Costo de Diamantes.

❖ Introducciò.

El valor de los diamantes tanto en el mundo mágico como en el mundo de los muggles, es muy importante para la economía, por lo que el robo de estos significa una gran perdida, determinemos el valor de estos.

❖ Datos.

En la tabla 1 se da un muestreo de los primeros datos correspondientes a las características de los diamantes.

	carat	cut	color	clarity	depth	table	price	x	y	z
0	0.23	Ideal	E	SI2	61.5	55.0	326	3.95	3.98	2.43
1	0.21	Premium	E	SI1	59.8	61.0	326	3.89	3.84	2.31
2	0.23	Good	E	VS1	56.9	65.0	327	4.05	4.07	2.31
3	0.29	Premium	I	VS2	62.4	58.0	334	4.20	4.23	2.63
4	0.31	Good	J	SI2	63.3	58.0	335	4.34	4.35	2.75

Tabla 1. Muestreo de datos.

Se valida el formato y datos faltantes.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 53930 entries, 0 to 53929
Data columns (total 10 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   carat    53930 non-null  float64
 1   cut      53930 non-null  object
 2   color    53930 non-null  object
 3   clarity  53930 non-null  object
 4   depth    53930 non-null  float64
 5   table    53930 non-null  float64
 6   price    53930 non-null  int64
 7   x        53930 non-null  float64
 8   y        53930 non-null  float64
 9   z        53930 non-null  float64
dtypes: float64(6), int64(1), object(3)
memory usage: 4.1+ MB
```

Tabla 2. Validación de datos.

❖ **Análisis.**

Para el modelo de predicción a usar, se determino que variable o variables usar de acuerdo a la matriz de correlación lineal.

price	1.000000
carat	0.921590
x	0.884431
y	0.865416
z	0.861245
table	0.127168
depth	-0.010633

Tabla 3. Correlación lineal.

Los atributos o características que se consideran en dicha matriz son “carat”, “x”, “y”, “z”, “table” y “depth” observando que el más cercano es “carat”.

En la tabla 1 nos muestra 10 atributos los 3 faltantes son datos categóricos por lo que no se consideran para una predicción y/o regresión.

De manera gráfica se determinó dicha decisión.

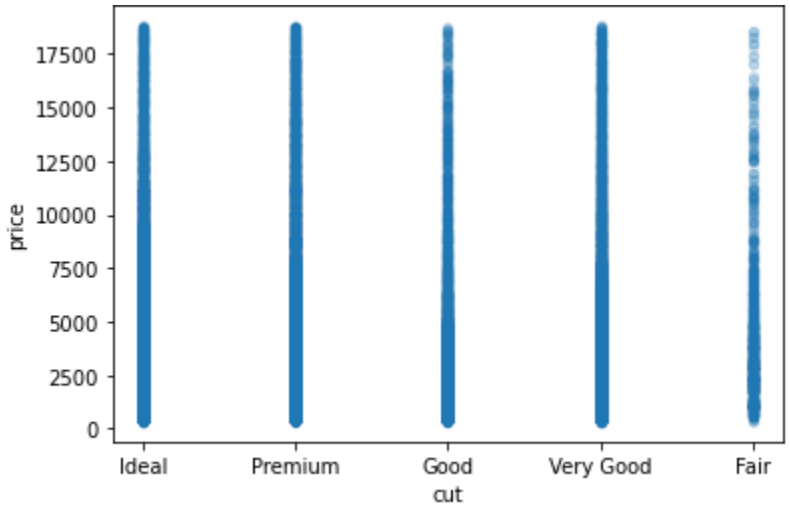

Grafico 1. Relación Cut vs Price.

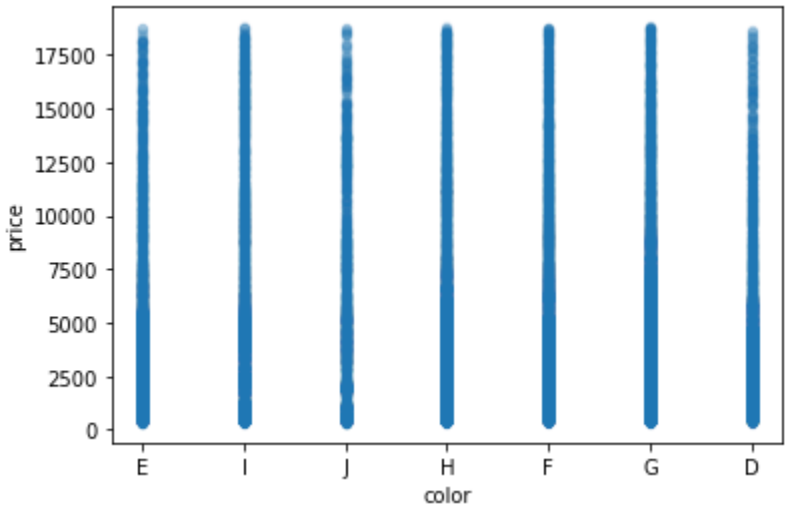

Grafico 2. Relación Color vs Price.

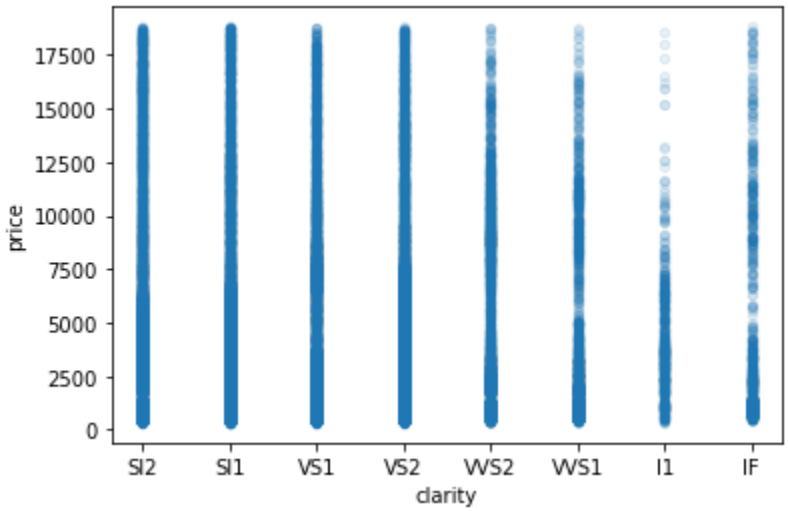

Grafico 3. Relación Darity vs Price.

- **Modelado.**

Se realizaron 3 modelos predictivos utilizando el atributo “carat” para determinar el “price”.

Se dividió el data set en 80% para valores de entrenamiento y 20% para valores de prueba.

Para el modelo regresión lineal, se obtuvo:

Coeficiente de determinación: 0.847632782481173

Raíz del error cuadrático medio: 1562.8016139550866

Desviación estándar: 36.9394430139873

Para el modelo de regresión árbol de decisiones, se obtuvo:

Coeficiente de determinación: 0.8748534804765188

Raíz del error cuadrático medio: 1416.3398371618766

Desviación estándar: 46.14879456566654

Para el modelo regresión polinomial, se obtuvo:

Coeficiente de determinación: 0.8489726864978453

Raíz del error cuadrático medio: 1555.914869646454

Desviación estándar: 39.10213091539103

En la tabla 4 se da un muestreo de los valores de predicción para cada modelo con los datos de prueba.

	carat	cut	color	clarity	depth	table	price	x	y	z	predic_lin	predic_tree	predic_lin2
34222	0.40	Premium	H	SI1	62.8	59.0	855	4.69	4.67	2.94	848.086206	940.028249	914.054991
30645	0.40	Ideal	I	VS1	61.9	57.0	736	4.73	4.74	2.93	848.086206	940.028249	914.054991
48886	0.56	Very Good	E	VVS2	62.1	58.0	2040	5.27	5.33	3.30	2090.729716	1831.033163	2071.716788
43004	0.42	Ideal	E	VVS1	59.7	56.0	1369	4.91	4.88	2.92	1003.416644	999.595782	1057.488320
29996	0.43	Ideal	E	SI2	61.9	57.0	716	4.83	4.86	3.00	1081.081864	995.180905	1129.341527
40797	0.41	Good	E	VVS1	60.4	61.0	1169	4.76	4.81	2.89	925.751425	975.059397	985.726141
22090	1.25	Ideal	D	VS2	62.6	56.0	10114	6.87	6.84	4.29	7449.629856	7018.156863	7331.073516
14073	1.03	Premium	E	SI1	62.9	56.0	5710	6.49	6.39	4.05	5740.995029	5537.577017	5607.115583
39075	0.38	Very Good	F	VVS2	61.9	59.7	1060	4.64	4.65	2.87	692.755767	906.465028	770.985774
19315	1.41	Ideal	H	SI2	62.0	56.0	8030	7.17	7.24	4.47	8692.273367	9627.212121	8612.533678

Tabla 4. Predicción de “price”.

❖ Resultados.

Del modelo de regresión lineal y regresión polinomial se obtuvieron valores de error muy cercanos, o iguales siendo mayores a los valores arrojados por la regresión de decisión de árbol.

Por lo que se propone optar por los valores de predicción de regresión de decisión de árbol.

Sin embargo, se podría proponer una variable dummy para cada atributo que se considero categórico y determinar un modelo que se pueda comparar con el seleccionado anteriormente.