UNIVERSIDAD AMERICANA
Facultad de Ingeniería y Arquitectura

Inteligencia de Negocio

Análisis de Datos de Fuentes Diversas.

Estudiante:
Lester Alejandro Rodríguez Cuevas

Docente:
Arlen Jeannette Lopez

Fecha:

18 de Septiembre 2024

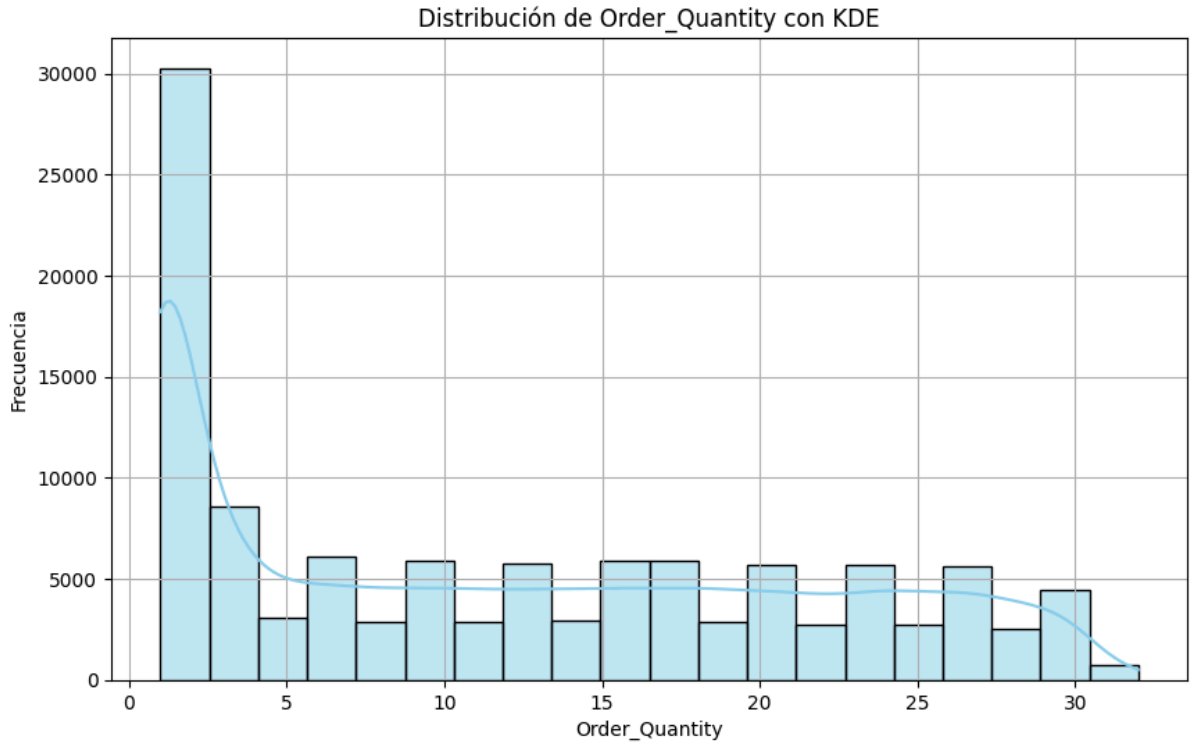
Recopilación y almacenamiento de datos

El dataset lo obtuvimos de Kaggle, verificamos la cantidad de registros y columnas para confirmar que fuera un rango responsable para el análisis, procesamiento y gráficas. Este dataset tiene 113036 registros y 18 columnas.

El tipo de archivo es un .CSV, comma separated values, un archivo en el que los valores están separados por comas.

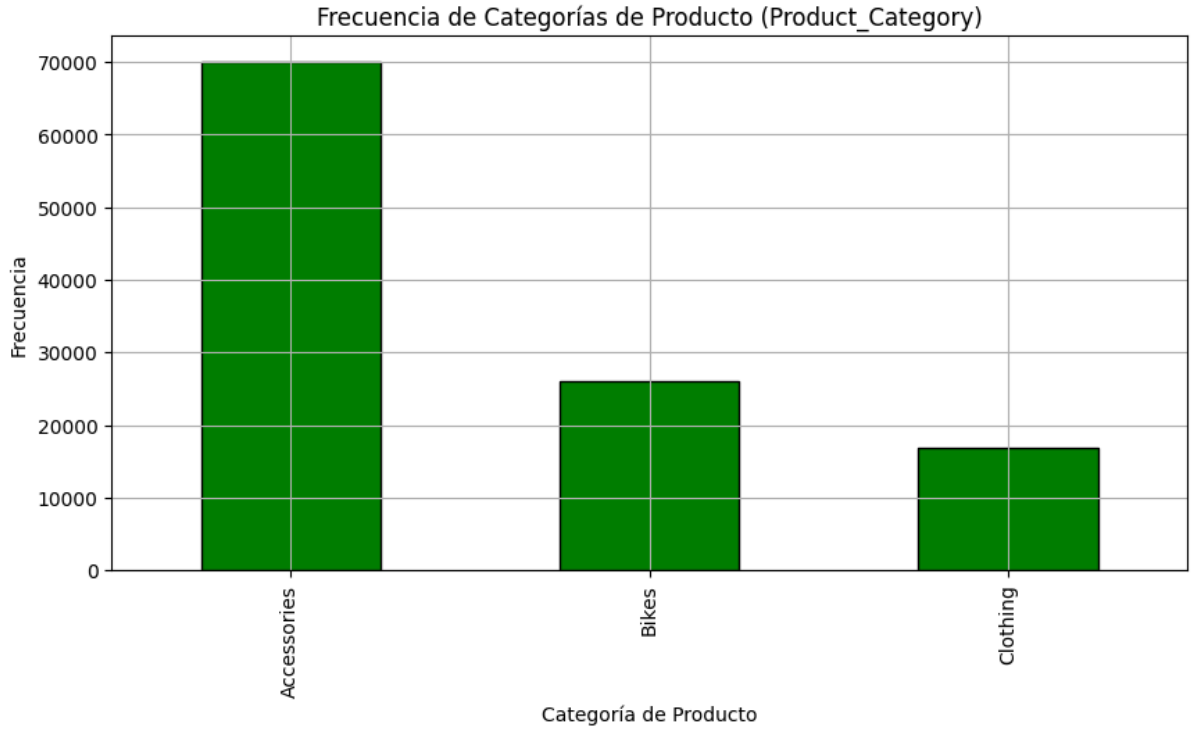
```
Product_Category
Accessories    70120
Bikes          25982
Clothing       16934
Name: count, dtype: int64
```

Se genera una tabla de frecuencias para la columna Product_Category en un dataset, utilizando la función value_counts() de pandas. El resultado muestra la cantidad de apariciones de cada categoría de producto: Accessories aparece 70,120 veces, Bikes 25,982 veces, y Clothing 16,934 veces. Esto proporciona una visión clara de la distribución de las distintas categorías de productos en el dataset, permitiendo identificar cuál es la categoría más frecuente.



En este histograma muestra la distribución de la cantidad de órdenes (Order_Quantity) con una curva KDE superpuesta, lo que indica que la mayoría de las órdenes tienen cantidades

pequeñas (entre 1 y 5), con una caída significativa en frecuencias a medida que la cantidad de órdenes aumenta.



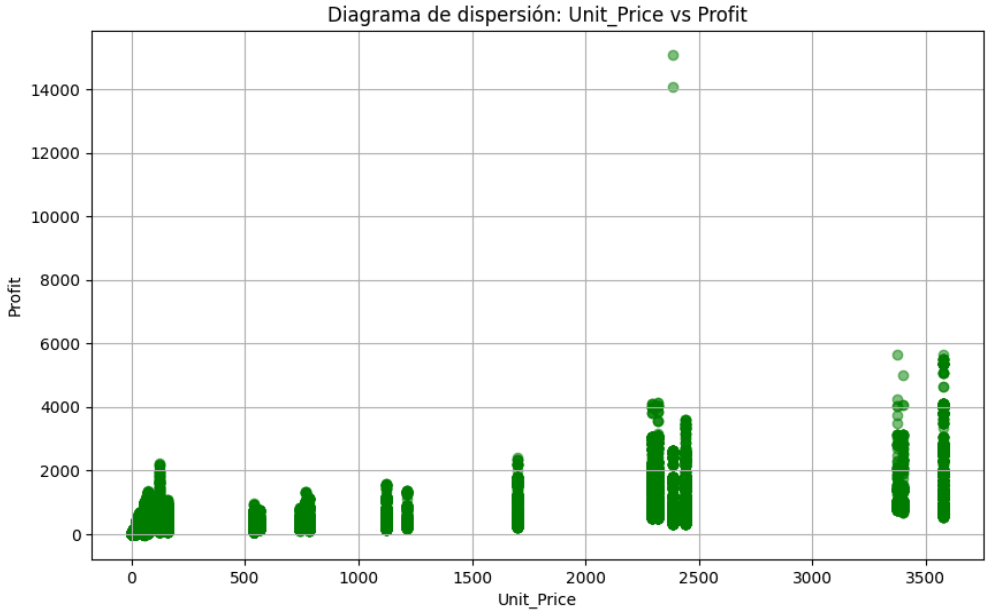
Es una tabla de frecuencias de la categoría de productos (Product_Category). Se observa que la categoría Accessories tiene una frecuencia mucho mayor (más de 70,000 apariciones), seguida por Bikes y Clothing, lo que indica que los accesorios son los productos más vendidos en comparación con las otras categorías.

```
{'correlation_Orden_Ganancia': -0.23886342119372153,
 'correlation_Costo_precio': 0.9978935825333143,
 'r_value_Orden_Ganancia': -0.23886342119371676,
 'p_value_Orden_Ganancia': 0.0,
 'r_value_costo_precio': 0.9978935825333042,
 'p_value_costo_precio': 0.0}
```

Relación entre Order_Quantity y Revenue:
Correlación: La correlación entre la cantidad de órdenes y los ingresos es de -0.313, lo que sugiere una relación negativa moderada. Esto significa que, a mayor cantidad de órdenes, los ingresos tienden a disminuir, lo cual podría ser consecuencia de descuentos o precios reducidos para órdenes grandes.
Regresión: El valor de r es -0.313 y el p-valor es muy bajo (0.0), lo que indica que esta relación es estadísticamente significativa.

Relación entre Customer_Age y Profit:
Correlación: La correlación entre la edad del cliente y las ganancias es prácticamente nula (0.004), lo que indica que no hay una relación significativa entre la edad del cliente y las ganancias obtenidas por la transacción.
Regresión: El valor de r es también muy bajo (0.004) y el p-valor es 0.146, lo que indica que no es estadísticamente significativo.

Mientras que la cantidad de órdenes tenga un impacto moderado en los ingresos, no hay evidencia de que la edad del cliente afecte significativamente las ganancias.

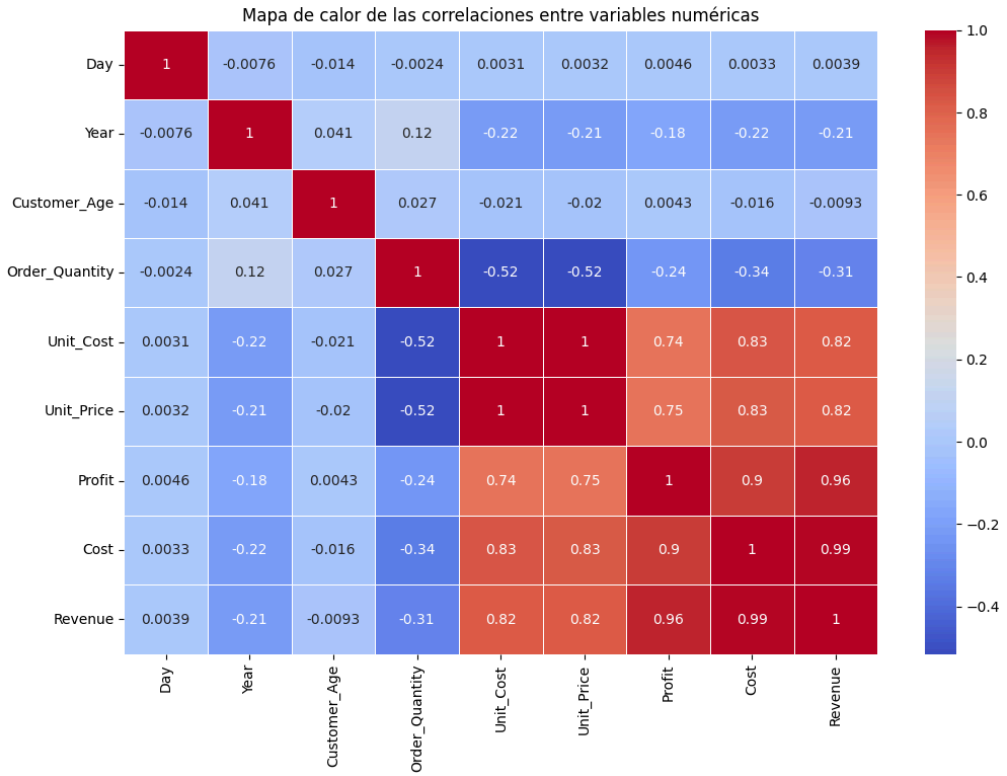


El diagrama de dispersión que realizamos, muestra que a medida que aumenta el precio unitario, hay una tendencia a que las ganancias sean mayores, aunque con bastante variabilidad.
Sin embargo, hay varios grupos de puntos (clusters) que indican que las ganancias no siempre siguen un aumento lineal con el precio unitario, ya que algunos productos con precios similares tienen ganancias muy diferentes. Esto podría estar influenciado por otros factores, como descuentos, promociones, o costos variables de los productos. Del mismo modo vemos que tenemos dos casos en los cuales los datos no son nada típicos, es probable que estos puntos representan outliers o transacciones atípicas, donde los productos fueron vendidos a precios muy altos y generaron ganancias inusualmente elevadas. Esto podría ser el resultado de un producto de lujo o de un volumen de ventas muy grande en una sola transacción.

	Unit_Price	Order_Quantity	Profit	Revenue	Cluster	Anomaly
0	120	8	590	950	0	1
1	120	8	590	950	0	1
2	120	23	1366	2401	2	-1
3	120	20	1188	2088	2	-1
4	120	4	238	418	0	1

Clustering (K-Means):
Los datos se agruparon en 4 clusters (columna Cluster), basados en las variables Unit_Price, Order_Quantity, Profit, y Revenue. Cada fila tiene asignado un número de clúster (0, 1, 2 o 3), que representa a qué grupo pertenece la transacción.

Detección de anomalías (Isolation Forest):
La columna Anomaly muestra si una transacción es un punto anómalo. Un valor de -1 indica que es un outlier o anomalía, mientras que 1 significa que la transacción es normal.
Como puedes ver en los resultados, algunas transacciones (como las filas 2 y 3) fueron detectadas como anomalías.



Aca podemos observar un mapa de calor, este muestra las correlaciones entre las variables numéricas en los datos de ventas de bicicletas. Podemos deducir y analizar que:

Cost y **Revenue** tienen una correlación muy alta (0.99), lo que es lógico ya que los ingresos están directamente relacionados con los costos de los productos vendidos.
Profit (ganancias) está altamente correlacionado con **Unit_Cost** (0.74) y **Unit_Price** (0.75), lo que indica que tanto el costo como el precio unitario tienen una fuerte influencia en las ganancias.
Order_Quantity tiene una correlación negativa moderada con **Unit_Cost** y **Unit_Price** (ambos en -0.52), lo que sugiere que, a medida que aumenta la cantidad de órdenes, los productos más baratos tienden a ser vendidos.

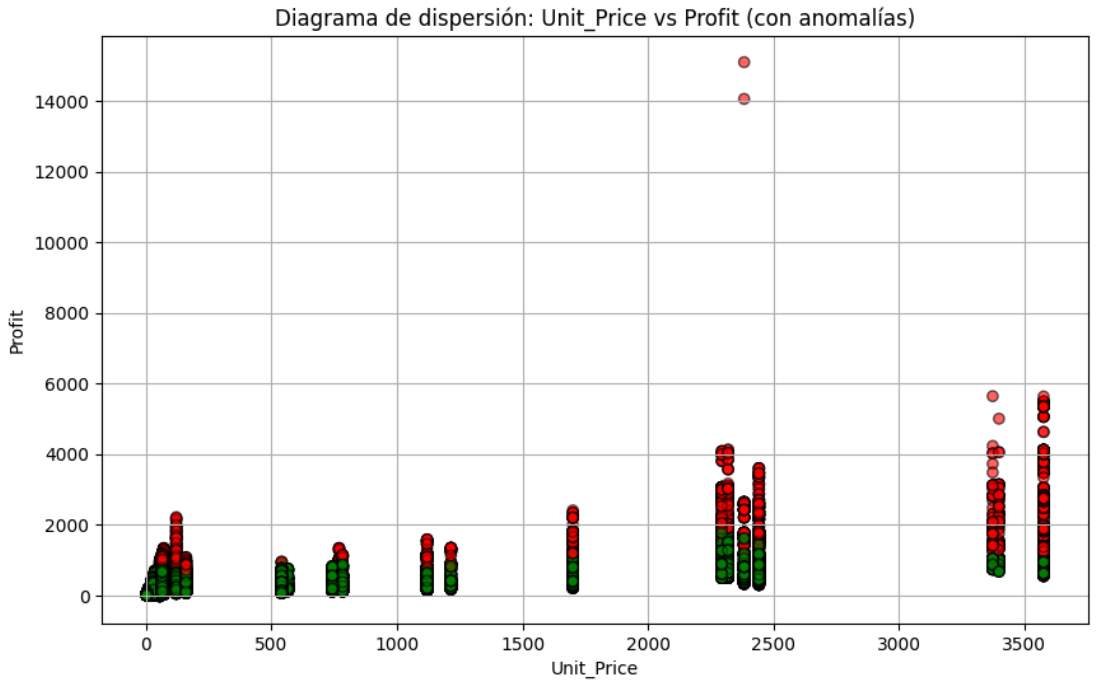


Gráfico de Distribución General:
La mayoría de los datos se concentran en precios unitarios bajos (menores a 1000) y ganancias moderadas (menores a 2000).
Hay un número considerable de puntos que se dispersan hacia la derecha, lo que representa transacciones con precios más altos y ganancias variables.

Los **Puntos de Anomalía (en rojo)** representan las transacciones detectadas como anomalías. Estos puntos se dispersan principalmente en dos áreas:

Zona de ganancias extremadamente altas: Observamos dos puntos dispersos muy notables, con precios unitarios de entre 2500 y 3000, y ganancias superiores a 14,000. Estos son los puntos más aislados en el gráfico y representan transacciones extremadamente inusuales. Son claramente outliers en comparación con el resto de los datos.

Zona de transacciones comunes: Dentro de las transacciones más frecuentes (con precios menores a 1000), algunos puntos rojos también aparecen, lo que sugiere que, aunque las ganancias y precios estén dentro de un rango razonable, hay algo inusual en estas transacciones en comparación con el resto.

Entonces podemos concluir que las anomalías detectadas en las transacciones de precios y ganancias extremadamente altos son las más evidentes y pueden estar relacionadas con productos exclusivos o condiciones especiales de venta. Las anomalías en la zona de precios bajos pueden indicar posibles errores de registro o promociones inusuales que afectaron el comportamiento normal de las ganancias.