

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

Alfonso Manuel Carvajal, Eider Ibiricu

16 de junio, 2023

1 Descripción del dataset

El dataset *Heart Attack Analysis & Prediction* kaggle contiene datos para realizar una clasificación de pacientes que tengan riesgo de sufrir un ataque al corazón.

Mediante este juego de datos es posible entrenar algoritmos que permitan un diagnóstico para futuros posibles pacientes.

<https://medium.com/mcd-unison/how-to-use-kaggle-api-to-download-datasets-in-r-312179c7a99c>

```
library(readr)
library(kagglr)
kgl_auth(creds_file = 'kaggle.json')

## <request>
## Options:
## * httpauth: 1
## * userpwd: eideribiricupera:2323ed93bf13610b7984e10597bedc72

response <- kgl_datasets_download_all(owner_dataset =
  "rashikrahmanpritom/heart-attack-analysis-prediction-dataset")

download.file(response[["url"]], "data/temp.zip", mode="wb")
unzip_result <- unzip("data/temp.zip", exdir = "data/", overwrite = TRUE)
unzip_result

## [1] "data//heart.csv"          "data//o2Saturation.csv"

heart_attack_data <- read_csv("data/heart.csv")
o2_saturation_data <- read_csv("data/o2Saturation.csv",header=F)

rm(response)
```

Las variables que encontramos en el dataset, según la descripción en *kaggle*:

- **Age:** Edad del individuo. (Variable numérica continua)
- **Sex:** Género del individuo (1 = masculino, 0 = femenino). (Variable categórica binaria)
- **cp:** Tipo de dolor en el pecho (categórica ordinal)
 - Value 1: typical angina
 - Value 2: atypical angina
 - Value 3: non-anginal pain
 - Value 4: asymptomatic

- **trtbps**: Presión arterial en reposo (en mm Hg) (Variable numérica continua)
- **chol**: Colesterol en mg/dl obtenido via sensor BMI (Variable numérica continua)
- **fbs**: Nivel de azúcar en sangre en ayunas (> 120 mg/dl, 1 = verdadero; 0 = falso). (Variable categórica binaria)
- **restecg**: Resultados electrocardiográficos en reposo. (Variable categórica ordinal)
 - Value 0: normal
 - Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
 - Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria
- **thalachh**: Máxima frecuencia cardíaca alcanzada. (Variable numérica continua)
- **exng**: Angina inducida por ejercicio (1 = sí; 0 = no). (Variable categórica binaria)
- **oldpeak**: Pico anterior, cambios en el segmento ST en un ECG (Variable numérica continua)
- **slp**: La pendiente del segmento ST en el pico de ejercicio (Variable numérica continua)
- **caa**: Número de vasos principales coloreados por fluoroscopia. (0-4) (categorical)
- **thall**: Talio en sangre.(Thallium Stress Test)(numerica)
- **output**: Diagnóstico de enfermedad cardíaca (estado del objetivo) (0 = menos probabilidad de ataque al corazón, 1 = más probabilidad de ataque al corazón). (Variable categórica binaria)

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

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

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

La recopilación y elección de información constituyen etapas fundamentales en cualquier tarea de análisis de datos. En relación a nuestro conjunto de información, estos procesos significarán identificar los factores que resultan más significativos para estimar un infarto al miocardio y optar por aquellos que nos brinden la mayor utilidad en nuestra investigación.

Dado que todas las variables en este conjunto de datos están directamente vinculadas a la salud cardiovascular y los riesgos asociados, todas podrían considerarse pertinentes. No obstante, puede que no todas estas variables contribuyan de la misma manera a la capacidad predictiva de un modelo de estimación de infartos al miocardio.

Por ejemplo, las variables **age**, **sex**, **cp**, **trtbps**, **chol**, **fbs**, **restecg**, **thalachh**, **exng**, **oldpeak**, **slp**, **caa**, y **thall** son todos posibles factores de riesgo para un infarto al miocardio y por lo tanto son de importancia para nuestro estudio. La variable **output** es la que nos gustaría pronosticar.

En consecuencia, el paso inicial en nuestro estudio será llevar a cabo un examen exploratorio de los datos para comprender de mejor manera la distribución y las relaciones de estas variables. Esto puede implicar visualizar la distribución de la información, calcular estadísticas descriptivas y analizar las correlaciones entre las diversas variables.

```
head(heart_attack_data) %>%
  kable_setup %>%
  kable_paper(full_width = F)%>%
  column_spec(c(3,7,14), width = "2 cm") %>%
  column_spec(c(1,5,9,12,13), width = "0.8 cm") %>%
  row_spec(0,bold=TRUE)
```

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
57	0	0	120	354	0	1	163	1	0.6	2	0	2	1
57	1	0	140	192	0	1	148	0	0.4	1	0	1	1

3 Limpieza de los datos

Primero asignamos los tipos de datos a cada variable. En los **comentarios** del dataset, hemos encontrado definiciones de los campos que nos ayudan a determinar el tipo de los datos:

```
# Data types

data <- heart_attack_data %>%
  mutate(
    sex = factor(sex, levels=c(1,0),labels = c("male","female")),
    cp = factor(cp,
      levels=c(0,1,2,3),
      labels=c("typical angina","atypical angina","non-anginal pain","asymptomatic")),
    fbs = factor(fbs,levels=c(0,1),labels = c(F,T)),
    restecg = factor(restecg,levels = c(0,1,2),
      labels = c("normal","ST-T wave abnormality","left ventricular hypertrophy")),
    exng = factor(exng),
    slp = factor(slp,levels = c(0,1,2),
      labels = c("unsloping","flat","downsloping")),
    caa = factor(caa),
    thall = factor(thall,
      levels = c(1,2,3),
      labels = c("fixed defect","normal","reversable defect")),
    output = factor(output,levels=c(0,1),
      labels = c("less chance of heart attack","more chance of heart attack"))
  )
```

Hacemos un resumen de los datos para identificar posibles valores nulos o atípicos. Esta tabla también nos permite entender los rangos en los que se mueven las variables.

También revisaremos si hay registros repetidos.

```
# Summary
summary(data)
```

##	age	sex	cp	trtbps
##	Min. :29.00	male :207	typical angina :143	Min. : 94.0
##	1st Qu.:47.50	female: 96	atypical angina : 50	1st Qu.:120.0
##	Median :55.00		non-anginal pain: 87	Median :130.0
##	Mean :54.37		asymptomatic : 23	Mean :131.6
##	3rd Qu.:61.00			3rd Qu.:140.0
##	Max. :77.00			Max. :200.0
##	chol	fbs	restecg	thalachh
##	Min. :126.0	FALSE:258	normal :147	Min. : 71.0
##	1st Qu.:211.0	TRUE : 45	ST-T wave abnormality :152	1st Qu.:133.5
##	Median :240.0		left ventricular hypertrophy: 4	Median :153.0
##	Mean :246.3			Mean :149.6
##	3rd Qu.:274.5			3rd Qu.:166.0

```
## Max.      :564.0                                Max.      :202.0
## exng      oldpeak      slp      caa      thall
## 0:204    Min.      :0.00    unsloping : 21    0:175    fixed defect    : 18
## 1: 99    1st Qu.:0.00    flat      :140    1: 65    normal      :166
##          Median :0.80    downsloping:142    2: 38    reversable defect:117
##          Mean   :1.04                                3: 20    NA's          : 2
##          3rd Qu.:1.60                                4: 5
##          Max.   :6.20
##
##          output
## less chance of heart attack:138
## more chance of heart attack:165
##
##
##
##
```

```
# Duplicates
data %>%
  unique() %>%
  nrow()
```

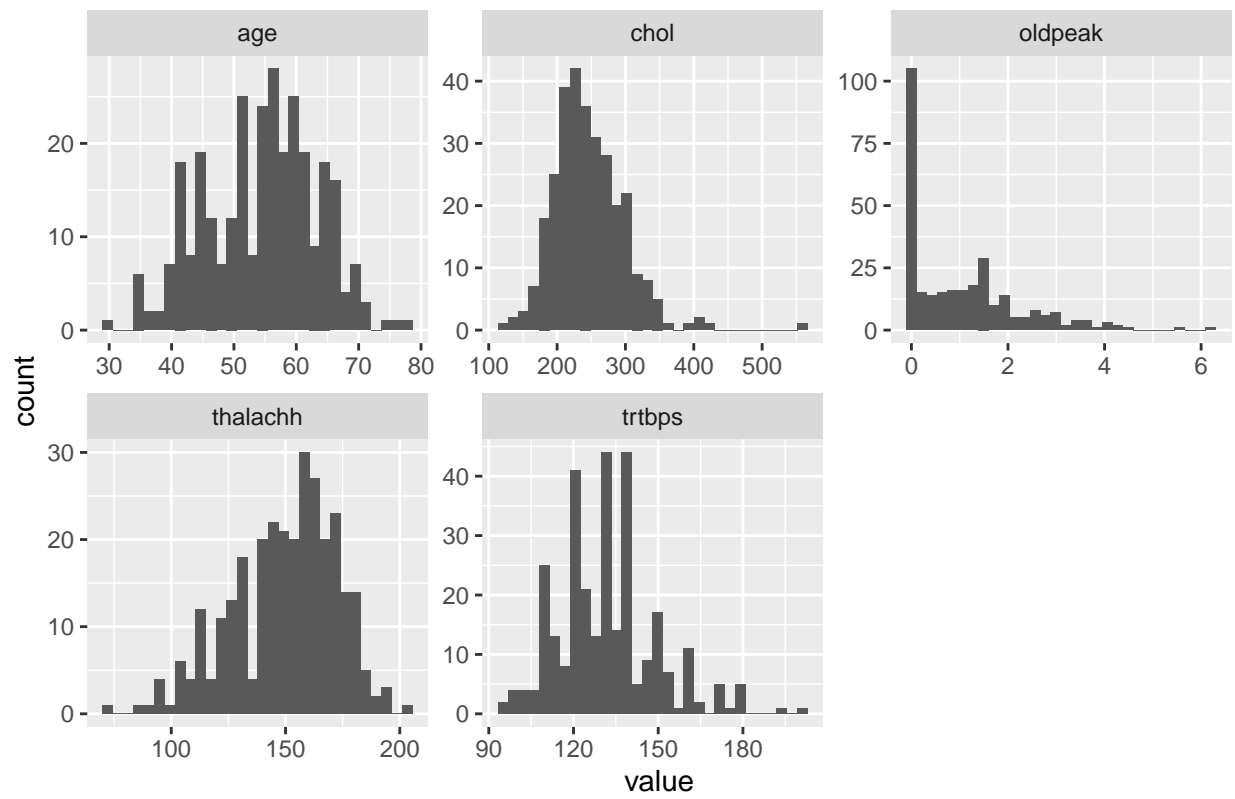
```
## [1] 302
```

```
data <- data %>%
  unique()
```

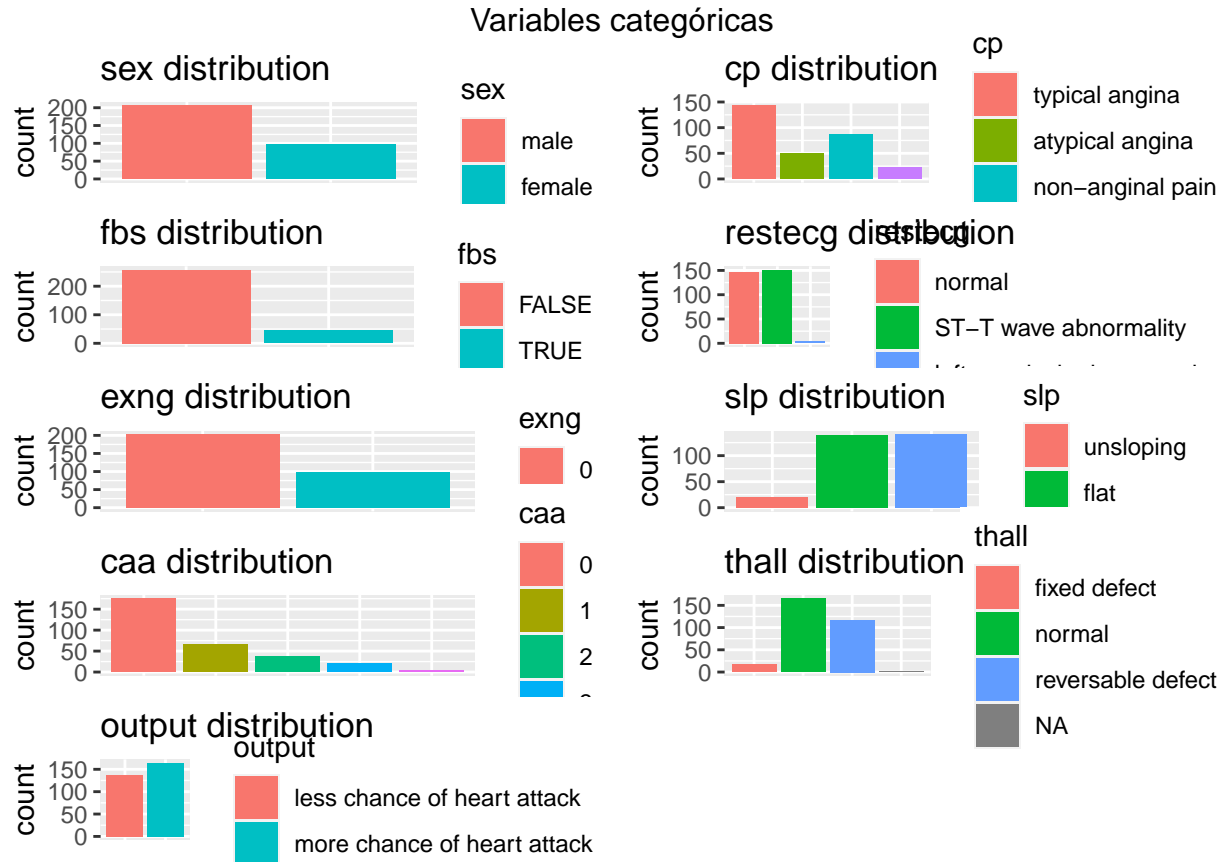
3.1 ¿Los datos contienen ceros o elementos vacíos?

Representamos las distribuciones de las variables:

Variables numéricas



En el caso de las variables numéricas



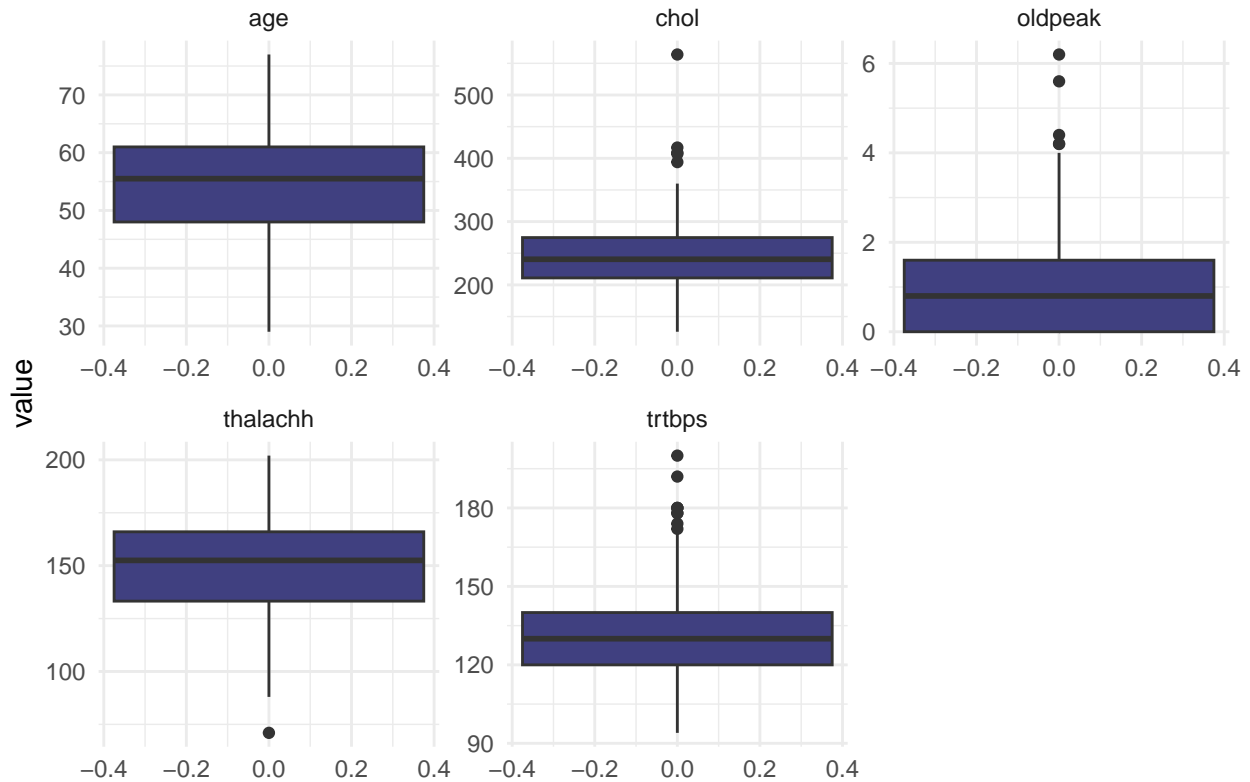
Vemos que la variable *thall* tiene valores nulos. Podemos considerar que estos registros como normales, por lo que les asignaremos el valor 2:

```
data <- data %>%
  mutate(
    thall = factor(if_else(is.na(as.numeric(thall)), 2, as.numeric(thall)),
                  levels = c(1, 2, 3),
                  labels = c("fixed defect", "normal", "reversable defect"))
  )
```

3.2 Identifica y gestiona los valores extremos.

Estudiaremos los boxplots de las variables numéricas para observar si existen valores atípicos:

Boxplots de variables numéricas

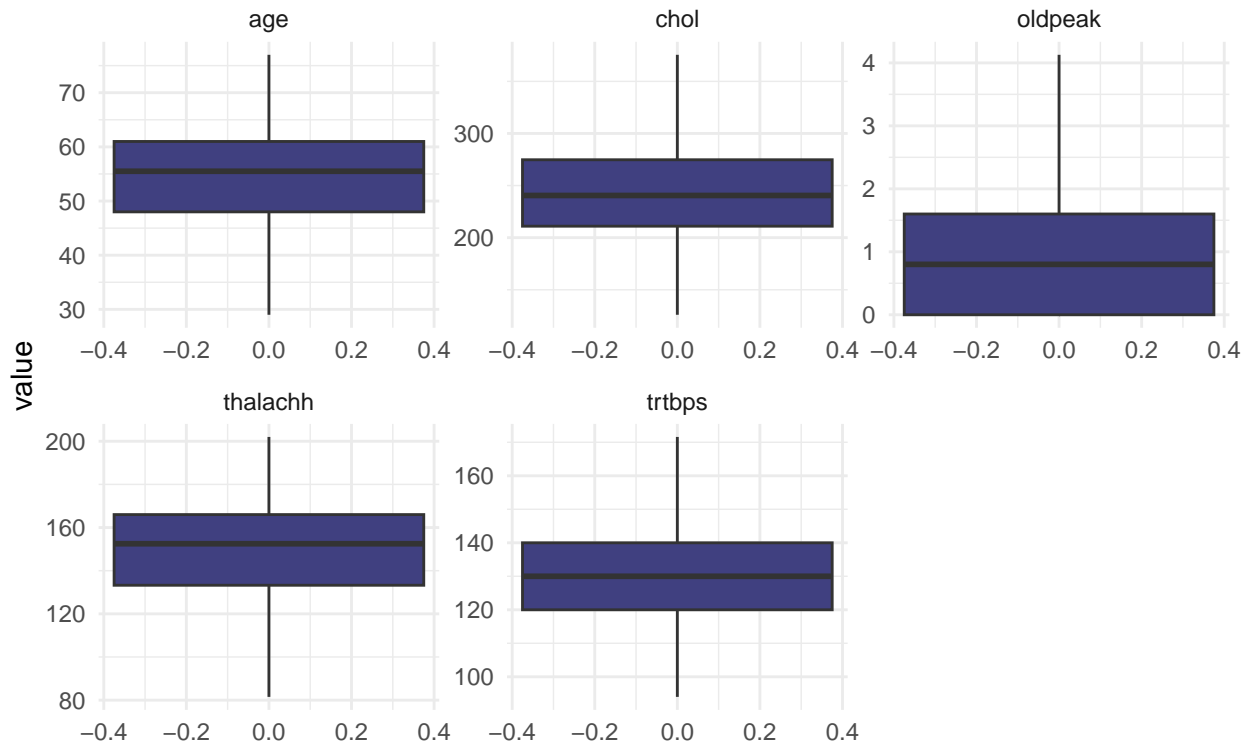


Vemos que todas las variables menos *age* muestran valores que están alejados de las medias y la mayoría de observaciones. Esta gráfica identifica como atípico el valor que $x < Q_1 - IQR \cdot 1.58$ o $x > Q_3 + IQR \cdot 1.58$ donde $IQR = Q_3 - Q_1$. La manera en la que vamos a tratarlos es asignarles el valor más cercano para que no sean considerados *outliers*.

```
for (col in colnames(data_num)){
  value = data[[col]][data[[col]] %in% boxplot.stats(data[[col]])$out]
  res <- quantile(data[[col]], probs = c(0,0.25,0.5,0.75,1))
  q1 <- res[[2]]
  q3 <- res[[4]]
  iqr <- q3 - q1
  min_thshld <- q1 - 1.58*iqr
  max_thshld <- q3 + 1.58*iqr
  data[[col]][data[[col]] < min_thshld] = min_thshld
  data[[col]][data[[col]] > max_thshld] = max_thshld
}
```

```
data %>%
  select_if(is.numeric) %>%
  pivot_longer(colnames(data_num)) %>%
  as.data.frame() %>%
  ggplot(aes(y = value)) + # Draw each column as histogram
  geom_boxplot(fill = "#404080", coef = 1.58) +
  theme_minimal() +
  facet_wrap(~ name, scales = "free") +
  labs(title="Boxplots de variables numéricas \n Outliers corregidos")
```

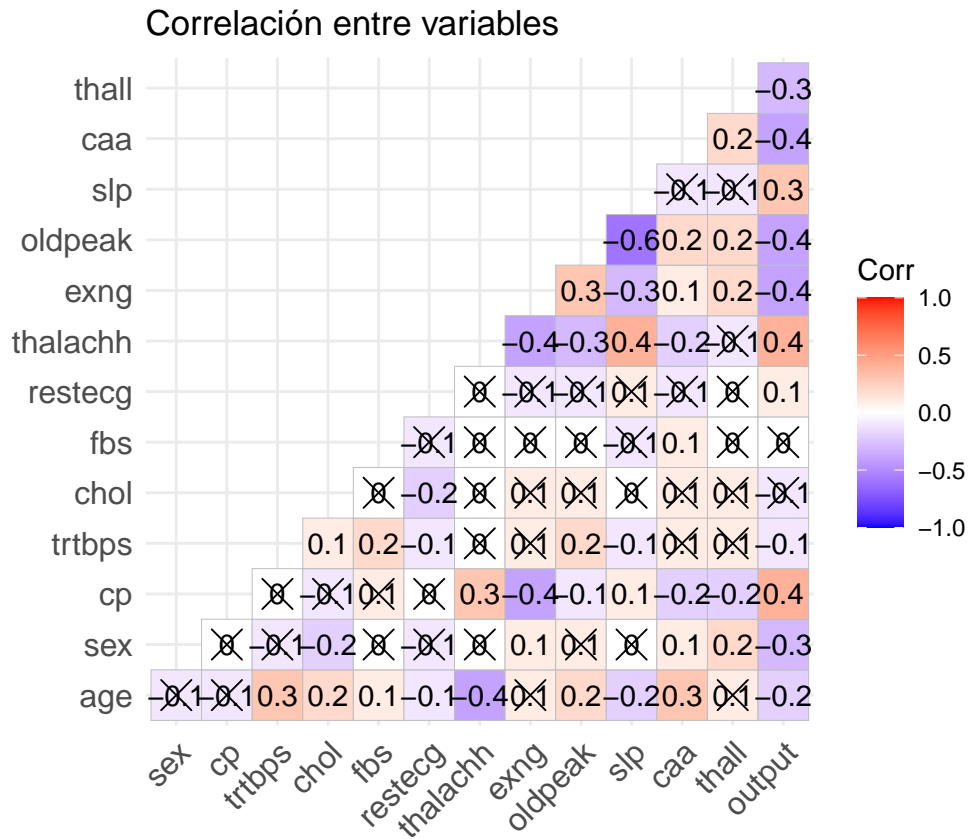
Boxplots de variables numéricas Outliers corregidos



4 Análisis de los datos

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

Lo primero que haremos será estudiar la correlación entre las variables y también la que tienen con la variable *output*.



Vemos que la variable *output* no tiene correlación significativa con *fbs* y *trtbps*, y con las que mayor correlación tiene son *caa*, *oldpeak*, *exng*, *thalachh* y *cp*.

Respecto al resto de variables, vemos parejas relacionadas como *oldpeak* y *slp*, *cp* y *exng*, *age* y *thalachh*, *thalachh* y *slp*.

Por otro lado, resulta curioso observar el la variable *chol* no se relaciona más que con *restecg*

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Vamos a analizar las variables recién mencionadas, para determinar si son aptas para aplicar tests de comparación de grupos.

```
data_compare <- data %>%
  select(
    age,
    cp,
    fbs,
    thalachh,
    exng,
    oldpeak,
    slp,
    caa,
    output
  )
head(data_compare)
```

age	cp	fbs	thalachh	exng	oldpeak	slp	caa	output
63	asymptomatic	TRUE	150	0	2.3	unsloping	0	more chance of heart attack
37	non-anginal pain	FALSE	187	0	3.5	unsloping	0	more chance of heart attack
41	atypical angina	FALSE	172	0	1.4	downsloping	0	more chance of heart attack
56	atypical angina	FALSE	178	0	0.8	downsloping	0	more chance of heart attack
57	typical angina	FALSE	163	1	0.6	downsloping	0	more chance of heart attack
57	typical angina	FALSE	148	0	0.4	flat	0	more chance of heart attack

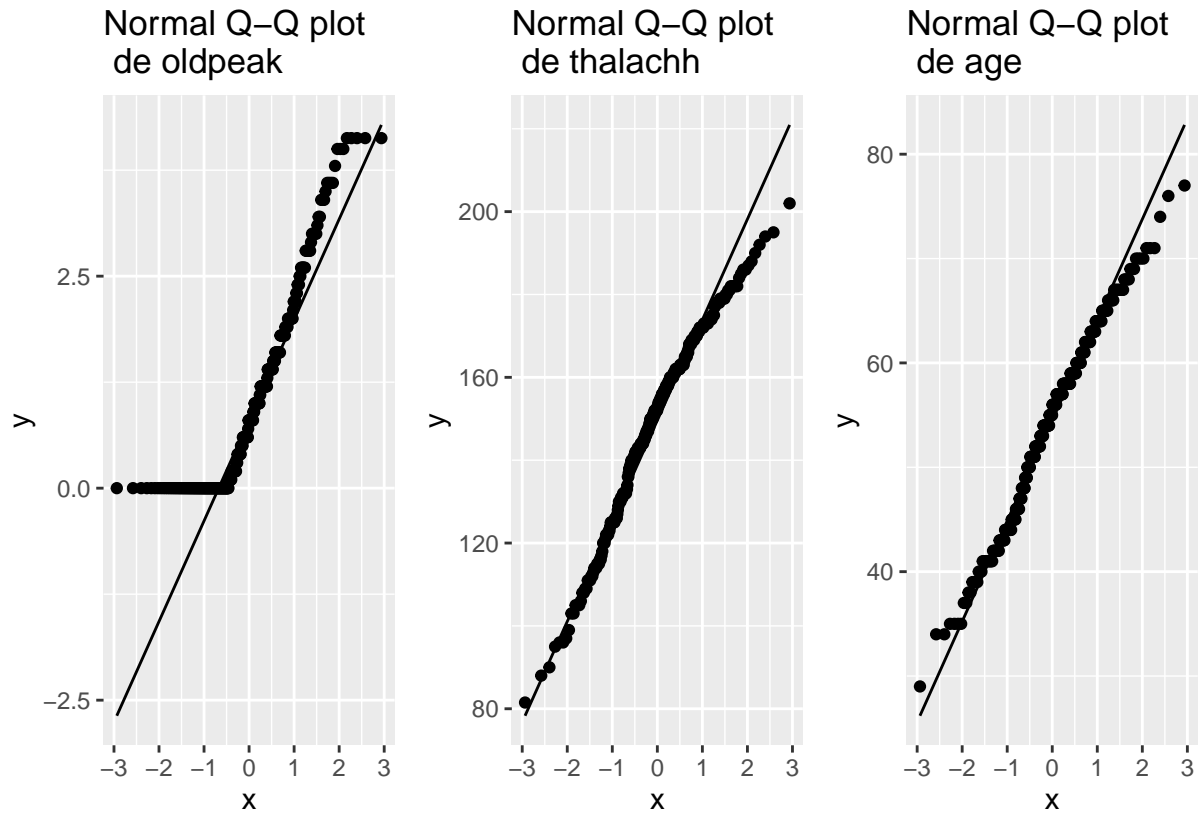
4.2.1 Normalidad

Vamos a analizar si las variables numéricas tienen distribuciones normales. Para ello aplicamos el test *Shapiro-Wilk*:

```
##
##  Shapiro-Wilk normality test
##
## data:  data_compare$oldpeak
## W = 0.85269, p-value = 2.615e-16

##
##  Shapiro-Wilk normality test
##
## data:  data_compare$thalachh
## W = 0.97705, p-value = 9.187e-05

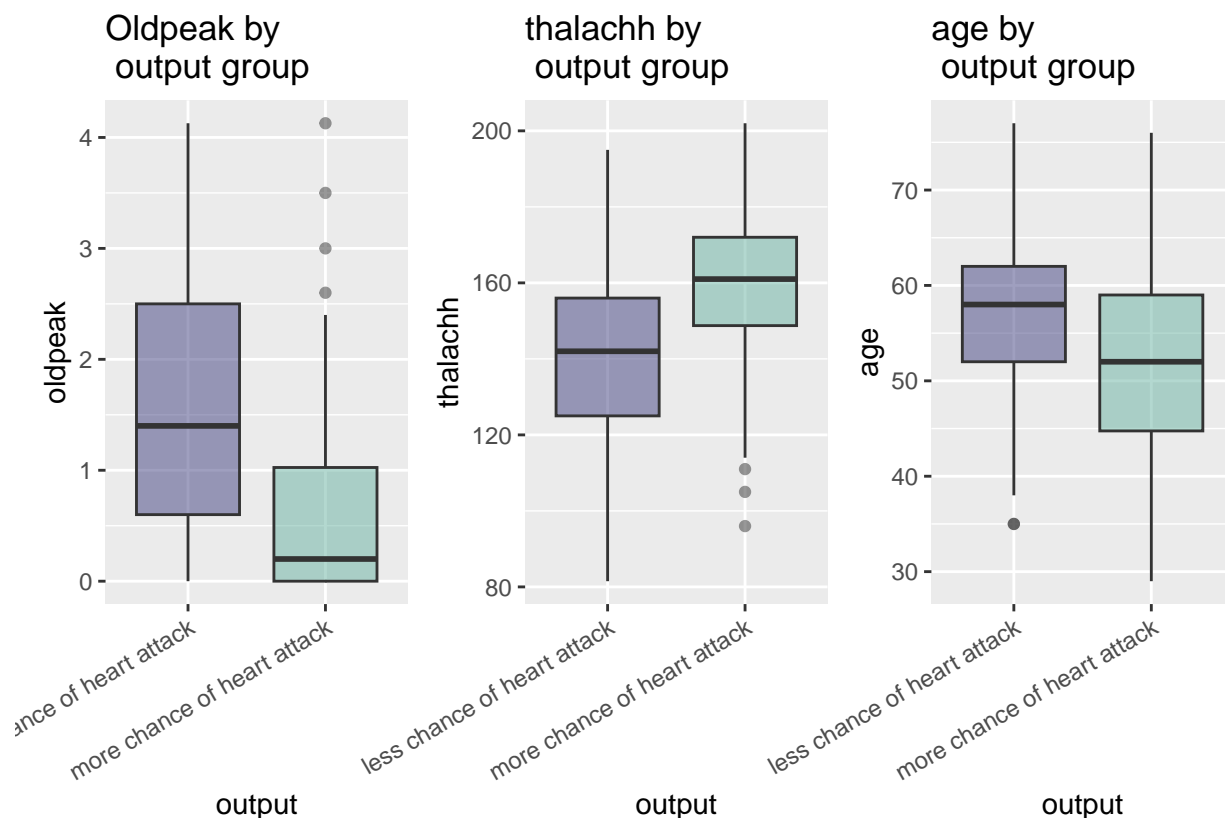
##
##  Shapiro-Wilk normality test
##
## data:  data_compare$age
## W = 0.98664, p-value = 0.006745
```



El test Shapiro-Wilk parte de la hipótesis de que la población está normalmente distribuida, por lo que si el resultado del p-valor es mayor que 0.05, podemos aceptar dicha hipótesis con un 95% de confianza.

En el caso anterior, vemos que no se puede aceptar que tanto *oldpeak* como *thalachh* estén normalmente distribuidas.

Podemos visualizar que las medias por los dos grupos objetivo no son iguales:



4.2.2 Homogeneidad de la varianza

Aplicando el test de Levene, analizamos si las varianzas de los grupos de la variable objetivo son homogéneas para las tres variables numéricas que estamos analizando:

```
print(leveneTest(oldpeak ~ output, data = data_compare,))

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value    Pr(>F)
## group  1 33.028 2.232e-08 ***
##      300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

print(leveneTest(thalachh ~ output, data = data_compare,))

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value    Pr(>F)
## group  1  5.0378 0.02553 *
##      300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

print(leveneTest(age ~ output, data = data_compare,))

## Levene's Test for Homogeneity of Variance (center = median)
##      Df F value    Pr(>F)
## group  1  7.6349 0.006079 **
```

```
##          300
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

En los tres casos, obtenemos que $p_{value} < 0.05$, por lo que rechazamos la hipótesis nula y concluimos que ninguna tiene varianzas homogéneas dentro de los grupos.

En resumen, ninguna de las variables analizadas cumplen con el supuesto de normalidad y homogeneidad de varianza.

4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos.

4.3.1 Prueba de chi-cuadrado

Ahora vamos a analizar la correlación entre las variables categóricas y la variable objetivo mediante el test *Chi-cuadrado*, el cual indica si existe correlación significativa entre dos variables cualitativas.

```
#categoricas Chi-quadado
# Variable "caa"
table_caa <- table(data_compare$caa, data_compare$output)
print(chisq.test(table_caa))
```

```
##
## Pearson's Chi-squared test
##
## data:  table_caa
## X-squared = 73.69, df = 4, p-value = 3.771e-15
```

```
# Variable "exng"
table_exng <- table(data_compare$exng, data_compare$output)
print(chisq.test(table_exng))
```

```
##
## Pearson's Chi-squared test with Yates' continuity correction
##
## data:  table_exng
## X-squared = 55.456, df = 1, p-value = 9.556e-14
```

```
# Variable "cp"
table_cp <- table(data_compare$cp, data_compare$output)
print(chisq.test(table_cp))
```

```
##
## Pearson's Chi-squared test
##
## data:  table_cp
## X-squared = 80.979, df = 3, p-value < 2.2e-16
```

1. Variable **cp**: El valor p es menor que 2.2e-16, que es menor que 0.05. Por lo tanto, rechazamos la hipótesis nula y concluimos que hay una asociación significativa entre las categorías de **cp**.
2. Variable **caa**: El valor p es 3.771e-15, que es menor que 0.05. Por lo tanto, rechazamos la hipótesis nula y concluimos que hay una asociación significativa entre las categorías de **caa**.
3. Variable **exng**: El valor p es 9.556e-14, que es menor que 0.05. Por lo tanto, rechazamos la hipótesis nula y concluimos que hay una asociación significativa entre las categorías de **exng**.

En resumen, todas las variables categóricas parecen tener una relación significativa con la variable de resultado.

4.3.2 Modelo de regresión logística

Al tratarse de una variable objetivo de dos clases, vamos ajustar un modelo de regresión logística.

```
# wilcox oldpeak
res <- wilcox.test(oldpeak ~ output, data = data_compare)
print(res)

##
## Wilcoxon rank sum test with continuity correction
##
## data: oldpeak by output
## W = 16721, p-value = 3.397e-13
## alternative hypothesis: true location shift is not equal to 0

# kruskal thalachh
res <- kruskal.test(thalachh ~ output, data = data_compare)
print(res)

##
## Kruskal-Wallis rank sum test
##
## data: thalachh by output
## Kruskal-Wallis chi-squared = 54.719, df = 1, p-value = 1.391e-13

# Regresión logística utilizando 'cp', 'caa' y 'exng' como predictores
model <- glm(output ~ cp + caa + exng, data = data_compare, family = binomial)
summary(model)

##
## Call:
## glm(formula = output ~ cp + caa + exng, family = binomial, data = data_compare)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.2751  -0.5887   0.3953   0.5236   2.3600
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)      0.6370     0.2973   2.143 0.032131 *
## cpatypical angina  1.8384     0.4967   3.701 0.000215 ***
## cpnon-anginal pain  1.8728     0.3974   4.713 2.45e-06 ***
## cpasymptomatic    1.2807     0.5694   2.249 0.024493 *
## caa1              -1.9302     0.3901  -4.947 7.52e-07 ***
## caa2              -2.3021     0.5238  -4.395 1.11e-05 ***
## caa3              -2.9253     0.7326  -3.993 6.52e-05 ***
## caa4              -0.1931     1.5149  -0.127 0.898573
## exng1             -1.4281     0.3554  -4.018 5.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 416.42  on 301  degrees of freedom
## Residual deviance: 259.31  on 293  degrees of freedom
## AIC: 277.31
##
```

Number of Fisher Scoring iterations: 5

5 Conclusiones

Revisión de correlaciones: En nuestro análisis, descubrimos que ‘output’ no parece estar fuertemente ligado con ‘fbs’ o ‘trtbps’. Sin embargo, observamos una fuerte relación entre ‘output’ y otras variables como ‘caa’, ‘oldpeak’, ‘exng’, ‘thalachh’ y ‘cp’. Esto podría indicar que estas últimas variables tienen más relevancia para prever el riesgo de ataque cardíaco.

Distribución y varianza de las variables: Según las pruebas de Shapiro-Wilk, ‘oldpeak’ y ‘thalachh’ no se distribuyen normalmente. Además, el test de Levene indica que la varianza entre ‘oldpeak’, ‘thalachh’ y ‘age’ no es uniforme.

Análisis de Chi-cuadrado: Según las pruebas de chi-cuadrado, existe una relación significativa entre ‘output’ y las variables categóricas ‘cp’, ‘caa’ y ‘exng’.

Regresión logística: Nuestro modelo sugiere que ‘cp’, ‘caa’ y ‘exng’ podrían ser buenos indicadores de ‘output’. Por ejemplo, se encontró que diferentes tipos de angina (‘cp’) y la cantidad de vasos visibles en la fluoroscopia (‘caa’) tienen una relación significativa con el riesgo de ataque cardíaco. Además, la angina inducida por ejercicio (‘exng’) parece estar inversamente relacionada con ‘output’, lo que sugiere que su presencia podría disminuir la probabilidad de un ataque cardíaco.

En resumen, ‘cp’, ‘caa’, ‘exng’, ‘oldpeak’ y ‘thalachh’ podrían ser importantes para predecir la probabilidad de un ataque cardíaco.

References

- Laia Subirats Maté, Mireia Calvo González, Diego Oswaldo Pérez Trenard. 2019. *Introducción a La Limpieza y Análisis de Los Datos*. FUOC.
- Ozdemir, Ozancan. 2022. *An Introduction to Ggplot2*.
- Xie, Dervieux, Yihui, and Emily Riederer. 2022. *R Markdown Cookbook*. Chapman & Hall/CRC.
- Zhu, Hao. 2019. *Create Awesome LaTeX Table with Knitr::kable and kableExtra*. Zhu, Hao.