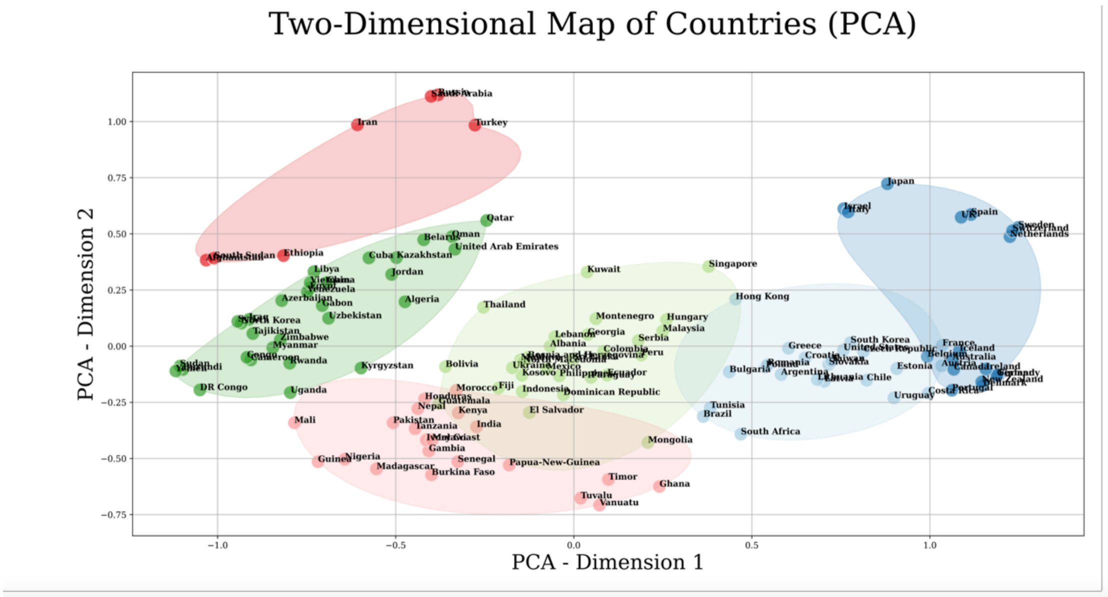
Short Answers:

1.

a. The dataset is about some evaluations of situations of different countries. The first variable is the names of the country. The second variable is the gini coefficient of the countries, it measures the extent to which the distribution of income among individuals with an economy deviates from an equal distribution. It has range from 0 to 100. The third variable is corruption perceptions index, which measures how corrupt their public sectors are perceived to be. It has range from 0 to 100. The fourth variable is freedom of the house. The range is 1 to 3. The fifth variable is hdi, which measures human development index. The range is 1 to 5. The fifth variable is press of freedom. The range is 1 to 5. The sixth variable is democracy economist. It has range from 1 to 5. The seventh variable is populism. The range is almost from 1 to 10. The eighth variable is effective coverage of health services index. The range is 0 to 100. The ninth variable is trust in news media. The range is from 0 to 100. The tenth variable is trust in government with range 0 to 100. The eleventh variable is trust in science with range 0 to 100. The last variable is colonized with 0 or 1.

4.



5.

a. For example, Uzbekistan, Gabon, Zimbabwe, Myanmar, Jordan are in the green cluster; Iran, Ethiopia, Turkey are in the red cluster; Albania, Georgia, Colombia, Malaysia are in the light green cluster; Gambia, Madagascar, Faso, Timor are in the light red cluster; Greece, South Africa, Brazil, Chile are in the light blue cluster; Japan, Italy, Spain, Netherland are in the blue cluster. I think they are clustered together based on overall how develop the country is and how good the life quantity is in the country.

b. Yes, the result will change. If we randomize different point to be the first center, the nearest distance from each point to the center will change. Then their weights to be selected to the next center will change. Thus, the initialized centers will change. When we do classifications, the elements in clusters will change due to the different centers. So every clusters' average will change, and thus influence the results.

c. I' m not sure this is the very accurate result, since there may be significant variables that we didn' t count them in. If we count them in, there should be changes in the cluster. But these clusters show some patterns, and they are overall constant with our common beliefs. Thus, I can say it is an accurate result, but changes may incur.

d. I don' t think PCA has affected the results overall, but there should be some differences. Since PCA is used when we have high-dimensional data, and it helps us to improve the accuracy of classification because of the simplification of the structure. If we have had a different number of principle components, our interpretation will not be too much different.