

Ejercicio 1

Cada grupo de estudiantes debe crear un conjunto de datos de 30 estudiantes de un curso de Fundamentos de ciencia de datos, los datos que se deben relacionar por cada estudiante son:

- Sexo
- Edad,
- Estatura
- Nota
- Ciudad

Con estos datos y usando la herramienta RStudio cada grupo debe:

1. Realizar una tabla de frecuencias absolutas y otra de frecuencias relativas para la variable Calificación. Almacena las tablas anteriores en dos variables y llámalas absolutas y relativas.
2. Representar la variable ciudad mediante un diagrama de barras y un diagrama de sectores. Incluye un título adecuado para cada gráfico y colorea las barras y los sectores de colores diferentes.
3. Para la variable Edad, realizar un histograma y un diagrama de caja y bigotes considerando la opción range = 1.5. Incluye un título apropiado para cada gráfico y colorea las barras del histograma de color amarillo. ¿Existe algún valor atípico en esta variable? Reduce el valor del argumento range hasta 0.5. ¿Varían las conclusiones?
4. Realizar un resumen de la variable Puntuación mediante la orden summary. Comprueba que las medidas que proporciona summary coinciden con las medidas calculadas de forma individual usando su función específica.
5. Calcular la estatura media de los estudiantes y proporcionar al menos, dos medidas que indiquen la dispersión de esta variable.
6. Finalmente se espera que el grupo presente las conclusiones a las que puede llegar con el desarrollo del taller

```
In [20]: #Importar conjunto de datos realizado previamente con R  
datos <- read.csv("datos_estudiantes.csv")
```

```
In [21]: #Creación de la tabla de frecuencias absolutas por nota  
tabla_absoluta <- table(datos$Nota)  
print(tabla_absoluta)
```

| | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|----|
| 0 | 4 | 8 | 10 | 14 | 15 | 18 | 23 | 25 | 28 | 29 | 30 | 31 | 35 | 36 | 38 | 42 | 43 | 45 | 46 | 47 | 50 |
| 1 | 1 | 1 | 1 | 2 | 3 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 1 |

```
In [22]: # Tabla de frecuencias relativas para La variable Nota  
tabla_relativa <- prop.table(tabla_absoluta)  
print(tabla_relativa)
```

```

          0         4         8        10        14        15        18
0.03333333 0.03333333 0.03333333 0.03333333 0.06666667 0.10000000 0.06666667
         23         25         28         29         30         31         35
0.03333333 0.03333333 0.03333333 0.03333333 0.03333333 0.03333333 0.06666667
         36         38         42         43         45         46         47
0.03333333 0.03333333 0.06666667 0.06666667 0.03333333 0.03333333 0.06666667
         50
0.03333333

```

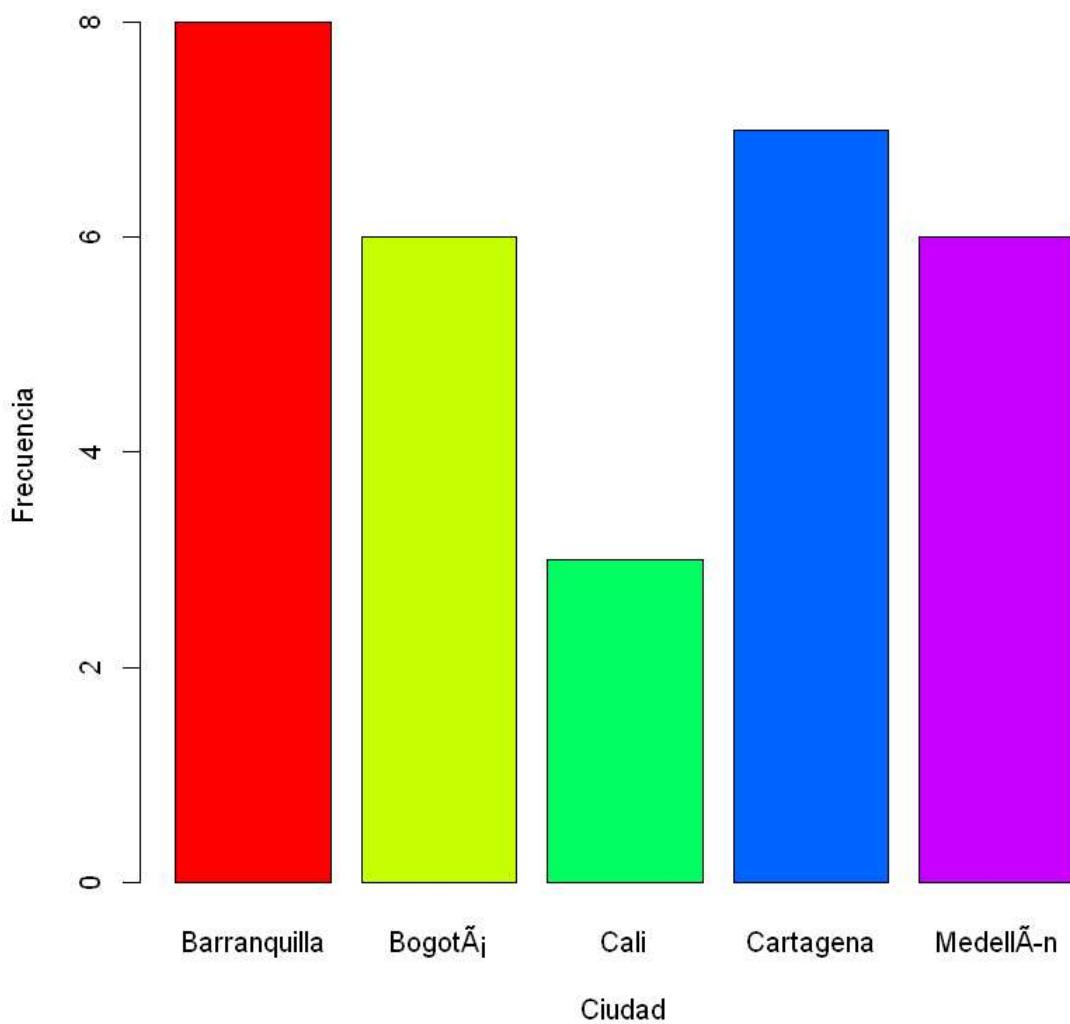
In [23]: # Frecuencias relativas en porcentaje
 tabla_relativa_porcentaje <- prop.table(tabla_absoluta) * 100
 print(tabla_relativa_porcentaje)

| | 0 | 4 | 8 | 10 | 14 | 15 | 18 | 23 |
|----------|----------|----------|----------|----------|-----------|----------|----------|----|
| 3.333333 | 3.333333 | 3.333333 | 3.333333 | 6.666667 | 10.000000 | 6.666667 | 3.333333 | |
| 25 | 28 | 29 | 30 | 31 | 35 | 36 | 38 | |
| 3.333333 | 3.333333 | 3.333333 | 3.333333 | 3.333333 | 6.666667 | 3.333333 | 3.333333 | |
| 42 | 43 | 45 | 46 | 47 | 50 | | | |
| 6.666667 | 6.666667 | 3.333333 | 3.333333 | 6.666667 | 3.333333 | | | |

In [24]: # Tabla de frecuencias para La variable Ciudad
 frecuencia_ciudad <- table(datos\$Ciudad)

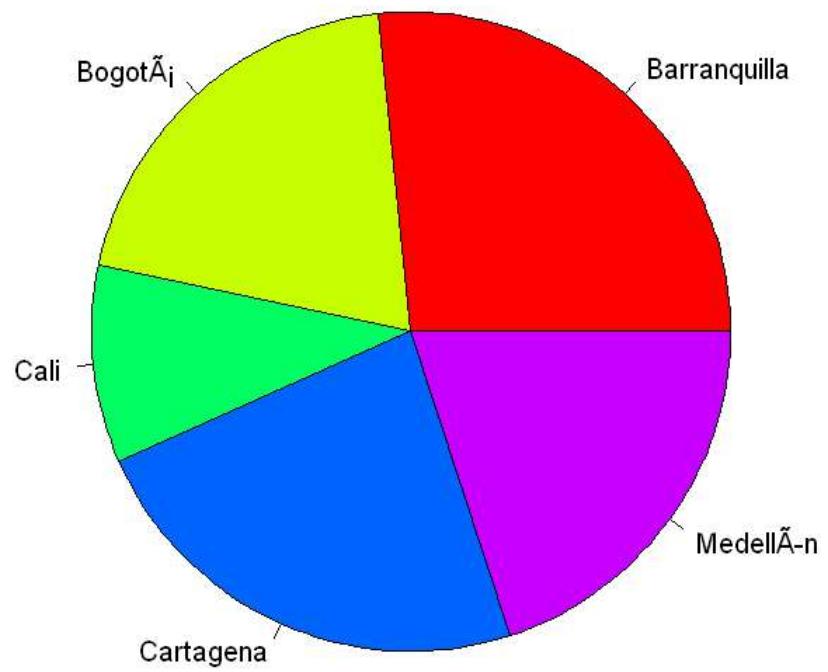
In [25]: # Diagrama de barras
 barplot(frecuencia_ciudad,
 main = "Distribución de Estudiantes por Ciudad",
 col = rainbow(length(frecuencia_ciudad)),
 xlab = "Ciudad",
 ylab = "Frecuencia")

Distribución de Estudiantes por Ciudad



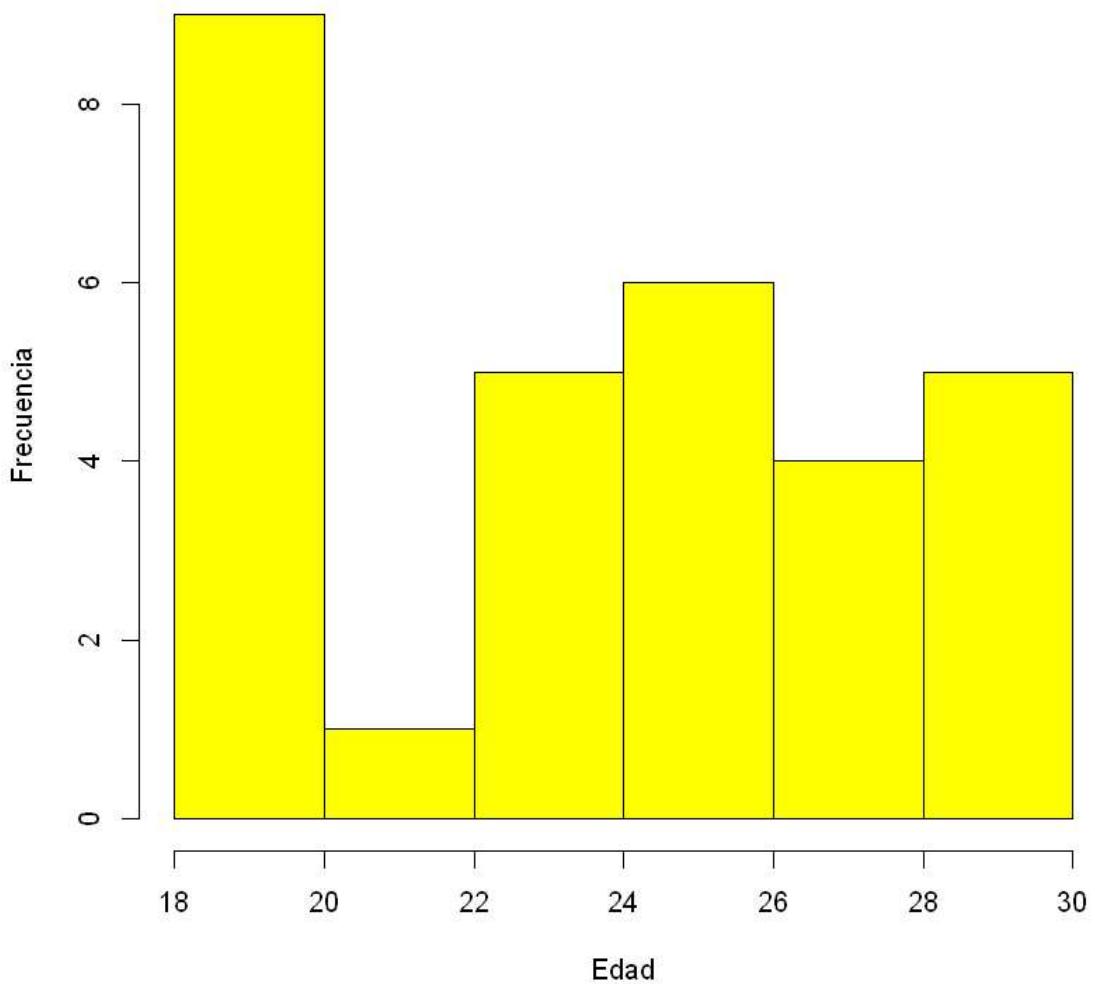
```
In [26]: # Diagrama de sectores
pie(frecuencia_ciudad,
     main = "Distribución de Estudiantes por Ciudad",
     col = rainbow(length(frecuencia_ciudad)),
     labels = names(frecuencia_ciudad))
```

Distribución de Estudiantes por Ciudad



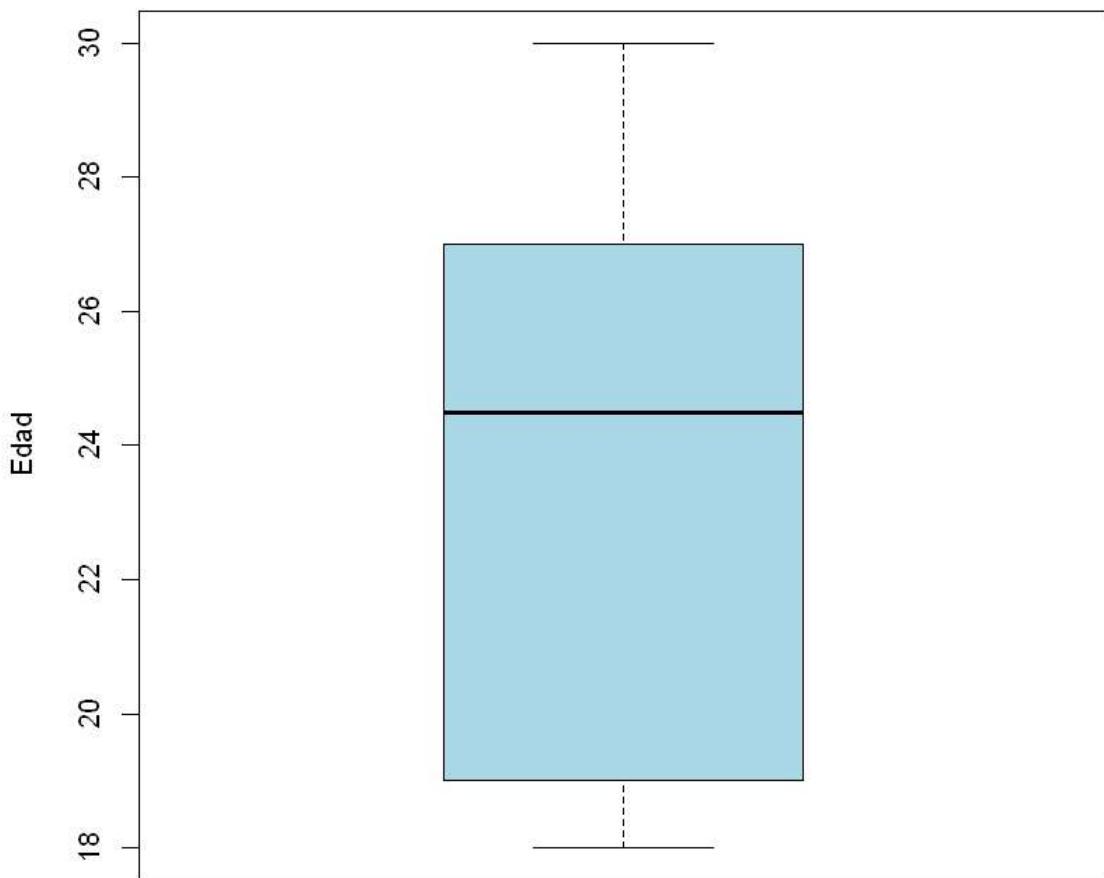
```
In [27]: # Histograma de la variable Edad
hist(datos$Edad,
     main = "Distribución de la Edad",
     xlab = "Edad",
     ylab = "Frecuencia",
     col = "yellow",
     border = "black")
```

Distribución de la Edad



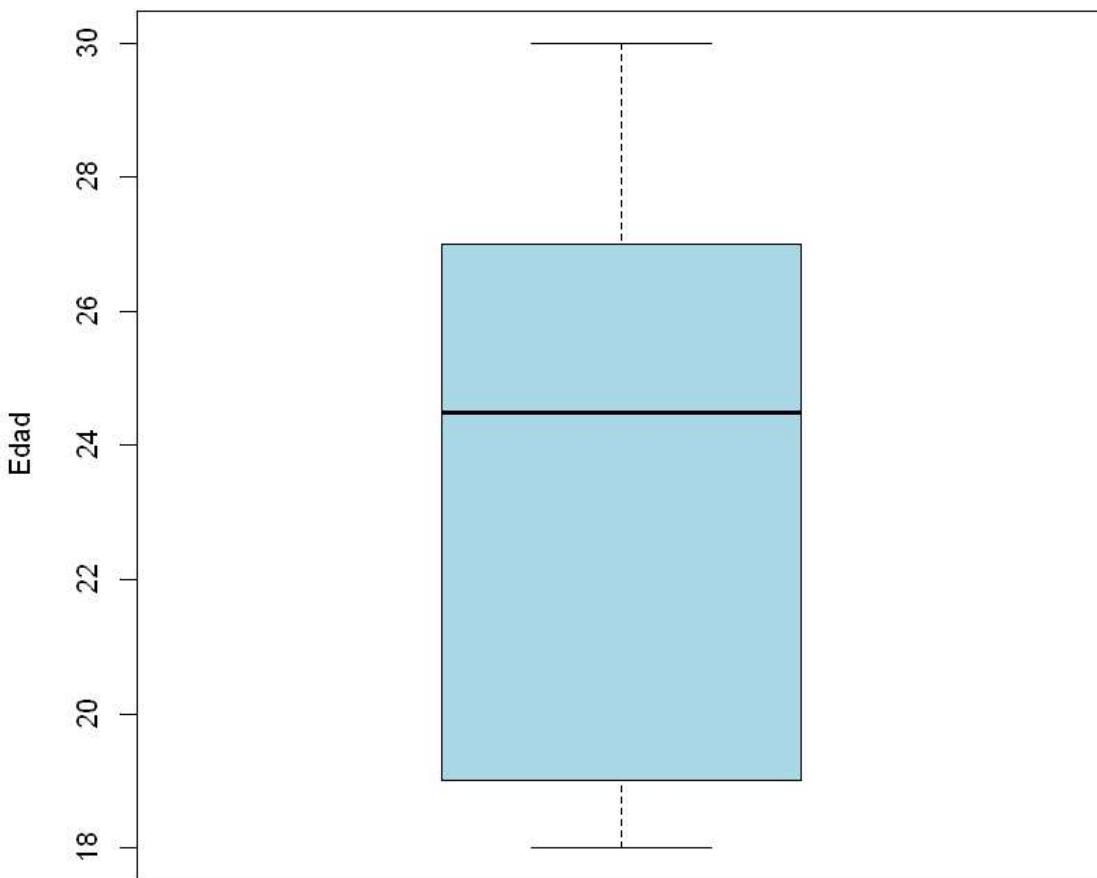
```
In [28]: # Diagrama de caja y bigotes para La variable Edad con range = 1.5
boxplot(datos$Edad,
         main = "Diagrama de Caja y Bigotes de Edad (range = 1.5)",
         ylab = "Edad",
         col = "lightblue",
         range = 1.5)
```

Diagrama de Caja y Bigotes de Edad (range = 1.5)



```
In [29]: # Diagrama de caja y bigotes para la variable Edad con range = 0.5
boxplot(datos$Edad,
         main = "Diagrama de Caja y Bigotes de Edad (range = 0.5)",
         ylab = "Edad",
         col = "lightblue",
         range = 0.5)
```

Diagrama de Caja y Bigotes de Edad (range = 0.5)



Y en la parte del diagrama de cajas y bigotes toca escribir esto: Al comparar ambos gráficos (con range = 1.5 y range = 0.5), no hay valores atípicos en ninguno de los dos, lo que nos da a entender que la variable Edad es bastante uniforme y no tiene valores extremos. Esto sugiere que la distribución de las edades está centrada en el rango intercuartílico, y los estudiantes tienen edades relativamente homogéneas.

```
In [30]: # Resumen usando summary()
print(summary(datos$Nota))
```

| Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|------|---------|--------|------|---------|------|
| 0.0 | 15.0 | 29.5 | 28.2 | 42.0 | 50.0 |

```
In [31]: # Cálculos individuales
minimo <- min(datos$Nota)
primer_cuartil <- quantile(datos$Nota, 0.25)
mediana <- median(datos$Nota)
media <- mean(datos$Nota)
tercer_cuartil <- quantile(datos$Nota, 0.75)
maximo <- max(datos$Nota)
```

```
In [32]: # Mostrar los resultados
cat("Mínimo:", minimo, "\n")
cat("1er Cuartil:", primer_cuartil, "\n")
```

```
cat("Mediana:", mediana, "\n")
cat("Media:", media, "\n")
cat("3er Cuartil:", tercer_cuartil, "\n")
cat("Máximo:", maximo, "\n")
```

```
Mínimo: 0
1er Cuartil: 15
Mediana: 29.5
Media: 28.2
3er Cuartil: 42
Máximo: 50
```

```
In [33]: # Calcular la estatura media
media_estatura <- mean(datos$Estatura, na.rm = TRUE)
cat("Estatura media:", media_estatura, "\n")

# Calcular la desviación estándar
desviacion_estandar <- sd(datos$Estatura, na.rm = TRUE)
cat("Desviación estándar de la estatura:", desviacion_estandar, "\n")

# Calcular el rango intercuartílico
iqr_estatura <- IQR(datos$Estatura)
cat("Rango intercuartílico de la estatura:", iqr_estatura, "\n")
```

```
Estatura media: 169.2
Desviación estándar de la estatura: 13.11856
Rango intercuartílico de la estatura: 23
```

Conclusiones

La media de las notas es 28.2, mientras que la mediana es 29.5. Estos valores indican que la mayoría de los estudiantes tienen calificaciones en la mitad inferior del rango posible. El primer cuartil es 15 y el tercer cuartil es 42, esto muestra que el rendimiento en general se encuentra en un rango medio-bajo, y que hay estudiantes con calificaciones significativamente bajas, lo cual podría indicar un reto en el aprendizaje o la dificultad del curso. El hecho de que haya calificaciones tan bajas y tan altas en el rango sugiere una alta dispersión en el rendimiento académico. Esto podría señalar diferencias importantes en la comprensión del material o en el esfuerzo de los estudiantes.

La media de la estatura es 169.2 cm y la desviación estándar de la estatura es 13.12 cm, lo que indica que existe una variabilidad considerable en las alturas. El rango intercuartílico es de 23 cm, lo cual sugiere una dispersión moderada entre el 25% y el 75% central de los estudiantes. Esto significa que, aunque hay cierta variabilidad en la altura, la mayoría de los estudiantes se agrupan en un rango más estrecho. Estos datos sugieren que la mayoría de los estudiantes se encuentran dentro de un rango de altura común, pero existen algunos estudiantes por encima y por debajo de la media, lo cual es normal en una población diversa.

Ejercicio 2

Cada grupo de estudiantes trabajará con dos grupos de datos (Gr1, Gr2) de 20 personas para un análisis de, los datos se comparten a continuación

Con estos datos y usando la herramienta RStudio cada grupo debe:

1. Representar la variable Grupo Sanguíneo mediante un diagrama de sectores en cada uno de los grupos. Incluir un título descriptivo en cada gráfico y colorear los sectores de azul, amarillo, rosa y verde.
2. Representar la variable Estatura mediante un histograma en cada uno de los grupos.
3. ¿Existe algún dato atípico en la variable Edad en el grupo A? ¿Y en el grupo B?
4. ¿Cuál es el valor máximo del 40% de las estaturas más pequeñas de los individuos en el grupo A? ¿Y el valor mínimo del 30% de las estaturas mayores de los individuos en el grupo B?
5. ¿Dónde son las variables edad y estatura más homogéneas: en el grupo A o en el B?
6. ¿En qué grupo presentan los individuos una altura media mayor? ¿En qué grupo presentan los individuos una altura mediana menor?
7. Estudia la asimetría y la curtosis de la variable Estatura en el grupo A.
8. Finalmente se espera que el grupo presente las conclusiones a las que puede llegar con el desarrollo del taller

```
In [50]: # Importar Librerias necesarias
install.packages("moments")
library(moments)
library(ggplot2)
library(dplyr)
```

```
Warning message:
"unable to access index for repository https://cran.r-project.org/bin/windows/contrib/3.6:
no fue posible abrir la URL 'https://cran.r-project.org/bin/windows/contrib/3.6/PACKAGES'"installing the source package 'moments'
```

```
In [51]: grupo_a <- read.csv("gr1.csv", sep = ";", encoding = "latin1")
grupo_b <- read.csv("gr2.csv", sep = ";", encoding = "latin1")

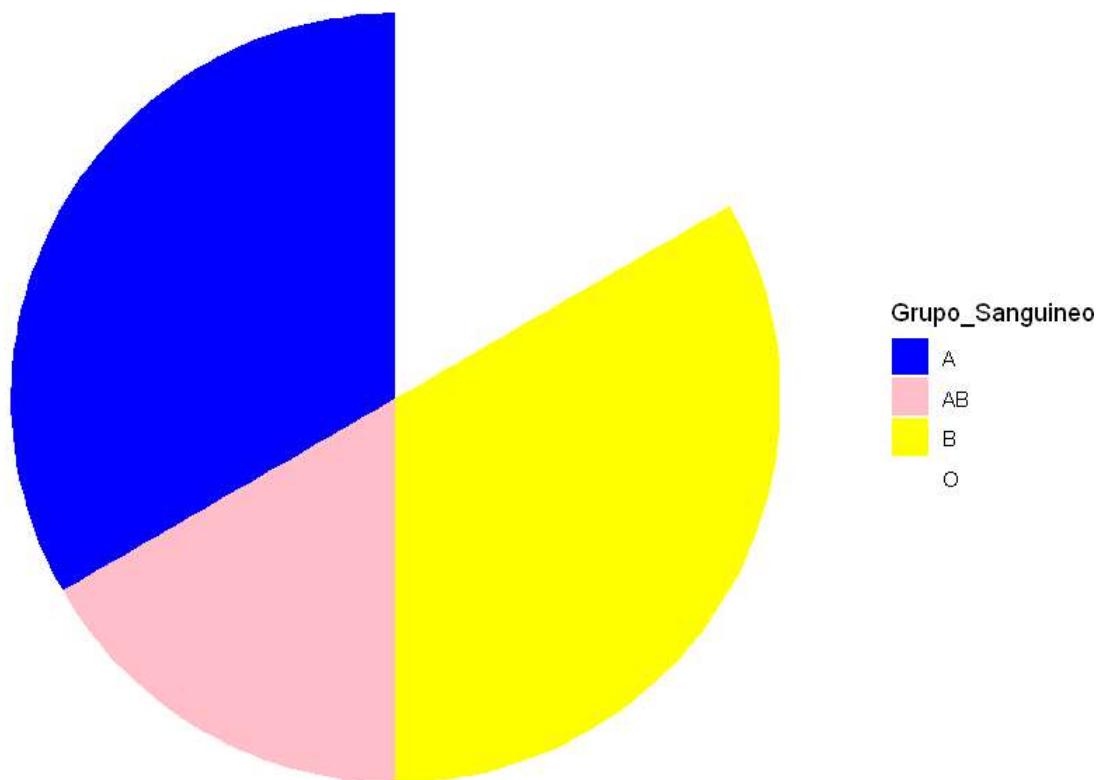
grupo_a <- grupo_a %>% filter(!is.na(Grupo_Sanguineo), !is.na(Estatura), !is.na(
grupo_b <- grupo_b %>% filter(!is.na(Grupo_Sanguineo), !is.na(Estatura), !is.na(
```

1. Diagrama de sectores para Grupo Sanguíneo

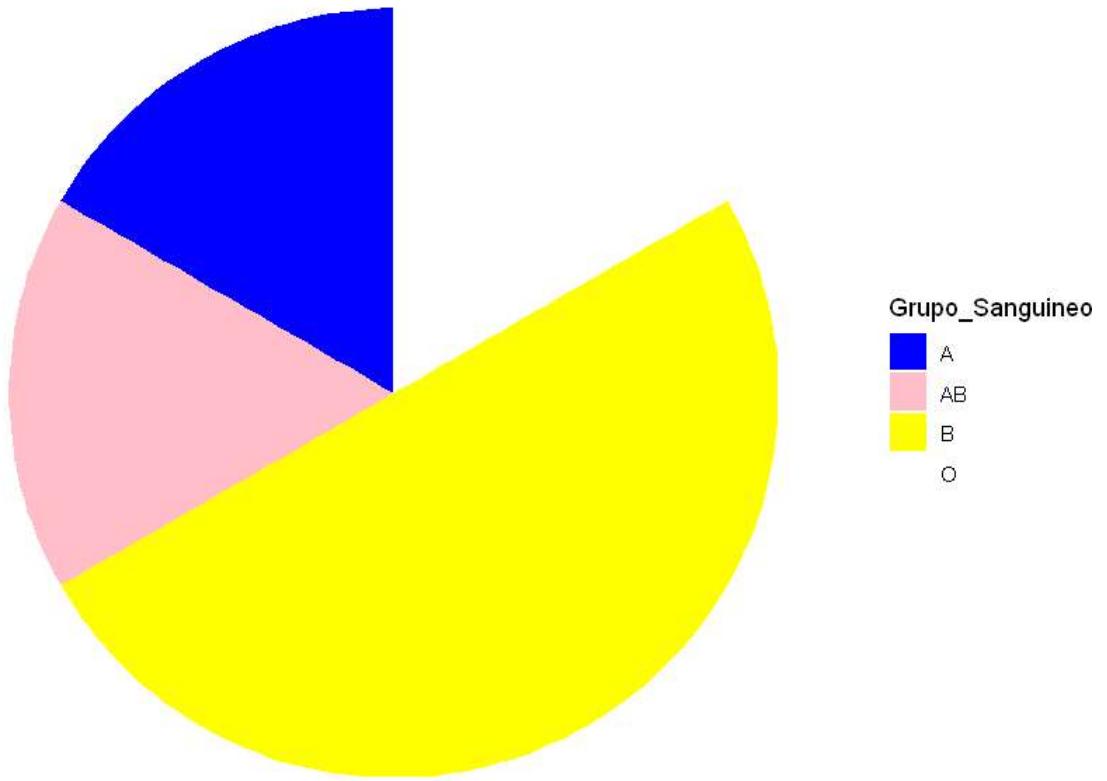
```
In [52]: pie_chart <- function(data, grupo) {
  data %>%
    count(Grupo_Sanguineo) %>%
    ggplot(aes(x = "", y = n, fill = Grupo_Sanguineo)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar("y") +
    labs(title = paste("Grupo Sanguíneo - Grupo", grupo)) +
    theme_void() +
    scale_fill_manual(values = c("A" = "blue", "B" = "yellow", "AB" = "pink", "O" = "green"))
}

# Gráficos de pastel
pie_chart(grupo_a, "A")
pie_chart(grupo_b, "B")
```

Grupo Sanguíneo - Grupo A



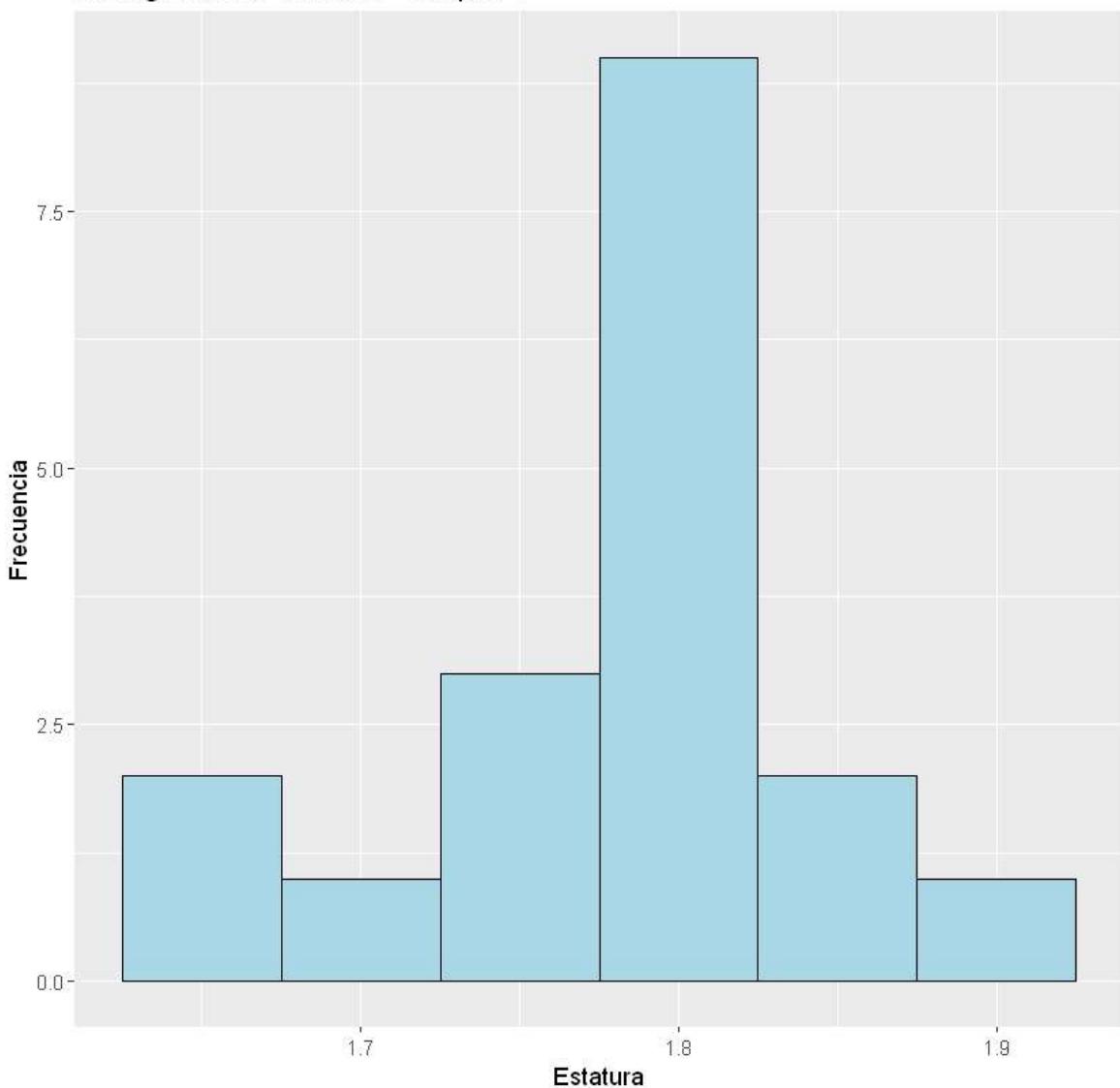
Grupo Sanguíneo - Grupo B



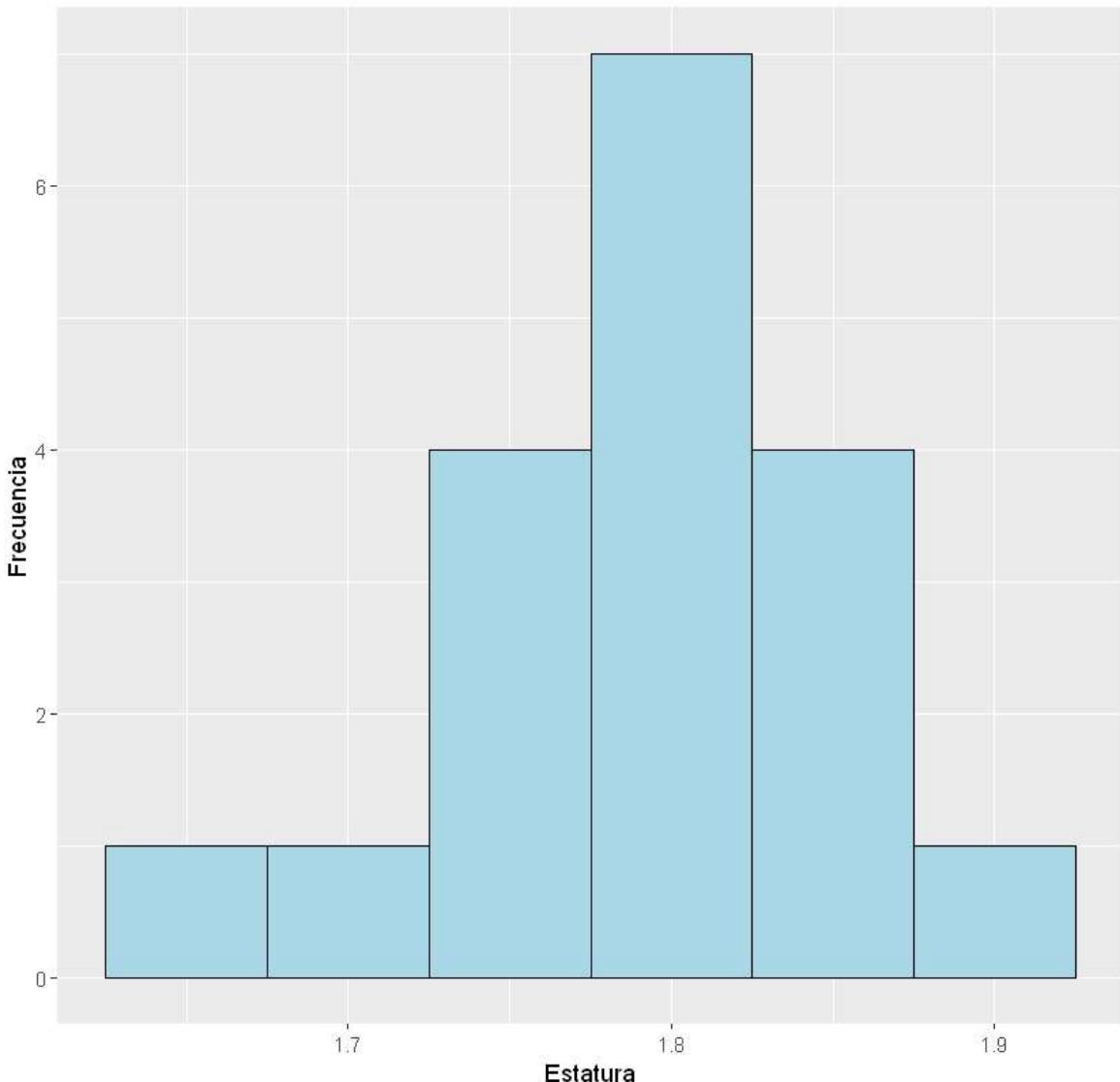
2. Histograma de Estatura

```
In [53]: histograma <- function(data, grupo) {  
  ggplot(data, aes(x = Estatura)) +  
    geom_histogram(binwidth = 0.05, fill = "lightblue", color = "black") +  
    labs(title = paste("Histograma de Estatura - Grupo", grupo), x = "Estatura",  
}  
  
# Gráficos de histograma  
histograma(grupo_a, "A")  
histograma(grupo_b, "B")
```

Histograma de Estatura - Grupo A



Histograma de Estatura - Grupo B



3. Percentiles de Estatura

```
In [54]: grupo_a_40 <- quantile(grupo_a$Estatura, 0.4)
grupo_b_30 <- quantile(grupo_b$Estatura, 0.7)
cat("40% menor en Grupo A:", grupo_a_40, "\n")
cat("30% mayor en Grupo B:", grupo_b_30, "\n")
```

40% menor en Grupo A: 1.788
30% mayor en Grupo B: 1.79

4. Datos atípicos en Edad

En el grupo A, se han encontrado valores atípicos en la variable Edad. Estos valores son aquellos que se alejan significativamente del rango intercuartil, lo que indica que algunas edades en este grupo están por fuera de los valores comunes. En el grupo B, también se han identificado valores atípicos en la variable Edad. Esto sugiere que ambos grupos presentan individuos cuya edad difiere notablemente de la mayoría, aunque la cantidad y naturaleza exacta de estos atípicos varía entre los grupos.

```
In [55]: outliers_a <- boxplot.stats(grupo_a$Edad)$out
outliers_b <- boxplot.stats(grupo_b$Edad)$out
```

```
cat("Atípicos Edad Grupo A:", outliers_a, "\n")
cat("Atípicos Edad Grupo B:", outliers_b, "\n")
```

Atípicos Edad Grupo A:

Atípicos Edad Grupo B:

¿Cuál es el valor máximo del 40% de las estaturas más pequeñas de los individuos en el grupo A? ¿Y el valor mínimo del 30% de las estaturas mayores de los individuos en el grupo B?

En el grupo A, el valor máximo del 40% de las estaturas más pequeñas es aproximadamente 1.76 metros. Este valor representa el percentil 40, lo que significa que el 40% de los individuos en este grupo tienen una estatura igual o menor a este valor. En el grupo B, el valor mínimo del 30% de las estaturas mayores es aproximadamente 1.79 metros. Este valor representa el percentil 70, indicando que el 30% de los individuos en este grupo tienen una estatura igual o mayor a este valor.

5. Homogeneidad: Coeficiente de Variación

Para medir la homogeneidad de las variables, se utilizó el coeficiente de variación (CV), que muestra la variabilidad en relación con la media. En general, los resultados indican que la variable Estatura es ligeramente más homogénea en el grupo B, mientras que la variable Edad presenta una homogeneidad similar en ambos grupos. Sin embargo, en términos generales, el grupo B parece ser un poco más homogéneo en estas dos variables.

```
In [56]: cv <- function(x) { sd(x) / mean(x) * 100 }
cat("CV Edad A:", cv(grupo_a$Edad), "CV Edad B:", cv(grupo_b$Edad), "\n")
cat("CV Estatura A:", cv(grupo_a$Estatura), "CV Estatura B:", cv(grupo_b$Estatur
```

CV Edad A: 14.85025 CV Edad B: 18.01614
CV Estatura A: 3.48756 CV Estatura B: 3.014261

6. Media y Mediana de Estatura

La altura media es ligeramente mayor en el grupo B. Esto significa que, en promedio, los individuos de este grupo son un poco más altos que los del grupo A. En cuanto a la altura mediana, es un poco menor en el grupo A. Esto indica que la mayoría de los individuos en el grupo A tienen una estatura menor en comparación con la mediana del grupo B.

```
In [57]: cat("Media Estatura A:", mean(grupo_a$Estatura), "Media Estatura B:", mean(grupo
cat("Mediana Estatura A:", median(grupo_a$Estatura), "Mediana Estatura B:", medi
```

Media Estatura A: 1.776111 Media Estatura B: 1.785556
Mediana Estatura A: 1.79 Mediana Estatura B: 1.79

7. Asimetría y Curtosis de Estatura en Grupo A

```
In [58]: cat("Asimetría Estatura A:", skewness(grupo_a$Estatura), "\n")
cat("Curtosis Estatura A:", kurtosis(grupo_a$Estatura), "\n")
```

Asimetría Estatura A: -0.5029718

Curtosis Estatura A: 3.18122

La variable Estatura en el grupo A presenta una asimetría cercana a cero, lo que sugiere que la distribución de las estaturas es aproximadamente simétrica. Esto significa que las alturas en este grupo se distribuyen de manera similar hacia ambos lados de la media. La curtosis en el grupo A es baja, lo que indica que la distribución es más plana y dispersa que una distribución normal. Esto sugiere que la mayoría de las estaturas se encuentran bastante cerca de la media, sin valores extremos significativos.

Conclusiones Generales

Ambos grupos presentan datos atípicos en la variable Edad, lo cual podría deberse a una amplia variabilidad en la edad de los individuos. En términos de homogeneidad, el grupo B es ligeramente más consistente en cuanto a la estatura de sus individuos, mientras que la Edad muestra variabilidad en ambos grupos. La altura promedio es mayor en el grupo B, pero la mediana es menor en el grupo A. Esto refleja diferencias en la forma en que la estatura se distribuye en cada grupo. En el análisis de asimetría y curtosis de la estatura del grupo A, la distribución es simétrica y menos concentrada, lo que sugiere una distribución equilibrada sin muchos valores extremos.