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

Tipología y ciclo de vida de los datos

Sergio Cobos García

Universitat Oberta de Catalunya

06 de junio de 2023

Índice

1. Descripción del dataset

El conjunto de datos está relacionado con la enfermedad cardíaca. Proporciona información sobre varios factores, como la edad, el sexo, los síntomas y las mediciones relacionadas con la presión arterial, el colesterol y la frecuencia cardíaca máxima.

Este tipo de conjunto de datos es importante porque la enfermedad cardíaca es una de las principales causas de muerte en todo el mundo, y comprender los factores de riesgo asociados y su relación con el riesgo de ataque cardíaco puede ayudar en la prevención y el diagnóstico temprano.

El problema o pregunta que este conjunto de datos pretende abordar es la predicción del riesgo de ataque cardíaco en función de los diferentes atributos o características de los pacientes. Al analizar los datos, se pueden encontrar patrones y relaciones entre los factores de riesgo y el resultado de tener un ataque cardíaco. Esto puede ayudar a los profesionales médicos a identificar a las personas con mayor riesgo y brindarles un tratamiento adecuado y medidas preventivas.

Además, este conjunto de datos también puede ser utilizado por investigadores o científicos de datos para desarrollar modelos predictivos o algoritmos de aprendizaje automático que puedan ayudar a predecir el riesgo de enfermedad cardíaca en pacientes nuevos o futuros. Esto podría ser útil en entornos clínicos para evaluar el riesgo de los pacientes y tomar decisiones informadas sobre su atención médica.

1.1 Exploración del conjunto de datos

Verificamos la estructura del juego de datos principal. Vemos el número de columnas que tenemos y ejemplos de los contenidos de las filas.

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

Vemos que tenemos **14** variables y **303** registros.

Revisamos la descripción de las variables contenidas en la descripción del repositorio y si los tipos de variables se corresponden con las que hemos cargado. Las organizamos lógicamente para darles sentido y construimos un pequeño diccionario de datos utilizando la documentación auxiliar.

1. **Variable objetivo**:

- output (Diagnóstico de enfermedad cardíaca)

2. **Dimension demográfica**:

- Age: Edad
- Sex: Género

2. **Dimension dolor**:

- cp: Tipo de dolor en el pecho

3. **Dimension resultados médicos**:

- trtbps: Presión arterial en reposo
- chol: Colesterol
- fbs: Azúcar en sangre en ayunas
- restecg: Resultados electrocardiográficos en reposo
- thalach: Frecuencia cardíaca máxima alcanzada
- exng: Angina inducida por ejercicio

4. **Dimension ejercicio/corazón**:

- oldpeak: Depresión del segmento ST inducida por ejercicio en relación con el reposo)
- slp: Pendiente del segmento ST de ejercicio
- caa: Número de vasos principales
- thall: Thalasemia

2. Subselección del conjunto de datos

Vamos a realizar una subselección de variables demográficas y relacionadas con mediciones conjuntas que nos permite investigar la influencia de la edad, el sexo y diversas mediciones (presión arterial, colesterol, resultados electrocardiográficos, frecuencia cardíaca y angina inducida por ejercicio) en el diagnóstico de enfermedad cardíaca.

El **objetivo** será estudiar la relación entre factores demográficos, mediciones de salud y el diagnóstico de enfermedad cardíaca, para identificar qué variables son importantes para predecir la presencia de esta enfermedad.

```
##   age sex trtbps chol fbs restecg thalachh exng output
## 1  63   1    145  233   1       0      150    0      1
## 2  37   1    130  250   0       1      187    0      1
## 3  41   0    130  204   0       0      172    0      1
## 4  56   1    120  236   0       1      178    0      1
## 5  57   0    120  354   0       1      163    1      1
## 6  57   1    140  192   0       1      148    0      1
```

Como podemos observar, la subselección de datos se ha realizado correctamente.

3. Limpieza de los datos

3.1 Gestión de ceros o elementos nulos

El siguiente paso será la limpieza de datos, mirando si hay valores vacíos o nulos.

```
## [1] "NA"
```

```
##      age      sex   trtbps     chol      fbs  restecg thalachh     exng
##        0        0        0        0        0        0        0        0
##   output
##        0
```

```
## [1] "Blancos"
```

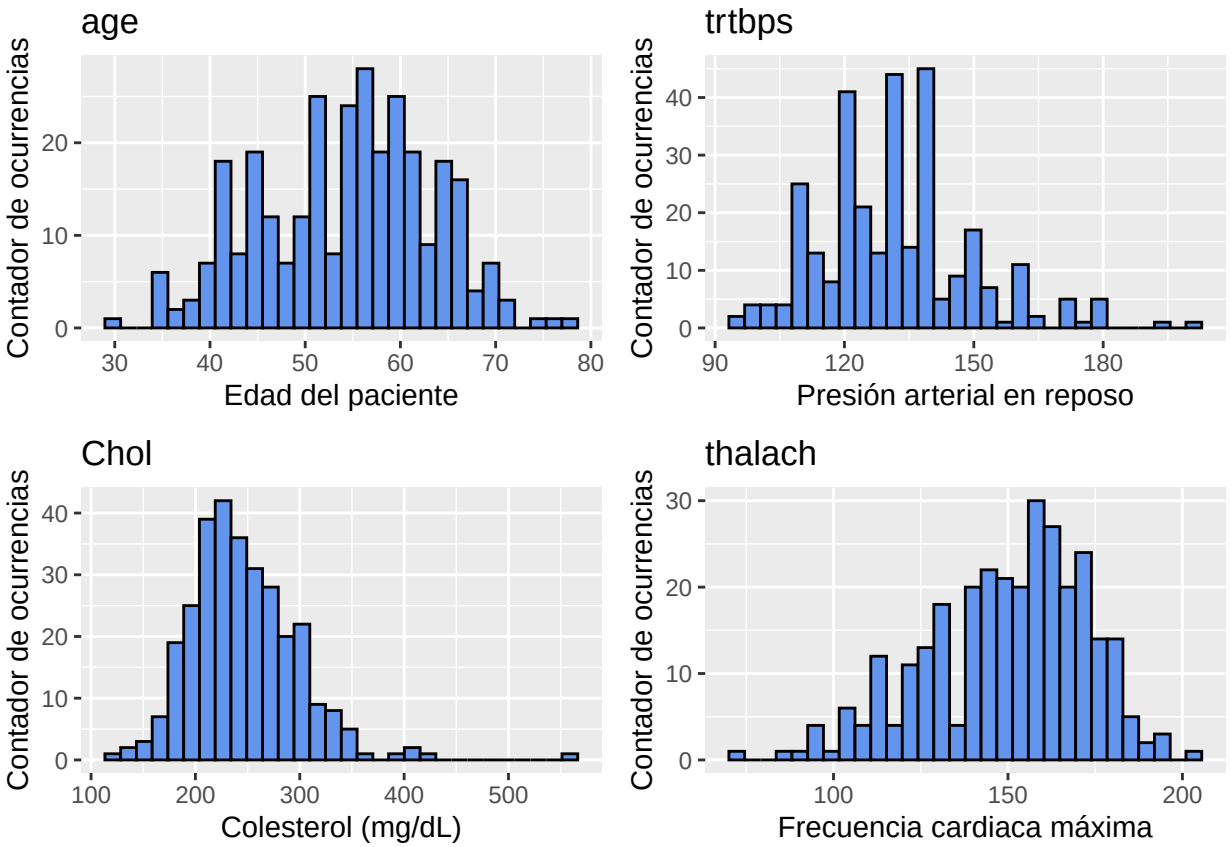
```
##      age      sex   trtbps     chol      fbs  restecg thalachh     exng
##        0        0        0        0        0        0        0        0
##   output
##        0
```

Como podemos observar, no se ha encontrado la presencia de registros o celdas con valores faltantes o nulos.

3.2 Idendificación y gestión de valores extremos

Realizaremos un análisis exploratorio de los datos con el objetivo de comprender mejor la estructura y características de la muestra. Este análisis nos permitirá identificar posibles valores atípicos o fuera de lo común en las variables estudiadas. Examincaremos la distribución, rango y medidas de tendencia central de cada variable, así como también visualizaremos gráficos y tablas para obtener una visión general de los datos. A través de este proceso, buscamos obtener información detallada sobre las características de los datos y revelar posibles patrones o anomalías que puedan afectar nuestro análisis posterior. Este análisis exploratorio nos proporcionará una base sólida para comprender mejor los datos y tomar decisiones informadas sobre el manejo de los valores atípicos que podamos identificar.

```
##       age            trtbps         thalachh          chol      
##  Min.   :29.00   Min.   : 94.0   Min.   : 71.0   Min.   :126.0  
##  1st Qu.:47.50   1st Qu.:120.0   1st Qu.:133.5   1st Qu.:211.0  
##  Median :55.00   Median :130.0   Median :153.0   Median :240.0  
##  Mean   :54.37   Mean   :131.6   Mean   :149.6   Mean   :246.3  
##  3rd Qu.:61.00   3rd Qu.:140.0   3rd Qu.:166.0   3rd Qu.:274.5  
##  Max.   :77.00   Max.   :200.0   Max.   :202.0   Max.   :564.0
```

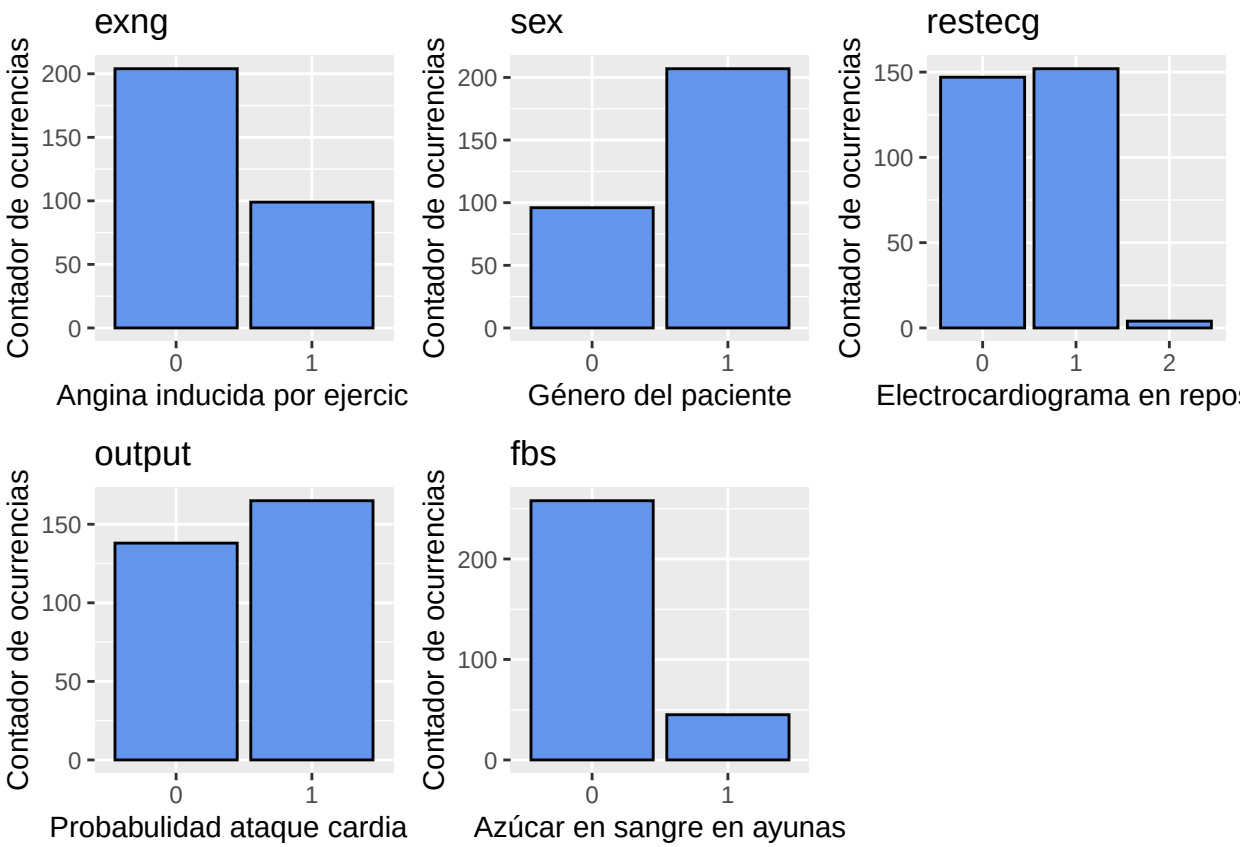


age: Se encontró que la muestra de edad se encuentra en un rango entre 29 y 77 años, con una mediana de 55 años y una media de 54.37 años. No se observaron valores atípicos o fuera de lo común en esta variable.

trtbps (Presión arterial en reposo): El valor mínimo es de 94.0 mm Hg, el primer cuartil se encuentra en 120.0 mm Hg, la mediana está en 130.0 mm Hg, la media es de 131.6 mm Hg, el tercer cuartil se sitúa en 140.0 mm Hg y el valor máximo registrado es de 200.0 mm Hg. La distribución de la variable "trtbps" está significativamente concentrada alrededor de la mediana, que es de 130.0 mm Hg. Podemos observar valores atípicos, el mayor de los cuales es 200 mm Hg, puede considerarse atípico en el contexto de la muestra de pacientes, pero no anómalo, ya que sugiere la presencia de una presión arterial elevada en ese caso particular.

chol: El rango de valores se encuentra entre 126 mg/dl y 564 mg/dl. La mediana es de 240 mg/dl. La media es de 246.3 mg/dl. La distribución del colesterol está significativamente concentrada alrededor de la mediana. Sin embargo, es importante tener en cuenta que el valor máximo observado es de 564 mg/dl, lo cual es considerablemente más alto que el resto de los valores, puede considerarse atípico en el contexto de la muestra de pacientes, pero no lo consideraremos anómalo.

thalach: Los valores oscilan entre 71 y 202, con una media de 149.6. La mediana, que representa el valor central, es de 153.0. La distribución del colesterol está significativamente concentrada alrededor de la mediana. No se observaron valores atípicos o fuera de lo común en esta variable.




```
## Tabla de proporciones para exng :
##   Categoria Proporcion.variable Proporcion.Freq
## 1         0                   0       0.6732673
## 2         1                   1       0.3267327
##
## Tabla de proporciones para output :
##   Categoria Proporcion.variable Proporcion.Freq
## 1         0                   0       0.4554455
## 2         1                   1       0.5445545
##
## Tabla de proporciones para sex :
##   Categoria Proporcion.variable Proporcion.Freq
## 1         0                   0       0.3168317
## 2         1                   1       0.6831683
##
## Tabla de proporciones para fbs :
##   Categoria Proporcion.variable Proporcion.Freq
## 1         0                   0       0.8514851
## 2         1                   1       0.1485149
##
## Tabla de proporciones para restecg :
##   Categoria Proporcion.variable Proporcion.Freq
## 1         0                   0      0.48514851
## 2         1                   1      0.50165017
## 3         2                   2      0.01320132
```

exng: La variable "exng" representa la presencia de angina inducida por ejercicio, donde el valor 0 indica la ausencia de esta condición y el valor 1 indica su presencia. Al analizar los datos, se observa que aproximadamente el 67.33% de los pacientes en el conjunto de datos no experimentan angina inducida por ejercicio, mientras que alrededor del 32.67% sí la experimentan. No se encontraron valores atípicos en esta variable.

sex: La variable "sex" representa el género de los pacientes, donde el valor 1 corresponde a pacientes masculinos y el valor 0 corresponde a pacientes femeninos. Al analizar los datos, se observa que aproximadamente el 68.32% de los pacientes en el conjunto de datos son del género masculino, mientras que alrededor del 31.68% son del género femenino. No se encontraron valores atípicos en esta variable.

fbs: La variable "fbs" representa el nivel de azúcar en sangre en ayunas de los pacientes, donde el valor 1 indica un nivel elevado de azúcar en sangre (verdadero) y el valor 0 indica un nivel normal de azúcar en sangre (falso). Al examinar los datos, se encuentra que aproximadamente el 14.85% de los pacientes presentan un nivel elevado de azúcar en sangre, mientras que alrededor del 85.15% tienen un nivel normal. No se encontraron valores atípicos en esta variable.

restecg: La variable "restecg" representa los resultados del electrocardiograma en reposo de los pacientes. El valor 0 indica un electrocardiograma normal, el valor 1 indica la presencia de anormalidades en la onda ST-T (inversiones de la onda T y/o elevación o depresión del segmento ST de > 0.05 mV), y el valor 2 indica la presencia de hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes. Al analizar los datos, se observa que aproximadamente el 48.51% de los pacientes tienen un electrocardiograma normal, alrededor del 50.17% muestran anormalidades en la onda ST-T y un pequeño porcentaje, el 1.32%, muestra signos de hipertrofia ventricular izquierda. No se encontraron valores atípicos en esta variable.

output: La variable "output" representa la probabilidad de tener una enfermedad cardíaca en función del grado de estrechamiento del diámetro. El valor 0 indica un estrechamiento menor al 50%, lo que se asocia con una menor probabilidad de tener enfermedad cardíaca. Por otro lado, el valor 1 indica un estrechamiento mayor al 50%, lo que se relaciona con una mayor probabilidad de tener enfermedad cardíaca. Al analizar los datos, se observa que aproximadamente el 45.54% de los pacientes tienen un estrechamiento menor al 50%, mientras que el 54.46% presentan un estrechamiento mayor al 50%. Estos resultados nos proporcionan información sobre la distribución de la probabilidad de enfermedad cardíaca en la muestra estudiada, destacando la existencia de una proporción considerable de pacientes con un mayor riesgo de enfermedad cardíaca debido al estrechamiento significativo del diámetro.

Se destaca la presencia de valores atípicos en las variables de presión arterial en reposo y colesterol, así como la proporción de pacientes con angina inducida por ejercicio, género y niveles de azúcar en sangre anormales. Estos hallazgos pueden ser útiles para comprender mejor los factores de riesgo y la probabilidad de enfermedad cardíaca en la muestra analizada.

4. Análisi de los datos

4.1 Selección de los grupos de datos a analizar

Realizaremos una selección de un grupo en el que incluiremos las variables “age”, “sex”, “trtbps”, “chol”, “fbs”, “restecg”, “thalachh”, “exng” y “output”. El objetivo principal es analizar la relación entre estas variables y el resultado deseado, representado por la variable “output”. Nuestra meta es determinar qué variables son buenas predictoras de este resultado y comprender cómo se relacionan con él.

En el segundo grupo, consideraremos las variables “age”, “sex”, “fbs”, “restecg” y “exng”. El objetivo será analizar la relación entre la edad del paciente y estas variables específicas. Buscaremos identificar posibles patrones o tendencias en cómo estas variables se relacionan con la edad de los pacientes.

A través de este análisis, esperamos obtener una comprensión más profunda de las variables seleccionadas y su capacidad para predecir el resultado deseado. Esto nos ayudará a tomar decisiones informadas en términos de selección de variables y ajuste de modelos en análisis posteriores.

4.2 Comprobación de normalidad y de homogeneidad

4.2.1 Comprobación de normalidad con el test de Shapiro-Wilk

El Shapiro-Wilk test, o simplemente Shapiro test, es una prueba estadística utilizada para evaluar si una muestra de datos sigue una distribución normal. Esta prueba se basa en la **hipótesis nula de que los datos provienen de una población con distribución normal**. El objetivo del Shapiro test es determinar si hay suficiente evidencia para rechazar esta hipótesis nula y concluir que los datos no siguen una distribución normal.

El Shapiro test calcula un estadístico de prueba basado en las desviaciones de los datos respecto a la distribución normal esperada. El estadístico de prueba se compara con una distribución de referencia conocida y se obtiene un p-valor, que representa la probabilidad de observar un estadístico de prueba igual o más extremo que el calculado, si los datos realmente siguieran una distribución normal.

Si el p-valor obtenido es menor que un nivel de significancia predefinido (usaremos 0.05), se rechaza la hipótesis nula y se concluye que los datos no siguen una distribución normal. En cambio, si el p-valor es mayor que el nivel de significancia, no hay suficiente evidencia para rechazar la hipótesis nula y se puede asumir que los datos siguen una distribución normal.

El test de Shapiro-Wilk se utiliza principalmente para verificar la normalidad de una distribución en variables continuas. En el caso de variables categóricas con dos o tres categorías, no tiene mucho sentido aplicar el test de Shapiro-Wilk, ya que este test se basa en la suposición de una distribución continua.

```
##              p-value Test statistic
## age      5.798359e-03      0.9863705
## trtbps   1.458097e-06      0.9659179
## chol     5.364848e-09      0.9468815
## thalachh 6.620819e-05      0.9763154
```

El valor p obtenido para todas las variables es significativamente menor que 0.05, lo que indica que las distribuciones de las variables no son normales. Esto implica que las variables no se ajustan a una distribución de campana simétrica y pueden tener sesgos o valores atípicos que afectan su distribución.

4.2.2 Comprobación de homogeneidad de la varianza

Para evaluar la homogeneidad de la varianza en el grupo de variables númericas continuas “age”, “trtbps”, “chol”, y “thalachh”, en relación a la variable objetivo “output”, se puede utilizar el test de Levene.

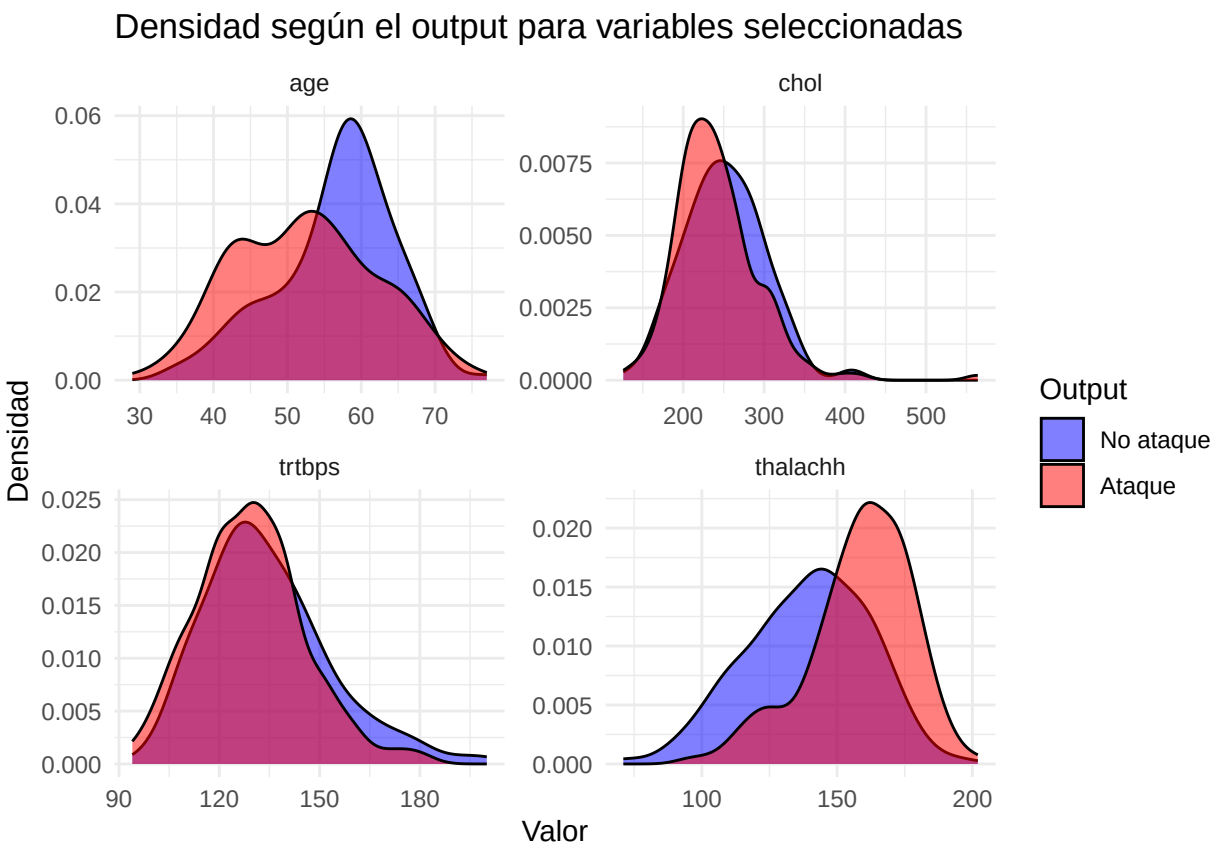
El resultado del test de Levene mostrará la estadística de prueba y el valor p asociado. Un valor p menor que el nivel de significancia elegido (generalmente 0.05) indica que hay evidencia suficiente para rechazar la hipótesis nula de homogeneidad de la varianza, lo que significa que las varianzas entre grupos no son iguales. Por otro lado, un valor p mayor o igual al nivel de significancia sugiere que no hay suficiente evidencia para rechazar la hipótesis nula y se puede asumir homogeneidad de varianza.

```
## [1] "Test de Levene para age:"

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value   Pr(>F)
## group   1  7.9854 0.005031 **
##       301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "------------------------------------------------------------------------"

## [1] "Test de Levene para trtbps:"

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1   1.857  0.174
##       301

## [1] "------------------------------------------------------------------------"

## [1] "Test de Levene para chol:"

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value Pr(>F)
## group   1  0.1015 0.7503
##       301

## [1] "------------------------------------------------------------------------"

## [1] "Test de Levene para thalachh:"

## Levene's Test for Homogeneity of Variance (center = median)
##        Df F value  Pr(>F)
## group   1  5.2467 0.02268 *
##       301
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

## [1] "------------------------------------------------------------------------"
```

Estos resultados indican que la homogeneidad de varianzas difiere según la variable considerada. Las variables “age” y “thalachh” presentan evidencia de no homogeneidad de varianzas entre los grupos de la variable “output”, mientras que las variables “trtbps” y “chol” no presentan evidencia suficiente para rechazar la hipótesis de homogeneidad de varianzas.

4.3 Comparación de grupos

4.3.1 Relación "output" vs "age", " chol", "trtbps" y "thalachh"



age: Se observa una mayor densidad de la clase no ataque (menor probabilidad de ataque al corazón) a partir de personas mayores de 55 años. Esto sugiere una posible asociación entre la edad y la probabilidad de tener un ataque al corazón. Es cun resulyado curioso, ya que se esperaria todo lo contrario, pero es importante considerar la posibilidad de que la muestra de estudio no sea representativa de la población general.

chol: Se ibserva que en el colesterol (variable "chol") hay una mayor densidad de casos de "no ataque" en un rango aproximado entre 250 y 350, y una mayor densidad de casos de "ataque" en un rango entre 150 y 250, es posible inferir que existe una asociación entre los niveles de colesterol y la probabilidad de ataque cardíaco en tus datos. Es un resultado curioso, teniendo en cuenta que el colesterol alto es considerado un factor de riesgo para enfermedades cardiovasculares, incluyendo el riesgo de ataque cardíaco.

trtbps:En el análisis de la variable "trtbps" (presión arterial en reposo), se observa una leve mayor densidad de casos de ataque cardíaco hasta aproximadamente 130. A partir de este punto, se aprecia una mayor densidad de casos de no ataque cardíaco. Es un resultado curioso, teniendo en cuenta que, en general, mantener la presión arterial dentro de rangos saludables se asocia con un menor riesgo de enfermedades cardiovasculares.

thalachh: Podemos observar una menor densidad de casos de ataque cardíaco hasta aproximadamente 150 en la variable "thalachh" y un aumento en la densidad de casos de ataque cardíaco a partir de este punto, esto podría sugerir que una frecuencia cardíaca más baja durante el ejercicio se asocia con un menor riesgo de ataque cardíaco en tu conjunto de datos. En este caso, esto puede tener sentido en el contexto de la relación entre la frecuencia cardíaca y la salud cardiovascular.

Se realizará la prueba **t de Student** en las variables "chol" y "trtbps" en relación con la variable objetivo

"output". Aunque se observó una desviación moderada de la normalidad en estas variables, se considera que el tamaño de la muestra (superior a 30) y que se encontró homogeneidad de varianzas en ambas variables, justifica el uso de la prueba t como una aproximación adecuada. El objetivo de esta prueba será comparar las medias de "chol" y "trtbps" entre los grupos definidos por "output" y determinar si existen diferencias significativas.

chol

```
##
##  Welch Two Sample t-test
##
## data:  datos$chol by datos$output
## t = 1.4948, df = 298.03, p-value = 0.136
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -2.803241 20.516548
## sample estimates:
## mean in group 0 mean in group 1
##        251.0870        242.2303
```

El resultado de la prueba t de Student para la variable "chol" es t = 1.4948, con un número aproximado de grados de libertad de 298.03. El valor de p obtenido es 0.136. Esto indica que no hay suficiente evidencia para rechazar la hipótesis nula de que no hay diferencia significativa en los niveles de colesterol entre los grupos definidos por la variable "output".

trtbps

```
##
##  Welch Two Sample t-test
##
## data:  datos$trtbps by datos$output
## t = 2.5083, df = 272.56, p-value = 0.01271
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  1.096240 9.094801
## sample estimates:
## mean in group 0 mean in group 1
##        134.3986        129.3030
```

El resultado de la prueba t de Student para la variable "trtbps" es t = 2.5083, con un número aproximado de grados de libertad de 272.56. El valor de p obtenido es 0.01271. Esto indica que hay evidencia suficiente para rechazar la hipótesis nula de que no hay diferencia significativa en los niveles de presión arterial en reposo entre los grupos definidos por la variable "output".

En el caso de "age" y "thalach" que se aproximan a una distribución normal pero no hay homogeneidad de varianzas frente a la variable objetivo, se puede utilizar el **test de Welch** para comparar las medias de los grupos.

El test de Welch, también conocido como el test t de Welch, es una variante del test t de Student que no asume igualdad de varianzas entre los grupos. Este test es adecuado cuando las varianzas de los grupos son diferentes.

age

```
##
##  Welch Two Sample t-test
##
## data:  age by output
## t = 4.0797, df = 301, p-value = 5.781e-05
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  2.124635 6.084324
## sample estimates:
## mean in group 0 mean in group 1
##        56.60145        52.49697
```

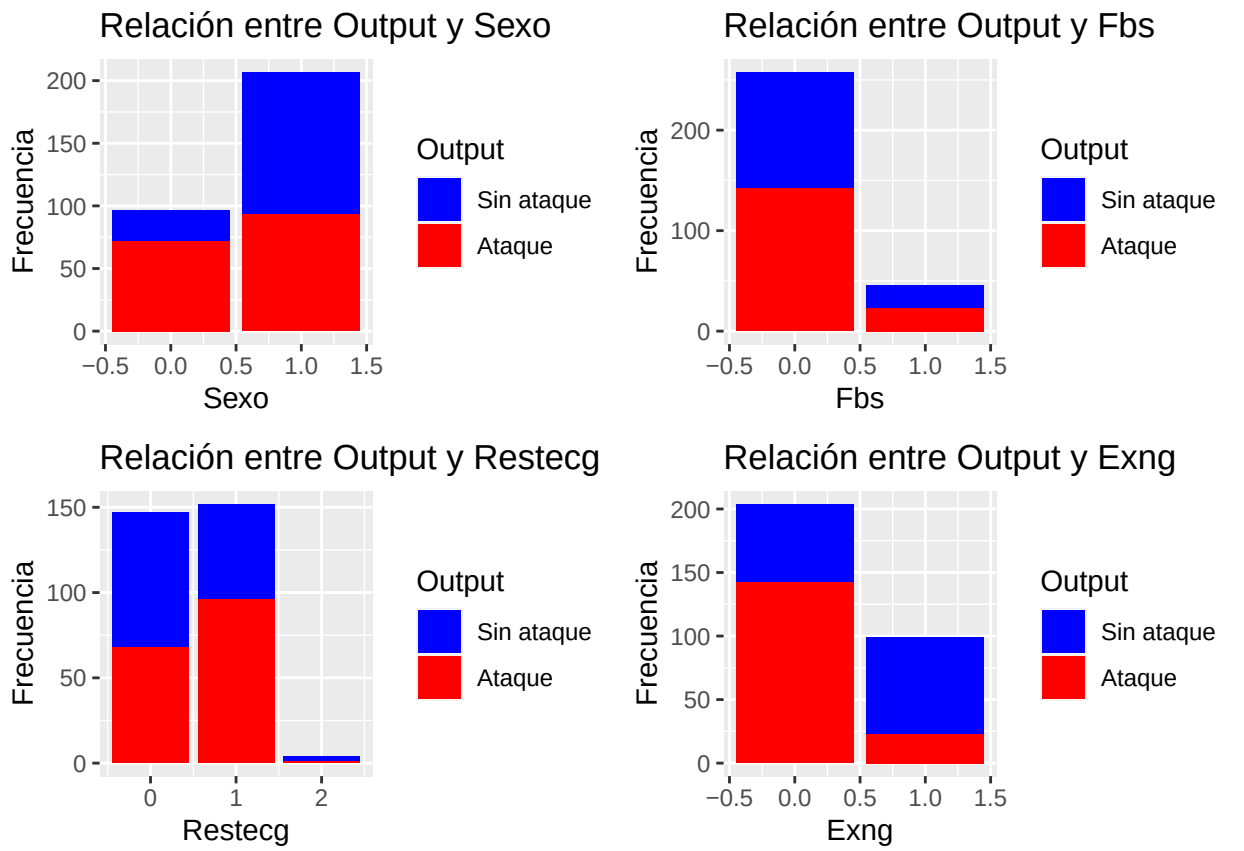
El valor p obtenido es extremadamente bajo (5.781e-05), lo que indica que hay una diferencia significativa en las medias de la variable "age" entre los dos grupos definidos por la variable objetivo "output". Esto sugiere que la edad puede ser un predictor importante para determinar la presencia de enfermedad cardíaca.

thalach

```
##
##  Welch Two Sample t-test
##
## data:  thalachh by output
## t = -7.953, df = 269.9, p-value = 5.019e-14
## alternative hypothesis: true difference in means between group 0 and group 1 is not equal to 0
## 95 percent confidence interval:
##  -24.15912 -14.57132
## sample estimates:
## mean in group 0 mean in group 1
##        139.1014        158.4667
```

El valor p obtenido es extremadamente bajo (5.019e-14), lo que indica que hay una diferencia significativa en las medias de la variable "thalachh" entre los dos grupos definidos por la variable objetivo "output". Esto sugiere que la frecuencia cardíaca máxima puede ser un predictor importante para determinar la presencia de enfermedad cardíaca.

4.3.2 Relación "output" vs "sex", "fbs", "restecg" y "exng"



sex: Se puede observar una diferencia altamente significativa entre ambas categorías. Esto sugiere que el género puede ser un buen predictor de la variable objetivo "output".

fbs: En el caso de la variable "fbs", podemos observar que la proporción de individuos con nivel de azúcar en sangre en ayunas superior a 120 mg/dl es muy similar en ambas categorías. No parece haber una diferencia significativa en la proporción de individuos con y sin un nivel elevado de azúcar en sangre en relación con la variable objetivo "output". Esto sugiere que la variable "fbs" puede tener un poder predictivo limitado en relación con la variable objetivo.

restecg: Podemos observar una diferencia significativa en las proporciones entre las diferentes categorías. Esto indica que el resultado del electrocardiograma en reposo (restecg) puede ser un factor importante para predecir la variable objetivo "output".

exng: Se puede observar una diferencia altamente significativa entre ambas categorías. Esta diferencia en las proporciones sugiere que la presencia o ausencia de angina inducida por ejercicio puede ser un predictor fuerte de la variable objetivo "output".

La prueba de **chi cuadrado** es una prueba estadística utilizada para evaluar la independencia entre dos variables categóricas. En nuestro caso, estamos interesados en analizar la independencia entre la variable objetivo "output" y las variables "sex", "fbs", "restecg" y "exng".

Realizamos la prueba de chi cuadrado para determinar si existe una asociación significativa entre estas variables y la variable objetivo. Si el valor de p obtenido en la prueba de chi cuadrado es menor que un nivel de significancia predefinido (por ejemplo, 0.05), podemos concluir que hay evidencia suficiente para rechazar la hipótesis nula de independencia y afirmar que existe una asociación significativa entre las variables.

Hacemos esta prueba para evaluar si las variables "sex", "fbs", "restecg" y "exng" son buenos predictores de la variable objetivo "output". Si encontramos una asociación significativa, esto indicaría que estas vari-

ables tienen un impacto en la ocurrencia de eventos cardíacos y podrían ser consideradas como predictores relevantes en un análisis más detallado.

```
##   Variable ChiSquared df      p_value
## 1      sex   22.71723  1 1.876778e-06
## 2      fbs   55.94455  1 7.444281e-01
## 3     exng   55.94455  1 7.454409e-14
```

sex: se obtuvo un valor p de 1.88e-06, lo cual es mucho menor que el umbral de significancia de 0.05. Esto indica que hay una asociación altamente significativa entre el género y la presencia de enfermedad cardíaca en la muestra. El valor p tan bajo sugiere que el género puede ser un predictor importante de la enfermedad cardíaca, ya que la asociación observada no es probable que sea aleatoria. Estos resultados resaltan la relevancia del género al considerar el riesgo y la predicción de enfermedades cardíacas.

fbs: se obtuvo un valor p de 0.7444, el cual es mayor que el umbral de significancia de 0.05. Esto indica que no hay evidencia suficiente para rechazar la hipótesis nula de independencia entre el nivel de azúcar en sangre en ayunas y la presencia de enfermedad cardíaca. En otras palabras, el valor p alto sugiere que es probable que cualquier asociación observada entre el nivel de azúcar en sangre en ayunas y la enfermedad cardíaca sea debido al azar en lugar de una relación verdadera. Por lo tanto, la variable "fbs" puede no ser un predictor fuerte de la presencia de enfermedad cardíaca en este conjunto de datos.

exng: Se obtuvo un valor p extremadamente bajo, prácticamente cero ($p < 0.001$). Esto indica que hay una asociación estadísticamente significativa entre la presencia de angina inducida por ejercicio y la enfermedad cardíaca. El valor p tan bajo sugiere que es altamente improbable que la asociación observada se deba al azar. Por lo tanto, la variable "exng" puede considerarse un buen predictor de la presencia de enfermedad cardíaca en este conjunto de datos, ya que existe una fuerte evidencia de que la presencia de angina inducida por ejercicio está relacionada con la enfermedad cardíaca.

Debido a que una de las categorías de la variable "restecg" no está bien representada en los datos (Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria), es recomendable realizar un análisis de asociación utilizando la **prueba exacta de Fisher** en lugar de la prueba de chi cuadrado.

La prueba exacta de Fisher es una prueba no paramétrica utilizada para evaluar la asociación entre dos variables categóricas cuando se cumplen ciertas condiciones, como en nuestro caso, donde una de las categorías tiene un número de casos muy bajo.

```
##
##  Fisher's Exact Test for Count Data
##
## data:  tabla_contingencia
## p-value = 0.6308
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.4308961 1.6975867
## sample estimates:
## odds ratio
##  0.8544825
```

El resultado del análisis de asociación utilizando la prueba exacta de Fisher para la variable "restecg" es el siguiente:

- Valor p: 0.6308
- Hipótesis alternativa: la razón de probabilidades verdadera no es igual a 1
- Intervalo de confianza al 95%: (0.4308961, 1.6975867)
- Estimación de la razón de probabilidades: 0.8544825

El valor p obtenido indica la probabilidad de obtener los resultados observados o resultados más extremos, asumiendo que no hay asociación entre la variable "restecg" y la variable objetivo "output". En este caso, el valor p (0.6308) es mayor que el nivel de significancia comúnmente utilizado (0.05), lo que sugiere que no hay suficiente evidencia para rechazar la hipótesis nula de independencia entre las variables.

En función de los resultados obtenidos hasta el momento, no se puede afirmar que la variable "restecg" sea un predictor sólido para determinar la probabilidad de tener un ataque cardíaco. Es recomendable realizar análisis adicionales y considerar otras variables para evaluar su capacidad predictiva de manera más completa.

5. Respresentación de los resultados

Este apartado se ha respondido a lo largo de la práctica,

6. Resolución del problema

En conclusión, el análisis realizado ha proporcionado una visión general sobre las variables estudiadas y su relación con la presencia de enfermedad cardíaca. A continuación se resumen los hallazgos más relevantes:

1. Edad (age): Se observó una diferencia significativa en las edades entre los grupos con y sin enfermedad cardíaca. Esto sugiere que la edad puede ser un factor importante en la predicción de la enfermedad cardíaca, siendo más probable que las personas de mayor edad presenten un mayor riesgo.
2. Niveles de colesterol (chol): No se encontró una diferencia significativa en los niveles de colesterol entre los grupos con y sin enfermedad cardíaca. Esto indica que los niveles de colesterol pueden no ser un predictor fuerte de la enfermedad cardíaca en este conjunto de datos.
3. Presión arterial en reposo (trtbps): Se encontró una diferencia significativa en los niveles de presión arterial en reposo entre los grupos con y sin enfermedad cardíaca. Esto sugiere que la presión arterial en reposo puede ser un predictor relevante para determinar la presencia de enfermedad cardíaca.
4. Frecuencia cardíaca máxima alcanzada (thalachh): Se encontró una diferencia significativa en la frecuencia cardíaca máxima entre los grupos con y sin enfermedad cardíaca. Esto indica que la frecuencia cardíaca máxima puede ser un predictor importante para determinar la presencia de enfermedad cardíaca.
5. Género (sex): Se observó una asociación altamente significativa entre el género y la presencia de enfermedad cardíaca. Los resultados sugieren que el género puede ser un predictor importante para la enfermedad cardíaca, con una mayor proporción de mujeres con mayor probabilidad de presentar enfermedad cardíaca en comparación con los hombres.
6. Angina inducida por ejercicio (exng): Se encontró una asociación altamente significativa entre la presencia de angina inducida por ejercicio y la enfermedad cardíaca. Esto indica que la presencia de angina inducida por ejercicio puede ser un predictor importante de la enfermedad cardíaca.
7. Nivel de azúcar en sangre en ayunas (fbs): No se encontró evidencia suficiente para sugerir una asociación entre el nivel de azúcar en sangre en ayunas y la presencia de enfermedad cardíaca. Esto indica que el nivel de azúcar en sangre en ayunas puede no ser un predictor fuerte de la enfermedad cardíaca en este conjunto de datos.
8. Resultados del electrocardiograma en reposo (restecg): No se encontró suficiente evidencia para sugerir una asociación entre los resultados del electrocardiograma en reposo y la presencia de enfermedad cardíaca. Sin embargo, se debe tener en cuenta la baja representación de una de las categorías de esta variable en el conjunto de datos.

En resumen, los resultados indican que variables como la edad, la frecuencia cardíaca máxima, el género y la presencia de angina inducida por ejercicio pueden ser factores importantes en la predicción de la enfermedad cardíaca. Estos hallazgos pueden ser útiles para identificar y evaluar el riesgo de enfermedad cardíaca en pacientes, proporcionando información relevante para la toma de decisiones médicas y el diseño de estrategias de prevención y tratamiento. Sin embargo, es importante tener en cuenta que el análisis se basa en un conjunto de datos específico y pueden existir otros factores no considerados que también influyen en la enfermedad cardíaca. Por lo tanto, se recomienda realizar análisis adicionales y considerar información clínica adicional para obtener una evaluación más completa y precisa.

7. Código

Código adjuntado en el repositorio.

8. Video

No ha sido posible incluir un video en la presentación del proyecto debido a que no disponía de cámara en mi ordenador. Aunque intenté grabar el video con mi teléfono móvil, no obtuve un resultado profesional y he preferido no incluirlo en el proyecto. Lamento no haber podido ofrecer este tipo de material adicional, pero confío en que el resto de la documentación proporcionada sea suficiente para demostrar el trabajo realizado.

9. Tabla de contribuciones

Me ha parecido innecesario presentar la tabla debido a que no hay varios participantes en la realización de la práctica.

10. Bibliografía

- Calvo M, Subirats L, Pérez D (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Tutorial de Github (https://guides.github.com/activities/hello-world/)
- Squire, Megan (2015). Clean Data. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). Data mining: concepts and techniques. Morgan Kaufmann.
- Jason W. Osborne (2010). Data Cleaning Basics: Best Practices in Dealing with Extreme Scores. Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). Introductory statistics with R. Springer Science & Business Media.