**Taller 3: introducción a la analítica de negocios**

**Estudiante:** Karen Marcela Pérez Castilla **C.C.** 1005605305

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**.

**R/** Sí se hizo necesario redefinir variables, debido a que “precio” y “variabilidad” están originalmente en formato *object,* y se observó que era más favorable el tenerlas en un formato numérico, esto con el fin de facilitar futuros procesos, como el de categorización teniendo en cuenta un intervalo o rango, por ejemplo. Y en general, se apreciaría más natural que su formato sea de números para poder se tratados como tal.

Entonces, para poder redefinirlas era necesario tener en cuenta algunos datos no registrados en estas dos columnas, los cuales se observaban como “n.d.”, estos se convirtieron en “NaN” y luego se procedió a cambiarles su formato a valor numérico con la función *pd.to\_numeric* de pandas.

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

**R/** Sí se categorizaron variables, en este caso “precio” y “variabilidad'”, debido a que los precios registrados pueden ser clasificados en Costoso, Asequible o Económicos dependiendo del punto de vista en que se analice (en este caso juicio propio y gracias a la utilización de límites del máximo y mínimo en cada columna se pudo hacer) y lo mismo para la variabilidad, ya sea Aumentó, Se Mantuvo, Disminuyó. Sin embargo, antes de pasar a realizar la categorización, se decidió imputar nulos y revisar duplicados previamente.

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

**R/** Se identificaron datos faltantes “NaN” en las columnas “precio”, “variabilidad”, “LATITUD” y “LONGITUD”, para el caso de las dos últimas columnas mencionas poseían la misma cantidad de nulos que era 805 y se descubrió que justamente pertenecían a la misma ciudad, Cúcuta, y que sin valores de latitud coincidían en que tampoco habían de longitud en las mimas filas. De esa forma, se procedió a imputarlos al reemplazar justamente esos datos numéricos por los reales consultados pertenecientes a dicha ciudad, es decir, latitud= 7.89391 y longitud= - 72.50782.

Luego, respecto al “precio” y “variabilidad” de los productos se decidió calcular la media por cada categoría de producto (por ejemplo, la media de precios de la arveja verde en vaina, yuca, coco, etc.) y de esa manera, reemplazarla dentro de esa misma categoría, pero en los valores faltantes. La justificación de lo anterior radica en que se observó que eliminar esas filas era improcedente, ya que era una cantidad de nulos potenciales para la estructura de la base de datos (más o menos ¼ de los datos de esta) y que al final son precios de las principales ciudades productoras a nivel nacional, por lo tanto, un promedio sería algo acertado entre todos los valores registrados por producto.

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

**R/** En algunas columnas como “producto”, “ciudad”, “fecha”, “LATITUD” y “LONGITUD” es justificable la cantidad de duplicados y esto se debe a que son los mismos productos, pero en distintas ciudades y fechas y así con todos, es decir, la forma en la que se registraron u organizaron los datos evidencia que la repetición de estos sea algo normal.

Aun así, para la variable “precio”, se notó en cierta proporción datos duplicados debido al paso previo, donde los datos faltantes fueron reemplazados por la media, la cual se repite tantas veces haya datos faltantes en cada categoría de productos. Lo anterior hizo que se aumentaran los duplicados respecto a la base de datos inicial (en la que se analizó este factor sin tener en cuenta los “n.d.”). Por lo tanto, se pasó de tener 5869 a 7126 duplicados.

Adicionalmente, al observar filas duplicadas (esto para detallar mejor si sí existía un problema de repetición estricta de registros completos) se obtuvo que se repiten datos como su variabilidad, precio y así con el resto, exceptuando la fecha, esto para la ciudad de Pereira y en ciertos productos como chócolo mazorca, zanahoria, entre otros. En este caso no es válido eliminarlas debido a que la fecha sí varía para esas filas duplicadas, a pesar de que el precio o variabilidad no, es decir, ciertos datos por columnas son iguales, pero el registro en sí es único (fila completa)

Al final, los duplicados encontrados no se eliminaron o imputaron, ya que para poder hacer eso se debía repetir estrictamente la misma información entre cada fila y no fue así, por lo tanto, puede que el precio sea igual en muchas filas, pero el resto de las características asociadas a este (como ciudad, producto, variabilidad, etc.), pueden variar sin problema. Y así para el resto de los casos.

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

R/ Para la variable “precio” se halló que el percentil 25, es decir, el 25% de los precios están por debajo de 1733.53, y está muy alejado del valor mínimo (339). Además, el percentil 75, es decir, el 75% de los precios están por debajo de 3964.31, y está muy alejado del valor máximo (10000). Por otro lado, en el histograma se observan muchos datos alejados del valor del percentil 75, lo cual confirma que sí puede haber existencia de outliers y el boxplot muestra datos igualmente fuera del 3RIC, pero no muy alejados de este. Y al aplicar Rango intercuartílico con k =1.5 a esta variable, se obtuvo 590 valores como puntos atípicos. Sin embargo, a partir de lo anterior no se podría determinar correctamente que sean outliers o no, debido a que estamos tratando con diferentes productos o categorías, por lo tanto, es lógico que existan valores atípicos o tan alejados entre sí, lo mejor en este caso es analizar cada producto o categoría.

Una vez mejor hecho el análisis descriptivo por categorías, se encontró que algunos productos sí presentaban distancias amplias entre sus valores límites y percentiles, por ejemplo, el limón común, donde su percentil 75 era 3705.0 y su valor máximo era de 7593.0, es decir, una diferencia de 3888.0, casi del 50 %. Esto sumado que, al realizar el histograma, efectivamente había ciertos valores muy alejados del resto. Para esta situación particular, no se utilizó ninguna estrategia para su tratamiento, debido a que estas variaciones en ocasiones tan amplias en los precios de ciertos productos dependen de razones a nivel macroeconómico, demanda, oferta e inflación, además, causas climatológicas de cada ciudad en particular, entre otras.

En cuanto a la columna de “variabilidad”, se observó nuevamente distancias amplias entre los percentiles límites y los extremos máximo y mínimo. Y efectivamente, el histograma al hacerlo muestra por ambos algunos datos muy alejados al resto, y lo mismo sucede para el boxplot, sobre todo fuera del 3RIC.Y al aplicar Rango intercuartílico con k =1.5 a esta variable, se obtuvo 1458 valores como posibles puntos atípicos, muchos más que para “precio”. Aun así, se tiene la misma justificación que para el caso anterior, del porqué no se debería hacer un tratamiento de estos.

1. **Justificación de la necesidad de normalizar o estandarizar variables.**

**R/** Para este caso se presentó la necesidad de estandarizar las variables “precio” y “variabilidad”, teniendo en cuenta que su rango de variación es distinto y que, de esta forma, podría resultar más sencillo la comparación entre ambas variables (en cómo la “variabilidad” se ve afectada por el “precio”, por ejemplo) y también en miras de un posible modelo predictivo a futuro.

Para ello, se usó la forma de estandarización *stats.zccore* de *scipy*, y así obtener un rango más homogéneo entre los datos.

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

**R/**

* ¿Cuáles son las ciudades con aguacates en una categoría de precio “Asequible” durante todo el mes de agosto?
* ¿Cuáles son los productos de la ciudad de Medellín con precios económicos y que tengan variabilidad que refleje su disminución?
* ¿Cuál es el precio más bajo y el más alto por ciudad?
* ¿Cuáles son los productos con los precios más económicos y los más caros?
* ¿Cuáles son los productos de la ciudad de Bucaramanga que tuvieron un precio menor o igual a 3000 en la fecha del 29 de agosto de 2023?

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

**R/**

* Gráfico de densidad, este al igual que el histograma permite determinar puntos outliers y observar la distribución de los datos.
* Gráfico de dispersión para Precio vs Variabilidad, mostrando cómo es la relación entre ambas variables y dónde se encuentra la mayor concentración de datos. Acá se logró identificar que entre mayor es el valor del precio de un producto, menor es la variabilidad o más bien, tiende a mantenerse.
* Diagrama de violín.
* Gráfico de dispersión con línea de regresión, siendo posible visualizar el ajuste a los datos gracias a la línea que se observa entre estos.
* *Q-Q plot,* permite observar la normalidad de los datos, al desviarse o no de la línea de igualdad.