

Correlaciones y Regresión Lineal Simple

**Análisis de Datos con
Python**

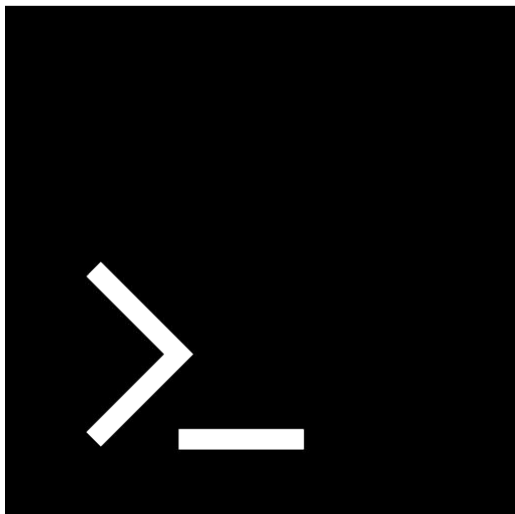
Eduardo Selim Martínez Mayorga



- Comprender el concepto de correlación entre variables y por qué es relevante.
- Comprender el significado del coeficiente de correlación e interpretarlo.
- Hacer matrices de correlaciones y a graficarlas usando *heatmaps*.
- Hacer gráficas de dispersión e interpretarlas.
- Aprender el concepto de Gráficas de Pares.
- Aprender el concepto de Regresión Lineal Simple y cómo funciona el proceso de entrenamiento e interpretación.



- Anotaciones
- Gráficas de Barras
- Moda
- Tablas de contingencia
- Múltiples gráficas
- Gráficas de caja/violín



¡No olvides hacer pull del repo!

El material de la sesión se encuentra ahí.

```
git pull origin master
```



1. ¿Qué significa que dos variables estén correlacionadas positivamente?
 - a. Que el aumento de valores en una de ellas está relacionado con la disminución de valores en la otra
 - b. Que un cambio en una variable causa un cambio en la otra variable
 - c. Que el aumento de valores en una de ellas está relacionado con el aumento de valores en la otra
 - d. Que las dos variables son totalmente independientes
 - e. Que la disminución de valores en una de ellas está relacionado con el aumento de valores en la otra



2. ¿Qué significa un valor de -1 en un coeficiente de correlación de Pearson?
 - a. Indica una correlación positiva perfecta entre nuestras variables: si una disminuye, la otra disminuye en la misma proporción
 - b. Indica una correlación negativa perfecta entre nuestras variables: si una aumenta, la otra disminuye en la misma proporción
 - c. Indica una correlación positiva perfecta entre nuestras variables: si una aumenta, la otra aumenta en la misma proporción
 - d. Indica una correlación negativa perfecta entre nuestras variables: si una aumenta, la otra aumenta en la misma proporción
 - e. Indica una correlación negativa perfecta entre nuestras variables: si una disminuye, la otra disminuye en la misma proporción



3. ¿Qué podemos concluir si obtenemos un coeficiente de correlación de Pearson con valor 0?
- a. Que el cambio en una de las variables causa un cambio en la otra
 - b. Que no existe ningún tipo de relación entre las dos variables
 - c. Que existe una correlación de tipo positivo perfecto
 - d. Que no existe una relación polinomial entre las dos variables
 - e. Que no existe una relación lineal entre las dos variables



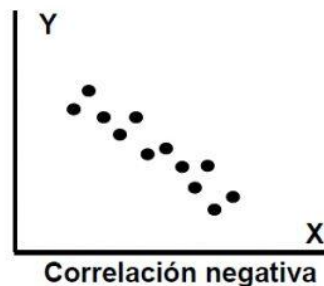
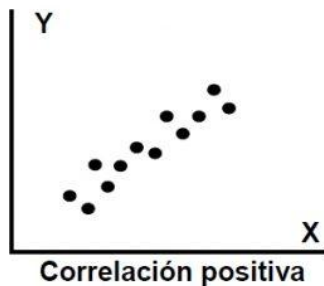
3. ¿Qué es una variable dependiente?
 - a. Es la variable que estamos intentando predecir usando un modelo de Regresión Lineal Simple
 - b. Es la variable que aumenta si todas las demás variables aumentan
 - c. Es la variable que causa cambios en las demás variables
 - d. Es la variable que usamos para predecir otra variable en un modelo de Regresión Lineal Simple
 - e. Es la variable que aumenta si todas las demás disminuyen



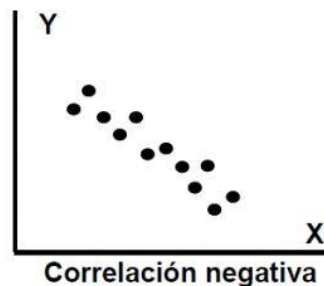
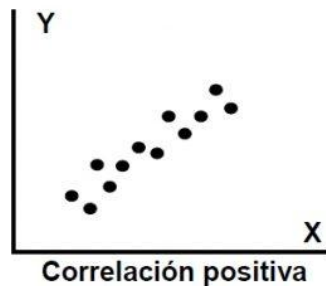
4. ¿Qué es una variable dependiente?
- a. Es la variable que estamos intentando predecir usando un modelo de Regresión Lineal Simple
 - b. Es la variable que aumenta si todas las demás variables aumentan
 - c. Es la variable que causa cambios en las demás variables
 - d. Es la variable que usamos para predecir otra variable en un modelo de Regresión Lineal Simple
 - e. Es la variable que aumenta si todas las demás disminuyen



3. ¿Qué explica el valor del coeficiente de determinación (R^2)?
 - a. Si la correlación entre dos variables es positiva o negativa
 - b. Qué tanta variación de la variable x podemos explicar utilizando la variable y
 - c. Si el coeficiente de correlación de Pearson es el algoritmo adecuado para nuestro problema
 - d. Qué tanta variación de la variable y podemos explicar utilizando la variable x
 - e. Qué tanta variación de la variable y podemos explicar utilizando la misma variable y



- Dos variables están **correlacionadas positivamente** si el aumento de valores en una de ellas está relacionado con el aumento de valores en la otra.
- También están **correlacionadas positivamente** si la disminución de valores en una de ellas está relacionado con la disminución de valores en la otra.
- Decimos que está **correlacionadas negativamente** si el aumento en los valores de una está relacionado a la disminución de los valores en la otra, y viceversa.



- Si dos variables están correlacionadas podemos intuir que existe cierto nivel de **dependencia directa o indirecta** entre ellas.

Correlación no implica causalidad

- El hecho de que nuestras variables estén correlacionadas no implica que el cambio de valores en una cause el cambio de valores en la otra.

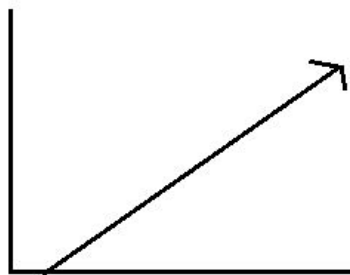
¡La relación podría ser un efecto de la aleatoriedad!

Coeficiente de correlación de Pearson

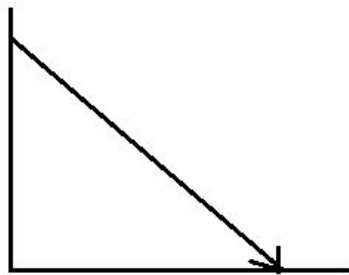


- Sirve para calcular la correlación entre dos variables numéricas, ya que depende de la desviación estándar de nuestras variables
- Es un valor entre -1 y 1.
- Un valor de -1 indica una correlación negativa perfecta entre nuestras variables,
- Un valor de 1 indica una correlación positiva perfecta entre las variables
- Un valor de 0 indica que no hay ninguna correlación entre las variables

Coeficiente de correlación de Pearson



Positive Linear Relationship



Negative Linear Relationship

Para que el coeficiente de correlación de Pearson sea capaz de encontrar la relación entre dos variables, la relación tiene que ser **lineal**.

[Ve al Ejemplo 1](#)

[Ve al Reto 1](#)

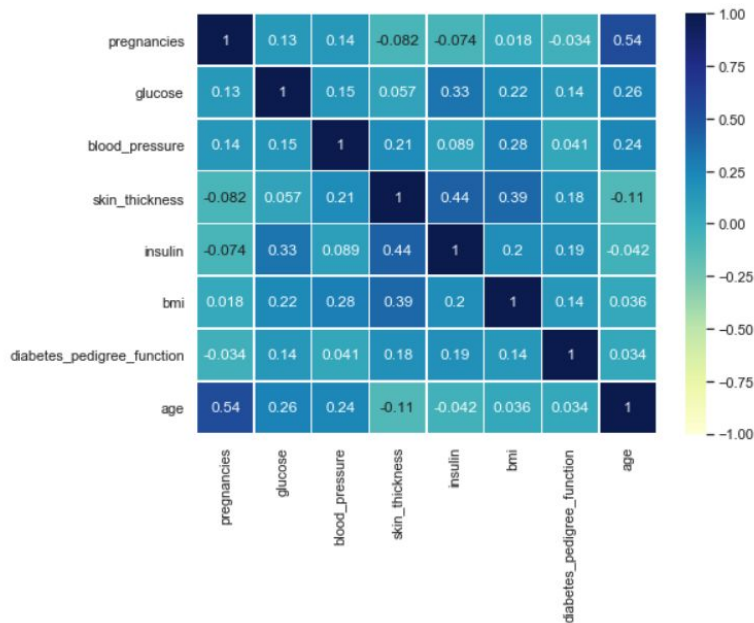
Matriz de correlación



- Muestra la correlación entre las variables de un dataframe.
- Cada celda muestra la intersección entre la columna y fila correspondiente.
- La diagonal tiene puros unos.

Municipio	Esparta	Jutiapa	La Masica	San Francisco	Tela	Arizona	Trujillo
Esparta	1	0.44670471	0.48570799	0.58155892	0.51878827	0.74529879	0.74704147
Jutiapa	0.44670471	1	0.4786205	0.32276876	0.09881743	0.16617166	0.80429634
La Masica	0.48570799	0.4786205	1	0.82319079	0.72245263	0.67522982	0.73169792
San Francisco	0.58155892	0.32276876	0.82319079	1	0.92416851	0.85814101	0.7031936
Tela	0.51878827	0.09881743	0.72245263	0.92416851	1	0.81531247	0.53862217
Arizona	0.74529879	0.16617166	0.67522982	0.85814101	0.81531247	1	0.61964299
Trujillo	0.74704147	0.80429634	0.73169792	0.7031936	0.53862217	0.61964299	1
Balfate	0.05868548	0.57765125	-0.06945436	-0.13968811	-0.31713462	-0.15884576	0.33289301
Iriona	-0.00852878	-0.0641272	-0.36171805	-0.25364428	-0.28336794	-0.23563747	-0.1191828
Limon	0.52306042	0.4893457	0.54705601	0.53130628	0.39536501	0.52469245	0.70188418
Saba	0.46494727	0.45910279	0.8631286	0.91691686	0.84797655	0.72306725	0.76491481
Santa Fe	0.15231857	-0.03639994	-0.34255118	-0.00493439	0.01970385	-0.00264652	-0.07523237
Santa Rosa de	0.88783618	0.50863078	0.56418688	0.59134663	0.51193188	0.73057676	0.86526373
Sonaguera	0.28669057	0.9336425	0.56919983	0.38959167	0.19778082	0.12436043	0.77327064

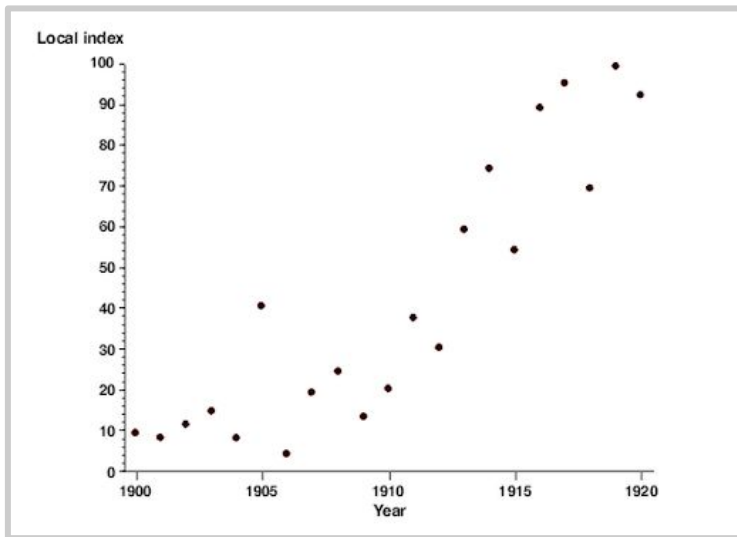
Mapas de calor



- La barra de la derecha muestra el rango y el significado por color.
- El coeficiente de correlación es utilizado principalmente con variables numéricas.

[Ve al Ejemplo 2](#)

[Ve al Reto 2](#)

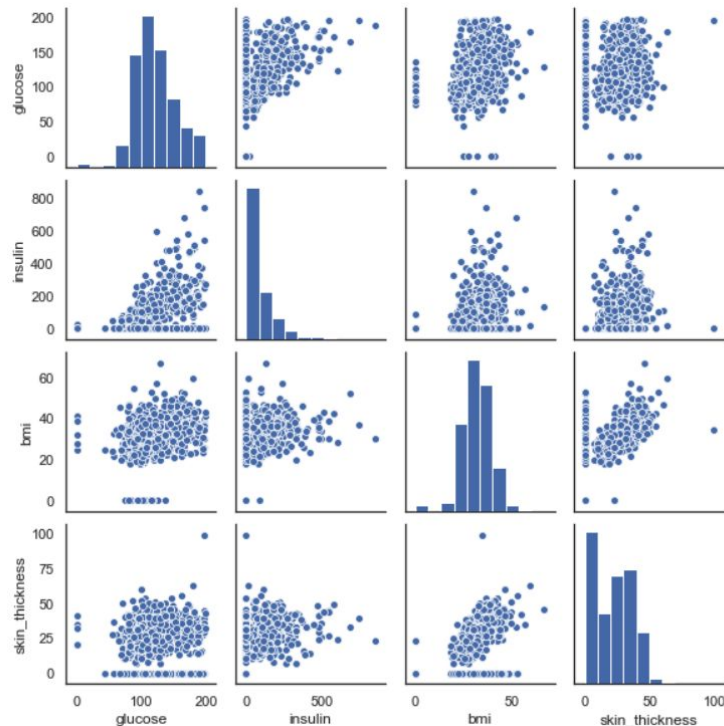


- Grafican una de las variables en el eje x y la otra variable en el eje y de un plano cartesiano.
- Cada muestra es un punto en el plano que tiene su respectivo valor para x y para y
- Muestran la relación entre las variables x y y

[Ve al Ejemplo 3](#)

[Ve al Reto 3](#)

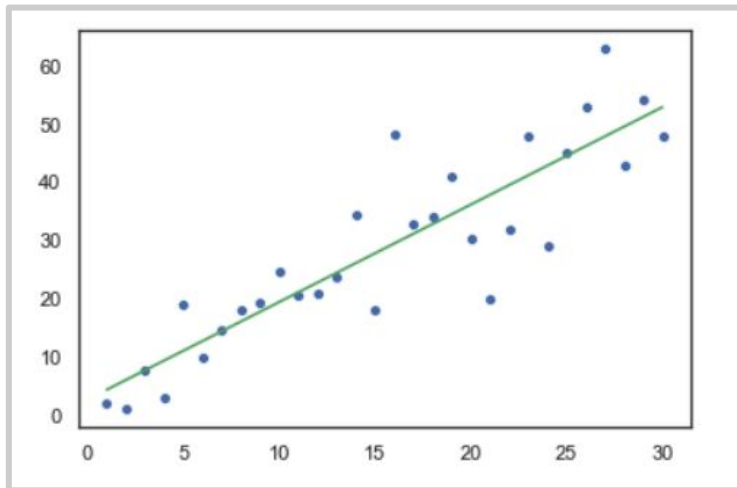
Gráficas de pares



- Visualmente son equivalentes a las matrices de correlación.
- En la diagonal se colocan histogramas.

Ve al Ejemplo 4

Regresión Lineal Simple



Ve al Ejemplo 5

Ve al Reto 4

- El objetivo es trazar una línea que pase por el mayor número de datos.
- Esto se puede analizar usando gráficas de dispersión y por supuesto, conociendo la correlación de nuestras variables.
- Ecuación de la recta:
$$y = mx + b$$
- El coeficiente de determinación (R^2) nos indica el margen de error.
- Esto nos permite generar predicciones, aunque en casi todos los casos nunca son del 100% :(



NO OLVIDES REVISAR TU
POSTWORK Y TU PREWORK



Preguntas

