Taller: Procesamiento y análisis
Estudiante: Dayana Sofía Del Valle Correa, CC.1003064124

1. **Identificación de datos duplicados, justificación y uso de la estrategia de tratamiento de datos duplicados.**
   Para identificar datos duplicados se utiliza la función *duplicated* de la librería Pandas, la cual arroja resultados de tipo booleano indicando en cuales filas hay datos duplicados (True) y en cuales no (False). En este caso el resultado obtenido señala que no hay filas que contengan datos duplicados, por lo tanto, no es necesario tratarlos.

2. **Identificación de datos faltantes, justificación y uso de la estrategia de tratamiento de datos faltantes.**
   Al utilizar la función 'info', es evidente que existen valores faltantes en las columnas 'precio', 'variabilidad', 'LATITUD' y 'LONGITUD', los cuales necesitan ser abordados. Esta observación se confirma al emplear la función 'isnull', que identifica las columnas específicas que contienen datos no disponibles. Además, para cuantificar la cantidad de valores faltantes en cada columna podemos utilizar 'isnull().sum()', lo que atribuye 2186 datos faltantes en 'precio', 2329 en 'variabilidad', 805 en 'LATITUD' y 805 en 'LONGITUD'. La cantidad de datos faltantes es significativa, por lo tanto, deberán ser imputados para una posterior interpretación adecuada.
   Para abordar los datos faltantes detectados, se aplicó la interpolación lineal a las variables 'precio' y 'variabilidad', un método que estima valores intermedios basados en los datos existentes. En el caso de las variables 'LATITUD' y 'LONGITUD', se llevó a cabo un proceso que inició con la identificación de las ciudades que tenían datos faltantes. Luego, se utilizó la función 'fillna' para sustituir estos valores faltantes con las respectivas coordenadas de latitud y longitud correspondientes a las ciudades.

3. **Identificación de datos outliers, justificación y uso de la estrategia de tratamiento de datos outliers.**
   Considerando que la variable ‘precio’ abarca una cantidad considerable de productos y que además esos productos se distribuyen en diversas ciudades, no es práctico eliminar los outliers identificados al graficar dicha variable, ya que en este caso se espera que haya puntos atípicos que se le atribuyen a los diversos precios propios de cada producto, para que el análisis sea significativo debería realizarse producto a producto.

4. **Justificación de la necesidad de normalizar o estandarizar variables.**
   La estandarización o normalización de variables es primordial cuando se están analizando medidas con diferentes unidades. Sin embargo, en este caso, dadas las características de las variables, no se justifica la normalización. Algunas variables son de tipo 'object', mientras que LATITUD y LONGITUD representan coordenadas geográficas, lo que haría que la normalización sea inapropiada ya que pueden perder su significado. La variable fecha por su parte, tiene un significado específico que no se beneficia de la normalización.

5. **Justificación de la necesidad de redefinir variables y redefinición si es necesaria.**

No es necesario redefinir variables ya que el tipo de datos que está asignado actualmente en la base de datos corresponde a la naturaleza de estas, por lo tanto, se pueden realizar análisis con ellas sin la necesidad de modificar su formato.

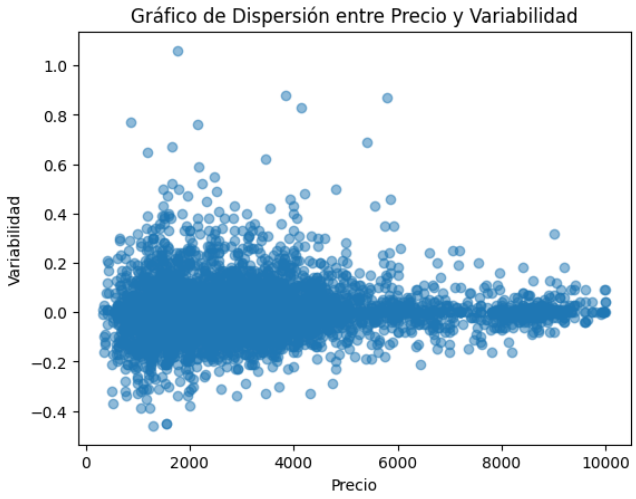
6. **Justificación de la necesidad de categorizar variables y categorización si es necesaria**

   La necesidad de categorizar variables depende del análisis que se quiera realizar, hay diversas formas de hacerlo. En este caso se optó por crear una categoría basada en el rango de precio de los productos, la cual clasifica cada producto en 'Económico', 'Moderado' y 'Costoso'.

   Para hacerlo, se identificaron los precios mínimos y máximos y se definieron los límites y los nombres de las categorías para posteriormente adjuntar la nueva variable al dataframe.

7. **Mínimo 5 preguntas que quiera resolver a partir del filtrado de columnas o filas.**
   - ¿Cuántos productos se encuentran en la categoría de precios económicos?
   - ¿En qué ciudad se encuentra el producto más caro?
   - ¿Cuáles productos tienen una variabilidad inferior al promedio?
   - ¿Cuál es el precio promedio de la Arveja verde en vaina en Neiva?
   - ¿En qué fecha se registró el precio más bajo y alto?

8. **Mínimo 5 gráficos que sean de interés para entender el problema.**



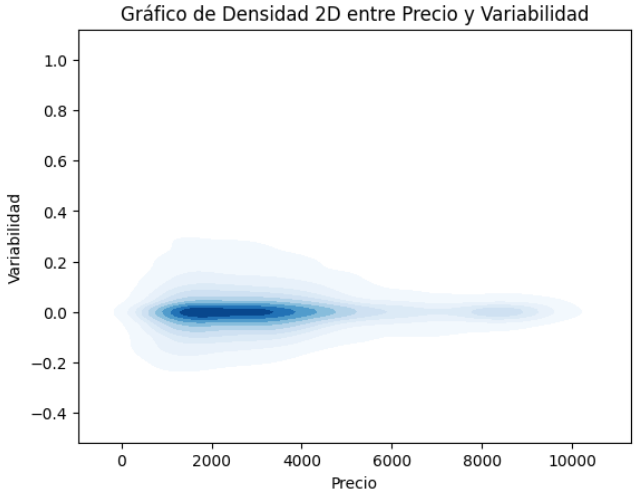


Gráfico de Densidad 2D entre Precio y Variabilidad
Variabilidad
Precio
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
2000
4000
6000
8000
10000

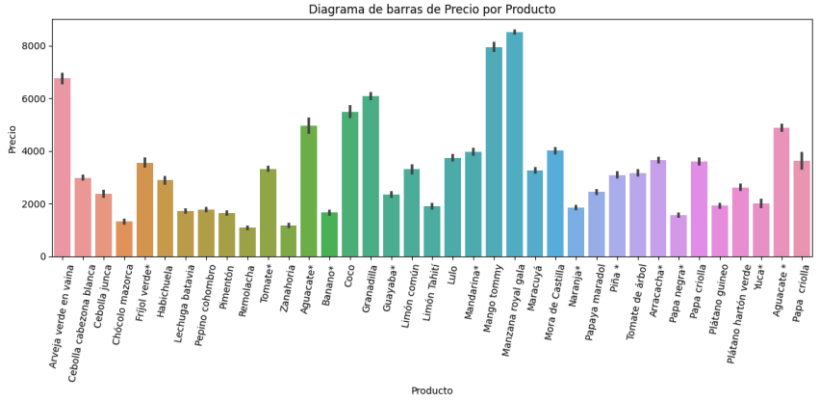
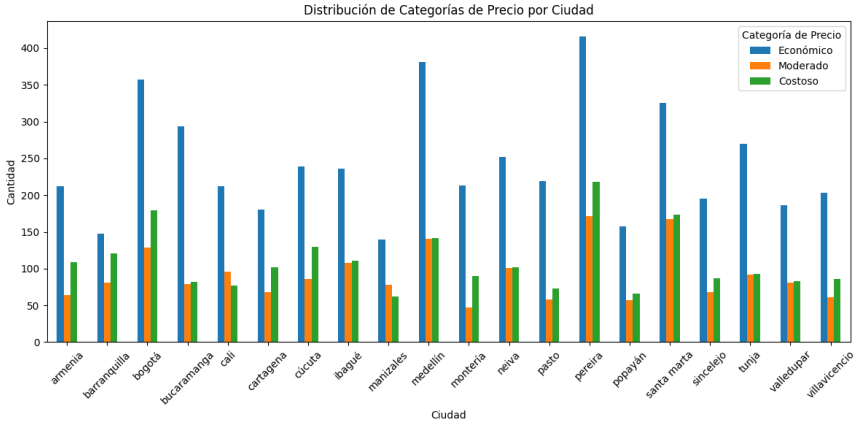
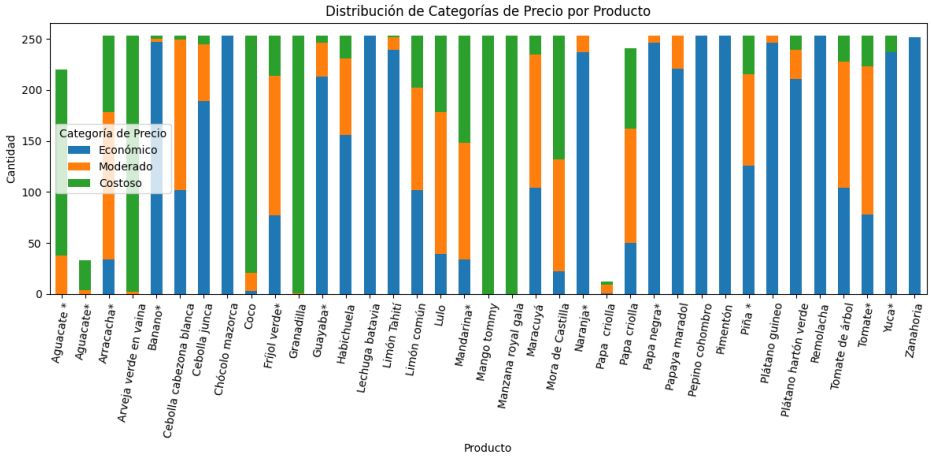


Diagrama de barras de Precio por Producto
Precio
Producto
0
2000
4000
6000
8000
Arveja verde en vaina
Cebolla cabezona blanca
Cebolla junca
Chócolo mazorca
Fríjol verde*
Habichuela
Lechuga batavia
Pepino cohombro
Pimentón
Remolacha
Tomate*
Zanahoria
Aguacate*
Banano*
Coco
Granadilla
Guayaba*
Limón común
Limón Tahití
Lulo
Mandarina*
Mango tommy
Manzana royal gala
Maracuyá
Mora de Castilla
Naranja*
Papaya maradol
Piña *
Tomate de árbol
Arracacha*
Papa negra*
Papa criolla
Plátano guineo
Plátano hartón verde
Yuca*
Aguacate *
Papa criolla


Distribución de Categorías de Precio por Ciudad
Cantidad
Ciudad
Categoría de Precio
Económico
Moderado
Costoso
0
50
100
150
200
250
300
350
400
armenia
barranquilla
bogotá
bucaramanga
cali
cartagena
cúcuta
ibagué
manizales
medellín
montería
neiva
pasto
pereira
popayán
santa marta
sincelejo
tunja
valledupar
villavicencio


Distribución de Categorías de Precio por Producto
Categoría de Precio
Económico
Moderado
Costoso
Cantidad
0
50
100
150
200
250
Aguacate *
Aguacate*
Arracacha*
Arveja verde en vaina
Banano*
Cebolla cabezona blanca
Cebolla junca
Chócolo mazorca
Coco
Fríjol verde*
Granadilla
Guayaba*
Habichuela
Lechuga batavia
Limón Tahití
Limón común
Lulo
Mandarina*
Mango tommy
Manzana royal gala
Maracuyá
Mora de Castilla
Naranja*
Papa criolla
Papa criolla
Papa negra*
Papaya maradol
Pepino cohombro
Pimentón
Piña *
Plátano guineo
Plátano hartón verde
Remolacha
Tomate de árbol
Tomate*
Yuca*
Zanahoria
Producto