Sales and Customer Analysis for a UK Online Retailer

By Axel Aranda Núñez
Aspiring Data Analyst &
Student of Industrial Organization

Index

Ask………pg.3

Prepare………pg.4

Process. …………………………………………………………………………………………………… pg.5

Analyse………pg.9

Share………pg.18

Recommendations & Next Steps………………………………………………………………………………………pg.19

Conclusions. ……………………………………………………………………………………………pg.19

Ask

Five questions will guide your case study:

1. **What type of company does your client represent, and what are they asking you to accomplish?**

My client is a mid-size e-commerce company based in the UK. They sell a wide range of consumer products online. The company wants to analyse its sales performance and customer behaviour over the past year to identify revenue trends, top-performing products, and key customer segments. Their goal is to make better decisions on inventory, marketing, and customer retention.

2. **What are the key factors involved in the business task you are investigating?**

The key factors include sales volume, pricing, geographic location (country), and customer purchasing behavior (frequency, basket size, returns). These will help determine which products and customers are most valuable to the business.

3. **What type of data will be appropriate for your analysis?**

I need detailed transaction data including invoice numbers, product IDs, quantity sold, sale dates, unit prices, customer IDs, and country of purchase.

4. **Where will you obtain that data?**

I am using a public dataset from Kaggle titled "Online Retail Dataset" provided by Ulrik Thyge Pedersen. It includes real sales transaction data from a UK-based online store between December 2010 and December 2011. Link:
https://www.kaggle.com/datasets/ulrikthygepedersen/online-retail-dataset

5. **Who is your audience, and what materials will help you present to them efectively?**

My audience is the company's management and business intelligence team. They are interested in clear, actionable insights presented through visual dashboards and summary reports. I will use tools like Tableau and PowerPoint to communicate the findings effectively.

Business Task
The business task is to analyze historical sales data from an online retail company to identify patterns in product performance, customer behavior, description and geographic trends. The goal is to provide actionable insights to help the company improve marketing strategies, inventory planning, and customer retention.

Prepare

1. Where is your data located?

The data is located on Kaggle and was downloaded as a CSV file. It is now stored in a dedicated project folder on my computer.

2. How is the data organized?

The dataset is structured as a flat table with one row per transaction. Each row includes:

- Invoice number
- Product code and description
- Quantity
- Invoice date
- Unit price
- Customer ID
- Country

It contains over 500,000 rows covering one year of sales.

3. Are there issues with bias or credibility in this data? Does your data ROCCC?

The dataset is publicly available and often used for educational purposes. It meets the ROCCC principles:

- **Reliable**: Consistent structure, numeric values where expected
- **Original**: Shared by Ulrik T. Pedersen but based on real-world data
- **Comprehensive**: Covers a wide range of products and customers
- **Current**: Although it's from 2010–2011, it's appropriate for practice
- **Cited**: Properly referenced on Kaggle with clear authorship

4. How are you addressing licensing, privacy, security, and accessibility?

The dataset is shared under Kaggle’s public data license. It does not contain personally identifiable information, and there are no privacy concerns. The data is stored in a secure folder on my local drive and backed up in Google Drive.

5. How did you verify the data’s integrity?

I reviewed the file for missing values, data types, and obvious errors (e.g., negative quantities, missing customer IDs). I also scanned for duplicate invoice entries and strange date formats. These issues will be addressed in the next phase (“Process”).

6. How does it help you answer your question?

The data allows me to analyze total revenue, product performance, customer segments, and sales trends over time and geography — all key to addressing the business question: “How can the company improve sales performance and retention?”

7. Are there any problems with the data?

Yes, the dataset contains some challenges:

- Missing customer IDs in some rows
- Negative quantities (likely returns)
- Some product descriptions are vague or duplicated
  These will be handled during the data cleaning phase.

Process

In SQL and R.

- **In SQL:**

Data cleaning

- Duplicate the table to clean it up.

```
CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean`  AS
SELECT *
FROM `enduring-wharf-464607-c7.online_retail.sales_raw`
```

- Remove null values

```
CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean_step1`
AS
SELECT *
FROM `enduring-wharf-464607-c7.online_retail.sales_clean`
WHERE CustomerID IS NOT NULL
AND InvoiceNo IS NOT NULL
AND Description IS NOT NULL;
```

- Remove negative or incorrect values. Table 2 is created to add a transaction type column based on whether the negative quantity is a return or a sale, and to remove values where the quantity is 0 and the price unit is negative.

```
 CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean_step2` AS
 SELECT *,
  CASE
   WHEN Quantity <0 THEN 'Return'
   ELSE 'Sale'
   END AS Transaction_type

FROM `enduring-wharf-464607-c7.online_retail.sales_clean_step1`
WHERE
Quantity != 0
AND UnitPrice >0;
```

- Remove duplicates and convert columns to strings to use the function.

```
CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean_step3`
AS
SELECT *
  EXCEPT (row_num)

FROM (
   SELECT *,
    ROW_NUMBER() OVER (PARTITION BY
      InvoiceNo,
      StockCode,
      Description,
       CAST(ROUND(Quantity, 0) AS INT64),
      InvoiceDate,

     CAST(CustomerID AS STRING),
      Country
    ) AS row_num
   FROM `enduring-wharf-464607-c7.online_retail.sales_clean_step2`)
WHERE row_num = 1;
```

- Clean up text. Capital letters and spaces.

```
CREATE TABLE `enduring-wharf-464607-
c7.online_retail.sales_clean_step4_clean_text` AS
SELECT
 InvoiceNo,
  StockCode,
  TRIM(UPPER(Description)) AS Description,
  Quantity,
  InvoiceDate,
  UnitPrice,
  CustomerID,
  TRIM(UPPER(Country)) AS Country,
  Transaction_type
FROM `enduring-wharf-464607-c7.online_retail.sales_clean_step3`
```

- Convert Customerid because it was in another format, with CAST.

```
CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean_step5` AS
SELECT *, CAST(CustomerID AS INT64) AS customerID_
FROM `enduring-wharf-464607-c7.online_retail.sales_clean_step4_clean_text`
```

- Create new columns for analysis

```
CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean_step7`
AS
SELECT *,
 Quantity * UnitPrice AS Total,
 EXTRACT(YEAR FROM InvoiceDate) AS Year,
 EXTRACT(MONTH FROM InvoiceDate) AS Month,
 EXTRACT(DAYOFWEEK FROM InvoiceDate) AS Day_of_week,
 EXTRACT(HOUR FROM InvoiceDate) AS Hour
 FROM `enduring-wharf-464607-c7.online_retail.sales_clean_step6`
```

- Rename table: `sales_clean_final`

```
CREATE TABLE `enduring-wharf-464607-c7.online_retail.sales_clean_final`
AS
SELECT *
FROM `enduring-wharf-464607-c7.online_retail.sales_clean_step7`
```

- **In R:**

- The numeric value of the months has been changed to text

```
df <- df %>%
+    mutate(Month_name = month.abb[Month])
```

- The numeric value of the hours has been changed to text

```
df <- df %>%
  mutate(Day_of_week_name = case_when(
year_frequency <- df %>%
  filter(Transaction_type == "Sale") %>%
  group_by(Year, Month_name) %>%
  summarise(monthly_sales = sum(Total), .groups = "drop")
    Day_of_week == 1 ~ "Sunday",
    Day_of_week == 2 ~ "Monday",
    Day_of_week == 3 ~ "Tuesday",
    Day_of_week == 4 ~ "Wednesday",
    Day_of_week == 5 ~ "Thursday",
    Day_of_week == 6 ~ "Friday",
    Day_of_week == 7 ~ "Saturday",
    TRUE ~ NA_character_
  ))
```

- Create a time slot
- df <- df %>%

```
  mutate(timeslot = case_when(
    Hour >= 6 & Hour < 12  ~ "Morning",
    Hour >= 12 & Hour < 18 ~ "Evening",
    Hour >= 18 & Hour < 24 ~ "Night",
    TRUE                   ~ "Early morning"
  ))

percentage_timeslot <- df %>%
  group_by(timeslot) %>%
  summarise(ventas = sum(Total, na.rm = TRUE)) %>%
  mutate(porcentaje = (ventas / sum(ventas)) * 100)
```

Analyse

In R

- General dataset exploration

```
str(df)
summary(df)
skimr::skim(df)   # skimr para resumen más complete
```

- Analysis: *Top 10 products by quantity sold (excluding returns)*

- **Formula**: top10_products <- df %>%
group_by(Description) %>%
summarise(Total_Quantity = sum(Quantity)) %>%
arrange(desc(Total_Quantity)) %>%
head(10)

-**Finding**: They are unrelated. They could be grouped together to see what percentage of the total they represent. It's not worthwhile now.

The most returned product by far is the Rotating Silver Angels T-light HLDR. 1,475 units were returned. 24 of the third product were returned. The second product is a discount. This product should be analyzed.
A total of 27,250 products were returned.

-Analysis: *Top Customers*

-**Formula**:

```
top_10customers <- df %>%
+  filter(Transaction_type == "Sale") %>%
+  group_by(customerID_) %>%
+  summarise(revenue = sum(Total), .groups = "drop") %>%
+  arrange(desc(revenue)) %>%
+  head(10)
```

```
sum_sales_topcustomers <- sum(top_10customers$revenue)
sum_sales_total <- sum(df$Total)
```

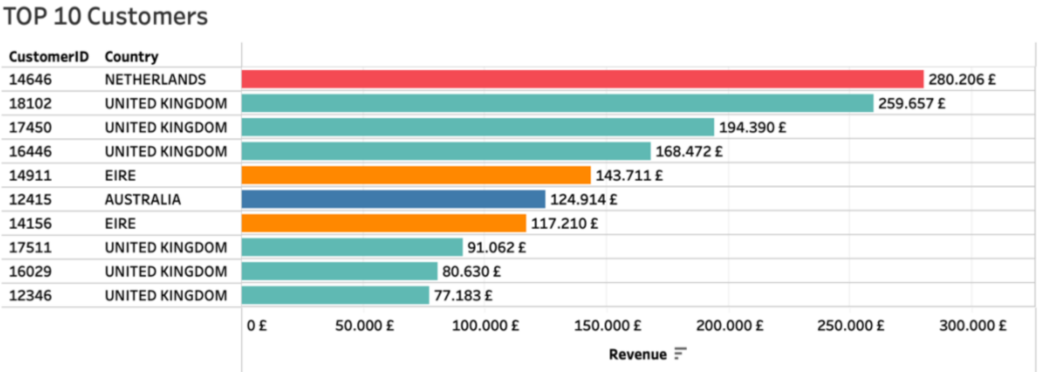
```
percentage_top_customers <- (sum_sales_topcustomers/ sum_sales_total) * 100
```

Datafram= top10.customerscountry

```
top_10_info <- top_10customers %>%
  inner_join(df %>% select(customerID_, Country) %>% distinct(), by = "customerID_")
```

-**Findings**:

The top 10 customers account for **18.57% of total sales**, which suggests a significant concentration of revenue among a small group of clients.
While most of these top customers are from the **UK (6 out of 10)**, the **highest-spending customer is from the Netherlands**, responsible for nearly **90% of all Dutch sales**—with only **9 customers total** in that country. This highlights the **disproportionate impact** a single international client can have.
Ireland has only **3 customers**, yet **2 are in the top 10**, and **Australia** appears in **6th place**, despite its geographical distance.



-Analysis: *Which countries generate the most money?*

-Formulas:

```
top_countries <- df %>%
  filter(Transaction_type == "Sale") %>%
  group_by(Country) %>%
  summarise(total_revenue = sum(Total), .groups = "drop") %>%
  arrange(desc(total_revenue))
```

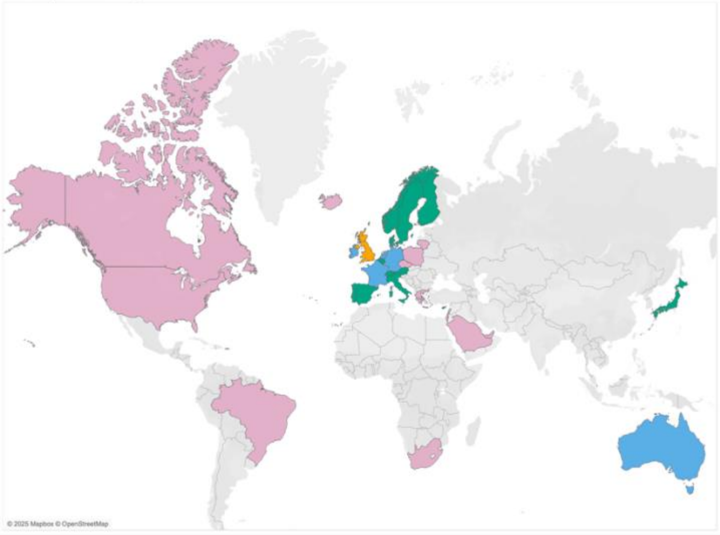
dataframe = top10_info

dataframe= customers_per_country; top_countries

- **Findings**:

The **United Kingdom** not only has the **highest revenue** but also the **largest number of clients**, which is expected for a domestic market.
However, countries like **Netherlands (9 clients)** and **Ireland (3 clients)** stand out: despite very few customers, they are among the **top 3 countries in sales**, showing **exceptionally high spending per client**.
On the other hand, **Germany, France, and Spain** have a larger number of customers but contribute **less revenue overall**. This suggests a **lower customer value** compared to smaller, more lucrative markets.

Sales per Country



- **Analysis:** *Months with the most sales*

-**Formula**:

```
sales_per_day <- df %>%
  filter(Transaction_type == "Sale") %>%
  group_by(Day_of_week_name) %>%
  summarise(total_sales = sum(Total), .groups = "drop") %>%
  arrange(desc(total_sales))
```

Fórmula para saber que representa del total las vents de un día:

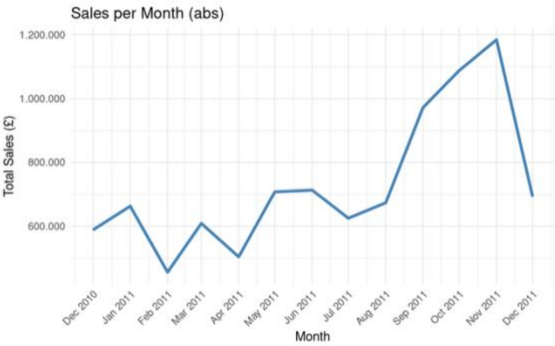
```
> sum(df$Total)
[1] 8280356

thursday <- (1971822/ 8280356) * 100

> tuesday  <- (1697055.6/ 8280356) * 100
```

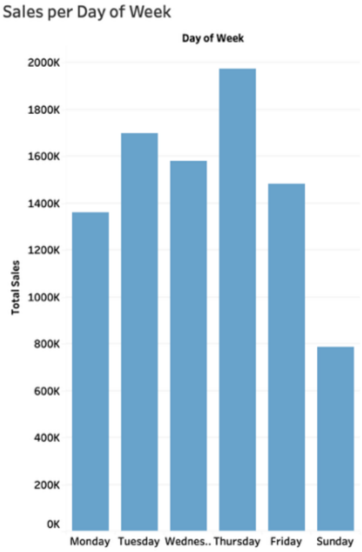
Dataframes: sales_per_day

Datafrma: sales_per_hour, percentage time

-**Findings**:

No sales are recorded on Saturdays, likely due to store closure or missing data. **Thursdays and Tuesdays** show the highest sales volume, while **Sundays and Mondays** perform worst.
Time-of-day analysis shows a **morning peak (10h–13h)** with nearly 60% of daily purchases.
Evenings (15h–18h) see another spike, while **night sales (post-18h)** are negligible (just 1.9%).
Across months, **September and October** show a strong increase in sales, with a notable **decline in December**, despite including holiday season.
February and April are consistently low in sales across both years. December 2010 likely underrepresented due to limited data.





-Analysis: *What percentage of transactions are returns? And Return Country*

-Formulas:

```
total_sales <- nrow(filter(df, Transaction_type == "Sale"))
total_returns <- nrow(filter(df, Transaction_type == "Return"))
return_rate <- total_returns / (total_sales + total_returns)
```

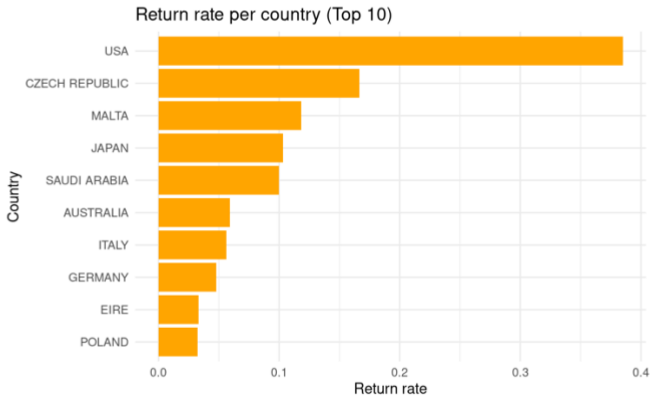
```
percentage_return <- (340 / sum(df$Transaction_type == "Return")) * 100
> View(percentage_return)
```

```
percentage_return_country <- (7476/ sum(df$Transaction_type == "Return" )) * 100
```

```
return_rate_country <- df %>%
  group_by(Country) %>%
  summarise(
    Sales = sum(Transaction_type == "Sale"),
    Return = sum(Transaction_type == "Return"),
    Return_Rate = Return / (Sales + Return),
    .groups = "drop"
  ) %>%
  arrange(desc(Return_Rate))

return_rate_product <- df %>%
  group_by(Description) %>%
  summarise(
    Sales = sum(Transaction_type == "Sale"),
    Return = sum(Transaction_type == "Return"),
    Return_Rate = Return / (Sales + Return),
    .groups = "drop"
  ) %>%
  arrange(desc(Return_Rate))
```

-**Findings**:

The overall return rate across all transactions is **2.2%**, calculated from sales and return entries.
19 products were returned but show no recorded sales in the dataset. This could be due to sales made **before December 2010** or a **data inconsistency** in the source.
The two most returned products are **"REGENCY CAKESTAND 3 TIER"** and **"MANUAL"**, which together make up **3.4% of all returns**.
The **United States** stands out sharply, representing **84.6% of total returns**, with **112 returns out of 179 total transactions** — an extremely high **return rate of 38.5%**.
The second-highest country in return volume is **Germany** with 452 returns (5.1%), followed by much lower values for other countries.
Countries like **Netherlands** appear low in returns (only 8) but rank high in revenue — an important insight for retention and customer quality.
The highest return rate is in the USA, 0.385. Sales are 179 and returns are 112. The difference in the return rate is more than double that of the latter. Something is happening in the US.



-**Analysis**: *Customer segmentation*

-Formulas:

```
# Calcular RFM por cliente
rfm <- df %>%
  filter(Transaction_type == "Sale") %>%
  group_by(customerID_) %>%
  summarise(
    Recency = as.numeric(difftime(analysis_date, max(InvoiceDate), units = "days")),
    Frequency = n_distinct(InvoiceNo),
    Monetary = sum(Total),
    .groups = "drop"
  )

# Para segmentar, puedes crear cuartiles o quintiles, por ejemplo:
rfm <- rfm %>%
  mutate(
    R_Score = ntile(-Recency, 4),   # Más reciente es mejor, por eso negativo
    F_Score = ntile(Frequency, 4),
    M_Score = ntile(Monetary, 4),
    RFM_Score = R_Score + F_Score + M_Score
  )

# Clientes con RFM_Score alto son los más valiosos
top_customers <- rfm %>% arrange(desc(RFM_Score))


top_rfm_count <- sum(rfm$RFM_Score == 12)
total_clients <- nrow(rfm)
percentage_top_rfm <- (top_rfm_count / total_clients) * 100

valuable_customers <- rfm_info %>%
  filter(RFM_Score == 12) %>%
  group_by(Country) %>%
  summarise(Valuable_customers = n()) %>%
  arrange(desc(Valuable_customers))


rfm_info %>%
  mutate(Grupo = ifelse(RFM_Score == 12, "Top", "Resto")) %>%
  group_by(Grupo) %>%
  summarise(Revenue = sum(Monetary))

top_rfm_count <- sum(rfm$RFM_Score == 12)
total_clients <- nrow(rfm)
percentage_top_rfm <- (top_rfm_count / total_clients) * 100

valuable_customers <- rfm_info %>%
  filter(RFM_Score == 12) %>%
  group_by(Country) %>%
  summarise(Valuable_customers = n()) %>%
  arrange(desc(Valuable_customers))

rfm_info %>%
  mutate(Group = ifelse(RFM_Score == 12, "Top", "Others")) %>%
  group_by(Group) %>%
  summarise(Revenue = sum(Monetary))

rfm_info_category <- rfm_info %>%
  mutate(Segmento = case_when(
    RFM_Score >= 10 ~ "Excellent Customers",
    RFM_Score >= 7 ~ "Loyal Customers",
    RFM_Score >= 4 ~ "Risk Customers",
    TRUE ~ "Lost Customers"
```

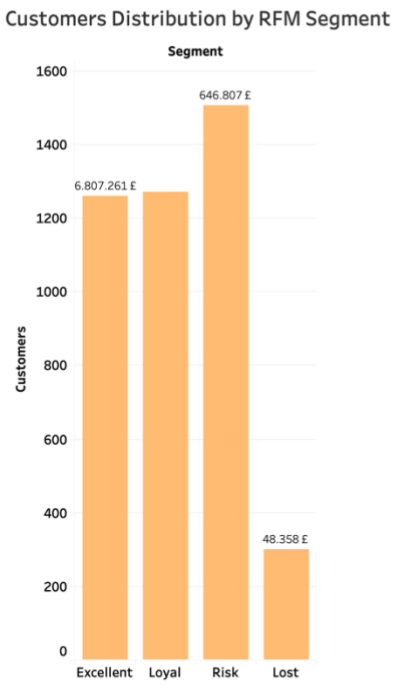
-**Findings**:

Customers were segmented using the **RFM model** (Recency, Frequency, Monetary).
The RFM score of **12/12** identifies the **most valuable clients**, representing
only **10.9%** of the customer base.
Despite their small size, these top customers generate nearly **half of the total revenue (49.3%)** — **£4.38M out of £8.87M** overall.
Segment breakdown:
Excellent: 1,260 clients, £6.8M
Loyal: 1,271 clients, £1.37M
At Risk: 1,507 clients, £646K

Lost: 300 clients, £48K

A **geo-distribution map** reveals that top RFM customers are highly concentrated in the **UK**, followed by **Germany**and **France**.

This insight supports the idea of focusing retention strategies on **high-value clients** and expanding in regions with similar profiles.



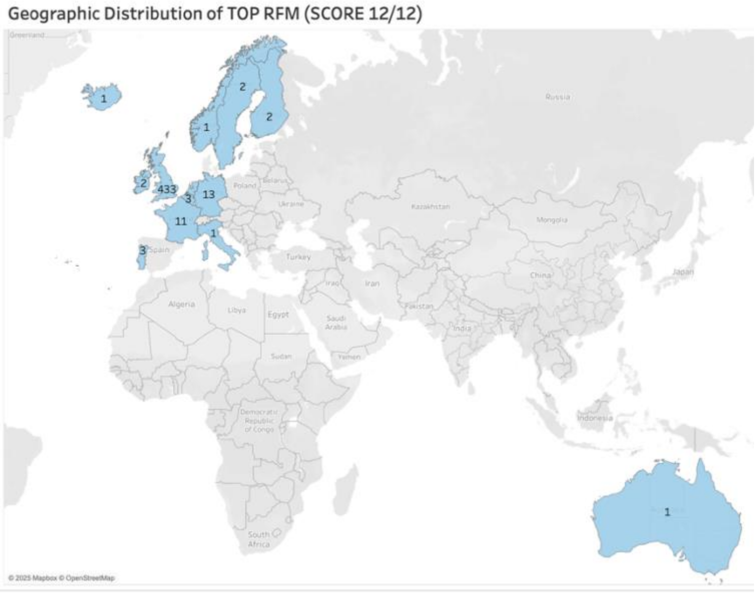
- **Analysis:** *Segmentation by country*

-Formula:

```
customers_per_country <- df %>%
  filter(Transaction_type == "Sale") %>%
  group_by(Country) %>%
  summarise(
    Total_Customers = n_distinct(customerID_),
    Total_Sales = sum(Total),
    .groups = "drop"
  ) %>%
  arrange(desc(Total_Sales))
```

-Findings:

-3,920 customers in the United Kingdom, and no foreign country has more than 100 customers. However, the Netherlands has the largest customer base, and Ireland only has three customers, and they are the top three sellers.



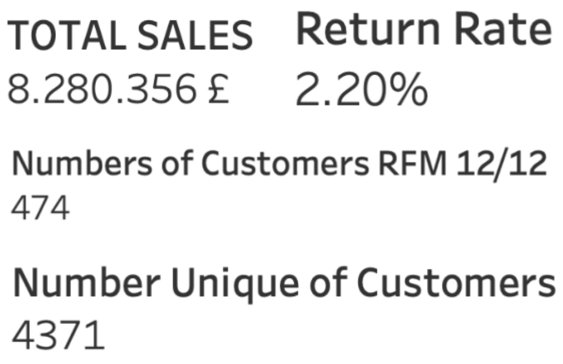
Geographic Distribution of TOP RFM (SCORE 12/12)

R graphics

```
df %>%
  group_by(Year, Month) %>%
  summarise(monthly_sales = sum(Total), .groups = "drop") %>%
  ggplot(aes(x = interaction(Year, Month, sep = "-"), y = monthly_sales)) +
  geom_line(group = 1, color = "steelblue") +
  labs(title = "Sales per Month",
       x = "Month",
       y = "Total Sales (£)") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

- **Key Performance Indicators (KPIs)**

The company generated **over £8.28M in total sales**, with **4,371 unique customers** during the analyzed period.
Out of these, **474 customers** scored the highest possible **RFM value (12/12)** — signaling key contributors to revenue.
The overall **return rate** was **2.2%**, with returns primarily concentrated in a few specific countries and products.
These metrics offer a quick, high-level overview of business performance and will be referenced throughout the analysis.



TOTAL SALES
8.280.356 £

Return Rate
2.20%

Numbers of Customers RFM 12/12
474

Number Unique of Customers
4371

Share

To communicate the findings effectively, a combination of data visualization and presentation tools was used:

- **R** was used to generate specific visualizations such as the return rate distribution and monthly sales trends.
- **Tableau** was the main tool for creating a wide variety of interactive and static visualizations, including:
  - Bar charts (vertical and horizontal)
  - Line and area charts
  - Geographic maps
  - Customer segmentation by RFM score
  - A comprehensive interactive **dashboard** summarizing KPIs and key analysis views
- **PowerPoint** was used to develop a presentation that organizes the business context, data preparation steps, analysis, and recommendations in a clear and visual format.

These tools enabled effective storytelling through data, allowing stakeholders and viewers to understand business patterns, performance insights, and potential improvement areas.

- *Sales per Country and Customer Unique Interactive Tableau Map:* https://public.tableau.com/app/profile/axel.aranda/viz/CustomersperCountry_17542072256270/Hoja2?publish=yes

- *Dashboard Sales Ecommerce:* https://public.tableau.com/app/profile/axel.aranda/viz/Dashboardecommerce

_17541573836780/Dashboard1

- *Power Point Presentation:* https://drive.google.com/drive/folders/1RRKErDu1KDoejitSMPpvf4U7Ey6VWAds

Recommendations & Next Steps

Propose targeted campaigns to retain the high-value RFM 12/12 customer segment. These customers generate nearly half of the company’s revenue and should be prioritized for loyalty initiatives.

Investigate the causes behind the unusually high return rates in certain countries, especially the U.S. Understanding the reasons behind these returns can help reduce lost revenue and improve customer satisfaction.

Use time-based sales insights to launch marketing campaigns during peak hours, particularly from 10:00 to 13:00 and again in the afternoon.

Capitalize on the profitability of customers from countries with a small but highly valuable customer base, such as the Netherlands or Ireland. These markets show strong revenue per customer and could be worth further development.

Conclusions

In summary, our analysis identified that a relatively small group of top customers drives a large portion of sales, especially concentrated in the UK but with important contributions from other countries like the Netherlands and Ireland.

We see clear temporal sales patterns, with Thursdays and Tuesdays being the strongest days and mornings accounting for the majority of transactions.

Returns are generally low, but the USA stands out with a notably high return rate, which warrants further investigation.

The RFM segmentation confirms that focusing on the top 10-11% of customers can capture almost half of total revenue, reinforcing the value of customer loyalty programs.

Finally, the geographic insights open possibilities for tailored campaigns targeting key regions to further increase sales and customer retention.