**JEREMY PETERSON DUTOYA**

**PR. MENG-SHEN HSIEH**

**PMBA-8358-OLA: DATA-DRIVEN STRATEGIES FOR BUSINESS**

Clustering Homework

**Objective:** Use R Data Miner to implement K Means Clustering to segment countries using several country attributes.

**Instructions**

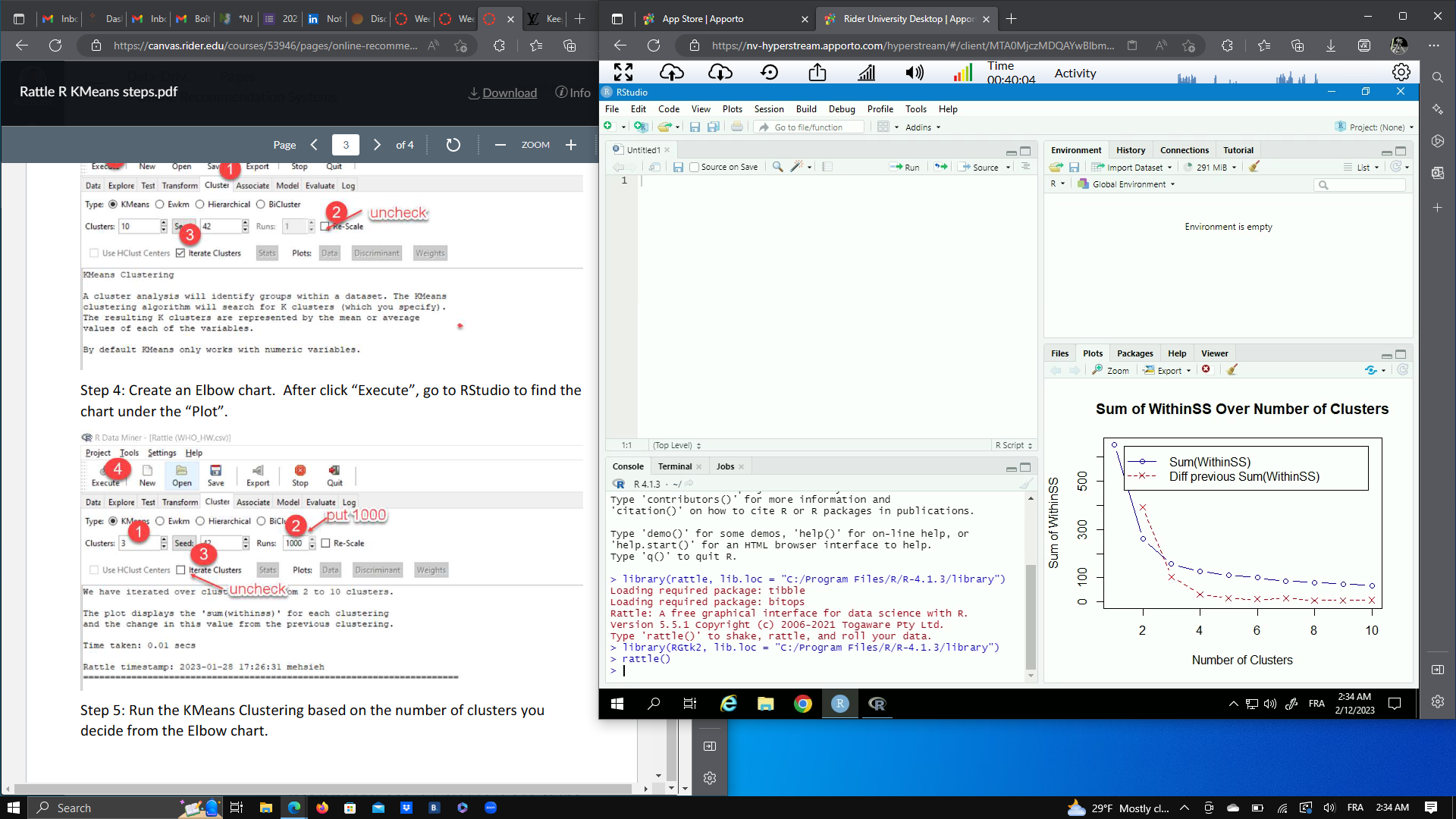
1. Upload the WHO\_HW.csv file to your Apporto local drive.
2. Launch the rattle R.
3. Use the following input variables to perform the K Means Clustering:

| 1. Under15 |
| --- |
| 1. Over60 |
| 1. FertilityRate |
| 1. LifeExpectancy |
| 1. CellularSubscribers |

Assign these variables as “Input” and the remaining variables “Ignore.

1. Standardize these variables

1. *Since we do not know how many clusters should be used for splitting the countries, we decide to create an* ***Elbow plot*** *and let the data helps us determine the number of optimal clusters.*



*Create an Elbow plot. Copy and Paste the plot here. Determine the number of clusters you will use. Justify your answer.*

I paid close attention to the videos you have added in the modules tab in order to understand how to determine the optimal number of clusters for our K Means clustering. In particular, the 3 first videos are very important. From these videos and additional research, I understood that the idea behind the elbow method is to fit the k-means clustering algorithm for different values of k, and for each value of k, calculate the sum of squared distances of samples to their closest cluster center (also known as the “within-cluster sum of squares”: WCSS).

The “optimal” number of clusters is determined by finding the “elbow” point on a graph of the WCSS versus the number of clusters, k. The “elbow” point is the point of inflection on the graph: that is where the WCSS starts to decrease at a slower rate. Based on your explanations and the Elbow Plot obtained (see above), we can see that the variation becomes smaller and more stable after the 3rd point, identifiable as the elbow of the graph: so the number of adequate clusters to use for our clustering analysis is 3.

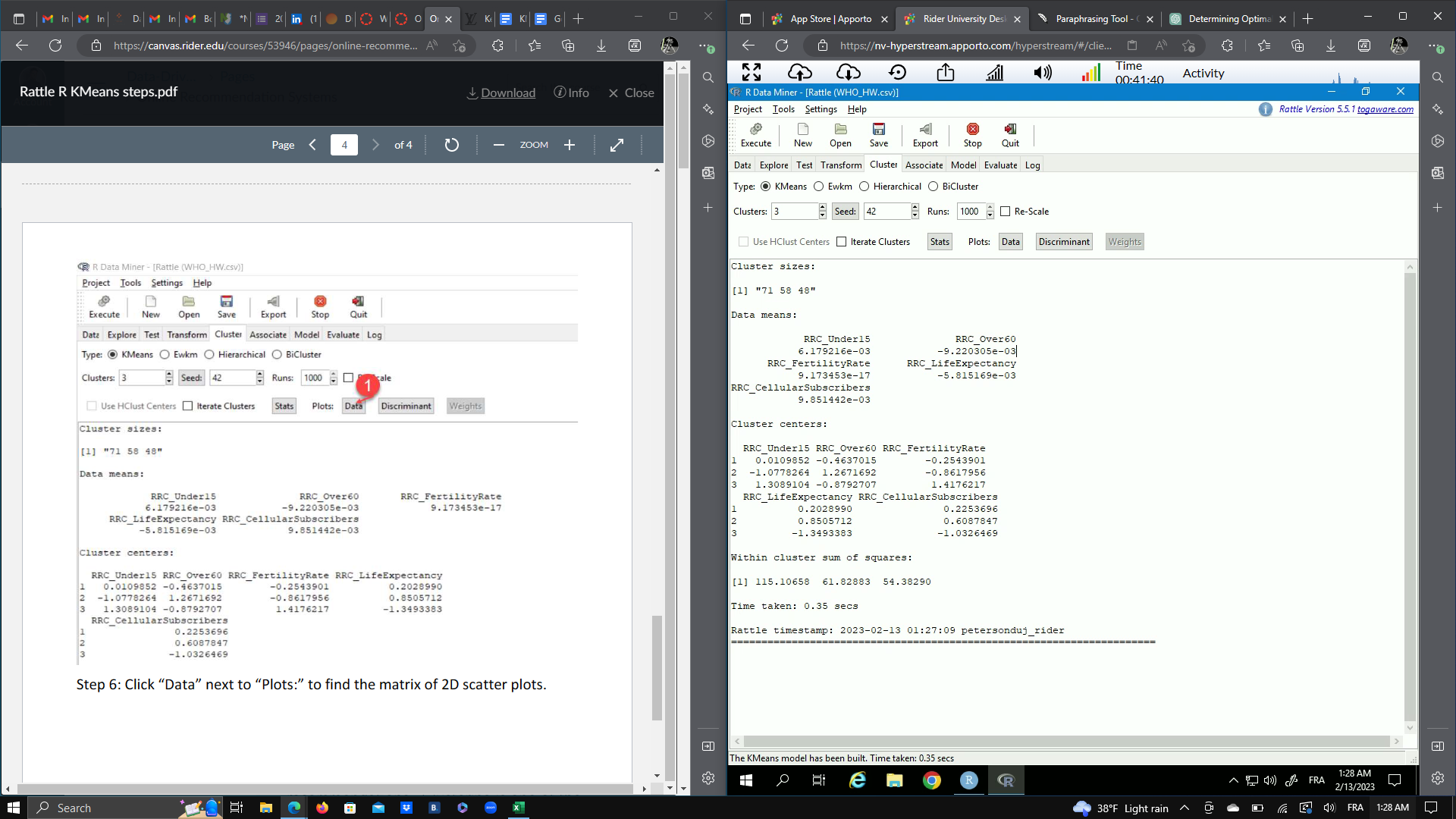
1. *Use the number of clusters you determined from the Elbow plot, and perform the K Means Clustering. How many countries are in each cluster?*

In Cluster 1 there are 71 countries.

In Cluster 2 there are 58 countries.

In Cluster 3 there are 48 countries.

See the attached image to justify my answer:



1. *Provides the centroid of each cluster. Each centroid contains five values, which are the average values of each input variable in a cluster.*

*Your answer should look like this:*

|  | RRC\_Under15 | RRC\_Over60 | RRC\_FertilityRate | RRC\_Life Expectancy | RRC\_CellularSubscribers |
| --- | --- | --- | --- | --- | --- |
| Cluster 1 | 0.0109852 | -0.4637015 | -0.2543901 | 0.2028990 | 0.2253696 |
| Cluster 2 | -1.0778264 | 1.2671692 | -0.8617956 | 0.8505712 | 0.6087847 |
| Cluster 3 | 1.3089104 | -0.8792707 | 1.4176217 | -1.3493383 | -1.0326469 |

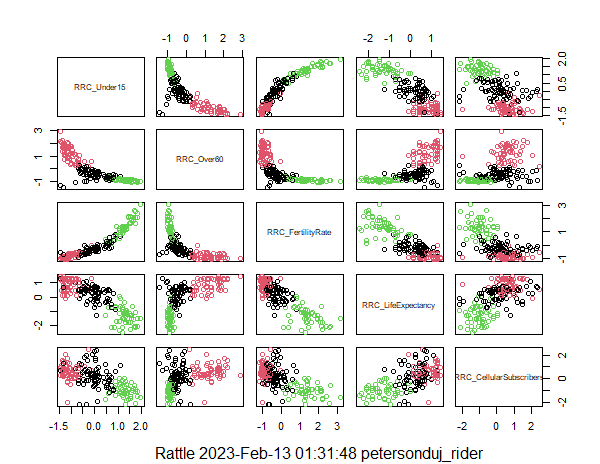
1. *Use the output from Q3, comment on the characteristics of each cluster. For example, your answer may look like*

|  | RRC\_Under15 | RRC\_Over60 | RRC\_FertilityRate | RRC\_Life Expectancy | RRC\_CellularSubscribers |
| --- | --- | --- | --- | --- | --- |
| Cluster 1 | High | Low | Low | High | High |
| Cluster 2 | Low | High | Low | High | High |
| Cluster 3 | High | Low | High | Low | Low |

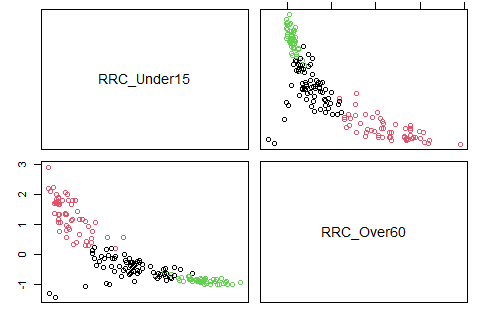
Based on the table, countries in Cluster 1 has a larger young population than the old population. They also have a higher Life Expectancy and population of Cellular Subscribers but a Lower Fertility Rate.

Based on the table above, countries in the Cluster 2 has a larger older population than a young population. They also have a Low fertility rate but a high Life Expectancy and a larger population of Cellular subscribers.

Based on the table above, Countries in the Cluster 3 have a larger young population than a old population. They have a high fertility rate but a low life expectancy and a smaller population of Cellular Subscribers.

1. *Create a matrix of 2D scatter plots, where each scatters plot shows the relationship between a pair of the standardized input variables.* 

*Each color represents a distinct cluster. Comment on the characteristics of each cluster. For example, a part of the matrix looks like:*

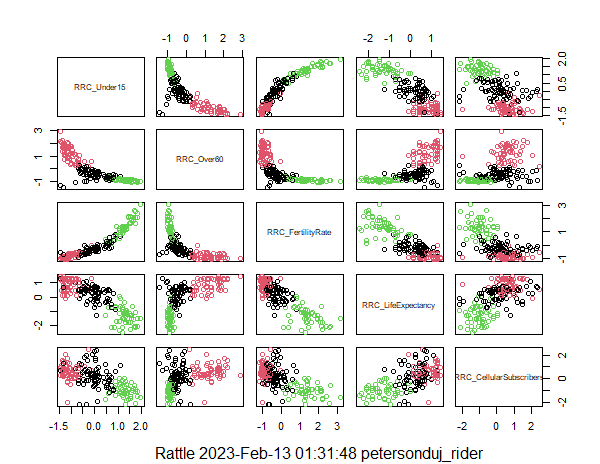


As I explained earlier. Cluster 3 includes countries respecting the following characterics:

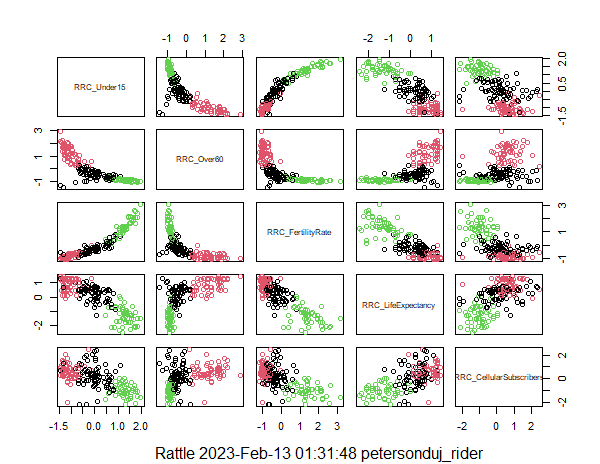
* the population under 15 years is the highest (1.3089104) of the 3 clusters,
* the population over 60 is the lowest (-0.879270) of the 3 clusters,
* the fertility rate is the highest (1.4176217) of the 3 clusters,
* life expectancy is the lowest (-1.3493383) of the 3 clusters,
* cellular subscriber is the lowest (-1.0326469) of the 3 clusters.

As a result, determining the color of this cluster should be easier in the first place because it is always at the extreme ends: either maximum or minimum.

By comparing the young and old populations, we may get a good idea of the color of the third cluster:



As we can see in this portion of the matrix of 2D Scatter plot I built. The green color seems to match. When RRC\_Under15 is the ordinate axis, green is above all others meaning that its population under 15 years is the highest and when RRC\_Over60 is the ordinate, green is below all others meaning that its population over 60 years is the lowest of 3 clusters.

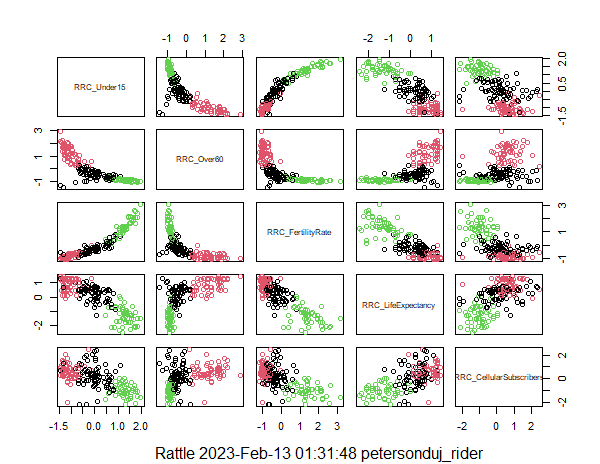
Even if we already have a good reason to conclude that cluster 3 is the green color, we will continue our analysis by checking and comparing the Fertility Rate with the Life Expectancy: 

This shows that the green color is that of Cluster 3. Indeed, when the RRC\_FertilityRate is the ordinate axis the green color is above all the other colors, showing that countries in this cluster have a higher fertility rate than other clusters. Conversely, when the RRC\_LifeExpectency is the ordinate, the green color is below all the others, showing that this cluster has the lowest life expectancy.

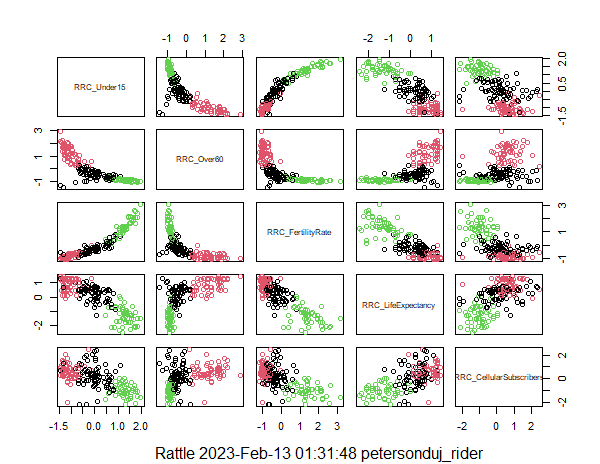
Now following the same reasoning, let’s analyze the color of Cluster 2 whose countries have:

* the population under 15 years is the lowest (-1.0778264) of the 3 clusters,
* the population over 60 is the highest (1.2671692) of the 3 clusters,
* the fertility rate is the lowest (-0.8617956) of the 3 clusters,
* life expectancy is highest (0.8505712) of 3 clusters,
* cellular subscriber is the highest (0.6087847) of the 3 clusters.

Let’s start by comparing the young and old populations, this will allow us to have a good idea of the color of cluster 2:



As we can see in this portion of the matrix of 2D Scatter plot I built. The red color seems to correspond to cluster 2. When RRC\_Under15 is the ordinate axis, red is below all others, meaning that its population under 15 is the lowest, and when RRC\_Over60 is the ordinate red is above all others meaning it has the highest population over 60 years old.

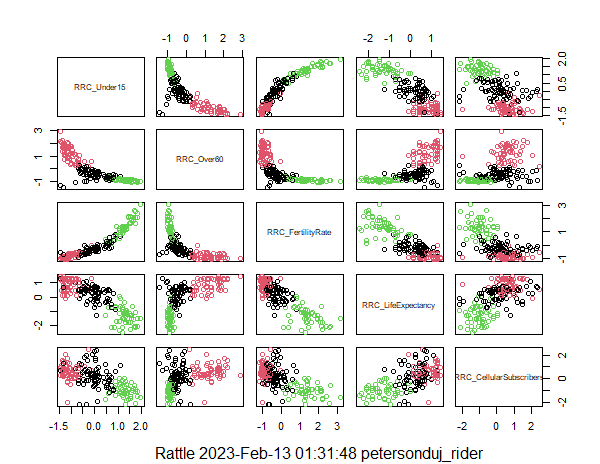
Even if we already have a good reason to conclude that cluster 2 is red, we will continue our analysis by this time comparing the Fertility Rate with the Life Expectancy: 

This shows that the red color is that of Cluster 2. Indeed, when the RRC\_FertilityRate is the ordinate axis the red color is below all the other colors, showing that countries in this cluster or a lower fertility rate than other clusters. Conversely, when the RRC\_LifeExpectency is the ordinate, the red color is above all the others, showing that this cluster has the highest expectancy life.

By Elimination, the black color is that of Cluster 1.

We can however verify this conclusion through a quick analysis, In terms of the figure obtained in Q3, Cluster 1 is always in the middle of the other two clusters. By analyzing the matrix of 2D scatter plots, we can quickly see that the black color is always in the middle in the different scatter plots while the red (cluster 2) and green (cluster 3) colors are always in the extremes and are diametrically opposite.

Let’s now analyze the proximity of the clusters between them. In terms of age, it seems the clusters are very heterogeneous and following a curve. However in terms of Fertility, life expectancy, and Cellular Subscriber, it has a greater mix between the clusters 1 and 2, and delimitation is no longer so clear and cut. For example Life Expectancy and Cellular Subscriber cluster 1 and 2 mix a little bit in each other, we see that the red and black dots are rather separated from the green ones but they mix however a lot being them. In general the clusters 1 and 2 are closer together, in particular in terms of Life expectancy, Fertility and Cellular subscribers and the Cluster 2 and 3 have the highest distance between clusters.



1. *Based on your clustering outcome, which of the following countries are in the same cluster as Canada? What are the characteristics of this clustering in terms of the properties of the input variables?*
2. *Argentina*
3. *Croatia*
4. *Brazil*
5. *Spain*
6. *Nigeria*

Based on my clustering outcome, Canada is in the Cluster 2. The characteristics of this clustering in terms of the properties of the input are that corresponding to countries that observe a larger old population (Canada:20.82) than a young population (Canada:16.37), a low Fertility Rate (Canada:1.66), and a High Life Expectancy (Canada:82) and Cellular subscribers (Canada:79.73). Argentina, Croatia, and Spain are in the Same Cluster as Canada and they observe the same characteristics.

Brazil is in Cluster 1: (High: Under 15, Life Expectancy, Cellular Subscribers / Low: Over 60, Fertility)

Nigeria in the Cluster 3 (High: Under 15 & Fertility Rate / Low: Over 60, Life Expectancy & Cellular Subscribers)