# Project Statement

The course of COVID-19 pandemic in US varies dramatically across regions and time. Popular discussions have emphasized the role of policies, income, age, belief in science, etc. However, the relationship between different factors and case growth is unlikely to be linear and there are complex interactions between factors which co-determines the pandemic evolution. Clear empirical evidence connecting the complex predictor network to case growth and timely scenario-based simulations are essential to steering policy responses.

In this project, we will provide new evidence on the power of different spatial, temporal and administrative factors in predicting virus transmission curves and provide multiple counterfactual simulations for policy implications. Specifically, we will take three steps to achieve this goal:

1. Prediction of COVID transmission curve.  
   Method: Random forest or Neural network.  
   We will calibrate the disease transmission parameter in Susceptible-Infected-Recovery (SIR) model on a monthly basis. There is a mathematical relationship between the growth rate of COVID cases and R0, the essential input parameter of SIR epidemiological model. We will thus predict the empirically derived COVID growth rate using policies, local socio-demographic factors, political ideology, climate conditions, days since COVID outbreak and their interactions.
2. Counterfactual simulations.   
   We will use the relationship we built between in previous step to simulate different counterfactual scenarios of the virus transmission curve given different policy scenarios (e.g., the timing of stay-at-home orders), various climate conditions (e.g., cold or hot), and heterogeneous local socio-demographic compositions (e.g., share of elderly population), etc. This can serve as an effective tool for policy makers of different jurisdictions to understand the risk and tailor their policy interventions.
3. Predict the onset of second/ third wave.   
   Method: LASSO penalized logistic regressions.   
   The above exercise will provide tools for controlling a given wave of pandemic peak. Yet in US and Europe, we see a second or even third wave which is even more severe than the first wave at the onset of pandemic. There is a desperate need to understand what causes the returns of pandemic and how can we foresee its coming to forestall or mitigate it. For the last step, we will use the policy portfolio, socio-demographic characteristics, mobility by places and weather conditions in the previous month of pandemic resurge to predict the onset of second/ third wave.

Putting together, we wish our predictions using data science approach can better capture the complex interactions between virus transmission moderators and inform the debate on optimal policy responses accounting for the local contexts of each US county.

# SIR model overview

SIR model is a classical theory model for the prediction of pandemic. It splits the population into three compartments including Susceptible, Infectious and Recovered. The theory allows us to describe the number of people in each compartment with the ordinary differential equation thus easy to get the step-wise simulation.

There are 2 determinant parameters in the model. β represents the number of people that an infected person could affect. γ represents the chance for an infected person to get recovered. Another important idea which is usually mentioned is R0, the basic reproduction number. This is an epidemiologic metric used to describe the contagiousness of a pandemic which could be calculated by:

In previous literature, can be calibrated using the growth rate of COVID cases[[1]](#footnote-1), and (i.e., recovery) are usually assumed using findings of epidemiological research. Our goal is to use rich predictors to predict the growth rate of COVID cases, transform into , and simulate the curve for a time period feeding back the parameter into SIR model.

# EDA

## Data

|  |  |  |
| --- | --- | --- |
| Item | Source | Content |
| County Level Mobility Index | [Google Mobility](https://www.google.com/covid19/mobility/index.html?hl=en) | The strength of mobility in six categories comparing to the baseline period. Six categories are: retail, groceries, parks, transit, workplaces and residential |
| State Level Policy Database | [Boston University Researchers](https://docs.google.com/spreadsheets/d/1zu9qEWI8PsOI_i8nI_S29HDGHlIp2lfVMsGxpQ5tvAQ/edit#gid=973655443) | The policy database collected comprehensive policies implemented within the US to control COVID-19 pandemic, including “shelter-in-place”, “stay-at-home”, “restaurant / bar / theater closure” etc. |
| County Level COVID-19 Database | [Johns Hopkins University](https://github.com/CSSEGISandData/COVID-19/tree/master/csse_covid_19_data/csse_covid_19_time_series) | The confirmed cases and deaths from January 21st to present. |
| County Level Demographic Database [1] | 2014-2018 American Community Survey extracted by tidycensus package | Female\_percentage, Median\_age, population, Latitude, Longitude |
| County Level Demographic Database [2] | [2017 American Community Survey](https://www.kaggle.com/muonneutrino/us-census-demographic-data?select=acs2017_county_data.csv) | Racial composition, Work Type, Income, Transportation Mode, Employment Types |

## Preprocessing

In our project group, each member was assigned the task to collect data for county social demographics, covid-19 cases, policies and mobility. To be integrated easily by others, each of us are supposed to aggregate the data and generate the processed csv/excel file which includes unique fips code.

**Mobility Index:**

Currently we choose to drop all NAs in our database and reserve qualified rows only.

**Policy:**

The raw data has multiple lines to represent each state and multiple columns to represent the start date or end date of policies. To be able to join with the daily cases data, we convert the policy data into dummy variables thus having a time series policy Dataframe.

**Socio-Demographics:**

State\_code is missing for Puerto Rico which we dropped; ChildPoverty is filled with median value for the missing counties in Hawaii; Crime variables are filled with median value for 7 counties in Alaska, 1 county in New Mexico and South Dakota.

**COVID-19 cases:**

We applied the 7-day moving average to conquer the problem of weekly effect. Two new columns are generated: cases\_7 and deaths\_7.

## Selected Figures

(Please refer to jupyter notebook for the codes and more detailed explanations of the EDA.)

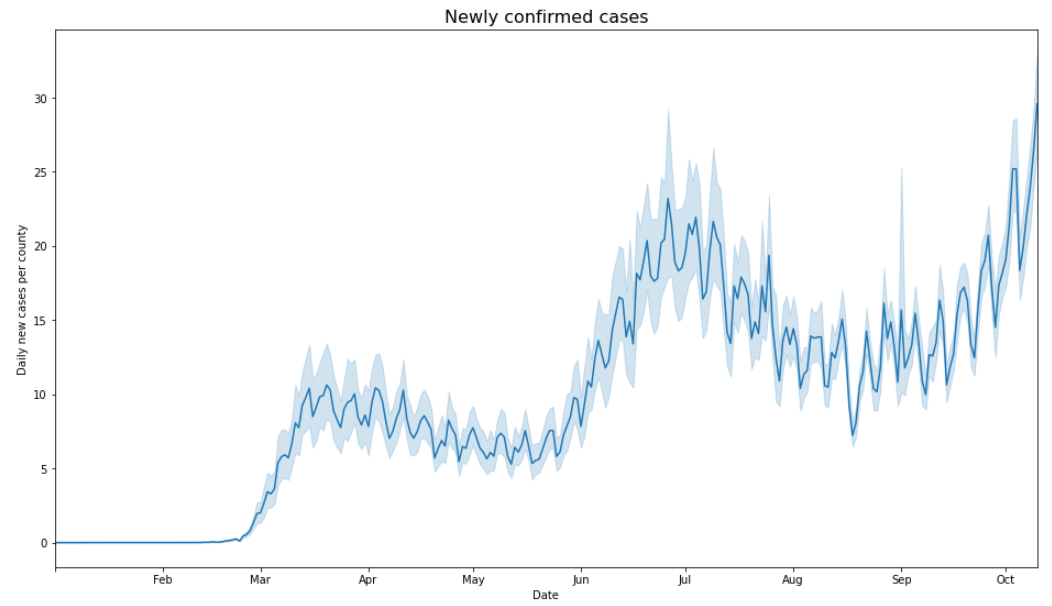


Fig 1 Weekly circular patterns in the newly confirmed cases

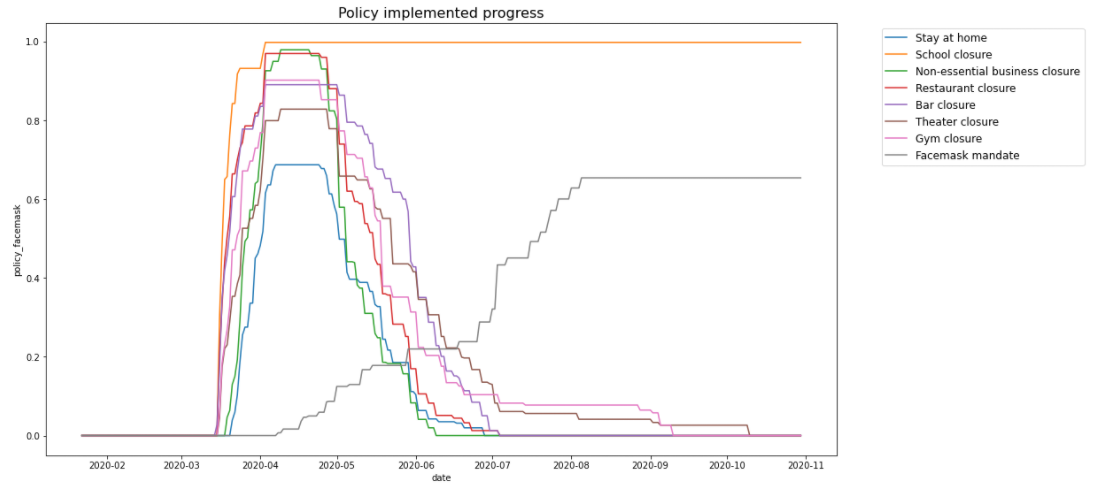
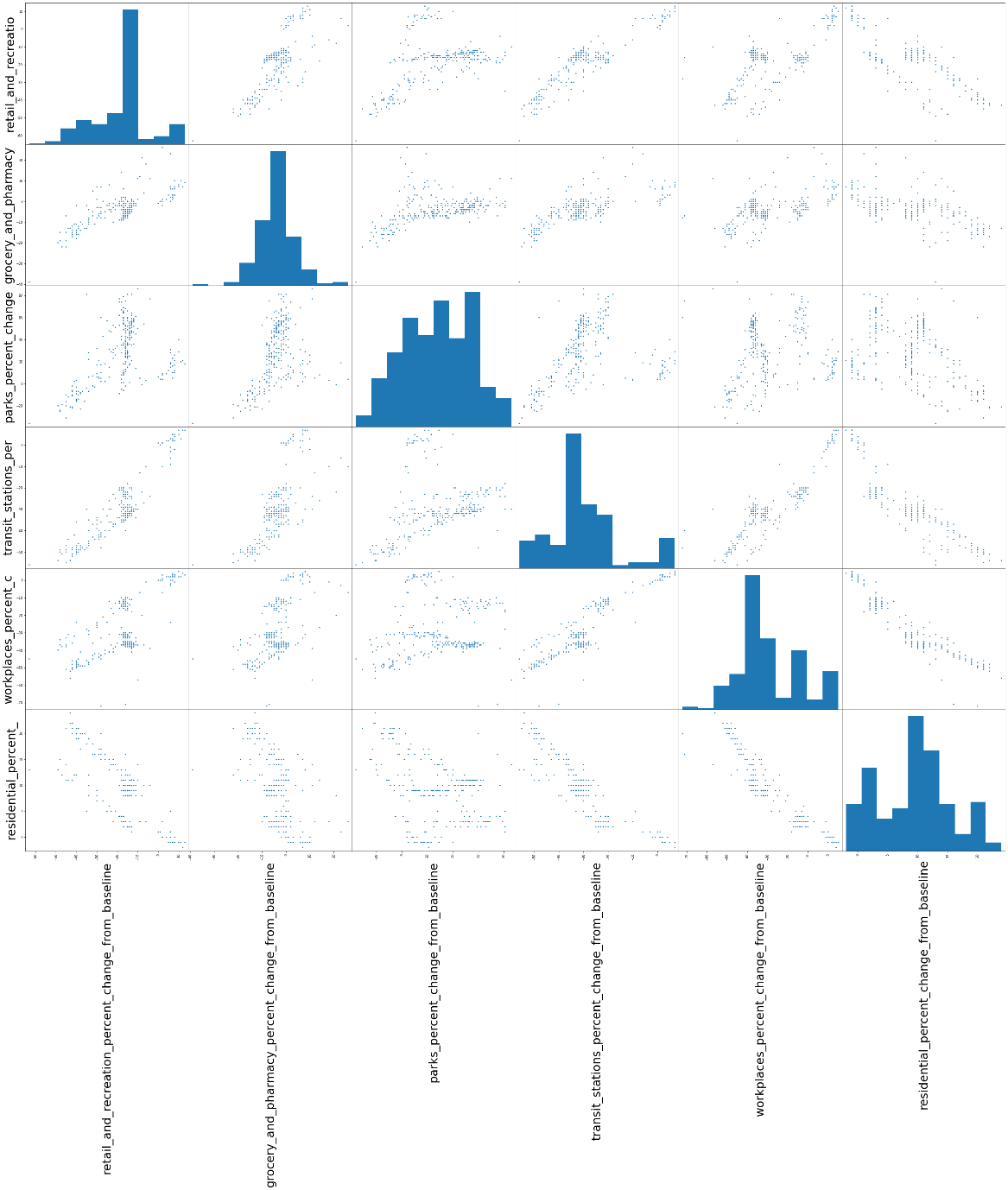


Fig 2 Except for facemask mandates, most policies are implemented around the same time period



Correlation between Mobility Index

Fig 3 We can see strong correlations between some index. For example, workplaces\_percent is strongly correlated to transit

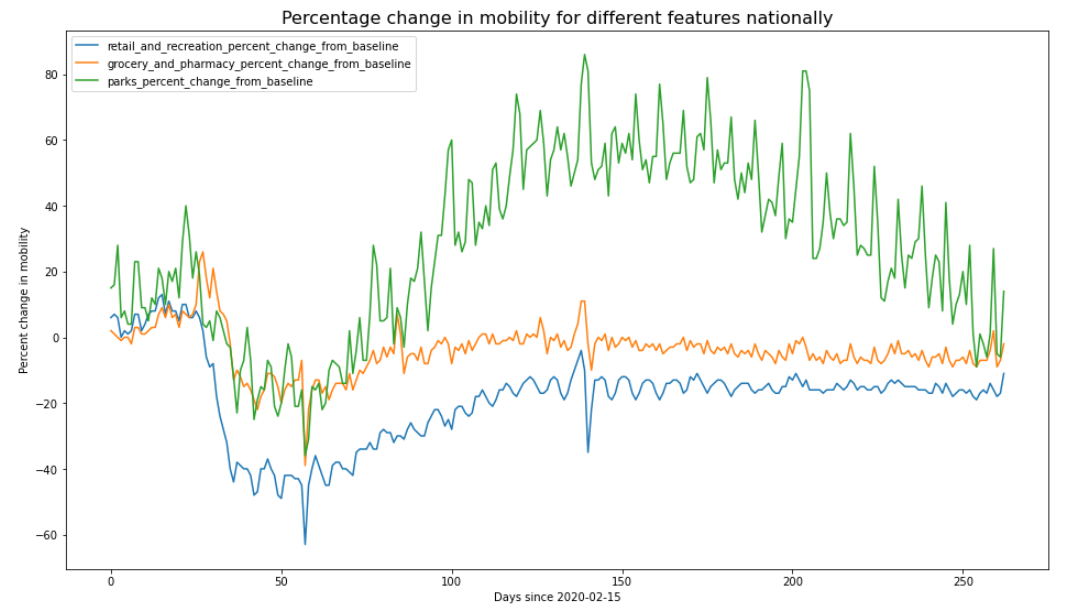


Fig 4 Comparing to other mobility index, park visitation experienced a clear increase and decrease

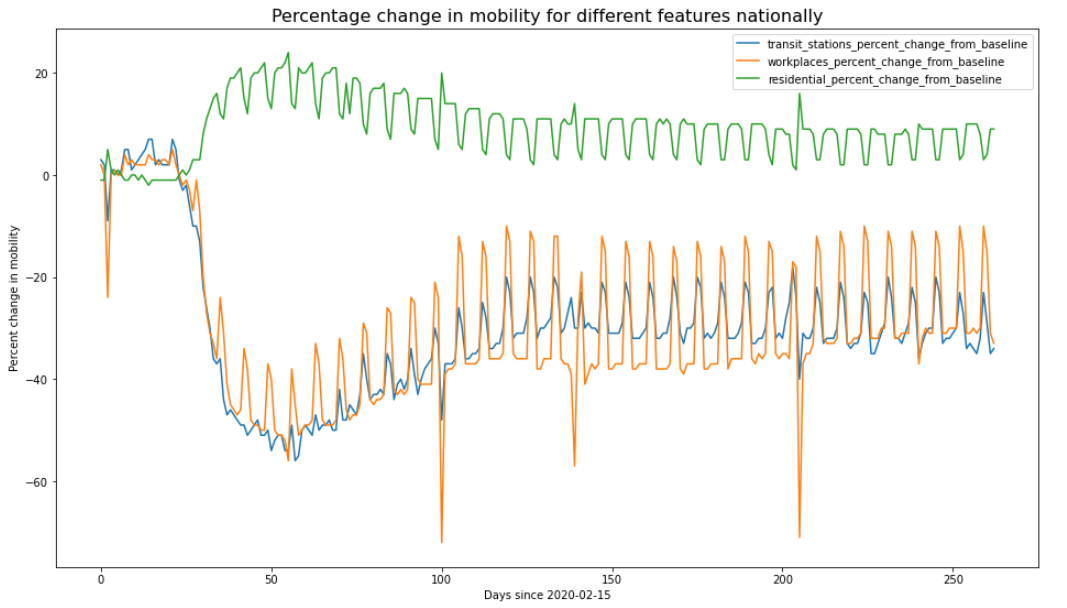


Fig 5 There is strong weekly pattern in the index related to work and transportation. The valley of workplaces might because of holidays.

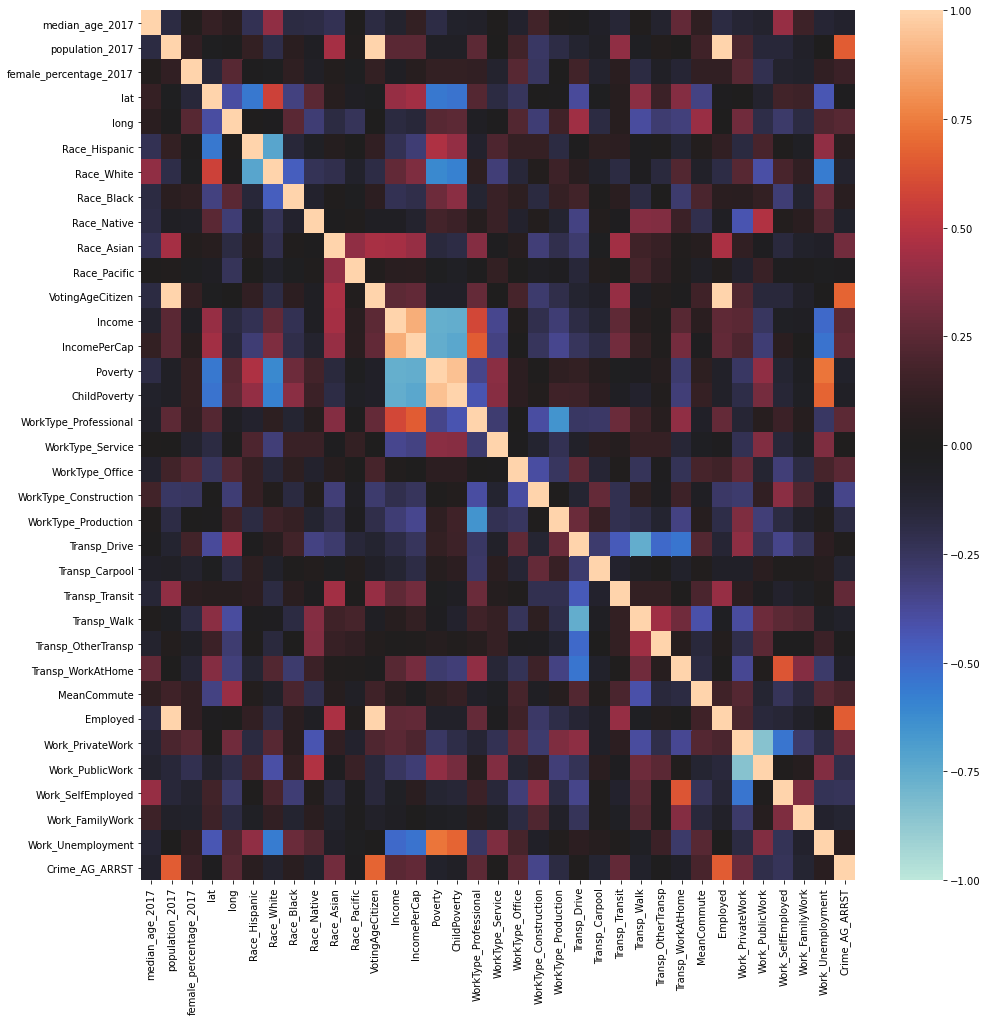


Fig 6 Correlation between demographic features

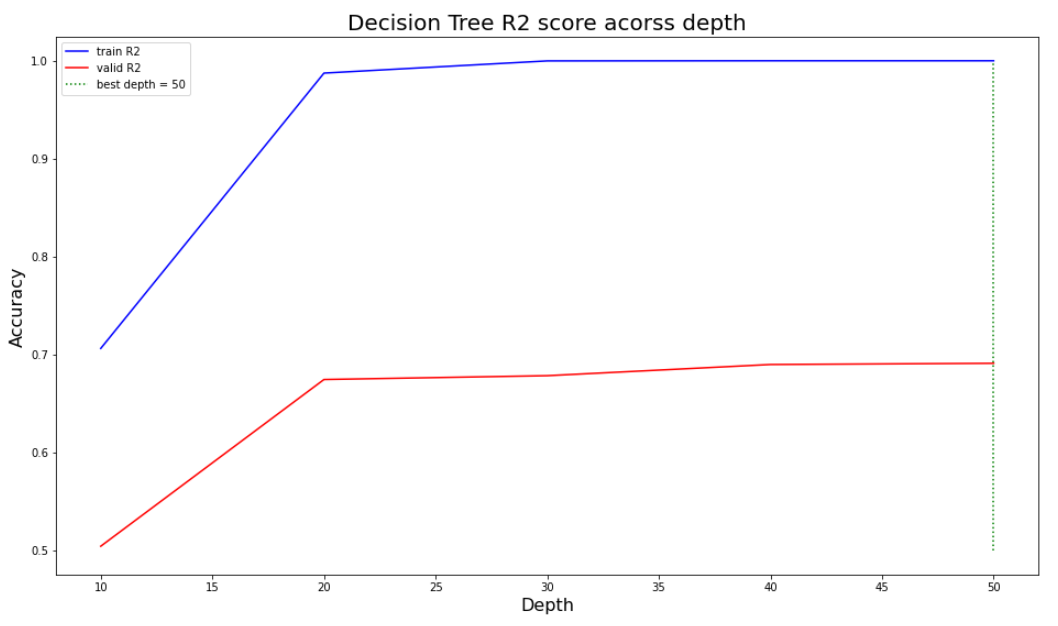


Fig 7 We implemented a decision tree regression to predict the COVID-19 growth rate (β). The R2 score achieves 0.69 on validation set.

1. See: Hsiang, S., Allen, D., Annan-Phan, S., Bell, K., Bolliger, I., Chong, T., Druckenmiller, H., Huang, L. Y., Hultgren, A., Krasovich, E., Lau, P., Lee, J., Rolf, E., Tseng, J., & Wu, T. (2020). The effect of large-scale anti-contagion policies on the COVID-19 pandemic. *Nature*. https://doi.org/10.1038/s41586-020-2404-8 [↑](#footnote-ref-1)