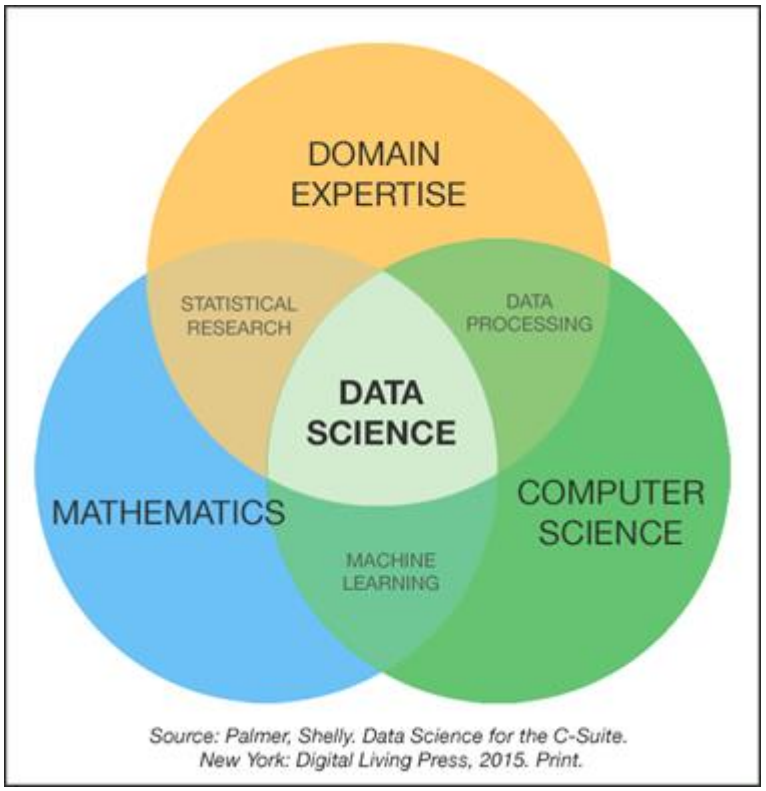
CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

CIENCIA DE DATOS:

APRENDE LOS FUNDAMENTOS DE MANERA PRÁCTICA



SESION 03

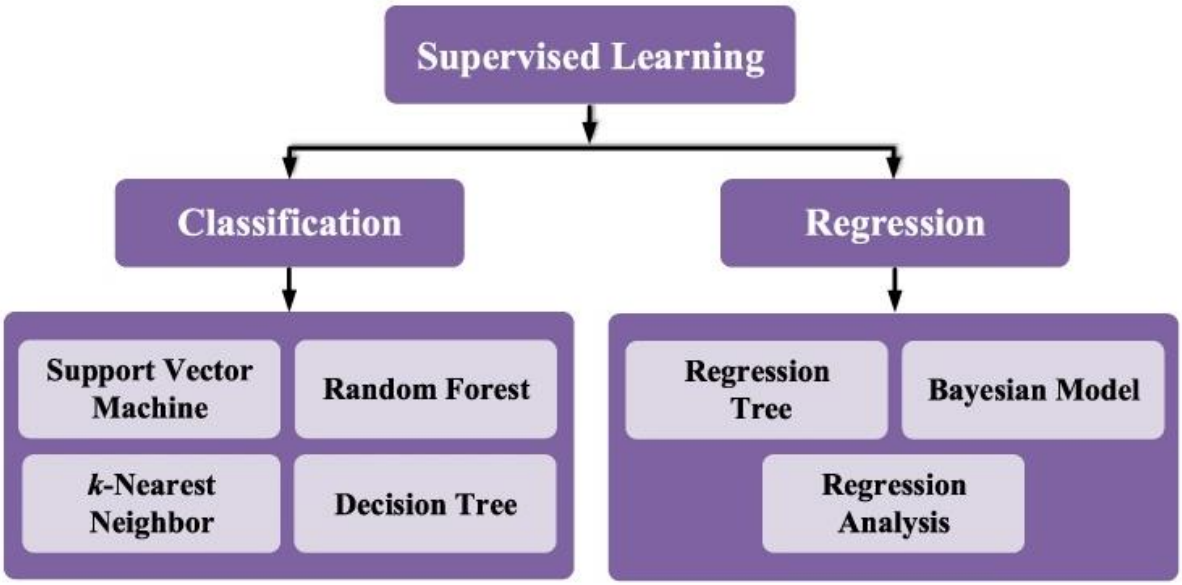
APRENDIZAJE SUPERVISADO

Juan Antonio Chipoco Vidal
jchipoco@gmail.com

CTIC UNI
CENTRO DE TECNOLOGÍAS DE INFORMACIÓN Y COMUNICACIONES
UNIVERSIDAD NACIONAL DE INGENIERÍA

ÍNDICE

OBJETIVO



El objetivo de esta sesión es profundizar en el análisis de la correlación lineal de dos variables, la cual cuantifica que tan relacionadas están las mismas. Esta técnica está estrechamente relacionada con la regresión lineal la cual da lugar a una ecuación que describe dicha relación en términos matemáticos.

En la práctica de esta sesión, continuación de la práctica de la sesión anterior, finalizaremos el análisis exploratorio de datos para poder ya aplicar diversos algoritmos de clasificación para obtener la variable objetivo buscada, en este caso la supervivencia o no de un pasajero del Titanic.

MEDIDAS DE DISPERSIÓN

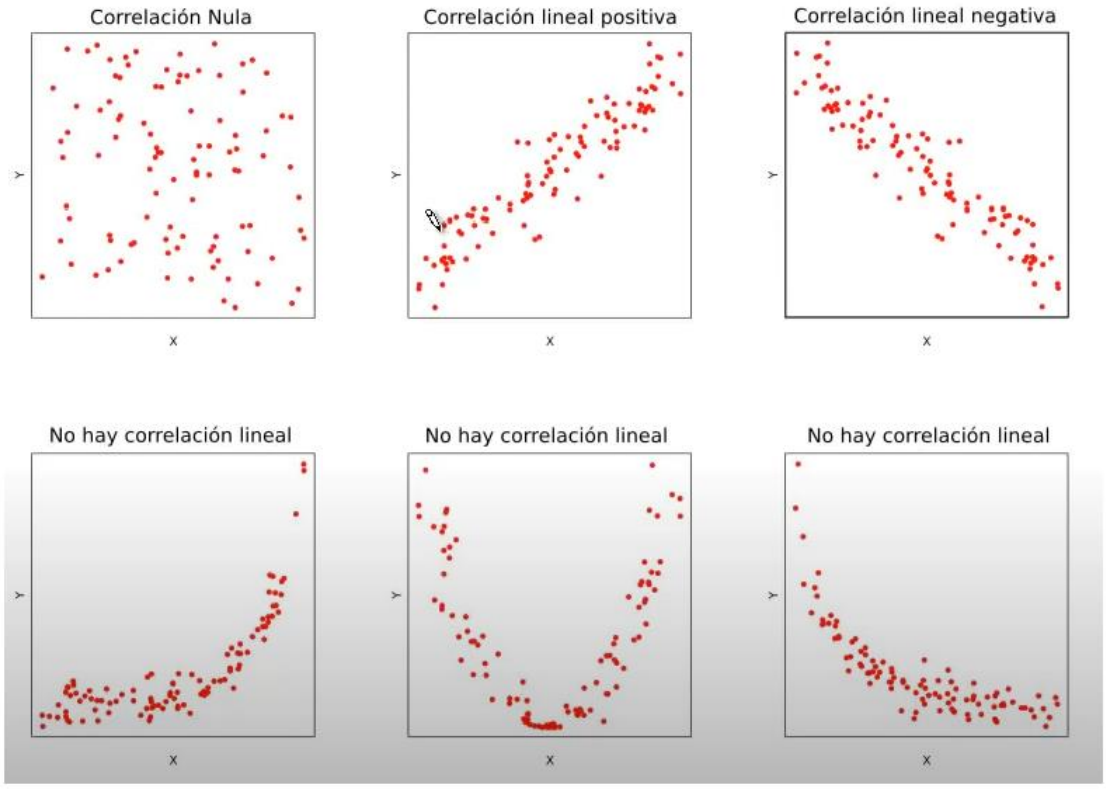
Las medidas de dispersión, se utiliza para describir la variabilidad en una muestra o población. Por lo general, se usa junto con una medida de tendencia central, como la media o la mediana, para proporcionar una descripción general de un conjunto de datos.

Correlación

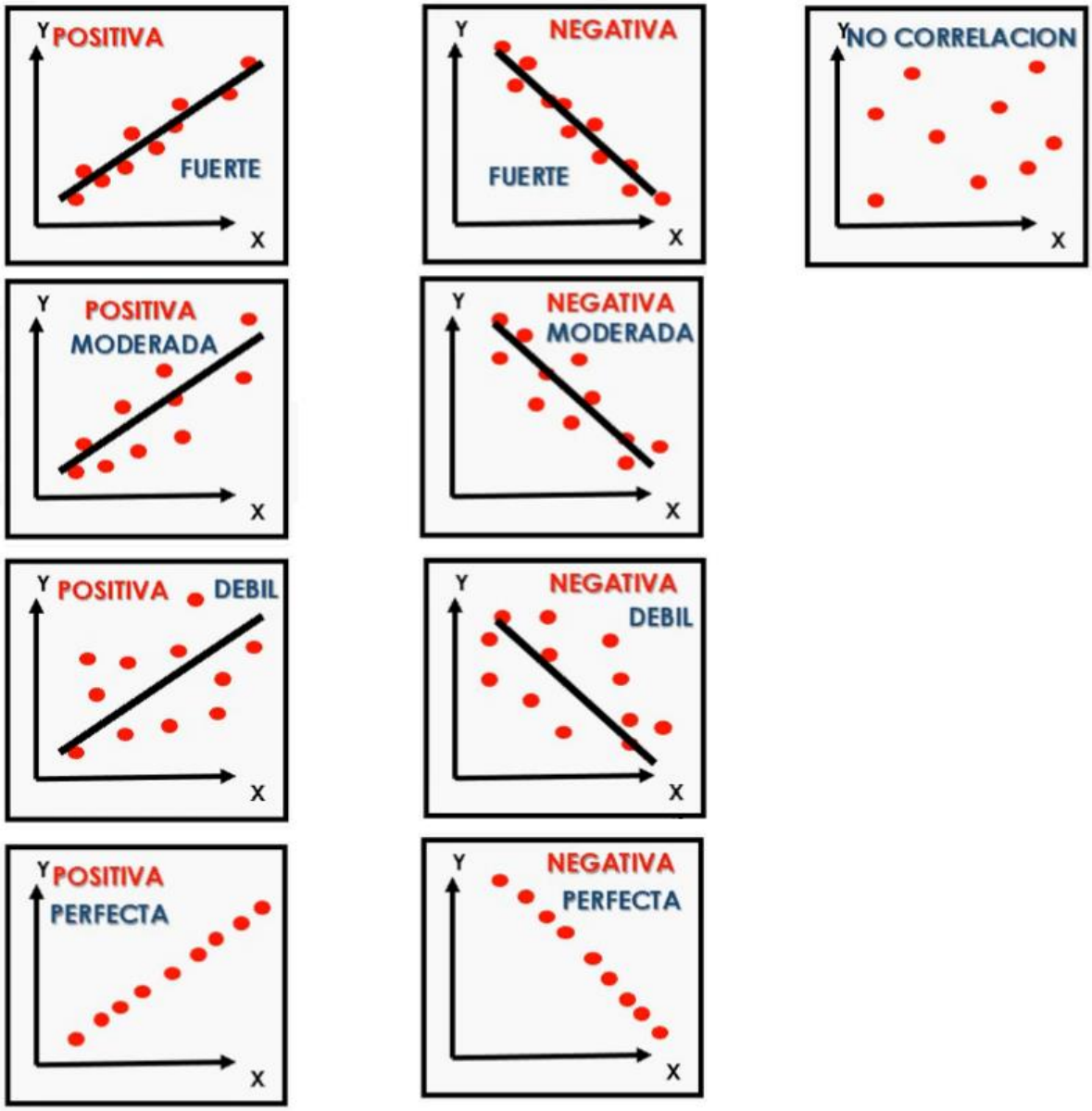
La correlación sirve para medir la relación que existe entre dos o más variables.

La correlación contesta preguntas como las siguientes:

- ¿La práctica de algún deporte está relacionada con una vida más longeva?
- ¿Existe una relación entre la cantidad de carne ingerida diariamente y el cáncer?
- ¿Mayor estudio implica mejores notas en un examen?



Si la correlación es lineal su dirección puede ser positiva o negativa. Su fuerza varía entre perfecta y nula.



Coeficiente de Correlación

Para cuantificar las relaciones anteriores tenemos el Coeficiente de Correlación al cual se le asignara un valor entre -1 y 1.

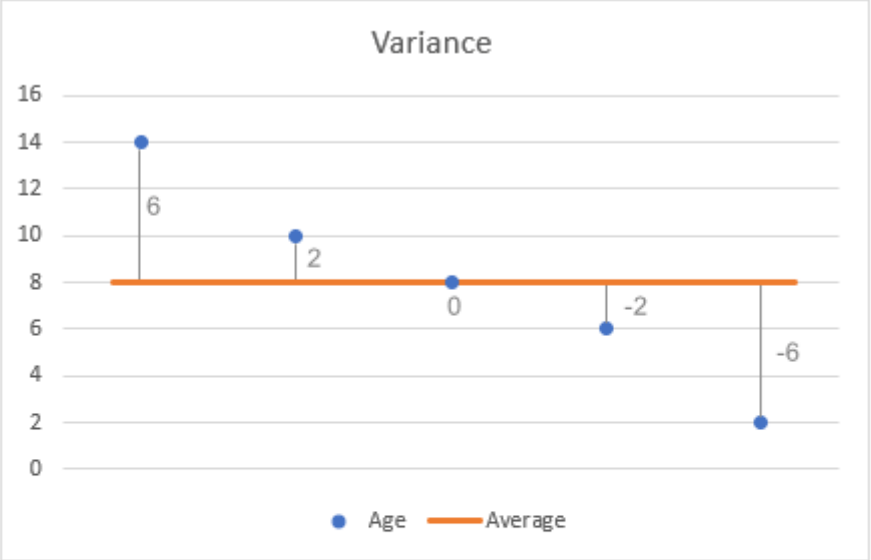
Este coeficiente nos da una medida de la fuerza y el sentido de una relación lineal entre variables cuantitativas.

Cuando el signo es positivo la asociación lineal es positiva lo que implica que cuando el valor de una variable x aumenta, también aumenta el valor de la otra variable y.

Cuando el signo es negativo la asociación lineal es negativa lo que implica que cuando el valor de una variable x aumenta, el valor de la otra variable y disminuye.

Coeficiente de Determinación R^2

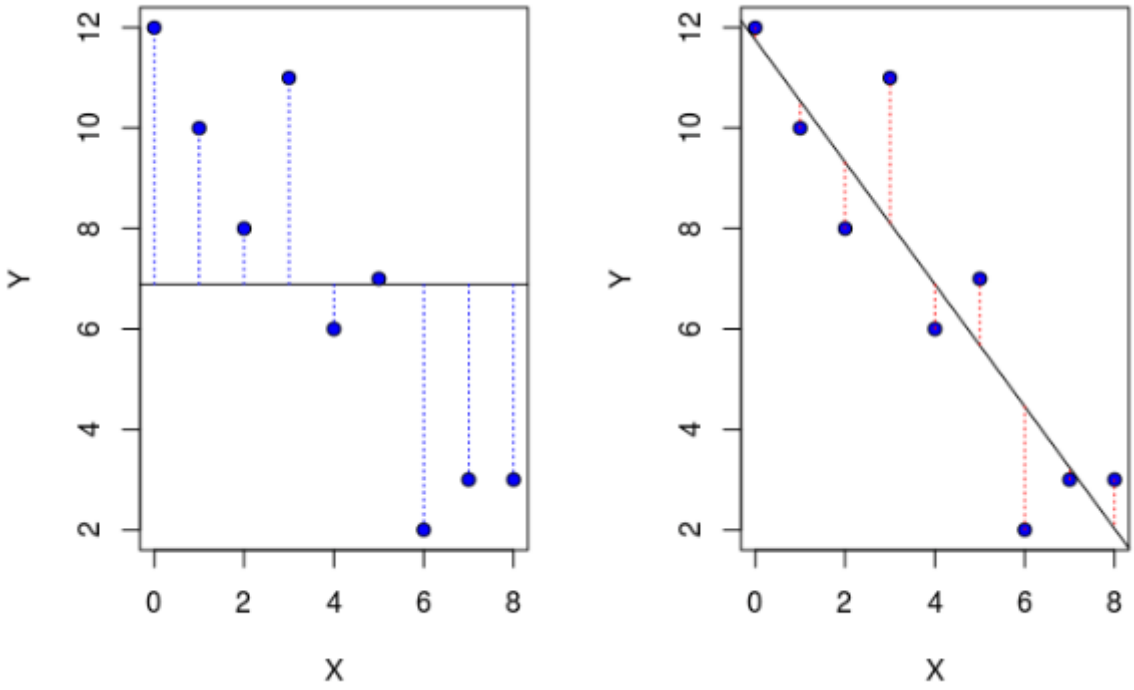
Recordemos que la varianza es la medida de la variabilidad de un conjunto de datos que indica hasta qué punto se distribuyen los diferentes valores. Matemáticamente, se define como la suma de los cuadrados de las diferencias entre una variable y su media, dividido entre el número de datos.



$$Mean = \frac{14 + 10 + 8 + 6 + 2}{5} = 8$$

$$Variance = \frac{6^2 + 2^2 + 0^2 + (-2)^2 + (-6)^2}{5} = 16$$

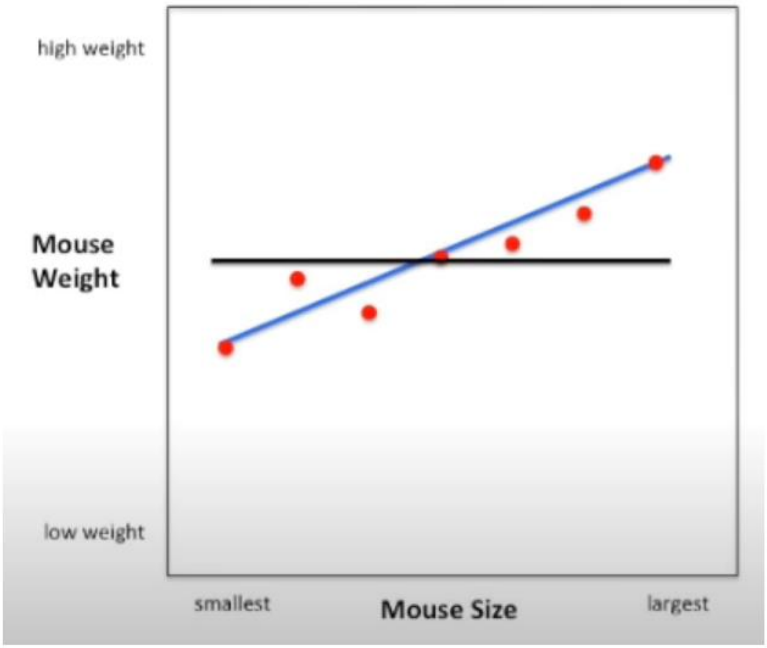
El 16 nos da una idea de la dispersión de los datos. Un valor de 0 indica que no hay variabilidad, mayor el valor, mayor la dispersión de los datos.



$$\text{Var(mean)} = \Sigma\,(y_i - \bar{y})^2 \qquad \text{Var(line)} = \Sigma\,(y_i - (mx_i + b))^2$$

En la gráfica anterior trataremos de averiguar que tan bien se ajusta la recta del lado derecho al conjunto de datos. ¿Cuál es la bondad del ajuste?

Variables correlacionadas



La suma total de cuadrados de los residuos de la imagen anterior Var(line) representa la variación del modelo ajustado, o variación no explicada por el modelo (recta de regresión).

Supongamos que: Var(mean) = 32 y Var(line) = 6

Por lo que Var(line)/Var(mean) nos indicara que porcentaje de la variación total en y (peso del ratón) no está explicada por la variación en x (tamaño del ratón).

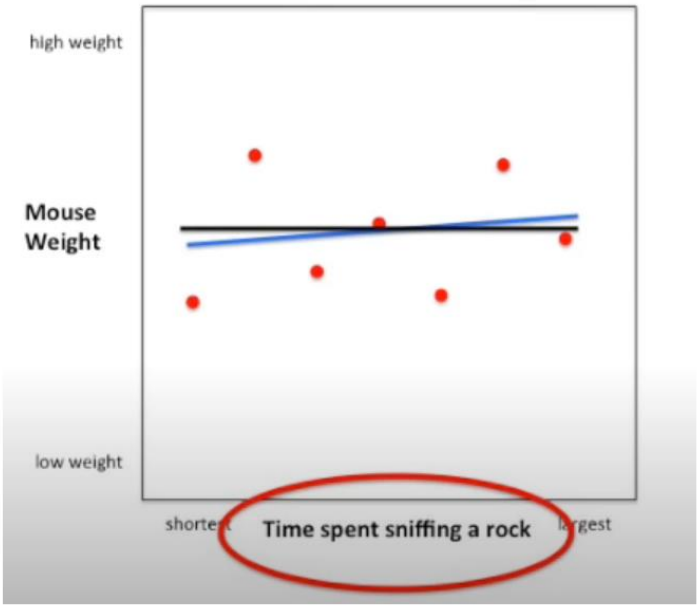
Var(line)/Var(mean) = 6/32 = 19%

Así pues, para saber qué porcentaje de la variación total en y (peso del ratón) esta explicada por la variación en x (tamaño del ratón) usamos 1 - Var(line)/Var(mean) = 81%

En otras palabras, la relación entre las dos variables explica el 81% de la variación de los datos. Esta relación es significativa.

A este último resultado se le conoce como coeficiente de determinación R^2.

Variables no correlacionadas



Var(mean) = 32 y Var(line) = 30

Var(line)/Var(mean) nos indicara que porcentaje de la variación total en y (pero del ratón) no está explicada por la variación en x (tiempo oliendo una roca).

Var(line)/Var(mean) = 30/32 = 94%

Así pues, para saber qué porcentaje de la variación total en y (peso del ratón) esta explicada por la variación en x (tiempo oliendo una roca) usamos 1 - Var(line)/Var(mean) = 6%

En otras palabras, la relación entre las dos variables explica el 6% de la variación de los datos. Esta relación no es significativa.

Si el coeficiente de correlación R = 0.9 entonces el coeficiente de determinación R2 = 0.81, la relación entre las dos variables explica el 81% de la variación de los datos.

R^2 es más fácil de interpretar, por ejemplo, que tan mejor es R = 0.7 que R = 0.5

$R^2 = 0.72 = 0.49$

$R^2 = 0.52 = 0.25$

Con R^2 es fácil ver que la primera correlación es el doble mejor que la segunda correlación.

COVARIANZA
Y EL COEFICIENTE DE CORRELACIÓN

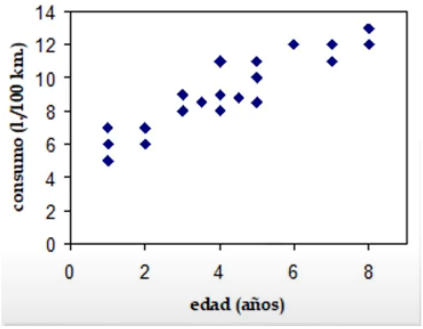
En el siguiente grafico tenemos información de la edad de 20 automóviles, así como el consumo de gasolina en litros por cada 100 km según la edad del automóvil.

Edad y consumo de gasolina de 20 automóviles

edad auto (años)	Consumo (l/100km)
7	11
5	10
3	8
2	7
7	12
8	12
5	11
4	11
4	8
8	13
1	7
6	12
1	6
3	9
2	6
3,5	8,5
1	5
4,5	8,75
5	8,5
4	9

Al graficar el diagrama de dispersión podemos ver que hay una relación lineal positiva o directa entre ambas variables. Nos da información sobre la covariación (variación conjunta) y sus características, si es lineal, su signo y su intensidad.

Nube de puntos
(diagrama de dispersión)



Ahora calculemos la covarianza y el coeficiente de correlación para estos datos:

edad auto x_i	consumo y_i	$x_i y_i$	x_i^2	y_i^2
7	11	77	49	121
5	10	50	25	100
3	8	24	9	64
2	7	14	4	49
7	12	84	49	144
8	12	96	64	144
5	11	55	25	121
4	11	44	16	121
4	8	32	16	64
8	13	104	64	169
1	7	7	1	49
6	12	72	36	144
1	6	6	1	36
3	9	27	9	81
2	6	12	4	36
3,5	8,5	29,75	12,25	72,25
1	5	5	1	25
4,5	8,75	39,375	20,25	76,5625
5	8,5	42,5	25	72,25
4	9	36	16	81
Totales	84	182,75	856,625	446,5	1770,0625

De la fórmula de covarianza tenemos:

$$\sigma_{XY} = \sum_{i=1}^{N} \frac{(X_i - \overline{X})(Y_i - \overline{Y})}{N}$$

$$\sigma_{XY} = 4.4548$$

$$S_{XY} \begin{cases} > 0 \Rightarrow \text{covariación lineal directa (positiva)} \\ < 0 \Rightarrow \text{covariación lineal inversa (negativa)} \\ = 0 \Rightarrow \text{no hay covariación lineal} \end{cases}$$

De la formula del coeficiente de correlación:

$$\rho_{XY} = \sigma_{XY}/(\sigma_X \sigma_Y)$$

$$\rho_{XY} = 0.9194$$

Se trata entonces de una relación directa o positiva y muy fuerte.