LABORATORIO I:

CONCEPTOS DE ESTADÍSTICA DESCRIPTIVA

# Introducción

Esta práctica inicial de laboratorio tiene como objetivo realizar una primera aproximación al Lenguaje R, utilizando el enfoque de análisis exploratorio de datos sobre un dataset, a efectos de repasar conceptos fundamentales de estadística descriptiva.

**Estadística descriptiva**

A partir del dataset *MPI\_subnational.csv[[1]](#footnote-1)* (Multidimensional Poverty Measures), se solicita trabajar sobre las siguientes consignas.

# Exploración de datos

Explore y explique en que consiste el dataset utilizando herramientas de exploración de datos.

## Características de los atributos

El dataset está compuesto por 984 observaciones y 8 variables:

* **ISO.country.code**: Código estándar, de tres letras, del país
* **Country**: País
* **Sub.national.region**: Provincia/Estado/Región
* **World.region**: Subregión a nivel mundial, que no necesariamente coincide con los países o los continentes. Las distintas regiones son:
  + Arab States
  + East Asia and the Pacific
  + Europe and Central Asia
  + Latin America and Caribbean
  + South Asia
  + Sub-Saharan Africa
* **MPI.National**: Índice multidimensional de pobreza del país
* **MPI.Regional**: Índice multidimensional de pobreza de la región
* **Headcount.Ratio.Regional**: Porcentaje de la población considerada pobre en la región
* **Intensity.of.deprivation.Regional**: Intensidad de privaciones. Distancia promedio de aquellos que se encuentran debajo de la línea de pobreza.

En R podemos ver qué tipo de dato tiene cada uno, y en caso de aquellos que son de tipo Factor, cuantos levels tienen.

|  |  |  |
| --- | --- | --- |
| ISO.country.code | Factor | 78 levels |
| Country | Factor | 78 levels |
| Sub.national.region | Factor | 928 levels |
| World.region | Factor | 6 leves |
| MPI.National | Num |  |
| MPI.Regional | Num |  |
| Headcount.Ratio.Regional | Num |  |
| Intensity.of.deprivation.Regional | Num |  |

### Estadísticos básicos de variables numéricas

MPI.National MPI.Regional Headcount.Ratio.Regional Intensity.of.

deprivation.

Regional

Min. :0.0060 Min. :0.0000 Min. : 0.00 Min. :33.30

1st Qu.:0.0660 1st Qu.:0.0530 1st Qu.:12.47 1st Qu.:41.40

Median :0.1740 Median :0.1550 Median :33.95 Median :45.60

Mean :0.2041 Mean :0.2113 Mean :40.18 Mean :47.18

3rd Qu.:0.3030 3rd Qu.:0.3415 3rd Qu.:66.72 3rd Qu.:51.90

Max. :0.6050 Max. :0.7440 Max. :99.00 Max. :75.90

NA's :1

### Registro nulo

El dataset tiene datos muy prolijos. Hay un solo null, que corresponde a Intensity.of.deprivation.Regional para la región de Callao en Perú.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **ISO.**  **Country.**  **Code** | **Country** | **Sub.**  **National.region** | **World.region** | **MPI.**  **National** | **MPI.**  **Regional** | **Headcount.**  **Ratio.**  **Regional** | **Intensity.of.**  **deprivation.**  **Regional** |
| PER | Peru | Callao | Latin America and Caribbean | 0.043 | 0 | 0 | NA |

## Cantidad de ciudades por región

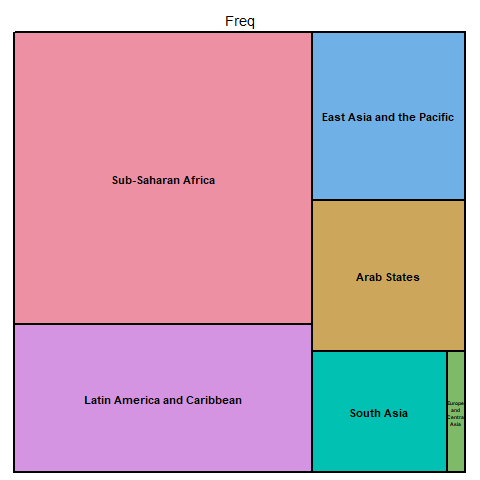
Represente gráficamente la cantidad de ciudades agrupados por Región.

Cantidad de ciudades por región:

* Arab States: 115
* East Asia and the Pacific: 128
* Europe and Central Asia: 11
* Latin America and Caribbean: 218
* South Asia: 81
* Sub-Saharan Africa: 431

Podemos ver que la mayor cantidad de registros se concentran en África subsahariana. Mientras que Europa y Asia central es la región con menos información disponible. Inclusive indagando aún más sobre este último, vemos que estos datos sólo corresponden a dos países:

* Tajikistan
* Uzbekistan

El siguiente gráfico muestra, de una forma más intuitiva, la proporción de ciudades correspondientes a cada región.

# Medidas de posición

Calcule las medidas de posición para los atributos numéricos y agrupe los cálculos de acuerdo a la Región.

1. Ordene los resultados del MPI resultante y concluya al respecto. Help(order).

## Medianas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **World.region** | **MPI.**  **National** | **MPI.**  **Regional** | **Headcount.**  **Ratio.**  **Regional** | **Intensity.of.**  **deprivation.**  **Regional** |
| Europe and Central Asia | 0.008 | 0.0160 | 4.10 | 36.0 |
| Latin America and Caribbean | 0.034 | 0.0375 | 8.95 | 41.4 |
| Arab States | 0.045 | 0.0520 | 12.50 | 40.6 |
| East Asia and the Pacific | 0.100 | 0.0995 | 23.00 | 44.7 |
| South Asia | 0.196 | 0.1990 | 45.20 | 46.6 |
| Sub-Saharan Africa | 0.307 | 0.3330 | 66.60 | 50.4 |

## Medias

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **World.region** | **MPI.**  **National** | **MPI.**  **Regional** | **Headcount.**  **Ratio.**  **Regional** | **Intensity.of.**  **deprivation.**  **Regional** |
| Europe and Central Asia | 0.029 | 0.025 | 6.43 | 37.08 |
| Latin America and Caribbean | 0.054 | 0.064 | 14.23 | 41.53 |
| Arab States | 0.110 | 0.115 | 23.10 | 42.65 |
| East Asia and the Pacific | 0.124 | 0.136 | 28.29 | 45.60 |
| South Asia | 0.209 | 0.219 | 44.11 | 47.63 |
| Sub-Saharan Africa | 0.332 | 0.337 | 61.52 | 51.87 |

## Modas

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **World.region** | **MPI.**  **National** | **MPI.**  **Regional** | **Headcount.**  **Ratio.**  **Regional** | **Intensity.of.**  **deprivation.**  **Regional** |
| Europe and Central Asia | 0.008 | 0.021 | 6.1 | 34.5 |
| Arab States | 0.014 | 0.006 | 2.2 | 40.6 |
| East Asia and the Pacific | 0.066 | 0.053 | 13.4 | 44.3 |
| Latin America and Caribbean | 0.072 | 0.030 | 1.4 | 41.6 |
| South Asia | 0.295 | 0.294 | 65.1 | 50.3 |
| Sub-Saharan Africa | 0.303 | 0.292 | 72.0 | 47.3 |

Tal vez esto se puede graficar con paralelos.

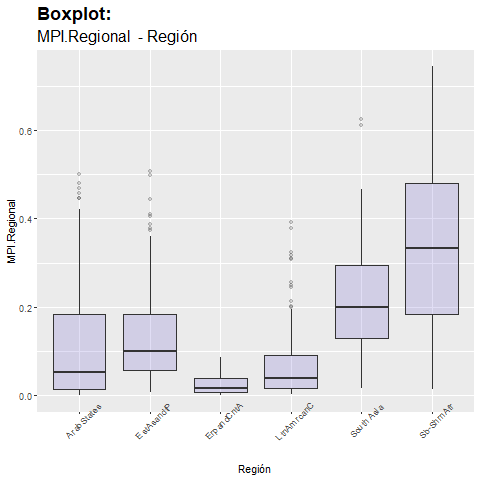
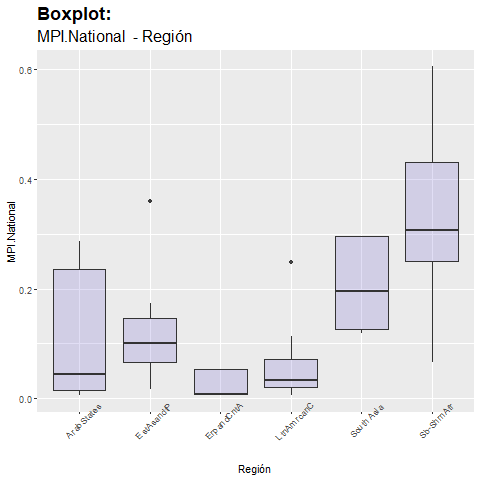
* 1. Grafique las variables y observe su comportamiento (graph: barplot, pie & hist).

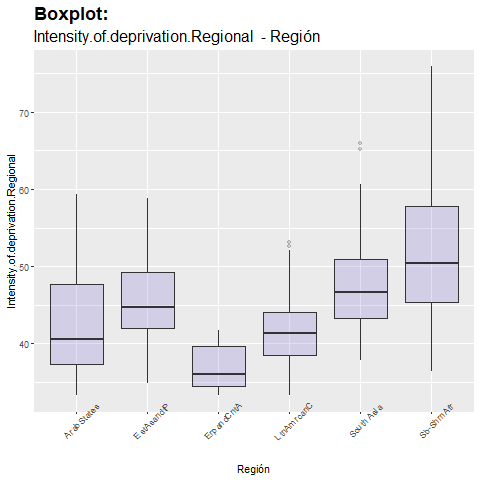
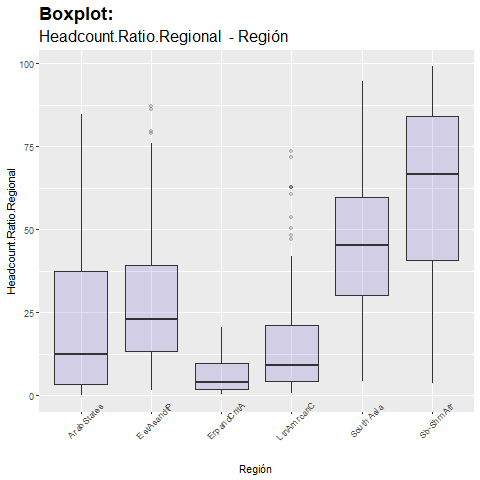
Histograma con media, mediana y moda.

# Medidas de dispersión

Calcular el desvío estándar, la varianza y el rango para cada una de las variables.

* 1. Realice diagramas de cajas y scatterplot’s. Documente las conclusiones.

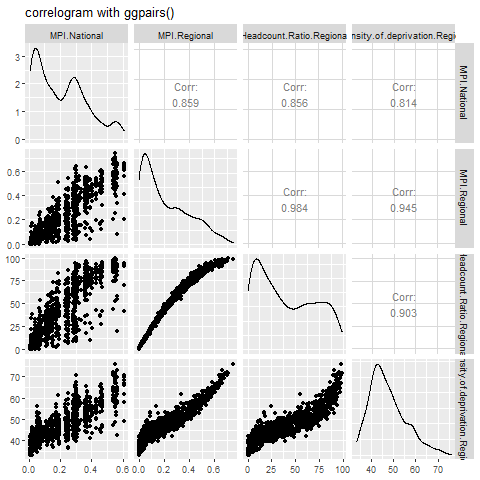




* 1. ¿Qué variable es la que presenta mayor dispersión? Tenga en cuenta que cada variable puede estar expresada en diferentes unidades y magnitudes.

# Medidas de asociación

Calcular el coeficiente de correlación de todas las variables y explique el resultado. ¿Qué tipo de gráficos describen mejor esta relación entre las variables?



Referencias sugeridas:

García, S., Luengo, J., & Herrera, F. (2016). Data preprocessing in data mining. Springer.

Han, J., Pei, J., & Kamber, M. (2011). Data mining: concepts and techniques. Elsevier.

An Introduction to R: <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>

1. Disponible en: <https://www.kaggle.com/ophi/mpi/data> [↑](#footnote-ref-1)