CSE 544 - Final Project
Andrew Burford 112251367
Daniel Billmann 114715276
Efthimios Vlahos 110540896

Part 1

We used pandas to read in the csv data and store it in a pandas DataFrame object. Data was cleaned and outliers removed independently for each column of data we are interested in. i.e. the daily cases in a specific state was cleaned separately from the daily cases in a different state. There were no empty rows within the CSV data we had to worry about. While decumulatizing the rows, we noticed an issue in the cumulative deaths data within MA:

	date	cases	deaths	state
17570	2022-03-06	1675767	23577	MA
17489	2022-03-07	1677619	23614	MA
30945	2022-03-08	1678247	23679	MA
19264	2022-03-09	1679034	23708	MA
20749	2022-03-10	1679921	23732	MA
17969	2022-03-11	1680673	23751	MA
16970	2022-03-12	1680673	23751	MA
16099	2022-03-13	1680673	23751	MA
26602	2022-03-14	1682693	19981	MA
30173	2022-03-15	1683224	19997	MA
18872	2022-03-16	1684036	20012	MA
25206	2022-03-17	1685055	20023	MA
30354	2022-03-18	1685937	20029	MA
30225	2022-03-19	1685937	20029	MA
21139	2022-03-20	1685937	20029	MA
15024	2022-03-21	1687720	20041	MA
30498	2022-03-22	1688534	20056	MA
21457	2022-03-23	1689750	20058	MA
17525	2022-03-24	1690983	20063	MA
18534	2022-03-25	1692177	20072	MA
25332	2022-03-26	1692177	20072	MA

You can see that on March 14th, 2022, the cumulative total drops. This results in a negative number of deaths for this day, but it ends up not mattering because this data point is dropped after applying Tukey's rule. Here is the results of our cleaning script:

```
Cleaning cases data in state MA
Missing 0 out of 833 days of data
Applying Tukey's rule
Min Q1 Q3 Max
0 193 2102 64715
Discarding non-zero values outside of range [-2670, 4965]
Number of rows with outliers: 79
```

```
Cleaning deaths data in state MA
Missing 0 out of 833 days of data
Applying Tukey's rule
Min Q1 Q3 Max
-3770 4 39 209
Discarding non-zero values outside of range [-48, 91]
Number of rows with outliers: 54
Cleaning cases data in state MS
Missing 0 out of 833 days of data
Applying Tukey's rule
Min Q1 Q3 Max
0 156 1033 22456
Discarding non-zero values outside of range [-1159, 2348]
Number of rows with outliers: 71
Cleaning deaths data in state MS
Missing 0 out of 833 days of data
Applying Tukey's rule
Min Q1 Q3 Max
0 3 20 126
Discarding non-zero values outside of range [-22, 45]
Number of rows with outliers: 46
Missing 0 days of vaccine data in MA
Applying Tukey's rule
Min Q1 Q3 Max
0 9696 44423 172038
Discarding non-zero values outside of range [-42394, 96513]
Rows with outliers in vaccine data: 14
Missing 0 days of vaccine data in MS
Applying Tukey's rule
Min Q1 Q3 Max
0 156 10780 119816
Discarding non-zero values outside of range [-15780, 26716]
Rows with outliers in vaccine data: 32
```

It is also worth noting that since we are taking the difference between cumulative totals each day, we do not have data for the first day in the data set so we always drop this day.

Part 2

a)

```
PART A for cases data in MA
ONE SAMPLE TESTS
Data size greater than or equal to 30, using Normal distribution for T test.
State: MA | For t test:
Size of data set: 31
Null Hypothesis mean: 2040.107142857143
Critical value: 2.359562
T statistic: -3.0973755870870985, p-value 0.0019524233115133887
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.
State: MA | For walds test:
Size of data set: 31
Null Hypothesis mean: 2040.107142857143
Critical value: 2.241403
T statistic: -37.76712261170382, p-value 0.0
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.
State: MA | For z test:
Size of data set: 31
Null Hypothesis mean: 2040.107142857143
Critical value: 2.241403
T statistic: -1.3902878663767728, p-value 0.16444148087806787
This indicates we SHOULD NOT reject the null hypothesis that
the mean #cases from March is the same as in February.
TWO SAMPLE TESTS
State: MA | For two sample t test:
Size of data sets: 28, 31
Critical value: 2.302158
T statistic: 0.03328753699154547, p-value 0.9734452922932615
This indicates we SHOULD NOT reject the null hypothesis that
the mean #cases from March is the same as in February.
State: MA | For two sample walds test:
Size of data sets: 28, 31
Critical value: 1.959964
T statistic: 24.977051287155554, p-value 1.0857455175684608e-137
```

```
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.

PART A for deaths data in MA
ONE SAMPLE TESTS
Data size greater than or equal to 30, using Normal distribution for T test.
State: MA | For t test:
Size of data set: 31
Null Hypothesis mean: 49.535714285714285
Critical value: 2.359562
T statistic: -12.297395933400132, p-value 9.3541424834056e-35
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
State: MA | For walds test:
Size of data set: 31
Null Hypothesis mean: 49.535714285714285
Critical value: 2.241403
T statistic: -15.987723214285715, p-value 1.5561523392868e-57
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
State: MA | For z test:
Size of data set: 31
Null Hypothesis mean: 49.535714285714285
Critical value: 2.241403
T statistic: -4.177116507987537, p-value 2.9522782593876546e-05
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
TWO SAMPLE TESTS
State: MA | For two sample t test:
Size of data sets: 28, 31
Critical value: 2.302158
T statistic: 4.2790551525500495, p-value 1.87688355457769e-05
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
State: MA | For two sample walds test:
Size of data sets: 28, 31
Critical value: 1.959964
T statistic: 9.802168601663972, p-value 1.1019402838883964e-22
```

```
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.

PART A for cases data in MS
ONE SAMPLE TESTS
Data size greater than or equal to 30, using Normal distribution for T test.
State: MS | For t test:
Size of data set: 31
Null Hypothesis mean: 664.9642857142857
Critical value: 2.359562
T statistic: -13.857382008534747, p-value 1.147849292650201e-43
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.
State: MS | For walds test:
Size of data set: 31
Null Hypothesis mean: 664.9642857142857
Critical value: 2.241403
T statistic: -110.62354526664869, p-value 0.0
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.
State: MS | For z test:
Size of data set: 31
Null Hypothesis mean: 664.9642857142857
Critical value: 2.241403
T statistic: -3.4539387639251684, p-value 0.0005524628056236125
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.
TWO SAMPLE TESTS
State: MS | For two sample t test:
Size of data sets: 28, 31
Critical value: 2.302158
T statistic: 0.505333368577627, p-value 0.6133247096873808
This indicates we SHOULD NOT reject the null hypothesis that
the mean #cases from March is the same as in February.
State: MS | For two sample walds test:
Size of data sets: 28, 31
Critical value: 1.959964
T statistic: 60.440726978860944, p-value 0.0
```

```
This indicates we SHOULD reject the null hypothesis that
the mean #cases from March is the same as in February.

PART A for deaths data in MS
ONE SAMPLE TESTS
Data size greater than or equal to 30, using Normal distribution for T test.
State: MS | For t test:
Size of data set: 31
Null Hypothesis mean: 11.75
Critical value: 2.359562
T statistic: -7.89970667147348, p-value 2.795605397300487e-15
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
State: MS | For walds test:
Size of data set: 31
Null Hypothesis mean: 11.75
Critical value: 2.241403
T statistic: -13.960027707547919, p-value 2.733275690513804e-44
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
State: MS | For z test:
Size of data set: 31
Null Hypothesis mean: 11.75
Critical value: 2.241403
T statistic: -3.0202524625465066, p-value 0.002525640644869188
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
TWO SAMPLE TESTS
State: MS | For two sample t test:
Size of data sets: 28, 31
Critical value: 2.302158
T statistic: 3.4883231839318523, p-value 0.00048606018299689836
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
State: MS | For two sample walds test:
Size of data sets: 28, 31
Critical value: 1.959964
T statistic: 7.725112129816948, p-value 1.1175492509153649e-14
```

```
This indicates we SHOULD reject the null hypothesis that
the mean #deaths from March is the same as in February.
```

With regard to the validity of these tests, the one sample t-tests are not valid for testing the hypothesis that the mean # of cases or deaths in February 2021 is different from the mean in March 2021 since they ignore the variance in the data in February. They could only test the null hypothesis that the mean # of cases/deaths in March 2021 is equal to a specific value. For the 1 sample Wald's test, the test is just about applicable to test this other hypothesis since we are using an asymptotically normal estimator and our data set size is either very close to or just above 30, when asymptotic normality kicks in. For 1 sample t-test, again this test is applicable to the other hypothesis because n is just about equal to 30 so it does not matter whether the underlying data is normally distributed. For the 1 sample z-test, we are using the much larger data set of #cases/deaths for the entire date range in the data set, so this is a good way of getting an accurate estimate of population standard deviation. This means the z-test is valid for testing this other hypothesis as well. The 2 sample t-test is also essentially valid since 28 is very close to 30 for the February sample even if the underlying data is not normal, so the variance of the difference in means will be approximately normal. The 2 sample Wald's test is valid for the same reason. Both 2 sample tests apply to testing the null hypothesis of equality of means between the two months.

b)

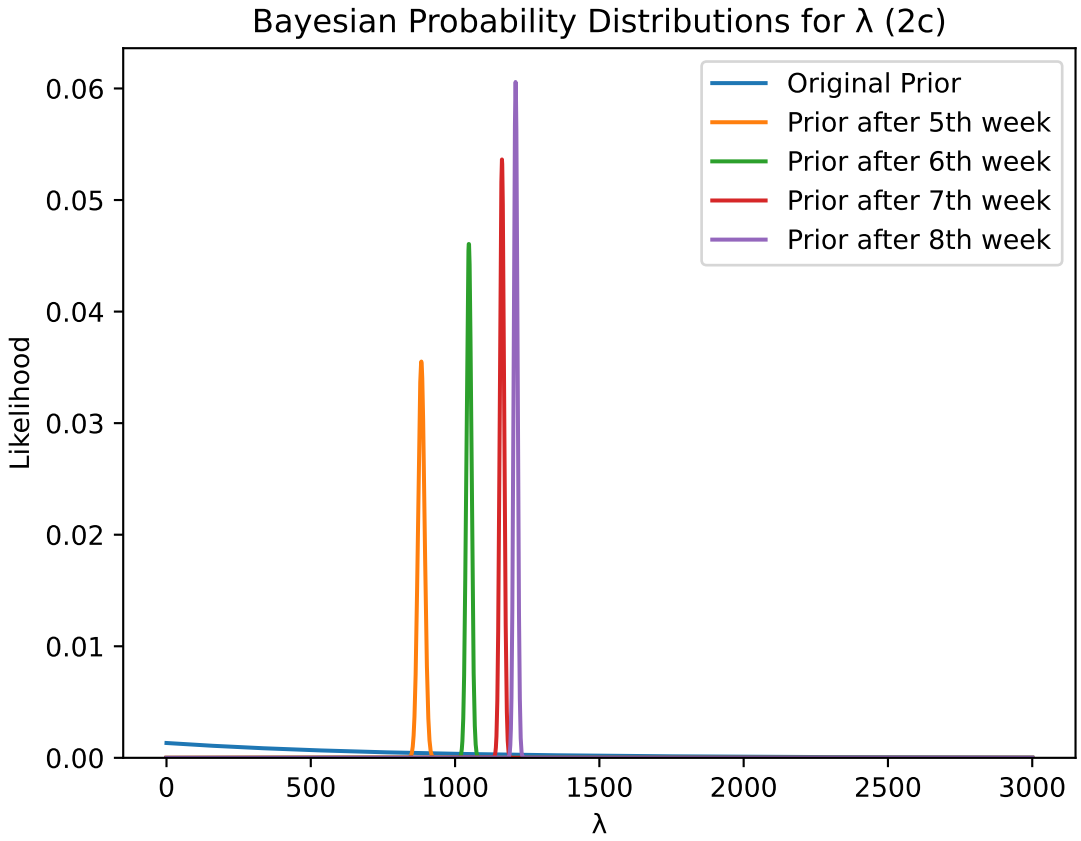
```
PART B
Analyzing cases data
1 Sample KS-test with Poisson distribution
K-S Statistic: 0.8761
p-val: 0.0000
We reject the null hypothesis that the distribution of daily #
of cases in MA and MS is equal
1 Sample KS-test with Geometric distribution
K-S Statistic: 0.3297
p-val: 0.0000
We reject the null hypothesis that the distribution of daily #
of cases in MA and MS is equal
```

```
MME for binomial distribution returned negative values
2 Sample KS-test
K-S Statistic: 0.1416
p-val: 0.2257
We fail to reject the null hypothesis that the distribution of daily #
of cases in MA and MS is equal
Permutation Test
p-val: 0.000000
We reject the null hypothesis that the distribution of daily #
of cases in MA and MS is equal
Analyzing deaths data
1 Sample KS-test with Poisson distribution
K-S Statistic: 0.5629
p-val: 0.0000
We reject the null hypothesis that the distribution of daily #
of deaths in MA and MS is equal
1 Sample KS-test with Geometric distribution
K-S Statistic: 0.2523
p-val: 0.0000
We reject the null hypothesis that the distribution of daily #
of deaths in MA and MS is equal
MME for binomial distribution returned negative values
2 Sample KS-test
K-S Statistic: 0.0984
p-val: 0.5940
We fail to reject the null hypothesis that the distribution of daily #
of deaths in MA and MS is equal
Permutation Test
p-val: 0.002000
We reject the null hypothesis that the distribution of daily #
of deaths in MA and MS is equal
```

c)

```
PART C
MAP for posterior after 5th week: 883
MAP for posterior after 6th week: 1047
```

```
MAP for posterior after 7th week: 1162
MAP for posterior after 8th week: 1209
```



d)

```
PART D for MA
Actual #vaccines administered during fourth week of May 2021:
90158 50738 35117 24316 41566 54022 58589
Predicted #vaccines administered during fourth week of May 2021 using AR(3):
56685 56887 61386 64270 63909 62220 61287
MAPE: 51.70%
MSE: 574056116.32
```

```
Predicted #vaccines administered during fourth week of May 2021 using AR(5):
59194 61139 55245 55830 60728 60225 59630
MAPE: 43.02%
MSE: 410284182.75
Predicted #vaccines administered during fourth week of May 2021 using EWMA(0.5):
61464 61464 61464 61464 61464 61464 61464
MAPE: 49.62%
MSE: 496006660.41
Predicted #vaccines administered during fourth week of May 2021 using EWMA(0.8):
65354 65354 65354 65354 65354 65354 65354
MAPE: 57.28%
MSE: 595320389.44

PART D for MS
Actual #vaccines administered during fourth week of May 2021:
327 305 28 22356 5116 195 12328
Predicted #vaccines administered during fourth week of May 2021 using AR(3):
7079 7507 7287 7167 7305 7314 7249
MAPE: 4879.32%
MSE: 66018795.69
Predicted #vaccines administered during fourth week of May 2021 using AR(5):
11346 1060 8452 6969 6549 3931 9322
MAPE: 5105.72%
MSE: 64963112.07
Predicted #vaccines administered during fourth week of May 2021 using EWMA(0.5):
7429 7429 7429 7429 7429 7429 7429
MAPE: 4971.29%
MSE: 65778635.61
Predicted #vaccines administered during fourth week of May 2021 using EWMA(0.8):
7105 7105 7105 7105 7105 7105 7105
MAPE: 4753.03%
MSE: 64835068.29
```

e)

Some of the days in this date range had outliers in one state but not the other. Since this is a paired t-test, we throw out any days with an outlier in either state. This results in less than 30 days of data.

```
PART E
Paired T-test for comparing the number of vaccines administered
each day in MA and MS during 2022-9:
Using 27 out of 30 days (missing days are outliers)
t = 2.2342
p-val = 0.034282
Paired T-test for comparing the number of vaccines administered
each day in MA and MS during 2022-11:
Using 26 out of 30 days (missing days are outliers)
t = 6.5531
p-val = 0.000001
```

In both cases, we reject the null hypothesis under alpha = 0.5 in favor of the alternative hypothesis that the number of vaccines administered each day is different in MA than in MS. The positive t statistics tell us that on average MA distributed more vaccines each day. The largest reason for this is likely that there are simply more people living in MA than in MS. MA has a population of about 6.9 million while MS is about 2.9 million.