KIBERNUM
IT Academy

Módulo 3
Clase 1

Introducción a la Ciencia de Datos

Análisis Multivariado

Objetivos

- Aprender Instrucciones básicas de Python
- Conocer sobre tipos y estructuras de datos, operadores y expresiones
- Conocer sobre flujos de control
- Codificar un programa creando funciones

KIBERNUM

Análisis Multivariado

El análisis multivariado está referido a las técnicas estadísticas para analizar la data que proviene de más de una variable. A continuación algunos ejemplos de visualizaciones. Para este ejemplo, utilizaremos el dataset Iris.

```
1 df = pd.read_csv('iris.csv')
```

```
1 df.head()
```

	sepal_length	sepal_width	petal_length	petal_width	species
0	5.1	3.5	1.4	0.2	setosa
1	4.9	3.0	1.4	0.2	setosa
2	4.7	3.2	1.3	0.2	setosa
3	4.6	3.1	1.5	0.2	setosa
4	5.0	3.6	1.4	0.2	setosa

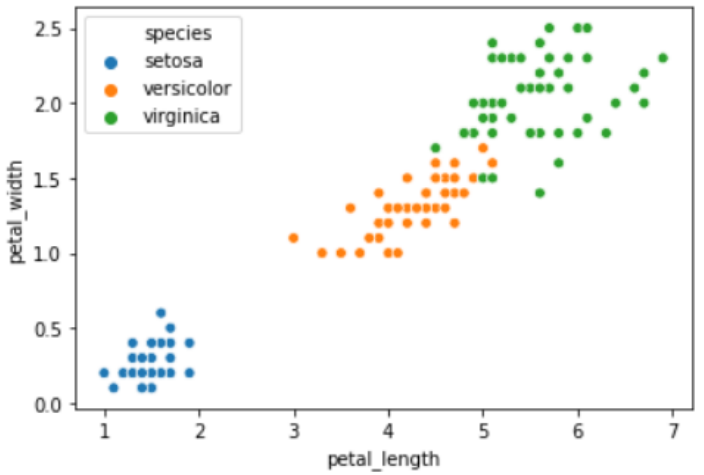
Diagrama de Dispersión

A continuación, vamos a utilizar la librería Seaborn para hacer un scaterplot en donde podamos distinguir cada especie de Iris.

```
import seaborn as sns
```

```
1 sns.scatterplot(data=df, x='petal_length',y='petal_width', hue='species')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d285e3688>
```



Como se puede apreciar, esta característica permite separar las distintas especies. Setosa tiene los tamaños más chicos de pétalos mientras que Virginica los más grandes

Diagrama de Barras

- Un diagrama de barras representa datos categóricos, desplegando barras rectangulares con largos proporcionales al valor que representan.

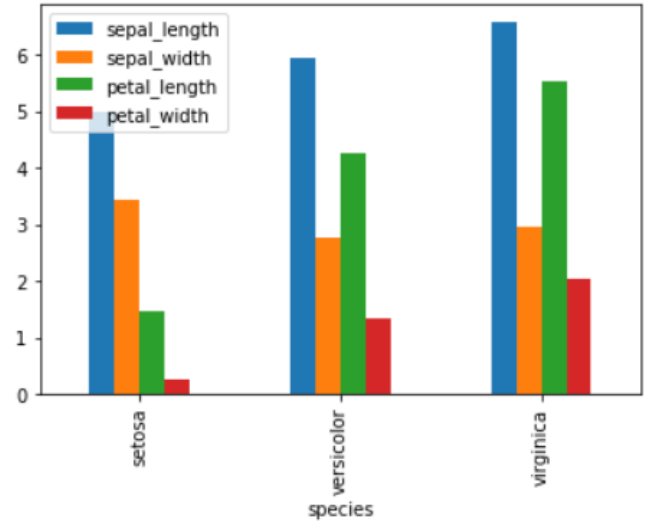
- En este caso, representamos el promedio de cada variable para cada una de las especies. Como se puede observar, la especie Virgínica es la que posee en promedio mayor tamaño tanto de largo y ancho del pétalo, así como el largo del sépalo

```
1 df.groupby('species').mean()
```

species	sepal_length	sepal_width	petal_length	petal_width
setosa	5.006	3.428	1.462	0.246
versicolor	5.936	2.770	4.260	1.326
virginica	6.588	2.974	5.552	2.026

```
1 df.groupby('species').mean().plot(kind='bar')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d26e30e48>
```



```
1 df.groupby('species').count()
```

species	sepal_length	sepal_width	petal_length	petal_width
setosa	50	50	50	50
versicolor	50	50	50	50
virginica	50	50	50	50

```
1 df.groupby('species').count()['petal_length'].plot(kind='pie', autopct='%.2f%%')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x19d2a498708>
```

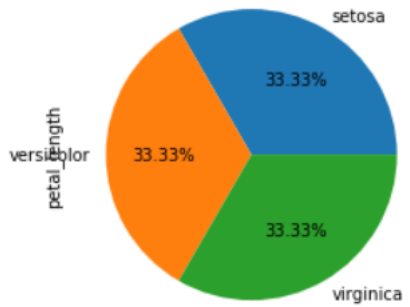


Diagrama de Torta

Un diagrama de torta permite una visualización de las proporciones de alguna variable categórica.
En el siguiente ejemplo, visualizamos la proporción de cada especie respecto a la cantidad de mediciones de cada una de ellas.

KIBERNUM

Correlación

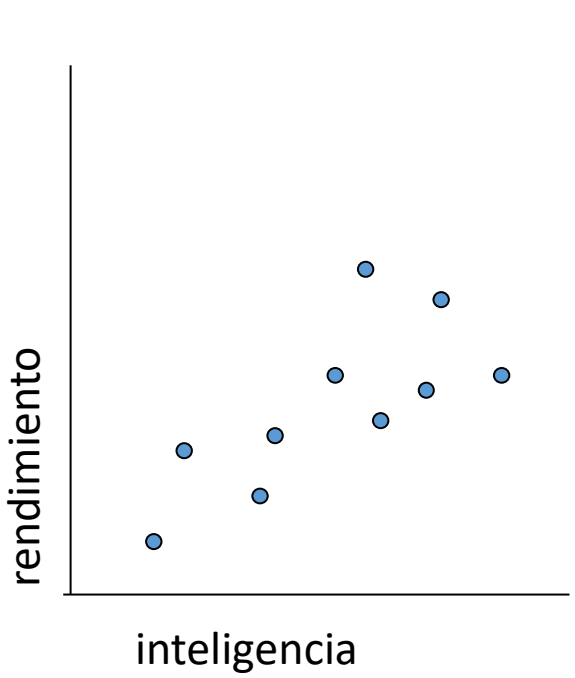
- Hasta ahora nos hemos centrado en medidas de tendencia central, variabilidad y asimetría de una única variable. No obstante, en la práctica es común examinar dos o más variables conjuntamente (v.g., relación entre inteligencia y rendimiento, etc.)

- En este tema, nos centraremos en la relación entre 2 variables (a partir de n observaciones apareadas) y calcularemos (en particular) un índice que nos dará el grado de relación/asociación entre ambas variables: el coeficiente de correlación

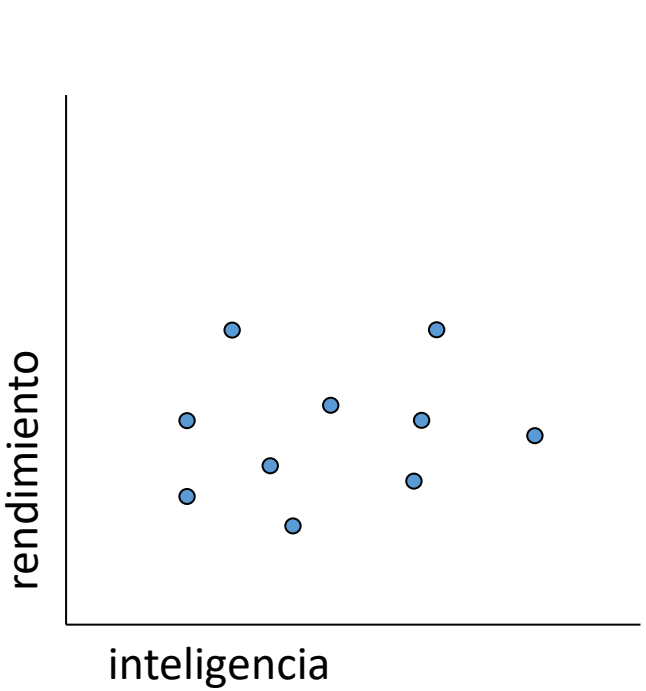
Una **correlación** es una medida o grado de relación entre dos variables.

KIBERNUM

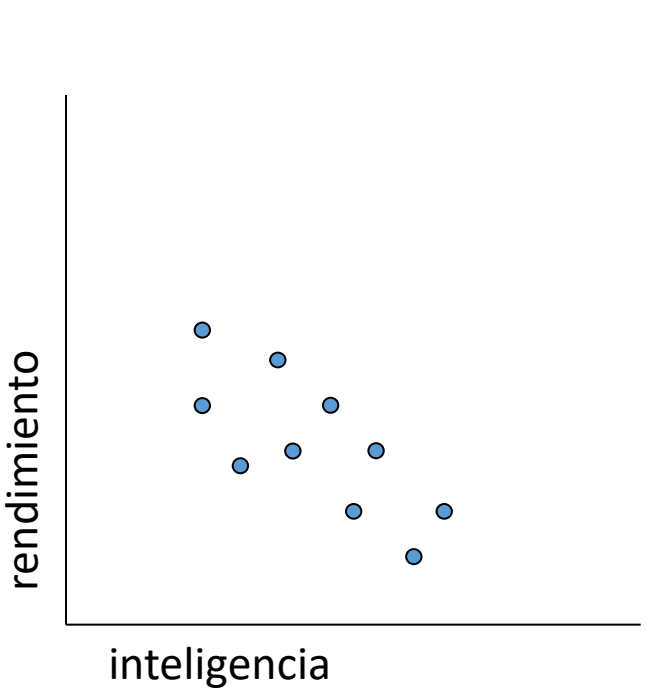
Representación gráfica de una relación



Relación lineal positiva



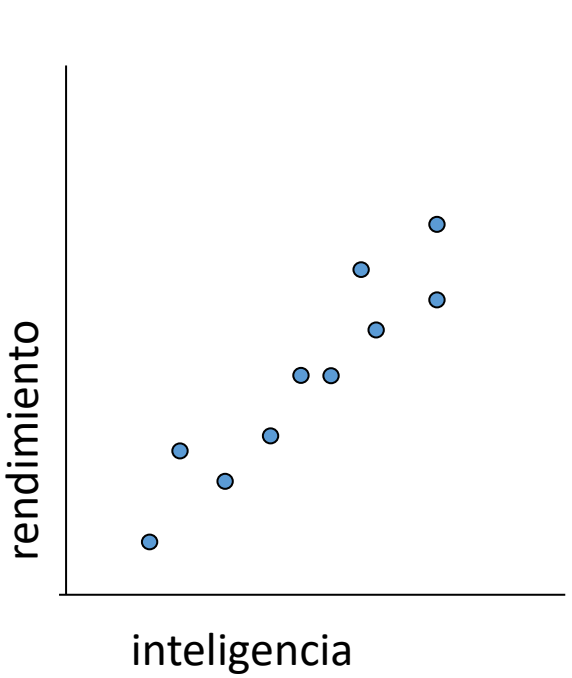
Sin relación



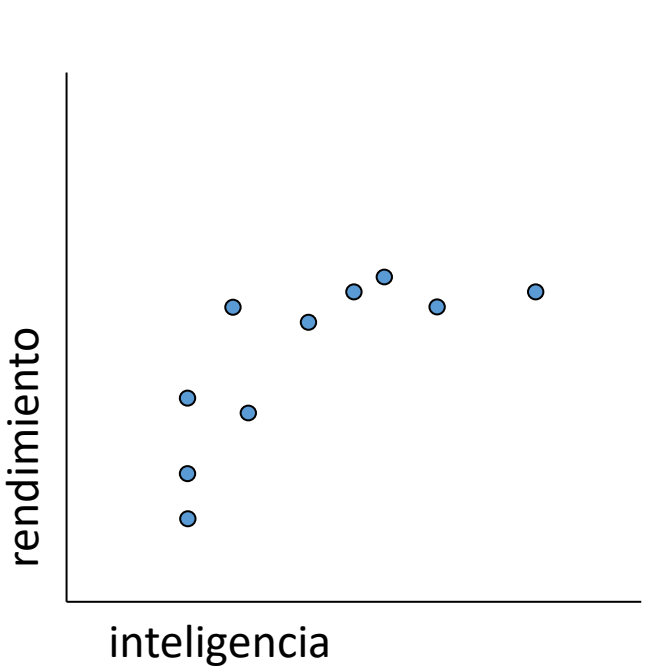
Relación lineal negativa

Nota: El coeficiente de correlación de Pearson mide relación LINEAL.

Representación gráfica de una relación



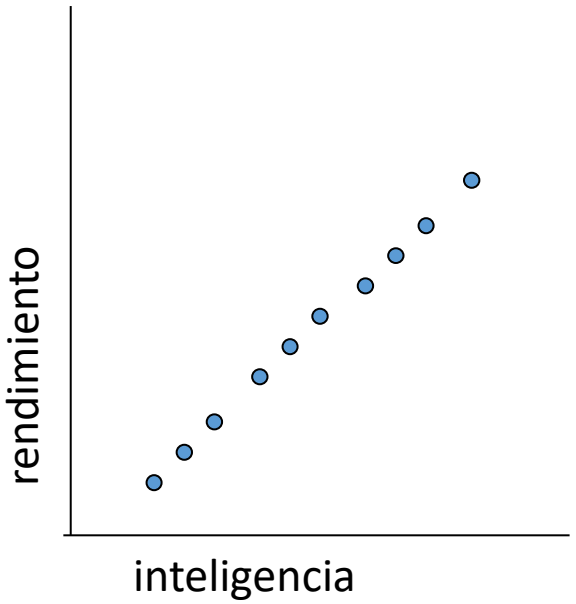
Relación lineal



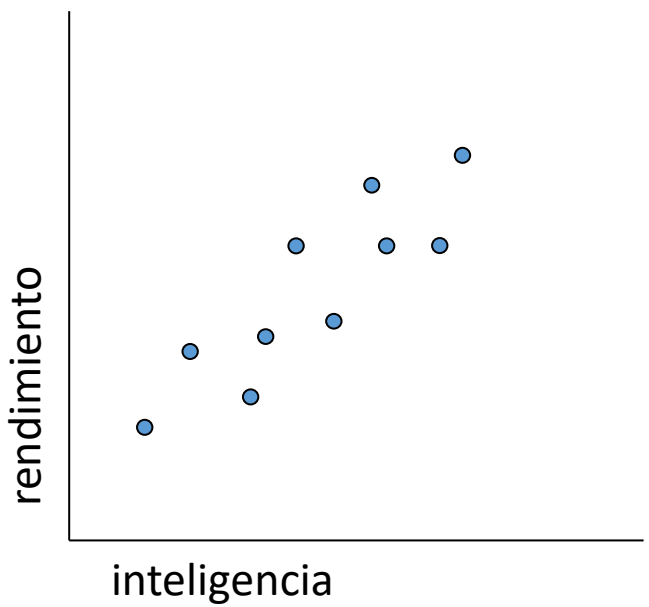
Relación no lineal

Representación gráfica de una relación

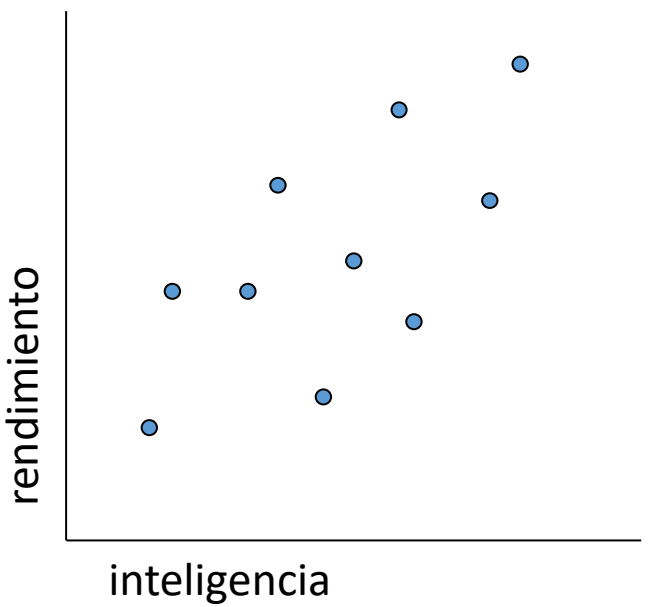
Ahora necesitamos un índice que nos informe tanto del grado en que X e Y están relacionadas, y si la relación es positiva o negativa



Relación lineal perfecta (casi perfecta)



Relación lineal fuerte/moderada



Relación lineal débil

Correlación vs Causalidad

Es importante indicar que “CORRELACIÓN NO IMPLICA CAUSACIÓN”. El que dos variables estén altamente correlacionadas (se mueven juntas) no implica necesariamente que X causa Y ni que Y causa X.

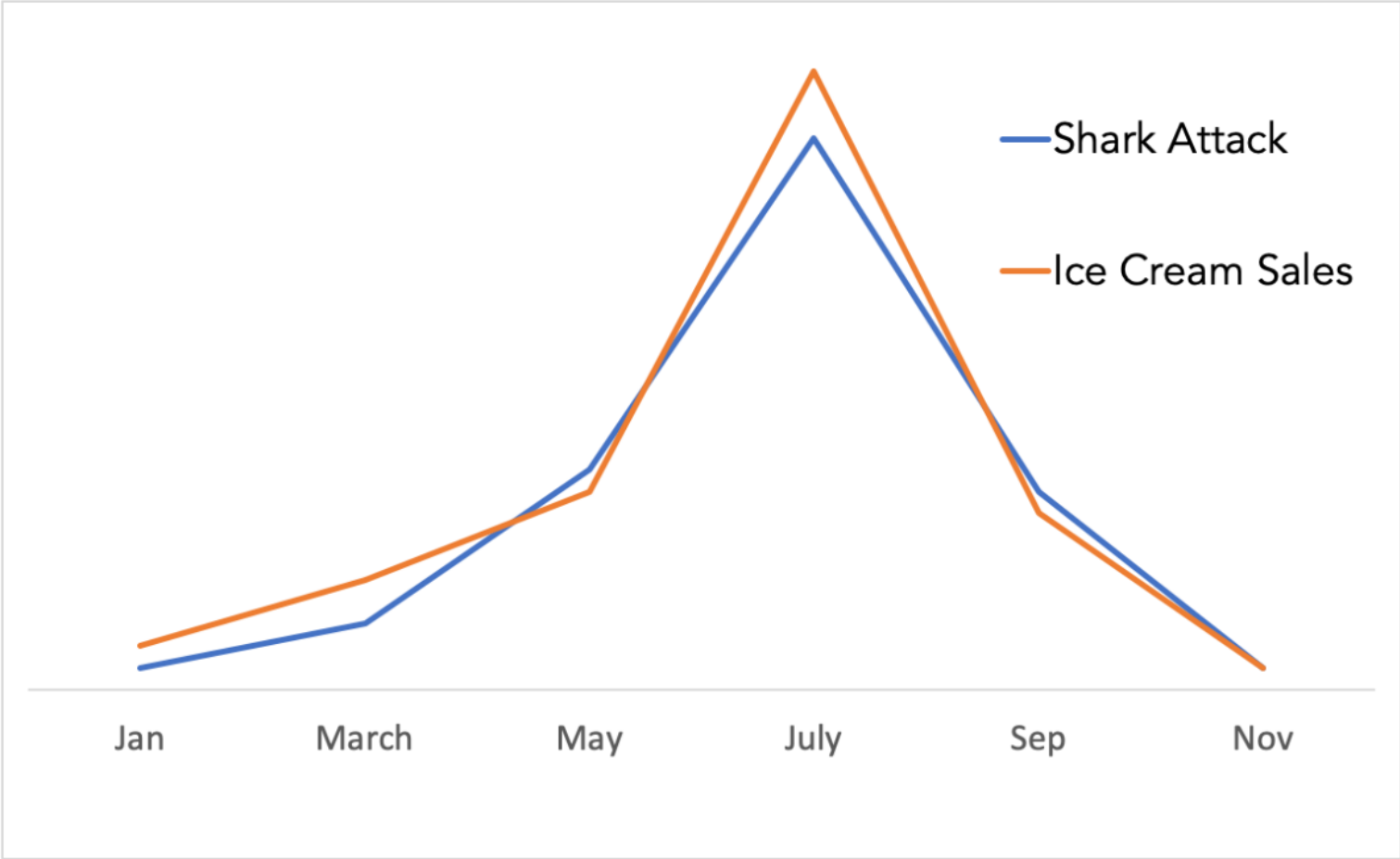
- (Esa es una de las razones empleadas por las tabaqueras en el tema de la correlación entre cáncer de pulmón y el hecho de fumar.)

Una correlación fuerte puede indicar causalidad, pero también es probable que existan otras explicaciones:

- Puede ser el resultado del azar: las variables parecen estar relacionadas, pero en realidad no hay una relación subyacente.

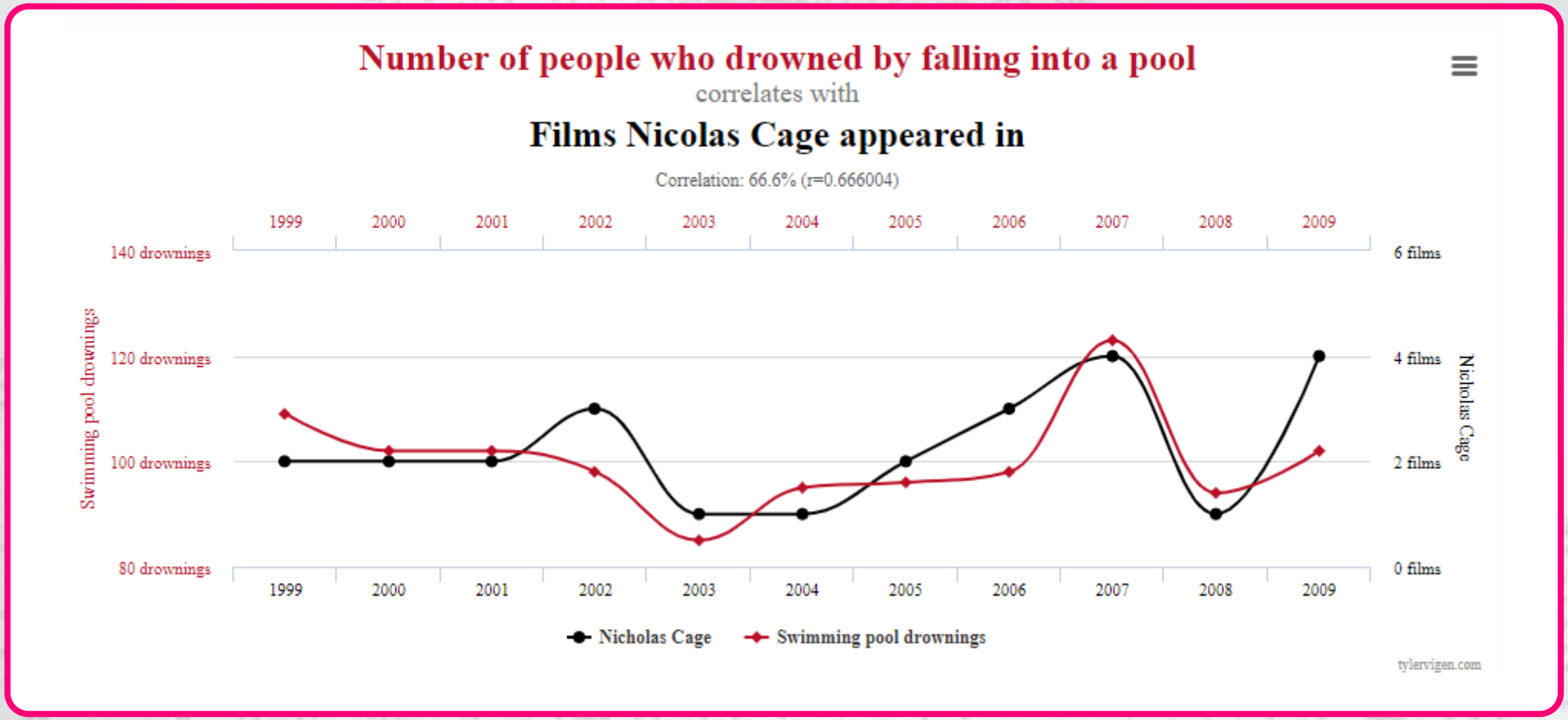
- Puede haber una tercera variable al acecho que haga que la relación parezca más fuerte (o más débil) de lo que realmente es.

Correlación y Causalidad



¡¡¡Correlación NO implica causalidad !!!

Correlación y Causalidad



Correlaciones Espurias

https://tylervigen.com/spurious-correlations

Covarianza

- La covarianza emplea el producto de las puntuaciones diferencias de X e Y :

- En el caso pendiente positiva, la covarianza será un valor positivo, y en el caso pendiente negativa, la covarianza será un valor negativo. Por tanto la covarianza nos da una idea de si la relación entre X e Y es positiva o negativa.

- Problema: la covarianza no es un índice acotado (v.g., cómo interpretar una covarianza de 6 en términos del grado de asociación), y no tiene en cuenta la variabilidad de las variables. Por eso se emplea el siguiente índice....

$$s_{xy} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n}$$

KIBERNUM

Coeficiente de correlación (lineal) de Pearson

El coeficiente de correlación de Pearson parte de la covarianza:

$$r_{xy} = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n \cdot s_x \cdot s_y} \qquad r_{xy} = \frac{s_{xy}}{s_x \cdot s_y}$$

$$r_{xy} = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}}$$

Ahora veremos varias propiedades del índice...

Coeficiente de correlación (lineal) de Pearson

Propiedad 1. El índice de correlación de Pearson no puede valer menos de -1 ni más de +1.

Un índice de correlación de Pearson de -1 indica una relación lineal negativa perfecta

Un índice de correlación de Pearson de +1 indica una relación lineal positiva perfecta.

Un índice de correlación de Pearson de 0 indica ausencia de relación lineal. (Observad que un valor cercano a 0 del índice no implica que no haya algún tipo de relación no lineal: el índice de Pearson mide relación lineal.)

Coeficiente de correlación (lineal) de Pearson

Propiedad 2. El índice de correlación de Pearson (en valor absoluto) no varía cuando se transforman linealmente las variables.

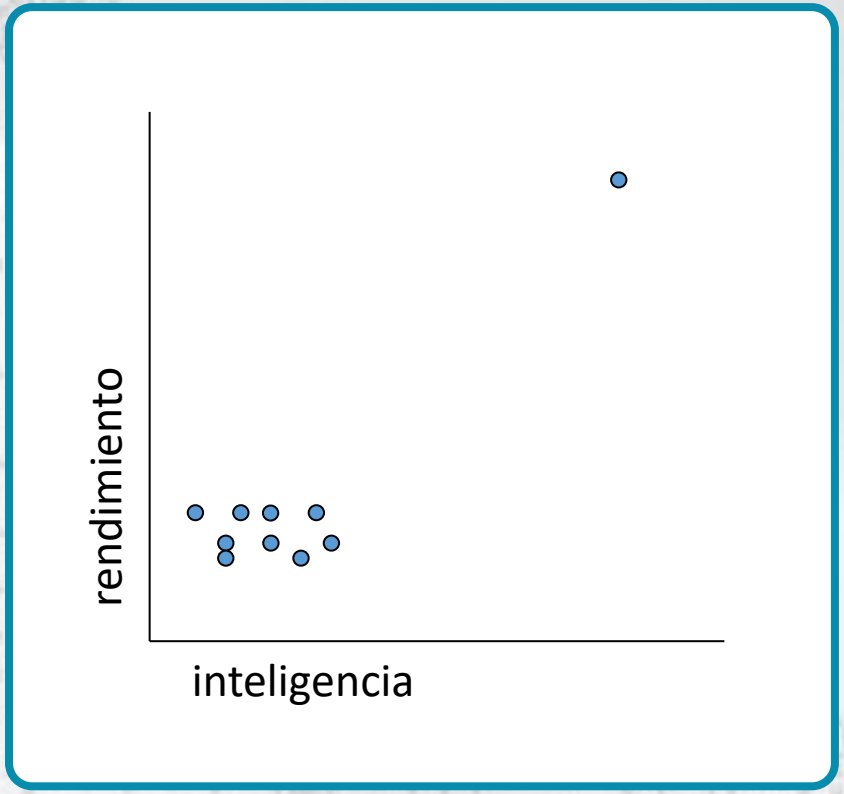
Por ejemplo, la correlación de Pearson entre la temperatura (en grados celsius) y el nivel de depresión es la misma que la correlación entre la temperatura (medida en grados Fahrenheit) y el nivel de depresión.

Coeficiente de correlación (lineal) de Pearson

- **Interpretación**

- Tener en cuenta qué es lo que estamos midiendo para poder interpretar cuán grande es la relación entre las variables bajo estudio. En muchos casos, depende del área bajo estudio.

- En todo caso, es muy importante efectuar el diagrama de dispersión. Por ejemplo, en el caso de la izquierda, es claro que no hay relación entre inteligencia y rendimiento. Sin embargo, si calculamos el índice de correlación de Pearson nos dará un valor muy elevado, causado por la puntuación atípica en la esquina superior derecha.



Otros coeficientes: variables semi-cuantitativas

Es posible obtener medidas del grado de relación de variables cuando éstas no sean cuantitativas.

El caso en que las variables X e Y sean ordinales

Cuando tenemos variables con escala ordinal, podemos establecer el orden entre los valores, pero no sabemos las distancias entre los valores. (Si supiéramos la distancia entre los valores ya estaríamos al menos en una escala de intervalo)

Podemos calcular:

- coeficiente de correlación de Spearman
- coeficiente de correlación de Kendall.

Coeficiente de correlación de Spearman

- Lo que tenemos ahora son 2 sucesiones de valores ordinales.
- El coeficiente de Spearman es un caso especial del coeficiente de correlación de Pearson aplicada a dos series de los n primeros números naturales (cuando no hay empates; si hay –muchos- empates hay otra fórmula

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

d_i es la diferencia entre el valor ordinal en X y el valor ordinal en Y del sujeto *i*

Coeficiente de correlación de Spearman

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)}$$

d_i es la diferencia entre el valor ordinal en X y el valor ordinal en Y del sujeto *i*

- Se encuentra acotado, como el coeficiente de Pearson entre -1 y +1.
- Un coeficiente de Spearman de +1 quiere decir que el que es primero en X es primero en Y, el que es segundo en X es segundo en I, etc
  Un coeficiente de Sperman de -1 quiere decir que el que es primero en X es último en Y, el segundo en X es el penúltimo en Y, etc.

Gracias

KIBERNUM