The University of Texas at Dallas

Naveen Jindal School of Management

Spring 2021

BA WITH R – BUAN 6356.002

Under Guidance of: Prof. Jianqing Chen

Project Final Report

Worldwide Unemployment

GROUP - 10 MEMBERS:

Anvita Kiran Shettennavar (AXS200128)

Hardik Subhash Agarwal (HSA200000)

Nitika Sharma (NXS200013)

Padmanabha Ajay Varma Mudunuri (PXM200018)

Sreedevi Rajitha Malladi (SXM190220)

TABLE OF CONTENTS:

1. Introduction

The World Bank Group is a global partnership that consists of five institutions that work together to find sustainable solutions for poverty and build shared prosperity in developing countries. The World Bank Group works in almost every major area of development. They provide a wide array of financial products and technical assistance and help developing countries to share and apply innovative knowledge and solutions to the challenges they face. Since 1947, the World Bank has funded over 12,000 development projects, via traditional loans, interest-free credits, and grants. The World Development Indicators is a compilation of relevant, high-quality, and internationally comparable statistics about global development and the fight against poverty. The World Bank collection of development indicators is compiled from officially recognized international sources. It consists of the current and accurate global development data available.

The World Bank collects and processes large amounts of data and generates them based on economic models. These data and models have gradually been made available to the public in a way that encourages reuse, whereas the recent publications describing them are available as open access under a Creative Commons Attribution License.

Unemployment is often used as a measure of the health of the economy. It is an economic condition in which individuals seeking jobs remain unhired. The official unemployment rate for the nation is the number of unemployed as a percentage of the labor force (the sum of the employed and unemployed). We are studying the various indicators of development to analyze how unemployment is being affected by these factors.

2. Project Motivation

Nowadays many people, who lose work due to a temporary decrease of a labor force demand, are under the threat of long-term unemployment that has serious consequences on a particular individual and the society in general, therefore the prevention of social exclusion is an important measure to be taken in the crisis. Covid-19 is one such reason for such a drastic increase in the unemployment rate which has led to millions of people losing their jobs. This motivated us to study the unemployment rate of the last 2 decades and how long it takes to get back to normality. This study can be useful for many local NGOs and nonprofits who aim to improve the literacy rate and unemployment conditions in their areas.

3. Project Objective

Our Objective of this project is to analyze the World Development Indicator Dataset using the K-Means Clustering model to observe trends in world unemployment over the years. We are also focusing on some developing regions in the Middle Eastern and Africa in order to observe trends for some other indicators such as Literacy, Urbanization, Industrialization, etc., and see if we can find any similar patterns between the indicators and observe the big picture.

4. Business Value

The information we gather from our observations for various World Development Indicators might help organizations such as the World Bank, UNICEF and their partner organizations, understand which countries might require their help the most. It can help them plan and implement measures to alleviate the situations in certain regions. This project will also help NGOs and CSR (Corporate Social Responsibility) Department of companies to understand the situation in certain countries and work on improving the conditions of such areas.

5. Executive Summary

To research about the unemployment in Africa and find out the different factors which are the cause of this unemployment. We tried to find a dataset that would contain the different indicators of unemployment or factors that are causing unemployment in this region. We took our data from Kaggle and explored all the variables and values in detail for data cleaning. We eliminated all the variables that were of less importance for our project analysis. We then preprocessed the data to filter out only the data that aligns with our analysis/data mining process. Our main variables are the country, indicator, indicator codes that help us in predicting the impact of that indicator in different regions. We then implement the K-means Clustering and Linear Regression to study the impact of the variables.

6. Data description

The World Bank database contains 1,400 time series indicators for 217 economies and more than 40 country groups, with data for many indicators going back more than 50 years. We have chosen a second-hand World Bank dataset from Kaggle that contains information about the

World Development Indicators, which will be used for observing Worldwide Unemployment and other World Development Indicators like Literacy, Food Deficit etc. The original data can be found on the World Bank site.

We are dealing with two different datasets:

- **Country dataset**: It contains information about countries around the world such as their names, codes, currency, region, income, etc. It has a total of 31 columns and 247 rows.

- **Indicators dataset**: It contains information about development indicators names, their codes, year, value, country codes and country names. It has a total of 6 columns and 5,656,458 rows. There are multiple rows for the same country with different indicators such as population total, youth literacy rate, depth of food deficit, etc.

```
> sapply(Country, class)
                         CountryCode                            ShortName                            TableName
                         "character"                          "character"                          "character"
                            LongName                           Alpha2Code                         CurrencyUnit
                         "character"                          "character"                          "character"
                        SpecialNotes                               Region                          IncomeGroup
                         "character"                          "character"                          "character"
                             Wb2Code                  NationalAccountsBaseYear              NationalAccountsReferenceYear
                         "character"                          "character"                          "character"
                    SnaPriceValuation                      LendingCategory                         OtherGroups
                         "character"                          "character"                          "character"
              SystemOfNationalAccounts              AlternativeConversionFactor                       PppSurveyYear
                         "character"                          "character"                          "character"
          BalanceOfPaymentsManualInUse              ExternalDebtReportingStatus                     SystemOfTrade
                         "character"                          "character"                          "character"
           GovernmentAccountingConcept              ImfDataDisseminationStandard                LatestPopulationCensus
                         "character"                          "character"                          "character"
         LatestHouseholdSurvey SourceOfMostRecentIncomeAndExpenditureData          VitalRegistrationComplete
                         "character"                          "character"                          "character"
            LatestAgriculturalCensus                   LatestIndustrialData                       LatestTradeData
                         "character"                             "numeric"                            "numeric"
           LatestWaterWithdrawalData
                           "numeric"
```

```
> sapply(Indicators, class)
  CountryName   CountryCode IndicatorName IndicatorCode          Year         Value
  "character"   "character"   "character"   "character"     "integer"     "numeric"
```

```
> describe(Country$CountryCode)
Country$CountryCode
       n  missing distinct
     247        0      247

lowest : ABW ADO AFG AGO ALB, highest: YEM ZAF ZAR ZMB ZWE
> describe(Country$ShortName)
Country$ShortName
       n  missing distinct
     247        0      247

lowest : Afghanistan        Albania            Algeria            American Samoa     Andorra
highest: West Bank and Gaza World              Yemen              Zambia             Zimbabwe
> describe(Country$Region)
Country$Region
       n  missing distinct
     214       33        7

lowest : East Asia & Pacific        Europe & Central Asia      Latin America & Caribbean  Middle East & North Africa North America
highest: Latin America & Caribbean  Middle East & North Africa North America              South Asia                 Sub-Saharan Africa

Value             East Asia & Pacific      Europe & Central Asia  Latin America & Caribbean Middle East & North Africa             North America
Frequency                          36                         57                         41                         21                         3
Proportion                      0.168                      0.266                      0.192                      0.098                     0.014

Value                      South Asia         Sub-Saharan Africa
Frequency                           8                         48
Proportion                      0.037                      0.224
```

6.1 Properties

- **Country dataset**: It has a total of 31 columns and 247 rows.
- **Indicator dataset**: It has a total of 6 columns and 5,656,458 rows.

6.2 Features of the Country dataset:

- Country code: contains the code of the country
- Short name: contains the short name for countries
- Long name: contains the full names
- Currency unit: mentions currency of each country
- Region: mentions the region of each country
- Income Group: contains income groups such as low, medium, high.

6.3 Features of Indicator dataset:

- Country name: Name of all the countries
- Country code: Code for each country
- Indicator name: World development indicators
- Indicator code: Code for each indicator
- Year: Year it was recorded
- Value: Value for the particular year

7. Data Cleaning

Our data set contains over a thousand annual indicators of economic development from hundreds of countries around the world. The Indicators data set consists of 6 columns and 5,656,458 rows that have complete information about the Indicators that are pointing towards a specific country. The country data set consists of 31 columns and 247 rows where there are some unnecessary columns such as national accounts reference year, Sna price valuation, lending category, ppp survey year, government accounting concept, system of trade, latest household survey, vital registration complete, agricultural data and industrial trade date, other groups. These columns do not provide relevant information for us to use in our project.

While examining the data we found there are many rows that have NA values and missing values where those columns have categorical variables which contain the region of a country and there were 33 rows that did not contain any data. We are omitting these rows as they do not impact our analysis of African regions or of unemployment data. We also found some NA values in columns such as LatestIndustrialData, LatestTradeData, and LatestWaterWithdrawalData. We are omitting these columns as they are not relevant for our objective.

We will also be focusing on the Middle Eastern and African regions in order to observe the different indicators . For this, we merged the Indicators data set and country data set by filtering it by its respective Country codes.

	Country Code	Country
2	ARB	Arab World
3	CEB	Central Europe and the Baltics
4	CSS	Caribbean small states
5	EAP	East Asia & Pacific (developing only)
6	EAS	East Asia & Pacific (all income levels)
7	ECA	Europe & Central Asia (developing only)
8	ECS	Europe & Central Asia (all income levels)
9	EMU	Euro area
10	EUU	European Union
11	FCS	Fragile and conflict affected situations
12	HIC	High income
13	HPC	Heavily indebted poor countries (HIPC)
14	LAC	Latin America & Caribbean (developing only)
15	LCN	Latin America & Caribbean (all income levels)
16	LDC	Least developed countries: UN classification
17	LIC	Low income
18	LMC	Lower middle income
19	LMY	Low & middle income
20	MEA	Middle East & North Africa (all income levels)
21	MIC	Middle income
22	MNA	Middle East & North Africa (developing only)
23	NAC	North America
24	NOC	High income: nonOECD
25	OEC	High income: OECD
26	OED	OECD members
27	OSS	Other small states
28	PSS	Pacific island small states
29	SAS	South Asia
30	SSA	Sub-Saharan Africa (developing only)
31	SSF	Sub-Saharan Africa (all income levels)
32	SST	Small states
33	UMC	Upper middle income

Codes:

```
> Country_Subset <- Country[c(1,8)]
```

Country_Subset

Filter

	CountryCode	Region
1	AFG	South Asia
2	ALB	Europe & Central Asia
3	DZA	Middle East & North Africa
4	ASM	East Asia & Pacific
5	ADO	Europe & Central Asia
6	AGO	Sub-Saharan Africa
7	ATG	Latin America & Caribbean
8	ARB	
9	ARG	Latin America & Caribbean
10	ARM	Europe & Central Asia
11	ABW	Latin America & Caribbean
12	AUS	East Asia & Pacific
13	AUT	Europe & Central Asia
14	AZE	Europe & Central Asia
15	BHR	Middle East & North Africa
16	BGD	South Asia
17	BRB	Latin America & Caribbean
18	BLR	Europe & Central Asia
19	BEL	Europe & Central Asia
20	BLZ	Latin America & Caribbean

```
Country_noEmpty <- Country_Subset[!(Country_Subset$Region == "" | is.na(Country_Subset$Region)), ]
```

Country_noEmpty

Filter

	CountryCode	Region
1	AFG	South Asia
2	ALB	Europe & Central Asia
3	DZA	Middle East & North Africa
4	ASM	East Asia & Pacific
5	ADO	Europe & Central Asia
6	AGO	Sub-Saharan Africa
7	ATG	Latin America & Caribbean
9	ARG	Latin America & Caribbean
10	ARM	Europe & Central Asia
11	ABW	Latin America & Caribbean
12	AUS	East Asia & Pacific
13	AUT	Europe & Central Asia
14	AZE	Europe & Central Asia
15	BHR	Middle East & North Africa
16	BGD	South Asia
17	BRB	Latin America & Caribbean
18	BLR	Europe & Central Asia
19	BEL	Europe & Central Asia
20	BLZ	Latin America & Caribbean
21	BEN	Sub-Saharan Africa

Showing 1 to 20 of 214 entries, 2 total columns

```
ue_merge <- merge(Indicators, Country_noEmpty, by="CountryCode")
```

ue_merge

Filter

	CountryCode	CountryName	IndicatorName	IndicatorCode	Year	Value	Region
1	ABW	Aruba	Consumer price index (2010 = 100)	FP.CPI.TOTL	2012	1.049714e+02	Latin America & Caribbean
2	ABW	Aruba	Manufacturing, value added (% of GDP)	NV.IND.MANF.ZS	2011	4.674974e+00	Latin America & Caribbean
3	ABW	Aruba	ICT goods exports (% of total goods exports)	TX.VAL.ICTG.ZS.UN	2011	4.195835e-01	Latin America & Caribbean
4	ABW	Aruba	Interest rate spread (lending rate minus deposit rate, %)	FR.INR.LNDP	2011	7.983333e+00	Latin America & Caribbean
5	ABW	Aruba	Manufacturing, value added (current LCU)	NV.IND.MANF.CN	2011	1.937800e+08	Latin America & Caribbean
6	ABW	Aruba	Employment in industry, male (% of male employment)	SL.IND.EMPL.MA.ZS	2011	2.440000e+01	Latin America & Caribbean
7	ABW	Aruba	International tourism, receipts (% of total exports)	ST.INT.RCPT.XP.ZS	2011	1.978686e+01	Latin America & Caribbean
8	ABW	Aruba	Employment in industry, female (% of female employ…	SL.IND.EMPL.FE.ZS	2011	3.300000e+00	Latin America & Caribbean
9	ABW	Aruba	Theoretical duration of primary education (years)	SE.PRM.DURS	1985	6.000000e+00	Latin America & Caribbean
10	ABW	Aruba	Household final consumption expenditure, PPP (curre…	NE.CON.PRVT.PP.CD	2011	1.736946e+09	Latin America & Caribbean
11	ABW	Aruba	Insurance and financial services (% of service imports,…	BM.GSR.INSF.ZS	2011	1.477597e+00	Latin America & Caribbean
12	ABW	Aruba	Rural population growth (annual %)	SP.RUR.TOTL.ZG	1967	6.129938e-01	Latin America & Caribbean
13	ABW	Aruba	Computer, communications and other services (% of c…	TX.VAL.OTHR.ZS.WT	2012	1.501365e+01	Latin America & Caribbean
14	ABW	Aruba	Employment in agriculture, male (% of male employm…	SL.AGR.EMPL.MA.ZS	2011	8.000000e-01	Latin America & Caribbean
15	ABW	Aruba	ICT service exports (% of service exports, BoP)	BX.GSR.CCIS.ZS	2012	1.395578e+01	Latin America & Caribbean
16	ABW	Aruba	Import value index (2000 = 100)	TM.VAL.MRCH.XD.WD	2011	2.291820e+02	Latin America & Caribbean
17	ABW	Aruba	Liner shipping connectivity index (maximum value in …	IS.SHP.GCNW.XQ	2011	6.210000e+00	Latin America & Caribbean
18	ABW	Aruba	Survival to age 65, female (% of cohort)	SP.DYN.TO65.FE.ZS	1985	8.395874e+01	Latin America & Caribbean
19	ABW	Aruba	Death rate, crude (per 1,000 people)	SP.DYN.CDRT.IN	2012	8.207000e+00	Latin America & Caribbean
20	ABW	Aruba	ICT service exports (BoP, current US$)	BX.GSR.CCIS.CD	2012	2.457542e+08	Latin America & Caribbean

8. BI model

8.1 Clustering:

Clustering is a process of grouping together data into smaller groups based on their similarity from a larger dataset. These smaller groups that are formed from the bigger data groups are called clusters. Here, we implement k-means clustering for analyzing our data. It is a form of unsupervised data mining.

8.2 Linear Regression:

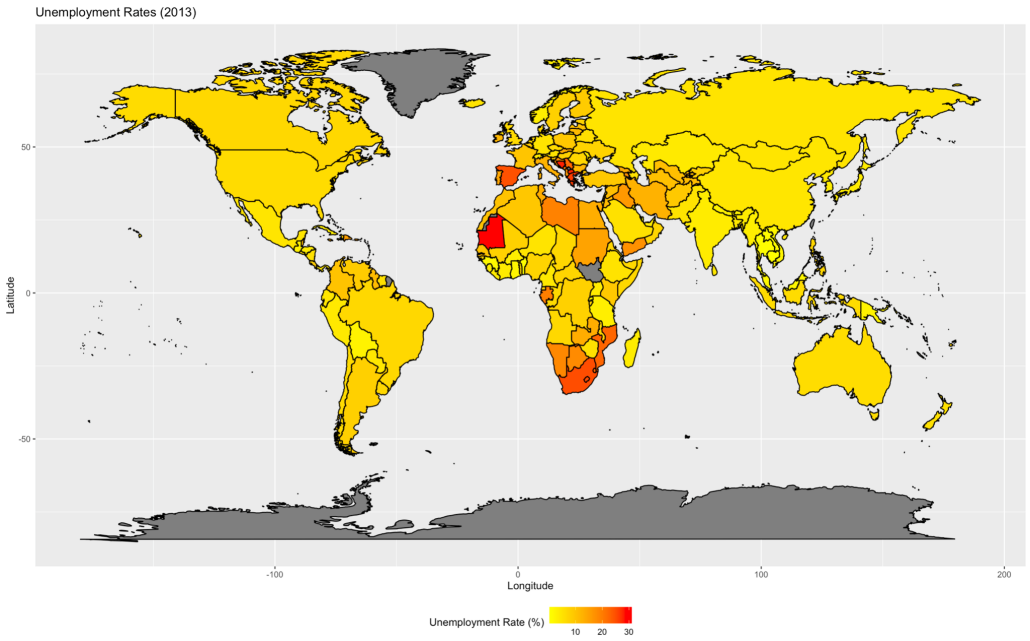
Linear Regression models are very popular tools used for prediction. It is a method used to summarize and analyze relationships between two continuous quantitative variables. We predict the outcome variable Y which is the dependent variable based on one or more input variables X which are the independent variables. They are also called as response (Y) and predictors (X). With the help of a straight line, we study the linear relationship to see how the scattered data is represented with respect to the original representation.

9. Findings

We observed the unemployment rate of the World for the year 2013. In order to observe the trends across the years 1991 to 2013, we ran a K- means clustering on the data to group the countries with similar trends. We then observed that from 1991 to 2013, there were almost 181 countries for which the unemployment rate didn't decrease. As we all know that the continent Africa is one of the least developed regions compared to the rest of the world, we decided to inspect the countries from Africa. We zoomed in to African continent in order to observe the unemployment rate and the other factors responsible for the underdevelopment of the countries of that region. We ran a K-means clustering to cluster countries with similar trends and analyzed how different indicators impact these countries of the African region.

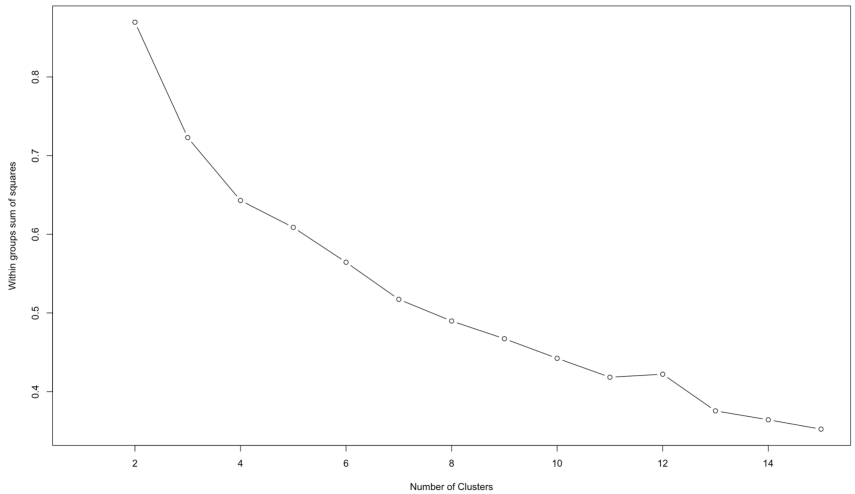
9.1 Unemployment Rate Of The World:

As we can see from the heat map, Yellow represents countries with low unemployment rate for Year 2013 and red shows countries with high unemployment rate for the Year 2013. Many countries from the African region showed a high unemployment rate for the Year 2013, the highest percentage being around 30%.



9.2 Sum of Square Plot for the World:

The below graph is the Sum of Square plot which helps in determining the number of clusters for K means clustering. The optimal number of clusters can be determined from the Elbow curve method. According to the Elbow curve method, the optimal point is the point after which the curve starts flattening. From the below curve, the optimal number can be 4. We took 2 clusters only because our data was huge and we were getting similar patterns in 3rd and 4th clusters. There were very few countries. Therefore, for our analysis 4 clusters was not relevant. So, we used 2 clusters.

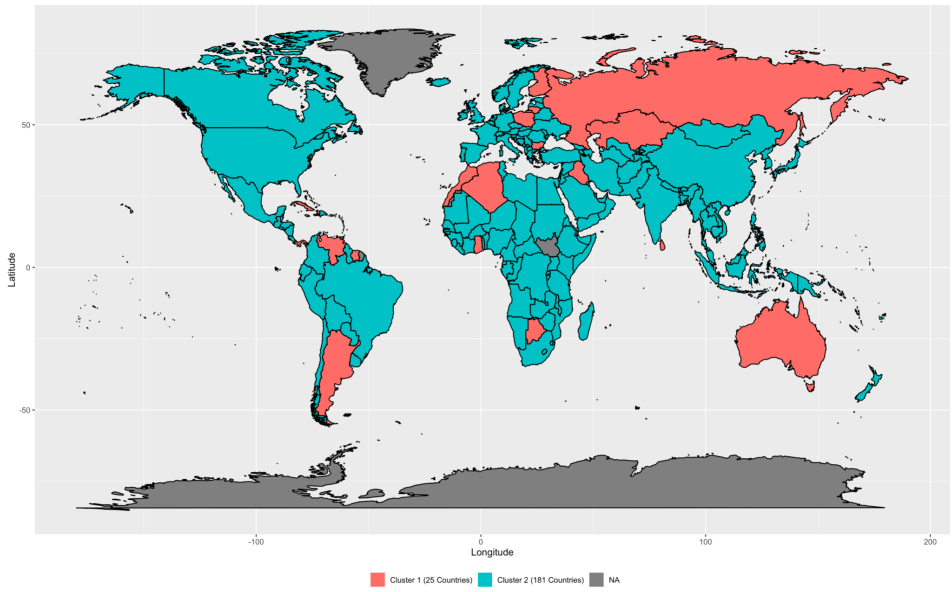


9.3 K means clustering for the World:

In the below K means clustering line graph across the year from 1991 till 2013, we can see that Cluster on the left has 25 countries with decreasing unemployment rates over a period of 20+ years. Whereas Cluster on the right have 181 countries with relatively stable unemployment rates over a period of 20+ years. We assume that the unemployment rate for developing countries might have remained stagnant while the unemployment rate for the developed countries might have decreased.

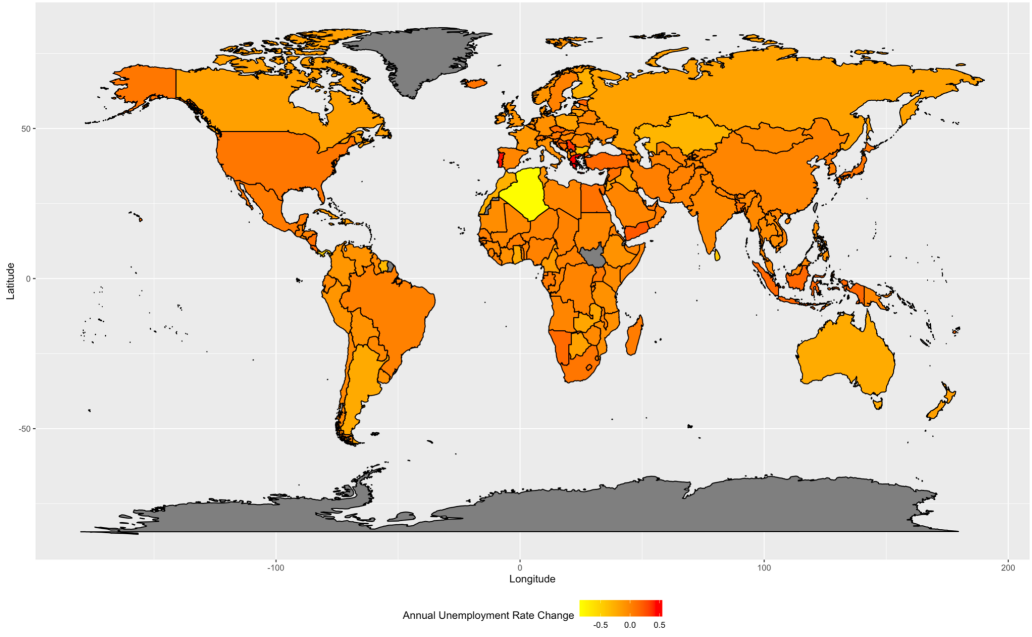
9.4 Location of the clusters:

From the below map, we can see two types of clusters. Cluster 2 which is blue color indicates countries with stagnant unemployment rate whereas Cluster 1 which is red color indicates countries with decreasing unemployment rate. Most of the countries from Africa have stagnant unemployment rates. Even in the developing countries of Asia, Latin America we see a stagnant unemployment rate. Some of the developed countries such as the USA, Norway, Germany, Ireland, Poland have stagnant unemployment rates but countries such as Russia have a decreasing unemployment rate.



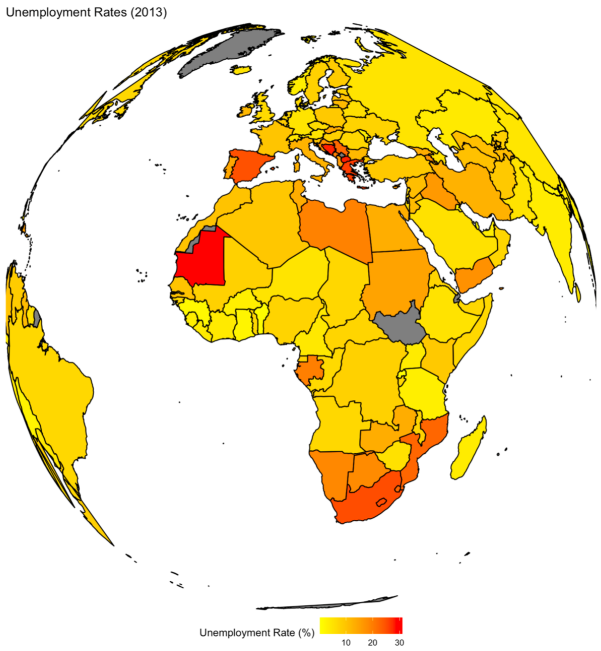
9.5 Changes in the unemployment rate from 1961 to 2013:

The below chart is the representation of the changes of unemployment in different countries from 1961 to 2013. We used linear regression for each country's employment against the time and then visualized it in the map. As we can see the unemployment rate changes -0.5 to 0.5. Most of the colour of the map is orange which indicates that the unemployment rate has increased approximately 0.2-0.3 in many developing countries of Africa, Latin America and Asia.



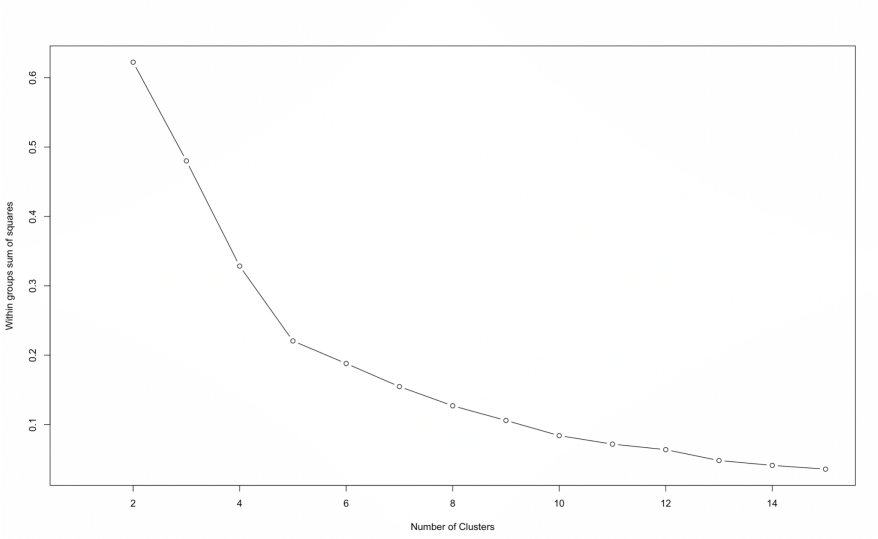
9.6 Unemployment rate for African Region:

Now, let's focus on the continent, Africa. The below heat globe represents the unemployment rate of the African region for the year 2013. Countries such as Mauritania, Senegal, South Africa have high unemployment rates almost ranging from 25% to 30%. Countries such as Libya, Namibia, Sudan, Zambia also have high unemployment rates ranging from 15% to 25%. Countries such as Kenya, Niger, Chad have low unemployment rates in comparison to the other countries.



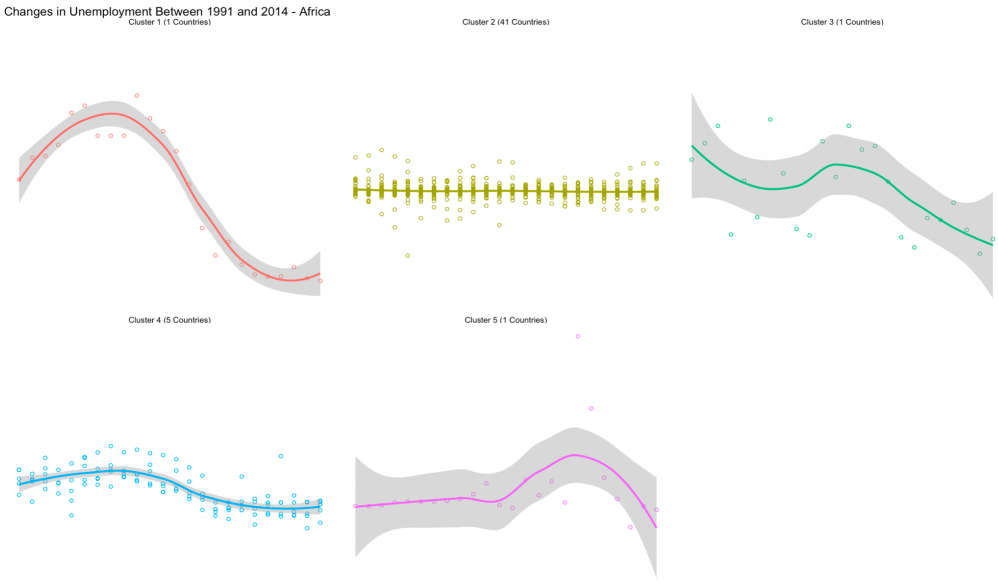
9.7 Sum of Square Plot for the African Region (Unemployment rate):

From the below graph, it can be observed that the optimal number of clusters for the African region for Unemployment rate can be 5 clusters as after 5 clusters the Sum of Square plot starts flattening. Therefore, the optimal number of clusters are 5.



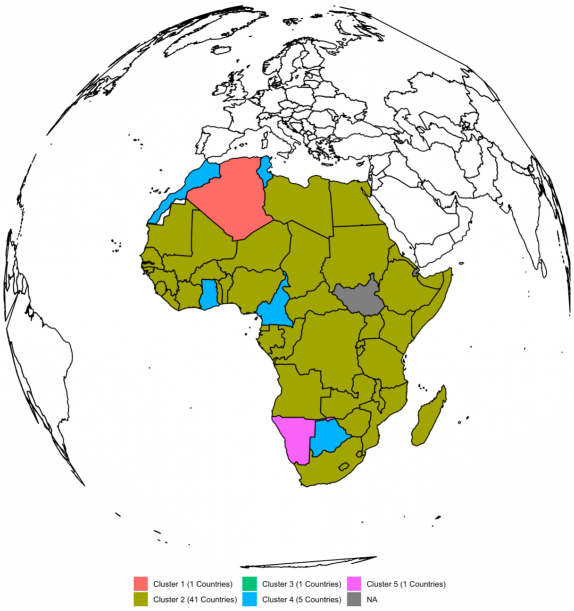
9.8 K means clustering for African Region for Unemployment Rate:

The below K means clustering line graph across the year from 1991 till 2013, we can see that in 5 clusters that one graph has a stagnant unemployment rate. Whereas in the rest of the 4 clusters there is a decreasing unemployment rate. So, almost 8 countries have had a decreasing unemployment rate for 20+ years. Whereas in the cluster 2 where there is a stagnant unemployment rate almost 41 countries have a stagnant unemployment rate of 20+ years.



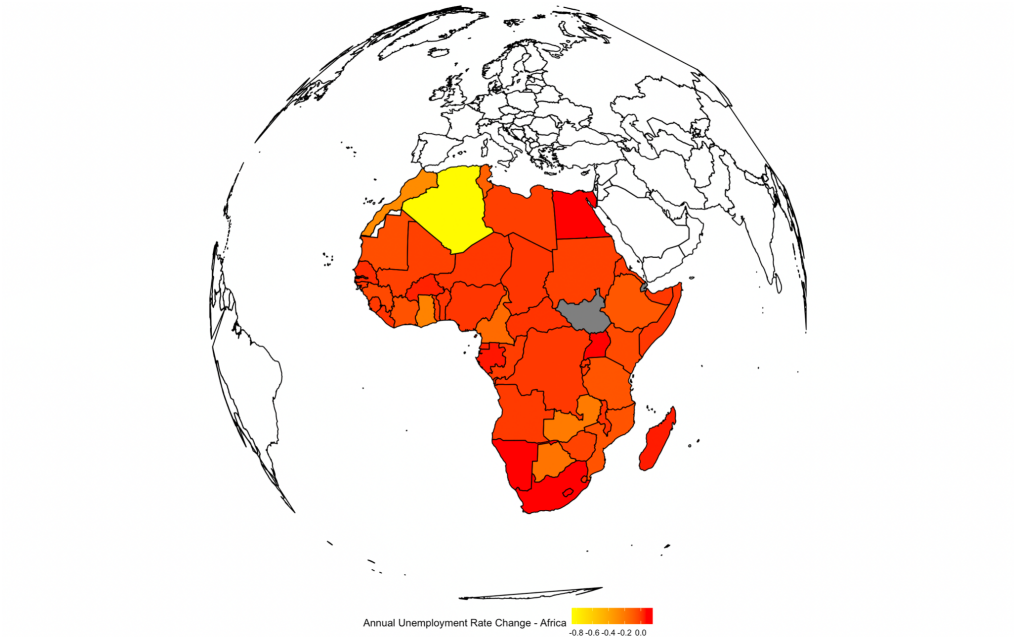
9.9 Location of the clusters in african region:

From the below map, we can see 5 types of clusters. The major color that can be seen is greenish yellow color, which represents stagnant unemployment rate. The rest of the color represents clusters which have decreasing unemployment rates. Countries such as Algeria, Morocco, Botswana have decreasing unemployment rates because they have good external financing. Economic stability, good supply of food etc. Whereas countries such as Nigeria, Libya, Egypt, Mali etc, come in third world countries, where even basic amenities are not available. The unemployment rate is still there, even after some interventions by UN agencies.



9.10 Changes in the unemployment rate for african region :

The below chart is the representation of the changes of unemployment in African region from 1961 to 2013. We used linear regression for each African countries' employment against the time and then visualized it in the map. As we can see the unemployment rate changes -0.8 to 0.3. Most of the colour of the map is red which indicates that the unemployment rate has increased approximately by 0.2 in many developing countries of Africa. As discussed above countries such as Algeria, Morocco, Botswana decreased their unemployment rate (-0.8 to -0.4). Whereas the rest of the countries faced an increase in the unemployment rate across the years from 1961 to 2013.



9.11 Other Indicators that might affect Unemployment directly or indirectly:

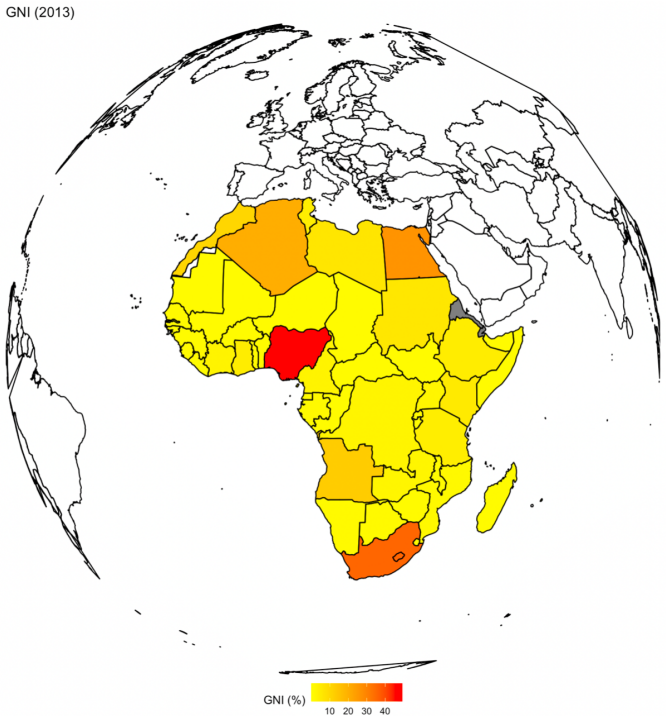
9.11.1 GNI(Gross National Income) for Africa region:

Gross national income (GNI), the sum of a country's gross domestic product (GDP) plus net income (positive or negative) from abroad. It represents the value produced by a country's economy in a given year, regardless of whether the source of the value created is domestic production or receipts from overseas.

The GNI is largely considered a better indicator to account for the income available to the dwellers of a country because it captures the incomes related to the mobility of factors of production (wages earned by cross-border workers, repatriated profits and dividends, etc.), the so-called Net Primary Incomes (NPI), in the Systems of National Accounts.

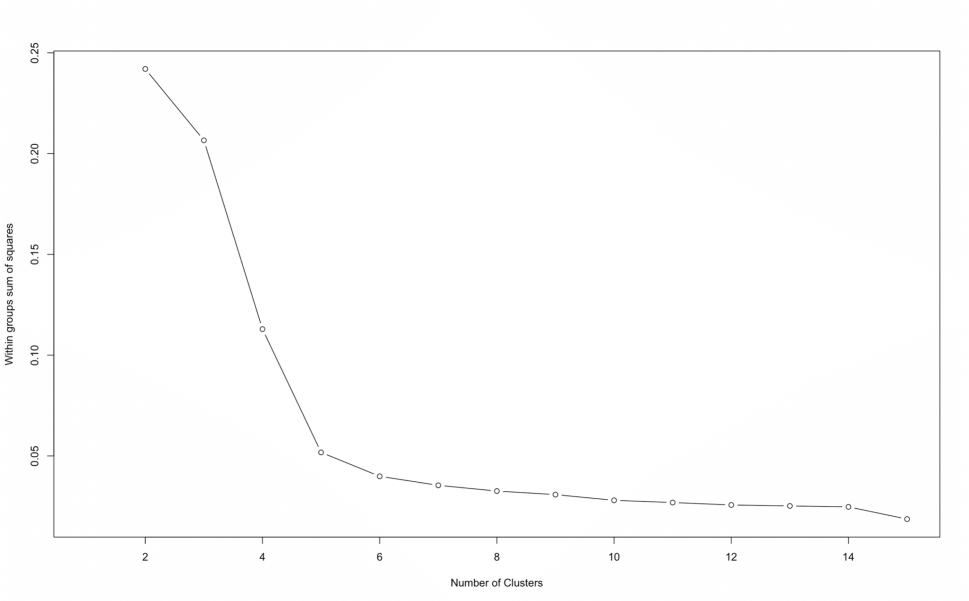
The income groupings use GNI per capita (in U.S. dollars, converted from local currency using the *Atlas* method) since they follow the same methodology used by the World Bank when determining it's operational lending policy. While it is understood that GNI per capita does not completely summarize a country's level of development or measure welfare, it has proved to be a useful and easily available indicator that is closely correlated with other, non monetary measures of the quality of life, such as life expectancy at birth, mortality rates of children, and enrollment rates in school.

The below heat globe represents the Gross National Income of the African region for the year 2013. Countries such as Algeria, Egypt, Morocco have high GNI ranging from 20% to 30%. Countries such as Niger, Malia, Chad, Congo have low GNI ranging from 5% to 10%. We can observe from the previous maps that most of the countries which have low GNI also have high unemployment rates. Therefore, where the unemployment rate is high people are earning less money.



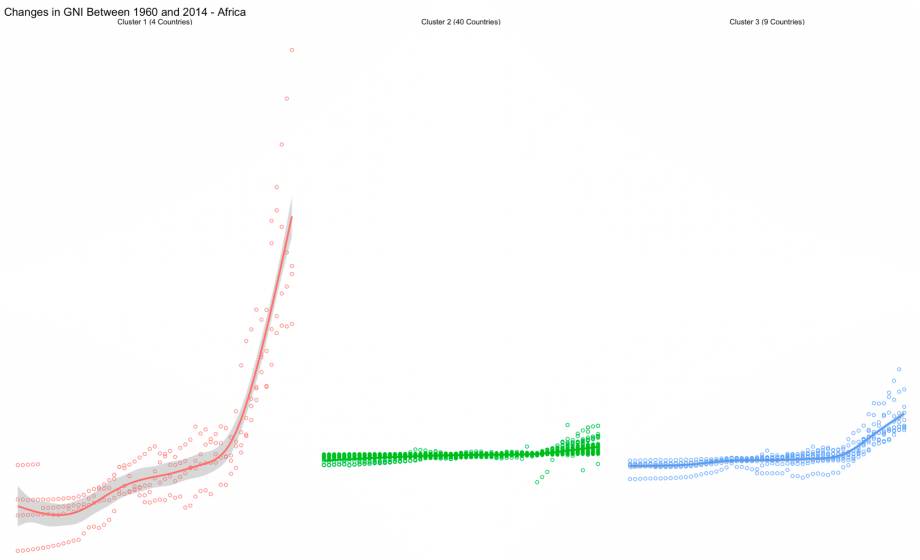
9.11.2 Sum of Square Plot for the African Region (GNI rate):

From the below graph, it can be observed that the optimal number of clusters for the African region for GNI can be 3. The elbow curve comes at 5th point but we were not getting the relevant clusters because the pattern was repetitive. Therefore, we are taking only 3 clusters.



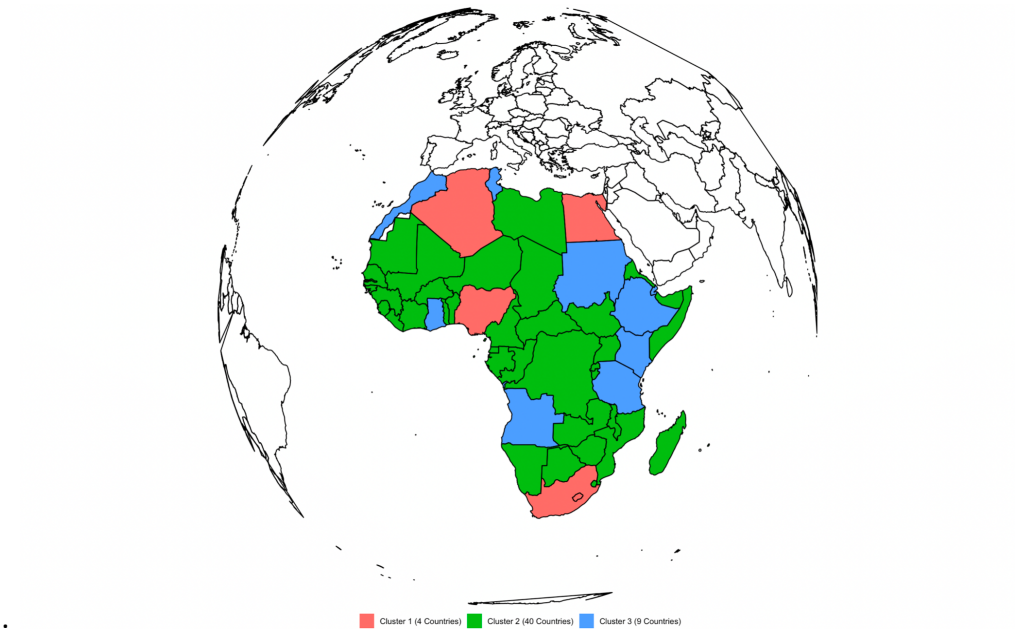
9.11.3 K- means clustering for GNI of African region:

The below K means clustering line graph across the year from 1991 till 2013, we can see that in 3 clusters that one graph has a stagnant GNI. Whereas in the rest of the 2 clusters GNI has increased. So, almost 13 countries have had an increasing GNI for 20+ years. Whereas in the cluster almost 40 countries have a stagnant GNI of 20+ years.



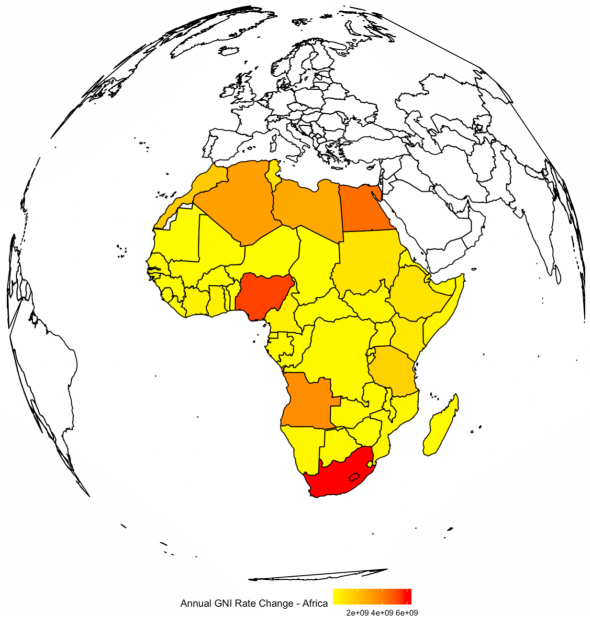
9.11.4 Location of the Clusters in african region (for GNI):

From the below map, we can see 3 types of clusters. The major colors that can be seen are green and blue. The color green represents cluster 2 where the GNI is stagnant. The color Blue represents cluster 3 where GNI has increased. Countries such as South Africa, Nigeria, Algeria and Egypt have increasing GNI because they have decreasing unemployment rate, people are earning more and their economy is stable. Whereas countries such as Namibia, Libya, Zambia, Mali etc, come in third world countries, where unemployment rate is high, people are not earning more, goods are not produced in abundance.



9.11.5 Change in the GNI rate for african region:

The below chart is the representation of the changes of GNI in African region from 1961 to 2013. We used linear regression for each African countries' GNI against the time and then visualized it in the map. As we can see the GNI changes from high to low across 50+ years. Most of the colour of the map is yellow which indicates that the GNI has not increased that much in many developing countries of Africa. As discussed above, countries such as South Africa, Nigeria, Egypt have a positive change in their GNI, which means that their GNI has increased in the past 50+ years. Whereas many countries such as Zambia, Tanzania, Burkina, Nigeria, Chad, etc did not see a significant increase in their GNI in the past 50+ years.

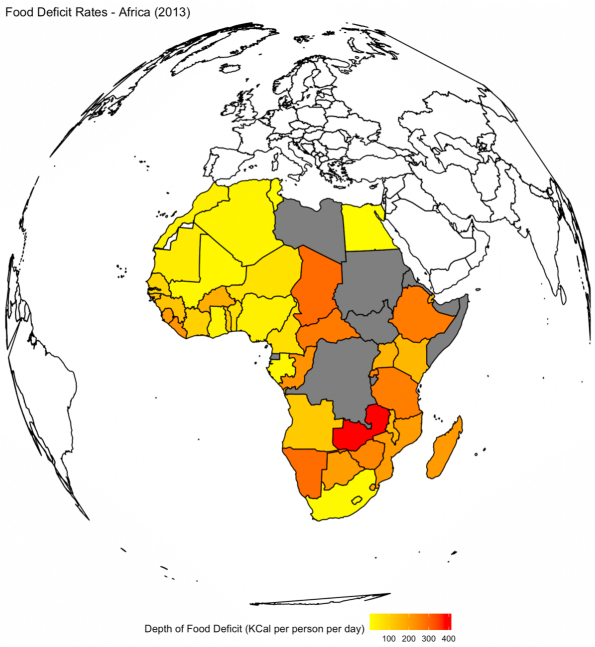


9.11.6 Food Deficit for African region:

Food deficit indicates how much food-deprived people fall short of minimum food needs in terms of dietary energy. Food deficiency is a major problem in African countries. There is a shortage of calories, protein, essential items in African countries. Due to this people do not get proper nutrition and this impairs their cognitive level.

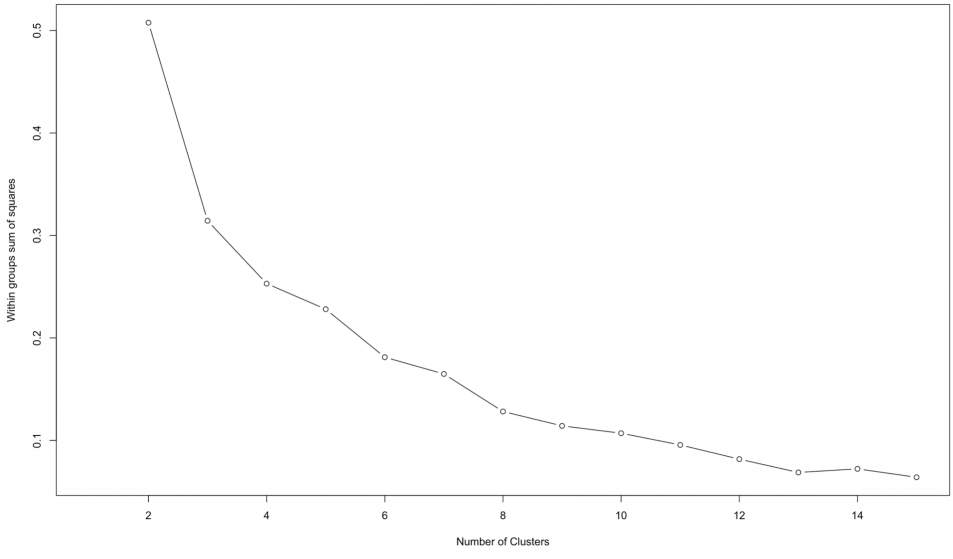
Food deficiency is not a direct factor of unemployment but it is an outcome of unemployment. Countries where unemployment is low, people also have less income. They are unable to afford basic facilities such as food, education, health services etc. Therefore, Gross National Income and Food deficiency are indirectly related to unemployment.

The below heat globe represents the Food deficiency rate of the African region for the year 2013. Countries such as Namibia, Zambia, Chad, Botswana have high Food deficiency rates ranging from 300 to 400 Kcal. Countries such as Algeria, Morocco, Mali, Niger have low food deficiency rates ranging from 50 to 150 Kcal. Kcal indicates Kilocalories which is a metric for how much amount of calories is needed to lift the undernourished people. In developing countries more Kilocalories are needed.



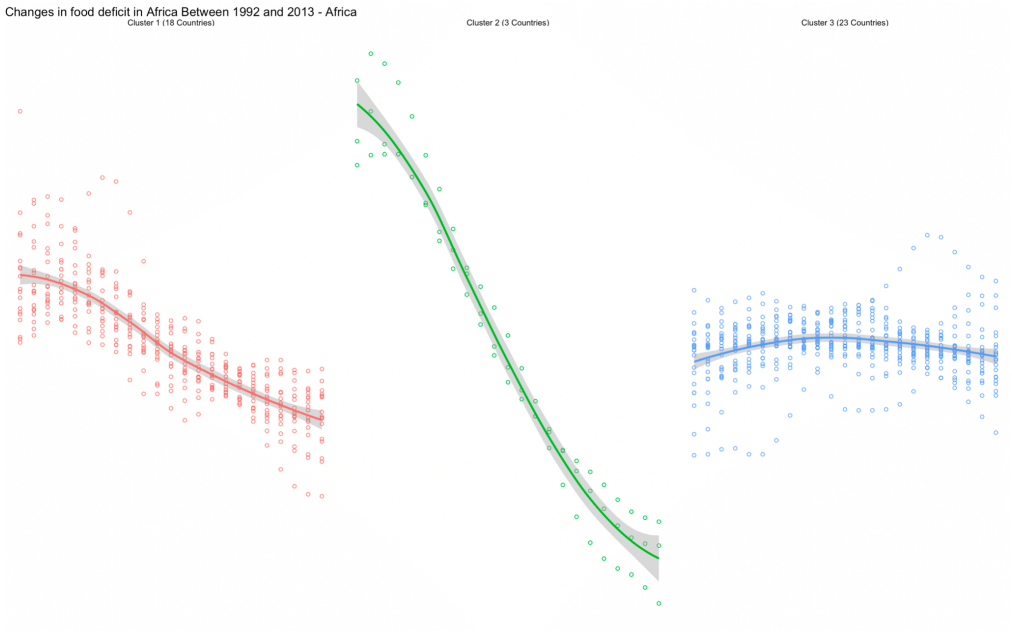
9.11.7 Sum of Square Plot for the African Region (Food Deficiency Rate):

From the below graph, it can be observed that the optimal number of clusters for the African region for Food Deficiency rate can be 3 clusters as after 3 clusters the Sum of Square plot starts flattening. Therefore, the optimal number of clusters are 3.



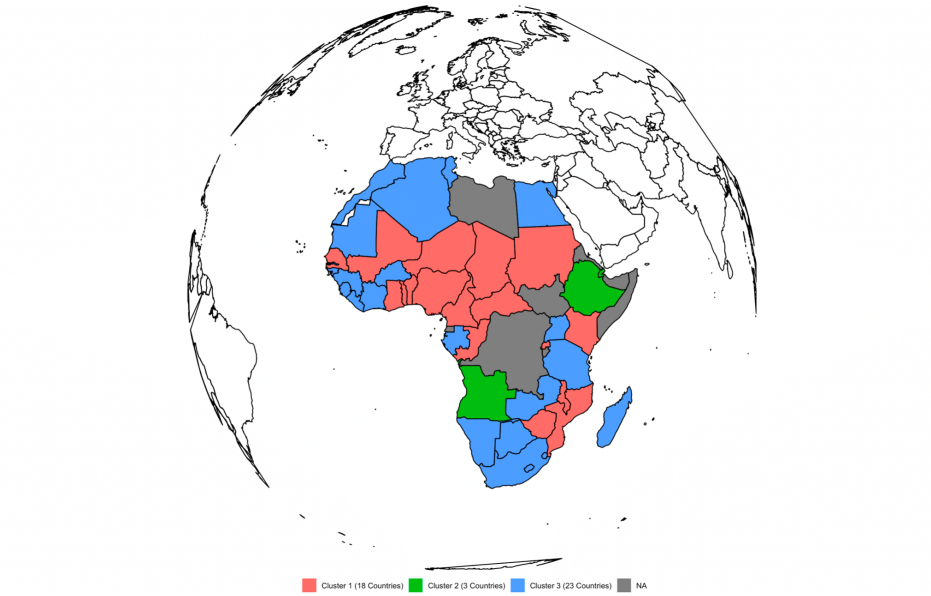
9.11.8 K means clustering for African Region for Food Deficit Rate:

In the below K means clustering line graph across the year from 1991 till 2013, we can see that in 2 clusters which have a total of 21 countries the Food deficiency has decreased tremendously during the period of 20+ years. Whereas in the third cluster, where there are 23 clusters , Food Deficit has not decreased that much for 20+ years.



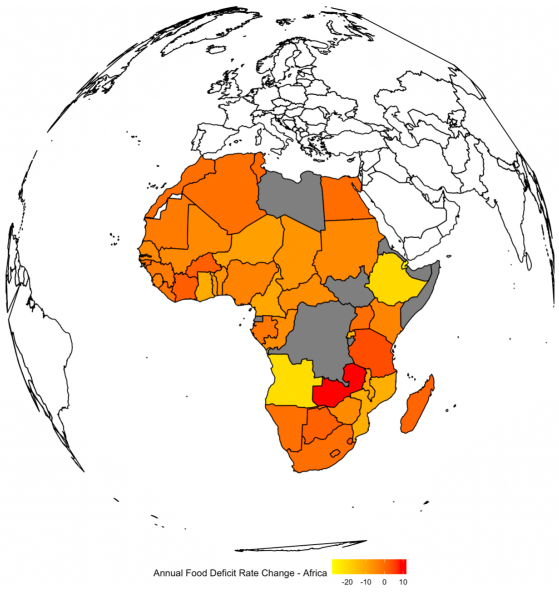
9.11.9 Location of the clusters in african region for food deficit:

From the below map, we can see 3 types of clusters. The major colors that can be seen are pink and blue. The color Pink represents clusters where the Food Deficit rate has decreased. The color Blue represents clusters where Food Deficit rate is somewhat stagnant. Countries such as Algeria, Morocco, Botswana have decreasing Food Deficit rates because they are able to afford basic amenities and their economy is stable whereas countries such as Nigeria, Libya, Egypt, Mali etc, come in third world countries, where even basic amenities are not available due to which their Food Deficit rate has also not decreased.



9.11.10 Changes in the Food Deficit Rate :

The below chart is the representation of the changes of Food Deficiency in African region from 1961 to 2013. We used linear regression for each African countries' Food deficiency against the time and then visualized it in the map. As we can see the Food Deficiency rate changes from -30 till 10 across 50+ years. Most of the colour of the map is orange which indicates that the Food deficiency rate has decreased approximately -0.5 in many countries of Africa. As discussed above countries such as Algeria, Morocco, Botswana decreased their unemployment rate (-0.8 to -0.4). Whereas some of the countries such as Zambia, Tanzania, Burkina faced an increase in the Food deficit rate across the years from 1961 to 2013 approximately by 5.



10. Next Steps

While working on this project, our team felt that we need to utilize these BI models and work further into gaining more insights about the unemployment rate across the world.

- Our next steps include focusing on different continents across the world for unemployment rates and using more indicators that affect the unemployment rate of a country directly or indirectly.
- We aim to gather dataset for the more recent years for all countries and use them to analyze the change in unemployment rate with respect to different times and obtain the reasons for the change in rates.
- We also aim to find the correlation between different countries and continents and how they are affecting each other's unemployment rate.

11. Challenges

Data cleaning is one of the challenges we faced as we have large volumes of data and preprocessing the data was difficult as the dataset contains rows and columns that do not impact our analysis.

We did not have much information on clustering, K-means, and linear regression at the beginning of the semester, we learned about them eventually as the classes progressed and it was challenging for us to implement them in the last phase of the project but with the help of class resources we were able to tackle it.

Due to the covid situation and lack of physical appearance, group work was challenging as we faced challenges in coordinating and collaborating. We were able to overcome this by working on teams and by planning weekly meetings in order to work on the project effectively.

12. Conclusion

From our insights and findings, we can observe that unemployment is a major issue in the world and especially in the African countries. In order to study the pattern of unemployment we used K means clustering and linear regression on the world and then zoomed in for African countries. In the heat maps, we could see that in most of the countries unemployment has increased across many years.

We were astounded to see that even after several innovations and interventions by UN agencies, the unemployment rate has not alleviated. Therefore, we observed other indicators such as GNI (Gross National Income) and Food Deficit Rate. We could see that for most of the African countries, Food Deficiency has increased and GNI has not increased that much across 50+ years. We observed that in most of the countries where unemployment rate was high, GNI was low and Food Deficit rate was high and vice versa.

From our observation, we can say that if steps are taken to reduce Food Deficiency and to improve Income, then indirectly unemployment can also be alleviated. Therefore, it shows that there is a need for proper intervention in certain regions of Africa in order to alleviate unemployment levels.

13. Rcodes

```
# title: "Worldwide Unemployment"

# indicator code: SL.UEM.TOTL.ZS


library(sf)

library(spData)

library(tmap)

library(leaflet)

library(cartogram)

library(data.table, warn.conflicts = FALSE, quietly = TRUE)

library(dplyr, warn.conflicts = FALSE, quietly = TRUE)

library(dtplyr, warn.conflicts = FALSE, quietly = TRUE)

library(ggplot2, warn.conflicts = FALSE, quietly = TRUE)

library(tidyr, warn.conflicts = FALSE, quietly = TRUE)

library(maps, warn.conflicts = FALSE, quietly = TRUE)

library(tidyverse)

library(tibble)

library(maps)

library(dplyr)  # %>% select() filter() bind_rows()

library(rgdal)  # readOGR() spTransform()

library(raster)  # intersect()

library(ggsn)  # north2() scalebar()

library(rworldmap)  # getMap()
```

```
library(viridis)

library(viridisLite)

require(maps)

require(viridis)

theme_set(

  theme_void()

)

indicators <- read.csv("Indicators.csv")

# What can the World Development Indicator dataset tell us about
unemployment?

# Where in the world are unemployment rates the highest?

# Correcting Country Names

correction <- c("Antigua and Barbuda"="Antigua", "Bahamas,
The"="Bahamas", "Brunei Darussalam"="Brunei", "Cabo Verde"="Cape
Verde", "Congo, Dem. Rep."="Democratic Republic of the Congo",
"Congo, Rep."="Republic of Congo", "Cote d'Ivoire"="Ivory
Coast", "Egypt, Arab Rep."="Egypt", "Faeroe Islands"="Faroe
Islands", "Gambia, The"="Gambia", "Iran, Islamic Rep."="Iran",
"Korea, Dem. Rep."="North Korea", "Korea, Rep."="South Korea",
"Kyrgyz Republic"="Kyrgyzstan", "Lao PDR"="Laos", "Macedonia,
FYR"="Macedonia", "Micronesia, Fed. Sts."="Micronesia", "Russian
Federation"="Russia", "Slovak Republic"="Slovakia", "St.
Lucia"="Saint Lucia", "St. Martin (French part)"="Saint Martin",
"St. Vincent and the Grenadines"="Saint Vincent", "Syrian Arab
Republic"="Syria", "Trinidad and Tobago"="Trinidad", "United
Kingdom"="UK", "United States"="USA", "Venezuela,
RB"="Venezuela", "Virgin Islands (U.S.)"="Virgin Islands",
"Yemen, Rep."="Yemen")



for (c in names((correction))) {

  indicators[indicators$CountryName==c,"CountryName"] =
correction[c]

}
```

```
ue2013<-indicators %>%
  filter(Year==2013
         & IndicatorCode =="SL.UEM.TOTL.ZS")


map.world <- merge(map_data(map="world"),
                   select(ue2013,CountryName,Value),
                   by.x='region',
                   by.y='CountryName',
                   all.x=TRUE,
                   fill=0)


map.world <- map.world[order(map.world$order),]


ggplot(map.world) +
  geom_map(map=map.world, aes(map_id=region, x=long, y=lat,
fill=Value)) +
  borders("world",colour="black")+
  scale_fill_gradient(low
="yellow",high="red",guide="colourbar",name="Unemployment Rate
(%)")+ labs(title="Unemployment Rates
(2013)",x="Longitude",y="Latitude")+theme(legend.position="botto
m")

# What clusters can a k-means analysis point out?

# * On the left, cluster 1 is 167 countries with relatively
stable unemployment rates over 20+ year period.

# * On the right, cluster 2 is 39 countries with decreasing
unemployment rates over 20+ year period.
```

```
ue <- indicators %>%
  filter(IndicatorCode == "SL.UEM.TOTL.ZS") %>%
  select(CountryName,Year,Value)  %>%
  spread(Year,Value,sep="")

country_name <- ue$CountryName
ue <- ue %>%
  select(-CountryName) %>%
  t() %>%
  scale(center=TRUE,scale=FALSE) %>%
  t()
ue <- data.frame(country_name,ue)

# Determine number of clusters
wss<-numeric()
for (i in 2:15) {
  km <-kmeans(select(ue,-country_name),centers=i)
  wss[i]<-sum(km$withinss)/sum(km$totss)
}
plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within
groups sum of squares")
set.seed(1001)
fit <- kmeans(select(ue,-country_name),
centers=2,iter.max=100,nstart=1)
ue <- data.frame(ue, fit$cluster)
```

```
ue <- ue %>%
  gather(Year,Value,-c(country_name,fit.cluster)) %>%
  group_by(fit.cluster) %>%
  mutate(cluster.count = n_distinct(country_name)) %>%
  ungroup() %>%
  mutate(Year = as.numeric(gsub("Year","",Year)),
         fit.cluster = paste("Cluster ",fit.cluster," 
(",cluster.count," Countries)",sep=""))
ggplot(ue,aes(Year,Value,colour=as.factor(fit.cluster)))+
  geom_point(shape=1)+geom_smooth()+
  scale_colour_discrete(guide=FALSE)+
  labs(title="Changes in World Unemployment Between 1991 and 
2013",x="Year",y="Unemployment Rate (scaled)")+
  facet_wrap(~fit.cluster)


# Where are these clusters located?
map.world2 <- merge(map_data(map="world"),
                    select(ue,country_name,fit.cluster),
                    by.x='region',
                    by.y='country_name',
                    all.x=TRUE,
                    fill=0)
map.world2 <- map.world2[order(map.world2$order),]
ggplot(map.world2) +
  geom_map(map=map.world2, aes(map_id=region, x=long, y=lat, 
fill=fit.cluster)) +
  borders("world",colour="black")+
```

```
  scale_fill_discrete(name="")+

  labs(x="Longitude",y="Latitude")+
theme(legend.position="bottom")

# Where are unemployment rates increasing and decreasing?

ue <- indicators %>%

  filter(IndicatorCode == "SL.UEM.TOTL.ZS") %>%

  select(CountryName,Year,Value) %>%

  group_by(CountryName) %>%

  do(ue.rate=coef(lm(Value~Year,data=.))[[2]])


# Convert from list to numeric.

ue$ue.rate<-as.numeric(unlist(ue$ue.rate))

map.world3 <- merge(map_data(map="world"),

                    ue, by.x='region',  by.y='CountryName',
all.x=TRUE,fill=0)

map.world3 <- map.world3[order(map.world3$order),]

ggplot(map.world3) +

  geom_map(map=map.world3, aes(map_id=region, x=long, y=lat,
fill=ue.rate)) +

  borders("world",colour="black")+

  scale_fill_continuous(low="yellow",high="red",name="Annual
World Unemployment Rate
Change")+labs(x="Longitude",y="Latitude")
+theme(legend.position="bottom")


# title: "Africa Unemployment"

# indicator code: SL.UEM.TOTL.ZS

indicators <- read.csv("Indicators.csv")
```

```
# Where in Africa are the unemployment rates highest?

correction <- c("Antigua and Barbuda"="Antigua", "Bahamas,
The"="Bahamas", "Brunei Darussalam"="Brunei", "Cabo Verde"="Cape
Verde", "Congo, Dem. Rep."="Democratic Republic of the Congo",
"Congo, Rep."="Republic of Congo", "Cote d'Ivoire"="Ivory
Coast", "Egypt, Arab Rep."="Egypt", "Faeroe Islands"="Faroe
Islands", "Gambia, The"="Gambia", "Iran, Islamic Rep."="Iran",
"Korea, Dem. Rep."="North Korea", "Korea, Rep."="South Korea",
"Kyrgyz Republic"="Kyrgyzstan", "Lao PDR"="Laos", "Macedonia,
FYR"="Macedonia", "Micronesia, Fed. Sts."="Micronesia", "Russian
Federation"="Russia", "Slovak Republic"="Slovakia", "St.
Lucia"="Saint Lucia", "St. Martin (French part)"="Saint Martin",
"St. Vincent and the Grenadines"="Saint Vincent", "Syrian Arab
Republic"="Syria", "Trinidad and Tobago"="Trinidad", "United
Kingdom"="UK", "United States"="USA", "Venezuela,
RB"="Venezuela", "Virgin Islands (U.S.)"="Virgin Islands",
"Yemen, Rep."="Yemen")



for (c in names((correction))) {

  indicators[indicators$CountryName==c,"CountryName"] =
correction[c]

}

africa <- c("Algeria","Angola","Benin",

            "Botswana","Burkina Faso","Burundi","Ivory
Coast","Cabo Verde","Cameroon",

            "Central African Republic","Chad","Comoros",

            "Democratic Republic of the Congo","Republic of
Congo","Djibouti","Egypt",

            "Equatorial
Guinea","Eritrea","Ethiopia","Gabon","Gambia","Ghana","Guinea",


"Guinea-Bissau","Kenya","Lesotho","Liberia","Libya","Madagascar"
,"Malawi",


"Mali","Mauritania","Mauritius","Morocco","Mozambique","Namibia"
,
```

```
            "Niger","Nigeria","Rwanda","Sao Tome and
Principe","Senegal","Seychelles",

            "Sierra Leone","Somalia","South Africa","South
Sudan","Sudan","Swaziland",


"Tanzania","Togo","Tunisia","Uganda","Zambia","Zimbabwe")



ue2013<-indicators %>%

  filter(Year==2013

         & IndicatorCode =="SL.UEM.TOTL.ZS"

         & CountryName %in% africa)



map.africa <- merge(map_data(map="world", region = africa),
select(ue2013,CountryName,Value),by.x='region',
by.y='CountryName', all.x=TRUE, fill=0)

map.africa <- map.africa[order(map.africa$order),]

ggplot(map.africa) +

  geom_map(map=map.africa, aes(map_id=region, x=long, y=lat,
fill=Value)) + borders("world",colour="black")+
scale_fill_gradient(low
="yellow",high="red",guide="colourbar",name="Unemployment Rate
(%)")+ labs(title="Unemployment Rates - Africa
(2013)",x="Longitude",y="Latitude")+
theme(legend.position="bottom") + coord_map("ortho", orientation
= c(10, 15, 0))



# What clusters can a k-means analysis point out?

ue <- indicators %>%

  filter(IndicatorCode == "SL.UEM.TOTL.ZS"

         & CountryName %in% africa) %>%

  select(CountryName,Year,Value)  %>%
```

```
  spread(Year,Value,sep="")

country_name <- ue$CountryName

ue <- ue %>%

  select(-CountryName) %>%

  t() %>%

  scale(center=TRUE,scale=FALSE) %>%

  t()

ue <- data.frame(country_name,ue)


# Determine number of clusters

wss<-numeric()

for (i in 2:15) {

  km <-kmeans(select(ue,-country_name),centers=i)

  wss[i]<-sum(km$withinss)/sum(km$totss)

}

plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within
groups sum of squares")


set.seed(1001)

fit <- kmeans(select(ue,-country_name),
centers=5,iter.max=100,nstart=1)

ue <- data.frame(ue, fit$cluster)


ue <- ue %>%

  gather(Year,Value,-c(country_name,fit.cluster)) %>%

  group_by(fit.cluster) %>%
```

```
  mutate(cluster.count = n_distinct(country_name)) %>%

  ungroup() %>%

  mutate(Year = as.numeric(gsub("Year","",Year)),

  fit.cluster = paste("Cluster ",fit.cluster," 
(",cluster.count," Countries)",sep=""))



ggplot(ue,aes(Year,Value,colour=as.factor(fit.cluster)))+

  geom_point(shape=1)+geom_smooth()+

  scale_colour_discrete(guide=FALSE)+

  labs(title="Changes in World Unemployment Between 1991 and 
2013 - Africa",x="Year",y="Unemployment Rate (scaled)")+

  facet_wrap(~fit.cluster)



# Where are these clusters located?

map.africa2 <- merge(map_data(map="world", region = africa),

                     select(ue,country_name,fit.cluster),

                     by.x='region',

                     by.y='country_name',

                     all.x=TRUE,

                     fill=0)

map.africa2 <- map.africa2[order(map.africa2$order),]

ggplot(map.africa2) +

  geom_map(map=map.africa2, aes(map_id=region, x=long, y=lat, 
fill=fit.cluster)) +

  borders("world",colour="black")+

  scale_fill_discrete(name="")+

  labs(x="Longitude",y="Latitude")+
```

```
  theme(legend.position="bottom") +   coord_map("ortho",
orientation = c(10, 15, 0))

# Where in africa are unemployment rates increasing and
decreasing?

ue <- indicators %>%

  filter(IndicatorCode == "SL.UEM.TOTL.ZS"

         & CountryName %in% africa) %>%

  select(CountryName,Year,Value) %>%

  group_by(CountryName) %>%

  do(ue.rate=coef(lm(Value~Year,data=.))[[2]])


# Convert from list to numeric.

ue$ue.rate<-as.numeric(unlist(ue$ue.rate))

map.africa3 <- merge(map_data(map="world", region = africa),

                     ue,

                     by.x='region',

                     by.y='CountryName',

                     all.x=TRUE,

                     fill=0)

map.africa3 <- map.africa3[order(map.africa3$order),]


ggplot(map.africa3) +geom_map(map=map.africa3,
aes(map_id=region, x=long, y=lat,
fill=ue.rate))+borders("world",colour="black")+scale_fill_contin
uous(low="yellow",high="red",name="Annual Unemployment Rate
Change - Africa")+  labs(x="Longitude",y="Latitude")
+theme(legend.position="bottom")  +  coord_map("ortho",
orientation = c(10, 15, 0))
```

```
# title: "Africa food deficit"

# indicator code: SN.ITK.DFCT

indicators <- read.csv("Indicators.csv")



# What can the World Development Indicator dataset tell us about
food deficit?

# Where in the world are food deficit rates the highest?

# Correcting Country Names

correction <- c("Antigua and Barbuda"="Antigua", "Bahamas,
The"="Bahamas", "Brunei Darussalam"="Brunei", "Cabo Verde"="Cape
Verde", "Congo, Dem. Rep."="Democratic Republic of the Congo",
"Congo, Rep."="Republic of Congo", "Cote d'Ivoire"="Ivory
Coast", "Egypt, Arab Rep."="Egypt", "Faeroe Islands"="Faroe
Islands", "Gambia, The"="Gambia", "Iran, Islamic Rep."="Iran",
"Korea, Dem. Rep."="North Korea", "Korea, Rep."="South Korea",
"Kyrgyz Republic"="Kyrgyzstan", "Lao PDR"="Laos", "Macedonia,
FYR"="Macedonia", "Micronesia, Fed. Sts."="Micronesia", "Russian
Federation"="Russia", "Slovak Republic"="Slovakia", "St.
Lucia"="Saint Lucia", "St. Martin (French part)"="Saint Martin",
"St. Vincent and the Grenadines"="Saint Vincent", "Syrian Arab
Republic"="Syria", "Trinidad and Tobago"="Trinidad", "United
Kingdom"="UK", "United States"="USA", "Venezuela,
RB"="Venezuela", "Virgin Islands (U.S.)"="Virgin Islands",
"Yemen, Rep."="Yemen")



for (c in names((correction))) {

  indicators[indicators$CountryName==c,"CountryName"] =
correction[c]

}

africa <- c("Algeria","Angola","Benin",

            "Botswana","Burkina Faso","Burundi","Ivory
Coast","Cabo Verde","Cameroon",

            "Central African Republic","Chad","Comoros",
```

```
            "Democratic Republic of the Congo","Republic of
Congo","Djibouti","Egypt",

            "Equatorial
Guinea","Eritrea","Ethiopia","Gabon","Gambia","Ghana","Guinea",


"Guinea-Bissau","Kenya","Lesotho","Liberia","Libya","Madagascar"
,"Malawi",


"Mali","Mauritania","Mauritius","Morocco","Mozambique","Namibia"
,

            "Niger","Nigeria","Rwanda","Sao Tome and
Principe","Senegal","Seychelles",

            "Sierra Leone","Somalia","South Africa","South
Sudan","Sudan","Swaziland",


"Tanzania","Togo","Tunisia","Uganda","Zambia","Zimbabwe")



fd2013<-indicators %>%

  filter(Year==2013

         & IndicatorCode =="SN.ITK.DFCT"

         & CountryName %in% africa)

map.africa <- merge(map_data(map="world", region = africa),

                    select(fd2013,CountryName,Value),

                    by.x='region',

                    by.y='CountryName',

                    all.x=TRUE,

                    fill=0)


map.africa <- map.africa[order(map.africa$order),]
```

```
ggplot(map.africa) +geom_map(map=map.africa, aes(map_id=region,
x=long, y=lat, fill=Value)) +
borders("world",colour="black")+scale_fill_gradient(low
="yellow",high="red",guide="colourbar",name="Food Deficit Rate
(%)")+labs(title="Food Deficit Rates - Africa
(2013)",x="Longitude",y="Latitude")+theme(legend.position="botto
m") + coord_map("ortho", orientation = c(10, 15, 0))


# What clusters can a k-means analysis point out?


fd <- indicators %>%

  filter(IndicatorCode == "SN.ITK.DFCT"

         & CountryName %in% africa) %>%

  select(CountryName,Year,Value)  %>%

  spread(Year,Value,sep="")


country_name <- fd$CountryName


fd <- fd %>%

  select(-CountryName) %>%

  t() %>%

  scale(center=TRUE,scale=FALSE) %>%

  t()


fd <- data.frame(country_name,fd)


# imputation

for(i in 1:ncol(fd)) {
```

```
  fd[ , i][is.na(fd[ , i])] <- mean(fd[ , i], na.rm = TRUE)
}

# Determine number of clusters
wss<-numeric()
for (i in 2:15) {
  km <-kmeans(select(fd,-country_name),centers=i)
  wss[i]<-sum(km$withinss)/sum(km$totss)
}
plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within
groups sum of squares")

set.seed(1001)
fit <- kmeans(select(fd,-country_name),
centers=3,iter.max=100,nstart=1)
fd <- data.frame(fd, fit$cluster)

fd <- fd %>%
  gather(Year,Value,-c(country_name,fit.cluster)) %>%
  group_by(fit.cluster) %>%
  mutate(cluster.count = n_distinct(country_name)) %>%
  ungroup() %>%
  mutate(Year = as.numeric(gsub("Year","",Year)),
         fit.cluster = paste("Cluster ",fit.cluster,"
(",cluster.count," Countries)",sep=""))

ggplot(fd,aes(Year,Value,colour=as.factor(fit.cluster)))+
```

```
  geom_point(shape=1)+geom_smooth()+

  scale_colour_discrete(guide=FALSE)+

  labs(title="Changes in food deficit in Africa Between 1992 and
2013 - Africa",x="Year",y="food deficit Rate (scaled)")+

  facet_wrap(~fit.cluster)



# Where are these clusters located?

map.africa2 <- merge(map_data(map="world", region =
africa),select(fd,country_name,fit.cluster), by.x='region',
by.y='country_name',  all.x=TRUE,fill=0)

map.africa2 <- map.africa2[order(map.africa2$order),]

ggplot(map.africa2) +

  geom_map(map=map.africa2, aes(map_id=region, x=long, y=lat,
fill=fit.cluster)) +

  borders("world",colour="black")+

  scale_fill_discrete(name="")+

  labs(x="Longitude",y="Latitude")+

  theme(legend.position="bottom") +

  coord_map("ortho", orientation = c(10, 15, 0))



# Where are food deficit rates increasing and decreasing?

# Below is a regression of each country's employment against
time (year).

# This is made simple with the dplyr library.



fd <- indicators %>%

  filter(IndicatorCode == "SN.ITK.DFCT"

         & CountryName %in% africa) %>%
```

```
  select(CountryName,Year,Value) %>%

  group_by(CountryName) %>%

  do(ue.rate=coef(lm(Value~Year,data=.))[[2]])


# Convert from list to numeric.

fd$ue.rate<-as.numeric(unlist(fd$ue.rate))


map.africa3 <- merge(map_data(map="world", region = africa),
fd,by.x='region',by.y='CountryName',all.x=TRUE, fill=0)

map.africa3 <- map.africa3[order(map.africa3$order),]

ggplot(map.africa3) +geom_map(map=map.africa3,
aes(map_id=region, x=long, y=lat, fill=ue.rate))
+borders("world",colour="black")+scale_fill_continuous(low="yell
ow",high="red",name="Annual Food Deficit Rate Change - Africa")+
labs(x="Longitude",y="Latitude") +
theme(legend.position="bottom")  + coord_map("ortho",
orientation = c(10, 15, 0))




# title: " GNI Current US Dollar - Africa"

# indicator code: NY.GNP.MKTP.CD

indicators <- read.csv("Indicators.csv")




# Correction

correction <- c("Antigua and Barbuda"="Antigua", "Bahamas,
The"="Bahamas", "Brunei Darussalam"="Brunei", "Cabo Verde"="Cape
Verde", "Congo, Dem. Rep."="Democratic Republic of the Congo",
"Congo, Rep."="Republic of Congo", "Cote d'Ivoire"="Ivory
Coast", "Egypt, Arab Rep."="Egypt", "Faeroe Islands"="Faroe
Islands", "Gambia, The"="Gambia", "Iran, Islamic Rep."="Iran",
```

```
"Korea, Dem. Rep."="North Korea", "Korea, Rep."="South Korea",
"Kyrgyz Republic"="Kyrgyzstan", "Lao PDR"="Laos", "Macedonia,
FYR"="Macedonia", "Micronesia, Fed. Sts."="Micronesia", "Russian
Federation"="Russia", "Slovak Republic"="Slovakia", "St.
Lucia"="Saint Lucia", "St. Martin (French part)"="Saint Martin",
"St. Vincent and the Grenadines"="Saint Vincent", "Syrian Arab
Republic"="Syria", "Trinidad and Tobago"="Trinidad", "United
Kingdom"="UK", "United States"="USA", "Venezuela,
RB"="Venezuela", "Virgin Islands (U.S.)"="Virgin Islands",
"Yemen, Rep."="Yemen")



for (c in names((correction))) {

  indicators[indicators$CountryName==c,"CountryName"] =
correction[c]

}

africa <- c("Algeria","Angola","Benin",

            "Botswana","Burkina Faso","Burundi","Ivory
Coast","Cabo Verde","Cameroon",

            "Central African Republic","Chad","Comoros",

            "Democratic Republic of the Congo","Republic of
Congo","Djibouti","Egypt",

            "Equatorial
Guinea","Eritrea","Ethiopia","Gabon","Gambia","Ghana","Guinea",


"Guinea-Bissau","Kenya","Lesotho","Liberia","Libya","Madagascar"
,"Malawi",


"Mali","Mauritania","Mauritius","Morocco","Mozambique","Namibia"
,

            "Niger","Nigeria","Rwanda","Sao Tome and
Principe","Senegal","Seychelles",

            "Sierra Leone","Somalia","South Africa","South
Sudan","Sudan","Swaziland",


"Tanzania","Togo","Tunisia","Uganda","Zambia","Zimbabwe")
```

```
gni2013<- indicators %>%

  filter(Year==2013

         & IndicatorCode =="NY.GNP.MKTP.CD"

         & CountryName %in% africa)



map.africa <- merge(map_data(map="world", region = africa),
select(gni2013,CountryName,Value),  by.x='region',
by.y='CountryName', all.x=TRUE, fill=0)

map.africa <- map.africa[order(map.africa$order),]

map.africa$Value <- map.africa$Value/10000000000

ggplot(map.africa) + geom_map(map=map.africa, aes(map_id=region,
x=long, y=lat, fill=Value)) +borders("world",colour="black") +
scale_fill_gradient(low
="yellow",high="red",guide="colourbar",name="GNI (%)") +
labs(title="GNI
(2013)",x="Longitude",y="Latitude")+theme(legend.position="botto
m") + coord_map("ortho", orientation = c(10, 15, 0))



# What clusters can a k-means analysis point out?

gni <- indicators %>%

  filter(IndicatorCode == "NY.GNP.MKTP.CD"

         & CountryName %in% africa) %>%

  select(CountryName,Year,Value)  %>%

  spread(Year,Value,sep="")

country_name <-gni$CountryName

gni <- gni %>%

  select(-CountryName) %>%

 t() %>%
```

```
  scale(center=TRUE,scale=FALSE) %>%
  t()
gni <- data.frame(country_name,gni)
for(i in 1:ncol(gni)) {
  gni[ , i][is.na(gni[ , i])] <- mean(gni[ , i], na.rm = TRUE)
}


# Determine number of clusters
wss<-numeric()
for (i in 2:15) {
  km <-kmeans(select(gni,-country_name),centers=i)
  wss[i]<-sum(km$withinss)/sum(km$totss)
}
plot(1:15, wss, type="b", xlab="Number of Clusters",
ylab="Within groups sum of squares")
set.seed(1001)
fit <- kmeans(select(gni,-country_name),
centers=3,iter.max=100,nstart=1)
gni <- data.frame(gni, fit$cluster)
gni <- gni %>%
  gather(Year,Value,-c(country_name,fit.cluster)) %>%
  group_by(fit.cluster) %>%
  mutate(cluster.count = n_distinct(country_name)) %>%
 ungroup() %>%
  mutate(Year = as.numeric(gsub("Year","",Year)),
         fit.cluster = paste("Cluster ",fit.cluster,"
(",cluster.count," Countries)",sep=""))
```

```
gni$Value <- gni$Value/100000000

ggplot(gni,aes(Year,Value,colour=as.factor(fit.cluster)))+

  geom_point(shape=1)+geom_smooth()+

  scale_colour_discrete(guide=FALSE)+

  labs(title="Changes in GNI Between 1960 and 2014 -
Africa",x="Year",y="GNI (scaled)")+

  facet_wrap(~fit.cluster)


# Where are these clusters located?

map.africa2 <- merge(map_data(map="world",region = africa),

                     select(gni,country_name,fit.cluster),

                     by.x='region',

                     by.y='country_name',

                     all.x=TRUE,

                     fill=0)

map.africa2 <- map.africa2[order(map.africa2$order),]

ggplot(map.africa2) +

  geom_map(map=map.africa2, aes(map_id=region, x=long, y=lat,
fill=fit.cluster)) +

  borders("world",colour="black")+

  scale_fill_discrete(name="")+

 labs(x="Longitude",y="Latitude")+

  theme(legend.position="bottom") +

  coord_map("ortho", orientation = c(10, 15, 0))


# Where in africa gni rates increasing and decreasing?
```

```
gni <- indicators %>%
  filter(IndicatorCode == "NY.GNP.MKTP.CD"
         & CountryName %in% africa) %>%
  select(CountryName,Year,Value) %>%
  group_by(CountryName) %>%
  do(gni.rate=coef(lm(Value~Year,data=.))[[2]])


# Convert from list to numeric.
gni$gni.rate<-as.numeric(unlist(gni$gni.rate))


map.africa3 <- merge(map_data(map="world", region = africa),
gni,  by.x='region', by.y='CountryName', all.x=TRUE, fill=0)
map.africa3 <- map.africa3[order(map.africa3$order),]
ggplot(map.africa3) +
  geom_map(map=map.africa3, aes(map_id=region, x=long, y=lat,
fill=gni.rate)) +
  borders("world",colour="black")+
scale_fill_continuous(low="yellow",high="red",name="Annual GNI
Rate Change - Africa")+
  labs(x="Longitude",y="Latitude")+
theme(legend.position="bottom")  + coord_map("ortho",
orientation = c(10, 15, 0))
```

14. References:

1) https://www.gfmag.com/global-data/economic-data/worlds-unemployment-ratescom
2) https://data.worldbank.org/about
3) https://datatopics.worldbank.org/world-development-indicators/
4) https://en.wikipedia.org/wiki/World_Bank#Open_data_initiative
5) https://www.bls.gov/cps/cps_htgm.htm
6) https://www.kaggle.com/benwatson/worldwide-unemployment?select=Indicators.csv
7) https://pubmed.ncbi.nlm.nih.gov/825171/
   https://www.mapsofworld.com/world-map-image.html#
8) https://www.kaggle.com/worldbank/world-development-indicators
9) https://www.kaggle.com/benhamner/r-ggplot-mapping-example
10) https://data.oecd.org/natincome/gross-national-income.htm
11) https://ontheworldmap.com/
12) https://www.thebalance.com/causes-of-unemployment-7-main-reasons-3305596