Práctica 2:

Tratamiento del dataset "Heart Attack Analysis & Prediction"

Lucía Fernández González

1. Descripción del dataset:

Para la realización de esta práctica me decanté por el dataset de Kraggle propuesto en el enunciado "Heart Attack Analysis & Prediction". Se trata de un conjunto de datos compuesto de dos ficheros, en el fichero heart se recogen una serie de 14 variables sobre 303 observaciones (pacientes analizados), los campos son:

- age: edad del paciente, variable entera.
- sex: sexo del paciente. Variable dicotómica que toma los valores 0 en el caso de mujer y 1 en el caso de paciente masculino.
- cp: tipo de dolor en el pecho. Es una variable categórica que toma 4 valores:
  - 0: angina típica
  - 1: angina atípica
  - 2: dolor sin ser de angina
  - 3: asintomático
- trtbps: presión sanguínea en reposo. Medida en mm Hg, es un nº entero.
- chol: nivel de colesterol en sangre. Medido en mg/dl, es una variable entera.
- fbs: es una variable dicotómica que toma el valor 1 si la prueba de azúcar en sangre arroja un resultado superior a 120 mg/dl ó 0 en caso contrario.
- rest_ecg: resultado del electrocardiograma en reposo. Variable categórica que toma los siguientes valores:
  - 0: normal
  - 1: anomalía de la onda ST-T
  - 2: posible o definitiva hipertrofia ventricular izquierda según el criterio de Estes
- thalach: frecuencia cardíaca máxima alcanzada. Es una variable entera.
- exang: angina inducida por el ejercicio. Variable dicotómica que toma el valor 1 en el caso afirmativo y 0 en caso contrario.
- old peak: pico previo. Depresión descendente del segmento ST inducida por el ejercicio. Es una variable decimal.
- Slope: pendiente del segmento ST de ejercicio máximo. Variable categórica que toma los siguientes valores:
  - 0: pendiente ascendente
  - 1: plana

2: pendiente descendente

- caa: número de vasos principales (0-3).
- thall: presencia de talasemia, trastorno sanguíneo hereditario que hace que el cuerpo tenga menos hemoglobina de lo normal, es una variable categórica que toma cuatro valores:
  - 0: nada
  - 1: defecto fijo
  - 2: normal
  - 3: defecto reversible
- output: diagnóstico de enfermedad cardíaca, variable dicotómica que toma los valores:
  - 0: estrechamiento del diámetro < 50%, menos probabilidades de tener un infarto
  - 1: estrechamiento del diámetro > 50%, más probabilidades de tener un infarto

En el fichero o2Saturation se recogen 3.586 observaciones de la variable saturación de oxígeno, desechamos este fichero en nuestro análisis ya que no podemos determinar a qué pacientes pertenecen estas mediciones ni en qué contexto se realizaron.

Este tipo de datasets se estudian de forma frecuente en los entornos sanitarios y pueden servir para la detección precoz, la correcta diagnosis y posterior tratamiento de enfermedades, así como para analizar la eficacia o efectos adversos de un fármaco o tratamiento.

En nuestro caso, nos servirá para construir un modelo predictivo que mida la probabilidad que una persona tiene de sufrir un ataque al corazón en función de sus características personales y otros factores relacionados con su salud cardiovascular, utilizando para ello la regresión logística. Estudiaremos así mismo la correlación existente entre las variables y realizaremos contrastes estadísticos para detectar si existen diferencias significativas entre grupos.

2. **Integración y selección de los datos de interés a analizar.** Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

En la práctica vamos a considerar solamente los registros del fichero heart, no es posible fusionar el fichero heart con el fichero o2Saturation ya que no hay manera de identificar a qué pacientes corresponde cada una de las mediciones del nivel de saturación de oxígeno, no hay ningún campo que permita fusionar ambos ficheros.
Empiezo cargando el fichero heart.csv en RStudio y cambiando el nombre de los campos por otro en castellano con el que resulte más fácil identificarlos.

A continuación compruebo la clase de variable que R asigna a cada campo, cambiando el tipo de variable de los campos sexo, tipo_dolor, azucar, result_electro, angina_ejer, pendiente_ST, talasemia y output de entero a factor, ya que son variables categóricas y depresion_ST la transformo en numérica.
El siguiente paso consiste en elegir el subconjunto de datos del dataset original con el que haré el análisis: elimino registros duplicados (había 1, con lo que resulta un dataset de 302 filas) y, en principio, voy a prescindir de la variable nº de vasos, ya que no parece una característica propia de las personas propensas a sufrir un infarto, aparte de que no está claro su significado.
En este apartado voy a aplicar también una técnica de conversión de datos, como es la discretización, al atributo edad, por si se quiere usar esta variable en futuros análisis. La discretización la realizo manualmente teniendo en cuenta la distribución de frecuencias y al mismo tiempo que la anchura de los intervalos tenga cierta simetría, de forma que los tramos queden equilibrados. En los tramos centrales tomo grupos decenales de edad y en las colas de 15 años. Me ayudo también de la función discretize() con el método equal-frequency para obtener una primera aproximación a los intervalos con frecuencias iguales.

3. Limpieza de los datos.

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

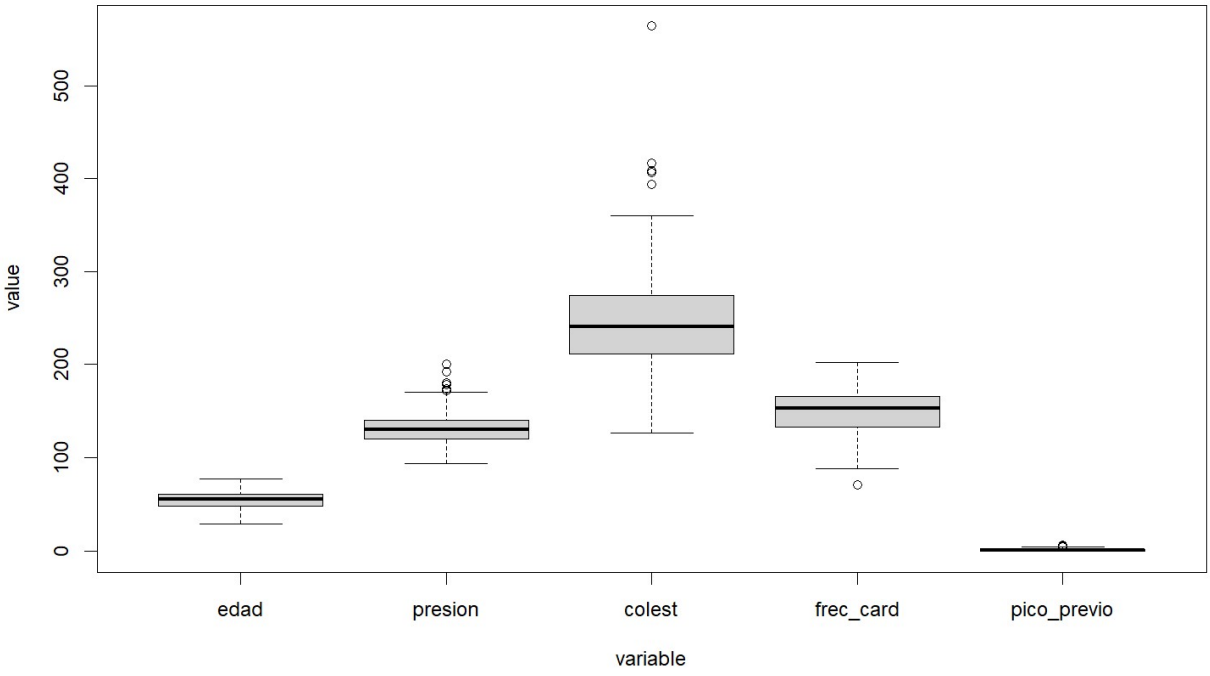
Con el comando status() hago un resumen con las principales características de los datos:
Compruebo que se ha asignado correctamente la clase de cada variable, si hay valores perdidos (NA), o presencia de ceros en las variables numéricas.

variable	q_zeros	p_zeros	q_na	p_na	q_inf	p_inf	type	unique
edad	0	0.00000000	0	0	0	0	integer	41
sexo	96	0.31683168	0	0	0	0	factor	2
tipo_dolor	143	0.47194719	0	0	0	0	factor	4
presion	0	0.00000000	0	0	0	0	integer	49
colest	0	0.00000000	0	0	0	0	integer	152
azucar	258	0.85148515	0	0	0	0	factor	2
result_electro	147	0.48514851	0	0	0	0	factor	3
frec_card	0	0.00000000	0	0	0	0	integer	91
angina_ejer	204	0.67326733	0	0	0	0	factor	2
pico_previo	99	0.32673267	0	0	0	0	numeric	40
pendiente_ST	21	0.06930693	0	0	0	0	factor	3
num_vasos	175	0.57755776	0	0	0	0	integer	5
talasemia	2	0.00660066	0	0	0	0	factor	4
output	138	0.45544554	0	0	0	0	factor	2
grupos_edad	0	0.000000000	0	0	0	0	factor	4

En el dataset hay una alta presencia de valores nulos en la variable pico_anterior, lo que puede interferir en algunos estadísticos para este campo. No hay valores missing

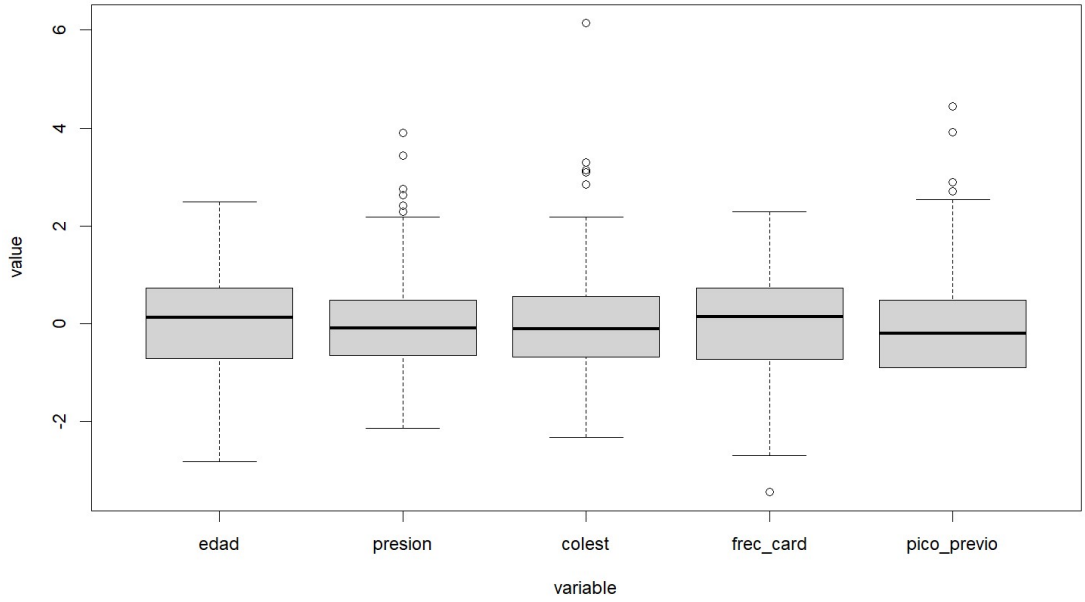
3.2 Identifica y gestiona los valores extremos.

Realizo el diagrama boxplot para visualizar los posibles outliers. Los boxplot o diagramas de caja muestran a simple vista la mediana y los cuartiles de los datos, y también pueden representarse sus valores atípicos. Los últimos valores que no son atípicos quedan representados en los extremos de los bigotes, que en el caso de que no hubiese outliers coincidirían con el máximo y el mínimo de los datos. También proporcionan una visión general de la simetría de la distribución de los datos; si la mediana no está en el centro del rectángulo, la distribución no es simétrica.



A simple vista hay presencia de outliers en las variables presión, colesterol, frecuencia cardíaca y pico previo.

En el siguiente diagrama se muestran los boxplot con los datos puestos en la misma escala para mejorar la comparabilidad. Sobre toda en la variable pico previo se marcan más los valores extremos.



Los atípicos de presión, frecuencia cardíaca y pico previo no los trato ya que se encuentran relativamente próximos entre sí y al bigote, en el caso de colesterol el valor problemático sería 564. Decido no prescindir de él, ya que no es un error al ser factible tener ese nivel de colesterol.

4. Análisis de los datos

En primer lugar, mediante describe(datos) obtengo una breve descripción de las variables, la distribución de frecuencias en las variables categóricas y estadísticos de centralización, posición y dispersión para las variables numéricas (media, percentiles, …), que dan una visión general del conjunto de datos.

Con profiling_num(datos) obtengo una descripción más pormenorizada de las variables numéricas:

variable	mean	std_dev	variation_coef	p_01	p_05	p_25	p_50	p_75	p_95	p_99
edad	54.420530	9.047970	0.1662602	35.00	40.00	48.00	55.5	61.00	68.00	71.00
presion	131.602649	17.563394	0.1334578	100.00	108.00	120.00	130.0	140.00	160.00	180.00
colest	246.500000	51.753489	0.2099533	149.00	175.05	211.00	240.5	274.75	326.95	406.87
frec_card	149.569536	22.903527	0.1531296	95.01	108.05	133.25	152.5	166.00	181.95	191.98
pico_previo	1.043046	1.161452	1.1135193	0.00	0.00	0.00	0.8	1.60	3.40	4.20

variable	skewness	kurtosis	iqr	range_98	range_80
edad	-0.2027299	2.461379	13.00	[35, 71]	[42, 66]
presion	0.7129775	3.887984	20.00	[100, 180]	[110, 152]
colest	1.1416259	7.447929	63.75	[149, 406.87]	[188.4, 308.9]
frec_card	-0.5300219	2.919037	32.75	[95.01, 191.98]	[116, 176.8]
pico_previo	1.2598751	4.522236	1.60	[0, 4.2]	[0, 2.8]

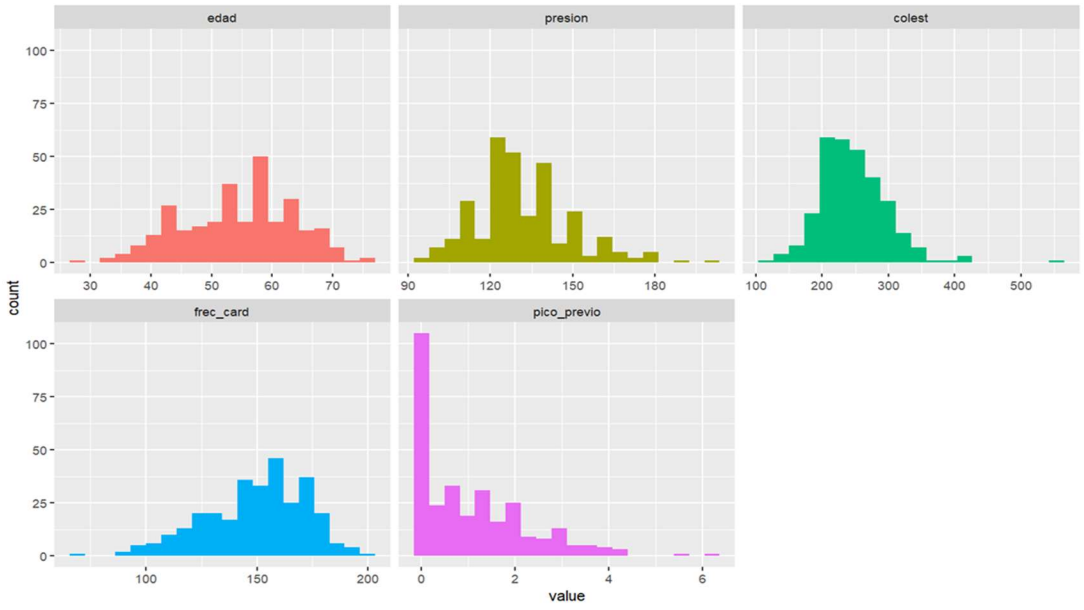
El coeficiente de variación (desv. típica/media) es independiente de la escala utilizada y sirve para medir la dispersión de los datos a la vez que permite establecer comparaciones entre distintas variables. Cuando supera el 30% significa que la media no es representativa de los datos (hay mucha heterogeneidad de datos).
kurtosis: describe las colas de la distribución; dicho en términos simples, un número alto puede indicar la presencia de valores atípicos
iqr: el rango intercuartil es el resultado de observar los percentiles 0.25 y 0.75, e indica, en la misma unidad de la variable, el largo de dispersión del 50% de los valores. Cuanto mayor sea el valor, más dispersa es la variable.

La variable pico_previo presenta una alta variabilidad (coeficiente de variación de 111,35%), eso es debido al alto nº de ceros que tiene, es preferible medirla normalizándola mediante min-max en (1,2):

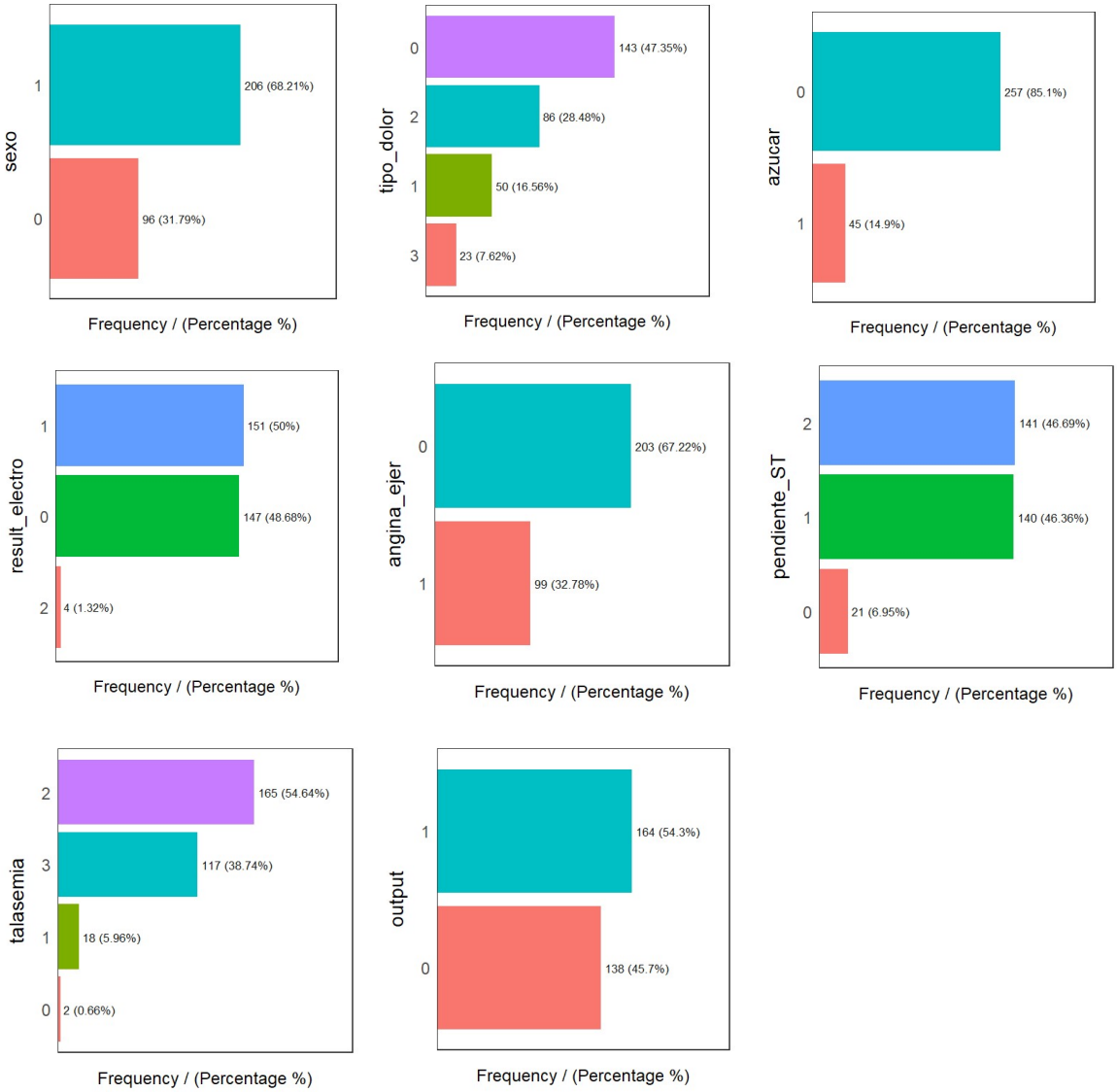
variable	mean	std_dev	variation_coef	p_01	p_05	p_25	p_50	p_75	p_95	p_99	skewness
pp_norm	1.168233	0.187331	0.1603541	1	1	1	1.129032	1.258065	1.548387	1.677419	1.259875

kurtosis	iqr	range_98	range_80
4.522236	0.2580645	[1, 1.6774193]	[1, 1.451612]

Así vemos que su coeficiente de variación es del 16,04%. Además según el rango intercuartílico las que presentan mayor dispersión son colesterol y frecuencia cardíaca.
Las variables que presentan más curtosis son colesterol y pico_previo, debido a la presencia de outliers, como vimos anteriormente. Siendo también la asimetría, que viene dada por el coeficiente skewness más marcada en ambas también, presentando asimetría positiva. Hay ausencia de normalidad en todas las variables.
Esto se distingue mejor en las gráficas:



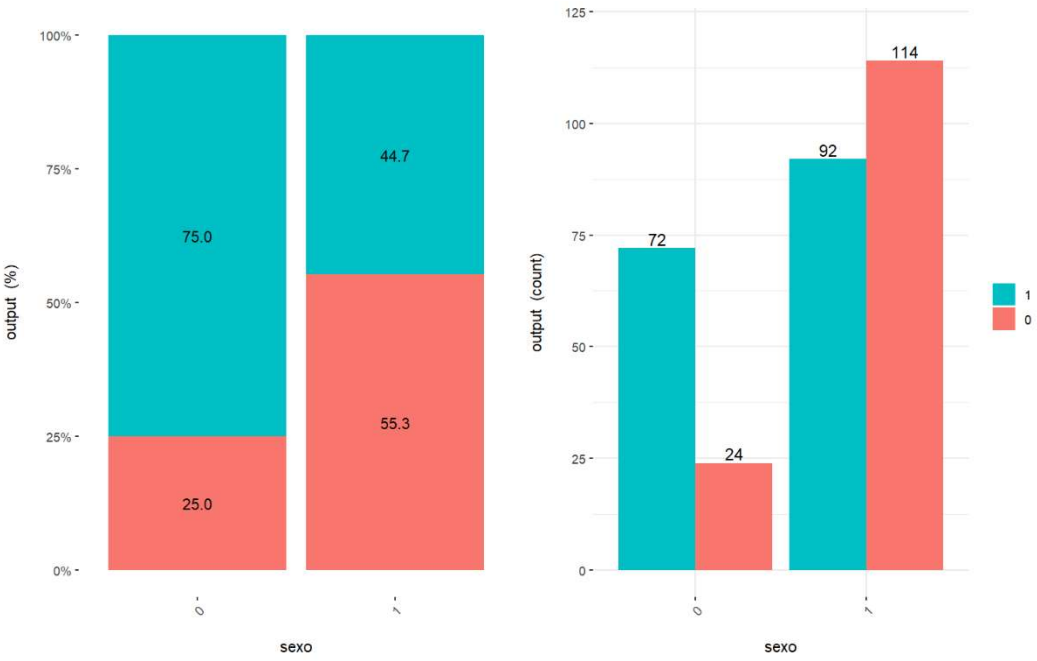
Analizaremos ahora la distribución de frecuencias de las variables categóricas:



Llama la atención que el número de hombres en la muestra es más del doble que el de mujeres.

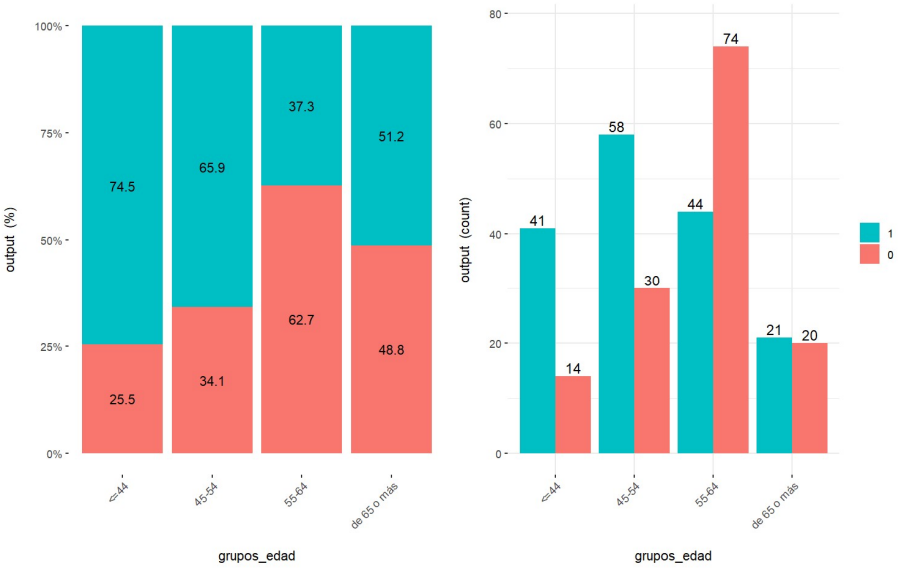
Comparación entre grupos:

- Entre sexo y variable objetivo: probabilidad de infarto



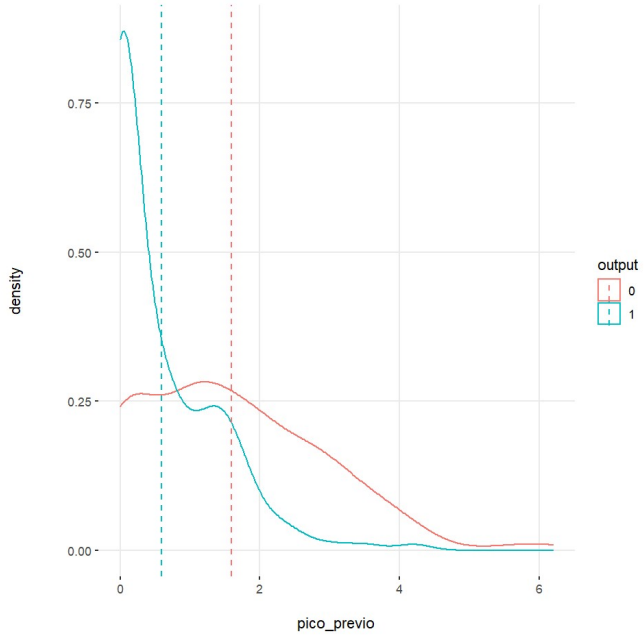
Las mujeres presentan mayor probabilidad de infarto

- Entre grupos de edad y probabilidad de infarto



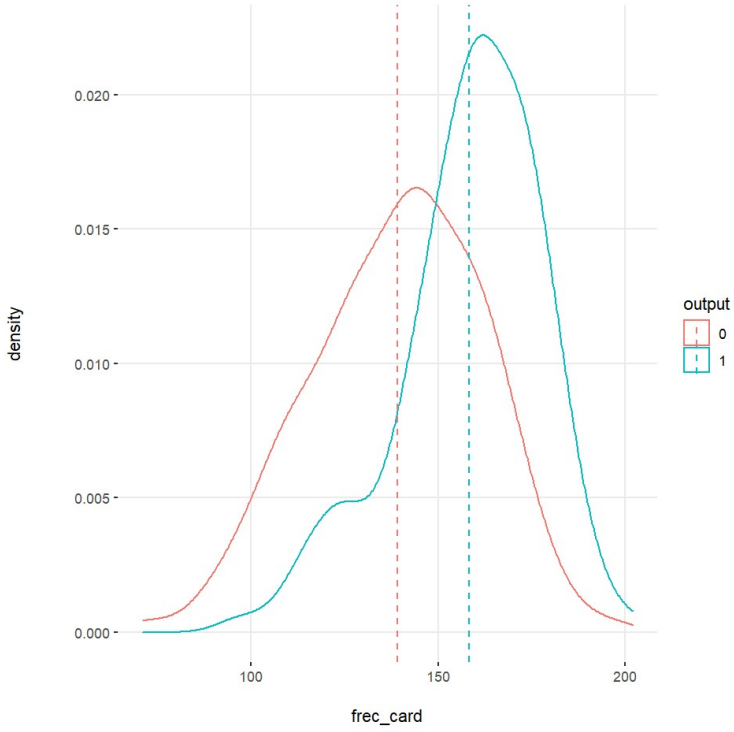
En contra de lo esperado, los grupos de edades más jóvenes presentan mayor probabilidad de infarto.

- Entre pico previo y probabilidad de infarto



Para menores valores de pico previo, la probabilidad de infarto es mayor.

- Entre frecuencia cardíaca y probabilidad de infarto.



Las personas con frecuencias cardíacas máximas más altas tienen más probabilidad de infarto.

4.2 Comprobación de la normalidad y homogeneidad de la varianza.

Para comprobar si en las variables numéricas el mecanismo que genera de los datos proviene de una distribución normal, voy a usar el test de Shapiro-Wilk.

H_0: la variable se distribuye normalmente

Para $\alpha = 0{,}05$, se rechazará H_0 si p-valor < α

Al aplicar el test en R obtengo los siguientes resultados:

variable	p.value	rechazo
edad	6,74E-03	1
presion	1,42E-06	1
colest	5,20E-09	1
frec_card	8,27E-05	1
pico_previo	9,99E-17	1

Para todas las variables se rechaza la hipótesis nula de que provengan de una distribución normal.

Para comprobar la homocedasticidad de las varianzas, es decir, si la varianza es constante (no varía) en los diferentes niveles de un factor, o lo que es lo mismo, entre diferentes grupos, utilizaré el test de Fligner-Killeen. Se trata de la alternativa no paramétrica al test de Levene cuando los datos no siguen una distribución normal, como comprobamos en el apartado anterior.

H_0: homocedasticidad

Para $\alpha = 0{,}05$, se rechazará H0 si p-valor < α

Comprobaré la homogeneidad de las varianzas de cada una de las variables numéricas en los diferentes niveles de output (probabilidad de infarto)

Al aplicar el test en R obtengo los siguientes resultados:

variable	p.value	rechazo
edad	8,10E-03	1
presion	2,54E-01	0
colest	4,01E-01	0
frec_card	2,13E-02	1
pico_previo	2,25E-08	1

Para las variables presión y colesterol se acepta la homocedasticidad, presentan igualdad de varianzas en los diferentes niveles de output.

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

Correlación entre variables

Comprobaré como están relacionadas entre sí las variables numéricas. Para ello utilizo la correlación de Spearman, que es la alternativa no paramétrica cuando las variables no siguen una distribución normal, a la correlación de Pearson.

H0: ρ = 0 (no hay correlación entre las variables)

Para α =0,05, se rechazará H0 si p-valor < α

Obtengo la siguiente matriz de correlaciones en R:

	edad	presion	colest	frec_card	pico_previo
edad	1	0.28970501	0.18890292	-0.39345342	0.26362540
presion	0.2897050	1	0.13021023	-0.04269948	0.15680732
colest	0.1889029	0.13021023	1	-0.04036747	0.03956479
frec_card	-0.3934534	-0.04269948	-0.04036747	1	-0.43049461
pico_previo	0.2636254	0.15680732	0.03956479	-0.43049461	1

Las varibles que presentan mayor correlación entre sí son:

- frecuencia cardíaca y pico previo (correlación negativa)
- frecuencia cardíaca y edad (correlación negativa)
- presión y edad (correlación positiva)

Chi-cuadrado para comparar factores

El test Chi-cuadrado de Pearson se utiliza para comparar si existen diferencias significativas en una variable categórica entre los diferentes grupos de otra variable categórica. Voy a comparar si existen diferencias significativas entre las distintas categorías de cada uno de los factores en cuanto a la probabilidad de infarto.

H0: no existen diferencias

Para α =0,05, se rechazará H0 si p-valor < α

Obtengo los siguientes resultados en R:

variable	p.value	rechazo
sexo	1,55E-06	1
tipo_dolor	1,89E-17	1
azucar	7,61E-01	0
result_electro	7,71E-03	1
angina_ejer	9,56E-14	1
pendiente_ST	6,58E-11	1
talasemia	3,15E-18	1

Solamente para la variable azúcar se acepta la hipótesis nula, es la única que no presenta diferencias significativas entre sus distintos niveles para la probabilidad de sufrir infarto.

Wilcoxon o Mann-Whitney para comparar un factor y una variable numérica

La prueba de Wilcoxon se utiliza para comparar si existen diferencias significativas en una variable numérica entre los diferentes grupos de una variable categórica, en el caso de que no se cumplan las hipótesis de normalidad y homocedasticidad. Voy a comparar si existen diferencias significativas entre las distintas variables numéricas en cuanto a la probabilidad de infarto.

H0: no existen diferencias
Para α =0,05, se rechazará H0 si p-valor < α
Obtengo los siguientes resultados en R:

variable	p.value	rechazo
edad	4,63E-05	1
presion	3,22E-02	1
colest	4,24E-02	1
frec_card	1,40E-13	1
pico_previo	3,35E-13	1

Existen diferencias estadísticamente significativas en todas las variables en términos de probabilidad de infarto.

Modelo de regresión logística

Mediante un modelo de regresión logística voy a predecir la probabilidad de tener un infarto en función de las distintas variables del dataset, al ser la variable respuesta una variable dicotómica.
Pruebo primero con una sola variable explicativa (presión):

```
Call:  glm(formula = output ~ presion, family = "binomial", data = datos_test)
Coefficients:
(Intercept)     presion
   2.42101     -0.01707

Degrees of Freedom: 301 Total (i.e. Null);  300 Residual
Null Deviance:      416.4
Residual Deviance: 409.9        AIC: 413.9
```

Añado otra variable explicativa al modelo, la frecuencia cardíaca.

```
Call:  glm(formula = output ~ frec_card + presion, family = "binomial",
    data = datos)
Coefficients:
(Intercept)    frec_card      presion
  -4.07408      0.04379     -0.01744

Degrees of Freedom: 301 Total (i.e. Null);  299 Residual
Null Deviance:      416.4
Residual Deviance: 353          AIC: 359
```

El criterio de información de Akaike es menor en el segundo modelo, por lo que este es preferible.

Realizo una predicción para frec_card = 150, presion = 140, obteniendo una probabilidad de infarto del 5,33%.