**Due date**: May 14th, 5:00pm

**Submission**: Submit a Github link in the [Google Sheet](https://docs.google.com/spreadsheets/u/2/d/11bu4D9h0ZS_PnH3h6thMMUxyFG05DfcF7pxVzG1dC9A/edit). Make your repo private, visible only to your group members. After the submission deadline, do not make any commits to the repo. Otherwise, you will lose points. Only AFTER the deadline, you will be asked to make the repo public so that TAs can grade your project.The github repo should include (i) all your py code with self-explanatory filenames and required comments; (ii) pdf of required answers, plots, and results of hypothesis (TAs are not required to run your code so all results should be in pdf); and (iii) your X dataset.

**Project weightage**: 10% of final grade

**Groups**: This is a group project. Groups can be the same as assignment groups or new groups can be formed, but with max size 6.

**COVID19 datasets:** <https://github.com/michaelofsbu/CSE-544-Datasets>

Will be assigned to each group by TAs.

**Project**: The project has two parts, a *mandatory hypothesis* part worth 70% of the project grade, and an *exploratory part* worth 30% of the project grade. For all code (only Python), please implement the logic yourself, as you have been doing in the assignments. Libraries for plotting, etc., can be used, as in assignments. You can reuse any code you developed for your assignments. All the tools/tests taught in class must be coded by you. For example, you must code the Permutation test as opposed to using an in-built python perm test. You can, of course, use helper libraries for lists, generating permutations, etc. But, the core logic of the test themselves should be implemented by you. Your code will be scrutinized while grading. As such, please document your code. For example, add in comments for each of the major steps, like “generating permutations”, “computing p-value”, etc. Basically, **document and comment your code well** so we understand what it is doing.

Please raise any doubts you have on piazza.

As expected, **plagiarism will not be tolerated**. Instances of plagiarism will result in a score of 0 for the entire project team and you will be reported to the academic judiciary.

* For the mandatory part, you will analyze a COVID19 dataset which will be unique to your group and assigned by the TAs. Each dataset has daily #cases and #deaths data for two states in the date range 01/22/2020 to 04/03/2021. The dataset can be found under the “States Data” folder. For your assigned dataset, the required tasks can be found below in the “Mandatory dataset analysis” section.
* For the exploratory part, you will use the US-all COVID19 data, which is made up of two files: US\_confirmed.csv and US\_deaths.csv. In addition, you will have to pick an X dataset yourself that is likely impacted by COVID19 or impacts COVID19. You will then come up with three exploratory hypotheses for the X dataset and test those. For example, X could be pollution in the US, energy usage of California, traffic patterns in NYC, crime data for the West Coast, weather data for Mid-Western states, etc. X should have at least one month’s worth of daily data, should be for US, and should be compatible with your hypotheses. You can only pick one X dataset. Examples of such datasets are provided at the end of the document. You can use these, or pick your own. You do not need to repeat the mandatory tasks on this dataset or on the US-all dataset. Some example exploratory hypotheses are also provided at the end of this document. You are encouraged to come up with your own, if possible. The more interesting the exploratory hypotheses, the better. Of course, you are only allowed to use techniques and tests discussed in the course so your exploratory hypotheses must be solvable by them.
* Add your group and X dataset info [here](https://docs.google.com/spreadsheets/d/11bu4D9h0ZS_PnH3h6thMMUxyFG05DfcF7pxVzG1dC9A/edit?usp=sharing). Only accessible via your @cs.stonybrook.edu email. Row 2 is shown as an example, leave it as is. Start from row 3 onwards. Do not edit other groups’ entries. We will fill in column B. To avoid duplicate X datasets, we will require every group’s X dataset to be unique. Once you have finalized your X dataset, add the dataset name and link in column C and the time at that moment in column D. We can see the history so we can double-check the time in column D, if needed. If there are duplicates between two groups, we will contact you to resolve it. Hopefully this is not needed and everyone can find a unique dataset. If someone has already picked a dataset, you cannot pick that. It is first-come-first-served. Add your final submission github link in column E. This should be done before the May 14th, 5pm deadline with the repo accessible by TAs.

**Mandatory tasks to be performed on your assigned COVID19 dataset from column B**:

1. Clean your dataset (remove missing values, sanitize data, etc.). Remove any outliers (except 0s) using the Tukey’s rule from class using the default values as in class. Report what you found (number of outliers). Comment on your findings both for data cleaning (what issues you found, how you dealt with them) and outlier detection. This will be 10% of the project grade.
2. Solve the required inferences below for your COVID19 dataset. Only use tools/tests learned in class. Show your work clearly and comment on results as appropriate. This will be 60% of the project grade, with 15% for each of the four tasks below.
   1. In this task, we want to predict COVID19 stats for each state. Use the COVID19 dataset to predict the COVID19 fatality and #cases for the fourth week in August 2020 using data from the first three weeks of August 2020. Do this separately for each of the two states. Use the following four prediction techniques: (i) AR(3), (ii) AR(5), (iii) EWMA with alpha = 0.5, and (iv) EWMA with alpha = 0.8. Report the accuracy (MAPE as a % and MSE) of your predictions using the actual fourth week data.
   2. In this step, we want to check, for each state, how the mean of monthly COVID19 stats has changed between Feb 2021 and March 2021. Apply the Wald’s test, Z-test, and t-test (assume all are applicable) to check whether the mean of COVID19 deaths and #cases are different for Feb’21 and March’21 in the two states. That is, we are checking, for each state separately, whether the mean of daily cases and the mean of daily deaths for Feb’21 is different from the corresponding mean of daily values for March’21. Use MLE for Wald’s test as the estimator; assume for Wald’s estimator purposes that daily data is Poisson distributed. Note, you have to report results for deaths and #cases in both states separately. After running the test and reporting the numbers, check and comment on whether the tests are applicable or not. First use one-sample tests for Wald’s, Z-test, and t-test by computing the sample mean of daily values from Feb’21 and using that as a guess for mean of daily values for March’21; here, your sample data for computing sample mean will be the 28 daily values in Feb’21 whereas your sample data for running the test will be the 31 daily values of March’21. Then, repeat with the two-sample version of Wald’s and two-sample unpaired t-test (here, your two samples will be the 28 values of Feb’21 and the 31 values of March’21). Use α=0.05 for all. For t-test, the threshold to check against is tn-1,α/2 for two-tailed, where n is the number of data points. You can find these values in online t tables, similar to z tables. For Z-test, use the corrected sample standard deviation of the entire COVID19 dataset you have for each state as the true sigma value.
   3. Inference the equality of distributions in the two states (distribution of daily #cases and daily #deaths) for the last three months of 2020 (Oct, Nov, Dec) of your dataset using K-S test and Permutation test. For the K-S test, use both 1-sample and 2-sample tests. For the 1-sample test, try Poisson, Geometric, and Binomial. To obtain parameters of these distributions to check against in 1-sample KS, use MME on the Oct-Dec 2020 data of the first state in your dataset to obtain parameters of the distribution, and then check whether the Oct-Dec 2020 data for the second state in your dataset has the distribution with the obtained MME parameters. For the permutation test, use 1000 permutations. Use a threshold of 0.05 for both K-S test and Permutation test.
   4. For this task, sum up the daily stats (cases and deaths) from both states. Assume day 1 is June 1st 2020. Assume the combined daily deaths are Poisson distributed with parameter λ. Assume an Exponential prior (with mean β) on λ. Assume β = λMME where the MME is found using the first four weeks data (so the first 28 days of June 2020) as the sample data. Now, use the fifth week’s data (June 29 to July 5) to obtain the posterior for λ via Bayesian inference. Then, use sixth week’s data to obtain the new posterior, using prior as posterior after week 5. Repeat till the end of week 8 (that is, repeat till you have posterior after using 8th week’s data). Plot all posterior distributions on one graph. Report the MAP for all posteriors.

**Exploratory tasks to be performed on the X dataset of your choice from column C**:

Propose three new inferences for your X dataset and solve them using tools learned in class. Each inference is expected to make use of the US-all dataset (full or subset, as compatible with your X). You will be graded on creativity/practicality of your inferences. For each inference you propose, provide a paragraph of text to explain why this inference is practical and useful. Also comment on the results of your inference, as appropriate. For each inference, use parameters that seem reasonable and explain why you are using those parameters. Where possible, comment on whether your inference technique is applicable. Only use tools/tests learned in class. This will be 30% of the project grade, with an additional 10% at the discretion of the TAs for especially creative and interesting hypotheses. This is designed to be open-ended. Creativity is encouraged when designing your inferences. Some sample inferences (you can use these or a subset of these, as needed):

1. Use your X dataset to check if COVID19 had an impact on the X data. State your hypothesis clearly and determine the best tool (from among those learned in class) to apply to your hypotheses. Also check whether the tool/test is applicable or not.
2. Check if COVID19 data changed after some local event or rule was enforced, like lockdown, stay-at-home, or vaccine etc. For this, compare COVID19 data before and after the event. Maybe take into account that COVID19 takes some time to show symptoms, so maybe give some time to allow the lockdown to show its effects.

**Some example X datasets (you can borrow these if needed, as long as no one else has used it already):**

**1. Energy data for different energy sources (Petroleum, natural gas, nuclear etc.):**

[**https://www.eia.gov/petroleum/data.php**](https://www.eia.gov/petroleum/data.php) **>> Sources and Uses tab.**

This site has very exhaustive data collection for various sources of energy. For each source they provide data specific to all the components involved in production of energy from that source. For e.g. for petroleum they provide daily/weekly data of crude oil reserve, import/export, transportation, production and consumption. Data is very clean and requires minimal processing.

**2. Crime data:**

1. **Chicago**

<https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2> [data at incident level i.e. whenever there is an incident, that corresponds to a row in the data with it’s time stamp]

1. **San Francisco:**

<https://data.sfgov.org/browse?q=crime&sortBy=last_modified&utf8=%E2%9C%93>

1. **NYC crime** data has not been updated since Feb 2020 (Most likely the data is updated quarterly level)

In fact most big cities have their own website for open data (crime traffic, environment etc, from which the students can use data as per their need). Some update their data regularly (e.g. SF, Chicago etc) while others don’t.

**3. Livestock & Meat Domestic Data (The data is till Feb 2020, the next update of the data (till March 2020 )will happen on 28-April)**

[**https://www.ers.usda.gov/data-products/livestock-meat-domestic-data/**](https://www.ers.usda.gov/data-products/livestock-meat-domestic-data/)

**4.** [**https://www.bea.gov/data/intl-trade-investment/international-trade-goods-and-services**](https://www.bea.gov/data/intl-trade-investment/international-trade-goods-and-services)

Check Tables only (excel format).

This contains trade data till Feb 2020. It contains import-export data of a lot of categories. Examples include, Food, industrial supply, petroleum, beverages. This is very extensive data, and clean as well.

**5.** [**https://explore.dot.gov/views/BorderCrossingData/Monthly?:isGuestRedirectFromVizportal=y&:embed=y**](https://explore.dot.gov/views/BorderCrossingData/Monthly?:isGuestRedirectFromVizportal=y&:embed=y)

Border-Crossing data (available till Feb-2020): Contains number of people travelled through ‘x’ (port-name) and by mode of transport. In a lot of cases, we can see the decrease in number in Feb 2020, compared to previous year.

**6.** [**https://www.tsa.gov/coronavirus/passenger-throughput**](https://www.tsa.gov/coronavirus/passenger-throughput)

This is Transport Security Administration travel numbers for 2020 and 2019

Available till today.

**7.** [**https://www.epa.gov/outdoor-air-quality-data/download-daily-data**](https://www.epa.gov/outdoor-air-quality-data/download-daily-data)

Air pollution data: Extensive data. Can select type of pollutant, State, County, etc. Available till April.

**8.** You could look at attendance data for major sporting events or concerts, etc. Data may be sparse, so may have to think about how to get enough data to make it meaningful.

**9.** There is a lot of information in the US Census Bureau, such as population data. May be useful.

**10.** <https://ourworldindata.org/>

Seems to have a lot of data that could be useful.

Below are datasets that are older but may have data from COVID19 overlap time, i.e., Feb-April 2020. If not, then the data is likely not useful. Please check. Nonetheless, it can give you examples of what X datasets to look for.

1. Traffic violations in USA. <https://www.kaggle.com/felix4guti/traffic-violations-in-usa>
2. Craig list trucks data: <https://www.kaggle.com/austinreese/craigslist-carstrucks-data>
3. P2P lending data: https://www.kaggle.com/skihikingkevin/online-p2p-lending