



Trabajo Final para la Especialización
en Machine Learning

MODELO PREDICTIVO DE ESTIMACIÓN DE PRECIOS DE DIAMANTES

Integrantes:

- Liner Lander Cullanco Canchaya
- Jair Gustavo Olivares Stuva

INTRODUCCIÓN

En el fascinante mundo de las piedras preciosas, cada diamante tiene una historia que contar a través de sus atributos únicos.

Nuestro proyecto se sumerge en un conjunto de datos de casi 54,000 diamantes, analizando factores clave como el precio, el peso en quilates, la calidad de corte, el color, y la claridad.

Además, examinamos medidas precisas como longitud, ancho y profundidad. Con esta rica información, buscamos predecir con precisión el valor de cada diamante, optimizando la estimación de precios y proporcionando una herramienta invaluable para la toma de decisiones en el mercado de joyas.

PROBLEMÁTICA

1. Variabilidad en los Precios:

Los precios de diamantes varían debido a factores como quilate, color, claridad y geometría, complicando la estimación precisa. Esto puede llevar a inconsistencias en las valoraciones y decisiones económicas erróneas.

2. Evaluación Manual:

La evaluación humana es subjetiva y lenta, con riesgo de errores, especialmente con grandes volúmenes de datos. Esto afecta la precisión y consistencia en la valoración de diamantes.

3. Escalabilidad Limitada:

Métodos tradicionales no escalan bien con grandes volúmenes de datos o análisis en tiempo real. Esto limita la capacidad para procesar y valorar eficientemente grandes inventarios.

PROPUESTA

1. Modelo Predictivo con Machine Learning:

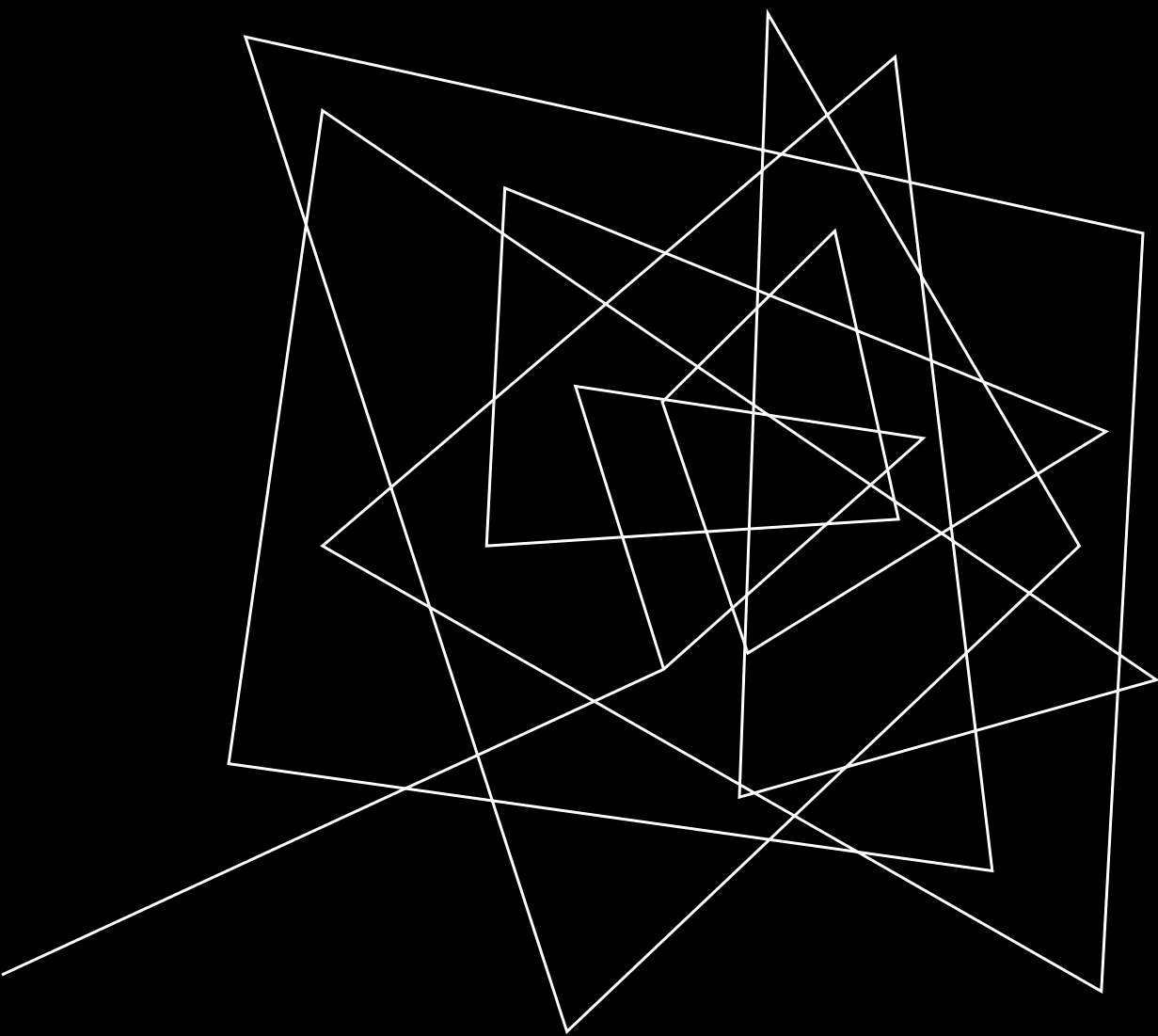
Desarrollar un modelo predictivo que estime el precio del diamante basado en quilate, color, claridad y geometría. Esto automatiza la valoración, reduce la subjetividad y mejora la consistencia.

2. Algoritmos de Regresión:

Usar regresión lineal múltiple para capturar relaciones complejas entre características y precio. Mejora la precisión de las predicciones.

3. Validación y Ajuste del Modelo:

Validar y ajustar el modelo usando datos de prueba y técnicas de validación cruzada. Asegura que el modelo sea preciso y generalice bien a nuevos datos.

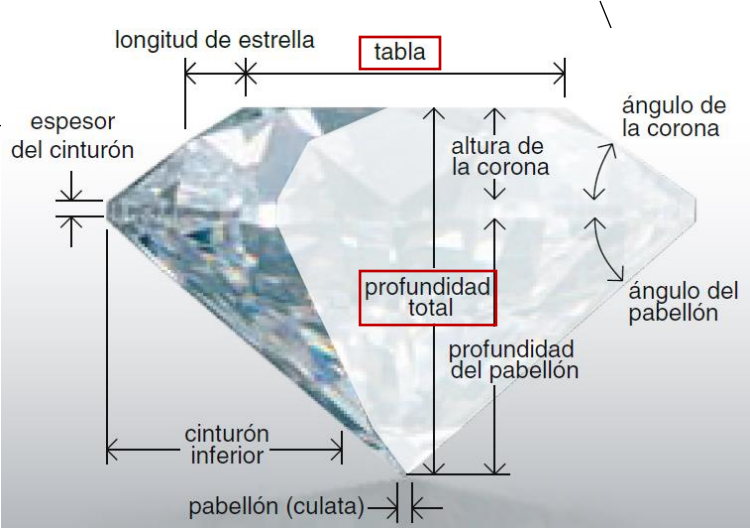


EXPLORACIÓN DE SET DE DATOS

COMPRENSIÓN DE DATOS

Se ha trabajado con un set de datos compuesto por 53 940 registros y 10 variables (tipo *numeric* y *object*). La variable “Price” es el target o variable objetivo.

Clave	Descripción	Tipo de dato
price	Precio en dólares estadounidenses	Int
carat	Peso en quilates del diamante	Float
cut	Calidad de corte del corte (Fair, Good, Very Good, Premium, Ideal)	Object
color	Color del diamante de J (peor) a D (mejor)	Object
clarity	Medida de qué tan claro es el diamante (I1 (peor), SI2, SI1, VS2, VS1, VVS2, VVS1, IF (mejor)).	Object
x	Longitud en mm	Float
y	Ancho en mm	Float
z	Profundidad en mm	Float
depth	Porcentaje de profundidad total	Float
table	Ancho de la parte superior del diamante en relación con el punto más ancho (en porcentaje)	Float



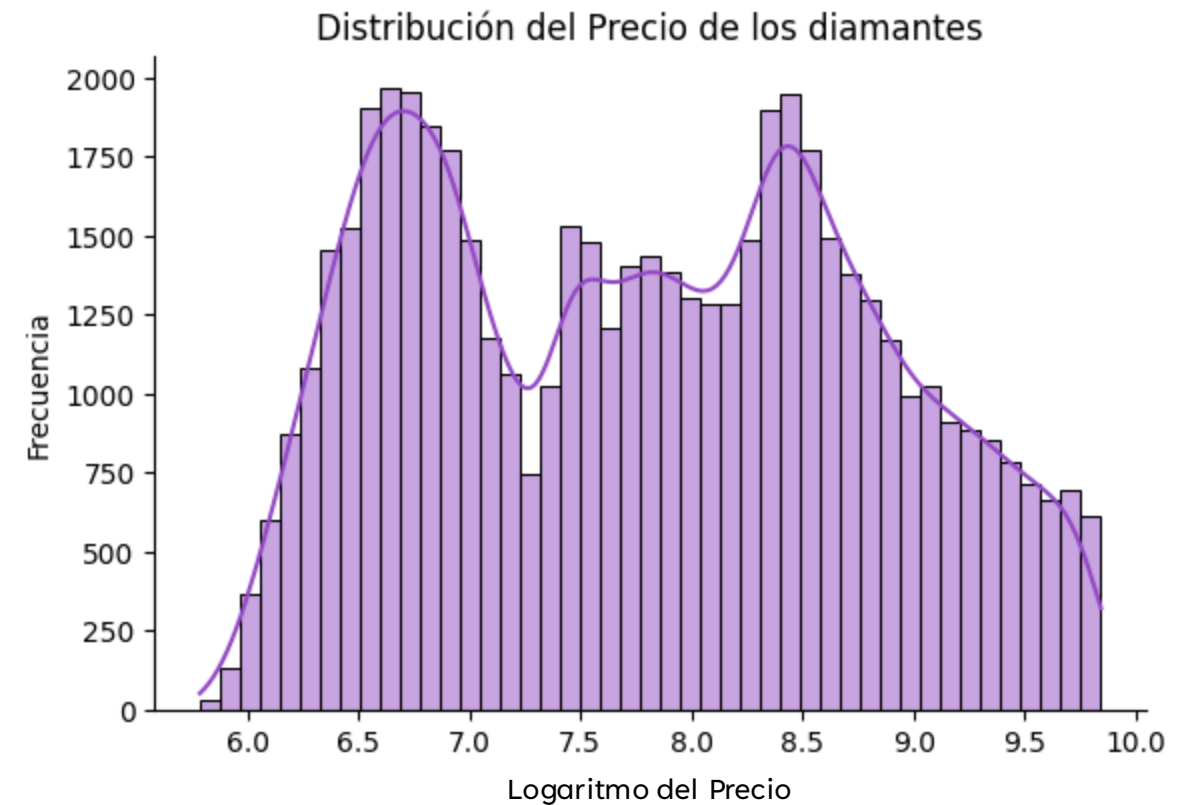
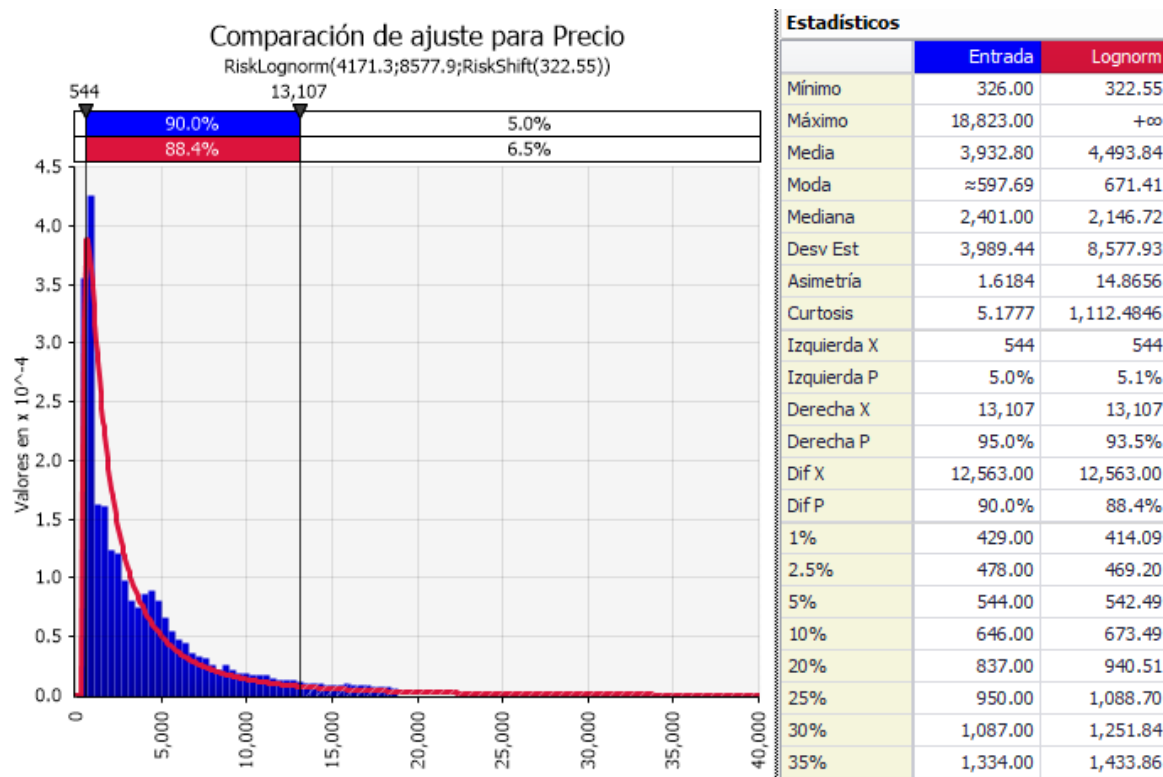
CARAT		CLARITY		COLOUR		CUT	
	1.25 CTS	FLAWLESS	FL	COLOURLESS	D		ROUND
	1 CT	INTERNALLY FLAWLESS	IF	NEAR COLOURLESS	E		PRINCESS
					F		
	0.75 CT	VERY, VERY SLIGHTLY INCLUDED	VVS ₁	FAINT	G		
			VVS ₂		H		MARQUISE
	0.50 CT	VERY, SLIGHTLY INCLUDED	VS ₁	VERY LIGHT	I		
			VS ₂		J		
	0.25 CT	SLIGHTLY INCLUDED	SI ₁	LIGHT	K		PEAR
			SI ₂		L		
	0.10 CT	INCLUDED	I ₁		M		
			I ₂		N		EMERALD
			I ₃		O		
	0.05 CT				P		
					Q		
					R		
					S		
					T		
					U		
					V		
					W		
					X		
					Y		
					Z		

*Se encontró 55 valores duplicados que fueron eliminados.

<https://www.goldandtime.org/noticia/83635/goldtime/cuales-son-los-criterios-para-evaluar-diamantes-que-empleamos-los-tasadores.html>

COMPRENSIÓN DE DATOS

La variable 'Precio' tiene naturaleza que se podría ajustar a un comportamiento de distribución lognormal, por ello, se ha trabajado con el valor logarítmico del precio para las correlaciones.














APLICACIÓN DE ORDINARY ENCODER

Se puede observar que las características de Claridad, Color y Corte sí tienen un orden, es decir son variables categóricas ordinales.

Debido a ello, se asigna una valoración a dichas propiedades con ‘ordinary recoder’:

OrdinalEncoder

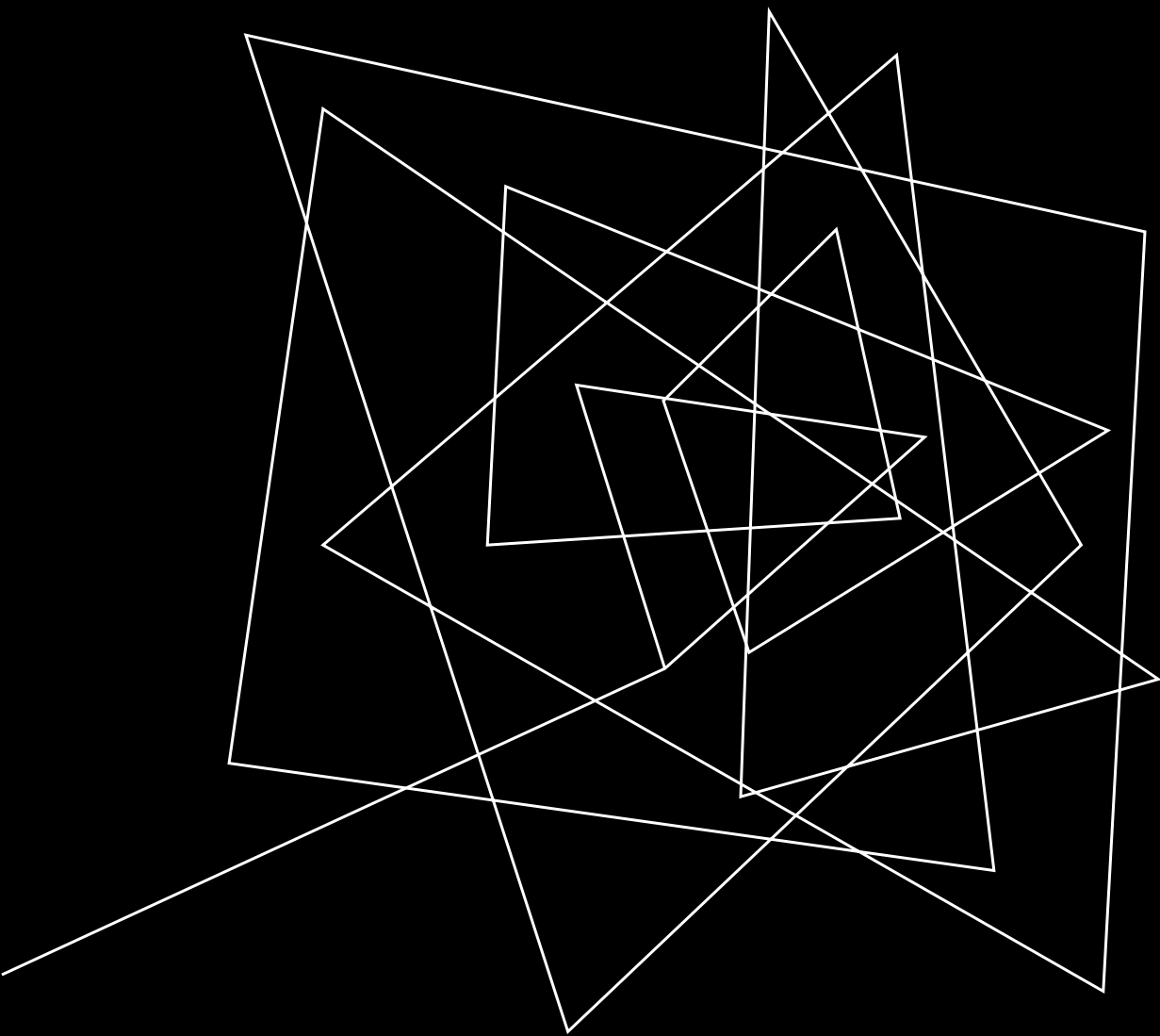


CARAT		CLARITY		COLOUR		CUT		
	1.25 CTS	FLAWLESS	FL		COLOURLESS	D		ROUND
	1 CT							
	0.75 CT	INTERNALLY FLAWLESS	IF		NEAR COLOURLESS	G		PRINCESS
						H		
						I		
	0.50 CT	VERY, VERY SLIGHTLY INCLUDED	VVS ₁		FAINT	K		MARQUISE
						L		
						M		
	0.25 CT	VERY, SLIGHTLY INCLUDED	VS ₁		VERY LIGHT	N		PEAR
						O		
						P		
	0.10 CT	SLIGHTLY INCLUDED	SI ₁		LIGHT	Q		EMERALD
						R		
						S		
	0.05 CT	INCLUDED	I ₁			T		
			I ₂			U		
			I ₃			V		
						W		
						X		
						Y		
						Z		

Claridad	Valor
IF	7
VVS1	6
VVS2	5
VS1	4
VS2	3
SI1	2
SI2	1
I1	0

Color	Valor
D	6
E	5
F	4
G	3
H	2
I	1
J	0

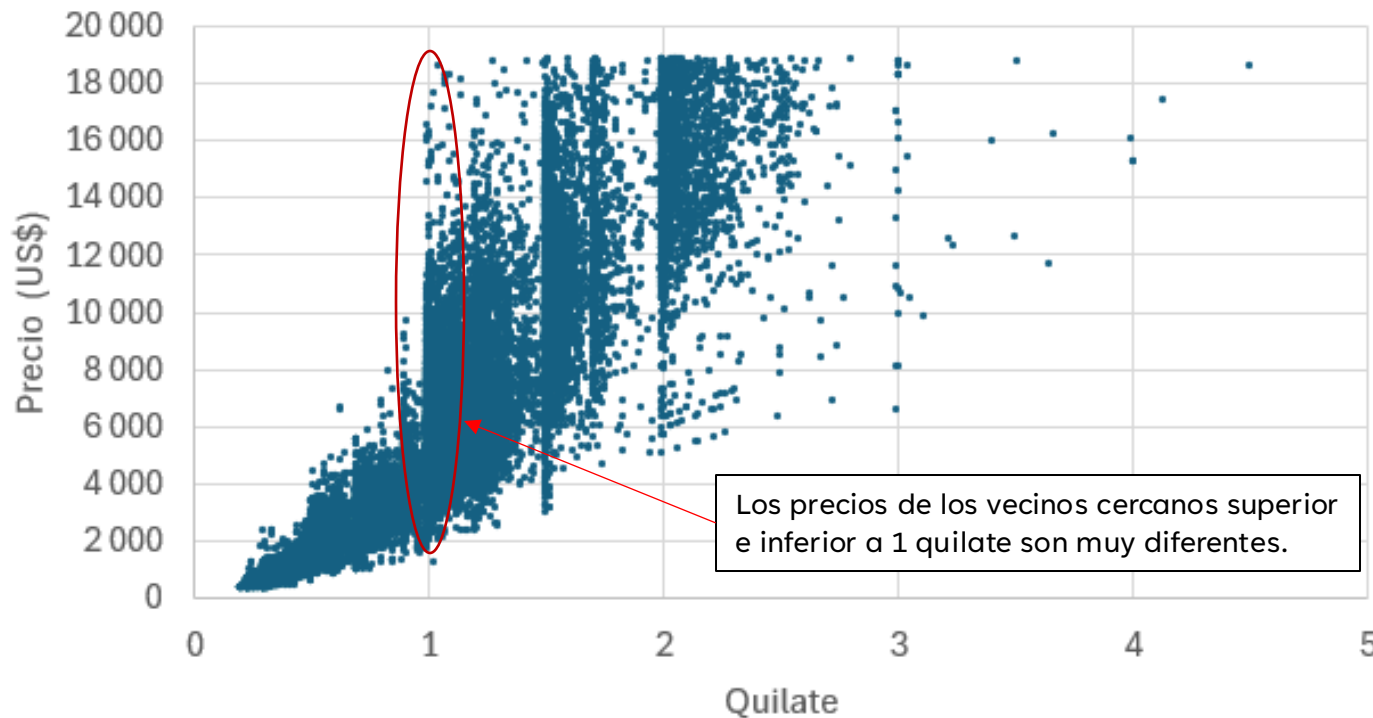
Corte	Valor
Ideal	4
Premium	3
Very Good	2
Good	1
Fair	0



SELECCIÓN DE MUESTRAS

SELECCIÓN DE MUESTRAS

Se observa una lógica tendencia positiva en la relación Quilate-Precio. Sin embargo, se observan importantes discontinuidades en los valores de 1, 1.5 y 2 quilates.



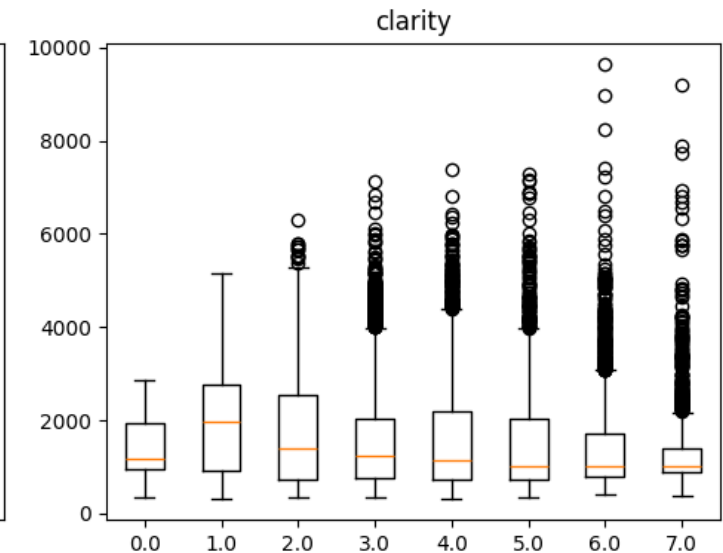
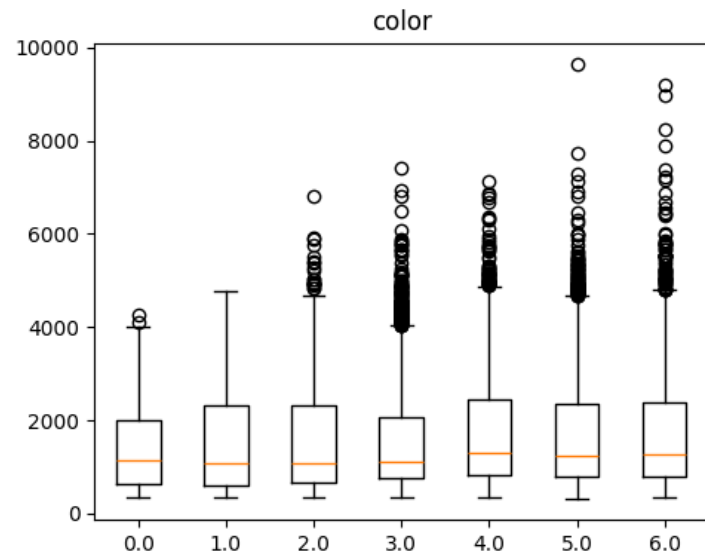
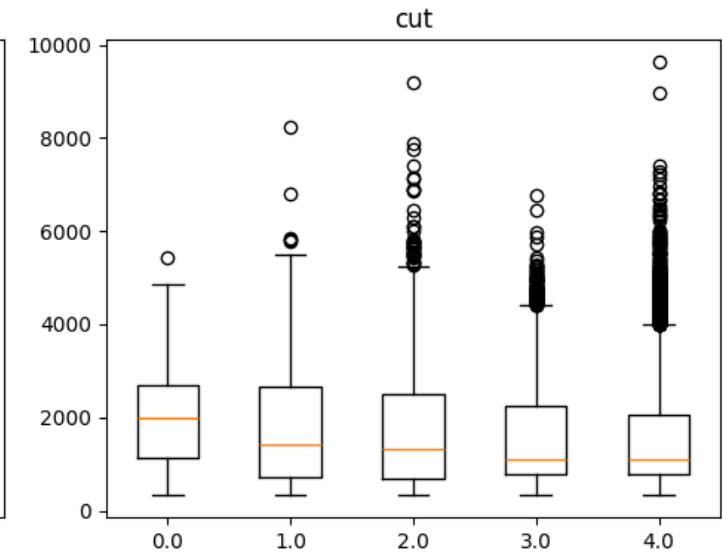
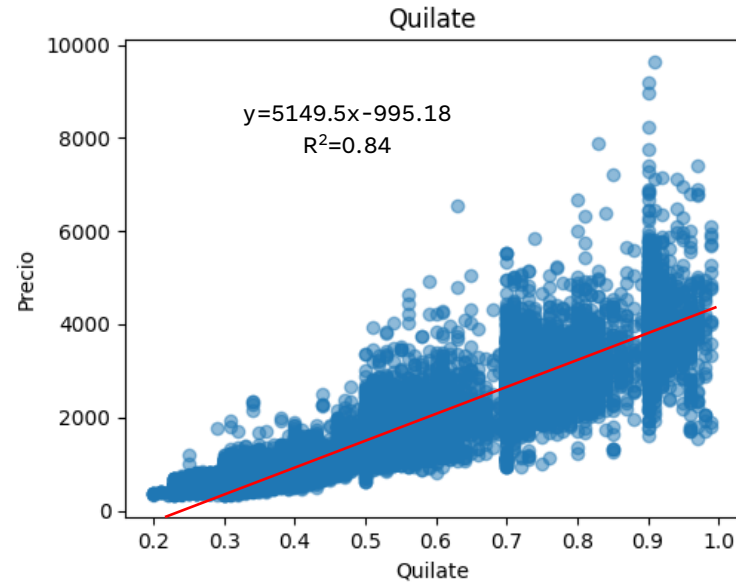
Se deduce que esto se debe a una clasificación del mercado. Por lo tanto, el análisis se realizará en los siguientes rangos de quilates:

- < 1.00 quilate
- $1.00 - 1.49$ quilates
- $1.50 - 1.99$ quilates
- ≥ 2.00 quilates

VALORES MENORES A 1 QUILATE

Menores a 1 quilate la variable que tiene mayor inferencia es el propio valor del quilate ($R^2=0.84$).

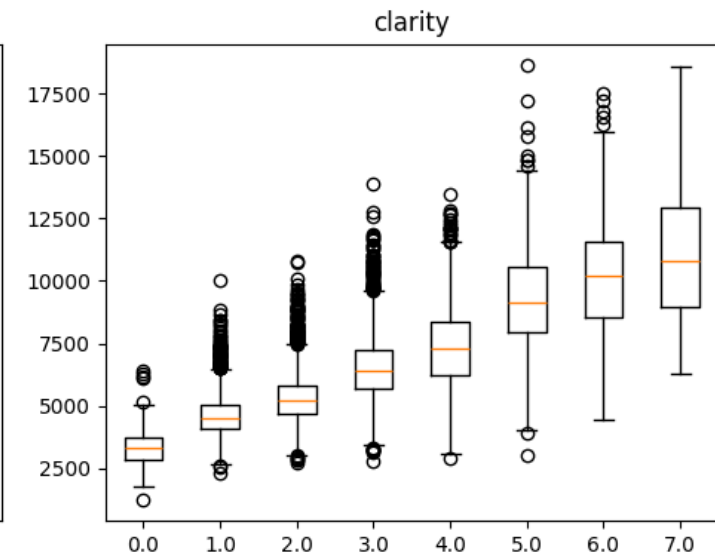
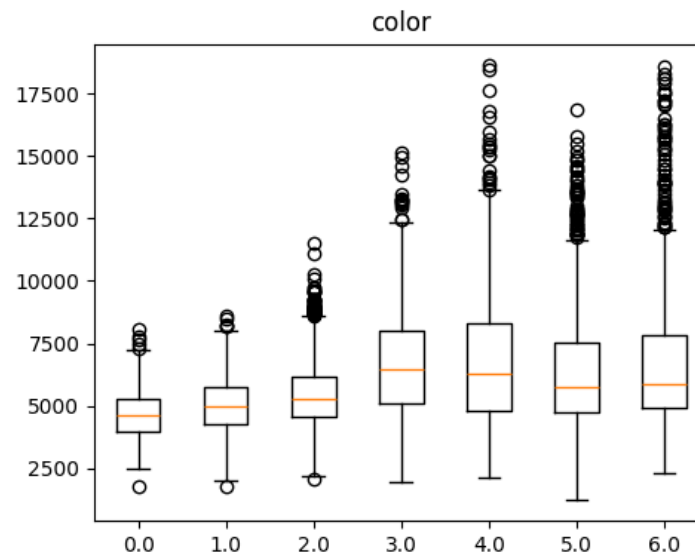
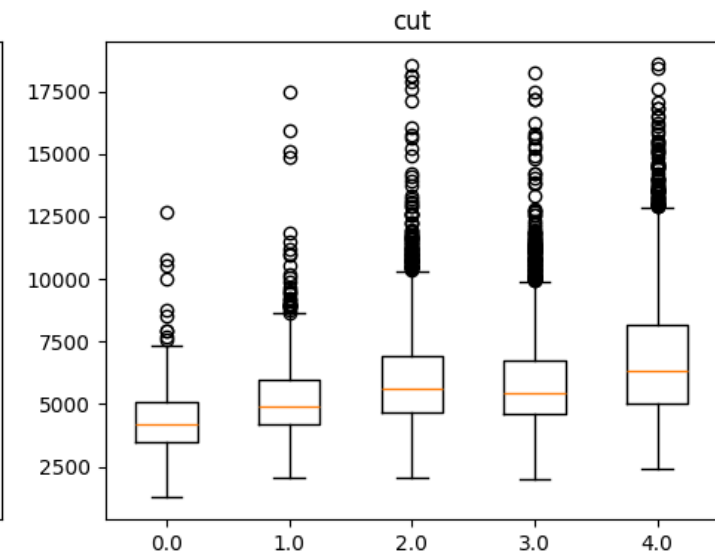
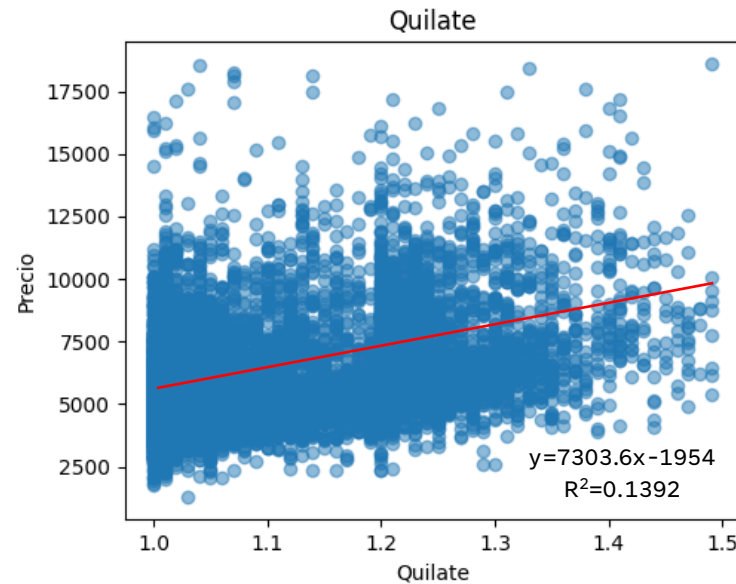
Las propiedades como la Claridad, Color y Corte no tienen relevancia.



VALORES DE 1.00 – 1.49 QUILATES

Para los diamantes de 1 a 1.5 quilates, el precio depende principalmente de la Claridad, ligeramente del Corte y Color.

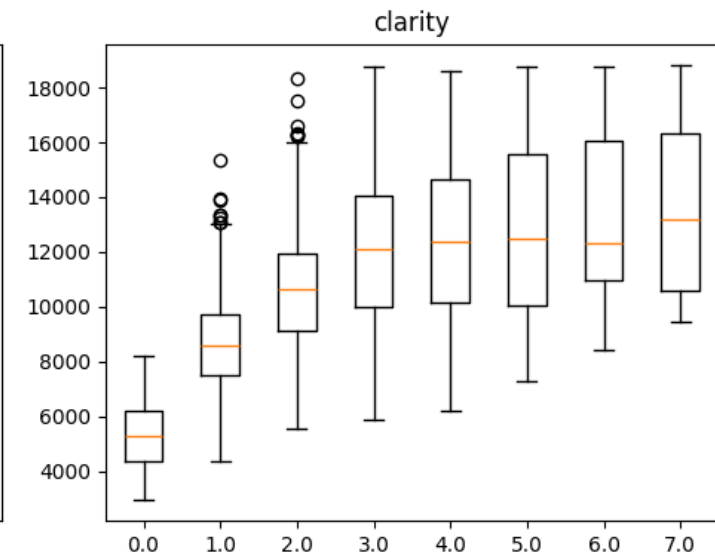
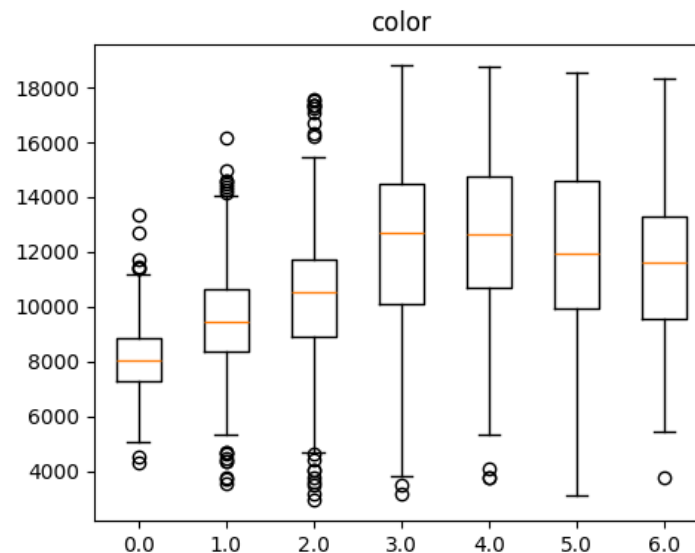
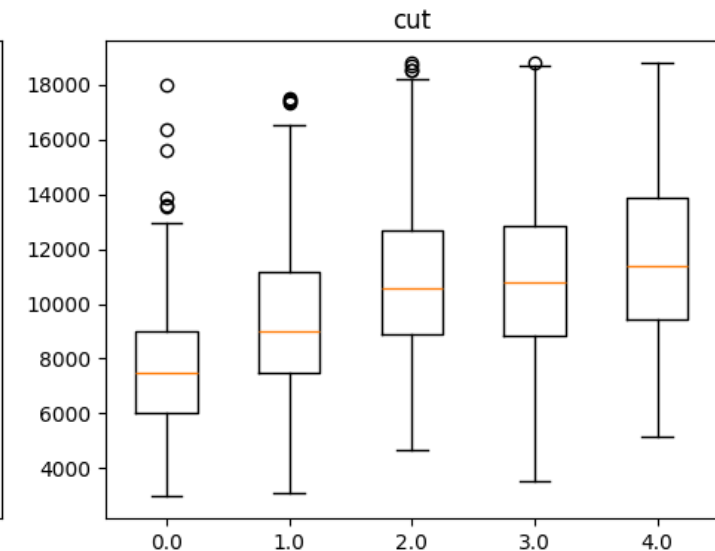
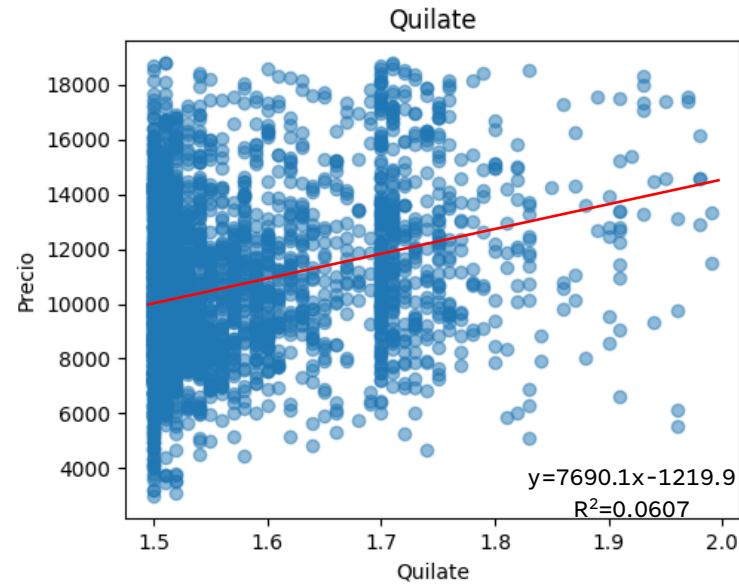
El valor del quilate no tiene injerencia.



VALORES DE 1.50 – 1.99 QUILATES

Para los diamantes de 1.5 a 2.0 quilates, el precio depende principalmente de la Claridad, y en menor medida, del Corte y Color.

El valor del quilate no tiene injerencia.



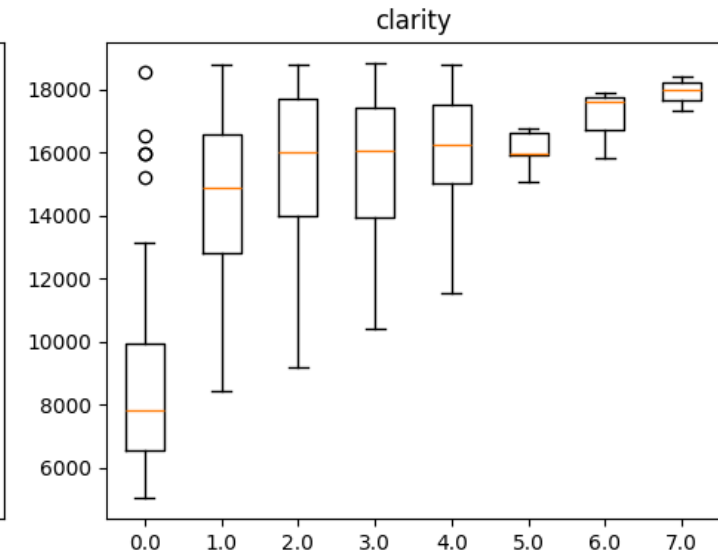
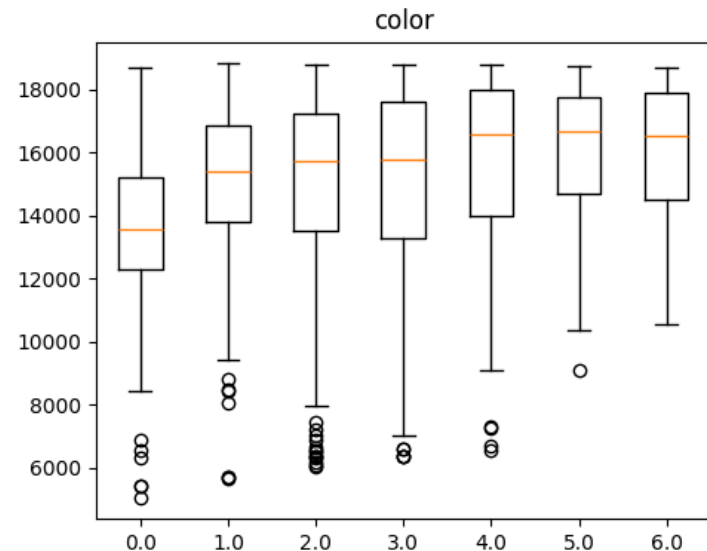
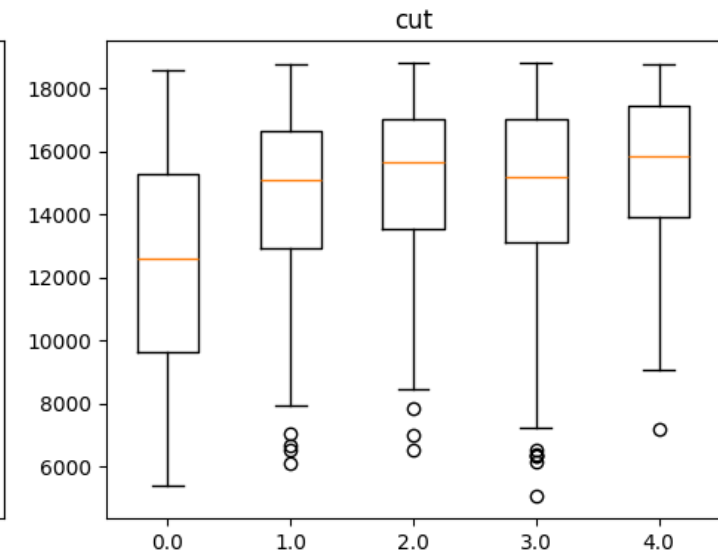
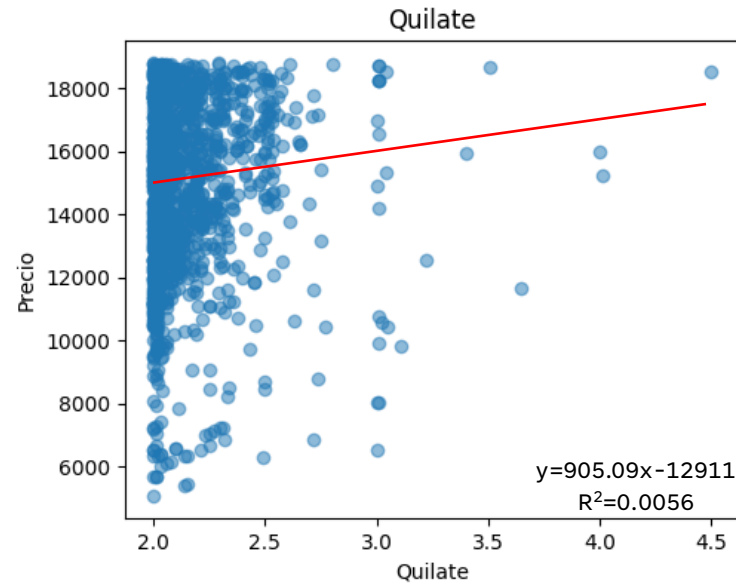
VALORES MAYOR O IGUAL A 2 QUILATES

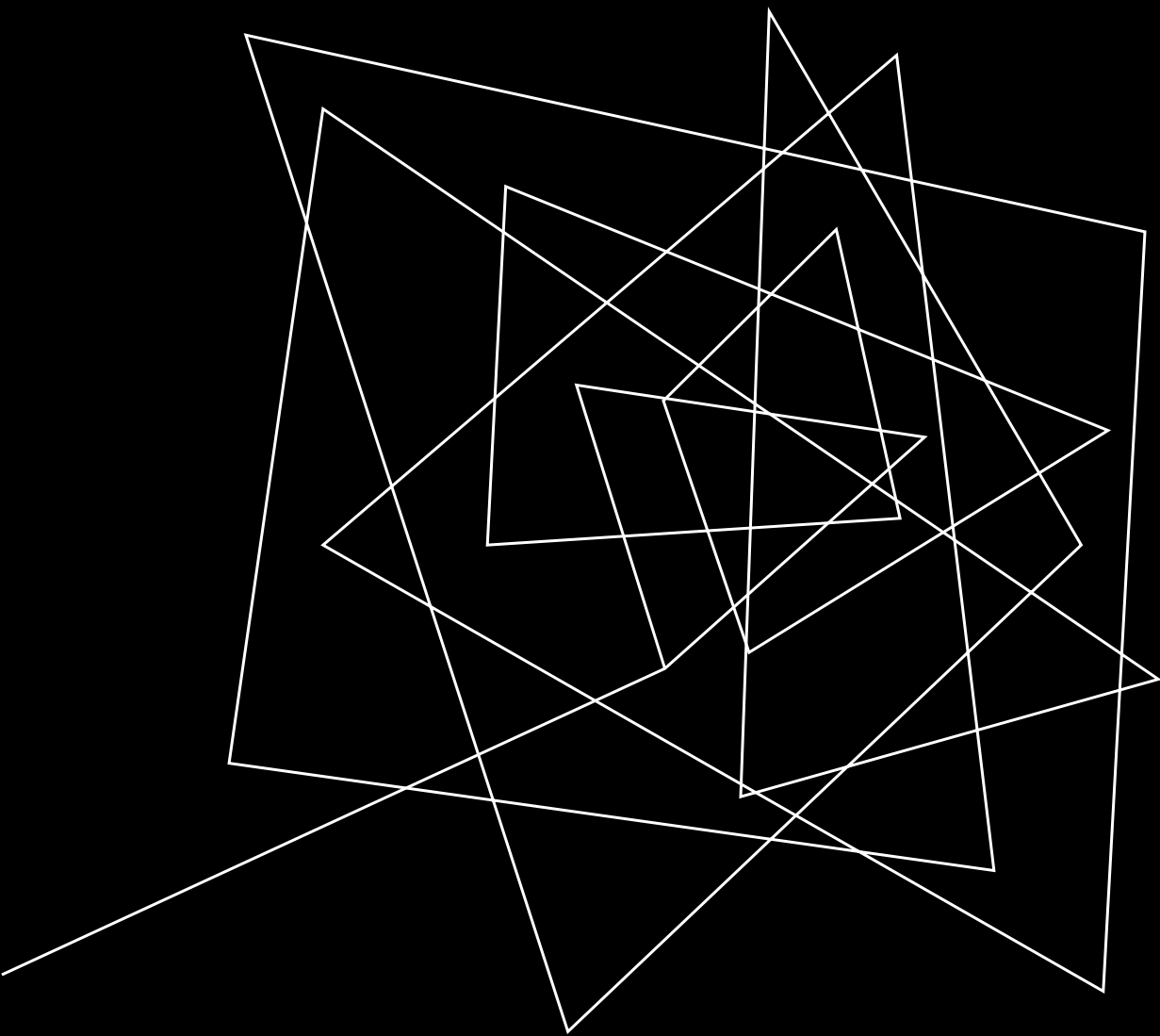
Para los diamantes mayor a 2.0 quilates, salvo para el menor valor de Claridad (I1), el precio depende muy ligeramente de la Claridad, del Corte y Color.

El valor del quilate no tiene relevante inferencia.

Las otras variables del dataset, como las geométricas (Profundidad, Tabla, X, Y, Z) no presentan relación con el precio.

Se deduce que debe haber otras características no presentes en el dataset que tengan inferencia en el precio del diamante. Probablemente, el establecimiento de venta.



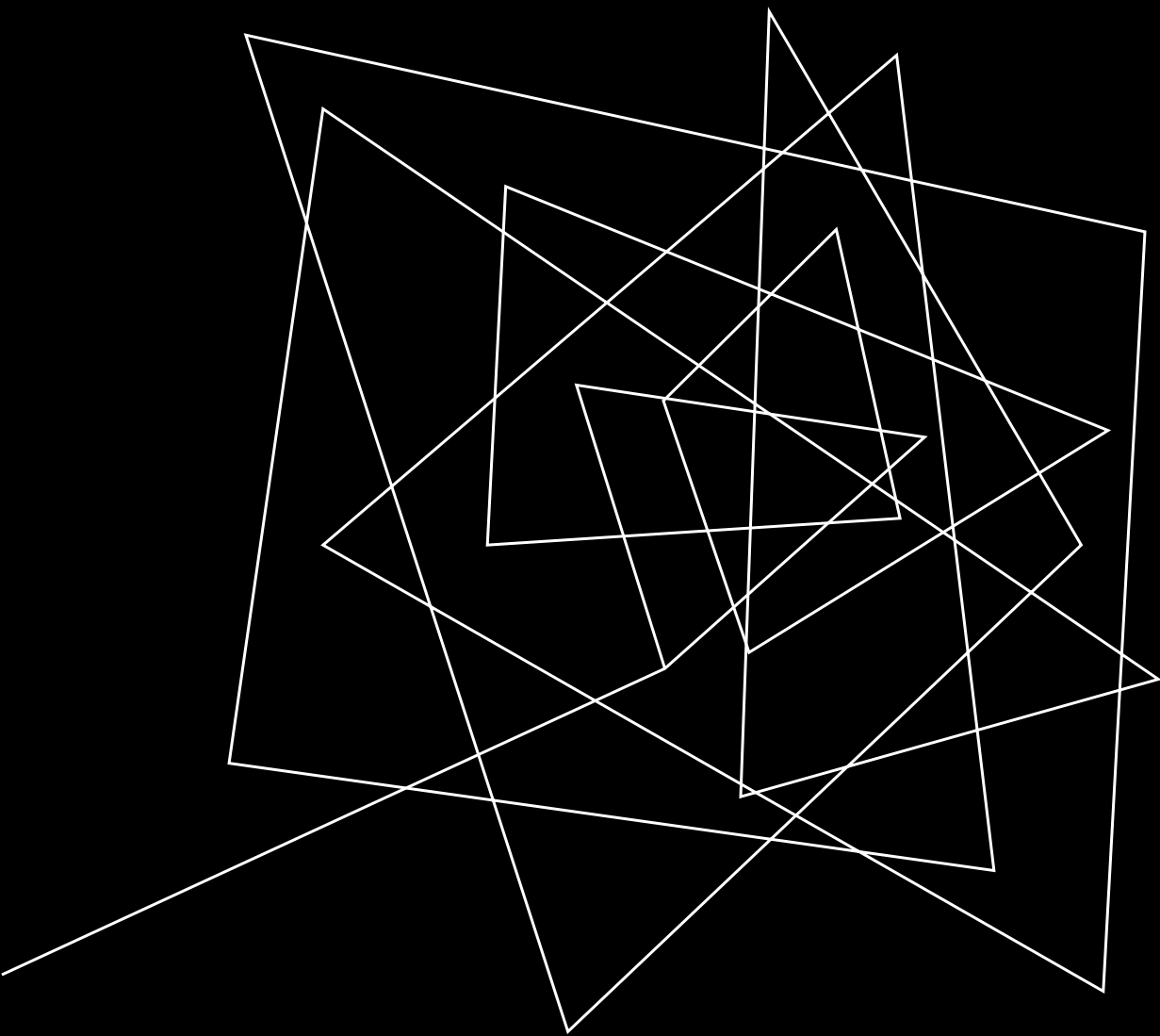


PRESELECCIÓN DE VARIABLES

PRESELECCIÓN DE VARIABLES

	feature	count	mean	std	min	Q1	median	Q3	max	missing_rate	corr	abs_corr
9	price	37719.0	7.792273	1.016902	5.786897	6.862758	7.785305	8.5897	9.842835	0.000000	1.000000	1.000000
1	x	30170.0	5.742950	1.125392	0.000000	4.720000	5.700000	6.5500	10.230000	0.200138	0.958041	0.958041
2	z	30207.0	3.542410	0.714943	0.000000	2.910000	3.530000	4.0400	31.800000	0.199157	0.924438	0.924438
3	carat	37719.0	0.800223	0.475464	0.200000	0.400000	0.700000	1.0400	4.500000	0.000000	0.920711	0.920711
0	y	30075.0	5.737424	1.167787	0.000000	4.720000	5.710000	6.5400	58.900000	0.202656	0.919583	0.919583
10	rango_carat	37719.0	1.513163	0.804918	1.000000	1.000000	1.000000	2.0000	4.000000	0.000000	0.798036	0.798036
6	clarity	37719.0	3.059466	1.650228	0.000000	2.000000	3.000000	4.0000	7.000000	0.000000	-0.214268	0.214268
8	table	37719.0	57.467523	2.221136	43.000000	56.000000	57.000000	59.0000	79.000000	0.000000	0.158284	0.158284
5	color	37719.0	3.402768	1.700173	0.000000	2.000000	3.000000	5.0000	6.000000	0.000000	-0.155016	0.155016

- Se puede observar que se ha seleccionado 9 variables de 11 las cuales pasan el proceso de preselección
- Las columnas con mejor correlación es carat, rango_carat , x, y, z
- Las columnas con menor correlación es table, clarity, color



PREPARACIÓN DE DATOS

PREPARACIÓN DE DATOS

El objetivo de esta fase es poder preparar los *features* para antes de poder utilizar los modelos de Machine Learning.

Dicho proceso consitión en las siguientes etapas:

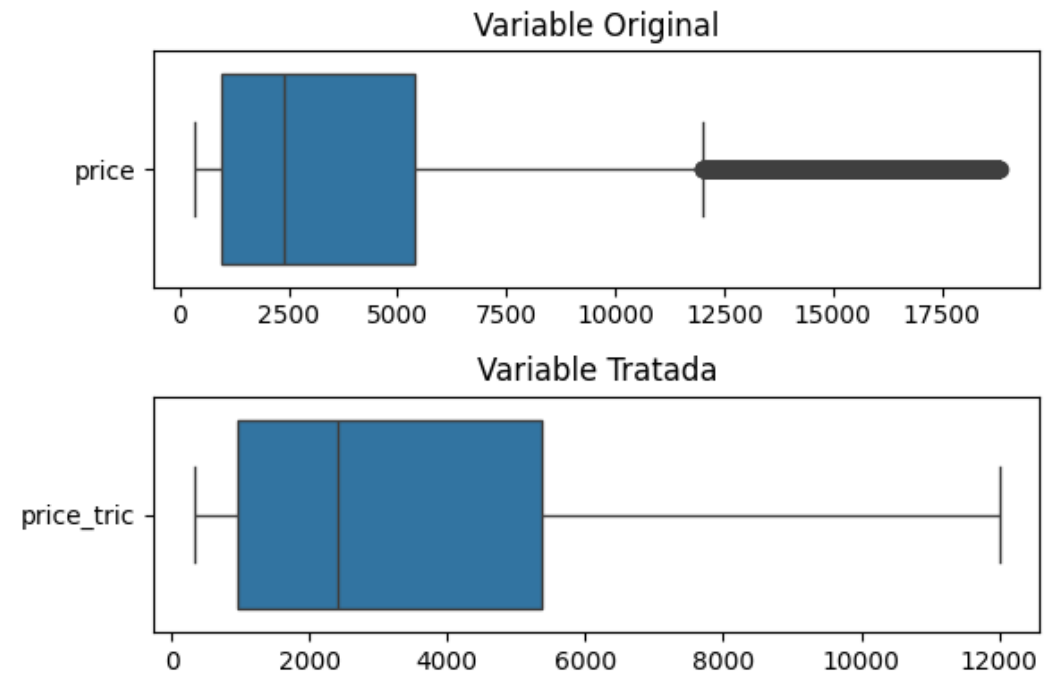
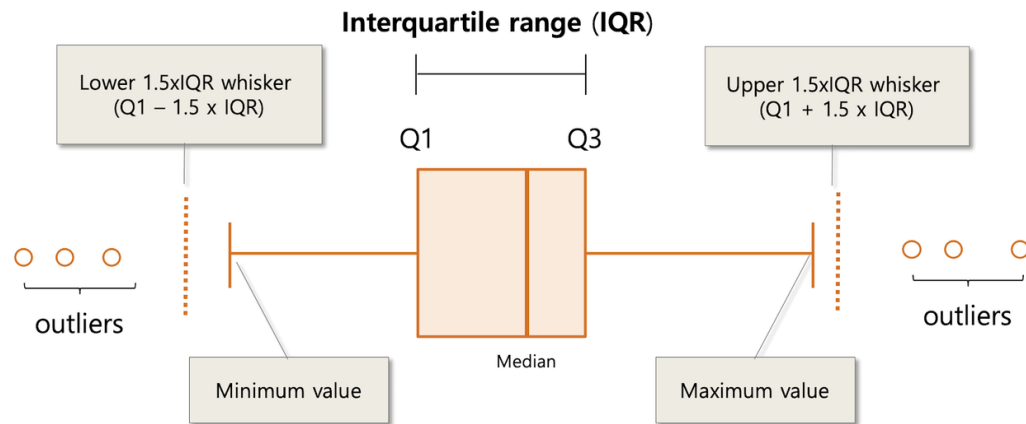
- 1.Tratamiento de *outliers*.
- 2.Tratamiento de *missings*.
- 3.Reescalamiento de datos por Z-Score.



TRATAMIENTO DE *OUTLIERS*

El tratamiento de *outliers*, utilizando el rango intercuartílico (IQR, por sus siglas en inglés) es una técnica común para identificar y manejar valores atípicos en un conjunto de datos.

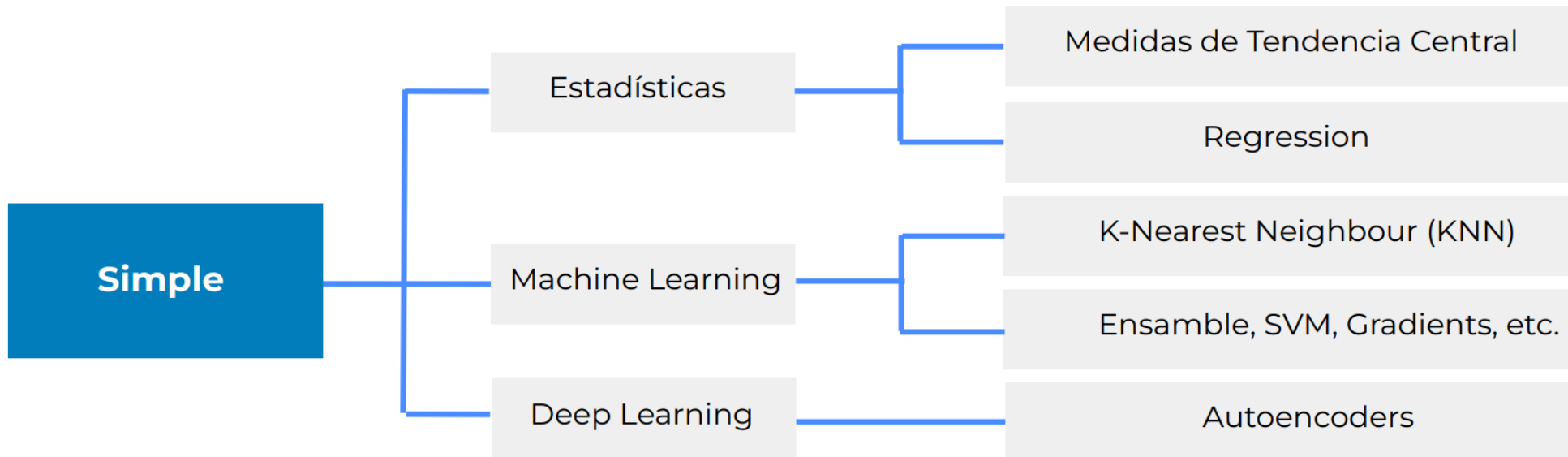
El rango intercuartílico es la diferencia entre el tercer cuartil (Q3) y el primer cuartil (Q1) de los datos.



TRATAMIENTO DE *MISSINGS*

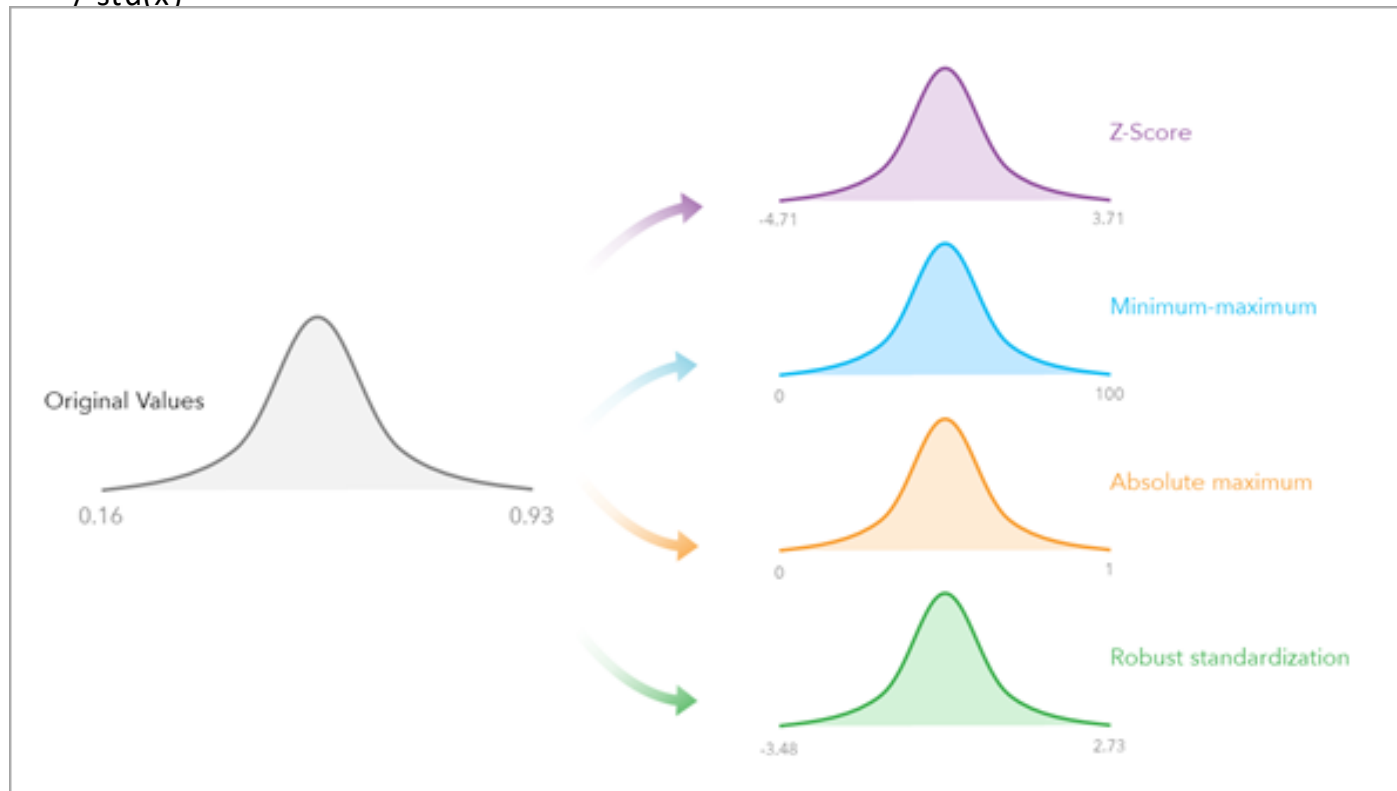
En lugar de eliminar los valores faltantes, se pueden reemplazar por valores estimados o imputados. Esto implica reemplazar los valores faltantes con medidas de tendencia central, como la media o la mediana, o utilizar métodos más avanzados como la regresión o el algoritmo K-NN para estimar los valores faltantes basándose en otros atributos del conjunto de datos.

En este caso los *missings* se trataron con Medidas de Tendencia Central, es decir la mediana.



REESCALAMIENTO DE DATOS POR Z-SCORE

- Es un proceso en el análisis de datos que consiste en transformar las variables para que tengan una escala común o un rango específico. El objetivo principal del reescalamiento de datos es colocar todas las variables en una misma escala numérica, lo cual puede facilitar la comparación y el análisis de las variables.
- Z-Score Scaling: También conocido como estandarización, transforma los datos para que tengan una media de 0 y una desviación estándar de 1. La fórmula para la estandarización es: $x_{std} = (x - \text{mean}(x)) / \text{std}(x)$





REGRESIÓN LINEAL POR MCO

MODELO INICIAL

OLS Regression Results

```
=====
Dep. Variable:          price_tric   R-squared:                0.939
Model:                  OLS          Adj. R-squared:            0.939
Method:                 Least Squares F-statistic:              7.207e+04
Date:                   Sun, 04 Aug 2024 Prob (F-statistic):        0.00
Time:                   21:26:22     Log-Likelihood:           -1524.1
No. Observations:      37719        AIC:                      3066.
Df Residuals:          37710        BIC:                      3143.
Df Model:               8
Covariance Type:       nonrobust
=====
               coef      std err          t      P>|t|      [0.025      0.975]
-----
const                7.7923      0.001    6005.943    0.000      7.790      7.795
x_tric_imp_std        0.0989      0.003     35.249    0.000      0.093      0.104
z_tric_imp_std        0.1005      0.003     36.752    0.000      0.095      0.106
carat_tric_imp_std    1.1265      0.006    182.746    0.000      1.114      1.139
y_tric_imp_std        0.1044      0.003     37.672    0.000      0.099      0.110
rango_carat_tric_imp_std -0.3610      0.004    -98.264    0.000     -0.368     -0.354
clarity_tric_imp_std  0.1866      0.001    130.748    0.000      0.184      0.189
table_tric_imp_std    -0.0044      0.001     -3.304    0.001     -0.007     -0.002
color_tric_imp_std     0.1471      0.001    107.350    0.000      0.144      0.150
=====
Omnibus:              1623.000   Durbin-Watson:           1.995
Prob(Omnibus):         0.000   Jarque-Bera (JB):        3436.851
Skew:                  -0.296   Prob(JB):                 0.00
Kurtosis:               4.355   Cond. No.                 11.8
=====
```

MODELO APLICANDO BACKWARD

OLS Regression Results

```

=====
Dep. Variable:    price_tric    R-squared:    0.939
Model:            OLS          Adj. R-squared: 0.939
Method:           Least Squares  F-statistic:  7.207e+04
Date:             Sun, 04 Aug 2024  Prob (F-statistic): 0.00
Time:             21:26:22      Log-Likelihood: -1524.1
No. Observations: 37719        AIC:            3066.
Df Residuals:     37710        BIC:            3143.
Df Model:         8
Covariance Type:  nonrobust
=====

```

	coef	std err	t	P> t	[0.025	0.975]
const	7.7923	0.001	6005.943	0.000	7.790	7.795
x_tric_imp_std	0.0989	0.003	35.249	0.000	0.093	0.104
z_tric_imp_std	0.1005	0.003	36.752	0.000	0.095	0.106
carat_tric_imp_std	1.1265	0.006	182.746	0.000	1.114	1.139
y_tric_imp_std	0.1044	0.003	37.672	0.000	0.099	0.110
rango_carat_tric_imp_std	-0.3610	0.004	-98.264	0.000	-0.368	-0.354
clarity_tric_imp_std	0.1866	0.001	130.748	0.000	0.184	0.189
table_tric_imp_std	-0.0044	0.001	-3.304	0.001	-0.007	-0.002
color_tric_imp_std	0.1471	0.001	107.350	0.000	0.144	0.150

```

=====
Omnibus:          1623.000    Durbin-Watson:    1.995
Prob(Omnibus):    0.000      Jarque-Bera (JB):  3436.851
Skew:             -0.296     Prob(JB):         0.00
Kurtosis:         4.355      Cond. No.         11.8
=====

```


MODELO CON CORRRELACIÓN NO MENOR AL 80%

OLS Regression Results

```
=====
Dep. Variable:      price_tric  R-squared:          0.897
Model:              OLS        Adj. R-squared:         0.897
Method:             Least Squares  F-statistic:       6.580e+04
Date:               Sun, 04 Aug 2024  Prob (F-statistic):    0.00
Time:               21:26:23    Log-Likelihood:     -11256.
No. Observations:   37719      AIC:                2.252e+04
Df Residuals:       37713      BIC:                2.258e+04
Df Model:           5
Covariance Type:    nonrobust
=====
```

```
=====
              coef      std err          t      P>|t|      [0.025      0.975]
-----
const                7.7923      0.002   4640.310      0.000      7.789      7.796
carat_tric_imp_std    0.9524      0.008   121.464      0.000      0.937      0.968
rango_carat_tric_imp_std -0.2928      0.005   -62.250      0.000     -0.302     -0.284
x_tric_imp_std         0.0971      0.004    26.743      0.000      0.090      0.104
z_tric_imp_std         0.0999      0.004    28.272      0.000      0.093      0.107
y_tric_imp_std         0.1084      0.004    30.241      0.000      0.101      0.115
=====
```

```
=====
Omnibus:             748.183  Durbin-Watson:          1.982
Prob(Omnibus):        0.000  Jarque-Bera (JB):      1547.748
Skew:                 0.079  Prob(JB):              0.00
Kurtosis:             3.980  Cond. No.              11.2
=====
```

PESOS DE WALD

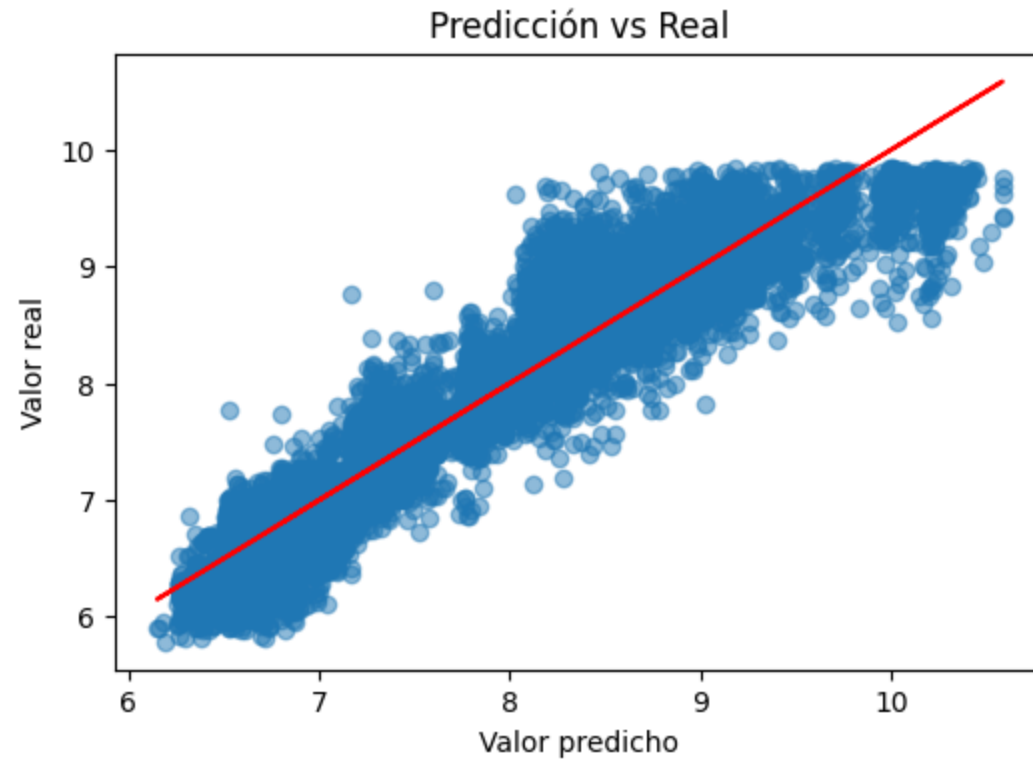
	features	F	Pesos_wald
0	carat_tric_imp_std	14753.617899	0.700629
1	rango_carat_tric_imp_std	3875.044534	0.184020
2	x_tric_imp_std	715.194809	0.033964
3	z_tric_imp_std	799.328873	0.037959
4	y_tric_imp_std	914.492825	0.043428



MÉTRICAS DE SEMEPEÑO

PRICE_TRIC VS PRICE_PRED

	price_tric	price_pred
33058	6.701960	6.534568
14626	8.683047	8.864513
48729	7.606885	7.293088
38826	6.954639	6.600312
48734	7.606885	7.302946
28906	6.527958	6.527120
10381	6.383507	6.467708
20566	9.091219	8.658088
12587	8.572249	8.844373
38428	6.933423	7.126886



PRICE_TRIC VS PRICE_PRED

	metric	train	test
0	r2	0.897156	0.897713
1	mape	0.032027	0.031803
2	mse	0.106347	0.104158
3	rmse	0.326109	0.322736
4	mae	0.250343	0.247888

Observaciones

- R^2 (R cuadrado): Train (0.897156) y Test (0.897713): El valor de R^2 indica la proporción de la varianza en la variable dependiente que es explicada por el modelo. Un valor cercano a 1 sugiere que el modelo explica muy bien la variabilidad de los datos. En este caso, ambos valores son muy altos y similares, lo que indica un buen ajuste del modelo tanto en el conjunto de entrenamiento como en el de prueba.

- MAPE (Mean Absolute Percentage Error): Train (0.032027) y Test (0.031803): El MAPE mide el error porcentual medio absoluto. Valores bajos indican que, en promedio, el modelo tiene un error pequeño en términos porcentuales. Aquí, ambos valores son bajos y muy similares, lo que sugiere que el modelo es preciso y consistente en ambos conjuntos.

- MSE (Mean Squared Error): Train (0.106347) y Test (0.104158): El MSE mide el promedio de los errores al cuadrado. Valores bajos indican que el modelo tiene pocos errores grandes. En este caso, los valores son bastante bajos y similares, lo que sugiere que el modelo tiene un buen desempeño tanto en entrenamiento como en prueba.

- RMSE (Root Mean Squared Error): Train (0.326109) y Test (0.322736): El RMSE es la raíz cuadrada del MSE y proporciona una medida de la magnitud promedio del error. Al igual que el MSE, valores bajos indican un buen rendimiento del modelo. Aquí, los valores son bajos y cercanos entre sí, indicando un buen rendimiento del modelo en ambos conjuntos.

- MAE (Mean Absolute Error): Train (0.250343) y Test (0.247888): El MAE mide el error absoluto medio entre las predicciones y los valores reales. Valores más bajos indican menos errores. En este caso, los valores son bajos y muy similares, lo que sugiere que el modelo tiene un rendimiento constante y preciso en ambos conjuntos.

CONCLUSIONES Y RECOMENDACIONES

- La variable 'Precio' presentó un comportamiento lognormal (algo que se esperaba, ya que hay mayor número de valores bajos y menor cantidad de valores altos). Este cambio de variable permitió un mejor ajuste en el modelo de regresión lineal.
- Es importante discernir de manera sustentada la no dependencia de la variable objetivo con respecto a otras variables, ya que ello permitirá que el modelo sea más limpio y eficiente.
- De los resultados obtenidos del modelo desarrollado se puede concluir que existe buena correlación de las variables empleadas.
- Debido al hallazgo de alto contraste de los precios alrededor de ciertos valores de quilate (1, 1.5 y 2), se recomienda desarrollar modelos para cada grupo de quilates.

A series of white, thin, overlapping geometric lines on a black background, forming various polygons and intersecting points, primarily located on the left side of the slide.

GRACIAS