Data Sources:

Data sources to be used on the project are:

1. Flat file (csv) - COVID-19 Corona Virus India Dataset | Kaggle. I will be using patients level data which tracks daily status of patients in different states in year 2020 in India. The dataset has 22 variables or columns.
2. HTML - https://indiacensus.net/literacy-rate.php#:~:text=Literacy%20Rate%201%20Most%20literate%20state%3A%20Kerala%202,4%20Least%20literate%20Union%20Territory%3A%20Jammu%20and%20Kashmir. The dataset has literacy levels of each of the states in India and has 5 variables or columns with literacy percentages and population for the state.
3. API - https://data.covid19india.org/v4/min/timeseries.min.json. The dataset a timeseries API that provides state level data. It tracks number of confirmed, recovered, deceased, tested and vaccinated by daily, 7 days average and total numbers up to that date.

Relationships:

State name will be used to create relationships between the three datasets. Two letter state codes will be used to get the alignment with the state names as the API link stores data only by two letter state codes.

Project Description

To detect the progression of covid cases in India in 2020 in different states and find if there is a correlation between literacy percentages or lack of it in each of the states and progression of cases in the initial months. In addition, if the primary reason for the spread was travel or through community spreading and that number of tests contributed to decrease in the spread of the virus.

Steps to achieve the milestones

Data cleaning such as removal of empty cells will be required. Incorrect entries such as wrong names entered for states will need to be removed as well. Average from the flat file will need to be calculated to correlate with the API dataset. Some of the columns which are not required in the analysis will be deleted. Summary statistics and timeseries data will be plotted for each of the states. Headers will need to be replaced in some cases. Some states may need to be clustered together to check if clustering helps in getting the correlations being targeted. Literacy percentages may need to be normalized by population.

I plan to utilize MySQL for creation of database for the merged transformed data. Thereafter ggplot2 will be primarily used for visualizations. Literacy data together with the overall covid numbers may provide some insight into the link between the two variables. I will also try to use R packages as this will help with the lessons learnt in the earlier class.

I plan to use regression to check the variables that contributed to the number of cases. Also, hypothesis testing may be required to test whether literacy contributed to whether covid growth in higher literacy states were lower.

Ethical Implications

Poor states may have lower literacy rates. Therefore, correlating covid growth with lower literacy rate may not be ethically correct.

Challenges

I anticipate the challenges during logistic regression and also the assumptions that will be made during the analysis. The correlation cause and effect are always very difficult to connect and is something I will try to be careful about.