Fecha:
Diciembre 2024

Memoria de

EDA TIENDA ONLINE

Dataset + 500k filas

Realizado por:
Antonio Herrera

1. Introducción

Una tienda online necesita ayuda para reconocer la proyección en ventas, los países donde están teniendo más éxito y una visión general de la tienda a partir de sus ventas/devoluciones. Hemos tomado de referencia un dataset +500k filas.

Fuente: https://archive.ics.uci.edu/dataset/502/online+retail+ii

En este proyecto analizaremos una tienda online con sus ventas y devoluciones. Teniendo en cuenta del país de donde procede dichos movimientos.

2. Hipótesis / Objetivos.

A. "EL Q4 FUE EL TRIMESTRE CON MÁS VENTAS"

- Se entiende que Navidades es el mes más fuerte.
- ¿Hay mucha diferencia con el resto de trimestres?

B. "EL REINO UNIDO ES EL PAÍS QUE MÁS COMPRA"

- Es cierto que es un conjunto de países pero, ¿cómo de dominada está la zona?
- ¿Hay mucha diferencia con el segundo país que más compra?

C. "LAS DEVOLUCIONES SON MENOS DEL 5 % FRENTE A LAS VENTAS"

- Buscan un bajo porcentaje de devoluciones.
- ¿Podemos verlo geográficamente?

QUEREMOS SABER MÁS

- ¿CUÁL ES EL PRODUCTO MÁS VENDIDO Y SU PROYECCIÓN DE VENTAS DURANTE EL AÑO?
- PRODUCTO MÁS DEVUELTO Y VERLO POR TRIMESTRE.
- PRECIO MEDIO DE COMPRA.
- EL PAÍS QUE MÁS DEVUELVE.

Se abordarán las hipótesis y objetivos mediante un enfoque estadístico y visual.

3. Exploración de los datos..

Atendiendo a nuestro dataset, vemos +500.000 filas y 8 columnas. En ellas vemos algunos fallos y algunos valores que manipular para hacer posteriormente un análisis más limpio.

- Añadimos las columnas “Trimestres”, “Devolucion”, “Ventas” y “Precio_Total”.
- Ponemos en positivo la columna “Quantity” para un mejor análisis.
- Corregimos los nulos.
- Dividimos el dataset en 4 archivos que hace referencia a los 4 trimestres del año.
- Limpiamos numerosos outliers que eran datos irreales.

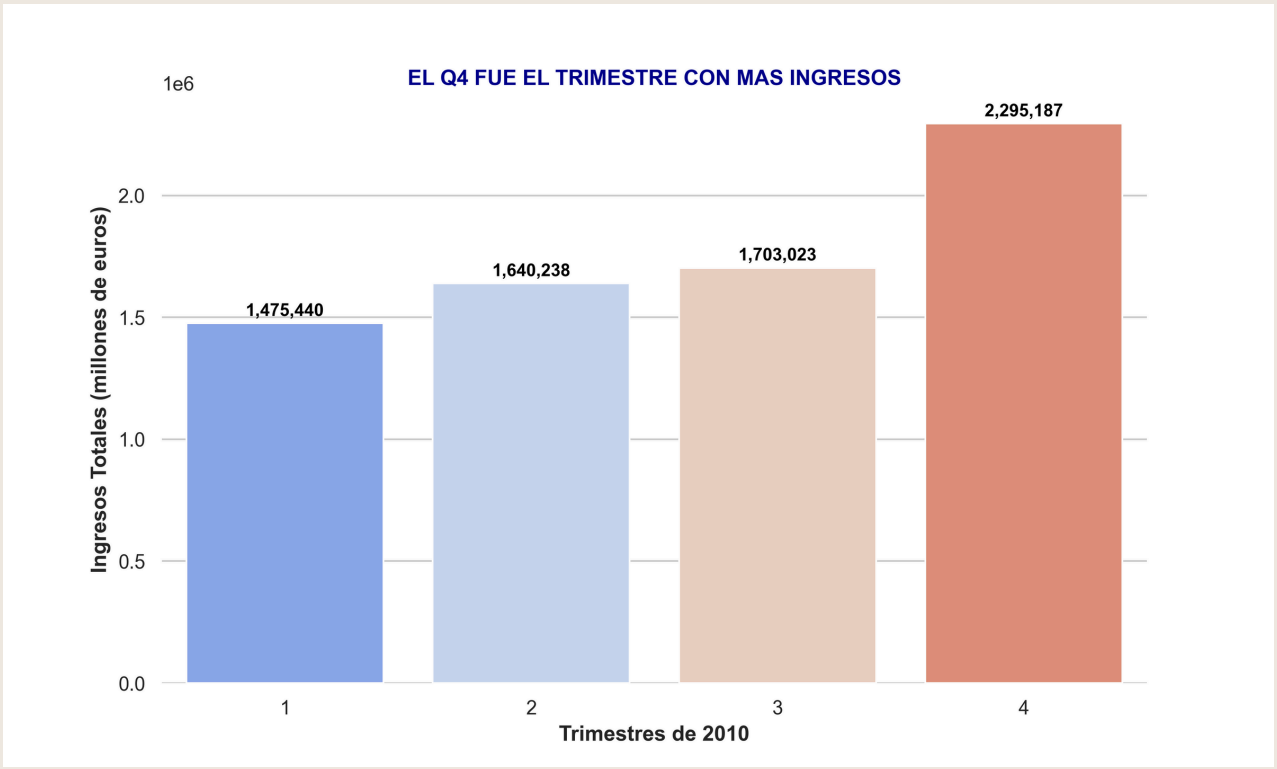
Finalmente nos quedamos con un dataset de 460.000 filas y 12 columnas habiendo sacado la tabla de variables y las medias/modas/medianas.

Columna/Variable	Descripción	Tipo_de_Variable	Importancia inicial	Nota
Invoice	Numero de factura o ticket	Numerica Discreta	3	Hay numeros y letras
StockCode	Numero identificativo del producto	Numerica Discreta	1	
Description	Nombre detallado del producto	Numerica Discreta	2	
Quantity	Cantidad de producto adquirido o devuelto	Numerica Discreta	3	
InvoiceDate	Fecha de la factura o ticket	Numerica Discreta	3	
Price	Precio del articulo por unidad	Numerica Discreta	3	
Customer_ID	Identificativo del cliente	Numerica Discreta	2	
Country	Pais de donde procede el movimiento	Numerica Discreta	0	
Devolucion	Devolucion	Binaria	0	True or False
Ventas	Ventas	Binaria	0	True or False
Precio_Total	Cuantia de la factura o ticket	Numerica Discreta	0	
Trimestres	Trimestre al que pertenece el movimiento	Categorica	0	Tenemos 4 trimestres

4. Análisis de los datos.

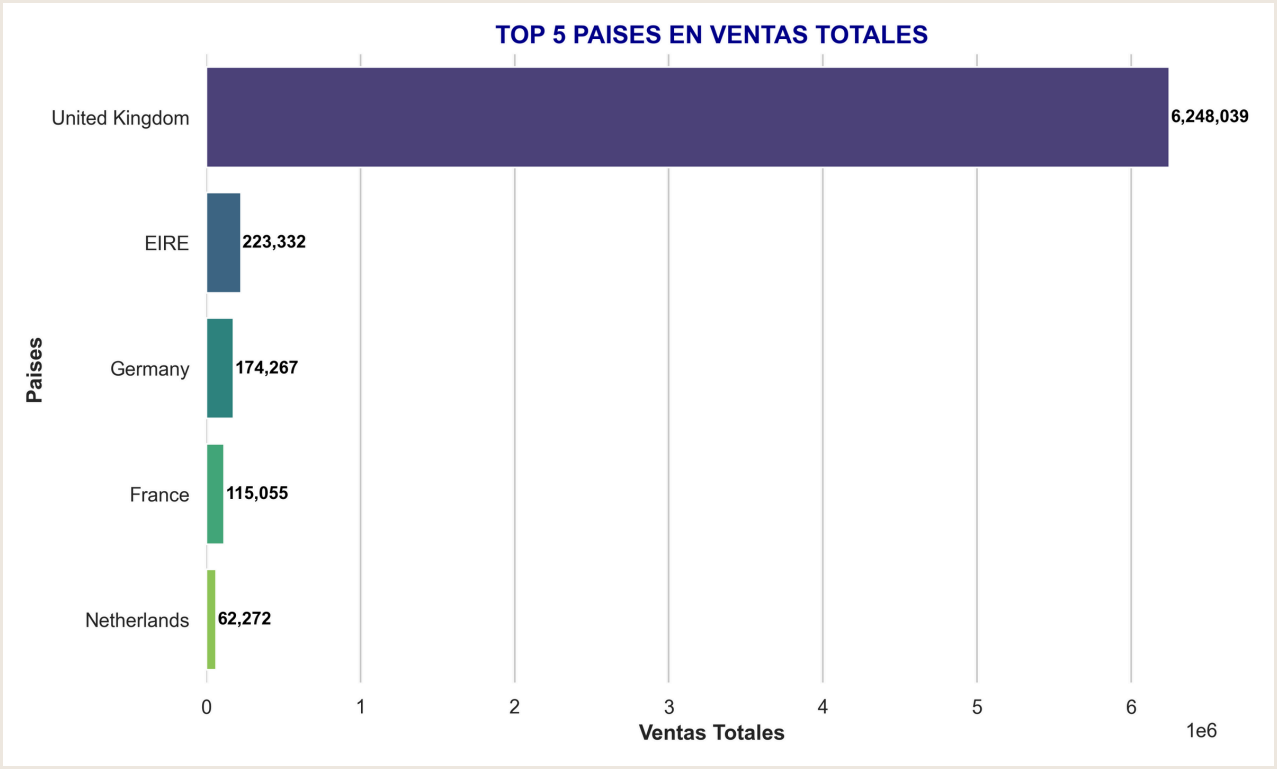
Análisis Univariante, Bivariante y Multivariante.

Como ya sabíamos los objetivos a seguir pudimos sacar las columnas necesarias para que el análisis fuera lo menos complejo posible. Asi que con las columnas “Trimestres” y “Precio_Total” conseguimos ver en una gráfica los ingresos totales por trimestre.



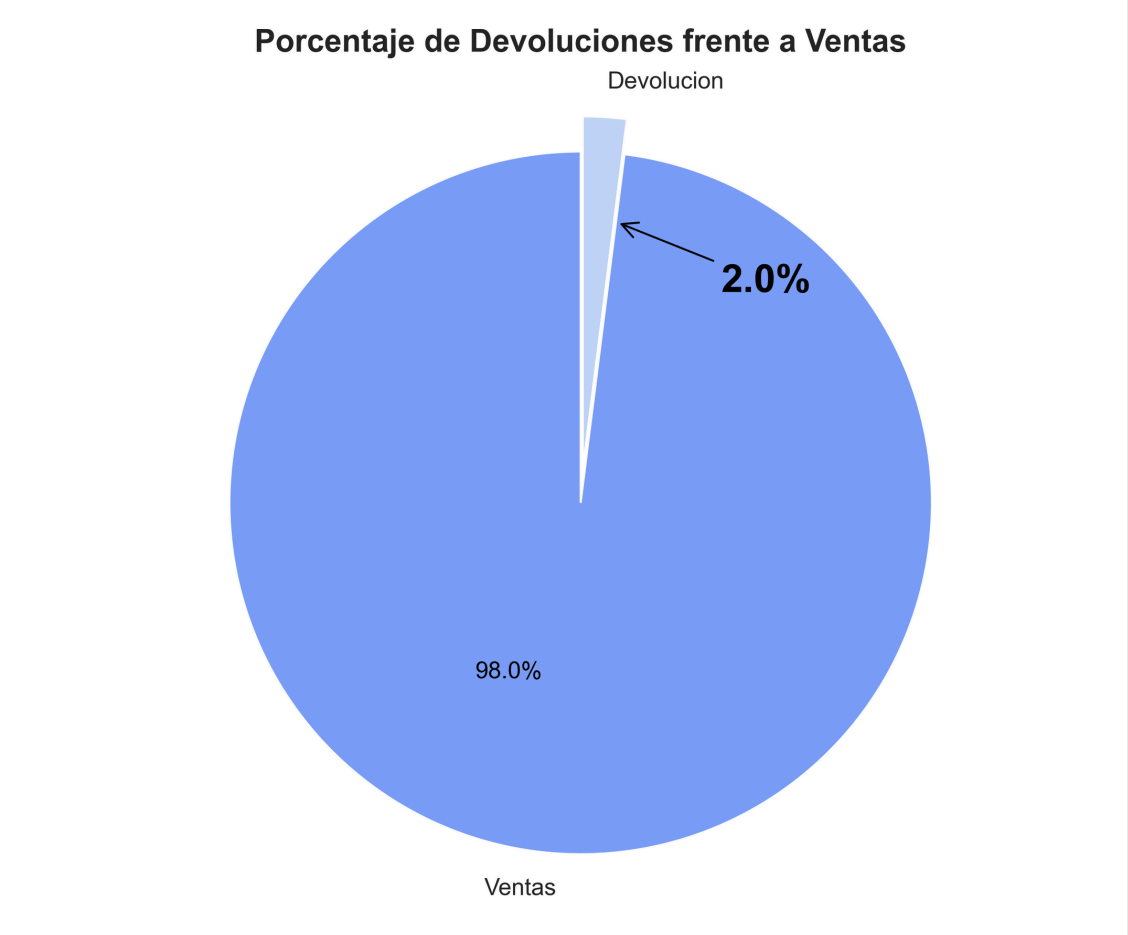
⬆ A. “EL Q4 FUE EL TRIMESTRE CON MÁS VENTAS” ✅

Para el siguiente teníamos que ver “Country”, las “Ventas” y “Precio_Total”. Tendremos que hacer un “sort” y graficar los primeros 5 países que más dinero se dejan en nuestra web.



⬆ B. “REINO UNIDO EL PAÍS QUE MÁS COMPRA” ✅

Para terminar con las hipótesis y pasar a los objetivos secundarios, nos queda por resolver las devoluciones frente a las ventas. Para ello tratamos las columnas “Ventas”, “Devolución” y calculamos sus porcentajes frente al DF completo. Graficamos y sale algo como esto:



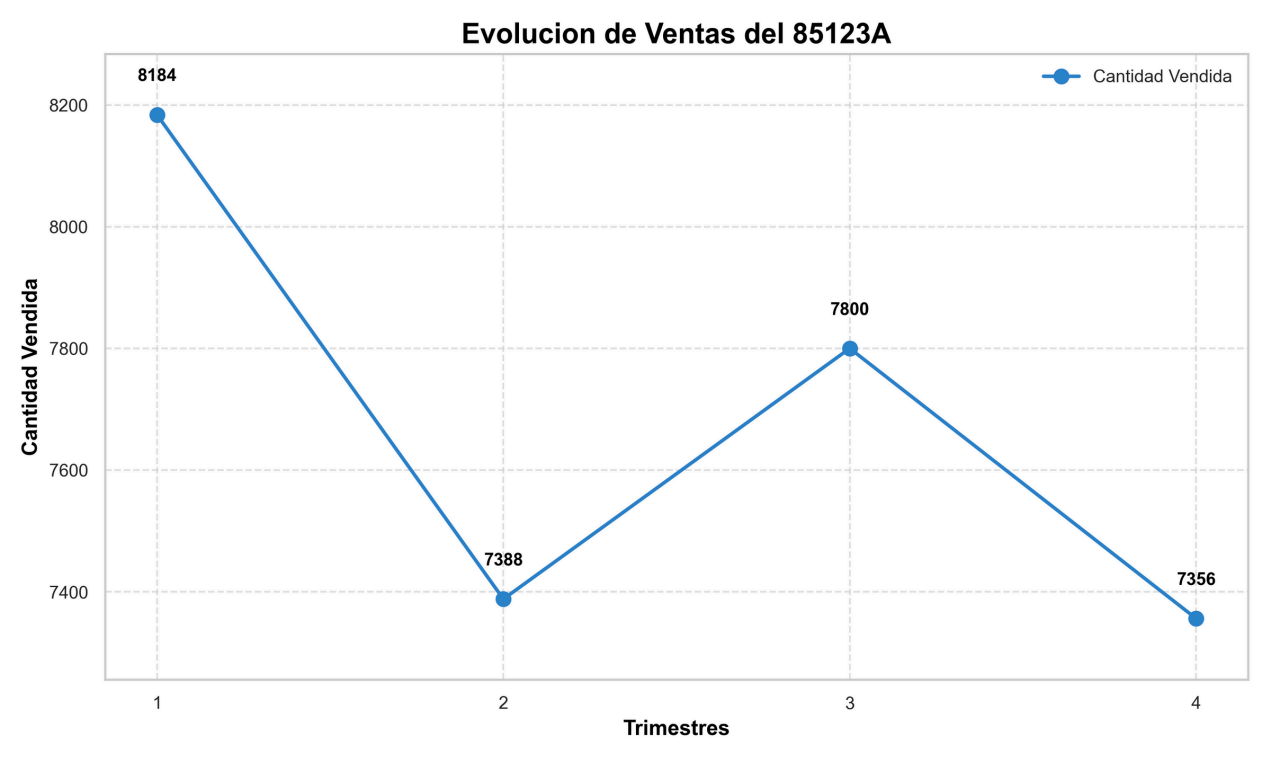
⬆ C. “LAS DEVOLUCIONES SON MENOS DEL 5 % FRENTE A LAS VENTAS” ✅

QUEREMOS SABER MÁS

- **¿CUÁL ES EL PRODUCTO MÁS VENDIDO Y SU PROYECCIÓN DURANTE EL AÑO?**

Miramos “Quantity”, “Ventas” y “StockCode”. Hacemos el groupby y sacamos el índice máximo y valor máximo. Cuando ya tenemos ambos valores, cogemos el “StockCode” y le hacemos un DF que posteriormente le daremos otro groupby pero esta vez con “Trimestres” y “Quantity”.

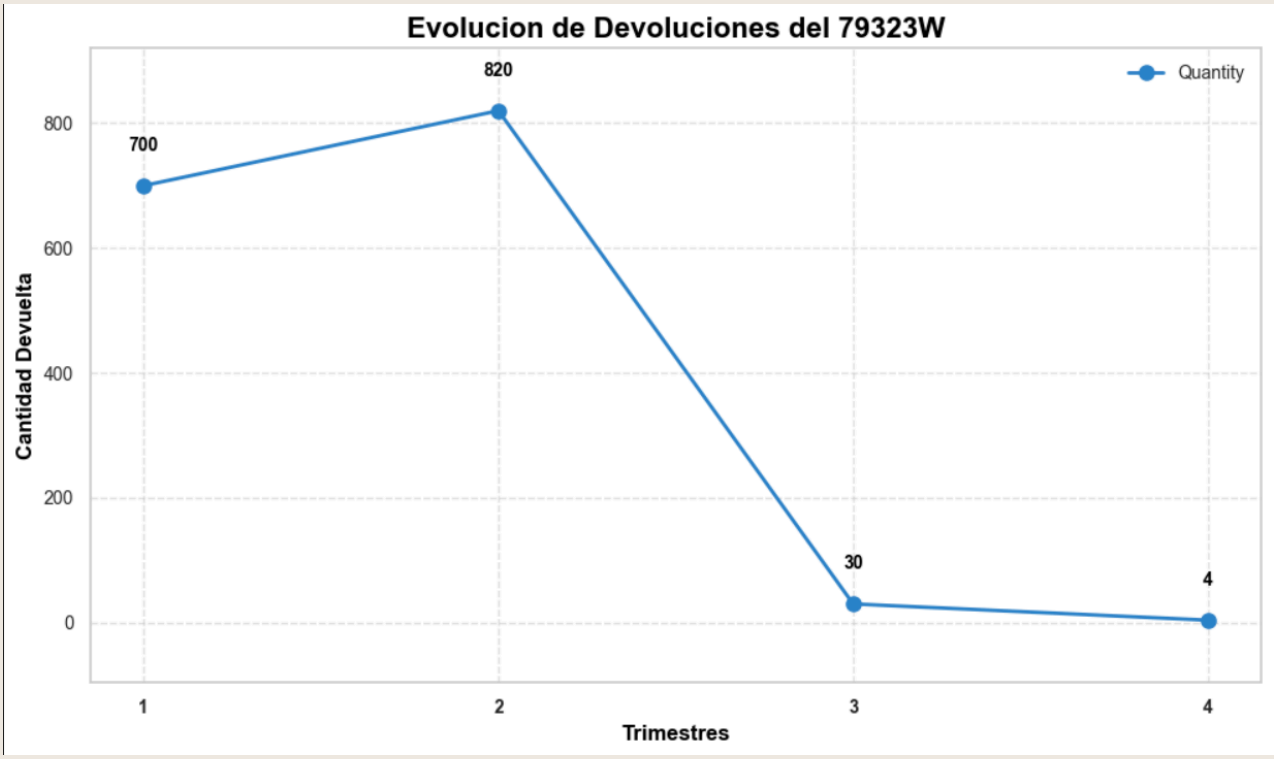
Hacemos el gráfico y obtenemos este resultado: “WHITE HANGING HEART T-LIGHT”



- **PRODUCTO MÁS DEVUELTO Y PROYECCIÓN POR TRIMESTRES.**

Hacemos algo muy parecido a lo anterior, pero esta vez siendo “Devolucion” la columna importante.

obtenemos este resultado: “WHITE CHERRY LIGHTS”



- **PRECIO MEDIO DE COMPRA.**

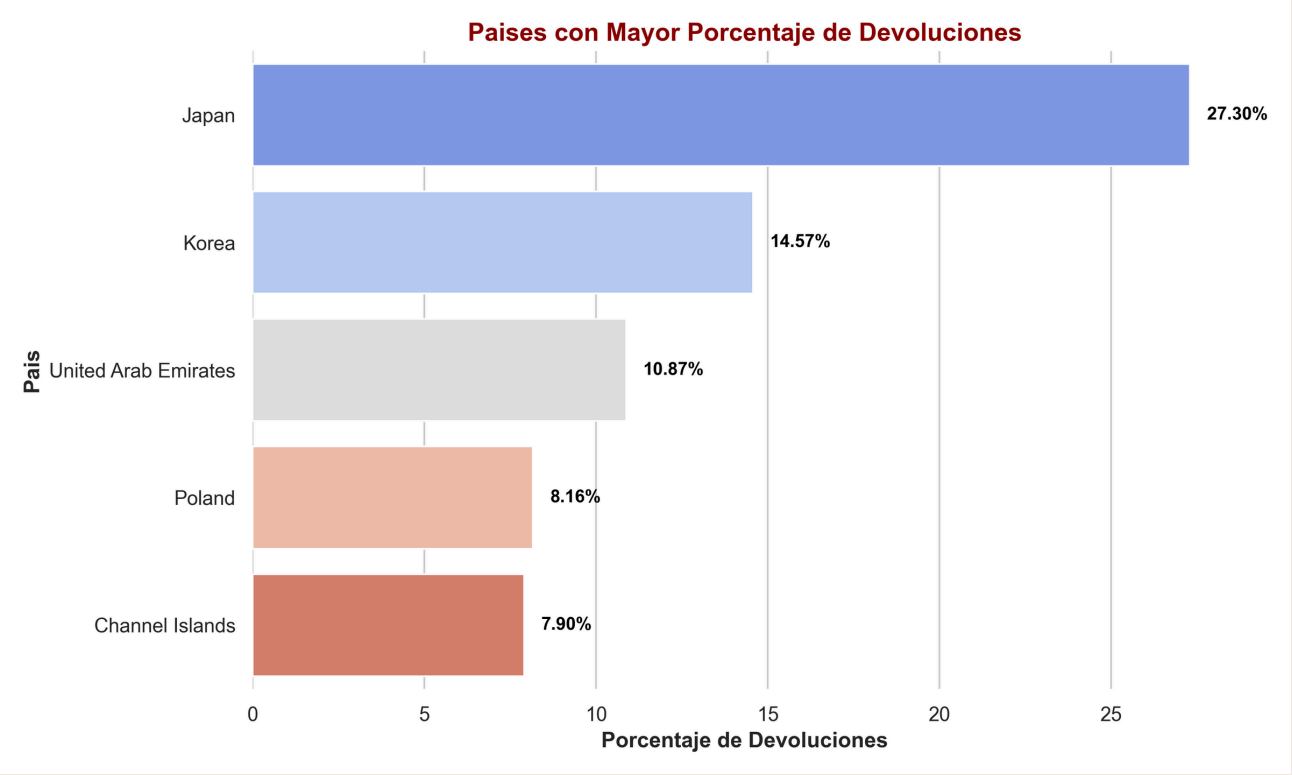
Agrupamos “Invoice” y “Precio_Total” para hacer la media.

Nos da como resultado: **15.62 euros de ticket medio**

- **EL PAÍS QUE MÁS DEVUELVE.**

Utilizamos “Devolucion”, “Country”, “Quantity” y sacamos indice máximo y valor máximo.

Nos damos cuenta de que Reino Unido es el que más devuelve pero es lógico ya que tienen muchas compras. Vamos a verlo desde una parte proporcional con las “Ventas” y las “Devolucion” de cada país. Para ver realmente que país está devolviendo más.



6. Conclusiones.

Tras analizar el dataset sacamos estas respuestas a los objetivos:

- EL Q4 ES EL MEJOR TRIMESTRE Y Q1-Q2-Q3 SON MUY PAREJOS.
- REINO UNIDO ES UN GRAN CLIENTE (+ EIRE)
- DEVOLUCIONES DEL 2 % FRENTE A LAS VENTAS
- LO MÁS VENDIDO ES “white hanging heart t-light holder”
- LO MÁS DEVUELTO ES “white cherry lights”
- EL PRECIO MEDIO DE COMPRA ES 15.60 euros
- REINO UNIDO DEVUELVE MÁS CANTIDAD PERO JAPÓN-KOREA A NIVEL % ES LIDER EN DEVOLUCIONES.

Dentro de mi recomendación, a partir de estos datos sería:

1. Ver que sucede en ASIA porque devuelven un gran porcentaje de productos.
2. El producto más devuelto fue por masividad en los trimestres 1 y 2. ¿Qué pasó?
3. El ticket medio para una tienda online es corto, podemos estudiar los productos que tenemos y generar categorías como “premium”, “basics” de tal forma que incitamos al cliente en pensarse si quiere una gama superior. Eso subiría el ticket medio.

Realizado por:
Antonio Herrera

EDA TIENDA ONLINE