María Fernanda Ledesma Martínez

A01734203

Actividad 1. Análisis de correlación utilizando plataformas de Analítica: Python

Para esta actividad se ocupó el DataFrame “diamonds” el cual tiene variables como precio, tamaño, quilate, profundidad, corte, claridad, entre otras; las cuales definen las características de diferentes diamantes. Con ello, realizamos un análisis de correlación con regresiones lineales en dónde la pregunta de investigación fue qué variables afectan el precio el precio de los diamantes.

Análisis

Modelo	Coeficiente de determinación (r^2)	Coeficiente de correlación	Análisis
Carat	0.849334334	0.921593367	Coeficientes muy altos
Depth	0.000113891	0.010671992	Resultado muy bajo
Table	0.016175846	0.127184299	Resultado muy bajo
Carat y Depth	0.85068029	0.922323311	Coeficientes muy altos
Carat y Table	0.851008664	0.922501308	Coeficientes muy altos
Depth y Table	0.01697079	0.130271985	Resultado muy bajo
Tamaño (x, y, z)	0.782559722	0.884624057	Coeficientes altos
Tamaño, carat y depth	0.856340782	0.925386828	Coeficientes muy altos
Tamaño, carat, depth y table	0.85921951	0.926940942	Coeficientes muy altos

Como se puede observar se probaron diferentes modelos de regresión lineal y se obtuvieron los coeficientes de determinación y de correlación para ver cuál modelo puede describir mejor nuestra variable respuesta. Dentro de estas, primero se realizaron los modelos de regresión simple y posteriormente los múltiples.

Cuando realizamos las regresiones simples, notamos que la variable carat es la que mayor correlación tiene, con un 92%, lo que nos dice que el quilate del diamante si influye o está

relacionado con el precio. Al mismo tiempo que obtuvimos un coeficiente de determinación de 84%, lo que explica que la variabilidad del precio es explicada por esta.

Por otro lado, con las otras regresiones simples con Depth y Table, la correlación y determinación salieron demasiado bajas, prácticamente nulas. Lo cual dice que estas dos variables no explican el precio. Y esto también lo podemos ver cuando hacemos una regresión múltiple con estas mismas variables, y nuestros resultados son igual de desfavorables.

Sin embargo, cuando juntamos cada una de estas variables anteriores con Carat (que salió muy correlacionada) en un modelo múltiple. Nuestros resultados resultan muy buenos e inclusive un poco mayores al modelo simple. Pero, al ser modelos múltiples la variabilidad podría aumentar y nuestro modelo hacerse más complejo. Como sabemos que Depth y table no nos generan valor, considero que no es el mejor modelo.

Y al igual que carat, tenemos otra variable que puede explicar el precio de los diamantes, este es el tamaño. Pues obtuvimos una correlación y determinación fuerte, lo que nos dice que el tamaño nos puede dar información sobre el precio.

Finalmente, si agregamos todas las variables a nuestro modelo, obtenemos mejores resultados debido a qué todas abonan (sea mínimo o no) a la relación y la varianza también aumenta, por lo que, mayor variabilidad (coeficiente de determinación) más podría explicar el modelo. En este caso, todas nuestras variables explican en cierta medida el precio de los diamantes.

A consideración propia, un mejor modelo es el más simple. Por ello, recomendaría uno que incluya las variables tamaño y carat con las que también obtenemos buenos resultados juntas y por sí mismas. Además, se recomendaría ver primero otros criterios de evaluación y los errores de cada modelo para elegir el mejor con mayor certeza.