Informe experimento en Kettle de Pentaho

Autor: Carlos Pérez Manzano

Introducción

Los datos tomados para el experimento se pueden encontrar en Kaggle. Contienen información sobre las ventas de verduras de un supermercado. Las fuentes de datos se componen de 4 archivos .csv que contienen los siguientes datos:

- **itemCategory.csv:** relaciones entre los productos y la categoría a la que pertenecen.
- **sales.csv:** ventas diarias de los productos desde el 01/07/2020 hasta el 30/06/2023.
- **wholesalePrice.csv:** precio al por mayor diario de los productos.
- **lossRateItem.csv:** tasa media de pérdida del artículo (en %). Este indicador mide el porcentaje del stock de un producto que se pierde durante un periodo determinado. Estas pérdidas pueden deberse debido a robos, productos caducados, pérdidas operativas, etc.

Nota: al realizar la lectura de los datos se han cambiado los nombres de los campos para evitar problemas posteriores debido a la presencia de espacios.

Análisis de la calidad de los datos

En primer lugar, realizamos transformaciones básicas para verificar la calidad de los datos originales.

- Verificación de existencia de valores nulos.
- Verificación de duplicidades.

Hacemos uso de los *Dummy* steps y comprobamos que no existen valores nulos ni duplicidades. Aparentemente los datos están limpios, por lo que no se ve necesario proseguir más en estas labores de análisis previo.

Creación de columnas

Mediante el *Modified JavaScript value* creamos las columnas Year-Month-Day para los flujos de DailySales y WholesalePrice, que van a ser útiles para estudios posteriores. También, para el flujo de DailySales, añadimos el tramo horario en el que se producen las ventas y el importe total de estas.

Análisis de Ventas del Supermercado

Se han calculado algunos indicadores típicos sobre las ventas totales del supermercado. El objetivo de estos KPIs es valorar de manera general el rendimiento del negocio. Estos son, número de ventas con descuentos y según horario; beneficio neto anual, mensual (en media) y por categoría. Para ello, hemos usado los steps de Sort rows, Group by y Merge join.

Destacamos los resultados de las ventas según horario. Claramente, el mayor número de ventas se produce por la mañana y conforme avanza el día las vetnas suelen ir decreciendo. Esto es algo a tener bastante en cuenta para la logística del personal del supermercado.

Posteriormente, se han realizado transformaciones para obtener dos indicadores más avanzados sobre las ventas según el producto. Estos van a permitir tomar decisiones basadas en datos en áreas clave: inventario, marketing, operaciones, etc. Estos KPIs escogidos son: productos con mayor número de ventas mensuales y productos con mayor beneficio neto mensual.

- **Productos con mayores ventas mensuales:** vamos a establecer los productos con mayores ventas como los productos que más veces aparezcan entre los 5 productos más vendidos por mes. Para ello, en primer lugar, calculamos el número de ventas por mes y producto. Posteriormente, con ayuda del step *add value fields changing sequence* tomamos los 5 productos con mayores ventas por cada mes. Por último, agrupamos por producto para obtener el número de apariciones en el ranking por mes.
  Destacan el brocoli y el pimiento verdeb Wuhu como los productos con mayores ventas.
  Nota: en este caso hemos considerado que los productos devueltos no constan como ventas. En los indicadores anteriores y en el que sigue no se va a considerar por simplicidad.

- **Productos con mayor beneficio neto total:** añadimos un campo para conocer el beneficio neto de cada venta. Este es (PrecioVenta - PrecioCompra) × KgVendidos × (1 - LossRate / 100).
  Para añadir este campo, hemos tenido que hacer el join de tres flujos de datos. En primer lugar, hacemos el join de DailySales con WholesalePrices para tomar el precio de compra del producto en el momento de la venta. Posteriormente, se realiza el join con LossRateItem para tomar la tasa media de pérdida del artículo vendido.

Una vez calculado el campo, solo falta agrupar por producto y realizar la suma de los beneficios por venta.
Destaca la mezcla de pimientos como el producto más rentable seguida de la raíz de loto Honghu en polvo.