Lab 2

Grupo:

- Eduardo Amorim Vilela de Salis - 11226805
- Débora Mauricio Kono - 9896754
- Douglas Silva Cardosos - 11766990

Biblioteca

```
library(WDI) # baixar os dados do World Bank
library(magrittr)
library(tidyverse)
```

```
── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
✔ dplyr     1.1.0     ✔ readr     2.1.4
✔ forcats   1.0.0     ✔ stringr   1.5.0
✔ ggplot2   3.4.1     ✔ tibble    3.2.0
✔ lubridate 1.9.2     ✔ tidyr     1.3.0
✔ purrr     1.0.1
── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
✖ tidyr::extract()   masks magrittr::extract()
✖ dplyr::filter()    masks stats::filter()
✖ dplyr::lag()       masks stats::lag()
✖ purrr::set_names() masks magrittr::set_names()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts
to become errors
```

```
library(ggplot2)
library(dplyr)
library(cluster)
library(fpc)


library(formattable)
```

```
pre_process_df <- function(df){
  df <- subset(df, region != "Aggregates")

  df$region |> as.character()

  dfi <- df[, lista_indicadores]
  row.names(dfi) <- df$country
  colnames(dfi) <- c("Inflacao", "PIB_per_Capita", "Crescimento_PIB", "Desemprego")

  dfi <- na.omit(dfi)
  dfi$Desemprego <- 100 - dfi$Desemprego
  names(dfi)[4] <- "Emprego"
```

```
  return(dfi)

}
```

KEANS

Análise para 2014

```
lista_indicadores <- c("FP.CPI.TOTL.ZG", # inflação (%)
                       "NY.GDP.PCAP.CD", # Pib per capita (USD)
                       "NY.GDP.MKTP.KD.ZG", # crescimento do PIB anual (%),
                       "SL.UEM.TOTL.ZS" # Desemprego (%)
)



df2014 <- WDI(indicator = lista_indicadores, country = "all", start = 2014, end = 2014
              extra = TRUE)
str(df2014 )
```

```
'data.frame':   266 obs. of  16 variables:
 $ country          : chr  "Afghanistan" "Africa Eastern and Southern" "Africa Western
and Central" "Albania" ...
 $ iso2c            : chr  "AF" "ZH" "ZI" "AL" ...
 $ iso3c            : chr  "AFG" "AFE" "AFW" "ALB" ...
 $ year             : int  2014 2014 2014 2014 2014 2014 2014 2014 2014 2014 ...
 $ status           : chr  "" "" "" "" ...
 $ lastupdated      : chr  "2023-05-10" "2023-05-10" "2023-05-10" "2023-05-10" ...
 $ FP.CPI.TOTL.ZG   : num  4.67 5.37 1.77 1.63 2.92 ...
  ..- attr(*, "label")= chr "Inflation, consumer prices (annual %)"
 $ NY.GDP.PCAP.CD   : num  628 1719 2243 4579 5516 ...
  ..- attr(*, "label")= chr "GDP per capita (current US$)"
 $ NY.GDP.MKTP.KD.ZG: num  2.72 4.04 5.93 1.77 3.8 ...
  ..- attr(*, "label")= chr "GDP growth (annual %)"
 $ SL.UEM.TOTL.ZS   : num  7.91 6.56 3.99 18.05 10.21 ...
  ..- attr(*, "label")= chr "Unemployment, total (% of total labor force) (modeled ILO
estimate)"
 $ region           : chr  "South Asia" "Aggregates" "Aggregates" "Europe & Central
Asia" ...
 $ capital          : chr  "Kabul" "" "" "Tirane" ...
 $ longitude        : chr  "69.1761" "" "" "19.8172" ...
 $ latitude         : chr  "34.5228" "" "" "41.3317" ...
 $ income           : chr  "Low income" "Aggregates" "Aggregates" "Upper middle
income" ...
 $ lending          : chr  "IDA" "Aggregates" "Aggregates" "IBRD" ...
```

```
dfi2014 <- pre_process_df(df2014)
```

```
dfi2014 |> str()
```

```
'data.frame':   171 obs. of  4 variables:
 $ Inflacao       : num  4.67 1.63 2.92 7.28 2.98 ...
 $ PIB_per_Capita : num  628 4579 5516 5059 4017 ...
 $ Crescimento_PIB: num  2.72 1.77 3.8 4.82 3.6 ...
 $ Emprego        : num  92.1 82 89.8 90.4 88.1 ...
 - attr(*, "na.action")= 'omit' Named int [1:45] 4 5 7 8 10 22 28 37 40 44 ...
  ..- attr(*, "names")= chr [1:45] "American Samoa" "Andorra" "Antigua and Barbuda"
"Argentina" ...
```

```
dfi2014 |> summary()
```

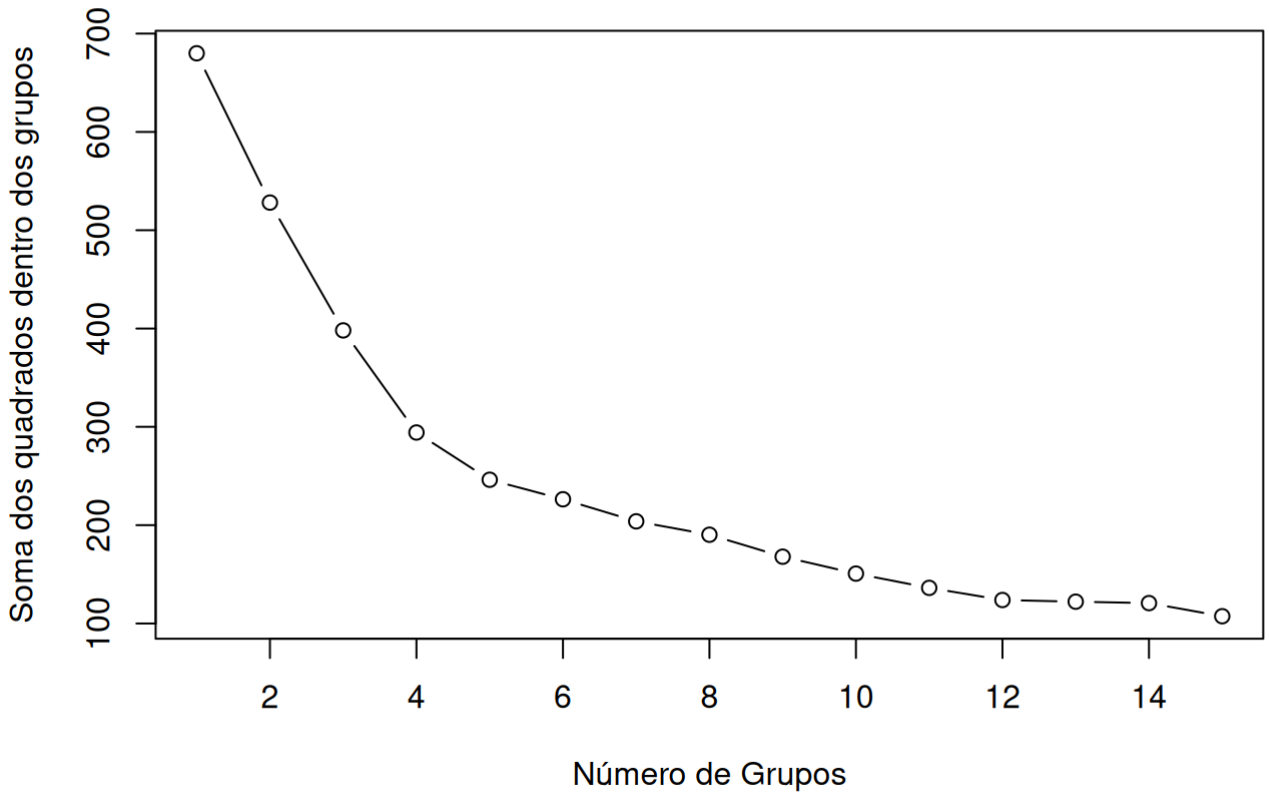
```
    Inflacao       PIB_per_Capita     Crescimento_PIB      Emprego
 Min.   :-1.5092   Min.   :   257.8   Min.   :-10.079   Min.   :71.62
 1st Qu.: 0.8682   1st Qu.:  1928.2   1st Qu.:  1.599   1st Qu.:89.36
 Median : 2.7592   Median :  5544.1   Median :  3.537   Median :93.45
 Mean   : 4.0680   Mean   : 15240.7   Mean   :  3.556   Mean   :91.97
 3rd Qu.: 5.6829   3rd Qu.: 16899.8   3rd Qu.:  5.397   3rd Qu.:96.43
 Max.   :62.1686   Max.   :123678.7   Max.   : 19.047   Max.   :99.80
```

```
dfi2014 |> head()
```

```
            Inflacao PIB_per_Capita Crescimento_PIB Emprego
Afghanistan 4.673996       628.1468        2.724543  92.090
Albania     1.625865      4578.6332        1.774449  81.950
Algeria     2.916927      5516.2306        3.800000  89.790
Angola      7.280387      5059.0804        4.820000  90.420
Armenia     2.981309      4017.2298        3.600000  88.138
Australia   2.487923     62513.4112        2.579017  93.920
```

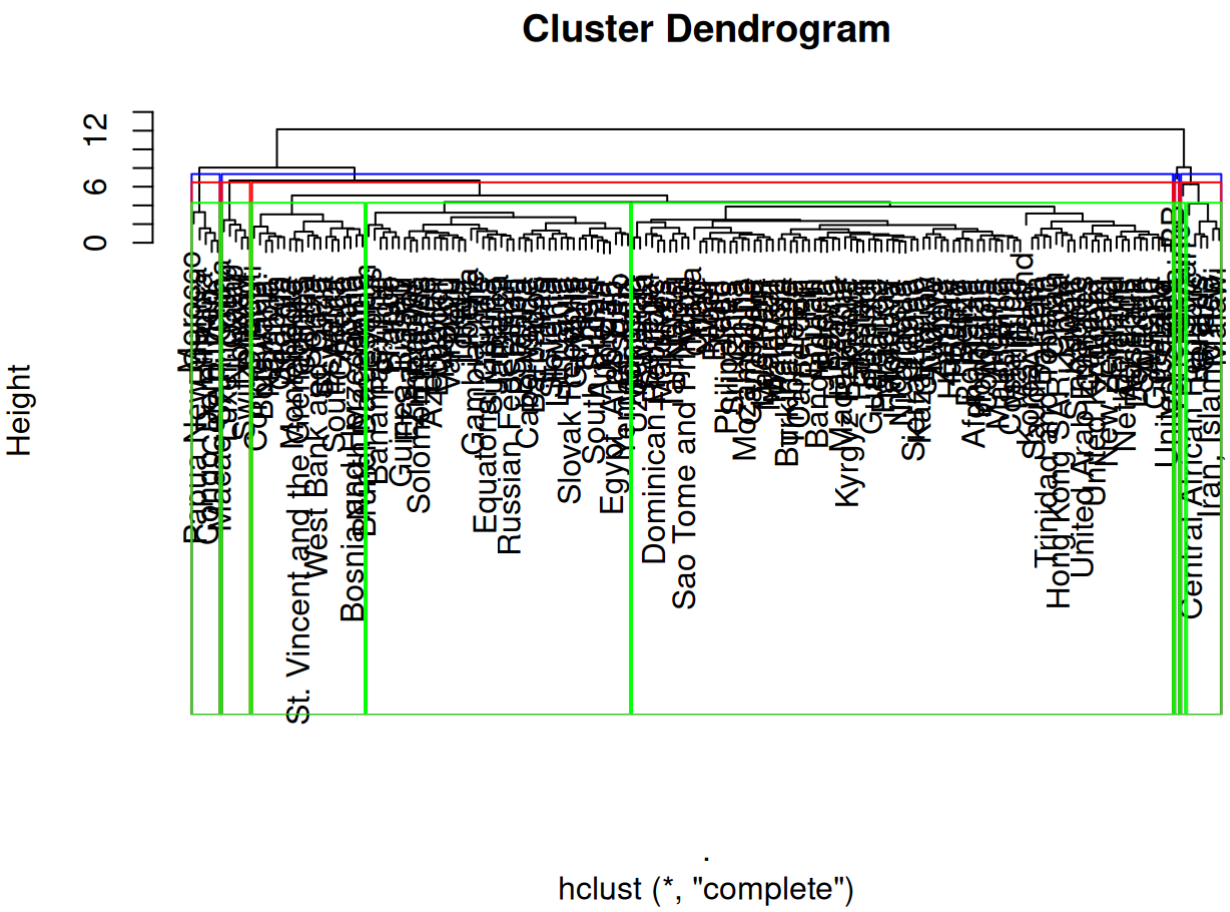
Determinando Quantidade de grupos

```
dfi2014_escala <- scale(dfi2014)



wss <- (nrow(dfi2014_escala)-1)*sum(apply(dfi2014_escala,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(dfi2014_escala,centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Número de Grupos",ylab="Soma dos quadrados dentro dos
```

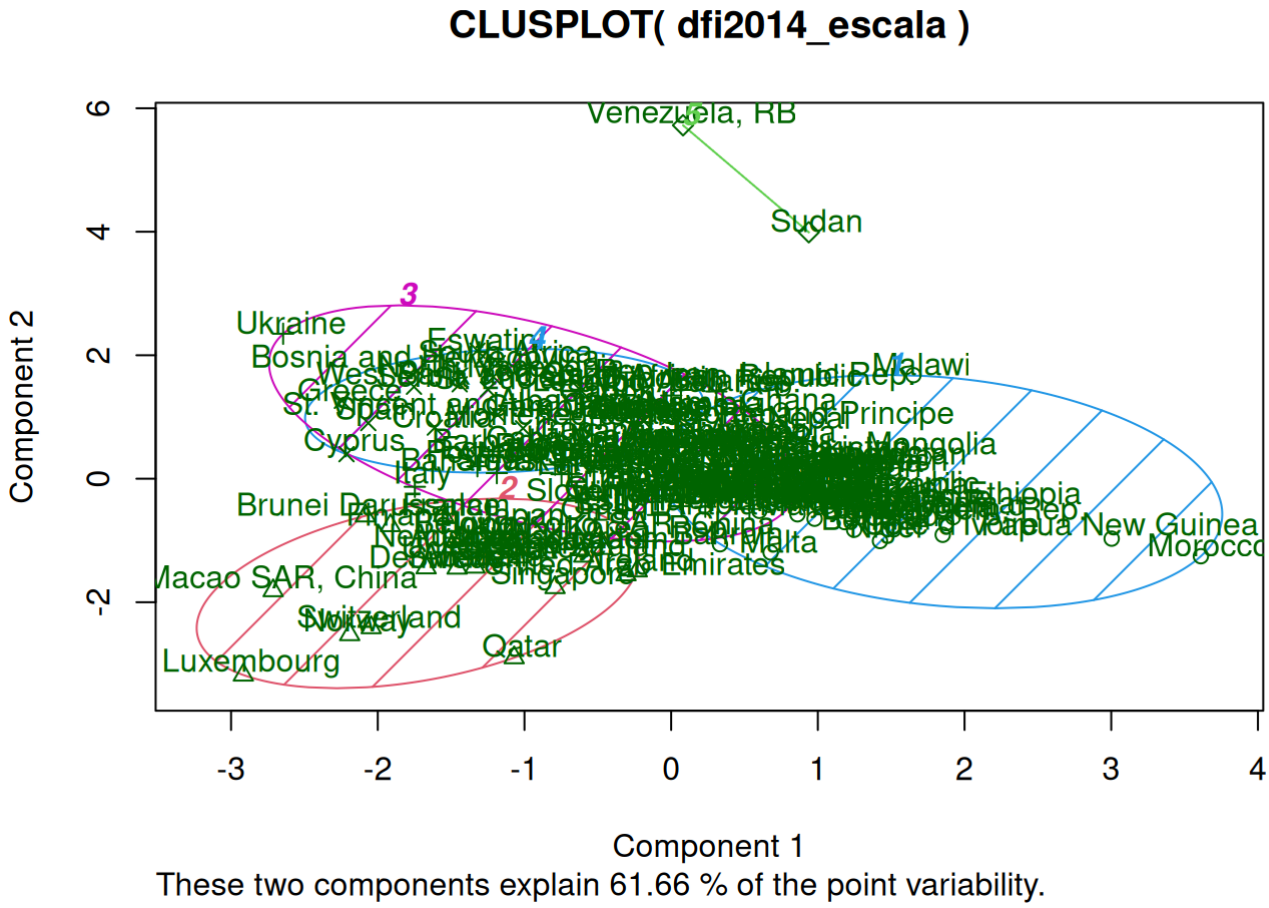


Plotando os clusters

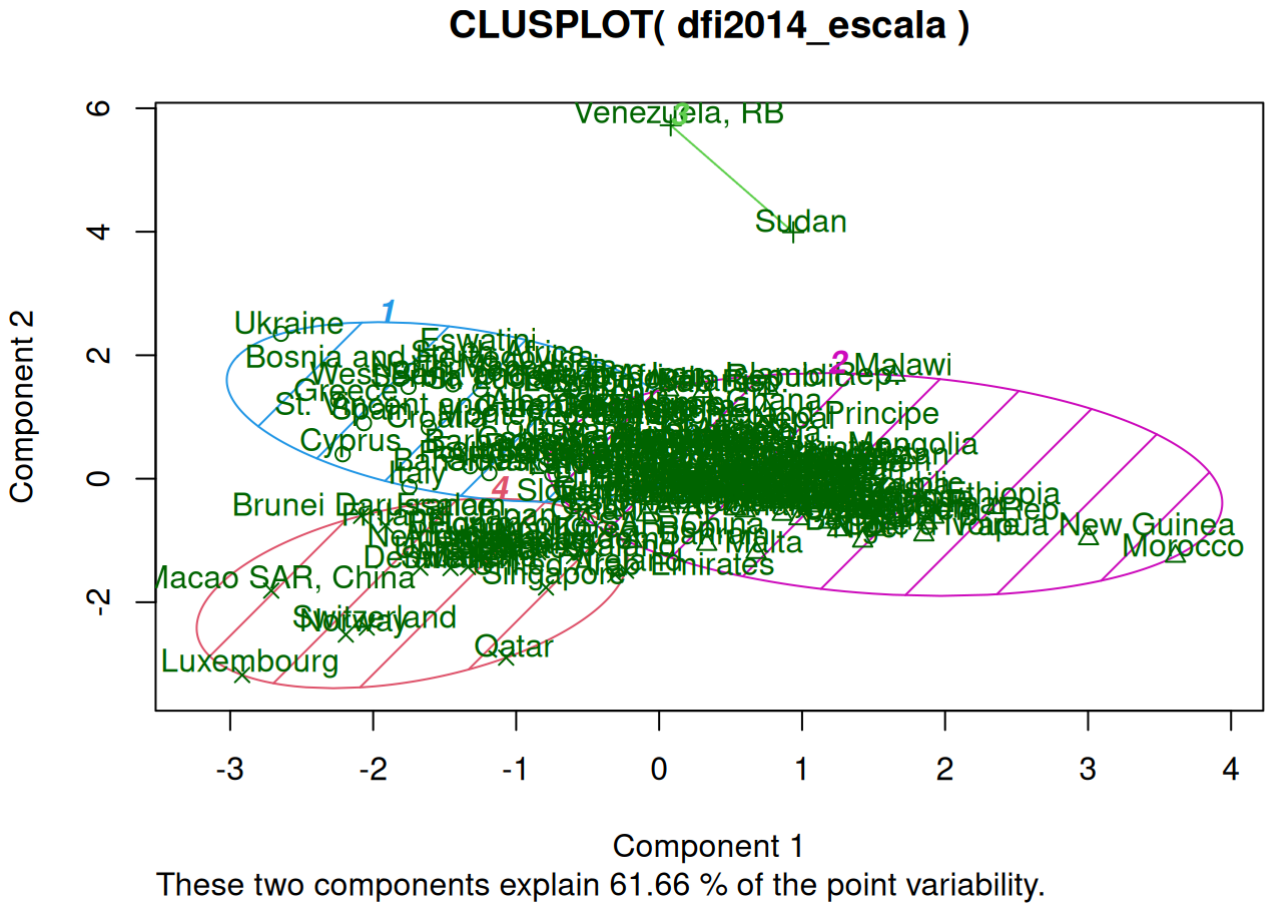
```
dendo <- dfi2014_escala %>% dist() %>% hclust()
plot(dendo)
rect.hclust(dendo, k = 4, border = "blue")
rect.hclust(dendo, k = 5, border = "red")
rect.hclust(dendo, k = 8, border = "green")
```




```
library(cluster)
library(fpc)

grupos <- kmeans(dfi2014_escala, centers=5)
clusplot(dfi2014_escala, grupos$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```




```
grupos <- kmeans(dfi2014_escala, centers=4)
clusplot(dfi2014_escala, grupos$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



Plotando os clusters

```
dfi2014_escala[c("Brazil", "Chile", "Colombia", "Norway", "United States"),] %>% dist(
```

```
               Brazil    Chile Colombia   Norway
Chile         0.499762
Colombia      1.435620 1.013618
Norway        4.017933 3.836570 4.241284
United States 2.185957 1.925989 2.311189 1.975246
```

Países com MENOR dissimilaridade em relação ao Brasil

```
mat_brasil <- dfi2014_escala %>% dist(diag = TRUE, upper = TRUE) %>% as.matrix()

mat_brasil[, "Brazil"] %>% sort() %>% head(5)
```

```
          Brazil Russian Federation  Equatorial Guinea           Suriname
       0.0000000          0.3694132          0.4409301          0.4849345
           Chile
       0.4997620
```

Países com MAIOR dissimilaridade em relação ao Brasil

```
mat_brasil[, "Brazil"] %>% sort() %>% tail(5)
```

```
Papua New Guinea       Luxembourg            Sudan          Morocco
        4.286916         5.243560         5.262285         6.075438
   Venezuela, RB
        8.814588
```

Estatística por cluster

```
set.seed(123)
lista_clusteres <- kmeans(dfi2014_escala, centers = 5)$cluster


dfi2014_com_cluster <- dfi2014 |>
  mutate(cluster = lista_clusteres)


stats_cluster <- dfi2014_com_cluster |>
  group_by(cluster) |>
  summarise(
    qtd = n(),
    Media_inflacao = mean(Inflacao, na.rm = TRUE),
    Media_pibpc = mean(PIB_per_Capita, na.rm = TRUE),
    Media_cresciemento = mean(Crescimento_PIB),
    Media_emprego = mean(Emprego, na.rm = TRUE))

stats_cluster
```

```
# A tibble: 5 × 6
  cluster   qtd Media_inflacao Media_pibpc Media_cresciemento Media_emprego
    <int> <int>          <dbl>       <dbl>              <dbl>         <dbl>
1       1     2          49.5        9026.              0.383          87.9
2       2    27           1.57      58757.              2.19           94.4
3       3     6           3.20       1761.             11.6            96.2
4       4    44           2.33       9828.              1.63           83.9
5       5    92           4.70       6072.              4.42           95.0
```

```
dfi2014_com_cluster["Brazil",]
```

```
       Inflacao PIB_per_Capita Crescimento_PIB Emprego cluster
Brazil  6.32904       12071.16       0.5039557   93.24       5
```

```
dfi2014_com_cluster |> filter(cluster == 3) # paises realmente parecidos
```

```
                 Inflacao PIB_per_Capita Crescimento_PIB Emprego cluster
Congo, Dem. Rep. 1.2430389       472.2662        9.470288  95.551       3
Cote d'Ivoire    0.4486821      2124.0196        9.372000  96.525       3
Ethiopia         6.8900195       557.5341       10.257493  97.588       3
Morocco          0.4423101      3430.5496       19.047279  90.300       3
```

```
Myanmar          4.9532992      1238.7287        8.199664  99.277       3
Papua New Guinea 5.2221172      2742.2333       13.543771  97.660       3
```

```
dfi2014_com_cluster |> filter(cluster == 2) # paises desenvolvidos
```

```
                      Inflacao PIB_per_Capita Crescimento_PIB Emprego cluster
Australia           2.48792271       62513.41       2.5790171  93.920       2
Austria             1.60581183       51786.38       0.6612728  94.380       2
Belgium             0.34000283       47764.07       1.5785331  91.480       2
Brunei Darussalam  -0.20710873       41037.07      -2.5083526  93.140       2
Canada              1.90663591       50956.00       2.8700361  93.090       2
Denmark             0.56402054       62548.98       1.6193938  93.070       2
Finland             1.04119621       50327.24      -0.3649082  91.340       2
France              0.50775882       43068.55       0.9561831  89.710       2
Germany             0.90679400       48023.87       2.2095434  95.020       2
Hong Kong SAR, China 4.42364532      40315.29       2.7624196  96.700       2
Iceland             2.04461482       54576.74       1.6872150  95.100       2
Ireland             0.18254232       55643.06       8.6493506  88.140       2
Israel              0.48649555       38259.68       3.9191269  94.110       2
Japan               2.75922671       38475.40       0.2962055  96.410       2
Kuwait              2.90892673       43234.82       0.5008770  97.100       2
Luxembourg          0.62854399      123678.70       2.6230860  94.150       2
Macao SAR, China    6.04823452       90873.93      -2.0483806  98.330       2
Netherlands         0.97603508       52900.54       1.4233954  92.580       2
New Zealand         1.22750751       44572.90       3.8154276  94.570       2
Norway              2.04170287       97019.18       1.9695443  96.520       2
Qatar               3.34972086       93126.15       5.3343233  99.800       2
Singapore           1.02514803       57562.53       3.9355403  96.260       2
Sweden             -0.17963849       60020.36       2.6577983  92.050       2
Switzerland        -0.01320254       88724.99       2.3498813  95.170       2
United Arab Emirates 2.34626866      46865.96       4.1656918  98.098       2
United Kingdom      1.45112016       47447.59       3.1997026  93.890       2
United States       1.62222298       55123.85       2.2877759  93.830       2
```

```
dfi2014_com_cluster |> filter(cluster == 5) # paises subdesenvolvidos / emergentes
```

```
                        Inflacao PIB_per_Capita Crescimento_PIB Emprego
Afghanistan           4.67399604       628.1468      2.72454336  92.090
Algeria               2.91692692      5516.2306      3.80000000  89.790
Angola                7.28038730      5059.0804      4.82000000  90.420
Azerbaijan            1.37344182      7891.3131      2.75050682  95.090
Bahrain               2.64755321     25464.7601      4.35039085  98.816
Bangladesh            6.99163889      1108.5151      6.06105936  95.607
Belarus              18.11955435      8341.3997      1.72638485  94.004
Belize                1.20139964      6068.0886      4.08958130  91.760
Benin                -0.54875755      1251.5048      6.35767910  97.889
Bhutan                8.27106094      2589.8998      5.77649568  97.370
Bolivia               5.76660075      3022.4629      5.46056951  97.980
Brazil                6.32904016     12071.1582      0.50395574  93.240
Burkina Faso         -0.25808952       767.3714      4.32684561  95.794
Burundi               4.40535234       257.8186      4.24065021  98.430
Cambodia              3.85568861      1098.0745      7.14257110  99.310
Cameroon              1.85489850      1631.7141      5.71981814  96.470
```

Central African Republic	14.89868418	394.8570	0.08107052	94.423
Chad	1.68197314	1017.7878	6.89998505	98.958
Chile	4.71867528	14666.3435	1.79264947	93.350
China	1.92164163	7636.1166	7.42576366	95.370
Colombia	2.89883788	8164.7145	4.49903000	91.430
Costa Rica	4.51920099	10737.6789	3.54210988	91.500
Dominican Republic	2.99864226	6533.6669	7.05046369	93.280
Ecuador	3.58922017	6374.6315	3.78886855	96.520
El Salvador	1.14134468	3638.5177	1.70870627	95.850
Estonia	-0.10617515	20261.0667	3.01136659	92.650
Fiji	0.51930647	5305.1909	5.60351489	95.723
Ghana	15.48961603	1942.9051	2.85624016	95.505
Guatemala	3.41836170	3779.6423	4.44397758	97.280
Guinea	6.15050557	774.5690	3.69655312	95.084
Guinea-Bissau	-1.50924461	605.1226	0.96456075	96.837
Honduras	6.12924930	2164.4202	3.05808056	92.920
Hungary	-0.22756627	14294.2584	4.23220981	92.270
India	6.66565672	1559.8645	7.41022761	92.019
Indonesia	6.39492541	3476.6249	5.00666843	95.950
Iran, Islamic Rep.	16.60655324	5757.5433	4.98477507	89.320
Jamaica	8.27407886	4991.5655	0.68982236	90.950
Kazakhstan	6.70657829	12807.2607	4.20000000	94.940
Kenya	6.87815499	1489.9191	5.02011100	97.204
Korea, Rep.	1.27477446	29249.5752	3.20245379	96.920
Kyrgyz Republic	7.53424730	1279.7698	4.02403863	96.717
Lao PDR	4.12924307	1984.5087	7.61196344	97.854
Lebanon	1.85460421	7665.3797	2.48406011	91.235
Liberia	9.86111286	713.7349	0.70139310	97.920
Madagascar	6.08040811	517.1362	3.33920311	98.609
Malawi	23.79206495	367.0243	5.70000001	95.032
Malaysia	3.14299051	11045.4451	6.00672195	97.120
Maldives	2.12000176	8872.1249	7.32962620	92.479
Mali	0.88381455	818.4304	7.08468388	98.651
Malta	0.31030647	26754.2623	7.63311934	94.270
Mauritania	3.53436856	1715.3888	4.27482327	89.934
Mauritius	3.21769192	10368.6134	3.82696982	92.530
Mexico	4.01861608	11076.0925	2.84977325	95.190
Moldova	5.08878555	3327.7868	4.99962592	96.270
Mongolia	12.25398081	4211.9395	7.88522548	95.200
Mozambique	2.55974876	680.3750	7.39851280	96.616
Nepal	8.36415470	827.7443	6.01148284	89.424
Nicaragua	6.03596862	1913.5213	4.78581617	95.480
Niger	-0.93028726	560.7545	6.64213665	99.480
Nigeria	8.04741088	3200.9531	6.30971866	96.056
Oman	1.02234314	23121.2064	1.29225229	96.450
Pakistan	7.18938403	1173.3925	4.67470798	98.170
Panama	2.62668365	12837.2480	5.06642235	95.581
Paraguay	5.02882767	6629.4170	5.30123859	94.970
Peru	3.41194580	6614.9333	2.38215737	96.790
Philippines	3.59782344	2935.9256	6.34798748	96.400
Poland	0.05382131	14182.1375	3.83695849	91.010
Romania	1.06830988	10031.2673	4.12067496	93.200
Russian Federation	7.82341184	14095.6484	0.73626722	94.840
Rwanda	2.35449053	724.3522	6.16716772	88.124

Sao Tome and Principe	6.99849944	1754.6005	6.54993496	86.360
Saudi Arabia	2.23629032	23543.5663	3.65248567	94.280
Senegal	-1.09025507	1417.0951	6.22407444	92.375
Sierra Leone	4.63931171	702.3354	4.55677237	95.320
Solomon Islands	5.16590238	2235.7473	1.18921747	99.266
Sri Lanka	3.17900228	3971.9187	6.37797890	95.810
Suriname	3.38341273	9199.1779	0.25550303	93.060
Tajikistan	6.10442765	1094.4227	6.70000069	91.773
Tanzania	6.13161433	1012.7669	6.73246187	97.880
Thailand	1.89514182	5822.3837	0.98446886	99.420
Timor-Leste	0.84883821	1221.5343	4.47196212	95.787
Togo	0.19087508	627.7094	5.92058857	97.855
Tonga	2.51087633	4125.4368	2.01870046	98.174
Trinidad and Tobago	5.68441815	20327.9835	3.32294441	97.520
Turkiye	8.85457271	12020.5826	4.93971516	90.120
Uganda	3.07570669	897.5097	5.10630732	97.677
Uruguay	8.87735333	16875.5062	3.23879122	93.450
Uzbekistan	9.28309356	2628.4600	6.87383844	94.910
Vanuatu	0.79886384	2861.2022	3.13686729	98.208
Vietnam	4.08455447	2558.7789	6.42224666	98.740
Zambia	7.80687554	1724.5762	4.69799236	91.867
Zimbabwe	-0.19778481	1407.0343	1.48454262	95.230

	cluster
Afghanistan	5
Algeria	5
Angola	5
Azerbaijan	5
Bahrain	5
Bangladesh	5
Belarus	5
Belize	5
Benin	5
Bhutan	5
Bolivia	5
Brazil	5
Burkina Faso	5
Burundi	5
Cambodia	5
Cameroon	5
Central African Republic	5
Chad	5
Chile	5
China	5
Colombia	5
Costa Rica	5
Dominican Republic	5
Ecuador	5
El Salvador	5
Estonia	5
Fiji	5
Ghana	5
Guatemala	5
Guinea	5
Guinea-Bissau	5

Honduras	5
Hungary	5
India	5
Indonesia	5
Iran, Islamic Rep.	5
Jamaica	5
Kazakhstan	5
Kenya	5
Korea, Rep.	5
Kyrgyz Republic	5
Lao PDR	5
Lebanon	5
Liberia	5
Madagascar	5
Malawi	5
Malaysia	5
Maldives	5
Mali	5
Malta	5
Mauritania	5
Mauritius	5
Mexico	5
Moldova	5
Mongolia	5
Mozambique	5
Nepal	5
Nicaragua	5
Niger	5
Nigeria	5
Oman	5
Pakistan	5
Panama	5
Paraguay	5
Peru	5
Philippines	5
Poland	5
Romania	5
Russian Federation	5
Rwanda	5
Sao Tome and Principe	5
Saudi Arabia	5
Senegal	5
Sierra Leone	5
Solomon Islands	5
Sri Lanka	5
Suriname	5
Tajikistan	5
Tanzania	5
Thailand	5
Timor-Leste	5
Togo	5
Tonga	5
Trinidad and Tobago	5
Turkiye	5

```
Uganda                         5
Uruguay                        5
Uzbekistan                     5
Vanuatu                        5
Vietnam                        5
Zambia                         5
Zimbabwe                       5
```

Análise para 2020

```
lista_indicadores <- c("FP.CPI.TOTL.ZG", # inflação (%)
                       "NY.GDP.PCAP.CD", # Pib per capita (USD)
                       "NY.GDP.MKTP.KD.ZG", # crescimento do PIB anual (%),
                       "SL.UEM.TOTL.ZS" # Desemprego (%)
)



df2020 <- WDI(indicator = lista_indicadores, country = "all", start = 2020, end = 2020
              extra = TRUE)
str(df2020 )
```

```
'data.frame':   266 obs. of  16 variables:
 $ country          : chr  "Afghanistan" "Africa Eastern and Southern" "Africa Western
and Central" "Albania" ...
 $ iso2c            : chr  "AF" "ZH" "ZI" "AL" ...
 $ iso3c            : chr  "AFG" "AFE" "AFW" "ALB" ...
 $ year             : int  2020 2020 2020 2020 2020 2020 2020 2020 2020 2020 ...
 $ status           : chr  "" "" "" "" ...
 $ lastupdated      : chr  "2023-05-10" "2023-05-10" "2023-05-10" "2023-05-10" ...
 $ FP.CPI.TOTL.ZG   : num  NA 6.36 2.44 1.62 2.42 ...
  ..- attr(*, "label")= chr "Inflation, consumer prices (annual %)"
 $ NY.GDP.PCAP.CD   : num  517 1364 1683 5332 3337 ...
  ..- attr(*, "label")= chr "GDP per capita (current US$)"
 $ NY.GDP.MKTP.KD.ZG: num  -2.35 -3.04 -0.9 -3.48 -5.1 ...
  ..- attr(*, "label")= chr "GDP growth (annual %)"
 $ SL.UEM.TOTL.ZS   : num  11.71 7.63 4.91 13.07 12.25 ...
  ..- attr(*, "label")= chr "Unemployment, total (% of total labor force) (modeled ILO
estimate)"
 $ region           : chr  "South Asia" "Aggregates" "Aggregates" "Europe & Central
Asia" ...
 $ capital          : chr  "Kabul" "" "" "Tirane" ...
 $ longitude        : chr  "69.1761" "" "" "19.8172" ...
 $ latitude         : chr  "34.5228" "" "" "41.3317" ...
 $ income           : chr  "Low income" "Aggregates" "Aggregates" "Upper middle
income" ...
 $ lending          : chr  "IDA" "Aggregates" "Aggregates" "IBRD" ...
```

```
dfi2020 <- pre_process_df(df2020)
```

```
dfi2020 |> str()
```

```
'data.frame':   159 obs. of  4 variables:
 $ Inflacao       : num  1.621 2.415 22.272 1.211 0.847 ...
 $ PIB_per_Capita : num  5332 3337 1604 4506 51720 ...
 $ Crescimento_PIB: num  -3.4816 -5.1 -5.6 -7.2 -0.0509 ...
 $ Emprego        : num  86.9 87.8 89.7 87.8 93.5 ...
 - attr(*, "na.action")= 'omit' Named int [1:57] 1 4 5 7 8 10 17 22 28 37 ...
  ..- attr(*, "names")= chr [1:57] "Afghanistan" "American Samoa" "Andorra" "Antigua
and Barbuda" ...
```

```
dfi2020 |> summary()
```

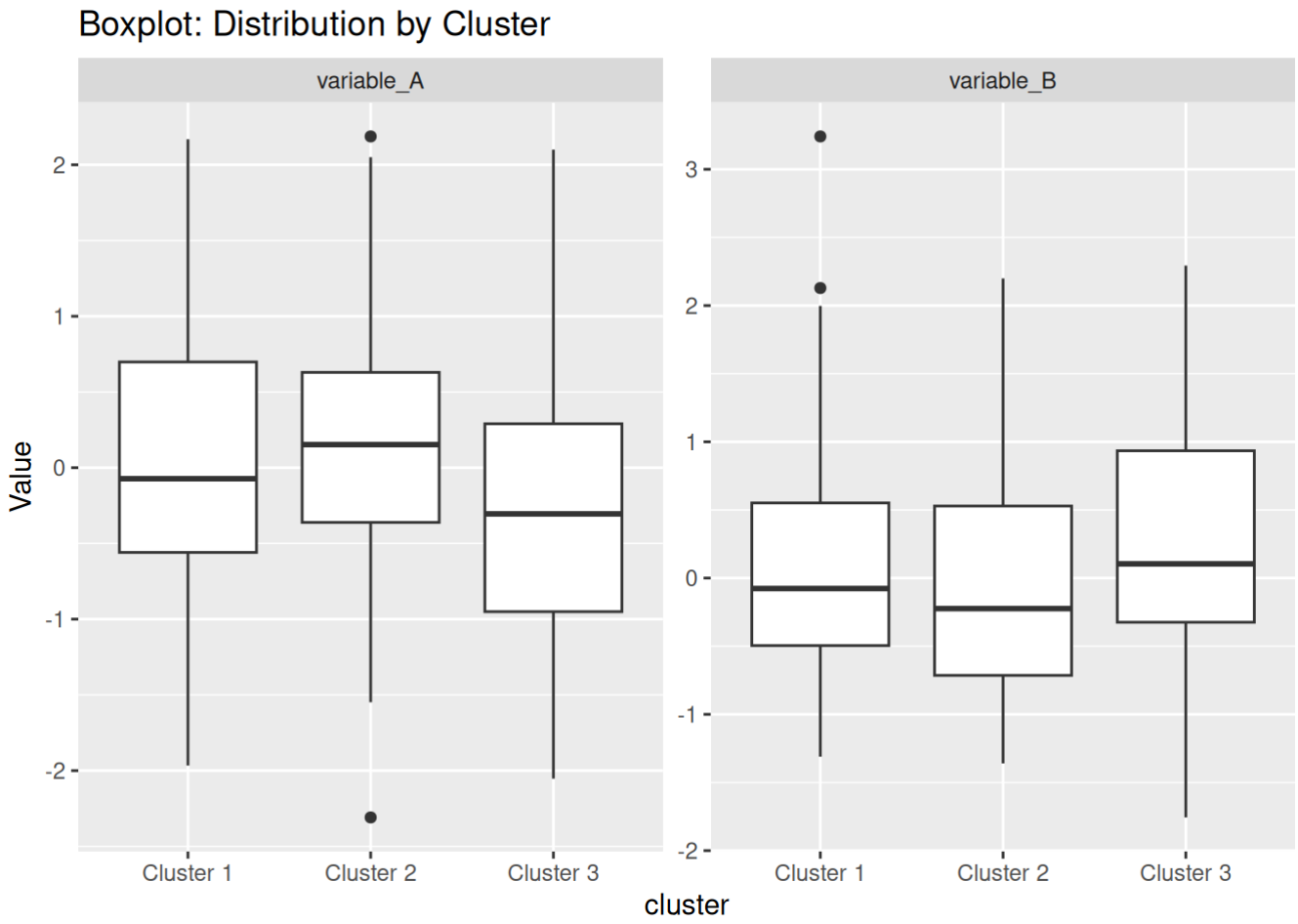
```
    Inflacao        PIB_per_Capita     Crescimento_PIB      Emprego
 Min.   : -2.5952   Min.   :   216.8   Min.   :-54.236   Min.   :71.95
 1st Qu.:  0.4573   1st Qu.:  2227.9   1st Qu.: -7.290   1st Qu.:89.48
 Median :  1.9403   Median :  5353.4   Median : -3.697   Median :93.97
 Mean   :  8.2186   Mean   : 14642.9   Mean   : -4.598   Mean   :91.96
 3rd Qu.:  4.3330   3rd Qu.: 18873.3   3rd Qu.: -1.086   3rd Qu.:95.77
 Max.   :557.2018   Max.   :117370.5   Max.   : 43.480   Max.   :99.86
```

```
dfi2020 |> head()
```

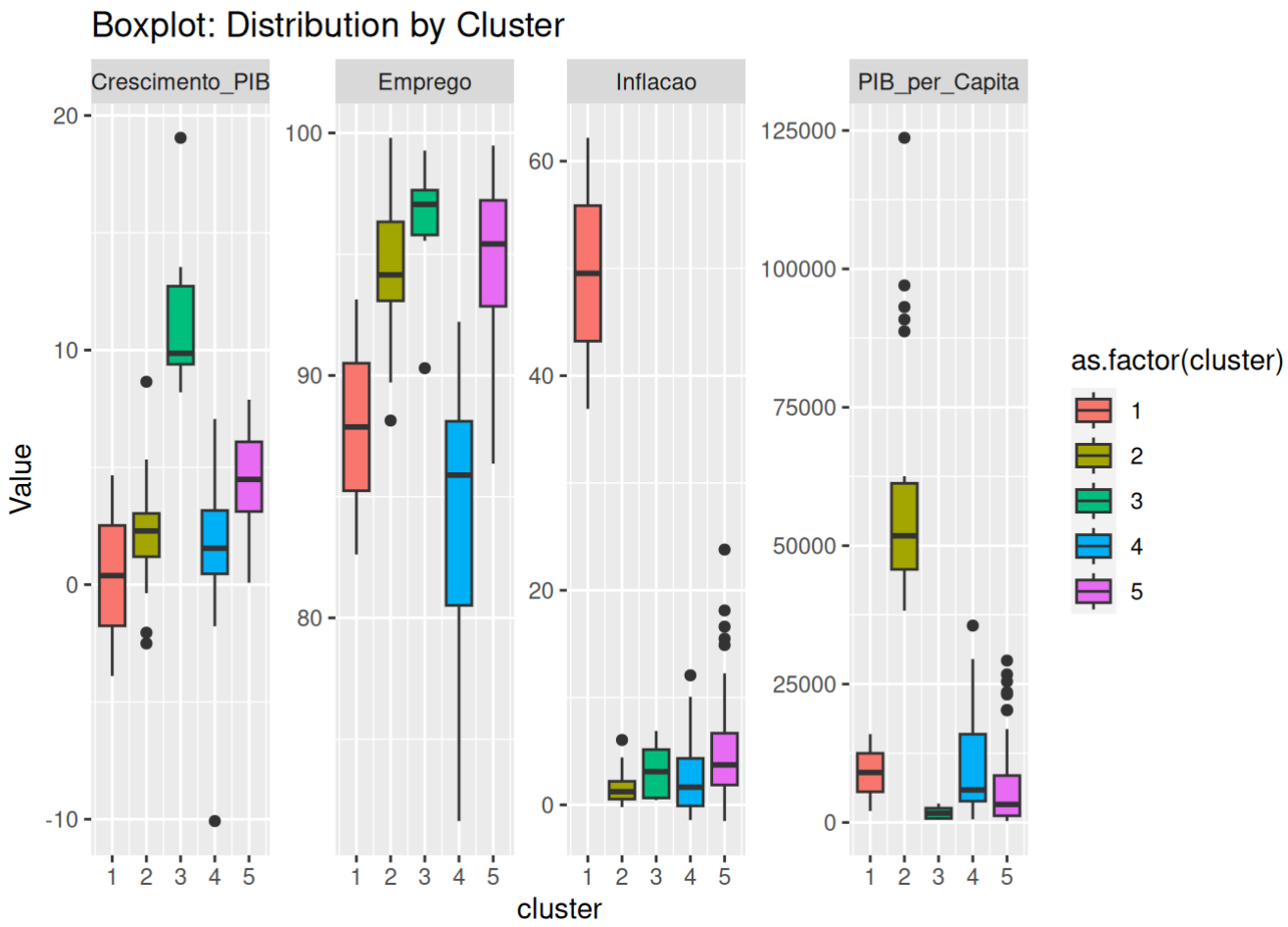
```
           Inflacao PIB_per_Capita Crescimento_PIB Emprego
Albania    1.6208866       5332.160     -3.48163037  86.933
Algeria    2.4151309       3337.253     -5.10000000  87.752
Angola    22.2715643       1603.993     -5.60000000  89.650
Armenia    1.2114358       4505.867     -7.20000000  87.820
Australia  0.8469055      51720.371     -0.05088534  93.540
Austria    1.3819106      48809.227     -6.45396847  94.640
```

```
# Load required libraries
library(ggplot2)
library(dplyr)

# Create a sample dataset
set.seed(123)
cluster <- rep(c("Cluster 1", "Cluster 2", "Cluster 3"), each = 50)
variable_A <- rnorm(150, mean = 0, sd = 1)
variable_B <- rnorm(150, mean = 0, sd = 1)
data <- data.frame(cluster, variable_A, variable_B)

# Reshape the data
data_long <- data %>%
  tidyr::pivot_longer(cols = c(variable_A, variable_B), names_to = "Variable", values_

# Boxplot with facet_wrap
ggplot(data_long, aes(x = cluster, y = Value)) +
  geom_boxplot() +
  ylab("Value") +
```

```
  ggtitle("Boxplot: Distribution by Cluster") +
  facet_wrap(~ Variable, scales = "free_y", nrow = 1)
```

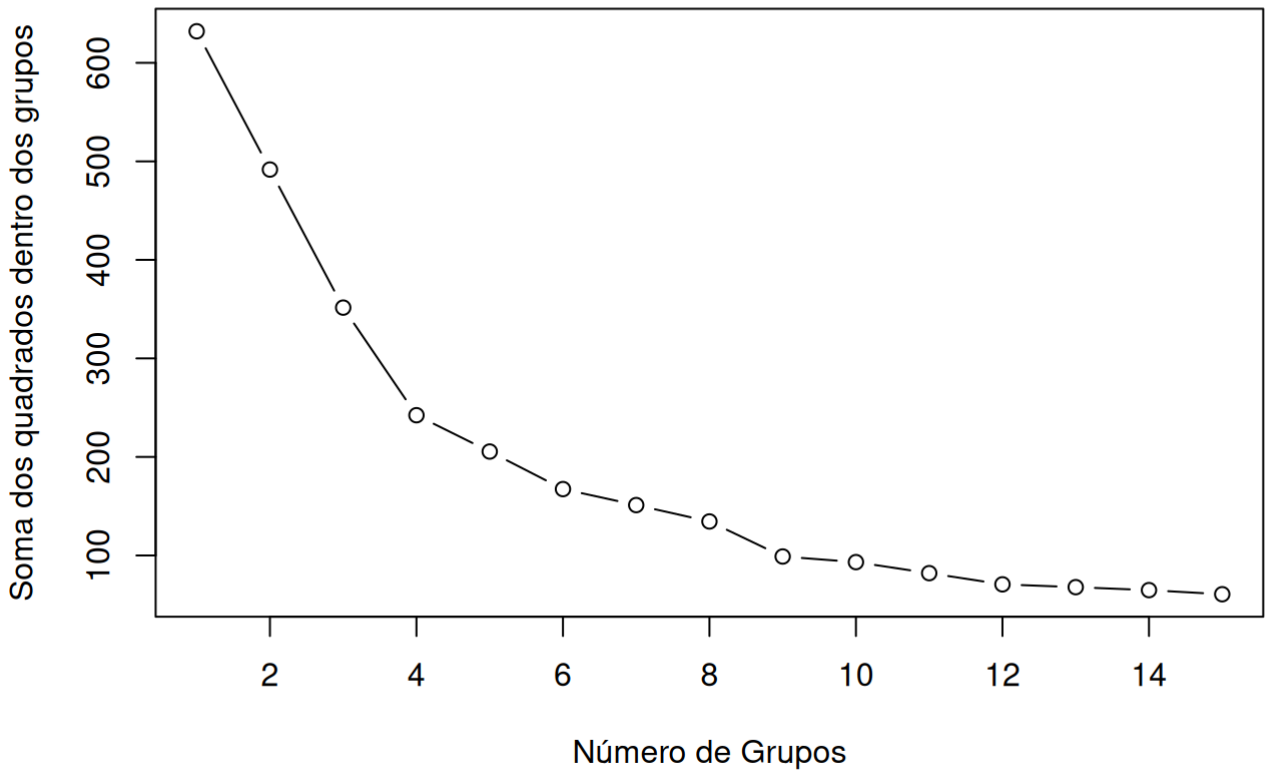



```
dfi2014_com_cluster_long <-  dfi2014_com_cluster |>
   tidyr::pivot_longer(cols = c("Inflacao" ,"PIB_per_Capita","Crescimento_PIB","Empreg

ggplot(dfi2014_com_cluster_long, aes(x = cluster, y = Value,group = cluster)) +
  geom_boxplot(aes(fill=as.factor(cluster))) +
  ylab("Value") +
  ggtitle("Boxplot: Distribution by Cluster") +
  facet_wrap(~ Variable, scales = "free_y", nrow = 1)
```



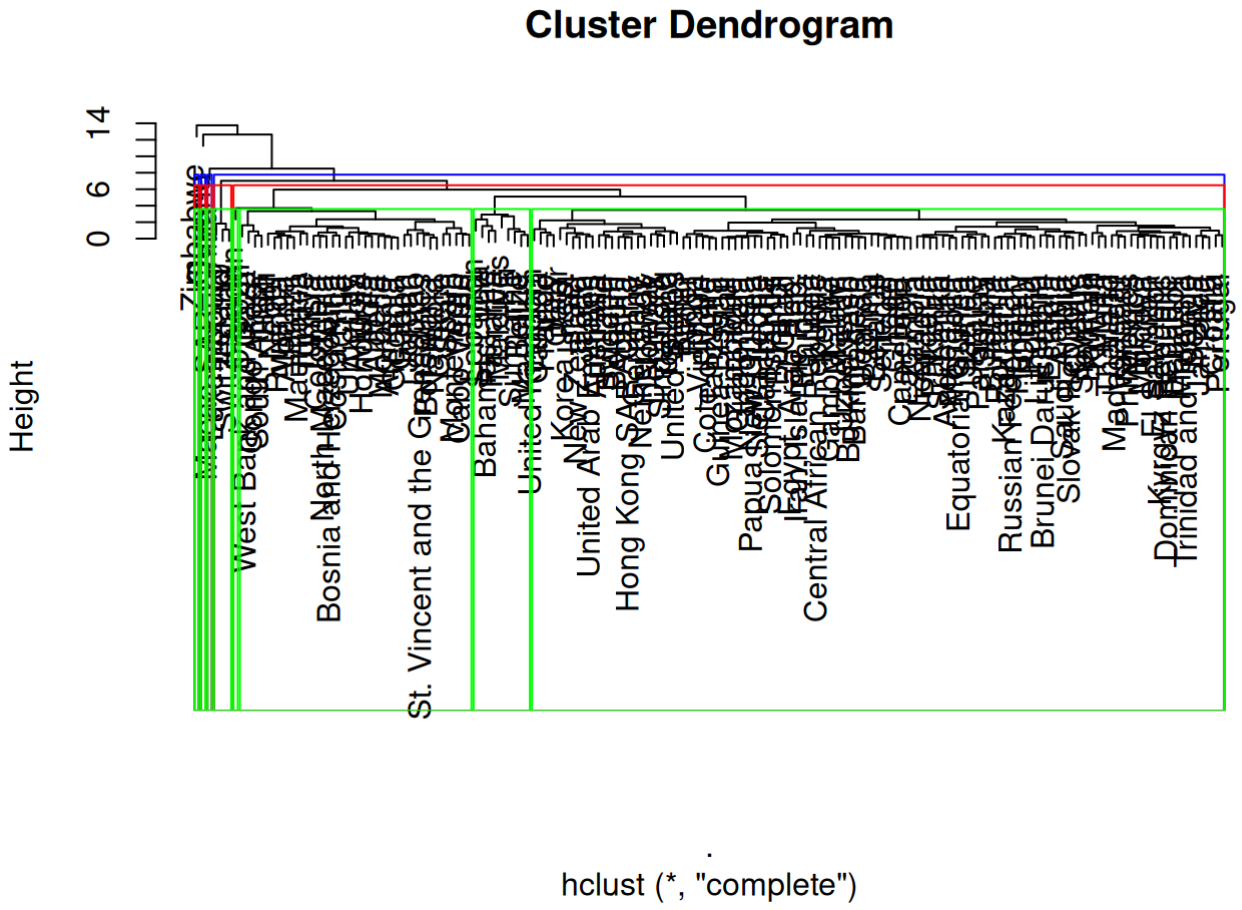
Determinando Quantidade de grupos

```
dfi2020_escala <- scale(dfi2020)



wss <- (nrow(dfi2020_escala)-1)*sum(apply(dfi2020_escala,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(dfi2020_escala,centers=i)$withinss)

plot(1:15, wss, type="b", xlab="Número de Grupos",ylab="Soma dos quadrados dentro dos
```

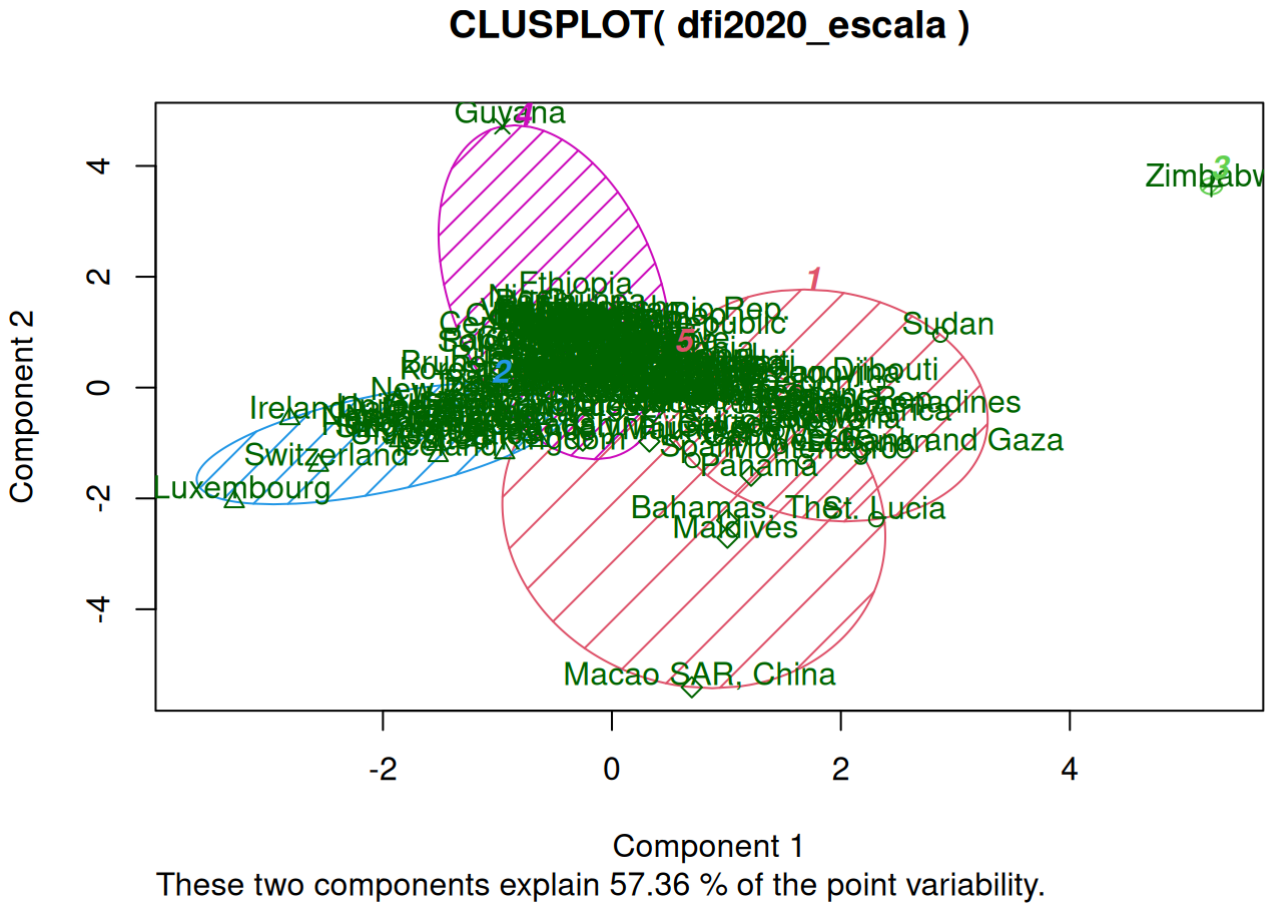


Plotando os clusters

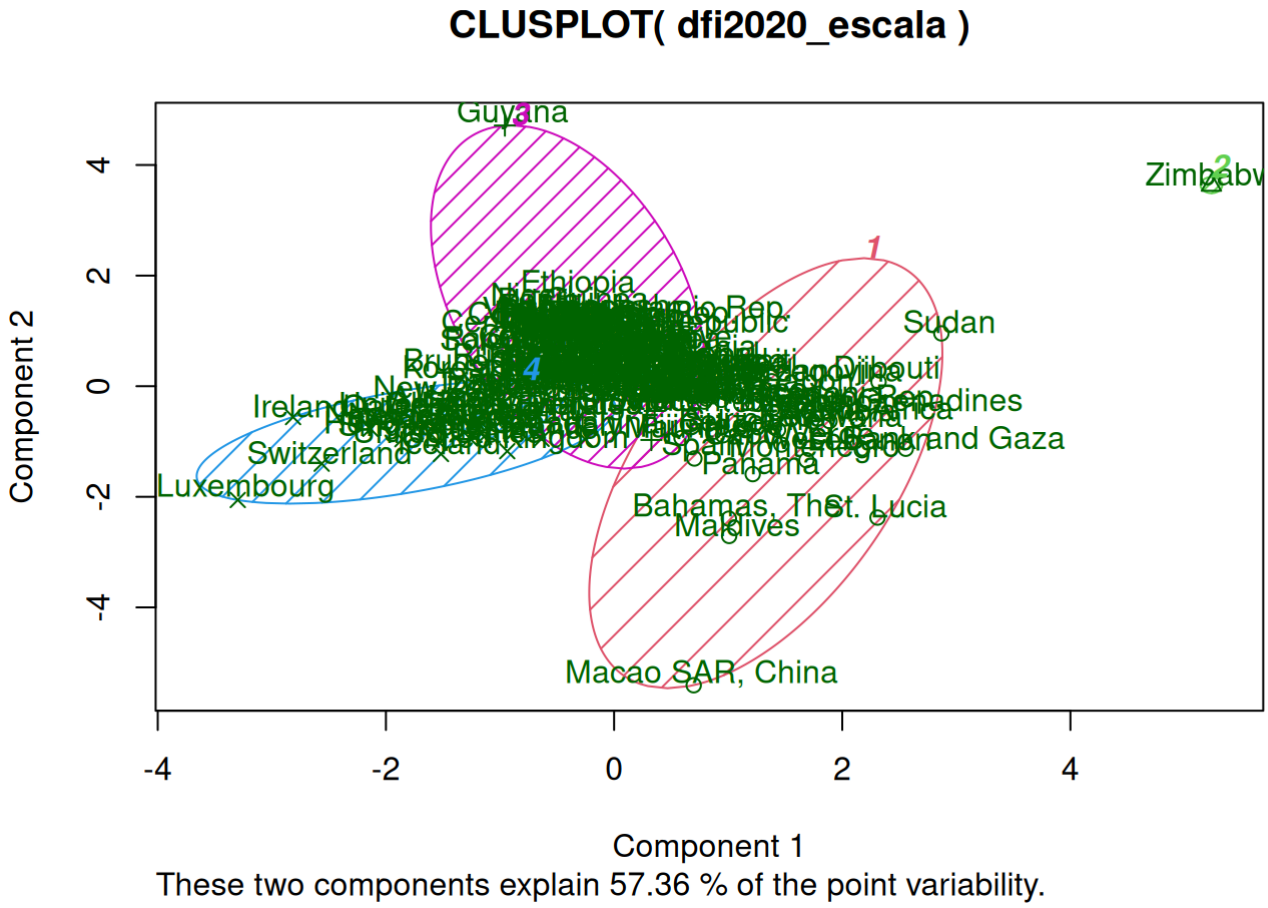
```
dendo <- dfi2020_escala %>% dist() %>% hclust()
plot(dendo)
rect.hclust(dendo, k = 4, border = "blue")
rect.hclust(dendo, k = 5, border = "red")
rect.hclust(dendo, k = 8, border = "green")
```




```r
grupos <- kmeans(dfi2020_escala, centers=5)
clusplot(dfi2020_escala, grupos$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```




```
grupos <- kmeans(dfi2020_escala, centers=4)
clusplot(dfi2020_escala, grupos$cluster, color=TRUE, shade=TRUE, labels=2, lines=0)
```



Plotando os clusters

```
dfi2020_escala[c("Brazil", "Chile", "Colombia", "Norway", "United States"),] %>% dist(
```

```
                Brazil     Chile  Colombia    Norway
Chile        0.6426576
Colombia     0.4520524 0.8039336
Norway       3.4752975 3.0242062 3.7025703
United States 3.0133136 2.6011301 3.1973846 0.7187515
```

Países com MENOR dissimilaridade em relação ao Brasil

```
mat_brasil <- dfi2020_escala %>% dist(diag = TRUE, upper = TRUE) %>% as.matrix()

mat_brasil[, "Brazil"] %>% sort() %>% head(5)
```

```
     Brazil               Albania Bosnia and Herzegovina
  0.0000000             0.1799763             0.2742855
      Nepal                Rwanda
  0.3739877             0.3760059
```

Países com MAIOR dissimilaridade em relação ao Brasil

```
mat_brasil[, "Brazil"] %>% sort() %>% tail(5)
```

```
        Ireland      Luxembourg           Guyana Macao SAR, China
       4.370194        5.661024        5.983287         6.840375
       Zimbabwe
      12.078602
```

Estatística por cluster

```
set.seed(123)
lista_clusteres <- kmeans(dfi2020_escala, centers = 5)$cluster


dfi2020_com_cluster <- dfi2020 |>
  mutate(cluster = lista_clusteres)


stats_cluster <- dfi2020_com_cluster |>
  group_by(cluster) |>
  summarise(
    qtd = n(),
    Media_inflacao = mean(Inflacao, na.rm = TRUE),
    Media_pibpc = mean(PIB_per_Capita, na.rm = TRUE),
    Media_cresciemento = mean(Crescimento_PIB),
    Media_emprego = mean(Emprego, na.rm = TRUE))

stats_cluster
```

```
# A tibble: 5 × 6
  cluster   qtd Media_inflacao Media_pibpc Media_cresciemento Media_emprego
    <int> <int>          <dbl>       <dbl>              <dbl>         <dbl>
1       1     1        557.          1373.             -7.82           92.1
2       2    72          4.31        6356.             -0.625          94.9
3       3    31          7.64        5852.             -6.73           82.6
4       4    25          0.400      54556.             -3.39           94.6
5       5    30          6.41       10796.            -12.8            92.3
```

```
dfi2020_com_cluster["Brazil",]
```

```
       Inflacao PIB_per_Capita Crescimento_PIB Emprego cluster
Brazil 3.211768       6794.489       -3.878676   86.07       3
```

```
dfi2020_com_cluster |> filter(cluster == 1)  # Zimbabwe é um outlier
```

```
         Inflacao PIB_per_Capita Crescimento_PIB Emprego cluster
Zimbabwe 557.2018       1372.697       -7.816951  92.102       1
```

```
dfi2020_com_cluster |> filter(cluster == 4)
```

	Inflacao	PIB_per_Capita	Crescimento_PIB	Emprego	cluster
Australia	0.84690554	51720.37	-0.05088534	93.540	4
Austria	1.38191063	48809.23	-6.45396847	94.640	4
Belgium	0.74079181	45517.79	-5.36138663	94.450	4
Canada	0.71699963	43258.26	-5.23302430	90.540	4
Denmark	0.42071197	60915.42	-1.99460757	94.360	4
Finland	0.29055456	49170.75	-2.20513937	92.240	4
France	0.47649885	39055.28	-7.78458649	91.990	4
Germany	0.14487793	46772.83	-3.69678871	96.140	4
Hong Kong SAR, China	0.25096202	46107.77	-6.54500824	94.190	4
Iceland	2.84792402	59200.18	-6.84229603	94.520	4
Ireland	-0.33458463	85420.19	6.18453804	94.380	4
Israel	-0.58843030	44846.79	-1.85728928	95.670	4
Japan	-0.02499583	39918.17	-4.50690454	97.200	4
Korea, Rep.	0.53728802	31721.30	-0.70941536	96.070	4
Luxembourg	0.81995730	117370.50	-0.79743559	93.230	4
Netherlands	1.27246038	52162.57	-3.88608392	96.180	4
New Zealand	1.71456170	41596.51	-1.25266453	95.400	4
Norway	1.28658491	67329.68	-0.71718267	95.580	4
Qatar	-2.54031503	52315.66	-3.64044980	99.860	4
Singapore	-0.18191667	60729.45	-4.14310562	95.900	4
Sweden	0.49736732	52837.90	-2.17021318	91.710	4
Switzerland	-0.72587493	85656.32	-2.37556328	95.180	4
United Arab Emirates	-2.07940318	37629.17	-4.95705244	95.710	4
United Kingdom	0.98948670	40318.56	-11.03085846	95.528	4
United States	1.23358440	63530.63	-2.76780251	91.950	4

```
dfi2020_com_cluster |> filter(cluster == 3)
```

	Inflacao	PIB_per_Capita	Crescimento_PIB
Albania	1.6208866	5332.1605	-3.481630
Algeria	2.4151309	3337.2525	-5.100000
Armenia	1.2114358	4505.8674	-7.200000
Bosnia and Herzegovina	-1.0512960	6012.0628	-3.119291
Botswana	1.8903592	5863.2032	-8.726409
Brazil	3.2117680	6794.4892	-3.878676
Cabo Verde	0.6057958	2924.1018	-14.783405
Colombia	2.5266350	5307.2152	-7.048151
Congo, Rep.	1.7953715	1838.4481	-6.239320
Costa Rica	0.7249115	12132.8769	-4.050908
Djibouti	1.7774078	2917.9963	1.202022
Gabon	1.3527611	6680.0827	-1.837761
Georgia	5.2024649	4255.7430	-6.760440
Greece	-1.2479836	17658.9473	-9.004044
Haiti	22.7963114	1283.1408	-3.343373
Iraq	0.5741627	4332.3041	-11.324199
Jordan	0.3332944	4042.7693	-1.569473
Lesotho	4.9780968	989.8472	-8.356396
Montenegro	-0.2556557	7677.1522	-15.306894
Namibia	2.2093824	4251.1728	-8.036214
Nepal	5.0523666	1139.1903	-2.369621
North Macedonia	1.2000735	5965.4502	-6.110887
Rwanda	9.8503990	774.6893	-3.358853

South Africa	3.2100360	5741.6431	-6.342471
Spain	-0.3227530	26959.6754	-11.325438
St. Lucia	-1.7558083	8458.1628	-24.364619
St. Vincent and the Grenadines	-0.6281313	8335.2565	-5.312646
Sudan	150.3227239	608.3325	-3.629801
Tunisia	5.6341512	3497.6814	-8.621135
Turkiye	12.2789574	8561.0709	1.940032
West Bank and Gaza	-0.7353320	3233.5686	-11.318466

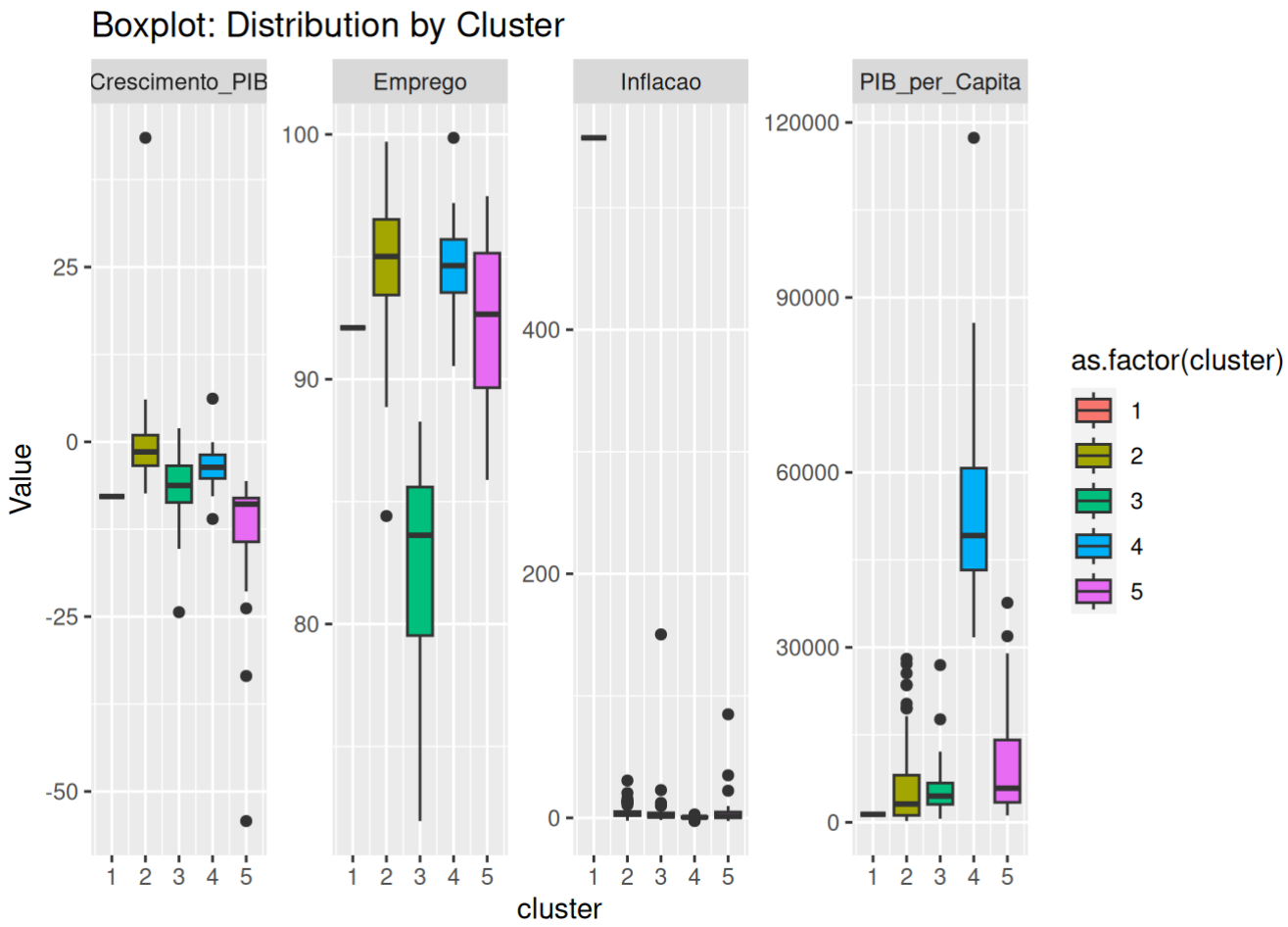
	Emprego	cluster
Albania	86.933	3
Algeria	87.752	3
Armenia	87.820	3
Bosnia and Herzegovina	84.735	3
Botswana	78.980	3
Brazil	86.070	3
Cabo Verde	85.122	3
Colombia	84.960	3
Congo, Rep.	77.483	3
Costa Rica	83.570	3
Djibouti	71.952	3
Gabon	78.274	3
Georgia	88.270	3
Greece	83.690	3
Haiti	84.915	3
Iraq	83.770	3
Jordan	80.790	3
Lesotho	81.539	3
Montenegro	82.120	3
Namibia	78.764	3
Nepal	86.922	3
North Macedonia	83.450	3
Rwanda	86.990	3
South Africa	75.660	3
Spain	84.470	3
St. Lucia	79.610	3
St. Vincent and the Grenadines	79.450	3
Sudan	80.708	3
Tunisia	83.627	3
Turkiye	86.890	3
West Bank and Gaza	74.110	3

```
dfi2020_com_cluster |> filter(cluster == 5)
```

	Inflacao	PIB_per_Capita	Crescimento_PIB	Emprego	cluster
Angola	22.27156431	1603.993	-5.600000	89.650	5
Bahamas, The	0.03852110	23862.711	-23.822608	87.133	5
Belize	0.12143464	5266.876	-13.402959	89.212	5
Bhutan	5.62936523	3009.924	-10.009699	94.970	5
Bolivia	0.94074215	3068.813	-8.737884	92.100	5
Chile	3.04549085	13094.460	-5.978224	88.860	5
Croatia	0.15481137	14198.754	-8.580343	92.490	5
Ecuador	-0.33887239	5645.199	-7.787607	93.890	5
El Salvador	-0.37159021	3903.396	-8.177217	94.980	5

Fiji	-2.59524326	4864.117	-17.000235	95.206	5
Honduras	3.46841178	2354.120	-8.964760	89.320	5
India	6.62343678	1910.421	-6.596081	89.805	5
Italy	-0.13770757	31911.036	-9.039953	90.840	5
Jamaica	5.22677779	4897.266	-10.000000	93.500	5
Kuwait	2.10172955	24300.329	-8.855279	96.669	5
Kyrgyz Republic	6.32542296	1182.522	-8.398364	95.370	5
Lebanon	84.86433305	5599.958	-21.399900	87.029	5
Macao SAR, China	0.81141051	37646.316	-54.235900	97.430	5
Maldives	-1.36977426	7282.358	-33.492796	94.660	5
Malta	0.63854496	28977.566	-8.324283	95.650	5
Mauritius	2.58080077	9007.419	-14.597398	91.370	5
Mexico	3.39683416	8655.001	-7.987912	95.550	5
Morocco	0.70596866	3258.121	-7.187080	88.886	5
Panama	-1.55027541	12569.172	-17.944894	85.886	5
Peru	2.00241206	6056.344	-10.952699	92.820	5
Philippines	2.39316239	3224.423	-9.518295	97.480	5
Portugal	-0.01243833	22242.406	-8.300516	93.200	5
Suriname	34.88978431	4796.533	-15.975196	90.476	5
Trinidad and Tobago	0.59898633	13871.798	-7.678331	95.790	5
Uruguay	9.75640636	15619.543	-6.121476	89.670	5

```
dfi2020_com_cluster <-  dfi2020_com_cluster |>
   tidyr::pivot_longer(cols = c("Inflacao" ,"PIB_per_Capita","Crescimento_PIB","Empreg

ggplot(dfi2020_com_cluster, aes(x = cluster, y = Value,group = cluster)) +
  geom_boxplot(aes(fill=as.factor(cluster))) +
  ylab("Value") +
  ggtitle("Boxplot: Distribution by Cluster") +
  facet_wrap(~ Variable, scales = "free_y", nrow = 1)
```



Seeds Agrupamento hierarquico

```
set.seed(786)
file_loc <- 'seeds.txt'
seeds_df <- read.csv(file_loc,sep = '\t',header = FALSE)


feature_name <-
c('area','perimeter','compactness','length.of.kernel','width.of.kernal
','asymmetry.coefficient','length.of.kernel.groove','type.of.seed')
colnames(seeds_df) <- feature_name
```

```
str(seeds_df)
```

```
'data.frame':   221 obs. of  8 variables:
 $ area                  : num  15.3 14.9 14.3 13.8 16.1 ...
 $ perimeter             : num  14.8 14.6 14.1 13.9 15 ...
 $ compactness           : num  0.871 0.881 0.905 0.895 0.903 ...
 $ length.of.kernel      : num  5.76 5.55 5.29 5.32 5.66 ...
 $ width.of.kernal
     : num  3.31 3.33 3.34 3.38 3.56 ...
 $ asymmetry.coefficient  : num  2.22 1.02 2.7 2.26 1.35 ...
 $ length.of.kernel.groove: num  5.22 4.96 4.83 4.8 5.17 ...
 $ type.of.seed          : num  1 1 1 1 1 1 1 5 NA 1 ...
```

```
summary(seeds_df)
```

```
     area         perimeter      compactness     length.of.kernel
 Min.   : 1.00   Min.   : 1.00   Min.   :0.8081   Min.   :0.8189
 1st Qu.:12.11   1st Qu.:13.43   1st Qu.:0.8577   1st Qu.:5.2447
 Median :14.13   Median :14.29   Median :0.8735   Median :5.5180
 Mean   :14.29   Mean   :14.43   Mean   :0.8713   Mean   :5.5639
 3rd Qu.:17.09   3rd Qu.:15.69   3rd Qu.:0.8877   3rd Qu.:5.9798
 Max.   :21.18   Max.   :17.25   Max.   :0.9183   Max.   :6.6750
 NA's   :1       NA's   :9       NA's   :14       NA's   :11
 width.of.kernal\n asymmetry.coefficient length.of.kernel.groove
 Min.   :2.630     Min.   :0.7651        Min.   :3.485
 1st Qu.:2.956     1st Qu.:2.6002        1st Qu.:5.045
 Median :3.245     Median :3.5990        Median :5.226
 Mean   :3.281     Mean   :3.6935        Mean   :5.408
 3rd Qu.:3.566     3rd Qu.:4.7687        3rd Qu.:5.879
 Max.   :5.325     Max.   :8.4560        Max.   :6.735
 NA's   :12        NA's   :11            NA's   :15
  type.of.seed
 Min.   :1.000
 1st Qu.:1.000
 Median :2.000
 Mean   :2.084
 3rd Qu.:3.000
 Max.   :5.439
 NA's   :15
```

```
any(is.na(seeds_df))
```

```
[1] TRUE
```

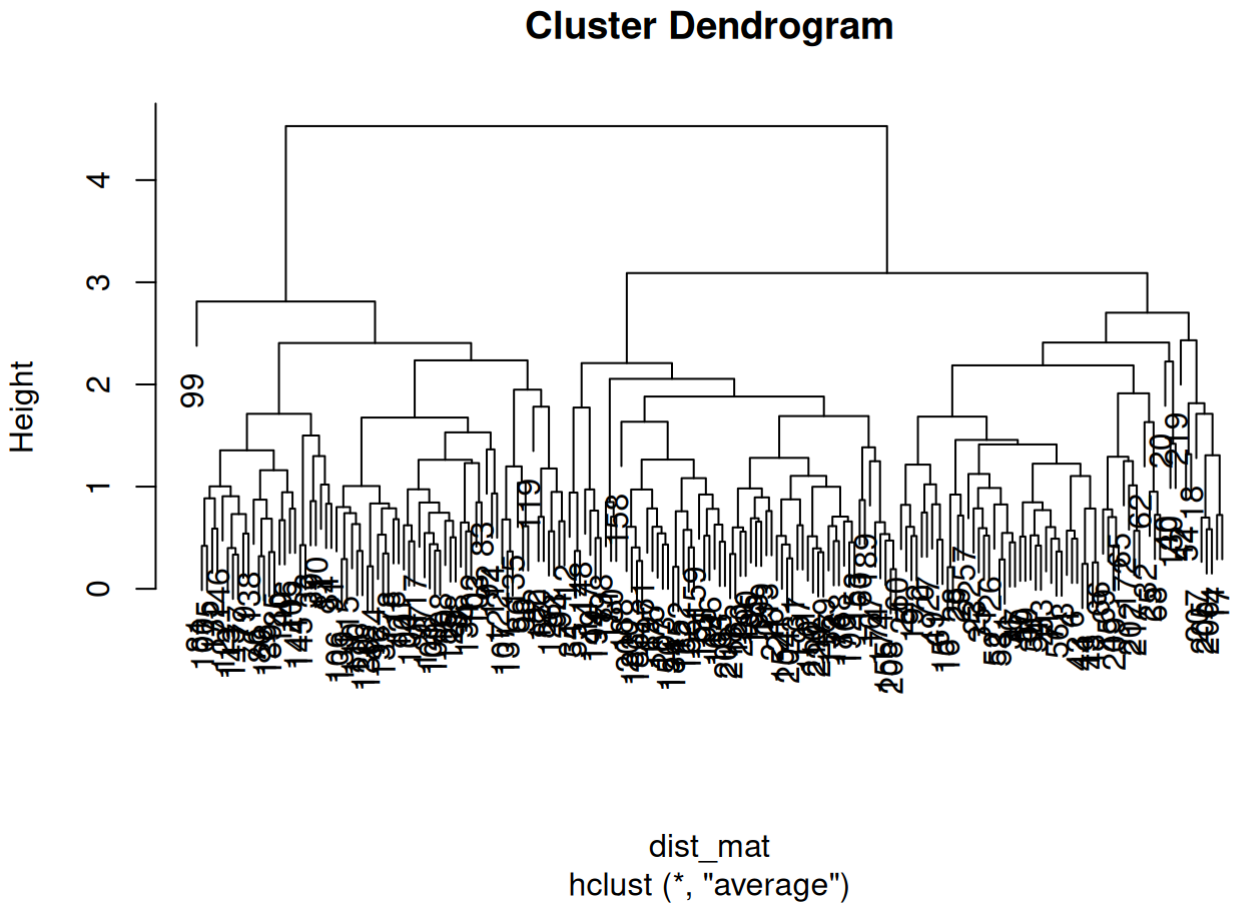
```
seeds_df <- na.omit(seeds_df)

seeds_label <- seeds_df$type.of.seed
seeds_df$type.of.seed <- NULL
str(seeds_df)
```

```
'data.frame':   199 obs. of  7 variables:
 $ area                   : num  15.3 14.9 14.3 13.8 16.1 ...
 $ perimeter              : num  14.8 14.6 14.1 13.9 15 ...
 $ compactness            : num  0.871 0.881 0.905 0.895 0.903 ...
 $ length.of.kernel       : num  5.76 5.55 5.29 5.32 5.66 ...
 $ width.of.kernal
      : num  3.31 3.33 3.34 3.38 3.56 ...
 $ asymmetry.coefficient  : num  2.22 1.02 2.7 2.26 1.35 ...
 $ length.of.kernel.groove: num  5.22 4.96 4.83 4.8 5.17 ...
 - attr(*, "na.action")= 'omit' Named int [1:22] 8 9 37 38 63 64 72 73 111 112 ...
  ..- attr(*, "names")= chr [1:22] "8" "9" "37" "38" ...
```

```
seeds_df_sc <- as.data.frame(scale(seeds_df))
summary(seeds_df_sc)
```

```
     area            perimeter        compactness      length.of.kernel
Min.   :-1.4825   Min.   :-1.6680   Min.   :-2.6891   Min.   :-1.6776
1st Qu.:-0.8866   1st Qu.:-0.8591   1st Qu.:-0.5879   1st Qu.:-0.8480
Median :-0.1674   Median :-0.1723   Median : 0.1110   Median :-0.2303
Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000
3rd Qu.: 0.8686   3rd Qu.: 0.9227   3rd Qu.: 0.6857   3rd Qu.: 0.8090
Max.   : 2.1443   Max.   : 2.0254   Max.   : 2.0364   Max.   : 2.3261
width.of.kernal\n  asymmetry.coefficient length.of.kernel.groove
Min.   :-1.67987   Min.   :-1.99450      Min.   :-1.8300
1st Qu.:-0.82214   1st Qu.:-0.76760      1st Qu.:-0.7604
Median :-0.05427   Median :-0.04637      Median :-0.3910
Mean   : 0.00000   Mean   : 0.00000      Mean   : 0.0000
3rd Qu.: 0.79025   3rd Qu.: 0.74759      3rd Qu.: 0.9302
Max.   : 2.02861   Max.   : 3.13764      Max.   : 2.2921
```

```r
dist_mat <- dist(seeds_df_sc, method = 'euclidean')
hclust_avg <- hclust(dist_mat, method = 'average')
plot(hclust_avg)
```

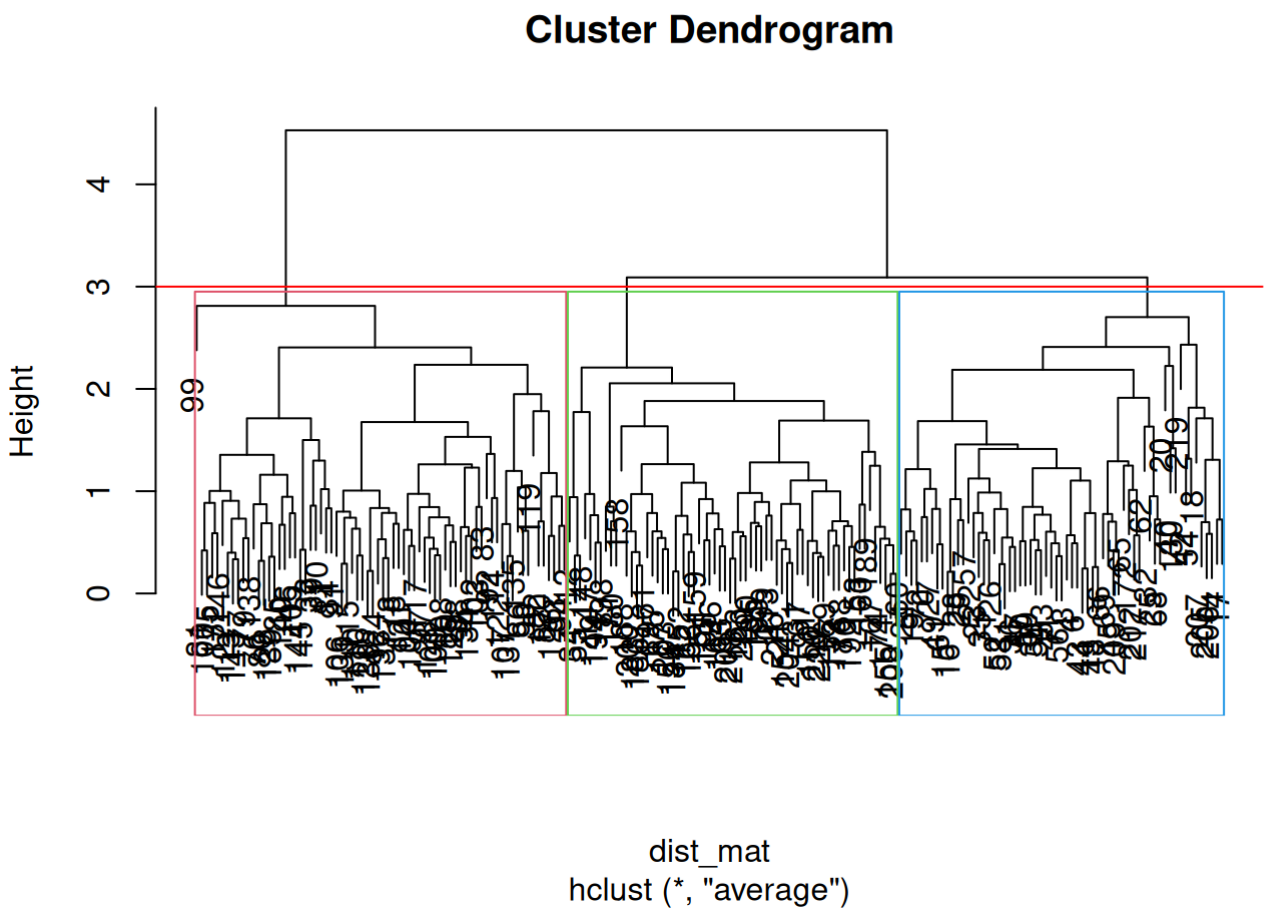



```r
cut_avg <- cutree(hclust_avg, k = 3)
```


```r
cut_avg <- cutree(hclust_avg, k = 3)
```


```r
plot(hclust_avg)
rect.hclust(hclust_avg , k = 3, border = 2:6)
```

```
abline(h = 3, col = 'red')
```

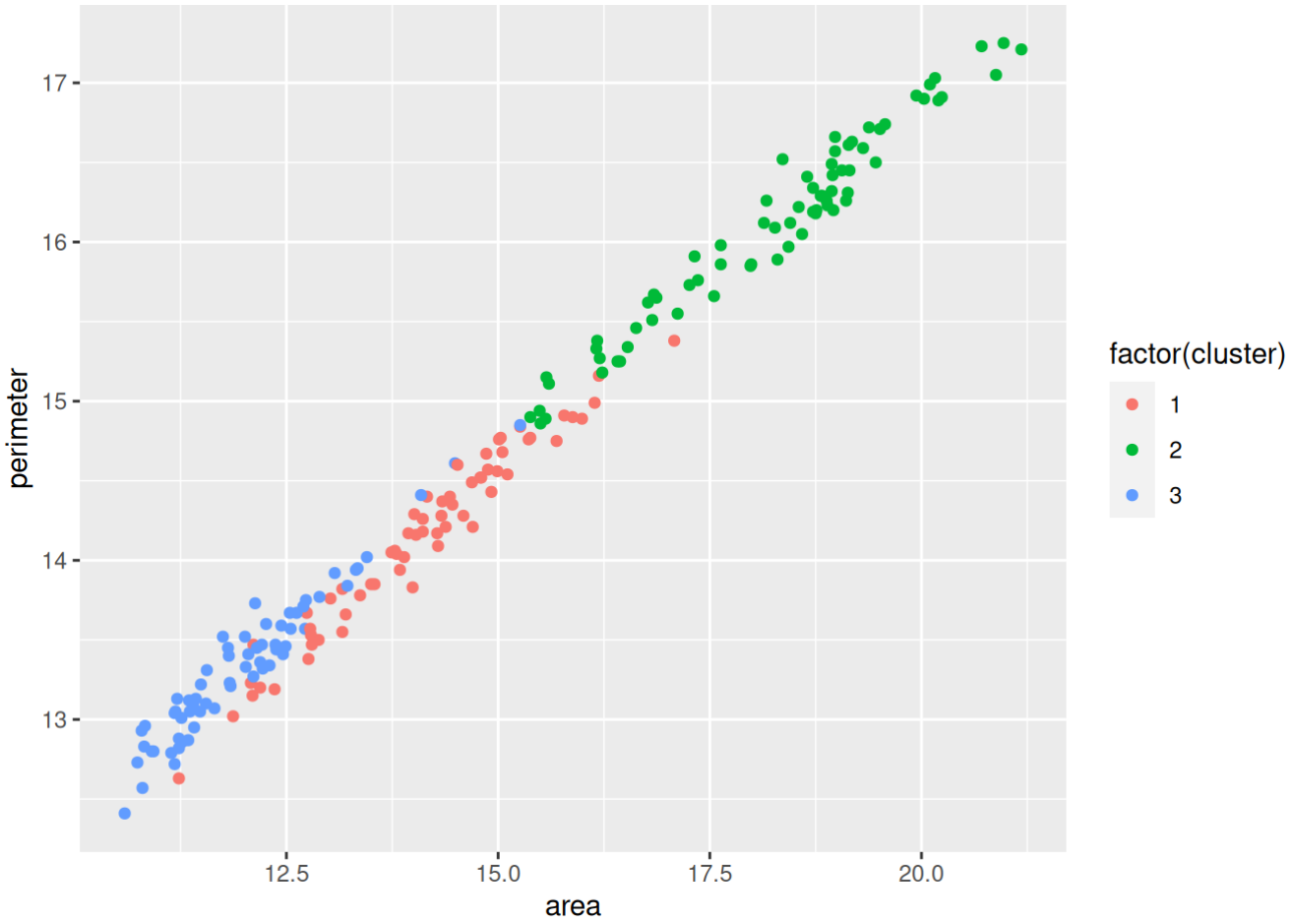



```
#install.packages("dendextend")
suppressPackageStartupMessages(library(dendextend))
avg_dend_obj <- as.dendrogram(hclust_avg)
avg_col_dend <- color_branches(avg_dend_obj, h = 3)
```

```
suppressPackageStartupMessages(library(dplyr))
seeds_df_cl <- mutate(seeds_df, cluster = cut_avg)
count(seeds_df_cl,cluster)
```

```
  cluster  n
1       1 63
2       2 72
3       3 64
```

```
suppressPackageStartupMessages(library(ggplot2))
ggplot(seeds_df_cl, aes(x=area, y = perimeter, color =
factor(cluster))) + geom_point()
```



Conclusão

Neste laboratório, pudemos compreender a importância da clusterização como um processo para identificar similaridades nos dados. Economistas e cientistas sociais podem se beneficiar da clusterização para compreender as verdadeiras relações entre os países, sem depender de classificações tradicionais baseadas em demarcações territoriais, regionais ou políticas. Da mesma forma, um gestor de ações pode agrupar seus ativos de forma não óbvia, identificando correlações “escondidas” e evitando a subjetividade das demarcações setoriais fornecidas por provedores como a Bolsa, Bloomberg, S&P, CVM, etc. Gerentes de marketing podem obter uma compreensão mais profunda de seus consumidores, identificando grupos nos quais uma campanha de marketing específica pode ser eficaz e conhecendo os concorrentes que possuem produtos semelhantes. Agricultores e biólogos podem compreender como diferentes grupos de sementes se comportam em diferentes condições climáticas e ambientais.

Ao aplicarmos o algoritmo K-means no dataset do Banco Mundial, pudemos capturar as reações dos países a diferentes dinâmicas econômicas. Observamos o seguinte:

- Em 2014, o K-means foi capaz de identificar países em situação extrema, como aqueles afetados por guerras, com baixo PIB per capita e alta inflação, agrupando-os nos clusters 3 e 1. Também identificamos países ricos/desenvolvidos que foram alocados no cluster 2 (EUA, Dinamarca, Canadá, etc.). Países subdesenvolvidos e emergentes foram agrupados no cluster 5.

- Em 2020, os clusters apresentaram uma distribuição um pouco mais uniforme (hipótese: “a pandemia pode ter aumentado a similaridade entre alguns países”). As regiões mais desenvolvidas (grupo 4) ainda não haviam sofrido os efeitos inflacionários da pandemia e experimentaram uma desaceleração menor. No entanto, é importante destacar que a principal distinção entre essas regiões foi o alto PIB per capita, um panorama histórico e anterior à pandemia. Independentemente do cluster, foi possível observar uma desaceleração generalizada entre os países devido à pandemia, ao contrário do que ocorreu em 2014.

- O Brasil, apesar de ter registrado uma queda no PIB abaixo da média, foi classificado no grupo de países emergentes/subdesenvolvidos. Em ambas as análises, foi possível identificar que o Brasil ficou agrupado com países emergentes/subdesenvolvidos, caracterizados por baixo crescimento e baixo PIB per capita.

Esta análise certamente será aprimorada com a inclusão de novas variáveis que nos permitam distinguir melhor os países. Algumas dessas variáveis podem ser o índice de alfabetização, notas no PISA (Programme for International Student Assessment), produtividade econômica, níveis de fome, índices de criminalidade, qualidade da saúde pública, índice de corrupção e uma variável indicadora para identificar se o país está em guerra, entre outras.

A adição dessas variáveis proporcionará uma visão mais abrangente e precisa da situação econômica, social e política de cada país, permitindo uma análise mais completa e aprofundada. Com um conjunto de dados mais abrangente, será possível identificar melhor os padrões, tendências e correlações entre as diferentes variáveis, o que contribuirá para uma compreensão mais precisa dos clusters e das relações entre os países.

Portanto, ao incorporar essas novas variáveis, será possível enriquecer a análise e obter insights mais valiosos sobre as semelhanças e diferenças entre os países, possibilitando uma tomada de decisão mais informada em várias áreas, como economia, política, educação, saúde e segurança.