Creado por:

Isabel Maniega

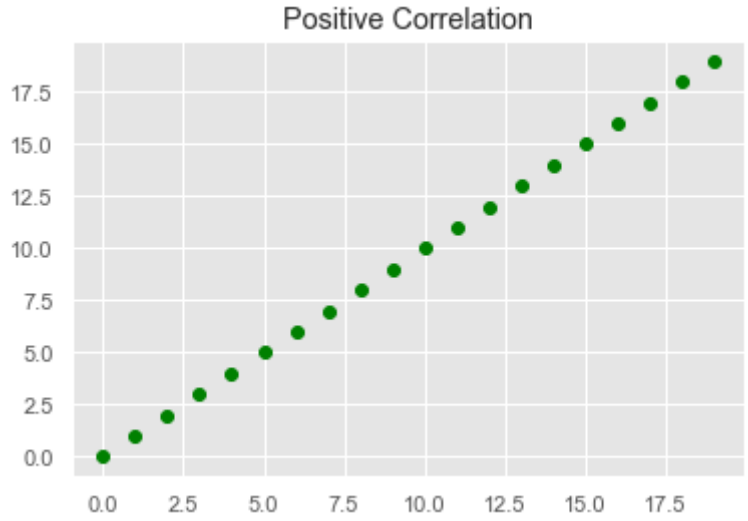
Correlación

La correlación es el análisis estadístico de la relación o dependencia entre dos variables. La correlación nos permite estudiar tanto la fuerza como la dirección de la relación entre dos conjuntos de variables.

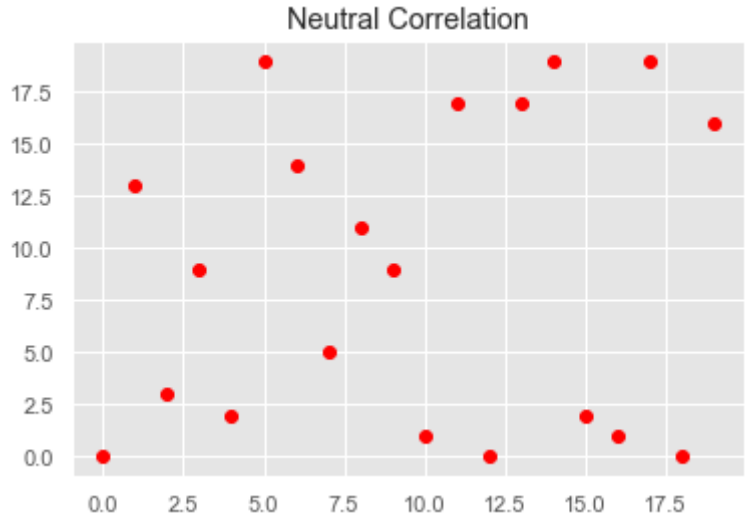
El estudio de la correlación es fundamental en el campo del aprendizaje automático. Por ejemplo, algunos algoritmos no funcionarán correctamente si dos o más variables están estrechamente relacionadas, lo que generalmente se conoce como multicolinealidad . La correlación también es la base del Análisis de Componentes Principales , una técnica de reducción de dimensionalidad lineal que es muy útil en proyectos de aprendizaje automático.

Tipos:

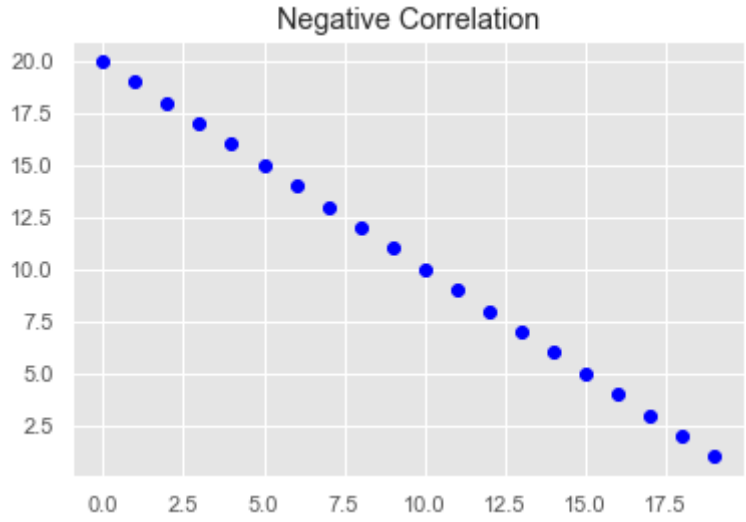
Correlación positiva: se dice que dos variables están correlacionadas positivamente cuando sus valores se mueven en la misma dirección. Por ejemplo, en la imagen a continuación, a medida que aumenta el valor de X, también lo hace el valor de Y a una tasa constante:



Correlación Neutral: No hay relación en el cambio de las variables X e Y. En este caso los valores son completamente aleatorios y no muestran ningún signo de correlación, como se muestra en la siguiente imagen:



Correlación negativa: finalmente, las variables X e Y estarán negativamente correlacionadas cuando sus valores cambien en direcciones opuestas, por lo que aquí, a medida que aumenta el valor de X, el valor de Y disminuye a una tasa constante:



Coeficientes de correlación

Un coeficiente de correlación es un resumen estadístico que mide la fuerza y la dirección con la que se asocian dos variables entre sí.

Una de las ventajas de los coeficientes de correlación es que estiman la correlación entre dos variables de forma estandarizada, por lo que el valor del coeficiente siempre estará en la misma escala, variando de -1,0 a 1,0.

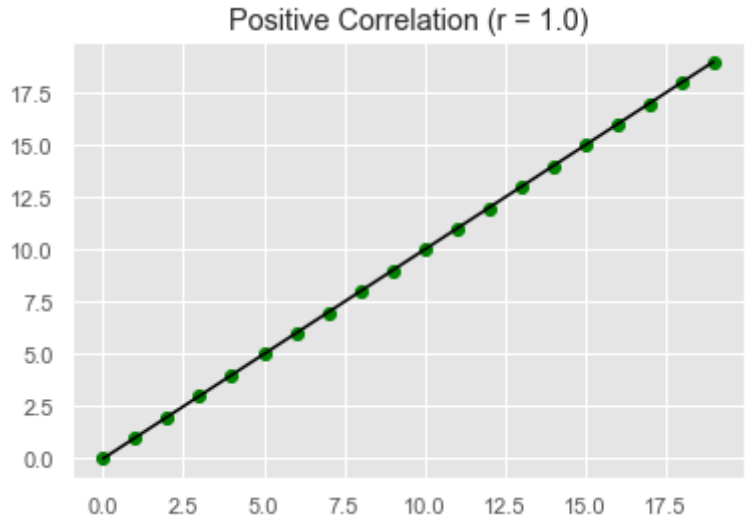
1. Coeficiente de correlación de Pearson

El coeficiente de correlación de Pearson (r) es una puntuación que mide la fuerza de una relación lineal entre dos variables. Se calcula dividiendo la covarianza de las variables X e Y por el producto de la desviación estándar de cada variable, como se muestra en la siguiente fórmula:

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

El coeficiente se basa en dos supuestos. Primero, asume que las variables siguen una distribución normal o gaussiana. Si los datos no se distribuyen normalmente, entonces otros coeficientes pueden ser más confiables.

En segundo lugar, supone que existe una relación lineal entre las dos variables, lo que significa que los cambios en los datos se pueden modelar mediante una función lineal (es decir, sus valores aumentan o disminuyen simultáneamente a una tasa constante). Si la relación entre las dos variables es más cercana a una línea recta, entonces su correlación (lineal) es más fuerte y el valor absoluto del coeficiente de correlación de Pearson es más alto. Por ejemplo, en la siguiente imagen, todos los puntos de datos se pueden modelar perfectamente utilizando una línea recta, lo que da como resultado un coeficiente de correlación igual a 1,0.



Un coeficiente de -1,0 indica una correlación negativa perfecta, mientras que un coeficiente de 1,0 muestra una correlación positiva perfecta. Por el contrario, un coeficiente de 0,0 indica que no existe una correlación lineal entre las variables.

Calculo:

In [1]:

```
experience = [1, 3, 4, 5, 5, 6, 7, 10, 11, 12, 15, 20, 25, 28, 30,35]

salary = [20000, 30000, 40000, 45000, 55000, 60000, 80000, 100000, 130000, 150000,
```

In [2]:

```
import pandas as pd

df = pd.DataFrame({"Experience": experience, "Salary": salary})
df
```

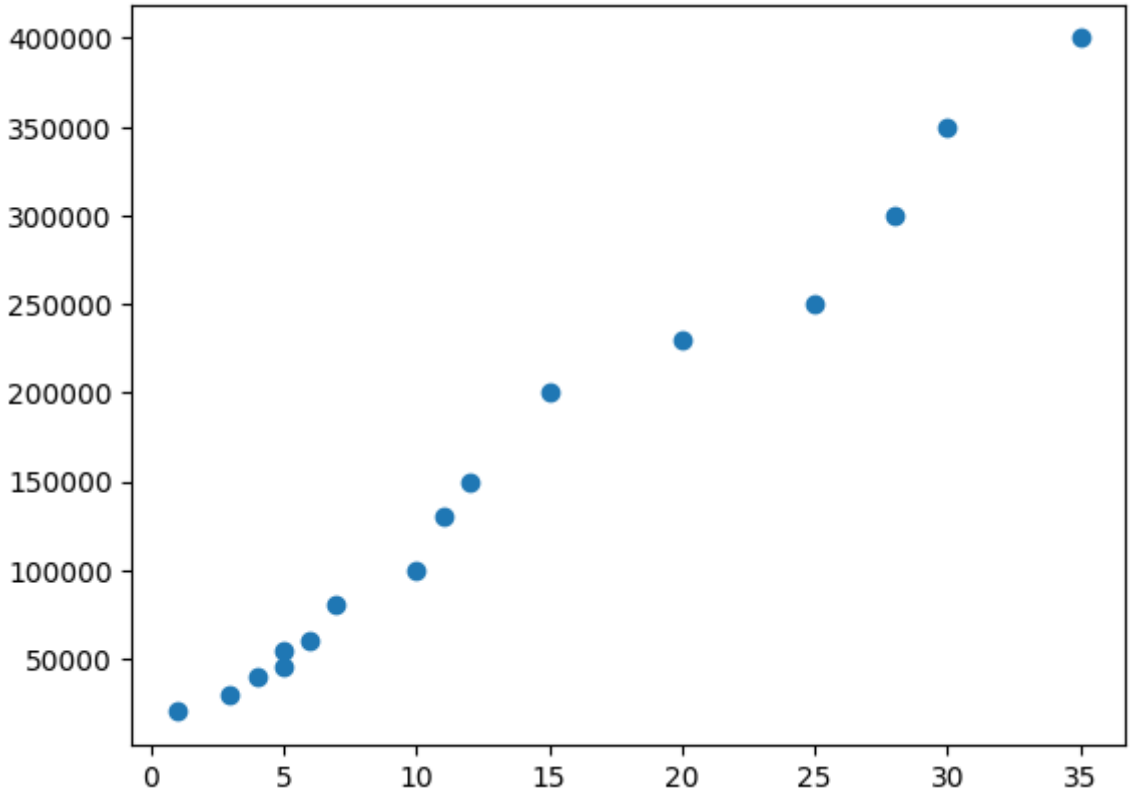
Out[2]:

	Experience	Salary
0	1	20000
1	3	30000
2	4	40000
3	5	45000
4	5	55000
5	6	60000
6	7	80000
7	10	100000
8	11	130000
9	12	150000
10	15	200000
11	20	230000
12	25	250000
13	28	300000
14	30	350000
15	35	400000

In [3]:

```
import matplotlib.pyplot as plt

plt.scatter(df.Experience, df.Salary)
#plt.plot(df.Experience, df.Salary, color='red', linewidth=2)
plt.show()
```



- scipy librería

In [4]:

```
import scipy.stats as stats

corr, _ = stats.pearsonr (experience, salary)
corr
```

Out[4]:

```
0.9929845761480396
```

In [5]:

```
# Otros coeficientes:

spearman_corr, _ = stats.spearmanr(experience, salary)
print("spearman:", spearman_corr)

kendall_corr, _ = stats.kendalltau(experience, salary)
print("Kendall:", kendall_corr)
```

```
spearman: 0.9992644353546791
Kendall: 0.9958246164193105
```

- Numpy librería

In [6]:

```
import numpy as np

np.corrcoef(df.Experience, df.Salary)
```

Out[6]:

```
array([[1.        , 0.99298458],
       [0.99298458, 1.        ]])
```

Una matriz de correlación es una tabla que muestra los coeficientes de correlación entre variables. Cada celda de la tabla muestra la correlación entre dos variables. La diagonal de la matriz incluye los coeficientes entre cada variable y ella misma, que siempre es igual a 1,0. Los demás valores de la matriz representan la correlación entre experiencia y salario. En este caso, como solo estamos calculando correlación para dos variables, los valores son los mismos.

- Pandas librería

In [7]:

```
df['Experience'].corr(df['Salary'])
```

Out[7]:

```
0.9929845761480398
```

In [8]:

```
print(df['Experience'].corr(df['Salary'], method='spearman'))
print(df['Experience'].corr(df['Salary'], method='kendall'))
```

```
0.9992644353546791
0.9958246164193105
```

In [9]:

```
df.corr()
```

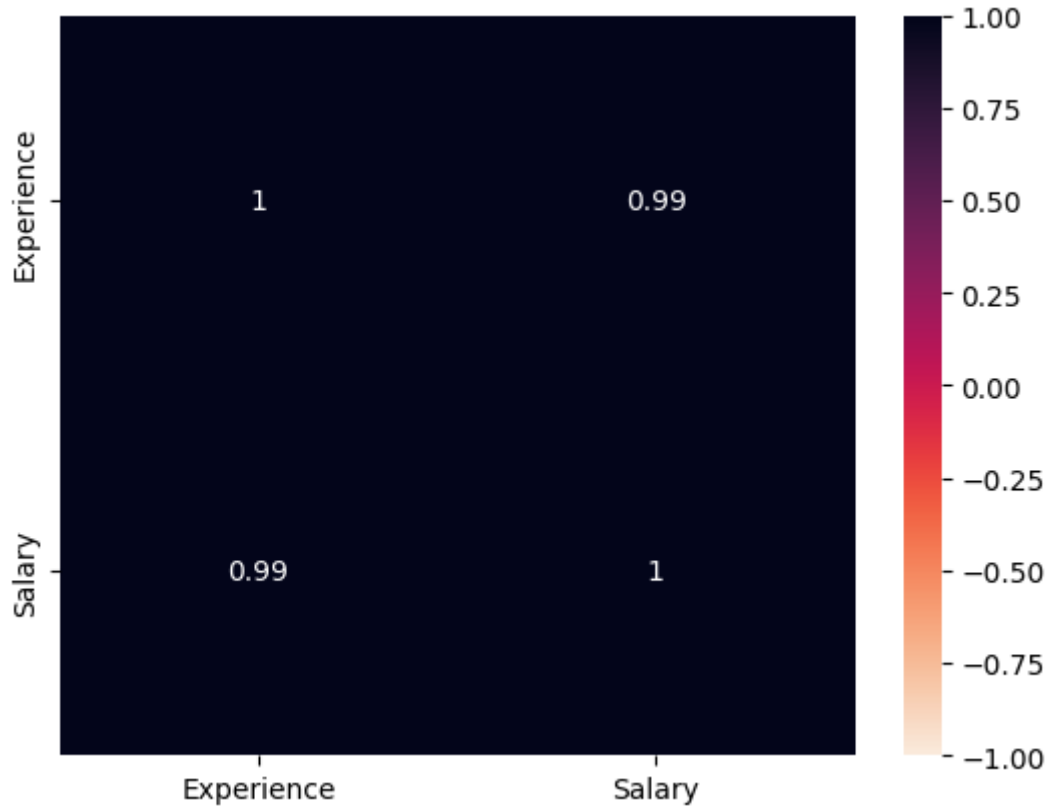
Out[9]:

	Experience	Salary
Experience	1.000000	0.992985
Salary	0.992985	1.000000

In [10]:

```
import seaborn as sns

sns.heatmap(df.corr(), vmin=-1, vmax=1,
annot=True,cmap="rocket_r")
plt.show()
```



Conclusión

La correlación solo cuantifica la fuerza y la dirección de la relación entre dos variables. Puede haber una fuerte correlación entre dos variables, pero no nos permite concluir que una causa la otra. Cuando las correlaciones fuertes no son causales, las llamamos correlaciones espurias.