National College of Ireland

An analysis of e-commerce purchase behaviour across the UK and Brazil

MSc Research Project
Data Analytics

Amol Upadhyay
Student ID: x20110812

School of Computing
National College of Ireland

Supervisor: Dr Giovani Estrada

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Amol Upadhyay
Student ID:	x20110812
Programme:	Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Dr Giovani Estrada
Submission Due Date:	15/12/2022
Project Title:	An analysis of e-commerce purchase behaviour across the UK and Brazil
Word Count:	6499
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	12th January 2023

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

An analysis of e-commerce purchase behaviour across the UK and Brazil

Amol Upadhyay
x20110812

Abstract

E-commerce has vastly evolved since CompuServe's 1969 launch. Technological advances and changing global economic circumstances will continue to boost internet trading. By 2023, online shopping might contribute for 22% of global retail sales, up from 14% in 2017. It is expected that by 2024, digital wallets will conduct more than 50% of internet payments. Different geographical locations are expected to share some common e-commerce patterns, and also possess unique characteristics. We analyzed two public e-commerce databases, one from the UK and one from Brazil. Recency, Frequency, and Monetary Value (RFM) analysis was performed to both datasets, as well as K-Means clustering to contrast the client pools. We found out a number of interesting customer behaviors. For instance, the purchase of common items, number of high-profile customers in terms of revenue, preferred shopping day, and share of loyal customers. We expect our findings are of interest to any e-commerce company from the UK expanding into Brazil, and vice versa.

1 Introduction

The growing global population and quick development of technology have led to fundamental changes in the retail market's buying decision-making process. Consumers face economic, social, political, cultural, and/or technical forces everyday. Millions of customers throughout the globe use the internet to search and make transactions. E-rapid shopping's growth provides companies and customers opportunity to develop and become more lucrative. The retail sector strives to accommodate changing customer needs and is impacted by technology. This condition changes regularly. E-commerce expands economic share in Europe, notably during the coronavirus epidemic (Kleisiari et al.; 2021).

E-commerce is the UK's fastest-growing retail industry. Rapid internet growth has led to extensive online sales, making the UK a leader in e-commerce. Surveys have led to major articles on the geo-demographics of e-commerce consumption. Geography is crucial in analyzing e-commerce sales, particularly when dis-aggregating by age and socioeconomic class. Fewer studies have analyzed retail organizations' online sales geography. This research examines e-commerce geography in UK food retailing. It will only cover business-to-business and business-to-consumer transactions. The study uses freshly obtained data from a large UK food store to explore the whereabouts of their online users (referred from here on as "partner data"). This will allow us to expand on prior survey-based analyses to study online spending drivers more thoroughly.

In this regard, Brazil provides a fascinating case study of the effects and effects of e-commerce and the local conditions that influence its acceptance and impact. An enormous developing nation, with some economically advanced and other economically disadvantaged sectors (Tigre and Dedrick; 2004). This research aims to examine the factors that influence consumers' decisions to make a purchase online in a developing economy (Wagner Mainardes et al.; 2019). Three hundred and forty-five e-commerce consumers were surveyed to get insight into the same. Inadequate customer trust was shown to have a detrimental impact on purchase intent, whereas positive influences included website identification and quality. Perceived risk and customer reluctance to innovation are seen as good influences, whereas prior experience is seen as negative. Further, the results show that customer resistance to innovation has a positive influence on perceived risk, whereas experience has a negative effect.

In this study, we tried to draw a comparison between the customer purchase behaviour patterns of the UK and Brazil. We have done a detailed Exploratory data analysis on two separate datasets. For Brazil, we have obtained 9 datasets from the Olist public dataset which is available on Kaggle. And for the UK, we have used a dataset of a UK non-retail online retail website. This UK dataset is also available in the public domain on UCI Library. The dataset of UK consist of 8 columns and Brazil dataset consists of 40 columns. We have also done RFM analysis for customer segmentation. Recency, frequency, and monetary value are abbreviated as RFM (Birant; 2011). Recency, frequency, and monetary investment (RFM) analysis is a marketing strategy that looks at consumer habits including how recently (recency) a consumer has made a purchase, how frequently (frequency), and how much (investment) they spend (monetary). Customers may be divided into subsets for use in future personalized services, and those most likely to react to marketing can be singled out for special attention.

1.1 Research Question

What are the similarities and differences of e-commerce purchase behaviour between Brazil and the UK customers

1.2 Research Objectives

1. Understanding the purchase behaviour of the customers across UK and Brazil.
2. Customer Segmentation using K-Means clustering on both the respective e-commerce markets.
3. Recency, Frequency and Monetary (RFM) analysis for drawing the comparison of the type of customer pool present in the UK and Brazil.
4. To get an insight into the products which are more likely to be bought in tandem with one another in both regions, using Market Basket Analysis.
5. Understanding the parallels and contrasts in client buying behaviour between the E-commerce markets of Brazil and the United Kingdom

2 Related Work

This section consist of all the relevant related work for our study. It is focus on the importance of understanding and improving customer experience.

2.1 The China example

As Internet technology has advanced, online shopping has evolved into a new kind of commerce channel. It ushers in a brand-new trade ecosystem and manner that varies greatly from the standard avenues of commerce. E-commerce is growing, and so does the number of people who purchase online. Therefore, the research on the efficiency and pricing strategy of Internet counterfeiting behaviour has significant practical value and provides insight into how to optimize the shopping environment of the Internet platform and the desire of Chinese customers to make purchases. Consumers' choices to buy on an e-commerce site are heavily influenced by their estimations of a product's worth and convenience to them, although the items' actual quality and usefulness also play a role. As a result of the e-commerce platform's effective and sufficient information display manifesting the quality of products, consumer trust in online shopping has increased, and the platform has contributed to a good global setting for the holistic development of Online economy, particularly in this age of mobile Internet, when innovation, knowledge exchange, and transmission are all developing at an incredible rate and exponentially disseminating. Therefore, in today's information era, product reputation on e-commerce marketplace is more crucial than in the conventional commerce model. In order to maintain growth, a company's infrastructure must prioritize the superiority of its wares (Wu et al.; 2020).

2.2 UK and EU E-Commerce Market

Different countries in the European Union have widely varying levels of e-commerce expertise and adoption. As in previous years, the United Kingdom, Germany, and France had the highest levels of internet sales in 2013. (51.1 billion). Sixty-one per cent of all sales in Europe and sixty-nine per cent of all EU sales came from these three regions, totalling $221.6 billion. With a total annual turnover of 96.2 billion euros, the British market is by far the biggest in Europe. France and Germany's combined online retail markets of $45 billion and $50 billion is almost identical. In contrast, in 2012, online sales in the United States reached an estimated $226 billion. The United States and the United Kingdom may anticipate a substantial growth (Żurek; 2015).

2.3 What role the last mile play the Brazilian E-Commerce market

The last mile delivery concept in the e-commerce market is a subset of urban freight transport (UFT) that entails "a number of actions and procedures that are essential for the delivery process from the last transit point to the ultimate drop point of the delivery chain" (Yuen et al.; 2018). It can be analyzed from the viewpoints of the people who use it (the "demand" side), the people who provide the services necessary for it to function (the "supply" side), and the people who regulate its physical environment (the "environment" side) (Bandeira et al.; 2018). People involved in last-mile deliveries

in urban areas have varying priorities and standards. E-shoppers (receivers) prioritize businesses (retailers) that can get them what they want quickly and cheaply; authorities (local government) prioritize the common good over the specific interests of businesses; shipping firms compete to provide the best possible service at the lowest possible price, regardless of the impact on the environment or traffic; and e-commerce services (shippers) compete to meet the specific needs of each customer. A sustainable strategy for city logistics of last mile delivery that takes into account the perspectives of key stakeholders (local government, shippers retailers, residents, receivers (e-customers), transportation providers/courier, express, parcel companies (CEP)) is required to address these issues.

2.4 Marketing tools which are relevant to this research

2.4.1 Use of RFM analysis in understanding customer behaviour

The RFM model provides an effective measure for the analysis of customers' consumption behavior, where three variables, namely, consumption interval, frequency, and money amount are used to quantify a customer's loyalty and contribution. The RFM model also provides an effective measure for the analysis of customers' contribution. Customers are able to be grouped into distinct groups based on the RFM value, and the information obtained from these groups is highly helpful in making decisions about the market (Dursun and Caber; 2016). Customers are first segmented using the Recency, Frequency, and Monetary (RFM) marketing analysis technique, and then again using the suggested expanded RFM analysis method with one extra parameter, termed Count Item. Based on the findings from the comparison of these methods, the clustering result is not affected by the addition of count Item as a new parameter to the RFM technique, hence CLV is computed using the weighted RFM method for each segment. The company's marketing and sales strategy may be better explained by looking at the computed CLV for various categories (Khajvand et al.; 2011).

RFM Ranking: Use of RFM Score for customer segmentation

Here, segmentation is done using behavioral data since it's readily accessible and changes over time and with purchases. RFM (Recency, Frequency, and Monetary) analysis evaluates consumers' purchase behavior. Recency, frequency, and monetary are scored. Finally, the scores of all three factors are aggregated as RFM score ranging from 555 to 111 (Haiying and Yu; 2010), which is utilized to anticipate future patterns by studying current and previous client histories. Recency, frequency, and monetary ratings are closely linked to customer lifetime and retention. Once the values of recency, frequency, and monetary have been computed, the K-Means technique is employed to cluster the customer base in accordance with the variables. The characteristic of each cluster is studied to identify the client segment that generates the most profit for the business((Christy et al.; 2021).

2.4.2 Scope of Cohort analysis in understanding E-Commerce Consumer behaviour

As described by (Fedushko and Ustyianovych; 2022), Cohort analysis has several applications in the fields of medical, pharmacy, sociology, and the environment. Even though it is widely used in e-commerce, not all company officials are able to evaluate the data and pay insufficient attention to the offered statistics on user behavior. In the ecommerce

industry, cohort analysis offers significant room for improvement in terms of utility, explanability, and investigation. Customer loyalty correlates with e-commerce representative credibility. It is characterized as an emotional connection between the company and the customer. The more a customer's loyalty, the more likely they are to engage with a certain e-business. Social commerce is intrinsically linked to consumer happiness, brand recognition, and brand loyalty for a particular e-business. It involves establishing customer confidence in e-commerce, media platforms, social aspects, and users.

2.4.3 The Pareto rule in Marketing

As per (Kruger; 2011), Pareto Principle is described as,a universal law that anticipates the link between inputs and outputs and even controls how consumers affect your business's profit is summarized by the 80/20 rule. The 80/20 rule states that a small percentage of your clientele will account for a disproportionately large share of your revenue (in this case, 80%) and be 16 times more lucrative than the rest of your clientele. Best Market Approach Applying this universal rule in business is done in three phases, which are shown by the 80/20 Rule.

- ✓ Divide your client base into lucrative segments and identify the 20% of your customers that are the most profitable.
- ✓ Decide to make this market sector your target audience.
- ✓ Put this target market's needs at the center of your strategic marketing strategy.

2.4.4 Market Basket analysis: The Concept

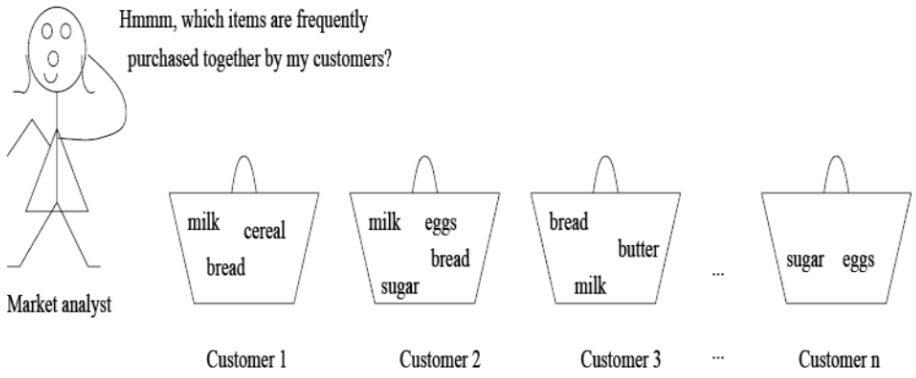


Figure 1: Market Basket Analysis. Source[1]

Figure 1 is an example of a Market basket analysis. This is an ideal illustration of association rule mining. It is a reality that all managers in shops and department stores would want to learn more about the purchasing habits of each individual consumer. This market basket analysis technique will assist managers in determining the sets of products that consumers are likely to buy. This analysis may be performed on all client transaction data collected by retail establishments. These outcomes will drive their marketing or advertising strategy planning. For instance, managers will be able to suggest new shop layouts with the assistance of market basket research. On the basis of this study, goods

that are often bought together might be put near together in order to encourage their sale together. If customers who buy computers are likely to also purchase anti-virus software, putting the hardware display near the software display will increase sales of both products (Dhanabhakyam and Punithavalli; 2011).

3 Methodology

This section consist of the methodology used in the research i.e., Knowledge discovery in database(KDD) with data mining tools such as RFM(Recencey, Frequency and Monetary) analysis and Market Basket Analysis using Apriori algorithm. Mentioned below are the steps used in methodology:

3.1 Knowledge discovery in database(KDD) as an effective data mining method

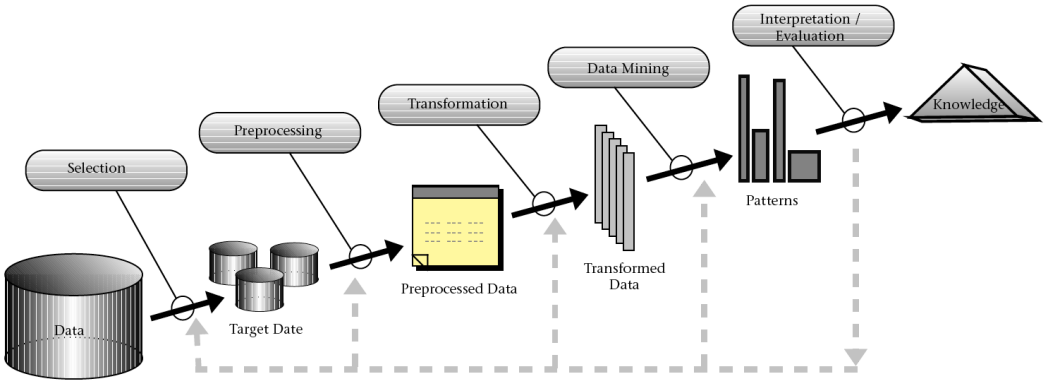


Figure 2: KDD Methodolgy as an effective data mining method. Source `https://infovis-wiki.net/w/images/4/4d/Fayyad96kdd-process.png`

As described and explained in the detail by HB (2012), the field of Knowledge Discovery in Databases, often known as KDD, was formally established with reference to the overarching idea of being comprehensive and comprehensive in one's goal of searching from data. After then, the concept of "data mining" came into being; data mining is a high-level application approach that is utilized to display and examine data for decision-makers. Whereas, a key component of knowledge discovery and development (KDD) is data mining, which uses algorithms to discover patterns and structures in data while maintaining reasonable computing efficiency.

Stages of KDD Methodology and its adaption in the research

The different stages of KDD is depicted in Figure 2, the explanation for the same, and also, how it is used in the research, is described below[2]:

[2] `https://www.datascience-pm.com/kdd-and-data-mining/`

- **Data Selection**: The targeted data and the factors that will be utilized to assess for knowledge discovery are both chosen based on how an existing database of collated data is acted upon. In our research we selected, one dataset for UK consisting of 8 features, and 9 separate datasets which consist of a total of 40 features.

- **Data Pre-processing**:At this step, the focus is on bettering the data that is being worked with, and the idea of data cleansing is included into the process. Following the same idea, we investigated and addressed the issues of null values, missing values and duplicate values in both datasets i.e. the UK and Brazil.

- **Data Transformation**:During this stage, the primary focus is on transforming the data that has been preprocessed into a format that can be used in its entirety. To do this, the scope is narrowed in terms of diversity, and the data qualities that will be used for the subsequent assessment are clearly determined. In this section, the material is arranged and classified, and it is often consolidated into a single category. Similarly, in our research, in order to consolidate the data, we merged the relevant columns, and renamed a few of them. We also tried to investigate a few anomalies in the data which we discussed in section 5.2.

- **Data Mining**:The data mining stage, which is the most well-known part of the process, consists of searching through the modified data for patterns of interest. This is the most well-known part of the process. These patterns are graphed, trended, and plotted in a format that is specifically beneficial to the process that the KDD is being undertaken for. We plotted different graphs in order to achieve our research objectives of drawing a comparison between the two different eCommerce markets. The methods of grouping, clustering, and regression are included into this phase of the process as part of the technique, with the selection of one (or more) of these methods depending on the output anticipated and sought from the process. The terms "association rules," "classification," "clustering," "prediction," and "sequence patterns" all refer to important data mining methods that have recently been created and used in various data mining initiatives HB (2012).
  We used the **Apriori algorithm** which seemed the best fit for performing market basket analysis, using the association rule, in our research. It has been suggested that the Apriori algorithm is among the most prominent algorithms for mining the frequent sets of items of Boolean association rules (Liu; 2010). The creation of an association rule mining algorithm may be broken down into two parts using this two-stage frequency-based approach:

  ✓ Find all collections that have more than the required number of "frequent items" to qualify.

  ✓ The above-obtained frequent set will serve as the basis for all the association rules that will be generated; specifically, for each frequent item set A, all the nonempty subset an of A will be identified if the ratio of support (A)/support(a) is larger than or equal to min-confidence. Meaning, employ the rules with confidence higher than the user-specified minimum confidence min-confidence, based on the frequent sets learned in the first step.

  K-Means clustering techniques were also used to identify different clusters of customers. This helped us in understanding the types of customers present in both

countries. We further carried out RFM(Retention, Frequency and Monetary) analysis for both the eCommerce markets. Improvements through RFM Analysis According to(Dawane et al.; 2021)RFM segmentation is said to be easy and effective. Some techniques and approaches are long-term, thus outcomes may be slow. We can enhance Segmentation based on requirement and data availability.

- **Data Interpretation and Evaluation**:After completing the previous step i.e., 'Data Mining', which was one of the most important parts for bringing out all the relevant customer behaviour patterns which now, would be used for drawing the comparison between these two distinct, but comparable markets. As it is well put in HB (2012), the advantages of data mining in marketing i.e., the practice of data mining may be of use to direct marketers by supplying them with helpful and accurate trends on the purchase behaviour of their clients. On the basis of these tendencies, marketers are able to more precisely concentrate their attention on their clients in their marketing efforts. Techniques like RFM Analysis and Market basket analysis are also of great help in customer segmentation and customer purchase behaviour analysis for the marketing domain. This section is equally important and may be of keen interest to direct marketers. We compared all the comparable results which we found during EDA, by keeping in mind the objectives of this research. The evaluation was also done on comparing results from RFM analysis and Market Basket Analysis which we found in Data Mining. Interpretations for the same, are discussed in Section 6.7 of this research.

4 Design Specification

This section talks about the design archetecture of the process which is followed in this research.

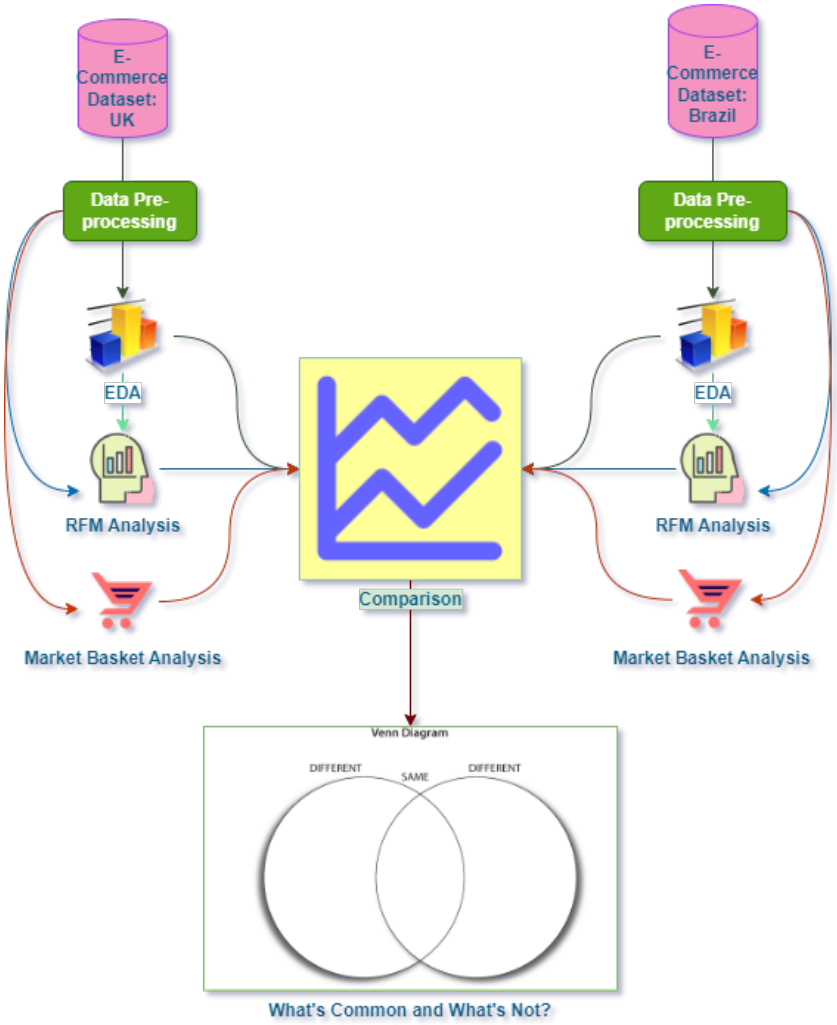


Figure 3: Design of the process flow

Figure 3 depicts the complete process which we followed during the course of this research. We have taken the data sets of two different E-commerce organizations based in UK and Brazil. The datasets for Brazil and the UK are publicly available on Kaggle and UCI website respectively. The UK data-set consists of a single CSV file whereas, Brazil data set consists of nine CSV files, which were merged for the analysis part. In Figure 3 it can be seen that there are two different datasets for each of the respective countries,used for the analysis. Then we moved on to data pre-processing for each of these data-sets. Here we cleaned the datasets and transformed them to get better outputs in the later stages of analysis.

Now when the data is cleaned and transformed we moved to the next phase of analysis i.e., Exploratory data analysis (EDA). Here we have visualised different features to have interesting insights from the data. Our focus was more on the comparative study for both countries. Although as the data sets for both these countries were different, containing different columns, we have also tried to understand different features ,even if, they are only present in one of the data sets. After visualising all the relevant features of both these datasets we did Recency,Frequency and Monetary(RFM) analysis on for each datasets to understand the different clusters or segments of customer present in both the countries, details of the finding for the same are mentioned in the Evaluation

6 section. After performing the RFM analysis, we then moved to the next analysis i.e., Market Basket Analysis. It was interesting to find out about the products in each of the respective countries which have the higher chances to be purchased together.We used Apriori algoritm to perform market basket analysis.
Later we compared all the results from the different finding we got from EDA,RFM and Market Basket analysis, in order to understand the customer purchase behaviour patterns and also about the different segments of customers in each of these countries. Our goal was to achieve our research objectives1.2.

5 Implementation

This section explains in detail about the process and the techniques which are implemented in this research. Figure 3 is the pictorial description about the architecture of the process which we took during the course of this study. We will try to touch on each section of the implementation process, but mostly we will try to focus more on the core process to have an understanding of the nitty-gritty of the techniques used. Below mentioned, are the sections of the process and techniques used for the implementation:

5.1 Selection of the Data-set and its description:

To start with, we have used two different data-sets for the comparative study. **UK data-set** consist of 8 Columns and more than 540000 rows (Refer Figure 4). This transnational data set is of a UK-based E-commerce company.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom

Figure 4: Structure of the UK E-commerce Dataset

The dataset is a publicly available dataset available on UCI website[3]. The **Brazilian dataset** consist of 40 columns (Refer Figure 5) after merging all the nine sub-datasets. The Brazilian E-Commerce dataset is also publicly available on Kaggle website[4].

5.2 Data Pre-processing (Data Cleaning and Transformation)

After importing the dataset into the data-frame, using Python through jupyter notebook, we imported all the relevant libraries such as, pandas,numpy,seaborn,matlplotlib.pyplot which are general to more specific ones like,Kmeans,KElbowVisualizer,seasonal_decompose, stats models.api,sklearn.metrics, etc for the UK dataset and scipy,urllib, unidecode, etc for the brazilian dataset. The data cleaning and data transformation is carried separately on both datasets, which are described below:

- **For UK Dataset:** We checked for the null values and found that, 1454 null values were present in "Description" column, and for all these null values in "Description",

[3]Dataset Source `https://archive.ics.uci.edu/ml/datasets/online+retail`
[4]Dataset Source `https://www.kaggle.com/datasets/olistbr/brazilian-ecommerce`

```
Index(['order_id', 'customer_id', 'order_status', 'order_purchase_timestamp',
       'order_approved_at', 'order_delivered_carrier_date',
       'order_delivered_customer_date', 'order_estimated_delivery_date',
       'customer_unique_id', 'customer_zip_code_prefix', 'customer_city',
       'customer_state', 'order_item_id', 'product_id', 'seller_id',
       'shipping_limit_date', 'price', 'freight_value',
       'product_category_name', 'product_name_lenght',
       'product_description_lenght', 'product_photos_qty', 'product_weight_g',
       'product_length_cm', 'product_height_cm', 'product_width_cm',
       'product_category_name_english', 'payment_sequential', 'payment_type',
       'payment_installments', 'payment_value', 'seller_zip_code_prefix',
       'seller_city', 'seller_state', 'review_id', 'review_score',
       'review_comment_title', 'review_comment_message',
       'review_creation_date', 'review_answer_timestamp'],
      dtype='object')
```

Figure 5: Brazil E-commerce Dataset features

there were no available "customer IDs" present and instead of a proper "UnitPrice", 0 was present. We dropped all these NaN values. After removing all these unknown customers we were left with 400000 observations.

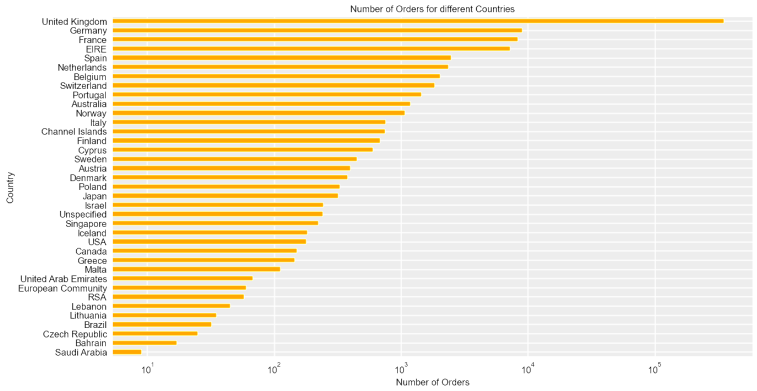
Then to check for more missing values like "NAN", "na", "?", "Unknown", and so on, we searched across the dataset with any string value less than 5, and found that in "InvoiceNo" there are few invoices starting with "C", which denotes cancelled invoices, we investigated little more into it and found out that there were no particular pattern for cancelled transactions. Similarly we checked for "StockCode" and 'Description' and found that, few of the stock codes such as "17107D", have more then one description, after further investigation we saw that they have a very little difference between them i.e., "," or "/", which means probably the same.

- **For Brazil Data-set:** As we had in total nine data-sets for the Brazilian E-Commerce company, that means the data pool was huge and lot of things can be done on this dataset. But in order to address our research objectives we need to filter the relevant information for this research analysis. To do that we build different data frames for each datasets. Then before merging them we change columns names as they consist of similar values e.g. 'customer_zip_code_prefix' and 'geolocation_zip_code_prefix' are both renamed as 'zip_code'. Then we finally joined the datasets with relevant columns. Then we checked for the null values and found that there were none for the Customer id. Although, Because the dataset is per order, there are duplicate 'customer unique id' values. Where 'client id' is the order dataset's primary key. Each order has a unique 'client id'. And that's how we did the data selection and transformation.

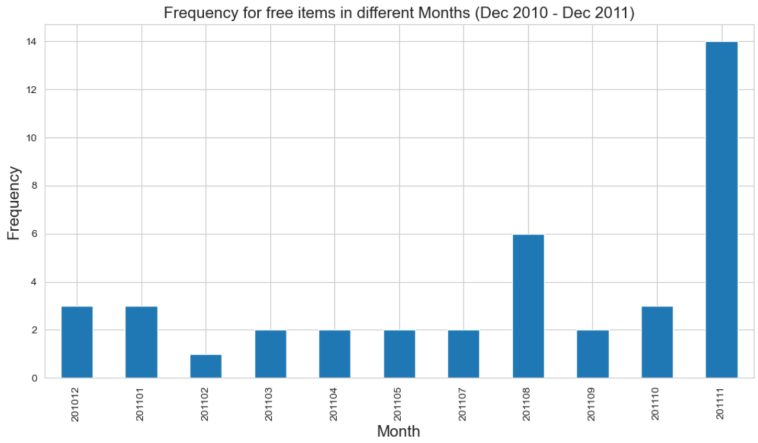
5.3 UK and Brazil E-Commerce datasets: The Implementation

We have done extensive EDA on both the data sets. But for the comparative study, we will try to focus more on those visualizations that are relatable and comparable. For most of the graphs in the EDA section we have used seaborn plots. We tried to Visualise maximum numbers of features with different pictorial representations, in order to understand the data. Lets focus on the visualization which are although not comparable due to the discprepancy in the features of both the datasets, but they might be relavent in regards to the customer behaviour patterns. Below are the different graphs for each of the data sets **which might be not comparable but might be relevant for Industry point of view**.

For UK Data-set: All the relevant,but Non-Comparable figures:



(a) Orders distribution across EU



(b) Distribution of free items(with 0 unit price) across the year

Figure 6: (a) Highest number of orders are from UK,as the E-Commerce firm is based is UK; (b)Most of the free items are sold in the year end

Figure 6(a) is showing the distribution of Orders across different countries. As the company is based in UK, around 350000 orders are from the UK out of the total number of orders. Figure 6(b) is the distribution of free items. As it can be see clearly that most of them are being distributed in the end of the year.

For Brazil Data-set: All the relevant, but Non-comparable figures:
Figure 7(a) shows the distribution of orders across the map (open source[5]) of Brazil. We used geo-location data which was available in 'zip_code','geolocation_city',ǵeolocation_state', 'geolocation_lat', and 'geolocation_lng', and plot-ed it using 'matplotlib.pyplot' library. Figure 7(b) is depicting top 10 states in terms of number of order. We used 'customer_state' and count id 'order_id' as an input and visualized it using 'seaborn' library. Figure 7(c) is showing the distribution of orders in a day. We used 'order_purchase_timestamp' and 'order_id' as an input and visualized this figure using 'seaborn' library.Figure 7(d) is the bar graph for the distribution of customers reviews for the products. We used 'countplot' function from the 'seaborn' library in python to visualize this graph. Figure 7(e)(Top 10 cities in terms of revenue generation) shows the distribution of the percent of total

[5]Image Source `https://i.pinimg.com/originals/3a/0c/e1/3a0ce18b3c842748c255bc0aa445ad41.jpg`

payments across cities. We took 'customer_city, payment_value as an input features and used 'seaborn' library to populate the graph.Figure 7(f) is depicting that 8.7%(357) of cities contribute to 80% revenue generation for the company. Figure 7(g) and Figure 7(h) shows the top 10 highest and the lowest cities in terms of time taken to deliver a product. For this also we used 'seaborn' library in python.

Comparable visualizations in EDA for UK and Brazil: The Implementation

We have used different marketing techniques such as RFM Analysis , Pareto Priciple(80/20 rule) and market basktet analysis to understand the patterns for both the markets, but we have also tried to compare the 'somewhat' similar features and insights we got from the EDA. Below mentioned are implementation for all the respective comparable feature between both the countres:

- **Top 10 Most selling products/product categories:** We visualised top 10 most selling products in UK and Top 10 most selling product categories of Brazil. We used 'stock_code','description' and the sum of the 'quantity' as an input, visualized a barplot using 'matplotlib.pyplot' to get top 10 most selling products in UK. Similarly, 'product_category_name_english', and 'order_id' as an input feature for the Brazil dataset.

- **Top 10 Customers by amount paid :** We mapped UK's top 10 customers by transactions and Brazil's top 10 by total paid value. We utilized 'customer_id' count and value as input, and visualized a barplot using 'matplotlib.pyplot' and'seaborn' to retrieve UK's top 10 customers by transaction count. Similarly, 'payment_value' and 'customer_unique_id' were input features and visualized to produce top 10 customer(ID)s by total amount paid for the Brazil dataset.

- **Order Distribution across days of week :** We mapped distribution of orders across the all the 7 days of a week, for both the datasets.

- **Pareto Principle-80/20 rule**: According to the Pareto principle, the most influential factors in most situations have a very little impact (about 20% of the total). That is to say, very few factors account for a disproportionate share of the final outcome[6]. We applied Pareto rule on UK dataset by creating a function called 'create_pareto_plot'. This funtion takes the commulative percentage as one of the core feature and slice the data into 80%. In Brazilian dataset we used another technique to visualize a similar figure. We took in consideration the proportion of customers who are contributing most in the total revenue of the company.

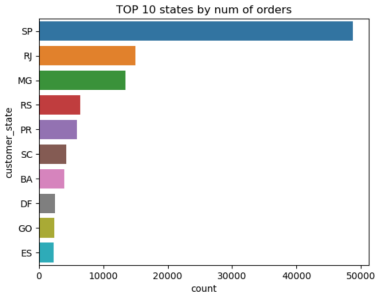
- **RFM Analysis for both the datasets:** Below are the steps taken for the implementation of RFM Analysis[7]:

    1. First, give each client Recency, Frequency, and Monetary values. **Recency** refers to the customer's most recent transaction (most businesses use days, though for others it might make sense to use months, weeks or even hours instead). **Frequency** is total customer transactions (during a defined period). Total client spending is **Monetary** (during a defined period).

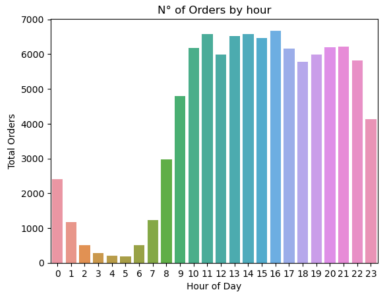
[6]Kindly refer to the link `https://asana.com/resources/pareto-principle-80-20-rule/`

[7]Kindly refer to the link`https://www.optimove.com/resources/learning-center/rfm-segmentation/`

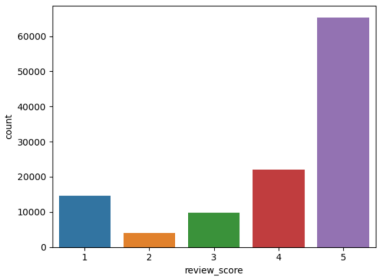
(a) Orders distribution across Brazil



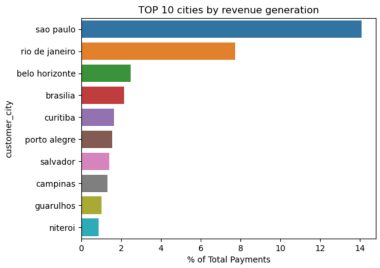
(b) Top 10 states in Brazil in terms of number of orders



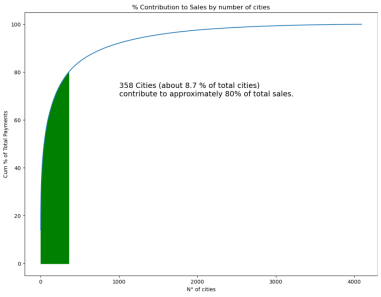
(c) Distribution of orders across the hours of a day



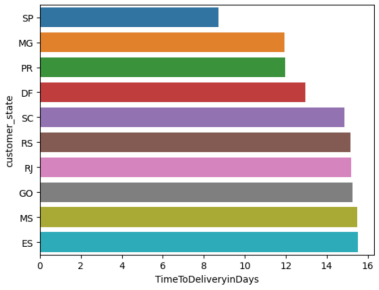
(d) Distribution of of review score by customers from 1 to 5



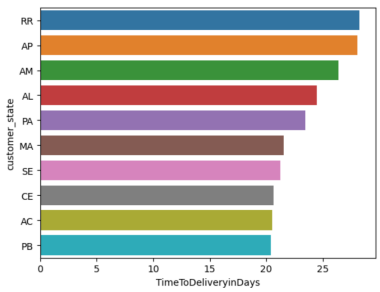
(e) Top 10 Cities in terms of revenue generation



(f) 8.7% of cities contribute aproximately 80% of total sales



(g) Top 10 cities having the lowest product delivery time(in days)



(h) Top 10 cities having the highest product delivery time(in days)

Figure 7: All the visualizations which are not directly comparable with ones in UK, but might be helpful in understanding the behaviour patterns

2. Using python, partition the customer base into tiers for R, F, and M. Software may do k-means cluster analysis, creating groups of clients with similar traits.
3. The third stage is to choose consumer groups who will get certain sorts of messages depending on the RFM clusters in which they appeared. It is useful to assign these clusters with appropriate names.

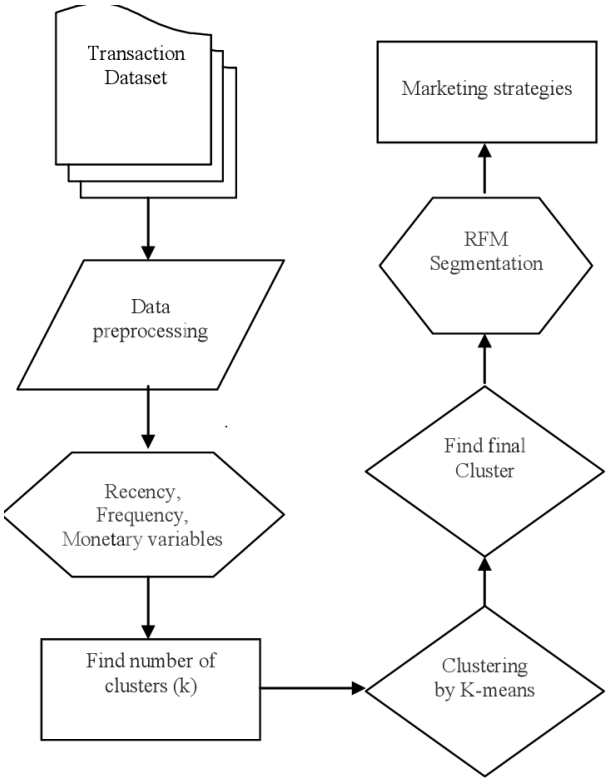


Figure 8: RFM Implementation Process flow 5.3

[8]

- **Apriori Algorithm for Market Basket analysis:**[9] The figure below depicts how the Apriori Algorithm begins by constructing the smallest itemset and then expands from there.

  procedure begins by constructing an itemset using the Join Step, that is, by combining K-itemsets into (K+1) itemsets. In the first iteration, the algorithm creates Cookie, Chocolate, and Cake, for instance. method then immediately moves on to the Prune Step, which consists of removing any candidate item set that does not match the minimal support criteria. For instance, the algorithm will eliminate Cake if Support(Cake) is below the minimal Support value.

[10]

[8]Figure 8 `https://www.semanticscholar.org/`

[9]Kindly refer to the link`https://towardsdatascience.com/data-mining-market-basket-analysis-with-apriori`

[10]Figure 9 `https://miro.medium.com/max/1100/1*oGmHkz3QXn-Dxf7WZeuYSg.webp`

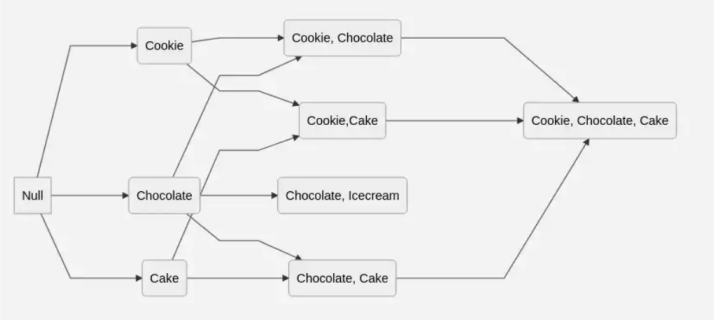
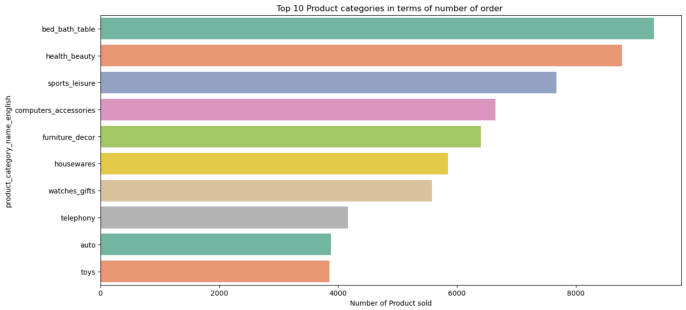


Figure 9: Apriori Algorithm Impementation concept 5.3

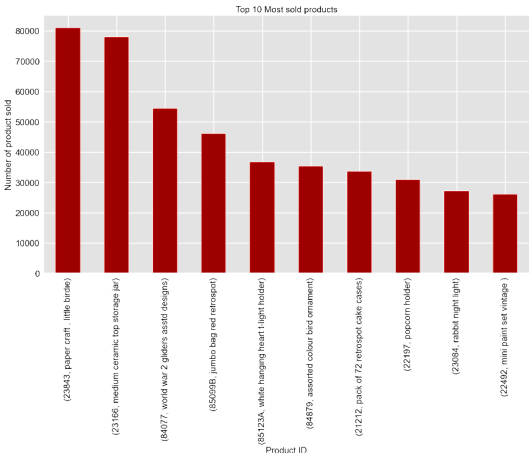
6 Evaluation

This sections talks about the findings and outcomes of this research. All the relevant comparable studies are carried out in the 6 different Case studies mentioned below. Discussion section talks about the interpretations of these findings.

6.1 Case Study 1: The Top selling product/categories across UK and Brazil



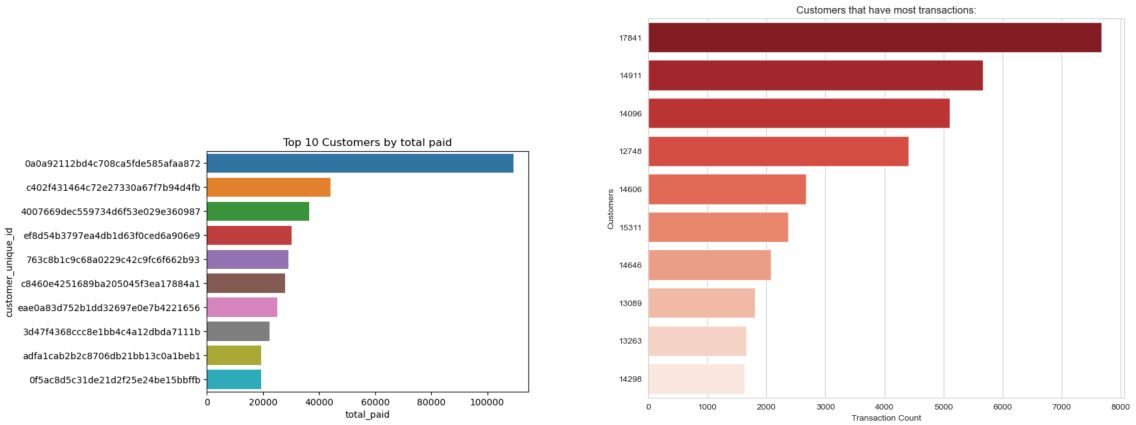
(a) Top 10 most selling product categories in Brazil



(b) Top 10 most selling products across UK

Figure 10: Top 10 Most selling Product/ Product category across UK and Brazil

6.2 Case Study 2: The top spending customer across the UK and Brazil

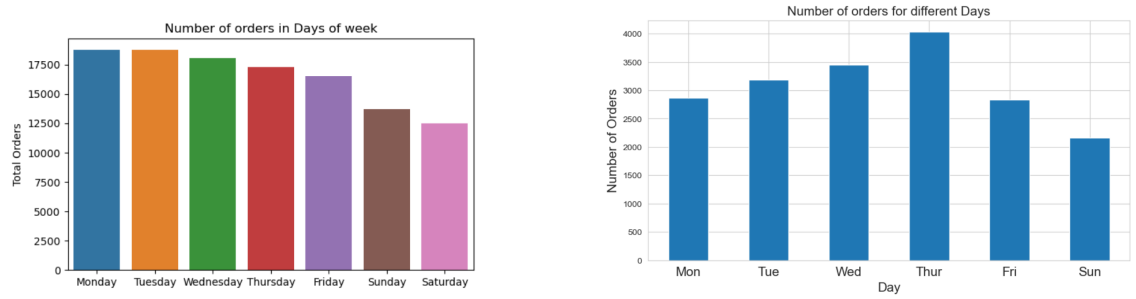


(a) Top 10 most spending customers across Brazil

(b) Top 10 most customers with highest number of transactions across UK

Figure 11: Top 10 most spending customers across UK and Brazil

6.3 Case Study 3:Most purchasing days in a week for Brazil and UK

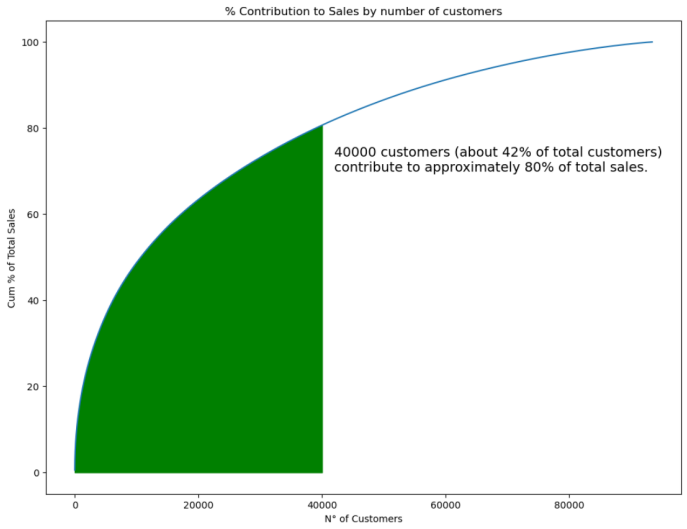


(a) Distribution of number of orders in a week in Brazil

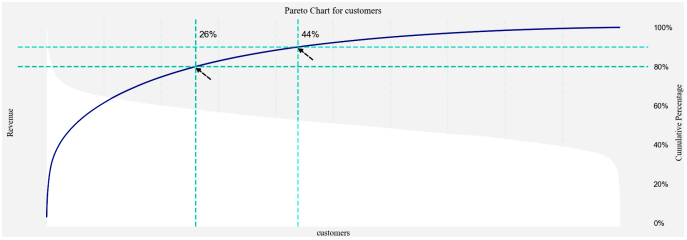
(b) Distribution of number of orders in a week in UK

Figure 12: Weekly order volume across the UK and Brazil

6.4 Case Study 4: The Pareto Rule outcomes for UK and Brazil



(a) Percentage of the influential base of the customers across Brazil



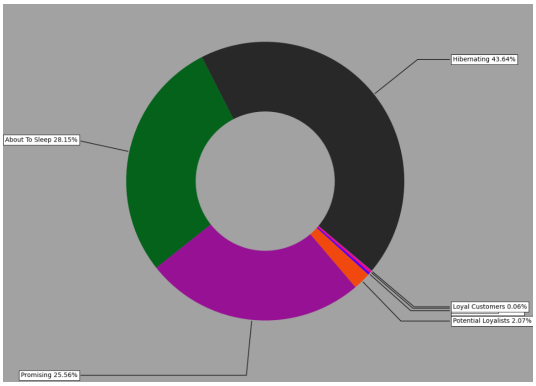
(b) Percentage of the influential base of the customers across UK

Figure 13: Most influential customer for revenue generation in percentage terms in UK and Brazil

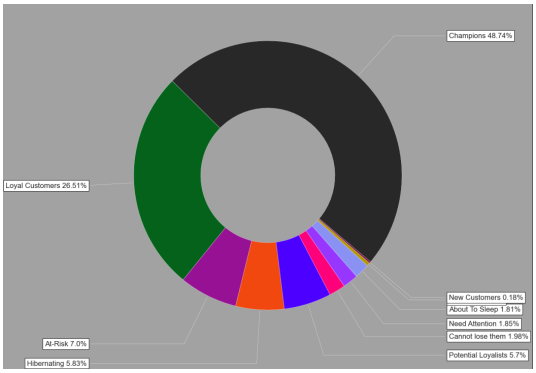
6.5 Case Study 5: Clusters of Customers in UK and Brazil

(a) Customer Clusters for Brazil-(Using rcParams)

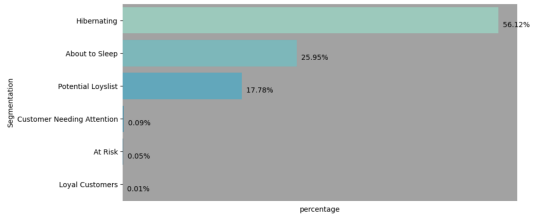
(b) Customer Clusters for UK-(Using rcParams)



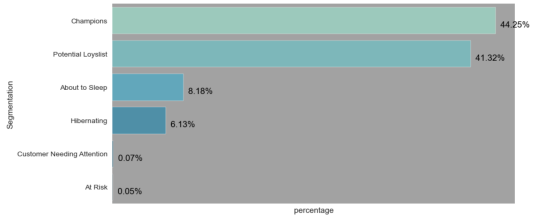
(c) Customer Clusters for Brazil-(Using Pie Chart)



(d) Customer Clusters for UK-(Using Pie Chart)



(e) Customer Clusters for Brazil-(Using bar plot)



(f) Customer Clusters for UK-(Using bar plot)

Figure 14: Customer clusters with different graphs for UK and Brazil

6.6 Case Study 6: Items which are mostly brought together

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
86	(53759a2ecddad2bb87a079a1f1519f73, 0bcc3eeca39...	(368c6c730842d78016ad823897a372db)	0.000331	0.014414	0.000166	0.5	34.689655	0.000161	1.971173
88	(0bcc3eeca39e1064258aa1e932269894, 368c6c73084...	(53759a2ecddad2bb87a079a1f1519f73)	0.000166	0.012757	0.000166	1.0	78.389610	0.000164	inf
92	(422879e10f46682990de24d770e7f83d, 0bcc3eeca39...	(389d119b48cf3043d311335e499d9c6b)	0.000166	0.013917	0.000166	1.0	71.857143	0.000163	inf
94	(0bcc3eeca39e1064258aa1e932269894, 389d119b48c...	(422879e10f46682990de24d770e7f83d)	0.000166	0.017396	0.000166	1.0	57.485714	0.000163	inf
98	(53759a2ecddad2bb87a079a1f1519f73, 0bcc3eeca39...	(389d119b48cf3043d311335e499d9c6b)	0.000331	0.013917	0.000166	0.5	35.928571	0.000161	1.972167
100	(0bcc3eeca39e1064258aa1e932269894, 389d119b48c...	(53759a2ecddad2bb87a079a1f1519f73)	0.000166	0.012757	0.000166	1.0	78.389610	0.000164	inf

(a) Products which are likely to be bought together in Brazil

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.043900	0.039802	0.030957	0.705185	17.717202	0.029210	3.256952
1	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.039802	0.043900	0.030957	0.777778	17.717202	0.029210	4.302452
2	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG PINK POLKADOT)	0.072841	0.055086	0.030632	0.420536	7.634188	0.026620	1.630668
3	(LUNCH BAG PINK POLKADOT)	(LUNCH BAG RED RETROSPOT)	0.055086	0.072841	0.030632	0.556080	7.634188	0.026620	2.088574
4	(JUMBO BAG RED RETROSPOT)	(JUMBO BAG PINK POLKADOT)	0.093197	0.052680	0.032908	0.353105	6.702899	0.027999	1.464412
5	(JUMBO BAG PINK POLKADOT)	(JUMBO BAG RED RETROSPOT)	0.052680	0.093197	0.032908	0.624691	6.702899	0.027999	2.416152
6	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG RED RETROSPOT)	0.064646	0.072841	0.031478	0.486922	6.684737	0.026769	1.807051
7	(LUNCH BAG RED RETROSPOT)	(LUNCH BAG BLACK SKULL.)	0.072841	0.064646	0.031478	0.432143	6.684737	0.026769	1.647164

(b) Products which are likely to be bought together in UK

Figure 15: Market Basket Analysis-Products which are more likely to be purchased together in UK and Brazil (Using Association rules of Apriori Algorithm)

6.7 Discussion

As we have got all the relevant outcomes for this comparative resarch for UK and Brazil E-commerce cutomers. Let's dive into the interpretation of these results. Will try to address the research objectives.

1. **Case Study 1** (Refer Figure 10) is the visualization for top most selling products or product categories in UK and Brazil.In the graph we can see that 'bed_bath_table' is the top most selling category in Brazil and product 'paper craft,little birdie' is the most selling product of UK, after a closer look into the graph it is quite evident that 'housewares' being a common product/product category in both UK and Brazil customers. Category 'housewares' comes at the 6th most selling product category in Brazil, and in UK, 'medium ceramic top storage jar', 'retrospot cake cases' and 'popcorn holder' are all moreover comes in the category of housewares. All these product are among the top most selling products in UK. **Which tells us one thing for sure, that category 'housewares' is popular among the users of UK and Brazil. (Limitation):** As product name is not available in the Brazilian dataset it was not possible to find out the list of exact top selling products, so, instead of that we used product categories for the same.

2. **Case Study 2**(Refer Figure 11) is the pictorial graph for the top most spending customers in UK an Brazil. Here we can see that, customer with customer ID ending with '..aa872' followed by '..4d4fb' and '..60987', are the top three customers in terms of spending, in Brazil dataset. Similarly, customers with cutomer ID, '17841',14911 and '14096' are the highest spender in the UK market. What is evident here that **the top most customers in both these graphs are contributing and generating the maximum chunk of revenue for both the markets.**

3. In **Case Study 3** (Refer Figure 12)we are visualizing the distribution of the orders across week for both UK and the Brazilian market. It is clear from the graph

that **'Mondays' and 'Tuesdays' are the shopping days for the users living in Brazil, whereas, in the Uk 'Thursday' is the day, when most of the customers do shopping.**

4. In **Case Study 4** (Refer Figure 13) We have applied Pareto rule on the datasets and we have found that **26% users in the UK and 42% users in Brazil contribute to 80% of revenue generation for the company. For Brazil 42% is quite a large number of users contributing to the revenue generation for the company, which is not good for the market, because it is telling us that the average purchase for the customers is not high.**

5. **Case Study 5** (Refer Figure 14) is visualized with different graphs and techniques. We have used RFM segmentation and K-Means clustering technique with the standard marketing segmentation rule for all three figure-comparison in Figure 14.Below are the interpretations from all the sub-figures of Figure 14:
   Comparison of Clusters using rcParams (Figure 14(a) and Figure 14(b)): We used one of the standard rfm coordinates rules to populate these graphs. In Figure 14(a) which is for Brazil, the highest chunk of customers i.e., 43.8% are in the 'Hibernating' cluster followed by the 'About to Sleep' cluster, consisting of 28.91%. The third-largest cluster is of 'Promising customers' which is 26.23%. This tells us that the majority of the customers are not active and are not frequent buyers of the Brazilian E-Commerce platform. Similarly, Figure 14(b) is for the UK, the biggest pool of customers consists of 'Hibernating', which is 24%, followed by 'Loyal customers' , 'Champions', and ' Potential Loyalist' consisting of around 50%. This tells us that most of the customers in the UK market are loyal and actively participate in the eCommerce purchase ecosystem.
   Comparison of Clusters using Pie-Chart (Figure 14(c) and Figure 14(d)): For the pie chart comparison, we made another RFM table by grouping the segment values present in the rfm table and aggregating with the mean of recency, frequency and monetary values for each segment with respective customer unique id, order to find the monetary value per segment. In Figure 14(c) which is for Brazil, the biggest pool of customers consists, of 'Hibernating' and 'About to sleep' customers which is around 70% and then followed by 'Promising' customers consisting of around 25%. This again tells us that the large pool of customer in Brazil are mostly in a dormant state. Figure 14(d), which is for UK, we see slightly better results than from the previous comparison. Here the biggest chunk of customers is 'Champions' and 'Loyal Customers' consisting of around 75%, which is followed by 'Hibernating' and 'At risk' customer pool of around 15%. This again confirms firmly that the customer base in UK is more loyal and are proactively participating in the eCommerce ecosystem.
   Comparison of Clusters Using Bar-graph (Figure 14(e) and Figure 14(f)): This comparison was done using K-Means clustering techniques. In Figure 14(e), the largest pool of customer segments i.e., around 80% is of 'Hibernating' and 'About to Sleep' customers which is then followed by 'Potential Loyalist' with around 18%. Which is similar to our previous finding and tells us that the customer pool in Brazil is not frequent purchasers. In Figure 14(f), which is for UK, the biggest customer segment is of 'Champions' and 'Potential loyalist' consisting of more than 80% , which is followed by 'About to sleep' and 'Hibernating' category consisting of around 15%. This again confirms that the customers in the UK are frequent

purchasers and are very much try to stick to the E-Commerce firm.
After analysing all the figures, it is evident that **the users in UK are more loyal than those present in Brazil. (Limitation):** Segmentation was done on the basis of standard marketing rules for customer segmentation. These rules may change on the basis of findings. As it is a comparative study, we used similar rules on both the datasets so that, they can be compared.

6. **Case Study 6** (Refer Figure 15) In this study we are trying to understand that which products are likely to be purchased together in UK and Brazil. This we achieved using the Association rule of Apriori Algorithm 5.3. In Brazilian dataset, **product with product ID starting with '53759..' and '368c6..' are likely to be purchased together and so on. Similarly in the UK, 'Roses regency teacup' and 'Green regency teacup' are likely to be purchased together by the customer.** We talked about the values present in the first rows of these association tables in Figure 15 This rule is also applicable on the rest of the products available in the table. **(Limitation):** Since, product name was not available for the Brazilian dataset, we cannot find, exactly what are those products which are possibly be purchased together, on the basis of product id.

7 Conclusion and Future Work

Customer shopping behaviour patterns are essential in understanding the market of any given country. Our research revolves around the objectives 1.2 of this paper:

- ✓ Most of our research talks about the customer behaviour patterns. e.g. the top selling product category is telling us that 'housewares' is one of the most selling product category in both the countries.
- ✓ We have achieved the second objectives related to customer segmentation and we have idetified that the customers are more loyal and are likely to stick with the company based in UK than, from the customers of Brazilian market.
- ✓ We have performed RFM analysis on both the datasets and segmented the customers into different clusters for both the markets.
- ✓ We performed market basket analysis using association rule of Apriori algorithm and figured out about the products in both the markets which are are likely to be purchased together.
- ✓ Both these markets are different in terms of culture, population, size and in so many other parameters, but both these markets have quite a lot of opportunity for an investor. Someone who has invested in one of these markets and is thinking to expand their business to either of these countries, for them this research could help in understanding the customer pool and parallels and contrast between these two countries before investing.

There is ample scope to extend this research, to bring out more interesting insights of the customers from both these nations. Due to the time constraint and limitations we have missed few interesting techniques such as "Customer Lifetime Value" (CLV) analysis.

We can also apply different Machine learning models on both these datasets. Models like FB Prophet and CatBoost can provide interesting predictions on "Churn rate". There is a scope of performing sentiment analysis specially on the Brazilian dataset which has a lot more features than the UK one.

References

Bandeira, R. A., D'Agosto, M. A., Ribeiro, S. K., Bandeira, A. P. and Goes, G. V. (2018). A fuzzy multi-criteria model for evaluating sustainable urban freight transportation operations, *Journal of cleaner production* **184**: 727–739.

Birant, D. (2011). Data mining using rfm analysis, *Knowledge-oriented applications in data mining*, IntechOpen.

Christy, A. J., Umamakeswari, A., Priyatharsini, L. and Neyaa, A. (2021). Rfm ranking–an effective approach to customer segmentation, *Journal of King Saud University-Computer and Information Sciences* **33**(10): 1251–1257.

Dawane, V., Waghodekar, P. and Pagare, J. (2021). Rfm analysis using k-means clustering to improve revenue and customer retention, *Proceedings of the International Conference on Smart Data Intelligence (ICSMDI 2021)*.

Dhanabhakyam, M. and Punithavalli, M. (2011). A survey on data mining algorithm for market basket analysis, *Global Journal of Computer Science and Technology* .

Dursun, A. and Caber, M. (2016). Using data mining techniques for profiling profitable hotel customers: An application of rfm analysis, *Tourism management perspectives* **18**: 153–160.

Fedushko, S. and Ustyianovych, T. (2022). E-commerce customers behavior research using cohort analysis: A case study of covid-19, *Journal of Open Innovation: Technology, Market, and Complexity* **8**(1): 12.

Haiying, M. and Yu, G. (2010). Customer segmentation study of college students based on the rfm, *2010 International Conference on E-Business and E-Government*, IEEE, pp. 3860–3863.

HB, B. K. (2012). An overview of data mining, *Oriental Journal of Computer Science and Technology* **5**: 69–73.

Khajvand, M., Zolfaghar, K., Ashoori, S. and Alizadeh, S. (2011). Estimating customer lifetime value based on rfm analysis of customer purchase behavior: Case study, *Procedia Computer Science* **3**: 57–63.

Kleisiari, C., Duquenne, M.-N. and Vlontzos, G. (2021). E-commerce in the retail chain store market: An alternative or a main trend?, *Sustainability* **13**(8): 4392.

Kruger, E. R. (2011). *Top Market Strategy: Applying the 80/20 Rule*, Business Expert Press.

Liu, Y. (2010). Study on application of apriori algorithm in data mining, *2010 Second international conference on computer modeling and simulation*, Vol. 3, IEEE, pp. 111–114.

Tigre, P. B. and Dedrick, J. (2004). E-commerce in brazil: Local adaptation of a global technology, *Electronic Markets* **14**(1): 36–47.

Wagner Mainardes, E., de Almeida, C. M. and de Oliveira, M. (2019). e-commerce: an analysis of the factors that antecede purchase intentions in an emerging market, *Journal of International Consumer Marketing* **31**(5): 447–468.

Wu, C.-H., Yan, Z., Tsai, S.-B., Wang, W., Cao, B. and Li, X. (2020). An empirical study on sales performance effect and pricing strategy for e-commerce: from the perspective of mobile information, *Mobile Information Systems* **2020**.

Yuen, K. F., Wang, X., Ng, L. T. W. and Wong, Y. D. (2018). An investigation of customers' intention to use self-collection services for last-mile delivery, *Transport Policy* **66**: 1–8.

Żurek, J. (2015). E-commerce influence on changes in logistics processes, *LogForum* **11**(2): 129–138.