**Taller 3**

A partir del archivo de [precios](https://raw.githubusercontent.com/ousuga/Proyecto1_IAN/main/precios_pdtos_agricolas.cvs) del Taller 2 realizar el preprocesamiento del conjunto de datos obtenido considerando:

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

Para el caso particular del conjunto de datos del taller 2, no hay necesidad de redefinir ninguna de las variables porque las que existen en el momento son las adecuadas para hacer el análisis pertinente. Estas proporcionan la información necesaria para cumplir con el objetivo del análisis. Sin embargo, para las variables de ‘precio’ y ‘variabilidad’ sí se hace un tratamiento a algunos registros porque aparecen como n.d. lo que indica que es información que no es numérica; entonces, se convierte todos los valores n.d. en valores NaN para después tratar esos valores de una forma más fácil como datos faltantes.

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

En la categorización, se crea una nueva columna de costo basada en el precio de los productos para determinar si en promedio dichos productos tienen un costo alto, medio o bajo, teniendo en cueta que dichos pueden variar por cada ciudad.

1. Identificación de datos faltantes, justificación y uso de la estrategia de tratamiento de datos faltantes

Para este apartado, se utilizaron métodos de detección de datos faltantes como **df.info()** y **df.isnull().any(),** pero personalmente considero que para este caso la función **df.isnull().sum()** es más precisa porque muestra el conteo de todos los datos faltantes porcada una de las variables. Con esta función se detectó que habían 2186, 2329, 805 y 805 datos faltantes en las variables de ‘precio’, ‘variabilidad’, ‘LATITUD’ y ‘LONGITUD’, respectivamente.

Para la estrategia de tratamiento de datos, se usó la imputación de datos faltantes a la columna de ‘precio’ y ‘variabilidad’ con el método de interpolación porque genera un valor numérico con que método lineal, lo que garantiza que no es un valor semejante a las medidas de tendencia central ya que para el contexto del dataframe no tendría sentido hacer una imputación de dichas medidas y tampoco se optó por eliminar los registros porque habría una pérdida de información importante. Finalmente, para las variables de ‘LATITUD’ y ‘LONGITUD’, los datos faltantes se reemplazaron por un valor numérico de longitud y latitud para cada ciudad, utilizando funciones como **.fillna**

1. Identificación de datos duplicados, justificación y uso de la estrategia de tratamiento de datos duplicados

Para la identificación de registros duplicados se usó la función **df.duplicated()**. Sin embargo, son 8855 registros y Python por defecto solo muestra los primeros 5 registros y los 5 últimos, lo que no da certeza de que en medio puedan haber o no datos duplicados. Entonces, para determinar si en su totalidad el df tiene duplicados, hice un ciclo implementando la función **.any** con el objetivo de buscar uno a uno entre los registros del df si había alguno repetido y el resultado fue que no halló ninguno. Por tanto, no hubo estrategia de tratamiento de datos debido a que no habían duplicados.

1. Identificación de datos outliers, justificación y uso de la estrategia de tratamiento de datos outliers

Para este punto, al aplicar la función df.isnull().sum() se puede apreciar que ya no hay datos faltantes de nungun tipo, lo que permite seguir con el siguiente paso que es determinar si hay datos outliers. Aquí se usó la función .**describe()** para identificar diferencias entre las medidas descriptivas. Lo que se pudo concluir al haber aplicado dicha función fue que la mediana está más cerca de los datos mínimos que de los máximos, lo que podría indicar que los datos máximos están más alejados del 50% de los datos y en ese sentido podría haber datos outliers, pero es importante hacer un análisis gráfico como un histograma.

Es importante resaltar que cuando se hizo el histograma completo de los precios, a pesar de que se pueden notar datos atípicos, no se pudo sacar ningún tipo de conclusión interesante puesto que se cuenta con el precio de todos los productos en su totalidad y todos son muy diversos, es decir, que no se deben relacionar justamente porque no tendría sentido comparar todos los precios de todos los productos de todas las ciudades de la base de datos. En ese sentido, lo más conveniente fue hacer un tipo de filtro por cada uno de los 37 productos para determinar si de forma individual presentaban datos outliers. Entonces lo que se hizo fue tomar un producto como la Mandarina para determinar su precio por cada ciudad y así tendría más sentido comparar y saber si existen outliers por ciudad. Una vez aplicado lo anterior, se procedió a aplicar las estrategias de tratamiento de datos atípicos y se llegó a la conclusión que la mejor estrategia es eliminarlos dado que al winzorizarlos no había ningún cambio y al imputarlos la mediana se desplazaba de forma considerable; en cambio al momento de eliminarlos, el boxplot forma una distribución más normal.

1. Justificación de la necesidad de normalizar o estandarizar variables

En el preprocesamiento de los datos es importante que las variables que se estén analizando estén en un rango específico según el requerimiento, por ejemplo, para hacer análisis predictivos. Sin embargo, para el contexto de la base de datos de los productos por ciudades no se tuvo como objetivo hacer dicho análisis, por lo tanto, no es necesario estandarizar, normalizar ni escalar los datos.

1. Mínimo 5 preguntas que quiera resolver a partir del filtrado de columnas o filas.

* ¿Cuál es la variabilidad de precios para un producto específico en diferentes ciudades?
* ¿Cuál es la ciudad con el precio promedio más alto para un producto en particular?
* ¿Cómo varían los precios de un producto a lo largo del tiempo?
* ¿Cuál es el producto más caro en promedio en todas las ciudades?
* ¿Cuál es la ciudad con la mayor variabilidad de precios en general?

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