

E-Commerce Data Analysis Report

Michalis Loizos

2025-05-27

I.Introduction

This independent project aims to clean, process, and analyze e-commerce data in order to identify correlations and uncover potential trends through data visualization. It also serves to demonstrate my skills in data analysis and reporting.

II. Data cleaning and considerations

In this report, e-commerce data is analyzed to derive business insights supported by data. The dataset is loaded into BigQuery, inspected, and cleaned prior to analysis. SQL queries are used to retrieve and aggregate relevant information, and R along with Tableau are used for data visualization.

Initially, the dataset was checked for missing values and duplicate entries. These were removed to ensure data quality. Entries with missing customer IDs were retained, as they represented a significant portion of the dataset and likely corresponded to unregistered or guest users.

```
# SELECT DISTINCT *  
# FROM  
# `[Project_Name].E_Commerce_Data.Retail_Data`  
  
# WHERE  
# Description IS NOT NULL  
# AND CustomerID IS NOT NULL  
# AND  
# Quantity > 0
```

It is worth mentioning that entries without customer ID were kept for the analysis, as they represented a significant portion of the dataset, and it is possible they were just not registered in the system.

III.Analysis Results

Monthly Revenue Trends

The analysis begins with a monthly revenue breakdown to identify potential trends. Data from December 2010 to December 2011 was used to calculate monthly revenue.

From the plotted results, it is evident that revenue peaks between September and November 2011, during which revenue nearly doubles. This indicates that the platform could benefit from targeted sales campaigns or special offers during this period to further boost revenue.

```

# SELECT
# EXTRACT(YEAR FROM InvoiceDate AT TIME ZONE 'UTC') AS year,
# EXTRACT(MONTH FROM InvoiceDate AT TIME ZONE 'UTC') AS month,
# SUM(Quantity * UnitPrice) AS Revenue

# FROM [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data

# GROUP BY
# year,
# month

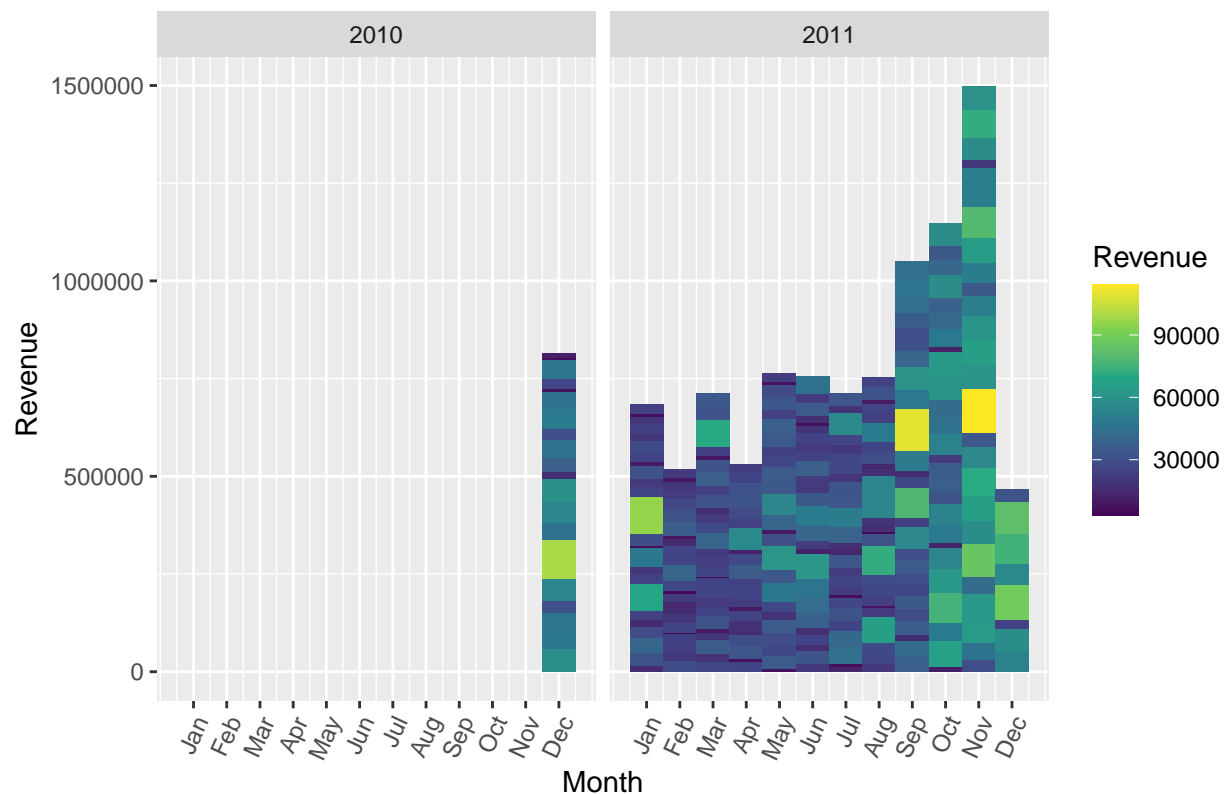
```

```

ecd1 <- read.csv("C:\\Users\\Michalis Loizos\\Desktop\\Data_Analysis_Programming\\E_Commerce_Data_Analy
ecd1_clean <- ecd1 %>%
  filter(!is.na(month) & !is.na(Day) & !is.na(Revenue))
ggplot(data = ecd1_clean) +
  geom_col(mapping = aes(x = month, y = Revenue, fill = Revenue),
    width = 1) +
  facet_wrap(~year) +
  labs(x = "Month", y = "Revenue",
    title = "Figure 1: Revenue generated per month") +
  scale_x_continuous(breaks = 1:12, labels = month.abb[1:12]) +
  theme(axis.text.x = element_text(angle = 65, hjust = 1)) +
  scale_fill_viridis_c()

```

Figure 1: Revenue generated per month



```

# SELECT
# Country,
# COUNT(DISTINCT(CustomerID)) AS Number_of_customers

# FROM
# [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data

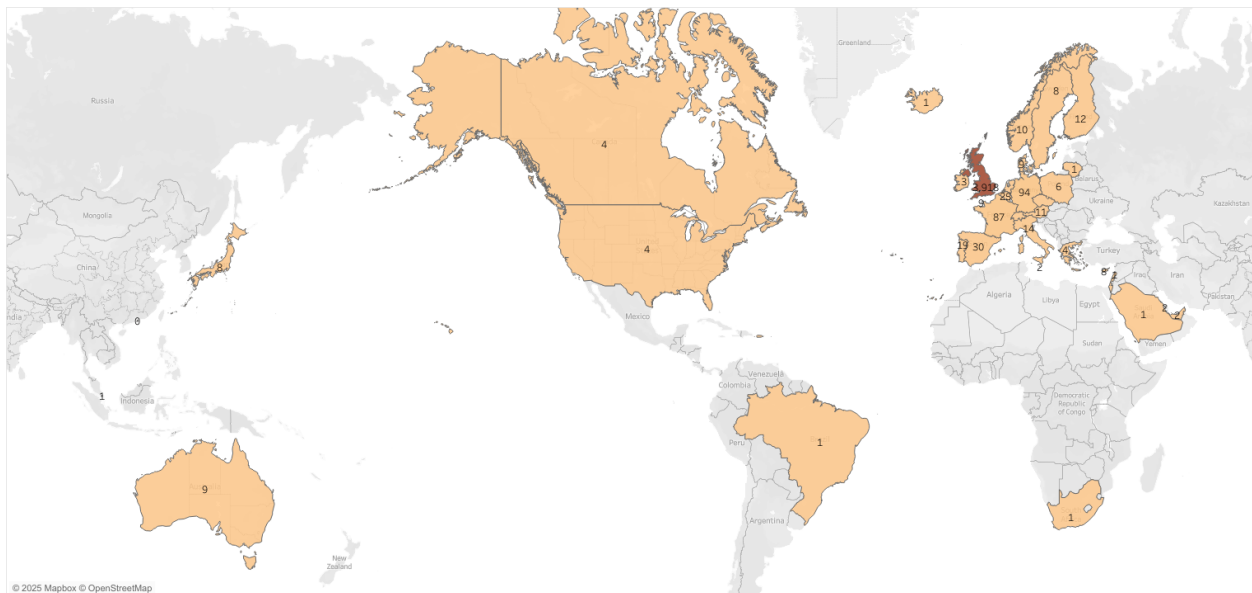
# GROUP BY
# Country

# ORDER BY
# Number_of_customers DESC

```

Customer Distribution by Country

Next, the country of origin for customers was examined. Great Britain has the highest number of customers, significantly surpassing other countries. This suggests that marketing efforts could be concentrated in this region to retain and further engage this dominant customer base.



Revenue by Country

To verify whether customer distribution aligns with financial performance, revenue per country was also analyzed. The results confirm that Great Britain contributes the majority of the platform's revenue, aligning with its high customer count.

```

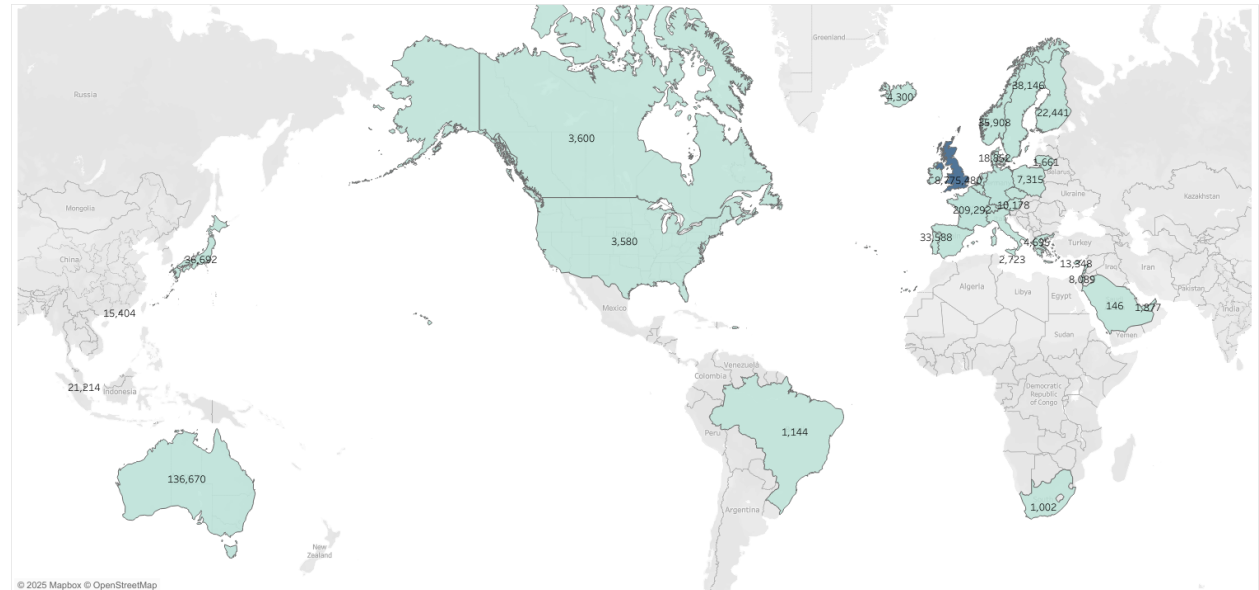
# SELECT
# Country,
# COUNT(DISTINCT(CustomerID)) AS Number_of_customers

# FROM
# [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data

# GROUP BY

```

```
# Country
# ORDER BY
# Number_of_customers DESC
```



Hourly Revenue Trends

Hourly trends were then explored to uncover patterns in revenue generation throughout the day. The data revealed that the time window from 10:00 to 15:00 consistently exhibits the highest average revenue.

These findings suggest that the platform could benefit from sending automated notifications, targeted emails, or launching special offers during peak hours to increase sales further and attract more customer engagement.

```
# SELECT
# Country,
# SUM(Quantity * UnitPrice) AS Revenue
# FROM
# [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data
# GROUP BY
# Country
# ORDER BY
# Revenue DESC
```

```
ecd4 <- read.csv("C:\\Users\\Michalis Loizos\\Desktop\\Data_Analysis_Programming\\E_Comme
ecd4_clean <- ecd4 %>%
  filter(!is.na(month) & !is.na(Day) & !is.na(Revenue))
data <- ecd4_clean %>%
  mutate(
    hour_of_day = as.numeric(Hour)
```

```

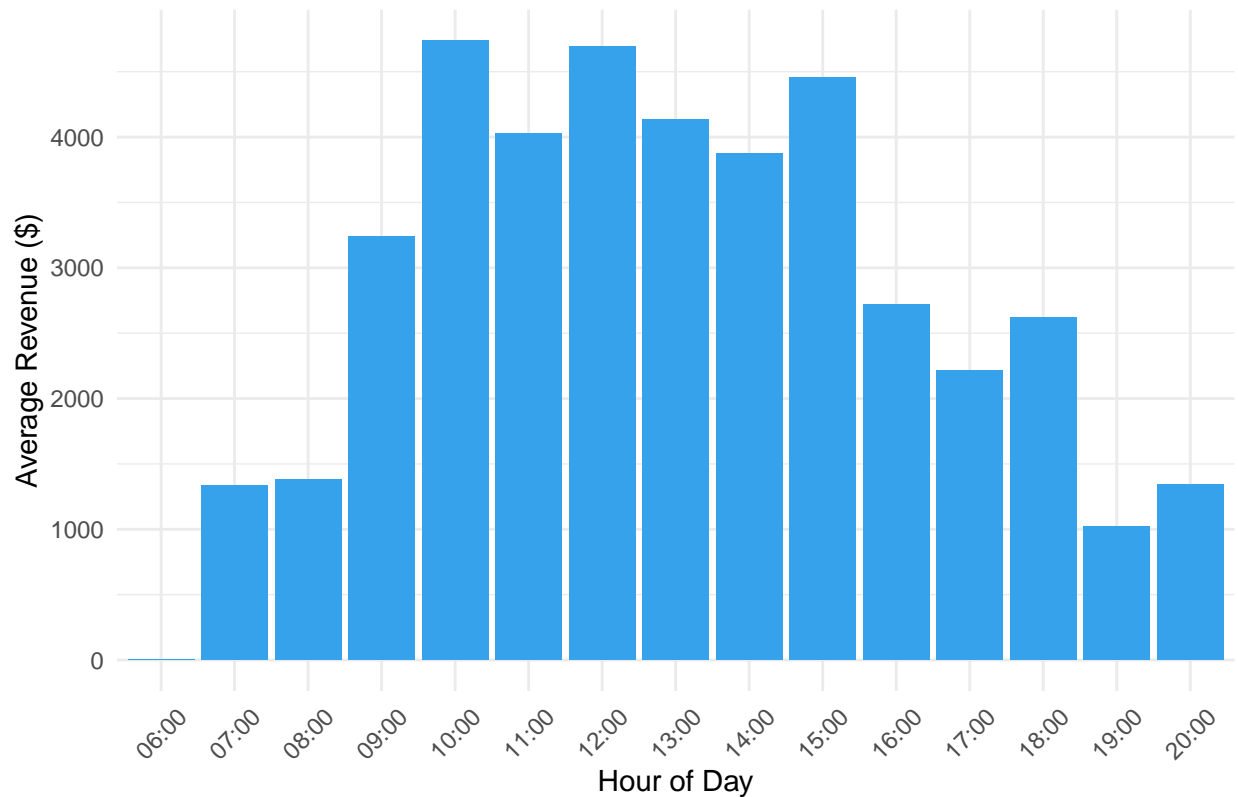
)

avg_revenue <- data %>%
  group_by(hour_of_day) %>%
  summarise(avg_revenue = mean(Revenue, na.rm = TRUE)) %>%
  ungroup()

ggplot(avg_revenue, aes(x = sprintf("%02d:00", hour_of_day), y = avg_revenue)) +
  geom_bar(stat = "identity", fill = "#36A2EB") +
  labs(
    title = "Figure 4: Average Revenue by Hour of Day (Dec 2010 - Dec 2011)",
    x = "Hour of Day",
    y = "Average Revenue ($)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, vjust = 0.5)
  )

```

Figure 4: Average Revenue by Hour of Day (Dec 2010 – Dec 2011)



Top Products by Revenue

A breakdown of revenue by product revealed the top-selling items. These products could be promoted more prominently or bundled into special offers to maximize their sales potential.

```

# SELECT
# EXTRACT(YEAR FROM InvoiceDate AT TIME ZONE 'UTC') AS year,
# EXTRACT(MONTH FROM InvoiceDate AT TIME ZONE 'UTC') AS month,
# EXTRACT(DAY FROM InvoiceDate AT TIME ZONE 'UTC') AS Day,
# EXTRACT(HOUR FROM InvoiceDate AT TIME ZONE 'UTC') AS Hour,
# SUM(Quantity * UnitPrice) AS Revenue

# FROM
# [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data

# GROUP BY
# year,
# month,
# Day,
# Hour
# ORDER BY
# Month

```

Top Customers by Revenue

The dataset also revealed which customers generated the most revenue. Identifying these high-value customers can support the development of loyalty programs, exclusive deals, or personalized outreach strategies to encourage continued spending.

```

# SELECT
# StockCode,
# SUM(Quantity * UnitPrice) AS Revenue

# FROM
# [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data

# GROUP BY

# StockCode

# ORDER BY
# Revenue DESC

```

```

# SELECT
# CustomerID,

# SUM(Quantity * UnitPrice) AS Revenue

# FROM
# [Project_Name].E_Commerce_Data.Cleaned_All_Customer_ID_Retail_Data

# GROUP BY
# CustomerID

# ORDER BY
# Revenue DESC

```

IV. Conclusions

In this project, the e-commerce dataset was cleaned, processed, and analyzed using BigQuery, R, and Tableau. The analysis revealed the following key insights:

- Revenue peaks between September and November, indicating seasonal sales opportunities.
- The most profitable time of day is between 10:00 and 15:00.
- Over 80% of both customers and total revenue originate from Great Britain.
- A subset of top customers and products generate a significant portion of the revenue.

These findings provide a solid foundation for implementing data-driven marketing, customer segmentation, and targeted advertising strategies.