Análisis Estadístico Multivariado

26/10/21

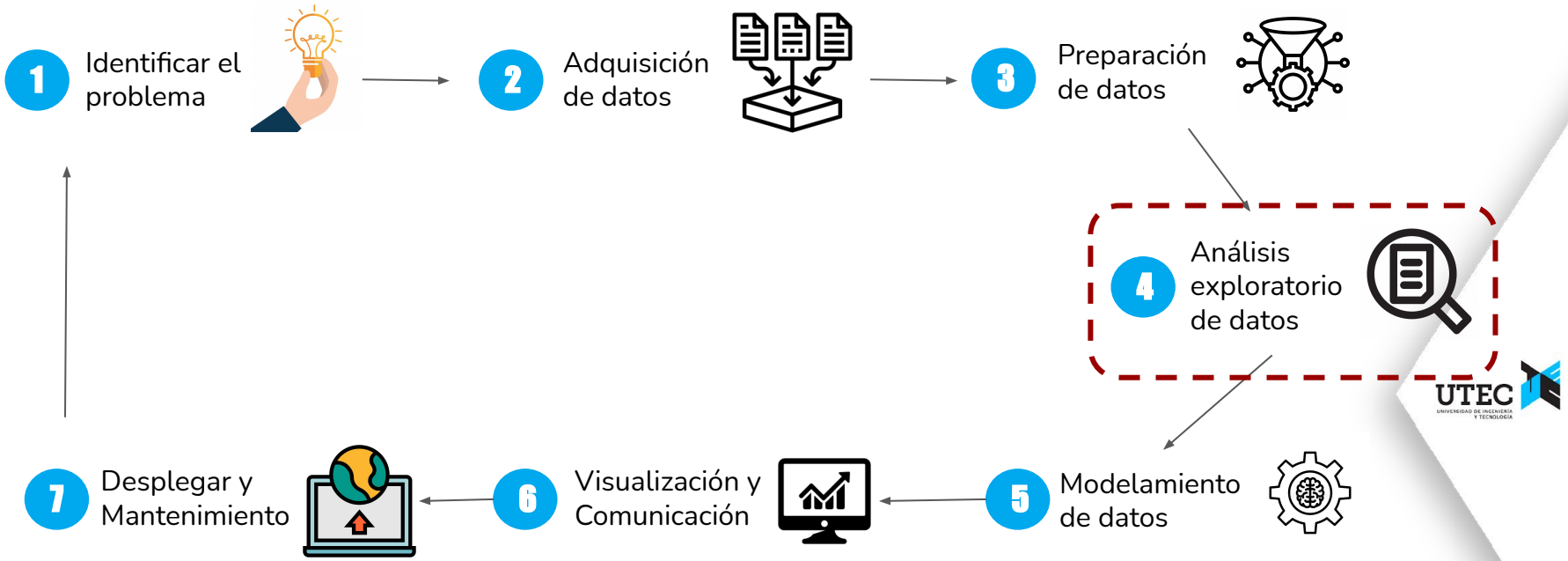
Unidad 4: Análisis Exploratorio de Datos

UTEC
UNIVERSIDAD DE INGENIERÍA Y TECNOLOGÍA

Índice

- Análisis Multivariado
  - Método No Gráfico
  - Método Gráfico

RECORDANDO:
Ciclo de Vida de un Proyecto en DS



Objetivos de la Sesión

Describir el análisis exploratorio de datos y su relación con el análisis estadístico univariado y multivariado de datos categóricos y numéricos.

¿Cuál es el objetivo del Análisis Exploratorio de Datos (EDA)?

¿Cuál de los siguientes atributos del Diamond dataset son categóricos?

- Cut, color, depth
- Color, price, cut
- Cut, color, clarity
- Color, table, x
- None of the above

¿Qué método de EDA se usa para analizar datos univariados cuantitativos ?

Bar Plot

Histogram

Diamonds Dataset

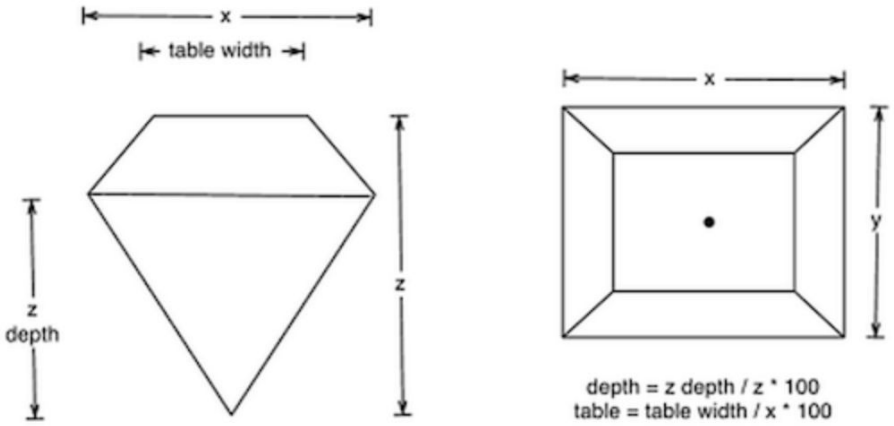
- Este conjunto de datos clásico contiene los precios y otros atributos de casi 54 000 diamantes. Es un gran conjunto de datos para principiantes que aprenden a trabajar con el análisis y la visualización de datos.

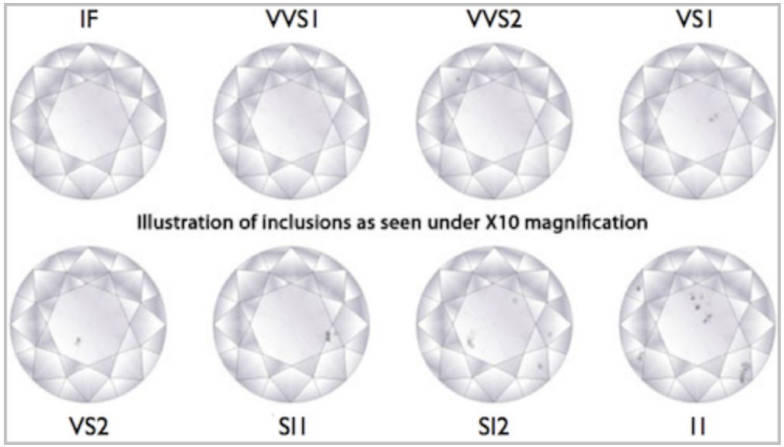
Diamonds Dataset

Atributo/ Columna	Descripción
Carat	Peso del diamante en quilates (0.2–5.01)
Cut	Describe la calidad del corte del diamante (Fair, Good, Very Good, Premium, Ideal)
Color	Color del diamante (D-J)
Clarity	Mide cuán claro es el diamante (I1 (worst), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (best))
Depth	Porcentaje de profundidad total del diamante (43-79)

Atributo/ Columna	Descripción
Table	Porcentaje de su diámetro promedio (43-95)
Price	El precio en dólares del diamante ($326 - $18 823)
X	Longitud del diamante en mm (0-10.74)
Y	Ancho del diamante en mm (0–58.9)
Z	Profundidad del diamante en mm (0–31.8)

Diamonds Dataset (2)





Clasificación de EDA

	UNIVARIADO	MULTIVARIADO

MÉTODO NO GRÁFICO	• Variable Categórica:
	  ○ Representación tabular de frecuencia.
	• Variable Cuantitativa:
	  ○ Ubicación (promedio, mediana)
	  ○ Dispersión (dev, varianza, std dev, rango, percentil)
		• Dos variables Categóricas:
		  ○ Tabulación Cruzada
		• Dos o más variables Cuantitativas:
		  ○ Correlación
		  ○ Covarianza

MÉTODO GRÁFICO	• Variable Categórica:
	  ○ Bar Plots
	• Variable Cuantitativa:
	  ○ Histogramas
	  ○ Boxplots
		• Bar Plots agrupados
		• Side-by-side boxplots
		• Histogramas
		• Scatterplot
		• Mapa de calor de correlación (Correlation Heatmap)

Yamilet Serrano Ll., PhD.

Análisis Multivariado

Método No Gráfico

Análisis Multivariado: Método No Gráfico
Datos Categóricos

- Los datos multivariados surgen de más de una variable.
- Los métodos de EDA no gráficos multivariados generalmente **muestran la relación entre dos o más variables de los datos** mediante:
  - tabulaciones cruzadas (cross-tabulations) o
  - estadísticas.

Análisis Multivariado: Método No Gráfico
Datos Categóricos (2)

- Para analizar los datos multivariados categóricos (y datos cuantitativos con pocos valores diferentes) se usa **tabulación cruzada (cross-tabulation)**.
- Para dos variables, la tabulación cruzada se realiza haciendo una tabla de dos factores con encabezados de columna que coinciden con los niveles de una variable y encabezados de fila que coinciden con los niveles de la otra variable, luego completando los recuentos de todos los sujetos que comparten un par de niveles.

Análisis Multivariado: Método No Gráfico

Datos Categóricos (2)

color cut	D	E	F	G	H	I	J	All
Ideal	0.052540	0.072358	0.070931	0.090545	0.057749	0.038802	0.016611	0.399537
Premium	0.029718	0.043326	0.043215	0.054208	0.043752	0.026474	0.014980	0.255673
Very Good	0.028050	0.044494	0.040119	0.042621	0.033815	0.022321	0.012570	0.223990
Good	0.012273	0.017297	0.016852	0.016148	0.013014	0.009677	0.005692	0.090953
Fair	0.003022	0.004153	0.005784	0.005821	0.005617	0.003244	0.002206	0.029848
All	0.125603	0.181628	0.176900	0.209344	0.153949	0.100519	0.052058	1.000000

Ejemplo: Tabulación cruzada entre el corte (cut) y el color de los diamantes

Análisis Multivariado: Método No Gráfico
Datos Cuantitativos

- Para dos variables cuantitativas, la estadística básica de interés de una muestra son:
    - **La covarianza:** Es una medida de la relación entre dos variables. ¿Cuando una variable cambia, habrá el mismo o un cambio similar en la otra variable?

Análisis Multivariado: Método No Gráfico

Datos Cuantitativos (2)

	carat	depth	table	price	x	y	z
carat	0.224687	0.019167	0.192365	1.742765e+03	0.518484	0.515248	0.318917
depth	0.019167	2.052404	-0.946840	-6.085371e+01	-0.040641	-0.048009	0.095968
table	0.192365	-0.946840	4.992948	1.133318e+03	0.489643	0.468972	0.237996
price	1742.765364	-60.853712	1133.318064	1.591563e+07	3958.021491	3943.270810	2424.712613
x	0.518484	-0.040641	0.489643	3.958021e+03	1.258347	1.248789	0.768487
y	0.515248	-0.048009	0.468972	3.943271e+03	1.248789	1.304472	0.767320
z	0.318917	0.095968	0.237996	2.424713e+03	0.768487	0.767320	0.498011

Ejemplos: Tablas de covarianza entre los datos cuantitativos

Análisis Multivariado: Método No Gráfico
Datos Cuantitativos

- Para dos variables cuantitativas, la estadística básica de interés de una muestra son:
  - **La covarianza:** Es una medida de la relación entre dos variables. ¿Cuando una variable cambia, habrá el mismo o un cambio similar en la otra variable?
  - **La correlación:** Proporciona una mejor comprensión de la covarianza. Es covarianza normalizada. La correlación nos dice qué tan correlacionadas están las variables entre sí.

Análisis Multivariado: Método No Gráfico

Datos Cuantitativos (2)

	carat	depth	table	price	x	y	z
carat	1.000000	0.028224	0.181618	0.921591	0.975094	0.951722	0.953387
depth	0.028224	1.000000	-0.295779	-0.010647	-0.025289	-0.029341	0.094924
table	0.181618	-0.295779	1.000000	0.127134	0.195344	0.183760	0.150929
price	0.921591	-0.010647	0.127134	1.000000	0.884435	0.865421	0.861249
x	0.975094	-0.025289	0.195344	0.884435	1.000000	0.974701	0.970772
y	0.951722	-0.029341	0.183760	0.865421	0.974701	1.000000	0.952006
z	0.953387	0.094924	0.150929	0.861249	0.970772	0.952006	1.000000

Ejemplos: Tablas de correlación entre datos cuantitativos

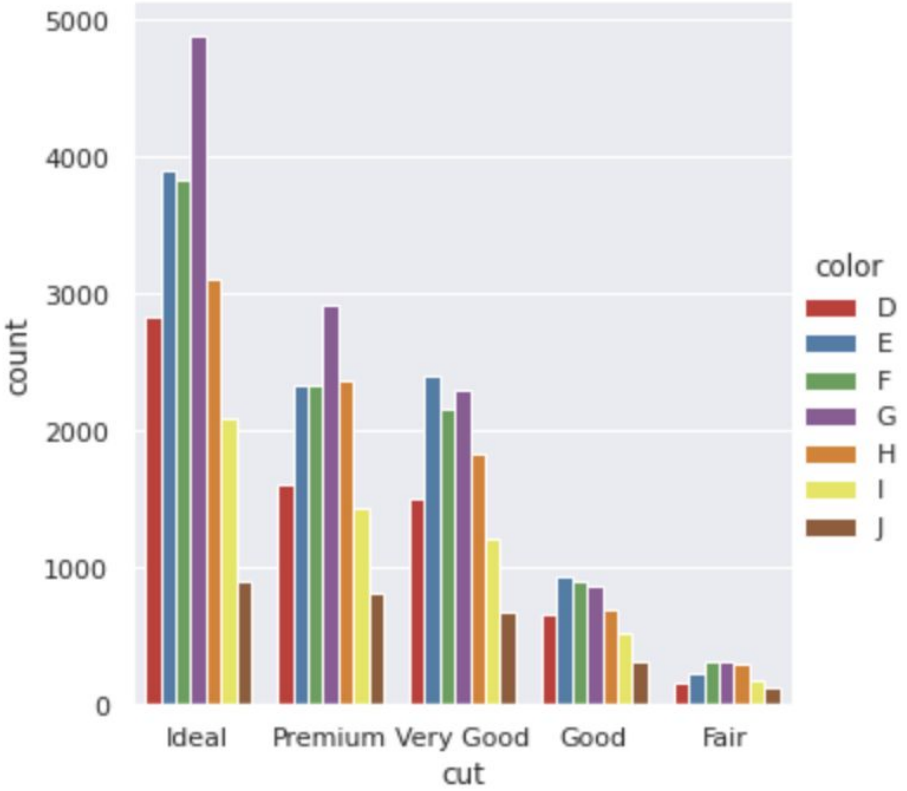
Método Gráfico

Análisis Multivariado: Método Gráfico

- Los datos multivariados utilizan gráficos para mostrar las relaciones entre dos o más conjuntos de datos.
- Los tipos de gráficos para análisis multivariado son:
  - Dos o más variables categóricas:
    - Bar Plots agrupados
  - Dos o más variables categóricas y una cuantitativa.
    - Bar Plots agrupados
  - Una variable categórica y una cuantitativa
    - Side-by-side boxplots
  - Una o más variables categóricas y una cuantitativa
    - Histogramas
  - Dos o más variables cuantitativas
    - Scatterplot
    - Mapa de calor de correlación (Correlation Heatmap)

Bar Plots agrupados

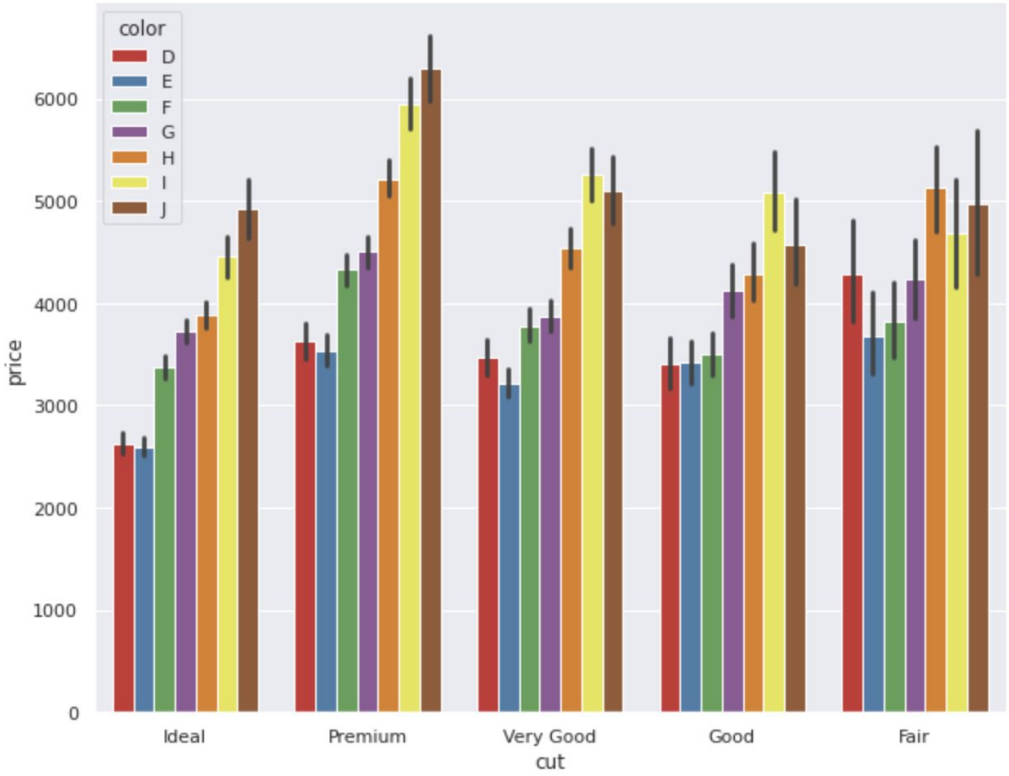
- Los Bar Plots agrupados son comúnmente usados para graficar la distribución de:
  - **Dos o más variables categóricas.**



Ejemplo: Pila de BarPlots que muestra la relación entre los cortes y colores de los diamantes

Bar Plots agrupados (2)

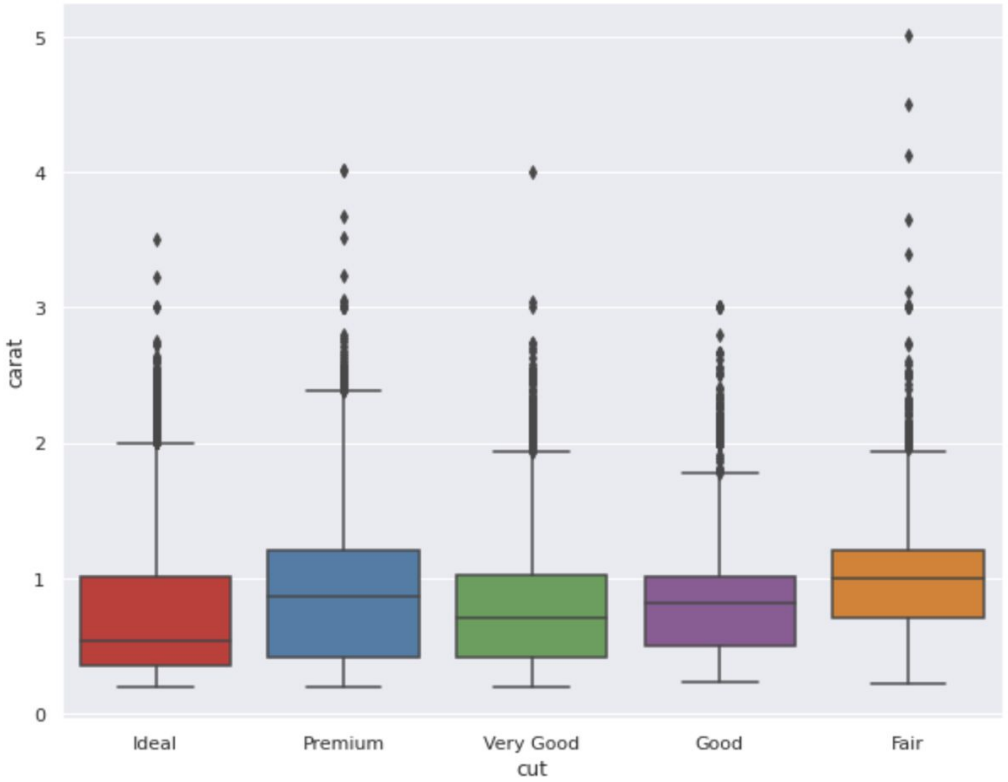
- Los Bar Plots agrupados son comúnmente usados para graficar la distribución de:
  - Dos o más variables categóricas.
  - **Dos o más variables categóricas y una cuantitativa.**



Ejemplo: BarPlots agrupados que muestra la relación entre los cortes y el precio diferenciando por su color de los diamantes

Side-by-Side Boxplots

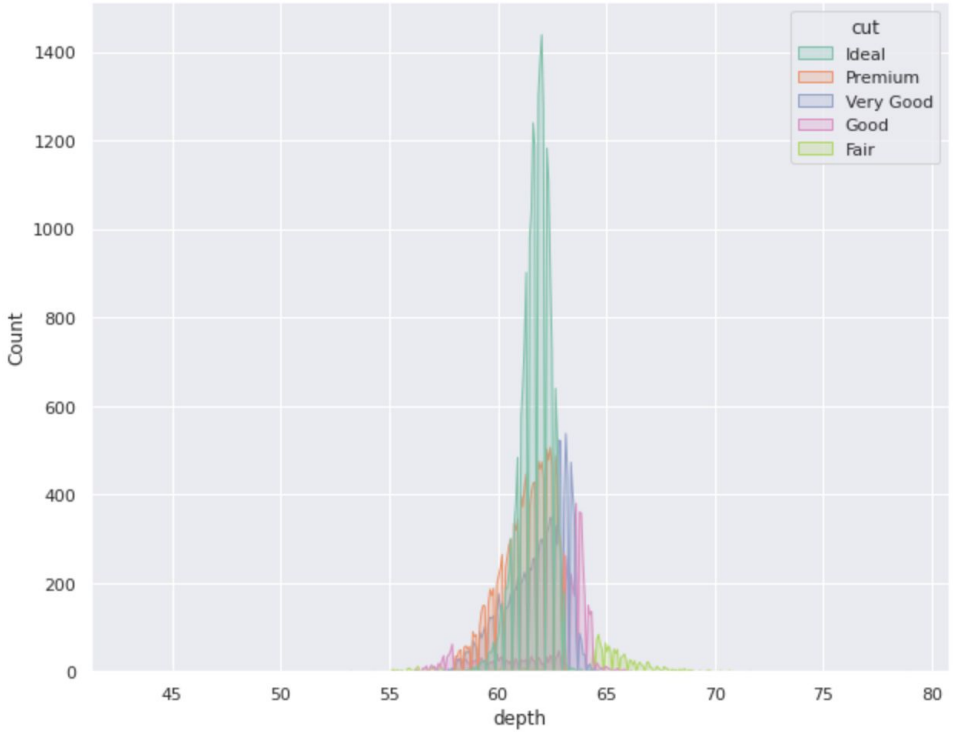
- Los *side-by-side Boxplots* son la mejor técnica de EDA gráfica para examinar la relación **entre una variable categórica y una variable cuantitativa**, así como la distribución de la variable cuantitativa en cada nivel de la variable categórica.



Ejemplo: Side-by-side boxplots que muestra la variación del quilate en función del color de los diamantes

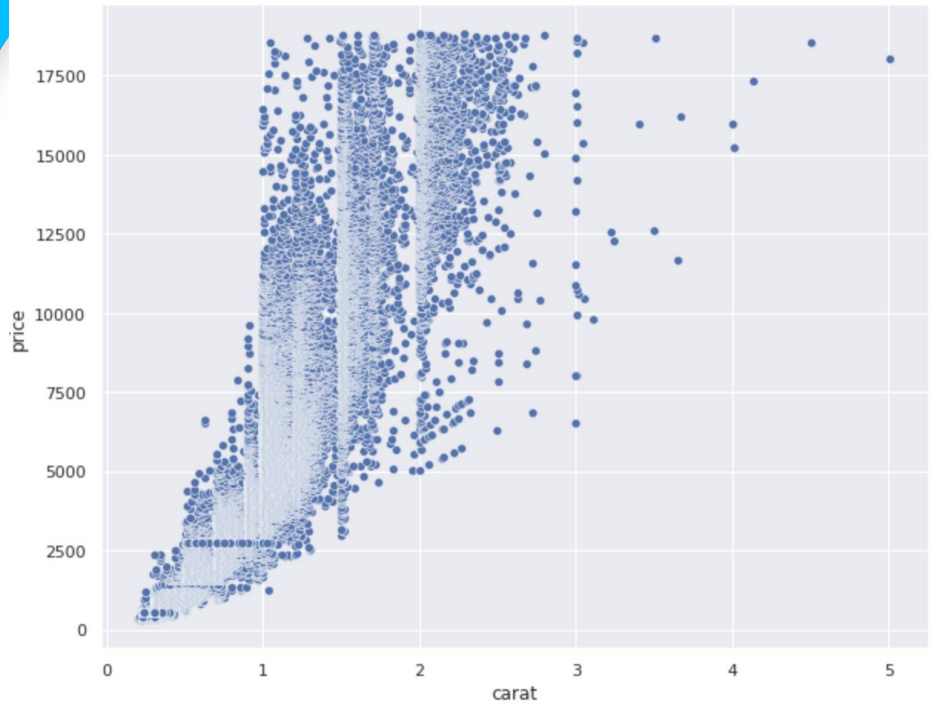
Histogramas

- Recordemos que los histogramas son para datos numéricos continuos (eje x). Sin embargo podemos usar los datos categóricos para distinguir la distribuciones.



Scatterplots

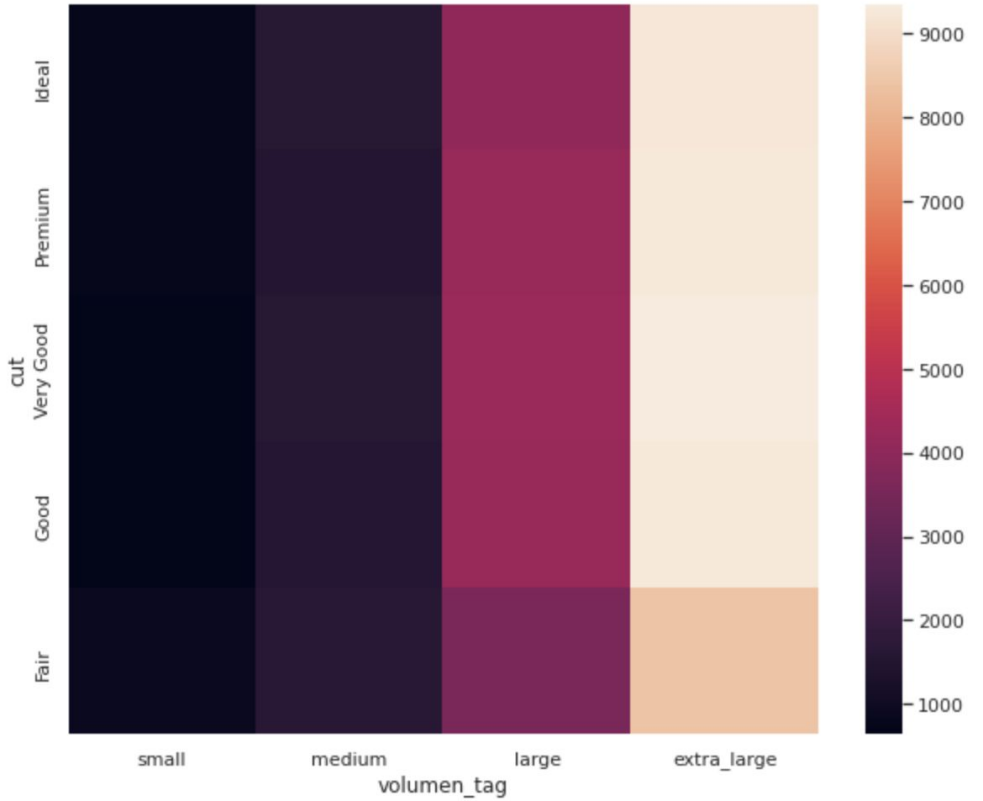
- Para dos variables cuantitativas, la técnica gráfica básica de EDA es el *scatterplot* (diagrama de dispersión) que tiene una variable en el eje x, una en el eje y y un punto para cada caso en su conjunto de datos.
- Si una variable es explicativa y la otra es el resultado, es una convención poner el resultado en el eje *y* (vertical).
- Los diagramas de dispersión permiten comprobar si existe un vínculo potencial entre dos o más variables cuantitativas. Por esta razón, los diagramas de dispersión se utilizan a menudo para visualizar **una posible correlación entre las variables.**

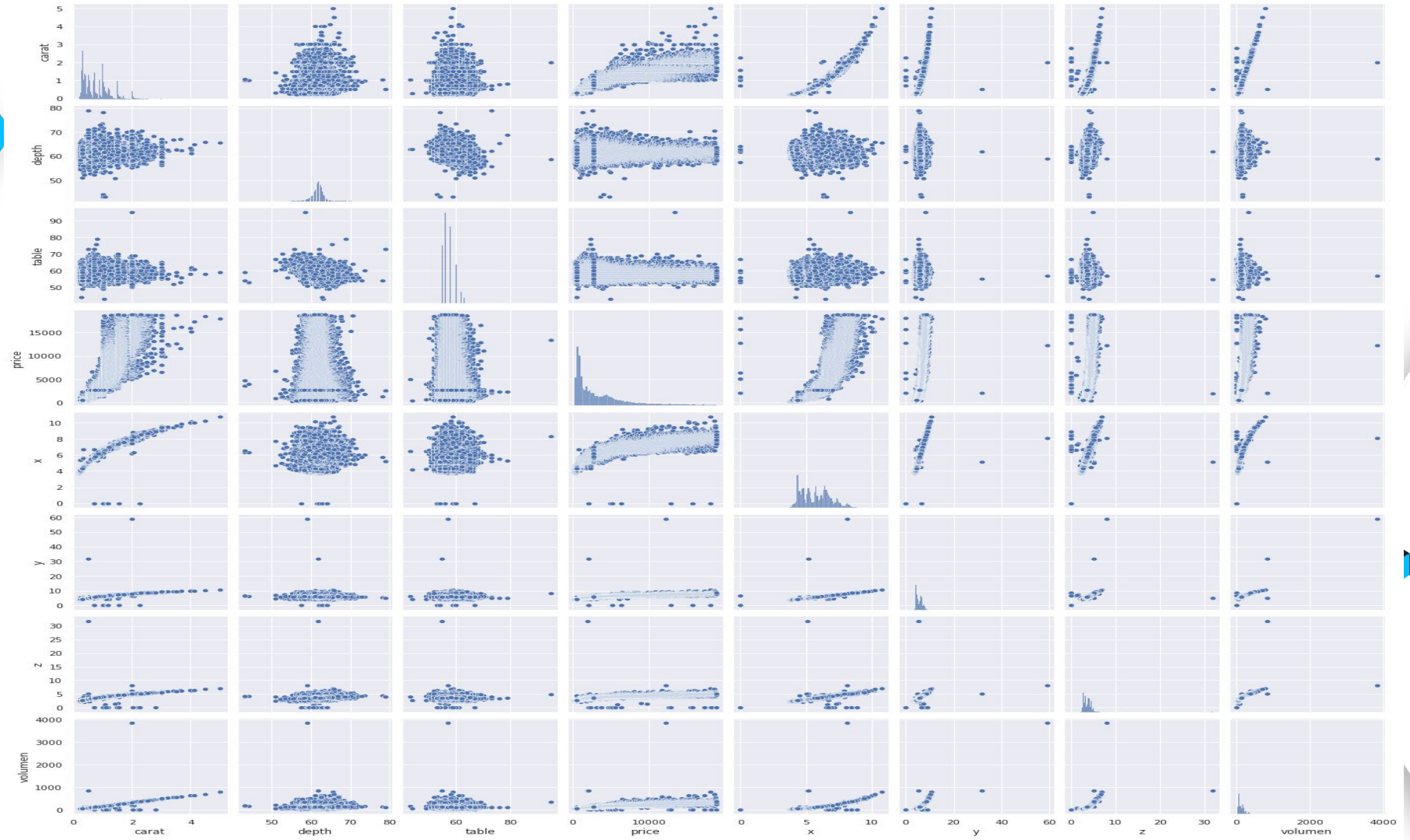


Ejemplos: Scatterplots que muestran la correlación entre los quilates y el precio de los diamantes. El gráfico de la derecha categoriza por la variable corte.

Mapa de Calor de Correlaciones

- Un mapa de calor de correlaciones es usado para ver las correlaciones entre distintos datos cuantitativos.




carat
depth
table
price
x
y
z
volumen

Time to code!

Q&A