STAT-6970-100 - Data Science Masters Project
COVID Data Analysis and Modelling

By (COVID Broncos)
Sri Surya Sameer Vaddhiparthy,
Humpi Chowdary Goli,
Corey Moon

Under the guidance of
Dr. Hyun Bin Kang

Department of Statistics
Western Michigan University
1903 W Michigan Ave
Kalamazoo MI 49008-5152 USA

Abstract

The project presents the COVID-19 situation in the United States until March 2022 using CDC and 2020 Census data. Various metrics used by the public when they are having a conversation about the COVID-19 situation have been identified. CDC COVID-19 data was aggregated and analyzed with the construction of a unique identifier (UID) based on state, year, and week. Vaccination data were compared against age levels, education levels, and poverty rates for all US states, D.C., and Puerto Rico. Visualizations were used to explore the breakdown of COVID-19 cases, deaths, and vaccinations by state. ARIMA modeling was used to predict cases, and beta regression was used to analyze the impact of demographics on vaccination rates. The ARIMA model tuning and analysis of the error metrics have pointed toward increasing the data resolution from weekly to daily data points which resulted in an accurate ARIMA model. After proper tuning, the ARIMA (2,1,3) model performed well at predicting cases with new data. The results of beta regression models indicated that the proportion of administered vaccines, and some age and education levels were important factors influencing complete vaccination rates, while completed vaccinations and some education levels were important factors influencing rates of vaccine booster shots. Further model tweaking could be done to improve model accuracy and interpretations.

Introduction

Literature review

Correlation

In data science one of the most important statistical terms used is correlation. To analyze the relationships between the variables, Correlation is a statistical technique used. It is used by estimating the power of connections between variables in data. A portion of the inquiry's relationship addresses the data are:

- Is there a connection between the variables?
- Does changing the value of one variable influence the value of the different variables?
- How solid is the connection between the variables?

Correlation refers to the degree to which two variables have a linear relationship with one another. A statistical method can show whether and how emphatically variables are connected. It is a scaled form of co-variance and values go from - 1 to +1. Variables inside a data set can be connected for lots of reasons. One variable could cause or rely upon the upsides of another variable, a variable could be delicately connected with another variable, or two variables could rely upon a third obscure variable.

It may very well be valuable in data analysis and modeling to even more likely to figure out the connections between variables. The statistical connection between two variables is referred to as their correlation. A correlation could be positive, meaning the two variables shift in a similar course, negative, implying that when one variable's worth expands, the other variables' qualities decline, or neutral, implying there is no relationship between the variables. Correlation can likewise be brain or zero, it is inconsequential to imply that the variables.

Stationarity

Stationarity is a significant idea in the field of time series analysis with a colossal impact on how the data is seen and anticipated. While forecasting the future, most time series models expect that each point is free of the other. The best indication of this is the point at which the dataset of past cases is stationary. For data to be stationary, the measurable properties of a framework don't change after some time. This doesn't mean that the qualities for every data point must be something similar, yet the general way of behaving of the data ought to stay consistent. From a simple visual assessment, time plots that don't show patterns or irregularity can be thought of as stationary. More mathematical variables on the side of stationarity incorporate a consistent mean and a constant variance.

Trend

When there is a long-term increase or decrease in the data

Seasonality

Reoccurring pattern at a fixed and known frequency based on a time of the year, week, or day.

A stationary interaction has the property that the mean, variance, and autocorrelation structure don't change after some time. Stationarity can be characterized in exact numerical terms, yet for our motivation, we mean a flat-looking series, without a pattern, consistent variance over the long run, a steady autocorrelation structure over a long time, and no periodic changes.

Dicky fuller test

Dickey-Fuller test tests the null hypothesis that a unit root is available in an autoregressive time series model. The elective hypothesis is different relying upon which form of the test is utilized, however, is typically stationarity or trend-stationarity.

A simple AR model can be represented as:

$$y_t = \rho y_{t-1} + u_t$$

where y_t is the variable of interest at the time t, ρ is a coefficient that defines the unit root and u_t is noise or can be considered as an error term.

If ρ = 1, the unit root is present in a time series, and the time series is non-stationary.

If a regression model can be represented as

$$\Delta y_t = (\rho - 1) y_{t-1} + u_t = \delta y_{t-1} + u_t$$

Where, Δ is a difference operator and δ = ρ-1

So here, if ρ = 1, which means we will get the differencing as the error term, and if the coefficient has some values smaller than one or bigger than one, we will see the changes according to the past observation.

There can be three versions of the test.

- $\Delta y_t = \delta y_{t-1} + u_t$ test for a unit root
- $\Delta y_t = a_0 + \delta y_{t-1} + u_t$ test for a unit root with constant
- $\Delta y_t = a_0 + a_1 t + \delta y_{t-1} + u_t$ test for a unit root with the constant and deterministic trends with time

So, assuming that a time series is non-stationary, it will tend to return an error term or a deterministic pattern with the time values. If the series is stationary, it will tend to return just an error term or deterministic pattern. In a stationary time, series, a huge worth tends to be trailed by a little worth, and a little worth tends to be trailed by an enormous worth. Also, in a non-stationary time series, the huge and the little worth will accumulate with probabilities that don't rely upon the ongoing worth of the time series.

Arima model [Autoregressive Integrated moving average]

It is a model utilized in statistics and econometrics to gauge occasions that occur throughout some period. The model is utilized to comprehend past data or predict future data in a series. It's used when a measurement is kept in ordinary spans, from parts of one moment to every day, week after week, or month to month periods. ARIMA is a kind of model known as a Box-Jenkins technique.

An autoregressive integrated moving average model is a type of regression analysis that checks the strength of one dependent variable relative to other changing variables. The model's goal is to predict future securities or financial market moves by examining the differences between values in the series instead of through real qualities.

An ARIMA model can be understood by outlining every one of its components as follows:

Autoregression (AR): alludes to a model that shows a changing variable that relapses all alone lagged, or prior, values.

Integrated (I): represents the differencing of crude observations to take into consideration the time series to become stationary

Moving average (MA): incorporates the dependency between an observation and a residual error from a moving average model applied to lagged observations.

ARIMA Parameters.

Every component in ARIMA functions as a parameter with a standard notation. For ARIMA models, a standard notation would be ARIMA with p, d, and q, where integer values substitute for the parameters to indicate the type of ARIMA model used. The parameters can be defined as:

p: the slack order.

d: the degree of difference.

q: the order of the moving average.

In a linear regression model, for example, the number and type of terms are included. A 0 worth, which can be used as a parameter, would imply that a particular component should not be used in the model. This way, the ARIMA model can be constructed to perform the function of an ARMA model, or even simple AR, I, or MA models.

Root Mean Square Error [RMSE]

RMSE is perhaps the most normally involved measure for assessing the nature of predictions. It shows how far predictions tumble from estimated genuine qualities utilizing Euclidean distance. To process RMSE, ascertain the residual (distinction among expectation and truth) for every data point, figure the standard of residuals for every data point, register the mean of residuals and take the square root of that mean. RMSE is generally utilized in supervised learning applications, as RMSE uses and needs obvious estimations at each predicted data point.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|^2}{N}},$$

Where N is the number of data points, y(i) is the i^{th} measurement, and y ̂(i) is its corresponding prediction.

Dataset description

CDC COVID-19 Vaccinations

This data set was obtained from the official public domain CSV file on the CDC's website, and it had close to 30,000 rows and 82 features and notable of which was date, MMWR week, vaccination statistics on the three vaccines in the US that are COVID-19 mRNA vaccines (Pfizer-BioNTech or Moderna) and Johnson & Johnson's Janssen. It contained the vaccination statistics based on age, population distribution, additional doses of vaccine, vaccines distributed per 100,000 of the population, etc. In the column State, there were 66 entries which signal that there were entries that didn't belong to any specific state in the US. The final set of features selected from this data set are:

'uid', 'state', 'date', 'mmwr week', 'week', 'distributed janssen' series complete moderna', 'distributed pfizer','series complete pfizer', 'distributed', 'dist per 100k', 'administered', 'admin per 100k', 'administered dose1 recip', 'administered dose1 pop pct', 'series complete yes', 'series complete pop pct', 'additional doses', 'additional doses vax pct', 'additional doses janssen', 'additional doses moderna', 'additional doses pfizer'.

US COVID-19 Cases and Deaths

This data set has also been sourced from the CDC's official website in form of a CSV file it had close to 48,000 rows and 15 columns and it had features like the submission date state total cases, total deaths, new cases, new deaths, etc. This data set is updated daily by the CDC and hence it was used for modeling the ARIMA model to predict new COVID19 cases using the daily frequency data and for visualizing the distribution and variation of various covid parameters the weekly frequency of this data was used. The following features were selected from this data set for further analysis: 'uid', 'cases total', 'cases new', 'death total', 'death new'.

COVID19 Hospital Statistics

This data set from the US Department of Health and Human Services is the largest of all the data sets in this analysis which had close to 42,000 rows and 109 features and it contains hospitalized data from across the US. The features in this data set are the hospital identifier hospital name, address and various admissions statistics, ICU, and regular bed occupancy statistics. The primary challenge is when the relevant features for the analysis had to be identified from the 109 available features in this data set and when aggregating the data set. The aggregation level for this analysis is weekly and state-level aggregation and this data set has hospital-level entries during various periods. Pivoting has been used to summarize the data to the state level after eliminating the nulls and negative values. The final features selected from this data set for analysis are total beds, total ICU beds, and the proportion of beds used from these two categories.

Census Data

The data used in the project was sourced from multiple tables of the 2020 US Census. Selected data regarding ages, education status, poverty and population counts were merged into a single data frame. Additional vaccination data from the CDC was merged as well. The final data contained information from all states, D.C., and Puerto Rico. The selected features were: the number of vaccines administered, the number of vaccines distributed, vaccine booster shot counts, number of people with a completed vaccination series, number of people under 18 years old, the number of people between 18 and 64 years old, the number of people 65 and older, the number of people with less than high school education, the number of people with high school through an associate's degree education, the number of people with at least a bachelor's degree, the number of people in poverty (whose status could be identified), and the estimated total population.

Research questions

This analysis was used to answer questions that fall under the public curiosity on covid-19 about the variation in the number of cases and deaths regarding the vaccination, availability of the hospital facilities, and demographics. Our goals were to investigate:

1. Trends in vaccination and boosters across states
2. Vaccine and booster preference state wise
3. Effect of vaccination on hospitalization and death
4. Vaccine distribution statistics across various states
5. Implementation and analysis with ARIMA
6. Short term predictions using ARIMA

Methods

Developed methods

Unique Identifier Generation

The three different data sets used in this analysis had rows with different timestamps and hence it has become impossible to join these rows about a specific date or week because they had different times hence the primary candidate considered was the MMWR week, which is the epidemiological week but it is a yearly feature that repeats yearly thereby eliminating the possibility of uniquely identifying a row with data of multiple years and hence the need for developing a unique identifier for the specific analysis in this project. This unique identifier must be unique across various years, state, month, and week. A unique identifier (UID) was specifically formulated to join and aggregate rows across various data sets into a master data set. This unique identifier consists of a state code followed by a year and week number.

Used methods

Data Loading and Exploration

The dataset obtained was loaded and explored using the below python methods to develop awareness and understanding of the dataset:

- *pd.read_csv* to load the data set from CSV file into the pandas data frame
- *df.head* to view the first few rows of the data frame
- df.shape to view the number of rows and columns
- *Df.columns* to view the list of the features of the data frame
- *df.dtypes* to view the data types of each feature
- *df["Location"].unique()* to print the unique values of the chosen feature
- *len(df["Location"].unique())* to print number of unique values
- *df.describe()* to print the statistical information of the data frame
- *df.isnull().sum()* to print the number of null values
- *df.tail()* to print last few rows of the data

Data Wrangling

The data obtained was not directly suitable for analysis and hence the following actions were performed using the below python methods.

- *df.drop_duplicates:* to drop rows that have the same values.
- *df.loc[['total_beds]]* to select based on feature name.
- *df.sort_values* to sort the values.
- *df.groupby('w_id').sum()* to sum the values across a chosen category in the groupby.
- *df.rename* to rename the feature names in a regular manner.
- *pivot_table* was used to aggregate the data in the hospital dataset from hospital level to weekly state level.
- *pd.merge* to join various data frames after completion of data wrangling.
- *df.to_csv* to back up to a CSV file and analyze the data in external software like Tableau to generate interactive visualizations etc.

Visualization techniques

The following are the different ways in which the data were visualized:

- *df.plot* to obtain line plots of the time series data.
- *sns.barplot* to obtain bar plots of the vaccination statistics.
- *sns.heatmap* to generation heatmaps that visualize the correlation.

Analysis techniques

The following were the methods incorporated in the project to analyze the different aspects of COVID-19 data.

- **Correlation** *df_ny.corr().round(decimals=2)* to compute correlation of various variables in the Newyork state subset dataset and round the value to 2 digits.
- **Dickey-fuller test:** *adfuller(df, autolag='AIC')* to evaluate if a time series is stationary.
- **ARIMA Model hyperparameter tuning:** *evaluate_arima_model* to perform grid search for best hyperparameter with lowest RMSE.

- **ARIMA Model implementation:**
  model = ARIMA(series, order= order)
  model_fit = model.fit()
- **ARIMA Forecasting:** *model_fit.forecast()[0]*

Demographic data analysis

To create better comparisons between states, D.C., and Puerto Rico a few adjustments were made to the demographic data. Each covariate was standardized by the estimated population of each state or territory, and ratios of administered to distributed vaccinations were constructed to reduce the dimensionality. The resulting demographic covariates were proportions (bounded between 0 -1). This created a dataset that was a good fit for beta regression.

Beta regression is a type of generalized linear model in which the error terms are assumed to follow some beta distribution. The beta regression performs well with proportional data and has been shown to outperform other models with transformed data bounded between 0 and 1. An additional benefit of beta regression is that coefficient interpretations are similar to that of logistic regression. Odds ratios can be extracted from the coefficient estimates to better understand which demographic categories are important factors for vaccination rates.

For this project, a custom best subset selection method was constructed to identify which combinations of demographic covariates modeled the data the best. The models were selected based on the lowest Akaike information criterion (AIC), which is a common method for model selection. Two types of models were constructed: Model A, modeling complete series vaccination rates using vaccine administration-distribution rates, education and age levels, and estimated poverty proportions. Model B, modeling vaccine booster shot rates using complete series vaccination rates, vaccine administration-distribution rates, education and age levels, and estimated poverty proportions. Chosen models based on the best subset selection were interpreted to identify significant demographic covariates.

Analysis

Each of the methods described in the previous section was implemented and the outcome of the application of those methods is summarized below.

Developed Methods

UID Generation

Unique Identifier generated using the same technique across all the datasets.

```
[11]: df_vaccine["uid"]=df_vaccine['Location']+df_vaccine['Date'].astype(str).str[:4]+df_vaccine["week"].astype(str)
      df_vaccine["uid"]

[11]: 0         DE202206
      1         KS202206
      2         ID202206
      3         US202206
      4         WI202206
                  ...
      27411     AS202650
      27412    LTC202650
      27413     VI202650
      27414     US202650
      27415     GU202650
      Name: uid, Length: 27416, dtype: object
```

Used Methods

Data loading and exploration

a) COVID Cases and Death Data

```
df_cases.dtypes

submission_date     object
state               object
tot_cases            int64
conf_cases         float64
prob_cases         float64
new_case             int64
pnew_case          float64
tot_death            int64
conf_death         float64
prob_death         float64
new_death            int64
pnew_death         float64
created_at          object
consent_cases       object
consent_deaths      object
dtype: object
```

```
df_cases.shape

(6563, 5)
```

b) COVID Vaccination Data

```
Index(['uid', 'state', 'date', 'mmwr_week', 'week', 'distributed_janssen',
       'series_complete_janssen', 'distributed_moderna',
       'series_complete_moderna', 'distributed_pfizer',
       'series_complete_pfizer', 'distributed', 'dist_per_100k',
       'administered', 'admin_per_100k', 'administered_dose1_recip',
       'administered_dose1_pop_pct', 'series_complete_yes',
       'series_complete_pop_pct', 'additional_doses',
       'additional_doses_vax_pct', 'additional_doses_janssen',
       'additional_doses_moderna', 'additional_doses_pfizer'],
      dtype='object')
```

```
[5]: df_vaccine.shape

[5]: (27416, 82)
```

```
len(df_vaccine["Location"].unique())

66
```

c) Hospital beds Data

```
hospital_pk
collection_week
state
ccn
hospital_name

total_personnel_covid_vaccinated_doses_one_7_day
total_personnel_covid_vaccinated_doses_all_7_day
previous_week_patients_covid_vaccinated_doses_one_7_day
previous_week_patients_covid_vaccinated_doses_all_7_day
is_corrected
Length: 109, dtype: object
```

```
df_hospitals.shape

(419754, 109)
```

Data wrangling

a) COVID Cases and Death Data

For visualization

	uid	cases_total	cases_new	death_total	death_new
0	PW202118	0	0	0	0
1	PW202124	2	0	0	0
2	AS202130	0	0	0	0
3	UT202109	374850	412	1976	1
4	CO202121	548917	460	6580	4
...	...	...	...	...	...
6558	PA202113	1049012	3254	25195	7
6559	RMI202150	4	0	0	0
6560	OH202018	19914	579	1038	17
6561	TN202047	332537	3412	4211	9
6562	SD202023	5438	71	65	0

6563 rows × 5 columns

For time series modeling

	date	state	cases_total	cases_new	death_total	death_new	week	uid
0	2022-01-14	KS	621273	19414	7162	21	02	KS202202
1	2022-01-02	AS	11	0	0	0	52	AS202252
2	2022-01-30	CO	1240361	0	11061	0	04	CO202204
3	2020-07-09	CO	36093	410	1706	2	28	CO202028
4	2022-01-26	CO	1222893	6962	10953	20	04	CO202204
...	...	...	...	...	...	...	...	...
46946	2020-06-07	SD	5438	71	65	0	23	SD202023
46947	2021-12-28	NY	1833465	13006	23446	42	52	NY202152
46948	2021-09-25	RMI	4	0	0	0	38	RMI202138
46949	2021-06-19	TN	863838	182	12518	2	24	TN202124
46950	2021-03-13	IA	342495	425	5633	3	10	IA202110

46951 rows × 8 columns

b) COVID Vaccination Data

	uid	state	date	mmwr_week	week	distributed_janssen	series_complete_janssen	distributed_moderna	series_complete_moderna	distributed_pfizer
0	DE202206	DE	2022-02-10	6	06	96000	56313	825380	229610	1239155
1	KS202206	KS	2022-02-10	6	06	253800	124431	2126560	620054	3249075
2	ID202206	ID	2022-02-10	6	06	155300	79440	1214860	324583	1786310
3	US202206	US	2022-02-10	6	06	30209400	16656790	246235940	74893111	398230385
4	WI202206	WI	2022-02-10	6	06	450400	297555	4002040	1316159	6403745
...	...	...	...	...	...	...	...	...	...	...
4076	AS202050	AS	2020-12-13	51	50	0	0	0	0	0
4077	LTC202050	LTC	2020-12-13	51	50	0	0	0	0	0
4078	VI202050	VI	2020-12-13	51	50	0	0	0	0	0
4079	US202050	US	2020-12-13	51	50	0	0	0	0	0
4080	GU202050	GU	2020-12-13	51	50	0	0	0	0	0

4081 rows × 24 columns

c) Hospital beds Data

```
df_icu.isnull().sum()

uid              0
total_beds       0
total_icu        0
used_inpatient   0
used_icu         0
dtype: int64
```

Pivot tables

The function *pivot_table* was used to aggregate the data in the hospital dataset from hospital level to weekly state level.

	uid	total_beds	total_icu	used_inpatient	used_icu
0	AK202031	11326.0	1413.0	4852.0	725.0
1	AK202032	10910.0	1342.0	4963.0	687.0
2	AK202033	10743.0	1512.0	4957.0	761.0
3	AK202034	11211.0	1524.0	4766.0	831.0
4	AK202035	11153.0	1526.0	4757.0	733.0
...	...	...	...	...	...
4620	WY202205	8563.0	793.0	3921.0	361.0
4621	WY202206	8130.0	772.0	3804.0	353.0
4622	WY202207	7826.0	756.0	3538.0	339.0
4623	WY202208	7098.0	715.0	3454.0	303.0
4624	WY202209	7317.0	623.0	3409.0	223.0

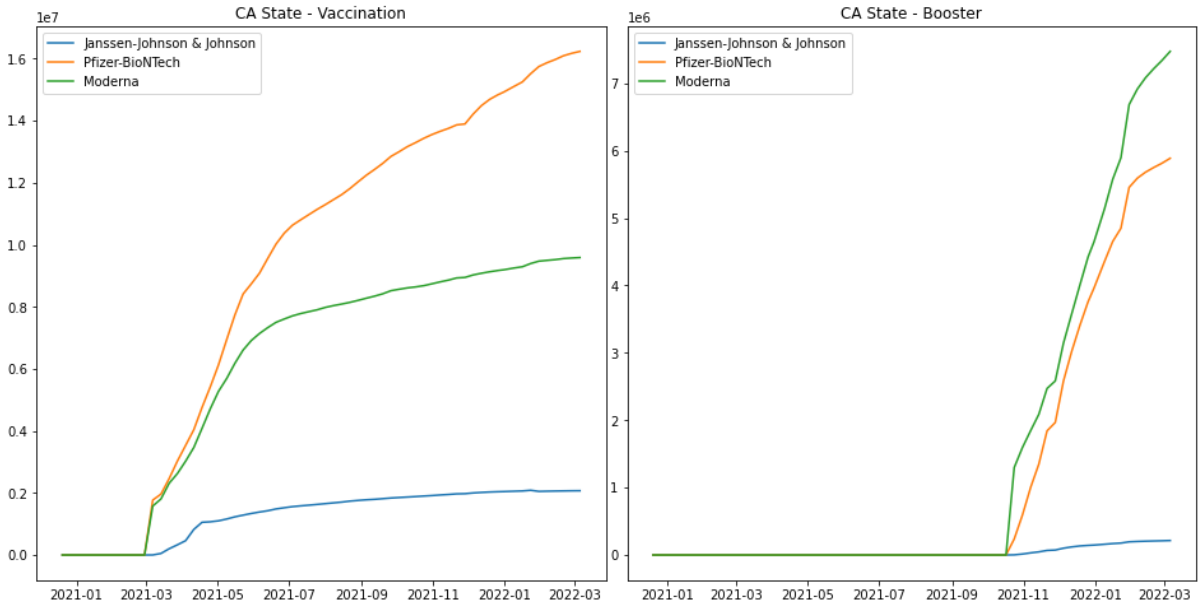
Joins

The function *pd.merge* to join various data frames after completion of data wrangling

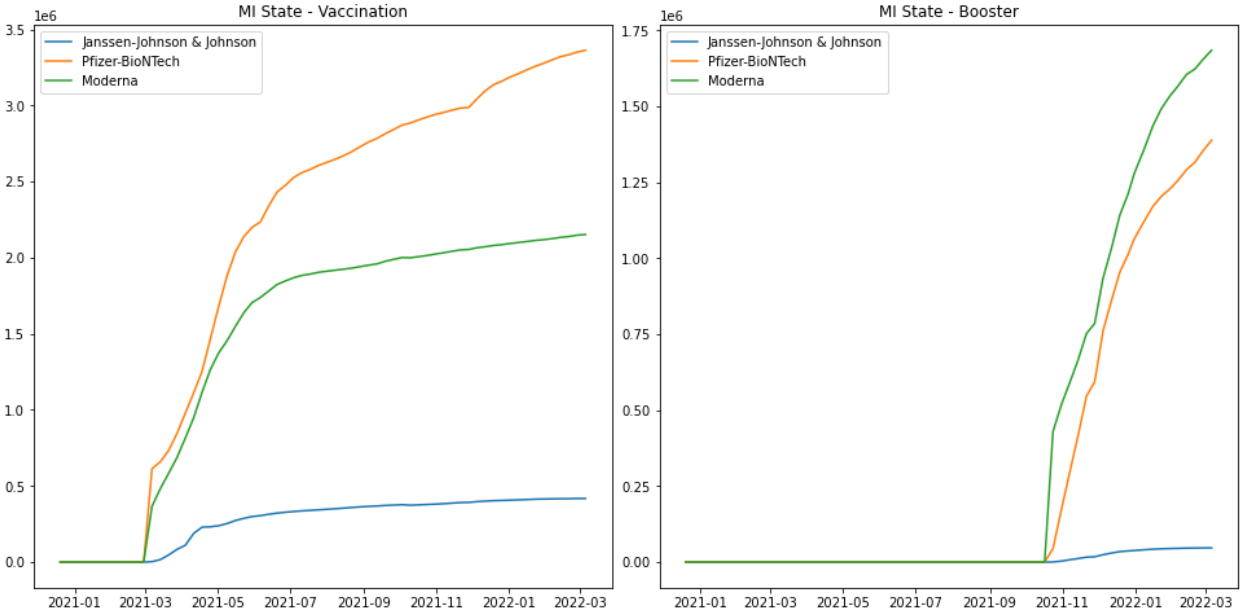
```
Index(['latitude', 'longitude', 'uid', 'stat_abr', 'date', 'mmwr_week', 'week',
       'cases_total', 'cases_new', 'death_total', 'death_new', 'occupied_icu',
       'total_icu', 'distributed_janssen', 'series_complete_janssen',
       'distributed_moderna', 'series_complete_moderna', 'distributed_pfizer',
       'series_complete_pfizer', 'distributed', 'dist_per_100k',
       'administered', 'admin_per_100k', 'administered_dose1_recip',
       'administered_dose1_pop_pct', 'series_complete_yes',
       'series_complete_pop_pct', 'additional_doses',
       'additional_doses_vax_pct', 'additional_doses_janssen',
       'additional_doses_moderna', 'additional_doses_pfizer'],
      dtype='object')
```

Visualization techniques

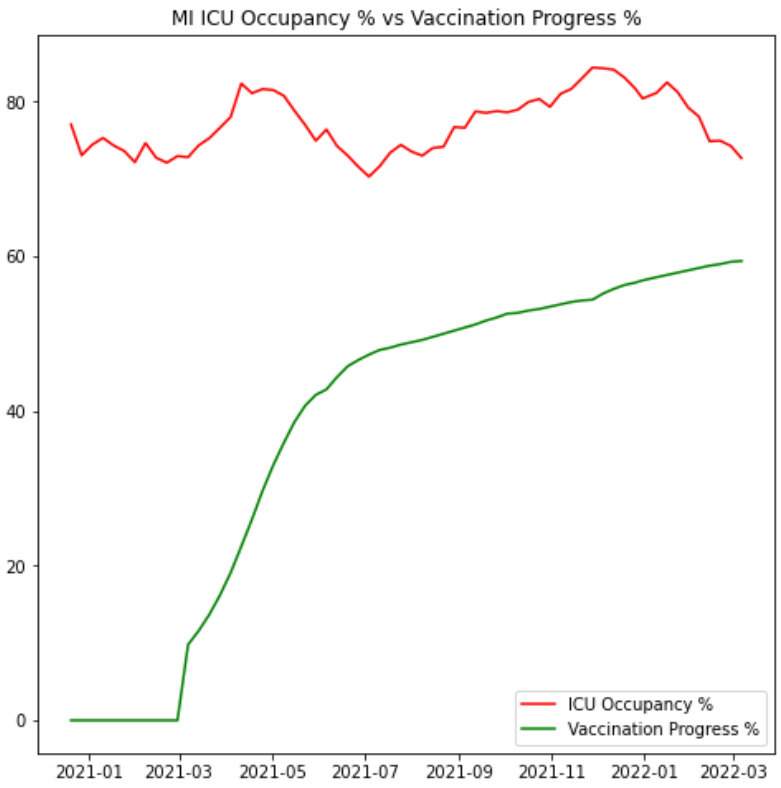
a.) Is there a preferred Vaccine/Booster for Californians?



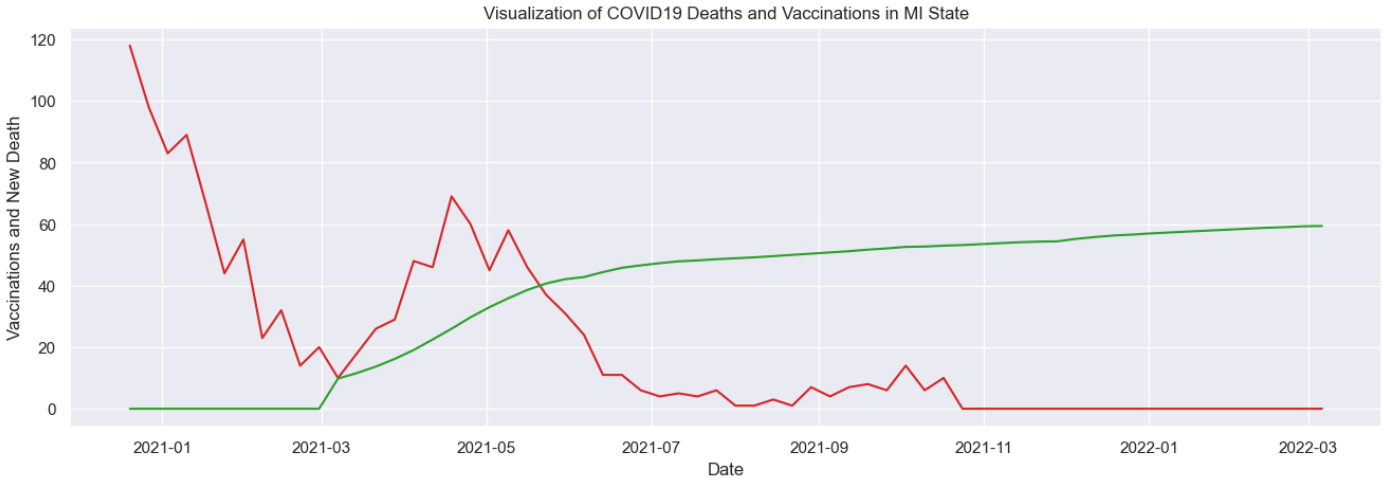
b.) Is there a preferred Vaccine/Booster for Michiganders?



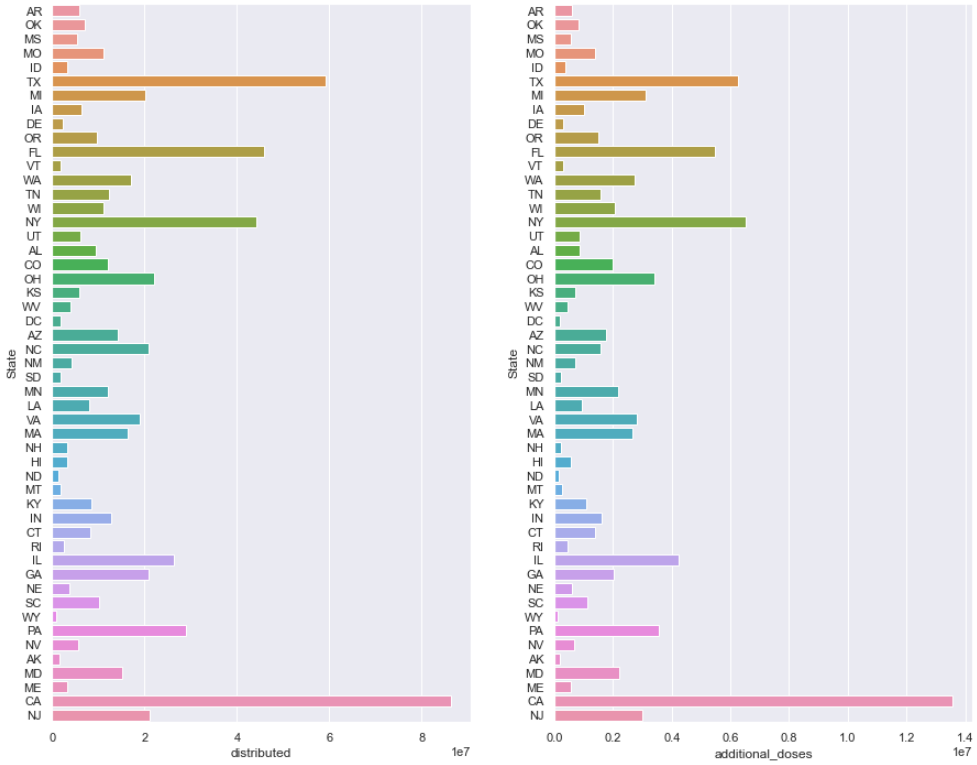
ICU Hospitalization & Vaccination for Michigan



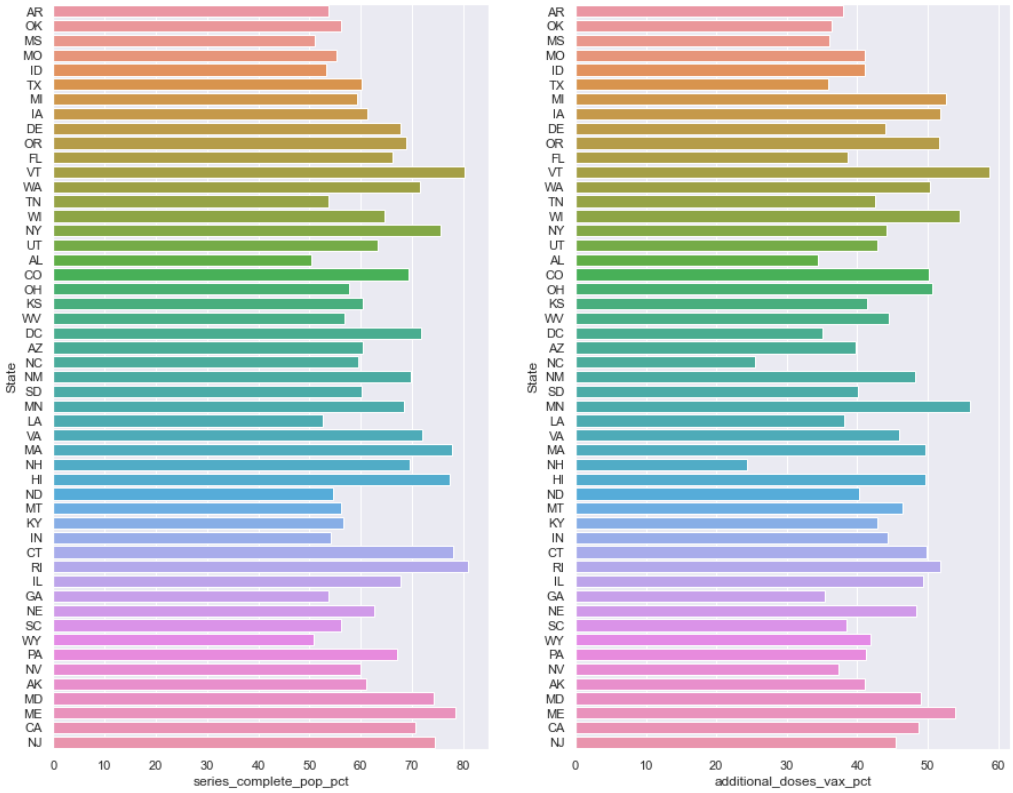
Did Vaccination prevent COVID death in MI?



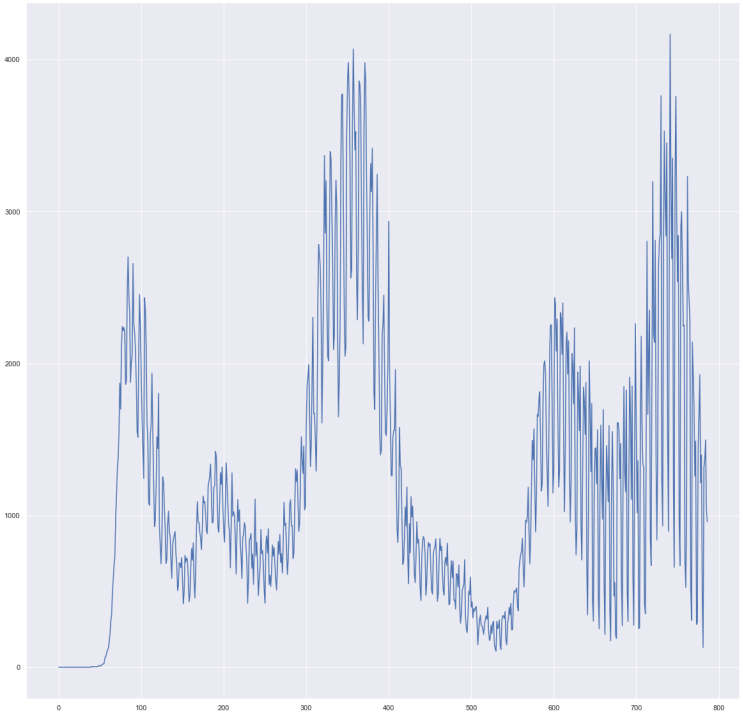
Distribution of Vaccines among States



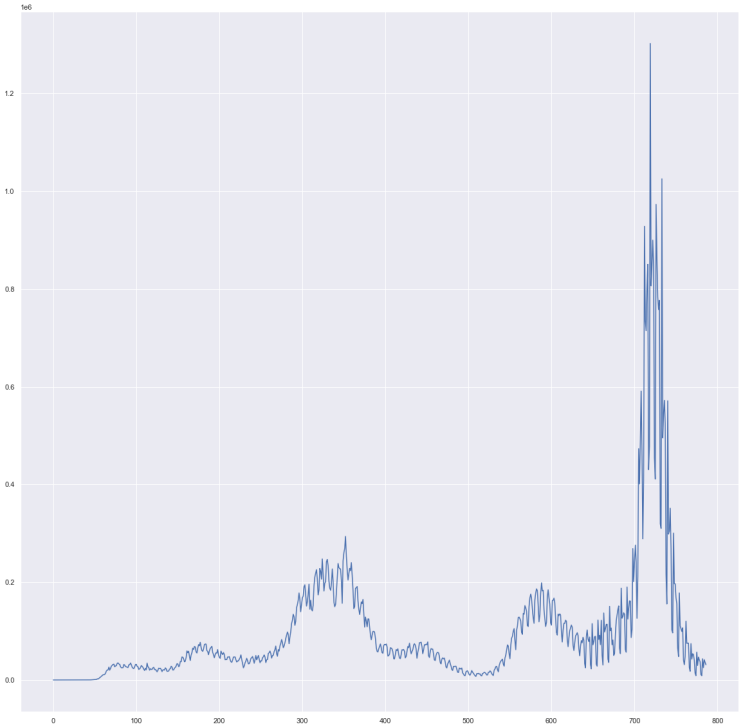
Distribution of Vaccines among States adjusted for population



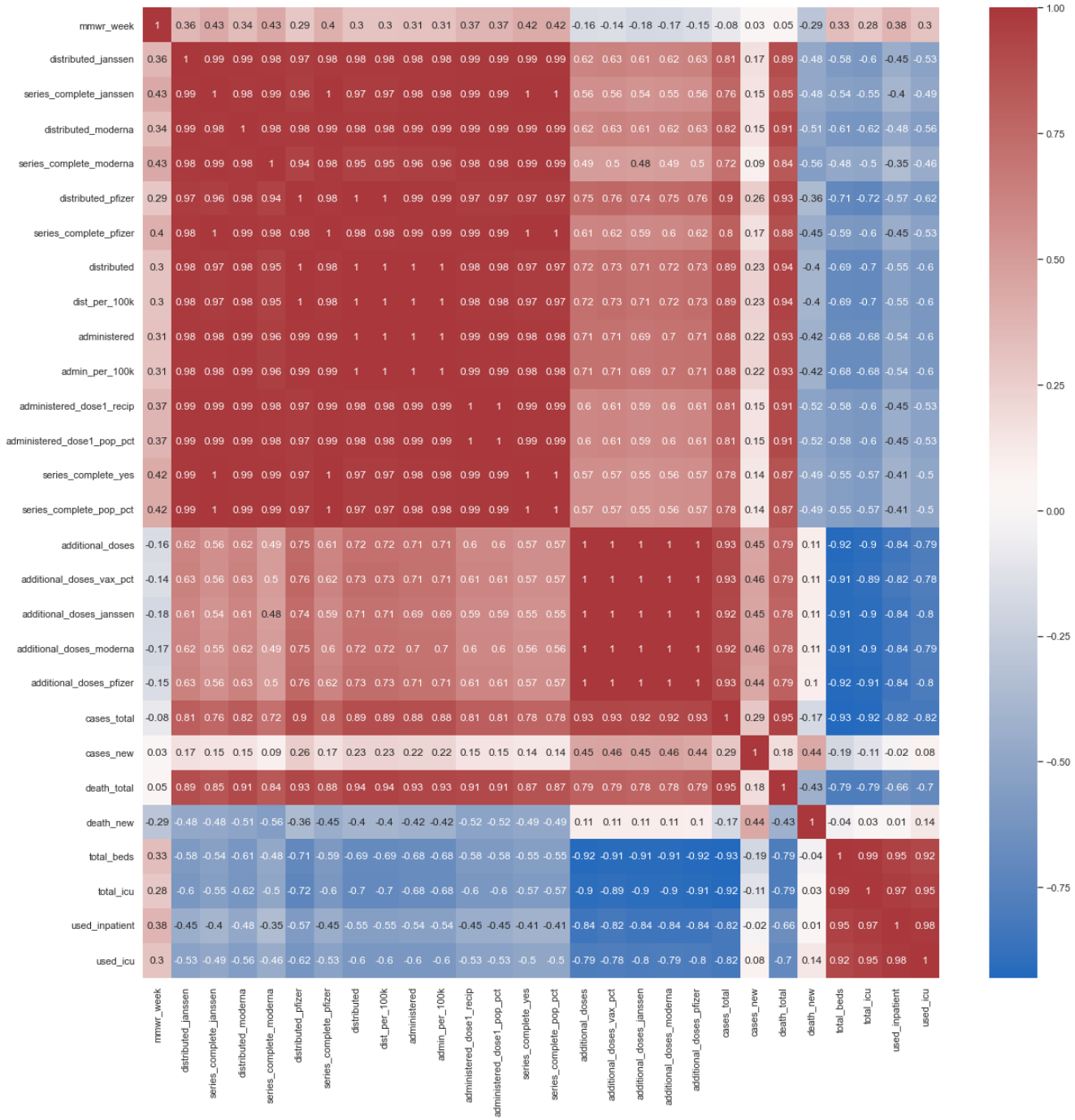
Timeline of COVID Deaths from the beginning of the pandemic

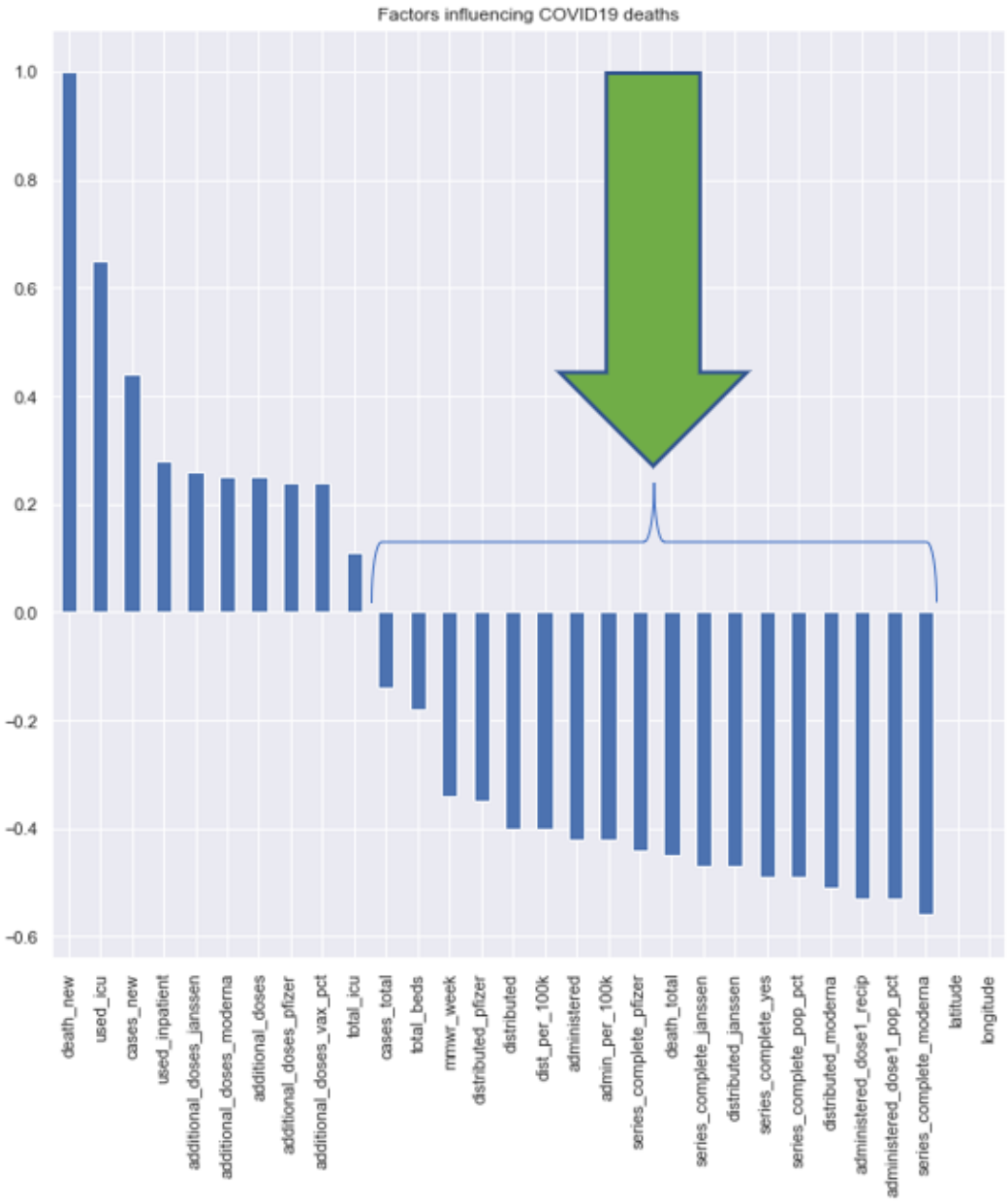


Timeline of COVID Cases from the beginning of the pandemic



Correlation






```
series_complete_moderna       -0.56
administered_dose1_pop_pct    -0.53
administered_dose1_recip      -0.53
distributed_moderna           -0.51
series_complete_pop_pct       -0.49
series_complete_yes           -0.49
distributed_janssen           -0.47
series_complete_janssen       -0.47
death_total                   -0.45
series_complete_pfizer        -0.44
admin_per_100k                -0.42
administered                  -0.42
dist_per_100k                 -0.40
distributed                   -0.40
distributed_pfizer            -0.35
mmwr_week                     -0.34
total_beds                    -0.18
cases_total                   -0.14
Name: death_new, dtype: float64
```

Dickey-fuller test

```
ADF Statistic: -0.131594
p-value: 0.946199
Critical Values:
        1%: -3.541
        5%: -2.909
        10%: -2.592
Time Series for cases_total is Non-Stationary
```

```
ADF Statistic: -0.896202
p-value: 0.789221
Critical Values:
        1%: -3.542
        5%: -2.910
        10%: -2.593
Time Series for death_total is Non-Stationary
```

```
ADF Statistic: -3.631895
p-value: 0.005180
Critical Values:
        1%: -3.546
        5%: -2.912
        10%: -2.594
Time Series for cases_new is Stationary
```

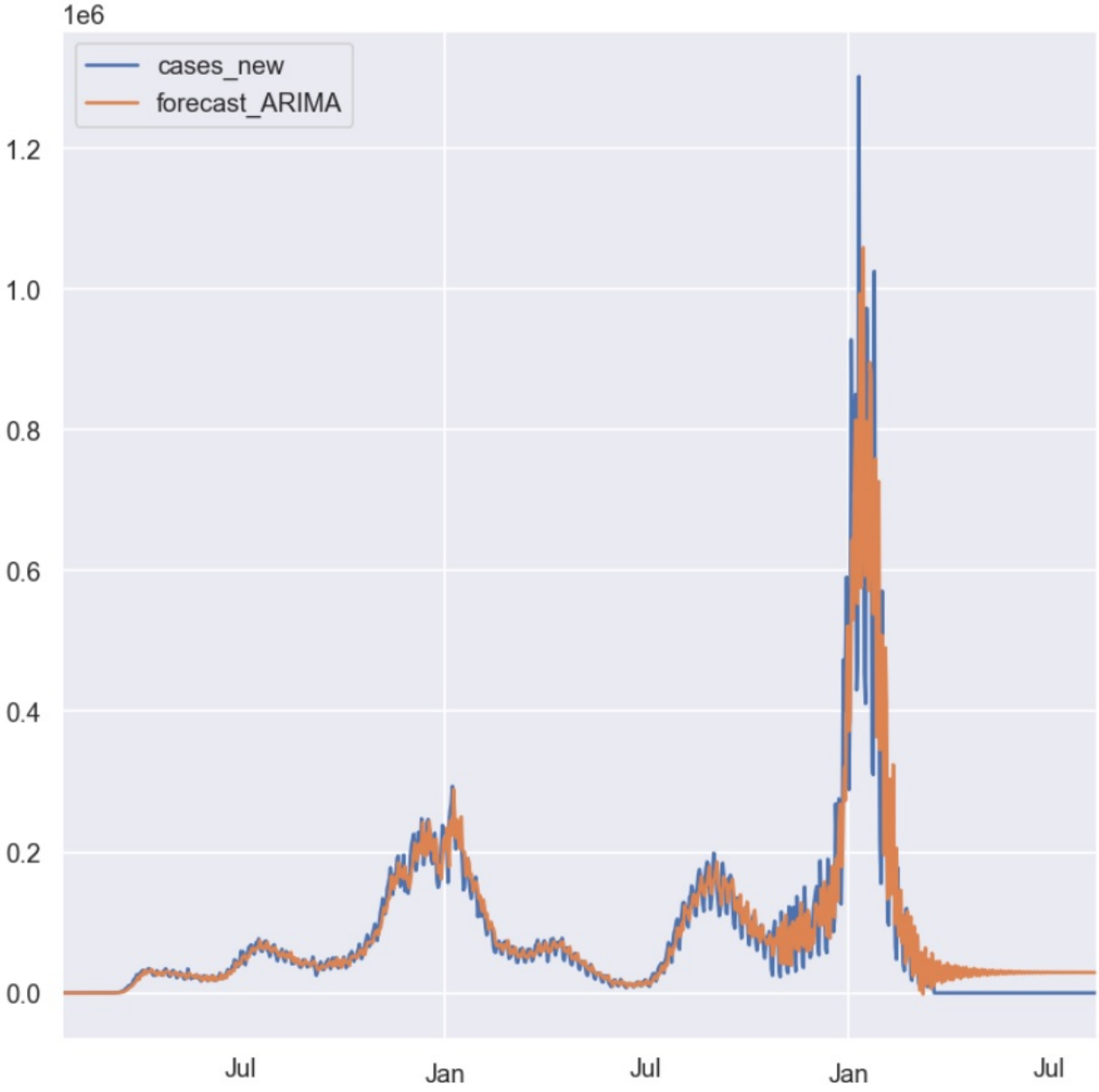
```
ADF Statistic: -2.384332
p-value: 0.146208
Critical Values:
        1%: -3.539
        5%: -2.909
        10%: -2.592
Time Series for death_new is Non-Stationary
```

ARIMA Model hyperparameter tuning

pmd.auto_arima to perform grid search for hyperparameters.

```
ARIMA(2, 1, 3) RMSE=94874.740
ARIMA(2, 1, 4) RMSE=103487.687
ARIMA(2, 2, 0) RMSE=144193.584
ARIMA(2, 2, 1) RMSE=107438.719
ARIMA(2, 2, 2) RMSE=108037.552
ARIMA(2, 2, 3) RMSE=110915.826
ARIMA(2, 2, 4) RMSE=105560.529
ARIMA(3, 0, 0) RMSE=106514.623
ARIMA(3, 0, 1) RMSE=109793.606
```

ARIMA model implementation



SARIMAX Results

Dep. Variable:	cases_new	No. Observations:	787
Model:	ARIMA(2, 1, 3)	Log Likelihood	-9638.465
Date:	Mon, 28 Mar 2022	AIC	19288.930
Time:	22:47:01	BIC	19316.931
Sample:	01-22-2020	HQIC	19299.695
	- 03-18-2022		
Covariance Type:	opg		

	coef	std err	z	P>|z|	[0.025	0.975]
ar.L1	-0.5038	0.007	-74.777	0.000	-0.517	-0.491
ar.L2	-0.9362	0.007	-132.528	0.000	-0.950	-0.922
ma.L1	0.0827	0.014	5.991	0.000	0.056	0.110
ma.L2	0.4486	0.015	29.936	0.000	0.419	0.478
ma.L3	-0.5493	0.016	-35.143	0.000	-0.580	-0.519
sigma2	2.929e+09	1.31e-12	2.24e+21	0.000	2.93e+09	2.93e+09

Ljung-Box (L1) (Q):	0.45	Jarque-Bera (JB):	49755.50
Prob(Q):	0.50	Prob(JB):	0.00
Heteroskedasticity (H):	309.72	Skew:	2.47
Prob(H) (two-sided):	0.00	Kurtosis:	41.66

ARIMA forecasting

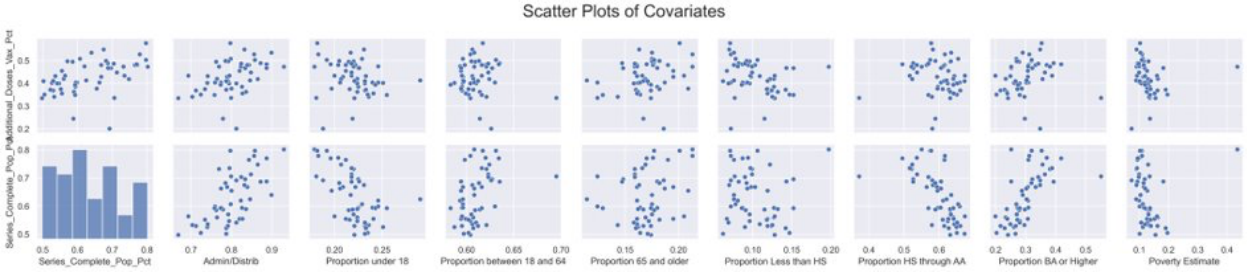
	cases_new	forecast_ARIMA
2022-03-19	0	9876.149889
2022-03-20	0	34086.353168
2022-03-21	0	44776.290649
2022-03-22	0	16724.378579
2022-03-23	0	20848.637971
2022-03-24	0	45033.907564
2022-03-25	0	28988.148654
2022-03-26	0	14428.981333
2022-03-27	0	36786.400690
2022-03-28	0	39153.509873
2022-03-29	0	17029.258808
2022-03-30	0	25959.260503

```
df_st['date'].max()
```

```
Timestamp('2022-03-18 00:00:00')
```

Covariate analysis

Covariate associations



It appeared that most of the demographic covariates had some relationship with each other and the target vaccination rates for model A (completed vaccination series) and model B (additional booster vaccination). The complete series vaccination data appeared to have stronger individual associations between the different demographic data when compared to the booster shot data. These plots indicated that vaccine administration-distribution rates, proportion of people under 18, people with at least a bachelor's degree, and estimated poverty rates would have the strongest impact on the models. There also appeared to be a few consistent outliers in the dataset.

Model selection

Although a best subset selection method was used to identify models of best fit for model A and B, in both cases the chosen model did not have the lowest AIC. This was largely due to the issue of multicollinearity between covariates and the practical significance of model interpretations.

Model A		
	Model AIC	Variables Included
Top Model	-164.0745	Administered-Distributed vaccine ratio, 65+ age, <HS education, HS-AA education, BS+ education
Chosen Model	-162.7897	Administered-Distributed vaccine ratio, 65+ age, <HS education, BS+ education

Model B		
	Model AIC	Variables Included
Top Model	-137.3185	Completed vaccination series, HS-AA education
Chosen Model	-137.0857	Completed vaccination series, <HS education, BS+ education

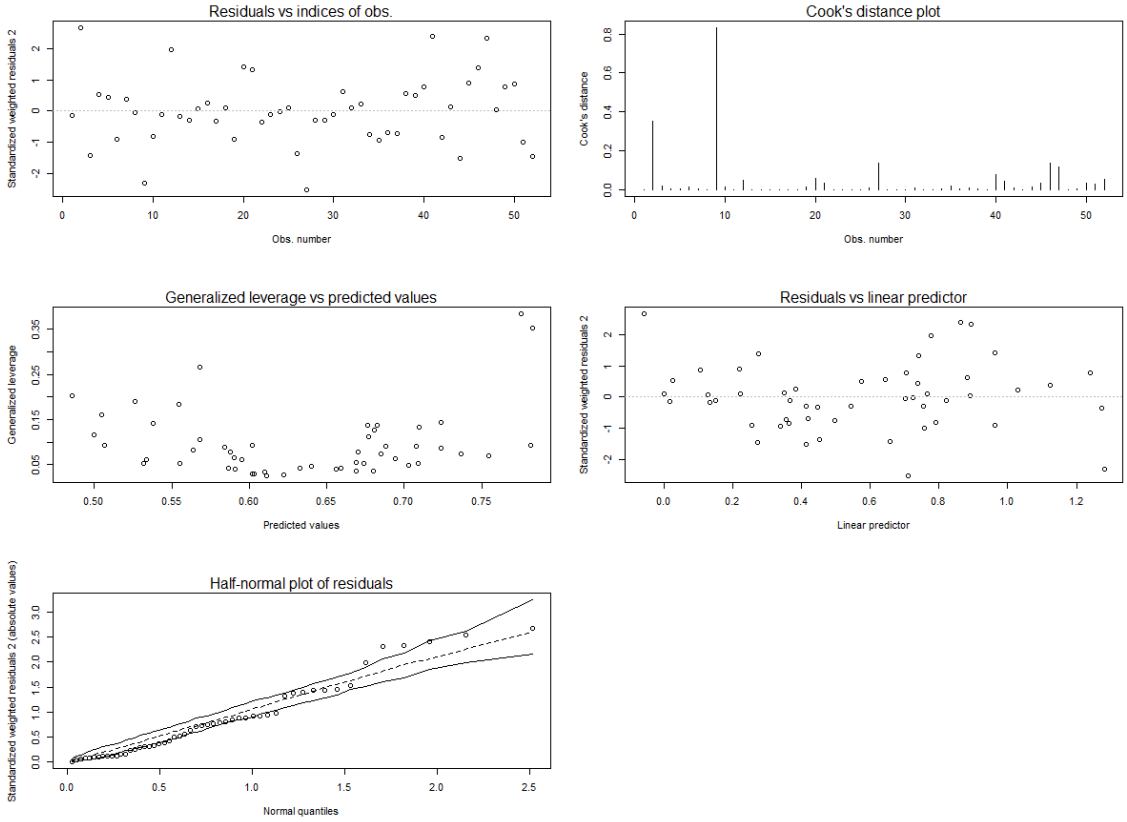
The choices for models A and B performed competitively with their respective top models but included more meaningful demographic data. The top choice for Model A also had some issues with multicollinearity between education levels.

Model implementation and diagnostics

Model A contained statistically significant coefficient estimates ($\alpha = 0.05$) and a good overall measure of fit for the data (Pseudo $R^2 = 0.739$). Variable inflation factors for each covariate included in the model were all low, indicating there were no issues with multicollinearity (VIF < 5).

Model A				
Variable	Coefficient Estimate (log-odds)	Odds Ratio	VIF	Pseudo R^2
Intercept	-4.3672	0.0127	---	0.739
Administration – Distribution Ratio	2.8421	17.1516	1.2659	
Age: 65+	6.3698	583.9139	1.0895	
Education: < High School	2.6211	13.7511	1.2490	
Education: Bachelor's +	4.3289	75.8613	1.5959	

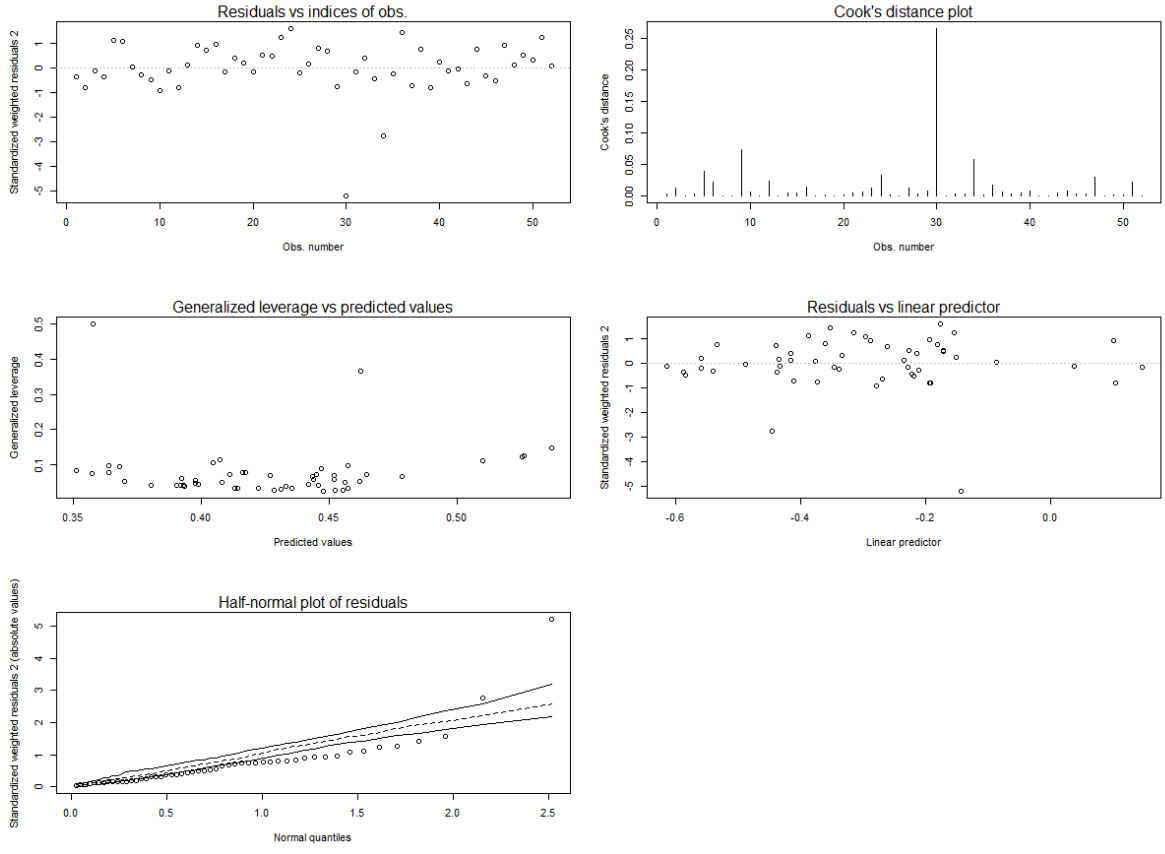
Diagnostic plots for model A confirm that were some issues with outliers in the data that impacted model performance. The residual plots show no distinct pattern, and the half-normal plot shows the standardized residuals were within normal expected variation.



Model B contained statistically significant estimated coefficients ($\alpha = 0.01$) but a weak overall measure of fit (Pseudo $R^2 = 0.339$). Variable inflation factors for each covariate included in the model were all low, indicating there were no issues with multicollinearity (VIF < 5).

Model B				
Variable	Coefficient Estimate (log-odds)	Odds Ratio	VIF	Pseudo R^2
Intercept	-0.7401	0.4771	---	0.339
Completed Vaccination Series	2.3963	10.9828	1.8162	
Education: < High School	-3.8969	0.0203	1.3127	
Education: Bachelor's +	-2.2429	0.1061	2.2121	

Diagnostic plots for model B also confirmed there were issues with outliers in the data. The residual plots showed there were some larger than expected residuals, and the half-normal plot showed a clear pattern in the standardized residuals.



Conclusions

COVID-19 exploratory analysis

Hospital data set was aggregated to state-level weekly data from hospital-level weekly data using pivot tables.

	uid	total_beds	total_icu	used_inpatient	used_icu
0	AK202031	11326.0	1413.0	4852.0	725.0
1	AK202032	10910.0	1342.0	4963.0	687.0
2	AK202033	10743.0	1512.0	4957.0	761.0
3	AK202034	11211.0	1524.0	4766.0	831.0
4	AK202035	11153.0	1526.0	4757.0	733.0

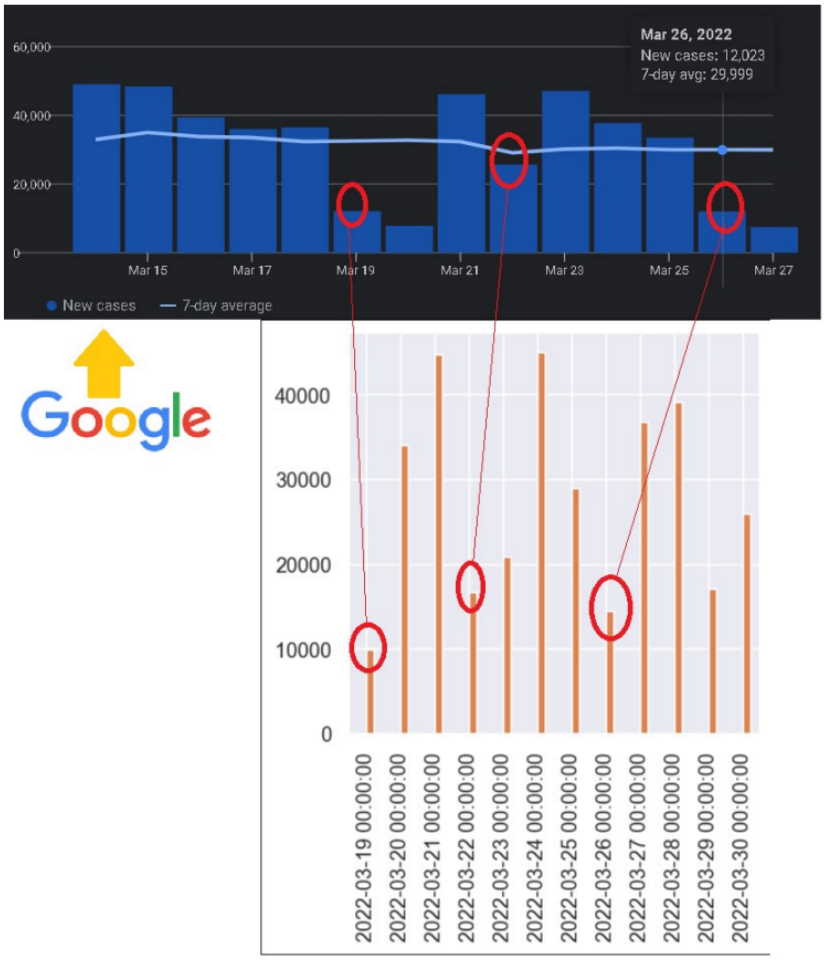
The unique identifier was successfully generated to identify rows across various data sets uniquely for a state for a specific week and has been utilized to join rows from multiple data sets matching a specific row for a specific state for a given week. The clean master data set that incorporates COVID19 cases deaths hospitalizations and vaccinations for a given week for a specific state was generated.

	uid	state	date	mmwr_week	week	distributed_janssen	series_complete_janssen	distributed_moderna	series_complete_moderna	distributed_pfizer	...
0	PR202209	PR	2022-03-06	10	09	213900	139409	2563560	950631	4522430	...
1	AR202209	AR	2022-03-06	10	09	251400	111894	2469560	658018	3182260	...
2	OK202209	OK	2022-03-06	10	09	321500	142105	2921880	842149	3768850	...
3	MS202209	MS	2022-03-06	10	09	224000	84654	2144080	579638	2951155	...
4	MO202209	MO	2022-03-06	10	09	423000	235711	4217400	1153591	6471845	...

The state of California preferred the Pfizer-BioNTech vaccine over the other two vaccines for the primary vaccination doses, and they preferred the Moderna vaccine over the other two for the booster doses. Michigan preferred the Pfizer-BioNTech vaccine over the other two vaccines for the primary series of vaccinations, and they preferred the Moderna vaccine for the booster dose over the other two vaccines. Vaccination in Michigan reduced the covid related deaths considerably. The distribution of vaccines among states appears to be showing that California has received the highest number of vaccines but when this distribution when adjusted for population and visualized we see that Vermont and Rhode Island have the highest vaccinated States and Vermont is the highest boosted state.

ARIMA modeling

We have seen a recent are you off Corvette case in January 2022 and currently, the case numbers seem to be at an average value of 30,000 cases as of March. Based on the correlation plots and value computations we see that vaccination is negatively correlated with the COVID-19 related death variable which implies that vaccination can reduce covid related deaths. Using the Dickey-Fuller test we can conclude that the new cases variable is stationary. Based on the ARIMA model hyperparameter tuning using the *autoarima* function the grid search resulted in ARIMA(2,1,3) as the model with the lowest RMSE error. The ARIMA(2,1,3) model generated was able to forecast COVID19 cases close to the actual cases as compared with the Google search results below for the selected period. It is interesting to note the model predicted the dip in cases compared to the neighboring days, on March 19th, March 22nd, and March 26th, and the value was close to the real world 7-day average of 30,000.



Covariate analysis

The results of the best subset model selection for model A indicated that the estimated poverty proportion and age levels of those under 18 and between 18 – 64 were all rarely used for modeling. For model B, the least used demographic factors were the vaccine administration – distribution ratio, and all age levels. These imply that none of these data would be important when discussing rates of complete series vaccinations or booster shots. Model A fit the data much better than model B. Both models suffered from influential outliers, diagnostic plots showed that outliers in model B were more severe and resulted in larger residuals.

For model A, all estimated coefficients corresponded to a positive increase in the odds of complete series vaccination rates and contained some substantially large odds ratios. The largest odds ratio was for the proportion of people 65 and older and corresponded to a 58,391% increase in odds of an increase in the average rate of completed vaccination series. In general, the interpretation of the odds ratios for model A is for every 1% increase in a given demographic covariate, and holding all other factors constant, there's a corresponding change in the odds of an increase in the average rate of completed vaccination series.

For model B, the rate of complete series vaccinations corresponded to a positive increase in the odds of booster shot rates, while education levels corresponded to a decrease in the odds of booster shot rates. While the interpretation for the rate of completed vaccination series is the same as for model A, the interpretation for education levels in model B is not. For model B, each additional increase in the rate of people with less than high school education or people with at least a bachelor's degree corresponds to a reduction (98% and 89% respectfully) in the odds of an increase in the average rate of booster vaccinations holding all other factors constant.

The estimated coefficient signs for the proportion of people with less than high school education in model A and the proportion of people with at least a bachelor's degree in model B appear to be opposite of what would be expected, given the scatter plots of all demographic factors in the final dataset. This result could largely be due to highly influential points (outliers) that skew the direction of the fitted model. However, both proportions of high and low levels of education were important for understanding rates of complete series vaccinations as well as rates of vaccine booster shots. This could indicate that average education levels are important when discussing vaccine hesitancy rates.

Future work

There are several aspects related to this project that would work on in the future. These covariates should be implemented into a time series model to possibly predict case surges, or potentially detect disease transmissions. Key biomarkers of COVID-19 could be used to predict genomic mutations or understand which environmental factors lead to genomic mutation. Models including demographic factors need to be updated to handle outliers and influential points to corroborate any results found in this project, including the addition of interaction terms. Lastly, these results may be useful in analyzing herd immunity prevalence.

References

Source code: https://github.com/SriSuryaSV/COVID-Data-Analysis

1. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
2. https://data.cdc.gov/Case-Surveillance/United-States-COVID-19-Cases-and-Deaths-by-State-o/9mfq-cb36
3. investopedia.com/terms/c/correlationcoefficient.asp
4. arxiv.org/abs/2111.03636
5. machinelearningplus.com/time-series/time-series-analysis-python/
6. machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
7. analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/
8. medium.com/analytics-vidhya/python-code-on-holt-winters-forecasting-3843808a9873
9. https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
10. https://www.machinelearningplus.com/time-series/arima-model-time-series-forecasting-python/
11. https://www.analyticsvidhya.com/blog/2020/10/how-to-create-an-arima-model-for-time-series-forecasting-in-python/
12. https://medium.com/swlh/temperature-forecasting-with-arima-model-in-python-427b2d3bcb53
13. https://www.ime.usp.br/~sferrari/beta.pdf
14. https://cran.r-project.org/web/packages/betareg/vignettes/betareg.pdf
15. http://applied-r.com/best-subset-regression-functions-r/