Análisis Exploratorio de los Datos: CUIDADOS DOMICILIARIOS

Alicia Perdices Guerra

21 de mayo, 2021

Contents

1.ANÁLISIS EXPLORATORIO POR PAISES Se procede en primer lugar a cargar todos los archivos para poder realizar el análisis.

```
cuidados<-read.csv("C:/temp/CuidadosDomiciliarios_clean.csv",sep= ",")
```

1.1.- EN RELACIÓN CON LOS CUIDADOS DOMICILIARIOS

- **1.1.1 Análisis Descriptivo**

Se procede a realizar el análisis descriptivo:

```
summary(cuidados)
```

```
##    ISCED11              GEO                UNIT               TIME
##  Length:396         Length:396         Length:396         Min.   :2014
##  Class :character   Class :character   Class :character   1st Qu.:2014
##  Mode  :character   Mode  :character   Mode  :character   Median :2014
##                                                           Mean   :2014
##                                                           3rd Qu.:2014
##                                                           Max.   :2014
##      SEX                Value        Value_imp
##  Length:396         Min.   : 0.100   Mode :logical
##  Class :character   1st Qu.: 1.475   FALSE:384
##  Mode  :character   Median : 2.550   TRUE :12
##                     Mean   : 3.411
##                     3rd Qu.: 4.600
##                     Max.   :15.900
```

Se filta el dataframe para que la variable GEO aparezcan solo los paises objeto de estudio.(Para cada archivo relacionado con Los Cuidados Domiciliarios y unificamos la información). Además se selecciona la información relevante de la variable SEX (Males, Females) y ISCED11(All ISCED 2011 levels)

```
#Estado de Salud (Años de Vida Sana)
#==================================

cuidados_paises<- filter(cuidados,
              +(GEO!="European Union - 27 countries (from 2020)")&
              +(GEO!="European Union - 28 countries (2013-2020)"))

cuidados_paises<-filter(cuidados_paises,ISCED11=="All ISCED 2011 levels ")

cuidados_males<-filter(cuidados_paises, SEX=="Males")
nrow(cuidados_males)
```

```
## [1] 31
```

```
cuidados_females<-filter(cuidados_paises, SEX=="Females")
nrow(cuidados_females)
```

```
## [1] 31
```

```
head(cuidados_females)
```

```
##                  ISCED11                                                GEO
## 1 All ISCED 2011 levels                                             Belgium
## 2 All ISCED 2011 levels                                            Bulgaria
## 3 All ISCED 2011 levels                                             Czechia
## 4 All ISCED 2011 levels                                             Denmark
## 5 All ISCED 2011 levels  Germany (until 1990 former territory of the FRG)
## 6 All ISCED 2011 levels                                             Estonia
##         UNIT TIME     SEX Value Value_imp
## 1 Percentage 2014 Females  12.0     FALSE
## 2 Percentage 2014 Females   3.3     FALSE
## 3 Percentage 2014 Females   2.6     FALSE
## 4 Percentage 2014 Females   6.2     FALSE
## 5 Percentage 2014 Females   3.7     FALSE
## 6 Percentage 2014 Females   1.4     FALSE
```

```
head(cuidados_males)
```

```
##                  ISCED11                                                GEO
## 1 All ISCED 2011 levels                                             Belgium
## 2 All ISCED 2011 levels                                            Bulgaria
## 3 All ISCED 2011 levels                                             Czechia
## 4 All ISCED 2011 levels                                             Denmark
## 5 All ISCED 2011 levels  Germany (until 1990 former territory of the FRG)
## 6 All ISCED 2011 levels                                             Estonia
##         UNIT TIME   SEX Value Value_imp
## 1 Percentage 2014 Males   7.5     FALSE
```

```
## 2 Percentage 2014 Males   2.6     FALSE
## 3 Percentage 2014 Males   1.6     FALSE
## 4 Percentage 2014 Males   3.6     FALSE
## 5 Percentage 2014 Males   1.5     FALSE
## 6 Percentage 2014 Males   1.1     FALSE
```

Se crea un Dataframe con toda la información:

```
year<-(cuidados_males$TIME)#Columna Year
country<-(cuidados_males$GEO)#Columna Paises


#Dataframe con toda la información relacionada
#con los Cuidados Domiciliarios en 2014 por Países
cuidados_df<-data.frame("TIME"=year,"Pais"=country,
                        "Cuidados_males"=
                          cuidados_males$Value,
                        "Cuidados_females"=
                          cuidados_females$Value)



#Generamos el fichero filtrado para utilizarlo en el siguiente análisis.
write.csv(cuidados_df, file="Cuidados_Domiciliarios_Analisis.csv", row.names = FALSE)
```

Se reescalan los datos:

```
cuidados_df["Cuidados_males_norm"]<-
  rescale(cuidados_df$Cuidados_males, to=c(0,1))
cuidados_df["Cuidados_females_norm"]<-
  rescale(cuidados_df$Cuidados_females, to=c(0,1))
```

- **1.1.2 Visualización y Distribución de la información"**

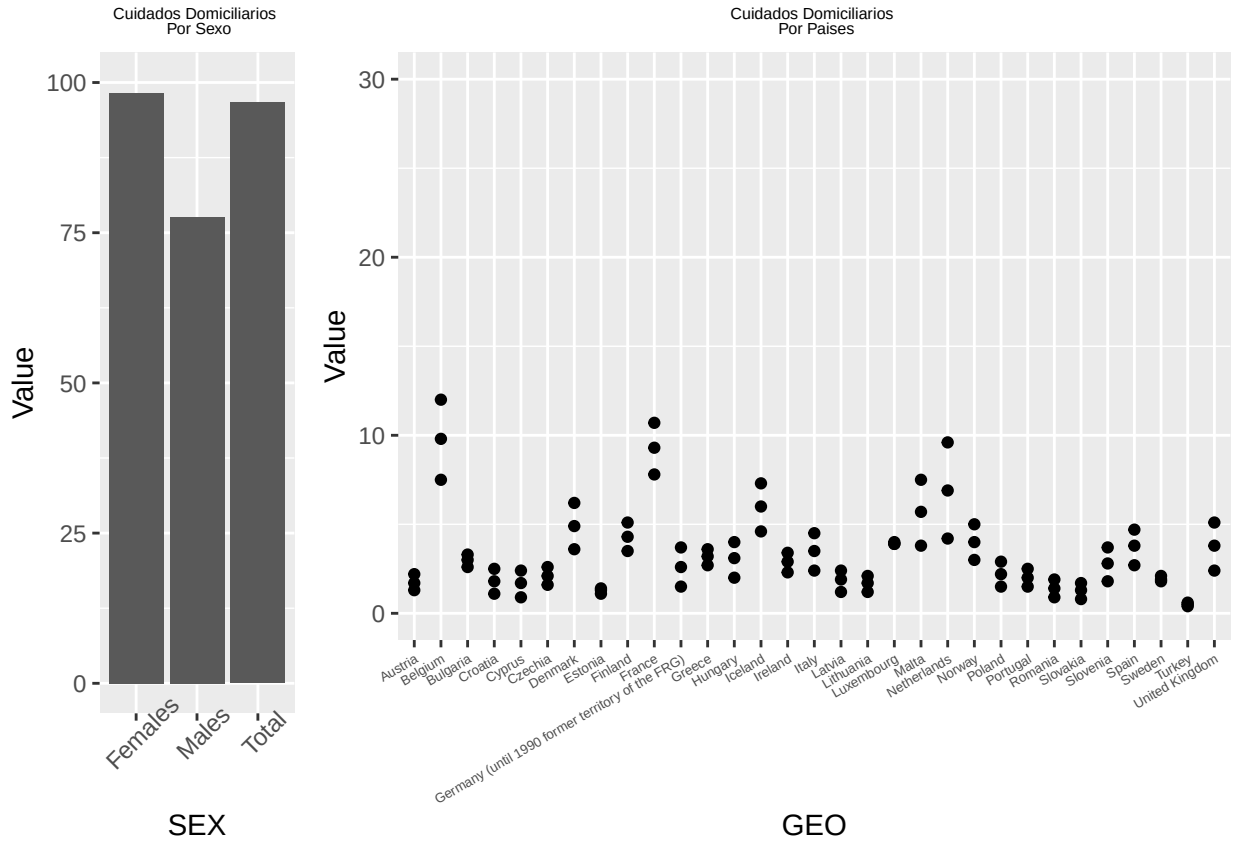
Se visualiza las variable que nos dan información sobre los cuidados domiciliarios realizada en función de TIME, y País.

```
#CUIDADOS DOMICILIARIOS
#================================================================================
#Gráfica de barras de la información sobre los cuidados Domiciliarios por Sexos"
plot1=ggplot(data=cuidados_paises)+
  geom_col(aes(x=SEX,y=Value))+
  theme(axis.text.x = element_text(angle = 45))+
  scale_y_continuous(limit=c(0,100))+
  ggtitle("Cuidados Domiciliarios \n Por Sexo")+
  theme (plot.title = element_text(size=rel(0.5), hjust = 0.5))

#Gráfica de puntos de la información sobre los cuidados Domiciliarios por Sexos"
plot2=ggplot(data=cuidados_paises)+
  geom_point(aes(x=GEO,y=Value))+
  theme(axis.text.x = element_text(size= 5,angle = 30,vjust=1,hjust = 1))+
  scale_y_continuous(limit=c(0,30))+
```

```
  ggtitle("Cuidados Domiciliarios \n Por Paises")+
  theme (plot.title = element_text(size=rel(0.5),hjust=0.5))

grid.arrange(plot1,plot2,widths=c(1,3),  ncol=2)
```



Se obtienen los 5 países con un mayor porcentaje en Cuidados Domiciliarios en 2014.

```
########################################
#Para "Cuidados_males"                 #
########################################

#Se ordena por "Cuidados_males"

cuidados_5paises_2014<-cuidados_males[with(cuidados_males, order(-cuidados_males$Value)),]

#Se crea una tabla para los Cuidados Domiciliarios en Hombres
#de los 5 Paises con un porcentaje más alto.

kable(cuidados_5paises_2014[0:5,c(2,6)], col.names = c("País","CD Hombres"),
      caption = "Países con un mayor porcentaje en cuidados domiciliarios en Hombres")
```

Table 1: Países con un mayor porcentaje en cuidados domiciliarios en Hombres

	País	CD Hombres
10	France	7.8
1	Belgium	7.5
28	Iceland	4.6
19	Netherlands	4.2
16	Luxembourg	3.9

```
########################################
#Para "Cuidados_females"                #
########################################

#Se ordena por "Cuidados_females"

cuidados_5paises_2014<-cuidados_females[with(cuidados_females, order(-cuidados_females$Value)),]

#Se crea una tabla para los Cuidados Domiciliarios en Mujeres
#de los 5 Paises con un porcentaje más alto.(En 2014)

kable(cuidados_5paises_2014[0:5,c(2,6)], col.names = c("País","CD Mujeres"),
      caption = "Países con un mayor porcentaje en cuidados domiciliarios en Mujeres en 2014")
```

Table 2: Países con un mayor porcentaje en cuidados domiciliarios en Mujeres en 2014

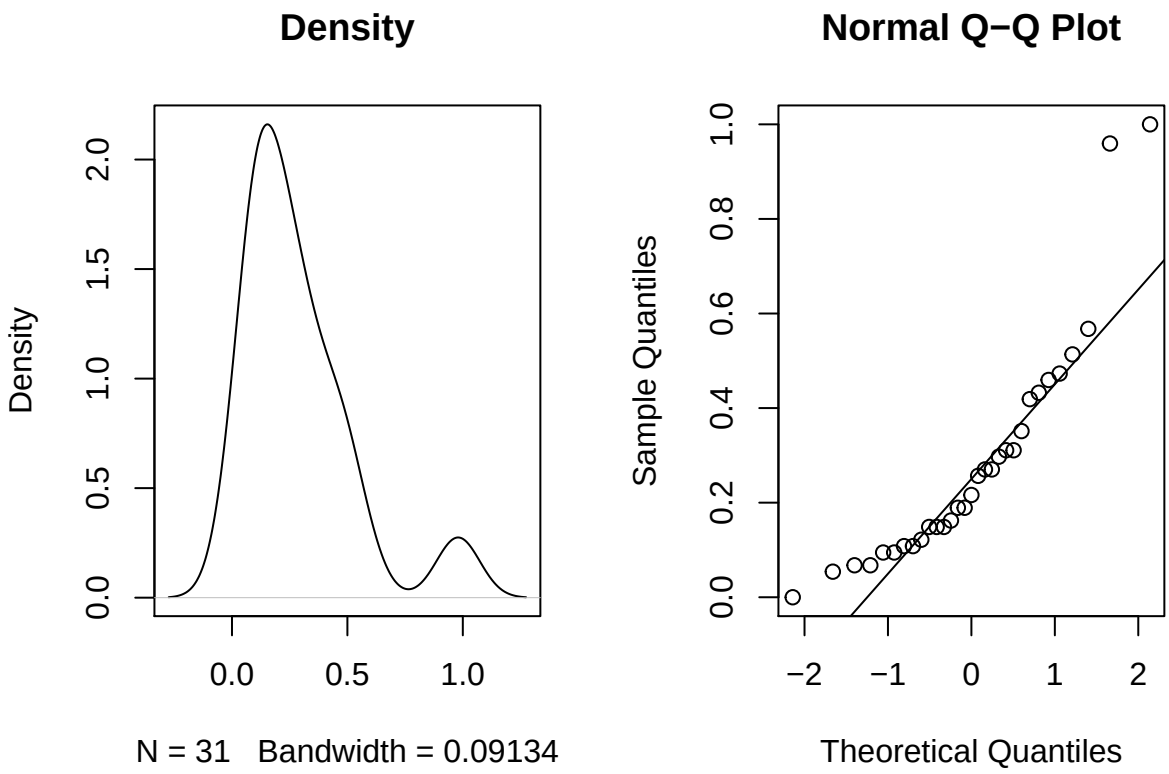
	País	CD Mujeres
1	Belgium	12.0
10	France	10.7
19	Netherlands	9.6
18	Malta	7.5
28	Iceland	7.3

- **1.1.3 Normalidad de la variable "Value (Cuidados_males,Cuidados_females)"**

Se comprueba con métodos visuales si la variable tiene una distribución normal.

Cuidados_males

```
par(mfrow=c(1,2))
plot(density(cuidados_df$Cuidados_males_norm) ,main="Density")
qqnorm(cuidados_df$Cuidados_males_norm)
qqline(cuidados_df$Cuidados_males_norm)
```



Para estudiar si una muestra proviene de una población con distribución normal, se disponen de tres herramientas:

- Histograma o Densidad
- Gráficos cuantil cuantil (QQplot)
- Pruebas de hipótesis.

Si en la prueba de Densidad se observa sesgo hacia uno de los lados de la gráfica, sería indicio de que la muestra no proviene de una población normal.Si por otra parte, sí se observa simetría, **NO** se garantiza que la muestra provenga de una población normal. En estos casos sería necesario utilizar otras herramientas como **QQplot y pruebas de hipótesis**.

En la gráfica Densidad de la variable "Cuidados_males_norm", se observa cierto sesgo hacia la izquierda,por lo que no se considera normalidad. Se puede confirmar observando la gráfica QQplot en la que la línea que grafica qqline sirve de referencia para interpretar el gráfico. Si se tuviese una muestra distribuída normalmente, se esperaría que los puntos del gráfico quantil quantil estuviesen perfectamente alineados con la línea de referencia, y observamos que para este caso, "Cuidados_males_norm" se alinea solo en la parte central.

Para confirmar, se realizan las pruebas de hipótesis:

- $h_0: La muestra proviene de una población normal.
- $h_1: La muestra NO proviene de una población normal.

Se aplica la prueba Shapiro-Wilk:

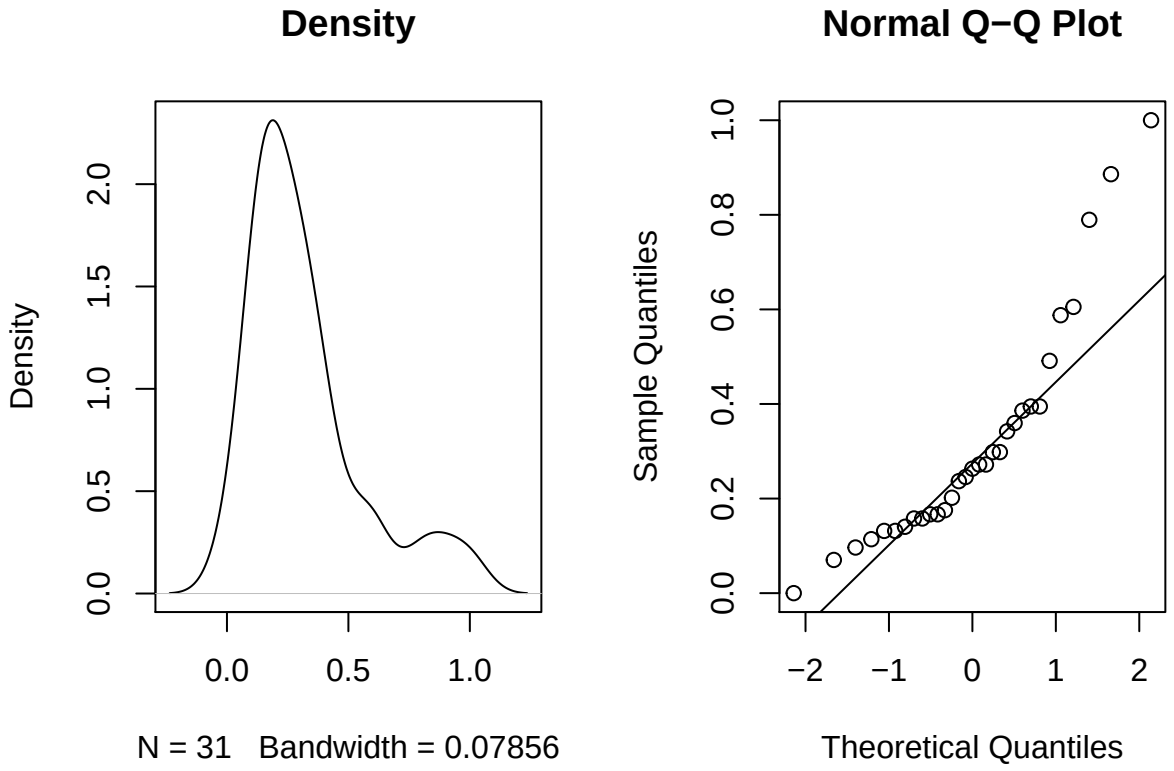
```
shapiro.test(cuidados_df$Cuidados_males_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cuidados_df$Cuidados_males_norm
## W = 0.8354, p-value = 0.0002515
```

Se observa un p-value pequeño, más pequeño que cualquier nivel de significación (como por ejemplo $alpha$=0.5) por lo que se rechaza la hipótesis nula y asumimos **NO Normalidad** en la muestra.

Cuidados_females

```
par(mfrow=c(1,2))
plot(density(cuidados_df$Cuidados_females_norm) ,main="Density")
qqnorm(cuidados_df$Cuidados_females_norm)
qqline(cuidados_df$Cuidados_females_norm)
```



En la gráfica Densidad de la variable "Cuidados_females_norm" , se observa sesgo hacia la izquierda por lo que no se considera normalidad. Se puede confirmar observando la gráfica QQplot en la que la línea que grafica qqline sirve de referencia para interpretar el gráfico, que no se alinea con los puntos de los valores de la variable "Cuidados_females_norm"(tan solo en la parte central).

Tras aplicar la prueba Shapiro-Wilk se comprueba:

```
shapiro.test(cuidados_df$Cuidados_females_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cuidados_df$Cuidados_females_norm
## W = 0.85945, p-value = 0.0008107
```

Se observa un p-value pequeño, más pequeño que cualquier nivel de significación (como por ejemplo $alpha$=0.5) por lo que, se rechaza la hipótesis nula y asumimos **NO Normalidad** en la muestra.