Correlaciones: ¿qué tiene que ver el deporte con la literatura?

José Calvo Tello, Universität Würzburg

Con el apoyo de Dariah-IE.

Universiteit Antwerpen, 8.10.2019

Presentación y materiales

- https://github.com/morethanbooks/taller-correlaciones

Contenido

- Parte teórica: ¿por qué correlaciones?
  - Ejemplos
  - ¿Por qué puede ser interesante para humanidades y humanidades digitales?
  - Principios
  - Causalidad != correlación
- Parte práctica: ¿cómo calcular correlaciones?
  - Entornos: Calc, Excel, Python, R...
  - Limpieza de datos
  - Visualización
  - Correlaciones
  - Hipótesis

Ejemplos

Ejemplos cotidianos:

- Cuanto más deporte hagas, más sano
- Cuanto más deporte hagas, menos pesarás
- Cuanto más fumes, mayores son las posibilidades de cáncer

Ejemplos de otras ciencias: Cigarros y cáncer

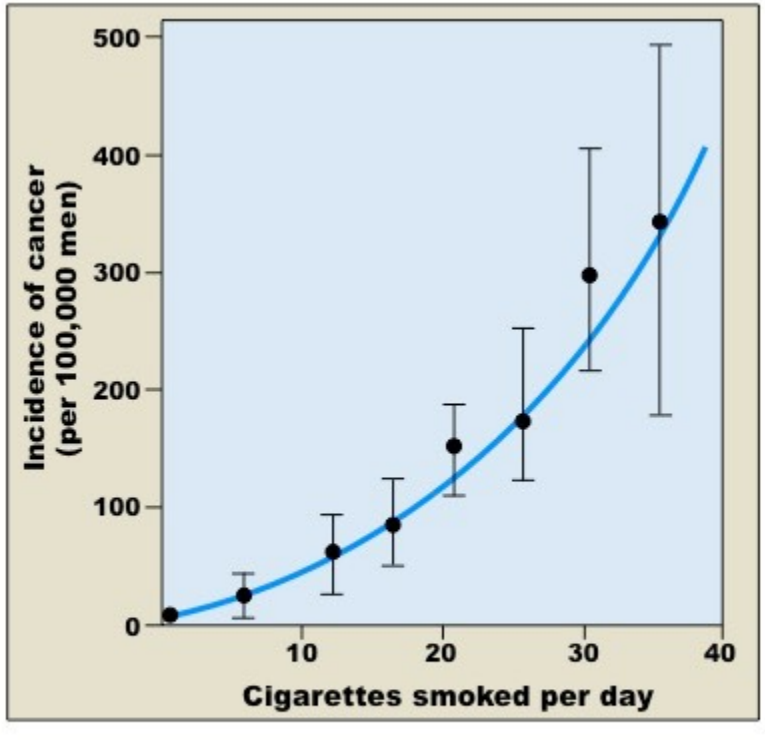


Figure 1: Smoking frequency vs Cancer rates

Source: *New Scientist, 1994*

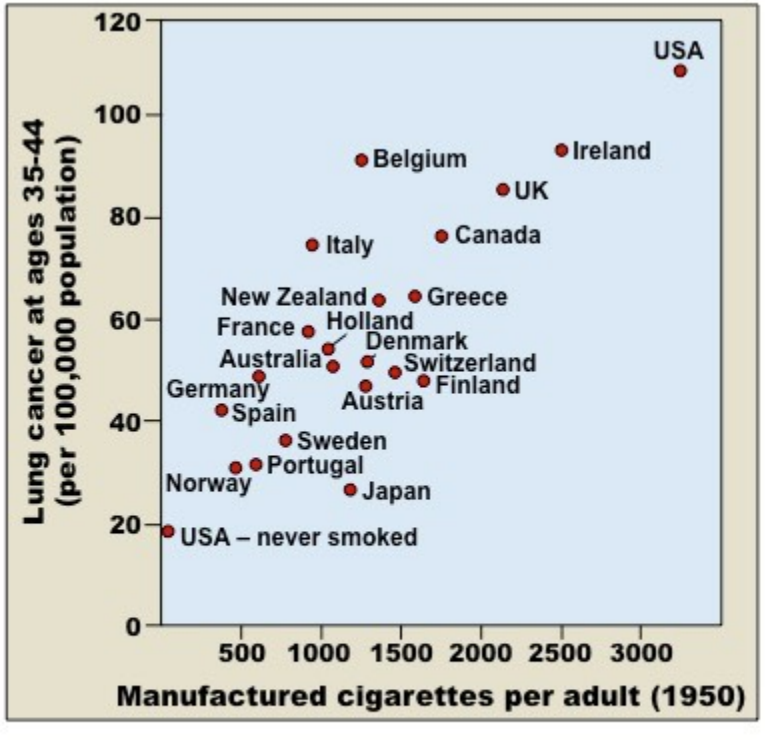
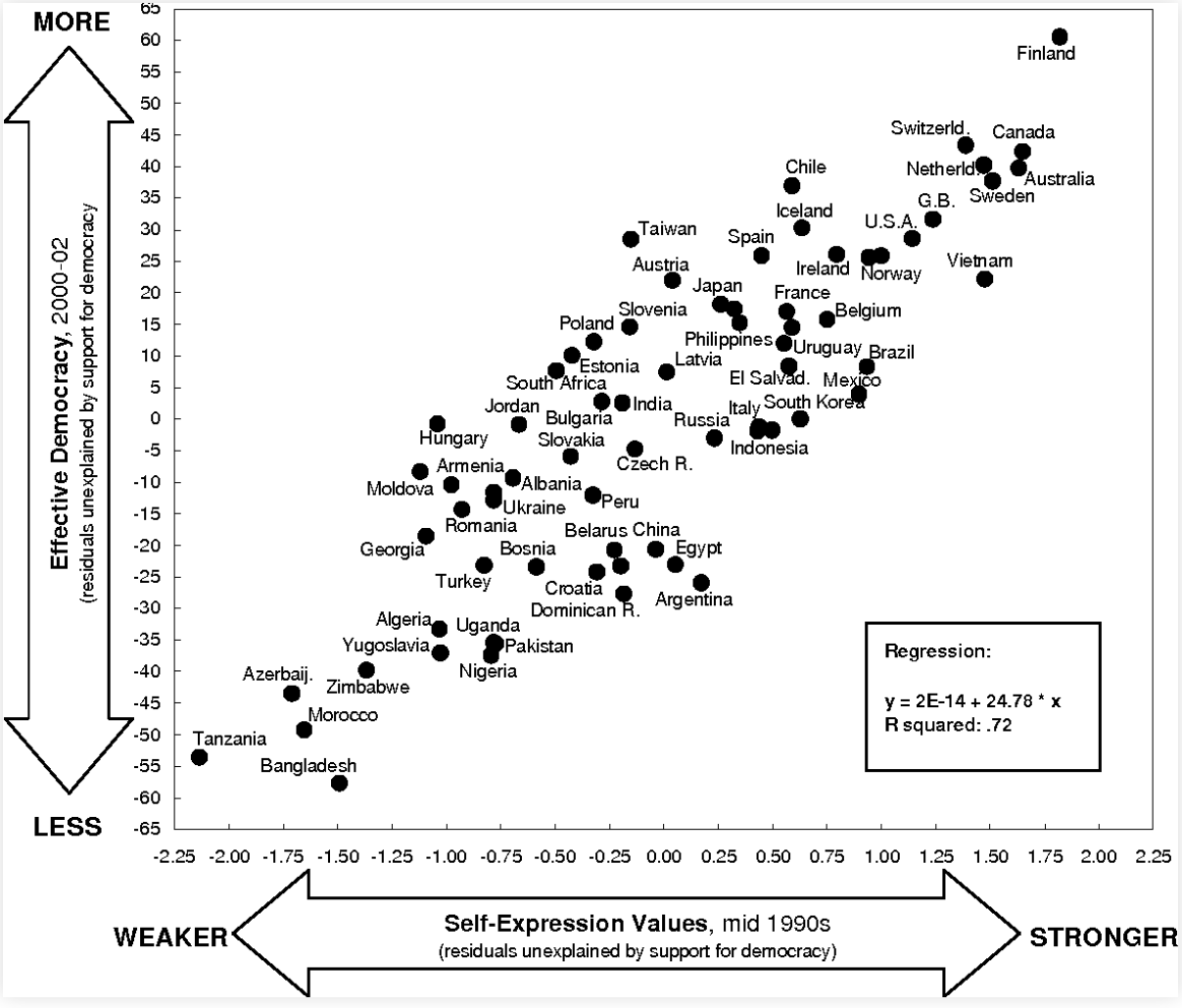


Figure 2: Cigarette availability vs Lung cancer rates

Source: *World Health Organization (WHO), 1977*

Fuente: https://ib.bioninja.com.au/_Media/smoking-and-cancer_med.jpeg

Ejemplos de otras ciencias: Democracia y valores



Fuente: https://www.semanticscholar.org/paper/Modernization%2C-Cultural-Change%2C-and-Democracy%3A-The-Inglehart-Welzel/975f7e0a0969b40fe39e40282580592afc68cce8

Si lo aceptamos en la vida normal y en otras ciencias,
¿por qué no en humanidades?

«Explicar (o juzgar) un hecho es unirlo a otro.»

«Tlön, Uqbar, Orbis Tertius»

Ficciones

Jorge Luis Borges

¿Por qué correlaciones?

¿Por qué es interesante para Humanidades?

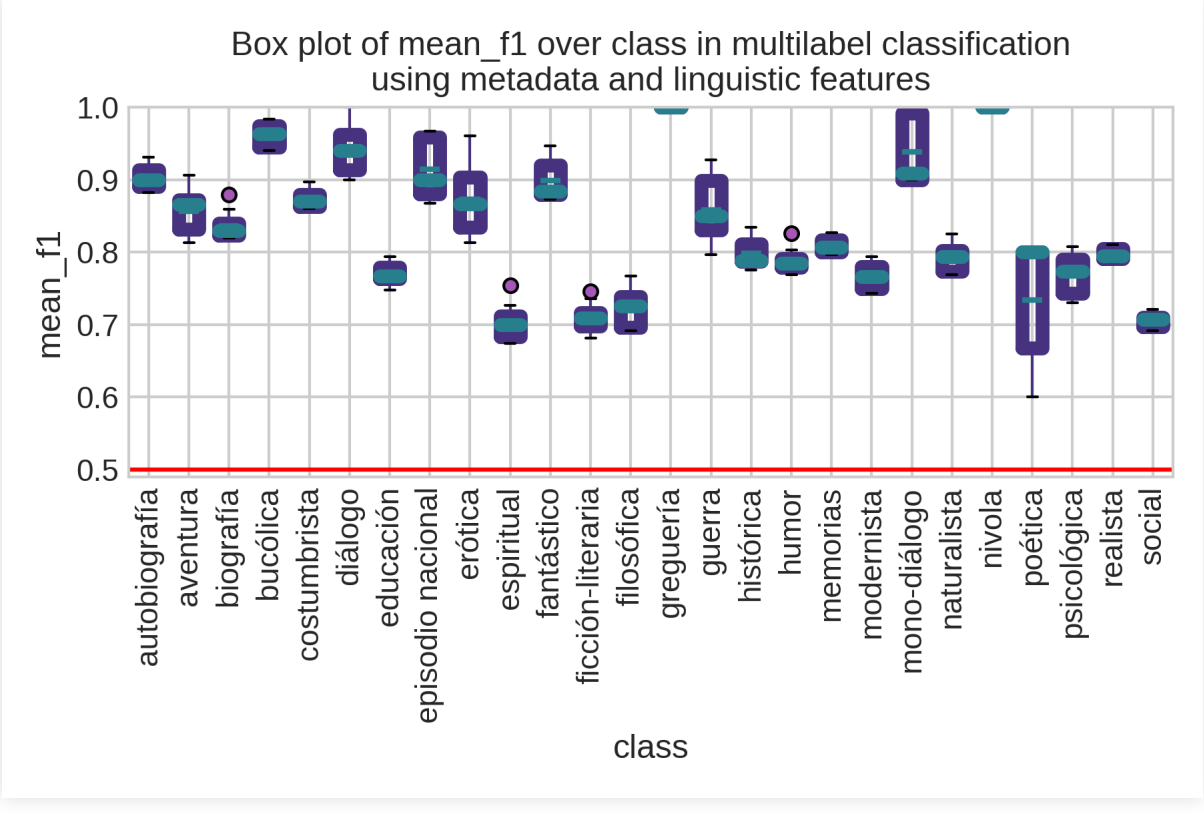
- Utilizar metodologías sólidas utilizadas en otras ciencias
- Trabajar con numerosos tipos de unidades
- Trabajar con cantidades tanto medianas como grandes de datos

¿Por qué es interesante para Humanidades Digitales?

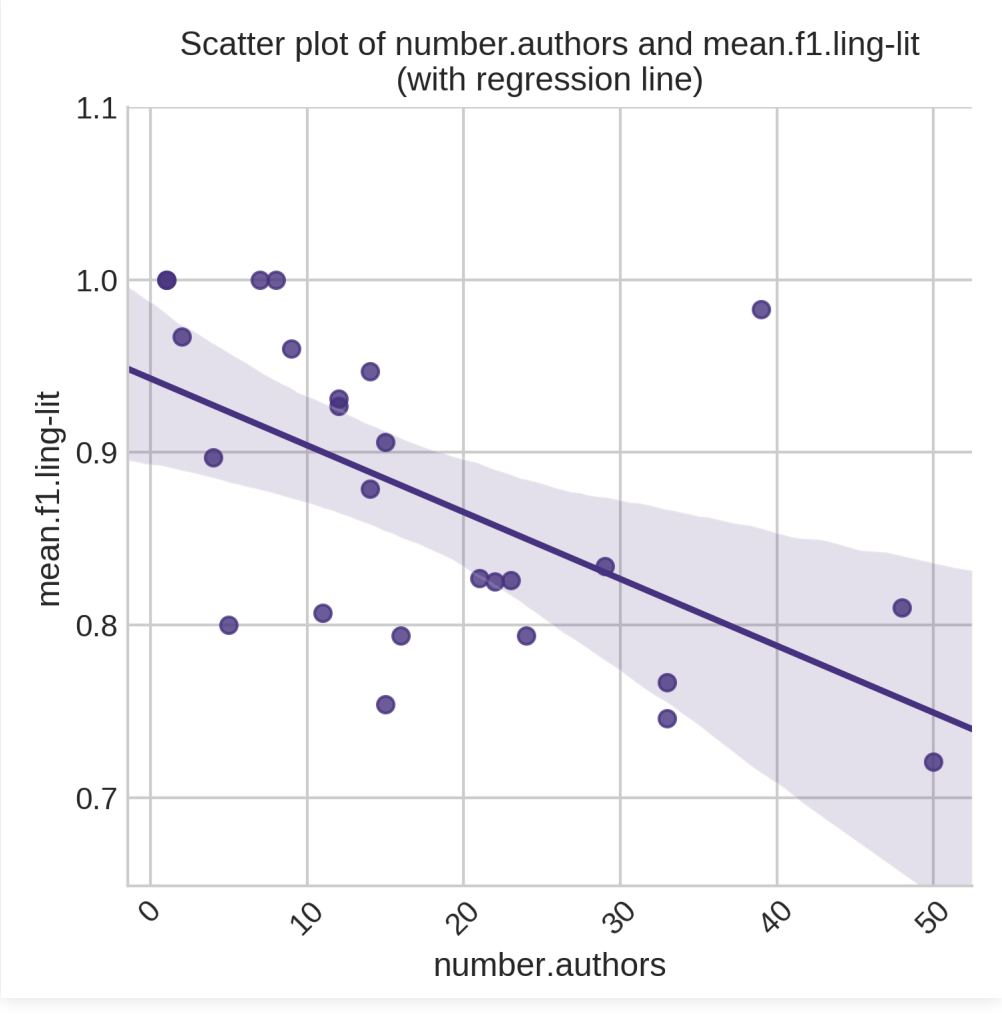
¿Por qué es interesante para Humanidades Digitales?

- Evaluar y consolidar resultados
- Ayudar a entender qué hacen los procesos/algoritmos
- Abrir la posibilidad de que cierto métodos pueden no ser útiles

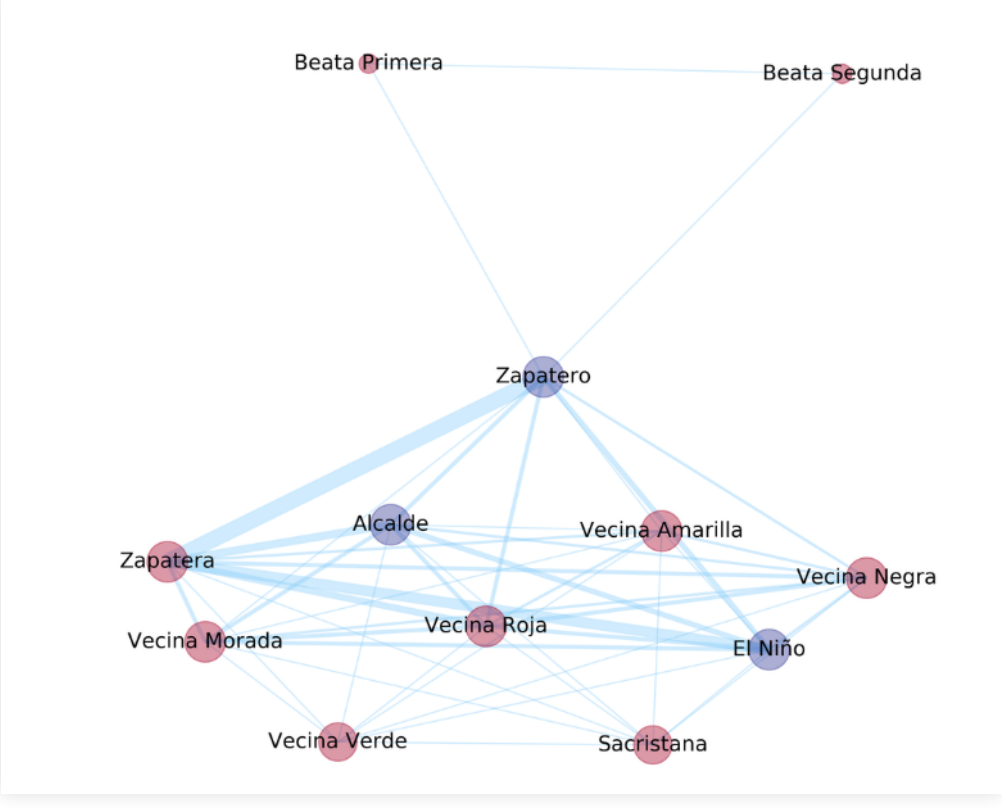
Clasificación de géneros literarios



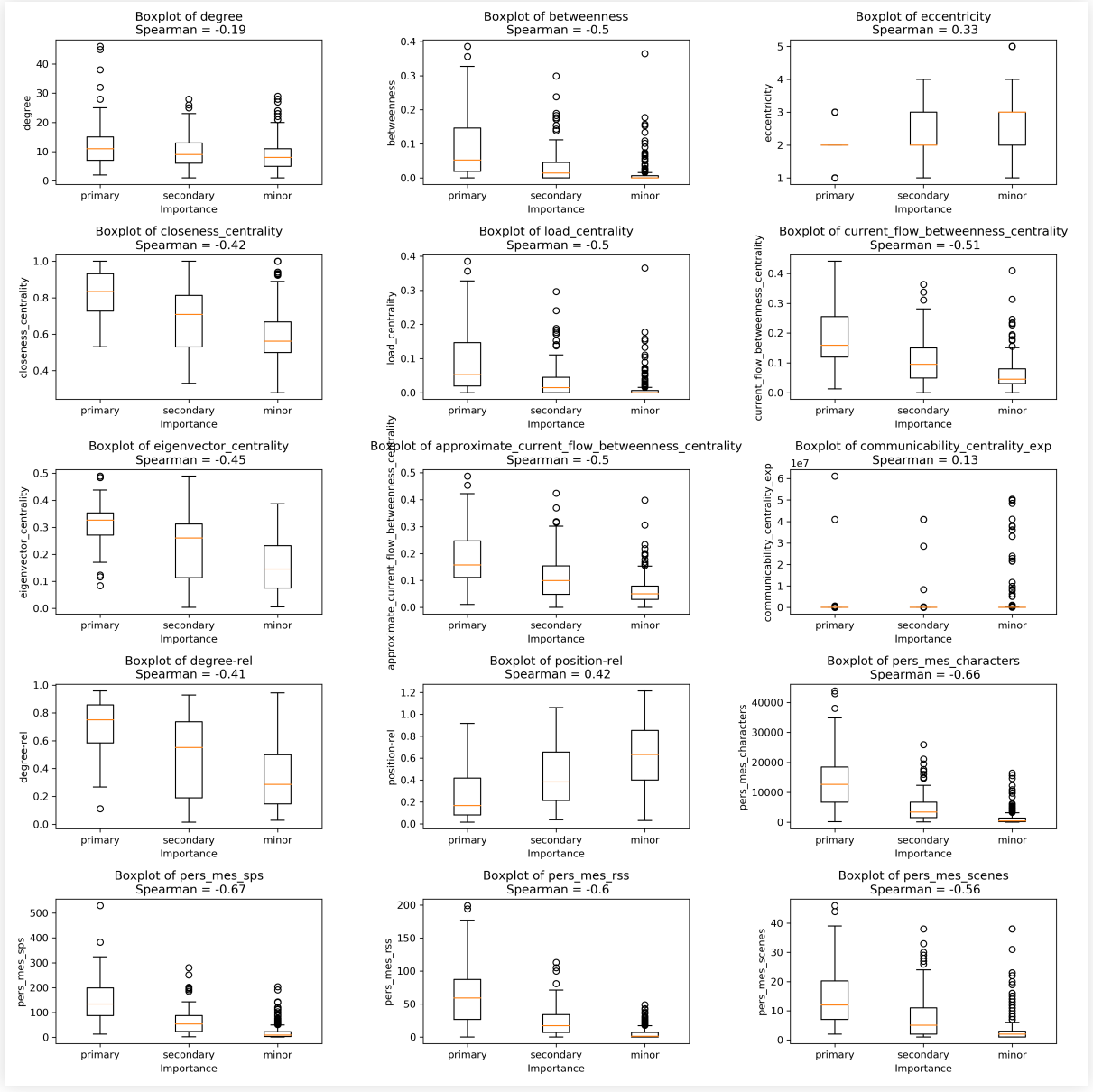
¿Por qué algunos géneros son más fáciles de clasificar?



Importancia de personajes y redes sociales



Fuente: Santa María, T., Martínez Carro, E., Jiménez, C., and Calvo Tello, J. (2018). ¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata? in DH2018 (México DF: ADHO), 494–498.


Boxplot of degree
Spearman = -0.19
degree
Importance
primary
secondary
minor
Boxplot of betweenness
Spearman = -0.5
betweenness
Boxplot of eccentricity
Spearman = 0.33
eccentricity
Boxplot of closeness_centrality
Spearman = -0.42
closeness_centrality
Boxplot of load_centrality
Spearman = -0.5
load_centrality
Boxplot of current_flow_betweenness_centrality
Spearman = -0.51
current_flow_betweenness_centrality
Boxplot of eigenvector_centrality
Spearman = -0.45
eigenvector_centrality
Boxplot of approximate_current_flow_betweenness_centrality
Spearman = -0.5
approximate_current_flow_betweenness_centrality
Boxplot of communicability_centrality_exp
Spearman = 0.13
1e7
communicability_centrality_exp
Boxplot of degree-rel
Spearman = -0.41
degree-rel
Boxplot of position-rel
Spearman = 0.42
position-rel
Boxplot of pers_mes_characters
Spearman = -0.66
pers_mes_characters
Boxplot of pers_mes_sps
Spearman = -0.67
pers_mes_sps
Boxplot of pers_mes_rss
Spearman = -0.6
pers_mes_rss
Boxplot of pers_mes_scenes
Spearman = -0.56
pers_mes_scenes

Resultado:

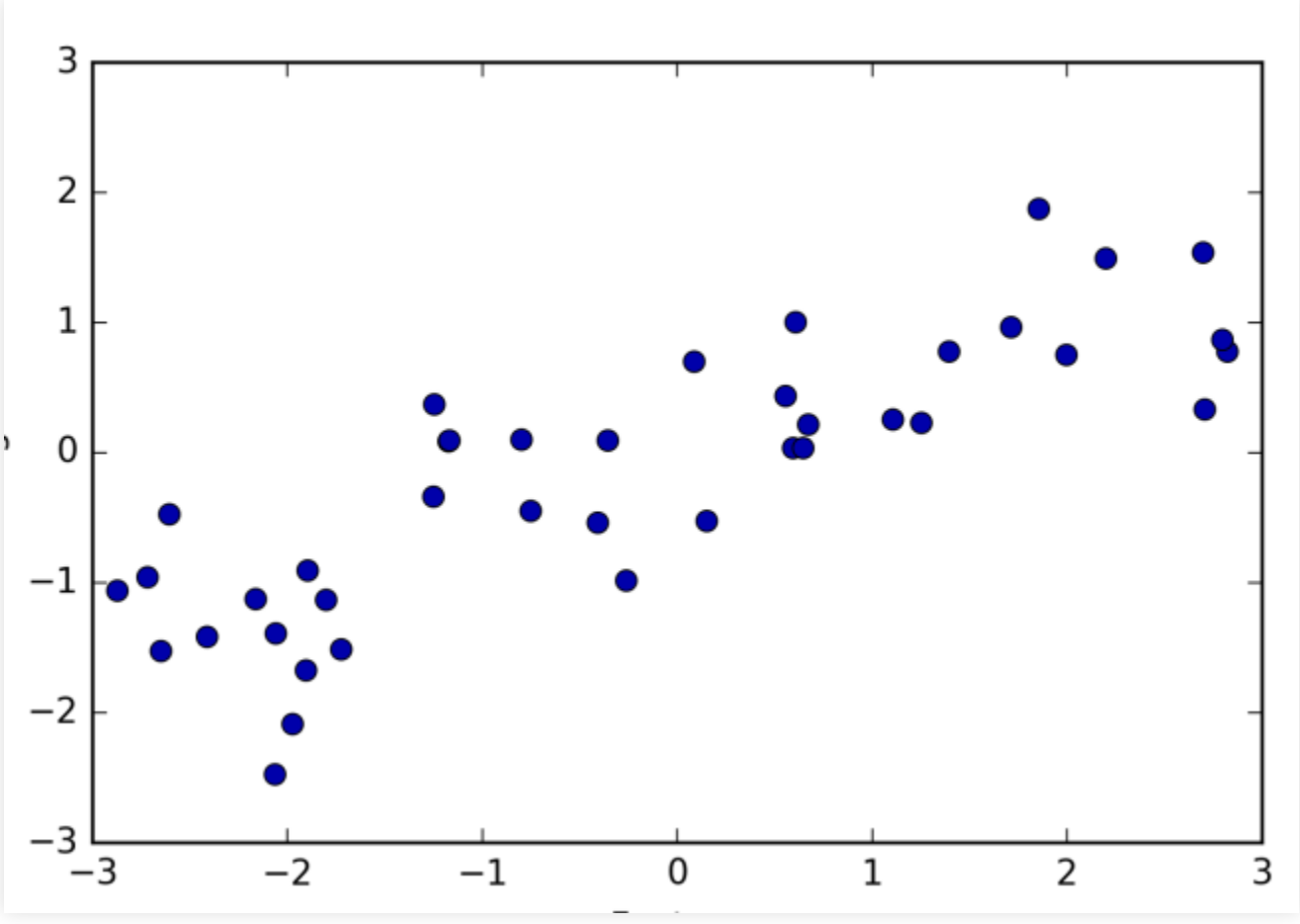
- Las redes nos dicen muy poco sobre la importancia de los personajes
- Mucho maś importante es saber cuántas veces habla el personaje

Hipótesis:

- hipótesis de autores
  - los autores más importantes son más digitalizados
  - cuanto más tarde nació un autor, más vivió
  - cuanto más vivió un autor, más publicó
- hipótesis de novelas
  - las novelas más largas representan historias más largas (en días)
  - las novelas más largas son más importantes
  - cuanto más tarde se publicó una novela, más corta
  - cuanto más rico sea el protagonista, más feliz será el final

Principios

¿Correlaciones?



- Una instancia
- Dos variables numéricas
- ¿Hay relación entre ellas?

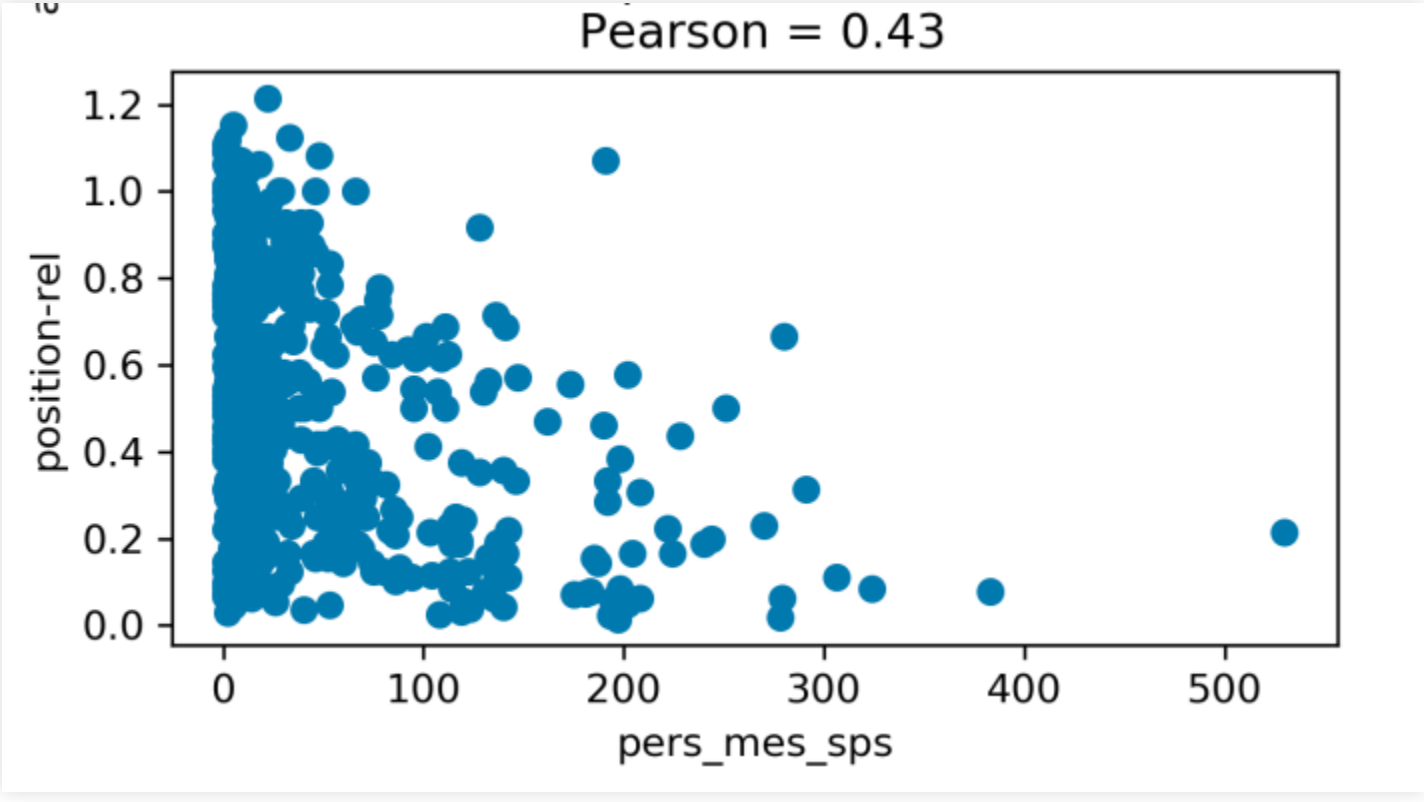
Una instancia, dos variables numéricas

- Instancias: País, género literario, persona, texto, personaje, autor...
- Las variables numéricas pueden ser tan sencillas, tan cuidadas, tan digitales como queramos

Limitaciones

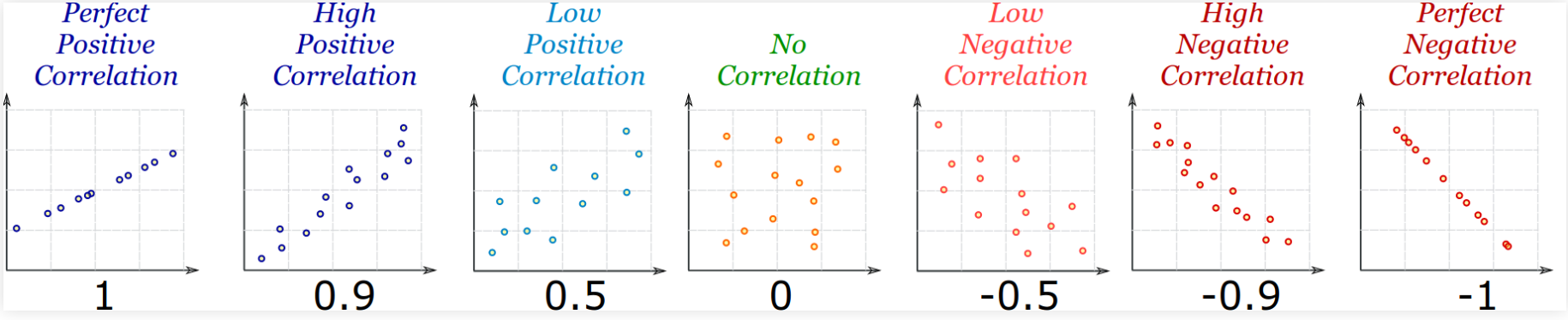
- **Correlaciones no pueden responder cualquier pregunta**
- Diferencias categoriales no se pueden representar de manera numérica. Por ejemplo, si queremos analizar diferencias de texto en cuanto al género de su autor (sexo) o género literario, no podremos utilizar correlaciones

Visualización mediante Scatter Plot



- Dos variables numéricas
- Eje horizontal (x): *variable independiente*
- Eje vertical (y): *variable dependiente*

Tipos de correlaciones



Fuente: https://www.mathsisfun.com/data/images/correlation-examples.svg

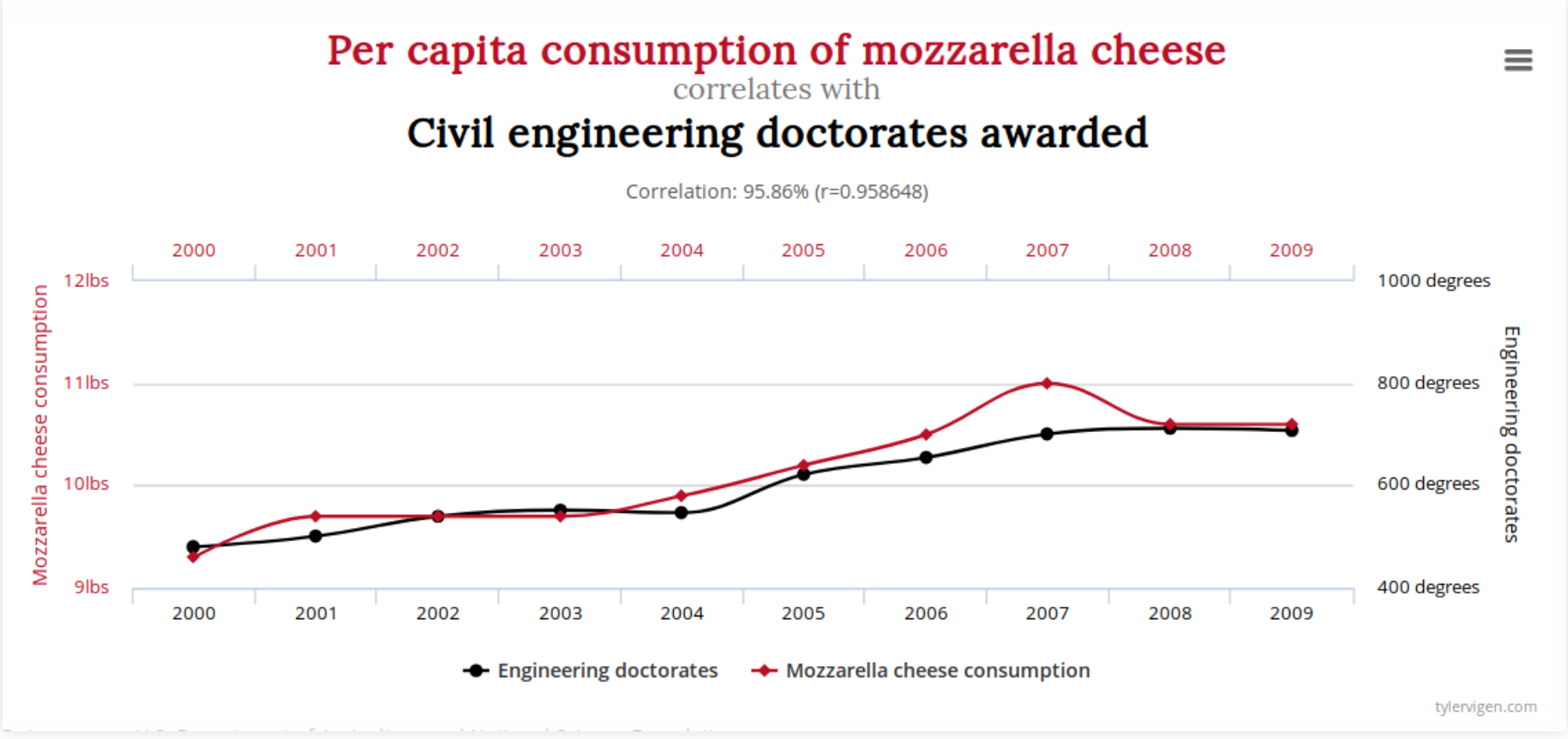
Tipos de datos numéricos:

- Intervalo: en unidades concretas. Ejs: altura, temperatura, cantidad de días que pasan en una novela.
- Ordinales: orden impuesto (la persona más alta, la segunda persona más alta... la persona más baja). Ejs: novela con final: muy feliz, feliz, neutro, triste, muy triste.

Tipos de relaciones

- Correlación entre **variables de intervalo**: Pearson's R
- Correlación entre **variables ordinales**: Spearman's Rho, Kendall' Tau
- Regresión: suponemos causalidad. Coeficiente (coefficient) y pendiente (slope).

Causalidad != Correlación



- Source: Spurious correlations

Causalidad != correlación

- Posibles variables de confusión
- Cantidad grande de datos (no absurdamente grandes)
- Motivado por hipótesis
- Por ejemplo: la clasificación de los géneros empeora cuantas más novelas contenga el género. Cuantas más novelas, peor es la clasificación. Sin embargo la cantidad de autores practicando un género tiene una correlación aún maś fuerte. Cantidad de autores es una variable de confusión. Cuantos más autores en un género, más difícil será reconocerlo.

Causalidad != Correlación

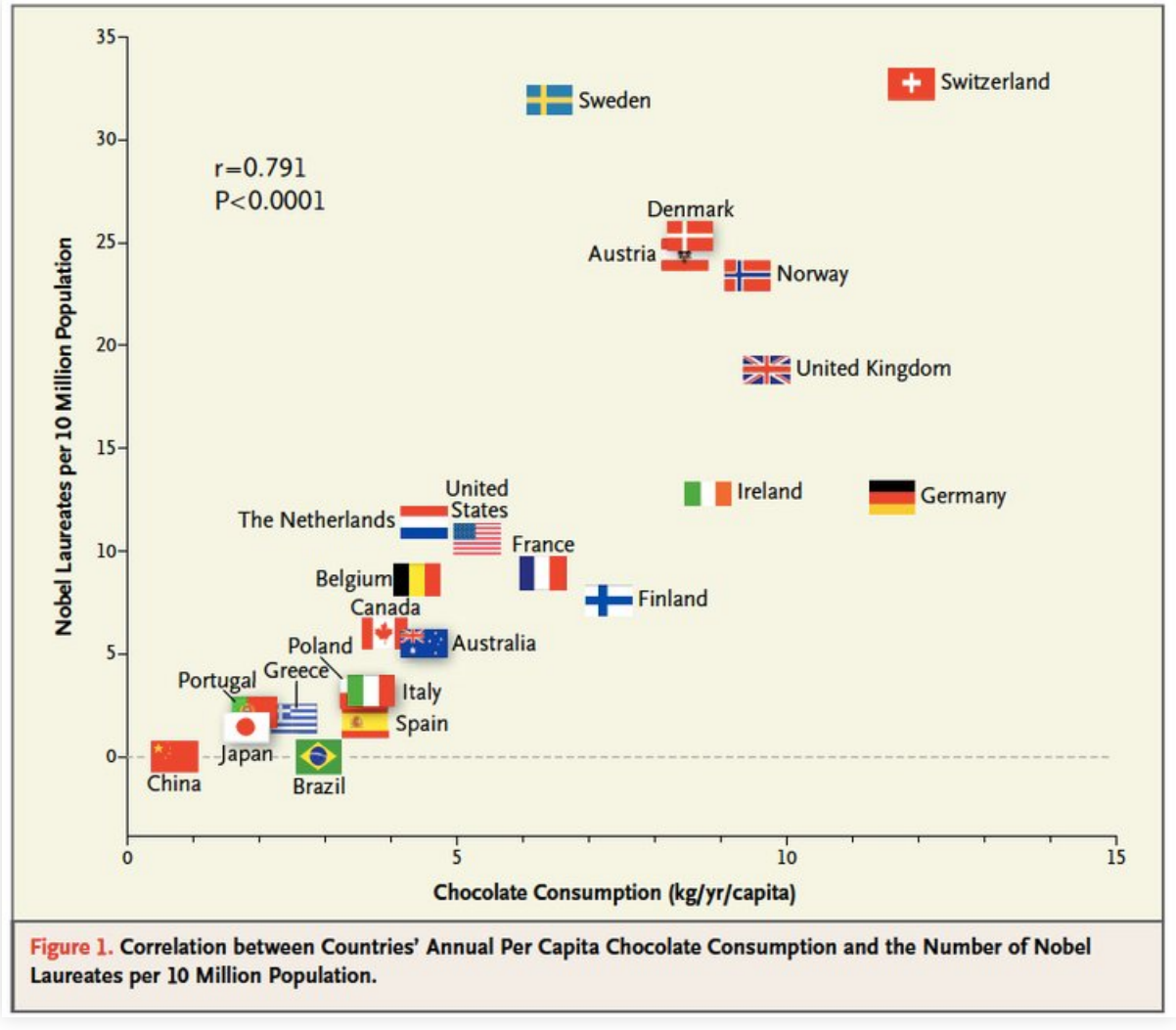


Figure 1. **Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.**

Parte práctica: ¿cómo calcular correlaciones?

Entornos

Entornos

- **Calc or Excel**: entornos familiares, no hace falta instalar ni aprender (casi) nada. Ventaja: sencillo. Desventaja: no todas las posibilidades, falta de cálcuo de valor p, limitaciones en cuanto a la visualización.
- **Python or R**: lenguas de programación. Mucho más potentes, posibilidades de evaluación y visualización.

Datos

Abriendo CSV (o TSV)

Explorando datos

Creando hoja de cálculo

Visualización

Funciones en Calc

Funciones en Calc

- =CORREL(series1, series2)
- =PEARSON(series1, series2)
- =SLOPE(seriesY, seriesX)

Interpretración de Pearson o Spearman

- ¿Cómo de fuerte es la correlación?
- Interpretación (Evans 1996):

If *r* is in this range:	Give *r* this label:
.80 to 1.00	Very strong
.60 to .79	Strong
.40 to .59	Moderate
.20 to .39	Weak
.00 to .19	Negligible to very weak

Hipótesis

Hipótesis:

- hipótesis de autores
  - los autores más importantes son más digitalizados
  - cuanto más tarde nació un autor, más vivió
  - cuanto más vivió un autor, más publicó
- hipótesis de novelas
  - las novelas más largas representan historias más largas
  - las novelas más largas son más importantes
  - cuanto más tarde se publicó una novela, más corta
  - cuanto más rico sea el protagonista, más feliz será el final