Dataset Cleaning

	CO Cases	CO Deaths	CO Vaccines	AR Cases	AR Deaths	AR Vaccines
Lower Bound	-1971.75	-22.0	-26358.25	-1224.125	-25.0	-11451.0
Upper Bound	4264.25	42.0	64007.75	2520.875	47.0	25725.0
# of Outliers	64	74	17	73	22	25

Notes: The dataset had no missing values for our target columns, which made data cleaning easier.

Covid Dataset

a) How did average stats change
   i) Wald's Test:

Wald's Test	1 Sample CO Cases	1 Sample CO Deaths	1 Sample AR Cases	1 Sample AR Deaths	2 Sample CO Cases	2 Sample CO Deaths	2 Sample AR Cases	2 Sample AR Deaths
Statistic Value	16.215	16.007	218.123	12.638	10.914	8.414	98.028	7.435
Accept / Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject	Reject

Z-Test	1 Sample CO Cases	1 Sample CO Deaths	1 Sample AR Cases	1 Sample AR Deaths	2 Sample CO Cases	2 Sample CO Deaths	2 Sample AR Cases	2 Sample AR Deaths
Statistic Value	0.567	3.562	5.661	3.777	0.362	2.430	3.863	2.577
Accept / Reject	Accept	Reject	Reject	Reject	Accept	Reject	Reject	Reject

T-Test	1 Sample CO Cases	1 Sample CO Deaths	1 Sample AR Cases	1 Sample AR Deaths	2 Sample CO Cases	2 Sample CO Deaths	2 Sample AR Cases	2 Sample AR Deaths
Statistic Value	1.532	8.188	26.188	6.272	1.116	4.027	5.129	2.902
Accept / Reject	Accept	Reject	Reject	Reject	Accept	Reject	Reject	Reject

Wald's Test is applicable. Wald's Test requires asymptotically normal parameters, and since we're using MLE, that condition holds.

For Z-Test, the main assumption is either the sample data is normally distributed or the size of the sample data is large, i.e. greater than or equal to 30. Since we are testing the mean of daily stats from February and March, we have size of sample data greater than 30, Z-test is applicable.

T-test is not applicable. T-test requires our data to follow a normal distribution and our data does not necessarily follow it.

The 2-Sample Wald's Test requires that the assumptions of asymptotically normal parameters holds for both samples, so since we're continuing to use MLE, the condition holds, so 2-Sample Wald's Test is applicable

Similarly, since our sample data isn't normally distributed, the assumptions of 2-Sample T Test won't be met, so 2-Sample T Test is not applicable.

b) Equality of distributions

i) K-S One Sample test

K-S One Sample	Poisson w/ Cases	Geometric w/ Cases	Binomial w/ Cases	Poisson w/ Deaths	Geometric w/ Deaths	Binomial w/ Deaths
Maximum distance	0.989	0.584	0.989	0.321	0.194	0.989
Accept / Reject	Reject	Reject	Reject	Reject	Reject	Reject

ii) K-S two sample test

Cases:

The max d = 0.9409427240339818
Reject Null Hypothesis.

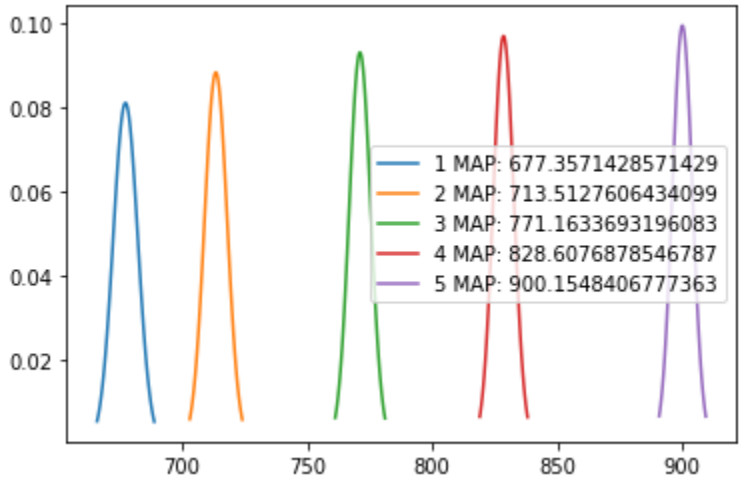
<u>Deaths:</u>
The max d = 0.3099885189437429
Reject Null Hypothesis.

iii) Permutation test:

<u>Cases:</u>
p-value for permutation test: 0.0
When n = 1000
Reject Null Hypothesis

<u>Deaths:</u>
p-value for permutation test: 0.016
When n = 1000
Reject Null Hypothesis

c) Bayesian inference



d) Vaccine predictions

1) Colorado:

Real vaccination data: [1933. 32323. 17252. 25360. 30459. 26579. 31154.]

AR3:

Predict values: [7146.0814922 39046.15952908 6782.25598024 28389.72870152 25268.48172531 23467.75185594 30188.19484083]

MAPE: 56.42%

SSE: 32675144.731596123

AR5:

Predict values: [5522.96513438 34429.93115638 17303.00179923 23663.51354647 24157.9185137 23440.90686539 26885.98705273]

MAPE: 35.06%

SSE: 12567838.138706962

EWMA(0.5):

Predict values: [1933.0, 1933.0, 17128.0, 17190.0, 21275.0, 25867.0, 26223.0]

MAPE: 25.09%

SSE: 157069133.85714287

EWMA(0.8):

Predict value of y: [1933.0, 1933.0, 26245.0, 19050.6, 24098.12, 29186.824, 27100.5648]

MAPE: 30.68%

SSE: 158275222.09571072

2) Arkansas

Real vaccination data: [12951. 9433. 4516. 2394. 9665. 8909. 8115.]

AR3:

Predict values: [11945.42896375 8339.68423668 5611.70697054 4425.88947015 10994.38901208 5247.93658972 8181.95357582]

MAPE: 26.31%

SSE: 3244400.610330465

AR5:

Predict values: [9892.80038557 6142.92502592 5331.03411388 4970.89678906 14074.58598089 3967.05459271 9822.63545348]

MAPE: 43.76%

SSE: 10609306.808366807

EWMA(0.5):
Predict values: [12951.0, 12951.0, 11192.0, 7854.0, 5124.0, 7394.5, 8151.75]
MAPE: 68.23%
SSE: 15667520.25892857

EWMA(0.8)
Predict values: [12951.0, 12951.0, 10136.6, 5640.12, 3043.2239999999997, 8340.6448, 8795.32896]
MAPE: 54.37%
SSE: 14162650.847965388

e) Equality of vaccine means

Paired T-test for number of vaccines administered daily in CO and AR for Sep.
Reject H0 with result of T-Test: 5.2594134699214266

Paired T-test for number of vaccines administered daily in CO and AR for Nov.
Reject H0 with result of T-Test: 6.664397605626252

From these two T-tests, we observe that the average daily vaccinations between Arkansas and Colorado do not seem to correlate. This holds for both September and November. This could potentially be a result of the population disparity between the two states, as Colorado has a population around 5.7 million, as compared to Arkansas's approximately 3 million. It could also imply that residents of Colorado and Arkansas had differing views on vaccination.

Exploratory Dataset

For our exploratory dataset, we have chosen a dataset of median house listing prices provided by Zillow.

1) Inference 1 - Correlation between Covid Stats and Listing Price:

Due to the nature of COVID, people were generally discouraged from making major lifestyle changes, such as buying a house, or traveling at all during periods COVID was rampant. We believe that as a result, changes in the amount of cases of COVID should have some sort of impact on the price of houses. Houses are priced based on the demand of people wanting to buy houses, so a large amount of COVID cases may affect the demand, and thus, the price of houses. We decide to test this using our two given states of Colorado and Arkansas and seeing if, over the course of the time period of October 2020 to May 2021 (since this gives a good snapshot of when COVID was most prevalent), there is a relationship between the weekly numbers of COVID cases or deaths and the weekly median prices of houses. Our hypotheses are as follows:

H_0: Confirmed weekly Covid-19 cases and deaths and house listing price are not correlated

H_1: Confirmed weekly Covid-19 cases and deaths and house listing price are correlated

We analyze the correlation using Pearson's correlation test, since this tells us if there is a relationship between the two values. We use a timeframe of October 2020 to May 2021.

Pearson Test	Colorado Cases	Colorado Deaths	Arkansas Cases	Arkansas Deaths
$\rho_{x,y}$ value	0.305	0.160	0.717	0.776
Accept / Reject H_0	Accept	Accept	Reject	Reject

It is applicable to apply Pearson's Correlation test. We used the plug-in estimator to perform the test. Based on the results from the Pearson's Correlation test, we conclude that median house listing data is positively linearly correlated with the weekly cases and deaths of Covid-19 in Arkansas. But for Colorado, it is not the same case. Some potential reasons could include the population and location of the state, which will need further investigation.

2) Inference 2 - K-S Test on the Listing Price:

In this inference, we check if the listing dataset and the Covid statistics are pulled from the same distribution; if this were the case, then there should be some strong relationship between the two metrics, and we'd also know exactly what distribution the two datasets follow, making calculating other statistics far more reliable. To do this, we perform a 2-sample K-S test. Because the home listing prices and the covid statistics have different scales, we scale the listing data by a constant to match the scale of the covid statistic data.

H_0: The distribution of listing prices matches the selected distribution

H_1: The distribution of listing prices does not match the selected distribution

For cases we divide the listing data array by 100, for vaccinations we divide by 1,000, and for deaths we divide by 10,000. This preserves the overall distribution of the listing data while allowing for comparison, since large differences in values will drastically impact the results of a K-S test.

2 Sample K-S Test	Colorado Cases	Colorado Deaths	Colorado Vaccinations	Arkansas Cases	Arkansas Deaths	Arkansas Vaccinations
Maximum Distance	0.829	0.486	0.958	0.571	0.857	0.917
Accept / Reject H_0	Reject	Reject	Reject	Reject	Reject	Reject

Conclusion: None of the covid statistic distributions matched up to the house listing distributions, which means that we can't conclude if the data are pulled from the same distribution or what that distribution even is. The distributions for deaths in Colorado and cases in Arkansas match slightly better than the other tests, but not enough for any meaningful conclusions. To interpret these results, it is useful to analyze the impact of the scaled listings dataset. By scaling, we ignore the least significant digits of the listing prices.

3) Inference 3 - χ^2 test for independence of datasets

For this inference, we will be testing the dependence of the listings datasets to various other Covid stats datasets. Just as with Inference 1, what we're looking to show here is that the prices of houses in a state aren't detached from the number of cases a state sees, since that will again show us that COVID cases and vaccinations may have had some impact on housing prices. For each of the tests, the null and alternative hypotheses will be as follows:

H_0 - Listing price is independent of this dataset

H_1 - Listing price is dependent on this dataset

	CO Cases	CO Deaths	CO Vaccinations	AR Cases	AR Deaths	AR Vaccinations
χ^2 Score	76253.32	596.96	682785.76	102281.44	1396.45	397899.05
Critical Value	48.60	48.60	35.17	48.60	48.60	35.17
P Value	0.0	0.0	0.0	0.0	0.0	0.0
Accept / Reject H_0	Reject	Reject	Reject	Reject	Reject	Reject

For all of our datasets, we have observed a p-value of < 0.05. Therefore, we reject the null hypothesis for all of our tests. This indicates that in all cases, the datasets are not independent. This is to be expected; Each of the datasets represents significant aspects of the Covid pandemic. Homeowners looking to sell their homes may ask for more or less depending on the level of uncertainty about the market, and surging/declining Covid numbers would be sure to impact such uncertainty.