Online Retail Data Analysis

Submitted by:

- Neeta Choudhry
- Cohort 8-DA

Objective

- Ecommerce is growing very fast and changing the way people shop for products and services.
- Analyze the online retail dataset to understand customer behavior, identify sales trends, and improve business strategies
- Suggest relevant products to customers based on their purchase history and preferences, enhancing the personalized shopping experience and increasing sales.

Stakeholder

- **Marketing Teams** benefit from the synergy of RFM analysis, enabling precise customer segmentation, and sales trend analysis, providing valuable insights into evolving consumer preferences, allowing for tailored marketing campaigns and strategies that boost sales and engagement.

Research Question

What are the sale trends over the period?

What are the most frequent purchased product

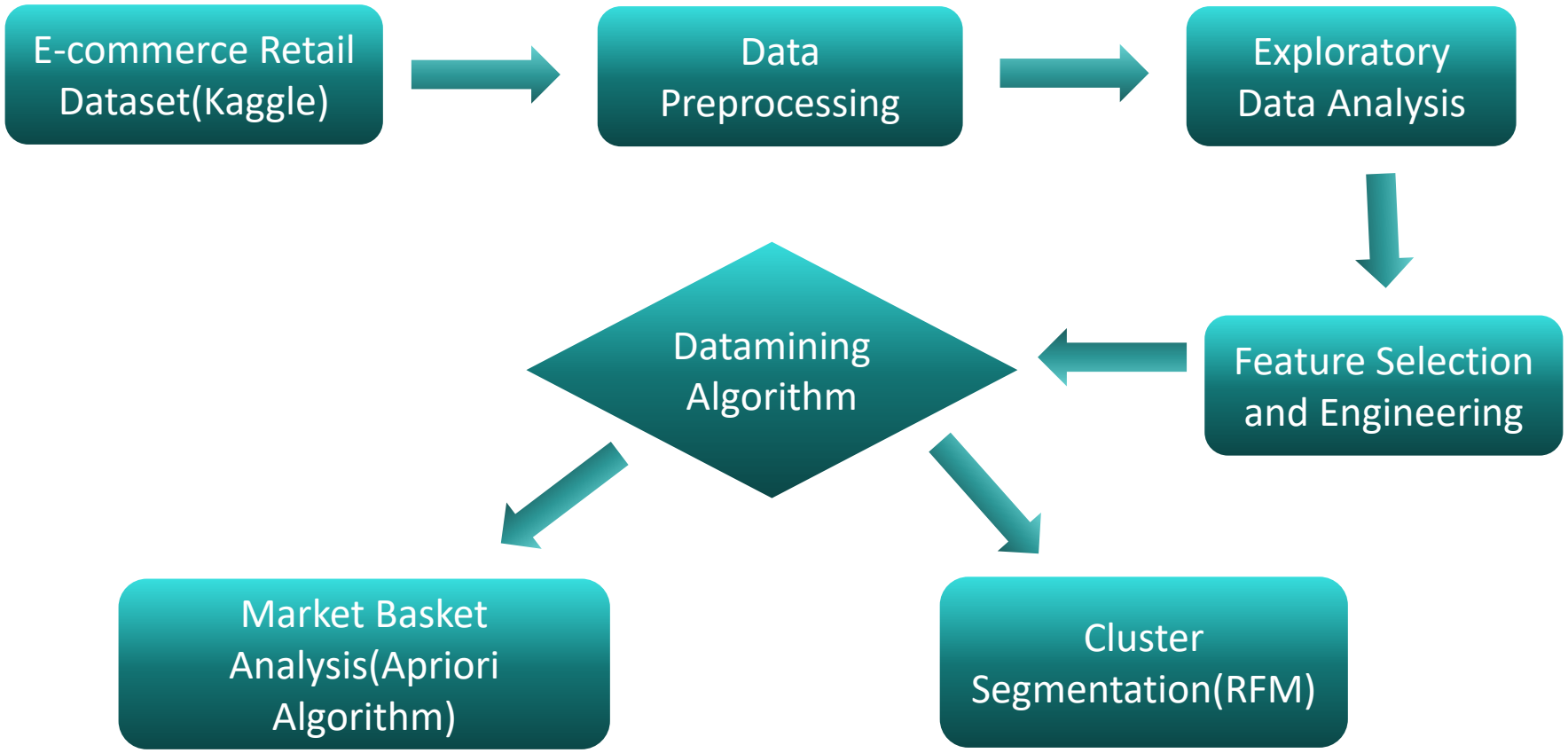
Any association between products

Can customer be divided into groups

Any strategy can increase the profit

What are the intervals between transactions

Flow Diagram



Dataset

UK-based ecommerce data for 2010-2011

Purchases from all over the world. Contains bulk orders

500k rows and 7 columns

Categorical:
InvoiceNo
StockCode
Description
CustomerId
Country

Numeric:
UnitPrice
Quantity

Letter ‘c’ cancelled orders

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	3.39	17850.0	United Kingdom

Data Preprocessing

We need to make sure the data is clean before starting analysis. As a reminder, we should check for:

- Duplicate records
- Consistent formatting
- Missing values (Remove the canceled order)
- Wrong values (Remove the transaction with negative quantity)

Data Preprocessing

- Total number of products, transactions and customers

```
[18]: pd.DataFrame([{'products': len(df['StockCode'].value_counts()),
                     'transactions': len(df['InvoiceNo'].value_counts()),
                     'customers': len(df['CustomerID'].unique().tolist()),
                    }], columns = ['products', 'transactions', 'customers'], index = ['quantity'])
```

[18]:

	products	transactions	customers
quantity	4070	25900	4373

- No. of Cancelled transaction:

```
n1 = df['order_canceled'].value_counts()[1]
n2 = df.shape[0]
print("Number of cancelled transactions:", n1)
print('Number of orders canceled: {}/{} ({:.2f}%) '.format(n1, n2, n1/n2*100))
df = df[df['order_canceled'] == 0]
```

```
Number of cancelled transactions: 9251
Number of orders canceled: 9251/536641 (1.72%)
```

Data Preprocessing

- Remove transaction with negative values for price and quantity.

```
Total number of transaction with negative quantity: 1336
Precentage of transactions with zero or negative quantity 0.25 %
Total number of transaction with negative Price: 2512
Precentage of transactions with zero or negative price 0.48 %
```

- Remove transaction with wrong values for description.

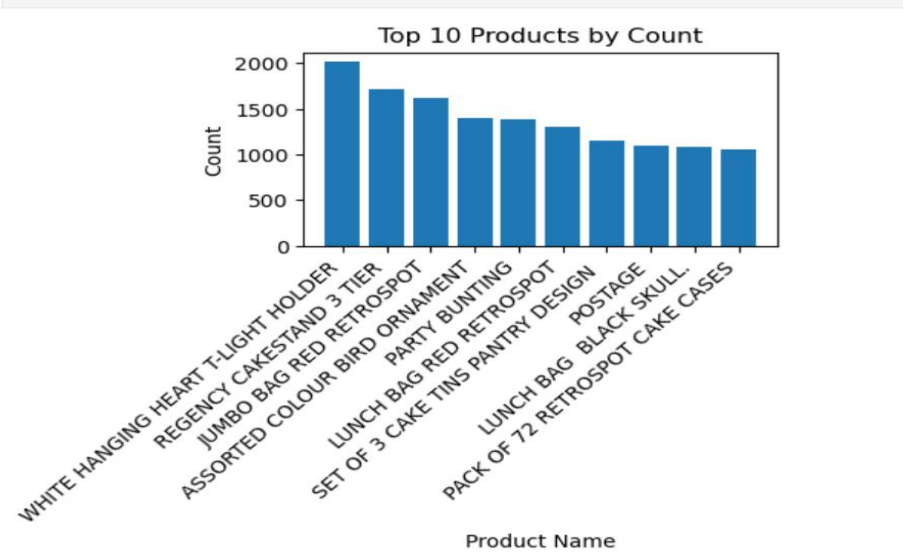
```
# Filter the DataFrame to display rows where 'InvoiceNo' column contains non-digit values
df[df['InvoiceNo'].str.isdigit() == False]
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Customer ID
299982	A563185	B	Adjust bad debt	1	2011-08-12 14:50:00	11062.06	NaN	United Kingdom	nan
299983	A563186	B	Adjust bad debt	1	2011-08-12 14:51:00	-11062.06	NaN	United Kingdom	nan
299984	A563187	B	Adjust bad debt	1	2011-08-12 14:52:00	-11062.06	NaN	United Kingdom	nan

Exploratory Data Analysis

- Top 10 product by count.

	Description	Count
0	WHITE HANGING HEART T-LIGHT HOLDER	2016
1	REGENCY CAKESTAND 3 TIER	1713
2	JUMBO BAG RED RETROSPOT	1615
3	ASSORTED COLOUR BIRD ORNAMENT	1395
4	PARTY BUNTING	1389
5	LUNCH BAG RED RETROSPOT	1303
6	SET OF 3 CAKE TINS PANTRY DESIGN	1152
7	POSTAGE	1099
8	LUNCH BAG BLACK SKULL.	1078
9	PACK OF 72 RETROSPOT CAKE CASES	1050

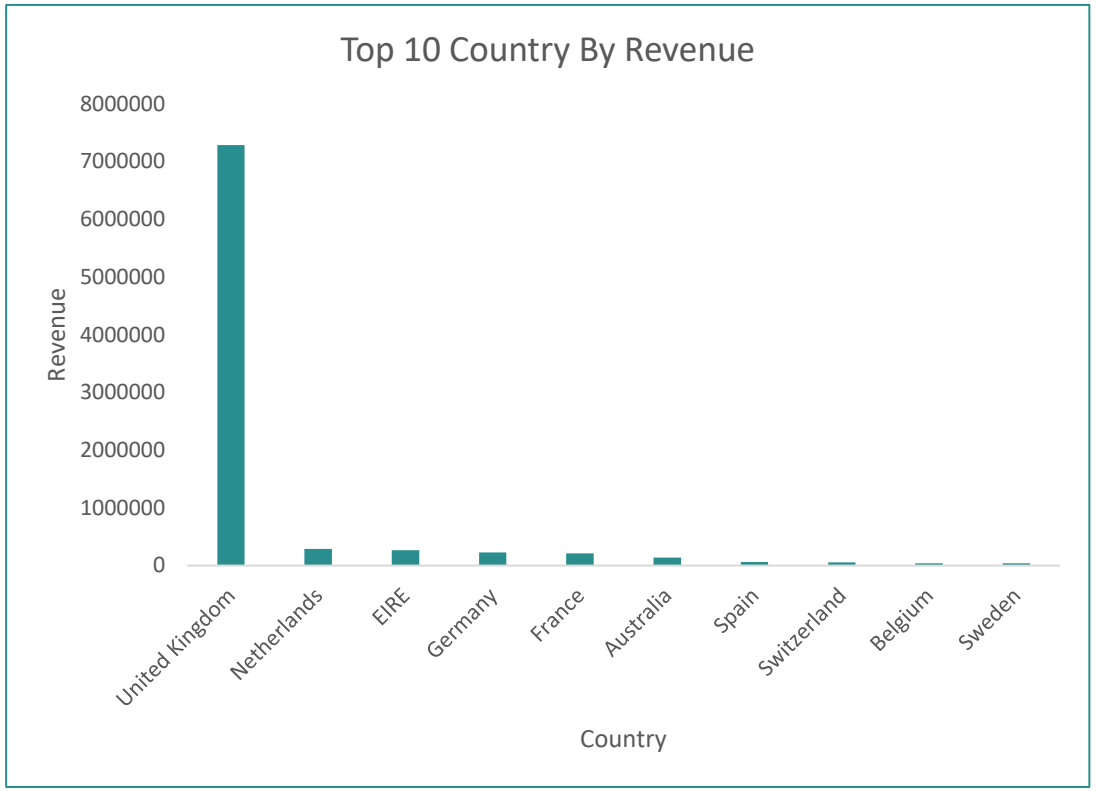


- WHITE HANGING HEART T-LIGHT HOLDER is the highest selling product almost 2016 units were sold.
- REGENCY CAKESTAND 3 TIER is the 2nd highest selling product almost 1713 units were sold.

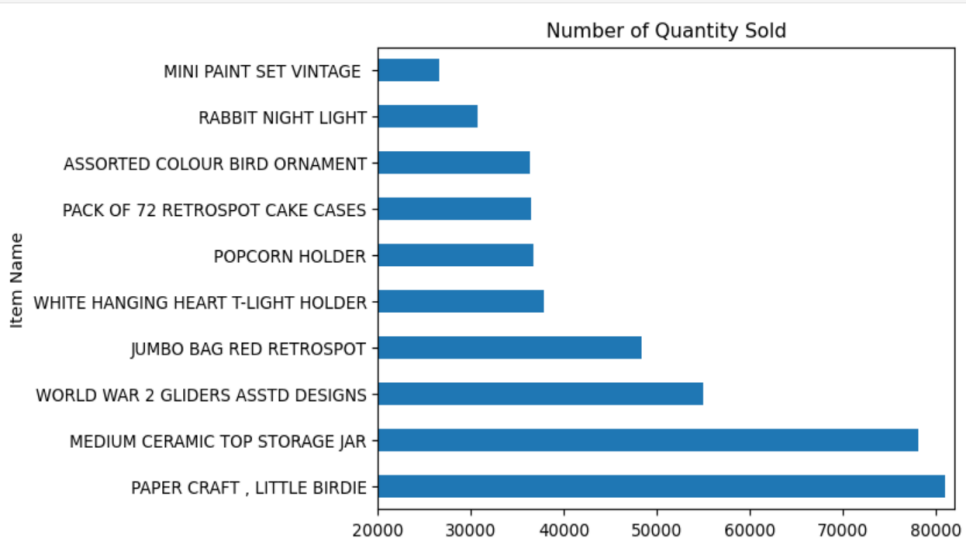
Exploratory Data Analysis

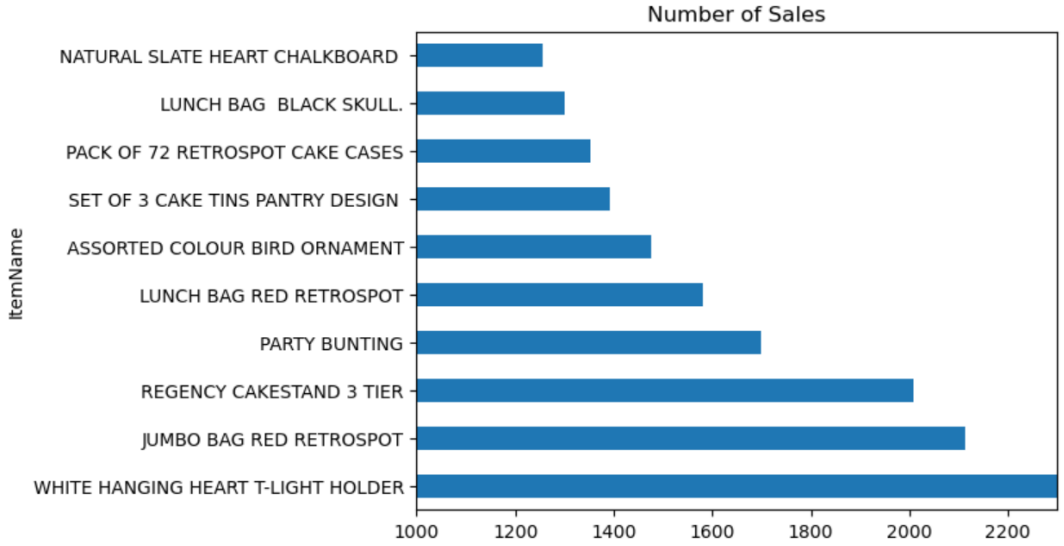
Countries Responsible for the Most Percentage of Revenue:

	Country	Revenue	Percentage_of_Revenue
35	United Kingdom	7285024.644	81.972020
23	Netherlands	285446.340	3.211878
10	EIRE	265262.460	2.984767
14	Germany	228678.400	2.573118
13	France	208934.310	2.350955
0	Australia	138453.810	1.557900
30	Spain	61558.560	0.692665
32	Switzerland	56443.950	0.635114
3	Belgium	41196.340	0.463546
31	Sweden	38367.830	0.431720



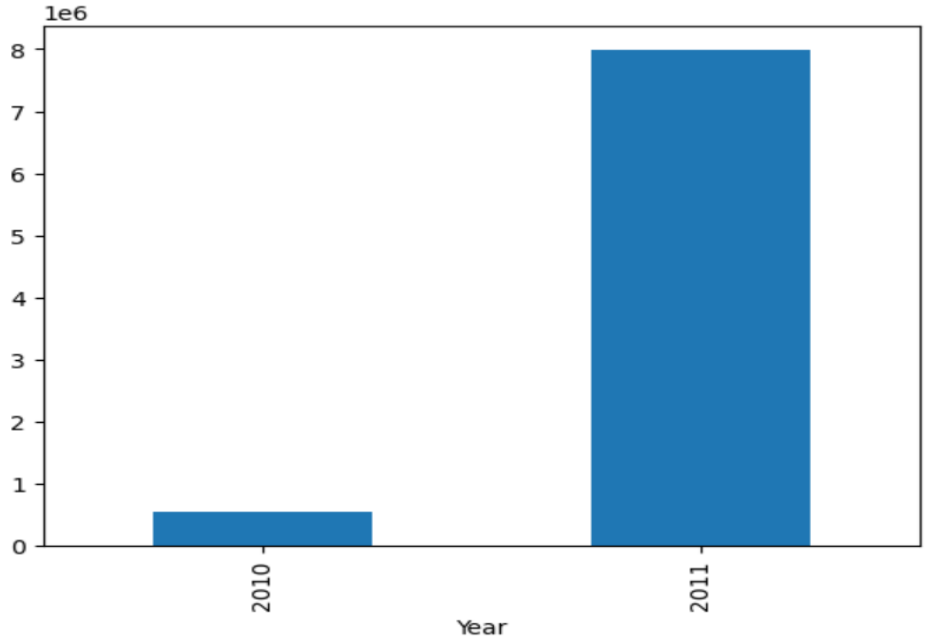
Exploratory Data Analysis

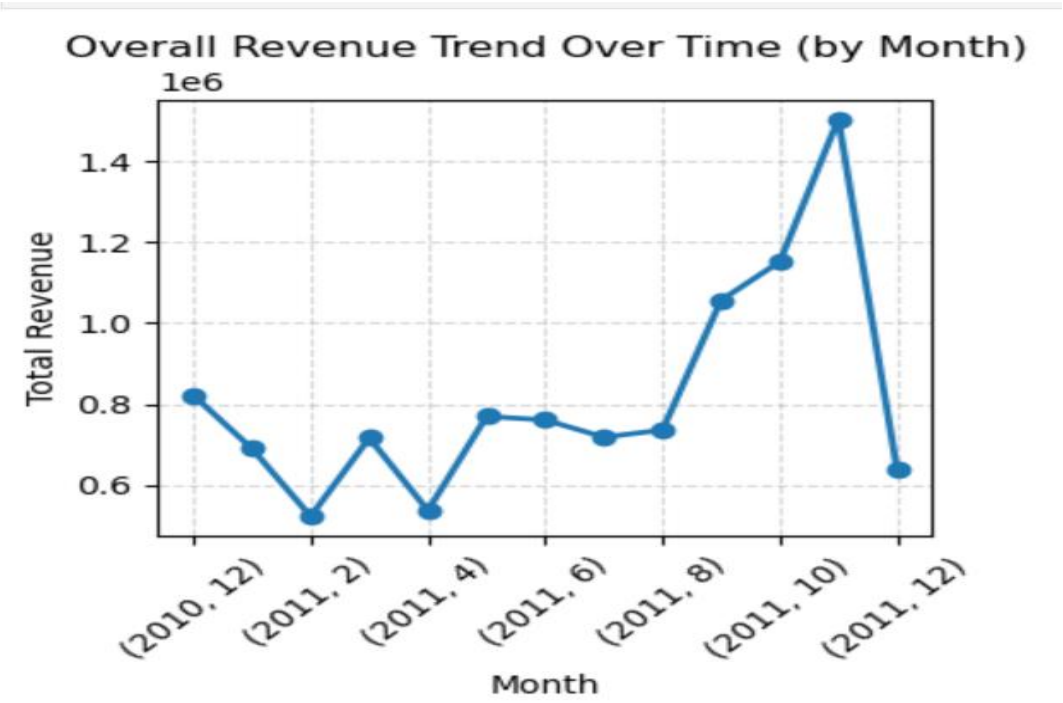




- This insight highlights the importance of considering both the quantity sold and the count of sales when analyzing the popularity and demand for different products.

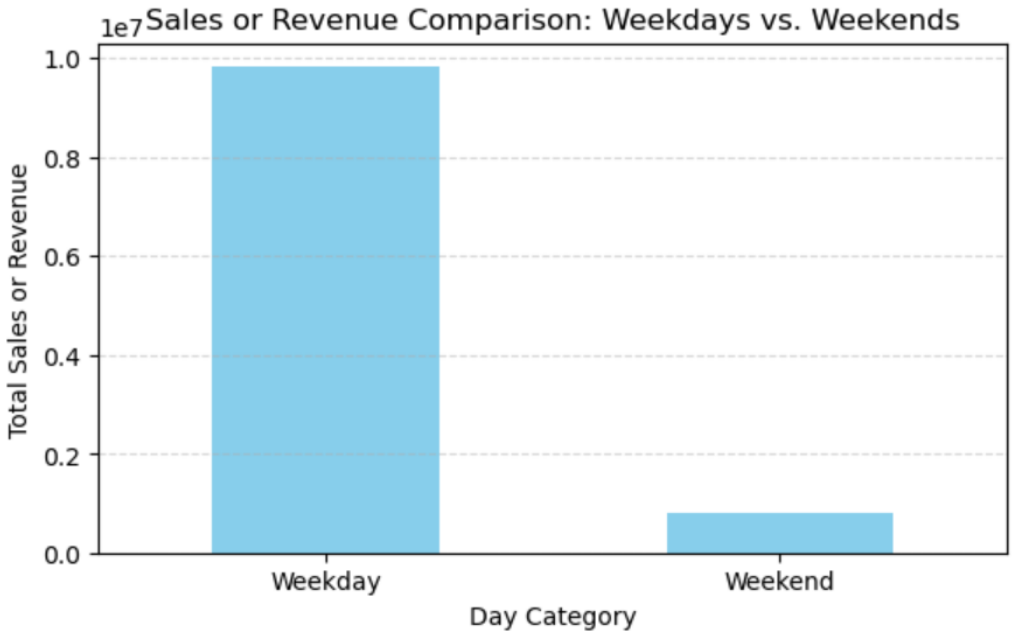
Exploratory Data Analysis

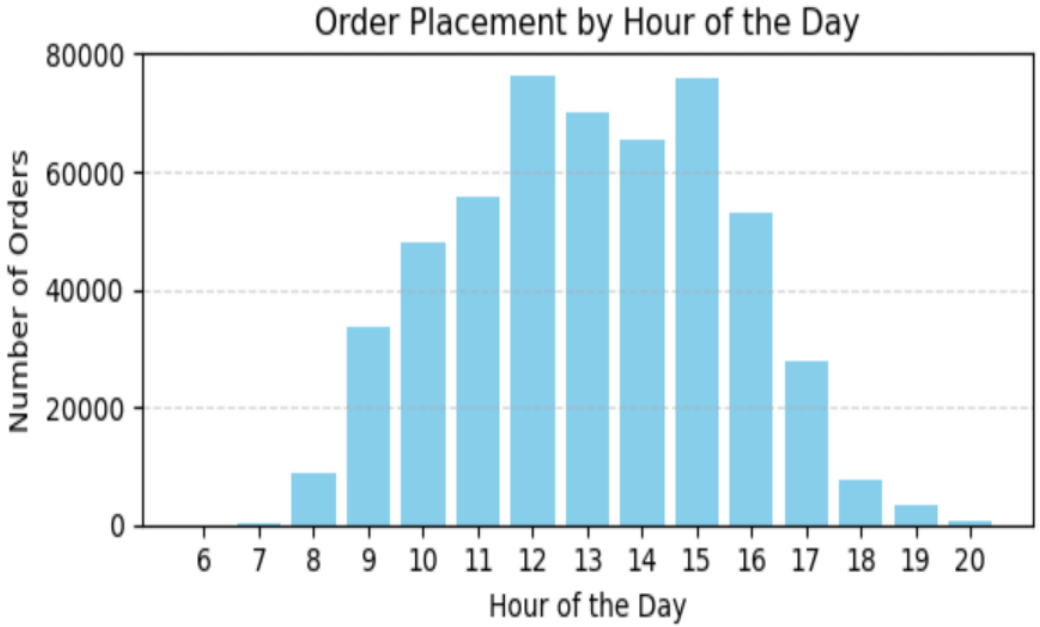




- In 2010 we have sales only for dec and in 2011 we have sales for all months

Exploratory Data Analysis





- Are there differences in sales or revenue between weekdays and weekends?
- What are the peak hours of the day when most orders are placed?

RFM Model

RFM is a basic customer segmentation algorithm based on their purchasing behavior. The behavior is identified by using only three customer data points:

R

*When is the **latest** purchase date?*

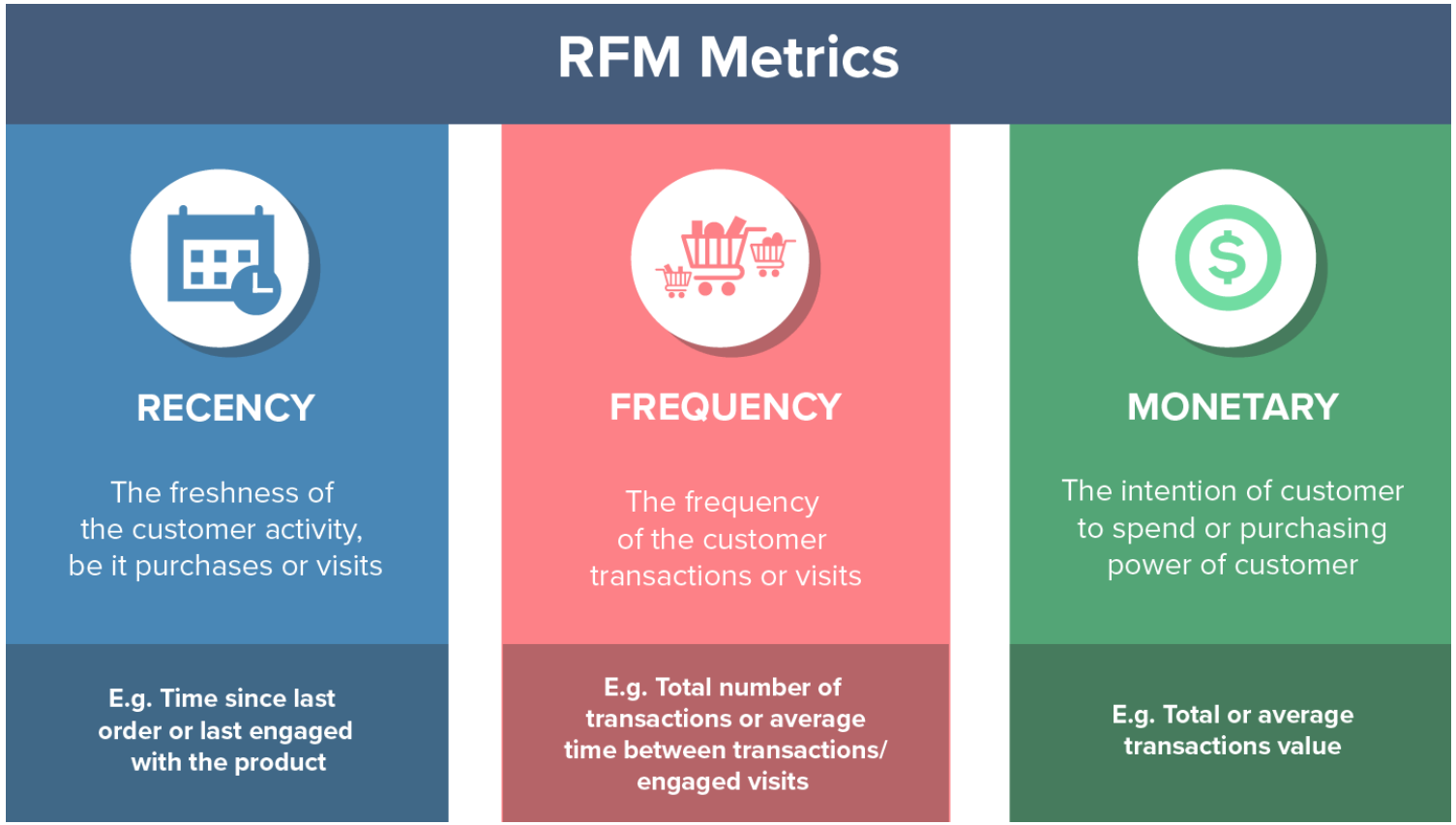
F

*How **frequently** do they make purchases?*

M

*How **large** their average ticket size is made?*

RFM Metrics


RFM Metrics
RECENCY
The freshness of the customer activity, be it purchases or visits
E.g. Time since last order or last engaged with the product
FREQUENCY
The frequency of the customer transactions or visits
E.g. Total number of transactions or average time between transactions/ engaged visits
MONETARY
The intention of customer to spend or purchasing power of customer
E.g. Total or average transactions value

RFM Analysis

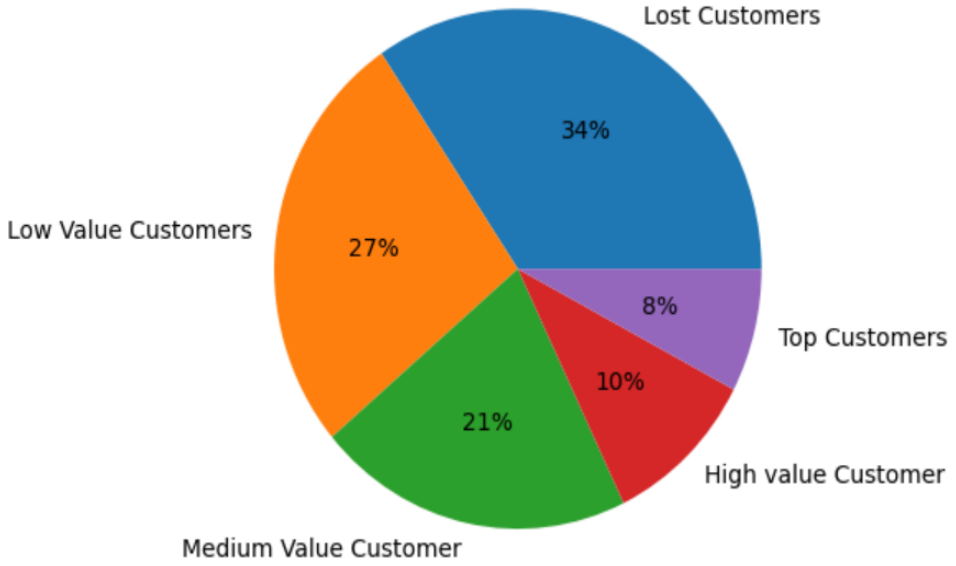
- Lost Customers (34%):
- - Implement reactivation campaigns to win back these customers.

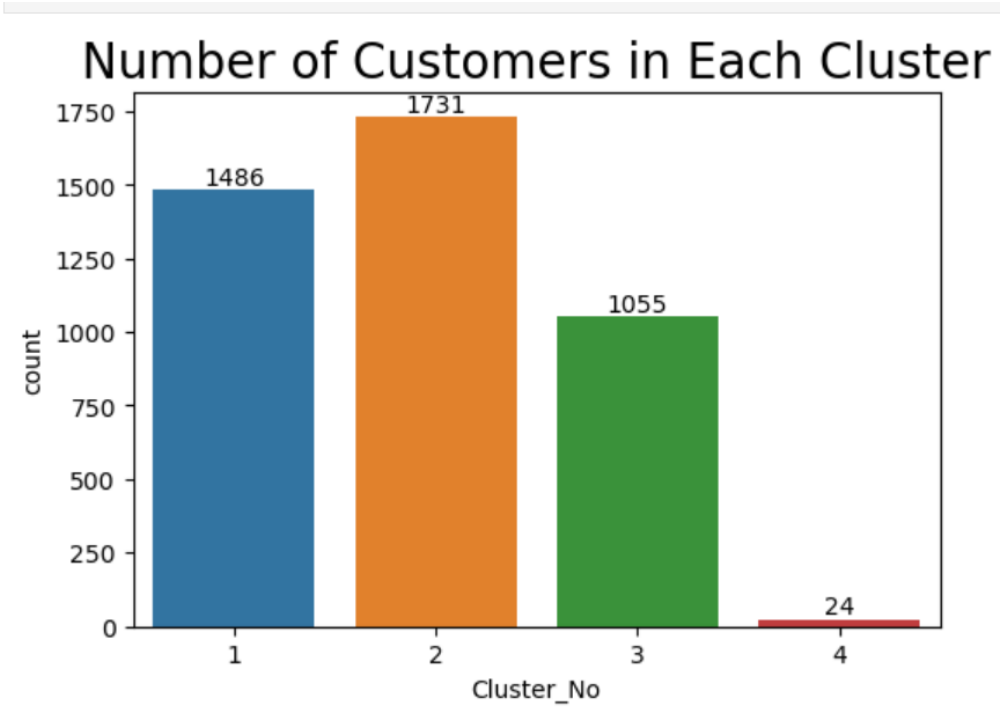
- Top Customers (8%):
- - Continue providing top-notch services and exclusive benefits to maintain their loyalty.

- High-Value Customers (10%):
- - Recognize and reward their loyalty with loyalty programs and special offers.

- Medium-Value Customers (21%):
- - Provide regular discounts, offers, and incentives to keep them engaged.

- Low-Value Customers (27%):
- - Target these customers with upselling and cross-selling strategies.





- Cluster 4: These customers have the highest monetary values, indicating they are high spenders.
- Cluster 2: While they have relatively high frequency and monetary values, they are not as recent as Cluster 4.
- Cluster 1: These customers have made purchases relatively recently, but their frequency and monetary values are moderate.
- Cluster 3: These customers have a low frequency and monetary value, and their recency is also not very recent.

Clustering with K-mean Algorithm

K-Means algorithm is an unsupervised learning algorithm that uses the geometrical principle to determine which cluster belongs to the data

MBA Model

Market basket analysis is a technique used mostly by retailers to identify which products clients purchase together most frequently.

Apriori algorithm:

It is an algorithm that uses frequent itemset to generate association rules.

It is based on the concept that a subset of a frequent itemset must also be frequent itemset.

MBA metrics

$$Rule: X \Rightarrow Y$$

$$Support = \frac{frq(X,Y)}{N}$$

$$Confidence = \frac{frq(X,Y)}{frq(X)}$$

$$Lift = \frac{Support}{Supp(X) \times Supp(Y)}$$

Example:

Rule	Support	Confidence	Lift
$A \Rightarrow D$	2/5	2/3	10/9
$C \Rightarrow A$	2/5	2/4	5/6
$A \Rightarrow C$	2/5	2/3	5/6
$B \& C \Rightarrow D$	1/5	1/3	5/9

Cross-Selling And Upselling Recommendation

```
Cross-Selling Recommendations:
Customers who bought 'PINK REGENCY TEACUP AND SAUCER' also bought 'GREEN REGENCY TEACUP AND SAUCER'.
Customers who bought 'PINK REGENCY TEACUP AND SAUCER' also bought 'ROSES REGENCY TEACUP AND SAUCER '.
Customers who bought 'GREEN REGENCY TEACUP AND SAUCER' also bought 'ROSES REGENCY TEACUP AND SAUCER '.
Customers who bought 'GARDENERS KNEELING PAD CUP OF TEA ' also bought 'GARDENERS KNEELING PAD KEEP CALM '.
Customers who bought 'ROSES REGENCY TEACUP AND SAUCER ' also bought 'GREEN REGENCY TEACUP AND SAUCER'.
```

```
Upselling Recommendations:
For customers who bought 'PINK REGENCY TEACUP AND SAUCER', recommend the following upgrades: GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER .
For customers who bought 'GREEN REGENCY TEACUP AND SAUCER', recommend the following upgrades: PINK REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER .
For customers who bought 'ROSES REGENCY TEACUP AND SAUCER ', recommend the following upgrades: GREEN REGENCY TEACUP AND SAUCER, PINK REGENCY TEACUP AND SAUCER.
```

Recommendations

1. We see that September to December we have very high sales. We can concentrate on improving the sales for the other 8 months
2. We find very less number of customers in Lithuania, Brazil, Czech Republic, Bahrain, Saudi Arabia
3. We have very less sales for Lebanon, Brazil, RSA, Bahrain, Saudi Arabia. We can concentrate on improving their sales

Recommendations

- Cluster 4: Continue providing premium services, offer exclusive deals, and engage them with loyalty programs to maintain their loyalty.

- Cluster 2: Provide regular offers and discounts to keep them engaged and encourage repeat purchases.

- Cluster 1:Engage them with personalized recommendations and incentives to increase their frequency and spending.

- Cluster 3: Target them with win-back campaigns and incentives to increase their spending and frequency.

Recommendations

-From MBA analysis, it indicates the possibility that customers buying the X product will buy the Y product. We need to make a decision for them. Maybe in our website, when the customer click on first one, we need to show them the other item.

Consclusion

In summary, this project highlighted the substantial significance of association rules analysis, encompassing sales trend analysis and RFM-based customer segmentation, with applications spanning various domains such as marketing, product recommendations, cross-selling strategies, and process optimization. By unraveling intricate item associations, we provided businesses with the tools to make data-driven decisions, fostering improved sales strategies, customer satisfaction, and marketing campaign refinement, ultimately driving substantial business growth. The combination of sales trend analysis and RFM segmentation allowed for a more holistic understanding of customer behavior and preferences, empowering businesses to tailor their strategies to meet specific needs and achieve remarkable results.

Thankyou