# ANALYSIS OF ONLINE RETAIL TRANSACTIONS

Emmanuel Attah Ackah

2026-01-07

## Analysis of Online Retail Transactions

**Author:** Emmanuel Attah Ackah
**Tools:** R, dplyr, ggplot2, tidyverse, lubridate
**Output:** PDF Report

### Project Overview

This project analyzes online retail transaction data to understand revenue trends, customer purchasing behavior, product performance, and geographic sales distribution. The goal is to generate actionable insights that support data-driven business and operational decisions.

### Business Questions

- How does revenue change over time?
- Which products generate the highest revenue?
- Are there identifiable customer segments based on spending behavior?
- Which countries contribute the most to overall revenue?
- Are there seasonal patterns in purchasing activity?

### Data Cleaning & Preparation

- Removed canceled transactions
- Excluded records with missing customer IDs
- Filtered out invalid quantities and prices
- Engineered revenue and time-based features (year, month)

### Key Analyses

- Monthly revenue and order trends
- Top products by revenue and quantity sold
- Customer segmentation based on spending behavior
- Country-level revenue contributions over time

## Key Insights

- Revenue exhibits strong seasonality, with noticeable peaks during specific months.
- A small group of high-value customers contributes a disproportionate share of total revenue.
- A limited number of products drive most sales, highlighting opportunities for inventory optimization.
- Revenue is concentrated in a few countries, suggesting geographic growth opportunities.

## Recommendations

- Prioritize retention strategies for high-value customers.
- Optimize inventory for top-performing products ahead of peak seasons.
- Expand targeted marketing in high-revenue countries.
- Develop loyalty programs to increase repeat purchases among mid-tier customers.

## Limitations & Next Steps

- Lack of customer demographic data limits deeper behavioral analysis.
- Future work could include cohort analysis and customer lifetime value modeling.
- Predictive models could be developed for revenue forecasting and churn prediction.

## Files

- `analysis.Rmd` – Full analysis workflow
- `analysis.pdf` – Final report output
- `data/` – Raw dataset

```r
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.1
## v readr     2.1.5
```

```
## -- Conflicts --------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
#install.packages("readxl")
library(readxl)
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
## Data Import and Overview
retail <- read_xlsx("//Users/ataka/Desktop/ONLINE_RETAIL_ANALYSIS/DATA/online_retail_II.xlsx")
head(retail)
```

```
## # A tibble: 6 x 8
##   Invoice StockCode Description Quantity InvoiceDate         Price `Customer ID`
##   <chr>   <chr>     <chr>          <dbl> <dttm>              <dbl>         <dbl>
## 1 489434  85048     "15CM CHRI~       12 2009-12-01 07:45:00  6.95         13085
## 2 489434  79323P    "PINK CHER~       12 2009-12-01 07:45:00  6.75         13085
## 3 489434  79323W    "WHITE CHE~       12 2009-12-01 07:45:00  6.75         13085
## 4 489434  22041     "RECORD FR~       48 2009-12-01 07:45:00  2.1          13085
## 5 489434  21232     "STRAWBERR~       24 2009-12-01 07:45:00  1.25         13085
## 6 489434  22064     "PINK DOUG~       24 2009-12-01 07:45:00  1.65         13085
## # i 1 more variable: Country <chr>
```

```
str(retail)
```

```
## tibble [525,461 x 8] (S3: tbl_df/tbl/data.frame)
##  $ Invoice    : chr [1:525461] "489434" "489434" "489434" "489434" ...
##  $ StockCode  : chr [1:525461] "85048" "79323P" "79323W" "22041" ...
##  $ Description: chr [1:525461] "15CM CHRISTMAS GLASS BALL 20 LIGHTS" "PINK CHERRY LIGHTS" "WHITE CHE
##  $ Quantity   : num [1:525461] 12 12 12 48 24 24 24 10 12 12 ...
##  $ InvoiceDate: POSIXct[1:525461], format: "2009-12-01 07:45:00" "2009-12-01 07:45:00" ...
##  $ Price      : num [1:525461] 6.95 6.75 6.75 2.1 1.25 1.65 1.25 5.95 2.55 3.75 ...
##  $ Customer ID: num [1:525461] 13085 13085 13085 13085 13085 ...
##  $ Country    : chr [1:525461] "United Kingdom" "United Kingdom" "United Kingdom" "United Kingdom" .
```

```
summary(retail)
```

```
##    Invoice            StockCode         Description           Quantity
##  Length:525461      Length:525461      Length:525461      Min.   :-9600.00
##  Class :character   Class :character   Class :character   1st Qu.:    1.00
##  Mode  :character   Mode  :character   Mode  :character   Median :    3.00
##                                                           Mean   :   10.34
##                                                           3rd Qu.:   10.00
```

```
##                                                        Max.   :19152.00
##
##   InvoiceDate                         Price            Customer ID
##  Min.   :2009-12-01 07:45:00.00   Min.   :-53594.36   Min.   :12346
##  1st Qu.:2010-03-21 12:20:00.00   1st Qu.:     1.25   1st Qu.:13983
##  Median :2010-07-06 09:51:00.00   Median :     2.10   Median :15311
##  Mean   :2010-06-28 11:37:36.84   Mean   :     4.69   Mean   :15361
##  3rd Qu.:2010-10-15 12:45:00.00   3rd Qu.:     4.21   3rd Qu.:16799
##  Max.   :2010-12-09 20:01:00.00   Max.   : 25111.09   Max.   :18287
##                                                       NA's   :107927
##    Country
##  Length:525461
##  Class :character
##  Mode  :character
##
##
##
##
```

```
colnames(retail)
```

```
## [1] "Invoice"     "StockCode"   "Description" "Quantity"    "InvoiceDate"
## [6] "Price"       "Customer ID" "Country"
```

```
## Data Cleaning and Feature Engineering
retail <- retail %>%
  clean_names()
#Used this to determine the total number of individual customers served
retail %>%
  summarise(
    total_customers = n_distinct(customer_id, na.rm = TRUE)
    )
```

```
## # A tibble: 1 x 1
##   total_customers
##             <int>
## 1            4383
```

```
#Used this step to determine the total number of countries the transactions were received from
retail %>%
  summarise(
    unique_countries = n_distinct(country, na.rm = TRUE)
    )
```

```
## # A tibble: 1 x 1
##   unique_countries
##              <int>
## 1               40
```

```
## Revenue and Order Trends
#Used this step to determine the total revenue generated by all the transactions recorded
retail %>%
```

```
  summarise(
    total_revenue = sum(quantity * price, na.rm=TRUE)
  )
```

```
## # A tibble: 1 x 1
##   total_revenue
##           <dbl>
## 1      9539485.
```

```
#Wanted to determine how many orders were placed
retail %>%
  summarise(total_orders = n_distinct(invoice))
```

```
## # A tibble: 1 x 1
##   total_orders
##          <int>
## 1        28816
```

```
#Total revenue generated per individual country
retail %>%
  group_by(country) %>%
  summarise(revenue = sum(quantity * price, na.rm = TRUE)) %>%
  arrange(desc(revenue))
```

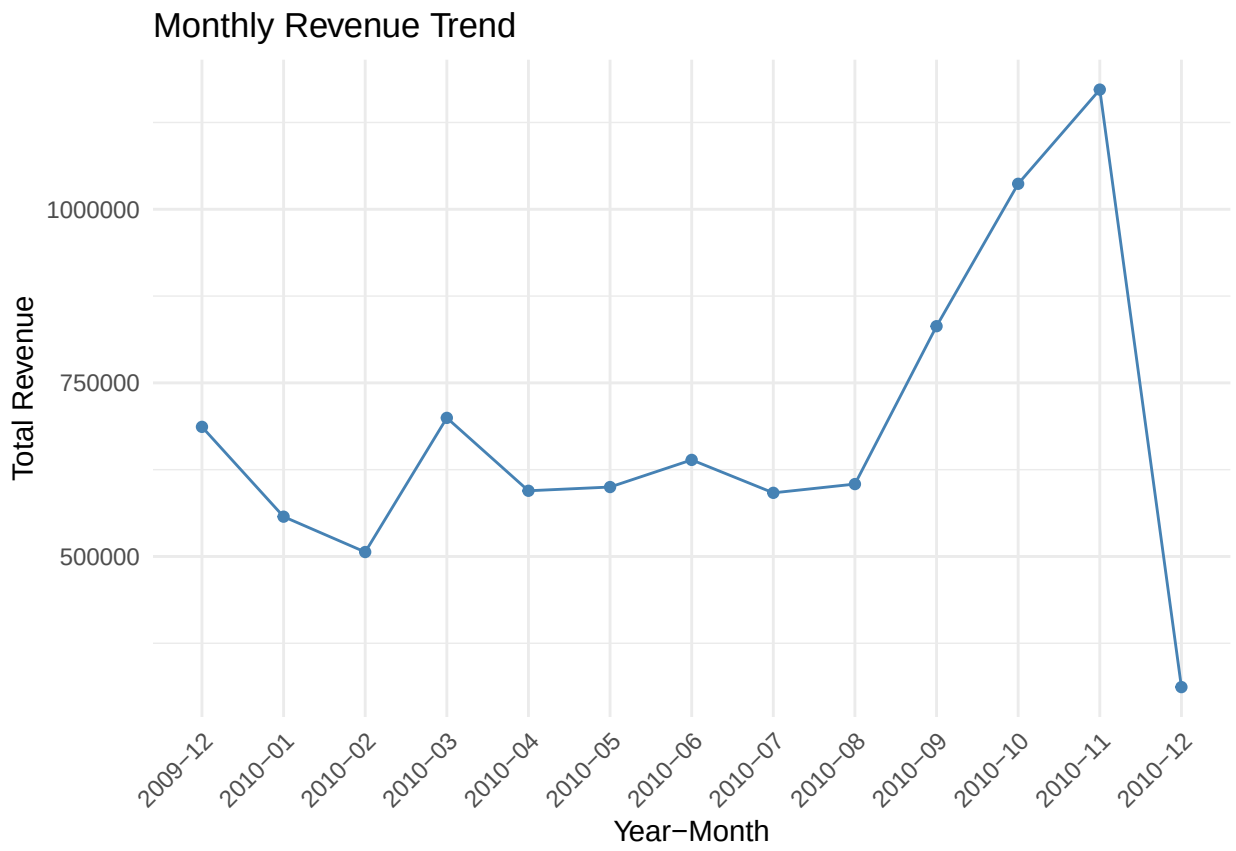
```
## # A tibble: 40 x 2
##    country         revenue
##    <chr>             <dbl>
##  1 United Kingdom 8194778.
##  2 EIRE            352243.
##  3 Netherlands     263863.
##  4 Germany         196290.
##  5 France          130770.
##  6 Sweden           51214.
##  7 Denmark          46973.
##  8 Switzerland      43343.
##  9 Spain            37085.
## 10 Australia        30052.
## # i 30 more rows
```

```
## Product Performance Analysis
# Remove cancelled transactions, missing customers, and invalid values
retail_clean <- retail %>%
  filter(
    !is.na(customer_id),
    quantity > 0,
    price > 0,
    !grepl("^C", invoice)
  )

# Create revenue variable
retail_clean <- retail_clean %>%
```

```r
  mutate(
    revenue = quantity * price
  )

retail_clean <- retail_clean %>%
  mutate(
    invoice_date = as.POSIXct(invoice_date),
    year = lubridate::year(invoice_date),
    month = lubridate::month(invoice_date, label = TRUE),
    year_month = format(invoice_date, "%Y-%m")
  )

monthly_revenue <- retail_clean %>%
  group_by(year_month) %>%
  summarise(
    total_revenue = sum(revenue),
    total_orders = n_distinct(invoice)
  ) %>%
  arrange(year_month)

monthly_revenue
```

```
## # A tibble: 13 x 3
##    year_month total_revenue total_orders
##    <chr>              <dbl>        <int>
##  1 2009-12          686654.         1512
##  2 2010-01          557319.         1011
##  3 2010-02          506371.         1104
##  4 2010-03          699609.         1524
##  5 2010-04          594609.         1329
##  6 2010-05          599986.         1377
##  7 2010-06          639067.         1497
##  8 2010-07          591637.         1381
##  9 2010-08          604243.         1293
## 10 2010-09          831615.         1689
## 11 2010-10         1036680.         2133
## 12 2010-11         1172336.         2587
## 13 2010-12          311878.          776
```

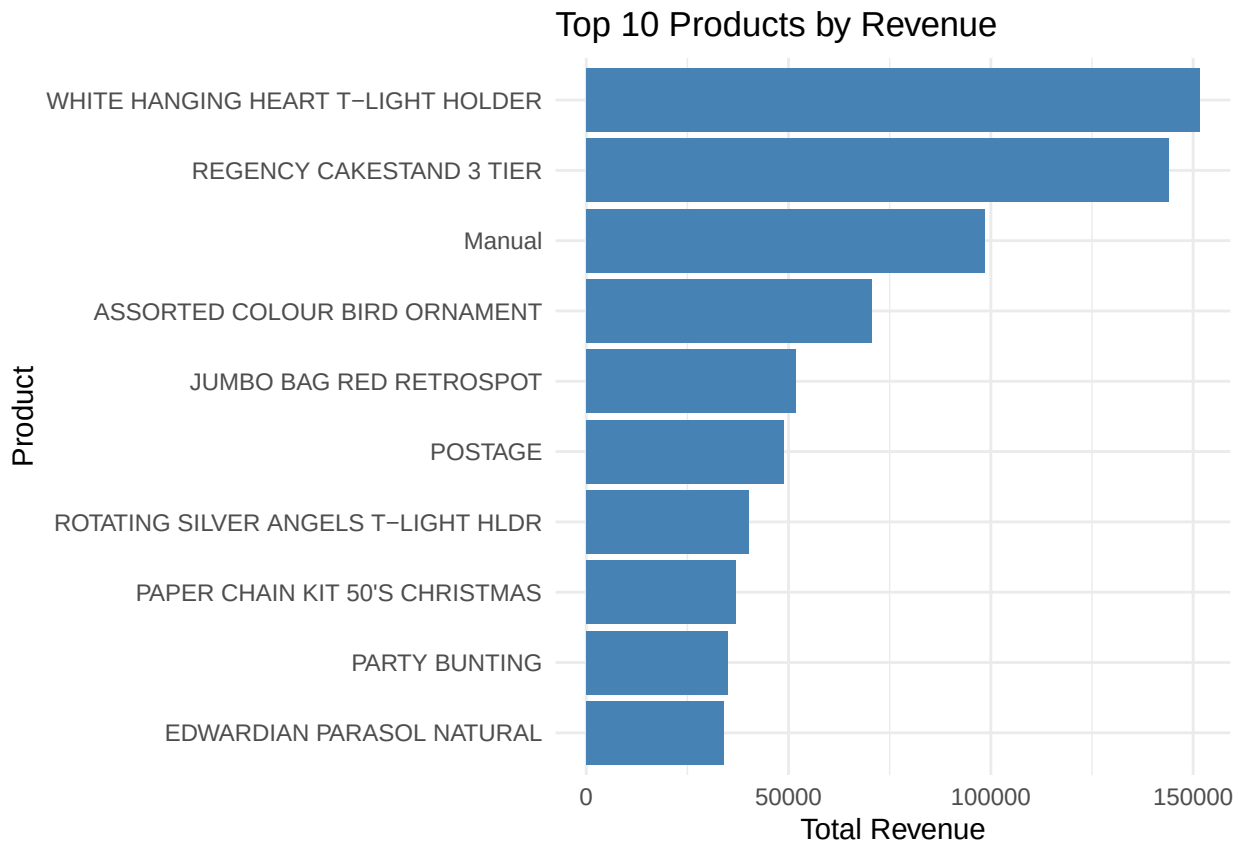
```r
ggplot(monthly_revenue, aes(x = year_month, y = total_revenue)) +
  geom_line(group = 1, color = "steelblue") +
  geom_point(color = "steelblue") +
  labs(
    title = "Monthly Revenue Trend",
    x = "Year-Month",
    y = "Total Revenue"
  ) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
top_products <- retail_clean %>%
  group_by(description) %>%
  summarise(
    total_revenue = sum(revenue),
    total_quantity = sum(quantity)
  ) %>%
  arrange(desc(total_revenue)) %>%
  slice_head(n = 10)

top_products
```

```
## # A tibble: 10 x 3
##    description                         total_revenue total_quantity
##    <chr>                                       <dbl>          <dbl>
##  1 WHITE HANGING HEART T-LIGHT HOLDER        151624.          56915
##  2 REGENCY CAKESTAND 3 TIER                  143893.          12497
##  3 Manual                                     98561.           2630
##  4 ASSORTED COLOUR BIRD ORNAMENT              70494.          44551
##  5 JUMBO BAG RED RETROSPOT                    51759.          29578
##  6 POSTAGE                                    48741.           2212
##  7 ROTATING SILVER ANGELS T-LIGHT HLDR        40187.          21591
##  8 PAPER CHAIN KIT 50'S CHRISTMAS             36934.          13860
##  9 PARTY BUNTING                              35036.           8316
## 10 EDWARDIAN PARASOL NATURAL                  34045.           7201
```

```
ggplot(top_products, aes(x = reorder(description, total_revenue), y = total_revenue)) +
  geom_col(fill = "steelblue") +
  coord_flip() +
  labs(
    title = "Top 10 Products by Revenue",
    x = "Product",
    y = "Total Revenue"
  ) +
  theme_minimal()
```

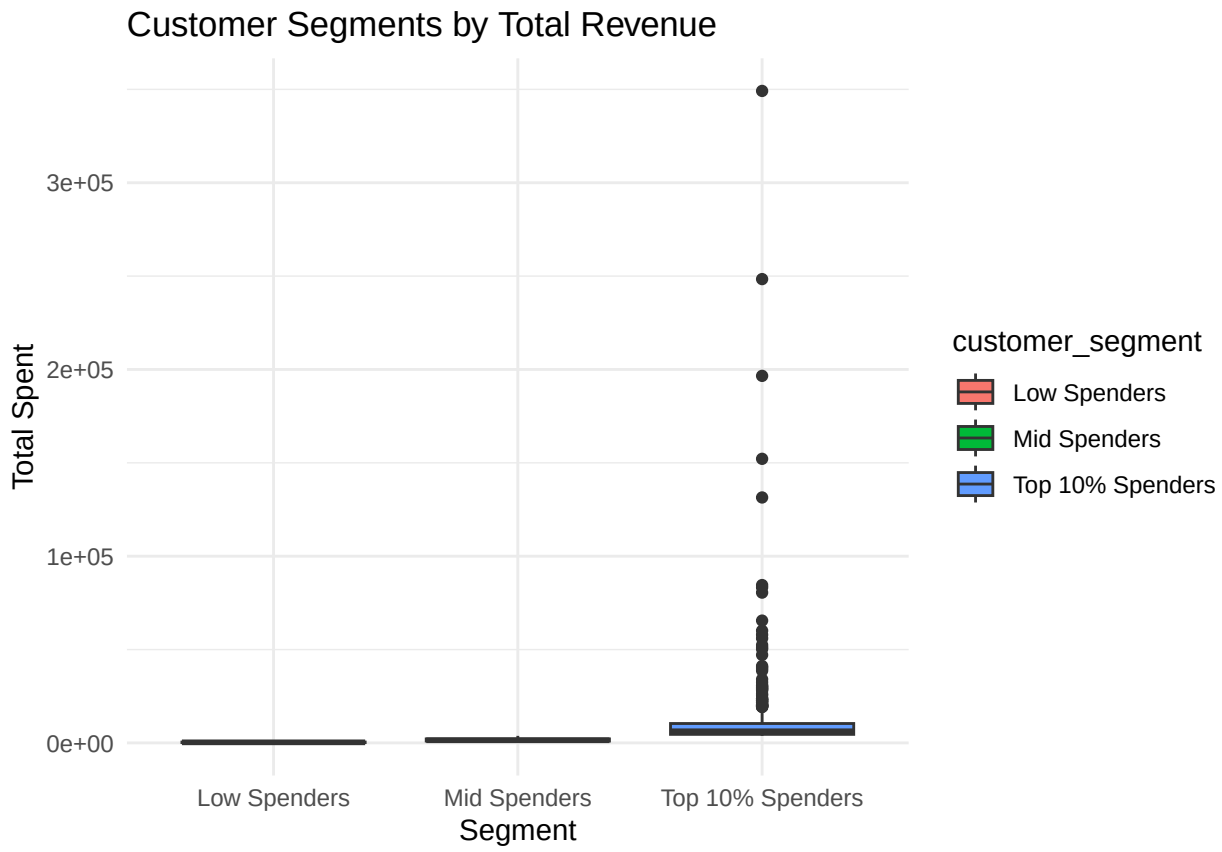



```
## Customer Segmentation
customer_summary <- retail_clean %>%
  group_by(customer_id) %>%
  summarise(
    total_spent = sum(revenue),
    number_of_orders = n_distinct(invoice),
    avg_order_value = total_spent / number_of_orders
  )

summary(customer_summary)
```

```
##   customer_id     total_spent     number_of_orders  avg_order_value
##  Min.   :12346   Min.   :     3   Min.   :  1.000   Min.   :    2.95
##  1st Qu.:13882   1st Qu.:   308   1st Qu.:  1.000   1st Qu.:  182.09
##  Median :15350   Median :   706   Median :  2.000   Median :  287.37
```

```
##  Mean   :15349   Mean   :  2048   Mean   :  4.456   Mean   :  378.32
##  3rd Qu.:16834   3rd Qu.:  1723   3rd Qu.:  5.000   3rd Qu.:  423.58
##  Max.   :18287   Max.   :349164   Max.   :205.000   Max.   :11880.84
```

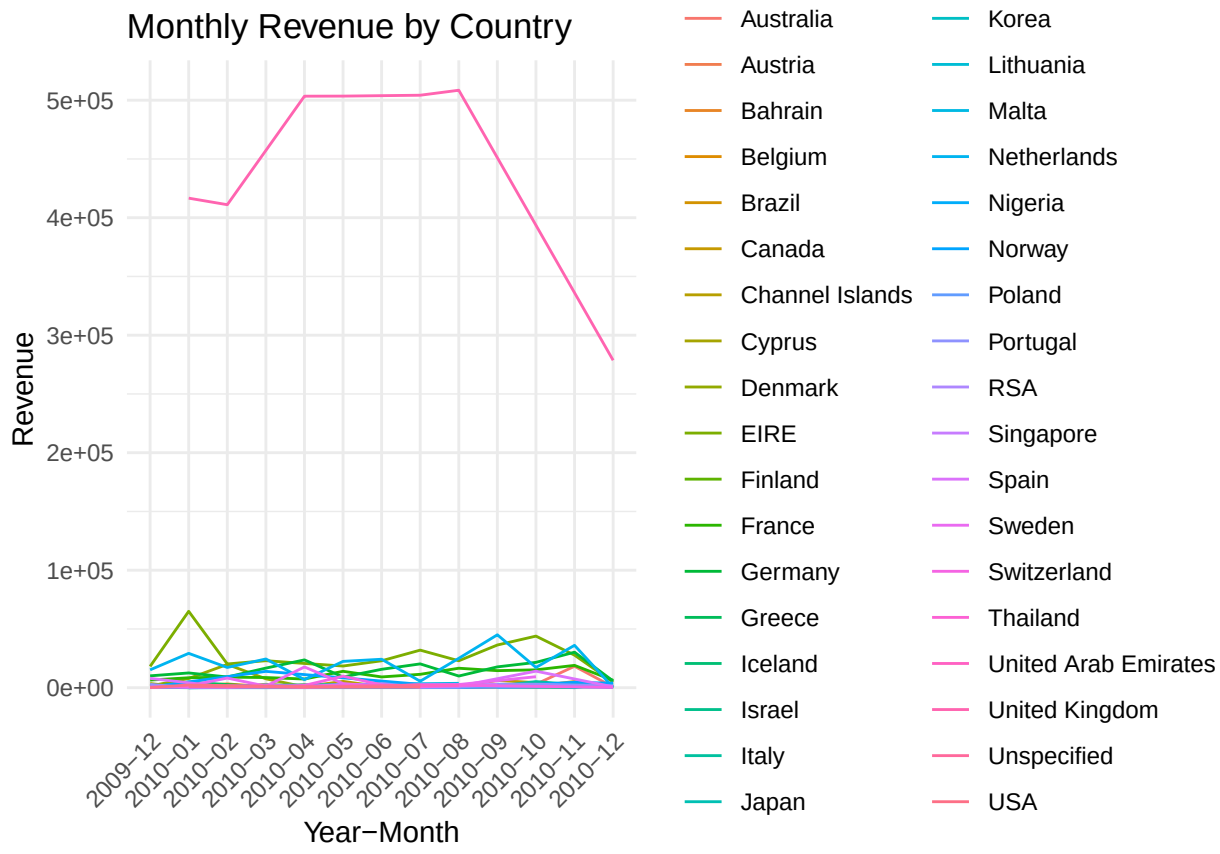
```r
customer_summary <- customer_summary %>%
  mutate(
    customer_segment = case_when(
      total_spent >= quantile(total_spent, 0.9) ~ "Top 10% Spenders",
      total_spent >= quantile(total_spent, 0.5) ~ "Mid Spenders",
      TRUE ~ "Low Spenders"
    )
  )

ggplot(customer_summary, aes(x = customer_segment, y = total_spent, fill = customer_segment)) +
  geom_boxplot() +
  labs(title = "Customer Segments by Total Revenue", x = "Segment", y = "Total Spent") +
  theme_minimal()
```



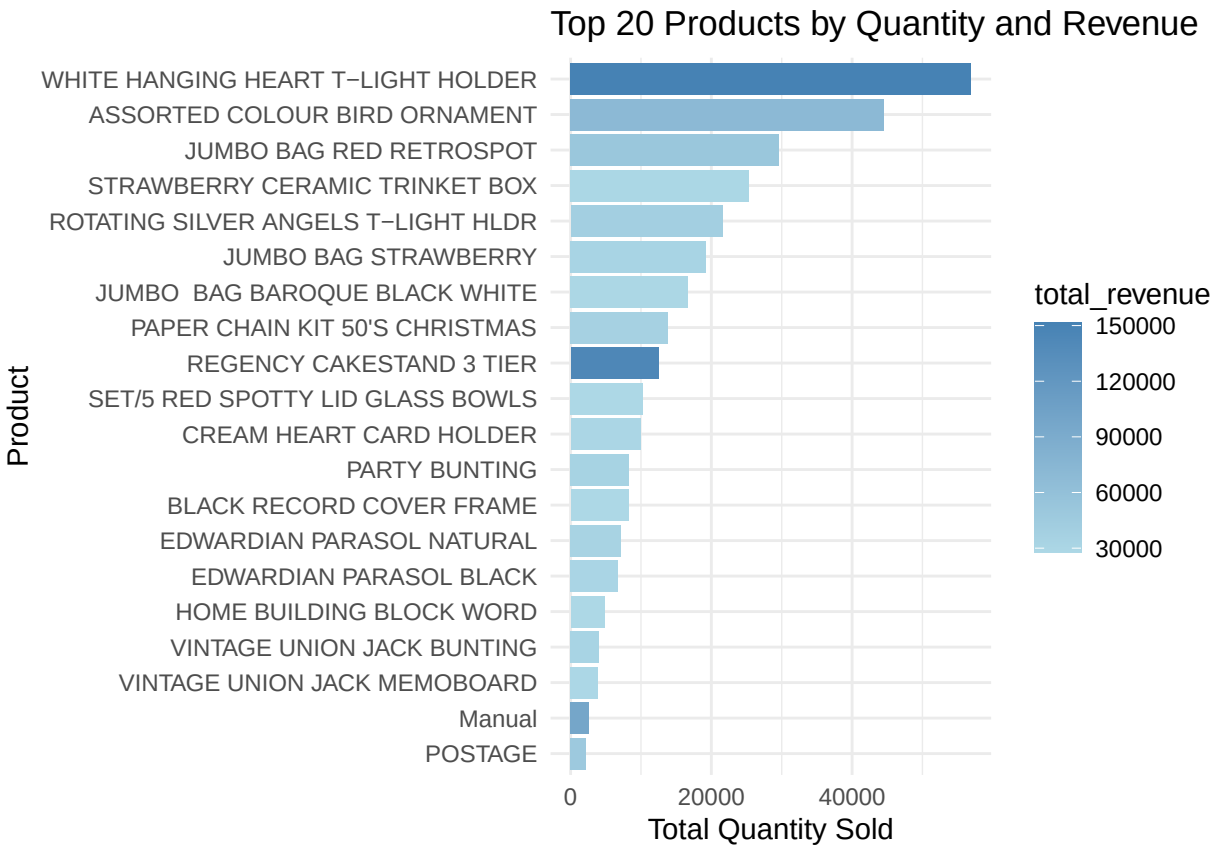

```r
## Geographic Revenue Patterns
#Determine which countries contribute the most revenue over time
monthly_country_revenue <- retail_clean %>%
  group_by(year_month, country) %>%
  summarise(total_revenue = sum(revenue)) %>%
  arrange(year_month) %>%
  slice_head(n=20)
```

```
## `summarise()` has grouped output by 'year_month'. You can override using the
## `.groups` argument.
```

```
ggplot(monthly_country_revenue, aes(x = year_month, y = total_revenue, color = country, group = country)
  geom_line() +
  labs(title = "Monthly Revenue by Country", x = "Year-Month", y = "Revenue") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```




```
#Determine top revenue products
top_products_plot <- retail_clean %>%
  group_by(description) %>%
  summarise(total_revenue = sum(revenue), total_quantity = sum(quantity)) %>%
  arrange(desc(total_revenue)) %>%
  slice_head(n = 20)

ggplot(top_products_plot, aes(x = reorder(description, total_quantity), y = total_quantity, fill = total
  geom_col() +
  coord_flip() +
  scale_fill_gradient(low = "lightblue", high = "steelblue") +
  labs(title = "Top 20 Products by Quantity and Revenue", x = "Product", y = "Total Quantity Sold") +
  theme_minimal()
```


Top 20 Products by Quantity and Revenue
Product
WHITE HANGING HEART T−LIGHT HOLDER
ASSORTED COLOUR BIRD ORNAMENT
JUMBO BAG RED RETROSPOT
STRAWBERRY CERAMIC TRINKET BOX
ROTATING SILVER ANGELS T−LIGHT HLDR
JUMBO BAG STRAWBERRY
JUMBO BAG BAROQUE BLACK WHITE
PAPER CHAIN KIT 50'S CHRISTMAS
REGENCY CAKESTAND 3 TIER
SET/5 RED SPOTTY LID GLASS BOWLS
CREAM HEART CARD HOLDER
PARTY BUNTING
BLACK RECORD COVER FRAME
EDWARDIAN PARASOL NATURAL
EDWARDIAN PARASOL BLACK
HOME BUILDING BLOCK WORD
VINTAGE UNION JACK BUNTING
VINTAGE UNION JACK MEMOBOARD
Manual
POSTAGE
0
20000
40000
Total Quantity Sold
total_revenue
150000
120000
90000
60000
30000