Práctica 2: ¿Cómo realizar la limpieza y análisis de datos?

Jordi Martínez Joyeux y Héctor Martínez Hidalgo

Junio 2023

Contents

Descripción del dataset

¿Por qué es importante y qué pregunta/problema pretende responder?

Este conjunto de datos es importante porque permite explorar la relación entre los niveles de colesterol y diversos factores de riesgo médicos en personas con una mayor probabilidad de sufrir un infarto. El colesterol alto es una preocupación médica que se asocia con un mayor riesgo de sufrir enfermedades cardiovasculares, incluyendo los infartos. Por lo tanto, entender cómo se relaciona el colesterol con otros factores de riesgo puede proporcionar información valiosa para gestionar y prevenir las enfermedades cardiacas. El problema que buscamos abordar con este conjunto de datos es el siguiente:

¿Cómo se relacionan los niveles de colesterol con factores como la edad, el tipo de dolor en el pecho, el nivel de azúcar en sangre y el sexo en personas que tienen una mayor probabilidad de sufrir un infarto? Al responder a esta pregunta, esperamos poder identificar patrones o correlaciones que podrían ser útiles para predecir y manejar los niveles de colesterol, y por ende, reducir el riesgo de infarto en estos individuos.

El dataset original puede descargarse desde Kaggle en el siguiente enlace:

https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

El código y este mismo dataset puede descargarse desde el repositorio Github:

https://github.com/hectormh88/TCVD-PRA2

Variables del dataset

El dataset contiene las siguientes variables:

- age: Edad de la persona (en años)
- sex: Sexo de la persona (0 = mujer; 1 = hombre)
- cp: Tipo de dolor de pecho (0 = angina típica; 1 = angina atípica; 2 = dolor no anginoso; 3 = asintomático)
- trtbps: Presión arterial en reposo (en mm Hg)
- chol: Nivel de colesterol en sangre (en mg/dl)
- fbs: Azúcar en sangre en ayunas > 120 mg/dl (1 = verdadero; 0 = falso)
- restecg: Resultados electrocardiográficos en reposo (0 = normal ; 1 = presenta anormalidad en la onda ST-T (inversiones de la onda T y/o elevación o depresión del segmento ST de > 0.05 mV); 2 = muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes)
- thalachh: Frecuencia cardíaca máxima alcanzada (en latidos por minuto)
- exng: Angina inducida por el ejercicio (0 = No se experimentó angina inducida por el ejercicio, 1 = Se experimentó angina inducida por el ejercicio)
- oldpeak: Depresión del ST inducida por el ejercicio en relación con el reposo (en múltiplos de la desviación estándar)

- slp: Pendiente del segmento ST durante el ejercicio (0 = Pendiente descendente; 1 = Pendiente plana; 2 = Pendiente ascendente)
- caa: Número de vasos sanguíneos principales obstruidos
- thall: Resultados de la prueba de estrés con talio (1 = Normal; 2 = Defecto fijo, 3 = Defecto reversible)
- output: Variable target que define la posibilidad de tener un infarto (0 = menos posibilidades de tener un infarto; 1 = más posibilidades de tener un infarto)

Lectura de datos

```
data <- read.csv("../data/heart.csv")
```

Estructura de datos

Podemos ver que el dataset contiene las variables descritas en el apartado anterior. Está compuesto por 14 variables y 303 observaciones, sin embargo, algunas variables no tienen asignada la clase correcta, lo cual corregiremos luego de seleccionar las variables de interés.

```
str(data)
```

```
## 'data.frame':    303 obs. of  14 variables:
##  $ age     : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ sex     : int  1 1 0 1 0 1 0 1 1 1 ...
##  $ cp      : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ trtbps  : int  145 130 130 120 120 140 140 120 172 150 ...
##  $ chol    : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ fbs     : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ restecg : int  0 1 0 1 1 1 0 1 1 1 ...
##  $ thalachh: int  150 187 172 178 163 148 153 173 162 174 ...
##  $ exng    : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ oldpeak : num  2.3 3.5 1.4 0.8 0.6 0.4 1.3 0 0.5 1.6 ...
##  $ slp     : int  0 0 2 2 2 1 1 2 2 2 ...
##  $ caa     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ thall   : int  1 2 2 2 2 1 2 3 3 2 ...
##  $ output  : int  1 1 1 1 1 1 1 1 1 1 ...
```

Integración y selección de los datos de interés a analizar

Para el objetivo de nuestro estudio, que es analizar la relación entre los niveles de colesterol y factores como la edad, el tipo de dolor en el pecho, el nivel de azúcar en sangre y el sexo en personas con una mayor probabilidad de tener un infarto, hemos seleccionado las siguientes variables: age, chol, cp, fbs, output y sex.

La variable 'age' es importante porque queremos examinar si hay una correlación entre la edad de la persona y su nivel de colesterol, que es medido por la variable 'chol'. Este análisis puede arrojar luz sobre si la edad es un factor que afecta los niveles de colesterol en aquellos con un mayor riesgo de tener un infarto.

Las variables 'sex', 'fbs' y 'cp' nos permiten realizar comparaciones entre diferentes grupos. En particular, vamos a comparar los niveles de colesterol entre hombres y mujeres, entre personas con y sin un nivel de

azúcar en sangre en ayunas superior a 120 mg/dl, y entre personas con diferentes tipos de dolores en el pecho. Estos análisis pueden ayudarnos a entender si estos factores están asociados con diferencias en los niveles de colesterol en personas con una mayor probabilidad de tener un infarto.

Finalmente, la variable 'output' es esencial para nuestro análisis, ya que nos permite seleccionar el subconjunto de personas en el dataset que tienen un mayor riesgo de tener un infarto.

Por otro lado, hemos decidido no incluir las otras variables del dataset en nuestro análisis porque no son directamente relevantes para nuestra pregunta de investigación. Mientras que variables como 'trtbps', 'restecg', 'thalachh', 'exng', 'oldpeak', 'slp', 'caa' y 'thall' pueden ser importantes en el estudio de la enfermedad cardíaca, no son necesarias para analizar la relación entre los niveles de colesterol y los factores que hemos identificado. Mantener nuestro análisis centrado en nuestras variables de interés puede ayudar a evitar la confusión y facilitar la interpretación de nuestros resultados.

Selección de variables de interés

```
data <- subset(data, select = c(age, chol, cp, fbs, output, sex))
```

Transformación de variables

Dado que ya tenemos definido el dataset con el que trabajaremos, ya podremos corregir la clase de las variables. En el actual dataset todas las variables son de tipo entero, pero por las características de 'cp', 'fbs', 'output' y 'sex', deberían ser variables categóricas de tipo factor.

```
str(data)
```

```
## 'data.frame':    303 obs. of  6 variables:
##  $ age   : int  63 37 41 56 57 57 56 44 52 57 ...
##  $ chol  : int  233 250 204 236 354 192 294 263 199 168 ...
##  $ cp    : int  3 2 1 1 0 0 1 1 2 2 ...
##  $ fbs   : int  1 0 0 0 0 0 0 0 1 0 ...
##  $ output: int  1 1 1 1 1 1 1 1 1 1 ...
##  $ sex   : int  1 1 0 1 0 1 0 1 1 1 ...
```

Corregimos la clase de las variables indicadas:

```
var <- c("age", "chol", "cp", "fbs", "output", "sex")
var_num <- c("age", "chol")
var_factor <- c("cp", "fbs", "output", "sex")

# Cambiar a tipo factor las variables de tipo carácter
data[var_factor] <- lapply(data[var_factor], factor)
```

Y a continuación podemos ver las características básicas del dataset con el que trabajaremos:

```
summary(data)
```

```
##       age             chol       cp      fbs     output  sex
##  Min.   :29.00   Min.   :126.0   0:143   0:258   0:138   0: 96
##  1st Qu.:47.50   1st Qu.:211.0   1: 50   1: 45   1:165   1:207
##  Median :55.00   Median :240.0   2: 87
##  Mean   :54.37   Mean   :246.3   3: 23
##  3rd Qu.:61.00   3rd Qu.:274.5
##  Max.   :77.00   Max.   :564.0
```

Limpieza de datos

¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos

Analizando variable a variable, podemos ver que el dataset no contiene valores perdidos categorizados como valores nulos:

```
# Conteo valores nulos
null_age <- sum(is.na(data$age))/length(data)*100
null_chol <- sum(is.na(data$chol))/length(data)*100
null_cp <- sum(is.na(data$cp))/length(data)*100
null_fbs <- sum(is.na(data$fbs))/length(data)*100
null_output <- sum(is.na(data$output))/length(data)*100
null_sex <- sum(is.na(data$sex))/length(data)*100

# Dataframe valores nulos
null_table <- data.frame(Variable = var,
                         Percentage = c(null_age, null_chol, null_cp, null_fbs, null_output, null_sex))

null_table
```

```
##   Variable Percentage
## 1      age          0
## 2     chol          0
## 3       cp          0
## 4      fbs          0
## 5   output          0
## 6      sex          0
```

Pero también puede darse el caso de valores perdidos que hayan sido expresados de otra forma, como ““,” “, “NA”, 0 o 99. Observemos nuevamente el resumen estadístico:

```
summary(data)
```

```
##       age             chol       cp      fbs     output  sex
##  Min.   :29.00   Min.   :126.0   0:143   0:258   0:138   0: 96
##  1st Qu.:47.50   1st Qu.:211.0   1: 50   1: 45   1:165   1:207
##  Median :55.00   Median :240.0   2: 87
##  Mean   :54.37   Mean   :246.3   3: 23
##  3rd Qu.:61.00   3rd Qu.:274.5
##  Max.   :77.00   Max.   :564.0
```

En el caso de las variables numéricas ‘age’ y ‘chol’ podemos afirmar que no hay registros nulos indicados mediante alguna cadena de caracteres como ““,” ” o “NA”, ni ninguna otra, ya que la estructura de datos ya la reconoce como números enteros (si hubiese una cadena de caracteres, la variable sería de tipo cadena de caracteres). Viendo los valores máximos y mínimos del resumen estadístico también se observa que no contienen valores con número 0 o número 99, que usualmente representan valores perdidos.

En cuanto a las variables categóricas ‘cp’, ‘fbs’, ‘output’ y ‘sex’, observando el resumen estadístico, podemos ver directamente que no contiene ningún valor que no corresponda a los niveles ya identificados en la descripción de las variables del set de datos.

Por lo tanto, en cuanto a valores nulos no será necesario realizar ninguna gestión.

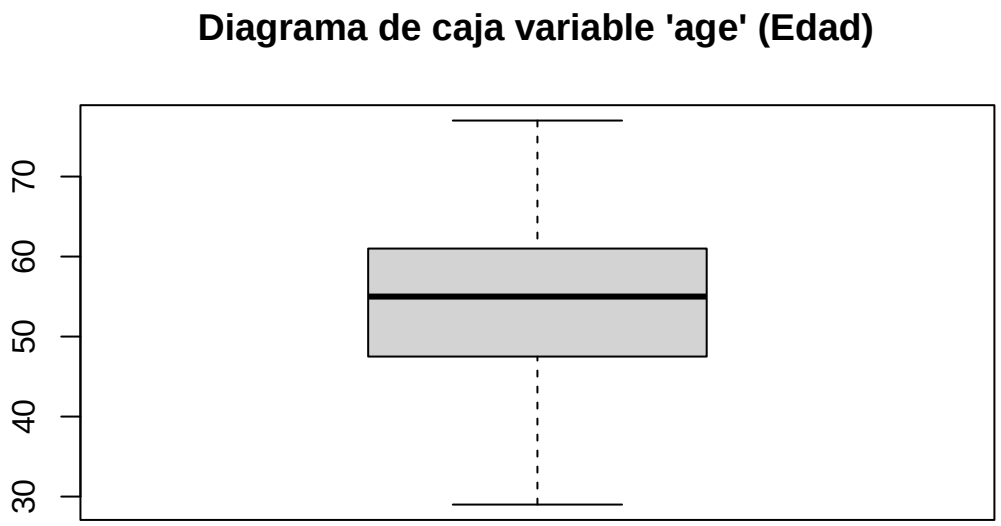
Identifica y gestiona los valores extremos

Podemos identificar los valores extremos o outliers mediante un diagrama de caja.

Edad

Podemos ver que en el caso de la edad no se ve presencia de valores extremos, los cuales figurarían bajo el bigote inferior y sobre el bigote superior.

```
# Diagrama de caja Edad
boxplot_age <- boxplot(data$age, main="Diagrama de caja variable 'age' (Edad)", xlab="Edad")
```




```
# Límite superior
upper_whisker_age <- boxplot_age$stats[5]
# Límite inferir
lower_whisker_age <- boxplot_age$stats[1]
# Número de outliers
n_outliers_age <- sum(data$age > upper_whisker_age | data$age < lower_whisker_age); n_outliers_age
```

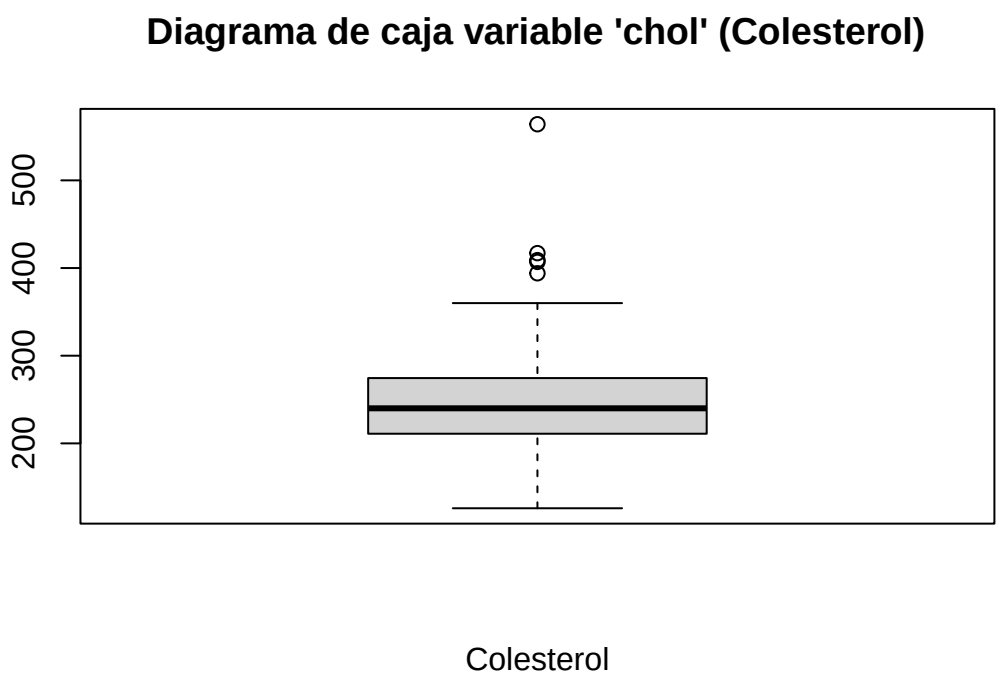
```
## [1] 0
```

Podemos corroborar que el número de valores extremos en la edad es 0, por lo que no es necesario realizar ninguna gestión al respecto.

Colesterol

Para el caso del colesterol, se observan algunos valores extremos sobre el límite superior.

```r
# Diagrama de caja Colesterol
boxplot_chol <- boxplot(data$chol, main="Diagrama de caja variable 'chol' (Colesterol)", xlab="Colester
```




```r
# Límite superior
upper_whisker_chol <- boxplot_chol$stats[5]
# Lñimite inferior
lower_whisker_chol <- boxplot_chol$stats[1]
# Número de outliers
n_outliers_chol <- sum(data$chol > upper_whisker_chol | data$chol < lower_whisker_chol); n_outliers_chol
```

```
## [1] 5
```

Se trata de 5 valores extremos, de los cuales se observa en el diagrama que uno de ellos es notoriamente más grande que el resto. Ya se ha detectado en la exploración inicial de los datos y corresponde a un valor de 564 mg/dL. Si bien es técnicamente posible que una persona tenga un nivel de colesterol tan alto, es muy raro y generalmente se asocia con condiciones médicas específicas como la hipercolesterolemia familiar.

Nuestro análisis tiene como objetivo explorar las relaciones entre el colesterol y otros factores en la población general y, en particular, en aquellos con mayor riesgo de sufrir un infarto. Por lo tanto, un valor tan extremo podría sesgar nuestros resultados y hacer que nuestras conclusiones sean menos aplicables a la población general (MaryAnn Fletcher, 1 de julio 2020, "When It Comes to Cholesterol, What's 'High' Truly Mean?" https://www.livestrong.com/article/230138-what-is-the-highest-number-that-cholesterol-can-go/).

Como tal, hemos tomado la decisión de eliminar este valor extremo de nuestro conjunto de datos. Esta decisión no se tomó a la ligera y solo después de considerar cuidadosamente el impacto potencial que este valor podría tener en nuestro análisis. Aunque eliminamos este valor, es importante tener en cuenta que las personas con hipercolesterolemia familiar existen y pueden tener niveles de colesterol extremadamente altos. Sin embargo, su inclusión en este análisis particular podría desviar la atención de las tendencias y relaciones que son más pertinentes para la población general.

Con la eliminación de este valor extremo, esperamos que nuestro análisis proporcionará una visión más precisa de las relaciones entre el colesterol y otros factores de riesgo en la población general. Este paso crítico en la limpieza de datos nos ayudará a garantizar que nuestras conclusiones sean válidas, confiables y aplicables a un contexto más amplio.

```
data <- data[data$chol != 564, ]
```

Análisis de los datos

Selección de los grupos de datos que se quieren analizar/comparar

Para nuestro estudio nos vamos a centrar en los datos de las personas con un riesgo alto de tener un infarto.

```
data <- data[data$output == 1,]
```

Renombramos los niveles de las variables categóricas para facilitar el análisis y la visualización de datos.

```
levels(data$cp) <- c("Angina típica", "Angina atípica", "Dolor no anginoso", "Asintomático")
levels(data$sex) <- c("Mujer", "Hombre")
levels(data$fbs) <- c("Menor o igual a 120", "Mayor a 120")
```

Resumen estadistico

```
summary(data)
```

```
##       age             chol                        cp
##  Min.   :29.00   Min.   :126.0   Angina típica    :39
##  1st Qu.:44.00   1st Qu.:208.0   Angina atípica   :41
##  Median :52.00   Median :234.0   Dolor no anginoso:68
##  Mean   :52.41   Mean   :240.3   Asintomático     :16
##  3rd Qu.:59.00   3rd Qu.:266.2
##  Max.   :76.00   Max.   :417.0
##                   fbs      output      sex
##  Menor o igual a 120:141   0:  0   Mujer :71
##  Mayor a 120        : 23   1:164   Hombre:93
##
##
##
##
```

Podemos observar las estadisticas que presentan las personas con una probabilidad alta de sufrir un ataque cardiaco.

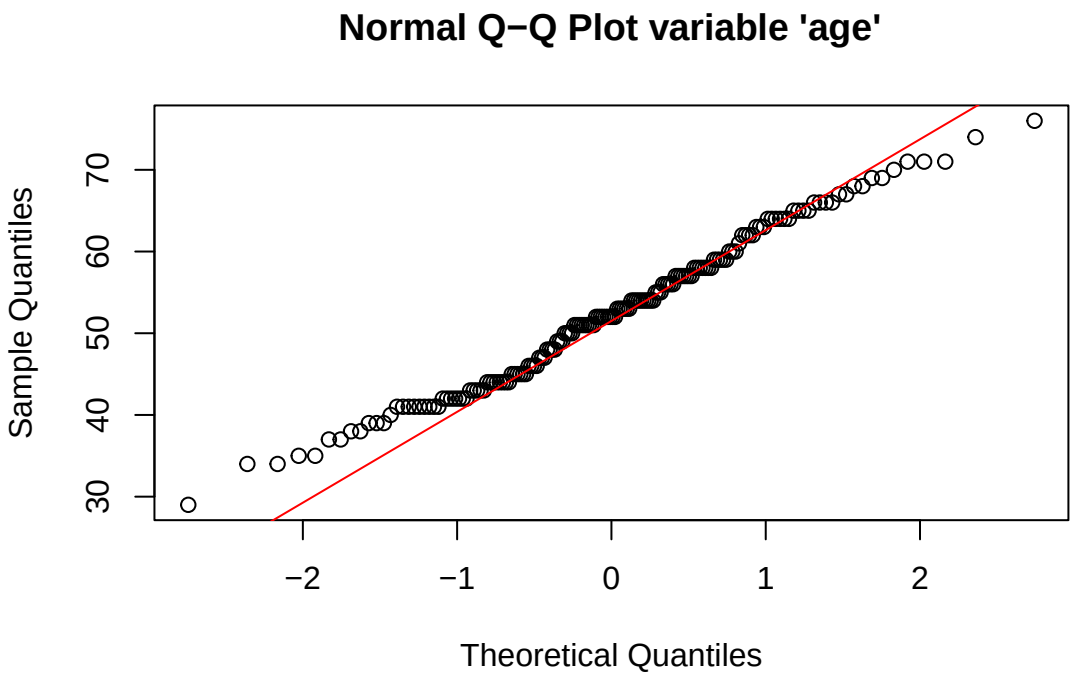
- Edad (age): La edad de estos pacientes varía entre 29 y 76 años, con una mediana de 52 años. Esto indica que la mitad de estos pacientes son menores de 52 años y la otra mitad son mayores de 52 años. Sin embargo, la edad promedio es de aproximadamente 52.41 años, lo que nos da una idea de la tendencia central de la distribución de la edad en este grupo de pacientes.
- Colesterol (chol): Los niveles de colesterol en estos pacientes varían entre 126 y 417 mg/dL, con una mediana de 234 mg/dL. Esto significa que la mitad de estos pacientes tienen niveles de colesterol menores a 234 mg/dL y la otra mitad tienen niveles mayores a 234 mg/dL. El nivel medio de colesterol es de 240.3 mg/dL. Cabe decir que el nivel de colesterol total considerado saludable para personas de 20 años o más, va de 125 a 200 mg/dL (https://medlineplus.gov/spanish/cholesterollevelswhatyouneedtoknow.html).
- Tipo de dolor en el pecho (cp): La variable de tipo de dolor en el pecho tiene 4 niveles (0, 1, 2, 3). En este grupo, la categoría más común es 2, seguida de 1 y 0. Solo un pequeño grupo de pacientes presenta el tipo de dolor en el pecho 3.
- Niveles de azúcar en sangre en ayunas (fbs): La mayoría de estos pacientes (141) tienen un nivel de azúcar en sangre en ayunas menor a 120 mg/dL (fbs = 0), mientras que solo una pequeña proporción (23) tiene un nivel de azúcar en sangre en ayunas mayor a 120 mg/dL (fbs = 1).
- Sexo (sex): Entre los pacientes con alta probabilidad de sufrir un infarto, hay más hombres (93) que mujeres (71).

En resumen, los pacientes con alta probabilidad de sufrir un infarto tienden a ser hombres alrededor de los 52 años de edad con niveles de colesterol alrededor de 240 mg/dL, y la mayoría tiene un nivel de azúcar en sangre en ayunas menor a 120 mg/dL.

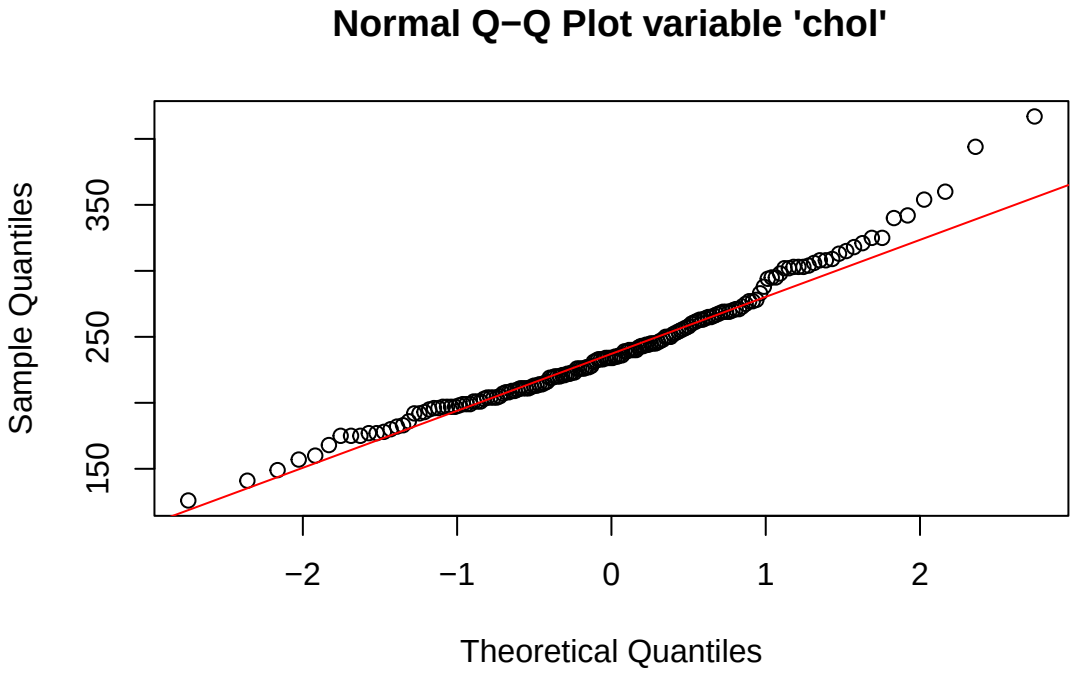
Comprobación de la normalidad

Primero, visualizaremos de manera gráfica similitud entre la distribución de las variables numéricas 'age' y 'chol' y la distribución normal, con la ayuda de Q-Q Plots.

```
qqnorm(data$age, main = "Normal Q-Q Plot variable 'age'")
qqline(data$age, col="red")
```




```
qqnorm(data$chol, main = "Normal Q-Q Plot variable 'chol'")
qqline(data$chol, col="red")
```



Se puede observar que, a priori, ambas variables tienen una distribución similar a una distribución normal,

sin embargo, no es del todo claro, por lo que para comprobar la normalidad en edad y colesterol vamos a usar la prueba de Shapiro-Wilk.

```
shapiro.test(data$age)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$age
## W = 0.98724, p-value = 0.1413
```

```
shapiro.test(data$chol)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  data$chol
## W = 0.96974, p-value = 0.001177
```

Para la edad, la prueba de Shapiro-Wilk nos dio un valor p de 0.1413. Como este valor es mayor que 0.05, no tenemos suficiente evidencia para rechazar la hipótesis nula de que los datos se distribuyen normalmente. Así que, parece que podemos asumir con seguridad que nuestros datos de **'age' están normalmente distribuidos**.

Para el colesterol, el valor p fue de 0.001177, que es menor que 0.05. Esto significa que tenemos que rechazar la hipótesis nula de que 'chol' se distribuye normalmente. Entonces, parece que nuestros datos de **'chol' no siguen una distribución normal**.

Comprobación de la homogeneidad de la varianza

Para comprobar la homocedasticidad u homogeneidad de varianza podemos utilizar leveneTest en el caso de variables con distribución normal, como el caso de la edad 'age', y fligner.test en el caso de variables sin distribución normal, como el nivel de colesterol 'chol'.

```
leveneTest(age ~ cp, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   3  0.0972 0.9615
##       160
```

```
leveneTest(age ~ fbs, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  4.1535 0.04317 *
##       162
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
leveneTest(age ~ sex, data = data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  2.5862 0.1097
##       162
```

En el test de Levene, sólo 'age' en función de 'fbs' obtuvo un p-valor menor al nivel de significancia, por lo tanto, en ese caso se rechaza la hipótesis nula de homocedasticidad, es decir, la edad tiene varianzas estadísticamente diferentes para las dos categorías de nivel de azucar en ayunas. Por el contrario, para 'age' en función de 'cp' y 'sex', no se rechaza la hipótesis nula de homocedasticidad, por lo que se concluye que la edad presenta iguales varianzas para los distintos tipos de dolor de pecho, por un lado, y para hombres y mujeres, por otro lado.

```
fligner.test(chol ~ cp, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chol by cp
## Fligner-Killeen:med chi-squared = 2.3543, df = 3, p-value = 0.5022
```

```
fligner.test(chol ~ fbs, data = data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chol by fbs
## Fligner-Killeen:med chi-squared = 0.0050521, df = 1, p-value = 0.9433
```

```
fligner.test(chol ~ sex, data = data)
```

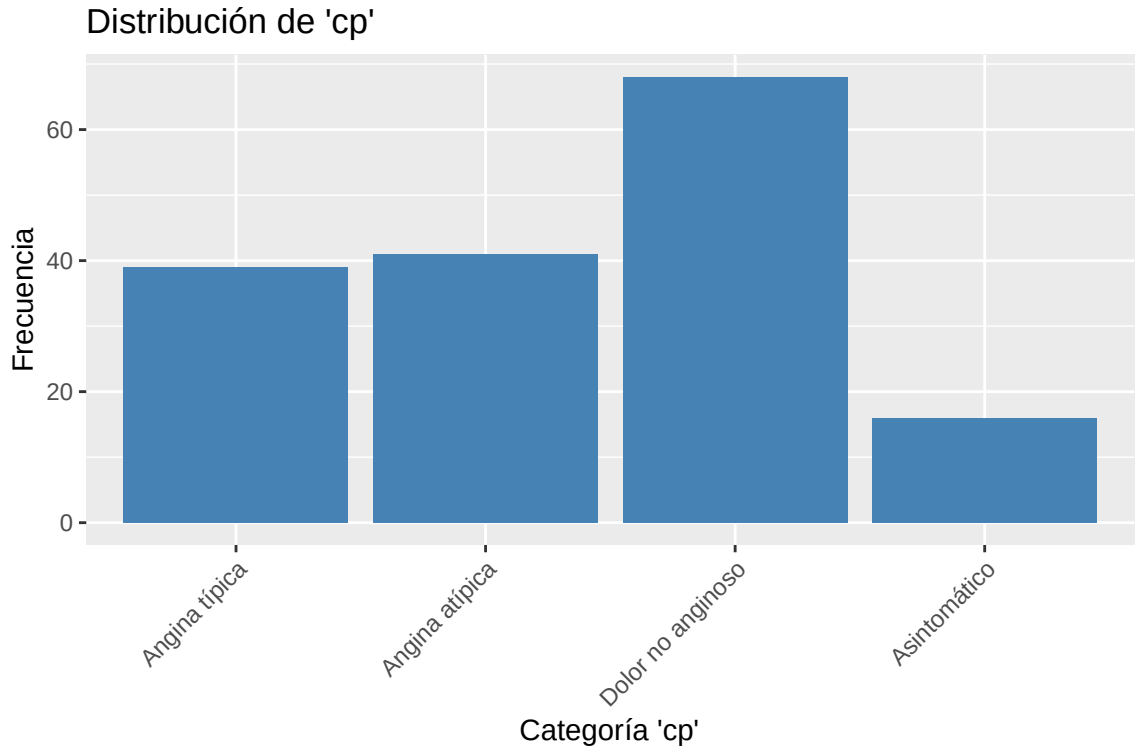
```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  chol by sex
## Fligner-Killeen:med chi-squared = 10.714, df = 1, p-value = 0.001063
```

Aplicando el test Fligner, sólo en el caso de 'chol' en función de 'sex' se rechazaría la hipótesis nula de homocedasticidad, por lo tanto, el nivel de colesterol tiene varianzas estadísticamente diferentes en hombres y mujeres, mientras que se puede asumir igualdad de varianzas para los diferentes grupos de tipo de dolor de pecho y de nivel de azucar en ayunas.

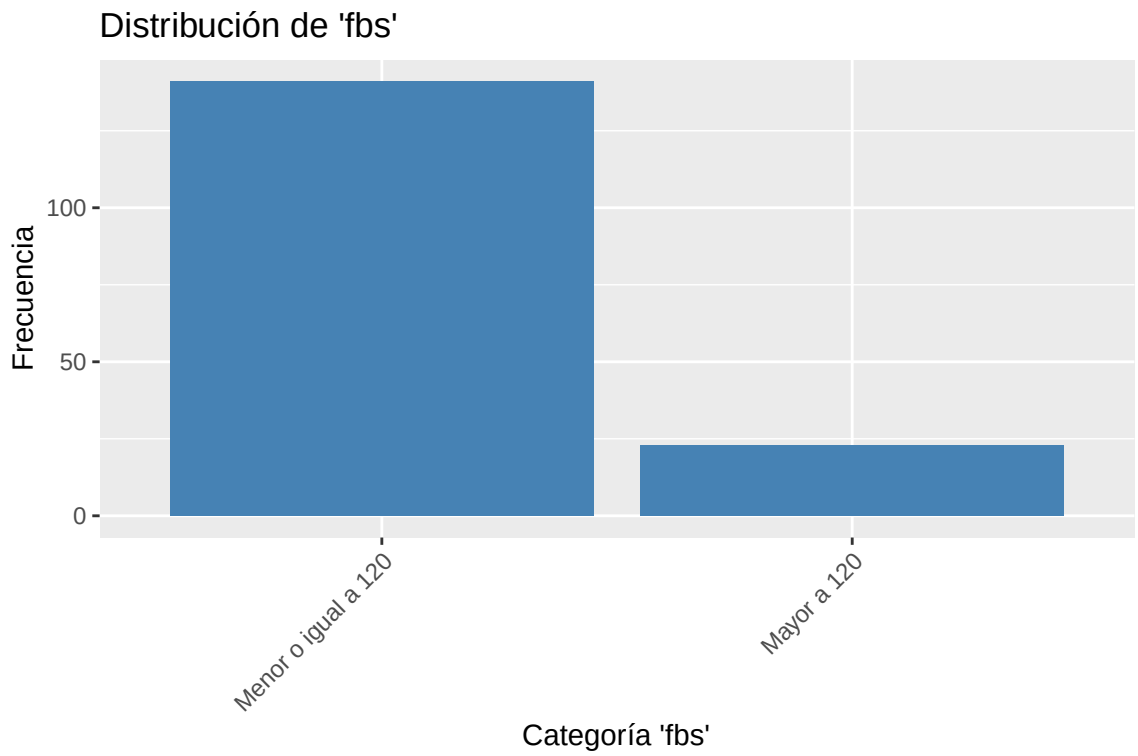
Análisis de la distribución de las variables categóricas

Vamos a explorar la distribución de las categorías utilizando tablas de frecuencia y gráficos de barras para entender cómo están distribuidas.

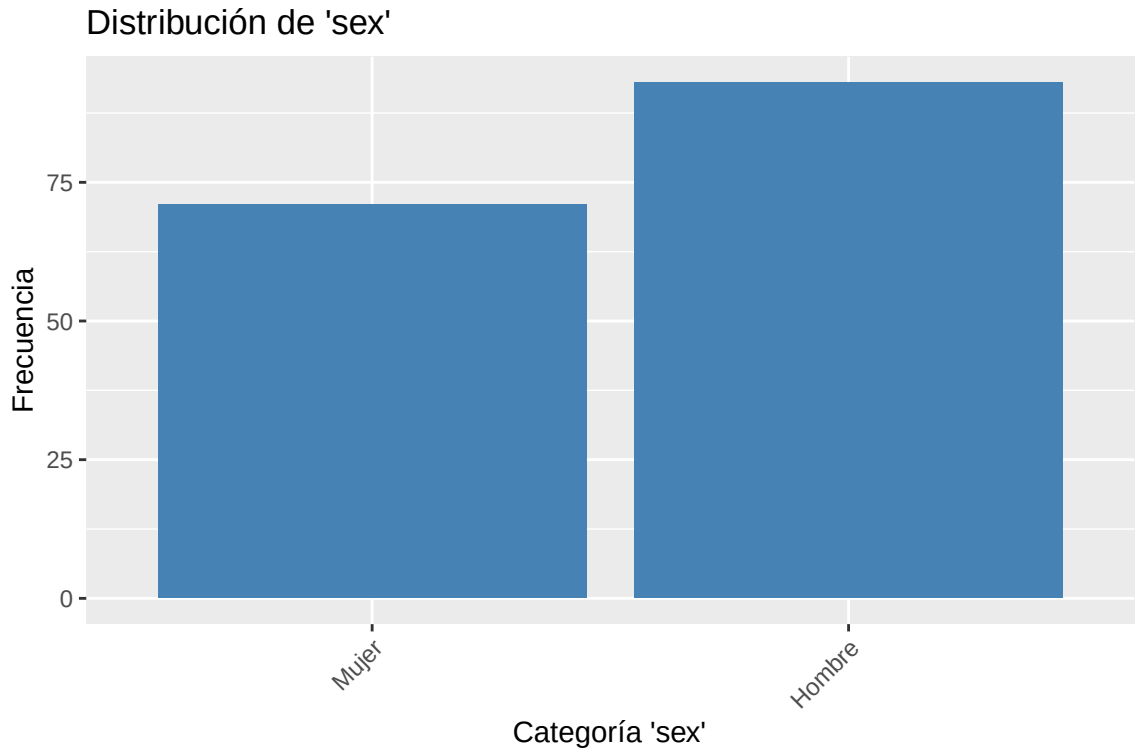
```r
# Tablas de frecuencias
cp_counts <- as.data.frame(table(data$cp))
fbs_counts <- as.data.frame(table(data$fbs))
sex_counts <- as.data.frame(table(data$sex))

# Gráficos de barras
ggplot(cp_counts, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x="Categoría 'cp'", y="Frecuencia", title="Distribución de 'cp'")
```




```r
ggplot(fbs_counts, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x="Categoría 'fbs'", y="Frecuencia", title="Distribución de 'fbs'")
```




```
ggplot(sex_counts, aes(x=Var1, y=Freq)) +
  geom_bar(stat="identity", fill="steelblue") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  labs(x="Categoría 'sex'", y="Frecuencia", title="Distribución de 'sex'")
```



En la variable 'cp', vemos que la categoría 2 es la más frecuente con 68 observaciones, seguida de cerca por la categoría 1 con 41 y la categoría 0 con 39. La categoría 3 es la menos común con solo 16 observaciones. Esto podría indicar que la variable 'cp' está un poco sesgada hacia la categoría 2, pero no hay una distribución uniforme en las categorías.

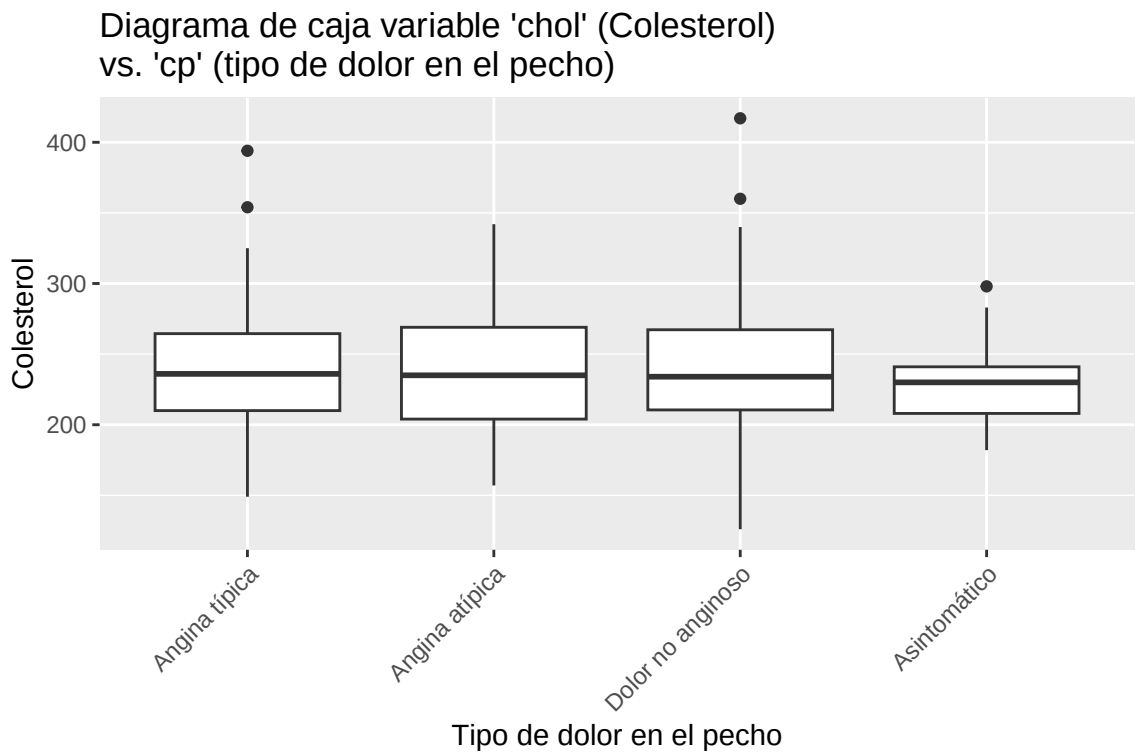
Analizamos 'fbs'. En este caso, la mayoría de las observaciones están en la categoría 0 con 141, y sólo 23 en la categoría 1. Esto sugiere que 'fbs' está muy sesgada hacia la categoría 0, lo que podría ser relevante cuando realicemos análisis posteriores.

Finalmente, observamos 'sex'. Aquí, la categoría 1 tiene 93 observaciones y la categoría 0 tiene 71. Aunque no es una distribución perfectamente uniforme, está bastante equilibrada.

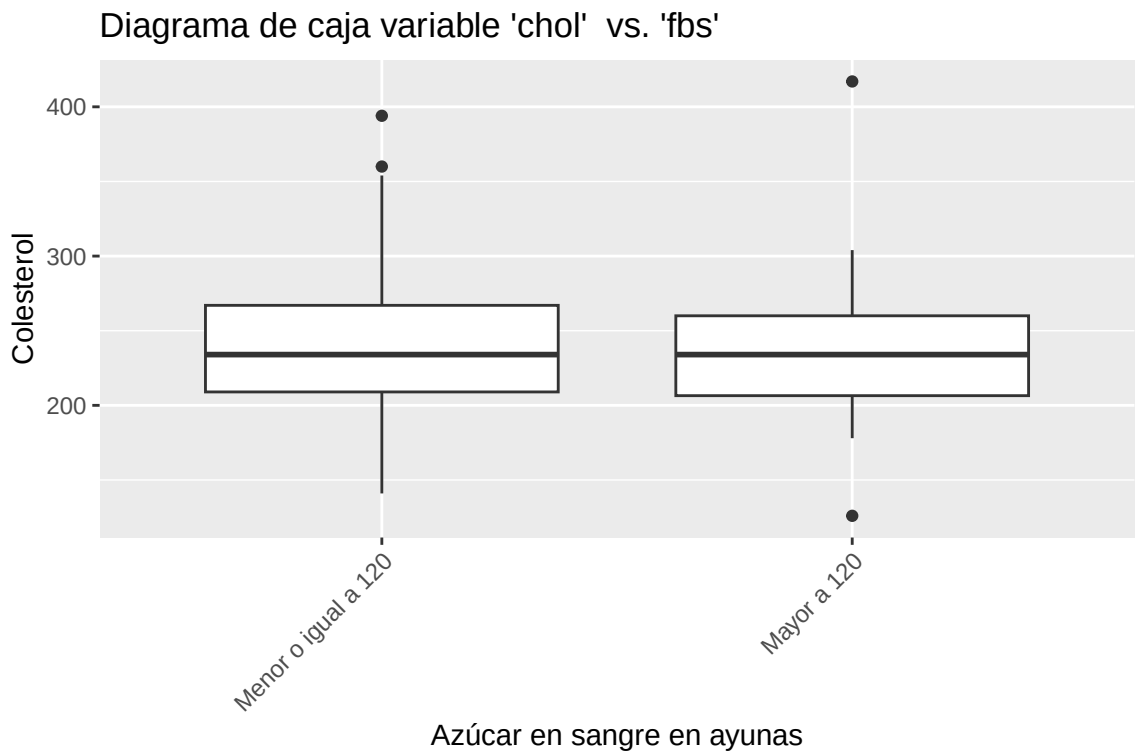
Relación entre el colesterol y el resto de variables

Antes de aplicar pruebas estadísticas para analizar las relaciones entre el colesterol y la edad, el tipo de dolor en el pecho, el nivel de azucar en sangre en ayunas y el sexo, inspeccionaremos los datos visualmente, con la ayuda de diagramas de caja y gráficos de dispersión.

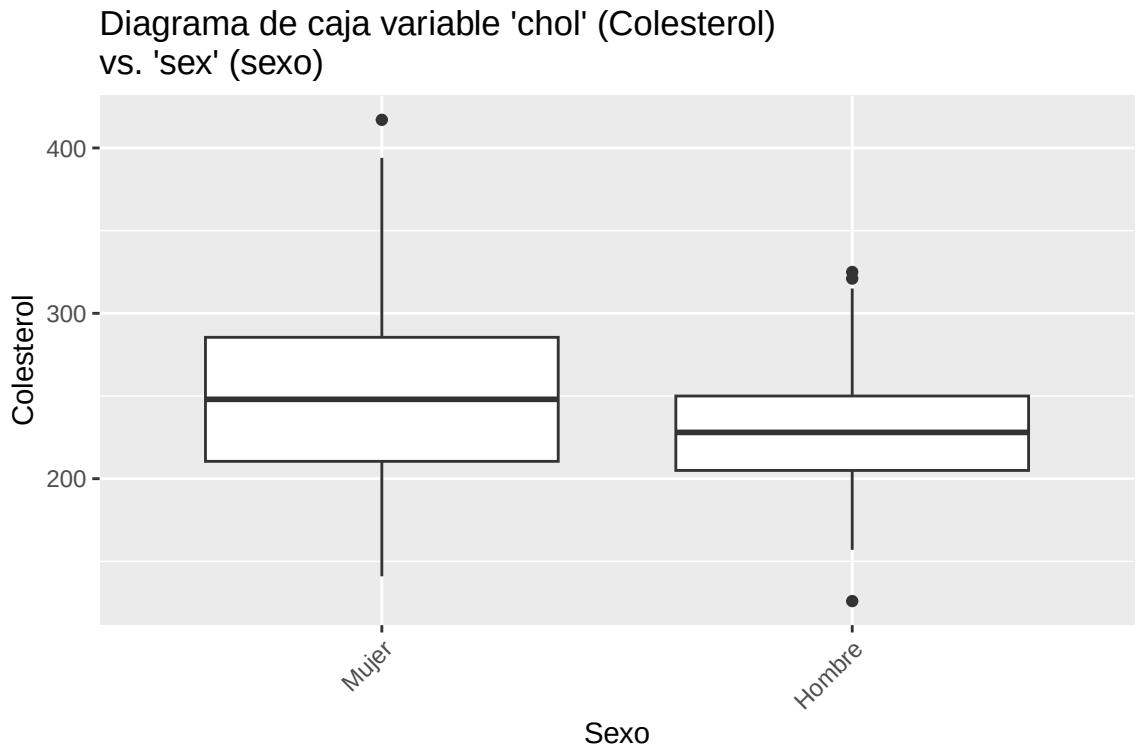
```
ggplot(data, aes(x=factor(cp), y=chol)) +
  geom_boxplot() +
  labs(x="Tipo de dolor en el pecho", y="Colesterol",
       title="Diagrama de caja variable 'chol' (Colesterol) \nvs. 'cp' (tipo de dolor en el pecho)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



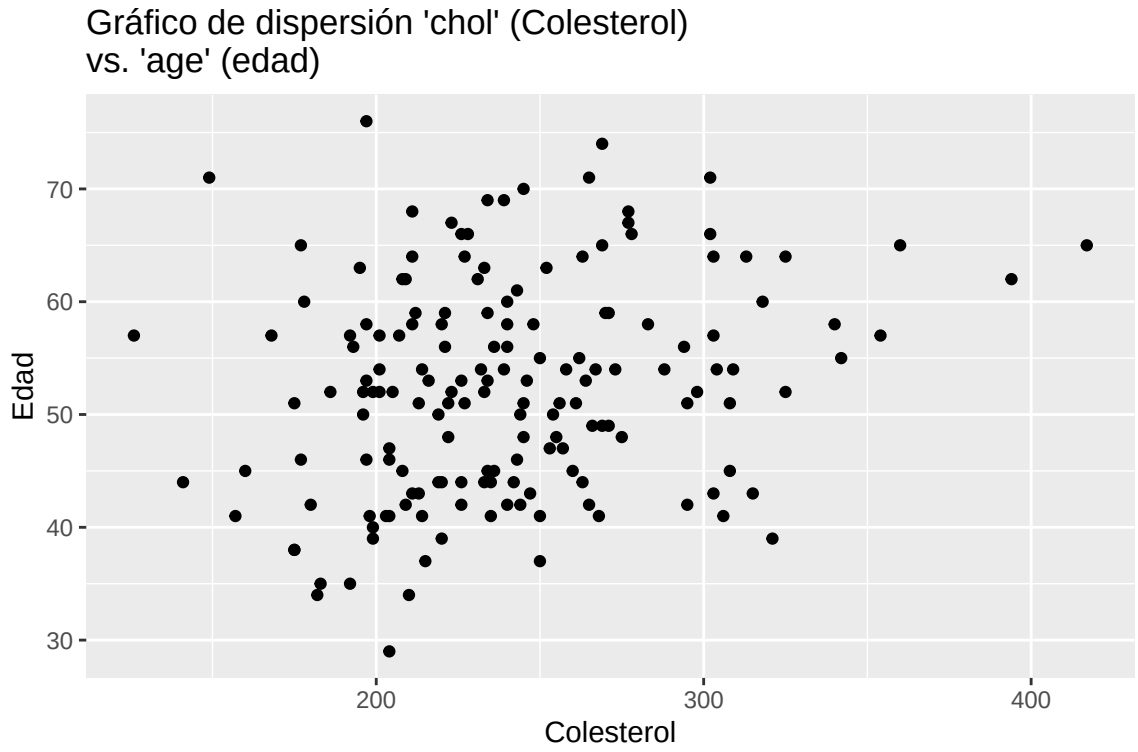

```
ggplot(data, aes(x=factor(fbs), y=chol)) +
  geom_boxplot() +
  labs(x="Azúcar en sangre en ayunas", y="Colesterol",
       title="Diagrama de caja variable 'chol'  vs. 'fbs' ") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(data, aes(x=factor(sex), y=chol)) +
  geom_boxplot() +
  labs(x="Sexo", y="Colesterol",
       title="Diagrama de caja variable 'chol' (Colesterol) \nvs. 'sex' (sexo)") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
ggplot(data, aes(x=chol, y=age)) +
  geom_point() +
  labs(x="Colesterol", y="Edad",
       title="Gráfico de dispersión 'chol' (Colesterol) \nvs. 'age' (edad)")
```



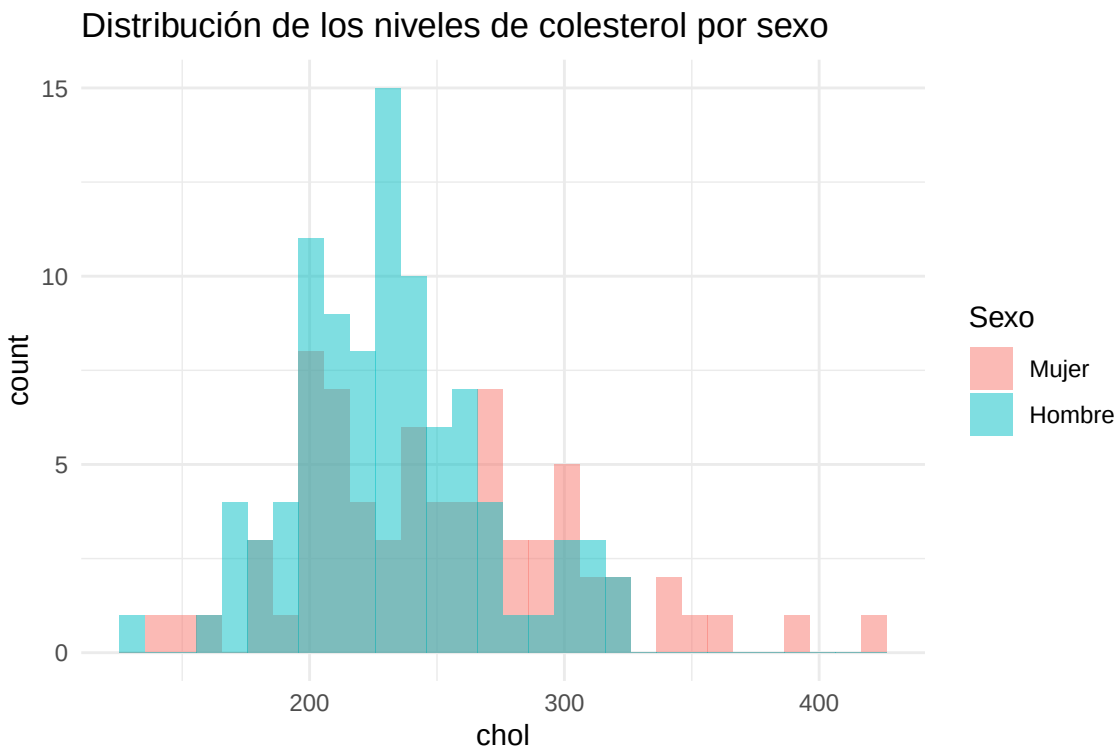
- Relación entre colesterol y el tipo de dolor de pecho: Salvo por algunos valores extremos, las muestras de tipo de dolor con angina típica, atípica y dolor no anginoso parecen comportarse de manera muy similar en cuanto a sus niveles de colesterol, no evidenciándose a simple vista diferencias en su distribución. Sin embargo, los pacientes asintomáticos parecen tener alguna diferencia, con un rango intercuartílico más reducido, un valor máximo menor y un valor mínimo mayor que eb el resto de tipos de dolor, aunque con una mediana muy similar al resto. Podría interpretarse como un indicio de una relación entre el tipo de dolor en el pecho y el nivel de colesterol.
- Relación entre colesterol y azucar en sangre en ayunas: Las medianas de colesterol y el rango intercuartílico son muy similares entre quienes tienen un nivel de azucar en sangre mayor a 120 mg/dL y quienes no, por lo que al menos el 50% de los datos se mueven en rangos similares. Pero cabe decir que el resto de los datos, los niveles que salen de rango intercuartílico, podrían tener diferencias, evidenciado en la posición de los bigotes.
- Relación entre colesterol y sexo: Se puede observar que hombres y mujeres podrían tener diferencias en los niveles de colesterol. Los hombres presentan una mediana menor y el 50% de los datos se mueve en un rango menor y más reducido.
- Relación entre colesterol y la edad: Si comparamos ambas variables numéricas a través de su dispersión, no parece haber algún patrón que indice un tipo de relación fuerte entre la edad y el colesterol, por lo que podría haber una relación débil o inexistente.

Sin embargo, este análisis es preliminar y debe corroborarse mediante la aplicación de tests estadísticos.

Resolución del problema

¿Existen diferencias significativas entre hombres y mujeres en el nivel de colesterol en sangre?

```
# histograma de los niveles de colesterol por sexo
ggplot(data, aes(x = chol, fill = as.factor(sex))) +
  geom_histogram(alpha = 0.5, position = 'identity', bins = 30) +
  labs(fill = "Sexo") +
  theme_minimal() +
  ggtitle("Distribución de los niveles de colesterol por sexo")
```



Parece que hay una distribución relativamente uniforme de niveles de colesterol en hombres y mujeres. Aunque hay algunos niveles de colesterol que son más comunes en hombres o en mujeres, no hay una tendencia clara que permita afirmar que un sexo tiene consistentemente niveles más altos o más bajos de colesterol.

En la prueba Shapiro-Wilk observamos que los datos de colesterol no siguen una distribución normal y en la prueba de Fligner vimos que no se puede asumir igualdad de varianza entre hombres y mujeres. Vamos a usar la prueba de Wilcoxon-Mann-Whitney, una prueba estadística no paramétrica para contestar la pregunta.

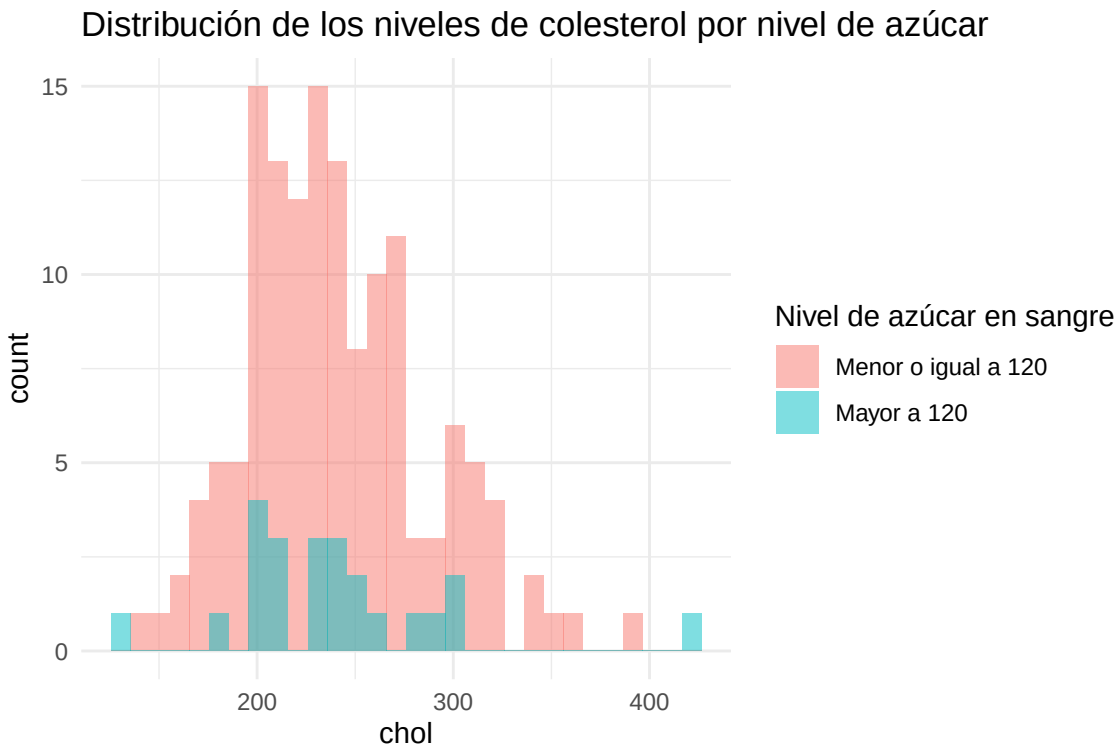
```
# Comparación de los niveles de colesterol por sexo
wilcox.test(chol ~ sex, data = data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  chol by sex
## W = 4047.5, p-value = 0.01335
## alternative hypothesis: true location shift is not equal to 0
```

El p-valor es 0.01335. Esto es menor que el nivel de significancia comúnmente aceptado de 0.05, por lo que podemos rechazar la hipótesis nula de que no hay diferencia en los niveles de colesterol entre hombres y mujeres. En otras palabras, hay una diferencia estadísticamente significativa en los niveles de colesterol entre hombres y mujeres en el conjunto de datos.

¿Existen diferencias significativas entre quienes sobrepasan y quienes no sobrepasan los 120 mg/dl de azúcar en sangre en ayunas, en el nivel de colesterol en sangre?

```
# histograma de los niveles de colesterol por nivel de azúcar en sangre
ggplot(data, aes(x = chol, fill = as.factor(fbs))) +
  geom_histogram(alpha = 0.5, position = 'identity', bins = 30) +
  labs(fill = "Nivel de azúcar en sangre") +
  theme_minimal() +
  ggtitle("Distribución de los niveles de colesterol por nivel de azúcar")
```



Parece que hay una distribución bastante uniforme de niveles de colesterol entre aquellos que sobrepasan y no sobrepasan los 120 mg/dl de azúcar en sangre en ayunas. Al igual que con el estudio de sexo, hay algunos niveles de colesterol que son más comunes en uno u otro grupo, pero no hay una tendencia clara que sugiera que un grupo tiene consistentemente niveles más altos o más bajos de colesterol.

Si bien sí existe igualdad de varianzas entre ambos grupos, recordemos también que el nivel de colesterol no sigue una distribución normal, por lo que, como en el caso anterior, aplicamos la prueba de Wilcoxon-Mann-Whitney

```
# Comparación de los niveles de colesterol entre aquellos que sobrepasan y no sobrepasan 120 mg/dL de a
wilcox.test(chol ~ fbs, data = data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  chol by fbs
## W = 1662, p-value = 0.8497
## alternative hypothesis: true location shift is not equal to 0
```

En este caso p-valor es 0.8497. Esto es mucho mayor que 0.05, por lo que no podemos rechazar la hipótesis nula de que no hay diferencia en los niveles de colesterol entre estos dos grupos. En otras palabras, no hay una diferencia estadísticamente significativa en los niveles de colesterol entre las personas que sobrepasan y las que no sobrepasan 120 mg/dL de azúcar en sangre en ayunas en el conjunto de datos.

En el grupo de personas que tiene mayores probabilidades de tener un infarto, ¿existen diferencias significativas entre los distintos tipos de dolores que presentan las personas en el pecho, en el nivel de colesterol en sangre?

Para resolver esta pregunta hemos seleccionado la prueba de Kruskal-Wallis debido a que la variable de interés, colesterol, aunque si se pueda asumir homocedasticidad, no sigue una distribución normal. La prueba de Kruskal-Wallis es una prueba no paramétrica que no asume una distribución normal de los datos, lo que la hace adecuada en esta situación. Además, esta prueba es útil para comparar más de dos grupos, en este caso, los cuatro tipos de dolores en el pecho.

```
#test de Kruskal
kruskal.test(chol ~ cp, data = data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  chol by cp
## Kruskal-Wallis chi-squared = 1.0444, df = 3, p-value = 0.7905
```

El resultado de la prueba de Kruskal-Wallis nos da un valor de p = 0.7905. Generalmente, si el valor de p es menor que 0.05, se considera que existe una diferencia estadísticamente significativa entre los grupos. Sin embargo, en este caso, el valor de p es mucho mayor que 0.05.

Por lo tanto, con base en esta prueba, concluimos que no hay diferencias significativas en los niveles de colesterol entre los distintos tipos de dolores en el pecho en el grupo de personas con un alto riesgo de infarto. En otras palabras, el tipo de dolor en el pecho que experimenta una persona no parece estar asociado con diferentes niveles de colesterol en este grupo de alto riesgo.

Existe correlación entre la edad y el nivel de colesterol en las personas que tienen mayores probabilidades de tener un infarto?

Para determinar si existe una correlación entre la edad y el nivel de colesterol, hemos decidido usar la prueba de correlación de Spearman. La correlación de Spearman es una prueba no paramétrica que mide la fuerza y dirección de la asociación entre dos variables clasificadas. Esta prueba es apropiada cuando al menos una de las variables no está normalmente distribuida.

```
# Correlación de Spearman entre la edad y el nivel de colesterol
cor.test(~ age + chol, data = data, method = "spearman")
```

```
##
##  Spearman's rank correlation rho
##
## data:  age and chol
## S = 572946, p-value = 0.004529
## alternative hypothesis: true rho is not equal to 0
## sample estimates:
##       rho
## 0.2206199
```

El resultado de la prueba de correlación de Spearman muestra que el coeficiente de correlación rho es 0.2206199. Este valor oscila entre -1 y 1, donde -1 indica una correlación negativa perfecta, 1 una correlación positiva perfecta y 0 ninguna correlación. En este caso, una rho de 0.2206199 indica una correlación positiva débil entre la edad y el nivel de colesterol.

El p-valor de la prueba es 0.004529, que es menor que 0.05, lo que indica que la correlación es significativa a nivel estadístico. En otras palabras, existe una correlación débil pero significativa entre la edad y el nivel de colesterol en las personas que tienen mayores probabilidades de tener un infarto.

Exportación de datos

```
write.csv(data, file = "../data/heart_out.csv", row.names = FALSE)
```

Tabla de contribuciones

```
contribuciones <- data.frame(
  "Contribuciones" = c("Investigación previa", "Redacción de las respuestas", "Desarrollo del código",
  "Firma" = c("JMJ, HMH", "JMJ, HMH", "JMJ, HMH", "JMJ, HMH")
)

kable(contribuciones)
```

Contribuciones	Firma
Investigación previa	JMJ, HMH
Redacción de las respuestas	JMJ, HMH
Desarrollo del código	JMJ, HMH
Participación en el vídeo	JMJ, HMH