Práctica 2: Limpieza y Análisis de datos

Ángel Carrasco Núñez - Inés Caro Molina

Semestre 2019.2

Contents

1 Descripción del dataset.

1.1 Origen y descripción del dataset.

El *dataset Official World Happiness Report*, donde se centrará la práctica, se ha extraído de la plataforma *kaggle*.

Los datos, provenientes del *Official World Happiness Report*, se han basado en una encuesta sobre el estado de la felicidad global que clasifica a 156 países según lo felices que se sienten sus ciudadanos.

1.2 ¿Por qué es importante y qué pregunta/problema pretende responder?

¿Cuáles son los factores que contribuyen a la felicidad?

¿Cómo evoluciona la felicidad a lo largo del tiempo?

¿Qué países o regiones son más felices?

A pesar de ser cuestiones con un alto grado de subjetividad y sujetas a multitud de interpretaciones, pretendemos dar una respuesta orientativa a partir de un *dataset*, mencionado anteriormente, elaborado con los datos provenientes de *Official World Happiness Reports*.

Este informe está basado en una encuesta sobre el estado de felicidad a nivel global y va ganando reconocimiento mundial a medida que los gobiernos, las organizaciones y la sociedad civil utilizan cada vez más los indicadores de felicidad para tomar decisiones políticas.

Las puntuaciones de felicidad y otras calificaciones subjetivas que recoge este *dataset* se han calculado a partir de las respuestas dadas por los ciudadanos en la **Encuesta Mundial Gallup (GWP)**.

Los datos provienen de muestras representativas a nivel nacional para los años 2005 a 2019. El *dataset* también incluye datos económicos y sociales que pueden estar relacionados con el nivel de felicidad.

A continuación, detallamos los campos del *dataset* que pueden resultar útiles para el desarrollo de la práctica:

Campo	Descripción
Country.name	Nombre del país.
Year	Año en el que se recogen los datos.
Life.ladder	Puntuación de felicidad o bienestar subjetivo. [1]
Log.GDP.per.capita	Logaritmo del PIB *per cápita* extraído a partir de los **Indicadores de Desarrollo Mundial (WDI)**.
Social.support	Media nacional de las respuestas dadas a la pregunta binaria, si se tiene familiares o amigos con los que puede contar el encuestado en caso de necesidad.
Healthy.life.expectancy.at.birth	Esperanza de vida al nacer, basada en datos de la **Organización Mundial de la Salud (OMS)**.
Freedom.to.make.life.choices	Media nacional de las respuestas dadas a la pregunta binaria, si está satisfecho con la libertad que tiene el encuestado de escoger qué hacer con su vida.
Generosity	Residuo de la media nacional de las respuestas dadas a la pregunta binaria, si ha donado dinero a una organización benéfica en el pasado mes, sobre el PIB *per cápita*.
Perceptions.of.corruption	Media nacional de las respuestas a las preguntas binarias, si la corrupción está extendida en el Gobierno y, también, si la corrupción está extendida en las empresas.

[1] Basada en la respuesta a la pregunta conocida como **Escalera de Cantril**. Esta pregunta pide a los encuestados que imaginen la mejor vida posible para ellos (equivalente a un 10 de puntuación) y la peor vida posible (equivalente a un 0 de puntuación). A partir de aquí, deben calificar sus propias vidas actuales en esa escala.

2 Integración y selección de los datos de interés.

Leemos el fichero de Official World Happiness Reports usando "," como separador de decimales, y asignando "NA" a los valores faltantes:

```
df <- read.csv("World_Happiness_Analysis-original.csv", sep = ";", na.strings = "NA", dec = ",")
```

Listamos las variables del *dataset*:

```
names(df)
```

```
##  [1] "Country.name"
##  [2] "year"
##  [3] "Life.Ladder"
##  [4] "Log.GDP.per.capita"
##  [5] "Social.support"
##  [6] "Healthy.life.expectancy.at.birth"
##  [7] "Freedom.to.make.life.choices"
##  [8] "Generosity"
##  [9] "Perceptions.of.corruption"
## [10] "Positive.affect"
## [11] "Negative.affect"
## [12] "Confidence.in.national.government"
## [13] "Democratic.Quality"
## [14] "Delivery.Quality"
## [15] "Standard.deviation.of.ladder.by.country.year"
## [16] "Standard.deviation.Mean.of.ladder.by.country.year"
## [17] "GINI.index..World.Bank.estimate."
## [18] "GINI.index..World.Bank.estimate...average.2000.2017..unbalanced.panel"
## [19] "gini.of.household.income.reported.in.Gallup..by.wp5.year"
## [20] "Most.people.can.be.trusted..Gallup"
## [21] "Most.people.can.be.trusted..WVS.round.1981.1984"
## [22] "Most.people.can.be.trusted..WVS.round.1989.1993"
## [23] "Most.people.can.be.trusted..WVS.round.1994.1998"
## [24] "Most.people.can.be.trusted..WVS.round.1999.2004"
## [25] "Most.people.can.be.trusted..WVS.round.2005.2009"
## [26] "Most.people.can.be.trusted..WVS.round.2010.2014"
```

Seleccionamos las variables de interés para el estudio y las renombramos para facilitar la lectura:

Descripción	**Nombre Original**	**Nombre Nuevo**
País	Country.name	Country
Año	year	Year
Puntuación de felicidad	Life.Ladder	Score
Log. de PIB per cápita	Log.GDP.per.capita	GDP
Apoyo social	Social.support	=
Esperanza de vida	Healthy.life.expectancy.at.birth	Life.expectancy
Libertad	Freedom.to.make.life.choices	Freedom
Generosidad	Generosity	=
Percepción de corrupción	Perception.of.corruption	Corruption

Importamos la librería **dplyr** para manipular *dataframes*:

```
library(dplyr)
```

Seleccionamos y renombramos campos:

```
df <- select(df, Country = Country.name, Year = year, Score = Life.Ladder,
             GDP = Log.GDP.per.capita, Social.support,
             Life.expectancy = Healthy.life.expectancy.at.birth,
             Freedom = Freedom.to.make.life.choices,
             Corruption = Perceptions.of.corruption, Generosity)
```

A continuación, añadimos al *dataset* un campo correspondiente al continente al que pertenece cada país. Para esta clasificación, tomamos como referencia, los datos de *countryRegions* incluidos en el paquete *rworldmap* de R. Este nuevo campo se nombrará como **Region**.

Importamos la librería **rworlmap** para mapear a nivel de país:

```
library(rworldmap)
```

Creamos el campo de región cruzando los datos:

```
pais_region <- rworldmap::countryRegions
df$Region <- pais_region$REGION[match(df$Country, pais_region$ADMIN)]
```

Hay que tener en cuenta que si no se encuentra el país en el fichero de referencia o bien está escrito de otro modo, obtendremos valores NA[2] en el campo *Region*. La gestión de dichos valores se realiza más adelante, en el apartado 3.1.

A continuación, inspeccionamos la estructura de las variables finales del *dataset*.

```
str(df)
```

```
## 'data.frame':    1848 obs. of  10 variables:
##  $ Country        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year           : int  2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 ...
##  $ Score          : num  3.72 4.4 4.76 3.83 3.78 ...
##  $ GDP            : num  7.14 7.31 7.42 7.39 7.48 ...
##  $ Social.support : num  0.451 0.552 0.539 0.521 0.521 ...
##  $ Life.expectancy: num  50.8 51.2 51.6 51.9 52.2 ...
##  $ Freedom        : num  0.718 0.679 0.6 0.496 0.531 ...
##  $ Corruption     : num  0.882 0.85 0.707 0.731 0.776 ...
##  $ Generosity     : num  0.179 0.201 0.132 0.173 0.247 ...
##  $ Region         : chr  "Asia" "Asia" "Asia" "Asia" ...
```

Vemos que el *dataset* está formado por 10 variables y 1.848 registros correspondientes a la información de diferentes países durante los años 2005 a 2019. De estas 10 variables, 7 de ellas son numéricas continuas (*Score*, *Social.support*, *Life.expectancy*, *Freedom*, *Corruption*, *Generosity*), 2 de ellas son categóricas nominales (*Region* y *Country*) y la variable *Year* es numérica discreta.

[2]Valores faltantes.

3 Limpieza de los datos.

3.1 Identificación y tratamiento de valores faltantes.

Identificamos los datos con valores faltantes (NA):

```
colSums(is.na(df))
```

```
##         Country            Year           Score             GDP  Social.support
##               0               0               0              29              13
## Life.expectancy         Freedom      Corruption      Generosity          Region
##              52              31             103              83             105
```

Identificamos los datos que contienen ceros:

```
colSums(df==0)
```

```
##         Country            Year           Score             GDP  Social.support
##               0               0               0              NA              NA
## Life.expectancy         Freedom      Corruption      Generosity          Region
##              NA              NA              NA              NA              NA
```

Se han obtenido **valores faltantes (NA)** en las variables correspondientes al *PIB*, el *apoyo social*, la *esperanza de vida*, la *libertad*, la *percepción de la corrupción* y la *generosidad*.

La cantidad de valores desconocidos es demasiado alta para omitir los registros correspondientes, ya que perderíamos mucha información. Por lo tanto, procedemos a asignar los valores mediante un método basado en la similitud o diferencia entre los registros, llamado *K vecinos más próximos (KNN)*.

Importamos la librería **VIM** para asignar y manipular valores faltantes:

```
library(VIM)
```

Asignamos los valores faltantes (NA) mediante la función **kNN()**:

```
df$GDP <- kNN(df)$GDP
df$Social.support <- kNN(df)$Social.support
df$Life.expectancy <- kNN(df)$Life.expectancy
df$Freedom <- kNN(df)$Freedom
df$Corruption <- kNN(df)$Corruption
df$Generosity <- kNN(df)$Generosity
```

Comprobamos si continua habiendo valores faltantes (NA):

```
colSums(is.na(df))
```

```
##         Country            Year           Score             GDP  Social.support
##               0               0               0               0               0
## Life.expectancy         Freedom      Corruption      Generosity          Region
##               0               0               0               0             105
```

Por otro lado, también se han obtenido **valores faltantes** en el campo correspondiente al continente (*Región*), tal y como se había previsto en el apartado anterior. En este caso, asignaremos manualmente el continente que corresponda.

Mostramos a qué países, les corresponden los valores faltantes (NA):

```
unique(df[is.na(df$Region), "Country"])
```

```
##  [1] "Congo (Brazzaville)"       "Congo (Kinshasa)"
##  [3] "Hong Kong S.A.R. of China" "North Cyprus"
##  [5] "Palestinian Territories"   "Serbia"
```

```
##  [7] "Somaliland region"          "Taiwan Province of China"
##  [9] "Tanzania"                   "United States"
```

Cruzamos los datos entre países y regiones:

```
df$Region <- pais_region$REGION[match(df$Country, pais_region$ADMIN)]
```

Asignamos a los países que carecían de regiones, sus correspondientes:

```
df$Region[df$Country %in% c("Somaliland region", "Congo (Brazzaville)",
                            "Congo (Kinshasa)", "Tanzania")] <- "Africa"
df$Region[df$Country %in% c("Hong Kong S.A.R. of China",
                            "Taiwan Province of China")] <- "Asia"
df$Region[df$Country %in% c("North Cyprus", "Palestinian Territories",
                            "Serbia")] <- "Europe"
df$Region[df$Country=="United States"] <- "North America"
```

3.2 Identificación y tratamiento de valores extremos.

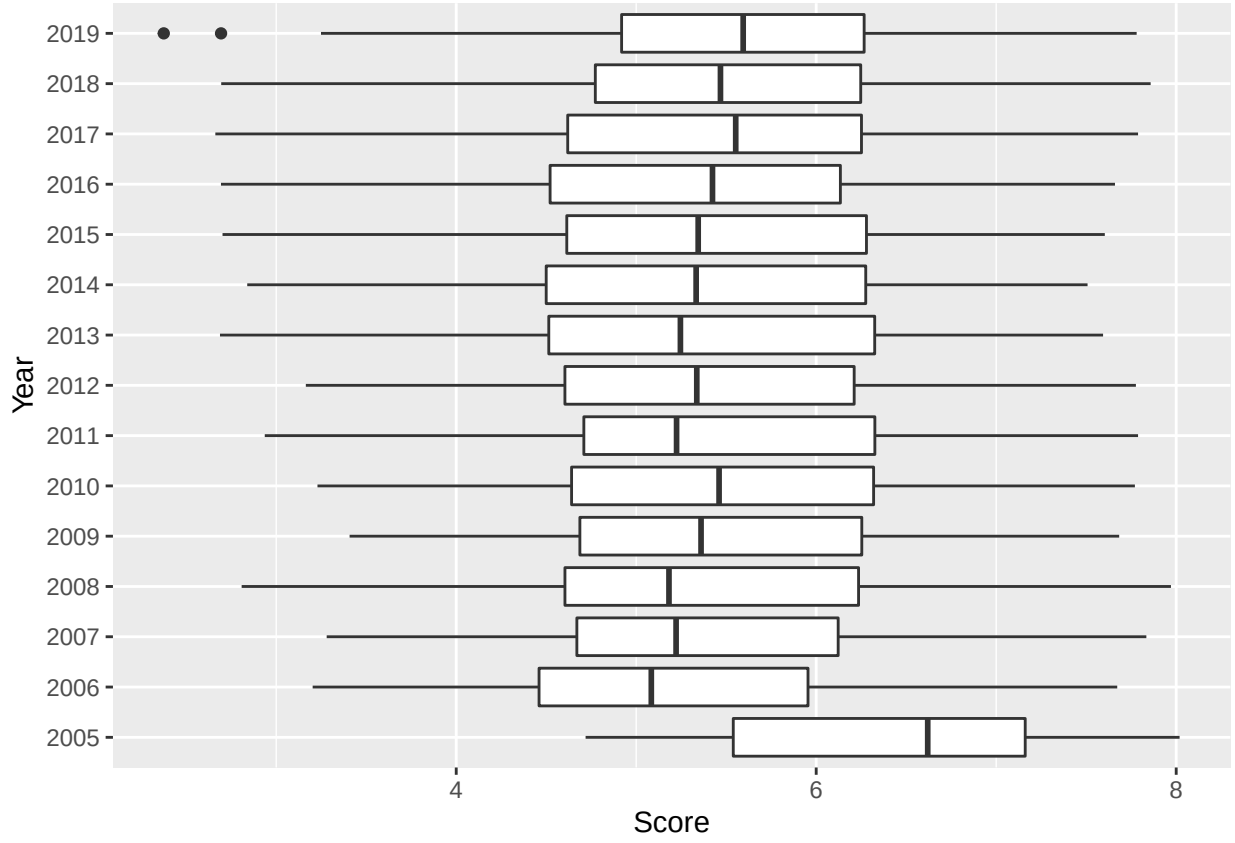
Para analizar los valores extremos, realizamos la gráfica de los datos de cada variable cuantitativa mediante un diagrama de caja o *boxplot*, tal como mostramos a continuación:

Importamos la librería **ggplot2** y **gridExtra** para manejar gráficos:

```
library(ggplot2)
library(gridExtra)
```

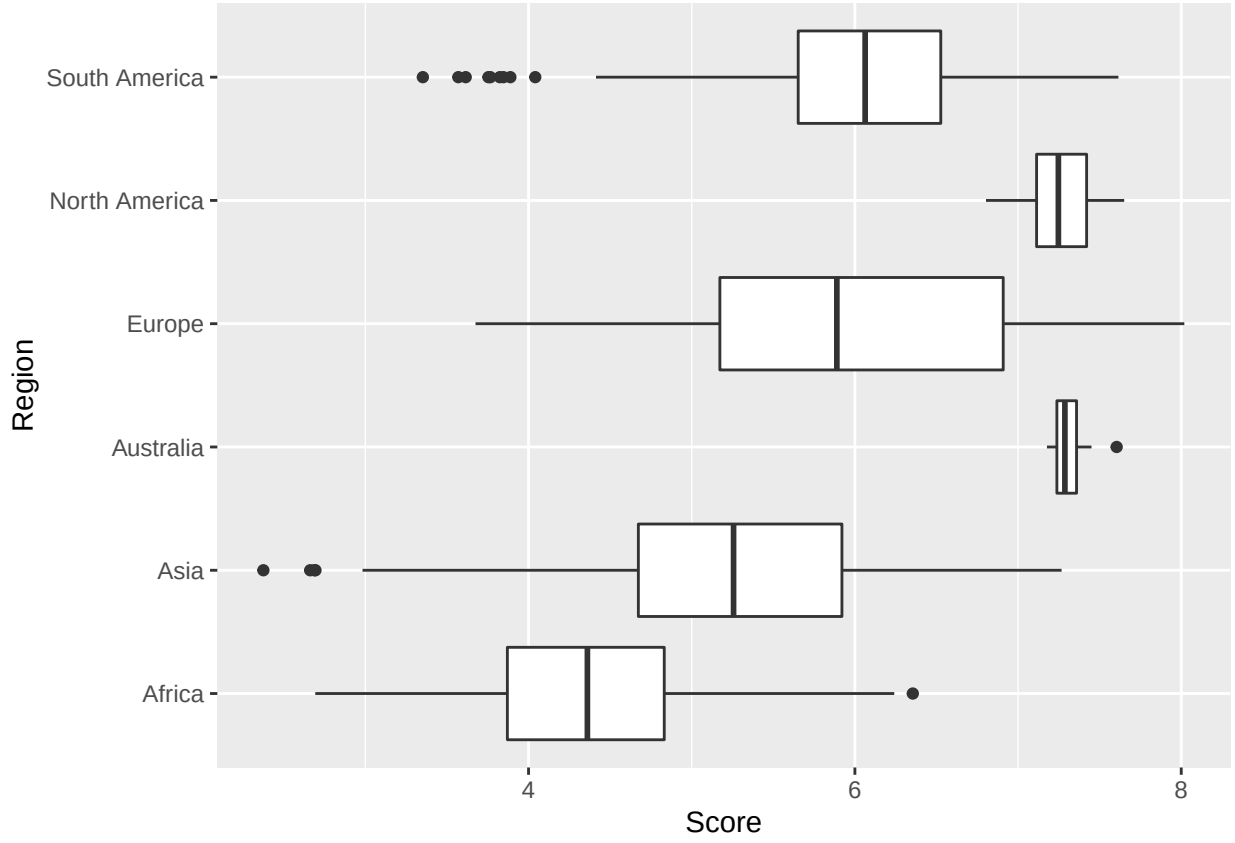
Realizamos un diagrama de caja de la variable *Score* por año:

```
df$Year <- as.factor(df$Year)
ggplot(df, aes(x=Score, y=Year)) + geom_boxplot()
```



Realizamos un diagrama de caja de la variable *Score* por continente:

```
df$Region <- as.factor(df$Region)
ggplot(df, aes(x=Score, y=Region)) + geom_boxplot()
```

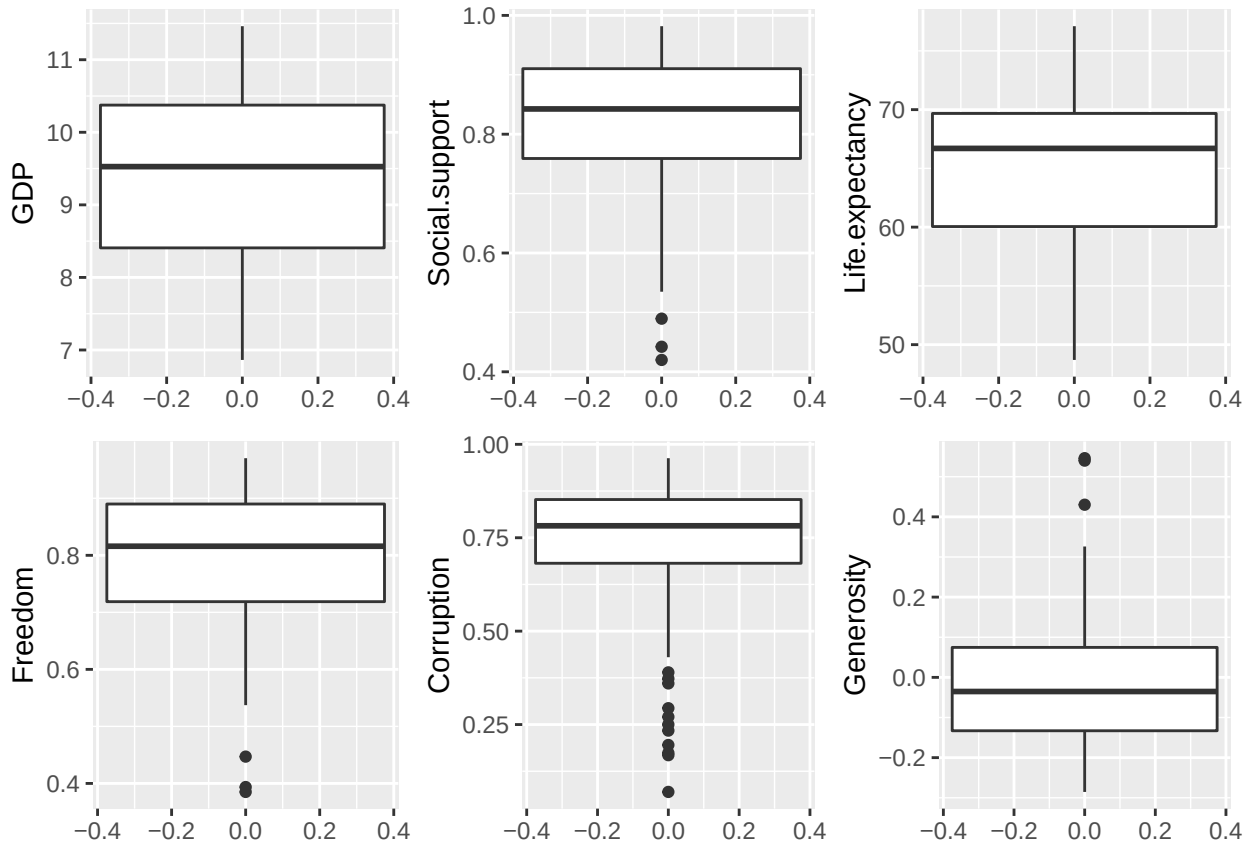


Realizamos un diagrama de caja del resto de variables para el año 2019:

```
b1 <- ggplot(data = df[df$Year=="2019",], aes(y = GDP)) + geom_boxplot()
b1 <- ggplot(data = df[df$Year=="2019",], aes(y = GDP)) + geom_boxplot()
b2 <- ggplot(data = df[df$Year=="2019",], aes(y = Social.support)) + geom_boxplot()
b3 <- ggplot(data = df[df$Year=="2019",], aes(y = Life.expectancy)) + geom_boxplot()
b4 <- ggplot(data = df[df$Year=="2019",], aes(y = Freedom)) + geom_boxplot()
b5 <- ggplot(data = df[df$Year=="2019",], aes(y = Corruption)) + geom_boxplot()
b6 <- ggplot(data = df[df$Year=="2019",], aes(y = Generosity)) + geom_boxplot()
```

Agrupamos los diagramas de caja del resto de variables:

```
grid.arrange(b1, b2, b3, b4, b5, b6, ncol=3)
```



Vemos que existen *outliers* en las variables *Social.support*, *Freedom*, *Corruption* y *Generosity* para el año 2019. Mantendremos dichos valores, ya que, a pesar de ser extremos, son posibles al estar basados en las respuestas subjetivas de la población. Sucede los mismo con la variable *Score* para el año 2019 y para los continentes de América del Sur, Asia, Oceanía y África.

4 Análisis de los datos.

Los análisis que vamos a realizar son los siguientes:

- Análisis 1º: Asociación entre variables. Queremos estudiar la relación que existe entre la variable *Score* y el resto de variables (*GDP*, *Social.support*, *Life.expectancy*, *Freedom*, *Corruption* y *Generosity*) para ver qué factores influyen en la felicidad y el modo y la medida en que lo hacen. Para este análisis, usaremos los datos disponibles más actuales[3]. Las pruebas que aplicaremos en este caso, se basarán en el cálculo de **correlaciones** entre cada par de variables.
- Análisis 2º: Predicción de la felicidad. Una vez analizada la correlación entre variables, queremos generar un **modelo de regresión lineal** que sea capaz de predecir la puntuación de felicidad (*Score*) a partir de las variables explicativas.
- Análisis 3º: Evolución de la felicidad. Queremos comparar la variable *Score* entre los años 2014 y 2019, para estudiar la evolución de la felicidad y responder a la pregunta: ¿Somos más o menos felices? Para este análisis, usaremos los datos de felicidad de los años 2014 y 2019. En este caso, aplicaremos un **contraste de hipótesis** para medias poblacionales con muestras pareadas, que nos permitirá descifrar si la media de felicidad en 2014 es igual o no a la de 2019 de un modo significativo.
- Análisis 4º: Comparación entre regiones. Queremos comparar la variable *Score* entre diferentes continentes. Concretamente queremos analizar, si el continente europeo es significativamente más feliz que el continente africano. Para este análisis, usaremos los datos disponibles más actuales[4]. En este caso, también aplicamos un **contraste de hipótesis** pero, esta vez, con muestras independientes, ya que las observaciones de cada muestra corresponden a casos (países) distintos.

A continuación, mostramos una tabla con el resumen de los análisis:

Análisis	**Tipo**	**Prueba estadística**
1º	Asociación de variables	Correlaciones
2º	Predicción de la felicidad	Modelo de regresión lineal
3º	Evolución de la felicidad	Contraste de medias (muestras pareadas)
4º	Comparación entre regiones	Contraste de medias (muestras independientes)

4.1 Selección de los grupos de datos que se quieren analizar/comparar.

Para llevar a cabo los análisis, crearemos grupos de datos divididos por año y por región.

Para analizar la evolución de la felicidad debemos disponer de los mismos países para los años que comparamos.

Creamos los grupos de datos por año:

```
table(df$Year)
```

```
##
## 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015 2016 2017 2018 2019
##   27   89  102  110  114  124  146  142  137  145  143  142  147  142  138
```

```
df.2005 <- df[df$Year=="2005",]
df.2006 <- df[df$Year=="2006",]
df.2007 <- df[df$Year=="2007",]
df.2008 <- df[df$Year=="2008",]
df.2009 <- df[df$Year=="2009",]
df.2010 <- df[df$Year=="2010",]
df.2011 <- df[df$Year=="2011",]
```

[3] Año 2019.
[4] Año 2019.

```
df.2012 <- df[df$Year=="2012",]
df.2013 <- df[df$Year=="2013",]
df.2014 <- df[df$Year=="2014",]
df.2015 <- df[df$Year=="2015",]
df.2015 <- df[df$Year=="2015",]
df.2016 <- df[df$Year=="2016",]
df.2017 <- df[df$Year=="2017",]
df.2018 <- df[df$Year=="2018",]
df.2019 <- df[df$Year=="2019",]
```

Hallamos los países existentes tanto en 2014 como en 2019:

```
common.countries <- intersect(df.2014$Country,df.2019$Country)
```

Creamos los grupos de datos por año dentro de los países existentes, a la vez, en los años 2014 y 2019:

```
df.comp <- df[(df$Year %in% c("2014","2019") & df$Country %in% common.countries),]
df.comp.2014 <- df.comp[df.comp$Year=="2014",]
df.comp.2019 <- df.comp[df.comp$Year=="2019",]
```

Observamos los valores por año:

```
table(df.2019$Region)
```

```
##
##        Africa          Asia     Australia        Europe North America
##            40            32             2            44             2
## South America
##            18
```

Creamos los grupos de datos por continentes en 2019:

```
df.cont <- df.2019[df.2019$Region %in% c("Europe","Africa"),]
df.AF <- df.2019[df.2019$Region=="Africa",]
df.AS <- df.2019[df.2019$Region=="Asia",]
df.OC <- df.2019[df.2019$Region=="Australia",]
df.EU <- df.2019[df.2019$Region=="Europe",]
df.NA <- df.2019[df.2019$Region=="North America",]
df.SA <- df.2019[df.2019$Region=="South America",]
```

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

4.2.1 Normalidad.

Considerando los análisis que vamos a realizar, debemos comprobar la normalidad de las siguientes variables cuantitativas:

- *GDP*, *Social.support*, *Life.expectancy*, *Freedom*, *Corruption*, *Generosity* para el año 2019.
- *Score* para los años 2014 y 2019.
- *Score* para los continentes de África y Europa en el año 2019.

Para comprobar que las variables cuantitativas de nuestro *dataset* siguen una **distribución normal**, representaremos los histogramas y realizaremos tests de normalidad mediante el paquete de R, *nortest*.

Para las muestras por regiones usaremos el test de *Shapiro-Wilk* (ya que el tamaño de la muestra es menor a 50) y para el resto usaremos el test de *Lilliefors*, que es una modificación del test *Kolmogorov-Smirnov*.

A continuación, desarrollamos los **Histogramas**:

Importamos la librería **psych** para los procedimientos para la investigación psicológica, psicométrica y de la personalidad:

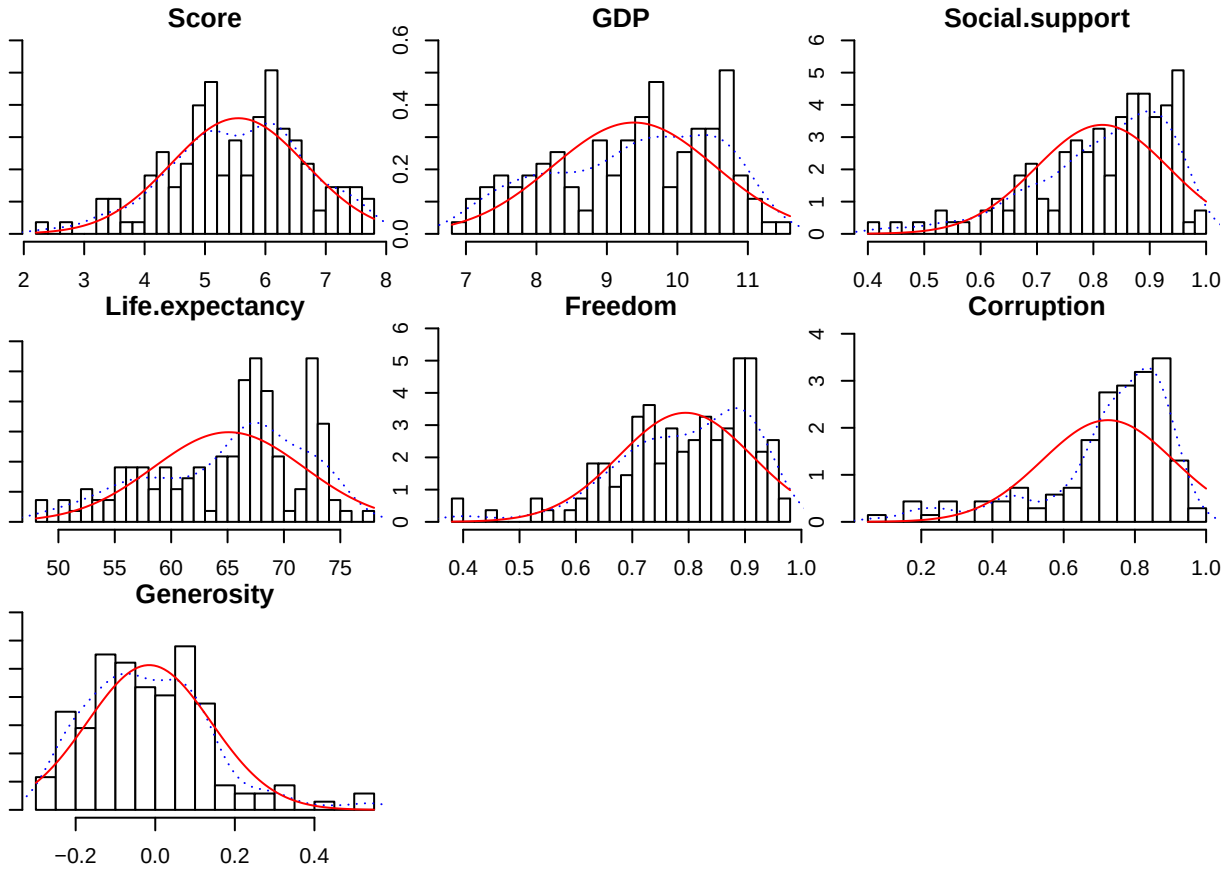
```
library(psych)
```

Importamos la librería **ggpubr** para manejar gráficos:

```
library(ggpubr)
```

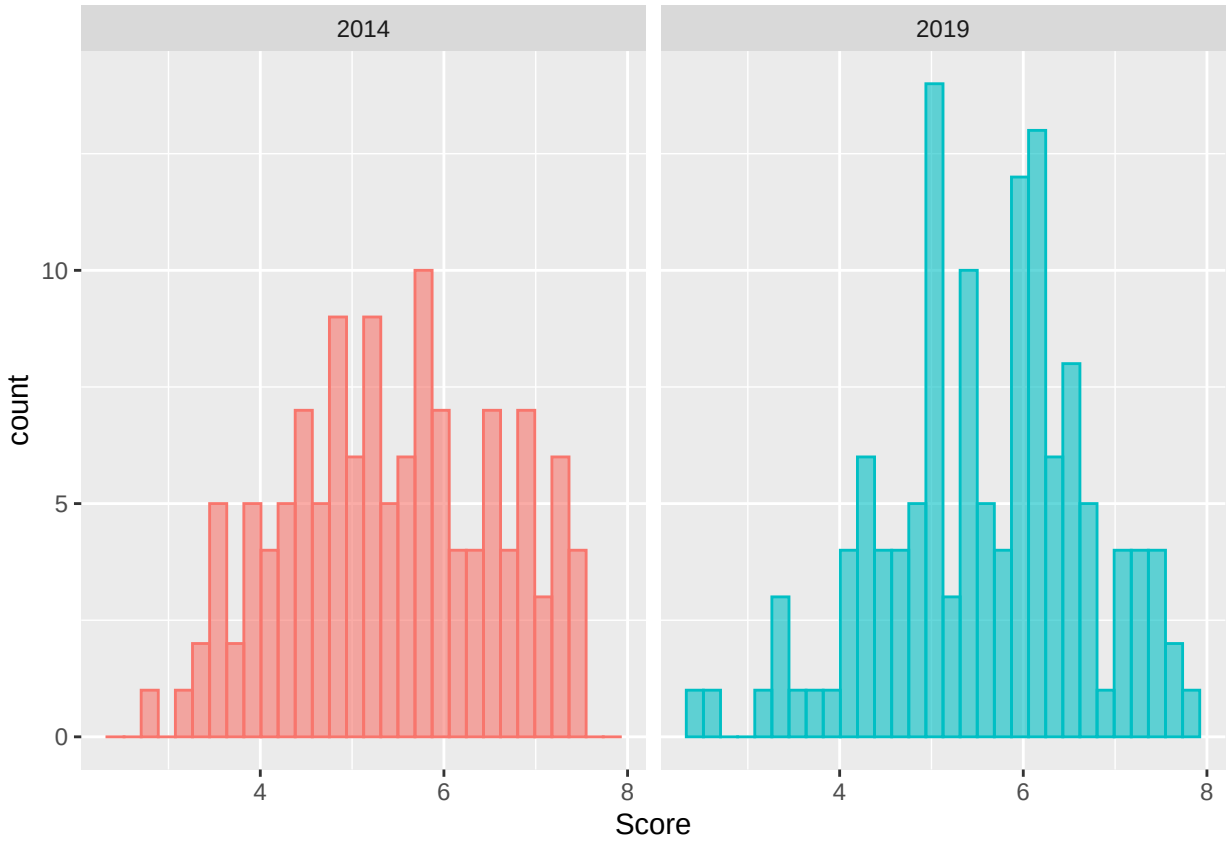
Generamos el histograma de las variables en 2019:

```
multi.hist(x = select_if(df.2019,is.numeric), dcol = c("blue", "red"),
           dlty = c("dotted", "solid"))
```



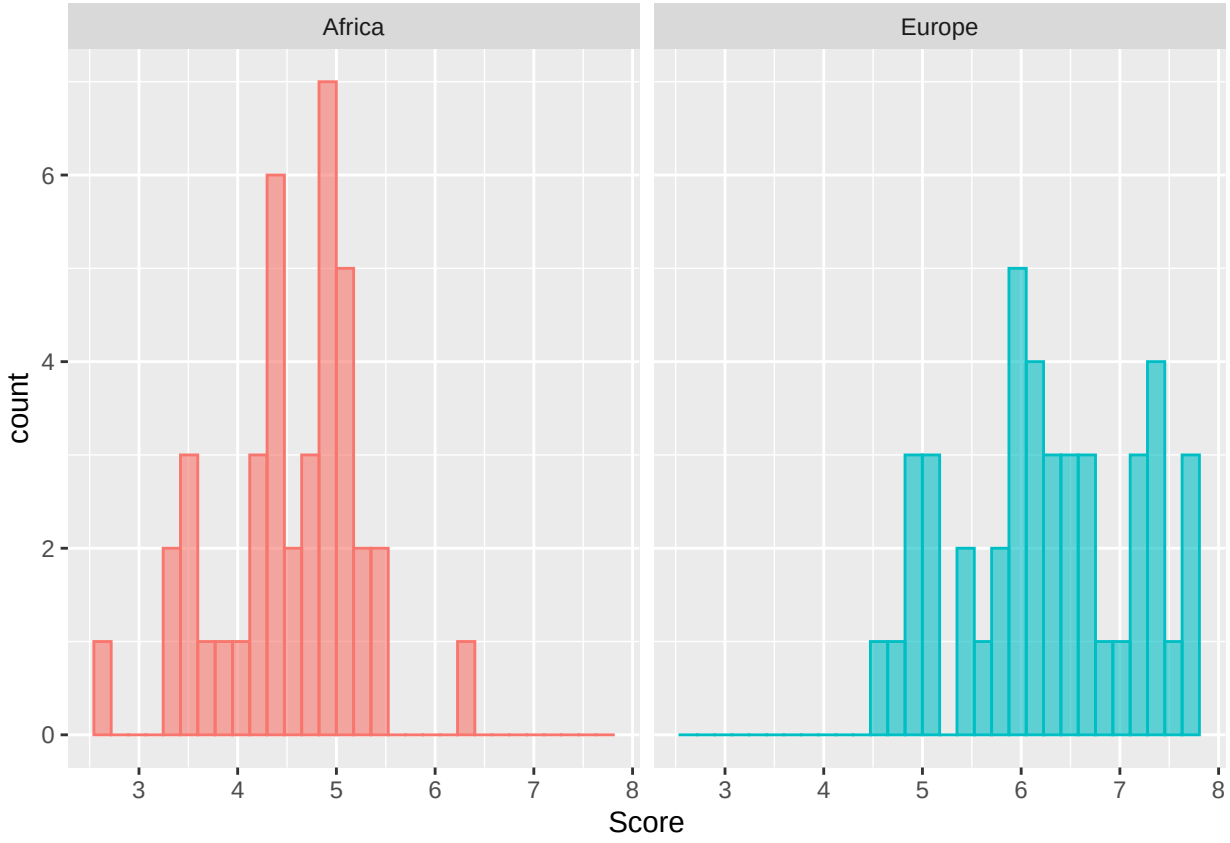
Generamos los histogramas para el año 2014 y 2019:

```
ggplot(df.comp, aes(x=Score, color=Year, fill=Year)) +
  geom_histogram(alpha=0.6) + theme(legend.position="none") + facet_wrap(~Year)
```



Generamos los histogramas para África y Europa en 2019:

```
ggplot(df.cont, aes(x=Score, color=Region, fill=Region)) +
  geom_histogram(alpha=0.6) + theme(legend.position="none") + facet_wrap(~Region)
```



A partir de los histogramas, obtenemos una representación visual de la distribución de los datos. Para asegurarnos que siguen una distribución normal, aplicamos los tests de normalidad. El test se basa en el siguiente contraste de hipótesis (fijando un nivel de confianza del 95%, $\alpha = 0,05$):

- H_0 : La muestra proviene de una distribución normal.
- H_1 : La muestra no proviene de una distribución normal.

A continuación, desarrollamos los **test de normalidad**:

Importamos librería **nortest** para realizar los test de normalidad:

```
library(nortest)
```

Realizamos el test de normalidad por años:

```
norm_by_year <- matrix(c(lillie.test(df.comp.2014$Score)$p.value,
                         lillie.test(df.comp.2019$Score)$p.value), ncol=1, byrow=TRUE)
colnames(norm_by_year) <- c("P-Value")
rownames(norm_by_year) <- c(2014, 2019)
norm_by_year <- as.table(norm_by_year)
norm_by_year
```

```
##         P-Value
## 2014 0.52511977
## 2019 0.05192206
```

El resultado del test nos indica que la variable *Score* en los años 2014 y 2019 sigue una distribución normal (P-Valor>0,05).

Realizamos los test de normalidad por regiones mediante el test *Shapiro*:

```
norm_by_region <- matrix(c(shapiro.test(df.EU$Score)$p.value,
                           shapiro.test(df.AF$Score)$p.value), ncol=1)
colnames(norm_by_region) <- c("P-Value")
rownames(norm_by_region) <- c("Europe", "Africa")
norm_by_region <- as.table(norm_by_region)
norm_by_region
```

```
##          P-Value
## Europe 0.1692556
## Africa 0.1979534
```

Para la variable *Score* agrupada por los continentes de Europa y África, obtenemos P-valores>0,05, por lo que podemos asumir normalidad.

Realizamos el test de normalidad para el año 2019:

```
norm_by_2019 <- matrix(c(lillie.test(df.2019$GDP)$p.value,
                         lillie.test(df.2019$Social.support)$p.value,
                         lillie.test(df.2019$Life.expectancy)$p.value,
                         lillie.test(df.2019$Freedom)$p.value,
                         lillie.test(df.2019$Corruption)$p.value,
                         lillie.test(df.2019$Generosity)$p.value), ncol=1, byrow=TRUE)
colnames(norm_by_2019) <- c("P-Value")
rownames(norm_by_2019) <- c("GDP", "Social Support", "Life Expectancy", "Freedom",
                            "Corruption", "Generosity")
norm_by_2019 <- as.table(norm_by_2019)
norm_by_2019
```

```
##                      P-Value
## GDP             3.984815e-02
## Social Support  4.125965e-04
## Life Expectancy 2.005189e-05
## Freedom         2.936840e-03
## Corruption      4.665930e-10
## Generosity      2.406285e-01
```

Las variables *GDP*, *Social.support*, *Life.expectancy*, *Freedom*, *Corruption* y *Generosity* para el año 2019 no siguen una distribución normal (P valor<0.05).

4.2.2 Homocedasticidad

Considerando los análisis que vamos a realizar, debemos comprobar la homogeneidad de la varianza entre los siguientes grupos de datos:

- *Score* para los años 2014 y 2019.
- *Score* para los continentes África y Europa en el año 2019.

Para comprobar la **homogeneidad de la varianza** (homocedasticidad) entre diferentes grupos, se debe tener en cuenta si las muestras cumplen con la condición de normalidad o no. En nuestro caso siguen una distribución normal por lo que realizaremos el test de *Bartlett* en ambos casos.

Para realizar este test, se realiza el siguiente contraste de hipótesis (fijando un nivel de confianza del 95%, $\alpha = 0,05$):

- H_0 : La varianza es igual entre los grupos.

- H_1 : La varianza no es igual entre los grupos.

Realizamos el Test de *Barlett* por año:

```
bartlett.test(Score ~ Year, data=df.comp)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Score by Year
## Bartlett's K-squared = 0.14838, df = 1, p-value = 0.7001
```

Realizamos el Test de *Bartlett* por continente:

```
bartlett.test(Score ~ Region, data=df.cont)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  Score by Region
## Bartlett's K-squared = 2.0445, df = 1, p-value = 0.1528
```

Tanto para los grupos creados en función del año como para los grupos creados en función del continente, obtenemos un P-valor>0,05, por lo que, podemos asumir igualdad de varianzas entre los años 2014 y 2019, y entre el continente de África y de Europa.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

4.3.1 Análisis 1. Asociación de variables.

Queremos responder a la siguiente pregunta: **¿Qué factores contribuyen a la felicidad?**

Para responder a esta pregunta usaremos los datos más actuales disponibles[5]. Comparamos cada par de variables cuantitativas mediante un diagrama de dispersión múltiple (*pairwise scatterplot*), para observar, si existe relación entre la puntuación de felicidad y el resto de variables. Posteriormente, realizaremos un test de correlación[6] para asegurarnos. También, se estudiará la relación entre variables para detectar una posible colinialidad.

En el diagrama de dispersión múltiple, se muestran *scatterplots* de cada par de variables en la parte izquierda de la figura, el coeficiente de correlación de *Pearson* en la parte derecha y la distribución de cada variable en la diagonal.

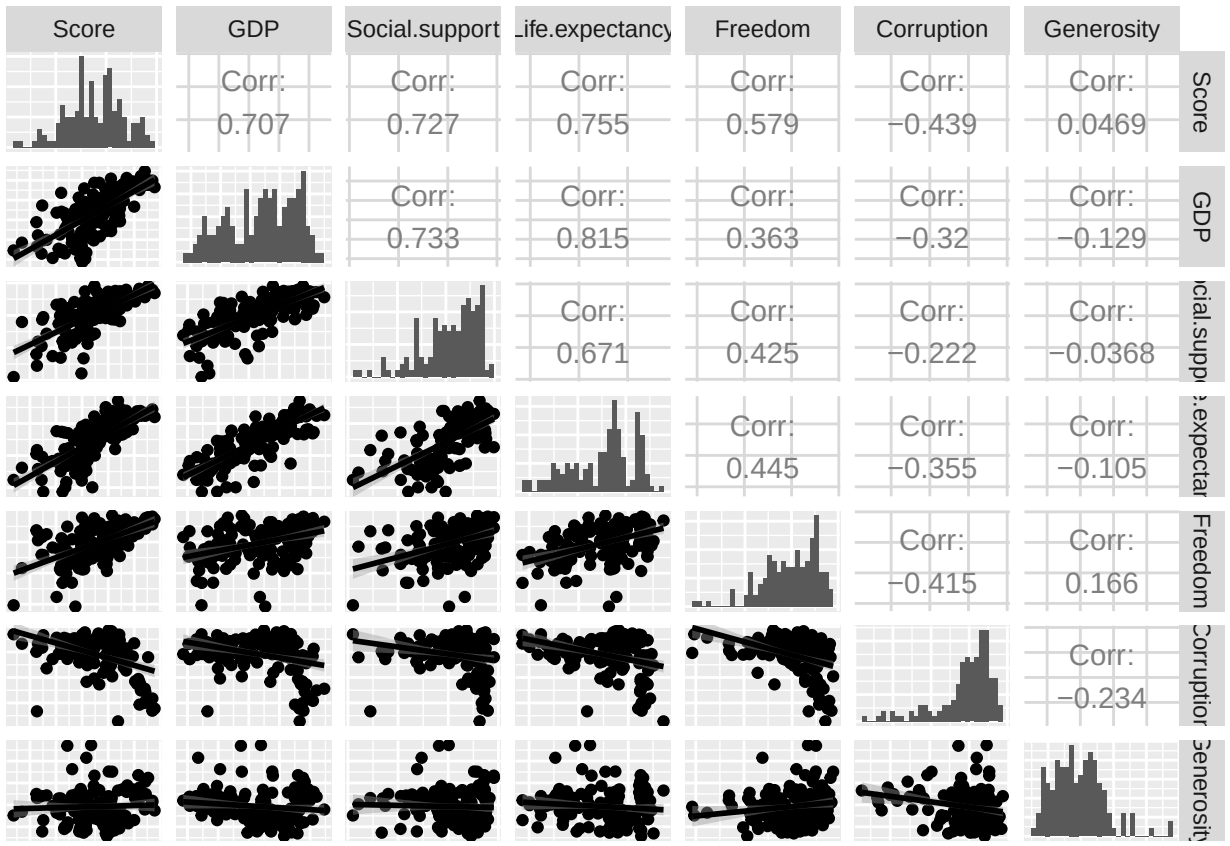
Importamos librería **GGally** para manejar diagramas de dispersión:

```
library(GGally)
```

Realizamos el diagrama de dispersión múltiple:

```
ggpairs(select_if(df.2019,is.numeric), lower = list(continuous = "smooth"),
        diag = list(continuous = "bar"), axisLabels = "none")
```

[5] Año 2019.
[6] Usaremos el test de *Kendall* ya que las variables no siguen una distribución normal



Observamos que existe relación lineal positiva entre la variable *Score* y las variables *GDP*, *social.support*, *life.expectancy* y en menor medida, *freedom*. Entre la variable *Score* y *Corruption* vemos una ligera relación lineal negativa mientras que entre la variable *Score* y *Generosity* no parece haber ningún tipo de relación (la nube de puntos no sigue ningún patrón). También hay indicios de colinealidad por la relación que existe entre las variables *GDP*, *social.support* y *life.expectancy*.

A continuación aplicamos el test *kendall* por cada variable:

```
test_kendall_1 <- cor.test(df.2019$Score,df.2019$GDP,method = "kendall")
test_kendall_2 <- cor.test(df.2019$Score,df.2019$Social.support,method = "kendall")
test_kendall_3 <- cor.test(df.2019$Score,df.2019$Life.expectancy,method = "kendall")
test_kendall_4 <- cor.test(df.2019$Score,df.2019$Freedom,method = "kendall")
test_kendall_5 <- cor.test(df.2019$Score,df.2019$Corruption,method = "kendall")
test_kendall_6 <- cor.test(df.2019$Score,df.2019$Generosity,method = "kendall")
```

Extraemos el P-Valor y los coeficientes de correlación:

```
correlaciones <- matrix(c(test_kendall_1$p.value, test_kendall_2$p.value,
                      test_kendall_3$p.value, test_kendall_4$p.value,
                      test_kendall_5$p.value, test_kendall_6$p.value,
                      test_kendall_1$estimate, test_kendall_2$estimate,
                      test_kendall_3$estimate, test_kendall_4$estimate,
                      test_kendall_5$estimate, test_kendall_6$estimate), nrow=6)
colnames(correlaciones) <- c("P Value","Coeficiente")
rownames(correlaciones) <- c("GDP","Social.support","Life.expectancy","Freedomm",
                             "Corruption","Generosity")
correlaciones <- as.table(correlaciones)
correlaciones
```

```
##                       P Value   Coeficiente
## GDP              3.097304e-21  5.436945e-01
## Social.support   2.842281e-22  5.578123e-01
## Life.expectancy  3.585527e-23  5.711879e-01
## Freedomm         2.526649e-14  4.379794e-01
## Corruption       1.253706e-05 -2.511111e-01
## Generosity       7.086546e-01  2.147694e-02
```

El P-Valor es menor a 0,05 en las variables *GDP*, *Life.expectancy*, *Social.support*, *Freedom* y *Corruption*, por lo que, podemos concluir que estas variables están significativamente correlacionadas con *Score*. A partir del coeficiente de correlación, vemos la fuerza y la dirección de estas relaciones. Los factores que más influyen en la felicidad son el PIB *per cápita*, el apoyo social, la esperanza de vida y la libertad. También, existe una ligera y negativa relación entre *Score* y *Corruption*.

4.3.2 Análisis 2. Predicción de la felicidad.

Con el fin de predecir el nivel de felicidad, crearemos un modelo de regresión lineal múltiple, donde la variable dependiente es la puntuación de felicidad (*Score*) y las variables independientes o explicativas son las correspondientes a la familia, el PIB *per cápita*, la esperanza de vida, la percepción de la corrupción, la libertad y la generosidad. Usaremos los datos más actuales[7] para generar el modelo. A partir de las correlaciones vistas en el apartado anterior, podemos intuir las variables que se podrían incluir en el modelo, sin embargo, para empezar, generaremos un modelo con todas ellas y analizaremos el resultado.

```
modelo <- lm(Score ~ GDP + Social.support + Life.expectancy + Freedom + Corruption +
               Generosity, data = df.2019)
summary(modelo)
```

```
##
## Call:
## lm(formula = Score ~ GDP + Social.support + Life.expectancy +
##     Freedom + Corruption + Generosity, data = df.2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.65812 -0.35689  0.07596  0.41139  1.35575
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.30520    0.71451  -3.226 0.001584 **
## GDP              0.08202    0.08583   0.956 0.341046
## Social.support   3.08926    0.66814   4.624 8.93e-06 ***
## Life.expectancy  0.05583    0.01390   4.015 9.94e-05 ***
## Freedom          1.87737    0.52705   3.562 0.000514 ***
## Corruption      -0.76734    0.32377  -2.370 0.019246 *
## Generosity       0.30330    0.35149   0.863 0.389763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.596 on 131 degrees of freedom
## Multiple R-squared:  0.7254, Adjusted R-squared:  0.7128
## F-statistic: 57.68 on 6 and 131 DF,  p-value: < 2.2e-16
```

El P-Valor del estadístico **F** es menor a 0,05, esto indica que, al menos, una de las variables del modelo es significativa. Para ver, cuáles lo son, debemos recurrir al P-Valor de los coeficientes de regresión parciales. Vemos que las variables *GDP* y *Generosity* tienen un P-Valor mayor a 0,05, por lo que, no las consideramos

[7] Año 2019.

significativas. Mientras que las variables *Social.support*, *Life.expectancy*, *Freedom* y *Corruption* sí serán significativas (P-Valores<0,05). En conjunto, el coeficiente de determinación ajustado $R^2 = 0,7171$, indica un buen ajuste siendo el modelo capaz de explicar el 71,71% de la variabilidad observada en el nivel de felicidad.

Puede sorprender que no se considere la variable *GDP* significativa a la hora de explicar el modelo. El motivo puede ser la relación que guarda con las variables *life.expectancy* y *social.support*. En el apartado anterior, ya hemos vistos indicios de colinealidad. A continuación, calculamos el **VIF** para medir la colinealidad de las variables. Vemos que el **VIF** para la variable *GDP* es bastante alto, aunque, al no llegar a 5, no lo consideraremos un motivo de preocupación.

Importamos la librería **faraway** para manejar la colinealidad:

```
library(faraway)
faraway::vif(modelo)
```

```
##             GDP  Social.support Life.expectancy         Freedom      Corruption
##        3.794463        2.397941        3.342460        1.489976        1.374886
##      Generosity
##        1.152074
```

Para escoger el mejor modelo, podemos emplear la estrategia de *stepwise mixto*. A partir del modelo con todas las variables como predictores, realizamos la selección mediante la **medición Akaike (AIC)**.

```
step(object = modelo, direction = "both", trace = 1)
```

```
## Start:  AIC=-136.01
## Score ~ GDP + Social.support + Life.expectancy + Freedom + Corruption +
##     Generosity
##
##                   Df Sum of Sq    RSS     AIC
## - Generosity       1    0.2645 46.799 -137.23
## - GDP              1    0.3244 46.859 -137.06
## <none>                         46.535 -136.01
## - Corruption       1    1.9953 48.530 -132.22
## - Freedom          1    4.5071 51.042 -125.26
## - Life.expectancy  1    5.7271 52.262 -122.00
## - Social.support   1    7.5941 54.129 -117.15
##
## Step:  AIC=-137.23
## Score ~ GDP + Social.support + Life.expectancy + Freedom + Corruption
##
##                   Df Sum of Sq    RSS     AIC
## - GDP              1    0.2632 47.062 -138.46
## <none>                         46.799 -137.23
## + Generosity       1    0.2645 46.535 -136.01
## - Corruption       1    2.5323 49.331 -131.96
## - Freedom          1    4.9212 51.720 -125.43
## - Life.expectancy  1    5.5382 52.337 -123.80
## - Social.support   1    7.8420 54.641 -117.85
##
## Step:  AIC=-138.46
## Score ~ Social.support + Life.expectancy + Freedom + Corruption
##
##                   Df Sum of Sq    RSS     AIC
## <none>                         47.062 -138.46
## + GDP              1    0.2632 46.799 -137.23
## + Generosity       1    0.2033 46.859 -137.06
```

```
## - Corruption       1    2.7489 49.811 -132.62
## - Freedom          1    4.7179 51.780 -127.27
## - Social.support   1   11.5106 58.573 -110.26
## - Life.expectancy  1   11.6502 58.712 -109.94
```

```
##
## Call:
## lm(formula = Score ~ Social.support + Life.expectancy + Freedom +
##     Corruption, data = df.2019)
##
## Coefficients:
##     (Intercept)   Social.support  Life.expectancy          Freedom
##        -2.12417          3.38879          0.06203          1.88719
##      Corruption
##        -0.86683
```

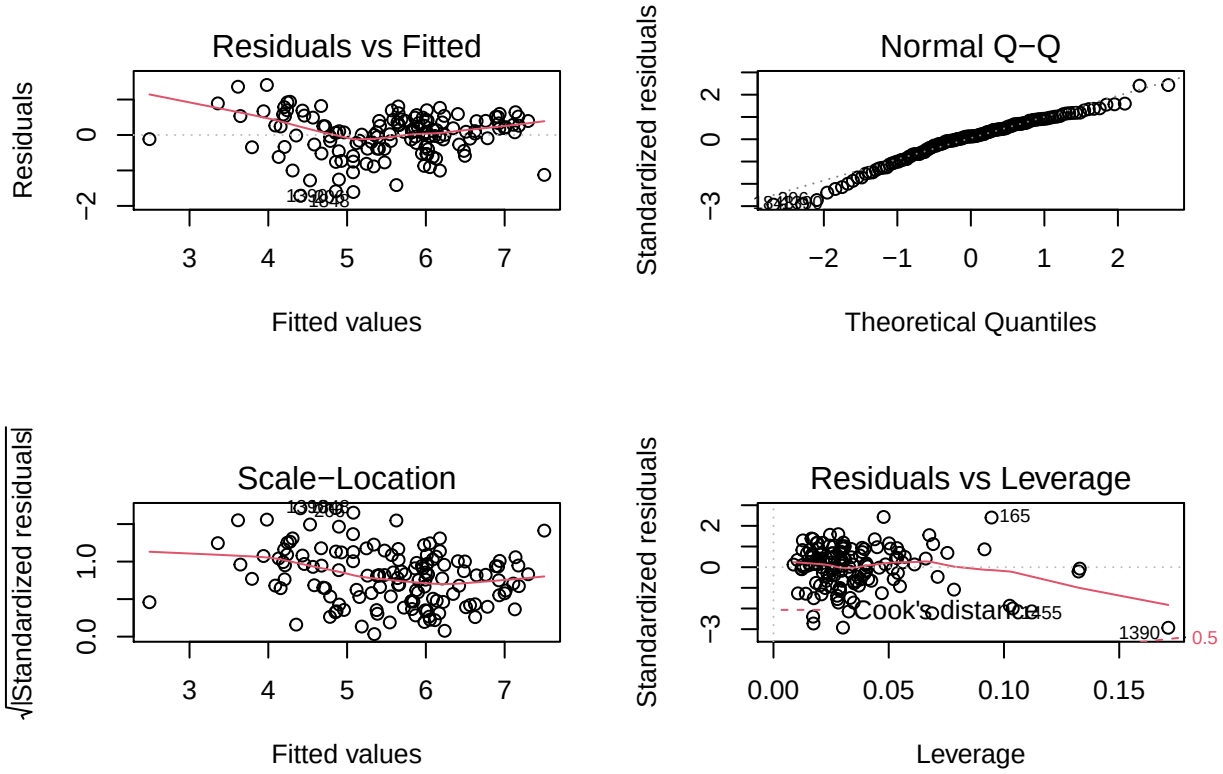
Obtenemos que el mejor modelo es: $Score = -2,071 + 3,392\ Social.support + 0,061\ Life.expectancy + 1,875\ Freedom - 0,9\ Corruption$

```
modelo_final <- lm(Score ~ Social.support + Life.expectancy + Freedom + Corruption,
                   data = df.2019)
summary(modelo_final)
```

```
##
## Call:
## lm(formula = Score ~ Social.support + Life.expectancy + Freedom +
##     Corruption, data = df.2019)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.71365 -0.34368  0.07399  0.39977  1.41048
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -2.12417    0.69589  -3.052 0.002741 **
## Social.support   3.38879    0.59416   5.703 7.26e-08 ***
## Life.expectancy  0.06203    0.01081   5.738 6.17e-08 ***
## Freedom          1.88719    0.51683   3.651 0.000374 ***
## Corruption      -0.86683    0.31101  -2.787 0.006096 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5949 on 133 degrees of freedom
## Multiple R-squared:  0.7223, Adjusted R-squared:  0.7139
## F-statistic: 86.48 on 4 and 133 DF,  p-value: < 2.2e-16
```

Finalmente, para validar el modelo, representamos gráficamente los residuos para comprobar la linealidad, la homocedasticidad y normalidad de residuos.

```
par(mfrow = c(2,2))
plot(modelo_final)
```



4.3.3 Análisis 3. Evolución de la felicidad.

Queremos responder a la siguiente pregunta: **¿Somos más felices ahora que hace unos años?**

Para responder a esta pregunta, comparamos los datos de felicidad de 2014 frente a los de 2019. Concretamente, queremos comparar la media de la puntuación de felicidad en 2014 frente a la media de la puntuación de felicidad en 2019. Se cumplen las condiciones de normalidad e igualdad de varianzas (tal y como hemos visto en el apartado anterior), aunque las muestras no son independientes. Por lo tanto, podemos aplicar un **contraste de hipótesis** sobre la diferencia de medias basado en la *t-Student* con una adaptación para muestras pareadas. Las hipótesis del test son las siguientes:

- $H_0 : \mu_{2014} - \mu_{2019} = 0$ (las medias son iguales).
- $H_1 : \mu_{2014} - \mu_{2019} < 0$ (la media de 2016 es menor a la media de 2019).

Fijamos el nivel de significación en $\alpha = 0,05$.

Aplicamos el *t-test* para muestras relacionadas:

```
t.test(x = df.comp.2014$Score, y = df.comp.2019$Score, alternative = "less", paired = T)
```

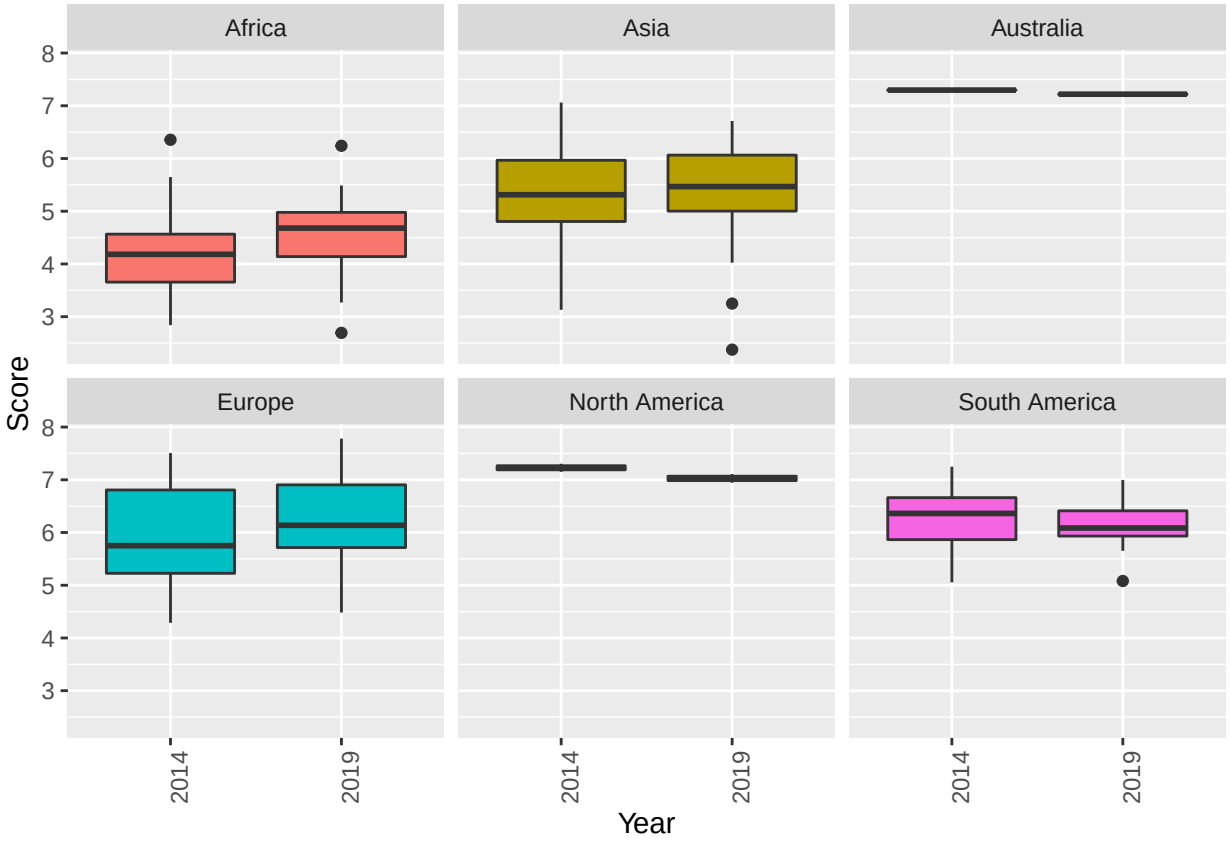
```
##
##  Paired t-test
##
## data:  df.comp.2014$Score and df.comp.2019$Score
## t = -2.6807, df = 127, p-value = 0.004161
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##         -Inf -0.05915668
## sample estimates:
```

```
## mean of the differences
##               -0.154905
```

Observamos que el P-Valor es menor a 0,05, por lo que, la media de 2014 es significativamente menor a la de 2019. Entonces, podemos decir que ha aumentado el nivel de felicidad entre los años 2014 y 2019 a nivel global.

A continuación, representamos visualmente la evolución de la felicidad entre los años 2014 y 2019.

```
ggplot(df.comp, aes(x=Year, y=Score, fill=Region)) + geom_boxplot()+facet_wrap(~Region)+
  theme(legend.position="none", axis.text.x = element_text(angle=90, hjust=1))
```



Hemos visto que ha habido una mejora significativa respecto al 2014 del nivel de felicidad a nivel global. Sin embargo, en el gráfico vemos como en algunas regiones aumenta, en otras disminuye y en otras se mantiene prácticamente constante.

4.3.4 Análisis 4. Comparación entre continentes. Contraste de hipótesis.

Queremos responder a la siguiente pregunta: **¿Somos más felices en Europa que en África?**

Para responder a esta pregunta, comparamos los datos de felicidad entre el continente europeo y el continente africano. Concretamente queremos comparar la media de la puntuación de felicidad en Europa frente a la media de la puntuación de felicidad en África, para los datos de 2019. Como hemos visto, se cumplen los supuestos de normalidad y homocedasticidad, así que también aplicamos un **contraste de hipótesis** (con $\alpha = 0,05$), pero esta vez, con muestras independientes, ya que las observaciones de cada muestra, corresponden a casos (países) distintos. Las hipótesis del test son las siguientes:

- $H_0 : \mu_{AF} - \mu_{EU} = 0$ (las medias son iguales).
- $H_1 : \mu_{AF} - \mu_{EU} < 0$ (la media de África es menor a la media de Europa).

Aplicamos el *t-test* para muestras independientes:

```
t.test(x = df.AF$Score, y = df.EU$Score, alternative = "less")
```

```
##
##  Welch Two Sample t-test
##
## data:  df.AF$Score and df.EU$Score
## t = -9.7395, df = 80.667, p-value = 1.436e-15
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##      -Inf -1.43775
## sample estimates:
## mean of x mean of y
##  4.531234  6.265232
```

Observamos que el P-Valor es menor a 0,05, por lo que, rechazamos la hipótesis nula y podemos decir que Europa es significativamente más feliz que África.

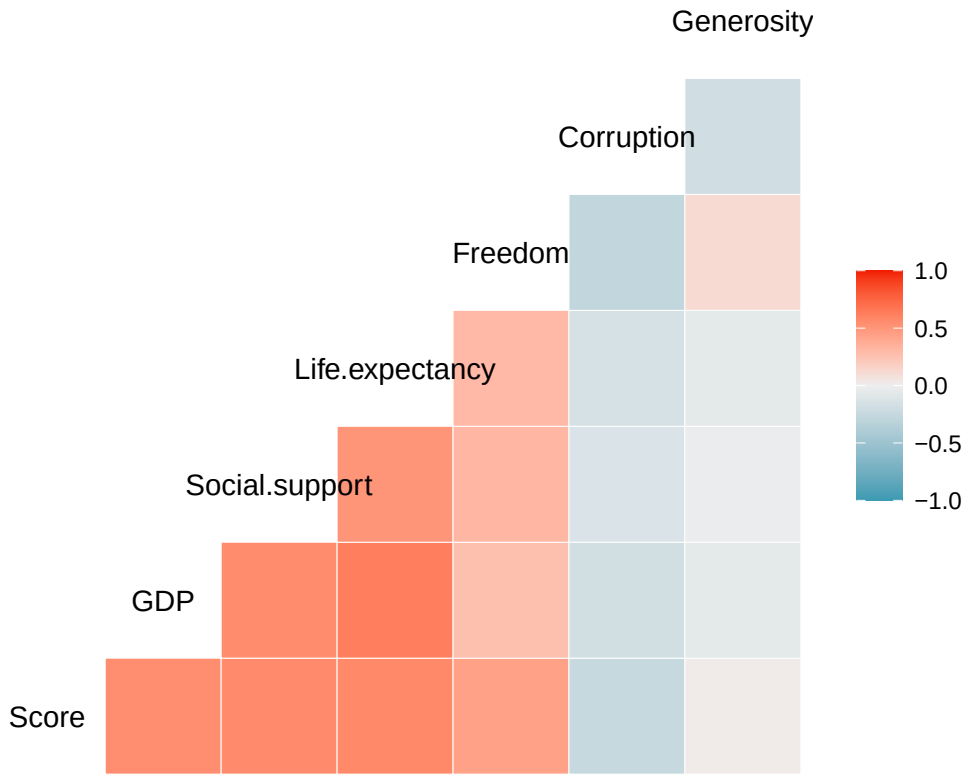
5 Representación de los resultados

En este apartado, complementamos los resultados del apartado anterior con algunas representaciones gráficas.

5.1 Análisis 1. Correlación.

Visualización de la correlación (coeficiente de correlación *Kendall*):

```
ggcorr(select_if(df.2019,is.numeric), method = c("everything", "kendall"))
```



5.2 Análisis 2. Predicción.

Ecuación y coeficientes de regresión del modelo de regresión múltiple:

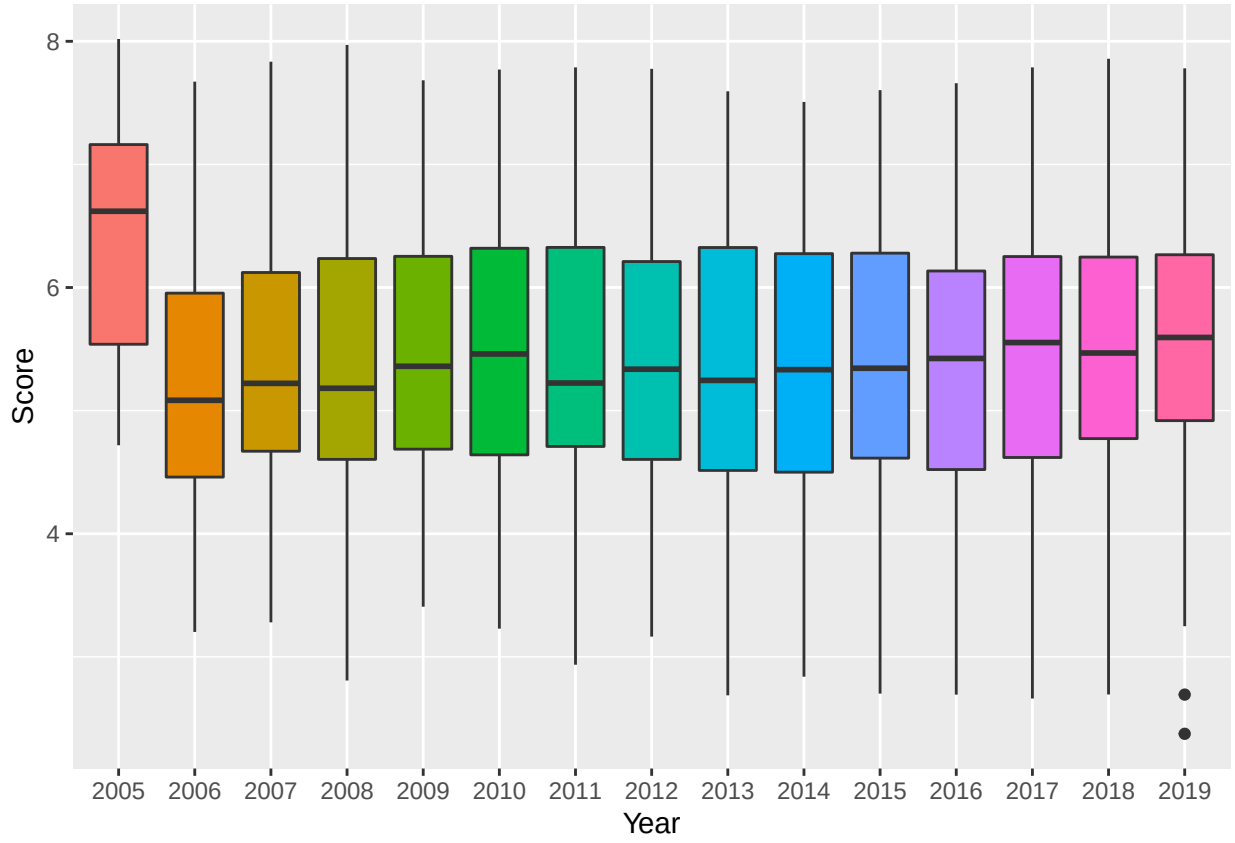
```
modelo_final
```

```
##
## Call:
## lm(formula = Score ~ Social.support + Life.expectancy + Freedom +
##     Corruption, data = df.2019)
##
## Coefficients:
##     (Intercept)   Social.support  Life.expectancy          Freedom
##         -2.12417          3.38879          0.06203          1.88719
##      Corruption
##         -0.86683
```

5.3 Análisis 3. Evolución felicidad.

Boxplot de felicidad entre años:

```
ggplot(df, aes(x=Year, y=Score, fill=Year)) + geom_boxplot() +
  theme(legend.position="none")
```



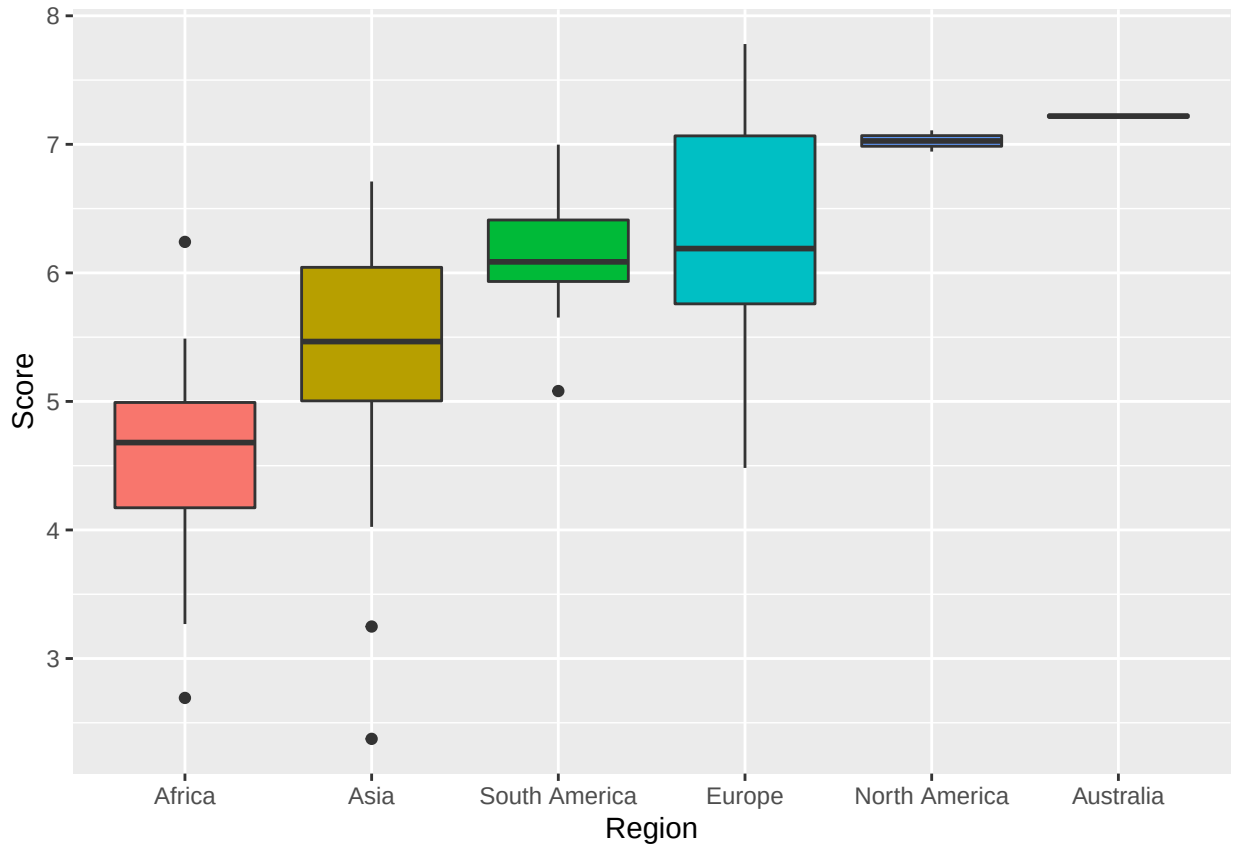
5.4 Análisis 4. Comparación entre regiones.

Ordenación de los continentes por *Score*:

```
df.2019$Region = with(df.2019, reorder(Region, Score, median))
```

Boxplot de felicidad entre continentes:

```
ggplot(df.2019, aes(x=Region, y=Score, fill=Region)) + geom_boxplot() +
  theme(legend.position="none")
```



Mapa de la felicidad mundial en el año 2019:

- Añadimos los datos al mapa:

```
mapdf <- joinCountryData2Map(df.2019, joinCode="NAME",
                             nameJoinColumn="Country", verbose=TRUE)
```

```
## 134 codes from your data successfully matched countries in the map
## 4 codes from your data failed to match with a country code in the map
##      failedCodes failedCountries
## [1,] NA          "Hong Kong S.A.R. of China"
## [2,] NA          "North Cyprus"
## [3,] NA          "Palestinian Territories"
## [4,] NA          "Taiwan Province of China"
## 109 codes from the map weren't represented in your data
```

- Renombramos los países que no se han vinculado anteriormente[8]:

[8] No se ha encontrado correspondiencia para North Cyprus.

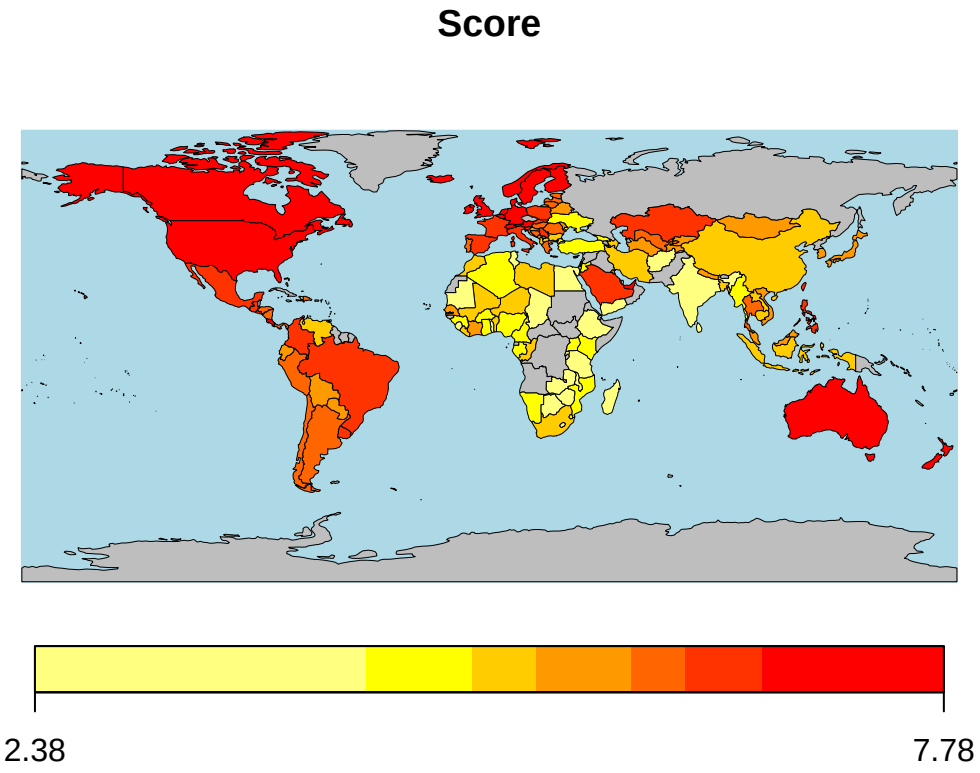
```
df.2019$Country <- as.character(df.2019$Country)
df.2019$Country[df.2019$Country=="Hong Kong S.A.R. of China"] <- "Hong Kong"
df.2019$Country[df.2019$Country=="Palestinian Territories"] <-
  "Palestinian Territory, Occupied"
df.2019$Country[df.2019$Country=="Taiwan Province of China"] <-
  "Taiwan, Province of China"
df.comp$Country <- as.character(df.comp$Country)
df.comp$Country[df.comp$Country=="Hong Kong S.A.R. of China"] <- "Hong Kong"
df.comp$Country[df.comp$Country=="Palestinian Territories"] <-
  "Palestinian Territory, Occupied"
df.comp$Country[df.comp$Country=="Taiwan Province of China"] <-
  "Taiwan, Province of China"

mapdf <- joinCountryData2Map(df.2019, joinCode="NAME",
                             nameJoinColumn="Country", verbose=TRUE)
```

```
## 137 codes from your data successfully matched countries in the map
## 1 codes from your data failed to match with a country code in the map
##      failedCodes failedCountries
## [1,] NA          "North Cyprus"
## 106 codes from the map weren't represented in your data
```

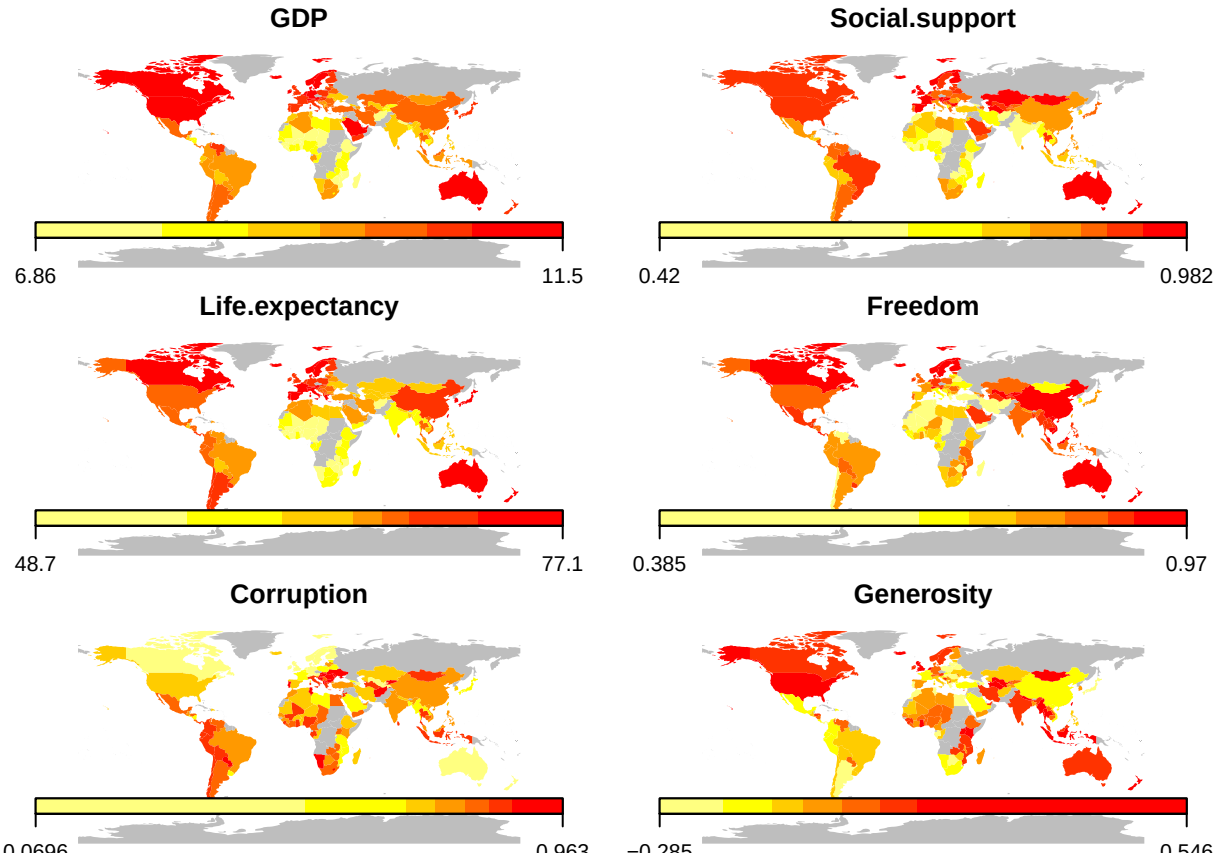
- Realizamos el mapa de felicidad:

```
mapCountryData(mapdf, nameColumnToPlot="Score", oceanCol = "lightblue",
               borderCol = "black", missingCountryCol="grey")
```



Realizamos los mapas de cada variable (*GDP*, *Social.support*, *Life.expectancy*, *Freedom*, *Corruption* y *Generosity*) con los valores del año 2019:

```
variables <- c("GDP", "Social.support", "Life.expectancy", "Freedom", "Corruption",
               "Generosity")
op <- par(mfcol=c(3,2), mai=c(0,0.2,0.3,0))
for (i in variables) {
  map_by_variable <- mapCountryData(mapdf, nameColumnToPlot=i, borderCol = "transparent",
                                    missingCountryCol="grey")
 }
```




```
par(op)
```

Mapa de la felicidad mundial entre los años 2014 y 2019:

- Importamos la librería **RColorBrewer** para gestionar los colores de los mapas:

```
library(RColorBrewer)
```

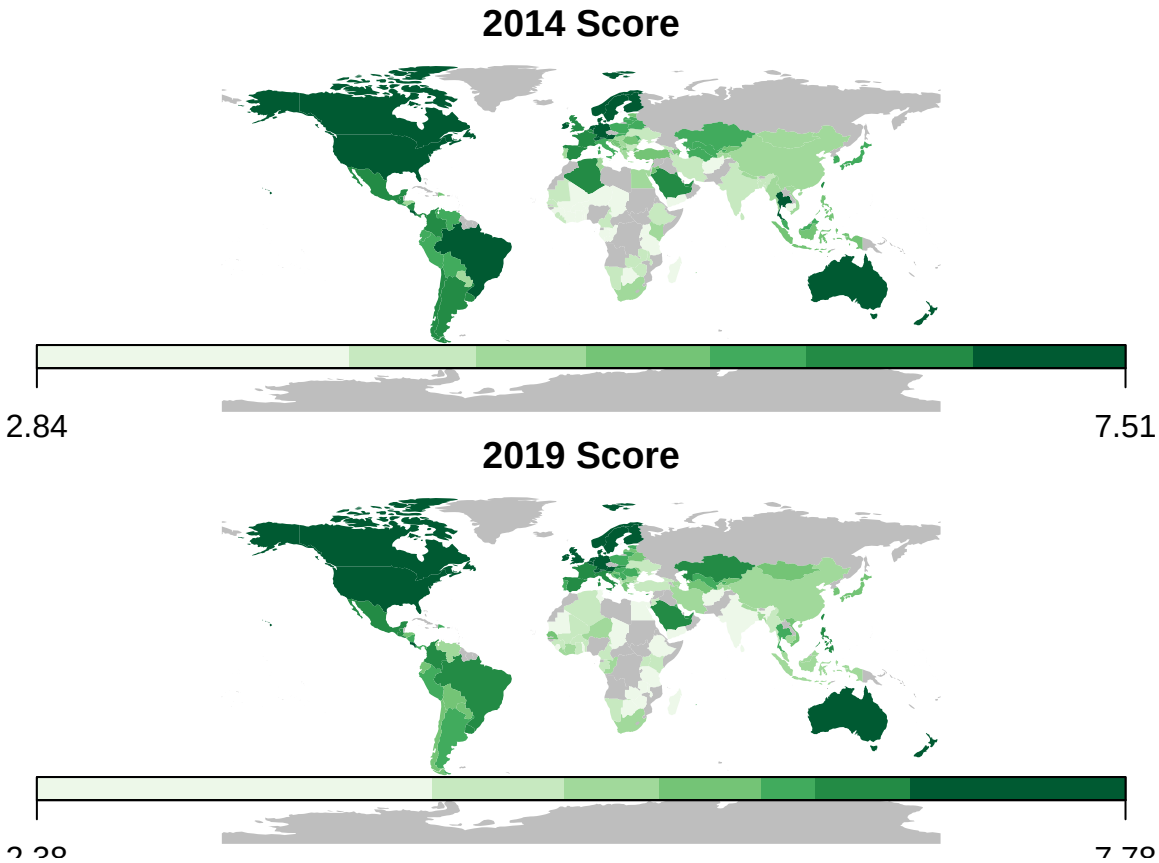
- Preparamos el próximo mapa de felicidad entre 2014 y 2019:

```
colourPalette <- RColorBrewer::brewer.pal(7,"Greens")
anyos <- c("2014", "2019")
op <- par(mfcol=c(2,1),mai=c(0,0.2,0.3,0))
for (i in anyos) {
  mapdf_anyos <- joinCountryData2Map(df.comp[df.comp$Year==i,], joinCode="NAME",
                                nameJoinColumn="Country", verbose=TRUE)
  map_by_year <- mapCountryData(mapdf_anyos,
                                nameColumnToPlot="Score",
```

```
                                borderCol = "transparent",
                                missingCountryCol="grey",
                                colourPalette = colourPalette,
                                mapTitle = paste(i,"Score"),)
 }
```

```
## 127 codes from your data successfully matched countries in the map
## 1 codes from your data failed to match with a country code in the map
##      failedCodes failedCountries
## [1,] NA          "North Cyprus"
## 116 codes from the map weren't represented in your data
```

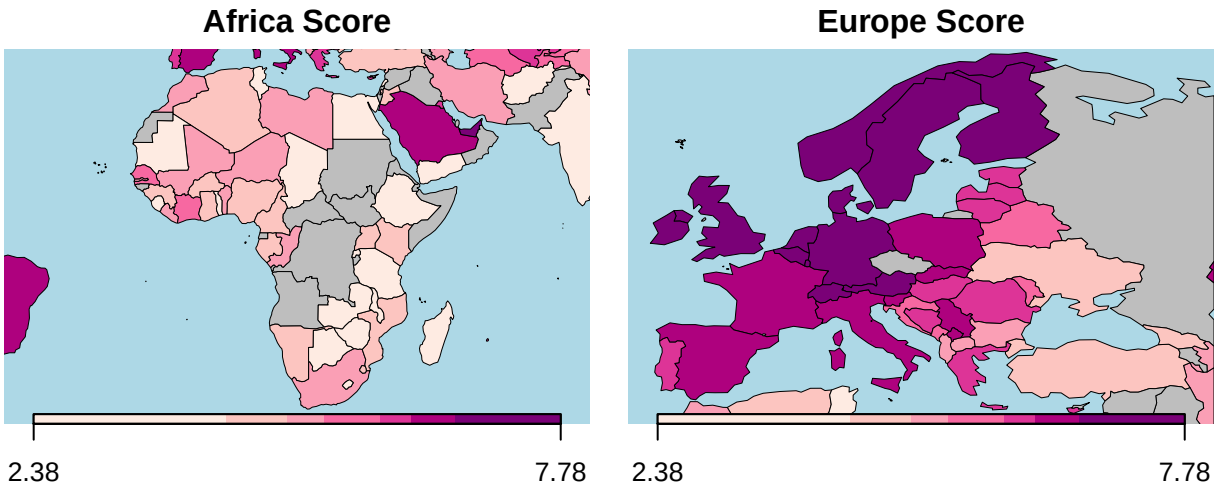
```
## 127 codes from your data successfully matched countries in the map
## 1 codes from your data failed to match with a country code in the map
##      failedCodes failedCountries
## [1,] NA          "North Cyprus"
## 116 codes from the map weren't represented in your data
```




```
par(op)
```

Mapa de la felicidad entre Europa y África:

```
colourPalette <- RColorBrewer::brewer.pal(7,"RdPu")
continentes <- c("Africa","Europe")
op <- par(mfcol=c(2,2),mai=c(0,0.2,0.3,0))
for (i in continentes) {
  map_by_cont <- mapCountryData(mapdf,
                nameColumnToPlot="Score",
                mapTitle = paste(i,"Score"),
                colourPalette = colourPalette,
                oceanCol = "lightblue",
                mapRegion=i,
                borderCol = "black",
                missingCountryCol="grey",
                addLegend=FALSE)
  do.call(addMapLegend, c(map_by_cont, legendWidth=0.5, legendIntervals="data",
                          legendMar=0))
}
par(op)
```



6 Resolución del problema

Primeramente, generamos en un fichero *csv* con los datos resultantes tras la integración, la selección de datos de interés (apartado 2) y la limpieza de los datos (apartado 3).

En términos generales, se han completado los valores faltantes mediante el **método kNN**, se ha seleccionado los campos que nos permiten realizar el análisis y se ha añadido un nuevo campo asociado al continente.

```
write.csv(df,"World_Happiness_Analysis-dataset-final.csv", row.names = FALSE)
```

Los resultados de los análisis nos han permitido responder a las preguntas que habíamos planteado en un inicio. Antes de realizar el estudio, hemos seleccionado grupos de datos por año y por continente que nos han facilitado el análisis. También, hemos comprobado la normalidad y homocedasticidad de estos grupos de datos para poder determinar qué tipo de pruebas estadísticas eran las más adecuadas.

¿Cuáles son los factores que contribuyen a la felicidad?

Para responder a esta pregunta, hemos realizado un estudio de correlaciones y posteriormente, hemos generado un modelo de regresión lineal múltiple.

Gracias al estudio de las **correlaciones** y el cálculo del coeficiente de correlación de *Kendall*, hemos determinado que el aspecto económico (PIB), la cohesión social entre los ciudadanos y la esperanza de vida del país, son los factores que más contribuyen positivamente a la felicidad de las personas. De hecho, estos tres factores también guardan relación entre ellos. La libertad de tomar decisiones, también influye positivamente, aunque, en menor medida. Por contra, la percepción de la corrupción influye ligera y negativamente en la percepción de la felicidad.

Por otro lado, la generación de un **modelo de regresión lineal** nos permite predecir el nivel de felicidad de un país conociendo los datos correspondientes al apoyo social, la esperanza de vida, la libertad de tomar decisiones y la percepción de la corrupción. Según la siguiente ecuación del modelo:

$$Score = -2,071 + 3,392\ Social.support + 0,061\ Life.expectancy + 1,875\ Freedom - 0,9\ Corruption$$

Queremos destacar, que el estudio podría haberse complementado con la incorporación de otros indicadores como el clima del país, la desigualdad social, los derechos civiles, la igualdad de género o la existencia de servicios públicos de sanidad y de educación.

¿Cómo evoluciona la felicidad a lo largo del tiempo?

Para responder a esta pregunta, hemos comparado la media de felicidad entre los años 2014 y 2019. Mediante un **contraste de hipótesis** hemos determinado que, a nivel global, somos más felices ahora, que hace 5 años. Sin embargo, hemos querido visualizar cómo ha evolucionado cada continente entre estos dos años, y hemos visto que Europa, África y ligeramente Asia sí que son más felices. No obstante, América del Norte, América del Sur y Oceanía eran más felices en 2014 que ahora.

¿Qué países o regiones son más felices?

En este punto nos hemos centrado concretamente en comparar los continentes de África y Europa mediante un **contraste de hipótesis**. Tras el resultado de la prueba, hemos comprobado que en el continente europeo son significativamente más felices que en el continente africano.

7 Contribución

Contribuciones	**Firma**
Investigación previa	Inés Caro Molina, Ángel Carrasco Núñez
Redacción de las respuestas	Inés Caro Molina, Ángel Carrasco Núñez
Desarrollo de código	Inés Caro Molina, Ángel Carrasco Núñez