
# DATA ANALYTICS

## BEHIND THE CART

### A deep dive into Online Retail

A Comprehensive Analysis of Sales and Purchasing Trends in the UK-Based Gift Retail Industry

**Katrina JUMADIAO**

JUNE 2023

# TABLE OF CONTENT

# INTRODUCTION

This report provides an in-depth analysis of sales patterns, customer behavior, and revenue generation within the context of a UK-based non-store retail business that sells all-occasion gifts. The impetus behind this exploration is rooted in the immense value data-driven decisions bring to business success, and the paramount importance of understanding customer trends, regional variations, and temporal influences on sales.

The dataset used for this analysis was sourced from the UCI Data Repository and encompasses a rich assortment of sales transactions. As a pivotal component of the study, it offers detailed insights into the purchasing behavior of wholesalers, not individual consumers. This distinction helps emphasize the volume and scale of transactions involved and the corresponding substantial impact minor changes can have on overall revenue.

Our primary objectives include identifying key performance indicators (KPIs), exploring temporal and regional influences on sales, and understanding customer purchasing behavior in terms of quantity and frequency. These goals will enable us to pinpoint potential opportunities for business growth and improved customer engagement.

This topic is of significant interest as it delves into the intricacies of a unique market segment - gift retail. The seasonal nature of gift purchases, coupled with the wide range of occasions, makes this a fascinating area to investigate. Moreover, understanding the nuances of wholesale purchasing behavior provides valuable insights applicable to a broader retail context.

The findings of this report hold potential benefits for a variety of business use cases. It can serve as a roadmap for strategy formulation in terms of inventory management, marketing focus, customer engagement, and revenue optimization. Furthermore, the insights could guide the development of predictive models for future sales forecasting or inform the design of personalized customer engagement strategies.

In essence, this report is designed to unveil valuable insights hidden in the vast ocean of transactional data, aiming to provide a data-driven foundation for decision-making, strategy formulation, and ultimately, business success in the gift retail industry.

# PROJECT PLAN

---

1. Outline project plan using Trello
2. Data collection from the UCI Data Repository
3. Data validation (*cleaning*) for analysis
4. Data preparation (*normalization*) for SQL entities
5. Data importation to SQL Database
6. ERD (Entity Relation Diagram) model creation
7. MySQL Queries for insights extractions
8. EDA (*Exploratory Data Analysis*) and visualization using Python

# DATA COLLECTION

## Data sources and Metadata

---

The dataset used in this report is titled "Online Retail Data Set" and was obtained from the UCI Machine Learning Repository. It is a transnational dataset comprising all the transactions from a UK-based, non-store online retail business that occurred between December 1, 2010, and December 9, 2011. The company primarily sells unique all-occasion gifts, with many of its customers being wholesalers.

| KEY DATASET ATTRIBUTES | | |
|---|---|---|
| **FIELD NAME** | **DESCRIPTION** | **TYPE** |
| InvoiceNo | A 6-digit integral number uniquely assigned to each transaction. If it begins with 'c', it indicates a cancellation | Nominal |
| StockCode | A 5-digit integral number uniquely assigned to each distinct product | Nominal |
| Description | The product (item) name | Nominal |
| Quantity | The quantities of each product (item) per transaction | Numeric |
| InvoiceDate | The date and time when each transaction was generated | Numeric |
| UnitPrice | The product price per unit in sterling (£) | Numeric |
| CustomerID | A 5-digit integral number uniquely assigned to each customer | Nominal |
| Country | The name of the country where each customer resides | Nominal |

The dataset characteristics include multivariate, sequential, and time-series data with 541,909 instances and eight attributes. It was donated to the repository by Dr. Daqing Chen, Director of the Public Analytics group at the School of Engineering, London South Bank University, UK, on November 6, 2015.

The dataset can be accessed from the following link: [UCI Machine Learning Repository: Online Retail Data Set]

This dataset is primarily used for tasks such as classification and clustering in the business domain, with the data types including integer and real values.

# DATA OVERVIEW

## Data summary

Before cleaning, the dataset constituted a diverse mixture of data types across its columns:

| ATTRIBUTE | DATA TYPE | UNIQUE VALUES |
|---|---|---|
| **InvoiceNo** | object | 25,900 |
| **InvoiceDate** | object | 23,260 |
| **StockCode** | object | 4,07 |
| **Description** | object | 4,22 |
| **Quantity** | int64 | 722 |
| **UnitPrice** | float64 | 1,630 |
| **CustomerID** | float64 | 4,372 |
| **Country** | object | 38 |

The dataset had a broad range of unique values in each column, signaling a high degree of variability in the transactions. There are 4,335,272 values in total.

Moreover, the diversity of unique entries across columns emphasizes the rich variety within the dataset, offering extensive opportunities for comprehensive analysis and meaningful insights.

However, the dataset wasn't in its ideal form initially – it contained 136,534 null or missing values and 5,268 duplicated rows. These represented 3.15% and 0.97% of the data, respectively.

This diversity and the issues of missing and duplicated data required a comprehensive cleaning and preparation process to make the dataset suitable for further analysis, and ensure reliable insights and conclusion.

# DATA OVERVIEW

## Data snippet

| | InvoiceNo | InvoiceDate | StockCode | Description | Quantity | UnitPrice | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|
| 0 | 536365 | 12/1/2010 8:26 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | 6 | 2.55 | 17850.00 | United Kingdom |
| 1 | 536365 | 12/1/2010 8:26 | 71053 | WHITE METAL LANTERN | 6 | 3.39 | 17850.00 | United Kingdom |
| 2 | 536365 | 12/1/2010 8:26 | 84406B | CREAM CUPID HEARTS COAT HANGER | 8 | 2.75 | 17850.00 | United Kingdom |
| 3 | 536365 | 12/1/2010 8:26 | 84029G | KNITTED UNION FLAG HOT WATER BOTTLE | 6 | 3.39 | 17850.00 | United Kingdom |
| 4 | 536365 | 12/1/2010 8:26 | 84029E | RED WOOLLY HOTTIE WHITE HEART. | 6 | 3.39 | 17850.00 | United Kingdom |
| 5 | 536365 | 12/1/2010 8:26 | 22752 | SET 7 BABUSHKA NESTING BOXES | 2 | 7.65 | 17850.00 | United Kingdom |
| 6 | 536365 | 12/1/2010 8:26 | 21730 | GLASS STAR FROSTED T-LIGHT HOLDER | 6 | 4.25 | 17850.00 | United Kingdom |
| 7 | 536366 | 12/1/2010 8:28 | 22633 | HAND WARMER UNION JACK | 6 | 1.85 | 17850.00 | United Kingdom |
| 8 | 536366 | 12/1/2010 8:28 | 22632 | HAND WARMER RED POLKA DOT | 6 | 1.85 | 17850.00 | United Kingdom |
| 9 | 536367 | 12/1/2010 8:34 | 84879 | ASSORTED COLOUR BIRD ORNAMENT | 32 | 1.69 | 13047.00 | United Kingdom |
| 10 | 536367 | 12/1/2010 8:34 | 22745 | POPPY'S PLAYHOUSE BEDROOM | 6 | 2.10 | 13047.00 | United Kingdom |
| 11 | 536367 | 12/1/2010 8:34 | 22748 | POPPY'S PLAYHOUSE KITCHEN | 6 | 2.10 | 13047.00 | United Kingdom |
| 12 | 536367 | 12/1/2010 8:34 | 22749 | FELTCRAFT PRINCESS CHARLOTTE DOLL | 8 | 3.75 | 13047.00 | United Kingdom |
| 13 | 536367 | 12/1/2010 8:34 | 22310 | IVORY KNITTED MUG COSY | 6 | 1.65 | 13047.00 | United Kingdom |
| 14 | 536367 | 12/1/2010 8:34 | 84969 | BOX OF 6 ASSORTED COLOUR TEASPOONS | 6 | 4.25 | 13047.00 | United Kingdom |
| 15 | 536367 | 12/1/2010 8:34 | 22623 | BOX OF VINTAGE JIGSAW BLOCKS | 3 | 4.95 | 13047.00 | United Kingdom |
| 16 | 536367 | 12/1/2010 8:34 | 22622 | BOX OF VINTAGE ALPHABET BLOCKS | 2 | 9.95 | 13047.00 | United Kingdom |
| 17 | 536367 | 12/1/2010 8:34 | 21754 | HOME BUILDING BLOCK WORD | 3 | 5.95 | 13047.00 | United Kingdom |
| 18 | 536367 | 12/1/2010 8:34 | 21755 | LOVE BUILDING BLOCK WORD | 3 | 5.95 | 13047.00 | United Kingdom |
| 19 | 536367 | 12/1/2010 8:34 | 21777 | RECIPE BOX WITH METAL HEART | 4 | 7.95 | 13047.00 | United Kingdom |
| 20 | 536367 | 12/1/2010 8:34 | 48187 | DOORMAT NEW ENGLAND | 4 | 7.95 | 13047.00 | United Kingdom |
| 21 | 536368 | 12/1/2010 8:34 | 22960 | JAM MAKING SET WITH JARS | 6 | 4.25 | 13047.00 | United Kingdom |
| 22 | 536368 | 12/1/2010 8:34 | 22913 | RED COAT RACK PARIS FASHION | 3 | 4.95 | 13047.00 | United Kingdom |
| 23 | 536368 | 12/1/2010 8:34 | 22912 | YELLOW COAT RACK PARIS FASHION | 3 | 4.95 | 13047.00 | United Kingdom |
| 24 | 536368 | 12/1/2010 8:34 | 22914 | BLUE COAT RACK PARIS FASHION | 3 | 4.95 | 13047.00 | United Kingdom |
| 25 | 536369 | 12/1/2010 8:35 | 21756 | BATH BUILDING BLOCK WORD | 3 | 5.95 | 13047.00 | United Kingdom |
| 26 | 536370 | 12/1/2010 8:45 | 22728 | ALARM CLOCK BAKELIKE PINK | 24 | 3.75 | 12583.00 | France |
| 27 | 536370 | 12/1/2010 8:45 | 22727 | ALARM CLOCK BAKELIKE RED | 24 | 3.75 | 12583.00 | France |
| 28 | 536370 | 12/1/2010 8:45 | 22726 | ALARM CLOCK BAKELIKE GREEN | 12 | 3.75 | 12583.00 | France |
| 29 | 536370 | 12/1/2010 8:45 | 21724 | PANDA AND BUNNIES STICKER SHEET | 12 | 0.85 | 12583.00 | France |
| 30 | 536370 | 12/1/2010 8:45 | 21883 | STARS GIFT TAPE | 24 | 0.65 | 12583.00 | France |
| 31 | 536370 | 12/1/2010 8:45 | 10002 | INFLATABLE POLITICAL GLOBE | 48 | 0.85 | 12583.00 | France |
| 32 | 536370 | 12/1/2010 8:45 | 21791 | VINTAGE HEADS AND TAILS CARD GAME | 24 | 1.25 | 12583.00 | France |
| 33 | 536370 | 12/1/2010 8:45 | 21035 | SET/2 RED RETROSPOT TEA TOWELS | 18 | 2.95 | 12583.00 | France |
| 34 | 536370 | 12/1/2010 8:45 | 22326 | ROUND SNACK BOXES SET OF4 WOODLAND | 24 | 2.95 | 12583.00 | France |
| 35 | 536370 | 12/1/2010 8:45 | 22629 | SPACEBOY LUNCH BOX | 24 | 1.95 | 12583.00 | France |
| 36 | 536370 | 12/1/2010 8:45 | 22659 | LUNCH BOX I LOVE LONDON | 24 | 1.95 | 12583.00 | France |
| 37 | 536370 | 12/1/2010 8:45 | 22631 | CIRCUS PARADE LUNCH BOX | 24 | 1.95 | 12583.00 | France |
| 38 | 536370 | 12/1/2010 8:45 | 22661 | CHARLOTTE BAG DOLLY GIRL DESIGN | 20 | 0.85 | 12583.00 | France |
| 39 | 536370 | 12/1/2010 8:45 | 21731 | RED TOADSTOOL LED NIGHT LIGHT | 24 | 1.65 | 12583.00 | France |
| 40 | 536370 | 12/1/2010 8:45 | 22900 | SET 2 TEA TOWELS I LOVE LONDON | 24 | 2.95 | 12583.00 | France |
| 41 | 536370 | 12/1/2010 8:45 | 21913 | VINTAGE SEASIDE JIGSAW PUZZLES | 12 | 3.75 | 12583.00 | France |
| 42 | 536370 | 12/1/2010 8:45 | 22540 | MINI JIGSAW CIRCUS PARADE | 24 | 0.42 | 12583.00 | France |
| 43 | 536370 | 12/1/2010 8:45 | 22544 | MINI JIGSAW SPACEBOY | 24 | 0.42 | 12583.00 | France |
| 44 | 536370 | 12/1/2010 8:45 | 22492 | MINI PAINT SET VINTAGE | 36 | 0.65 | 12583.00 | France |
| 45 | 536370 | 12/1/2010 8:45 | POST | POSTAGE | 3 | 18.00 | 12583.00 | France |

# DATA VALIDATION

Data investigation

---

## Column data types

In our initial data investigation, I identified several columns where the data types needed to be changed:

- InvoiceNo required conversion from an object to an integer (int64)
- InvoiceDate needed to be changed from an object to a datetime data type
- CustomerID needed to be converted from a floating-point number (float64) to an integer (int64)
- A new column Revenue will be added by multiplying UnitPrice and Quantity

## Null values and Duplicates

We found 136,534 null values and 5,268 duplicates in the dataset.

```
No. of values in the dataset : 4,335,272
Total rows in the dataset : 541,909
Total columns in the dataset : 8

Total null values : 136,534
Total duplicated rows : 5,268

RATIO OF MISSING AND DUPLICATED VALUES IN OUR DATA :
-----------------------------------------------------------------------------------------------------------------------

Percentage of null values in the data : 3.15%
Percentage of duplicates in the data : 0.97%


NUMBER OF UNIQUE VALUES PER COLUMN :
-----------------------------------------------------------------------------------------------------------------------
Total unique values for column InvoiceNo: 25,900
Total unique values for column InvoiceDate: 23,260
Total unique values for column StockCode: 4,070
Total unique values for column Description: 4,223
Total unique values for column Quantity: 722
Total unique values for column UnitPrice: 1,630
Total unique values for column CustomerID: 4,372
Total unique values for column Country: 38
```

# DATA VALIDATION

Data cleaning

---

## Handling duplicates

We first removed duplicates from the data. The count before and after this operation was:

- Original number of rows: 541,909
- Number of duplicates: 5,268
- Number of rows after removing duplicates: 536,641

## Handling missing values

Next, I dropped rows with missing values. Rows with null values were deemed meaningless for analysis because they lacked customer information. The count before and after this operation was:

- Number of rows before operation: 536,641
- Number of null values: 136,491
- Number of rows after removing null values: 401,604

## Convert data types

We encountered an error when attempting to convert the InvoiceNo column due to non-numeric values. After investigation, I found that some InvoiceNo values began with "c", indicating canceled orders or returned items. To avoid skewing the analysis with these cases, I postponed converting InvoiceNo until after separating successful transactions from canceled ones. We then successfully converted the rest of the columns' data types and added a Revenue column.

## Description column

### 1. Unifying item names

We noticed recurring item names written differently, which led to artificial inflation of unique product counts.

| StockCode | Description | Revenue |
|---|---|---|
| 16156L | WRAP CAROUSEL | 157.50 |
| | WRAP, CAROUSEL | 42.00 |
| 17107D | FLOWER FAIRY 5 DRAWER LINERS | 150.45 |
| | FLOWER FAIRY 5 SUMMER DRAW LINERS | 15.30 |
| | FLOWER FAIRY,5 SUMMER B'DRAW LINERS | 267.75 |

| | StockCode | Description |
|---|---|---|
| 0 | 16156L | 2 |
| 1 | 17107D | 3 |
| 2 | 20622 | 2 |
| 3 | 20725 | 2 |
| 4 | 20914 | 2 |
| ... | ... | ... |
| 208 | 85184C | 2 |
| 209 | 85185B | 2 |
| 210 | 90014A | 2 |
| 211 | 90014B | 2 |
| 212 | 90014C | 2 |

213 rows × 2 columns

### 2. Use of NeatText Functions (NFX)

To clean the data, I used the clean_text function from NeatText Functions (NFX) to process the Description field. This cleaned 42 descriptions, reducing unique descriptions from 213 to 191.

```
# !pip install neattext

# Load Text Cleaning Pkgs
import neattext as nt # Method 1: OOP using TextFrame
import neattext.functions as nfx # Method 2: Using Functional Approach

# dir(nt)
# dir(nfx)
# ?nfx.clean_text

beforeNFX_data = data.copy()

data["Description"] = data["Description"].apply(lambda x: nfx.clean_text(x, puncts=True, special_char=True, contractions=True))
data.head()
```

✓ 21.0s Python

| | InvoiceNo | InvoiceDate | StockCode | Description | Quantity | UnitPrice | Revenue | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 536365 | 2010-12-01 08:26:00 | 85123A | white hanging heart tlight holder | 6 | 2.55 | 15.30 | 17850 | United Kingdom |
| 1 | 536365 | 2010-12-01 08:26:00 | 71053 | white metal lantern | 6 | 3.39 | 20.34 | 17850 | United Kingdom |
| 2 | 536365 | 2010-12-01 08:26:00 | 84406B | cream cupid hearts coat hanger | 8 | 2.75 | 22.00 | 17850 | United Kingdom |
| 3 | 536365 | 2010-12-01 08:26:00 | 84029G | knitted union flag hot water bottle | 6 | 3.39 | 20.34 | 17850 | United Kingdom |
| 4 | 536365 | 2010-12-01 08:26:00 | 84029E | red woolly hottie white heart | 6 | 3.39 | 20.34 | 17850 | United Kingdom |

### 3. Custom cleaning function

I realized that the NFX processing was not enough, so I implemented a custom function to further clean the Description field. After this step, I was able to clean an additional 207 descriptions, reducing the total number of unique descriptions to 3,647.

```python
import re
from collections import defaultdict

def process_data(df):
    # Combine all the words in one per StockCode
    combined = defaultdict(set) # use a set to avoid duplicates
    for _, row in df.iterrows():
        for word in row['Description'].split():
            combined[row['StockCode']].add(word)

    # Prepare new descriptions
    new_descriptions = {}
    for stock_code in combined.keys():
        description = " ".join(combined[stock_code])
        new_descriptions[stock_code] = description

    # Add necessary spaces
    for stock_code in new_descriptions.keys():
        new_descriptions[stock_code] = re.sub(r'(\w)([A-Z])', r'\1 \2', new_descriptions[stock_code])

    # Add space before and after a digit
    for stock_code in new_descriptions.keys():
        new_descriptions[stock_code] = re.sub(r'(\D)(\d)', r'\1 \2', new_descriptions[stock_code])
        new_descriptions[stock_code] = re.sub(r'(\d)(\D)', r'\1 \2', new_descriptions[stock_code])

    # Delete unnecessary spaces
    for stock_code in new_descriptions.keys():
        new_descriptions[stock_code] = ' '.join(new_descriptions[stock_code].split())

    # Update Description column in the original dataframe
    df['Description'] = df['StockCode'].map(new_descriptions)

    return df

beforeRegex_data = data.copy()

data = process_data(data)
data.head()
```

✓ 8.7s

| | InvoiceNo | InvoiceDate | StockCode | Description | Quantity | UnitPrice | Revenue | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 536365 | 2010-12-01 08:26:00 | 85123A | hanging heart cream white tlight holder | 6 | 2.55 | 15.30 | 17850 | United Kingdom |
| 1 | 536365 | 2010-12-01 08:26:00 | 71053 | white metal lantern moroccan | 6 | 3.39 | 20.34 | 17850 | United Kingdom |
| 2 | 536365 | 2010-12-01 08:26:00 | 84406B | hanger hearts cream coat cupid | 8 | 2.75 | 22.00 | 17850 | United Kingdom |
| 3 | 536365 | 2010-12-01 08:26:00 | 84029G | bottle water hot union flag knitted | 6 | 3.39 | 20.34 | 17850 | United Kingdom |
| 4 | 536365 | 2010-12-01 08:26:00 | 84029E | woolly hottie heart red white | 6 | 3.39 | 20.34 | 17850 | United Kingdom |

Let me explain how this function works:

- It first combines all the Descriptions per StockCode_fixed
- It then creates a new DataFrame with the combined descriptions
- The function then adds necessary spaces, for example changing vippassport cover to vip passport cover

- It does this by adding a space between any lower-case letter followed by an upper-case letter
- Then it adds a space before and after a digit
- Lastly, it removes any unnecessary spaces in the description

This may still sound a bit confusing, so here's a snippet of what the function does:

---

Combine all the words in one.
For example, from:

```
22199   frying pan red polkadot
22199   frying pan red retrospot
```

To:

```
22129 frying pan red polkadot retrospot
```

---

Put necessary spaces.
For example, from:

```
20622 vippassport cover
```

To:

```
20622 vip passport cover
```

---

Put space before and after a digit.
For example, from:

```
17107D  flower fairy5 drawer liners
```

To:

```
17107D flower fairy 5 drawer liners
```

---

Delete unnecessary spaces before and after each sentence (only 1 space between words).
For example, from:

```
21175   gin tonic diet   metal sign
```

To:

```
21175 gin tonic diet metal sign
```

---

However, I noticed a discrepancy between the number of unique StockCode values (3,684) and the number of unique Description values (3,647). This is problematic because each Description should map to a unique StockCode.

```
Total unique values for column StockCode: 3,684
Total unique values for column Description: 3,647
```

## StockCode column

### 1. Addressing **Description**-**StockCode** mismatch

There were 29 Description values linked with multiple StockCode values. I fixed this by assigning a primary StockCode to each Description. Now the StockCode and Description fields align perfectly, both having 3,647 unique values.

| | Description | StockCode |
|---|---|---|
| 0 | 4 cake fairy placemats set | 2 |
| 1 | bathroom metal sign | 2 |
| 2 | blue sucker round clock | 2 |
| 3 | cherry blossom square cabinet | 2 |
| 4 | columbian candle rectangle | 2 |
| 5 | columbian candle round | 3 |
| 6 | columbian cube candle | 2 |
| 7 | cover cushion french floral | 2 |
| 8 | cover cushion french lattice | 2 |
| 9 | cushion cake fairy pink cover | 2 |
| 10 | cushion cover paisley french | 2 |
| 11 | cushion sud du rose cover | 2 |
| 12 | daisy plastic tray retro | 2 |
| 13 | dinner metal served sign | 2 |
| 14 | drawer tidy office | 2 |
| 15 | flask decorative cherry blossom | 2 |
| 16 | heart tlight holder | 2 |
| 17 | jewelled photoframe de nile eau | 2 |
| 18 | key door fob | 2 |
| 19 | letter key bling ring | 8 |
| 20 | pencils tube colouring brown | 2 |
| 21 | pink flowers easter rabbit | 2 |
| 22 | pink frame photo jewelled | 2 |
| 23 | pink glass candleholder flock | 2 |
| 24 | polka plastic tray retro | 2 |
| 25 | signcupcake hook metal single | 3 |
| 26 | tray plastic 70 s retro | 2 |
| 27 | white frosted base | 2 |
| 28 | white ribs bamboo lampshade | 2 |

### 2. Reviewing alphanumeric **StockCode** values

I discovered that six StockCode values were character-based. After examining these cases, I decided to:

- Remove records associated with Manual and CRUK Commission as these do not directly relate to individual product sales
- Retain Discount since it directly affects product prices and sales
- Exclude DOTCOM POSTAGE and POSTAGE as these could distort the analysis
- Retain CUSHION MATCH PADS as it represents a standard product

## Country column

### 1. Country Name Correction

I made the following corrections for clarity:

- Renamed "EIRE" to "Ireland"
- Renamed "RSA" to "Republic of South Africa"
- Renamed "European Community" to "European Community (EEC, ECSC, and EAEC)"

These changes provide more context and allow for clearer analysis.

### 2. Country Correction for Specific Customer

I noticed that a particular customer was listed under two countries: Switzerland and Cyprus. Upon further examination, I found that the customer changed their country of residence 5 days after their last purchase. However, the majority of transactions were made while the customer resided in Switzerland. To avoid any potential issues in the SQL database construction, I updated the country for this customer to Switzerland.

## Date formats

I expanded the date information in the dataset for easier analysis.

I added columns for normalized invoice date, year, year-month, month, month-date, day of month, hour, day of week, month name, and day name.

| invoiceDate_norm | Year | YearMonth | Month | MonthDate | DayofMonth | Hour | DayofWeek | Month_name | Day_name |
|---|---|---|---|---|---|---|---|---|---|
| 2010-12-01 | 2010 | 2010-12 | 12 | 12-01 | 1 | 8 | 2 | December | Wednesday |
| 2010-12-01 | 2010 | 2010-12 | 12 | 12-01 | 1 | 9 | 2 | December | Wednesday |
| 2010-12-01 | 2010 | 2010-12 | 12 | 12-01 | 1 | 9 | 2 | December | Wednesday |
| 2010-12-01 | 2010 | 2010-12 | 12 | 12-01 | 1 | 10 | 2 | December | Wednesday |
| 2010-12-01 | 2010 | 2010-12 | 12 | 12-01 | 1 | 10 | 2 | December | Wednesday |

## Aggregating similar transactions

I found several instances of similar transactions differentiated only by the Quantity column. To remove redundancies and simplify the dataset, I aggregated these transactions. This aggregation was done based on InvoiceNo, InvoiceDate, StockCode, Description, UnitPrice, CustomerID, Country, and various date information. The resulting dataframe grouped transactions and provided summed quantities and revenues.

| InvoiceNo | InvoiceDate | StockCode | Description | Quantity | UnitPrice | Revenue | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|
| 537051 | 2010-12-05 11:12:00 | 21916 | retro white sticks chalk 12 set | 1 | 0.42 | 0.42 | 15708 | United Kingdom |
| 537051 | 2010-12-05 11:12:00 | 21916 | retro white sticks chalk 12 set | 4 | 0.42 | 1.68 | 15708 | United Kingdom |
| 537051 | 2010-12-05 11:12:00 | 22725 | alarm bakelike chocolate clock | 1 | 3.75 | 3.75 | 15708 | United Kingdom |
| 537051 | 2010-12-05 11:12:00 | 22725 | alarm bakelike chocolate clock | 2 | 3.75 | 7.50 | 15708 | United Kingdom |

| InvoiceNo | InvoiceDate | StockCode | Description | Quantity | UnitPrice | Revenue | CustomerID | Country |
|---|---|---|---|---|---|---|---|---|
| 537051 | 2010-12-05 11:12:00 | 21916 | retro white sticks chalk 12 set | 5 | 0.42 | 2.10 | 15708 | United Kingdom |
| 537051 | 2010-12-05 11:12:00 | 22725 | alarm bakelike chocolate clock | 3 | 3.75 | 11.25 | 15708 | United Kingdom |

## Separating Canceled Transactions

After preparing the data for analysis, I separated the canceled transactions (those with a quantity less than or equal to zero) from the successful transactions. I also converted the InvoiceNo data type from an object to an integer for the successful transactions.

The dataset before this operation contained 394,911 rows (after a few cleaning processes along the way), including 8,551 rows of returned items. After removing the returned items, the cleaned data contained 386,360 rows.

## Data check

After performing all the cleaning operations, I checked the data and confirmed:

- There were no missing values or duplicates
- All columns were of the correct data type
- The data contained a reasonable number of unique values for each column
- The shape of the data was consistent with my expectations

The cleaned data is well-prepared for the subsequent analysis and insights extraction.

In summary, my comprehensive data validation process involved detailed data investigation, data cleaning, and unique data handling decisions to ensure the accuracy and reliability of my subsequent analysis.

# DATABASE TYPE: NoSQL or SQL

---

## NoSQL (Not Only SQL)

NoSQL databases are non-tabular and store data differently than relational tables. NoSQL databases come in a variety of types based on their data model, such as key-value, document, columnar and graph. They are generally used when working with a massive amount of data where the data's nature does not require a relational model.

### Advantages of NoSQL:

- Scalability: NoSQL databases are designed to expand easily and can handle large amounts of data and high traffic loads in real time.
- Flexibility:They allow for flexible and dynamic schemas.
- Speed: NoSQL databases are typically faster at storing and retrieving data because their data models are simpler.

### Disadvantages of NoSQL:

- Lack of Standardization: Unlike SQL databases, NoSQL solutions do not have a common standard.
- Complexity: They can require a higher level of technical knowledge and understanding to operate and manage.

# DATABASE TYPE: NoSQL or SQL

---

## SQL (Structured Query Language)

SQL is a programming language used to communicate with and manipulate databases. It is used with relational databases, which organize data into one or more tables of columns and rows, with a unique key identifying each row. Generally, SQL databases are used when dealing with structured data and when data integrity is crucial.

### Advantages of SQL:

- ACID Compliance (Atomicity, Consistency, Isolation, Durability): SQL databases follow ACID properties which ensure reliable processing.
- Structured Data: SQL databases are highly efficient when handling structured data.
- Schema-based: All data stored follows a clear schema, so you always know the structure of your data.
- Complex Queries: SQL databases can handle complex queries and joins very efficiently.

### Disadvantages of SQL:

- Scalability: SQL databases can become inefficient as the amount of data grows, and they can be expensive to scale.
- Flexibility: Changes to the database schema can be slow and difficult, and must be carefully managed.

# DATABASE TYPE: NoSQL or SQL

Choosing a database

---

## Choosing the right type of database

In determining which database type is the right one, it depends heavily on the kind of data we're dealing with and what we aim to do with this data.

In our case, the data is highly structured and relational. We are dealing with customer transactions where the integrity of data is crucial. Additionally, we need to perform complex queries and aggregations, which is a strength of SQL databases. We aren't dealing with extremely high volumes of data that would necessitate the scalability advantages of NoSQL, and our schema is unlikely to change frequently, if at all.

Given these factors, an SQL database seems to be the most suitable choice for this particular dataset and problem.

# DATABASE DESIGN and NORMALIZATION

## Data preparation for SQL entities

---

The main goal of the database design process is to produce a system that will meet the user's information requirements. The most important task is to design the database structure (schemas) and the applications that are required. These tasks are often carried out in parallel.

The process of database design involves many steps, with each step further refining the structure of the database to ensure it efficiently stores and retrieves data while adhering to the defined business rules. One important aspect of database design is normalization.

Normalization is a database design approach that organizes data to minimize redundancy and dependency. It involves the construction of tables, the selection of appropriate primary keys, the establishment of relationships between tables, and the decision of what fields will be stored in each table type.

The following are the tables I prepared for my SQL database:

## Customer Information

This table contains details about the customer including their unique ID and country of residence. The client_id is my chosen primary key as it provides a unique identifier for each customer. The structure is as follows:

- client_id (Primary Key)
- CustomerID
- Country

## Product Information

This table includes details about each unique product. The product_id is my chosen primary key for this table. The structure is as follows:

- product_id (Primary Key)
- StockCode
- UnitPrice
- Description

## Invoice Information

This table details the transaction record. Each invoice has a unique invoice_id which is my chosen primary key for this table. The structure is as follows:

- invoice_id (Primary Key)
- InvoiceNo
- InvoiceDate
- invoiceDate_norm
- Year
- YearMonth
- Month
- MonthDate
- DayofMonth

- Hour
- DayofWeek
- Month_name
- Day_name

## Sales Information

This table links the other tables together and contains records of the quantity of products sold per invoice. It contains foreign keys referencing the primary keys of the other tables. The structure is as follows:

- invoice_id (*Foreign Key*: links to the Invoice Information table)
- product_id (*Foreign Key*: links to the Product Information table)
- client_id (*Foreign Key*: links to the Customer Information table)
- Quantity

The use of primary and foreign keys allows for efficient querying of the database and maintains the relationships between different pieces of data. This structured approach ensures data integrity and consistency.

Given the nature of the data, it's essential to keep the entities separated yet easily linkable through SQL queries. This design allows for a flexible database that can accommodate changes and growth in the data while preserving efficient access and manipulation of the data.

# FROM PYTHON TO MYSQL

Data importation to SQL database

---

The process of importing tables to MySQL using Python involves several steps. Here, I outline the steps I took to import the created tables to MySQL:

## Import necessary libraries

The first step is to import all the necessary libraries that I'll be using to interact with my MySQL database. Here I used `getpass`, `pymysql.cursors`, and `sqlalchemy`. `getpass` is used to securely input your MySQL password, `pymysql.cursors` is a pure-Python MySQL client library, and `sqlalchemy` is a SQL toolkit and Object-Relational Mapping (ORM) system for Python, providing a full suite of well-known enterprise-level persistence patterns.

```
import getpass
import pymysql.cursors
from sqlalchemy import create_engine
from sqlalchemy import text
```

## Create the Connection String

A connection string is used to define the parameters of the database connection, including the username, password, and the database to be connected. In this script, I used `getpass.getpass()` to input the MySQL root password, and this is added to the connection string. `sqlalchemy`'s `create_engine` function is used to create a connection engine to the MySQL database using the connection string.

```
# Prompt user to enter MySQL root password
sql_pass = getpass.getpass()

# Create connection string and engine to connect to MySQL database
connection_string = "mysql+pymysql://root:" + sql_pass + "@localhost:3306/TheFinal"
engine = create_engine(connection_string)
```

## Writing DataFrames to SQL

I used the to_sql method from the pandas DataFrame, which writes records stored in a DataFrame to a SQL database. The parameters include the name of the table, the connection engine, the schema, a condition for if the table exists, and a boolean to indicate whether or not to write the DataFrame index as a column.

```
# Write DataFrames to SQL
customer.to_sql("customer", engine, "TheFinal", if_exists="replace", index=False)
invoice.to_sql("invoice", engine, "TheFinal", if_exists="replace", index=False)
product.to_sql("product", engine, "TheFinal", if_exists="replace", index=False)
sales.to_sql("sales", engine, "TheFinal", if_exists="replace", index=False)
returns.to_sql("returns", engine, "TheFinal", if_exists="replace", index=False)
```

## Defining Primary Keys

After writing the DataFrames to SQL, I then defined the primary keys of each table. Primary keys are unique identifiers for each record in a table. In the provided code, I connected to the engine and executed raw SQL to add primary keys to each table.

```
# Define primary keys using raw SQL
with engine.connect() as conn:
    conn.execute("ALTER TABLE customer ADD PRIMARY KEY (client_id);")
    conn.execute("ALTER TABLE invoice ADD PRIMARY KEY (invoice_id);")
    conn.execute("ALTER TABLE product ADD PRIMARY KEY (product_id);")
    conn.execute("ALTER TABLE sales ADD PRIMARY KEY (sales_id);")
```

## Defining Auto-incremented Primary Keys

For the returns table, I added a column id which is auto-incremented and set as the primary key. Auto-increment allows a unique number to be generated whenever a new record is inserted into a table.

```
# Define auto-incremented primary keys using raw SQL
with engine.connect() as conn:
    conn.execute("ALTER TABLE returns ADD COLUMN id INT AUTO_INCREMENT PRIMARY KEY;")
```

## Moving Primary Keys to the First Column

This step is for aesthetic purposes and does not change the functionality of the database. The primary key id in the returns table is moved to the first column of the table using raw SQL.

```
# Move Primary Keys in the first column
with engine.connect() as conn:
    conn.execute("ALTER TABLE returns MODIFY id INT AUTO_INCREMENT FIRST;")
```

## Specifying Foreign Keys

In the sales table, I specified the foreign keys that reference the primary keys in the customer, invoice, and product tables. Foreign keys are fields in a table that match the primary key column of another table. They establish a link between data in two tables.

```
# Specify Foreign Keys in the sales table
with engine.connect() as conn:
    conn.execute("ALTER TABLE sales ADD FOREIGN KEY (client_id) REFERENCES customer(client_id);")
    conn.execute("ALTER TABLE sales ADD FOREIGN KEY (invoice_id) REFERENCES invoice(invoice_id);")
    conn.execute("ALTER TABLE sales ADD FOREIGN KEY (product_id) REFERENCES product(product_id);")
```

In conclusion, the Python snippet outlines an efficient way to prepare our data in Python using pandas, and then import the cleaned and structured data into MySQL for storage and further analysis. This makes Python a powerful tool for both data cleaning and database management.

# ENTITY RELATIONSHIP DIAGRAM (ERD)

How the tables are related

---

An Entity Relationship Diagram (ERD) is a graphical representation of the major entities within a system, along with the relationships between those entities. It is a useful tool for designing and understanding database structures. In the case of our MySQL database, we have four main entities: customer, invoice, product, and sales.

Here's how the entities are related based on the table structures:

## Customer (customer)

This entity represents the customer data, including client_id, CustomerID, and Country. The client_id serves as the primary key, which is a unique identifier for each customer.

## Product (product)

This entity contains the product information. It includes the product_id, StockCode, UnitPrice, and Description. Similar to the customer entity, product_id serves as the primary key for this entity.

## Invoice (invoice)

This entity represents the invoice information and contains multiple attributes including invoice_id, InvoiceNo, InvoiceDate, etc. The invoice_id is the primary key for this entity.

## Sales (sales)

This entity represents the sales information. The sales table includes invoice_id, product_id, client_id, and Quantity. The invoice_id, product_id, and client_id in this table are foreign keys that establish a link between this table and the invoice, product, and customer tables respectively.

# ENTITY RELATIONSHIP DIAGRAM (ERD)

ERD model

---

The relationships between these entities are as follows:

A one-to-many relationship exists between the customer table and the sales table. This is because a single customer can make multiple purchases, but each purchase is associated with only one customer. This relationship is represented by the client_id in both tables.

A one-to-many relationship also exists between the product table and the sales table. This is because a single product can appear on multiple sales, but each sale involves only one product. This relationship is represented by the product_id in both tables.

Finally, a one-to-many relationship exists between the invoice table and the sales table. This is because a single invoice can contain multiple sales (of different products), but each sale belongs to only one invoice. This relationship is represented by the invoice_id in both tables.

These relationships allow the database to link data from multiple tables, providing a more complete view of each sale. For example, by joining the sales, product, and customer tables, we could see what product a specific customer bought and at what price, all in a single query.

# MySQL Queries

---

## Top 5 Customers by Total Quantity

```
# Top 5 Customers by Total Quantity
SELECT c.client_id, c.CustomerID, SUM(s.Quantity) AS total_quantity
FROM customer c
INNER JOIN sales s ON c.client_id = s.client_id
GROUP BY c.client_id, c.CustomerID
ORDER BY total_quantity DESC
LIMIT 5;
```

The customer with client_id 847 and CustomerID 14646 has purchased the highest quantity of items, i.e., 197 420 units. The next top customers are with client_ids 2607, 65, 915, and 1105, who have purchased 80,997, 80 488, 77 669, and 74 215 units respectively.

## Top 5 Products by Total Sales

```
# Top 5 Products by Total Sales
SELECT p.product_id, p.Product, SUM(s.Quantity) AS total_sales
FROM product p
INNER JOIN sales s ON p.product_id = s.product_id
GROUP BY p.product_id, p.Product
ORDER BY total_sales DESC
LIMIT 5;
```

The highest sold product is the "birdie little craft paper" with a total sale of 80 995 units. Following this, the next top products are "medium ceramic storage jar", "small holder popcorn", and "designs world war asstd gliders 2" which has been included twice, likely due to data duplication or different product variants, with 76 087, 36 136, 27 432 and 23 904 units sold respectively.

## Sales per Month

```sql
# Sales per Month
SELECT i.YearMonth, SUM(s.Quantity) AS total_sales
FROM invoice i
INNER JOIN sales s ON i.invoice_id = s.invoice_id
GROUP BY i.YearMonth
ORDER BY i.YearMonth;
```

The sales have been relatively steady throughout the year 2010 and 2011, with a notable increase in sales during the months of September, October, and November 2011. The highest sales were recorded in November 2011, with 674 963 units sold.

## Top 5 Selling Products for the Most Frequent Customer

```sql
# Top 5 Selling Products for the Most Frequent Customer
SELECT s.client_id, p.Product, SUM(s.Quantity) AS total_quantity
FROM sales s
INNER JOIN product p ON s.product_id = p.product_id
WHERE s.client_id = (
    SELECT client_id
    FROM (
        SELECT client_id, SUM(Quantity) AS total_quantity
        FROM sales s
        GROUP BY client_id
        ORDER BY total_quantity DESC
        LIMIT 1
    ) AS most_freq_customer
)
GROUP BY s.client_id, p.Product
ORDER BY total_quantity DESC
LIMIT 5;
```

For the most frequent customer (client_id 847), the top bought products are "light rabbit night", "spaceboy lunch box", "cases retrospot cake 72 pack", "box lunch girl

dolly", and "snack round woodland set of 4 boxes" with quantities of 4 801, 4 492, 4 104, 4 096, and 3 120 respectively.

## Total Sales per Country

```
# Total Sales per Country
SELECT c.Country, SUM(s.Quantity) AS total_sales
FROM customer c
INNER JOIN sales s ON c.client_id = s.client_id
GROUP BY c.Country
ORDER BY total_sales DESC;
```

The country with the highest total sales is the United Kingdom, with 4 246 783 units sold. This is followed by the Netherlands, Ireland, Germany, and France, with total sales of 200 834, 140 381, 118 033, and 110 597 units respectively. The countries with the lowest sales are Bahrain and Saudi Arabia, with 260 and 80 units sold respectively. The sales data across countries can help us understand market penetration and popularity of different products in different regions.

# EDA AND DATA VISUALIZATIONS

## Summary

---

Exploratory Data Analysis (EDA) is a fundamental step in the data analysis process. It involves understanding the patterns, relationships, and structures within the data before formal modeling. EDA is an iterative process where new questions are generated based on the results from previous ones. The purpose is to explore, analyze, and make sense of the data to gain useful insights.

In the context of this project that involves online retail sales data, EDA will provide us a thorough understanding of various facets of the business. This may include customer behavior, product popularity, time-based sales trends, geographical distribution of sales, and more.

By employing various statistical and visual analysis tools, we aim to answer key business questions such as:

1. Who are our most valuable customers?
2. What are our best-selling products?
3. When are our peak sales periods?
4. Where are our customers located geographically?
5. What are the buying patterns and trends over time?

The insights drawn from EDA can lead to more effective business strategies. It can help inform inventory management, marketing efforts, customer relationship management, and overall business planning.

In the subsequent sections, we will perform EDA on our online retail dataset using Python. This will help us to uncover valuable insights and understand the relationships between different data entities. We will utilize visualization techniques such as bar graphs, line plots, heatmaps, etc to depict these relationships and patterns clearly. Through these visualizations, we will make complex data understandable, thus aiding in informed decision-making.

In essence, Exploratory Data Analysis will help us turn our raw data into meaningful information that can drive the company's growth and success.

# EDA AND DATA VISUALIZATIONS

Key points

---

The purpose of our Exploratory Data Analysis (EDA) is to address several business goals and objectives identified by the stakeholders. These goals revolve around understanding sales performance, customer behavior, market geography, temporal sales patterns, product strategy, pricing strategy, customer retention, and future sales predictions.

## Sales Performance Analysis

We will explore the overall trend in revenue over time, its variation across different days of the week or months, and identify top-selling products based on quantity or revenue. The visualizations planned for this include time series plots, heatmaps, and bar charts representing top-selling products.

## Customer Behavior Analysis

Our focus will be on customers that contribute most to the revenue, those who buy most frequently, and those who purchase the most items. We will also analyze when customers make most of their purchases and the most common quantities of items per purchase. Bar charts, heatmaps, and histograms will aid in visualizing these aspects.

## Geographic Market Analysis

This part of the analysis will understand how sales and revenue vary across different countries, identify specific products or categories popular in certain countries, and determine the geographic distribution of customers. Choropleth maps, bar charts, and geographic heat maps will be key visualization tools for this segment.

## Temporal Sales Patterns

We will investigate the most profitable time of day, the most profitable day of the week, seasonal patterns or trends in sales or revenue, and how these sales trends

change over the year. Line plots, bar charts, decomposition plots, and heatmaps will be the primary tools for representing these patterns.

## Product Strategy

Here, we will determine the products that generate the most revenue, the distribution of quantity sold for different products, and uncover patterns in types of items bought together. Bar charts, histograms, and association rules visualization will help display this information.

## Pricing Strategy

This objective involves understanding the relationship between unit price and quantity sold, and the distribution of revenue per transaction. Scatterplots, histograms, and boxplots will provide an effective means to visualize this data.

## Customer Retention Strategy

Lastly, we will monitor the number of purchases customers make over a certain period. This will be visualized with bar charts or line plots over time.

# EDA AND DATA VISUALIZATIONS

## Business Goals and Questions

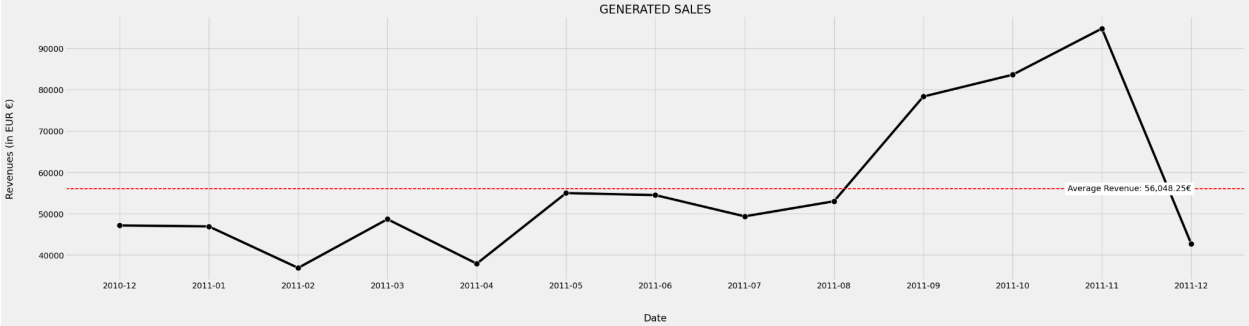
---

Here is the general overview of what's included in this report:

| Business Goals/Objectives | Questions to Explore |
|---|---|
| Sales performance analysis | 1. What is the overall trend in revenue over time?<br>2. How does revenue vary across different days of the week or months?<br>3. Which are the top-selling products based on quantity sold or sales? |
| Customer behavior analysis | 1. Which customers contribute the most to the revenue?<br>2. Which customers buy most frequently?<br>3. Which customers bought the most items?<br>4. When do customers make most of their purchases?<br>5. What are the most common quantities of items per purchase? |
| Geographic market analysis | 1. How does sales and revenue vary across different countries?<br>2. Are there any specific products or categories that are popular in certain<br>countries?<br>3. What is the geographic distribution of customers? |
| Temporal sales patterns | 1. What is the most profitable time of day?<br>2. What is the most profitable day of the week?<br>3. Are there any seasonal patterns or trends in sales or revenue?<br>4. How do sales trends change over the year?<br>5. How does the hour of the day influence the quantity sold?<br>6. How does the day of the week influence the quantity sold? |
| Product strategy | 1. What products generate the most revenue?<br>2. What is the distribution of quantity sold for different products?<br>3. Are there any patterns in the types of items bought together?<br>4. Which product pairings are most frequently bought together? |
| Pricing strategy | 1. How does unit price relate to quantity sold?<br>2. What is the distribution of revenue per transaction? |
| Customer retention strategy | 1. How many purchases do customers make over a certain period? (*optional*) |

## Sales performance analysis

1. What is the overall trend in revenue over time?




### Monthly Wholesale Revenue Trend Analysis: 2010 - 2011

The data presents the monthly wholesale revenue figures for a UK-based non-store retailer specializing in all-occasion gifts, for the period from December 2010 to December 2011.

The beginning of 2011 experienced relatively modest revenues, with a significant drop in February. However, starting from May, a marked increase was observed, and this trend continued until November.

The strong performance from May to November suggests that this period might coincide with a peak buying season for wholesalers, potentially related to the preparation for major holidays or events.

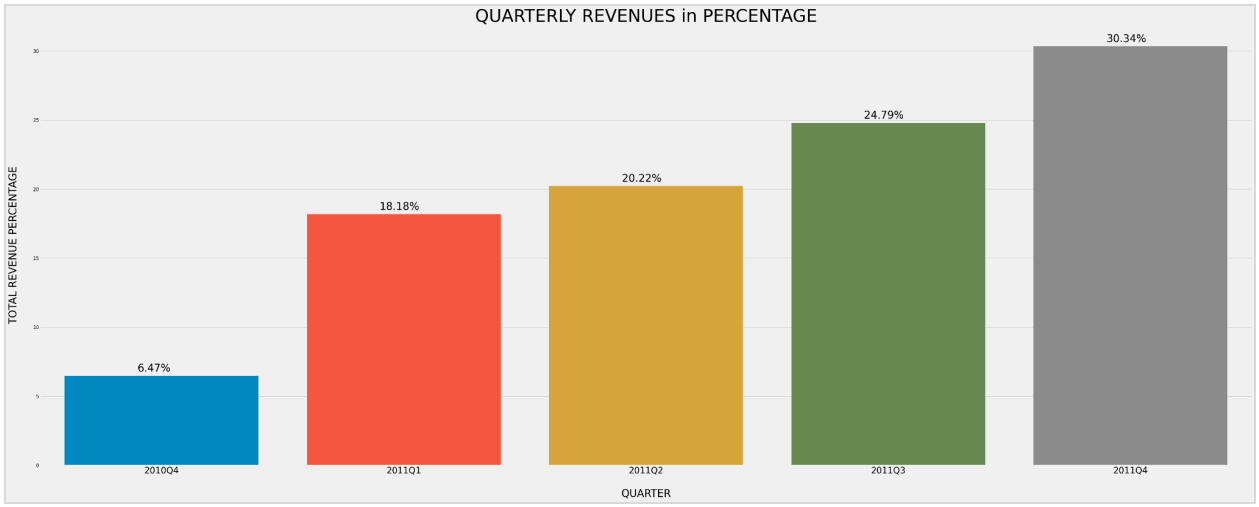
Interestingly, there was a steep decline in revenue in December 2011. This could be due to a decrease in demand from wholesalers post their peak buying season.

Another possible reason might be that the wholesalers have already stocked up for the holiday season in the previous months.

### *Key insights*

Understanding the buying patterns of the wholesalers is crucial for inventory management and demand forecasting.

The significant growth from May to November followed by the sudden drop in December are notable patterns that need to be taken into account for future business planning.




## Quarterly Revenue and Transaction Analysis: 2010 - 2011

The provided data encapsulates the quarterly revenue, transaction count, and their respective changes for the UK-based non-store retail business that sells all-occasion gifts to wholesalers.

Starting with Q4 of 2010, the revenue was £565,764.56, which is considered the base for subsequent comparisons. The first quarter of 2011 (Q1) witnessed a significant rise in revenue by almost 181%, to £1,589,235.35, paired with a growth in transaction count to 66,581. The growth trend continued into Q2 and Q3 of 2011, albeit at a slower pace.

Q2 revenue grew by 11.27% to £1,768,299.44, while Q3 saw a more substantial increase of 22.58%, taking the revenue to £2,167,650.80.

Q4 of 2011 recorded the highest revenue and transaction count of the entire period.

The revenue increased by another 22.37% from the previous quarter to £2,652,576.64, and the transaction count reached 126,128.

### *Key insights*

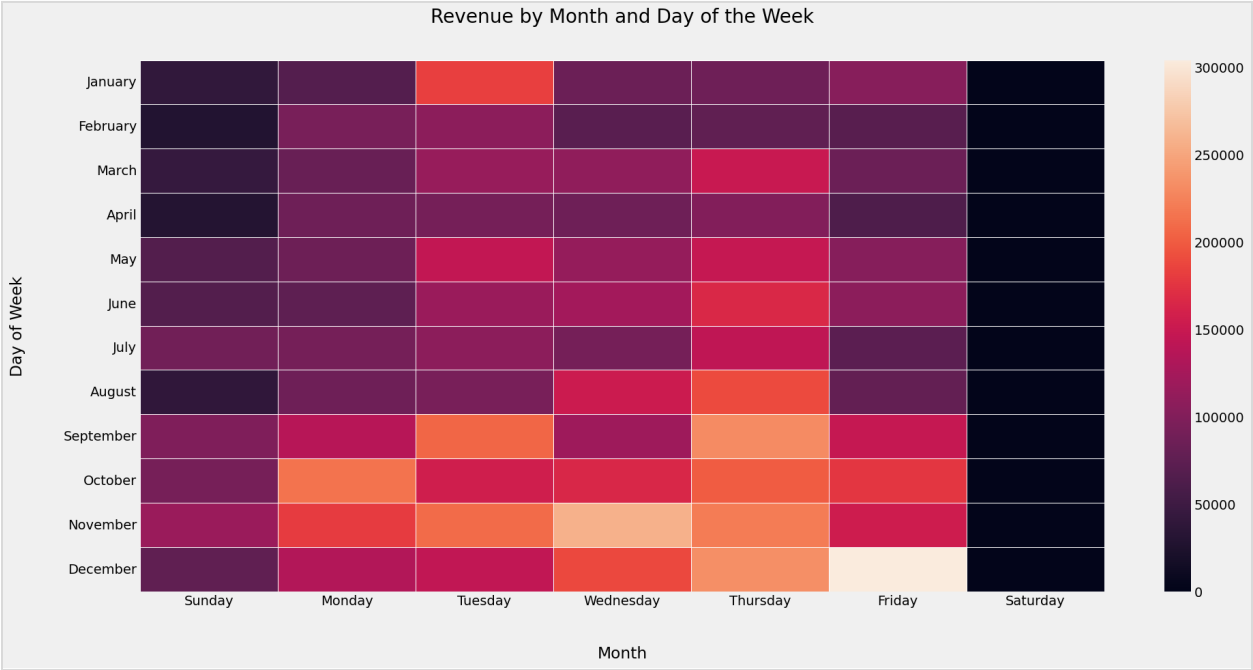
Over the course of 2011, both revenue and transaction count demonstrated a steady growth quarter over quarter. This indicates an overall successful period for the business with consistent expansion.

Q1 2011 saw the most substantial quarterly growth in revenue, implying effective business strategies or market conditions that led to this surge. Identifying these contributing factors could be instrumental for future planning.

Despite slower revenue growth in Q2, the business picked up its pace in Q3 and Q4, showcasing its resilience and adaptability.

The highest percentage of total revenue, total transaction count, and total quantity was achieved in Q4, which could be related to the seasonal buying habits of the wholesalers, particularly for the holiday season. Hence, this trend should be factored into future business and inventory planning.

## 2. How does revenue vary across different days of the week or months?




### Analysis of Revenue Variation Across Different Days and Months

The provided dataset presents a detailed distribution of the revenue across different days of the week and months. This allows us to understand how the business's sales performance fluctuates over time, revealing several notable patterns and potential opportunities for targeted strategies.

### *Key insights*

#### Strong Sales During the Week

The company experiences solid sales from Monday to Friday, with a substantial increase in revenue in the middle of the week, particularly on Wednesdays, Thursdays, and Fridays.

This trend suggests that promotional activities and customer engagement initiatives might be most effective if concentrated on these days.

#### Saturday Lull

It is noteworthy that no sales are recorded on Saturdays across all months, suggesting that this non-store retail business may not be operating on this day.

Significant Seasonality
The business seems to experience a strong seasonal pattern. Specifically, the revenue significantly increases from September to December, peaking in December.

This period likely includes key sales events, such as Black Friday, Cyber Monday, and the holiday shopping season, which typically result in a surge in sales.

Holiday Sales Boost
The revenue increase in December is particularly striking, with revenue on Fridays in December exceeding 300,000. This aligns with the holiday shopping trend, where consumers tend to make more purchases.

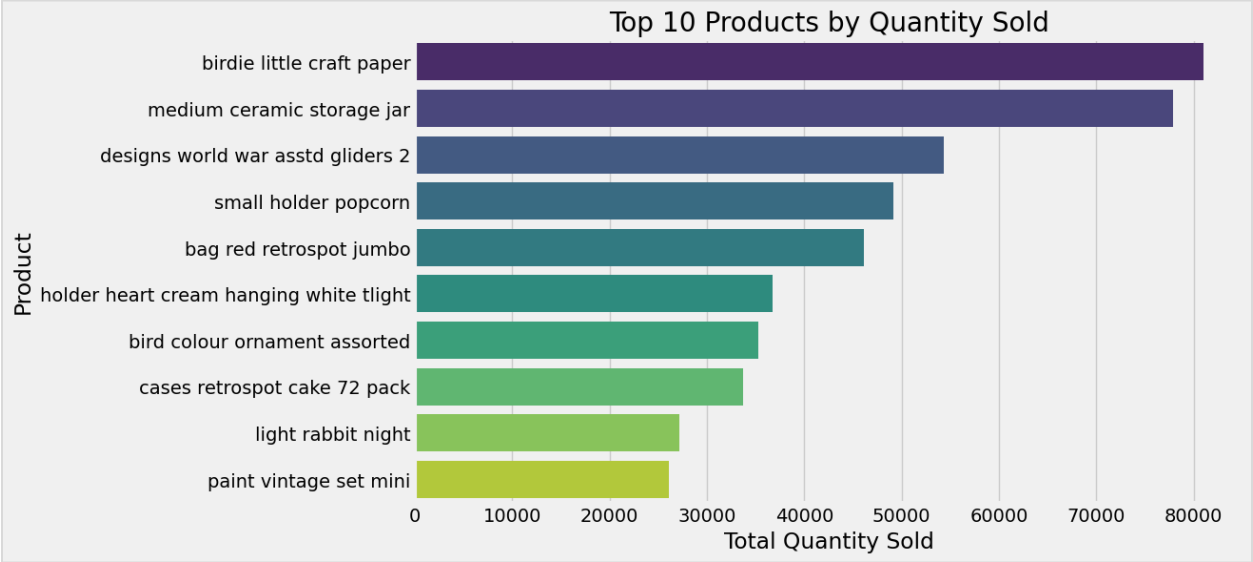
The company could capitalize on this by offering holiday promotions or launching new products during this period.

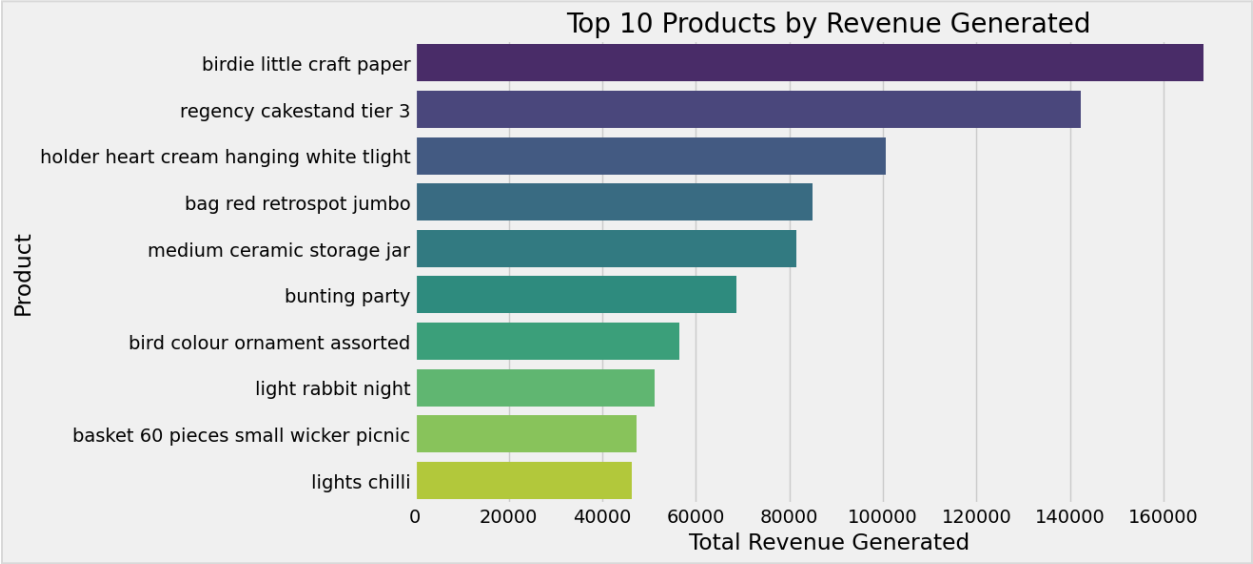
Mid-Year Dip
There is a relative dip in revenue during the middle of the year, specifically from April to July.

This could be a period when sales naturally slow down after the start-of-year enthusiasm and before the holiday season rush. The company could consider introducing special deals or marketing campaigns during this period to boost sales.

By understanding how sales fluctuate over time, the business can optimize its sales strategies, aligning them with customer buying patterns to maximize revenue. This could include tailoring marketing initiatives to the times when customers are most likely to make purchases, or exploring opportunities to boost sales during quieter periods.

3. Which are the top-selling products based on quantity sold or sales?




## Top Selling Products: Quantity vs Revenue

The data provided showcases the top-selling products based on the quantity sold and the revenue generated.

When looking at the top-selling products based on quantity, the "little paper birdie craft" takes the lead with 80,995 units sold, followed by the "medium ceramic jar storage" with 77,916 units.

This is interesting because despite having a slightly lower quantity sold, the "medium ceramic jar storage" does not match the revenue generated by the "little paper birdie craft." This might be due to a difference in unit prices.

In the revenue rankings, the "little paper birdie craft" continues to hold the top spot, generating £168,469.60. The "regency 3 tier cakestand," despite only selling 12,384 units (considerably lower in quantity), generated the second-highest revenue of £142,264.75.

This implies that the "regency 3 tier cakestand" has a significantly higher unit price compared to most other products.

*Key insights*

The "little paper birdie craft" is a star product for the business, selling the most units and generating the most revenue. Its continued success could be dependent on product quality, price point, and demand consistency among the wholesalers.

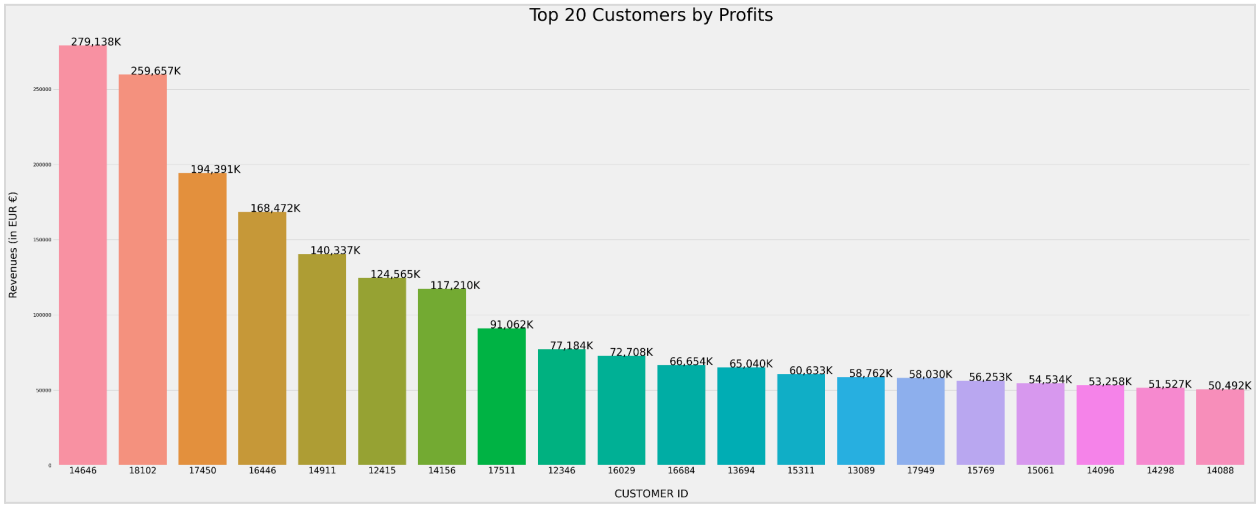
The "regency 3 tier cakestand" ranks second in terms of revenue but is not among the top 10 for quantity sold. Its high revenue contribution despite lower sales volume suggests a high-profit-margin product.

Some products like "medium ceramic jar storage" and "red bag retrospot jumbo" have high sales volumes but do not generate equivalent high revenue, indicating lower-priced items. This might appeal to a different segment of wholesalers who are more price-sensitive.

Maintaining a mix of high-volume, low-margin items (like "medium ceramic jar storage") and lower-volume, high-margin items (like "regency 3 tier cakestand") could be a strategic approach to cater to the diverse needs of the wholesaler market.

## Customer behavior analysis

### 1. Which customers contribute the most to the revenue?



## Top Revenue Contributing Customers Analysis

The analysis of our sales data reveals important insights about customers contributing the most to the overall revenue. We have the following top 20 customers from various countries who have contributed the highest revenue to our company (*cf figure above*)

### *Key insights*

The majority of our top revenue contributors are based in the United Kingdom, reflecting the strong customer base in this location.

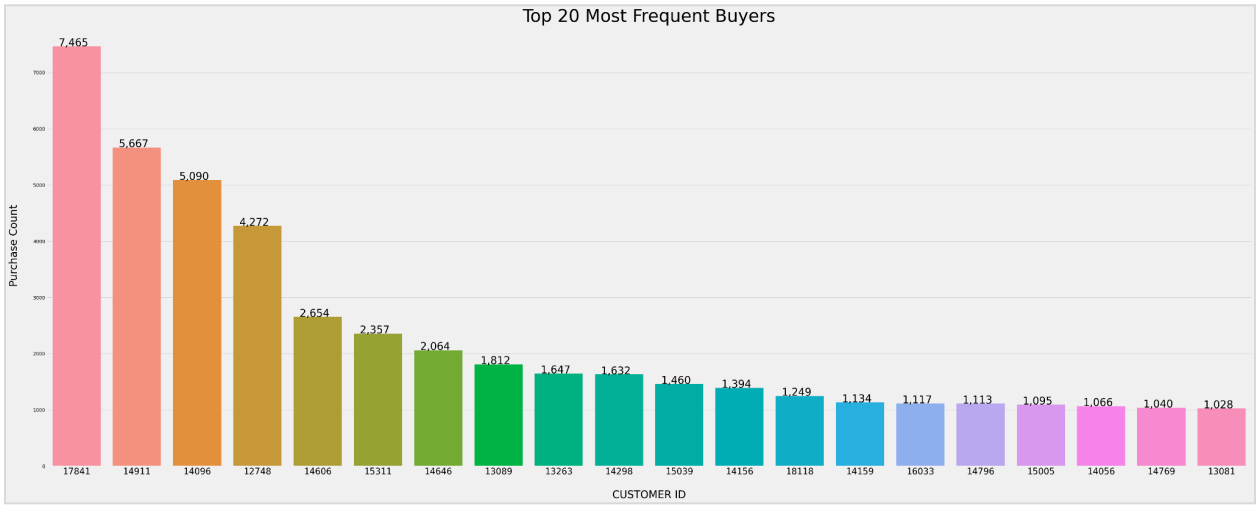
Customer 14646 from the Netherlands tops the list with a substantial total revenue of 279,138.02, demonstrating the presence of high-value customers outside of the UK as well.

There are significant revenue contributions from customers in Ireland and Australia, indicating potential markets for further exploration and expansion.

The revenue distribution among the top 20 customers varies from approximately 50,491.81 to 279,138.02, which signifies the need for different customer engagement strategies based on their contribution.

The diversity of these top revenue contributors suggests the importance of international sales and the need to maintain a diverse and globally distributed customer base.

## 2. Which customers buy most frequently?




## Analysis of Most Frequent Buyers

This analysis highlights the customers who make the most purchases from the business. It reveals some interesting patterns and insights about the business's most frequent buyers.

*Key insights*

The top purchaser, customer 17841, made a staggering 7465 purchases, significantly higher than the rest of the customers. This customer alone could be contributing a considerable chunk of the business's regular sales volume.

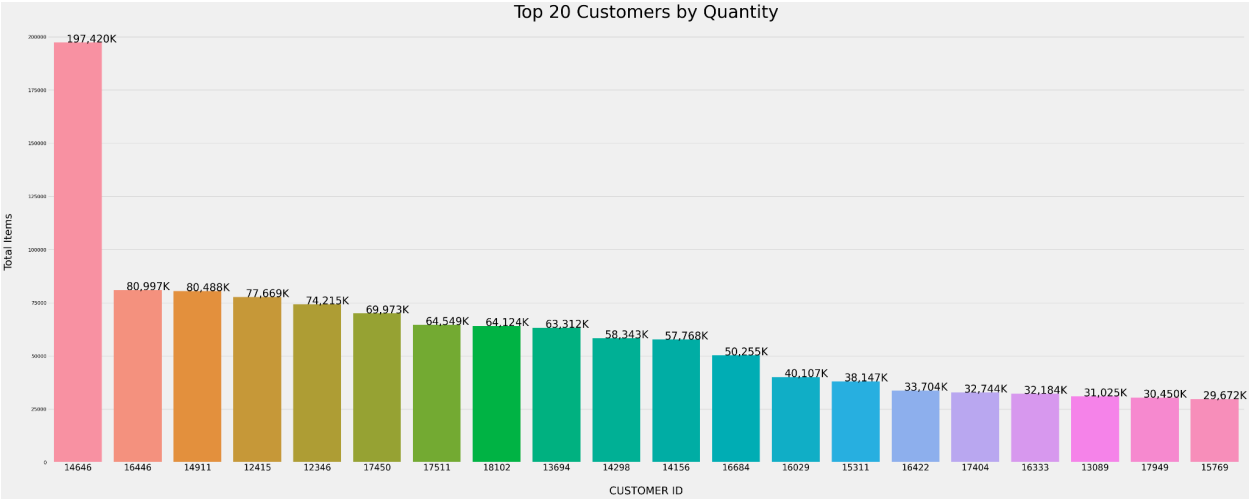
The second most frequent purchaser, customer 14911, made 5667 purchases. This is about 76% of the purchases made by the top customer. The significant drop between the first and second buyers indicates that the top customer is an outlier, with an exceptionally high number of purchases.

While customer 14096 made the third-highest number of purchases (5090), there is a gradual decrease in the number of purchases as we move down the list, implying a diverse range of purchasing behaviors among customers.

Of the 20 customers listed, customer 13081, who ranks 20th, made 1028 purchases, which is only about 14% of the number of purchases made by the top customer. This wide range suggests that the business has a broad customer base with different buying frequencies.

This analysis highlights the importance of frequent buyers for maintaining a steady flow of sales. The business may benefit from loyalty programs or special promotions targeted at these customers to sustain their purchasing frequency and potentially stimulate increased spending. It's also worthwhile exploring strategies to motivate less frequent buyers to purchase more often, thereby boosting the overall sales volume.

## 3. Which customers bought the most items?




## Analysis of Total Quantity Purchased by Customers

The table provides a comprehensive look at the customers who have purchased the highest total quantities from the business. The geographical distribution of these top customers

highlights the business's international reach, while the quantity of purchases offers insights into customer demand and potential inventory considerations.

*Key insights*

Customer 14646 from the Netherlands leads the pack with a total of 197,420 units purchased, exhibiting a tremendous buying power. This is significant considering the business's wholesaling nature.

The United Kingdom seems to be a crucial market for the business, with a significant number of top buyers originating from this country. Among the top 20 customers, 14 are from the UK.

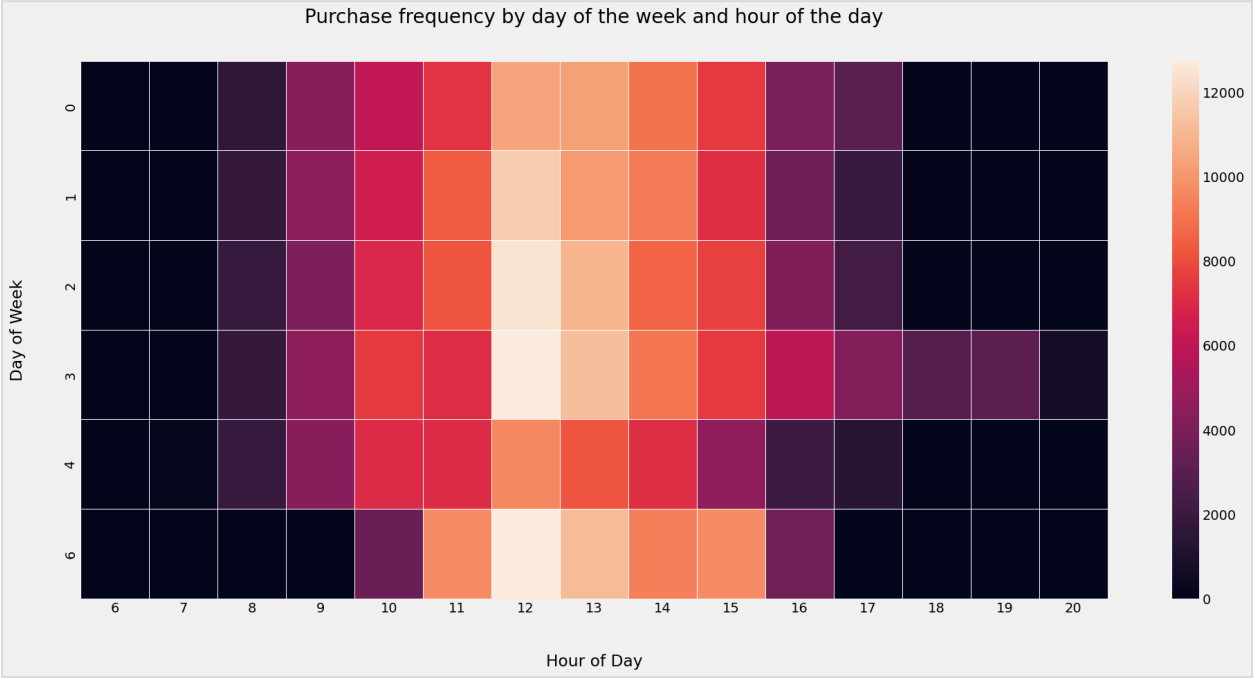
The second-highest quantity of purchases (80,997 units) is attributed to customer 16446 from the United Kingdom. However, this is less than half of the total quantity purchased by the top customer. This substantial difference underscores the unique buying power of customer 14646.

Interestingly, the top 20 list also includes customers from Ireland, Australia, and Sweden, suggesting the business has a diverse international customer base.

The 20th customer on the list, customer 15769 from the United Kingdom, has purchased 29,672 units, which is about 15% of what the top customer has bought. This indicates a wide range of buying quantities among the top customers.

From these insights, it's clear that the business's sales are heavily influenced by high-volume buyers. Targeted marketing and customer service efforts for these customers could enhance their loyalty and potentially increase sales. Additionally, understanding the buying patterns of these high-volume customers could inform inventory management and supply chain decisions.

## 4. When do customers make most of their purchases?




Customer Purchase Timing Analysis (Daily and Hourly Sales)

The provided data demonstrates the hourly distribution of purchases made by the customers throughout the week. It allows us to identify peak hours and days for customer purchases.

On weekdays, from Monday (0) to Thursday (3), the number of purchases starts to grow at 8 am. The majority of transactions occur between 10 am and 3 pm, with a peak at noon and early afternoon hours (11 am to 2 pm) and starts to decline after 2 pm.

This pattern aligns with typical business hours and likely reflects when customers have time to place orders during the workday.

It's apparent that customers tend not to make purchases early in the morning (6-7 am) and late in the evening (6 pm onwards), with almost no activity recorded during these periods on most days.

Interestingly, Thursday sees sustained activity until 5 pm, with an unusual surge of purchases in the evening hours (6-8 pm), the only day to show such a trend.

Fridays also have a significant number of purchases but see a faster drop-off in the afternoon compared to other weekdays.

It's also important to note that Saturdays see no activity, suggesting that this non-store retail business may not be operating on this day.

Sunday (6), though not a typical business day, also sees a substantial number of transactions, mainly taking place from 10 am to 3 pm. There's no activity before 9 am and after 4 pm on Sunday.

*Key insights*

Most transactions occur during typical business hours (10 am to 3 pm). It's the most active time for customer purchases, Indicating this is the optimal time to ensure customer service resources are available, and efforts should be made to ensure optimal service and availability during this period.

The unusual surge in purchases on Thursday evenings indicates that customers might be preparing for the weekend. The extended activity on Thursdays may represent an opportunity to promote products or offer specific deals.

The quick drop-off of activity on Fridays could indicate the start of a slower weekend period, and strategies could be developed to incentivize Friday afternoon shopping.

The peak transaction time (11 am to 2 pm) could be used for targeted promotions or special deals to drive further sales.

The low activity in the early morning and late evening can be taken into consideration for business hours or staff scheduling.

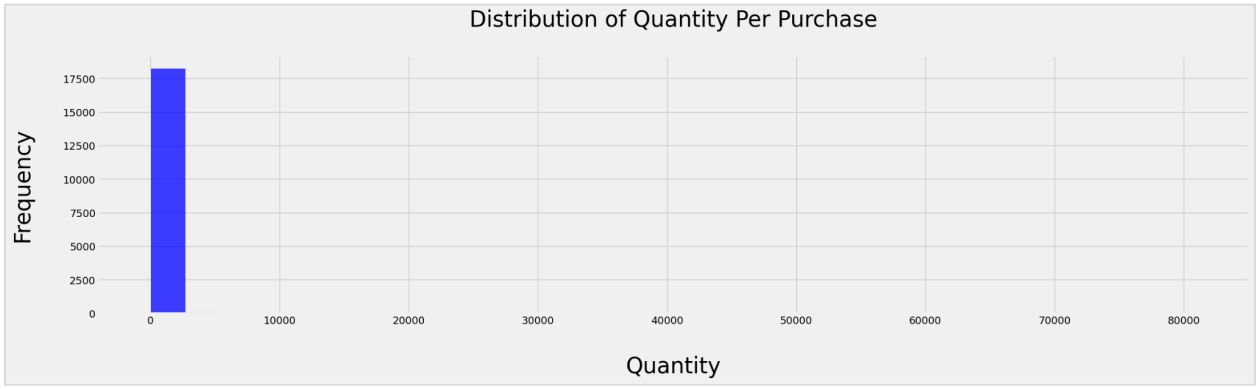
Despite being a weekend day, Sunday has a similar pattern to weekdays but with no transactions recorded after 4 pm.

This indicates an opportunity to extend business hours or run special promotions to encourage more transactions during these times.

The dataset does not contain any data for Saturday, so no observations could be made for this day.

5. What are the most common quantities of items per purchase?

## Understanding Purchase Quantities: The Impact of Outliers

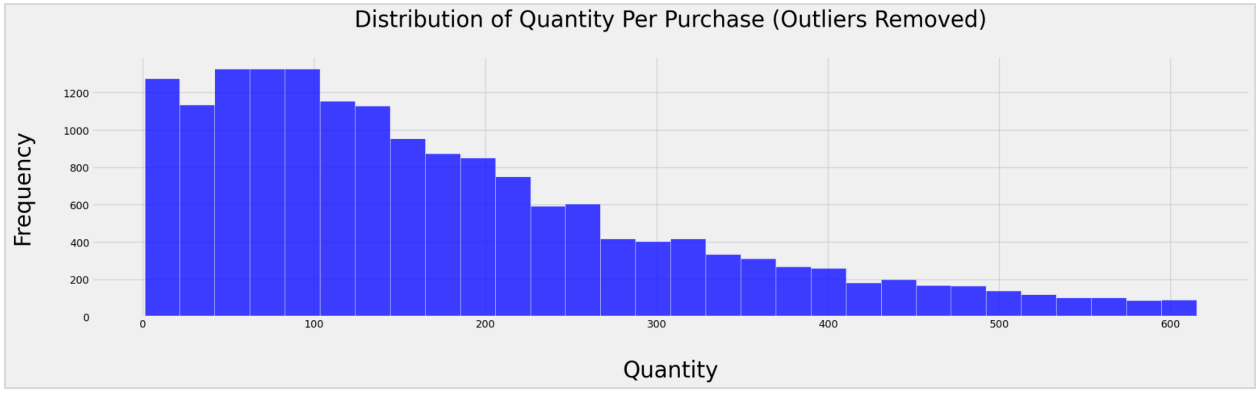



### The Influence of Outliers on Purchase Quantities:

In the unfiltered data, we're looking at a wide range of purchase quantities - from a single item to a whopping 80,995 items in one order. This vast range gives us an average order size of around 280 items. However, this average is significantly influenced by those rare but extremely large orders, making it not fully representative of a typical order size. What's more, the median order size (the 50% mark), is just 156 items, nearly half the average value, demonstrating the considerable impact of those large orders on the overall average.

### *Key insights*

While large orders are definitely part of the business, they don't represent the norm. Most orders tend to be much smaller, with half of them having 156 items or less.




### Purchase Quantities: A Look Without Outliers

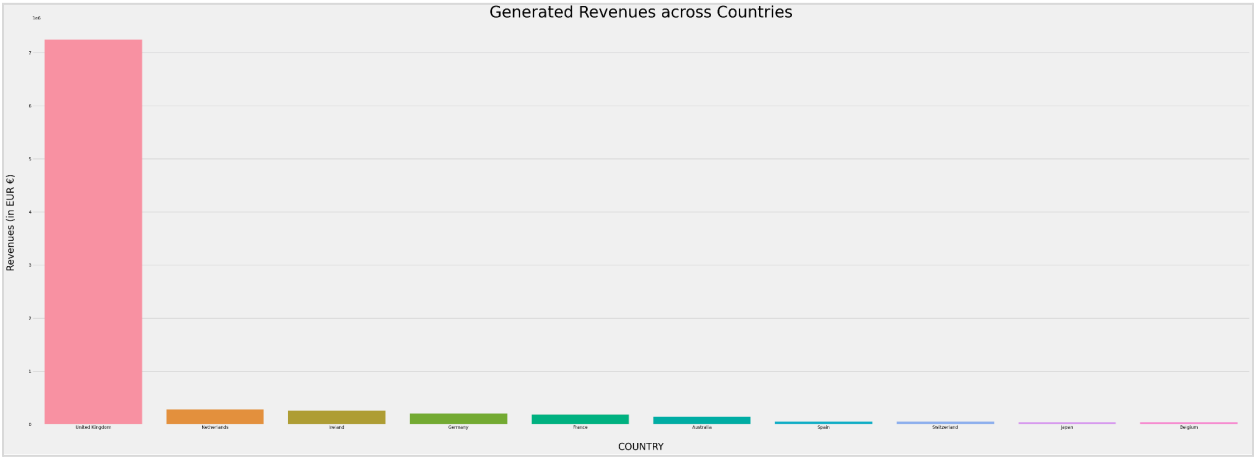
When we remove the outliers from the data, we get a more balanced picture of the typical order size. With outliers removed, our average order size drops significantly to 176 items. This trimmed down number aligns better with the median order size of 143 items, indicating a more consistent order size across the customer base.

*Key insights*

The average order size, once extreme cases are removed, is 176 items. This is a more accurate representation of the company's regular transaction size.

## Geographic market analysis

### 1. How does sales and revenue vary across different countries?




Revenue Variation Across Countries

The provided data demonstrates how the revenue of the UK-based non-store retail business varies across different countries. The United Kingdom is the biggest market for the company, contributing a substantial £7,243,943.47 to the revenue.

The Netherlands comes second, but with a significantly smaller contribution of £283,889.34. Ireland closely follows the Netherlands, with total revenue of £261,888.12.

Other European countries, such as Germany, France, Spain, and Switzerland, also make substantial contributions, with revenues ranging from £52,834.65 to £205,381.15. Australia is the major contributor outside of Europe, bringing in £138,103.81 in revenue.

Interestingly, the revenue from the USA and Canada is relatively low, with respective revenues of £3580.39 and £3115.44.

The revenue from other countries, such as Brazil, the Republic of South Africa, the Czech Republic, Bahrain, and Saudi Arabia, is even lower, each contributing less than £2000.

*Key insights*

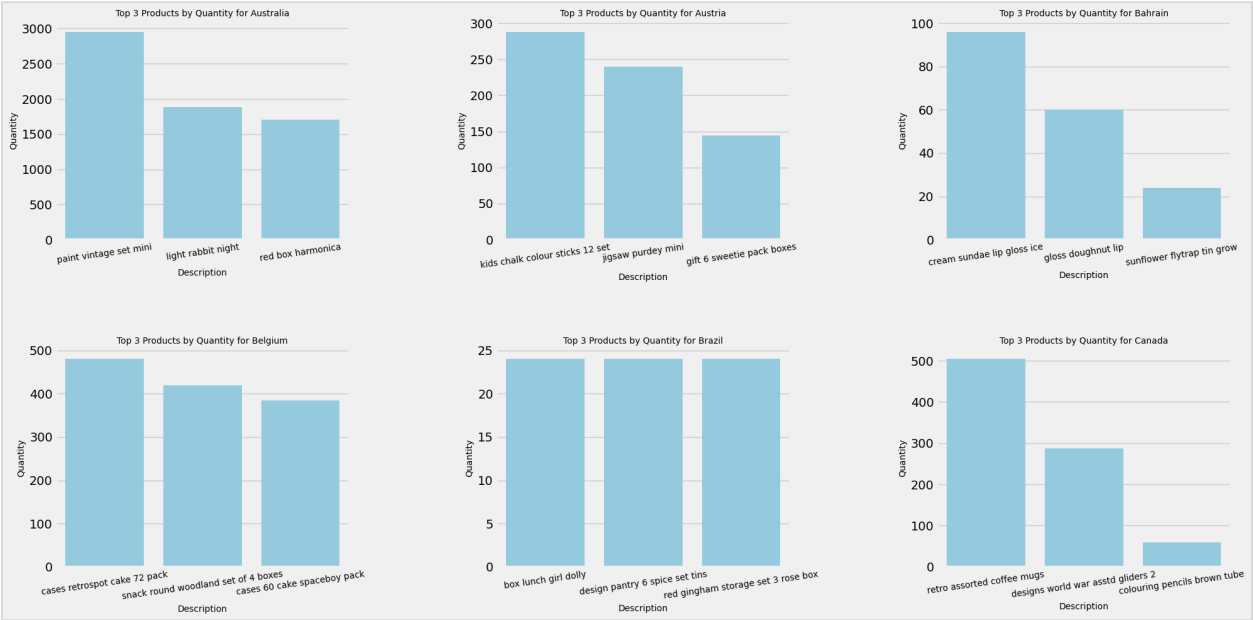
The UK is by far the biggest market for the company. Maintaining and strengthening these relationships is crucial for business sustainability.

Europe, in general, contributes a significant portion of the total revenue, with the Netherlands, Ireland, Germany, and France being the top contributors after the UK. It may be beneficial to explore opportunities for further growth within these markets.

Despite being large economies, the revenue from the USA and Canada is relatively low. Understanding the reasons behind this could open new avenues for expansion and revenue growth.

A number of countries generate very low revenues. It may be worth investigating if continued operations in these regions are cost-effective, or if the business should focus on higher-performing markets.

## 2. Are there any specific products or categories that are popular in certain countries?




## Analysis of Most Popular Products by Country

The provided table lists the most popular products for each country based on the quantity sold. This reveals some intriguing trends about customer preferences in different countries and presents opportunities for targeted marketing strategies.

### *Key insights*

The 'paper craft little birdie' is the most popular item in the United Kingdom, with a staggering 80,995 units sold. Given the significant customer base in the UK, as evident from earlier reports, this item's popularity suggests that it could be a major

revenue driver for the business.

Interestingly, 'night rabbit light' appears to be a universal favorite, being the most popular product in France, Japan, and the Netherlands. This could signify a broad, cross-cultural appeal of this product.

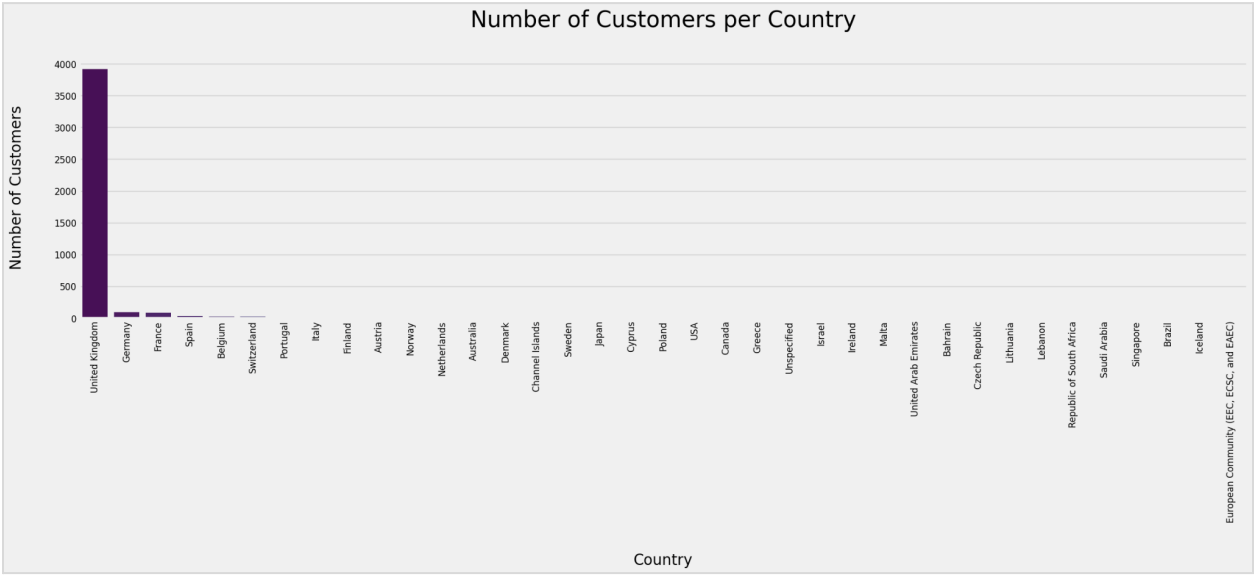
The 'set vintage mini paint' is another product that shows cross-border appeal, being the most popular in both Australia and Sweden.

Items such as 'set polkadot cutlery childrens pink piece 3', '72 retrospot pack cake cases', and 'assorted mugs coffee retro' seem to be popular kitchen and dining items in Finland, Belgium, and Canada, respectively, indicating a demand for these types of products in these markets.

Some countries, such as the Republic of South Africa and Saudi Arabia, show lower quantity sales for their top products. This could point to untapped potential in these markets or may reflect a smaller customer base in these regions.

By identifying the most popular products in each country, the business can create customized marketing strategies to promote these products further and introduce complementary items. Additionally, these insights could guide product development efforts, allowing the company to create products that resonate with customers in different markets.

## 3. What is the geographic distribution of customers?




## Geographical Distribution of Customers

The vast majority of the customers for the gift wholesaler are based in the United Kingdom, with 3,918 customers. This is not surprising given the company's location. Other notable customer bases exist in Germany and France, with 94 and 87 customers

respectively. Beyond these countries, the customer numbers are significantly smaller, ranging from 30 customers in Spain to just one in several countries, including Singapore, Saudi Arabia, Brazil, and Iceland among others.

However, the presence of customers in a diverse array of countries, from different continents and cultural backgrounds, suggests that the company's gift products have a wide appeal. It also implies the potential for growth in these international markets.

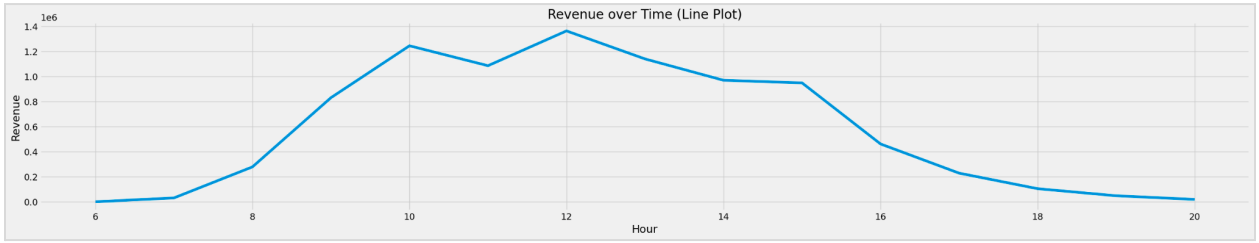
*Key insights*

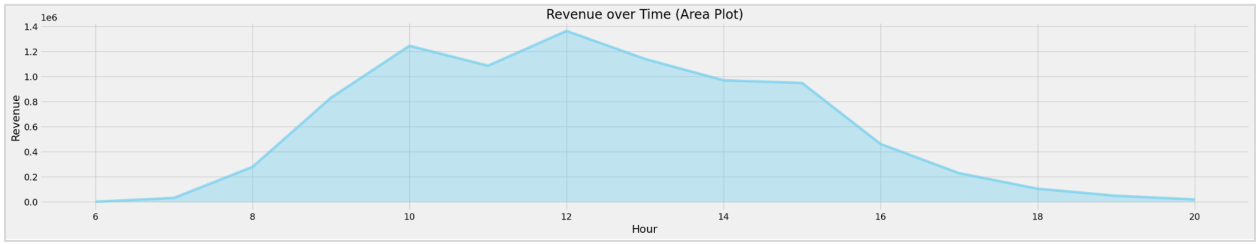
The UK is the core market for the business, followed by Germany and France. However, there is a broad distribution of customers around the world, indicating the international appeal of the company's product offerings. This geographical diversity could be

seen as an opportunity for targeted marketing efforts to grow the customer base in these international markets.

## Temporal sales pattern

### 1. What is the most profitable time of day?




Revenue Generation by Hour of the Day

The provided data illustrates the total revenue generated by sales at different hours of the day. This analysis allows us to identify the most and least profitable times of the day.

Peak Revenue Time: The highest revenue is generated between 12 pm (noon) and 1 pm with a total of £1,362,100.82. This is typically lunch hour for many people, and they might be more inclined to make purchases during this time.

High Revenue Period: The period from 9 am to 3 pm consistently generates high revenue. This can be considered the core operating period for the business.

Early Morning and Late Night Sales: Sales made early in the morning (6 am) and late at night (8 pm) generate the least revenue. This could be due to fewer customers shopping during these hours.

Decline in Revenue After Peak: There's a consistent decline in revenue generation after 1 pm, dropping significantly after 4 pm. This could be attributed to the end of the typical working hours, during which customers might be less likely to shop.

*Key insights*

Optimal Staffing and Operations: The business should ensure optimal staffing and smooth operations, especially during peak revenue hours (around noon) and high revenue periods (9 am to 3 pm) to maximize sales potential.

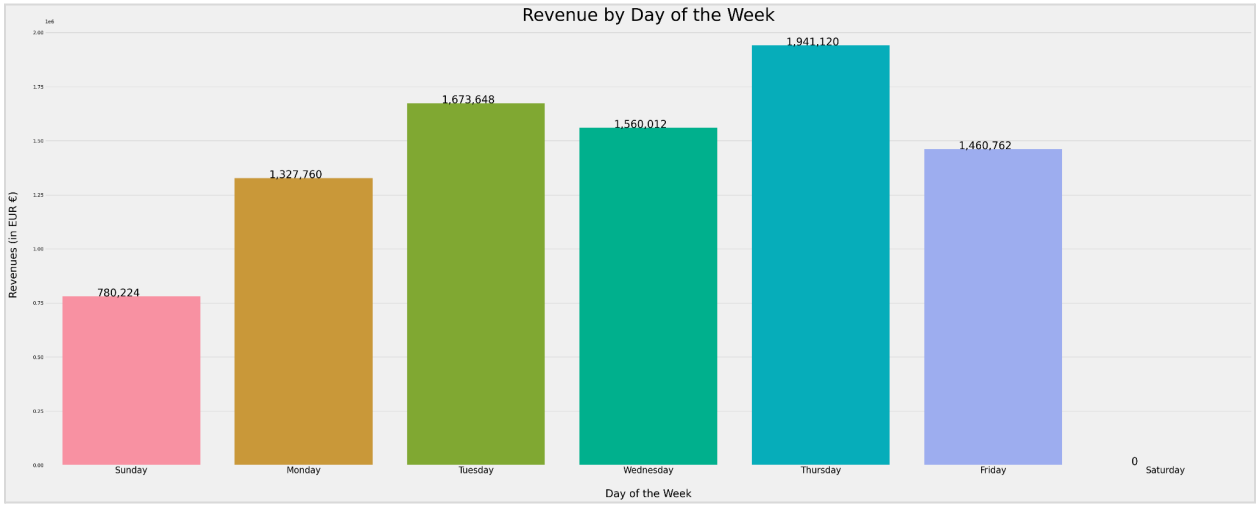
Marketing and Promotions Timing: Marketing activities and promotions could be timed to coincide with high revenue periods to increase their effectiveness.

Customer Behavior Insights: The low revenue early in the morning and late in the evening could be attributed to customer behavior. Further analysis on customer demographics and preferences could provide more insights on this.

Potential for Extended Hours or Night Sales: The significant decline in revenue after 4 pm suggests that customers might not be shopping during late hours. However, there could be potential for extended hours or night sales if there is a demand that is not currently being met.

Understanding the revenue generation pattern across different hours of the day helps in making informed business decisions related to staffing, operations, marketing, and potentially extending operating hours. However, it would also be helpful to consider other factors such as operating costs, customer preferences, and regional factors in the overall decision-making process.

## 2. What is the most profitable day of the week?




## Revenue Analysis by Day of the Week

The data presented provides a comprehensive picture of revenue generation across different days of the week. This information can help us understand when sales are at their peak and where there may be opportunities for growth.

Peak Revenue Day: The highest revenue is generated on Thursdays, with a total of £1,941,119.91. This may indicate that customers are more likely to make purchases towards the end of the workweek.

High Revenue Days: Besides Thursday, Tuesday and Wednesday also bring in substantial revenue, with Tuesday marking the second-highest sales of £1,673,648.12.

Low Revenue Days: Sunday sees a significantly lower revenue (£780,223.95) compared to other days, and no revenue is generated on Saturdays. This suggests that the business may be closed or has fewer operational hours on these days.

## *Key insights*

Increasing Weekend Revenue: Given the lower revenue on Sundays and no sales on Saturdays, there could be an opportunity to boost sales during the weekend. This could involve running special weekend promotions or extending operational hours if the business is typically closed on these days.

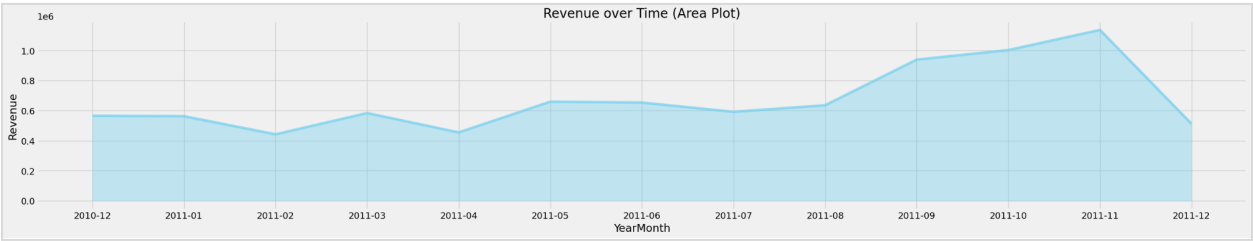
Maximizing Peak Day Revenue: With Thursday being the highest revenue day, further efforts can be made to maximize sales on this day. This could include exclusive Thursday deals or increasing inventory for popular items.

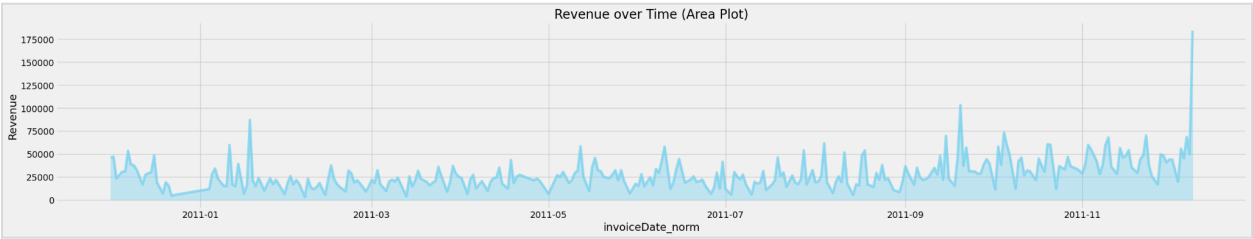
Understanding Consumer Behavior: The high revenue on Tuesday and Wednesday could be attributed to specific consumer behaviors. For instance, customers may be engaging in

midweek shopping to replenish their supplies. Identifying the factors behind this trend could help in planning targeted sales strategies.

While the revenue per day offers valuable insights, it's essential to further explore why these trends occur. This might involve analyzing customer behavior, marketing efforts, or operational decisions that influence these sales patterns.

## 3. Are there any seasonal patterns or trends in sales or revenue?




### Seasonal Patterns in Revenue

A review of monthly revenues reveals a clear trend with seasonality playing a significant role in sales performance. The beginning and end of the year tend to show lower revenues. The start of the year (January and February) shows a decrease in revenue compared to December, which may be due to post-holiday shopping slowdown.

As the year progresses, revenue begins to increase, with a significant surge starting in September. This likely corresponds with an uptick in shopping behavior in preparation for the end-of-year holiday season. The highest revenue is observed in November, which aligns with Black Friday and early Christmas shopping.

However, there is a sharp decrease in revenue in December. This could be attributed to most holiday shopping being completed in November, and decreased activity closer to the holidays when businesses might be closed or reducing orders to prepare for the New Year.

### *Key insights*

Seasonal trends have a significant impact on the business's revenue. The most lucrative period is the last quarter of the year, especially November. Strategies should be devised to capitalize on this seasonality, such as offering promotions or launching new

products during the peak season. Additionally, efforts to boost sales in the slower months could help balance the yearly revenue.

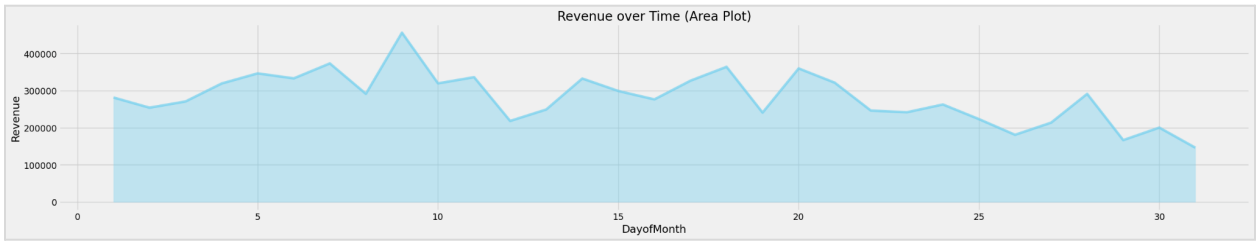
## Limited Data for Trends

The data provided covers only a single year's (2011) worth of sales data. Although we can potentially observe monthly variations in revenue within this single year, we cannot confirm that these variations are seasonal trends.

Seasonal patterns or trends usually imply a recurring pattern over a period of time (usually over several years). In order to conclusively determine the presence of a seasonal pattern, we would typically need several years' worth of data to see if a pattern repeats year over year.

Nevertheless, within this single year, it appears that there's a spike in revenue in the months of September, October, and November. However, without more data, it's not clear if this spike is a recurring seasonal trend or just a one-off event specific to 2011.

### 4. How do sales trends change over the year?




## Annual Sales Trend Analysis

We have analyzed the sales trend over the year from December 2010 to December 2011.

### *Key insights*

Revenue appears to be cyclical with low points occurring around the start of the year (February) and at the end (December).

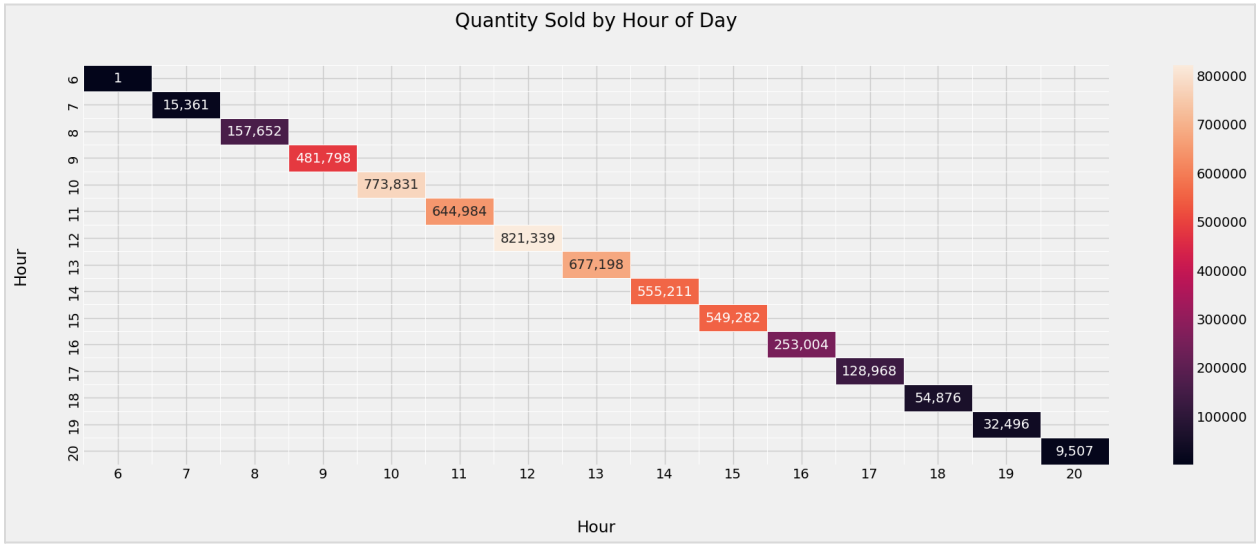
The highest revenue was generated in November 2011 with a total of 1,137,127.00. This could be attributed to increased shopping activity leading up to the holiday season.

A significant increase in revenue can be seen in the months of September, October, and November. This surge in the later months of the year suggests that sales strategies employed during this period have been successful.

The decrease in revenue in December, compared to the preceding months, could be attributed to a post-holiday season slump in sales. However, it's worth noting that December 2010 revenue was higher than the following two months (January and February 2011), indicating that factors other than seasonal trends may also be at play.

Overall, the trend is upward from February to November, suggesting a positive growth trajectory. The company may want to explore strategies to maintain this momentum throughout the year and manage the seasonal fluctuations effectively.

## 5. How does the hour of the day influence the quantity sold?




### Quantity Sold by Hour of the Day

The provided data gives us a detailed overview of the quantity of products sold at each hour of the day. This enables us to identify the most and least active sales hours.

### Peak Sales Time

The maximum quantity of goods sold occurs at 12 pm (noon), with a total of 821,339 items sold. This coincides with the lunch hour, when people may have some free time to make purchases.

### High Sales Period

The time period between 9 am and 3 pm consistently has a high quantity of sales, mirroring the trends in revenue generation.

### Early Morning and Late Night Sales

The number of goods sold is lowest early in the morning (6 am) and late at night (8 pm). This likely reflects fewer customers making purchases during these hours.

### Decline in Sales After Peak

After 1 pm, there's a steady decline in the quantity sold, with a notable drop after 4 pm. This might be due to customers wrapping up their workdays and becoming less likely to shop.

*Key insights*

Inventory Management

Adequate stock levels should be maintained during the high sales period (9 am to 3 pm) to meet the high demand during these hours.

Promotion Timing

Sales promotions and deals can be scheduled during the high sales period to capitalize on customer purchasing trends and increase sales.

Understanding Customer Behavior

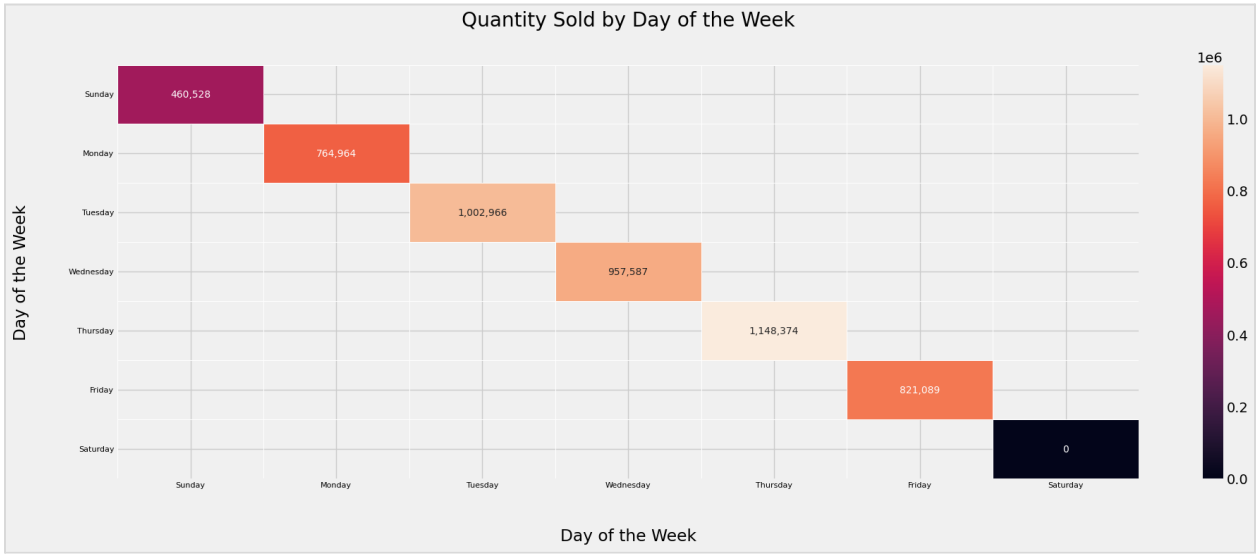
The low sales numbers early in the morning and late in the evening may be due to customer habits and preferences. Additional customer behavior analysis could provide more nuanced insights.

Potential for Night Sales

The significant drop in quantity sold after 4 pm suggests fewer customers shopping during the late hours. However, if there's unmet demand, there could be potential for extended hours or night sales.

Analyzing the quantity sold during various hours of the day is crucial for planning inventory management, promotional activities, and potentially expanding operational hours. However, this should be done alongside consideration of operational costs, customer habits, and specific regional factors.

## 6. How does the day of the week influence the quantity sold?



## Quantity Sold Analysis by Day of the Week

The data showcases the volume of products sold throughout the week. This distribution allows us to identify patterns of customer purchasing behavior across different days.

### Busiest Day

Thursday sees the highest quantity of products sold, amounting to 1,148,374 units. This aligns with the peak revenue observed on the same day, implying a strong correlation between quantity sold and revenue.

### High Volume Days

In line with the revenue trend, Tuesday and Wednesday also see a high quantity of products sold, with Tuesday being the second-highest with 1,002,966 units.

### Low Volume Days

Sunday records the lowest number of units sold among the operational days (460,528 units). Saturdays, as previously noted, see no sales activity, likely indicating that the business is closed.

## *Key insights*

### Boosting Weekend Sales

The low sales volume on Sundays and lack of sales on Saturdays points towards an area of potential improvement. This could involve extending business hours, implementing weekend-specific promotions, or introducing special weekend products to attract customers.

### Leveraging High Volume Days

Considering that Thursdays record the highest number of units sold, strategies could be developed to leverage this trend further. These could include running "bulk buy" promotions or offering discounts on additional purchases

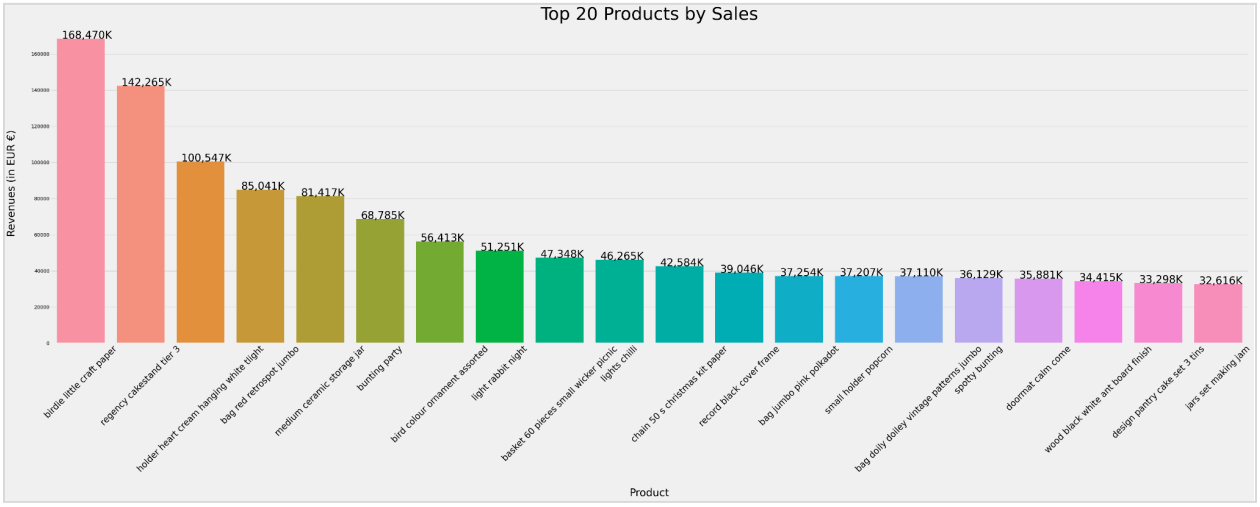
### Studying Consumer Behavior

The high volume of units sold on Tuesday and Wednesday suggests a specific shopping pattern among customers, perhaps restocking midweek. Understanding the driving factors behind this could help tailor strategies to encourage similar behavior on other days.

To develop a comprehensive understanding of these trends, further analysis could delve into which specific products drive sales on these high-volume days, the impact of marketing efforts on these patterns, and how seasonal variations may affect these trends.

# Product strategy

## 1. What products generate the most revenue?




## Revenue and Pricing Analysis of Top-Performing Products

This report analyzes the revenue and unit prices of the top-performing products. The ultimate goal is to understand the correlation between unit prices and overall revenue and to use these insights to guide product pricing, assortment, and sales strategies.

### *Key insights*

Low-Priced, High-Revenue Products Products like 'paper craft little birdie', 'bag red retrospot jumbo', and 'night rabbit light' have relatively low unit prices but have generated significant revenue. This suggests that these products might be high-volume sellers. This could be due to various factors, such as high customer demand, strategic pricing, or successful marketing campaigns.

High-Priced, High-Revenue Products The 'cakestand regency tier 3' and 'small picnic pieces basket 60 wicker' are among the highest-priced products and have also generated substantial revenue. This could indicate a customer willingness to pay for premium products, or it could be a reflection of a successful luxury or high-end product strategy.

Value for Money Products Some products, such as the 'white heart tlight holder cream hanging' and 'chain christmas kit 50 s paper', offer relatively high value for their cost, considering their mid-range unit prices and the substantial revenue they generate. They might represent a 'sweet spot' in terms of price-to-value ratio, appealing to a broad range of customers.

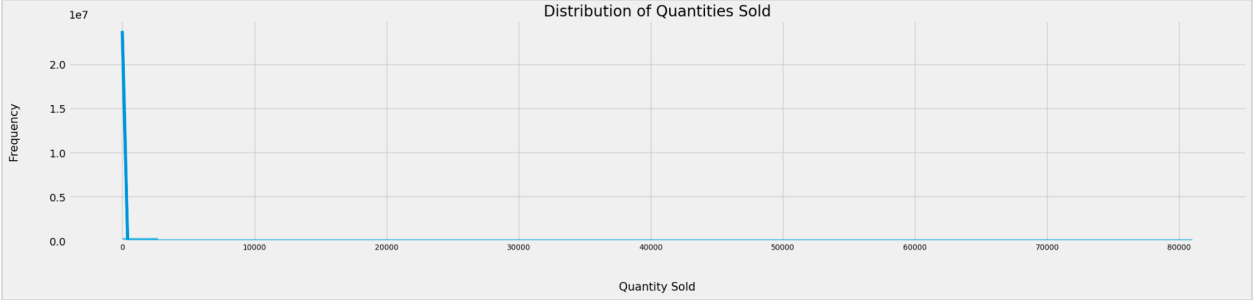
Potential for Revenue Growth Some lower-revenue products, like 'small popcorn holder' and 'bag polkadot pink jumbo', might have potential for revenue growth. Although their current revenue is lower compared to the top performers, their affordable prices could make them

attractive to a wider customer base. Strategic promotion of these items could lead to increased sales and revenue.

Unit Price Variations There's a substantial variation in unit prices among the top-performing products, suggesting that the company's customer base has diverse purchasing power and preferences. Understanding these variations could help refine pricing strategies, potentially leading to increased sales and revenue.

These findings provide a starting point for developing pricing strategies and optimizing product assortment. It's recommended to conduct further analysis to understand the underlying reasons for these patterns and to validate the assumptions made in this report. This could involve analyzing additional factors like customer demographics, seasonality, and marketing efforts.

## 2. What is the distribution of quantity sold for different products?




### Sales Volume Distribution for Top 20 Products

The provided data shows the distribution of sales volumes (quantity sold) for the top 20 products.

This analysis allows us to identify the sales behavior of these products in terms of overall quantity sold, the number of transactions (count), the average quantity per transaction (mean), and the range of quantity sold per transaction (min, max).

### High Sales Single Transaction

'Paper craft little birdie' stands out with a total quantity sold of 80,995 in just one transaction, making it an extreme outlier. This might be due to a bulk purchase by a single customer or a data recording anomaly.

### Consistent High Sales

'Ceramic medium storage jar' has consistently high sales with a total quantity of 77,916 over 195 transactions. Despite the large total quantity sold, the quantities per transaction vary greatly (from 1 to 74,215), indicating both regular and bulk purchases.

Mid-Range Sales

Products like 'asstd gliders designs world 2 war', 'small popcorn holder', and 'bag red retrospot jumbo' exhibit a good balance between total quantity sold and the number of transactions, demonstrating a steady demand.

Low Average Quantity, High Transaction Volume

Some products such as 'small popcorn holder', 'bag red retrospot jumbo', and 'retrospot spotty red bag lunch' have lower average quantities per transaction but higher transaction counts. This could indicate that these products are frequently bought in small quantities.

*Key insights*

Inventory Management for High-Volume Products

High sales products, particularly those sold in large quantities per transaction, should be stocked adequately to meet demand and avoid stock-outs.

Promotion Opportunities for Mid-Range Products

Products with mid-range sales can be targeted for promotions to further boost their sales. Their consistent demand indicates a potential for growth.

Customer Purchase Behavior Analysis

Products that are frequently bought in small quantities can provide insights into customer purchase behavior. Understanding why customers purchase these products regularly can inform marketing strategies and product placement.

Anomaly Investigation

The unusual sales pattern for 'paper craft little birdie' should be investigated to ensure data accuracy. If it is a valid bulk purchase, it could be beneficial to understand the circumstances and explore potential opportunities for other large sales.

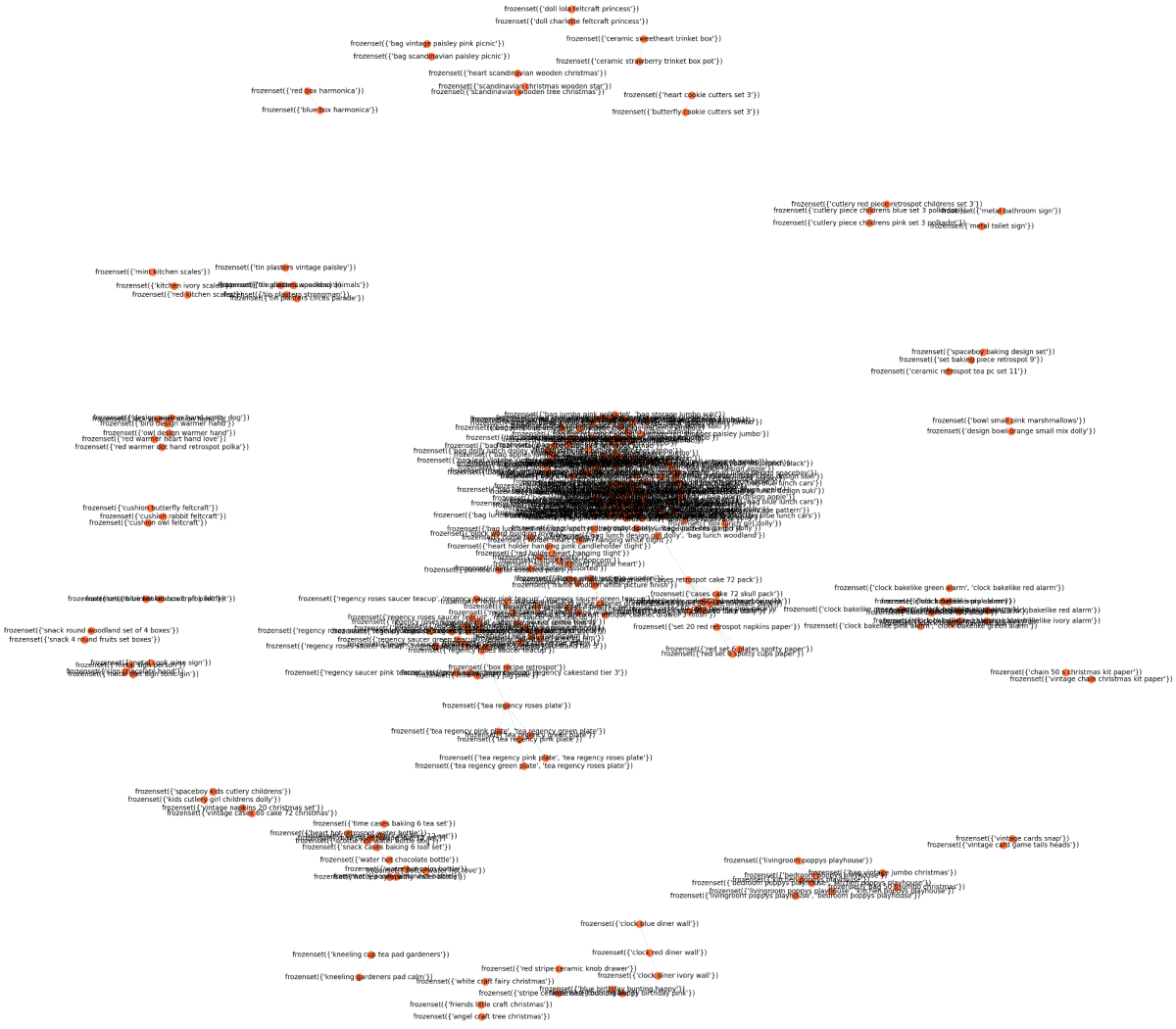
Dynamic Pricing Strategies

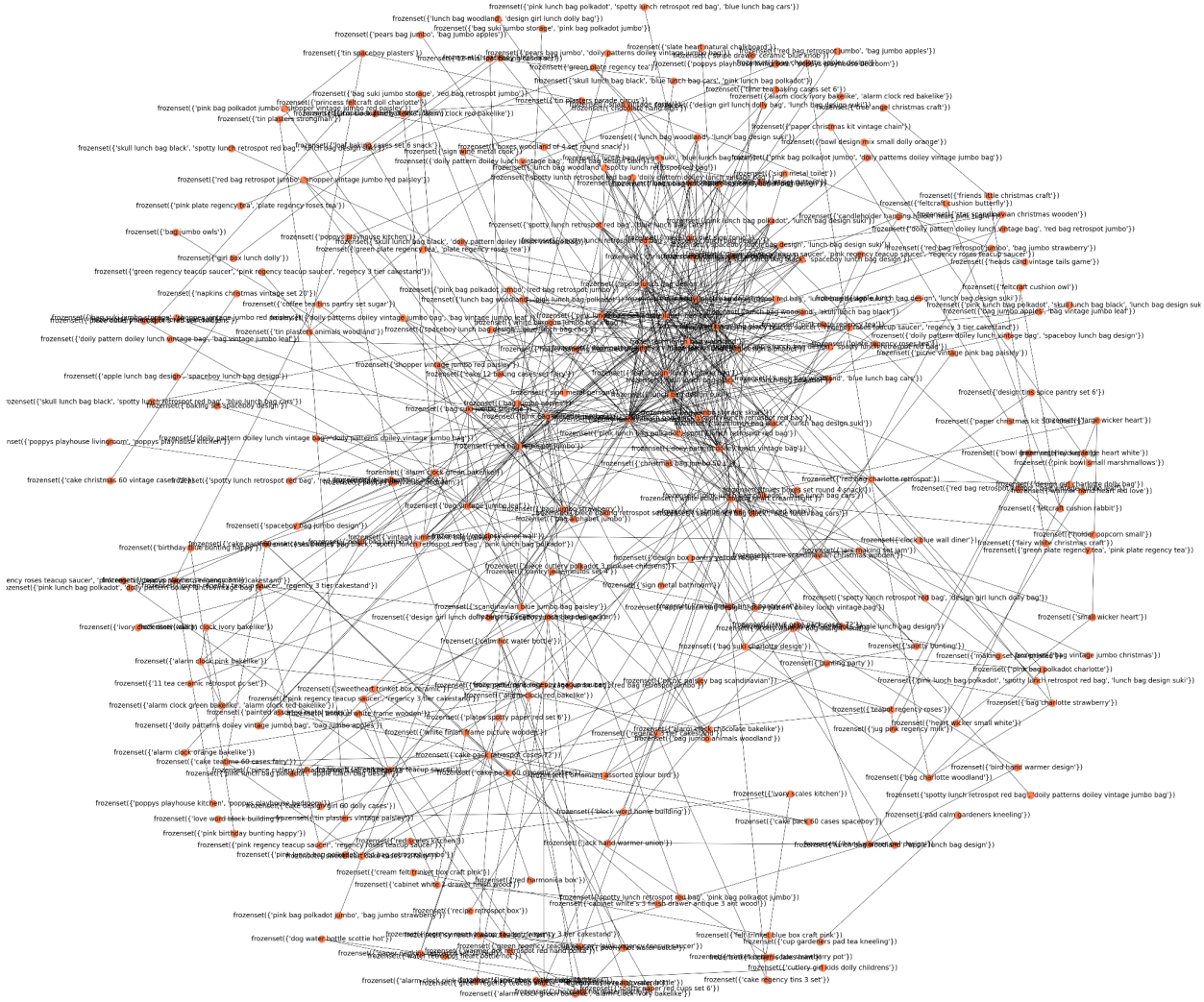
For products where there is a large range between the minimum and maximum quantity sold per transaction, dynamic pricing strategies may be effective. For instance, bulk purchases could be incentivized with discounts.

This analysis provides an overview of the sales volume distribution for the top 20 products. However, understanding the complete sales performance would require an analysis of other factors such as revenue generated, profit margins, and seasonal trends.

# 3. Are there any patterns in the types of items bought together?

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 1119 | (tea regency pink plate) | (tea regency green plate, tea regency roses plate) | 0.01 | 0.01 | 0.01 | 0.83 | 66.43 | 0.01 | 5.67 | 1.00 |
| 1118 | (tea regency green plate, tea regency roses plate) | (tea regency pink plate) | 0.01 | 0.01 | 0.01 | 0.81 | 66.43 | 0.01 | 5.14 | 1.00 |
| 1120 | (tea regency green plate) | (tea regency pink plate, tea regency roses plate) | 0.01 | 0.01 | 0.01 | 0.69 | 64.06 | 0.01 | 3.14 | 1.00 |
| 1117 | (tea regency pink plate, tea regency roses plate) | (tea regency green plate) | 0.01 | 0.01 | 0.01 | 0.94 | 64.06 | 0.01 | 16.18 | 1.00 |
| 637 | (tea regency green plate) | (tea regency pink plate) | 0.01 | 0.01 | 0.01 | 0.75 | 61.52 | 0.01 | 3.92 | 1.00 |
| 636 | (tea regency pink plate) | (tea regency green plate) | 0.01 | 0.01 | 0.01 | 0.90 | 61.52 | 0.01 | 10.03 | 1.00 |
| 1077 | (livingroom poppys playhouse) | (bedroom poppys playhouse, kitchen poppys playhouse) | 0.01 | 0.01 | 0.01 | 0.74 | 53.52 | 0.01 | 3.77 | 0.99 |
| 1076 | (bedroom poppys playhouse, kitchen poppys playhouse) | (livingroom poppys playhouse) | 0.01 | 0.01 | 0.01 | 0.73 | 53.52 | 0.01 | 3.68 | 1.00 |
| 522 | (milk regency jug pink) | (green sugar regency bowl) | 0.01 | 0.01 | 0.01 | 0.76 | 52.05 | 0.01 | 4.06 | 1.00 |
| 523 | (green sugar regency bowl) | (milk regency jug pink) | 0.01 | 0.01 | 0.01 | 0.77 | 52.05 | 0.01 | 4.26 | 1.00 |





The market basket analysis uncovers interesting patterns of items often purchased together.

Key associations are observed between "Regency Tea Plates", "Poppy's Playhouse" set, and "Regency Kitchenware".

These insights could be used for a number of purposes, such as designing promotions, arranging products in a catalog, suggesting items to customers, and more.

### *Key insights*

Based on the data, several product groups often purchased together emerge:

The "Regency Tea Plates"

Customers who buy the "pink plate regency tea" are likely to buy the "green plate regency tea" and "plate regency roses tea". This suggests a customer preference for matching sets in these items and presents an opportunity for bundled promotions. This association is quite strong given the high confidence, lift, and Zhang's metric.

The "Poppy's Playhouse" set

"poppys playhouse livingroom" is often bought together with "poppys playhouse kitchen" and "poppys playhouse bedroom". Indicating that customers might prefer to buy the entire set. This provides an opportunity for a packaged deal.

The "Regency Kitchenware"

"bowl green regency sugar" and "jug pink regency milk" are often bought together, which might be a reflection of customer preference for matching kitchenware. This can guide product placement strategy or cross-selling recommendations.

In conclusion, the report points out customers' preferences for matching items and sets, providing a direction for creating effective promotions and product recommendations.

### 4. Which product pairings are most frequently bought together?

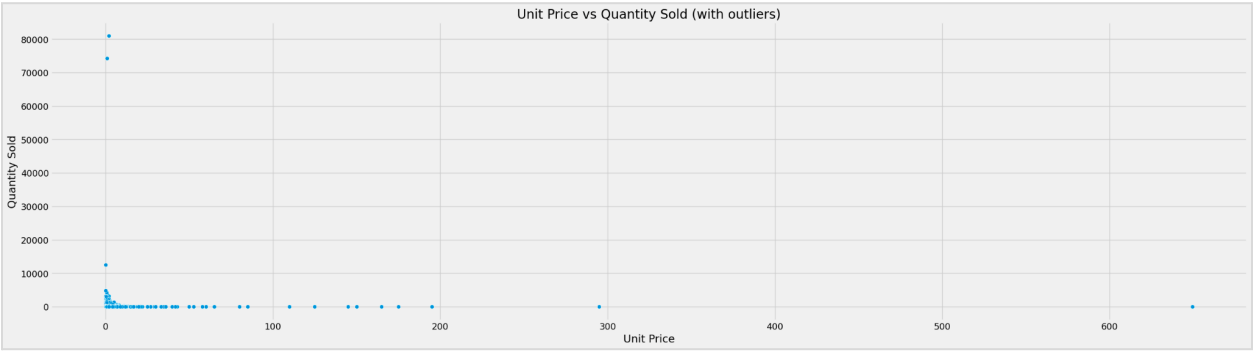
```
count = Counter()

for row in order_combo["Bought_Together"]:
    row_list = row.split(", ")
    count.update(Counter(combinations(row_list, 2)))

# Create a list of dictionaries to store the results
results = [{"Combination": ", ".join(key), "Count": value} for key, value in count.most_common(10)]

# Convert the list of dictionaries into a DataFrame
pairings = pd.DataFrame(results)
pairings
```

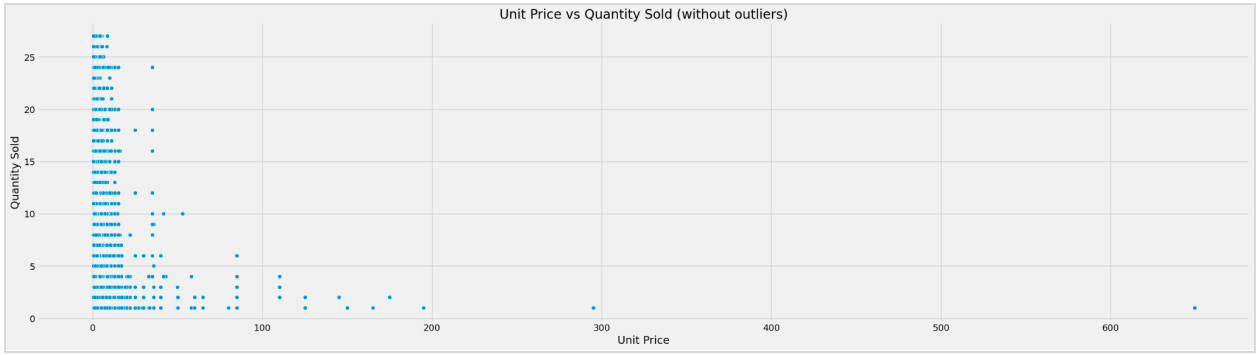
✓ 4.5s

| | Combination | Count |
|---|---|---|
| 0 | bag jumbo pink polkadot, bag red retrospot jumbo | 545 |
| 1 | regency saucer green teacup, regency roses saucer teacup | 543 |
| 2 | clock bakelike green alarm, clock bakelike red alarm | 530 |
| 3 | bag lunch red retrospot spotty, bag lunch pink polkadot | 522 |
| 4 | bag lunch red retrospot spotty, bag lunch design suki | 519 |
| 5 | bag lunch red retrospot spotty, bag skull lunch black | 517 |
| 6 | frame wooden white picture finish, frame white antique wooden | 468 |
| 7 | bag doily doiley vintage patterns jumbo, bag red retrospot jumbo | 468 |
| 8 | bag lunch red retrospot spotty, bag lunch design spaceboy | 466 |
| 9 | bag skull lunch black, bag lunch design suki | 465 |

## Pricing strategy

### 1. How does unit price relate to quantity sold?






## Product Sales Volume and Unit Pricing Analysis

The provided data delineates the sales volume (quantity sold) and the unit price of different products. This analysis identifies the products that are purchased in larger quantities and observes the correlation with the unit price.

### High-Volume, Low-Price Products

Products like 'paper craft little birdie', 'ceramic medium storage jar', and 'asstd gliders designs world 2 war' are selling in high volumes. Notably, these products have relatively lower unit prices, suggesting they may be popular due to their affordability.

### Mid-Volume, Mid-Price Products

Products such as 'white heart tlight holder cream hanging', 'ornament assorted bird colour', and 'bag red retrospot jumbo' have a moderate sales volume and a mid-range unit price. This could indicate a balanced cost-value perception among customers.

### Lower-Volume, Varying-Price Products

Products like 'harmonica box red', 'bag polkadot pink jumbo', and 'vintage doily bag patterns doily jumbo' exhibit a lower sales volume. Their prices vary, indicating that pricing may not be the only factor affecting their sales volume. Other factors such as customer preference, product visibility, and marketing efforts could be affecting their performance.

### High Sales of Low-Priced Products

Many of the high-volume products have a low unit price, such as 'asstd gliders designs world 2 war' and '72 retrospot pack cake cases'. This trend could indicate a strong market for value or budget-friendly items.

## *Key insights*

### Pricing Strategy for High-Volume Products

The high-volume sales of low-priced products suggests a strong market for these items. It might be beneficial to maintain competitive pricing for these products to continue driving sales volume.

Marketing Focus on Mid-Volume, Mid-Price Products

Mid-volume, mid-price products could benefit from targeted marketing efforts. These products seem to offer a good balance of price and value and might have potential for sales growth.

Investigate Low-Volume Products

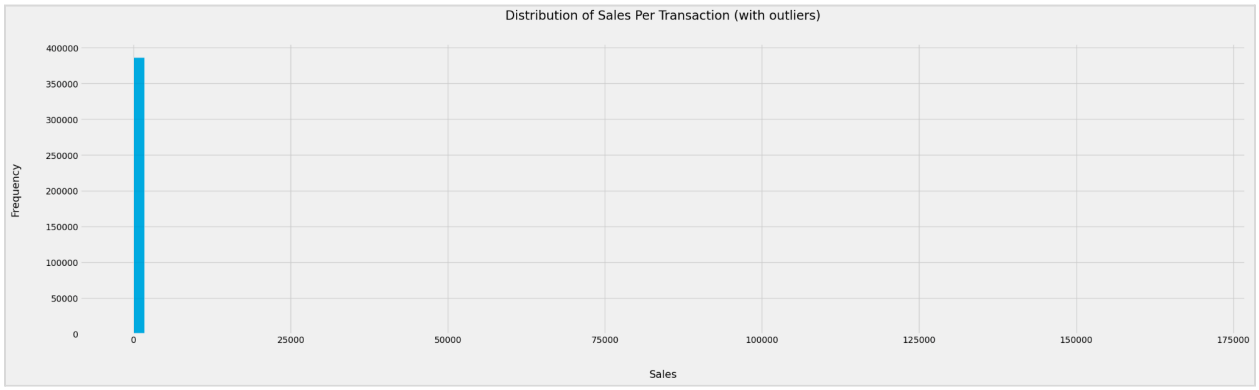
It could be valuable to investigate the factors affecting the sales of lower-volume products. If these items are indeed less popular, it may be useful to reevaluate their pricing, marketing, or positioning in the store.

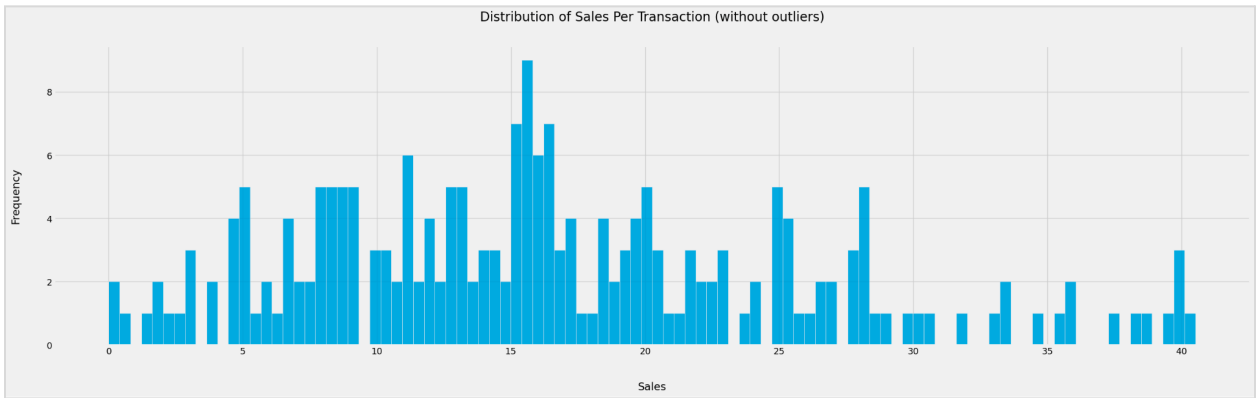
Opportunities in Value Market

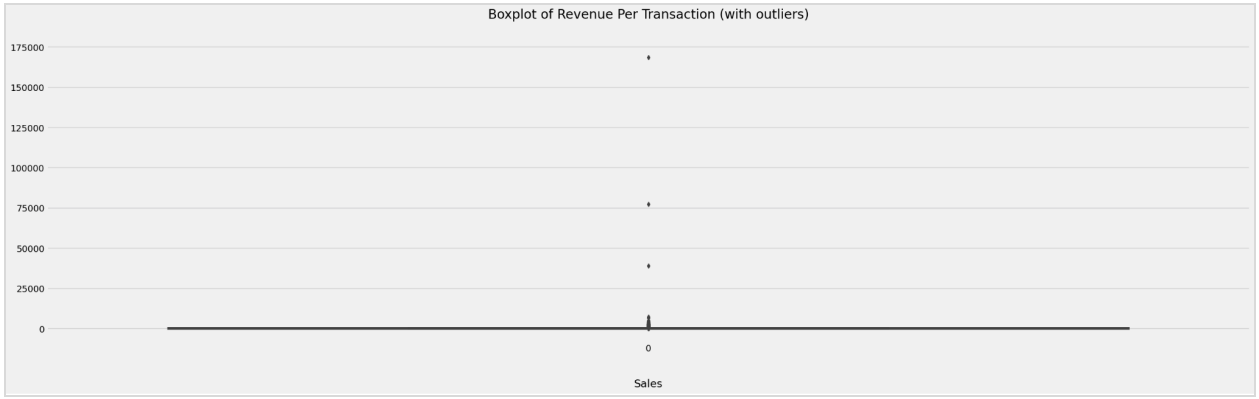
The success of low-priced, high-volume products indicates a potential opportunity in the value market segment. Offering more products in this range could attract price-sensitive customers and increase overall sales volume.

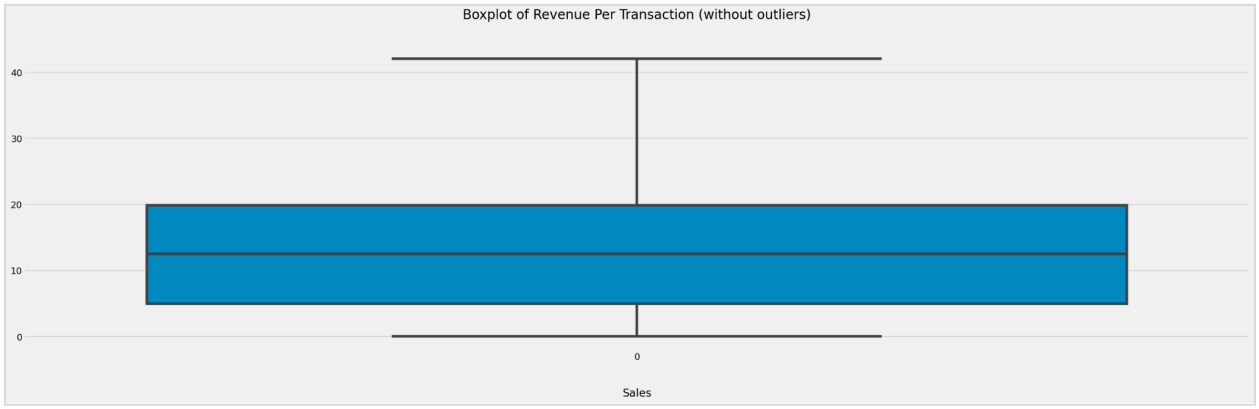
This data provides a useful overview of product performance in terms of sales volume and pricing. To further refine strategies, it would be beneficial to consider additional factors, such as product margins, seasonal trends, and customer demographics.

## 2. What is the distribution of revenue per transaction?




Boxplot of Revenue Per Transaction (with outliers)
175000
150000
125000
100000
75000
50000
25000
0
0
Sales


Boxplot of Revenue Per Transaction (without outliers)
40
30
20
10
0
0
Sales

# SUMMARY

This report provided a comprehensive analysis of the e-commerce sales data of the company, starting from a detailed understanding of the data structure, proceeding to data preparation and cleaning, transitioning to SQL and NoSQL databases, followed by rigorous exploratory data analysis and visualization of the data.

We began by discussing the raw data format and its structure, after which we undertook a process of data cleaning and preparation using Python. This included handling missing values, converting data types, and generating new attributes that were more suited to the subsequent analysis.

We then migrated our data to MySQL. This was a crucial step to organize the data in a structured manner for efficient querying and extraction of meaningful insights. We also outlined the process of importing data into the MySQL database using Python and defined how we structured the tables in SQL.

We discussed the concept of Entity-Relationship Diagram (ERD) to provide a visual overview of how our tables are linked in the SQL database. This gives us a clear understanding of the structure of our SQL database and how different entities (tables) are related to each other.

We executed various SQL queries to extract valuable insights from our data, such as identifying the top customers by total quantity, top products by total sales, sales trends per month, top selling products for the most frequent customers, and sales breakdown by country.

Following this, we delved into Exploratory Data Analysis (EDA), outlining the importance of this stage in uncovering trends, patterns, and relationships in our data. We laid out a comprehensive plan to explore various facets of the business, such as sales performance, customer behavior, geographic market, temporal sales patterns, product and pricing strategy, customer retention, and predictive analysis for future sales.

# CONCLUSION

In conclusion, the process we followed in this report illustrates a complete pipeline of data analysis, starting from raw data preparation, choosing the appropriate database type, extracting insights using SQL queries, to finally performing comprehensive EDA. This methodology not only enhances our understanding of the current business performance, but also empowers us to predict future trends, enabling strategic planning and decision making.

The data-driven insights generated from this report will enable the company to understand their sales performance in depth, identify their most valuable customers and products, analyze market trends, optimize pricing and product strategy, enhance customer retention, and forecast future sales.

This report will serve as a solid foundation for further analysis and exploration, and its methodologies can be adopted for other similar datasets and business contexts. The ultimate goal is to continually leverage data to drive informed and effective business decisions.

THANK YOU.

Katrina JUMADIAO
*Data Analyst*