Due date: May 12th, 5:30pm, via google colab

Project weightage: 15% 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.

Deliverables format: jupyter notebook + word/pdf if needed

Brief google colab tutorials (if needed):
https://towardsdatascience.com/getting-started-with-google-colab-f2fff97f594c:
https://medium.com/lean-in-women-in-tech-india/google-colab-the-beginners-guide-5ad3b417dfa

Project: analyze a COVID19 + X dataset, where COVID19 dataset can be anything you want such as COVID19 data from NY, from Spain, from a given county in NY, etc. The requirement is that the dataset should span at least one month. The +X refers to an associated dataset that is likely impacted by COVID19. For example, X could be pollution, energy usage, traffic patterns, crime data, weather data, etc. X should have at least one month's worth of data and that time period must coincide with the COVID19 dataset time period. Additionally, X should also have data from before the COVID19 time period, say from last year, to allow comparison pre- and post-COVID19. Examples of such datasets are provided at the end of the document. You can use these, or pick your own.

You must finalize your dataset and ensure it is not the same as other groups. We will ask you to fill in your dataset info and group info here
https://docs.google.com/spreadsheets/d/15rVWd1eWr9dfr_2U437EHc2DKTLunwqC7Ox-_phj4b4/edit?usp=sharing
Sign in with your @cs email. Also add in the last column the date and time when you entered your dataset info. Dataset choice will be first-come-first-served; the COVID19 dataset should be unique and no two groups can have the same COVID19 dataset. Do not edit other groups' entries. Finding a unique dataset is part of the project.

Once finalized, you will have the following tasks:

1. Clean your dataset (remove missing values, sanitize data, etc.). Remove any outliers using the Tukey's rule from 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. Provide basic visualization of the COVID19 and X datasets to explain the general trends in data. Use histograms, timeline plots, etc., to convey any meaningful information as you need to. Comment on your findings from the graphs. This will be 10% of the project grade.

3. Solve the required inferences for your dataset. See “Required inferences” section below. Only use tools/tests learned in class. Show your work clearly and comment on results as appropriate. This will be 50% of the project grade.
4. Propose three new inferences for your dataset and solve them using tools learned in class. 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. See “Sample inferences section below for ideas. Only use tools/tests learned in class. This will be 30% of the project grade.

   For #3 and #4 above, 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 for #3 and #4 above will be carefully 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.

   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.

Required inferences:

1. Use your COVID19 dataset to predict the COVID19 fatality and #cases for the next one week. 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. Make sure that your dataset allows you to verify the one week prediction. For example, use the first three weeks of data to predict the fourth week, and report the accuracy of your predictions using the actual fourth week data. Use metrics learned in class (MAPE as a % and MSE) to report accuracy numbers.
2. 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 from the first week to the last week in your dataset. Use MLE for Wald’s test as the estimator. Note, you have to report results for deaths and #cases separately, so think of this as two inferences. After running the test and reporting the numbers, check and comment on whether the tests are applicable or not. First use one-sample tests by computing the mean of the first week data and using that as guess for last week data. Then, repeat with a two-sample version of Wald and t-tests. For t-test, use both paired and unpaired tests. Use alpha value of 0.05 for all. For t-test, the threshold to check against is $t_{n-1,\ alpha/2}$ for two-tailed and $t_{n-1,\ alpha}$ for one-tailed, where n is the number of data points. You can find these values in online t tables, similar to z tables.
3. Repeat inference 2 above but for equality of distributions (distribution of first week and last week), 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 first week's data to obtain parameters of the distribution, and then check whether the last week's data has the distribution with the obtained MME parameters. Use a threshold of 0.05 for both K-S test and Permutation test.

4. Report the Pearson correlation value for #deaths and your X dataset, and also for #cases and your X dataset over one month of data. Use the most relevant column in X to compare against the covid numbers.
5. Assume the daily deaths are Poisson distributed with parameter lambda. Assume an Exponential prior on lambda. Use first week's data to obtain the posterior for lambda via Bayesian inference. Now, use second week's data to obtain the new posterior, using prior as posterior after week 1. Repeat till the end of week 4. Plot all posterior distributions on one graph. Report the MAP for all posteriors.

Sample inferences (edit as needed depending on your data):

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.

1. (**highly recommended**) 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. (**highly recommended**) Check if COVID19 data changed after some local event or rule was enforced, like lockdown or stay-at-home, 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.
3. (**highly recommended**) Use linear regression to find the impact of age, gender, underlying conditions, etc., on the severity of covid19 symptoms or duration.
4. Use Chi-square independence test to check if COVID19 impacted your X dataset in some way.
5. Check if distribution of covid #cases or #deaths are different based on gender, age group, etc., or distribution of age of covid patients is different for different genders, etc.
6. Check if distribution of covid #cases or #deaths are different based on symptoms of patients and/or based on patients' gender, age, etc.

COVID19 datasets:

https://github.com/CSSEGISandData/COVID-19
Links to JHU covid dataset

https://datarepository.wolframcloud.com/resources/Patient-Medical-Data-for-Novel-Coronavirus-COVID-19

(Patient record including || age|| sex|| location|| date of onset|| symptoms(list of symptoms) || travel history || chronic diseases || and date of discharge || or death ||)

https://github.com/nytimes/covid-19-data (**Time series format data of USA**)
https://developer.nytimes.com/covid
NYTimes Covid dataset

https://ourworldindata.org/grapher/covid-confirmed-cases-since-100th-case
Covid **Time series format data** at country level.

https://www1.nyc.gov/site/doh/covid/covid-19-data.page
NYC specific data, not sure about format

https://data.europa.eu/euodp/en/data/dataset/covid-19-coronavirus-data
This contains World-wide (from all continents) COVID data till April 20th.

https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports
Could be useful but data is not aggregated together

Typing “coronavirus stats” in google search also gives some data, though the source is not clear. Could be useful if you can figure out the data source.

X datasets:

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

a. **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]
b. **San Francisco:**
   https://data.sfgov.org/browse?q=crime&sortBy=last_modified&utf8=%E2%9C%93
c. **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/

4. 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
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
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
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