Project 1 Proposal

Arthur, Jon, and Sonny

The project will study how COVID-19 impacted airport traffic based on information in 2020 as compared to the baseline of pre-Covid traffic capacity. We will first look at data from the TSA showing the passenger throughput in the USA between March 1st, 2020 and Dec 1st, 2020. This data will then be compared to the same timepoints in 2019. This data was sourced from <https://www.tsa.gov/coronavirus/passenger-throughput?page=1>

After looking at how Covid affected travel numbers we will attempt to find relationships between travel numbers and total number of cases using data from the New York Times from March 1st to July 28th. <https://www.kaggle.com/roche-data-science-coalition/uncover>

We will find the correlations between this data and the travel numbers and also explore the deviations in case numbers between the two datasets.

We also will look through a few of the US’s most frequently traveled airports and look at individual trip numbers through a proportion of number of trips compared to the baseline of March 15th, 2020. We will obtain this data from <https://www.kaggle.com/terenceshin/covid19s-impact-on-airport-traffic>

We will look for the relationship between this data and the individual state data case numbers from the NYT dataset. We will examine the Atlanta airport and Georgia, Boston Logan and Massachusetts, San Francisco and California, LAX and California, Dallas-Fort Worth and Texas.

Task Breakdown:

Arthur: creates group project repo and invite team members (partially done)  
Sonny and Arthur: research more data source and come up with questions  
Jon: create architecture for analysis

Research Questions:

1. How have traveler numbers been affected by Covid in the US as a whole?
2. How does this compare to more normal years?
3. How have case numbers changed throughout the year?
4. Is there a discrepancy between different sources and their case number reports?
5. Is there any quantitative difference we can find or explore in these discrepancies?\*
6. How do case numbers correlate with traveler numbers?
7. What do case numbers look like in particular states?
8. What do traveler numbers look like in specific airports?
9. What is the relationship between an airport and its state?

Coding Methodology

The first step will be taking the TSA data and putting it into an Excel document and saving it as a csv.  Then we can load this data with pandas, and using Matplotlib make a line graph of 2019, 2020, and both.  After this convert dates from numbers into strings

For the NYT data:

First load the csv into pandas.  Convert dates into numbers from strings.  Change the dataframe to only include the relevant dates and relevant columns of date, state, and case totals.  Group by date and take the sum of cases from all the states.  Make a new dataframe with the dates and a new column of the sum of cases for each data total.  Then make a line graph of the data.

Similar process for WHO data:

Make a new dataframe where you only use USA as nation.

(usa\_data = total\_data.loc[total\_data[“Nation”] == ”USA”])

Convert dates into numbers from strings.  Also only include relevant dates, total cases, and nation as columns.  Then make a line graph of the data.

For correlations:  create new dataframe from TSA data of just 2020 data.  Merge this data with NYT data.  Make a line graph with both lines and then use .corr function to find correlation.  Repeat these steps with WHO data.

For advanced correlation work we should test the data with some lag in dates to see if there is a delay present in case numbers affecting travel numbers.  To do this create a for loop for example for I from -15 to 15.  Add I to the date for case numbers and test these dates against travel number dates.  Remember to account for smaller date size during this.  Find correlation of the numbers here, and test the square of this correlation to best correlation squared, initially set at 0 squared.  If new correlation squared is greater than current best correlation set new correlation for x dates as new best correlation.  Iterate through all test dates, recording correlation for each segment and show graphs for best correlation period.  Also potentially test for smaller time period where correlation may be stronger as Covid was taken more seriously.

For individual states, take NYT data and (usa\_data = total\_data.loc[total\_data[“State”] == ”State wanted”]).  Make new dataframe with only the dates you want and case numbers.

Use airport traffic data to obtain similar info for related airport to the state.  Find correlation and make line graph with both lines.  Potentially use advanced correlation method.  Do this for all 5 samples.