Kanishaka Sharma – Assignment 2 Report (#1007900229)

From the iterated values of n_init, from 2 to 100, we can see the high-risk cluster vary immensely. A value of two means only two iterations to determine the optimal cluster, so it would be the most likely to converge to a suboptimal solution or a local instead of a global minima. However, even comparing the high risk cluster for 20 and 50 iterations shows massive in-cluster variance, with the high-risk countries with 20 iterations being classified as medium or low risk in the 50 and 100 cluster iterations. This suggests a stable solution isn't being found with more iterations, suggesting the features used might not be the most effective in distinguishing between high and medium risk countries. Even from the visualizations for the basic 3-feature analysis, there is significant overlap in cluster assignment, with countries like Paraguay or Israel being clustered within the high risk-groups even while being classified as medium risk. This suggests the possibility of more features being appropriate for a more accurate analysis, given that radically different countries are assigned the same cluster with high within-cluster sum of square values. More features would allow for more dimensions to be analyzed to determine appropriate clusters, which should potentially lead to less instability in the high-cluster countries.

Moving on to part C, we can see the difference in a country's risk profile depending on whether corruption was or was not included in the analysis. Countries like Argentina, who are in the high-risk cluster for both feature sets, seem to have strong indicators like negative GDP growth and legal score that align them with higher risk. The inclusion of corruption would likely just have strengthened this relationship rather than providing any new information. Most countries, with n_init being constant across varying features, tend to fall into the same cluster. This provides evidence against the claim that more features should be used, given that the cluster variance is almost minimal between different features levels. The inclusion of other features that are uncorrelated with the existing ones might provide more of a benefit, but the inclusion of corruption did not make much of a difference in the conclusions already drawn from this dataset.

In part D, we use a new clustering method called Agglomerative clustering, which builds clusters bottom-up by identifying which individual data points have the smallest distance measure and aggregating clusters together until the desired number is reached. There are different measures for aggregating clusters, such as ward and complete linkage, all of which were used independently for comparison. Looking at the results, the previous variance and instability in clusters is still present, though this is to be expected from the different distance metrics. An interesting observation, however, is the consistency of Albania in the high-risk cluster through different distance metrics. Through ward, complete, and average, Albania is considered high risk, compared to middle risk using kmeans clustering and low risk using single linkage. This suggests that Albania's metrics are very robust in agglomerative clustering, but the overall variability using different clustering methods still indicates a less obvious structure to the data.

Another interesting observation is that the high-risk countries identified using the Ward method are extremely diverse, both geographically and economically. This would suggest there are core similarities between these countries that might be obscured when using other linkage or clustering methods. Given that the ward metric analyses the variance between clusters before merging, compared to measuring the distance directly, there seems to be hidden similarities that require further analysis. To expand on this and investigate further, I would want to redo this clustering method with more features to see if this pattern still holds or if the results shown are just spurious correlations that don't hold much weight.

In part E, Venezuela was added to see if the presence of this outlier would affect the results of the previous analysis. Initially, after comparing the inertia value with the initial analysis without Venezuela, this model would seem to be more accurate given the lower value of 147 compared to 161. When comparing agglomerative clustering methods against previous analysis, the results also seem to be more consistent. Ward and complete linkage methods result in the same immediate clustering, with greater variance with the average and single linking as before. However, when we look at the visualization, it's clear to see that the inclusion of this outlier has dramatically altered the structure of these clusters. Venezuela has its own high-risk cluster that eliminates any other countries from this analysis. Given the lower legal and GDP growth values, the distance between Venezuela and any other country is too large for the cluster to include more countries, even after scaling to standardize the magnitude of different values.

Looking at the graphs without Venezuela, countries tend to be spread more evenly across the full area of the graph, with the clusters being more or less accurate to the three major groups. Looking at the distribution of peace and legal, countries in upper left quadrant tend to have low to medium peace and a high legal score, suggesting low-risk countries that are relatively stable. The middle of the graph is a mix of different countries from all clusters, suggesting a more complex risk profile, which would also explain the variation in clustering depending on the method used, since most countries in the dataset tend to shift between two clusters. By comparison, the same peace and legal graph with Venezuela included has the same number and coordinates of clusters, but Venezuela is categorized as the only high risk country. This indicates the sensitivity of kmeans clustering to outliers, with it seeming like Venezuela is pulling the cluster center towards itself and skewing the countries to be categorized as lower risk when compared to Venezuela. For the purposes of this analysis, it would make more sense to isolate a few countries with a similar risk profile or just examine Venezuela's specific feature space independently. Venezuela's high economic distress, as indicated by its extremely low GDP growth, distorts the clustering model and unbalances the risk analysis of different countries. This effect emphasizes the need for more robust clustering techniques or separate analysis for countries that have an extreme risk profile