**Covid-19 Insights**

**Overview:**

Use feature data to determine what features best predict Covid-19 case & death rates. From there create a webpage where users can specify feature values and the code can specify the predicted case and death rates.

**Data Sources:**

* COVID-19 US County-level Summaries – Kaggle.com
* US counties COVID 19 dataset – Kaggle.com

**Approach:**

* Import data from CSVs into either Pandas or Apache Spark.
* Use transposes, pivots, and/or joins to combine all data into one master table.
* Import Scikit-Learn and IterativeImputer for Machine Learning purposes.
* Use Train\_test\_split to split features (Covid-19 predictors) & outputs (Covid-19 case & death rates).
* Create & train a Linear Regression model to draw correlations between the features & outputs.
* Obtain the R^2 score for the model.
  + If the R^2 score is sufficiently high (>= ~0.7), then go to next step.
  + If the R^2 score is too low, either use hyper-parameter tuning to improve the R^2 (if possible) or find better feature data.
  + Repeat the parent step.
* Tabulate the predicted vs actual case & death rates.
* Use ANOVA analysis with F-values and p-values to rank the features by importance and tabulate the results.
* Create a website that has fields where users can enter feature values, see the ANOVA feature importance table, and obtain Covid-19 case & death rate predictions.