Exploration and modeling of COVID-19 data

For this assignment, you will conduct EDA and build models of COVID-19 dataset(s) that intrigue you.
This assignment may be completed individually or in teams of 2-4 students

By the due date for this assignment, **each team member will upload**:

- their own scripts: R/(Rmd + html) OR Python+Pandas/JupyterNotebook (25 pts).

Also, by the appropriate due dates **one designated team member will upload**:

- a team description of your proposed dataset(s) (3 pts)
- a team report of your EDA and modeling: Word OR (Rmd + html) OR JupyterNotebook (20 pts)
- an account of which team member performed what tasks (2 pts).

The **team description of dataset (due April 20th)** will be uploaded by the designated team member. You are encouraged to discuss your proposed dataset with your instructor during office hours and during the limited time set aside for discussion on Wed April 15th. Your team description should include a brief overview/description of the dataset(s) you plan to use, and a list of your team members. You may choose your dataset(s) from Johns Hopkins datasets on confirmed infections, deaths and recoveries that we've been following in class.

Any dataset of COVID-19 that intrigues you is fine to propose. You may investigate and model the association of statewide Shelter-In-Place orders with flattening of the curve of new COVID-19 cases. You may explore data and predict new COVID-19 deaths in specific region(s) 2 or 3 weeks into the future using your own simulation and compare with a model that interests you (ex: regression, SVM, tree-based model, ANN). For predictions, **bonus points for correctly predicting the future**.

For your report, let your comments be brief interpretations of your visualizations and summary output.

Please be sure to include the following steps:

1. Perform any needed data cleaning/preparation.
2. Create two subsets for further exploration (train-test split)
3. Display aggregate information on your data, perhaps stratified (ex: age, gender). (Ex: ***summary*** in R or ***describe*** in Pandas.
4. Create visualizations of univariate distributions of key variables.
5. Create visualizations of bivariate or trivariate relationships that tell the story of your data (ex: scatter plot of regional mortality vs time-to shelter-in-place order, aligned according to date of 20th case, colored by a measure of wealth.
6. Fit 3 different types of classic ML models (ex: regression, SVM, tree-based model, ANN) OR just one type of model to compare with your own approach ex: (simulation). Be sure to evaluate your models using appropriate visualizations (ex: Residuals vs Fitted plots, ROC curves) and statistics (ex: sensitivity/recall, specificity, precision, confusion matrix, AUC).
7. In a team report, briefly analyze/evaluate/summarize your best findings in EDA and modeling. Compare your models, evaluate as to quality: which are best. Have your best plots tell much of the story of the data. A key part of your report is describing the insights your EDA, plots and models reveal.