Predictive Modelling in Retail

Supervised & Unsupervised Machine Learning

IST 498 – Capstone Info Science

By: Aya Ibrahim

Outline

1. Dataset Overview
    - Source & Significance
    - Key Attributes & Timeframe
2. Goals & Objectives
    - Expected Outcomes
3. Data Analysis
    - Data Wrangling & Pre-processing
    - Descriptive Statistics & Visualization of Key Metrics
4. Predictive Modelling
    - Supervised: Time Series Forecasting
        - Methodology and Predictions for Annual Revenue Trends
        - Model Accuracy
    - Customer Segmentation
        - Application of RFM Analysis
        - Unsupervised: Clustering with K-means and Details on Customer Groups
5. Results
    - Key Findings and Actionable Strategies

1. Dataset Overview

Source & Significance:

- Sourced from the University of California at Irvine (UCI) Machine Learning Repository, one of many resources for getting access to real-world datasets.
- It involves detailed transactions from a UK-based online retail store that specializes in unique gift items, mainly selling to wholesale customers.

Key Attributes & Timeframe:

- Data includes a variety of attributes over two years, from December 2009 to December 2011.

Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085.0	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085.0	United Kingdom
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085.0	United Kingdom
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085.0	United Kingdom

2. Goals & Objectives

Expected Outcomes:

- The retail store is looking to increase its revenue. Therefore:
  - The project's goal is to get deep insights into sales trends and customer behavior to inform strategic decision-making.
  - The model aims to forecast annual revenue using a time series manner, enabling the retailer to plan better for the future.
  - By segmenting customers, this model strives to personalize marketing strategies and enhance customer engagement.
  - This analysis is expected to pave the way for annual revenue growth and an enhanced understanding of market dynamics.

3. Data Analysis

Data Wrangling & Pre-processing:

- Eliminated records with missing or NULL "Description" or "Customer ID" fields, as well as ~34k duplicate records.
- Filtered out all canceled transactions. These were identifiable by an "InvoiceNo" beginning with "C". This step ensures that our dataset only reflects completed sales.
- Converted the "Customer ID" from a numerical float to a string object. This change is crucial for accurate categorization as IDs are identifiers, not quantities.
- Feature engineering of a new column, "Revenue". This attribute is calculated by multiplying "Price" and "Quantity". It offers an additional layer for our analysis, sales performance at the transaction level.

Before

#	Column	Non-Null Count	Dtype
0	Invoice	1067371 non-null	object
1	StockCode	1067371 non-null	object
2	Description	1062989 non-null	object
3	Quantity	1067371 non-null	int64
4	InvoiceDate	1067371 non-null	datetime64[ns]
5	Price	1067371 non-null	float64
6	Customer ID	824364 non-null	float64
7	Country	1067371 non-null	object

After

#	Column	Non-Null Count	Dtype
0	InvoiceDate	779495 non-null	datetime64[ns]
1	Invoice	779495 non-null	object
2	StockCode	779495 non-null	object
3	Description	779495 non-null	object
4	Quantity	779495 non-null	int64
5	Price	779495 non-null	float64
6	Customer ID	779495 non-null	object
7	Country	779495 non-null	object
8	Revenue	779495 non-null	float64

Descriptive Statistics for Categorical Attributes:

- There are ~37k unique invoices and ~5800 unique Customer IDs.
- There are ~4500 unique stock codes, which indicates the variety of products being sold.
- There are 41 unique country entries in the dataset, with the United Kingdom being the most frequent country listed. This suggests that the dataset is heavily skewed towards the UK market, as it appears 700,434 times out of 779,495 which makes sense since the store is UK-based.

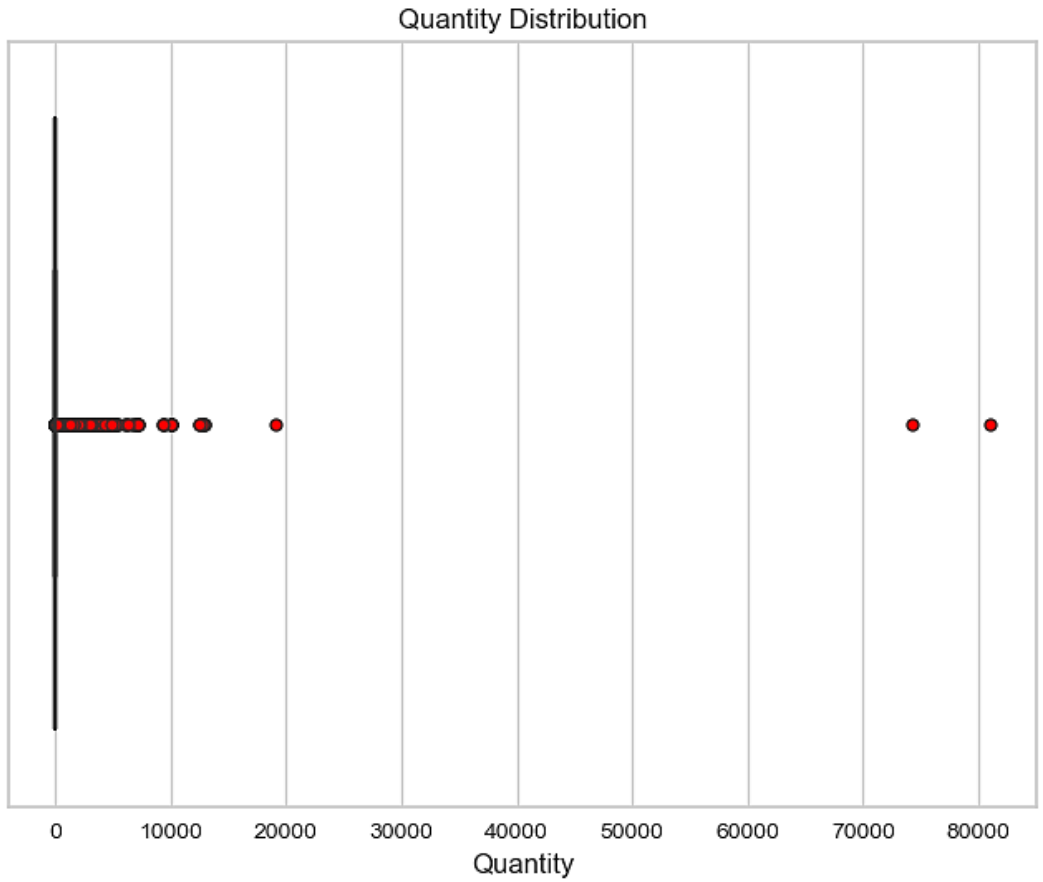
	Invoice	StockCode	Description	Customer ID	Country
count	779495	779495	779495	779495	779495
unique	36975	4631	5283	5881	41
top	576339	85123A	WHITE HANGING HEART T-LIGHT HOLDER	17841.0	United Kingdom
freq	542	5023	5016	12435	700434

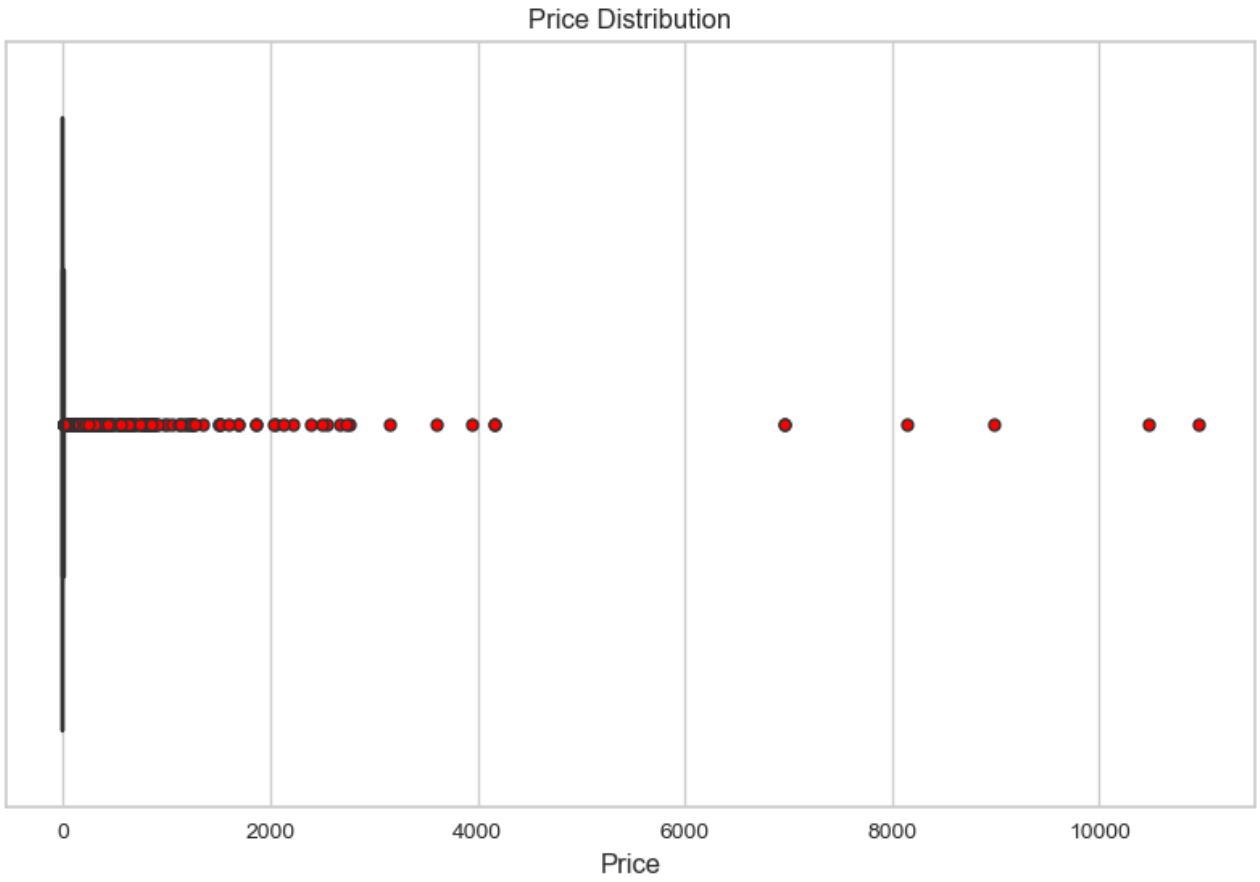
Descriptive Statistics for Numerical Attributes:

- On average, each transaction includes about 13 items, but this average is influenced by a wide variation in quantities, as indicated by a high standard deviation of ~146.
- At least one transaction includes up to 80,995 items, which indicates bulk purchases or extreme outliers.
- The average price per unit is roughly £3, with a standard deviation of about £29, pointing to a significant spread in the unit prices across different products. The median price is £1.95, which means half of the products cost less than this amount.
- Prices start at £0 and go up to an unusually high of £10,935.5, which could signify potential outliers.
    - Box plots are needed to further explore the outlier's issue.

	count	mean	std	min	25%	50%	75%	max
Quantity	779495.0	13.507085	146.540284	1.0	2.00	6.00	12.00	80995.0
Price	779495.0	3.218199	29.674823	0.0	1.25	1.95	3.75	10953.5

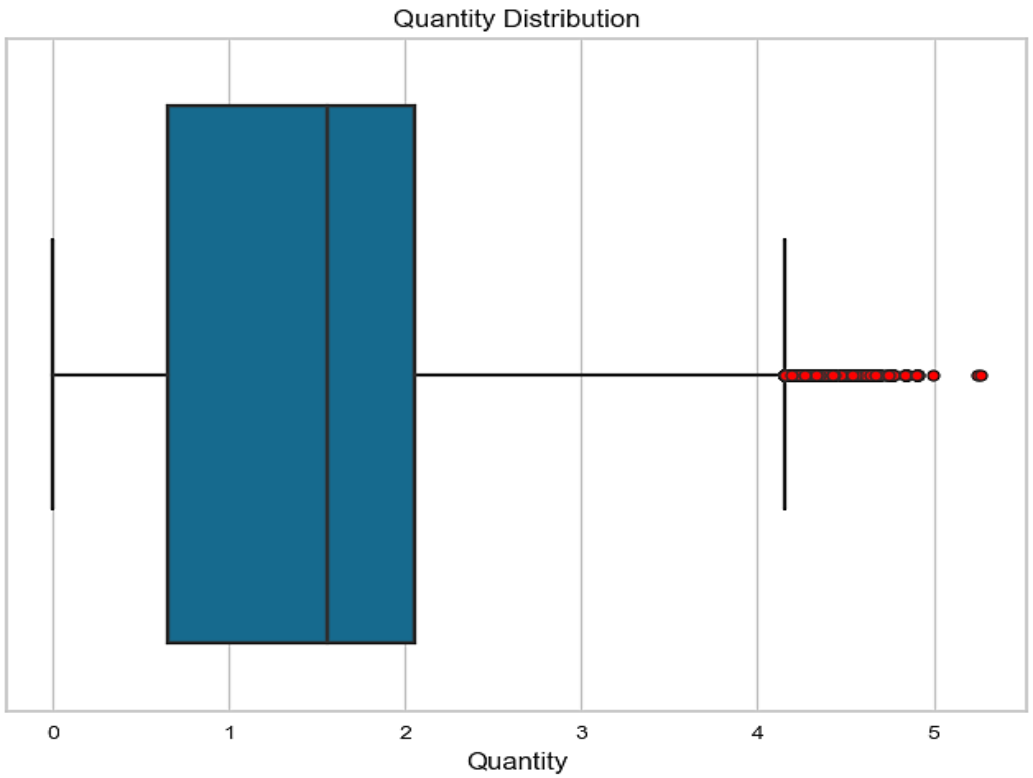
- Both Quantity and Price distributions are extremely right-skewed with a concentration of data near the lower end and outliers that suggest high-quantity/high-price transactions.
- The extreme outliers in both plots could potentially distort any predictive models and might need to be addressed through further data cleaning or transformation.

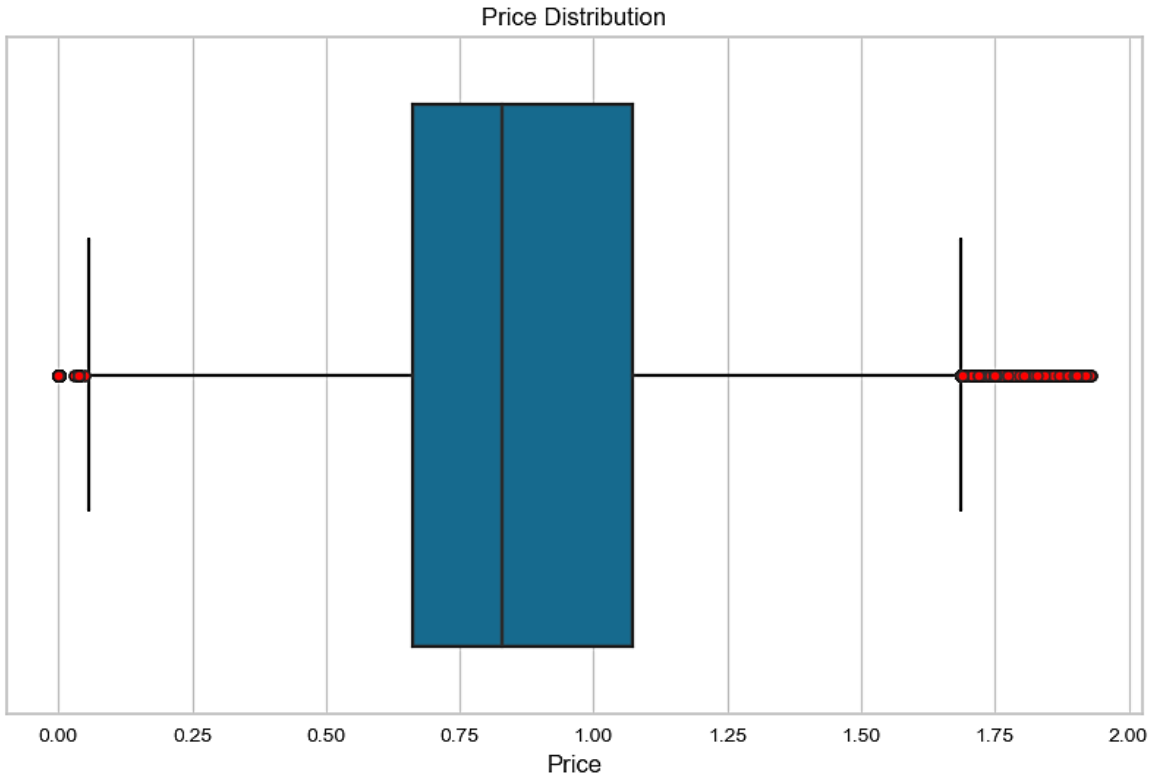




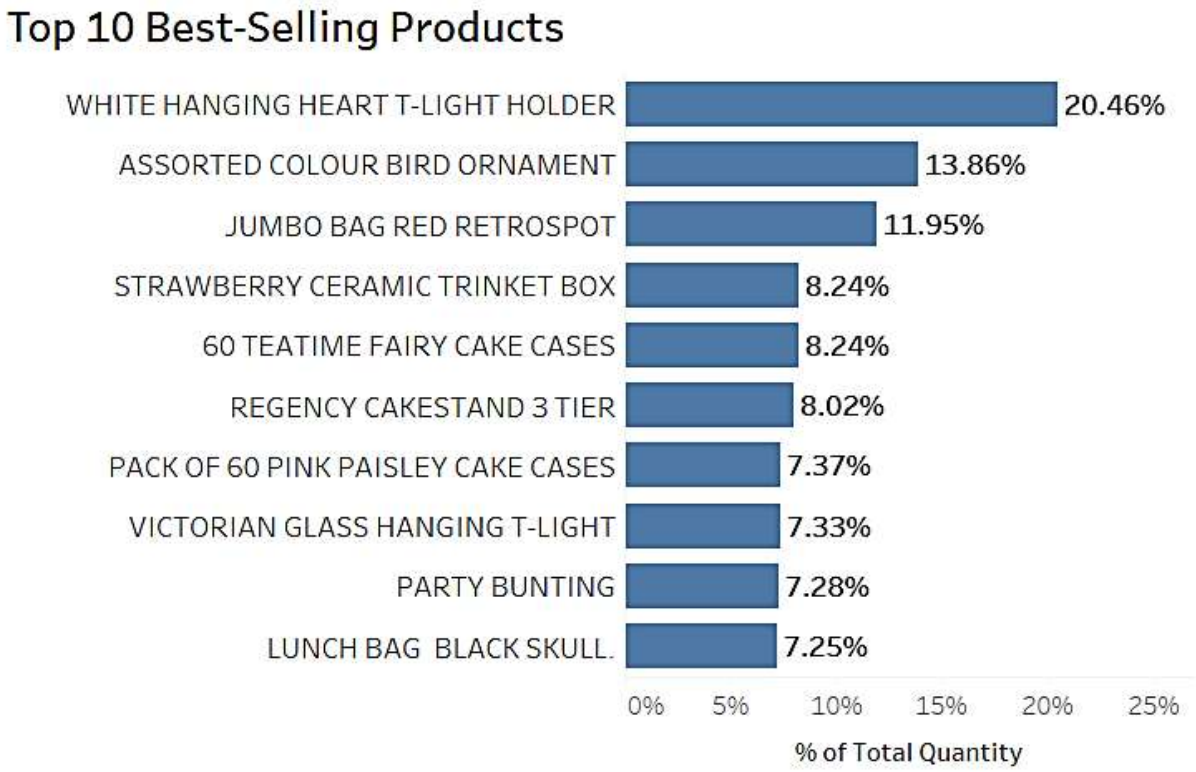
- Applied the Box-Cox method to adjust the scales of the data to reduce skewness. The λ parameters below indicate that the data required a moderate transformation to approach a normal distribution.
- As we can see from the plot, after transformation, the data might still have some outliers, but they are less extreme compared to the original data.

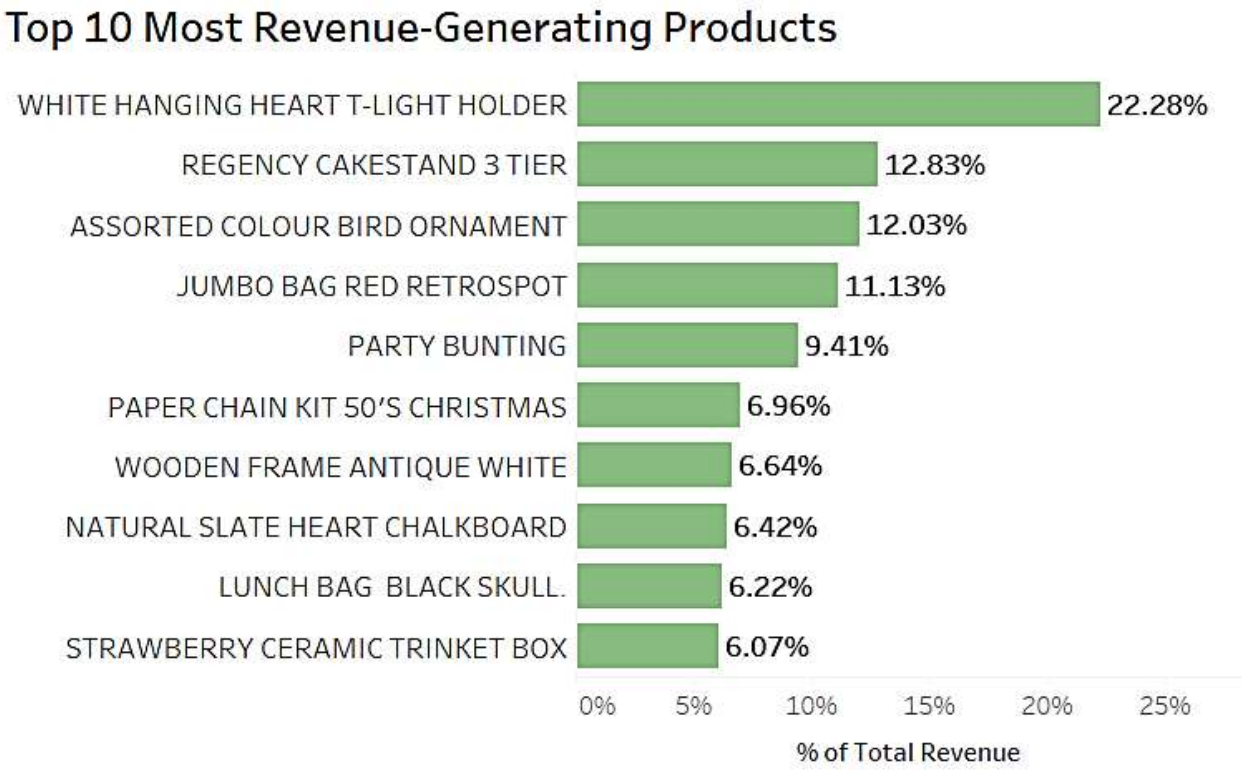
```
Fitted Lambda for Quantity: -0.15818977777916324
Fitted Lambda for Price: -0.5139244435400433
```



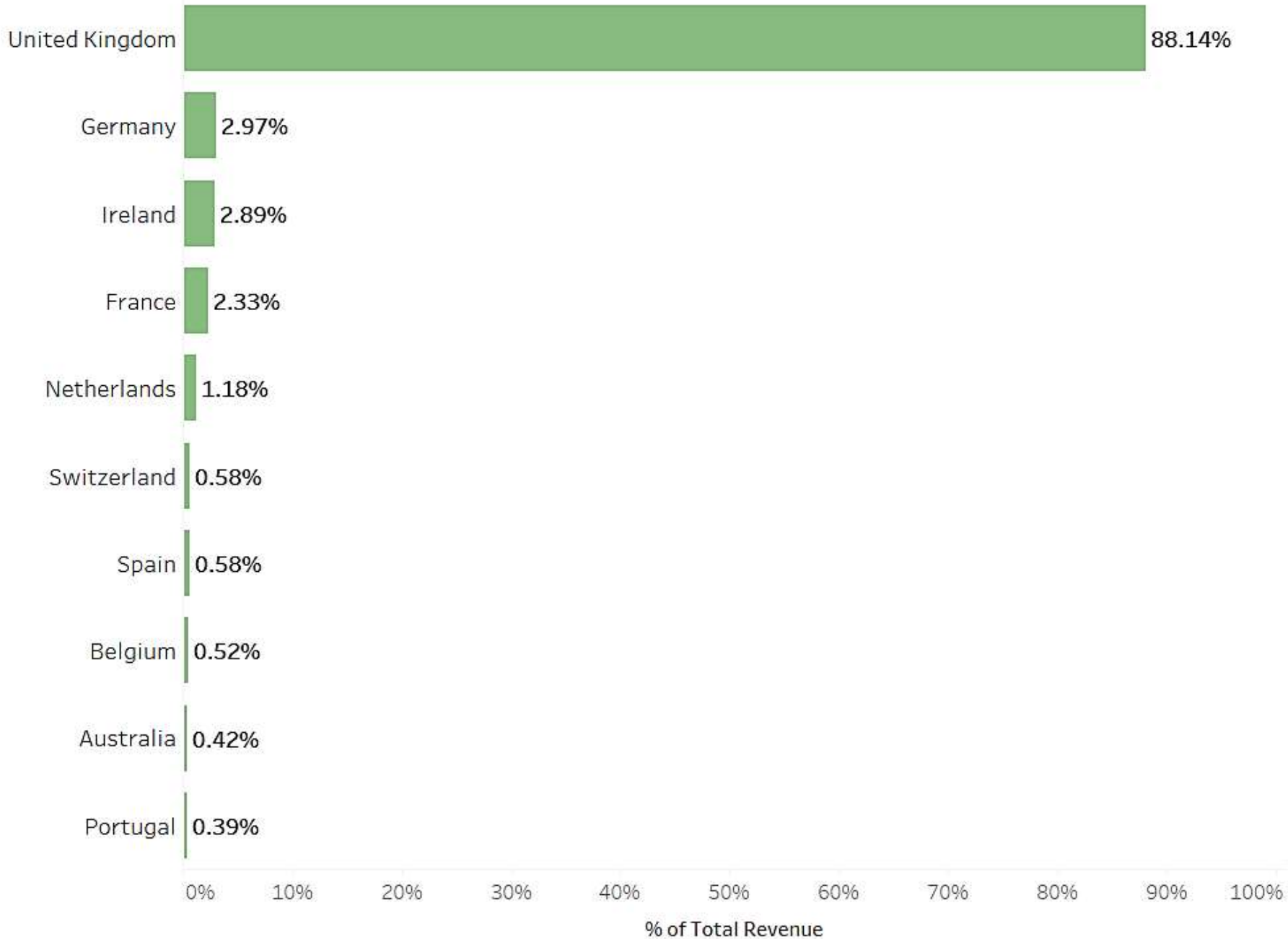


Visualization of Key Metrics

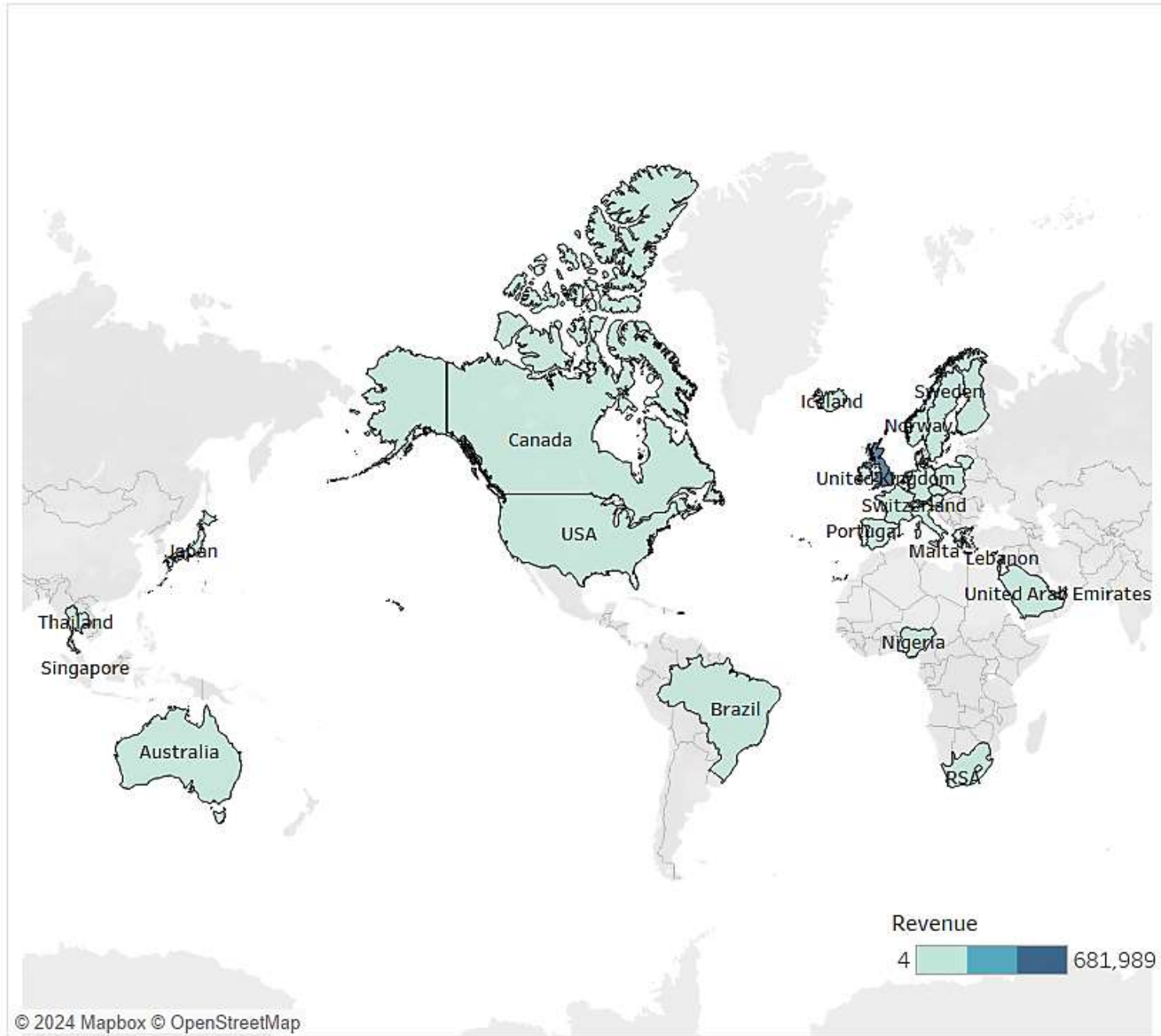




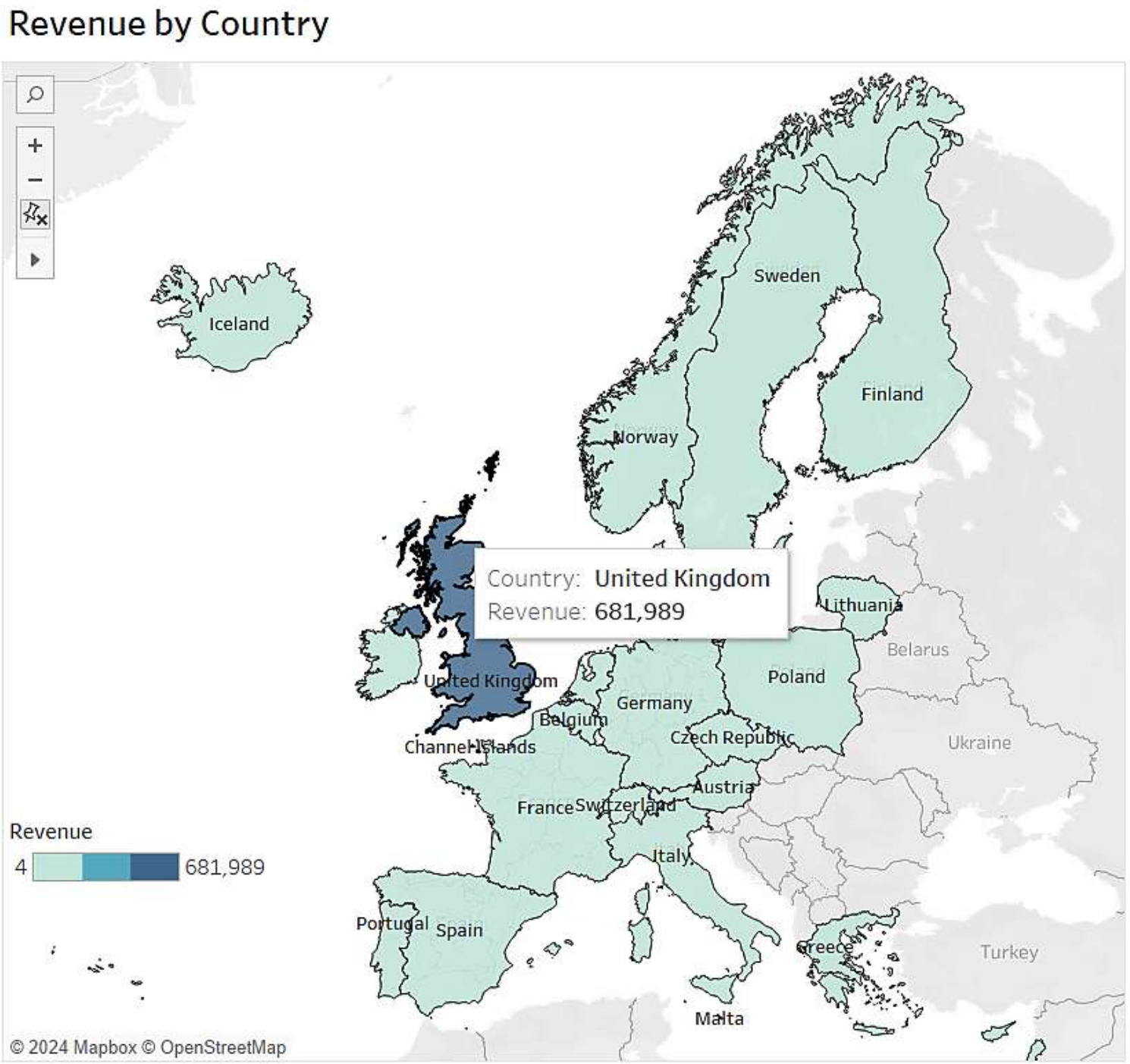
Top 10 Most Revenue-Generating Countries



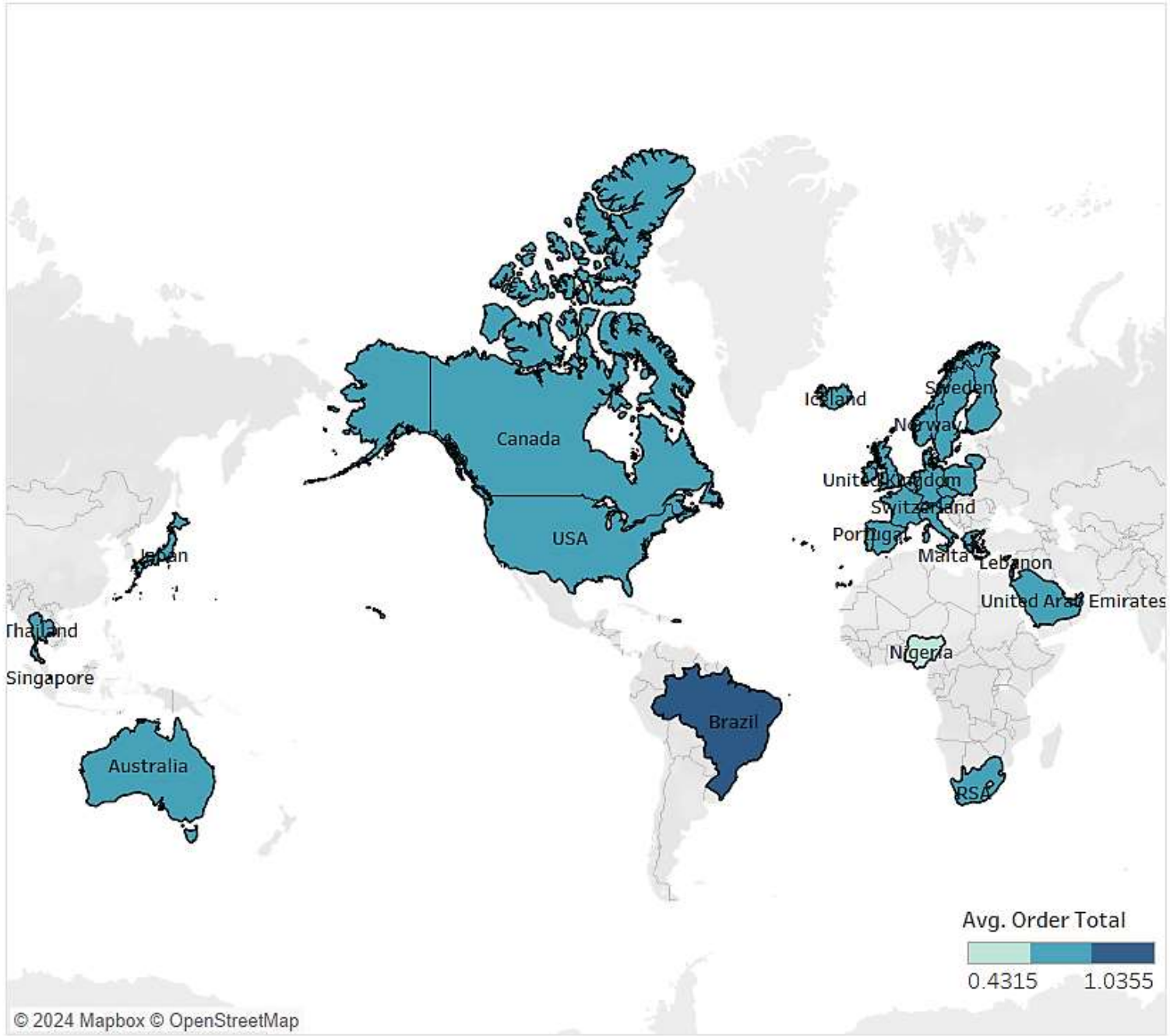
Revenue by Country



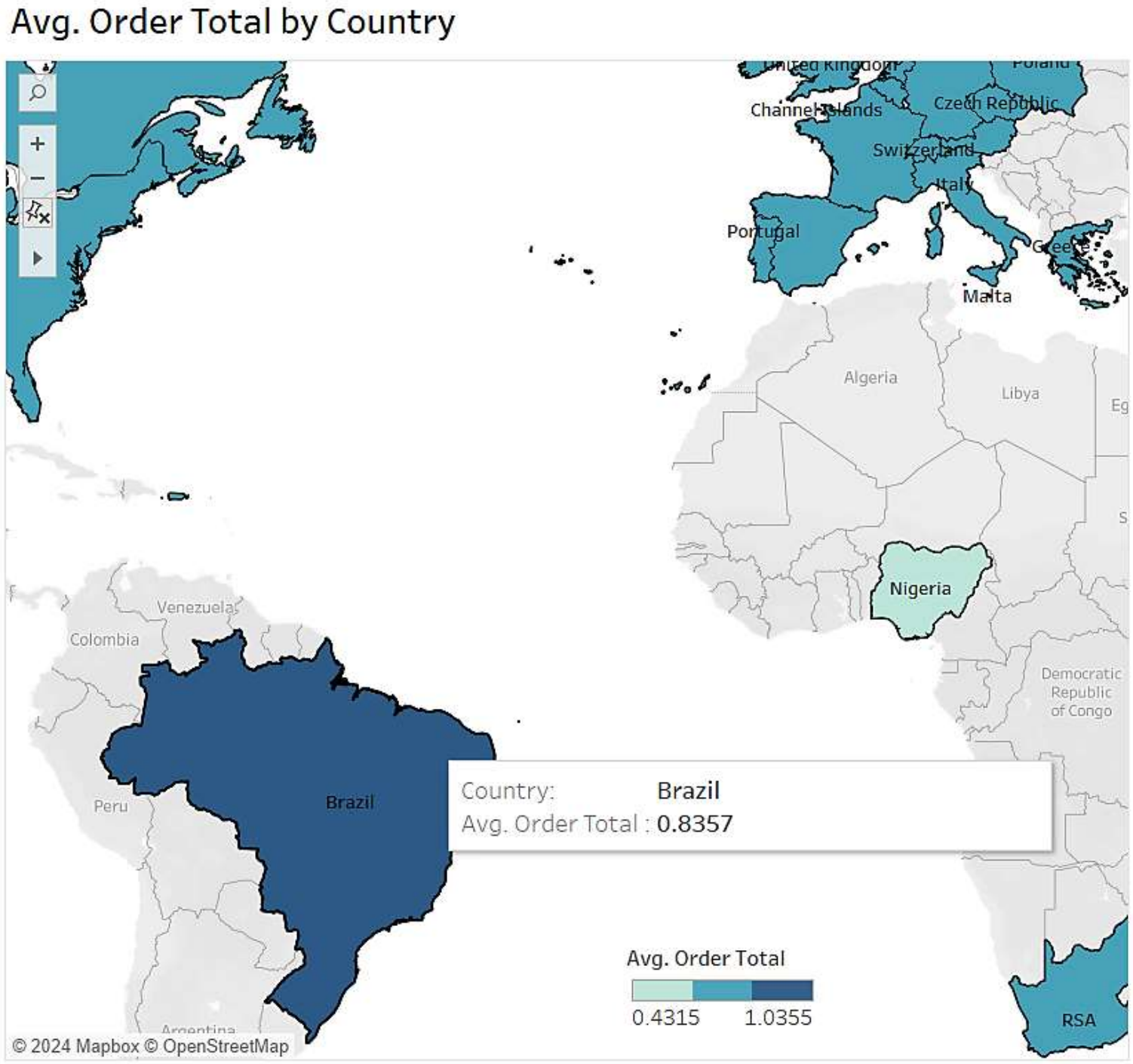
A closer look:



Avg. Order Total by Country



A closer look:

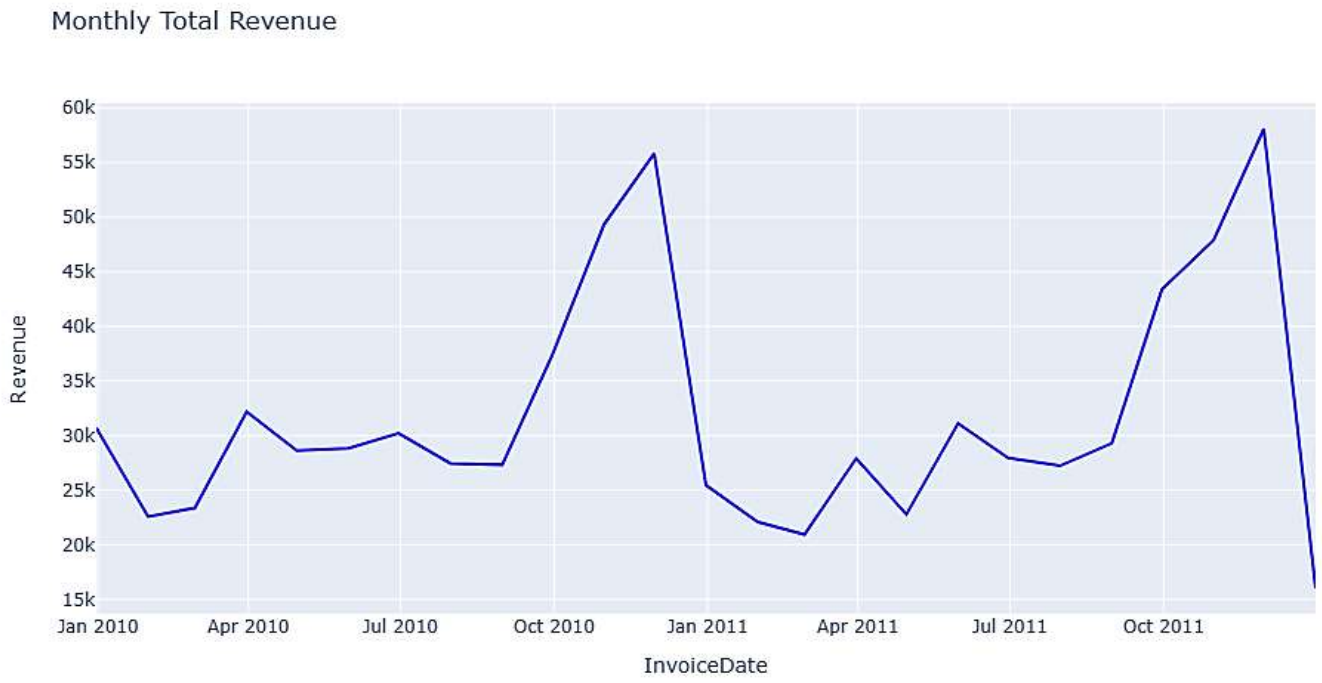


4. Predictive Modelling

InvoiceDate	Revenue
2009-12-31	30686.763563
2010-01-31	22575.973726
2010-02-28	23377.060658
2010-03-31	32173.947300
2010-04-30	28631.424498
2010-05-31	28833.369061
2010-06-30	30206.577936
2010-07-31	27450.031860
2010-08-31	27328.357241
2010-09-30	37379.568207
2010-10-31	49337.813580
2010-11-30	55831.263165
2010-12-31	25470.848184
2011-01-31	22092.918111
2011-02-28	20940.332543
2011-03-31	27911.204366
2011-04-30	22782.549860
2011-05-31	31117.802547
2011-06-30	27941.895974
2011-07-31	27234.185680
2011-08-31	29292.085126
2011-09-30	43374.243437
2011-10-31	47892.772283
2011-11-30	58083.732227
2011-12-31	16034.980337

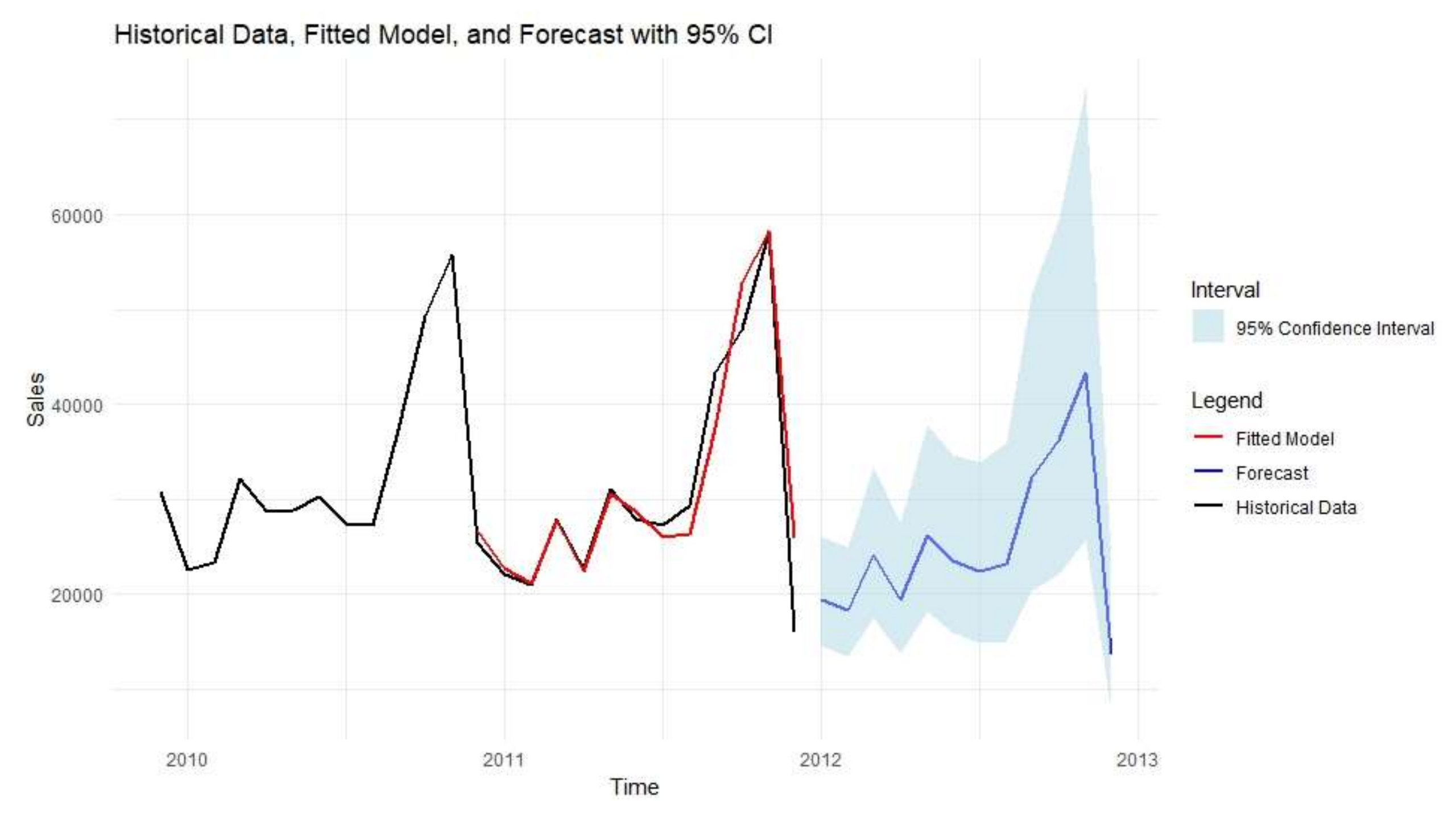
Time Series Forecasting: (Supervised Model; relies on historical data)

- Methodology:
  - What is a time series model?
    - A collection of points over time that record certain events. For instance, in this case scenario, it is the total monthly revenue/sales a retail store generated over a span of 2 years.



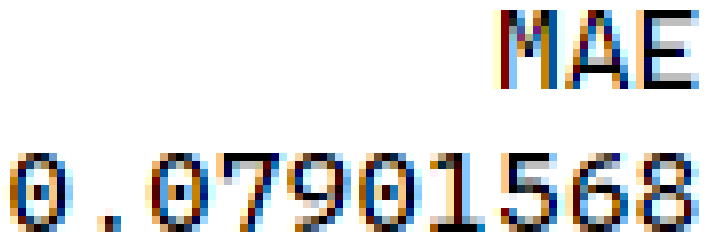
A closer look:

Notice the seasonal trend occurring at the same time of the year in November.


Historical Data, Fitted Model, and Forecast with 95% CI
60000
40000
20000
Sales
2010
2011
2012
2013
Time
Interval
95% Confidence Interval
Legend
Fitted Model
Forecast
Historical Data

Model Accuracy

Given that the Mean Absolute Error (MAE) is quite low, the model has a good average performance which indicates that the model’ s predictions on average are generally close to the actual revenue values, making it reliable.



Customer Segmentation:

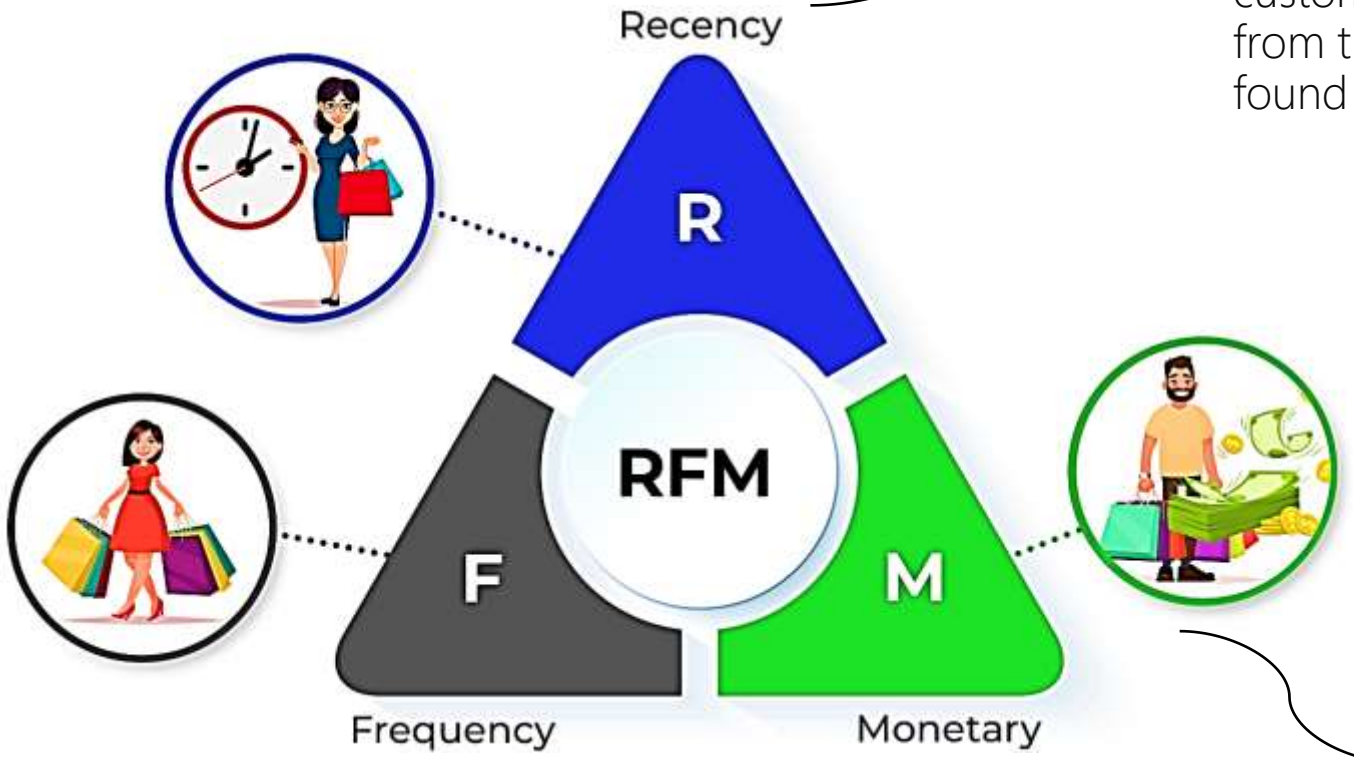
- With the information we have so far, the retail store is facing an issue in that:
    - United Kingdom: drives all sales and contributes to most of the revenue but has a very low average transaction size.
    - All other countries: contribute much less to total revenue but have very high average transaction totals.

Application of RFM Analysis:

- Therefore, we will be splitting our customers into two groups:
    - those who are only from the UK.
    - those who are from any of the 40 other countries.
- After that, we will be proposing a marketing tool for each group known as the RFM (Recency, Frequency, Monetary) in hopes of resolving the issue.

How often does each customer purchase?
counts how many unique transactions each customer has made.

How recent was the customer’ s last purchase?
finds the most recent date of purchase for each customer and subtracts it from the latest date found in the dataset.



How much does each customer spend?
adds up all the revenue values from transactions associated with each customer.

After calculating Recency, Frequency, & Monetary values for each customer, we add an RFM Score which is a composite score combining these three factors.

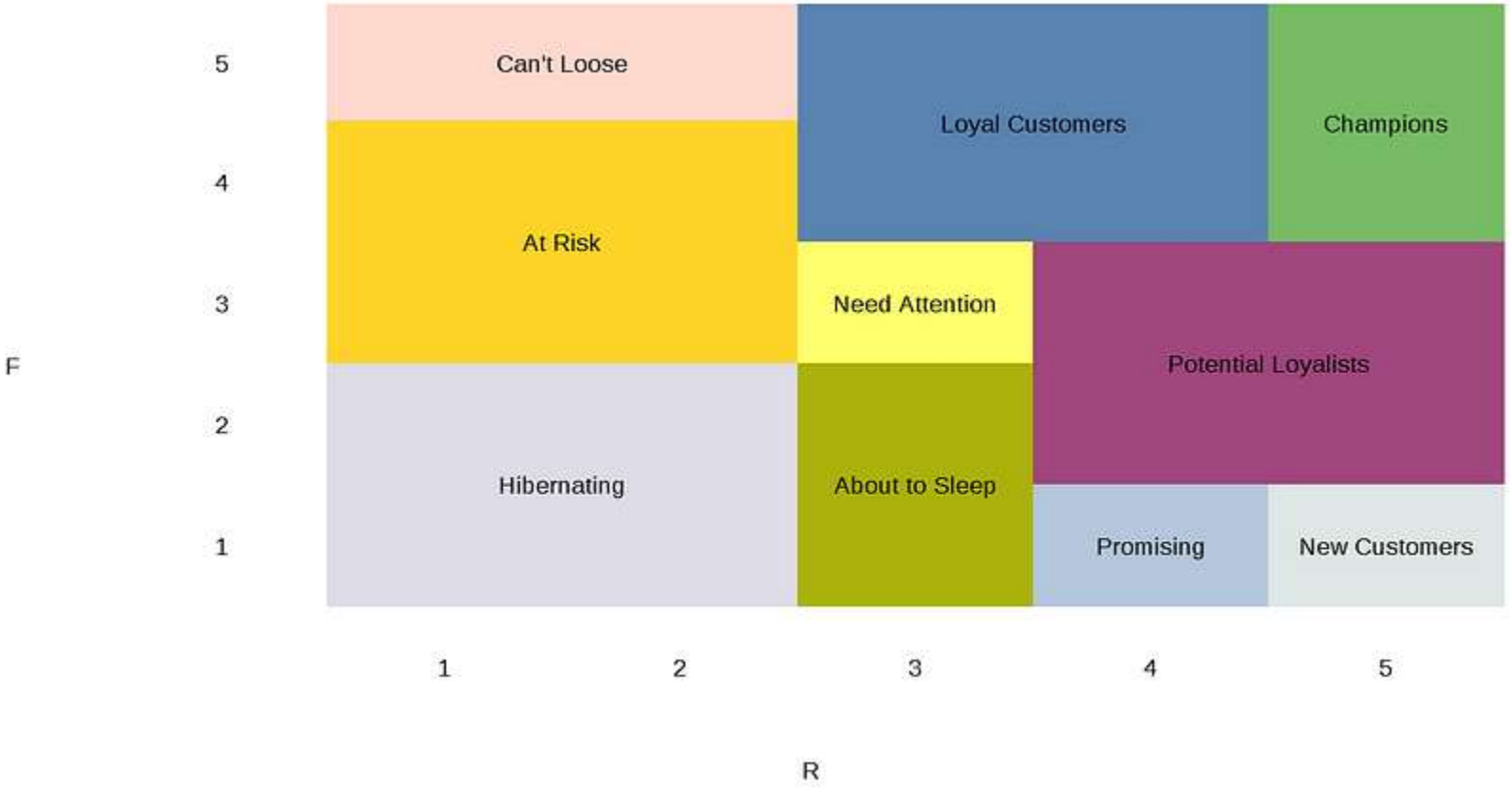
- For recency, each customer is first assigned a label from 1 to 5, with 5 being the most recent and 1 being the least recent.
- For frequency & monetary, customers with the lowest values get a label of 1, and those with the highest get a label of 5.

Customer ID	Recency	Frequency	Monetary	Cluster	R	F	M	RFM_Score
12347.0	1.202204	1.506548	6.625626	0	5	3	3	53
12348.0	6.272055	1.249917	4.465685	4	3	3	2	33
12349.0	3.770935	1.113484	6.215840	0	4	3	3	43
12350.0	9.458893	0.000000	3.458756	3	1	1	2	11
12351.0	9.946331	0.000000	3.503964	1	1	1	2	11

Segment Mapping:

Each customer is then assigned to a segment based on their RFM Composite Score. This segmentation allows us to tailor marketing strategies appropriate to each group's behaviors and preferences.

```
map = {
    r'[1-2][1-2]': 'Hibernating',
    r'[1-2][3-4]': 'At Risk',
    r'[1-2]5': 'Can''t Loose',
    r'3[1-2]': 'About to Sleep',
    r'33': 'Need Attention',
    r'[3-4][4-5]': 'Loyal Customers',
    r'41': 'Promising',
    r'51': 'New Customers',
    r'[4-5][2-3]': 'Potential Loyalists',
    r'5[4-5]': 'Champions'
}
```



Example demo:

RFM_Score	Segment
53	Potential Loyalists
33	Need Attention
43	Potential Loyalists
11	Hibernating
11	Hibernating

Segment Descriptions:

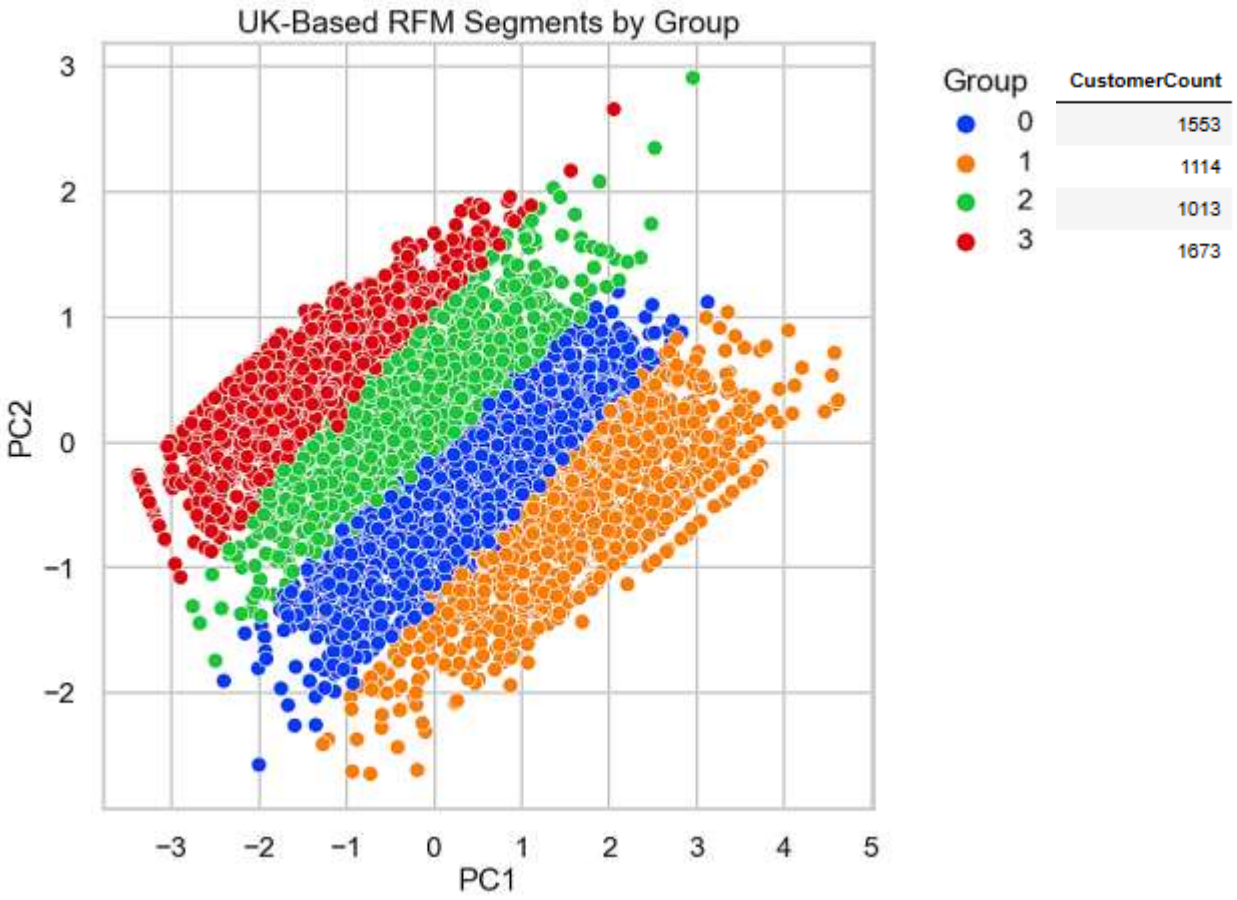
Customer Segment	Activity
Champions	Bought recently, buy often and spend the most!
Loyal Customers	Spend good money with us often. Responsive to promotions.
Potential Loyalist	Recent customers, but spent a good amount and bought more than once.
Recent Customers	Bought most recently, but not often.
Promising	Recent shoppers, but haven't spent much.
Customers Needing Attention	Above average recency, frequency and monetary values. May not have bought very recently though.
About To Sleep	Below average recency, frequency and monetary values. Will lose them if not reactivated.
At Risk	Spent big money and purchased often. But long time ago. Need to bring them back!
Can't Lose Them	Made biggest purchases, and often. But haven't returned for a long time.
Hibernating	Last purchase was long back, low spenders and low number of orders.

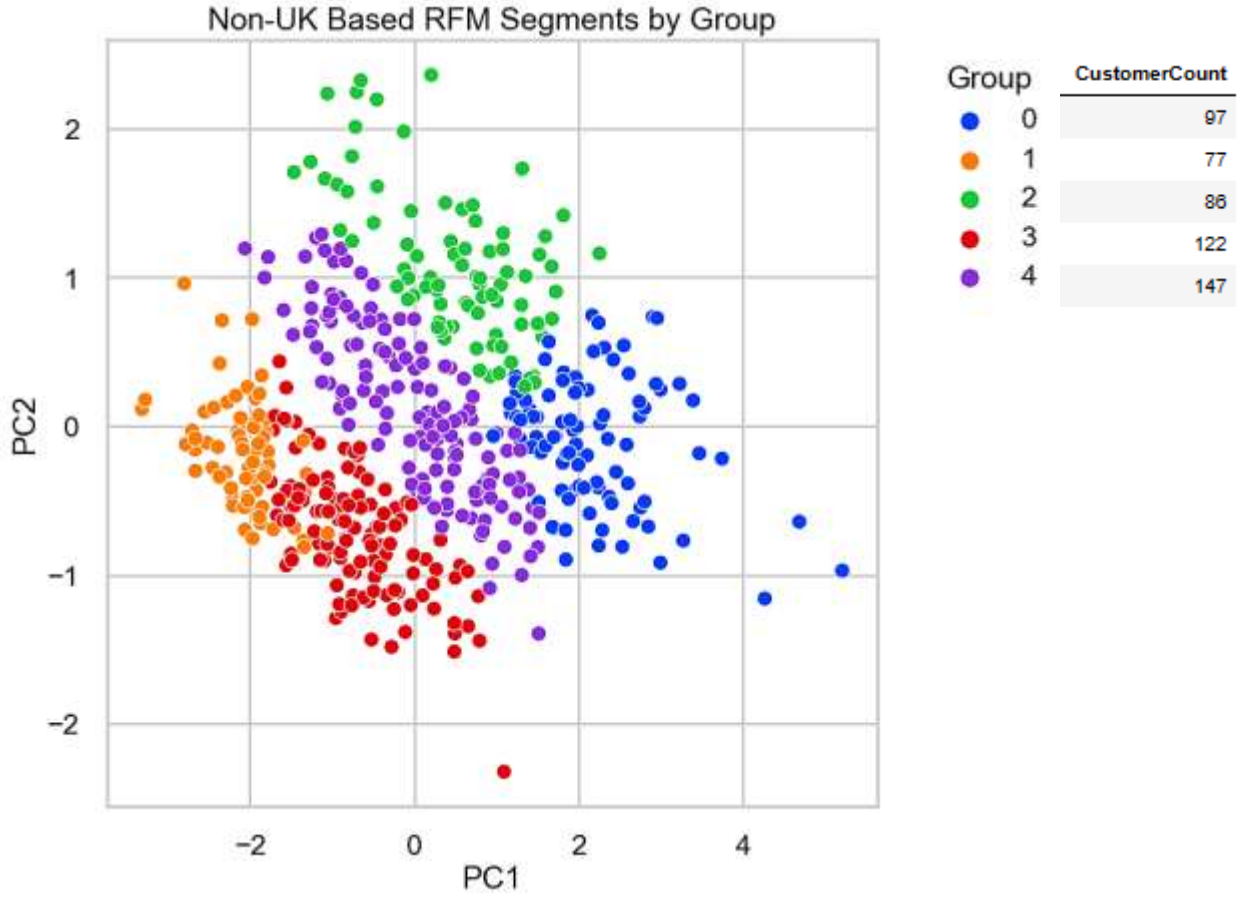
This RFM segmentation will readily answer key questions for your business:

- Who are my best customers?
- Which customers are at the verge of churning?
- Who has the potential to be converted into more profitable customers?
- Who can you view as lost customers?
- Which customers must you retain?
- Who are your loyal customers?
- Which group of customers is most likely to respond to your current campaign?

Clustering with K-means: (Unsupervised model; discovers hidden patterns)

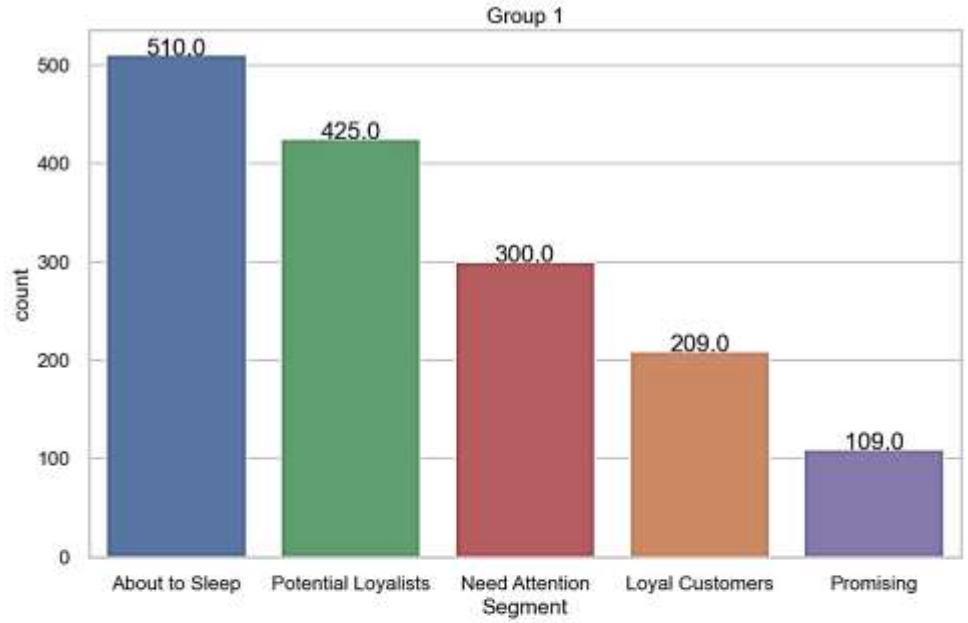
- Now that I have both groups' RFM Tables, I applied a model known as K-means which functions as a way to organize our big, mixed group of data points (customers) into smaller, more similar groups based on common features that the algorithm detects.

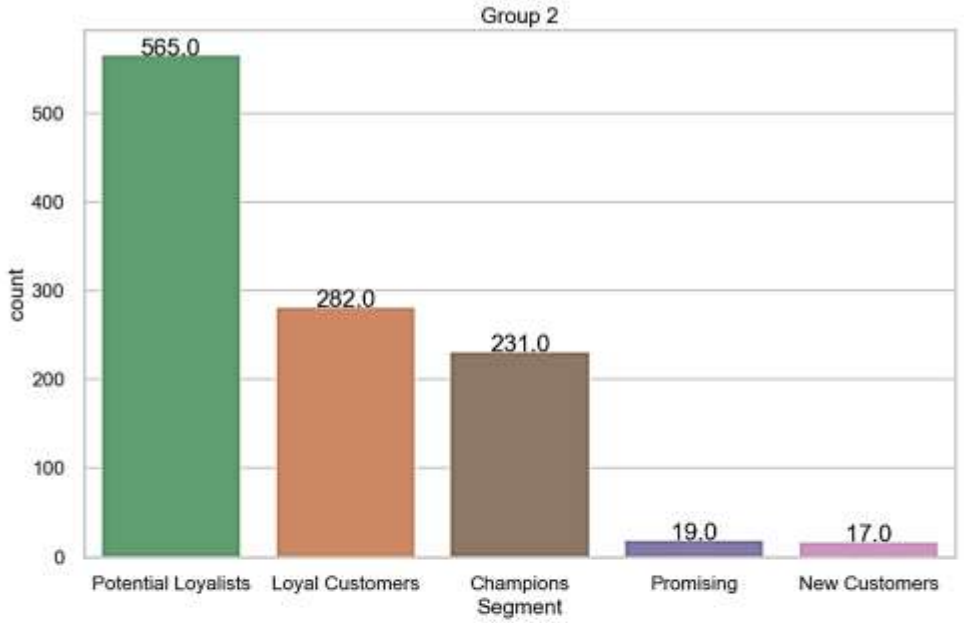


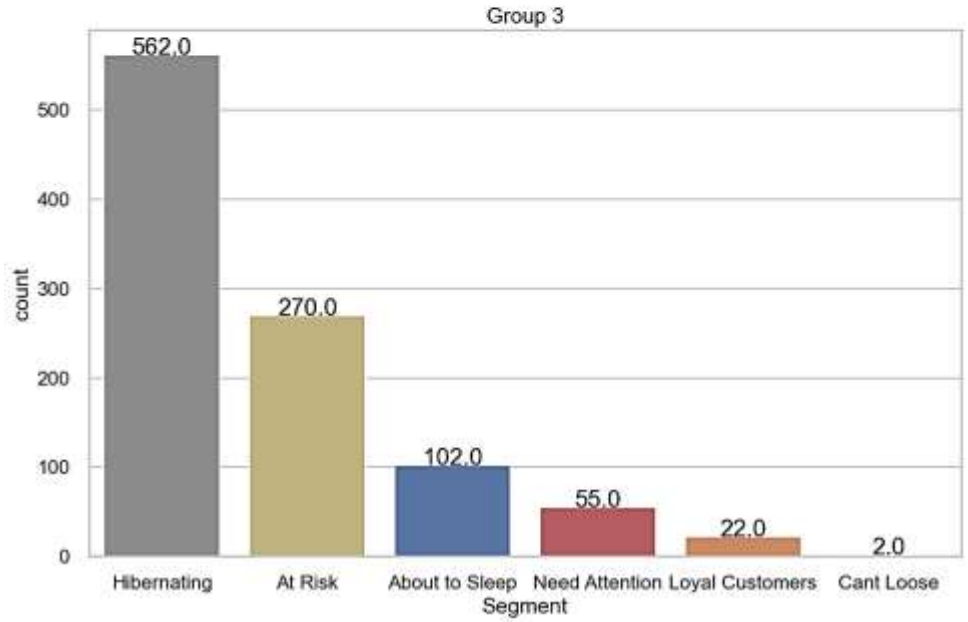


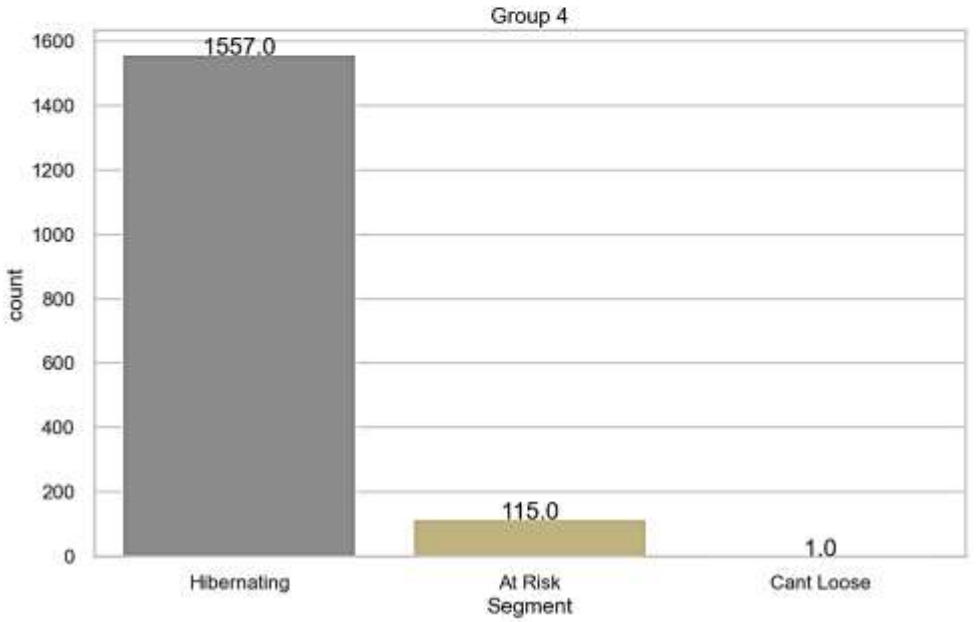
Details on Customer Groups:

- UK-Based:

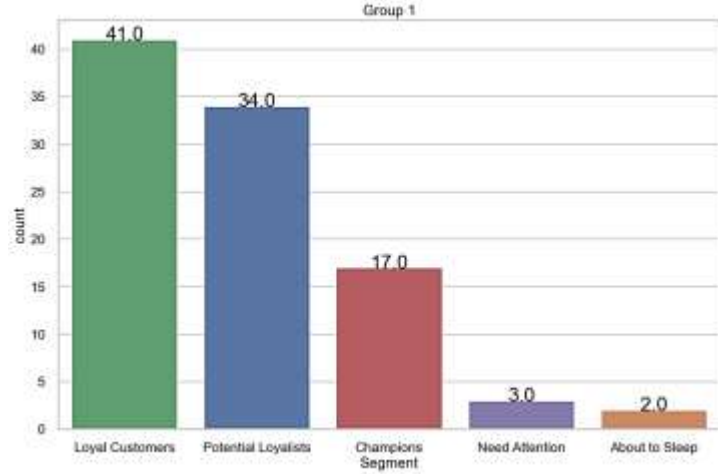


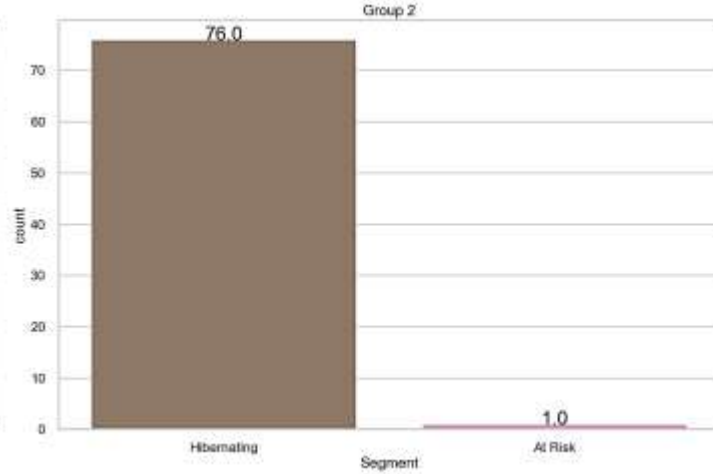


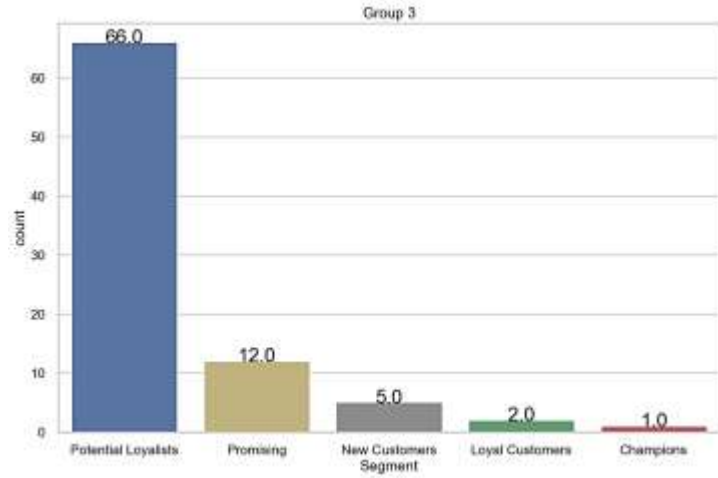


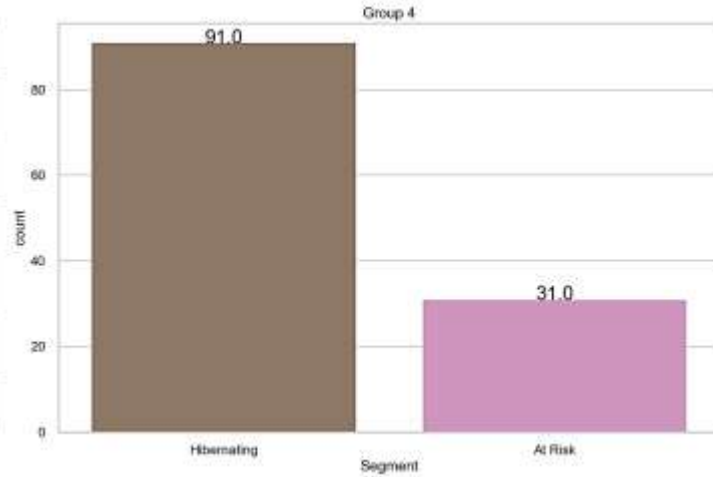


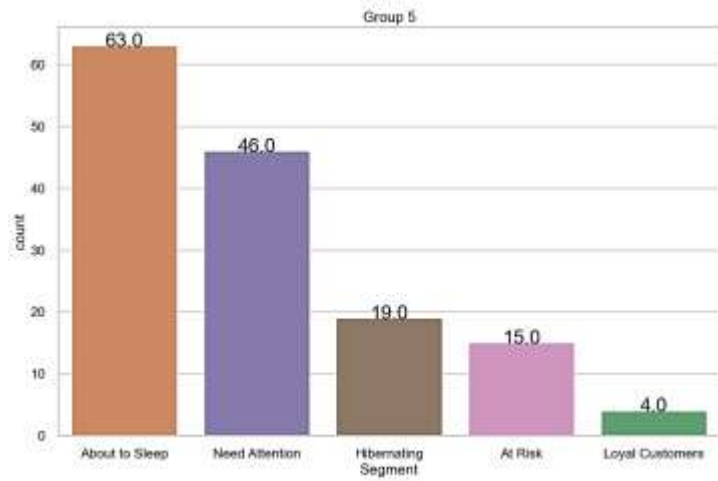
- Non-UK-Based:


Group 1
count
41.0
34.0
17.0
3.0
2.0
Loyal Customers
Potential Loyalists
Champions
Need Attention
About to Sleep
Segment


Group 2
count
76.0
1.0
Hibernating
At Risk
Segment


Group 3
count
66.0
12.0
5.0
2.0
1.0
Potential Loyalists
Promising
New Customers
Loyal Customers
Champions
Segment


Group 4
count
91.0
31.0
Hibernating
At Risk
Segment


Group 5
count
63.0
46.0
19.0
15.0
4.0
About to Sleep
Need Attention
Hibernating
At Risk
Loyal Customers
Segment

Moving forward, the store’ s strategies should be improved to cater to the specific needs and behaviors of each segment as shown in the table on the right.

To implement those strategies, the store will need to pull out the Customer IDs pertaining to each segment to get more information on how to reach them and a range of communication strategies such as:

1. Use insights from RFM analysis to personalize communication and offers.
2. Reach out through their preferred channels.
3. Collect feedback after each campaign to refine the approach continually.
4. Monitor key performance indicators to measure the success of targeted strategies for each segment.

Customer Segment	Actionable Tip
Champions	Reward them. Can be early adopters for new products. Will promote your brand.
Loyal Customers	Upsell higher value products. Ask for reviews. Engage them.
Potential Loyalist	Offer membership / loyalty program, recommend other products.
Recent Customers	Provide on-boarding support, give them early success, start building relationship.
Promising	Create brand awareness, offer free trials
Customers Needing Attention	Make limited time offers, Recommend based on past purchases. Reactivate them.
About To Sleep	Share valuable resources, recommend popular products / renewals at discount, reconnect with them.
At Risk	Send personalized emails to reconnect, offer renewals, provide helpful resources.
Can't Lose Them	Win them back via renewals or newer products, don't lose them to competition, talk to them.
Hibernating	Offer other relevant products and special discounts. Recreate brand value.

5. Results

Here's what we've learned:

- Time series forecasting indicates a strong seasonal pattern. The model predicts a similar pattern to continue into the future year with high confidence.
- The RFM analysis has revealed key customer groups, each with distinct behaviors. The UK market, while high in transactions, shows smaller average spending. The international segments contribute less in volume but more in transaction size.
    - UK-Based: Loyal Customers and Potential Loyalists represent significant segments, suggesting the value in improving these relationships.
    - Non-UK-Based: There's a prevalence of Hibernating and At Risk segments, highlighting a need for re-engagement strategies.

Actionable Strategies:

- The RFM table provides clear, actionable tips for each segment, and we will leverage these to deliver tailored marketing efforts and communication strategies, ensuring our retail store thrives in the competitive market.