**Coronavirus Linear Regression**

The project aimed to observe virus protein mutations and if there is any linear relation with death growth rate change via correlation analysis. Also aim was extrapolating death growth rate by countries based on the linear model.

For this purpose, protein sequence data of Covid-19 was imported from GISAID at 24.05.2020 and Covid-19 statistics with new deaths, new cases, stringency indexes, female and male smokers, population older than 65, GDPs per country and much more parameters by countries and by dates could be found and used to get pandas dataframe from “Our World in Data” and combined with other statistics taken from “OECD”, was also calculated manually. Data cleaning and preparation took long time by filling missing values generally with values obtained from reliable references. Protein sequence scores were combined with Covid-19 data frame by their dates according to countries. Sequence scores from same dates and same countries were utilized by taking mean values of the scores. These are progressively done in general\_data\_preprocess.py, protein\_data\_preprocess.py and combined\_data\_preprocess.py. Functions in these scripts are not general, most of them is used for once or twice, to correctly manipulate the non-typical/none- values.

To calculate death growth rate, weekly and monthly data frames were generated with functions that are explained in analysis functions(df\_summer, dataframe\_selector\_all, dataframe\_selector\_obo), which are basically work on boolean principles. Then the weekly dates and monthly dates, score were adjusted again by taking mean values and clades belonging to countries were added. Afterwards, general data frame was optionally standardized and normalized via functions in analysis\_functions.py(clade\_remover, scaler\_function).

Each dataframe is processed to a regression class in the Machine\_Learning.py, which takes both world data and country-filtered dataframe.

For the purpose estimation analysis were performed via scikit-learn linear regression tools by forming train and test sets and fitting 1,000 times and measuring accuracy. After training the data on world data, model is used for country death growth rate prediction.

Unfortunately, results generally have low accuracies which indicate the measure of variance proportion for a dependent variable which can be explained by independent variables in a regression model. Due to low level of accuracy, our model was interpreted as did not fit into a linear model. Mutation of Covid-19 was also observed non-significantly as not usually affecting death growth rate change. Although some of the protein mutation scores have shown as decreasing growth rate. Therefore, more accurate observations are needed. This may be due to instability of the proteins that are mutated. However, our regression model is not able to correctly assess the mutations. More elaborate models such as vector embedding, polynomial regression models may be used.