1. Data Cleaning

- The given data was the cumulative cases and death in MT and NC.

- We first obtained the daily data from this to better analyse the data cleaning and inferences.

- We first noticed that the data had negative daily values for the number of cases in both the states.
  They were replaced by taking the mean of the surrounding values so that the trend is not disturbed and data's consistency is maintained. The number of negative values obtained are:

```
MT_cases_daily = 2
NC_cases_daily = 1
MT_death_daily = 3
NC_death_daily = 4
```

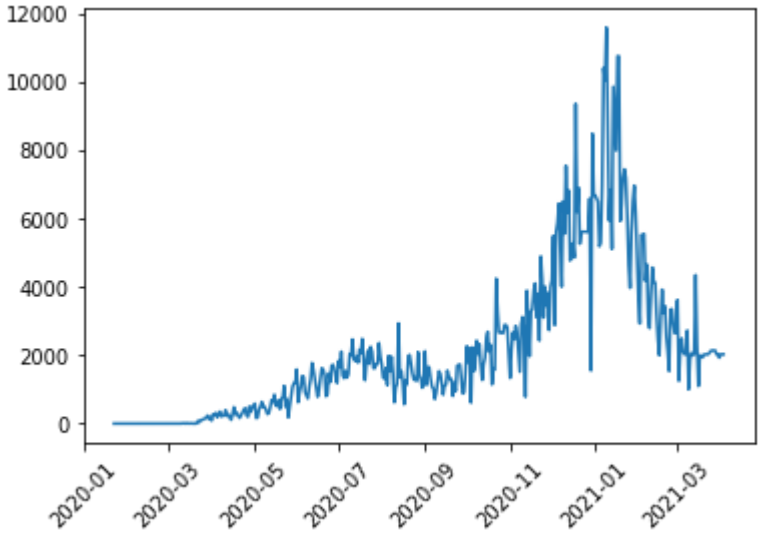
- We also noticed that for some days the value of the covid cases/deaths were 0. This can be acceptable for the initial months when the impact of COVID was minimal. But after the first few months, the cases and deaths started rising. Getting a 0 value for a day, especially when the trend from the data clearly shows that value lies in a range of tens or hundreds (or even thousands), can be interpreted as missing values rather than the true figure.
  These missing values were also replaced by previous days' values in such a manner that trend is not disturbed and data's consistency is maintained.

- Next, we applied the Tukey's rule to find the outliers and remove them accordingly.
  We faced 2 major issues with applying Tukey's rule on the whole dataset:
  1) As the COVID date is increasing (likely in a geometric distribution manner), and the cases are very less in the initial months and are very high in the last few months, if the Tukey's rule is applied on the whole dataset at once, then most of the points of the final few months will be classified as outliers.
  2) Also, if there is a high value in the first few months or a low value in the final few months, then it should be flagged. But if Tukey's rule is applied on the whole data at once, this case won't be flagged.

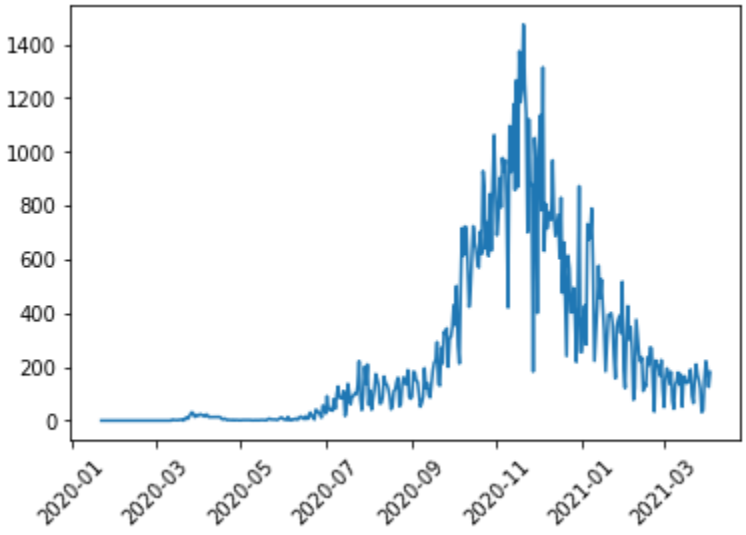
  To deal with this issue, we applied Tukey's rule on periods of 30 days each. This prevented the values in final few months to be marked as outliers. It also helped in removing unusually high or low values from each 30 day period.
  Here are the outliers in the data:

```
MT_cases_daily = 12
NC_cases_daily = 20
MT_death_daily = 28
NC_death_daily = 12
```

- The cumulative cases and deaths were calculated from this cleaned data and added back in.
- Example Plot after cleaning:
  NC daily cases



  MT daily cases



2A. AR and EWMA

- **AutoRegression Results**

  MSE
- MSE MT CASES (parameter=3) = 1637.7142857142858
- MSE NC CASES (parameter=3) = 69543.0

- MSE MT DEATH (parameter=3) = 6.142857142857143
- MSE NC DEATH (parameter=3) = 128.0
- MSE MT CASES (parameter=5) = 2622.0
- MSE NC CASES (parameter=5) = 87424.14285714286
- MSE MT DEATH (parameter=5) = 6.428571428571429
- MSE NC DEATH (parameter=5) = 131.42857142857142

  MAPE in %
- MAPE MT CASES (parameter=3) = 37.855703551158726
- MAPE NC CASES (parameter=3) = 13.772278472463457
- MAPE MT DEATH (parameter=3) = 78.57142857142857
- MAPE NC DEATH (parameter=3) = 118.04078196935339
- MAPE MT CASES (parameter=5) = 45.695593830064304
- MAPE NC CASES (parameter=5) = 19.06492766270813
- MAPE MT DEATH (parameter=5) = 140.47619047619045
- MAPE NC DEATH (parameter=5) = 112.84872534872534

  (Note:- Please see the graphs in the Python Notebook PDF file to see the graphs for AR. As the variance in data is high, the time series methods(AR & EWMA) of predicting the values might not give the best outputs.)

- **EWMA Results**

  MSE
- MSE MT CASES (alpha=0.5) = 2424.0
- MSE NC CASES (alpha=0.5) = 129229.71428571429
- MSE MT DEATH (alpha=0.5) = 5.857142857142857
- MSE NC DEATH (alpha=0.5) = 198.28571428571428
- MSE MT CASES (alpha=0.8) = 2613.8571428571427
- MSE NC CASES (alpha=0.8) = 166645.2857142857
- MSE MT DEATH (alpha=0.8) = 7.285714285714286
- MSE NC DEATH (alpha=0.8) = 206.28571428571428

  MAPE in %
- MAPE MT CASES (alpha=0.5) = 48.351725562418544
- MAPE NC CASES (alpha=0.5) = 17.894404613881967
- MAPE MT DEATH (alpha=0.5) = 119.04761904761902
- MAPE NC DEATH (alpha=0.5) = 103.358929430358
- MAPE MT CASES (alpha=0.8) = 45.971274316881846
- MAPE NC CASES (alpha=0.8) = 19.747692005707478
- MAPE MT DEATH (alpha=0.8) = 119.04761904761905
- MAPE NC DEATH (alpha=0.8) = 81.58833230261801

2B.

One Sample Wald Test for Montana Cases: **Reject**
Two Sample Wald Test for Montana Cases: **Reject**
One Sample Z Test for Montana Cases: **Accept**
One Sample T Test for Montana Cases: **Reject**
Two Sample T Test for Montana Cases: **Reject**

One Sample Wald Test for North Carolina Cases: **Reject**
Two Sample Wald Test for North Carolina Cases: **Reject**
One Sample Z Test for North Carolina Cases: **Reject**
One Sample T Test for North Carolina Cases: **Reject**
Two Sample T Test for North Carolina Cases: **Reject**

One Sample Wald Test for Montana Deaths: **Reject**
Two Sample Wald Test for Montana Deaths: **Reject**
One Sample Z Test for Montana Deaths: **Accept**
One Sample T Test for Montana Deaths: **Reject**
Two Sample T Test for Montana Deaths **Reject**

One Sample Wald Test for North Carolina Deaths: **Reject**
Two Sample Wald Test for North Carolina Deaths: **Reject**
One Sample Z Test for North Carolina Deaths: **Reject**
One Sample T Test for North Carolina Deaths: **Reject**
Two Sample T Test for North Carolina Deaths: **Reject**

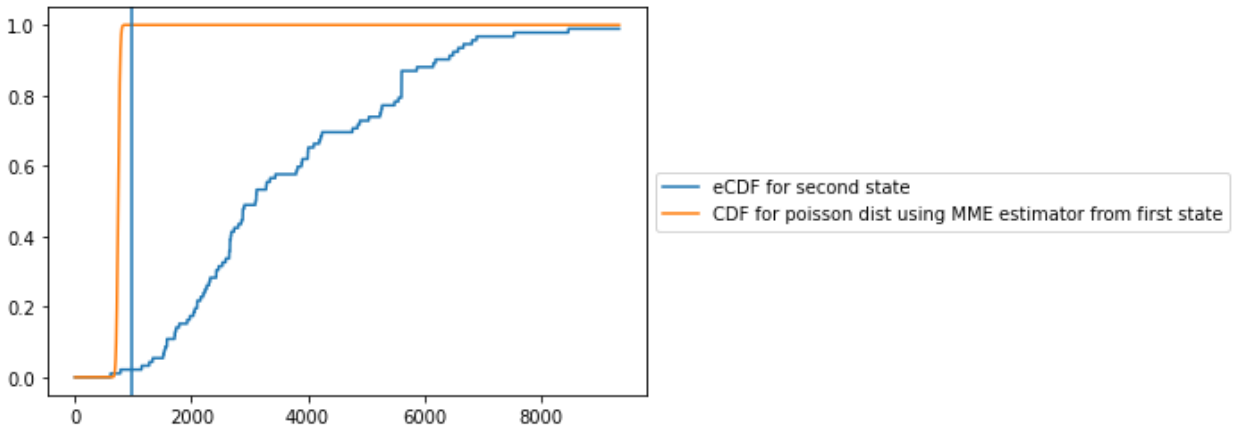
The Wald's test is applicable for our dataset since we are using Poisson MLE as our estimate, which is Asymptotically Normal which is enough for Wald's test. Z-test is also applicable since the true standard deviation is known to us and the number of samples is 31 (>30), which is our usual value for CLT to be applicable. The T-test however is not valid since the data points are expected to follow a Normal distribution but the given distribution to us is Poisson.

2C. Infer the equality of distributions in the two states using KS and Permutation test.

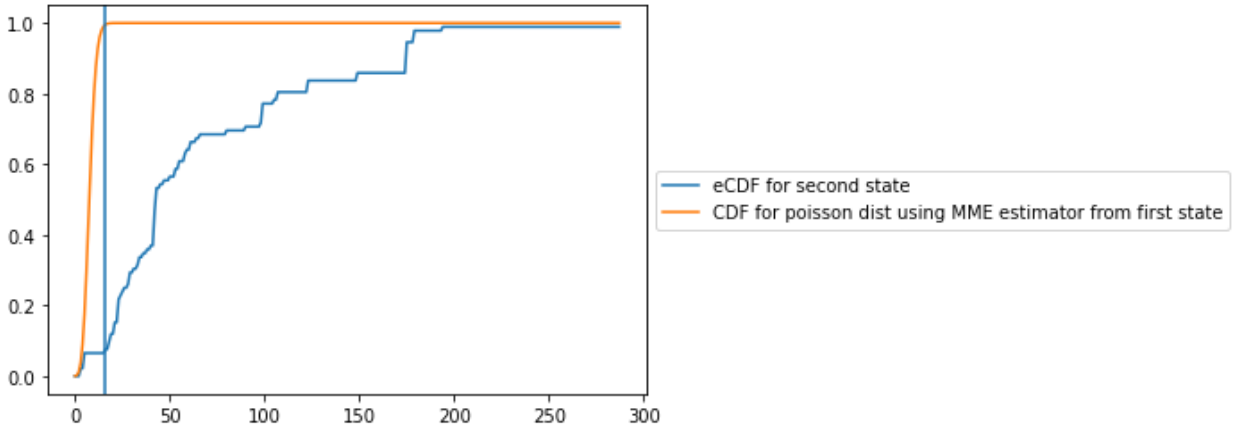
- **1 Sample KS test (H0: Fd~Fc H1:Fd!~Fc)**
  - Poisson Distribution

  Results

```
poisson_KS(MT_daily_cases, NC_daily_cases)
980
0.9782608695652174
Hypothesis Rejected
```



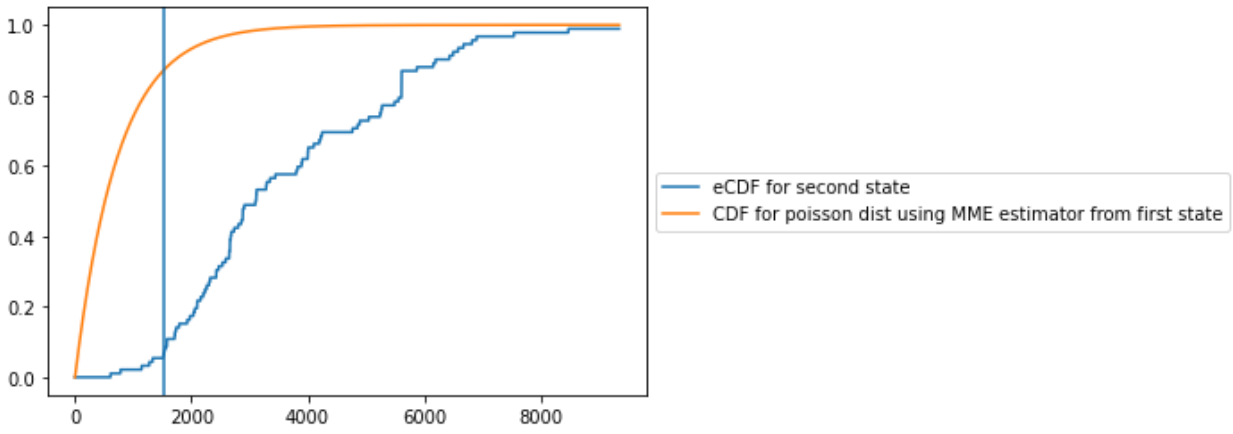

```
poisson_KS(MT_daily_death, NC_daily_death)
16
0.930208449695998
Hypothesis Rejected
```



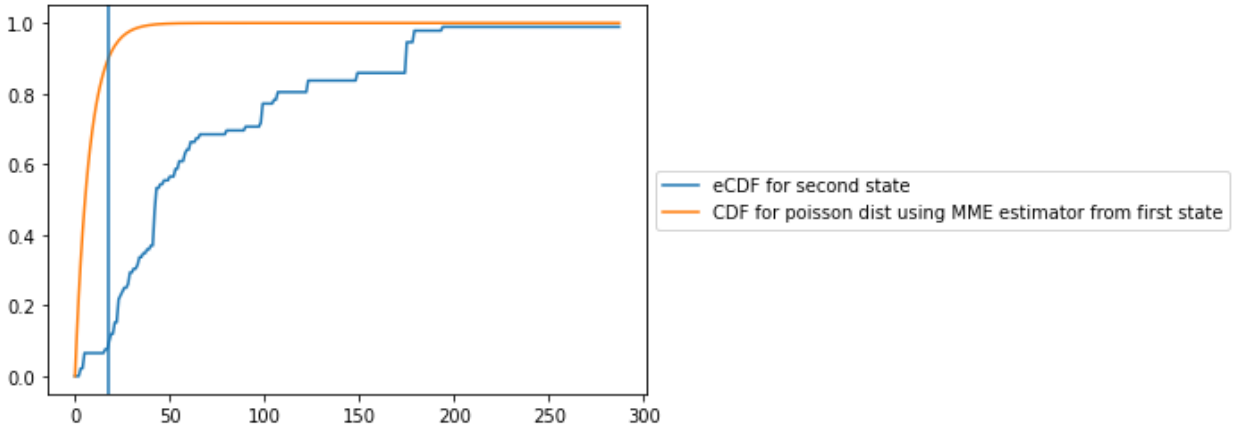
- Geometric Distribution

Results

```
geometric_KS(MT_daily_cases, NC_daily_cases)
0.8157381468179481
1515
Hypothesis Rejected
```




```
geometric_KS(MT_daily_death, NC_daily_death)
0.828442430917523
18
Hypothesis Rejected
```

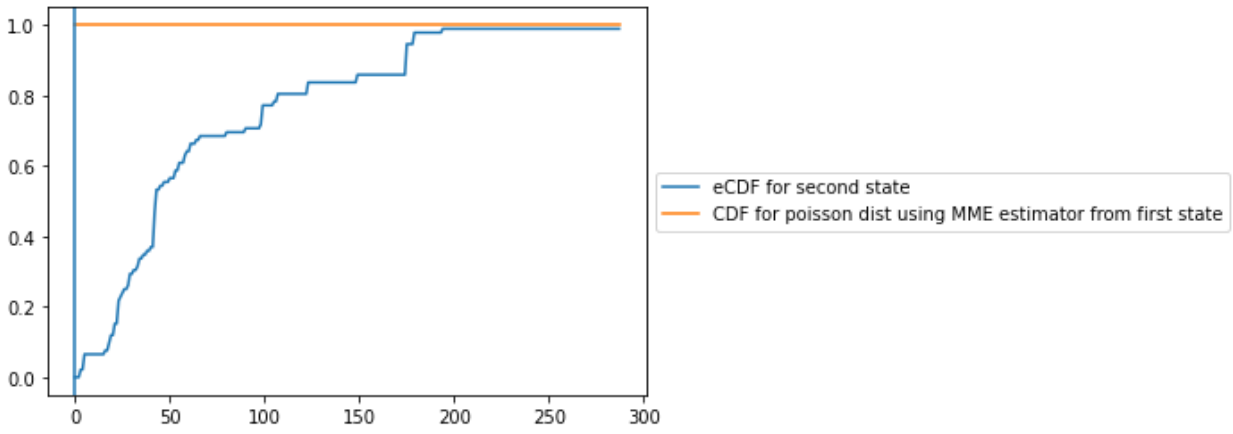


(Note:- Even though the difference in CDF and eCDF is high, the geometric distribution seems to fit the spread of COVID the best.)

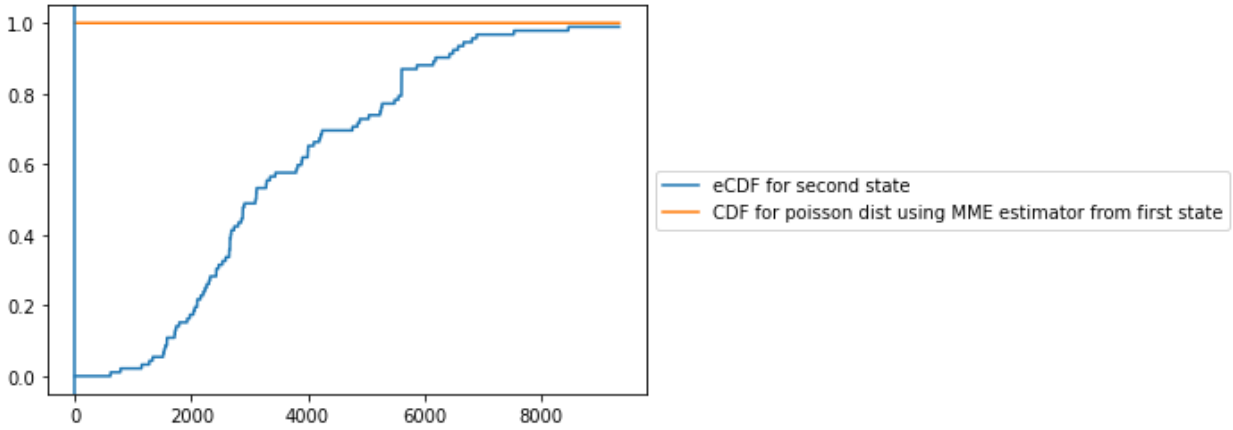
- Binomial Distribution

Results

```
binomial_KS(MT_daily_death, NC_daily_death)
1.0
0
Hypothesis Rejected
```



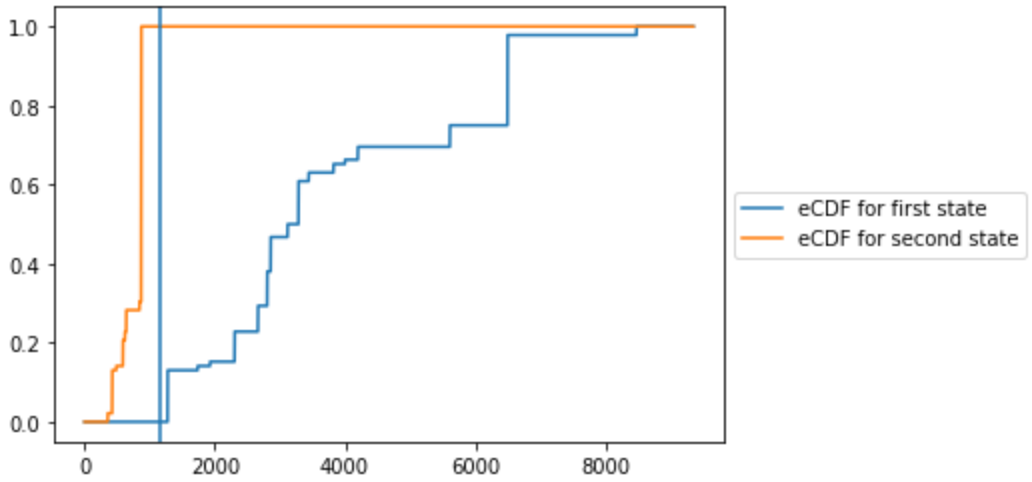

```
binomial_KS(MT_daily_cases, NC_daily_cases)
1.0
0
Hypothesis Rejected
```



(Note:- The binomial dist has negative N and P values from the dataset of the first state. This shows that the binomial distribution is not a good fit for the given data. It is further confirmed by the fact that these negative values(of N and P) produce a CDF which has a constant value of 1 in the entire domain of the second state's data.)

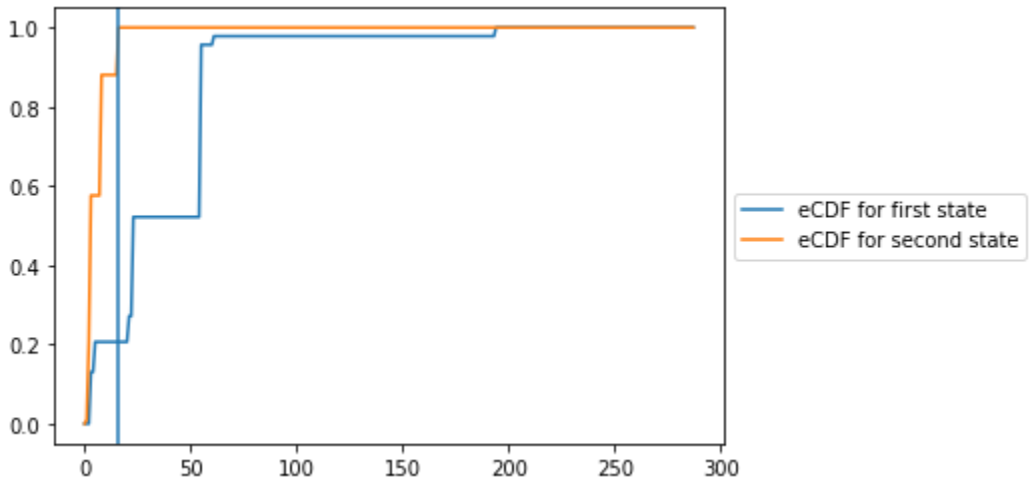
- **2 Sample KS test(H0: Fx~Fy H1:Fx!~Fy)**

```
ks_2_sample(NC_daily_cases, MT_daily_cases)
1.0
1144
Hypothesis Rejected
```



(Note:- Due to the huge difference in the population of the states, there is vast difference in the COVID cases and deaths in the same time period. The minimum number of cases in NC is more than the maximum number of cases/deaths in MT. As a result, the max difference in eCDF is 1.)

```
ks_2_sample(NC_daily_death, MT_daily_death)
0.7934782608695652
16
Hypothesis Rejected
```



- **Permutation test(H0: X~Y H1:X!~Y)**

```
permutation_test(NC_daily_cases, MT_daily_cases)
0
Hypothesis rejected
```

```
permutation_test(NC_daily_death, MT_daily_death)
0
Hypothesis rejected
```

2D. Bayesian Inference

MAP for 5th Week Posterior: 10.914917127071824
MAP for 6th Week Posterior: 13.145349696880082
MAP for 7th Week Posterior: 15.81879668887543
MAP for 8th Week Posterior: 17.58287991416967