1. Data Cleaning

We use Tukey's rule to eliminate outlier data points from the input data-set
We document the results for each

```
Column: CT confirmed
Detected 43 outlier rows
IQR = 838.0
lower threshold = -1257.0
upper threshold = 2095.0

Column: DC confirmed
Detected 19 outlier rows
IQR = 111.0
lower threshold = -130.5
upper threshold = 313.5

Column: CT deaths
Detected 38 outlier rows
IQR = 26.0
lower threshold = -39.0
upper threshold = 65.0

Column: DC deaths
Detected 15 outlier rows
IQR = 4.0
lower threshold = -6.0
upper threshold = 10.0
```

Note: We do not drop the outlier rows for the following reason (as instructed on Piazza):

1. We lose about 20% of the data-set, which may affect inferences
2. We lose large chunks of contiguous data-points, which may cause issues when performing time-series analysis
3. We lose 12 data points in the range of Feb 2021 to March 2021, which are all needed for the 2b) task to perform the 3 hypothesis tests.

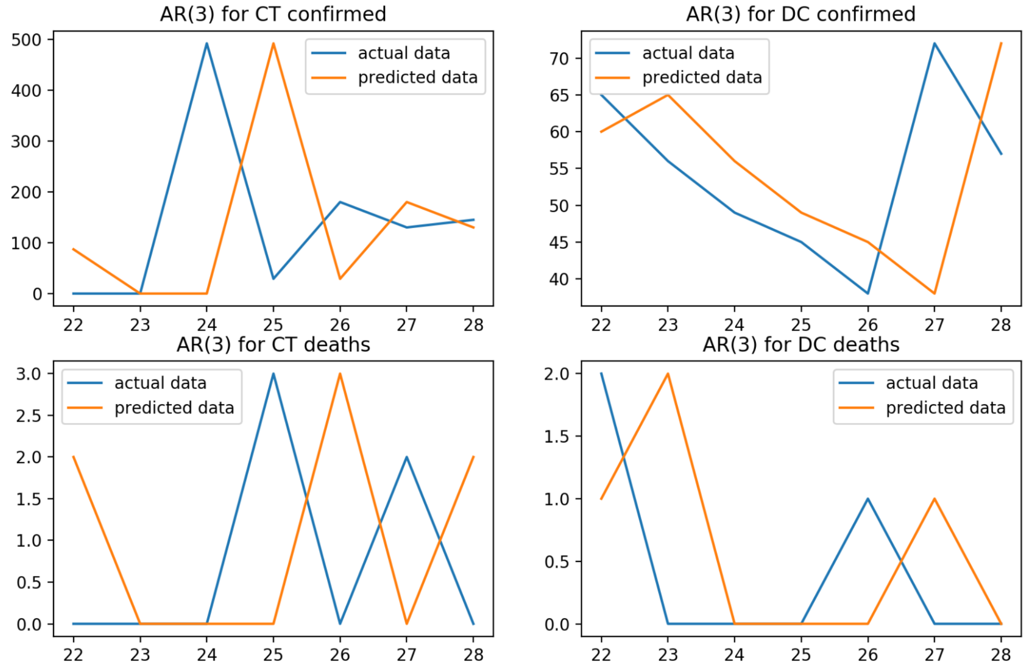
2. Mandatory Tasks

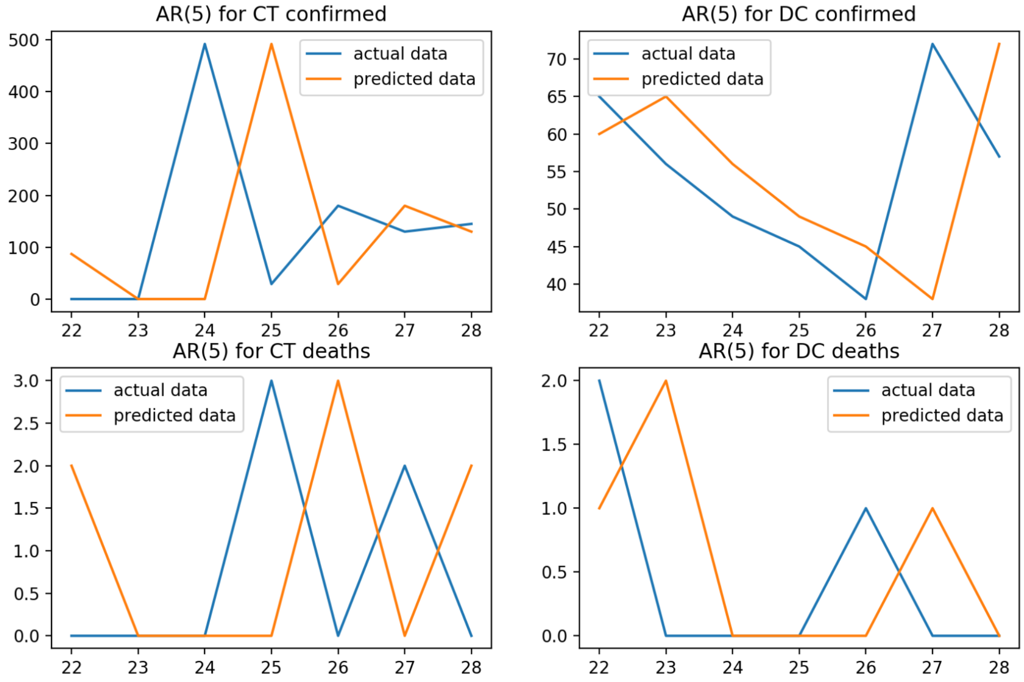
a. Time-series analysis

We perform time-series analysis and predict the number of cases and deaths for CT and DC states for the 4th week of August 2020, using the data from the 1st 3 weeks of August.

I. Auto-regression:

We perform auto-regression with parameters 3 and 5 to predict the number of cases and deaths for the 2 states in the 4th week of August. Below are the results, along with MSE and MAPE values

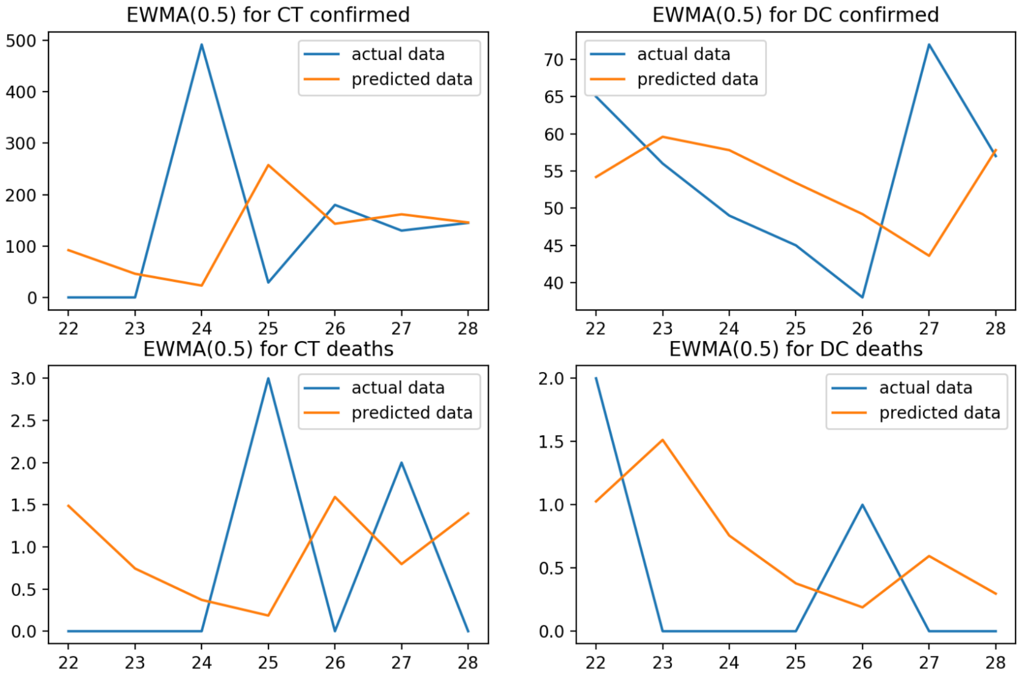


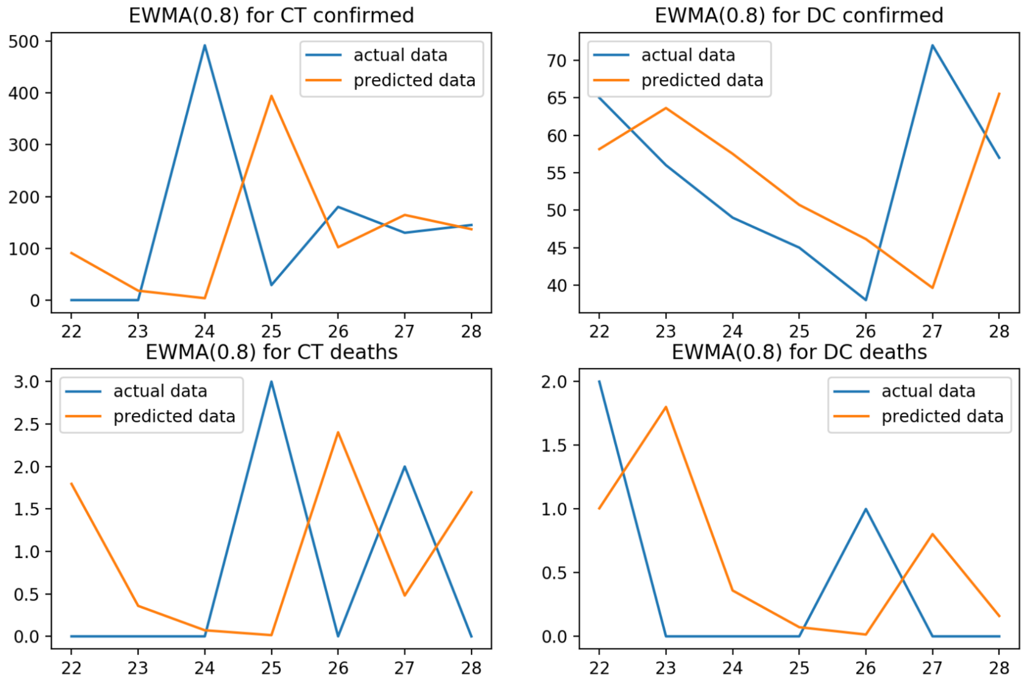


```
Predicting CT confirmed on 2020-08-22 using AR(3): [87.]
Predicting CT confirmed on 2020-08-23 using AR(3): [-1.36557432e-13]
Predicting CT confirmed on 2020-08-24 using AR(3): [-3.86357613e-14]
Predicting CT confirmed on 2020-08-25 using AR(3): [492.]
Predicting CT confirmed on 2020-08-26 using AR(3): [29.]
Predicting CT confirmed on 2020-08-27 using AR(3): [180.]
Predicting CT confirmed on 2020-08-28 using AR(3): [130.]
MSE for CT confirmed: [69932.57142857]
MAPE for CT confirmed: [365.84939581]

Predicting DC confirmed on 2020-08-22 using AR(3): [60.]
Predicting DC confirmed on 2020-08-23 using AR(3): [65.]
Predicting DC confirmed on 2020-08-24 using AR(3): [56.]
Predicting DC confirmed on 2020-08-25 using AR(3): [49.]
Predicting DC confirmed on 2020-08-26 using AR(3): [45.]
Predicting DC confirmed on 2020-08-27 using AR(3): [38.]
Predicting DC confirmed on 2020-08-28 using AR(3): [72.]
MSE for DC confirmed: [228.71428571]
MAPE for DC confirmed: [19.84248625]

Predicting CT deaths on 2020-08-22 using AR(3): [2.]
Predicting CT deaths on 2020-08-23 using AR(3): [1.55431223e-15]
Predicting CT deaths on 2020-08-24 using AR(3): [0.]
Predicting CT deaths on 2020-08-25 using AR(3): [0.]
Predicting CT deaths on 2020-08-26 using AR(3): [3.]
Predicting CT deaths on 2020-08-27 using AR(3): [6.66133815e-16]
Predicting CT deaths on 2020-08-28 using AR(3): [2.]
MSE for CT deaths: [4.28571429]
MAPE for CT deaths: [100.]

Predicting DC deaths on 2020-08-22 using AR(3): [1.]
Predicting DC deaths on 2020-08-23 using AR(3): [2.]
Predicting DC deaths on 2020-08-24 using AR(3): [-4.4408921e-16]
Predicting DC deaths on 2020-08-25 using AR(3): [-8.8817842e-16]
Predicting DC deaths on 2020-08-26 using AR(3): [0.]
Predicting DC deaths on 2020-08-27 using AR(3): [1.]
Predicting DC deaths on 2020-08-28 using AR(3): [0.]
MSE for DC deaths: [1.]
MAPE for DC deaths: [75.]

Predicting CT confirmed on 2020-08-22 using AR(5): [87.]
Predicting CT confirmed on 2020-08-23 using AR(5): [-3.06421555e-13]
Predicting CT confirmed on 2020-08-24 using AR(5): [-9.92261828e-14]
Predicting CT confirmed on 2020-08-25 using AR(5): [492.]
Predicting CT confirmed on 2020-08-26 using AR(5): [29.]
Predicting CT confirmed on 2020-08-27 using AR(5): [180.]
Predicting CT confirmed on 2020-08-28 using AR(5): [130.]
MSE for CT confirmed: [69932.57142857]
MAPE for CT confirmed: [365.84939581]

Predicting DC confirmed on 2020-08-22 using AR(5): [60.]
Predicting DC confirmed on 2020-08-23 using AR(5): [65.]
Predicting DC confirmed on 2020-08-24 using AR(5): [56.]
Predicting DC confirmed on 2020-08-25 using AR(5): [49.]
Predicting DC confirmed on 2020-08-26 using AR(5): [45.]
Predicting DC confirmed on 2020-08-27 using AR(5): [38.]
Predicting DC confirmed on 2020-08-28 using AR(5): [72.]
MSE for DC confirmed: [228.71428571]
MAPE for DC confirmed: [19.84248625]

Predicting CT deaths on 2020-08-22 using AR(5): [2.]
Predicting CT deaths on 2020-08-23 using AR(5): [-1.44328993e-15]
Predicting CT deaths on 2020-08-24 using AR(5): [2.3869795e-15]
Predicting CT deaths on 2020-08-25 using AR(5): [-9.43689571e-16]
Predicting CT deaths on 2020-08-26 using AR(5): [3.]
Predicting CT deaths on 2020-08-27 using AR(5): [1.99840144e-15]
Predicting CT deaths on 2020-08-28 using AR(5): [2.]
MSE for CT deaths: [4.28571429]
MAPE for CT deaths: [100.]

Predicting DC deaths on 2020-08-22 using AR(5): [1.]
Predicting DC deaths on 2020-08-23 using AR(5): [2.]
Predicting DC deaths on 2020-08-24 using AR(5): [-1.11022302e-15]
Predicting DC deaths on 2020-08-25 using AR(5): [-7.77156117e-16]
Predicting DC deaths on 2020-08-26 using AR(5): [1.83186799e-15]
Predicting DC deaths on 2020-08-27 using AR(5): [1.]
Predicting DC deaths on 2020-08-28 using AR(5): [0.]
MSE for DC deaths: [1.]
MAPE for DC deaths: [75.]
```

II. EWMA:

We apply EWMA estimation to predict the same results for the 4th week of august, with alpha values 0.5 and 0.8. The results are documented below:





```
Predicting CT confirmed on 2020-08-22 for alpha=0.5 using EWMA(0.5): 91.94676208496094
Predicting CT confirmed on 2020-08-23 for alpha=0.5 using EWMA(0.5): 45.97338104248047
Predicting CT confirmed on 2020-08-24 for alpha=0.5 using EWMA(0.5): 22.986690521240234
Predicting CT confirmed on 2020-08-25 for alpha=0.5 using EWMA(0.5): 257.4933452606201
Predicting CT confirmed on 2020-08-26 for alpha=0.5 using EWMA(0.5): 143.24667263031006
Predicting CT confirmed on 2020-08-27 for alpha=0.5 using EWMA(0.5): 161.62333631515503
Predicting CT confirmed on 2020-08-28 for alpha=0.5 using EWMA(0.5): 145.81166815757751
MSE for alpha = 0.5: 40728.850485233765
MAPE for alpha = 0.5: 185.7079850408238

Predicting DC confirmed on 2020-08-22 for alpha=0.5 using EWMA(0.5): 54.18722438812256
Predicting DC confirmed on 2020-08-23 for alpha=0.5 using EWMA(0.5): 59.59361219406128
Predicting DC confirmed on 2020-08-24 for alpha=0.5 using EWMA(0.5): 57.79680609703064
Predicting DC confirmed on 2020-08-25 for alpha=0.5 using EWMA(0.5): 53.39840304851532
Predicting DC confirmed on 2020-08-26 for alpha=0.5 using EWMA(0.5): 49.19920152425766
Predicting DC confirmed on 2020-08-27 for alpha=0.5 using EWMA(0.5): 43.59960076212883
Predicting DC confirmed on 2020-08-28 for alpha=0.5 using EWMA(0.5): 57.799800381064415
MSE for alpha = 0.5: 172.91308694416338
MAPE for alpha = 0.5: 18.569675412600834

Predicting CT deaths on 2020-08-22 for alpha=0.5 using EWMA(0.5): 1.487192153930664
Predicting CT deaths on 2020-08-23 for alpha=0.5 using EWMA(0.5): 0.743596076965332
Predicting CT deaths on 2020-08-24 for alpha=0.5 using EWMA(0.5): 0.371798038482666
Predicting CT deaths on 2020-08-25 for alpha=0.5 using EWMA(0.5): 0.185899019241333
Predicting CT deaths on 2020-08-26 for alpha=0.5 using EWMA(0.5): 1.5929495096206665
Predicting CT deaths on 2020-08-27 for alpha=0.5 using EWMA(0.5): 0.7964747548103333
Predicting CT deaths on 2020-08-28 for alpha=0.5 using EWMA(0.5): 1.3982373774051666
MSE for alpha = 0.5: 2.3947289513285797
```

```
MAPE for alpha = 0.5: 76.98981414238612

Predicting DC deaths on 2020-08-22 for alpha=0.5 using EWMA(0.5): 1.026616096496582
Predicting DC deaths on 2020-08-23 for alpha=0.5 using EWMA(0.5): 1.513308048248291
Predicting DC deaths on 2020-08-24 for alpha=0.5 using EWMA(0.5): 0.7566540241241455
Predicting DC deaths on 2020-08-25 for alpha=0.5 using EWMA(0.5): 0.37832701206207275
Predicting DC deaths on 2020-08-26 for alpha=0.5 using EWMA(0.5): 0.18916350603103638
Predicting DC deaths on 2020-08-27 for alpha=0.5 using EWMA(0.5): 0.5945817530155182
Predicting DC deaths on 2020-08-28 for alpha=0.5 using EWMA(0.5): 0.2972908765077591
MSE for alpha = 0.5: 0.7217998941424745
MAPE for alpha = 0.5: 64.87642228603363

Predicting CT confirmed on 2020-08-22 for alpha=0.8 using EWMA(0.8): 90.7661870313488
Predicting CT confirmed on 2020-08-23 for alpha=0.8 using EWMA(0.8): 18.15323740626976
Predicting CT confirmed on 2020-08-24 for alpha=0.8 using EWMA(0.8): 3.6306474812539493
Predicting CT confirmed on 2020-08-25 for alpha=0.8 using EWMA(0.8): 394.3261294962508
Predicting CT confirmed on 2020-08-26 for alpha=0.8 using EWMA(0.8): 102.06522589925015
Predicting CT confirmed on 2020-08-27 for alpha=0.8 using EWMA(0.8): 164.41304517985003
Predicting CT confirmed on 2020-08-28 for alpha=0.8 using EWMA(0.8): 136.88260903597003
MSE for alpha = 0.8: 55408.54640536814
MAPE for alpha = 0.8: 286.87484164956584

Predicting DC confirmed on 2020-08-22 for alpha=0.8 using EWMA(0.8): 58.147485241647374
Predicting DC confirmed on 2020-08-23 for alpha=0.8 using EWMA(0.8): 63.62949704832948
Predicting DC confirmed on 2020-08-24 for alpha=0.8 using EWMA(0.8): 57.52589940966591
Predicting DC confirmed on 2020-08-25 for alpha=0.8 using EWMA(0.8): 50.70517988193318
Predicting DC confirmed on 2020-08-26 for alpha=0.8 using EWMA(0.8): 46.141035976386625
Predicting DC confirmed on 2020-08-27 for alpha=0.8 using EWMA(0.8): 39.62820719527733
Predicting DC confirmed on 2020-08-28 for alpha=0.8 using EWMA(0.8): 65.52564143905545
MSE for alpha = 0.8: 199.61460286584585
MAPE for alpha = 0.8: 19.369467486595347

Predicting CT deaths on 2020-08-22 for alpha=0.8 using EWMA(0.8): 1.7958732272773372
Predicting CT deaths on 2020-08-23 for alpha=0.8 using EWMA(0.8): 0.35917464545546723
Predicting CT deaths on 2020-08-24 for alpha=0.8 using EWMA(0.8): 0.07183492909109343
Predicting CT deaths on 2020-08-25 for alpha=0.8 using EWMA(0.8): 0.01436698581821868
Predicting CT deaths on 2020-08-26 for alpha=0.8 using EWMA(0.8): 2.402873397163644
Predicting CT deaths on 2020-08-27 for alpha=0.8 using EWMA(0.8): 0.4805746794327286
Predicting CT deaths on 2020-08-28 for alpha=0.8 using EWMA(0.8): 1.6961149358865457
MSE for alpha = 0.8: 3.3189416528734883
MAPE for alpha = 0.8: 87.74618325054482

Predicting DC deaths on 2020-08-22 for alpha=0.8 using EWMA(0.8): 1.0049558160353946
Predicting DC deaths on 2020-08-23 for alpha=0.8 using EWMA(0.8): 1.8009911632070794
Predicting DC deaths on 2020-08-24 for alpha=0.8 using EWMA(0.8): 0.3601982326414157
Predicting DC deaths on 2020-08-25 for alpha=0.8 using EWMA(0.8): 0.07203964652828314
Predicting DC deaths on 2020-08-26 for alpha=0.8 using EWMA(0.8): 0.014407929305656625
Predicting DC deaths on 2020-08-27 for alpha=0.8 using EWMA(0.8): 0.8028815858611316
Predicting DC deaths on 2020-08-28 for alpha=0.8 using EWMA(0.8): 0.16057631717222628
MSE for alpha = 0.8: 0.8586299856898202
MAPE for alpha = 0.8: 74.1557081338323
```

b. Wald's, Z and T Hypothesis Tests

I. Wald's 1-Sample Test:

A. - Wald's 1-Sample Test Result for comparing means of **daily confirmed positive** cases in **CT** across Feb'21 and March'21
   - Sample mean = 998.129, True mean = 1068.679, std error estimate = 5.674
   - Rejected Null Hypothesis: We reject the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in CT, as **W-statistic = 12.433 exceeds threshold 1.96**

B. - Wald's 1-Sample Test Result for comparing means of **daily confirmed positive** cases in **DC** across Feb'21 and March'21
   - Sample mean = 126.290, True mean = 133.071, std error estimate = 2.018
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in DC, as **W-statistic = 3.36 exceeds threshold 1.96**

C. - Wald's 1-Sample Test Result for comparing means of **daily deaths** in **CT** across Feb'21 and March'21
   - Sample mean = 8.581, True mean = 20.750, std error estimate = 0.526
   - Rejected Null Hypothesis: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in CT, as **W-statistic = 23.131 exceeds threshold 1.96**

D. - Wald's 1-Sample Test Result for comparing means of **daily deaths** in **DC** across Feb'21 and March'21
   - Sample mean = 1.903, True mean = 3.286, std error estimate = 0.248
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in DC, as **W-statistic = 5.58 exceeds threshold 1.96**

II. 1-Sample Z Test:

A. - 1-Sample Z Test Result for comparing means of **daily confirmed positive** cases in **CT** across Feb'21 and March'21

   - Sample mean = 998.129, True mean = 1068.679, True std deviation = 1243.712

   - **Failed to reject Null Hypothesis**: We accept the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in CT, as **Z-statistic = 0.316 does not exceed threshold 1.96**

B. - 1-Sample Z Test Result for comparing means of **daily confirmed positive** cases in **DC** across Feb'21 and March'21

   - Sample mean = 126.290, True mean = 133.071, True std deviation = 92.760

   - **Failed to reject Null Hypothesis**: We accept the hypothesis that the mean of **daily confirmed positive** cases is same for Feb’21 and March’21 in DC, **as Z-statistic = 0.407 does not exceed threshold 1.96**

C. - 1-Sample Z Test Result for comparing means of **daily deaths** in **CT** across Feb'21 and March'21

   - Sample mean = 8.581, True mean = 20.750, True std deviation = 28.598

   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in CT, **as Z-statistic = 2.369 exceeds threshold 1.96**

D. - 1-Sample Z Test Result for comparing means of **daily deaths** in **DC** across Feb'21 and March'21

   - Sample mean = 1.903, True mean = 3.286, True std deviation = 3.837

   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in DC, **as Z-statistic = 2.006 exceeds threshold 1.96**

III. 1-Sample T Test:

A. - 1-Sample T Test Result for comparing means of **daily confirmed positive** cases in **CT** across Feb'21 and March'21

   - Sample mean = 998.129, True mean = 1068.679, Sample std deviation = 930.035

   Failed to reject Null Hypothesis: We accept the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in CT, as **T-statistic = 0.422 does not exceed threshold 2.042**

B. - 1-Sample T Test Result for comparing means of **daily confirmed positive** cases in **DC** across Feb'21 and March'21

   - Sample mean = 126.290, True mean = 133.071, Sample std deviation = 55.815

   - **Failed to reject Null Hypothesis**: We accept the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in DC, as **T-statistic = 0.676 does not exceed threshold 2.042**

C. - 1-Sample T Test Result for comparing means of **daily deaths** in **CT** across Feb'21 and March'21

   - Sample mean = 8.581, True mean = 20.750, Sample std deviation = 7.536

   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in CT, as **T-statistic = 8.991 exceeds threshold 2.042**

D. - 1-Sample T Test Result for comparing means of **daily deaths** in **DC** across Feb'21 and March'21

   - Sample mean = 1.903, True mean = 3.286, Sample std deviation = 2.574

   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in DC, as **T-statistic = 2.991 exceeds threshold 2.042**

IV. Wald's 2-Sample Test

A. - Wald's 2-Sample Test Result for comparing means of **daily confirmed positive cases** in **CT** across Feb'21 and March'21
   - Feb Mean = 1068.679, March Mean = 998.129, std error estimate = 8.388
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in CT, as **W-statistic = 8.41 exceeds threshold 1.96**

B. - Wald's 2-Sample Test Result for comparing means of **daily confirmed positive cases** in **DC** across Feb'21 and March'21
   - Feb Mean = 133.071, March Mean = 126.290, std error estimate = 2.971
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in DC, as **W-statistic = 2.282 exceeds threshold 1.96**

C. - Wald's 2-Sample Test Result for comparing means of **daily deaths** in **CT** across Feb'21 and March'21
   - Feb Mean = 20.750, March Mean = 8.581, std error estimate = 1.009
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in CT, as **W-statistic = 12.062 exceeds threshold 1.96**

D. - Wald's 2-Sample Test Result for comparing means of **daily deaths** in **DC** across Feb'21 and March'21
   - Feb Mean = 3.286, March Mean = 1.903, std error estimate = 0.423
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in DC, as **W-statistic = 3.27 exceeds threshold 1.96**

V. Unpaired 2-Sample T Test

A. - Unpaired 2-Sample T Test Result for comparing means of **daily confirmed positive cases** in **CT** across Feb'21 and March'21
   - Feb Mean = 1068.679 with sample size 28, March Mean = 998.129 with sample size 31, Sample pool std deviation = 8.473
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in CT, as **T-statistic = 8.326 exceeds threshold 2.002**

B. - Unpaired 2-Sample T Test Result for comparing means of **daily confirmed positive** cases in DC across Feb'21 and March'21
   - Feb Mean = 133.071 with sample size 28, March Mean = 126.290 with sample size 31, Sample pool std deviation = 2.363
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily confirmed positive cases is same for Feb’21 and March’21 in DC, as **T-statistic = 2.869 exceeds threshold 2.002**

C. - Unpaired 2-Sample T Test Result for comparing means of **daily deaths** in **CT** across Feb'21 and March'21
   - Feb Mean = 20.750 with sample size 28, March Mean = 8.581 with sample size 31, Sample pool std deviation = 0.993
   - **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in CT, as **T-statistic = 12.253 exceeds threshold 2.002**

D. - Unpaired 2-Sample T Test Result for comparing means of **daily deaths** in **DC** across Feb'21 and March'21
   - Feb Mean = 3.286 with sample size 28, March Mean = 1.903 with sample size 31, Sample pool std deviation = 0.423
   - Rejected Null Hypothesis: We reject the hypothesis that the mean of daily deaths is same for Feb’21 and March’21 in DC, as **T-statistic = 3.271 exceeds threshold 2.002**

VI. Applicability of Tests

A. Wald's Test -

- Wald's Test is applicable for Asymptotically Normal(AN) estimators. For the 1-Sample test, since the sample size M = 31, which is just over the minimum number of observations needed for Asymptotic Normality to kick in via Central Limit Theorem(CLT), the 1-Sample test may just be applicable.
- For the 2-Sample Test, the Feb data has only 28 observations, and since AN is needed for both estimators, the 2-Sample Test may just not be applicable.

B. Z-Test -

- For Z-test, since the true standard deviation is unknown and being calculated by us, the sample size is not large and since we do not know if the sample data is normally distributed (given to be Poisson), Z-Test may not be applicable.

C. T-Test -

- T-Tests work well when the true standard deviation is unknown and for small sample sizes. However, a requirement is that the data should be Normally distributed. Since that is not the case, we can say that both 1 and 2-Sample (Unpaired) T-Tests are not applicable.

c. 1/2 - Sample KS and Permutation Tests

I. Permutation Test

Note - The data for the two states for our mandatory tasks is skewed and higher for the first state so in general the outcome is rejecting the hypothesis

A) Permutation test for confirmed cases for CT and DC -1000 permutations

T_obs = 1244.5760869565217
p_value_1000 = 0.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily confirmed positive cases is same in both **CT** and **DC**, as Permutation test **p-value = 0.0 does not exceed threshold 0.05**

B) Permutation test for deaths cases for CT and DC -1000 permutations

T_obs = 14.434782608695654
p_value_1000 = 0.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily deaths is same in both **CT** and **DC**, as Permutation test **p-value = 0.0 does not exceed threshold 0.05**

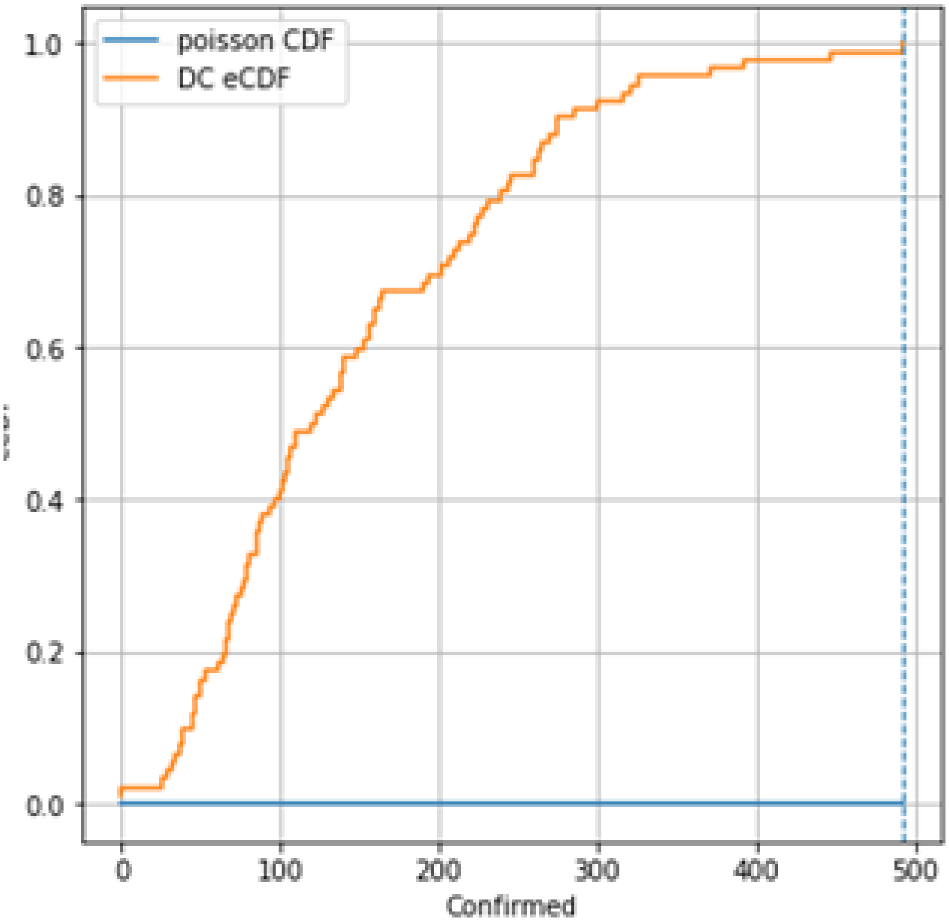
II. KS 1-Sample Tests

Note - The data for the two states for our mandatory tasks is skewed and higher for the first state so in general the outcome is rejecting the hypothesis

Note - As the data is skewed for both states and mme parameters are calculated for the first state and these parameters are used to calculate the KS statistic for the second state, the distribution cdfs are not as expected and we tend to reject the null hypothesis

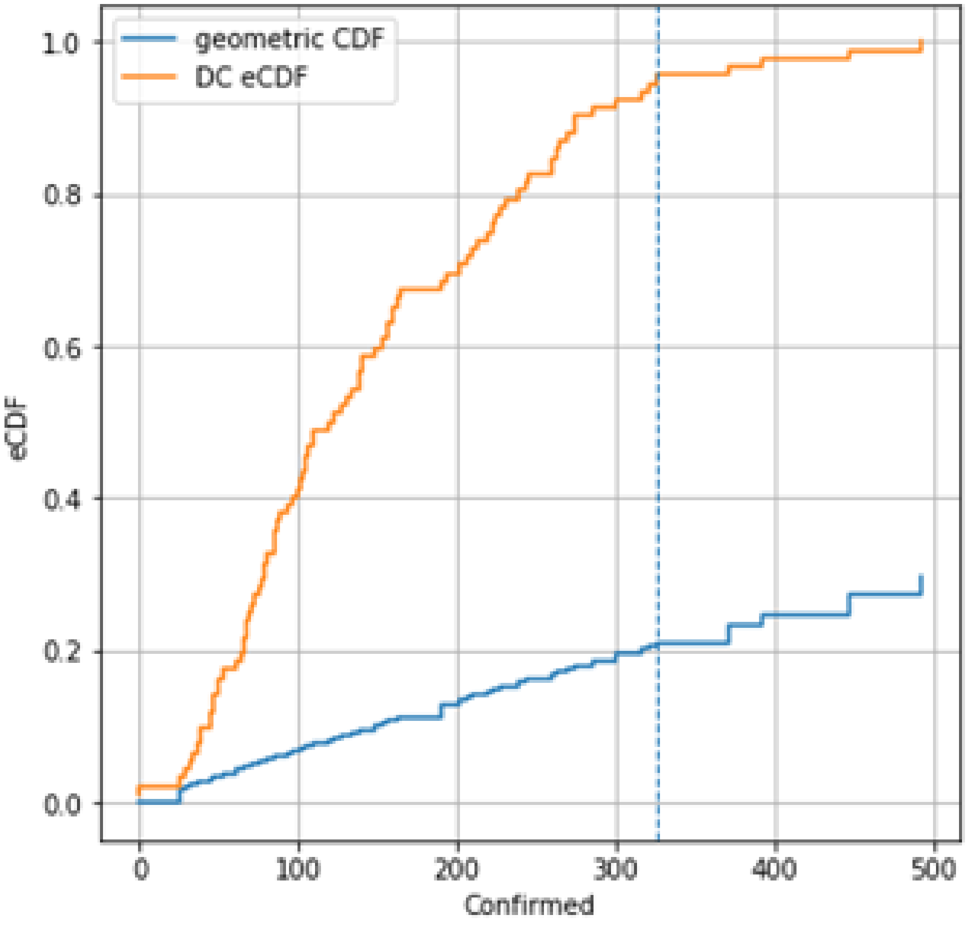
A) KS_1_sample_test for poisson distribution on DC confirmed cases

lambda_mme = 1393.0217391304348
KS Statistic is 1.0 at x = 492.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily confirmed positive cases in **DC** is **poisson**, as KS Statistic d = **1.0 exceeds threshold 0.05**



B) KS_1_sample_test for geometric distribution on DC confirmed cases

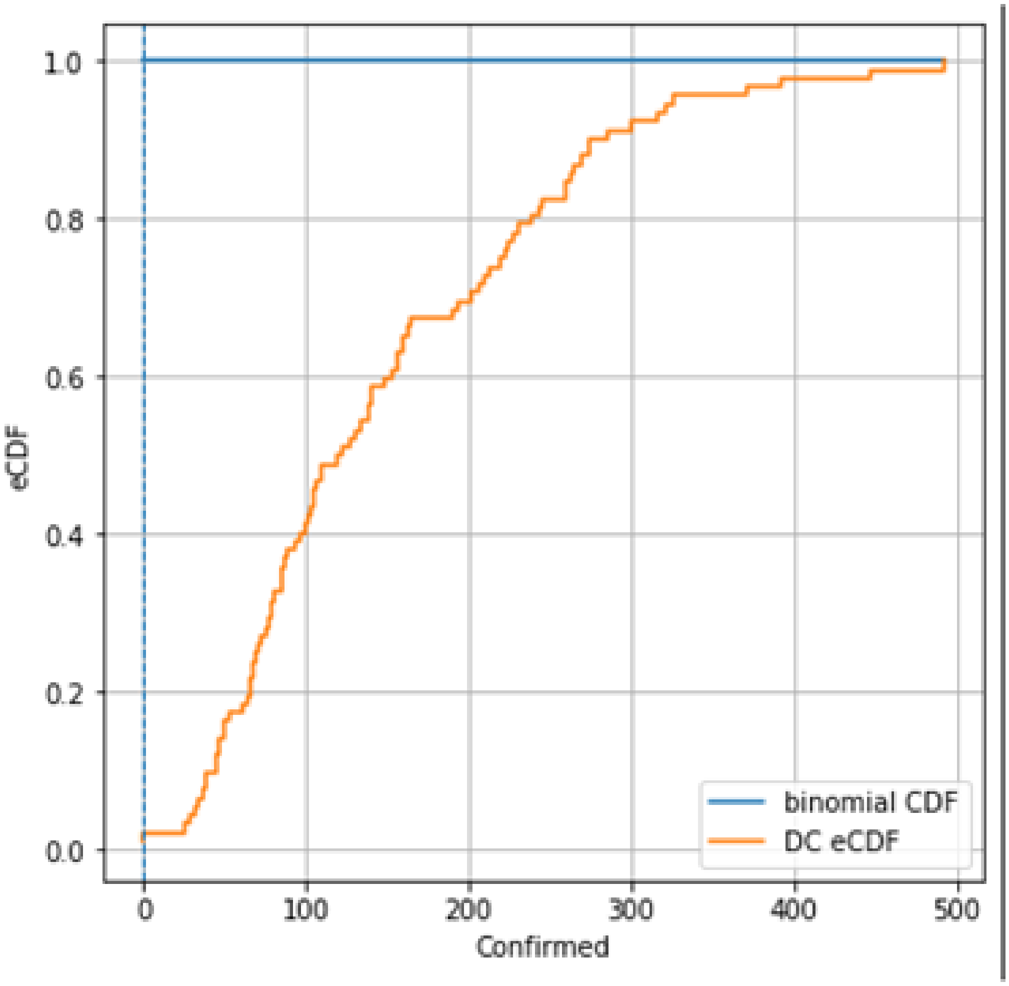
p_mme = 0.0007178638867647748
KS Statistic is 0.7478 at x = 326.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution daily confirmed positive cases in **DC** is **geometric**, as KS Statistic d = **0.7478 exceeds threshold 0.05**



C) KS_1_sample_test for binomial distribution on DC confirmed cases

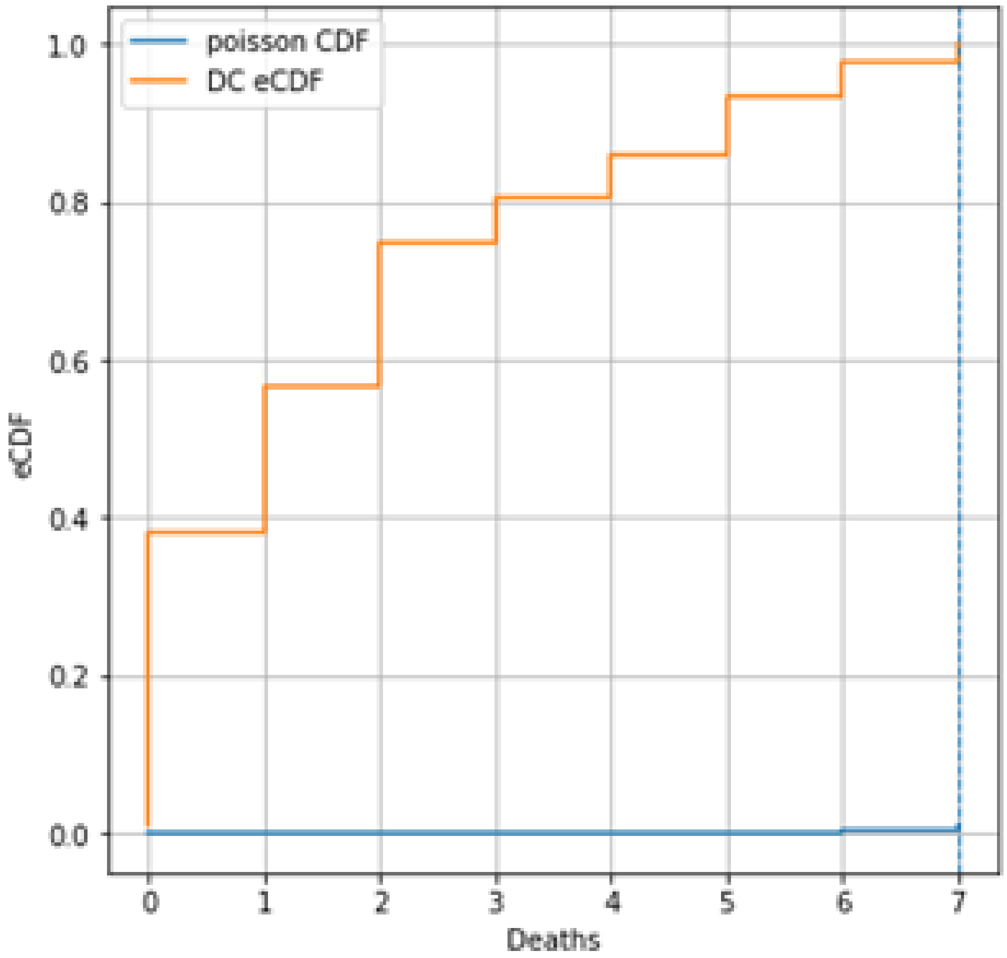
Note - The assumption that data is binomial distributed may be wrong

n_mme = -0.614697664831132 , p_mme = -2266.1900619276335
KS Statistic is 1.0 at x = 0.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily confirmed positive cases in **DC** is **binomial**, as KS Statistic d = **1.0 exceeds threshold 0.05**



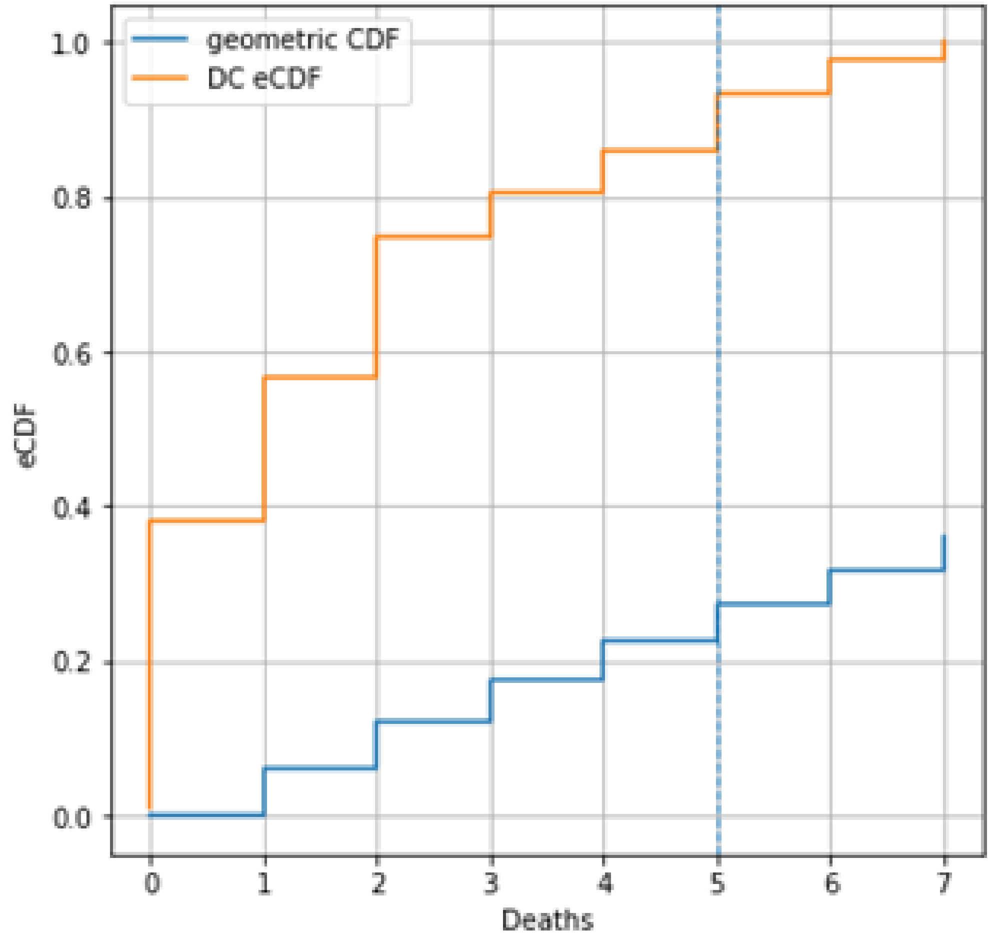
D) KS_1_sample_test for poisson distribution on DC death cases

lambda_mme = 16.16304347826087
KS Statistic is 0.9909 at x = 7.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily deaths in **DC** is **poisson**, as KS Statistic d = **0.9909 exceeds threshold 0.05**



E) **KS_1_sample_test for geometric distribution on DC death cases**

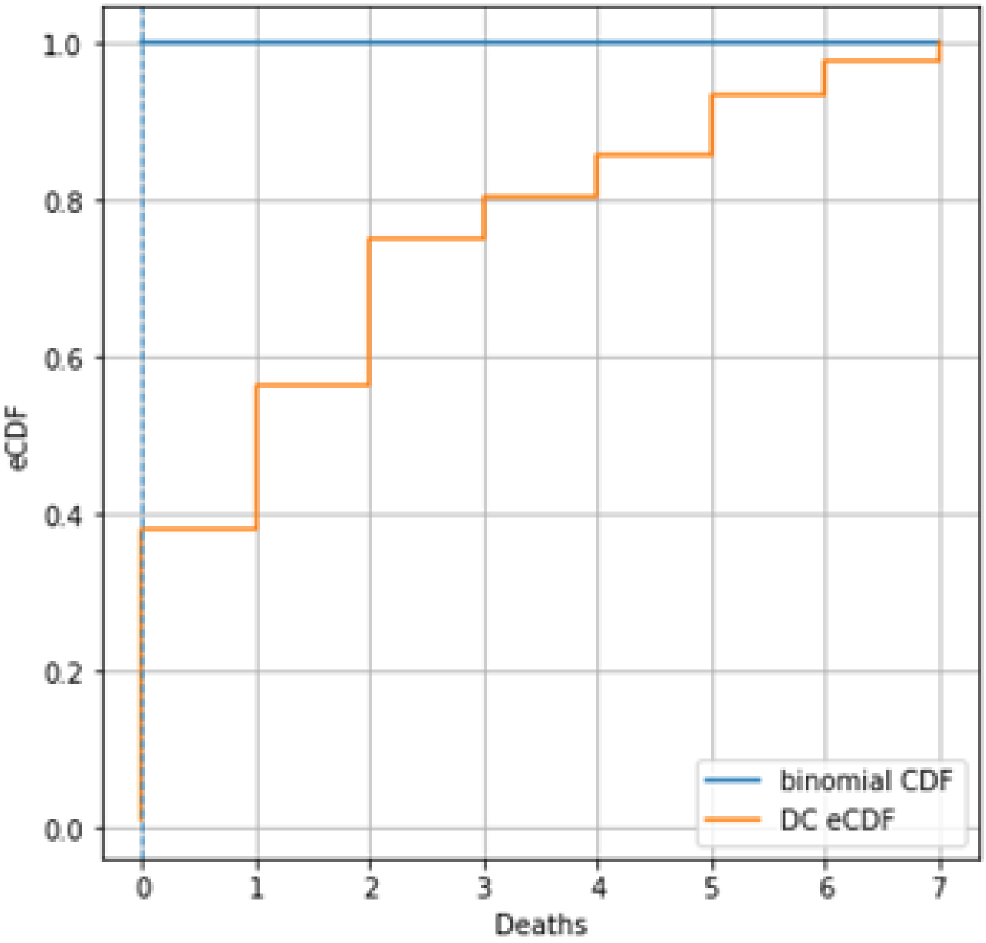
p_mme = 0.061869535978480154
KS Statistic is 0.6614 at x = 5.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily deaths in **DC** is **geometric**, as KS Statistic d = **0.6614 exceeds threshold 0.05**



F) KS_1_sample_test for binomial distribution on DC death cases

Note - The assumption that data is binomial distributed may be wrong

n_mme = -0.5386527499763096 , p_mme = -30.00642525072365
KS Statistic is 1.0 at x = 0.0
Rejected Null Hypothesis: We reject the hypothesis that the distribution of daily deaths in DC is binomial, as KS Statistic d = 1.0 exceeds threshold 0.05

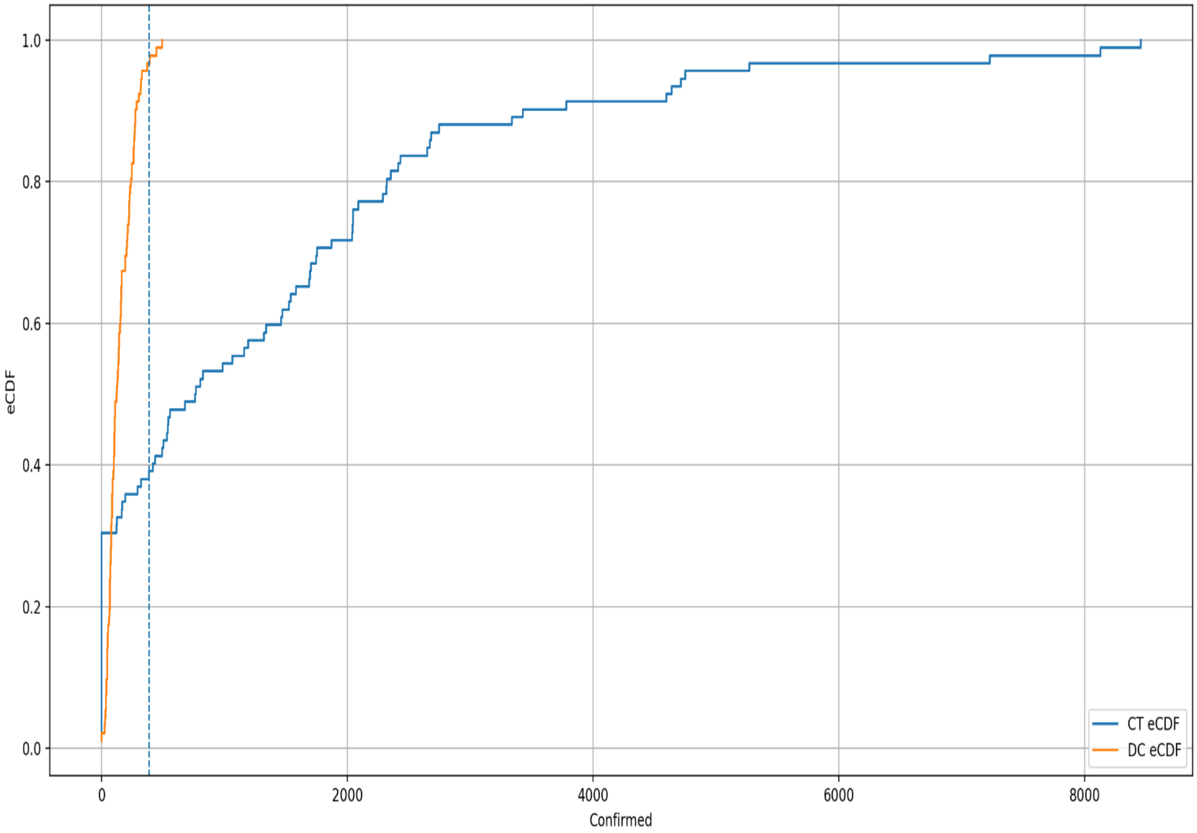


III. KS 2-Sample Tests

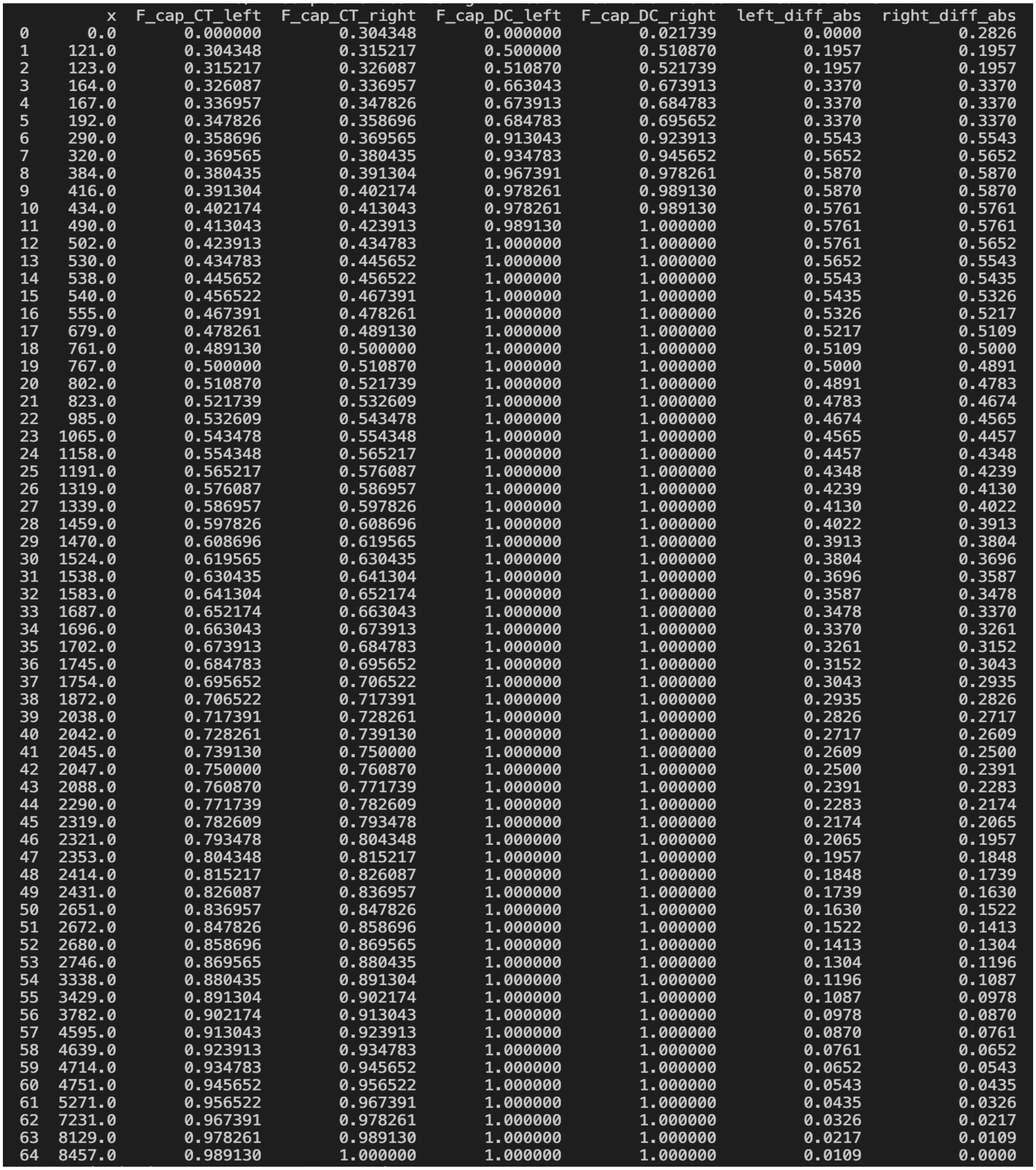
A. Daily Confirmed Cases in CT and DC

- KS Statistic is **0.587** at **x = 384.0**
- **Rejected Null Hypothesis**: We reject the hypothesis that the distribution of daily confirmed positive cases is same in both CT and DC, as **KS Statistic d = 0.587 exceeds threshold 0.05**

KS eCDF Graph



2- Sample KS - Table for CT - DC Confirmed Cases

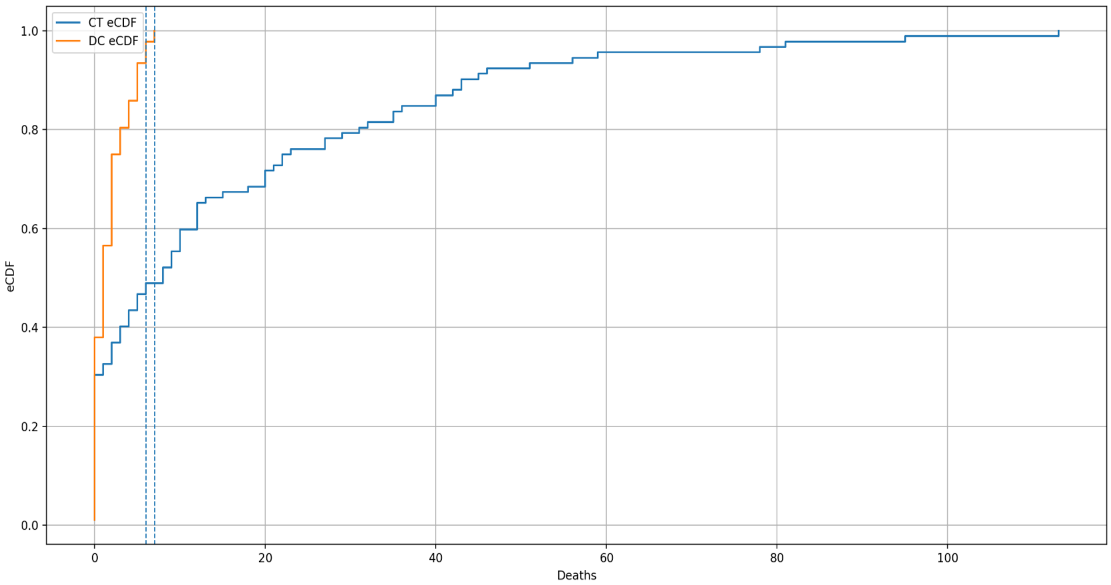


```
         x  F_cap_CT_left  F_cap_CT_right  F_cap_DC_left  F_cap_DC_right  left_diff_abs  right_diff_abs
0      0.0       0.000000        0.304348       0.000000        0.021739         0.0000          0.2826
1    121.0       0.304348        0.315217       0.500000        0.510870         0.1957          0.1957
2    123.0       0.315217        0.326087       0.510870        0.521739         0.1957          0.1957
3    164.0       0.326087        0.336957       0.663043        0.673913         0.3370          0.3370
4    167.0       0.336957        0.347826       0.673913        0.684783         0.3370          0.3370
5    192.0       0.347826        0.358696       0.684783        0.695652         0.3370          0.3370
6    290.0       0.358696        0.369565       0.913043        0.923913         0.5543          0.5543
7    320.0       0.369565        0.380435       0.934783        0.945652         0.5652          0.5652
8    384.0       0.380435        0.391304       0.967391        0.978261         0.5870          0.5870
9    416.0       0.391304        0.402174       0.978261        0.989130         0.5870          0.5870
10   434.0       0.402174        0.413043       0.978261        0.989130         0.5761          0.5761
11   490.0       0.413043        0.423913       0.989130        1.000000         0.5761          0.5761
12   502.0       0.423913        0.434783       1.000000        1.000000         0.5761          0.5652
13   530.0       0.434783        0.445652       1.000000        1.000000         0.5652          0.5543
14   538.0       0.445652        0.456522       1.000000        1.000000         0.5543          0.5435
15   540.0       0.456522        0.467391       1.000000        1.000000         0.5435          0.5326
16   555.0       0.467391        0.478261       1.000000        1.000000         0.5326          0.5217
17   679.0       0.478261        0.489130       1.000000        1.000000         0.5217          0.5109
18   761.0       0.489130        0.500000       1.000000        1.000000         0.5109          0.5000
19   767.0       0.500000        0.510870       1.000000        1.000000         0.5000          0.4891
20   802.0       0.510870        0.521739       1.000000        1.000000         0.4891          0.4783
21   823.0       0.521739        0.532609       1.000000        1.000000         0.4783          0.4674
22   985.0       0.532609        0.543478       1.000000        1.000000         0.4674          0.4565
23  1065.0       0.543478        0.554348       1.000000        1.000000         0.4565          0.4457
24  1158.0       0.554348        0.565217       1.000000        1.000000         0.4457          0.4348
25  1191.0       0.565217        0.576087       1.000000        1.000000         0.4348          0.4239
26  1319.0       0.576087        0.586957       1.000000        1.000000         0.4239          0.4130
27  1339.0       0.586957        0.597826       1.000000        1.000000         0.4130          0.4022
28  1459.0       0.597826        0.608696       1.000000        1.000000         0.4022          0.3913
29  1470.0       0.608696        0.619565       1.000000        1.000000         0.3913          0.3804
30  1524.0       0.619565        0.630435       1.000000        1.000000         0.3804          0.3696
31  1538.0       0.630435        0.641304       1.000000        1.000000         0.3696          0.3587
32  1583.0       0.641304        0.652174       1.000000        1.000000         0.3587          0.3478
33  1687.0       0.652174        0.663043       1.000000        1.000000         0.3478          0.3370
34  1696.0       0.663043        0.673913       1.000000        1.000000         0.3370          0.3261
35  1702.0       0.673913        0.684783       1.000000        1.000000         0.3261          0.3152
36  1745.0       0.684783        0.695652       1.000000        1.000000         0.3152          0.3043
37  1754.0       0.695652        0.706522       1.000000        1.000000         0.3043          0.2935
38  1872.0       0.706522        0.717391       1.000000        1.000000         0.2935          0.2826
39  2038.0       0.717391        0.728261       1.000000        1.000000         0.2826          0.2717
40  2042.0       0.728261        0.739130       1.000000        1.000000         0.2717          0.2609
41  2045.0       0.739130        0.750000       1.000000        1.000000         0.2609          0.2500
42  2047.0       0.750000        0.760870       1.000000        1.000000         0.2500          0.2391
43  2088.0       0.760870        0.771739       1.000000        1.000000         0.2391          0.2283
44  2290.0       0.771739        0.782609       1.000000        1.000000         0.2283          0.2174
45  2319.0       0.782609        0.793478       1.000000        1.000000         0.2174          0.2065
46  2321.0       0.793478        0.804348       1.000000        1.000000         0.2065          0.1957
47  2353.0       0.804348        0.815217       1.000000        1.000000         0.1957          0.1848
48  2414.0       0.815217        0.826087       1.000000        1.000000         0.1848          0.1739
49  2431.0       0.826087        0.836957       1.000000        1.000000         0.1739          0.1630
50  2651.0       0.836957        0.847826       1.000000        1.000000         0.1630          0.1522
51  2672.0       0.847826        0.858696       1.000000        1.000000         0.1522          0.1413
52  2680.0       0.858696        0.869565       1.000000        1.000000         0.1413          0.1304
53  2746.0       0.869565        0.880435       1.000000        1.000000         0.1304          0.1196
54  3338.0       0.880435        0.891304       1.000000        1.000000         0.1196          0.1087
55  3429.0       0.891304        0.902174       1.000000        1.000000         0.1087          0.0978
56  3782.0       0.902174        0.913043       1.000000        1.000000         0.0978          0.0870
57  4595.0       0.913043        0.923913       1.000000        1.000000         0.0870          0.0761
58  4639.0       0.923913        0.934783       1.000000        1.000000         0.0761          0.0652
59  4714.0       0.934783        0.945652       1.000000        1.000000         0.0652          0.0543
60  4751.0       0.945652        0.956522       1.000000        1.000000         0.0543          0.0435
61  5271.0       0.956522        0.967391       1.000000        1.000000         0.0435          0.0326
62  7231.0       0.967391        0.978261       1.000000        1.000000         0.0326          0.0217
63  8129.0       0.978261        0.989130       1.000000        1.000000         0.0217          0.0109
64  8457.0       0.989130        1.000000       1.000000        1.000000         0.0109          0.0000
```

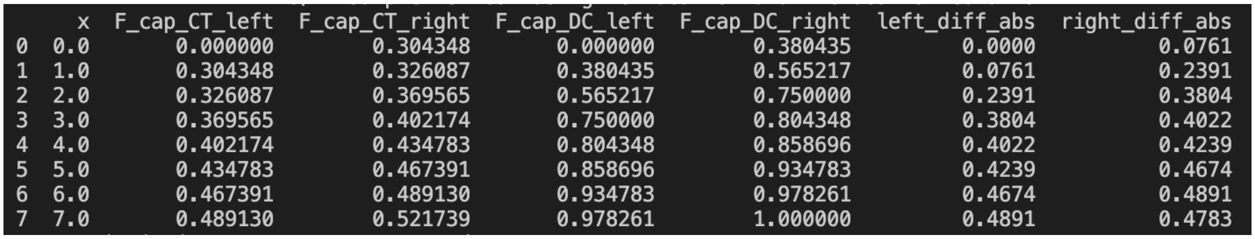
B. On Daily Deaths in CT and DC

- KS Statistic is **0.4891** at **x = 7.0 and 6.0**
- **Rejected Null Hypothesis**: We reject the hypothesis that the distribution of daily deaths is same in both CT and DC, as **KS Statistic d = 0.4891 exceeds threshold 0.05**

KS eCDF Graph



2- Sample KS - Table for CT - DC Deaths

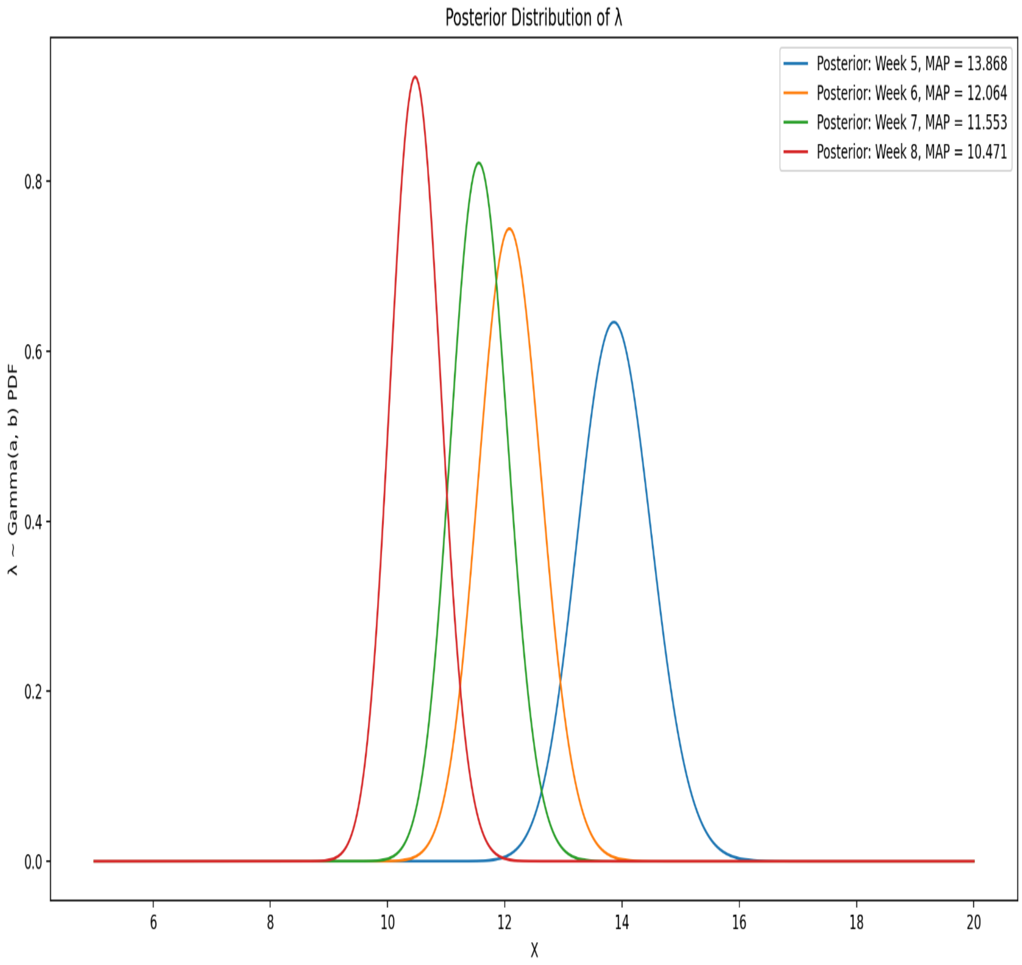


```
     x  F_cap_CT_left  F_cap_CT_right  F_cap_DC_left  F_cap_DC_right  left_diff_abs  right_diff_abs
0  0.0       0.000000        0.304348       0.000000        0.380435         0.0000          0.0761
1  1.0       0.304348        0.326087       0.380435        0.565217         0.0761          0.2391
2  2.0       0.326087        0.369565       0.565217        0.750000         0.2391          0.3804
3  3.0       0.369565        0.402174       0.750000        0.804348         0.3804          0.4022
4  4.0       0.402174        0.434783       0.804348        0.858696         0.4022          0.4239
5  5.0       0.434783        0.467391       0.858696        0.934783         0.4239          0.4674
6  6.0       0.467391        0.489130       0.934783        0.978261         0.4674          0.4891
7  7.0       0.489130        0.521739       0.978261        1.000000         0.4891          0.4783
```

d. Posterior Distributions of λ

- **Results:**

MAP for week 5 Posterior Distribution of λ = **13.868**
MAP for week 6 Posterior Distribution of λ = **12.064**
MAP for week 7 Posterior Distribution of λ = **11.553**
MAP for week 8 Posterior Distribution of λ = **10.471**



The posterior distribution is found for each successive week by making use of the Conjugate Priors for Poisson / Exponential distributions, called Gamma distributions. The derivation is shown below

2d) $D = \{x_1, \ldots x_n\} \sim$ Poisson (λ)

Prior of $\lambda \sim$ Exp(λ_1) with mean β.

→ For exponential, mean $= E[X] = 1/\lambda_1$

∴ we have $\lambda_1 = 1/\beta$

⇒ Prior of $\lambda \sim$ Exp$(1/\beta)$

now, $f(\lambda | D) \propto f(D|\lambda) \cdot f(\lambda)$

i) $f(D|\lambda) = \mathcal{L}(D) = f(\{x_1, \ldots x_n\} | \lambda)$

$$= \prod_{i=1}^{n} f(x_i | \lambda)$$

$$= \prod_{i=1}^{n} \frac{\lambda^{x_i} \cdot e^{-\lambda}}{x_i!} = \frac{\lambda^{\Sigma x_i} \cdot e^{-n\lambda}}{\prod_{i=1}^{n} x_i!}$$

ii) $f(\lambda) = \text{Exp}(\lambda | \frac{1}{\beta})$

$$= \frac{1}{\beta} \cdot e^{-\frac{1}{\beta} \cdot \lambda}$$

constant wrt λ, ignore

$$\Rightarrow f(\lambda | D) \propto \boxed{\frac{1}{\beta \cdot \prod_{i=1}^{n} x_i!}} \cdot \lambda^{\Sigma x_i} \cdot e^{-n\lambda} \cdot e^{-1/\beta \cdot \lambda}$$

$$\propto \lambda^{\Sigma x_i} \cdot e^{-\lambda(n + 1/\beta)} \qquad (1)$$

Now, the conjugate prior of Poisson is a Gamma distribution of the form

$$\text{Gamma}(\lambda | a, b) = \frac{\lambda^{a-1} \cdot e^{-\lambda \cdot \frac{1}{b}}}{b^a \cdot \Gamma a}$$

$$\propto \lambda^{a-1} \cdot e^{-\lambda \cdot \frac{1}{b}} \qquad (2)$$

(from (1) & (2))

$$\Rightarrow a = \Sigma x_i + 1, \quad b = \frac{1}{n + \frac{1}{\beta}}$$

$\therefore$ Posterior$_1(\lambda)$ is of the form -

$$\text{Gamma}\left(\sum_1^n X_i + 1, \frac{1}{n + 1/\beta}\right)$$

$\rightarrow$ Posterior$_2(\lambda) \propto \prod_{i=n}^{n+m} \lambda^{X_i} \cdot e^{-\lambda} \cdot \lambda^{\sum_1^n X_i} \cdot e^{-\lambda(n+\frac{1}{\beta})}$

$$\cancel{\propto \lambda^{\sum_n^{n+m}}}$$

$$\propto \lambda^{\sum_n^{n+m} X_i} \cdot e^{-m\lambda} \cdot \lambda^{\sum_1^n X_i} \cdot e^{-\lambda(n+\frac{1}{\beta})}$$

$$\propto \lambda^{\sum_1^{n+m} X_i} \cdot e^{-\lambda(m+n+1/\beta)}$$

$$= \text{Gamma}\left(\sum_1^{n+m} X_i + 1, \frac{1}{m+n+1/\beta}\right)$$

Similarly, Posterior$_3(\lambda) = \text{Gamma}\left(\sum_1^{n+m+l} X_i + 1, \frac{1}{m+n+l+\frac{1}{\beta}}\right)$

where $\beta = \lambda_{MME} = \frac{\sum X_i}{n}$ (for first 4 weeks)

3. Exploratory Analysis using X dataset of US Domestic Flights Cancellation data (Jan-Jun 2020)

a. <u>Chi-square test for independent samples</u>

We perform a chi-square test to check whether the presence of covid cases affected the number of flight cancellations.

We take the count of flights cancelled in the months with *lowest*, and *highest* covid cases for *NY state*, and use them to test our hypothesis.

H_0: covid case count is independent of number of flight cancellations
H_1: covid case count is dependent of number of flight cancellations

Our chi-square table will look like:

	Month with min cases	**Month with max cases**
# cancelled flights	669	11413
# on-time flights	60739	9390

p-value = **0.0** < 0.05

Hence, **reject** H_0

<u>We conclude that the covid case count and number of flight cancellations.</u>
<u>are **not independent** of each other</u>

b. Pearson’s correlation test

We perform Pearson’s correlation test with number of new cases and number of flight cancellations per day as the two datasets.

We look at the daily number of new cases and number of flight cancellations (includes departing as well as arriving flights) in the state of **New York**.

Hypothesis -

H_0: number of new cases and number of flight cancellations are not linearly correlated
H_1 : number of new cases and number of flight cancellations are linearly correlated

The pearson correlation coefficient computed over the period of 22 Jan 2020 to 30 Jun 2020 is **0.7064795731186282**

Since correlation coefficient is greater than 0.5 we **reject H_0**.

This implies a positive correlation between the number of new cases and the number of cancelled flights i.e. as the number of cases increases the number of flight cancellations also increases and as the number of cases decreases the number of flight cancellations also decreases.

Plotting the graphs of cases and flight cancellations over this date range shows that this observation holds.

Fig: Cases Vs Days from 22 Jun 2020

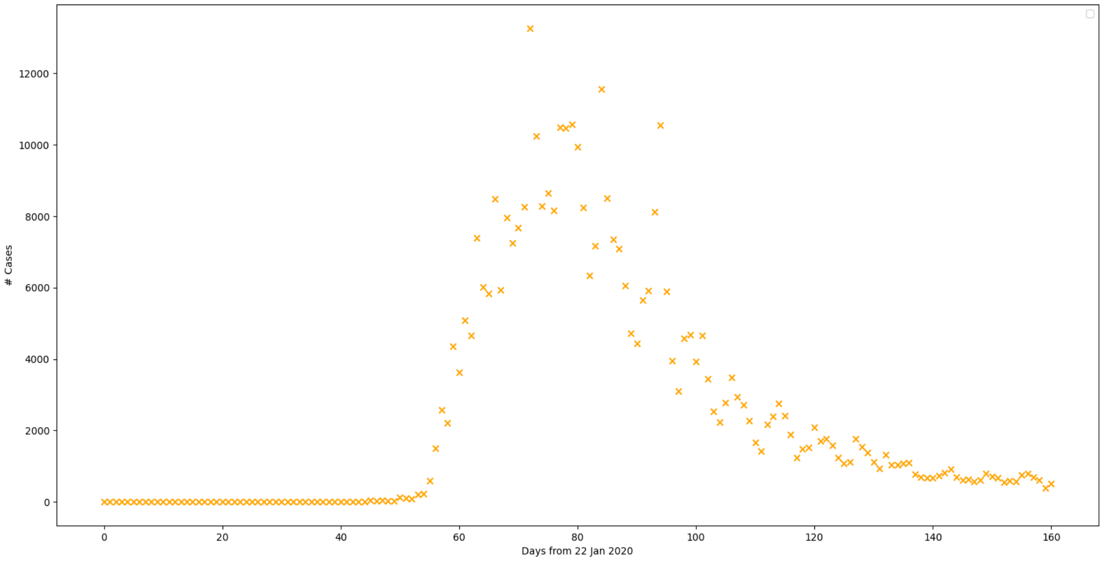
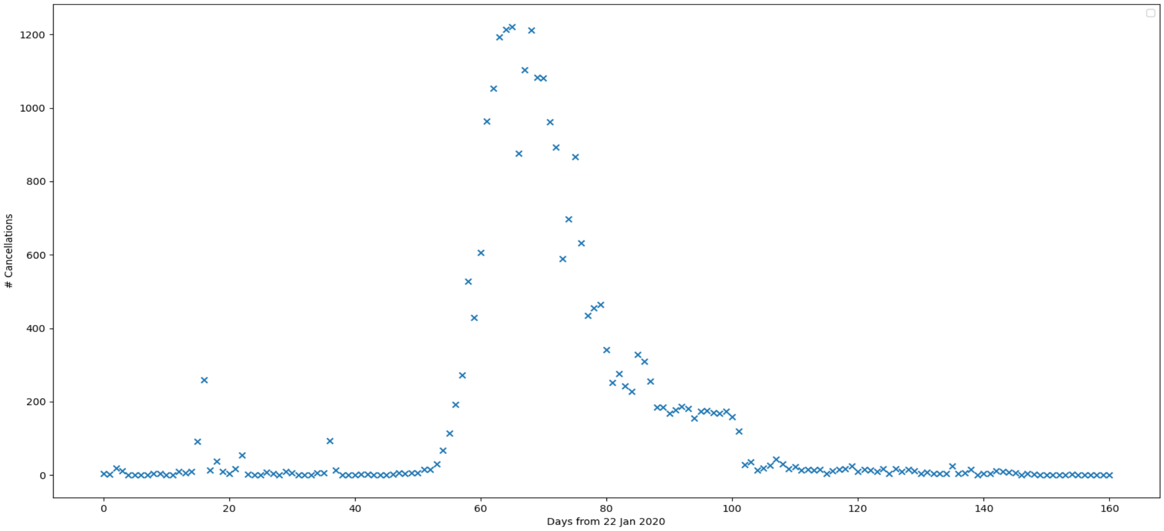


Fig: Flight cancellations Vs Days from 22 Jun 2020



c. One-Tailed Unpaired T-Test

Test Hypothesis -
H_0 : Mean of daily NY flight cancellation ratio in the month with no/least covid cases (Feb 2020) >= Mean of daily flight cancellation ratio in month with highest covid cases (April 2020)
H_1 : Mean of daily NY flight cancellation ratio in the month with no/least covid cases (found to be Feb 2020 in our tests) < Mean of daily flight cancellation ratio in month with highest covid cases (found to be April 2020 in our tests)

Test Applicability / Assumptions -

- We assume the daily cancellation ratio (cancelled flights / total scheduled flights) to be distributed normally.
- Since T-tests work well when true standard deviation of the data is unknown and for small number of observations, we can say that T-Test is hence applicable for our use-case

Results:

- Sample size of X = 29, Sample size of Y = 30, Sample pool std deviation = 0.059
- **Rejected Null Hypothesis**: We reject the hypothesis that the mean of daily flight cancellation ratio in a month with no covid cases is greater than or equal to that in a month with the highest number of covid cases, as **T-statistic = -8.544 is less than threshold -2.002**
- Since the **p-value 4.3237151577116935e-12** <<< **alpha = 0.05** (significance level), we can say that we can **reject the null hypothesis with a great degree of confidence**

Inference:

- Based on the above results and due to the extremely small p-value that we have obtained from the tests, we can say with a good deal of certainty that since the mean of cancellation ratio in the month with a large number of covid positive cases is found to be greater than one with almost no cases, **COVID-19 has had an impact on the increase in the flight cancellations in the heavily impacted months**.