National College of Ireland

Configuration Manual

MSc Research Project
Fintech

Salimat Abdullahi
Student ID: 19107111

School of Computing
National College of Ireland

Supervisor: Noel Cosgrave

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Salimat Abdullahi
Student ID:	19107111
Programme:	MSc. Fintech **Year:** 2019-2020
Module:	MSc. Research Project
Lecturer:	Noel Cosgrave
Submission Due Date:	15th of August 2019
Project Title:	An Investigation into the Level of Financial Inclusion among Regional Blocs in Africa
Word Count:	2638 **Page Count:** 19

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: .....

Date: .....

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Salimat Abdullahi
Student ID: 19107111

1 Introduction

This user configuration manual provides details of the technical specifications and procedures necessary to reproduce the analysis under the research titled: "*An investigation into the Level of Financial Inclusion among Regional Blocs in Africa".*

2 System Requirements

2.1. Hardware

- System type: Windows 10, 64-bit Operating System, x64-based processor
- Processor: Intel ® Core™ i5-8265U CPU @ 1.60GHz 1.80GHz
- Installed memory (RAM): 8.00 GB
- Computer name: LAPTOP-O53J3QER

2.2. Software

- R programming language and R studio – Version 4.0.2

3 Data

Findex data should be obtained from World bank's Global Findex Database, 2017[1]. On the other hand, Financial Access Survey (FAS) data should be collected from International Monetary Fund database[2]. Download both datasets in form of csv and save in the document folder as 'project final data.' and 'project final data 2.' respectively.

4 Analysis

4.1. Install Packages

- install.packages("zoo") – For processing time series data[3]

- install.packages("xts") – Similar to *zoo* package

[1] https://globalfindex.worldbank.org/
[2] https://data.imf.org/?sk=E5DCAB7E-A5CA-4892-A6EA-598B5463A34C&sId=1460043522778
[3] https://www.rdocumentation.org/packages/xts/versions/0.12-0/topics/xts

- install.packages("devtools") – This package makes work easier by providing functions that simplify common tasks in R[4]

- install.packages("visdat") – Used to visualize the datasets in terms of missing and observed values.

- install.packages("dplyr") –To manipulate the data[5]

- install.packages("psych") – To perform multivariate analysis and scale construction using principal component analysis[6]. Bartlett's and KMO test can be performed using it.

- install.packages("factoextra") – Provides functions that extract and also visualize the result derived from multivariate data analyses e.g. PCA[7]

- install.packages("missMDA") – For imputing missing values with PCA or MCA[8]

- install.packages("rrcov") – To perform robust PCA, using 'PcaCov'

- install.packages("BiocManager")

- install.packages("pcaMethods")

'pcaMethods' is a package used to aid PPCA imputation. It failed to install directly under this version of R studio. Installation can be done following the process below.

```
if (!requireNamespace("BiocManager", quietly =
TRUE))
  install.packages("BiocManager")

BiocManager::install("pcaMethods")
```

The code was gotten from[9]

[4] https://www.rdocumentation.org/packages/devtools/versions/1.13.6
[5] https://www.rdocumentation.org/packages/dplyr/versions/0.7.8
[6] https://cran.r-project.org/web/packages/psych/index.html#:~:text=A%20general%20purpose%20toolbox%20for,others%20provide%20basic%20descriptive%20statistics.
[7] https://cran.r-project.org/web/packages/factoextra/index.html
[8] https://cran.r-project.org/web/packages/missMDA/missMDA.pdf
[9] https://www.bioconductor.org/packages/release/bioc/html/pcaMethods.html

4.2. Data Pre-processing and Transformation

Although, the purpose is to merge both datasets. Before doing that, each of them should be pre-processed and transformed individually before merging.

4.2.1. Findex Data

- Import dataset

```
project.final.data. <- read.csv("~/project final data..csv", header=TRUE)
```

  - ✓ Store Findex data as 'project.final.data.'

- Remove unwanted columns and rows from the dataset

  - ✓ The columns that contain variables not considered as indicators in the study should be removed.

  - ✓ Also, the rows made up of the countries which do not belong to any of the regional blocs in Africa as well as those countries that are not in FAS data should be removed.

- Rename columns

  - ✓ The first and second column which indicates year and economy are originally stored as 'X' and 'X.2' respectively. To address that, they have to be renamed appropriately as 'Year' and 'Economy'.

- Change index number

  - ✓ Removal of the unwanted rows will alter the numbering of the index numbers. As a result, the row index numbering should be re-ordered in an ascending manner.

- Removal of symbol

  - ✓ The data was collated as percentage. However, the percentage (%) symbol must be removed, and figures should reflect percentage by dividing through by 100. Before the division, be sure of the structure of the dataset. Findex data are stored as character, so, the numeric variables should be transformed from character to numeric while the 'Year' and 'Economy' variables should be left as character.

- Rename Countries

  - ✓ Although, Eswatini was originally called 'Swaziland', but in this dataset, collated data for the country did not reflect its new name, rather, it was stored

as its former name, Swaziland. This must be addressed by renaming the country as Eswatini – just as it appears in FAS data.

- ✓ Rename Egypt, Democratic Republic of Congo, and Congo Republic as they appear in FAS data, to ensure uniformity in both datasets.

```
project.final.data.$Economy[project.final.data.$Economy == "Swaziland"] <-
"Eswatini, Kingdom of"
project.final.data.$Economy[project.final.data.$Economy == "Egypt, Arab Rep."] <-
"Egypt"
project.final.data.$Economy[project.final.data.$Economy == "Congo, Dem. Rep."] <-
"Congo, Democratic Republic of"
project.final.data.$Economy[project.final.data.$Economy == "Congo, Rep."] <-
"Congo, Republic of"
```

- Check the degree and pattern of missingness
  - ✓ Use the 'vis_miss' function to examine the missingness.
- Check for outliers
  - ✓ Use boxplot to examine if there are outliers

4.2.2. FAS Data

- Import dataset

```
project.final.data.2. <- read.csv("~/project final data 2..csv", header=TRUE)
```

- ✓ Store FAS data as 'project.final.data.2.'

- Remove unwanted columns and rows from the dataset
  - ✓ The columns that contain variables not considered as indicators in the study should be removed.
  - ✓ Also, the rows made up of the countries which do not belong to any of the regional blocs in Africa as well as those countries that are not in Findex data should be removed. In addition, data for years other than 2011, 2014, and 2017 must be removed because they are beyond the scope of the study.

- Re-order column

  - ✓ The first two columns in Findex data are ‘Year’ and ‘Economy’, but the reverse is the case for FAS data.

  - ✓ Change the order so that ‘Year’ would also come first in the data, followed by ‘Economy’.

- Change index number

  - ✓ Removal of the unwanted rows will alter the numbering of the index numbers. As a result, the row index numbering should be re-ordered in an ascending manner.

- Change data structure

  - ✓ Change the structure of ‘Year’ from integer to character. Leave the other variables as they are originally stored as their data type.

- Convert NAs in mobile money to zero

  - ✓ Change NAs in both ‘number of mobile money agent outlets per 100,000 adults’ and ‘number of mobile money transactions during the reference year per 1,000 adults’ to zero, with the exclusion of those countries stated in the report to have started mobile money services as at a particular period but have missing values.

```
for (i in
c(1:6,8:9,11:15,19:26,28:38,40:43,46:51,53:63,65:80,82,84:87,89:96,98:108,110:122,124
:135,137:nrow(project.final.data.2.))){
  print
  if (is.na(project.final.data.2.[i,7])==TRUE){
    project.final.data.2.[i,7]=0
  }
}

for (j in
c(1:6,8:12,14,15,19:26,28:36,38:43,45:51,53:63,65:75,77:79,82,84:87,89:93,95,96,98:10
2,104:108,110:122,124:nrow(project.final.data.2.))){
  print
  if (is.na(project.final.data.2.[j,10])==TRUE){
    project.final.data.2.[j,10]=0
  }
}
```

  - ✓ ‘i’ should contain all the row numbers, excluding that of the countries identified to have started mobile money services at a particular period but have

NAs during and after the year of commencement in the variable called 'Number of mobile money agent outlets per 100,000 adults'.

- ✓ 7 indicates the 7th column, which is 'Number of mobile money agent outlets per 100,000 adults.'

- ✓ 'j' should contain all the row numbers, excluding that of the countries identified to have started mobile money services at a particular period but have NAs during and after the year of commencement in the variable called 'Number of mobile money transactions during the reference year per 1,000 adults.'

- ✓ 10 indicates the 10th column, which is 'Number of mobile money transactions during the reference year per 1,000 adults.'

- Check the degree and pattern of missingness

  - ✓ Use the 'vis_miss' function to examine the missingness.

- Check for outliers

  - ✓ Use boxplot to examine if there are outliers

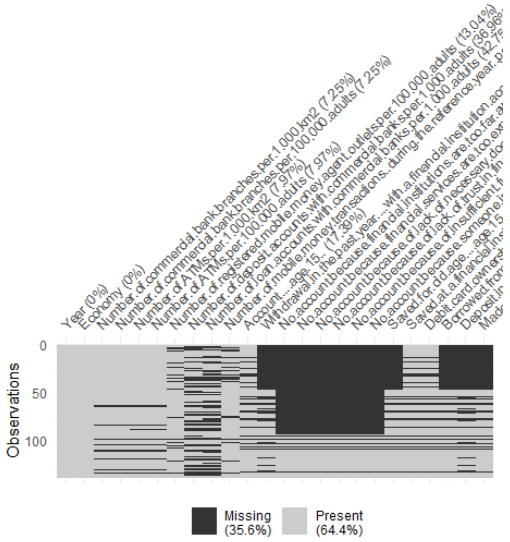
4.2.3. Merged Data (Findex and FAS)

- Combine both datasets

  - ✓ Since the variables 'Year' and 'Economy' are present in both datasets and they contain the same countries, merge both datasets by 'Year' and 'Economy'

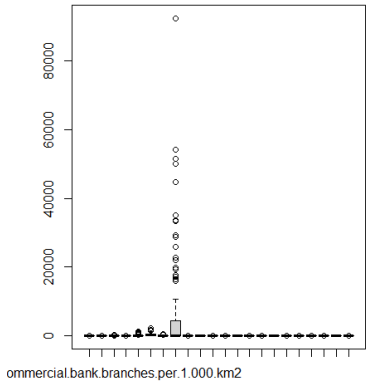
- Check the degree and pattern of missingness

  - ✓ Use the 'vis_miss' function to examine the missingness.



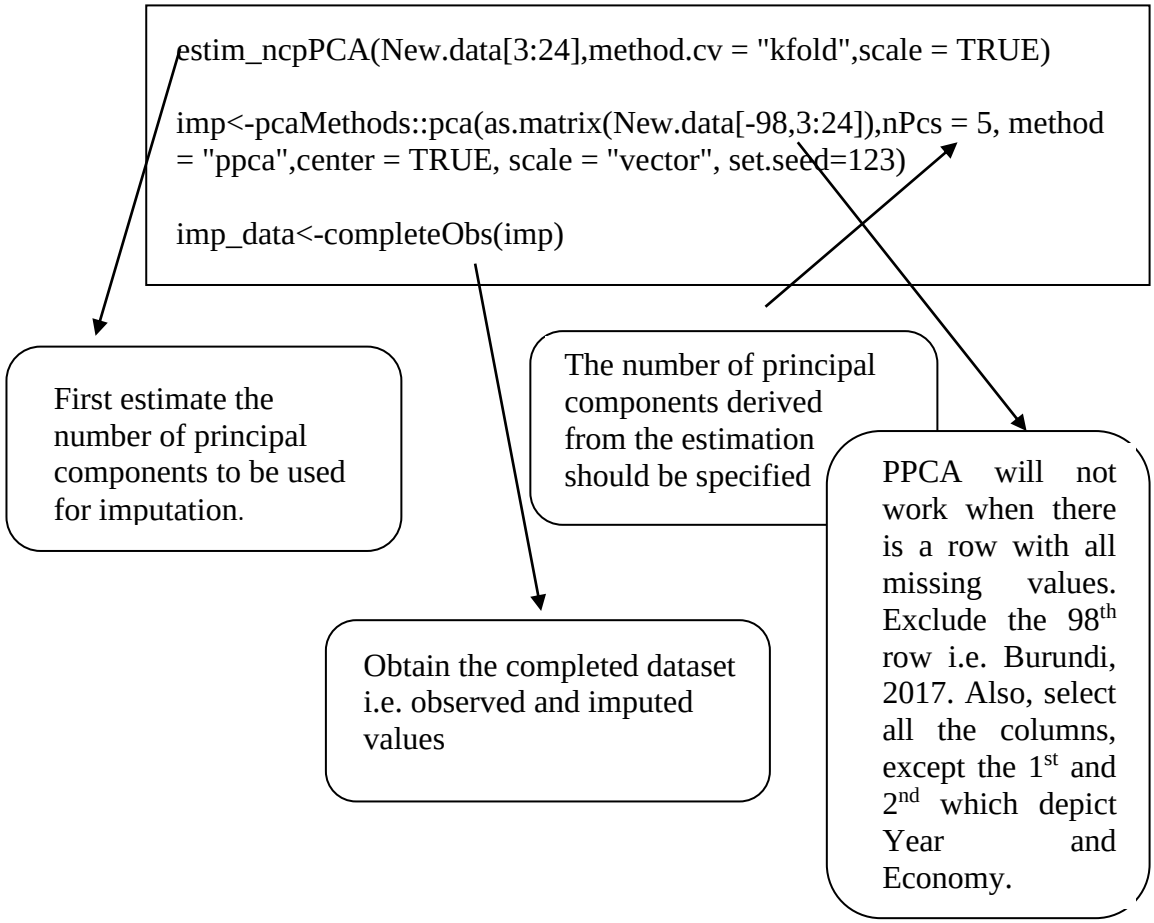
- Check for outliers

  - ✓ Use boxplot to examine if there are outliers



- Impute missing values

  - ✓ Since it has been ascertained that the data is missing at random, with 35.6% missingness, utilize 'PPCA' imputation technique which is capable of handling data that are missing at random and missingness up to 50% (Hegde *et al.*, 2019).



- ✓ After obtaining the completed data, bind it with the 'Year' and 'Economy' columns of the merged data, with the exclusion of the 98th row.

```
#Bind
comp_data = cbind(imp_data, New.data[-98, 1:2])
```

- Re-order column and index number

  - ✓ After binding, 'Year' and 'Economy' will become the last two columns. Re-order so that both variables will occupy the 1st and 2nd columns, respectively.

  - ✓ Removal of the 98th column will alter the numbering of the rows, re-arrange the index numbers in ascending order.

- Create a new column and re-order

  - ✓ Create a new column that will specify the regional blocs each country belongs to. After that, re-arrange the columns so that the newly added column (Region) will be the third variable.

- Create new variables from existing ones

    - ✓ All the variables which make up ‘account’ as a usage indicator should be added together to create a new variable called ‘Account’, as specified in the research report.

    - ✓ The variables which make up ‘savings’ as a usage indicator should be added together to create a new variable called ‘Savings’, as specified in the research report.

    - ✓ The variables which make up ‘loan’ as a usage indicator should be added together to create a new variable called ‘Loan’, as specified in the research report.

    - ✓ To avoid duplicates, the individual variables that are added up should be removed from the dataset.

- Create regional blocs

    - ✓ Define the countries that belong to each regional bloc, and respectively assign them using the variable called ‘Region’.
      e.g.

```
AMU = list(Algeria=")
'AMU_CEN-SAD'=list(Mauritania=",Morocco=",Tunisia=")
'AMU_CEN-SAD_COMESA'=list(Libya=")
```

```
completedData$region[completedData$Economy %in% names(AMU)] <- "AMU"
completedData$region[completedData$Economy %in% names(`AMU_CEN-
SAD`)] <- "AMU,CEN-SAD"
completedData$region[completedData$Economy %in% names(`AMU_CEN-
SAD_COMESA`)] <- "AMU,CEN-SAD,COMESA"
```

    - ✓ Some countries belong to more than one regional bloc. To simplify the process, create a subset of the data set for each regional bloc. In essence, each of the subsets will contain only the countries that make up each bloc over the period of three years.

e.g.

```
AMU_data <-
completedData[c(1,24,28,30,43,47,70,74,76,89,93,115,119,121,134),c(1,
3:17)]

SADC_data<-
completedData[c(2,4,10,11,16,22,25,26,29,31,32,38,41,45,46,48,50,56,5
7,62,68,71,72,75,77,78,84,87,91,92,94,96,101,102,107,113,
116,117,120,122,123,129,132,136,137), c(1,3:17)]
```

- ✓ In each of the subsets i.e. 8 subsets (because there are 8 regional blocs), sum up each of the blocs by year.
  e.g. AMU

```
AMU_data$region='AMU'
AMU_sum <- AMU_data %>%
  group_by(Year, region) %>%
  summarise(across(.cols=everything(),.fns=sum))
```

- ✓ Bind the different subsets of the data.

```
Final_data <-
rbind(AMU_sum,SADC_sum,EAC_sum,ECCAS_sum,ECOWAS_sum,CEN_SAD_su
m,COMESA_sum,IGAD_sum)
```

4.3. Data Mining

4.3.1. Suitability of data

- Check correlation of the data
  - ✓ Use the ‘cor’ function in R.
- Compute Bartlett test
  - ✓ Use the cortest.bartlett function in psych package.

```
cor.d <-cor(Final_data[,3:16])
View(cor.d)
cortest.bartlett(cor.d,n =nrow(Final_data))
```

Output:

```
$chisq
[1] 833.8622

$p.value
[1] 2.093446e-120

$df
[1] 91
```

With p-value, less than 0.05, there is actual correlation in the data. Hence, it is good for analysis

- Perform KMO test

```
KMO(cor.d)
```

✓ Minimum overall measure of sampling adequacy value required to carry out PCA is 0.6[10].

4.3.2. Robust Principal Component Analysis (RPCA)

a) First Stage RPCA

The analysis should be carried out on each of the three dimensions of financial inclusion.

- Access dimension

✓ Use the function 'PcaCov' to perform RPCA on the access dimension

```
pca1 <- PcaCov(Final_data[, c(3:7, 13)], scale = TRUE,center = TRUE)
```

Column number of the six indicators that make up the access dimension

[10] https://www.researchgate.net/post/What_should_be_ideal_KMO_value_for_factor_analysis

✓ Derive the scores, which is equivalent to $P^{a}_{kr\ ,p}$.

```
x<-pca1$scores
```

Output

x[1,1]

x[1,5]

```
          PC1         PC2         PC3         PC4         PC5
1  -0.35620655 -0.60322507 -0.17278082 -0.03928842 -0.048220968
2  -0.07623090 -0.66755104 -0.13436277 -0.06486269 -0.102424386
3   0.19604107 -0.75900565 -0.12107911 -0.01107157 -0.058436890
4   6.74659202  0.26011368  1.42961884  2.11138059  0.987914295
5   8.08708662  0.55320931  1.22709622  2.16027978  1.057422410
6   8.64355421  1.24244131  1.18375994  2.69100896  1.040428856
7  -1.08808450  0.03443905  0.14438782 -0.04540232  0.172695601
8  -0.64138260  0.62784712 -0.02790334 -0.12580615  0.073515639
9  -0.58639158  1.18341638 -0.06121817 -0.13691653  0.043740738
10 -1.07195189 -0.33420830  0.09913235 -0.07624884  0.097987937
11 -0.06468088 -0.03459994  0.53808149  0.06586782  0.028745497
12 -0.43745959  0.11693030 -0.26048113 -0.15020038 -0.295771666
13 -0.76355765 -0.20428586  0.07651270 -0.11872049 -0.214382890
14  0.42432032  0.10053339  0.26789073  0.09299970 -0.169430431
15  1.35440632  1.84082390 -0.15265226  0.05710034 -0.080140947
16  1.79528596 -0.62649367 -0.19248625 -0.08093214  0.096274032
17  3.31409420 -0.36939281  0.08288591  0.04455958  0.084088945
18  4.82014370  1.18638488 -0.38487768  0.18888450  0.354440779
19  5.45585477  1.11445613  2.50089167  1.23157993  0.965595665
20  6.95357899  1.69331213  2.76487043  1.31320525  0.956347687
21  8.43144556  3.05934734  2.74065542  2.00015020  0.984953530
22 -1.47852077 -0.25604223 -0.16714637  0.18130781  0.015280412
23 -0.95714054  0.06774472 -0.17926192  0.25741390  0.060707712
24  0.08728937  0.62052289  0.56580713  0.38060362  0.006418249
          PC6
1  -3.611061e-02
2   2.276221e-02
3   6.014192e-02
4   3.000648e-01
5   6.792814e-01
6   7.684434e-01
7  -4.096957e-02
8   9.077498e-02
9  -3.066451e-02
10  1.412682e-02
11  9.279625e-05
12  5.981335e-01
13 -1.990097e-02
14 -2.075283e-02
15 -4.467951e-03
16 -6.721258e-02
17  3.232921e-02
18  1.012529e-01
19 -1.256803e+00
20 -1.201409e+00
21 -1.332091e+00
22 -1.795276e-03
23  1.646357e-03
24 -2.520653e-01
```

NOTE

Row 1 refers to the PCs for AMU region in 2011
Row 2 refers to the PCs for AMU region in 2014
Row 3 refers to the PCs for AMU region in 2017

The cycle continues like that in the following order:
AMU ➔ SADC ➔ EAC ➔ ECCAS ➔ ECOWAS ➔ CEN-SAD ➔ COMESA ➔ IGAD
*The 24 rows are as a result of 8 regions for 3 years (8 * 3 =24).

Hence, Row 22 refers to the PC values for IGAD region in 2011
Row 23 refers to the PC values for IGAD region in 2014
Row 24 refers to the PC values for IGAD region in 2017

** In essence, the scores produce the principal components for each region at a particular period.

**The scores above represent that of the ACCESS dimension = $P^{a}_{kr,p}$.

- ✓ Derive the eigen values which is equivalent to λ^{a}_{j}

```
variance_a = pca1$eigenvalues
```

Output

```
[1] 4.417440206 1.398267567 0.112910701 0.035526153 0.031486833
                                [6] 0.004368541
```

variance_a[1]

variance_a[6]

- ✓ Substitute the values obtained in the formula below to derive the access score for each region over the period of 2011, 2014, and 2017.

$$A^{a}_{r,p} = \frac{\sum^{p}_{j,k=1} \lambda^{a}_{j} P^{a}_{kr,p}}{\sum^{p}_{j=1} \lambda^{a}_{j}} \qquad (5)$$

- Usage dimension

  ✓ Use the function ‘PcaCov’ to perform RPCA on the usage dimension

```
pca12 <- PcaCov(Final_data[, c(14:16)], scale = TRUE,center = TRUE)
```

Column number of the three indicators that make up the usage dimension

  ✓ Derive the scores, which is equivalent to $P^{u}_{kr,p}$

```
y<-pca12$scores
```

Output

y[1,1]

y[1,3]

	PC1	PC2	PC3
1	-0.64856242	-0.583232507	0.584335666
2	-0.62173856	-0.581563681	0.502588104
3	-0.24381477	-0.756554642	0.957437641
4	1.86878677	-0.929994935	0.587450788
5	2.13557284	-0.239258048	0.561182913
6	2.12313672	1.419449022	0.209507548
7	-1.00862346	-0.202530584	-0.050215353
8	-0.45659062	0.787349063	0.052320634
9	-0.52034659	1.507371607	-0.091766251
10	-0.86367968	-0.583560419	0.062204283
11	-0.67556956	-0.490534700	0.032605980
12	-0.55017627	0.177692810	0.042075108
13	-0.37747558	-0.422678373	-0.084837551
14	0.02319743	-0.348560554	0.044865918
15	0.34823180	1.745189579	-0.438259954
16	0.99633542	-0.778173226	0.005225124
17	1.44123008	-0.708404539	0.168865108
18	2.14973844	1.153474745	0.086058711
19	0.83038830	-0.625241467	-0.264152000
20	1.29240554	-0.005462348	-0.091885662
21	2.02725527	1.930537234	-0.317547547
22	-1.10675888	-0.244942685	0.094586058
23	-0.75823108	0.150811573	0.005679968
24	-0.41584348	0.633389097	-0.011630076

NOTE

Row 1 refers to the PCs for AMU region in 2011
Row 2 refers to the PCs for AMU region in 2014
Row 3 refers to the PCs for AMU region in 2017

The cycle continues like that in the following order:
AMU ➔ SADC ➔ EAC ➔ ECCAS ➔ ECOWAS ➔ CEN-SAD ➔ COMESA ➔ IGAD
*The 24 rows are as a result of 8 regions for 3 years (8 * 3 =24).

Hence, Row 22 refers to the PC values for IGAD region in 2011
Row 23 refers to the PC values for IGAD region in 2014
Row 24 refers to the PC values for IGAD region in 2017

** In essence, the scores produce the principal components for each region at a particular period.

**The scores above represent that of the USAGE dimension = $P^u_{kr,p}$.

- ✓ Derive the eigen values which is equivalent to λ^u_j

```
variance_u = pca12$eigenvalues
```

Output

```
[1] 2.02424953 0.95534902 0.02040145
```

variance_u[1]

variance_u[3]

- ✓ Substitute the values obtained in the formula below to derive the usage score for each region over the period of 2011, 2014, and 2017.

$$A^u_{r,p} = \frac{\sum^p_{j,k=1} \lambda^u_j P^u_{kr,p}}{\sum^p_{j=1} \lambda^u_j} \qquad (6)$$

- Quality dimension

  ✓ Use the function 'PcaCov' to perform RPCA on the quality dimension

```
pca13 <- PcaCov(Final_data[, c(8:12)], scale = TRUE,center = TRUE)
```

Column number of the five indicators that make up the quality dimension

  ✓ Derive the scores, which is equivalent to $P^{q}_{kr\,,p}$

```
z<-pca13$scores
```

Output

	PC1	PC2	PC3	PC4	PC5
1	-1.6325055	-0.124393311	-0.027998661	0.0150767687	6.040508e-05
2	-1.6881417	-0.119614818	-0.042223724	0.0161798517	-1.507989e-04
3	-1.8763650	-0.255230884	-0.186175257	-0.0444599603	2.367551e-02
4	0.7345031	-0.136198970	-0.031201842	-0.0206770955	-1.085186e-03
5	0.3377367	-0.048418983	-0.135629485	-0.0253396257	-1.775411e-04
6	0.9834588	-0.076099449	0.237673858	-0.0452672664	4.871964e-02
7	-1.3330421	0.035279654	0.061281834	-0.0064698704	-4.190770e-04
8	-1.5521254	0.069276264	-0.018445765	-0.0058969772	3.327893e-03
9	-1.5666675	-0.031956755	0.001682632	0.1174312218	6.074680e-02
10	-0.3045797	0.003747999	0.089114530	-0.0008814180	-1.037567e-03
11	-0.7304184	0.126838715	-0.054991836	0.0114476306	-2.260745e-03
12	-0.5175948	-0.194098111	-0.048309165	-0.0099126389	-4.385959e-02
13	0.5846577	0.008045043	0.091334532	0.0005457616	3.904021e-04
14	0.2939548	0.042560336	-0.026987544	0.0167134050	3.841131e-04
15	0.4296855	0.091510334	0.114135173	-0.0249523271	-4.203403e-02
16	2.8765150	-0.083963740	0.071961534	0.0038186304	1.175606e-03
17	2.4471515	-0.009562173	-0.076811780	0.0180085654	4.233605e-04
18	2.5276594	-0.187194782	-0.072038392	-0.0470880900	2.249981e-02
19	1.5840766	0.041543410	0.077512330	-0.0046029371	-7.702335e-04
20	1.0561344	0.168918679	-0.070445910	-0.0049034696	5.020573e-05
21	1.1485232	0.026651603	0.042209975	-0.0743030233	7.118080e-02
22	-1.2363126	0.002620914	0.066350228	-0.0027051279	-6.117913e-04
23	-1.4376042	0.023320983	0.027181559	-0.0103140920	7.009542e-04
24	-1.3254658	-0.063861038	-0.006415705	0.0897050969	-1.725202e-02

NOTE

Row 1 refers to the PCs for AMU region in 2011
Row 2 refers to the PCs for AMU region in 2014
Row 3 refers to the PCs for AMU region in 2017

The cycle continues like that in the following order:
AMU ➔ SADC ➔ EAC ➔ ECCAS ➔ ECOWAS ➔ CEN-SAD ➔ COMESA ➔ IGAD
*The 24 rows are as a result of 8 regions for 3 years (8 * 3 =24).

Hence, Row 22 refers to the PC values for IGAD region in 2011
Row 23 refers to the PC values for IGAD region in 2014
Row 24 refers to the PC values for IGAD region in 2017

** In essence, the scores produce the principal components for each region at a particular period.

**The scores above represent that of the QUALITY dimension = $P^{q}_{kr\,,p}$.

- ✓ Derive the eigen values which is equivalent to λ^{q}_{j}

```
variance_q = pca13$eigenvalues
```

Output

```
[1] 4.971513e+00 1.702666e-02 1.107270e-02 3.846168e-04 3.324665e-06
```

variance_q[1]

variance_q[3]

- ✓ Substitute the values obtained in the formula below to derive the quality score for each region over the period of 2011, 2014, and 2017.

$$A^{q}_{r\,,p} = \frac{\sum^{p}_{j,k=1} \lambda^{q}_{j} P^{q}_{kr\,,p}}{\sum^{p}_{j=1} \lambda^{q}_{j}} \qquad (7)$$

b) Second Stage RPCA

- Create new data frame

  ✓ The values obtained from the equations above in relation to the index scores of each dimension for all the regions over the three years period should be used to create another data frame. However, the new data frame should consist of 24 rows and 5 variables (Region, Year, Access, Usage and Quality).

- Perform RPCA on the new variables

  ✓ The new variables are Access, Usage and Quality.

  ✓ Like the first stage RPCA, use the function 'PcaCov' to perform RPCA on the new variables.

```
pca_final <- PcaCov(dimensions[, c(3:5)])
```

Column number of the three dimensions (access, usage, and quality).

  ✓ Derive the loadings, which is equivalent to φ

```
vec<-pca_final$loadings
```

Output

Vec[1,3] or φ_{13}

```
            PC1        PC2         PC3
Access  0.4495781  0.7033650 -0.55059710
Usage   0.3110010  0.4545793  0.83464729
Quality 0.8373517 -0.5464754 -0.01437834
```

  ✓ To determine the $P_{kr,p}$ values in this stage, the loadings as well as the values estimated in the first stage RPCA should be used in the formula below:

$$P_{1r,p} = \varphi_{11}A^a_{r,p} + \varphi_{12}A^u_{r,p} + \varphi_{13}A^q_{r,p} \quad (9)$$

$$P_{2r,p} = \varphi_{21}A^a_{r,p} + \varphi_{22}A^u_{r,p} + \varphi_{23}A^q_{r,p} \quad (10)$$

$$P_{3r,p} = \varphi_{31}A^a_{r,p} + \varphi_{32}A^u_{r,p} + \varphi_{33}A^q_{r,p} \quad (11)$$

- ✓ Derive the eigen values which is equivalent to λ_j

```
variance_final<-pca_final$eigenvalues
```

Output

```
[1] 1.80196373 0.87091087 0.09312332
```

variance_final[1]

variance_final[3]

- Estimate Financial inclusion index score
  - ✓ The financial inclusion index score for each region should be calculated by substituting the values estimated in the formula below.

$$FI_{r,p} = \frac{\sum_{j=1}^{3} \lambda_j (\varphi_{j1} A_{r,p}^{a} + \varphi_{j2} A_{r,p}^{u} + \varphi_{j3} A_{r,p}^{q}}{\sum_{j=1}^{3} \lambda_j} \qquad (12)$$

References

Hegde, H. *et al.* (2019) 'MICE vs PPCA: Missing data imputation in healthcare', *Informatics in Medicine Unlocked*, 17, p. 100275. doi: 10.1016/j.imu.2019.100275.