Task 1)

	mean	Standard deviation
Developing countries	67.11146523178817	9.004228045064881
Developed countries	79.19785156249996	3.9271012579848645

Please explain why standard deviation can be important to characterize the life expectancy statistics:

Standard deviation is really important because it tells you the spread or distribution of the data, which can help us calculate and determine different life expectancies across countries. If the standard deviation is really high, that means there is a lot of variability in the data, and that might make us wonder if there's a certain region in a country that is bringing a lot of outliers to the average life expectancy of the country. Additionally, if the standard deviation is really high, then that tells us that the life expectancy is "unstable" and that maybe some people are dying young and others are dying very late in their lives. In other words, everyone seems to be dying at completely different times, which makes it very hard to predict when the average person will die. If the standard deviation is low, then it makes it easier to make predictions about when the average person will die because there is not much variability between the deaths of people, which means that on average everyone will die at around the same time.

Task 3)

Overall Model Description:

When designing a machine learning model to predict the labels, I would first make sure to filter out any potential outliers and missing data that could potentially mess with the accuracy of my model. I would remove the outliers depending on the type of distribution the data was in. If it was a normal distribution, Q1-1.5*IQP and or Q3+1.5*IQP would be considered outliers. It would be important to first remove these so that the model isn't skewed in any way and it can make it more interpretable. This will, in general, lead to my model being more generalizable, for outliers can cause overfitting in the training data. After that, I would want to determine what data I was going to use to train the model and which algorithm I would use. The data or features that I would train it on would be the 20 predicting variables. Since we haven't covered a lot of ML algorithms, I might use something common such as a decision tree, but I might change this once we learn more about algorithms. Once I have my training data, it is important to gather my testing data so that I'm able to see the model's performance. The testing data would be an accurate collection of countries with their corresponding labels that the model has not used or seen during training. After training, I can then compare its performance to the testing data by using precision, recall, F1, and AUC/ROC calculations. If there is a lot of error, I will need to

change the model's parameters and keep training it until I reach satisfactory results with precision, recall, F1, and AUC/ROC calculations. It is difficult to determine exactly when the model would be considered accurate enough, but it would likely be accurate enough when there is minimal error. The final step is making it available for other people to use once we are satisfied that it is accurate enough.

What features will you use the predict the label? Why they can be important

Some of the important features that I can use to predict the label include GDP, death rate, birth rate, any kind of diseases and their count, income, etc. These are very important because they can tell me all kinds of different background information about each country that my model will need to take into account when predicting a label. For example, a country with a very high GDP, population, and average age of death, for example, is much more probable and makes more sense to be placed under "Long", whereas a country that is the opposite of this one may be placed under "Short".

As each country has multiple years of data, how to preprocess your data to generate the features that you need?

There are several possible ways of handling countries with variable years of data. One way is selecting the most recent five years of data and doing the label prediction on that set of data. Another way is calculating the mean for each of the 20 variables out of all the years of data available so that each country is narrowed down to only one row. Finally, I could select a time frame of data that I am going to collect for each country, assuming it has available data, for example, 2012-2020.

Can you use dynamic information (e.g., GDP is increasing/decreasing rates) to enhance the model that you just created?

Of course, it would be really important to consider this information. By being able to see the change in GDP over a period of time, you can calculate the rate, how it is changing, and by how much. For example, if a country's GDP has been decreasing by 50% over the past couple of years, then it would be more probable to place that under "Short". Likewise, if a country previously had a poor GDP 50 years ago, but in the last 10 years it has been increasing by 10% each year, maybe the country would be considered "Short" 50 years ago, but now it can be considered "Normal". Of course, there are other features that the model would need to take into account to be able to firmly make this prediction, but knowing how different data is changing over time and not just within one year is crucial because it tells you how the country is performing over time.