Taller 3

1. **Redefinir variables**

   En esta sección modifiqué los tipos de datos de las variables, producto y ciudad como strings, precio y variabilidad como numéricos, fecha como datetime, y aunque latitud y longitud contiene datos numéricos, en el ejercicio los traté como string, pues son variables que numéricamente no les realizaría análisis.

   Además, por sentido de concordancia en formato, todos los nombres de las columnas los dejé en minúscula, Cambiando LATITUD y LONGITUD por latitud y longitud respectivamente. Esta fue una decisión más por estética, porque el desempeño de la base de datos no cambia por dicho formato, pero sí hizo más fácil de recordar al momento de mencionarlas en algún código.

2. **Categorizar variables**

   Para la primera categorización que realicé, fue la categoría del producto, frutas, verduras y hortalizas, y tubérculos, raíces y plátanos, logrando incluir a todos los productos en una categoría. Con esto se pueden realizar análisis por tipo de producto, pues es más acertado comparar los productos de esta manera.

   Además, para usar el componente geográfico que da la base de datos, la segunda categorización fue según la zona del país, a partir de la latitud, quedando norte, centro o sur; es decir, poder comparar los productos no solo por ciudad, sino por el conjunto de ciudades que conforman la zona.

3. **Datos faltantes**

   Para que el análisis de la base de datos sea más completo, no eliminé los datos faltantes para tener la mayor cantidad de información. Primero, los faltantes de latitud y longitud se imputaron con los valores correspondientes de las principales centrales de abasto de la respectiva ciudad. Los de variabilidad se imputaron con la media de la variable según el producto. Los de precio con la media del precio del producto según la zona.

4. **Datos duplicados**

   Como solo identifiqué 13 filas duplicadas, aproximadamente el 0.14%, un valor no relevante, por lo que se eliminaron todos los duplicados.

5. **Datos outliers**

   La identificación de estos datos fue con respecto a la categoría, estadísticamente con el método de z-score, gráficamente, presenté solo aquellos productos que presentaron outliers según la prueba mencionada. El tratamiento fue winsorizar con límites del 15%

6. **Normalizar o estandarizar variables**

   Con el fin de mantener la integridad de los datos, elegí estandarizar precio y variabilidad