Actividad 1: Análisis exploratorio y descriptivo de datos

A continuación, analizaremos el dataset “cancer.csv” mediante código de Python y crearemos gráficos para visualizar resultados numéricos. Durante esta actividad, abordaremos el análisis siguiendo 4 pautas:

1. Caracterización de las distribuciones
2. Medidas de dispersión
3. Valores atípicos
4. Correlación entre variables

1. Caracterización de las distribuciones

Representamos la **tabla de frecuencias relativas** para la variable `Uniformity of Cell Size`:

Uniformity of Cell Size	count
1	0.546120
10	0.098097
3	0.076135
2	0.065886
4	0.055637
5	0.043924
8	0.040996
6	0.036603
7	0.027818
9	0.008785

Figura 1. Tabla de frecuencias relativas de Uniformity of Cell Size

> Como podemos ver, la mayoría de observaciones en nuestro dataset (el 54,6%) presenta una uniformidad de tamaño de célula de 1. Seguido de un escaso 9% con uniformidad de tamaño de 10.

Para visualizar mejor como se distribuyen estos valores, creamos un **histograma** de esta misma variable:

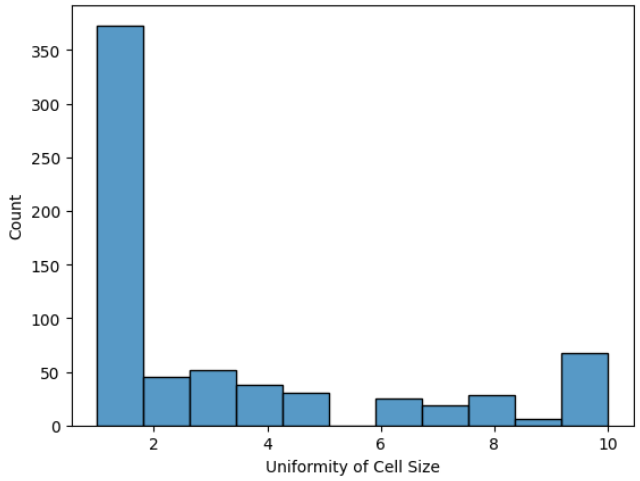

Figura 2. Histograma de Uniformity of Cell Size

> Este es un histograma asimétrico y unimodal (moda=1). Observamos que los demás valores se distribuyen en el gráfico de manera más uniforme.

2. Medidas de dispersión

Para saber cómo de dispersos se encuentran los datos respecto a la media, primero obtenemos información numérica de la variable `Uniformity of Cell Size`:

```
Media:  3.150805270863836
Varianza:  9.39511298695165
Desviación típica:  3.0651448557860443
Coeficiente de variación:  97.28131675194555 %
```

Figura 3. Código de Python de medidas estadísticas de Uniformity of Cell Size

El **coeficiente de variación cercano al 100%** sugiere que los **datos están muy dispersos** respecto a la media, es decir, presentan una alta variabilidad. Esto es lógico ya que tenemos muchos datos concentrados en el valor 1, lo que hará que la media sea menor, en este caso 3. Consecuentemente, la mayoría de valores estarán alejados de la media.

3. Valores atípicos

Antes de proceder al estudio de los valores atípicos de nuestras dos variables elegidas, `Bland Chromatin` y `Single Epithelial Cell Size`, representaremos sus

respectivos **diagramas de caja y bigotes** para identificar de manera visual la existencia de dichos valores atípicos.

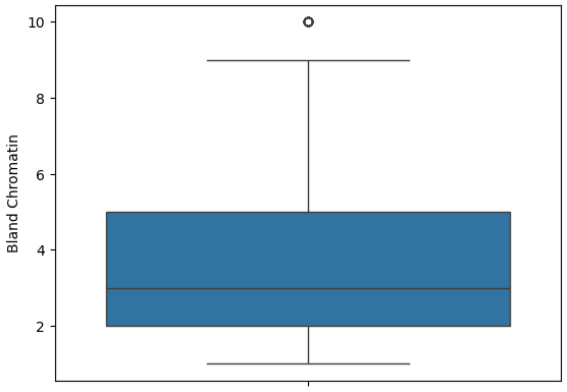

Figura 4. Boxplot de Bland Chromatin

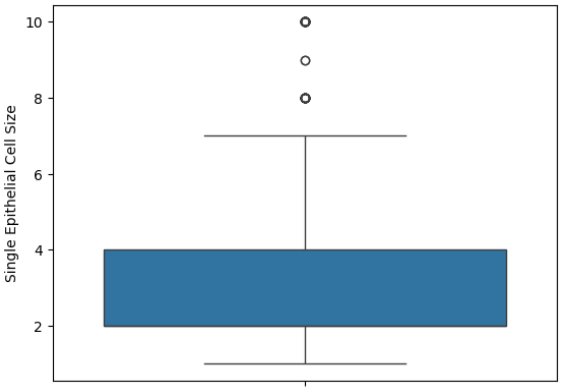

Figura 5. Boxplot de Single Epithelial Cell Size

Como vemos, ambas variables presentan valores atípicos por encima del extremo izquierdo.

Para estudiar estos valores de manera numérica, primero calcularemos el **rango intercuartílico**. Así, identificaremos los valores que están fuera de este rango: los **valores atípicos**. Serán los que superen el valor del extremo izquierdo, ya que no presentan valores inferiores al extremo derecho.

```
1º cuartil:  2.0
3º cuartil:  5.0
Rango intercuarítlico:  3.0
Extremo izquierdo:  9.5
Valores atípicos:  Bland Chromatin
10    20
```

Figura 6. Código de Python para calcular IQR de Bland Chromatin

En el caso de Bland Chromatin, existen 10 valores atípicos en los que Bland Chromatin=20.

También concluimos que el 50% de los datos centrales se concentran entre 2 y 5.

```
1º cuartil:  2.0
3º cuartil:  4.0
Rango intercuarítlico:  2.0
Extremo izquierdo:  7.0
Valores atípicos:  Single Epithelial Cell Size
10    31
8     21
9      2
```

Figura 7. Código de Python para calcular IQR de Single Epithelial Cell Size

En el caso de Single Epithelial Cell Size, existen en total 54 valores atípicos en los que Single Epithelial Cell Size = {8,9,10}, siendo 10 el valor atípico más presente en la variable.

También concluimos que el 50% de los datos centrales se concentran entre 2 y 4.

4. Correlación entre variables

Para estudiar si existe correlación entre las variables de nuestro dataset, elaboramos en Python un gráfico de coordenadas paralelas.

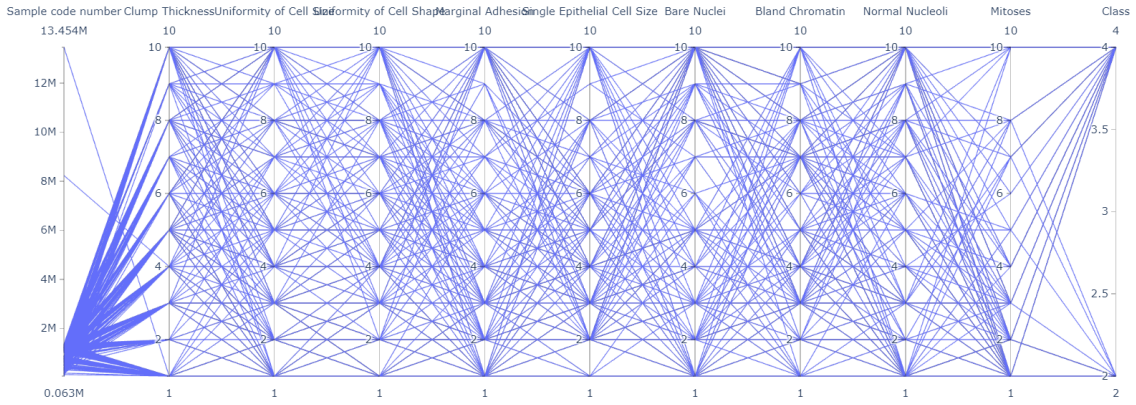

Figura 8. Gráfico de coordenadas paralelas de las variables de cancer.csv

Este gráfico sugiere una **baja correlación positiva de las variables**, ya que las líneas que las unen siguen patrones aleatorios, cruzándose de manera indistinta entre sí.

Obviamos la **nula correlación** entre el código de muestra (`Sample code number`) y las demás variables ya que intuitivamente no tiene nada que ver con los datos sobre cáncer de los pacientes.

La **correlación positiva más baja** parece encontrarse en la variable `Mitoses`, puesto que las líneas se unen de forma aún más aleatoria.

Pero, entre las variables `Uniformity of Cell Size` y `Uniformity of Cell Shape`, vemos que las líneas se unen sin formar tantos “nudos”, lo que sugiere una **fuerte correlación positiva**.

Para ver de manera más clara estas relaciones, volvemos a representar este gráfico utilizando distintos códigos:

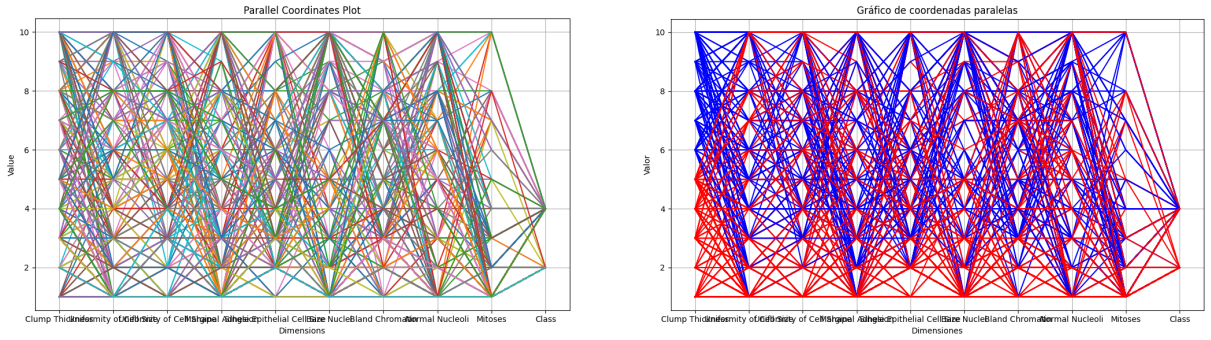

Figura 9. Variaciones de Figura 8.

Como vemos, los gráficos producen mucho ruido visual al contener bastantes variables y datos. En el de la izquierda, uso un código de color rojo si Clump Thickness < 5. Se puede apreciar que este grupo tiende a relacionarse con cáncer de clase 2.

Es decir, si el grosor de la agrupación de células cancerígenas es menor que 5, sugiere que el cáncer del paciente se encuentra en la clase 2. En cambio, si es mayor que 5, el cáncer se encontrará en etapa 4 (clase 4) en el 100% de los casos.

Para corroborar las correlaciones entre variables expuestas, utilizaremos la **matriz de correlaciones**. La elaboraremos en Python usando un sistema de colores:

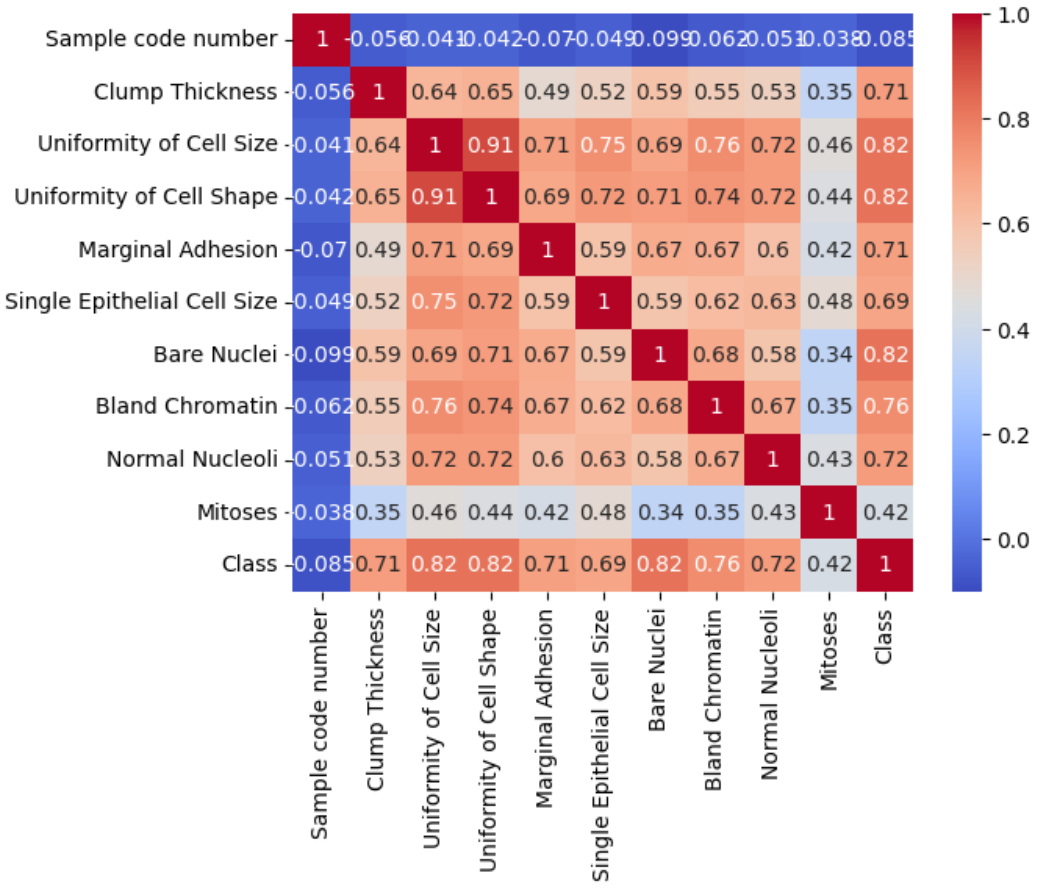

Figura 10. Matriz de correlaciones de variables de cancer.csv

Mediante este gráfico, confirmamos las siguientes **correlaciones positivas fuertes**:

- Uniformity of Cell Size y Uniformity of Cell Shape (0.91)
- La variable Class con: (0.82)
  - Uniformity of Cell Size
  - Uniformity of Cell Shape
  - Bare Nuclei

Además de confirmarse los supuestos de **baja correlación** en la variable Mitoses.

También podemos observar **correlaciones positivas importantes** de Uniformity of Cell Size con:

- Bland Chromatin (0.76)
- Single Epithelial Cell Size (0.75)

> En conclusión, podemos afirmar que cuanto más grande sea la célula más gruesa será, y tendrá mayor probabilidad de ser un cáncer de mayor clase.

Y que, a mayor tamaño de célula, más cromatina blanda habrá en su núcleo, así como mayor serán las células epiteliales.

En definitiva, la mayoría de variables están correlacionadas de manera positiva ya que tratamos con aspectos relacionados con las células.