TEAM 154

Combined Impact of Weather and Social Determinants of Health on COVID-19 in the US

Emilee Nathan | **Renee Yaldoo** | **Songeun Lee** | **Jaedyn Kwon** | **Thomas Han** | **Rohan Smith**

Summary

In this project, Team 154 aims to determine how weather changes and/or social determinants of health have a combined impact on COVID-19 and its effects in the United States.

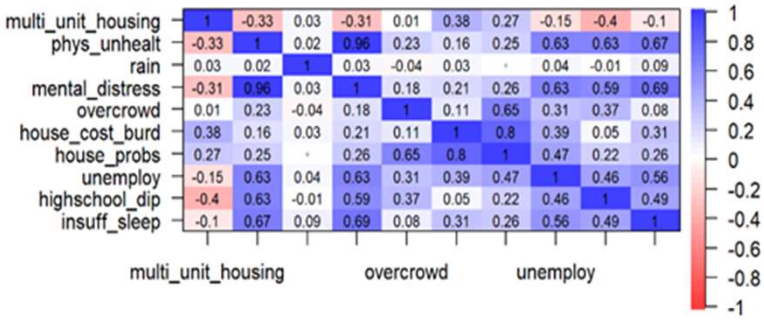
Proposed method consists of **two multiple linear regression models** and an **interactive Tableau dashboard** that allows users to explore the dataset and visualize the regression models.

Data [Source: Kaggle]

- US_counties_COVID19_health_weather_data.csv
  - 1.37GB, 790,331 rows X 227 cols

Exploratory Data Analysis

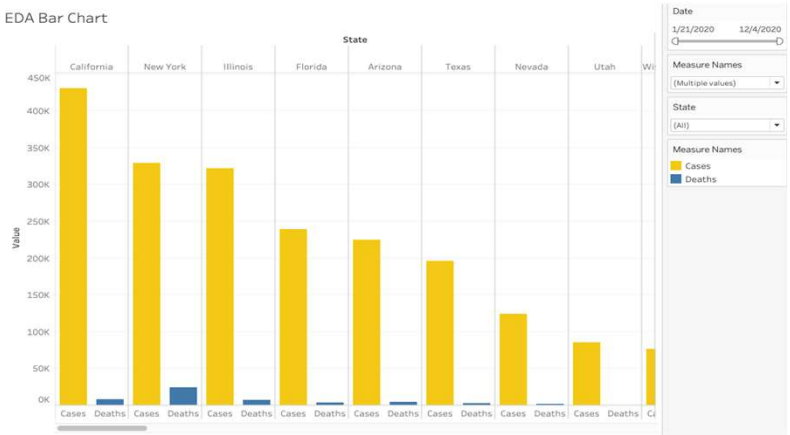
[Correlation Plot between Top 10 significant coefficients - R]



High Correlation
- Avg. number of Physically Unhealthy Days vs. Percent Frequency of Mental Distress
- Severe House Cost Burden vs. Percent Severe Housing Problems

Low Correlation
- Rain vs. Percent Population Overcrowding

[Compare Values of Variables within each States - Tableau]
- Order of states sorted by number of max Cases descending



[Findings via EDA]

- Deaths and Cases in dataset were cumulative numbers, so adjusted dataset to show number of deaths and cases per day recorded for modeling purposes.
- Kept original Deaths and Cases columns to use for EDA purposes.
- Just by looking at the correlation plot, there seems to be no direct effect of weather on socio-economic factors

Data Preparation (1.19 GB, 773,676 rows X 199 cols)

Many of the independent variables in the data set contained missing values. Therefore, we decided to use KNN to impute missing data with five neighbors (k = 5) to perform a more meaningful analysis with more data points.

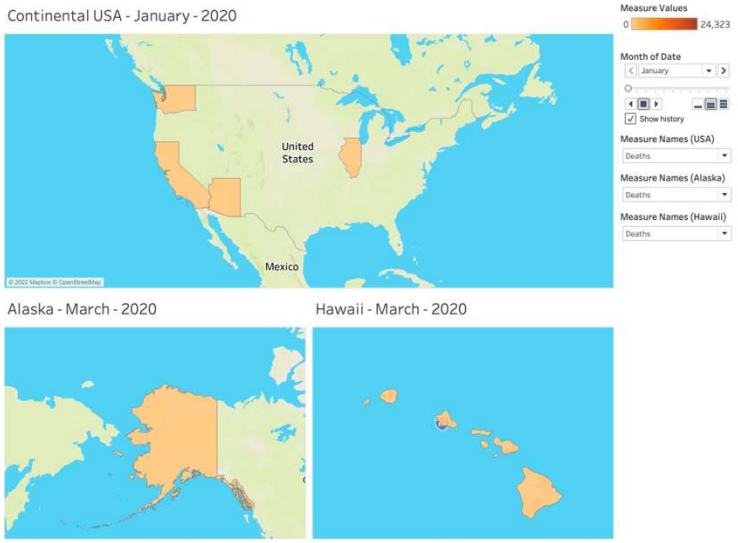
The cumulative “cases” and ”deaths” columns were broken down into the number of cases and deaths for that county/state for that specific date.
This was necessary to use linear regression.

Modeling in R

The multiple linear regression models investigate the relationship between the number of COVID-19 cases and the number of COVID-19 deaths and varying weather and socio/economic factors.

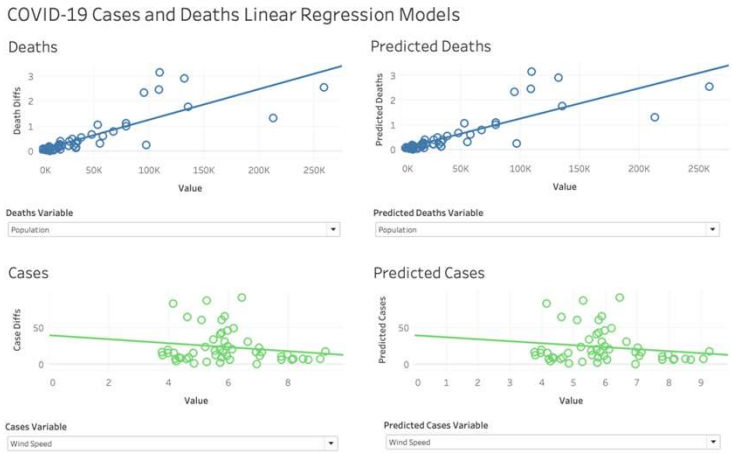
Tableau

[USA Dashboard]
- Automated progression of the variables picked over the months of 2020



- Highest amount of **Deaths – New York / Cases – California**

[ML Visual Dashboard]
- Comparison between our independent variables vs. the model predictions (predicted deaths and predicted cases) and the actual response variables (deaths and cases)



Connected Tableau with R using Rserve to run the **Multiple Linear Regression models**
- As Population increases, the number of predicted Deaths also increases on average
- As the Wind Speeds are higher, there is a lesser effect on predicted Cases on average

The analysis performed by Team 154 lead to the conclusion of the existence of a weak relationship between COVID-19 cases and deaths and most weather-related factors and social determinants of health. This conclusion was drawn by the low R-squared values deduced from the models, however, there is still awareness to be gained from this analysis.