DATA SCIENCE

Federico Baiocco
baioccofede@gmail.com
3512075440

Clase 6 - Agenda

SEGUIMOS CON PANDAS

- Correlación
- Pandas

¿ Dudas de la clase pasada ?

Correlación, casualidad e independencia estadística

Estos 3 conceptos tratan sobre la relación entre 2 variables aleatorias y son muy fáciles de confundir.

Correlación, Causalidad e Independencia estadística

La correlación es una medida estadística que expresa hasta qué punto dos variables están relacionadas linealmente (esto es, cambian conjuntamente a una tasa constante).
Sin embargo, observar que dos variables se mueven conjuntamente no significa necesariamente que una variable sea la causa de la otra. Por eso solemos decir que "la correlación no implica causalidad".
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.

http://tylervigen.com/spurious-correlations

Es importante saber que la correlación no nos informa sobre causas y efectos!!!

¿ Cómo se mide la correlación?

Para medir la correlación lineal entre 2 variables, es muy común utilizar el coeficiente de correlación de Pearson.

Describimos la correlación mediante una medida sin unidades llamada coeficiente de correlación, que va desde -1 a +1 y se indica mediante la letra r.

- Cuanto más se aproxima r a cero, más débil es la relación lineal.
- Los valores de r positivos indican una correlación positiva, en la que los valores de ambas variables tienden a incrementarse juntos.
- Los valores de r negativos indican una correlación negativa, en la que los valores de una variable tienden a incrementarse mientras que los valores de la otra variable descienden.

Coeficiente de correlación de Pearson

La fórmula del coeficiente de correlación de Pearson es la siguiente:

$$r = \frac{\sum z_X z_Y}{N}$$

Donde:

- x es la variable número uno
- y es la variable número dos
- Zx es la desviación estándar de la variable uno
- Zy es la desviación estándar de la variable dos
- N es es número de datos.

Abrimos notebook clase 6

Kahoot!