**Taller 3**

**Por: Laura Betancourt**

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:

Link Github: <https://github.com/LauBetancourt31/Analitica/blob/main/Taller_3_Betancourt.ipynb>

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

Se hace necesario redefinir las variables ‘precio’, ‘variabilidad’, ‘LATITUD’ y ‘LONGITUD’, debido a que ‘precio’ y ‘variabilidad’ son numéricas y se encuentran definidas como ‘Object’, por lo que se redefinieron como ‘Float’ ya que más adelante con ellas se realizara análisis estadístico. Mientras que las variables de latitud y longitud se recategorizaron a ‘Object’ por ser denominadas descriptivas, esto se debe realizar para evitar futuros problemas con el procesamiento de los datos y que al realizar análisis estadístico sean tomadas en cuenta.

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

Se realizaron dos categorizaciones de los datos, una categorización intuitiva y mecánica de los datos, en una nueva variable llamada ‘clasificación’ se agregaron las categorías ‘Fruta’, ‘Verdura’ y ‘Tubérculo’. La otra categorización realizada fue la comparación del precio diario de los productos de cada ciudad en la nueva variable ‘Precio día’ con las categorías ‘Alto’ si el precio aumentaba más del 20% con respecto a la media del precio del producto, ‘Bajo’ disminución del 20% del precio con respecto a la media, y ‘Normal’ si se mantenía en el rango de 20% por encima y por debajo de la media. La necesidad de categorizar estas variables es facilitar el análisis de los datos al tratarlos para dar respuesta las preguntas que se planteen para el problema

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

Al inicio del código se realizo el cambio de datos que estaban dados como ‘n.d.’ a ‘NaN’ para un mejor tratamiento y reconocimiento de datos faltantes en las base de datos. Para identificar los datos faltantes primero se sumaron los datos nulos por cada variable, con el fin de identificar en que variables se presentaban faltantes y cuantos eran.

Interfaz de usuario gráfica, Texto

Descripción generada automáticamente

Como anteriormente se definió el tipo de datos de cada variable se realizo el tratamiento de los datos en dos partes. En la primera parte se trataron las variables ‘LATITUD’ y ‘LONGITUD’ que contaban con la misma cantidad de faltantes, por lo que se decidió identificar las ciudades a las que correspondían estos datos faltantes, los cuales se imputaron de la base de datos ‘DIVIPOLA\_Municipios.csv' del taller 2, extrayendo los valores de latitud y longitud de las ciudades correspondientes a estos datos, y reemplazando los NaN por los valores extraídos.

Tras eliminar los datos coincidentes, aún había valores nulos en la variable 'variabilidad' en casos donde sí existía un valor de 'precio'. Esto sugiere que el producto fue vendido, pero la información de variabilidad faltaba. Para abordar esta situación, optamos por imputar los datos faltantes de 'variabilidad' con el promedio de variabilidad correspondiente al mismo producto. Esta decisión se basó en la existencia de más datos de variabilidad del mismo producto en otras instancias, lo que permitió estimar un valor razonable para llenar los vacíos en la información.

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

Para identificar los datos duplicados se hizo uso del código, para visualizar las filas que tenían todos su valores repetidos:

Captura de pantalla de un celular

Descripción generada automáticamente

Se decide eliminar los valores duplicados que arroja el código pues se determina que los datos de la ciudad de Pereira se encuentran dos veces en la base de datos y el mantenerlos puede llegar a afectar la precisión de los análisis.

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

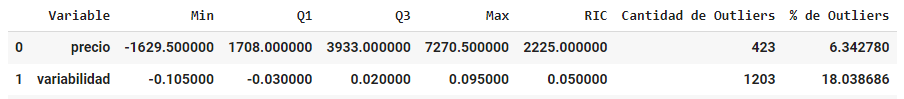
Se identificaron los outliers de cada producto con relación a las variables ‘precio’ y ‘variabilidad’ utilizando dos métodos. Primero el método visual graficando, por medio de boxplots, los outliers de cada producto para ambas variables, se obtuvo como resultado lo siguiente:

Gráfico

Descripción generada automáticamenteGráfico

Descripción generada automáticamente

El segundo método utilizado fue el Rango Intercuartílico (RIC), para conocer más información acerca de los outliers, y es por el RIC, por el que se identifica el porcentaje de estos datos atípicos en la base de datos para el posterior tratamiento.



Identificados los outliers se paso al tratamiento de estos y lo que se decidió fue winsorizarlos con un limite del 10% con el fin de corregir los valores que fuesen más extremos y tratando de reducir el impacto de los outliers, con una menor alteración de la distribución de los datos que pueden llegar a ser significativos en el análisis estadístico.

Gráfico de dispersión

Descripción generada automáticamente

Imagen que contiene Gráfico

Descripción generada automáticamente

1. Justificación de la necesidad de normalizar o estandarizar variables

Para las variables que se relacionaron y estudiaron no se encontró necesario la normalización o estandarización de los datos, ya que las variables ‘precio’ y ‘variabilidad’ si bien depende una de la otra, no están expresadas en las mismas unidades, y no se hace necesario para el análisis de la información. Otra razón es que se pierde la variabilidad de precio por lo que se hace más difícil el análisis del problema.

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

* ¿Cómo varía el precio del Limón Thaití de la ciudad de Medellín a Bogotá?
* ¿En qué ciudad se alcanza un precio más alto en las verduras?
* ¿Cómo es la variabilidad de la zanahoria en la ciudad de Cúcuta?
* ¿En qué ciudad es más cara la Yuca\*?
* ¿Cómo es la participación de verduras, frutas y tubérculos en las ciudades?

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

* ¿Cómo varía el precio del Limón Thaití de la ciudad de Medellín a Bogotá?

Gráfico, Gráfico de líneas

Descripción generada automáticamente

* ¿En qué ciudad se alcanza un precio más alto en las verduras?

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

* ¿Cómo es la variabilidad de la zanahoria en la ciudad de Cúcuta?

Gráfico, Gráfico de líneas

Descripción generada automáticamente

* ¿En qué ciudad es más cara la Yuca\*?

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

* ¿Cómo es la participación de verduras, frutas y tubérculos en las ciudades?

Gráfico, Gráfico circular

Descripción generada automáticamente

Deberá entregar un notebook y un documento con las respuesta a cada uno de los items 1 a 8. Estos documentos los debe subir a Ingenia y a su GitHub.