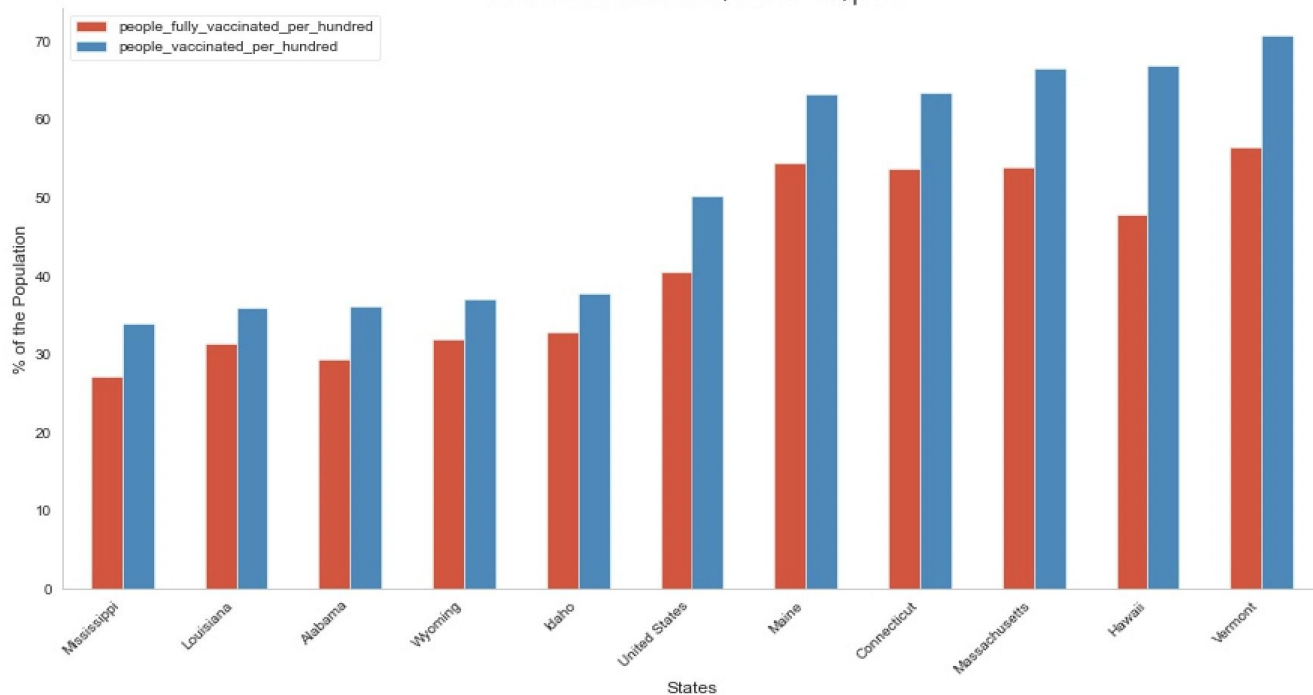
Let us take a quick look at the PDF Generation

(Demo)

Regression Analysis

We are trying to find the correlation for the no of deaths reported each day.

As with the analysis, we are again using the 7-day-rolling-average - to be able to get a clearer picture.

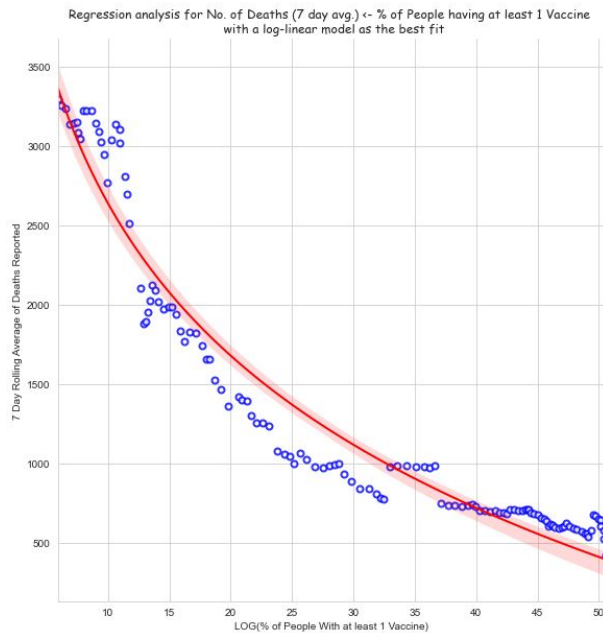We are looking at the following dependent variables:

- Vaccination % (at least 1 dose)
- Vaccination % (fully vaccinated)
- No. of cases reported in the last 14 days (prior to current date)
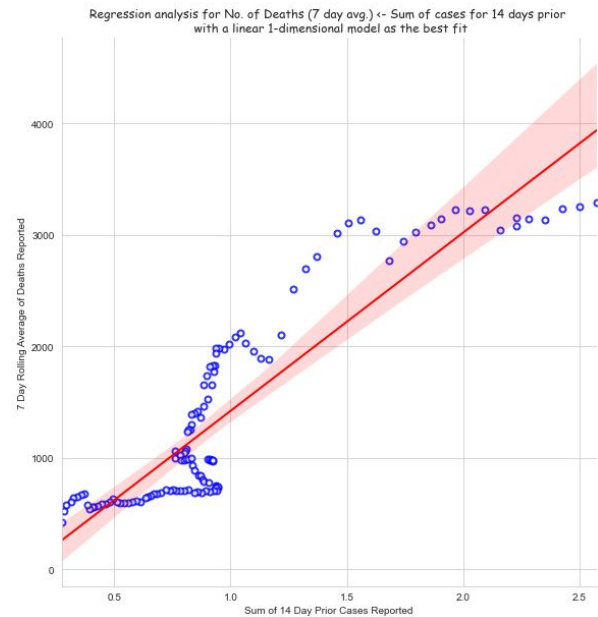 [Because there is always a lead time from infection to death]

Regression Data

Deaths <- Log(% of People Fully Vaccinated)
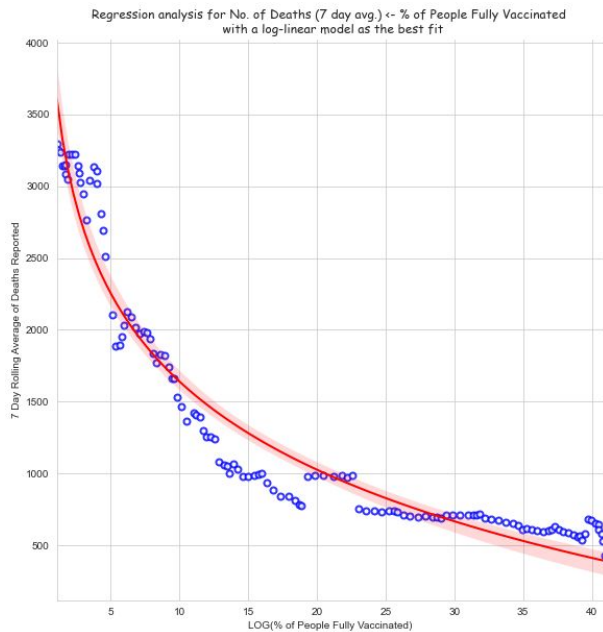
Deaths <- % of At least 1 Vaccination

Deaths <- Last 14 days total cases reported

Regression analysis for No. of Deaths (7 day avg.) <- % of People Fully Vaccinated with a log-linear model as the best fit

Regression analysis for No. of Deaths (7 day avg.) <- % of People having at least 1 Vaccine with a log-linear model as the best fit

Regression analysis for No. of Deaths (7 day avg.) <- Sum of cases for 14 days prior with a linear 1-dimensional model as the best fit

Regression Parameters

We have 2 models that explain the data:

1. deaths_7_day_rolling_avg = -70.61 * log(fully_vaccinated_per_hundred) + 1576.83 * 14_day_rolling_sum (rolling sum of cases for previous 14 days, in Millions)
2. deaths_7_day_rolling_avg = 4753.56 - 1149.9 * log(vaccinated_per_hundred) + 329.62 * 14_day_rolling_sum (rolling sum of cases for previous 14 days, in Millions)

The R^2 score in both cases is above 0.94 The P-Value is below 0.05

So there is a strong correlation between the vaccination drive & the drop in deaths seen over time.

OLS Regression Results

Dep. Variable:	deaths_7_day_rolling_avg	R-squared:	0.954
Model:	OLS	Adj. R-squared:	0.953
Method:	Least Squares	F-statistic:	1301.
Date:	Fri, 04 Jun 2021	Prob (F-statistic):	2.17e-84
Time:	21:32:03	Log-Likelihood:	-854.70
No. Observations:	128	AIC:	1715.
Df Residuals:	125	BIC:	1724.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	4753.5595	250.097	19.007	0.000	4258.586	5248.533
log_vaccinated_per_hundred	-1149.8999	57.658	-19.943	0.000	-1264.013	-1035.787
14_day_rolling	329.6189	72.220	4.564	0.000	186.686	472.552

Thank you!

Questions?