Trabajo 1

Ivan Santiago Rojas Martinez

Estudiante de Pregrado en Estadística

Docente

Rene Iral Palomino

Asignatura

Introducción al Análisis Multivariado

Sede Medellín
Septiembre 2 de 2023

Índice

Índice de figuras

Índice de cuadros

Trabajo 1

Selección de la muestra de datos

Se incluye el código propuesto por el docente, con la intención de validar la extracción de la muestra

1 Primer Punto

Para todas sus variables realice un análisis exploratorio gráfico e identifique posibles valores atípicos u otro tipo de anomalías. (Para las variables Categóricas diagramas de barras, para las continuas o discretas, use Histogramas y/o Boxplot). Comente brevemente.

1.1 Análisis Descriptivos

Breve descripción de la base de datos: La base de datos corresponde a las medidas antropométricas de la población laboral colombiana (ACOPLA). Esta base de datos cuenta con 200 observaciones y 9 variables de interés, las cuales son:

-**Sexo:** Variable categórica (Hom, Muj)

-**P1: Masa Corporal** Variable continua (kg)

-**P7: Perímetro muslo mayor** Variable continua (cm)

-**P16: Perímetro abdominal cintura** Variable continua (cm)

-**P22: Anchura de las caderas** Variable continua (cm)

-**P27: Longitud promedio de los pies** Variable continua (cm)

-**P29: Longitud promedio de las manos** Variable continua (cm)

-**P38: Estatura** Variable continua (cm)

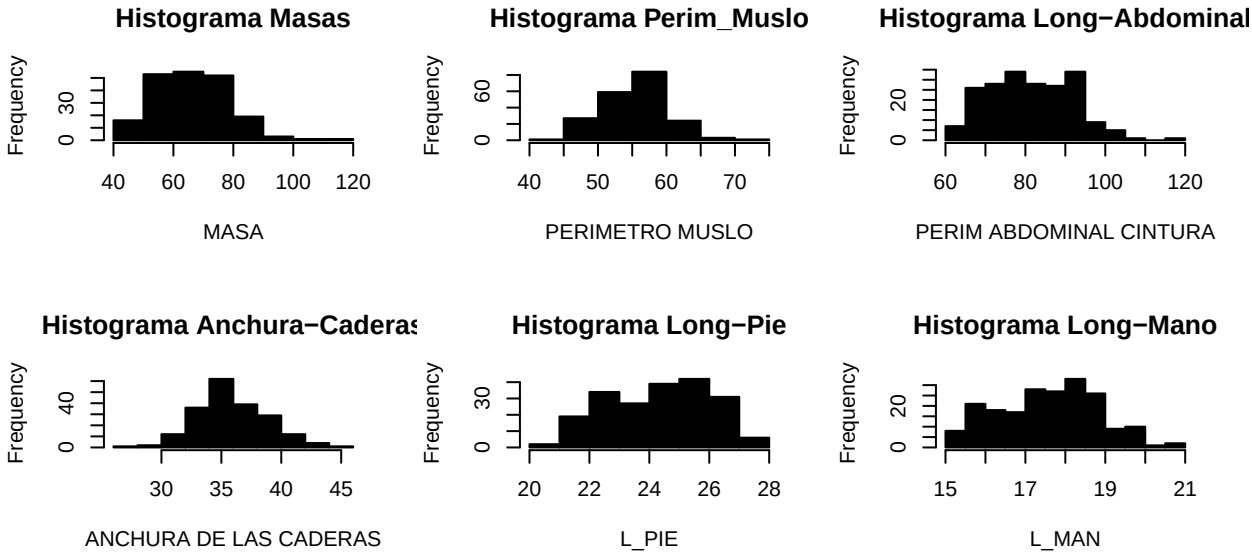
-**CAT_IMC: Categoría del indice de masa corporal** Variable categórica (DELGADO, NORMAL Y OBESO)

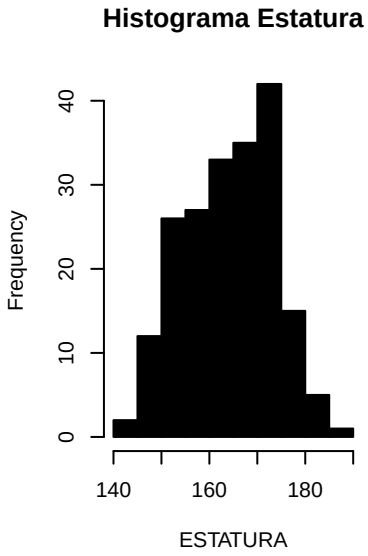
1.2 Resumen Numerico

Cuadro 1: Tabla de resúmenes estadísticos de variables continuas

Variable	Media	Mediana	SD	Q1	Q2	Q3	Rango.intercuartil	Rango
P1	66.11600	64.80	12.140225	56.475	64.80	74.225	17.75	71.5
P7	55.47688	55.50	4.741592	52.300	55.50	58.650	6.35	29.2
P16	81.65200	81.70	10.163656	74.125	81.70	90.075	15.95	56.1
P22	35.87626	35.60	2.963242	34.000	35.60	37.800	3.80	16.3
P27	24.33600	24.50	1.681380	22.800	24.50	25.700	2.90	7.2
P29	17.63400	17.70	1.246480	16.700	17.70	18.500	1.80	6.0
P38	164.14949	164.95	9.178836	156.700	164.95	171.450	14.75	43.6

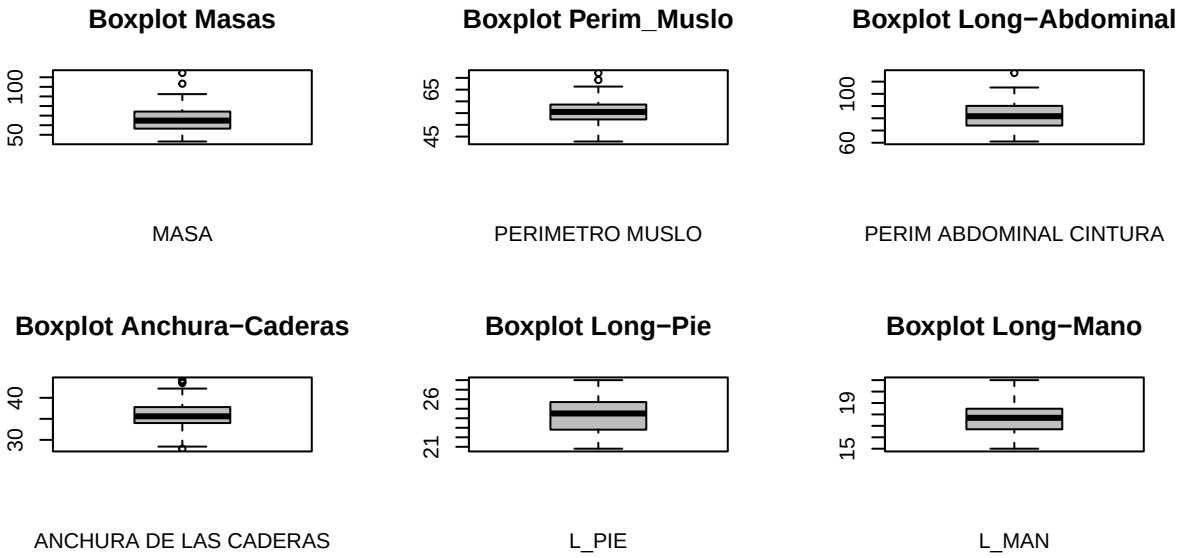
1.3 Histogramas



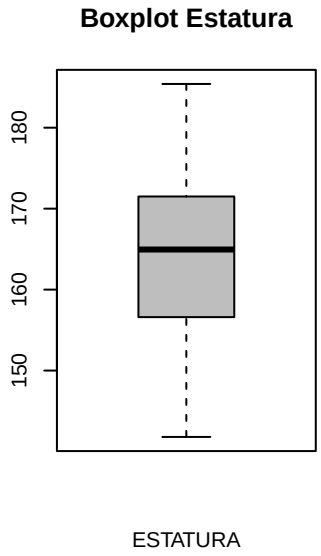


De los anterios histogramas se puede observar algún tipo de bi-modalidad, esto puede ser debido a que en la base de datos se encuentra la variable genero, la cual puede ser un factor discriminante en las variables.

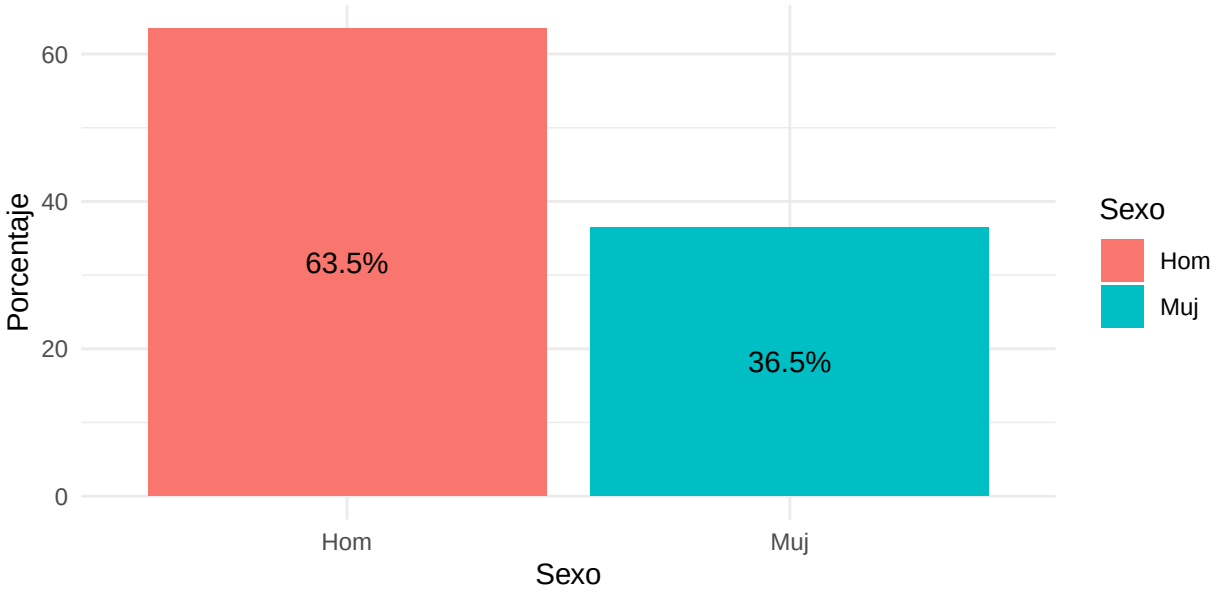
1.4 Boxplots



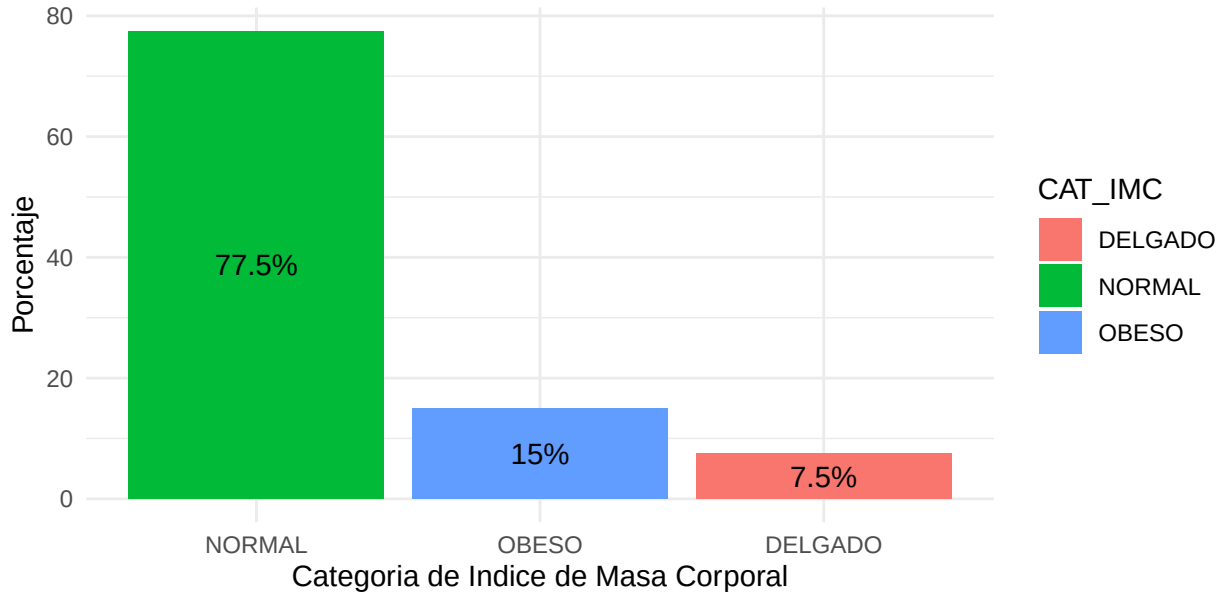
De los anteriores box-plots se pueden observar datos atípicos en las variables: **MASA**, **PERIMETRO MUSLO**, **PERIMETRO ABDOMINAL CINTURA** y **ANCHURA DE LAS CADERAS**.



1.5 Diagrama de Barras



Del anterior grafico de barras se puede observar que el **63.5%** de las observaciones pertenecen al genero **Hombre** y el **36.5%** restante pertenece al genero **Mujer**.



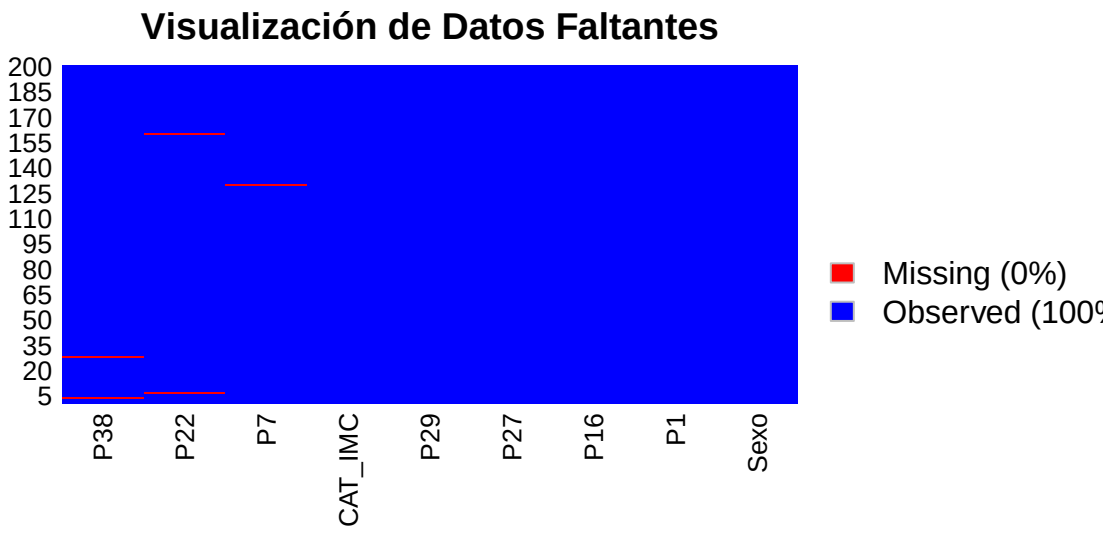
Del anterior grafico de barras se puede observar que el **77.5%** de las observaciones pertenecen a la categoría de indice de masa corporal **Normal**, el **15%** pertenece a la categoria **Obeso** y el **7.5%** restante pertenecen a la categoría **Delgado**.

2 Segundo Punto

Realice el respectivo proceso de imputación para los datos faltantes en su base de datos. Explique cómo realiza dicha imputación, cuál criterio utiliza y muestre un par de ejemplos ilustrativos.

2.1 Grafico de datos faltantes

Procederemos realizando un gráfico de datos faltantes, el cual nos permitirá determinar el porcentaje de datos ausentes en cada variable.



Se puede observar que el porcentaje de datos faltantes es aproximadamente 0. Esto nos indica que la base de datos no presenta muchos problemas con valores faltantes (missing o NA). Sin embargo, las variables P38, P22 y P7 contienen datos faltantes, los cuales son:

- **P38:** observación 173 y 197.
- **P22:** observación 41 y 194.
- **P7:** observación 71.

2.2 Imputación de datos faltantes

Como se observó en el análisis descriptivo previo, el género parece ser un factor discriminante en las variables de estatura (P38), anchura de las caderas (P22) y el perímetro del muslo mayor (P7), indicando medidas promedio mayores o menores dependiendo del genero. Por lo tanto, el criterio de imputación de datos se basara en el promedio de la variable respecto al genero de la observación que cuenta con un dato faltante en alguna de las anterios variables.

Dos ejemplos de como se realizo la imputación de los datos para la variable P38

- Para la observación **173** que tiene un valor faltante en la variable *P38* (Estatura) se procede a calcular el promedio para hombres y mujeres los cuales son: 169.3024 cm y 155.1319 cm respectivamente. Como la observación **173** es una **mujer** el valor a imputar es **155.1319**
- Para la observación **197** que tiene un valor faltante en la variable *P38* (Estatura) se procede a calcular el promedio para hombres y mujeres los cuales son: 169.3024 cm y 155.1319 cm respectivamente. Como la observación **197** es un **hombre** el valor a imputar es **169.3024**

De esta manera se imputan los datos faltantes en las demás variables.

Se anexa el código usado en R para hacer la imputación de datos.

```
data <- data %>%
  group_by(Sexo) %>%
  mutate(P38 = ifelse(is.na(P38), mean(P38, na.rm = TRUE), P38),
         P22 = ifelse(is.na(P22), mean(P22, na.rm = TRUE), P22),
         P7 = ifelse(is.na(P7), mean(P7, na.rm = TRUE), P7))
```

3 Tercer Punto

Considere las variables P1, P29 y P38. ¿Se puede afirmar que cada variable por separado permitiría discriminar entre Hombres y Mujeres? Elabore los resúmenes numéricos y gráficos que considere pertinentes para responder la pregunta.

3.1 Resumen Numerico

Cuadro 2: Resumen Numerico para P1

Sexo	Promedio	DesviacionEstandar	Mediana
Hom	70.98898	11.557182	71.3
Muj	57.63836	7.671857	56.7

Se puede observar que el promedio de la Masa corporal **(P1)** es mayor para los hombres respecto que el de las mujeres, tambien se observa que los hombres son un poco mas disperso que las mujeres. Que el promedio en la masa corporal de los hombres sea mayor nos da un indicio descriptivo de que dicha variable puede discriminar hombres y mujeres.

Cuadro 3: Resumen Numerico para P29

Sexo	Promedio	DesviacionEstandar	Mediana
Hom	18.30787	0.9103990	18.3
Muj	16.46164	0.8058174	16.4

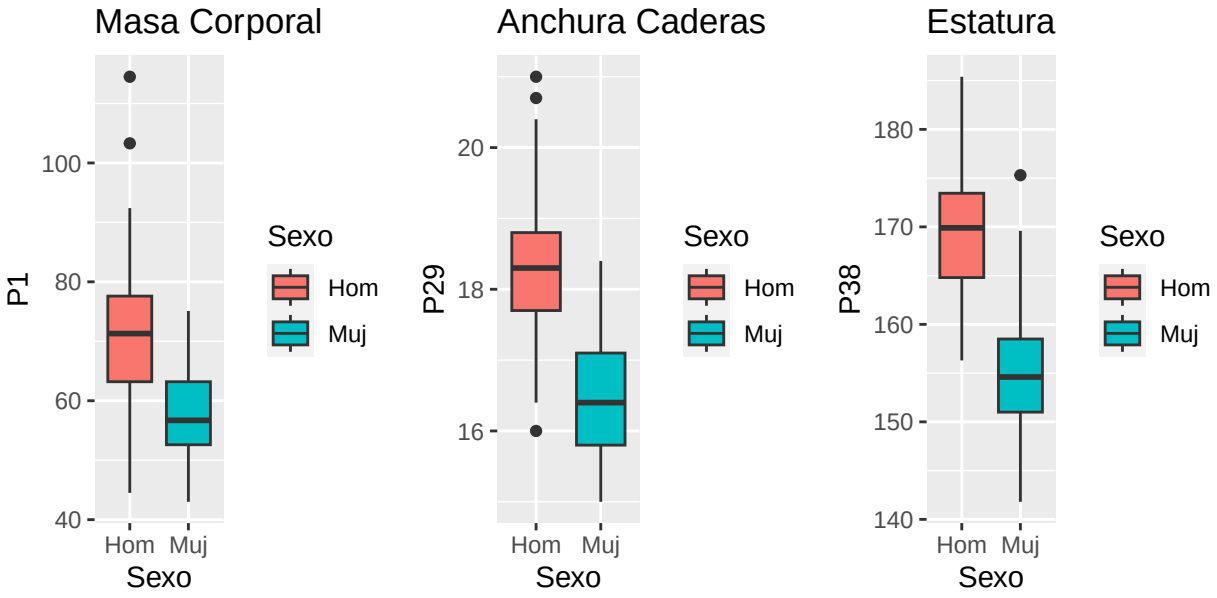
Se puede observar que la longitud promedio de las manos **(P29)** es un poco mayor para los hombres respecto que el de las mujeres, ambos géneros tiene un comportamiento muestral muy similar.

Cuadro 4: Resumen Numerico para P38

Sexo	Promedio	DesviacionEstandar	Mediana
Hom	169.3024	6.098132	169.9
Muj	155.1319	6.136770	154.6

Se puede observar que la estatura promedio **(P38)** es mayor para los hombres respecto que el de las mujeres. Que el promedio en las estaturas de los hombres sea mayor nos da un indicio descriptivo de que dicha variable puede discriminar hombres y mujeres.

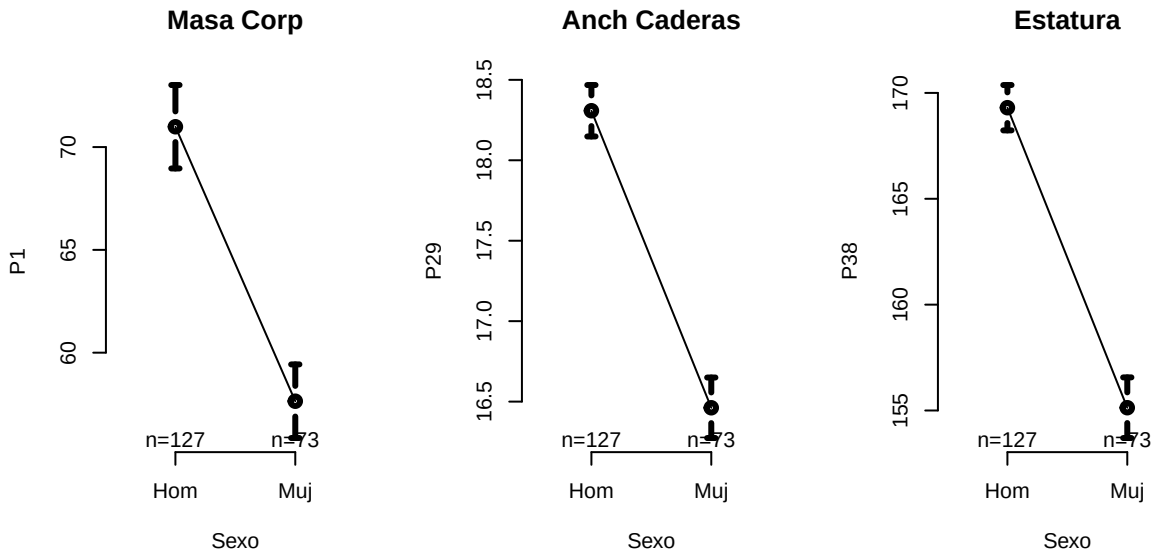
Ahora, vamos a crear gráficos de caja (boxplot) que nos ayudarán a visualizar de manera más clara los comentarios descriptivos previos.



En los boxplots anteriores, se observa que las cajas de los géneros no se traslapan. Esto indica que existen diferencias muestrales entre los géneros en lo que respecta a la masa corporal, la anchura de las caderas y la estatura.

Procederemos a realizar un grafico de diferencias de medias que nos permitira si existe diferencia estadistica entre el sexo y las anterios varirables de interes.

3.2 Grafico de diferencias de medias



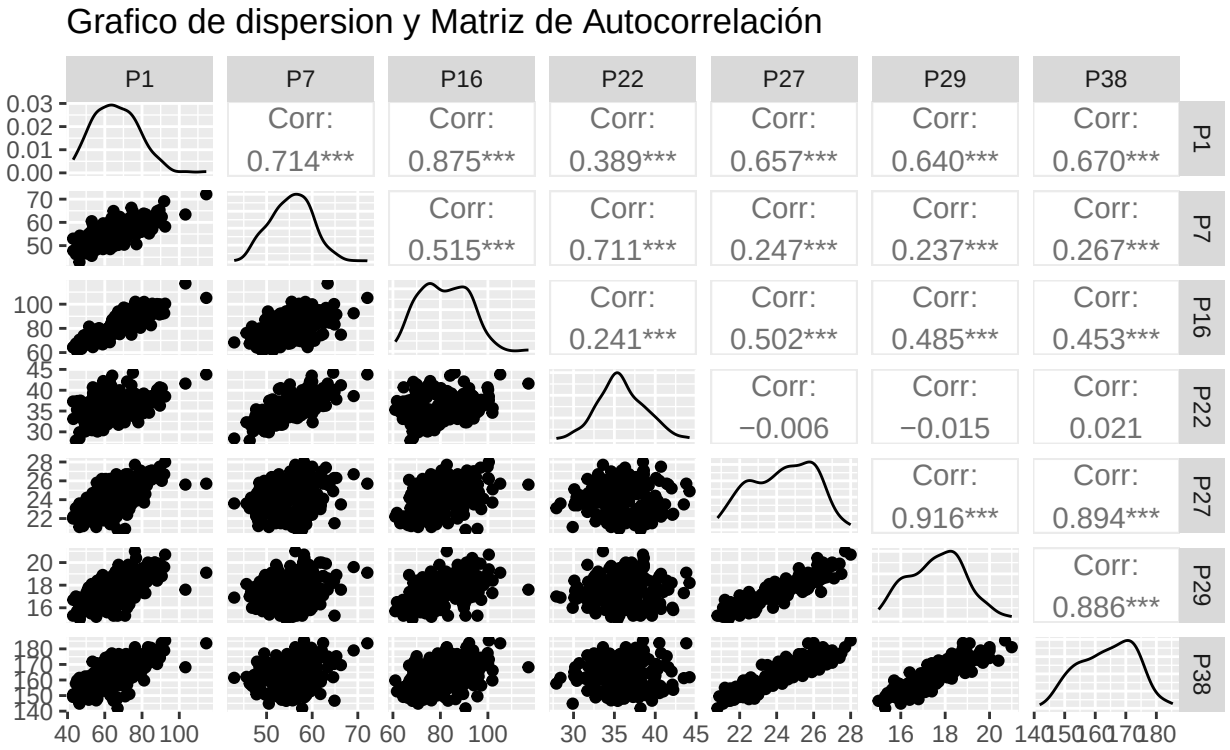
Este gráfico de diferencias de medias nos permite concluir y afirmar que cada variable por separado permitiría discriminar entre Hombres y Mujeres, dado que hay diferencias estadísticas significativas. Indicando que los hombres tienen en promedio una mayor masa corporal, una mayor anchura en las caderas y una mayor estatura.

4 Cuarto Punto

Usando las variables continuas, realice un gráfico de dispersión para identificar posibles relaciones entre sus variables. Explique si lo que se observa gráficamente tiene sentido o es coherente a la luz de sus datos. Corrobore lo observado con el cálculo de la matriz de correlaciones. Comente. Repita el proceso discriminando por SEXO. ¿Hay cambios en las estructuras de Covarianzas para ambos grupos? Comente

4.1 Relaciones entre variables

Procederemos a realizar un grafico de dispersión para tratar de identificar posibles relaciones entre las variables continuas.



Se invita al lector en fijarse en los diagramas de dispersión los cuales nos indican de manera descriptiva algún tipo de relación o tendencia lineal entre las variables. Se mencionaran las variables que tiene una forma de tendencia lineal mas notoria, las cuales son:

- P1 y P7
- P1 y P16
- P7 y P22
- P27 y P29
- P27 y P38
- P29 y P38

Luego procederemos a mirar la matriz de correlación en el mismo gráfico anterior buscando y mencionando en este trabajo escrito, las relaciones lineal fuerte con un coeficiente de autocorrelacion por encima del 80%, las cuales son:

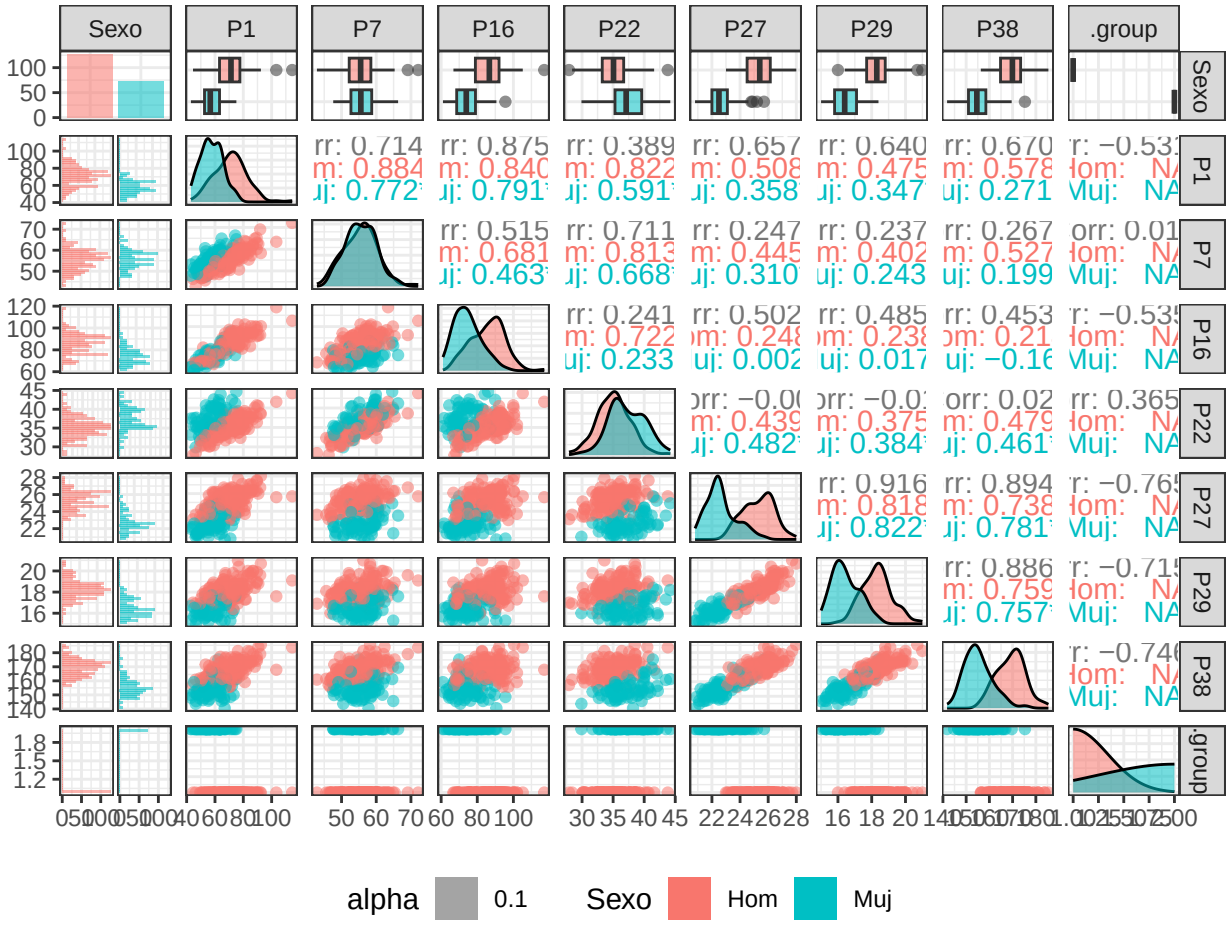
-P27 y P29: con un coeficiente de autocorrelación de 91.6% -P27 y P38: con un coeficiente de autocorrelación de 89.4% -P38 y P29: con un coeficiente de autocorrelación de 88.6%

A la luz de los datos, las relaciones anteriormente mencionadas tiene demasiado sentido. Dado que estas relaciones como:

- La variable P27 longitud promedio de los pies y P29 longitud promedio de las manos.
- La variable P27 longitud promedio de los pies y P38 la estatura.
- La variable P38 estatura y P29 longitud promedio de las manos.

Guardan una estrecha relación antropometricas.

4.2 Gráfico de analisis general multivariado



Se puede evidenciar, como era de esperarse, que existen cambios en la estructura de covarianza cuando discriminamos las variables por género. Dado que esta segmentacíon ayuda a explicar mejor las variables como la estatura, la longitud promedio de las manos y la longitud promedio de los pies.

5 Quinto Punto

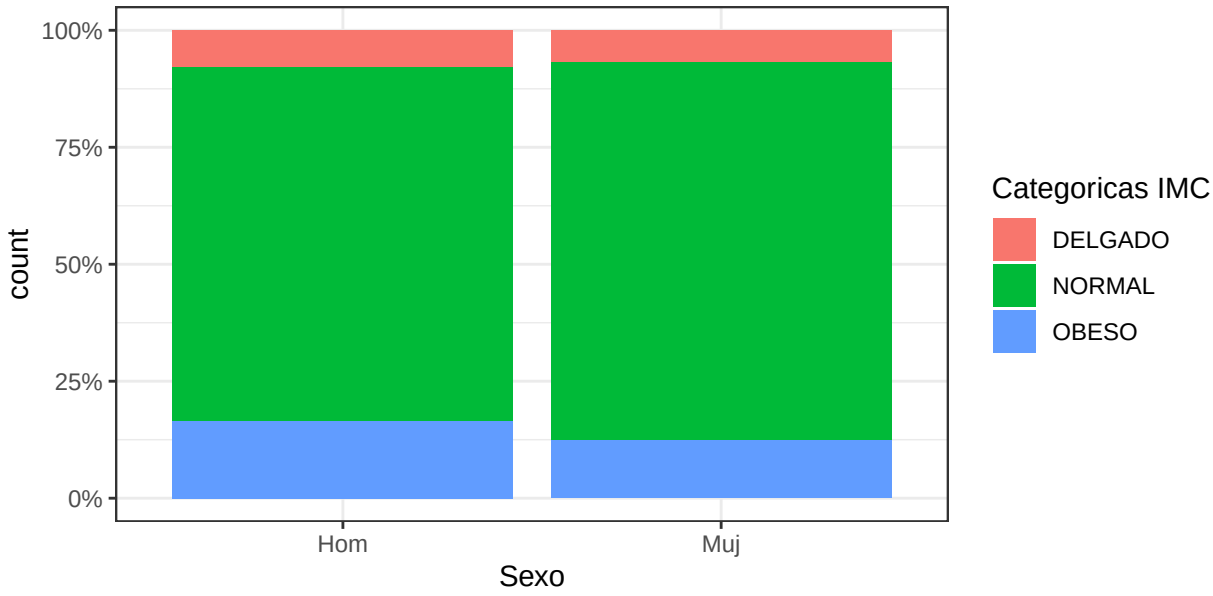
Elabore una tabla de porcentajes de doble entrada con las variables CAT_IMC y SEXO. Luego presente la información gráficamente. ¿Se puede afirmar que la distribución porcentual de la variable CAT_IMC es diferente para hombre y mujeres? Justifique su respuesta.

5.1 Distribución porcentual y tabla de contingencia.

Cuadro 5: Tabla de contingencia porcentual de doble entrada.

	Hom	Muj	Sum
DELGADO	5.0	2.5	7.5
NORMAL	48.0	29.5	77.5
OBESO	10.5	4.5	15.0
Sum	63.5	36.5	100.0

En la anterior tabla vemos el comportamiento porcentual de los individuos en base a las variables categóricas del problema. Se espera que del 15% de personas con obesidad, el 70% son hombres y el 30% son mujeres, del 7.5% de pesonas delgadas, el 66.67% son hombres y el 33.33% son mujeres,



Segun el anterior grafico la distribución porcentual de la variable CAT_IMC no cambia para hombre y mujeres.

6 Sexto Punto

Se tienen los siguientes datos de 5 personas, de las cuales se desconoce su CAT_IMC.

P1	P7	P16	P22	P27	P29	P38	CAT_IMC
66.1	53.9	73.8	34.7	27.6	20.9	181.6	
55.8	50.1	76.9	39.5	24.7	17.3	154.5	
62.8	54.3	80.4	37.5	23.5	16.5	156.6	
63.9	50.6	75.6	31.5	24.9	18.6	173.1	
50.7	46.3	72.7	30.4	23.5	16.7	159.5	

Usando la distancia estadística, determine a cuál de las tres categorías pertenece cada sujeto. Explique claramente el proceso empleado para clasificar los sujetos. Anexe el código empleado.

6.1 Clasificación usando la distancia estadística

Se anexa el codigo usado para realizar la clasificación de la categorias.

```
library(stats)

datos <- data[, 2:8]
k <- 3


predecir_categoria <- function(nueva_observacion) {

  distancias <- mahalanobis(datos, nueva_observacion, cov_matrix)

  df_distancias <- data.frame(Distancia = distancias, Categoria = data$CAT_IMC)

  df_distancias <- df_distancias[order(df_distancias$Distancia), ]

  vecinos_cercanos <- df_distancias[1:k, ]

  categoria_predicha <- names(sort(table(vecinos_cercanos$Categoria),
                                   decreasing = TRUE))[1]

  return(categoria_predicha)
}


n1 <- c(66.1, 53.9, 73.8, 34.7, 27.6, 20.9, 181.6)
n2 <- c(55.8, 50.1, 76.9, 39.5, 24.7, 17.3, 154.5)
n3 <- c(62.8, 54.3, 80.4, 37.5, 23.5, 16.5, 156.6)
n4 <- c(63.9, 50.6, 75.6, 31.5, 24.9, 18.6, 173.1)
n5 <- c(50.7, 46.3, 72.7, 30.4, 23.5, 16.7, 159.5)
```

```
categoria_predicha1 <- predecir_categoria(n1)
categoria_predicha1

categoria_predicha2 <- predecir_categoria(n2)
categoria_predicha2

categoria_predicha3 <- predecir_categoria(n3)
categoria_predicha3

categoria_predicha4 <- predecir_categoria(n4)
categoria_predicha4

categoria_predicha5 <- predecir_categoria(n5)
categoria_predicha5
```

Las nuevas observaciones se clasifican en:

- Normal
- Normal
- Normal
- Normal
- Delgado