COSC 311 Project 2

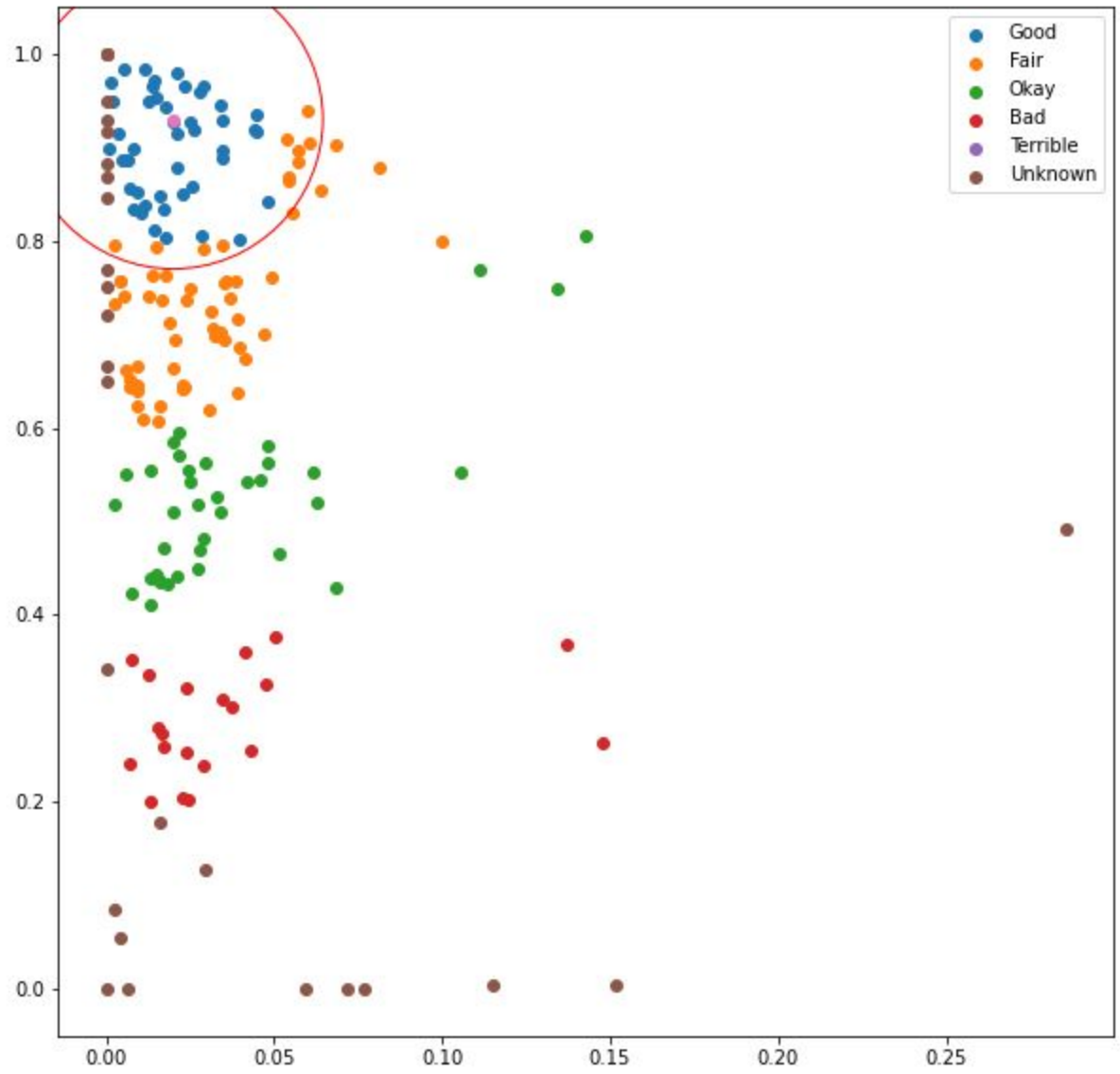
Members: Declan Sheehan, Jack Stoetzel

1. The data we will use for our machine learning algorithms is the COVID-19 dataset here: Kaggle C19 (full_grouped.csv). It comprises every day from January 21st, 2020 to July 26th, 2020 with 187 countries that each have current total cases, deaths, recoveries, active along with new cases, deaths, and recoveries. It is fair to say that this data represents the beginning of the Covid-19 cases trend for nearly every country. Unfortunately it is only the beginning, because as a new day passes, the nation hits new record highs in cases and deaths. The data was gathered from the worldometer website, and CSSEGIS Covid-19 repository.
2. The only two classes this dataset seems to have is W.H.O. region, and season. The regions are six regions: Americas, Europe, Africa, Eastern Mediterranean, South-East Asia, and Western Pacific, and the seasons would be: winter, spring, summer. Other classes would have to be observed by data itself. Observing the small chart below, it is evident that the Americas are having the worst time (followed by Europe) dealing with the pandemic… Or is it?

WHO Region	New cases	New Deaths	New Recoveries
Americas	8842455	342732	4468616
Europe	3316928	211144	1993723
South-East Asia	1835296	41349	1156933
Eastern Mediterranean	1490854	38339	1201400
Africa	723540	12223	440645
Western Pacific	291879	8232	206742

   Dividing the values by population of each region would put this data into a whole new perspective. In addition, there likely exists a large error bar, since not all tests are accurate, and many countries lack the necessary amount of tests to fully encompass the total Covid-19 cases. Nonetheless, it is evident that features for America, Europe are high case and death rates. Countries will specifically correspond with region class, however there are not many certain attributes that distinguish a dataset to a class.
3. If we were to establish different levels of Covid-19 for each country ranging from low (minimum cases/deaths, high recovery rate) to high (maximum cases/deaths, low recovery rate), we could use cases, deaths, recoveries, new cases, new deaths, and new recoveries to train a machine learning algorithm to classify a country to a specific level.
6. The k-nearest-neighbor algorithm performs well with two dimensions (death rate & recover rate). The predictions with two dimensions unfortunately classify more countries as “Unknown”, but the situation improves with three dimensions. Here are both the two-dimensional figure, and three-dimensional figure:

The country's Covid-19 status is: Good



The country's Covid-19 status is: Fair