

TICS-411 Minería de Datos

Clase 1: Calidad de los Datos y Feature Engineering

Alfonso Tobar-Arancibia

alfonso.tobar.a@edu.uai.cl

Avisos

Avisos

Ayudantía

Tenemos (posible) ayudante, pero tenemos un problema de horario.

- **Horario Actual:** Viernes 20:00 a 21:10 hrs.
- **Horario Propuesto:** Lunes 11:45 a 12:55 hrs.

Tarea 1

- Entrega el **7 de Abril**: Parejas inscribirse en Webcursos.
- Plazo para inscribir parejas: **Este Domingo**.

Fechas de Prueba

- **Prueba 1:** Martes 30 de Abril 18:30 a 21:00
- **Prueba 2:** Martes 11 de Julio 18:30 a 21:00

Datos Tabulares

Tipos de Datos: Datos Tabulares

	siteEvt	site	time	ml	depth	dist	azmth	lat	lon	peak	evid	epoch	Id
0	2431	1	2011-07-16 07:39:00 (UTC)	2.99	14.8	34.88	168	32.791	-115.453	10019300.0	15018412	1310801940	2431
1	4028	1	2010-04-06 03:09:53 (UTC)	3.05	10.0	88.57	171	32.312	-115.380	610667.0	14616764	1270523393	4028
2	7393	1	2008-01-18 09:52:29 (UTC)	2.70	19.0	47.68	164	32.686	-115.389	1863890.0	14343464	1200649949	7393
3	12447	1	2005-09-01 01:38:23 (UTC)	2.01	4.6	11.20	315	33.169	-115.615	13613300.0	14178480	1125538703	12447
4	7284	1	2009-12-30 18:48:57 (UTC)	5.80	6.0	77.43	156	32.464	-115.189	736620000.0	14565620	1262198937	7284



- Filas: Observaciones, registros, instancias. (Normalmente independientes).
- Columnas: Variables, Atributos, Features.



- Probablemente el tipo de datos más amigable.
- Requiere conocimiento de negocio (**Domain Knowledge**)



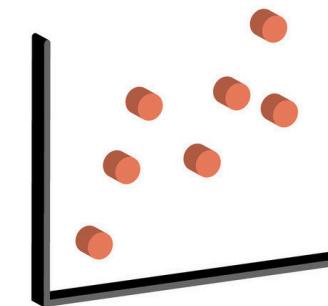
- Es un % bajísimo del total de datos existentes en el Mundo.
- Distintos tipos, por lo que normalmente requiere de algún tipo de **preprocesamiento**.

Data Types: Numéricos

Numéricos

Valores a los que se les puede aplicar alguna operación matemática.

- **Discretas:** Número finito o contable de valores. Integers (Enteros). Ej: [Número de Hijos](#), [Cantidad de Productos](#), [Edad](#).
- **Continuas:** Existen infinitos puntos entre dos puntos. Floats (punto flotando o decimales). Ej. [Temperatura](#), [Peso](#).



Discrete
Data

VS



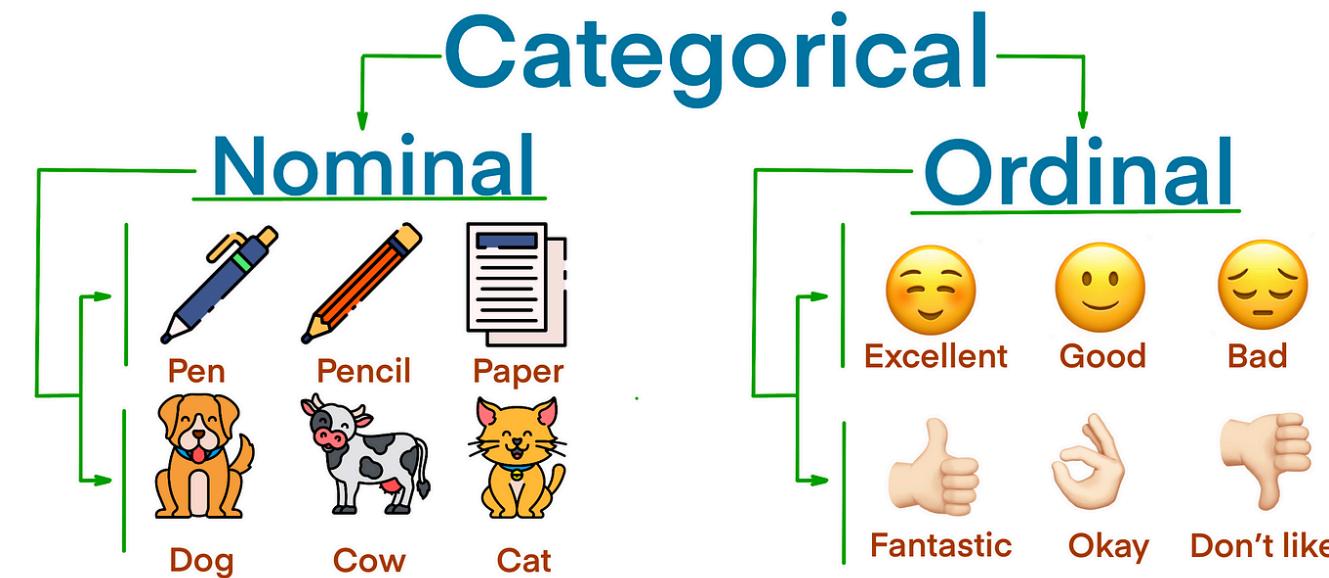
Continuous
Data

Data Types: Categóricos

Categóricos

Datos que representan una categoría.

- **Nominales:** Sólo nombres que no representan ningún orden. Ej: **Nacionalidad, género, ocupación.**
- **Ordinales:** Que tienen un orden o jerarquía inherente. Ej: **Nivel de Escolaridad, tamaño.**



No todas las operaciones matemáticas son aplicables. Ej: Media, Mediana, Sumas, Restas, etc.

Data Types: Otros

Strings

Datos de texto, los cuales podrían eventualmente ser tratados y representar algo. Ej: **Rescatar comunas de una dirección, rescatar sexo desde el nombre, etc.**

Fechas

Datos tipo fecha, los cuales podrían eventualmente ser tratados y representar variables de algún tipo. Ej: **Rescatar Años, meses, días, semanas, trimestres (quarters), etc.**

Datos Geográficos

Datos que representan la ubicación geográfica de un elemento. Ej: **Latitud, Longitud, Coordenadas.**



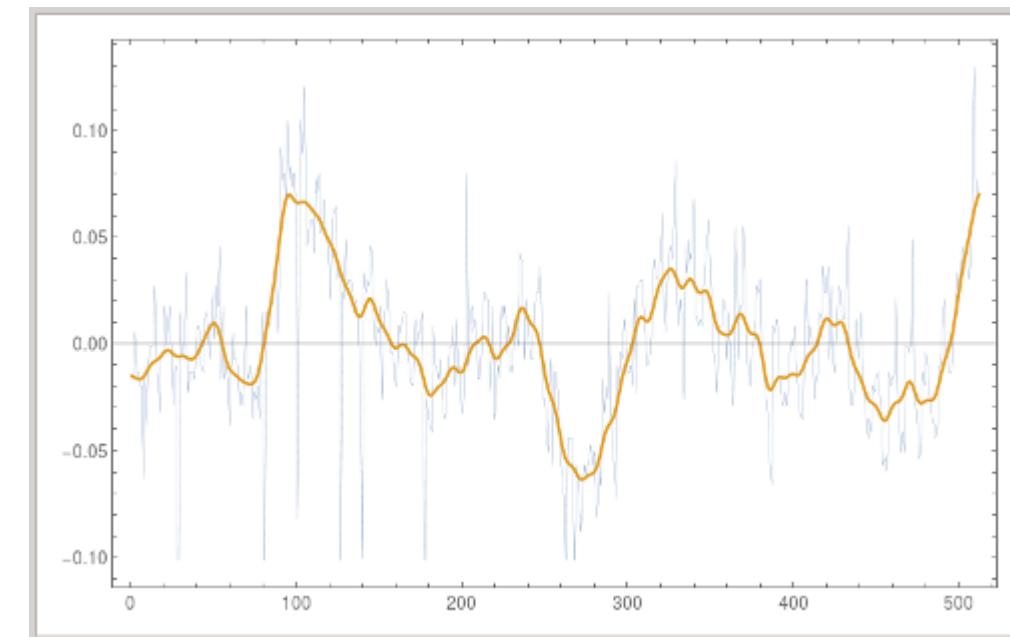
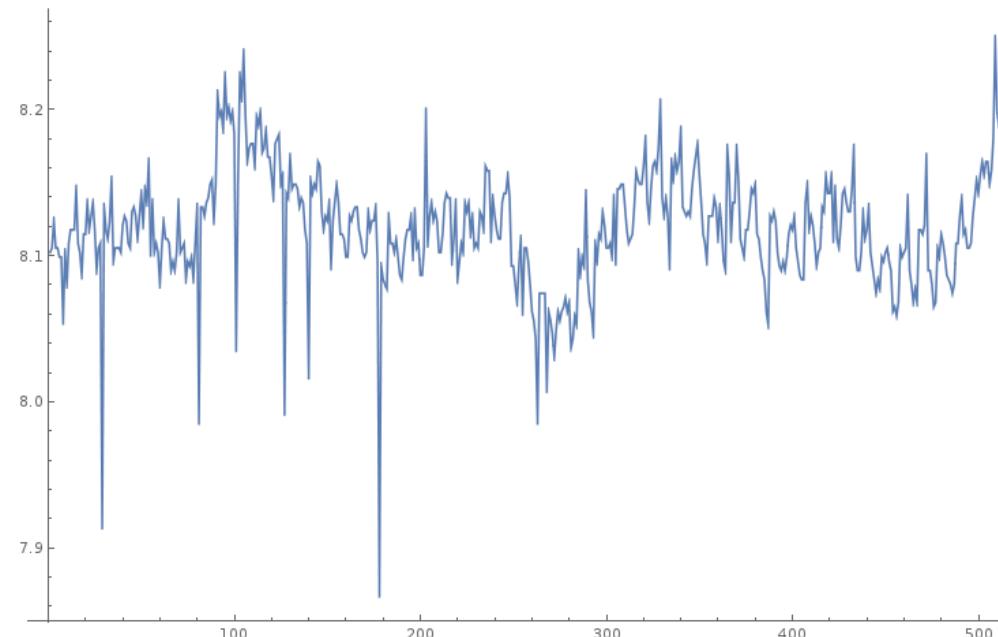
Sin importar el tipo de dato el mayor problema es su **calidad**.

Calidad de los Datos

Calidad de los Datos: Ruido

Ruido

Corresponde al error y extrema variabilidad en la medición en los datos. Este error puede ser aleatorio o sistemático.

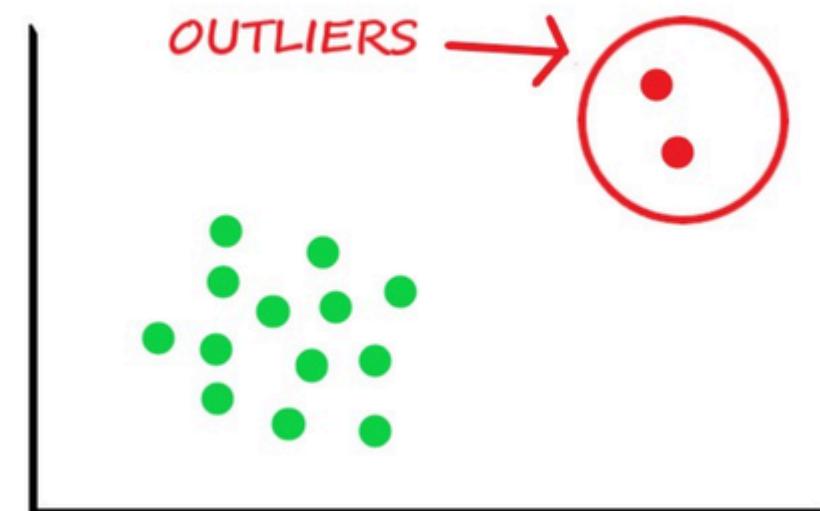


Se le llama **Señal** a la tendencia principal y representa la información significativa y valiosa de los datos.

Calidad de los Datos: Outliers

Outliers

Son datos considerablemente diferentes a la mayoría del dataset. Dependiendo del caso pueden indicar casos "interesantes" o errores de medición.



- Es importante notar que dependiendo del caso puede ser una buena idea deshacerse de ellos. ¿En qué casos podría no ser necesario eliminarlos?

Calidad de los Datos: Valores Faltantes

Missing Values

Son valores que por alguna razón no están presentes.

- **Missing at Random (MAR)**: Son valores que no están presentes por causas que no se pueden controlar. Ej: No se registró, no se preguntó, fallas en el sistema de recolección de datos, etc.
- **Informative Missing**: Es un valor no aplicable. Ej: Sueldo en niños, Precio de la entrada de un concierto si es que **NO** compró entrada.

Missing value

:	loan_amnt	term	int_rate	sub_grade	emp_length	home_ownership	annual_inc	loan_status	addr_state	dti	mths_since_recent_inq	revol_util	bc_open_to_buy	bc_util	num_op_rev_tl
0	3600	36 months	14	C4	10+ years	MORTGAGE	55000	Fully Paid	PA	6	4	30	1506	37	4
1	24700	36 months	12	C1	10+ years	MORTGAGE	65000	Fully Paid	SD	0	19	57830	27	20	
2	20000	60 months	11	B4	10+ years	MORTGAGE	63000	Fully Paid	IL	10	56	2737	56	4	
3	35000	60 months	15	C5	10+ years	MORTGAGE	60000	Current	NJ	12	54962	12	10		
4	10400	36 months	22	F1	3 years	MORTGAGE	104483	Fully Paid	PA	1	64	4567	78	7	
5	10400	36 months	13	C3	4 years	RENT	34000	Fully Paid	GA	10	68	844	91	4	
6	20000	36 months	9	B2	10+ years	MORTGAGE	50000	Fully Paid	MN	15	84	103	9		
7	20000	36 months	8	B1	10+ years	MORTGAGE	85000	Fully Paid	SC	18	6	13674	6	3	
8	10400	36 months	6	A2	6 years	RENT	85000	Fully Paid	PA	13	34	50	13		
9	10400	36 months	11	B5	10+ years	MORTGAGE	42000	Fully Paid	RI	35	10	39	9966	41	5

Calidad de los Datos: Datos Duplicados

Duplicates

Se refiere a registros que pueden estar total o parcialmente duplicados.

S.N°	First Name	Last Name	Title	Company
1	Mary	Sue	Senior Marketing Manager	ABC Ltd.
2	Janet	Martin	Marketing Executive	ABC Ltd.
3	Bryan	Oscar	SEO Manager	ABC Ltd.
4	Jude	Taylor	Marketing Manager	ABC Ltd.
5	Mary S	Sue	Senior Marketing Manager	ABC Ltd.



Esto genera problemas en la confiabilidad de los datos. ¿Cuál es el registro correcto?

Ej: Caso particular de una Jooycar (una startup de seguros).

Calidad de los Datos: Dominio del Problema



- Por lejos el problema de calidad más difícil de encontrar.
- Se requiere experiencia y conocimiento profundo del negocio para detectarlo.

Ej: Caso de Super Avances en Cencosud.

Feature Engineering

Feature Engineering

Feature Engineering

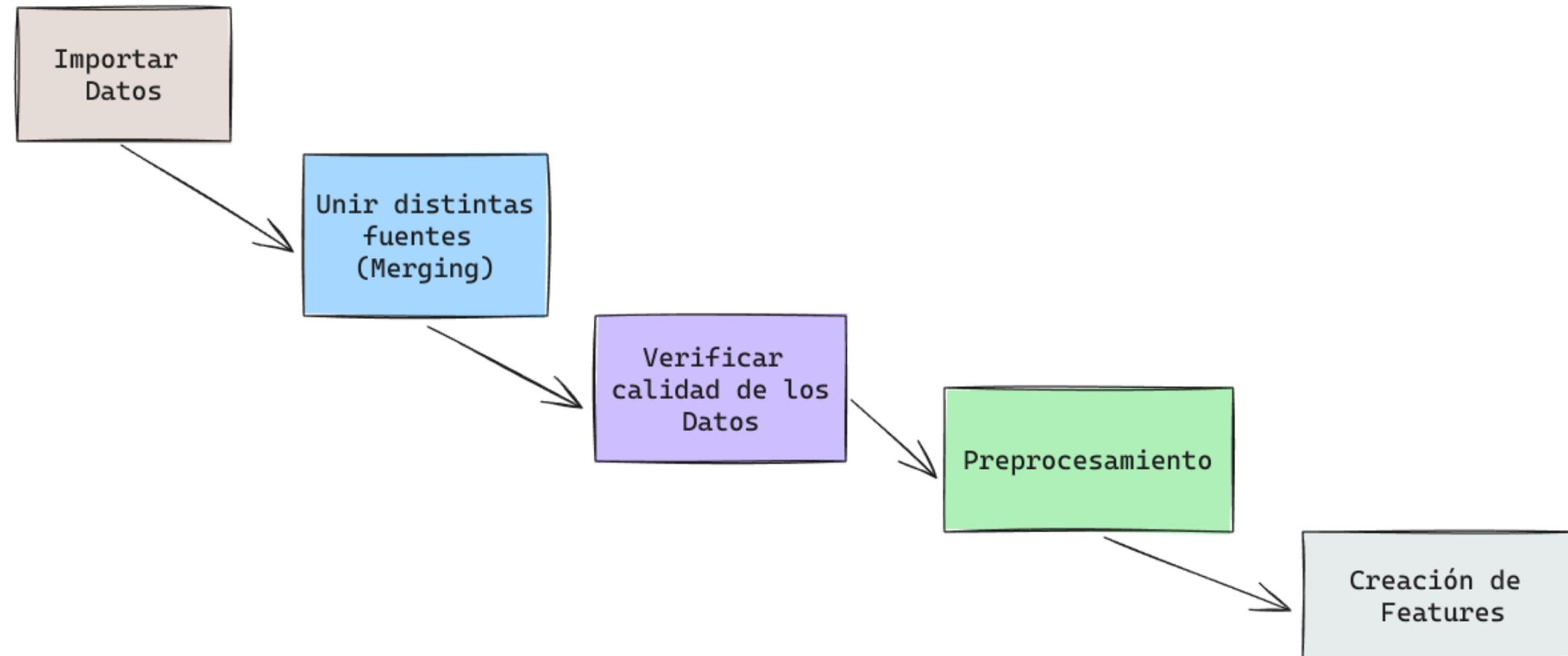
También conocida como Ingeniería de Atributos, es el **arte** de trabajar las **features** existentes para *limpiar* o *corregir* variables existentes o *crear* nuevas variables.

Preprocesamiento

Se refiere al proceso de preparación de los datos para su ingreso a un modelo. En una primera parte puede incluir limpieza de datos corruptos, redundantes y/o irrelevantes. Por otra parte, también hace referencia a la transformación de datos para que puedan ser consumidos por un algoritmo.

Feature Engineering

- No existe un procedimiento estándar.
- Revisar los datos y ver potenciales errores que puedan afectar el funcionamiento de un modelo.



Preprocesamiento: Valores Faltantes

Imputación: Se refiere al proceso de llenar datos faltantes.

Numerical Variables

- Mean/ Median Imputation
- Arbitrary Value Imputation
- End of tail Imputation
- Mode Imputation

Categorical Variable

- Frequent category Imputation
- Adding a "Missing" category

Both

- Complete Case Analysis
- Adding a "Missing" Indicator
- Random Sample Imputation

Mean (Download Speed) = 130

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	130	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	130	95%
8	Lite	76	77%
9	Fast+	180	95%

Median (Download Speed) = 155

Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	N/A	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	N/A	95%
8	Lite	76	77%
9	Fast+	180	95%



Mobile ID	Mobile Package	Download Speed	Data Limit Usage
1	Fast+	157	80%
2	Lite	99	70%
3	Fast+	167	10%
4	Fast+	155	80%
5	Lite	76	70%
6	Fast+	155	10%
7	Fast+	155	95%
8	Lite	76	77%
9	Fast+	180	95%



Dependiendo del nivel de valores faltantes, es necesario evaluar la eliminación de registros o atributos completos de ser necesario.

Preprocesamiento: Manejo de Outliers

Capping

Se refiere al proceso de acotar un atributo eliminando los valores extremos o atípicos (outliers).

	genotype	yield	yield_cleaned
Group B	B	21.77	21.77
	B	26.30	22.85
	B	22.85	22.85
	B	22.31	22.31
Group F	F	20.09	20.09
	F	15.01	19.52
	F	20.38	20.38
	F	19.52	19.52



Al igual que en el caso anterior, es necesario evaluar la eliminación de registros si es que representan valores atípicos.

Preprocesamiento: Manejo de Variables Categóricas

La mayoría de los modelos no tienen la capacidad de poder lidiar con variables categóricas por lo que deben ser transformadas en una representación numérica antes de ingresar a un modelo.

Color
Red
Red
Yellow
Green
Yellow



	Red	Yellow	Green
Red	1	0	0
Red	1	0	0
Yellow	0	1	0
Green	0	0	1
Yellow	0	1	0

One Hot Encoder

color
red
green
blue
red



color
0
1
2
0

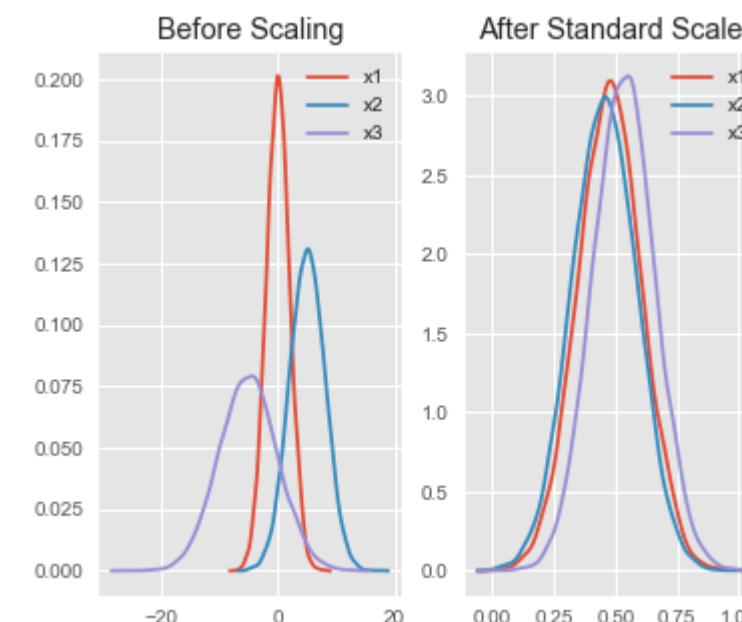
Ordinal Encoder

- **One Hot Encoder** suele dar mejores resultados en modelos lineales modelos que dependan de distancias.
- **Ordinal Encoder** suele dar mejores resultados en modelos de árbol.

¿Son necesarias todas las columnas en un One Hot Encoder?

Preprocesamiento: Escalamiento

El **escalamiento** se refiere al proceso de llevar distintas variables a una misma escala.



StandardScaler (Normalización)

$$x_j = \frac{x_j - \mu_x}{\sigma_x}$$

- Este proceso fuerza (en la medida de lo posible) a tener media 0 y std 1.
- Notar que σ_x hace referencia a la varianza poblacional.

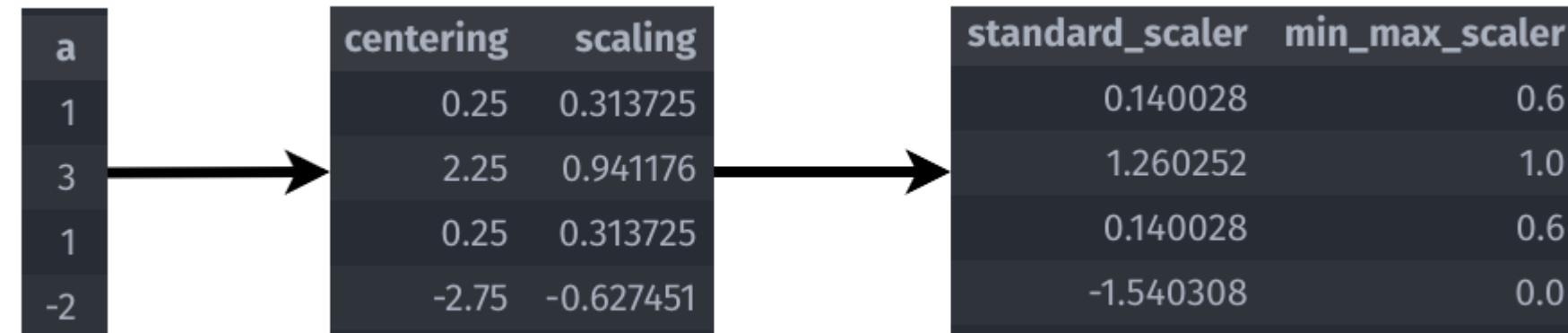
MinMax Scaler

$$x_j = \frac{x_j - \min(x_j)}{\max(x_j) - \min(x_j)}$$

- Evitar que la escala de una “sobre-importancia” a una cierta variable.
- Permitir una mejor convergencia de los algoritmos.

- Este proceso fuerza a los datos a distribuirse entre 0 y 1.

Preprocesamiento: Escalamiento



- i**
- Media: 0.75
 - Std: 3.1875
 - Min: -2
 - Max: 3

- !**
- **Centering (Centrado):** Se le llama a la diferencia entre la variable y su media.
 - **Scaling (Escalado):** Se le llama al cuociente entre la variable y su Desviación Estándar.
 - **StandardScaler (Normalización):** Es Centrado y Escalado.

Creación de Variables

Combinación

Combinar 2 o más variables. Ej: Calcular el área de un sitio a partir del ancho y largo.

Transformación

Aplicar una operación a una variable. Ej: El logaritmo de las ganancias.

Discretización (Binning)

Generar categorías a partir de una variable continua.

AGE	AGE_bins
10	[10,21]
15	[10,21]
16	[10,21]
18	[10,21]
20	[10,21]
30	[22,33]
35	[34,45]
42	[34,45]
48	[46,55]
50	[46,55]
52	[46,55]
55	[46,55]

Creación de Variables

Ratios

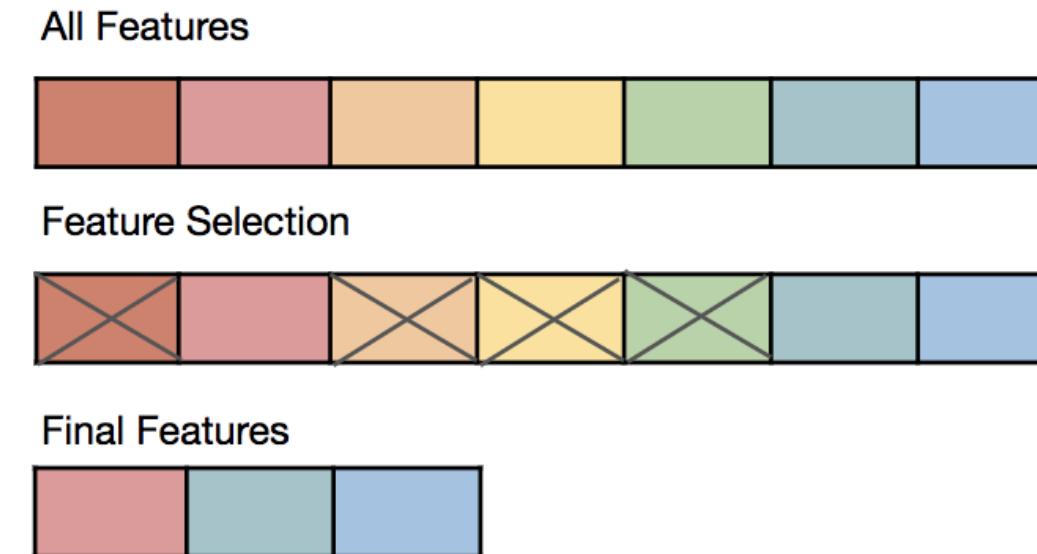
Es una medida que expresa la relación entre dos cantidades. Ej: Puntos por partido, cantidad de transacciones por mes, etc.

Agregación

Agregar o agrupar información resumida de ciertas variables. Ej: Promedio de tiempo en aprobar un tipo de crédito.

Selección de Variables

Se refiere al proceso de eliminar variables que pueden ser irrelevantes o poco significativas.

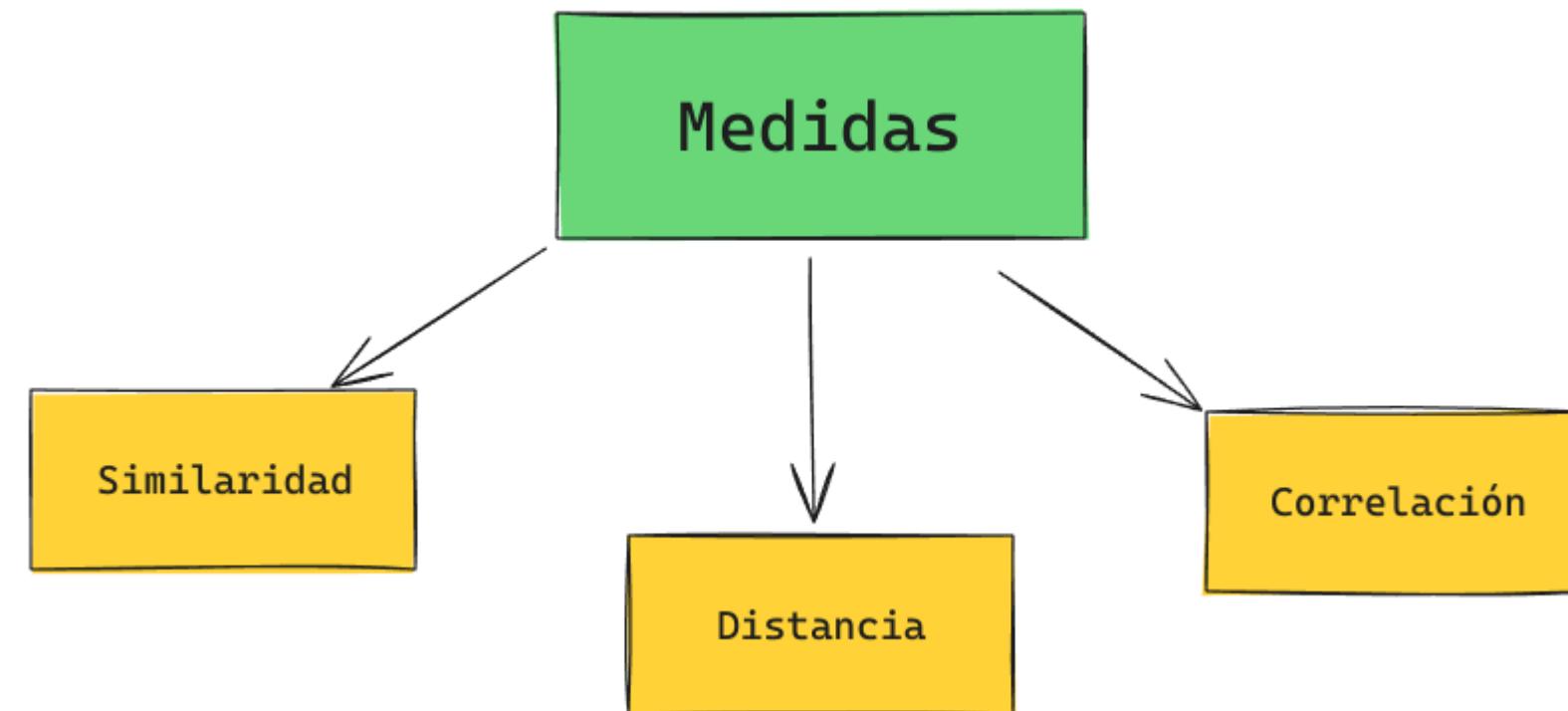


- Procesos Manuales.
- Procesos Automáticos:
 - PCA (Principal Component Analysis).
 - Recursive Feature Elimination.
 - Recursive Feature Addition.
 - Eliminación mediante alguna medida.

Medidas

Medidas

Son métricas que permiten cuantificar la relación existente entre dos o más objetos.



Medidas: Similaridad

Similaridad

- Medida numérica que indica cuán iguales son los datos entre de dos objetos
- Es mayor cuando los objetos con semejantes
- Normalmente se encuentran en el rango [0,1]

Disimilaridad

- Medida numérica que indica cuan diferente son los datos entre de dos objetos
- Es menor cuando los datos son parecidos
- El valor mínimo es normalmente cero
- El límite superior puede variar



Medidas: Similaridad Nominal

- Disimilaridad:

$$D = \begin{cases} 0, & \text{if } p = q \\ 1, & \text{if } p \neq q \end{cases}$$

- Similaridad:

$$S = \begin{cases} 1, & \text{if } p = q \\ 0, & \text{if } p \neq q \end{cases}$$



$$S(p, q) = 0$$

$$D(p, q) = 1$$

Medidas: Similaridad Ordinal

- Disimilaridad:

$$D = \frac{|p - q|}{n}$$



- Similaridad:

$$S = 1 - \frac{|p - q|}{n}$$

$$S(p, q) = 1 - \frac{5 - 4}{5} = 0.8$$

Medidas: Similaridad Intervalo o Ratio

- Disimilaridad:

$$D = |p - q|$$

- Similaridad:

$$S = -D$$

$$S = \frac{1}{1 + D}$$

Sea $p = 35^\circ C$ y $q = 40^\circ C$. Luego:

$$S(p, q) = -5$$

$$S(p, q) = \frac{1}{1 + 5} = 0.17$$

Medidas: Similaridad Datos Categóricos

Sea p y q vectores de dimensión m con sólo *atributos categóricos*. Para calcular la similaridad entre vectores se usa lo siguiente:

$$Sim(p, q) = \sum_{i=1}^m S(p_i, q_i)$$

- Overlap:
- Frecuencia de Ocurrencia Inversa
- Medida de Goodall

$$S(p_{a_i}, q_{a_i}) = \begin{cases} 1, & \text{if } p_{a_i} = q_{a_i} \\ 0, & \text{if } p_i \neq q_i \end{cases}$$

$$S(p_i, q_i) = \frac{1}{p_k(p_i)^2}$$

$$S(p_i, q_i) = 1 - p_k(p_i)^2$$

- ! • $p_k()$ se refiere a la probabilidad de ocurrencia del atributo k.
 • Todas estas medidas son 0 si $p_i \neq q_i$

Medidas: Similaridad Datos Categóricos

Datos		
item	shape	color
1	circle	red
2	square	red
3	triangle	red
4	circle	red
5	square	yellow
6	triangle	yellow
7	circle	blue
8	square	blue

Atributos	
p1	shape
p2	color
circle	3/8
square	3/8
triangle	2/8
red	4/8
yellow	2/8
blue	2/8

Ejercicio Propuesto: ¿Cuánto vale la similaridad entre los siguientes registros?

- 1-4
- 2-5
- 7-8

Medidas: Similaridad Datos Binarios

Sea p y q vectores de dimensión m con sólo *atributos binarios*. Para calcular la similaridad entre vectores se usa lo siguiente:

$$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}}$$

$$JC = \frac{M_{11}}{M_{01} + M_{10} + M_{11}}$$

- Simple Matching Coefficient = Número de Coincidencias / Total de Atributos
- Jaccard Coefficient = Número de Coincidencias 11 / Número de Atributos distintos de Ceros.

M_{01} = El número de atributos donde p era 0 y q era 1

M_{10} = El número de atributos donde p era 1 y q era 0

M_{00} = El número de atributos donde p era 0 y q era 0

M_{11} = El número de atributos donde p era 1 y q era 1

Medidas: Similaridad Datos Binarios

name	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
p_i	1	0	0	0	0	0	0	0	0	0
q_i	0	0	0	0	0	0	1	0	0	1

$$SMC = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}} = \frac{7 + 0}{7 + 2 + 1 + 0} = 0.7$$

$$JC = \frac{M_{11}}{M_{01} + M_{10} + M_{11}} = \frac{0}{2 + 1 + 0} = 0$$

Medidas: Similaridad (Distancia Coseno)

Sean d_1 y d_2 dos vectores. La distancia coseno se calcula como:

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \|d_2\|}$$

name	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
d_1	3	2	0	5	0	0	0	2	0	0
d_2	1	0	0	0	0	0	1	1	0	2
d_3	6	4	0	10	0	0	0	4	0	0

Ejercicio Propuesto: ¿Cuánto vale $\cos(d_1, d_2)$ y $\cos(d_1, d_3)$?

Distancias

Distancias

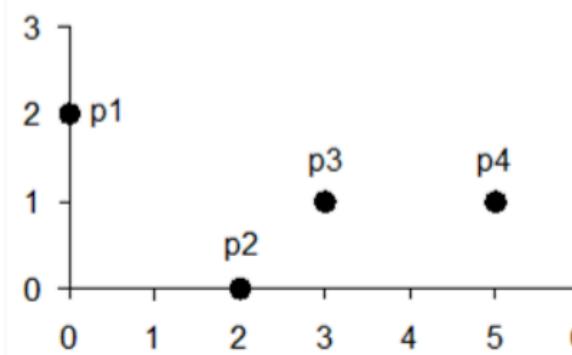
Una métrica o función de distancia es una función que define una distancia para cada par de elementos de un conjunto. Sean dos puntos x e y , una métrica o función de distancia debe satisfacer las siguientes condiciones:

- No Negatividad:
 - $d(x, y) \geq 0$
- Identidad:
 - $d(x, y) = 0 \Leftrightarrow x = y$
- Simetría:
 - $d(x, y) = d(y, x)$
- Desigualdad Triangular:
 - $d(x, z) \leq d(x, y) + d(y, z)$

Distancias: Distancia Minkowski

$$d(p, q) = \left(\sum_{k=1}^m |p_k - q_k|^r \right)^{1/r}$$

Datos		
point	x	y
p1	0	2
p2	2	0
p3	3	1
p4	5	1



- $r = 1 \rightarrow$ Distancia Manhattan (L1).
- $r = 2 \rightarrow$ Distancia Euclideana (L2).
- $r = \infty \rightarrow$ Distancia Chebyshev ($L\infty$).

$$D_{ch}(p, q) = \max_k |p_k - q_k|$$

Resolvamos en Colab

- ! • Se denomina **Matriz de Distancias** a la Matriz que contiene la distancia $d(p_i, p_j)$ en la coordenada i, j .

Distancias: Distancia Minkowski (Resultados)

L1	p1	p2	p3	p4
p1	0	4	4	6
p2	4	0	2	4
p3	4	2	0	2
p4	6	4	2	0

L2	p1	p2	p3	p4
p1	0	2.828	3.162	5.099
p2	2.828	0	1.414	3.162
p3	3.162	1.414	0	2
p4	5.099	3.162	2	0

L∞	p1	p2	p3	p4
p1	0	2	3	5
p2	2	0	1	3
p3	3	1	0	2
p4	5	3	2	0

Distancias: Distancia Mahalanobis

$$d(p, q) = \sqrt{(p - q)^T \Sigma^{-1} (p - q)}$$

donde Σ es la **Matriz de Covarianza** de los datos de entrada.

$$cov(x, y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

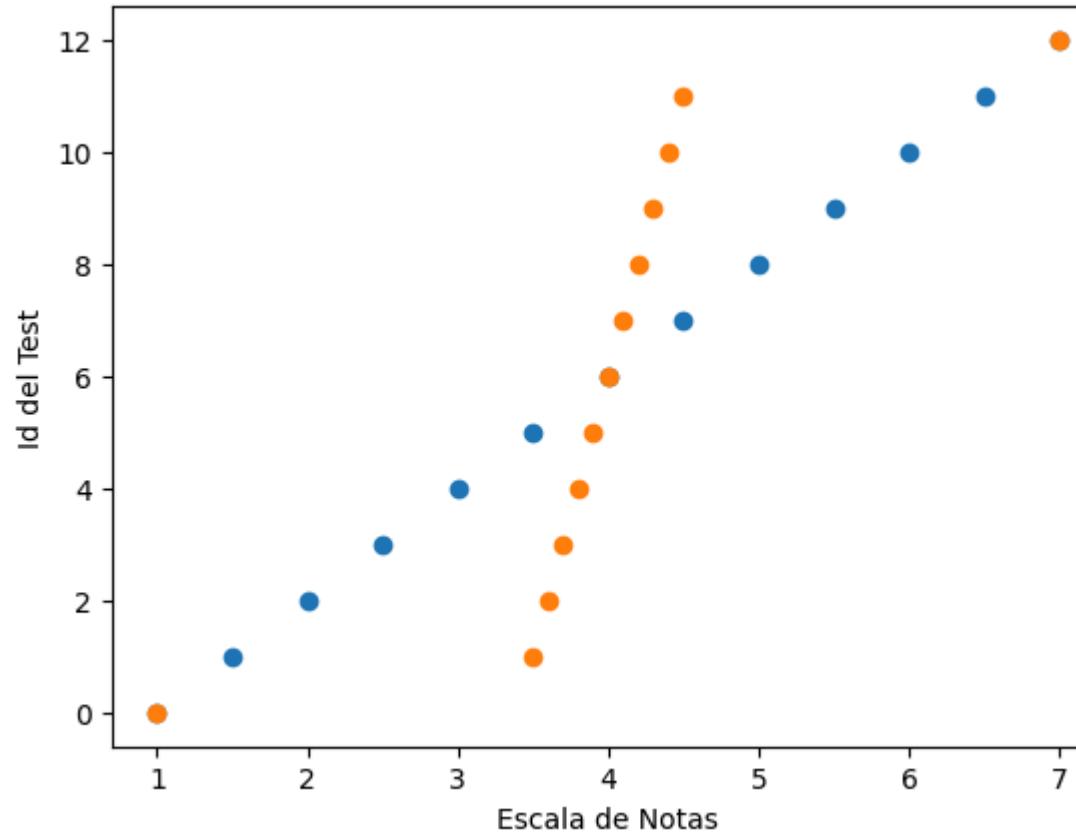
- Para 2 variables p y q:

$$\Sigma = \begin{bmatrix} cov(p, p) & cov(p, q) \\ cov(q, p) & cov(q, q) \end{bmatrix}$$

Ejercicio: Supongamos las siguientes escalas de notas. Calcular la distancia entre la nota (1.0 y 7.0)

- test #1: 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0, 4.5, 5.0, 5.5, 6.0, 6.5, 7.0
- test #2: 1.0, 3.5, 3.6, 3.7, 3.8, 3.9, 4.0, 4.1, 4.2, 4.3, 4.4, 4.5, 7.0

Distancias: Distancia Mahalanobis (Resultados)



- test #1:

$$d(7.0, 1.0) = \sqrt{(7 - 1) \frac{1}{3.79} (7 - 1)} = 3.08$$

- test #2:

$$d(7.0, 1.0) = \sqrt{(7 - 1) \frac{1}{1.59} (7 - 1)} = 4.76$$



- Es importante notar que la correlación existente entre los datos **influye en la distancia**.

Correlación

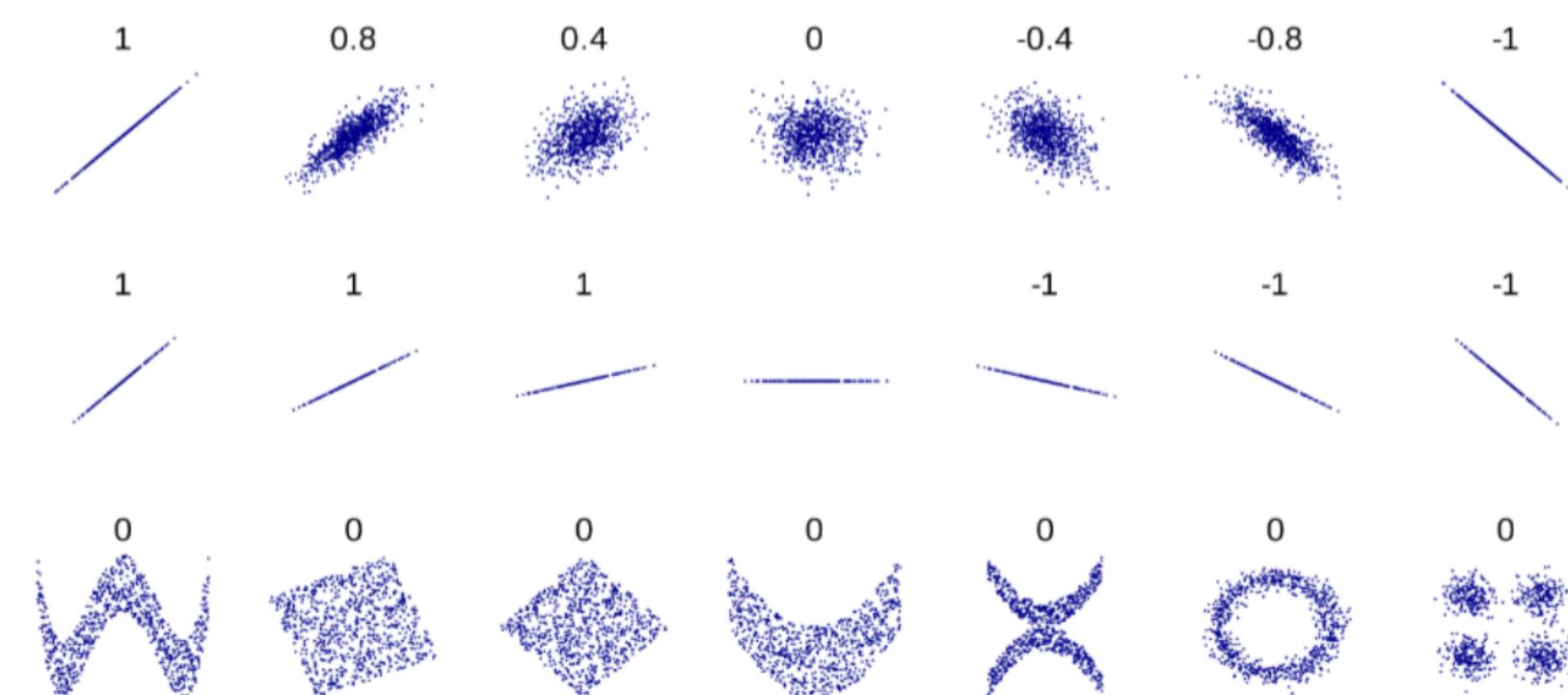
La correlación mide la relación lineal entre 2 atributos.

Correlación Poblacional

$$\rho(X, Y) = \text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

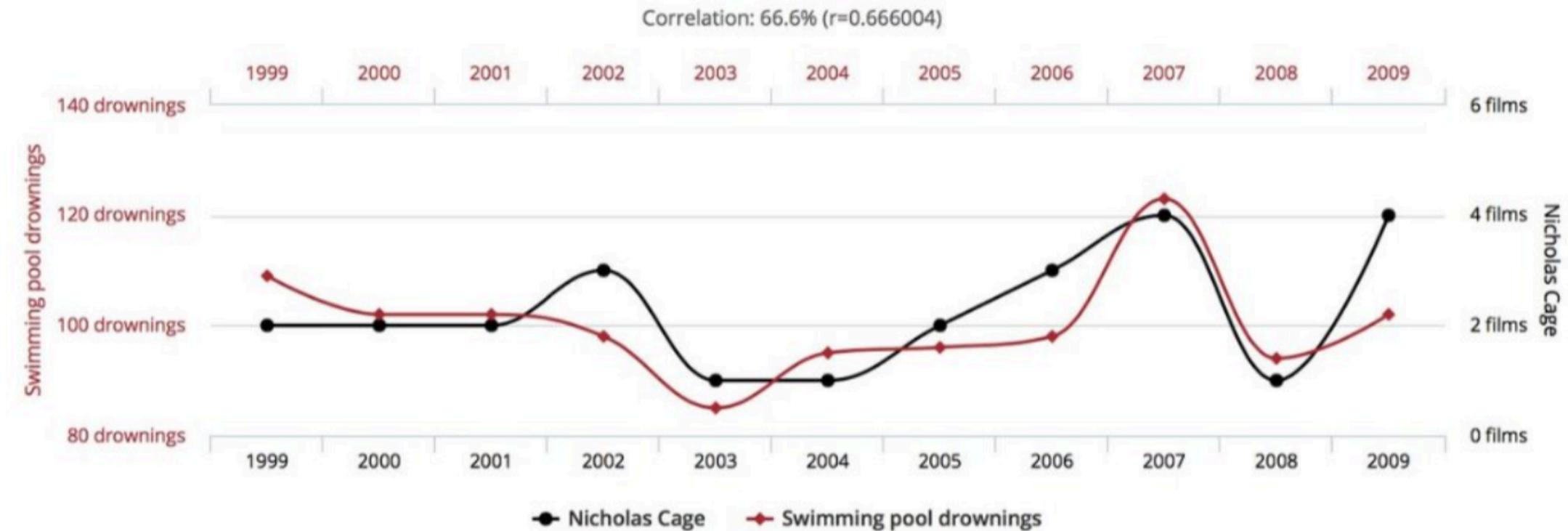
Correlación Muestral o Pearson

$$r(X, Y) = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{S_x S_y}$$



Correlación

Número de personas ahogadas al caer a una piscina
correlaciona con
Películas de Nicolás Cage



- Es importante recalcar que **Causalidad** no es igual a **Correlación**.
- La **Correlación** no se ve afectada por la escala de los datos.

Preguntas para terminar

- ¿En qué se diferencia un estimador muestral de uno poblacional?
- ¿Cuándo es preferible utilizar la Distancia de Mahalanobis?
- ¿Cuál es la diferencia entre Covarianza y Correlación?

Danke Schön

Medidas: Similaridad Datos Categóricos

Atributos

p ₁	shape	p ₂	color
circle	3/8	red	4/8
square	3/8	yellow	2/8
triangle	2/8	blue	2/8

Medidas de Similaridad

Sim	overlap	IOF	Goodall
1-4	1+1	64/9+64/16 ≈ 11.11	55/64+48/64 ≈ 1.61
2-5	1+0	64/9+0 ≈ 7.11	55/64+0 ≈ 0.85
7-8	0+1	0+64/4 = 16.00	0+60/64 = 0.94

frecuencia de ocurrencia inversa (IOF) estima la similaridad entre atributos semejantes. Sea $p_k(x)$ la fracción de registros en los cuales el k -ésimo atributo toma el valor x en el data set.

La medida Goodall da un peso mayor a los valores infrecuentes

$$S(p_1, q_1) + S(p_2, q_2) = \left(\frac{8}{3}\right)^2 + \left(\frac{8}{4}\right)^2 = 11.11$$

$$S(p_1, q_1) + S(p_2, q_2) = 1 - \left(\frac{3}{8}\right)^2 + 1 - \left(\frac{4}{8}\right)^2 = 1.61$$

Medidas: Similaridad (Distancia Coseno)

name	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
d_1	3	2	0	5	0	0	0	2	0	0
d_2	1	0	0	0	0	0	1	1	0	2
d_3	6	4	0	10	0	0	0	4	0	0

$$d_1 \cdot d_2 = 5$$

$$d_1 \cdot d_3 = 84$$

$$\|d_1\| = \sqrt{42} = 6.481$$

$$\|d_2\| = \sqrt{6} = 2.449$$

$$\|d_3\| = \sqrt{168} = 12.962$$

$$\cos(d_1, d_2) = 0.3150$$

$$\cos(d_1, d_3) = 0.9999$$