Country Risk Case Study

The aim of this analysis is to use different methods of clustering, specifically k-means and agglomerative hierarchal clustering, to cluster countries dependent upon their risk profile in terms of foreign investment. The following measures are considered - GDP growth rate, Corruption Index, Peace Index, Legal Risk, and we look at 122 countries.

We use z-score scaling by subtracting the observation by the mean and dividing the result by its standard deviation. The mean comes out to be zero and has a standard deviation of 1. This allows to have a similar distribution of data across the features and hence, allow for clustering. We see that the correlation between Corruption and Legal Index is 0.94, which means that they are almost perfectly correlated and hence, we do not consider Corruption Index in our initial analysis.

Starting with k-means clustering, we will use 3 clusters. The k-means algorithm looks to identify centres of these clusters. The default algorithm (which runs 10 times), produces the following centroid seeds as its best output:

```
                   Peace        Legal    GDP Growth
High Risk      [ 1.22506036 -0.83385901 -1.07842464]
Moderate Risk  [-0.85097477  1.02149992 -0.23897931]
Low Risk       [ 0.23006626 -0.54045468  0.65506397]
```

We see that risk for foreign investment and GDP growth are positively correlated. Now, when we look to change the number of times the algorithm runs, we see that when we run the algorithm twice, we get 70 countries as high-risk countries. As we increase that number to 10 and above, we see the same 22 countries being included in the 'high-risk' cluster. As the algorithm will run with different centroid seeds, the result will be the best output in terms of inertia and hence, this points to the fact that more the number of times the k-means algorithm runs, the more accurate results will be.

Let’s look at k-means clustering with all four features.

```
                 Peace        Legal      GDP Growth   Corruption
High Risk      [-0.89877793  1.12417837 -0.26007806  1.17949284]
Moderate Risk  [ 0.17066495 -0.47838646  0.5929059  -0.49863571]
Low Risk       [ 1.22506036 -0.83385901 -1.07842464 -0.8835607]
```

We see that Peace is positively correlated to risk, Legal and Corruption is negatively correlated. This shows that clustering using four features will have different centres, and hence, will cluster countries differently. Comparing the lists of the countries that are included in the ‘high-risk’ cluster, we see two completely different lists. Clustering using all four features flag 41 countries as high-risk and cluster countries differently dependent on their values. In the initial clustering, with three features we saw a relation with GDP growth, and hance, can infer that clustering was dependent upon that feature (higher the GDP growth value, lower te risk profile of the country). On the other hand, clustering with four features saw relations with the other indices (countries with high Peace values and, low legal and corruption indices).

We now use agglomerative hierarchical clustering to identify ‘high-risk’ clusters and compare it with the clusters from our initial k-means clustering with three features. We also use different linkage criteria to cluster the data. This determines the distance to use between sets of observations. ‘Ward’ minimizes the variance of the clusters being merged and provides 73 ‘high-risk countries; ‘Average’ uses the average of the distances between clusters and provides 81 countries; ‘Complete’ uses the maximum distances between all observations and provides 106 countries; and lastly, ‘Single’ uses the minimum of the distances between all observations and provides only 3 countries.

Considering the first three measures of clustering, we see that 25 of the countries are included in all three ‘high-risk’ clusters provided by the three linkage methods. When we compare that list to the k-means’ list, we see that, Ecuador, Liberia, Paraguay are the countries that are present in both those lists. This provides the inference that the countries mentioned above are likely to be in the ‘high-risk’ cluster.

We will conclude by including an outlier into the data and observe how it affects the k-mean analysis. The clustering includes 77 countries in the high-risk cluster and have no observable relation among the features and risk. The value of inertia reduces as well. Comparing the two lists, we see that all 22 countries that were identified in the earlier three feature clustering without Venezuela. The inference that can be derived from this is that including outliers will confound the values of the centroid and may not provide accurate results. This is also because k-means clustering is highly sensitive to outliers and can highly influence the clusters.