Universidad Politécnica Metropolitana de Hidalgo

Programa Educativo de Maestría en Inteligencia Artificial

"Estadística Aplicada"

Reporte: Tarea 2

Flores García Katherine Itzel
(253220100)
Morales Hernández Emmanuel
(253220003)

26 de septiembre de 2025

Índice

Índice de figuras

Índice de cuadros

1. Introducción

El análisis estadístico constituye una herramienta fundamental para comprender el comportamiento de los datos y extraer información significativa. En este estudio, se examina la forma de construir un análisis estadístico robusto de un dataset, aplicando para ello técnicas y herramientas estadísticas diversas, resolviéndolo con el lenguaje de programación Python y usando librerías como: *Pandas, NumPy, MatPlotLib, Seaborn, Scikit*.
El objetivo de esta actividad, es aprender a conocer la estructura cuantitativa y cualitativa de un dataset, aplicando técnicas estadísticas usando el lenguaje de programación de Python. Y a su vez, aprender a extraer información de un dataset, poder transformar dichos datos, limpiarlos y posteriormente cargarlos; con el fin de poder interpretar dichos resultados obtenidos.
A continuación se adjunta el link (Código Python) del código de la actividad, para que se pueda visualizar, ejecutar el código y entender como funciona el programa. El código fue elaborado con el lenguaje de Python y ejecutado en Google Colab.

2. Parte I

2.1. Búsqueda en Kaggle

Para la siguiente actividad, se utilizó la página web *Kaggle* [1], para buscar un dataset adecuado a la actividad; con el objetivo de poder construir un análisis estadístico robusto de un dataset elegido por el equipo, aplicando para ello técnicas y herramientas estadísticas diversas y Python, con la finalidad de conocer la estructura cuantitativa y cualitativa de dicho conjunto.

2.2. Elección de Dataset

Para la elección del dataset, se tuvo que seguir unas condiciones que tenia que contar el datset, las cuales son:

- Mínimo 5 variables numéricas.
- 2 variables categóricas.
- Una variable numérica que pueda ser considerada como la variable dependiente.
- Contenga valores faltantes en algunas variables.

tomando esto en cuenta, el dataset que se decidió utilizar para esta actividad fue **Diamond Online Marketplace**, ya que cumple con las especificaciones anteriormente mencionadas.

2.3. Descripción del Dataset

Este dataset contiene información detallada sobre más de 6400 diamantes, incluyendo sus medidas físicas (como longitud, anchura, altura y peso en quilates), características de calidad (como talla, color, claridad y fluorescencia) y precio. También incluye atributos como el tipo de certificación, las proporciones y la simetría, lo que lo convierte en un excelente recurso para analizar cómo diversos factores influyen en el precio de los diamantes.[2]

1. Shape: Forma geométrica del diamante.
2. Cut: Grado de calidad del corte del diamante.
3. Color: Grado de color del diamante de D a H.
4. Clarity: Grado de claridad basado en imperfecciones.
5. Carat Weight: Peso del diamante en quilates.
6. Length/Width Ratio: Proporción entre el largo y el ancho.
7. Depth %: Profundidad del diamante como porcentaje de su ancho.
8. Table %: Ancho de la faceta superior como porcentaje.
9. Polish: Calidad del acabado superficial del diamante.
10. Symmetry: Precisión de la forma del diamante.
11. Girdle: Grosor del borde del diamante.
12. Culet: Tamaño de la faceta inferior.
13. Length: Longitud del diamante en milímetros.
14. Width: Ancho del diamante en milímetros.
15. Height: Altura del diamante en milímetros.
16. Price: Precio del diamante en dólares estadounidenses ($).
17. Type: Certificación o tipo de origen del diamante.
18. Fluorescence: Nivel de fluorescencia UV del diamante.

3. Parte II

3.1. Separación de dataset en variable independiente y dependiente

Para poder hacer la separación de datos de variables independientes y dependientes, debemos tener en cuenta que la única variable que debemos encontrar/buscar es la dependiente (ya que esta variable va a depender de variables independientes); por ende, encontrando la variable dependiente, las demás variables serán independientes.
Y para este caso, el dataset cuenta con 18 variables, y su variable dependiente es la columna *"Price"*, ya que de esta columna se determina el valor monetario (el precio) de los diamantes, y su valor va a depender de las demás columnas, como su tamaño, corte, precisión, grosor, calidad, etc.
Por lo tanto, la variable dependiente es *"Price"* y las variables independientes son las demás columnas, y para hacer la separación de variables, utilizamos el siguiente código:

```
# Separando dataset en una parte que tenga variables independientes y otra la variable dependiente.

df_indep = df.drop(columns='Price') # Dataframe con variables independientes
df_dep = df['Price'] # Dataframe con variable independiente (Price)
```

3.2. Llenado de valores nulos (variables numéricas)

Para el llenado de valores nulos en variables numéricas, lo primero que debemos hacer es identificar las columnas que son numéricas y posteriormente ver que columnas cuentan con valores nulos. Y para hacerlo, usamos el siguiente código:

```
print(df.info()) # Informacion del dataset
display(df.isnull().sum()) # Observamos la cantidad de valores nulos por columna
```

Después de identificar las variables numéricas con valores nulos, vamos a llenar esos valores con el promedio de los datos de cada variable. Y para eso, usamos el siguiente código:

```
"""Cambiar el valor nulo (NaN), por el promedio de los datos (variables numericas)"""
df['Carat Weight'] = df_indep['Carat Weight'].fillna(df_indep['Carat Weight'].mean())
df['Length/Width Ratio'] = df_indep['Length/Width Ratio'].fillna(df_indep['Length/Width Ratio'].mean())
df['Depth %'] = df_indep['Depth %'].fillna(df_indep['Depth %'].mean())
df['Table %'] = df_indep['Table %'].fillna(df_indep['Table %'].mean())
df['Length'] = df_indep['Length'].fillna(df_indep['Length'].mean())
df['Width'] = df_indep['Width'].fillna(df_indep['Width'].mean())
df['Height'] = df_indep['Height'].fillna(df_indep['Height'].mean())
```

4. Parte III

Después de rellenar los valores nulos, con el promedio de los datos, el siguiente paso es realizar el cambio a las variables categóricas en forma numérica o dummy, con el fin de rellenar los espacios nulos.

4.1. Cambio de forma numérica las variables categóricas independientes

Para realizar el cambio de forma numérica, para las variables categóricas, primero es identificar que columnas/datos es recomendable aplicar este cambio, y posteriormente ejecutar el siguiente código:

```
# Cambiar a forma numerica las variables categoricas independientes

for col in ['Cut', 'Clarity']:
    df[col] = LabelEncoder().fit_transform(df[col])
```

Las variables categóricas *Çut"* y *Çlarity"*, se les aplico el cambio numérico, porque los datos que contienen son valores ordinales, es decir; son variables que dan orden a la información, por lo que su valor es muy importante a la hora de decidir el precio de los diamantes.

4.2. Cambio de variables dummy las variables categóricas independientes

Para realizar el cambio de forma dummy, anteriormente identificamos los motivos del cambio para cada columna, y en el caso de las variables dummies, solo datos que contienen las demás columnas a excepción de *"Price", Çut", Çlarity"*, cuentan con valores nominales, es decir; no tienen un orden natural, la información que dan no es tan clara o precisa como para asignarle un valor numérico, por lo tanto se usa dummy con el fin de arrojar valores de si o no, pero en este caso representados por 0 y 1. Para la elaboración de este cambio, se usa el siguiente código:

```
# Cambiar a variables dummy las variables categoricas independientes

df_dummy = pd.get_dummies(df, columns=['Shape', 'Color', 'Polish', '
    Symmetry', 'Girdle', 'Culet', 'Type', 'Fluorescence'], dtype=int)
```

5. Parte IV

El Análisis Estadístico Exploratorio (EDA) es un enfoque de análisis de datos que tiene como objetivo examinar, resumir y visualizar un conjunto de datos para entender sus características principales y patrones subyacentes (Tukey, 1970). Es un proceso de investigación que utiliza estadísticas de resumen y herramientas gráficas para llegar a conocer los datos

y comprender lo que se puede averiguar de ellos. [3]
La finalidad del EDA es múltiple:

- Entender la estructura y distribución de los datos
- Identificar patrones, relaciones y anomalías
- Detectar valores atípicos que podrían influir en análisis posteriores
- Generar hipótesis para investigaciones más profundas
- Facilitar la toma de decisiones informadas sobre técnicas de modelado
- Comunicar resultados de manera efectiva a diferentes audiencias

5.1. Número de valores por cada variable

Esta medida proporciona información sobre la completitud de los datos y ayuda a identificar posibles problemas de calidad. El código a utilizar es el siguiente:

```
display(df_dummy.count())
```

donde:

- **df_dummy** $\longrightarrow$ es la variable que contiene el dataset ya con los valores modificados (realizados en la Parte II y III).

Analizando los resultados (*Cuadro 1*), se observa que los resultados muestran que todas las variables tienen 6,485 observaciones completas después del procesamiento (variables dummy y codificación), lo que indica que el proceso de imputación de valores faltantes se realizó correctamente. Este nivel de integridad evita sesgos por eliminación de registros y asegura que todas las métricas reflejen la misma población de diamantes.

5.2. Medidas de tendencia central

- **Media**: Indica el valor promedio de cada variable.
- **Mediana**: Representa el valor central que divide los datos en dos mitades iguales.
- **Moda**: Muestra el(los) valor(es) más frecuente(s).

El código a utilizar es el siguiente:

```
print("Media")
display(df_dummy.mean())
print("Mediana")
display(df_dummy.median())
print("Moda")
display(df_dummy.mode())
```

Analizando los resultados (*Cuadro 2 y 3*), se observa lo siguiente:

Cut	6485
Clarity	6485
Carat Weight	6485
Length/Width Ratio	6485
Depth %	6485
Table %	6485
Length	6485
Width	6485
Height	6485
Price	6485
Shape_Round	6485
Shape_Oval	6485
Type_GIA	6485
Type_GIA Lab-Grown	6485
Type_IGI Lab-Grown	6485
Fluorescence_Faint	6485
Fluorescence_Medium	6485
Fluorescence_Strong	6485

Cuadro 1: Número de valores por cada variable

- **Cut (Media = 3.27, Mediana = 4.0)**: La diferencia entre media y mediana (3.27 vs 4.0) indica que la distribución está sesgada negativamente (hacia la izquierda), esto significa que hay más diamantes con calificaciones de corte altas (valores cercanos a 4) que con calificaciones bajas y en términos prácticos, la mayoría de diamantes en el dataset tienen cortes de buena a excelente calidad. El sesgo negativo sugiere que hay algunos diamantes con cortes de menor calidad que "jalan" la media hacia abajo.
- **Clarity (Media = 2.84, Mediana = 2.0)**: Aquí la media es mayor que la mediana (2.84 vs 2.0), indicando sesgo positivo (hacia la derecha), esto significa que aunque la mayoría de diamantes tienen claridad en niveles bajos-medios (alrededor de 2), hay algunos diamantes con claridad muy alta que elevan el promedio. En términos prácticos: la mayoría de diamantes tienen claridad moderada, pero existe una cola de diamantes con claridad excepcional.
- **Carat Weight (Media = 1.24, Mediana = 1.04)**: La media es mayor que la mediana (1.24 vs 1.04), indicando distribución sesgada positivamente. Esto es típico en el mercado de diamantes, ya que hay muchos diamantes pequeños (cerca de 1 quilate) pero algunos diamantes grandes que elevan el promedio, la diferencia de 0.2 quilates sugiere la presencia de diamantes significativamente más grandes que el valor central.
- **Moda (0 para la mayoría de variables dummy)**: Esto indica que para cada característica específica (como "Shape_Round"), la categoría no está presente en la

mayoría de observaciones, y por esta razón se entiende, porque cada variable dummy representa una sola categoría de una variable multi-categórica. Por ejemplo, si "Shape_Round"tiene moda = 0, significa que la mayoría de diamantes NO son redondos.

Media	
Cut	3.270933
Clarity	2.835929
Carat Weight	1.235772
Length/Width Ratio	1.329548
Depth %	64.206499
Table %	59.223301
Length	6.848149
Width	5.255819
Height	4.196134
Price	4184.208139
Type_GIA Lab-Grown	0.193369
Type_IGI Lab-Grown	0.289283
Fluorescence_Faint	0.107170
Fluorescence_Medium	0.047957
Fluorescence_Strong	0.033153

Cuadro 2: Medidas de tendencia central - Media

Mediana	
Cut	4.00
Clarity	2.00
Carat Weight	1.04
Length/Width Ratio	1.25
Depth %	64.20
Table %	59.00
Length	6.72
Width	5.20
Height	4.13
Price	2640.00
Type_GIA Lab-Grown	0.00
Type_IGI Lab-Grown	0.00
Fluorescence_Faint	0.00
Fluorescence_Medium	0.00
Fluorescence_Strong	0.00

Cuadro 3: Medidas de tendencia central - Mediana

5.3. Medidas de dispersión

- **Varianza y desviación estándar**: Indican qué tan dispersos están los datos respecto a la media.
- **Rango**: Muestra la amplitud total de los datos.
- **Coeficiente de variación**: Permite comparar la variabilidad relativa entre variables con diferentes escalas.

El código a utilizar es el siguiente:

```
print("Varianza")
display(df_dummy.var())
print("Desviacion estandar")
display(df_dummy.std())
print("Rango")
display(df_dummy.max() - df_dummy.min())
print("Coeficiente de variacion (%)")
display((df_dummy.std() / df_dummy.mean()*100).round(2))
```

Analizando los resultados (*Cuadro 4, 5, 6 y 7*), observamos que las medidas de dispersión revelan que "Depth %" y "Table %" son estrechas (desviaciones estándar 4.77 y 3.41 respectivamente) en relación con su media, lo cual es deseable porque refleja poca variabilidad en proporciones críticas de tallado. Por otro lado, "Price" presenta una desviación estándar de 6344 y un coeficiente de variación de 151 %, lo que evidencia una enorme variabilidad absoluta y relativa en precios: el mercado contiene diamantes de valores muy diversos, desde unos pocos cientos hasta decenas de miles de dólares. El rango de "Price" (1010 – 41395) refuerza esta dispersión extrema.

Varianza	
Cut	1.177201
Clarity	1.332114
Carat Weight	0.256493
Length/Width Ratio	0.106003
Depth %	22.758766
Table %	11.619000
Length	1.771511
Width	0.714045
Height	0.745912
Price	40257828.49
Type_GIA Lab-Grown	0.156002
Type_IGI Lab-Grown	0.205630
Fluorescence_Faint	0.095700
Fluorescence_Medium	0.045664
Fluorescence_Strong	0.032059

Cuadro 4: Medidas de dispersión – Varianza

Desviación estándar	
Cut	1.084989
Clarity	1.154173
Carat Weight	0.506452
Length/Width Ratio	0.325581
Depth %	4.770615
Table %	3.408013
Length	1.331282
Width	0.845027
Height	0.863888
Price	6344.755662
Type_GIA Lab-Grown	0.394970
Type_IGI Lab-Grown	0.453464
Fluorescence_Faint	0.309354
Fluorescence_Medium	0.213691
Fluorescence_Strong	0.179051

Cuadro 5: Medidas de dispersión – Desviación estándar

Rango (max – min)	
Cut	4.00
Clarity	6.00
Carat Weight	8.38
Length/Width Ratio	1.58
Depth %	30.90
Table %	75.00
Length	15.00
Width	10.00
Height	10.00
Price	41395.00

Cuadro 6: Medidas de dispersión – Rango

Coeficiente de variación (%)	
Cut	33.17
Clarity	40.70
Carat Weight	40.98
Length/Width Ratio	24.49
Depth %	7.43
Table %	5.75
Length	19.43
Width	16.08
Height	20.97
Price	151.47
Type_GIA Lab-Grown	204.26
Type_IGI Lab-Grown	156.75
Fluorescence_Faint	288.66
Fluorescence_Medium	445.59
Fluorescence_Strong	540.07

Cuadro 7: Medidas de dispersión – Coeficiente de variación

5.4. Medidas de posición

Los cuartiles dividen los datos en cuatro partes iguales, permitiendo entender la distribución de los valores. El código a utilizar es el siguiente:

```
print("Cuartiles")
display(df_dummy.quantile([0.25, 0.5, 0.75]))
```

Cuartiles			
Variable	**Q1 (25 %)**	**Q2 (50 %)**	**Q3 (75 %)**
Cut	2.00	4.00	4.00
Clarity	2.00	2.00	3.00
Carat Weight	1.01	1.03	1.20
Length/Width Ratio	1.02	1.27	1.51
Depth %	61.20	63.10	67.60
Table %	58.00	61.00	65.00
Length	6.33	7.42	8.64
Width	5.39	5.64	6.58
Height	3.46	3.67	4.10
Price	1210.00	3320.00	4390.00

Cuadro 8: Medidas de posición – Cuartiles

Analizando los resultados (*Cuadro 8*), los cuartiles de "Price" (1210, 3320 y 4390) muestran que el 50 % de los diamantes cuesta menos de 3320, y sólo el 25 % supera los

4390. Esta concentración en precios bajos explica el sesgo positivo de 2.15 y la curtosis alta de 15.44; por lo tanto, la distribución es asimétrica con una cola larga hacia valores elevados y picos agudos en la parte inferior.

5.5. Sesgo y curtosis

- **Sesgo**: Mide la asimetría de la distribución. Un sesgo positivo indica una cola más larga hacia valores altos, mientras que un sesgo negativo indica lo contrario.
- **Curtosis**: Mide la concentración de datos alrededor de la media y el "peso"de las colas. Valores positivos indican distribuciones más puntiagudas, mientras que valores negativos indican distribuciones más aplanadas.

El código a utilizar es el siguiente:

```
print("Sesgo")
display(df_dummy.skew())
print("Curtosis")
display(df_dummy.kurtosis())
```

Analizando los resultados (*Cuadro 9 y 10*), vemos que la variable "Carat Weight" exhibe un sesgo muy marcado (4.43) y curtosis extremadamente elevada (32.25), lo que indica que la mayoría de los diamantes pesan menos de 1 ct, pero existen unos pocos ejemplares muy pesados que deforman la distribución y crean colas pesadas. En contraste, "Cut" y "Clarity" tienen sesgos negativos o moderados (-1.02 y 0.86) y curtosis cercana a cero, lo cual significa que sus distribuciones son más planas o ligeramente asimétricas sin valores atípicos extremos.

Sesgo	
Cut	-1.015087
Clarity	0.860936
Carat Weight	4.426270
Length/Width Ratio	0.821662
Depth %	0.361669
Table %	-0.056857
Length	0.302189
Width	0.386690
Height	0.431127
Price	2.147104
Type_GIA Lab-Grown	1.553155
Type_IGI Lab-Grown	0.929652
Fluorescence_Faint	2.540464
Fluorescence_Medium	4.232109
Fluorescence_Strong	5.216289

Cuadro 9: Medidas de forma – Sesgo

Curtosis	
Cut	-0.587725
Clarity	-0.520118
Carat Weight	32.252981
Length/Width Ratio	-0.269737
Depth %	-0.302364
Table %	0.085284
Length	-0.155237
Width	-0.293283
Height	-0.330445
Price	15.441257
Type_GIA Lab-Grown	0.412419
Type_IGI Lab-Grown	-1.136098
Fluorescence_Faint	4.455329
Fluorescence_Medium	15.915656
Fluorescence_Strong	25.217445

Cuadro 10: Medidas de forma – Curtosis

5.6. Correlación entre pares de variables cuantitativas

La matriz de correlación identifica relaciones lineales entre pares de variables cuantitativas, ayudando a entender cómo se relacionan las características de los diamantes.

Analizando los resultados (*Figura 1*), se revela varias relaciones interesantes. En primer lugar, "Carat Weight" guarda una fuerte correlación positiva con "Length" y "Width" (colores rojizos cercanos a 0.8 – 0.85), lo cual es lógico ya que los diamantes más pesados suelen ser físicamente más grandes. Por su parte, la variable "Depth %" aparece fuertemente correlacionada negativamente con "Length" y "Carat Weight" (tonos azulados de alrededor de –0.5), lo que sugiere que a medida que el porcentaje de profundidad aumenta, el peso y el tamaño horizontal disminuyen. La diagonal roja intensa muestra la correlación perfecta de cada variable consigo misma, mientras que los cuadros casi grises entre "Cut" y "Price" señalan una correlación muy débil (cercana a 0.1), confirmando que el peso y las dimensiones ejercen un papel más determinante en el precio que la calidad del corte.

5.7. Diagramas de caja entre variables categóricas y cuantitativas

Los diagramas de caja, muestran la distribución de variables cuantitativas según categorías, identificando valores atípicos y diferencias entre grupos.

Analizando los resultados (*Figura 2*), observamos que los precios según la variable Shape_Round muestra con claridad que los diamantes redondos (Shape_Round = 1) tienden a agruparse en un rango de valores más estrecho y, en general, con un precio medio ligeramente inferior comparado con los diamantes de otras formas (Shape_Round = 0). Mientras

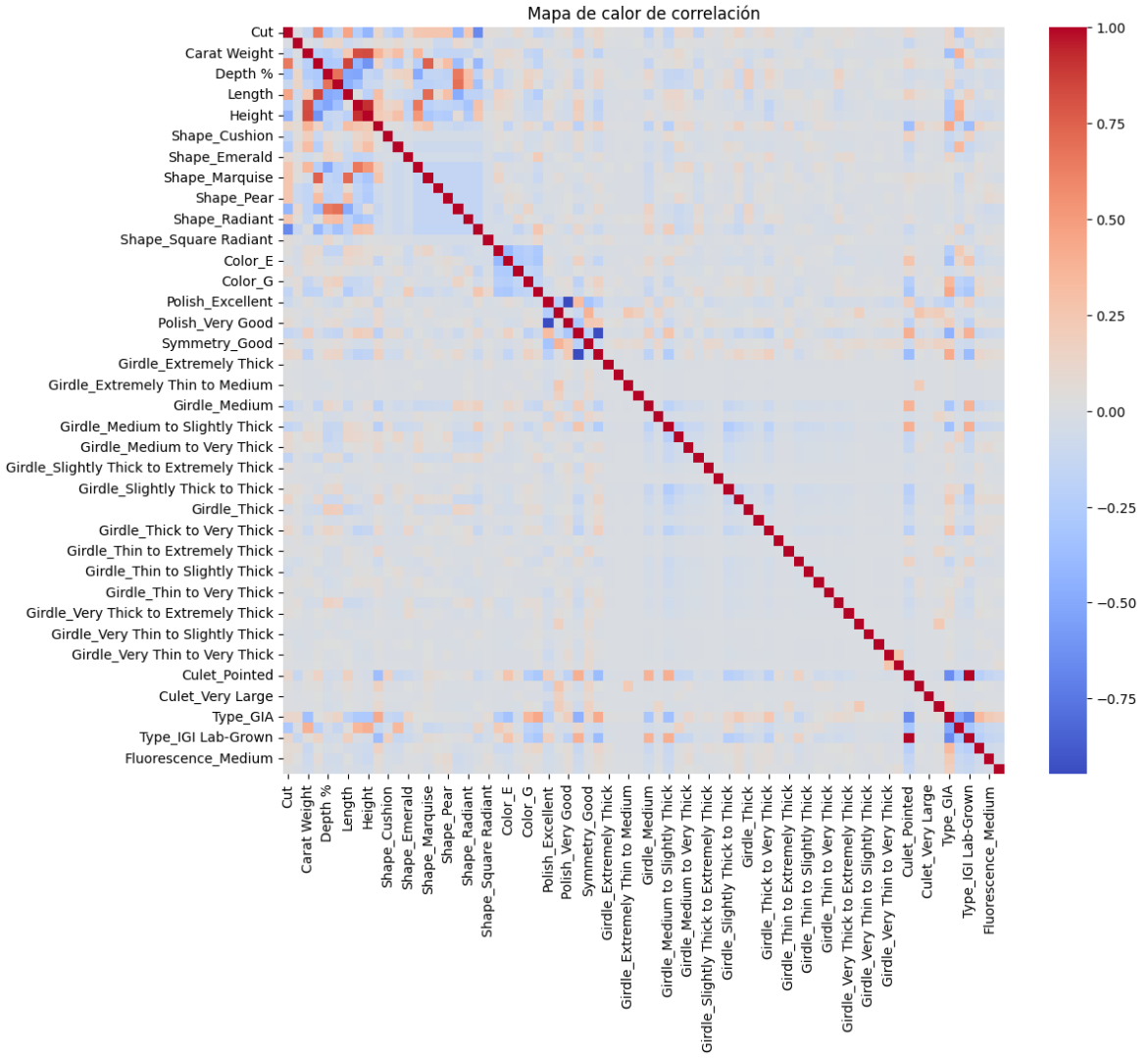

Figura 1: Gráfico de correlación

ambos grupos presentan valores atípicos elevados (puntos por encima del bigote superior), los diamantes no redondos (0) muestran una cola de valores extremos mucho más pronunciada, con múltiples ejemplares por encima de 20000 USD, e incluso cerca de 40000 USD. Esto indica que las formas distintas de "round" incluyen algunos ejemplares muy caros que no se observan –o se observan con menor frecuencia– entre los diamantes redondos.

5.8. Tablas de distribución de frecuencias

Las tablas de frecuencia, proporcionan un resumen cuantitativo de la distribución de datos categóricos El código a utilizar es el siguiente:

```
x = df_dep
n = len(x)
k = int(1 + 3.322 * np.log10(n))
```

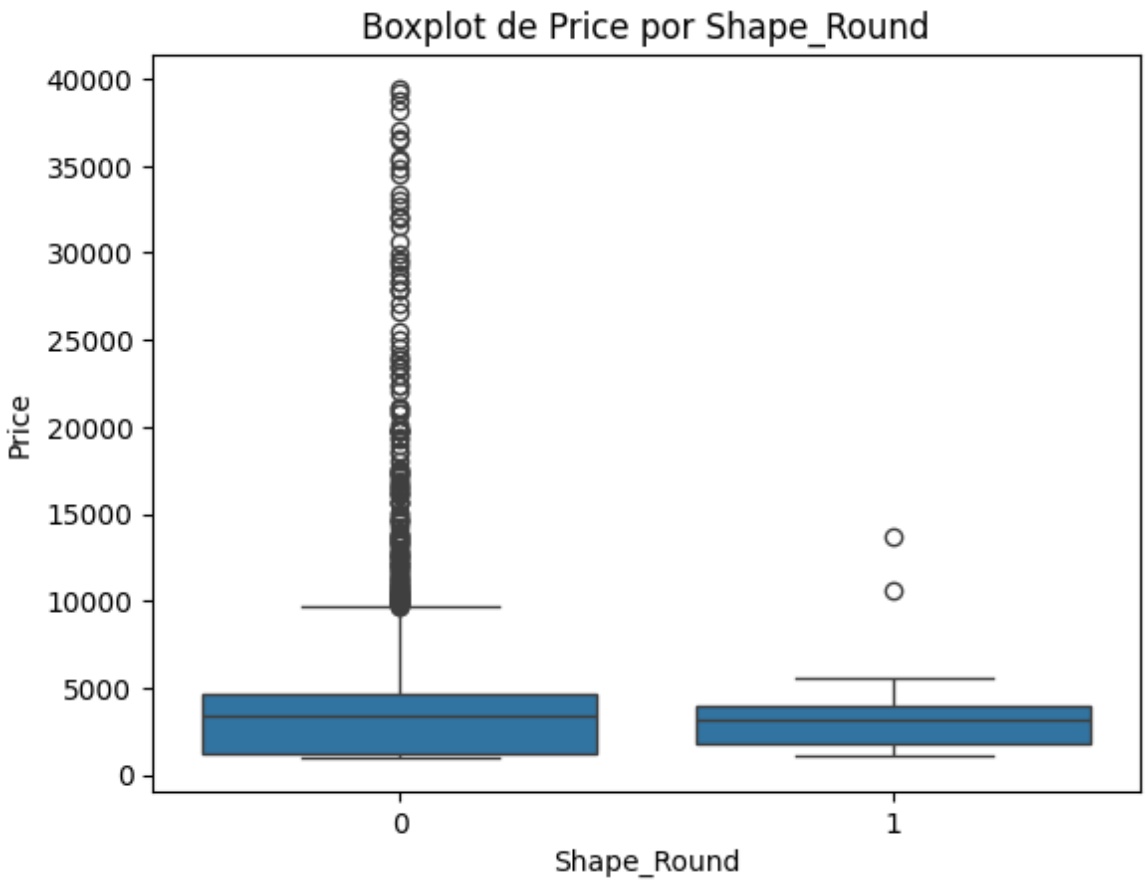

Figura 2: Diagrama de caja

```
rango = x.max() - x.min()
amplitud = round(rango / k, 2)
bins = np.arange(x.min(), x.max() + amplitud, amplitud)

# Asignar intervalos
x_cut = pd.cut(x, bins=bins, right=False)
tabla_freq = x_cut.value_counts().sort_index()
marca_clase = [(interval.left + interval.right) / 2 for interval in
    tabla_freq.index]
freq_abs = tabla_freq.values
freq_rel = np.round(freq_abs / n, 4)
freq_acum = np.cumsum(freq_abs)
freq_rel_acum = np.cumsum(freq_rel)

tabla_frecuencias = pd.DataFrame({
    "Intervalo": tabla_freq.index.astype(str),
    "Marca de clase": marca_clase,
    "Frecuencia": freq_abs,
    "Frecuencia relativa": freq_rel,
    "Frecuencia acumulada": freq_acum,
```

```
    "Frecuencia relativa acumulada": freq_rel_acum
})
display(tabla_frecuencias)
```

Analizando los resultados (*Cuadro 11*), podemos observar que la tabla de frecuencias muestra que el 64.49 % de los diamantes se ubica en el primer intervalo de precio ([1010; 3967)), y sólo el 0.02 % en el último ([39460; 42418)). Este patrón confirma una distribución sesgada a la derecha, donde unos pocos diamantes de alto valor marcan la cola larga.

Distribución de frecuencias - Price					
Intervalo	Marca de clase	Frecuencia	Freq. relativa	Freq. acumulada	Freq. rel. acum.
[1010.0, 3967.69)	2488.845	4182	0.6449	4182	0.6449
[3967.69, 6925.38)	5446.535	1884	0.2905	6066	0.9354
[6925.38, 9883.07)	8404.225	236	0.0364	6302	0.9718
[9883.07, 12840.76)	11361.915	78	0.0120	6380	0.9838
[12840.76, 15798.45)	14319.605	35	0.0054	6415	0.9892
[15798.45, 18756.14)	17277.295	25	0.0039	6440	0.9931
[18756.14, 21713.83)	20234.985	14	0.0022	6454	0.9953
[21713.83, 24671.52)	23192.675	10	0.0015	6464	0.9968
[24671.52, 27629.21)	26150.365	6	0.0009	6470	0.9977
[27629.21, 30586.9)	29108.055	6	0.0009	6476	0.9986
[30586.9, 33544.59)	32065.745	3	0.0005	6479	0.9991
[33544.59, 36502.28)	35023.435	3	0.0005	6482	0.9996
[36502.28, 39459.97)	37981.125	2	0.0003	6484	0.9999
[39459.97, 42417.66)	40938.815	1	0.0002	6485	1.0000

Cuadro 11: Tabla de distribución de frecuencias para Price

5.9. Histogramas

El histograma, revela la forma de la distribución de cada variable. Y analizando los resultados (*Figura 3*), se visualiza que los precios exhibe una distribución altamente sesgada a la derecha: la mayoría de los diamantes (más de 4000 observaciones) se concentran en la franja de precios bajos, por debajo de 5000 USD, y luego la frecuencia cae de forma abrupta. Apenas un pequeño porcentaje de diamantes supera los 10000 USD y únicamente unos pocos llegan a 40000 USD. La forma de campana asimétrica del histograma coincide con el sesgo positivo y la curtosis elevada identificados en las tablas, mostrando una "cola" larga de precios altos y un pico muy marcado en los valores menores. Todas estas visualizaciones confirman que el mercado de diamantes en este dataset está dominado por productos de precio moderado, con unos pocos ejemplares de lujo que generan los valores extremos en las distribuciones.

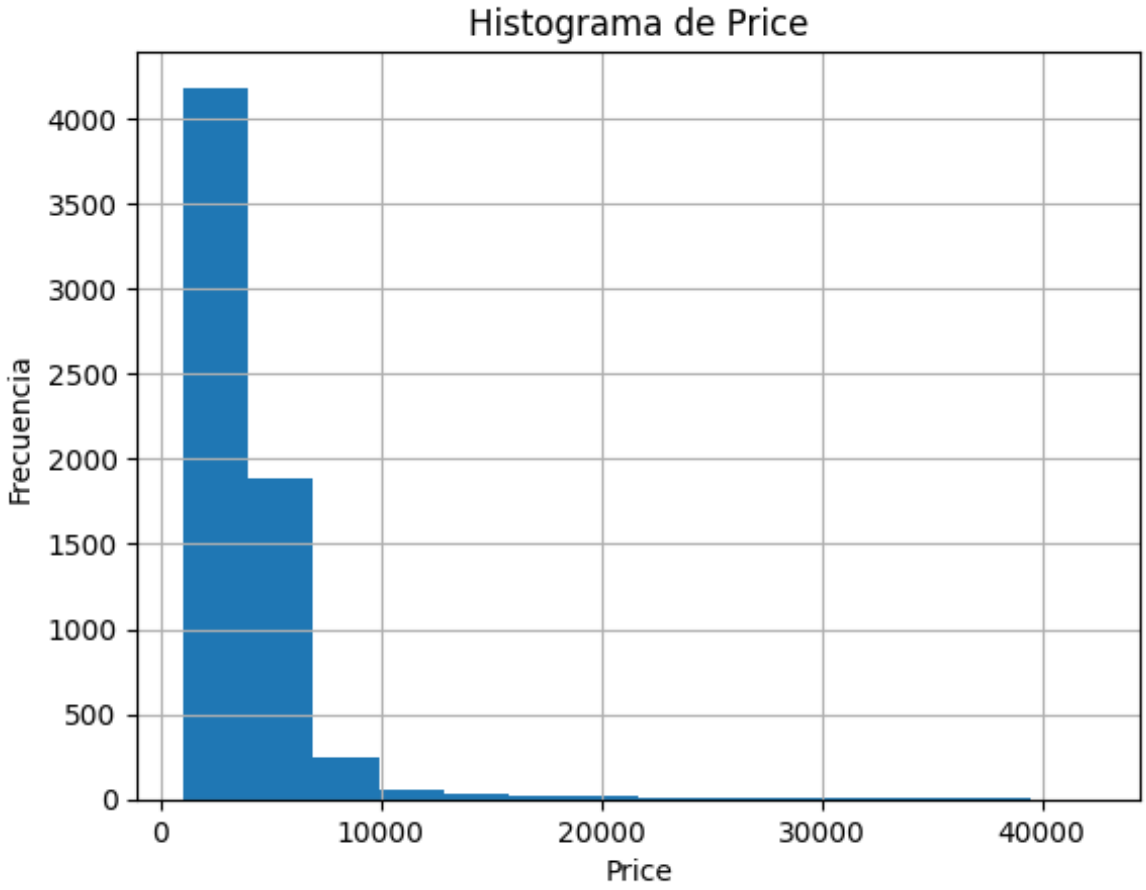

Figura 3: Histograma

6. Conclusión

En conclusión, podemos determinar que el análisis revela como las variables relacionadas con la calidad del diamante (corte y claridad) muestran cierta variabilidad y aportan información valiosa, son el peso en quilates y las dimensiones físicas las que ejercen la influencia más decisiva sobre el precio. Ya que la distribución sesgada de los precios, con una concentración de la mayoría de las piezas en rangos bajos y una cola larga de ejemplares de alto valor, muestra la situación del mercado, demostrando que: unos pocos diamantes de gran tamaño y rareza elevan significativamente los valores máximos, mientras que la mayoría mantiene precios moderados. Además, de que las correlaciones fuertes entre peso y tamaño confirman que estas características van de la mano, mientras que la influencia débil del corte sobre el precio sugiere que, para la mayoría de los consumidores, el peso y la apariencia física dominan la percepción de valor.

En resumen, este estudio EDA (Análisis Estadístico Exploratorio) muestra que, en este dataset, el precio de los diamantes depende fundamentalmente de su peso y dimensiones, con la calidad del corte y la claridad aportando un segundo nivel de diferenciación.

Referencias

[1] https://www.kaggle.com/

[2] beridzeg45. (2025). Diamond Online Marketplace [Conjunto de datos]. Kaggle. https://www.kaggle.com/datasets/beridzeg45/diamonds-prices-prediction

[3] Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley.