Online Retail Report

- **Introduction**

  Online Retail, an e-commerce company, recently invested a significant portion of its revenue in an advertising campaign aimed at increasing brand and product awareness. Despite these efforts, the campaign only achieved an acquisition response rate of 3%, falling short of the projected 6%. Management suspects that the underperformance of the campaign resulted from its broad and costly approach, which did not take into account the diverse purchasing behaviors of customers.

  To improve future results, the company intends to focus its marketing efforts on customers most likely to drive revenue growth. With the next campaign scheduled in six months, management aims to achieve the following objectives:

  - **Customer Value Analysis:** Assess the commercial value of each customer just before the campaign launch.
  - **Customer Segmentation:** Develop a segmentation strategy based on purchasing behaviors to identify key customer groups.
  - **Marketing Enablement Tool:** Equip the marketing team with a tool to implement and sustain a targeted marketing strategy.

  The success of this project will be measured by achieving the targeted response rate of 6%, a key performance indicator set by management.

  The Data Science team has been assigned to lead this project. They will collaborate with the marketing team responsible for promotions, the technology team, and a representative from the management committee. Although the company's database contains some gaps due to past system migrations, it will serve as the foundation for this initiative.

- **Dataset**

  The Online Retail dataset can be downloaded from Kaggle by clicking here. It initially consists of the following columns:

  - **CustomerID:** Unique number to identify each customer.
  - **InvoiceDate:** Date of product purchases by a customer, covering all transactions from the first purchase up to the present date.
  - **InvoiceNo:** Invoice number recording customer purchases.
  - **StockCode:** Unique code identifying the purchased product.
  - **Description:** Product description supplementing the StockCode.
  - **Quantity:** Number of units of a product purchased in a single transaction.

- **UnitPrice:** Unit price of the selected product.
- **Country:** Indicates the country of the purchasing customer.

- **Data Cleaning and Processing**

  To effectively analyze the dataset, the following data preparation steps were necessary:

  - Convert specific columns into appropriate formats:
    - **InvoiceDate:** Object → Datetime
    - **Country:** Object → Category
  - Remove records where **CustomerID** is missing. A total of **135,080** rows were removed.
  - Introduce new columns to enhance the analysis of the company's activities over time and geographically, as well as to understand customer behavior:
    - **Continent:** Continent where the customer is located.
    - **Month, Year, Day:** Breakdown of purchase dates.
    - **Recency:** Date of the customer's last purchase.
    - **Frequency:** Number of purchases made by the customer.
    - **Revenue:** Revenue generated per customer.
  - Standardize country names:
    - **'USA'** and **'RSA'** were replaced by **'United States'** and **'Republic of South Africa'**, respectively.
    - **'European Community'** was renamed **'Other European Country'** to avoid confusion.

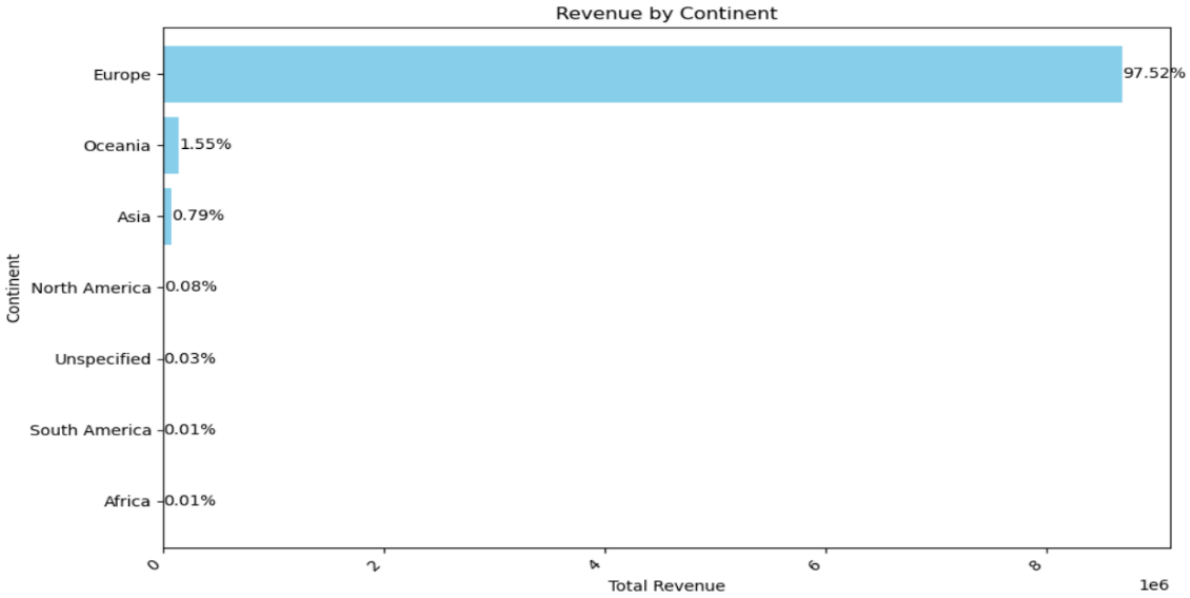
- **Exploratory Data Analysis**

  The dataset covers a period of **373 days**, from **December 1, 2010, to December 9, 2011**. The key business activity insights are:

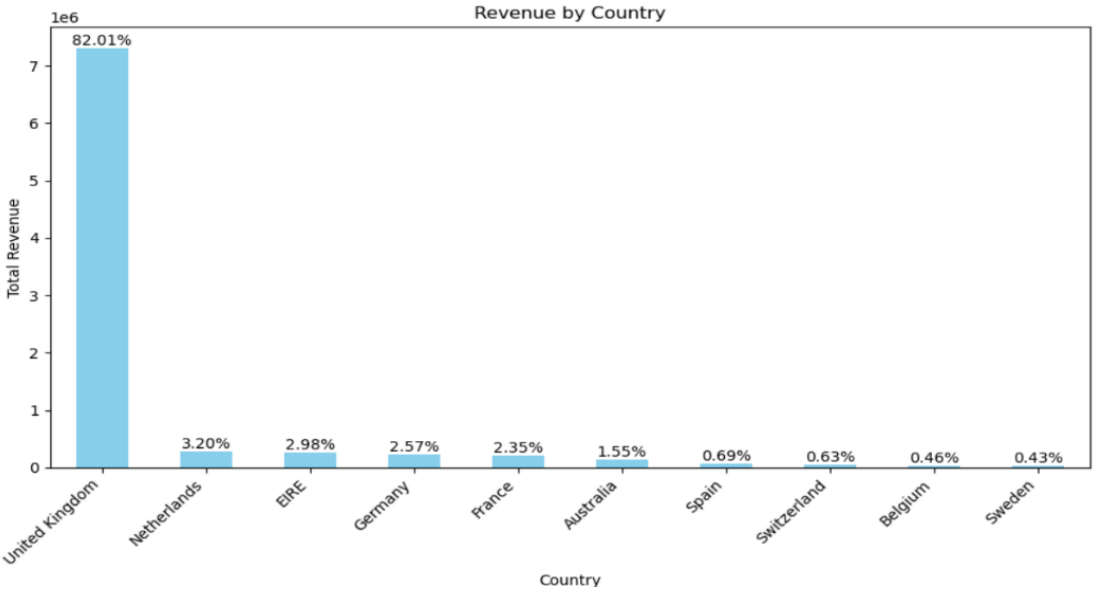
Number of Customers	Total Items Sold	Total Sales Revenue	Number of Unique Item Types sold	Number of Countries
4,372	4,908,888	$8,300,066	3,896	37

- **Market Analysis**

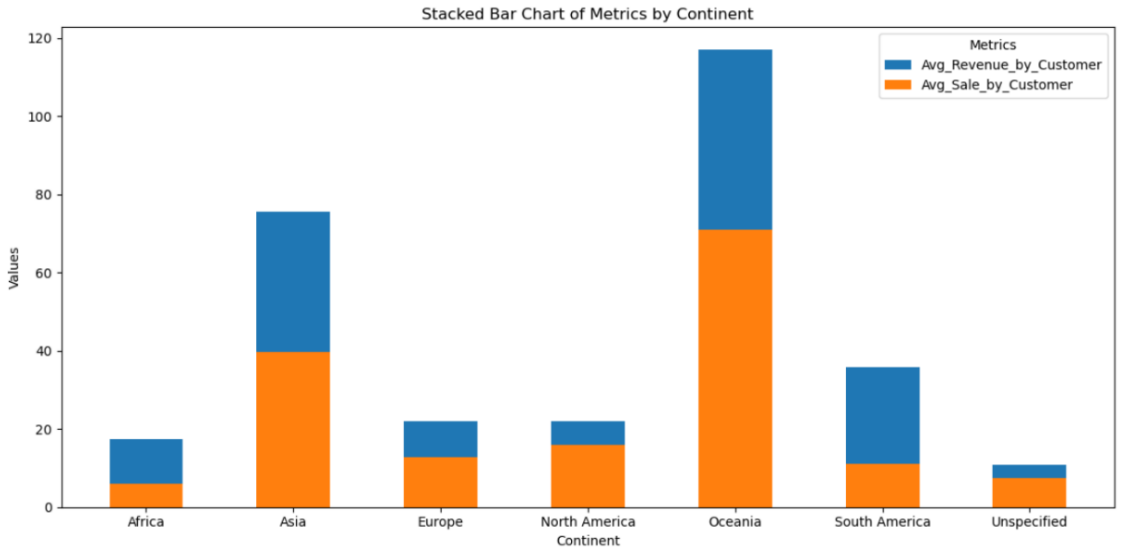
  The European market dominates Online Retail's revenue, accounting for **97.52%** of total revenue, followed by Oceania (**1.55%**) and Asia (**0.79%**).



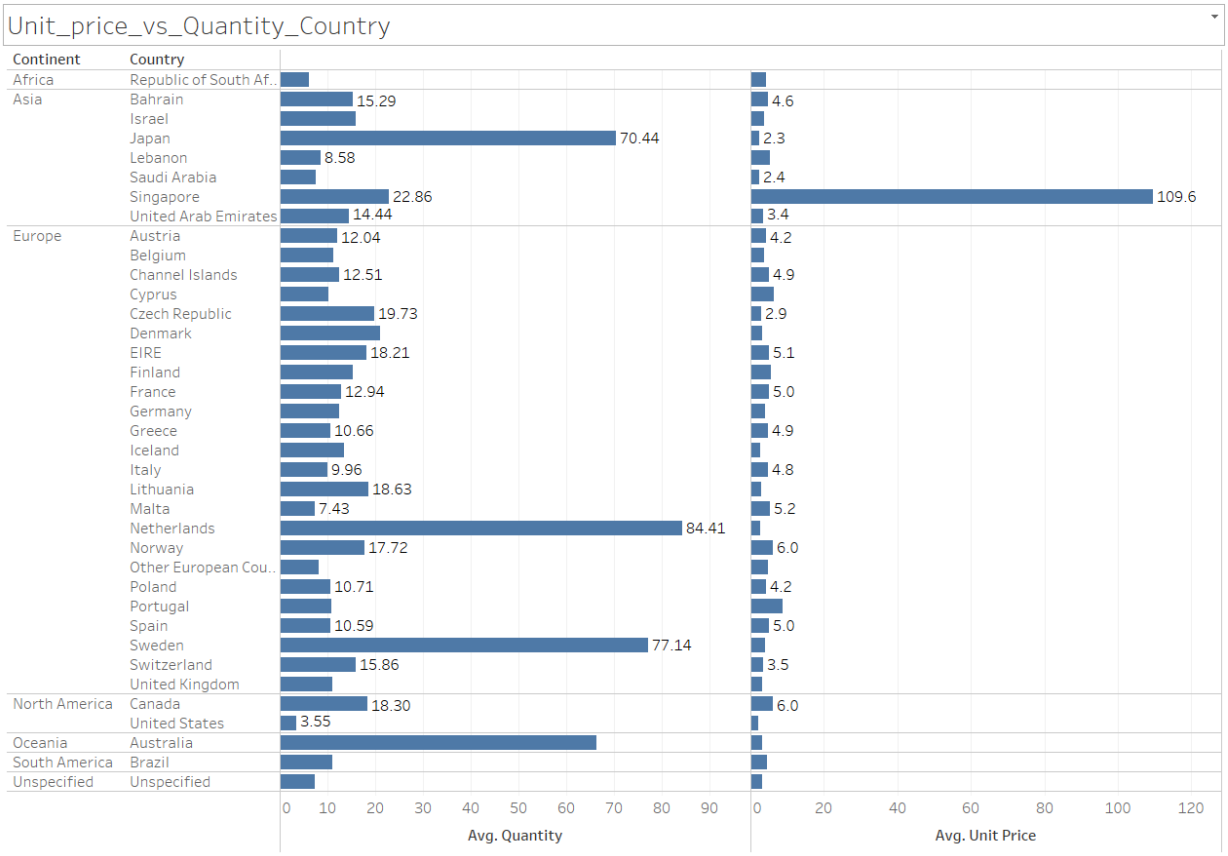
  The UK alone contributes **82%** of the company's revenue, making it the primary market focus.



- Despite their lower contribution to total revenue, customers from Oceania and Asia generate the highest **average revenue per customer**, making them attractive for future market strategies.



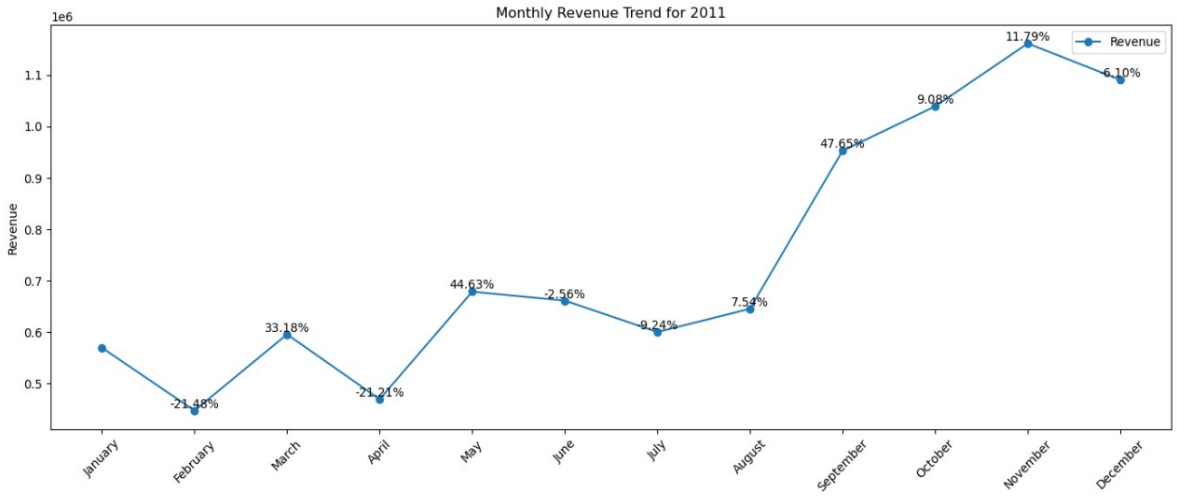
Analyzing the two key components of income at the country level reveals the following insights:



- Netherlands, Switzerland, Japan and Australia are the countries that order the most
- Singapore customers buy the most expensive products.

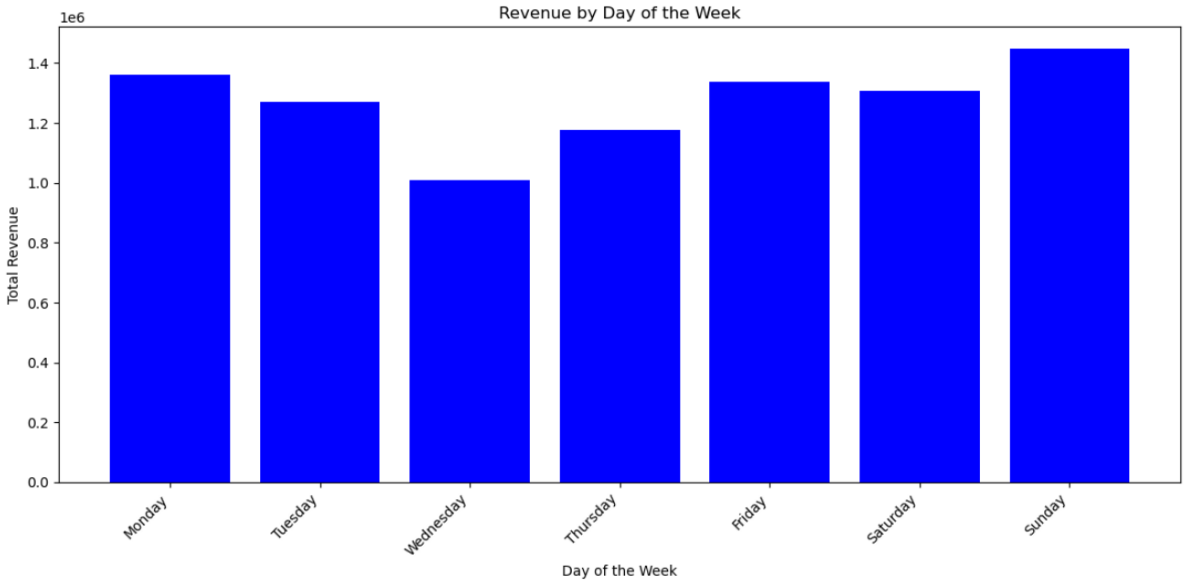
- **Revenue Trends**

  Revenue grew consistently throughout the year, peaking in **November**.



  A sharp growth of **47.65%** per month occurred between **August and September**, driven by demand in Europe.

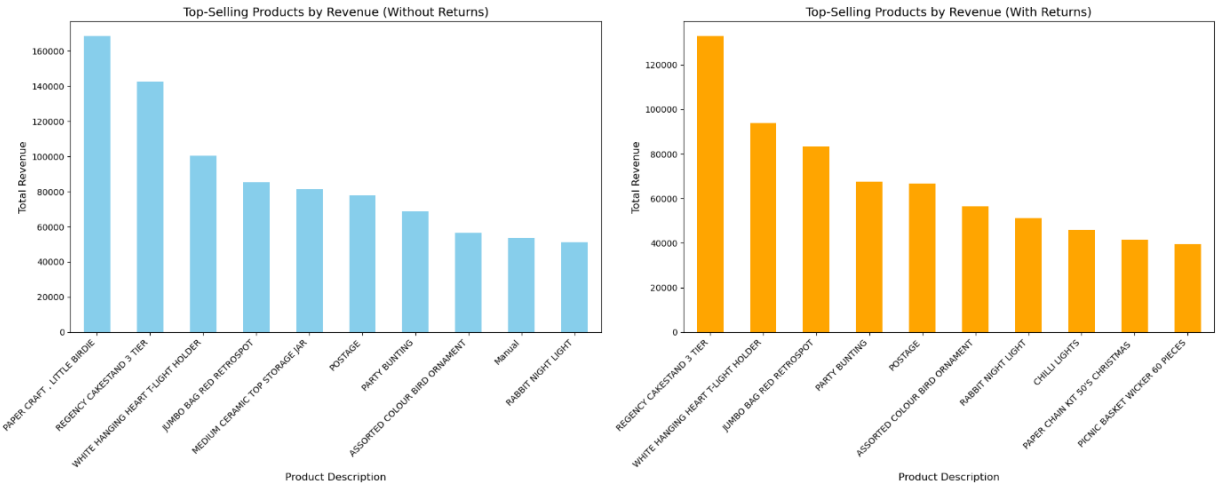
  On a weekly scale, **Sunday** generates the highest revenue, likely due to increased online shopping during weekends. **Wednesday** shows the lowest sales activity.



- Negative revenue

  Negative revenue values indicate **returns or error corrections**. Some high-demand products, such as **'PAPER CRAFT, LITTLE BIRDIE'** and **'MEDIUM CERAMIC TOP STORAGE JAR'**, have a **100% return rate**, making them critical points for further investigation.

The following graph shows the presence of these two products in the Top 10 product list when returns are not taken into account (left image) and their absence from this list when returns are introduced.



- **Customer Segmentation**

  The **RFM (Recency, Frequency, Monetary)** model was used to segment customers into three distinct groups:

	Recency	Frequency	MonetaryValue	
	mean	mean	mean	count
General_Segment				
Bronze	219.7	10.9	168.0	783
Gold	25.2	195.1	4130.3	1690
Silver	98.9	36.1	625.8	1899

- **Predicting Customer Lifetime Value (LTV)**

  Customer Lifetime Value (LTV) prediction was approached as a regression problem, later transformed into a **multi-class classification** problem by segmenting LTV into four categories:

  - **Top LTV**
  - **High LTV**
  - **Medium LTV**
  - **Low LTV**

  Three machine learning models were tested to predict the LTV category for each customer. The evaluation of the models led to the following results :

	Model	Accuracy_train	Accuracy_test	Accuracy
0	KNN	0.948898	0.949580	0.949580
1	Random Forest	0.982659	0.982143	0.982143
2	Gradient Boosting	0.984105	0.986870	0.986870

The evaluation results showed that **Gradient Boosting** outperformed the other models and was selected as the final predictive model.

- **Recommendations**

  The model allows for:

  - **More effective customer segmentation** based on revenue potential.
  - **Targeted marketing campaigns** tailored to different customer segments.
  - **Improved revenue forecasting and customer retention strategies**.

  Future actions should focus on:

  - Refining segmentation by incorporating behavioral data.
  - Expanding into high-value but underdeveloped markets (Oceania & Asia).
  - Investigating high return rates on certain products to improve profitability.

- This report outlines the Data Science team's findings and recommendations for optimizing Online Retail's customer engagement and revenue growth strategies.