Facultad de Ingeniería
Ingeniería Civil Informática

Informe Proyecto Ciencia de Datos

Retail Online

Autores:
Marcelo Guzmán Escobar
Sebastián Herrera Liberona
Pedro Hurtado Lagos
Cristian Meza Cesped
Diego Muñoz Cabrera

Profesor:
Billy Peralta Márquez.

08 de Mayo del 2020

May 7, 2020

Contents

1 Introducción

Definición de proyecto

Este proyecto consiste en encontrar los patrones de compra dentro de un conjunto de datos de ventas online de retail y las distintas relaciones entre los datos de esta. Se implementarán estrategias de procesamiento al dataset como eliminación de valores nulos y duplicados, con el fin de reducir los tiempos de análisis de los datos.

Como primera instancia, el presente informe mostrará una descripción general de los datos del DataSet. Luego, se mostrarán los resultados del análisis 1D y 2D sobre los datos, explicando su representación. Posterior a esto, se explicará el pre-procesamiento implementado con su fundamento, para finalmente mencionar los experimentos que planteamos en este trabajo y los resultados esperados.

La ejecución del programa fue realizada en un computador con procesador Intel Core i5 6300HQ 2.30GHz, con memoria RAM de 16 GB y con un sistema operativo Windows 10.

2 Descripción de los datos

A continuación, se explicará cada columna del DataSet para un mejor entendimiento del proyecto:

Número de filas totales: 541909
Número de columnas totales: 8

- **InvoiceNo:** Número de transacción. Si comienza con una C, significa que es una transacción cancelada.
- **StockCode:** Código de identificación para cada producto.
- **Description:** Nombre del producto
- **Quantity:** Cantidad del producto por transacción.
- **InvoiceDate:** Día y hora en la que la transacción fue efectuada.
- **UnitePrice:** Precio por unidad del producto en libras esterlinas.
- **CustomerID:** Código de identificación del cliente.
- **Country:** País donde vive el cliente.

3 Preprocesamiento de datos

Dentro del análisis al Dataset hecho, se encontraron valores 0, "nulls" valores negativos. Por lo que dentro de la etapa de procesamiento se aplican técnicas de reducción de datos para eliminar estas columnas mencionadas.
Se dice que es una reducción de datos y una limpieza de datos ya que la definición de reducción está dada por "Obtener representación reducida en volumen, produciendo los mismos resultados o similares en el análisis" [1], y además la limpieza habla de "identificar o remover los datos inconsistentes" [1] Donde se obtiene un dataset de menor volumen que el inicial, pero estos datos eliminados no representan una relación, por lo tanto se denominan como inconsistentes para el análisis que se está efectuando .

4 Análisis de datos

Un análisis de los datos permite una mejor comprensión del contexto e instancia en que se sitúa la entidad. Para esta sección, se tomó en consideración el conjunto de datos resultante luego del preprocesamiento.

4.1 Evaluación 1D de los datos

El análisis 1D se manifiesta como una representación numérica o categórica del comportamiento general de una variable, donde se puede utilizar métricas tales como media, mediana, moda, etc. según sea el caso.
Durante la evaluación 1D de los datos se apuntó a obtener distintas medidas de tendencia central, para el análisis de los datos numéricos y no numéricos
Numéricos:

Columna	Promedio	Mediana	Desviación Estandar	Máximo	Mínimo
Quantity	13.153371805709746	6.0	181.588141959350687	80995	1
UnitPrice	3.12559955307896963	1.95	22.240725281425334	8142.75	1
CustomerId	15287.734821710479	15150.0	1713.5677729984204	18287.0	12346.0

Table 1: Tendencia central en columnas numéricas

De la tabla numérica analizaremos los datos de la columna Quantity e ItemPrice, y descartaremos a la columna CustomerID por el momento. De lo anterior se despendre el promedio, mediana, desviación estandar, máximo y mínimo.

Para la columna Quantity, podemos observar que las compras por producto en promedio eran de 13 unidades, no obstantes, también se cuenta con una desviación estandar bastante elevada, de 181 aproximadamente. Esto puede influir en el análisis del dato. Lo anterior se puede entender al tener valores tan distintos dentro de la columna, ya que el mínimo es de 1 y el máximo alcance los 80995 productos.

Para la columna UnitPrice se cuenta con un promedio de 3.126 libras esterlinas aproximadamente, pero los precios oscilan mucha, el máximo alcanza las 8142.75 libras esterlinas, mientras el mínimo es de 1 libra. Es por esto que se tiene una desviación estandar que bordea las 22 libras.

En cuanto a CustomerID solo se puede analizar que el cliente que más productos distintos compro fue el correspondiente a la id 18287.

No Numéricos:

Columna	Moda
InvoiceNo	576339
StockCode	85123A
Description	WHITE HANGING HEART T-LIGHT HOLDER
InvoiceDate	11/14/2011 15:27
Country	United Kingdom

Table 2: Tendencia central en columnas no numéricas

De los datos obtenidos en el análisis 1D de las columnas no numéricas se puede concluir que la compra correspondiente al código 576339 fue el que más productos distintos llevó en una sola compra. El StockCode 85123A fue el más repetido y se relaciona con la columna Description ya que corresponde a "WHITE HANGING HEART T-LIGHT HOLDER", como el nombre del producto más vendido. Por último se obtiene que la mayoría de los clientes recide en United Kingdom y la fecha donde más productos se compraron corresponde al 14/11/2011

4.2 Analisis 2D

Por otro lado, es posible realizar un análisis 2D del DataSet representado en Mapas de calor, Diagramas de dispersión e Histogramas, entre otros.
Para la representación de la frecuencia de compra de los productos es adecuado llevar su representación a un histograma para su mejor comprensión

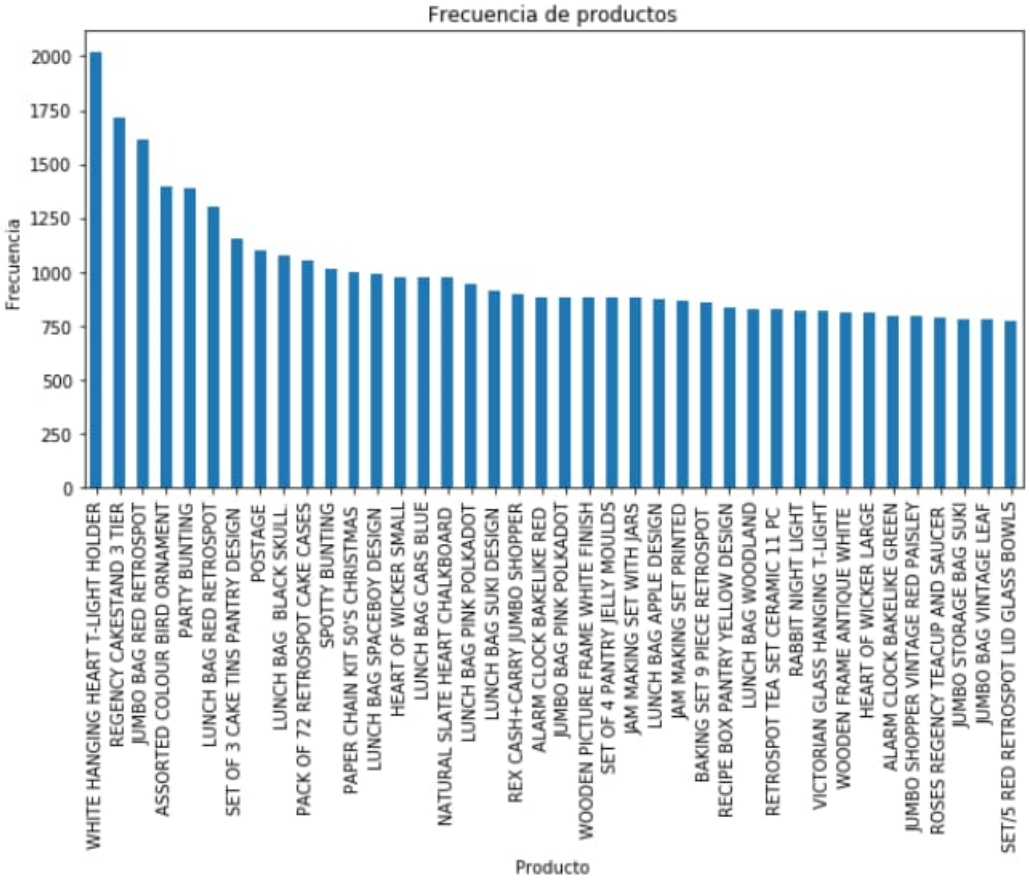

Figure 1: Histograma frecuencia de productos

Con lo representado en el diagrama anterior, podemos observar que el producto que más se compra es “WHITE HANGING HEART T-LIGHT HOLDER”.
Por otro lado, también podemos representar el país en donde reside el cliente que más frecuencia posee

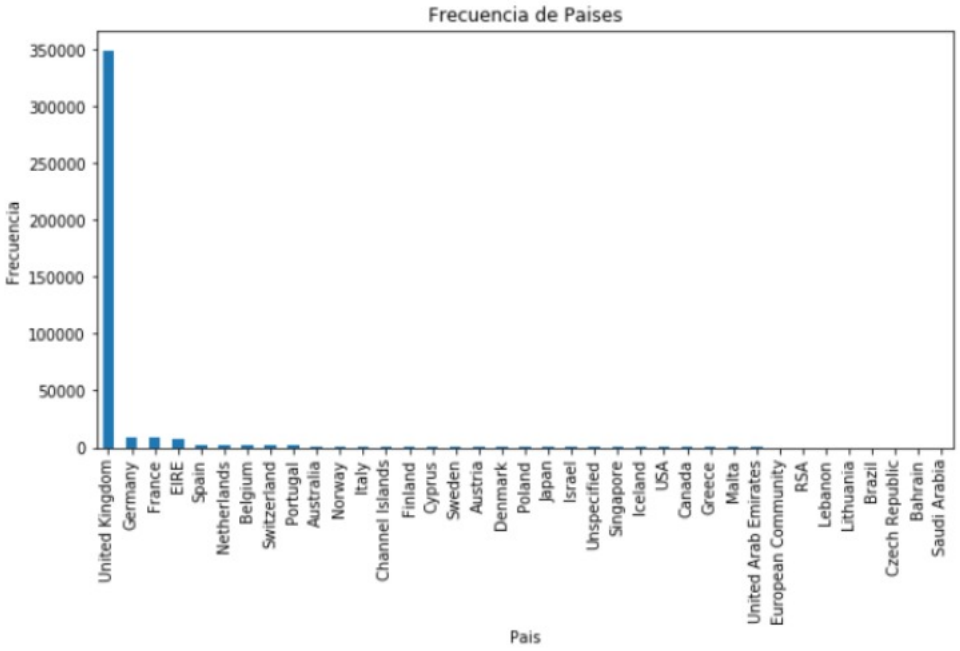

Figure 2: Histograma frecuencia de paises

En este diagrama podemos observar que el país que más se repite es United Kingdom, eso significa que gran parte de las operaciones son hechas por personas con residencia en el Reino Unido, entregando información que podría ser valiosa a la hora de entrablar relaciones

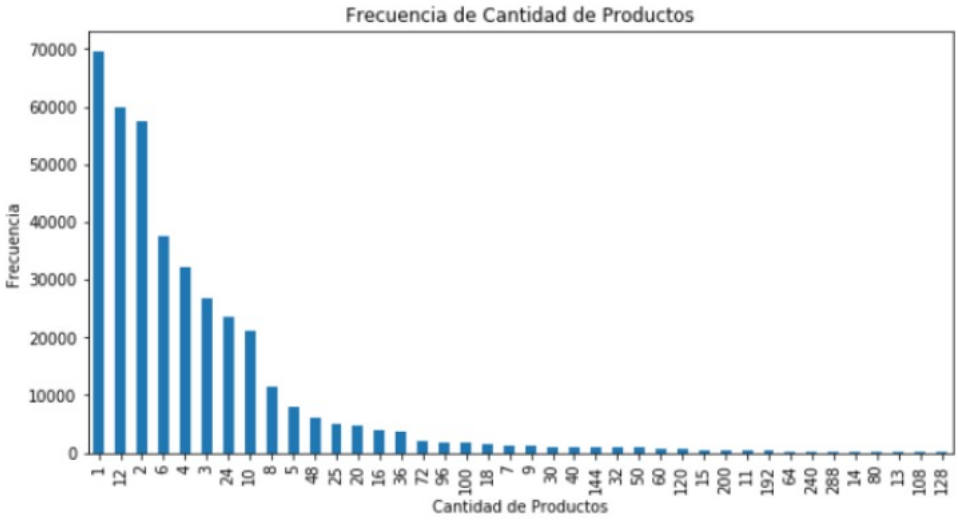

Figure 3: Histograma frecuencia de cantidad de productos

La figura 3 muestra que las transacciones que más se repiten es la de llevar solo 1 producto, en segundo lugar 12 productos y en tercero solo 2 productos, de lo anterior se puede deducir que las compras generalmente son de pocos productos y más frecuentes, ya que las compras de mayor cantidad de productos tienen un frecuencia mucho menor. Esto se ayuda a complementar lo obtenido en el análisis 1D de la columna Quantity, y favorece a su mejor comprensión a través de un desglose de la cantidad de productos, ya que en la primera revisión de los datos se obtuvo un promedio de 13 prodcutos, cuando la moda de esta columna es de 1 producto.

Con el uso de un mapa o matriz de calor, podemos observar la correlación entre multiples variables. A continuación se mostrarán resultados obtenidos del análisis de variables como Quality, UnitPrice y CustomerID

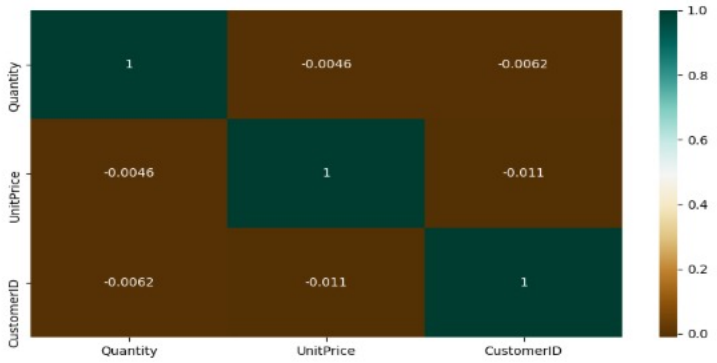

Figure 4: Matriz de calor

La matriz anterior nos describe la correlación entra estas variables, donde se pueden identificar todas como correlaciones negativas, unas más fuertes que otras, pero en su mayoria cercanas a 0, lo que las definiría como una relación débil o nula entre estás variables.

En base a Diagramas de dispersión se puede observar el flujo de compras por día, como es el caso de las figura 5, donde se realizó un modificación a la columna InvoiceDate y se obtuvo solo la fecha de esta, para poder agrupar según día en que se realizo la compra

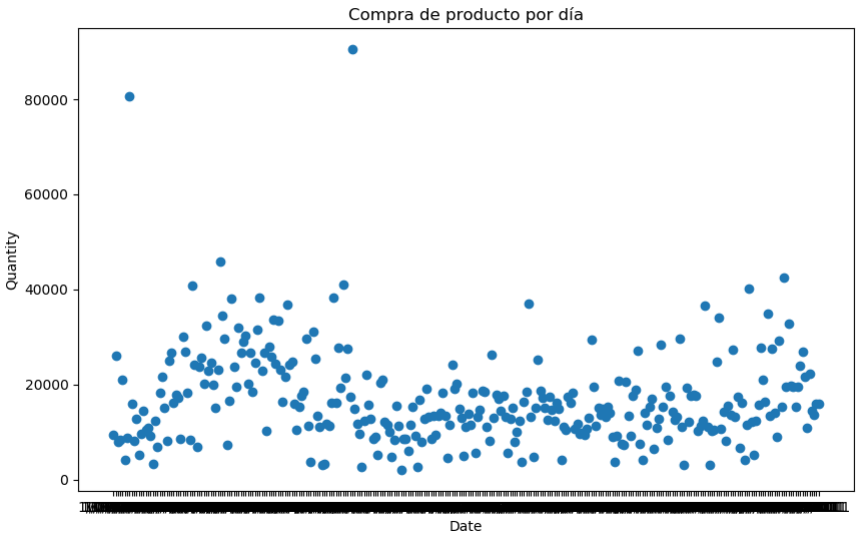

Figure 5: Diagrama de dispersión de compras por día

Del anterior diagrama es muy dificil sacar una conclusión ya que se maneja una gran cantidad de datos aun en la columna Date (305 tuplas). En base a lo anterior es que se decidió refinar más a la columna en base a obtener un diagrama concluyente para un análisis

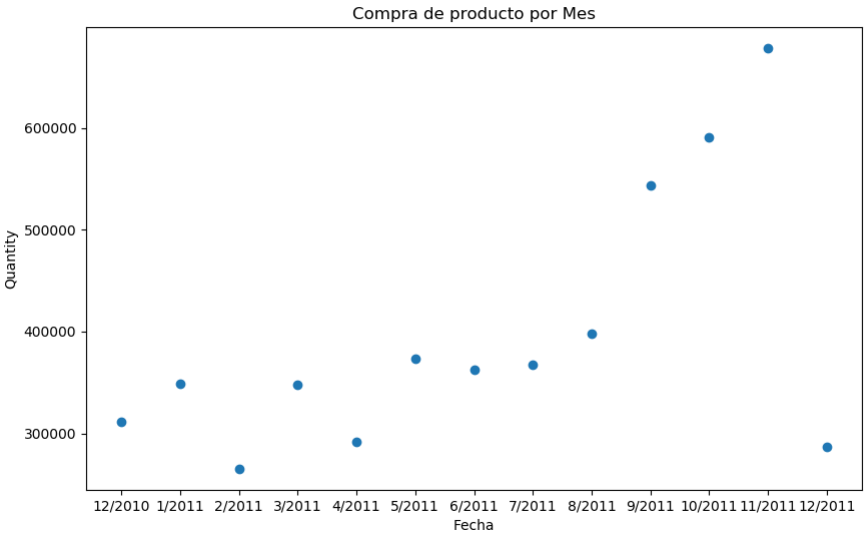

Figure 6: Diagrama de dispersión de compras por mes

Como se puede observar, el diagrama se ha agrupado según los meses del estudio, correspondiente a 13 meses. La mayor cantidad de compras se registró durante los meses de Septiembre, Octubre y Noviembre del año 2011. También se puede apreciar que los peores meses en cuanto a cantidad fueron Febrero, Abril y Diciembre 2011.Esta información podría ser muy beneficiosa a la hora de generar ofertas o estrategias de ventas

5 Diseño de experimentos

Posibles experimentos:

- Encontrar las fechas en que se obtuvo mayor ganancia de dinero. Relacionar Quantity, UnitPrice e InvoiceDate
- Uso de árboles de decisión para formar clusters ya que no existen etiquetas de salida (aprendizaje no supervisado). Cómo se relacionan los clientes con los productos.
- Recomendaciones de productos
- Horarios de mayor fluctuación de datos (mayor cantidad de compras)
- Relaciones entre los productos con el uso de algoritmo a-priori (cálculo de confianza, soporte y lift)

References

[1] Claudia Hernández Jorge Enrique Rodríguez Rodríguez. *Preprocesamiento de datos estructurados*. 2008.