ACÁMICA

# ¡Bienvenidos/as a Data Science!

# Agenda

¿Cómo anduvieron?

Repaso: Probabilidad

Explicación: Probabilidad y Estadística, Pandas

Break

Explicación: valores faltantes

¿Sabías que…?

Hands-on training

Cierre

# Proyecto 1: Análisis Exploratorio de Datos (EDA)

# Análisis Exploratorio de Datos (EDA)

| fase | ADQUISICIÓN Y EXPLORACIÓN | | MODELADO | | | | DEPLOY |
|---|---|---|---|---|---|---|---|
| entrega | Exploración de datos | Feature Engineering | Regresión | Optimización de parámetros | Procesam. del lenguaje natural | Sistema de recomendación | Publicación de modelos |
| tiempo | SEM 1 | SEM 5 | SEM 7 | SEM 11 | SEM 13 | SEM 18 | SEM 22 |
| | SEM 2 | SEM 6 | SEM 8 | SEM 12 | SEM 14 | SEM 19 | SEM 23 |
| | SEM 3 | | SEM 9 | | SEM 15 | SEM 20 | SEM 24 |
| | SEM 4 | | SEM 10 | | SEM 16 | SEM 21 | |
| | | | | | SEM 17 | | |

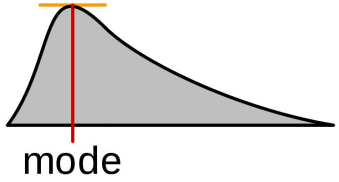
# Proyecto EDA: Hoja de ruta

| | | Usted Está Aquí | Entrega 1 | | Entrega 2 | |
|---|---|---|---|---|---|---|
| SEM 1 | SEM 2 | SEM 3 | SEM 4 | SEM 5 | SEM 6 | SEM 7 |
| • Python | • Numpy<br>• Pandas<br>• Estadística | • Pandas<br>• Estadística<br>• Matplotlib | • Seaborn<br>• Ejercicio/s Integradores | | | |

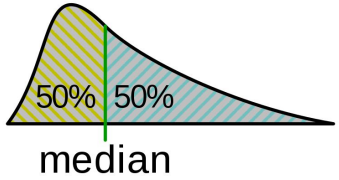
# ¿Cómo anduvieron?

# Tarea

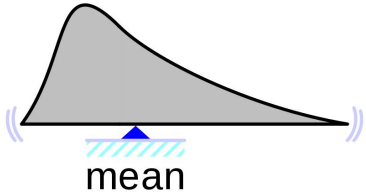
La **moda** es el valor de una serie de datos que aparece con más frecuencia.




La **mediana** es el valor medio de una secuencia ordenada de datos.



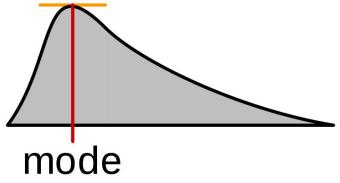

La **media aritmética** puede o no coincidir con alguna de ellas.



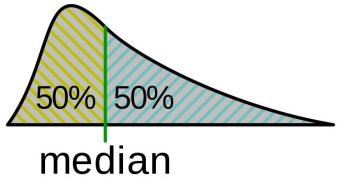

Fuente: Wikipedia

# Tarea

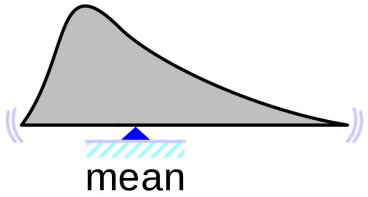
La **moda** es el valor de una serie de datos que aparece con más frecuencia.




La **mediana** es el valor medio de una secuencia ordenada de datos.




La **media aritmética** puede o no coincidir con alguna de ellas.




**EN LA NORMAL LAS TRES COINCIDEN. ¿POR QUÉ?**

Fuente: Wikipedia

# Repaso: Probabilidad

# Probabilidad: Definición

**Frecuentista**: si hacemos un experimento muchas (!) veces, la probabilidad está asociada a la **frecuencia** con que ocurre cada posible valor de la variable aleatoria.

**Bayesiana**: medida de la *confianza* o *certidumbre* de que un suceso ocurra. La mejor medida de la incertidumbre es la probabilidad.

# Probabilidad: Variables aleatorias

***X* variable aleatoria**. Posibles resultados de un proceso aleatorio:

$X_{moneda}$: {cara, ceca}

$X_{dado}$: {1,2,3,4,5,6}

$X_{clima}$ : {lluvia, no lluvia}

$X_{clima}$ : {cuánto llovió}

$X_{avión}$ : {accidente, no-accidente}

# Probabilidad: Variables aleatorias

**PROBABILIDAD**

## Variables **discretas**

- Son aquellas que se *cuentan*
- Pueden estar acotadas o no

Ejemplo: edades (en años), número de hijos, cantidad de dormitorios en una casa, etc.

## Variables **continuas**

- Son aquellas que se *miden*
- Pueden estar acotadas o no

Ejemplo: altura de una persona, temperaturas, edades (medidas en tiempo transcurrido desde el nacimiento), etc.

# Variables discretas: Distribución

La distribución de probabilidad de una variable aleatoria es una función que asigna a cada suceso definido sobre la variable la probabilidad de que dicho suceso ocurra.

Fuente: Wikipedia

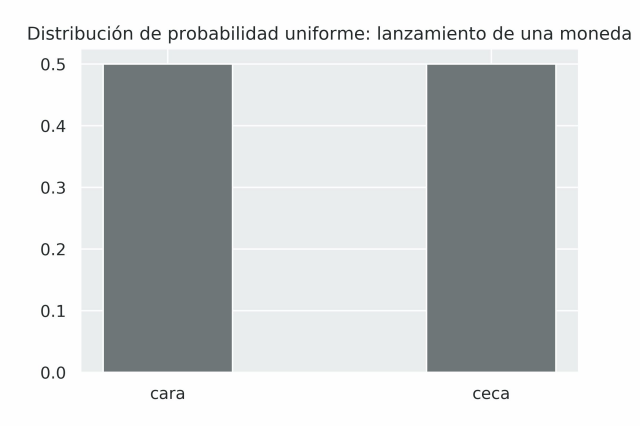
# Variables discretas: Distribución **uniforme**

La distribución de probabilidad **uniforme** asigna la misma probabilidad para todo un rango de valores.

Ejemplos: moneda, dado.

$X_{moneda}$: {cara, ceca}

P(X = cara, ceca) = 1/2

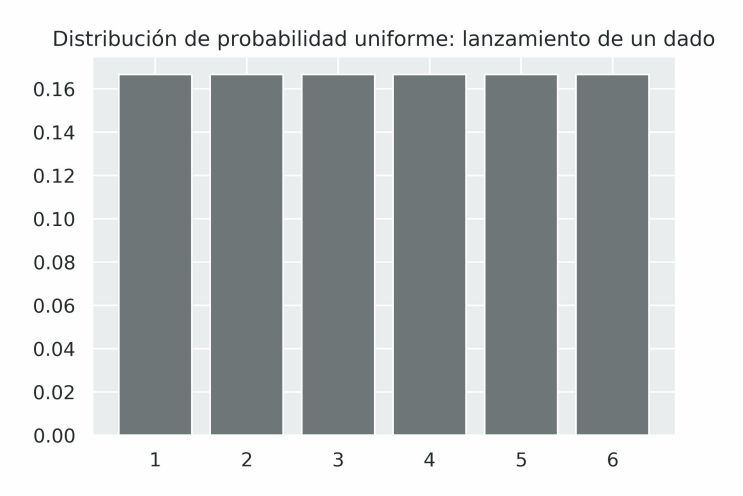


# Variables discretas: Distribución **uniforme**

La distribución de probabilidad **uniforme** asigna la misma probabilidad para todo un rango de valores.

Ejemplos: moneda, dado.

> $X_{dado}$: {1,2,3,4,5,6}
>
> P(X = 1,2,3,4,5,6) = 1/6



# Variables discretas: Distribución binomial

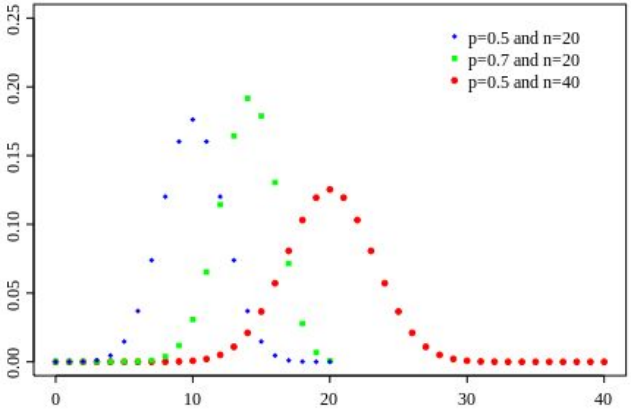
Si tiro *n* veces una moneda con probabilidad *p* de sacar cara (o ceca), ¿cuál es la probabilidad de sacar x caras (o cecas)?

# Variables discretas: Distribución binomial

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad 0 \leq p \leq 1$$

donde $x = \{0, 1, 2, \ldots, n\}$

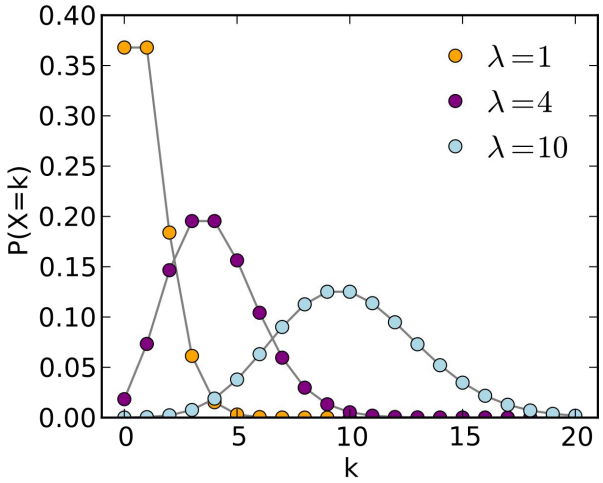
$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$ (combinatorio)




Fuente: Wikipedia

# Variables discretas: Distribución **Poisson**

$$P(k \text{ events in interval}) = e^{-\lambda}\frac{\lambda^k}{k!}$$




La distribución de probabilidad **Poisson** sirve para describir la probabilidad de *k* cantidad de eventos en cierto intervalo de tiempo.

Ejemplos: número de inundaciones en una ciudad por año o década, número de mensajes que mandamos por WhatsApp por día, número de goles en un partido del Mundial de fútbol.

Fuente: Wikipedia

Para variables **continuas**…

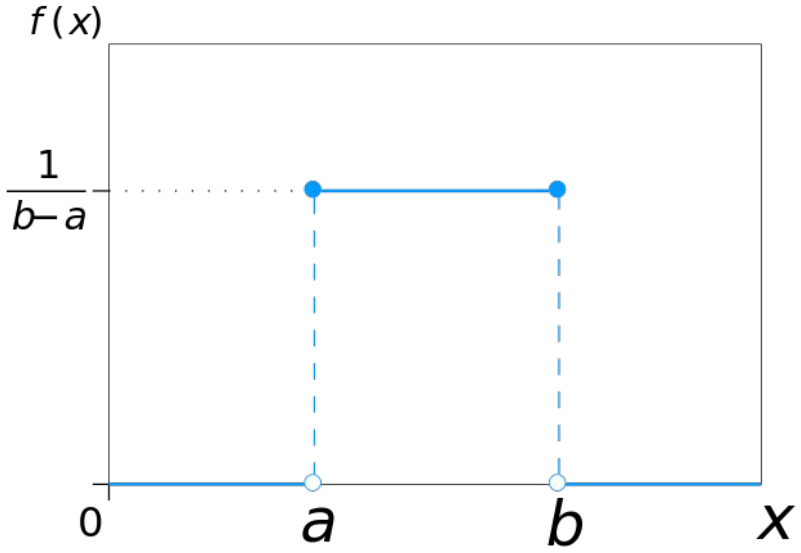
**¿qué concepto de probabilidad usamos?**

Para variables **continuas**,

¡usamos el concepto de **densidad** de probabilidad!

# Probabilidad: Densidad uniforme

Muy parecida a su versión discreta.



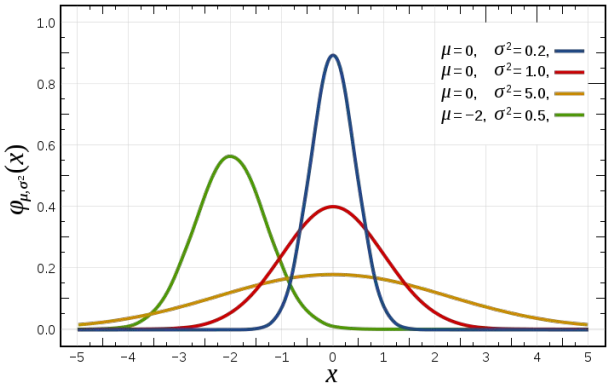

Fuente: Wikipedia

# Probabilidad: Densidad normal o Gaussiana

¡La más famosa de las distribuciones!

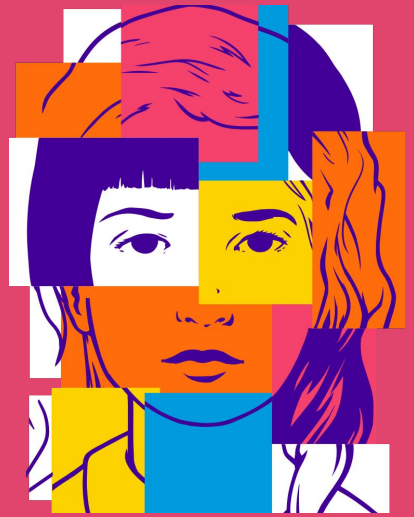
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Parámetros**:
$\mu$: valor medio
$\sigma$: desviación estándar




Fuente: Wikipedia

# + Probabilidad



# Probabilidad y Estadística

| | |
|---|---|
| **PROBABILIDAD** | Qué espero ver.<br>Modelos sobre la naturaleza o nuestro problema |
| **ESTADÍSTICA** | Lo que vi. **Preguntas**: ¿tiene sentido con mi modelo?¿Qué puedo aprender?<br>(Lo que mido en el laboratorio) |
| **NUEVA PROBABILIDAD** | ¿Entendí lo que estaba pasando?<br>(Lo que aprendí sobre la naturaleza. ¿Nuevos modelos?) |

# Nos ponemos técnicos: **Esperanza**

Sea *X* una variable aleatoria cuya distribución (discreta) es P(x).

$$E[X] = \sum_{x \in X} x P(x)$$

¡Esperanza, valor esperado y muchos nombres más!

# Nos ponemos técnicos: **Esperanza**

Sea *X* una variable aleatoria cuya distribución (discreta) es P(x).

$$E[X] = \sum_{x \in X} x P(x)$$

¡Esperanza, valor esperado y muchos nombres más!

¡Es menos difícil de lo que parece! Veamos un ejemplo, **el dado.**
Al tratarse de un caso finito (acotado),

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i \, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

¿Cuánto valen y qué representa $k$, $x_i$ y $p_i$?

Fuente: Wikipedia

# Nos ponemos técnicos: **Esperanza**

Sea *X* una variable aleatoria cuya distribución (discreta) es P(x).

$$E[X] = \sum_{x \in X} x P(x)$$

¡Esperanza, valor esperado y muchos nombres más!

¡Es menos difícil de lo que parece! Veamos un ejemplo, **el dado.**
Al tratarse de un caso finito (acotado),

$$\mathrm{E}[X] = \sum_{i=1}^{k} x_i \, p_i = x_1 p_1 + x_2 p_2 + \cdots + x_k p_k$$

¿Cuánto valen y qué representa ***k***, $\boldsymbol{x_i}$ y $\boldsymbol{p_i}$?

$$\mathrm{E}[X] = 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + 3 \cdot \frac{1}{6} + 4 \cdot \frac{1}{6} + 5 \cdot \frac{1}{6} + 6 \cdot \frac{1}{6} = 3.5$$

¡Es un promedio pesado/ponderado!

Fuente: Wikipedia

# Y seguimoooos...**Varianza**

Sea *X* una variable aleatoria cuya distribución (discreta) es P(x).
La varianza se calcula como:

$$\operatorname{Var}(X) = \operatorname{E}\left[(X - \mu)^2\right] \qquad \mu = \operatorname{E}[X]$$

Fuente: Wikipedia

# Y seguimoooos...**Varianza**

La varianza nos da una idea de cuán dispersos están los valores de una distribución con respecto a su valor medio.

En el caso del dado,

$$\operatorname{Var}(X) = \sum_{i=1}^{n} p_i \cdot (x_i - \mu)^2$$

$$\operatorname{Var}(X) = \sum_{i=1}^{6} \frac{1}{6}\left(i - \frac{7}{2}\right)^2$$

$$= \frac{1}{6}\left((-5/2)^2 + (-3/2)^2 + (-1/2)^2 + (1/2)^2 + (3/2)^2 + (5/2)^2\right)$$

$$= \frac{35}{12} \approx 2.92.$$

Fuente: Wikipedia

# **Desviación estándar** y Varianza

La desviación estándar es, simplemente, la raíz cuadrada de la varianza:

$$SD = \sqrt{Var(x)}$$

Fuente: Wikipedia

# Esperanza y Varianza: Caso continuo

Sea *X* una variable aleatoria cuya densidad de probabilidad es f(x).

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx$$

$$\mathrm{Var}(X) = \int_{\mathbb{R}} (x-\mu)^2 f(x)\, dx$$

Fuente: Wikipedia

# Esperanza y Varianza: Caso continuo

Sea *X* una variable aleatoria cuya densidad de probabilidad es f(x).

$$\mathrm{E}[X] = \int_{\mathbb{R}} x f(x)\, dx \qquad \mathrm{Var}(X) = \int_{\mathbb{R}} (x-\mu)^2 f(x)\, dx$$

Si hacemos estos cálculo para una **Gaussiana**, se obtiene que:

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad \mu = \mathrm{E}[X] \qquad \mathrm{Var}(X) = \sigma^2$$

Y por eso se suelen usar esas símbolos para representar el valor medio, la varianza y la desviación estándar.

Fuente: Wikipedia

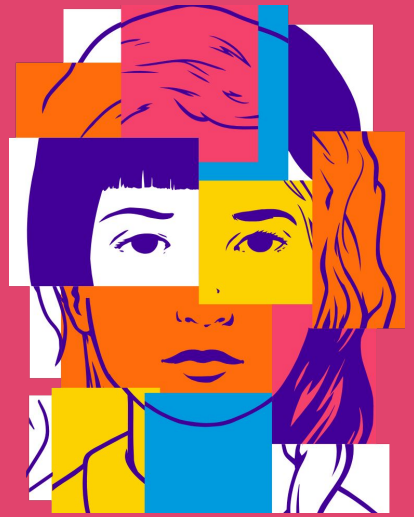
ENTONCES...

Si conozco la distribución de probabilidad
(o densidad de probabilidad)
con las fórmulas que mostramos,
**podemos calcular su esperanza, varianza y más.**

Pero **en general** no conocemos las distribuciones,
sino que tenemos **datos**.

Ahí es donde entra la **Estadística.**

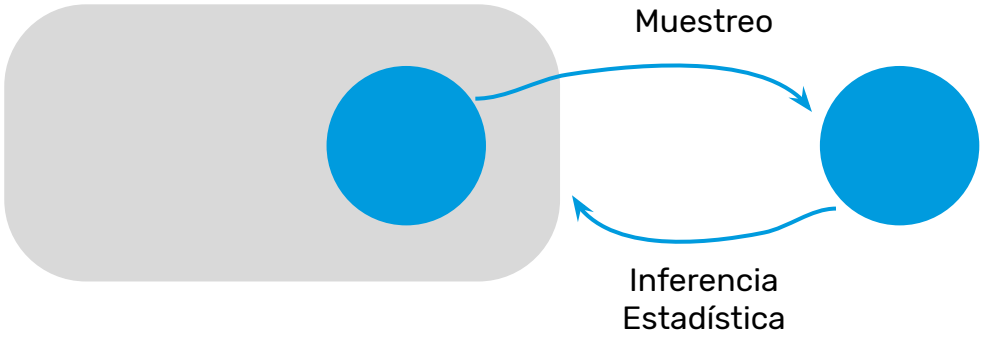
# Estadística



# Población y Muestra

**POBLACIÓN**
(Parámetros)

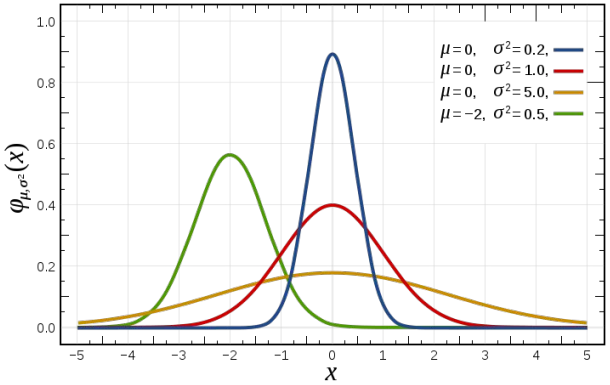
**MUESTRA**
(Estadísticos)



# Relación entre estadísticos y parámetros

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Parámetros**:
$\mu$: valor medio
$\sigma$: desviación estándar

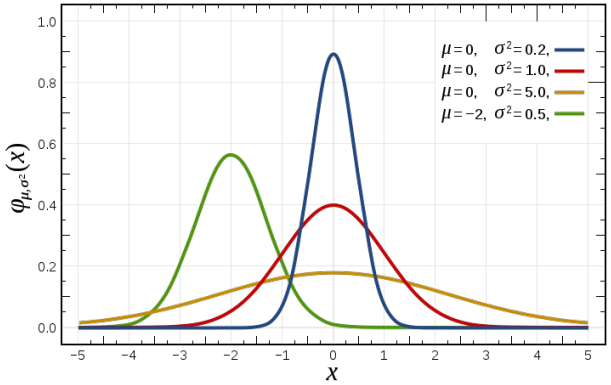



Fuente: Wikipedia

# Relación entre estadísticos y parámetros

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

**Parámetros**:
$\mu$: valor medio
$\sigma$: desviación estándar




Si nuestros datos tienen una distribución Gaussiana

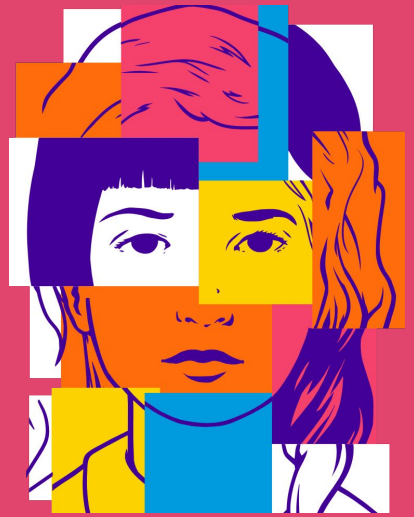
| Parámetro | Estadístico |
|---|---|
| $\mu$ | Promedio de los datos |
| $\sigma$ | Desviación Estándar Calculada de los datos |

Fuente: Wikipedia

# Volvemos al ejemplo del dado cargado...

- **Adquisición**: ¿qué valores anoto de los lanzamientos?
- **Organización**: ¿cómo guardo y agrupo estos valores?
- **Análisis**: ¿en qué rango están estos valores? ¿Cómo son de precisos o de dispersos?
- **Interpretación**: ¿qué dicen mis datos sobre si el dado está cargado o no?
- **Presentación**: ¿cómo comunico mis conclusiones? ¿Qué números o gráficas utilizo para que se me entienda bien?

# Pandas



# DATASET

Es el conjunto de datos que utilizaremos en el workflow de data science. Los podemos generar, obtener de terceros o simular.

| | |
|---|---|
| datasets **estructurados** | similar a planilla de cálculo. Información pre-procesada. Suelen venir en .txt, .csv, .xlsx, .json, etc. |
| datasets **no estructurados** | audio, imágenes, texto en crudo<br>humanos / redes neuronales |

# DATASET

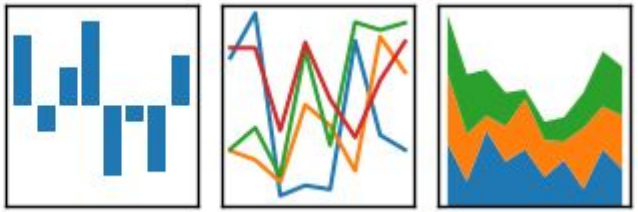
datasets **estructurados**

similar a planilla de cálculo. Información pre-procesada. Suelen venir en .txt, .csv, .xlsx, .json, etc.

Para trabajar con datasets estructurados (y bueno, más), la librería estándar de Python es:

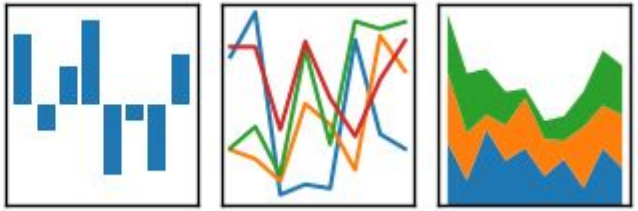
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



# ¿Qué aprendieron en los videos de la plataforma?

# Repaso

1. ¿Cuál es la diferencia entre *loc* e *iloc*? Respuestas: acá y acá.
2. ¿Qué hacen *describe*, *info* y *shape*?
3. ¿Cómo se relacionan *unique* y *value_counts*?
4. ¿Cuál es la diferencia entre *concat*, *append* y *merge*? Respuestas: acá y acá.
5. ¿Cuál es la diferencia entre *apply* y *applymap*? Respuestas: acá y acá.

# ¡BREAK!

Ph. Credit: Drew Coffmann

# ¿Por qué hay valores faltantes?

**MCAR**
(Missing Completely At Random)

***P(missing)*** es la misma para todas las instancias y no depende de las medidas de esa u otras variables.
Ej: Se perdió la respuesta para una encuesta.
En general, no es el caso.

# ¿Por qué hay valores faltantes?

**MCAR**
(Missing Completely At Random)

***P(missing)*** es la misma para todas las instancias y no depende de las medidas de esa u otras variables.
Ej: Se perdió la respuesta para una encuesta.
En general, no es el caso.

**MAR**
(Missing At Random)

***P(missing)*** no depende del valor faltante, pero sí de otras variables observables.
Ej: ¿Cuánto gana? Tal vez no responden porque consideran la pregunta inapropiada, independiente del monto que ganen.

# ¿Por qué hay valores faltantes?

**MCAR**
(Missing Completely At Random)

***P(missing)*** es la misma para todas las instancias y no depende de las medidas de esa u otras variables.
Ej: Se perdió la respuesta para una encuesta.
En general, no es el caso.

**MAR**
(Missing At Random)

***P(missing)*** no depende del valor faltante, pero sí de otras variables observables.
Ej: ¿Cuánto gana? Tal vez no responden porque consideran la pregunta inapropiada, independiente del monto que ganen.

**MNAR**
(Missing Not At Random)

***P(missing)*** depende de la variable que queremos medir.
Ej: ¿Cuánto gana? (si es muy alta, quizás no contestan).

Valores Faltantes

# Qué hacemos con los valores faltantes?

## * Eliminar datos con problemas:

- **Por Fila**: eliminamos las instancias que tienen algún valor faltante. Puede sesgar nuestros resultados.
  (Ej: eliminamos aquellas personas que no respondieron cuánto ganan porque ganan mucho).
- **Por Columna**: eliminamos aquella columna/variable que tiene muchos valores faltantes. Podemos perder información relevante.

## *Imputación

Rellenamos los valores faltantes con estadísticos obtenidos de los datos que sí tenemos. Por ejemplo, con el promedio, la mediana o la moda.

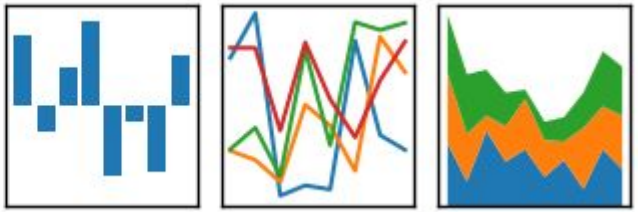
## *Agregar una variable categórica binaria (¿qué es eso?)

por atributo que indique si hay un valor faltante o no.

# Valores Faltantes con

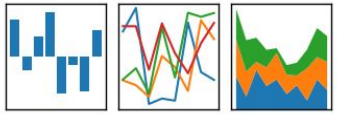
pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



# Valores faltantes con pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



**¿Qué hace cada una de las siguientes instrucciones?**
**¿Cuáles son algunos de sus argumentos?**

1. *df.isna(), df.isnull()*

2. *df.dropna(), df.dropna(subset=...)*

3. *df.fillna(), df.fillna(inplace = ...)*

## Comentarios

1. La *mejor* técnica para lidiar con los valores faltantes depende del mecanismo con el que se hayan generado. Y del problema[1].

[1]Se van a cansar de escucharnos decir “depende del problema”

## Comentarios

1. La *mejor* técnica para lidiar con los valores faltantes depende del mecanismo con el que se hayan generado. Y del problema[1].

2. Esto no quiere decir que siempre apliquemos la *mejor* técnica. Existen MUCHÍSIMAS técnicas para lidiar con valores faltantes. No las conocemos todas.

## Comentarios

1. La *mejor* técnica para lidiar con los valores faltantes depende del mecanismo con el que se hayan generado Y del problema[1].

2. Esto no quiere decir que siempre apliquemos la *mejor* técnica. Existen MUCHÍSIMAS técnicas para lidiar con valores faltante. No las conocemos todas.

3. Se pueden usar técnicas de aprendizaje automático para imputar valores faltantes.

## Comentarios

1. La *mejor* técnica para lidiar con los valores faltantes depende del mecanismo con el que se hayan generado Y del problema[1].

2. Esto no quiere decir que siempre apliquemos la *mejor* técnica. Existen MUCHÍSIMAS técnicas para lidiar con valores faltante. No las conocemos todas.

3. Se pueden usar técnicas de aprendizaje automático para imputar valores faltantes.

4. El análisis estadístico con valores faltantes es toda una área de la estadística. Como profesionales, profundizaremos o no dependiendo de nuestros intereses y necesidades.

# Hands-on training

**Hands-on training**

# DS_Clase_05_Pandas.ipynb

# Sabías que...

# ¿Qué es un censo?

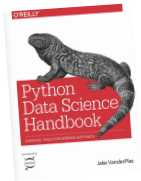
The New York Times Magazine

## The Census: Why We Can't Count

By James Gleick

# Recursos

## Recursos




- Valores faltantes: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4
- Capítulo 3, “Data Manipulation With Pandas”, de Python Data Science Handbook
- Estadística en general: Serie de cinco artículos sobre Estadística en Data Science: https://towardsdatascience.com/statistics-is-the-grammar-of-data-science-part-2-8be5685065b5

# Para la próxima

1. Averiguar qué es la asimetría estadística (*skewness*) y la curtosis (*kurtosis*)

2. Ver los videos de la plataforma “Visualización de datos”

3. Completar Notebooks atrasados.

ACÁMICA