CSE 544 Probability & Statistics Project

Team Members:
Yasharth Sharma (114362612)
Ayush Dayani (114241407)
Shubham Jindal (113772070)
Manikanta Sathwik Yeluri (113785603)

Mandatory Tasks:

Data Set:

Covid Cases and Deaths
Vaccination
X-dataset - NYC Green Cab Data for Year 2020 and 2021

Inference/Observation:

1) Dataset Cleaning and Preprocessing -

Step1 - After loading the dataset convert the date columns to datetime objects, filter the dataset according to the assigned states and sort the dataset in ascending order for preview and further preprocessing.
Step2 - Since the dataset was provided in an cumulative manner, we had to reduce it to everyday data for each date for both vaccination as well as cases dataset. There were no NaN values in the dataset.
Step3 - Checking for Outlier using Tukey's rule. The number of outliers found was 269 rows and after cleaning the final dataset size was 571 rows for the cases dataset and 457 rows for the vaccination dataset.

2) Tasks -

Task A -

Null Hypothesis is that mean of Covid 19 deaths and cases are same for the month of february and march

Wald's Test (1 Sample)

Hawaii Feb Cases vs Hawaii March Cases
W-val = 10.722503138986319
reject null hypothesis
Iowa Feb Cases vs Iowa March Cases
W - val = 23.70761970984968
reject null hypothesis
Hawaii Feb Deaths vs Hawaii March Deaths
W - val = 1.1681075300867827
accept null hypothesis

Iowa Feb Deaths vs Iowa March Deaths

W - val = 5.498717658953222

reject null hypothesis

T-Test (1 Sample)

Hawaii feb cases vs Hawaii March Cases

2.2868273523121077

reject null hypothesis

Iowa feb cases vs Iowa March Cases

3.409790776514401

reject null hypothesis

Hawaii feb deaths vs Hawaii March deaths

1.1612159800881006

accept null hypothesis

Iowa feb deaths vs Iowa March deaths

2.442096164544151

reject null hypothesis

Z-Test (1 Sample)

Hawaii Feb Cases vs Hawaii March Cases

Z-val = 5.673777858813078

reject null hypothesis

Iowa Feb Cases vs Iowa March Cases

Z - val = 4.729779904419943

reject null hypothesis

Hawaii Feb Deaths vs Hawaii March Deaths

Z - val = 3.9252309873283546

reject null hypothesis

Iowa Feb Deaths vs Iowa March Deaths

Z - val = 11.34859953130403

reject null hypothesis

Wald’s Test (2 Sample)

Hawaii Feb Cases vs Hawaii March Cases

W-val = 7.119049455593124
reject null hypothesis

Iowa Feb Cases vs Iowa March Cases

W - val = 13.708691681336964
reject null hypothesis

Hawaii Feb Deaths vs Hawaii March Deaths

W - val = 0.667011318708645
accept null hypothesis

Iowa Feb Deaths vs Iowa March Deaths

W - val = 3.020212097633475
reject null hypothesis

T-Test (2 Sample)

Hawaii feb cases vs Hawaii March Cases

T val = 1.8078276860478544
accept null hypothesis

Iowa feb cases vs Iowa March Cases

T val = 2.019654741334597
accept null hypothesis

Hawaii feb deaths vs Hawaii March deaths

T val = 0.605871042839492
accept null hypothesis

Iowa feb deaths vs Iowa March deaths

T val = 1.1234241936106828
accept null hypothesis

Test Applicable or not?

Wald's Test - These tests are not applicable here due to the fact that the data is not asymptotically normal estimators and the data size is small hence according to this test is not applicable for neither 1 sample nor 2 sample.
T-Test - Since the data is not normally distributed and also not large enough and therefore the test is not appropriate.
Z-Test - THis test is not applicable since the dataset is not large enough and also not

Task B -

KS - Test -

1-Sample :-

Poisson Distribution -

Cases

Null Hypothesis is that there is equality of distributions between Hawaii cases and Iowa cases
Poisson parameter: 134.36111111111111
rejecting null hypothesis
KS statistic : 0.5277777777777777

Deaths

Null Hypothesis is that there is equality of distributions between Hawaii deaths and Iowa deaths
Poisson parameter: 2.4722222222222223
rejecting null hypothesis
KS statistic : 0.9156028986548875

Binomial Distribution-

Cases

Null Hypothesis is that there is equality of distributions between Hawaii cases and Iowa cases
Geometric parameter : 0.007442629729170974
rejecting null hypothesis
KS statistic : 0.4722222222222224

Deaths

Null Hypothesis is that there is equality of distributions between Hawaii deaths and Iowa deaths
Geometric parameter : 0.4044943820224719
rejecting null hypothesis
KS statistic : 1.0000000000000002

Geometric Distribution -

Cases

```
Null Hypothesis is that there is equality of distributions between Hawaii cases and Iowa cases
Binomial parameters(n,p) :  -34.92026164059449 -3.847654765418983
rejecting null hypothesis
KS statistic :  1.0
```

Deaths

```
Null Hypothesis is that there is equality of distributions between Hawaii deaths and Iowa deaths
Binomial parameters(n,p) :  -0.3367665418227215 -7.341056533827621
rejecting null hypothesis
KS statistic :  1.0
```

2-Sample -

Cases -

```
Null Hypothesis is that there is equality of distributions between Hawaii cases and Iowa cases
Reject Null Hypothesis
KS statistic :  0.5000000000000002
```

Deaths -

```
Null Hypothesis is that there is equality of distributions between Hawaii deaths and Iowa deaths
Reject Null Hypothesis
KS statistic :  0.7222222222222225
```

Permutation Test:

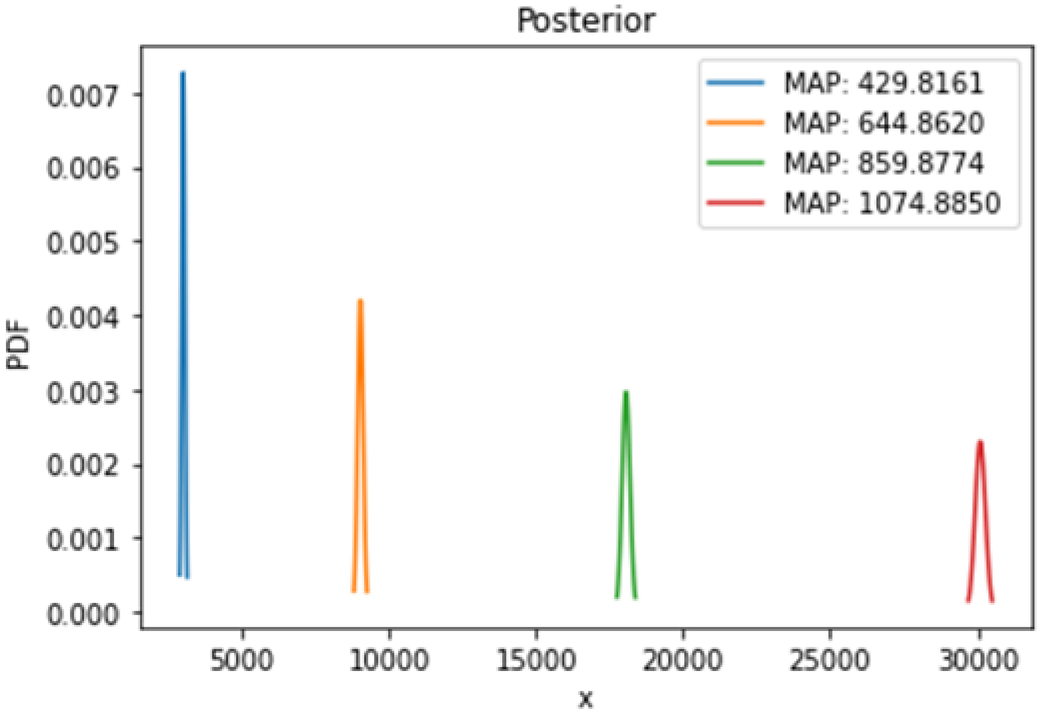
Cases -

```
Null Hypothesis is that both Hawaii_cases and Iowa_cases have same distribution
T_observed is :  530.5555555555555
p statistic :  0.0
reject the null hypothesis
```

Deaths -

```
Null Hypothesis is that both Hawaii_deaths and Iowa_deaths have same distribution
T_observed is :  2.4722222222222223
p statistic :  0.0
reject the null hypothesis
```

Task C



Since the MAP value is increasing, hence the number of total (deaths+cases) are also increasing.

This trend says that as the time progresses, total stats(deaths+cases) will increase, but we can’t say this for sure since the time frame for data is limited.

Task D

MAPE				
States	AR(3)	AR(5)	EWMA (alpha 0.5)	EWMA (alpha 0.8)
Hawaii	290.75	362.24	341.90	321.68
Iowa	59.48	59.96	40.92	38.20

MSE				
States	AR(3)	AR(5)	EWMA (alpha 0.5)	EWMA (alpha 0.8)
Hawaii	30489191.75	38747520.59	30578548.38	40305170.73
Iowa	33538503.47	65312647.60	19095522.14	15420840.32

Task E - Paired T-test

null hypothesis is that means of the #vaccines administered between the two states for the months of september and november are same

September 2021
T-val = 2.045 (alpha = 0.05)
t value for september is: 2.2113837108049004
reject null hypothesis

November 2021
T-val = 2.069 (alpha = 0.05)
t value for november is: 2.697930261674824
reject null hypothesis

Exploratory Tasks

Dataset Used - NYC Green Cab Data for Year 2020 and Year 2021

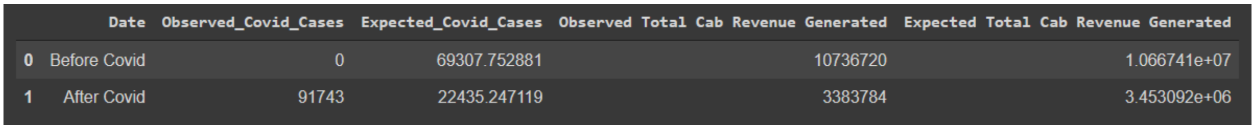
Inference 1 - The cab revenue generated was not affected due to rise in covid cases (null hypothesis). Our assumption is that cab fare shouldn't have been affected that much due to covid cases, because even if there were less people preferring cabs, the cost won't shoot up because of a steady demand and supply of cab transport.

Dataset Used - Combined Cab data (2020 & 2021) split into two parts -

1) before covid (2020-03-04) [42 rows available in cases dataset]
2) After covid (2020-03-04) [42 rows taken due to before dataset row limitation]

Test Done - Chi Square Test to determine whether the two distributions are same for cab revenue before and after covid.

Results -



	Date	Observed_Covid_Cases	Expected_Covid_Cases	Observed Total Cab Revenue Generated	Expected Total Cab Revenue Generated
0	Before Covid	0	69307.752881	10736720	1.066741e+07
1	After Covid	91743	22435.247119	3383784	3.453092e+06

Q_expected : 285257.08942954347

Degrees of freedom : 1

Using the P-value calculator and significance value of 0.05, we get The P-Value is < .00001. The result is significant at p < .05. Hence we reject the Null Hypothesis (which says that cab revenue generation was not affected due to rise in covid cases).

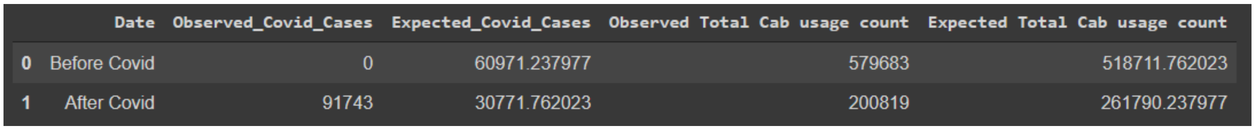
Hence we can clearly see from the above inference that cab revenue were affected due to rise in covid cases. This could be because there was a sudden shortage of demand for the cabs.

Inference 2 - The Null hypothesis is that cab usage was not affected due to rise in covid cases. It is our null hypothesis that cab usage should not be affected as 1st case of covid 19 is found. People need a cab especially in NYC to commute to hospitals to check themselves. Also, Initially lockdown didn't happen so the cab frequency shouldn't get affected so much.

Dataset Used - Combined Cab data (2020 & 2021) split into two parts -

1) before covid (2020-03-04) [42 rows available in cases dataset]
2) After covid (2020-03-04) [42 rows taken due to before dataset row limitation]

Test Done - Chi Square Test to determine whether the two distributions are same for cab revenue before and after covid.



	Date	Observed_Covid_Cases	Expected_Covid_Cases	Observed Total Cab usage count	Expected Total Cab usage count
0	Before Covid	0	60971.237977	579683	518711.762023
1	After Covid	91743	30771.762023	200819	261790.237977

Results -
Q_expected : 203146.83190642967
Degrees of freedom : 1

Using the P-value calculator and significance value of 0.05, we get The P-Value is < .00001. The result is significant at p < .05. Hence we reject the Null Hypothesis (which says that cab usage was not affected due to rise in covid cases).
Hence we can clearly see from the above inference that cab usage was affected due to rise in covid cases. Reason could be that social distancing was enforced and people maybe didn’t like the idea of traveling via public transport vehicle.

Inference 3 - Null Hypothesis - vaccination rate and cab usage are directly proportional.
Taking threshold value as 0.5.
Vaccination Started after 2020-12-14 so taking data after that only.

By logic, we can assume that as the vaccination rate goes up in the state of New york, the cab usage should also increase. Hence it looks like the relation between vaccination and cab usage is directly proportional hence it is our null hypothesis.

Dataset Used - Combined Cab data (2020 & 2021) right join with Vaccination dataset (327 rows)

Test Done - Pearson Correlation Coefficient.

Results -
Pearson Correlation Coefficient Value is: -0.24

Since the Pearson correlation coefficient value is < than the threshold, we'll accept the null hypothesis and hence higher the vaccination, higher the cab usage.