Table of Contents

[Introduction 1](#_Toc169924586)

[Dataset 2](#_Toc169924587)

[Data Processing Outcomes 2](#_Toc169924588)

[Data Clean & Preparation 3](#_Toc169924589)

[Further Data Cleaning and Initial Analysis 5](#_Toc169924590)

[Further analysis initiated with df\_without\_c. 6](#_Toc169924591)

[Further Analysis and Data Visualisation 6](#_Toc169924592)

[Analysis and Data Visualization for Dataframe df\_without\_c 6](#_Toc169924593)

[Analysis and Data Visualization for Dataframe df\_with\_c 11](#_Toc169924594)

[Data quality 13](#_Toc169924595)

[Conclusion 14](#_Toc169924596)

# Introduction

This dataset includes transactions from December 1, 2010, to December 9, 2011, for a UK online retailer specializing in unique gifts. Many customers are wholesalers. The purpose of this data analysis and visualization task is to gain insights into sales transactions, customer behaviour, and product performance. Key questions include: How do sales change over time and across different regions? What is the customer purchasing patterns, and how do they relate to seasons and regions? What are the popular products, and which products generate the most revenue?

# Dataset

There are eight attributes in this dataset including invoice numbers, product codes, descriptions, quantities, invoice dates, unit prices, customer IDs, and countries.

Dataset Source: <https://archive.ics.uci.edu/dataset/352/online+retail>

A screenshot of a computer

Description automatically generated

There are 541,909 rows and 8 columns in this dataset, and data types including strings, integers, dates, and floats.

# Data Processing Outcomes

The libraries that were employed during the analysis include pandas, numpy, matplotlib and seaborn.

A close-up of a white background

Description automatically generated

Three steps are used for data process purpose:

* Data Clean & Preparation
* Further Data Cleaning and Initial Analysis
* Further Analysis and Data Visualisation

## Data Clean & Preparation

Operations including head and tail are used to get an overview of the dataset.A screenshot of a computer

Description automatically generated

Summary statistics were generated at the beginning of the data cleaning process, and various approaches were applied accordingly.A screenshot of a computer

Description automatically generated

According to the summary statistics, several points need attention: there are negative values in quantity and unit price, and some extremely large values in both. The format of the invoice date needs to be changed to avoid hours and minutes. Additionally, the customer ID should be in string format, not integers.

The following modifications were made:

changing the invoice date format to year-month-day, changing the data type of customer ID to string, and conducting data exploration of the irregular values and outliers in quantity and unit price.

At the end of this step, all the column names were renamed and transformed to lower case to facilitate the analysis process.

Operations used for changing the data types include astype, fillna, dt.date, and dtypes (to check the result). Operations used for column reformatting include rename and str.lower.



## Further Data Cleaning and Initial Analysis

This step begins with two for loops: one calculates the sum and average of 'unit\_price', and the other converts all 'description' values to lowercase.A screenshot of a computer

Description automatically generated

Exploration and data cleaning are applied to the columns 'invoice\_no' and 'stock\_no'. To efficiently observe these columns, a function named detect\_non\_number is created to detect any non-numerical characters. It is noted that invoice numbers with 'a' are entries for adjusting bad debt, and all three rows with 'a' are dropped from the dataset.

Next, exploration and data cleaning are employed for the 'unit\_price' column, focusing on outliers. Methods used include selecting rows from the dataframe by specific conditions and dropping rows from the dataframe by index. The groupby operation is used to locate 'stock\_code' and descriptions for rows that are not cancellation products but have corresponding 'invoice\_no' containing 'c'.

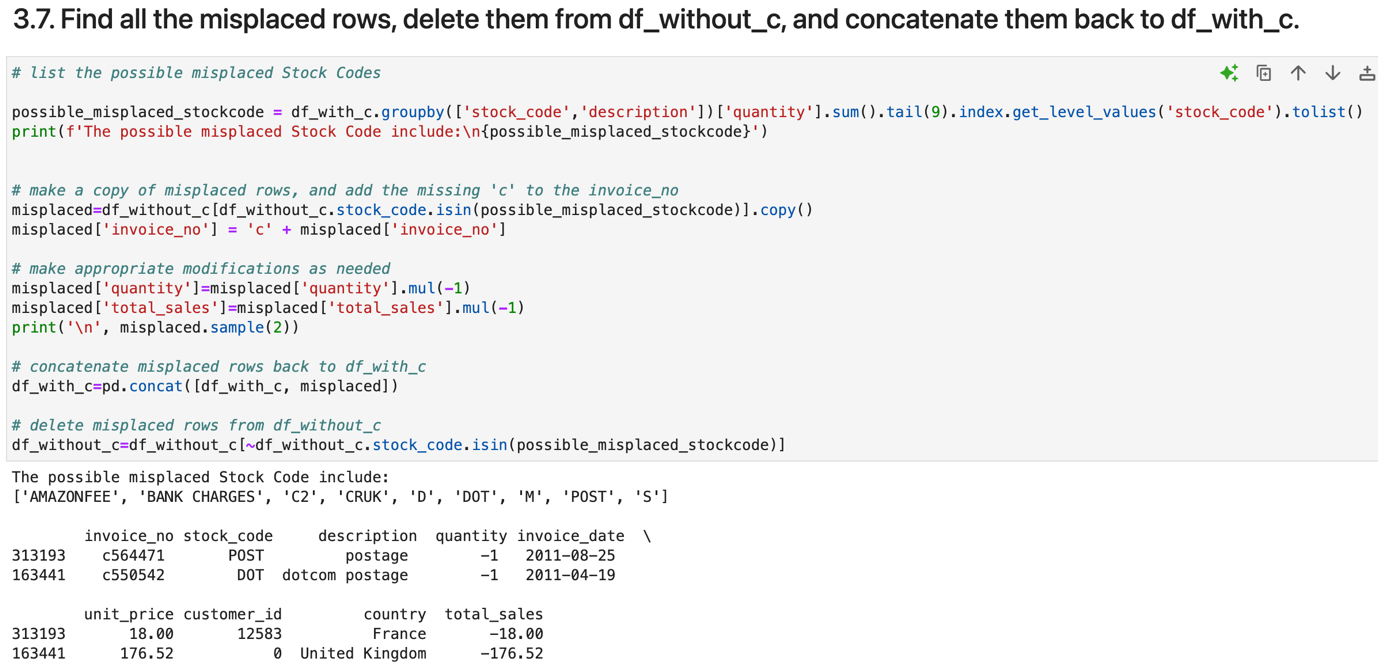
Due to the complexity and size of the dataset, an analytic strategy is implemented to split the dataset into two parts: one with invoices containing 'c' and the other without. Exploration and analysis will be conducted separately for each. A new column, 'total\_sales', is created for analysis and visualization purposes.

A screenshot of a computer code

Description automatically generated

Operations and methods used in this process include assign, startswith, and copy.

### Further analysis initiated with df\_without\_c.

For data cleaning and analysis purposes, the distribution of 'unit\_price' needs to be examined repeatedly. Hence, a function called plot\_box is defined to visualize the summary statistics for 'unit\_price'. This provides a clear view of the distribution of outliers. Upon examination, all the outliers are rows that have been misplaced from df\_with\_c. All the misplaced rows are deleted from df\_without\_c and concatenated back to df\_with\_c.

Operations and methods used in this process include:

groupby, sum, index.get\_level\_value, tolist, mul, sample, concat, isin.

The data clean has completed by now.

## Further Analysis and Data Visualisation

### Analysis and Data Visualization for Dataframe df\_without\_c

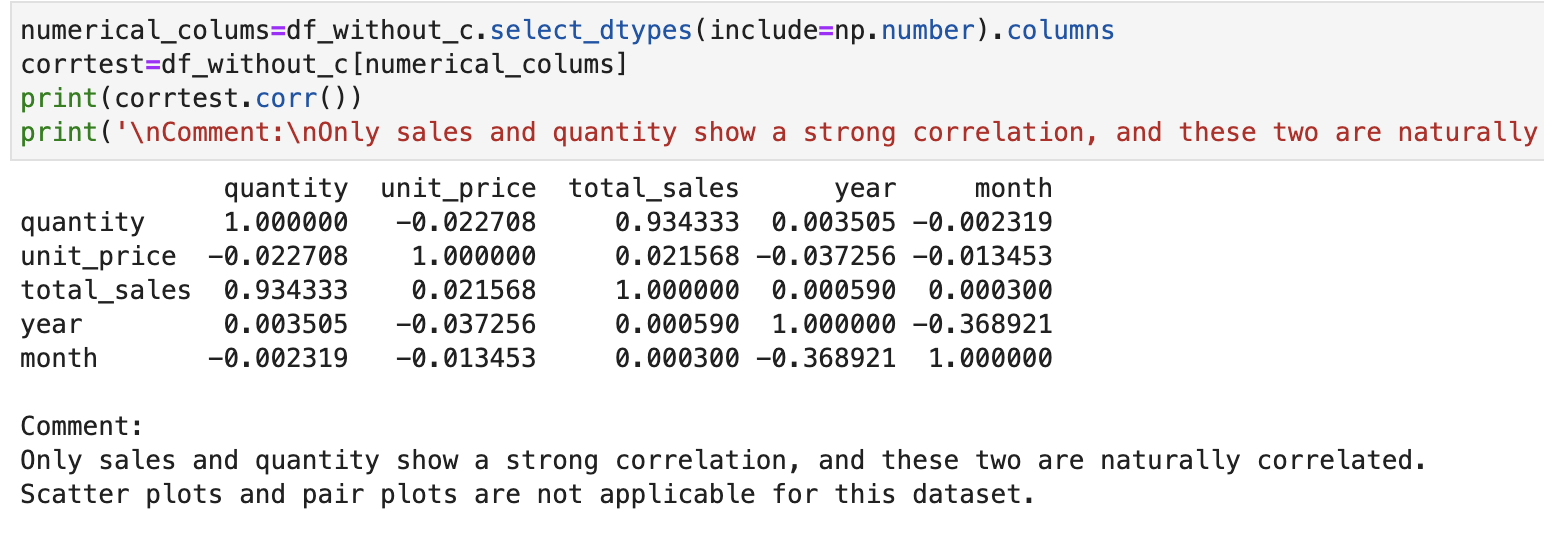
The total sales trends and the correlation relationships between numerical columns are examined first. Two new columns, ‘year’ and ‘month’ are created by extracting values from 'invoice\_date'.

Operations and methods used include:

to\_datetime, dt.year, dt.month, groupby, sum, plot, plt.figure, plt.xlabel, plt.ylabel, plt.title, plt.xticksA screenshot of a computer code

Description automatically generated

A graph showing the growth of sales

Description automatically generated

#### Customer Segmentation Analysis by Region

To facilitate the analysis and visualisation, countries with small contributions are grouped as 'Other Country'.

A screenshot of a computer code

Description automatically generated

Analysis and visualisation of the proportion of count values of different regions:

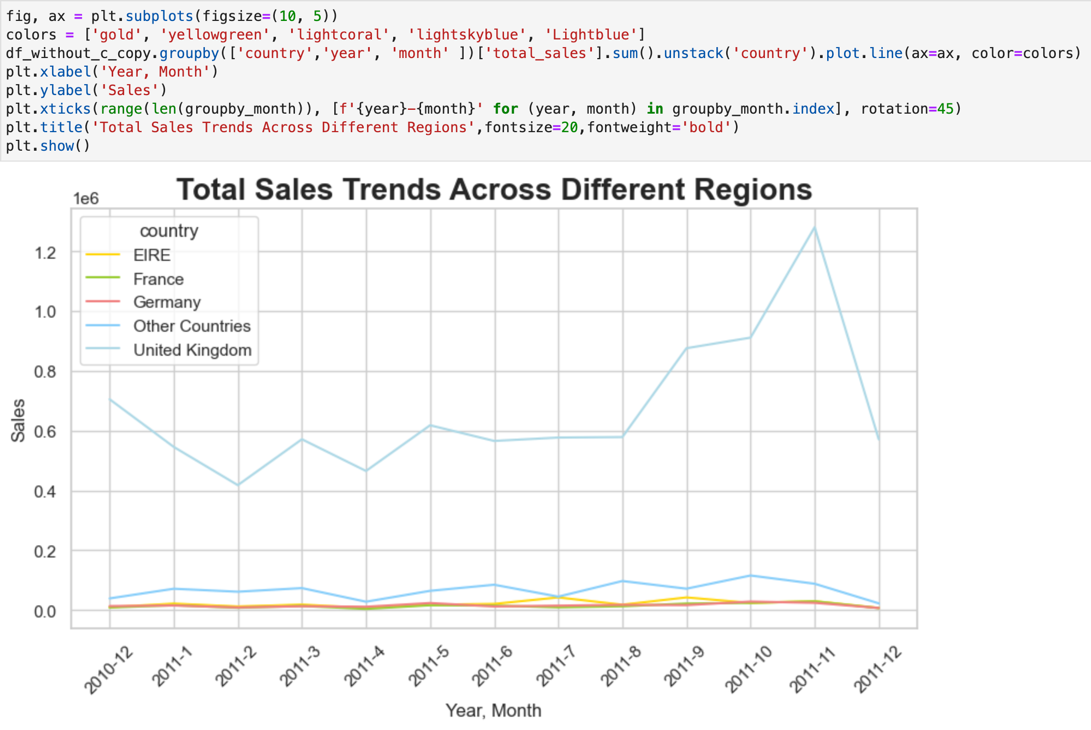
A screenshot of a graph

Description automatically generated

Analysis and visualisation of the percentage of total sales by region: A screenshot of a computer screen

Description automatically generated

Analysis and visualisation of total sales trends across different regions:



#### Product Segmentation Analysis by Price Range

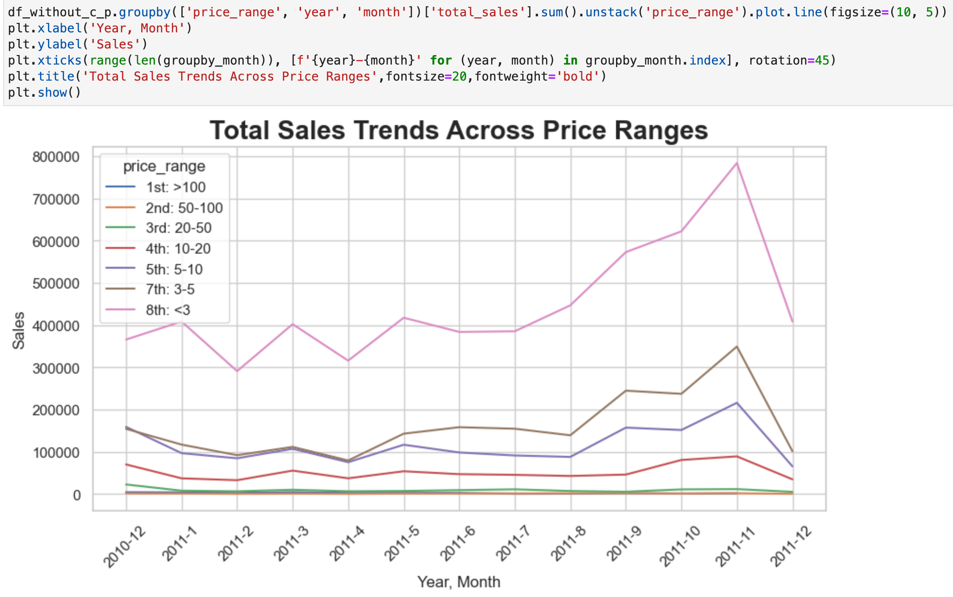
A new column, 'price\_range', is created to segment all products into different price ranges based on the summary statistics for 'unit\_price'.

A function called assign\_price\_range is defined and applied using the apply operation.

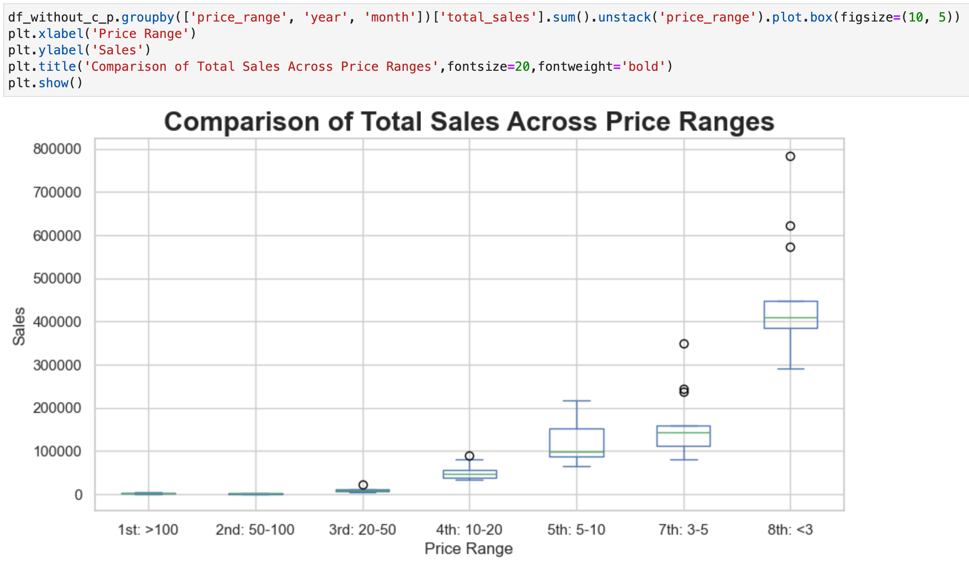
A screenshot of a computer program

Description automatically generated

Analysis and visualisation of total sales trends across different price ranges:



Analysis and visualisation of the total sales between different price ranges:



### Analysis and Data Visualization for Dataframe df\_with\_c

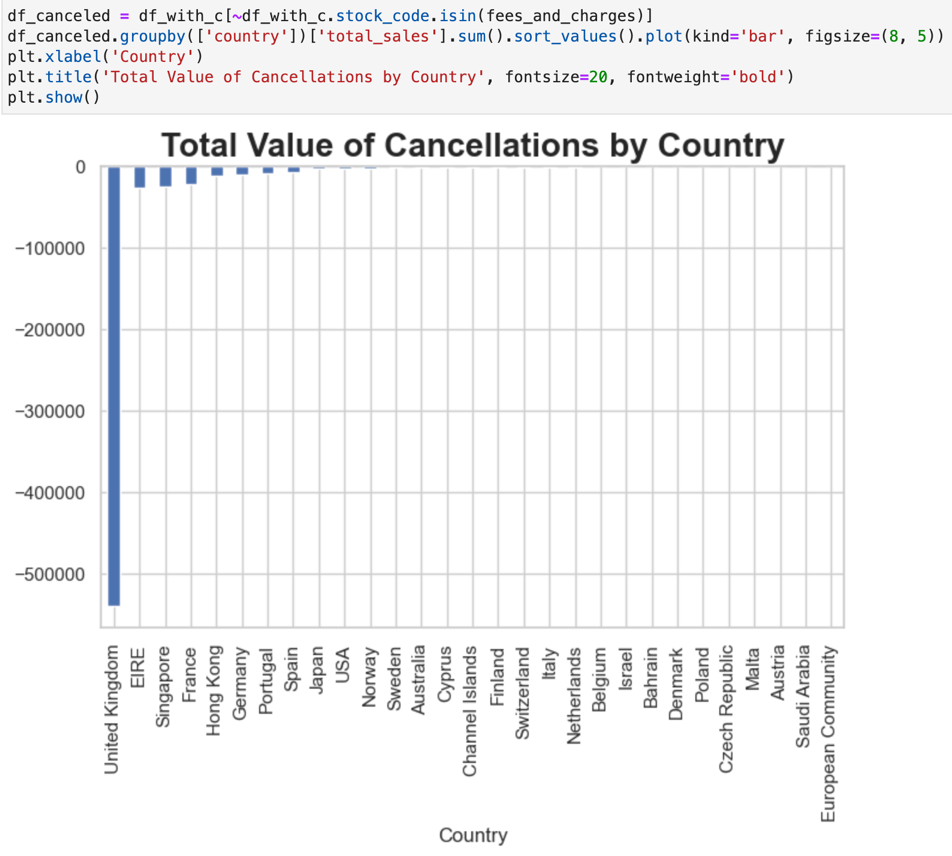
In the previous exploration, the dataset includes two types of information: one concerning fees and charges, and the other related to cancelled orders. Analyses and visualizations have been conducted separately for each category.

Summary of fees and charges:

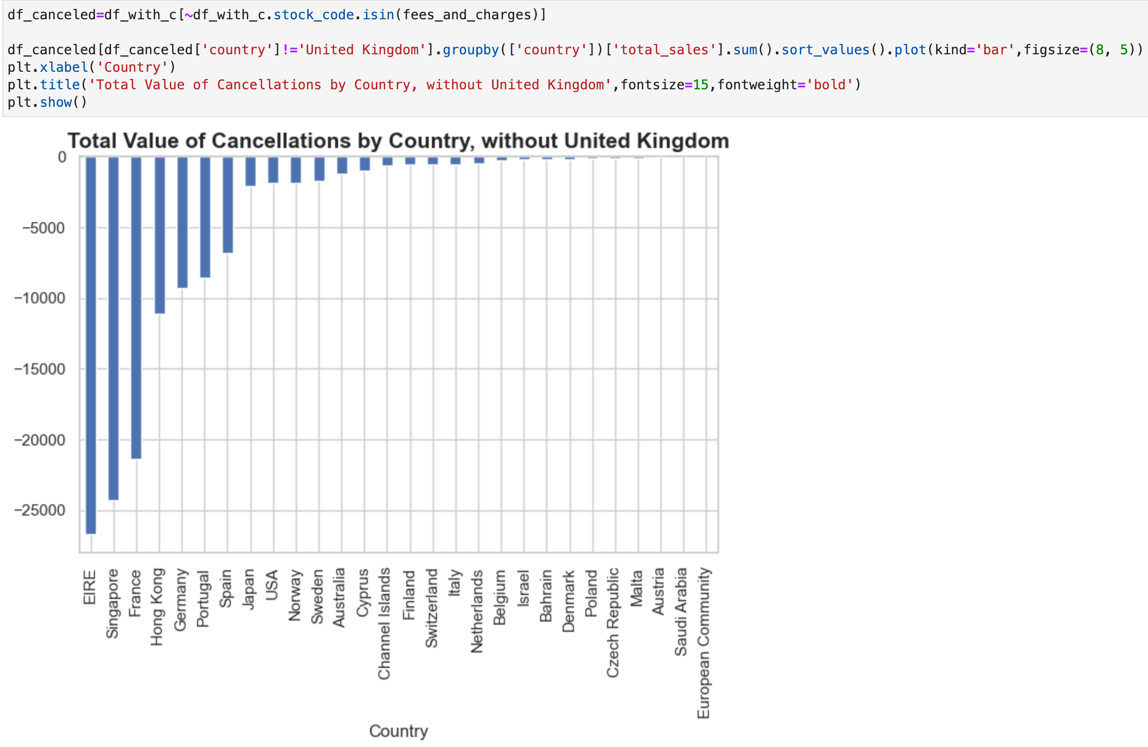
A screenshot of a computer

Description automatically generated

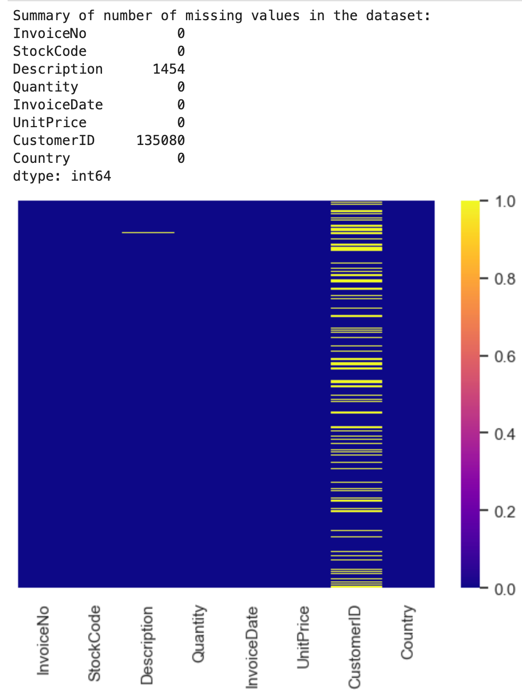
Summary of total cancellation value by country (with UK):



Summary of total cancellation value by country (without UK):



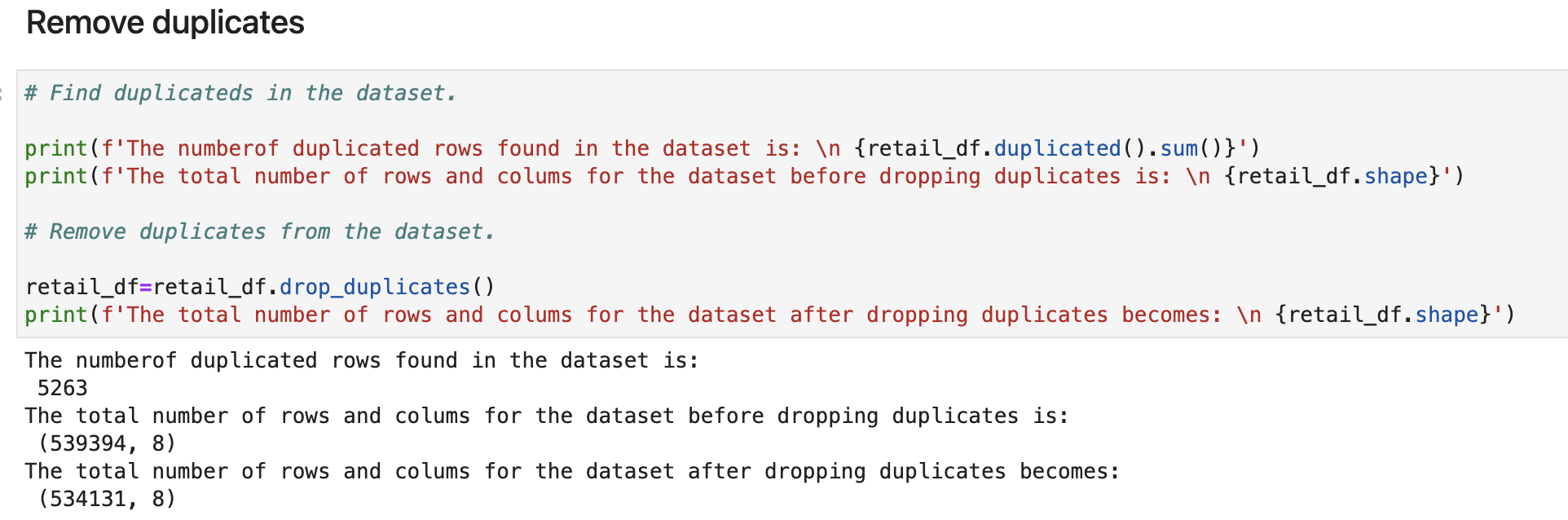
# Data quality

The code ‘retail\_df.isnull().sum()’ generates a summary of the total missing values in each column. This finding is visualized using ‘sns.heatmap’.

Upon exploration, rows with missing descriptions are considered invalid inputs, as their unit price equals 0. Therefore, all rows with missing descriptions have been removed from the dataset. The percentage of these invalid inputs is 0.5%.

Although there were a large number of rows with missing Customer ID, these rows have complete entries for all other columns. Given the wholesale nature of the business, it is reasonable to assume that the Customer ID was missing due to small purchases. These null values were retained and replaced with empty strings during the data cleaning process.

There were 5263 duplicates in the dataset and is removed by the below process:



# Conclusion

The dataset maintains consistency and integrity, with only 5% invalid inputs.

In terms of sales trends, total sales showed minimal fluctuations between late 2010 and the first five months of 2011. Subsequently, they began to rise, peaking in November 2011, before sharply declining toward the year's end.

Customer segmentation analysis reveals that the majority of customers are from the UK, followed by Germany, France, EIRE, and 34 other countries. Orders from the UK represent 91.6% of the total, contributing 85.1% to overall sales. Conversely, orders from other countries constitute 3.7% of orders but contribute 8.4% to total sales. Sales from the UK follow the overall sales pattern, while sales from other countries exhibit a consistent flat trend throughout the year.

Product segmentation analysis shows that products priced under $3 generated the highest sales, with sales decreasing as prices increased.

Among fees and charges, the largest is the Amazon fee, followed by dotcom postage and postage costs.

Regarding total cancellations by value, the UK, as the dominant country, also leads in cancellation value. Apart from the UK, EIRE has the highest cancellation value, followed by Singapore and France. Interestingly, this ranking does not align with their respective sales rankings.