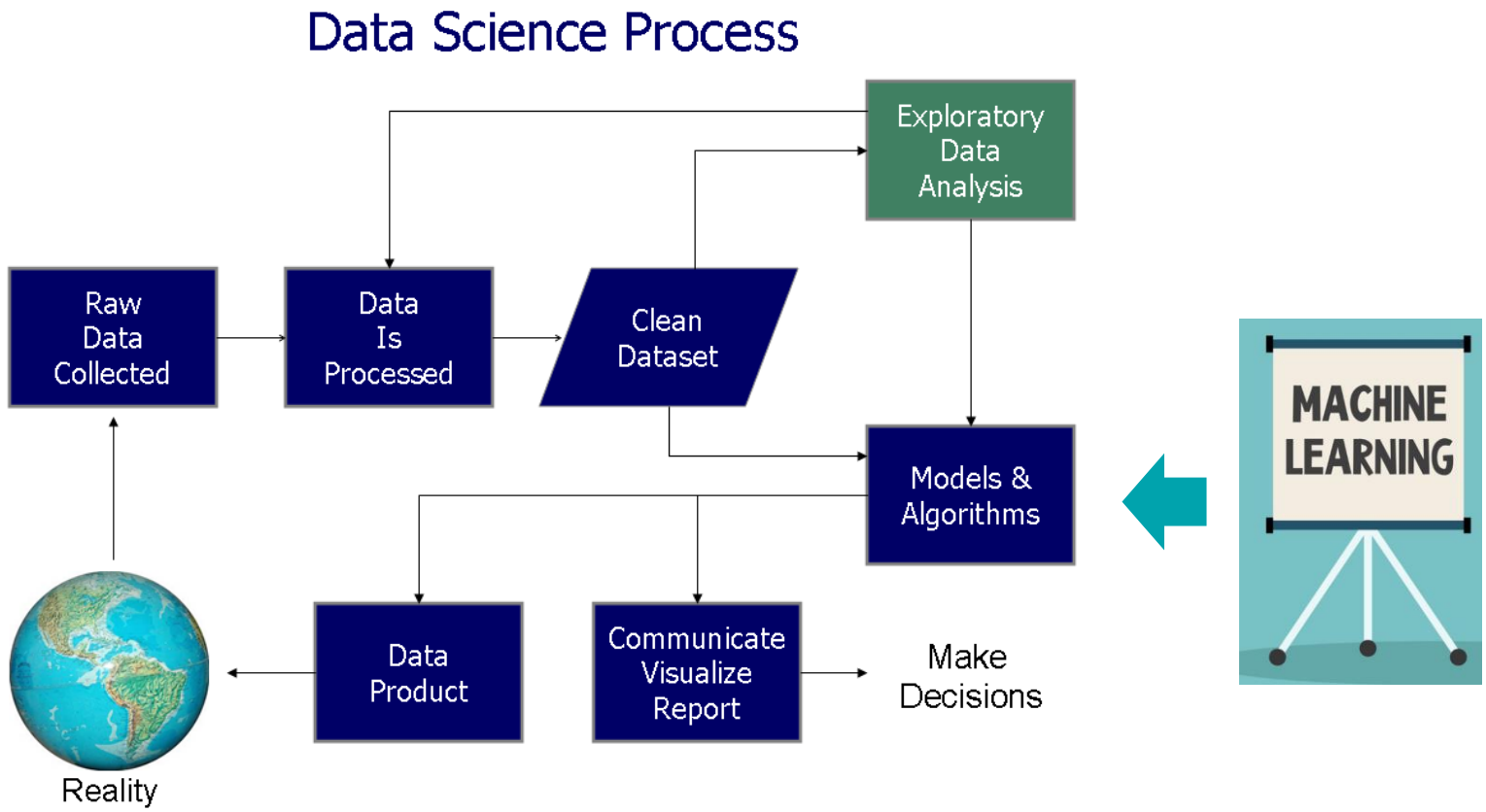
Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Análisis Exploratorio

Fundamentos del Aprendizaje Automático

Proceso de análisis de datos



Análisis exploratorio

Análisis exploratorio

- EDA = Exploratory Data Analysis
- Empezó a coger fuerza en 1977 con un libro de Tukey

EDA

Mondragon Unibertsitatea
Goi Eskola Politeknikoa

- ¿Qué problemas estamos tratando de resolver?
  - Validar o invalidar una hipótesis
- ¿Qué clase de datos tenemos?
- ¿Cómo son los datos?
  - Preproceso
    - Datos faltantes, normalización, reducción de dimensionalidad…
  - Outlayers o datos atípicos
- Análisis de características o atributos
  - ¿Sobran?
  - ¿Podemos deducir nuevos atributos?

- …

EDA: elementos de datos estructurados

- Tipos de da

	Tipo de variable	Descripción	Sinónimos
Datos Númericos	Continua	Datos que pueden tener cualquier valor dentro de un intervalo	Intervalo, float, numérico…
	Discreta	Datos que solo pueden tener valores enteros	Integer, count
Datos categóricos	Categórica	Un dato que suele puede tener un número de valores específicos que representen una categoría. Atributo color: rojo, amarillo…	Enums, factors, nominal…
	Binaria	Caso especial del dato categoríco donde solo existen dos categorías: 0/1, true/false	Boolean, lógica…
	Ordinal	Datos categóricos que tienen un orden especial	Ordered factor

EDA: datos rectangulares

- La estructura de referencia para el análisis en la ciencia de datos es un objeto de datos rectangulares
- Conceptos clave:

Tipo de variable	Descripción	Sinónimos
Data Frame	Dato rectangular (parecido a un spreadsheet). Es la estructura básica para estádistica y modelos de machine learning	
Feature (columna)	Las columnas de una tabla se denominan como feature	Atribute, input, predictor, variable
Outcome (output)	Muchos proyectos de ciencia de datos requieren predecir una respuesta:	Dependent variable, response, target, output
Records (filas)	Cada fila de la tabla se denomina como registro o record	Example, insntance, observation, pattern, sample

EDA: datos rectangulares

import **pandas** as pd

```
In [33]: data
```

Out[33]:

	Area Abbreviation	Area Code	Area	Item Code	Item	Element Code	Element	Unit	latitude	longitude	...	Y2004	Y2005	Y2006	Y2007	Y2008	Y2009
0	AF	2	Afghanistan	2511	Wheat and products	5142	Food	1000 tonnes	33.94	67.71	...	3249.0	3486.0	3704.0	4164.0	4252.0	4538.0
1	AF	2	Afghanistan	2805	Rice (Milled Equivalent)	5142	Food	1000 tonnes	33.94	67.71	...	419.0	445.0	546.0	455.0	490.0	415.0
2	AF	2	Afghanistan	2513	Barley and products	5521	Feed	1000 tonnes	33.94	67.71	...	58.0	236.0	262.0	263.0	230.0	379.0
3	AF	2	Afghanistan	2513	Barley and products	5142	Food	1000 tonnes	33.94	67.71	...	185.0	43.0	44.0	48.0	62.0	55.0
4	AF	2	Afghanistan	2514	Maize and products	5521	Feed	1000 tonnes	33.94	67.71	...	120.0	208.0	233.0	249.0	247.0	195.0
5	AF	2	Afghanistan	2514	Maize and products	5142	Food	1000 tonnes	33.94	67.71	...	231.0	67.0	82.0	67.0	69.0	71.0
6	AF	2	Afghanistan	2517	Millet and products	5142	Food	1000 tonnes	33.94	67.71	...	15.0	21.0	11.0	19.0	21.0	18.0
7	AF	2	Afghanistan	2520	Cereals, Other	5142	Food	1000 tonnes	33.94	67.71	...	2.0	1.0	1.0	0.0	0.0	0.0
8	AF	2	Afghanistan	2531	Potatoes and products	5142	Food	1000 tonnes	33.94	67.71	...	276.0	294.0	294.0	260.0	242.0	250.0
9	AF	2	Afghanistan	2536	Sugar cane	5521	Feed	1000 tonnes	33.94	67.71	...	50.0	29.0	61.0	65.0	54.0	114.0
10	AF	2	Afghanistan	2537	Sugar beet	5521	Feed	1000 tonnes	33.94	67.71	...	0.0	0.0	0.0	0.0	0.0	0.0

EDA: Medidas de posición

- Media:
  - La suma de todos los valores dividida por el número de valores

$$\text{Mean} = \bar{x} = \frac{\Sigma_i^n x_i}{n}$$

  - Números = {3 5 1 2}
  - Media Números = 3 + 5 + 1 + 2 / 11 = 2.75

EDA: Medidas de posición

- Media recortada
  - Una variedad de la media, donde se hace el cálculo de la media eliminando un número determinado de valores en los extremos.
  - Partiendo de un grupo de valores ordenados $x_{(1)}, x_{(2)}, x_{(3)}, \ldots x_{(n)}$
  - Ejemplo:

$$\text{Trimmed mean} = \bar{x} = \frac{\sum_{i=p+1}^{n-p} x_{(i)}}{n-2p}$$

    - Salto (natación):
      - *"For a five-judge panel, the highest and lowest scores are discarded and the middle three are summed and multiplied by the Degree of Difficulty"*

EDA: Medidas de posición

- Media ponderada:
  - La suma de todos los valores multiplicados por un peso dividida por la suma de todos los pesos

$$\text{Weighted mean} = \bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i}^{n} w_i}$$

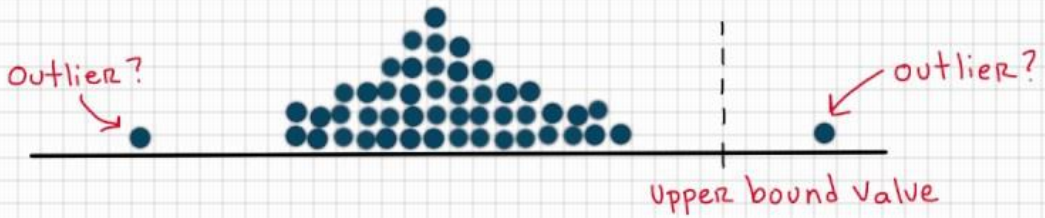
  - Motivación:
    - Algunos valores pueden ser intrínsecamente más variables que otras
      - Menos peso
      - Ejemplo: un sensor de menor precisión
    - Los datos no representan de forma equitativa a todos los grupos
      - Ejemplo: experimento online donde el dataset no refleja de forma precisa los usuarios de la bbdd => podríamos dar mayor peso a los grupos subrepresentados

EDA: Medidas de posición

- Mediana:
  - El valor de la variable de posición central en un conjunto de datos ordenados
  - Muy dependiente de los valores centrales
    - Menos sensitivo a los outliers
  - Ejemplo

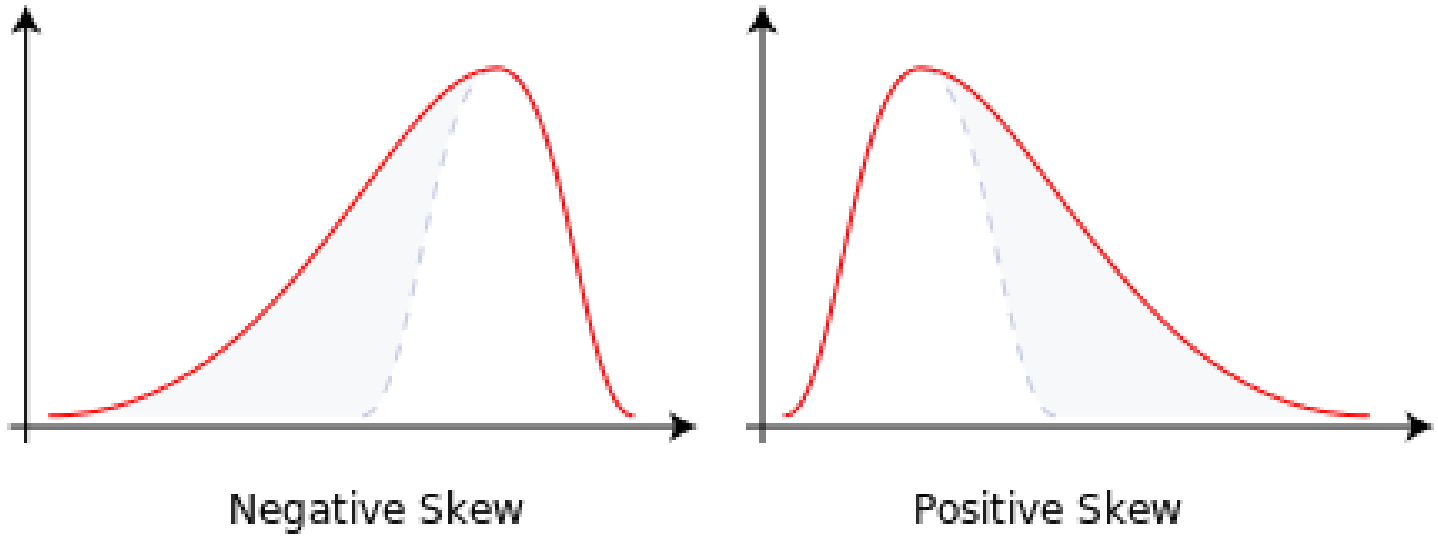
EDA: Medidas de posición

Def: An outlier are data value(s) that lie outside of the overall pattern of the distribution.



EDA: medidas de posición

- Mediana:
  - Muy dependiente de los valores centrales



EDA: Medidas de posición

- Ejemplo:

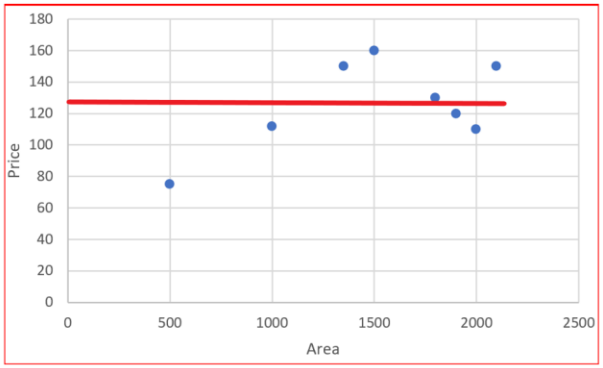
```
1  import pandas as pd
2  from scipy.stats import trim_mean
3  import numpy as np
4
5
6
7  df = pd.read_csv("./../data/report.csv")
8
9  print(df["population"].mean())
10 print(trim_mean(df["population"].values, 0.1))
11 print(df["population"].median())
```

```
795698.0891304348
620192.463576159
536614.5
```

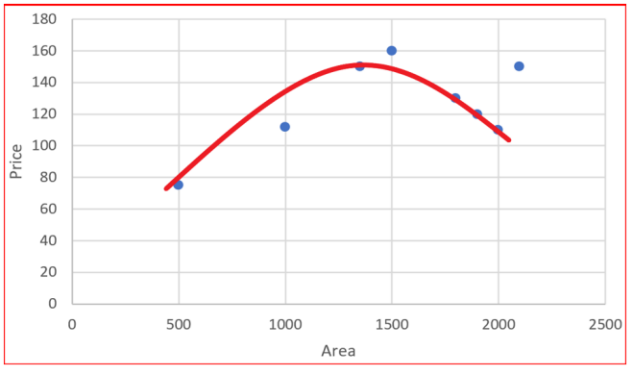
- La media es mayor que la media recortada, ya que en esta última hemos eliminado el 10% de valores de cada extremo.

EDA: Medidas de variabilidad

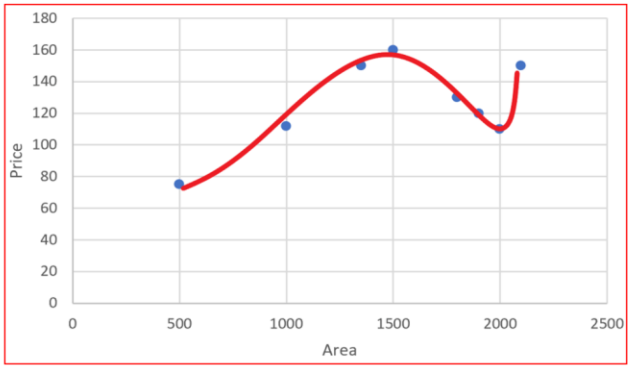
- Variabilidad, esencial en la estadística
  - ¿Cómo se mide?
  - ¿Cómo reducirla?
  - ¿Cómo distinguir la aleatoriedad de la variabilidad?



High Bias - underfit



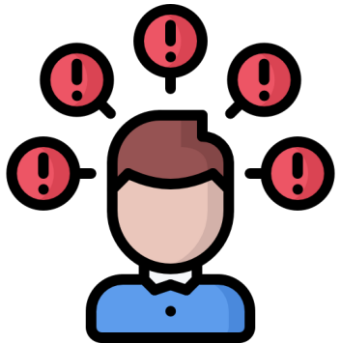
Just Fit



High Variance – overfit

EDA: Medidas de variabilidad

- Desviación estándar y estimadores relacionados
  - Diferencias o desviaciones entre un estimador de posición y el dato observado
  - Ejemplo:
    - {1, 4, 4} => media = 3, mediana = 4
    - Desviación de la media = 1-3, 4-3, 4-3 = -2,1,1
      - Estas desviaciones nos dicen como de dispersos están los datos respecto al valor central
  - Calcular variabilidad:
    - Calcular valor típico para estas desviaciones
      - ¿Media?



EDA: Medidas de variabilidad

- Desviación media absoluta

$$\text{Mean absolution deviation} = \frac{\Sigma_{i=1}^{n} |x_i - \bar{x}|}{n}$$

- Varianza

$$\text{Variance} = s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$

- Desviación estándar

$$\text{Standard deviation} = s = \sqrt{\text{Variance}}$$

  - La desviación estándar es más fácil de interpretar que la varianza
    - Misma escala que los datos originales
  - Medidas no robustas frente a outliers

EDA: Medidas de variabilidad

- MAD (*Median absolute deviation from the median*)
  - Más robusto frente a valores extremos

$$\text{Median absolute deviation} = \text{Median}(|x_1 - m|, |x_2 - m|, ..., |x_N - m|)$$

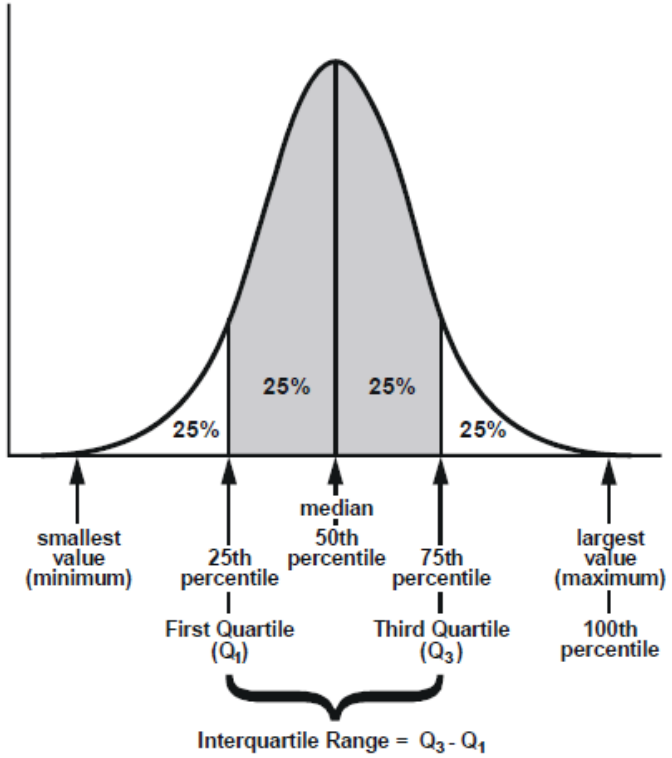
EDA: Medidas de variabilidad

- Estimadores basados en percentiles
- Definiciones:
  - Estadísticos de orden: estadísticas basadas en datos ordenados
  - Rango: diferencia entre el mayor y el menor valor
  - Percentil núm. P:
    - Valor donde al menos el P% de los valores tiene un valor menos y un (100 – P)% de los valores tienen un valor igual o mayor

EDA: Medidas de variabilidad

- Una medida común de variabilidad:
  - Rango intercuartil (IQR): Diferencia entre el percentil núm. 25 (Q1) y el percentil núm. 75 (Q3)

  - {3,1,5,3,6,7,2,9} → Ordenar → {1,2,3,3,5,6,7,9}
    - Percentil 25 => 8 x 25 / 100 o 8 + 1 /4 (No hay uniformidad) => 2.5
    - Percentil 75 => 6.5
    - IQR = 6.5 – 2.5 = 4



EDA: medidas de variabilidad

Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Estimadores de variabilidad de la población

	report_year	agency_code	agency_jurisdiction	population	violent_crimes	homicides	rapes	assaults	robberies	months_reported	crimes_percapita	ho
0	1975	NM00101	Albuquerque, NM	286238.0	2383.0	30.0	181.0	1353.0	819.0	12.0	832.52	
1	1975	TX22001	Arlington, TX	112478.0	278.0	5.0	28.0	132.0	113.0	12.0	247.16	
2	1975	GAAPD00	Atlanta, GA	490584.0	8033.0	185.0	443.0	3518.0	3887.0	12.0	1637.44	
3	1975	CO00101	Aurora, CO	116656.0	611.0	7.0	44.0	389.0	171.0	12.0	523.76	
4	1975	TX22701	Austin, TX	300400.0	1215.0	33.0	190.0	463.0	529.0	12.0	404.46	
...	...	...	...	...	...	...	...	...	...	...	...	
2824	2015	OK07205	Tulsa, OK	401520.0	3628.0	55.0	365.0	2354.0	854.0	NaN	903.57	
2825	2015	VA12800	Virginia Beach, VA	452797.0	626.0	19.0	103.0	234.0	270.0	NaN	138.25	
2826	2015	DCMPD00	Washington, DC	672228.0	8084.0	162.0	494.0	4024.0	3404.0	NaN	1202.57	
2827	2015	KS08703	Wichita, KS	389824.0	3839.0	27.0	349.0	2730.0	733.0	NaN	984.80	
2828	2015	NaN	United States	NaN	1197704.0	15696.0	NaN	NaN	NaN	NaN	372.60	

```python
import pandas as pd

Q1 = df['population'].quantile(0.25)
Q3 = df['population'].quantile(0.75)

stdDev = df['population'].std()
IQR = Q3 - Q1  #también se puede calcular directamente con scipy

d = abs(df['population'] - df['population'].median())
MAD = d.median()
```

```python
print(stdDev)
print(IQR)
print(MAD)
```

```
1012450.5695786542
438924.75
180837.5
```

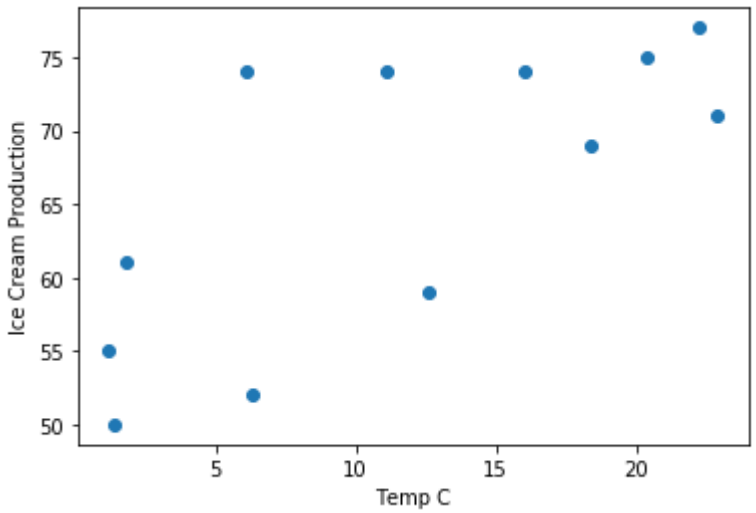
EDA: correlación

- La correlación describe como se relaciona una variable con otra
  - ¿Existen relación entre ellas?
  - ¿Existe causalidad?
    - Hacer predicciones

EDA: correlación

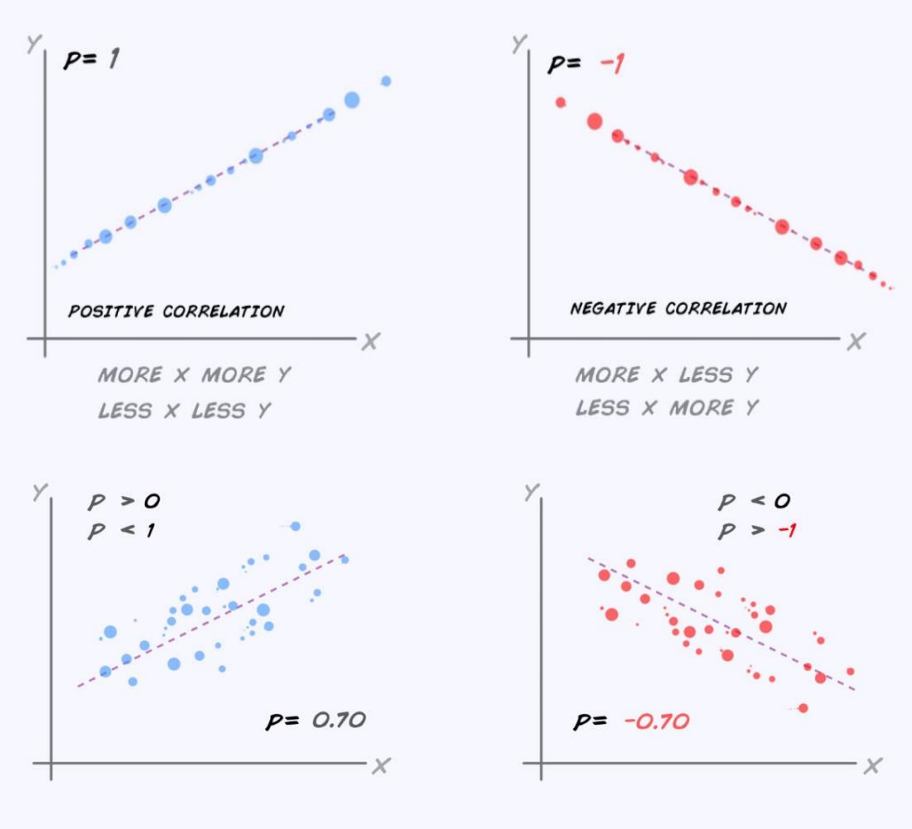
- Temperatura vs producción de helados

Scatter plot



  - PCC = 0.72

- Coeficiente de correlación
  - PCC = Pearson's correlation coefficient
  - Desde -1 a 1



EDA: correlación

- Paradoja de Simpson
  - Pregunta: ¿Qué alumnos son más agradables?

Territorio	Núm. alumnos	Media de núm de amigos
Gipuzkoa	101	8.2
Bizkaia	103	6.4

EDA: correlación

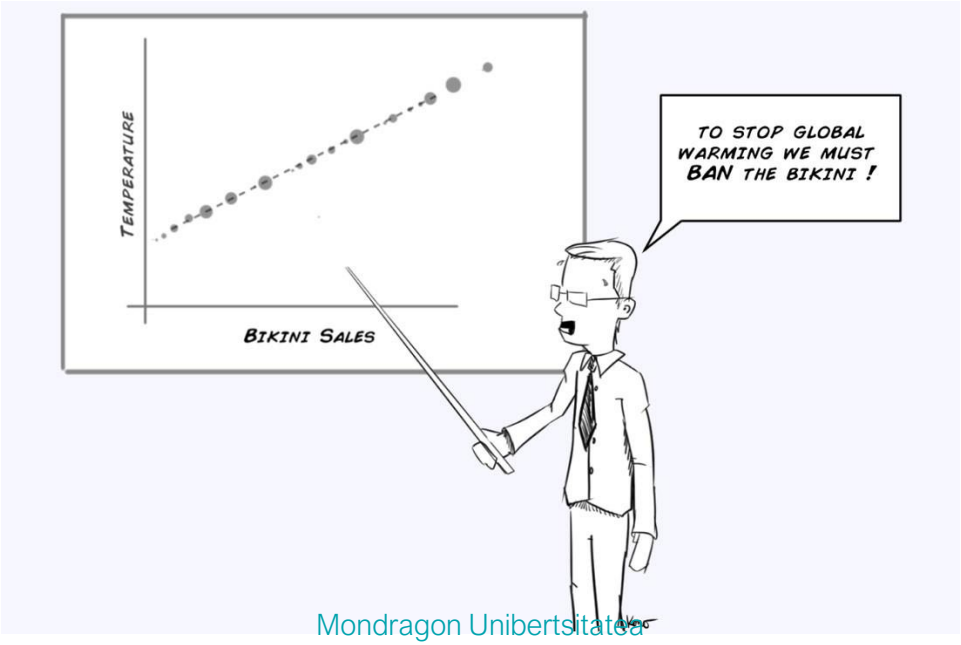
- Paradoja de Simpson
  - Pregunta: ¿Qué alumnos son más agradables?

Territorio	Grado	Núm. alumnos	Media de núm. de amigos
Gipuzkoa	Infor	35	3.1
Bizkaia	Infor	70	3.2
Gipuzkoa	Teleco	66	10.9
Bizkaia	Teleco	33	13.4

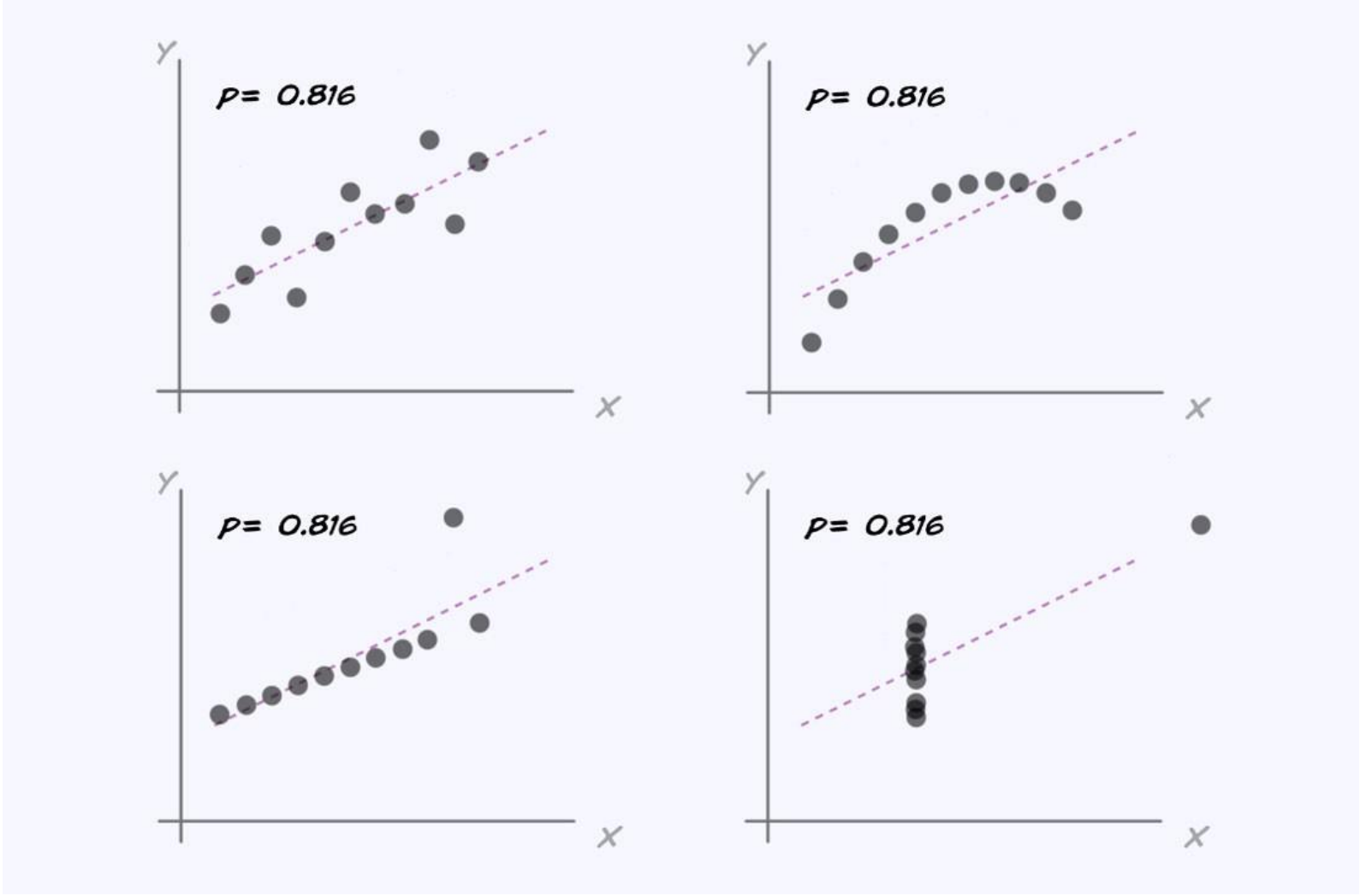
KNOW YOUR DATA!!!

EDA: correlación

- Correlación y causalidad
  - "*Correlation doesn't imply causation*"
  - Ejemplo:
    - Producción de helados – Venta de trajes de baño
  - Otros ejemplos: https://www.datasciencecentral.com/profiles/blogs/spurious-correlations-15-examples



EDA: correlación



Mondragon Unibertsitatea
Goi Eskola Politeknikoa

Aitor Agirre
aaguirre@mondragon.edu