Overview of Analysis

The analysis uses 121 samples of countries for clustering analysis focusing on variables contributing the overall risk profile of a country (e.g., legal index, peace index, GDP growth rate). Given the nature of unsupervised learning, the results need additional interpretation to gain a more meaningful categorization. The objective is to increase the accuracy of clustering and predictive ability.

The cluster centers are 3-D vectors. Intuitively, the sum of all entries would indicate the risk level. The higher the sum, the lower risk there is, given that most variables are negatively correlated with the risk, except for corruption. There can also be modification such as creating some form of linear or non-linear combination of these variables that would give us an index of risk which is quantifiable and comparable.

As the data is in a clean format, no significant cleaning is required. The data containing measurements of corruption, legal, peace, and GDP growth is normalized for the purpose of clustering. In most cases, we exclude the “corruption” variable, the only one that’s positively correlated with risk. Therefore, the cluster centers that are more “negative” would indicate that the country is not robust and is riskier. Riskiest is taken to be the smallest sum of cluster center’s entries, using sorted(key=sum) method.

Task 1: Impact of Iteration Number and Clustering Results

The impact of the number of clusters (k) directly changes the number of categories of risk, n_init changes the number of iterations in the clustering function. Initially we have $k = 3$, which can result in the centroids of (high-risk, medium-risk, low-risk). Intuitively, increasing the iteration number should most likely yield a more accurate result of clustering. The findings shows that inertia and Silhouette Score converge to a certain point (161.13 and 0.356). Form the point where $n\_init = 10$, additional iterations would not increase accuracy significantly. Surprisingly, Silhouette score is highest when $n\_init = 2$, which is around 0.38, though inertia was 168.24 (See Appendix). Convergence is expected as for k-means clustering, centroids are initially arbitrarily set, and then iterated based on the relative distances of points in proximity. Initially, high-risk centroid is (0.71, -0.96, -3.44) and converges towards (1.23, -0.83, -1.08), where risk evaluation is more holistic and not concentrated on GDP growth.

Task 2: 4D vs 3D clustering (Adding the Corruption Variable)

With iteration numbers set to default (10), and number of clusters, k set to 4, inertia is approximately 194.4. This is higher than the inertia for the result of using 3 clusters which may be caused by the fact that corruption index is negatively correlated with the other metrics and positively correlated with risk. Therefore, the measurement of inertia using Euclidean distance would naturally be greater. On the other hand, the Silhouette Score has improved where for $k = 4$, the score is 0.42, compared to $k = 3$, where the score is 0.356. This suggests that using 4D cluster will yield a better result.

The definition of a “high-risk” country also changes between 4 clusters and 3 clusters. Considering $C_{Highrisk}$ as the cluster center for high-risk countries where $C_{Highrisk} = [Corruption, Legal, Peace, GDP\ Growth]$. The main difference between the two cluster centers is that for the cluster where $k = 4$, Legal score and peace scores are -0.9 and 1.12, whereas for $k = 3$, its 1.23 and -0.83, which are contradicting (See Appendix). This suggests for the two different models, the determinants of a country’s risk are different. In this case, we could say that the model yielded by $k = 4$ is more accurate, meaning that the legal index is major determinant of a country's risk. A final judgement needs to be made when choosing the model, it ranges from using the heuristics elbow method to optimization of the Silhouette Score.

Task 3: K-means vs Agglomerative Clustering (k=3)

Agglomerative Clustering is also used to generate the clusters of high, medium, and low risk countries (n clusters = 3). The clusters generated is more accurate considering that the silhouette scores are higher than using k-means (for all numbers of iterations from 2 to 100) across all linkage methods (i.e., ward, average, complete, single). For the agglomerative clusters, the number of high-risk countries using "ward" method is 15, then 106 across all other methods. "Average" and "Complete" method yield the highest Silhouette Scores (i.e., 0.481) (See Appendix). For k-means method, highest silhouette is yielded using n=2 iterations, with Silhouette Score being 0.38, with 72 high-risk countries. The number of high-risk clusters is much lower for the other methods (See Appendix). The high-risk centroid for agglomerative cluster is like k-means using ward method. However, with other methods, high-risk countries tend to have lower growth, centroid GDP growth is roughly -3.4% (See Appendix)

Task 4: Impact of Outliers on Clustering Results

By fixing the parameters of the clustering function and adding Venezuela into the dataset, the result yields higher accuracy of clustering. The overall inertia is 147.36, Silhouette is approximately 0.56, higher than all previous methods. These are all positive indicators of the clustering result so far. However, the implied high-risk country dataset only includes Venezuela given its extreme values takes the high-risk cluster to a very extreme point. This radically shift of the high-risk cluster center, and the numerical definition of "high-risk" countries. The inference yielded is that sensitivity of clustering result by adding outlier data points is significant.

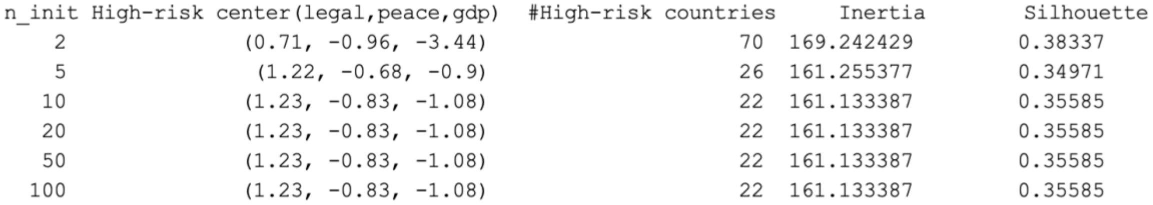
By including Venezuela, high-risk cluster center is at (Peace=1.43, legal=-2.03, GDP Growth = -8.77), compared to the previous (1.23, -0.83, -1.08) from k-means with the same remaining parameters. Qualitatively, the interpretation of this might not be meaningful as we are trying to have a group of countries based on risk levels. Having the extreme high-risk cluster group leads to unmeaningful results as a lot more countries are indicated to be either medium or low risk. The predictive ability of the clustering method is thus reduced. If we a new dataset of a moderately risky country, it would very likely be grouped into either medium or low risk.

Overall Findings and Conclusions

The additional "tweaks" of parameters and methods (e.g., using 4D clusters, adding outliers) increased accuracy, considering metrics such as Silhouette Score and inertia. However, there is a lack of consistency for the cluster centers such that the numerical definition of a high-risk group suffers variability with different approaches of clustering. Therefore, given the nature of unsupervised learning and unlabeled data, extra interpretation and determination of these results' purpose is needed to select the "best" method. For instance, solely put Venezuela in the high-risk group is reasonable given its relative extremeness. But that in itself does not offer significant insights.

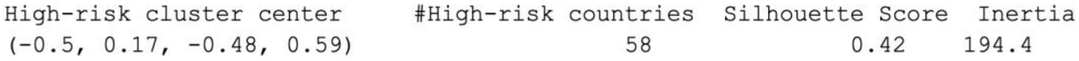
Appendix: Clustering Results with Different Parameters

K-means Clustering Result vs Number of Iterations (k=3)



```
 n_init High-risk center(legal,peace,gdp)  #High-risk countries     Inertia        Silhouette
     2              (0.71, -0.96, -3.44)                      70  169.242429        0.38337
     5               (1.22, -0.68, -0.9)                      26  161.255377        0.34971
    10              (1.23, -0.83, -1.08)                      22  161.133387        0.35585
    20              (1.23, -0.83, -1.08)                      22  161.133387        0.35585
    50              (1.23, -0.83, -1.08)                      22  161.133387        0.35585
   100              (1.23, -0.83, -1.08)                      22  161.133387        0.35585
```

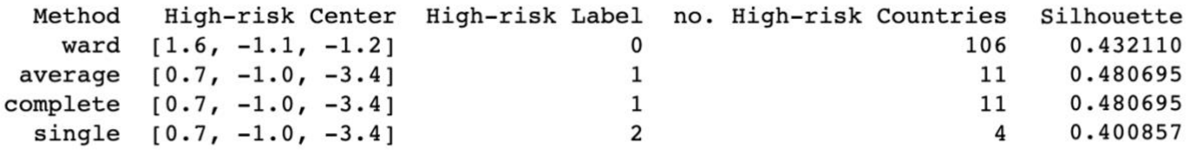
K-means Clustering Result (variables=4)



```
High-risk cluster center     #High-risk countries  Silhouette Score  Inertia
(-0.5, 0.17, -0.48, 0.59)                      58              0.42    194.4
```

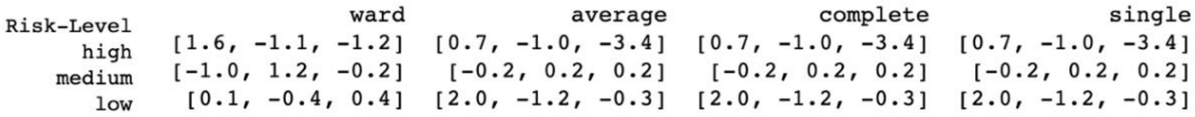
Cluster Center = [corruption, legal, peace, GDP growth] (Normalized)

Agglomerative Clustering Result vs Linkage Methods



```
  Method   High-risk Center  High-risk Label  no. High-risk Countries  Silhouette
    ward  [1.6, -1.1, -1.2]                0                      106    0.432110
 average  [0.7, -1.0, -3.4]                1                       11    0.480695
complete  [0.7, -1.0, -3.4]                1                       11    0.480695
  single  [0.7, -1.0, -3.4]                2                        4    0.400857
```

Agglomerative Clustering Centers vs Linkage Methods



```
                          ward            average           complete             single
Risk-Level
      high  [1.6, -1.1, -1.2]  [0.7, -1.0, -3.4]  [0.7, -1.0, -3.4]  [0.7, -1.0, -3.4]
    medium  [-1.0, 1.2, -0.2]   [-0.2, 0.2, 0.2]   [-0.2, 0.2, 0.2]   [-0.2, 0.2, 0.2]
       low   [0.1, -0.4, 0.4]  [2.0, -1.2, -0.3]  [2.0, -1.2, -0.3]  [2.0, -1.2, -0.3]
```

Cluster Center = [Legal, Peace, GDP Growth] (Normalized)