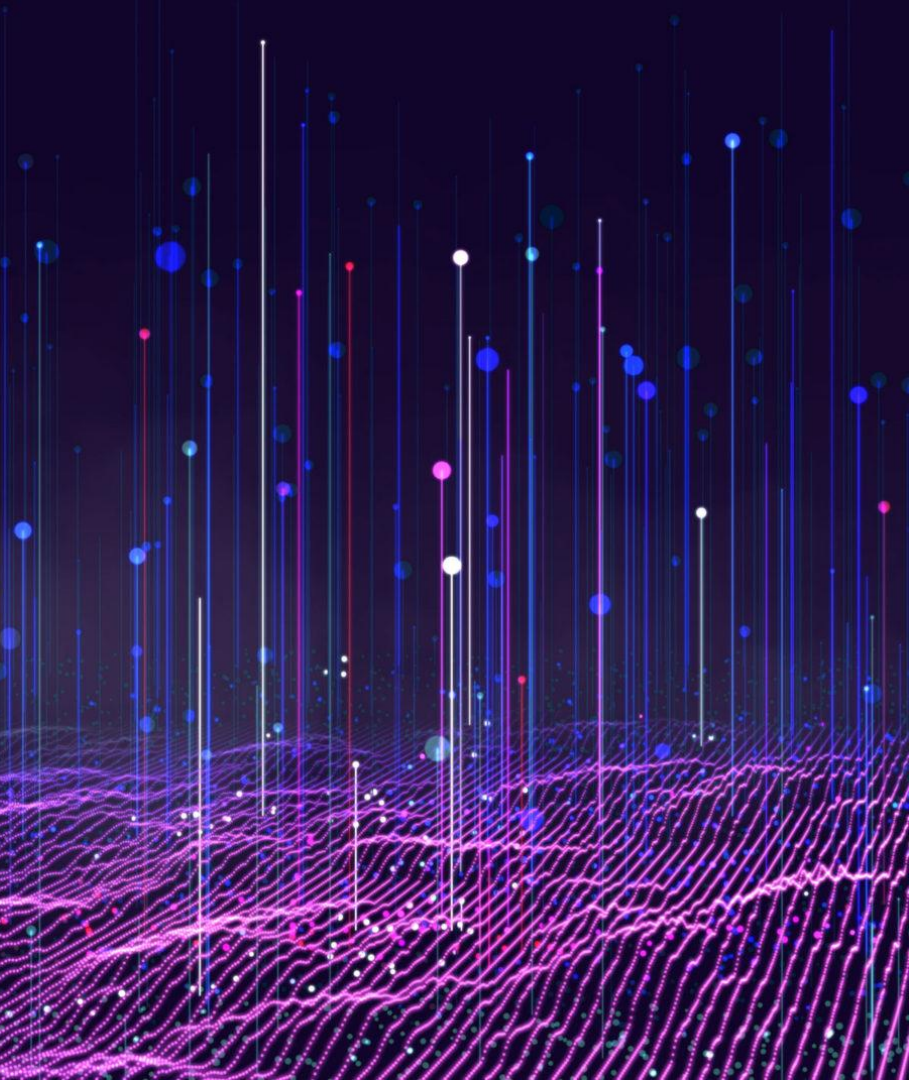


# Estimados de Locación y Variabilidad

## Análisis de Datos con Python

L. en C.C. Manuel Soto Romero

# ¡Bienvenidos!

# Acerca de mí

- Soy Licenciado en Ciencias de la Computación por la UNAM.
- Estudiante de la Maestría en Ciencia e Ingeniería de la Computación
- Desarrollador de Bases de Datos, Analista de Datos
- Profesor en UNAM, UACM

**Áreas de interés:**

Lenguajes de Programación, Lógica Computacional, Ciencia de Datos, Clasificación y Reconocimiento de imágenes

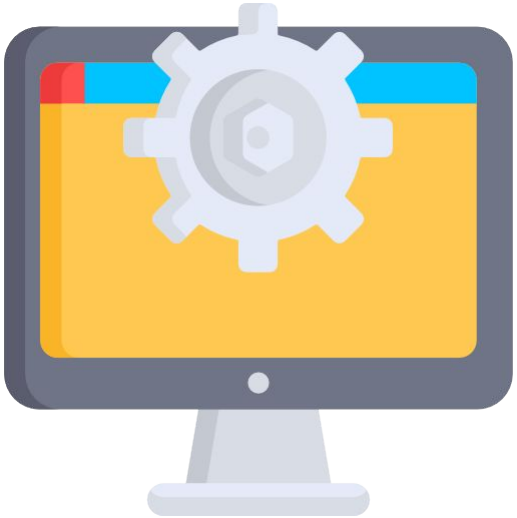
# Objetivos del módulo

Utilizar **Python** y sus bibliotecas para realizar análisis robustos aplicando modelos **estadísticos** y matemáticos que permitan encontrar patrones y relaciones en los datos con el fin de generar visualizaciones de análisis univariados, bivariados y multivariados con **Seaborn** y **Matplotlib** y aplicar modelos de **regresión, clasificación y predicción**.

# Objetivos de la sesión

- Utilizar Jupyter Notebook / Google Colab en conjunción con Google Drive y GitHub.
- Identificar los tipos de datos estructurados que existen.
- Identificar valores típicos y atípicos
- Realizar cálculos estadísticos robustos
- Identificar los estimados de variabilidad y en qué momento son útiles
- Identificar los estadísticos de orden

# Utilización de software



- Usaremos Jupyter Notebook para los ejemplos y retos de la sesión
- Necesitarás clonar el repositorio y los conjuntos de datos que se encuentran en Google Drive
- Necesitarás tener instalada la biblioteca pandas en tu ambiente de trabajo (Conda o MiniConda)

```
git clone https://github.com/manu-msr/adp092020
```

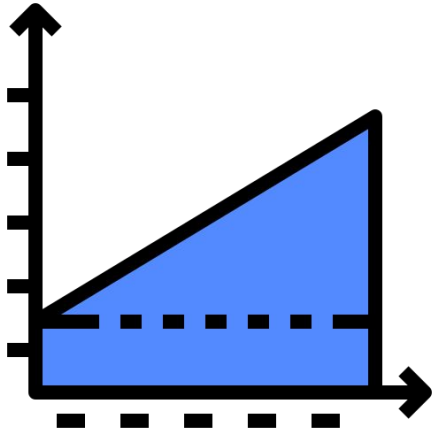
```
pip install jupyter
```

```
pip install pandas
```

```
jupyter notebook
```

Esta instrucción debes ejecutarla donde clonaste el repo

# Clasificación de Variables



VE AL RETO 1

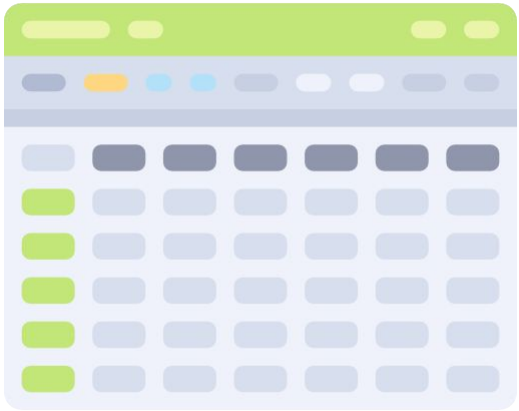
Al conjunto de los distintos valores numéricos que adopta un carácter cuantitativo se llama **variable estadística**.

- Variables **cualitativas** o **categóricas**
- Variables **cuantitativas**
  - **Discretas**
  - **Continuas**

Cuando se estudia a una variable ocupamos los siguientes conceptos:

- **Individuo:** Un elemento
- **Población:** Todos los elementos
- **Muestra:** Una parte de los elementos

# Medidas de tendencia central



- Permiten determinar qué valor **describe mejor** un conjunto de datos.
- A este valor le llamamos **valor típico** o **tendencia central**.
- También son conocidas como **estimados de localidad**.

Las medidas más comunes son:

- **Media, mediana** y moda

# Suma

Por sumatoria se entiende la suma de un conjunto finito de números, que se denota por la letra sigma mayúscula ∑.

$$S = \sum_{i=k}^{k+n} x_i = x_k + x_{k+1} + x_{k+2} + \cdots + x_{k+n-1} + x_{k+n}$$

**JUGUEMOS CON PANDAS**

```
pandas: sum()
```

# Media aritmética

La media aritmética de n valores, es igual a la suma de todos ellos dividida entre n.

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

1. Es una medida totalmente numérica.
2. En su cálculo se toman en cuenta todos los valores de la variable.
3. Es lógica desde el punto de vista algebraico.
4. Es afectada por extremos.
5. La media aritmética es única, o sea, un conjunto de datos numéricos tiene una y sólo una media aritmética

**JUGUEMOS CON PANDAS**

```
pandas: mean()
```

# Mediana

La mediana es el punto central de una serie de datos ordenados de forma ascendente o descendente.

Hay dos formas de calcularla:

- Para número par
- Para número impar

1. En su cálculo no se incluyen todos los valores de la variable.
2. La Mediana no es afectada por valores extremos.
3. No es lógica desde el punto de vista algebraico.

JUGUEMOS CON PANDAS

```
pandas: median()
```

VE AL EJEMPLO 1

VE AL RETO 2

# Tomemos un descanso

## EN 10 MINUTOS CONTINUAMOS

# Trabalenguas

Compré pocas copas,
pocas copas compré,
como compré pocas copas,
pocas copas pagaré.

# Media truncada

Sirve para volver más **robusto** el promedio.

1. Se ordena el conjunto.
2. Se trunca un porcentaje en partes iguales al inicio y final. Ejemplo 50%, quitamos un 25% al inicio y final respectivamente.
3. Se obtiene el promedio con los datos resultantes.

VE AL EJEMPLO 2

`scipy: trim_mean()`

# Desviación estándar

La **desviación estándar** o **desviación típica** se define como la raíz cuadrada de los cuadrados de las desviaciones de los valores de la variable respecto a su media. Esto es:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

Una gran desviación estándar indica que la población está muy dispersa respecto de la media. Una desviación estándar pequeña indica que la población está muy compacta alrededor de la media.

# Paso a paso...

Hallar la desviación estándar de la siguiente serie: 10, 18, 15, 12, 3, 6, 5, 7

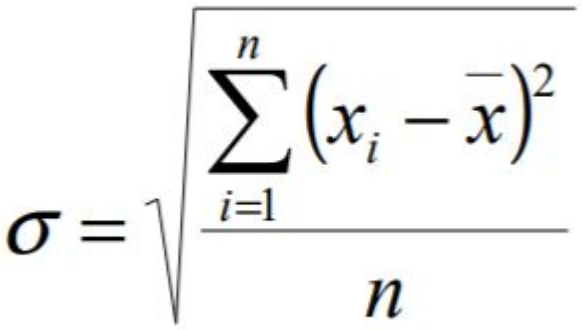
- Promedio: 9.5
- Restamos de cada individuo el promedio:

  (10 - 9.5), (18 - 9.5), …, (7 - 9.5)
  0.5, 8.5, …, -2.5

- Elevamos cada uno al cuadrado:

  0.25, 72.25, …, 6.25

- Sumamos: 190
- Dividimos entre el número de elementos: 190/8 = 23.75
- Ráíz cuadrada: 4.873



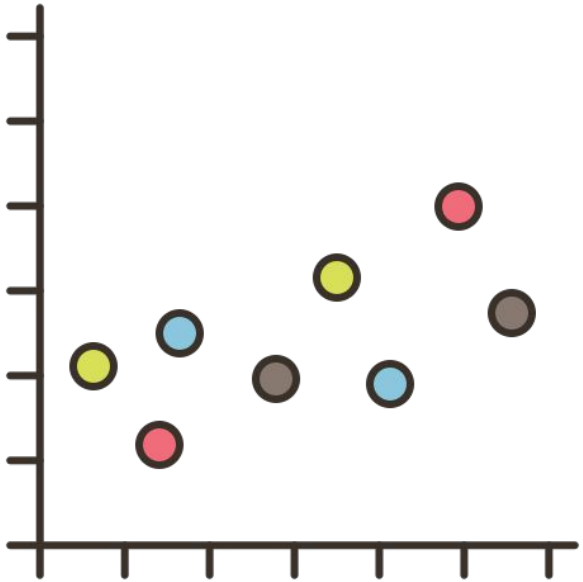

**JUGUEMOS CON PANDAS**

```
pandas: std()
```

**VE AL EJEMPLO 3**

**VE AL RETO 3**

# Medidas de dispersión



- La dispersión **mide** que tan alejados están un conjunto de valores respecto a su **promedio**.
- Cuanto menos disperso sea el conjunto, más cerca del valor medio se encontrarán sus valores.
- Este aspecto es de vital importancia para el estudio de investigaciones.

Se llaman medidas de dispersión aquellas que permiten retratar la distancia de los valores de la variable a un cierto valor central, o que permiten identificar la concentración de los datos en un cierto sector del recorrido de la variable.

# Rango

Es la diferencia entre el valor máximo y el valor mínimo de la variable estadística. Para su cálculo, basta con ordenar los valores de menor a mayor.

1. A medida que el rango es menor, el grado de representatividad de los valores centrales se incrementa.
2. A medida que el rango es mayor, la distribución está menos concentrada o más dispersa.
3. Su cálculo es extremadamente sencillo.
4. Tiene gran aplicación en procesos de control de calidad.
5. Tiene el inconveniente de que sólo depende de los valores extremos.

# Ejemplo

- Si se conoce que el valor promedio de días de espera para obtener una licencia de manejo, es de 5 días en la oficina A, y de 7 días en la oficina B, con esta única información no es posible hacer una elección adecuada.
- Sin embargo, si se sabe que en la oficina A, el número mínimo de días de espera es de 3 y el máximo de 15, mientras que en la oficina B, los valores son 3 y 8 días respectivamente, se podrá tomar una decisión más adecuada para acudir a obtener la licencia, gracias a esta información adicional.

# Medidas de posición

Medidas de posición para datos agrupados y no agrupados

- Los cuantiles son los valores de la distribución que la dividen en partes iguales, es decir, en intervalos que comprenden el mismo número de valores.
- Cuando la distribución contiene un número alto de intervalos o de marcas y se requiere obtener un promedio de una parte de ella. Generalmente, se divide la distribución en cuatro, en diez o en cien partes.
- Los cuantiles más usados son los **percentiles**, cuando dividen la distribución en cien partes, los **deciles**, cuando dividen la distribución en diez partes y los **cuartiles**, cuando dividen la distribución en cuatro partes.

# Percentiles

Son números que dividen en 100 partes iguales un conjunto de datos ordenados. Es decir, El percentil k es un valor que deja aproximadamente el k por ciento de los datos por abajo de él. Se denota por medio de P(k%).

# Ejemplo

- En un estudio de ingresos mensuales de la población económicamente activa, revela que el percentil 90 (P90) es $20,000. Esto significa que aproximadamente el 90% de las personas tienen ingresos que son menores o iguales a $20,000, y por supuesto, el 10% tiene ingresos mayores o iguales a dicho valor.
- En el ejemplo anterior se tomó el percentil 90 pero se podría haber considerado cualquier valor, por ejemplo, 70, 80 entre otros. Fundamentalmente cuando la distribución de frecuencia es asimétrica, puede ser más útil e informativo, resumir la distribución de la variable en estudio, mediante los percentiles.

# Rango intercuartil

Para un cálculo de rangos **más eficiente**, se eliminan los valores extremadamente alejados aplicando el rango intercuartil que es una medida de variabilidad adecuada cuando la medida de posición central empleada ha sido la mediana y se define como

**Rango Intercuartil = P(75) – P(25)**

VE AL EJEMPLO 4

VE AL RETO 4

# Postwork / Prework

**NO OLVIDES REVISAR TU POSTWORK Y TU PREWORK**

# Preguntas