Intro to Data Science | Project Proposal | Frances Chen

COVID-19 Risk Assessment: *Predicting COVID-19 case rates in the United States*

The U.S. response to the Coronavirus disease (COVID-19) pandemic has been poor and chaotic, in part due to the president's muddled reactions. President Trump's responses to the pandemic have been riddled with misinformation[1] and false claims,[2] which have contributed to immense delays and confusion in instituting wide-spread preventative measures and testing across the country.[3]

Social distancing has been increasingly promoted on a national scale to slow the spread of the disease. However, states have reacted to social distancing measures differently; states with high confirmed cases like New York, California, and Washington have seen drastic drops in travel while states in the South and Midwest lag behind in reducing their travel.[4] According to the European Centre for Disease Prevention and Control, risk of occurrence of widespread national community transmission of COVID-19 in the coming weeks will be moderate if effective mitigation measures are in place, but will be very high if insufficient mitigation measures persist.[5]

Considering the unique political polarization and political economy that exist in the United States, I am interested in analyzing **which political, demographic, health care, and economic state characteristics best predict COVID-19 case rates across the United States**:

- **Demographic**: State population size; urban/rural status; population age, race, education, occupation type (e.g. essential vs. non-essential) distribution; income distribution
- **Political**: party affiliation; Trump approval ratings; state-level policies (e.g. shelter-in-place; remote work mandates)
- **Economic**: State GDP per capita
- **Health care system**: Number of hospitals available; Medicare acceptance; hospital capacity
- **Preventative measures**: Social distancing adherence, testing prevalence

More specifically, I am keen to explore whether state-level political affiliation or loyalty to President Trump (as expressed through approval ratings) influence citizens' adherence to preventive measures that is captured through growth in state-level COVID-19 reported cases.

Data

I plan to compile various sources of data, obtained through API and non-API sources, and standardize the data on a **state-level** (using packages *readxl / httr*) to run machine learning analyses to determine how state-level characteristics are correlated with COVID-19 case rates.

<u>API data sources:</u>

- Hospital General Information (U.S. Department of Health & Human Services)
    - List of hospitals registered with Medicare
- 2019 Novel Coronavirus COVID-19 (2019-nCoV) Data Repository (Johns Hopkins CSSE)
    - US city name, state/province name
    - Number of confirmed cases, deaths, recovered cases.

[1] https://www.politifact.com/factchecks/2020/mar/12/donald-trump/trump-wrongly-said-health-insurance-companies-will/

[2] https://www.politifact.com/factchecks/2020/mar/11/donald-trump/donald-trumps-wrong-claim-anybody-can-get-tested-c/

[3] https://www.nytimes.com/interactive/2020/03/17/us/coronavirus-testing-data.html

[4] https://www.nytimes.com/interactive/2020/03/23/opinion/coronavirus-economy-recession.html

[5] https://www.ecdc.europa.eu/en/publications-data/download-todays-data-geographic-distribution-covid-19-cases-worldwide

- American Community Survey (ACS) (United States Census Bureau):
    - Detailed US demographic data at various geographic resolutions

Non-API data sources:

- States Reporting Cases of COVID-19 to CDC
    - Updated daily, so will need to take daily pulls of table data to track growth in coming weeks
- European Centre for Disease Prevention and Control (ECDC)
    - Country-level data on the geographic distribution of COVID-19 cases worldwide
- Trump Approval Ratings by State (*not sure of the best source for this)
- Political Composition of the 50 U.S. States (Gallup)
- Social Distancing Scoreboard (unacast) (*need to find source that measures scores over time)

Proposed Methodology

Data wrangling component

- Standardize all data at **state-level**
- Scrape reported state cases from CDC website on a daily basis for one week (unless can identify data source with reported state cases over time)
- Monitor and track implementation of state-level policies (e.g. shelter-in-place mandates)

Data visualizations

- Descriptive:
    - State average adherence to social distancing over time
    - Health system preparedness: Number of hospitals per capita in each state
    - Average state Trump approval ratings and number of confirmed cases in state over time
- Machine learning:
    - **Correlation plot**: visualize correlation between variables and COVID-19 cases
    - **Prediction error (RMSE) plot**: visualize machine learning strategy performances

Machine learning component

- **Ordinary least squares (OLS) linear regression model**: which independent variables are most highly correlated with COVID-19 spread
- **K -Nearest Neighbors (KNN)**: train data on existing COVID-19 cases and characteristics in states to predict future COVID-19 spread
- **Random Forest (RF):** low correlation between states can contribute to building a highly predictive national-level model for COVID-19 spread

Project Success

Success in this project will be defined by effectively executing each machine learning component (OLS, KNN, RF) conditional upon the cleaning, manipulation, and standardization of data at the state-level to best predict COVID-19 spread within each state. I will know the project has been successful if these models can pinpoint which characteristics best predict COVID-19 spread.

Further Citations

- citation(readxl)
- citation(httr)

Further Resources

- Mapping 2019-nCoV (JHU)
- Ventilator Availability (JHU)
- US ICU Resource Availability for COVID19 (Society of Critical Care Medicine)
- COVID-19 Research Database (provided by the WHO)
- LitCOVID (provided by the NIH)
- COVID-19 Resource Page (provided by Microsoft Academic)
- COVID-19 Research Export File (provided by Dimensions)
- Day-Level COVID-19 Dataset (hosted on Kaggle)
- COVID-19 Global Cases (provided by Johns Hopkins University)
- COVID-19 Open Patent Dataset (hosted by Lens.org)
- Blog Post: Computer Scientists Are Building Algorithms to Tackle COVID-19
- Our World in Data - COVID19
- Corona Data Scraper: pulls COVID-19 Coronavirus case data from verified sources, finds the corresponding GeoJSON features, and adds population data.