UNIVERSIDAD EAFIT

Científico de Datos

Nivel Básico

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación

UNIVERSIDAD EAFIT

Tema:

Estadística descriptiva y Pandas

Aliados:

Microsoft

idata Advanced analytics for business

Vigilada Mineducación

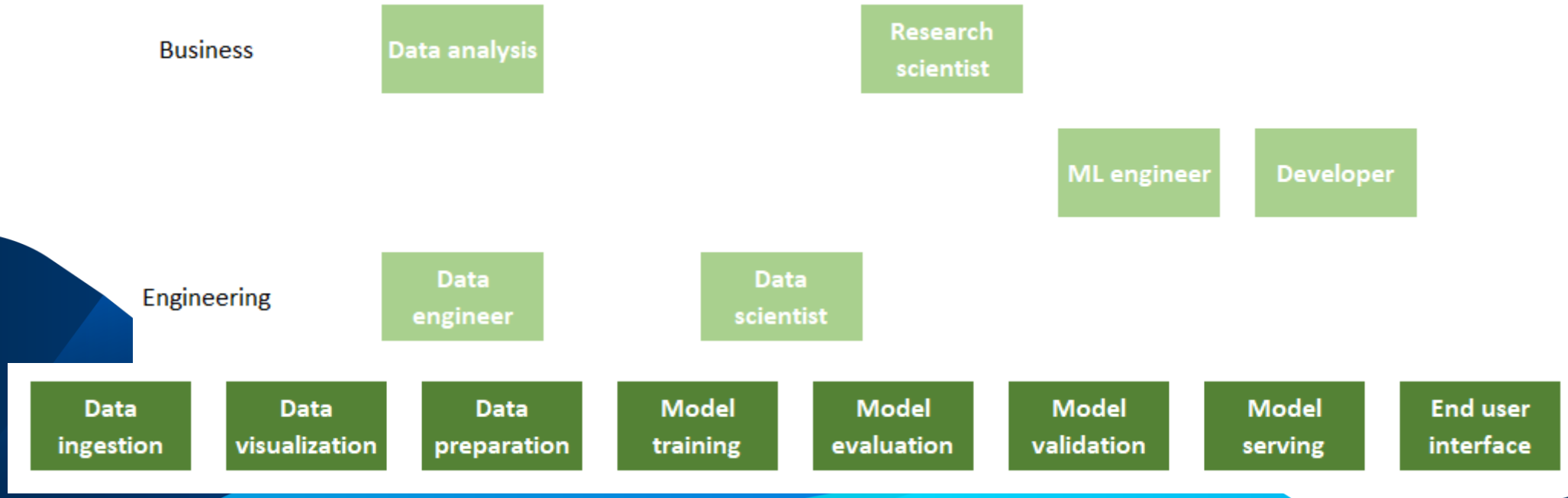
Estadística descriptiva vs Estadística diferenciales

- Diferencial: resumir (a través de métricas).
- Inferencial: deducir, sacar conclusiones y predicciones. Cosas que pueden pasar.

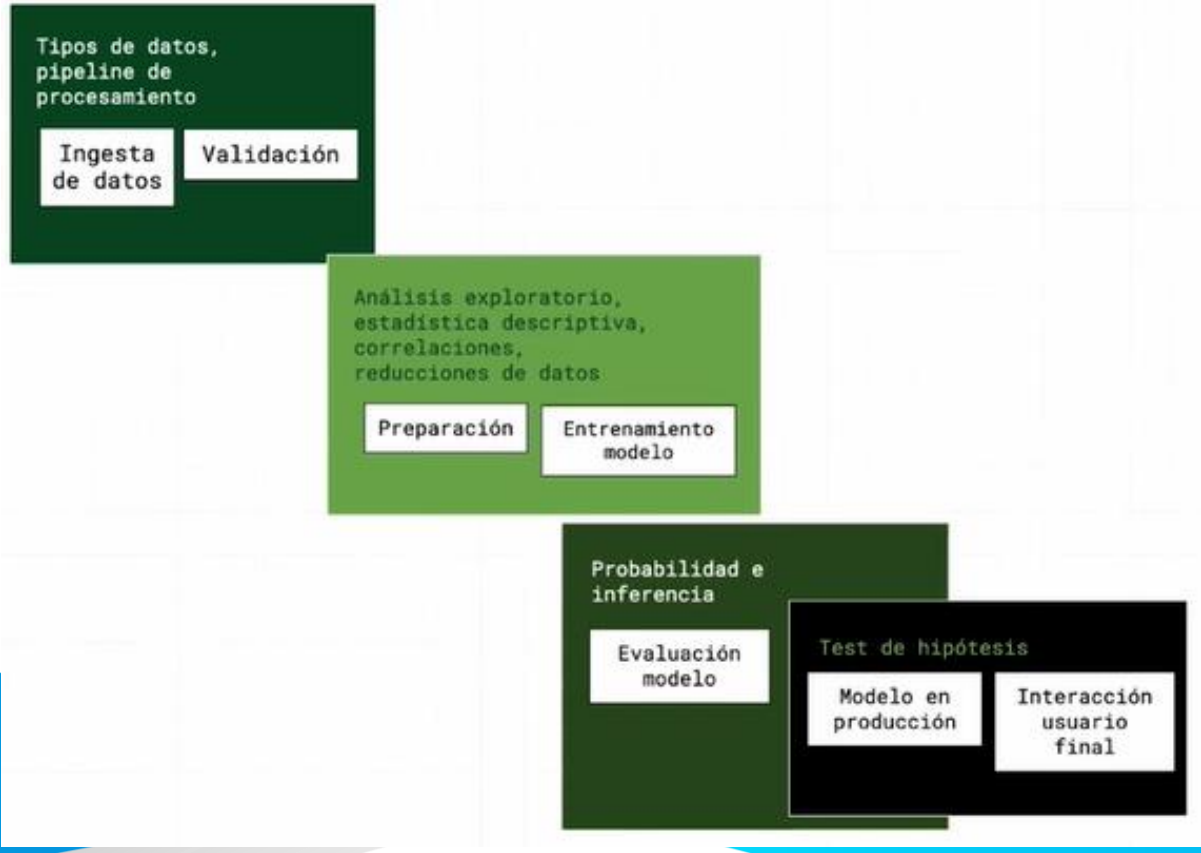
Por qué estudiar estadística:

1. Resumir grandes cantidades de información
2. Tomas mejores decisiones
3. Reconocer patrones en los datos (inferencial – para ver qué va a pasar).

Cómo se usa Estadística en Data Science



Cómo se usa Estadística en Data Science



Medidas de tendencia central

Media: es el valor promedio estándar (lo que siempre conocimos por promedio).

Mediana: es el valor medio exacto en un conjunto de datos ordenados. Es decir, el 50% de los valores son menores que la media y el 50% son mayores.

Moda: el valor con mayor frecuencia en un conjunto de datos.

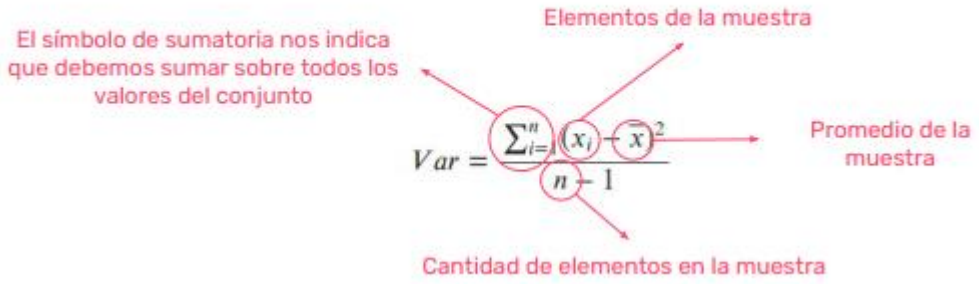
Ejemplo
Muestra: {5, 6, 7, 6, 7, 8, 6, 5, 6}

- **Media = 6.22**
- **Mediana = 6**
  5, 5, 6, 6, 6, 6, 7, 7, 8
- **Moda = 6**
  5, 5, 6, 6, 6, 6, 7, 7, 8

Medidas de dispersión

Varianza y Desviación estándar

Mide la variabilidad o dispersión de un conjunto de números (*muestra*).



- Muestra: {5, 10, 8, 20} ➡
  - N es 4
  - El promedio, $\bar{X}$ es 10,75

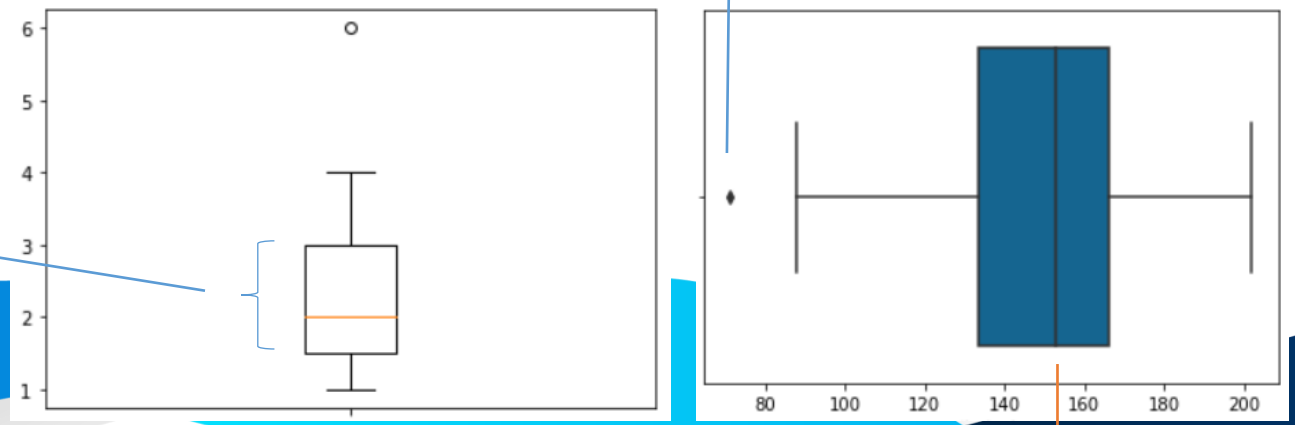
$$Var = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$\Rightarrow \quad Var = \frac{(5-10.75)^2 + (10-10,75)^2 + (8-10,75)^2 + (20-10,75)^2}{4-1}$$

Medidas de dispersión

Rango: diferencia entre valor mínimo y máximo

Rango intercuartil:



Tipos de datos asociado a variables

Categóricos
(género, categoría de película, método de pago)

- ordinal
- nominal

Numéricos
(edad, altura, temperatura)

- discretos
- continuos

Int

Float

[Variables categóricas]

Aquí vemos que los tipos de datos se identifican de la siguiente manera:

- Categoricos: `object`, `bool`
- Numéricos: `int64` (discreto), `float64` (contínuo)

Algunas denominaciones:

- Datos transaccionales
- Datos demográficos
- Datos de comportamiento

Pandas

- Es una librería para manejar conjuntos de dat
- Trae una estructura de datos: los **arrays.**

Ver Documentación Pandas

Ver Cómo instalar librerías

Material complemetario (DS)

- Deepnote

Contenido asincrónico

- Azure – Core solutions and Management tools
- Azure – General security and network security features
- Resto de actividades en plataforma Interactiva Virtual

UNIVERSIDAD EAFIT

¡Gracias!

Aliados:

Microsoft

idata
Advanced analytics for business

Vigilada Mineducación