Brilliant Deluxe

2023

Realizado por:

MANUEL JOSÉ LÓPEZ MUÑOZ

ÍNDICE

Introducción

DESCRIPCIÓN DE LA EMPRESA

Brilliance Deluxe es una prestigiosa empresa especializada en la venta de diamantes de alta calidad. Con una sólida reputación en el mercado internacional, la empresa se distingue por su amplia gama de diamantes exclusivos, que atraen tanto a coleccionistas privados como a diseñadores de joyas de élite.

OBJETIVO DE BRILLIANCE DELUXE

Brilliance Deluxe busca optimizar su estrategia de precios y de adquisiciones. El objetivo principal es desarrollar un modelo que pueda predecir el precio de los diamantes con precisión, teniendo en cuenta diversas características tales como el quilate, la profundidad, la mesa, el color, la claridad y la calidad del corte. Esta herramienta analítica servirá para dos propósitos principales:

- **Estrategia de precios:**
  - Ajustar los precios de venta de manera dinámica, tomando como base las características de cada diamante.
  - Identificar oportunidades de inversión en diamantes, cuyas características sugieran un valor de reventa más alto.
- **Selección y Adquisición:**
  - Ayudar en la toma de decisiones para la compra de nuevos diamantes, seleccionando los que ofrezcan el mejor retorno de inversión.
  - Determinar qué características de los diamantes influyen más en su valor en el mercado.

RESUMEN DEL DATASET

El Dataset utilizado para este análisis contiene una muestra total de 53.940 diamantes e incluye las siguientes variables:

- **Carat (quilates)**: Es la unidad de medida del peso del diamante, un quilate equivale a 200 miligramos.
- **Depth (profundidad)**: Se refiere a la altura del diamante, se expresa como porcentaje del ancho del diamante.
- **Table (mesa)**: Es la parte superior plana del diamante, se expresa como porcentaje del ancho del diamante.
- **Precio**: Es el precio del diamante.
- **xmm**: Es la longitud del diamante en milímetros.
- **ymm**: Es el ancho del diamante en milímetros.
- **zmm**: Es la profundidad del diamante en milímetros.

INTERVALOS AÑADIDOS

El dataset original incluía tres variables relacionadas con la claridad, la calidad del corte y el color del diamante. Estas variables estaban marcadas con las letras correspondientes al intervalo en el que se encuentran.

Por ello, al tratarse de clasificaciones clasificadas de mejor a peor calidad, hemos sustituido las letras por números correspondientes a cada intervalo.

Escala de color:

- 1(D): Incoloro, sin tintes de color; es el grado de color más alto y es extremadamente raro.
- 2 (E): Incoloro. Sólo minúsculas trazas de color pueden ser detectadas por un gemólogo experto.
- 3 (F): Incoloro. Una pequeña diferencia con E, pero aún considerado incoloro
- 4 (G-H): Casi incoloro. El color es difícil de detectar a menos que se compare lado a lado con diamantes de mejor calidad.
- 5 (I - J): Casi incoloro. Un ligero tinte de color, pero todavía es un diamante de calidad.
- 6 (K-M): Ligeramente tintado, Un tinte visible de color.
- 7 (N-R): Muy ligeramente tintado. Normalmente un tinte amarillo o marrón visible.
- 8 (S-Z): Tintado. El color es notablemente visible.

Escala de claridad:

- **1 (FL o Flawless)**: Sin inclusiones ni defectos visibles bajo una ampliación de 10x. Es extremadamente raro.
- **2 (IF o Internally Flawless)**: Sin inclusiones visibles bajo una ampliación de 10x, pero pueden tener algunas imperfecciones superficiales.
- **3 (VVS1-VVS2 o Very, Very Slightly Included)**: Inclusiones muy pequeñas que son difíciles de ver, incluso bajo una ampliación de 10x.
- **4 (VS1-VS2 o Very Slightly Included)**: Inclusiones pequeñas que son algo fáciles de ver bajo una ampliación de 10x, pero generalmente no a simple vista.
- **5 (SI1-SI2 o Slightly Included)**: Inclusiones que son visibles bajo una ampliación de 10x y a veces a simple vista
- **6 (I1-I2-I3 o Included)**: Inclusiones y/o defectos que son obvios bajo una ampliación de 10x y generalmente visibles a simple vista. Pueden afectar la transparencia y brillo del diamante.

Escala de calidad del corte:

- **1 (Ideal):** Este es considerado el corte de mayor calidad. Un diamante con un corte ideal reflejará casi toda la luz que entra en él y, por lo tanto, tendrá un brillo excepcional. Los diamantes con corte ideal tienen proporciones precisas y están bien pulidos
- **2 (Premium)**: Un corte premium reflejará una gran cantidad de luz y tendrá un brillo muy bueno. En la escala, "Premium" se ubicaría entre "Ideal" y "Muy Bueno".
- **3 (Very Good)**: Refleja gran parte de la luz que entra. Estos diamantes tienen proporciones muy similares a las de un corte excelente, pero a un costo menor.
- **4 Bueno (Good)**: Refleja la mayoría de la luz que entra. Estos diamantes tienen proporciones que están fuera del rango ideal, pero aún así son de calidad y más asequibles que los cortes de mayor grado.
- **5 (Fair)**: Refleja menos luz que los cortes de mejor calidad y puede parecer menos brillante. A menudo tiene proporciones que están significativamente fuera del rango ideal.

Limpieza y preparación de datos

TRATAMIENTO DE VALORES ATÍPICOS O NULOS

Teniendo en cuenta el objetivo, nuestra estrategia de manejo de datos debe orientarse a preservar la máxima cantidad de información válida y relevante.

Por ello, realizaremos un análisis básico para el conjunto con datos atípicos y posteriormente un análisis más profundo con estos eliminados, ya que, los diamantes con características atípicas pueden representar oportunidades de mercado importantes o tendencias interesantes (por ejemplo, diamantes muy grandes o de calidad única).

En el caso de los datos nulos o con valores erróneos, hemos procedido a eliminarlos, ya que eran claramente errores.

IDENTIFICACIÓN INTEGRAL DE VALORES ATÍPICOS

Mediante el análisis de cajas y bigotes y el cálculo del Rango Intercuartílico (IQR), identificamos una cantidad significativa de valores atípicos en varias variables.

Establecimos intervalos superiores e inferiores basados en el IQR para las variables:

- carat
- depth
- table
- precio
- x
- y
- z

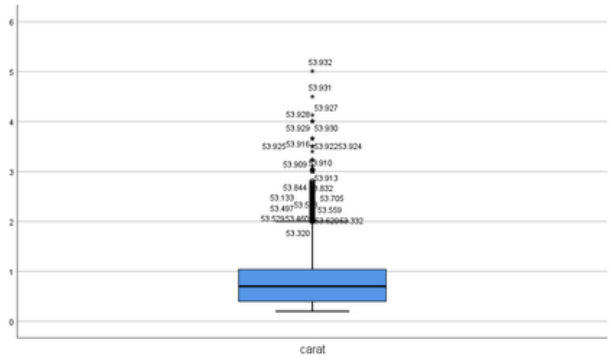
Posteriormente creamos indicadores para cada variable, señalando la presencia de valores atípicos.

RESULTADOS DE LA IDENTIFICACIÓN

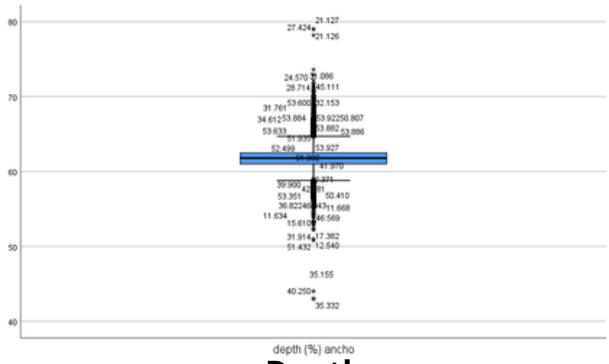
Los valores atípicos varían significativamente entre las variables, con la mayor cantidad observada en depth y la menor en z. Algunos registros contienen valores atípicos en múltiples variables, lo que indica la presencia de casos extremos.

GRÁFICOS DE CAJAS Y BIGOTES

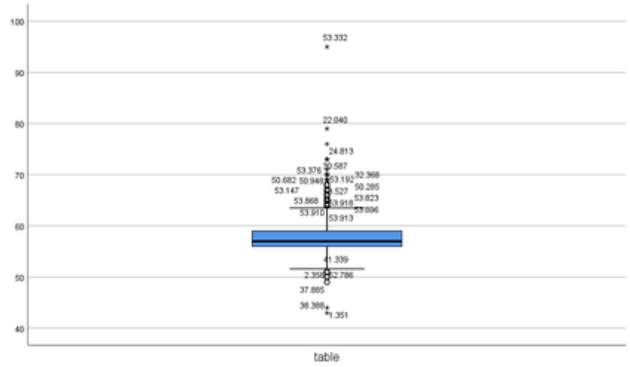
A continuación mostraremos los gráficos de cajas y bigotes que muestran los casos atípicos en cada variable.



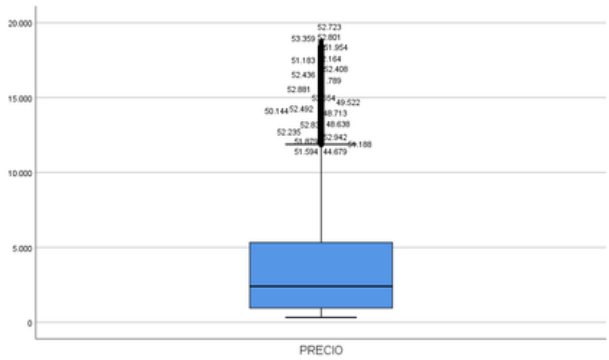
Carat



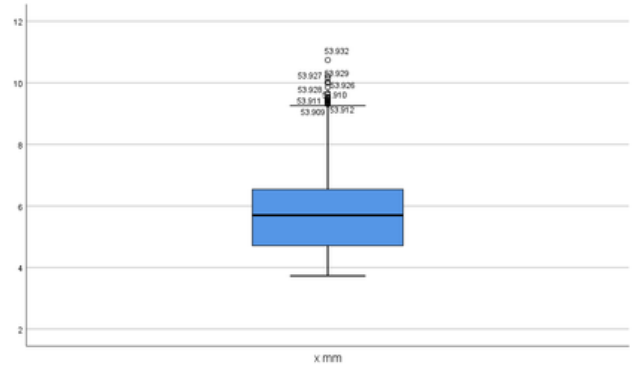
Depth



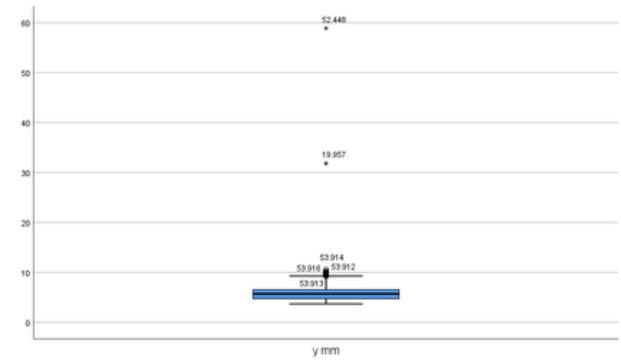
Table



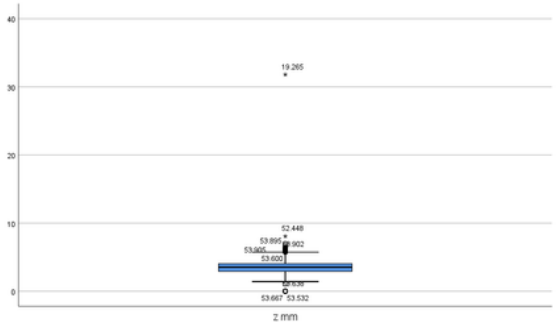
Precio



X



Y



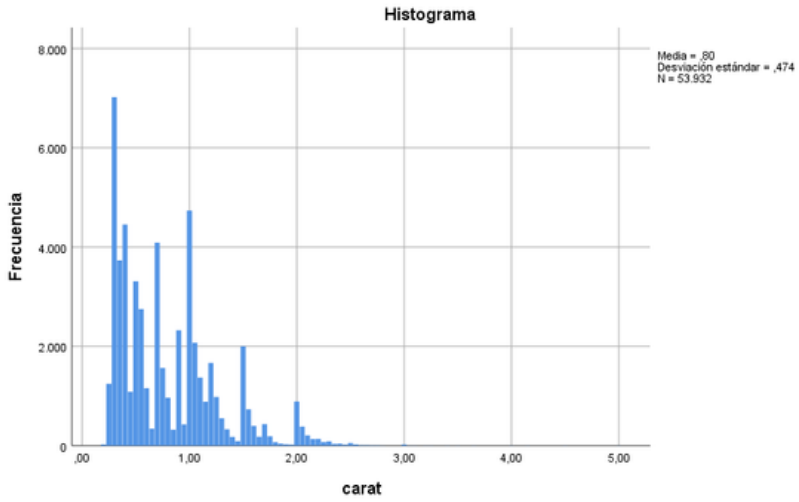
Z

Análisis descriptivo (Todos los datos)

Análisis por variable

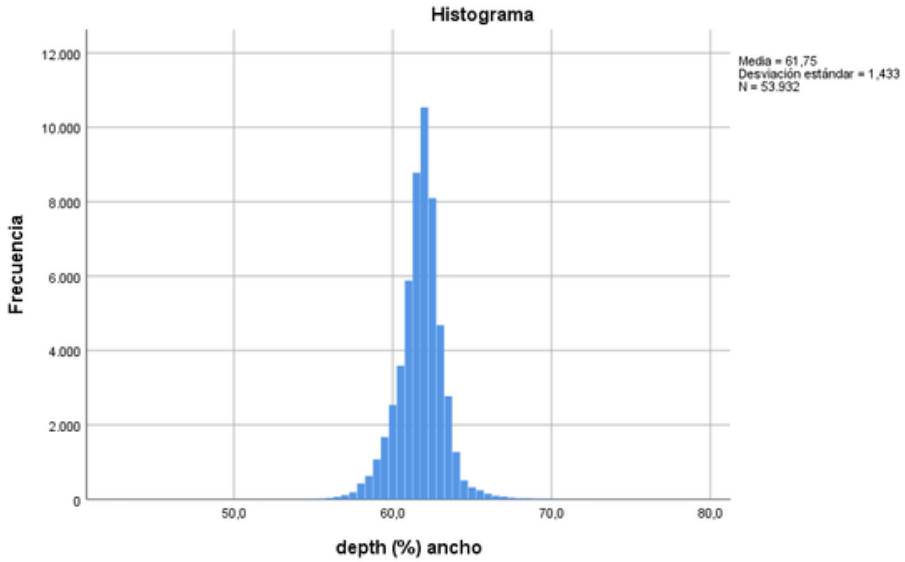
Quilates (carat):

- Media: 0.80
- Mediana: 0.70
- Rango: 0.20 a 5.01
- Desviación Estándar: 0.474
- Distribución: Ligera asimetría hacia la derecha.



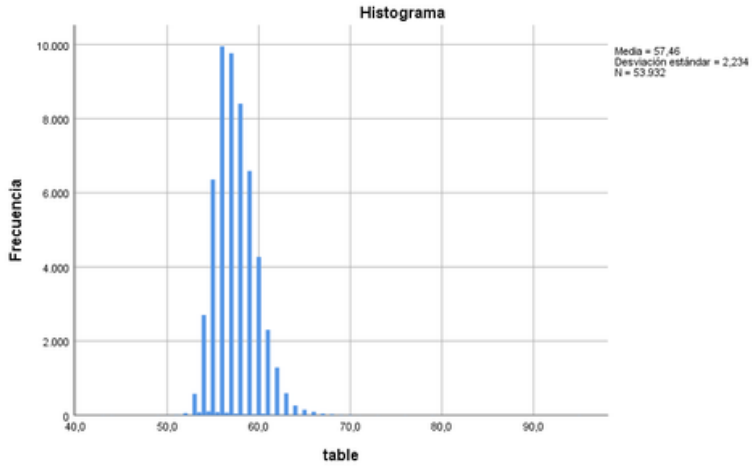
Profundidad (Depth %)

- Media: 61.75%
- Mediana: 61.8%
- Rango: 43.0% a 79.0%
- Desviación Estándar: 1.43%
- Distribución: Relativamente simétrica, con ligera tendencia a valores menores.



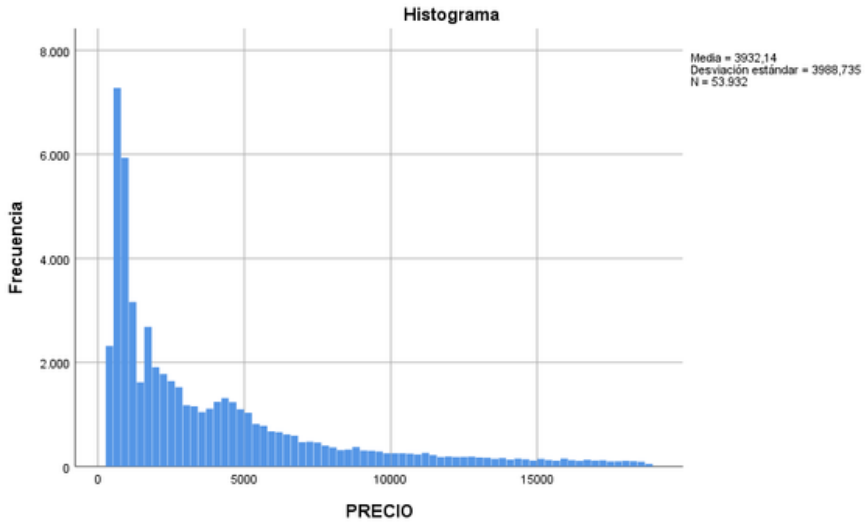
Mesa (table):

- Media: 57.46
- Mediana: 57.0
- Rango: 43 a 95
- Desviación Estándar: 2.23
- Distribución: Hay una mayor concentración de datos en los valores menores de la mesa



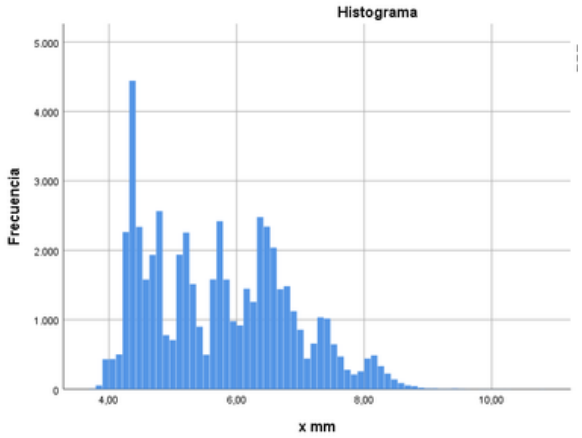
Precio

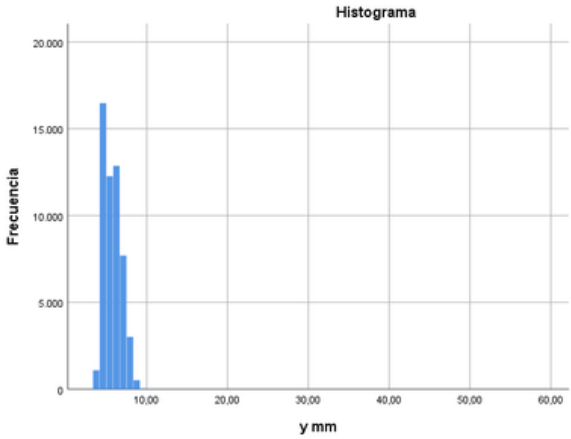
- Media: $3,932.14
- Mediana: $2,401
- Rango: $326 a $18,823
- Desviación Estándar: $3,988.74
- Distribución: Asimétrica, indicando una concentración de diamantes en rangos de precio más bajos.

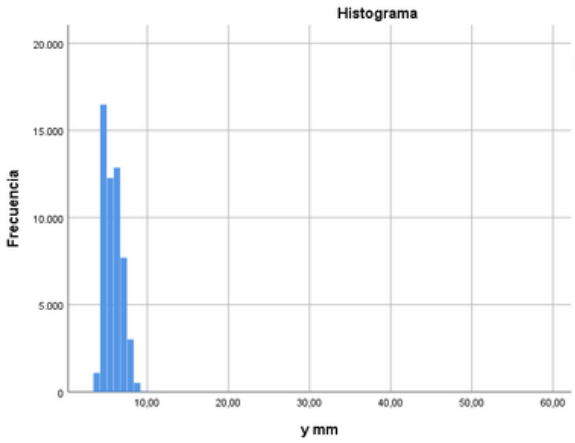


Dimensiones (x, y, z):

- x (mm): Media = 5.73, Mediana = 5.70
- y (mm): Media = 5.74, Mediana = 5.71
- z (mm): Media = 3.54, Mediana = 3.53
- Distribución: La variable 'y' muestra una asimetría notable y curtosis alta, indicando la presencia de valores extremos.

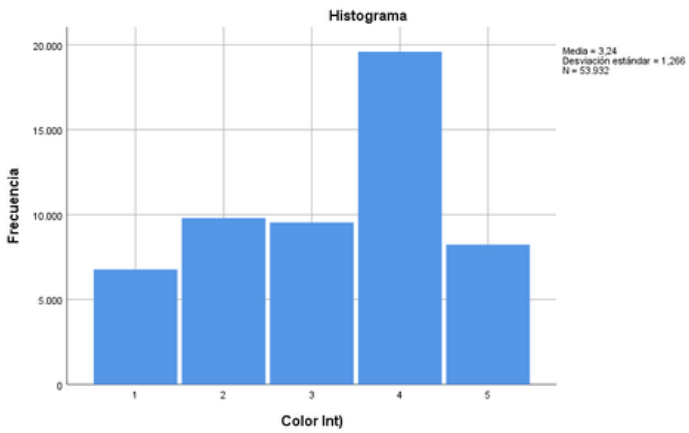






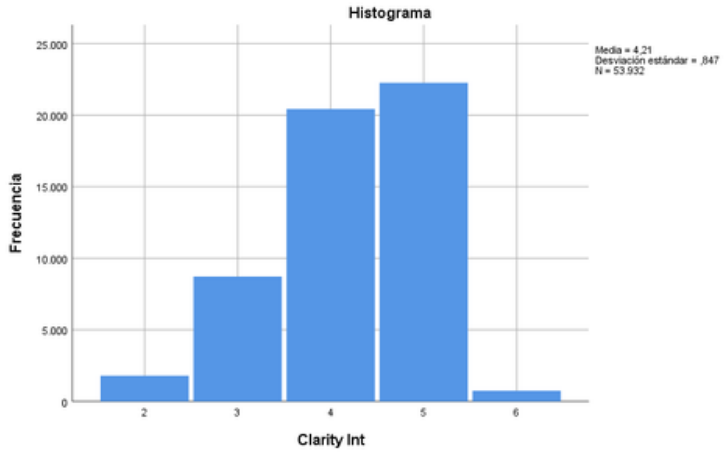
Color

- Media: 3.24 (en una escala de 1 a 5)
- Mediana: 4
- Distribución: Ligeramente asimétrica hacia valores inferiores.



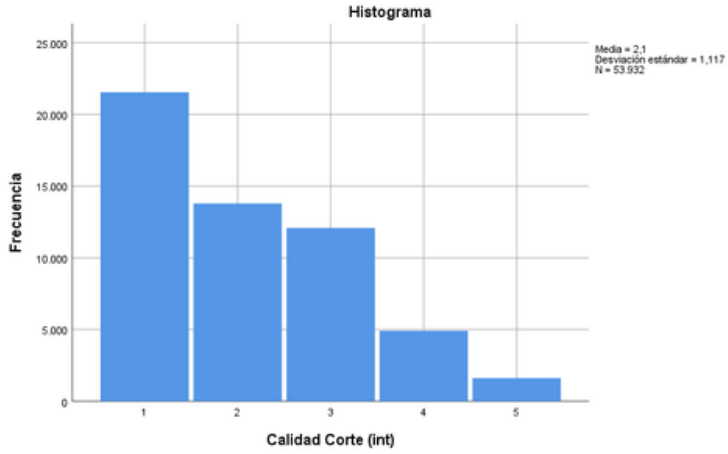
Claridad:

- Media: 4.21 (en una escala de 2 a 6)
- Mediana: 4
- Distribución: Ligeramente asimétrica hacia valores inferiores.



Calidad del Corte:

- Media: 2.10 (en una escala de 1 a 5)
- Mediana: 2
- Distribución: Asimétrica hacia valores mayores.



Observaciones Generales:

- **Valores Atípicos**: La presencia de valores atípicos en varias variables sugiere una diversidad en la calidad y características de los diamantes. Esto es especialmente notable en las dimensiones y el precio.

- **Distribuciones**: La mayoría de las variables muestran cierta asimetría y varían en curtosis, lo que sugiere diferencias en la distribución de los datos.

Análisis descriptivo (Datos atípicos eliminados)

Quilates (carat):

Media: La media de 0.6183 quilates sugiere que, en promedio, los diamantes de la muestra son de tamaño modesto. Podemos observar la diferencia de media al análisis con datos atípicos que era de casi 0.8, aún así, con un coeficiente de variación del 48,39% la media no es representativa.

Mediana: La mediana de 0.53 quilates, al ser menor que la media, indica una tendencia hacia la presencia de más diamantes pequeños en la muestra. Esto sugiere que más de la mitad de los diamantes tienen menos de 0.53 quilates.

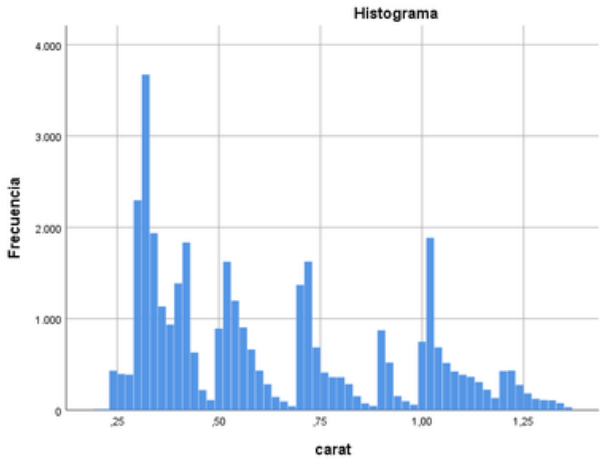
Varianza y desviación estándar: La varianza es de 0,089 y la desviación estándar es de 0,29913, lo que indica la dispersión de los datos alrededor de la media. El coeficiente de variación es de 48,39%

Asimetría: Una asimetría positiva de 0.628 refleja un sesgo en la distribución hacia diamantes más pequeños. Esto se traduce en que hay una cola más larga en el lado derecho del histograma, lo que indica una mayor concentración de diamantes en tamaños más pequeños y menos en tamaños más grandes.

Curtosis: Una curtosis de -0.900, siendo negativa, señala que la distribución es más plana en comparación con una distribución normal. Esto implica que los valores están más distribuidos y no se concentran tanto alrededor de la media.

Observación visual: El histograma muestra claramente múltiples picos o modas, lo que sugiere la presencia de múltiples subgrupos o categorías dentro de la muestra.

Esta distribución multinodal sugiere que sería adecuado realizar un análisis más detallado para identificar y entender estos subgrupos. Además, sería interesante realizar un análisis cluster para asignar estos grupos, ya que, esto disminuiría el error en las predicciones.



Profundidad (Depth):

Media: La media de profundidad es 61.655%, lo que indica que, en promedio, los diamantes de la muestra tienen una proporción estándar entre la profundidad y el ancho. Esta media se encuentra entre 57.5% y 63% que es lo recomendado por expertos.

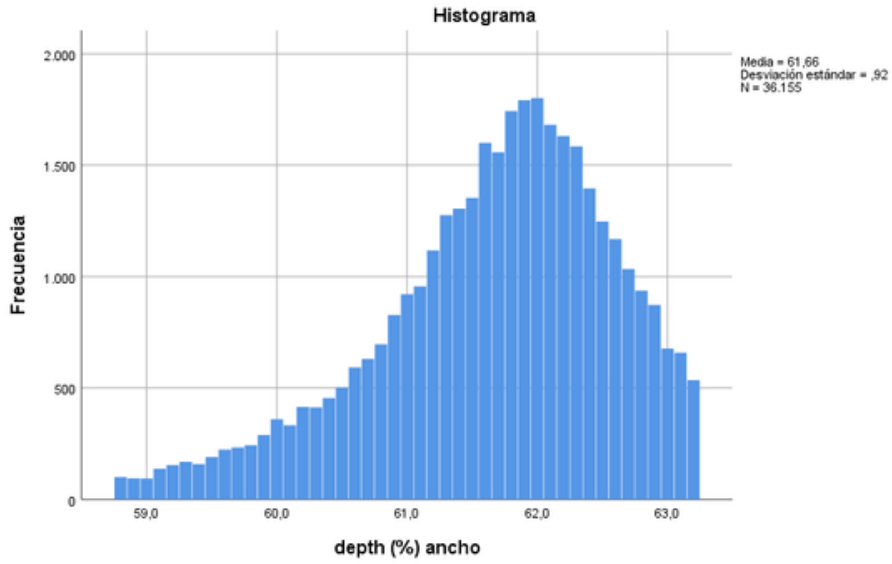
Mediana: La mediana es 61.800%, esta es bastante similar a la media, lo que indica poca asimetría, además se encuentra en la profundidad recomendada.

Varianza y desviación estándar: La varianza es de 0.846 y la desviación estándar es de 0,9196, lo que indica la dispersión de los datos alrededor de la media. El coeficiente de variación es de 1.49%

Asimetría: Una asimetría de -0.702 refleja un sesgo en la distribución hacia diamantes con menor profundidad. Esto se traduce en que hay una cola más larga en el lado izquierdo del histograma.

Curtosis: Un exceso de curtosis de 0.186, siendo positiva pero cercana a cero, señala que la distribución tiene una forma similar a una distribución normal en términos de "picosidad". No es ni demasiado puntiaguda ni demasiado plana.

Observación visual: El histograma muestra una forma unimodal y asimétrica hacia la izquierda, indicando una mayor concentración de diamantes con profundidades alrededor de la media y la mediana.



Mesa (Table):

Media: La media de la variable "table" es de 56,911. Es un valor moderado que cae dentro del rango generalmente aceptado por expertos para un diamante de corte ideal

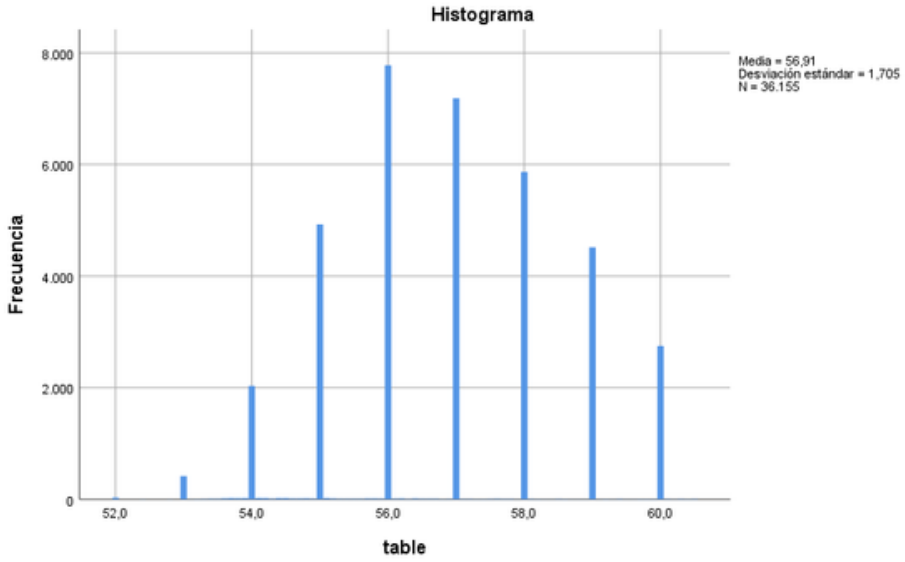
Mediana: La mediana es de 57,000, muy cercana a la media, lo que sugiere una distribución simétrica de los datos, sin asimetrías pronunciadas hacia ninguno de los lados.

Varianza y desviación estándar: La varianza es de 2.906 y la desviación estándar es de 1.7046, indicando una dispersión moderada de los datos alrededor de la media.

Asimetría: Una asimetría muy baja de 0.030 refleja una distribución casi simétrica. Esto significa que la distribución de los datos en torno a la media es bastante uniforme a ambos lados del histograma

Curtosis: Un exceso de curtosis de -0.689 indica que la distribución es platicúrtica, es decir, tiende a ser más plana en comparación con una distribución normal. Esto sugiere que los valores están distribuidos de manera más uniforme y no se concentran tan fuertemente alrededor de la media

Observación visual: El histograma muestra una forma unimodal, sin asimetrías pronunciadas y con un pico en torno a la mediana. Además, se puede notar que la mayoría de los valores están concentrados entre 54 y 60, siendo este el rango intercuartil, lo que indica que el 50% de los datos se encuentran dentro de este intervalo.



Precio (PRECIO):

Media: La media del precio es de 2365.55, la presencia de diamantes con precios elevados está afectando a la media, que es superior a la mediana.

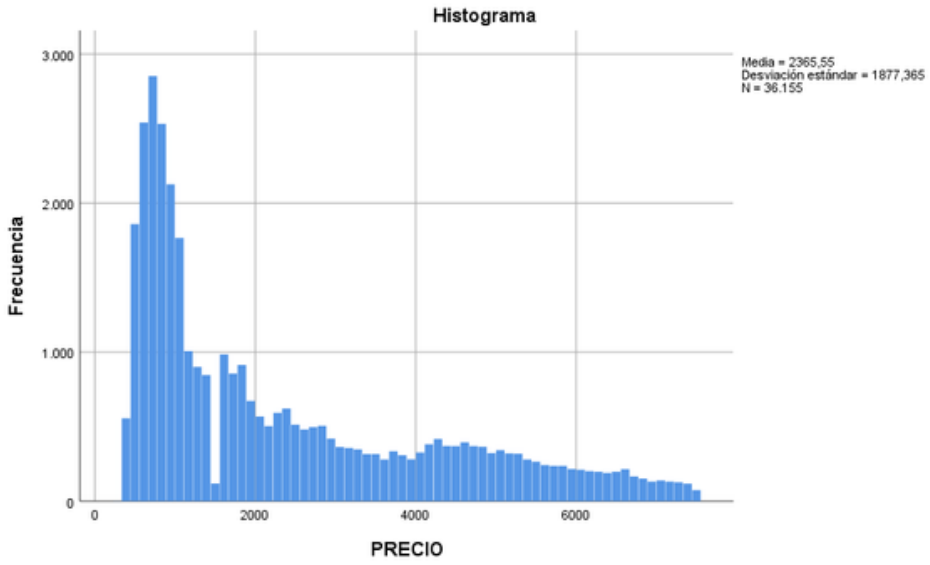
Mediana: La mediana es de 1665.00, lo que indica que el 50% de los diamantes tienen un precio inferior a este valor y el otro 50% un precio superior.

Varianza y desviación estándar: La varianza es de 3524500.009 y la desviación estándar es de 1877.365, lo que refleja una considerable dispersión de los precios alrededor de la media. El coeficiente de variación, que es de aproximadamente 79.38%, indica que la dispersión relativa es bastante alta.

Asimetría: Una asimetría de 0.967 refleja un sesgo positivo en la distribución del precio, indicando que hay un número significativo de diamantes con precios muy elevados.

Curtosis: Un exceso de curtosis de -0.228 indica que la distribución es ligeramente platicúrtica, es decir, más plana en comparación con una distribución normal.

Observación visual: El histograma muestra una fuerte concentración de diamantes con precios bajos, formando una asimetría hacia la derecha. También se puede observar que la mayoría de los diamantes tienen precios por debajo de 4000. También se observa un rango más bajo de lo normal, lo que puede indicar la existencia de dos subconjuntos de distribuciones.



Longitud (X):

Media: La media de la medida x (longitud en mm) de los diamantes en la muestra es de 5,3346 mm.

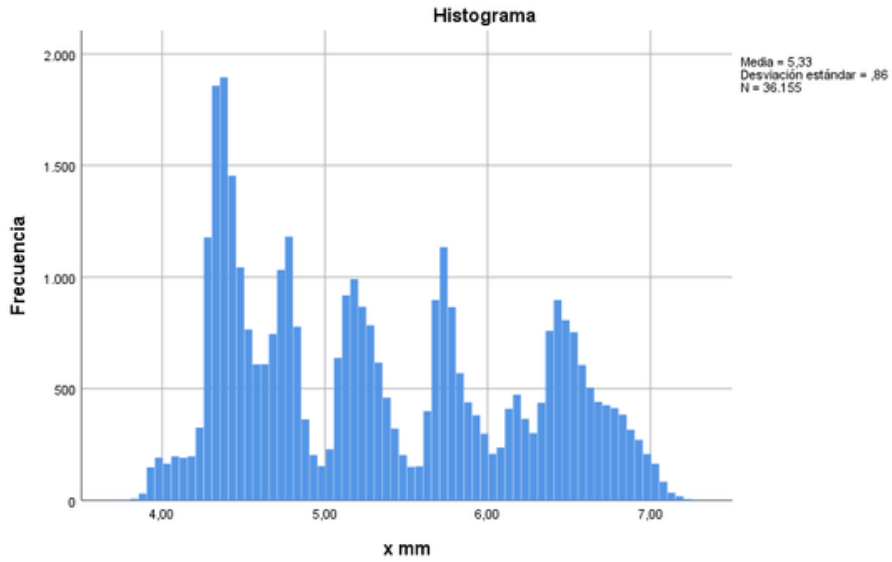
Mediana: La mediana es de 5,2000 mm. Al ser menor que la media, indica que hay una ligera inclinación de la distribución hacia valores mayores en la medida x.

Varianza y desviación estándar: La varianza es de 0,740 y la desviación estándar es de 0,86028 mm. Estos valores muestran que hay una dispersión moderada, el coeficiente de variación es de 16.13%

Asimetría: Una asimetría de 0,317 indica una ligera asimetría hacia la derecha. Esto se traduce en una concentración de diamantes con longitudes menores y una cola más larga en el lado derecho del histograma

Curtosis: Un valor de curtosis de -1,206 indica que la distribución es leptocúrtica, es decir, tiene colas más pesadas y un pico más agudo en comparación con una distribución normal.

Observación visual: El histograma muestra tres picos, con el pico más alto alrededor de los 4,5 mm. La distribución tiende a inclinarse ligeramente hacia valores mayores, como lo indica la asimetría. La mayoría de los diamantes tienen una longitud entre 4,5 mm y 6,5 mm, con menos diamantes con longitudes superiores a 6,5 mm. Podemos apreciar varios subconjuntos de distribuciones.



Altura (Y):

Media: La media de la medida "y" es de 5.3420 mm

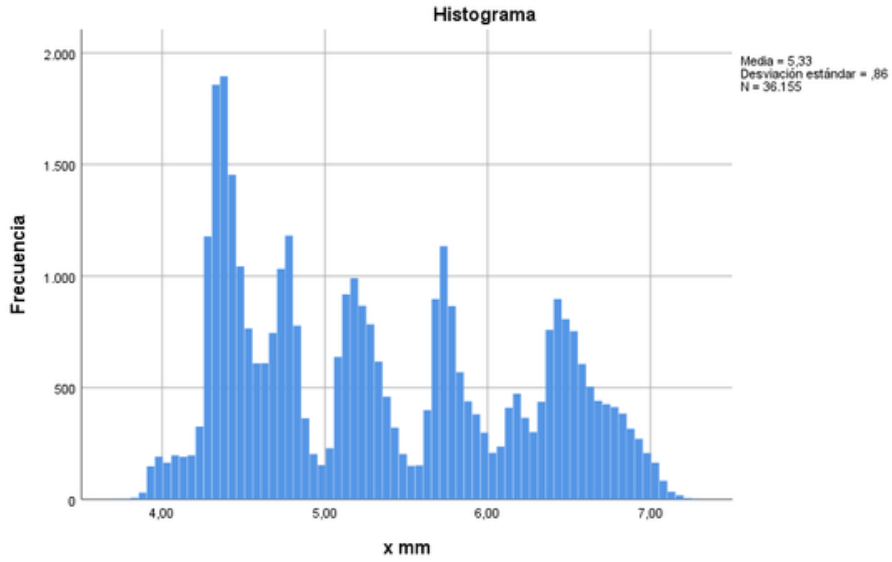
Mediana: La mediana de 5.2100 mm, al ser menor que la media, refleja una ligera tendencia hacia la presencia de diamantes con medidas menores en la dimensión "y".

Varianza y desviación estándar: La varianza de 0.732 y la desviación estándar de 0.85544 mm nos dan una idea de la dispersión de los datos alrededor de la media y el coeficiente es de 16.01%

Asimetría: La asimetría positiva de 0.307 indica un sesgo leve en la distribución hacia diamantes con medidas menores en la dimensión "y". Esto se traduce en una cola más larga en el lado derecho del histograma

Curtosis: Una curtosis de -1.224, siendo negativa, indica que la distribución es más plana en comparación con una distribución normal.

Observación visual: Al observar el histograma, se puede notar que hay múltiples picos o modas, lo que sugiere la presencia de subgrupos o categorías distintas dentro de la muestra en cuanto a la dimensión "y" de los diamantes.



Profundidad (Z):

Media: La media de la profundidad "z" es de 3.2909 mm. Esto sugiere que, en promedio, los diamantes de la muestra tienen una profundidad de alrededor de 3.29 mm

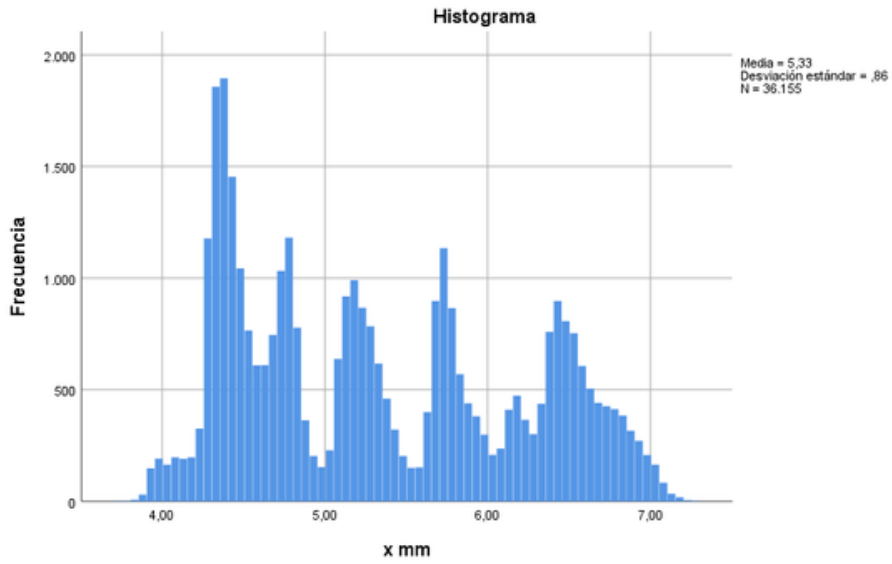
Mediana: La mediana es de 3.2100 mm. Al ser menor que la media, refleja una ligera tendencia hacia la presencia de diamantes con profundidades menores en la muestra.

Varianza y desviación estándar: La varianza es de 0.280 y la desviación estándar de 0.52955 mm.

Asimetría: Una asimetría positiva de 0.314 señala un sesgo leve hacia diamantes con profundidades menores. Esto implica una cola más larga en el lado derecho del histograma

Curtosis: Una curtosis de -1.211, siendo negativa, sugiere que la distribución es más plana en comparación con una distribución normal

Observación visual: El histograma muestra claramente una distribución con varios picos, lo que indica la presencia de subgrupos o categorías distintas dentro de la muestra en cuanto a la profundidad de los diamantes.



Color (Intervalo)

Media: La media de la intensidad del color es de 3.14. Esto sugiere que, en promedio, los diamantes en la muestra están cerca del valor medio en la escala de 1 a 5.

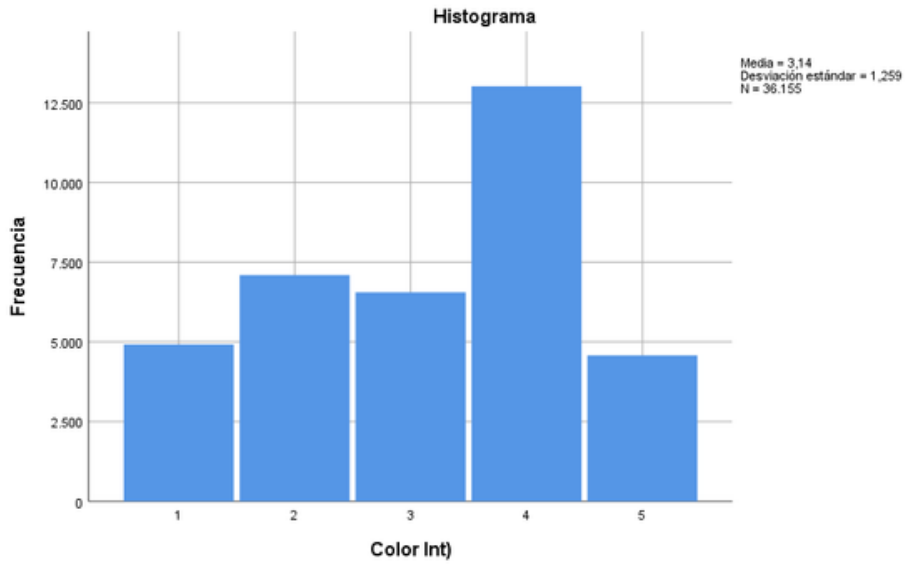
Mediana: La mediana es de 3.00. Coincide con la mitad exacta de la escala.

Varianza y desviación estándar: a varianza es de 1.585 y la desviación estándar es de 1.259. Estos valores muestran una variabilidad notable en la intensidad del color de los diamantes en la muestra.

Asimetría: La asimetría es negativa (-0.302), lo que indica un sesgo leve hacia las intensidades de color más altas, es decir, hay más diamantes en la muestra con intensidades de color mayores que menores.

Curtosis: Una curtosis de -1.048, siendo negativa, sugiere que la distribución es más plana en comparación con una distribución normal. Esto significa que los valores están más dispersos y hay menos concentración alrededor de la media.

Observación visual: El histograma muestra una distribución con una notable concentración en la intensidad de color 4. Esto sugiere que muchos diamantes en la muestra tienen una intensidad de color cercana a 4.



Claridad (Intervalo)

Media: La media de claridad es de 4.13, lo que indica que en promedio los diamantes en la muestra tienen una claridad cercana a 4 en una escala de 1 a 6

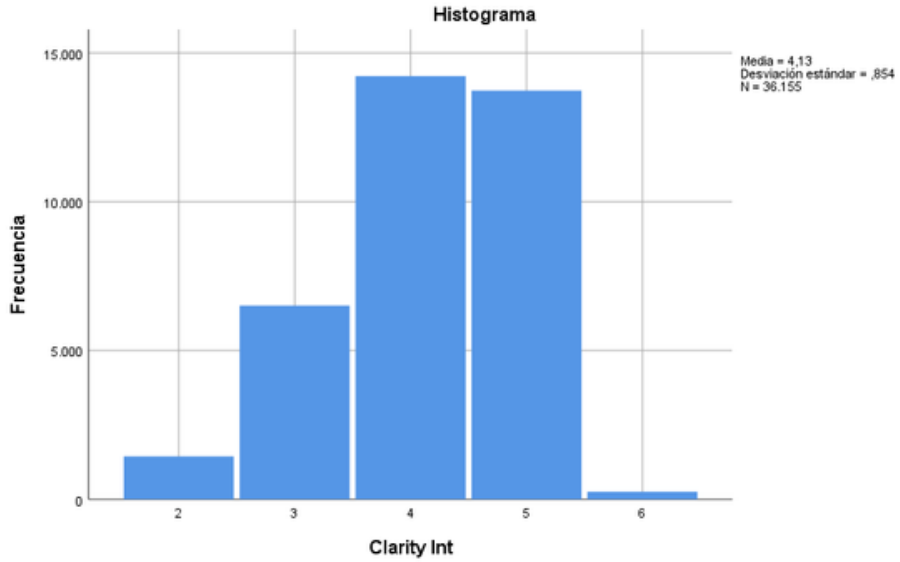
Mediana: La mediana es 4.00, lo que indica que la mitad de los diamantes tienen una claridad por debajo de 4 y la otra mitad por encima.

Varianza y desviación estándar: La varianza es 0.729 y la desviación estándar es 0.854, mostrando una dispersión moderada en la claridad de los diamantes en la muestra.

Asimetría: La asimetría es negativa (-0.576), lo que indica un sesgo hacia claridades más altas

Curtosis: La curtosis de -0.324 sugiere que la distribución es levemente más plana que una distribución normal, indicando una dispersión de los datos.

Observación visual: El histograma muestra que la mayoría de los diamantes en la muestra caen en las categorías de claridad 4 y 5.



Calidad del corte (Intervalo)

Media: La media de calidad del corte es de 1.74, lo que indica que, en promedio, los diamantes en la muestra tienen una calidad de corte cercana a 2 en una escala de 1 a 5.

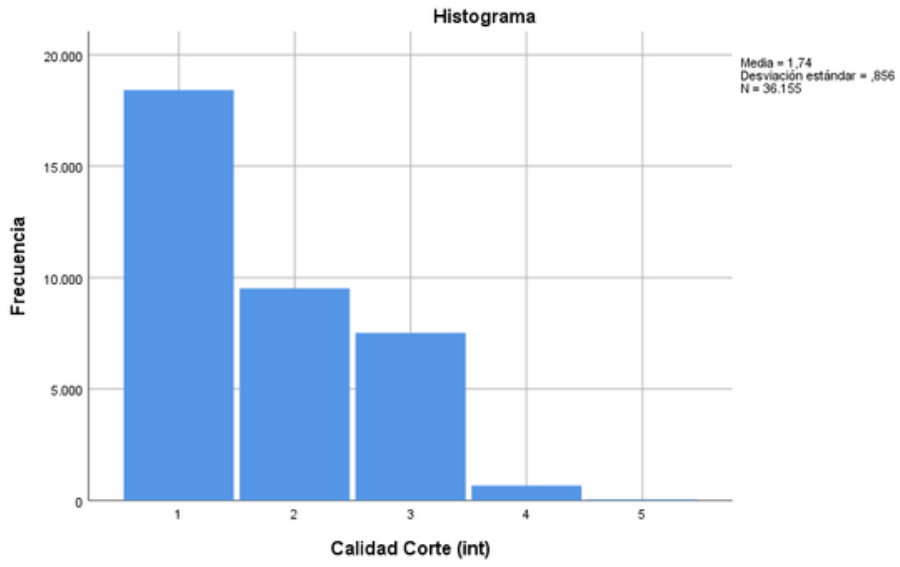
Mediana: La mediana es 1.00, lo que indica que la mitad de los diamantes tienen una calidad de corte de 1 y la otra mitad tiene una calidad de corte superior a 1.

Varianza y desviación estándar: La varianza es 0.732 y la desviación estándar es 0.856, lo que muestra una dispersión moderada en la calidad del corte de los diamantes en la muestra.

Asimetría: La asimetría es positiva (0.743), lo que indica un sesgo hacia calidades de corte más bajas.

Curtosis: La curtosis de -0.619 sugiere que la distribución es levemente más plana que una distribución normal, indicando una mayor dispersión de los datos.

Observación visual: El histograma muestra que la mayoría de los diamantes en la muestra tienen una calidad de corte de 1, seguido por 2 y 3. Es menos común encontrar diamantes con una calidad de corte de 4, y aún menos común para una calidad de corte de 5. Así que generalmente los diamantes se cortan bien.



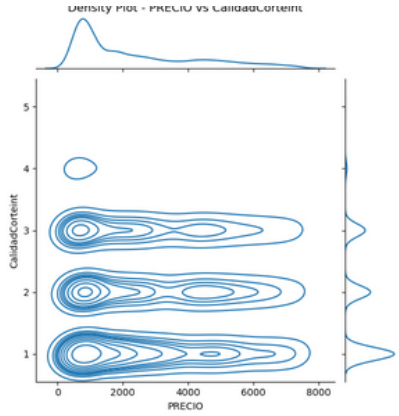
Identificación del número de clusters

Realizamos la prueba de Kolmogorov Smirnoff y las hipótesis nulas se rechazan en todas las variables, por lo que las distribuciones no son normales.

Tras la realización de un análisis clúster jerárquico a las variablespara analizar el número de subgrupos que encontramos en esta variable, para ello usamos el método de Ward, que minimiza la varianza total dentro de los clústeres. Esta técnica es robusta y no requiere estrictamente que los datos sigan una distribución normal.

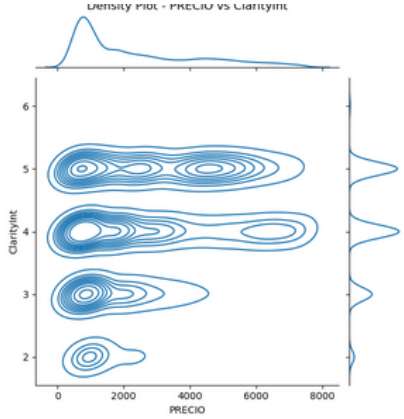
Tras observar el dendrograma proponemos realizar un modelo de clasificación de Cluster K-means con

Graficos de densidad



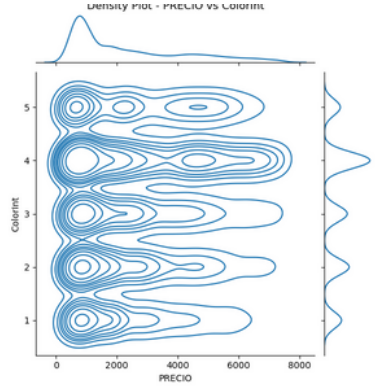
- **Precio -Calidad Corte:**

En este gráfico vemos que los diamantes de mejor calidad (categorías mas bajas) se concentran más en precios altos que aquellas categorías con más defectos. Como era de esperar.



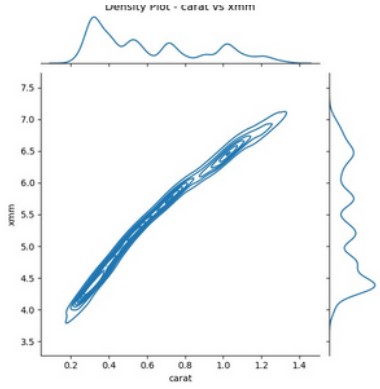
- **Precio - ClarityInt:**

Aquí observamos una contradicción, las categorías con menos defectos tienen precios mas baratos y aquellas con mayores defectos tienen precios más altos. Por ejemplo, existe una concentración de diamantes con precios mayores a 4000, que se encuentran en la categoría 5.



- **Color - Precio:**

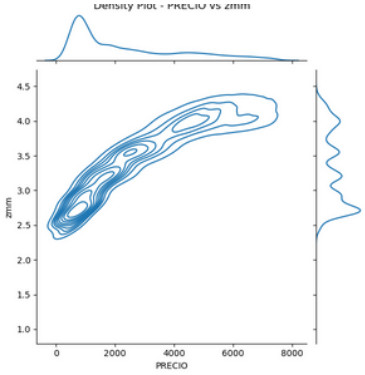
La densidad del color con el precio es más confusa, vemos concentraciones en niveles altos en la categoría 4 y algo en la categoría 5.

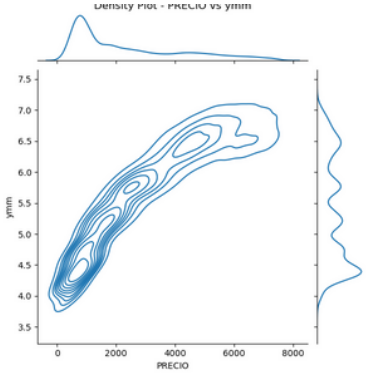


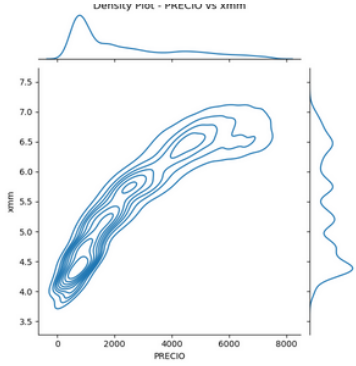
- **Carat - Precio:**

Observamos que a medida que aumenta el peso en quilates, el rango de precios también tiende a aumentar.

Las densidades más altas parecen estar centradas alrededor de 0.2-0.4 carat y 61-62 de depthancho, con otras áreas menos densas dispersas a lo largo del gráfico.







- **Z, Y, X - Precio:**

Podemos observar como en las dimensiones, el precio aumenta a medida que aumenta el tamaño del diamante. Además, podemos notar que sigue una distribución bastante lineal.

Correlaciones

Correlaciones

		carat	depth (%) ancho	table	PRECIO	x mm	y mm	z mm	Color Int)	Clarity Int	Calidad Corte (int)
carat	Correlación de Pearson	1	,008	,174**	,943**	,990**	,989**	,989**	,193**	,427**	,127**
	Sig. (bilateral)		,126	,000	,000	,000	,000	,000	,000	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
depth (%) ancho	Correlación de Pearson	,008	1	-,238**	,004	-,039**	-,039**	,056**	,050**	,021**	,035**
	Sig. (bilateral)	,126		,000	,453	,000	,000	,000	,000	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
table	Correlación de Pearson	,174**	-,238**	1	,149**	,167**	,160**	,141**	-,009	,155**	,449**
	Sig. (bilateral)	,000	,000		,000	,000	,000	,000	,085	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
PRECIO	Correlación de Pearson	,943**	,004	,149**	1	,926**	,926**	,925**	,106**	,271**	,098**
	Sig. (bilateral)	,000	,453	,000		,000	,000	,000	,000	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
x mm	Correlación de Pearson	,990**	-,039**	,167**	,926**	1	,998**	,994**	,175**	,432**	,110**
	Sig. (bilateral)	,000	,000	,000	,000		,000	,000	,000	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
y mm	Correlación de Pearson	,989**	-,039**	,160**	,926**	,998**	1	,994**	,175**	,428**	,116**
	Sig. (bilateral)	,000	,000	,000	,000	,000		,000	,000	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
z mm	Correlación de Pearson	,989**	,056**	,141**	,925**	,994**	,994**	1	,180**	,432**	,116**
	Sig. (bilateral)	,000	,000	,000	,000	,000	,000		,000	,000	,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
Color Int)	Correlación de Pearson	,193**	,050**	-,009	,106**	,175**	,175**	,180**	1	-,088**	,004
	Sig. (bilateral)	,000	,000	,085	,000	,000	,000	,000		,000	,447
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
Clarity Int	Correlación de Pearson	,427**	,021**	,155**	,271**	,432**	,428**	,432**	-,088**	1	,132**
	Sig. (bilateral)	,000	,000	,000	,000	,000	,000	,000	,000		,000
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155
Calidad Corte (int)	Correlación de Pearson	,127**	,035**	,449**	,098**	,110**	,116**	,116**	,004	,132**	1
	Sig. (bilateral)	,000	,000	,000	,000	,000	,000	,000	,447	,000	
	N	36155	36155	36155	36155	36155	36155	36155	36155	36155	36155

**. La correlación es significativa en el nivel 0,01 (bilateral).

Vamos a analizar las correlaciones mostradas en la tabla:

- **carat**:
  - Tiene una correlación muy fuerte y positiva con "PRECIO" (0.943) lo que indica que, a medida que aumenta el tamaño del diamante (carat), el precio también tiende a aumentar. Esto es consistente con lo que uno esperaría en el mercado de diamantes.
  - También tiene fuertes correlaciones positivas con "x mm", "y mm" y "z mm", lo que es lógico ya que estas medidas indican las dimensiones físicas del diamante.
- **depth** (% ancho):
  - Presenta una correlación negativa, aunque débil, con "y mm" y "z mm" (-0.039 y -0.056 respectivamente).

- **table**:
    - Tiene una correlación positiva con "PRECIO" (0.149) aunque es bastante débil.
    - Presenta correlaciones negativas débiles con "y mm" y "z mm" (-0.039 y -0.056 respectivamente).
- **Precio**:
    - Además de la fuerte correlación con "carat", tiene correlaciones positivas muy fuertes con las dimensiones "x mm", "y mm" y "z mm" (0.990, 0.989 y 0.989 respectivamente).
    - La correlación con "Clarity Int" es positiva pero moderada (0.427). Esto puede ser la razón por la cual vimos en el gráfico anterior que los diamantes con más defectos (mayor valor en "Clarity Int") tenían precios más altos. Sin embargo, esto es solo una correlación y no implica causalidad directa.
- **Dimensiones (x mm, y mm, z mm)**:
    - Estas tres dimensiones tienen correlaciones muy fuertes entre sí (alrededor de 0.990 o más). Esto es de esperar ya que las dimensiones de un diamante tienden a aumentar juntas a medida que crece en tamaño.
- **Color Int**:
    - Tiene una correlación negativa débil con "PRECIO" (-0.009), lo que indica que hay poca o ninguna relación lineal entre el color y el precio.
- **Clarity Int**:
    - Aparte de la correlación con "PRECIO", tiene correlaciones positivas débiles con las dimensiones y "table", y una correlación negativa débil con "depth (% ancho)".
- **Calidad Corte (int)**:
    - Tiene correlaciones débiles con la mayoría de las variables, con la correlación más fuerte siendo con "table" (0.449).

En general, lo más destacado de estas correlaciones es la fuerte relación entre "carat" y "PRECIO", así como las dimensiones del diamante

Estudio de las Características para un Modelo Multivariante

Selección de características.

Basado en los análisis anteriores, algunas variables que parecen ser particularmente significativas son "carat", las dimensiones "x mm", "y mm", "z mm", y "Clarity Int"

Propuesta de Modelos:

Clustering y Regresión Múltiple: Podemos utilizar técnicas de clustering, como el K-means, para identificar subgrupos de diamantes que tienen características similares, tal y como hemos visto en los histogramas.

Posteriormente, podemos entrenar un modelo de regresión múltiple para cada cluster, lo que nos permitiría tener predicciones más específicas según las características de cada grupo.

Regresión Múltiple Directa: Como alternativa, podemos utilizar una regresión múltiple que tome en cuenta todas las variables o solo aquellas que sean estadísticamente más significativas

Propuesta de Implementación del Modelo

Proceso de Entrenamiento:

Para entrenar el modelo, podemos dividir los datos en un conjunto de entrenamiento y un conjunto de prueba, típicamente en una proporción de 70:30. El conjunto de entrenamiento se utilizará para construir el modelo y el conjunto de prueba para evaluar su rendimiento.

Validación y Ajuste:

Podemos emplear técnicas de validación cruzada, como k-fold cross-validation, para obtener una estimación más robusta del rendimiento del modelo. Además, se puede hacer una búsqueda en malla (grid search) para ajustar los hiperparámetros del modelo

Medición del Rendimiento:

Los criterios para medir el rendimiento del modelo incluirán el error cuadrático medio (MSE) para evaluar la precisión de las predicciones y el coeficiente de determinación (R^2) para evaluar la proporción de la variabilidad explicada por el modelo

Conclusiones

Conclusiones Generales:

A partir de nuestro análisis descriptivo y gráfico, es evidente que el peso del diamante ("carat") y sus dimensiones son factores significativos en la determinación del precio.

Además, se han detectado subgrupos en las variables significativas, por lo que esto apoya la idea de realizar primero un modelo de clasificación en estas variables y posteriormente desarrollar el modelo de regresión multivariante en cada uno de los cluster para mejorar la predicción.

Potencial del Modelo:

El modelo propuesto podría ofrecer a los comerciantes de diamantes una herramienta valiosa para predecir el precio de un diamante en función de sus características, optimizando así la toma de decisiones relacionada con la compra, venta o tasación de diamantes.

Limitaciones y Recomendaciones:

Es importante recordar que correlación no implica causalidad. Además, se recomienda explorar otros factores que pueden influir en el precio, como la marca, el lugar de compra, entre otros. Se podría considerar la adición de más datos o la utilización de técnicas de aprendizaje más avanzadas para mejorar el modelo.

A pesar de haber realizado un primer análisis con datos atípicos, se ha llegado a la conclusión de que es preferible dejarlos fuera de este modelo. Si se desea generar un modelo que cubra estos atípicos, la recomendación es agruparlos por separado.