Mario Bru Barrero
Marcos Gómez Vázquez

Tipología y ciclo de vida de los datos

Práctica 2

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

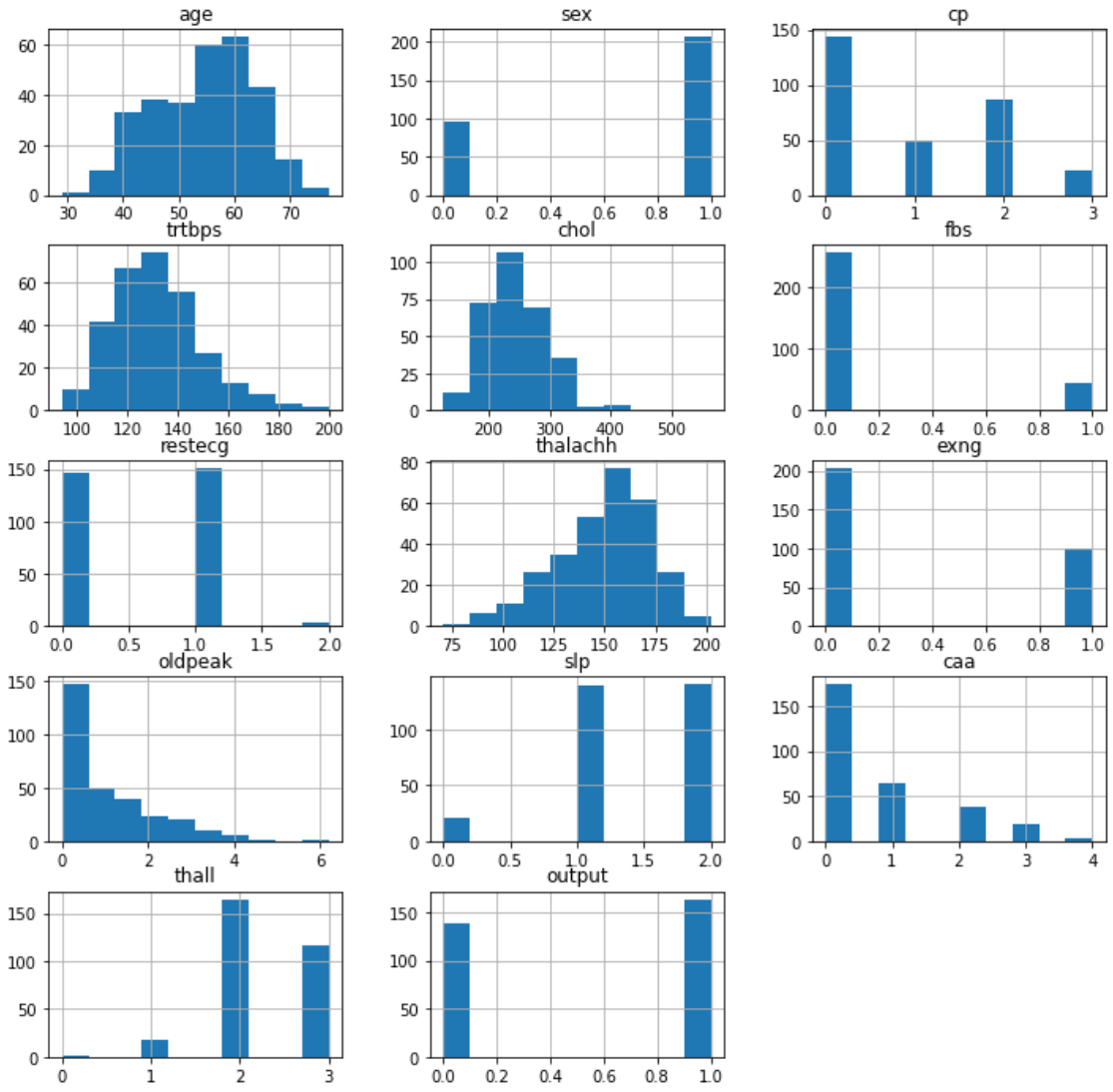
El dataset **Heart Attack Analysis & Prediction Dataset** es un conjunto de datos que almacena diferentes características de un grupo de personas con el objetivo de ayudar a analizar las causas de las enfermedades cardíacas. Por lo tanto, el principal objetivo de este dataset es descubrir nuevos factores o patrones que incrementen las posibilidades de ataques de corazón y ser capaces de predecirlos en base a las características que presenten futuros individuos.

Este conjunto de datos cuenta con **14 atributos y 303 muestras**, pesando un total de **11KB**. La siguiente tabla muestra una descripción de cada atributo del dataset.

Campo	Descripción	Tipo	Ejemplo
age	edad de la persona	int	29
sex	género de la persona • 0: mujer • 1: hombre atributo categórico binario	int	0
cp	tipo de dolor en el pecho • 0: angina típica • 1: angina atípica • 2: dolor no-anginal • 3: asintomático atributo categórico	int	1
trtbps	presión arterial en reposo (en mmHg)	int	120
chol	nivel de colesterol, obtenido a través de sensor BMI (en mg/dl)	int	233
fbs	indica si el nivel de azúcar en sangre en ayunas es mayor/menor que 120 mg/dl • 0: menor • 1: mayor atributo categórico binario	int	0
restecg	resultados electrocardiográficos en reposo • 0: normal • 1: presenta anomalías en la onda ST-T (inversiones de la onda T y/o elevación o depresión del ST > 0.05 mV) • 2: muestra hipertrofia ventricular izquierda probable o definitiva según los criterios de Estes atributo categórico	int	2

thalachh	frecuencia cardíaca máxima alcanzada	int	150
exng	angina inducida por el ejercicio ● 0: no ● 1: sí atributo categórico binario	int	1
oldpeak	Depresión del ST inducida por el ejercicio en relación con el reposo	float	2.3
slp	la pendiente del segmento ST de ejercicio máximo ● 0: pendiente hacia arriba ● 1: sin pendiente (plano) ● 2: pendiente hacia abajo atributo categórico	int	2
caa	número de vasos principales (0, 1, 2, o 3) atributo categórico	int	3
thall	talasemia ● 0: nula ● 1: defecto fijo ● 2: normal ● 3: defecto reversible atributo categórico	int	2
output	diagnóstico de enfermedad cardíaca (estado de enfermedad angiográfico) ● 0: menos del 50% de estrechamiento del diámetro, menos posibilidades de enfermedades del corazón ● 1: más del 50% de estrechamiento del diámetro, más posibilidades de enfermedades del corazón atributo categórico binario esta es la variable objetivo del dataset	int	1

Los siguientes histogramas muestran cómo se distribuyen los valores de cada variable:



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

En este apartado, decidimos crear una variable nueva a partir de 'age'. El objetivo es discretizarla para poder tratarla como una variable categórica. En este caso, la decisión es dividir las edades en rangos.

```
df['age_range'] = pd.cut(x=df['age'], bins=[20, 29, 39, 49, 59, 69,
79], labels=[2, 3, 4, 5, 6, 7])
```

Para nuestro análisis, no queremos escoger ningún subconjunto de los datos, ya que todos son de interés para resolver nuestro problema.

3. Limpieza de los datos.

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

El dataset no contiene elementos vacíos. Esto lo podemos ver inspeccionando la información del DataFrame que almacena los datos:

```
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   age       303 non-null    int64  
 1   sex       303 non-null    int64  
 2   cp        303 non-null    int64  
 3   trtbps    303 non-null    int64  
 4   chol      303 non-null    int64  
 5   fbs       303 non-null    int64  
 6   restecg   303 non-null    int64  
 7   thalachh  303 non-null    int64  
 8   exng      303 non-null    int64  
 9   oldpeak   303 non-null    float64
 10  slp       303 non-null    int64  
 11  caa       303 non-null    int64  
 12  thall     303 non-null    int64  
 13  output    303 non-null    int64  
```

Como vemos, todas las columnas tienen 303 valores no nulos, que se corresponde con el número de filas del dataset.

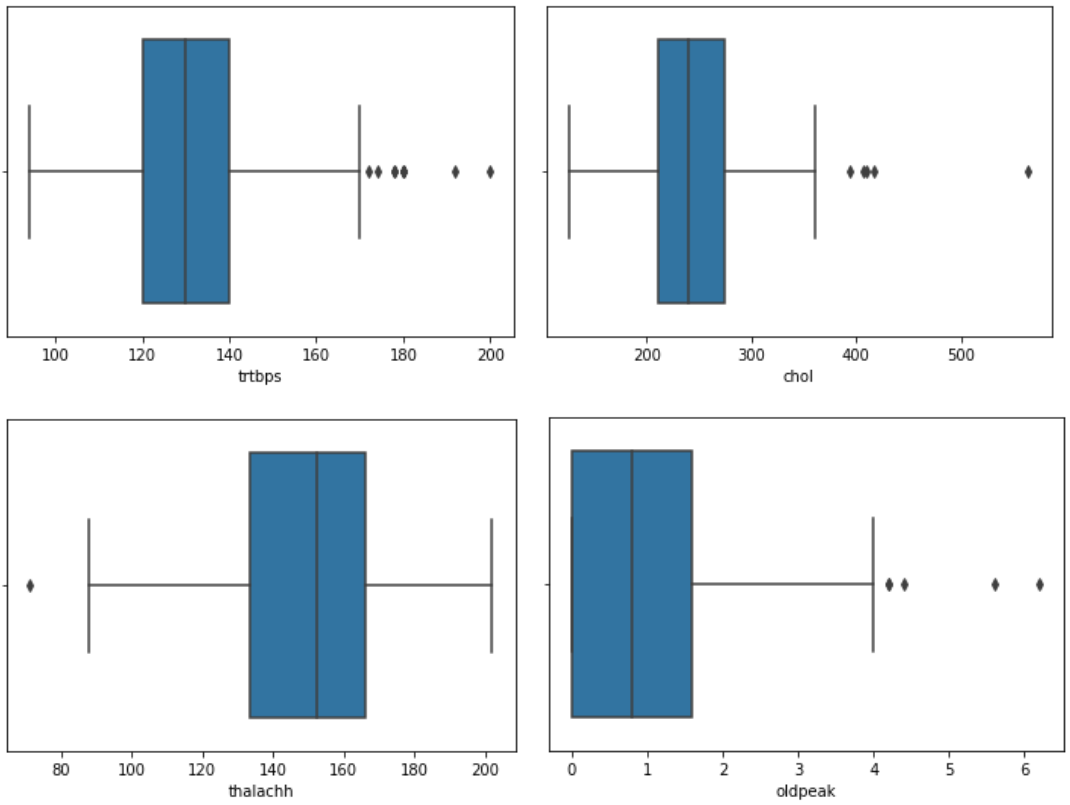
Lo que sí se ha detectado es la presencia de registros duplicados:

age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
38	1	2	138	175	0	1	173	0	0.0	2	4	2	1

Por lo tanto, se ha procedido a eliminar el duplicado, pasando a tener 302 muestras.

3.2. Identifica y gestiona los valores extremos.

Visualizando las variables individualmente, se ha detectado la presencia de outliers, como muestran estos gráficos boxplot:



Para eliminarlos, se ha realizado una normalización *Z-score*. *Z-Score* mide esencialmente a cuántas desviaciones estándar está el valor real del valor medio, en este caso hemos definido un umbral de 3 para definir un valor como "atípico". El código Python correspondiente es el siguiente:

```
z = np.abs(stats.zscore(df))
df_normalized = df[(z<3).all(axis=1)]
```

Una vez realizado este proceso, el número de muestras total ha descendido de 302 a 287.

4. Análisis de los datos.

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

Vamos a comparar los grupos de hombres y mujeres. Consideramos que es interesante analizar los datos en base al género, ya que puede haber diferencias en las causas de los ataques al corazón para cada género. También vamos a comparar la edad con el tipo de enfermedad cardiovascular y con el nivel de colesterol. Por último, analizamos cómo influyen el nivel de azúcar en sangre en ayunas y la frecuencia máxima alcanzada en cada tipo de enfermedad cardiovascular.

Una vez definidos los diferentes grupos, vamos a ver la normalidad y la homocedasticidad de las variables, vamos a observar las correlaciones existentes entre

variables y vamos a realizar regresión logística para ver si los ataques al corazón son predecibles en base a los datos.

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

Comprobación de la normalidad

Para realizar la comprobación de la normalidad, aplicaremos 2 tests: **Shapiro-Wilk** y **D'Agostino's K-squared**. En ambos, se considera como hipótesis nula que los datos proceden de una distribución normal.

El p-value de estos test indica la probabilidad de obtener unos datos como los observados si realmente procediesen de una población con una distribución normal con la misma media y desviación que estos. Por lo tanto, si el p-value es menor que un determinado valor (típicamente 0.05), entonces se considera que hay evidencias suficientes para rechazar la normalidad.

Test de Shapiro-Wilk:

Estadístico = 0.98, p-value = 0.01 (age)
Estadístico = 0.58, p-value = 9.55e-26 (sex)
Estadístico = 0.79, p-value = 7.49e-19 (cp)
Estadístico = 0.97, p-value = 3.22e-05 (trtbps)
Estadístico = 0.99, p-value = 0.25 (chol)
Estadístico = 0.41, p-value = 1.85e-29 (fbs)
Estadístico = 0.68, p-value = 6.19e-23 (restecg)
Estadístico = 0.97, p-value = 9.15e-05 (thalachh)
Estadístico = 0.59, p-value = 1.72e-25 (exng)
Estadístico = 0.85, p-value = 1.52e-15 (oldpeak)
Estadístico = 0.73, p-value = 6.05e-21 (slp)
Estadístico = 0.71, p-value = 8.91e-22 (caa)
Estadístico = 0.74, p-value = 7.63e-21 (thall)
Estadístico = 0.63, p-value = 2.24e-24 (output)

Test de D'Agostino's K-squared:

Estadístico = 7.81, p-value = 0.02 (age)
Estadístico = 430.34, p-value = 3.56e-94 (sex)
Estadístico = 147.73, p-value = 8.32e-33 (cp)
Estadístico = 13.32, p-value = 0.001 (trtbps)
Estadístico = 3.04, p-value = 0.21 (chol)
Estadístico = 112.15, p-value = 4.41e-25 (fbs)
Estadístico = 456.95, p-value = 5.93e-100 (restecg)
Estadístico = 11.75, p-value = 0.002 (thalachh)
Estadístico = 6885.94, p-value = 0.0 (exng)
Estadístico = 35.62, p-value = 1.83e-08 (oldpeak)
Estadístico = 21.13, p-value = 2.56e-05 (slp)
Estadístico = 49.16, p-value = 2.10e-11 (caa)

Estadístico = 10.8, p-value = 0.004 (thall)
Estadístico = 1401.66, p-value = 4.29e-305 (output)

En ambos tests, los resultados muestran un **p-value superior a 0.05** únicamente en la variable **chol**. Esto significa que hay evidencias para descartar la normalidad de todas las variables a excepción de 'chol'.

Homocedasticidad

Vamos a comprobar la homocedasticidad (homogeneidad de las varianzas) de cada variable respecto 2 grupos: hombres y mujeres. Es decir, mediante el test de **Levene** y el de **Fligner-Killeen** vamos a comprobar si la varianza de cada una de estas 3 variables entre estos 2 grupos es homogénea o no.

Al igual que en los tests de comprobación de normalidad, en estos se considera como hipótesis nula que los datos proceden de distribuciones con la misma varianza (homocedasticidad). Por lo tanto, si el p-value es menor que un determinado valor (típicamente 0.05), entonces se considera que hay evidencias suficientes para rechazar la homocedasticidad en favor de la heterocedasticidad.

Existe otro test (Barteltt), pero la decisión de utilizar los 2 tests mencionados viene dada por la falta de normalidad de las variables, que como vimos anteriormente no ha sido demostrada.

[Levene] Estadístico = 0.79, p-value = 0.37 (age)
[Fligner-Killeen] Estadístico = 1.09, p-value = 0.29 (age)

[Levene] Estadístico = 0.60, p-value = 0.43 (cp)
[Fligner-Killeen] Estadístico = 1.36, p-value = 0.24 (cp)

[Levene] Estadístico = 0.61, p-value = 0.43 (trtbps)
[Fligner-Killeen] Estadístico = 0.51, p-value = 0.47 (trtbps)

[Levene] Estadístico = 3.54, p-value = 0.06 (chol)
[Fligner-Killeen] Estadístico = 3.45, p-value = 0.06 (chol)

[Levene] Estadístico = 0.97, p-value = 0.32 (fbs)
[Fligner-Killeen] Estadístico = 0.97, p-value = 0.32 (fbs)

[Levene] Estadístico = 0.49, p-value = 0.48 (restecg)
[Fligner-Killeen] Estadístico = 0.58, p-value = 0.44 (restecg)

[Levene] Estadístico = 3.79, p-value = 0.05 (thalachh)
[Fligner-Killeen] Estadístico = 3.92, p-value = 0.04 (thalachh)

[Levene] Estadístico = 6.28, p-value = 0.01 (exng))
[Fligner-Killeen] Estadístico = 6.16, p-value = 0.01 (exng)

[Levene] Estadístico = 8.36, p-value = 0.004 (oldpeak)
[Fligner-Killeen] Estadístico = 11.06, p-value = 0.0008 (oldpeak)

[Levene] Estadístico = 0.03, p-value = 0.84 (slp)
[Fligner-Killeen] Estadístico = 0.03, p-value = 0.84 (slp)

[Levene] Estadístico = 5.33, p-value = 0.02 (caa)
[Fligner-Killeen] Estadístico = 5.64, p-value = 0.01 (caa)

[Levene] Estadístico = 63.2, p-value = 4.32e-14 (thall)
[Fligner-Killeen] Estadístico = 51.9, p-value = 5.73e-13 (thall)

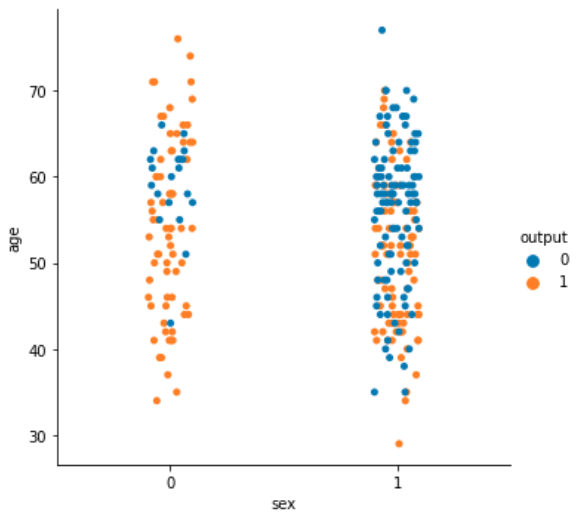
Resaltados en negrita, vemos las variables que no muestran evidencias para rechazar la hipótesis de que los grupos de hombres y mujeres tienen la misma varianza

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

Primero procedemos con un análisis visual de los grupos definidos en el punto 4.1 para descubrir si existe alguna relación entre dichas variables, así como sacar conclusiones si las hubiere.

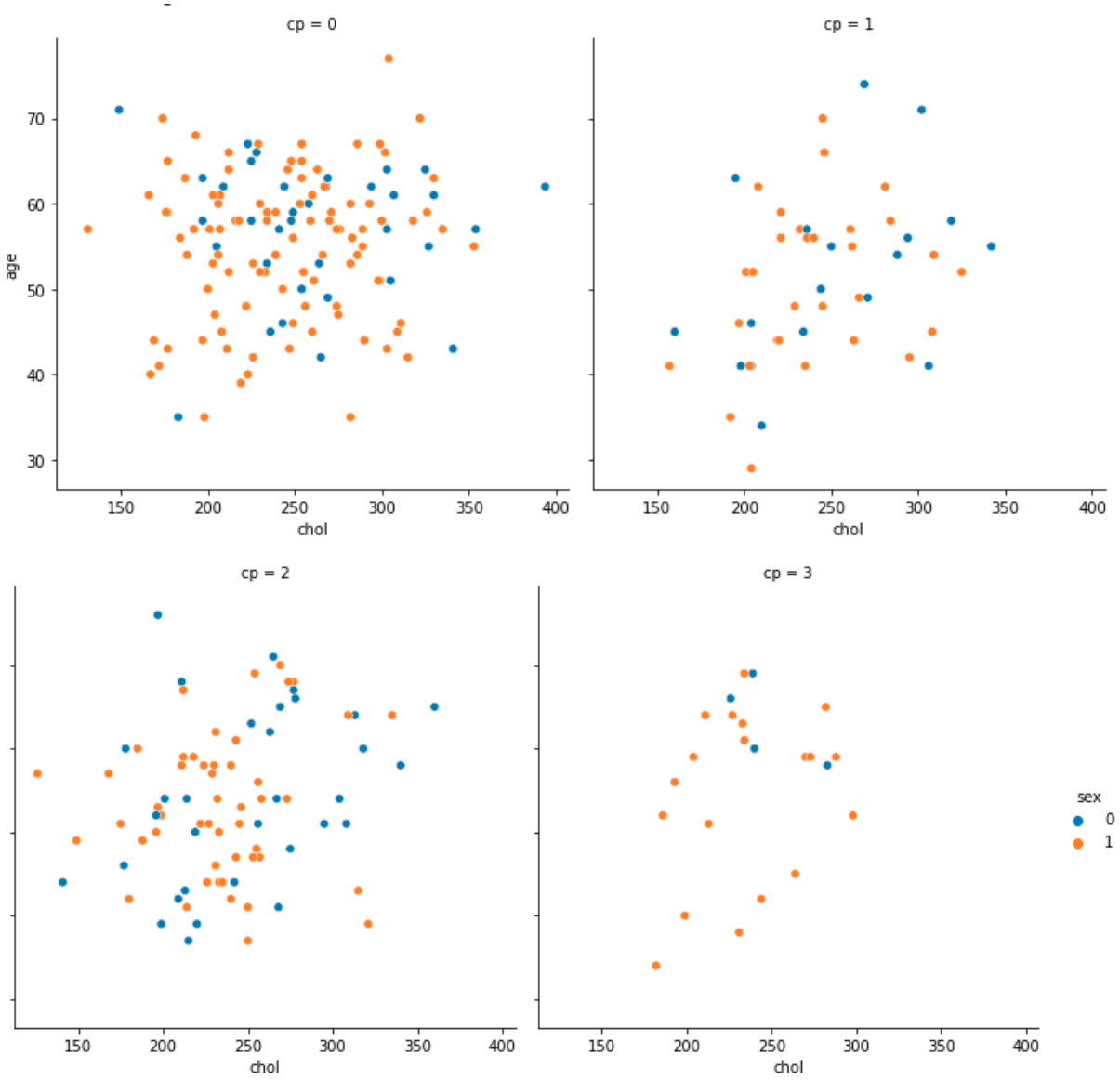
```
[ ] sns.catplot(data=df, x='sex', y='age', hue='output')
```

```
<seaborn.axisgrid.FacetGrid at 0x7fb8f1a5c430>
```



Analizando la gráfica superior descubrimos que, en general, los hombres son menos propensos a tener enfermedades cardiovasculares que las mujeres, quizás relacionado por una una mayor actividad física en dicho sexo.

Otro análisis de grupos interesante podría ser también por edades, por tipo de enfermedad cardiovascular y por índice de colesterol:

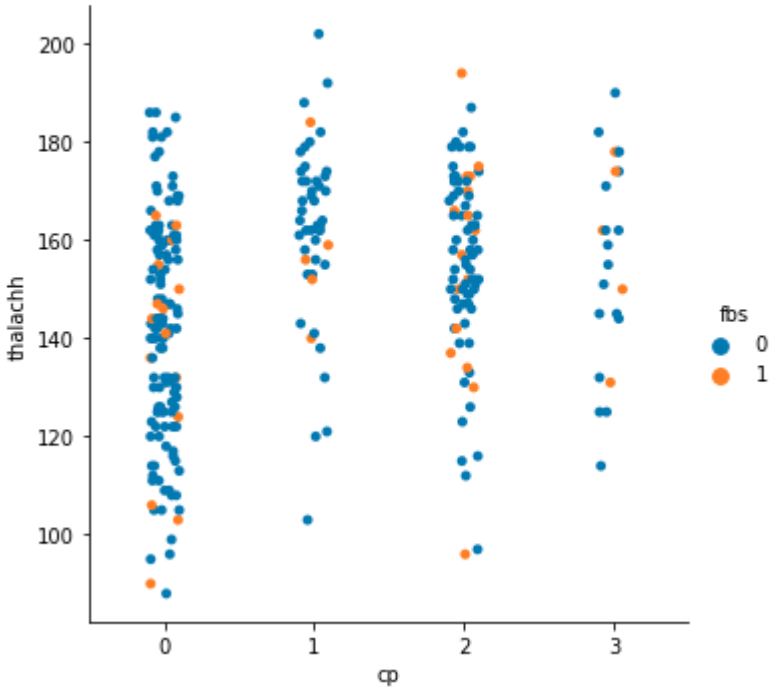


Del resultado anterior podemos deducir que las personas entre 50 y 65 años experimentan cp = 0 (angina típica), que es el dolor más común entre la lista.

Por último, analizamos cómo influyen el nivel de azúcar en sangre en ayunas y la frecuencia máxima alcanzada en cada tipo de enfermedad cardiovascular:

```
sns.catplot(data=df, x='cp', y='thalachh', hue='fbs')
```

```
<seaborn.axisgrid.FacetGrid at 0x7f745232f6d0>
```



Del gráfico anterior podemos deducir que cp[0] es el dolor en el pecho más habitual que experimentan los pacientes, seguido de cp[2], cp[1] y, por último, cp[3]. También podemos ver que el nivel de azúcar en sangre en ayunas (fbs) está por debajo de 120, lo que significa que la mayoría de los fbs del paciente no están por encima de 120.

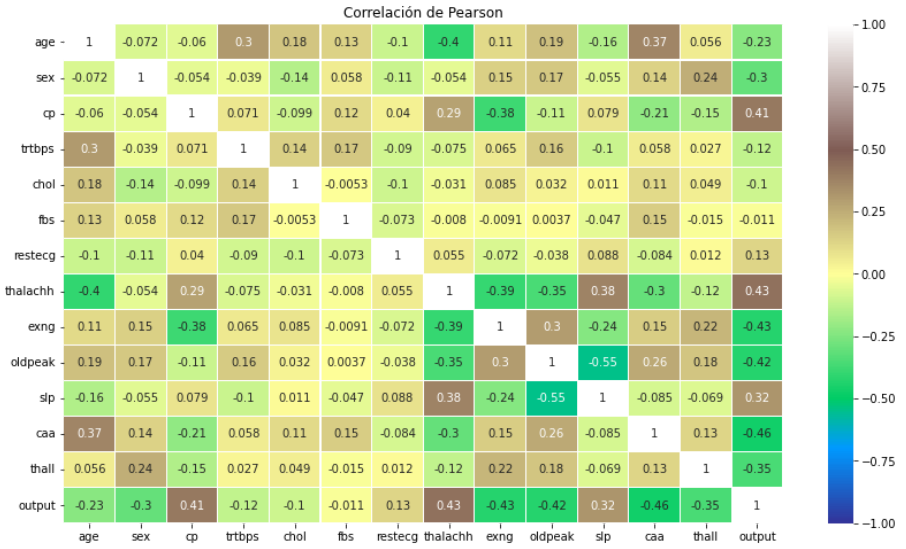
Análisis de correlación

Utilizaremos los dos tipos de correlación explicados en la asignatura:

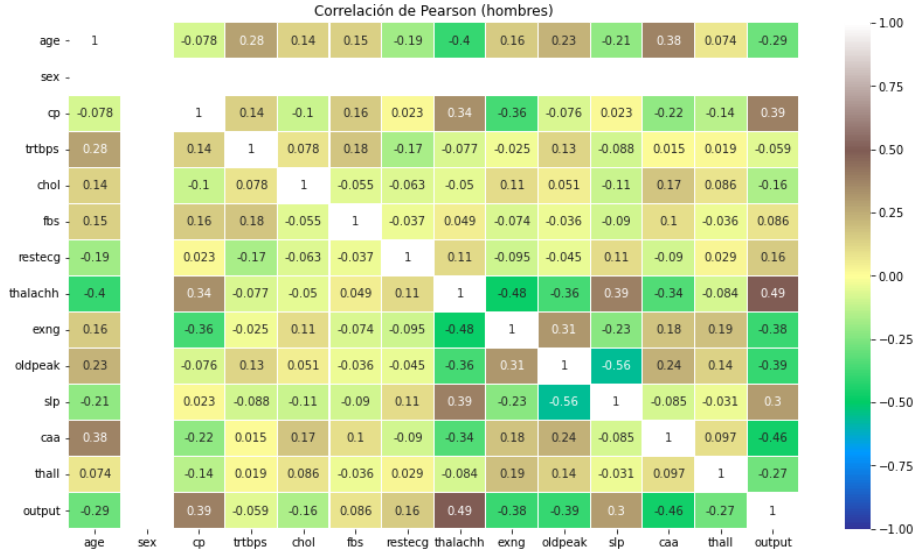
- Correlación de Pearson
- Correlación de Spearman

Pearson

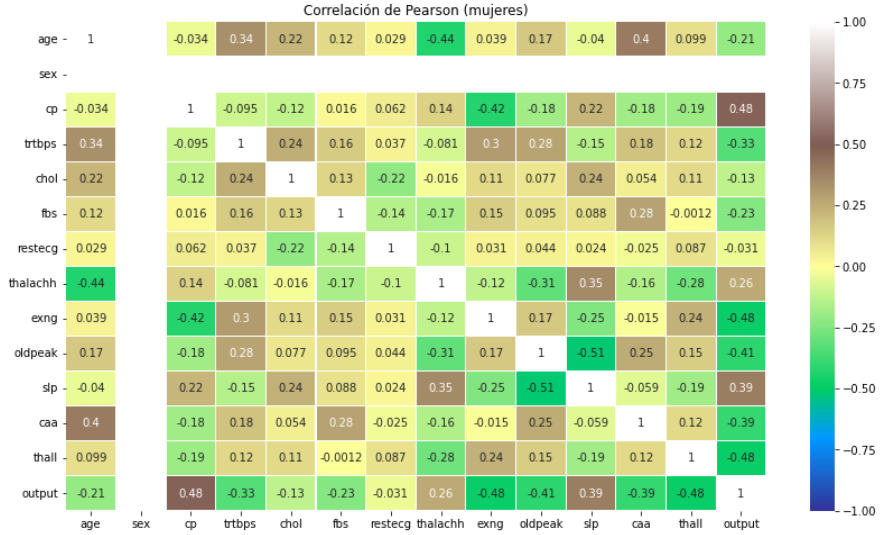
A continuación se muestran los resultados de la correlación de Pearson. La primera matriz muestra la correlación entre cada par de variables, teniendo en cuenta todos los datos (hombres y mujeres).



La siguiente muestra la misma matriz pero solamente habiendo utilizado los hombres para el análisis.

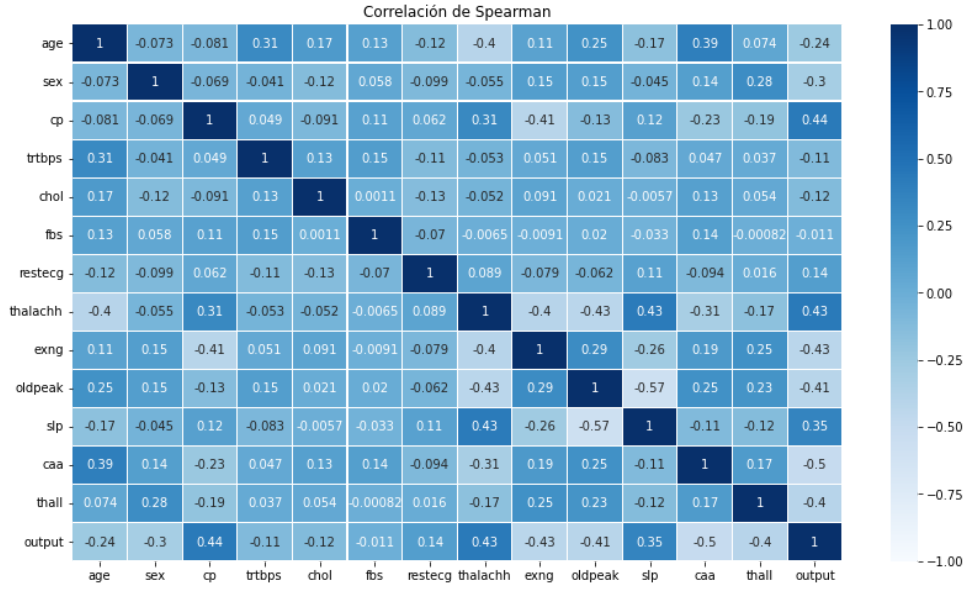


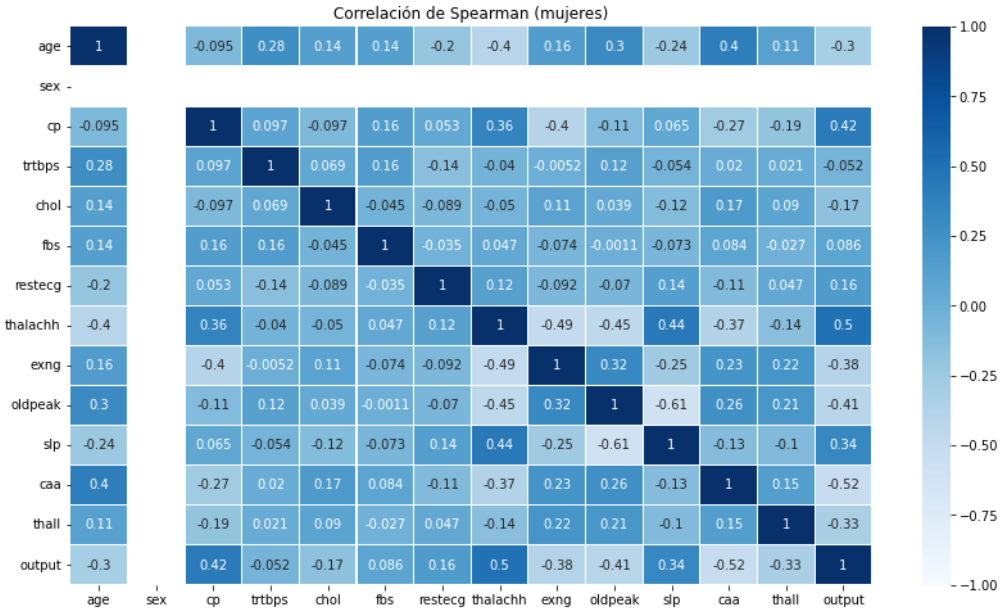
Y finalmente, la misma matriz para las mujeres.

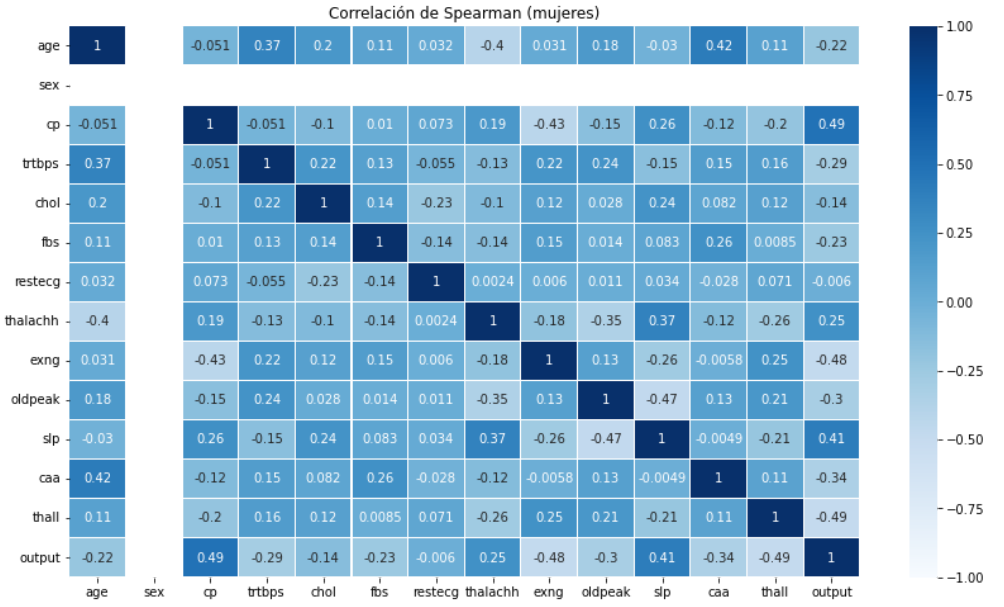


Spearman

Ahora visualizaremos la matriz de correlación de Spearman, al igual que anteriormente, primero para todos los datos, y luego para hombres y mujeres de forma separada.







Se han obtenido unos resultados muy similares tanto con Pearson como con Spearman y probando con todos los datos y con hombres y mujeres por separado. Las gráficas de este apartado muestran que las variables más correlacionadas con 'output' son 'cp', 'thalachh' y 'slp'. Lo que se ha comprobado con este análisis es que los 2 grupos que hemos escogido presentan correlaciones muy similares. Así que en este análisis hemos aprendido (o más bien corroborado) que las causas de ataques al corazón son casi iguales en ambos géneros.

Regresión logística

En este análisis, el objetivo es crear un modelo de regresión lineal (modelo de clasificación) y entrenarlo con una parte de los datos (datos de entrenamiento). Después, con la parte sobrante de los datos (datos de testeo) evaluaremos qué tan bien es capaz de clasificar este modelo nuevas entradas. La pregunta que pretende responder este análisis es: ¿Es posible predecir un ataque al corazón con los datos de los que disponemos? Hay que experimentar con ellos para saber si las variables son representativas de lo que puede producir un ataque al corazón y si contamos con las muestras necesarias para llevar a cabo esta tarea.

Para los datos de entrenamiento, seleccionamos un 80% de las muestras y para testeo el 20% restante. Tenemos que separar también la variable objetivo a predecir (Y) del resto de variables (X)

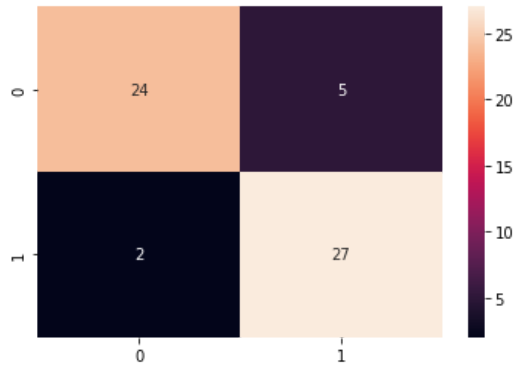
```
y = df['output']
x = df.loc[:, ['age', 'sex', 'cp', 'trtbps', 'chol', 'fbs',
'restecg', 'thalachh', 'exng', 'oldpeak', 'slp', 'caa', 'thall']]
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=
0.2, random_state= 0)
```

Después, solamente tenemos que inicializar el modelo y empezar el entrenamiento.

```
scaler = StandardScaler()
x_train = scaler.fit_transform(x_train)
x_test = scaler.transform(x_test)
model = LogisticRegression()
model.fit(x_train, y_train)
```

Finalmente, procedemos a evaluar el modelo con los datos que hemos reservado para esta fase de testeo. Para visualizar los resultados utilizamos la métrica de precisión (accuracy) y una matriz de confusión que muestra los true positives / true negatives / false positives / false negatives obtenidos con los datos de testing. El eje vertical indica los valores reales de la variable output y el eje horizontal el valor que el modelo ha predicho.

Accuracy: **87.93%**



La conclusión de este análisis es que probablemente en un 87% de los casos, podríamos ser capaces de diagnosticar correctamente a una persona con una enfermedad cardíaca. Aun así, es importante recalcar que la tasa de errores, pese a ser baja, en un caso real debe ser muy controlada, ya que no es lo mismo equivocarse al decir que una persona tiene una enfermedad (y por tanto no la tenga) que lo contrario, ya que sería muy peligroso para su vida.

5. Representación de los resultados a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

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

Tras analizar este dataset en profundidad hemos observado que:

- La mayoría de los pacientes tienen edad (50-60). En la que el número máximo de pacientes tiene 56 años.
- Alrededor del 68 % son pacientes masculinos y el 32 % son pacientes femeninas.
- La mayoría de los pacientes tienen dolor de pecho tipo 1 que es un dolor de angina típico.
- La presión arterial de la mayoría de los pacientes se encuentra entre (130-140).
- El nivel de colesterol de la mayoría de los pacientes se encuentra entre (200-250).
- La frecuencia cardíaca de la mayoría de los pacientes se encuentra entre (155-165).

En base a los resultados obtenidos a lo largo de este análisis, estas son nuestras conclusiones:

- No existe una relación sólida entre la edad y el ataque cardíaco. Por lo tanto, no podemos decir que con el aumento de la edad existe una alta o baja probabilidad de ataque cardíaco.
- Parece que el sexo sí es un factor diferencial en el resultado, ya que hemos visto que los hombres en general son menos propensos a tener enfermedades cardiacas que las mujeres. Aunque hemos comprobado que las causas en ambos géneros son las mismas.
- Observamos que el aumento de la frecuencia cardíaca tiene relación con un alto riesgo de enfermedad cardiaca.
- El aumento de la presión arterial también está directamente relacionado con una alto riesgo de enfermedad cardiaca.
- El aumento del nivel de colesterol también aumenta el riesgo de ataques al corazón.

Los resultados obtenidos nos responden muy bien al problema. El objetivo era analizar las causas de enfermedades cardíacas y lo hemos conseguido. Respecto a ser capaces de prever una enfermedad cardíaca en base a nuestro historial de datos, creemos que lo hemos logrado de forma aceptable con una solución muy simple. Creemos que la naturaleza de los datos puede permitir realizar análisis y predicciones aún más elaboradas si se desea.

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

8. Vídeo. Realizar un breve vídeo explicativo de la práctica (máximo 10 minutos), donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC (https://drive.google.com/…), junto con enlace al repositorio Git entregado.

CONTRIBUCIONES

Contribuciones	Firma
Investigación previa	MGV, MBB
Redacción de las respuestas	MGV, MBB
Desarrollo del código	MGV, MBB
Participación en el vídeo	MGV, MBB

BIBLIOGRAFÍA

- Calvo M, Subirats L, Pérez D (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.