**Fundamentos de Ciencia de Datos**

**Taller 1:** Adquisición, procesamiento y visualización de datos

**Dataset:** Online Retail Data Set, link: <https://archive.ics.uci.edu/dataset/352/online+retail>

**Parte1: Adquisición y limpieza de datos**

Primero se revisó la información que contenía el archivo OnlineRetail.xlsx con el programa Excel, para identificar características especiales que los datos pueden tener. Existen 541909 registros de compras en la tienda en línea y las compras realizadas van entre el 1 de diciembre de 2010 al 9 de diciembre de 2011 (visualizados en Python con Pandas).

Para realizar la limpieza y después de la revisión de los datos del mismo, se tomó las siguientes decisiones considerando que el caso expuesto es una tienda en línea para compra de productos:

* Se eliminaron los registros duplicados dejando solo un valor único.
* Los campos **CustomerID**, **UnitPrice** y **Quantity** contienen datos nulos, valores negativos o ceros, lo que hace imposible procesarlos como una transacción real. Por ello, se decidió eliminar dichos registros. Sin embargo, para no descartar información sin justificación, se verificó la columna **InvoiceID**, donde se confirmó que estos datos correspondían a registros cancelados.
* El criterio más importante para esta limpieza fue considerar los percentiles de Quantity y UnitPrice ya que es probable que por error se hayan puesto valores exagerados en dichos campos, se aplicó el 1% del percentil de los mismos para eliminar la información.
* Además, se identificó que existen productos de diferentes modelos o subproductos y que el StockCode excedía del tamaño permitido según el caso, por lo que se volvió a asignar códigos de 5 dígitos entre los intervalos de 00001 hasta 10000 (no existen productos en este rango).
* Posteriormente, se aplicó cambios más de forma que de fondo, pero no menos importantes, como: homologar la data de UnitPrice y Description a partir del StockCode, los datos de la columna InvoiceDate se homologaron al formato datetime, se cambió a mayúsculas los campos Description y Country, etc, reemplazo de campo vacío en CustomerID por 99999, C2 por 99988, DOT por 99977, etc.
* Se crearon las columnas **TotalValue** para almacenar el precio total del registro, donde el valor calculado es cantidad \* precio y la columna **InvoiceDay** que contiene la fecha en formato aaaa-mm-dd para poder agrupar posteriormente y mostrar en las gráficas.

**Parte 2: EDA**

**Realizar un análisis y una descripción (incluir como archivo de texto) de que información contiene el set de datos. ¿Qué puede decir o cómo puedes describir este set de datos?**

Después de la limpieza aplicada con los criterios antes mencionados al DataSet **Online Retail**, el número de registros final es: 391551. Esta tienda en línea vende productos principalmente para regalos y está ubicada en Reino Unido. De la lectura realizada se puede describir lo siguiente:

|  |  |  |  |
| --- | --- | --- | --- |
| **Columnas** | **Tipo de dato al cargar** | **Tamaño requerido** | **Descripción** |
| InvoiceNo | Object | 5 caracteres | Número único de la factura, si empieza con la letra C significa que fue cancelada. |
| StockID | Object | 5 dígitos | Código identificador único de cada producto. |
| Description | Object | N/A | Nombre del producto. |
| Quantity | int64 | 5 dígitos | Cantidad comprada. |
| InvoiceDate | datetime64[ns] | N/A | Fecha y hora de la compra. |
| UnitPrice | float64 | # enteros, 2 decimales | Precio del producto (en libras). |
| CustomerID | float64 | 5 dígitos | Cliente. |
| Country | object | N/A | País donde se realizó la compra. |
| InvoiceDay | datetime64[ns] | N/A | Fecha de la compra. |
| TotalValue | float64 | # enteros, 2 decimales | Cantidad total del registro: Precio\*Cantidad. |

Con esta data podemos generar varias gráficas que permitirán analizar la información y poder implementar estrategias de marketing e incluso formar alianzas con otras empresas de transporte, bancarias, etc.

A continuación, se plasma algunas ideas de qué datos se deben considerar para algunas gráficas de análisis:

* IvoiceNo, Customer: compras canceladas por determinado usuario, si existe algún producto en especial que se devuelve más o se cancela la compra.
* StockID, Quantity: este campo lo usaría para conocer la información de los productos especiales como: valores cancelados por transporte, por uso de otras plataformas en línea para envíos, tarifas de recarga por transacciones, comisiones especiales como CRUCK Commision, descuentos, etc. Con esa información se podría realizar convenios con empresas de transporte en la época de mayor demanda, crear una tienda física en el país donde más se envían los productos.
* Description, TotalValue, InvoiceDay, Country: los productos más vendidos en una época del año, en cierto país.
* InvoideDay, Quantity: las ventas realizadas por mes.
* UnitPrice, Description: los productos más caros, o más baratos.
* CustomerID: un top de los clientes que más compran, se podría aplicar descuentos por ser cliente VIP.
* Country: los países donde más se venden productos.
* TotalValue, InvoideDay: en qué mes hubo mayor o menor venta.

**Parte 3: Visualización**

**Gráfico 1.**

Al ser Reino Unido el país donde se localiza la tienda es totalmente apreciable que el mayor número de ventas se lo realiza en este país, seguido de Países Bajos, Irlanda, Alemania, Francia, Australia y Suiza, en referencia al top 7 de países con mayores ventas.

**Gráfico 2.**

Observando a más detalle, podemos encontrar el producto más vendido en cada país que a su vez está en el top 5 de lugares donde más compras se realizan. Aquí se observa que en Reino Unido se vendió aproximadamente **40000** productos y a su vez el producto más comprado es **JUMBO BAG RED RETROSPOT.**

**Gráfico 3.**

Las ventas iniciaron bajas en el mes de febrero de 2011 pero crecieron progresivamente, alcanzando su pico en **noviembre**, posiblemente debido a la cercanía de la época navideña.  Hay que considerar que los datos del mes de diciembre del 2011 no están completos por lo que no se debería considerar ese mes.

**Gráfico 4.**

Es importante conocer cuáles son los clientes más frecuentes que la empresa tiene, este es el caso del cliente nro. 57656, que es el principal comprador. Para conservar y premiar su fidelidad se podría aplicar descuentos en ciertos productos.

**Gráfico 5.**

Finalmente, se tiene el servicio de envío de productos a diferentes países, donde destaca Alemania con más de 1000 productos comprados. A futuro se podría analizar si es factible un convenio con una empresa de transporte para que así los clientes paguen menos.