1. Descripción del dataset

El dataset a tratar en la práctica es https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset.

Se trata de un dataset con 14 variables, las cuales estan asociadas a pancientes con síntomas coronarios y cuyo objetivo es la predicción de la probabilidad (mayor o menor) de tener un ataque al corazón bajo esos síntomas. El dataset es de mucha relevancia, dado que poder entrerar un modelo que anticipe esta circunstancia podría ayudar a salvar muchas vidas.

El dataset consta de las siguientes variables:

- **age** : Edad del paciente
- **sex**: Sexo del paciente codificado como 0 o 1. Se desconoce su traducción a Hombre o Mujer.
- **cp** : Chest Pain type chest pain type. Tipo de dolor en el pecho. Puede tomar los valores:
  - Valor 0: typical angina
  - Valor 1: atypical angina
  - Valor 2: non-anginal pain
  - Valor 3: asymptomatic

    > Aunque en el dataset de referenca de Kaggle de indica que el rango de valores de la variable es [1..4], analizando el fichero se observa que es [0..3]
- **trtbps** : resting blood pressure (in mm Hg). Presion sanguinea en reposo.
- **chol** : cholestoral in mg/dl fetched via BMI sensor. Colesterol en sangre.
- **fbs** : (fasting blood sugar > 120 mg/dl) (1 = true; 0 = false). Azugar en sangre en ayunas por encima de 129 mg/dl. Codificado como 1 Verdadero, 0 Falso.
- **restecg** : resting electrocardiographic results. Resultados del electrocardiograma en reposo. Puede tomar los siguientes valores:
  - Valor 0: normal
  - Vaslor 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)
  - Valor 2: showing probable or definite left ventricular hypertrophy by Estes' criteria.
- **thalachh** : maximum heart rate achieved. Frecuencia cardiaca máxima.

- **exng**: "exercise induced angina". Angina inducida por el ejercicio. Codificado como 1 "Si" 0 "No".
- **caa**: number of major vessels. Numero de vasos sanguineos mayores. Codificado de 0 a 4.

> En la descripción del dataset de Kaggle se indica que que el rango de la variable es 0..3, sin embargo observando el fichero, se codifica de 0..4. Esto es mas coherente, ya que los vasos mayores del corazón son 5.

- **target** : Variable objetivo. 0 = menor posibilidad de ataque al corazón 1 = mayor posibilidad de ataque al corazon.

El dataset consta de otras tres variables que no están descritas y no conocemos su significado (oldpoeak, slp y thall) que no utilizaremos por pruedencia.

A continuación, vamos a visualizar los primeros datos del dataset

In [190…
```
import pandas as pd
import matplotlib.pyplot as plt
```

In [191…
```
df = pd.read_csv("./datos/heart.csv")
```

In [192…
```
df.head()
```

Out[192]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	oldpeak	slp	caa	thall	output
0	63	1	3	145	233	1	0	150	0	2.3	0	0	1	1
1	37	1	2	130	250	0	1	187	0	3.5	0	0	2	1
2	41	0	1	130	204	0	0	172	0	1.4	2	0	2	1
3	56	1	1	120	236	0	1	178	0	0.8	2	0	2	1
4	57	0	0	120	354	0	1	163	1	0.6	2	0	2	1

Borramos las columnas que no vamos a utlizar

In [193…
```
df = df.drop(["oldpeak", "slp", "thall"], axis=1)
```

2. Integración y selección

Vamos a integrar los ficheros que contienen las descriciones de los campos categoricos del dataset, de manera que sea más fácil su interpretación y trabajo con los datos.

Después de la integración de cada fichero, borraremos la columna de cruce del fichero integrado, para no duplicar la columna.

Exang

In [194…
```
df_exang = pd.read_csv("./datos/exang.csv")
```

In [195…
```
df = df.merge(df_exang, how="left", left_on="exng", right_on="id_exang").drop(["
```

Chest pain

In [196…
```
df_cp = pd.read_csv("./datos/chest_pain.csv")
```

In [197…
```
df = df.merge(df_cp, left_on="cp", right_on="id_cp", how="left").drop(["id_cp"],
```

Visualizamos el dataset final

In [234…
```
df.head()
```

Out[234]:

	age	sex	cp	trtbps	chol	fbs	restecg	thalachh	exng	caa	output	desc_exang	d
0	63	1	3	145	233	1	0	150	0	0	1	no	asympt
1	37	1	2	130	250	0	1	187	0	0	1	no	non-a
2	41	0	1	130	204	0	0	172	0	0	1	no	a
3	56	1	1	120	236	0	1	178	0	0	1	no	a
4	57	0	0	120	354	0	1	163	1	0	1	yes	typical

3. Limpieza de los datos.

NULOS

Obenemos un listado de las columnas que tienen nulos y el porcentaje que representan sobre el total de datos.

In [257…
```
df.apply(lambda x: sum(x.isnull())/len(x))
```

Out[257]:
```
age          0.0
sex          0.0
cp           0.0
trtbps       0.0
chol         0.0
fbs          0.0
restecg      0.0
thalachh     0.0
exng         0.0
caa          0.0
output       0.0
desc_exang   0.0
desc_cp      0.0
dtype: float64
```

No hay datos nulos

PERDIDOS

Vamos a identificar valores extraños que puedan significar perdida de datos:

age

In [203…
```
df.age.describe()
```

Out[203]:
```
count    303.000000
mean      54.366337
std        9.082101
min       29.000000
25%       47.500000
50%       55.000000
75%       61.000000
max       77.000000
Name: age, dtype: float64
```

Minimo y máximo en rangos coherentes. La información parece correcta.

sex

In [202…
```
df.sex.unique()
```

Out[202]:
```
array([1, 0], dtype=int64)
```

Correcta

cp

In [205…
```
df.cp.unique()
```

Out[205]:
```
array([3, 2, 1, 0], dtype=int64)
```

Valores en rango. Parece correcta.

trtbps

In [206…
```
df.trtbps.describe()
```

Out[206]:
```
count    303.000000
mean     131.623762
std       17.538143
min       94.000000
25%      120.000000
50%      130.000000
75%      140.000000
max      200.000000
Name: trtbps, dtype: float64
```

El rango de presión arterial parece normal. No vemos valores raros.

chol

In [208…
```
df.chol.describe()
```

Out[208]:
```
count    303.000000
mean     246.264026
std       51.830751
min      126.000000
25%      211.000000
50%      240.000000
75%      274.500000
max      564.000000
Name: chol, dtype: float64
```

No se observan valores extraños.

fbs

In [210…
```
df.fbs.unique()
```

Out[210]:
```
array([1, 0], dtype=int64)
```

Valores esperados. Todo ok.

restecg

In [212…
```
df.restecg.unique()
```

Out[212]:
```
array([0, 1, 2], dtype=int64)
```

Valores esperados. Todo ok.

thalachh

In [214…
```
df.thalachh.describe()
```

Out[214]:
```
count    303.000000
mean     149.646865
std       22.905161
min       71.000000
25%      133.500000
50%      153.000000
75%      166.000000
max      202.000000
Name: thalachh, dtype: float64
```

El rango [71..202] parece normal.

exng

In [216…
```
df.exng.unique()
```

Out[216]:
```
array([0, 1], dtype=int64)
```

Valores esperados. Todo ok

caa

In [219…
```
df.caa.unique()
```

Out[219]:
```
array([0, 2, 1, 3, 4], dtype=int64)
```

Valores en rango. Todo correcto

output

In [222…
```
df.output.unique()
```

Out[222]:
```
array([1, 0], dtype=int64)
```

Valores en rango. Todo ok.

Conclusión: No hay ni nulos ni valores que representen pérdida de información. No se realizará ningún tratamiento sobre los datos

---

VALORES EXTREMOS

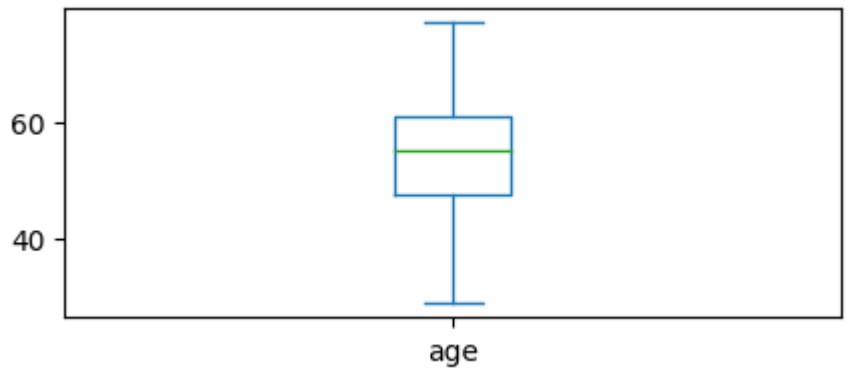
Vamos a revisar las variables numéricas, para identificar posibles valores extremos, que identifiquen un error de dato. Vamos a realizar el análisis a traves de diagramas de caja.

Las variables categóricas las hemos revisado buscando datos perdidos y están correctas.

age

In [233…
```
df.age.plot.box(figsize=(5, 2))
```

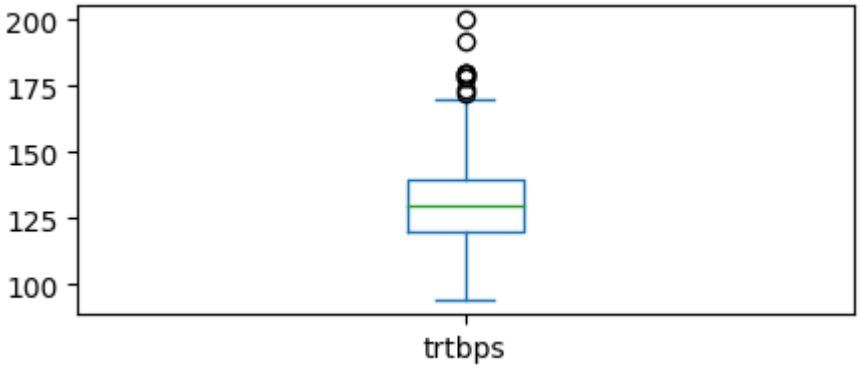
Out[233]:
```
<AxesSubplot: >
```



No se observan valores atípicos.

trtbps (presión sanguinea en reposos)

In [237…
```
bp = df.trtbps.plot.box(figsize=(5, 2))
```



Se observan valores atípicos superiores. Vamos a listarlos:

In [240…
```
from matplotlib.cbook import boxplot_stats
```

In [241…
```
boxplot_stats(df.trtbps)
```

Out[241]:
```
[{'mean': 131.62376237623764,
  'iqr': 20.0,
  'cilo': 128.1961171325887,
  'cihi': 131.8038828674113,
  'whishi': 170,
  'whislo': 94,
  'fliers': array([172, 178, 180, 180, 200, 174, 192, 178, 180], dtype=int64),
  'q1': 120.0,
  'med': 130.0,
  'q3': 140.0}]
```

Los valores de presión aún siendo extremos para el conjunto de datos, son coherentes para la medida. Según Osborne en "*Best Practices in Data Cleaning: A Complete Guide to Everything You Need to Do Before and After Collecting Your Data*", corresponderían a la categoría *6 Extreme Scores as Legitimate Cases Sampled From the Correct Population*.

La recomendación en este caso es eliminarlos, ya que afectan sensiblemente a las estadísticas de la población y los test que se realicen en ella.

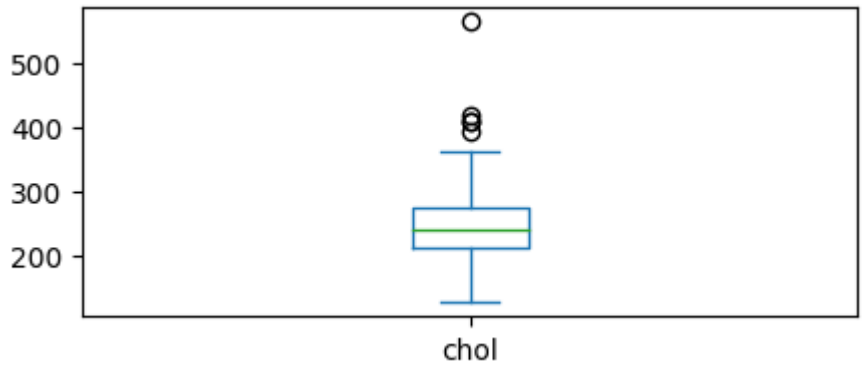
__TODO__:

- Realizar análisis multivariable para determinar si son atípicos realmente.
- Decidir si eliminarlos (son sólo 9/303) o vaciarlos e imputar mediante clustering (mejor esto segundo)

chol (Colesterol en sangre)

In [250…
```
df.chol.plot.box(figsize=(5, 2))
```

Out[250]:
```
<AxesSubplot: >
```



Se observan atípicos. Vamos a listaros:

In [254…
```
boxplot_stats(df.chol)
```

Out[254]:
```
[{'mean': 246.26402640264027,
  'iqr': 63.5,
  'cilo': 234.27267189596918,
  'cihi': 245.72732810403082,
  'whishi': 360,
  'whislo': 126,
  'fliers': array([417, 564, 394, 407, 409], dtype=int64),
  'q1': 211.0,
  'med': 240.0,
  'q3': 274.5}]
```

Nuevamente, los valores son atípicos, pero coherentes para la medida. También los eliminaremos.

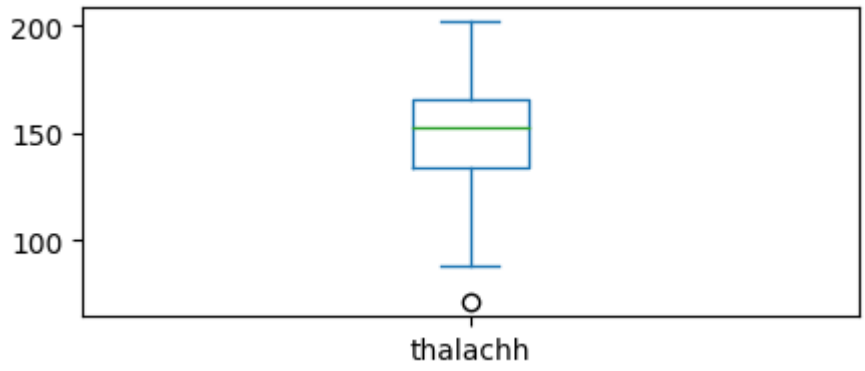
__TODO__:

- Eliminar atípicos e imputar por clustering.
- Verificar la media y varianza antes y despues de eliminar los atípicos

thalachh (pulso máximo)

In [252…
```
df.thalachh.plot.box(figsize=(5,2))
```

Out[252]:
```
<AxesSubplot: >
```



Se observan atípicos. Vamos a listarlos:

In [256…
```
boxplot_stats(df.thalachh)
```

Out[256]:
```
[{'mean': 149.64686468646866,
  'iqr': 32.5,
  'cilo': 150.06869034045667,
  'cihi': 155.93130965954333,
  'whishi': 202,
  'whislo': 88,
  'fliers': array([71], dtype=int64),
  'q1': 133.5,
  'med': 153.0,
  'q3': 166.0}]
```

El valor de 71 para pulso máximo parece un error de medición. Es muy baja para ser un pulso en esfuerzo. Según Osborne, sería un atípico del tipo *1. Extreme Scores From Data Errors*. Como no podemos corregirlo, lo eliminaremos.

__TODO__:

- Eliminar e imputar

4. Análisis de los datos.

Las variables que vamos a analizar son:

- age (edad)
- sex (sexo)
- trtbps (presión sanguinea en reposo)
- chol (colesterol en sangre)

El objetivo es determinar si existe Correlación entre las variables numéricas y si a partir de ellas podemos realizar una modelo de Regresión que prediga la variable objetivo. También verificaremos que si el sexo presenta diferencias estadísticas con respecto a las otras variables y así realizar estudios separados.

Antes de todo, realizaremos pruebas de normalidad y homocedasticidad a las variables numéricas, para determinar qué tipo de análisis estadístico realizaremos.

Test de normalidad

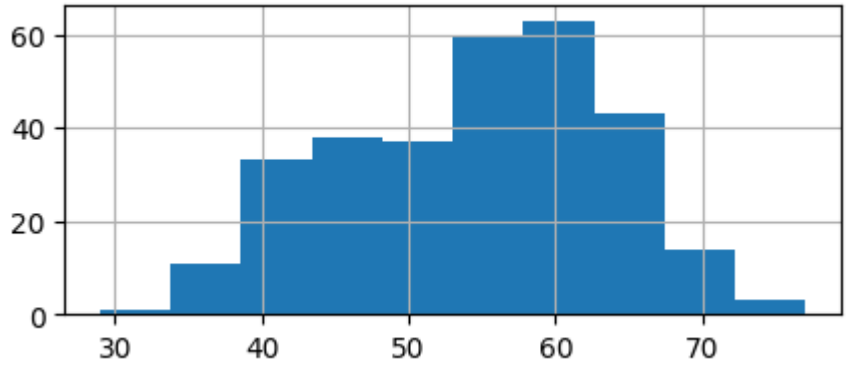
In [261…
```
from scipy.stats import shapiro
```

In [263…
```
shapiro(df.age)
```

Out[263]:
```
ShapiroResult(statistic=0.9863712787628174, pvalue=0.005800595041364431)
```

In [265…
```
df.age.hist(figsize=(5,2))
```

Out[265]:
```
<AxesSubplot: >
```



El valor pvalue del test de Shaphiro es menor a 0.05, lo que implica que no hay normalidad. Deberemos transformar los datos antes de comprobar la correlación (Pearson) con las demás variables.

En el histograma se aprecia una inclinación hacia la derecha.

In [ ]:

5. Representación de los resultados

Se realizará durante toda la práctica

6. Resolución del problema

In [ ]: