CSE 544 Project Report

By Group 11
Akhil Arradi(SBU ID:114353508)
Akhila Juturu (SBU ID: 114777498)
Sai Vikas Balabadhrapatruni(SBU ID:114777850)
Varshith Adavala(SBU ID:114778433)
Chirumamilla Swati Sree(SBU ID:114778895)
Sai Sujith Bezawada(SBU ID:114778433)

Note: The execution of the python script has been performed in google colab for running it faster than the local machine.

Given Dataset Links :

Cases dataset -
https://data.cdc.gov/api/views/9mfq-cb36/rows.csv?accessType=DOWNLOAD

Vaccinations dataset -
https://data.cdc.gov/api/views/unsk-b7fc/rows.csv?accessType=DOWNLOAD

Chosen Dataset Link(X) :

We have chosen-
https://www.marketwatch.com/investing/stock/jnj/download-data

States Allocation :

Each team has been assigned with 2 states. Similarly we have been assigned to California and Alaska.

Data Cleaning :

1) Check for missing values **and** remove them
3) Convert Date column to the desired date time format
2) We used Tukey's rule to **eliminate** outliers

Data Preprocessing :

- Removing all the null values

```
def remove_nulls(data):
  data=data[data.select_dtypes(include=[np.number]).ge(0).all(1)]
  return data.dropna(axis=0,how="any")
```

- Outliers can be detected using Tukey's rule

It is one of the most popular simple outlier detectors for one-dimensional number arrays. This approach assumes that for a given sample, we calculate first and third quartiles (Q1 and Q3), and mark all the sample elements outside the interval as outliers. Typical recommendation for k is 1.5 for “regular” outliers and 3.0 for “far outliers”.

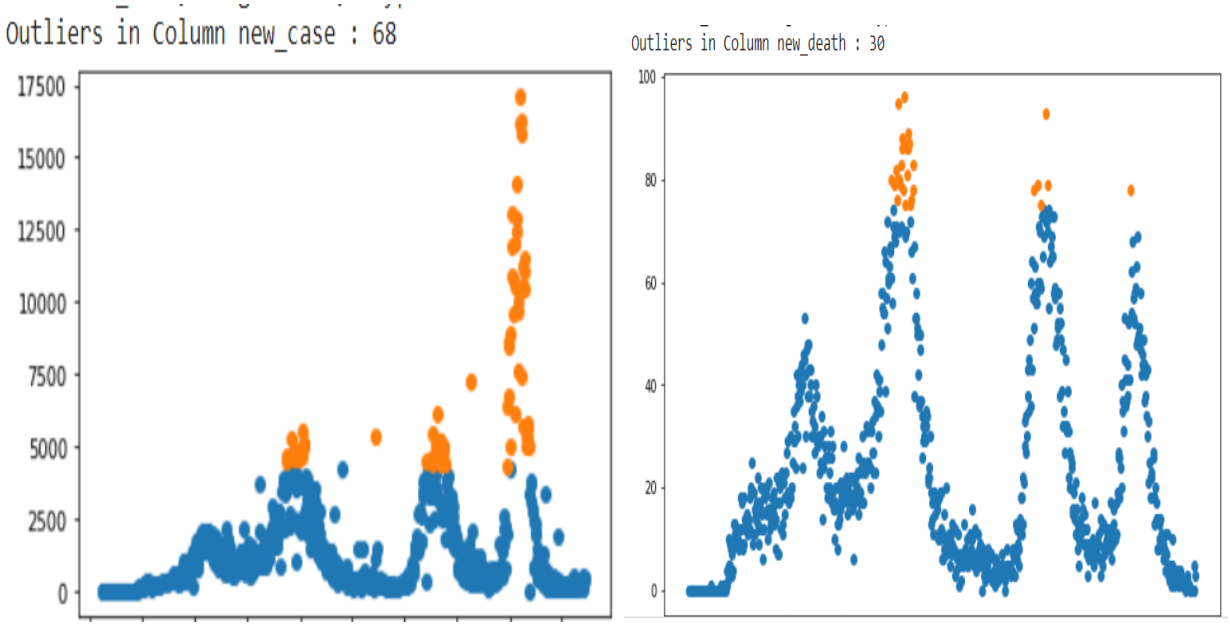
$$[Q_1 - k(Q_3 - Q_1),\ Q_3 + k(Q_3 - Q_1)]$$

Outlier’s Detection Code:

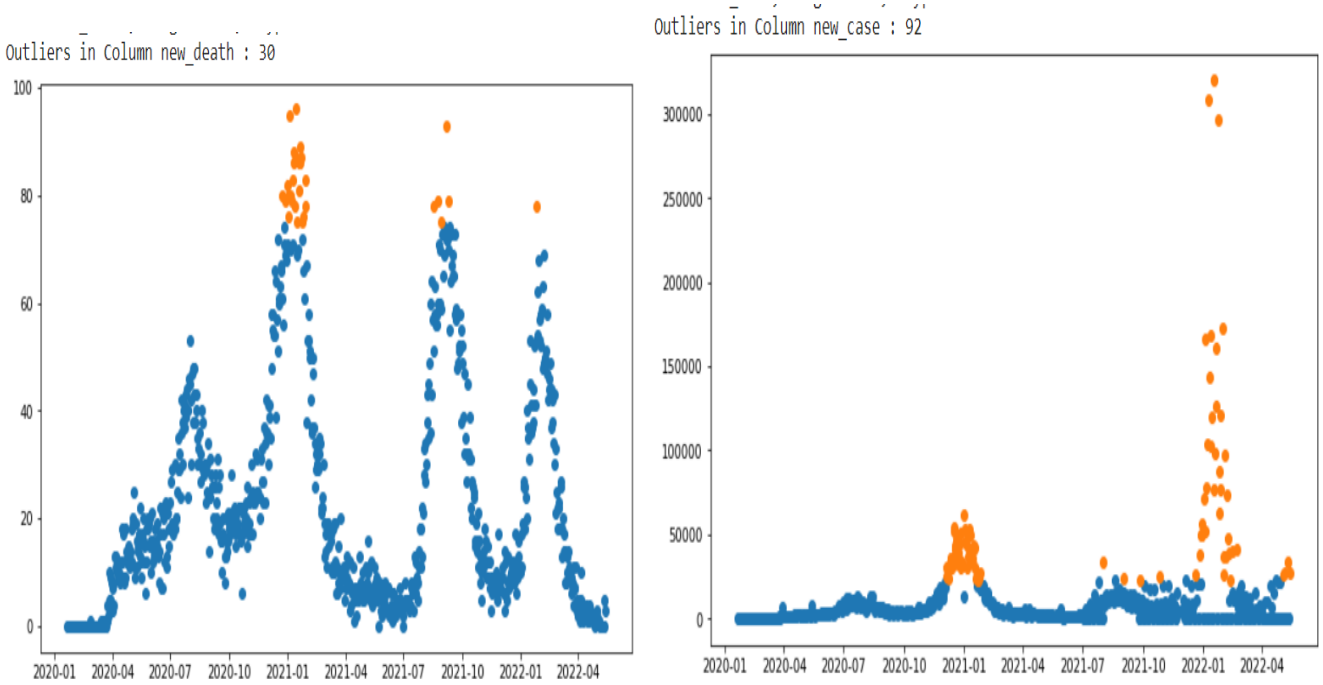
```
def remove_outliers(value,min_threshold,max_threshold):
  if value < min_threshold or value > max_threshold:
    return np.nan
  return value

def remove_all_outliers(data,columns):
  for column in columns:
    temp_data=data[column]
    q1=np.percentile(temp_data,25)
    q3=np.percentile(temp_data,75)
    iqr=q3-q1
    min_threshold,max_threshold=q1-1.5*iqr,q3+1.5*iqr
    data[column] = data.apply(lambda x: remove_outliers(x[column],min_threshold, max_threshold), axis=1)
  return remove_nulls(data).reset_index(drop=True)
```

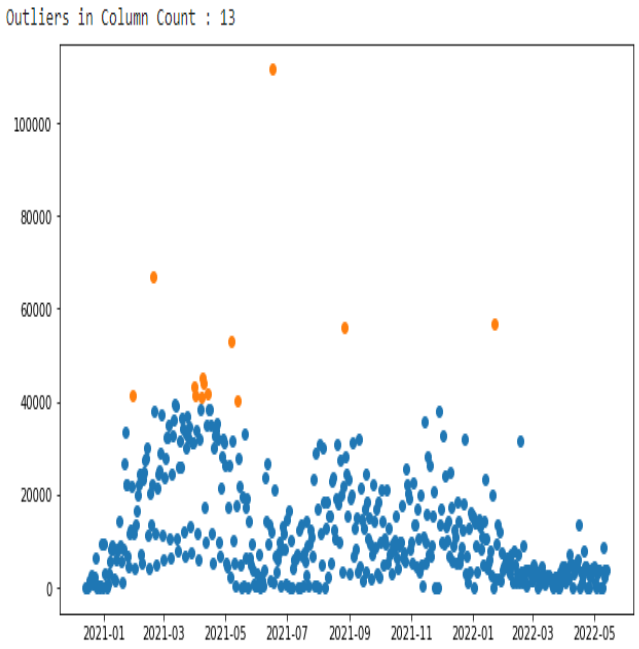
Dataset outliers(Alaska):

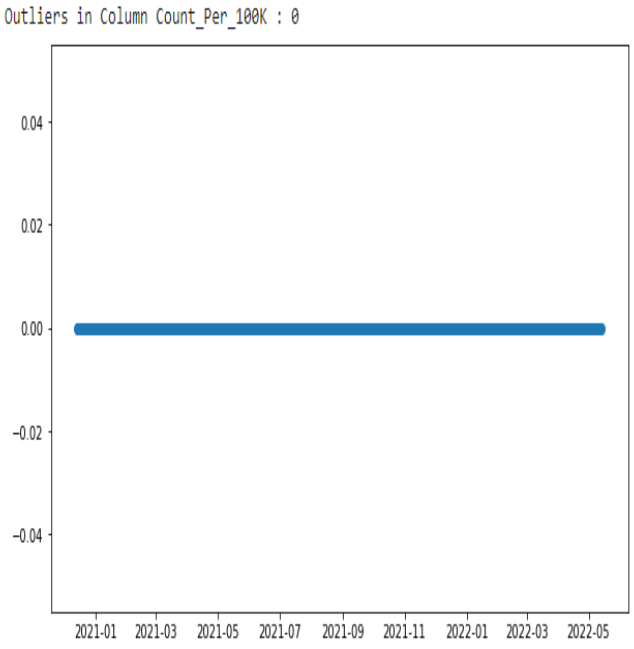


Dataset outliers(California):

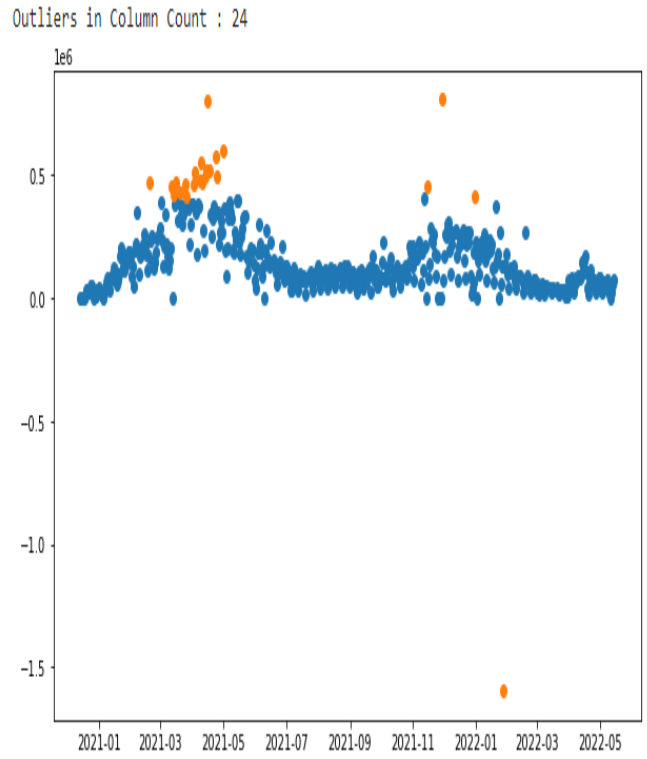


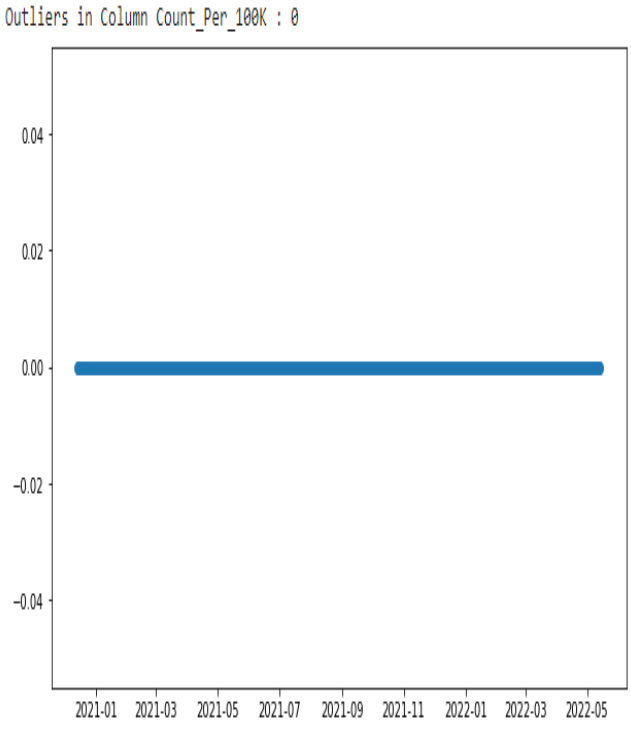
Vaccines dataset outliers:(Alaska)





Vaccines dataset outliers:(California)



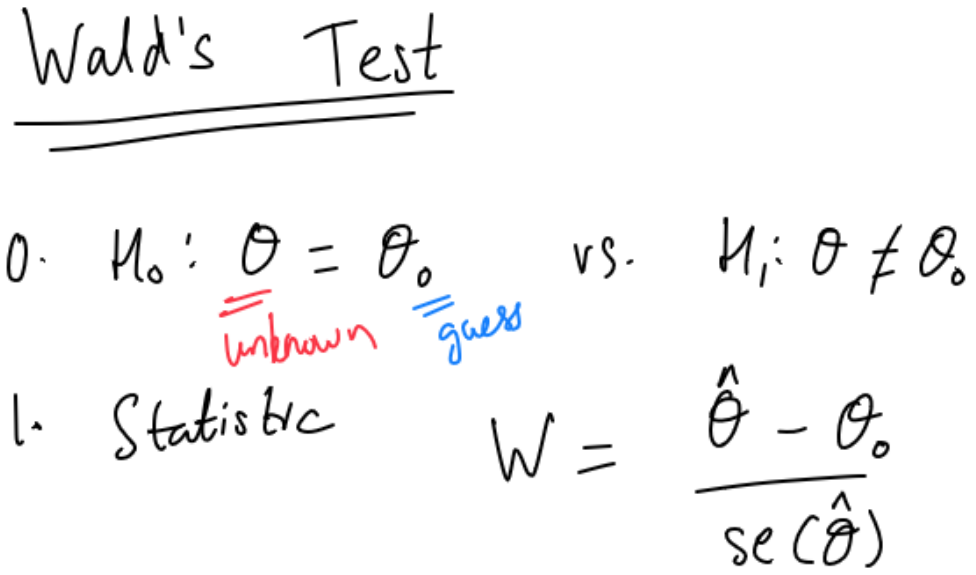


Inferences for COVID 19 dataset :

Task A :

Use different Hypothesis Test to identify how the mean of monthly COVID19 deaths and cases are different for Feb'21 and March'21 in the two states

Result of Walds 1 sample testing for mean of cases and death in CA:



Null hypothesis (H0) : Mean of daily cases/deaths in CA for March'21 is different from the corresponding mean of daily values for Feb'21.

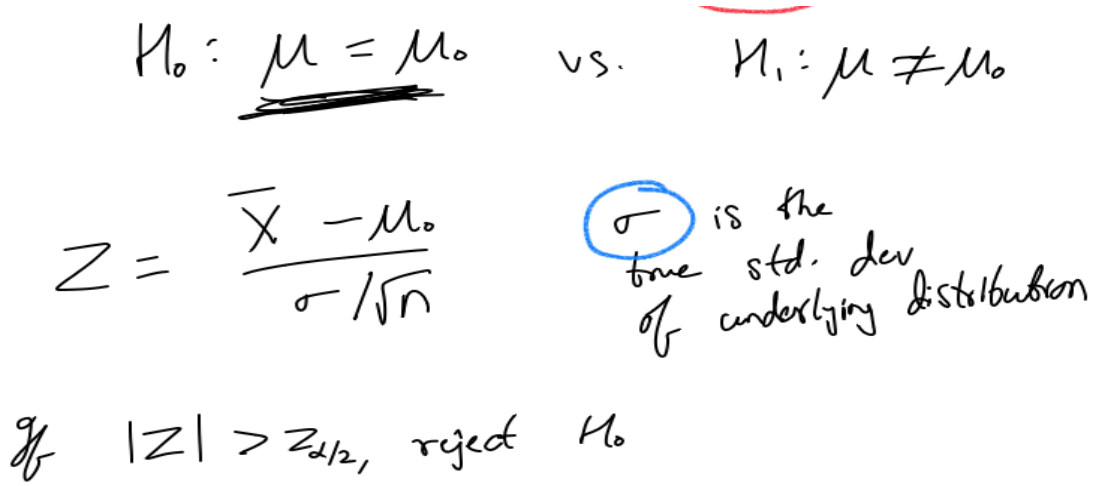
Alternate hypothesis(H1) : Mean of March'21 cases/deaths in CA is same as mean of Feb'21 cases/deaths.

Procedure : We took the u0 as mean of Feb'21 cases/deaths and alpha = 0.05 as given in documentation and the MLE estimator for mean of March'21 cases/deaths becomes the sample mean of cases/deaths .The standard error of the estimator is calculated in Wald's function(code file).

Result : As the w value for mean of March'21 death =**89.471** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

As the w value for mean of March'21 cases =**538.925** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

Result of Z testing for mean of cases and death for California (CA):



Null hypothesis (H0) : Mean of March'21 cases/deaths in CA different from Mean of Feb'21 cases/deaths in CA.

Alternate hypothesis(H1) : Mean of March'21 cases/deaths in CA is same as Mean of Feb'21 cases/deaths in CA.

Result /Inference:

As the Calculated Z value for mean of March'21 deaths in **CA = 10.616** which is greater than **1.96** we are **accepting** the NULL hypothesis.

As the Calculated Z value for mean of March'21 cases in **CA = 30.323** which is greater than **1.96** we are **accepting** the NULL hypothesis.

Is the Test Applicable ?

The main Assumptions of Z-test are the sample size has to be large or the sample data has to be normally dustributed.Here we can clearly see sample size is around 30 which satisfies the size threshold and through CLT data behaves as normal .

Hence ,We can conclude the Z Test is applicable on a given dataset.

Result of T 1 sample testing for mean of cases/death in CA:

$H_0: \mu = \mu_0 \quad \text{vs.} \quad H_1: \mu \neq \mu_0$

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

$S/\sqrt{n}$ → corrected sample std. dev. $= \sqrt{\frac{\sum (X_i - \bar{X})^2}{(n-1)}}$

Null hypothesis (H0) : Mean of March'21 cases/deaths is different from the mean of Feb'21 cases/deaths in CA.

Alternate hypothesis(H1) : Mean of March'21 cases/deaths is same as mean of Feb'21 cases/deaths in CA.

Procedure : We have taken the alpha = 0.05,n =30 as we took 30 days of data as given in the task and calculated the numerator and denominator of T in the above T_one_sample_testing function.

Result : As the calculated value for mean of March'21 deaths in **CA = 10.443** which is greater than **T value 2.045** we are **accepting** the NULL hypothesis.

As the calculated value for mean of March'21 cases in **CA = 29.830** which is greater than **T value 2.045** we are **accepting** the NULL hypothesis.

Is T-1 sample Test Applicable ?

The main assumption of T-test is the data is normally distributed, here it is not but as size n is around 30 and through CLT data behaves as T distribution.

	Statistical value for cases(with mod)	Result(cases) Accept/Reject Null Hypothesis	Statistical value for deaths	Result(deaths) Accept/Reject Null Hypothesis
Wald's test	**538.925**	Accept	**89.471**	Accept
Z-test	**30.323**	Accept	**10.616**	Accept
T-test	**29.830**	Accept	**10.443**	Accept

Result of Walds 1 sample testing for mean of cases and death in Alaska(AL):

Null hypothesis (H0) : Mean of daily cases/deaths in AL for March'21 is different from the corresponding mean of daily values for Feb'21.

Alternate hypothesis(H1) : Mean of March'21 cases/deaths in AL is same as mean of Feb'21 cases/deaths.

Procedure : We took the u0(guess value) as mean of Feb'21 cases/deaths and alpha = 0.05 as given in task and the MLE estimator for mean of March'21 cases/deaths becomes the sample mean of cases/deaths .The standard error of the estimator is calculated in above walds function.

Result :

As the w value for mean of March'21 **death =34.325** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

As the w value for mean of March'21 **cases =100.517** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

Result of Z testing for mean of cases and death for Alaska (AL)

Null hypothesis(H0) : Mean of March'21 cases/deaths in AL different from Mean of Feb'21 cases/deaths in AL.

Alternate hypothesis(H1) : Mean of March'21 cases/deaths in AL is same as Mean of Feb'21 cases/deaths in AL.

Result /Inference :

As the Calculated Z value for mean of March'21 deaths in **AL = 27.235** which is greater than 1.96 we are **accepting** the NULL hypothesis.

As the Calculated Z value for mean of March'21 cases in **AL = 3.499** which is greater than 1.96 we are **accepting** the NULL hypothesis.

Is the Test Applicable ?

The main Assumptions of Z-test are the sample size has to be large or the sample data has to be normally dustributed.Here we can clearly see sample size is around 30 which satisfies the size threshold and through CLT data behaves as normal .

Hence ,We can conclude the Z Test is applicable on a given dataset.

Result of T 1 sample testing for mean of cases/death in AL:

Null hypothesis(H0) : Mean of March'21 cases/deaths is different from the mean of Feb'21 cases/deaths in AL.

Alternate hypothesis(H1) : Mean of March'21 cases/deaths is same as mean of Feb'21 cases/deaths in AL.

Procedure : We have taken the alpha = 0.05,n =30 as we took 30 days of data as given in task and calculated the numerator and denominator of T in the above T_one_sample_testing function

Result : As the calculated value for mean of March'21 **deaths** in AL = **26.792** which is greater than T value 2.045 we are **accepting** the NULL hypothesis.

As the calculated value for mean of March'21 **cases** in AL= **3.442** which is greater than T value 2.045 we are **accepting** the NULL hypothesis.

Is T-1 sample Test Applicable ?

The main assumption of T-test is the data is normally distributed, here it is not but as size n is around 30 and through CLT data behaves as T distribution.

	Statistical value for cases	Result(cases) Accept/Reject Null Hypothesis	Statistical value for deaths	Result(deaths) Accept/Reject Null Hypothesis
Wald's test	**100.517**	Accept	**34.325**	Accept
Z-test	**3.499**	Accept	**27.235**	Accept
T-test	**3.442**	Accept	**26.792**	Accept

Result of Walds 2 sampled test for mean of cases and death in Alaska(AL) and California(CA):

Null hypothesis (H0):

Mean of daily cases/deaths in AL/CA for March'21 is different from the corresponding mean of daily values for Feb'21

Alternate hypothesis(H1):

Mean of March'21 cases/deaths in AL/CA is same as mean of Feb'21 cases/deaths

Procedure :

We took the u0(guess value) as mean of Feb'21 cases/deaths and alpha = 0.05 as given in documentation and the MLE estimator for mean of March'21 cases/deaths becomes the sample mean of cases/deaths .The standard error of the estimator is combination of the standard error of both the time frame data March and Feb.

Result:

As the w statistic value for deaths in **AL =17.1367** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

As the w statistic value for cases in **AL =59.526** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

As the w statistic value for death in **CA =48.730** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

As the w statistic value for cases in **CA =267.295** which is greater than 1.96 we are **Accepting** the NULL hypothesis.

Result of T 2-sample unpaired testing for mean of cases and death for CA/AL

Null hypothesis (H0):

Mean of March'21 cases/deaths in CA/AL is different from Mean of Feb'21 cases/deaths.

Alternate hypothesis(H1):

Mean of March'21 cases/deaths in CA/AL is the same as the mean of Feb'21 cases/deaths.

Procedure :

We have taken the alpha = 0.05 n=30, m=28 as given in the task and calculated the numerator and denominator of T value in the above t_2Sampled_test function .

Result:

As the T-Statistic value for cases in **CA = 7.767** which is greater than 2.18 we are **accepting** the NULL hypothesis.

As the T-Statistic value cases in **AL = 2.760** which is less than 2.18 we are **accepting** the NULL hypothesis.

As the T-statistic value for deaths in **AL = 9.343** which is greater than 2.18 we are **accepting** the NULL hypothesis.

As the T-statistic value for deaths in **CA = 4.811** which is greater than 2.18 we are **accepting** the NULL hypothesis.

Task B:

Inference the equality of distributions between the two states (distribution of daily #cases and daily #deaths) for the last three months of 2021 (Oct, Nov, Dec)

Permutation Test Function :

The permutation test is used to check whether two data samples follow the same distribution.

$D_1 = \{X_1 \dots X_n\} \overset{iid}{\sim} X \qquad D_2 = \{Y_1 \dots Y_m\} \overset{iid}{\sim} Y$

$H_0 : X \overset{d}{=} Y \quad \text{vs.} \quad H_1 : X \overset{d}{\neq} Y$

Assumptions : None!

Result of Permutation test for two state cases:

Null hypothesis(H0) : Distribution of state Alaska cases equals distribution of state California cases

Alternate hypothesis(H1) : Distribution of state Alaska cases not equals to distribution of state California cases

Procedure : We arranged all Alaska state records and California records data 1000 ways. We take alpha = 0.05, as stated in the documentation, and calculate it.

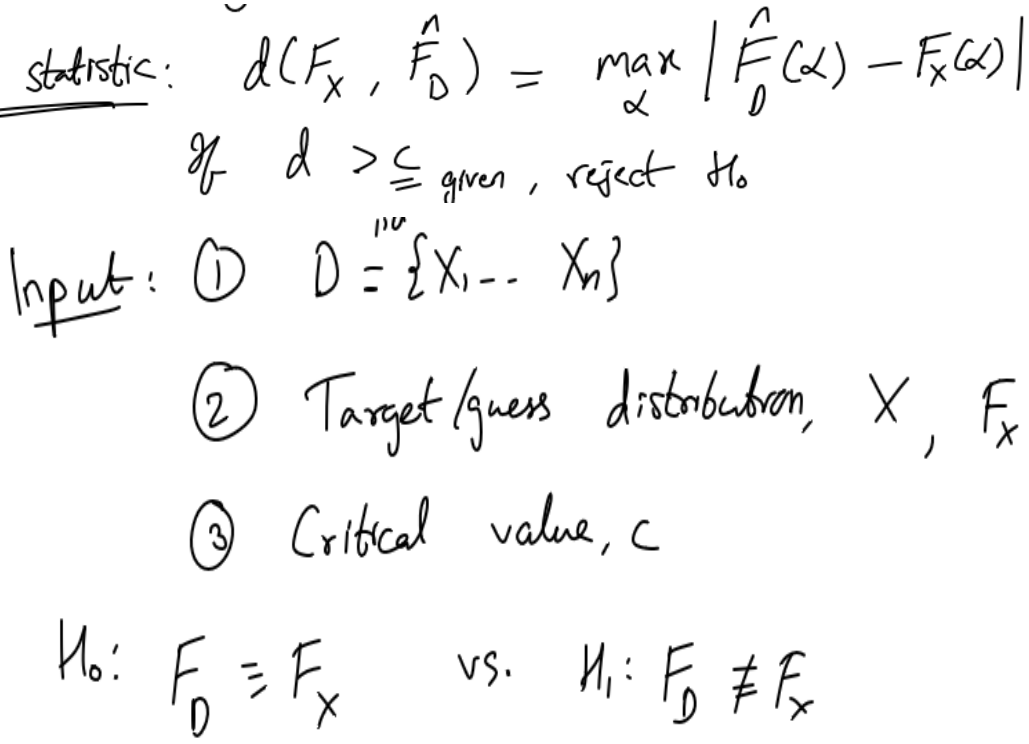
Is the Permutation Test applicable ?

There are no assumptions under the Permutation test, hence the test is applicable.

	P-value for daily cases	Result(cases) Accept/Reject Null Hypothesis	P-value for daily deaths	Result(cases) Accept/Reject Null Hypothesis
Permutation Test	0	Reject	0	Reject

K-S One Sample Test

The KS test is used to check whether two data samples follow the same distribution. It has zero assumptions



KS test for distribution of daily cases

1. Result of 1 sample KS test for for the last three months covid cases of 2021 with Poisson distribution

$$P(x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

λ- mean number of successes over a given interval

$Var(X) = \lambda$

E[X] for Poisson Distribution = lambda

Null hypothesis(H0) : Distribution of Oct-Dec 2021 cases equals poisson distribution

Alternate hypothesis(H1) : Distribution of Oct-Dec 2021 cases not equals poisson distribution

Procedure : We obtained the parameters for the Poisson distribution using MME on the second state data. We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Is KS testing applicable?

There are no assumptions in the KS test, so the test is applicable

Result : Since the statistical value of the KS test is ***1.0001***, greater than 0.05, we **reject** the null hypothesis.

Oct-Dec 2021 data for the second state does not follow the Poisson distribution.

2.Result of 1 sample KS test for for the last three months covid cases of 2021 with geometric distribution

Null hypothesis(H0) : Distribution of Oct-Dec 2021 cases equals geometric distribution.

Alternate hypothesis(H1) : Distribution of Oct-Dec 2021 cases not equals geometric distribution

Procedure : We obtained the parameters for the Geometric distribution using MME on the second state data. We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Result : Since the statistical value of the KS test is ***0.70382***, greater than 0.05, we **reject** the null hypothesis.

Oct-Dec 2021 data for the second state does not follow the geometric distribution.

3.Result of 1 sample KS test for for the last three months covid cases of 2021 with Binomial distribution

Result : Since the statistical value of the KS test is ***1***, greater than 0.05, we **reject** the null hypothesis.

Oct-Dec 2021 data for the second state does not follow the binomial distribution.

KS test for distribution of daily deaths :

1.Result of 1 sample KS test for for the last three months covid deaths of 2021 with Poisson distribution

Null hypothesis(H0) : Distribution of Oct-Dec 2021 deaths equals poisson distribution.

Alternate hypothesis(H1): Distribution of Oct-Dec 2021 deaths not equals poisson distribution.

Procedure : We obtained the parameters for the Poisson distribution using MME on the second state data. We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Is KS testing applicable?

There are no assumptions in the KS test, so the test is applicable.

Result : Since the statistical value of the KS test is ***0.994***, greater than 0.05, we **reject** the null hypothesis.

Oct-Dec 2021 data for the second state deaths data does not follow the poisson distribution.

2.Result of 1 sample KS test for for the last three months covid deaths of 2021 with geometric distribution

Since the statistical value of the KS test is ***0.908***, greater than 0.05, we **reject** the null hypothesis.

Oct-Dec 2021 data for the second state deaths data does not follow the geometric distribution.

3.Result of 1 sample KS test for for the last three months covid deaths of 2021 with binomial distribution

Since the statistical value of the KS test is ***1.0***, greater than 0.05, we **reject** the null hypothesis.

Oct-Dec 2021 data for the second state deaths data does not follow the binomial distribution.

KS Two Sample Test

KS test for distribution of daily cases :

1.Result of 2 sample KS test for the equality of distributions between the two states(cases)

Null hypothesis(H0) : Distribution of Oct-Dec 2021 cases of state Alaska equals distribution of Oct-Dec 2021 cases of state California.

Alternate hypothesis(H1) : Distribution of Oct-Dec 2021 cases of state Alaska not equals distribution of Oct-Dec 2021 cases of state California.

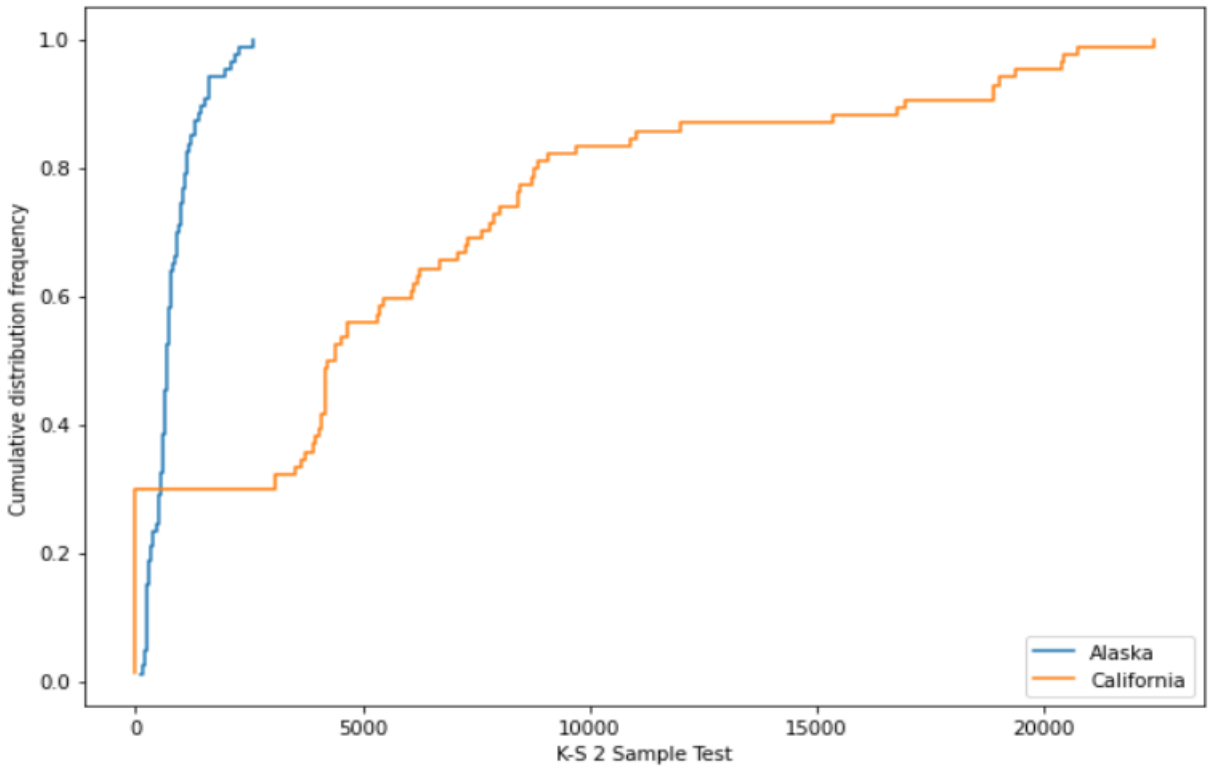
Procedure : We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Is the KS Test applicable ?

There are no assumptions under KS test, hence the test is applicable.

Result : Since the statistical value of the KS test is ***1.00***, greater than 0.05, we **reject** the null hypothesis.

The distribution of Oct-Dec 2021 cases of state Alaska does not equal the distribution of Oct-Dec 2021 cases of state California.



K-S statistic is 1.000000000000001 and point of max difference is 2582.0

KS test for distribution of daily deaths :

Result of 2 sample KS test for the equality of distributions between the two states(deaths)

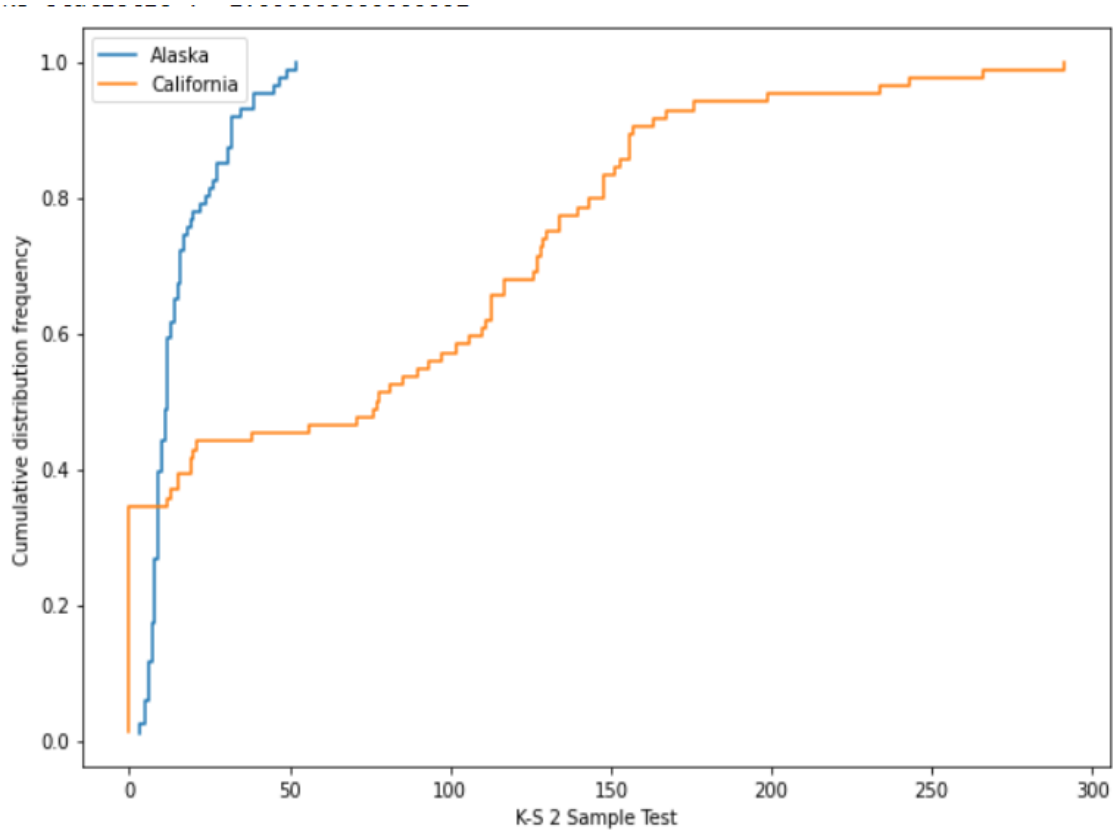
Null hypothesis(H0) : Distribution of Oct-Dec 2021 deaths of state Alaska equals distribution of Oct-Dec 2021 deaths of state California.

Alternate hypothesis(H1) : Distribution of Oct-Dec 2021 deaths of state Alaska not equals distribution of Oct-Dec 2021 deaths of state California.

Procedure : We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Result : Since the statistical value of the KS test is ***1.00***, greater than 0.05, we **reject** the null hypothesis.

The distribution of Oct-Dec 2021 deaths of state Alaska does not equal the distribution of Oct-Dec 2021 deaths of state California.

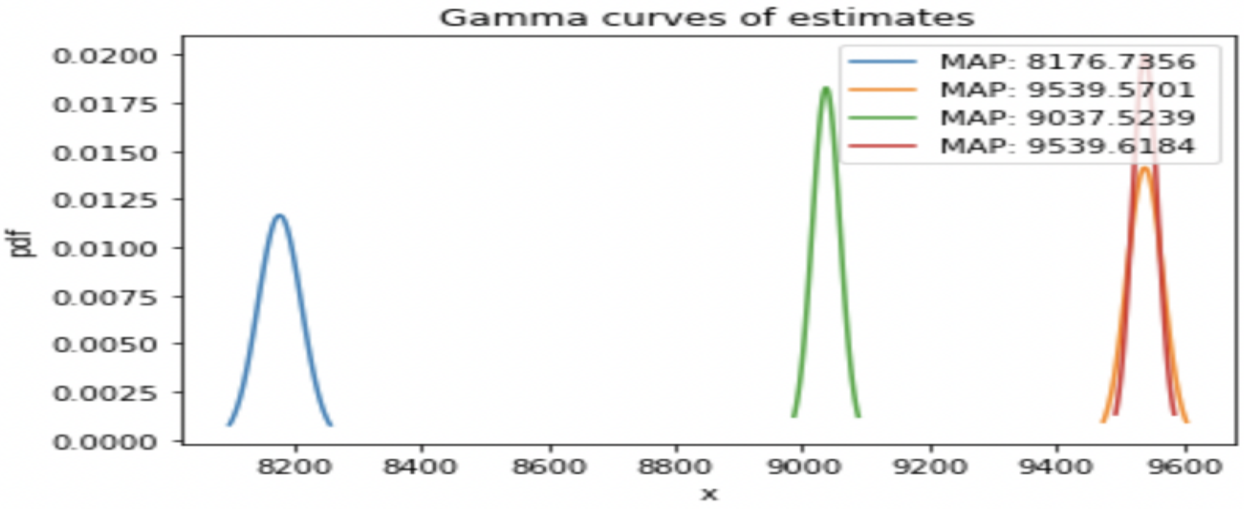


K-S statistic is 1.000000000000001 and point of max difference is 52.0

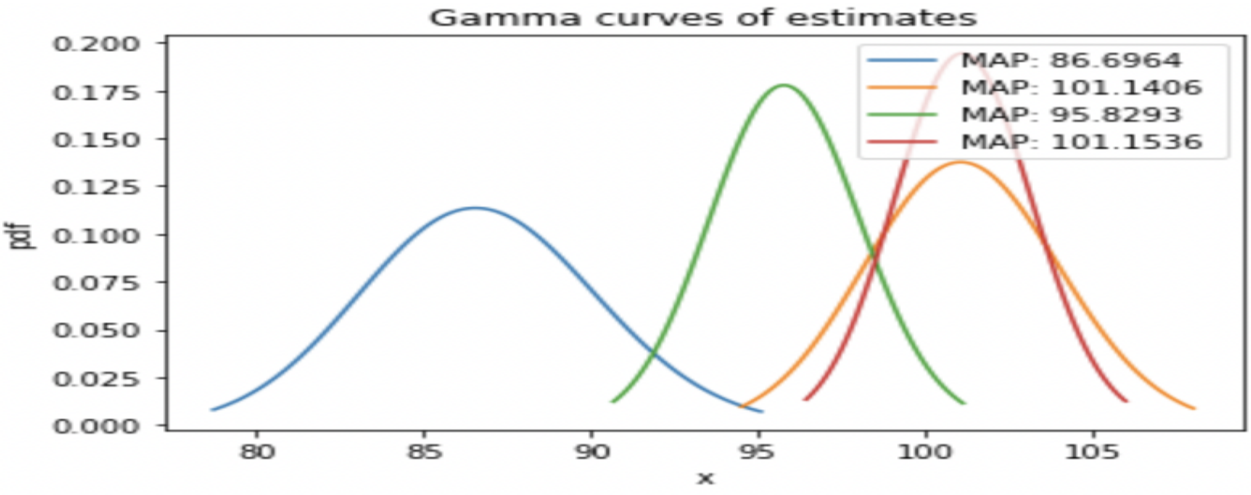
Task C:

Used the fifth week's data (June 29 to July 5) to obtain the posterior for λ via Bayesian inference. Then, used the sixth week's data to obtain the new posterior, using prior as posterior after week 5. Repeated till the end of week 8 and plotted the four week data along with their MAP for the posteriors.

Posterior Gamma distributions for California/Alaska 'Cases':



Posterior Gamma distributions for California/Alaska Deaths:



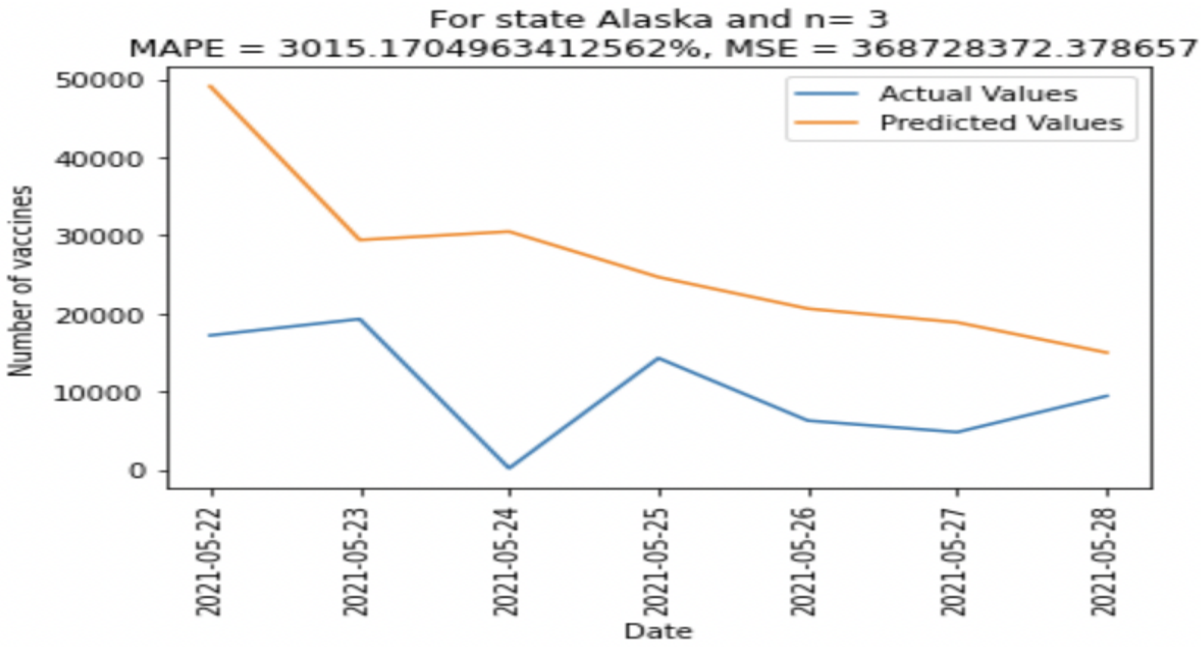
Task D:

Predicted the #vaccines for each state using the vaccinations dataset to predict the COVID19 #vaccines administered for the fourth week in May 2021 using data from the first three weeks of May 2021. Reported MAPE as a % and MSE for each forecasting technique and plotted the respective plots for the same.

Auto Regression:

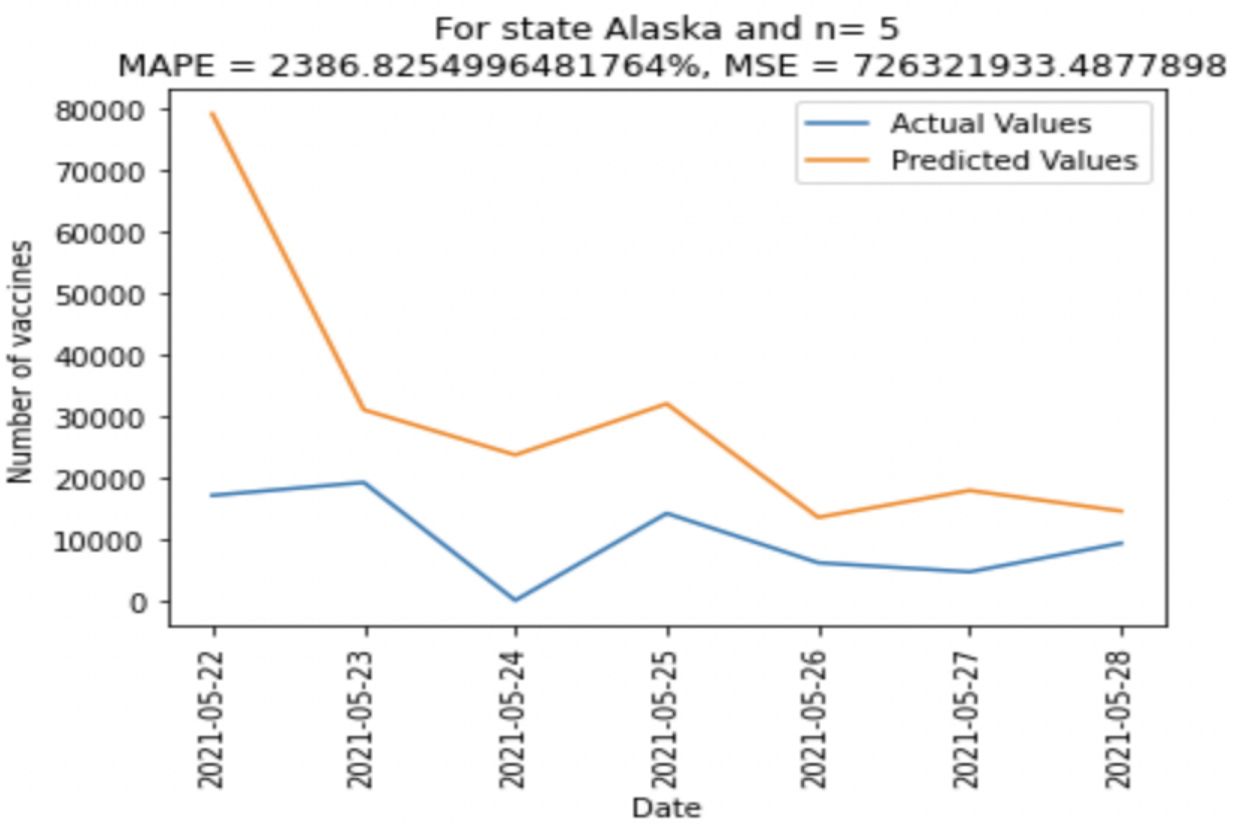
Predictions for #vaccines for Alaska using AR(3):

	Date	Actual Value	Predicted Value
0	2021-05-22	17173.0	49073.900452
1	2021-05-23	19270.0	29395.413647
2	2021-05-24	150.0	30469.637078
3	2021-05-25	14265.0	24637.811889
4	2021-05-26	6261.0	20597.314981
5	2021-05-27	4780.0	18839.548943
6	2021-05-28	9416.0	14968.539307



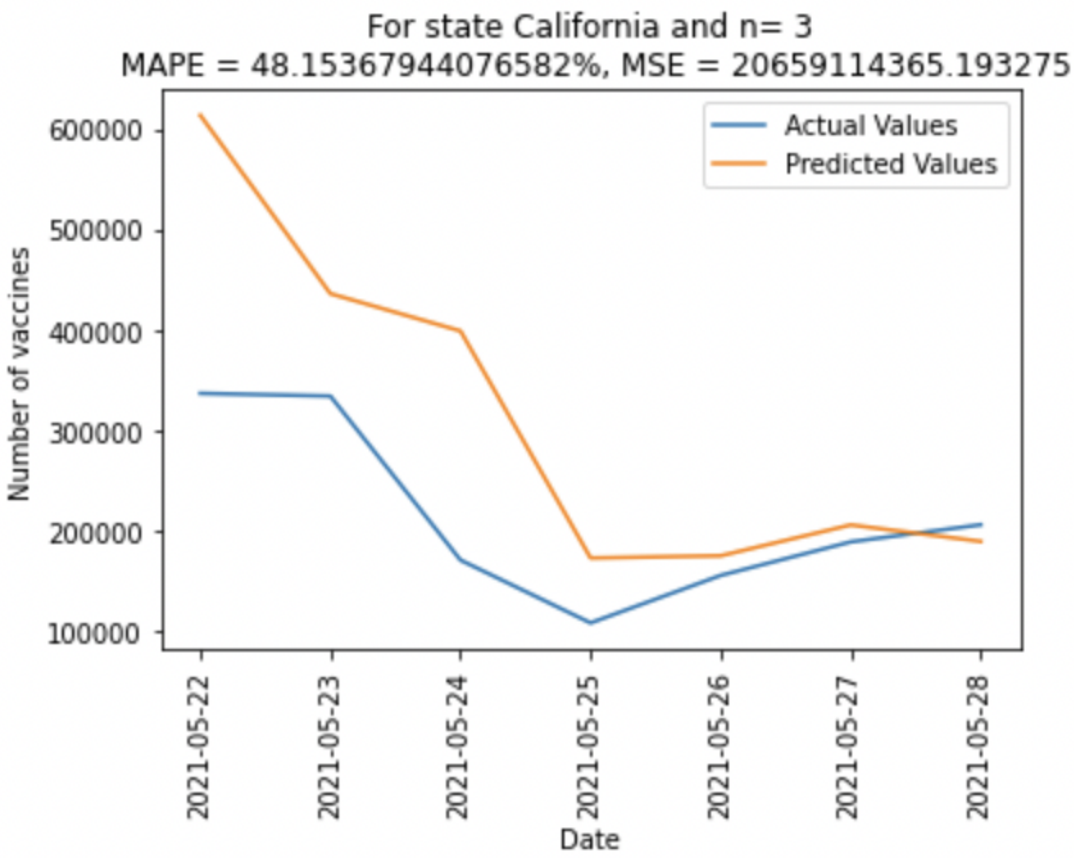
Predictions for #vaccines for Alaska using AR(5):

	Date	Actual Value	Predicted Value
0	2021-05-22	17173.0	78976.463523
1	2021-05-23	19270.0	31040.148743
2	2021-05-24	150.0	23720.445976
3	2021-05-25	14265.0	32021.272336
4	2021-05-26	6261.0	13607.287108
5	2021-05-27	4780.0	17972.148863
6	2021-05-28	9416.0	14631.090883



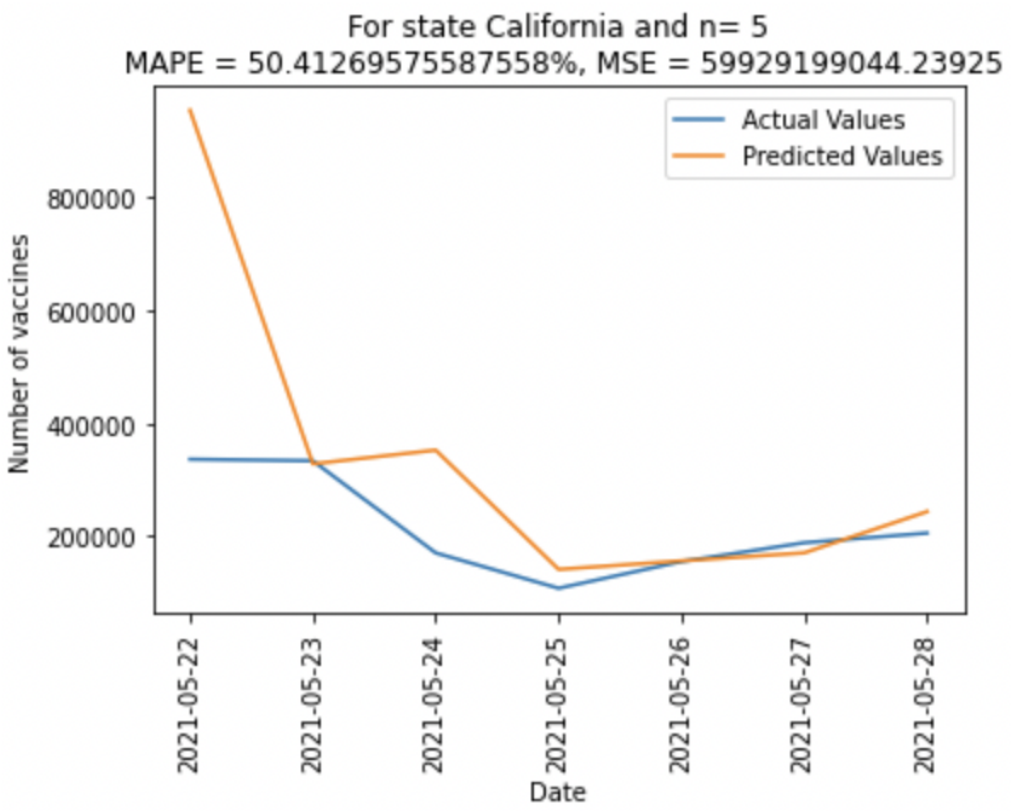
Predictions for #vaccines for California using AR(3):

	Date	Actual Value	Predicted Value
0	2021-05-22	336400.0	613689.756305
1	2021-05-23	333684.0	435770.667805
2	2021-05-24	170194.0	398607.372668
3	2021-05-25	107545.0	172242.444442
4	2021-05-26	154889.0	174467.637850
5	2021-05-27	188350.0	205221.218169
6	2021-05-28	205413.0	188796.301537



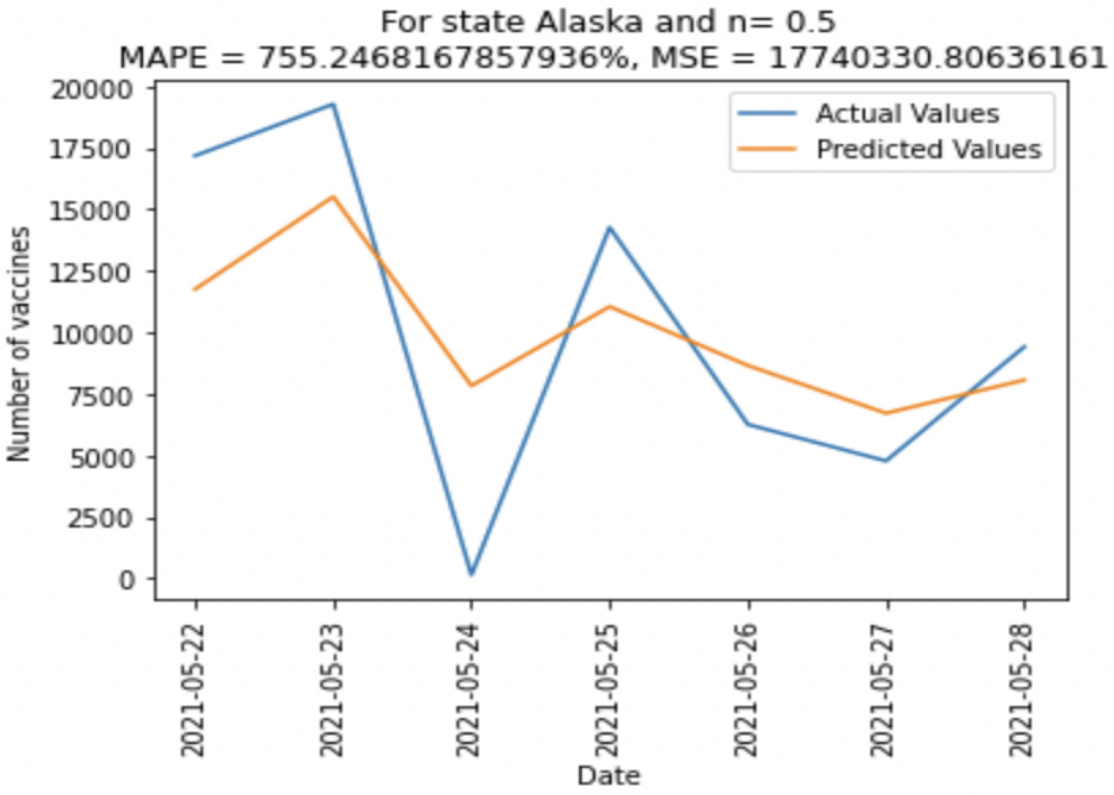
Predictions for #vaccines for California using AR(5):

	Date	Actual Value	Predicted Value
0	2021-05-22	336400.0	955525.070589
1	2021-05-23	333684.0	328173.202335
2	2021-05-24	170194.0	352557.254005
3	2021-05-25	107545.0	141027.300687
4	2021-05-26	154889.0	156102.773203
5	2021-05-27	188350.0	170357.281980
6	2021-05-28	205413.0	243564.531669



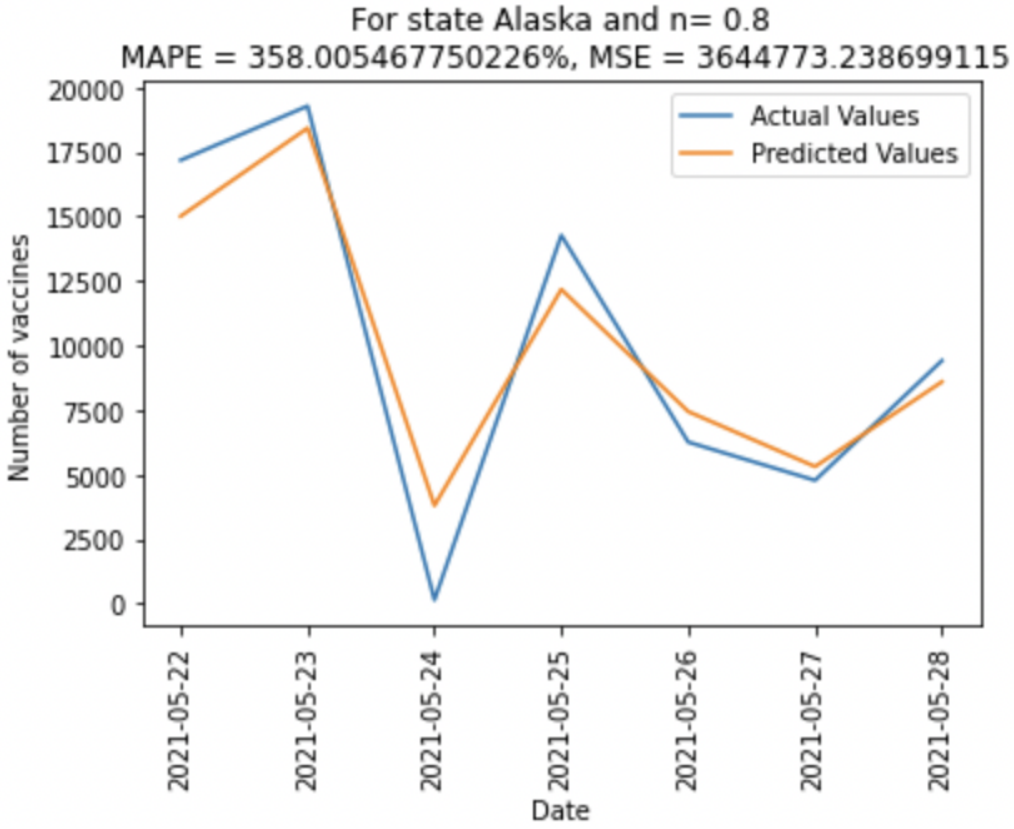
Predictions for #vaccines for Alaska using EWMA(n=0.5):

	Date	Actual Value	Predicted Value
0	2021-05-22	17173.0	11750.0000
1	2021-05-23	19270.0	15510.0000
2	2021-05-24	150.0	7830.0000
3	2021-05-25	14265.0	11047.5000
4	2021-05-26	6261.0	8654.2500
5	2021-05-27	4780.0	6717.1250
6	2021-05-28	9416.0	8066.5625



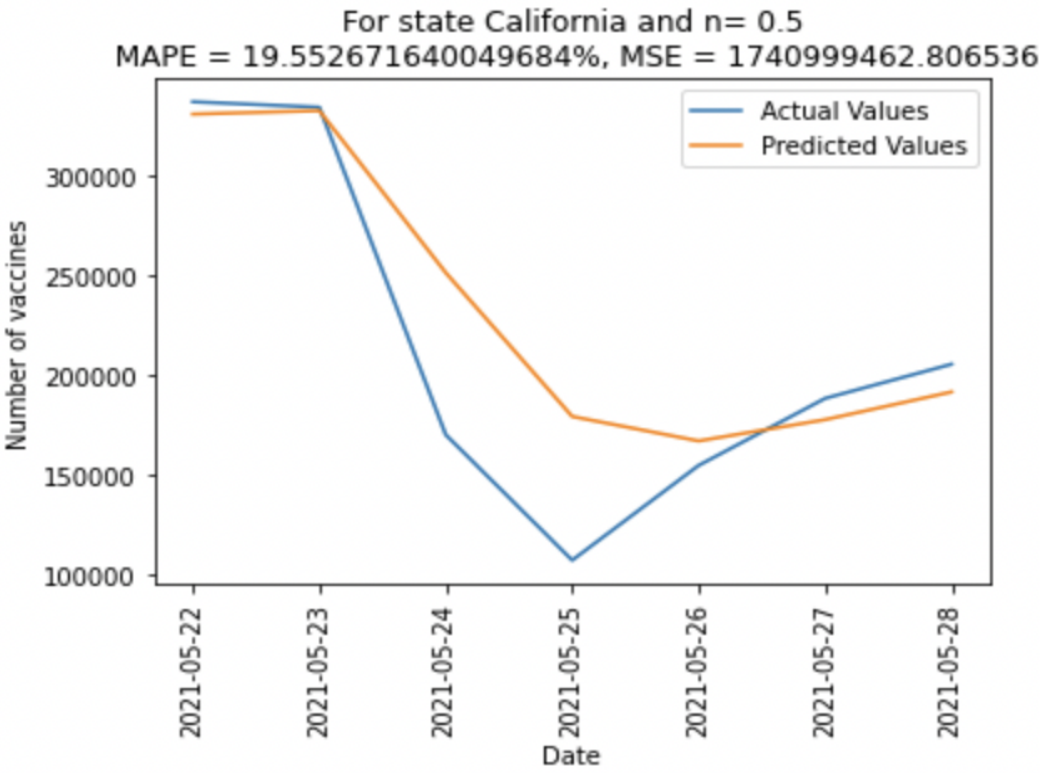
Predictions for #vaccines for Alaska using EWMA(n=0.8):

	Date	Actual Value	Predicted Value
0	2021-05-22	17173.0	15003.800000
1	2021-05-23	19270.0	18416.760000
2	2021-05-24	150.0	3803.352000
3	2021-05-25	14265.0	12172.670400
4	2021-05-26	6261.0	7443.334080
5	2021-05-27	4780.0	5312.666816
6	2021-05-28	9416.0	8595.333363



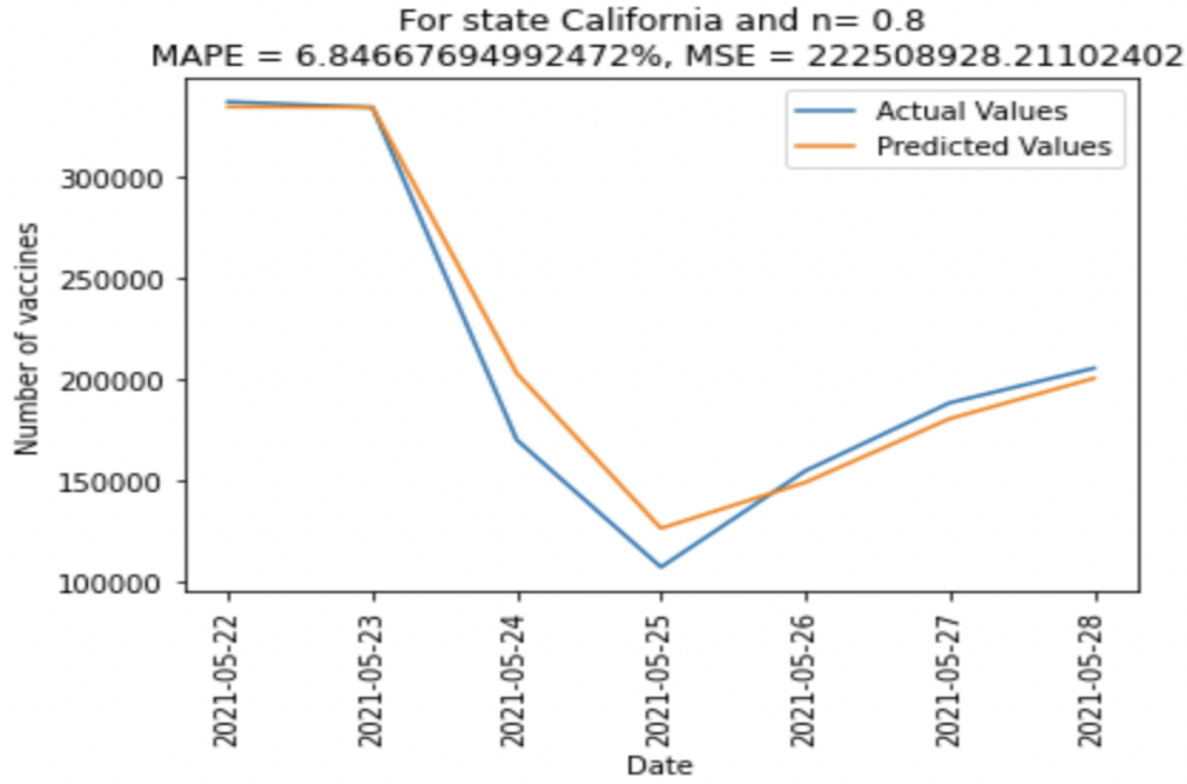
Predictions for #vaccines for California using EWMA(n=0.5):

	Date	Actual Value	Predicted Value
0	2021-05-22	336400.0	330235.000000
1	2021-05-23	333684.0	331959.500000
2	2021-05-24	170194.0	251076.750000
3	2021-05-25	107545.0	179310.875000
4	2021-05-26	154889.0	167099.937500
5	2021-05-27	188350.0	177724.968750
6	2021-05-28	205413.0	191568.984375



Predictions for #vaccines for California using EWMA(n=0.8):

	Date	Actual Value	Predicted Value
0	2021-05-22	336400.0	333934.0000
1	2021-05-23	333684.0	333734.0000
2	2021-05-24	170194.0	202902.0000
3	2021-05-25	107545.0	126616.4000
4	2021-05-26	154889.0	149234.4800
5	2021-05-27	188350.0	180526.8960
6	2021-05-28	205413.0	200435.7792



Task E:

Result of T 2-sample Paired testing for mean of no:of vaccines administered in CA/Al:

Null hypothesis (H0): Mean of Sept'21 vaccines administered in CA/Al = Mean of Nov'21 vaccines administered.

Alternate hypothesis(H1): Mean of Sept'21 vaccines administered in CA/Al is different from Mean of Nov'21 vaccines administered.

Procedure :

We have taken the alpha = 0.05 n=30 as given in Task and calculated the numerator and denominator of T value in the above t_test_paired function .

Result:

As the T statistic value for vaccines administered in CA/Al around September'21 = 13.184 which is greater than 2.45 we are rejecting the NULL hypothesis.

As the T statistic value for vaccines administered in CA/Al around Nov'21 = 7.68 which is greater than 2.45 we are rejecting the NULL hypothesis.

Is the paired T Test applicable ?

Paired T-Test Assumes that the data X and Y are dependent and diff of(X,Y) data set is normally distributed and as these two assumptions are failing Paired T test is not applicable.

Exploratory Tasks

Task-1

Use your X dataset to check if COVID19 cases/vaccinations had an impact on the X data. State your hypothesis clearly and determine the best tool (from among those learned in class) to apply to your hypotheses. Also, check whether the tool/test is applicable or not.

Here we have the Johnson & Johnson stock market data set as our X dataset to check if the Covid 19 case is affecting the X data.Johnson & Johnson is **the world's largest health care company**. It is also the highest paid drug company in the world. J&J remains at the top of the Big Pharma list of powerful corporations with more than $82 billion in annual revenue .
The recent health crisis has impacted almost all financial markets worldwide, in particular, stock and share prices trend dropped continuously and significantly. We want to see the correlation between the number of Covid-19 cases and vaccination count reported and Johnson & Johnson's stock price affected by this pandemic.

Here, we are loading from 3 datasets namely cases.csv,vaccination.csv and stock market data.csv. cases.csv file contains the data related to the covid cases, similarly vaccination.csv contains data related to the vaccinations. Both cases.csv and vaccinations.csv are the given datasets for this project. Whereas stock market data.csv file is selected as a part of the inference.

Load Data

```
[ ] total_cases=pd.read_csv("/content/drive/MyDrive/Prob Stats Project Data/cases.csv")#cases dataset
    total_vaccinations=pd.read_csv("/content/drive/MyDrive/Prob Stats Project Data/vaccinations.csv")#vaccination dataset
    data=pd.read_csv("/content/drive/MyDrive/Prob Stats Project Data/stock market data.csv")#x dataset
```

The location column from the vaccination dataset is being renamed to state and the submission_date from cases dataset is renamed to date.

```
total_vaccinations.rename(columns={"Location":"state"},inplace=True)#renaming the column from location to state in vaccinations file
total_cases.rename(columns={"submission_date":"Date"},inplace=True)#renaming the columns from submission_date to date in cases file
```

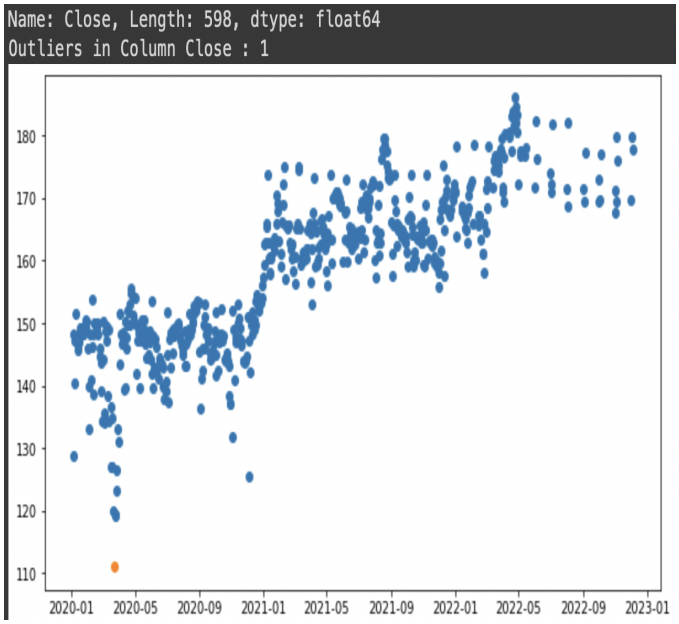
We are filtering the data based on the Alaska and California states and using some columns from both cases and vaccination dataset.

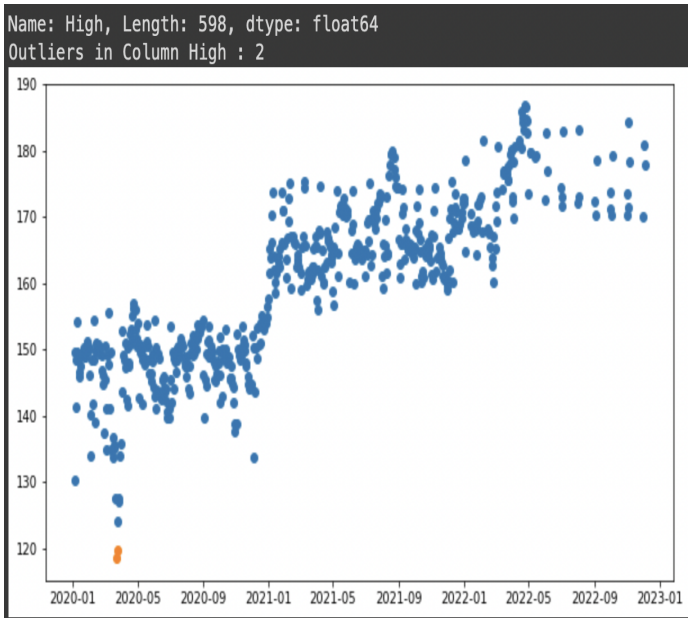
DATA PREPROCESSING:

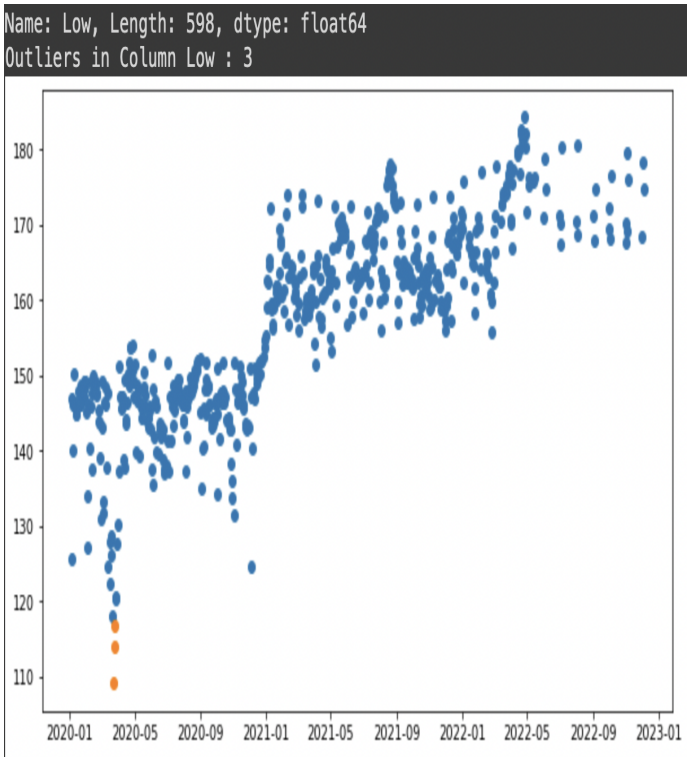
- Removing the null values
- Formatting the date

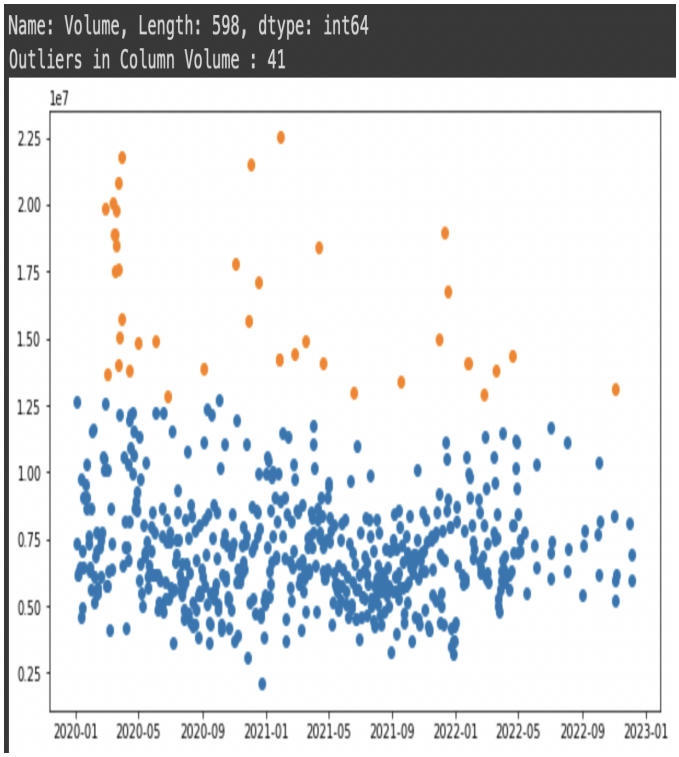
```
def convert_date(data):#converting the date format
  data['Date'] =  pd.to_datetime(data['Date'], format=r"%m/%d/%Y")
  return data
```

- Performing sorting for assigned states
- Outliers can be detected using Tukey's rule as performed earlier.









Since we have no outliers in this dataset, there is no need to remove the value.

- Performing merge operation on date column when state is Alaska and California

Alaska:

	Date	Open	High	Low	Close	Volume	state	new_case	new_death	tot_cases	month
0	2020-02-01	145.87	146.02	145.08	145.97	5634469.0	AL	0.0	0.0	0	2
1	2020-02-03	134.78	140.13	134.01	140.02	11508230.0	AL	0.0	0.0	0	2
2	2020-02-04	129.12	134.00	127.22	133.15	11594460.0	AL	0.0	0.0	0	2
3	2020-02-06	147.06	148.35	146.12	148.25	6078311.0	AL	0.0	0.0	0	2
4	2020-02-07	141.25	141.84	140.33	140.97	5152238.0	AL	0.0	0.0	0	2

California:

	Date	Open	High	Low	Close	Volume	state	new_case	new_death	tot_cases	month
0	2020-02-01	145.87	146.02	145.08	145.97	5634469.0	CA	0.0	0.0	0	2
1	2020-02-03	134.78	140.13	134.01	140.02	11508230.0	CA	3.0	0.0	6	2
2	2020-02-04	129.12	134.00	127.22	133.15	11594460.0	CA	0.0	0.0	6	2
3	2020-02-06	147.06	148.35	146.12	148.25	6078311.0	CA	0.0	0.0	6	2
4	2020-02-07	141.25	141.84	140.33	140.97	5152238.0	CA	0.0	0.0	6	2

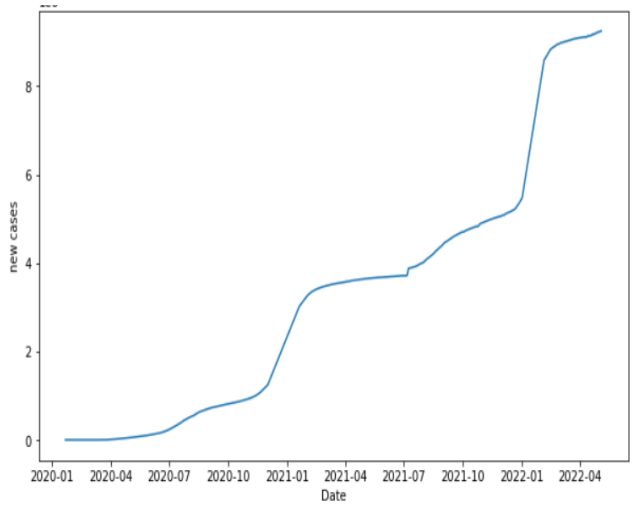
Pearson Correlation Test :

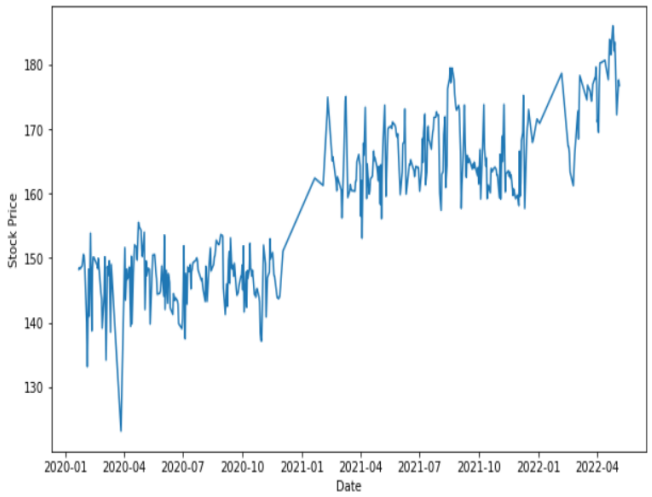
1.Result of Pearson Test for California State

We got the correlation value as 0.8626. From the result we can say that they are positively correlated.

Result: As the coefficient value for California is greater than 0.5,they are positively correlated.

California:- 0.8626

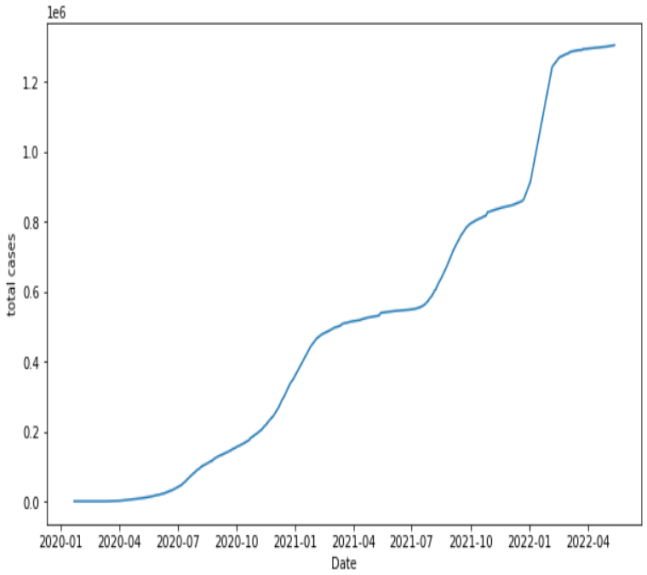


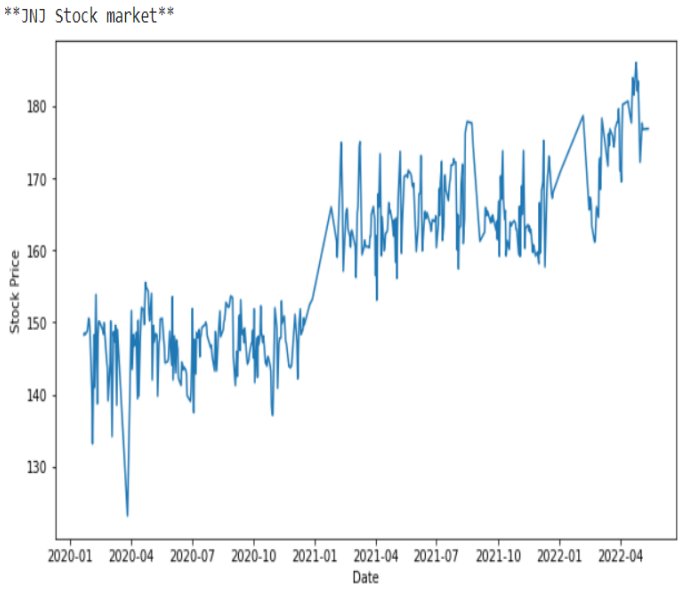


2.Result of Pearson Test for Alaska State

Result: We got the correlation value as 0853. As the coefficient value for California is greater than 0.5,they are positively correlated.

Alaska: 0.8537





By looking at the above plots, we can see that they are linearly correlated.

K-S Test :

1.Result of K-S Test for California State

Null hypothesis (H0): Distribution of Covid cases in California same as Distribution of Stock price for J&J in the timeframe 2020-2022

Alternate hypothesis(H1): Distribution of Covid cases in California is different from Distribution of Stock price for J&J in the timeframe 2020-2022

Procedure : We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Is the KS Test applicable ? There as no assumptions under KS test, hence the test is applicable

Result:

Since the statistical value of the KS test is ***0.424***, greater than 0.05, we **reject** the null hypothesis.

2.Result of K-S Test for Alaska State

Null hypothesis (H0): Distribution of Covid cases in Alaska same as Distribution of Stock price for J&J in the timeframe 2020-2022

Alternate hypothesis(H1): Distribution of Covid cases in Alaska sme as Distribution of Stock price for J&J in the timeframe 2020-2022

Procedure : We take c = 0.05 as reported in the literature and calculate the maximum difference in the CDF of the distributions at all points.

Is the KS Test applicable ? There as no assumptions under KS test, hence the test is applicable

Result:

Since the statistical value of the KS test is ***0.701***, greater than 0.05, we **reject** the null hypothesis.

Permutation Test :

The permutation test is used to check whether two data samples follow the same distribution.

1.Perumtation test for Alaska state

Permutation test result when performed on total cases and close stock price for Alaska state.

Null hypothesis (H0): Distribution of Covid cases in Alaska same as Distribution of Stock price for J&J in the timeframe 2020-2022

Alternate hypothesis(H1): Distribution of Covid cases in Alaska sme as Distribution of Stock price for J&J in the timeframe 2020-2022

Procedure : We arranged all Alaska state records and California records data 5000 ways. We take threshold = 0.05, and calculate it.

Is the Permutation Test applicable ?

There are no assumptions under the permutation test, hence the test is applicable.

```
print("--------For Alaska--------")
permutation_test_function(total_cases_abs, close_abs)
```

```
--------For Alaska--------
The p-value is:  0.0
==> Reject the Null Hypothesis
```

Since the null value is less than the threshold value for Alaska the hypothesis is rejected.

2.Perumtation test for California state

Permutation test result when performed on total cases and close stock price for California state.

Null hypothesis (H0): Distribution of Covid cases in Alaska same as Distribution of Stock price for J&J in the timeframe 2020-2022

Alternate hypothesis(H1): Distribution of Covid cases in Alaska sme as Distribution of Stock price for J&J in the timeframe 2020-2022

Procedure : We arranged all Alaska state records and California records data 5000 ways. We take threshold = 0.05, and calculate it.

Is the Permutation Test applicable ?

There are no assumptions under the permutation test, hence the test is applicable.

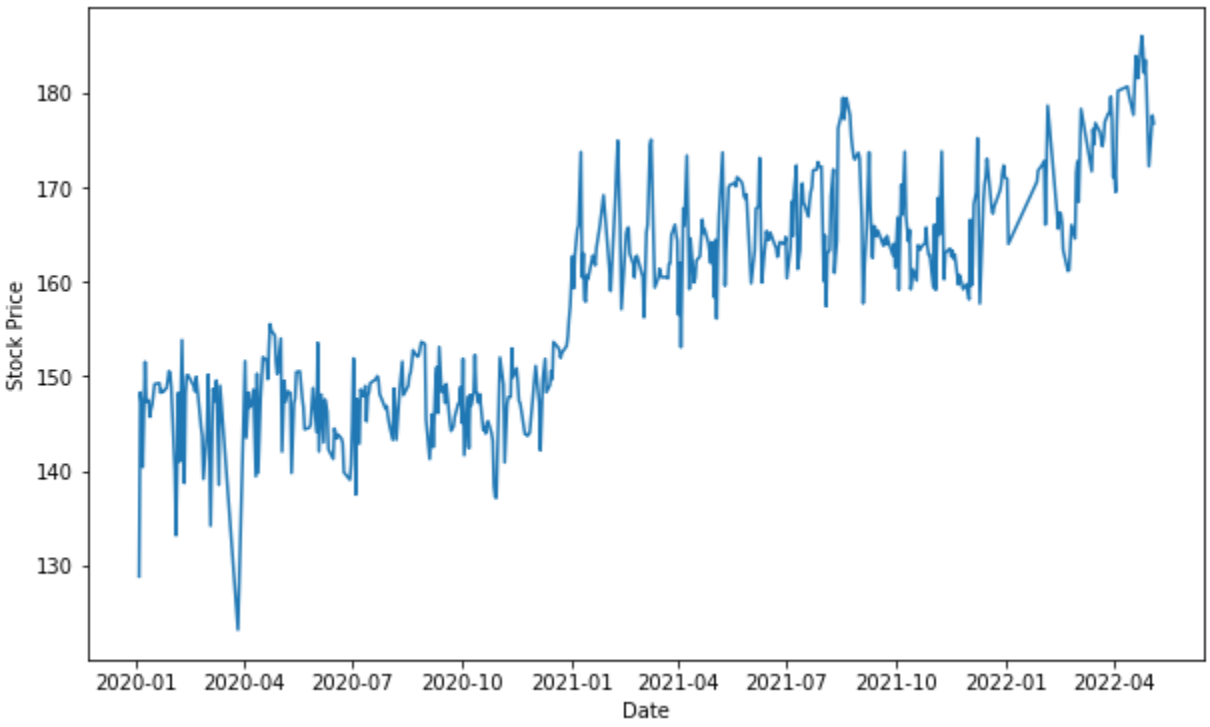
```
print("--------For Cali--------")
permutation_test_function(total_cases_cali_abs, close_abs)
```

```
--------For Cali--------
The p-value is:  0.0
==> Reject the Null Hypothesis
```

Since the null value is less than the threshold value for California the hypothesis is rejected.

Overall inference from the exploratory analysis :

Our null hypothesis was that the prices of the J&J stocks depends on the number of covid cases. This assumption was based on the thinking that the negative news of covid cases and death affects the stock market in a negative way.
But, based on the permutation, KS tests our null hypothesis was rejected. So the covid cases and closing prices of stocks do not follow a similar distribution.
Because of considering a few random permutations from a large number of possible permutations, this result also might be inaccurate, but we assume this to be true.
Upon further analysis , the reason for the rejection of the null hypothesis was understood. Although the bad news of covid and lockdown affected the prices of shares for a few days, the prices recovered gradually. We can also observe the increase of prices of stocks over the period of time.
This is because, covid was no longer an influencing factor in the stock market after its initial outburst. This is like, Elon Musk influencing the price of DogeCoin with a single tweet. Although it was effective initially, its price was constant even though Elon Musk tweeted after a period of time. So, we thought this is similar to that and the same kind of news may not be impactful over a long period of time, especially when covid was being referred to as the new normal.
In addition to that, when we were expecting the decrease in prices of stocks, there was an increase to it.
This is because, covid was no longer an influencing factor in the stock market after its initial outburst. This is like, Elon Musk influencing the price of DogeCoin with a single tweet. Although it was effective initially, its price was constant even though Elon Musk tweeted after a period of time. So, we thought this is similar to that and the same kind of news may not be impactful over a long period of time, especially when covid was being referred to as the new normal.
In addition to that, when we were expecting the decrease in prices of stocks, there was an increase to it.



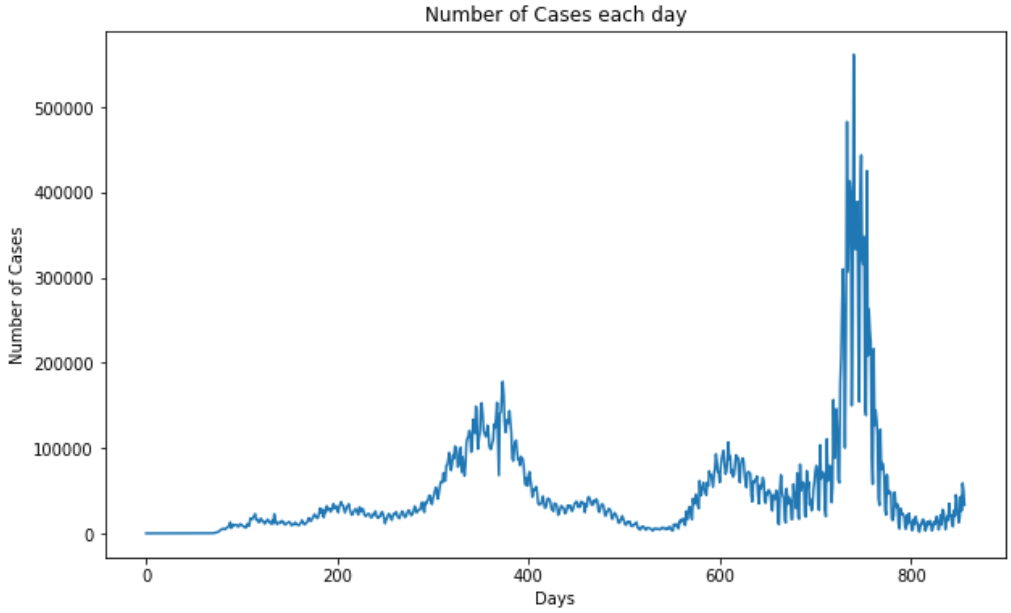
In the above plot, we can see that the price of J&J stock increased over time. A huge dip can be observed in March 2020, which might be because of the announcement of lockdown.
In the same way, a big jump in the price has also been observed in February 2021, that is when the J&J Vaccine got Emergency Use Authorization.
So, the relationship between the situations created by the pandemic and the stock price of J&J can be clearly observed.

As the number of covid cases also increased gradually, we observed a positive correlation value between the number of covid cases and the closing price of stocks.
But we know that, Correlation does not mean Causation. The increase in covid cases are not the reason for this, but the situations created by the pandemic.

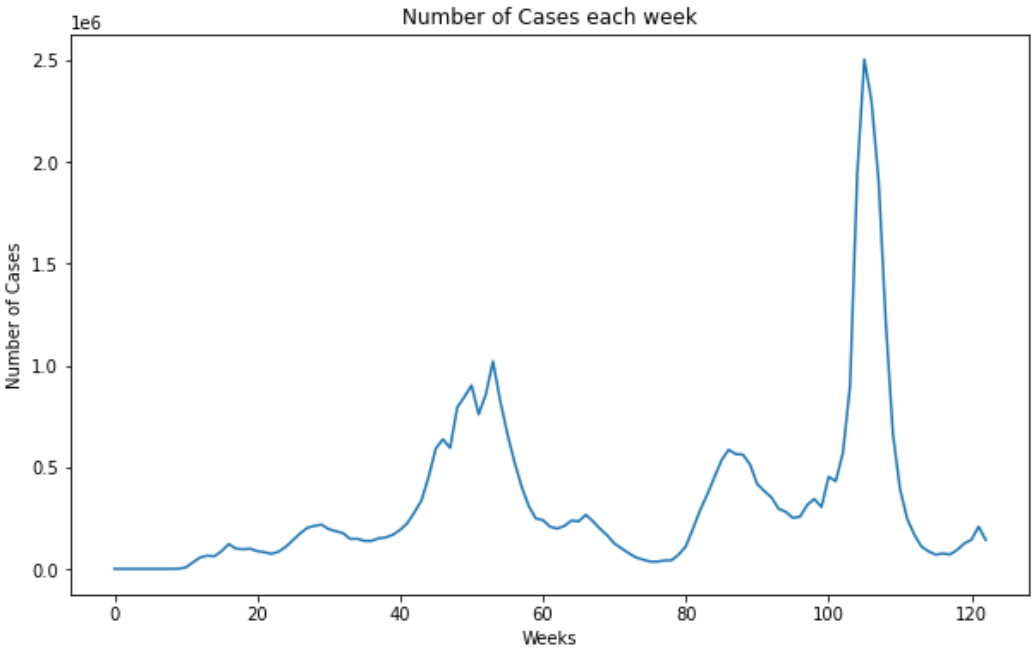
After the pandemic hit and the situation got worse day by day, people were eagerly waiting for the vaccine for COVID. J&J is one of the very few companies that came up with a vaccine. This boosted their business along with their stock value. That explains the increase of stock value of J&J over a period of time. These were our observations and the reason we understood for the observation to be as they were.

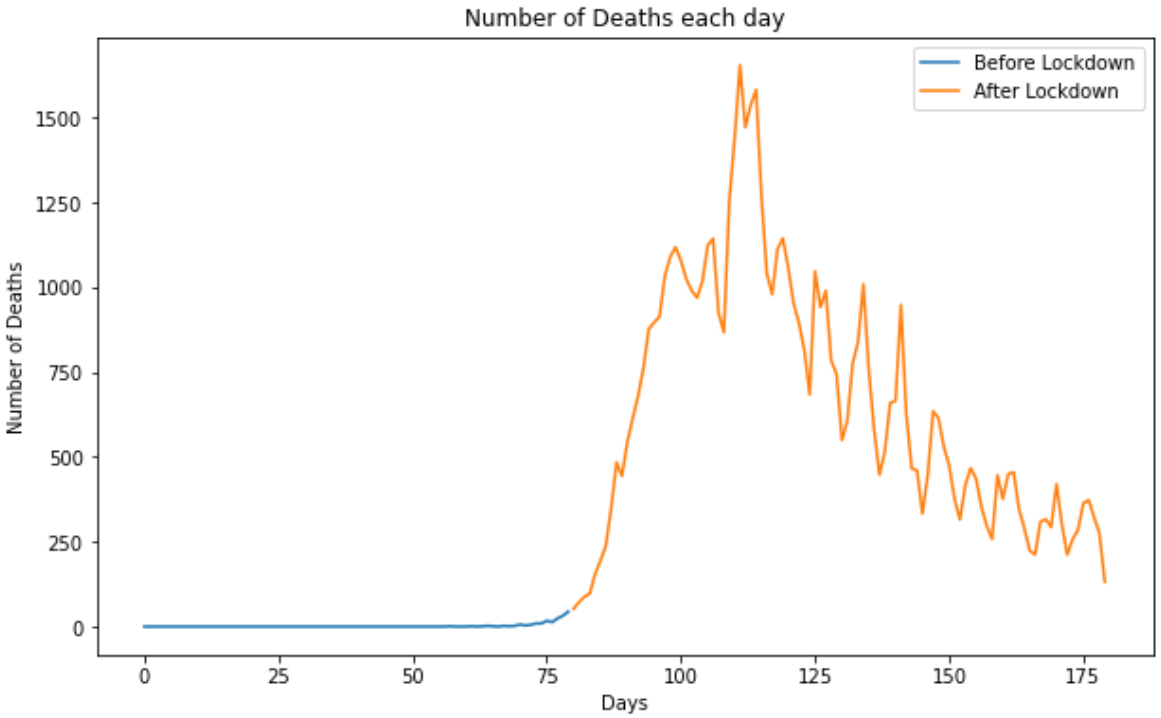
Exploratory Analysis Part 2 :

We analyzed the daily cases in the entire country and made some interesting observations from the daily cases and deaths patterns.
From the above plot, we can observe that two peaks which are clear and one part which seems like a peak but not that noticeable. These peaks indicate that the number of covid cases increased drastically for a period of time. We use the term “WAVE” to refer this situation. The number of covid waves differ in each country. In the United States, we see 2 (or 3) waves according to this data.

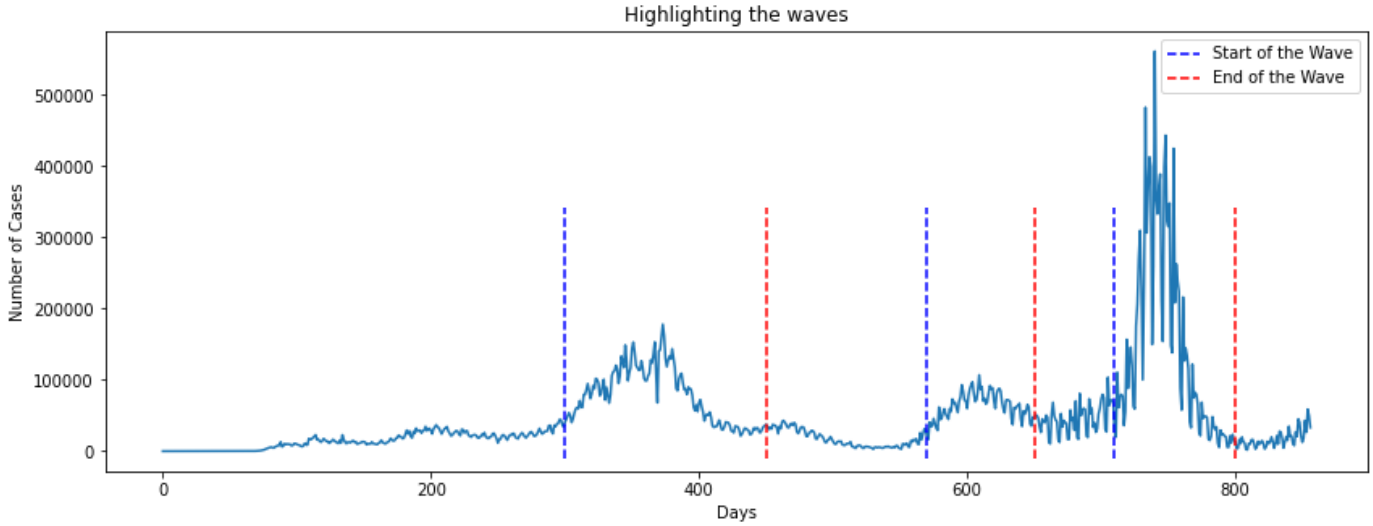


The weekly covid cases data also supports the above made statement.



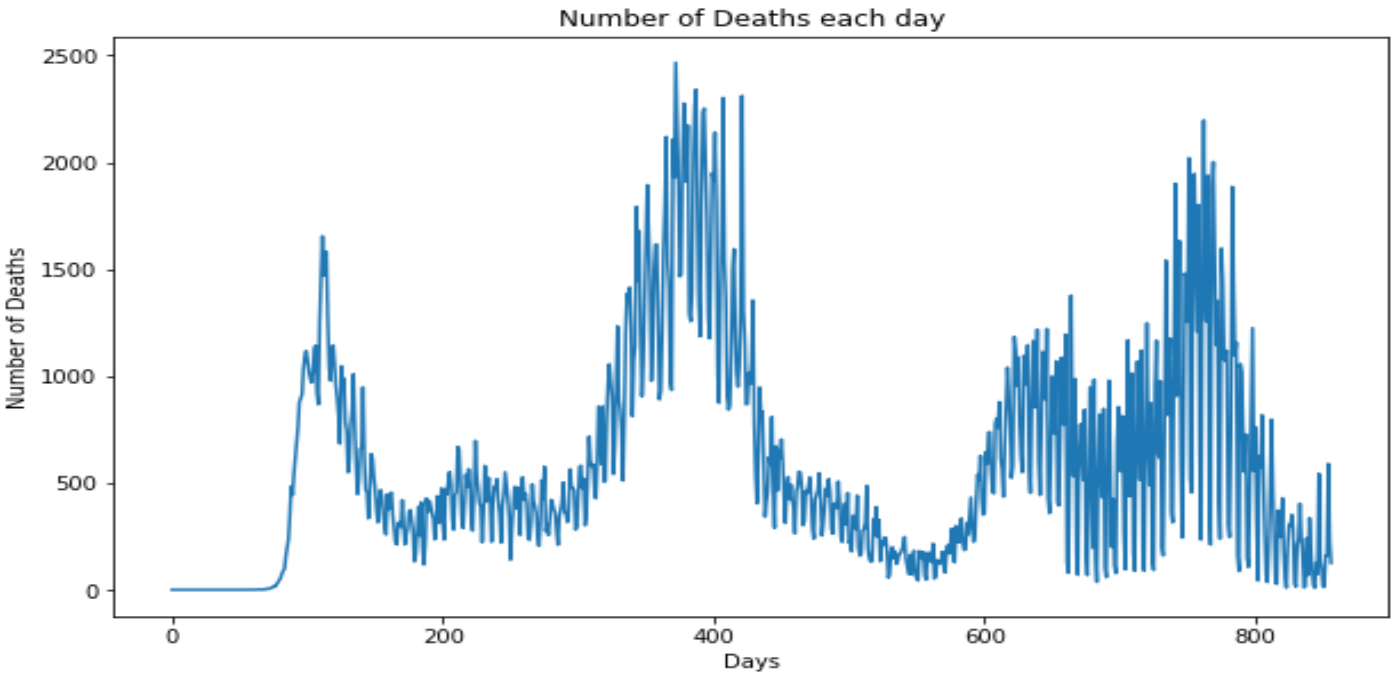


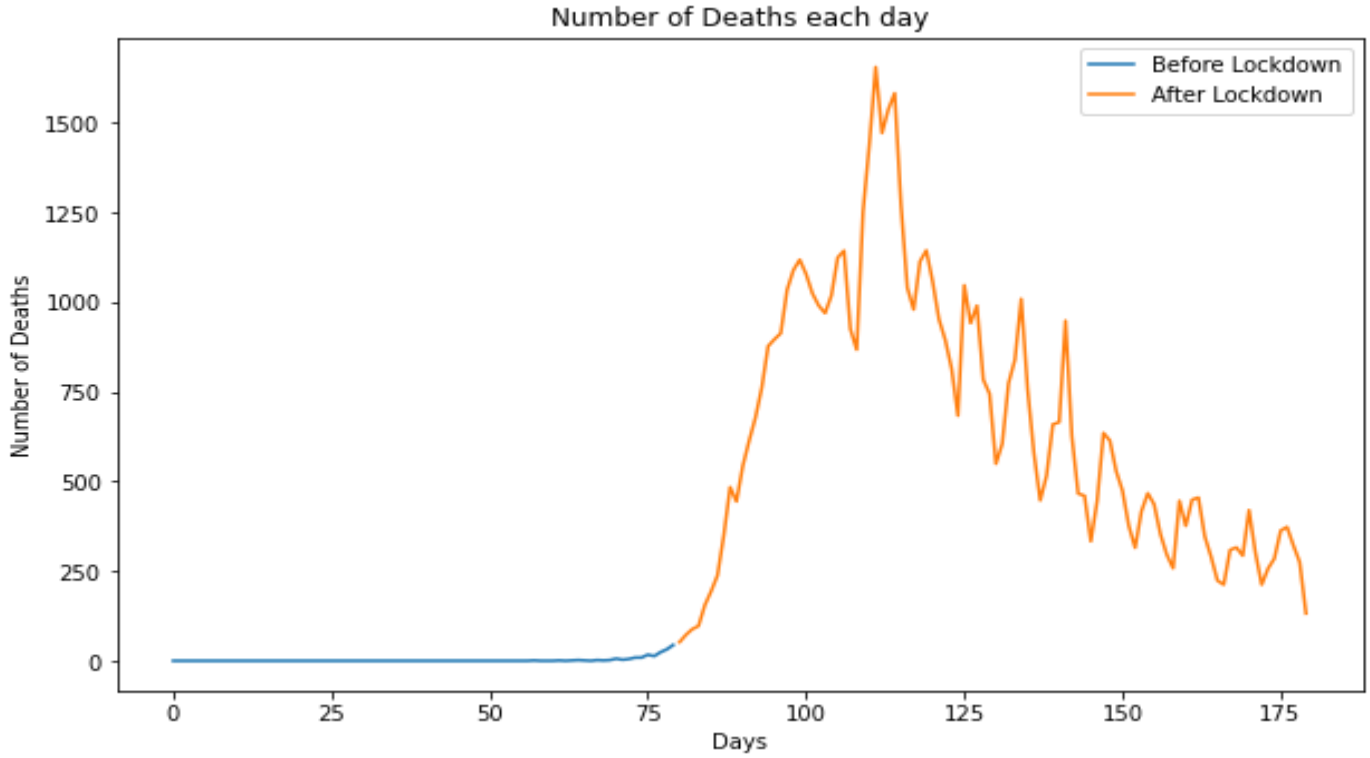
This plot shows the number of covid cases before and after the lockdown was announced. We observe that the number of cases continues to increase in spite of the lockdown. But after a period of time, the cases starts to decrease.After observing the peaks(waves) in the number of cases, all of them occurred around the same time when there were some relaxations announced in the lockdown. The rules had to be made strict everytime the number of cases again starts to increase. So, this indicates that if the lockdown was not announced, the situation would be unimaginable today.



The above plot shows the covid data highlighting the Waves in between the blue and red lines.
The green line indicates the start of the wave and the red line indicates the end of the wave.
The peaks can be clearly seen in between these lines.
The most recent wave we observed was because of the omicron variant.

Deaths





The above plots are of the daily deaths from the given data. We can observe three peaks in the same dates where we observed the peaks in covid cases.
The second and third peak are close to each other. And the same observation goes for the daily deaths before and after lockdown.

The Deadly Delta Variant

In the covid peaks, the world has experienced the peak because of the delta variant was more dangerous. This could be observed and proved using our data.
If we observe the second peak in the cases and deaths graphs, that was the time the delta variant flourished.
But, if we observe the plot of daily deaths and daily cases on each day, even though the peak during the delta variant time is not that huge in the cases plot, it is relatively high in the death plot. The same was announced by WHO about the delta variant.