Assisting Development Policy-Making Process in Mexico

Oguzhan Gurbuz

2023-03-20

#1. INTRODUCTION

#This project uses Principal Component Analysis (PCA) and K-means clustering method to group the regions with similar development levels in Mexico to assist regional development policy-making. Regional development is an important element of a country's development strategy, and policymakers must understand regional differences in order to make sound decisions about resource distribution, infrastructure development, and social programmes. In this analysis, I will use the data from Mexico at the regional level and try to give insights into the policy-making processes. I will weigh three main indicators (explained below) differently to design a development model which will be used to cluster the Mexican regions. The insights gained from this analysis can help policymakers identify regional disparities and make informed decisions about development policies.

#Mexico is a developing country with more than 126 million of population. It is one of the members of the OECD and is a developing country. For the research, I collected the data from the OECD Regional Statistics page, including a member country database. Available at: https://www.oecd.org/regional/regional-statistics/ (https://www.oecd.org/regional/regional-statistics/)

#The data mainly includes several indicators and key points on development (OECD, 2018, p. 13) of the member countries, listed below:

#1. Material conditions

#1.1. Income #1.2. Jobs #1.3. Housing

#2. Quality of Life

#2.1. Health #2.2. Education #2.3. Environment #2.4. Safety #2.5. Civic Engagement

#3. Subjective Well-Being

#3.1. Community #3.2. Life satisfaction

#2. DATA AND BACKGROUND #INSTALLING AND LOADING THE PACKAGES

```
library(readxl)
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## ── Attaching packages
## ─────────────────────────────────────
## tidyverse 1.3.2 ──
```

```
## ✓ ggplot2 3.4.0     ✓ purrr   0.3.5
## ✓ tibble  3.1.8     ✓ stringr 1.4.1
## ✓ tidyr   1.2.1     ✓ forcats 0.5.2
## ── Conflicts ──────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(tidyr)
library(FactoMineR)
library(factoextra)
```

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

```
library(ggplot2)
library(ggrepel)
```

#EXPLORING THE DATA

```
#Loading the data
mexico_data <- read_excel("Mexico-Regional Level Data.xlsx")

#A few observations from the data
head(mexico_data)
```

```
## # A tibble: 6 × 27
##   Provinces  Popul…¹ Educa…²  Jobs Income  Safety Health Envir…³ Civic…⁴ Acces…⁵
##   <chr>        <dbl>   <dbl> <dbl>  <dbl>   <dbl>  <dbl>   <dbl>   <dbl>   <dbl>
## 1 Aguascali… 1425607    1.70  5.94  0.176 8.42      3.02    5.62 1.51       4.16
## 2 Baja Cali… 3769020    1.63  6.79  0.438 0.00224   2.43    4.33 0.00997    4.92
## 3 Baja Cali…  798447    2.20  7.14  0.338 7.64      2.95    7.06 1.36       4.54
## 4 Campeche    928363    1.31  6.84  0.01  8.10      2.41    4.73 4.43       3.63
## 5 Coahuila   3146771    1.41  5.93  0.211 8.04      2.62    5.77 3.16       4.22
## 6 Colima      731391    1.76  7.83  0.270 0.00147   2.24    6.37 2.19       4.06
## # … with 17 more variables: Housing <dbl>, Community <dbl>,
## #   `Life satisfaction` <dbl>, `Secondary education` <dbl>,
## #   `Employment rate` <dbl>, `Unemployment rate` <dbl>,
## #   `Income per capita` <dbl>, `Homicide rate` <dbl>, `Mortality rate` <dbl>,
## #   `Life expectancy` <dbl>, `Air pollution (level of PM2.5)` <dbl>,
## #   `Voter turnout` <dbl>, `Broadband access` <dbl>,
## #   `Internet download speed 2021-Q4` <dbl>, …
```

```
#Number of observations and categories
dim(mexico_data) #The data set includes 32 regions in Mexico and 26 variables, excluding t
he column of the region names.
```

```
## [1] 32 27
```

```
#Titles of the Columns
colnames(mexico_data)
```

```
##  [1] "Provinces"
##  [2] "Population"
##  [3] "Education"
##  [4] "Jobs"
##  [5] "Income"
##  [6] "Safety"
##  [7] "Health"
##  [8] "Environment"
##  [9] "Civic engagement"
## [10] "Accessiblity to services"
## [11] "Housing"
## [12] "Community"
## [13] "Life satisfaction"
## [14] "Secondary education"
## [15] "Employment rate"
## [16] "Unemployment rate"
## [17] "Income per capita"
## [18] "Homicide rate"
## [19] "Mortality rate"
## [20] "Life expectancy"
## [21] "Air pollution (level of PM2.5)"
## [22] "Voter turnout"
## [23] "Broadband access"
## [24] "Internet download speed 2021-Q4"
## [25] "Number of rooms per person"
## [26] "Perceived social network support"
## [27] "Self assessment of life satisfaction"
```

```
#Missing values
colSums(is.na(mexico_data)) #There's no missing values
```

```
##                            Provinces                           Population
##                                    0                                    0
##                            Education                                 Jobs
##                                    0                                    0
##                               Income                               Safety
##                                    0                                    0
##                               Health                          Environment
##                                    0                                    0
##                    Civic engagement              Accessiblity to services
##                                    0                                    0
##                              Housing                            Community
##                                    0                                    0
##                    Life satisfaction                  Secondary education
##                                    0                                    0
##                     Employment rate                    Unemployment rate
##                                    0                                    0
##                    Income per capita                        Homicide rate
##                                    0                                    0
##                       Mortality rate                       Life expectancy
##                                    0                                    0
##       Air pollution (level of PM2.5)                         Voter turnout
##                                    0                                    0
##                     Broadband access      Internet download speed 2021-Q4
##                                    0                                    0
##           Number of rooms per person     Perceived social network support
##                                    0                                    0
## Self assessment of life satisfaction
##                                    0
```

```
#Structure of the data
class(mexico_data) #It's a data frame
```

```
## [1] "tbl_df"     "tbl"        "data.frame"
```

#GENERAL INFORMATION ABOUT MEXICO

#With a population of almost 130 million, Mexico is the world's 11th most populous country and 3rd in the Americas after the U.S. and Brazil. The majority of the country's population lives in the central and southern regions, and Mexico City, as the primary metropolitan hub.

#Mexico's economy is the second-largest in Latin America after Brazil. The industrial sector, particularly the car industry, dramatically contributes to the country's economy. It has more than 1.2 Trillion USD Gross Domestic Product (GDP) (2021). However, the country has a variety of challenges, including income disparity, poverty, and a significant informal sector.

#ASSIGNING THE VALUES

```
#I will assign the variables to be used in the study.
regions <- mexico_data$Provinces
population <- mexico_data$Population
education <- mexico_data$Education
jobs <- mexico_data$Jobs
income <- mexico_data$Income
safety <- mexico_data$Safety
health <- mexico_data$Health
environment <- mexico_data$Environment
civic_engagement <- mexico_data$`Civic engagement`
accesibility <- mexico_data$`Accessiblity to services`
housing <- mexico_data$Housing
community <- mexico_data$Community
life_satisfaction <- mexico_data$`Life satisfaction`
income_pct <- mexico_data$`Income per capita`
mortality <- mexico_data$`Mortality rate`
life_expectancy <- mexico_data$`Life expectancy`
```

#3. INCOME

#Income is an important indicator of understanding poverty. To carry out an evaluation, it'd be good to look at the income-based ranking of the regions in Mexico. We can do this by using the 'dplyr' package.

```
#Regions with the highest income per capita (USD) --> Sorting the data in descending order
and keeping the top five rows
highest_income <- mexico_data %>%
  arrange(desc(income_pct)) %>%
  select(all_of(c("Provinces", "Income per capita"))) %>%
  top_n(5)
```

```
## Selecting by Income per capita
```

```
#Regions with the lowest income per capita (USD) --> Sorting the data in ascending order a
nd keeping the bottom five rows
lowest_income <- mexico_data %>%
  arrange(income_pct) %>%
  select(all_of(c("Provinces", "Income per capita"))) %>%
  top_n(-5)
```

```
## Selecting by Income per capita
```

#CHECKING THE RESULTS

```
#Top 5 Regions
print(highest_income)
```

```
## # A tibble: 6 × 2
##   Provinces           `Income per capita`
##   <chr>                             <dbl>
## 1 Baja California                    5744
## 2 Nuevo Leon                         5523
## 3 Baja California Sur                5257
## 4 Distrito Federal                   5069
## 5 Colima                             4924
## 6 Chihuahua                          4924
```

```
#Baja California, Nuevo Leon, Baja California Sur, Distrito Federal, Colima and Chihuahua

#Bottom 5 Regions
print(lowest_income)
```

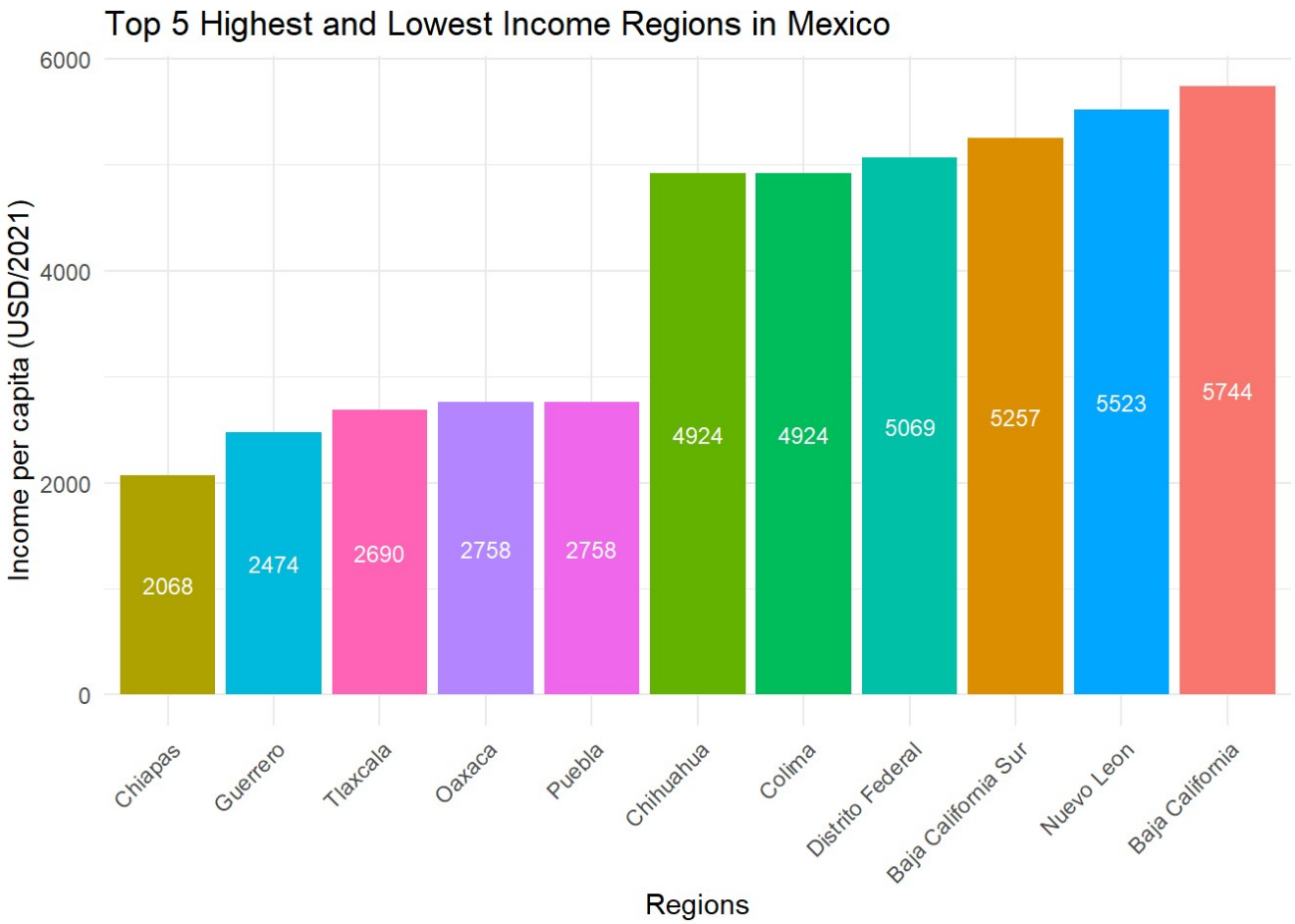
```
## # A tibble: 5 × 2
##   Provinces `Income per capita`
##   <chr>                   <dbl>
## 1 Chiapas                  2068
## 2 Guerrero                 2474
## 3 Tlaxcala                 2690
## 4 Oaxaca                   2758
## 5 Puebla                   2758
```

```
#Chiapas, Guerrero, Tlaxcala, Oaxaca, Puebla
```

#VISUALIZING THE RESULTS

```
#Combining the highest and lowest-income regions
income_levels <- rbind(highest_income, lowest_income)

#Creating a bar chart
ggplot(income_levels, aes(x = reorder(Provinces, `Income per capita`), y = `Income per cap
ita`, fill = Provinces)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(`Income per capita`, 2)), position = position_stack(vjust =
0.5), color = "white", size = 3) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Regions", y = "Income per capita (USD/2021)", title = "Top 5 Highest and Lowes
t Income Regions in Mexico") +
  guides(fill = FALSE)
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none" instead as
## of ggplot2 3.3.4.
```



#OBSERVATIONS

#Top 5 Regions: 1. Baja California 2. Nuevo Leon 3. Baja California Sur 4. Distrito Federal (Ciudad de Mexico) 5. Colima and Chihuahua

#5 Regions with the Lowest Income Values: 1. Chiapas 2. Guerrero 3. Tlaxcala 4. Oaxaca 5. Puebla

#4. POPULATION

#Population is also a significant indicator of developing public policies in a country and is an essential aspect of a country's development planning. In other words, a high population can influence many facets of a country's economic, social, and political development. Policymakers must understand their nation's demographic trends to make sound decisions about resource allocation, infrastructure development, and social programmes. A rising and youthful population can bring a demographic dividend to a country, but an elderly population might provide issues to the labour market and healthcare system.

#To this end, we can rank the regions based on their population. I will filter out the regions with more than 1 million population and this will help us understand which regions are more populous in the country.

```
#Filtering the regions by the population
bigger_pops <- mexico_data %>%
  arrange(desc(population)) %>%
  select(c("Provinces", "Population")) %>%
  filter(population > 1000000)

#Checking the results
print(bigger_pops) #26 out of 33 provinces have more than 1 million population
```

```
## # A tibble: 29 × 2
##    Provinces        Population
##    <chr>                 <dbl>
##  1 State of Mexico    16992418
##  2 Distrito Federal    9209944
##  3 Puebla              6583278
##  4 Nuevo Leon          5784442
##  5 Chiapas             5543828
##  6 Michoacan           4748846
##  7 Oaxaca              4132148
##  8 Baja California     3769020
##  9 Chihuahua           3741869
## 10 Guerrero            3540684
## # … with 19 more rows
```

#5. CREATING A NEW DEVELOPMENT MODEL

#I will create a model to make it easy to see what different Mexican regions look like in terms of development levels. This will be useful to show policymakers regional disparities and how regions differ and cluster. The model contains three different parts: #i. Material conditions ii. Quality of Life, and iii. Subjective Well-Being. I will make each component weigh differently in the model.

#First, “material conditions” considers household disposable income, employment and unemployment rate, and the number of rooms per person. It has a value between 0 and 10. Second, “Quality of Life” considers health, education, environment, safety, civic engagement, and access to services (0 to 10). Finally, “Subjective Well-Being” measures the percentage of people having friends to rely on and the life satisfaction of a population. It also has values between 0 and 10.

#Overall, the model is calculated by finding the average value from these three sections, providing a comprehensive overview of development and well-being.

#I will weigh the components differently as shown in the following model: #Our Development Model = Material Conditions * 0.4 + Quality of Life * 0.4 + Subjectivity * 0.2

#P.S: Tough some indicators are generally measured by percentage values (like the share/rate of X), the OECD data has values between 0 and 10 for some indicators.

#MODEL FOR DEVELOPMENT POLICY

```
#The three components in the model and their average values
#1. Material conditions

material = (income * 0.45 + jobs * 0.45 + housing * 0.1)

  #1.1. Income: Household disposable income per capita (in real USD PPP) - 0 to 10
  #1.2. Jobs: Employment and Unemployment Rate - 0 to 10
  #1.3. Housing: Number of rooms per person (ratio)

#2. Quality of Life

quality = (health * 0.35 + education * 0.3 + environment * 0.25 + safety * 0.05 + civic_en
gagement * 0.05)

  #2.1. Health: Life expectancy at birth (years) and Age-adjusted mortality rate (per 1 00
0 people) - 0 to 10
  #2.2. Education: Share of the labour force with at least secondary education - 0 to 10
  #2.3. Environment: Estimated average exposure to air pollution in PM2.5 (µg/m³), based o
n satellite imagery data-0 to 10
  #2.4. Safety: Homicide rate (per 100 000 people) - 0 to 10
  #2.5. Civic Engagement: Voter turnout - 0 to 10


  #3. Subjective Well-Being

subjectivity = (community * 0.3 + life_satisfaction * 0.7)

  #3.1. Community: Percentage of people who have friends or relatives to rely on in case o
f need
  #3.2. Life satisfaction: Average self-evaluation of life satisfaction on a scale from 0
to 10

#I will assign weights to each component to show their importance in the development model
and score.
#0.4 (40%) for material conditions, 0.40 (40%) for quality conditions, 0.2 (20%) for subje
ctivity conditions

#Final model for our analysis
development_model = material * 0.4 + quality * 0.4 + subjectivity * 0.2
```

#Adding the Model

```
#Adding updated model to the data set for our analysis
mexico_data <- mexico_data %>%
  mutate(
       development_model = (material * 0.4) +
                    (quality * 0.4) +
                    (subjectivity * 0.2)
  )
```

#Ranking the regions based on our new model

```
#Regions with the highest development scores
highest_dev_scores <- mexico_data %>%
  arrange(desc(development_model)) %>%
  select(Provinces, development_model) %>%
  top_n(5)
```

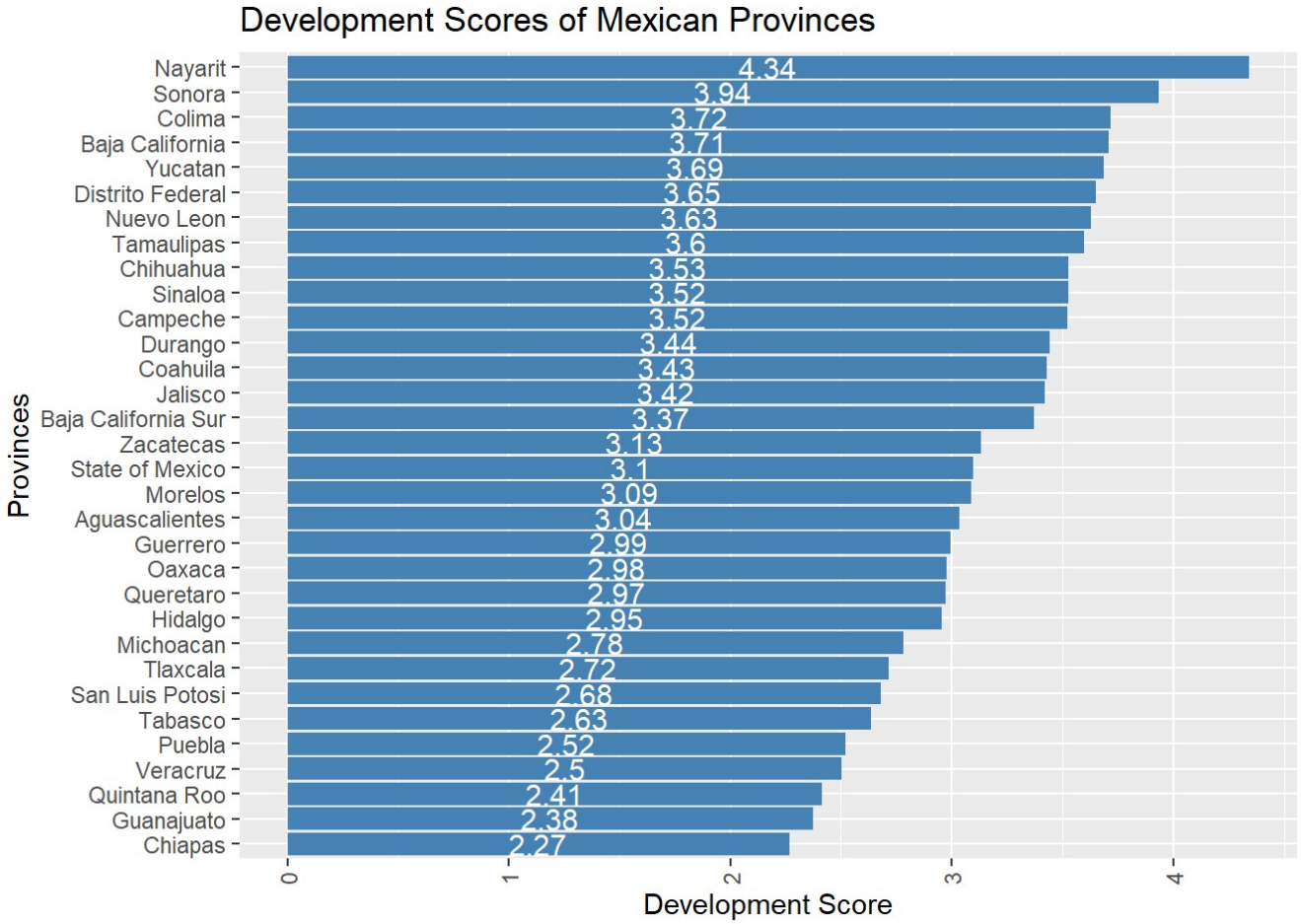
```
## Selecting by development_model
```

```
#Regions with the lowest development scores
lowest_dev_scores <- mexico_data %>%
  arrange(development_model) %>%
  select(Provinces, development_model) %>%
  top_n(-5)
```

```
## Selecting by development_model
```

#All Regions

```
#All regions sorted by development score
all_dev_scores <- mexico_data %>%
  arrange(desc(development_model)) %>%
  select(Provinces, development_model)

#Creating the bar plot for all regions
ggplot(all_dev_scores, aes(x = reorder(Provinces, development_model), y = development_mode
l)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(development_model, 2)), position = position_stack(vjust = 0.
5), color = "white", size = 4) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust = 1)) +
  labs(title = "Development Scores of Mexican Provinces",
       x = "Provinces",
       y = "Development Score") +
  coord_flip()
```



#Top and Bottom 5 Scores

```
#Top 5 Regions with the Highest Development scores
print(highest_dev_scores) #Our model demonstrates that the top 5 regions are: Nayarit, Son
ora, Colima, Baja California, Yucatan
```

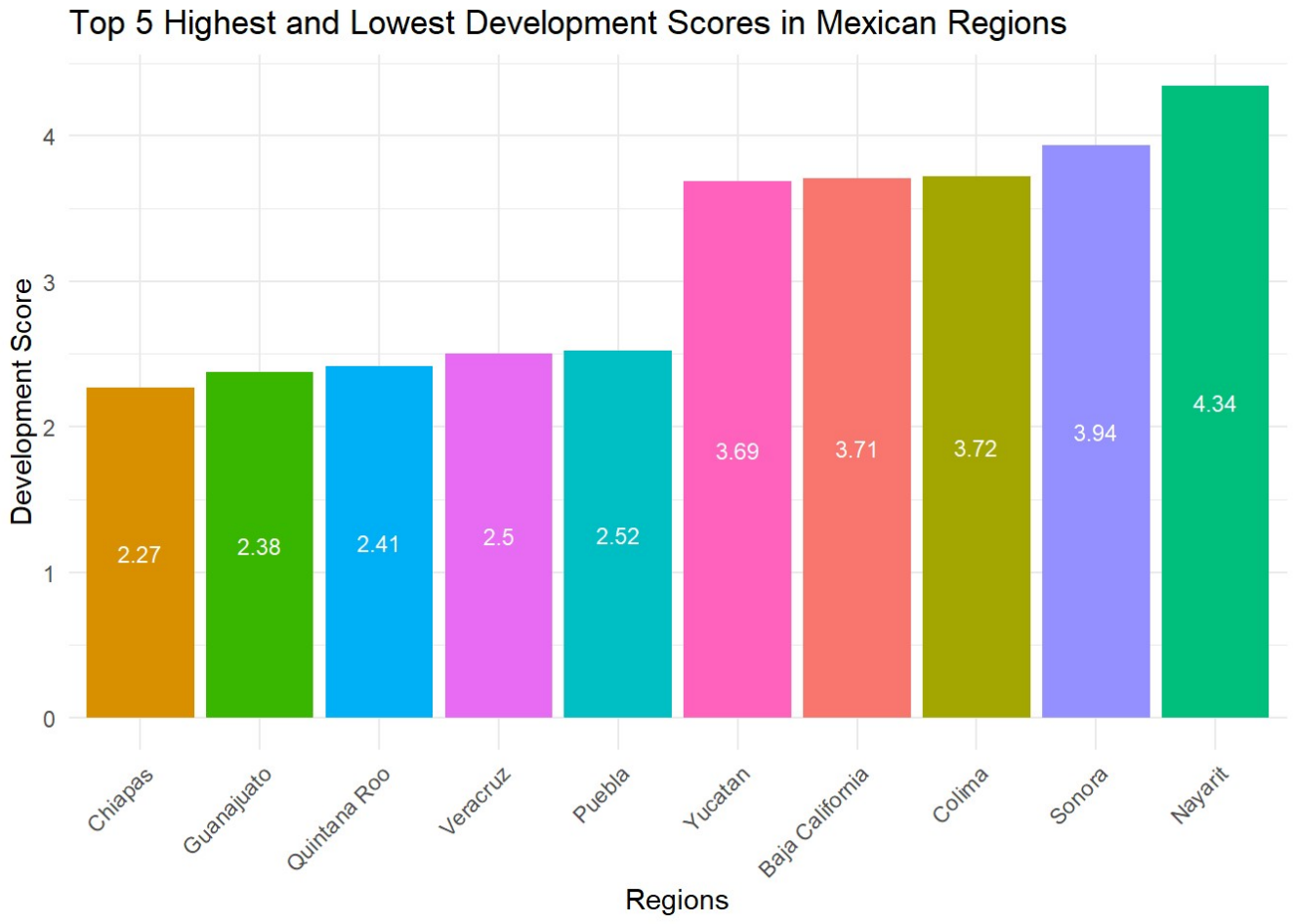
```
## # A tibble: 5 × 2
##   Provinces       development_model
##   <chr>                       <dbl>
## 1 Nayarit                      4.34
## 2 Sonora                       3.94
## 3 Colima                       3.72
## 4 Baja California              3.71
## 5 Yucatan                      3.69
```

```
#Bottom 5 Regions with the Lowest Development scores
print(lowest_dev_scores) #Our model demonstrates that the 5 regions with the lowest scores
are: Chiapas, Guanajuato, Quintana Roo, Veracruz, Puebla
```

```
## # A tibble: 5 × 2
##   Provinces    development_model
##   <chr>                    <dbl>
## 1 Chiapas                   2.27
## 2 Guanajuato                2.38
## 3 Quintana Roo              2.41
## 4 Veracruz                  2.50
## 5 Puebla                    2.52
```

#Visualizing the development ranking

```
#Combining the highest and lowest development scores
high_low_regions <- rbind(highest_dev_scores, lowest_dev_scores)

#Creating a bar chart
ggplot(high_low_regions, aes(x = reorder(Provinces, development_model), y = development_mo
del, fill = Provinces)) +
  geom_bar(stat = "identity") +
  geom_text(aes(label = round(development_model, 2)), position = position_stack(vjust = 0.
5), color = "white", size = 3) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x = "Regions", y = "Development Score", title = "Top 5 Highest and Lowest Developme
nt Scores in Mexican Regions") +
  guides(fill = FALSE)
```



#6. PRINCIPAL COMPONENT ANALYSIS (PCA)

#Principal Component Analysis (PCA) is a powerful and widely used technique to reduce dimensionality and create data visualizations. It allows us to visualize the relationships between observations (provinces) in a lower-dimensional space while retaining as much of the original information as possible. It is also a great technique before applying the k-means clustering.

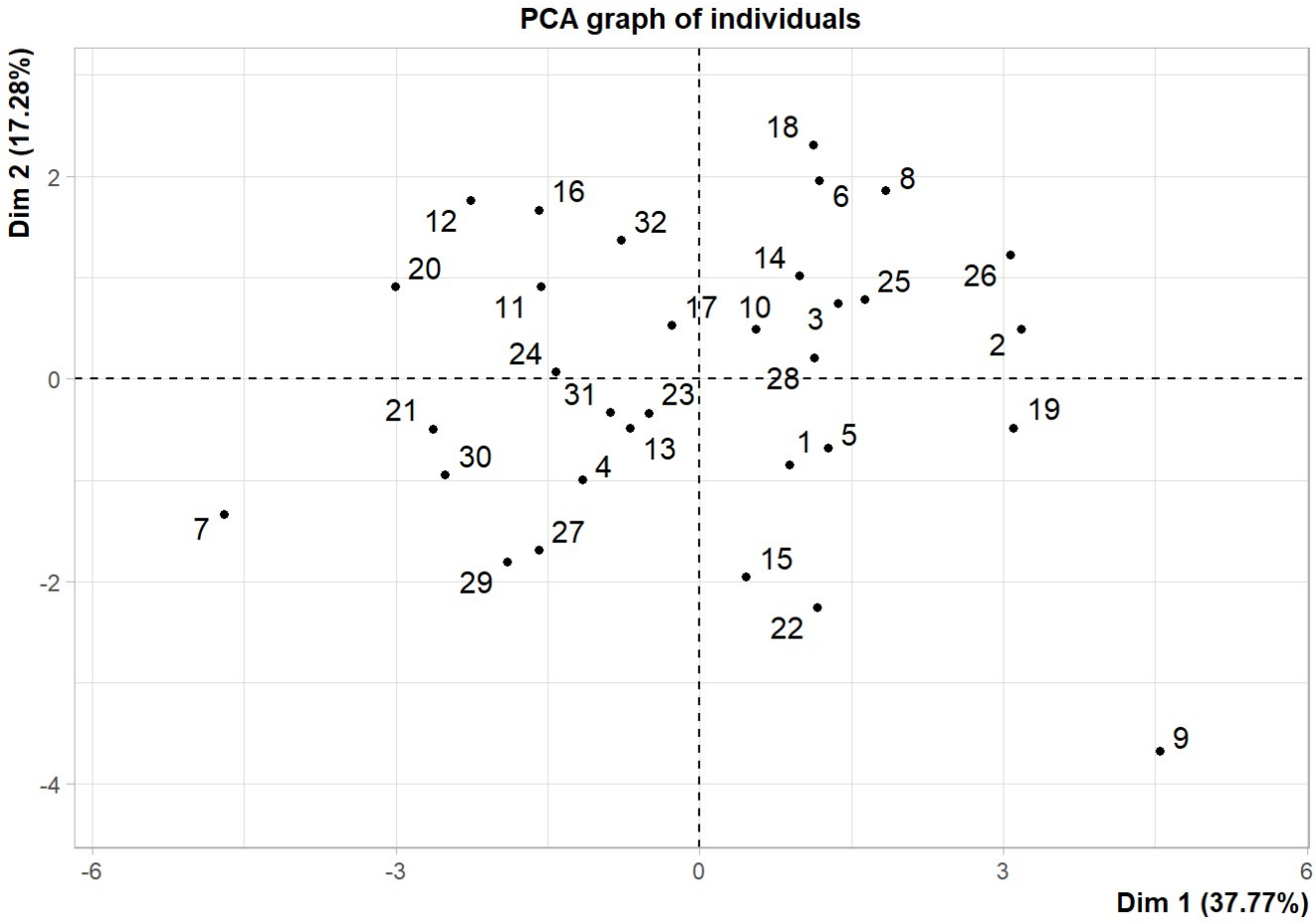
#First, we need to choose the numeric columns and assign them as a new matrix.

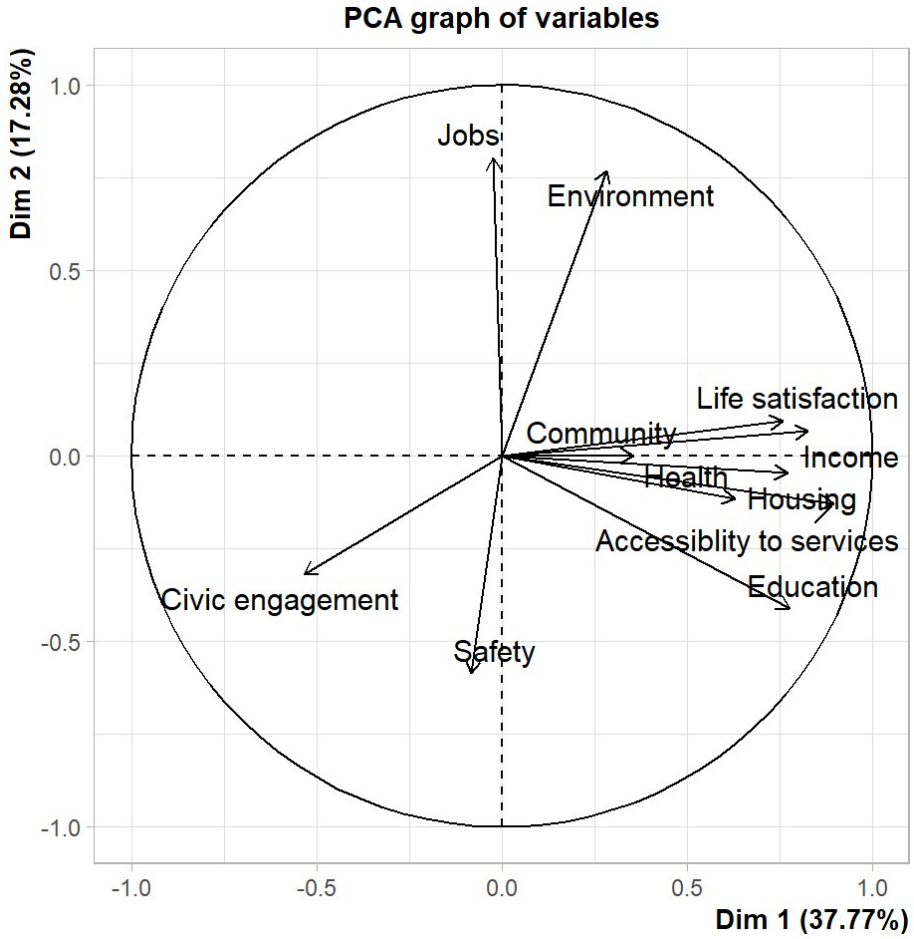
```
# Selecting numeric columns and casting to matrix
mexico_mat  <- mexico_data  %>%
  select(3:13) %>%
  as.matrix()

# Checking a few lines of the numeric columns
head(mexico_mat)
```

```
##      Education     Jobs   Income   Safety   Health Environment Civic engagement
## [1,]  1.700443 5.942970 0.176240 8.423913 3.017920    5.621891         1.513648
## [2,]  1.629428 6.785714 0.437541 0.002240 2.430376    4.328358         0.009974
## [3,]  2.203604 7.142387 0.338202 7.635870 2.950353    7.064677         1.357997
## [4,]  1.309072 6.836815 0.010000 8.097826 2.413337    4.726368         4.432664
## [5,]  1.413771 5.934971 0.211325 8.043478 2.618684    5.771144         3.158132
## [6,]  1.758074 7.831357 0.270276 0.001470 2.239718    6.368159         2.188134
##      Accessiblity to services  Housing Community Life satisfaction
## [1,]                 4.155700 1.292135  1.081081          3.076923
## [2,]                 4.922864 1.235955  5.225225          7.692308
## [3,]                 4.542863 0.505618  0.180180          3.076923
## [4,]                 3.629144 0.010000  6.216216          5.000000
## [5,]                 4.219586 1.685393  0.009986          6.538462
## [6,]                 4.062717 0.842697  2.657658          6.153846
```

#PCA

```
#Applying Principal Component Analysis (PCA)
mexico_pca  <- PCA(mexico_mat, scale.unit = TRUE) #Performing PCA with the FactoMineR pack
age and scaling units to avoid large variances
```





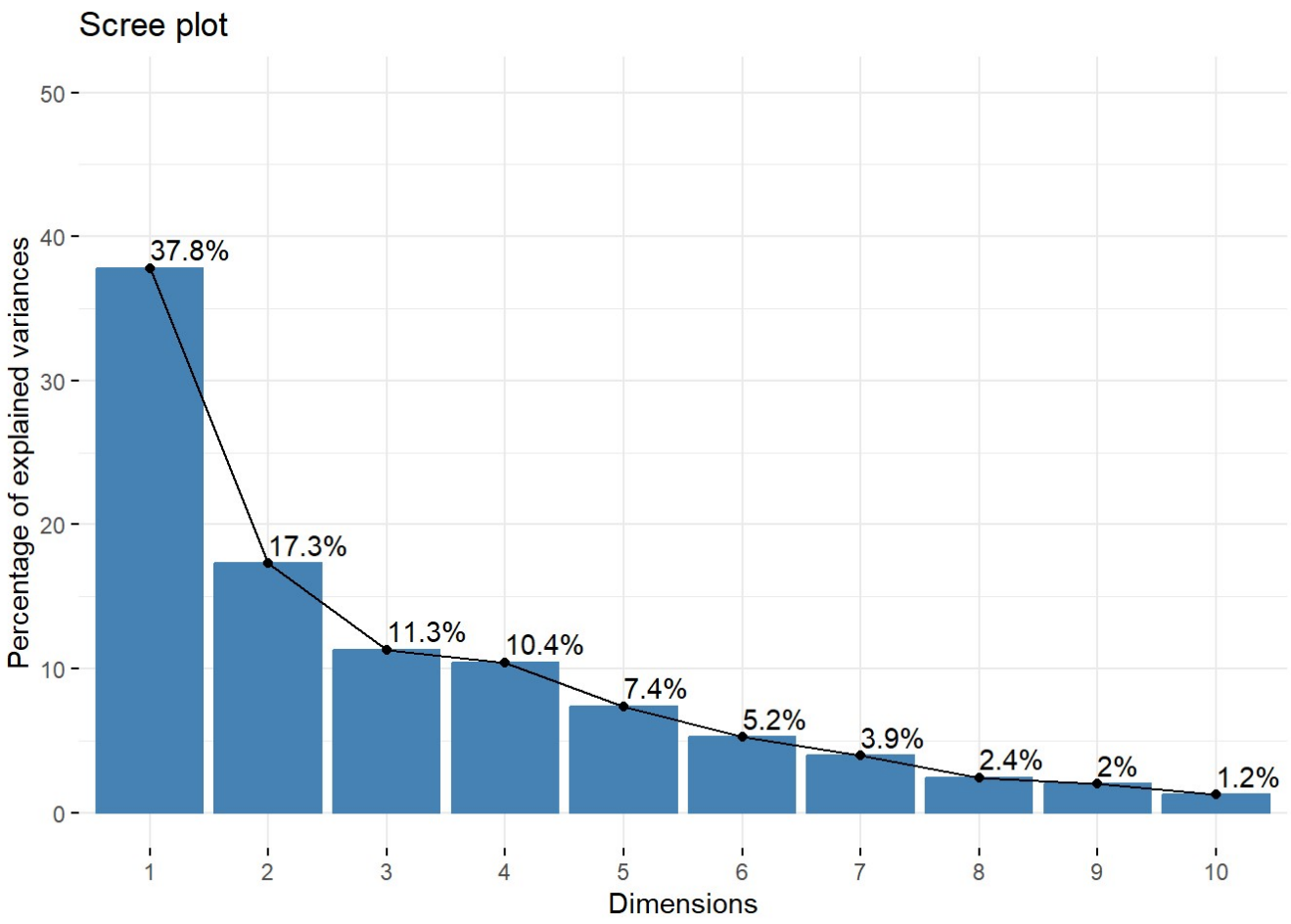

```
#Checking the results
print(mexico_pca)
```

```
## **Results for the Principal Component Analysis (PCA)**
## The analysis was performed on 32 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"
```

#VISUALIZING THE PCA

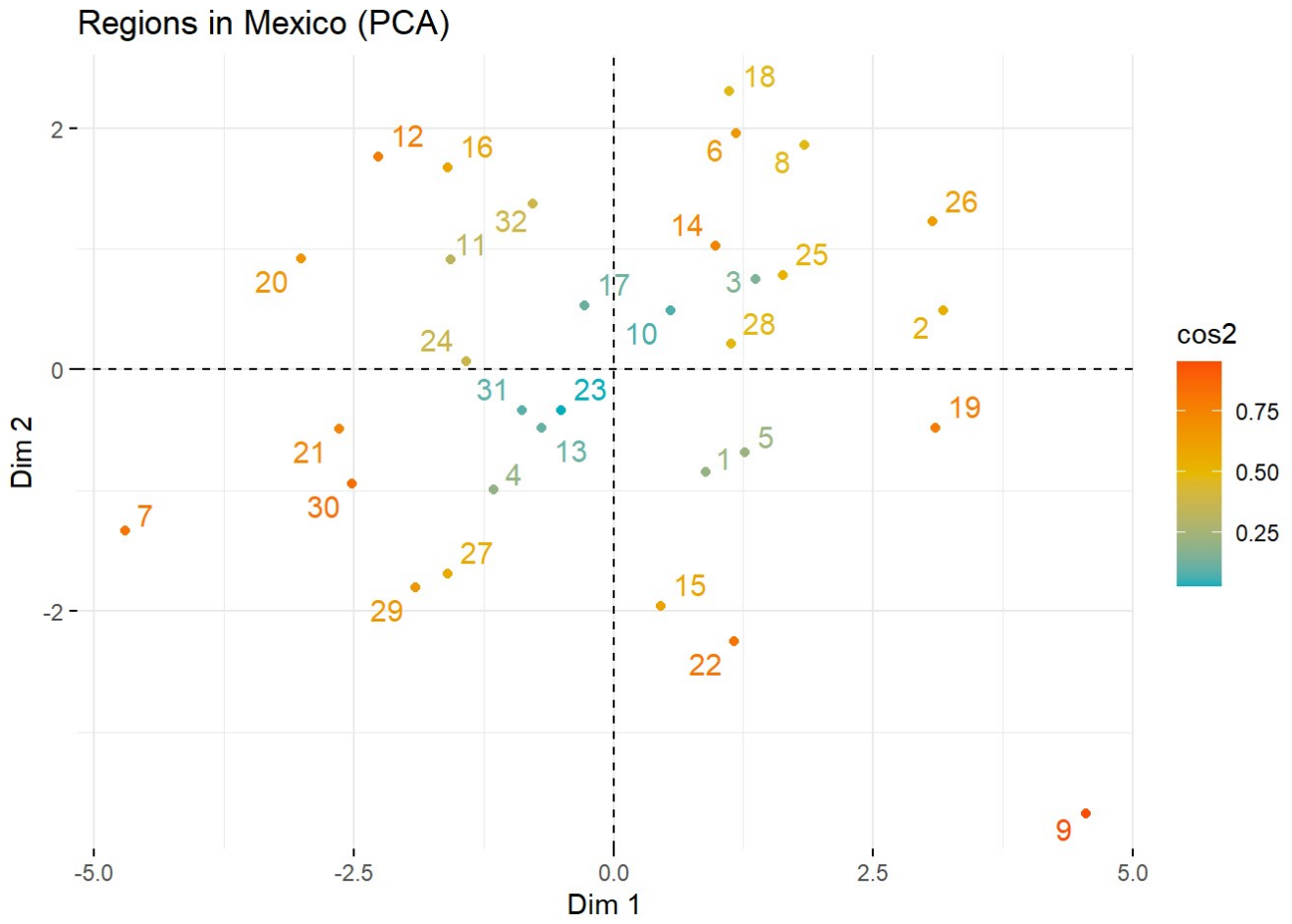
#Scree Plot

```
#Visualizing the results with a scree plot (eigenvalues)
fviz_eig(mexico_pca, addlabels = TRUE, ylim = c(0, 50))
```



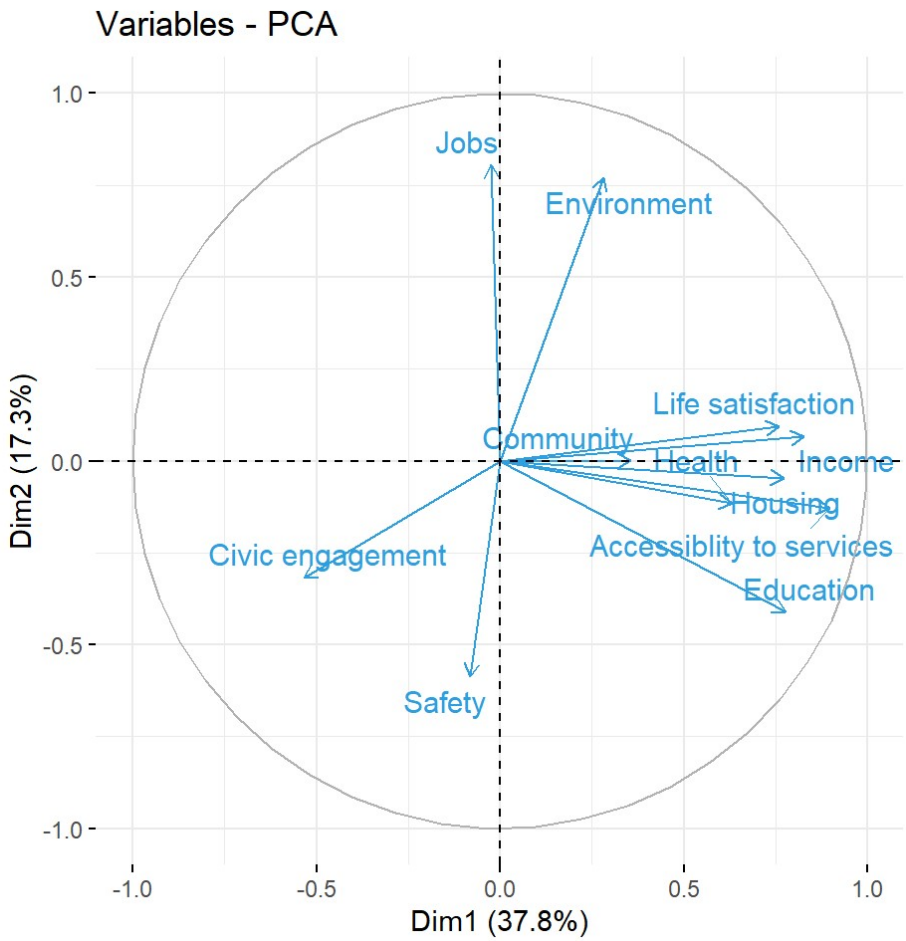
#Visualizing the Mexican Provinces on the first two principal component axes #We can plot the Mexican regions based on their principal component coordinates.

```
#Visualize the individuals (regions) on the principal component axes
fviz_pca_ind(mexico_pca,
             col.ind = "cos2", #Color by the quality of representation
             gradient.cols = c("#00AFBB", "#E7B800", "#FC4E07"),
             repel = TRUE, #Avoid text overlapping
             title = "Regions in Mexico (PCA)",
             xlab = "Dim 1",
             ylab = "Dim 2")
```



#The scatter plot of the two principal components (Dim1 and Dim2) allows us to visualize the distribution of the data points. The scatter plot provides insight into the relationship between the original variables and the first two components and can help us understand the variability these two components explain. The position of the data points on the scatter plot allows us to identify any potential outliers or groups within the data. In the plot, each point represents a single region. It also shows the relationship between the provinces based on their scores/colours on the first two principal components. The areas being close to each other and having the same colours demonstrate similar well-being/development values.

```
#Visualizing the results with a Variables plot
fviz_pca_var(mexico_pca, col.var = "#2E9FDF", repel = TRUE)
```



#Comments on PCA and Visualizations #The Dim1 and Dim2 values show the principal components in the three different PCA plots. These two percentage values (Dim1 is 37.8% & Dim2 is 17.3%) demonstrate the variance values of the data explained by the two principal components.

#In other words, Dim1 is 37.8% meaning the first principal component explains 37.8% of the total variance in the dataset (the largest explaining variance). Besides, Dim2 is 17.3% indicating the second principal component explains 17.3% of the total variance (second largest explaining variance).

#Finally, these two values explain 55.1% or more than half of the total variance (37.8% + 17.3%).

#Based on the variables plot, we can say health, housing, community, income and life satisfaction are correlated.

```
#Showing differently: let's sum the variances (Dim1 and Dim2 as two components).
variance_first_two_pca <- mexico_pca$eig[1, 2] + mexico_pca$eig[2, 2]

variance_first_two_pca #55.1%
```

```
## [1] 55.05064
```

#7. CLUSTERING WITH K-MEANS

#Clustering is an unsupervised machine learning technique to group comparable data based on their features. The K-means clustering method splits the data into specified groups. In this section, we will attempt to identify distinct groups of regions with comparable elements in order to design policies and implement policy interventions.

#K-means Clustering

#In this section, we will apply k-means clustering to group comparable data and we will use the first two principal components as features to cluster the observations.

```
#Let's set the seed to 1234 for reproducibility
set.seed(1234)

#A few lines of the data
head(mexico_pca$ind$coord)
```

```
##        Dim.1      Dim.2       Dim.3      Dim.4      Dim.5
## 1  0.8843723 -0.8455517  0.56405897 -2.0747692 -0.4525334
## 2  3.1740636  0.4874337 -1.30077873  1.9074663  0.7964506
## 3  1.3672198  0.7442913  1.01466887 -2.8569479 -0.5082131
## 4 -1.1565213 -0.9913798  2.08378097  1.8919028 -0.5558714
## 5  1.2669947 -0.6824698  0.80701916 -1.5485495  1.7977894
## 6  1.1840537  1.9538596 -0.04561048  0.2284943  0.9121276
```

#PCA Scores and Dimensions

```
#Extracting the PCA scores
pca_scores <- mexico_pca$var$coord

#Displaying the PCA scores for Dim1 and Dim2
mexican_pca1_pca2 <- pca_scores[, c(1, 2)]
mexican_pca1_pca2
```

```
##                              Dim.1        Dim.2
## Education                 0.77743677 -0.411114850
## Jobs                     -0.02383758  0.804987080
## Income                    0.82613049  0.066518282
## Safety                   -0.08262809 -0.586703993
## Health                    0.63045628 -0.116204959
## Environment               0.28268994  0.770069248
## Civic engagement         -0.53413336 -0.318220923
## Accessiblity to services  0.89584237 -0.128813074
## Housing                   0.77135944 -0.046674606
## Community                 0.35514933  0.001251356
## Life satisfaction         0.75742797  0.094962594
```

##The result tells us excluding jobs, safety and civic engagement, all other components contribute positively to the Dimension 1. This implies that Dimension 1 could be seen as a measure of overall development, given higher scores in this dimension are typically associated with better socioeconomic conditions.

#On the other hand, Dimension 2 has low scores in Education, Safety, and Civic Participation but high scores in Employment, Environment, and Community. This shows that Dimension 2 may be a combination of socioeconomic difficulties, with Education, Safety, and Civic Participation influencing well-being adversely and Employment, Environment, and Community influencing well-being positively.

#Intermediate data frame

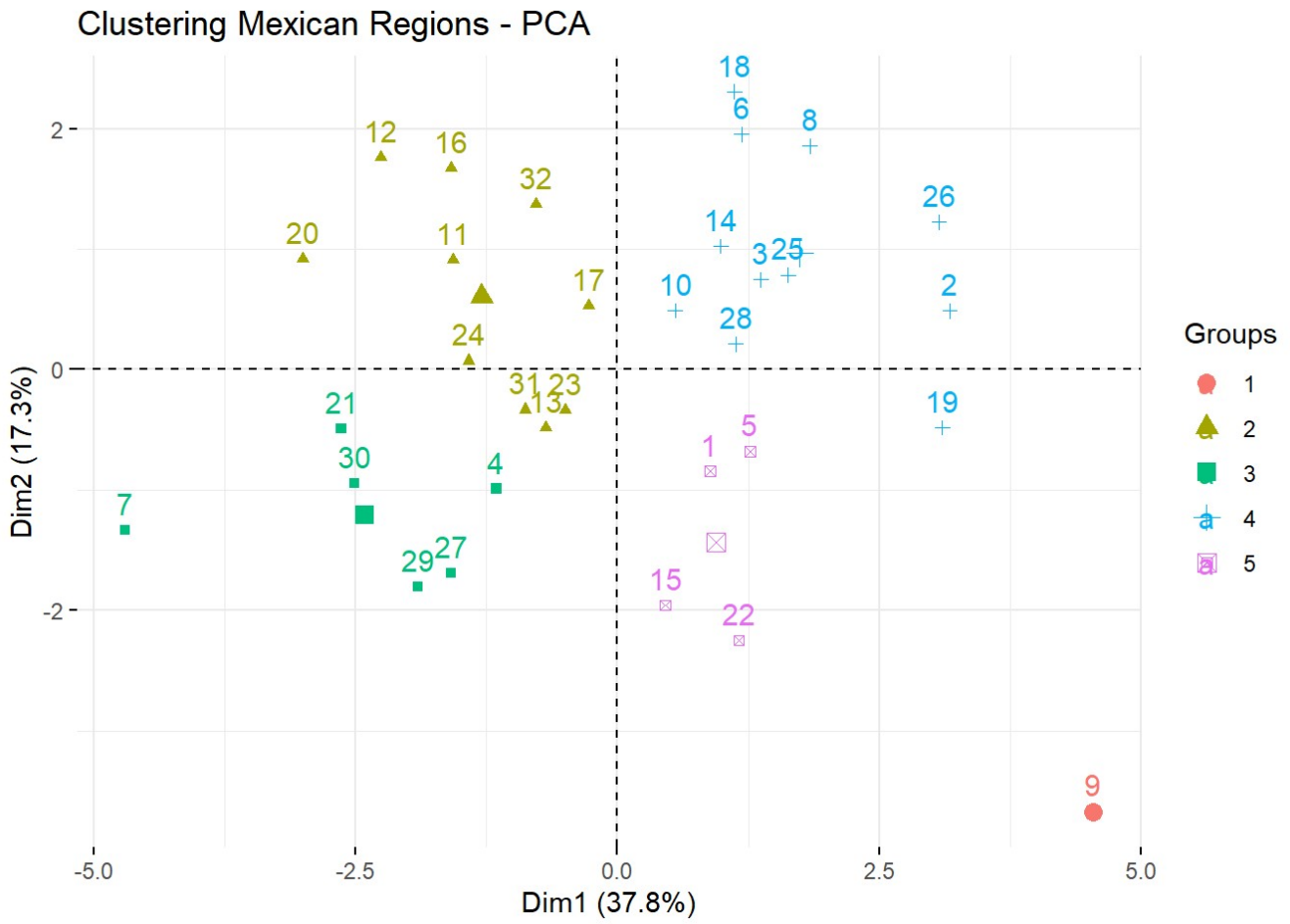
```
#Let's set the seed to 1234 for reproducibility
set.seed(1234)

#Creating an intermediate data frame (including PCA1 and PCA2)
mexico_comps <- tibble(pca_1 = mexico_pca$ind$coord[ ,1],
                       pca_2 = mexico_pca$ind$coord[ ,2]
                      )
```

#Visualizing the Clusters

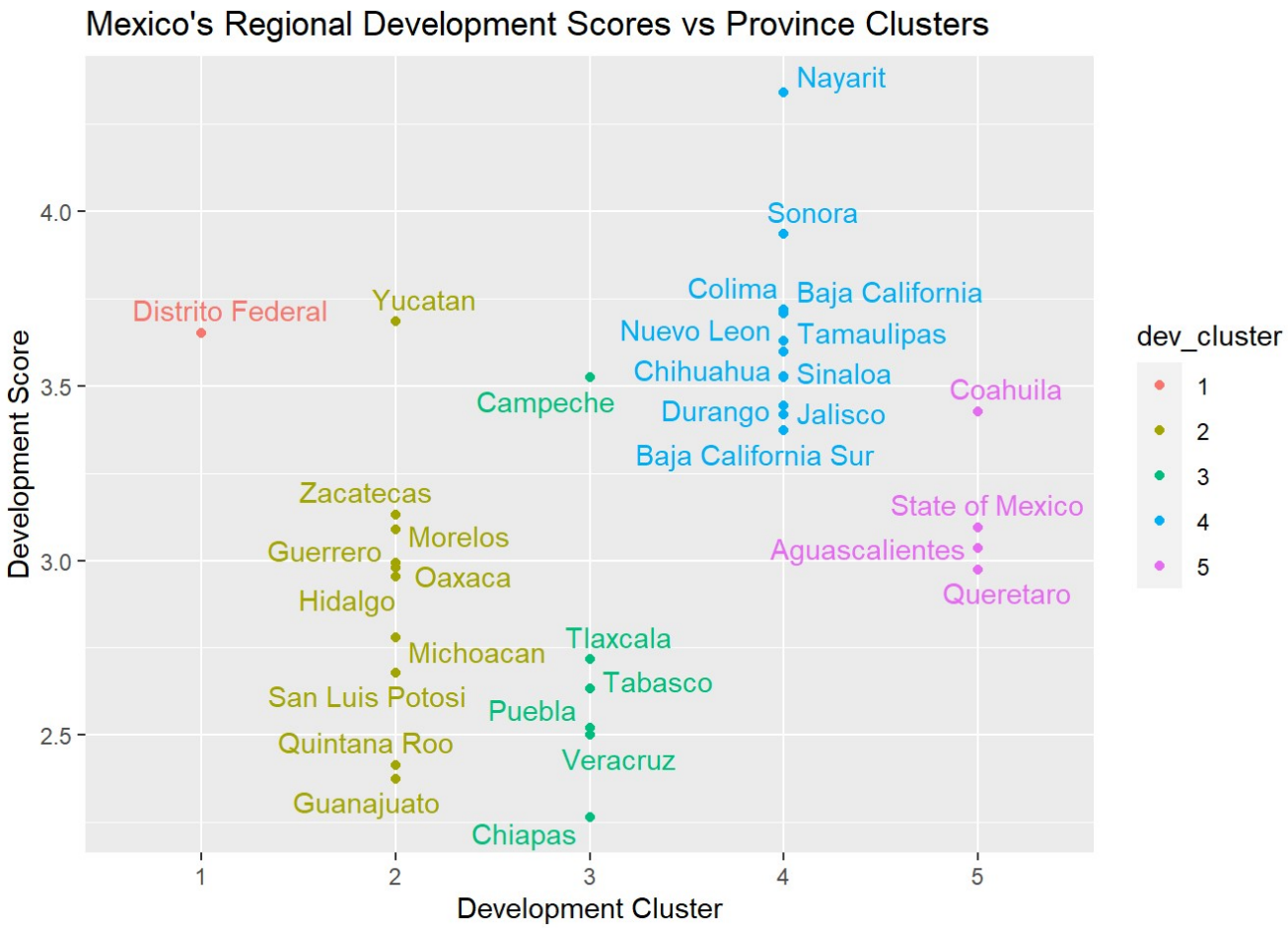
#Following the clustering for each Mexican province, it'd be useful to plot the regions according to their principal components' coordinates, and colour them by the clusters.

#First, we will assign development scores to a new matrix and apply k-means clustering on the development scores. We will then add a cluster column to the Mexican data and create a scatterplot of development scores vs. cluster, colored by cluster. I will divide the regions into five clusters

```
#Assigning development scores to a new matrix
dev_matrix <- as.matrix(mexico_data$development_model)

#Setting seed for reproducibility
set.seed(1234)

#Applying k-means clustering on development scores ==> 5 Clusters
mexico_km <- kmeans(mexico_comps, centers = 5, nstart = 20, iter.max = 50)

#Converting assigned clusters to factor
dev_clusters_as_factor <- factor(mexico_km$cluster)

#Adding the 'cluster' column to the Mexican data
mexico_data_with_wb_clusters <- mexico_data %>%
  mutate(dev_cluster = dev_clusters_as_factor)

#Regions colored by cluster
fviz_pca_ind(mexico_pca,
             title = "Clustering Mexican Regions - PCA",
             habillage = dev_clusters_as_factor)
```




```
#Making a scatterplot of development scores vs. cluster, coloured by cluster
ggplot(mexico_data_with_wb_clusters, aes(y = development_model, x = dev_cluster, color = d
ev_cluster)) +
  geom_point() +
  geom_text_repel(aes(label = Provinces), show.legend = FALSE) +
  labs(x = "Development Cluster", y = "Development Score") +
  ggtitle("Mexico's Regional Development Scores vs Province Clusters")
```



#8. DISCUSSION AND RESULTS

#According to the results of the PCA analysis and K-Means Clustering, there are 5 clusters having different development scoreas and weighted values. There are mainly 5 clusters listed below:

#Cluster 1: Distrito Federal #Cluster 2: Guanajuato, Quintana Roo, Hidalgo, Michoacan, Morelos, Oaxaca, Zacatecas, Guerrero, San Luis Potosi, Yucatan #Cluster 3: Tabasco, Puebla, Veracruz, Chiapas, Campeche, Tlaxcala #Cluster 4: Baja California, Durango, Sonora, Chihuahua, Jalisco, Sinaloa, Baja California Sur, Nuevo Leon, Tamaulipas, Nayarit, Colima #Cluster 5: State of Mexico, Aguascalientes, Queretaro, Coahuila

#Key Insights

#Cluster 1, which solely includes the Distrito Federal, demonstrates distinct development features. This region stands apart from the rest of Mexico, suggesting that the Distrito Federal may experience higher levels of socio-economic growth compared to other regions.

#Cluster 2, characterized by the most negative values in Dimension 2, indicates that these areas may struggle with education, health, income and life satisfaction. Policymakers must concentrate on enhancing these aspects in the regions within Cluster 2.

#Cluster 2 and 3 are in need of development interventions aiming at creating livelihoods opportunities. The 10 regions with the lowest income are listed below, including their clusters:

##Chiapas (Cluster 3), Guerrero (Cluster 2), Tlaxcala (Cluster 2), Puebla (Cluster 3), Oaxaca (Cluster 2), Veracruz (Cluster 3), Hidalgo (Cluster 2, Zacatecas (Cluster 2), Tabasco (Cluster 3) and Morelos (Cluster 2)

#Cluster 3, displaying the highest positive values in Dimension 1, signifies that these regions grapple with low-income generating opportunities, health, environment, housing and life satisfaction while enjoying a high civic engagement rate. This cluster is clearly in need of development support.

#Cluster 4 enjoys high education, income, health, environment, housing and life satisfaction rates while it needs an intervention in safety and civic engagement.

#Finally, Cluster 5, exhibiting relatively small absolute values in both dimensions, can be considered a cluster with average or mixed development levels. For these regions, pinpointing specific areas of improvement and devising tailored policies to tackle their unique development challenges is of utmost importance.

#In summary, recognizing the distinct development characteristics of each cluster will enable policymakers to devise more effective strategies and allocate resources more efficiently to cater to the specific requirements of each area.

#It is also clear that the most members of the Cluster 3 have big populations (Please see the 'population' section), and policymakers should consider their population and population density while developing policies. Besides, based on income levels, Cluster 3 also has regions having low income levels (Please see the 'population' section)

#For future use, we can assign new names for clusters to make regional decisions to alleviate poverty and better address regional disparities, as seen below: #Cluster 1 = MXC1, Cluster 2 = MXC2, Cluster 3 = MXC3, Cluster 4 = MXC4, Cluster 5 = MXC5

#Long story short, MXC2 and MXC3 are the first two regional development group/cluster to focus on the policymakers based on the OECD regional economic, social and well-being data.

#Considering the Cluster 3, our results are in line with the Human Development Index released in 2021 by the UNDP; #UNDP Human Development Index-Mexican Regions Having the Lowest Values in 2021 (Global Data Lab)
#28 = Veracruz 0.723 #29 = Puebla 0.721 #30 = Guerrero 0.694 #31 = Oaxaca 0.689 #32 = Chiapas 0.677

#REFERENCES

#Global Data Lab (2023) Subnational Human Development Index, Mexico, Available at: https://globaldatalab.org/shdi/table/shdi/MEX/?levels=1+4&years=2021&interpolation=0&extrapolation=0 (https://globaldatalab.org/shdi/table/shdi/MEX/?levels=1+4&years=2021&interpolation=0&extrapolation=0)

#OECD (2018) OECD Regional Well-Being: A user's guide, Available at: https://www.oecdregionalwellbeing.org///assets/downloads/Regional-Well-Being-User-Guide.pdf (https://www.oecdregionalwellbeing.org///assets/downloads/Regional-Well-Being-User-Guide.pdf)