CSE 544: Probability & Statistics for Data Science, Spring 2021

Final Project Report

Submitted by:

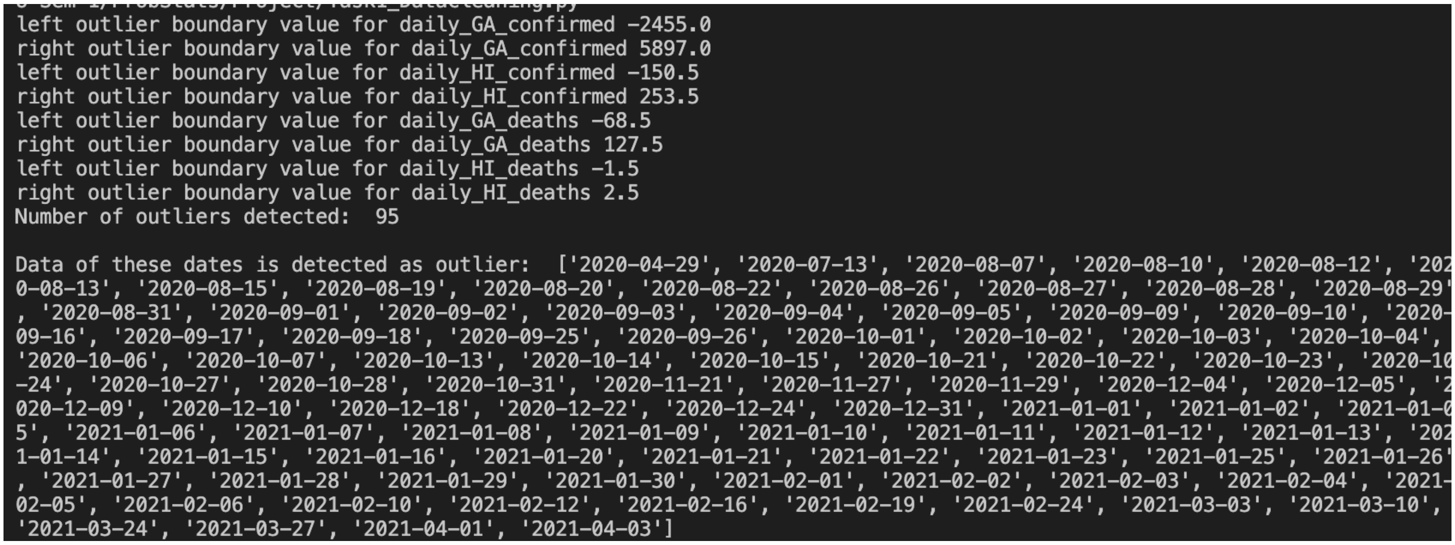
Harmanpreet Singh Khurana, SBU ID: 113262379,
Mayank Jain, SBU ID: 113263864,
Rajadorai DS, SBU ID: 113259773,
Venkatesan Ravi, SBU ID: 113263484

Mandatory Task

Task 1 - Data Cleaning

(Python file used for this: Task1_DataCleaning.py):

- For data cleaning, first we have calculated the daily stats for each column from the original cumulative data and formed a new data file.
- Then on the new daily data we have run the Outlier Detection and detected the outliers using Tukey's rule. We have saved all the dates for which data was an outlier in a list. Below output is shown after running the code.



```
o Sem 1/ProbStats/Project/Task1_DataCleaning.py
left outlier boundary value for daily_GA_confirmed -2455.0
right outlier boundary value for daily_GA_confirmed 5897.0
left outlier boundary value for daily_HI_confirmed -150.5
right outlier boundary value for daily_HI_confirmed 253.5
left outlier boundary value for daily_GA_deaths -68.5
right outlier boundary value for daily_GA_deaths 127.5
left outlier boundary value for daily_HI_deaths -1.5
right outlier boundary value for daily_HI_deaths 2.5
Number of outliers detected:  95

Data of these dates is detected as outlier:  ['2020-04-29', '2020-07-13', '2020-08-07', '2020-08-10', '2020-08-12', '202
0-08-13', '2020-08-15', '2020-08-19', '2020-08-20', '2020-08-22', '2020-08-26', '2020-08-27', '2020-08-28', '2020-08-29'
, '2020-08-31', '2020-09-01', '2020-09-02', '2020-09-03', '2020-09-04', '2020-09-05', '2020-09-09', '2020-09-10', '2020-
09-16', '2020-09-17', '2020-09-18', '2020-09-25', '2020-09-26', '2020-10-01', '2020-10-02', '2020-10-03', '2020-10-04',
'2020-10-06', '2020-10-07', '2020-10-13', '2020-10-14', '2020-10-15', '2020-10-21', '2020-10-22', '2020-10-23', '2020-10
-24', '2020-10-27', '2020-10-28', '2020-10-31', '2020-11-21', '2020-11-27', '2020-11-29', '2020-12-04', '2020-12-05', '2
020-12-09', '2020-12-10', '2020-12-18', '2020-12-22', '2020-12-24', '2020-12-31', '2021-01-01', '2021-01-02', '2021-01-0
5', '2021-01-06', '2021-01-07', '2021-01-08', '2021-01-09', '2021-01-10', '2021-01-11', '2021-01-12', '2021-01-13', '202
1-01-14', '2021-01-15', '2021-01-16', '2021-01-20', '2021-01-21', '2021-01-22', '2021-01-23', '2021-01-25', '2021-01-26'
, '2021-01-27', '2021-01-28', '2021-01-29', '2021-01-30', '2021-02-01', '2021-02-02', '2021-02-03', '2021-02-04', '2021-
02-05', '2021-02-06', '2021-02-10', '2021-02-12', '2021-02-16', '2021-02-19', '2021-02-24', '2021-03-03', '2021-03-10',
'2021-03-24', '2021-03-27', '2021-04-01', '2021-04-03']
```

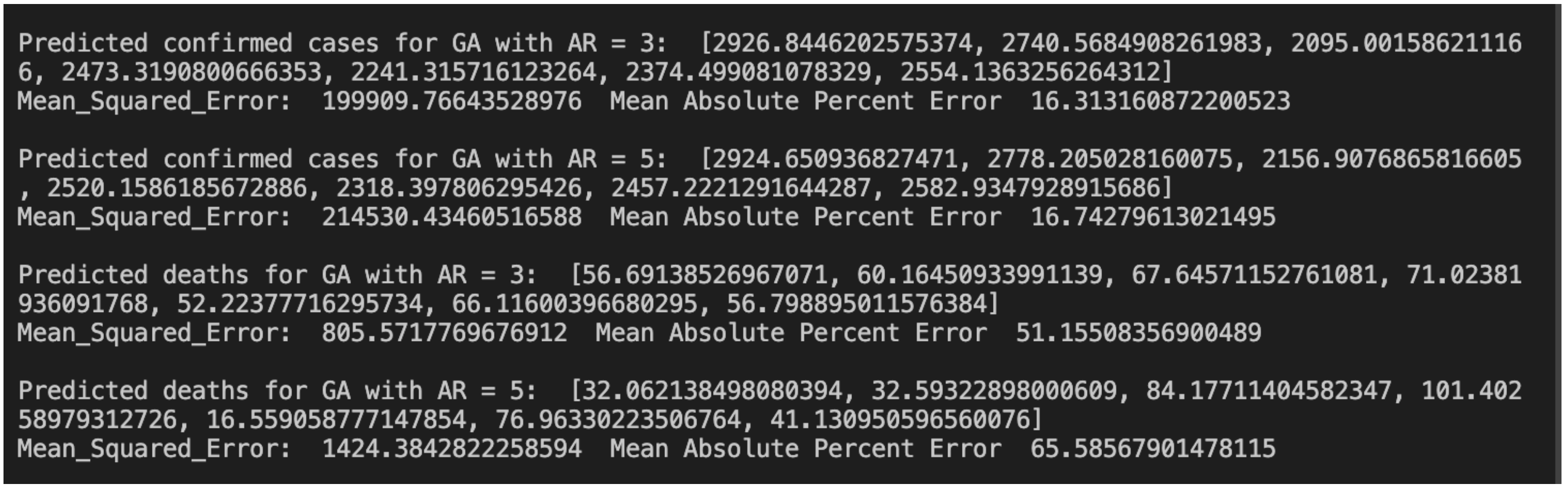
- Then out of the total **95 outliers** we have eliminated the outlier dates except for the dates which were in August, October, November, December, February and March as this data is going to be used in further questions.
- Apart from this no noisy or missing value was there in the data.

Task 2

Part 1: Auto-Regression and EWMA

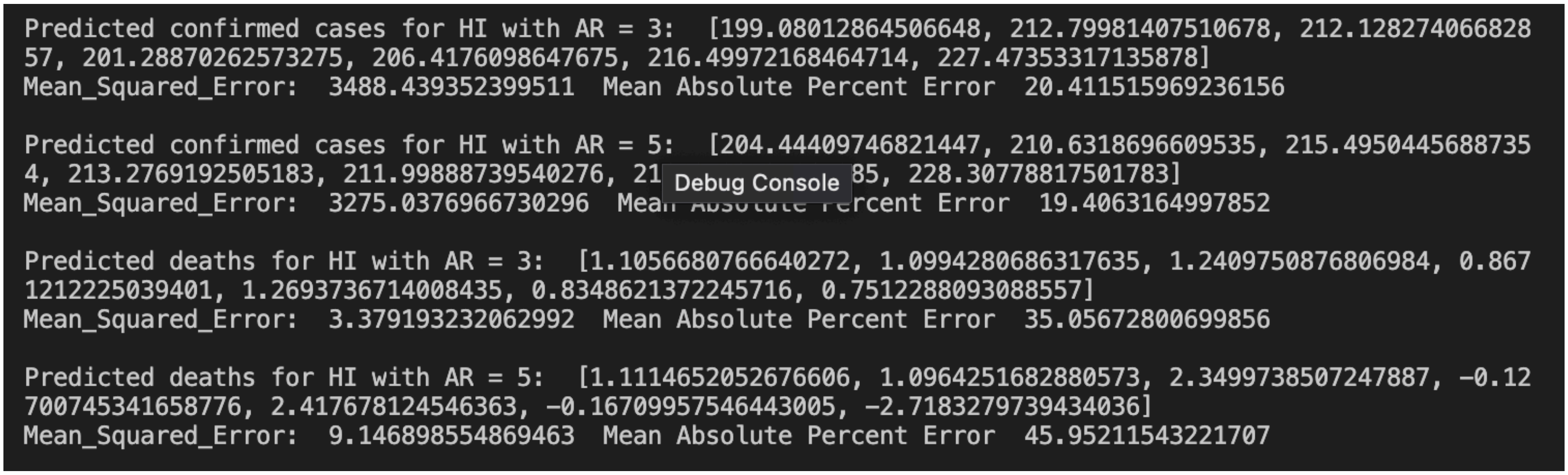
(Python file used for this: Task2_Step1_AR_EWMA.py):

a. Predicted confirmed cases, deaths and MSE,MAPE% for GA using AR = 3 and AR = 5:



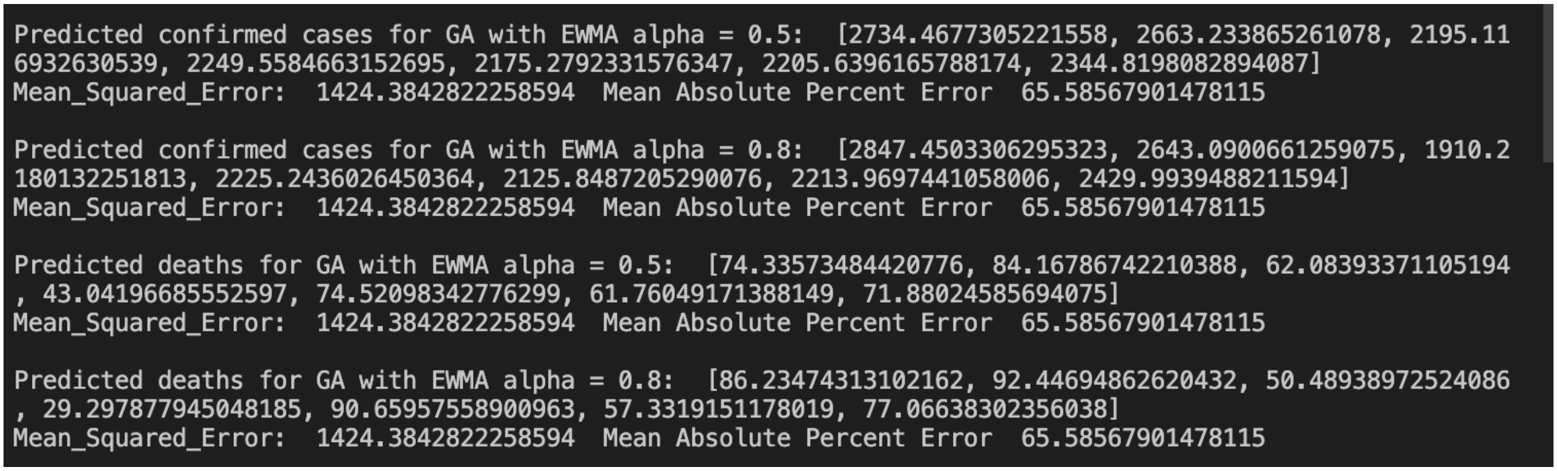
```
Predicted confirmed cases for GA with AR = 3:  [2926.8446202575374, 2740.5684908261983, 2095.00158621116
6, 2473.3190800666353, 2241.315716123264, 2374.499081078329, 2554.1363256264312]
Mean_Squared_Error:  199909.76643528976  Mean Absolute Percent Error  16.313160872200523

Predicted confirmed cases for GA with AR = 5:  [2924.650936827471, 2778.205028160075, 2156.9076865816605
, 2520.1586185672886, 2318.397806295426, 2457.2221291644287, 2582.9347928915686]
Mean_Squared_Error:  214530.43460516588  Mean Absolute Percent Error  16.74279613021495

Predicted deaths for GA with AR = 3:  [56.69138526967071, 60.16450933991139, 67.64571152761081, 71.02381
936091768, 52.22377716295734, 66.11600396680295, 56.798895011576384]
Mean_Squared_Error:  805.5717769676912  Mean Absolute Percent Error  51.15508356900489

Predicted deaths for GA with AR = 5:  [32.062138498080394, 32.59322898000609, 84.17711404582347, 101.402
58979312726, 16.559058777147854, 76.96330223506764, 41.130950596560076]
Mean_Squared_Error:  1424.3842822258594  Mean Absolute Percent Error  65.58567901478115
```

b. Predicted confirmed cases, deaths and MSE,MAPE% for HI using AR = 3 and AR = 5:



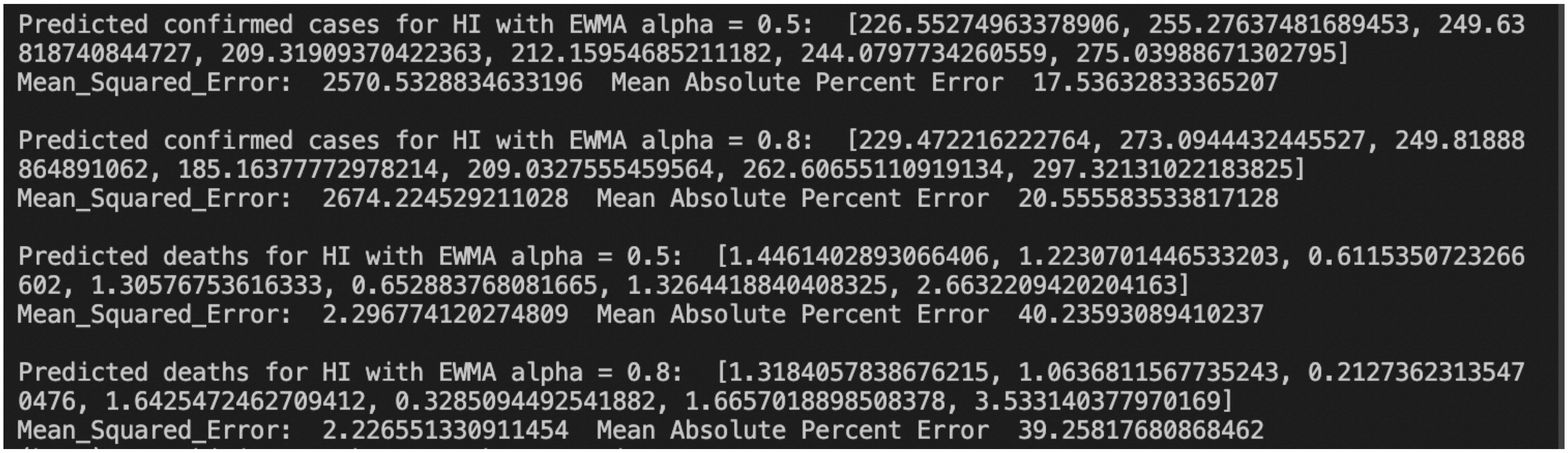
```
Predicted confirmed cases for HI with AR = 3:  [199.08012864506648, 212.79981407510678, 212.128274066828
57, 201.28870262573275, 206.4176098647675, 216.49972168464714, 227.47353317135878]
Mean_Squared_Error:  3488.439352399511  Mean Absolute Percent Error  20.411515969236156

Predicted confirmed cases for HI with AR = 5:  [204.44409746821447, 210.6318696609535, 215.495044568873
54, 213.2769192505183, 211.99888739540276, 21[illegible]85, 228.30778817501783]
Mean_Squared_Error:  3275.0376966730296  Mean Absolute Percent Error  19.4063164997852

Predicted deaths for HI with AR = 3:  [1.1056680766640272, 1.0994280686317635, 1.2409750876806984, 0.867
1212225039401, 1.2693736714008435, 0.8348621372245716, 0.7512288093088557]
Mean_Squared_Error:  3.379193232062992  Mean Absolute Percent Error  35.05672800699856

Predicted deaths for HI with AR = 5:  [1.1114652052676606, 1.0964251682880573, 2.3499738507247887, -0.12
700745341658776, 2.417678124546363, -0.16709957546443005, -2.7183279739434036]
Mean_Squared_Error:  9.146898554869463  Mean Absolute Percent Error  45.95211543221707
```

c. Predicted confirmed cases, deaths and MSE,MAPE% for GA using EWMA with alpha = 0.5 and 0.8:



```
Predicted confirmed cases for GA with EWMA alpha = 0.5:  [2734.4677305221558, 2663.23386526078, 2195.11
6932630539, 2249.5584663152695, 2175.2792331576347, 2205.6396165788174, 2344.8198082894087]
Mean_Squared_Error:  1424.3842822258594  Mean Absolute Percent Error  65.58567901478115

Predicted confirmed cases for GA with EWMA alpha = 0.8:  [2847.4503306295323, 2643.0900661259075, 1910.2
180132251813, 2225.2436026450364, 2125.8487205290076, 2213.9697441058006, 2429.9939488211594]
Mean_Squared_Error:  1424.3842822258594  Mean Absolute Percent Error  65.58567901478115

Predicted deaths for GA with EWMA alpha = 0.5:  [74.33573484420776, 84.16786742210388, 62.08393371105194
, 43.04196685552597, 74.52098342776299, 61.76049171388149, 71.88024585694075]
Mean_Squared_Error:  1424.3842822258594  Mean Absolute Percent Error  65.58567901478115

Predicted deaths for GA with EWMA alpha = 0.8:  [86.23474313102162, 92.44694862620432, 50.48938972524086
, 29.297877945048185, 90.65957558900963, 57.3319151178019, 77.06638302356038]
Mean_Squared_Error:  1424.3842822258594  Mean Absolute Percent Error  65.58567901478115
```

d. Predicted confirmed cases, deaths and MSE,MAPE% for HI using EWMA with alpha = 0.5 and 0.8:



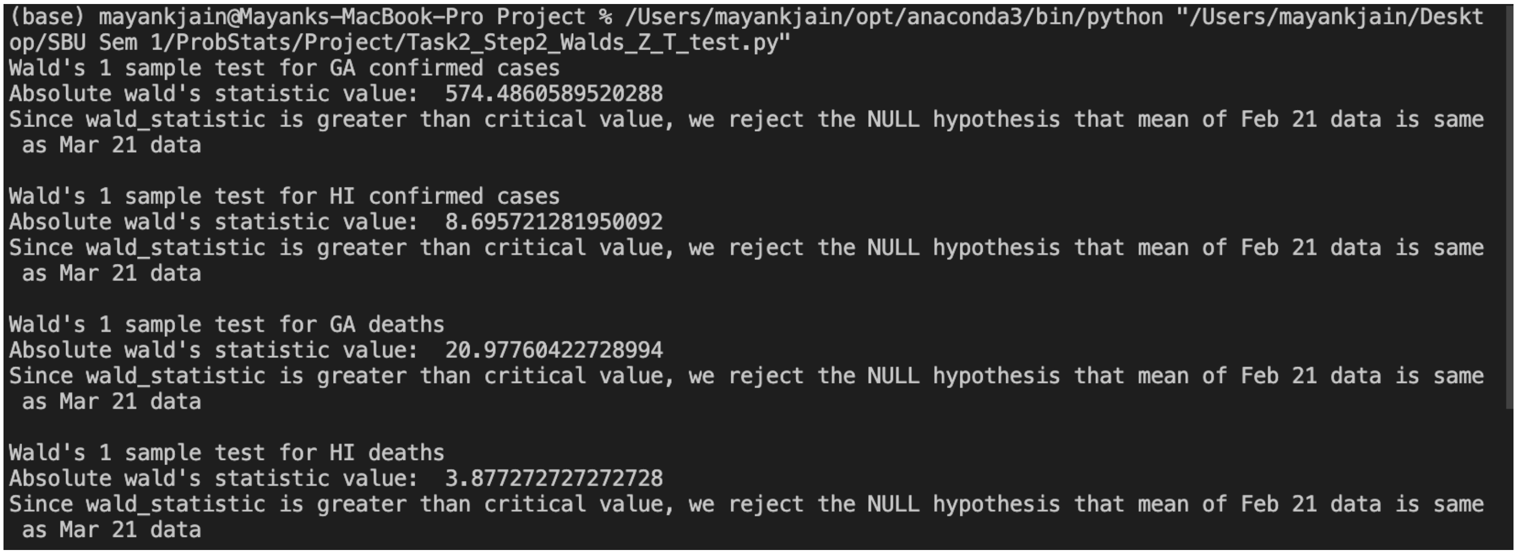
```
Predicted confirmed cases for HI with EWMA alpha = 0.5:  [226.55274963378906, 255.27637481689453, 249.63
818740844727, 209.31909370422363, 212.15954685211182, 244.0797734260559, 275.03988671302795]
Mean_Squared_Error:  2570.5328834633196  Mean Absolute Percent Error  17.53632833365207

Predicted confirmed cases for HI with EWMA alpha = 0.8:  [229.472216222764, 273.0944432445527, 249.81888
864891062, 185.16377772978214, 209.0327555459564, 262.60655110919134, 297.32131022183825]
Mean_Squared_Error:  2674.224529211028  Mean Absolute Percent Error  20.555583533817128

Predicted deaths for HI with EWMA alpha = 0.5:  [1.4461402893066406, 1.2230701446533203, 0.6115350723266
602, 1.30576753616333, 0.652883768081665, 1.3264418840408325, 2.6632209420204163]
Mean_Squared_Error:  2.296774120274809  Mean Absolute Percent Error  40.23593089410237

Predicted deaths for HI with EWMA alpha = 0.8:  [1.3184057838676215, 1.0636811567735243, 0.2127362313547
0476, 1.6425472462709412, 0.3285094492541882, 1.6657018898508378, 3.533140377970169]
Mean_Squared_Error:  2.226551330911454  Mean Absolute Percent Error  39.25817680868462
```

Task 2

Part 2: Wald’s Test, Z-Test and T-test

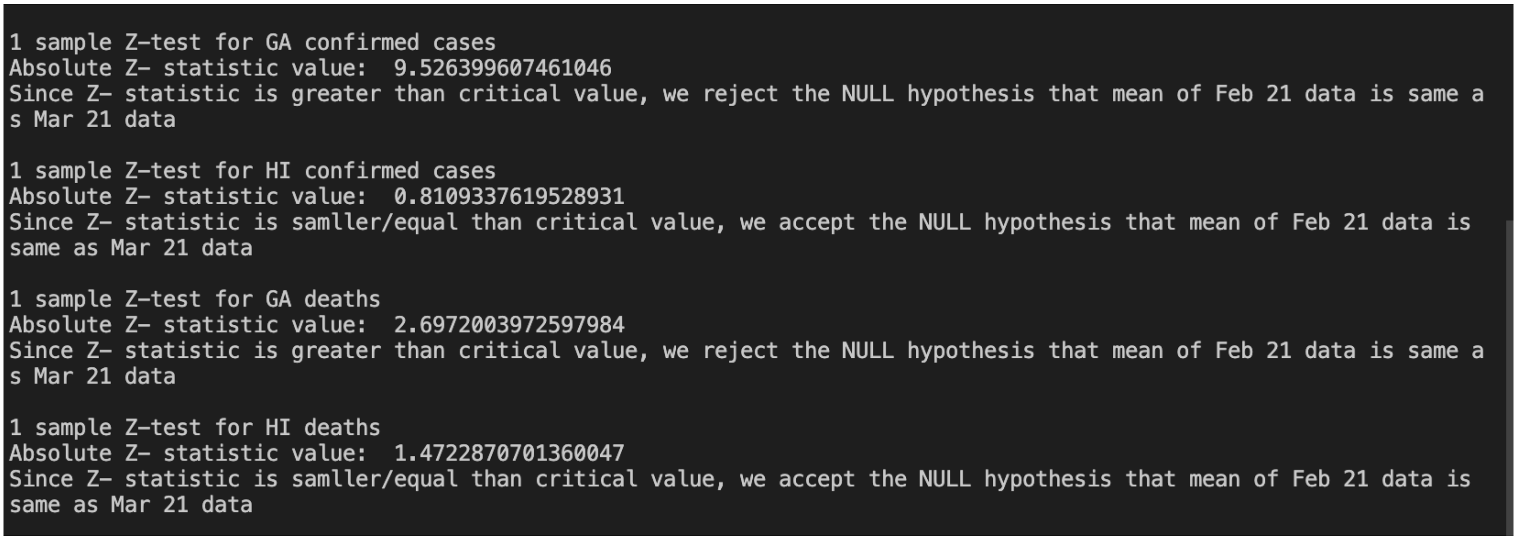
(Python file used for this: Task2_Step2_Walds_Z_T_test.py):

a. 1 sample Wald’s test results for GA and HI confirmed cases and deaths:



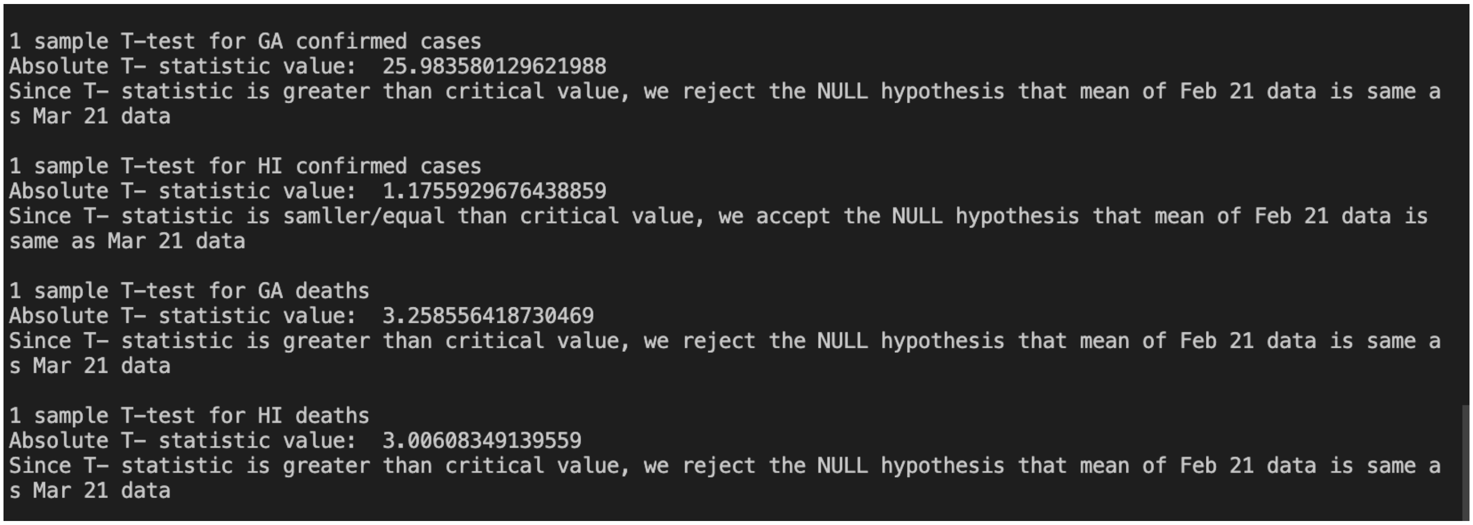
```
(base) mayankjain@Mayanks-MacBook-Pro Project % /Users/mayankjain/opt/anaconda3/bin/python "/Users/mayankjain/Deskt
op/SBU Sem 1/ProbStats/Project/Task2_Step2_Walds_Z_T_test.py"
Wald's 1 sample test for GA confirmed cases
Absolute wald's statistic value:  574.4860589520288
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data

Wald's 1 sample test for HI confirmed cases
Absolute wald's statistic value:  8.695721281950092
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data

Wald's 1 sample test for GA deaths
Absolute wald's statistic value:  20.97760422728994
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data

Wald's 1 sample test for HI deaths
Absolute wald's statistic value:  3.877272727272728
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data
```

b. 1 sample Z- test results for GA and HI confirmed cases and deaths:



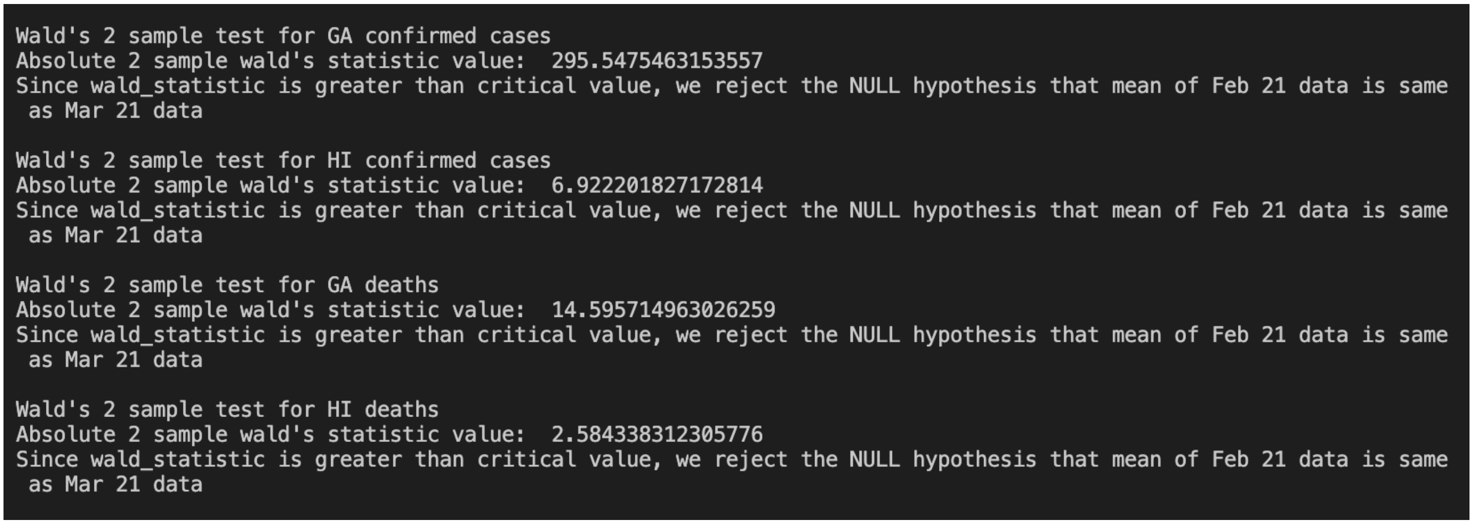
```
1 sample Z-test for GA confirmed cases
Absolute Z- statistic value:  9.526399607461046
Since Z- statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same a
s Mar 21 data

1 sample Z-test for HI confirmed cases
Absolute Z- statistic value:  0.8109337619528931
Since Z- statistic is samller/equal than critical value, we accept the NULL hypothesis that mean of Feb 21 data is
same as Mar 21 data

1 sample Z-test for GA deaths
Absolute Z- statistic value:  2.6972003972597984
Since Z- statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same a
s Mar 21 data

1 sample Z-test for HI deaths
Absolute Z- statistic value:  1.4722870701360047
Since Z- statistic is samller/equal than critical value, we accept the NULL hypothesis that mean of Feb 21 data is
same as Mar 21 data
```

c. 1 sample T- test results for GA and HI confirmed cases and deaths:



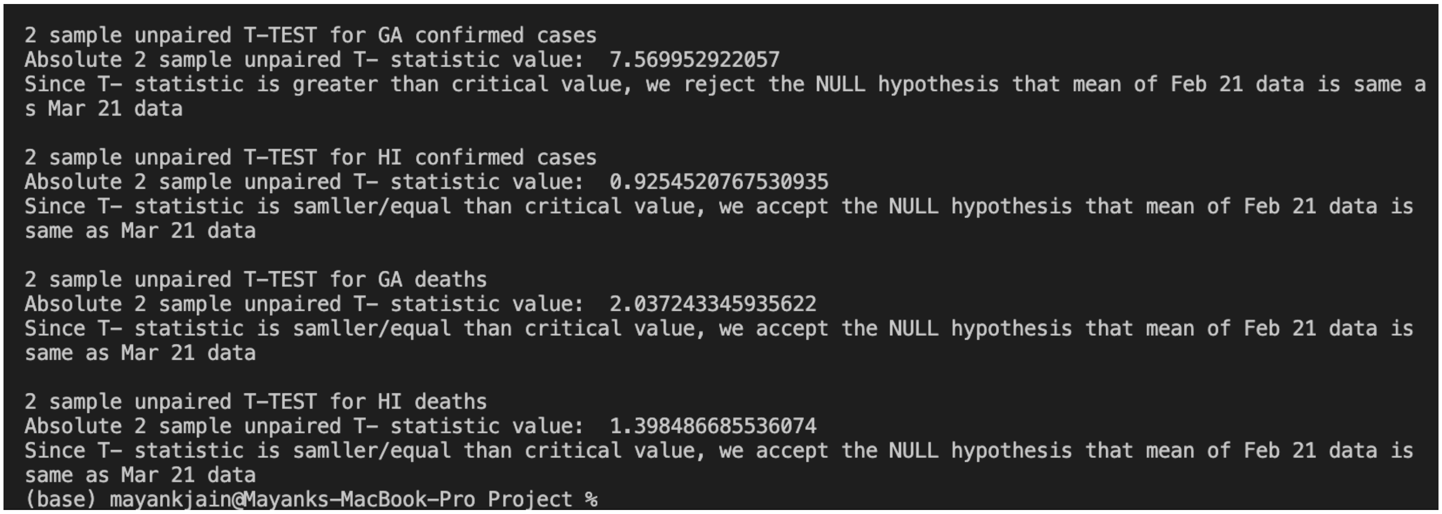
```
1 sample T-test for GA confirmed cases
Absolute T- statistic value:  25.983580129621988
Since T- statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same a
s Mar 21 data

1 sample T-test for HI confirmed cases
Absolute T- statistic value:  1.1755929676438859
Since T- statistic is samller/equal than critical value, we accept the NULL hypothesis that mean of Feb 21 data is
same as Mar 21 data

1 sample T-test for GA deaths
Absolute T- statistic value:  3.258556418730469
Since T- statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same a
s Mar 21 data

1 sample T-test for HI deaths
Absolute T- statistic value:  3.00608349139559
Since T- statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same a
s Mar 21 data
```

d. 2 sample Wald’s test results for GA and HI confirmed cases and deaths:



```
Wald's 2 sample test for GA confirmed cases
Absolute 2 sample wald's statistic value:  295.5475463153557
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data

Wald's 2 sample test for HI confirmed cases
Absolute 2 sample wald's statistic value:  6.922201827172814
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data

Wald's 2 sample test for GA deaths
Absolute 2 sample wald's statistic value:  14.595714963026259
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data

Wald's 2 sample test for HI deaths
Absolute 2 sample wald's statistic value:  2.584338312305776
Since wald_statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same
 as Mar 21 data
```

e. 2 sample unpaired T-test results for GA and HI confirmed cases and deaths:

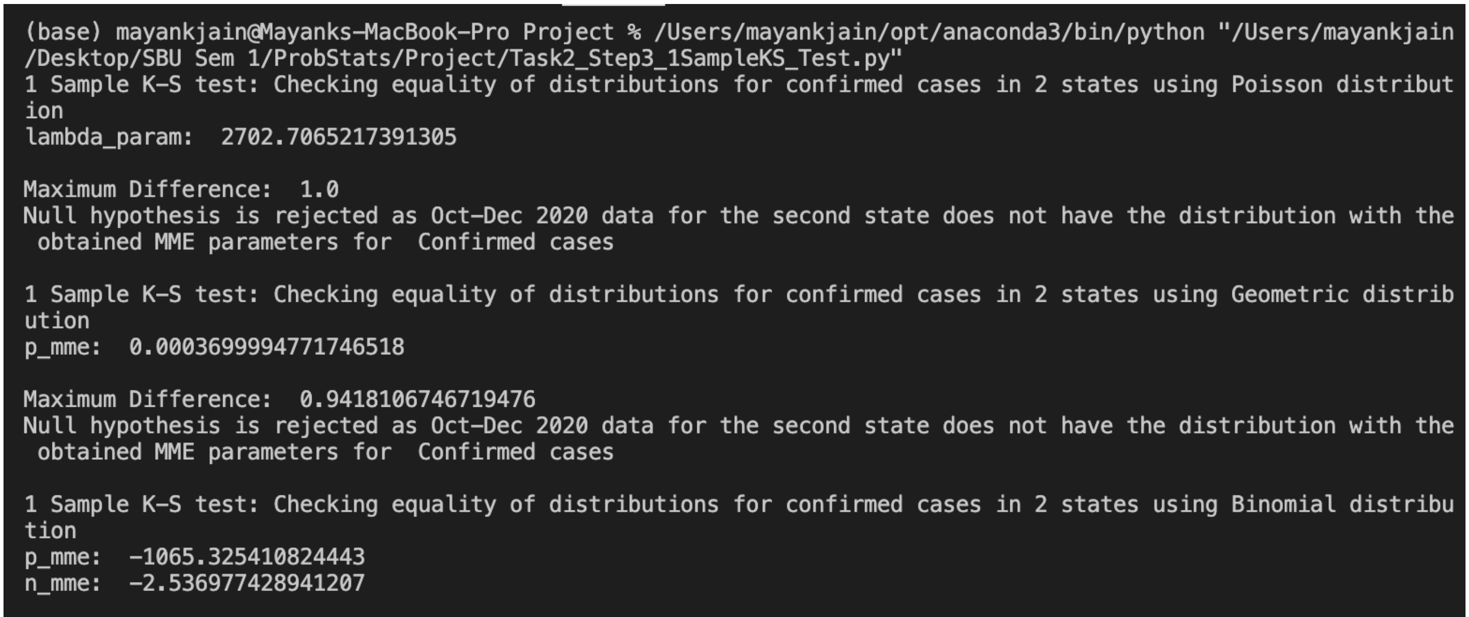


```
2 sample unpaired T-TEST for GA confirmed cases
Absolute 2 sample unpaired T- statistic value:  7.569952922057
Since T- statistic is greater than critical value, we reject the NULL hypothesis that mean of Feb 21 data is same a
s Mar 21 data

2 sample unpaired T-TEST for HI confirmed cases
Absolute 2 sample unpaired T- statistic value:  0.9254520767530935
Since T- statistic is samller/equal than critical value, we accept the NULL hypothesis that mean of Feb 21 data is
same as Mar 21 data

2 sample unpaired T-TEST for GA deaths
Absolute 2 sample unpaired T- statistic value:  2.037243345935622
Since T- statistic is samller/equal than critical value, we accept the NULL hypothesis that mean of Feb 21 data is
same as Mar 21 data

2 sample unpaired T-TEST for HI deaths
Absolute 2 sample unpaired T- statistic value:  1.398486685536074
Since T- statistic is samller/equal than critical value, we accept the NULL hypothesis that mean of Feb 21 data is
same as Mar 21 data
(base) mayankjain@Mayanks-MacBook-Pro Project %
```

Task 2

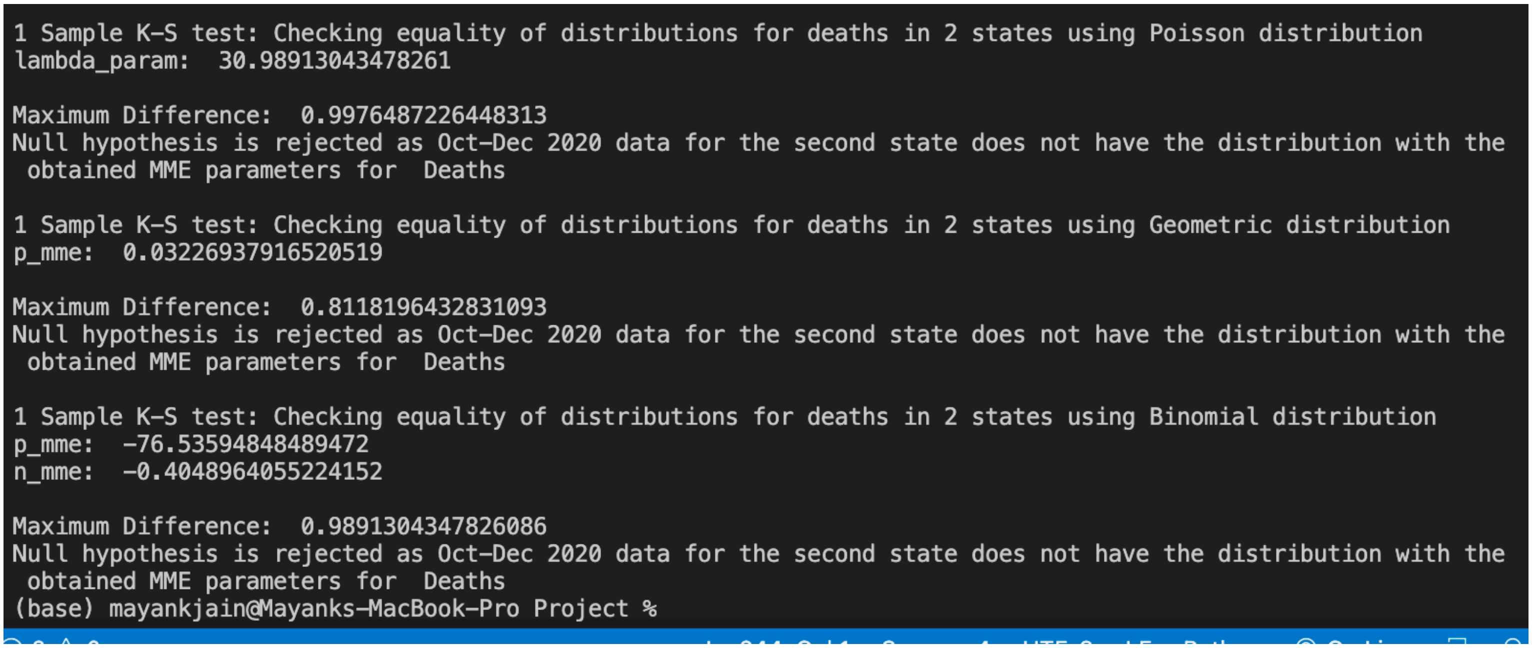
Part 3: One Sample K-S Test, 2 sample K-S test and Permutation test.
(Python files used for this: Task2_Step3_1SampleKS_Test.py, Task2_Step3_2SampleKS_test.py, Task2_Step3_Permutation_test.py):

a. 1 Sample K-S Test for confirmed cases in 2 states using Poisson, Geometric and Binomial distribution:



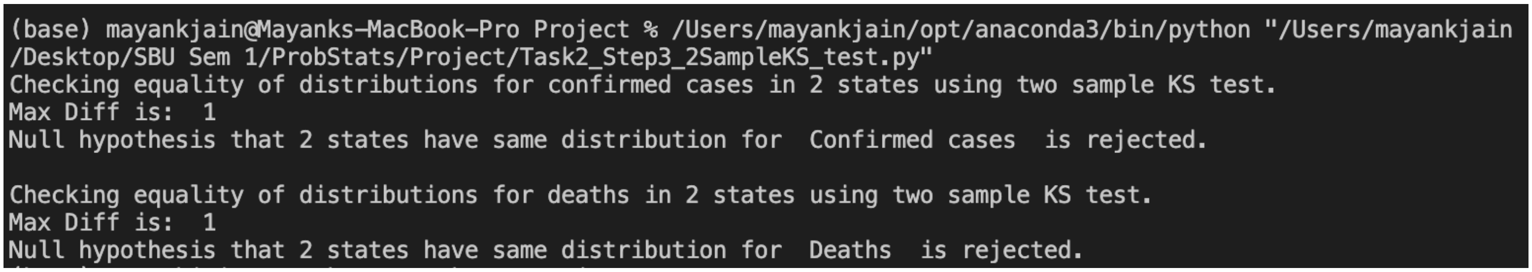
```
(base) mayankjain@Mayanks-MacBook-Pro Project % /Users/mayankjain/opt/anaconda3/bin/python "/Users/mayankjain
/Desktop/SBU Sem 1/ProbStats/Project/Task2_Step3_1SampleKS_Test.py"
1 Sample K-S test: Checking equality of distributions for confirmed cases in 2 states using Poisson distribut
ion
lambda_param:  2702.7065217391305

Maximum Difference:  1.0
Null hypothesis is rejected as Oct-Dec 2020 data for the second state does not have the distribution with the
 obtained MME parameters for  Confirmed cases

1 Sample K-S test: Checking equality of distributions for confirmed cases in 2 states using Geometric distrib
ution
p_mme:  0.0003699994771746518

Maximum Difference:  0.9418106746719476
Null hypothesis is rejected as Oct-Dec 2020 data for the second state does not have the distribution with the
 obtained MME parameters for  Confirmed cases

1 Sample K-S test: Checking equality of distributions for confirmed cases in 2 states using Binomial distribu
tion
p_mme:  -1065.325410824443
n_mme:  -2.536977428941207
```

b. 1 Sample K-S Test for deaths in 2 states using Poisson, Geometric and Binomial distribution:



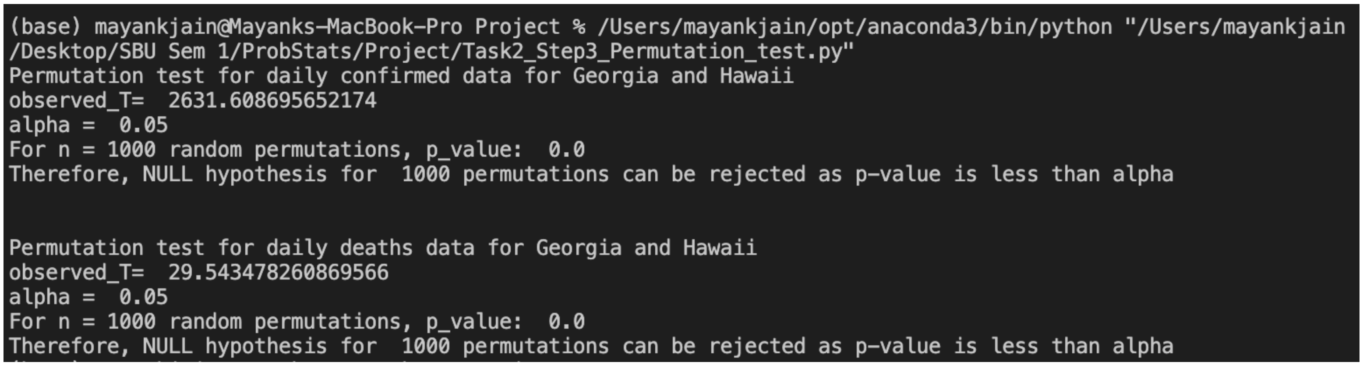
```
1 Sample K-S test: Checking equality of distributions for deaths in 2 states using Poisson distribution
lambda_param:  30.98913043478261

Maximum Difference:  0.9976487226448313
Null hypothesis is rejected as Oct-Dec 2020 data for the second state does not have the distribution with the
 obtained MME parameters for  Deaths

1 Sample K-S test: Checking equality of distributions for deaths in 2 states using Geometric distribution
p_mme:  0.03226937916520519

Maximum Difference:  0.8118196432831093
Null hypothesis is rejected as Oct-Dec 2020 data for the second state does not have the distribution with the
 obtained MME parameters for  Deaths

1 Sample K-S test: Checking equality of distributions for deaths in 2 states using Binomial distribution
p_mme:  -76.53594848489472
n_mme:  -0.4048964055224152

Maximum Difference:  0.9891304347826086
Null hypothesis is rejected as Oct-Dec 2020 data for the second state does not have the distribution with the
 obtained MME parameters for  Deaths
(base) mayankjain@Mayanks-MacBook-Pro Project %
```

c. 2 sample KS Test for confirmed cases and deaths in 2 states:



```
(base) mayankjain@Mayanks-MacBook-Pro Project % /Users/mayankjain/opt/anaconda3/bin/python "/Users/mayankjain
/Desktop/SBU Sem 1/ProbStats/Project/Task2_Step3_2SampleKS_test.py"
Checking equality of distributions for confirmed cases in 2 states using two sample KS test.
Max Diff is:  1
Null hypothesis that 2 states have same distribution for  Confirmed cases  is rejected.

Checking equality of distributions for deaths in 2 states using two sample KS test.
Max Diff is:  1
Null hypothesis that 2 states have same distribution for  Deaths  is rejected.
```

d. Permutation test for confirmed cases and death:



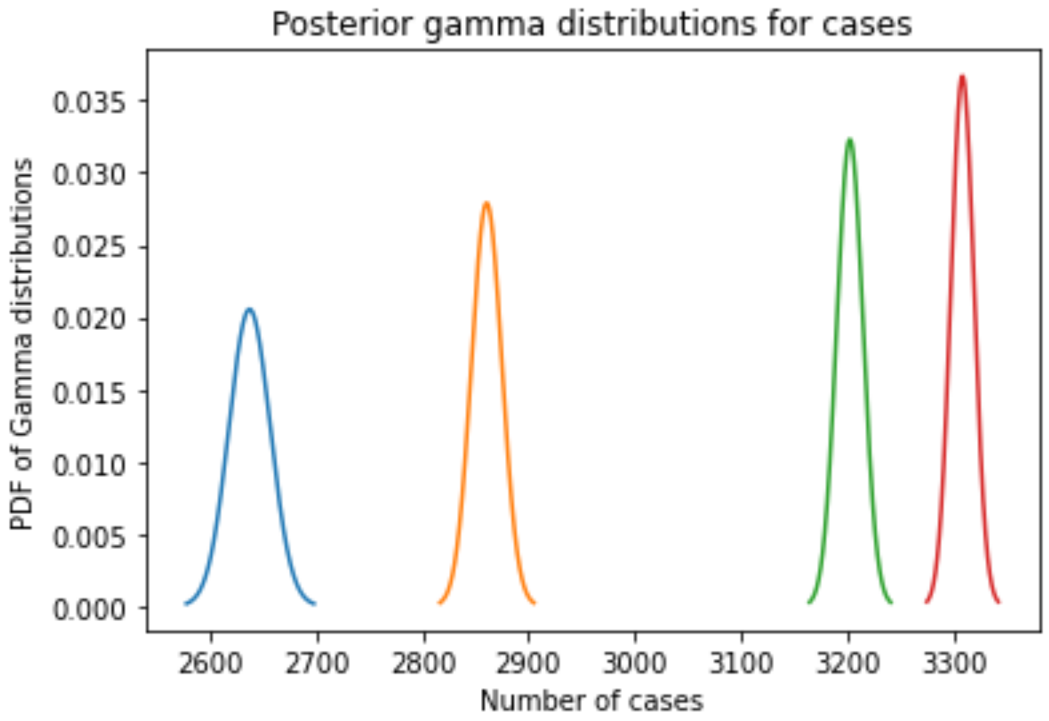
```
(base) mayankjain@Mayanks-MacBook-Pro Project % /Users/mayankjain/opt/anaconda3/bin/python "/Users/mayankjain
/Desktop/SBU Sem 1/ProbStats/Project/Task2_Step3_Permutation_test.py"
Permutation test for daily confirmed data for Georgia and Hawaii
observed_T=  2631.608695652174
alpha =  0.05
For n = 1000 random permutations, p_value:  0.0
Therefore, NULL hypothesis for  1000 permutations can be rejected as p-value is less than alpha


Permutation test for daily deaths data for Georgia and Hawaii
observed_T=  29.543478260869566
alpha =  0.05
For n = 1000 random permutations, p_value:  0.0
Therefore, NULL hypothesis for  1000 permutations can be rejected as p-value is less than alpha
```

Task 2

Part 4: Bayesian Inference

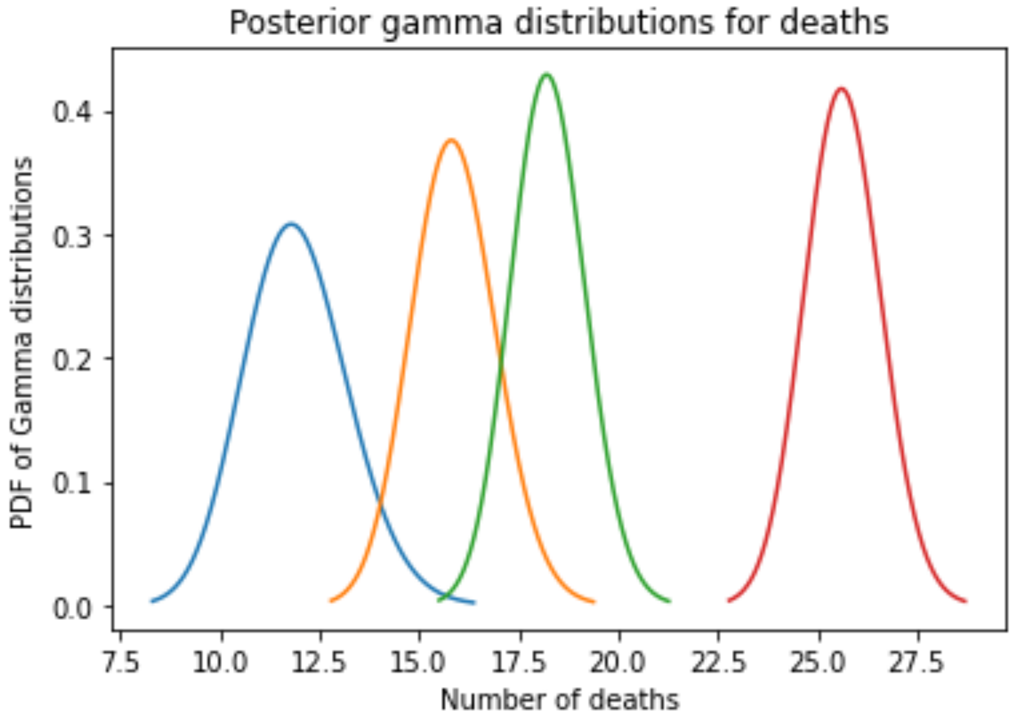
<u>Cases:</u>

```
Week : 1
alpha : 18462.0 MAP : 2636.949135763326
Week : 2
alpha : 40049.0 MAP : 2860.3491518957435
Week : 3
alpha : 67260.0 MAP : 3202.6616995534105
Week : 4
alpha : 92648.0 MAP : 3308.6822040973166
```



Deaths:

```
Week : 1
alpha : 84.0 MAP : 11.795617469873322
Week : 2
alpha : 223.0 MAP : 15.815427119459661
Week : 3
alpha : 384.0 MAP : 18.20677163839773
Week : 4
alpha : 719.0 MAP : 25.606578699932847
```



Exploratory Task

Hypothesis 1

Null Hypothesis (H0): Pollution caused by roadside traffic pollutants (here, Nitrous Oxide - NO) has a high correlation with COVID19 cases in Dallas, Texas

H1 (Rejecting the null hypothesis): Pollution caused by traffic pollutants (here, Nitrous Oxide - NO) does not have a high correlation with COVID19 cases in Dallas, Texas

We are testing the hypothesis using **Pearson's correlation technique** to determine whether the magnitude of pollutants produced (Nitrous Oxide) have been affected by the rise in Covid19 cases in 2020. We test the effect of Covid19 cases on roadside pollution with Nitrous Oxide as NO is the primary pollutant in the case of vehicular emissions.

The outbreak of COVID-19 has significantly inhibited global economic growth and impacted the environment. Some evidence suggests that lockdown strategies have significantly reduced traffic-related air pollution (TRAP) in regions across the world. In this hypothesis, we assessed the influence of the COVID-19 lockdown on the levels of traffic-related air pollutants in Dallas, Texas.

Code output:

```
Pearson Coefficient for Covid Cases in 2020 vs NO2 emission in 2020 in TX):
 0.481163941094839
```

Inference

After using the Pearson's test, we obtained the Pearson's correlation coefficient between Covid19 Cases and Roadside traffic air pollutants (Nitrous Oxide) = **0.48**. This shows that the effect of Covid19 cases (and hence, lockdown) has a **low-moderate correlation** with the magnitude of roadside pollutants such as Nitrous Oxide.

Hence, we can conclude by **rejecting the Null hypothesis** (pearson's coefficient value < 0.5).

Hypothesis 2

Null Hypothesis (H0): Pollution caused by industrial pollutants (here we compare with, Sulphur Dioxide - SO2) has a correlation with COVID19 cases

H1 (Rejecting the null hypothesis): Pollution caused by industrial pollutants (here, Sulphur Dioxide - SO2) has no correlation with COVID19 cases

We are testing the hypothesis using **Pearson's correlation technique** to determine whether the magnitude of pollutants produced through industrial emissions (here we use, Sulphur Dioxide - SO2) have been affected by the rise in Covid19 cases in 2020.

With large scale industries being primarily shut down during the time of rise in Covid19 cases, we were interested in evaluating the correlation between industrial pollutants (such as Sulphur Dioxide) and the COVID-19 outbreak in Dallas, Texas. We employed Pearsons's correlation test to analyze the association of SO2 with COVID-19 cases in the Greater Dallas area.

Code Output:

```
Pearson Coefficient for Covid Cases in 2020 vs SO2 emission in 2020 in TX):  0.2874
```

Inference

On using the Pearson's test, we obtained the Pearson's correlation coefficient between Covid19 Cases and industrial air pollutants (Sulphur dioxide) = **0.28**. We arrive at a surprising inference that the effect of Covid19 cases (and hence, lockdown) has **no correlation** with the magnitude of industrial pollutants such as Sulphur dioxide.

Hence, we can conclude by **rejecting the Null hypothesis** ((pearson's coefficient value = 0.28) < 0.5).

Hypothesis 3

Null Hypothesis (H0): Pollution caused by domestic pollutants (here we compare with, Carbon Monoxide - CO) has a correlation with COVID19 cases

H1 (Rejecting the null hypothesis): Pollution caused by domestic pollutants (here, Carbon Monoxide - CO) has no correlation with COVID19 cases

We are testing the hypothesis using **Pearson's correlation technique** to determine whether the magnitude of pollutants produced through domestic emissions (here we use, Carbon Monoxide - CO) have been affected by the rise in Covid19 cases in 2020.

With the '**Work from home**' mode going on a rise in 2020 during the time of rise in Covid19 cases, we were interested to evaluate the correlation between domestic pollution determinants (such as Carbon Monoxide) and the COVID-19 outbreak in Dallas, Texas. We employed Pearsons's correlation test to analyze the association of CO with COVID-19 cases in the Greater Dallas area.

Code Output:

```
Pearson Coefficient for Covid Cases in 2020 vs CO concentration in 2020 in TX):
 0.19322677026846036
```

Inference

On using the Pearson's test, we obtained the Pearson's correlation coefficient between Covid19 Cases and domestic emissions/pollutants (Carbon Monoxide) = **0.193**. We arrive at a surprising inference that the effect of Covid19 cases (and hence, lockdown) has **no correlation** with the magnitude of domestic pollutants such as Carbon Monoxide.

Hence, we can conclude by **rejecting the Null hypothesis** ((pearson's coefficient value = 0.193) < 0.5).

Conclusion and Key takeaway from our hypotheses

Considering that in all **three** of our hypotheses, the null hypothesis was rejected indicating that the roadside-traffic/industrial/domestic produced pollutant had no/significantly low correlation with the rise in Covid19 cases. This is an unexpected but alarming reveal. This shows that there is an urgent need to reduce the pollution levels across sectors (traffic related, domestic and industrial).

Overall, our study is a useful supplement to encourage regulatory bodies to promote changes in environmental policies as pollution source control can reduce the harmful effects of environmental pollutants.