TATA online retail data analysis

1. Context

TATA is a global enterprise based in India, established in 1868. TATA has operated in over 40 countries across 6 continents, comprising 96 companies engaged in 7 commercial sectors. Retail is one of the 7 commercial sectors that TATA is currently operating in. The group's main revenue comes from 2 countries outside of India: the UK and the USA.

The management of TATA online retail wants to identify the major factors contributing to revenue and plan strategically for next year. They want to view metrics from both operations and marketing perspectives and seek guidance on areas performing well. They also want to view demographic-based metrics.

Your role is pivotal in understanding and optimizing the company's revenue generation. You'll be diving deep into the available data to uncover insights that will guide the company's strategic decisions for the upcoming year. The ultimate goal is to identify the key drivers of revenue growth, both from operational and marketing perspectives.

2. Dataset

Data source: https://www.kaggle.com/datasets/ishanshrivastava28/tata-online-retail-dataset

Description:

This dataset is about online retail orders of TATA in 2010 and 2011, which is available in both csv and xlsx formats. The given dataset contains 8 attributes:

1. InvoiceNo: Invoice number.
2. StockCode: Product code. Unique for each product.
3. Description: Description for products in each invoice.
4. Quantity: The quantities of each product per transaction.
5. InvoiceDate: Invoice date and time. The day and time when each transaction was generated.
6. UnitPrice: Product price per unit.
7. CustomerID: Customer number, uniquely assigned to each customer.
8. Country: Country name.

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

Data exploration:

```
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
```

```
The percentage of null values across the total records:
InvoiceNo      0.000000
StockCode      0.000000
Description    0.002683
Quantity       0.000000
InvoiceDate    0.000000
UnitPrice      0.000000
CustomerID     0.249267
Country        0.000000
```

Overall:

- There’re 541909 records in this dataset.

- There're 2 features with null values: Description and CustomerID. Null values in CustomerID account for approximately 25% of the data.

- Three numerical features are Quantity, UnitPrice and CustomerID. InvoiceDate should be a datetime type. The remaining 4 attributes can be considered as categorical or discrete, depending on the number of unique data points.

- Both Quantity and UnitPrice have negative values. This is abnormal. The values of these two variables also have extreme outliers (max too far from Q3).

Details:

- InvoiceNo:

+ There're 25900 unique orders in this dataset.

+ Almost invoice numbers in this dataset are represented by string of number, some contain letters. Invoice numbers which have letter contain ‘A’ or ‘C’ and their position are at the begining of string. After analysis, it can be observed that invoice numbers starting with ‘C’ represent canceled orders, those starting with ‘A’ indicate bad debts.

+ All canceled orders have negative quantities.

+ There’re also some orders without cancellation but have negative unit prices and quantities. Records with negative unit prices that are not cancelled are considered bad debts and customer id of them are NaN. Records with negative quantities that are not cancelled could be due to shortages or damages, based on the description. Besides, unit price of them seem to be all zeros and customer id are NaN.

3. Data preprocess and exploration

Data cleaning:

- Remove duplicated records.

- As negative unit price, negative quantity and null customer id don’t contribute to total revenue, I’ll drop all of them. Besides, I’ll fill in null value in description with ‘Unknow description’.

- As I have mentioned, invoice date should be datetime type. Therefore, I’ll also convert the datatype of this attribute and separate the hour, day, month and year from it.

- Afterall, I create a new column call Revenue for analysis convinience.

Data exploration:

Because invoice no are essential for data cleaning purposes, I have explored them first. Now, I’ll proceed to explore the remaining data fields.

- StockCode:

    + There're 3665 unique stock code in this dataset.

    + Top 5 stock code by quantity are: 85123A, 22423, 85099B, 84879, 47566.

- Description:

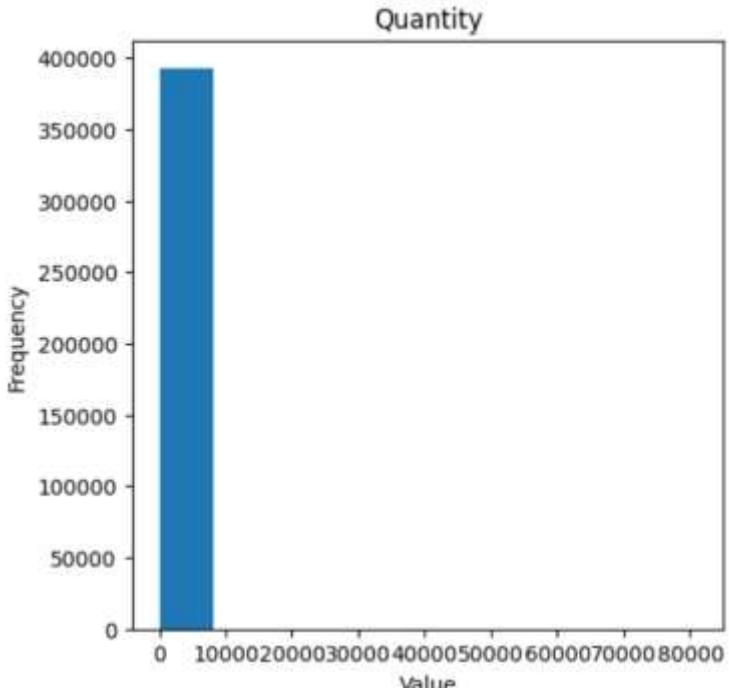
    + There're 3866 unique description in this dataset.

    + Although description may seem to describe the product name and quantity, there are also descriptions that refer to damage or deficiencies, which account for the discrepancies with the stock code.

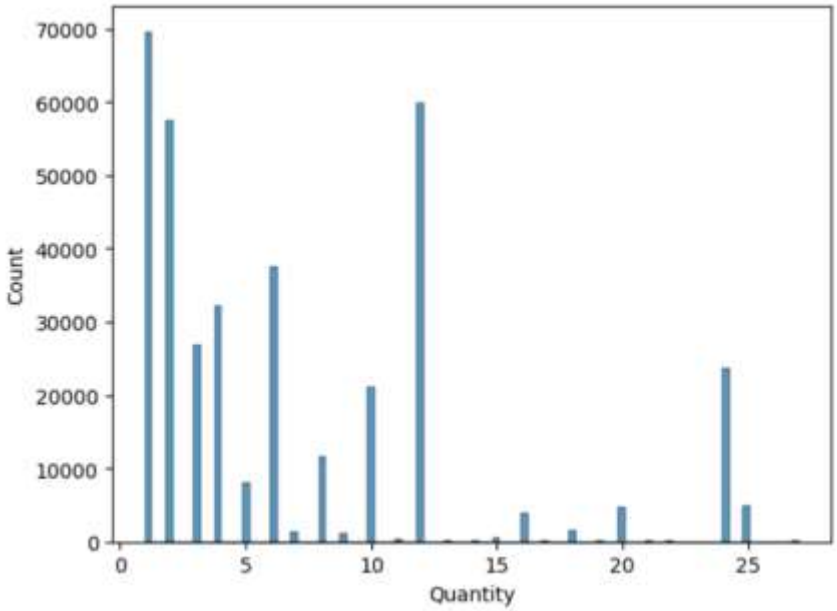
- Country:

    + There're 37 unique country in this dataset. This means that TATA business network is quite extensive. In which, UK is the largest market.

- Quantity:



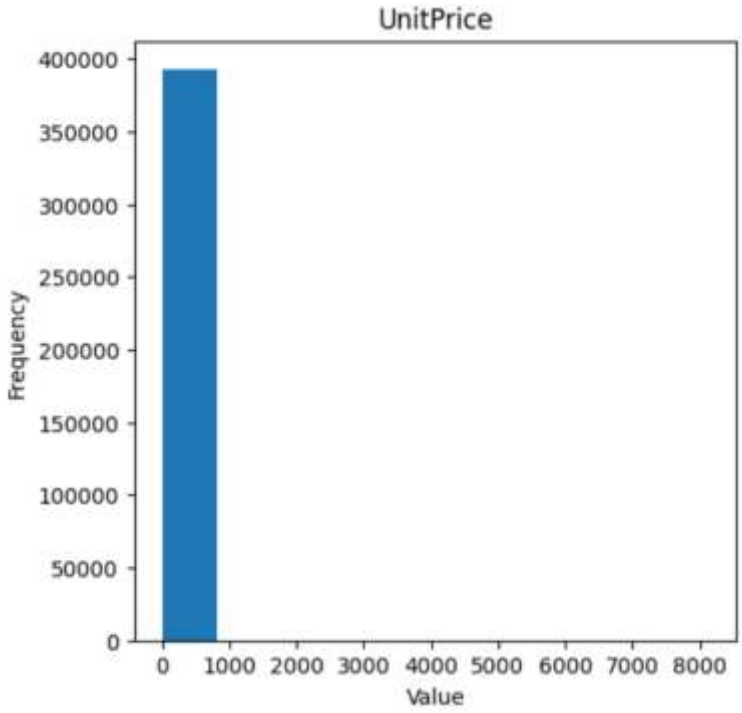
    + From histogram of quantity, it can be observed that this field have extreme outliers. I'll remove some outliers to see the true distribution of this attribute. The technique I use here is IQR. IQR defines outliers as those values that under (Q1 – 1.5IQR) and above (Q3 + 1.5IQR) with IQR = Q3-Q1.

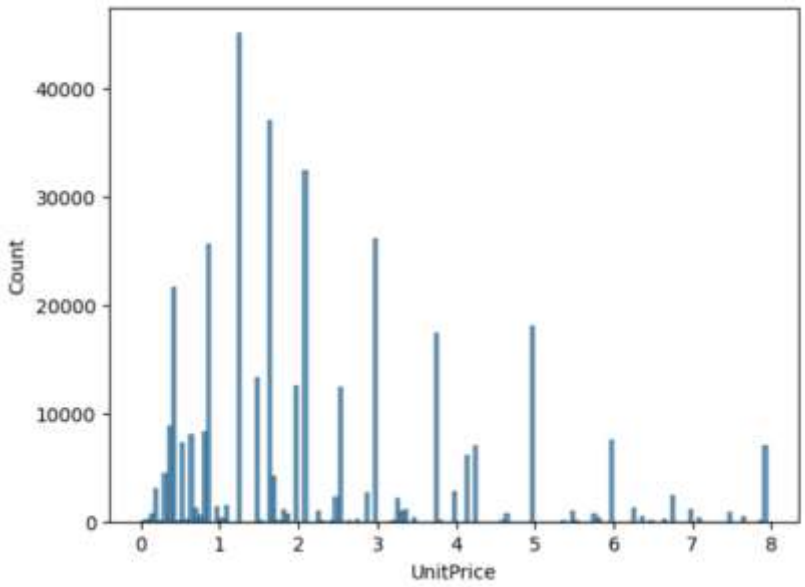


+ This is the distribution of quantity after removing outliers based on IQR.

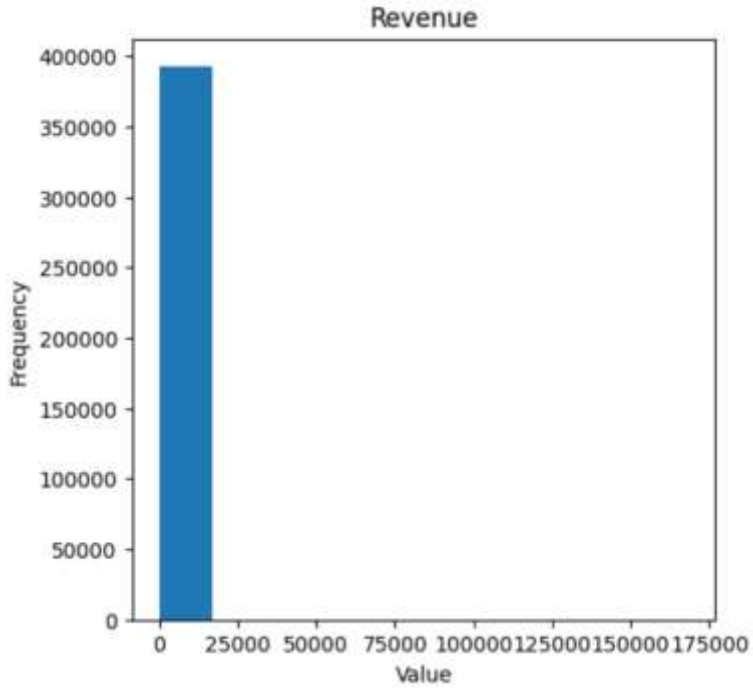
UnitPrice and revenue is the same as quantity.

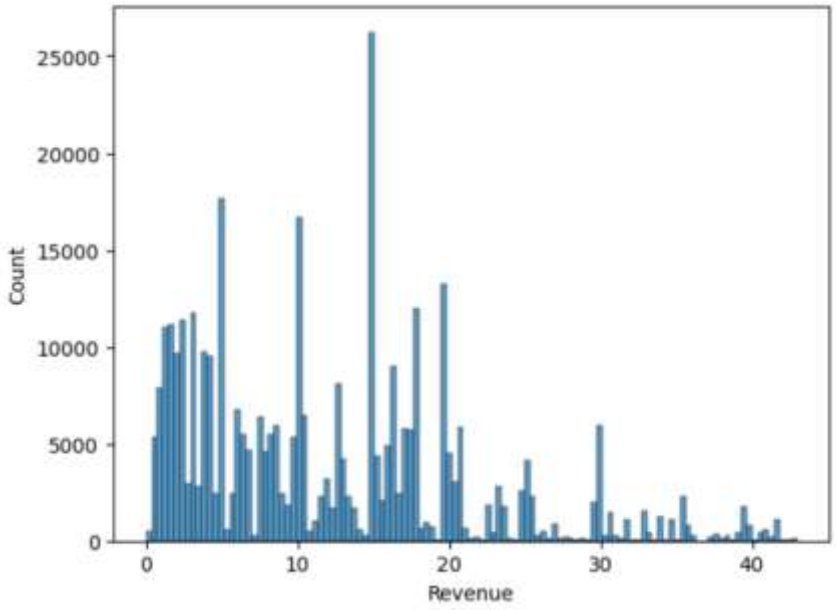
- UnitPrice:




40000
30000
20000
10000
0
Count
0
1
2
3
4
5
6
7
8
UnitPrice

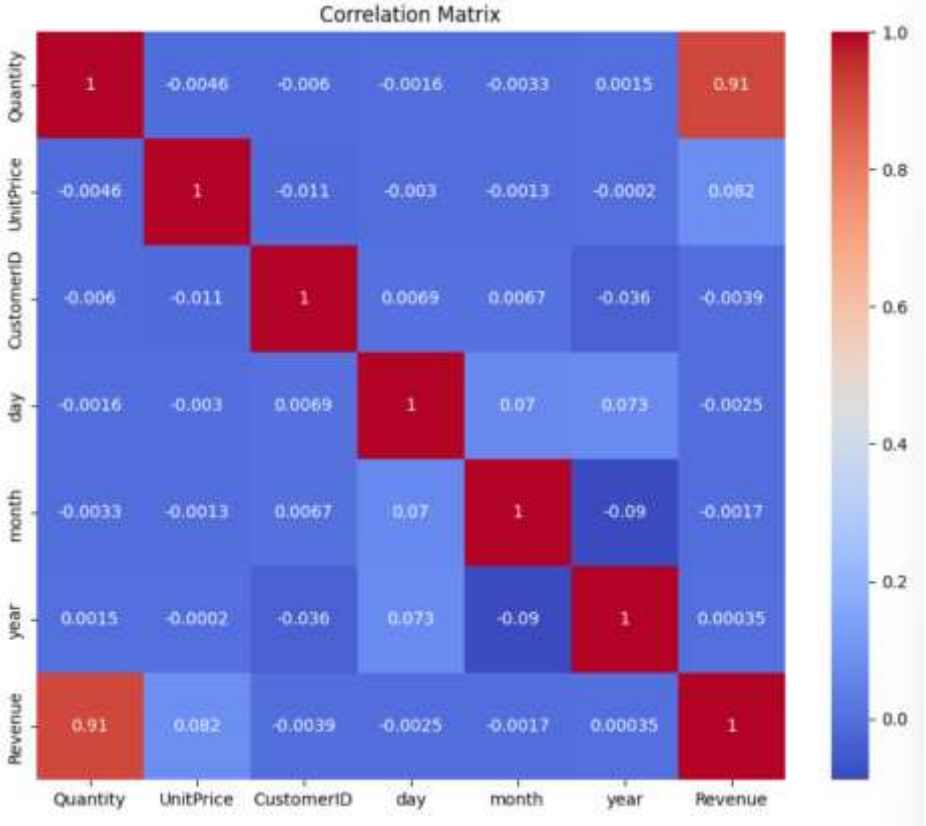
- Revenue:


Revenue
400000
350000
300000
250000
200000
150000
100000
50000
0
Frequency
0
25000
50000
75000
100000
125000
150000
175000
Value



- InvoiceDate:

  + This dataset contains orders in 2010 and 2011. Orders in 2011 are much more than in 2010.



The heatmap above shows correlation of numerical features. As we can see that quantity and revenue have a strong positive linear relationship (r = 0.91).

4. Revenue analysis

Objectives:

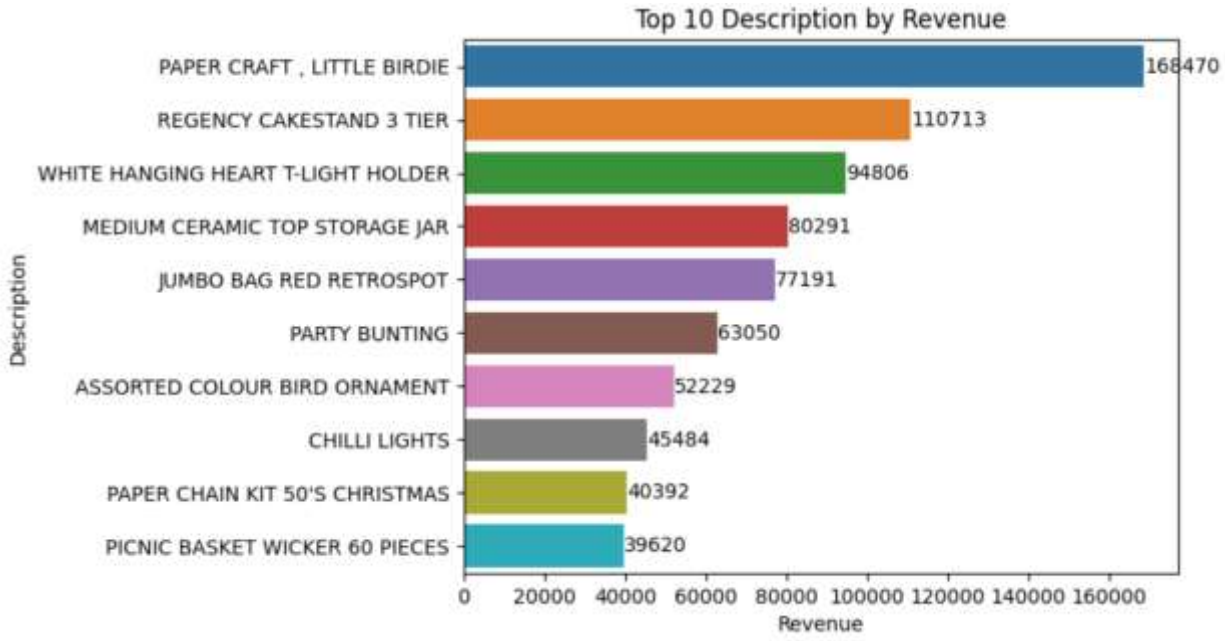
- Identify revenue source, involves examining various aspects, such as product categories, geographic regions…

Details:

Top performing products: top product that drive the most revenue

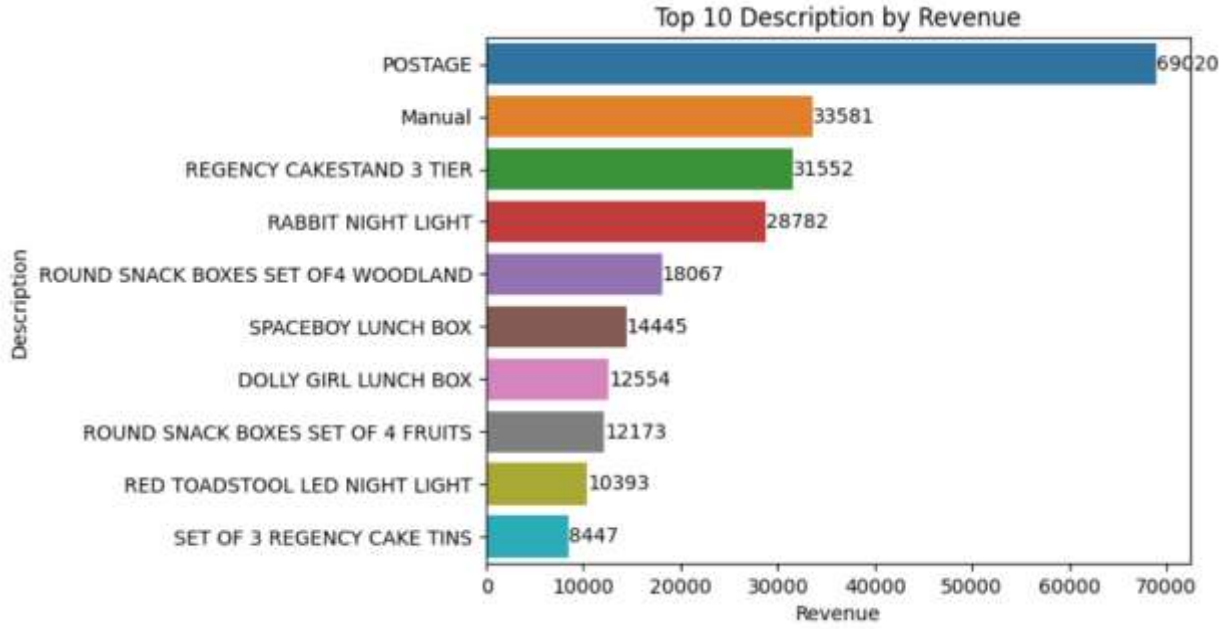
As the UK is a significant market for TATA, I'll analyze the revenue of this market separately compared to the other markets.

Below is bar plot of revenue by product in UK.

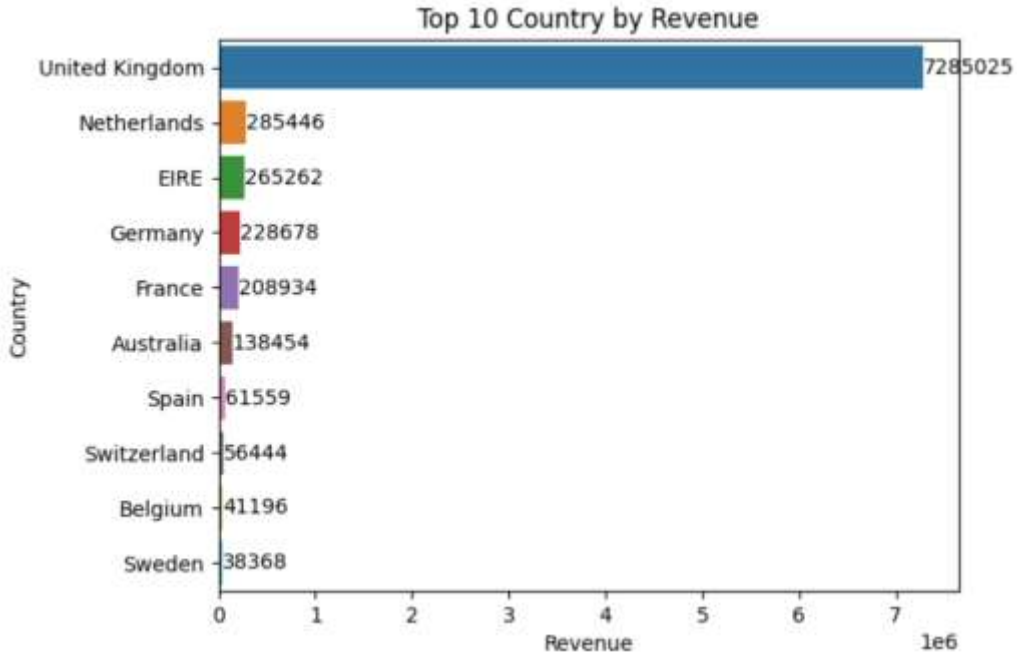


It can be observed that handmade crafts and decorations are the products with the highest revenue in UK.

The following chart is revenue by product in other countries (except UK).

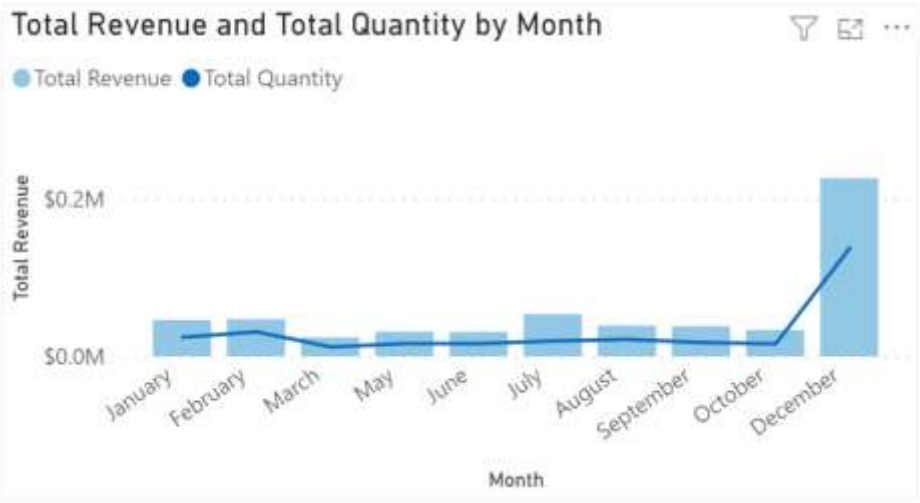


Top performing countries: top country that drive the most revenue

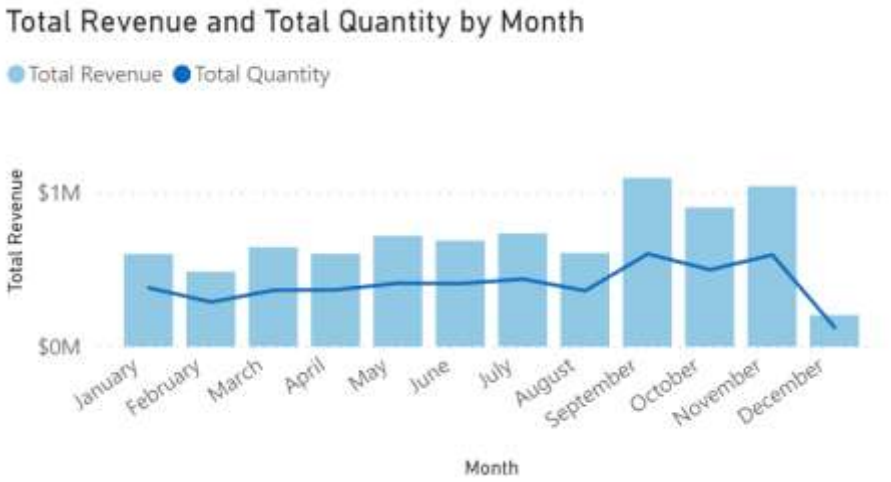


Because the UK has outstanding revenue compared to other markets, we will have a dedicated marketing plan for this region. Additionally, there are some potential areas like Netherlands, EIRE, Germany, France and Australia that also need to be promoted.

Revenue by month



Revenue by month in 2010



Revenue by month in 2011

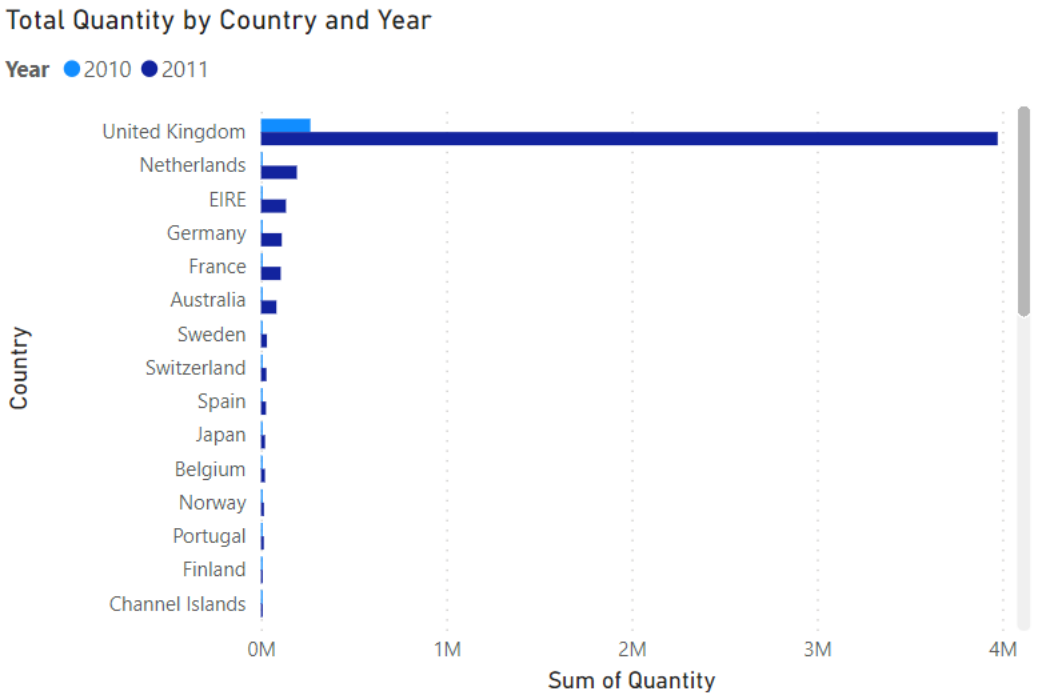
There is a difference in the revenue trend between 2010 and 2011.

In 2010, revenue seems to increase dramatically in the end of year. However, in 2011, revenue hit its peak in September and December is the month with lowest revenue.

However, there's one point that needs to be considered here: the revenue range in both years. The revenue range in 2011 is significantly higher than in 2010. Therefore, it can be observed that revenue experienced remarkable growth in December 2010 and

continued to maintain its growth trend in 2011, but experienced a sudden decline in December 2011.

Revenue = Quantity * Unit Price. Therefore, the decline or raise in revenue is related to a decrease or increase in the quantity of products.



As we can see, the total quantity of products increases much more in 2011 rather than in 2010. Especially, in the largest market UK, total quantity in 2011 is 3974732 and in 2010 is 266573, fifteen times higher.

5. Customer segmentation

Objectives:

Segmenting customers into multiplesegments based on buying behaviour. This task is essential to tailor appropriate marketing strategies for each customer group.

Technique:

Using RFM model. RFM stands for Recency, Frequency and Monetary – a model used to construct customer segments based on buying behaviour.

R – Recency: How recency customer bought? R = now – last purchase date

F – Frequency: How often they bought? F = total number of orders

M – Monetary: How much they bought? M = total money spent

Details:

Since the data is too old, without losing generality, I'll assume that the current year is 2012.

After caculating recency, frequency and monetary score for each customer, I have below table.

	CustomerID	Recency	Frequency	Monetary
0	12346.0	347	1	77183.60
1	12347.0	61	7	4310.00
2	12348.0	97	4	1797.24
3	12349.0	40	1	1757.55
4	12350.0	332	1	334.40
...	...	...	...	...
4333	18280.0	181	1	180.60
4334	18281.0	25	1	80.82
4335	18282.0	237	2	178.05
4336	18283.0	31	16	2045.53
4337	18287.0	21	3	1837.28

I plan to divide customers into 11 segments as below:

Segment	Description	Recency Score	Frequency Score	Monetary Score
Champions	Bought recently, buy often and spend the most.	**4 - 5**	**4 - 5**	**4 - 5**
Loyal Customers	Spend good money. Responsive to promotions.	**2 - 4**	**3 - 4**	**4 - 5**
Potential Loyalists	Recent customers, spent good amount, bought more than once	**3 - 5**	**1 - 3**	**1 - 3**
New Customers	Bought more recently but not often	**4 - 5**	**< 2**	**< 2**
Promising	Recent shoppers but haven't spent much	**3 - 4**	**< 2**	**< 2**
Need Attention	Above average recency, frequency & monetary values	**3 - 4**	**3 - 4**	**3 - 4**
About to Sleep	Below average recency, frequency & monetary values	**2 - 3**	**< 3**	**< 3**
At Risk	Spent big money, purchased often but long time ago	**< 3**	**2 - 5**	**2 - 5**
Can't Lose Them	Made big purchases and often but long time ago	**< 2**	**4 - 5**	**4 - 5**
Hibernating	Low spenders, low frequency and purchased long time ago	**2 - 3**	**2 - 3**	**2 - 3**
Lost	Lowest recency, frequency & monetary values	**< 2**	**< 2**	**< 2**

Therefore, I'll divide each range of R, F, M into 5 equal intervals. Afterall, I have table as below:

	CustomerID	Recency	Frequency	Monetary	RScore	FScore	MScore	rfmscore	segment
0	12346.0	347	1	77183.60	1	1	5	111	Lost
1	12347.0	61	7	4310.00	4	5	5	455	Champions
2	12348.0	97	4	1797.24	3	4	4	344	Loyal Customers
3	12349.0	40	1	1757.55	4	1	4	411	Potential Loyalists
4	12350.0	332	1	334.40	1	1	2	111	Lost

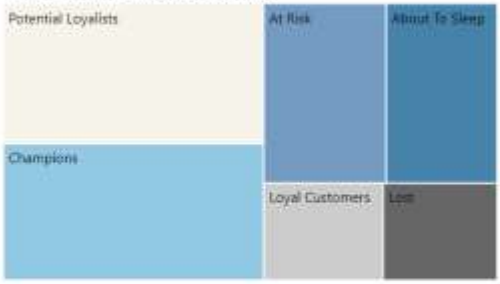
Power BI Dashboard

Customer Segmentation Dashboard

4338
Total customers

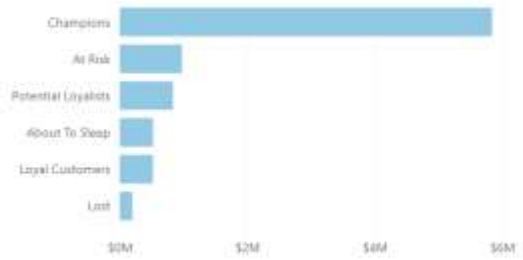
6
Total segments

Total customers in each segments



CustomerID	Recency	Frequency	Monetary	rfmscore	segment
12346	347	1	77,183.60	111	Lost
12347	61	7	4,310.00	455	Champions
12348	97	4	1,797.24	344	Loyal Customers
12349	40	1	1,757.55	411	Potential Loyalists
12350	332	1	334.40	111	Lost
12352	94	8	2,506.04	355	At Risk
12353	226	1	89.00	111	Lost
12354	254	1	1,079.40	111	Lost
12355	117	1	459.40	211	About To Sleep

Total revenue in each segments



Total Customer by Country and Segment