Práctica 2:
Limpieza y análisis de datos

Mariana Tolivar Baqué
DNI: 78948019J

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El dataset escogido se obtiene de los datos recogidos del Índice de mujeres emprendedoras y el índice de emprendedores global, en 2015 en los países de la OCDE. El dataset consta de 51 observaciones y 9 columnas.

Vamos a tratar de analizar si existe alguna relación entre el índice de mujeres emprendedoras y otras variables que componen el dataset, en concreto el hecho de que el país pertenezca o no a la UE y que el país sea un país desarrollado o en desarrollo, para ello más adelante dividimos el conjunto de datos en subgrupos.

Consta de 9 columnas:

- La primera "No", es simplemente el número que se le asigna
- La segunda es "Country", en ella encontramos los países pertenecientes a la OCDE.
- La tercera recoge si el país está desarrollado o no, tenemos dos opciones "Developed", que hace referencia a que está desarrollado y "Developing" que nos indica que está en desarrollo. Se trata de una variable nominal.
- La cuarta columna es "European Union", que nos indica si un país es ("Member") o no es ("Not Member") miembro de la Unión Europea.
- La quinta columna "Currency" nos indica si utilizan el euro o la moneda nacional ("National Currency").
- La sexta es "Women Entrepreneurship Index", que nos muestra el índice de mujeres emprendedoras en el país.
- La séptima es "Entrepreneurship Index", el índice de emprendimiento del país.
- La octava es la tasa de inflación del país.
- Y, por último, la novena columna nos indica la participación femenina en el mundo laboral.

2. Integración y selección de los datos de interés a analizar

Si hablamos de integración de datos nos referimos al proceso de combinar los datos que obtenemos de distintas fuentes, para obtener una estructura de datos coherente y única que contenga mayor cantidad de información. Si hablamos de selección, nos referimos en cambio al filtrado de datos de interés, para reducir el número de datos. A continuación, en primer lugar, en el dataset escogido, voy a realizar el proceso de selección, para analizar tan solo los países desarrollados y miembros de la Unión Europea.

CÓDIGO EN R:

```
dataset3 <- read.csv("C:/Users/maria/OneDrive/Escritorio/Dataset3.csv", sep=";")

desarrollados <- subset(dataset3,Level.of.development == "Developed")

datafinal <- subset(desarrollados, European.Union.Membership == "Member")

datafinal
```

3. Limpieza de los datos:

3.1. ¿Los datos contienen casos o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

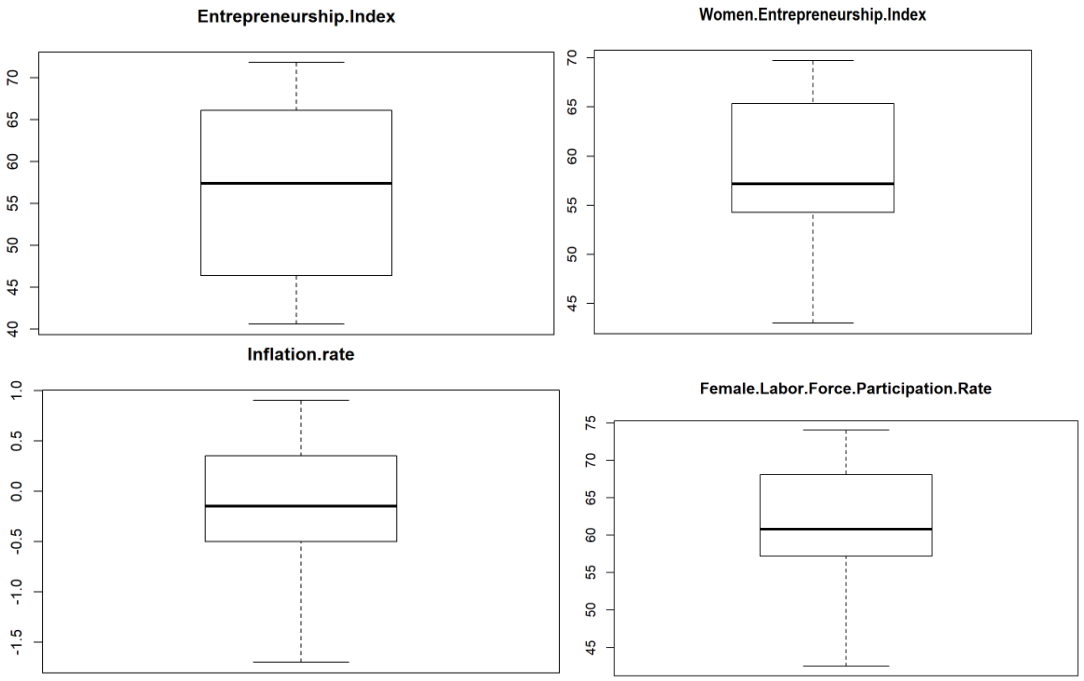
CÓDIGO EN R:

```
sapply(datafinal, function(x) sum(is.na(x)))
```

En este conjunto de datos en concreto, no obtenemos ningún caso o elemento vacío. En caso de que lo obtuviéramos, tendríamos distintas soluciones para ello. En primer lugar, si supiéramos la información y no nos llevara excesivo tiempo, podríamos introducir manualmente los registros que faltan. Podríamos reemplazarlos por una constante o por una etiqueta, como por ejemplo “Desconocido”. Otra de las opciones, es reemplazar los registros por la media o la mediana del atributo. O por último, podríamos recurrir a métodos para predecir los valores perdidos, métodos como las regresiones o los árboles de decisión.

3.2. Identificación y tratamiento de valores extremos

Los valores extremos son aquellos datos que se encuentran muy alejados de la distribución normal de una variable o población. Para comprobar si en alguna de las variables que tenemos en el dataset, tenemos valores extremos, realizamos diagramas de caja en las variables numéricas para comprobarlo.



Como vemos en los gráficos resultantes, no existen casos de valores extremos en el dataset, pero vamos a mencionar que haríamos en caso de que tuviésemos que gestionarlos.

En algunos casos, debemos fijarnos si se trata de errores o tal vez, que las unidades en las que está expresado no son las correctas (por ejemplo, si en una columna todo está reflejado en tanto

por ciento, y uno de los valores en tanto por mil), en ese caso deberíamos corregirlo. En caso de que sean errores, los trataríamos como valores perdidos, eliminándolos o corrigiéndolos mediante los métodos que hemos mencionado arriba.

4. Análisis de los datos: perseguimos explicar las características de los mismos para tratar de dar respuesta a las preguntas planteadas.

4.1. Selección de los grupos de datos que se quieren a analizar/comparar (planificación de los análisis a aplicar).

En este punto, ya hemos definido uno de los grupos que queremos usar para comparar, el de los países desarrollados pertenecientes a la unión europea (grupo 1). Además, vamos a crear un segundo subgrupo que es el de los países que no son miembros de la unión europea y que se encuentran en proceso de desarrollo (grupo 2). Queremos comparar el índice de mujeres emprendedoras en un grupo y otro.

CÓDIGO EN R

```
grupo1 <- subset(desarrollados, European.Union.Membership == "Member")

no.desarrollados <- subset(dataset3,Level.of.development == "Developing")

grupo2 <- subset(no.desarrollados, European.Union.Membership == "Not Member")
```

4.2. Comprobación de la normalidad y homogeneidad de la varianza

COMPROBACIÓN DE NORMALIDAD

Comprobamos si en el grupo 1 y en el grupo 2, la variable del índice de mujeres emprendedoras, sigue una distribución normal. Para ello realizamos el test de normalidad de Shapiro-Wilk.

H0: La muestra proviene de una distribución normal.

H1: La muestra no proviene de una distribución normal.

Si P < Alfa Se rechaza Ho

Si p >= Alfa No se rechaza Ho

CÓDIGO EN R

```
shapiro.test(grupo1[,6])

shapiro.test(grupo2[,6])
```

```
 Shapiro-Wilk normality test

data:  grupo1[, 6]
W = 0.9486, p-value = 0.3464
```

```
 Shapiro-Wilk normality test

data:  grupo2[, 6]
W = 0.96132, p-value = 0.4654
```

Como vemos en los resultados, concluiríamos que, en ambos casos, los datos siguen una distribución normal con un nivel de significación de 0.05.

COMPROBACIÓN DE LA HOMOCEDASTICIDAD

Al hablar de homocedasticidad estamos hablando de la comprobación de que las varianzas de ambos grupos son iguales.

CÓDIGO EN R

```
var.test(grupo1$Women.Entrepreneurship.Index, grupo2$Women.Entrepreneurship.Index)
```

El test no encuentra diferencias significativas entre las varianzas de ambos grupos.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos, y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

Queremos comprobar si el hecho de que el país pertenezca a la UE y se desarrollado infiere en el índice de mujeres emprendedoras.

Calculamos, por un lado, el índice de correlación entre el índice de mujeres emprendedoras y el nivel de desarrollo del país, y por otro lado la correlación entre el índice de mujeres emprendedoras y el hecho de que le país sea miembro o no de la Unión Europea.

```
dataset3$Level.of.development <- as.numeric(dataset3$Level.of.development =="Developed")

cor(dataset3$Women.Entrepreneurship.Index,dataset3$Level.of.development)

dataset3$European.Union.Membership <- as.numeric(dataset3$European.Union.Membership == "Member")

cor(dataset3$Women.Entrepreneurship.Index,dataset3$European.Union.Membership)
```

En el primero de los casos, obtenemos un índice de correlación de 0.86. Lo que nos indica que existe una fuerte relación entre el nivel de desarrollo del país y el índice de mujeres emprendedoras. Y en el segundo caso, obtenemos un índice de correlación de 0.62, que nos indica que, aunque es menos fuerte, también existe relación entre la variable que nos indica si el país forma parte de la Unión Europea y el índice de mujeres emprendedoras.

Contribuciones	Firma
Investigación previa	Mariana Tolivar Baqué
Redacción de las respuestas	Mariana Tolivar Baqué
Desarrollo código	Mariana Tolivar Baqué