Cluster 2020

Ciencia de Datos en Ingeniería Industrial

clase_01

Análisis exploratorio de datos. Descripción estadística.

AI & Art




Obra de Robbie Barrat, artista. Imágenes creadas por una Generative Adversarial Network.

https://robbiebarrat.github.io

agenda_clase_01

- Boxplot
- Outliers utilizando quantiles
- Correlaciòn Lineal (Pearson)
- EDA Subtes (continuaciòn)
- EDA GooglePlay

Quantiles

Los cuantiles suelen usarse como límites entre los grupos que dividen la distribución de **una** variable aleatoria en partes iguales; entendidas estas como intervalos que comprenden la misma proporción de valores.

Los mas populares son:

- Cuartiles, dividen la distribuciòn en 4 partes iguales (0.25, 0.5, 0.75)
- Quintiles, dividen la dist. en 5 partes iguales (0.2, 0.4, 0.6, 0.8)
- Deciles, dividen la dist. en 10 partes iguales (0.1,0.2…...0.9)
- Percentiles, dividen la dist. en 100 partes iguales (0.01…….0.99)

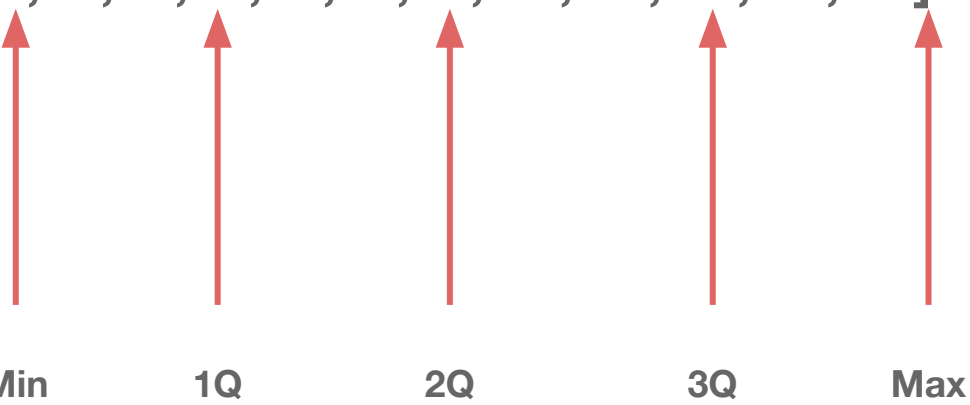
Quantiles, Cuartil

Datos Originales de una variable aleatoria

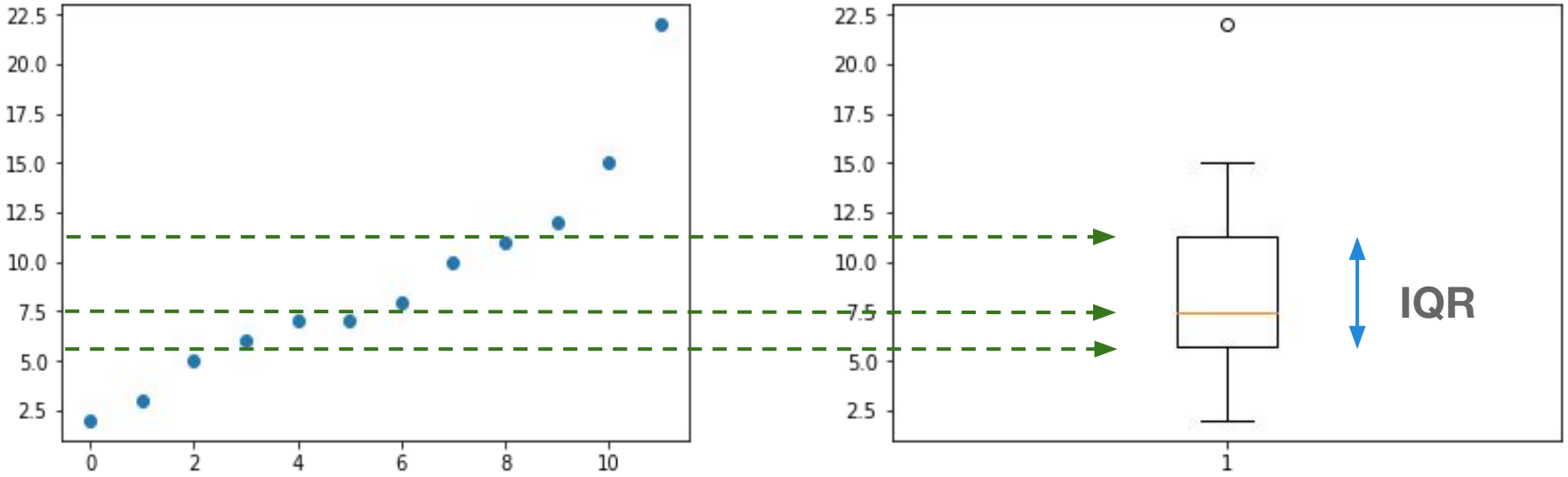
[15, 7, 3, 22, 10, 8, 6, 7, 2, 11, 5, 12]

Datos ordenados

[2, 3, 5, 6, 7, 7, 8, 10, 11, 12, 15, 22]



Boxplot



En este caso por ejemplo tenemos una variable/feature que se mide en un lapso de 11 segundos. Queremos entender cómo se distribuyen los valores de la variable en cuestión.

Cuantiles y Boxplots

En otras palabras, si ordenamos los datos de menor a mayor:

- El 25% de los datos será menor al 1er cuartil
- El 50% de los datos será menor al 2do cuartil (mediana)
- El 75% de los datos será menor al 3er cuartil
- Los valores que esten sobre el percentil 0.01 y 0.99 podrian considerarse outliers.

Mean, median & outliers

Mean, median & outliers

Filtrar por Cuantiles

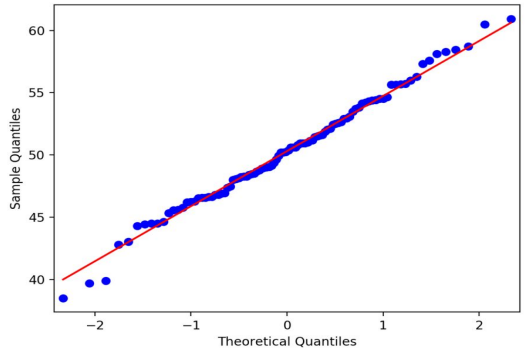
Muchas veces, con el fin de quitar outliers de la distribuciòn de datos que deseamos analizar, lo que podemos realizar es:

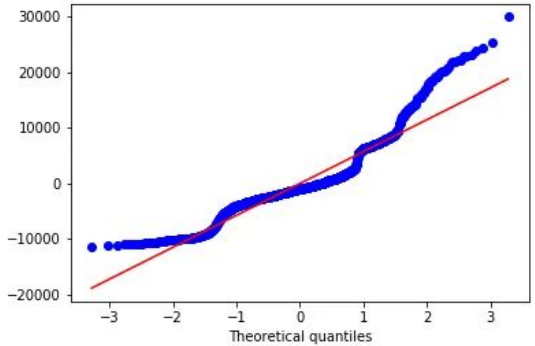
- Quitar todos los datos que estén por encima del Percentil 99
- Quitar todos los datos que estén por debajo del Percentil 1
- Quitar todos los datos que estén por fuera del 1.5 * IQR (Inter Quartile Range).

Cuidado! Quitar datos del dataset dependerá de cada caso, es importante entender las consecuencias de quitar instancias consideradas anomalías.

QQ-Plot (quantile - quantile plot)

Un gráfico Cuantil-Cuantil permite observar cuán cerca está la distribución de un conjunto de datos a alguna distribución ideal ó comparar la distribución de dos conjuntos de datos.





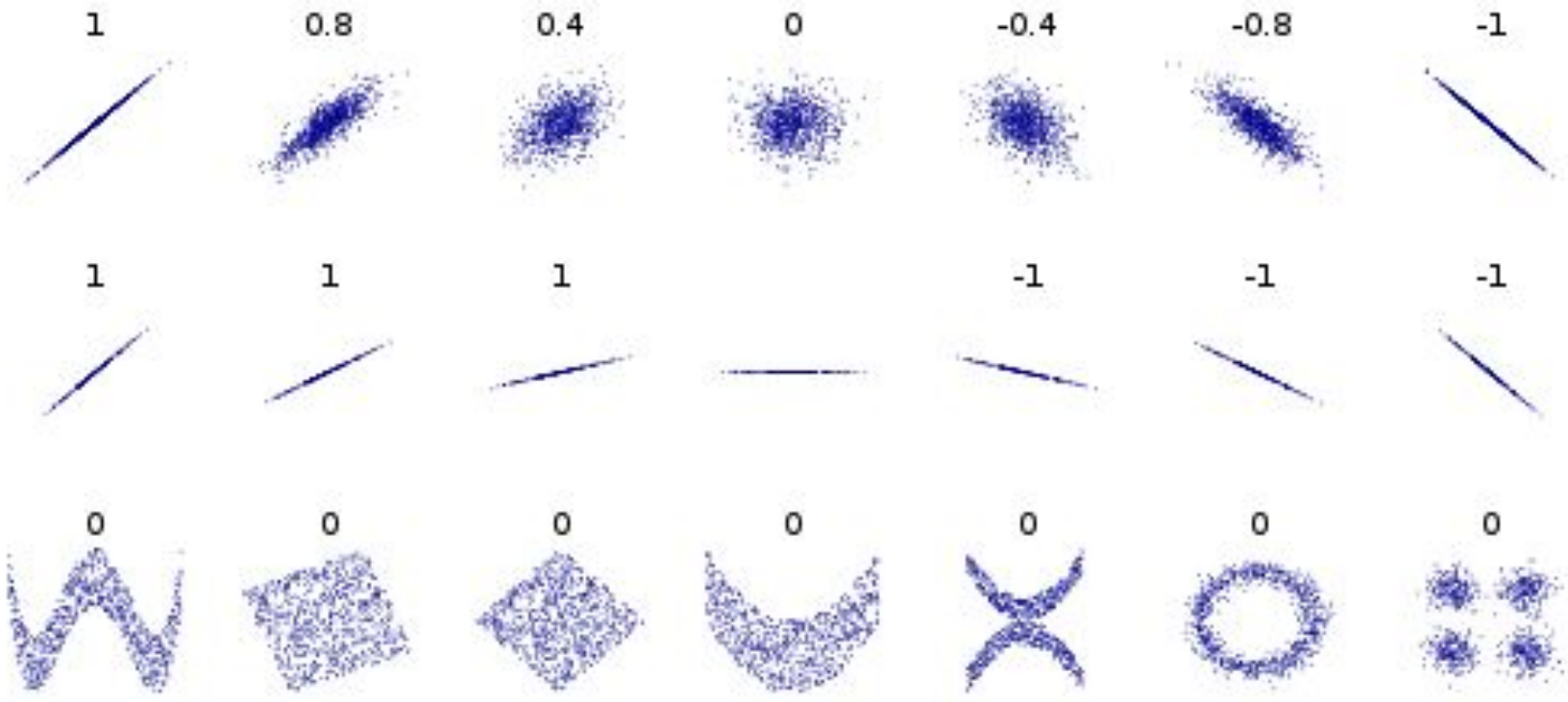
A la izquierda se muestra como la distribución de la muestra (azul) se aproxima mucho a la distribución teórica de una normal gaussiana (rojo). A la derecha se observa que la distribución de la muestra no se aproxima lo suficiente a una gaussiana.

Correlación lineal (Pearson)

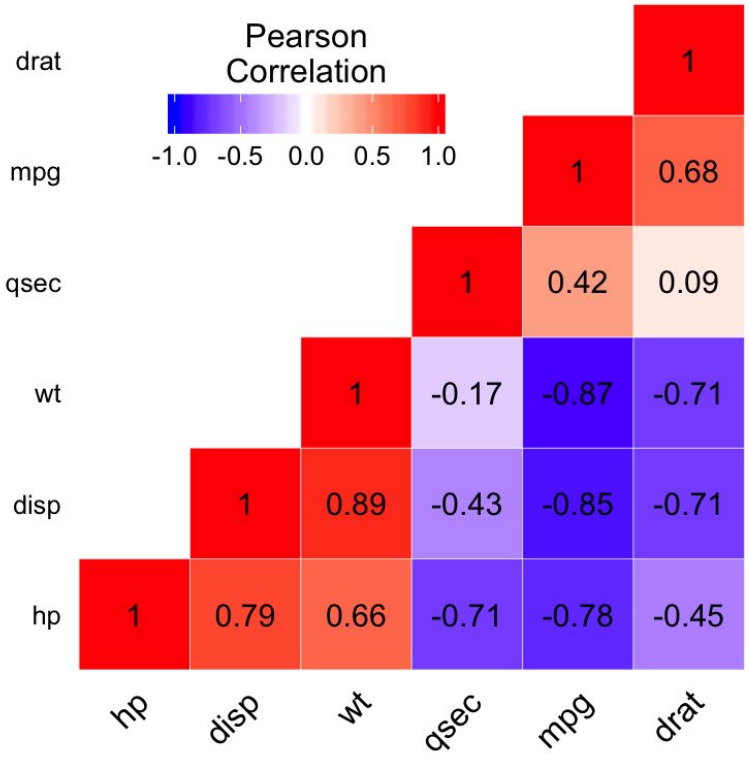
Es una forma de medir cuán cercanas están dos variables (features) a tener una relación lineal entre ellas.

$$r = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\left[\sum_i^n (x_i - \overline{x})^2 (y_i - \overline{y})^2\right]^{1/2}}$$

Correlaciòn lineal (Pearson)



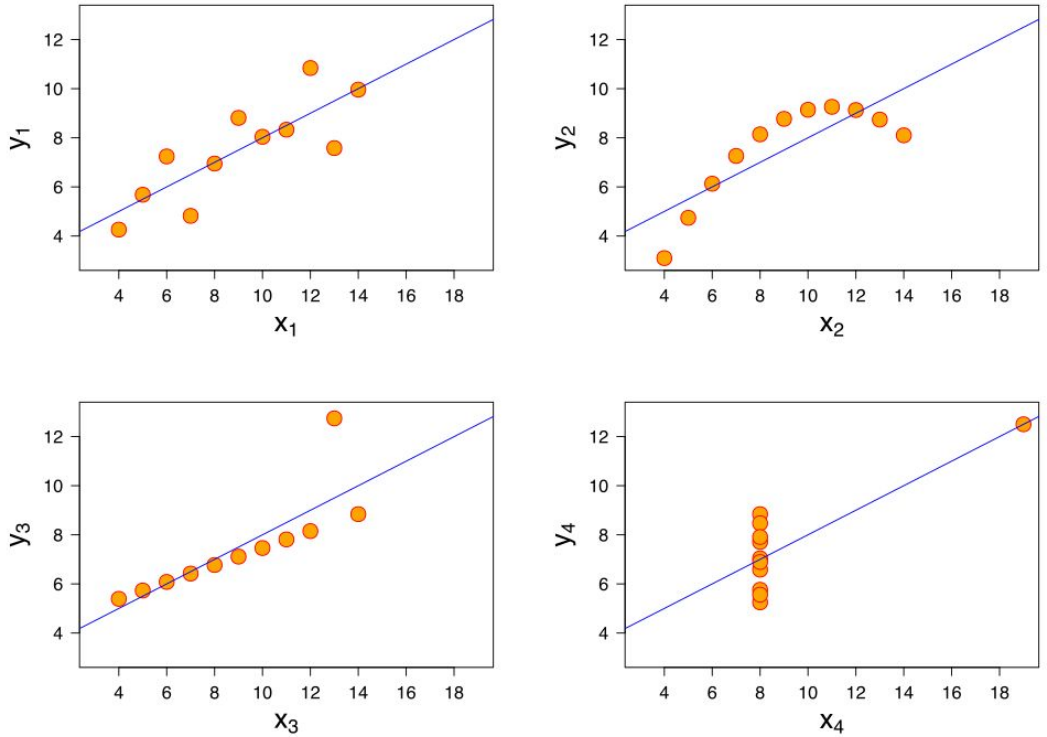
Correlación pairwise entre variables



En el ejemplo tenemos 6 variables/features. Podemos calcular la correlación lineal de Pearson par-a-par y visualizarla con un heatmap.

Atención: la correlación de Pearson **sólo** mide relación lineal entre variables. Que no exista correlación lineal no quiere decir que no exista relación alguna. Puede existir relación no lineal.

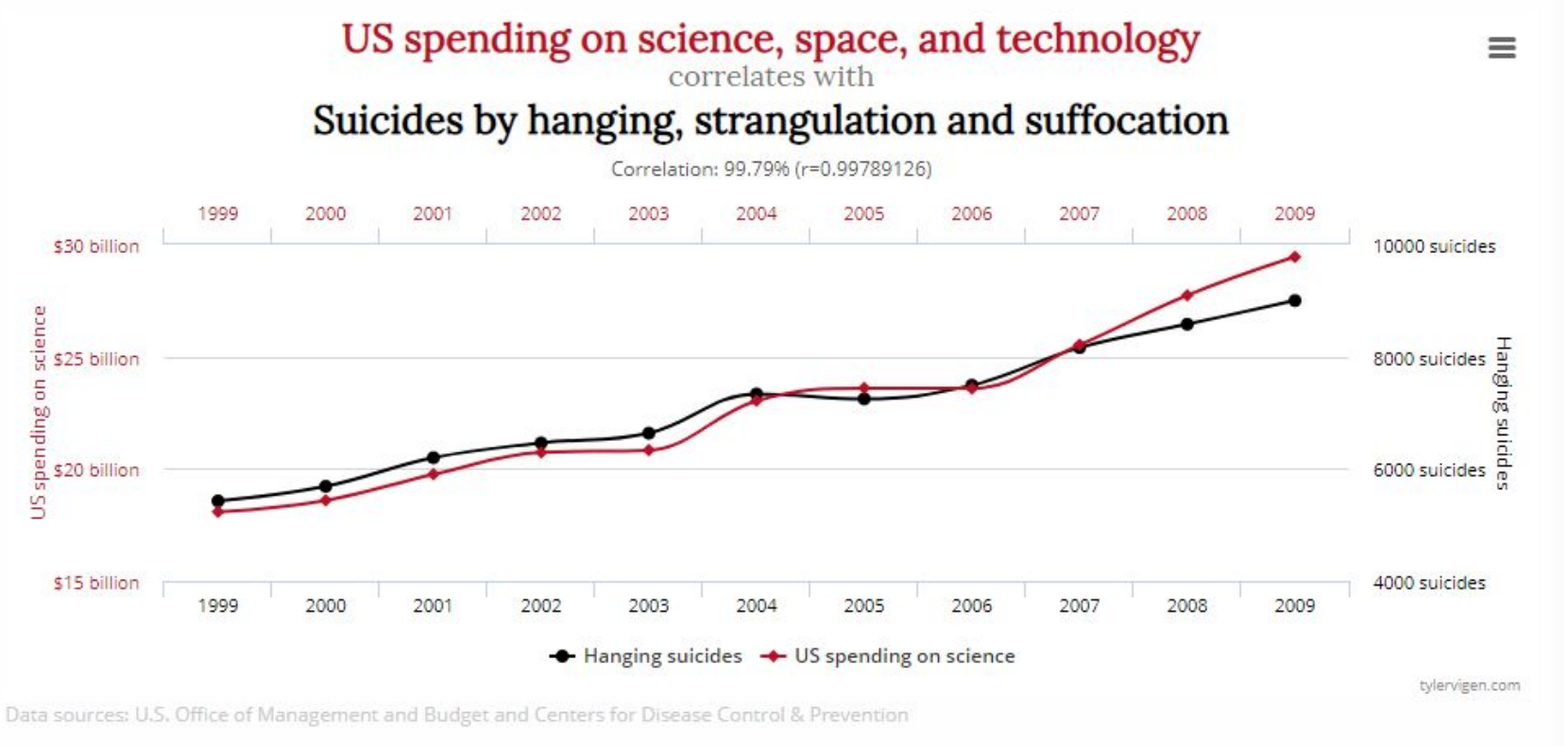
Correlaciòn lineal: trampas



Los 4 datasets tienen las mismas estadísticas descriptivas, sin embargo se ven muy distintos cuando se visualizan:

Media X = 9
Rxy = 0.81

Correlation is not causation



A agarrar la PyLA