UOW MALAYSIA KDU

PART OF THE UNIVERSITY OF WOLLONGONG AUSTRALIA GLOBAL NETWORK

SCHOOL OF COMPUTING & CREATIVE MEDIA

Please tick ✓ or click if using MS WORD

☐FOUNDATION ☐DIPLOMA ☒ DEGREE ☐ MASTER

Assignment Coversheet

Please complete all details required clearly. For softcopy submissions, please ensure this cover sheet is included at the start of your document or in the file folder.

Assignment & Course Details:

Subject Code: *(e.g. XCAT1234)* XBDS 2014		**Subject Name** *(e.g. Fundamentals of Computing)***:** Introduction to Data Science	
Course *(e.g. Bachelor in Computing)* : Bachelor of Computer Science			
Lecturer Name: Dr. Law Foong Li			
Assessment Due Date: *(dd/mm/yy)*	09/08/2021	**Assessment Title:**	Group Assignment

I/We declare that:

- *This assignment is my/our own original work, except where I/we have appropriately cited the original source.*
- *This assignment or parts of it has not previously been submitted for assessment in this or any other subject.*
- *I/We allow the assessor of this assignment to test any work submitted by me/us, using text comparison software for plagiarism.* ***(For more information, Please read the Academic Integrity Guidelines)***

Name : Thean Jun Chao **Student ID:** 0127122 **Email :** 0127122@kdu-online.com **Mobile No:** 0173479830 **Signature:** Chao. **Date:** 08/08/2021	**Name :** Taashwin Reddy A/L Ramamoorthy **Student ID:** 0128664 **Email :** 0128664@kdu-online.com **Mobile No:** 0172632461 **Signature:** **Date:** 08/08/2021	**Name :** Bernard Lee Hanlin **Student ID:** 0131545 **Email :** 0131545 @kdu-online.com **Mobile No:** 0163477019 **Signature:** **Date:** 08/08/2021
Name : **Student ID:** **Email :** **Mobile No:** **Signature:** **Date:**	**Name :** **Student ID:** **Email :** **Mobile No:** **Signature:** **Date:**	**Name :** **Student ID:** **Email :** **Mobile No:** **Signature:** **Date:**

For office use only – Lecturer comments (if applicable)	**Marks Breakdown**	

Table of Contents

1 Introduction

1.1 Problem Statement

The covid-19 pandemic has adversely affected people's lives and livelihoods around the world. Economies worldwide are severely affected by the pandemic as companies have gone bankrupt or have to fire some employees to survive. Atradius, a Dutch insurance company predicted a 26% increase in bankruptcies globally. LegalJobs reported that as of September 2020, 470 companies had gone bankrupt, including NPC International Inc; the largest franchisee of PizzaHut restaurants. Besides, the healthcare sector is suffering tremendously as front liners are getting infected or killed. A shortage of oxygen tanks in some countries, causing more distraught among the people as people are dying at a much faster rate since newer mutated variants of the virus are more contagious and dangerous. Frontliners are overworked for more than a year to standby to aid infected patients. Ensuring vaccinations progress at a steady increasing rate worldwide is necessary as vaccines are the main solution to overcoming this pandemic. However, some people are hesitant to get vaccinated mainly due to the questions regarding the vaccine safety. A study conducted on the acceptance of the covid-19 vaccination in China found that despite having a 91.3% acceptance rate among 2058 participants from various provinces, only 52.2% wanted to get vaccinated as soon as possible while the other 47.8% would delay until the safety of the vaccines is confirmed (Wang et al., 2020). To persuade and educate the public to get vaccinated as soon as possible, strong evidence of the effectiveness of vaccination must be displayed to the. Therefore, scientific research shall be carried out on the effectiveness of the vaccination in reducing the positive cases and death cases due to covid-19, as well as when a nation can reach herd immunity.

1.2 Limitations of Existing Solutions

Several observational studies have been taken to assess the vaccine effectiveness by the Centers for Disease Control and Prevention (CDC, 2021).

i. Case-control studies

Two groups of participants are asked if they have been vaccinated. The two groups are the case group, referring to people infected with covid-19 and the control, referring to people who have not been infected with covid-19 (CDC, 2021).

ii. Cohort studies

Observing groups of vaccinated and unvaccinated people for some time to see if they get infected by the covid-19 virus later (CDC, 2021).

iii. Screening method assessments

Collecting the vaccination status among a group of people infected with the covid-19 virus (CDC, 2021).

iv. Ecologic analysis assessments

Collecting data from people from different locations or at different times to find a relationship between those vaccinated and those infected with covid-19 (CDC, 2021).

One of the limitations of observational studies is the long observational time and high cost is needed to conduct the study (WHO, 2021). The different situations in different locations and rapid vaccination rollout makes it harder to compare vaccination rates between vaccinated and unprotected locations (WHO, 2021). Recall bias can occur as the participants may have forgotten the vaccine brand and the dates of their vaccination (WHO, 2021). Also, the study can be further enhanced by predicting the future number of new cases, new deaths, and when herd immunity can be attained.

1.3 Proposed Solution and Approach

To convince more people to get vaccinated as soon as possible, we will use machine learning to train trusted and publicly available data to identify the trend of the vaccination against the new cases and new deaths. A decreasing trend is expected as the number of people who received their vaccination increases. Furthermore, we will also compare the effectiveness of partial vaccination (1 dose) against (2 doses) in terms of reducing the new cases and new deaths. Lastly, we would also allow the user to predict roughly when the United States can attain herd immunity.

1.4 Hypothesis & Research Question

This paper aims to study the following research questions:

- RQ1: How does one dose of vaccination compare with two doses of vaccination in the effectiveness of reducing the number of positive Covid-19 cases and the number of deaths due to the Covid-19 in the USA?
    - Research hypothesis: Two doses of vaccination will be more effective in reducing the number of positive Covid-19 cases and the number of deaths due to the Covid-19 in the USA.

- RQ2: When will the US attain herd immunity (70% vaccinated)?
    - Research hypothesis: The USA can attain herd immunity by winter 2021 (D’souza and Dowdy, 2021)

2 Methodology

2.1 Data Collection

The dataset used in this study was collected from a public source in Github. The data contained within the dataset is sourced from a variety of legitimate sources, including the official data collated by Our World in Data; the COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University; a peer study from Franciscco Arroyo-Marioli, Francisco Bullano, Simas Kucinskas, and Carlos Rondon-Moreno; Oxford COVID-19 Government Response Tracker, and International organizations including but not limited to UN, World Bank, OECD (Organisation for Economic Co-operation and Development) and IMHE (Institute for Health Metrics and Evaluation). The dataset includes 60 columns of data, each representing one variable of the dataset. The date range is between 2020-01-22 and 2021-07-31.

2.2 Programming Language

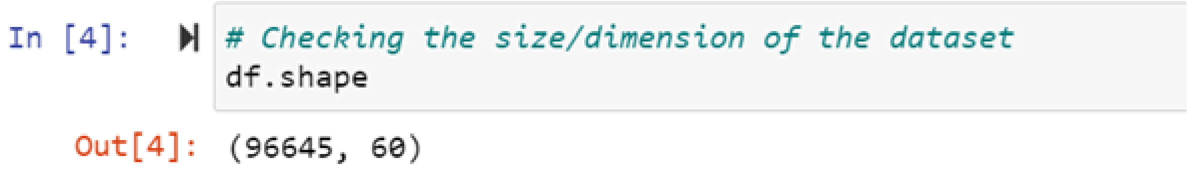
The study is conducted by implementing a project using python programming, a general-purpose language that is easy to learn as the high-level syntax that can be understood easily.

2.3 Data Understanding

The initial step taken to understand the dataset is by performing the exploratory data analysis (EDA). The EDA is conducted to determine the significance between variables; several libraries are imported into python. These libraries are:

- Pandas: Library for data manipulation and analysis
- Numpy: Support library for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.
- Matlplotlib: Plotting library that provides an object-oriented API for embedding plots into applications.
- Seaborn: Data visualization library built on top of matplotlib and closely integrated with pandas data structure in Python. Visualisation is the central part of seaborn which helps in the exploration and understanding of data.

As shown in figure 1, the dataset contains 96,654 rows and 60 columns.



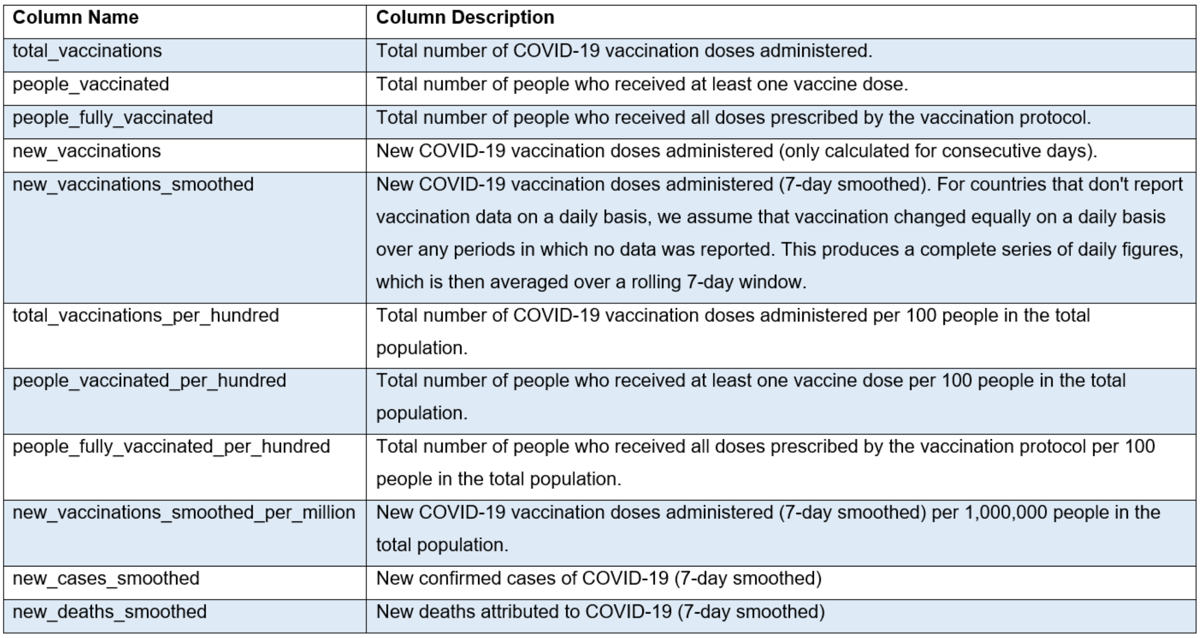
```
In [4]:   # Checking the size/dimension of the dataset
          df.shape

Out[4]:   (96645, 60)
```

Figure 1: The shape of the dataset

Eleven columns were selected to perform the initial EDA process. These columns are listed in table 1 below.

Table 1: Columns used to perform the initial EDA process.



Column Name	Column Description
total_vaccinations	Total number of COVID-19 vaccination doses administered.
people_vaccinated	Total number of people who received at least one vaccine dose.
people_fully_vaccinated	Total number of people who received all doses prescribed by the vaccination protocol.
new_vaccinations	New COVID-19 vaccination doses administered (only calculated for consecutive days).
new_vaccinations_smoothed	New COVID-19 vaccination doses administered (7-day smoothed). For countries that don't report vaccination data on a daily basis, we assume that vaccination changed equally on a daily basis over any periods in which no data was reported. This produces a complete series of daily figures, which is then averaged over a rolling 7-day window.
total_vaccinations_per_hundred	Total number of COVID-19 vaccination doses administered per 100 people in the total population.
people_vaccinated_per_hundred	Total number of people who received at least one vaccine dose per 100 people in the total population.
people_fully_vaccinated_per_hundred	Total number of people who received all doses prescribed by the vaccination protocol per 100 people in the total population.
new_vaccinations_smoothed_per_million	New COVID-19 vaccination doses administered (7-day smoothed) per 1,000,000 people in the total population.
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed)
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed)

Two python functions namely generate_heat_map and plot_scatter_plot is defined to perform the EDA. The scatter plots help to identify the trends of the data. Heatmap on the other hand displays the correlation of the target variable against the other variables to identify the relevant dependent variables for the independent variable. Figure 2 below is the heatmap generated.

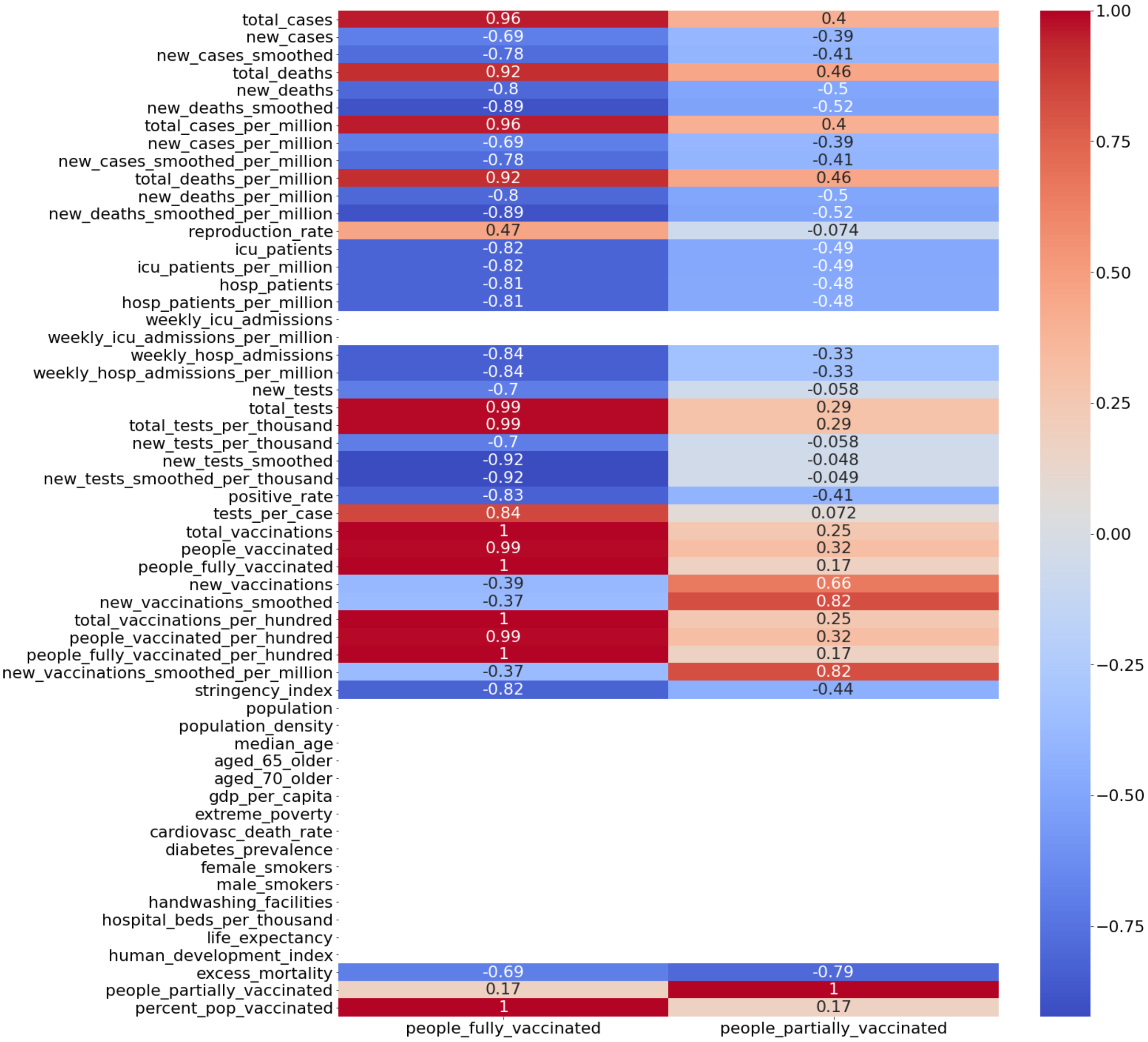


Figure 2: Heatmap

2.4 Data Preparation & Feature Engineering

The data preparation process involves cleaning and transforming the raw data into the desired format, and performing feature engineering for processing. Through EDA, a number of columns shown in table 2 below are to be used in this study.

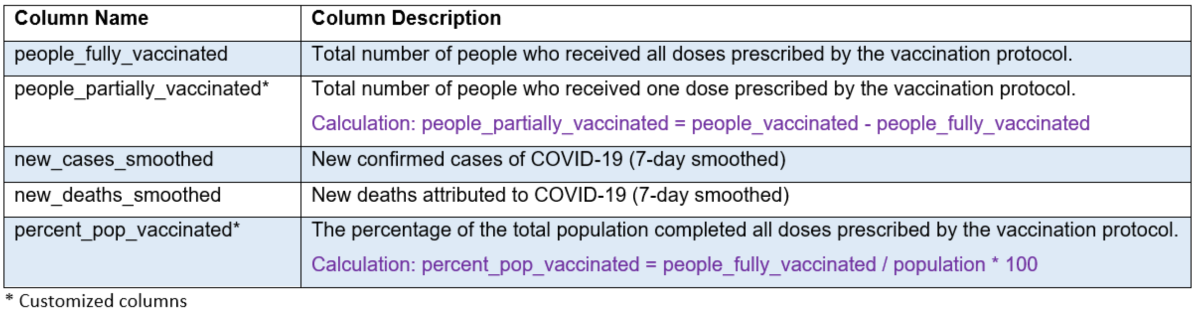
Table 2: Independent & dependent variables (columns)

RESEARCH QUESTION	*INDEPENDENT VARIABLE*	*DEPENDENT VARIABLE*
RQ1: How does one dose of vaccination compare with two doses of vaccination in the effectiveness of reducing the number of positive Covid-19 cases and the number of deaths due to the Covid-19 in the USA?	people_fully_vaccinated	new_cases_smoothed
		new_deaths_smoothed
	people_partially_vaccinated*	new_cases_smoothed
		new_deaths_smoothed
RQ2: When will the US attain herd immunity (70% vaccinated)?	date	percent_pop_vaccinated*

* Customized columns

Two customized columns have been created to satisfy the needs of our research questions namely people_partially_vaccinated and percent_pop_vaccinated. The calculations are included in the column description.

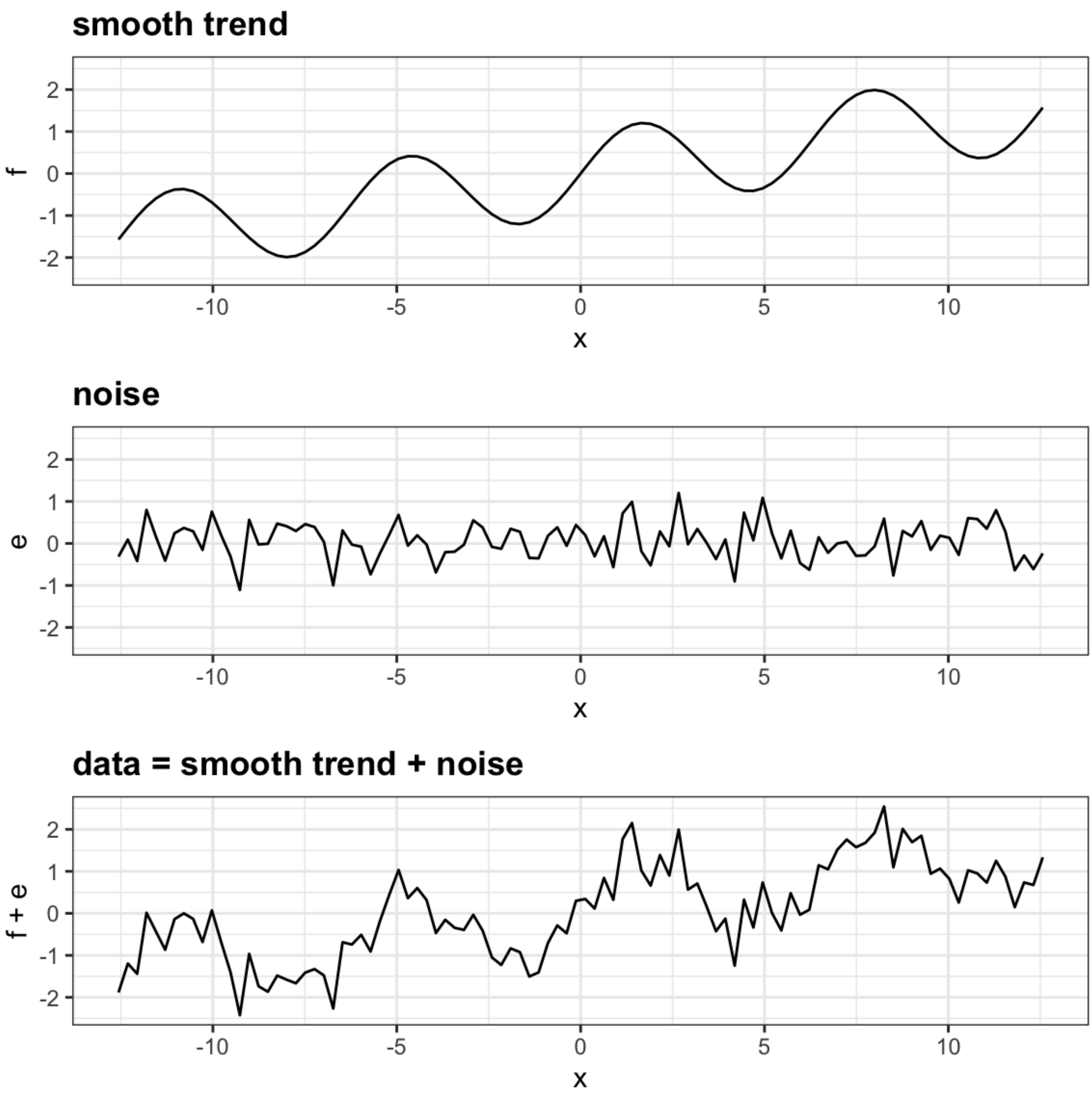
Table 3: Descriptions of the columns used



Column Name	Column Description
people_fully_vaccinated	Total number of people who received all doses prescribed by the vaccination protocol.
people_partially_vaccinated*	Total number of people who received one dose prescribed by the vaccination protocol. Calculation: people_partially_vaccinated = people_vaccinated - people_fully_vaccinated
new_cases_smoothed	New confirmed cases of COVID-19 (7-day smoothed)
new_deaths_smoothed	New deaths attributed to COVID-19 (7-day smoothed)
percent_pop_vaccinated*	The percentage of the total population completed all doses prescribed by the vaccination protocol. Calculation: percent_pop_vaccinated = people_fully_vaccinated / population * 100

* Customized columns

Furthermore, new_cases_smoothed and new_deaths_smoothed columns are used instead of new_cases and new_deaths. They are already present in the dataset. The smoothed data refers to the value obtained by averaging the values from the past 7 days. Smoothed data is used to eliminate the noises in the dataset so that the trends can be easily identified (Irizarry, 2021). The diagram below shows how the data becomes cleaner after removing the noises by smoothing.

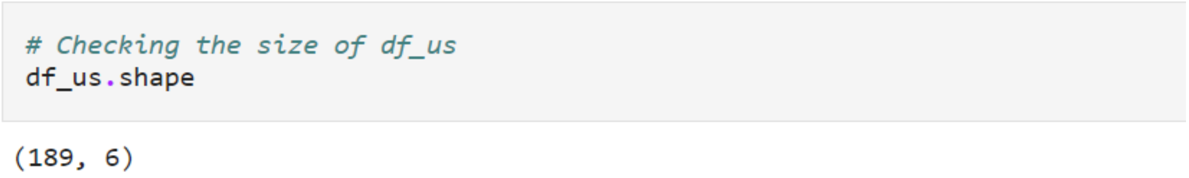

smooth trend

noise

data = smooth trend + noise

Figure 3: Data smoothing

The study will focus on the data of the United States of America as it is inappropriate to use every country in the world due to different population sizes and different situations in different countries. Figure 4 shows that the number of records available is 189 and 6 columns/features are kept for modelling. Figure 5 below shows a scatter plot of many trend lines due to the different population sizes of each country.



```
# Checking the size of df_us
df_us.shape
```

```
(189, 6)
```

Figure 4: The shape of df_us

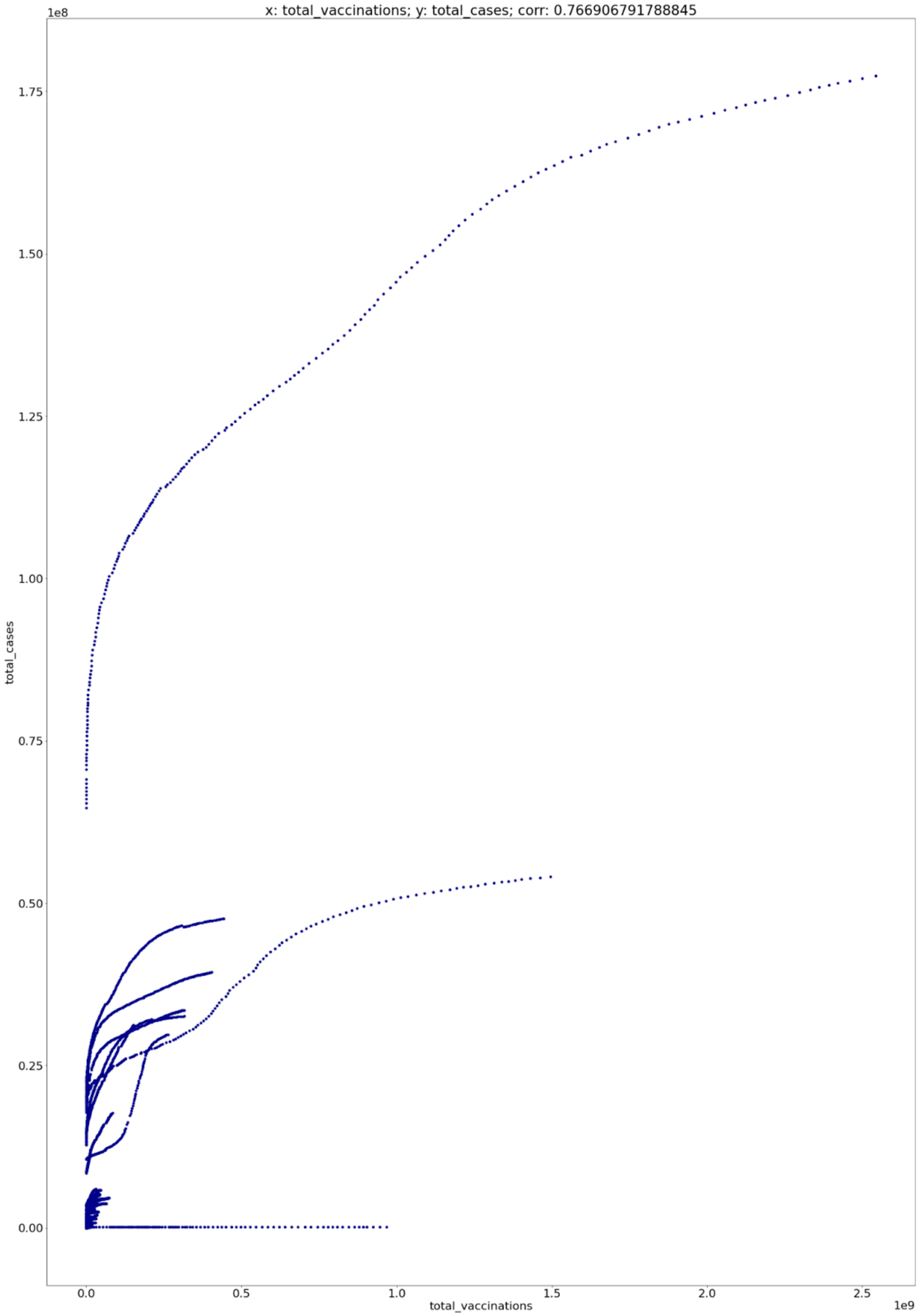


Figure 5: Sample scatter plot

Two functions are defined to perform the data cleaning process, namely “display_missing_val_dates” and “group_missing_val_by_month”. These are used to discover missing values hidden in the data. From the discovery, the missing values of the variables "people_fully_vaccinated", "people_partially_vaccinated", and "percent_pop_vaccinated" are from the beginning of January 2020 until 18 January 2021, totalling up to 365 each. It indicates that the vaccination only starts taking place in early 2021. Therefore, all data before the date 2021-01-18 have to be discarded. Besides, four other missing values were found after the date 2021-01-18. . Since there are only four missing data, case deletion to handle the missing data whereby those columns are eliminated from the dataset as it is the easiest way and deleting a few data will not affect the overall performance of the model. Both "new_cases_smoothed" and "new_deaths_smoothed" only has 6 missing values, which is the first six days of the record that have been deleted earlier. Therefore, the actual data range used is between 2021-01-18 and 2021-07-31.

2.5 Modelling

Machine learning is used in modelling. The data have been randomly split into a train set and a test set in the 7:3 ratio. The machine learning algorithms used for modelling are polynomial regression and linear regression. The polynomial regression of power 4 (quartic) is used in RQ1 as the graphs have non-linear trends while linear regression is used in RQ2 as the graph presents a positive linear relationship between the date and the percentage of people vaccinated. Table 4 below summarizes the models used to investigate each research question and the machine learning algorithm used.

Table 4: Machine learning algorithm used

Research Question	**Model**	**Machine Learning Algorithm**
RQ1	Model 1: people fully vaccinated vs new cases smoothed	Polynomial regression
	Model 2: people fully vaccinated vs new deaths smoothed	
	Model 3: people partially vaccinated vs new cases smoothed	
	Model 4: people partially vaccinated vs new deaths smoothed	
RQ2	Model 5: Date vs Percentage of People Fully Vaccinated	Linear Regression

The following equations are used in modelling:

RQ1: $y \ = \ ax^4 + bx^3 + cx^2 + dx + e$

RQ2: $y \ = \ mx + c$

2.6 Platform

Jupyter Notebook is used from the data understanding process right up to the model evaluation process. It’s a free open-source platform that supports Python and helps document the entire data science process. Jupyter allows for codes to be run block by block which makes it easier to highlight possible bugs or errors that might be present.

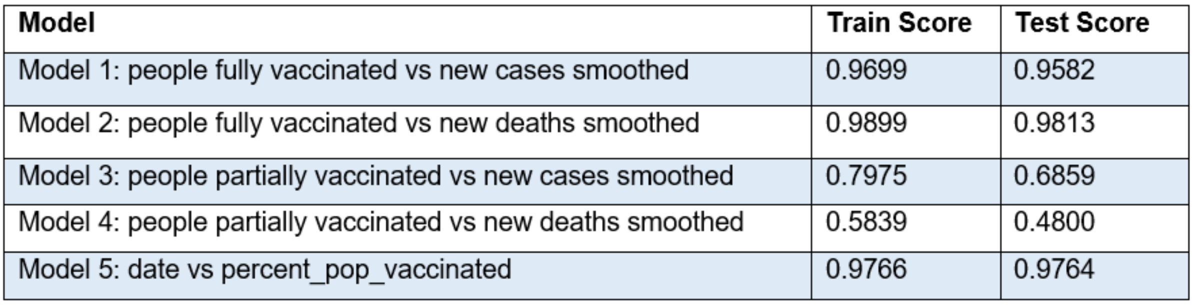
The spyder platform is used to develop the user interface as it supports the Streamlit library. Streamlit is an open-source application framework made to develop machine learning and data science applications using Python. It allows for a seamless transition of the inputted script and converts it into shareable and visually appealing web applications, making it easier for those who are not accustomed to data science to view the data patterns or trends through visualizations to better understand the data and make better decisions.

3 Results

3.1 Evaluation

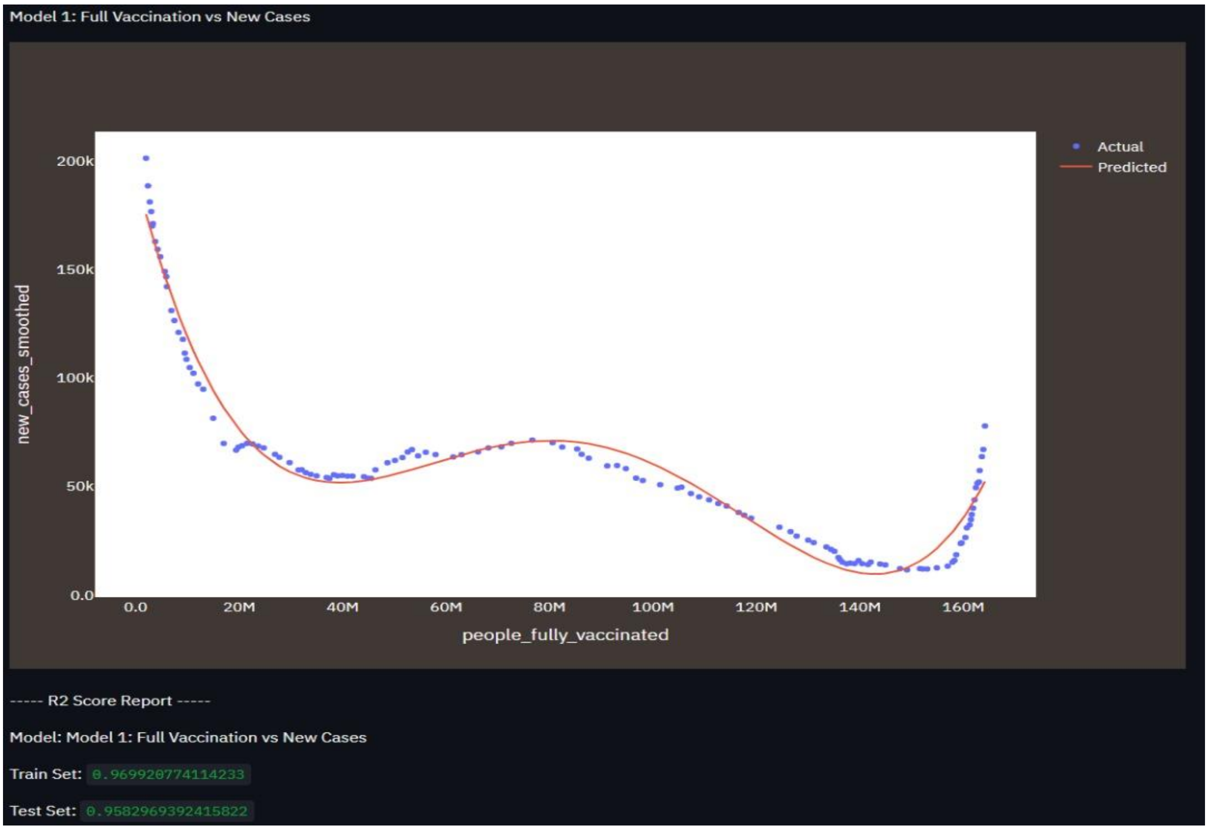
Model 1 graphs the number of people fully vaccinated against the new positive cases. It has obtained an accuracy of 0.9699 and 0.9582 on the training set and testing set respectively. Model 2 graphs the number of people fully vaccinated against the new death cases. It has obtained an accuracy of 0.9899 and 0.9813 on the training set and testing set respectively. Model 3 graphs the number of people partially vaccinated against the new positive cases. It has obtained an accuracy of 0.7975 and 0.6859 on the training set and testing set respectively. Model 4 graphs the number of people partially vaccinated against the new deaths cases. It has obtained an accuracy of 0.5839 and 0.4800 on the training set and testing set respectively. Model 5 graphs the date against the percentage of the population who received 2 doses of vaccinations. It has obtained an accuracy of 0.9766 and 0.9764 on the training set and testing set respectively. Table 5 below summarizes the accuracy scores for all models.

Table 5: Model accuracy scores



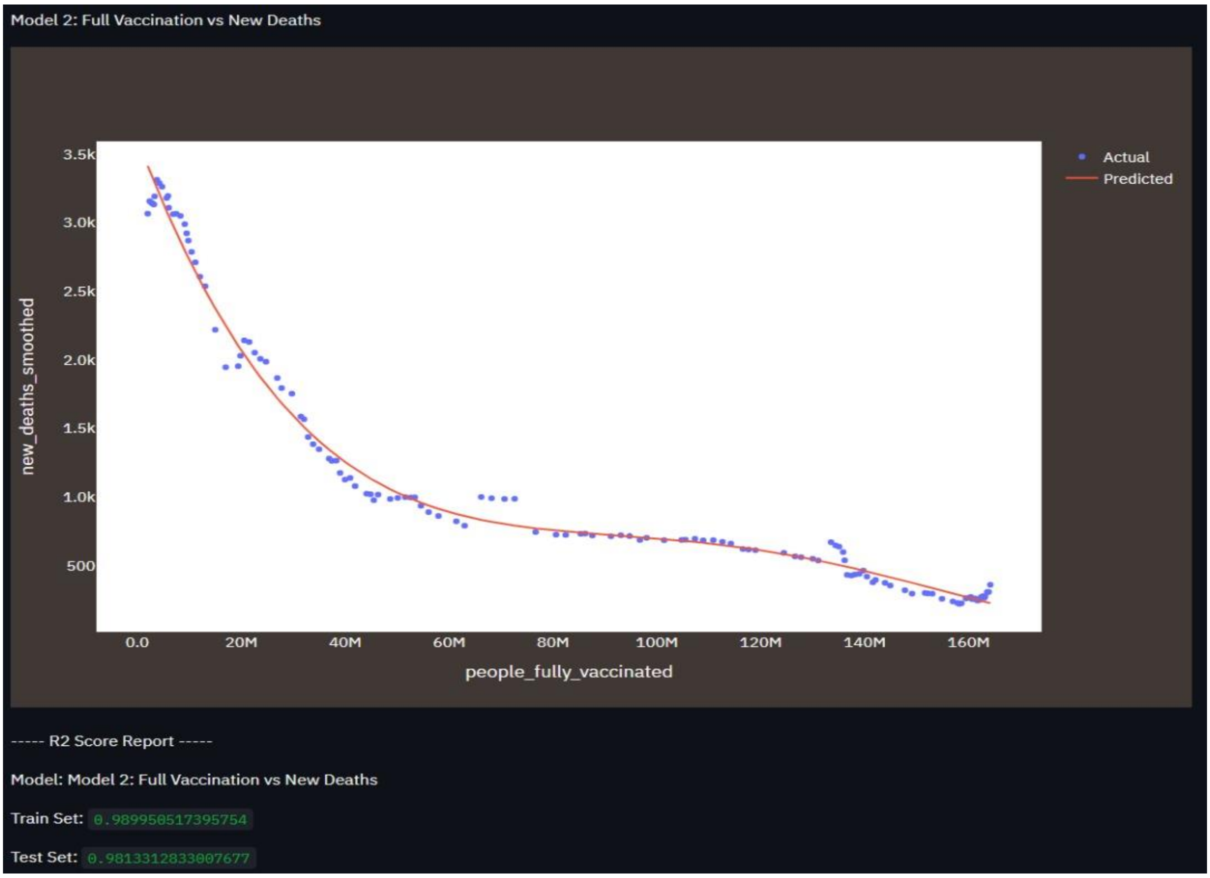
Model	Train Score	Test Score
Model 1: people fully vaccinated vs new cases smoothed	0.9699	0.9582
Model 2: people fully vaccinated vs new deaths smoothed	0.9899	0.9813
Model 3: people partially vaccinated vs new cases smoothed	0.7975	0.6859
Model 4: people partially vaccinated vs new deaths smoothed	0.5839	0.4800
Model 5: date vs percent_pop_vaccinated	0.9766	0.9764

The figures below show models 1 to 5. The scatter plots are the actual values while the red predicted line is the predicted regression line.



```
----- R2 Score Report -----
Model: Model 1: Full Vaccination vs New Cases
Train Set: 0.969920774114233
Test Set: 0.9582969392415822
```

Figure 6: Model 1 scatter plot and regression curve



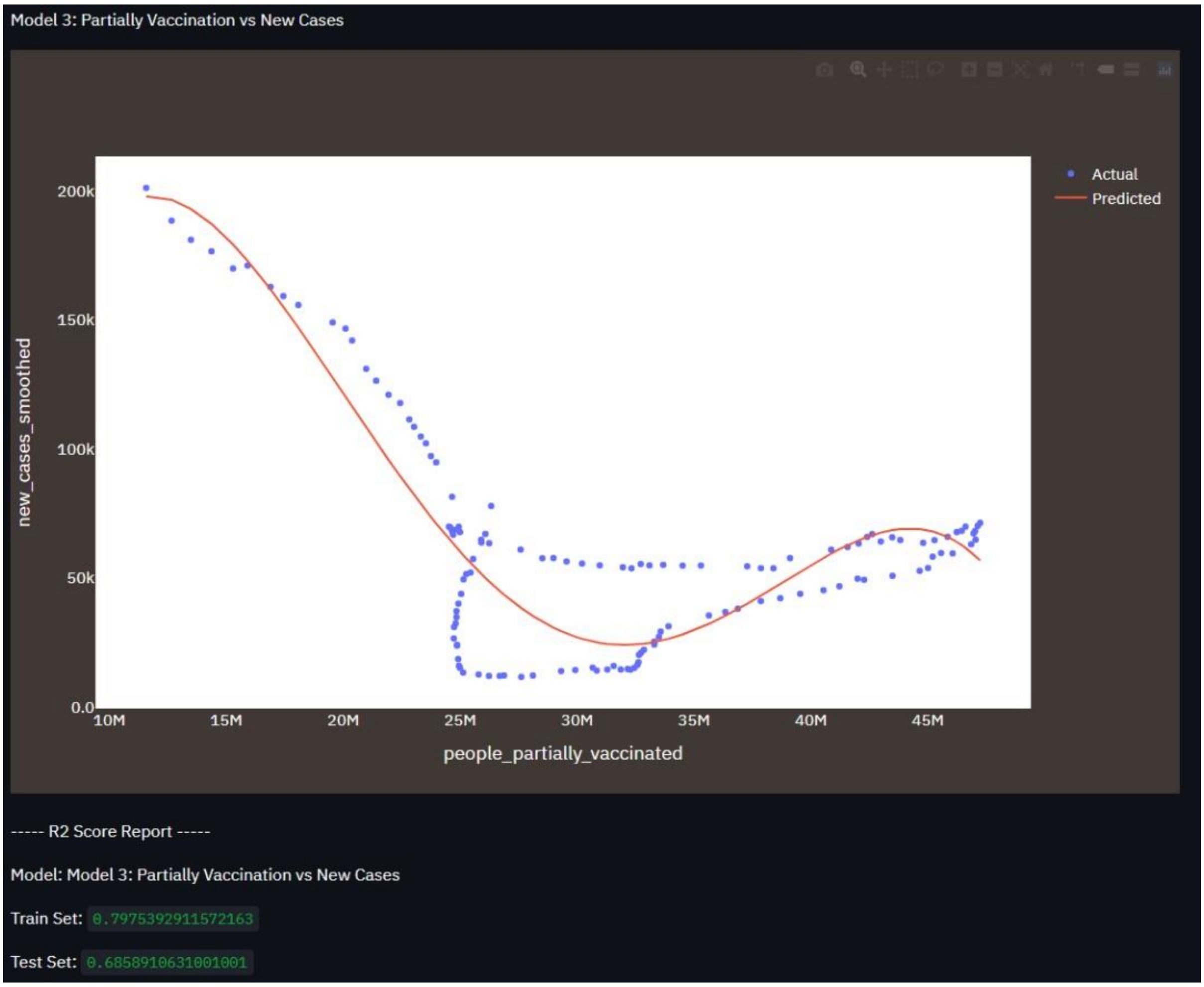
```
----- R2 Score Report -----

Model: Model 2: Full Vaccination vs New Deaths

Train Set: 0.989950517395754

Test Set: 0.9813312833007677
```

Figure 7: Model 2 scatter plot and regression curve



Model 3: Partially Vaccination vs New Cases

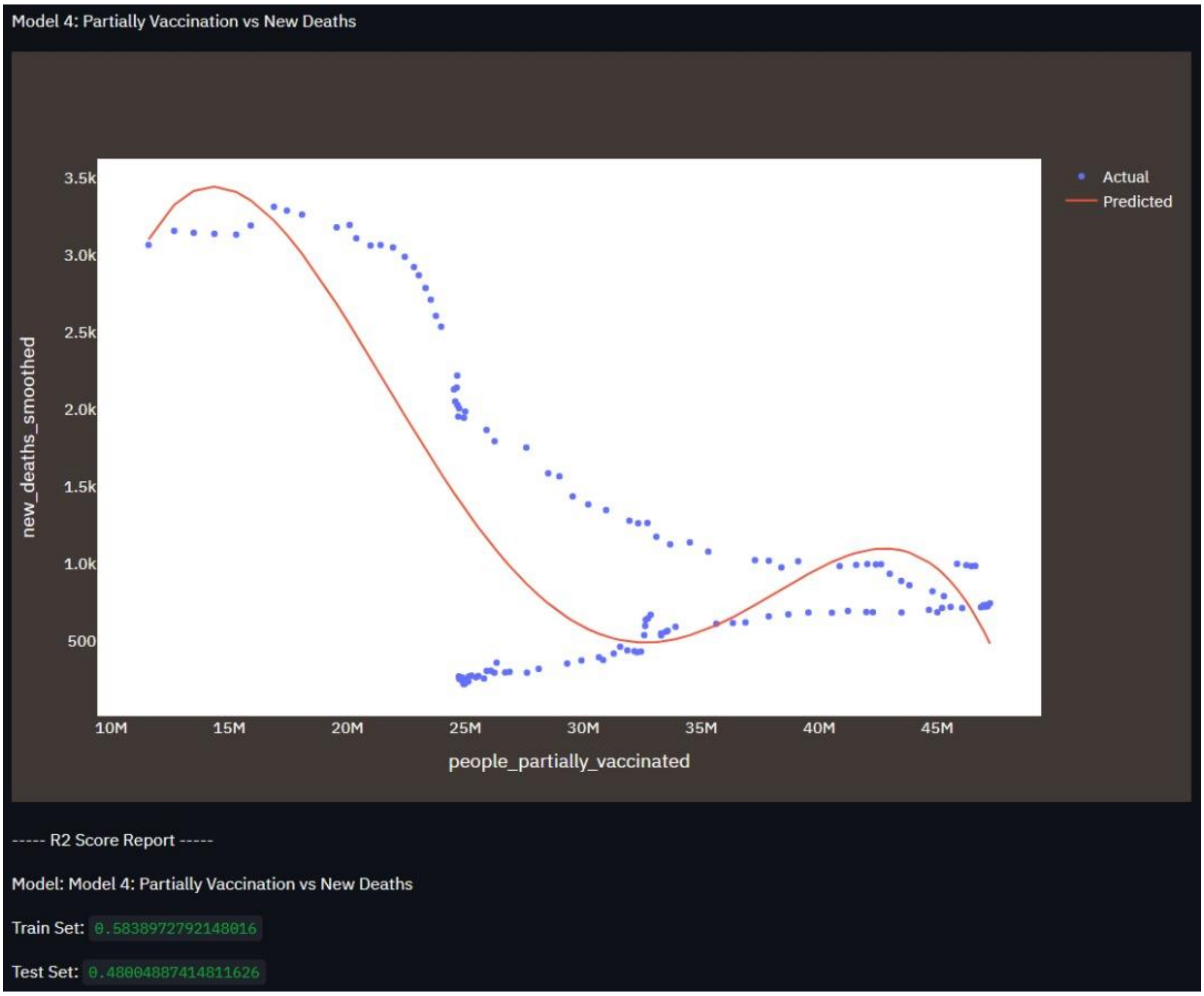
```
----- R2 Score Report -----

Model: Model 3: Partially Vaccination vs New Cases

Train Set: 0.7975392911572163

Test Set: 0.6858910631001001
```

Figure 8: Model 3 scatter plot and regression curve



Model 4: Partially Vaccination vs New Deaths

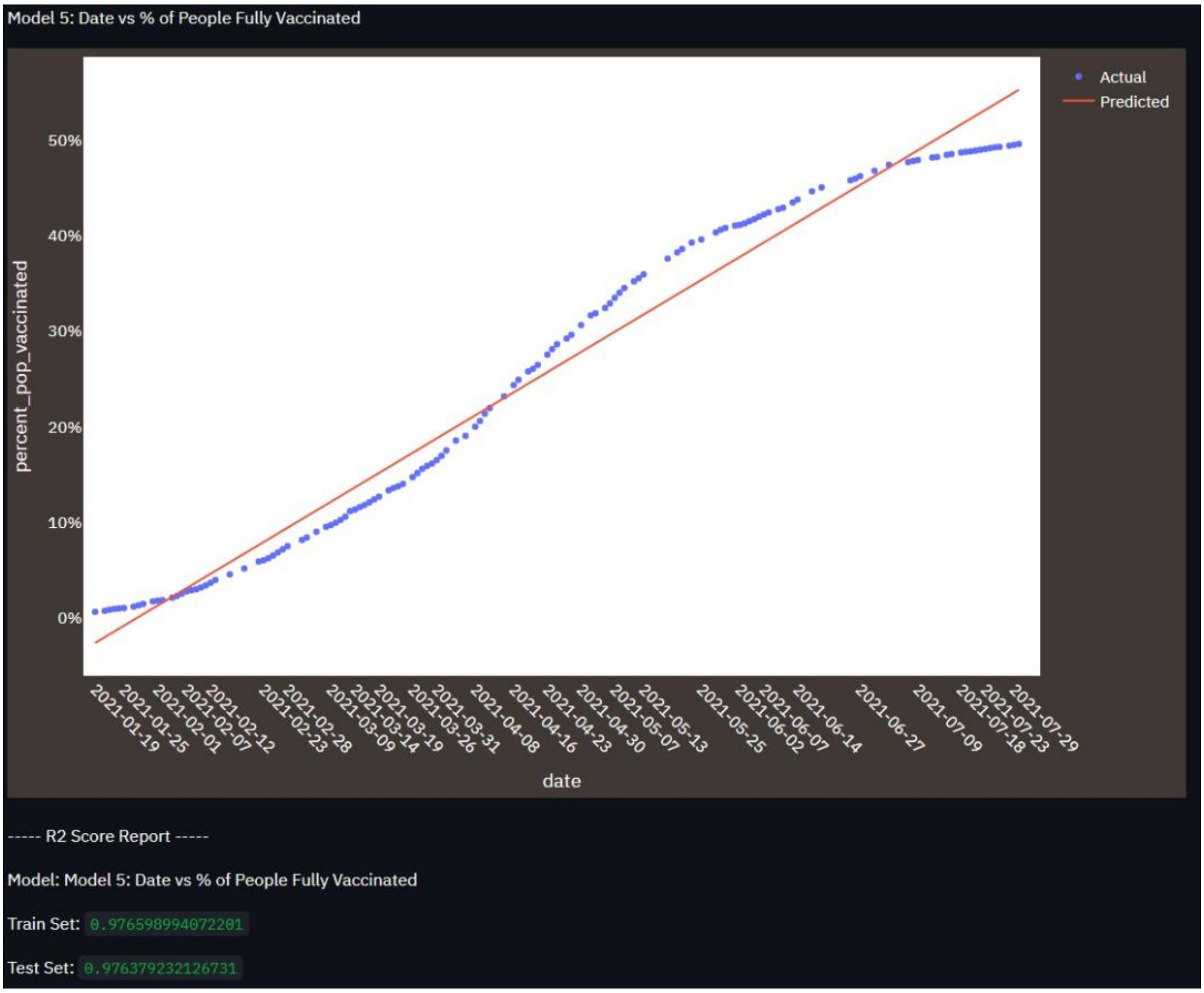
```
----- R2 Score Report -----

Model: Model 4: Partially Vaccination vs New Deaths

Train Set: 0.5838972792148016

Test Set: 0.48004887414811626
```

Figure 9: Model 4 scatter plot and regression curve



```
----- R2 Score Report -----

Model: Model 5: Date vs % of People Fully Vaccinated

Train Set: 0.976598994072201

Test Set: 0.976379232126731
```

Figure 10: Model 5 scatter plot and regression curve

The figures below show comparisons between models 1 and 3 as well as models 2 and 4.

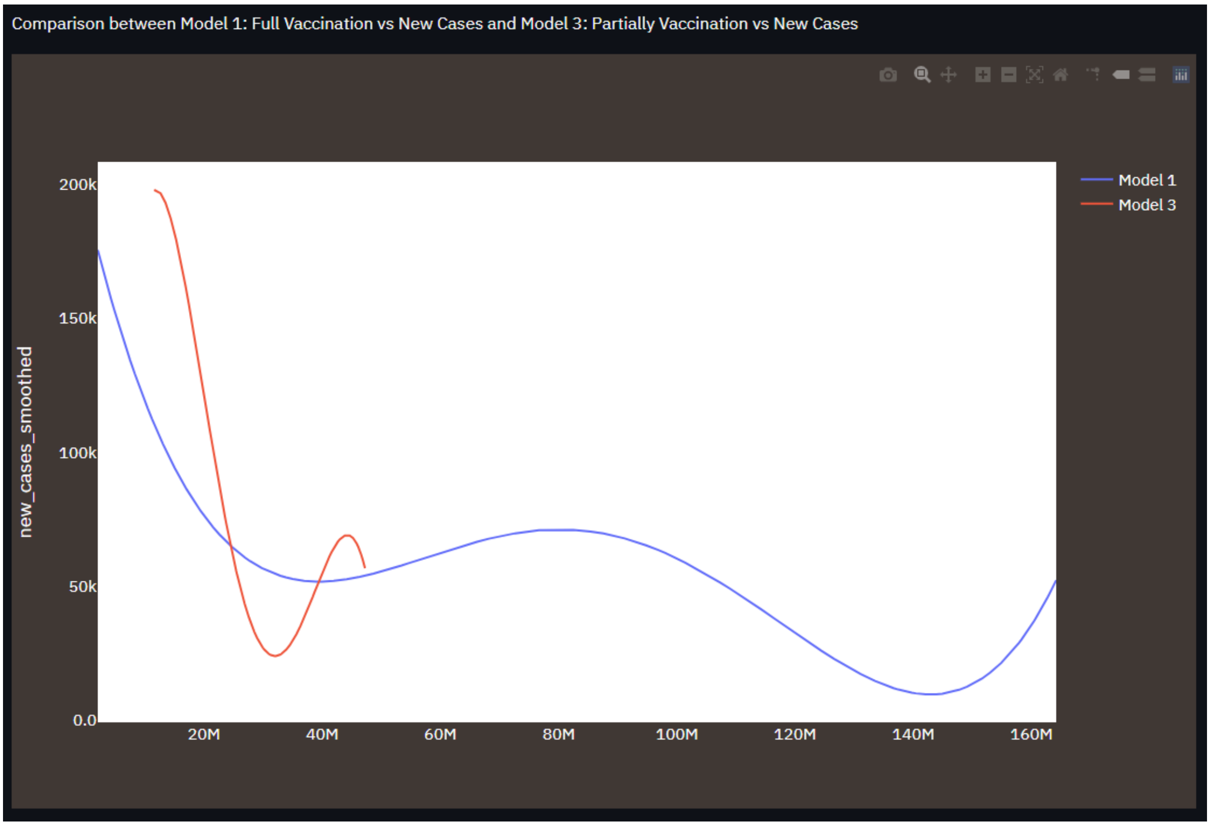


Figure 11: Comparison graph between model 1 and model 3

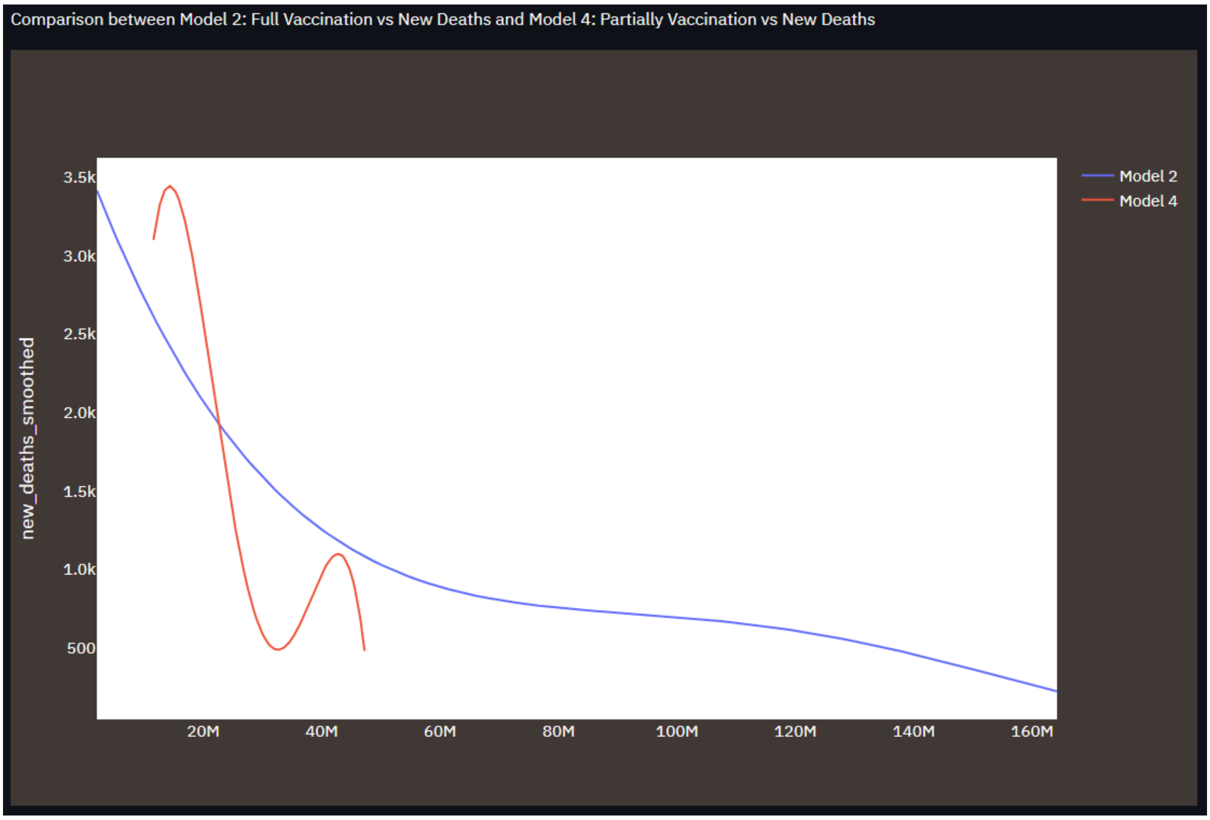
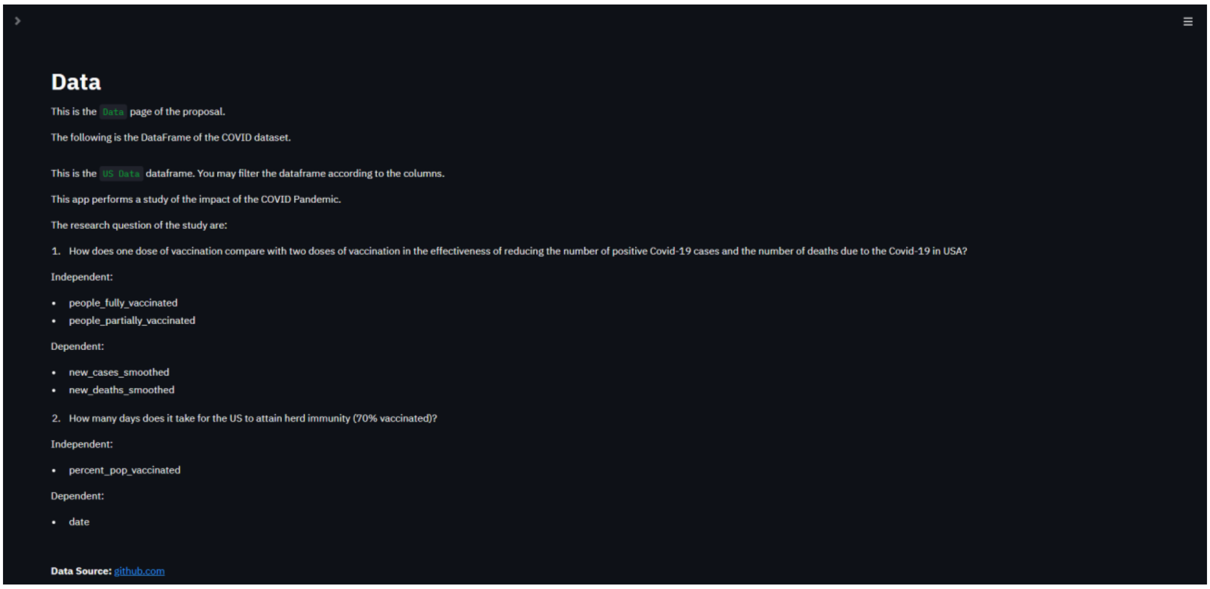


Figure 12: Comparison graph between model 2 and model 4

3.2 Samples Screenshot of the Application



Data

This is the `Data` page of the proposal.

The following is the DataFrame of the COVID dataset.

This is the `US Data` dataframe. You may filter the dataframe according to the columns.

This app performs a study of the impact of the COVID Pandemic.

The research question of the study are:

1. How does one dose of vaccination compare with two doses of vaccination in the effectiveness of reducing the number of positive Covid-19 cases and the number of deaths due to the Covid-19 in USA?

Independent:

- people_fully_vaccinated
- people_partially_vaccinated

Dependent:

- new_cases_smoothed
- new_deaths_smoothed

2. How many days does it take for the US to attain herd immunity (70% vaccinated)?

Independent:

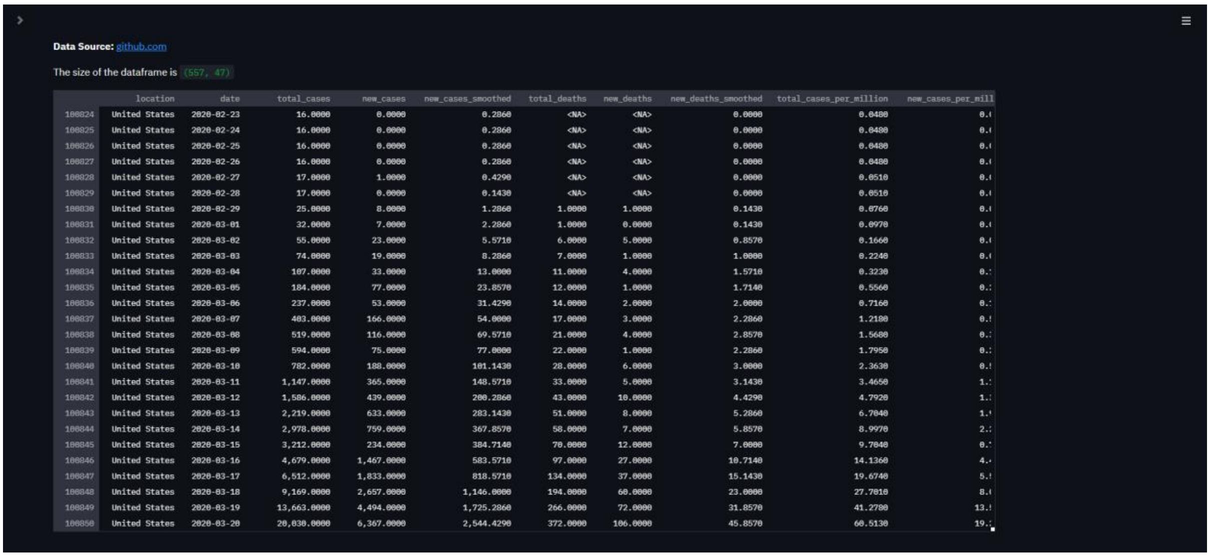
- percent_pop_vaccinated

Dependent:

- date

Data Source: github.com

Figure 13: The introduction Data page of the Streamlit app

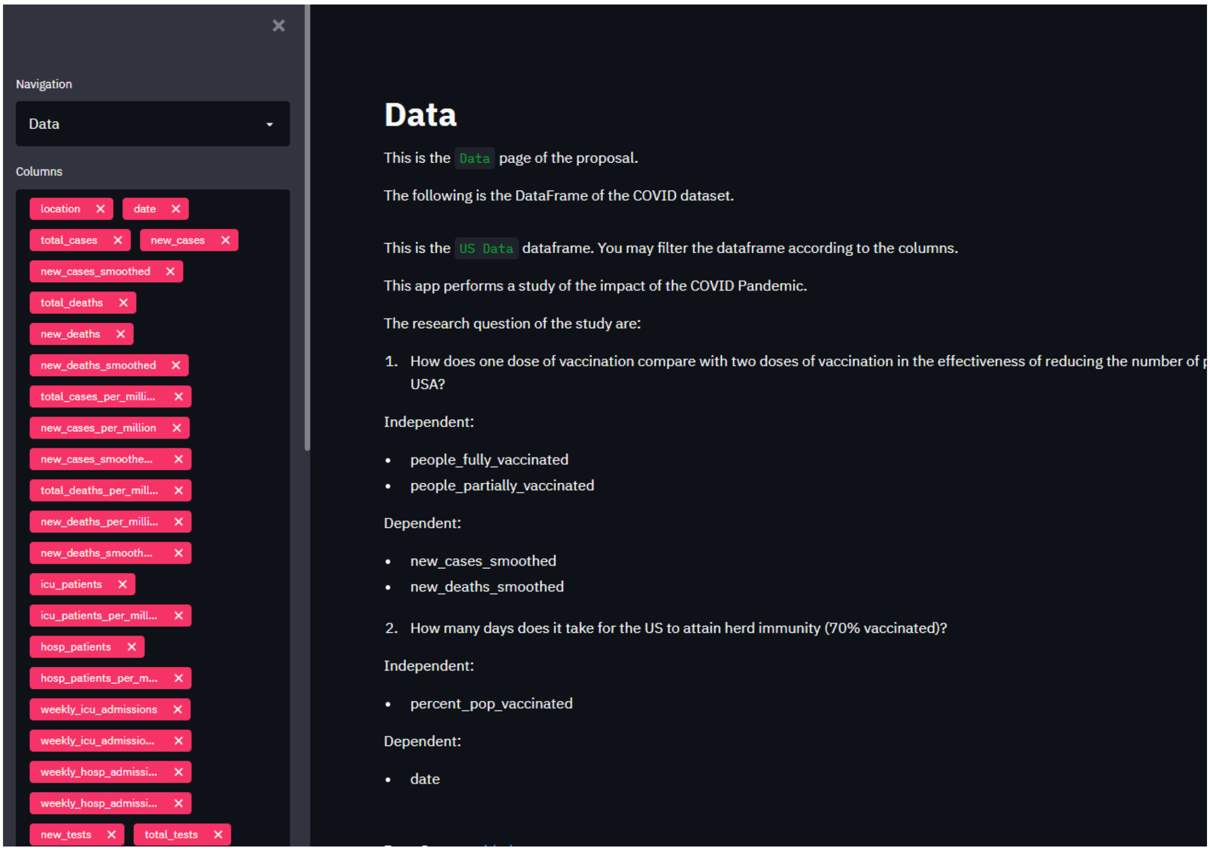


Data Source: github.com

The size of the dataframe is `(557, 47)`

	location	date	total_cases	new_cases	new_cases_smoothed	total_deaths	new_deaths	new_deaths_smoothed	total_cases_per_million	new_cases_per_mill
100824	United States	2020-02-23	16.0000	0.0000	0.2860	<NA>	<NA>	0.0000	0.0480	0.
100825	United States	2020-02-24	16.0000	0.0000	0.2860	<NA>	<NA>	0.0000	0.0480	0.
100826	United States	2020-02-25	16.0000	0.0000	0.2860	<NA>	<NA>	0.0000	0.0480	0.
100827	United States	2020-02-26	16.0000	0.0000	0.2860	<NA>	<NA>	0.0000	0.0480	0.
100828	United States	2020-02-27	17.0000	1.0000	0.4290	<NA>	<NA>	0.0000	0.0510	0.
100829	United States	2020-02-28	17.0000	0.0000	0.1430	<NA>	<NA>	0.0000	0.0510	0.
100830	United States	2020-02-29	25.0000	8.0000	1.2860	1.0000	1.0000	0.1430	0.0760	0.
100831	United States	2020-03-01	32.0000	7.0000	2.2860	1.0000	0.0000	0.1430	0.0970	0.
100832	United States	2020-03-02	55.0000	23.0000	5.5710	6.0000	5.0000	0.8570	0.1660	0.
100833	United States	2020-03-03	74.0000	19.0000	8.2860	7.0000	1.0000	1.0000	0.2240	0.
100834	United States	2020-03-04	107.0000	33.0000	13.0000	11.0000	4.0000	1.5710	0.3230	0.
100835	United States	2020-03-05	184.0000	77.0000	23.8570	12.0000	1.0000	1.7140	0.5560	0.
100836	United States	2020-03-06	237.0000	53.0000	31.4290	14.0000	2.0000	2.0000	0.7160	0.
100837	United States	2020-03-07	403.0000	166.0000	54.0000	17.0000	3.0000	2.2860	1.2180	0.
100838	United States	2020-03-08	519.0000	116.0000	69.5710	21.0000	4.0000	2.8570	1.5680	0.
100839	United States	2020-03-09	594.0000	75.0000	77.0000	22.0000	1.0000	2.2860	1.7950	0.
100840	United States	2020-03-10	782.0000	188.0000	101.1430	28.0000	6.0000	3.0000	2.3630	0.
100841	United States	2020-03-11	1,147.0000	365.0000	148.5710	33.0000	5.0000	3.1430	3.4650	1.
100842	United States	2020-03-12	1,586.0000	439.0000	200.2860	43.0000	10.0000	4.4290	4.7920	1.
100843	United States	2020-03-13	2,219.0000	633.0000	283.1430	51.0000	8.0000	5.2860	6.7040	1.
100844	United States	2020-03-14	2,978.0000	759.0000	367.8570	58.0000	7.0000	5.8570	8.9970	2.
100845	United States	2020-03-15	3,212.0000	234.0000	384.7140	70.0000	12.0000	7.0000	9.7040	0.
100846	United States	2020-03-16	4,679.0000	1,467.0000	583.5710	97.0000	27.0000	10.7140	14.1360	4.
100847	United States	2020-03-17	6,512.0000	1,833.0000	818.5710	134.0000	37.0000	15.1430	19.6740	5.
100848	United States	2020-03-18	9,169.0000	2,657.0000	1,146.0000	194.0000	60.0000	23.0000	27.7010	8.
100849	United States	2020-03-19	13,663.0000	4,494.0000	1,725.2860	266.0000	72.0000	31.8570	41.2780	13.
100850	United States	2020-03-20	20,030.0000	6,367.0000	2,544.4290	372.0000	106.0000	45.8570	60.5130	19.

Figure 14: Dataframe displaying our dataset



Data

This is the `Data` page of the proposal.

The following is the DataFrame of the COVID dataset.

This is the `US Data` dataframe. You may filter the dataframe according to the columns.

This app performs a study of the impact of the COVID Pandemic.

The research question of the study are:

1. How does one dose of vaccination compare with two doses of vaccination in the effectiveness of reducing the number of p
USA?

Independent:

- people_fully_vaccinated
- people_partially_vaccinated

Dependent:

- new_cases_smoothed
- new_deaths_smoothed

2. How many days does it take for the US to attain herd immunity (70% vaccinated)?

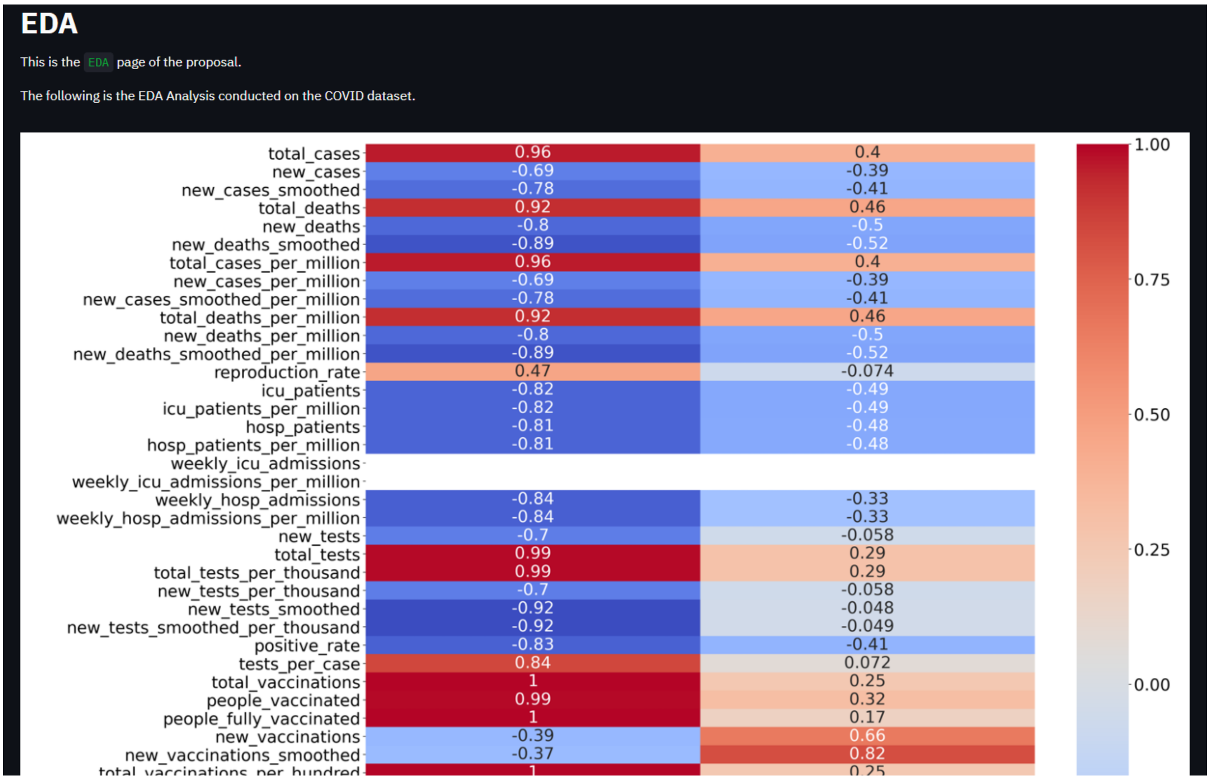
Independent:

- percent_pop_vaccinated

Dependent:

- date

Figure 15: The sidebar to change page and choose columns in the dataframe



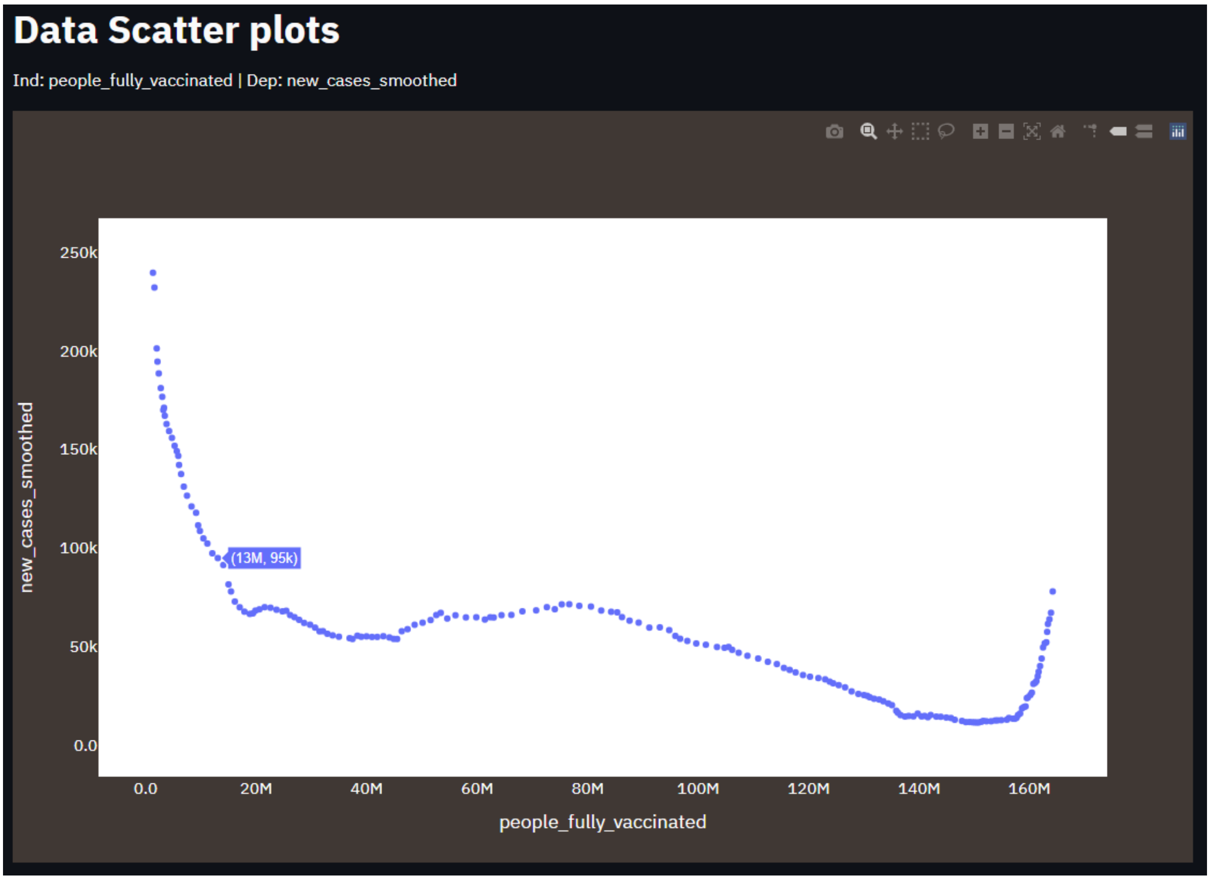
EDA

This is the `EDA` page of the proposal.

The following is the EDA Analysis conducted on the COVID dataset.

total_cases	0.96	0.4
new_cases	-0.69	-0.39
new_cases_smoothed	-0.78	-0.41
total_deaths	0.92	0.46
new_deaths	-0.8	-0.5
new_deaths_smoothed	-0.89	-0.52
total_cases_per_million	0.96	0.4
new_cases_per_million	-0.69	-0.39
new_cases_smoothed_per_million	-0.78	-0.41
total_deaths_per_million	0.92	0.46
new_deaths_per_million	-0.8	-0.5
new_deaths_smoothed_per_million	-0.89	-0.52
reproduction_rate	0.47	-0.074
icu_patients	-0.82	-0.49
icu_patients_per_million	-0.82	-0.49
hosp_patients	-0.81	-0.48
hosp_patients_per_million	-0.81	-0.48
weekly_icu_admissions		
weekly_icu_admissions_per_million		
weekly_hosp_admissions	-0.84	-0.33
weekly_hosp_admissions_per_million	-0.84	-0.33
new_tests	-0.7	-0.058
total_tests	0.99	0.29
total_tests_per_thousand	0.99	0.29
new_tests_per_thousand	-0.7	-0.058
new_tests_smoothed	-0.92	-0.048
new_tests_smoothed_per_thousand	-0.92	-0.049
positive_rate	-0.83	-0.41
tests_per_case	0.84	0.072
total_vaccinations	1	0.25
people_vaccinated	0.99	0.32
people_fully_vaccinated	1	0.17
new_vaccinations	-0.39	0.66
new_vaccinations_smoothed	-0.37	0.82
total_vaccinations_per_hundred	1	0.25

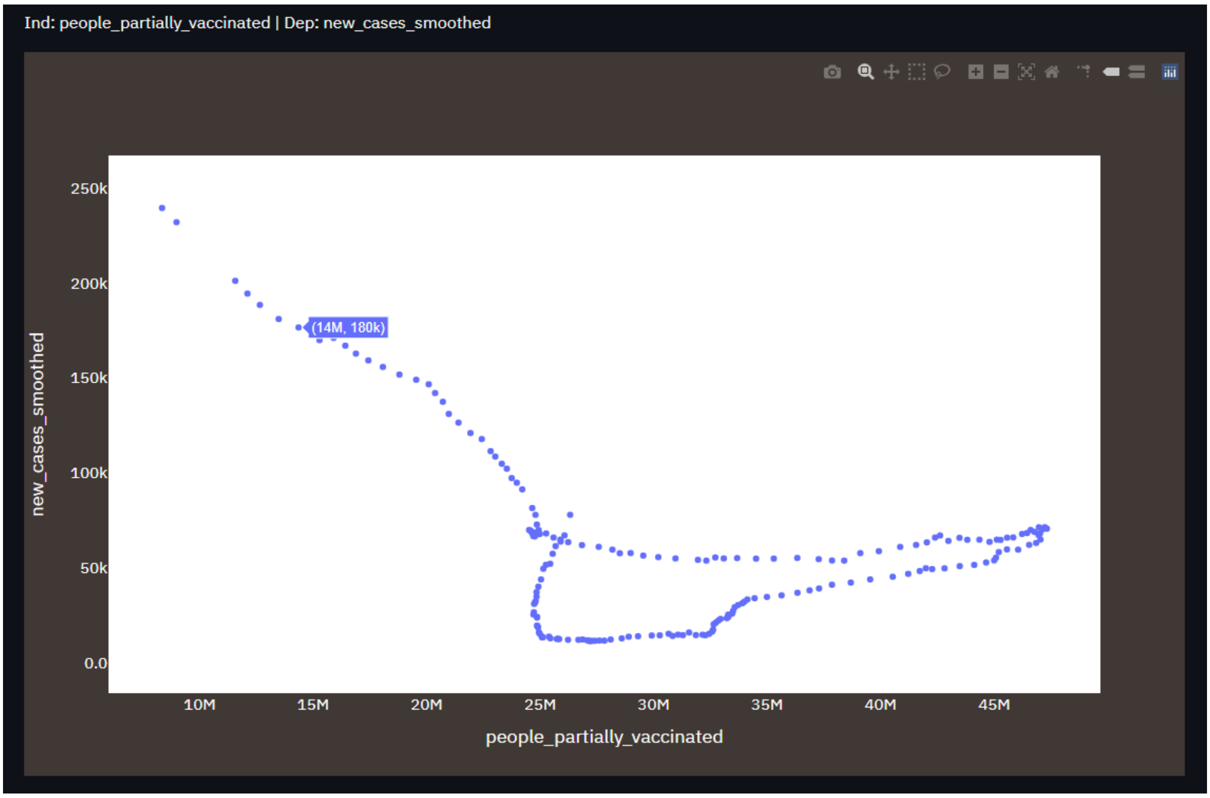
Figure 16: The EDA Page. The heatmap of correlation between independent variables and other variables.



Data Scatter plots

Ind: people_fully_vaccinated | Dep: new_cases_smoothed

Figure 17: People_fully_vaccinated against new_cases_smoothed.



Ind: people_partially_vaccinated | Dep: new_cases_smoothed

Figure 18: people_partially vaccinated against new_cases_smoothed

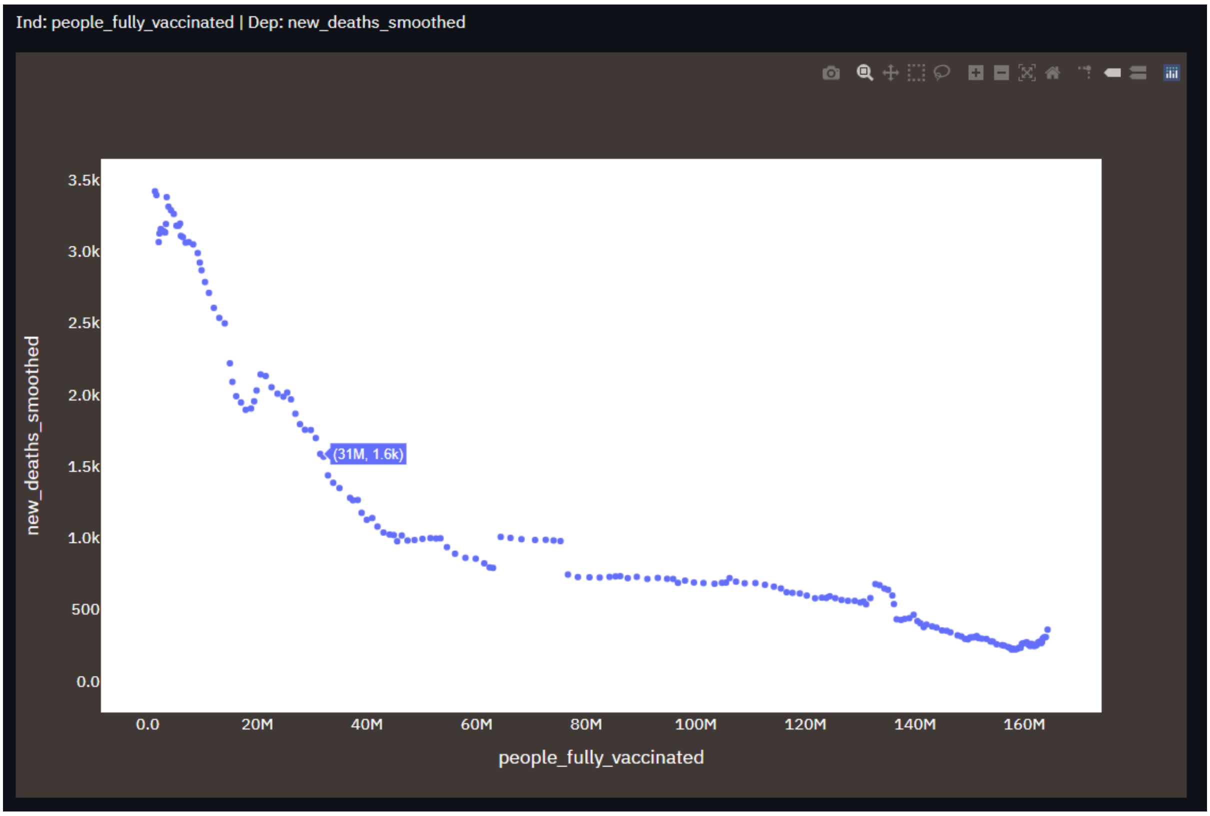


Figure 19: for people_fully_vaccinated against new_deaths_smoothed.

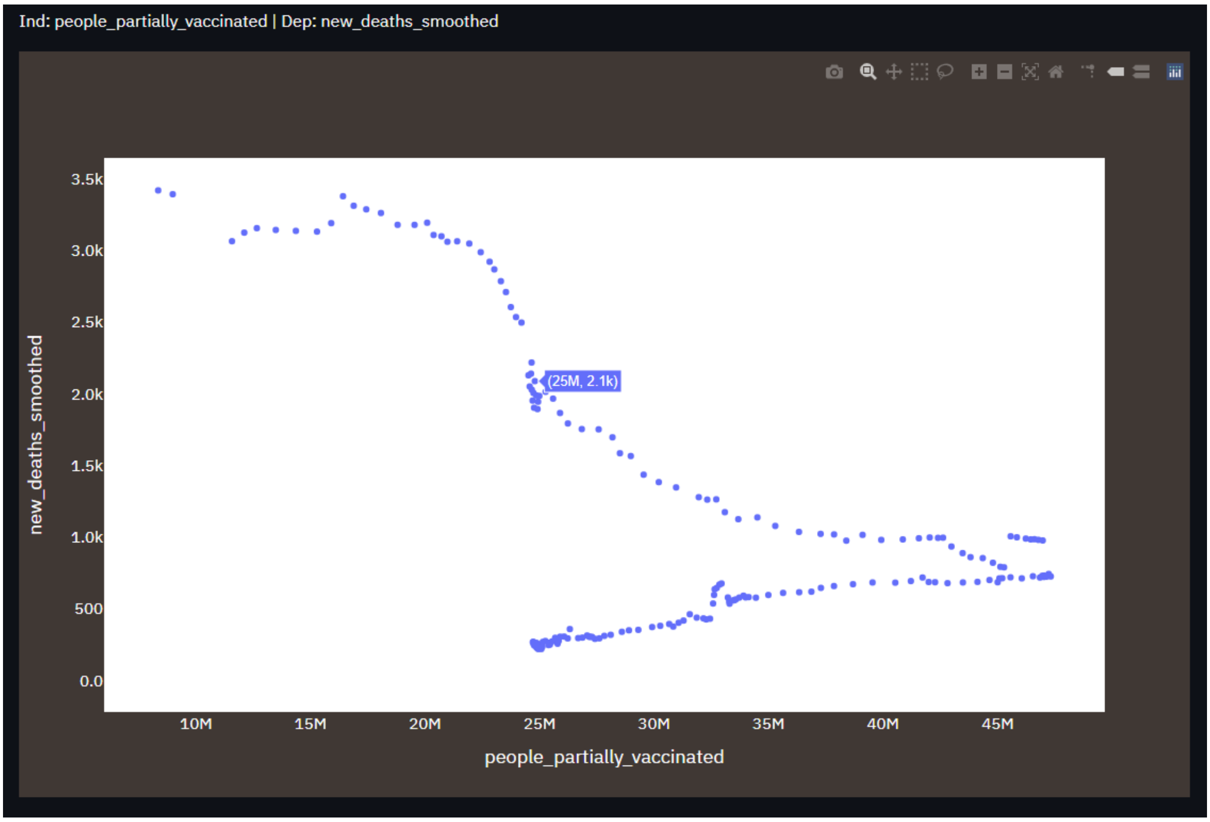
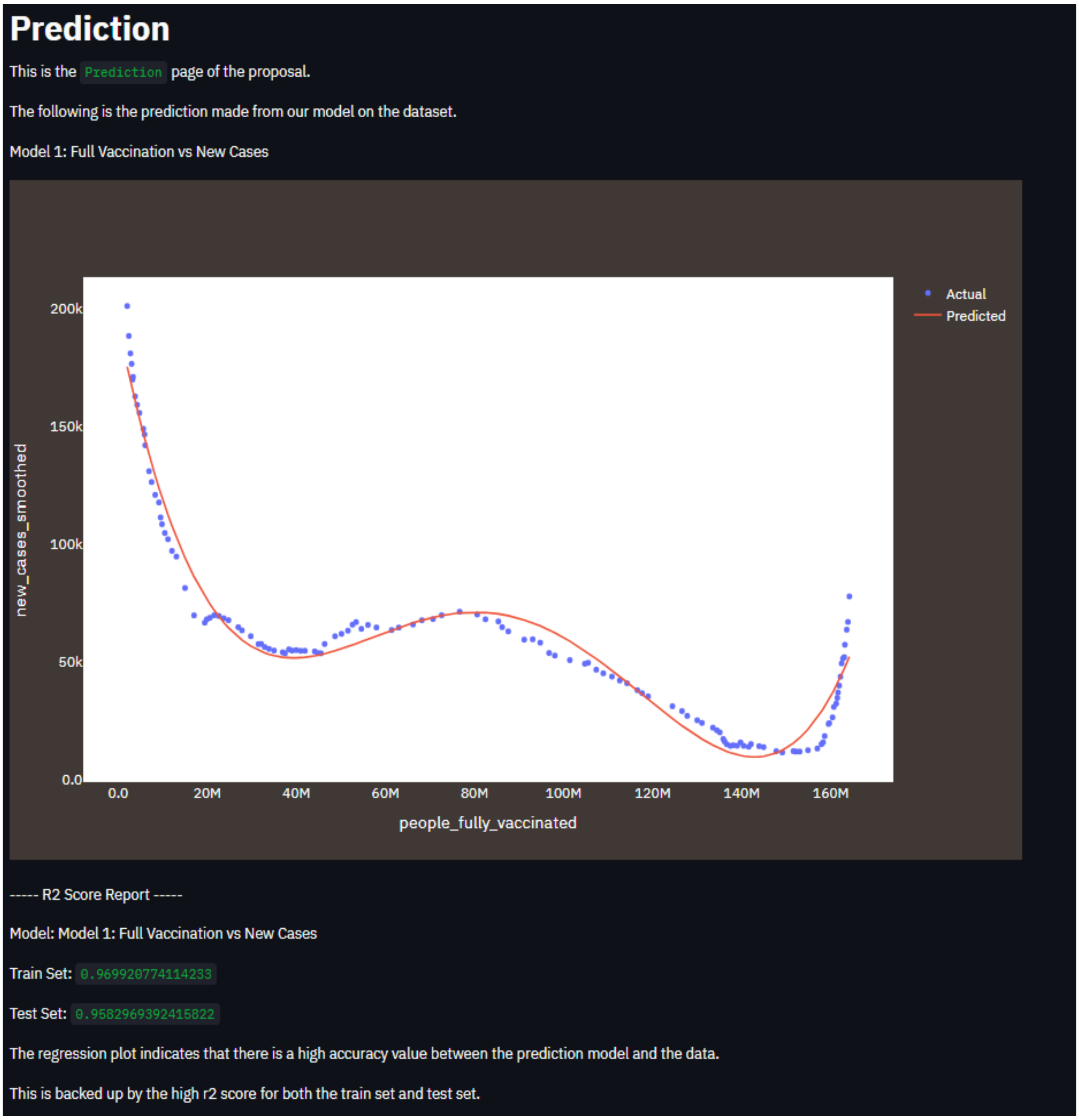


Figure 20: people_partially_vaccinated against new_deaths_smoothed.



Prediction

This is the `Prediction` page of the proposal.

The following is the prediction made from our model on the dataset.

Model 1: Full Vaccination vs New Cases

----- R2 Score Report -----

Model: Model 1: Full Vaccination vs New Cases

Train Set: `0.969920774114233`

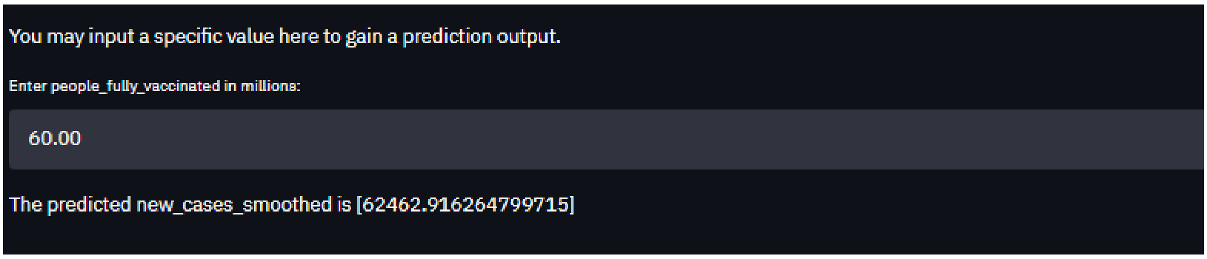
Test Set: `0.9582969392415822`

The regression plot indicates that there is a high accuracy value between the prediction model and the data.

This is backed up by the high r2 score for both the train set and test set.

Figure 21: Prediction Page, Prediction model between people_full_vaccinated against new_cases_smoothed, R2 Score Report.

- The graph allows users to select charts in the legends.



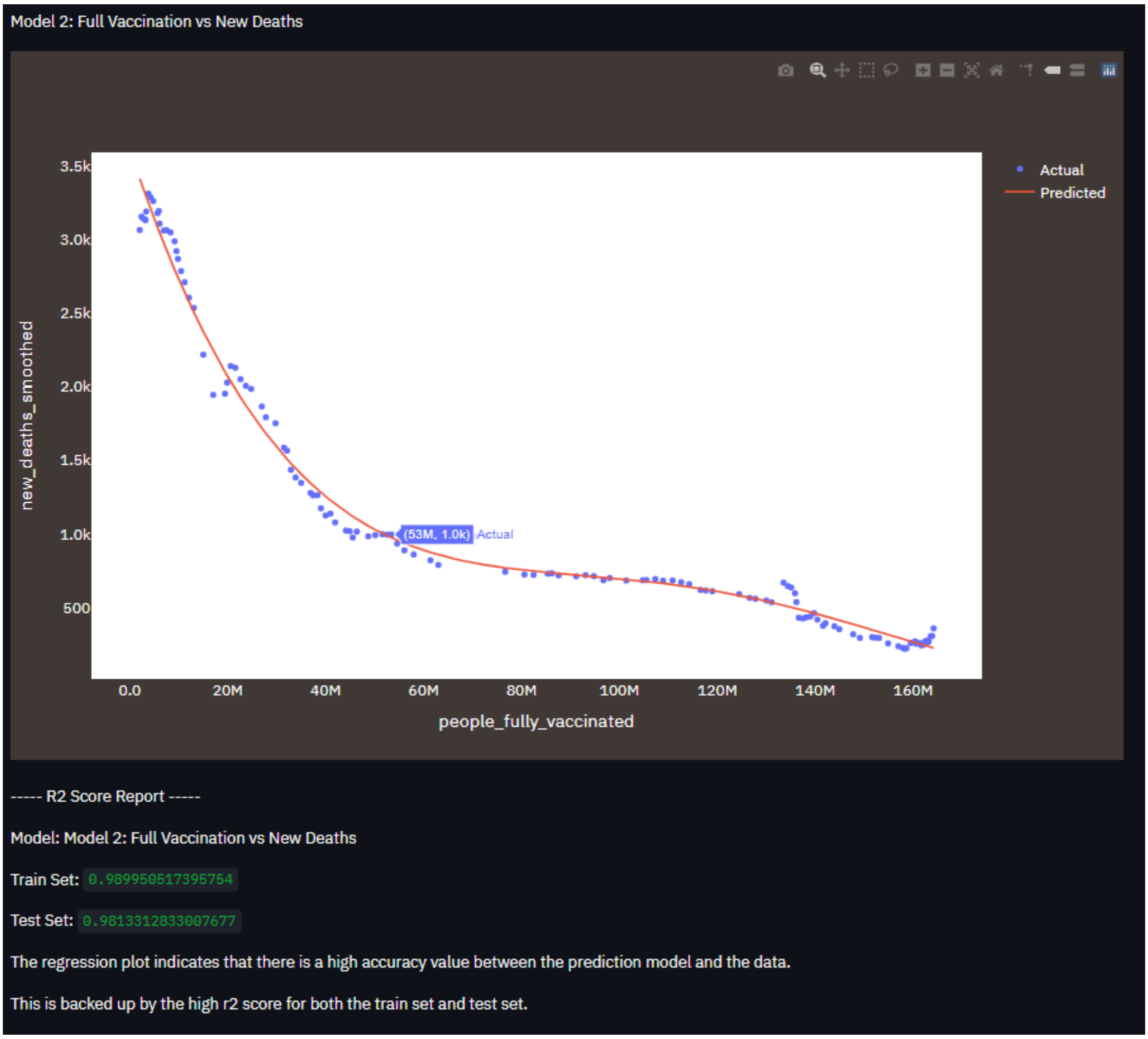
You may input a specific value here to gain a prediction output.

Enter people_fully_vaccinated in millions:

60.00

The predicted new_cases_smoothed is [62462.916264799715]

Figure 22: User input prediction for new_cases_smoothed based on people_fully_vaccinated.



Model 2: Full Vaccination vs New Deaths

----- R2 Score Report -----

Model: Model 2: Full Vaccination vs New Deaths

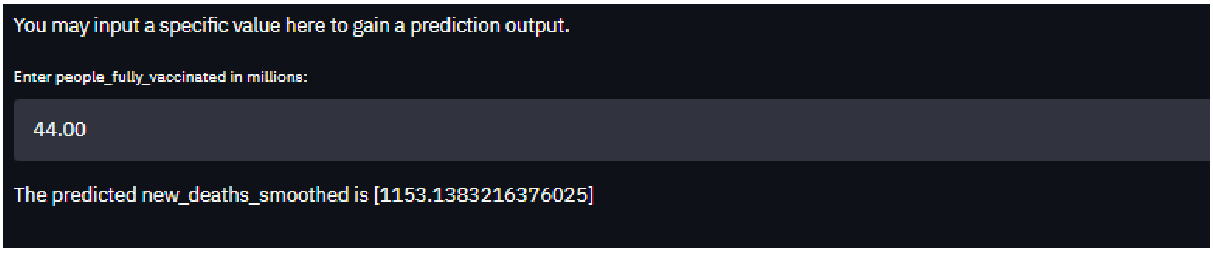
Train Set: `0.989950517395754`

Test Set: `0.9813312833007677`

The regression plot indicates that there is a high accuracy value between the prediction model and the data.

This is backed up by the high r2 score for both the train set and test set.

Figure 23: Prediction model between people_full_vaccinated against new_deaths_smoothed, R2 Score Report.



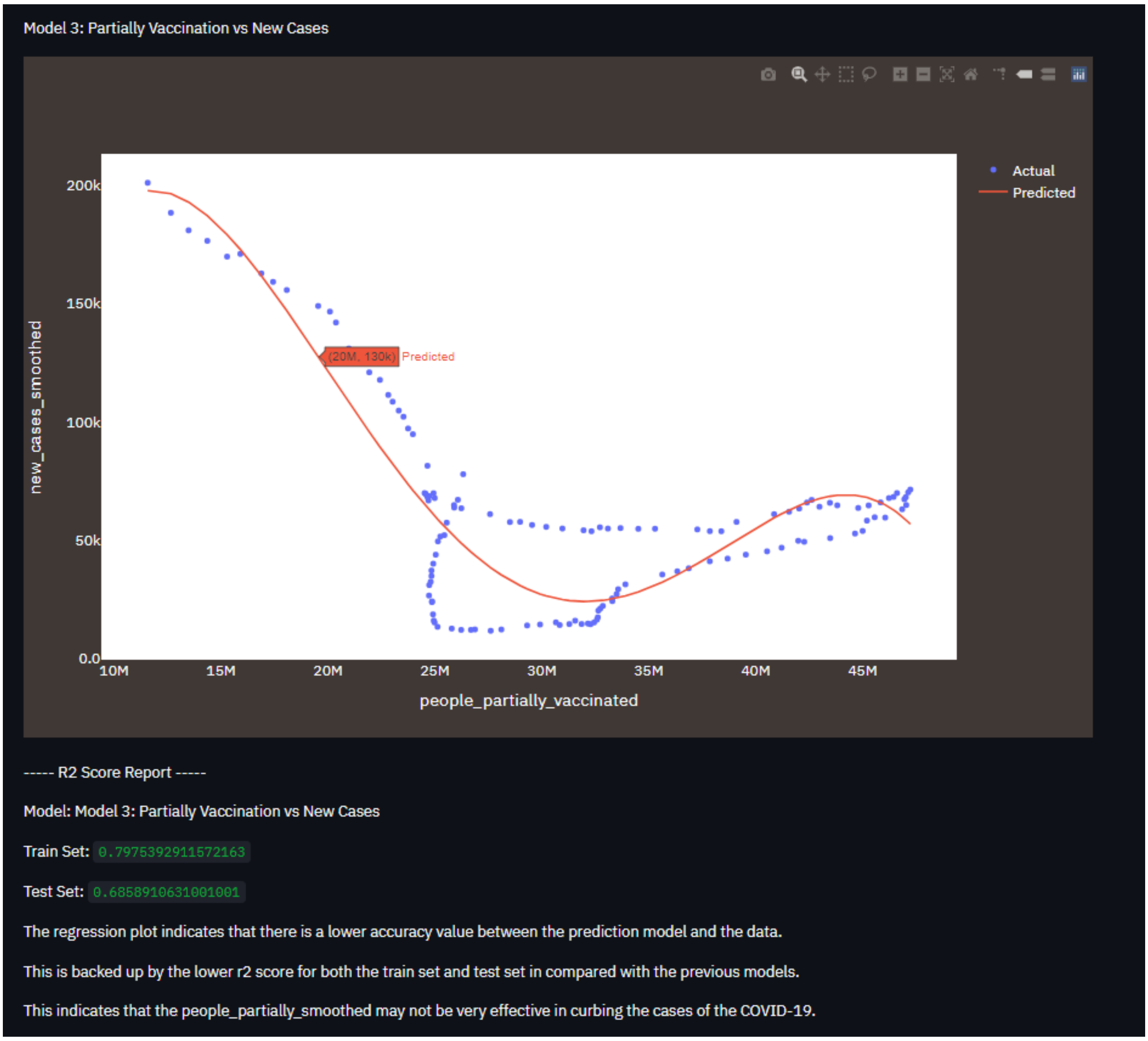
You may input a specific value here to gain a prediction output.

Enter people_fully_vaccinated in millions:

44.00

The predicted new_deaths_smoothed is [1153.1383216376025]

Figure 24: User input prediction for new_deaths_smoothed based on people_fully_vaccinated.


Model 3: Partially Vaccination vs New Cases

----- R2 Score Report -----

Model: Model 3: Partially Vaccination vs New Cases

Train Set: `0.7975392911572163`

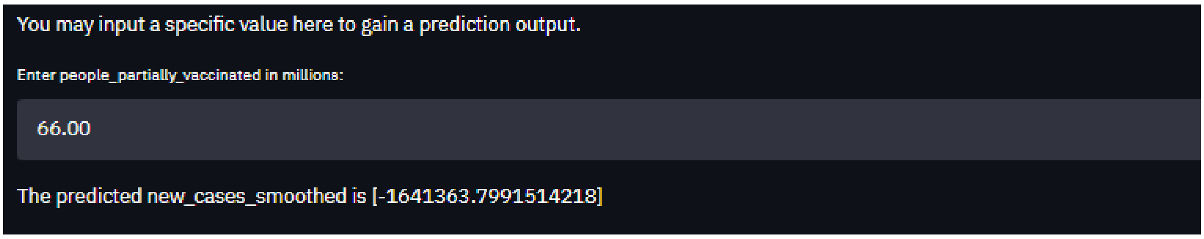
Test Set: `0.6858910631001001`

The regression plot indicates that there is a lower accuracy value between the prediction model and the data.

This is backed up by the lower r2 score for both the train set and test set in compared with the previous models.

This indicates that the people_partially_smoothed may not be very effective in curbing the cases of the COVID-19.

Figure 25: Prediction model between people_partially_vaccinated against new_cases_smoothed, R2 Score Report.

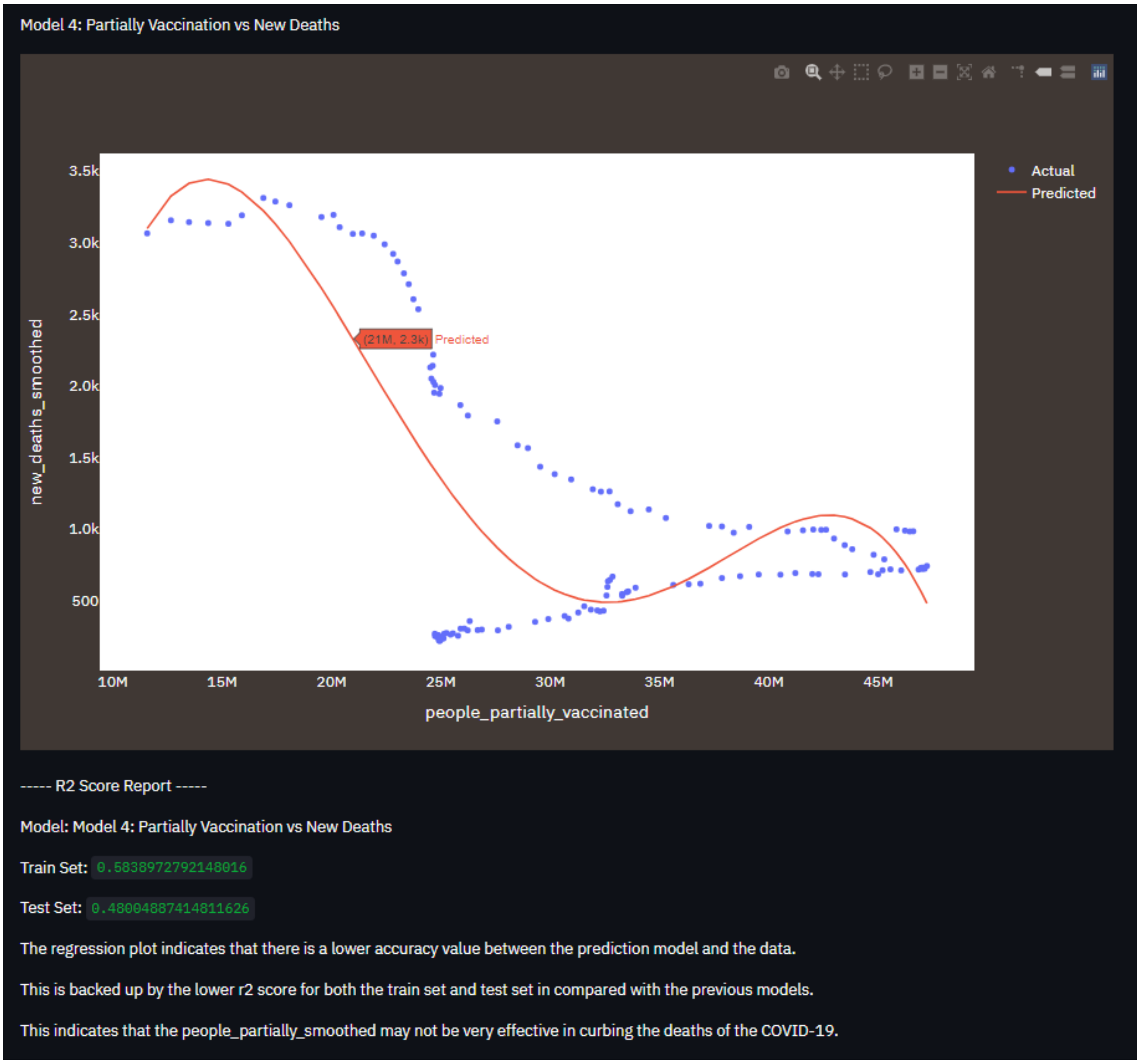

You may input a specific value here to gain a prediction output.

Enter people_partially_vaccinated in millions:

66.00

The predicted new_cases_smoothed is [-1641363.7991514218]

Figure 26: User input prediction for new_cases_smoothed based on people_partially_vaccinated.



Model 4: Partially Vaccination vs New Deaths

----- R2 Score Report -----

Model: Model 4: Partially Vaccination vs New Deaths

Train Set: `0.5838972792148016`

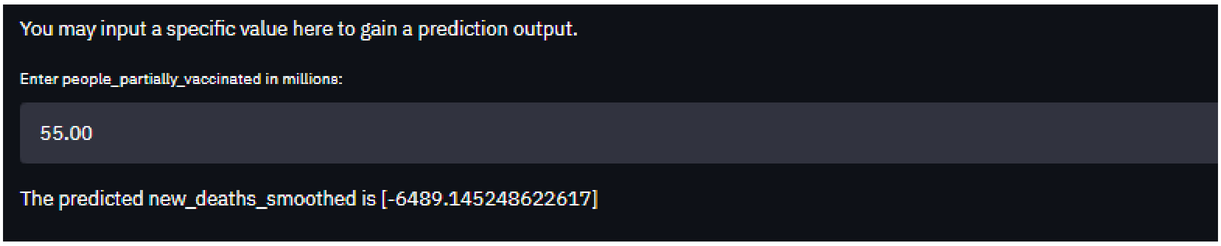
Test Set: `0.48004887414811626`

The regression plot indicates that there is a lower accuracy value between the prediction model and the data.

This is backed up by the lower r2 score for both the train set and test set in compared with the previous models.

This indicates that the people_partially_smoothed may not be very effective in curbing the deaths of the COVID-19.

Figure 27: Prediction model between people_partially_vaccinated against new_deaths_smoothed, R2 Score Report.



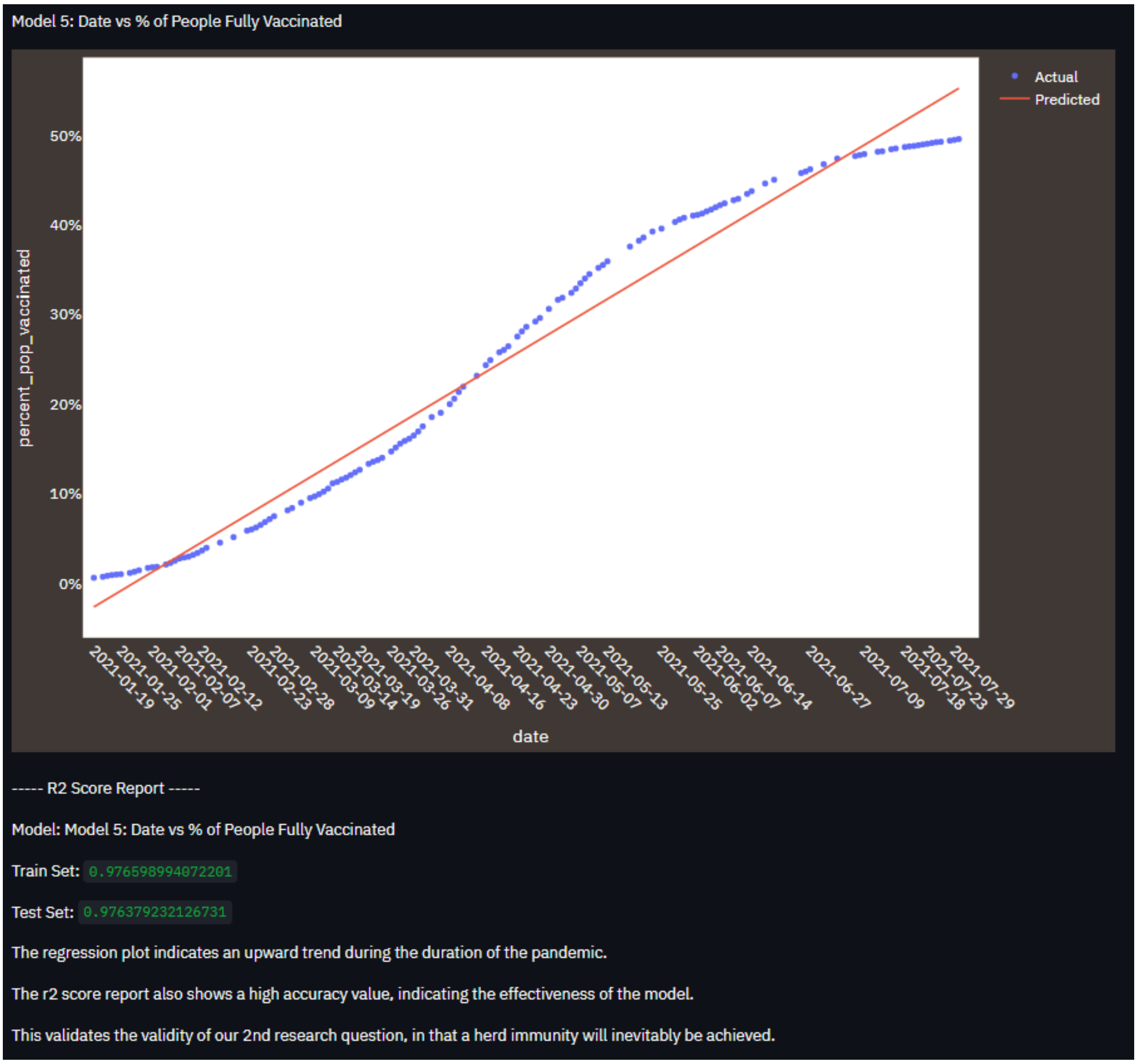
You may input a specific value here to gain a prediction output.

Enter people_partially_vaccinated in millions:

55.00

The predicted new_deaths_smoothed is [-6489.145248622617]

Figure 28: User input prediction for new_deaths_smoothed based on people_partially_vaccinated.



Model 5: Date vs % of People Fully Vaccinated

----- R2 Score Report -----

Model: Model 5: Date vs % of People Fully Vaccinated

Train Set: `0.976598994072201`

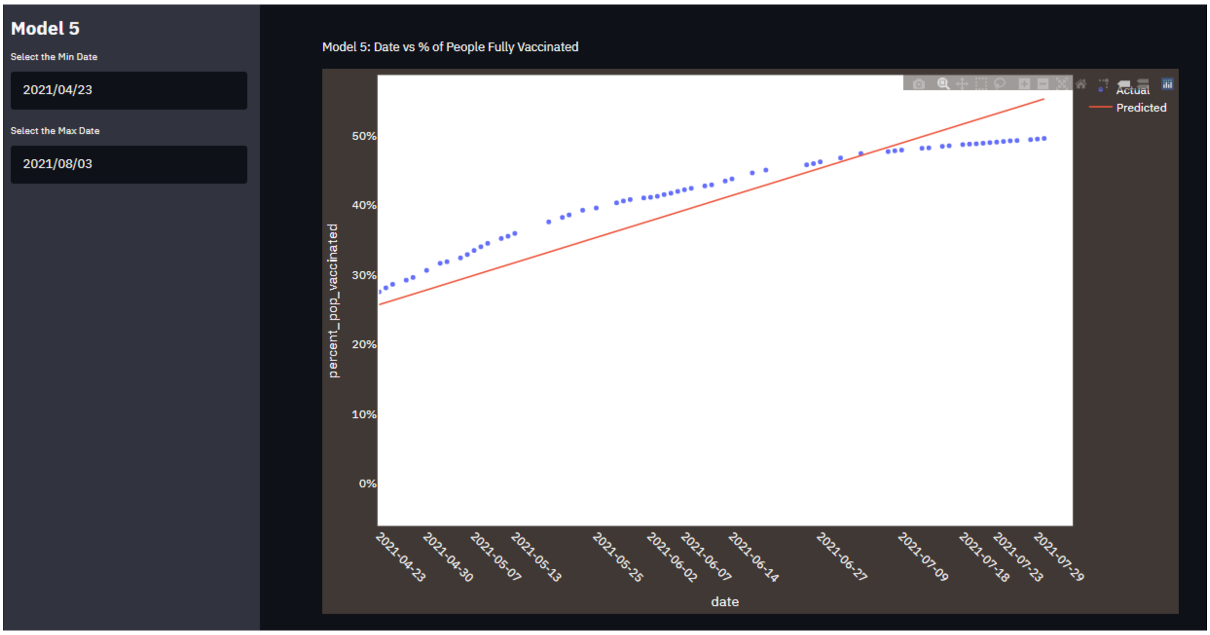
Test Set: `0.976379232126731`

The regression plot indicates an upward trend during the duration of the pandemic.

The r2 score report also shows a high accuracy value, indicating the effectiveness of the model.

This validates the validity of our 2nd research question, in that a herd immunity will inevitably be achieved.

Figure 29: Prediction model between date against people_pop_vaccinated, R2 Score Report.



Model 5

Select the Min Date

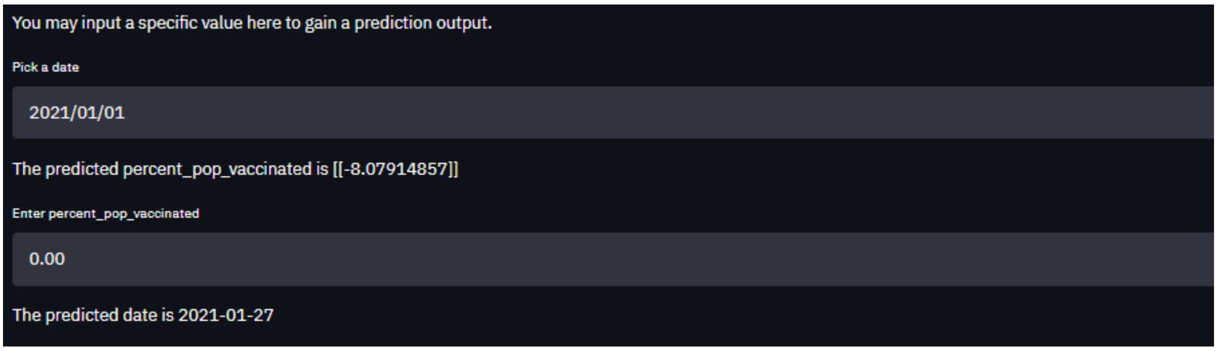
2021/04/23

Select the Max Date

2021/08/03

Figure 30: Date picker for model 5

- Date input is limited to the range of the dataset, 12-1-2021 to 3/8/2021.



You may input a specific value here to gain a prediction output.

Pick a date

2021/01/01

The predicted percent_pop_vaccinated is [[-8.07914857]]

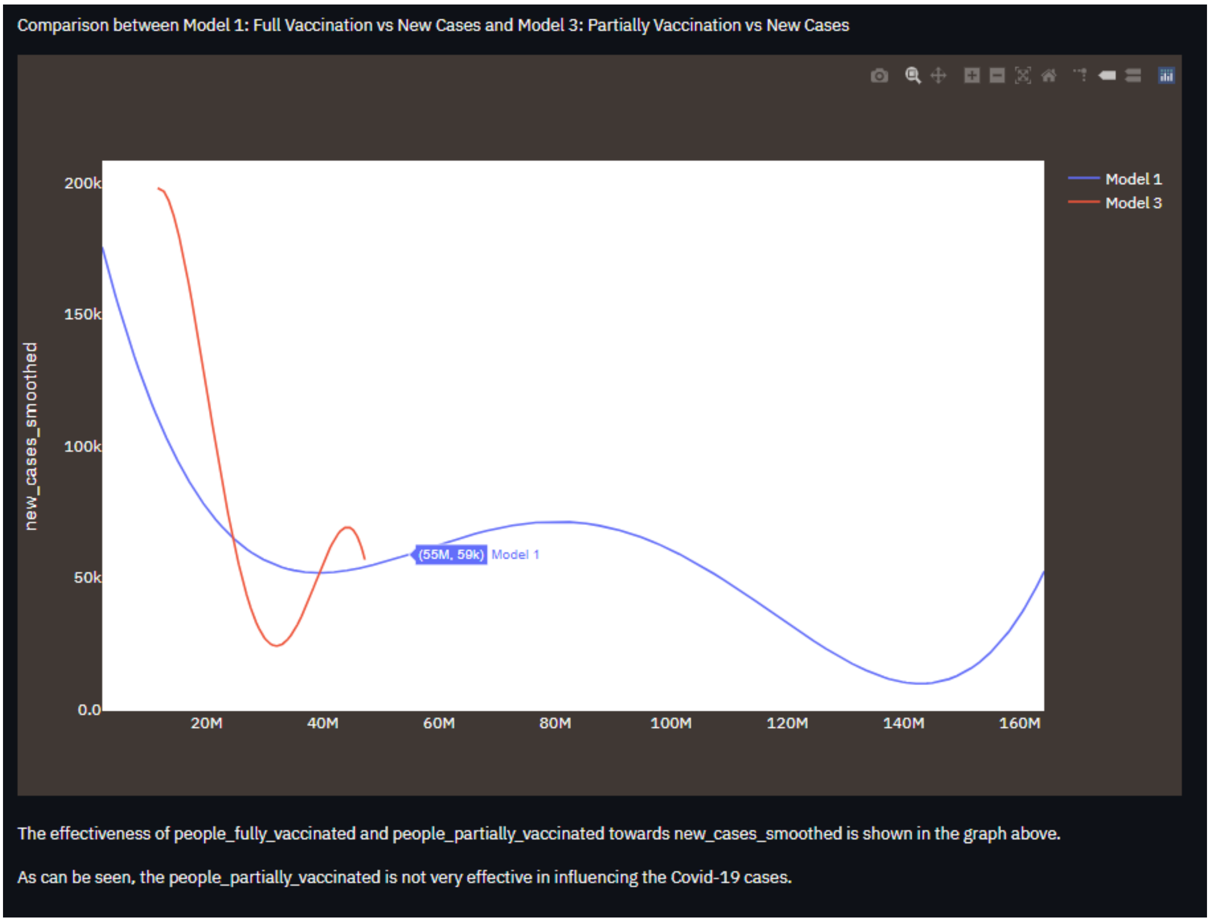
Enter percent_pop_vaccinated

0.00

The predicted date is 2021-01-27

Figure 31: User input prediction for percent_pop_vaccinated based on date, and date prediction based on percent_pop_vaccinated.

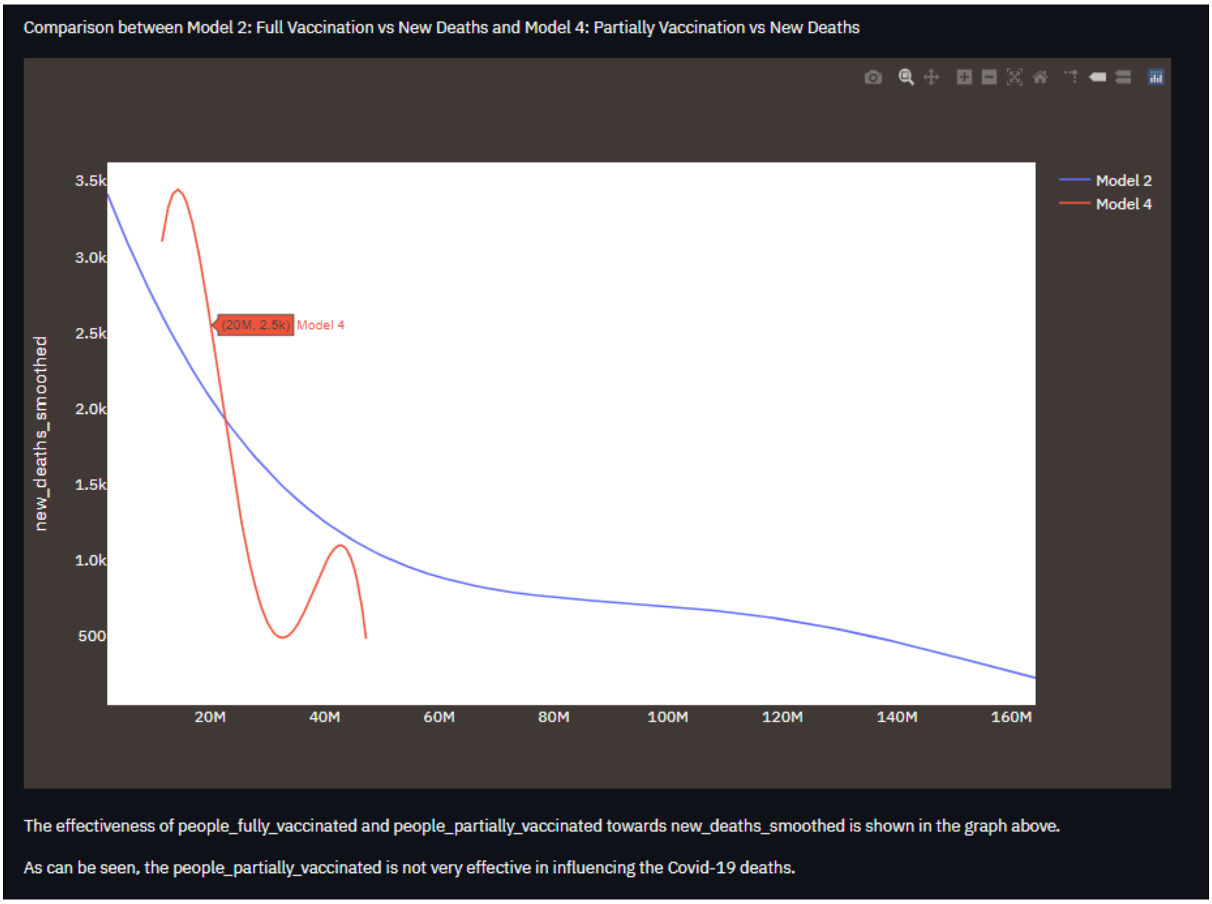
- Date input is limited to 1-1-2021 to 1-1-2023


Comparison between Model 1: Full Vaccination vs New Cases and Model 3: Partially Vaccination vs New Cases

The effectiveness of people_fully_vaccinated and people_partially_vaccinated towards new_cases_smoothed is shown in the graph above.

As can be seen, the people_partially_vaccinated is not very effective in influencing the Covid-19 cases.

Figure 32: Regression line comparison between model 1 and model 3.



Comparison between Model 2: Full Vaccination vs New Deaths and Model 4: Partially Vaccination vs New Deaths

The effectiveness of people_fully_vaccinated and people_partially_vaccinated towards new_deaths_smoothed is shown in the graph above.

As can be seen, the people_partially_vaccinated is not very effective in influencing the Covid-19 deaths.

Figure 33: Regression line comparison between model 2 and model 4.

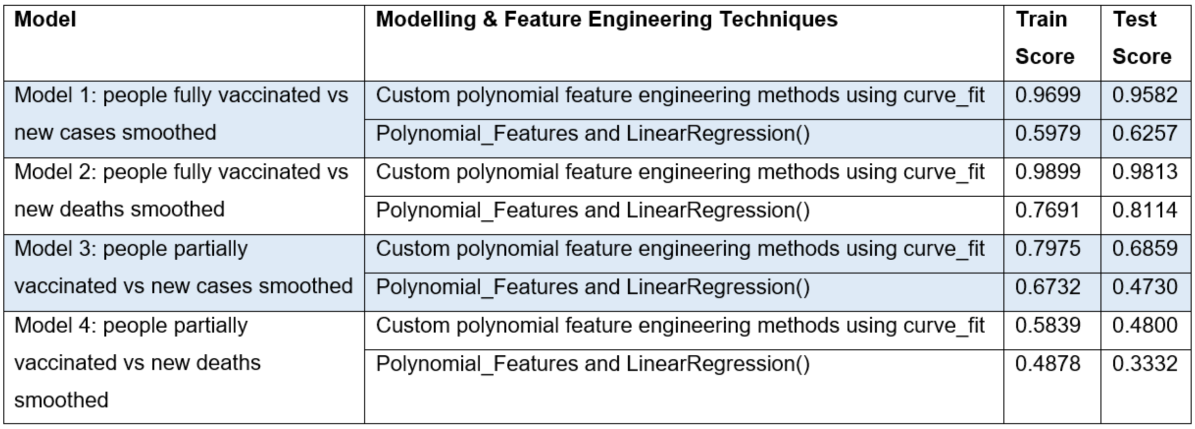
4 Discussion

4.1 Changes in Tools and Methods Used

A custom feature engineering technique has been used to apply polynomial regression to the models under the RQ1. In this case, a quartic function $y = ax^4 + bx^3 + cx^2 + dx + e$ is defined and the scipy.optimize.curve_fit function is used. The scipy.optimize.curve_fit is a library that uses a nonlinear least square method to fit a function by returning a series of optimized coefficients for the curve. It is typically used to sketch best-fit curves for scatter plot graphs with a nonlinear trend. The curve_fit parameter requires 3 variable inputs, including the polynomial function, the x-array, and the y-array. The generated optimized curve coefficients and the x_train values will be used in the quartic function to train and predict the model.

The reason for using a customized feature engineering technique rather than using the conventional PolynomialFeatures together with the LinearRegression modules in scikit-learn is because the conventional way does not train our model well. Table 6 below compares the result of the train and test score using the conventional method and the custom method.

Table 6: PolynomialFeatures vs custom features

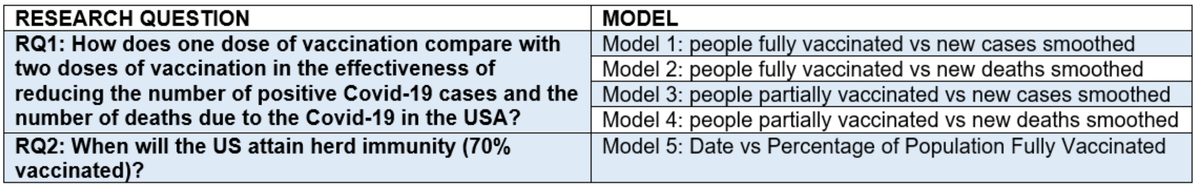


Model	Modelling & Feature Engineering Techniques	Train Score	Test Score
Model 1: people fully vaccinated vs new cases smoothed	Custom polynomial feature engineering methods using curve_fit	0.9699	0.9582
	Polynomial_Features and LinearRegression()	0.5979	0.6257
Model 2: people fully vaccinated vs new deaths smoothed	Custom polynomial feature engineering methods using curve_fit	0.9899	0.9813
	Polynomial_Features and LinearRegression()	0.7691	0.8114
Model 3: people partially vaccinated vs new cases smoothed	Custom polynomial feature engineering methods using curve_fit	0.7975	0.6859
	Polynomial_Features and LinearRegression()	0.6732	0.4730
Model 4: people partially vaccinated vs new deaths smoothed	Custom polynomial feature engineering methods using curve_fit	0.5839	0.4800
	Polynomial_Features and LinearRegression()	0.4878	0.3332

As observed above, our custom features performed better compared to the conventional polynomial regression model.

4.2 Answer to the Research Question

Table 7: Models used in each research question.



RESEARCH QUESTION	MODEL
RQ1: How does one dose of vaccination compare with two doses of vaccination in the effectiveness of reducing the number of positive Covid-19 cases and the number of deaths due to the Covid-19 in the USA?	Model 1: people fully vaccinated vs new cases smoothed
	Model 2: people fully vaccinated vs new deaths smoothed
	Model 3: people partially vaccinated vs new cases smoothed
	Model 4: people partially vaccinated vs new deaths smoothed
RQ2: When will the US attain herd immunity (70% vaccinated)?	Model 5: Date vs Percentage of Population Fully Vaccinated

4.2.1 Research Question 1 (RQ1)

Models 1 and 2 show an exceptionally high train and test score that is over 95% accuracy. Model 3 shows a lower but acceptable accuracy train and test scores, 0.7975 and 0.6859 respectively. However, model 4 has a low accuracy train and test score that is below 0.5.

Models 3 and 4 faced serious data fluctuations to the point that they almost looked like a one-to-many relation (see figure 12). A weird trend is observed in figures 10 and 11 whereby those partially vaccinated regression lines in red show a greater decrease in the number of new cases and new deaths as compared to the fully vaccinated. Besides, the regression lines of models 3 and 4 came to a stop at around 50 million people as compared to 160 million people in models 1 and 2. This is because those who have received their first dose of vaccine will have their second dose appointment scheduled in approximately three weeks after their first dose, therefore the number of people partially vaccinated is lesser than the number of people fully vaccinated.

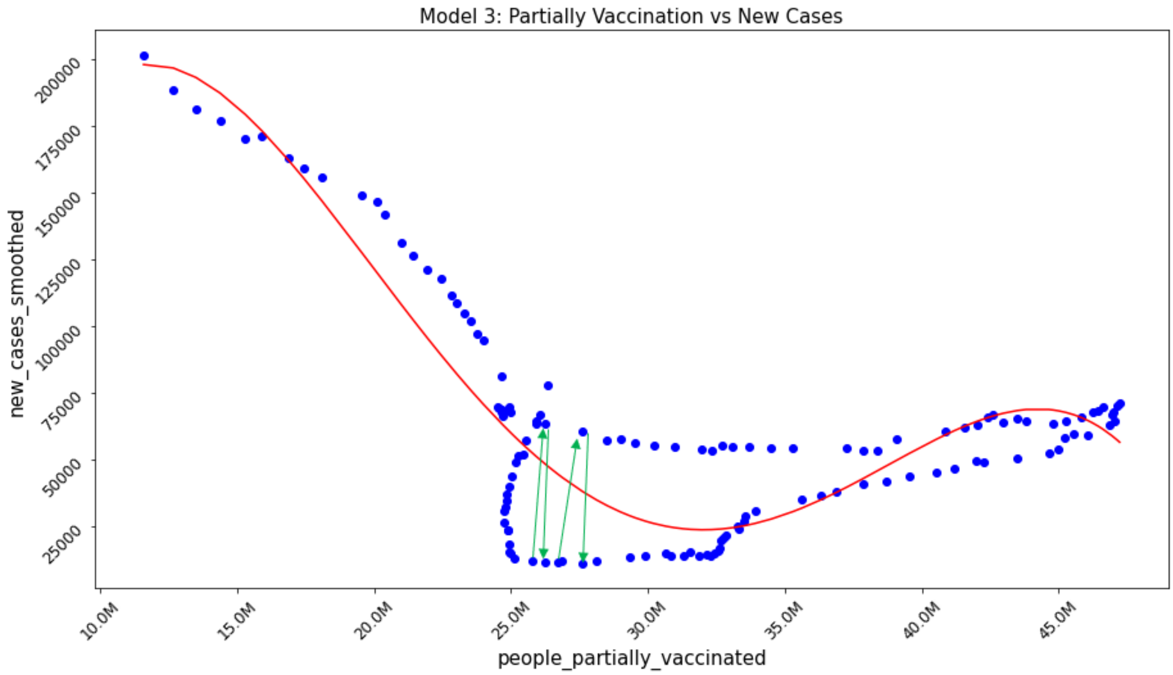
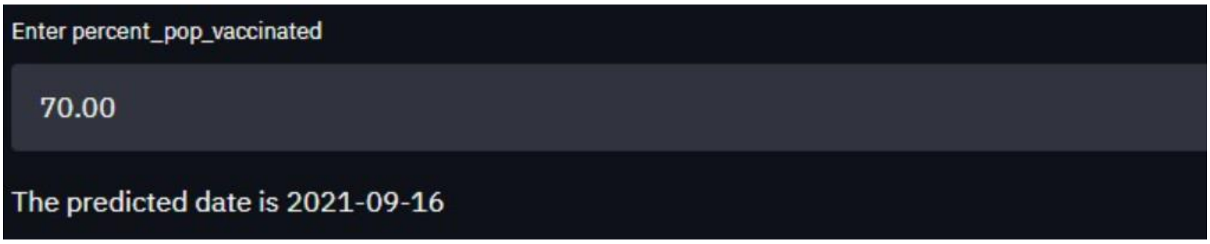


Figure 34: Data fluctuation

It is difficult to make a direct comparison between the number of people fully vaccinated (models 1 and 2) and the number of people partially vaccinated (models 3 and 4) in terms of their effectiveness in controlling the new cases and new deaths. However, we can safely say that the number of people fully vaccinated is certainly more reliable data to investigate the effectiveness of vaccination in controlling the new cases (model 1) and new deaths (model 2) as the data are more stable. Model 1 does fluctuate at some point but it is not as severe as compared to model 3. The fluctuation could be caused by some other extraneous variables that have some influence on the new cases, like mutated delta variant of the covid-19 virus, which is more contagious; and the loosen of lockdown restrictions. The death cases under model 1 show a constant decreasing trend without any fluctuation. It shows that although completing two doses of vaccination may not fully prevent one from getting infected by the covid-19 virus, however, it is effective in reducing the severity of the disease since it reduces the number of deaths. In conclusion, there is no strong evidence to claim that the number of people fully vaccinated is more effective than the number of people partially vaccinated in reducing the new cases and death cases; however, the number of people fully vaccinated is a more reliable data to investigate the effectiveness of vaccines.

4.2.2 Research Question 2 (RQ2)

As for RQ2, model 5 has a very high train and test accuracy score of 0.9766 and 0.9764 respectively. Figure 35 below shows a prediction made by the system in which 70% of the population will be fully vaccinated by 2021-09-16. Therefore, according to our model, the hypothesis claiming that the USA can attain herd immunity by winter 2021 is valid as September is the fall season.



Enter percent_pop_vaccinated

70.00

The predicted date is 2021-09-16

Figure 35: Model 5 herd immunity prediction

5 Conclusion

5.1 Implications of the Findings to Machine Learning Area

The use of the *scipy.optimize.curve_fit* to get the best-fit curve coefficients and train the model by fitting the training dataset with the coefficients could be a new potential way of modelling polynomial regression when the traditional method of using *sklearn.preprocessing.PolynomialFeatures* and *sklearn.linear_model.LinearRegression* does not produce the desired result. The curve_fit package is normally used to sketch the best fit curve for non-linear data. To the best of our knowledge, we have not come across a researcher using this method to model a polynomial regression.

5.2 Limitations & Future Enhancement

Our study focuses on the USA; therefore, the training data is limited. The initial 96645 records as shown in figure 1 reduced to 189 records as shown in figure 4 after data cleaning and narrowing down to the USA. The 189 records are further divided into train set and test set in the ratio of 7:3, leaving only 132 records to train the model. As discussed in section 2.4 and demonstrated in figure 5, focusing on one country was necessary due to different population sizes. Perhaps feature scaling on the population size of other countries as well as on the independent and dependent variables can be done to get more training data to train our data more accurately. However, accuracy cannot be guaranteed as different countries are in different situations. Lastly, this study does not support uncontrollable factors such as the emergence of new mutated variants of covid-19 viruses that could be potentially more contagious and fatal.

---Total word count: 3288 (including titles and picture captions)---

6 References

COVID-19 Vaccination. (2021). Retrieved 8 August 2021, from https://www.cdc.gov/coronavirus/2019-ncov/vaccines/effectiveness/how-they-work.html

Dowdy, D., & D'Souza, G. (2021). What is Herd Immunity and How Can We Achieve It With COVID-19?. Retrieved 8 August 2021, from https://www.jhsph.edu/covid-19/articles/achieving-herd-immunity-with-covid19.html

Evaluation of COVID-19 vaccine effectiveness. (2021). Retrieved 8 August 2021, from https://www.who.int/publications/i/item/WHO-2019-nCoV-vaccine_effectiveness-measurement-2021.1

Gerryn, C. (2021). Bankruptcies expected to increase 26% globally | Atradius. Retrieved 8 August 2021, from https://group.atradius.com/press/press-releases/bankrupticies-expected-to-grow-twenty-six-percent-in-2021.html

Irizarry, R. (2021). Chapter 28 Smoothing | Introduction to Data Science. Retrieved 8 August 2021, from https://rafalab.github.io/dsbook/smoothing.html

Kuadli, J. (2021). 11+ Mind-Blowing Bankruptcy Statistics for 2021. Retrieved 8 August 2021, from https://legaljobs.io/blog/bankruptcy-statistics/

Mathieu, E., Ritchie, H., Ortiz-Ospina, E., Roser, M., Hasell, J., & Appel, C. et al. (2021). A global database of COVID-19 vaccinations. Nature Human Behaviour, 5(7), 947-953. doi: 10.1038/s41562-021-01122-8

Wang, J., Jing, R., Lai, X., Zhang, H., Lyu, Y., Knoll, M., & Fang, H. (2020). Acceptance of COVID-19 Vaccination during the COVID-19 Pandemic in China. Vaccines, 8(3), 482. doi: 10.3390/vaccines8030482

ASSESSMENT RUBRIC

CRITERIA	MARKS					
	16-20	13-15	10-12	8-9	0-7	Comments
Methodology (30%)	Excellent in documenting the methodology. • Generates complete, clear and unambiguous requirements specification. • Identifies ambiguity in givens and states necessary assumptions. • Uses appropriate diagrams to describe implementation clearly including the design decisions. Overall contents comprehensively articulates all relevant and pertinent issues	Good in documenting the methodology. • Generates requirements specification with minor residual ambiguity. • Identifies ambiguity in givens however necessary assumptions are not fully stated. • Uses appropriate diagrams to describe implementation however contain a small number of errors, omissions or additions.	Satisfactory in documenting the methodology. • Generates requirements specification with some residual ambiguity. • Omits ambiguity in givens and states necessary assumptions ambiguously. • Uses appropriate diagrams to describe implementation however contain a number of errors, omissions or additions.	Weak in documenting the methodology. • Generates requirements specification with substantial ambiguity. • Omits ambiguity in given and necessary assumptions. • Uses inappropriate diagrams to describe implementation and contain large number of errors, omissions or additions. It is possible that the methodology is	Unsatisfactory in documenting the methodology. • Generates requirements specification with substantial ambiguity. • Omits ambiguity in given and necessary assumptions. • No diagrams to describe software architecture. It is possible that the methodology is weak or above in some areas and unsatisfactory in others. Unsatisfactory in two or more areas.	

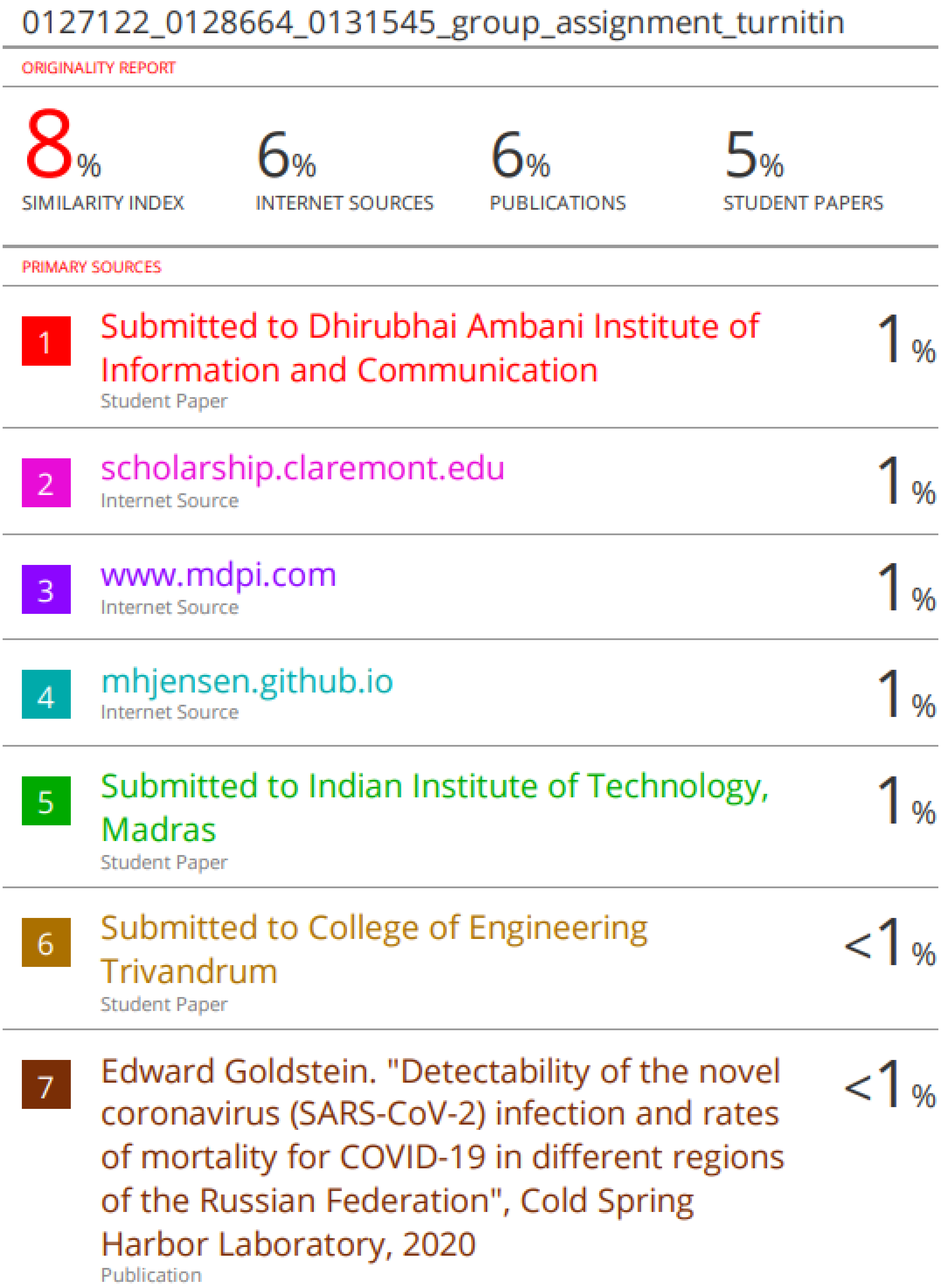
	related to the overall solution. All areas are at least good. May be outstanding is some areas and good in others and hence is on balance excellent. Good or above in all areas. Likely to contain minor errors, omissions or additions which prevent the methodology from being outstanding. Overall an excellent methodology.	It is possible that the methodology is outstanding or excellent in some areas and satisfactory in others but on balance is good. Satisfactory or above in all areas. Likely to contain a small number of errors, omissions or additions which prevent the methodology from being excellent. Overall a good methodology.	It is possible that the methodology is good or above in some areas and satisfactory in others. Weak in no more than two areas. Likely to contain a number of errors, omissions or additions which prevent the methodology from being good. Overall satisfactory.	satisfactory or above in some areas and unsatisfactory in others. Likely to be weak in more than three areas. It might be unsatisfactory in one area but no more. Likely to contain errors, omissions, additions, or misunderstandings which prevent the methodology from being satisfactory. Overall poor.	Likely to contain errors, omissions, additions, or misunderstandings which prevent the design model from being weak. May not be recognisable as a methodology, might have majors errors in content or a combination of the two.	
	16-20	**13-15**	**10-12**	**8-9**	**0-7**	
Implementation (30%)	Excellent in the implementation, comments, indentation, consistency and syntax. Correct following of object-oriented concepts. The work shows	Good in the implementation, comments, indentation, consistency and syntax. It is possible that the application is outstanding or	Satisfactory in the implementation, comments, indentation, consistency and syntax. It is possible that the application is good or above in some	Weak in the implementation, comments, indentation, consistency and syntax. It is possible that the application is satisfactory or	Unsatisfactory in the implementation, comments, indentation, consistency and syntax. It is possible that the application is weak	

	particular insight or originality in its approach. Excellent functionalities which identifies the underlying principles behind the problem. All areas are at least good. May be outstanding is some areas and good in others and hence is on balance excellent. Good or above in all areas. Likely to contain minor errors, omissions or additions which prevent the implementation from being outstanding. Overall an excellent implementation.	excellent in some areas and satisfactory in others but on balance is good. Satisfactory or above in all areas. Likely to contain a small number of errors, omissions or additions which prevent the implementation from being excellent. Overall a good implementation.	areas and satisfactory in others. Likely to be weak in no more than two areas. Likely to contain a number of errors, omissions or additions which prevent the implementation from being good. Overall a satisfactory implementation.	above in some areas and unsatisfactory in others. Likely to be weak in more than three areas. It might be unsatisfactory in one area but no more. Likely to contain errors, omissions, additions, or misunderstandings which prevent the implementation from being satisfactory. Still recognisable as an object-oriented application of the problem in focus. Overall poor but satisfactory.	or above in some areas and unsatisfactory in others. Unsatisfactory in two or more areas. Likely to contain errors, omissions, additions, or misunderstandings which prevent the implementation from being weak. May not be recognisable as an object-oriented application, might have majors errors in implementation or a combination of the two.	
Results (20%)	**16-20**	**13-15**	**10-12**	**8-9**	**0-7**	
	Excellent in the results discussion. The explanation and justification of how it meets specified requirements shows	Good in the results discussion. The explanation and justification of how it meets specified requirements shows	Satisfactory in the results discussion. The explanation and justification of how it meets specified	Weak in the areas of the results discussion. The explanation and justification of how it meets specified	Unsatisfactory in the areas of the results discussion. It conveys little understanding of solution of the	

	outstanding insight into the issues involved and alternatives available. The presentation clearly and concisely demonstrates a deep understanding of the project implementation. May be outstanding is some areas and good in others and hence is on balance excellent. Good or above in all areas. Likely to contain minor errors, omissions or additions which prevent the results discussion from being outstanding. Overall an excellent results discussion.	good understanding into the issues involved and alternatives available. The presentation demonstrates a good understanding of the project implementation. It is possible that the presentation is outstanding or excellent in some areas and satisfactory in others but on balance is good. Satisfactory or above in all areas. Likely to contain a small number of errors, omissions or additions which prevent the results discussion from being excellent. Overall a good results discussion.	requirements covers relevant aspects into the issues involved and alternatives available, but is not outstanding in any respect. It is possible that the presentation is good or above in some areas and satisfactory in others. Likely to be weak in no more than two areas. Likely to contain a number of errors, omissions or additions which prevent the results discussion from being good. Overall a satisfactory results discussion.	required is inadequate. It is possible that the presentation is satisfactory or above in some areas and unsatisfactory in others. Likely to be weak in more than three areas. It might be unsatisfactory in one area but no more. Likely to contain errors, omissions, additions, or misunderstandings which prevent the results discussion from being satisfactory. Overall poor but satisfactory.	problem or no useful explanation and justification of how it meets specified requirements. It is possible that the presentation is weak or above in some areas and unsatisfactory in others. Unsatisfactory in two or more areas. Likely to contain errors, omissions, additions, or misunderstandings which prevent the results discussion from being weak.	
Communication (15%)						

	16-20	13-15	10-12	8-9	0-7	Comments
	Excellent, well-directed presentation, logically and coherently structured. It is free or almost free grammatical errors. The format is clear and consistent with appropriate use of headings and paragraphs. English usage is easily understandable. References and quotations are utilized appropriately to indicate sources.	Good presentation, logically structured. There are occasional spelling and grammatical errors, but the reader does not struggle to interpret the writer's intended meaning. The writing would benefit from the use of organizational tools (e.g. headings, paragraphs) and more consistent use of references to sources.	Satisfactory presentation, well structured. It contains number of spelling and grammatical errors, but the reader does not struggle to interpret the writer's intended meaning. It is possible that the use of organizational tools (e.g. headings and paragraphs) are good or above in some areas and satisfactory in others.	Weak presentation and structure. Spelling and grammatical errors force the reader to struggle to determine the intended meaning. Organizational tools such as headings, paragraphs are used inconsistently. References are not used properly to indicate the sources of material.	Unsatisfactory presentation and structure. Numerous spelling and grammatical errors and a lack of clear consistent organization interfere with the writer's ability to communicate to the key points. The reader frequently cannot determine the intended meaning. There are no references to indicate material taken from other sources.	

Turnitin Report



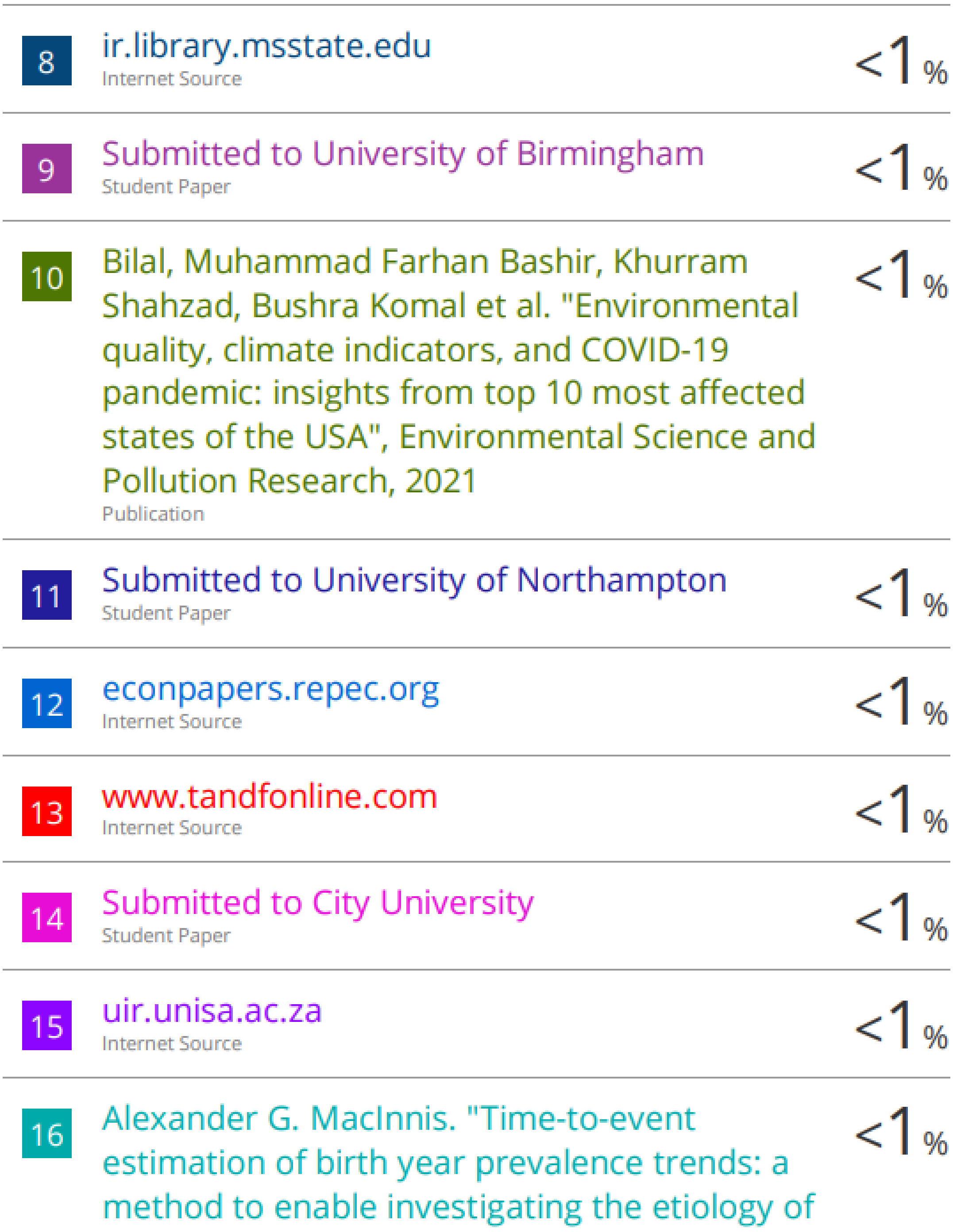
0127122_0128664_0131545_group_assignment_turnitin

ORIGINALITY REPORT

8%	6%	6%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Dhirubhai Ambani Institute of Information and Communication Student Paper	1%
2	scholarship.claremont.edu Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	mhjensen.github.io Internet Source	1%
5	Submitted to Indian Institute of Technology, Madras Student Paper	1%
6	Submitted to College of Engineering Trivandrum Student Paper	<1%
7	Edward Goldstein. "Detectability of the novel coronavirus (SARS-CoV-2) infection and rates of mortality for COVID-19 in different regions of the Russian Federation", Cold Spring Harbor Laboratory, 2020 Publication	<1%



8	ir.library.msstate.edu Internet Source	<1 %
9	Submitted to University of Birmingham Student Paper	<1 %
10	Bilal, Muhammad Farhan Bashir, Khurram Shahzad, Bushra Komal et al. "Environmental quality, climate indicators, and COVID-19 pandemic: insights from top 10 most affected states of the USA", Environmental Science and Pollution Research, 2021 Publication	<1 %
11	Submitted to University of Northampton Student Paper	<1 %
12	econpapers.repec.org Internet Source	<1 %
13	www.tandfonline.com Internet Source	<1 %
14	Submitted to City University Student Paper	<1 %
15	uir.unisa.ac.za Internet Source	<1 %
16	Alexander G. MacInnis. "Time-to-event estimation of birth year prevalence trends: a method to enable investigating the etiology of	<1 %

	childhood disorders including autism", Cold Spring Harbor Laboratory, 2021 Publication	
17	Francisco Benita, Francisco Gasca-Sanchez. "The main factors influencing COVID-19 spread and deaths in Mexico: A comparison between Phases I and II", Cold Spring Harbor Laboratory, 2021 Publication	<1 %
18	José J. Ortiz-García, Seósamh B. Costello, Martin S. Snaith. "Derivation of Transition Probability Matrices for Pavement Deterioration Modeling", Journal of Transportation Engineering, 2006 Publication	<1 %
19	ueaeprints.uea.ac.uk Internet Source	<1 %
20	"Advances in Computing and Network Communications", Springer Science and Business Media LLC, 2021 Publication	<1 %
21	arxiv.org Internet Source	<1 %

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	Off		