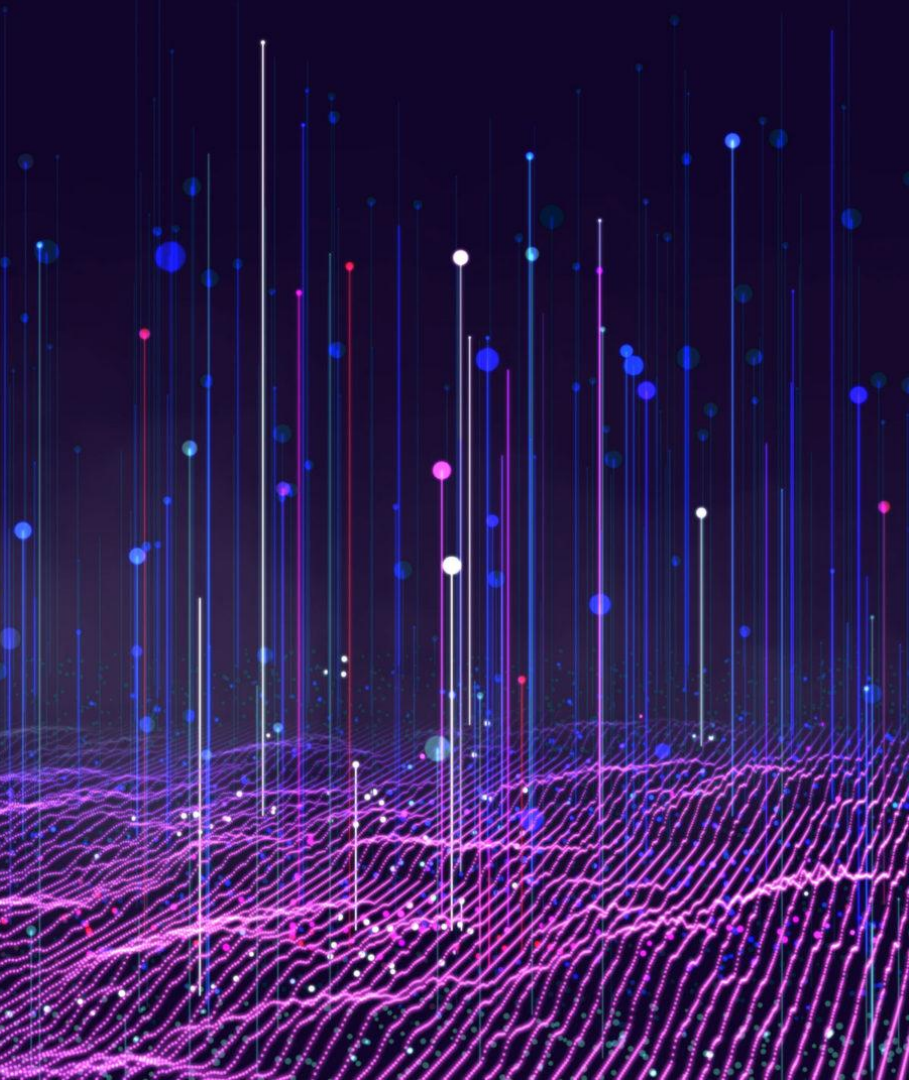


Introducción a la Visualización de Datos

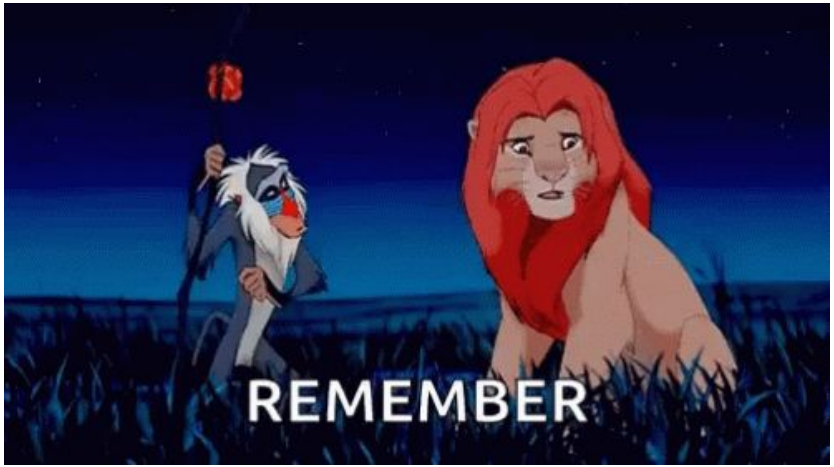
Análisis de Datos con Python

Eduardo Selim Martínez Mayorga

Objetivos de la sesión

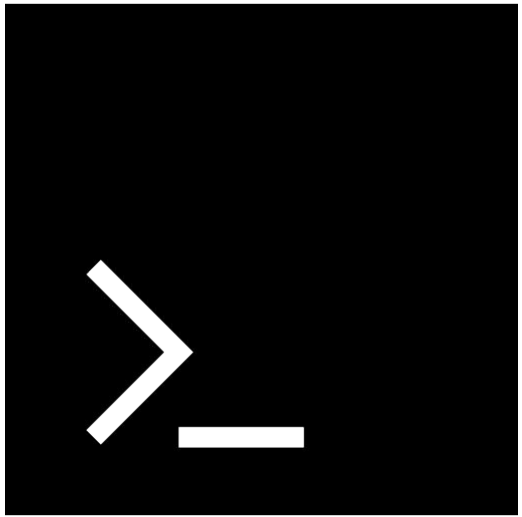
- Comprender el concepto de distribución e identificar la distribución de nuestros datos junto con su importancia.
- Utilizar la biblioteca Seaborn.
- Conocer los boxplots y aprender a generarlos.
- Conocer las tablas de frecuencias y los histogramas como maneras de visualizar distribuciones.
- Clasificar algunas de las formas que generan los histogramas.
- Conocer las gráficas de densidad como una alternativa a los histogramas clásicos.

Recuerda...



- Tipos de datos estructurados
- Medidas de tendencia central
- Desviación estándar
- Medidas de dispersión
- Medidas de posición

Antes de empezar



¡No olvides hacer pull del repo!

El material de la sesión se encuentra ahí.

```
git pull origin master
```

Prework

1. En un diagrama de caja y bigotes, ¿qué percentiles representan los bordes de la caja?
   a. 25 - 50
   b. 25 - 75
   c. 0 - 100
   d. 50 - 75
   e. 50 - 100

Prework

2. En un diagrama de caja y bigotes, ¿cuál es el tamaño máximo de los bigotes?
   a. 1.5 * Rango Intercuartílico
   b. 1.5 * Rango Total
   c. 1.2 * Rango Intercuartílico
   d. Rango Intercuartílico ^ 2
   e. 1.5 * Mediana

Prework

3. En un histograma, ¿qué se grafica en el eje y?
    a. El rango de los valores
    b. Los percentiles
    c. El Rango Intercuartílico
    d. La frecuencia de los valores
    e. El logaritmo de los valores

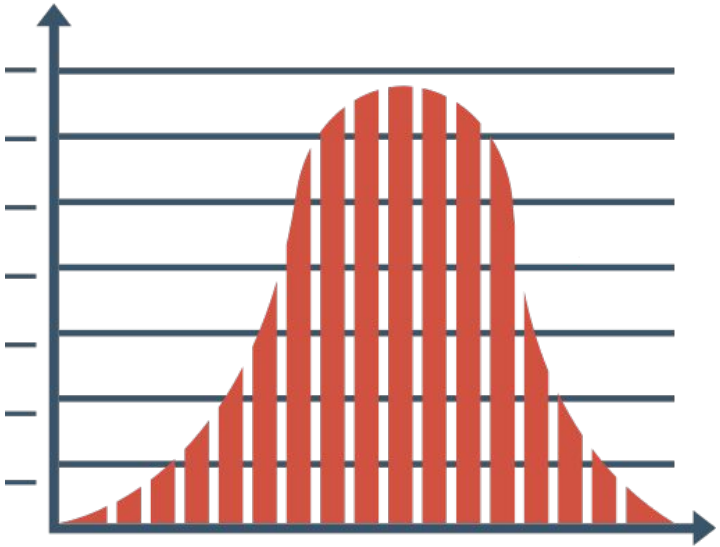
Prework

4. ¿En qué caso decimos que una distribución tiene asimetría positiva?
    a. Cuando la "cola" a la izquierda de la media es más larga que a la derecha
    b. Cuando tenemos dos aglomeraciones de datos
    c. Cuando las colas se extienden mucho más allá de la mayoría de los datos
    d. Cuando la distribución tiene un promedio de 0 y una desviación estándar de 1
    e. Cuando la "cola" a la derecha de la media es más larga que a la izquierda

Prework

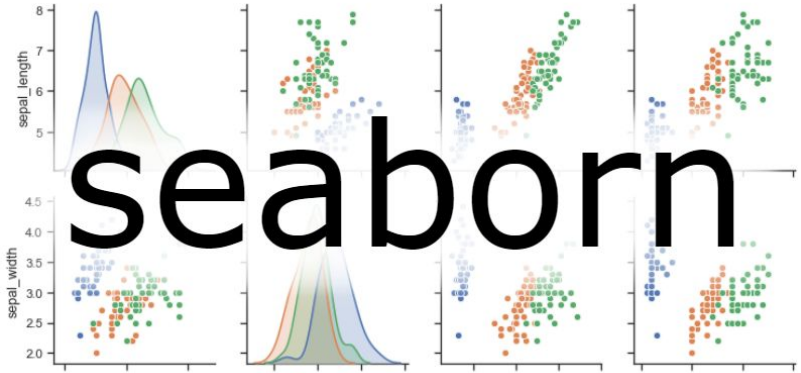
5. ¿Qué características tiene una distribución uniforme?
    a. Es aquélla donde es mucho más probable obtener datos cercanos a la media
    b. Es aquélla donde es mucho más probable obtener datos a la izquierda de la media
    c. Es aquélla donde la probabilidad de obtener alguno de los valores dentro del rango total es prácticamente la misma
    d. Es aquélla donde es mucho más probable obtener datos a la derecha de la media
    e. Es aquélla donde la probabilidad de obtener alguno de los valores dentro del Rango Intercuartílico es prácticamente la misma

Distribuciones



- Como vimos en la sesión anterior, los datos pueden adoptar muchas formas:
  - Pueden estar cerca del promedio
  - Cerca del valor mínimo
  - Cerca del valor máximo
  - Completamente dispersos
- Ya los analizamos con métodos estadísticos
- Ahora los analizaremos mediante visualización

Seaborn



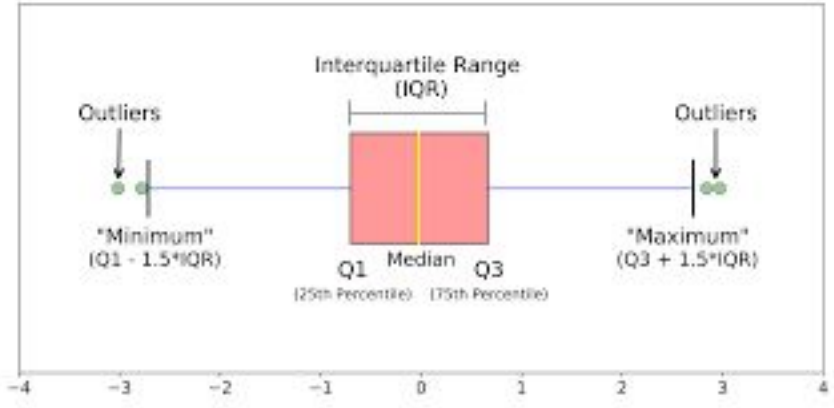
- Es una biblioteca de **Python** para la visualización de datos.
- Está basada en **matplotlib.**

https://seaborn.pydata.org/

https://towardsdatascience.com/data-visualization-using-seaborn-fc24db95a850

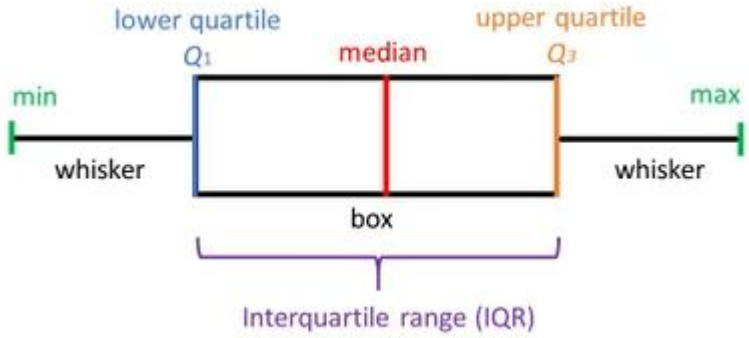
```
pip install seaborn
```

Diagrama de caja



- Es un método de representar gráficamente una serie de datos a través de sus cuartiles.
- Muestra la mediana y los cuartiles asociados.
- También permite revisar de cerca algunos de los valores atípicos a través de los bigotes.

Diagrama de caja



El rango intercuartil es el rango entre el percentil 25 y el percentil 75.

Los bigotes en general se calculan como 1.5 * RIC

Se componen de:

- Rango (sin datos atípicos)
- Datos atípicos.
- Rango intercuartil (también conocido como RIC)
- Cuartiles (denotados como Q1, Q2 y Q3)
- Mediana (Q2)
- Mínimo y máximo.

Ve al Ejemplo 1

Ve al Reto 1

Tabla de frecuencias

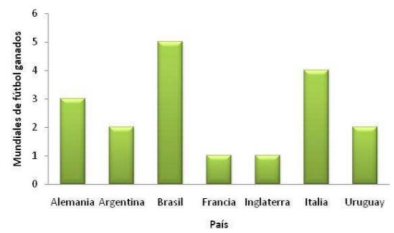
Number of Pets	Number of Students
0	2
1	7
2	3
3	1
4	2

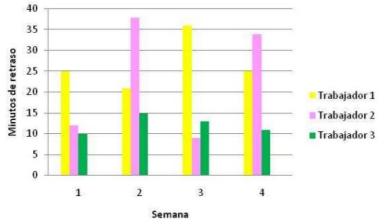
- Permiten seccionar los datos en segmentos.
- Una forma es contabilizar cuántos datos hay por cada posible valor de la columna.
- Otra forma es calcular el porcentaje.
- Lo más idóneo con variables numéricas es seccionar por segmentos de varios valores.

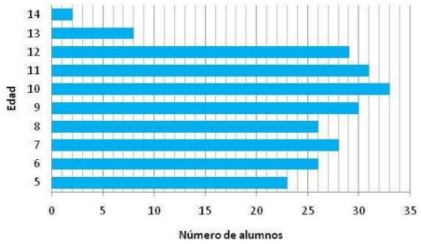
Ve al Ejemplo 2

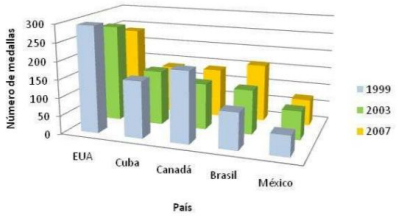
Ve al Reto 2

Histogramas





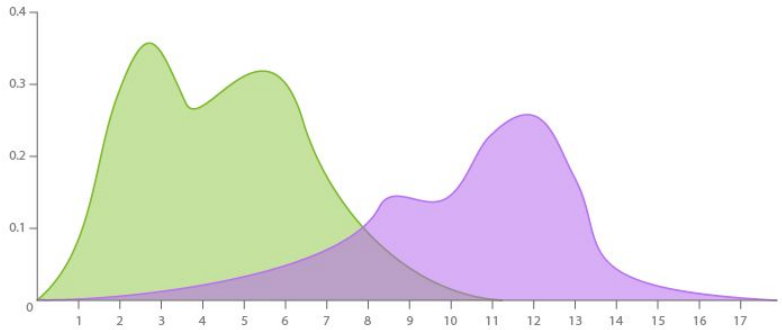




- Cada barra sólida, ya sea vertical u horizontal representa un intervalo/cajita.
- La barra con mayor altura representa la mayor frecuencia.
- La suma de las alturas de las columnas equivale al 100% de los datos.

Ve al Ejemplo 3

Gráficas de densidad



- Permiten visualizar la distribución de datos en un intervalo o período de tiempo continuo.
- Este gráfico es una variación de un Histograma que usa el suavizado de cerner para trazar valores, permitiendo distribuciones más suaves al suavizar el ruido.
- Los picos de un gráfico de densidad ayudan a mostrar dónde los valores se concentran en el intervalo.

Ve al Ejemplo 5

Ve al Reto 4

Postwork / Prework

NO OLVIDES REVISAR TU POSTWORK Y TU PREWORK

Preguntas