**Population Attitudes – Clustering Lab**

Numerous studies have attempted to analyze attitudes of populations, and countries which have the most content and happiest population. Typically, the happier the population, the more productive the workforce is, the higher the investment into the community is, and the stronger the loyalty is to the country. As such, having a happier population is a very important calculation for countries to make.  
  
As might be expected, it is difficult to accurately calculate a countries level of happiness. A number of efforts have been made to use various drivers to draw conclusions on which countries are happiest.

The data set **Clustering Lab.jmp** contains this database. In general, the higher the values, the better they were perceived. The dataset contains the variables below:

**Country** – The name of the country

**Score** – The sum of Investment Potential, Diet, Freedom of Press, Recreation, Education System, Healthcare, Infrastructure Development, Currency Stability, Transportation, Political Stability

**Investment Potential** – Perceived attractiveness of investing into the country (0.00 – 1.00)

**Diet** – Perceived quality of the diet of the population (0.00 – 1.00)

**Freedom of Press** – Perceived freedom of press (0.00 – 1.00)

**Recreation** – Perceived ability to engage in recreational activities (0.00 – 1.00)

**Education System** – Perceived quality of education (0.00 – 1.00)

**Healthcare** – Perceived accessibility to healthcare facilities (0.00 – 1.00)

**Infrastructure Development** – Perceived level of infrastructure development (0.00 – 1.00)

**Currency Stability** – Perceived confidence in currency (0.00 – 1.00)

**Transportation** – Perceived accessibility to transportation (0.00 – 1.00)

**Political Stability** – Perceived political stability (0.00 – 1.00)

Using k-Means clustering, answer the following questions:

1. If the data set asked for respondents to explain their favorite feature of a country (e.g. mountains, beach, activities etc. etc.) would this be helpful for us to cluster the data?

No, the criteria for clustering is to use the variable that is stable, easily accessible, and numerical, so the favorite feature of a country which is not numerical would not be helpful to cluster the data.

1. Using a cluster model with 6 clusters, which cluster has the highest score on average?

Cluster 6

1. What is the Cluster Means of this cluster?
2. Does this cluster have the highest number of data points?

No, cluster 6 has 22 data points. Cluster 4 has the highest number of data points of 37.

1. If someone valued Investment Freedom of Press and Education Systems, which cluster of countries would they want to live in?

Cluster 6

1. Let’s assume that we are wanting to use the statistical best number of clusters. If we used a range of clusters (3-9), which number of clusters would be best to use? Why?

3, because it has optimal CCC.

1. 7 Cluster, 8 Cluster and 9 Cluster models all have clusters with 1 data point in them. What does this mean?

It means this 1 data point is so unique that cannot be group into any of the other clusters, which forms its own cluster.

1. Do the number of clusters matter? For example, explain in simple terms what it would mean if we used 3 clusters as opposed to 23 clusters.

Yes. The reason we want to cluster the data into a set of groups is that we want to find similarities among the categories. Hence, the smaller number of the clusters is better in order to identify simpler similarities to interpret. The bigger number of the clusters will become harder to interpret the character of each cluster.

For questions 9, 10, 11 use hierarchical clustering (with all values except ‘Country’)

1. Look at difference in distance between going from a 3 cluster model to a 4 cluster model, compared to going from a 4 cluster model to a 5 cluster model. Based on your understanding of distance and length of the lines in the dendogram, which model (3 clusters or 4 clusters) would be best to use? Why?

4 clusters model, the longer the line, the more the distance that exists, so the huge distance between 3 clusters and 4 clusters model means 4 clusters model separates information out a lot more and is more valuable. In addition, comparing to the distance between 4 clusters and 5 clusters model, the further jump between 3 clusters and 4 clusters model indicates 5 cluster model is not that necessary.

1. Create a model with 5 clusters. What is the cluster means of this model?

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | **Column Labels** |  |  |  |  |  |
| **Values** | **1** | **2** | **3** | **4** | **5** | **Grand Total** |
| Average of Score | 6.654927743 | 5.690829 | 4.966572 | 4.118809 | 3.977101 | 5.038193249 |
| Average of Investment Potential | 76.91% | 40.38% | 60.25% | 34.05% | 23.54% | 0.528646696 |
| Average of Diet | 81.52% | 50.67% | 68.80% | 39.26% | 22.85% | 0.587916149 |
| Average of Freedom of Press | 87.36% | 76.10% | 60.23% | 25.13% | 60.89% | 0.625259513 |
| Average of Recreation | 48.54% | 54.47% | 19.91% | 34.07% | 28.03% | 0.296442708 |
| Average of Education System | 87.36% | 76.10% | 60.23% | 25.13% | 60.89% | 0.625259513 |
| Average of Healthcare | 83.19% | 51.70% | 70.20% | 40.06% | 23.32% | 0.599914438 |
| Average of Infrastructure Development | 48.54% | 54.47% | 19.91% | 34.07% | 28.03% | 0.296442708 |
| Average of Currency Stability | 50.63% | 72.73% | 47.87% | 51.42% | 49.52% | 0.503088323 |
| Average of Transportation | 43.99% | 72.92% | 42.06% | 69.42% | 52.50% | 0.486079624 |
| Average of Political Stability | 57.45% | 19.55% | 47.19% | 59.28% | 48.14% | 0.489143576 |

1. Are there any areas in which the cluster with the lowest score performs better in, compared to the cluster with the highest score?

Yes, the average of transportation is performed better in the cluster 5 with the lowest score than cluster 1 with the highest score.

For the remaining questions, use hierarchical clustering (with all values except ‘Country’), and create a model using 8 clusters

1. A data analyst suggests that if a country wants happier residents, they should invest heavily in healthcare, as this leads to a higher level of investment potential, and therefore happier residents. Do you agree?

No, I don’t agree. Although there seems to be a correlation between investment potential, healthcare and happiness, it doesn’t mean there is a casual relationship.

1. Which cluster contains the lowest average score?

cluster 7

1. Fill out the following table for this cluster:

|  |  |
| --- | --- |
| Average of Diet | 22.39% |
| Average of Freedom of Press | 16.84% |
| Average of Currency Stability | 48.65% |
| Average of Transportation | 46.67% |

1. An analyst suggests that the lower the number of clusters, would give us the strongest and most meaningful result, as opposed to a large number of clusters. Comment on whether this suggestion is valid, by using a 1 cluster, 15 cluster, and 152 cluster model as a basis for comparison.

No, the suggestion is not valid. If there is only 1 cluster, then we cannot achieve our goal of separating data. If there are 152 clusters, then there would be mostly only one data point in one cluster, which is not the result we want. Since we want to cluster the data into a set of groups is that we want to find similarities among the categories, 15 clusters model would be the best in this situation, which we could have meaningful inference. Thus, the suggestion that the lower the number of clusters didn’t draw the strongest and most meaningful result we want and is not valid.