Instituto Politécnico Nacional

Escuela Superior de Cómputo

Práctica 5

Coeficientes de correlación de Pearson, Spearman, Kendall y Chi cuadrada..

Alumno: López Fabián Jesús Manuel
Profesor: Flores Estrada Ituriel Enrique
Materia: Análitica y visualización de datos
Grupo: 5AV1

28 de junio de 2024

1. Limpieza o transformación de datos.

Para un analisis de correlación lineal es necesario eliminar las dimensiones categoricas, ya que los coeficientes de correlación lineal, como el de Pearson, están diseñados específicamente para medir relaciones lineales entre variables numéricas continuas. Incluir variables categóricas sin una relación ordinal en dicho análisis puede producir resultados engañosos y falsos. Al convertir categorías en números mediante técnicas como Label Encoding, se introduce una pseudo-relación ordinal que puede llevar a interpretaciones incorrectas, como asumir una jerarquía o proporción que no existe.

Una relación ordinal es una relación entre categorías donde estas pueden ser ordenadas o clasificadas de manera significativa según un criterio específico. Las categorías tienen un orden secuencial, aunque las distancias entre ellas no son necesariamente uniformes. Por ejemplo, en una escala de satisfacción (Muy Insatisfecho, Insatisfecho, Neutral, Satisfecho, Muy Satisfecho), hay una secuencia lógica en las respuestas.

Se identificaron cuatro dimensiones categóricas:

- Formation
- Well Name
- Facies
- NM_M

	Facies	Formation	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	NM_M	RELPOS
0	3	A1 SH	SHRIMPLIN	2793.0	77.45	0.664	9.9	11.915	4.6	1	1.000
1	3	A1 SH	SHRIMPLIN	2793.5	78.26	0.661	14.2	12.565	4.1	1	0.979
2	3	A1 SH	SHRIMPLIN	2794.0	79.05	0.658	14.8	13.050	3.6	1	0.957
3	3	A1 SH	SHRIMPLIN	2794.5	86.10	0.655	13.9	13.115	3.5	1	0.936
4	3	A1 SH	SHRIMPLIN	2795.0	74.58	0.647	13.5	13.300	3.4	1	0.915

Figura 1: Primeros cinco datos del dataset previo a eliminar variables categóricas.

	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	RELPOS
0	2793.0	77.45	0.664	9.9	11.915	4.6	1.000
1	2793.5	78.26	0.661	14.2	12.565	4.1	0.979
2	2794.0	79.05	0.658	14.8	13.050	3.6	0.957
3	2794.5	86.10	0.655	13.9	13.115	3.5	0.936
4	2795.0	74.58	0.647	13.5	13.300	3.4	0.915

Figura 2: Primeros cinco datos del dataset después de eliminar variables categóricas.

2. Obtención de coeficientes de correlación lineal.

A continuación se procede a calcular los coeficientes de correlación lineal (Pearson, Spearman y Kendall) entre cada par de variables, para su representación se usara una matriz de correlación y se presentara como un mapa de calor.

Una matriz de correlación es una tabla que muestra los coeficientes de correlación entre múltiples variables. Cada celda en la matriz contiene el coeficiente de correlación entre las dos variables correspondientes.

2.1. Pearson

El coeficiente de correlación lineal de Pearson es una medida estadística que evalúa la fuerza y la dirección de la relación lineal entre dos variables numéricas. Este coeficiente, también conocido como r, varía entre -1 y 1, donde:

- 1 indica una correlación positiva perfecta.
- -1 indica una correlación negativa perfecta.
- 0 indica que no hay correlación lineal.

El coeficiente de correlación de Pearson r se calcula utilizando la siguiente fórmula:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

donde:

- n es el número de pares de valores.
- x_i y y_i son los valores individuales de las variables X e Y.
- $\bar{x}$ y $\bar{y}$ son las medias de las variables X e Y.

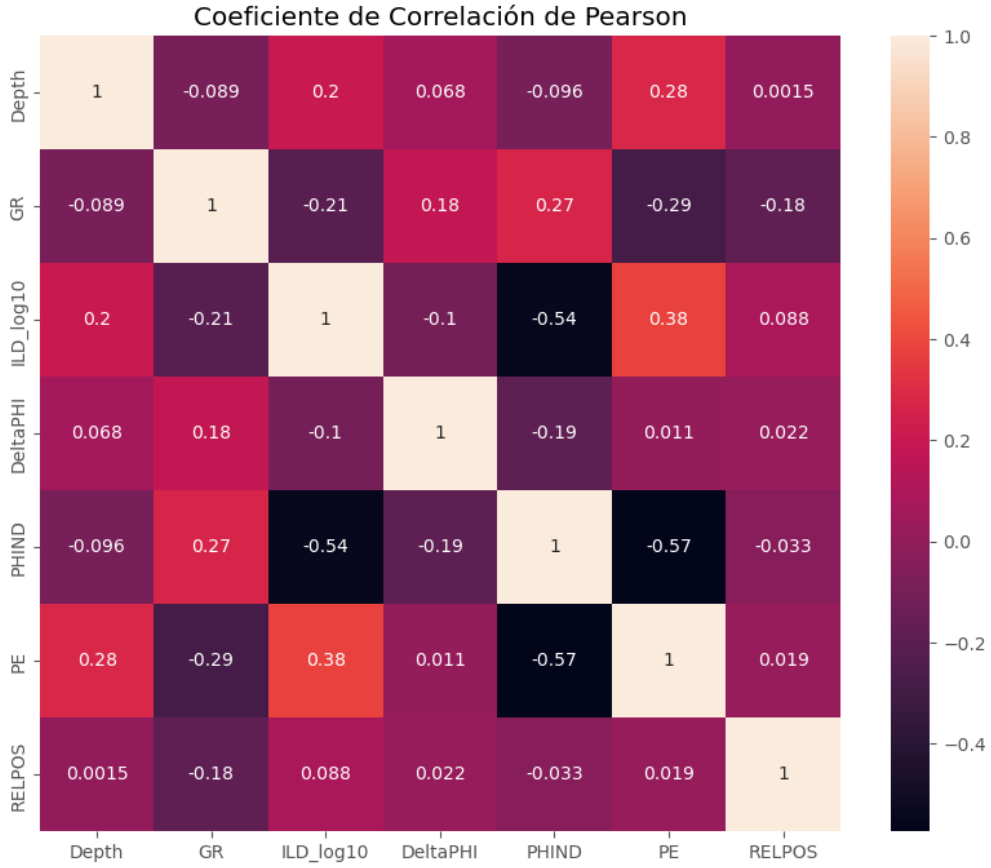

Figura 3: Mapa de calor para coeficientes de correlación de Pearson

En la Figura 3 podemos observar una correlación positiva débil entre los pares de dimensiones siguientes:

- PHIND y GR
- PE y ILD_log10
- PE y Depth

Una correlación positiva débil sugiere que hay una relación positiva entre dos variables, pero esta relación no es fuerte. En términos prácticos, a medida que una variable aumenta, la otra variable tiende a aumentar también, pero esta tendencia no es consistente y puede estar influenciada por una cantidad significativa de variabilidad o ruido en los datos.

En cambio, los pares de variables siguientes presentan una correlación negativa moderada:

- PHIND y PE
- PHIND y ILD_log10

Una correlación negativa moderada sugiere que hay una relación inversa entre dos variables, donde a medida que una variable aumenta, la otra tiende a disminuir de manera consistente, aunque no de forma perfecta.

2.2. Sperman

El coeficiente de correlación de Spearman, también conocido como $\rho(rho)$, es una medida no paramétrica de la relación monótona entre dos variables. A diferencia del coeficiente de correlación de Pearson, que mide la relación lineal entre dos variables, el coeficiente de Spearman evalúa cómo una variable tiende a cambiar cuando la otra cambia, sin suponer una relación lineal específica.

El coeficiente de correlación de Spearman ρ se calcula utilizando la siguiente fórmula:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

donde:

- d_i es la diferencia entre los rangos de las dos variables para la i-ésima observación.
- n es el número de observaciones.

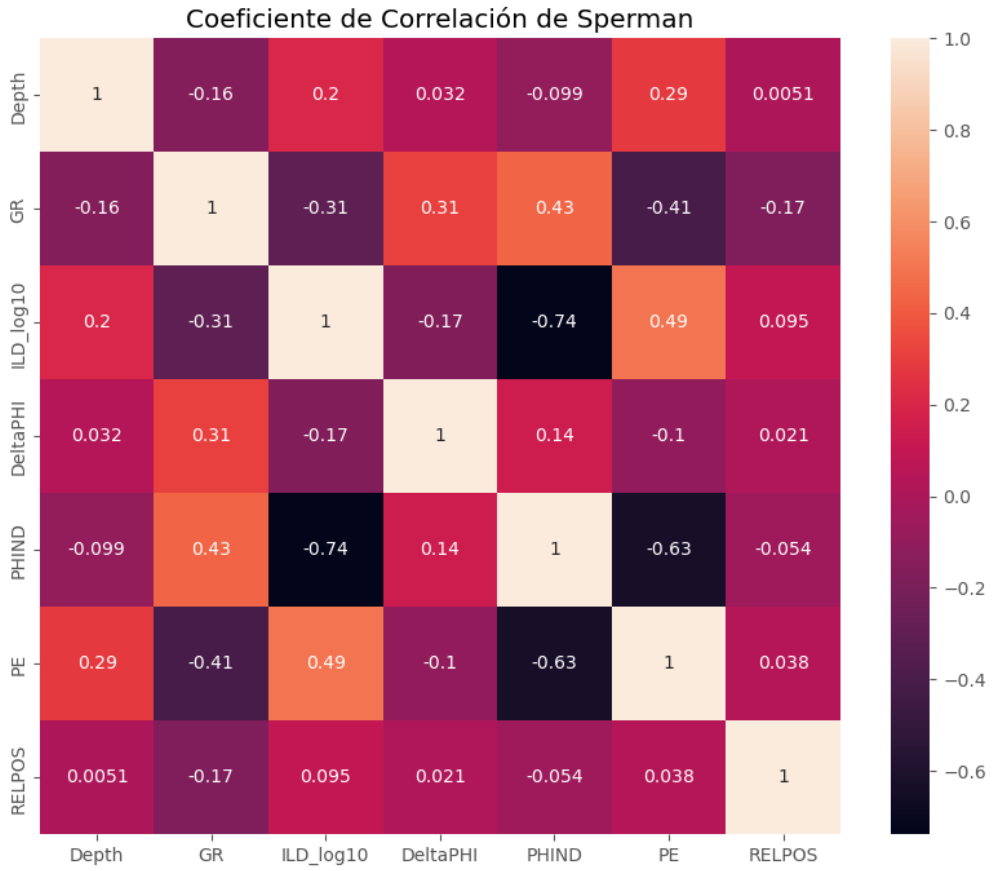

Figura 4: Mapa de calor para coeficientes de correlación de Spearman

2.3. Kendall

El coeficiente de correlación de Kendall, también conocido como $\tau(tau)$, es una medida no paramétrica de la relación entre dos variables ordinales. Evalúa la asociación entre dos variables midiendo la concordancia y discordancia de los pares de observaciones.

En el contexto del coeficiente de correlación de Kendall, los términos **concordancia** y **discordancia** se refieren a la relación entre pares de observaciones en dos variables.

Un par de observaciones (x_i, y_i) y (x_j, y_j) es concordante si el orden de los valores de x coincide con el orden de los valores de y. Es decir, si tanto x_i es mayor que x_j y y_i es mayor que y_j, o si tanto x_i es menor que x_j y y_i es menor que y_j.

Matemáticamente, un par de observaciones es concordante si:

$$(x_i - x_j)(y_i - y_j) > 0$$

Un par de observaciones (x_i, y_i) y (x_j, y_j) es discordante si el orden de los valores de x no coincide con el orden de los valores de y. Es decir, si x_i es mayor que x_j y y_i es menor que y_j, o si x_i es menor que x_j y y_i es mayor que y_j.

Matemáticamente, un par de observaciones es discordante si:

$$(x_i - x_j)(y_i - y_j) < 0$$

El coeficiente de correlación de Kendall τ se calcula utilizando la siguiente fórmula:

$$\tau = \frac{(C - D)}{\frac{1}{2}n(n-1)}$$

donde:

- C es el número de pares concordantes.
- D es el número de pares discordantes.
- $\frac{1}{2}n(n-1)$ es el número total de pares posibles.

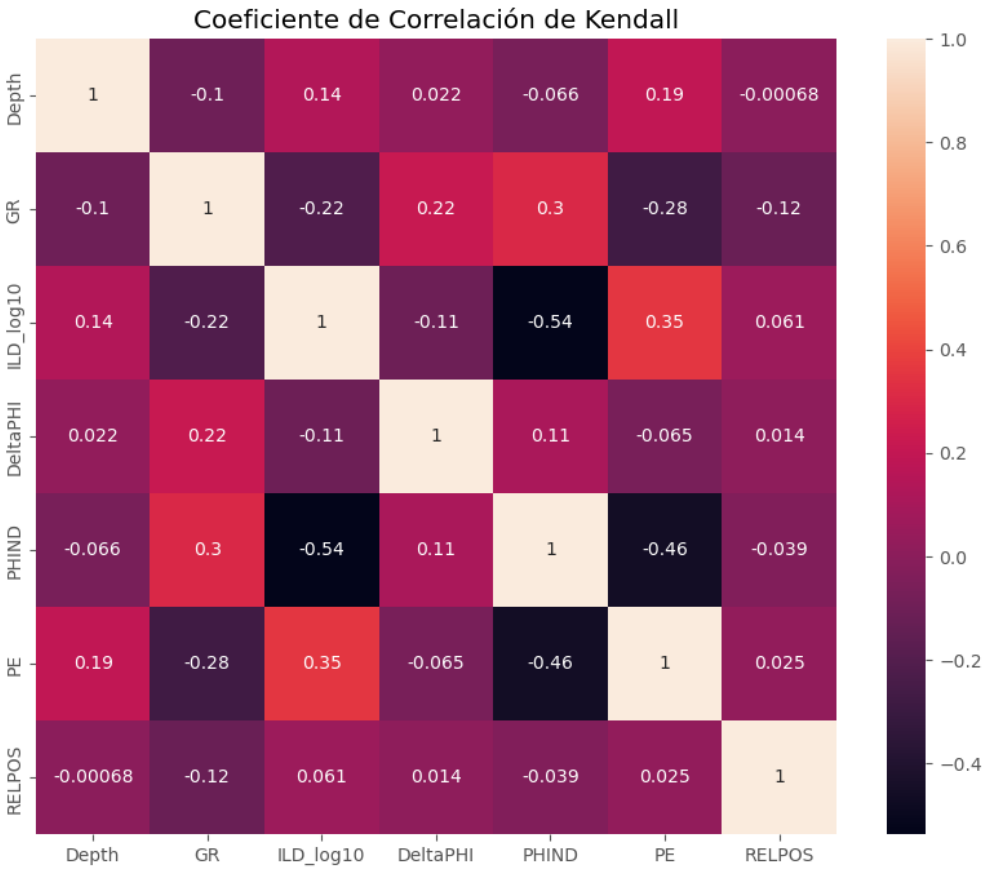

Figura 5: Mapa de calor para coeficientes de correlación de Kendall.

3. Prueba de independencia Chi-cuadrada

La prueba de chi-cuadrado es una prueba estadística utilizada para determinar si hay una asociación significativa entre las variables categóricas en una tabla de contingencia. La prueba se basa en la comparación entre las frecuencias observadas y las frecuencias esperadas bajo la hipótesis nula de independencia.

El estadístico de prueba chi-cuadrado se calcula como:

$$\chi^2 = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

donde:

- O_{ij} son las frecuencias observadas en cada celda de la tabla.
- E_{ij} son las frecuencias esperadas bajo la hipótesis nula de independencia.

El valor esperado E_{ij} para cada celda se calcula como:

$$E_{ij} = \frac{(r_i \cdot c_j)}{n}$$

donde:

- r_i es la suma de las frecuencias de la fila i.
- c_j es la suma de las frecuencias de la columna j.
- n es el total de observaciones en la tabla de contingencia.

El valor crítico de chi-cuadrado se compara con la distribución chi-cuadrado con $(r-1)\cdot(c-1)$ grados de libertad, donde r es el número de filas y c es el número de columnas en la tabla de contingencia.

Para hacer el analisis de independencia chi-cuadrada se procedio a eliminar aquellas dimensiones que presentan valores numericos continuos en sus observaciones, como lo son: .ªge", "balance", "day","duration", çampaign", "pdaysƶ"previous"

	Variable1	Variable2	Chi2	P-value	Significativo
0	job	marital	373.181838	1.372525e-65	True
1	job	education	2840.042901	0.000000e+00	True
2	job	default	9.306352	5.936374e-01	False
3	job	housing	380.003636	1.069462e-74	True
4	job	loan	47.191298	1.988689e-06	True
...	...	...	...	...	...
85	y	housing	48.884628	2.714700e-12	True
86	y	loan	21.871822	2.914829e-06	True
87	y	contact	87.869857	8.304301e-20	True
88	y	month	250.500063	2.195355e-47	True
89	y	poutcome	386.877387	1.539883e-83	True

Figura 6: Prueba chi-cuadrada para cada par de variables.

True (Verdadero): Significa que la prueba de chi-cuadrado de independencia encontró evidencia estadísticamente significativa para rechazar la hipótesis nula. Esto indica que existe una asociación significativa entre las dos variables categóricas analizadas.

False (Falso): Significa que la prueba de chi-cuadrado de independencia no encontró evidencia estadísticamente significativa para rechazar la hipótesis nula. Esto indica que no hay suficiente evidencia para concluir que existe una asociación significativa entre las dos variables categóricas analizadas.

En total tenemos 70 pares de variables que pasaron la prueba de independencia chi-cuadrado y 20 que no pasaron esta prueba estadística.

4. Conclusiones

Al analizar un mismo conjunto de datos utilizando los coeficientes de correlación de Pearson, Spearman y Kendall, se pueden observar discrepancias significativas en los resultados debido a las diferentes características y enfoques de cada medida. Estas diferencias revelan cómo cada coeficiente está diseñado para capturar aspectos específicos de las relaciones entre variables. Mientras que Pearson es ideal para relaciones lineales claras, Spearman y Kendall ofrecen mejores resultados cuando la relación es monotónica pero no necesariamente lineal. Considerar la presencia de valores atípicos y la estructura ordinal de los datos es crucial para interpretar correctamente los resultados y obtener conclusiones significativas del análisis de correlación.

La prueba de chi-cuadrado es crucial por su capacidad para evaluar de manera estadísticamente rigurosa si las diferencias observadas entre las frecuencias esperadas y las observadas en variables categóricas son significativas, proporcionando una medida objetiva de la asociación entre estas variables. Esto la convierte en una herramienta fundamental en el análisis de datos categorizados, facilitando la identificación de relaciones subyacentes que pueden ser relevantes para la toma de decisiones informadas en diversas disciplinas científicas y aplicaciones prácticas.