Limpieza y Análisis de Datos

Lucas Gómez Torres y Joan Amengual Mesquida

13 de enero, 2023

Índice General

1 Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

Actualmente cada vez sufren más personas ataques al corazón originados por diferentes factores como pueden ser el exceso de colesterol, el nivel de azúcar en la sangre, el consumo de tabaco, la presión arterial, la obesidad, la edad o la falta de ejercicio, entre muchos otros más, que pueden dar lugar a un daño permanente en el corazón como la insuficiencia cardiaca o a la muerte.

Por ello, los ataques al corazón son un problema muy grave que hay que intentar prevenir, analizando las diferentes variables que pueden influir a la hora de que una persona sufra un ataque al corazón o no, pudiendo responder a preguntas como por ejemplo:

- ¿Los hombres son más probables a sufrir un ataque que las mujeres?
- ¿El nivel de azúcar en sangre es determinante para que una persona pueda padecer un ataque?
- ¿Hay diferencias significativas en el nivel de colesterol según padezca o no un ataque?
- ¿Hubo algún indicio de sufrir más fácilmente un ataque al corazón según el dolor de pecho del paciente ?
- ¿Qué factores son los más influyentes para sufrir un ataque?

El conjunto de datos está dividido en dos subconjuntos de datos:

- *heart.csv*: contiene toda la información sobre los pacientes, incluyendo si finalmente sufrieron un ataque al corazón o no. Tiene 303 observaciones y 14 atributos. De estos 14 atributos, 13 son variables independientes y 1 la variable dependiente (variable objetivo que podría servir para construir un modelo de aprendizaje supervisado que nos permita predecir si un paciente tendrá un ataque al corazón o no). A continuación, se describen todos los atributos de este dataset:
  - **age**: Variable de tipo numérica. Determina la edad de la persona.
  - **sex**: Variable de tipo numérica. Refleja el género de la persona *(1 = masculino, 0 = femenino)*.
  - **cp**: Variable de tipo numérica. Identifica el tipo de dolor en el pecho *(0 = angina típica, 1 = angina atípica, 2 = dolor no anginoso, 3 = asintomático)*.
  - **trtbps**: Variable de tipo numérica. Indica la presión arterial en reposo en mg/dl.
  - **chol**: Variable de tipo numérica. Hace referencia al nivel de colesterol en mg/dl.
  - **fbs**: Variable de tipo numérica. Indica si el nivel de azúcar en sangre en ayunas es mayor a 120 mg/dl *(1 = verdadero, 0 = falso)*.
  - **restecg**: Variable de tipo numérica. Muestra los resultados electrocardiográficos en reposo *(0 = normal, 1 = anomalía de onda ST-T (inversiones de onda T y/o elevación o depresión ST de > 0,05 mV), 2 = hipertrofia ventricular izquierda probable o definida por los criterios de Estes)*.
  - **thalachh**: Variable de tipo numérica. Determina la frecuencia cardiaca máxima alcanzada.
  - **exng:**: Variable de tipo numérica. Indica si la angina ha sido inducida por el ejercicio *(1 = sí, 0 = no)*.
  - **oldpeak**: Variable de tipo numérica. Señala la depresión ST inducida por el ejercicio en relación con el descanso.
  - **slp**: Variable de tipo numérica. Muestra la pendiente del segmento ST de ejercicio máximo *(0 = inclinación hacia abajo, 1 = plano, 2 = inclinación hacia arriba)*.
  - **caa**: Variable de tipo numérica. Indica el número de vasos principales *(0, 1, 2, 3)*.

  - **thall**: Variable de tipo numérica. Señala el ratio de un trastorno sanguíneo llamado talasemia *(0 = no tiene, 1 = defecto fijo (sin flujo sanguíneo en alguna parte del corazón), 2 = flujo sanguíneo normal, 3 = defecto reversible (se observa un flujo sanguíneo, pero no es normal)).*
  - **output**: Variable de tipo numérica. Indica si el paciente sufre un ataque al corazón o no (0 = No, 1 = Sí). Se trata de la variable objetivo o dependiente que se puede utilizar para predecir.

- *o2Saturation.csv*: contiene 3585 observaciones sobre los niveles de oxígeno en la sangre de distintos pacientes y solo tiene 1 atributo.

2 Integración y selección de los datos de interés a analizar.

Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

En este apartado se van a cargar ambos conjuntos de datos, para decidir si se van a unificar ambos o no, o si nos vamos a centrar en unos pacientes concretos limitando el número de registros o de características con el fin de reducir el dataset. Además, en el dataset de *heart.csv* se van a renombrar los atributos para que se entiendan mejor y sean más intuitivos a la hora de utilizarlos más adelante.

```
# Se carga el dataset
heart_data <- read.csv("heart.csv", header = TRUE)

# Modificamos los nombres de las variables para que sean más intuitivos
colnames(heart_data) <- c("age","sex","chest_pain_type","resting_blood_pressure",
                          "cholesterol",    "fasting_blood_sugar","rest_ecg_type",
                          "max_heart_rate_achieved","exercise_induced_angina",
                          "st_depression", "st_slope_type", "num_major_vessels",
                          "thalassemia_type","heart_attack")

# Dimensión del dataset
dim(heart_data)
```

```
## [1] 303  14
```

```
# Se carga el dataset
O2_saturation <- read.csv("o2Saturation.csv", header = TRUE)

# Dimensión del dataset
dim(O2_saturation)
```

```
## [1] 3585    1
```

Podemos observar que ambos conjuntos de datos tienen dimensiones diferentes. El que contiene los niveles de oxígeno en la sangre consta de 3.585 observaciones, es decir, diferentes niveles de oxígeno para 3.585 pacientes, en cambio, el otro, contiene información sobre 303 pacientes y 14 características distintas. Como ya tenemos suficientes características en el dataset de *heart.csv* con las que poder realizar un estudio detallado y completo a las preguntas que hemos planteado al principio, se va a optar por descartar el otro conjunto y perder este atributo adicional de los pacientes.

En el caso de haber querido unificarlos y por lo tanto añadir otro atributo al dataset de *heart.csv* (saturación de oxígeno), se podría haber utilizado la función *merge* permitiéndonos fusionarlos de forma horizontal. Posteriormente, se podría comprobar que no existen inconsistencias ni duplicidades en los registros con la función *duplicated* o *unique*. No obstante, no existe un identificador único para cada uno de los pacientes como podría ser un id o un nombre, por lo que suponemos que podría haber dos pacientes con los mismos valores de atributos. Asimismo comprobaremos si hay muchos registros duplicados con el fin de que no pueda afectar significamente en los análisis posteriores.

```
# Comprobamos si existen registros duplicados con los mismos valores en todos los campos
# (dado que no tenemos identificador) Y contamos cuántos son

nrow(heart_data[duplicated(heart_data), ])
```

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

```
# Vemos los registros que están duplicados
heart_data[duplicated(heart_data), ]
```

```
##     age sex chest_pain_type resting_blood_pressure cholesterol
## 165  38   1               2                    138         175
##     fasting_blood_sugar rest_ecg_type max_heart_rate_achieved
## 165                   0             1                     173
##     exercise_induced_angina st_depression st_slope_type num_major_vessels
## 165                       0             0             2                 4
##     thalassemia_type heart_attack
## 165                2            1
```

Dado que solo existe un registro duplicado, con los mismos valores en todos los campos, no se va a eliminar porque es un porcentaje muy bajo del total y no afectará de manera significativa a los resultados que obtendremos más adelante. Además, al ser solo un registro, podría ser el caso de que esos dos pacientes fueran distintos y tuvieran las mismas características. Si tuviéramos muchos más, entonces seguramente serían los mismos pacientes y tendríamos que eliminarlos.

A continuación, se muestran algunos registros e información general de los datos que servirá para posteriomente proceder a la limpieza y conversión de los datos.

```
# Mostramos los tipos de datos de las variables tal y como las interpreta R
sapply(heart_data,class)
```

```
##                      age                      sex         chest_pain_type
##                "integer"                "integer"                "integer"
##  resting_blood_pressure              cholesterol     fasting_blood_sugar
##                "integer"                "integer"                "integer"
##           rest_ecg_type max_heart_rate_achieved exercise_induced_angina
##                "integer"                "integer"                "integer"
##           st_depression           st_slope_type       num_major_vessels
##                "numeric"                "integer"                "integer"
##        thalassemia_type             heart_attack
##                "integer"                "integer"
```

```
# Mostramos un resumen de los datos
summary(heart_data)
```

```
##       age             sex         chest_pain_type resting_blood_pressure
##  Min.   :29.00   Min.   :0.0000   Min.   :0.000   Min.   : 94.0
##  1st Qu.:47.50   1st Qu.:0.0000   1st Qu.:0.000   1st Qu.:120.0
##  Median :55.00   Median :1.0000   Median :1.000   Median :130.0
##  Mean   :54.37   Mean   :0.6832   Mean   :0.967   Mean   :131.6
##  3rd Qu.:61.00   3rd Qu.:1.0000   3rd Qu.:2.000   3rd Qu.:140.0
##  Max.   :77.00   Max.   :1.0000   Max.   :3.000   Max.   :200.0
##   cholesterol    fasting_blood_sugar rest_ecg_type    max_heart_rate_achieved
##  Min.   :126.0   Min.   :0.0000      Min.   :0.0000   Min.   : 71.0
##  1st Qu.:211.0   1st Qu.:0.0000      1st Qu.:0.0000   1st Qu.:133.5
##  Median :240.0   Median :0.0000      Median :1.0000   Median :153.0
##  Mean   :246.3   Mean   :0.1485      Mean   :0.5281   Mean   :149.6
```

```
##  3rd Qu.:274.5   3rd Qu.:0.0000      3rd Qu.:1.0000   3rd Qu.:166.0
##  Max.   :564.0   Max.   :1.0000      Max.   :2.0000   Max.   :202.0
##  exercise_induced_angina st_depression  st_slope_type   num_major_vessels
##  Min.   :0.0000          Min.   :0.00   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.0000          1st Qu.:0.00   1st Qu.:1.000   1st Qu.:0.0000
##  Median :0.0000          Median :0.80   Median :1.000   Median :0.0000
##  Mean   :0.3267          Mean   :1.04   Mean   :1.399   Mean   :0.7294
##  3rd Qu.:1.0000          3rd Qu.:1.60   3rd Qu.:2.000   3rd Qu.:1.0000
##  Max.   :1.0000          Max.   :6.20   Max.   :2.000   Max.   :4.0000
##  thalassemia_type  heart_attack
##  Min.   :0.000    Min.   :0.0000
##  1st Qu.:2.000    1st Qu.:0.0000
##  Median :2.000    Median :1.0000
##  Mean   :2.314    Mean   :0.5446
##  3rd Qu.:3.000    3rd Qu.:1.0000
##  Max.   :3.000    Max.   :1.0000
```

```
# Se muestran las 4 primeras observaciones de los datos
head(heart_data,4)
```

```
##   age sex chest_pain_type resting_blood_pressure cholesterol
## 1  63   1               3                   145         233
## 2  37   1               2                   130         250
## 3  41   0               1                   130         204
## 4  56   1               1                   120         236
##   fasting_blood_sugar rest_ecg_type max_heart_rate_achieved
## 1                   1             0                     150
## 2                   0             1                     187
## 3                   0             0                     172
## 4                   0             1                     178
##   exercise_induced_angina st_depression st_slope_type num_major_vessels
## 1                       0           2.3             0                 0
## 2                       0           3.5             0                 0
## 3                       0           1.4             2                 0
## 4                       0           0.8             2                 0
##   thalassemia_type heart_attack
## 1                1            1
## 2                2            1
## 3                2            1
## 4                2            1
```

Por último, para nuestro análisis no se van a descartar registros porque no nos vamos a centrar en un tramo de edad concreto, sexo o una cantidad de colesterol, sino que se van a considerar a todos los pacientes con todas sus características para extraer el mayor número de conclusiones posibles teniendo en cuenta todos los atributos.

3 Limpieza de los datos.

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

3.1.1 Caso: Ceros

```
# Analisis de las columnas que contienen ceros en sus valores
cols_with_zeros <- which(apply(heart_data, 2, function(x) sum(x == 0)) > 0)
```

```r
colnames(heart_data)[cols_with_zeros]
```

```
##  [1] "sex"                    "chest_pain_type"
##  [3] "fasting_blood_sugar"    "rest_ecg_type"
##  [5] "exercise_induced_angina" "st_depression"
##  [7] "st_slope_type"          "num_major_vessels"
##  [9] "thalassemia_type"       "heart_attack"
```

Las variables que contienen algún valor igual a cero son variables que esperan reflejar este valor tal y como se ha definido en el enunciado, por lo tanto no se va a realizar una limpieza de datos para este caso en particular. Véase a continuación las variables que aparecen con algún valor cero son las siguientes:

- "sex": Refleja el género de la persona (1 = masculino, 0 = femenino).
- "chest_pain_type": Identifica el tipo de dolor en el pecho (0 = angina típica, 1 = angina atípica, 2 = dolor no anginoso, 3 = asintomático).
- "fasting_blood_sugar": Indica si el nivel de azúcar en sangre en ayunas es mayor a 120 mg/dl (1 = verdadero, 0 = falso).
- "rest_ecg_type": Muestra los resultados electrocardiográficos en reposo (0 = normal, 1 = anomalía de onda ST-T (inversiones de onda T y/o elevación o depresión ST de $>$ 0,05 mV), 2 = hipertrofia ventricular izquierda probable o definida por los criterios de Estes).
- "exercise_induced_angina": Indica si la angina ha sido inducida por el ejercicio (1 = sí, 0 = no).
- "st_depression": Señala la depresión ST inducida por el ejercicio en relación con el descanso.
- "st_slope_type": Muestra la pendiente del segmento ST de ejercicio máximo (0 = inclinación hacia abajo, 1 = plano, 2 = inclinación hacia arriba).
- "num_major_vessels": Indica el número de vasos principales (0, 1, 2, 3).
- "thalassemia_type": Señala el ratio de un trastorno sanguíneo llamado talasemia (0 = no tiene, 1 = defecto fijo (sin flujo sanguíneo en alguna parte del corazón), 2 = flujo sanguíneo normal, 3 = defecto reversible (se observa un flujo sanguíneo, pero no es normal)).
- "heart_attack": Indica si el paciente sufre un ataque al corazón o no (0 = No, 1 = Sí).

3.1.2 Caso: Elementos Vacíos

A continuación se realiza la comprobación de si hay elementos vacíos en el dataset, para cada columna se realiza el conteo de elementos vacíos existentes.

```r
# Elementos vacíos de las variables del dataset
colSums(is.na(heart_data))
```

```
##                      age                      sex          chest_pain_type
##                        0                        0                        0
##   resting_blood_pressure              cholesterol      fasting_blood_sugar
##                        0                        0                        0
##           rest_ecg_type  max_heart_rate_achieved  exercise_induced_angina
##                        0                        0                        0
##           st_depression           st_slope_type        num_major_vessels
##                        0                        0                        0
##        thalassemia_type             heart_attack
##                        0                        0
```

Como se visualiza en los resultados anteriores no existen elementos vacíos en el conjunto de datos. Con ello, no será necesario realizar ningún procedimiento de limpieza de datos para valores vacíos de las variables del dataset.

3.1.3 Conversión y adaptación de los datos

Se van a realizar algunas conversiones de los tipos de algunas variables para realizar un análisis más eficiente y que nos facilite la interpretación de los resultados.

Primero convertiremos las siguientes variables numéricas a categóricas:

```
# Transformamos a tipo factor las siguientes variables
heart_data$sex <- factor(heart_data$sex, levels = c(0,1), labels=
                           c("Femenino", "Masculino"))

heart_data$chest_pain_type <- factor(heart_data$chest_pain_type, levels = c(0,1,2,3),
                                     labels=
                                       c("Angina típica", "Angina atípica",
                                        "Dolor no anginoso","Asintomático"))

heart_data$fasting_blood_sugar <- factor(heart_data$fasting_blood_sugar, levels = c(0,1),
                                         labels=
                                           c("Azúcar Bajo", "Azúcar Alto"))

heart_data$rest_ecg_type <- factor(heart_data$rest_ecg_type, levels = c(0,1,2), labels=
                                     c("Normal", "Anomalía de onda ST-T",
                                      "Hipertrofia ventricular izquierda"))

heart_data$exercise_induced_angina <- factor(heart_data$exercise_induced_angina,
                                             levels = c(0,1), labels= c("No", "Sí"))

heart_data$st_slope_type <- factor(heart_data$st_slope_type, levels = c(0,1,2),
                                   labels= c("Baja", "Normal","Alta"))
heart_data$thalassemia_type <- factor(heart_data$thalassemia_type, levels = c(0,1,2,3),
                                      labels= c("Inexistente", "Fijo",
                                          "Normal","Reversible"))
heart_data$heart_attack <- factor(heart_data$heart_attack, levels = c(0,1),
                                      labels= c("No", "Yes"))
```

También se pueden aplicar otro tipo de conversiones como por ejemplo la normalización *z-score* que resta la media a la variable y la divide por su desviación estándar.

En el caso de las variables que no presenten una distribución normal, como será el caso de *cholesterol*, una opción sería realizar transformaciones de tipo Box-Cox para poder mejorar su normalidad y su homocedasticidad.

Asimismo, para algunas variables como por ejemplo la edad del paciente, sería interesante realizar un proceso de discretización. Esto nos permitiría agrupar las edades en diferentes grupos y poder sacar conclusiones que nos aporten un valor simbólico más allá de solo un número, aportándonos mayor información.

3.2 Identifica y gestiona los valores extremos

En primer lugar se realiza la visualización de los valores extremos para las variables: *"age", "cholesterol", "max_heart_rate_achieved", "resting_blood_pressure", "st_depression".*

```
outlier_info <- function(var, name_var, show_plot = TRUE) {
  # Valores extremos en formato boxplot de la variable
  if (show_plot) {
    boxplot(var, main = name_var,
            ylab="Valor", col = "lightblue", horizontal = FALSE, outline = TRUE)
  }
```

```
  # Identificar los valores atípicos
  outliers <- boxplot.stats(var)$out

  # Imprimir los valores máximo y mínimo de los valores atípicos
  stats <- boxplot.stats(var)$stats
  cat("Valor mínimo:", stats[1], "\n")
  cat("Primer cuartil:", stats[2], "\n")
  cat("Media:", stats[3], "\n")
  cat("Tercer cuartil:", stats[4], "\n")
  cat("Valor máximo:", stats[5], "\n")

  if (length(outliers) == 0) {
    cat("No se han identificado valores atípicos", "\n")
  } else {
    # Imprimir el número de valores atípicos
    cat("Outliers identificados:", unique(outliers), "\n")
  }
}
```

```
par(mfrow=c(2, 3))
outlier_info(heart_data$age, "age")
```

```
## Valor mínimo: 29
## Primer cuartil: 47.5
## Media: 55
## Tercer cuartil: 61
## Valor máximo: 77
## No se han identificado valores atípicos
```

```
outlier_info(heart_data$cholesterol, "cholesterol")
```

```
## Valor mínimo: 126
## Primer cuartil: 211
## Media: 240
## Tercer cuartil: 274.5
## Valor máximo: 360
## Outliers identificados: 417 564 394 407 409
```

```
outlier_info(heart_data$max_heart_rate_achieved, "max_heart_rate_achieved")
```

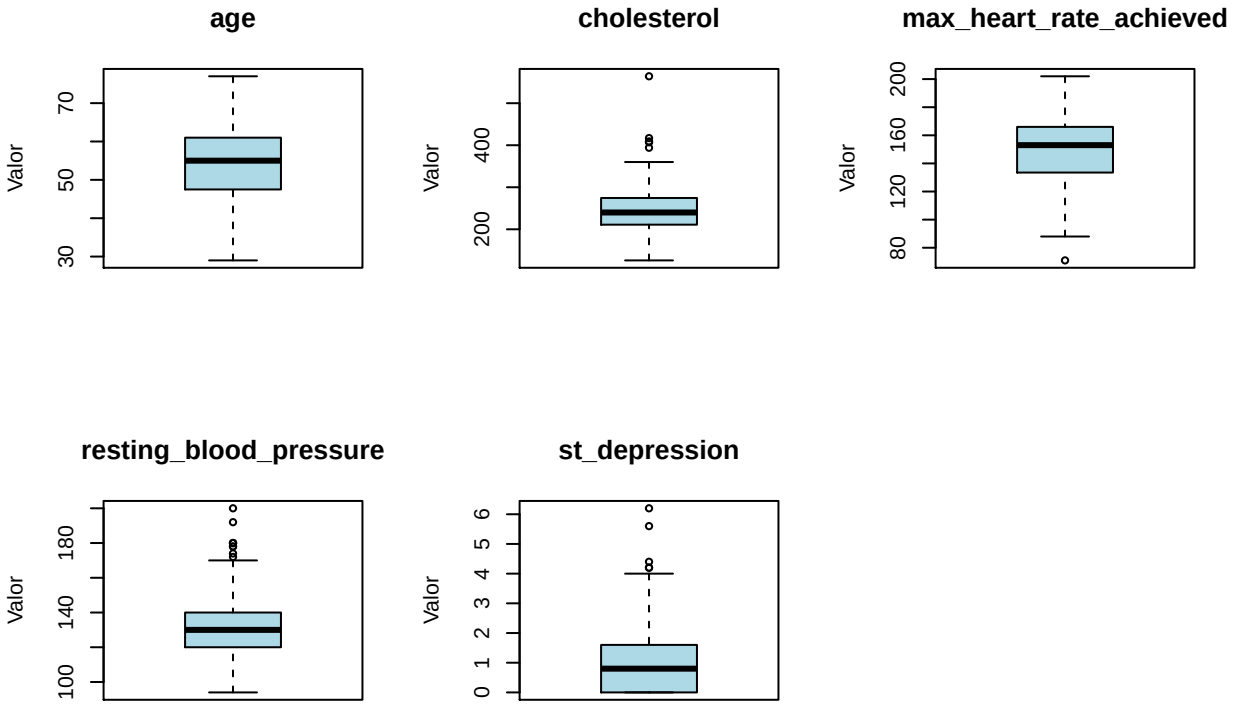
```
## Valor mínimo: 88
## Primer cuartil: 133.5
## Media: 153
## Tercer cuartil: 166
## Valor máximo: 202
## Outliers identificados: 71
```

```
outlier_info(heart_data$resting_blood_pressure, "resting_blood_pressure")
```

```
## Valor mínimo: 94
## Primer cuartil: 120
## Media: 130
## Tercer cuartil: 140
## Valor máximo: 170
## Outliers identificados: 172 178 180 200 174 192
```

```
outlier_info(heart_data$st_depression, "st_depression")
```

```
## Valor mínimo: 0
## Primer cuartil: 0
## Media: 0.8
## Tercer cuartil: 1.6
## Valor máximo: 4
## Outliers identificados: 4.2 6.2 5.6 4.4
```



A continuación vamos a extraer las conclusiones pertinentes respectos a los valores extremos detectados en los resultados y los gráficos previos:

- En la variable "age", no se han identificado valores atípicos. Los valores máximo y mínimo de la variable son 29 y 77, respectivamente.
- En la variable "cholesterol", se han identificado 5 valores atípicos (outliers). Los valores máximo y mínimo de los outliers identificados son 126 y 564, respectivamente.
- En la variable "max_heart_rate_achieved", se ha identificado 1 valor atípico (outlier). Los valores máximo y mínimo de los outliers identificados son 71 y 202, respectivamente.
- En la variable "resting_blood_pressure", se han identificado 6 valores atípicos (outliers). Los valores máximo y mínimo de los outliers identificados son 94 y 200, respectivamente.
- En la variable "st_depression", se han identificado 4 valores atípicos (outliers). Los valores máximo y mínimo de los outliers identificados son 0 y 6.2, respectivamente.

Estos resultados indican que algunas de las variables tienen valores extremos que se alejan significativamente del resto y que pueden afectar el rendimiento de algunos algoritmos de análisis de datos.

3.3 Corrección de los outliers

Es importante destacar que un valor extremo no tiene por que ser no válido, para determinar si un valor extremo es válido o no hemos realizado una investigación sobre las variables y los posibles valores que estas pueden tener. De todos los outliers detectados simplemente nos centramos en el caso de la variable *cholesterol*.

Realizando una búsqueda sobre los valores comúnes y menos comúnes de cholesterol (mg / dL), en 300 mg/dl o más ya se considera un nivel muy alto. Para casos más elevados se habla de sufrir hipertrigliceridemia. Nosotros hemos establecido que para un valor mayor a 550 se va a realizar una corrección de este valor.

```
# Número de outliers que superan el valor establecido en la variable tratada
num_outliers_var = nrow(heart_data[heart_data$cholesterol >550,])

# Número de NA's en la variable tratada
num_nas_variable_inicio = sum(is.na(heart_data$cholesterol))

# Se substituyen esos valores atipicos por el valor NA
heart_data = heart_data %>% mutate(cholesterol = ifelse(cholesterol > 550, NA, cholesterol))

# Numero de NA's final después del tratado en la variable
num_nas_variable_final = sum(is.na(heart_data$cholesterol))

# Visualización de los resultados
print(paste("Número de outliers que superan el valor establecido en la variable:",
            num_outliers_var))
```

```
## [1] "Número de outliers que superan el valor establecido en la variable: 1"
```

```
print(paste("Número inicial de NA's en la variable:", num_nas_variable_inicio))
```

```
## [1] "Número inicial de NA's en la variable: 0"
```

```
print(paste("Número final de NA's en la variable después del tratado:",
            num_nas_variable_final))
```

```
## [1] "Número final de NA's en la variable después del tratado: 1"
```

3.4 Imputación de valores

Se va a imputar la media aritmética a esos valores NA's. La imputación de valores por media aritmética es un método utilizado para reemplazar valores faltantes o perdidos en un conjunto de datos. Este método consiste en reemplazar el valor faltante con la media aritmética de los valores presentes en el conjunto de datos.

```
# Se calcula la media aritmética
mean_cholesterol = mean(heart_data$cholesterol,na.rm=T)
# Se redondea la media
mean_cholesterol=round(mean_cholesterol,2)
mean_cholesterol
```

```
## [1] 245.21
```

```
# Imputamos la media aritmética en los valores nulos
heart_data$cholesterol[is.na(heart_data$cholesterol)] <- mean_cholesterol
```

Finalmente se puede observar que dicha variable ya no contiene valores nulos, ya que estos se han remplazado por la media aritmética.

```
# Comprobación de no existencia de valores nulos
sum(is.na(heart_data$cholesterol))
```

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

4 Análisis de los datos.

4.1 Selección de los grupos de datos que se quieren analizar/comparar (p.ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?).

Como comentamos al principio de la práctica, queremos responder a las siguientes preguntas:

- ¿Los hombres son más probables a sufrir un ataque que las mujeres?
- ¿El nivel de azúcar en sangre es determinante para que una persona pueda padecer un ataque?
- ¿Hay diferencias significativas en el nivel de colesterol según padezca o no un ataque?
- ¿Hubo algún indicio de sufrir más fácilmente un ataque al corazón según el dolor de pecho del paciente ?
- ¿Qué factores son los más influyentes para sufrir un ataque?

En nuestro caso, se va a analizar el conjunto de datos *heart.csv* para intentar respuesta a las preguntas anteriores. Para ello, se harán diferentes contrastes de hipótesis realizando diferentes análisis estádisticos como la *prueba t de student*, el *test de Wilcoxon* y el *test de chi cuadrado*. Además, se construirá un modelo de regresión logística para poder analizar qué variables son las que más influyen a la hora de un paciente sufra o no un ataque al corazón, tomando como variable dependiente *heart_attack*.

Se comparará el valor de *cholesterol*, la edad de los pacientes (*age*) y la presión arterial en reposo (*resting_blood_pressure*) entre sufrir o no un ataque (*heart_attack*). También se comparará el valor de *cholesterol* entre hombres y mujeres (*sex*), y se analizará si hay diferencias significativas entre las variables categóricas *sex*, *fasting_blood_sugar* y el dolor de pecho (*chest_pain_type*) con *heart_attack* y entre el nivel de azúcar en sangre (*fasting_blood_sugar*) y *sex*.

A la hora de aplicar algunos tests estadísticos, se tendrá que tener en cuenta la normalidad y la homocedasticidad de las variables como se verá en el apartado siguiente.

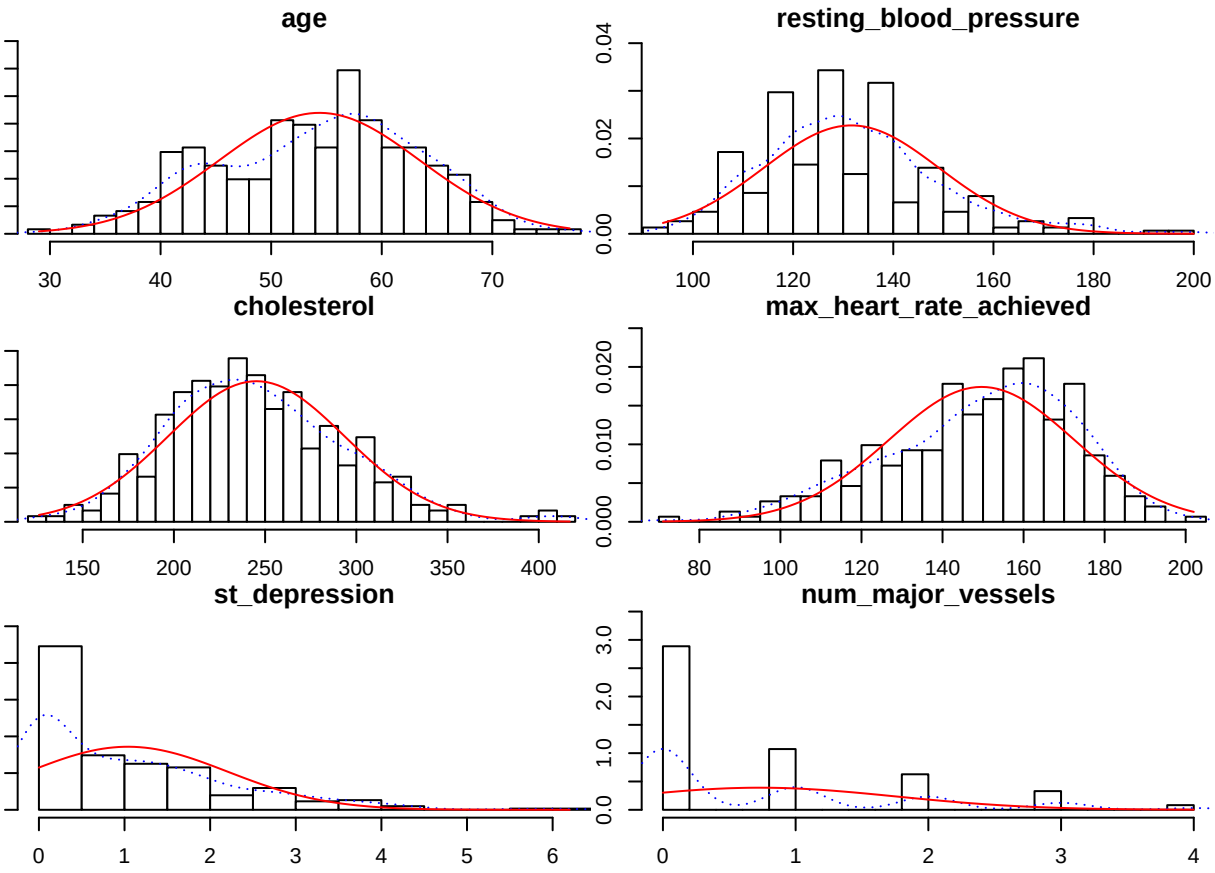
4.2 Comprobación de la normalidad y homogeneidad de la varianza.

4.2.1 Normalidad

Se va a analizar la normalidad y la homocedasticidad de las variables cuantitativas que nos servirán para dar respuesta a las preguntas anteriores.

Para ello, vamos a representar mediante histogramas la distribución de los datos de las variables en comparación con la normal teórica para poder ver visualmente si siguen una distribución normal. No obstante, después lo verificaremos mediante el test de *Shapiro Wilk* y mediante el gráfico *Q-Q plot* mediante las funciones de R `qqnorm` y `qqline`.

```
# Indices de las variables cuantitativas
idx_var_cuant <- c(1,4,5,8,10,12)
# Histograma de la distribución de la variable VS la distribucion normal teorica
multi.hist(x = heart_data[,idx_var_cuant],
           dcol = c("blue", "red"),
           dlty = c("dotted", "solid"),
           global=FALSE)
```



Podemos comprobar como visualmente no siguen una distribución normal ninguna de las variables, no obstante, algunas variables como *age*, *cholesterol*, *max_heart_rate_achieved* y *resting_blood_pressure* no se le alejan mucho de la normal.

A continuación, se va a ratificar lo anterior aplicando el test de Shapiro Wilk a cada una de las variables.

```
# Devuelve el p-valor aplicando el test de Shapiro Wilk
p_value_shapiro_wilk <- function(x) {
  p_value <- shapiro.test(x)["p.value"]
  return (p_value)
}

# Se crea una dataframe con los p-valores obtenidos para cada variable
df_p_value_shapiro_wilk <- data.frame(
"P-Value" = sapply(heart_data[,idx_var_cuant],p_value_shapiro_wilk))

# Se le añade el nombre a las variables
colnames(df_p_value_shapiro_wilk) <- c("Age","Resting_Blood_Pressure",
                                       "Cholesterol",
                                       "Max_heart_rate_achieved",
                                       "st_depression",
                                       "num_major_vessels")

# Se visualiza el dataframe creado
kable(df_p_value_shapiro_wilk,digits=3,
      caption="P-Valores de las variables cuantitativas aplicando Shapiro Wilk")
```

Viendo los resultados del test con unos p-valores inferiores al nivel de signficancia de 0.05, se rechaza la

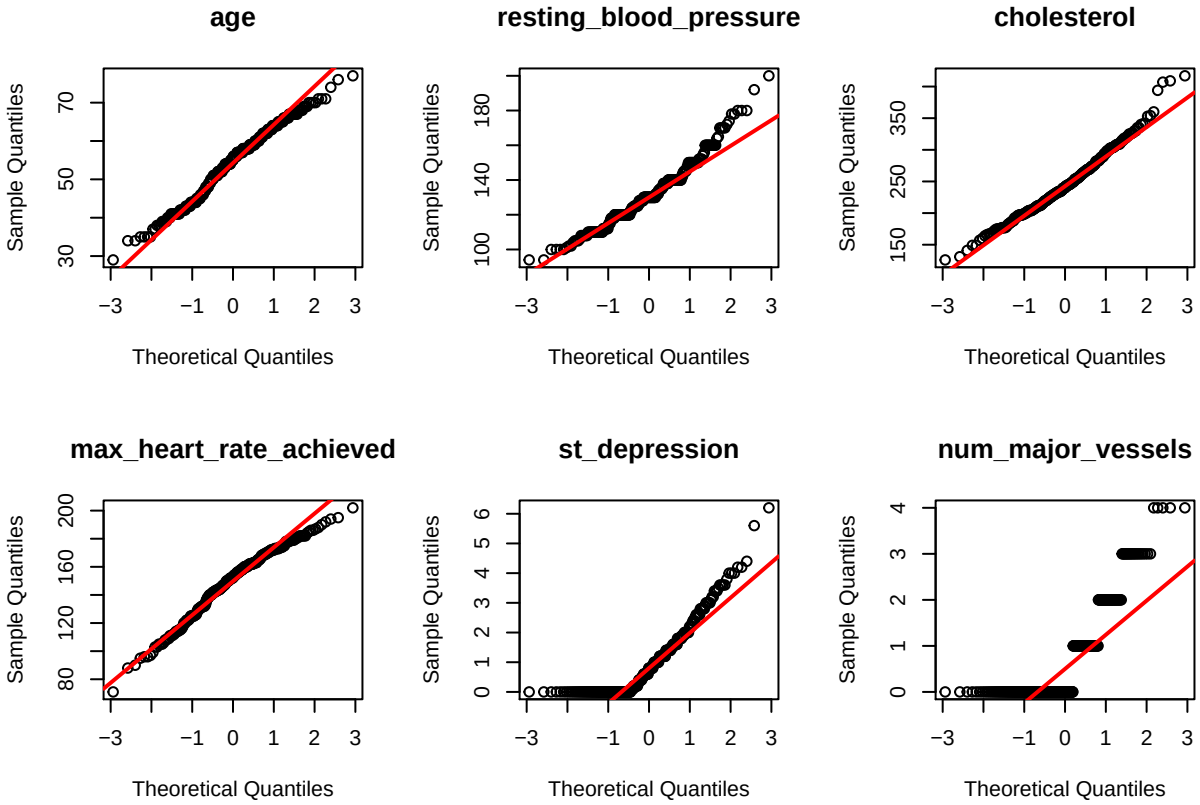
Table 1: P-Valores de las variables cuantitativas aplicando Shapiro Wilk

Age	Resting_Blood_Pressure	Cholesterol	Max_heart_rate_achieved	st_depression	num_major_vessels
0.006	0	0.001	0	0	0

hipótesis nula y se confirma que la distribución de las variables no siguen una distribución normal al 95% de confianza.

Por último, se muestra el *Q-Q plot* que representa en el eje X los cuantiles teóricos (la variable normal estándar) y en el eje Y los valores ordenados de la muestra de cada variable, con el fin de ver la similitud entre la distribución de la muestra y una distribución normal con media 0 y desviación estándar 1.

```
par(mfrow = c(2, 3))
for (var in idx_var_cuant){
  qqnorm(heart_data[,var], main=colnames(heart_data)[var], pch=1)
  qqline(heart_data[,var],col='red', lwd=2) }
```



Como podíamos comprobar con el histograma y con el test de *Shapiro Wilk*, las variables *age*, *cholesterol*, *max_heart_rate_achieved* y *resting_blood_pressure* no se ajustan del todo a una distribución normal porque presentan un gran número de muestras en los extremos izquierdo y derecho que se encuentran fuera de la recta de regresión, sin embargo, se puede ver que la mayoría de las muestras sí. Las demás variables, *st_depression* y *num_major_vessels* sí que se alejan mucho de una distribución normal.

Por lo tanto, *ninguna variable sigue una distribución normal*. Sin embargo, para las que más se acercan se intentará transformar los datos para que sean normales con la transformación de *Box-Cox* y volviendo a aplicar el test de *Shapiro Wilk* para verificar la transformación.

```
# Se aplica la transformación de Box Cox a las variables age,
# cholesterol, max_heart_rate_achieved y resting_blood_pressure
```

```r
# Age
age_norm <- BoxCox(heart_data$age,
                   lambda = BoxCoxLambda(heart_data$age))
shapiro.test(age_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  age_norm
## W = 0.98786, p-value = 0.01216
```

```r
# cholesterol
cholesterol_norm <- BoxCox(heart_data$cholesterol,
                           lambda = BoxCoxLambda(heart_data$cholesterol))
shapiro.test(cholesterol_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  cholesterol_norm
## W = 0.99668, p-value = 0.7855
```

```r
# max_heart_rate_achieved
max_heart_rate_achieved_norm <- BoxCox(heart_data$max_heart_rate_achieved,

                                       lambda = BoxCoxLambda(heart_data$max_heart_rate_achieved))
shapiro.test(max_heart_rate_achieved_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  max_heart_rate_achieved_norm
## W = 0.99146, p-value = 0.07686
```

```r
# resting_blood_pressure
resting_blood_pressure_norm <- BoxCox(heart_data$resting_blood_pressure,

                                      lambda = BoxCoxLambda(heart_data$resting_blood_pressure))
shapiro.test(resting_blood_pressure_norm)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resting_blood_pressure_norm
## W = 0.99029, p-value = 0.04192
```

Podemos comprobar como después de aplicar la transformación de Box-Cox las variables *cholesterol* y *max_heart_rate_achieved* siguen una distribución normal al tener el p-valor superior a 0.05 aceptando la hipótesis nula de que la muestra es normal. En cambio, *age* y *resting_blood_pressure_* siguen sin ser normales.

Por último, incluimos al dataset la variable normal transformada de *cholesterol* para que se pueda aplicar con ella tests de tipo paramétricos. No incluimos *max_heart_rate_achieved* aunque sea normal porque no se utilizará para nuestros contrastes de hipótesis.

```r
# Sustituimos en la variable colesterol la variable transformada para que siga normal
heart_data$cholesterol <- cholesterol_norm
```

4.2.2 Homogeneidad de varianzas

Para comprobar si las variables que usamos para responder a las preguntas previamente planteadas tienen o no homogeneidad de varianzas, se aplicará el *test de Levene* (parámetrico) si las variables cuantitativas son normales (en este caso es *cholesterol*) y el *test de Fligner* (no paramétrico) si no lo son (el resto de variables). Para ambos tests, la hipótesis nula asume igualdad de varianzas en los diferentes grupos de datos, por lo que un p-valor inferior al nivel de significancia indicará heterocedasticidad.

```
# Comprobación de homocedasticidad Cholesterol - Heart attack
LeveneTest(cholesterol ~ heart_attack,data = heart_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.4198 0.5175
##       301
```

```
# Comprobación de homocedasticidad Cholesterol - Sex
LeveneTest(cholesterol ~ sex,data = heart_data)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  4.5177 0.03436 *
##       301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Comprobación de homocedasticidad Age - Heart attack
fligner.test(age ~ heart_attack,data = heart_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  age by heart_attack
## Fligner-Killeen:med chi-squared = 7.2992, df = 1, p-value = 0.006898
```

```
# Comprobación de homocedasticidad resting_blood_pressure - Heart attack
fligner.test(resting_blood_pressure ~ heart_attack,data = heart_data)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  resting_blood_pressure by heart_attack
## Fligner-Killeen:med chi-squared = 1.367, df = 1, p-value = 0.2423
```

Las conclusiones de estos tests son las siguientes:

- La variable *cholesterol* presenta homocedasticidad con el hecho de si sufrieron un ataque al corazón y heterocedastidad con el sexo del paciente, es decir, la varianza variará entre los hombres y las mujeres y será similar cuando se tiene en cuenta si un paciente tiene un ataque o no.
- La variable *age* presenta heterocedastidad con la variable *heart_attack*, por lo que la varianza de la edad de los pacientes no será constante con el hecho de si un paciente sufre o no un ataque.
- La variable *resting_blood_pressure* presenta homocedasticidad para padecer o no un ataque cardiaco, concluyendo que la varianza de la presión arterial es similar entre padecer un ataque o no.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

4.3.1 Contrastes de hipótesis

Con los resultados anteriores, se van a aplicar varios tests estadísticos con la finalidad de responder a las preguntas planteadas al principio del enunciado.

En este caso, podemos aplicar tanto pruebas paramétricas como no paramétricas dado que tenemos variables normales y no normales.

Para el caso de la variable *cholesterol*, como ya es normal, y se ha visto que presenta homocedasticidad con la variable *heart_attack* (más adelante será la variable dependiente en los modelos de regresión logística), se va a aplicar la prueba de *t de student*, donde la hipótesis nula asume que las medias de los grupos de los datos son las mismas.

```
# Se aplica el test t de student cholesterol y heart_attack
t.test(cholesterol ~ heart_attack, data = heart_data)
```

```
##
##  Welch Two Sample t-test
##
## data:  cholesterol by heart_attack
## t = 1.8921, df = 287.59, p-value = 0.05948
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.002683574  0.136055304
## sample estimates:
##  mean in group No mean in group Yes
##         6.912523          6.845837
```

Viendo el resultado del test para el caso de la variable cholesterol con heart_attack, como el p-valor es mayor al nivel de significancia de 0.05, se puede observar que no hay diferencias estadísticamente significativas entre las medias de los grupos de datos de heart_attack. Esto se puede corroborar mediante el siguiente boxplot, donde toman valores parecidos de colesterol pacientes que sufren y que no sufren ataques al corazón.

```
# Diagrama de cajas Cholesterol - Heart attack
ggplot(heart_data,aes(x=heart_attack,y=cholesterol, fill=heart_attack)) +
  geom_boxplot() +
  # Barras de error
  stat_boxplot(geom = "errorbar",width = 0.25) + # Ancho
  # Etiqueta Eje x y leyenda
  scale_fill_hue(labels = c("NO", "YES"))+
  # Título del gráfico
  ggtitle ("Box Plot Cholesterol - Heart attack")+
  # Características del gráfico
    theme (plot.title = element_text(
        hjust = 0.5,
        size=rel(1.2),
        face="bold",
        color="black"))
```

Los siguientes tests a aplicar serán no parámetricos dado que las variables no son normales o no presentan homocedasticidad y por lo tanto no cumplen las supociones requeridas por los tests paramétricos. Se aplicará el *test de Wilcoxon o Mann-Whitney* (ambos se aplican igual con la misma función *wilcox.test*) donde la hipótesis nula asume igualdad de distribución para los diferentes grupos de la variable categórica.

```
# Se aplica el test no paramétrico con el resto de variables

# cholesterol vs sex
wilcox.test(cholesterol ~ sex, data = heart_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  cholesterol by sex
## W = 11715, p-value = 0.01219
## alternative hypothesis: true location shift is not equal to 0
```

```
# age vs heart_attack
wilcox.test(age ~ heart_attack, data = heart_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  age by heart_attack
## W = 14530, p-value = 3.439e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
# resting_blood_pressure vs heart_attack
wilcox.test(resting_blood_pressure ~ heart_attack, data = heart_data)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  resting_blood_pressure by heart_attack
## W = 12986, p-value = 0.03465
## alternative hypothesis: true location shift is not equal to 0
```

En los 3 casos se puede ver que no se puede determinar que la distribución de las variables sea la misma en los diferentes grupos, tanto de la variable *heart_attack* como de *sex*.

Otro test que se va a realizar va a ser el de χ^2 para comprobar si existen diferencias significativas entre las variables categóricas *heart_attack* y *sex*, entre *fasting_blood_sugar* y *sex*, entre *fasting_blood_sugar* y *heart_attack*, y entre *chest_pain_type* y *heart_attack*. La hipótesis nula que asume es que no existen diferencias significativas entre los grupos de ambas variables.

```
# Se comprueba la proporción de hombres y mujeres que sufrieron un ataque
table(heart_data$sex, heart_data$heart_attack)
```

```
##
##              No Yes
##   Femenino   24  72
##   Masculino 114  93
```

```
chisq.test(table(heart_data$sex, heart_data$heart_attack))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(heart_data$sex, heart_data$heart_attack)
## X-squared = 22.717, df = 1, p-value = 1.877e-06
```

```
# Tuvo alguna influencia el nivel de azúcar en sangre
table(heart_data$fasting_blood_sugar, heart_data$heart_attack)
```

```
##
##                No Yes
##   Azúcar Bajo 116 142
##   Azúcar Alto  22  23
```

```
chisq.test(table(heart_data$fasting_blood_sugar, heart_data$heart_attack))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table(heart_data$fasting_blood_sugar, heart_data$heart_attack)
## X-squared = 0.10627, df = 1, p-value = 0.7444
```

```
# Tiene alguna relacion el nivel de azúcar en sangre con el sexo del paciente
table(heart_data$fasting_blood_sugar, heart_data$sex)
```

```
##
##                Femenino Masculino
##   Azúcar Bajo        84       174
##   Azúcar Alto        12        33
```

```
chisq.test(table(heart_data$fasting_blood_sugar, heart_data$sex))
```

```
##
##  Pearson's Chi-squared test with Yates' continuity correction
```

```
##
## data:  table(heart_data$fasting_blood_sugar, heart_data$sex)
## X-squared = 0.3724, df = 1, p-value = 0.5417
```

```
# Tuvo alguna influencia el tipo de dolor de pecho
table(heart_data$chest_pain_type, heart_data$heart_attack)
```

```
##
##                      No Yes
##   Angina típica     104  39
##   Angina atípica      9  41
##   Dolor no anginoso  18  69
##   Asintomático        7  16
```

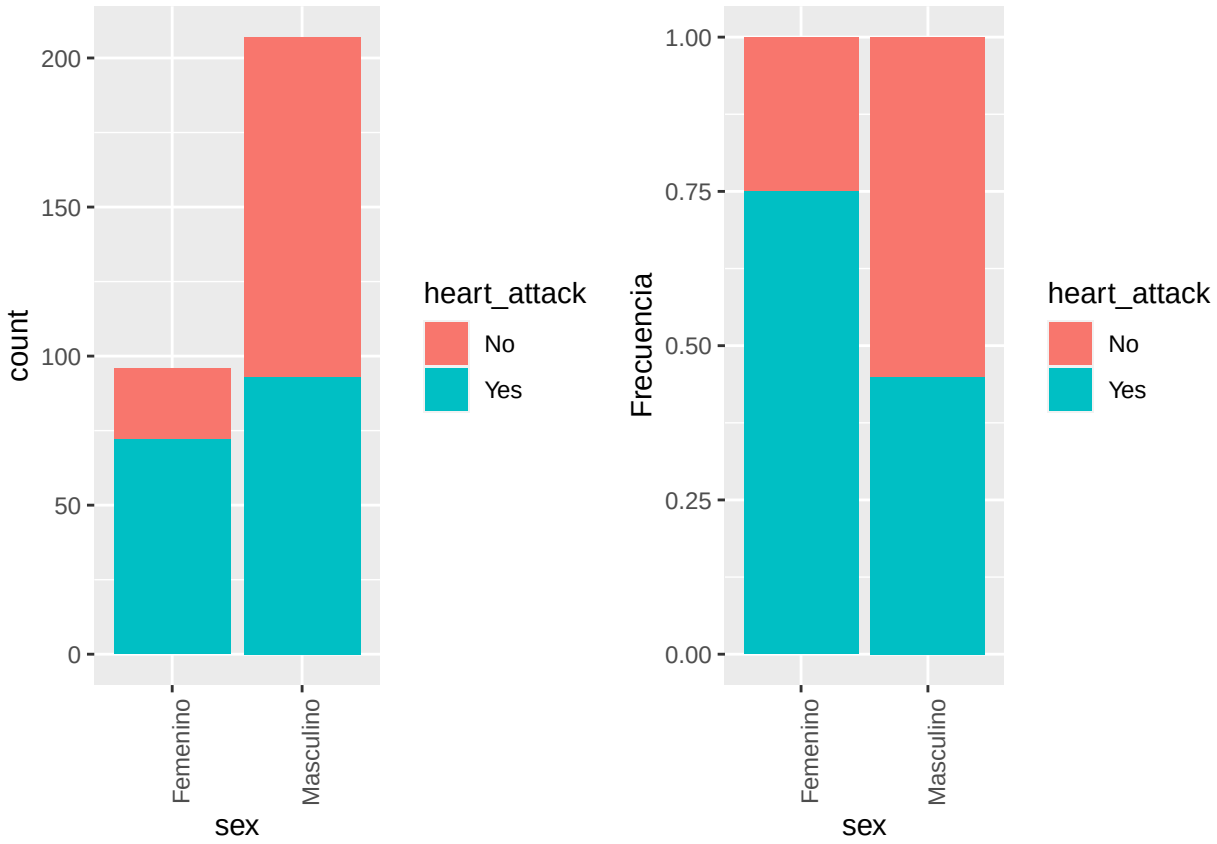
```
chisq.test(table(heart_data$chest_pain_type, heart_data$heart_attack))
```

```
##
##  Pearson's Chi-squared test
##
## data:  table(heart_data$chest_pain_type, heart_data$heart_attack)
## X-squared = 81.686, df = 3, p-value < 2.2e-16
```

Viendo los resultados podemos decir:

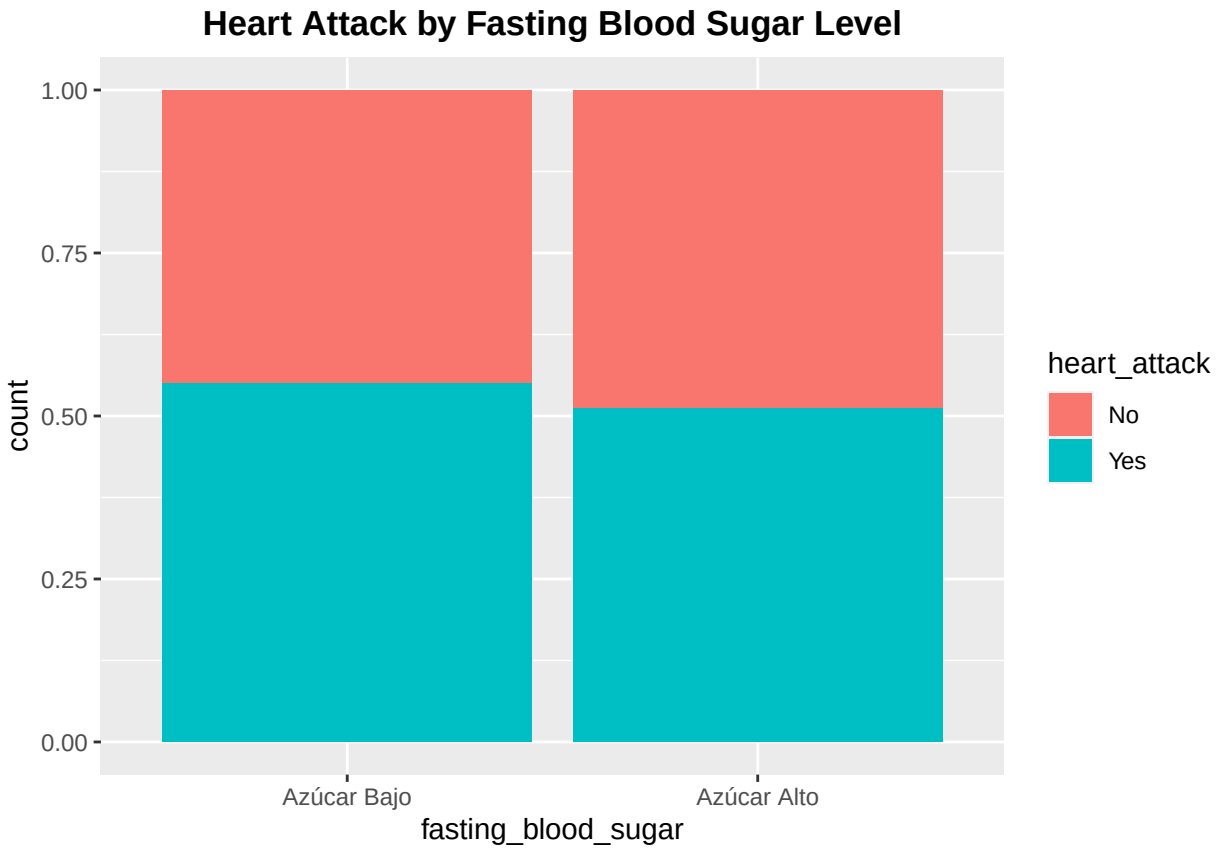
- El hecho de ser hombre o mujer y el tipo de dolor de pecho muestra diferencias significativas con padecer un ataque puesto que no se cumple la hipótesis nula, por lo tanto el sexo y el tipo de dolor de pecho son variables que repercuten a la hora de sufrir un ataque al corazón, siendo dependientes con la variable heart_attack. Véase en el siguiente gráfico que el sexo femenino es más propenso a sufrir ataques al corazón.

```
# Dibujar gráfico de barras
g1 <- ggplot(data = heart_data, aes(x=sex, fill=heart_attack))+ geom_bar()  +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
g2 <- ggplot(data = heart_data, aes(x=sex, fill=heart_attack)) +
  geom_bar(binwidth = 3,position="fill") + ylab("Frecuencia")  + theme(axis.text.x = element_text(angle
grid.arrange(g1, g2, nrow = 1)
```



- No hay una diferencia significativa en la proporción de personas con azúcar en sangre baja o alta que han padecido un ataque al corazón, ya que el valor p del test de chi-cuadrado es muy grande (p-value = 0.7444). Véase en el siguiente gráfico donde se observa que no existe una diferencia destacable.

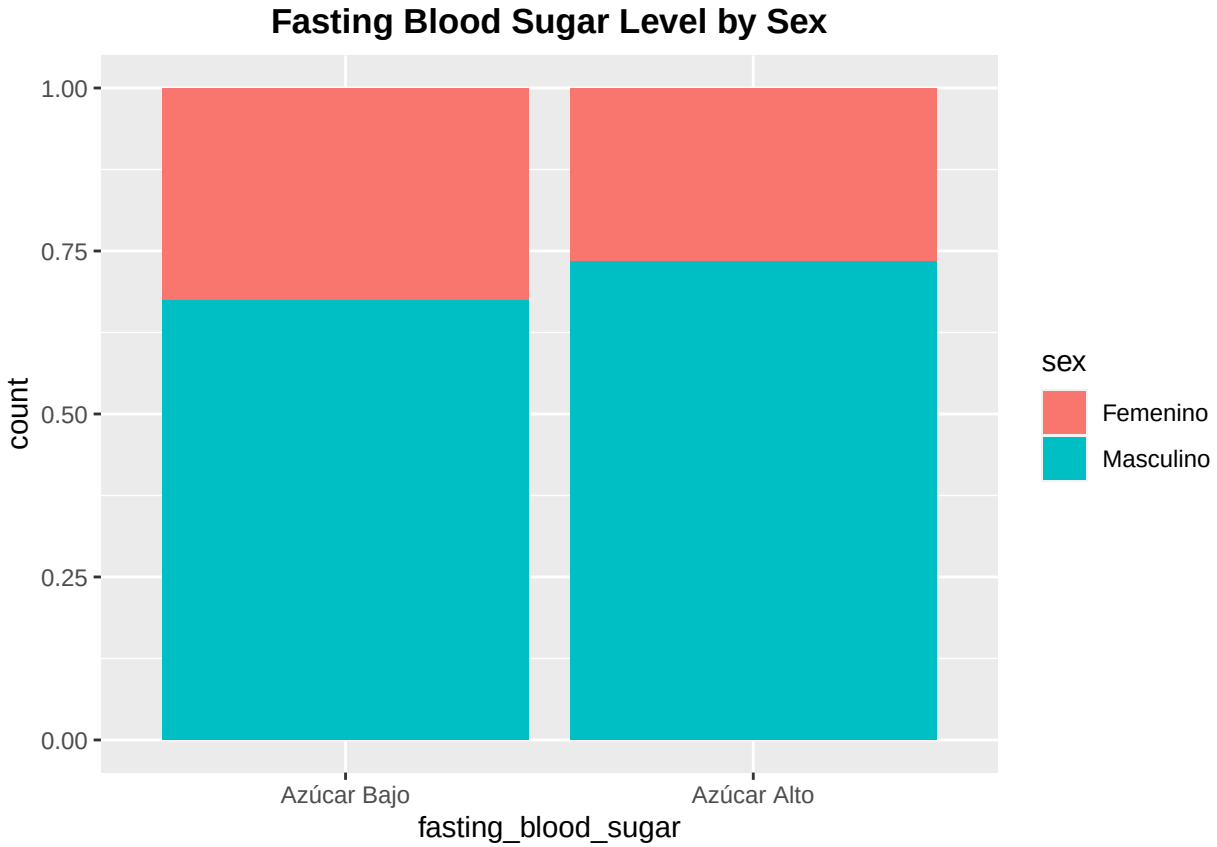
```
# Dibujar gráfico de barras apiladas
ggplot(heart_data, aes(x=fasting_blood_sugar, fill=heart_attack)) +
  geom_bar(position="fill") +
  ggtitle("Heart Attack by Fasting Blood Sugar Level") +
  theme (plot.title = element_text(
    hjust = 0.5,
    size=rel(1.2),
    face="bold",
    color="black"))
```



- No hay una diferencia significativa en la proporción de personas de diferentes sexos con azúcar en sangre baja o alta, ya que el valor p del test de chi-cuadrado es muy grande (p-value = 0.5417). Véase en el siguiente gráfico donde se observa que no existe una diferencia destacable.

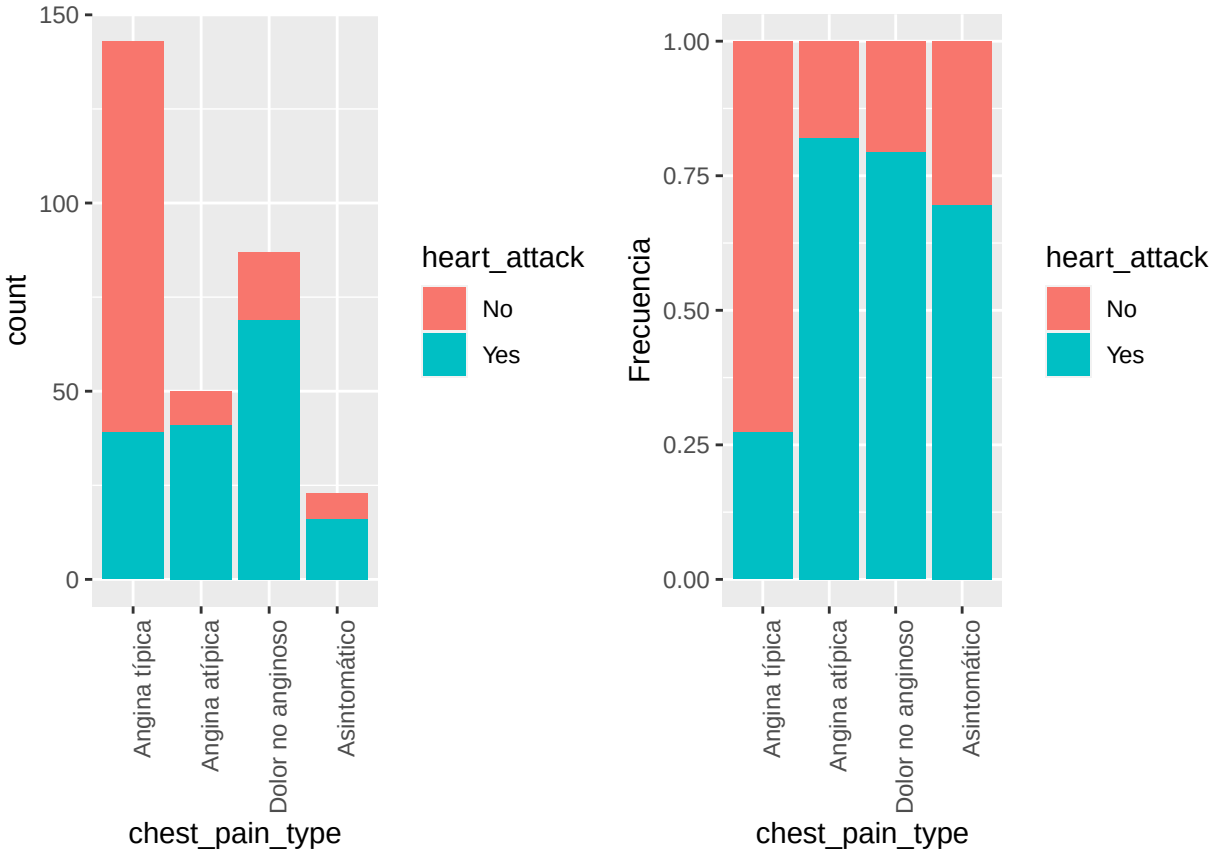
```
# Crear tabla de contingencia
sugar_sex_table <- table(heart_data$fasting_blood_sugar, heart_data$sex)

# Dibujar gráfico de barras apiladas
ggplot(heart_data, aes(x=fasting_blood_sugar, fill=sex)) +
  geom_bar(position="fill") +
  ggtitle("Fasting Blood Sugar Level by Sex") +
  theme (plot.title = element_text(
    hjust = 0.5,
    size=rel(1.2),
    face="bold",
    color="black"))
```



- Hay una diferencia significativa en la proporción de personas con diferentes tipos de dolor de pecho que han padecido un ataque al corazón, ya que el valor p del test de chi-cuadrado es muy pequeño (p-value $<$ 2.2e-16). El siguiente gráfico demuestra que sufrir un dolor en el pecho como angina atípica, dolor no anginoso o asintomático puede repercutir en un ataque del corazón con una alta probabilidad. Por tanto el dolor en el pecho tiene una importancia significativa para los casos de ataque al corazón.

```
g1 <- ggplot(data = heart_data, aes(x=chest_pain_type, fill=heart_attack)) + geom_bar() +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
g2 <- ggplot(data = heart_data, aes(x=chest_pain_type, fill=heart_attack)) +
  geom_bar(binwidth = 3,position="fill") +
ylab("Frecuencia") +
theme(axis.text.x = element_text(angle = 90, hjust = 1))
grid.arrange(g1, g2, nrow = 1)
```



4.3.2 Modelos de regresión logística

En este apartado se van a construir varios modelos de regresión logística para analizar la influencia de algunas de las variables de forma que se pueda ver cuáles son las más significativas a la hora de determinar si un paciente sufre o no un ataque al corazón. De esta forma, sabremos la relación existente entre los diferentes atributos sobre la variable dicotómica dependiente *heart_attack*. Además, se calcularán las odds-ratio y se interpretarán junto con los coeficientes del modelo, de esta forma sabremos si la probabilidad del suceso de la variable dependiente va a aumentar o disminuir según el signo de estos coeficientes.

```
# Se estima el modelo de regresión logística
model_rg_1 <- glm(formula=heart_attack~.,data=heart_data,
                  family=binomial(link=logit))
summary(model_rg_1)
```

```
##
## Call:
## glm(formula = heart_attack ~ ., family = binomial(link = logit),
##     data = heart_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7435  -0.3504   0.1582   0.5201   2.6276
##
## Coefficients:
##                                                 Estimate Std. Error z value
## (Intercept)                                    7.0473730  5.4589594   1.291
## age                                           -0.0007433  0.0236370  -0.031
```

```
## sexMasculino                                      -1.5086509  0.5126370  -2.943
## chest_pain_typeAngina atípica                      1.0070109  0.5665294   1.778
## chest_pain_typeDolor no anginoso                   1.8866144  0.4787523   3.941
## chest_pain_typeAsintomático                        1.9995462  0.6520120   3.067
## resting_blood_pressure                            -0.0158789  0.0107826  -1.473
## cholesterol                                       -1.0509816  0.7151901  -1.470
## fasting_blood_sugarAzúcar Alto                     0.2106506  0.5712255   0.369
## rest_ecg_typeAnomalía de onda ST-T                 0.5609417  0.3738693   1.500
## rest_ecg_typeHipertrofia ventricular izquierda -0.3145765  2.3123159  -0.136
## max_heart_rate_achieved                            0.0171244  0.0107101   1.599
## exercise_induced_anginaSí                         -0.7494929  0.4275154  -1.753
## st_depression                                     -0.4926132  0.2257259  -2.182
## st_slope_typeNormal                               -0.6999261  0.8574112  -0.816
## st_slope_typeAlta                                  0.2085676  0.9321546   0.224
## num_major_vessels                                 -0.8357204  0.2063651  -4.050
## thalassemia_typeFijo                               1.8181465  2.3174120   0.785
## thalassemia_typeNormal                             1.9328592  2.2299344   0.867
## thalassemia_typeReversible                         0.5162475  2.2384356   0.231
##                                                  Pr(>|z|)
## (Intercept)                                       0.19671
## age                                               0.97492
## sexMasculino                                      0.00325 **
## chest_pain_typeAngina atípica                     0.07548 .
## chest_pain_typeDolor no anginoso                 8.12e-05 ***
## chest_pain_typeAsintomático                       0.00216 **
## resting_blood_pressure                            0.14085
## cholesterol                                       0.14169
## fasting_blood_sugarAzúcar Alto                    0.71230
## rest_ecg_typeAnomalía de onda ST-T                0.13352
## rest_ecg_typeHipertrofia ventricular izquierda    0.89179
## max_heart_rate_achieved                           0.10984
## exercise_induced_anginaSí                         0.07958 .
## st_depression                                     0.02908 *
## st_slope_typeNormal                               0.41431
## st_slope_typeAlta                                 0.82295
## num_major_vessels                                5.13e-05 ***
## thalassemia_typeFijo                              0.43271
## thalassemia_typeNormal                            0.38606
## thalassemia_typeReversible                        0.81760
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 200.70  on 283  degrees of freedom
## AIC: 240.7
##
## Number of Fisher Scoring iterations: 6
```

Se puede observar como las variables más significativas son *num_major_vessels*, *chest_pain_type*, *sex* y *st_depression*, tal y como vimos aplicando el test de chi cuadrado para el caso de *chest_pain_type* y *sex*, por lo tanto, será sobre estas variables sobre las que se centrará este análisis.

A continuación, se estiman otros modelos de regresión con la combinación de las variables regresoras anteri-

ores para ver cómo afectan a la variable dependiente heart_attack y se calculan sus valores AIC para poder compararlos.

```
# Se estima varios modelos de regresión logística
model_rg_2 <- glm(formula=heart_attack~chest_pain_type,data=heart_data,
                  family=binomial(link=logit))

model_rg_3 <- glm(formula=heart_attack~chest_pain_type + num_major_vessels,
                  data=heart_data,
                  family=binomial(link=logit))

model_rg_4 <- glm(formula=heart_attack~chest_pain_type + num_major_vessels + sex ,
                  data=heart_data,
                  family=binomial(link=logit))

model_rg_5 <- glm(formula=heart_attack~chest_pain_type + num_major_vessels + sex +
                    st_depression ,data=heart_data,
                  family=binomial(link=logit))

# Guardamos los valores de una tabla
indices_AIC <- data.frame( c(2:5), c(model_rg_2$aic,model_rg_3$aic,model_rg_4$aic,
                                     model_rg_5$aic))
colnames(indices_AIC) <- c("Modelo", "AIC")

# Se muestran en una tabla los resultados de los valores AIC de cada modelo
indices_AIC %>% kable() %>% kable_styling(latex_options = "hold_position")
```

Modelo	AIC
2	339.6969
3	307.9877
4	291.5398
5	263.2847

Comparando el valor AIC de cada modelo (aquel que relaciona su bondad de ajuste junto con su complejidad) a medida que se han ido añadiendo variables regresoras, se puede ver que ha ido disminuyendo y por lo tanto han ido mejorando los modelos, es decir, todas ellas son significativas para el hecho de sufrir un ataque al corazón.

Por lo tanto, nos quedamos con el modelo *model_rg_5*.

```
summary(model_rg_5)
```

```
##
## Call:
## glm(formula = heart_attack ~ chest_pain_type + num_major_vessels +
##     sex + st_depression, family = binomial(link = logit), data = heart_data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2542  -0.5818   0.2233   0.5885   2.2965
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       1.4346     0.3878   3.700 0.000216 ***
## chest_pain_typeAngina atípica     1.7514     0.4645   3.771 0.000163 ***
## chest_pain_typeDolor no anginoso  2.4184     0.4068   5.945 2.77e-09 ***
```

```
## chest_pain_typeAsintomático        2.3194     0.5837   3.974 7.07e-05 ***
## num_major_vessels                 -0.7361     0.1638  -4.494 7.00e-06 ***
## sexMasculino                      -1.3944     0.3770  -3.699 0.000217 ***
## st_depression                     -0.8677     0.1755  -4.944 7.64e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 417.64  on 302  degrees of freedom
## Residual deviance: 249.28  on 296  degrees of freedom
## AIC: 263.28
##
## Number of Fisher Scoring iterations: 5
```

Se puede ver como todas las variables regresoras son estadísticamente significativas ya que $Pr(>|z|) < 0.05$ y tienen una repercusión fuerte en la variable *heart_attack*.

Si calculamos sus odds ratio obtenemos lo siguiente:

```
# Cálculo de las Odds-Ratio
exp(coefficients(model_rg_5))
```

```
##                       (Intercept)    chest_pain_typeAngina atípica
##                         4.1980325                        5.7628558
## chest_pain_typeDolor no anginoso      chest_pain_typeAsintomático
##                        11.2283857                       10.1698527
##                num_major_vessels                      sexMasculino
##                         0.4789933                        0.2479817
##                     st_depression
##                         0.4199293
```

Si comentamos los resultados de los coeficientes de los regresores y sus odds-ratio, para el caso de la variable *st_depression* que se ha obtenido un coeficiente estimado negativo y una odds-Ratio de 0.41, va a indicar que por cada unidad que aumente la variable, la probabilidad de sufrir un ataque es 0.41 veces menor. Para la variable *num_major_vessels*, con una odds-Ratio de 0.47 y un coeficiente estimado negativo, se interpreta de forma que cuántos más vasos principales tenga el paciente, la probabilidad de sufrir un ataque es 0.47 veces menor.

Para la variable categórica, *chest_pain_type*, obteniendo varios coeficientes estimados positivos respecto al nivel de referencia *angina típica*, y unas odds-ratio de 10.15, 5.75, 11.11, nos indican que la probabilidad para que un paciente sufra un ataque con el resto de tipos de anginas de pecho comparado con una angina típica son de 10.15, 5.75, 11.11 respectivamente veces mayor.

Por último, para la variable *sex* obteniendo un coeficiente negativo respecto al nivel de referencia *femenino*, y una odd-ratio de 0.25, nos muestra que la probabilidad de que un paciente hombre sufra un ataque comparado con una paciente mujer es 0.25 veces menor.

En definitiva, la probabilidad para que el paciente sufra un ataque al corazón aumenta teniendo dolor de pecho asintomático, angina atípica y dolor no anginoso, mientras que disminuye siendo hombre, a mayor número de vasos principales y con la depresión ST inducida por el ejercicio en relación con el descanso.

4.4 Generación del archivo con los datos tratados

En este punto se genera el fichero con los datos tratados y limpiados tal y como se pide en la práctica. Se exporta con el nombre `clean_data.csv`.

```r
# Dataframe tratado
df_heart_final <- heart_data
# Se exporta a formato csv
write.csv(df_heart_final, file = "clean_data.csv", row.names = FALSE, col.names = TRUE)
```

5 Representación de los resultados a partir de tablas y gráficas.

La representación de los resultados a partir de tablas y gráficas ha sido realizada durante el desarrollo de la práctica para poder situarlos en su contexto específico.

6 Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

A continuación se van a extraer las conclusiones observadas durante el desarrollo de este proyecto. Los resultados nos han permitido extraer muchas conclusiones y dar respuestas a nuestras preguntas iniciales.

- **¿Los hombres son más probables a sufrir un ataque que las mujeres?**

Los hombres son menos probables a sufrir un ataque que las mujeres. Es decir, el sexo si que tiene importancia en cuanto a los ataques del corazón, se ha demostrado que las mujeres sufren muchos más ataques al corazón que los hombres.

- **¿El nivel de azúcar en sangre es determinante para que una persona pueda padecer un ataque?**

El nivel de azúcar en sangre no es determinante para que una persona pueda padecer un ataque, ya que no se encontraron diferencias significativas entre esta variable y el padecimiento de un ataque.

- **¿Hay diferencias significativas en el nivel de colesterol según padezca o no un ataque?**

Curiosamente el nivel de colesterol no tiene un impacto directo en que una persona sufra un ataque o no lo sufra tal y como se ha demostrado en este análisis y mediante su visualización gráfica.

- **¿Hubo algún indicio de sufrir más fácilmente un ataque al corazón según el dolor de pecho del paciente?**

Sí hubo indicios de que el dolor de pecho del paciente es un factor determinante para sufrir un ataque, ya que se encontraron diferencias significativas entre los tipos de dolor de pecho y el padecimiento de un ataque.

- **¿Qué factores son los más influyentes para sufrir un ataque?**

Los factores más influyentes para sufrir un ataque, son el tipo de dolor de pecho, el número de vasos principales y el sexo del paciente.

7 Código.

Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código de la práctica en el lenguaje R se encuentra en el repositorio de GitHub: https://github.com/LucasGomezTorres/Limpieza_Analisis_Datos.

8 Vídeo.

El vídeo realizado para detallar el proyecto se encuentra en los siguientes enlaces, tanto de Google Drive como Youtube:

1. Vídeo en Google Drive de la UOC: https://drive.google.com/file/d/1tz2x5cRMS535CEUYc3imyvPUGqwW04fc/view?usp=sharing
2. Vídeo en Youtube: https://youtu.be/zQspF4SBH-k

9 Contribuciones de los integrantes

A continuación se presentan las contribuciones del proyecto firmadas por los integrantes del grupo.

Contribución	Firma
Investigación previa	Lucas Gómez, Joan Amengual
Redacción de las respuestas	Lucas Gómez, Joan Amengual
Desarrollo del código	Lucas Gómez, Joan Amengual
Participación en el vídeo	Lucas Gómez, Joan Amengual

El trabajo ha sido dividido de forma modular para que cada integrante pueda recoger parte de las tareas a resolver y enfocarse de manera individual sacar el trabajo adelante. Continuamente se han ido realizando reuniones de grupo (mediante Zoom) para comentar todos los puntos importantes y tomar las decisiones necesarias en cuanto a las tareas que se han necesitado realizar.