# Cluster 2018

## Ciencia de Datos en Ingeniería Industrial

## clase_01

Análisis exploratorio de datos. Descripción estadística.

# agenda_clase_01

- Boxplot
- Outliers utilizando quantiles
- Correlaciòn Lineal (Pearson)
- P-values, tests estadísticos
- EDA Subtes (continuaciòn)
- EDA House Prices
- 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)

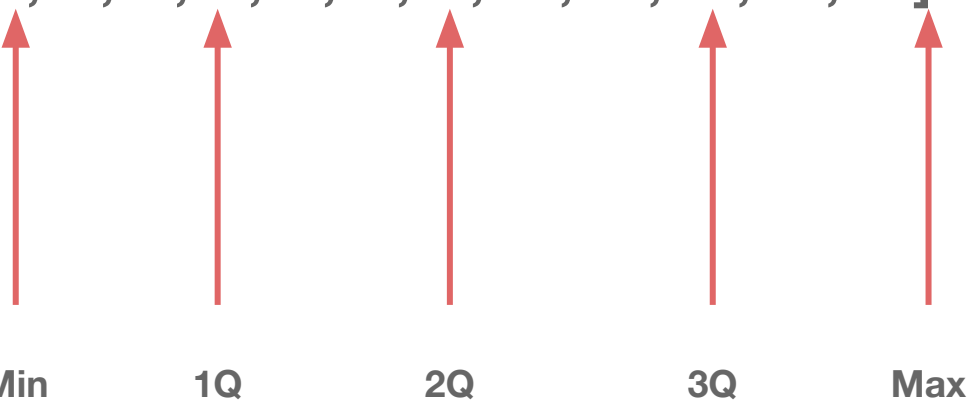
# Quantiles, Cuartil

**Datos Originales**

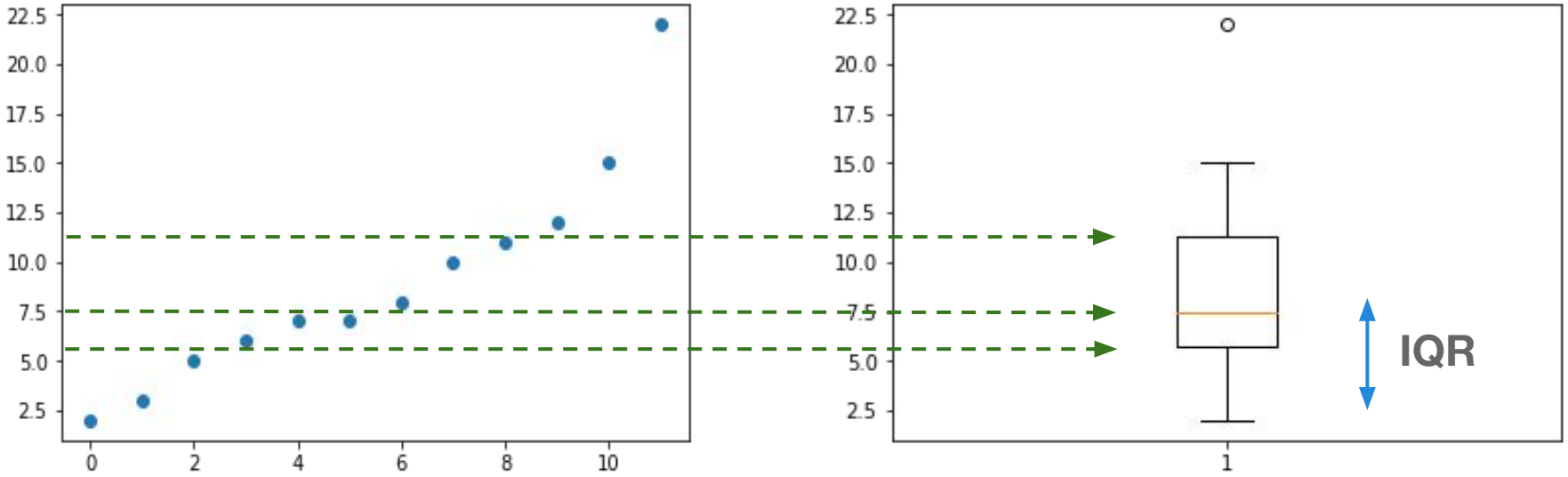
[15, 7, 3, 22, 10, 8, 6, 7, 2, 11, 5, 12]

**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

# Mean, median & outliers

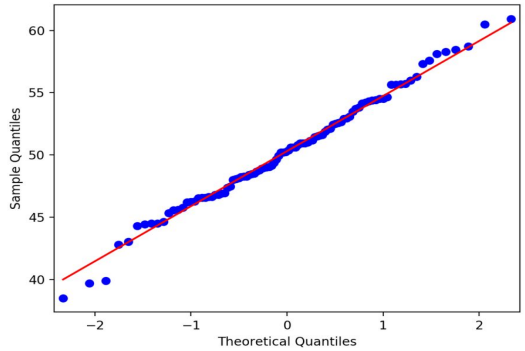
# Filtrar por Cuantiles

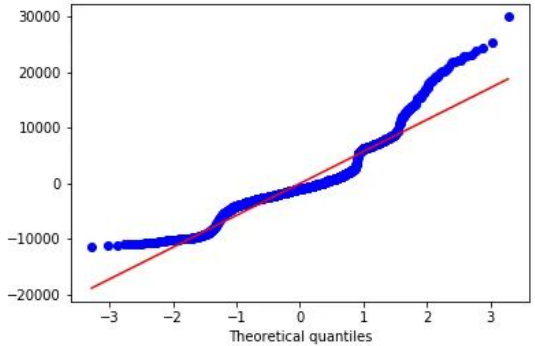
Muchas veces, con el fin de quitar outliers de la distribuciòn de datos que deseamos analizar, lo que podemos realizar es:

- Quitar todos los datos que estén por encima del Percentil 90
- Quitar todos los datos que estén por debajo del Percentil 10
- Quitar todos los datos que estén por fuera del 1.5 * IQR (Inter Quartile Range).

# 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.



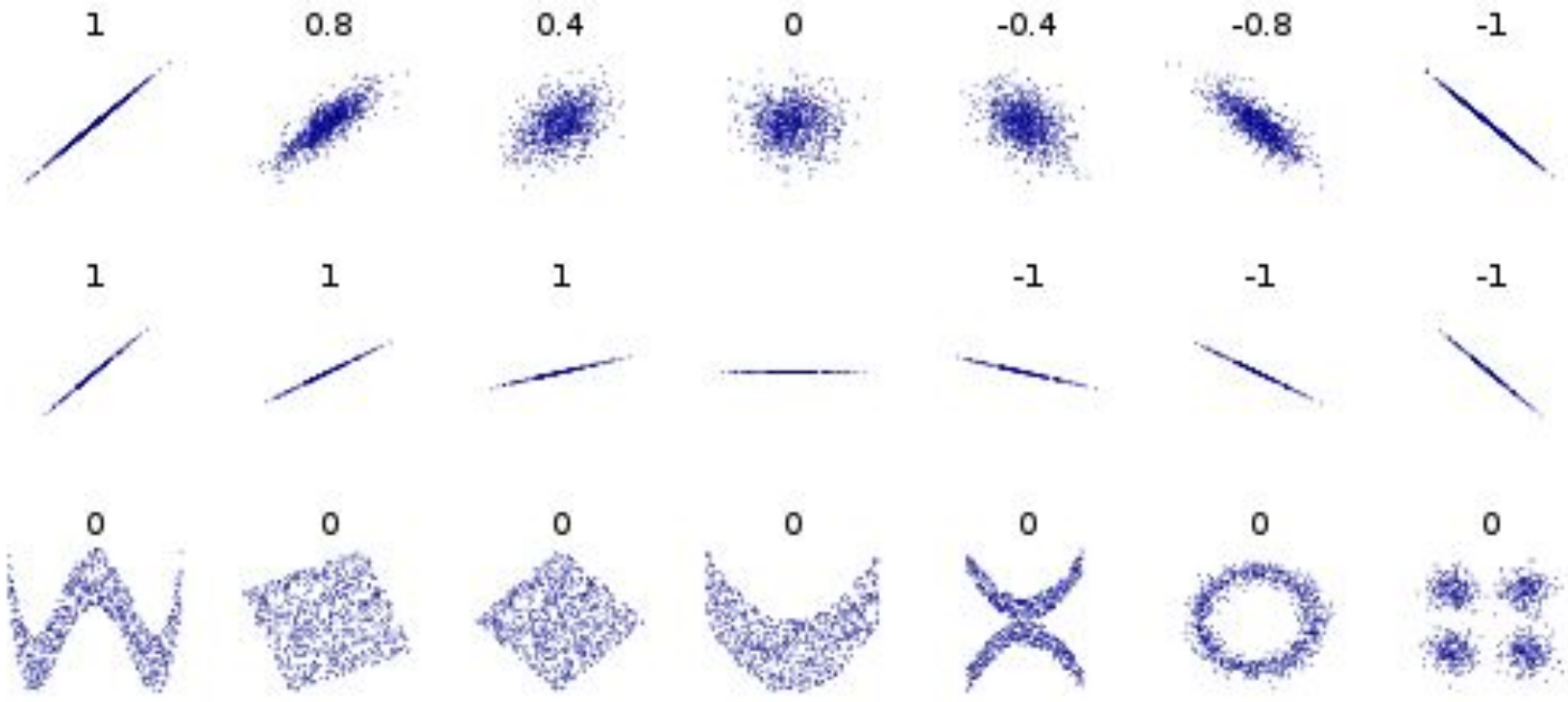

A la izquierda se muestra como la distribución de la muestra (azul) se aproxima mucho a la distribución teórica de una normal gaussiana (rojo). A la derecha se observa que la distribución de la muestra no se aproxima lo suficiente a una gaussiana.

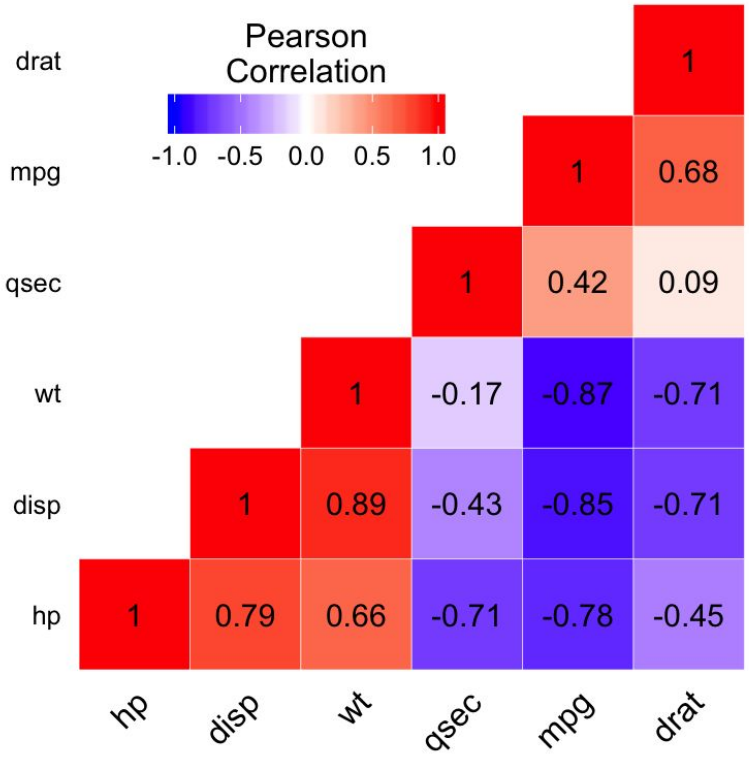
# Correlación lineal (Pearson)

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)

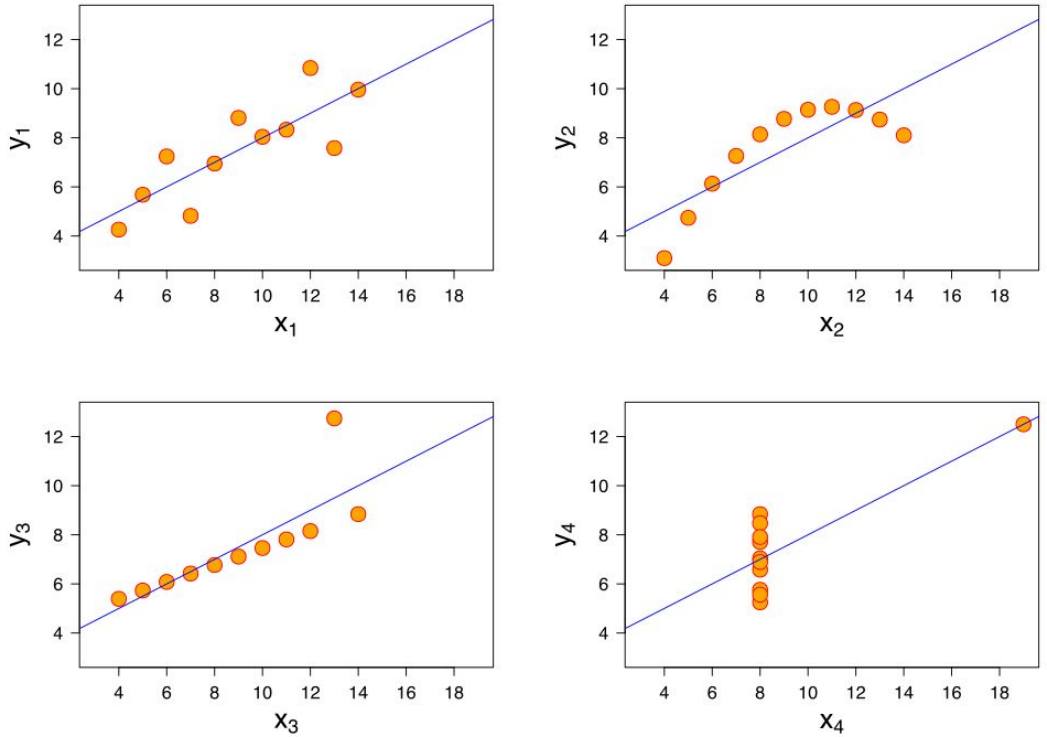





En el ejemplo tenemos 6 variables/features. Podemos calcular la correlación lineal de Pearson par-a-par y visualizarla con un heatmap.

**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.

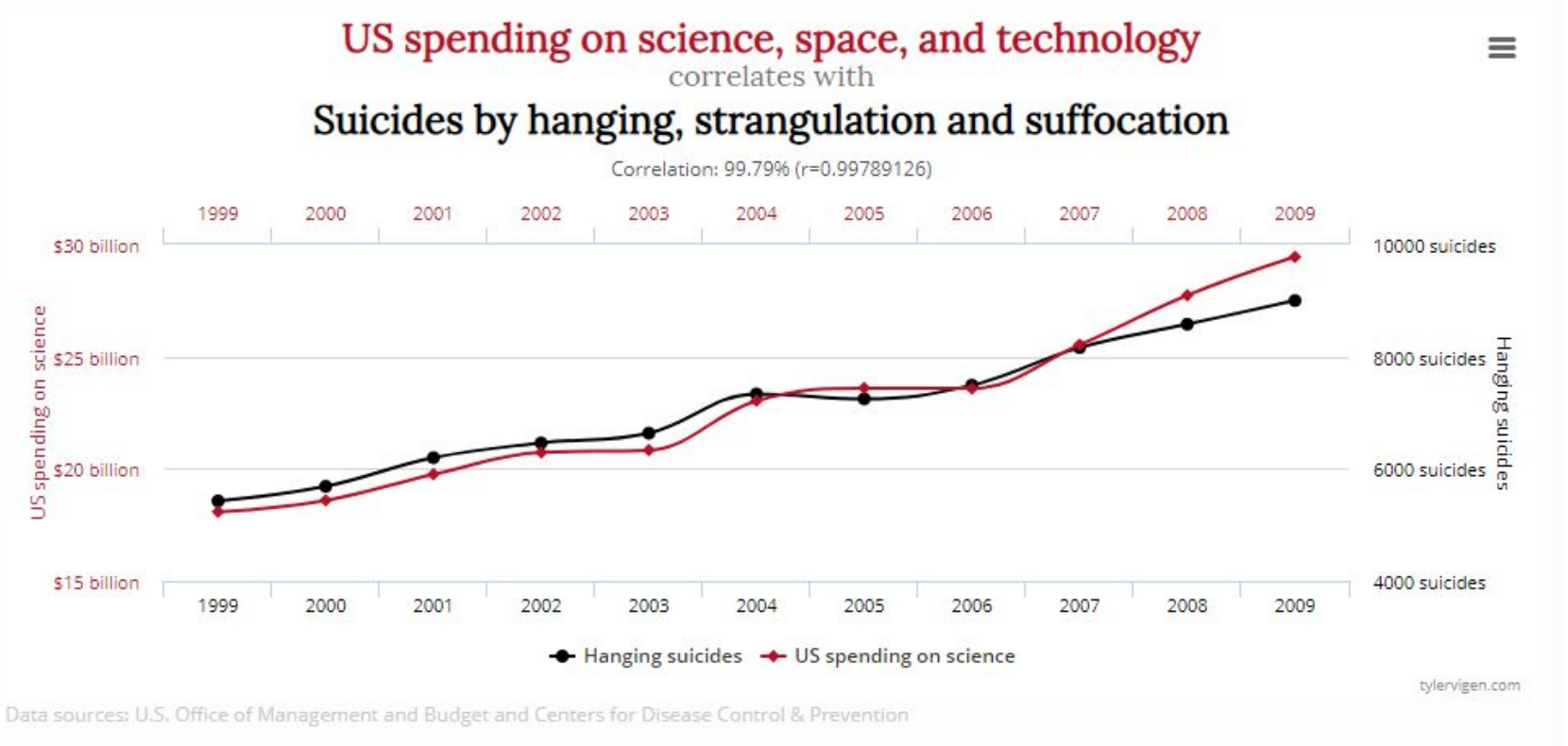
# Correlaciòn lineal: trampas




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



# P-values, tests estadísticos

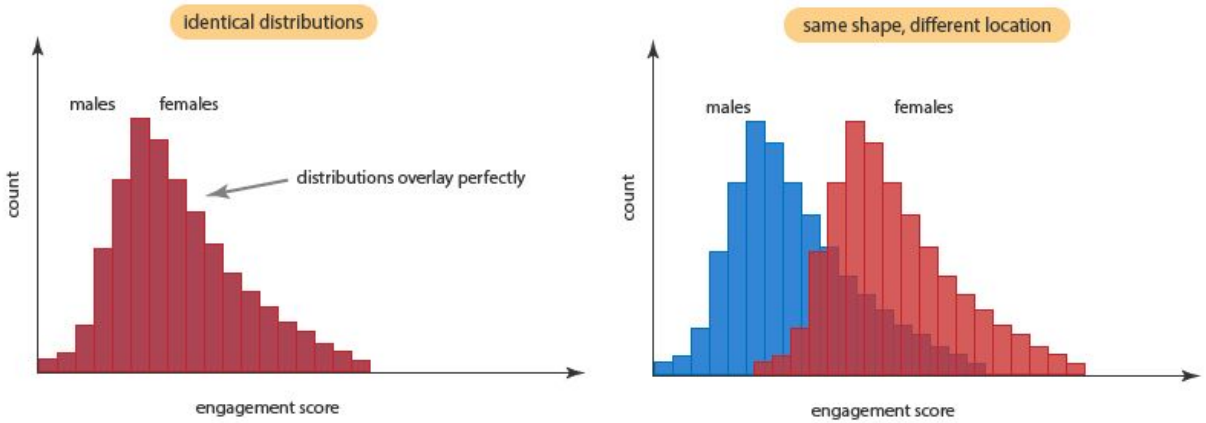
Existen muchos casos donde se quiera asegurar con evidencia que existen diferencias entre dos grupos de mediciones (caso muy común en la ciencia o en estrategias de marketing como un AB test).

**Ejemplo 01**: Se quiere determinar que una campaña de marketing es significativamente más efectiva que otra (llamada control) respecto a la captación de usuarios (AB test).

**Ejemplo 02**: Se quiere mostrar que una droga es efectiva y muestra una mejora significativa en un tratamiento contra células tumorales, midiendo la superficie de tejido dañado en los pacientes con el tratamiento regular y en los pacientes con la droga de prueba.

# P-values, tests estadísticos: Mann Whitney U Test

Entonces si tenemos dos conjuntos de datos, y en cada conjunto medimos la misma variable, vamos a querer saber si existe una diferencia significativa en el valor de la variable entre los dos grupos -> podemos usar el **Mann Whitney U test**



# P-values, tests estadísticos: Mann Whitney U Test

El **Mann Whitney U test** tiene como input dos conjuntos de datos. Calcula el estadístico correspondiente y arroja un valor de “p-value”.

- Queremos saber si la media de cada grupo es significativamente distinta o no. Como todo test estadístico, existe una hipótesis nula y una alternativa.
- Si el P valor obtenido es menor a 0.05, estamos en condiciones de decir que tenemos suficiente evidencia para rechazar la hipótesis nula y afirmar que ambas muestras vienen de poblaciones con distintas medias.
- El P valor es una medida de probabilidad que indica cual es la chance de que las hipótesis nula se rechaze solo por azar. Obviamente cuanto menor sea el P valor mas fuerte es la evidencia para rechazar la H0.
- Por convención, todo p valor menor a 0.05 se dice que es estadisticamente significativo.

H0 = Las medias de las muestras son iguales, provienen de la misma distribución.
H1 = Una de las medias excede a la otra

**P value < 0.05 --> rechazamos H0**

# A agarrar la PyLA