Práctica 2

Autor: Christian Rolando Oyola Flores

Tipología y ciclo de vida de los datos - Semestre 2023

Contents

---

1 Introducción

1.1 Presentación

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de los mismos

1.2 Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.
- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos

1.3 Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

---

2 Planteamiento de objetivos analíticos y descripción detallada de la metodología para su resolución

---

El propósito fundamental de este desarrollo es examinar e identificar las variables de riesgo asociadas con problemas cardíacos en una población específica. La identificación de estos factores es crucial para desarrollar modelos predictivos que puedan contribuir en la detección temprana y en la generación de alertas sobre problemas cardíacos. En este contexto, se plantea suministrar información crucial a las instituciones de salud pública sobre los factores relacionados con problemas cardíacos. Esta información está destinada a facilitar la creación de políticas o programas preventivos en el ámbito de la salud, enfocados en mitigar y manejar eficazmente esta problemática.

En este contexto, el análisis estadístico se centrará en el examen de datos históricos de pacientes que hayan experimentado cuadros clínicos asociados a problemas cardíacos. Esto permitirá identificar perfiles de riesgo basándose en varias dimensiones de interés: demográficas, análisis clínicos y clasificación de riesgos. Cada

una de estas dimensiones aportará información vital para comprender mejor los factores que contribuyen a estos problemas de salud y facilitar así la toma de decisiones informadas.

La metodología para abordar el objetivo planteado incluye los siguientes procesos:

a. Análisis exploratorio de datos:

- Se llevará a cabo un análisis inicial de las variables definidas en el conjunto de datos, incluyendo su tipo (categórica, numérica, binaria), y su distribución.
- Identificar y tratar los valores faltantes o datos atípicos que puedan afectar al estudio.
- En esta misma línea, se llevará a cabo un proceso de identificación de las relaciones entre las variables, apoyados sobre gráficos y técnicas estadísticas para obtener una comprensión optima del conjunto de datos.

b. Preprocesamiento de datos:

- Dentro de esta etapa se llevará a cabo una codificación (según corresponda) de las variables categóricas.
- Así también, es relevante para el caso de estudio escalar o estandarizar las variables numéricas, esto con el objetivo de asegurar que todas las variables tengan un rango comparable.

2.1 Selección del juego de datos

2.1.1 Justificación de elección del set de datos

La elección del conjunto de datos Heart Attack Analysis & Prediction Dataset del repositorio kaggle, se basa en su idoneidad para aplicar una variedad de técnicas de análisis y exploración de informacíón para llevar a cabo un proyecto analítico,

Este único conjunto de datos contiene información sobre factores vinculados con ataques cardiacos en una población, vinculando adicionalmente variables numéricas y categóricas, así momo, la vinculación de variables de salida para la predicción de pacientes que puede presentar esta problemática.

Con una variable objetivo (target) claramente definida y un conjunto diverso de variables predictoras relacionadas con los pacientes, este conjunto de datos se presta para un análisis más exhaustivo. Este análisis podría incluir la aplicación de algoritmos supervisados de clasificación y regresión, orientados a predecir y entender los factores que inciden en los ataques cardíacos.

2.1.2 Descripción del dataset

La selección de este dataset permitirá llevar a cabo un análisis de factores que contribuyen a los ataques cardíacos, este conjunto de datos puede ayudar a desarrollar modelos predictivos, partiendo de una identificación inicial de input's o variables vinculadas. Estos modelos podrían ser utilizados por los profesionales de la salud para identificar a los individuos con alto riesgo de sufrir un ataque cardíaco, permitiendo llevar a cabo intervenciones preventivas y diagnósticos tempranos. Las preguntas o problemas específicos que se pretende responder o abordar con este conjunto de datos incluyen:

- ¿Cuales son lor principales factores predictivos de un ataque cardiaco?
- ¿Cómo afectan el ejercicio y la dieta al riesgo de ataques cardíacos?
- ¿Cómo difieren los síntomas y factores de riesgo entre hombres y mujeres?
- ¿Cómo cambia el riesgo y la presentación de enfermedades cardíacas con la edad?

2.1.3 Integración y selección de los datos de interes

Para el desarrollo de esta práctica, se integrarán las 14 variables presentes en el conjunto de datos original, utilizando así todas las características disponibles. Para facilitar este proceso, crearemos un diccionario de datos basándonos en la documentación auxiliar proporcionada. Este diccionario será una herramienta clave para mejorar la comprensión de la importancia y el significado de cada variable en el contexto del estudio constituido en 3 dimensiones de análisis:

Dimensión Demográfica:

- **age** - edad en años
- **sex** - sexo (1 = masculino; 0 = femenino)

Dimensión Análisis Clinico:

- **cp** - tipo de dolor torácico (1 = angina típica; 2 = angina atípica; 3 = dolor no anginoso; 0 = asintomático)
- **trestbps** - presión arterial en reposo (en mm Hg al ingreso al hospital)
- **chol** - colesterol sérico en mg/dl
- **fbs** - glucosa en sangre en ayunas > 120 mg/dl (1 = cierto; 0 = falso)
- **restecg** - resultados electrocardiográficos en reposo (1 = normal; 2 = anormalidad de onda ST-T; 0 = hipertrofia)
- **thalachh** - máxima frecuencia cardíaca alcanzada
- **exng** - angina inducida por ejercicio (1 = sí; 0 = no)
- **oldpeak** - depresión del segmento ST inducida por ejercicio relativo al reposo
- **slp** - la pendiente del segmento ST del ejercicio máximo (2 = ascendente; 1 = plano; 0 = descendente)
- **caa** - número de vasos principales (0-3) coloreados por fluoroscopia
- **thall** - 2 = normal; 1 = defecto fijo; 3 = defecto reversible

Dimensión Clasificación:

- **output: 0** = menor probabilidad de ataque al corazón 1 = mayor probabilidad de ataque al corazón

Después de definir el diccionario y especificar los tipos de datos, procederemos a realizar un análisis más detallado de los datos que conforman dicho diccionario.

2.2 Limpieza de datos

Para comenzar el análisis, importamos el conjunto de datos en una estructura de datos adecuada utilizando bibliotecas especializadas en manipulación y análisis de datos.

Cargamos los datos de un directorio local:

```
heart_csv <- read_delim("Data/heart.csv", #lectura del archivo empleando la función read_delim()
            delim = ",", #especificación del delimitador en la lectura del archivo.
            escape_double = FALSE, #no se deben "escapar" las comillas dobles presentes en el archivo.
            trim_ws = TRUE) #se deben eliminar los espacios en blanco de los valores en el archivo
```

2.2.1 Inspección inicial de datos

Realizamos una inspección inicial de los datos con el fin de obtener una visión general de la estructura y el formato de los atributos. Durante esta etapa, verificamos el número de filas y columnas, revisamos los nombres de las columnas y observamos los primeros registros del DataFrame. Este proceso nos permite familiarizarnos con los datos y nos ayuda a considerar lo siguiente:

- **¿Qué representa cada atributo?**
- **¿Qué tipo de dato debería tener cada columna?**
- **¿Que granularidad o atomicidad tiene la data?**

a) Mostramos la estructura del conjunto de datos:

```
data_heart <- heart_csv
glimpse(data_heart) # vista resumida de la estructura del dataset
```

```
## Rows: 303
## Columns: 14
## $ age      <dbl> 63, 37, 41, 56, 57, 57, 56, 44, 52, 57, 54, 48, 49, 64, 58, 5~
## $ sex      <dbl> 1, 1, 0, 1, 0, 1, 0, 1, 1, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 0, 1~
## $ cp       <dbl> 3, 2, 1, 1, 0, 0, 1, 1, 2, 2, 0, 2, 1, 3, 3, 2, 2, 3, 0, 3, 0~
## $ trtbps   <dbl> 145, 130, 130, 120, 120, 140, 140, 120, 172, 150, 140, 130, 1~
## $ chol     <dbl> 233, 250, 204, 236, 354, 192, 294, 263, 199, 168, 239, 275, 2~
## $ fbs      <dbl> 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0~
## $ restecg  <dbl> 0, 1, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, 0, 1, 1, 1, 1, 1, 1~
## $ thalachh <dbl> 150, 187, 172, 178, 163, 148, 153, 173, 162, 174, 160, 139, 1~
## $ exng     <dbl> 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0~
## $ oldpeak  <dbl> 2.3, 3.5, 1.4, 0.8, 0.6, 0.4, 1.3, 0.0, 0.5, 1.6, 1.2, 0.2, 0~
## $ slp      <dbl> 0, 0, 2, 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 1, 2, 1, 2, 0, 2, 2, 1~
## $ caa      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 0~
## $ thall    <dbl> 1, 2, 2, 2, 2, 1, 2, 3, 3, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 3~
## $ output   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
```

Se puede evidenciar que para este dataset se cuenta con un total de **303** registros y **14** variables o caracteristicas. Estos datos nos proporcionarán la información necesaria para respaldar nuestro modelo de predicción.

2.3 Preprocesamiento y gestión de características

2.3.1 Valores faltantes

El siguiente paso será la limpieza de datos, identificando la presencia de valores vacíos o nulos.

```
colSums(is.na(data_heart))
```

```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

```r
colSums(data_heart=="")
```

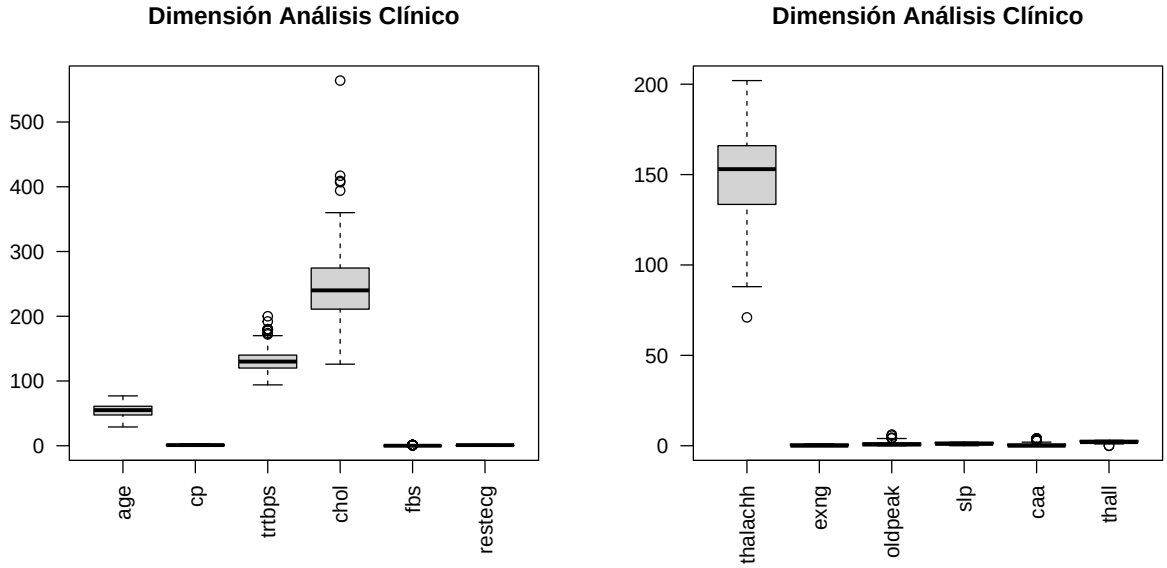
```
##      age      sex       cp   trtbps     chol      fbs  restecg thalachh
##        0        0        0        0        0        0        0        0
##     exng  oldpeak      slp      caa    thall   output
##        0        0        0        0        0        0
```

En un primer análisis se puede identificar que en el conjunto de datos, todas las columnas son continuas, no hay columnas completamente vacías, y todas las filas están completas, es decir, sin observaciones faltantes.

2.3.2 Valores Outliers

Dentro del proceso de limpieza de datos, se efectúa un análisis de los valores extremos en las características. El objetivo es identificar y tratar adecuadamente estos valores para evitar sesgos en el análisis. Este paso es crucial, ya que los valores extremos pueden distorsionar los resultados y afectar la precisión de los modelos, para ello usaremos una representación mediante diagrama de cajas:

```r
# Ajustar el layout para tener 1 fila y 2 columnas
par(mfrow=c(1, 2), las=2)
# Primer boxplot
boxplot(data_heart[, c("age", "cp", "trtbps", "chol", "fbs", "restecg")],
        main="Dimensión Análisis Clínico")
# Segundo boxplot
boxplot(data_heart[, c("thalachh", "exng", "oldpeak", "slp", "caa", "thall")],
        main="Dimensión Análisis Clínico")
```

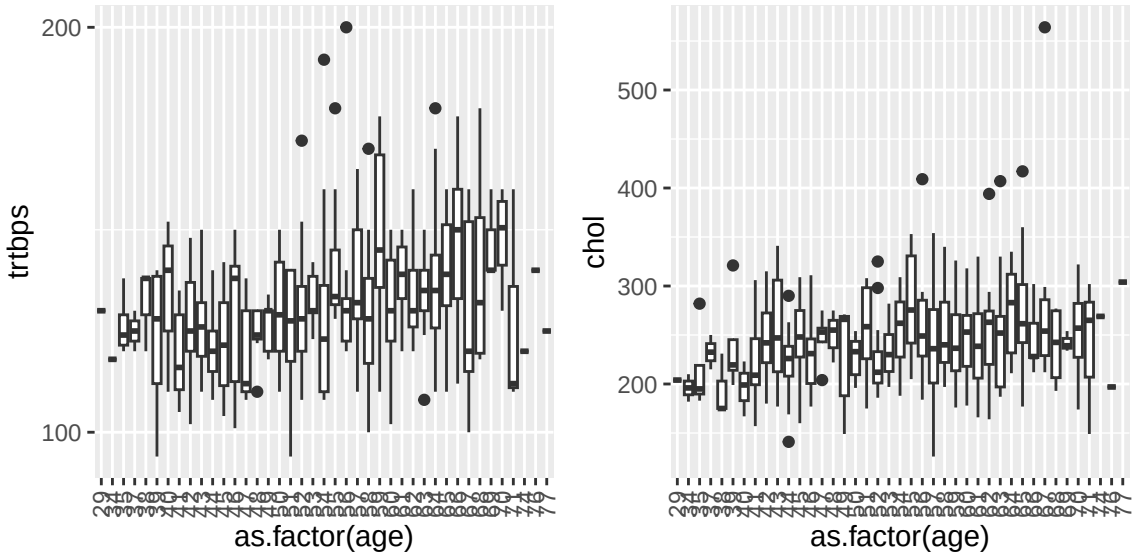


La observación gráfica inicial permite evidenciar la presencia de valores extremos en las variables trtbps, chol, thalachh y oldpeak. Para complementar y confirmar estos hallazgos, es esencial acceder a los detalles estadísticos referentes a los boxplots, que incluyen los cuartiles, la mediana y los valores extremos específicos.

```r
chol_outliers <- boxplot.stats(data_heart$chol)$out # valores atípicos para 'chol'
trtbps_outliers <- boxplot.stats(data_heart$trtbps)$out #valores atípicos para 'trtbps'
thalachh_outliers <- boxplot.stats(data_heart$thalachh)$out #valores atípicos para 'thalachh'
oldpeak_outliers <- boxplot.stats(data_heart$oldpeak)$out
caa_outliers <- boxplot.stats(data_heart$caa)$out
outliers_list <- list(
  Cholesterol_Outliers = chol_outliers,
  Resting_BP_Outliers = trtbps_outliers,
  Max_HR_Outliers = thalachh_outliers,
  oldpeak = oldpeak_outliers,
  Max_HR_Outliers = caa_outliers
)
```

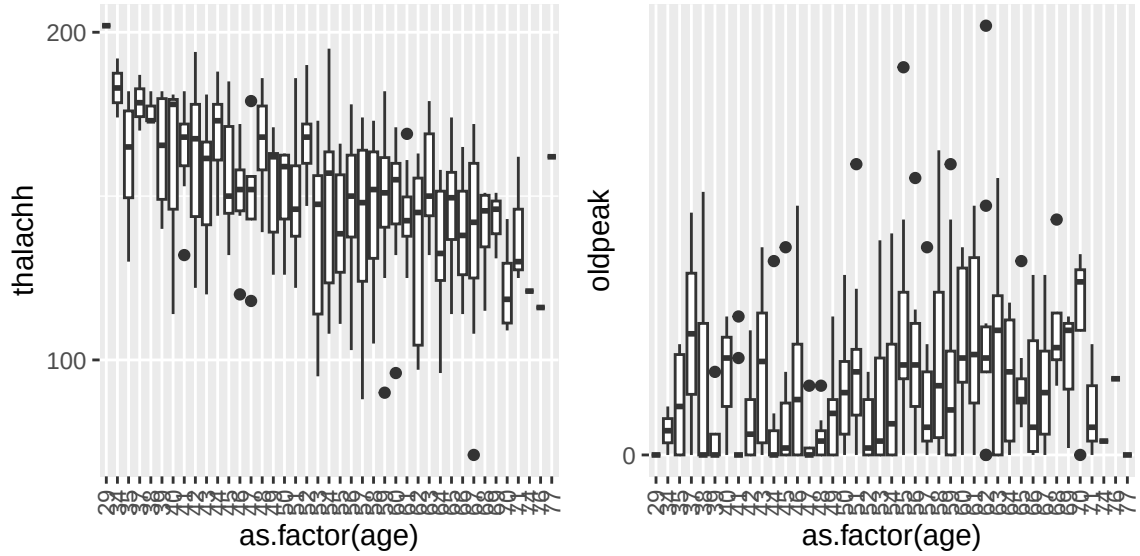
Se analiza la influencia de los outliers en relación con la edad de los pacientes para tomar un criterio más acertado de eliminación o no de los valores extremos.

```r
# Crear el primer boxplot para 'trtbps' en función de 'age'
bp1 <- ggplot(data = data_heart, aes(x = as.factor(age), y = trtbps)) +
  geom_boxplot() + scale_y_continuous(breaks = seq(from = 0, to = 1500, by = 100)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
# Crear el segundo boxplot para 'chol' en función de 'age'
bp2 <- ggplot(data = data_heart, aes(x = as.factor(age), y = chol)) +
  geom_boxplot() + scale_y_continuous(breaks = seq(from = 0, to = 1500, by = 100)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
grid.arrange(bp1, bp2, ncol = 2)
```




```r
# Crear el primer boxplot para 'thalachh' en función de 'age'
bp1 <- ggplot(data = data_heart, aes(x = as.factor(age), y = thalachh)) +
  geom_boxplot() + scale_y_continuous(breaks = seq(from = 0, to = 1500, by = 100)) +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
# Crear el segundo boxplot para 'oldpeak' en función de 'age'
bp2 <- ggplot(data = data_heart, aes(x = as.factor(age), y = oldpeak)) +
  geom_boxplot() + scale_y_continuous(breaks = seq(from = 0, to = 1500, by = 100)) +
```

```
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
grid.arrange(bp1, bp2, ncol = 2)
```



Tras este análisis, se llega a la conclusión de que es pertinente incluir los valores atípicos (outliers) detectados en las variables en el estudio. Esto se debe a que dichos valores podrían representar casos clínicos atípicos, como episodios de hipertensión, o respuestas inusuales frente a las mediciones. Se decide conservar estos valores debido a la falta de información que confirme si son errores en la entrada de datos o anomalías en las mediciones.

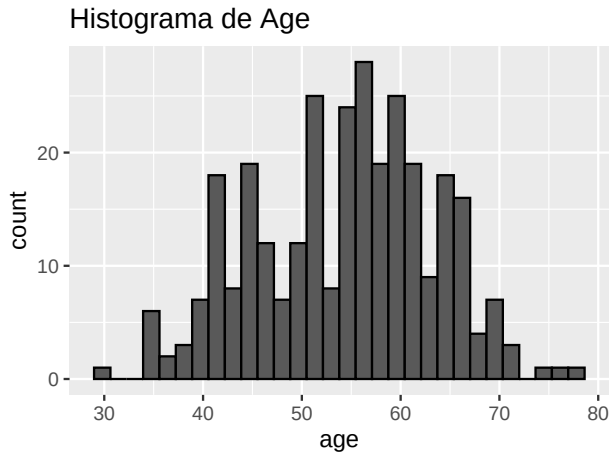
2.4 Análisis de los datos

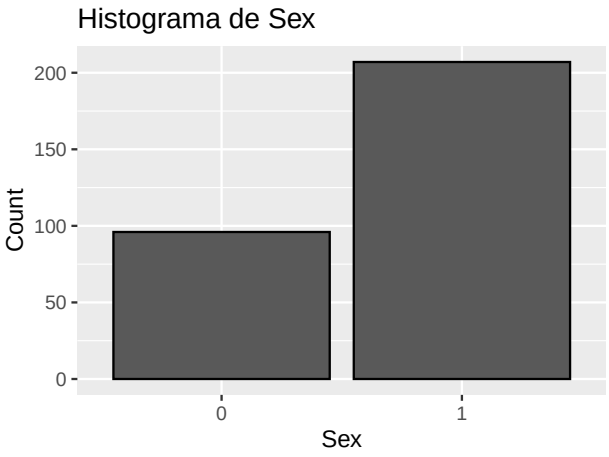
Con la intención de tener un primer acercamiento con las variables y su distribución, se empleará una representación gráfica basado en histogramas:

AGE, **SEX**

```
# Crear el primer gráfico (Histograma de Age)
p1 <- ggplot(data_heart, aes(x = age)) +
  geom_histogram(color = "black") + ggtitle("Histograma de Age")
# Preparar los datos para el segundo gráfico (Histograma de Sex)
sex_counts <- table(data_heart$sex)
# Crear el segundo gráfico (Histograma de Sex)
p2 <- ggplot(as.data.frame(sex_counts), aes(x = Var1, y = Freq)) +
  geom_bar(stat = "identity", color = "black") +
  xlab("Sex") + ylab("Count") + ggtitle("Histograma de Sex")
grid.arrange(p1, p2, ncol = 2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



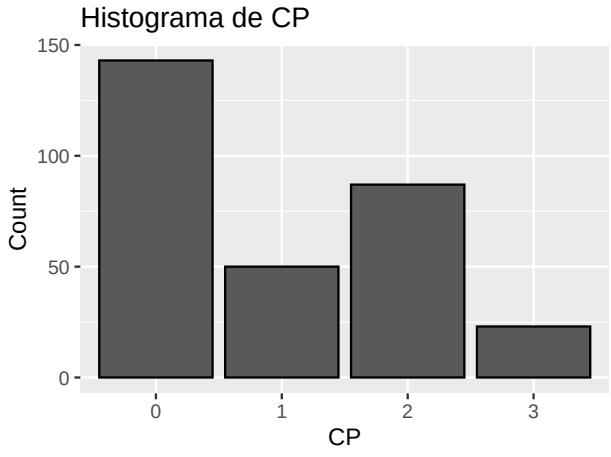


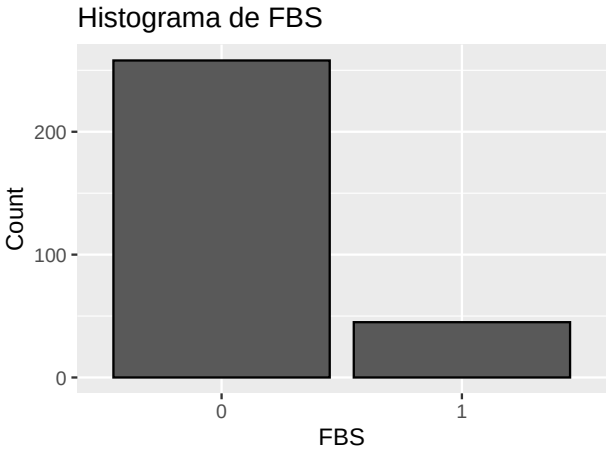
AGE: La distribución de las edades es multimodal, lo que sugiere la presencia de varios grupos de edad distintos. Así mismo, se identifica que la mayoría de los individuos en el conjunto de datos están en sus 50 y 60 años. Hay menos individuos jóvenes y de avanzada edad.

SEX: Esta distribución puede reflejar que los hombres están más representados en los casos de ataques cardíacos dentro de la población de datos, al rededor de 200 registros.

CP, FBS

```
cp_counts <- table(data_heart$cp)
p3 <- ggplot() + geom_bar(data = as.data.frame(cp_counts), aes(x = Var1, y = Freq),
          stat = "identity", color = "black") + xlab("CP") + ylab("Count") +
  ggtitle("Histograma de CP")
fbs_counts <- table(data_heart$fbs)
p4 <- ggplot() + geom_bar(data = as.data.frame(fbs_counts), aes(x = Var1, y = Freq),
           stat = "identity", color = "black") + xlab("FBS") + ylab("Count") +
  ggtitle("Histograma de FBS")
grid.arrange(p3, p4, ncol = 2)
```



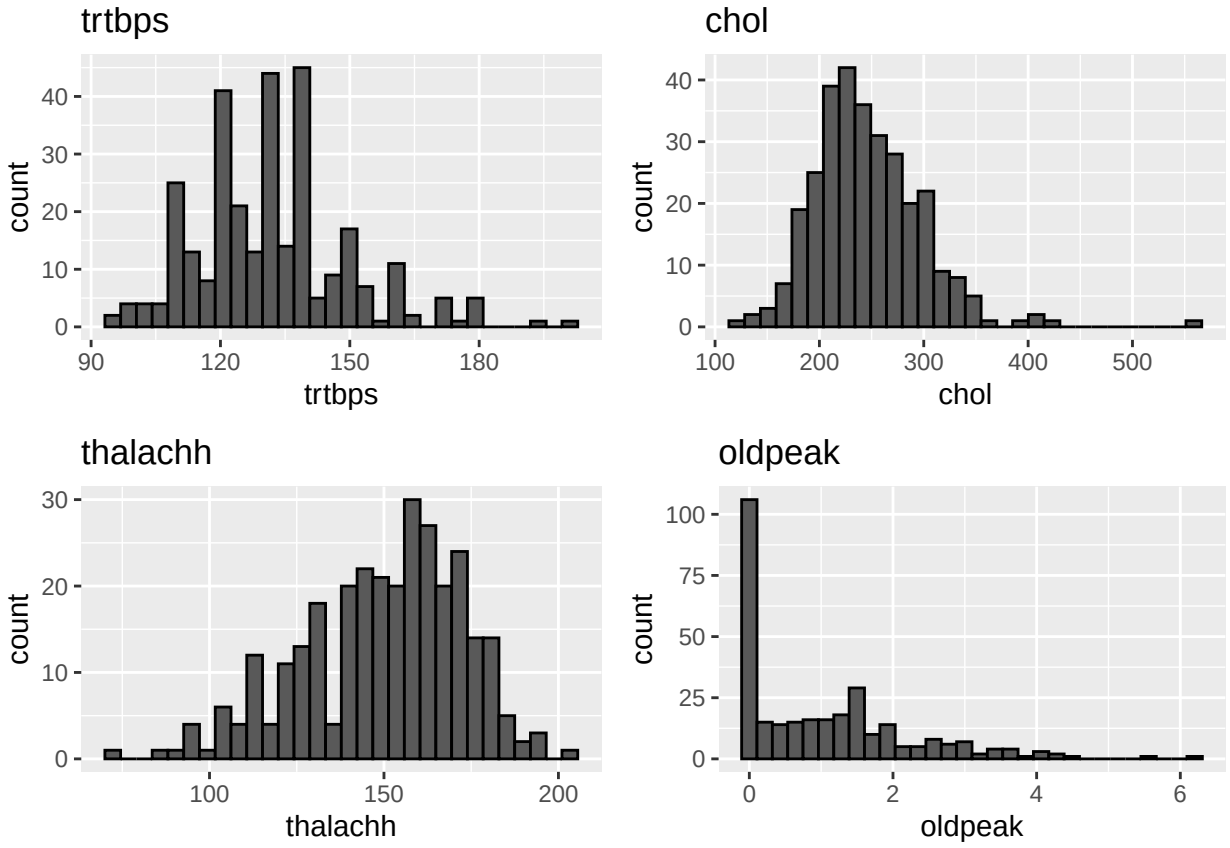


CP: A partir del análisis previo, se puede evidenciar que la mayoría de las observaciones en el conjunto de datos corresponden a individuos sin síntomas de dolor torácico. Las barras para '1' (angina típica) y '2' (angina atípica) son más bajas, con '1' siendo la menos común de las categorías sintomáticas. La categoría '3' (dolor no anginoso) tiene la menor cantidad de observaciones.

FBS: Representa la distribución de los valores de glucosa en sangre en ayunas: Categoría 0: Representa a los individuos con un nivel de glucosa en sangre en ayunas de 120 mg/dl o menos. Categoría 1: Representa a los individuos con un nivel de glucosa en sangre en ayunas mayor de 120 mg/dl. Existe la presencia de más individuos con niveles normales o bajos de glucosa en sangre en ayunas (< 120 mg/dl) que aquellos con niveles altos (> 120 mg/dl), lo cual puede ser relevante para la evaluación de riesgos cardíacos.

TRTBPS, CHOL, THALACHH, OLDPEAK

```
histList <- list()
n = c("trtbps", "chol", "thalachh", "oldpeak")
datosAux = data_heart %>% select(all_of(n))
for(y in 1:ncol(datosAux)) {
  col <- names(datosAux)[y]
  ggp <- ggplot(datosAux, aes_string(x = col)) +
    geom_histogram(bins = 30, color = "black") +
    labs(title = paste(col))
  histList[[y]] <- ggp # Añadimos cada plot a la lista vacía
}
ggplot2.multiplot(plotlist = histList, cols = 2)
```



TRTBPS: Presión arterial en reposo (mmHg). La distribución muestra varios picos, lo que puede indicar rangos comunes de presión arterial en reposo o posibles agrupaciones de datos.

CHOL: Nivel de colesterol sérico (mg/dl). La distribución parece aproximadamente normal con un pico central, indicando que la mayoría de los individuos tienen niveles de colesterol en un rango intermedio.

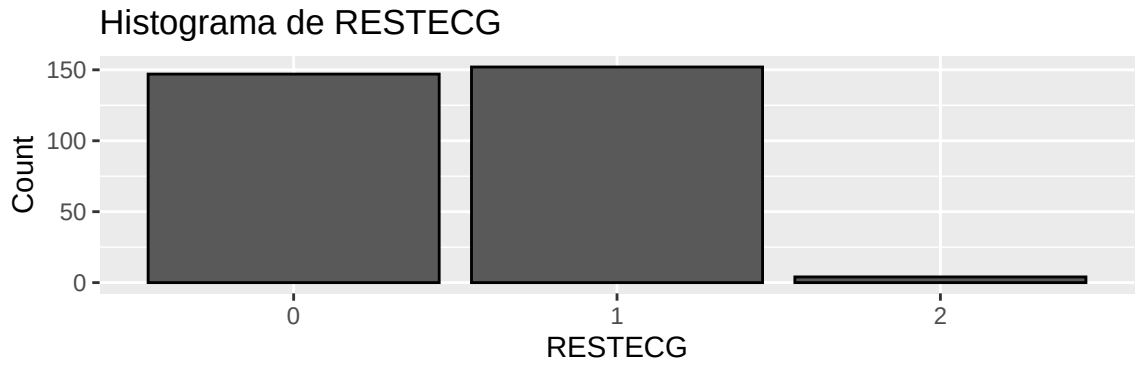
THALACHH: Máxima frecuencia cardíaca alcanzada. Esta distribución también muestra una tendencia aproximadamente normal, con la mayoría de los individuos alcanzando una frecuencia cardíaca máxima en

el rango medio.

OLDPEAK: Depresión del ST inducida por el ejercicio en relación con el reposo. La distribución está sesgada hacia la izquierda, con la mayoría de los individuos mostrando valores bajos de depresión del ST y algunos pocos con valores más altos.

RESTECG

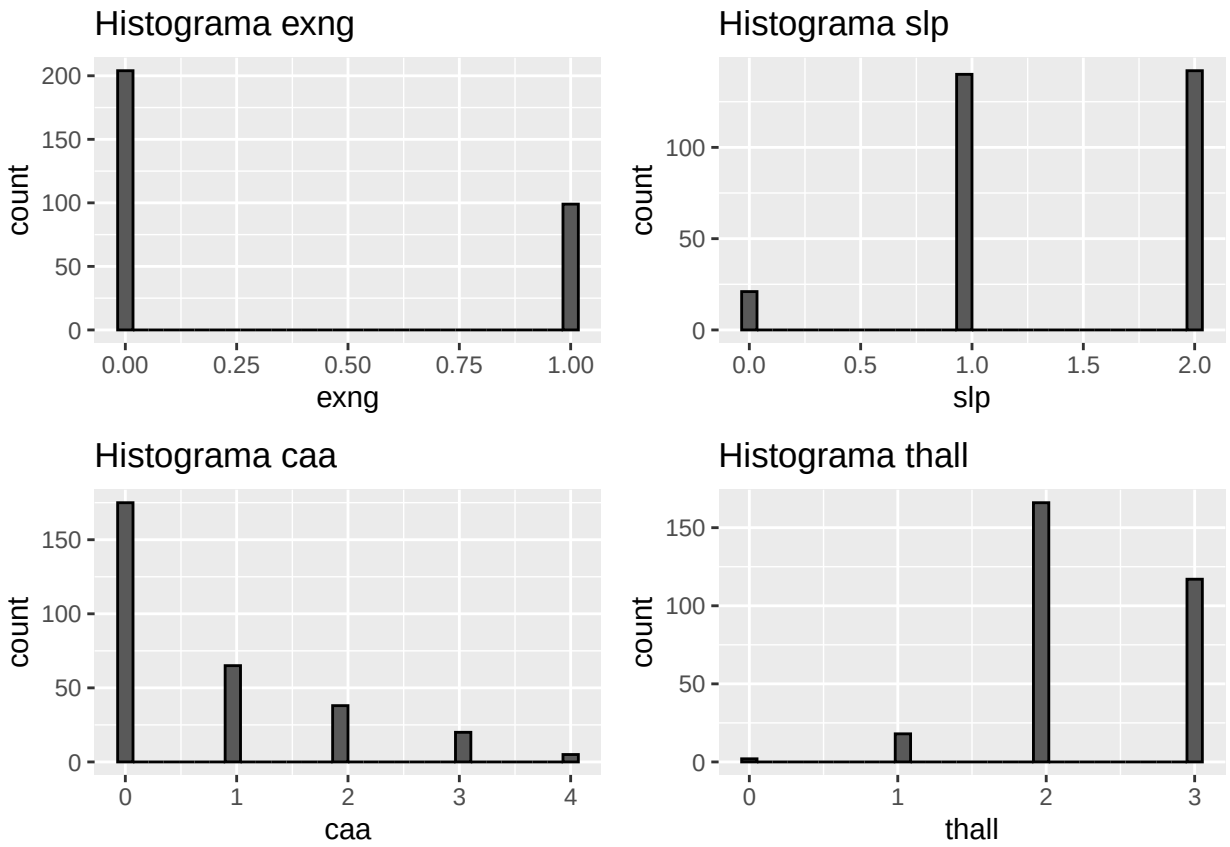
```
restecg_counts <- table(data_heart$restecg)
# Histograma de los conteos de "restecg"
ggplot() +
  geom_bar(data = as.data.frame(restecg_counts), aes(x = Var1, y = Freq),
           stat = "identity", color = "black") +
  xlab("RESTECG") +
  ylab("Count") +
  ggtitle("Histograma de RESTECG")
```



RESTECG: representa la distribución de los resultados electrocardiográficos en reposo: 0: Normal, 1: Anormalidades en la onda ST-T 2: Hipertrofia ventricular izquierda probable o definitiva. El histograma indica que la cantidad de individuos con resultados ECG normales y con anormalidades de onda ST-T es similar, mientras que aquellos con hipertrofia ventricular izquierda son significativamente menos en el conjunto de datos. Información relevante en la identificación de enfermedades cardíacas.

EXNG, SLP, CAA, THALL

```
histList <- list()
n = c("exng", "slp", "caa", "thall")
datosAux = data_heart %>% select(all_of(n))
for(y in 1:ncol(datosAux)) {
  col <- names(datosAux)[y]
  ggp <- ggplot(datosAux, aes_string(x = col)) +
    geom_histogram(bins = 30, color = "black") +
    labs(title = paste("Histograma",col))
  histList[[y]] <- ggp # Añadimos cada plot a la lista vacía
}
ggplot2.multiplot(plotlist = histList, cols = 2)
```



exng (Angina inducida por ejercicio): Se observan dos barras, una para '0' (no) y otra para '1' (sí). La barra para '0' es más alta, indicando que más individuos no experimentaron angina durante el ejercicio.

slp (Pendiente del segmento ST del ejercicio máximo): Hay tres barras, que corresponden a '0' (descendente), '1' (plano), y '2' (ascendente). La pendiente plana es la más común, seguida por la ascendente, y la descendente es la menos común.

caa (Número de vasos principales coloreados por fluoroscopia): Las barras representan el conteo de vasos principales, desde '0' hasta '3' o más. La mayoría de los individuos no tienen vasos coloreados (barra '0'), y hay una disminución progresiva en el número a medida que aumenta el número de vasos afectados.

thall (Resultados de la prueba de talio): Con barras para '1' (defecto fijo), '2' (normal) y '3' (defecto reversible). La categoría '2' (normal) es la más prevalente, seguida de la '3' (defecto reversible), y la '1' (defecto fijo) es la menos común.

2.4.1 Comprobación de la normalidad y homogeneidad de la varianza.

```
data_heart$output <- as.factor(data_heart$output)
selected_vars <- c("age", "cp", "trtbps", "chol", "fbs", "restecg", "thalachh", "exng",
                   "oldpeak", "slp", "caa", "thall")
# Prueba de normalidad con la prueba de Shapiro-Wilk
normality_tests <- lapply(data_heart[selected_vars], shapiro.test)
# dataframe para almacenar los resultados de la prueba de Shapiro-Wilk
normality_results_df <- do.call(rbind, lapply(normality_tests, function(test, var) {
  data.frame(
    Variable = var,
```

```
    Shapiro_W = test$statistic,
    P_Value = test$p.value
  )
}, names(normality_tests)))
# Prueba de homogeneidad de la varianza con la prueba de Levene
homogeneity_tests <- lapply(selected_vars, function(var) {
  leveneTest(as.formula(paste(var, '~ output')), data = data_heart)
})
# Mostramos los resultados de las pruebas de normalidad
head(normality_results_df, 4)
```

```
##       Variable Shapiro_W     P_Value
## age.1      age 0.9863705 0.005798359
## age.2       cp 0.9863705 0.005798359
## age.3   trtbps 0.9863705 0.005798359
## age.4     chol 0.9863705 0.005798359
```

```
# Mostramos los resultados de las pruebas de homogeneidad
#print(homogeneity_tests)
```

Edad (age): W = 0.98637, p-value = 0.005798 La distribución de la edad no es normal al nivel de significancia estándar ($p < 0.05$).

Tipo de dolor en el pecho (cp): W = 0.79016, p-value < 2.2e-16 La distribución de esta variable es significativamente no normal.

Presión arterial en reposo (trtbps): W = 0.96592, p-value = 1.458e-06 La distribución de la presión arterial en reposo no es normal.

Colesterol sérico (chol): W = 0.94688, p-value = 5.365e-09 La distribución del colesterol sérico no es normal.

Azúcar en la sangre en ayunas (fbs): W = 0.42399, p-value < 2.2e-16 La distribución del azúcar en la sangre en ayunas es significativamente no normal.

Resultados electrocardiográficos en reposo (restecg): W = 0.67932, p-value < 2.2e-16 La distribución de los resultados del ECG en reposo no es normal.

Máxima frecuencia cardíaca alcanzada (thalachh): W = 0.97632, p-value = 6.621e-05 La distribución de la frecuencia cardíaca máxima alcanzada no es normal.

Angina inducida por el ejercicio (exng): W = 0.59126, p-value < 2.2e-16 La distribución de la angina inducida por el ejercicio es significativamente no normal.

Depresión del ST inducida por el ejercicio (oldpeak): W = 0.84418, p-value < 2.2e-16 La distribución de la depresión del ST inducida por el ejercicio no es normal.

Pendiente del segmento ST del ejercicio máximo (slp): W = 0.74465, p-value < 2.2e-16 La distribución de la pendiente del segmento ST no es normal.

Número de vasos principales coloreados (caa): W = 0.72812, p-value < 2.2e-16 La distribución del número de vasos principales coloreados no es normal.

Defectos identificados en la prueba de talio (thall): W = 0.75058, p-value < 2.2e-16 La distribución de los defectos identificados en la prueba de talio no es normal.

La mayoría de las variables muestran p-valores significativamente menores a 0.05, lo que indica que las distribuciones de estas variables no son normales según la prueba de Shapiro-Wilk. Este es un hallazgo

importante que afecta la elección de las pruebas estadísticas a utilizar; para datos no normales, se prefieren métodos no paramétricos.

Age: F = 7.9854, p-valor = 0.005031. La varianza no es homogénea ($p < 0.05$). Hay una diferencia significativa en las varianzas entre los grupos.

Cp: F = 12.158, p-valor = 0.0005617. La varianza no es homogénea ($p < 0.001$). Existe una diferencia significativa en las varianzas entre los grupos.

Trtbps: F = 1.857, p-valor = 0.174. Las varianzas son homogéneas ($p > 0.05$). No hay una diferencia significativa en las varianzas entre los grupos.

Chol: F = 0.1015, p-valor = 0.7503. Las varianzas son homogéneas. No hay una diferencia significativa en las varianzas entre los grupos.

Fbs: F = 0.2369, p-valor = 0.6268. Las varianzas son homogéneas. No hay una diferencia significativa en las varianzas entre los grupos.

Restecg: F = 0.2724, p-valor = 0.6021. Las varianzas son homogéneas. No hay una diferencia significativa en las varianzas entre los grupos.

Thalachh: F = 5.2467, p-valor = 0.02268. La varianza no es homogénea ($p < 0.05$). Hay una diferencia significativa en las varianzas entre los grupos.

Exng: F = 40.27, p-valor < 0.0001. La varianza no es homogénea. Existe una diferencia significativa en las varianzas entre los grupos.

Oldpeak: F = 32.916, p-valor < 0.0001. La varianza no es homogénea. Existe una diferencia significativa en las varianzas entre los grupos.

Slp: F = 1.0924, p-valor = 0.2968. Las varianzas son homogéneas. No hay una diferencia significativa en las varianzas entre los grupos.

Caa: F = 26.235, p-valor < 0.0001. La varianza no es homogénea. Existe una diferencia significativa en las varianzas entre los grupos.

Thall: F = 13.614, p-valor = 0.0002664. La varianza no es homogénea. Existe una diferencia significativa en las varianzas entre los grupos.

2.4.2 Dimensión demográfica vs Dimensión Análisis Clinico

AGE vs Dimensión Análisis Clinico

Vamos a proceder a analizar la relación entre la variable AGE y la Dimensión Análisis Clinico.

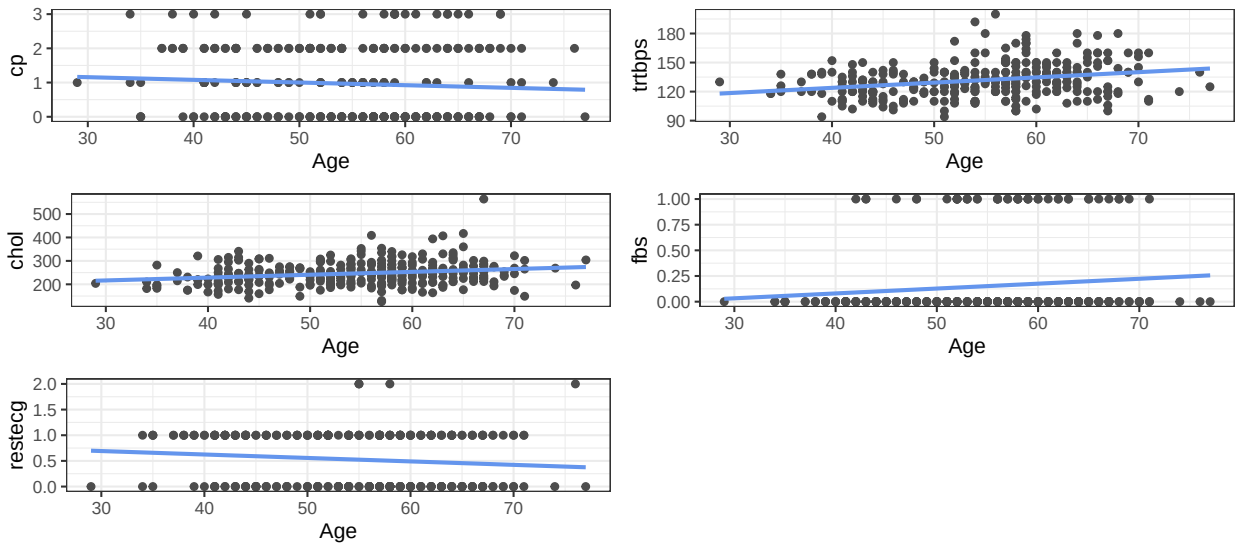
```
selected_vars <- c("cp", "trtbps",  "chol", "fbs", "restecg", "thalachh", "exng", "oldpeak", "slp", "ca
filtered_data <- data_heart %>% select(age, all_of(selected_vars))
# Calculo correlaciones
correlations <- sapply(filtered_data[, -1], function(var) cor(filtered_data$age, var))
correlations
```

```
##          cp      trtbps        chol         fbs     restecg    thalachh
## -0.06865302  0.27935091  0.21367796  0.12130765 -0.11621090 -0.39852194
##        exng     oldpeak         slp         caa       thall
##  0.09680083  0.21001257 -0.16881424  0.27632624  0.06800138
```

```
first_half_vars <- selected_vars[1:5]
second_half_vars <- selected_vars[6:11]
first_half_graphList <- vector('list', length(first_half_vars))
```

```r
second_half_graphList <- vector('list', length(second_half_vars))
create_graphs <- function(vars, graphList) {
  for (i in seq_along(vars)) {
    var <- vars[i]
    graphList[[i]] <- ggplot(data = data_heart, aes_string(x = "age", y = var)) +
      geom_point(color = "gray30") + geom_smooth(method = "lm", color = "cornflowerblue", se = FALSE) +
      labs(x = "Age", y = var) + theme_bw()
  }
  return(graphList)
}
first_half_graphList <- create_graphs(first_half_vars, first_half_graphList)
second_half_graphList <- create_graphs(second_half_vars, second_half_graphList)
grid.arrange(grobs = first_half_graphList, ncol = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

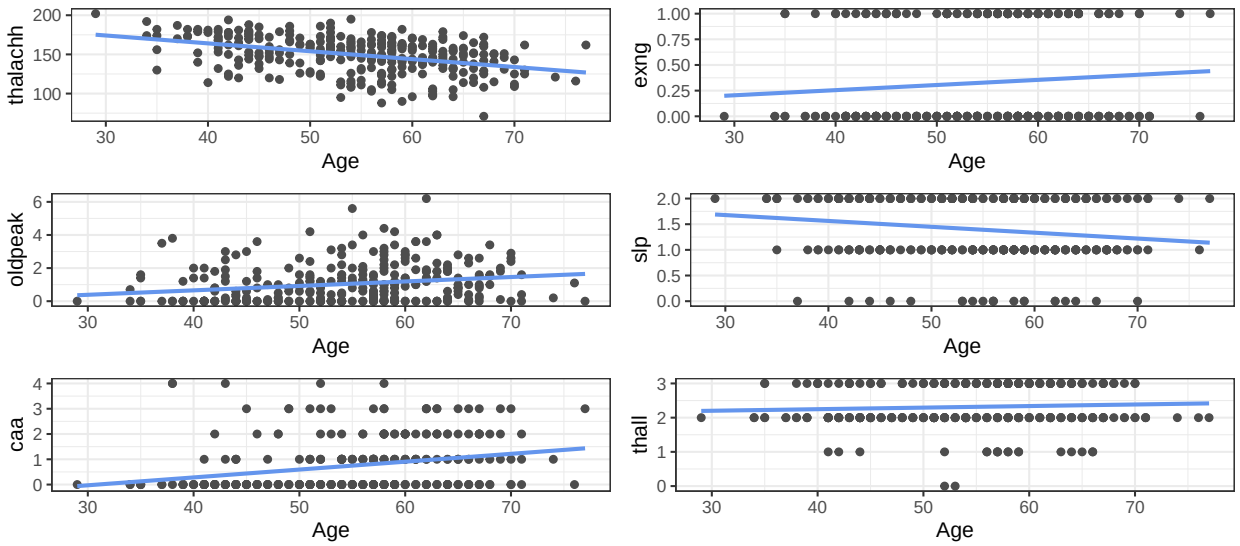


- **cp (dolor en el pecho):** Con un coeficiente de -0.06865302, hay una correlación negativa muy débil con la edad, lo que sugiere que no hay una relación lineal significativa entre la edad y el tipo de dolor en el pecho experimentado.
- **trtbps (presión arterial en reposo):** El coeficiente de 0.27935091, correlación positiva débil a moderada con la edad, lo que significa que la presión arterial en reposo tiende a aumentar ligeramente a medida que la gente envejece.
- **chol (colesterol sérico):** Una correlación de 0.21367796 positiva entre el nivel de colesterol y la edad, lo que podría indicar que los niveles de colesterol tienden a ser más altos en individuos de mayor edad.
- **fbs (azúcar en la sangre en ayunas):** Con un coeficiente de 0.12130765, hay una correlación positiva muy débil, lo que sugiere que hay poca o ninguna relación lineal directa entre la edad y los niveles de azúcar en la sangre en ayunas.

- **restecg (electrocardiográficos en reposo):** El valor de -0.11621090, una correlación negativa muy débil con la edad, implicando que no hay una relación lineal clara entre los resultados del ECG en reposo y la edad.

```
grid.arrange(grobs = second_half_graphList, ncol = 2)
```

```
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
## `geom_smooth()` using formula = 'y ~ x'
```

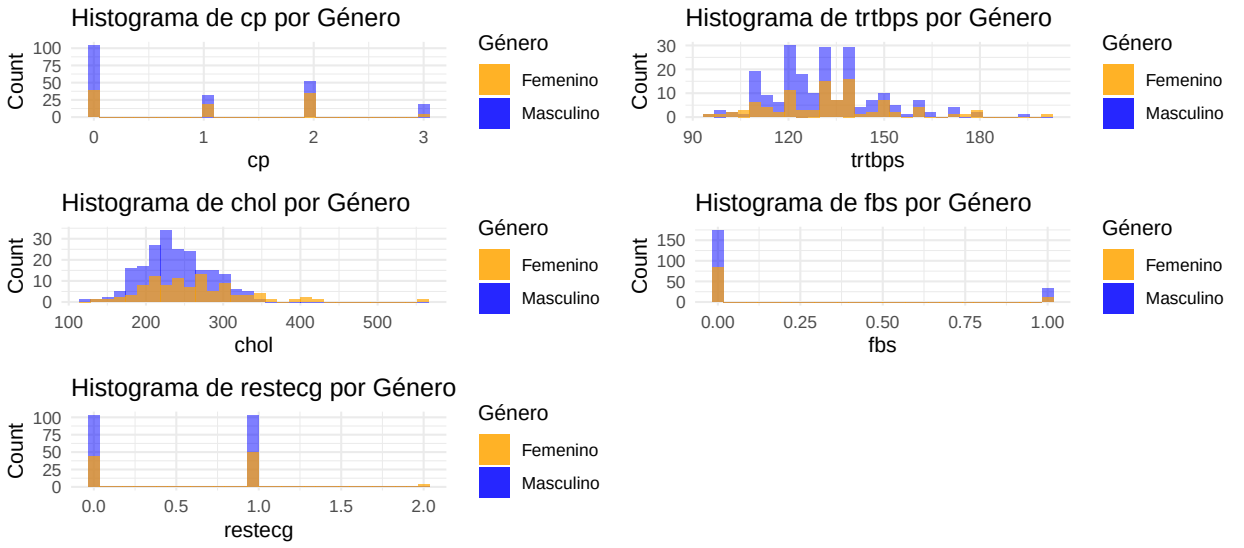


- **thalachh (máxima frecuencia cardíaca alcanzada):** Una correlación negativa -0.39852194 indica que a medida que la edad aumenta, la frecuencia cardíaca máxima alcanzada tiende a disminuir.
- **exng (angina inducida por el ejercicio):** La correlación positiva de 0.09680083 indica que no hay una relación lineal significativa entre la angina inducida por el ejercicio y la edad.
- **oldpeak (depresión del ST inducida por el ejercicio en relación con el reposo):** Con un coeficiente de 0.21001257, hay una correlación positiva débil, lo que podría significar que hay una leve tendencia a que la depresión del ST sea mayor en personas de mayor edad.
- **slp (la pendiente del segmento ST del ejercicio máximo):** El valor de -0.16881424 muestra una correlación negativa débil, sugiriendo que la pendiente del segmento ST durante el ejercicio máximo tiende a disminuir ligeramente con la edad.
- **caa (número de vasos principales coloreados por fluoroscopia):** Una correlación de 0.27632624 indica una relación positiva débil a moderada con la edad, lo que podría implicar que la probabilidad de tener una mayor cantidad de vasos coloreados aumenta con la edad.
- **thall (defectos identificados en la prueba de talio):** La correlación positiva muy débil de 0.06800138 sugiere que apenas hay una relación entre los defectos detectados por la prueba de talio y la edad.

2.4.3 Genero vs Dimensión laboratorio

Vamos a proceder a analizar la relación entre la variable de genero y la Dimensión laboratorio.

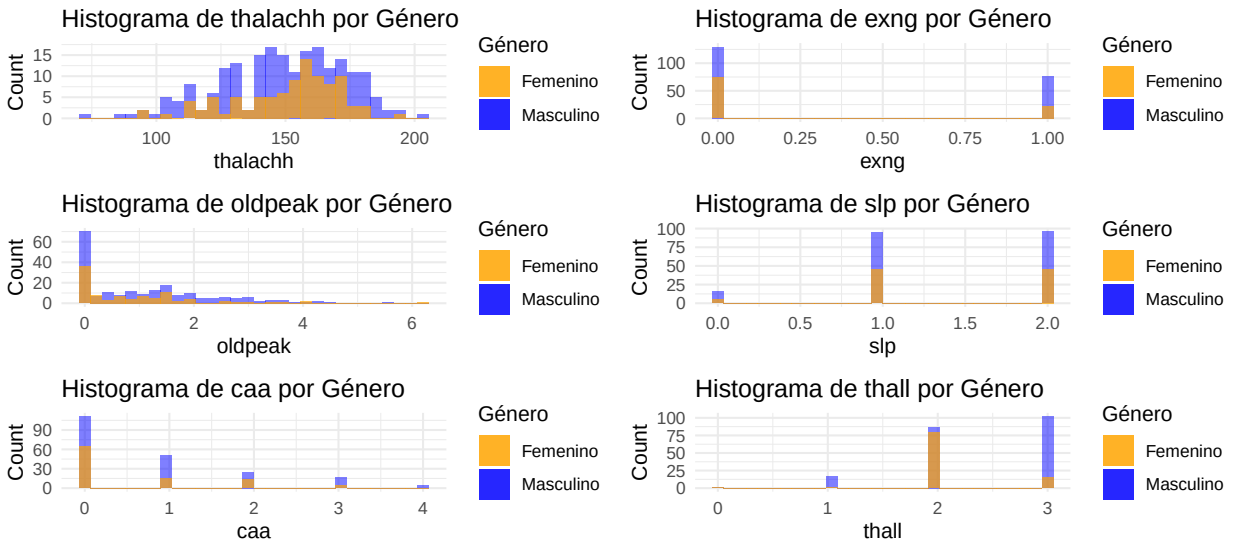
```
data_heart$sex <- factor(data_heart$sex, levels = c(0, 1), labels = c("Femenino", "Masculino"))
selected_vars <- c("cp", "trtbps", "chol", "fbs", "restecg", "thalachh", "exng", "oldpeak", "slp", "caa
filtered_data <- data_heart %>% select(sex, all_of(selected_vars))
first_group_vars <- selected_vars[1:5]
second_group_vars <- selected_vars[6:11]
hist_list <- lapply(first_group_vars, function(var) {
  ggplot(filtered_data, aes_string(x = var, fill = "sex")) +
    geom_histogram(data = subset(filtered_data, sex == "Masculino"), bins = 30, alpha = 0.5) +
    geom_histogram(data = subset(filtered_data, sex == "Femenino"), bins = 30, alpha = 0.7) +
    scale_fill_manual(values = c("Femenino" = "orange", "Masculino" = "blue")) +
    labs(x = var, y = "Count", fill = "Género") + ggtitle(paste("Histograma de", var, "por Género")) + t
})
first_group_grid <- do.call(grid.arrange, c(hist_list, ncol = 2))
```



- **Distribución de cp (dolor en el pecho):** La mayoría de los individuos, parecen tener un tipo de dolor en el pecho tipo '0', lo que podría indicar el tipo más común en el conjunto de Datos. Hay menos incidencias de las categorías '1', '2' y '3', con los hombres mostrando ligeramente más casos en estas categorías que las mujeres.
- **Distribución de trtbps (presión arterial en reposo):** La distribución de la presión arterial en reposo para los hombres muestra una tendencia a valores más altos en comparación con las mujeres. Ambos presentan un pico al rededor de los 120-140 mmHg, que está en el rango de la presión arterial normal.
- **Distribución de chol (colesterol sérico):** Tanto hombres como mujeres presentan una distribución similar del colesterol sérico, con un pico claro alrededor de 200-250 mg/dl. Esto sugiere que la mayoría de los individuos en la muestra tienen niveles de colesterol dentro de lo que se consideraría un rango normal o ligeramente elevado.
- **Distribución de fbs (azúcar en la sangre en ayunas):** La gran mayoría de los individuos de ambos géneros tienen un azúcar en la sangre en ayunas inferior a 0.25, lo que probablemente indica niveles normales de glucosa en ayunas.

- **Distribución de restecg (resultados electrocardiográficos en reposo):** La categoría '0' y '1, son las más comúnes para los resultados del ECG en reposo en ambos géneros, lo que podría indicar una presencia de hipertrofia y una ausencia general de anomalías notables en el ECG de la mayoría de la muestra.

```
hist_list <- lapply(second_group_vars, function(var) {
  ggplot(filtered_data, aes_string(x = var, fill = "sex")) +
    geom_histogram(data = subset(filtered_data, sex == "Masculino"), bins = 30, alpha = 0.5) +
    geom_histogram(data = subset(filtered_data, sex == "Femenino"), bins = 30, alpha = 0.7) +
    scale_fill_manual(values = c("Femenino" = "orange", "Masculino" = "blue")) +
    labs(x = var, y = "Count", fill = "Género") + ggtitle(paste("Histograma de", var, "por Género")) +
second_group_grid <- do.call(grid.arrange, c(hist_list, ncol = 2))
```



- **Distribución de thalachh (máxima frecuencia cardíaca alcanzada):**Ambos géneros muestran una distribución similar con un pico en la frecuencia cardíaca alrededor de 150 a 160 bpm (latidos por minuto). Los hombres tienden a tener un rango ligeramente más amplio de frecuencias cardíacas máximas comparado con las mujeres.
- **Distribución de exng (angina inducida por el ejercicio):** La mayoría de los individuos en ambos géneros no experimentan angina inducida por el ejercicio, como se muestra por las barras altas en la categoría '0'.
- **Distribución de oldpeak (depresión del ST inducida por el ejercicio en relación con el reposo):** Los valores de 'oldpeak' para ambos géneros están mayormente concentrados cerca de 0, lo que indica que la muestra o de los sujetos no tienen una significativa depresión del segmento ST. Hay algunos casos, más visible en hombres, donde su valor de 'oldpeak' es mayor, lo que podría indicar algún grado de enfermedad relacionada con el corazón.
- **Distribución de slp (la pendiente del segmento ST del ejercicio máximo):** La pendiente del segmento ST parece ser más comúnmente de tipo '2' para ambos géneros, lo cual puede considerarse como normal.
- **Distribución de caa (número de vasos principales coloreados por fluoroscopia):** La mayoría de los sujetos de ambos géneros parecen no tener vasos principales coloreados (categoría '0'), lo que podría ser indicativo de ausencia de enfermedad relacionada con el corazón relevante. Hay una presencia notable de individuos masculinos con mayor número de vasos coloreados en comparación con las mujeres.

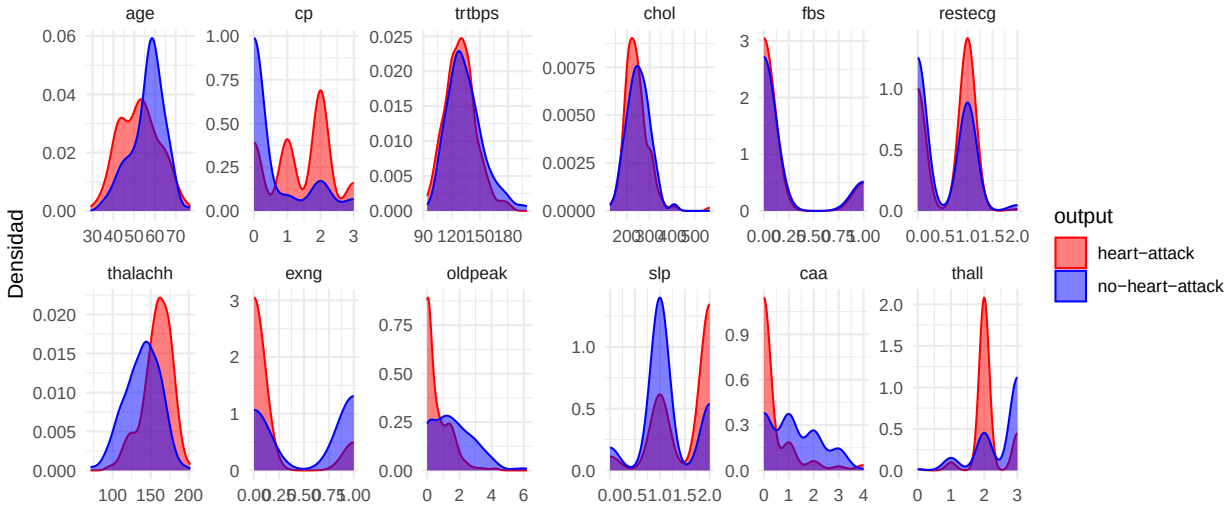
- **Distribución de thall (defectos identificados en la prueba de talio):** La categoría '2' parece ser la más prevalente en hombres y mujeres, lo que podría corresponder a un hallazgo normal o a la ausencia de defectos reversibles en la prueba de talio. Hay menos casos en las categorías '1' y '3', con los hombres mostrando una distribución más amplia entre estas categorías en comparación con las mujeres.

2.4.4 Relación entre variables

Vamos a buscar las relaciones en función de la variable OUTPUT y unas variables elegidas que creemos que pueden ayudar a predecir los inconvenientes de ataque al corazón:

Debido a que dentro de OUTPUT, cada clase tiene una representación notablemente distinta, realizaremos un analisis mediante gráficos de densidad:

```
# Ajustar el factor de la variable CLASS con los nombres de etiquetas deseadas
data_heart$output <- factor(data_heart$output, levels = c("1", "0"),
                            labels = c("heart-attack", "no-heart-attack"))
# Variables seleccionadas
selected_vars <- c("age", "cp", "trtbps", "chol", "fbs", "restecg", "thalachh", "exng", "oldpeak", "slp
# Crear el grid de gráficos de densidad
ggplot(data_heart, aes(fill = output)) +
  facet_wrap(~ variable, scales = "free", nrow = 2) +
  geom_density(data = reshape2::melt(data_heart, id.vars = "output", measure.vars = selected_vars),
               aes(x = value, color = output), alpha = 0.5) +
  labs(x = "", y = "Densidad", fill = "output") +
  scale_fill_manual(values = c("heart-attack" = "red", "no-heart-attack" = "blue")) +
  scale_color_manual(values = c("heart-attack" = "red", "no-heart-attack" = "blue")) +
  theme_minimal()
```



- **Edad (age):** Los pacientes que han sufrido un ataque cardíaco tienden a ser mayores en comparación con aquellos que no han sufrido uno, como lo indica la curva roja desplazada hacia la derecha.
- **Presión arterial en reposo (trtbps):** La presión arterial en reposo de los pacientes que han sufrido un ataque cardíaco parece tener una distribución similar a la de aquellos que no, aunque hay una ligera tendencia hacia valores más altos en el grupo que ha sufrido un ataque cardíaco, lo que podría sugerir una correlación entre la presión arterial en reposo más alta y los ataques cardíacos.

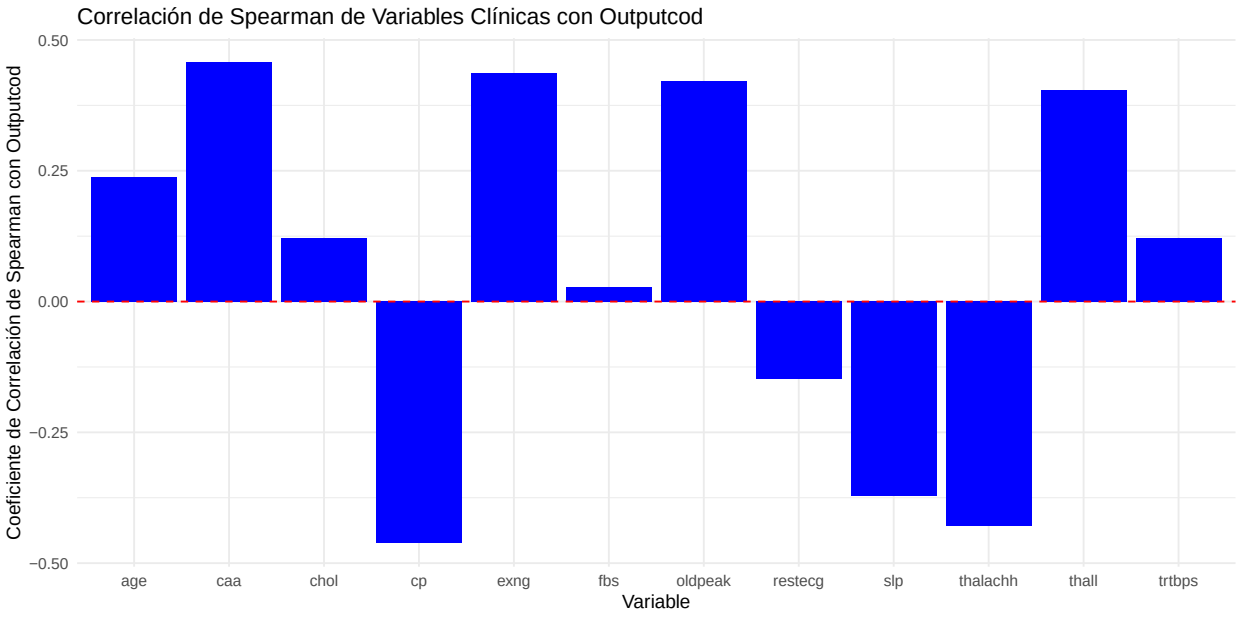
- **Colesterol sérico (chol):** Los niveles de colesterol sérico parecen ser más altos en el grupo de pacientes que han sufrido un ataque cardíaco, aunque las distribuciones son bastante similares.
- **Máxima frecuencia cardíaca alcanzada (thalachh):** Los individuos que no han sufrido un ataque cardíaco tienden a alcanzar frecuencias cardíacas máximas más altas, como se muestra por la curva azul con un pico más a la derecha.
- **Depresión del ST inducida por el ejercicio en relación con el reposo (oldpeak):** Los pacientes que han tenido un ataque cardíaco muestran valores más altos de 'oldpeak', indicando mayor depresión del ST.
- **Número de vasos principales coloreados por fluoroscopia (caa):** Los individuos con ataques cardíacos (curva roja) tienden a tener una mayor cantidad de vasos principales coloreados, lo que sugiere que este puede ser un indicador importante de riesgo cardíaco.

Vamos a analizar las correlaciones entre estas variables:

```
data_heart$outputcod <- as.numeric(factor(data_heart$output))
selected_vars <- c("age", "cp", "trtbps", "chol", "fbs", "restecg", "thalachh", "exng", "oldpeak", "slp
# Calculamos la correlación de Spearman
spearman_correlation_r <- cor(data_heart[, c(selected_vars, "outputcod")], method = "spearman")
# Mostramos la correlación de Spearman con 'output'
spearman_with_output_r <- spearman_correlation_r["outputcod", ]
spearman_with_output_r <- spearman_with_output_r[!names(spearman_with_output_r) %in% "outputcod"]
correlation_df <- data.frame(
  Variable = names(spearman_with_output_r),
  Correlation = spearman_with_output_r
)
correlation_df
```

```
##          Variable Correlation
## age           age  0.23840007
## cp             cp -0.46086018
## trtbps     trtbps  0.12159275
## chol         chol  0.12088824
## fbs           fbs  0.02804576
## restecg   restecg -0.14861154
## thalachh thalachh -0.42836989
## exng         exng  0.43675708
## oldpeak   oldpeak  0.42148706
## slp           slp -0.37146048
## caa           caa  0.45760748
## thall       thall  0.40329932
```

```
ggplot(correlation_df, aes(x = Variable, y = Correlation)) +
  geom_col(fill = "blue") + geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
  theme_minimal() +  xlab("Variable") + ylab("Coeficiente de Correlación de Spearman con Outputcod") +
  ggtitle("Correlación de Spearman de Variables Clínicas con Outputcod")
```

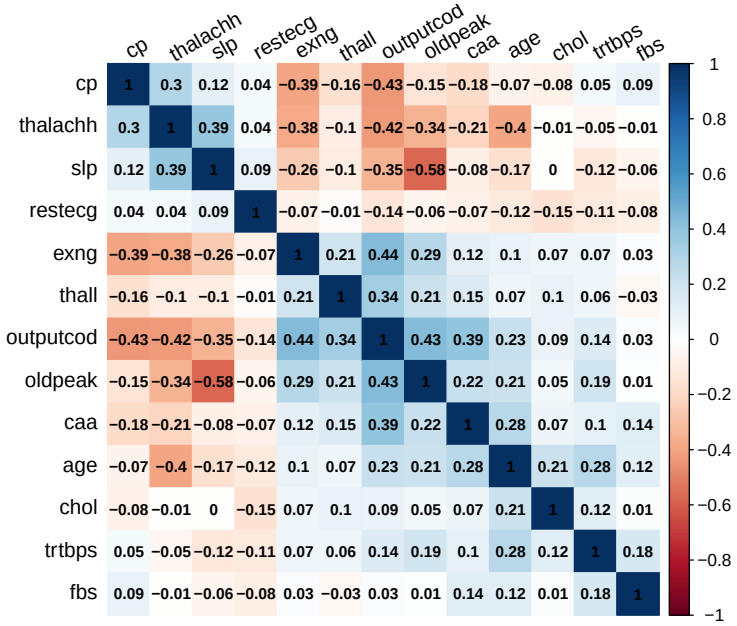


- Las variables **caa (número de vasos principales coloreados por fluoroscopia)**, **exng (angina inducida por el ejercicio)**, **oldpeak (depresión del ST inducida por el ejercicio en relación con el reposo)**, y **thall (defectos identificados en la prueba de talio)** muestran correlaciones positivas con output Esto indica que hay una asociación en la que valores más altos de estas variables tienden a ir acompañados de un valor más alto en output. Se puede indicar de manera preliminar que estas variables pueden considerarse factores de riesgo o indicadores de mayor probabilidad de dicho evento.

- Por otro lado, **cp (tipo de dolor en el pecho)**, **slp (la pendiente del segmento ST del ejercicio máximo)** y **thalachh (máxima frecuencia cardíaca alcanzada)**, muestran correlaciones negativas con output. Esto significa que valores más altos en estas variables están asociados con valores más bajos en output, lo cual podría interpretarse como un indicador de menor riesgo.

- Las variables **chol (colesterol sérico)**, **fbs (azúcar en la sangre en ayunas)**, **restecg (resultados electrocardiográficos en reposo)** y **trtbps (presión arterial en reposo)**, presentan correlaciones muy débiles con output, ya sean positivas o negativas. Estas débiles correlaciones sugieren que no hay una fuerte relación lineal entre estas variables y la variable de salida (output).

```
n = c("outputcod", "age", "cp", "trtbps", "chol", "fbs", "restecg", "thalachh", "exng", "oldpeak", "slp
factores= data_heart %>% select(all_of(n))
res<-cor(factores)
corrplot(res,method="color",tl.col="black", tl.srt=30, order = "AOE",
number.cex=0.75,sig.level = 0.01, addCoef.col = "black")
```



3 Construcción del conjunto de datos final

3.1 Recodificación de variables

Se aplica la codificación one-hot en la variable OUTPUT para optimizar su interpretación por los algoritmos de aprendizaje automático. Este enfoque garantiza que las categorías se manejen de manera equitativa, sin asumir jerarquías o secuencias entre ellas.

```
# Codificación one-hot de la variable "OUTPUT"
data_heart <- cbind(data_heart, model.matrix(~ output - 1, data_heart))
head(data_heart,4)
```

```
##   age       sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
## 1  63 Masculino  3    145  233   1       0      150    0     2.3   0   0     1
## 2  37 Masculino  2    130  250   0       1      187    0     3.5   0   0     2
## 3  41  Femenino  1    130  204   0       0      172    0     1.4   2   0     2
## 4  56 Masculino  1    120  236   0       1      178    0     0.8   2   0     2
##         output outputcod outputheart-attack outputno-heart-attack
## 1 heart-attack         1                  1                     0
## 2 heart-attack         1                  1                     0
## 3 heart-attack         1                  1                     0
## 4 heart-attack         1                  1                     0
```

3.2 Discretización y codificación de la variable AGE

Procederemos a discretizar la variable AGE definiendo grupos como “jóvenes” (menores de 30 años), “Mediana_edad” (30-50 años), y “Mayores” (mayores de 50 años):

```r
# Discretización basada en criterios de dominio
data_heart$age_group <- cut(data_heart$age, breaks = c(0, 30, 50, Inf),
                      labels = c("Jóven", "Mediana_edad", "Mayor_edad"))
head(data_heart,4)
```

```
##   age       sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
## 1  63 Masculino  3    145  233   1       0      150    0     2.3   0   0     1
## 2  37 Masculino  2    130  250   0       1      187    0     3.5   0   0     2
## 3  41  Femenino  1    130  204   0       0      172    0     1.4   2   0     2
## 4  56 Masculino  1    120  236   0       1      178    0     0.8   2   0     2
##         output outputcod outputheart-attack outputno-heart-attack    age_group
## 1 heart-attack         1                  1                     0   Mayor_edad
## 2 heart-attack         1                  1                     0 Mediana_edad
## 3 heart-attack         1                  1                     0 Mediana_edad
## 4 heart-attack         1                  1                     0   Mayor_edad
```

```r
# Codificación one-hot de la variable "AGE_Group"
data_heart <- cbind(data_heart, model.matrix(~ age_group - 1, data_heart))
# Verificar la codificación one-hot
head(data_heart,4)
```

```
##   age       sex cp trtbps chol fbs restecg thalachh exng oldpeak slp caa thall
## 1  63 Masculino  3    145  233   1       0      150    0     2.3   0   0     1
## 2  37 Masculino  2    130  250   0       1      187    0     3.5   0   0     2
## 3  41  Femenino  1    130  204   0       0      172    0     1.4   2   0     2
## 4  56 Masculino  1    120  236   0       1      178    0     0.8   2   0     2
##         output outputcod outputheart-attack outputno-heart-attack    age_group
## 1 heart-attack         1                  1                     0   Mayor_edad
## 2 heart-attack         1                  1                     0 Mediana_edad
## 3 heart-attack         1                  1                     0 Mediana_edad
## 4 heart-attack         1                  1                     0   Mayor_edad
##   age_groupJóven age_groupMediana_edad age_groupMayor_edad
## 1              0                     0                   1
## 2              0                     1                   0
## 3              0                     1                   0
## 4              0                     0                   1
```

4 Aplicación de pruebas estadisticas

4.1 Regresión logística

```r
# Establecer una semilla para reproducibilidad
set.seed(123)
# Dividir los datos en conjuntos de entrenamiento y prueba
split <- sample.split(data_heart$trtbps, SplitRatio = 0.7)
train_data <- subset(data_heart, split == TRUE)
test_data <- subset(data_heart, split == FALSE)
# Ajustar el modelo de regresión logistica con los datos de entrenamiento
ModlgF <- glm(output ~ age + sex + trtbps + chol + fbs + exng + oldpeak + caa + thall +cp + trtbps + res
                      data = train_data, family = binomial())
summary(ModlgF)
```

```
##
## Call:
## glm(formula = output ~ age + sex + trtbps + chol + fbs + exng +
##     oldpeak + caa + thall + cp + trtbps + restecg + slp + thalachh,
##     family = binomial(), data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.986500   3.161683  -0.628 0.529804
## age          -0.006412   0.029600  -0.217 0.828518
## sexMasculino  1.822565   0.580850   3.138 0.001702 **
## trtbps        0.017612   0.011666   1.510 0.131113
## chol          0.004454   0.004432   1.005 0.314962
## fbs           0.239072   0.602680   0.397 0.691602
## exng          0.911284   0.501707   1.816 0.069314 .
## oldpeak       0.422185   0.234587   1.800 0.071909 .
## caa           0.944318   0.246665   3.828 0.000129 ***
## thall         0.994650   0.337759   2.945 0.003231 **
## cp           -0.818427   0.220288  -3.715 0.000203 ***
## restecg      -0.230784   0.406824  -0.567 0.570523
## slp          -0.501202   0.404063  -1.240 0.214825
## thalachh     -0.029569   0.012650  -2.338 0.019412 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 296.37  on 214  degrees of freedom
## Residual deviance: 156.58  on 201  degrees of freedom
## AIC: 184.58
##
## Number of Fisher Scoring iterations: 6
```

```r
set.seed(123)
split <- sample.split(data_heart$trtbps, SplitRatio = 0.7)
# Ajustar el modelo de regresión logistica con los datos de entrenamiento
ModlgF <- glm(output ~ age + sex + exng + oldpeak + caa + thall +cp + slp + thalachh,
                      data = train_data, family = binomial())
summary(ModlgF)
```

```
##
## Call:
## glm(formula = output ~ age + sex + exng + oldpeak + caa + thall +
##     cp + slp + thalachh, family = binomial(), data = train_data)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.81088    2.89925  -0.280 0.779717
## age           0.01490    0.02673   0.557 0.577211
## sexMasculino  1.55337    0.50692   3.064 0.002182 **
## exng          0.88628    0.48856   1.814 0.069669 .
## oldpeak       0.46042    0.22596   2.038 0.041586 *
## caa           0.88740    0.23863   3.719 0.000200 ***
## thall         1.03036    0.32506   3.170 0.001525 **
```

```
## cp           -0.75470    0.20684  -3.649 0.000264 ***
## slp          -0.49841    0.39594  -1.259 0.208102
## thalachh     -0.02282    0.01147  -1.989 0.046735 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 296.37  on 214  degrees of freedom
## Residual deviance: 160.55  on 205  degrees of freedom
## AIC: 180.55
##
## Number of Fisher Scoring iterations: 6
```

Según el modelo con las variables con correlación que muestran una fuerte relación con la variable dependiente "output", y que se pueden considerar estadisticamente significativas son: **sexMasculino**, **exng**, **oldpeak**, **caa**, **thall**, **cp**, **thalachh** lo que puede sugerir que estos factores son relevantes en la predicción de 'output'.

4.2 Matríz de confusión

```r
# probabilidades predichas por el modelo
probabilidades <- predict(ModlgF, newdata = test_data, type = "response")
# conversión de las probabilidades en predicciones binarias usando el umbral de 0.7
predicciones <- ifelse(probabilidades >= 0.7, 1, 0)
# matriz de confusión frente a valores reales
matriz_confusion <- table(predic = predicciones, val = test_data$output)
# sensibilidad y especificidad
sensibilidad <- matriz_confusion[2, 2] / sum(matriz_confusion[2, ])
especificidad <- matriz_confusion[1, 1] / sum(matriz_confusion[1, ])
print(matriz_confusion)
```

```
##       val
## predic heart-attack no-heart-attack
##      0           45               8
##      1            3              32
```

Verdaderos Negativos: 45 casos donde el modelo predijo correctamente que no ocurriría un ataque cardíaco (no-heart-attack) y efectivamente no ocurrió.

Falsos Positivos: 8 casos donde el modelo predijo incorrectamente que ocurriría un ataque cardíaco (heart-attack), pero en realidad no sucedió.

Falsos Negativos: 3 casos dobde el modelo falló al predecir un ataque cardíaco; es decir, predijo que no habría ataque cardíaco (no-heart-attack), pero realmente sí ocurrió.

Verdaderos Positivos: 32 casos donde el modelo predijo correctamente la ocurrencia de un ataque cardíaco (heart-attack) y este realmente sucedió.

```r
print(paste("Sensibilidad:", sensibilidad))
```

```
## [1] "Sensibilidad: 0.914285714285714"
```

```r
print(paste("Especificidad:", especificidad))
```

```
## [1] "Especificidad: 0.849056603773585"
```

El modelo predice de manera satisfactoria los casos en los cuales se presenta un ataque cardiaco relacionado con las variables identificadas que pueden poternciar esta problemática, así mismo, el modelo responde a identificar de manera relativamente óptima, los casos en los que no se presenta este escenario.

5 Conclusiones

Principales Factores Predictivos de un Ataque Cardíaco:

Los factores que mostraron una asociación estadísticamente significativa con la variable dependiente output incluyen sexMasculino, cp, thalachh, caa y thall.

- *sexMasculino:* Ser hombre se asoció con un aumento en las odds de output.
- *cp:* Una disminución en esta variable (tipo de dolor en el pecho) se asoció con un aumento en las odds de output, lo que sugiere que la mayoría de escenarios en los que se da un ataque cardiaco el paciente es asintomático, o presenta un dolor torácico relacionado con una angina típica.
- *thalachh:* Una disminución en la máxima frecuencia cardíaca alcanzada se asoció con un aumento en las odds de output. Cabe indicar que esta variable está relacionada con la actividad física, lo que puede sugerir que los individuos que realizan un esfuerzo físico presentan esta patología.
- *caa:* Aunque en la gráfica de densidad observamos un pico rojo en caa = 0, el modelo de regresión logística reveló que un mayor número de caa se asocia con un aumento significativo en las odds de output. Esto puede indicar que, si bien tener caa = 0 es común entre los pacientes que han tenido un ataque cardíaco, tener un número mayor de vasos coloreados está aún más fuertemente asociado con el riesgo de ataque cardíaco.
- *thall:* Un aumento en esta variable se asoció con un aumento en las odds de output.

Efecto del Ejercicio y la Dieta en el Riesgo de Ataques Cardíacos:

El ejercicio, representado por variables como thalachh (máxima frecuencia cardíaca alcanzada) y exng (angina inducida por el ejercicio), mostró que una menor frecuencia cardíaca máxima y la presencia de angina están asociadas con un mayor riesgo, aunque la significancia de exng es marginal. La dieta, que podría influir en variables como chol (colesterol sérico) y fbs (azúcar en la sangre en ayunas), no mostró una relación estadísticamente significativa con el riesgo de ataque cardíaco en este modelo.

Diferencias en Síntomas y Factores de Riesgo entre Hombres y Mujeres:

La variable sexMasculino mostró que ser hombre está asociado con un mayor riesgo de ataque cardíaco en este modelo. Esto sugiere que puede haber diferencias en los factores de riesgo y posiblemente en los síntomas entre hombres y mujeres.

Cambio en el Riesgo y Presentación de Enfermedades Cardíacas con la Edad:

La variable age no mostró una asociación significativa con output en este modelo, lo que sugiere que, dentro del rango de edad del conjunto de datos, la edad por sí sola no es un predictor significativo. Sin embargo, esto no significa que la edad no sea un factor en el riesgo general de enfermedades cardíacas.