Analisis Estadistico de Datos

Grupo 05

2025-03-22

Contents

Detalles

Taller 1: Análisis Estadístico de Datos
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

Parte 1

Pregunta 1

$w_1 = 7$

Importe la base de datos a R y determine el número de UE y de variables en la base de datos. ¿Cuántas filas y cuántas columnas tiene la tabla?, ¿Cuántos registros fueron medidos o recolectados a medio día (hora = 12:00)?

```r
library(tidyverse)
library(ggExtra)
library(ggplot2)

datos <- read_csv2("../datasets/ori.csv")
tabla=read.csv2("/Users/aosoriom/Repos/MER_AEDatos/datasets/ori.csv")
MedioDia = tabla[tabla$Hora=="12:00",]
```

```r
summary(tabla)
```

```
##     Cod_Div           Latitud            Longitud           Region
##  Min.   :   99001   Length:4543        Length:4543        Length:4543
##  1st Qu.:50270000   Class :character   Class :character   Class :character
##  Median :81065000   Mode  :character   Mode  :character   Mode  :character
##  Mean   :66126186
##  3rd Qu.:85263000
##  Max.   :99773000
##  Departamento        Municipio            Fecha              Hora
##  Length:4543        Length:4543        Length:4543        Length:4543
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Temperatura        Velocidad_del_Viento Direccion_del_Viento
##  Length:4543        Length:4543          Length:4543
##  Class :character   Class :character     Class :character
##  Mode  :character   Mode  :character     Mode  :character
##
##
##
##    Presion           Punto_de_Rocio     Cobertura_total_nubosa
##  Length:4543        Length:4543        Length:4543
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Precipitacion_mm_h Probabilidad_de_Tormenta   Humedad
##  Length:4543        Min.   : 0.000           Length:4543
##  Class :character   1st Qu.: 0.000           Class :character
##  Mode  :character   Median : 0.000           Mode  :character
##                     Mean   : 1.197
##                     3rd Qu.: 0.000
##                     Max.   :60.000
##   Pronostico
##  Length:4543
##  Class :character
##  Mode  :character
##
##
##
```

UE= Orinoquia=1 Variables=18 Filas:4543 Columnas:18 Registros medio dia: 118

Pregunta 2

$w_2 = 7$

Para los registros recolectados a medio día ¿cuál es el pronóstico climático más frecuente? ¿cuál es el menos frecuente?

```
library(dplyr)

tabla %>%
  count(Pronostico, sort = TRUE)
```

```
##                                Pronostico    n
## 1                    Parcialmente Nublado 1691
## 2                                 Nublado 1318
## 3                          Nublado - Lluvia  470
## 4    Parcialmente Nublado - Lluvia Fuerte  353
## 5                        Nublado - Llovizna  236
## 6          Parcialmente Nublado - Llovizna  213
## 7                                Despejado  156
## 8  Parcialmente Nublado - Tormenta Ligera   48
## 9          Parcialmente Nublado - Tormenta   43
## 10                Nublado - Lluvia Fuerte    9
## 11           Parcialmente Nublado - Lluvia    6
```

El más frecuente es parcialmente Nublado mientras que el menos frecuente es Parcialmente Nublado-Lluvia.

Pregunta 3

$w_3 = 7$

Clasifique las variables 1) Pronóstico, 2) Municipio, 3) Temperatura y 4) Velocidad del viento según su escala y clase, teniendo en cuenta que:

- Pronóstico Pronóstico del clima.
  - Cualitativa
  - Clase: Politómicas.
  - Escala: Ordinal.

- Municipio Municipio de medición
  - Cualitativa
  - Clase: Politómicas.
  - Escala: Nominal

- Temperatura Temperatura medida en grados centígrados.
  - Cuantitativa
  - Clase: Discreta

  - Escala: Intervalo

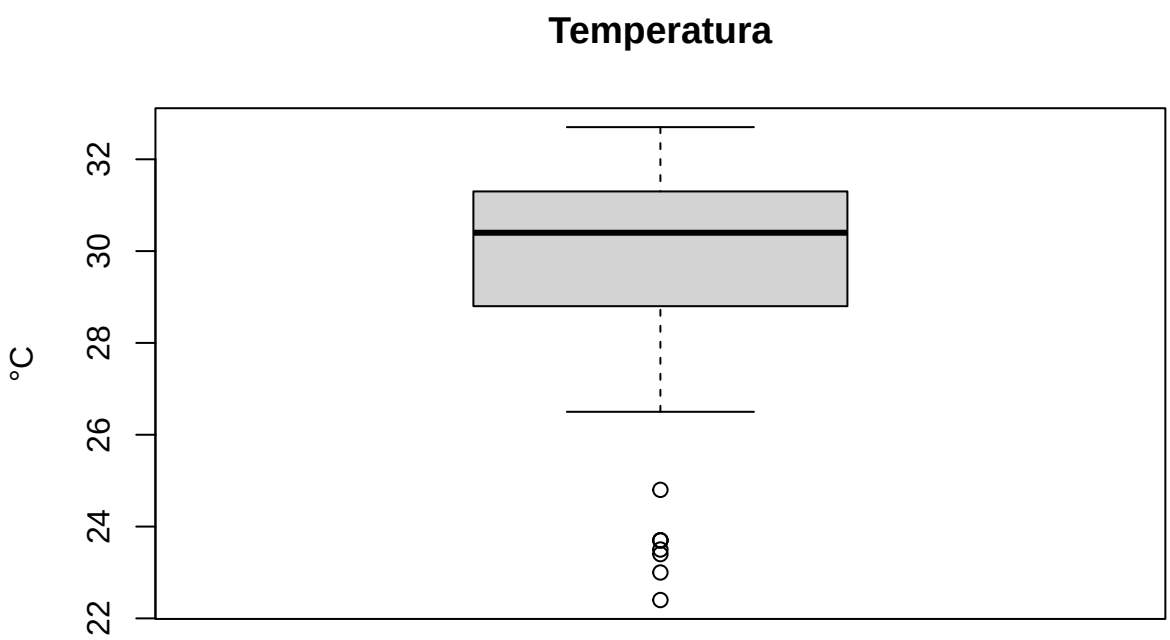
- Velocidad del viento Velocidad del viento medida en mph.
  - Cuantitativa
  - Clase: Continua
  - Escala: Razon

Pregunta 4

$w_4 = 7$

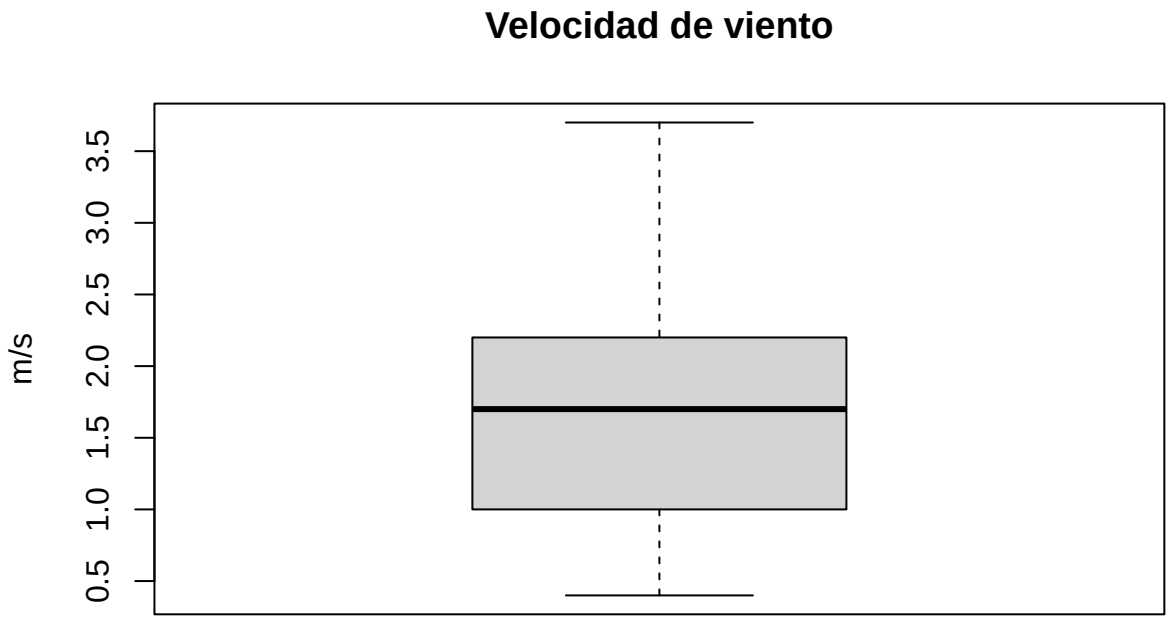
Para los registros recolectados a medio día, describa de manera gráfica la temperatura y la velocidad del viento univariadamente a partir de un diagrama de caja. ¿Qué características observa los datos en términos de centro, localización y dispersión?.

```
MedioDia$Temperatura <- gsub(",", ".", MedioDia$Temperatura)
MedioDia$Temperatura <- as.numeric(MedioDia$Temperatura)
boxplot(MedioDia$Temperatura, horizontal = FALSE, main = 'Temperatura', ylab = '°C') #punto 4
```



Del diagrama `boxplot` para los datos de medio día se observa que la temperatura media es aproximadamente 30°C, y que el histograma de los datos estaría sesgado hacia la izquierda, es decir, los datos están más concentrados en temperaturas entre 30°C y 33°C. En contraposición, las temperaturas por debajo de 30°C tienen más dispersión. Del diagrama se puede concluir que las temperaturas más probables a medio día se encuentran aproximadamente entre 29°C y 31°C, que son los valores que delimitan la caja entre Q1 y Q3 del boxplot. De otra parte, no se observan valores altos de temperatura atípicos, mientras que sí se observan varios valores atípicos por debajo de 27°C aproximadamente.

```r
MedioDia$Velocidad_del_Viento  <- gsub(",", ".", MedioDia$Velocidad_del_Viento )
MedioDia$Velocidad_del_Viento  <- as.numeric(MedioDia$Velocidad_del_Viento )
boxplot(MedioDia$Velocidad_del_Viento ,horizontal = FALSE, main = 'Velocidad de viento', ylab = 'm/s')
```



A continuación se muestra el diagrama boxplot de la velocidad de viento. La primera característica que salta a la vista es que no se encuentran valores atípicos en el conjunto de datos. En segundo lugar, la velocidad media del viento está entre 1.5m/s y 2m/s, visualmente puede indicarse aproximadamente 1.7m/s. Las velocidades de viento más usuales están entre aproximadamente 1m/s y 2.2m/s, y ligeramente con mayor concentración entre la media 1.7m/s y 2.2m/s.

Pregunta 5

$w_5 = 7$

Para los registros recolectados a medio día, describa de manera numérica mediante la función `summary()` la temperatura y la velocidad del viento univariadamente. ¿Concuerda su descripción con los gráficos de boxplot previamente presentados?.

```r
summary(MedioDia$Temperatura) # Punto 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   22.40   28.82   30.40   29.75   31.30   32.70
```

Estos datos concuerdan con lo analizado a partir del boxplot. Sin embargo, en el boxplot se hace una aproximación de los valores de la temperatura, mientras que aquí la tabla presenta valores exactos. De esta

manera, aunque el boxplot da más información sobre los datos, la función summary puede complementar dando la cifra exacta. Se confirma la temperatura media a medio día observada en el boxplot. También se confirma el valor máximo de temperatura, de casi 33°C. En cuanto al valor mínimo, los 22.4°C coinciden con lo que se presenta en el boxplot, sin embargo, observando el boxplot podemos identificar que ese mínimo es un outlier.

```
summary(MedioDia$Velocidad_del_Viento) # Punto 5
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.400   1.000   1.700   1.719   2.175   3.700
```

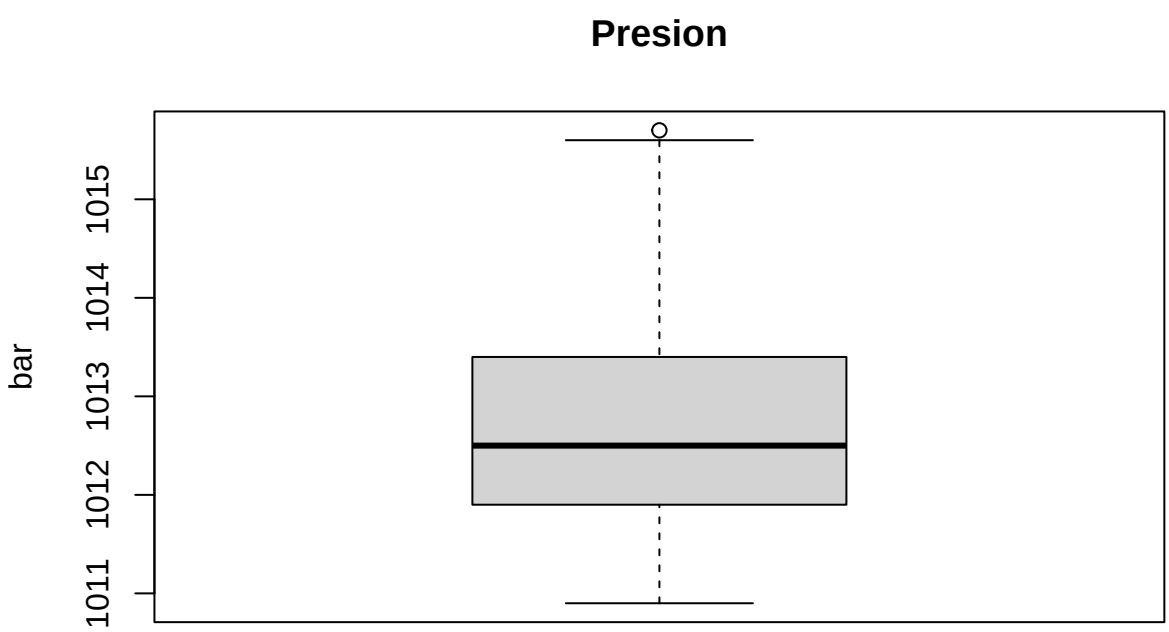
Así como para la temperatura, los datos aquí presentados son consistentes con lo analizado a partir del boxplot. Se observa que el mínimo y máximo coincide con las líneas del boxplot (o "bigotes") que delimitan el valor máximo y mínimo no atípicos. La media de la velocidad de viento se sitúa cercana a 1.7m/s, como se observó en el boxplot, y así mismo, los valores del primer y tercer cuartil corresponden con los identificados en el Punto 4. Se resalta, no obstante, que los valores identificados en el punto 4 no eran exactos, mientras que en este punto s{i se observa el valor exacto de los datos.

Pregunta 6

$w_6 = 10$

Para los registros recolectados a medio día, determine si existen o no observaciones atípicas univariadas para la variables Presión y Punto de Rocio usando el diagrama de caja de Tukey. ¿cuál variable presenta mayor número de atipicidades?

```
MedioDia$Presion  <- gsub(",", ".", MedioDia$Presion )
MedioDia$Presion  <- as.numeric(MedioDia$Presion )
boxplot(MedioDia$Presion, horizontal=FALSE, main='Presion', ylab = 'bar')
```

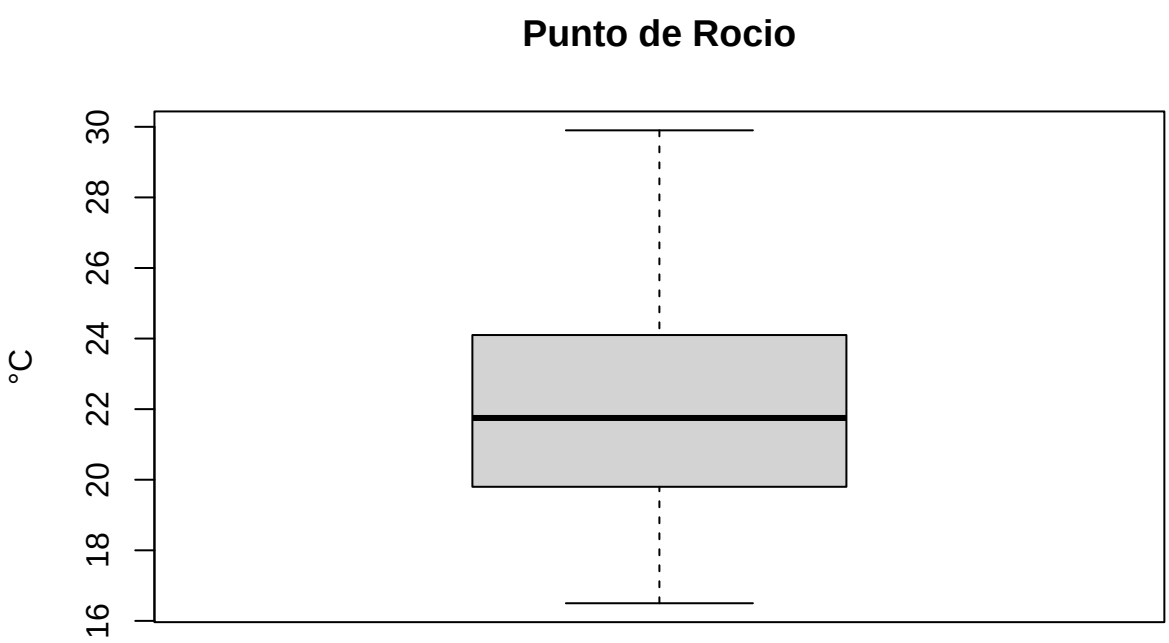


Outlier (círculo) Hay un valor atípico por encima de los 1015 mmHg, lo que indica que hubo un registro de presión inusualmente alto comparado con la tendencia de los datos.

Rango intercuartílico (caja gris) La caja representa el rango intercuartílico (Q1 a Q3), es decir, donde se encuentra el 50% central de los datos. Aproximadamente va de 1012 mmHg a 1013.5 mmHg.

Bigotes El bigote inferior llega cerca de 1011 mmHg.
El bigote superior llega hasta un poco más de 1014 mmHg.

```
MedioDia$Punto_de_Rocio  <- gsub(",", ".", MedioDia$Punto_de_Rocio )
MedioDia$Punto_de_Rocio  <- as.numeric(MedioDia$Punto_de_Rocio )
boxplot(MedioDia$Punto_de_Rocio, horizontal=FALSE, main='Punto de Rocio', ylab = '°C')
```



Mediana La línea negra dentro de la caja indica que la mediana está aproximadamente en 21-22°C.

Rango intercuartílico

- Primer cuartil (Q1) alrededor de 19°C
- Tercer cuartil (Q3) cerca de 24-25°C
- Esto indica que el 50% central de los datos está entre 19°C y 25°C

Distribución La caja está ligeramente desplazada hacia la parte baja, mostrando una posible leve concentración de datos en valores bajos.

Pregunta 7

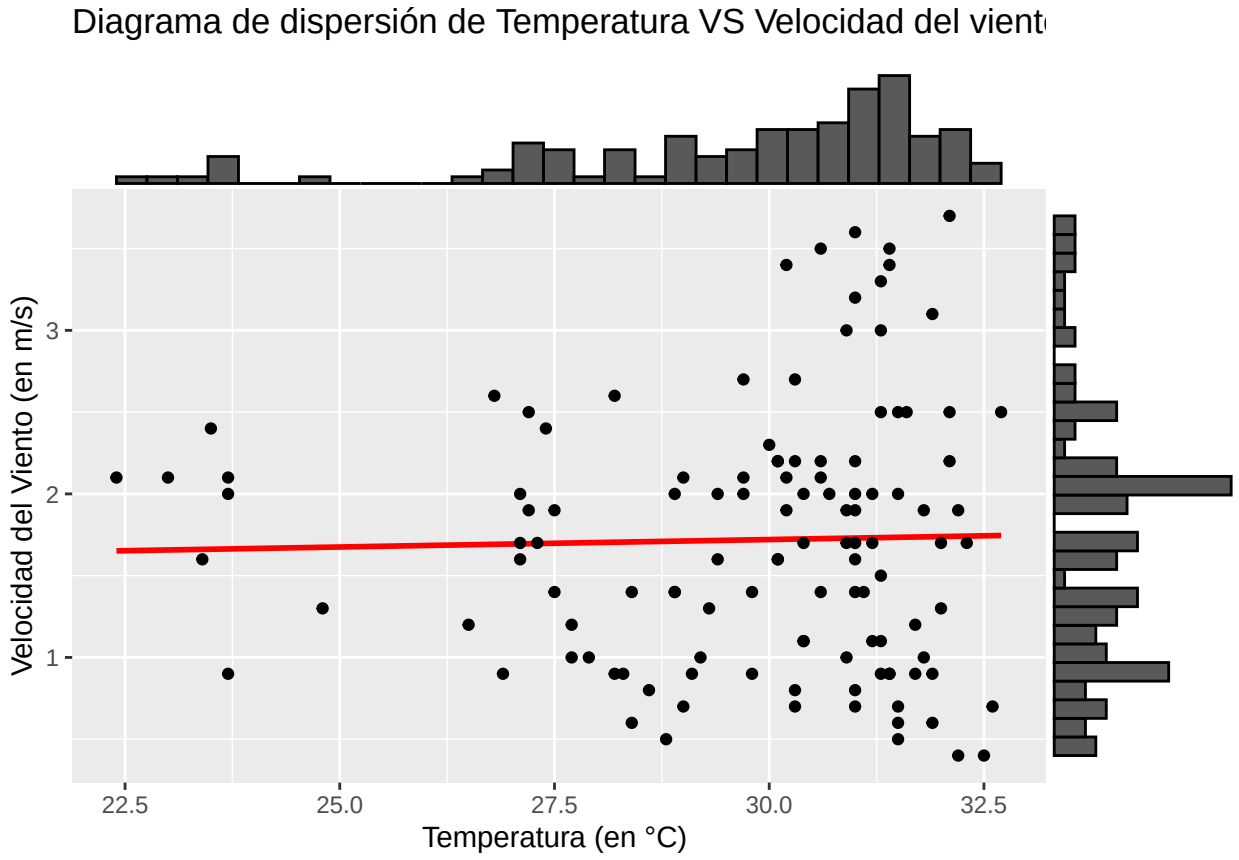
$w_7 = 10$

Para los registros recolectados a medio día, describa de manera gráfica la asociación entre la temperatura y velocidad del viento a partir de un diagrama de dispersión entre las variables. Añada el histograma marginal de cada variable. ¿Parecen correlacionarse las dos variables?

```
g_base = ggplot(MedioDia, aes(x = Temperatura,
                                  y = Velocidad_del_Viento)) +
```

```
  labs(title="Diagrama de dispersión de Temperatura VS Velocidad del viento") +
  xlab("Temperatura (en °C) ") +
  ylab("Velocidad del Viento (en m/s)") + geom_smooth(method = "lm", se = FALSE, color = "red") +
  geom_point()

##Histograma marginal
g1 = ggMarginal(g_base, type = "histogram")
g1
```



Relación entre Temperatura y Velocidad del Viento

- La línea de tendencia roja, generada por una regresión lineal, muestra una pendiente muy baja, lo que sugiere que no hay una correlación fuerte entre la temperatura y la velocidad del viento.
- El coeficiente de la pendiente parece ser cercano a cero, indicando que un aumento en la temperatura no implica un cambio significativo en la velocidad del viento.

Distribución de los datos

- Se observa una alta dispersión de los puntos, lo que sugiere una relación débil o inexistente entre ambas variables.
- La velocidad del viento varía en un rango amplio incluso para temperaturas similares, lo que refuerza la falta de un patrón claro.

Histogramas marginales

- Se observa que la temperatura tiene una distribución concentrada entre 27°C y 32°C, con una mayor frecuencia en ese rango.
- La velocidad del viento parece seguir una distribución asimétrica con valores más dispersos.

Pregunta 8

$w_8 = 10$

Para los registros recolectados a medio día, describa de manera numérica la asociación entre temperatura y velocidad del viento a partir de la covarianza y el coeficiente de correlación de Pearson. ¿Parecen correlacionarse las dos variables? ¿en qué unidades se encuentran cada una de las dos medidas anteriormente mencionadas?

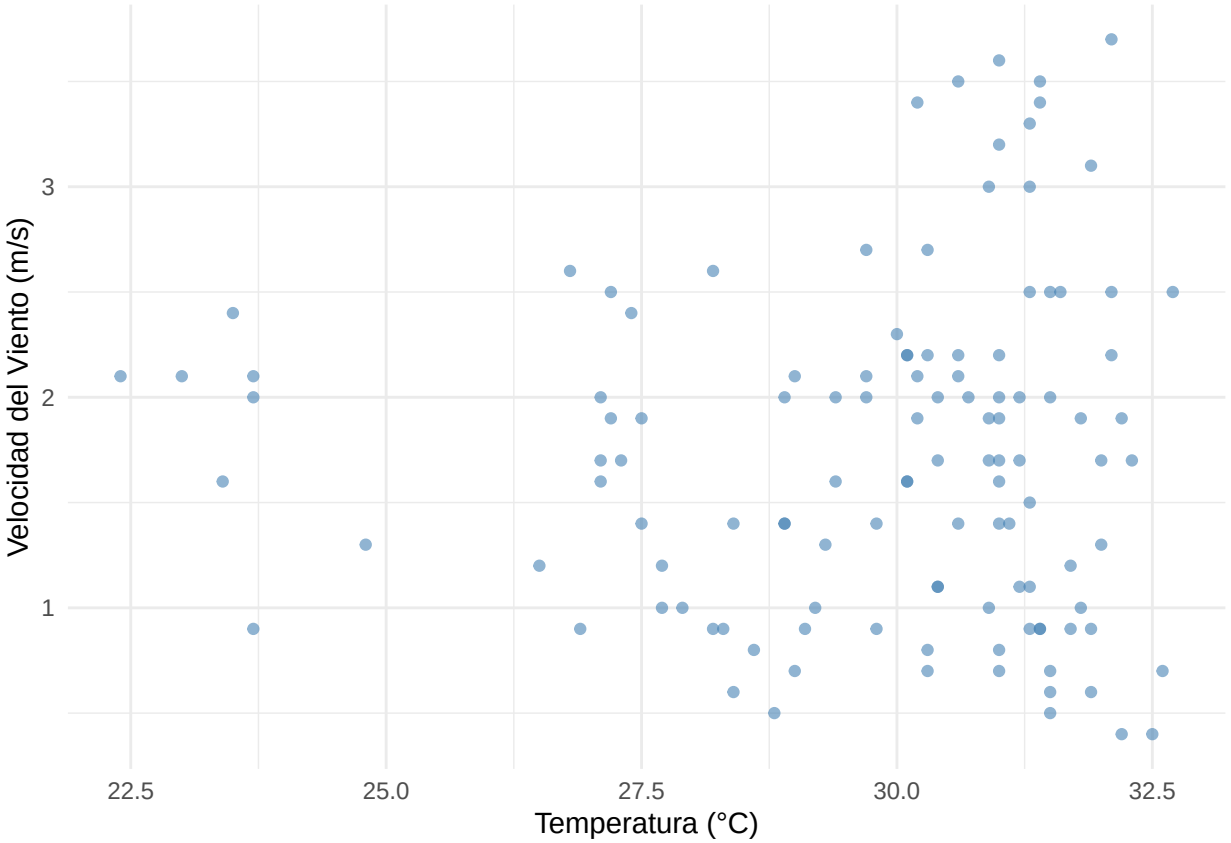
```
covar_Temp_VelViento <- cov(MedioDia$Temperatura, MedioDia$Velocidad_del_Viento, use = "complete.obs")
corrPearson_Temp_VelViento <- cor(MedioDia$Temperatura, MedioDia$Velocidad_del_Viento, use = "complete.
```

La covarianza entre la temperatura (°C) y la velocidad del viento (m/s) a medio día es 0.047 y tendria unidades de °C/(m/s) para mostrar como cambia la temperatura por cada variacion en la velocidad. El coeficiente de correlación de Pearson es 0.026 y este es adimensional. Ambos valores son muy cercanos a cero (0), lo que indica que la asociación (relacion lineal) entre ambas variables es muy debil casi nula; en otras palabras no se correlacionan. Es preciso mencionar que el coeficiente de Pearson al ser adimensional y estar en el rango de -1 a 1 facilita y agiliza la interpretacion de la correlacion entre las dos variables.

En la siguiente grafica se puede visualizar como para un mismo valor de temperatura pueden exisitir multiples velocidades de viento; tanto altos como bajos, y viceversa. En ambos casos, para un mismo valor de una de las variables se evidencian agrupaciones de la otra variable tanto altas como bajas. En otras palabras, para dias con la misma temperatura a medio dia puede haber viento (Velocidades cercanas a los 3.5m/s) o no (Velocidades cercanas a los 0.5m/s), es decir, cuando cambia una variable no necesariamente la otra cambia principalmente ambas estan influenciadas por otras variables climaticas y geograficas del punto de medida.

```
library(ggplot2)

ggplot(MedioDia, aes(x = Temperatura, y = Velocidad_del_Viento)) +
  geom_point(color = "steelblue", alpha = 0.6) +
  labs(x = "Temperatura (°C)",
       y = "Velocidad del Viento (m/s)") +
  theme_minimal()
```



En esta imagen tambien se puede mejorar el entendimeinto del `boxplot` del punto 4, al verse como la mayoria de los datos de velocidad se encuentran por debajo de 2.175m/s.

Pregunta 9

$w_9 = 5$

Para los registros recolectados a medio día, responda. Son estos datos ¿univariados, bivariados o multivariados?

Estos datos son multivariados, ya que para el medio día, en varias ubicaciones de Colombia, se realizaron mediciones de múltiples variables como: temperatura, velocidad del viento, dirección del viento, presión, punto de rocío, cobertura total nubosa, precipitación, probabilidad de tormenta, humedad y pronóstico del clima. Esto quiere decir que para cada ubicación a las 12 del día, todas estas variables mencionadas fueron medidas, y ello hace que estos datos sean multivariados.

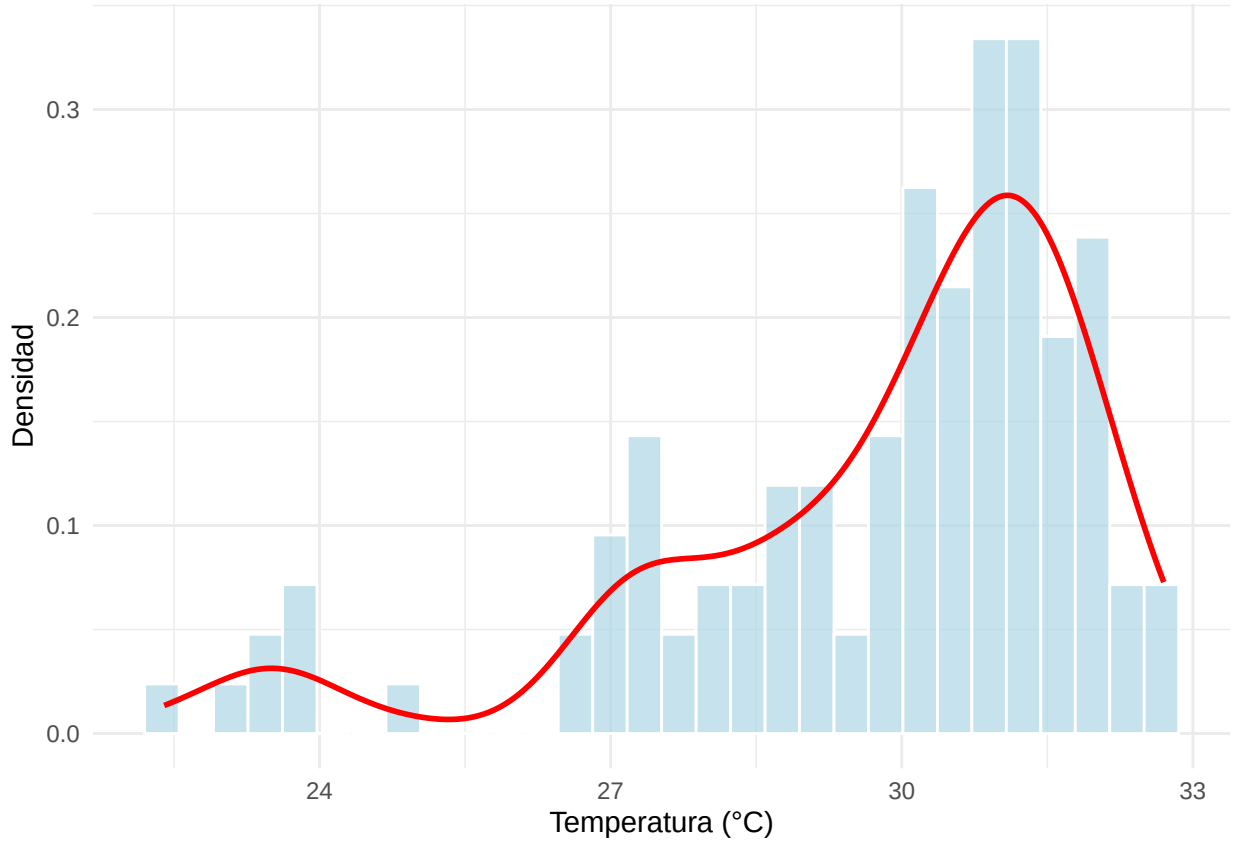
Parte 2

Pregunta 10

$w_{10} = 10$

Para los registros recolectados a medio día, elabore la estimación histograma de la densidad y la estimación kernel de la densidad para la variable Temperatura. ¿Podría afirmar que los datos provienen de una distribución normal?

```r
ggplot(MedioDia, aes(x = Temperatura)) +
  geom_histogram(aes(y = ..density..),
                 bins = 30, fill = "lightblue", color = "white", alpha = 0.7) +
  geom_density(color = "red", size = 1) +
  labs(x = "Temperatura (°C)",
       y = "Densidad") +
  theme_minimal()
```



Con base en la anteror grafica se evidencia que la variable muestra tres grupos con desidades significativas, hasta se podria indicar que tiene 3 con modas; y aunque hay una moda principal, las otras dos tambien son representativas. Se puede indicar, con base en lo anterior, que la variable temperatura es multimodal y no proviene de una distribucion normal (unimodal).

Esto se debe a que en la muestra se tienen diferentes puntos de medida ubiados en zonas dierentes y que en cada una de estas hay dias calurosos, dias frescos y dias frios, asi mismo que estas condiciones se pueden presentar en una misma hora del dia.

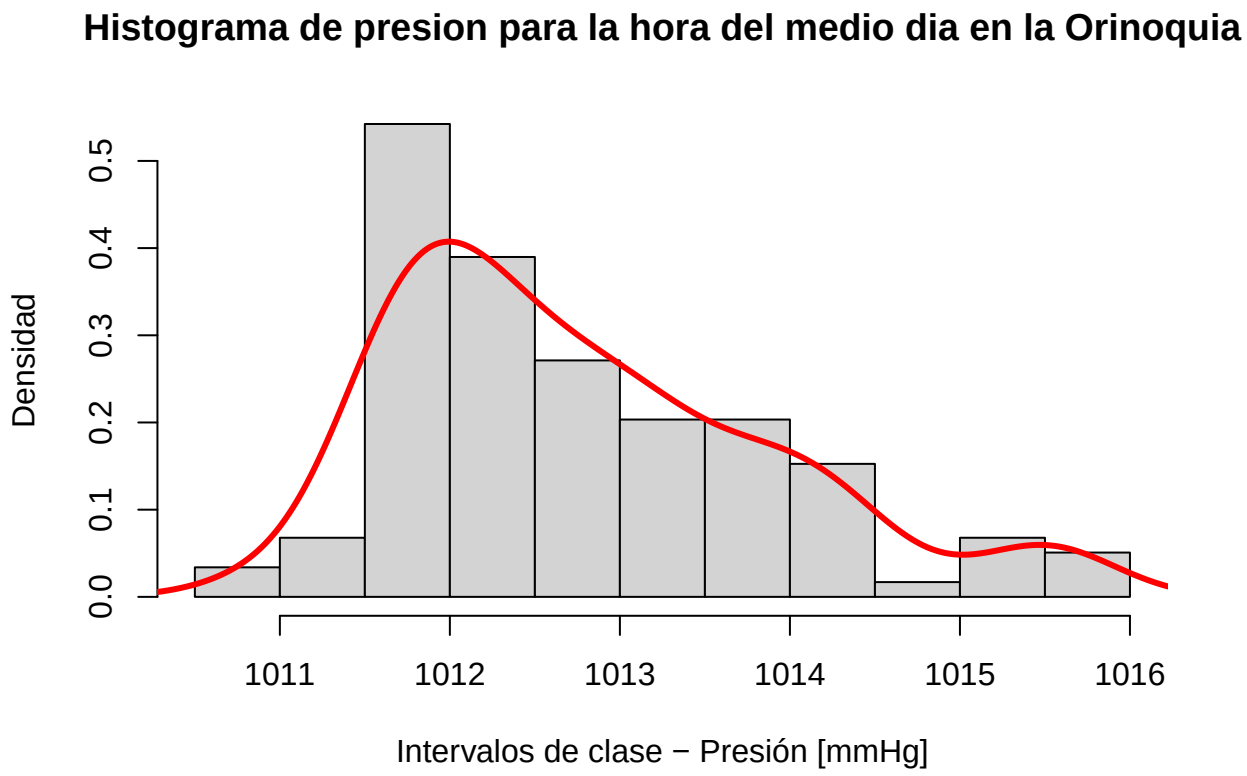
El grafico tambien muestra una estimación kernel de la densidad; tambien llamado curva suavizada. Esta curva muestra la diferencia respecto de una distribucion normal.

Pregunta 11

$w_{11} = 10$

Para los registros recolectados a medio día, elabore la estimación histograma de la densidad y la estimación kernel de la densidad para la variable Presión ¿Es la distribución estimada simétrica? explique.

```r
Hmedd2_dens = hist(MedioDia$Presion, prob=TRUE,
                   main = 'Histograma de presion para la hora del medio dia en la Orinoquia',
                   xlab = 'Intervalos de clase - Presión [mmHg]',
                   ylab = 'Densidad')
lines(density(MedioDia$Presion),
      lwd = 3,col="red")
```



En este caso, la estimación de Kernel de la densidad, tiene una distribución estimada asimétrica con un sesgo a la derecha, es decir que la densidad de los datos de la derecha se encuentra mayormente alejados de la media que se encuentra en 1012 de acuerdo al gráfico.

Pregunta 12

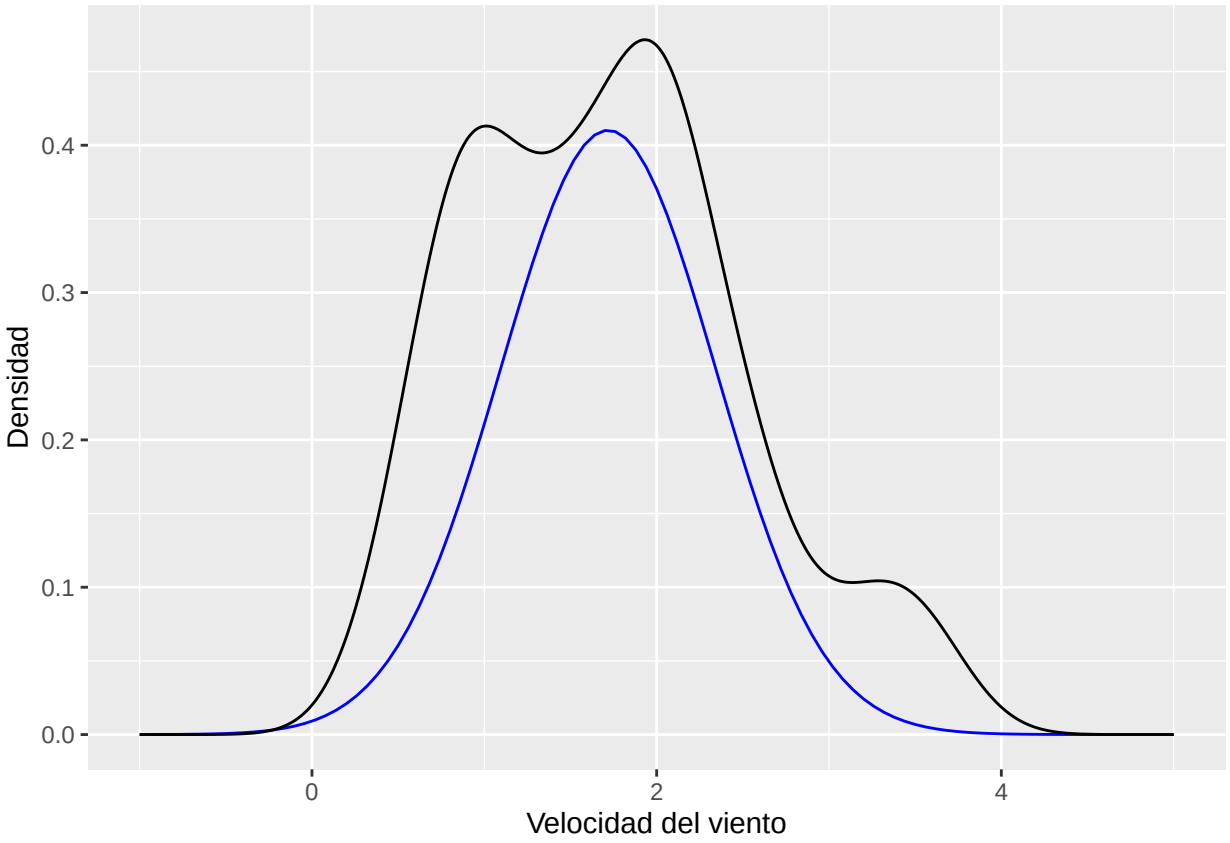
$w_{12} = 10$

Para los registros recolectados a medio día, elabore la estimación kernel de la densidad para la variable Velocidad del viento. Sobreponga en su gráfico la función de densidad normal con la media y desviación estándar observada en sus datos. ¿Se aproxima el modelo normal estimado a la estimación de la densidad encontrada?

Ayuda: Desde R, sumar `geom_density(data=datos_noon,aes(Velocidad_del_Viento))` a un `ggplot` preexistente, añade la kde de la variable de interés sobre el gráfico anterior. Revise el cuaderno de clase 1.

```r
mean_Vel=mean(MedioDia$Velocidad_del_Viento)# encontrar el valor promedio de los datos = a mu
Var_Vel=var(MedioDia$Velocidad_del_Viento)# encontrar el valor de la varianza = a sigma
f_normal_Velviento = function(x,mu,sigma) {(1/(2* pi * sigma^2)) * (exp(-0.5*((x-mu)/sigma)^2))}
```

```r
ggplot() + xlim(-1,5) +
geom_function(fun = f_normal_Velviento,args=list(mu=mean_Vel,sigma=Var_Vel) ,color = "blue")+
 geom_density(data=MedioDia,aes(x=MedioDia$Velocidad_del_Viento)) +
 labs(x="Velocidad del viento", y="Densidad")
```



Se aproxima, pero se encuentran unas diferencias marcadas en la información adicional que nos brinda con los picos que grafica Kernel, donde se encuentra las mayores densidades de los intervalos de clase, en la gráfica negra se pueden observar 3 de estos picos. Por otro lado, la gráfica de densidad normal muestra como las densidades se distribuyen de manera simétrica indicando que los datos no se encuentran tan dispersos dado que la curva no es aplanada, sino por el contrario, semejante a una distribución mesocúrtica, con un pico marcado o verosimilitud a una velocidad del viento estimada de 1.71 m/s que concuerda con el valor promedio de los datos de la velocidad del viento a las 12 de la tarde.

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