Analisis Estadistico de Datos

Grupo 05

2025-05-04

Contents

Taller 02

Multivariada
Maestría en Energías Renovables
Escuela de Ingeniería, Ciencia y Tecnología
Universidad del Rosario
Integrantes:
Zahira Itzel González Cleves
Diego Alejandro Mejía Montañez
Daniel Felipe Russi Aragón
Iván Camilo Granados Niño
Andrés Alfonso Osorio Marulanda

Desarrollo Taller

```
library(corrplot)
library(devtools)
library(dplyr)
library(factoextra)
library(FactoMineR)
```

```r
library(ggExtra)
library(GGally)
library(ggplot2)
library(kableExtra)
library(knitr)
library(MASS)
library(plotly)
library(psych)
library(tidyverse)
library(usethis)
library(writexl)

#carga de datos separados con sep=;
df_pais <- read.table("/Users/aosoriom/Repos/MER_AEDatos/datasets/datos_taller_02.csv",sep=";",header =

# Visualiza la tabla de datos de mejor manera y con scroll
# Para tablas pequeñas usar: kable(head(mi_dataframe), caption = "Primeras filas del dataset")
#kable(head(df_pais), caption = "Primeras filas del dataset") %>%
#  scroll_box(width = "100%", height = "auto")
#summary(df_pais)
```

Normal Multivariada

Pregunta 1

$w_1 = 5$

Encuentre la estimación del vector de medias y la matriz de varianzas y covarianzas para 5 variables de su interés del conjunto de datos.

Vector de Medias

```r
datos_taller_02 <- df_pais
# Seleccionar las 5 variables de interés
vars_interes <- datos_taller_02[, c(12, 16, 17, 40, 58)]
# Estimación del vector de medias
vector_medias <- colMeans(vars_interes)
print(vector_medias)
```

```
##                     Kerosene_consumption_TBPD
##                                      8.224625
##              Motor_gasoline_consumption_TBPD
##                                    214.132432
##               Motor_gasoline_production_TBPD
##                                    211.467568
## Fossil_fuels_electricity_net_generation_BKWH
##                                    132.584084
##         Wind_electricity_net_generation_BKWH
##                                      6.084685
```

Unidades en las que se miden las variables a estudiar: - TBPD (Thousand Barrels Per Day) - BKWH (Billion Kilowatt-hours)

Se observa que el promedio de consumo de gasolina es muy alto en comparación con otras variables, y que la generación eólica todavía es baja en promedio.

Matriz de varianzas y covarianzas

```
# Estimación de la matriz de varianzas y covarianzas
matriz_covarianzas <- cov(vars_interes)
print(matriz_covarianzas)
```

```
##                                                Kerosene_consumption_TBPD
## Kerosene_consumption_TBPD                                      1111.20579
## Motor_gasoline_consumption_TBPD                                3652.59338
## Motor_gasoline_production_TBPD                                 4047.14314
## Fossil_fuels_electricity_net_generation_BKWH                   4495.08919
## Wind_electricity_net_generation_BKWH                             91.56329
##                                                Motor_gasoline_consumption_TBPD
## Kerosene_consumption_TBPD                                             3652.593
## Motor_gasoline_consumption_TBPD                                     790595.046
## Motor_gasoline_production_TBPD                                      834206.418
## Fossil_fuels_electricity_net_generation_BKWH                        318100.939
## Wind_electricity_net_generation_BKWH                                18042.266
##                                                Motor_gasoline_production_TBPD
## Kerosene_consumption_TBPD                                            4047.143
## Motor_gasoline_consumption_TBPD                                    834206.418
## Motor_gasoline_production_TBPD                                     890035.397
## Fossil_fuels_electricity_net_generation_BKWH                       335168.583
## Wind_electricity_net_generation_BKWH                                19190.037
##                                                Fossil_fuels_electricity_net_generation_BKWH
## Kerosene_consumption_TBPD                                                          4495.089
## Motor_gasoline_consumption_TBPD                                                  318100.939
## Motor_gasoline_production_TBPD                                                   335168.583
## Fossil_fuels_electricity_net_generation_BKWH                                     223824.714
## Wind_electricity_net_generation_BKWH                                              10183.494
##                                                Wind_electricity_net_generation_BKWH
## Kerosene_consumption_TBPD                                                   91.56329
## Motor_gasoline_consumption_TBPD                                          18042.26614
## Motor_gasoline_production_TBPD                                           19190.03732
## Fossil_fuels_electricity_net_generation_BKWH                             10183.49430
## Wind_electricity_net_generation_BKWH                                       563.17167
```

Las covarianzas (situadas fuera de la diagonal de la matriz) indican cómo dos variables interactuan. Los valores positivos indican relación directa o positiva, los valores negativos indican una relación indirecta o negativa, es decir, mientra una aumenta la otra disminuye. Y por último, si el valor es cero o cercano a cero, indica que las variables no tienen relación.

Podemos concluir que:

- Motor_gasoline_consumption y production: Covarianza “
approx” 834206.42 m Covarianza muy alta y positiva: los países que consumen más gasolina también tienden a producir más.

- Kerosene y gasoline consumption: Covarianza “
approx” 3652.59 Covarianza: positiva pero pequeña → Existe relación entre las variables, pero no muy fuerte.
- Wind electricity vs. cualquier otra: Covarianzas bajas (ej. con fósil “
approx” 10183.49, con gasolina “
approx” 18042.27), la generación eólica no tiene una fuerte relación lineal con las demás variables.

A continuación, se muestra la diagonal de la matriz de varianzas y covarianzas, que representan la varianza de cada variable, es decir, qué tanto se dispersan los datos respecto a su media.

```
# Varianza de cada una de las variables
print (diag(matriz_covarianzas))
```

```
##                     Kerosene_consumption_TBPD
##                                     1111.2058
##              Motor_gasoline_consumption_TBPD
##                                   790595.0460
##               Motor_gasoline_production_TBPD
##                                   890035.3975
## Fossil_fuels_electricity_net_generation_BKWH
##                                   223824.7141
##         Wind_electricity_net_generation_BKWH
##                                      563.1717
```

- La producción y consumo de gasolina tienen altísima varianza, lo que sugiere grandes diferencias entre países.
- La generación eólica tiene baja varianza, lo que indica que la mayoría de los países tienen valores similares (probablemente bajos).
- La generación eléctrica por fósiles también muestra alta varianza, lo cual refleja diferencias estructurales entre países en sus sistemas eléctricos.

ACP

Pregunta 2

$w_2 = 5$

Realice un análisis descriptivo univariado y bivariado del conjunto de datos. Considere nuevamente las mismas 5 variables de su interés para este punto.

Análisis univariado

Resumen estadístico

```
summary(vars_interes)
```

```
##  Kerosene_consumption_TBPD Motor_gasoline_consumption_TBPD
##  Min.   :  0.000           Min.   :   0.8
##  1st Qu.:  0.000           1st Qu.:  13.0
##  Median :  0.500           Median :  33.0
##  Mean   :  8.225           Mean   : 214.1
##  3rd Qu.:  2.300           3rd Qu.: 108.5
##  Max.   :298.000           Max.   :8921.0
##  Motor_gasoline_production_TBPD Fossil_fuels_electricity_net_generation_BKWH
##  Min.   :   0.0                 Min.   :   0.00
##  1st Qu.:   5.6                 1st Qu.:   4.25
##  Median :  37.0                 Median :  21.00
##  Mean   : 211.5                 Mean   : 132.58
##  3rd Qu.: 106.5                 3rd Qu.:  80.00
##  Max.   :9571.0                 Max.   :3985.00
##  Wind_electricity_net_generation_BKWH
##  Min.   :  0.000
##  1st Qu.:  0.000
##  Median :  0.100
##  Mean   :  6.085
##  3rd Qu.:  1.450
##  Max.   :182.000
```

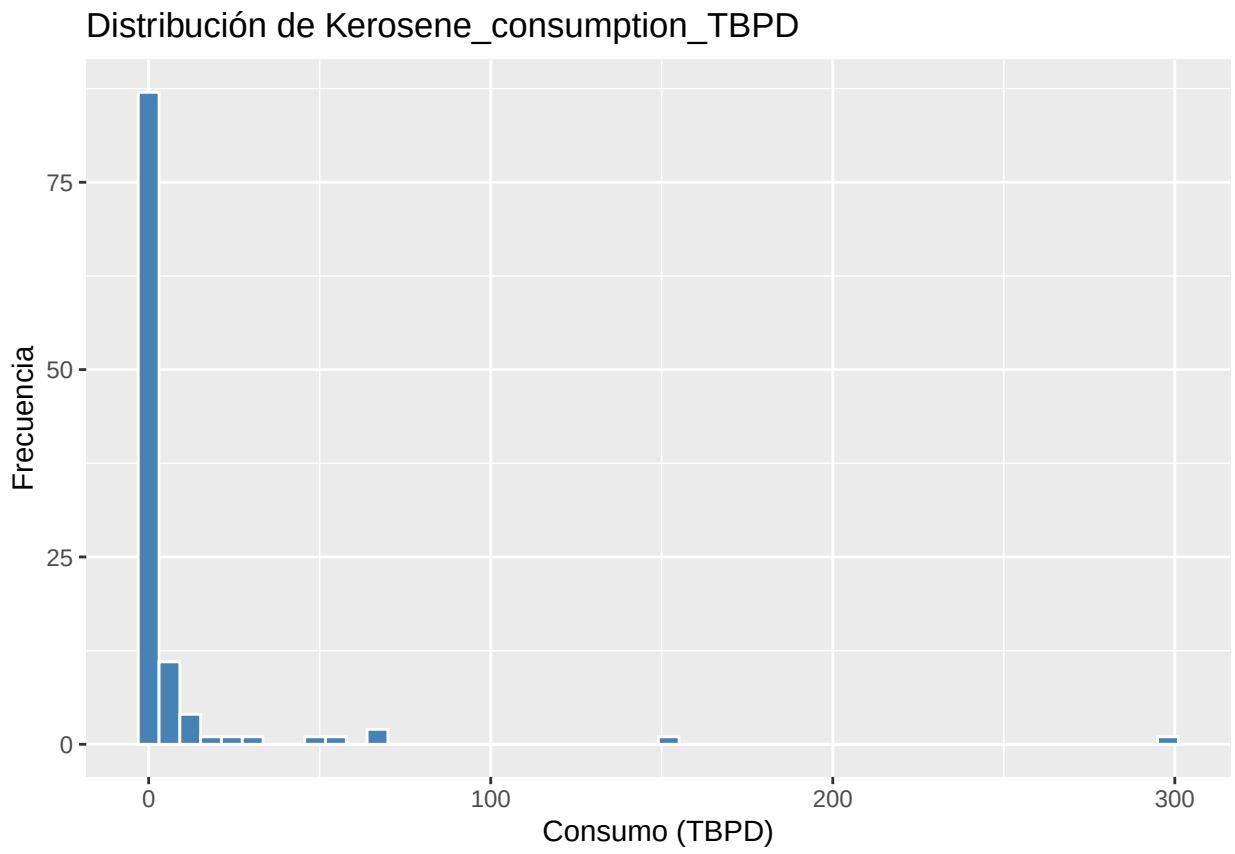
Desviación estándar para cada variable

```
sapply(vars_interes, sd, na.rm = TRUE)
```

```
##                     Kerosene_consumption_TBPD
##                                      33.33475
##              Motor_gasoline_consumption_TBPD
##                                     889.15412
##               Motor_gasoline_production_TBPD
##                                     943.41687
## Fossil_fuels_electricity_net_generation_BKWH
##                                     473.10117
##         Wind_electricity_net_generation_BKWH
##                                      23.73124
```

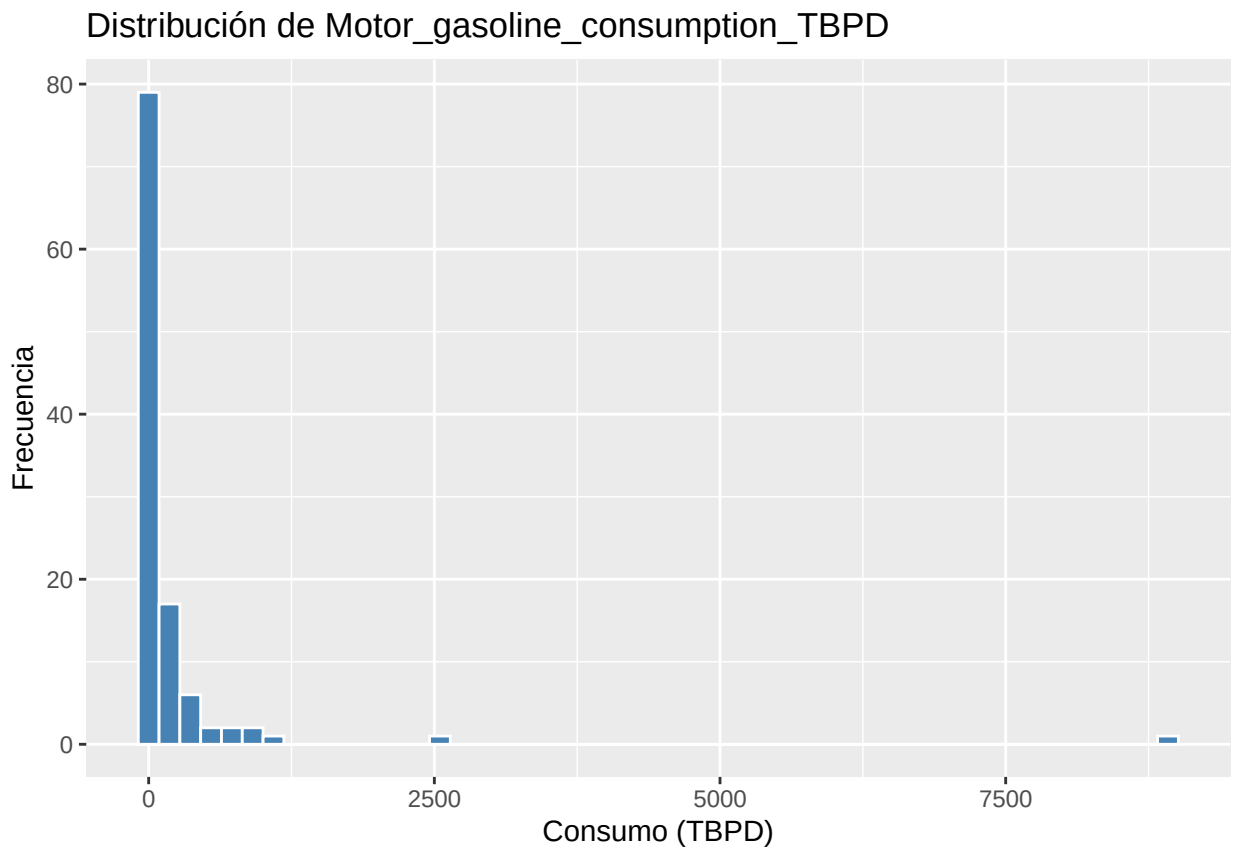
Histograma de distribución para cada variable Histograma de distribución Kerosene_consumption_TBPD

```
ggplot(vars_interes, aes(x = Kerosene_consumption_TBPD)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(title = "Distribución de Kerosene_consumption_TBPD",
       x = "Consumo (TBPD)",
       y = "Frecuencia")
```



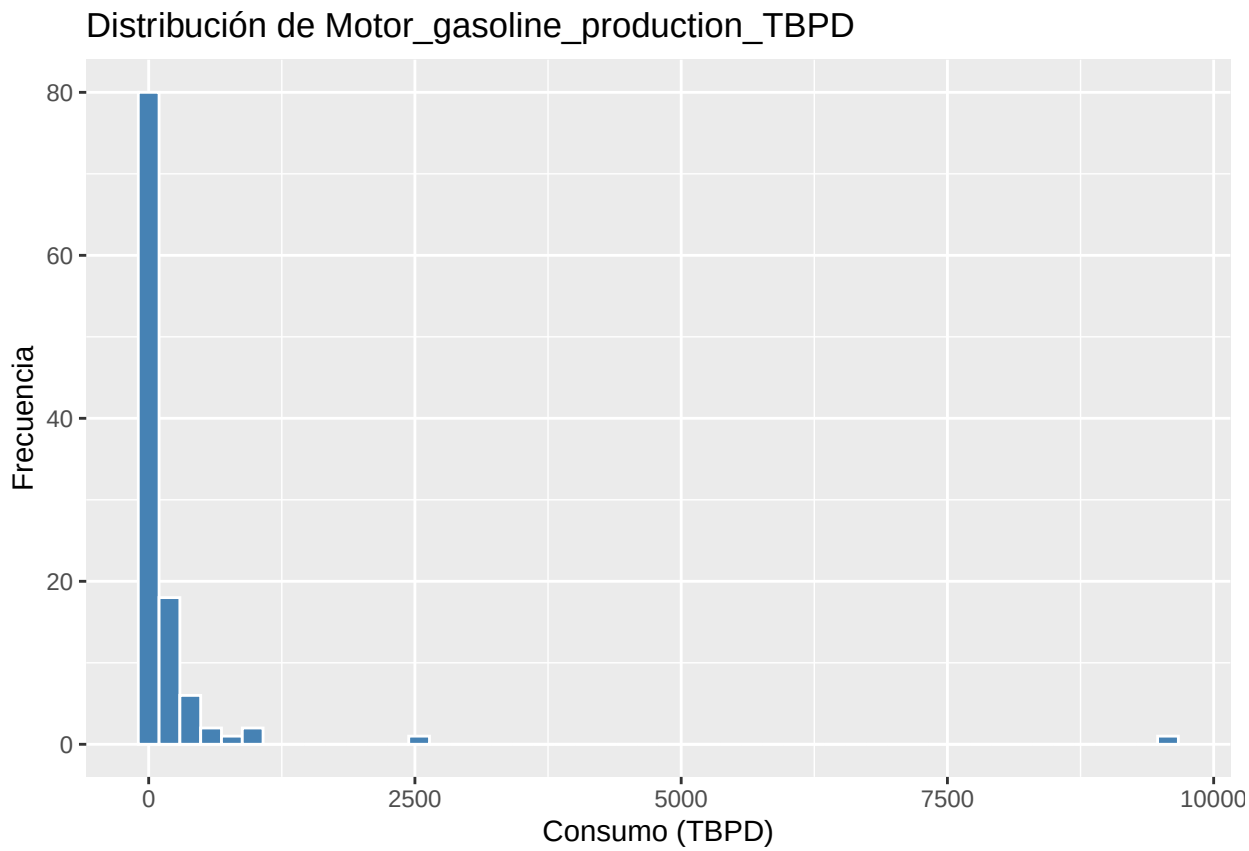
Histograma de distribución Motor_gasoline_consumption_TBPD

```
ggplot(vars_interes, aes(x = Motor_gasoline_consumption_TBPD)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(title = "Distribución de Motor_gasoline_consumption_TBPD",
       x = "Consumo (TBPD)",
       y = "Frecuencia")
```



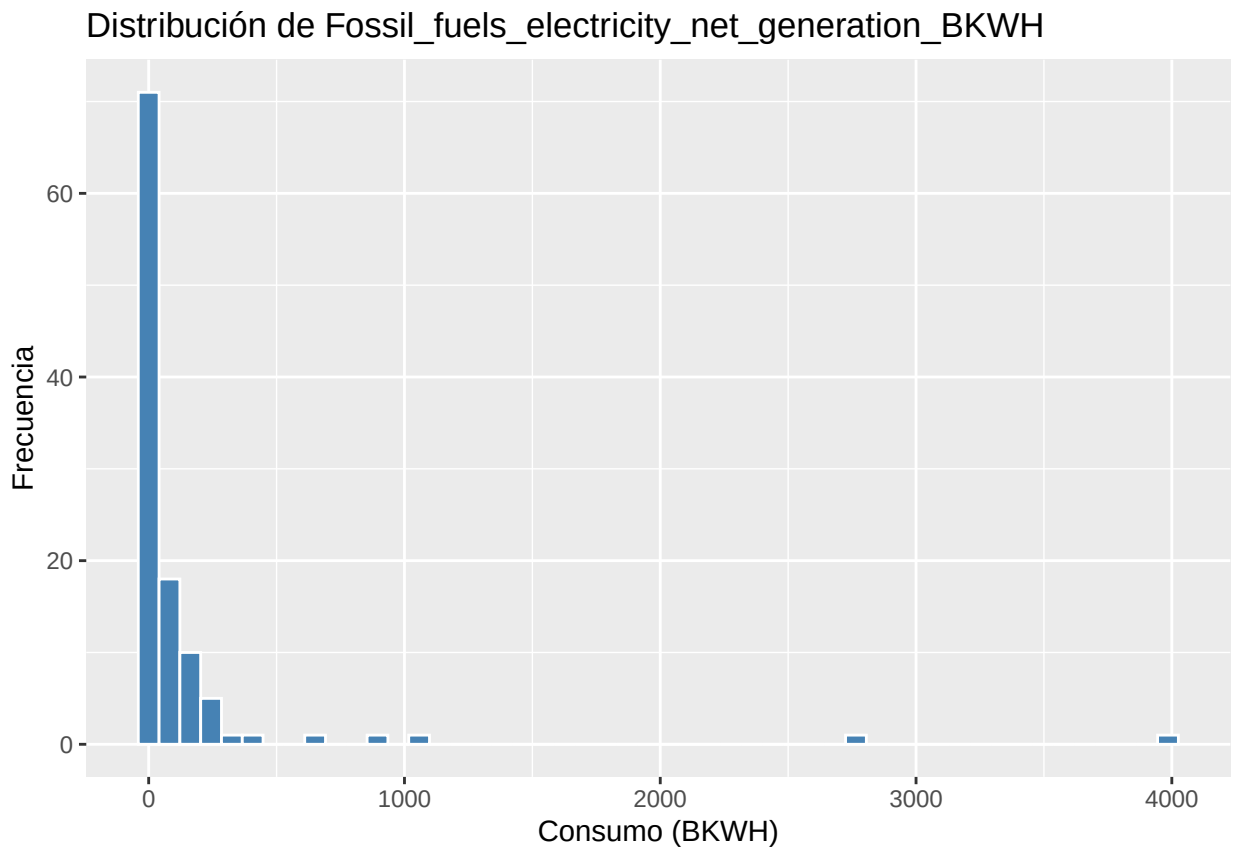
Histograma de distribución Motor_gasoline_production_TBPD

```
ggplot(vars_interes, aes(x = Motor_gasoline_production_TBPD)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(title = "Distribución de Motor_gasoline_production_TBPD",
       x = "Consumo (TBPD)",
       y = "Frecuencia")
```



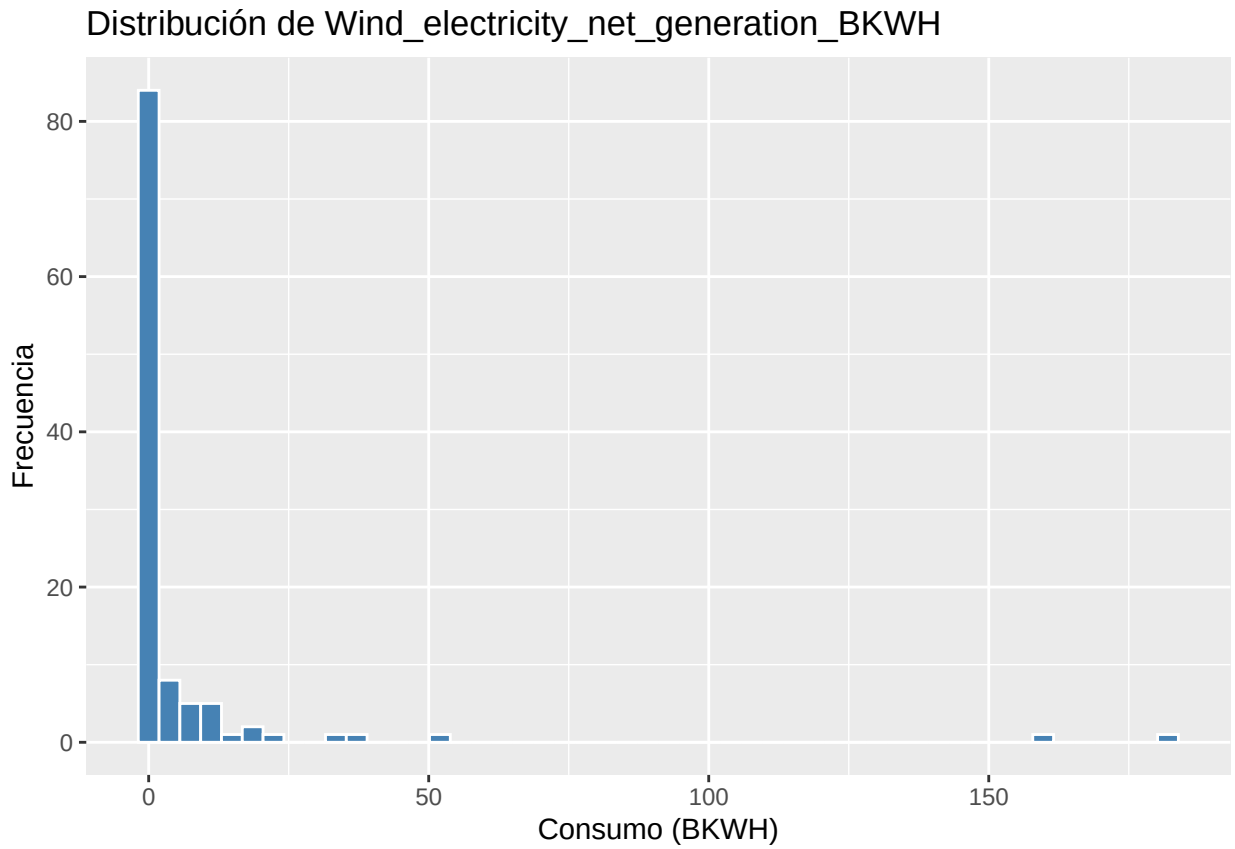
Histograma de distribución Fossil_fuels_electricity_net_generation_BKWH

```
ggplot(vars_interes, aes(x = Fossil_fuels_electricity_net_generation_BKWH)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(title = "Distribución de Fossil_fuels_electricity_net_generation_BKWH",
       x = "Consumo (BKWH)",
       y = "Frecuencia")
```



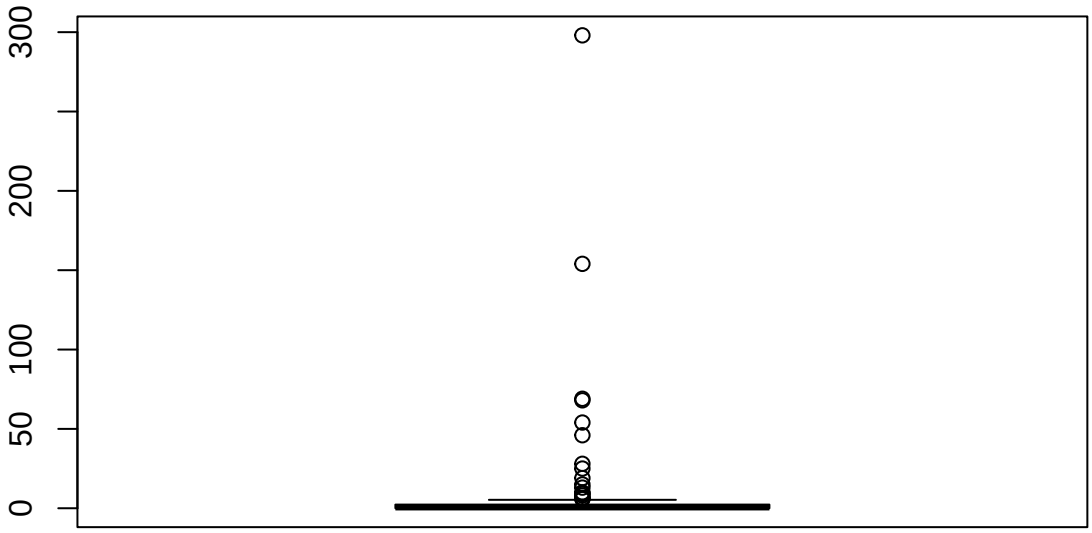
Histograma de distribución Fossil_fuels_electricity_net_generation_BKWH

```
ggplot(vars_interes, aes(x = Wind_electricity_net_generation_BKWH)) +
  geom_histogram(fill = "steelblue", color = "white", bins = 50) +
  labs(title = "Distribución de Wind_electricity_net_generation_BKWH",
       x = "Consumo (BKWH)",
       y = "Frecuencia")
```



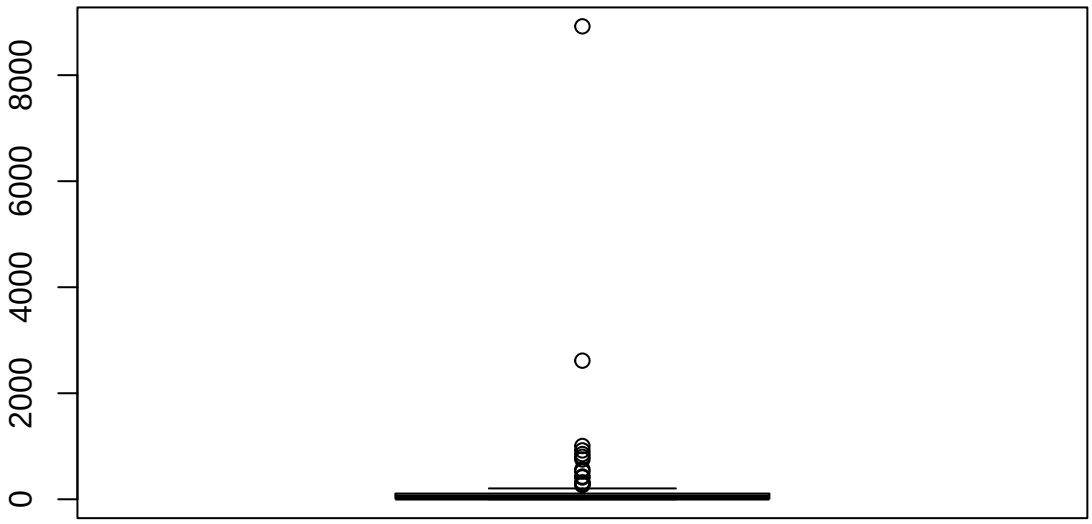
Boxplots Boxplot Kerosene_consumption_TBPD

```
boxplot(vars_interes$Kerosene_consumption_TBPD)
```



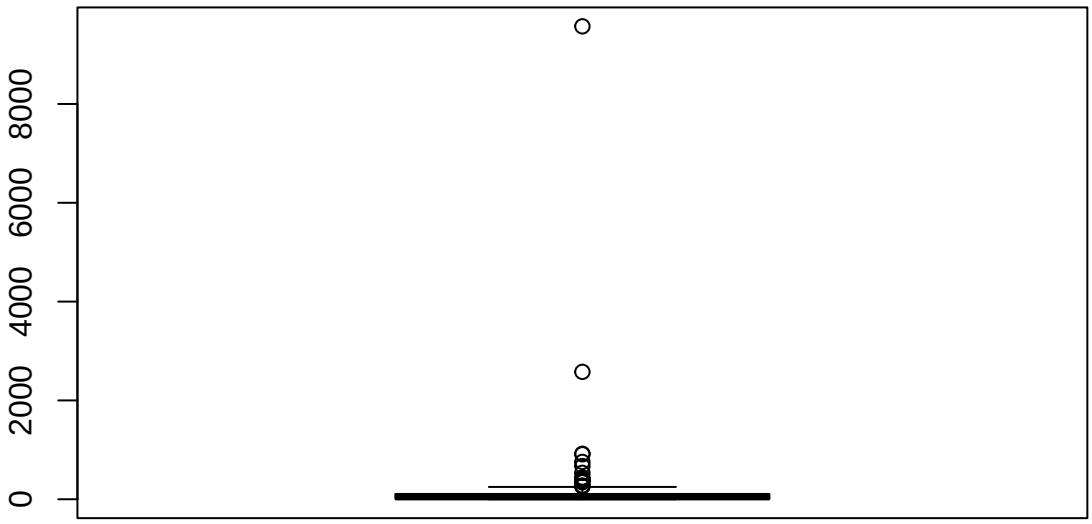
Boxplot Motor_gasoline_consumption_TBPD

```
boxplot(vars_interes$Motor_gasoline_consumption_TBPD)
```



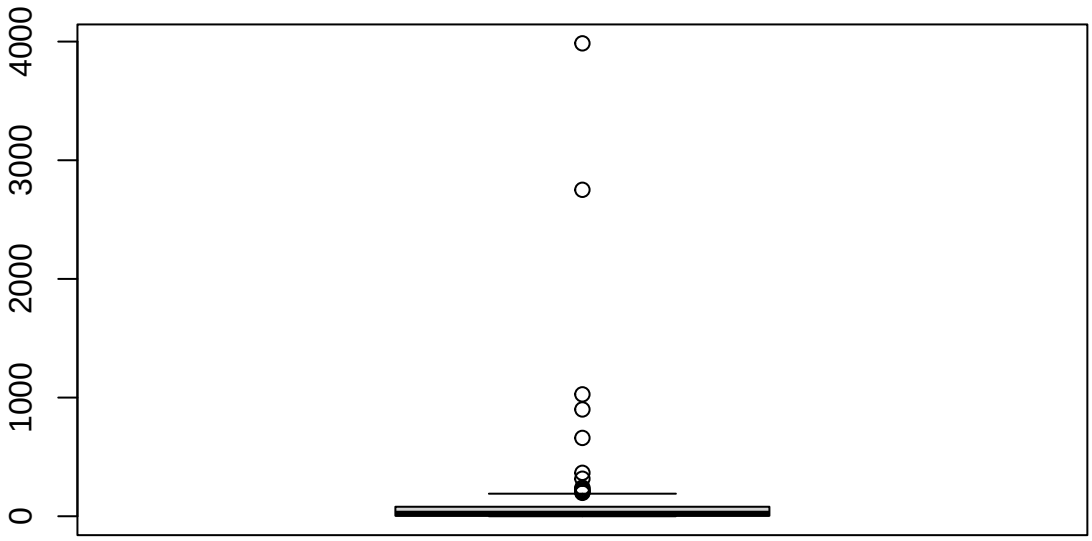
Boxplot Motor_gasoline_production_TBPD

```
boxplot(vars_interes$Motor_gasoline_production_TBPD)
```



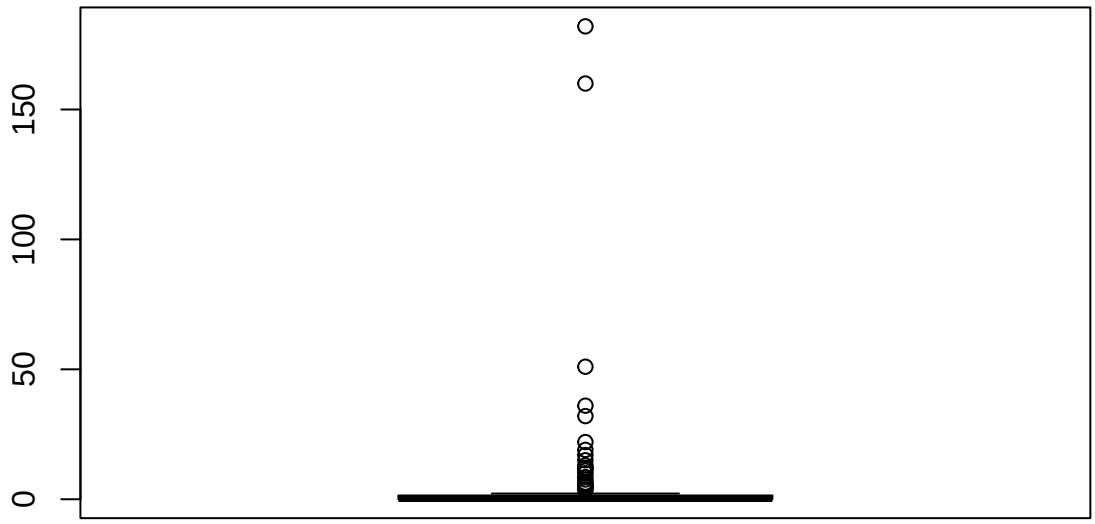
Boxplot Fossil_fuels_electricity_net_generation_BKWH

```
boxplot(vars_interes$Fossil_fuels_electricity_net_generation_BKWH)
```



Boxplot Wind_electricity_net_generation_BKWH

```
boxplot(vars_interes$Wind_electricity_net_generation_BKWH)
```



Análisis multivariado

La matriz de correlación indica la fuerza y dirección de las relaciones lineales entre las variables de un conjunto de datos.

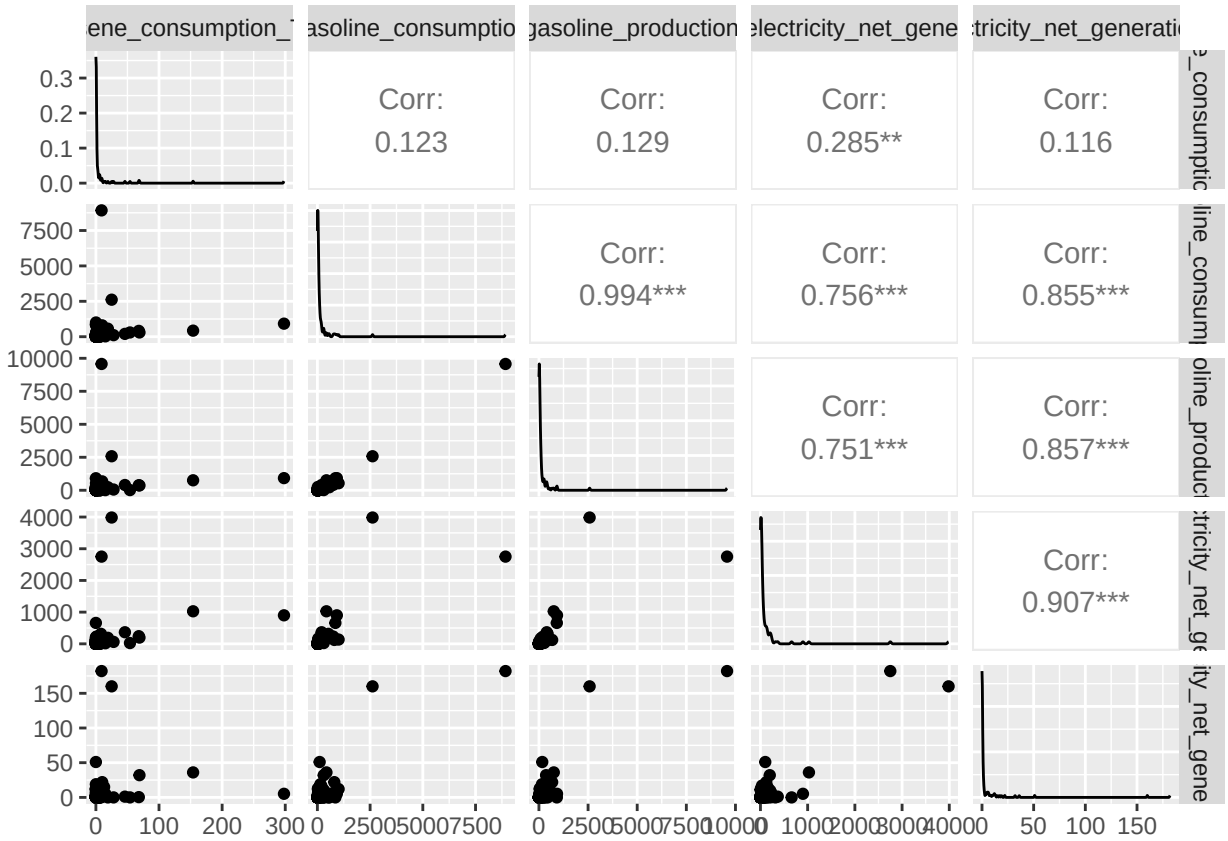
```
correlaciones <- cor(vars_interes, use = "complete.obs")
print(correlaciones)
```

```
##                                          Kerosene_consumption_TBPD
## Kerosene_consumption_TBPD                                1.0000000
## Motor_gasoline_consumption_TBPD                          0.1232330
## Motor_gasoline_production_TBPD                           0.1286909
## Fossil_fuels_electricity_net_generation_BKWH             0.2850277
## Wind_electricity_net_generation_BKWH                     0.1157454
##                                          Motor_gasoline_consumption_TBPD
## Kerosene_consumption_TBPD                                      0.1232330
## Motor_gasoline_consumption_TBPD                                1.0000000
## Motor_gasoline_production_TBPD                                 0.9944726
## Fossil_fuels_electricity_net_generation_BKWH                   0.7561952
## Wind_electricity_net_generation_BKWH                           0.8550542
##                                          Motor_gasoline_production_TBPD
## Kerosene_consumption_TBPD                                     0.1286909
## Motor_gasoline_consumption_TBPD                               0.9944726
## Motor_gasoline_production_TBPD                                1.0000000
## Fossil_fuels_electricity_net_generation_BKWH                  0.7509407
```

```
## Wind_electricity_net_generation_BKWH                                0.8571400
##                                                Fossil_fuels_electricity_net_generation_BKWH
## Kerosene_consumption_TBPD                                                         0.2850277
## Motor_gasoline_consumption_TBPD                                                   0.7561952
## Motor_gasoline_production_TBPD                                                    0.7509407
## Fossil_fuels_electricity_net_generation_BKWH                                      1.0000000
## Wind_electricity_net_generation_BKWH                                              0.9070316
##                                                Wind_electricity_net_generation_BKWH
## Kerosene_consumption_TBPD                                                 0.1157454
## Motor_gasoline_consumption_TBPD                                           0.8550542
## Motor_gasoline_production_TBPD                                            0.8571400
## Fossil_fuels_electricity_net_generation_BKWH                              0.9070316
## Wind_electricity_net_generation_BKWH                                      1.0000000
```

- Alta correlación entre Motor_gasoline_consumption_TBPD y Motor_gasoline_production_TBPD: 0.9945. Esto indica una relación casi perfecta. Los países que más gasolina producen también la consumen mucho.
- Alta correlación entre entre Fossil_fuels_electricity_net_generation_BKWH y Wind_electricity_net_generation_BK 0.9070. Los países que generan más electricidad con fósiles también tienen alta generación eólica.
- Correlación fuerte entre Motor_gasoline_production_TBPD y Wind_electricity_net_generation_BKWH: 0.8571. Los paises con alta producción de gasolina, también tienen una alta generación neta de electricidad eólica.
- Kerosene_consumption_TBPD tiene correlaciones débiles con todas las demás (< 0.29).

```
ggpairs(vars_interes)
```



La mayoría de las variables tienen distribuciones sesgadas hacia valores bajos con algunos valores muy altos (ver diagramas de dispersión).

Esto sugiere que la mayoría de los países tienen baja producción o consumo, mientras que unos pocos tienen valores extremos.

Pregunta 3

$w_3 = 8$
El conjunto de datos incluye 57 variables, analizarlas de forma individual podría representar un gasto computacional y de tiempo bastante agotador; por lo cuál una técnica de reducción de dimensionalidad sería ideal. Ejecute un ACP sobre estos datos, compare la contribución de las variables sobre el primer plano factorial de un ACP normado (escalado) y uno sin normalizar (sin escalar).

El conjunto de datos incluye 57 variables, analizarlas de forma individual podría representar un gasto computacional y de tiempo bastante agotador; por lo cual una técnica de reducción de dimensionalidad sería ideal. Ejecute un ACP sobre estos datos, compare la contribución de las variables sobre el primer plano factorial de un ACP normado (escalado) y uno sin normalizar (sin escalar).

```r
#carga de datos separados con sep=;
pais = read.table("/Users/aosoriom/Repos/MER_AEDatos/datasets/datos_taller_02.csv",sep=";",header = TRU

#se crea el data frame país
datos<-as.data.frame(pais)

#nueva variable y se le quita la 1ra columna
columnas_a_convertir <- names(datos)[-1]
```

```r
#transformación de datos a numéricos str(Datos)
datos[columnas_a_convertir] <- lapply(datos[columnas_a_convertir], as.numeric)
# str(datos) --> se comenta para no sacar todos los datos, se debe quitar.

#selección de las 5 variables
select_data <- datos[, c("Refined_petroleum_products_consumption_TBPD", "Refined_petroleum_products_proc
summary(select_data)
```

```
##  Refined_petroleum_products_consumption_TBPD
##  Min.   :   15.0
##  1st Qu.:   71.5
##  Median :  202.0
##  Mean   :  811.4
##  3rd Qu.:  750.0
##  Max.   :19100.0
##  Refined_petroleum_products_production_TBPD Electricity_exports_BKWH
##  Min.   :    0.0                            Min.   : 0.000
##  1st Qu.:   39.5                            1st Qu.: 0.000
##  Median :  172.0                            Median : 0.600
##  Mean   :  777.3                            Mean   : 5.158
##  3rd Qu.:  555.0                            3rd Qu.: 5.550
##  Max.   :19653.0                            Max.   :75.000
##  Electricity_imports_BKWH Electricity_installed_capacity_MK
##  Min.   : 0.000           Min.   :   0.40
##  1st Qu.: 0.000           1st Qu.:   3.50
##  Median : 0.700           Median :  11.00
##  Mean   : 5.671           Mean   :  51.63
##  3rd Qu.: 7.550           3rd Qu.:  33.00
##  Max.   :67.000           Max.   :1380.00
```

```r
#se excluye la primera columna
columnas_numericas <- names(datos)[-1]

#se aplica pca a datos normalizados o escalados

#"prcom()p" es la forma rápida de implementar un PCA sobre un conjunto de datos, es necesaria la biblio
pca_normalizado <- prcomp(datos[, columnas_numericas], center = TRUE, scale = TRUE)

#se aplica pca a datos sin normalizar
pca_sin_normalizar <- prcomp(datos[, columnas_numericas], center = FALSE, scale = FALSE)

#se comparan las contirbuciones de las variables seleccionadas
contribucion_normalizado <- pca_normalizado$sdev^2
contribucion_sin_normalizar <- pca_sin_normalizar$sdev^2

#creación de tabla de datos para luego ver las contribuciones
contribuciones_df <- data.frame(
  Contribucion_Normalizado = contribucion_normalizado,
  Contribucion_Sin_Normalizar = contribucion_sin_normalizar)
#print(format(contribuciones_df, scientific=FALSE))
head(contribuciones_df, n=10)
```

```
##    Contribucion_Normalizado Contribucion_Sin_Normalizar
```

```
## 1              35.3158793              27524045.470
## 2               4.4099873               4204746.774
## 3               3.9984709                470085.023
## 4               2.7056000                134152.350
## 5               2.4548859                109320.329
## 6               1.8678717                 42942.129
## 7               1.3074352                 22302.380
## 8               1.1151341                 13868.073
## 9               0.9019626                 12746.905
## 10              0.7565284                  9601.909
```

Con las variables seleccionadas que presentan valores bajos con respecto a otros países se observa que las contribuciones tienen poca influencia en la variación de los datos después de ser normalizados, por lo que se puede decir que dichas variables no contribuyen con aportes significativos al primer componente principal con datos escalados. Para los datos sin normalizar, las contribuciones son mayores comparadas con las normalizadas debido a los rangos o escalas de las variables seleccionadas no normalizadas.

Pregunta 4

$w_4 = 8$
Una vez calculado el PCA normado, ¿qué número de componentes principales deberíamos seleccionar? ¿Bajo que criterio seleccionó este número de componentes?

Se toma como referencia dos criterios 1) el Criterio de Kaiser que define que se deben conservar o seleccionar aquellos cuyos valores propios son iguales o superiores a 1 y 2) Porcentaje de la varianza acumulada, en el cual se calcula el porcentaje de la varianza de cada componente y se realiza el calculo del acumulado con el aporte de cada componente hasta que se evidencia un valor significativo de este porcentaje acumulado.

1) Criterio Kaiser: se selecciona hasta el componente principal 8, el cual es $PC_1 = 1.11$ y cumple son ser mayor que 1. En otras palabras, se seleccionana del componente 1 al 8 por ser mayores a 1.

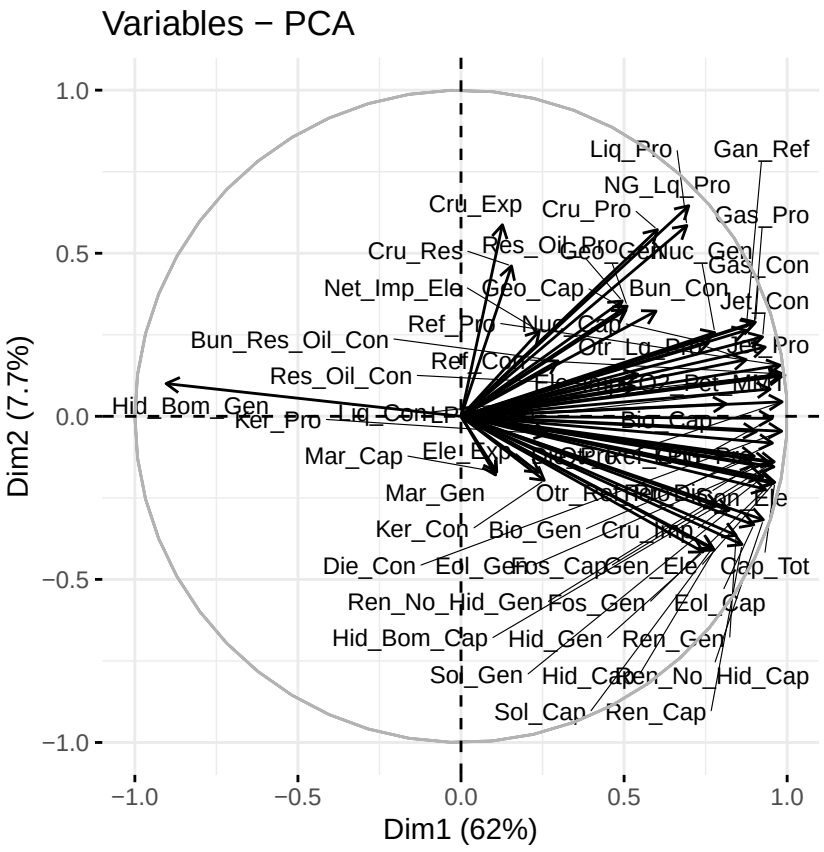
2) Criterio de % de Varianza acumulada:usualmente se toman los componentes principales que totalicen una varianza acumulada entre 70% y 90%. Para este ejercicio se usara el principio de Pareto identificando los componentes que alcancen una varianza acumulada igual o superior al 80%, los cuales serian de los componentes del 1 al 4; estos 4 componentes equivalen al 7% del total y representan el 81.8% de la varianza acumulada.

Pregunta 5

$w_5 = 8$
Analice la representación de las variables originales sobre las dos primeras componentes en el primer plano factorial. ¿Que conclusiones podemos sacar de este análisis?

En el siguiente diagrama fasorial se grafican los 2 componentes principales, la misma es un poco difusa porque se están graficando las 57 variables.

```
#agrupa los datos por componentes principales escalados y los guarda en la variable asignada acp_c (par
# La sintaxis [, -1] selecciona todas las filas (espacio antes de la coma)
# y todas las columnas excepto la columna en el índice 1 (el -1 después de la coma)
datos_rec <- read.table("/Users/aosoriom/Repos/MER_AEDatos/datasets/datos_taller_02_rec.csv",sep=";",hea
datos_solo_numericos <- datos_rec[, -1]
acp_c <-prcomp(datos_solo_numericos, scale = TRUE)

fviz_pca_var(acp_c,axes=c(1,2),repel=TRUE, labelsize=3, arrowsize=0.5, arrowwidth=0.5)
```

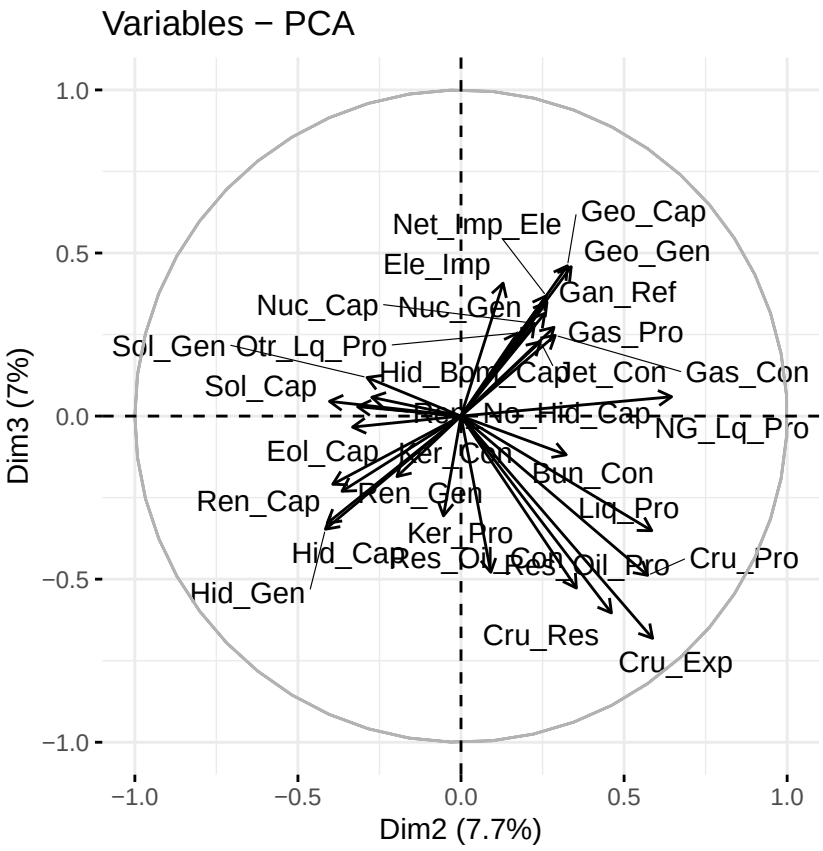


Por esto se procede a graficar únicamente las 30 variables que más aportan a los datos como se aconseja en la sección de ayuda, mediante `fviz_pca_var(acp_c,axes=c(1,2),repel=TRUE,select.var = list(contrib = 30))` obteniendo la siguiente grafica, de esta podemos concluir que dimension 1 o PCA 1, agrupa el 62% de los datos y el PCA 2 el 7,7%, siendo necesario apoyarnos en mas PCAs para lograr tomar un 80% de los datos y tener una mejor representacion de los datos de las 57 variables. Por otro lado, las flechas asignadas para cada variable indican la contribución en cada PCA, dando como resultado que en su mayoria aportan a la dimension 1 con una longitud muy parecida o agrupada entre cada variable dado que eliminamos las que menos contribuían, teniendo 2 casos visiblemente diferentes como lo son las variables Hid_Bom_Gen (Hydroelectric_pumped_storage_electricity_net_generation_BKWH), la cual se ubica en el cuadrante 2 aportanto al PCA1 negativo, y la variable NG_Lq_Pro (Natural_gas_plant_liquids_production_TBPD), la cual aporta en la dimension 1 y 2.

En contraste al ejercicio de clase, en este contamos con una cantidad mayor de componentes relevantes para el análisis, por lo cual, realice el mismo gráfico anterior pero ahora sobre el plano factorial conformado por los componentes 2 y 3. ¿Qué conclusiones podemos sacar de este gráfico?

Utilizando la linea de codigo `fviz_pca_var(acp_c, axes = c(2, 3),repel=TRUE,select.var = list(contrib = 30))`, Los datos de las variables en estos componentes tienen una menor contribución y estos se puede evidenciar en la magnitud de las flechas, junto a esto, se puede ver que hay una mayor distribución en los diferentes cuadrantes del diagrama fasorial, es importante resaltar como el PCA3 tiene 7% de los datos, con estos 3 PCAs agruparíamos el 76.7% de los datos, sería necesario tomar más componentes principales para llegar al 80%, lo cual nos daría una representación objetiva de los datos.

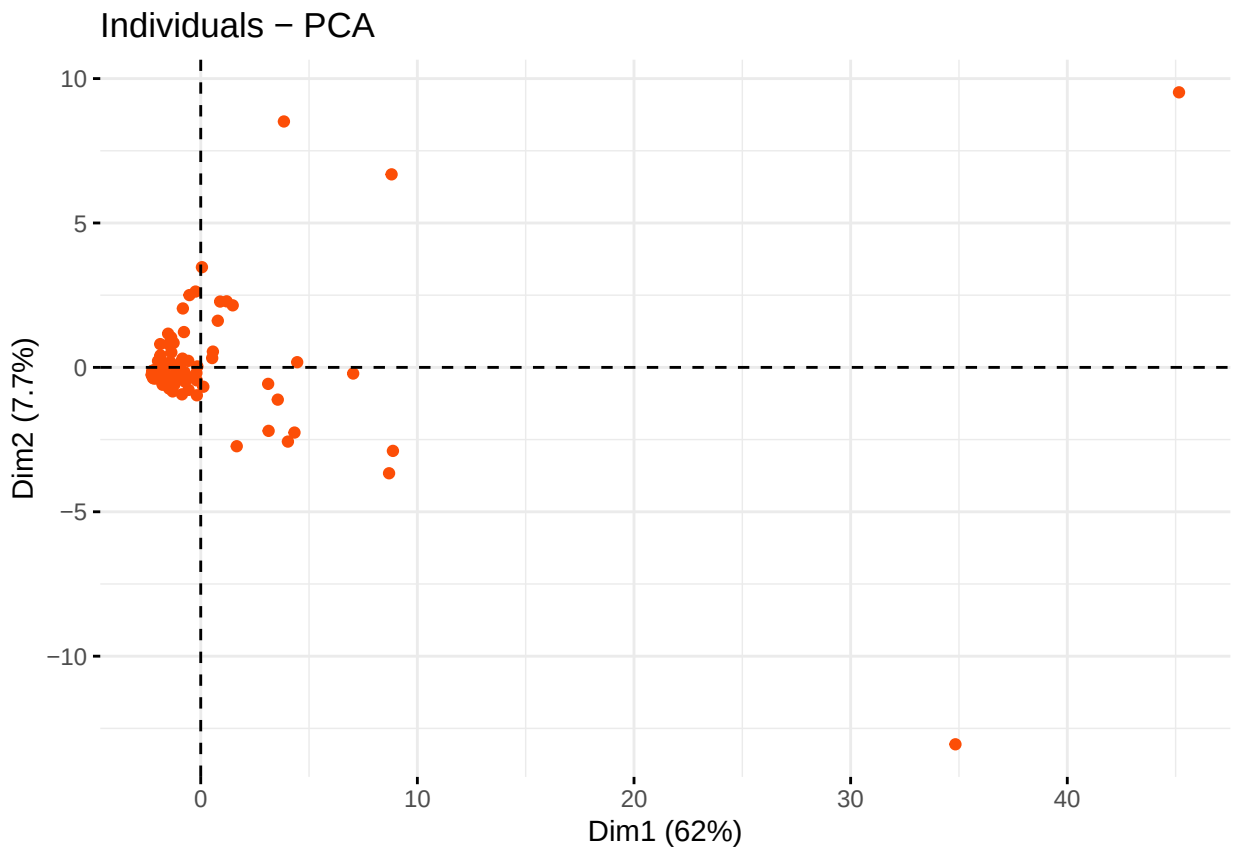
```
fviz_pca_var(acp_c, axes = c(2, 3),repel=TRUE,select.var = list(contrib = 30))
```



Pregunta 6

$w_6 = 10$
Ahora examine con detenimiento el mapa de individuos sobre los planos factoriales que conforman las componentes (1,2) y (2,3). ¿Qué conclusiones puede sacar según la cercanía de algunas UE?

```
res.pca <- PCA(datos[,-1], graph = FALSE)
fviz_pca_ind(res.pca, geom.ind = "point",
             col.ind = "#FC4E07",
             axes = c(1, 2),
             pointsize = 1.5)
```



Dim1 (62%):

- Este eje captura la mayor parte de la variabilidad, sugiriendo que una única dimensión ordena eficientemente a los países.
- Rango de valores: Los países se distribuyen desde -10 (extremo izquierdo) hasta 40 (extremo derecho), indicando una fuerte polarización.

Dim2 (7.7%):

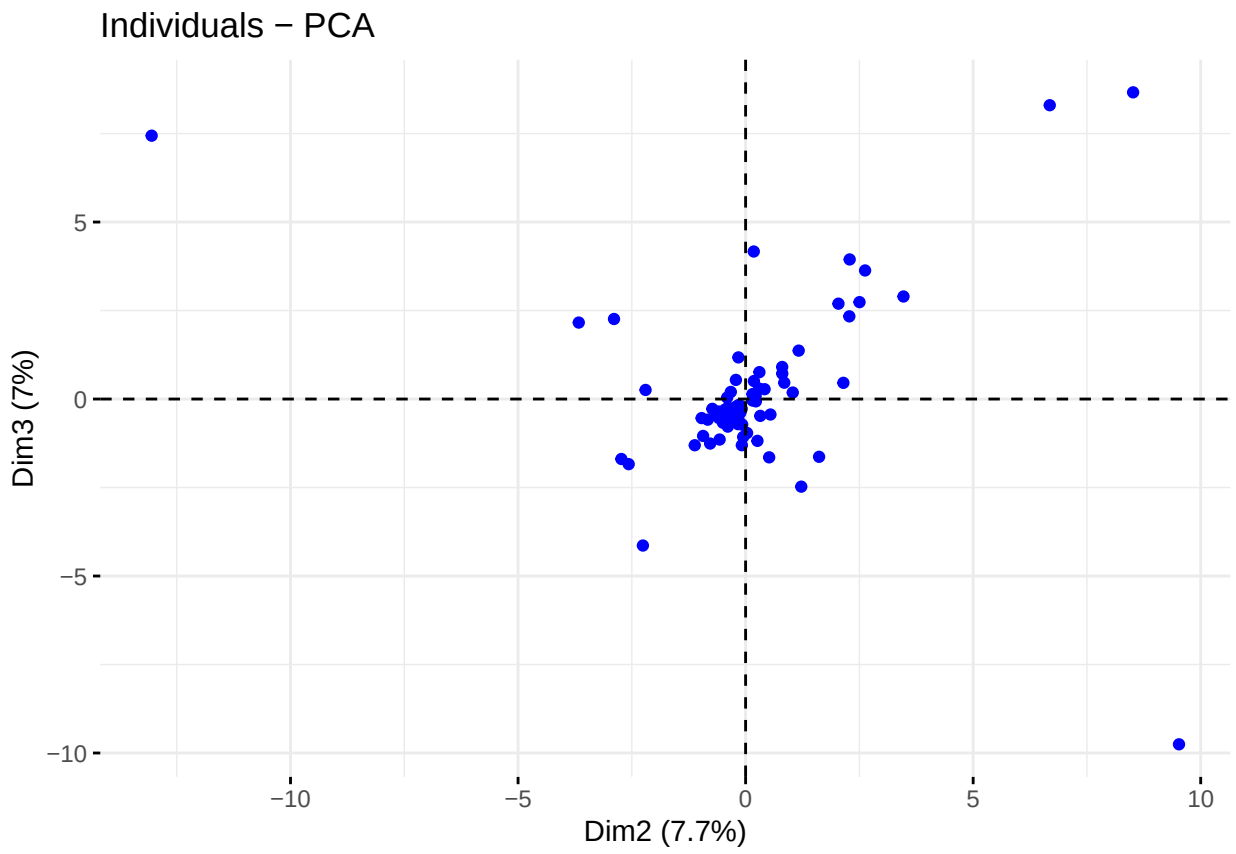
- Su contribución parece marginal, por lo que el análisis se centra en Dim1 (62%).

Distribución de Países

Patrón de dispersión:

- Extremo derecho: Países con características inusuales.
- Centro (valores cercanos a 0): Países con perfiles energéticos similares.
- Extremo izquierdo (valores negativos): Países con patrones atípicos.

```
fviz_pca_ind(res.pca, geom.ind = "point",
             col.ind = "blue",
             axes = c(2, 3),
             pointsize = 1.5)
```



Dim3 (7%) y Dim2 (7.7%):

Ambos componentes explican una proporción baja de varianza total (14.7% combinada). Esto sugiere que:

- La estructura subyacente de los datos es alta-dimensional (muchas variables influyentes no capturadas en estos ejes).

Distribución de Países

- Los puntos (países) están dispersos en un rango de -10 a 10 en ambos ejes, sin agrupamientos claros. Esto indica:
- No hay perfiles dominantes que agrupen a múltiples países.
- Outliers potenciales: Países en extremos (ej., cerca de (-10, 10)) podrían ser casos atípicos.

Pregunta 7

$w_7 = 10$
Como hemos observado en clase, el ACP es una técnica bastante sensible a datos atípicos, ejecute nuevamente el ACP retirando del conjunto de datos a Estados Unidos, China, Arabia Saudi y Rusia. ¿En que cambia el ACP al excluir estos países?. ¿Se perciben clusters de países con mayor claridad?. Calcule únicamente el ACP normado.

```r
#se retiran los países propuestos
retirados <- c("UnitedStates", "China", "SaudiArabia", "Russia")

#creación de la nueva tabla de datos que ya no tienen los países retirados
paises_filtrados <- datos[!datos$Country %in% retirados, ]
columnas_numericas <- names(paises_filtrados)[-1]

pca_normalizado <- prcomp(paises_filtrados[, columnas_numericas], center = TRUE, scale = TRUE)
pca_sin_normalizar <- prcomp(paises_filtrados[, columnas_numericas], center = FALSE, scale = FALSE)

contribucion_normalizado <- pca_normalizado$sdev^2
contribucion_sin_normalizar <- pca_sin_normalizar$sdev^2

contribuciones_df <- data.frame(
  Contribucion_Normalizado = contribucion_normalizado,
  Contribucion_Sin_Normalizar = contribucion_sin_normalizar
)

varianza_explicada <- pca_normalizado$sdev^2

varianza_acumulada <- cumsum(varianza_explicada) / sum(varianza_explicada)

head(contribuciones_df, n=10)
```

```
##    Contribucion_Normalizado Contribucion_Sin_Normalizar
## 1                27.8996281                 4098283.841
## 2                 5.8387300                 1445037.166
## 3                 4.6204396                   86544.805
## 4                 3.7753312                   41993.071
## 5                 2.8210627                   26544.386
## 6                 2.2286043                   14081.658
## 7                 2.0557117                   10294.495
## 8                 1.6001717                    8629.487
## 9                 1.3866329                    5178.434
## 10                0.8667378                    3747.178
```

Los países retirados presentan datos atípicos ya que son grandes consumidores de recursos energéticos como la electricidad o el petróleo y por otra parte, también se tienen grandes productores de aquellos recursos.

Las variaciones de las contribuciones de los datos normalizados si se comparan con los primeros resultados cuando estaban todos los países, muestran menos saltos, de manera similar sucede con los datos sin normalizar.

Lo anterior debido a que los países retirados presentaban datos altos si se comparaban con otros, lo que hacía que la escala de datos fuera mucho mayor. Al ser retirados estos países con datos atípicos, contribuye a buscar más eficazmente países con comportamientos más comunes entre sí y permite sean agrupados de o subdivididos de manera geográfica, por ejemplo.

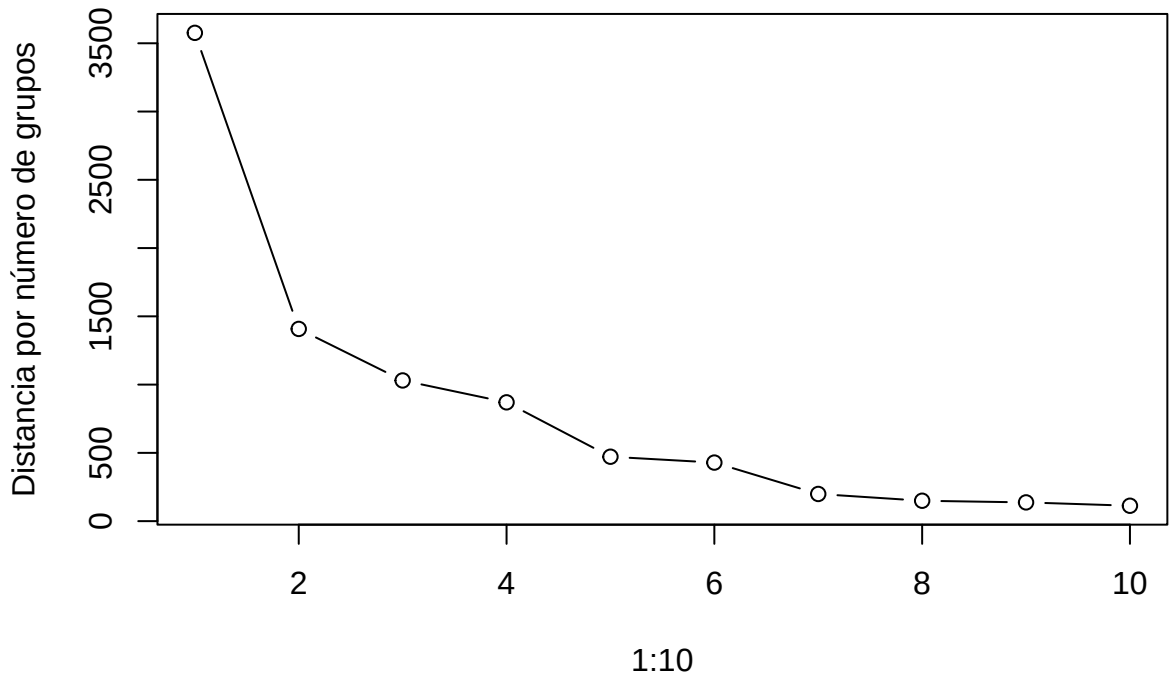
Clustering

Los puntos 8 a 11 se realizarán excluyendo de la base de datos los países atípicos mencionados en el punto 7 (es decir, retirando 'UnitedStates','China','SaudiArabia','Russia').

Pregunta 8

$w_8 = 8$
De acuerdo a lo aprendido en la clase de agrupamiento, utilice el primer plano factorial para determinar de manera aproximada el número de grupos en el análisis. ¿Cuántos clusters espera que existan en el conjunto de datos?

```
PrimerasComponentes <- pca_normalizado$x[, c(1, 2)] #Seleccion de la ubicación de cada país en las dos
distancia_total <- numeric(10)  # Crear vector de longitud 10

for (i in 1:10)
  {
    kmeans_model <- kmeans(PrimerasComponentes, centers = i) #agrupar los datos ubicados en el plano en
    distancia_total[i] <- kmeans_model$tot.withinss #sumar la distancia total de la agrupación i
  }
  plot(1:10,distancia_total,ylab="Distancia por número de grupos", xlab = '1:10', type='b')
```



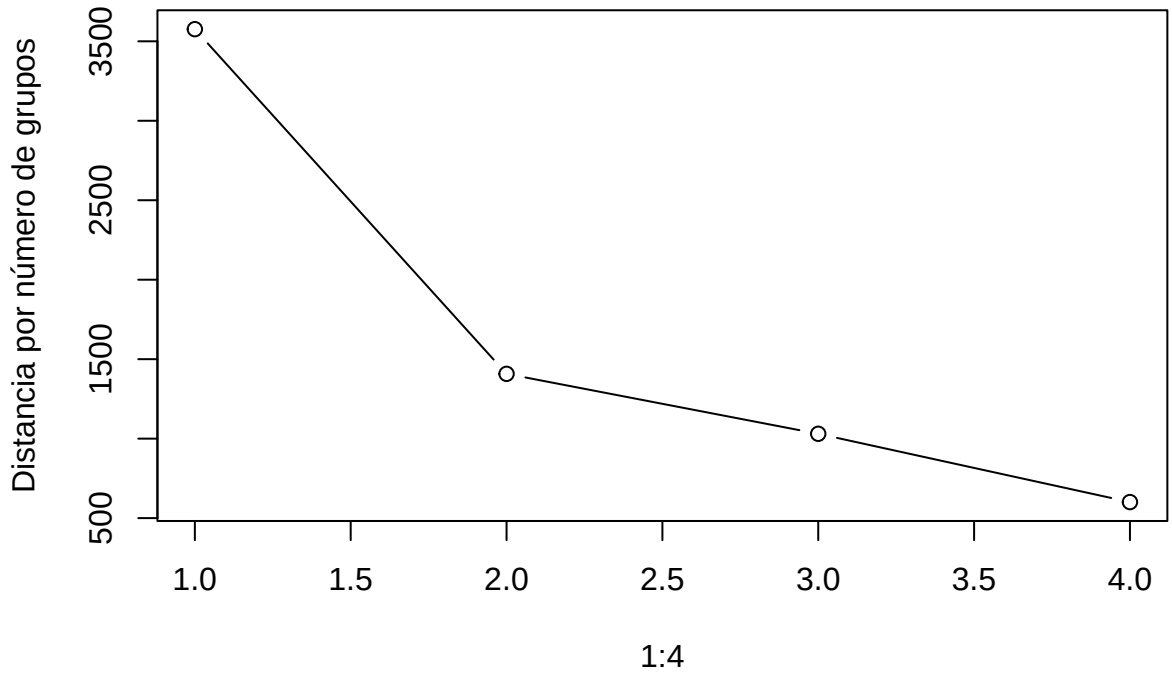
De acuerdo con la imagen anterior, se pueden esperar 4 clústeres o grupos ya que con esta cantidad se obtiene una reducción de distancia significativa y con cantidades mayores las reducciones son pequeñas.

Pregunta 9

$w_9 = 12$
Mediante el método de reducción de varianza explicada, determine un número fijo de clusters para su análisis. No tiene que ser el mismo número que el elegido por su compañero, recuerde que ninguno conoce las etiquetas reales de los grupos de los países.

Respecto de la curva de codo de obtenida en el punto anterior, se considera que un numero ideal de clusters para analisis es k=4, dado que, luego de este, la disminucion se reduce significativamente la mejor en valores superiores de k no es representativa

```
#PrimerasComponentes <- pca_normalizado$x[, c(1, 2)] #Seleccion de la ubicación de cada país en las dos
distancia_total <- numeric(4)  # Crear vector de longitud 10
for (i in 1:4)
  {
    kmeans_model <- kmeans(PrimerasComponentes, centers = i) #agrupar los datos ubicados en el plano en
    distancia_total[i] <- kmeans_model$tot.withinss #sumar la distancia total de la agrupación i
  }
  plot(1:4,distancia_total,ylab="Distancia por número de grupos", xlab = '1:4', type='b')
```



Pregunta 10

$w_{10} = 12$
Determine mediante k-means y aglomeración jerárquica (usando enlace promedio) la clasificación en grupos para los países de estudio. Considere todas las variables de estudio ¿son diferentes los resultados de los dos análisis cluster?

```
set.seed(100)
kmeans_m <- kmeans(pca_normalizado$x, centers = 4)# aplicar K-means con 4 clusters
clusters <- kmeans_m$cluster
```

```r
 # mostrar tabla con los 4 clusters
 table(clusters)
```

```
## clusters
##  1  2  3  4
##  8 10  2 87
```

```r
# aplicar metodo de aglomeracion promedio
d_aglomerados <- hclust(dist(pca_normalizado$x), method = "average")

#lleva el arbol jerarquico a 4 grupos para poder comparar con el resultado de kmeans
clusters_aglomerados <- cutree(d_aglomerados, k = 4)

#resultados de de los 4 grupos aglomerados
table(clusters_aglomerados)
```

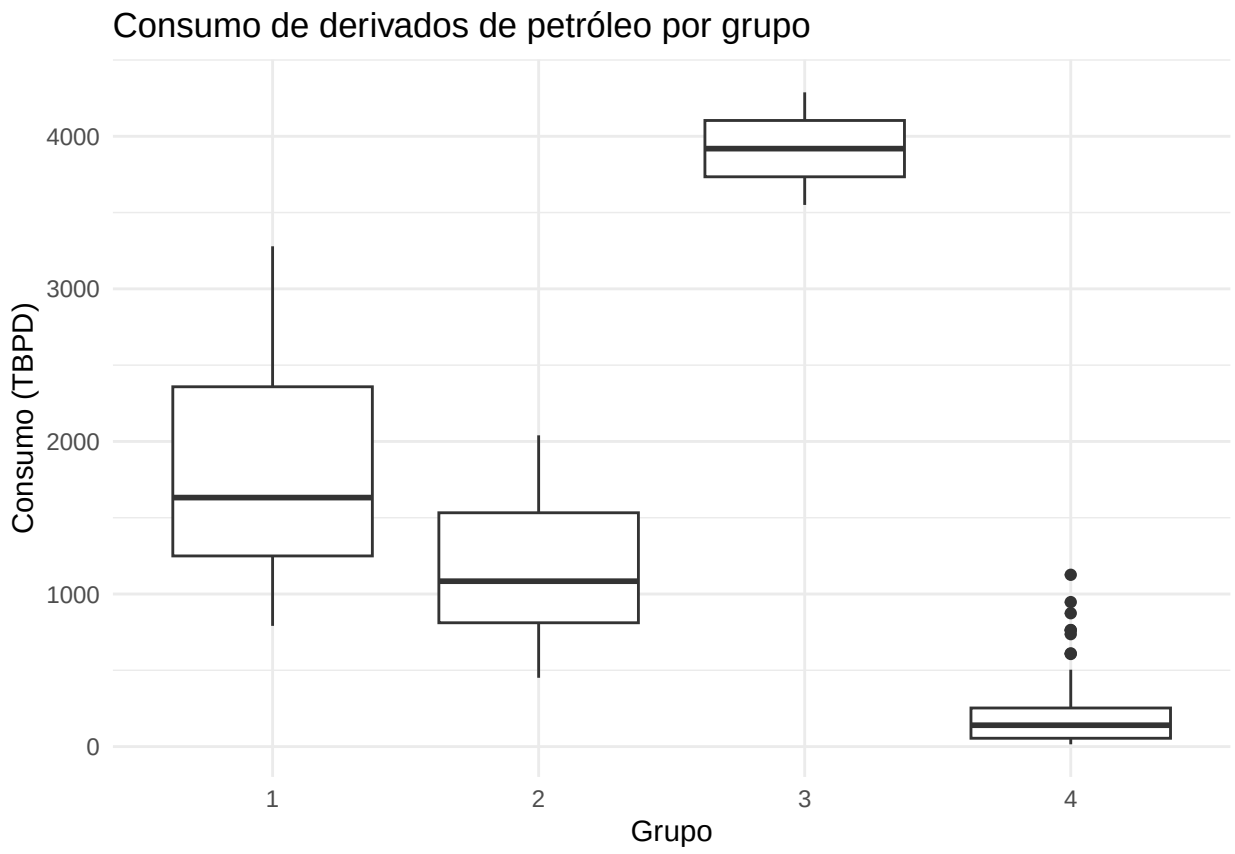
```
## clusters_aglomerados
##   1   2   3   4
## 101   2   2   2
```

Si, Son diferentes los resultados entre k-means y aglomeración jerárquica promedio, podemos observar en los resultados que en el caso de K-means el cluster 3 contiene 87 paises, y en el caso de agrupamiento promedio, el grupo 3 contiene 2 paises. El agrupamiento jerarquico promedio agrupa 101 países en el grupo 1, como se vio en clase, los diferentes métodos de agrupamiento (clusters) difieren al momento de tomar las distancias entre los diferentes valores para irlos agrupando, lo cual queda demostrado en la comparación de los resultados anteriores.

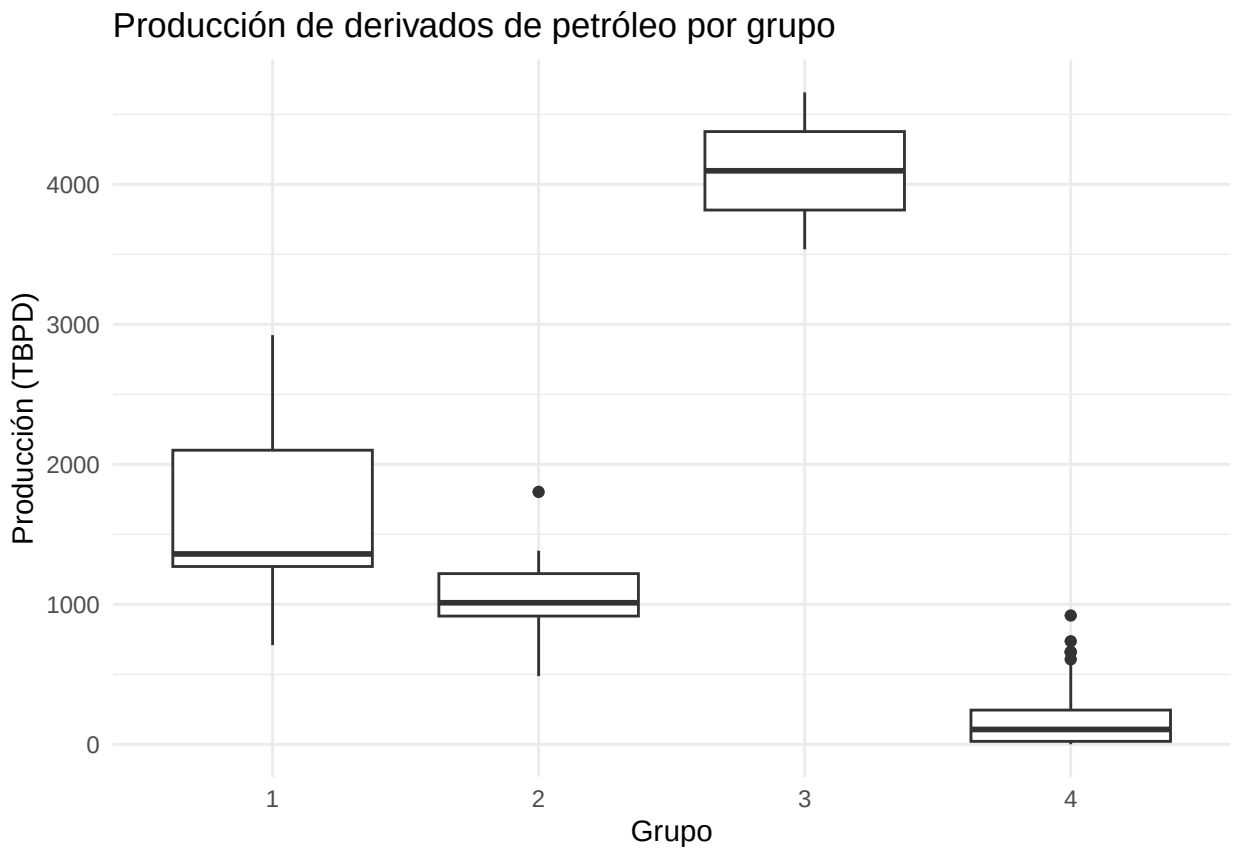
Pregunta 11

$w_{11} = 12$
Grafique mediante boxplot la distribución de las 5 variables seleccionadas en el punto 2 diferenciando por los grupos encontrados mediante k-means (es decir, obtenga un boxplot por variable y por grupo: si por ejemplo k-means indica un total de 3 grupos, realice 3 boxplots, uno por grupo, para cada una de las 5 variables). ¿Observa diferencias importantes entre las distribuciones por grupos de la misma variable?

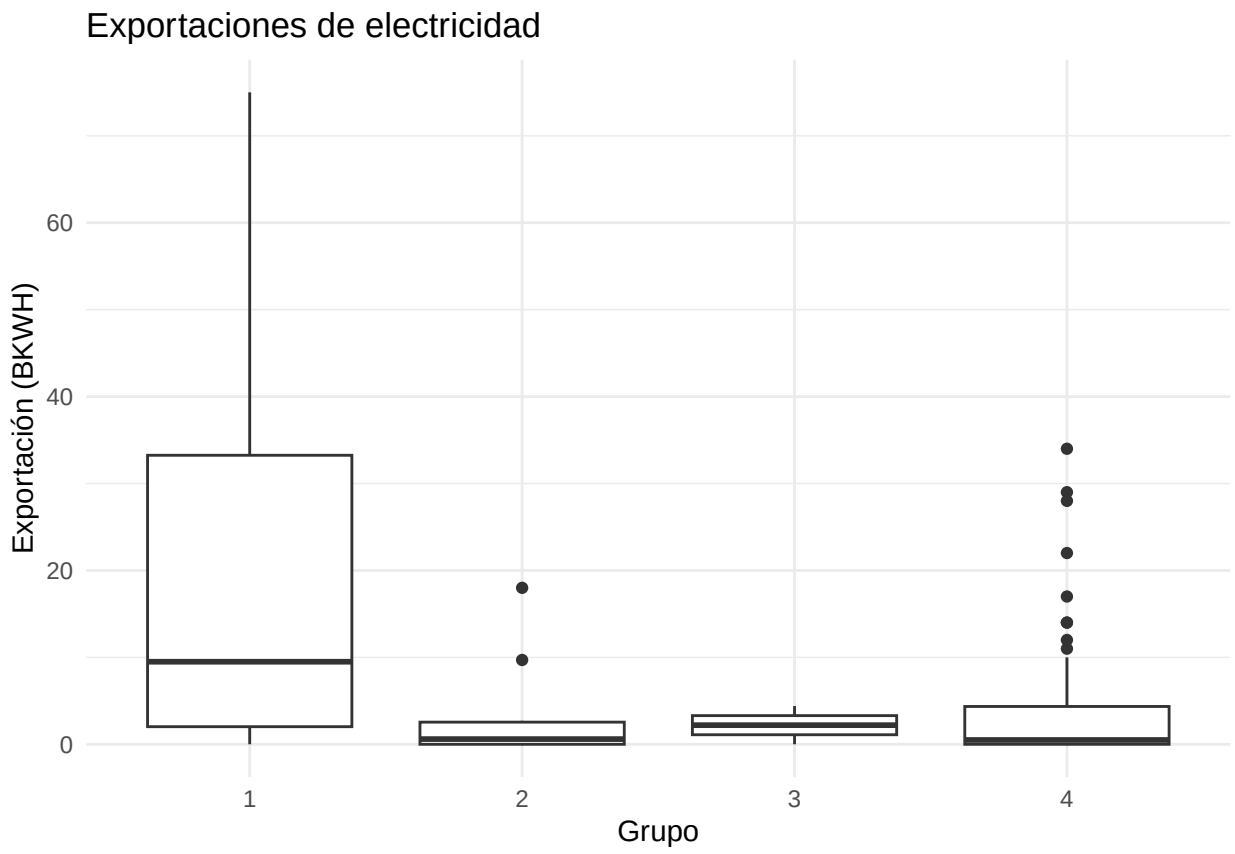
```r
Variables_Escogidas = paises_filtrados[, c("Country" ,"Refined_petroleum_products_consumption_TBPD", "R
Variables_Escogidas$Grupo <- clusters # agregar al set de datos una columna con el grupo

# boxplot de Refined_petroleum_products_consumption_TBPD
ggplot(Variables_Escogidas, aes(x = factor(Grupo), y = Refined_petroleum_products_consumption_TBPD)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Consumo de derivados de petróleo por grupo",
    x = "Grupo",
    y = "Consumo (TBPD)" )
```



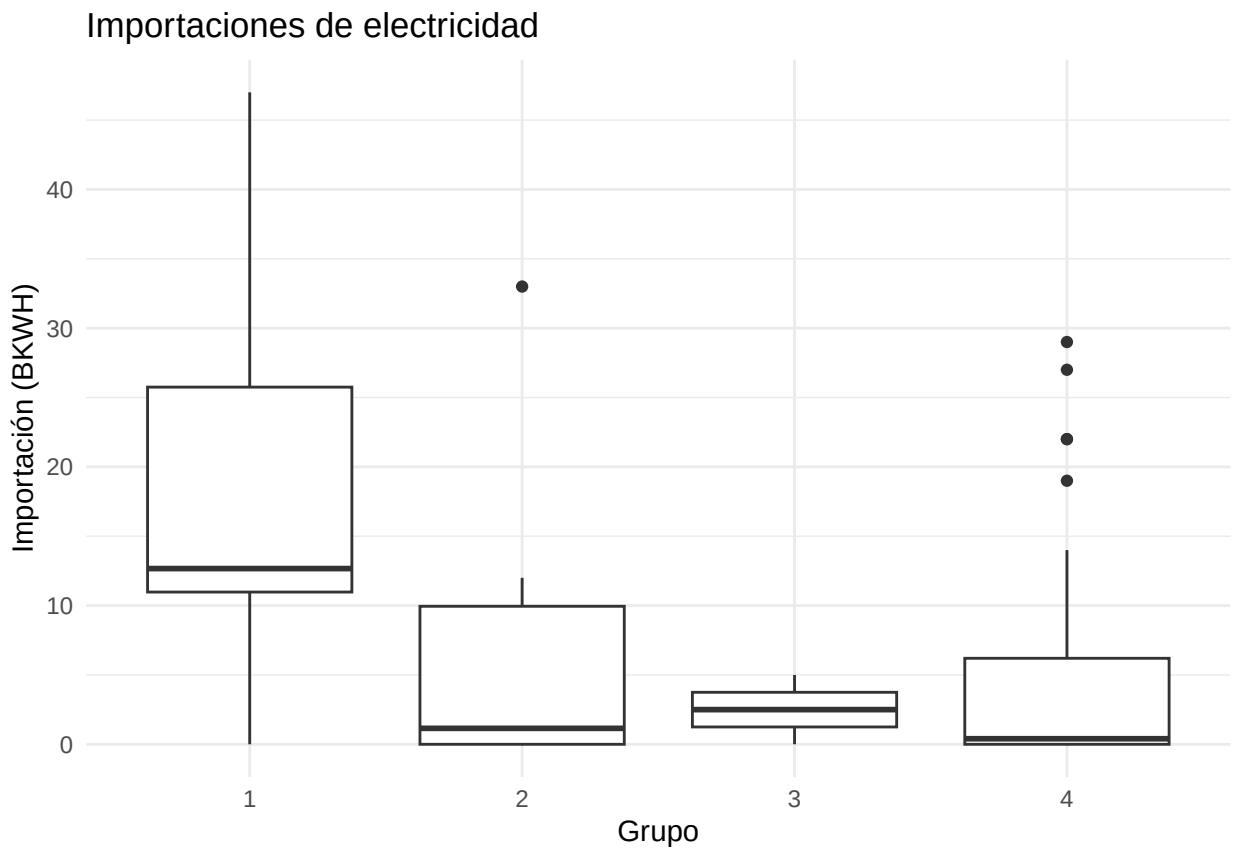

```r
# boxplot de Refined_petroleum_products_production_TBPD
ggplot(Variables_Escogidas, aes(x = factor(Grupo), y = Refined_petroleum_products_production_TBPD)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Producción de derivados de petróleo por grupo",
    x = "Grupo",
    y = "Producción (TBPD)" )
```



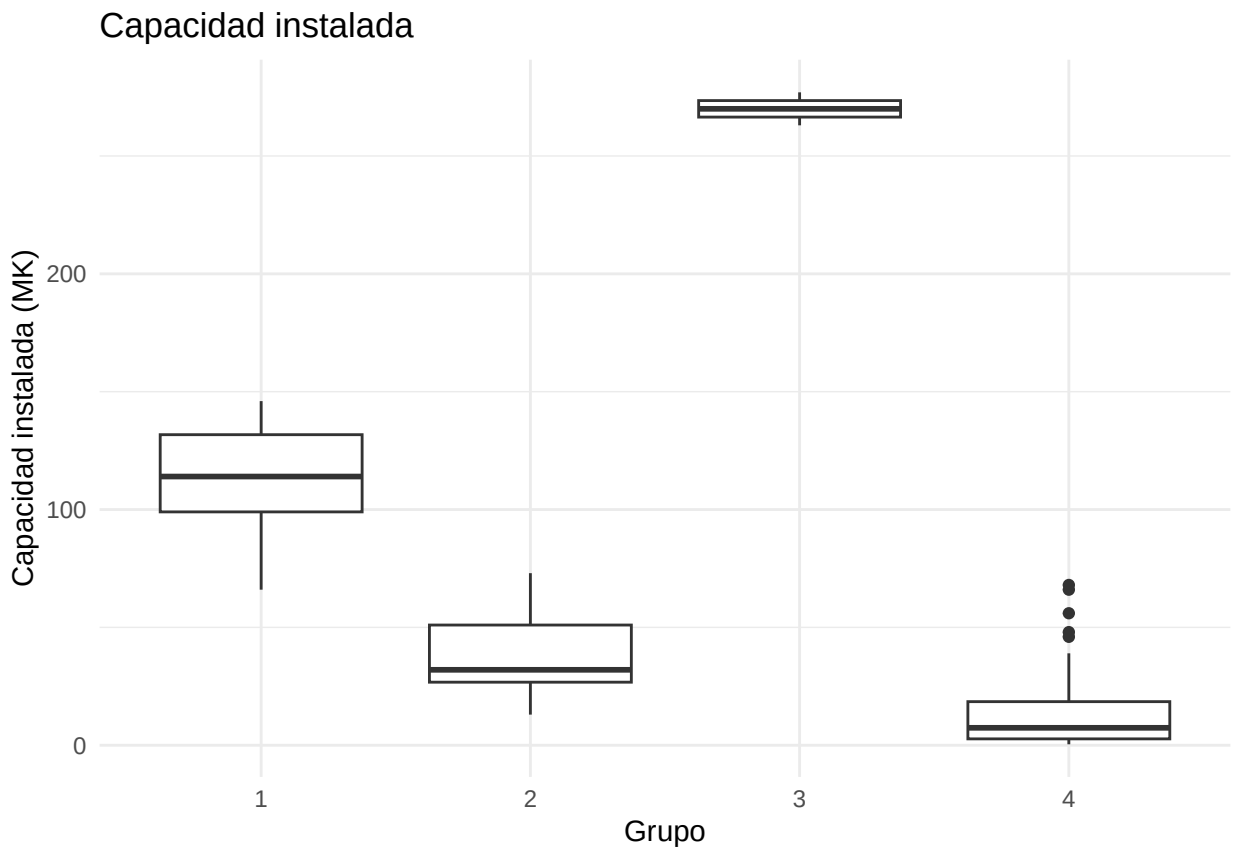

```r
# boxplot de Electricity_exports_BKWH
ggplot(Variables_Escogidas, aes(x = factor(Grupo), y = Electricity_exports_BKWH)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Exportaciones de electricidad",
    x = "Grupo",
    y = "Exportación (BKWH)")
```




```
# boxplot de Electricity_imports_BKWH
ggplot(Variables_Escogidas, aes(x = factor(Grupo), y = Electricity_imports_BKWH)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Importaciones de electricidad",
    x = "Grupo",
    y = "Importación (BKWH)")
```




```
# boxplot de Electricity_installed_capacity_MK
ggplot(Variables_Escogidas, aes(x = factor(Grupo), y = Electricity_installed_capacity_MK)) +
  geom_boxplot() +
  theme_minimal() +
  labs(
    title = "Capacidad instalada",
    x = "Grupo",
    y = "Capacidad instalada (MK)")
```



Referencias

1. Introduction to R markdown
2. MarkDown Guide - Basic Syntax
3. Data Visualization CheatSheet
4. Repositorio del Taller en Github
5. Pagina web del taller en Gihub Pages