Data summary:

This dataset contains 536,350 records with 8 columns. This is a sales transaction data set of UK-based e-commerce (online retail) for one year. This London-based shop has been selling gifts and homeware for adults and children through the website since 2007. But this data was collected throughout the year 2019 and 2018. And the dataset contains 38 countries.

Columns:

Variable name	Description	Data type	Example value
TransactionNo	a six-digit unique number that defines each transaction. The letter "C" in the code indicates a cancellation.	Object	581482
Date	Indicates when a transaction occurred	Object	01/01/2019
ProductNo	a five or six-digit unique character used to identify a specific product.	object	90214Z
ProductName	Indicates the name of the specific product	Object	Organiser Wood Antique White
Price	Indicates how much the product costs in pound sterling	Float	5.13
Quantity	Indicates the amount of a specific product, negative values indicate that the transaction was canceled	Integer	9
CustomerNo	a five-digit unique number that identifies each customer.	Float	12004.0
Country	The name of the country that the transaction occurred in	Object	United Kingdom

Target Variables**:**

If predicting **sales behavior**, potential targets include:

TotalSales = Quantity * UnitPrice (predict revenue per transaction)

Customer Churn or Segmentation based on CustomerID

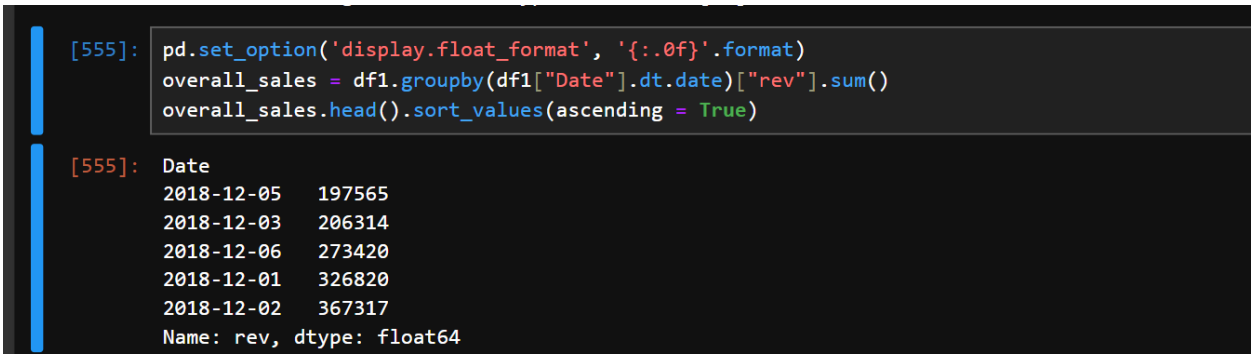
Product Demand based on Quantity

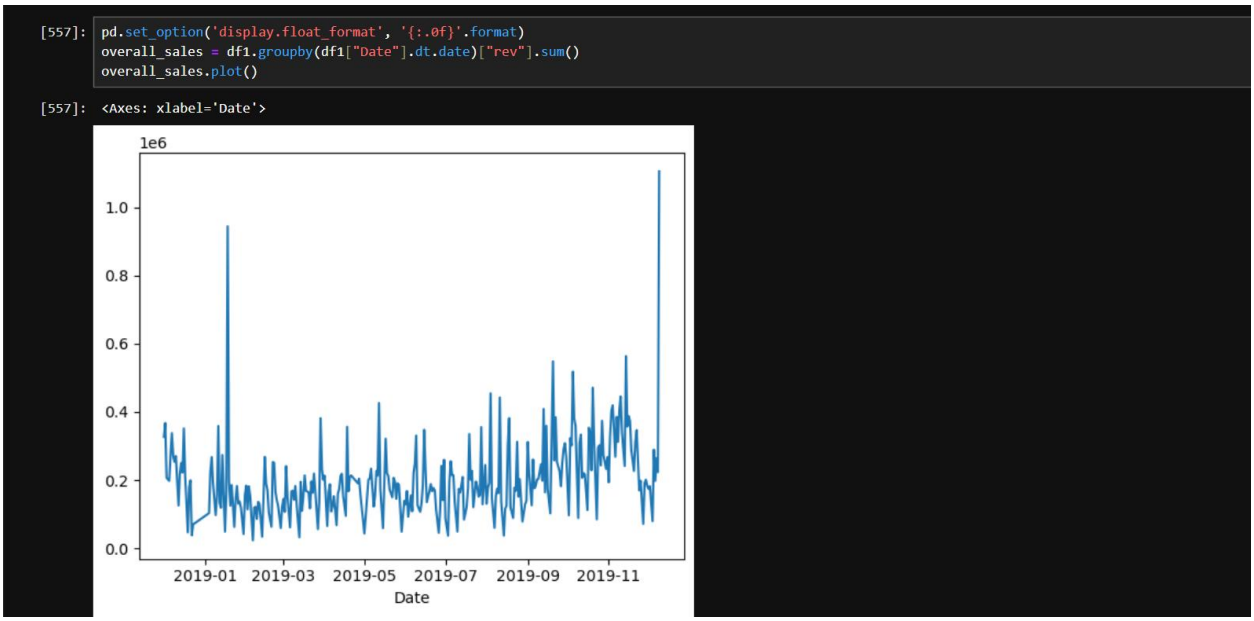
Data Exploration plan:

1. Understand overall sales trends
2. Discover the number of canceled transactions per transaction
3. See which country has the most transactions
4. Discover which products are most popular per country
5. Analyze customer behavior, in terms of repeat purchases and canceled transactions.

Exploratory Data Analysis results:

1. There's an increase in sales from 2018 to 2019, also the month that sales tend to spike is December in both years, where December 5th 2018 having the most sales at 197565





2. The number of canceled orders was 8585

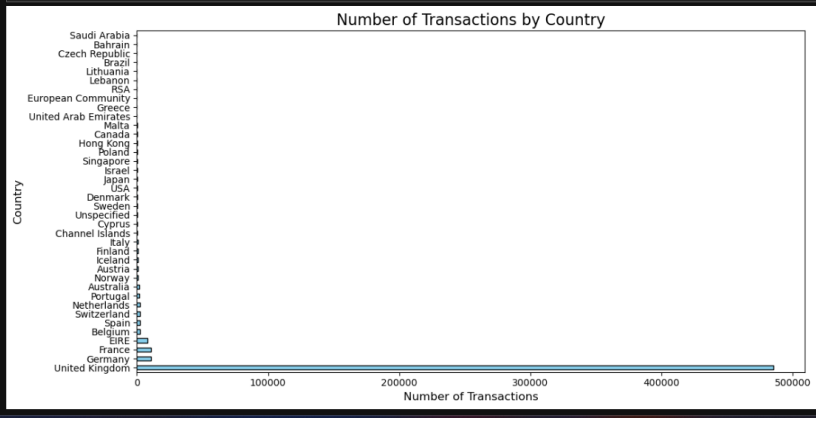
```
[205]: canceled_orders = df[df["Quantity"] <= 0]
       canceled_orders["Canceled_orders"].count()

[205]: 8585
```

3. 3.The country with the most transactions is the United Kingdom with 485095 transactions, and Saudi Arabia having the least transactions with only 10

```
[209]: country = df.groupby("Country")["TransactionNo"].count().sort_values(ascending=False)
       plt.figure(figsize=(12, 6))
       country.plot(kind="barh", color="skyblue", edgecolor="black")

       # Make it look better
       plt.title("Number of Transactions by Country", fontsize=16)
       plt.xlabel("Number of Transactions", fontsize=12)
       plt.ylabel("Country", fontsize=12)
       plt.xticks(rotation=0)
       plt.tight_layout()
       plt.show()
```



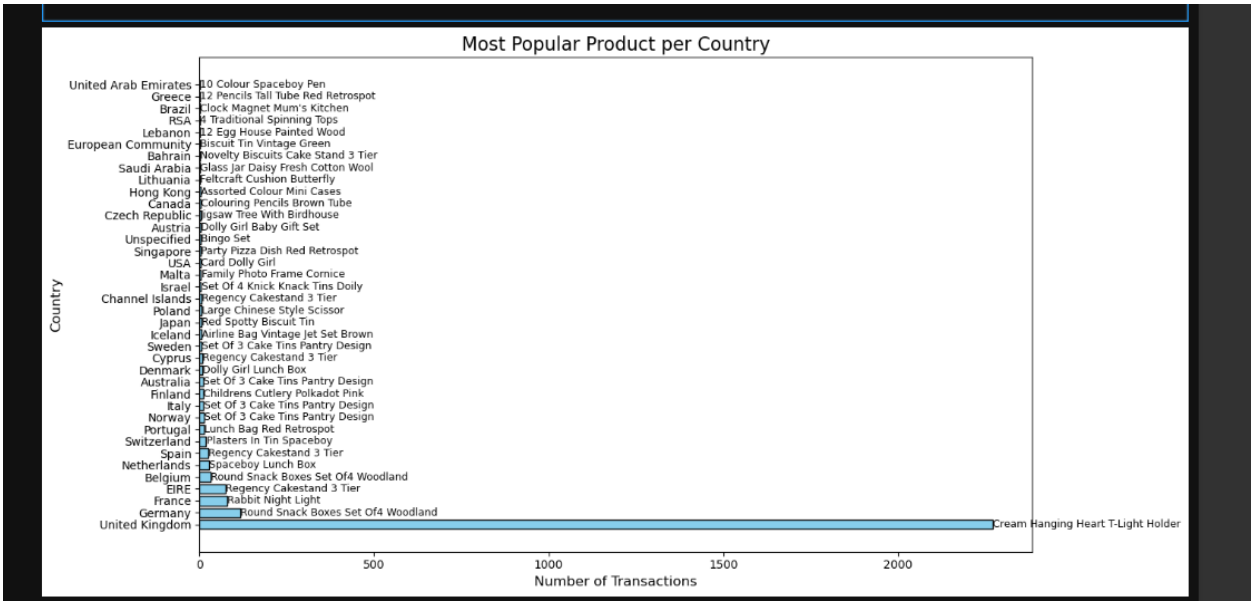
4. The top ten most popular products per country were:

Country	**Top Product**	**Transactions count**
United Kingdom	Cream Hanging Heart T-Light Holder	2271
Germany	Round Snack Boxes (Set of 4)	117
France	Rabbit Night Light	79

Ireland (EIRE)	Regency 3-Tier Cakestand	75
Belgium	Round Snack Boxes (Set of 4)	33
Netherlands	Spaceboy Lunch Box	28
Spain	Regency 3-Tier Cakestand	25
Switzerland	Plasters in Tin (Spaceboy)	19
Portugal	Lunch Bag (Red Retrospot)	14
Norway	Set of 3 Cake Tins	13
Italy	Set of 3 Cake Tins	12
Finland	Polkadot Pink Cutlery (Children's)	10
Australia	Set of 3 Cake Tins	10
Denmark	Dolly Girl Lunch Box	9
Cyprus	Regency 3-Tier Cakestand	8
Sweden	Set of 3 Cake Tins	7
Iceland	Vintage Jet Set Airline Bag	7
Japan	Red Spotty Biscuit Tin	7
Poland	Large Chinese-Style Scissors	7
Channel Islands	Regency 3-Tier Cakestand	6
Israel	Knick Knack Tins (Set of 4)	5
Malta	Family Photo Frame	5

USA	Card Dolly Girl	4
Singapore	Party Pizza Dish (Red Retrospot)	4
Unspecified	Bingo Set	4
Austria	Dolly Girl Baby Gift Set	4
Czech Republic	Jigsaw Tree with Birdhouse	3
Canada	Colouring Pencils (Brown Tube)	3
Hong Kong	Mini Cases (Assorted Colours)	3
Lithuania	Feltcraft Cushion (Butterfly)	2
Saudi Arabia	Daisy Fresh Cotton Wool Jar	2
Bahrain	Novelty 3-Tier Biscuit Stand	2
European Community	Vintage Green Biscuit Tin	2
Lebanon	12-Egg Wooden House	1
South Africa (RSA)	Traditional Spinning Tops (Set of 4)	1
Brazil	Mum’s Kitchen Clock Magnet	1
Greece	12 Pencils (Red Retrospot Tube)	1
UAE	10-Colour Spaceboy Pen	1

```
[237]: top_products = products.sort_values(["Country","TransactionNo"], ascending=[True,False])
       top_products = top_products.groupby("Country").head(1)
       top_products = top_products.sort_values("TransactionNo",ascending= False)
       top_products
```



5. The country with the most repeat purchases is the United Kingdom with 4324 purchases, and the customer with the most purchases in the UK is customer number 17841

```
[379]: repeat_cust = df.groupby(["CustomerNo","Country"])["TransactionNo"].count().sort_values(ascending=False)
       repeat_cust

[379]: CustomerNo  Country
       17841.0     United Kingdom    7967
       14911.0     EIRE              5800
       14096.0     United Kingdom    5093
       12748.0     United Kingdom    4627
       14606.0     United Kingdom    2773
                                     ...
       14025.0     United Kingdom       1
       16953.0     United Kingdom       1
       12791.0     Netherlands          1
       13674.0     United Kingdom       1
       16579.0     United Kingdom       1
       Name: TransactionNo, Length: 4738, dtype: int64
```

```
[381]: repeat_cust = repeat_cust.groupby("Country").count().sort_values(ascending=False)
       repeat_cust

[381]: Country
       United Kingdom          4324
```

Also, the country with the most cancelled orders is the UK with 7324 cancelled orders

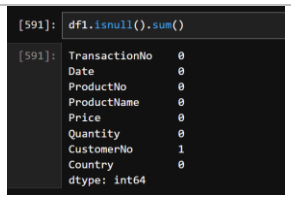
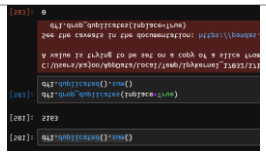
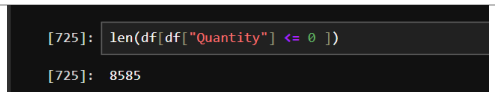
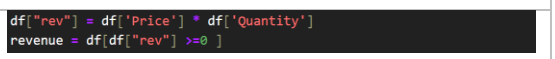
```
[205]: canceled_orders = df[df["Quantity"] <= 0]
       canceled_orders["Canceled_orders"].count()

[205]: 8585

[389]: canceled_orders_country = canceled_orders.groupby("Country")["Quantity"].count().sort_values(ascending=False)
       canceled_orders_country

[389]: Country
       United Kingdom        7324
       Germany                435
       EIRE                   241
       France                 133
       USA                    112
       Australia               73
       Spain                   44
       Italy                   37
       Japan                   34
       Switzerland             33
       Belgium                 32
       Malta                   14
       Norway                  11
       Portugal                10
       Sweden                   9
       Channel Islands          9
       Denmark                  8
       Cyprus                   6
       Finland                  6
       Czech Republic           4
       Netherlands              4
       Austria                  3
       Poland                   1
       Saudi Arabia             1
       European Community       1
       Name: Quantity, dtype: int64
```

Data Cleaning and Feature Engineering:

Step	Description	Example/Screenshot
Handling Missing Values	Dropped rows without Description; removed CustomerID using known purchase history.	
Removing Duplicates	Summed all duplicates, the result was 5163 duplicated rows, then removed them.	
Fixing Negative Quantities	Flagged negative quantities as canceled orders. There were 8585 rows that were canceled orders.	
Encoding Categorical Data	Applied One-Hot Encoding for Country and product nam	
Feature Creation	Created Rev = Quantity * UnitPrice.	

Date column	Converted the Date column from an object to a date data type	`df1["Date"] = pd.to_datetime(df1["Date"])` `df1["Date"]`

Key Insights Summary:

1. Sales Growth & Seasonality
    - Sales increased from 2018 to 2019.
    - December shows consistent spikes in demand across both years, with December 5th, 2018 recording the single highest sales volume (197,565 transactions).
    - This highlights strong seasonality around holiday periods.
2. Cancellations
    - A total of 8,585 transactions were canceled, with the UK accounting for 7,324 of them.
    - Returns or cancellations are a significant issue, concentrated heavily in the UK, the company's primary market.
3. Geographic Concentration
    - The dataset spans 38 countries, but the UK dominates with 485,095 transactions (90% of total).
    - Countries like Saudi Arabia (10 transactions) and others have very minimal sales presence.
4. Product Preferences by Country
    - Product popularity varies by country. For example:
        - UK: *Cream Hanging Heart T-Light Holder* (2,271 transactions).
        - Germany: *Round Snack Boxes (Set of 4)* (117 transactions).
        - France: *Rabbit Night Light* (79 transactions).
    - This indicates localized product preferences and potential for tailored marketing strategies.
5. Customer Behavior
    - The UK has the highest number of repeat customers (4,324), with customer 17841 being the most frequent purchaser.

- However, the UK also leads in cancellations, suggesting both loyalty and return issues are concentrated in the same market.

Hypotheses:

1. Seasonality Effect
    - H1: Sales transactions in December are significantly higher than in other months due to holiday-driven demand.
2. Cancellations by Market
    - H2: The United Kingdom has a significantly higher cancellation rate compared to other countries.
3. Repeat Customer Value
    - H3: Repeat customers contribute a disproportionately larger share of total sales compared to one-time buyers.

Significance Test Discussion: Repeat Customers vs. One-Time Buyers

Context and Rationale

Customer loyalty is one of the most valuable assets for any e-commerce business. While acquiring new customers drives growth, repeat customers often bring sustained profitability. Our dataset revealed that the United Kingdom had the highest number of repeat buyers, suggesting that customer retention could play a key role in revenue generation. To test this formally, we conducted a statistical test comparing the total sales generated by repeat customers versus one-time buyers.

Hypothesis

- Null Hypothesis (H_0): There is no significant difference in average total sales between repeat customers and one-time buyers.
- Alternative Hypothesis (H_1): Repeat customers generate significantly higher total sales than one-time buyers.

Methodology

We calculated the revenue per customer (Revenue = Quantity × Price). Customers were classified into two groups:

- One-time buyers: customers with only one transaction in the dataset.
- Repeat buyers: customers with two or more transactions.

An independent two-sample t-test (Welch's t-test) was applied to compare the mean total sales of the two groups. This test was chosen because it does not assume equal variances and is robust for large sample sizes.

Results

- t-statistic: 11.33
- p-value: 2.37×10^{-29} (extremely small, effectively 0)

With such a low p-value, the probability of observing this difference by random chance is practically zero. Therefore, we reject the null hypothesis and conclude that repeat customers generate significantly higher sales than one-time buyers.

Interpretation and Insights

This result is not only statistically significant but also strategically important. The positive t-statistic (11.33) indicates that repeat customers contribute substantially more value than one-time buyers. This finding aligns with common retail wisdom — that customer retention is often more profitable than customer acquisition.

Interestingly, the UK, which has the highest volume of repeat buyers, also has the highest number of cancellations. This paradox suggests that while loyal customers contribute more revenue, they may also be more demanding, returning products at higher rates. Understanding and managing this tension could be a key competitive advantage.