République du Sénégal

Un peuple - Un But - Une fois

****

Ministère de l'Economie, du Plan et de la Coopération

****

Agence nationale de la Statistique et de la Démographique

****

Ecole nationale de la Statistique et de l'Analyse économique - Pierre NDIAYE

****

TP3 - Logiciel statistique R

Table des matières

I. Importation et pré-traitements des données

I.1. Importation

Afin d'assurer un travail de qualité, il est essentiel d'importer les données nécessaires à notre étude. Pour ce faire, nous utiliserons la librairie readr qui offre la fonction read_csv, conçue spécifiquement pour lire les bases de données au format CSV. Pour garantir une manipulation fluide des données par la suite, nous remplacerons les valeurs manquantes (NA) par des valeurs vides. Cette option est spécifiée lors de l'importation des données avec la fonction read_csv(). Les trois dataframes relatives aux données sur le taux de croissance annuel du revenu national par habitant, l'indice d'inégalité entre les sexes et le taux d'accroissement annuel de la population sont affectés respectivement aux objets dfGrowth, dfGindex, dfPopGrowth. Pour avoir un aperçu sur les dataframes, nous pouvons utiliser la fonction head() pour afficher les premières lignes et tail() pour les dernières lignes. Nous nous limiterons à l'affichage des 5 premières et dernières lignes de chacune des bases.

```
library(readr)

dfGrowth <- read_csv("..//Donnees//annual-growth-in-gni-per-capita.csv")
# tail(dfGrowth, 5)
head(dfGrowth, 5)
```

Region Name	Region Alpha-3 Code	Region Alpha-2 Code	Start Year	End Year	Value	Unit	Source
Albania	ALB	AL	1998	1998	10.151616	%	World Bank
Albania	ALB	AL	1999	1999	12.983405	%	World Bank
Albania	ALB	AL	2000	2000	8.391977	%	World Bank
Albania	ALB	AL	2001	2001	9.565283	%	World Bank
Albania	ALB	AL	2002	2002	4.141344	%	World Bank

```
dfGindex <- read_csv("..//Donnees//gender-inequality-index.csv")
# tail(dfGindex, 5)
head(dfGindex, 5)
```

Region Name	Region Alpha-3 Code	Region Alpha-2 Code	Start Year	End Year	Value	Unit	Source
Afghanistan	AFG	AF	2005	2005	0.748	NA	UNDP
Afghanistan	AFG	AF	2006	2006	0.749	NA	UNDP
Afghanistan	AFG	AF	2007	2007	0.752	NA	UNDP
Afghanistan	AFG	AF	2008	2008	0.755	NA	UNDP
Afghanistan	AFG	AF	2009	2009	0.755	NA	UNDP

```
dfPopGrowth <- read_csv("..//Donnees//population-growth-annual.csv")
# tail(dfPopGrowth, 5)
head(dfPopGrowth, 5)
```

Region Name	Region Alpha-3 Code	Region Alpha-2 Code	Start Year	End Year	Value	Unit	Source
Afghanistan	AFG	AF	1961	1961	1.898499	%	World Bank
Afghanistan	AFG	AF	1962	1962	1.965805	%	World Bank
Afghanistan	AFG	AF	1963	1963	2.029830	%	World Bank
Afghanistan	AFG	AF	1964	1964	2.090208	%	World Bank
Afghanistan	AFG	AF	1965	1965	2.147639	%	World Bank

I.2. Pré-traitement

Dans cette partie, nous allons de fond en combre explorer les bases de données pour voir la nécissité de faire des traitements avant de passer aux traçages des graphiques.\ En affichant certaines observations, nous remarquons que les données sont disposées sous le même format. Elle comporte 8 variables dont les lables sont décrites dans le tableau ci-après:

Nom de la variable	**Description**	**Exemples**
Region Name	Nom du pays	Albania,Afghanistan,Algeria...
Region Alpha-3 Code	Code à 3 lettres du pays	ALB,AFG,DZA...
Region Alpha-2 Code	Code à 2 lettres du pays	

Nom de la variable	Description	Exemples
Start Year (End Year)	Année de début (année de fin)	2002, 2007
Value	Valeur de l'indicateur	0.297, 12.7854
Unit	Unité de mesure de l'indicateur	en %
Source	Source de la donnée	UNDP, World Bank...

Les données sur le taux de croissance annuel du revenu national par habitant se présentent sous la forme de 5404 enregistrements distincts. Chaque enregistrement correspond à un pays spécifique pour une année particulière. Cette répétition des pays sur plusieurs lignes est inhérente à la nature temporelle des données, qui couvrent la période allant de 1961 à 2021. Ces données proviennent de la Banque mondiale et fournissent le taux de croissance annuel du revenu national par habitant pour chaque pays et chaque année.

```
summary(dfGrowth)
```

```
##  Region Name        Region Alpha-3 Code Region Alpha-2 Code   Start Year
##  Length:5404        Length:5404         Length:5404         Min.   :1961
##  Class :character   Class :character    Class :character    1st Qu.:1989
##  Mode  :character   Mode  :character    Mode  :character    Median :2003
##                                                             Mean   :2000
##                                                             3rd Qu.:2012
##                                                             Max.   :2021
##     End Year        Value              Unit              Source
##  Min.   :1961   Min.   :-47.3916   Length:5404        Length:5404
##  1st Qu.:1989   1st Qu.: -0.4354   Class :character   Class :character
##  Median :2003   Median :  2.0808   Mode  :character   Mode  :character
##  Mean   :2000   Mean   :  1.8405
##  3rd Qu.:2012   3rd Qu.:  4.4471
##  Max.   :2021   Max.   : 45.9738
```

En ce qui concerne les données relatives à l'indice d'inégalité entre les sexes, elles comportent 4889 observations et sont présentées da la même manière que la base ci-dessus. Elles proviennent essentiellement de l'

```
summary(dfGindex)
```

```
##  Region Name        Region Alpha-3 Code Region Alpha-2 Code   Start Year
##  Length:4889        Length:4889         Length:4889         Min.   :1990
##  Class :character   Class :character    Class :character    1st Qu.:1999
```

```
##  Mode  :character   Mode  :character    Mode  :character    Median :2007
##                                                            Mean   :2006
##                                                            3rd Qu.:2014
##                                                            Max.   :2021
##     End Year        Value          Unit             Source
##  Min.   :1990   Min.   :0.0130   Mode:logical   Length:4889
##  1st Qu.:1999   1st Qu.:0.2430   NA's:4889      Class :character
##  Median :2007   Median :0.4360                  Mode  :character
##  Mean   :2006   Mean   :0.4134
##  3rd Qu.:2014   3rd Qu.:0.5790
##  Max.   :2021   Max.   :0.8220
```

La dernière base de données renseigne sur les taux de croissance de tous les pays de la période 1961 - 2021. Elle est constitué de 13070 observations et est présentée de manière semblable aux deux autres.

```
summary(dfPopGrowth)
```

```
##  Region Name        Region Alpha-3 Code Region Alpha-2 Code   Start Year
##  Length:13070       Length:13070        Length:13070        Min.   :1961
##  Class :character   Class :character    Class :character    1st Qu.:1976
##  Mode  :character   Mode  :character    Mode  :character    Median :1991
##                                                            Mean   :1991
##                                                            3rd Qu.:2006
##                                                            Max.   :2021
##     End Year        Value               Unit               Source
##  Min.   :1961   Min.   :-10.9551   Length:13070       Length:13070
##  1st Qu.:1976   1st Qu.:  0.6943   Class :character   Class :character
##  Median :1991   Median :  1.6771   Mode  :character   Mode  :character
##  Mean   :1991   Mean   :  1.7709
##  3rd Qu.:2006   3rd Qu.:  2.6578
##  Max.   :2021   Max.   : 28.0410
```

```
# library(forcats)
# fct_unique(dfGrowth[[1]])
sum(duplicated(dfGrowth))
```

```
## [1] 0
```

```
update = function(data){
  # Librarie
  library(dplyr)
```

```
  # renommer les colonnes en termes plus simples
  data = data |> dplyr::rename(region = `Region Name`,
                      C3region = `Region Alpha-3 Code`,
                      year = `End Year`)

  # Sélection des colonnes nécessaire
  data = data |> dplyr::select(c("region","C3region","year","Value"))

  attach(data)
  # Trier selon l'année et la région
  df = data[order(year,region),]

  return(df)
}

dfGrowth = update(dfGrowth)
```

```
## Cette Fonction permet d'enregistrer la base avec les variables du tracé
rebase = function(df){
  # Librairie
  library(tidyverse)

  # Transposer la base en format temporelle
  df <- df %>% pivot_wider(
                            names_from = "year",
                            values_from = "Value")

  # Index des lignes
  df  <- df %>% remove_rownames %>%
    column_to_rownames(var="C3region")

  # Base de données temporelle ordonnées
  new_df = df[,order(names(df))]

  return(df)
}

## Application
dfGrowth = rebase(dfGrowth)
```

```r
## Pays de l'Afrique de l'Ouest
WAfrica = c("Benin","Burkina Faso","Cabo Verde","Côte d'Ivoire","Gambia",
            "Guinea","Guinea-Bissau","Mali","Mauritania","Niger","Nigeria",
            "Senegal","Sierra Leone","Togo")
```

I.3 Création de nouvelles observations

```r
# Déterminer les valeurs des lignes du monde et de l'Afrique de
# l'ouest dans la base de données par lecalcul des moyennes
lineWorld = colMeans(dfGrowth[-1], na.rm = T)
lineWAfrica = colMeans((dfGrowth |> filter(region %in% WAfrica))[-1],
                       na.rm = T)

# Dimension de la base de données
n = dim(dfGrowth)[1]
p = dim(dfGrowth)[2]

# Ajout de la ligne World
dfGrowth[n+1,1] = "World"
dfGrowth[n+1,2:p] = lineWorld

# Ajout de la ligne Western Africa
dfGrowth[n+2,1] = "Western Africa"
dfGrowth[n+2,2:p] = lineWAfrica

# Définir les pays pour les graphes
countries = c("Niger", "Western Africa", "World")

# Filtrer selon les pays définis ci-dessus
dfGrowth = dfGrowth %>% filter(region%in%countries)
```

```r
## Librairie
library(tidyr)

## Transposer pour avoir le format long
dfGrowth = gather(dfGrowth,key = "year", value = "Value", -region)

## Convertir la variable year en numérique
dfGrowth$year = as.numeric(dfGrowth$year)
```

```r
## Trier selon les années
dfGrowth <- dfGrowth[order(dfGrowth$year), ]

## Affichage des premières masques
head(dfGrowth)
```

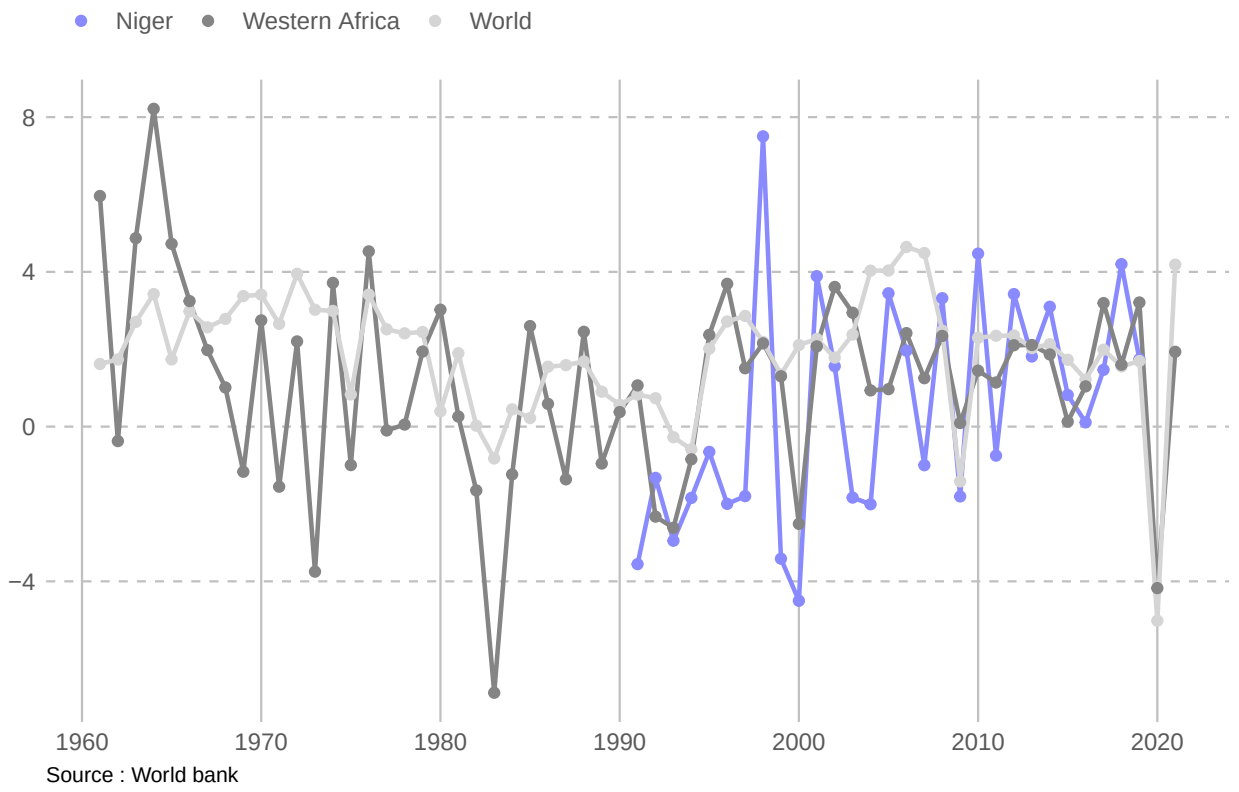
region	year	Value
Niger	1961	NA
World	1961	1.6189927
Western Africa	1961	5.9612408
Niger	1962	NA
World	1962	1.7365680
Western Africa	1962	-0.3742891

II. Graphiques

```r
library(ggplot2)
library(ggthemes)

ggplot(dfGrowth, mapping = aes(x = year, y = Value, col = region)) +
  # Ajout de la courbe et des points
  geom_line(size = 0.8, show.legend = F) + theme_bw() +
  geom_point() +
  scale_x_continuous(n.breaks = 6) +
  labs(
    title = "Figure : Income growth and distribution (Gini Index)",
    x = "", y = "", caption = "Source : World bank", colour = "",) +
  # Ajout d'un thème format excel
  theme_excel_new() +
  # Changer les couleurs
  scale_color_manual(values = c("#8A8AFF","#858585","#D5D5D5")) +
  theme(
    panel.grid.major.y = element_line(linetype = 'dashed',
                                      size = 0.4, color = "#C0C0C0"),
    panel.grid.major.x = element_line(linetype = 'solid',
                                      size = 0.4, color = "#C0C0C0"),
    plot.title = element_text(hjust=0, size = 10,
                              color = "#0F4761", face = "italic"),
```

```
    plot.caption = element_text(hjust=0, size = 8, color = "black"),
    legend.position = 'top', legend.justification = 'left',
  )
```

Figure : Income growth and distribution (Gini Index)



Source : World bank

Le traçage des deux autres graphiques se fait de manière analogue. Nous créons donc une fonction gérant à la fois le traitement ainsi que le tracé.

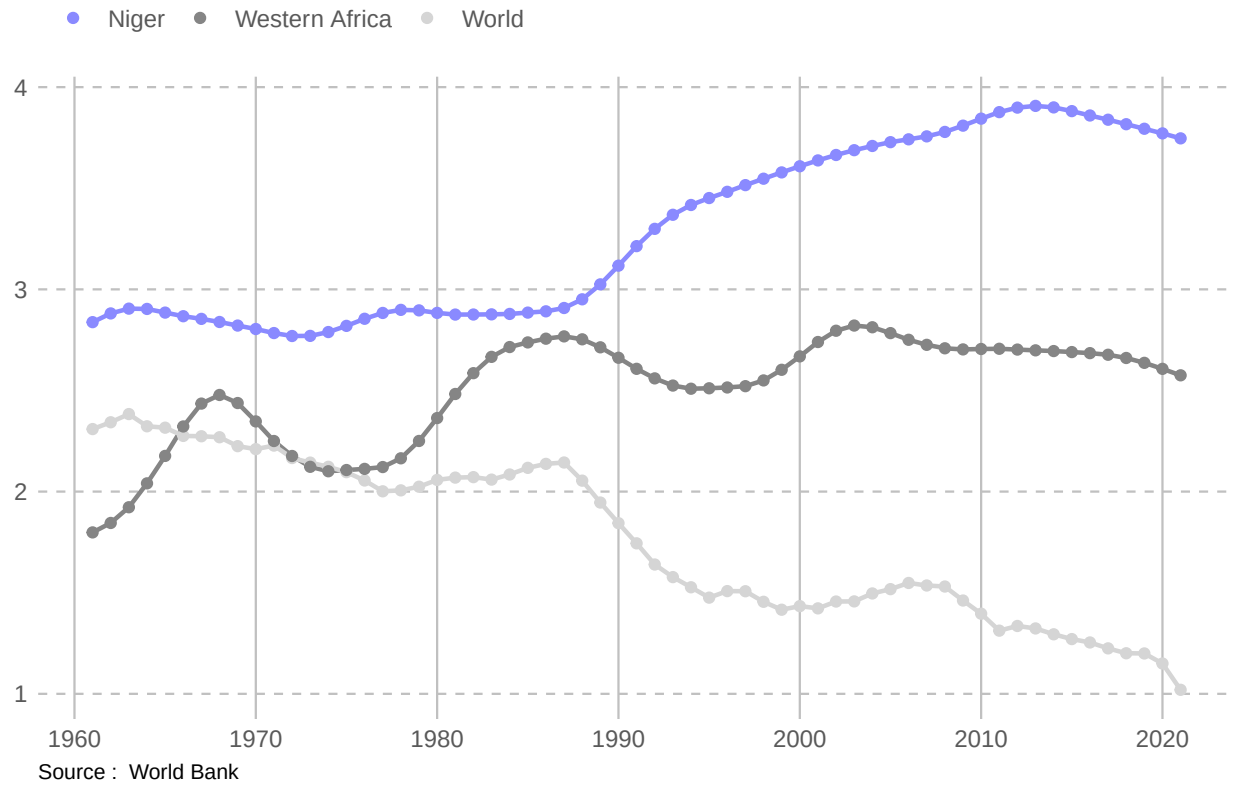
```
TP2_graph <- function(df, titre, source, display_box = F){

  ## Librairies
  library(tidyverse)
  library(tidyr)
  library(ggplot2)
  library(ggthemes)

  ## Pays de l'Afrique de l'Ouest
  WAfrica = c("Benin","Burkina Faso","Cabo Verde","Côte d'Ivoire","Gambia",
          "Guinea","Guinea-Bissau","Mali","Mauritania","Niger","Nigeria",
          "Senegal","Sierra Leone","Togo")

  ## Zones géographiques pour le tracé des graphes
```

```r
countries = c("Niger", "Western Africa", "World")

## Transformer la base avec la fonction update
df = update(df)

## Reconstruire la base
df = rebase(df)

## Ajout des statistiques moyennes dans le monde
lineWorld = colMeans(df[-1], na.rm = T)
df[dim(df)[1]+1,1] = "World"
df[dim(df)[1],2:dim(df)[2]] = lineWorld

## Ajout des statistiques moyennes dans l'Afrique de l'Ouest
lineWAfrica = colMeans((df |> filter(region %in% WAfrica))[-1], na.rm = T)
df[dim(df)[1]+1,1] = "Western Africa"
df[dim(df)[1],2:dim(df)[2]] = lineWAfrica

## Filtrer les zones pour les courbes
new_df = df %>% filter(region%in%countries)

## Transformation de la base en format long
new_df = gather(new_df,key = "year", value = "Value", -region)

## Convertir la variable année en numérique
new_df$year = as.numeric(new_df$year)

## Trier selon l'année
new_df <- new_df[order(new_df$year), ]

## Courbes
plot <- ggplot(new_df, mapping = aes(x = year, y = Value, col = region)) +
geom_line(size = 0.8, show.legend = F) + theme_bw() + geom_point() +
scale_x_continuous(n.breaks = 6) +
labs(
  title = paste("Figure : ",titre),
  x = "", y = "", caption = paste("Source : ",source), colour = "",
) + theme_excel_new() +
  scale_color_manual(values = c("#8A8AFF","#858585","#D5D5D5")) +
```

```
  theme(
    panel.grid.major.y = element_line(linetype = 'dashed',
                                      size = 0.4, color = "#C0C0C0"),
    panel.grid.major.x = element_line(linetype = 'solid',
                                      size = 0.4, color = "#C0C0C0"),
    plot.title = element_text(hjust=0, size = 10,
                              color = "#0F4761", face = "italic"),
    plot.caption = element_text(hjust=0, size = 8, color = "black"),
    legend.position = 'top', legend.justification = 'left'
  )

  ## Ajout de la bulle narrative
  if (display_box == T){
    label_data <- data.frame(
      x = 2002,
      y = 0.4,
      label = "World: "
    )
    plot = plot +
      geom_rect(aes(xmin = 2002 - 2, xmax = 2002 + 2,
                    ymin = 0.4 + 0.07, ymax = 0.4 + 0.1),
            fill = "#D5D5D5", color = NA, shadow = 2,
            alpha = 0.8, show.legend = F) +
      geom_text(data = label_data,
            aes(label = paste(label,0.4), x = x, y = y+0.085),
            size = 4, colour = "black") +
      annotate("polygon",
         x = c(2002 - 0.3, 2002 + 0.3, 2002),
         y = c(0.4 + 0.07, 0.4 + 0.07, 0.4 + 0.05),
         fill = "#D5D5D5",
         alpha = 0.8)
  }
  return(plot)
}
```

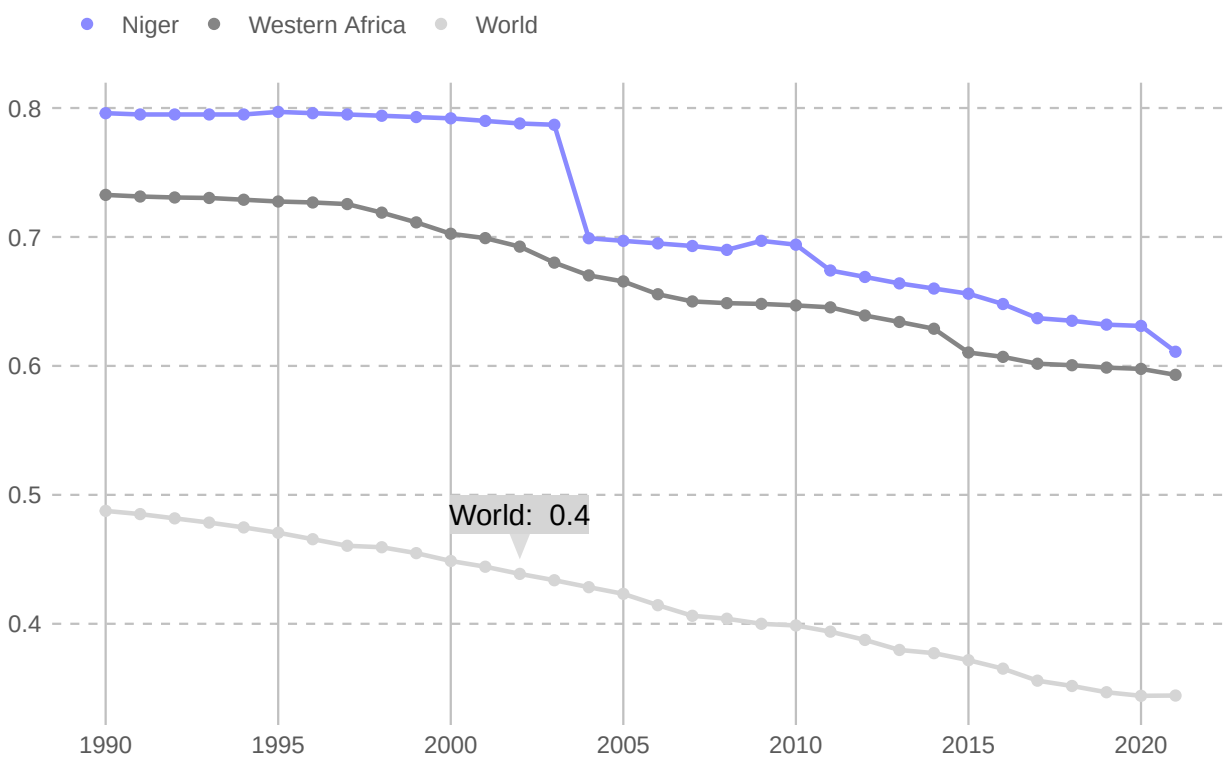
```
TP2_graph(dfPopGrowth,titre="Annual population growth (%)",
          source="World Bank")
```

Figure : Annual population growth (%)




```
TP2_graph(dfGindex,titre="Gender index inequality",
          source="UNDP", display_box = T)
```

Figure : Gender index inequality



Source : UNDP