Retail Store Data-EDA

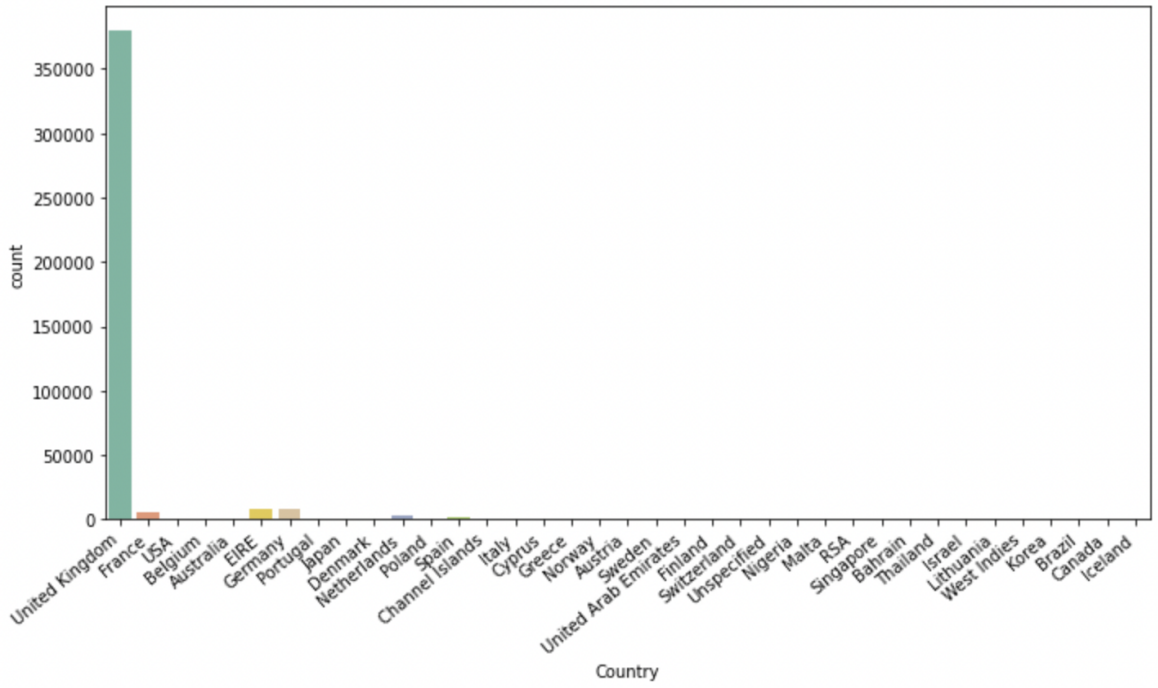
- The data we are given is the dataset for a multinational retail store which operates in many countries worldwide.

- Let us start by observing the first few entries of the dataset:

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	12/1/2009 7:45	6.95	13085.0	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	12/1/2009 7:45	6.75	13085.0	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	12/1/2009 7:45	6.75	13085.0	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	12/1/2009 7:45	2.10	13085.0	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	12/1/2009 7:45	1.25	13085.0	United Kingdom

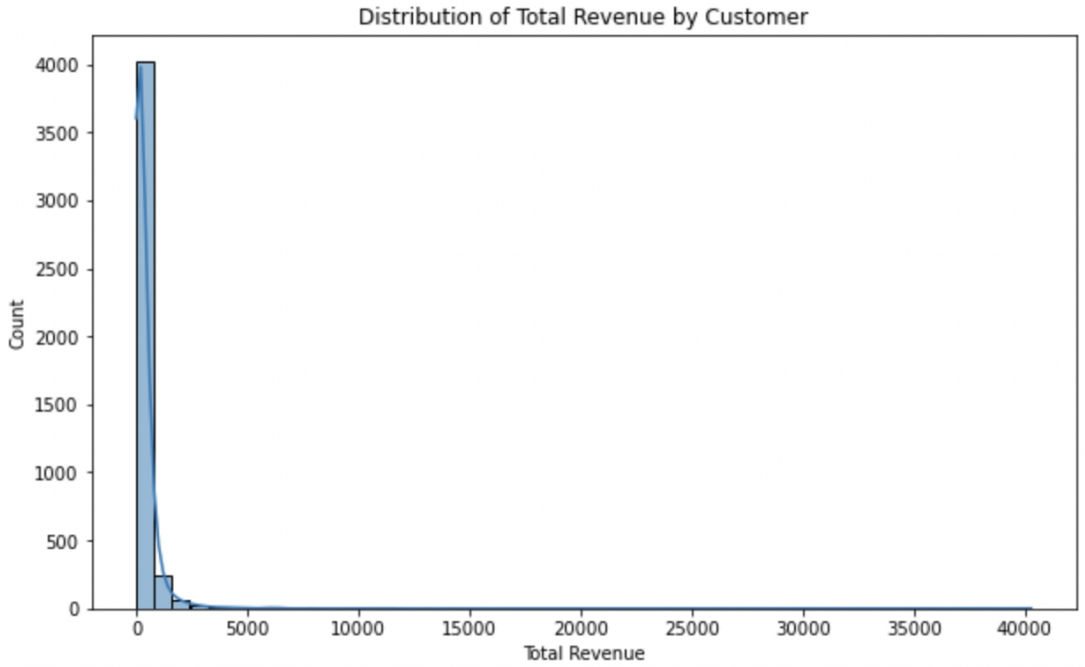
  By observation, The data columns given are Invoice, StockCode, Description, Quantity, InvoiceDate, Price, Customer ID, and Country. A particular row of this dataset defines a single transaction made by a particular customer and everything about it.

- Now, after some cleaning and restructuring of the data, let us start with the exploratory data analysis of the dataset:
  1. Top countries in terms of customers



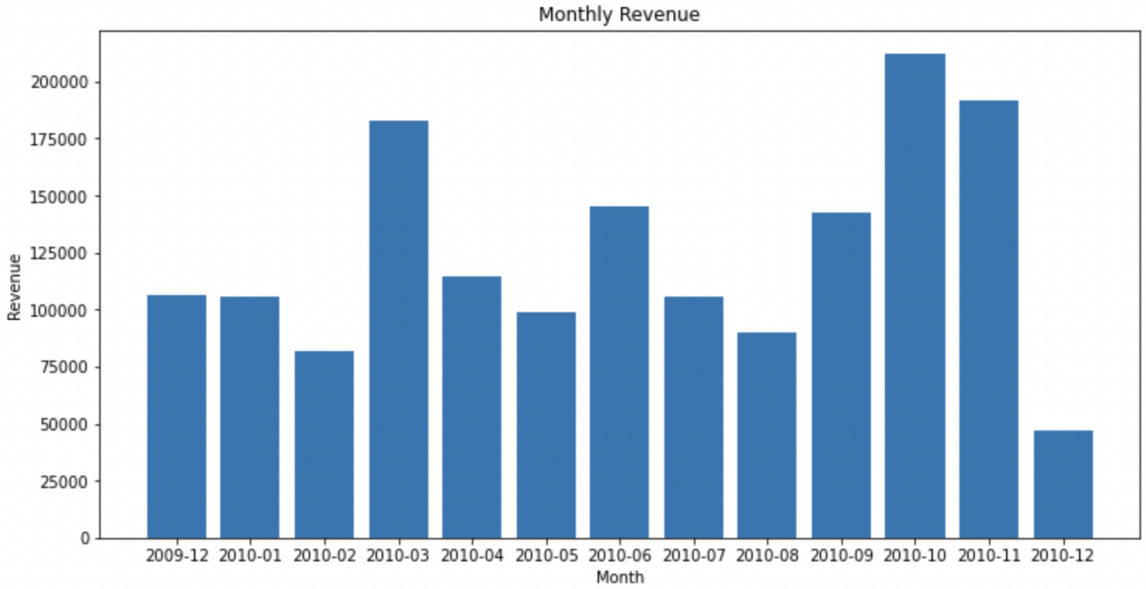
The above figure shows that the majority of customers of the retail chain are from the **United Kingdom.** This can also mean that there is simply more number of stores of the chain in the UK as supposed to the other countries.

  2. Distribution of total revenue by customers
     We can perform customer segmentation based on the metrics such as the total number of unique invoices, total quantity purchased, and total revenue generated by each customer.



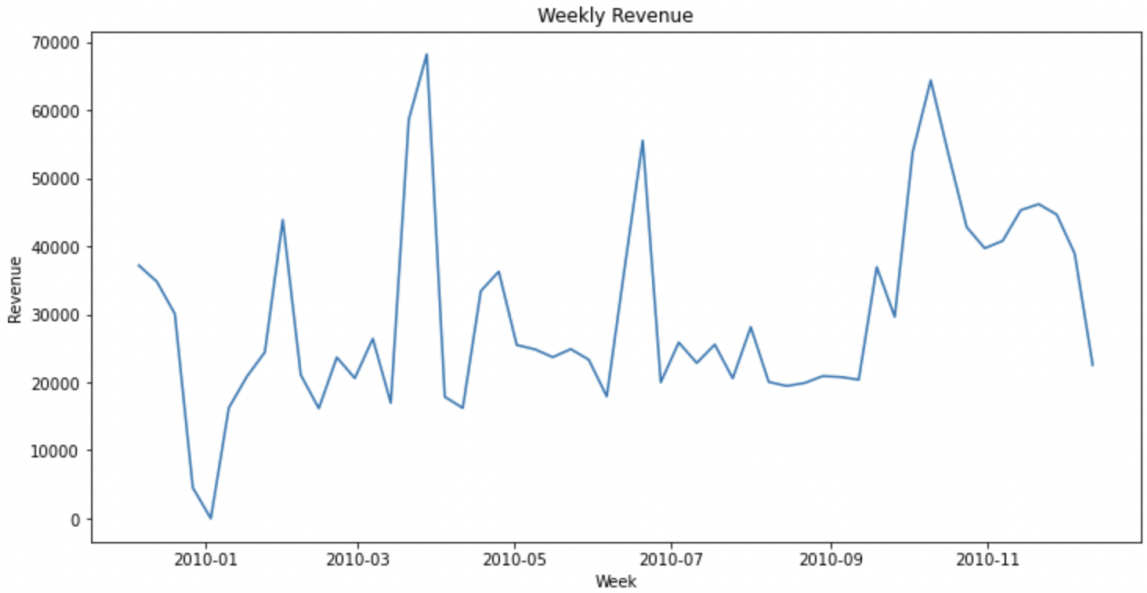
Here, we can observe that there is an **inverse relation between the total revenue** and **the number of customers**, which means there are a large number of customers in the lower revenue market while there is a very niche category of customers in the higher revenue category.

3. Monthly Revenue



We can see that the most popular time of year is around **November** and **October**, which is the holiday season. The sales peak during these months and then **decline in January**, which is typical for retail businesses. There is also a slight increase in sales during the summer months, which could be due to seasonal products or promotions. Overall, the holiday season is the most important time of year for this store, and they could potentially increase their marketing efforts during this time to boost sales even further.

4. Seasonality in data

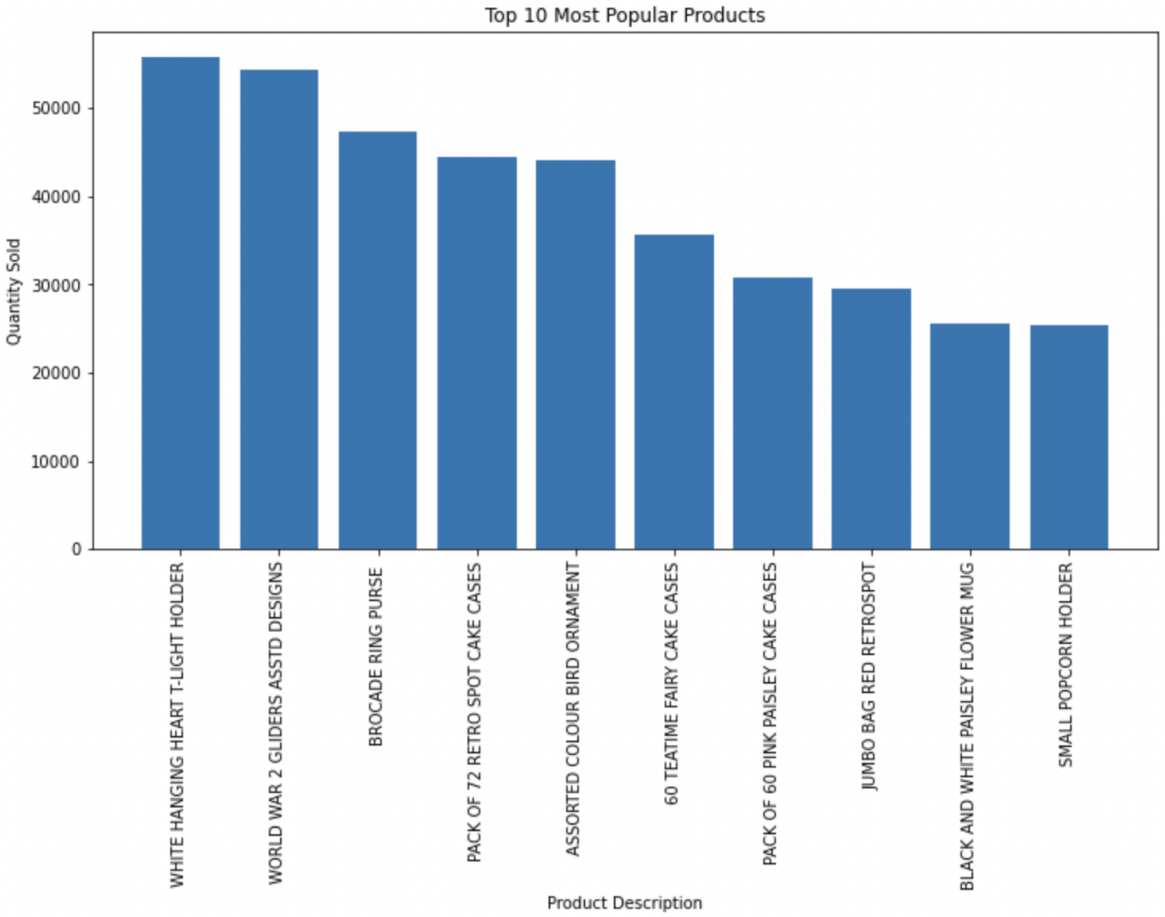


We can see that there is a clear weekly pattern in the sales data. Sales **peak during the weekdays** and **decline during the weekends**. This is a common pattern in retail businesses where weekdays are busier due to working days and weekends being relatively slower.

5. Relation between the quantity of items and price of those items

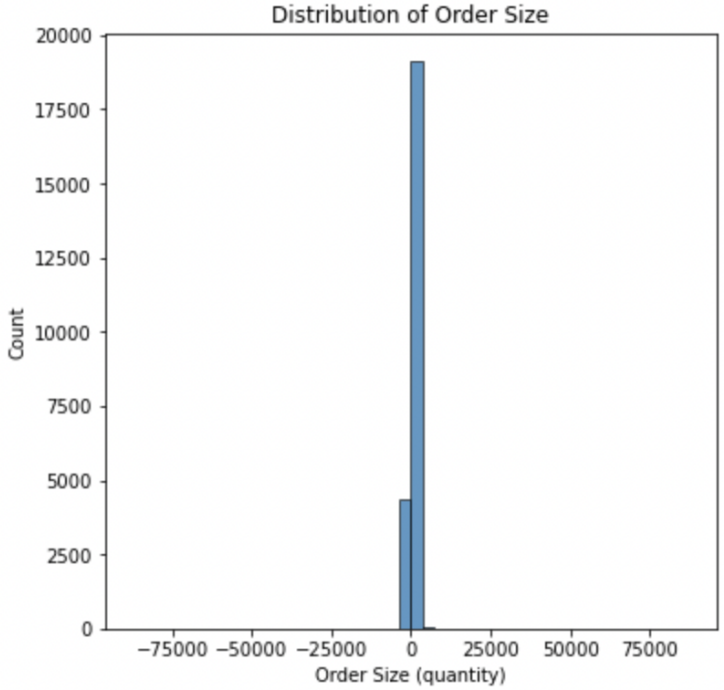
Here, we can see that as the price of the **product increases**, the number of **quantities becomes lesser** which was also very evident by the distribution of customer vs revenue graph which we plotted above in the 2nd point

6. Most popular products



Now, the above observation is critical for **inventory management** as the graph clearly describes what all products are popular and what are the plausible market size for it in the total sales.

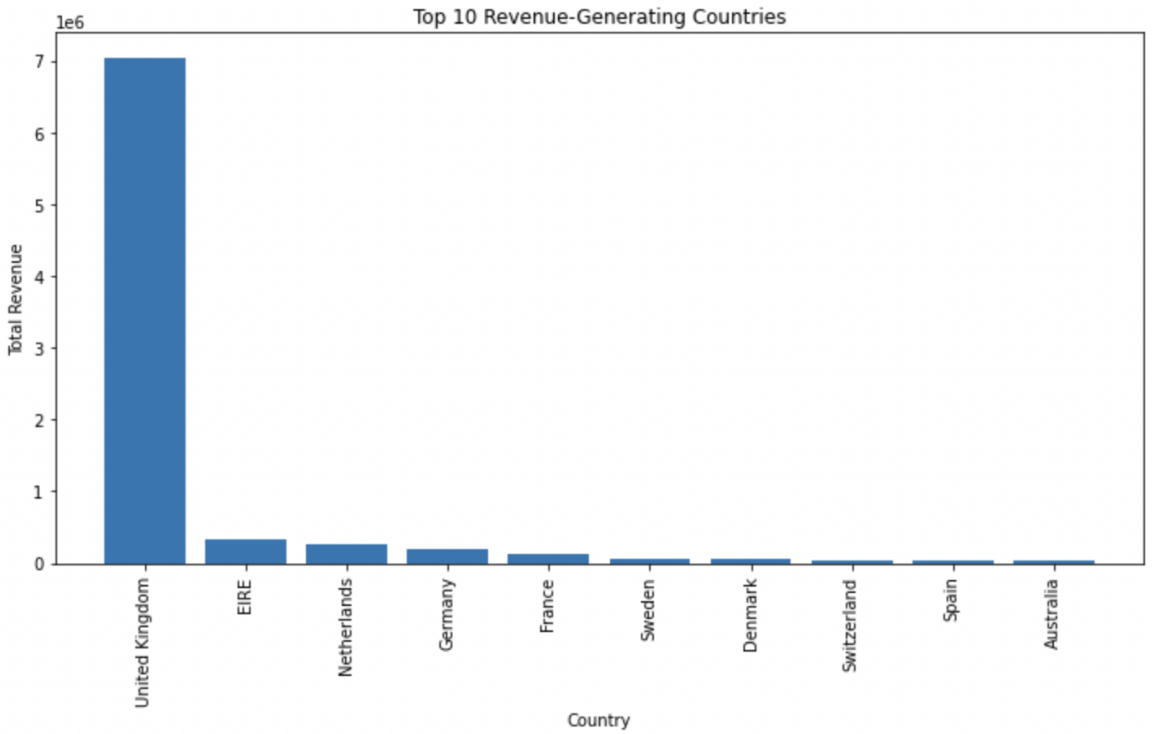
7. Average order size



From the above graph, we can see that the majority of the customers do not buy very large quantities of goods from the retail chain. The calculated average order size and the average revenue are:-

Average order size (quantity): 225.8548776868614
Average order size (revenue): 353.93261432145886

8. Top countries in terms of revenue



This result was obvious if we see the maximum number of customers in the retail chain, UK has the highest amount of market share both in terms of revenue and number of customers for the retail chain.

9. Customer lifetime value
Now, if the customer lifetime value is calculated, we get the following insights:

Count	4383
Mean	1904.107136
Standard deviation	3797.491782
Minimum	-165092.669277
25th percentile	975.644393
50th percentile	1512.396222
75th percentile	2319.549936
Maximum	78110.491773

There are 4,383 customers in the dataset with an average CLV of 1,904.10 and a standard deviation of $3,797.49. The minimum CLV is negative, which means that some customers may have returned more products than they purchased, resulting in a loss for the company.

The quarter range (IQR) of the CLV distribution is between 975.64 and 2,319.55, indicating that the majority of customers fall within this range.

How can the above data be used to improve the business?

- **Customer segmentation on the basis of CLV(customer lifetime value)**
  The customer ID column can be used to segment customers based on their purchasing behaviour, such as total amount spent, frequency of purchases, and recency of purchases. This information can be used to tailor marketing campaigns and promotions to specific customer groups.

- **Inventory management(on the basis of most popular products)**
  The stock code and description columns can be used to analyze which products are selling well and which are not. This information can be used to optimize inventory levels and make better purchasing decisions.

- **Sales forecasting(on the basis of past trends)**
  The quantity and price columns can be used to forecast future sales and revenue. This information can be used to make more accurate sales predictions and adjust pricing strategies accordingly.

- **Geographic analysis(on the basis of the largest segment of customers)**
  The country column can be used to analyze sales trends in different geographic regions. This information can be used to tailor marketing and sales strategies to specific regions and countries.