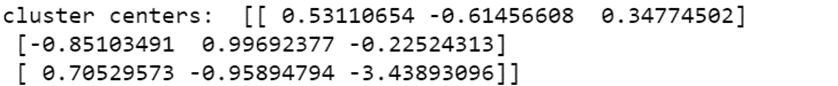
**(B) switching n\_init in kmeans and compare high risk cluster results:**

n\_init shows the number of times the kmeans algorithm is ran in order to calculate the cluster centroid. Through switching the input value for n\_init from 2, 5, 11, 12, I discovered that the clusters for high risk , low risk , medium risk for the countries only varies when n\_init is in between 1 to 9. When n\_init is greater than or equal to 10, the risk clusters no longer changes. That means that the k means algorithm for this country risk data set with 3 features ( Peace, Legal, GDP growth) is stable after running the algorithm for 10 times.

For all the cluster observations in the report: The order of the list in the cluster center is in the order of Peace index, legal index, and GDP growth. **A high-risk country is a country with a high Peace index, low legal index, and low GDP growth.**

**When n\_init = 2**,

0 = medium risk , 1 = low risk , 2 = high risk

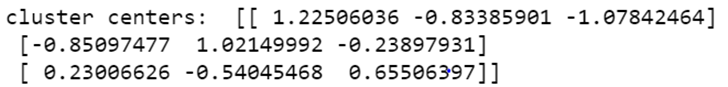
Text, letter

Description automatically generated4 countries are in high risk, 70 countries are in medium risk, and 47 countries are in low risk

**When n\_init = 5,**

0 = high risk, 1 = low risk, 2 = medium risk

26 countries are in high risk , 49 countries in medium risk, 46 countries in low risk

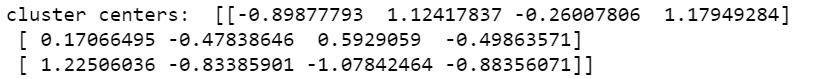
**When n\_init = 10 or greater ( 11, 12)**

0 = high risk , 1 = low risk , 2 = medium risk

22 countries are in high risk, 46 countries in low risk , 52 countries in medium risk

As n\_init increase between 1- 9 , the countries in the high risk cluster increases. However, the amount of countries eventually stabilizes at 22 when the algorithm is ran 10 times or more.

**(C) Comparing 4 features with 3 features both with n\_init = default**

When there are 4 features ,

0 = low risk, 1 = medium risk , 2 = high risk

There are 41 countries in low risk , 58 countries in medium risk , 22 countries in high risk

The quantity of the countries in the high risk cluster does not change because the feature that was added back (Corruption index) is strongly correlated with one of the three features ( Legal index). When 2 features are strongly correlated with each other ( a correlation of close to 1) , then adding the correlated feature back does not change the cluster centers by much.

Text

Description automatically generated**(D) Using hierarchical clustering and compare results with k means**

When using linkage = ward

0 = high risk, 1 = low risk, 2 = medium risk

55 countries in the high risk cluster

Text

Description automatically generatedWhen using linkage = complete

0 = medium risk, 1 = high risk, 2 = low risk

59 countries in the high risk cluster

Text

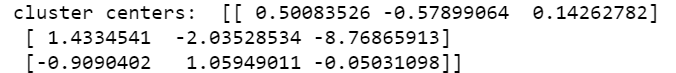
Description automatically generated When using average = average

0 = high risk, 1 = low risk , 2 = medium risk

46 countries in the high risk cluster

Based on the comparison between the hierchical clustering approach and the k means approach, we can see the hierarchical clustering method return much more countries than the k means method for the high risk cluster.

(E) Insert Venezuela’s index information and analyze outliers

When Venezuela’s data is inserted into the original raw data, the cluster centers changes to

Table

Description automatically generatedIn this case, 0 = medium risk , 1 = high risk, and 2 = low risk.

And when we check how many countries fall into the high risk cluster , only one country falls into the high risk cluster , which is Venezuela. Based on this change, we can conclude that k means is very sensitive to outliers. The reason why k means is very sensitive to outliers is because its algorithm relies heavily on the calculation of means. An outlier will affect the mean of a data set significantly. So when mean of the data set is affected, the clustering process will also be affected.