Provinces of Argentina

Shreya Rao

With almost 40 million inhabitants and a diverse geography that encompasses the Andes mountains, glacial lakes, and the Pampas grasslands, Argentina is the second largest country (by area) and has one of the largest economies in South America. It is politically organized as a federation of 23 provinces and an autonomous city, Buenos Aires.

Here I will analyze ten economic and social indicators collected for each province. Because these indicators are highly correlated, I will use principal component analysis (PCA) to reduce redundancies and highlight patterns that are not apparent in the raw data. After visualizing the patterns, I will use k-means clustering to partition the provinces into groups with similar development levels.

These results can be used to plan public policy by helping allocate resources to develop infrastructure, education, and welfare programs.

```
# Load the tidyverse
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.5
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(ggrepel)
# Read in the dataset
argentina <- read.table("C:/Users/Shreya/Documents/Projects/argentina.txt", sep = ",", header = T)
# Inspect the first rows of the dataset
nrow(argentina)
```

```
## [1] 22
```

```
head(argentina)
```

```
##       province       gdp illiteracy   poverty deficient_infra school_dropout
## 1 Buenos Aires 292689868    1.38324  8.167798        5.511856      0.7661682
## 2    Catamarca   6150949    2.34414  9.234095       10.464484      0.9519631
## 3     CÃ³rdoba  69363739    2.71414  5.382380       10.436086      1.0350558
## 4   Corrientes   7968013    5.60242 12.747191       17.438858      3.8642652
```

```
## 5        Chaco   9832643    7.51758 15.862619       31.479527      2.5774621
## 6       Chubut  17747854    1.54806  8.051752        8.044618      0.5863094
##   no_healthcare birth_mortal      pop movie_theatres_per_cap doctors_per_cap
## 1       48.7947          4.4 15625084           6.015968e-06     0.004835622
## 2       45.0456          1.5   367828           5.437324e-06     0.004502104
## 3       45.7640          4.8  3308876           1.118204e-05     0.010175359
## 4       62.1103          5.9   992595           4.029841e-06     0.004495288
## 5       65.5104          7.5  1055259           2.842904e-06     0.003604802
## 6       39.5473          3.0   509108           1.571376e-05     0.004498063
```

Argentina ranks third in South America in total population, but the population is unevenly distributed throughout the country. 60% of the population resides in the Pampa region (Buenos Aires, La Pampa, Santa Fe, Entre Ríos and Córdoba) which only encompasses about 20% of the land area.

GDP is a measure of the size of a province's economy. To measure how rich or poor the inhabitants are, economists use per capita GDP, which is GDP divided by the province's population.

```
# Add gdp_per_capita column to argentina
argentina <- argentina %>%
  mutate(gdp_per_cap = gdp / pop)
# Find the four richest provinces
( rich_provinces  <- argentina %>%
    arrange(desc(gdp_per_cap)) %>%
     select(province, gdp_per_cap) %>%
     top_n(4) )
```

```
## Selecting by gdp_per_cap
```

```
##     province gdp_per_cap
## 1 Santa Cruz    42.57398
## 2   NeuquÃ©n    40.93143
## 3     Chubut    34.86069
## 4   San Luis    27.25093
```

```
# Find the provinces with populations over 1 million
( bigger_pops <- argentina %>%
     arrange(desc(pop)) %>%
     select(province, pop) %>%
     filter(pop > 1000000) )
```

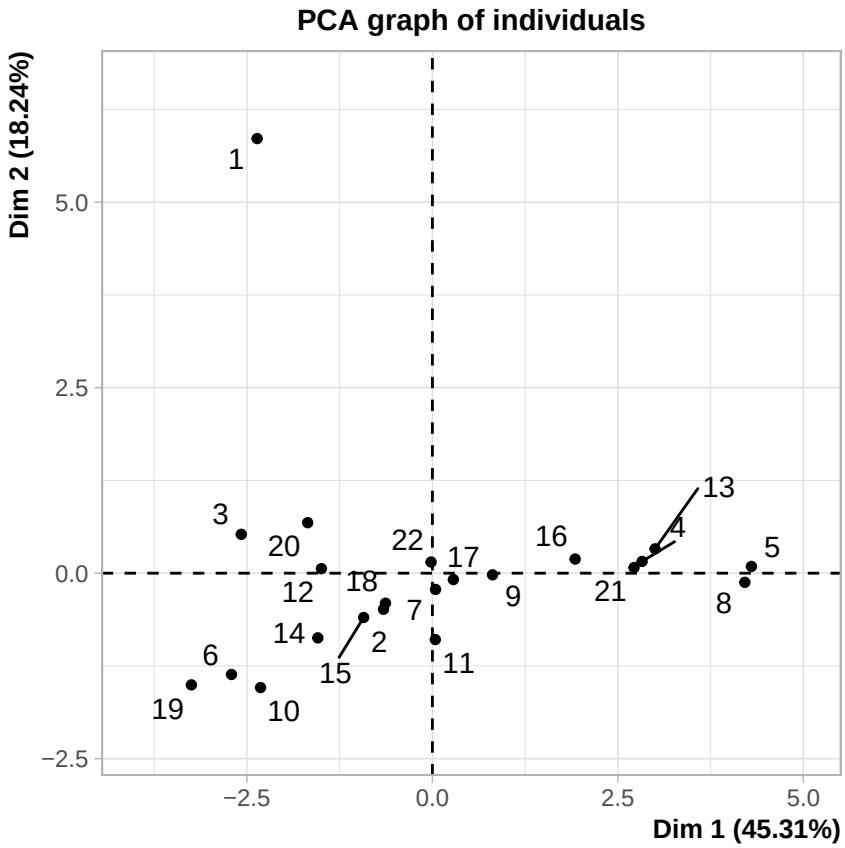
```
##       province      pop
## 1 Buenos Aires 15625084
## 2     CÃ³rdoba  3308876
## 3     Santa Fe  3194537
## 4      Mendoza  1738929
## 5     TucumÃ¡n  1448188
## 6  Entre RÃos  1235994
## 7        Salta  1214441
## 8     Misiones  1101593
## 9        Chaco  1055259
```

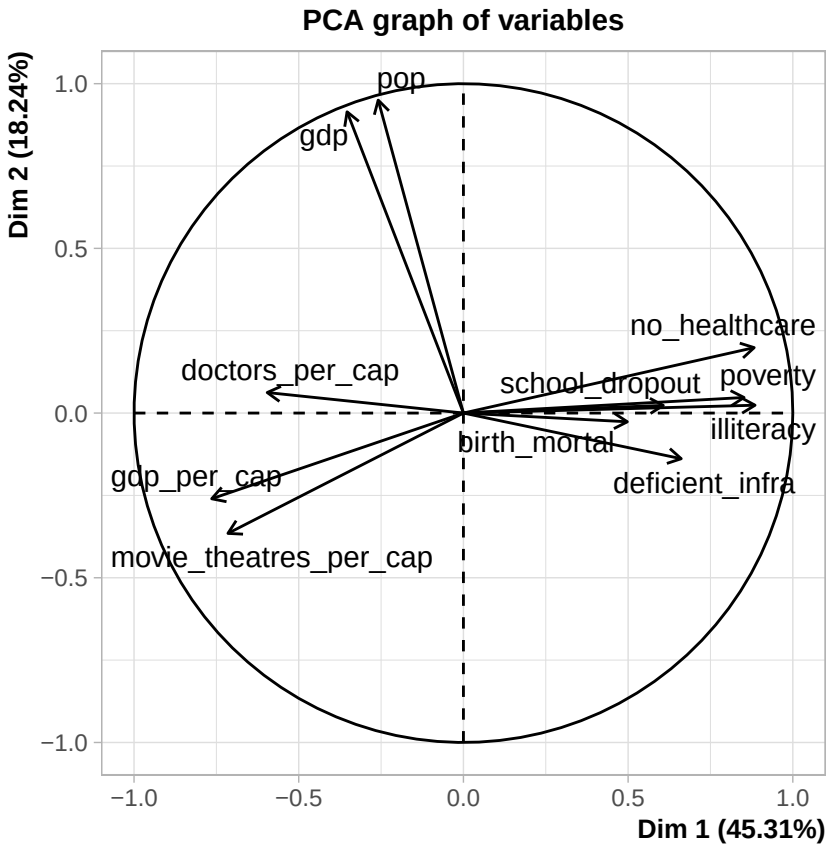
PCA

```r
# Select numeric columns and cast to matrix
argentina_matrix  <- argentina  %>%
  select_if(is.numeric) %>%
  as.matrix()
# Print the first lines of the result
head(argentina_matrix)
```

```
##            gdp illiteracy   poverty deficient_infra school_dropout
## [1,] 292689868    1.38324  8.167798        5.511856      0.7661682
## [2,]   6150949    2.34414  9.234095       10.464484      0.9519631
## [3,]  69363739    2.71414  5.382380       10.436086      1.0350558
## [4,]   7968013    5.60242 12.747191       17.438858      3.8642652
## [5,]   9832643    7.51758 15.862619       31.479527      2.5774621
## [6,]  17747854    1.54806  8.051752        8.044618      0.5863094
##      no_healthcare birth_mortal      pop movie_theatres_per_cap doctors_per_cap
## [1,]       48.7947          4.4 15625084           6.015968e-06     0.004835622
## [2,]       45.0456          1.5   367828           5.437324e-06     0.004502104
## [3,]       45.7640          4.8  3308876           1.118204e-05     0.010175359
## [4,]       62.1103          5.9   992595           4.029841e-06     0.004495288
## [5,]       65.5104          7.5  1055259           2.842904e-06     0.003604802
## [6,]       39.5473          3.0   509108           1.571376e-05     0.004498063
##      gdp_per_cap
## [1,]   18.732051
## [2,]   16.722352
## [3,]   20.962931
## [4,]    8.027456
## [5,]    9.317753
## [6,]   34.860686
```

```r
# Load FactoMineR
library(FactoMineR)
# Apply PCA and print results
( argentina_pca  <- PCA(argentina_matrix, scale.unit = TRUE) )
```


PCA graph of individuals
Dim 2 (18.24%)
Dim 1 (45.31%)
5.0
2.5
0.0
−2.5
−2.5
0.0
2.5
5.0
1
3
20
22
17
16
13
4
5
12
18
9
21
8
7
14
2
11
15
6
19
10

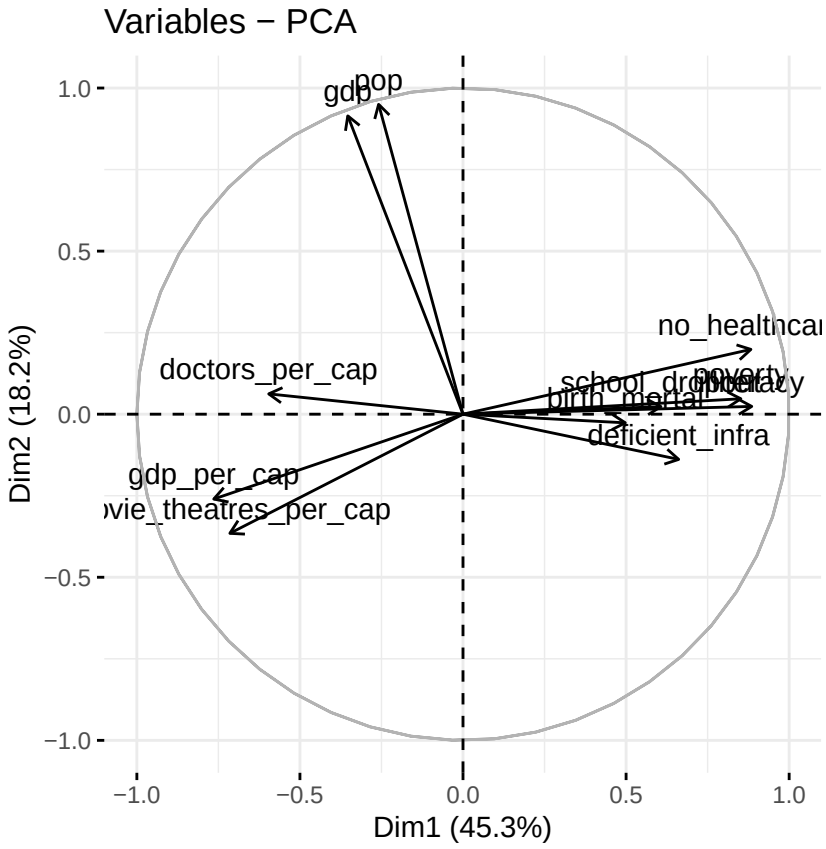



```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 22 individuals, described by 11 variables
## *The results are available in the following objects:
##
##    name               description
## 1  "$eig"             "eigenvalues"
## 2  "$var"             "results for the variables"
## 3  "$var$coord"       "coord. for the variables"
## 4  "$var$cor"         "correlations variables - dimensions"
## 5  "$var$cos2"        "cos2 for the variables"
## 6  "$var$contrib"     "contributions of the variables"
## 7  "$ind"             "results for the individuals"
## 8  "$ind$coord"       "coord. for the individuals"
## 9  "$ind$cos2"        "cos2 for the individuals"
## 10 "$ind$contrib"     "contributions of the individuals"
## 11 "$call"            "summary statistics"
## 12 "$call$centre"     "mean of the variables"
## 13 "$call$ecart.type" "standard error of the variables"
## 14 "$call$row.w"      "weights for the individuals"
## 15 "$call$col.w"      "weights for the variables"
```

```r
# Load factoextra
library(factoextra)
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
# Set the size of plots in this notebook
options(repr.plot.width=7, repr.plot.height=5)
# Plot the original variables and the first 2 components and print the plot object.
( pca_var_plot <- fviz_pca_var(argentina_pca) )
```

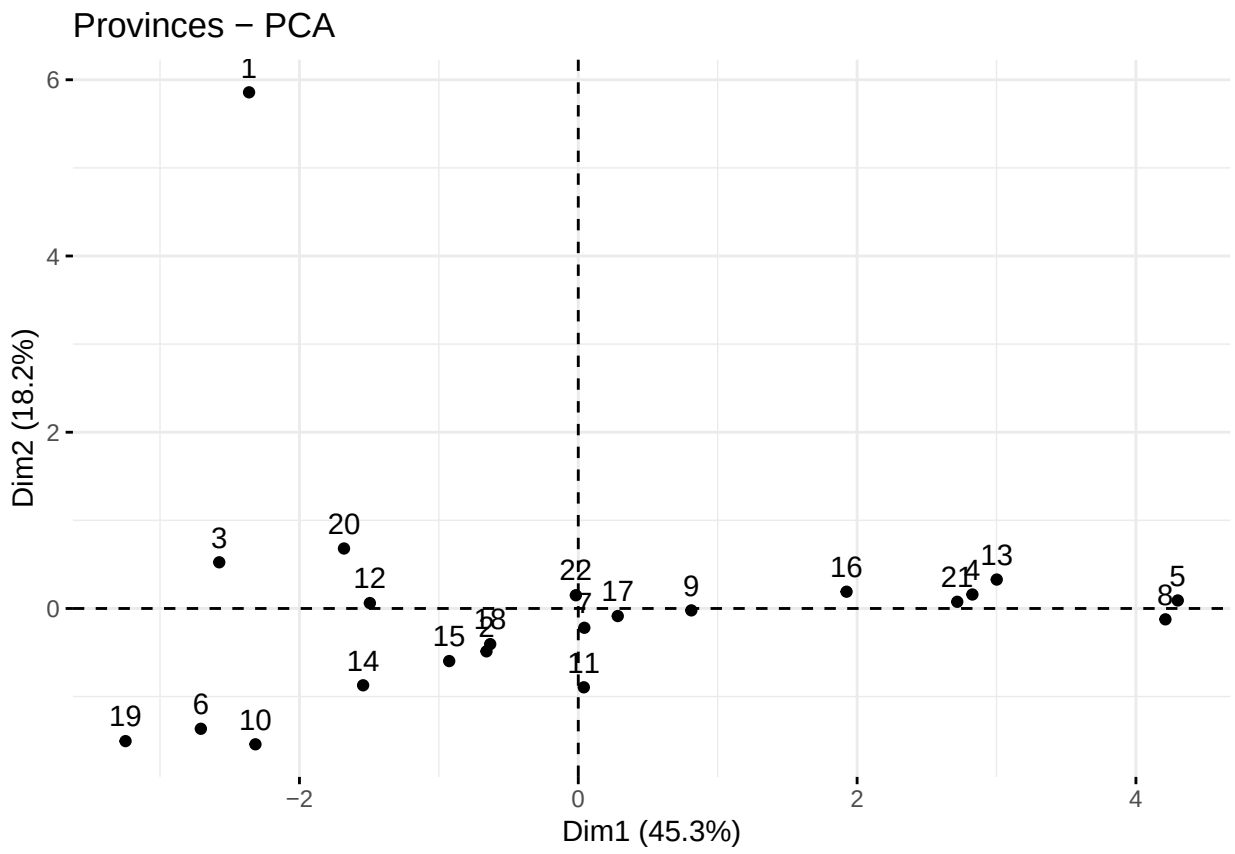



```
# Sum the variance preserved by the first two components. Print the result.
( variance_first_two_pca <- argentina_pca$eig[1, 2] + argentina_pca$eig[2, 2] )
```

```
## [1] 63.54897
```

Visualizing The Components:

```
# Visualize Dim2 vs. Dim1
fviz_pca_ind(argentina_pca, title = "Provinces - PCA")
```



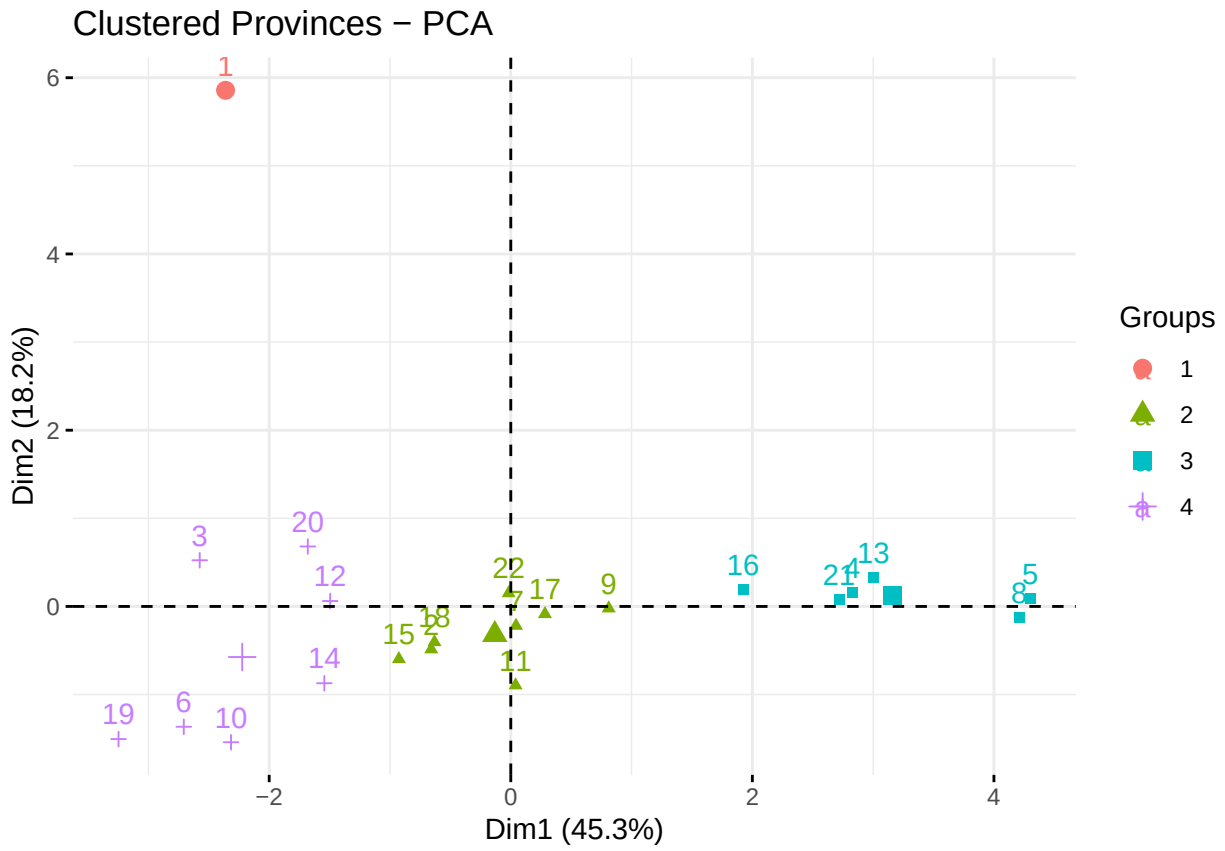
K-Means Clustering:

```
# Create an intermediate data frame with pca_1 and pca_2
argentina_comps <- tibble(pca_1 = argentina_pca$ind$coord[ ,1],
                          pca_2 = argentina_pca$ind$coord[ ,2])
# Cluster the observations using the first 2 components and print its contents
( argentina_km <- kmeans(argentina_comps, centers = 4, nstart = 20, iter.max = 50) )
```

```
## K-means clustering with 4 clusters of sizes 1, 8, 6, 7
##
## Cluster means:
##        pca_1      pca_2
## 1 -2.3614699  5.8572297
## 2 -0.1320515 -0.3199319
## 3  3.1637648  0.1200775
## 4 -2.2235295 -0.5740342
##
## Clustering vector:
##  [1] 1 2 4 3 3 4 2 3 2 4 2 4 3 4 2 3 2 2 4 4 3 2
##
## Within cluster sum of squares by cluster:
## [1] 0.000000 3.109136 4.375350 8.403846
##  (between_SS / total_SS =  89.7 %)
##
## Available components:
##
```

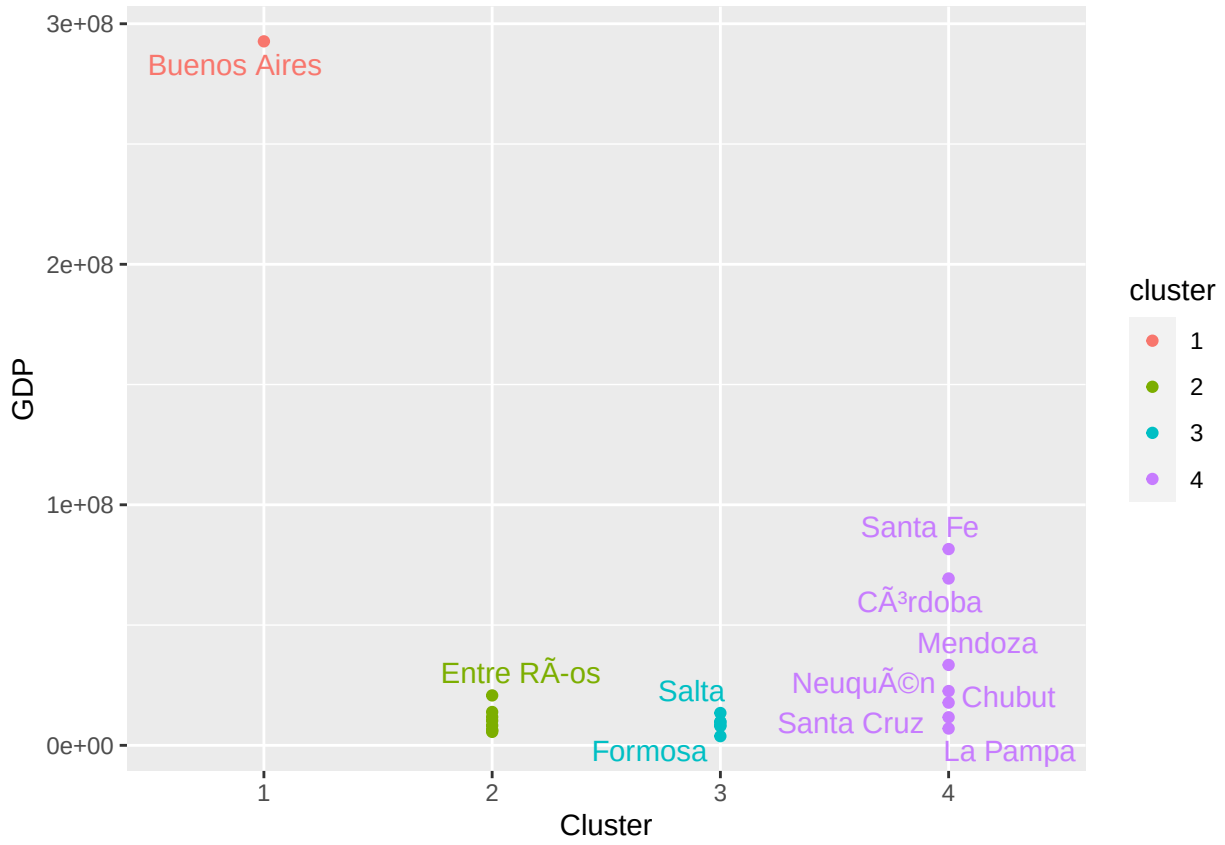
```
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
# Convert assigned clusters to factor
clusters_as_factor = factor(argentina_km$cluster)
# Plot individulas colored by cluster
fviz_pca_ind(argentina_pca,
             title = "Clustered Provinces - PCA",
             habillage = clusters_as_factor)
```

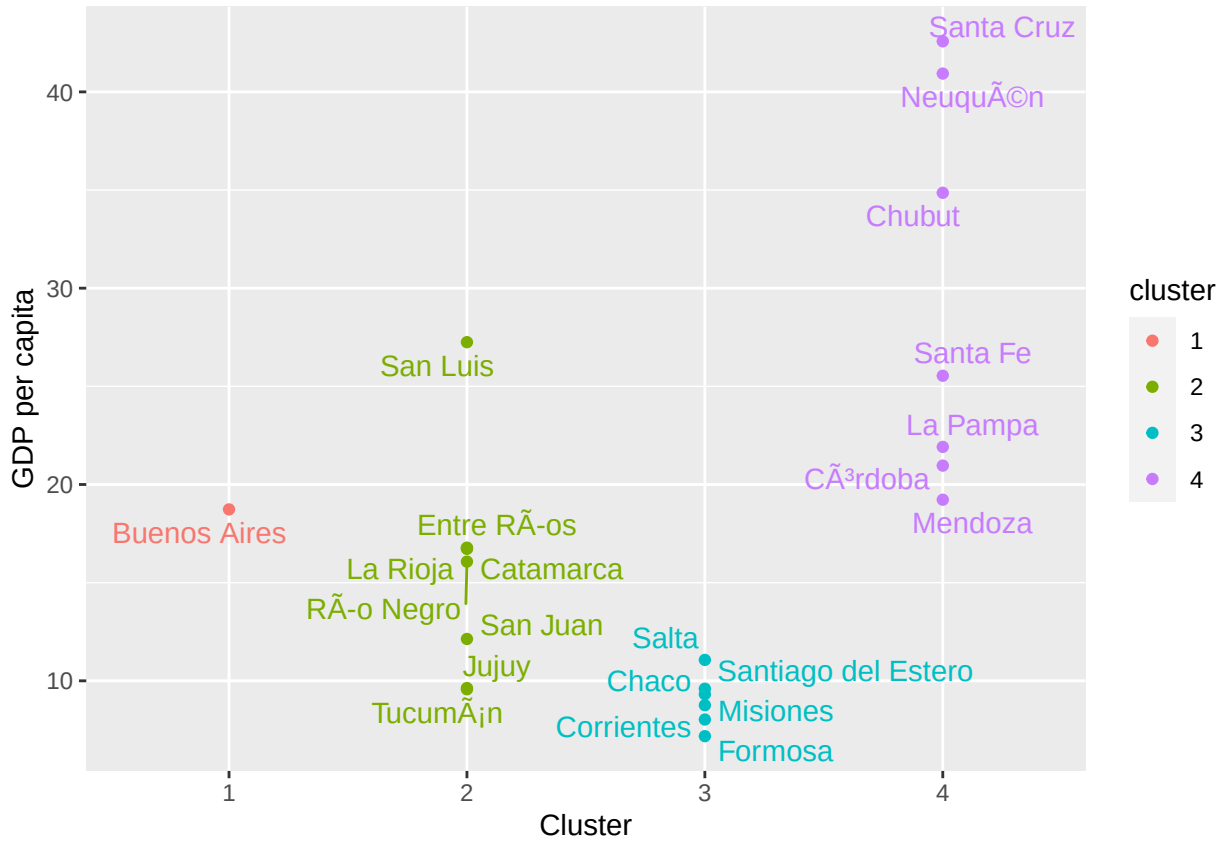



```
# Load ggrepel
library(ggrepel)
# Add cluster column to argentina
argentina <- argentina %>%
             mutate(cluster=clusters_as_factor)
# Make a scatterplot of gdp vs. cluster, colored by cluster
ggplot(argentina, aes(cluster, gdp, color = cluster)) +
  geom_point() +
  geom_text_repel(aes(label = province), show.legend = FALSE) +
  labs(x = "Cluster", y = "GDP")
```

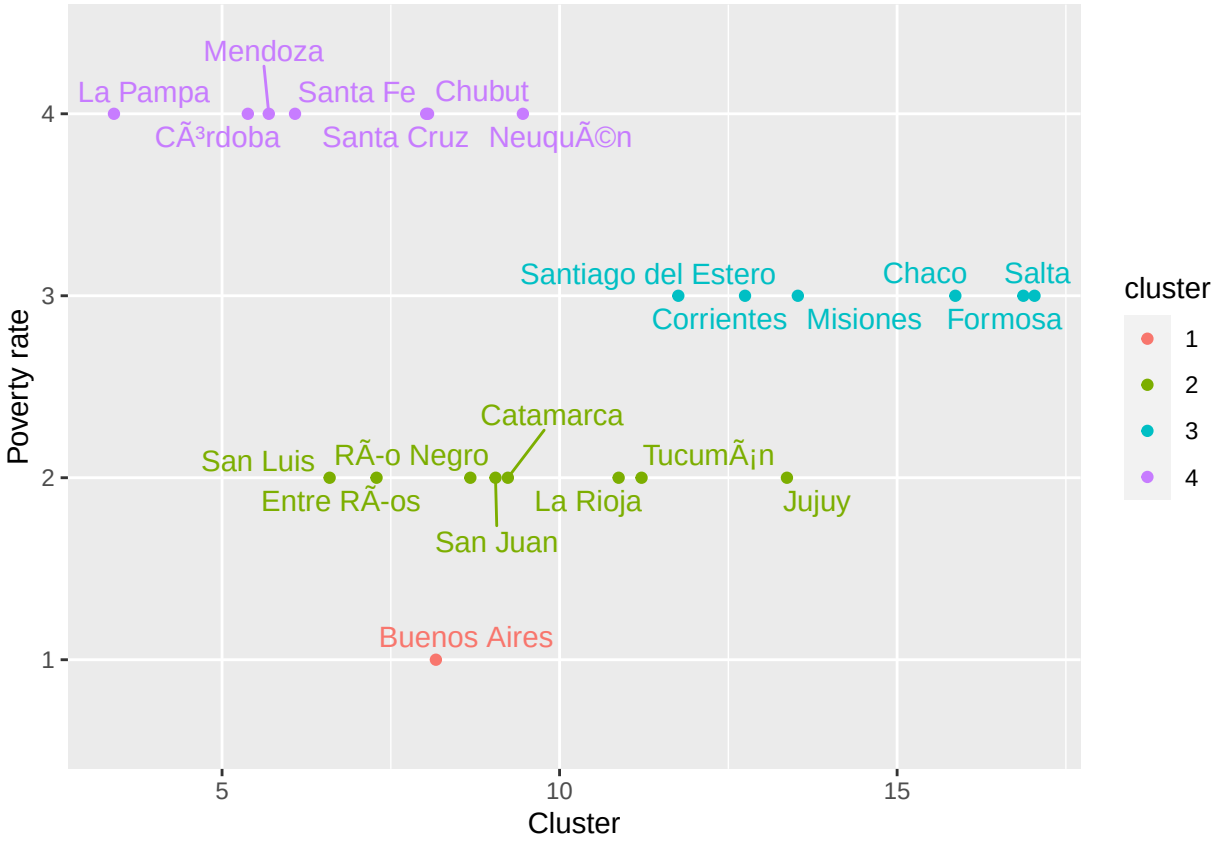
```
## Warning: ggrepel: 11 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
# Make a scatterplot of GDP per capita vs. cluster, colored by cluster
ggplot(argentina, aes(cluster, gdp_per_cap, color = cluster)) +
  geom_point() +
  geom_text_repel(aes(label = province), show.legend = FALSE) +
  labs(x = "Cluster", y = "GDP per capita")
```




```
# Make scatterplot of poverty vs. cluster, colored by cluster
ggplot(argentina, aes(poverty, cluster, color = cluster)) +
  geom_point() +
  labs(x = "Cluster", y = "Poverty rate") +
  geom_text_repel(aes(label = province), show.legend = FALSE)
```


Mendoza
La Pampa
Santa Fe
Chubut
CÃ³rdoba
Santa Cruz
NeuquÃ©n
Santiago del Estero
Chaco
Salta
Corrientes
Misiones
Formosa
Catamarca
San Luis
RÃ-o Negro
TucumÃ¡n
Entre RÃ-os
La Rioja
Jujuy
San Juan
Buenos Aires
Poverty rate
Cluster
cluster
1
2
3
4