# Online Retail Company Dataset

This is a multinational data set which contains all the transactions happening between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail. The company mostly sells exclusive all-occasion gifts with many wholesaler’s customers.



Table 1: Raw Data

# Data Analysis, Understanding and Pre-processing

The Data is analyses and pre-processed in order to make further algorithms. This is an important step as raw data will give unconditional results. The quality of the data is checked before applying any machine learning algorithms. The referred dataset comprises of 55K rows and 8 features columns. The dataset has been analysed and count against all the countries has been calculated. The same is shown in Appendix 1

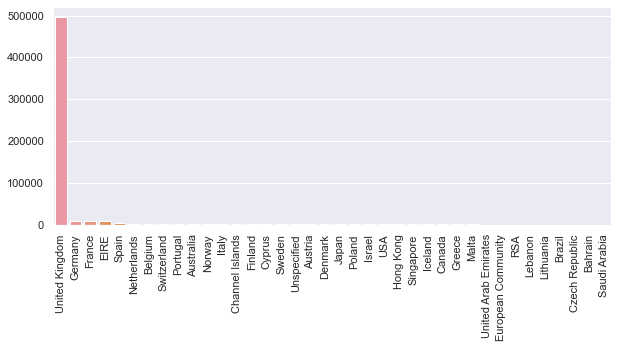
The data includes the following features:

* InvoiceNo
* StockCode
* Description
* Quantity
* InvoiceDate
* UnitPrice
* CustomerID
* Country

# Statistical exploration

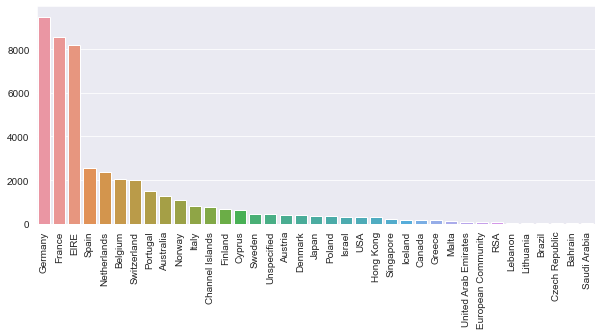
Statistical exploration states to the initial step in data analysis where data is visualized with statistical techniques to describe dataset characterizations like as volume, quantity, and accuracy. This gives better understanding of the data. Below is an overview for the country-level online retail.

The below graph shows the percentage of orders from the 40 countries, sorted by the number of orders. This displays as more than 90% of orders are coming from United Kingdom (UK) and no other country even makes up 3% of the orders in the data.



Graph 1: Number of orders in each country

Just to heck the data over other countries, United Kingdom is dropped and below visualisation is generated.



Graph 2: Number of orders in each country **(without UK)**.

Dataset patterns are spotted in the report using the Correlation. A positive correlation result refers to the increase in relation of both variables, whereas a negative correlation means one variable is inversely proportional to another variable. Correlation has a very vital role in calculating the degree at which one domain tend to diverge with variations in the data points of another domain.

# Transformation & Elimination of Redundant Variables

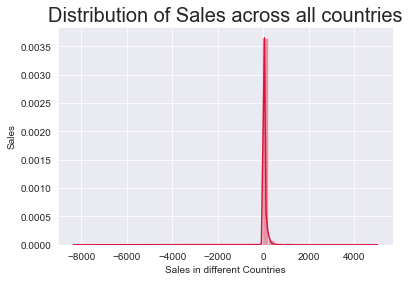
In order to make use of any data, statistical treatment is very essential. To draw relatable and valuable conclusion, data treatment is necessary. Python offer lots of methods which can perform many operations on a value. In technical terms, report has considered data in objects, integer and float values. At required point, string data type has been dropped out. (Appendix 2)

* Most Invoices look as if as normal transactions with positive quantity and prices still there are some prefixed “C” or “A” which denote unalike transaction types.
* InvoiceNo are required to be broken down into Number and Code
* Empty or Null values are present
* Negative values are present

The data set has many missing values and outliers which cannot justify the required results. Hence the dataset was trimmed by removing the unwanted values. Null values were dropped within the dataset and outliers were removed. (Appendix 3)

# Data visualisation

The graphical representation of information and data for the given dataset has been identified using statical method and Kmeans to design clustering. To understand trends, outliers, and patterns in data, below graph showcase the distribution of sale across all the countries.



Graph 3: Distribution of Sale across all countries

# Management of missing values

Data comprise of many cancelled transactions like Receipt No with ‘C’, numerous NANs (Null) in Customer Id and zero value in UnitPrice, All the missing values are operated using the following SQL statement -

*#Qoute*

Cleandata\_set = pd.read\_sql (''' SELECT customerid,

MAX(invoicedate) AS last\_order\_date,

COUNT(\*) AS count\_order,

SUM(unitprice\*quantity) AS totalprice

FROM data

WHERE InvoiceNo NOT LIKE '%C%'

AND customerid IS NOT NULL

AND unitprice != 0

GROUP BY customerid ''', conn)

Cleandata\_set.to\_sql ("cleandata\_set", conn)

*#unQoute*

# RFM Segmentation

Based on marketing principles, the RFM model helps businesses to distinguish between existing and new customers. The RFM model even helps to create a personalized and relatable understanding of user behaviour. With the application RFM model, report has identified the criteria based on the transaction history of customer’s transaction.

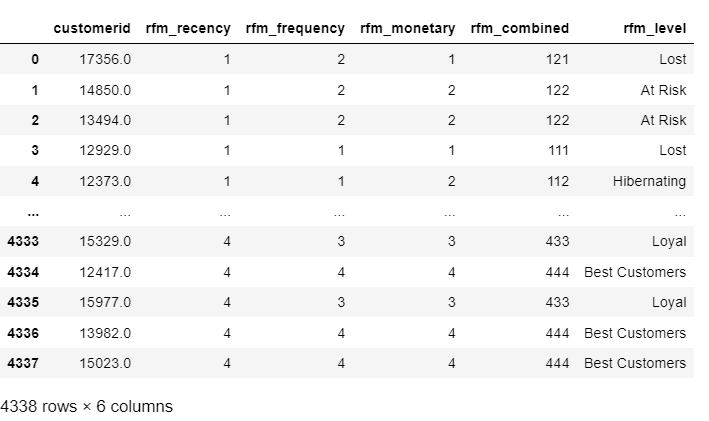


Table 2: RFM Segmentation

RFM factors demonstrate below facts:

* The customer is more responsible to the most recent purchase
* The customer satisfaction is based on the frequent buying of product

# RFM Level

Customer KPIs involves analysis about active customers count and segregation of new and old customers in terms of revenue. In the report, the entire customer list is divided into tiered groups against all the three dimensions (R, F and M). Dividing the customer list, customer insights have been observed. The model supported to gain perspective on what their brand means to the existing customers.

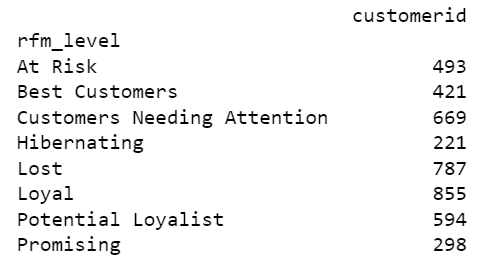


Figure 1: RFM Level

# RFM Segments Mapping

RFM analysis has revealed the data anomalies that helped to understand the most important groups of customers acquisition costs against the margin they generate. It’s an automatically display of customer groups such as Loyal, Active, At Risk, and Lost customers and easy to identify the pattern. Below Segmentation has divided potential customers into distinct groups over which the business can have a firm discussion.

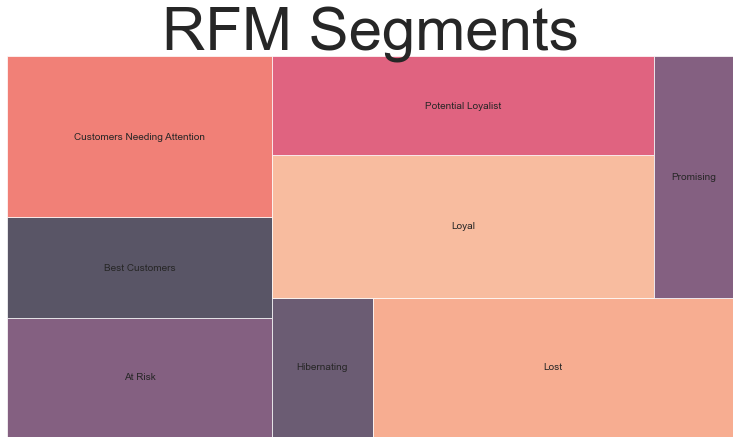


Figure 2: Mapping of RFM Segments

# K-Means Model & Justification

Once the data is set up, run\_kmeans() has been assigned against each customer to a cluster based over their recency, frequency, and monetary value. Further groupby() is used to examine each cluster and agg() to compute some statistics for examining the scope of the clusters and the mean values within.

In this report K-Means clustering as an unsupervised learning method has taken into consideration to identify a specific number of clusters. The steps followed are:

* Report has aim for K=8 and arbitrarily, the number of clusters has been set.
* To normalise the R, F, M values of data, StandardScaler method is used.
* Optimise K value based on manifold runs for a range of Ks and assess using the Elbow graph
* Reviewing final customer segments solution

Below is the formation of clusters which give the idea over the customer data. The cluster formation is based on Unit Price and Quantity.

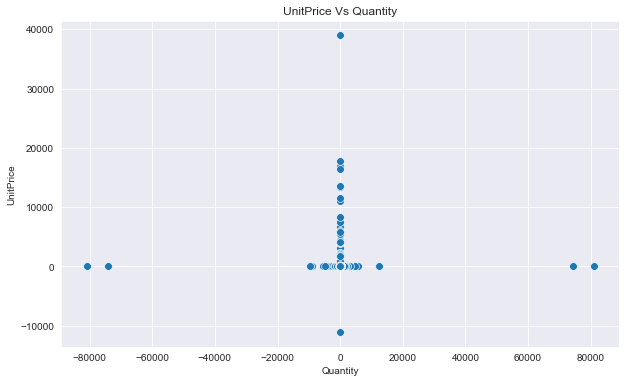


Figure 3: Cluster

In this report the model defines customer segments based on a count of unlike features and is used to assign each group with a different cluster label. Further the Within Cluster Sum of Squared Errors (WSS) has been calculated for different values of k. The k for which WSS first starts to reduce, is value of K giving the best count of clusters to make from the raw data. The elbow graph with x-axis being the number of clusters and number of clusters is taken at the elbow joint point.

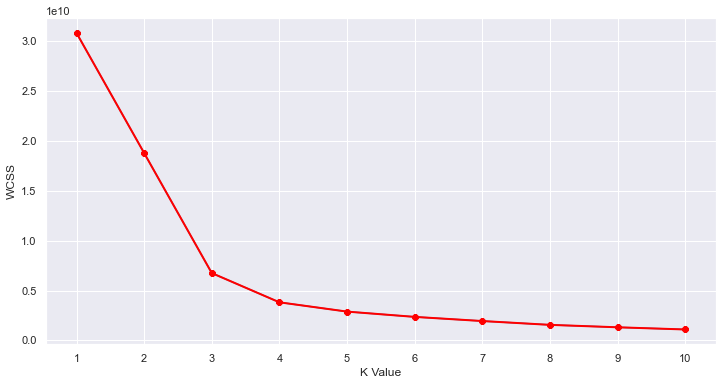
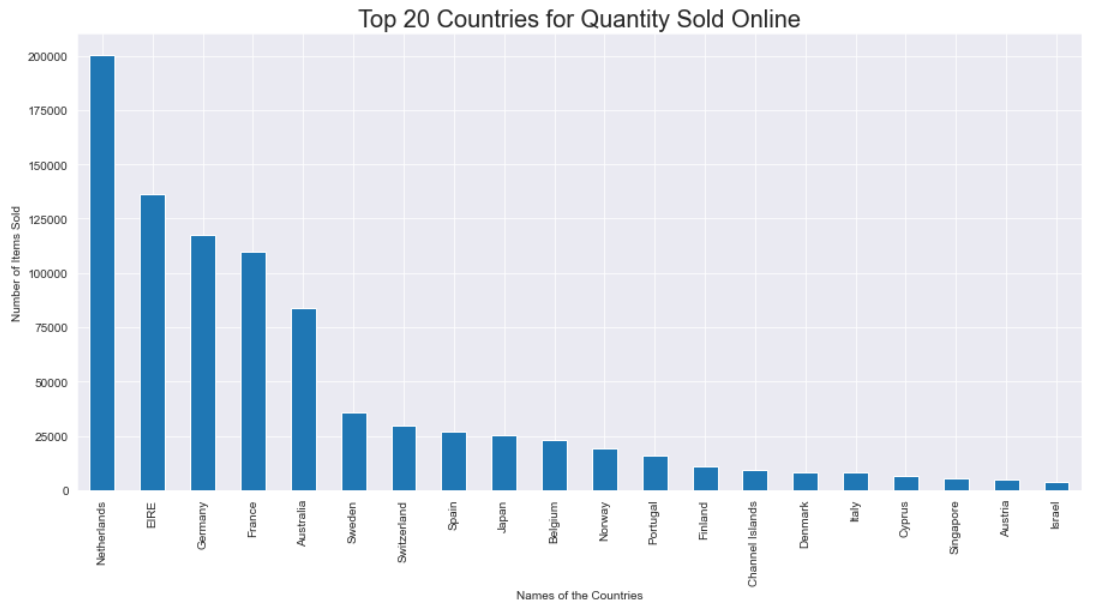


Figure 4: Clusters

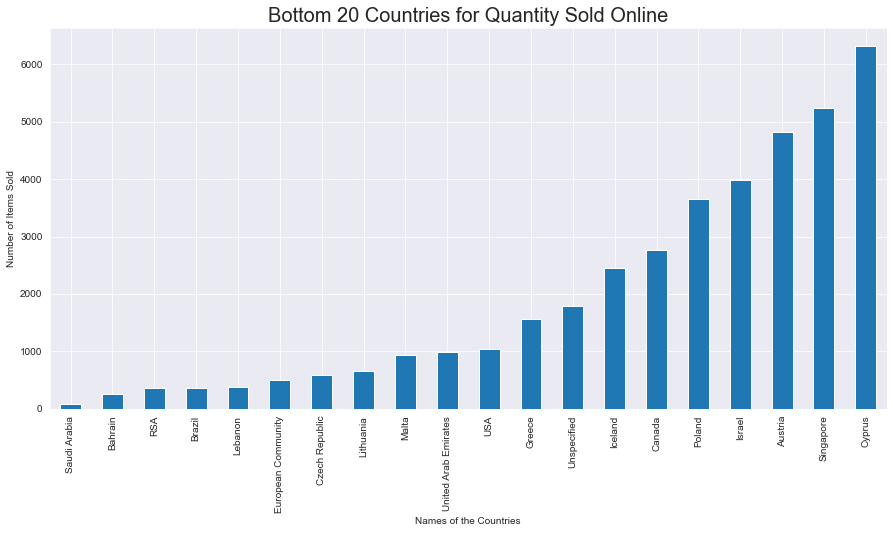
# Review of Results

Below is the general review obtained over the data set.

* The customer with highest number of orders comes from the United Kingdon
* The customer with highest money spent on purchases comes from Netherlands
* Top 5 countries (including UK) that place the highest number of orders are as follow → United Kingdom, Netherlands, Ireland (EIRE), Germany, France



Graph 4: Top 20 countries for Quantity Sold Online



Graph 5: Bottom 20 countries for Quantity Sold Online

Customer value here is defined as getting what is desired by receiving quality for what is paid and receiving somewhat in return for what is give. It is also defined as the satisfaction level received by the customer in respects of the money, data, time and knowledge. The customer value is seen in terms execution acknowledged in return of the payment made for the product or service.

# Identification & Justification of Business Values

This RFM and Kmeans is simple & reasonable model, by helping increase in revenue by targeting specific groups. One size doesn’t have to fit all means those who have bought frequently are likely to continue doing the same and by giving them an incentive or loyalty/referral program to encourage recurrent purchasing.

The goal of K means is achieved by grouping the data points into distinct and non-overlapping subgroups. Application of K means clustering is for the segmentation of customers and to get a better understanding to result the pattern into high revenue of the company

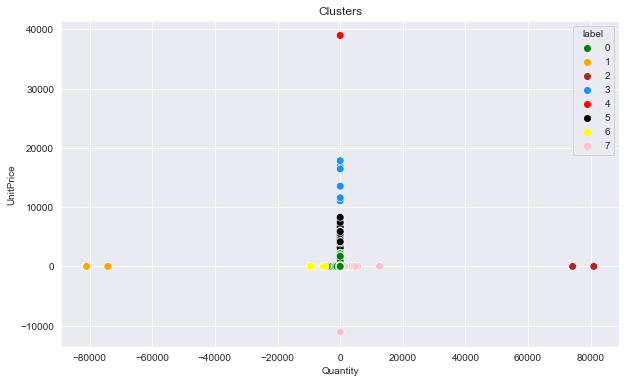


Figure 5: Coloured Clustering to Identify the Customers

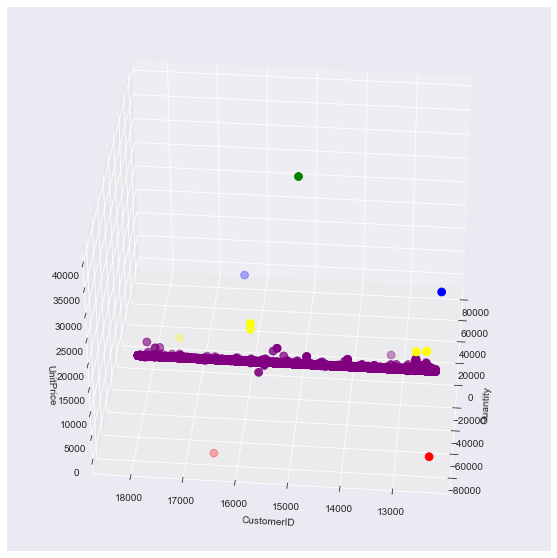


Figure 5: 3D Clustering to Identify the Customers

# Data Mart Design

Data mart is a simple form of data warehouse focused on a solitary subject or say a particular line of business. Data mart can provide access to all the data to gain insights much faster and since teams don’t have to spend too much time searching within a more complex data warehouse or even to do manual aggregation of data from different sources.

# Identification & Justification of Dimensions

The entire report is dedicated to the transnational data set which has all the details of transactions for online store across multiple countries. The calculated dimensions for the correct scenario are represented in the below image.

Stock code, Invoice Number, Invoice Date, Customer ID and Country are the selected dimensions for the transnational online retail data. These dimensions are great enough to provide all the required insights of the data mart. Majorly, the country and invoice date are used to identify the pattern of sale across the countries.

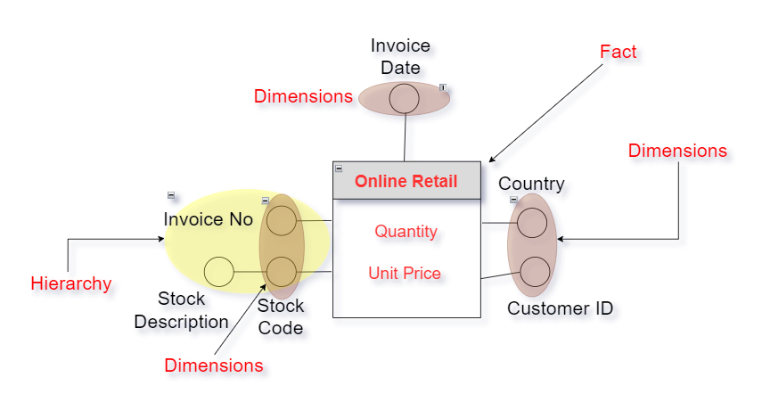


Figure 6: Identification of Dimensions

# Identification & Justification of Measures

The measure for the transnational online retail data are quantity and unit price. Below is the fact design

Unit price and quantity helps in identifying the total revenue for the online retaining. Across any given country, invoice or date the revenue can be calculated using the information from quantity of the products and unit price of the stock. Amount spent = Quantity \* Unit price

A customer-oriented company always keen to know much more about customer needs and ensure to deliver as per their expectation. It is important for them to develop deeper relationships in order to get successful business.

The selected Dimensions are Stock code, Invoice Number, Invoice Date, Customer ID and Country. These give direct justification for from where the business is coming and what this the relation of customers with the company.

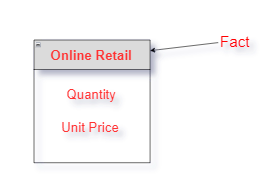


Figure 7: 3D Identification of Fact and Measure

Insights from customer segmentation are cast-off to develop marketing campaigns and for designing strategies.

# Further analysis

Key consideration by any company is the segmentation of customer which helps company to understand how its customers behave in terms of purchasing and what is important to them. All these information can be used to design personalized content for different customers. Studies have found that customers like such individual attention and will more likely to respond to buy the product. Following analysis can also help to get more details about the customers -

* Analysing the number of days by when the first transaction was made by each customer. This will help to understand how long the customer was in the system.
* Steering profounder segmentation on customers based on the geographical location.
* Using the data from Google Analytics account to track many important business metrics and incorporate them in your business values such as Customer Life time value, traffic source, Page visit and view, exit rate on website etc