**BUSINESS UNDERSTANDING**

**Business Overview**

Given data of trade done on an e-commerce retailer in the United Kingdom, we as data scientists are tasked to study and analyze the dataset and get insights on the consumer behaviour and trends throughout the year. This will help the company’s management in allocation of resources to their products catalogue and define the right amount across the different countries the e-commerce operates.

**Business Objectives**

The main idea of the analysis is to draw insights that will assist in policy formulations, marketing strategy, stock inventory management and assist the company in its growth and prediction of consumer behaviour.

**Business Objective Question**

Which products, country, month should the e-commerce retailer focus on and which products should it do away with in a bid to improve its marketing, inventory management and yearly turnover?

**Business Success Criteria**

To determine the most effective growth strategy for the company to optimize expenditure  
 and returns.

**Requirements, Assumptions and Constraints**

1. **Resources**

i) Personnel (Data Analysts, Data Miners)

ii) The [project dataset.](https://www.kaggle.com/carrie1/ecommerce-data/download)

iii) Softwares (Jira, Google Colaboratory Notebook, GitHub)

iv) Computing resources

1. **Assumptions**

i) The data recorded is an actual representation of the transactions done.

ii) The data follows a seasonal trend.

1. **Risks and Contingencies**

The data could be missing important information.

1. **Cost/ Benefit Analysis**

The potential revenues associated with discerning consumer behaviour and stock inventory management is totally incomparable to the consultancy and project implementation costs.

This is however dependent on the resources allocated to the marketing and the project implementation resources.

However getting to know the right products and peak months, hours, days results in more revenues and profits.

**Data Mining Goals**

We are concerned about the consumer behaviour in different times of the year, the products with the highest sales and returns across the countries.

The potential questions for consideration include:-

i) Which is the peak month of the year? Which is the most popular day for purchases?

ii)Which are the top three countries with the most sales? Which product is most popularly bought in these three countries? Which item has brought in most returns? Overall and in the three countries? Is the most popular product the one that brings in the most revenue?

iii) Which items have brought in the least returns? Overall and in the individual countries?

Does the least popular product bring in the least returns?

iv) Which products are more likely to have cancelling issues? The country with the highest cancelling issues,the popular product cancelled least and the highest.

v) Which country do we have the most customers from? Is it the country bringing the most revenue? What is the average expenditure of a consumer in the UK? Germany? France?

vi) Which country is seen to require more marketing to optimize its expenditure?

**Project Plan**

The Cross-Industry Standard Process for Data Mining (CRISP-DM) will be used as a guideline for conducting our research. Below is the overview plan:

|  |  |  |  |
| --- | --- | --- | --- |
| Phase | Time | Resources | Risks |
| Business Understanding | 1 Hour | The project dataset/  Data Scientists |  |
| Data Understanding | 1 Hour | The project dataset/  Data Scientists |  |
| Data Preparation | 2 Hours | The project dataset/  Data Scientists |  |
| Data Analysis | 3 Hours | The project dataset/  Data Scientists | Some assumptions |
| Recommendation | 1 Hour | The project dataset/  Data Scientists |  |
| Evaluation | 1 Hour | The project dataset/  Data Scientists |  |

# **DATA UNDERSTANDING**

**Overview**

The dataset was obtained from the kaggle website from the following link <https://www.kaggle.com/carrie1/ecommerce-data>. The existing dataset file contains an E-commerce retail company based in the United Kingdom.The initial dataset contains one table with 541,909 records of transactions in 8 columns.

The dataset to be analysed for our study is provided in the *link* below.

[project dataset](https://www.kaggle.com/carrie1/ecommerce-data/download)

**Collecting initial data**

The data set was downloaded from the website as a csv file into our local computers and then loaded the data into our notebook for further analysis.

**Describing and exploring data**

The dataset has been presented in columns and rows. Each column has a specific attribute as follows:

Invoice No - the invoice number of the transactions

Stock code - the unique identifier of the product

Customer ID - the unique identifier of the person who purchased the product

Description - the product description

Quantity - the amount of products purchased

Unit price - the price of each product

Invoice date - the date the transactions were made

Country - the country where the transaction was made

The data did not have any data description with it.

**Verifying data quality**

While verifying our data, we checked for

* Validity - all the columns were well defined and relevant for our analysis
* Accuracy - we came across some negative values in our quantity column which were accompanied by a unique Invoice number indicating the transactions that did not go through hence were cancelled
* Completeness - The dataset contained numerous missing values mostly on the CustomerID column and a few on the Description column
* Consistency- the dataset contained duplicate values of Invoice numbers and CustomerID’s

**DATA PREPARATION**

Steps that were taken during data preparation are as follows:

**Selecting Data**

The following columns were used for analysis in this project based on the relevance of our goals and data quality:

* InvoiceNo, StockCode, Description, Quantity, UnitPrice, InvoiceDate, CustomerID and Country.

All the records were also necessary for our analysis hence no data was left out.

The data was loaded as a dataframe from a csv file as df1, previewed, examined basic properties such as the size of the dataframe, unique values from the categorical data, datatypes of each columns e.t.c

**Data Cleaning**

Data cleaning procedures performed during analysis included the following:

To ensure our dataset was complete, we checked for null values which were 135,080 records from our CustomerID column and 1,454 records in our description column. These records were dropped as they would have caused serious misanalysis of our data

To ensure accuracy of our analysis, the records with the cancelled transactions were derived from the dataset as a new data frame - df4 - which consisted of 8,905 records. This dataframe was used for its own set of analysis.

For consistency, we checked for duplicate values which were mainly in the columns InvoiceNo and CustomerID. However, we decided to retain the values as the contents were of great value to our analysis.

For uniformity of our dataset, the InvoiceDate was changed to a datetime datatype as it was necessary for our analysis and the date and time were both split to allow for individual manipulation

## **Constructing New Data**

A new dataframe df3 was created, it contained the records and/or transactions that actually did go through. This contained 397,924 records which we used for most of our analysis. In addition to this, there was the df4 that had been mentioned earlier that contained the records with the cancelled transactions which was also used for a different set of analysis.

New rows were obtained from the dataset as follows:

* Total Expenditure - this was obtained by multiplying the Quantity and the UnitPrice columns. This was necessary as our analysis required calculations of returns by the e-commerce company. A box plot was used to check for outliers in our total expenditure column.
* Date and Time - Was a result of splitting of the InvoiceDate column to accommodate part of our analysis that required specific elements from the column i.e month and days with the highest returns
* Year, months and day - Further splitting of the Date column was conducted as we needed to access specific months in our analysis
* Hour, minutes and seconds - Further splitting of the Time column was also conducted as specific hours had to be accessed to allow for analysis

## **DATA ANALYSIS**

The following questions were looked into during our analysis;

**Question one**

a. Which were the peak months of the year?

months

11 1161817.380

12 1090906.680

10 1039318.790

9 952838.382

5 678594.560

6 661213.690

8 645343.900

7 600091.011

3 595500.760

1 569445.040

4 469200.361

2 447137.350

b. Which are the peak days in the peak month - 11?

days

23 71979.93

10 70513.29

9 61489.18

3 60672.11

14 58777.71

4 56099.24

17 55885.30

28 51831.67

22 49664.89

29 48851.68

16 48439.76

15 47729.00

21 45333.13

6 42941.34

30 41481.23

2 38734.70

24 38579.11

8 38295.12

11 37081.37

18 36751.25

20 30190.92

1 29132.81

7 28779.24

13 28607.78

25 26674.66

27 17300.96

**Question two**

a. Which top three countries had the most sales?

Country

United Kingdom 354345

Germany 9042

France 8342

EIRE 7238

Spain 2485

Netherlands 2363

Belgium 2031

Switzerland 1842

Portugal 1462

Australia 1185

Norway 1072

Italy 758

Channel Islands 748

Finland 685

Cyprus 614

Sweden 451

Austria 398

Denmark 380

b. Which products were most popularly bought in these countries?

* in the United Kingdom

Description

WHITE HANGING HEART T-LIGHT HOLDER 1940

JUMBO BAG RED RETROSPOT 1464

REGENCY CAKESTAND 3 TIER 1426

ASSORTED COLOUR BIRD ORNAMENT 1333

PARTY BUNTING 1308

* in France

Description

POSTAGE 300

RABBIT NIGHT LIGHT 73

RED TOADSTOOL LED NIGHT LIGHT 70

PLASTERS IN TIN WOODLAND ANIMALS 68

PLASTERS IN TIN CIRCUS PARADE 66

* in Germany

Description

POSTAGE 374

ROUND SNACK BOXES SET OF4 WOODLAND 113

ROUND SNACK BOXES SET OF 4 FRUITS 72

PLASTERS IN TIN WOODLAND ANIMALS 64

REGENCY CAKESTAND 3 TIER 63

c. Which products brought in most returns? Overall and in the three countries with the highest sales?

* Overall

Description

PAPER CRAFT , LITTLE BIRDIE 168469.600

REGENCY CAKESTAND 3 TIER 142592.950

WHITE HANGING HEART T-LIGHT HOLDER 100448.150

JUMBO BAG RED RETROSPOT 85220.780

MEDIUM CERAMIC TOP STORAGE JAR 81416.730

* in the United Kingdom

Description

PAPER CRAFT , LITTLE BIRDIE 168469.60000

PICNIC BASKET WICKER 60 PIECES 19809.75000

TEA TIME TEA TOWELS 3022.50000

DOTCOM POSTAGE 744.14750

HALL CABINET WITH 3 DRAWERS 625.8825

* in France

Description

Manual 1582.061667

MINI WOODEN HAPPY BIRTHDAY GARLAND 835.200000

PINK HAPPY BIRTHDAY BUNTING 232.500000

PINK PAINTED KASHMIRI CHAIR 171.800000

JUMBO BAG STRAWBERRY 132.766667

* in Germany

Description

STOOL HOME SWEET HOME 318.250000

SET OF 16 VINTAGE BLACK CUTLERY 262.800000

Manual 255.138889

COLOURING PENCILS BROWN TUBE 212.000000

REGENCY CAKESTAND 3 TIER 143.840476

**Question 3**

a. Which products are the least popular, overall and in the three countries?

* Overall

Description

PINK BAROQUE FLOCK CANDLE HOLDER 1

BLACK CHERRY LIGHTS 1

CRYSTAL CHANDELIER T-LIGHT HOLDER 1

BLACK 3 BEAD DROP EARRINGS 1

PINK POLKADOT KIDS BAG 1

* In the UK

Description

LETTER "O" BLING KEYRING 1

BAKING MOULD CUPCAKE CHOCOLATE 1

GLASS AND PAINTED BEADS BRACELET OL 1

GLASS AND BEADS BRACELET IVORY 1

SET/3 TALL GLASS CANDLE HOLDER PINK 1

* in France

Description

50'S CHRISTMAS GIFT BAG LARGE 1

MEASURING TAPE BABUSHKA BLUE 1

MEASURING TAPE BABUSHKA RED 1

MEDIUM PINK BUDDHA HEAD 1

METAL MERRY CHRISTMAS WREATH 1

* in Germany

Description

CHRISTMAS GINGHAM HEART 1

MEDIUM MEDINA STAMPED METAL BOWL 1

DARK BIRD HOUSE TREE DECORATION 1

I'M ON HOLIDAY METAL SIGN 1

DANISH ROSE PHOTO FRAME 1

b. Which items have brought in the least returns? Overall and in the three countries?Does the least popular product bring in the least returns?

* Overall

Description

PADS TO MATCH ALL CUSHIONS 0.00075

HEN HOUSE W CHICK IN NEST 0.42000

60 GOLD AND SILVER FAIRY CAKE CASES 0.55000

SET 12 COLOURING PENCILS DOILEY 0.65000

CHAMPAGNE TRAY BLANK CARD 0.76000

* in the United Kingdom?

Description

PADS TO MATCH ALL CUSHIONS 0.00075

HEN HOUSE W CHICK IN NEST 0.42000

60 GOLD AND SILVER FAIRY CAKE CASES 0.55000

WINE BOTTLE DRESSING LT.BLUE 0.76000

CHAMPAGNE TRAY BLANK CARD 0.76000

* in France?

Description

BLUE EGG SPOON 0.360000

GLITTER HEART DECORATION 0.390000

MIXED NUTS LIGHT GREEN BOWL 0.420000

TRAVEL CARD WALLET RETRO PETALS 0.420000

TRAVEL CARD WALLET VINTAGE ROSE 0.420000

* in Germany?

Description

ROUND CAKE TIN VINTAGE GREEN 0.000000

SWALLOW SQUARE TISSUE BOX 0.390000

CHERUB HEART DECORATION GOLD 0.830000

SANDALWOOD FAN 0.850000

FOLKART ZINC HEART CHRISTMAS DEC 0.850000

**Question four**

a. Which products are more likely to have cancelling issues?

Description

PAPERCRAFT , LITTLE BIRDIE -80995

MEDIUM CERAMIC TOP STORAGE JAR -74494

ROTATING SILVER ANGELS T-LIGHT HLDR -9367

Manual -3995

FAIRY CAKE FLANNEL ASSORTED COLOUR -3150

b. Which products are least likely to have cancelling issues?

Description

FUNKY WASHING UP GLOVES ASSORTED -1

BLACK HEART CARD HOLDER -1

RECYCLED ACAPULCO MAT PINK -1

RECYCLED ACAPULCO MAT TURQUOISE -1

BLACK BAROQUE WALL CLOCK -1

c. Which countries have the most and least cancelled products?

Country

United Kingdom -540518.16

EIRE -15260.68

France -12311.21

Singapore -12158.90

Germany -7168.93

Spain -6802.53

Portugal -4380.08

Japan -2075.75

USA -1849.47

Sweden -1782.42

Australia -1444.04

Norway -1001.98

Netherlands -784.80

Switzerland -704.55

Cyprus -644.09

Italy -592.73

Channel Islands -364.15

Belgium -285.38

Israel -227.44

Malta -220.12

Finland -219.34

Denmark -187.20

Poland -121.51

Czech Republic -119.02

Greece -50.00

Austria -44.36

Saudi Arabia -14.75

European Community -8.50

**Question five**

a. Which countries do we have the most and the least customers from?

Country

United Kingdom 354345

Germany 9042

France 8342

EIRE 7238

Spain 2485

Netherlands 2363

Belgium 2031

Switzerland 1842

Portugal 1462

Australia 1185

Norway 1072

Italy 758

Channel Islands 748

Finland 685

Cyprus 614

Sweden 451

Austria 398

Denmark 380

Poland 330

Japan 321

Israel 248

Unspecified 244

Singapore 222

Iceland 182

USA 179

Canada 151

Greece 145

Malta 112

United Arab Emirates 68

European Community 60

RSA 58

Lebanon 45

Lithuania 35

Brazil 32

Czech Republic 25

Bahrain 17

Saudi Arabia 9

b. Countries bringing in the most and least revenue?

Country

United Kingdom 7.308392e+06

Netherlands 2.854463e+05

EIRE 2.655459e+05

Germany 2.288671e+05

France 2.090240e+05

Australia 1.385213e+05

Spain 6.157711e+04

Switzerland 5.644395e+04

Belgium 4.119634e+04

Sweden 3.837833e+04

Japan 3.741637e+04

Norway 3.616544e+04

Portugal 3.343989e+04

Finland 2.254608e+04

Singapore 2.127929e+04

Channel Islands 2.045044e+04

Denmark 1.895534e+04

Italy 1.748324e+04

Cyprus 1.359038e+04

Austria 1.019868e+04

Poland 7.334650e+03

Israel 7.221690e+03

Greece 4.760520e+03

Iceland 4.310000e+03

Canada 3.666380e+03

USA 3.580390e+03

Malta 2.725590e+03

Unspecified 2.667070e+03

United Arab Emirates 1.902280e+03

Lebanon 1.693880e+03

Lithuania 1.661060e+03

European Community 1.300250e+03

Brazil 1.143600e+03

RSA 1.002310e+03

Czech Republic 8.267400e+02

Bahrain 5.484000e+02

Saudi Arabia 1.459200e+02

c. What is the average expenditure of consumers per country?

Country

Netherlands 120.798282

Australia 116.895620

Japan 116.561900

Singapore 95.852658

Sweden 85.096075

Denmark 49.882474

Lithuania 47.458857

Lebanon 37.641778

EIRE 36.687745

Brazil 35.737500

Norway 33.736418

Czech Republic 33.069600

Finland 32.913985

Greece 32.831172

Bahrain 32.258824

Switzerland 30.642752

Israel 29.119718

United Arab Emirates 27.974706

Channel Islands 27.340160

Austria 25.624824

Germany 25.311562

France 25.056827

Spain 24.779521

Malta 24.335625

Canada 24.280662

Iceland 23.681319

Italy 23.064960

Portugal 22.872702

Poland 22.226212

Cyprus 22.134169

European Community 21.670833

United Kingdom 20.625073

Belgium 20.283772

USA 20.002179

RSA 17.281207

Saudi Arabia 16.213333

Unspecified 10.930615

**RECOMMENDATIONS**

From our analysis,the following recommendations were provided,

1. The best time for the company to launch marketing campaigns should be around September, as most sales are recorded towards the end of the year i.e November, December and October respectively. Additionally, the company can employ research on why customers purchase more on these seasons for improving sales on the other seasons.
2. A product analysis on the most cancelled product(Paper Craft Little Birdie) should be launched concentrating on the product’s quality, cost, product description and other key features as possible reasons why the product is being cancelled
3. Countries like the Netherlands,Japan, Eire and Australia have more promising indications on bringing in returns hence marketing strategies to increase the customer base should be launched to maximize returns
4. For products with the least sales overall the company can carry out a customer survey to find out why these products have low sales.Then in turn a marketing strategy can be implemented to increase sales. For example, advertisements, giving discounts, incentives.
5. The company should also maximise on sales of the top popular products being bought by the customers across all the countries. This is more likely to help improve revenue generation.
6. For inventory management, best time for the company to stock up is also towards the end of the year as it is when they receive most purchases
7. The company should embrace customer feedback platforms on all their products across all countries. This will in turn lead to identification of market gaps and product diversification as per customers suggestions.
8. For the UK, Germany and France, the company should try to standardize prices to reduce the gap revealed on returns and to maximize returns based on the sales they are making

**EVALUATION**

Our business success criteria has been successful as we have been able to cover the aspects of marketing strategies, inventory management and yearly turn over as has been indicated in our business objectives. This has been highlighted in the recommendations part based on insights obtained from our analysis. Our analysis was based on questions influenced by our need to investigate the dataset to gain knowledge and insights from the data. Also, comparing the products sold to the particular countries and the corresponding returns gained resulted in powerful insights that actually led to more questions and the need for further research.