¿Caro o Barato? Análisis de Precios de Almacén en un Contexto Inflacionario

Diplomatura en Ciencia de Datos, Aprendizaje Automático y sus aplicaciones

Práctico 2 - Exploración y Curación

Dario Yvanoff

Leandro Sepulveda

Gonzalo Torres

Respuestas a consignas disparadoras.

1) Todas las variables tienen el tipo apropiado? Analizar las features con tipo Objeto. Qué decisiones tomarán al respecto?

*Las variables en general tienen el tipo apropiado. Aquellas que son esencialmente un valor númerico con el que se precisa hacer un cálculo son numéricas (float64), mientras que el resto son objetos (strings)*

2) Todos los producto\_id en el dataset de precios se encuentran en la entidad de productos? Tomamos decisiones al respecto?

*No, se decide eliminar las observaciones de precios sin producto.*

3)Todos los sucursal\_id en el dataset de precios se encuentran en la entidad de sucursales? Tomamos decisiones al respecto?

*No, se definen eliminar las observaciones sin datos de sucursal.*

4) Cuántos precios reporta cada sucursal? Todas las sucursales reportan precios en todas las fechas?

*No todas las sucursales reportan los mismos productos en todas las fechas.*

Listado de acciones realizadas y comentarios.

*Los conjuntos de datos (precios, sucursales y productos) sobre los que se realiza alguna acción se detallan al principio en “negrita”. Cuando se menciona a dos o más conjuntos Productos/Precios, por ejemplo, se refiere al conjunto adecuadamente combinado.*

**Precios**: Se eliminaron 2273 id de productos únicos que están sin precios en el conjunto de datos . Representan un total de 3.15% del total de los 72038 productos en el dataset de productos.

**Sucursales**: Existen 5 sucursales que presentaron precios en Nan. En las sucursales en donde aparecen productos con precios en NaN no aparecen en la entidad precios. Entendemos que hay un error en el dataset para los datos en esas sucursales (todos son NaN). Dado el porcentaje de NaN en precios y el hecho de que pertenencen a sucursales en donde no hay productos con precios informados consideramos eliminar los NaN.

**Precios**: No consideramos adecuado calcular o agregar los valores faltantes en precio, porque representaban un porcentaje muy pequeño del dataset.

**Precios**: Existe 7 productos informados en el conjunto de precios que no se encuentran en el conjunto de datos de productos. Dado que no existe forma de identificar nombre y marca de dichos productos se los elimina del conjunto de precios.

**Precios**: Existen 17998 filas en el conjunto de datos de precios cuya sucursal\_id no existe en sucursales. Una alternativa podría ser incorporar las sucursales inexistentes tomando el nombre a partir del comercioId y banderaId. Sin embargo, cabe aclarar que si utilizamos este método no vamos a disponer de la ubicación de la sucursal y no podríamos inferir la ciudad de la sucursal pues hay cadenas que tienen presencia en distintas ciudades y regiones del país. Considerando que el objetivo final de este proyecto es encontrar una forma de identificar si un producto está caro o barato, no en una sucursal puntual, sino dada una región o ciudad debemos proceder a eliminar las filas que no posean información de sucursales.

**Productos**: Se eliminaron 2 productos que no tienen presentación ni nombre

**Productos/Precios**: Se genera una nueva unidad de medida “pack”, para productos que son una combinación de 2 o más.

**Productos/Precios:** Se eliminan 124.246 productos del dataset por ser outliers aplicando IQR y hay 27 outliers adicionales si aplicamos Z score.

*Nota: bajo el metodo de eliminacion de outliers utilizando z-score se eliminan 16003 filas outliers de productos mientras que utilizando IQR se eliminan 124219 filas. Existian dentro de las 16003 filas de z score 27 filas que no estaban dentro de las 124219 filas del metodo IRQ.*

**Productos/Precios:/Sucursales**: Se incorpora metadata al conjunto de datos.

**Productos**: Se extrajo del nombre la marca del producto del nombre (ejemplo: *Radicheta atada La Anónima* es remplazado por *Radicheta Atada*)

**Productos**: Obtenemos aquellos productos cuya unidad de medida en el nombre no coincide con la unidad de medida de la presentación y la unidad de medida del nombre es una unidad de presentación válida. Existen 1207 casos donde se realizaron estas modificaciones.

Adicionalmente se actualizaron las um\_depurada y las cantidad\_depurada en función de las siguientes reglas que definimos:

1- Cuando hay 1kg en nombre y 1 unidad en presentacion entonces dejamos 1 kg en la presentacion

2- Cuando hay gr en mombre y unidad en presentacion entonces dejamos la cantidad de gramos por cantidad de unidades en presentacion

3- Cuando hay mililitro en nombre y litro en presentacion entonces dejamos ml en presentacion.

Cabe destacar que existe un vino cuyo producto\_id=7791540049415 que tiene 750 lt en la presentacion en vez de 750 ml

En función de las reglas antes mencionadas se actualizaron 72036 filas con las um\_depurada y cantidad\_depurada de acuerdo a las reglas antes especificadas

**Productos**: Se computa el factor de conversión a UM homogénea llevando cc y ml a litros, cm a mt y gr a kg.

**Productos**: Se realiza la extracción del nombre de dígitos numéricos, caracteres especiales y las preposiciones del castellano.

**Productos**: Se tokenizaron las palabras componente de los nombres y se calculó la frecuencia de aparición de las mismas. Se crearon variables dummies que identifican si el nombre de un producto está dentro de las más frecuentes (palabras frecuentes). La definición para la creación de las variables dummies fue tomar el conjunto de palabras mas frecuentes que represente el 20% del total de palabras. Esto representa aproximadamente 60 palabras que contienen 63.000 coincidencias (1000 coincidencias por palabra) quedando fuera 14.000 palabras con un promedio de conciencia por palabra de 22.

**Productos**: Se crea una variable dummy si el nombre del producto contiene palabras no definidas como frecuentes.

**Productos**: Se tokenizan las 30 marcas más frecuentes y se crea una variable de clasificación que replica la marca frecuente si el producto tiene esa característica o la clasifica como *otra* si no la presenta.

**Productos:** Se crean las variables dummies para las diferentes marcas frecuentes.

**Sucursales**: Se controla que no haya provincias faltantes y se generan dummies por provincia.

**Precio/Producto/Sucursal**: Se crea un nuevo dataset que integra los tres anteriores.

**Precio/Producto/Sucursal**: como se menciono anteriormente se hizo el análisis de outliers utilizando IRQ y z score.

**Precio/Producto/Sucursal:** Se crea una nueva variable bien numérico que será el numerario para calcular los precios relativos. Se controla que la variable se encuentre presente en todas las provincias y en todas las fechas. Producto elegido: 7793100111563.

**Precio/Producto/Sucursal:** Se crea una nueva variable que es el precio promedio (por unidad homogénea) del bien numérico, para cada provincia y fecha. Se calculan los precios relativos de todos los productos por provincia y por fecha utilizando la variable mencionada.

**Precio/Producto/Sucursal:** Se realiza el estudio de la distribución de precios por cuartiles y no se detectan mayores diferencias entre precios relativos y precios absolutos. Tampoco se detectan mayores diferencias en la detección de outliers.