Summary Report

Approach

1. Shortlisting the domains in which COVID-19 might have had an effect
2. Corresponding changes on climate
3. Types of datasets which measure this information (Geography no bar)
4. Machine learning, correlations, statistical A/B testing.
5. Theorize and conduct research

Effects of COVID-19

Domains with direct change

- Reduced demand, productions and emissions - factories, industries (Water pollution, Carbon, NO2, PM)
- Reduced travels - trains, flights, cars
- Reduced deforestation and pollution - effect of climate change slowed, regrowth, improved AQI

Domains with indirect change

- Economy in shambles - less purchases, less production, less travel
- Reduced energy consumption - less fuel burning

Hence I will be focussing on Air Travel/Deforestation, for which I will require data pertaining to:

- OUTPUT: Air pollution, AQI, Water Pollution, Emissions
- INPUT: Deforestation, Transport
- OTHER: COVID-19

Data Sources

- Deforestation: India Deforestation Rates and Statistics
- COVID Dataset: COVID-19 Dataset | Kaggle
- AQI Dataset: Airdata | US EPA
- Air Travel Dataset: Transtats Data Elements

Solution

For Deforestation, I looked at the data for India only, in the hopes that I could correlate tree cover loss and carbon data. Both these datasets came from the same data source, and were joined at a State level. The dataset involved a lot of preprocessing in the form of converting columns to rows. Taking out a simple correlation, it was found that as we increase the tree cover loss (i.e. increasing deforestation), the carbon levels decrease, which does not make sense. Hence I scrapped this dataset and did not pursue this further.

For Air Traffic, I was aspiring to check how COVID would've led to a decrease in Air Traffic, and in-turn the amount of pollution in the atmosphere. I looked at different variables for AQI, namely CO, NO2, Ozone, PM10, PM2.5, which was acquired through data preprocessing. Statistical t-tests were conducted to validate the inferences, alongside general correlations.

To understand the clear impact of air traffic on the climate, a regression line was fit.

The code is well-commented and annotated for further details of the implementation.

Key Takeaways

- We have established a significant correlation between the amount of Air Travel and COVID-19 using the t-test and log-correlation value.
- We have also established a significant correlation between the amount of Air Travel and Air Quality using the t-test and correlation value.
- The highest correlation was experienced between Air Travel and NO2, among other AQI Emissions.
- Using non-linearities leads to improved accuracy of ML Model.
- We were able to fit a linear regression equation to capture the amount of NO2 pollution given the number of flights taken, with a low MSE value of 54.

Improvements Required

The biggest improvement required would be in the state of Geospatial Datasets, and the documentation and accessibility to it. Geospatial datasets are often in different formats, which can be hard to combine with other satellite data, and even harder to join with ground-level CSVs. Sometimes, even when a good Geospatial dataset was found, it was not overlapping with the COVID duration. Even when I managed to find a NetCDF file for NO2 data, the data was difficult to load on Python.

Here, in retrospect, I should've attempted this problem using Google's Earth Engine Datasets, which can directly be loaded on Colab. Creating a stellar Remote-Sensing Dataset to cater to a specific problem statement takes a lot of time and effort. Even with the simple data-points such as COVID and Air Traffic, the data had to be changed midway due to the inconsistencies such as granularity, missing data, etc.