# DOM 207- Introduction to DS using R and Python

## Mini Project 3- Cluster Analysis

Instructor- Dr Jaideep Ghosh

Gargi Kaushik (2210110283)

Rekha Suresh (2210110502)

Reetu Rana (2210120501)

# INTRODUCTION

As Data Science Consultants at CX, we used the data provided to us by the World Bank to analyse the nine individual issues faced by employees in the ICT industry across 37 countries. Our job was to group countries based on the proximity of the magnitude of issues faced by the employees using cluster analysis. We also performed Cluster Analysis on the two different scores and tried to provide reasons for the dis/similarity of the groups formed in both cases. We have justified the groupings keeping in mind various macroeconomic and business factors. Our aim was to highlight the differences and similarities in the problems faced by employees in different countries and understand the reasoning behind the same.

We used the agglomerative hierarchical clustering method and the reasons for the same have been elaborated in the report further.

# DATA PROCESSING

We separated the First and the Second Scores in two different Excel files from the main data file provided to us by the World Bank.

The following cleaning steps were undertaken: -

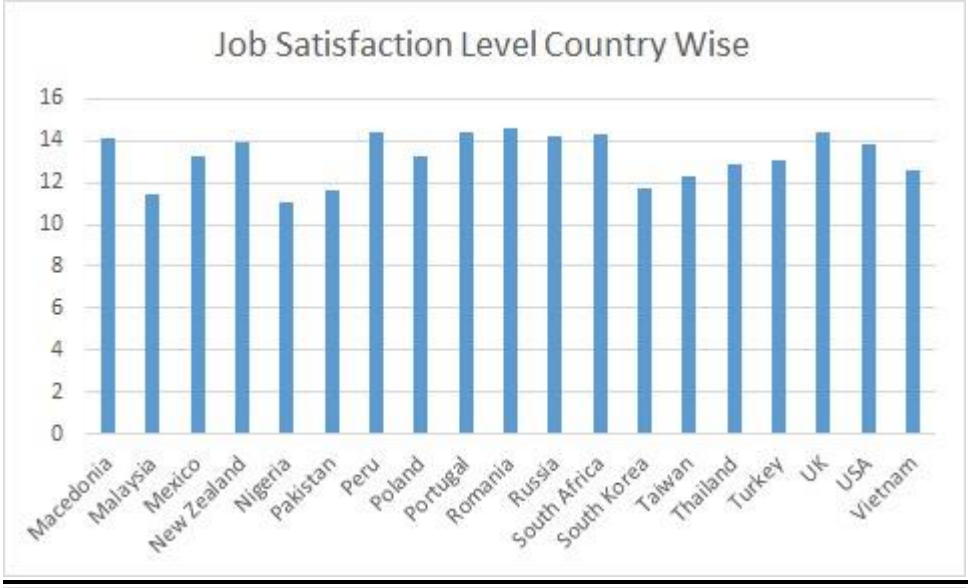
1) There were no missing values.
2) The negative values were accepted as we didn't have information on how the scores were computed.
3) There were no relevant outliers.

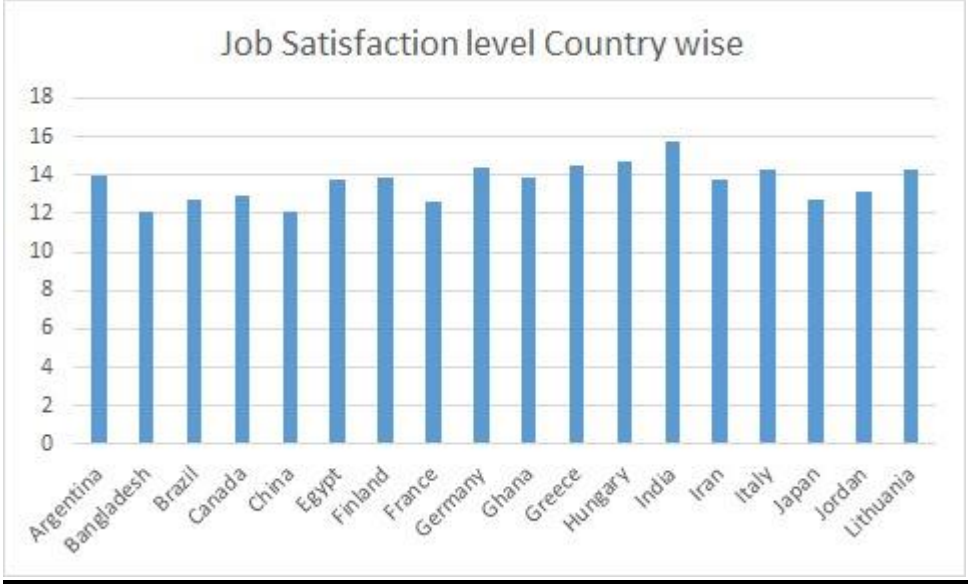
After reading the CSV files of the data in our R session we standardized the data using the scale() function to further continue with Agglomerative Clustering.

We also made sure the data before standardization was a numeric matrix by creating separate data frames for the two scores containing only the numerical values.

# DATA VISUALISATION

Comparing Job satisfaction levels of the countries used in our analysis.



# DATA ANALYSIS

We opted for hierarchical clustering over K means clustering for the following reasons –

1) Since our data set was small and K means is generally suited for larger data sets we considered hierarchical clustering.
2) K means also requires an advanced knowledge of clusters to pre-specify the number of clusters hence we chose hierarchical clustering to ease the process.
3) Moreover, we found that the dendrograms produced in hierarchical clustering were relatively easier to interpret.

We further opted to go for agglomerative hierarchical clustering over divisive clustering due to the following reasons-

1) Since we are dealing with a small data set agglomerative clustering didn't pose any complex computational problems.
2) Further the dendograms produced in agglomerative clustering are generally easier to interpret as they produce smaller clusters.
3) It is also better at handling outliers since outliers don't get absorbed in smaller clusters.

After choosing agglomerative hierarchical clustering we went about selecting the linkage for our clustering process. After undergoing a trial and error process, we picked the complete linkage method for the following reasons-

1) The complete linkage method gave the most clear-cut, tight and isolated clusters
2) It differentiates clusters based on the maximum distance which is better for our analysis since we want to find out employees in which countries face the most similar problems.

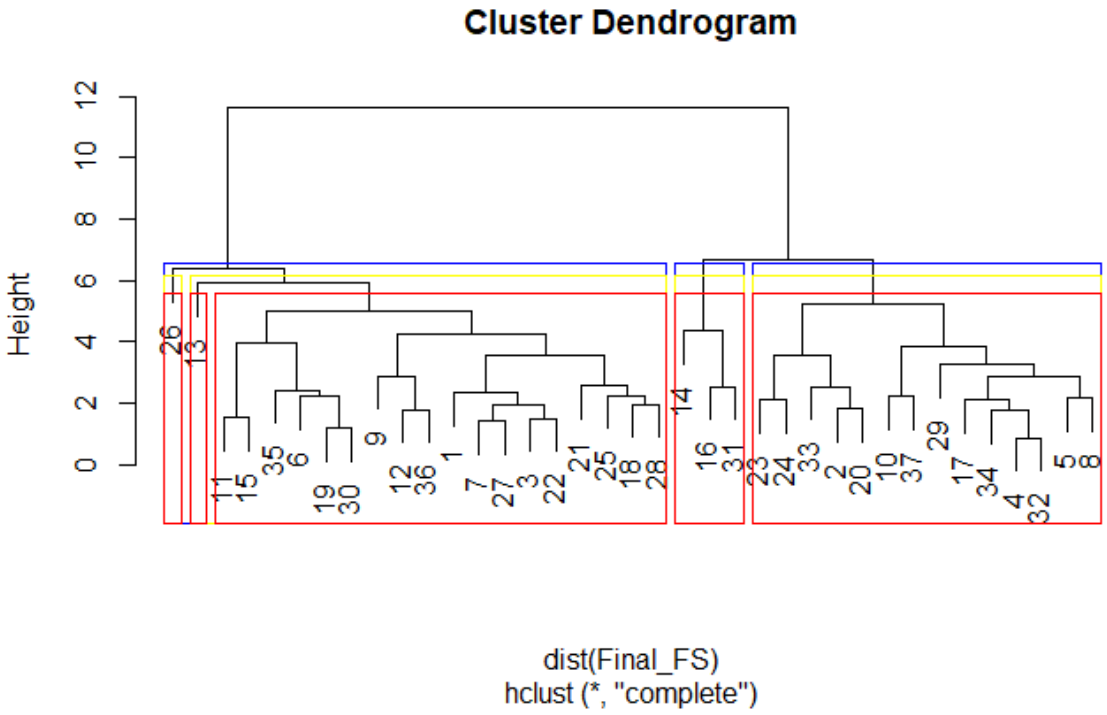
Finally, we performed the clustering method according to the Agglomerative Hierarchical Method using the complete linkage function.

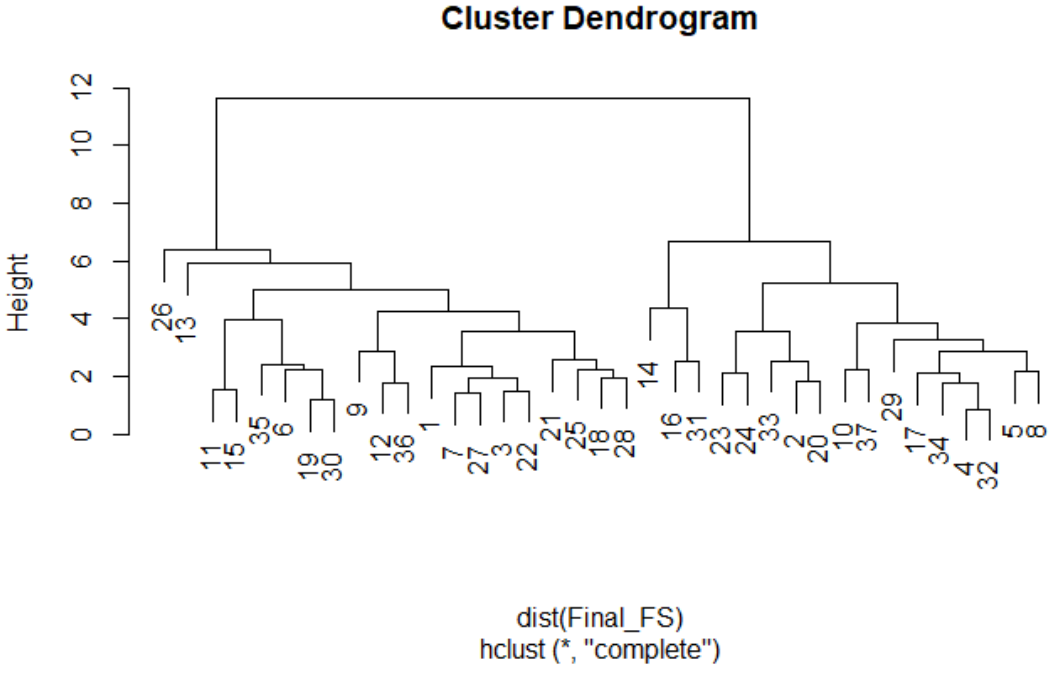
The R code for clustering process-

```
dendogram_fs<-hclust(dist(Final_FS), method="complete")
plot(dendogram_fs, mains="complete")
rect.hclust(dendogram_fs,k=3,border="blue")
rect.hclust(dendogram_fs,k=4,border="yellow")
rect.hclust(dendogram_fs,k=5,border="red")

dendogram_s<- hclust(dist(Final_S), method = "complete")
plot(dendogram_s,mains= "complete")
rect.hclust(dendogram_s,k=3, border= "blue")
rect.hclust(dendogram_s,k=4, border= "yellow")
rect.hclust(dendogram_s,k=5, border= "red")
```

RESULTS-


Cluster Dendrogram
Height
0
2
4
6
8
10
12
26
13
11
15
35
6
19
30
9
12
36
1
7
27
3
22
21
25
18
28
14
16
31
23
24
33
2
20
10
37
29
17
34
4
32
5
8
dist(Final_FS)
hclust (*, "complete")


Cluster Dendrogram
Height
0
2
4
6
8
10
12
26
13
11
15
35
6
19
30
9
12
36
1
7
27
3
22
21
25
18
28
14
16
31
23
24
33
2
20
10
37
29
17
34
4
32
5
8
dist(Final_FS)
hclust (*, "complete")

We decided to group countries into 5 clusters since our dataset was very small we realised using 5 clusters would give us sharper insights and more accurate results thus preventing underutilisation of the data provided.

Interpretation-

We will analyse each cluster one by one and work our way up:

Cluster 1- This cluster includes only **Poland** since it is extremely far apart from the other countries in the dendrogram. Poland has a great geographical location, the cost of living is much lower than most European countries and Polish people have access to universal healthcare and the work weeks are much smaller. Their GDP Per Capita is above 70% of the EU average. Poland was the only country in EU that was not affected by the 2008 recession. Overall they have a pretty stable economy that contributes to favourable working conditions for its citizens. Hence we think these indicators make Poland different from the rest of the countries in the analysis.

Cluster 2- This cluster includes **India** which similar to Poland is far apart from the rest of the countries in the dendrogram. India is a country with about 59% of job dissatisfaction among working citizens. Due to a high working-age population, the job market is extremely competitive making it difficult to find quality jobs. Moreover, unemployment rates are extremely high. India provides a cheaper workforce to international companies that outsource work which adds to the overall unhappiness of the workers.

Cluster 3- We will analyse the closest leaves first:

**Greece & Italy**- Low cost of living, access to good healthcare, and good weather are some factors that are common to both countries which add to employee satisfaction. However, both of these countries report high unemployment rates of 8.2% and 12.45 respectively and low working age population. We think these countries are pretty similar in terms of both business and macroeconomic concepts hence their grouping makes sense.

**North Macedonia & South Africa**- Both of these countries have a high corruption index, almost equal average income of 6640$ and 6780$, cost of living is very close at 38.57% and 42.55% (Compared to the United States which is taken as 100%) and their GDPs are also almost equal at 6.59Mn $ and 6.78 Mn $. We think these factors are important indicators of employee work satisfaction and these countries should be grouped together.

**Egypt is the closest to North Macedonia & South Africa**. The major reason for this grouping would be low costs of living in all three countries and a high score on the corruption index for all three countries. These factors add to the dis/satisfaction of employees. And hence this grouping makes sense.

**UK is the closest to Egypt, South Africa and North Macedonia**. While UK's extremely strong economy differentiates it from the three countries. A high inflation rate in all four countries is what ties them together.

According to the dendogram, **UK, Egypt, South Africa and North Macedonia are close to Greece and Italy**. This grouping is valid since four of these countries are from the European Union, they enjoy political and economic stability to some extent. In addition, South Africa and Egypt are from the African continent. While geographical similarity is one factor. We also think the growing ICT sector and high inflation rates in all countries add to the dis/similarity of the issues faced by employees in the respective countries.

**Hungary & USA** – Both countries show that their ICT sector is growing and is expected to grow at a CAGR of 8.58% and 6.70% respectively and outsourcing is very prevalent in the industry. However , since this sector is fast-growing it might make the job market excessively competitive, these factors imply that the employees face similar levels of job dis/satisfaction in both the countries which influences the decision of grouping these countries together.

**Germany is close to Hungary& USA**. Both Germany and USA are extremely strong economies. Being a part of the European Union Hungary enjoys political and economic stability. In addition, workers from Hungary are allowed to migrate to different European Countries for quality jobs. Booming ICT sector in all three countries, good healthcare facilities, and equally low unemployment rates at 3%, 3.4% and 3.6% respectively produces happy workers. Thus the grouping is valid.

**Finland & Portugal** – Both countries have similar costs of living and access to good healthcare facilities, they have fairly low unemployment rates at 6.8% and 5.8% respectively and high levels of civil rights along with a relatively low corruption index. We think these factors contribute to employees work satisfaction and the grouping of these countries together is valid.

**Brazil & New Zealand**- Both the countries have a growing ICT sector with CAGR of 8.34% and 7.20% respectively which could imply higher levels of job satisfaction among employees in the same industry. This factor could contribute to the grouping of the countries together.

And according to the dendogram **Finland-Portugal are close to New Zealand and Brazil**. Booming ICT sectors, good civil rights and low corruption are some factors that justify this grouping.

Now, **Argentina is close to Finland-Portugal and New Zealand-Brazil**. Argentina and Brazil have geographical proximity. Both countries have high corruption, equally moderate civil rights, similar costs of living. Argentina is currently facing an inflation crisis. We think the growing ICT sector ties these countries together.

**Lithuania & Romania**- These European countries have fairly low unemployment rates at 5.6% and 5.4% respectively, similar inflation rates and cost of living. We feel these indicators might influence work did/satisfaction among employees and the grouping of the 2 countries is valid.

**Peru is close to Lithuania and Romania**. Low unemployment rates (Peru is 3.7%), high inflation are some macro factors that tie these countries together.

**Mexico is close to Peru and Lithuania-Romania**. All four of these countries have a weak economy, high inflation, bad quality of life, high corruption. These factors influence the issues people face at work. It would be safe to say that the grouping is valid and employees from these four countries might face very similar issues at work.

Finally, Cluster 3 is made up of countries with both strong and weak economies. It would not be safe to compare them based on macroeconomic factors. However, the Booming global ICT sector in these countries might influence the working conditions and workers might face similar issues and benefits of working in these respective countries.

Cluster 4-

**Japan and South Korea** are the closest to each other since in both countries majority of the working population reports stressful working conditions and a highly competitive job market which causes difficulty in finding quality job opportunities. Besides this these countries have very low unemployment rates at

2.6% and 2.8% respectively and extremely similar costs of living. These factors contribute to the job dis/satisfaction and make the grouping valid.

**Iran to Japan & South Korea** are closer to the group containing both Japan and South Korea according to the dendrogram. Iran is similar to them in terms of its cost of living.

Thus, similar cost of living is the common factor that affects the lives of workers in the countries of the 4$^{th}$ cluster.

Cluster 5: -

This cluster includes the following countries

**Nigeria and Pakistan**- The grouping of these countries makes a lot of sense. The quality of civil rights is low in the two countries, the cost of living is also almost equal, the unemployment rates in the countries are very close at 5.8% and 6.4%, both countries are facing high inflation at 18.85% and 19.87%, the two countries also face high corruption levels. Since these factors are important to determine employee dis/satisfaction we can say that the employees in the two countries may face similar issues at work.

**Bangladesh and Malaysia**- These 2 Asian countries are pretty similar in terms of their cost of living which is 35.78% and 35.19% respectively, their fairly low unemployment rates at 4.7% and 3.7% respectively and both the countries score badly in the corruption index. We feel that these factors influence the job dis/satisfaction is the country and thus contribute to a grouping of these countries.

**Thailand is close to Bangladesh and Malaysia**. Low costs of living, high corruption, and low unemployment rate are factors common to these three countries that influence worker's lives.

**Ghana & Vietnam**- These countries are similar in terms of their cost of living and both countries score very badly in the corruption index. These are the factors that contributed to the grouping of these countries together.

**Canada & Taiwan**- The ICT industry in both of the countries is constantly growing and expanding. Canada and Taiwan's ICT sectors are expected to grow at a CAGR of 9.91 % and 7.41% respectively which implies the creation of job opportunities in that field for the citizens thus creating job satisfaction.

**Turkey is close to Canada & Taiwan**. Turkey has 288 thousand employees working in its ICT sector which is fast growing. We think the growing ICT sector makes the issues faced by employees similar in the three countries.

**Jordan is close to Turkey, Canada and Taiwan**. Jordan and Turkey both belong to the middle east, apart from geographical proximity. Turkey has a fast growing ICT sector which makes the grouping valid.

**China & France** - The ICT Market is booming in both these countries. China and France's ICT sector is expected to grow at a CAGR of 13.3% and 11.5%. The ICT sector makes up 55% of the Chinese economy. The ICT sector was valued at USD 800.24 billion and USD 114.24 billion in China and France respectively. The industry's growth creates more job opportunities and provides better salaries which translates to a higher cost of living and better job satisfaction. It would be safe to say that employees working in fast-growing ICT industries will face similar issues and benefits thus the grouping of the two countries is valid.

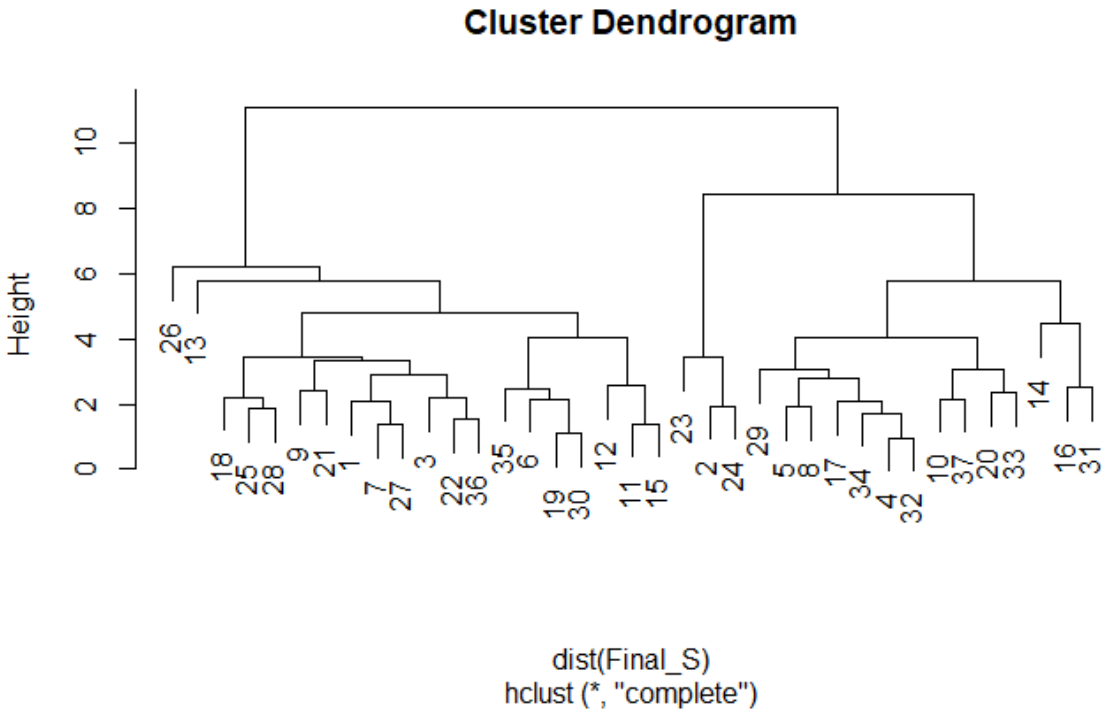
**Jordan, Turkey, Canada and Taiwan are close to China & France**. These countries have very different economies and political situations. They can be grouped on the basis the growing ICT sector which leads to similar issues and benefits faced by employees.

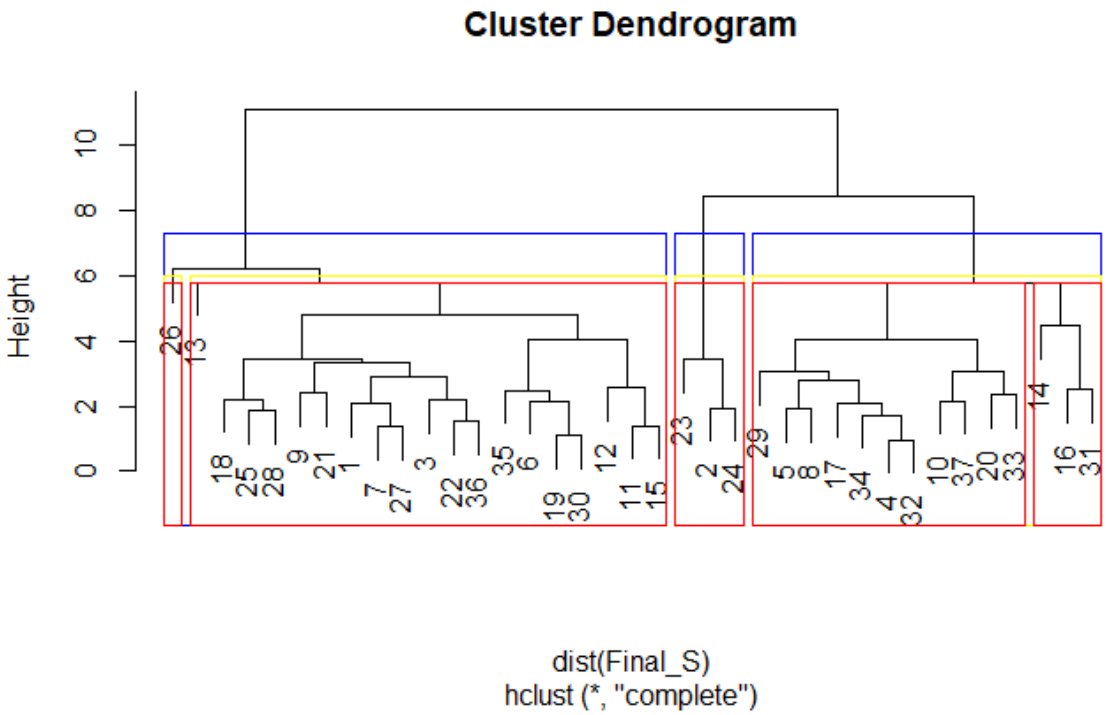
**Russia is similar to Jordan, Turkey, Canada -** Taiwan and China – France. Russia and China have similar political situation and a fast-growing economy. It could be grouped with these countries on the basis of fast-growing ICT sector.

**Ghana- Vietnam is close to Russia, Jordan, Turkey, Canada - Taiwan and China – France.**

Finally, cluster 5 can be justified on the basis of the fast-growing ICT sector.

Now we move on to clustering these countries based on the second score.




Similar to the previous clustering process we chose to create 5 clusters using the given data in order to get sharper and accurate results.

Cluster 1- This cluster includes only **Poland** since it is extremely far apart from the other countries in the dendrogram. Poland has a great geographical location, the cost of living is much lower than most European countries and Polish people have access to universal healthcare and the work weeks are much smaller. Their GDP Per Capita is above 70% of the EU average. Poland was the only country in EU that was not affected by the 2008 recession. Overall they have a pretty stable economy that contributes to favourable working conditions for its citizens. Hence we think these indicators make Poland different from the rest of the countries in the analysis.

Cluster 2- In order to understand the grouping of this cluster we start by looking at countries which are closest to each other-

**Peru & Romania**- These countries have similar costs of living at 49.66% and 43.66% and both of them score badly on the corruption index. Moreover, the ICT industry in both countries is growing at a CAGR of 6.49% and 9.46% respectively. We feel that these factors might play a role in job satisfaction among employees and influence the decision to group these countries together.

**Lithuania is close to Peru-Romania**. Low unemployment rates (Peru is 3.7%), high inflation are some macro factors that tie these countries together.

**India is close to Lithuania, Peru and Romania:** We think the high corruption, high inflation, the nature of developing countries are the common factors in all these countries.

**Germany and Mexico**- The two countries are similar in terms of their low unemployment rates and their ICT industry seems to be growing at a similar pace with their CAGR being 6.80% and 7.53%. These factors can possibly contribute to the grouping of these two countries together.

**Finland & Portugal**- The two countries have similar unemployment rates at 6.8% and 5.8% and similar cost of living. We feel that these similarities have played a role in the grouping of the countries together.

**Argentina is close to Finland-Portugal.** Argentina is currently facing an inflation crisis. We think the growing ICT sector ties these countries together.

**NZ & USA**- These 2 countries have are similar in terms of their cost of living and their low unemployment rates at 3.3% and 3.6%. The ICT industry in both

the countries is expected to grow at a CAGR of 7.20% and 6.80%. We assume that these similarities have influenced the decision of the grouping of countries.

**Brazil is close to NZ&USA**. High Inflation is the macroeconomic indicator that we used to justify this grouping.

**Argentina, Finland-Portugal are close to Brazil, NZ&USA. USA, Finland and New Zealand** are very close in terms of economic growth and stability. Argentina, Portugal and Brazil are relatively weaker economies. We think the developing and developed ICT sectors in all these countries is the common point that would make employees in these countries face similar issues being in the workforce.

**Germany & Mexico are close to Argentina, Finland-Portugal, Brazil, NZ&USA**. While these countries are different in terms of their economies, they all seem to have low unemployment rates. This means that jobs are easier to find and secure. We think this couples with growing ICT sectors makes employee lives similar in these countries.

**India, Lithuania, Peru and Romania are close to Germany & Mexico, Argentina, Finland-Portugal, Brazil, NZ&USA.** Lithuania, Peru, Romania, Mexico, Argentina, Brazil are weaker economies that face political and economic uncertainties. However, the fast-growing ICT business is what makes the countries similar in terms of employee happiness.

**Macedonia & South Africa-**Both of these countries have a high corruption index, almost equal average income of 6640$ and 6780$, cost of living is very close at 38.57% and 42.55% and their GDPs are also almost equal at 6.59Mn $ and 6.78 Mn $. We think these factors are important indicators of employee work satisfaction and these countries should be grouped together.

**Egypt is the closest to North Macedonia & South Africa**. The major reason for this grouping would be low costs of living in all three countries and a high score on the corruption index for all three countries. These factors add to the dis/satisfaction of employees. And hence this grouping makes sense.

**UK is the close to Egypt, South Africa and North Macedonia**. While UK's extremely strong economy differentiates it from the three countries. A high inflation rate in all four countries is what ties them together.

**Greece & Italy**- Low cost of living, access to good healthcare, and good weather are some factors that are common to both countries which add to employee satisfaction. However, both of these countries report high unemployment rates of 8.2% and 12.45 respectively and low working age population. We think these

countries are pretty similar in terms of both business and macroeconomic concepts hence their grouping makes sense.

**Hungary is close to Greece & Italy.** Hungary is similar to Greece and Italy on many macroeconomic parameters. All three countries are a part of the EU and enjoy political and economic stability due to the same. All three countries face heavy corruption, high inflation that makes lives of workers difficult. However, they have access to good healthcare and quality civil rights. These factors heavily influence the issues and benefits workers face.

**Hungary, Greece & Italy are closer to UK, Egypt, South Africa and North Macedonia.** High inflation is one macro indicator and growing ICT sectors is another to justify the groupings.

Finally, cluster 2 can be justified based on the growing needs of the ICT sector.

Cluster 3- This cluster includes **Nigeria, Bangladesh and Pakistan**. We will begin by analysing the countries which appear to be the closest with each first which will be Bangladesh and Pakistan. The 2 countries are very similar in terms of their cost of living and they stand on the same scale in the corruption index and the quality of civil rights offered are also pretty much the same.

Now while comparing this group with Nigeria, we can see that in all three countries, the job satisfaction level among employees isn't that high even though the ICT industry is constantly growing in all the countries. A possible reason for this could be the existence of high competition for the quality jobs available and lack of civil rights that make the lives of employees difficult and more demanding.

Cluster 4- We repeat the same process by focusing on the leaves which are closest

**China & France-** The ICT Market is booming in both these countries. China and France's ICT sector is expected to grow at a CAGR of 13.3% and 11.5%. The ICT sector makes up 55% of the Chinese economy. The ICT sector was valued at USD 800.24 billion and USD 114.24 billion in China and France respectively. The industry's growth creates more job opportunities and provides better salaries which translates to a higher cost of living and better job satisfaction. It would be safe to say that employees working in fast-growing ICT industries will face similar issues and benefits thus the grouping of the two countries is valid.

**Canada & Taiwan-** The ICT industry in both of countries is constantly growing and expanding. Canada and Taiwan's ICT sectors are expected to grow at a CAGR of 9.91 % and 7.41% respectively which implies the creation of job opportunities in that field for the citizens thus creating job satisfaction.

Now since **Turkey is the closest to the Canada & Taiwan,** we will look into it now. Similar to Canada and Taiwan, Turkey is also showing great growth in the ICT sector of about 14.65% CAGR. This is the factor that helps us group these countries together.

After grouping these countries together, we notice that **Jordan is the next country closest to this new group**. The reason for this could be similar growth trends in the ICT sector of 8.63% and the probable satisfaction level of employees here.

**This particular group created is the closest to China & France** and that is so because of the fast-growing ICT sectors which could be the reason for grouping China and France with them.

**Ghana & Vietnam-** These countries are similar in terms of their cost of living and both countries score very badly in the corruption index. These are the factors that contributed to the grouping of these countries together.

**Malaysia & Thailand**- Both the countries have a similar cost of living of 35.19% and 34.67% and they both score badly on the corruption index. Further the ICT sector of the countries is constantly growing at a CAGR of 9.84% and 8% respectively. These factors might contribute to the grouping of the 2 countries together.

Now **Ghana & Vietnam are the closest to Malaysia & Thailand** and the factors that are joining them together besides their location is the high corruption index and the cost of living along with their performance in ICT sector.

The countries in this cluster don't usually lie on the same economic scale and it is difficult to compare them based on a few macroeconomic factors. The only thing that is very common among them is their constant improvement in the ICT sector across the years

Cluster 5- **Japan and South Korea** are the closest to each other since in both countries majority of the working population reports stressful working conditions and difficulty in finding quality job opportunities. Besides this these countries have very low unemployment rates at 2.6% and 2.8% respectively and extremely similar costs of living. Even though Iran is not very close to the individual countries

**Iran to Japan & South Korea** are closer to the group containing both Japan and South Korea according to the dendrogram. Iran is similar to them in terms of its cost of living.

COMPARING CLUSTERING USING BOTH SCORES-

After completing the clustering process using both the scores we can notice the following –

1) Poland is the outlier in both scores.
2) Iran, Japan, and South Korea are always in the same cluster.
3) Ghana and Vietnam, Canada and Taiwan, China and France, Italy and Greece, Macedonia and South Africa, have always been grouped together.
4) India is an outlier in the first score is included in the $2^{nd}$ cluster in the second score.
5) Nigeria, Bangladesh and Pakistan form a separate cluster in the $2^{nd}$ score whereas they were a part of the $5^{th}$ cluster in the $1^{st}$ score.