STA160 Final Project

Group Members

Sophia Tierney – sftierney@ucdavis.edu, Student ID: 917308455

Aris Briones – anbriones@ucdavis.edu, Student ID: 917170309

HungHsu (Allen) Chen – chhchen@ucdavis.edu, Student ID: 915145076

Amay Kharbanda – akharbanda@ucdavis.edu, Student ID: 914798754

June 9, 2021

Abstract

This project is about the administration of vaccines for COVID-19. It focuses on the total and weekly administer of three brands of COVID vaccine, Moderna, Pfizer, and J&J, for each state. The data is collected from the CDC and several data sets were merged into the table that has been used in this project by the weeks and states. By analyzing the table, it shows that the new cases and deaths of COVID-19 are decreasing over time as the amount of people getting vaccinated increases. Meanwhile, by fitting the data into the regression models such as multiple linear regression model and random forest regression model, the data can be used to predict the vaccine allocation amount. As a result, both the multiple linear regression model and the random forest regression model that uses COVID cases, deaths, and population has the best performance with a 98.8 percent accuracy on predicting the vaccine allocation amount. On the other hand, by applying the dimension reduction with PCA and using several different clustering methods, the data shows some clustering. This could be due to the different vaccine brands.

1 Introduction

As COVID-19 vaccines are finally being administered to the population, there are a lot of people in different states have received the vaccines. Since we are all wondering when we will be able to go back to the normal without wearing a mask, it is important to understand how the vaccines popularize in different states and between different races. It is also important to understand the new case and new death trends since what we want is not only that everyone get vaccinated – but to also eliminate the threat of COVID-19 entirely. In order to understand about what is going on with the vaccines and COVID-19, we have gathered the data from CDC that are about the population that have fully vaccinated in different states. There are 3 brands of vaccines that are considered - Moderna, Pfizer, and J&J. We have combined the data for all three brands of vaccine as our table that used to analyze and to model.

2 Data Description

We've gathered data from multiple different CDC sources that we found would bring different and interesting perspectives to our analysis and exploration. In our Python data analysis and modeling, we worked with many data frames – all gathered, cleaned, and combined neatly into a total of 4. All of the data is continuously updated, either daily or weekly by the CDC.

The **df** dataframe contains weekly data of vaccination allocations per state, for each of the 3 major COVID-19 vaccine manufacturers – Pfizer [1], Moderna [2], and Janssen (Johnson & Johnson).[3] Along with the vaccine allocations per state, we combined a column containing the weekly numbers of COVID-19 deaths per state. For the Janssen second dose column, we have NA values for all weeks because it is a single dose only vaccine. For the other two manufacturers, Pfizer and Moderna, the number of allocations per state for each week is exactly the same for the first dose as it is second dose. This makes sense because the manufacturers only allocated the amount of vaccines they could to each state because they had to make sure they had enough doses for each state population's second dose. Hence, throughout our analysis we consider the first dose column as both the first and second doses, for simplicity of computations.

From the CDC COVID-19 Data Tracker website [4], we pulled observations from US Trend Data. This was loaded into the **state_data** dataframe, with columns of interest consisting of cumulative cases and deaths (one column for the total numbers for the entire population of each state, and one column for the total numbers scaled down per $100K$ people, daily new cases and deaths, and average new cases and deaths. This data is updated daily, so we have values for each day ranging from Jan 22, 2020 to the present (Jun 4, 2021 today). However, because our primary interest is in the analysis of vaccine data, we ignore the state data from 2020, and keep only 2021 data in the **state_data_21** dataframe. We concatenated the 2021 state data with vaccine allocation data from **df** to create **full_df**, which contains weekly data for all the states

and narrowed down the time series to dates ranging from Mar 1, 2021 (the first date vaccines began to be distributed to all the states) to present.

Since our **state_data_21** dataframe only tells us about the weekly vaccine allocations, we searched for vaccine data that would be more comprehensive and help us gain insight to more than just the allocations alone. Also pulled from the CDC COVID-19 Data Tracker website [4], **df_vacc_totals** dataframe contains columns of cumulative totals for each state including:

- the estimated population of the state according to 2019 Census data
- total number of vaccines administered to the population of the state and total number of vaccines administered per 100K - independent of brand/dosing break down. each value in this variable counts 2 doses for Pfizer and Moderna as 1 vaccine, or 1 dose of Janssen
- total number of each Moderna, Pfizer, and Janssen doses administered
- cumulative number of Moderna, Pfizer, and Janssen vaccines distributed AND administered to each state's population per 100K people (this takes into account first doses and second doses)
- total number of the population who is fully vaccinated (completed full set of doses) - not people who have only had their first doses for Moderna, Pfizer, and Janssen

All data in this dataset is daily, and contains a breakdown for each US State.

Finally, we found it difficult to find data that contains information about proportions of ethnic groups vaccinated, so we created the **vacc_demo** dataframe. The **vacc_demo** data is accessed through an API to the JSON file found here [5]. This contains insightful information about the percentages of administered doses and the percentages of fully vaccinated population in the US pertaining to the following demographic categories:

- Ethnic Groups
  - non-Hispanic Asian (NHAsian)
  - non-Hispanic Native Hawaiian, Other or Pacific Islander (NHNHOPI)
  - non-Hispanic American Indian or Alaska Native (NHAIAN)
  - non-Hispanic Black (NHBlack)
  - Hispanic
  - non-Hispanic multiracial or other (NHMult_Oth)
  - non-Hispanic White (NHWhite)

- Age Groups
  - $18-24$
  - $18-29$
  - $30-39$
  - $40-49$
  - $50-64$
  - $65-74$
  - $75+$
- Sex
  - Male
  - Female

At the end of gathering and combining all of our data, we have a total of 4 data frames used in our data analysis and modeling which are: **full_df**, **df_vacc_totals**, **state_data21**, **vacc_demo**.

2.1 Exploratory Data Analysis

2.1.1 Understanding the table full_df

Data Types

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2219 entries, 0 to 2218
Data columns (total 18 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   state                        2219 non-null   object
 1   week                         2219 non-null   datetime64[ns]
 2   week_num                     2219 non-null   UInt32
 3   company                      2219 non-null   object
 4   1st_dose                     2219 non-null   object
 5   2nd_dose                     1708 non-null   object
 6   covid_19_deaths              2219 non-null   object
 7   cum_cases                    2219 non-null   int64
 8   daily_new_cases              2219 non-null   int64
 9   daily_new_deaths             2219 non-null   int64
 10  avg_new_cases                2219 non-null   int64
 11  cum_deaths                   2219 non-null   int64
 12  avg_new_deaths               2219 non-null   int64
 13  cum_new_cases_per100K        2219 non-null   float64
 14  cum_new_deaths_per100K       2219 non-null   float64
 15  date                         2219 non-null   datetime64[ns]
 16  historical_new_total_cases   0 non-null      float64
 17  historical_new_total_deaths  0 non-null      float64
dtypes: UInt32(1), datetime64[ns](2), float64(4), int64(6), object(5)
memory usage: 322.9+ KB
```

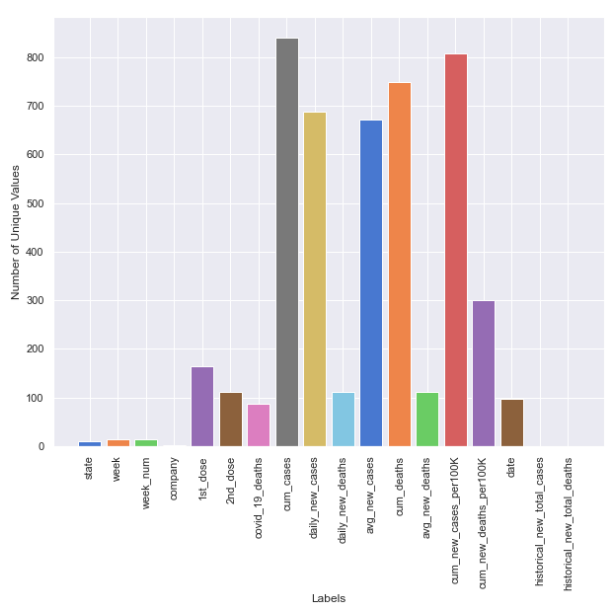
Figure 1: Data types

These are the 18 variables in the table **full_df**.

- **state** includes each state in the U.S..

- **week** includes the date of the first day of the week.
- **week_num** includes the count of the week.
- **company** includes the brands of vaccine.
- **1st_dose** includes the number of first dose administered.
- **2nd_dose** includes the number of second dose administered.
- **covid_19 death** includes the number of deaths caused by COVID-19 that day.
- **cum_cases** includes the cumulative COVID-19 cases.
- **daily_new_cases** includes the new COVID-19 cases in that day.
- **daily_new_deaths** includes the new death caused by COVID-19 in that day.
- **avg_new_cases** includes the average new cases of COVID-19 in a week.
- **cum_deaths** includes the cumulative death caused by COVID-19.
- **avg_new_deaths** includes the average new death caused by COVID-19 in a week.
- **cum_new_cases_per100K** includes cumulative new COVID-19 cases per 100k population.
- **cum_new_deaths_per100K** includes cumulative new death caused by COVID-19 per 100k population.
- **date** includes the date of the day
- **historical_new_total_cases** includes only null values so it got removed.
- **historical_new_total_deaths** includes only null values so it got removed.

Unique and Null Values



(a) Unique Value Count

state	0
week	0
week_num	0
company	0
1st_dose	0
2nd_dose	511
covid_19_deaths	0
cum_cases	0
daily_new_cases	0
daily_new_deaths	0
avg_new_cases	0
cum_deaths	0
avg_new_deaths	0
cum_new_cases_per100K	0
cum_new_deaths_per100K	0
date	0
historical_new_total_cases	2219
historical_new_total_deaths	2219

dtype: int64

(b) Null values in the data

There are a lot of null values under **2nd_dose** because J&J only requires one dose, so all the rows for J&J have a null value for the **2nd_dose**. The **historical_new_total_cases** and **historical_new_total_deaths** only include null values and got removed later.

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
week_num	2219.0	1.476972e+01	3.830100e+00	9.00	12.000	14.00	18.00	22.00
cum_cases	2219.0	1.040029e+06	1.090859e+06	40684.00	330970.000	534575.00	1109958.00	3690868.00
daily_new_cases	2219.0	1.226598e+03	1.456476e+03	0.00	263.000	702.00	1600.50	8925.00
daily_new_deaths	2219.0	2.146237e+01	2.903718e+01	0.00	2.000	9.00	28.00	159.00
avg_new_cases	2219.0	1.263990e+03	1.394442e+03	21.00	320.000	802.00	1514.00	6651.00
cum_deaths	2219.0	1.810222e+04	1.821386e+04	358.00	5972.000	10920.00	20368.00	62473.00
avg_new_deaths	2219.0	2.232853e+01	2.765358e+01	0.00	5.000	9.00	33.00	162.00
cum_new_cases_per100K	2219.0	9.972535e+01	6.135628e+01	10.31	48.415	83.50	145.71	268.03
cum_new_deaths_per100K	2219.0	1.612014e+00	1.113602e+00	0.00	0.860	1.31	2.05	6.57

Figure 2: Statistical description of the data

Cleaning the Data Set

For cleaning this data set, null values were replaced with 0's (mainly for the J&J's 2nd dose) and the last two columns were dropped as they had no values at all. Moreover, there were no duplicate observations in

our data set. Besides these changes, there was not much data cleaning required.

Boxplots

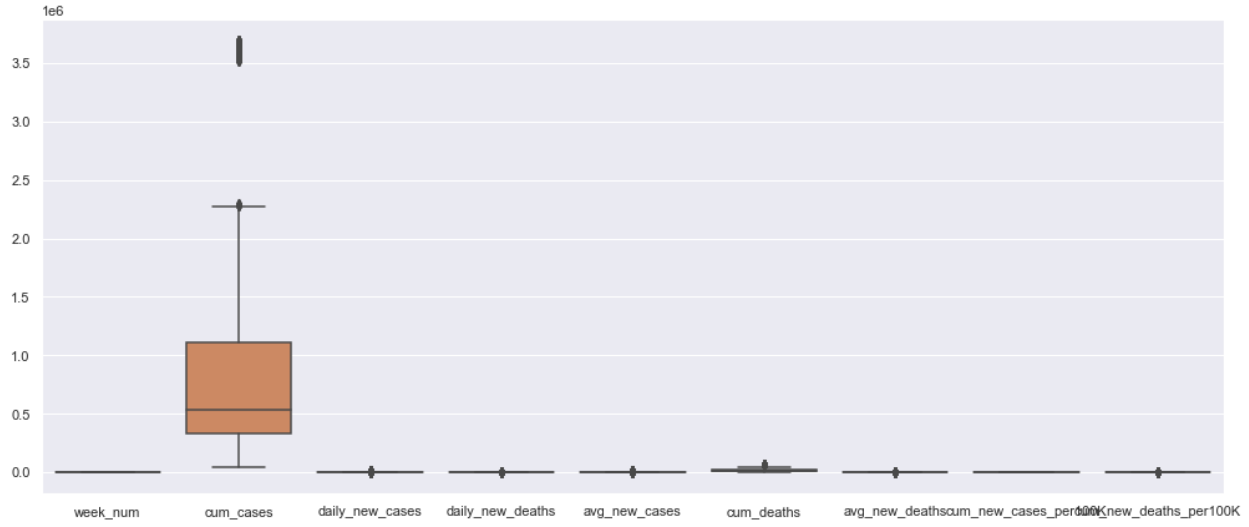
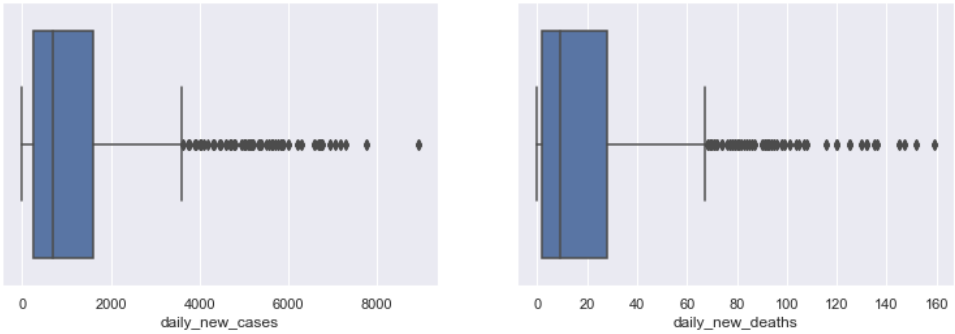


Figure 3: Boxplots of the data

We can see that the variables vary highly, so it might be useful to observe selected boxplots of the variables.

daily_new_cases and daily_new_deaths

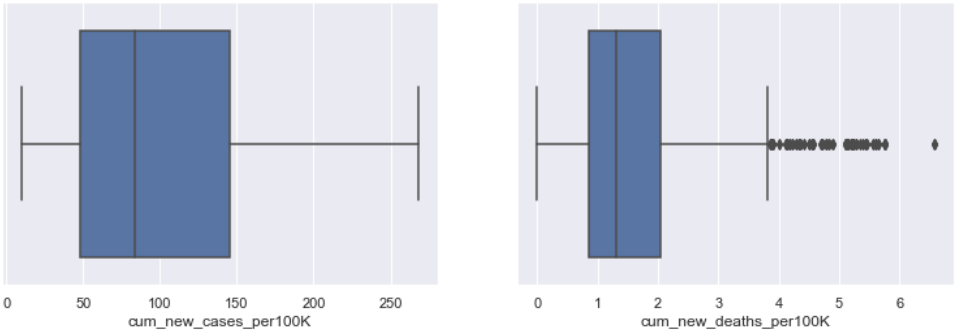


(a) daily_new_cases

(b) daily_new_deaths

Daily new cases were mainly spread from a few hundreds to 2 thousands cases, and sometimes there were more than 3 thousands new cases daily. There were 0 to 60 new deaths daily that were caused by COVID-19, and sometimes there were more than 60 deaths daily caused by COVID.

"cum_new_cases_per100K" and "cum_new_deaths_per100K"



(a) cum_new_cases_per100K (b) cum_new_deaths_per100K

There are mainly about 50 to 150 cumulative new cases per 100K population. There are mainly about 1 to 2 cumulative new deaths that are caused by COVID-19 per 100K population.

2.1.2 Visualizing Relationships in the Data

Count

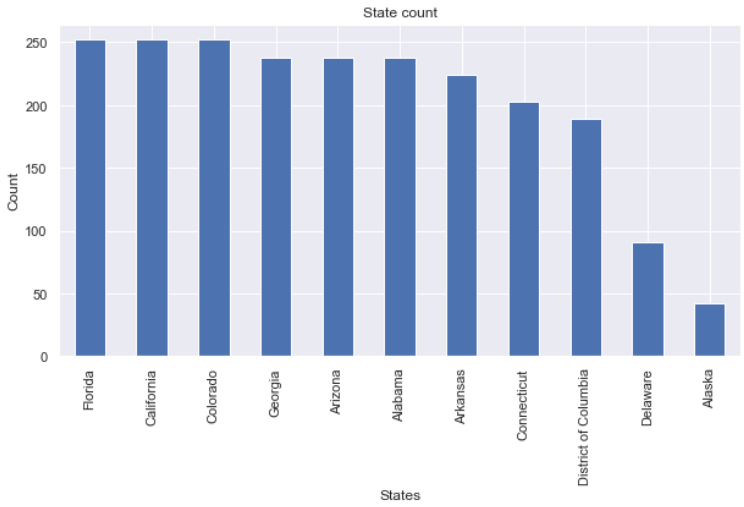


Figure 4: Number of cases per state

This bar chart shows the occurrence of each state in the data set. From the values on this graph, we can speculate that states such as Florida, California, and Colorado received higher number of doses, and had a higher amount of cases relative to the other states in the US.

Correlation heat map

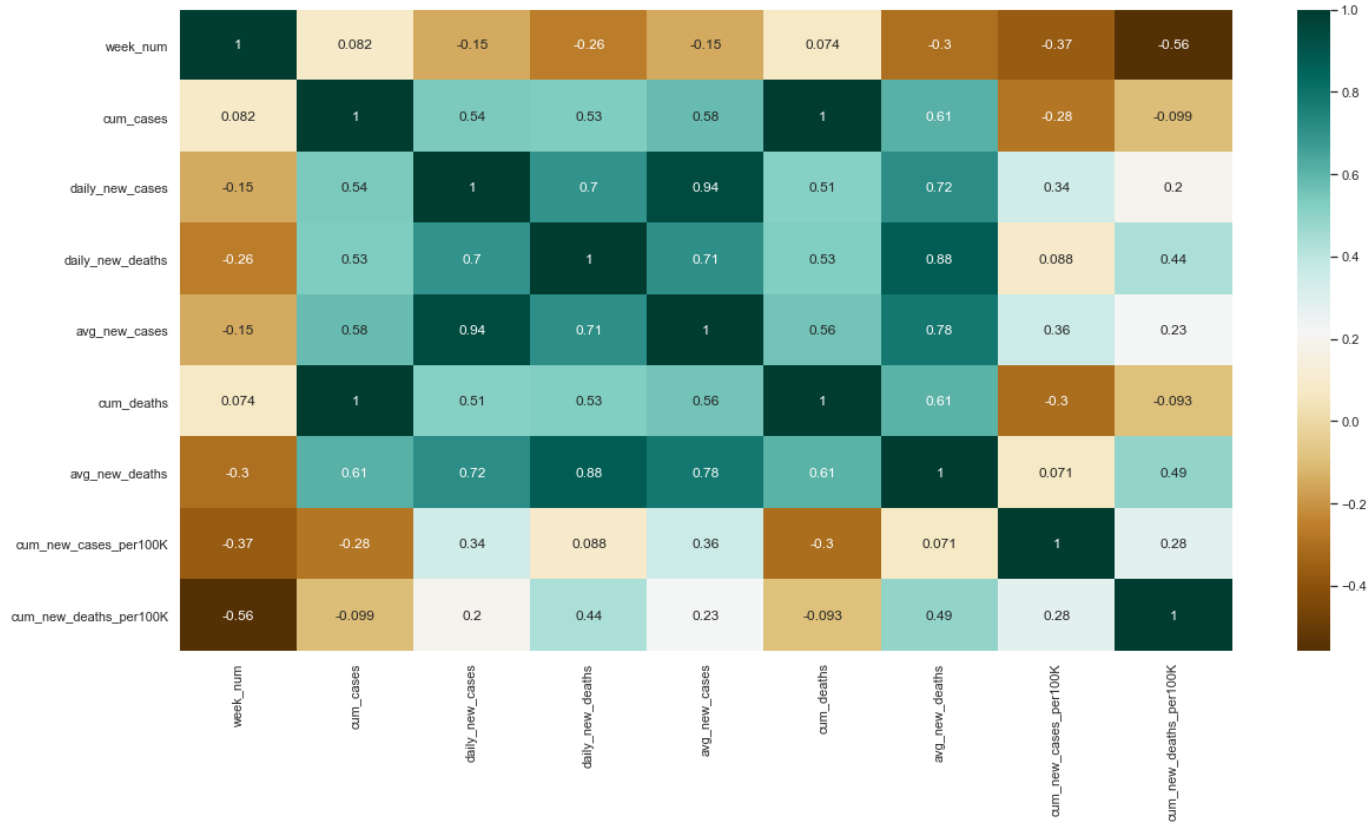
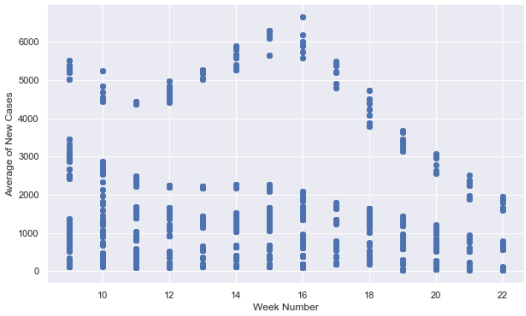


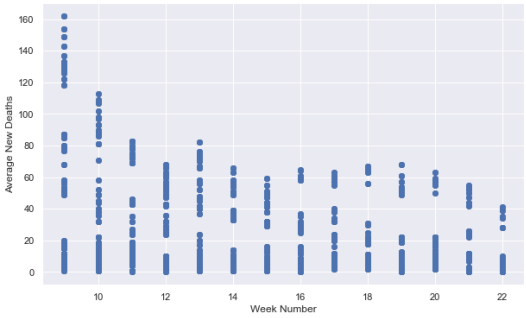
Figure 5: Map of correlation between variables

This heat map shows the correlation between each variables. From the heat map, the variables that relate to COVID-19 cases have strong positive relationships with the variables that relate to death caused by COVID-19. This make sense that as more people got infected, there would be more people died because of the infection. Also, it shows the negative correlation between the week number and the cumulative new cases and deaths per 100k population. This shows that there is less people getting infected or died because of COVID-19 over time, and this also suggest that vaccines are having a strong effect on preventing the spread of COVID-19.

Average new cases vs week number and Average new deaths vs week number



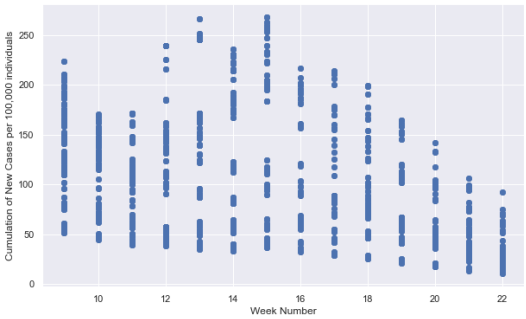
(a) Average new cases vs week number



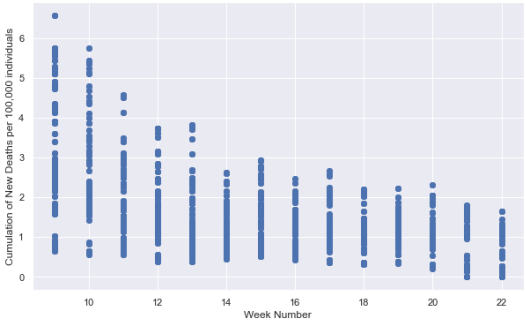
(b) Average new deaths vs week number

It shows a trend of decreasing average new cases. This is a good sign that the average new cases are decreasing especially for the states that had a large amount of average new cases. It also shows a decreasing trend for average new deaths for most of the states.

Cumulation of new cases and deaths per 100,000 individuals vs week number (respectively)



(a) Cumulation of New Cases per 100,000 individuals vs week number



(b) Cumulation of New Deaths per 100,000 individuals vs week number

The cumulative new cases per 100K population as well as the cumulative new deaths per 100K population are showing a decreasing trend.

2.1.3 Understanding the table df_vacc_totals

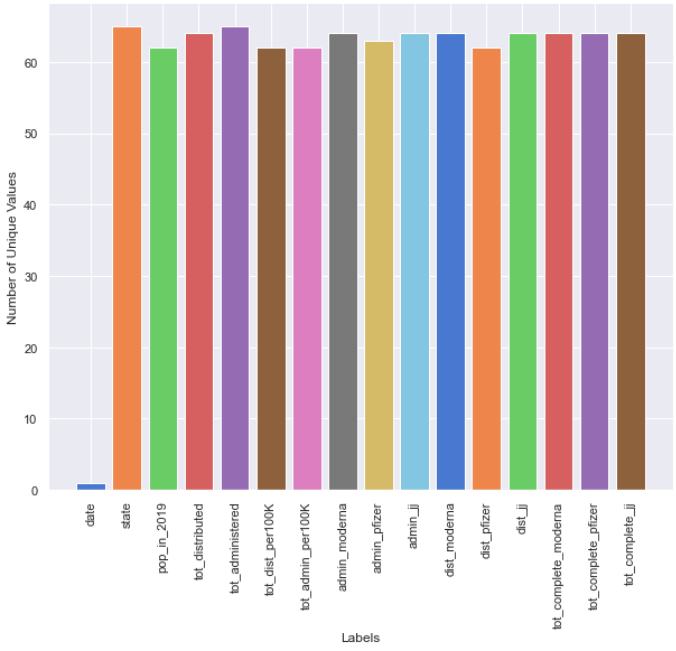
Data Types

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 65 entries, 0 to 64
Data columns (total 16 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   date                  65 non-null     object
 1   state                 65 non-null     object
 2   pop_in_2019           64 non-null     float64
 3   tot_distributed       64 non-null     float64
 4   tot_administered      65 non-null     int64
 5   tot_dist_per100K      64 non-null     float64
 6   tot_admin_per100K     64 non-null     float64
 7   admin_moderna         64 non-null     float64
 8   admin_pfizer          64 non-null     float64
 9   admin_jj              64 non-null     float64
 10  dist_moderna          64 non-null     float64
 11  dist_pfizer           64 non-null     float64
 12  dist_jj               64 non-null     float64
 13  tot_complete_moderna  64 non-null     float64
 14  tot_complete_pfizer   64 non-null     float64
 15  tot_complete_jj       64 non-null     float64
dtypes: float64(13), int64(1), object(2)
memory usage: 8.2+ KB
```

Figure 6: Data types

These are the 15 variables in the table **df_vacc_totals**. The most of the variables were described in the **Data Description** section.

Unique and Null Values



(a) Unique Value Count

```
date                  0
state                 0
pop_in_2019           1
tot_distributed       1
tot_administered      0
tot_dist_per100K      1
tot_admin_per100K     1
admin_moderna         1
admin_pfizer          1
admin_jj              1
dist_moderna          1
dist_pfizer           1
dist_jj               1
tot_complete_moderna  1
tot_complete_pfizer   1
tot_complete_jj       1
dtype: int64
```

(b) Null values in the data

There is one observation that contains mostly null values.

Descriptive Statistics

	count	mean	std	min	25%	50%	75%	max
pop_in_2019	64.0	5.219985e+06	7.005948e+06	0.0	854009.75	3117613.0	6756457.75	39512223.0
tot_distributed	64.0	6.016159e+06	8.091460e+06	0.0	1211580.00	3633817.5	7122878.75	48056890.0
tot_administered	64.0	5.029028e+06	6.609794e+06	24369.0	1139818.50	2955154.5	6177784.50	39764396.0
tot_dist_per100K	64.0	1.034121e+05	3.197087e+04	0.0	95690.00	106831.5	121679.25	148731.0
tot_admin_per100K	64.0	8.385512e+04	2.810675e+04	0.0	74693.75	86032.5	101100.75	136086.0
admin_moderna	64.0	2.062154e+06	2.753788e+06	0.0	404922.75	1221639.0	2575914.25	16583915.0
admin_pfizer	64.0	2.660779e+06	3.626408e+06	0.0	516283.75	1494563.0	3138254.00	21763289.0
admin_jj	64.0	1.797589e+05	2.529258e+05	0.0	32420.50	105412.0	208010.25	1411693.0
dist_moderna	64.0	2.490721e+06	3.287392e+06	0.0	501915.00	1567550.0	3051090.00	19577860.0
dist_pfizer	64.0	3.178668e+06	4.339131e+06	0.0	652860.00	1877167.5	3819513.75	25863630.0
dist_jj	64.0	3.467703e+05	4.718127e+05	0.0	64750.00	187000.0	433025.00	2615400.0
tot_complete_moderna	64.0	9.195704e+05	1.207414e+06	0.0	184078.00	549953.5	1146024.00	7212225.0
tot_complete_pfizer	64.0	1.167270e+06	1.557689e+06	0.0	222521.50	668661.5	1389235.50	9288296.0
tot_complete_jj	64.0	1.771821e+05	2.484839e+05	0.0	32866.75	104535.0	203857.00	1399720.0

Figure 7: Statistical description of the data

Cleaning the Data Set

For cleaning this data set, null values were replaced with 0's (only one observation) and the first row, which contained cumulative data of all the states in the U.S., was dropped as it made our data irregular . Moreover, there were no duplicate observations in our data set. Besides these changes, there was not much data cleaning required.

Boxplots

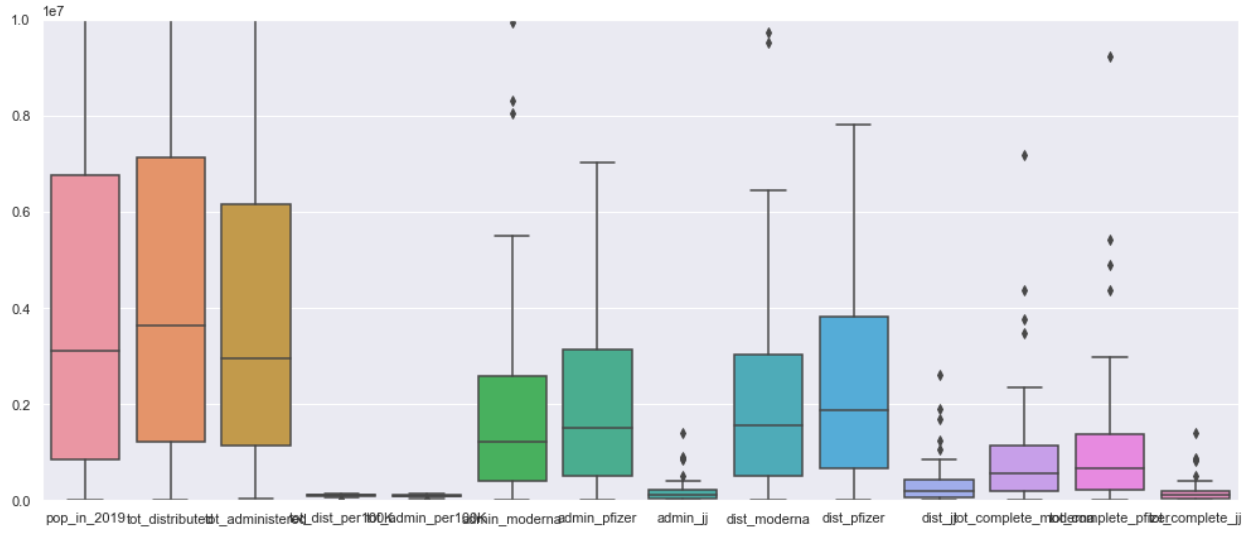
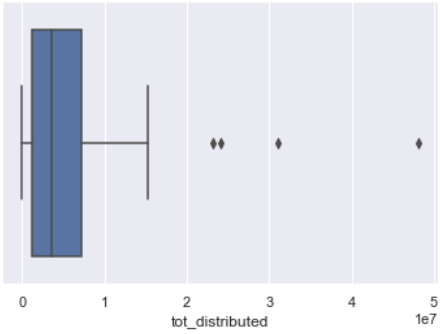


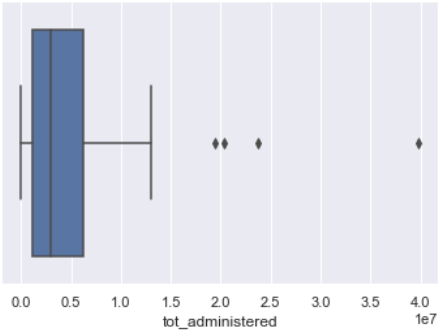
Figure 8: Boxplots of all variables in the data

We can see that the variables vary highly, so it might be useful to observe selected boxplots of the variables.

tot_distributed and tot_administered



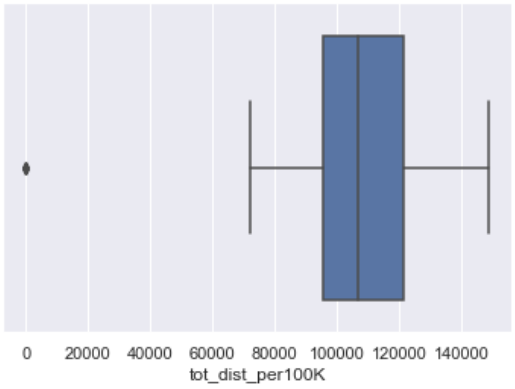
(a) Boxplot of the total distribution



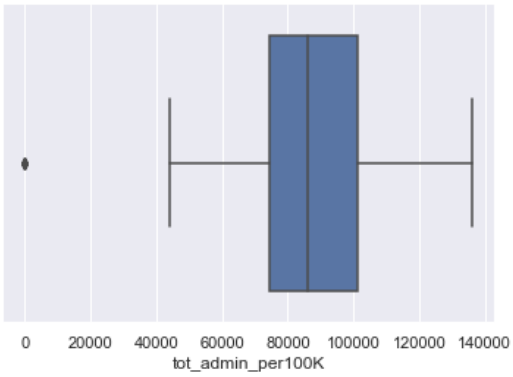
(b) Boxplot of the total administration

There are about 0 to 10,000,000 vaccines are distributed to each state for the distribution plot whereas there are about 1,000,000 to 6,000,000 people received the vaccines for each state in the administration plot.

tot_dist_per100K and tot_admin_per100K



(a) Boxplot of the total distribution per 100,000 individuals



(b) Boxplot of the total administration per 100,000 individuals

There are about 90,000 to 120,000 vaccines distributed per 100K people in each state whereas there are about 70,000 to 100,000 received vaccines per 100K people in each state.

2.1.4 Visualizing Relationships in the Data

Correlation heat map

Figure 9: Map of correlation between variables

This correlation heat map shows the correlation between each variables. It shows that the population, vari-

ables relate to distribution, and variables relate to administration are highly correlated. That is, the states that have large population received a large amount of vaccines and those vaccines are mostly administered.

Total distribution per 100,000 individuals vs total administration per 100,000 individuals

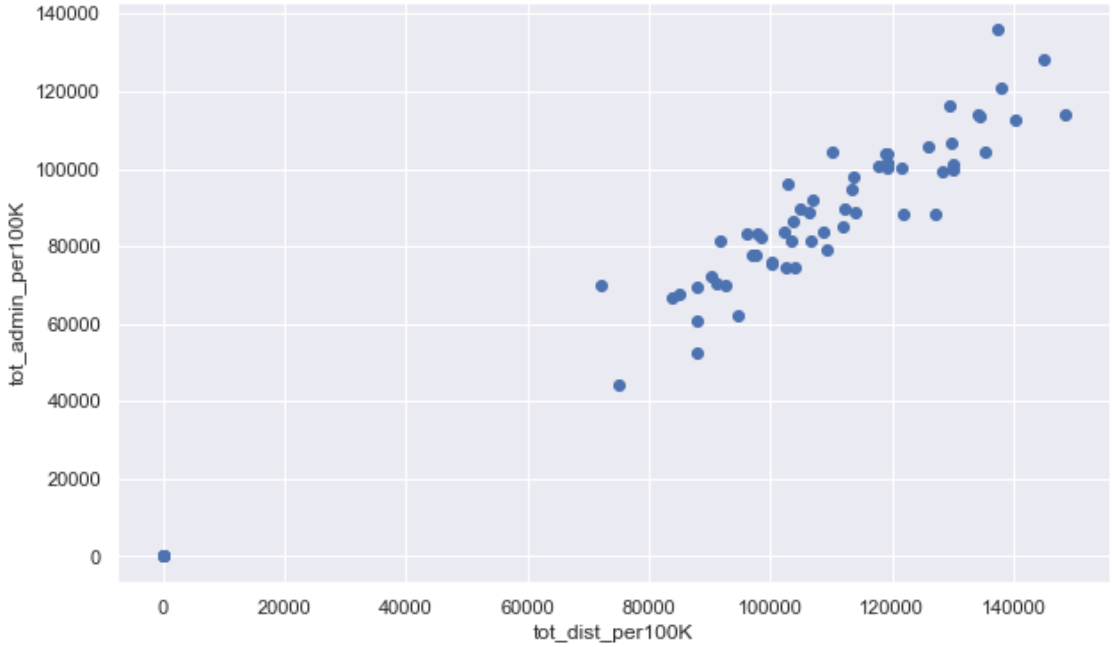


Figure 10: Total distribution per 100K vs total administration per 100K

This plot shows the positive correlation between the total distribution per 100K people and total administration per 100K. That is, as there are more vaccines distributed, there are more people getting vaccinated.

Administration vs distribution of Moderna

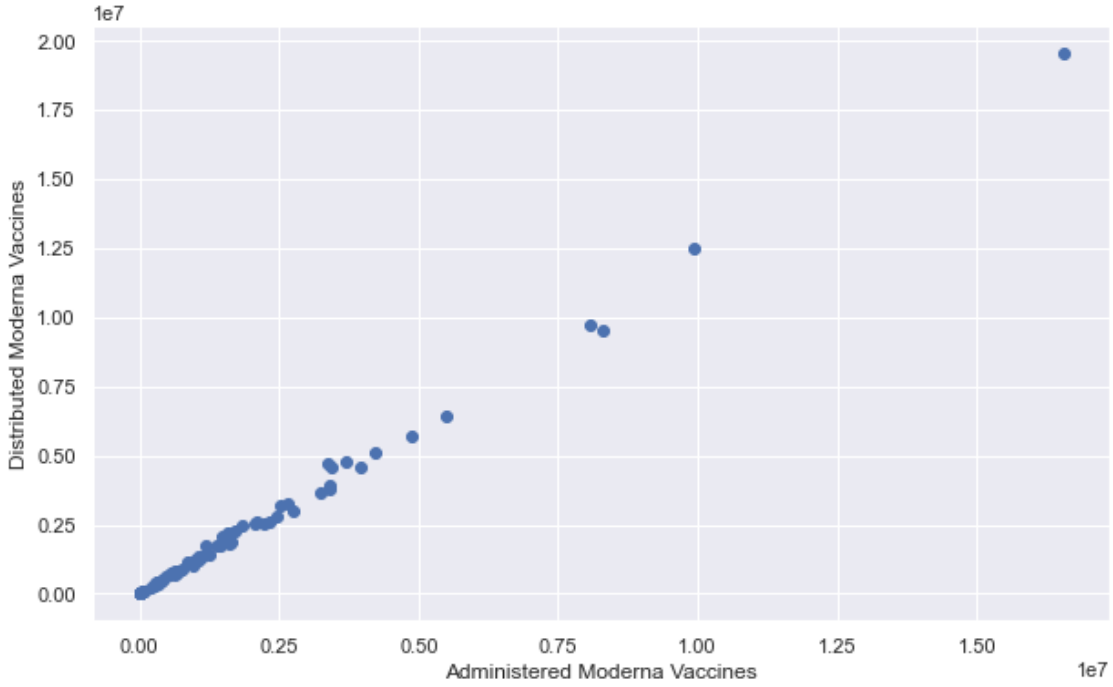


Figure 11: Administration vs distribution of Moderna

This plot shows the positive relationship between the distribution of Moderna and administration of Moderna. This indicates that there are increasing amount of Moderna that are distributing to each state and people are willing to receive them.

Administration vs distribution of Pfizer

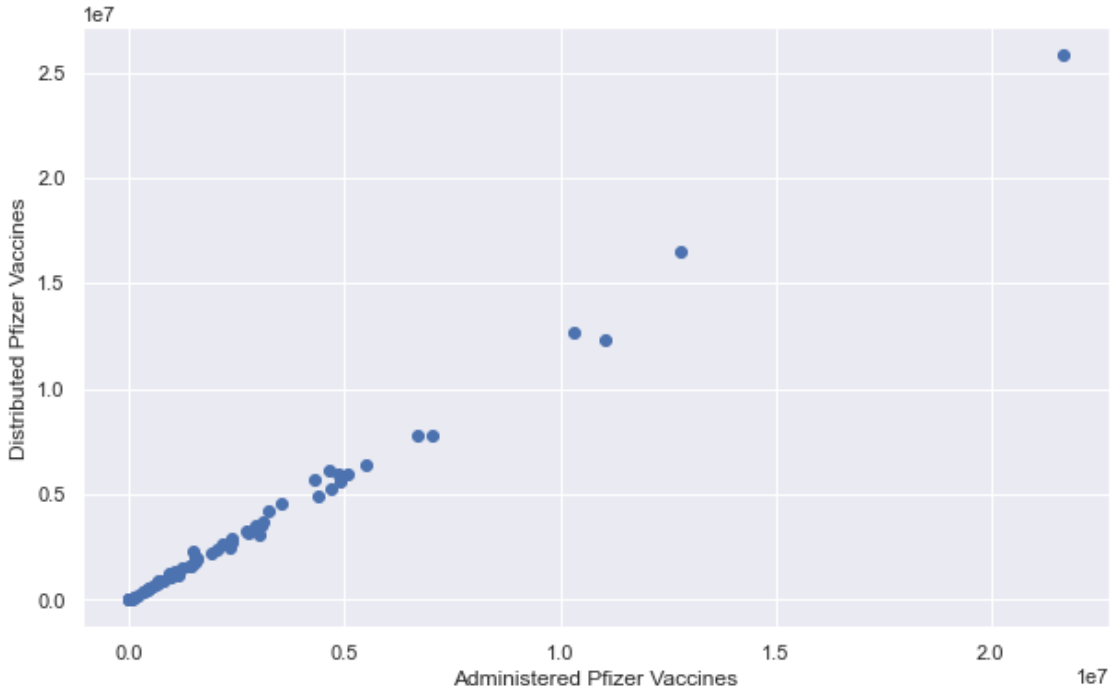


Figure 12: Administration vs distribution of Pfizer

Similar to the Moderna, Pfizer also have a positive relationship between the distribution and administration. That is, people are willing to receive Pfizer as there are more of them rolling out.

Administration vs distribution of Janssen

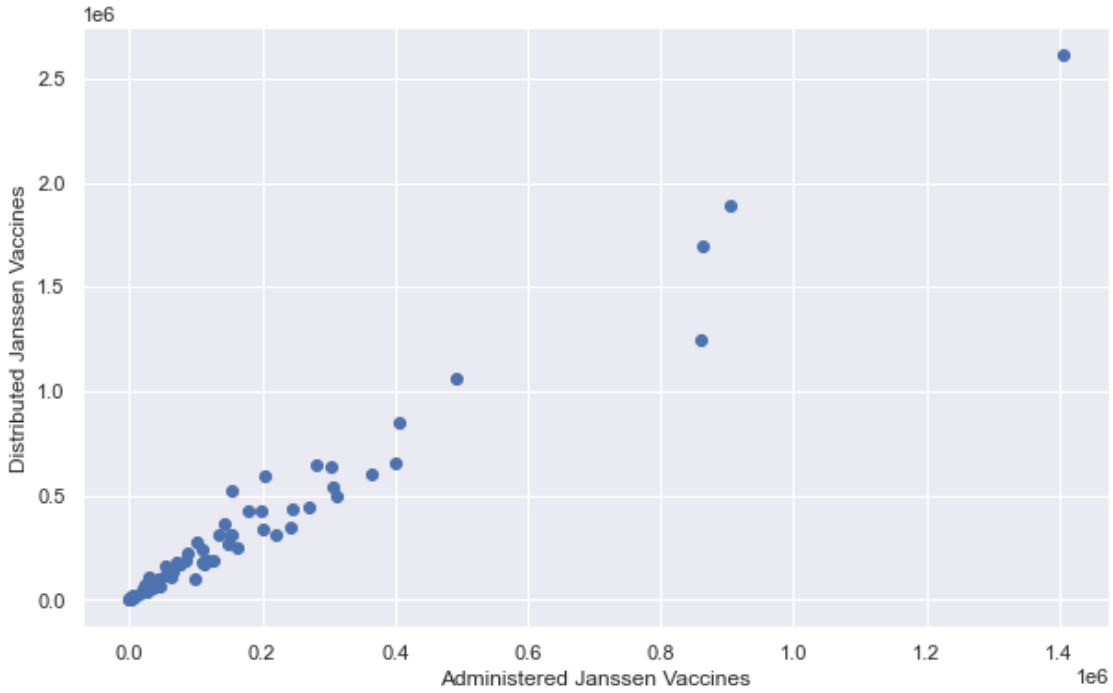


Figure 13: Administration vs distribution of Jannsen

Slightly different than Moderna and Pfizer that the distribution and administration are having a slightly weaker relationship for J&J. This may shows that there are people not willing to receive them maybe because of the side effects.

2.1.5 Understanding the table vacc_demo

Data Types

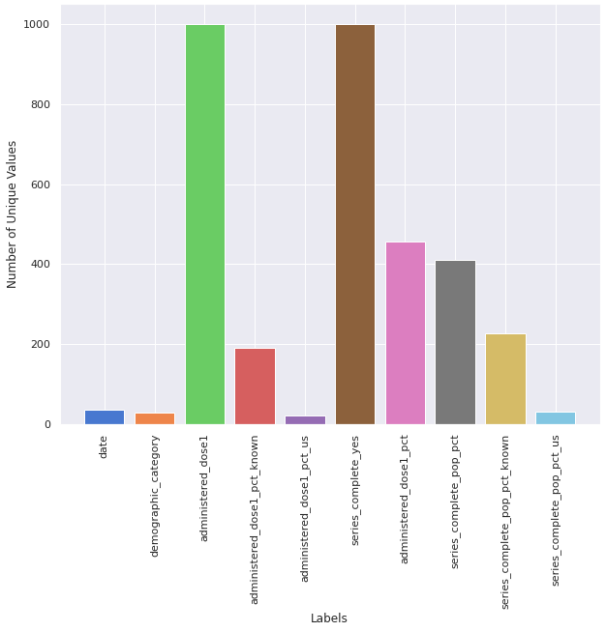
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 10 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   date                          1000 non-null   object
 1   demographic_category          1000 non-null   object
 2   administered_dose1            1000 non-null   object
 3   administered_dose1_pct_known  1000 non-null   object
 4   administered_dose1_pct_us     1000 non-null   object
 5   series_complete_yes           1000 non-null   object
 6   administered_dose1_pct        1000 non-null   object
 7   series_complete_pop_pct       1000 non-null   object
 8   series_complete_pop_pct_known 1000 non-null   object
 9   series_complete_pop_pct_us    1000 non-null   object
dtypes: object(10)
memory usage: 78.2+ KB
```

Figure 14: Data types

These are the 10 variables in the table **vacc_demo**. Most of the variables were described in the **Data**

Description section.

Unique and Null Values



(a) Unique Value Count

```
date                              0
demographic_category              0
administered_dose1                0
administered_dose1_pct_known      0
administered_dose1_pct_us         0
series_complete_yes               0
administered_dose1_pct            0
series_complete_pop_pct           0
series_complete_pop_pct_known     0
series_complete_pop_pct_us        0
dtype: int64
```

(b) Null values in the data

There are no null values in this data set.

2.1.6 Understanding the table state_data21

Data Types

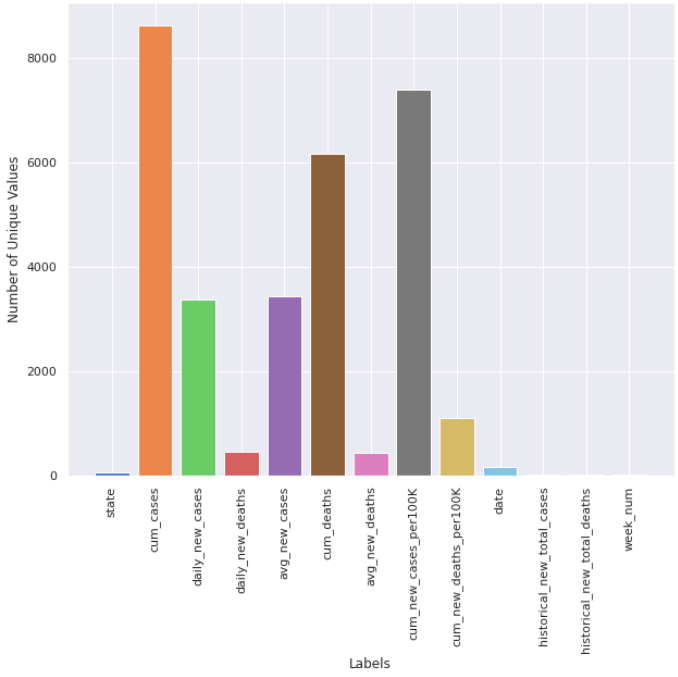
```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 9540 entries, 345 to 30239
Data columns (total 13 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   state                       9540 non-null   object
 1   cum_cases                   9540 non-null   int64
 2   daily_new_cases             9540 non-null   int64
 3   daily_new_deaths            9540 non-null   int64
 4   avg_new_cases               9540 non-null   int64
 5   cum_deaths                  9540 non-null   int64
 6   avg_new_deaths              9540 non-null   int64
 7   cum_new_cases_per100K       9540 non-null   float64
 8   cum_new_deaths_per100K      9540 non-null   float64
 9   date                        9540 non-null   datetime64[ns]
 10  historical_new_total_cases  159 non-null    float64
 11  historical_new_total_deaths 159 non-null    float64
 12  week_num                    9540 non-null   UInt32
dtypes: UInt32(1), datetime64[ns](1), float64(4), int64(6), object(1)
memory usage: 1015.5+ KB
```

Figure 15: Data types

These are the 13 variables in the table **vacc_demo**. Most of the variables were described in the **Data**

Description section.

Unique and Null Values



(a) Unique Value Count

```
state                          0
cum_cases                      0
daily_new_cases                0
daily_new_deaths               0
avg_new_cases                  0
cum_deaths                     0
avg_new_deaths                 0
cum_new_cases_per100K          0
cum_new_deaths_per100K         0
date                           0
historical_new_total_cases     9381
historical_new_total_deaths    9381
week_num                       0
dtype: int64
```

(b) Null values in the data

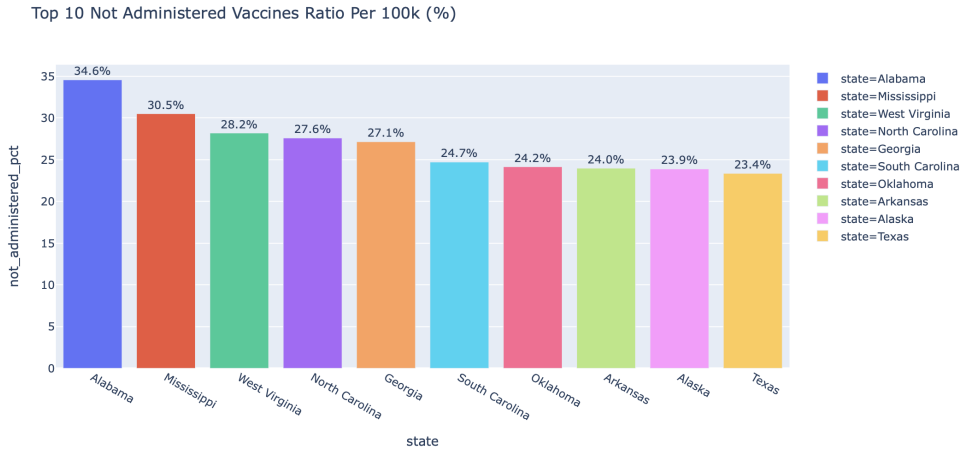
There are two columns with null values but those columns were dropped as they did not contain valuable information.

3 Algorithms

3.1 Regression Models

Before running the modeling, we were curious to see how much of the allocated doses to each state were actually administered to the population. The top 10 states that had the highest percentages of vaccines that were allocated to their population, but went unused are displayed below.

Figure 16: Top 10 States with the Highest Percentages of Vaccines that were not Administered



We thought these 10 states possibly didn't have the resources to administer the amount of doses they were allocated, however we believe it might have more to do with possibly a higher negative sentiment towards vaccines among the states' population.

For our models, **Multiple Linear Regression** and **Random Forest Regression** will be used to predict allocation based on cases, deaths, and population. This can be an effective tool since allocations of vaccines are announced a week beforehand and using the present data, we can predict how much vaccines are gonna be allocated based on those variables. It also works well for new data if we wanted to predict how many vaccines would be allocated to a state/country that is outside of our data based on cases, deaths, and population.

We will also implement a model that takes into account the weeks and state which should be good for predicting an already existing state in our data and their vaccine allocation amount.

3.1.1 Model Comparison

Predictors/Regr Algo	Random Forest Reg	Multiple Linear Reg
Case/Death/Pop	0.9653117576551982	0.974387150090667
Case/Death	0.7296342774823763	0.37746392440824494
Pop	0.98887970942369	0.9881943295992618
Case/Death/Pop/State/ Week	0.9875087866713682	0.9830530029206256

Figure 17: MLR and Random Forest Scores

We can predict with vaccine allocation amount with a 98.8 percent accuracy for an already existing state in our data. Also, from the table, it shows that the regression model that only uses population have a high accuracy. this indicates that population are having a strong influence on vaccine allocation. Overall, both random forest regression and multiple linear regression have great performances on predicting the vaccine allocation amount.

3.2 Clustering Models

First, we will use PCA for dimension reduction. Then, we will apply clustering methods and see which give the best silhouette scores.

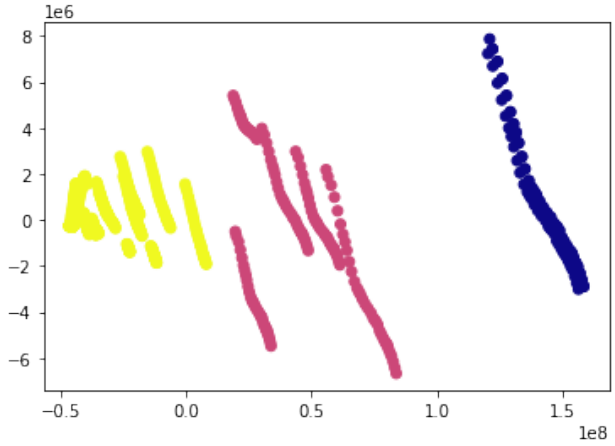


Figure 18: DBSCAN

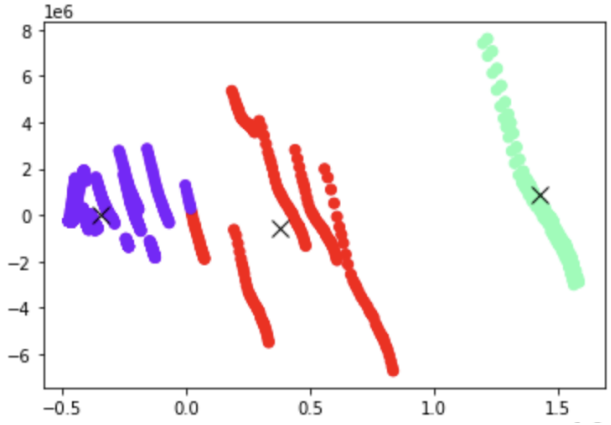


Figure 19: KMeans

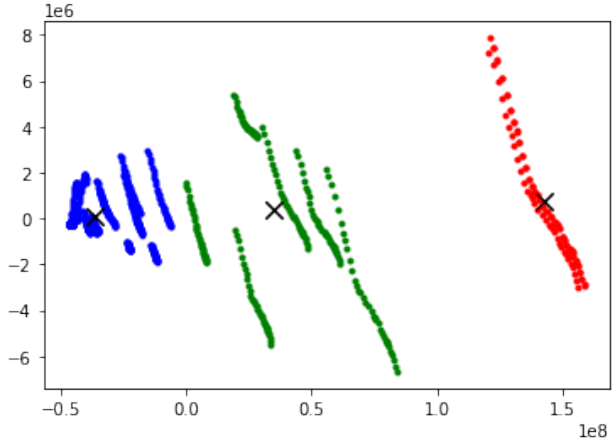


Figure 20: Meanshift

We find that our clustering methods all give similar clusters no matter the technique. We now check our silhouette scores to determine the goodness of our clusters.

3.2.1 Model Comparison

KMeans	0.7384270167826079
DBScan	0.7580215599239616
Meanshift	0.7533056436471283

Figure 21: Silhouette Scores

After comparing clustering models, our silhouette scores are all above .70 showing that we've recovered most of the shape and that our clusters are good. Meanshift and DBScan perform the best slightly having an edge over Kmeans. Our clustering tells us that there are trends in our data that can be discovered amongst our different demographic groups.

4 Conclusion

Overall, in this project, we have analyze and model the CDC COVID vaccine data. Through the analyzing, we find out that the infection of COVID-19 are getting controlled as vaccines are popularized. Moreover, since the death caused by COVID-19 are highly correlate to the new COVID-19 cases positively, the new deaths are also decreasing. These features are all showing the sign of the end of pandemic. Meanwhile, in order to use the data to help the allocation of COVID vaccines, we draw regression models that can be used to predict the vaccine allocation amount for a state. On the other hand, we use clustering methods and dimension reduction on the data, and we find out that the data are likely to have 3 clusters which agree to the amount of the vaccine brands.

4.1 Discussion/Future Work

In this project, we have analyze the state of vaccination in each state. Maybe in future, we can corporate the data into a map to see if the geography of the states are correlate to the vaccination, and we can also consider about the political states in each state that may can strongly influence people's will of getting the vaccine or not.

References

[1] Centers for Disease Control and Prevention. *COVID-19 Vaccine Distribution Allocations by Jurisdiction - Pfizer*. 2021. URL: https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/saz5-9hgg (visited on 06/01/2021).

[2] Centers for Disease Control and Prevention. *COVID-19 Vaccine Distribution Allocations by Jurisdiction - Moderna*. 2021. URL: https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/b7pe-5nws (visited on 06/01/2021).

[3] Centers for Disease Control and Prevention. *COVID-19 Vaccine Distribution Allocations by Jurisdiction - Janssen*. 2021. URL: https://data.cdc.gov/Vaccinations/COVID-19-Vaccine-Distribution-Allocations-by-Juris/w9zu-fywh (visited on 06/01/2021).

[4] Centers for Disease Control and Prevention. *COVID Data Tracker*. 2021. URL: https://covid.cdc.gov/covid-data-tracker/#datatracker-home (visited on 06/01/2021).

[5] Centers for Disease Control and Prevention. *Demographic Trends of People Receiving COVID-19 Vaccinations in the United States*. 2021. URL: https://covid.cdc.gov/covid-data-tracker/#vaccination-demographics-trends (visited on 06/01/2021).