Actividad 3A. Análisis exploratorio de datos univariados.

Materia: Estadística

Alumno: Luis Fernando Izquierdo Berdugo

Fecha: 5 de Septiembre de 2024

Instrucciones:

1. Partiendo de la actividad anterior, expresar de manera breve cuáles son las aplicaciones en el ámbito profesional del análisis exploratorio de datos univariados y multivariados.
2. Con las bases de datos proporcionadas a continuación, debes generar reportes con los resúmenes estadísticos correspondientes que evidencien su capacidad de análisis de datos univariados.
3. Considera los siguientes elementos en su elaboración:

- Los reportes deberán ir acompañados con el código R correspondiente, así como gráficas y tablas resumen.
- Todo reporte deberá contener su respectiva interpretación.

Inciso 1

El análisis exploratorio de datos permite entender a fonto los datos con los que se están trabajando, lo cual permite tomar decisiones informadas sobre los análisis que se aplicarán a estos datos.

Para el ámbito univariado se pueden observar aplicaciones en el análisis de ventas, como los productos más vendidos, las temporadas de mayor demanda y los clientes más fieles; también para el análisis de las campañas de marketing y la segmentación del mercado.

Las aplicaciones del ámbito multivariado se pueden encontrar en la industra, como la segmentación de clientes conforme a características similares para personalizar las estrategias de marketing. También en la Investigación Científica puede ser útil como un método de análisis de características de imagenes para su clasificación o segmentación

Inciso 2

Ejercicio 1.

En una clase de Física, 40 alumnos obtuvieron las siguientes puntuaciones (sobre 50).

49, 48, 47, 44, 42, 41, 39, 39, 38, 38, 38, 37, 36, 36, 35, 35, 34, 34, 34, 33, 32, 32, 31, 29, 28, 28, 27, 26, 25, 24, 23, 22, 20, 17, 15, 15, 13, 13, 11, 10

Calcule las medidas de tendencia central y de dispersión, además encuentre el primer y tercer Cuartil, así como el cuarto y sexto decil. Por último elabore el análisis gráfico de estos datos.

Lo primero será la declaración de las calificaciones

```
calificaciones <- c(49, 48, 47, 44, 42, 41, 39, 39, 38, 38, 38, 37, 36, 36, 35, 35, 34,
                    34, 34, 33, 32, 32, 31, 29, 28, 28, 27, 26, 25, 24, 23, 22, 20, 17,
                    15, 15, 13, 13, 11, 10)
```

Ya con la variable de calificaciones, se procede a hacer las **medidas de tendencia central**:

- Tamaño de Muestra

```
n <- length(calificaciones)
n
```

```
## [1] 40
```

- Valor mínimo

```r
min(calificaciones)
```

```
## [1] 10
```

- Valor máximo

```r
max(calificaciones)
```

```
## [1] 49
```

- Media

```r
me <- mean(calificaciones)
me
```

```
## [1] 30.45
```

- Mediana

```r
median(calificaciones)
```

```
## [1] 32.5
```

- Moda

```r
library(modeest)
mfv(calificaciones)
```

```
## [1] 34 38
```

Se aplicaran de igual manera las **medidas de dispersión**:

- Varianza

```r
var(calificaciones)
```

```
## [1] 109.8436
```

- Desviación Estándar

```r
s <- sd(calificaciones)
s
```

```
## [1] 10.48063
```

- Rango

```r
range(calificaciones)
```

```
## [1] 10 49
```

- Rango intercuartílico

```r
quantile(calificaciones, 0.75) - quantile(calificaciones, 0.25)
```

```
##   75%
## 14.25
```

- Coeficiente de Variación

```r
sd(calificaciones) / mean(calificaciones)*100
```

```
## [1] 34.41914
```

Se busca el **primer y tercer cuartil**:

```r
primer <- quantile(calificaciones, 0.25)
tercero <- quantile(calificaciones, 0.75)

primer
```

```
##   25%
## 23.75
```

```r
tercero
```

```
## 75%
##  38
```

Se procede a buscar el **cuarto y sexto decil**:

```r
cuarto <- quantile(calificaciones, 0.4)
sexto <- quantile(calificaciones, 0.6)

cuarto
```

```
##  40%
## 28.6
```

```r
sexto
```

```
##  60%
## 34.4
```

Medidas de Forma:

- Coeficiente de Asimetría

```r
(sum(calificaciones-me))^3/(n*s^3)
```

```
## [1] 4.985727e-46
```

- Coeficiente de Curtosis

```r
((sum(calificaciones-me))^4/(n*s^4))-3
```

```
## [1] -3
```

Análisis

Gracias a las medidas de tendencia central, se sabe que los alumnos promediaron poco más de la mitad de los puntos posibles (Media de 30.45), con la mediana de 32.5, sabemos que la mitad de los alumnos sacaron 32.5 o más de calificación y la otra mitad sacó menos que eso.

MTC	Tamaño de Muestra	Valor mínimo	Media	Mediana	Moda
Resultado	40	10	30.45	32.5	34 y 38

Con la desviación estándar se observa que los puntajes de los estudiantes se desvían 10.48063 de la media de 30.45, la varianza con valor alto 109.8436 indica que los datos están muy dispersos alrededor de la media.

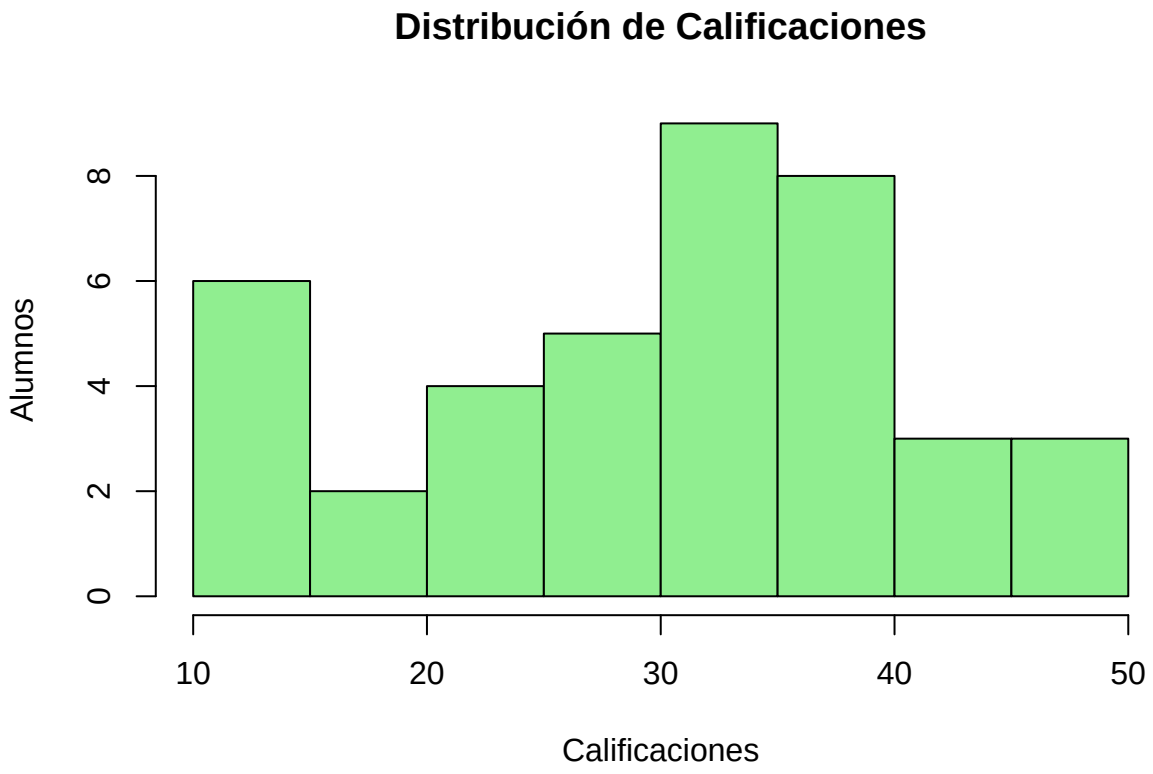
En el caso de no tener los datos a la vista, el rango mostraría que los estudiantes obtuvieron calificaciones entre 10 y 49. El coeficiente de Variación es más alto que la media indica una distribución bastante dispersa con respecto a esta

Medidas de Dispersión	Varianza	Desviación Estándar	Rango	Rango Intercuartílico	Coeficicente de Variación
Resultado	109.8436	10.48063	10 - 49	75% 14.25	34.41914

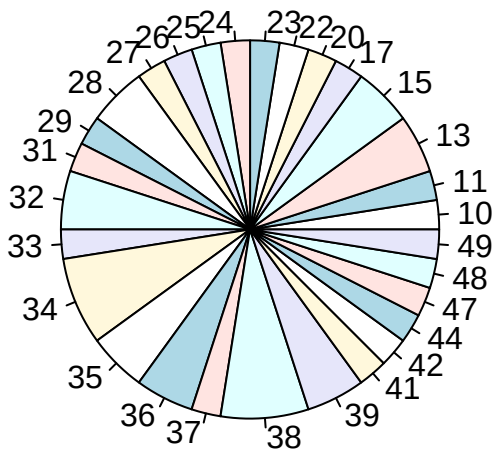
Como el coeficiente de asimetría es computacionalmente cero ($4.985727e-46$) se sabe que los datos son simétricos. De igual manera los datos tienen una distribución leptocurtica (Al hacer un gráfico de campana, el pico central es más pronunciado y las colas se extienden más hacia los extremos), esto se sabe porque el coeficiente de curtosis es menor a 0 (-3)

Análisis Gráfico:

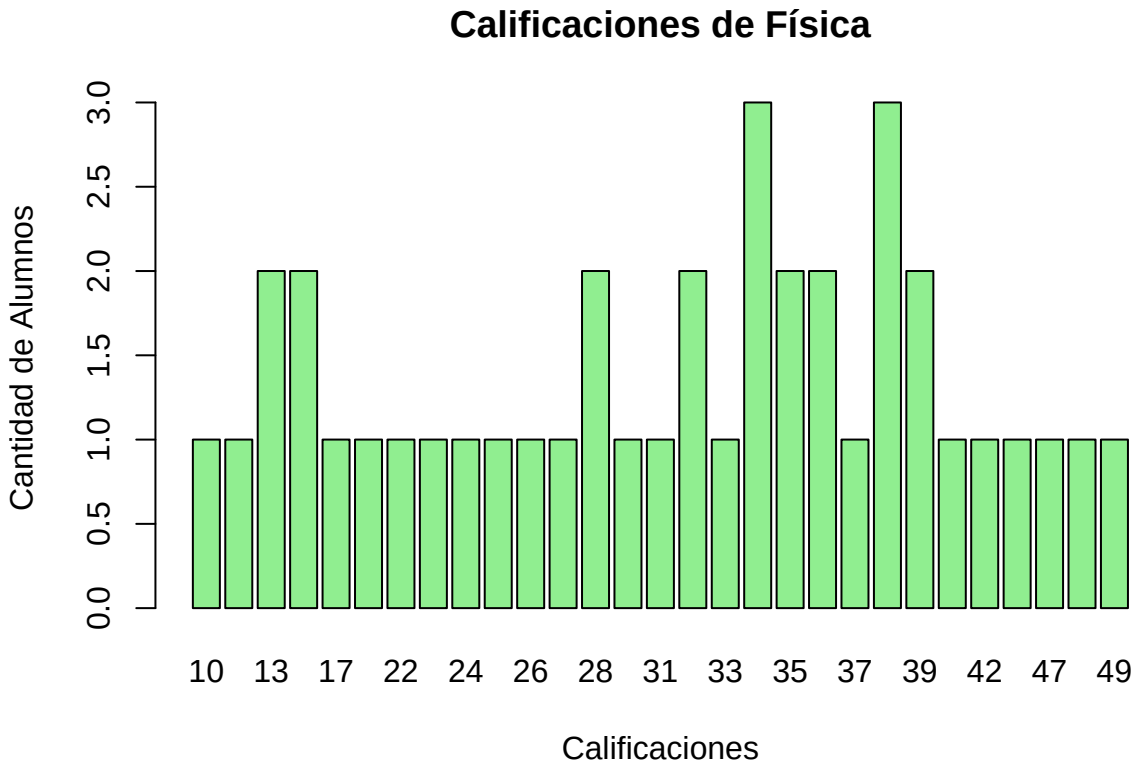
```
hist(calificaciones, main = "Distribución de Calificaciones", xlab = "Calificaciones",
     ylab = "Alumnos", col = "lightgreen", breaks=10)
```




```
pie(c(table(calificaciones)), main="Calificaciones de Física")
```




```
barplot(c(table(calificaciones)), col="lightgreen", main = "Calificaciones de Física",
        ylab = "Cantidad de Alumnos", xlab="Calificaciones")
```



Ejercicio 2.

En una compañia aseguradora de autos en los dos últimos años tuvo que entregar las siguientes cantidades como indemnización (en miles de pesos) en el rubro de robo.

500, 500, 495, 490, 473, 441, 429, 419, 405, 400, 390, 390, 390, 376, 353, 350, 300, 258, 240, 220, 210, 200, 192, 190, 150, 130, 125, 110, 100, 100

Calcule las medidas de tendencia central y de dispersión, además encuentre el percentil 0.16 y 0.84, así como

el primer, quinto y noveno decil.

Por último elabore el análisis gráfico de estos datos.

Lo primero será generar el conjunto de datos:

```
ind <- c(500, 500, 495, 490, 473, 441, 429, 419, 405, 400, 390, 390, 390, 376,
         353,350, 300, 258, 240, 220, 210, 200, 192, 190, 150, 130, 125, 110,
         100, 100)
```

Ya con la variable de indemnizaciones, se procede a hacer las **medidas de tendencia central**:

- Tamaño de Muestra

```
n <- length(ind)
n
```

```
## [1] 30
```

- Valor mínimo

```
min(ind)
```

```
## [1] 100
```

- Valor máximo

```
max(ind)
```

```
## [1] 500
```

- Media

```
me <- mean(ind)
me
```

```
## [1] 310.8667
```

- Mediana

```
median(ind)
```

```
## [1] 351.5
```

- Moda

```
library(modeest)
mfv(ind)
```

```
## [1] 390
```

Se aplicaran de igual manera las **medidas de dispersión**:

- Varianza

```
var(ind)
```

```
## [1] 18598.53
```

- Desviación Estándar

```
s <- sd(ind)
s
```

```
## [1] 136.3764
```

- Rango

```r
range(ind)
```

```
## [1] 100 500
```

- Rango intercuartílico

```r
quantile(ind, 0.75) - quantile(ind, 0.25)
```

```
##   75%
## 221.5
```

- Coeficiente de Variación

```r
sd(ind) / mean(ind)*100
```

```
## [1] 43.86975
```

Se busca el **primer, quinto y noveno decil**:

```r
primer <- quantile(ind, 0.1)
quinto <- quantile(ind, 0.5)
noveno <- quantile(ind, 0.9)

primer
```

```
##   10%
## 123.5
```

```r
quinto
```

```
##   50%
## 351.5
```

```r
noveno
```

```
##   90%
## 490.5
```

Se procede a buscar los **percentiles 0.16 y 0.84**:

```r
p1 <- quantile(ind, 0.16)
p2 <- quantile(ind, 0.84)

p1
```

```
##   16%
## 142.8
```

```r
p2
```

```
##    84%
## 452.52
```

Medidas de Forma:

- Coeficiente de Asimetría

```r
(sum(ind-me))^3/(n*s^3)
```

```
## [1] -1.544832e-46
```

- Coeficiente de Curtosis

```r
((sum(ind-me))^4/(n*s^4))-3
```

```
## [1] -3
```

Análisis

MTC	Tamaño de Muestra	Valor mínimo	Valor máximo	Media	Mediana	Moda
Resultado	30	100	500	310.8667	351.5	390

El tamaño de la muestra de 30 da una idea general de cuantos autos se robaron en los últimos 2 años. Con la media se puede observar que en 2 años de la aseguradora se entrega un promedio de 310.8667 miles de pesos como indemnización por cuestión de robo de vehículos.

Gracias a la mediana se sabe que el 50% de las indemnizaciones son iguales o menores a 351.5 miles de pesos y el otro 50% son iguales o mayores. Con la moda se nota que el costo más común en los casos de indemnización es de 390 mil pesos, esto podría indicar que hay cierto modelo de coche que se roba más, sin embargo, asumirlo solamente con este dato podría ser un error.

Medidas de Dispersión	Varianza	Desviación Estándar	Rango	Rango Intercuartílico	Coeficicente de Variación
Resultado	18598.53	136.3764	100-500	75% 221.5	43.86975

Al efectuar la varianza, se encuentra un valor de 18598.53, debido al valor elevado, se puede asumir que existen grandes diferencias entre todos los montos de indemnización.

Con la desviación estándar se observa que, en promedio, los montos de las indemnizaciones se desvían 136.3764 miles de pesos de la media.

El rango describe la amplitud del conjunto de datos, en donde encontramos una diferencia de 400 mil pesos entre el monto mínimo de indemnización (100 mil pesos) y el monto máximo (500milpesos) presentados en los dos años analizados.

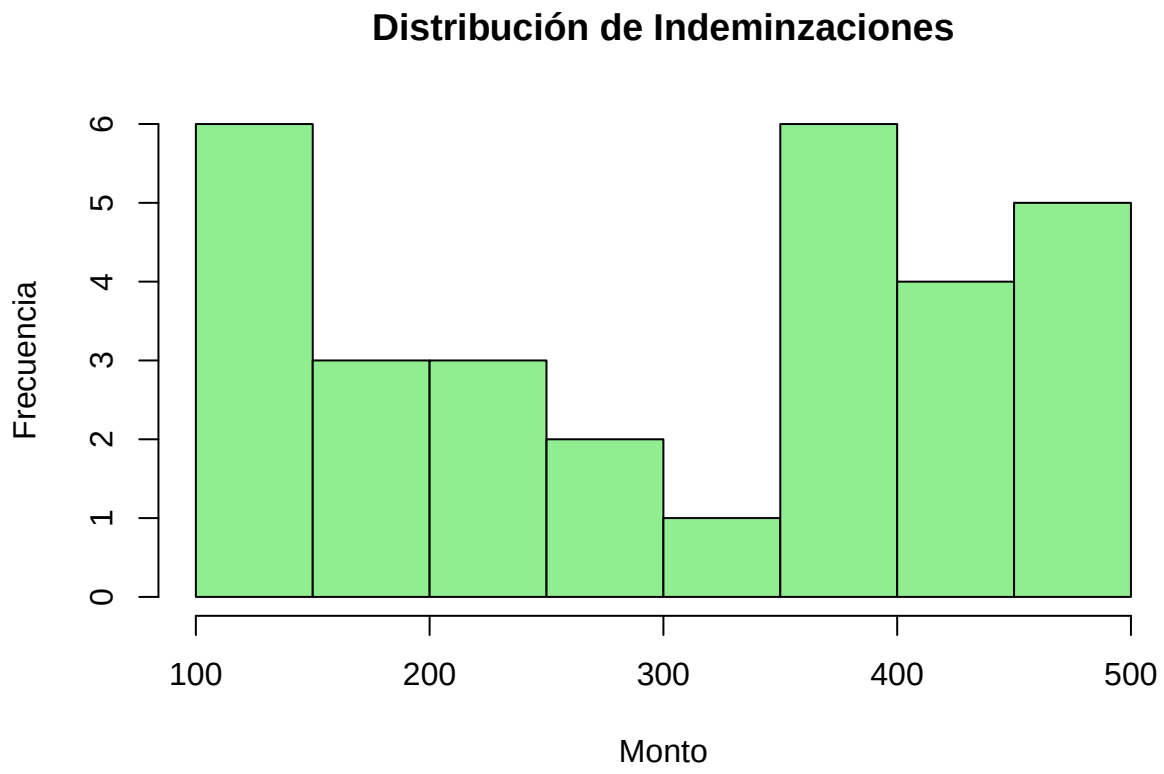
Se encuentra que los datos centrales están más concentrados al analizar el rango intercuartílico, que es la diferencia entre el tercer cuartil y el primero.

El coeficiente de variación de 43.86975 % indica que existe una variabilidad considerable entre los datos de indemnización con respecto a su media, por lo que se observa,hay indemnizaciones mucho más bajas que la media (310.8667), al mismo tiempo que hay mucho más altas.

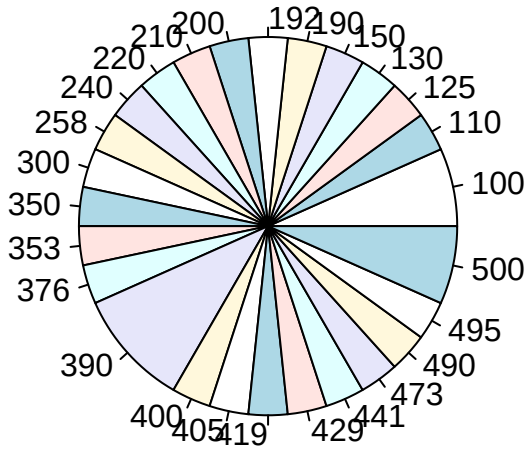
Como el coeficiente de asimetría es computacionalmente cero ($-1.544832e-46$) se sabe que los datos son simétricos. De igual manera los datos tienen una distribución leptocurtica, esto se sabe porque el coeficiente de curtosis es menor a 0 (-3)

Análisis Gráfico:

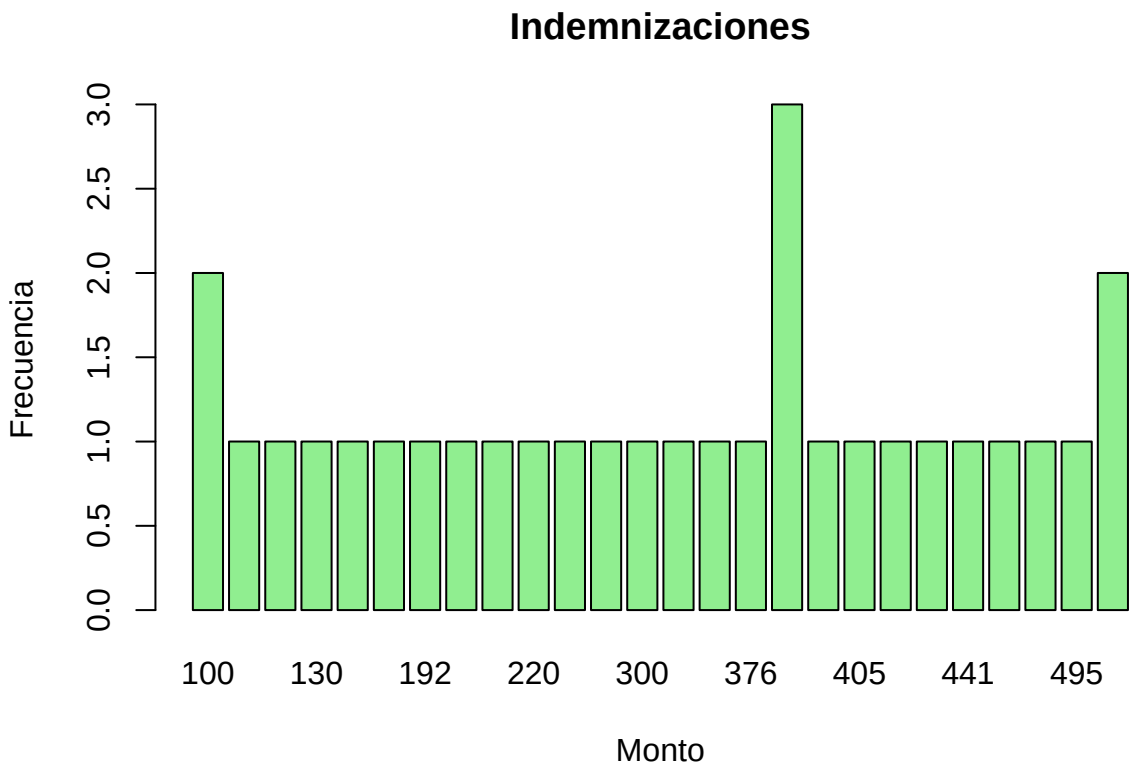
```r
hist(ind, main = "Distribución de Indeminzaciones", xlab = "Monto",
     ylab = "Frecuencia", col = "lightgreen", breaks=10)
```




```
pie(c(table(ind)), main="Indemnizaciones")
```




```
barplot(c(table(ind)), col="lightgreen", main = "Indemnizaciones",
        ylab = "Frecuencia", xlab="Monto")
```



Ejercicio 3.

En una empresa trabajan 36 personas, le hicieron 2 preguntas ¿Sexo? y ¿Estado Civíl?, los resultados fueron los siguientes:

Sexo	Edo. Civil	Sexo	Edo. Civil	Sexo	Edo. Civil	Sexo	Edo. Civil
Hombre	Soltero	Hombre	Casado	Mujer	Soltero	Hombre	Casado
Mujer	Casado	Mujer	Casado	Hombre	Casado	Mujer	Soltero
Hombre	Soltero	Mujer	Soltero	Mujer	Soltero	Hombre	Casado
Hombre	Casado	Hombre	Soltero	Mujer	Soltero	Mujer	Casado
Hombre	Soltero	Mujer	Soltero	Mujer	Soltero	Hombre	Soltero
Mujer	Casado	Hombre	Soltero	Hombre	Casado	Mujer	Soltero
Mujer	Casado	Mujer	Casado	Hombre	Soltero	Mujer	Casado
Mujer	Soltero	Hombre	Soltero	Mujer	Casado	Mujer	Casado
Hombre	Soltero	Hombre	Soltero	Mujer	Soltero	Hombre	Casado

Realice la tabla de contingencia correspondiente, así como la tabla marginal y condicional por Sexo y por Estado Civìl, concluya el análisis.

Primero se genera el dataframe con los datos proporcionados.

```
Sexo <-c("Hombre","Mujer","Hombre","Hombre","Hombre","Mujer","Mujer","Mujer","Hombre","Hombre","Mujer",
EdoCivil <- c("Soltero","Casado","Soltero","Casado","Soltero","Casado","Casado","Soltero","Soltero","Ca

df <- data.frame(Sexo,EdoCivil)
```

Tablas de Contingencia

Total de Hombres y Mujeres que están o no casados.

```r
ftable(df)
```

```
##        EdoCivil Casado Soltero
## Sexo
## Hombre                7      10
## Mujer                 9      10
```

Total de datos en la tabla

```r
sum(ftable(df))
```

```
## [1] 36
```

Proporción de hombres y mujeres que están o no solteros (respecto al total de personas).

```r
prop.table(table(df))
```

```
##         EdoCivil
## Sexo        Casado   Soltero
##   Hombre 0.1944444 0.2777778
##   Mujer  0.2500000 0.2777778
```

Condicional de casados o solteros según el sexo.

```r
prop.table(table(df),1)
```

```
##         EdoCivil
## Sexo        Casado   Soltero
##   Hombre 0.4117647 0.5882353
##   Mujer  0.4736842 0.5263158
```

Condicional de hombres y mujeres según si están solteros o casados

```r
prop.table(table(df),2)
```

```
##         EdoCivil
## Sexo     Casado Soltero
##   Hombre 0.4375  0.5000
##   Mujer  0.5625  0.5000
```

Análisis

Como se puede observar en las distintas tablas de contingencia, se observa que hay una mayor cantidad de personas solteras, siendo este porcentaje equitativo entre ambos sexos. De igual manera, se concluye que existe la misma cantidad de hombres y mujeres solteros en la oficina, mientras que los casados siguen una relacion 56.25% de mujeres contra 43.75% hombres.

En este caso no se pueden obtener la **Covarianza** ni la **Correlación** debido a que ambas asumen variables numéricas y continuas. Incluso asignando valores (como 0 y 1), no habría una interpretación clara en términos de variables.

Ejercicio 4.

Con los datos del dataset `mtcars`. Calcule la matriz de varianza-covarianza, la matriz de correlación. Además obtenga los gráficos multivarados correspondientes.

```r
df_cars <- datasets::mtcars[,c(1,3:6)]
```

Se tendrán en cuenta los siguientes datos:

- mpg = Millas por galón
- cyl = Cilindros

- disp = Desplazamiento del motor
- hp = Caballos de fuerza
- am = Transmisión (manual o automática)

Medidas de asociación entre variables

Matriz de Varianza-Covarianza

```
cov(df_cars)
```

```
##              mpg        disp         hp        drat          wt
## mpg    36.324103  -633.09721 -320.73206   2.1950635  -5.1166847
## disp -633.097208 15360.79983 6721.15867 -47.0640192 107.6842040
## hp   -320.732056  6721.15867 4700.86694 -16.4511089  44.1926613
## drat    2.195064   -47.06402  -16.45111   0.2858814  -0.3727207
## wt     -5.116685   107.68420   44.19266  -0.3727207   0.9573790
```

Matriz de correlación

```
cor(df_cars)
```

```
##             mpg       disp         hp       drat         wt
## mpg   1.0000000 -0.8475514 -0.7761684  0.6811719 -0.8676594
## disp -0.8475514  1.0000000  0.7909486 -0.7102139  0.8879799
## hp   -0.7761684  0.7909486  1.0000000 -0.4487591  0.6587479
## drat  0.6811719 -0.7102139 -0.4487591  1.0000000 -0.7124406
## wt   -0.8676594  0.8879799  0.6587479 -0.7124406  1.0000000
```

```
andrews.function <- function (xs, no.pts=101){
n <- length(xs)
xpts <- seq(0, 2*pi, length=no.pts)
ypts <- c()
for (p in xpts){
y <- xs[1]
for (i in 2:n){
if (i %% 2 == 1) {y <- y + xs[i]*sin((i %/% 2)*p)}
else {y <- y + xs[i]*cos((i %/% 2)*p)}
}
ypts <- c(ypts, y)
}
return(ypts)
}
andrews.curves <- function(xdf, cls, npts=101, title="Classes"){
n <- nrow(xdf)
clss <- as.factor(cls)
xpts <- seq(-pi, pi, length=npts)
X <- xpts
for (i in 1:n){
xi <- unname(unlist(xdf[i, ]))
ys <- andrews.function(xi, npts)
X <- cbind(X, ys)
}
ymin <- min(X[, 2:(n+1)])
ymax <- max(X[, 2:(n+1)])
plot(0, 0, type="n", xlim=c(-pi, pi), ylim=c(ymin, ymax),
main="Curvas de Andrews", xlab="", ylab="")
clrs <- as.integer(clss)
```

```
for (i in 2:(n+1)){
lines(X[, 1], X[, i], col=clrs[i-1])
}
legend(4, ymax, levels(clss), col=c(1:nlevels(clss)), lty=1)
}
par(mfrow=c(2,2))
```

Análisis

Con la matriz de covarianza vemos las siguientes relaciones:

- mpg y disp: A mayor desplazamiento del motor (disp), menor es el consumo de combustible (mpg)
- mpg y hp: A mayor potencia (hp), mayor consumo (mpg).
- mpg y wt: Los coches más pesados (wt) son menos eficientes en el consumo de combustible (mpg).
- disp y hp: Mientras más potencia, mayor será el desplazamiento
- wt y hp: Loches más pesados tienden a tener motores más potentes.

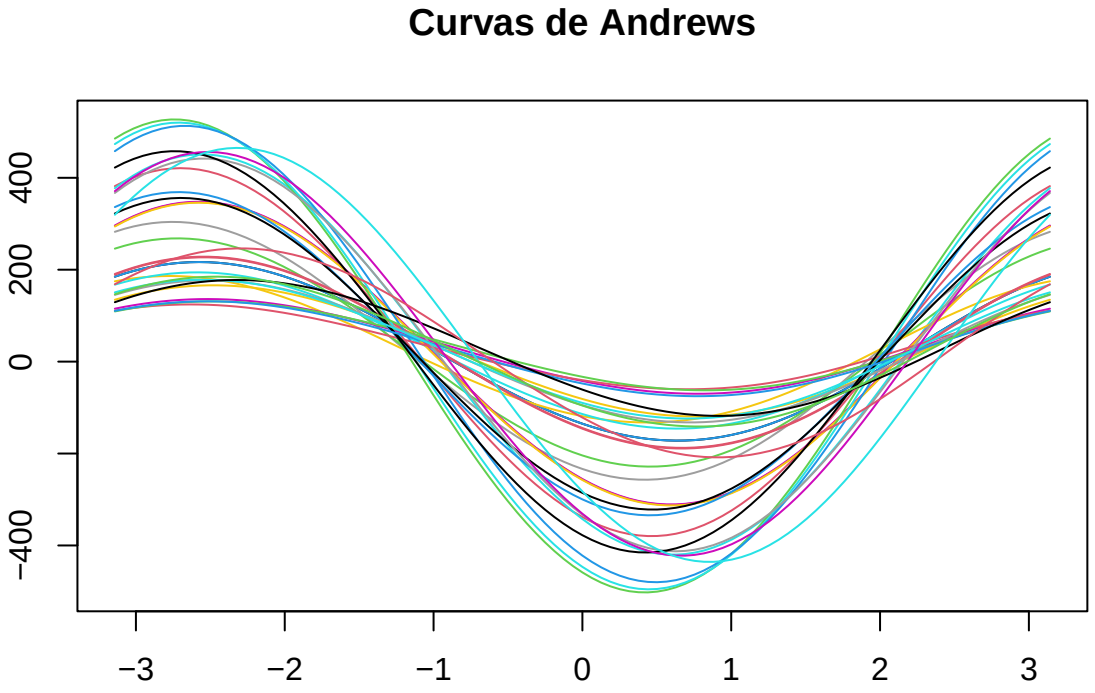
Con la matriz de correlación se pueden obtener las siguientes conclusiones: - mpg y disp: La correlación es -0.847, lo que confirma la fuerte relación negativa observada en la matriz de covarianza. - mpg y hp: La correlación es -0.776, confirmando lo observado previamente. - mpg y wt: La correlación es -0.867, confirma que coches más pesados son menos eficientes en el consumo de combustible. - disp y hp: La correlación es 0.791, lo que indica que los motores con mayor desplazamiento suelen tener más potencia. - wt y hp: La correlación es 0.659, confirma lo observado en la matriz de covarianza.

Análisis Gráfico

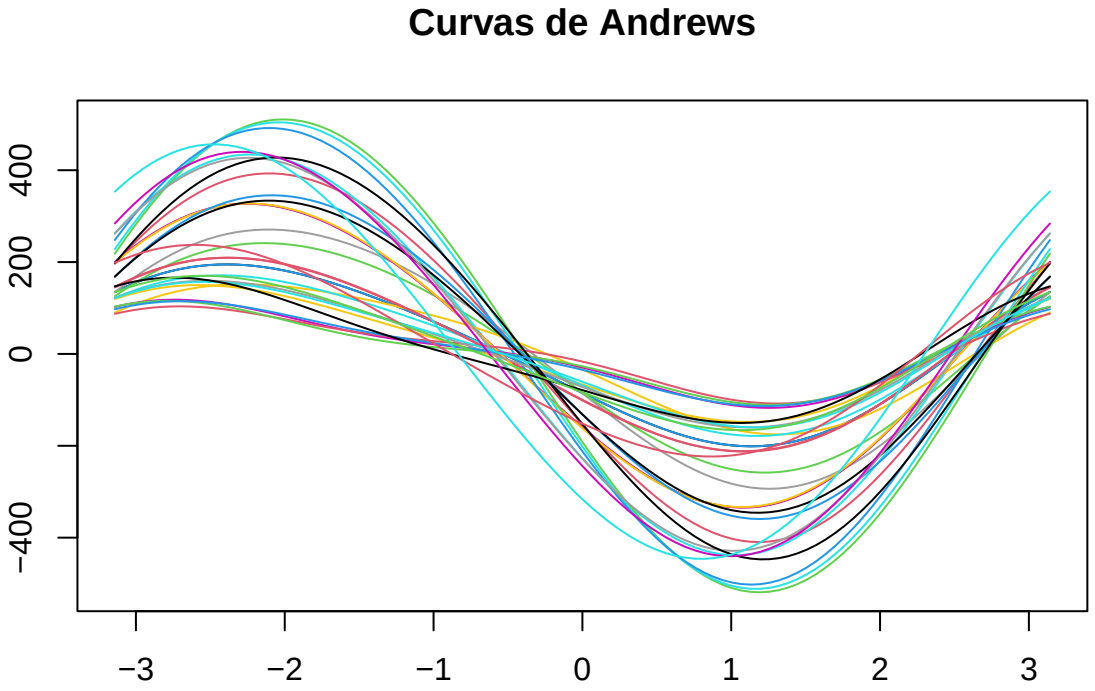
Se harán cuatro curvas de Andrews, las cuales serán

- mpg, cyl, disp, hp vs. am
- hp, disp, cyl, mpg vs. am
- hp, cyl, mpg, disp vs. am
- mpg, disp, cyl, hp vs. am

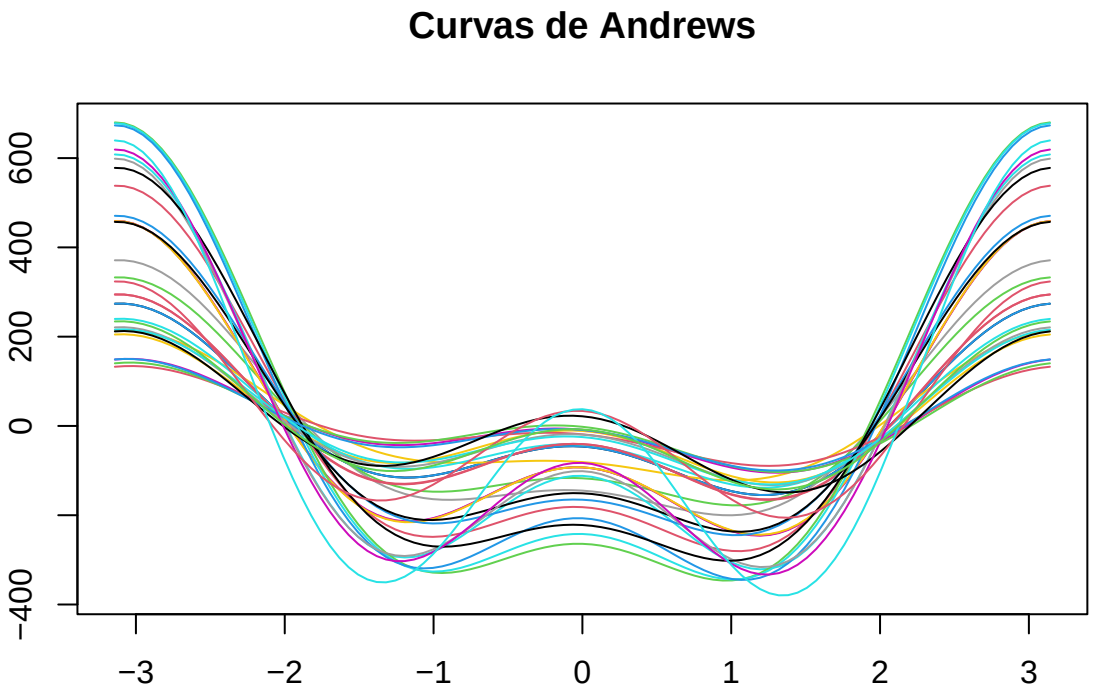
```
andrews.curves(df_cars[,1:4], df_cars[,5])
```



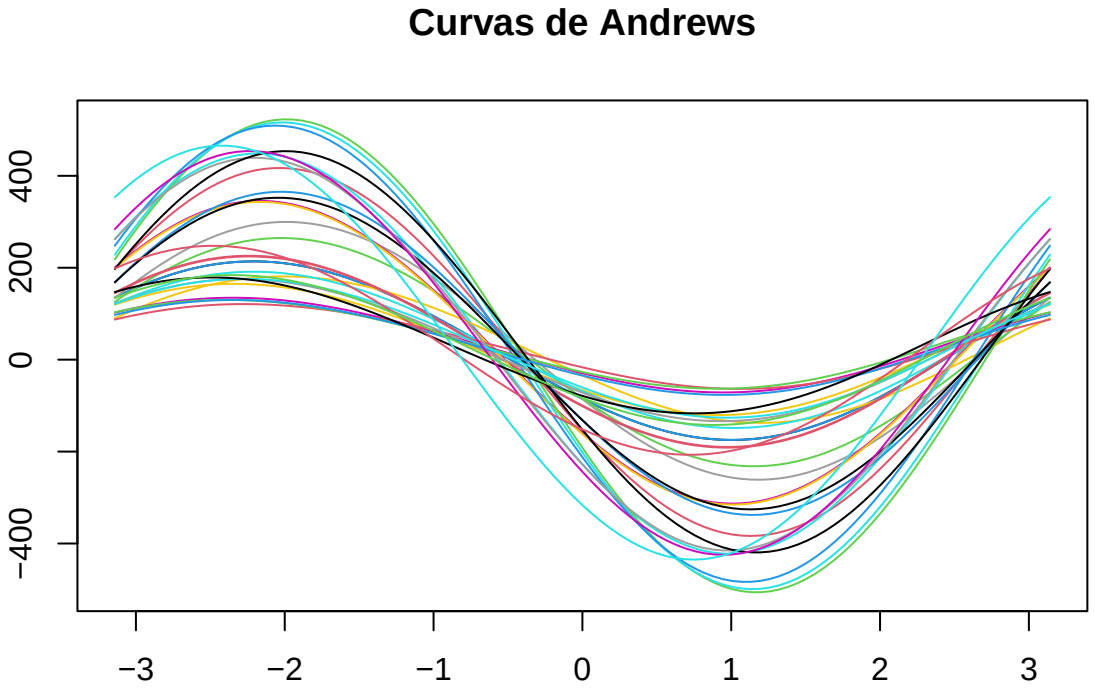
```
andrews.curves(df_cars[,4:1], df_cars[,5])
```




```
andrews.curves(df_cars[,c(4,2,1,3)], df_cars[,5])
```




```
andrews.curves(df_cars[,c(1,3,2,4)], df_cars[,5])
```



Se observa un comportamiento similar de las curvas en cada uno de los datos, lo cual puede implicar que la transmisión manual o automática no representa un factor diferencial en los modelos.

Ejercicio 5.

Con los datos del dataset `USArrests` calcule la matriz de varianza-covarianza, la matriz de correlación. Además obtenga los gráficos multivarados correspondientes.

```
states_selected <- c("Arizona", "Connecticut", "Illinois", "Iowa", "Kansas", "Kentucky", "Michigan","Mis
data("USArrests")
df_arrests <- USArrests[rownames(USArrests) %in% states_selected, ]
```

Medidas de asociación entre variables

Matriz de Varianza-Covarianza

```
cov(df_arrests)
```

```
##             Murder   Assault  UrbanPop      Rape
## Murder    17.07473  286.6973 -10.28867  16.94913
## Assault  286.69733 7239.7233 132.82167 411.86217
## UrbanPop -10.28867  132.8217 193.64333  42.98267
## Rape      16.94913  411.8622  42.98267  58.01427
```

Matriz de correlación

```
cor(df_arrests)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder    1.0000000 0.8154282 -0.1789291 0.5385216
## Assault   0.8154282 1.0000000  0.1121777 0.6355116
## UrbanPop -0.1789291 0.1121777  1.0000000 0.4055316
## Rape      0.5385216 0.6355116  0.4055316 1.0000000
```

Análisis

Gracias a la Matriz de Varianza-Covarianza se encontraron las siguientes relaciones: - Assault y Murder:

Indica que los estados con mayor número de asesinatos suelen tener altas tasas de asalto. - Assault y UrbanPop: Indica que los estados de mayor población tienen más número de asalto. - Murder y UrbanPop: La covarianza es negativa, indica que los estados con mayor población tienen menor número de asesinatos, sin embargo, la diferencia es muy ligera.

La Matriz de Correlacion nos indica que lo mostrado en la de Varianza-Covarianza es correcto: - Murder y Assault: La correlación de 0.815 indica una fuerte correlación positiva. - Assault y UrbanPop: La correlación de 0.112 indica una correlación positiva débil. - Murder y UrbanPop: La correlación de -0.179 indica una correlación negativa débil.

Análisis Gráfico Se efectúa una gráfica de Curvas de Andrews de las diferentes razones de arrestos contra la población de cada ciudad. En esta se observa un patrón bastante consistente y la forma de U sugiere una relación no lineal entre las razones de arresto y la población; el mínimo alrededor de $x = 0$ podría indicar que los estados con un nivel de urbanización medio (cerca de 0) tienden a tener las tasas de crimen más bajas para estas variables específicas y los máximos en los extremos sugieren que tanto los estados muy urbanizados como los muy rurales podrían tener tasas de crimen más altas.

```
andrews.curves(df_arrests[,c(1,2,4)], df_arrests[,3])
```