#  Predicting World Happiness

## Overview

The selected topic for the project is predicting world happiness utilizing the [world happiness dataset](https://www.kaggle.com/ajaypalsinghlo/world-happiness-report-2021) from Kaggle.  This dataset consists of numerous continuous variables presenting the opportunity to utilize a multiple regression machine learning model to predict the happiness of a country or region of the world.  Additionally, using backwards elimination, the variables that are the least predictive of the happiness score can be removed to improve the performance of the model

The source data includes the following:

Country (categorical)

Region of the World (categorical)

Year (categorical)

Happiness Score (continuous)

Economic Production (continuous)

Social Support (continuous)

Life Expectancy (continuous)

Freedom (continuous)

Absence of Corruption (continuous)

Generosity (continuous)

In this project, we are aiming to answer the following questions:

1. What variables are most predictive of happiness by country?
2. What variables are most predictive of happiness by region of the world?
   1. Which region’s happiness has gone up the most? Which has gone down?

## GitHub

Each member of the team has a branch in the World\_Happiness\_Report where code will be committed before merging with the main branch.  A pull request will be created each time someone from the team wants to work on code from the main branch and any code being merged to the main branch will also be reviewed to avoid conflicts during the process.

Communication with the team will also occur outside of GitHub via slack and video to ensure proper use of GitHub.

## Machine Learning Model

Given that the data set includes all continuous variables that would be used for predictive modeling, a multiple linear regression machine learning model was chosen. To access this data, the machine learning model will connect to our provisional database, likely in PostgreSQL.

There are 6 independent variables that will be utilized in the model – economic production, social support, life expectancy, freedom, absence of corruption, and generosity – that will be used to predict the dependent variable – happiness score.  All of these variables will be used to train the model.

The R-squared score, as an output of the model, will help determine the strength of the initial multiple linear regression model and whether it is predictive of the happiness score.  A score higher than .75 will be ideal; however, a lower score may be accepted as some of the variables measure human perception, which adds complexity to the model.

To identify which variable(s) are the most predictive of a country’s happiness score, backwards elimination will be used and any independent variable with a p-value >.05 will be removed from the multiple linear regression model.  This approach will likely increase the overall R-squared score of the multiple linear regression.

##  Database

A provisional database, likely to be in PostgreSQL, will be created and referenced in the machine learning model overview. The database will house each variable and value in the main happiness report csv file which shows the progression/regression of each country’s scores between 2007-2020 and 2021 world happiness report csv file which shows the happiness of each country in 2021. By creating the world happiness report 2021 table in pgAdmin we are going to add 20 variables that identify social, political, psychological, and etc aspects that correlate to overall happiness of a country. A table will also be created for the world happiness report which holds 11 variables of similar context. The two tables will then be joined to create one table for the machine learning model, visualizations, and other analyses.

 Do we want to stick with the country?

 What else would folks add about source data description?

 This is a placeholder and the questions I’ve put in here can definitely be changed!