COVID-19 PREDICTION

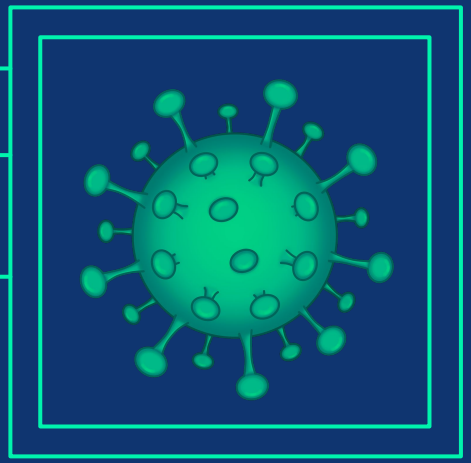
Team: Alisiya Balayan, Bishoy Abdelmalik, Arian Dehghani, Ariel Kohanim, Natalie Weingart, Sergio Ramirez

TABLE OF CONTENTS

TABLE OF CONTENTS

01

ABOUT THE PROJECT



OUR PROJECT

We are examining datasets about the coronavirus pandemic, and trying to predict when the pandemic will come to an end using data about vaccinations, herd immunity, and hospitalizations

DATA MINING GOAL

Our goal is to predict when 70% of the U.S. population will have CV-19 immunity, which will assume herd immunity, using the datasets acquired.

02

DATASETS

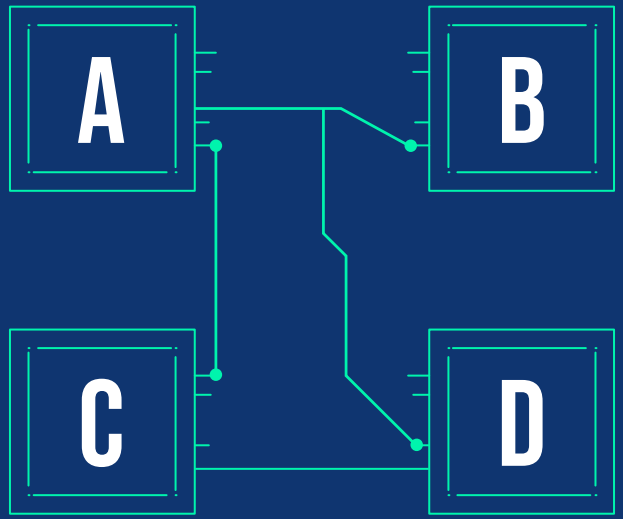
OUR DATASETS

COVID-19 WORLD VACCINATION PROGRESS

This dataset tells us about: which country is using what vaccine, which country's vaccination programme is more advanced, and where the rate of vaccinated people per day is higher in terms of percent from the entire population.

COUNTRIES POPULATION BY YEAR 2020

This dataset provides us with the world population and top 20 countries' live clock. It contains population data for the past, present and future.



NOVEL CORONAVIRUS 2019 DATASET

This dataset has daily level information on the number of affected cases, deaths and recovery from COVID-19. It provides us with data about each country and their cases, deaths and recoveries.

COVID-19 DAILY VACCINATION

This data set contains vaccination data for countries showing how many people are vaccinated daily.

03

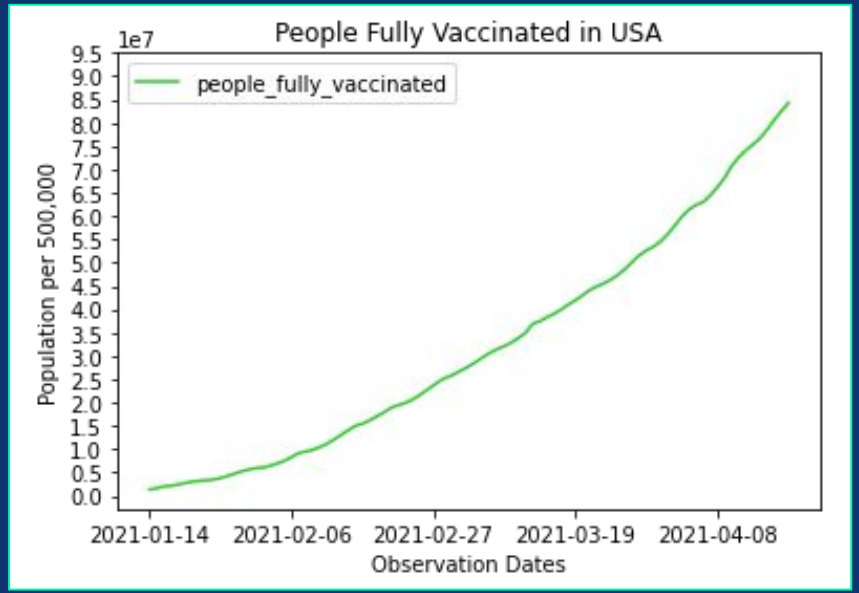
DATA EXPLORATION

DATA EXPLORATION

FULLY VACCINATED INDIVIDUALS IN THE USA

The mean, standard deviation, min, max, and the quartiles are found through Dataset A

```
count          91.00000
mean     31929907.90110
std      24208365.96282
min       1342086.00000
25%       9679222.00000
50%      27795980.00000
75%      49418470.50000
max      84263408.00000
dtype: object
```

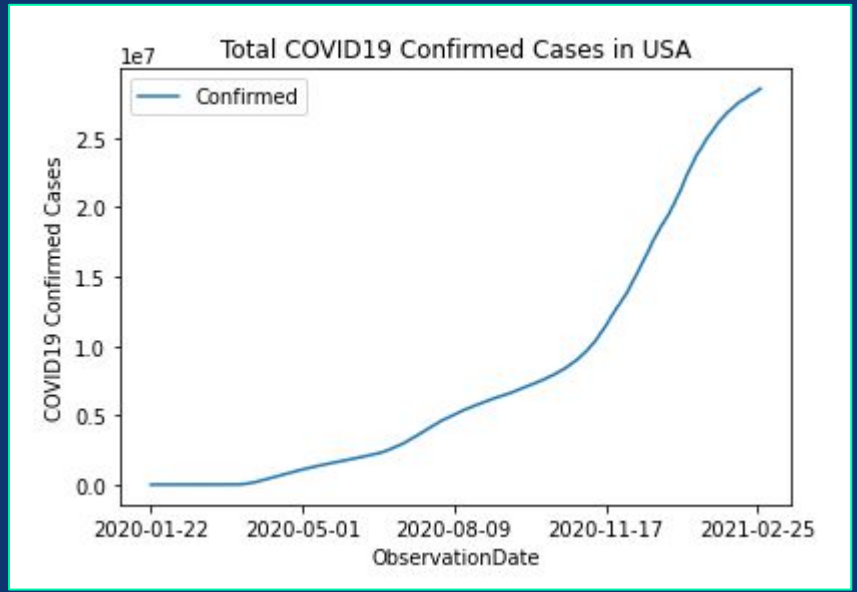


DATA EXPLORATION

USA CONFIRMED COVID-19 CASES

The mean, standard deviation, min, max, and the quartiles are found through Dataset B

count	21462.00000
mean	149160.90355
std	309144.87948
min	0.00000
25%	2710.25000
50%	35865.50000
75%	156865.25000
max	3563578.00000
dtype:	object

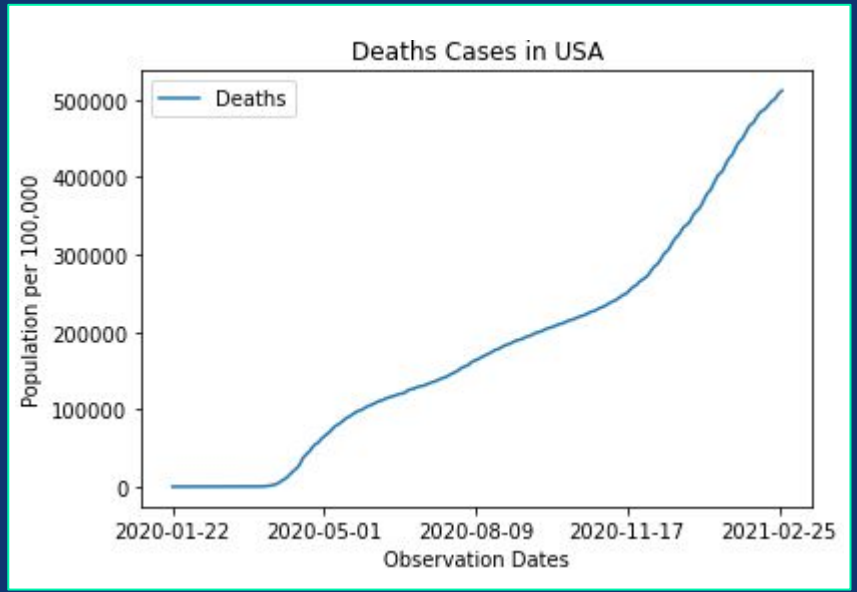


DATA EXPLORATION

USA DEATHS CAUSED BY COVID-19

The mean, standard deviation, min, max, and the quartiles are found through Dataset B

	Deaths
count	403.000000
mean	179213.498759
std	142776.472362
min	0.000000
25%	66075.000000
50%	164041.000000
75%	253826.000000
max	511994.000000

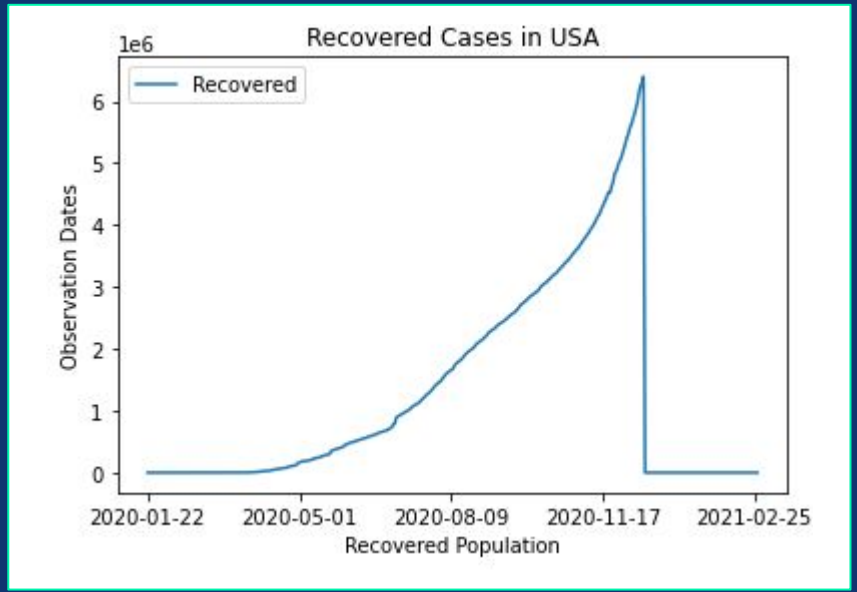


DATA EXPLORATION

USA COVID-19 RECOVERED CASES

The mean, standard deviation, min, max, and the quartiles are found through Dataset B

```
count          403.00000
mean       1249059.44417
std        1630170.41782
min              0.00000
25%              3.00000
50%         391508.00000
75%        2292820.50000
max        6399531.00000
dtype: object
```

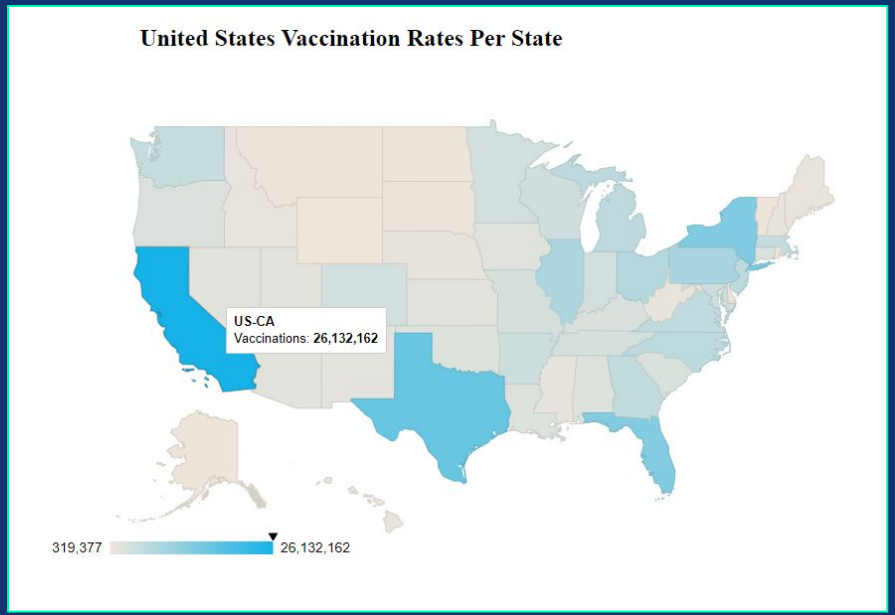


DATA EXPLORATION

USA DAILY VACCINE BY STATE

The mean, standard deviation, min, max, and the quartiles are found through Dataset D

count	50.00000
mean	4089885.82000
std	4695572.64397
min	319377.00000
25%	1110997.00000
50%	2653070.00000
75%	5102783.50000
max	26132162.00000
dtype:	object



04

STATISTICAL TECHNIQUES

STATISTICAL TECHNIQUES

Our group looked at multiple statistical techniques to evaluate the datasets, those techniques are:

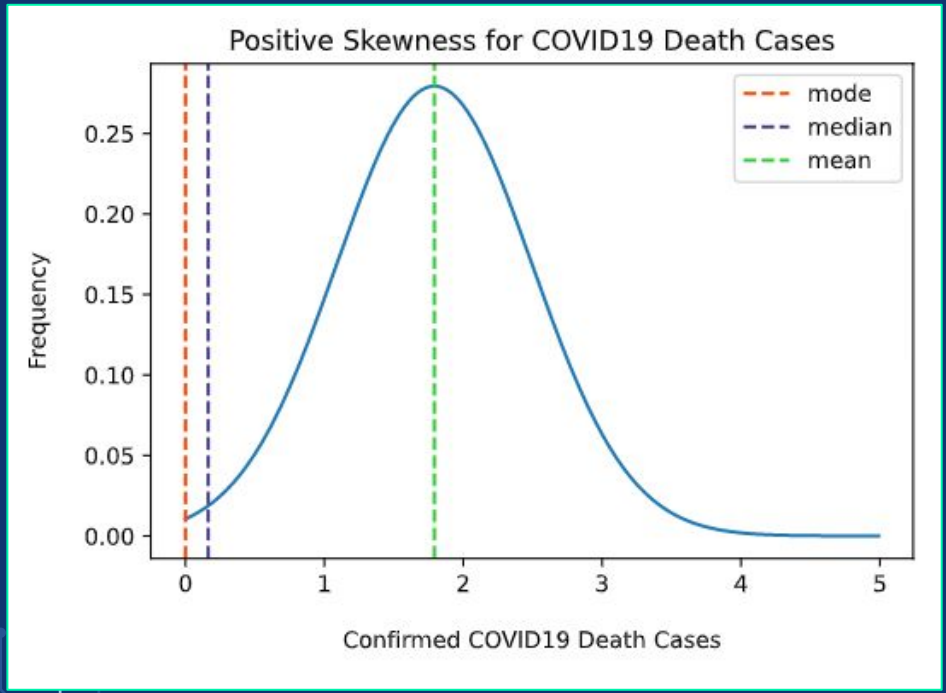
1. Feature skewness and looking at dataset characteristics such as mean, mode, median and standard deviation for:
   a. Confirmed Positive CV-19 cases
   b. Recovered CV-19 cases
   c. Death by CV-19 cases

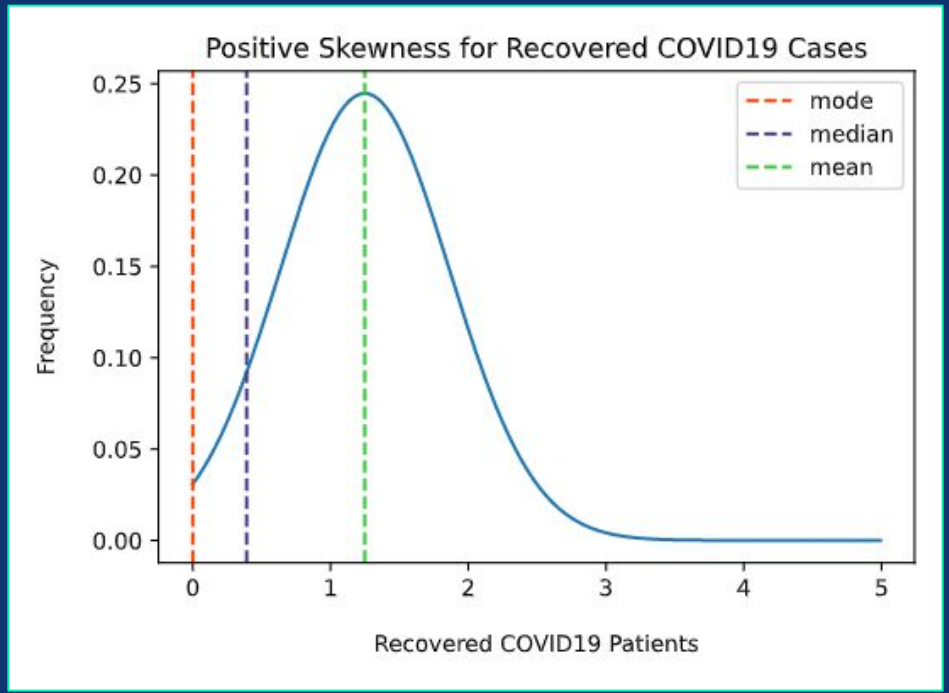
2. Correlations between attributes and using heatmaps to model it.

3. Kernel Density Estimation to estimate the probability density function of the random variables to make inferences about the population based on the finite data sample that we have.

STATISTICAL TECHNIQUES

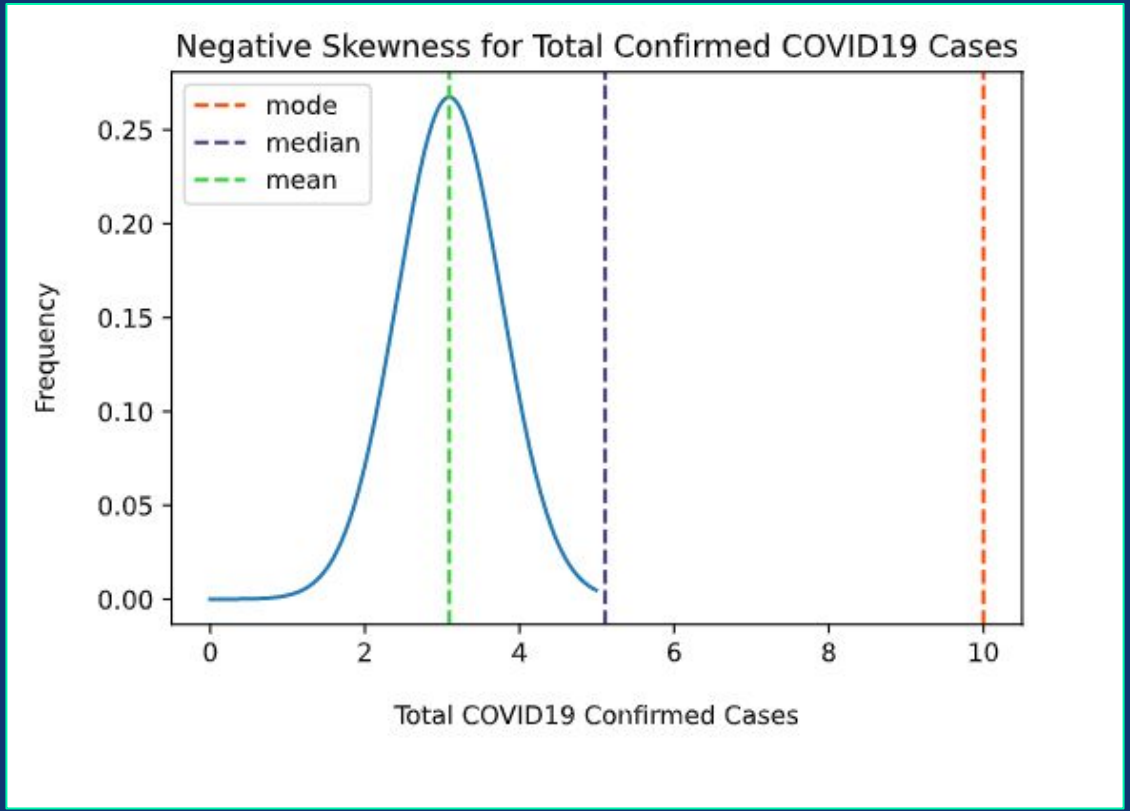
FEATURE SKEWNESS





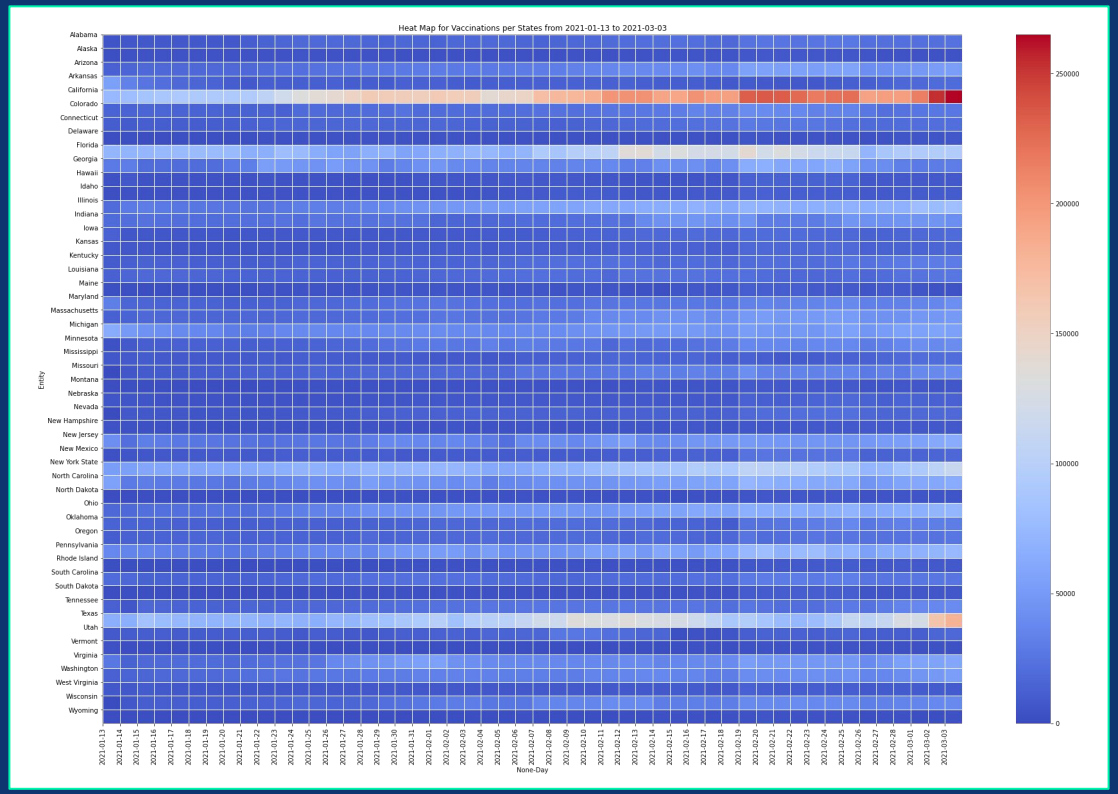
STATISTICAL TECHNIQUES

FEATURE SKEWNESS



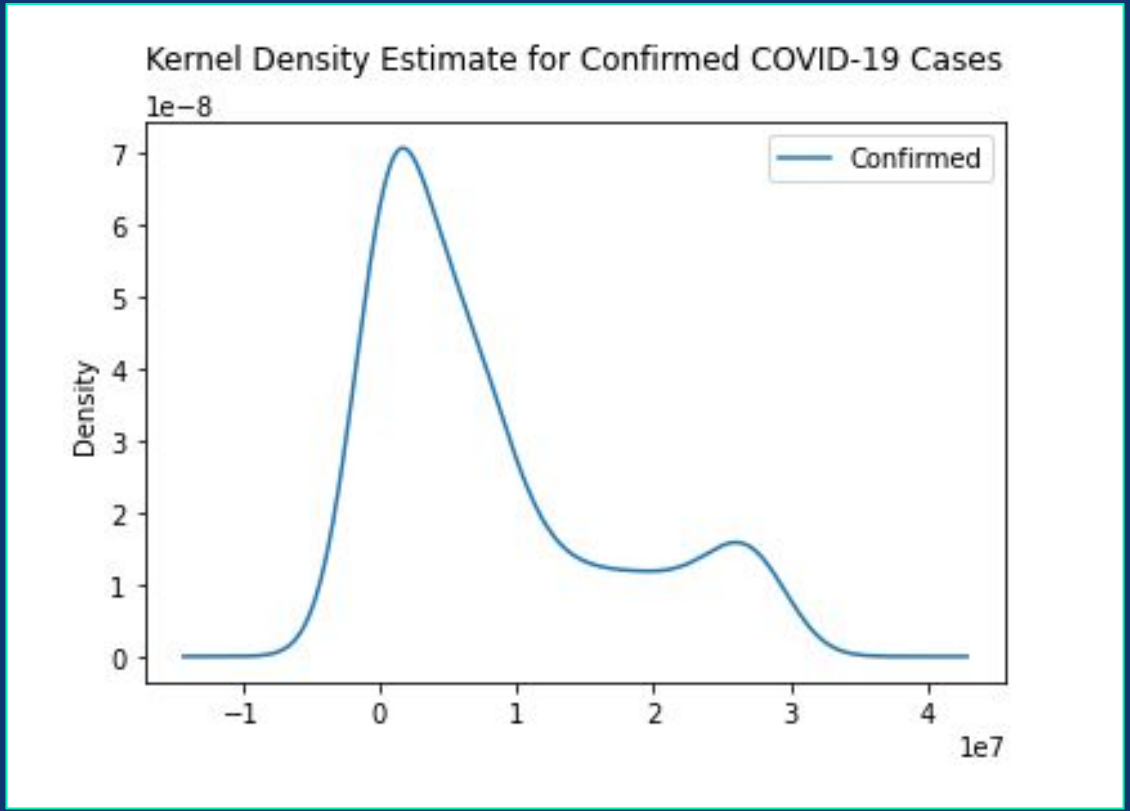
STATISTICAL TECHNIQUES

CORRELATIONS BETWEEN ATTRIBUTES



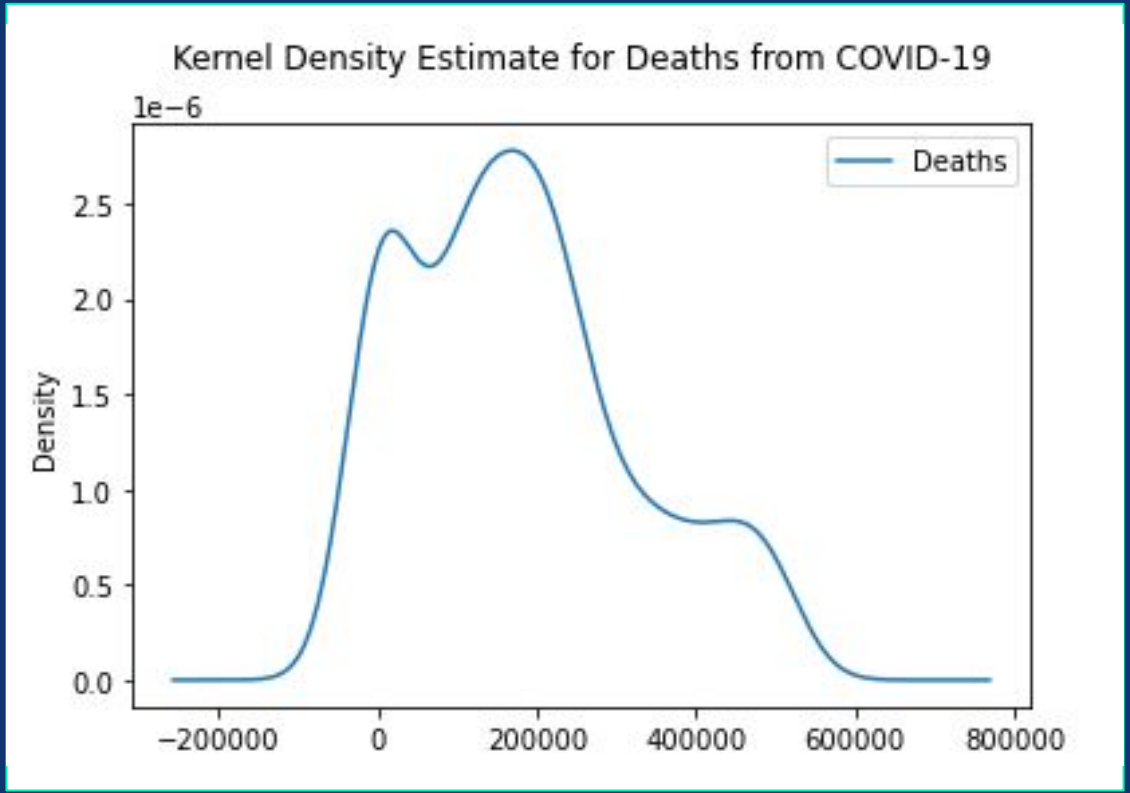
STATISTICAL TECHNIQUES

KERNEL DENSITY ESTIMATION



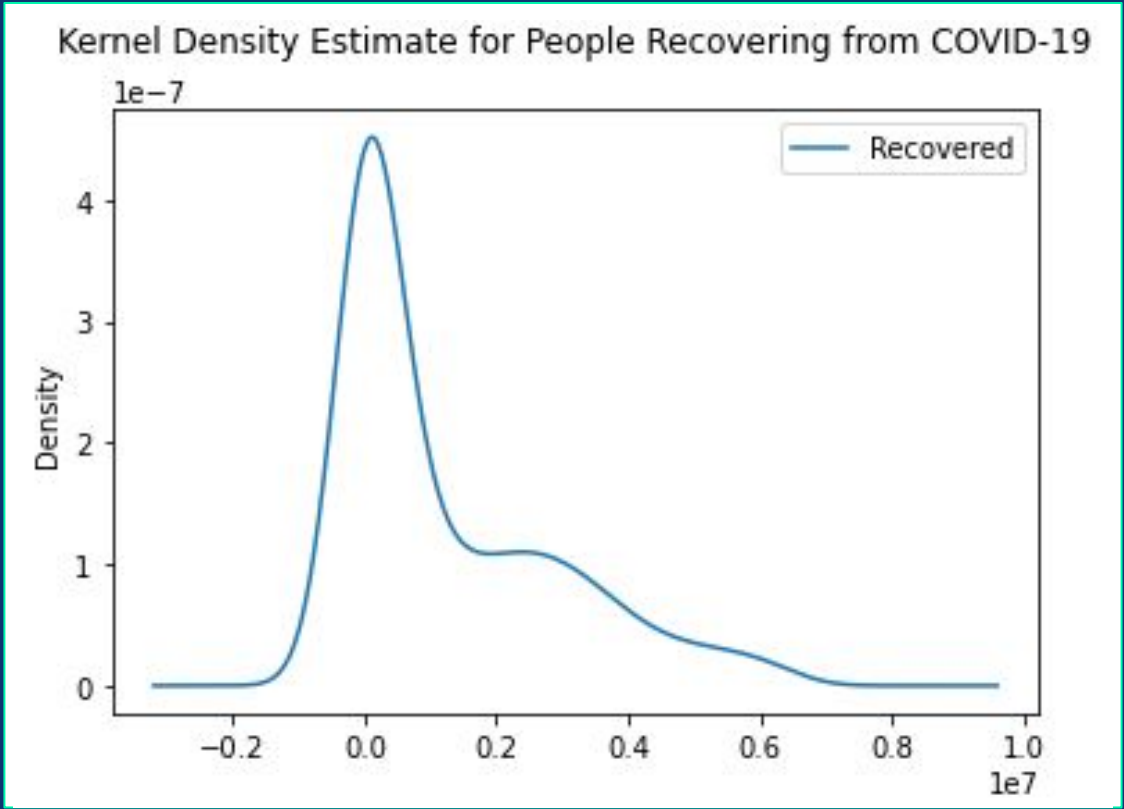
STATISTICAL TECHNIQUES

KERNEL DENSITY ESTIMATION



STATISTICAL TECHNIQUES

KERNEL DENSITY ESTIMATION



05

DATA VISUALIZATIONS

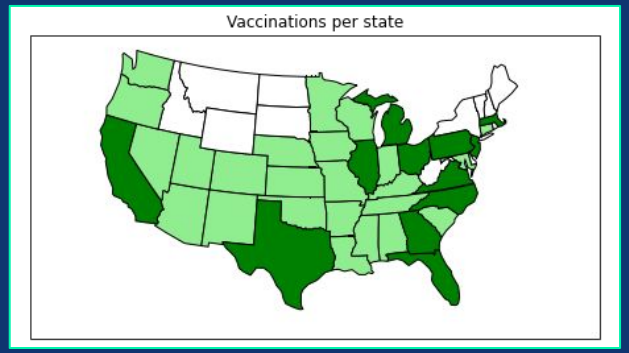
DATA VISUALIZATION

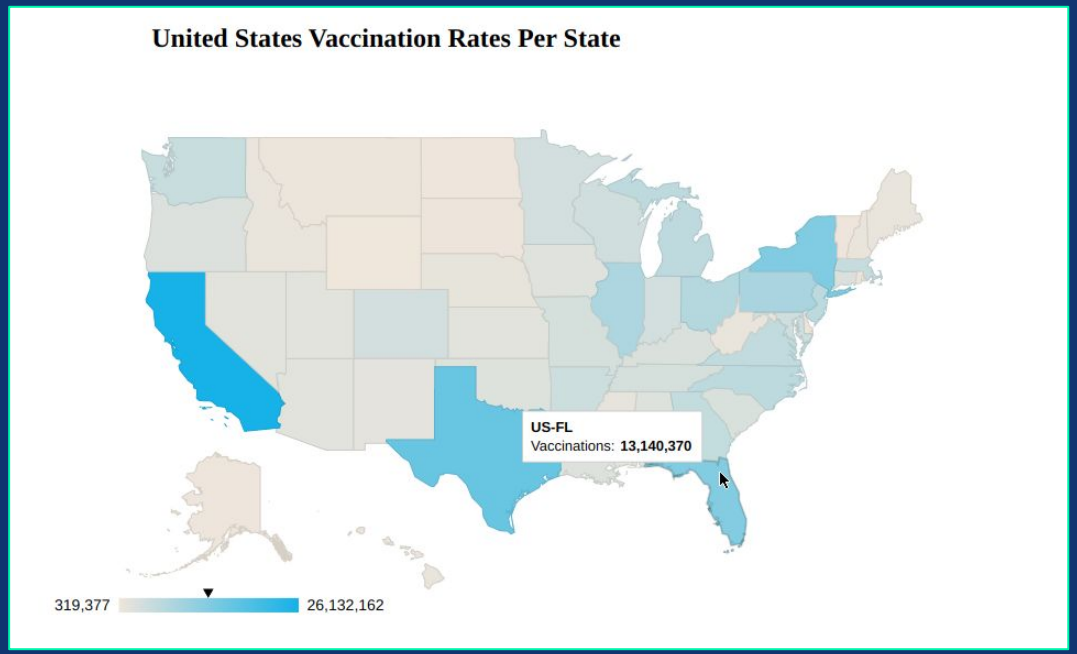
GOOGLE CHARTS

Google charts was used for creating an interactive map that shows the USA vaccination progress per state.

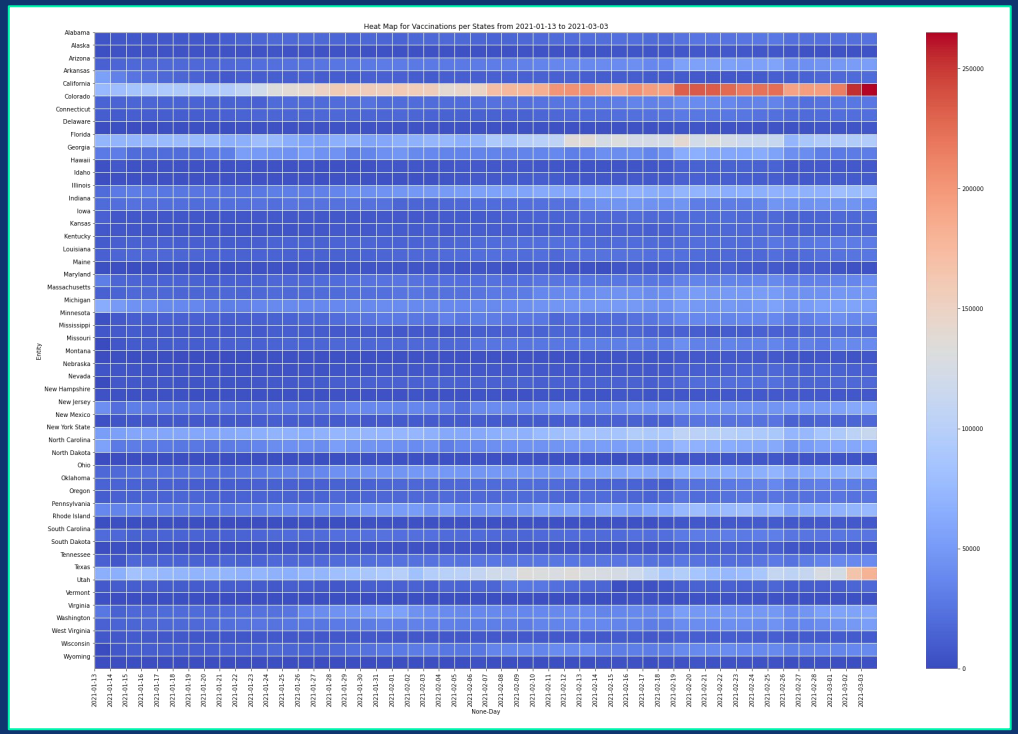
CARTOPY

The Cartopy library was used for easy visualization of Geomaps.





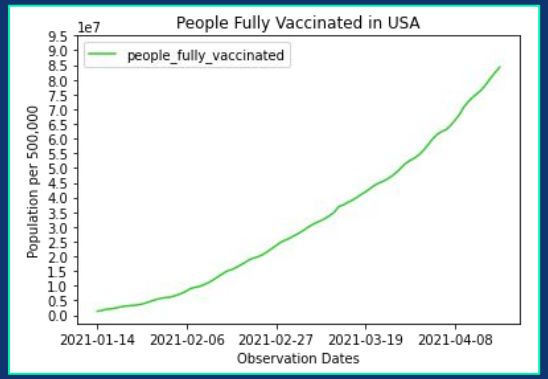
DATA VISUALIZATION



SEABORN

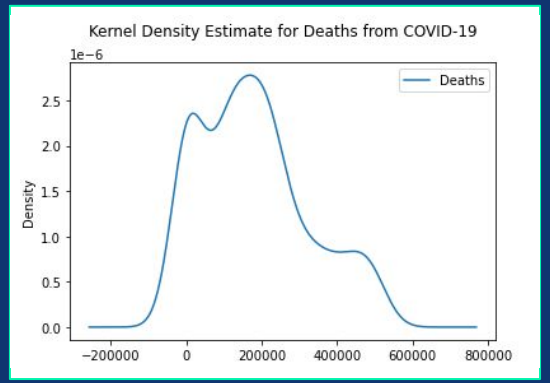
The Seaborn library was used to create heat maps to show the correlation between the attributes

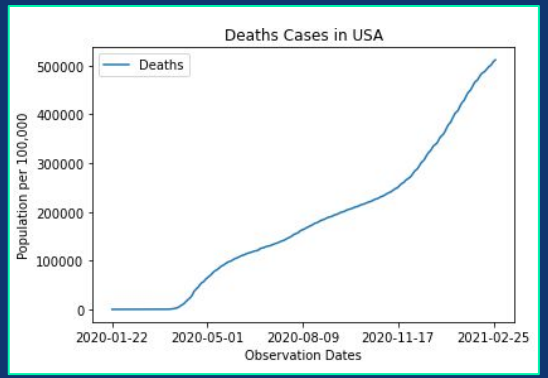
DATA VISUALIZATION



MATPLOT

The team used the famous matplot library for creating every other 2 dimensional graph such as skewness, rate of fatalities, and density





06

NEXT STEPS

DATA CLEANING AND TRANSFORMATION (NEXT STEP)

Here we can see one of datasets, Countries population by year 2020.csv, has N.A. values in addition to NaN. We can choose to drop these rows or fill in those values with other data such as 0 for NaN.

Example of a process we may follow is that we will drop all columns containing N.A., parse percentage values into float types, and fill NaN values as 0. We would also like to note that we will normalize our data as well.

NaN count by attribute for country_vaccinations.csv

230	Montserrat	4991	0.06 %	3	50	100	NaN	N.A.	N.A.	10 %	0.00 %
231	Falkland Islands	3458	3.05 %	103	0	12170	NaN	N.A.	N.A.	66 %	0.00 %
232	Niue	1624	0.68 %	11	6	260	NaN	N.A.	N.A.	46 %	0.00 %
233	Tokelau	1354	1.27 %	17	136	10	NaN	N.A.	N.A.	0 %	0.00 %
234	Holy See	801	0.25 %	2	2003	0	NaN	N.A.	N.A.	N.A.	0.00 %

196	Aruba	106675	0.43	452	593	180	201.0	1.9	41	44.0	0.00
197	Tonga	105449	1.15	1201	147	720	-800.0	3.6	22	24.0	0.00
198	U.S. Virgin Islands	104456	-0.15	-153	298	350	-451.0	2.0	43	96.0	0.00
199	Seychelles	98224	0.62	608	214	460	-200.0	2.5	34	56.0	0.00
200	Antigua and Barbuda	97764	0.84	811	223	440	0.0	2.0	34	26.0	0.00

```
NaN value by attribute
 country                                0
iso_code                                0
date                                    0
total_vaccinations                     14
people_vaccinated                      15
people_fully_vaccinated                29
daily_vaccinations_raw                 24
daily_vaccinations                      1
total_vaccinations_per_hundred         14
people_vaccinated_per_hundred          15
people_fully_vaccinated_per_hundred    29
daily_vaccinations_per_million          1
vaccines                                0
source_name                             0
source_website                          0
dtype: int64
```