

## HEART DATASET

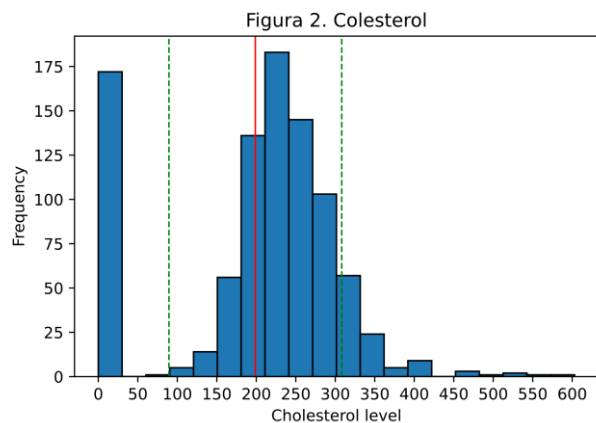
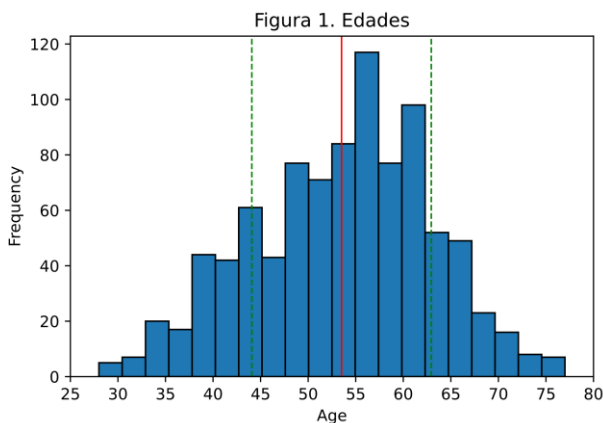
El *dataset* que he empleado para este entregable consiste en un tamaño de muestra de 918 pacientes y contamos con un total de 12 variables diferentes. 7 de las variables son de carácter cuantitativo y se trata de: *Age*, *RestingBP*, *Cholesterol*, *FastingBS*, *MaxHR*, *Oldpeak* y *HeartDiseas*. Y las 5 variables restantes son de carácter nominal y se trata de: *Sex*, *ChestPainType*, *RestingECG*, *ExerciseAngina* y *ST\_Slope*.

Las variables cuantitativas las hemos identificado porque sus valores se reflejan directamente con números. En este caso he decidido escoger como variables cuantitativas: la edad (*Age*) y el colesterol (*Cholesterol*).

Para obtener valores estadísticos de ambos, hemos utilizado un *describe* y hemos obtenido los siguientes valores:

- Edad: media de 53.5 y std 9.43
- Colesterol: media de 198.9 y std 109.38

También ha sido de gran ayuda saber los mínimos y máximos a la hora de elegir en el gráfico el comienzo del valor de la x en su eje y su extensión, así como el valor en *steps*.



En la Figura 1 se muestra el gráfico de la Edades en el que podemos apreciar como la media de edad (línea roja) se sitúa en 53 años. También observamos como la mínima edad es de 28 años y su máxima es de 77. De la misma forma también podemos ver que la edad que mas veces se repite es el grupo perteneciente al rango de entre 55 y 57 años.

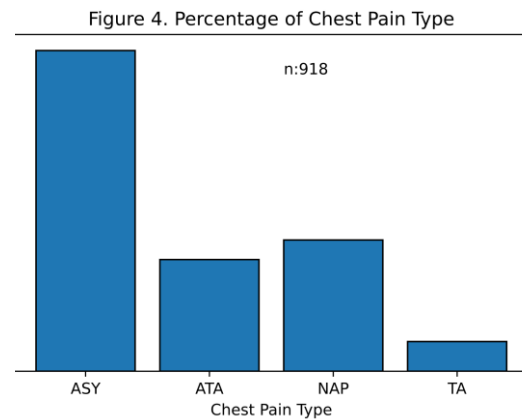
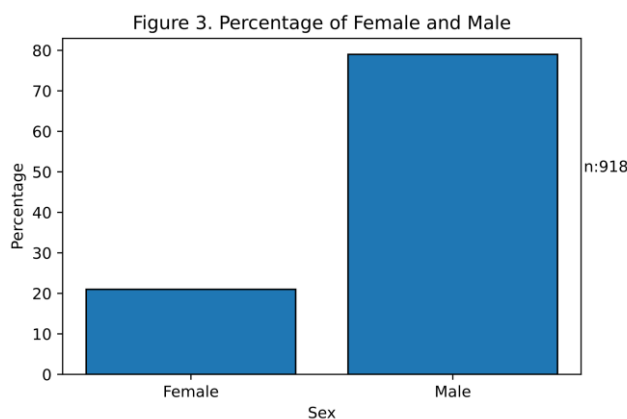
El segundo histograma, el gráfico de la variable cuantitativa colesterol, en la cual observamos a simple vista como hay un gran numero de pacientes con 0 colesterol, el mínimo valor. Con un máximo valor se encuentra un único paciente, con una cifra de 603 de nivel de colesterol. La media, indicada como en el anterior gráfico, con una línea roja, se sitúa sobre los 198.8. Los niveles de colesterol que mas frecuencia tienen se reflejan entre los 200 y 250.

Una vez comentadas numérica y gráficamente las dos variables cuantitativas escogidas, analizamos las variables nominales que hemos decidido estudiar. Las variables nominales que utilizamos en este caso son: el sexo (*Sex*) y el tipo de dolor pectoral (*Chest Pain Type*).

Las variables nominales las hemos identificado porque sus valores se reflejan con características o datos no numéricos. El sexo, cuyos valores son: femenino y masculino; y el tipo de dolor en el pecho, los cuales son: ASY, ATA, NAP y TA.

Para obtener valores estadísticos de ambos, hemos utilizado también un describe y además una tabla y obtenemos lo siguiente:

- Sexo: 21% mujeres y 79% hombres.
- Tipo de dolor en el pecho: el más repetido es el *Chest Pain Type* ASY, con una frecuencia de 496 pacientes, suponiendo así el 54%; NAP en un 22.1%; ATA en un 18.8% y TA con un 5%.



En el tercer gráfico hallamos una gran diferencia entre la cantidad de pacientes según su género, en el que los hombres superan con creces al número de mujeres. De los 918 pacientes que definen el tamaño de la muestra, 725 pertenecen al género masculino, superando casi 5 veces más al género femenino.

Y en el cuarto y último gráfico, es objeto de estudio la variable nominal *Chest Pain Type*, en el que apreciamos una clara superioridad del paciente con un dolor en el pecho de tipo ASY. La cantidad de pacientes ASY son de 496, siendo el tipo de dolor en el pecho que más predomina, aun sumando a todos los pacientes con diferente tipo de dolencia en el pecho. Por el contrario, la dolencia menos común es de tipo TA, que comprende únicamente el 5% de los pacientes.