Introducción a la Analítica de Negocios
Taller - Seguimiento 3

Santiago Restrepo Olarte
Estudiante de Ingeniería Industrial
Universidad de Antioquia, Colombia

A partir del archivo de precios del Taller 2 realizar el preprocesamiento del conjunto de datos.

Descripción general del dataframe

Inicialmente el dataframe posee 9240 filas y 7 columnas, donde las filas están ordenadas por producto (alfabéticamente) y fecha (ver *Ilustración 1*). Además, en primera instancia, el dataframe posee algunas observaciones repetidas y datos faltantes en algunas columnas.

	producto	ciudad	precio	variabilidad	fecha	LATITUD	LONGITUD
0	Arveja verde en vaina	armenia	7200	0.1	01ago2023	6.163684	-75.809955
1	Arveja verde en vaina	armenia	7200	0.1	01ago2023	4.499501	-75.724900
2	Arveja verde en vaina	bogotá	7925	0.11	01ago2023	NaN	NaN
3	Arveja verde en vaina	bucaramanga	6860	0.08	01ago2023	7.155834	-73.111570
4	Arveja verde en vaina	cali	7733	-0.03	01ago2023	3.399044	-76.576493

Ilustración 1. Visualización cabeza del dataframe

1. Redefinición de variables

Gracias al $.info()$ del dataframe es posible observar (ver *Ilustración 2*) que los tipos de datos que existen son 'object' y 'float'. Sin embargo, para el caso del precio y la variabilidad es pertinente realizar una redefinición del tipo de dato, dado que más adelante se realiza la manipulación de dichos datos y no es posible realizarlo si se mantienen en formato 'object', por ende, se procede a redefinir 'precio' como tipo 'int' y 'variabilidad como tipo 'float'.

```
 0   producto      9240 non-null   object
 1   ciudad        9240 non-null   object
 2   precio        9240 non-null   object
 3   variabilidad  9240 non-null   object
 4   fecha         9240 non-null   object
 5   LATITUD       7105 non-null   float64
 6   LONGITUD      7105 non-null   float64
dtypes: float64(2), object(5)
```

$.info()$

Ilustración 2. Tipos de datos del dataframe

2. Categorización de variables

En este caso, no se encuentra la necesidad de realizar categorización de variables para el análisis que se realiza, dado que solo se hace tratamiento de datos faltantes y duplicados acompañado de análisis exploratorio de los datos. En caso de que los datos fueran a ser utilizados para calibrar un modelo predicción o realizar análisis mucho más detallados si se pudiesen realizar categorizaciones, por ejemplo, por tipo de producto (frutas, verduras u hortalizas) o por rango de precios (bajo, medio, alto).

3. Datos faltantes

Dentro del dataframe se evidencian los datos faltantes como 'NaN', sin embargo, parece que algunos datos faltantes fueron registrados de forma manual como 'n.d.', por ende, es necesario convertir todos los datos faltantes en 'NaN'. Para ello se hace uso de un $.replace('n.d.', np.nan)$ que reemplaza los faltantes 'n.d.' por datos faltantes de numpy ('NaN').

Una vez realizado el reemplazo, se identifican datos faltantes en las columnas 'precio', 'variabilidad', 'LATITUD' y 'LONGITUD' (ver *Ilustración 3*).

```
producto        False
ciudad          False
precio           True
variabilidad     True
fecha           False
LATITUD          True
LONGITUD         True
dtype: bool
```

$.isnull().any()$

```
producto          0
ciudad            0
precio         2236
variabilidad   2384
fecha             0
LATITUD        2135
LONGITUD       2135
dtype: int64
```

$.isnull().sum()$

Ilustración 3. Datos faltantes del dataframe

Para la imputación de duchos datos faltantes se procede de la siguiente forma:

- Para la columna 'LATITUD' Y 'LONGITUD', gracias a un filtrado se identificó que las ciudades que no poseían el dato de latitud y longitud eran Bogotá, Cúcuta, Santa Marta y Cartagena. Dichos datos se imputaron con el dato de longitud y latitud correspondiente por medio de un $.at[\,]$ para modificar el valor especifico de esas columnas a través de un ciclo.
- En cuanto a la columna 'precio' se decide utilizar un $.dropna()$ para eliminar todas las observaciones que tengan dato faltante en dicha columna, dado que si dicho dato está vacío quiere decir que ese producto no obtuvo registro en ese día especifico.
- Para el caso de la columna 'variabilidad' se identificaron inicialmente 2384 datos faltantes, sin embargo, al eliminar las observaciones con datos faltantes en la columna 'precio' observamos que quedan solo 148 datos faltantes y se decide imputarlos por medio de un $.fillna()$ con el valor de 0. Además, cabe mencionar que no sería adecuado eliminar dichas observaciones dado que sí poseen dato en su precio, por ende, sí se obtuvo registro para ese producto en ese día.

  Después de realizar el tratamiento de estos datos se obtiene un dataframe con 0 datos faltantes (ver *Ilustración 4*).

```
producto        0
ciudad          0
precio          0
variabilidad    0
fecha           0
LATITUD         0
LONGITUD        0
dtype: int64
```

$.isnull().sum()$

Ilustración 4. Datos faltantes del dataframe después de tratamiento

4. Datos duplicados

Inicialmente, gracias a la exploración a través de un $.duplicated()$, se identificaron 13 observaciones duplicadas (ver *Ilustración 5*).

	producto	ciudad	precio	variabilidad	fecha	LATITUD	LONGITUD
897	Chócolo mazorca	pereira	n.d.	n.d.	03ago2023	4.803663	-75.795791
1017	Zanahoria	pereira	n.d.	n.d.	03ago2023	4.803663	-75.795791
1332	Plátano guineo	pereira	n.d.	n.d.	03ago2023	4.803663	-75.795791
2857	Chócolo mazorca	pereira	n.d.	n.d.	10ago2023	4.803663	-75.795791
2977	Zanahoria	pereira	n.d.	n.d.	10ago2023	4.803663	-75.795791
3022	Coco	pereira	n.d.	n.d.	10ago2023	4.803663	-75.795791
3292	Plátano guineo	pereira	n.d.	n.d.	10ago2023	4.803663	-75.795791
5307	Chócolo mazorca	pereira	n.d.	n.d.	17ago2023	4.803663	-75.795791
5427	Zanahoria	pereira	n.d.	n.d.	17ago2023	4.803663	-75.795791
5742	Plátano guineo	pereira	n.d.	n.d.	17ago2023	4.803663	-75.795791
7267	Chócolo mazorca	pereira	n.d.	n.d.	24ago2023	4.803663	-75.795791
7387	Zanahoria	pereira	n.d.	n.d.	24ago2023	4.803663	-75.795791
7702	Plátano guineo	pereira	n.d.	n.d.	24ago2023	4.803663	-75.795791

Ilustración 5. Datos duplicados en el dataframe inicial

Sin embargo, al realizar la eliminación de las filas que poseen el precio y la variabilidad del precio como dato faltante se eliminan 12 de ellas, y la única observación duplicada que queda es eliminada.

5. Datos outliers

Haciendo el análisis de outliers (una vez preprocesado el dataframe) para las variables numéricas 'precio' y 'variabilidad' a través de un $.describe()$ (ver *Ilustración 6*) nos damos cuenta de que las dos variables poseen su media alejada de sus valores máximos.

```
count     7003.000000          count    7003.000000
mean      3155.019849          mean        0.002519
std       1932.068856          std         0.092890
min        339.000000          min        -0.460000
25%       1700.000000          25%        -0.030000
50%       2775.000000          50%         0.000000
75%       3942.000000          75%         0.020000
max      10000.000000          max         1.060000
Name: precio, dtype: float64   Name: variabilidad, dtype: float64
```

$.describe()$

Ilustración 6. Descripción de variables 'precio' y 'variabilidad'

Sin embargo, no se decide realizar tratamiento dado que, a pesar de que los valores máximos no son cercanos a la media, tampoco son datos atípicos derivados de errores de codificación o cálculo. Además, se presentan gráficos (ver *Ilustración 7 y* 8) que permiten apreciar la distribución de los datos 'precio' y 'variabilidad'.

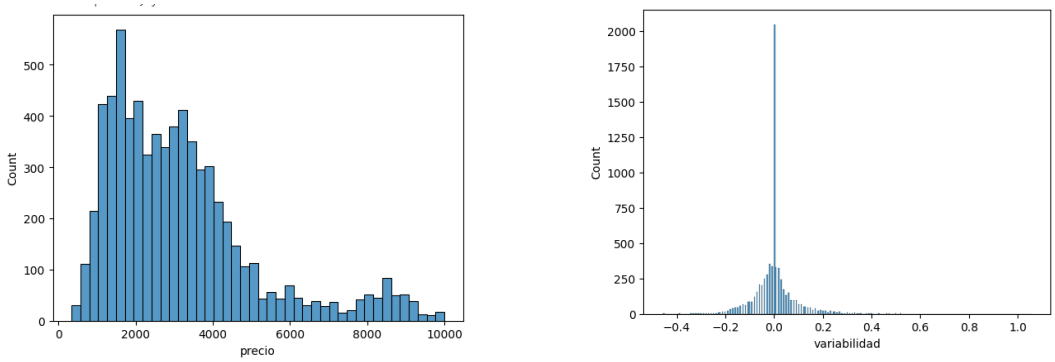

Ilustración 7. Histogramas de variables 'precio' y 'variabilidad'

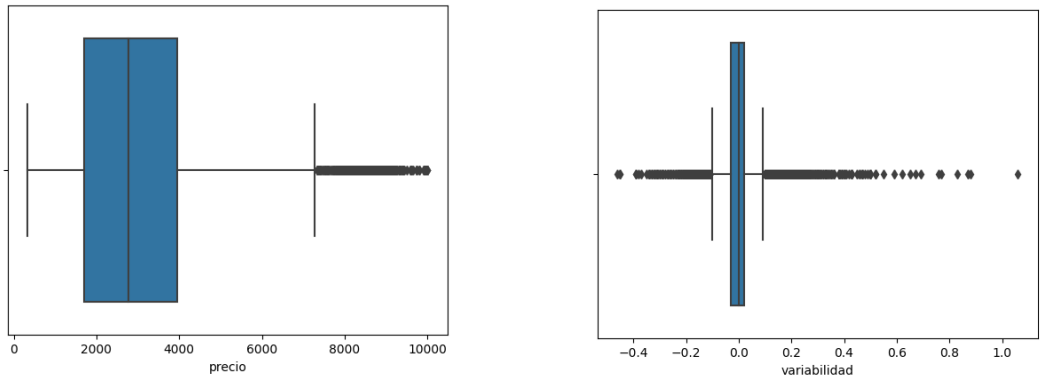

Ilustración 8. Boxplots de variables 'precio' y 'variabilidad'

6. Normalización o estandarización de datos

Dado que, en este momento el interés no se basa en calibrar o entrenar modelos predictivos ni realizar comparaciones directas entre las variables, entonces no se realiza normalización o estandarización de las mismas.

7. Preguntas a partir del filtrado

Las preguntas formuladas y solucionadas a partir del filtrado de columnas se observan en la *Tabla 1*.

	Pregunta	Respuesta
1	¿Cuáles son los productos que alcanzaron un precio superior a $9.000 en algún momento de agosto en la ciudad de Bogotá y cuál fue su precio máximo?	**Manzana royal gala** con un precio máximo de $9397 **Mango tommy** con un precio máximo de $9010
2	¿Cuáles son los productos que alcanzaron un precio inferior a $500 en algún momento de agosto en la ciudad de Medellín y cuál fue su precio promedio?	**Zanahoria** Precio promedio: $396

3	¿En qué ciudades del país el aguacate alcanzó un precio inferior a $3.500 en algún momento de agosto?	**Armenia**
4	¿Qué productos obtuvieron la variabilidad de precio más alta durante el mes de agosto y en qué ciudades?	**Pimentón:** con una variabilidad de 1.06 en la ciudad de Neiva **Cebolla cabezona blanca:** con una variabilidad de 0.88 en la ciudad de Pereira **Cebolla cabezona blanca:** con una variabilidad de 0.83 en la ciudad de Popayán **Habichuela:** con una variabilidad de 0.87 en la ciudad de Cali
5	¿El precio de la papa fue inferior en la ciudad de Pasto (territorio papero) comparado con la ciudad de Cartagena (territorio no papero) durante el mes de agosto?	**Sí;** obteniendo un precio promedio de los diferentes tipos de papa: **Pasto**: $2000 **Cartagena**: $3416

Tabla 1. Preguntas formuladas y solucionadas a partir del filtrado de datos

Cabe mencionar que las preguntas formuladas se pueden realizar para diferentes productos, ciudades y valores, pero en este caso se realizan de manera que lleguen a preguntas y conclusiones específicas.

8. Gráficos de interés

i. En primer lugar, se podría realizar un **'barplot'** que relaciones variables categóricas con variables numéricas. En este caso podríamos usar como variable categórica un producto en específico (eje X) y el precio promedio durante el mes de ese producto (eje Y).

ii. Se puede realizar un **'line chart'** como una serie de tiempo sobre el precio de un producto especifico o la variabilidad de un producto a través del tiempo. Donde el eje Y sería el precio o variabilidad del producto y el eje X es el tiempo en días (en este caso el mes de agosto). Así mismo, si se realiza la normalización o estandarización de los datos numéricos se podría realizar un análisis comparativo entre ellos mismos.

iii. Teniendo en cuenta que contamos con datos como la ciudad y sus coordenadas geoespaciales se podría realizar un **'bubble map'** para determinar geográficamente el estado de una variable en específico. Por ejemplo, se podrían ubicar las burbujas en las ciudades que contiene el dataset y dicha burbuja será más grande o tendrá otro color donde el precio de algún producto en particular sea mayor.

iv. Para relacionar el precio y la variabilidad de un producto se puede hacer uso de un **'scatterplot',** donde precio puede ser el eje X y variabilidad puede ser el eje Y.

v. Se podría hacer uso de un **'density plot'** para identificar la distribución de las variables númericas como precio y variabailidad, y así identificar tendencias o patrones del precio o variabilidad de un producto en específico.

Gracias al preprocesamiento de datos realizado se obtiene un dataframe con 7003 filas y 7 columnas, sin datos faltantes ni datos duplicados.