Data Interaction HW 1

Ram M Kripa

2024-10-05

The Dataset

The dataset used in this assignment is the Customer Segmentation dataset from Kaggle. It is available at the following link: https://www.kaggle.com/datasets/yasserh/customer-segmentation-dataset

```
library(readxl)
library(here)
```

```
## here() starts at /Users/ramkripa/Desktop/Uchicago/Data Interaction/data_interaction_hw1
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
```

```
## v ggplot2 3.4.4     v purrr   1.0.2
## v tibble  3.2.1     v dplyr   1.1.4
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
```

```
## Warning: package 'dplyr' was built under R version 4.2.3
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
dataset <- read_excel(here("OnlineRetail.xlsx"))
summary(dataset)
```

```
##   InvoiceNo          StockCode         Description           Quantity
##  Length:541909      Length:541909      Length:541909      Min.   :-80995.00
##  Class :character   Class :character   Class :character   1st Qu.:     1.00
##  Mode  :character   Mode  :character   Mode  :character   Median :     3.00
##                                                           Mean   :     9.55
##                                                           3rd Qu.:    10.00
##                                                           Max.   : 80995.00
##
##   InvoiceDate                     UnitPrice           CustomerID
##  Min.   :2010-12-01 08:26:00.00   Min.   :-11062.06   Min.   :12346
##  1st Qu.:2011-03-28 11:34:00.00   1st Qu.:     1.25   1st Qu.:13953
##  Median :2011-07-19 17:17:00.00   Median :     2.08   Median :15152
##  Mean   :2011-07-04 13:34:57.16   Mean   :     4.61   Mean   :15288
##  3rd Qu.:2011-10-19 11:27:00.00   3rd Qu.:     4.13   3rd Qu.:16791
##  Max.   :2011-12-09 12:50:00.00   Max.   : 38970.00   Max.   :18287
##                                                       NA's   :135080
##    Country
##  Length:541909
```

```
##  Class :character
##  Mode  :character
##
##
##
##
```

Here, we can see the columns in the dataset. Using the Data Types from lecture, we can classify the variables in the columns as follows:

Column Name	Data Type	Description
InvoiceNo	Nominal	The invoice number of the transaction
StockCode	Nominal	The stock code of the product
Description	Nominal	The description of the product
Quantity	Quantitative	The quantity of the product
InvoiceDate	Quantitative (Interval)	The date of the transaction
UnitPrice	Quantitative (Ratio)	The unit price of the product
CustomerID	Nominal	The ID of the customer
Country	Nominal	The country of the customer

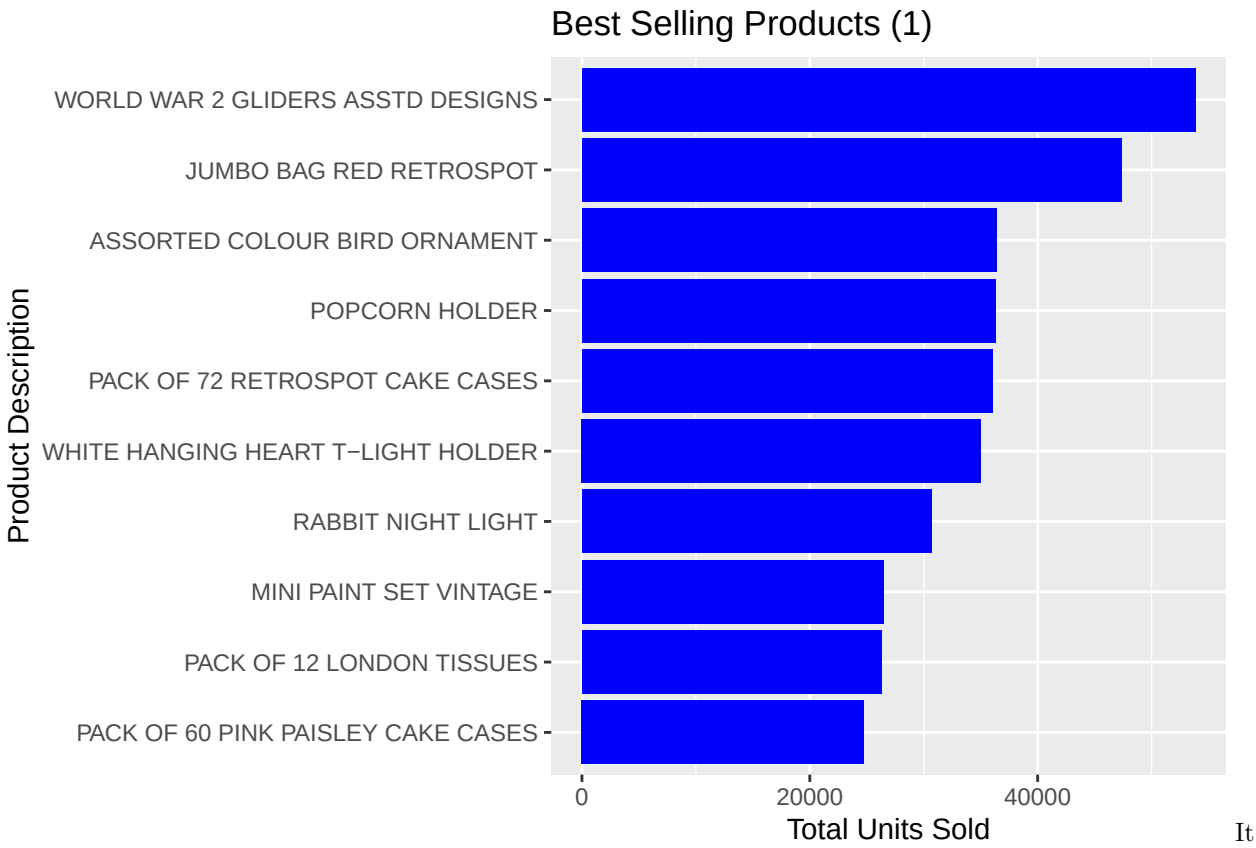
Now that we have our variables, we can ask some basic questions of the data

What were the best selling products in the dataset?

A natural question to ask is "What were the best selling products in the dataset?" One way we can answer this question by getting the total number of units sold for each product. We can then sort the products in descending order of the total quantity and display the top 10 products.

```
dataset %>%
  group_by(StockCode, Description) %>%
  summarise(TotalQuantity = sum(Quantity)) %>%
  arrange(desc(TotalQuantity)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(Description, TotalQuantity), y = TotalQuantity)) +
  geom_bar(stat = "identity", fill = "blue") +
  labs(x = "Product Description",
       y = "Total Units Sold",
       title = "Best Selling Products (1) ") +
  coord_flip()
```

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



It looks like consumers bought a lot of World war 2 Gliders Kit, and the second best selling product was Jumbo Bag Red Retrospot. But are these items really popular, or do a few customers buy them in bulk?

Another way to think about "best-sellers" is to count the total number of unique customers that bought the product.

```
dataset %>%
  group_by(StockCode, Description) %>%
  summarise(TotalCustomers = n_distinct(CustomerID)) %>%
  arrange(desc(TotalCustomers)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(Description, TotalCustomers), y = TotalCustomers)) +
  geom_bar(stat = "identity", fill = "red") +
  labs(x = "Product Description",
       y = "No. of Customers who bought it",
       title = "Best Selling Products (2) ") +
  coord_flip()
```

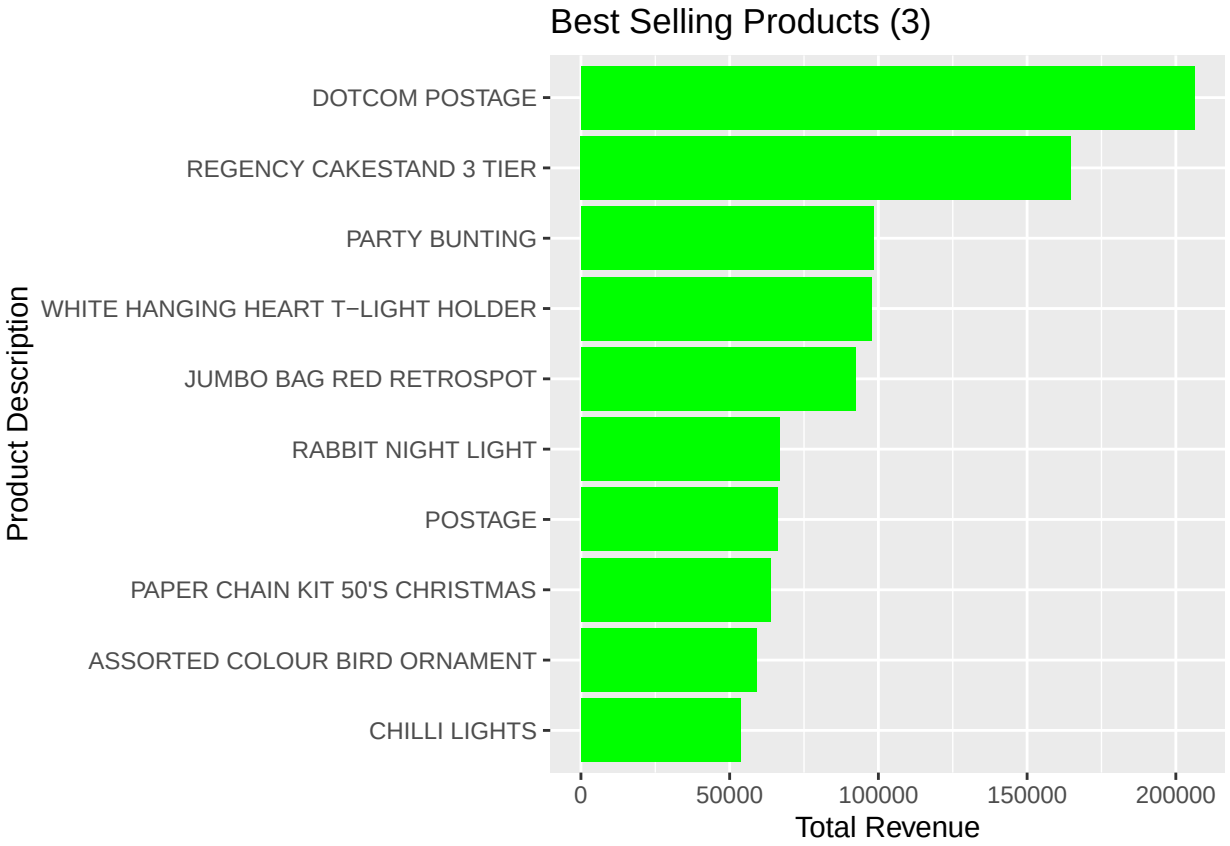
```
## `summarise()` has grouped output by 'StockCode'. You can override using the
## `.groups` argument.
```

From this visualization, we can see that a lot of customers bought the Regency Cakestand, and the White Hanging Heart T-Light Holder was the second most popular product. But did they generate revenue for the store?

Another way of determining what is the best-selling product is to ask which product generated the most revenue. We can calculate the total cost by multiplying the quantity sold by the unit price, and summing them up to get total revenue.

```
dataset %>%
  mutate(TotalCost = Quantity * UnitPrice) %>%
  group_by(StockCode, Description) %>%
  summarise(TotalRevenue = sum(TotalCost)) %>%
  arrange(desc(TotalRevenue)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(Description, TotalRevenue), y = TotalRevenue)) +
  geom_bar(stat = "identity", fill = "green") +
  labs(x = "Product Description",
       y = "Total Revenue",
       title = "Best Selling Products (3) ") +
  coord_flip()
```

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

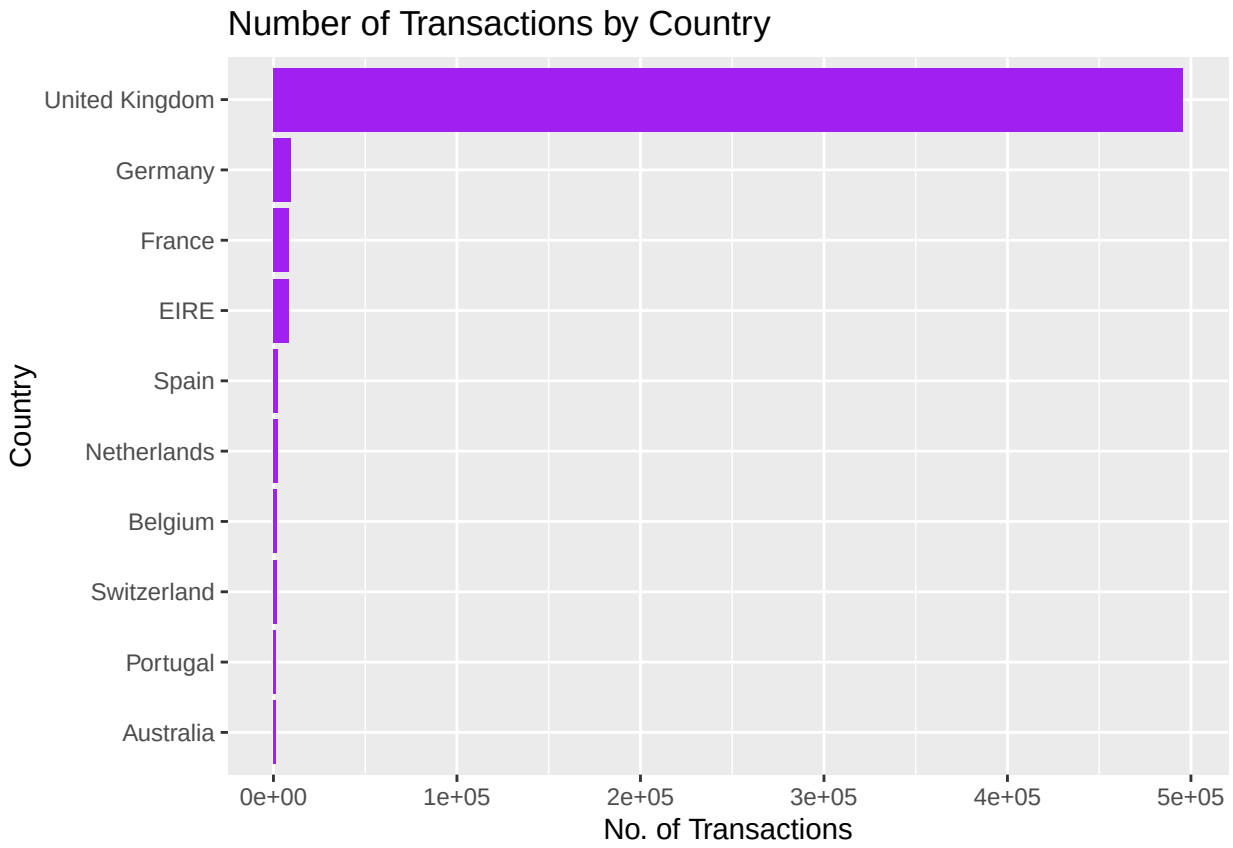


Here, we see that Postage generated the maximum revenue, and the Regency Cake stand was the second highest revenue generating product.

But, are different products popular in different countries? Let's keep revenue, and units as the two key metrics, and also split the products by country. Given that revenue and units are numeric, we should use position to represent them. We can use a bar chart to represent the data, since product description is nominal. Additionally, we can use a facet grid to separate the three metrics, and the country since they are also nominal. There are likely many products, so let us focus only on the top 2 by units sold per country. There are also 38 countries in the dataset, so we will focus on a select few.

First let's see which countries we have the most data for.

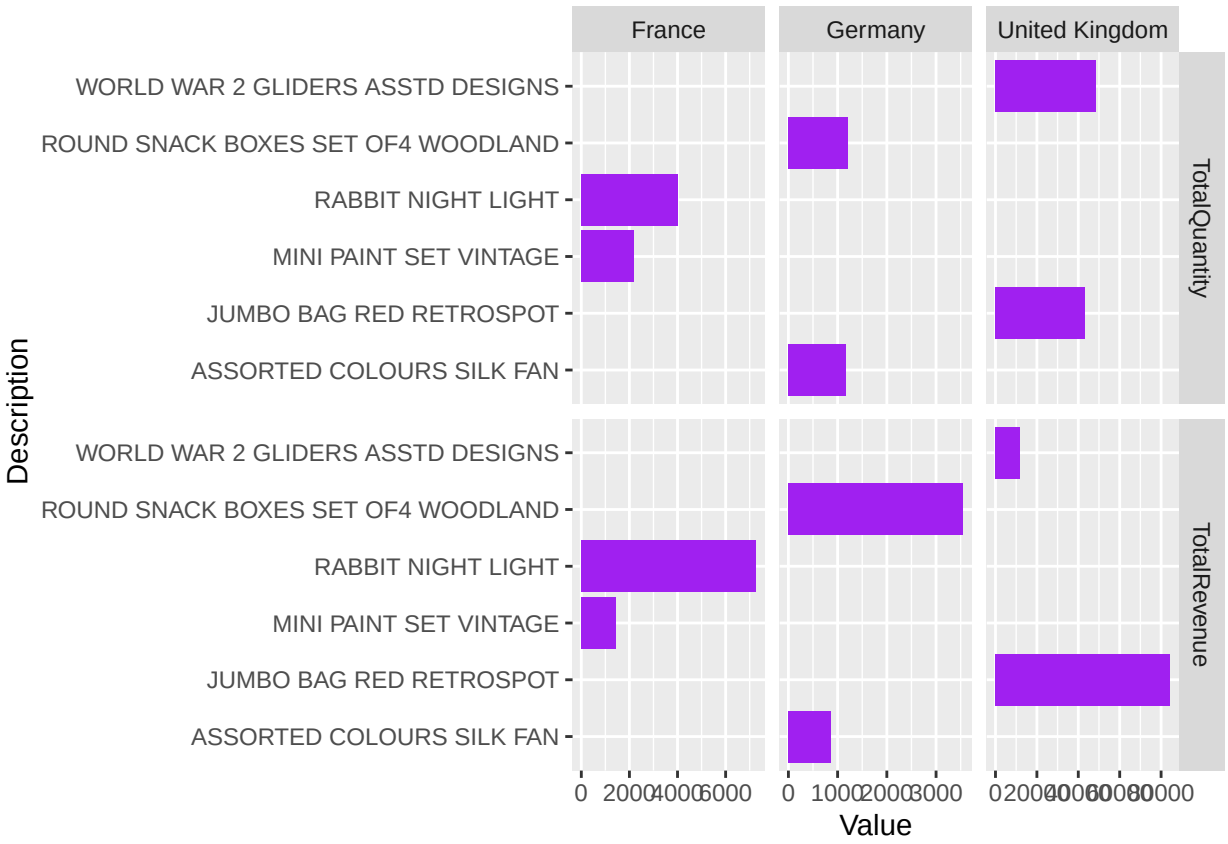
```
dataset %>%
  group_by(Country) %>%
  summarise(NoOfTransactions = n()) %>%
  arrange(desc(NoOfTransactions)) %>%
  head(10) %>%
  ggplot(aes(x = reorder(Country, NoOfTransactions), y = NoOfTransactions)) +
  geom_bar(stat = "identity", fill = "purple") +
  labs(x = "Country",
       y = "No. of Transactions",
       title = "Number of Transactions by Country") +
  coord_flip()
```



It looks like we have the most data for UK, Germany, and France. Let's focus on these three countries. From this data, it is obvious that the United Kingdom dataset far outstrips the other two countries in terms of the number of transactions, and the total revenue generated. Hence, we should use free scales for the y-axis to make the data more interpretable.

```
dataset %>%
  filter(Country %in% c("United Kingdom", "Germany", "France")) %>%
  mutate(TotalCost = Quantity * UnitPrice) %>%
  group_by(Country, StockCode, Description) %>%
  summarise(TotalQuantity = sum(Quantity), TotalRevenue = sum(TotalCost)) %>%
  group_by(Country) %>%
  top_n(2, TotalQuantity) %>%
  pivot_longer(cols = c(TotalQuantity, TotalRevenue), names_to = "Metric", values_to = "Value") %>%
  ggplot(aes(x = Description, y = Value)) +
  geom_bar(stat = "identity", fill = "purple") +
  coord_flip() +
  facet_grid(Metric ~ Country, scales = "free_x")
```

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



From these plots, it is clear that different products are popular in different countries.

What did sales look like in these three countries over time? Both in terms of number of units and total revenue

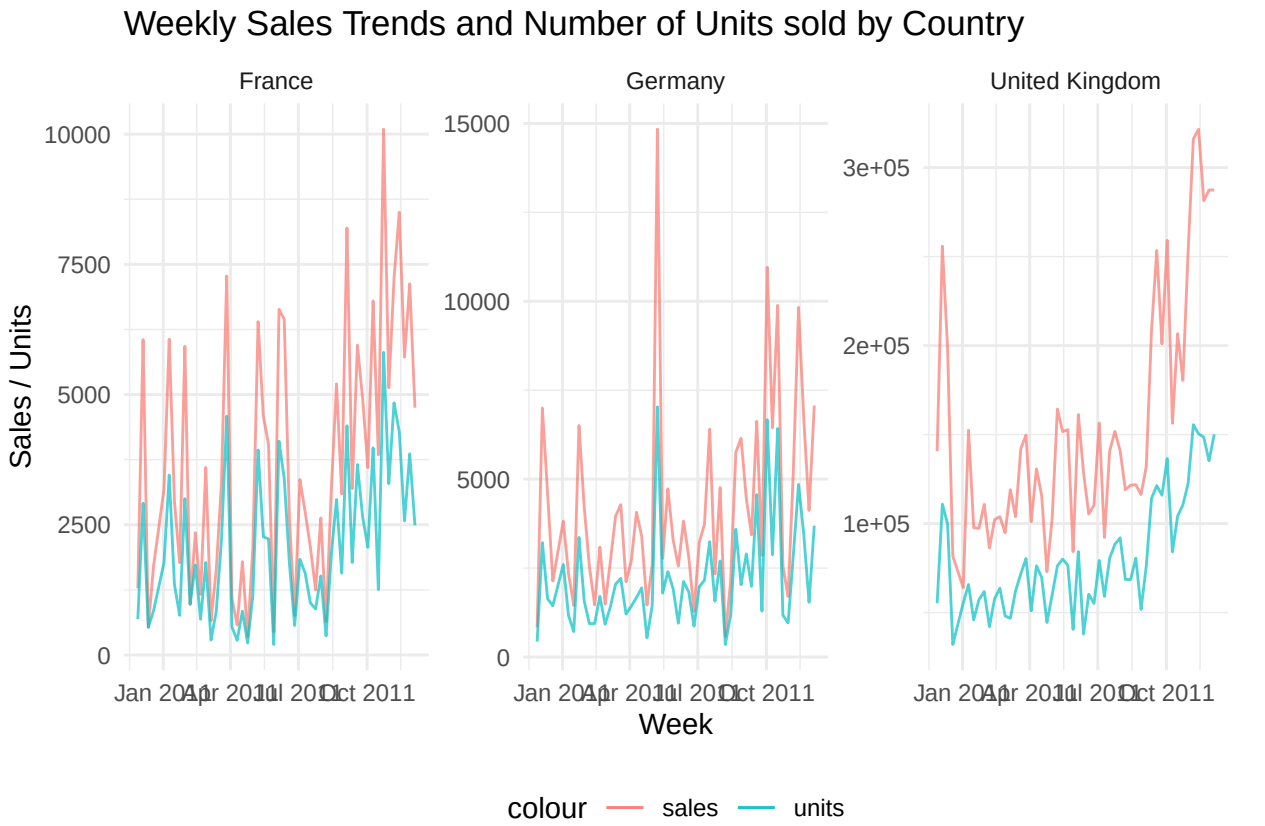
```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

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

```
# Data Preparation: Aggregate sales and number of units over time
sales_data <- dataset %>%
  filter(Country %in% c("United Kingdom", "Germany", "France")) %>%
  mutate(Week = floor_date(InvoiceDate, "week")) %>%
  group_by(Week, Country) %>%
  summarise(TotalSales = sum(Quantity * UnitPrice, na.rm = TRUE),
            TotalUnits = sum(Quantity, na.rm = TRUE),
            .groups = "drop")

# Plotting the data
ggplot(sales_data, aes(x = Week)) +
  geom_line(aes(y = TotalSales, color = "sales"), alpha = 0.7) + # Line plot for Total Sales
  geom_line(aes(y = TotalUnits, color = "units"), alpha = 0.7) + # Points for Total Units
  facet_wrap(~ Country, scales = "free_y") + # Facet by country
  labs(title = "Weekly Sales Trends and Number of Units sold by Country",
       x = "Week",
```

```
      y = "Sales / Units"
      ) +
  theme_minimal() +
  theme(legend.position = "bottom")
```



Here, I spent a bunch of time tweaking the visualization to play around with trying to visualize both sales and units together. I think reducing the alpha from 1 made it easier to not get confused when the lines overlapped. I also used a facet grid to separate the data by country, since the data is nominal. I also used position to represent the two metrics, since they are numeric. I think this visualization is effective because it allows us to see how sales and units sold have changed over time in the three countries. It also allows us to compare the trends in the three countries side by side. The choice to free the y scales was because otherwise the UK graph would have skewed the axes completely. This way, we can see the trends in the other two countries more clearly.

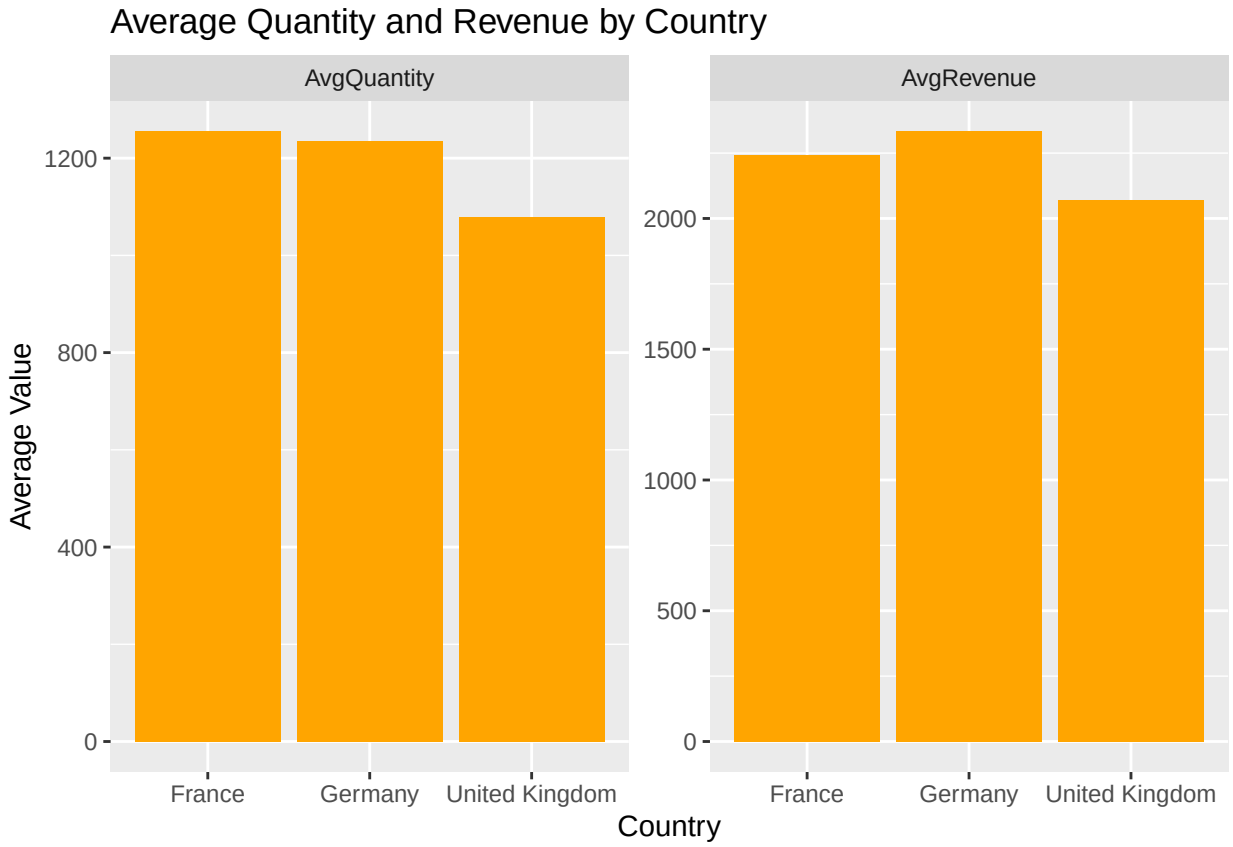
What are the Customers in different countries like?

To get a sense of how similar or different customers in different countries are, we can use the same dataset and look at the average quantity and revenue generated by customers in different countries. We can use a bar chart to represent the data, since country is nominal, and use position to represent the two metrics. Again, let's keep the analysis focused on the three countries above.

```
dataset %>%
  filter(Country %in% c("United Kingdom", "Germany", "France")) %>%
  mutate(TotalCost = Quantity * UnitPrice) %>%
  group_by(Country, CustomerID) %>%
  summarise(TotalQuantity = sum(Quantity), TotalRevenue = sum(TotalCost)) %>%
  group_by(Country) %>%
  summarise(AvgQuantity = mean(TotalQuantity), AvgRevenue = mean(TotalRevenue)) %>%
```

```
  pivot_longer(cols = c(AvgQuantity, AvgRevenue), names_to = "Metric", values_to = "Value") %>%
  ggplot(aes(x = Country, y = Value)) +
  geom_bar(stat = "identity", fill = "orange") +
  labs(x = "Country",
       y = "Average Value",
       title = "Average Quantity and Revenue by Country") +
  facet_wrap(~Metric, scales = "free_y")
```

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



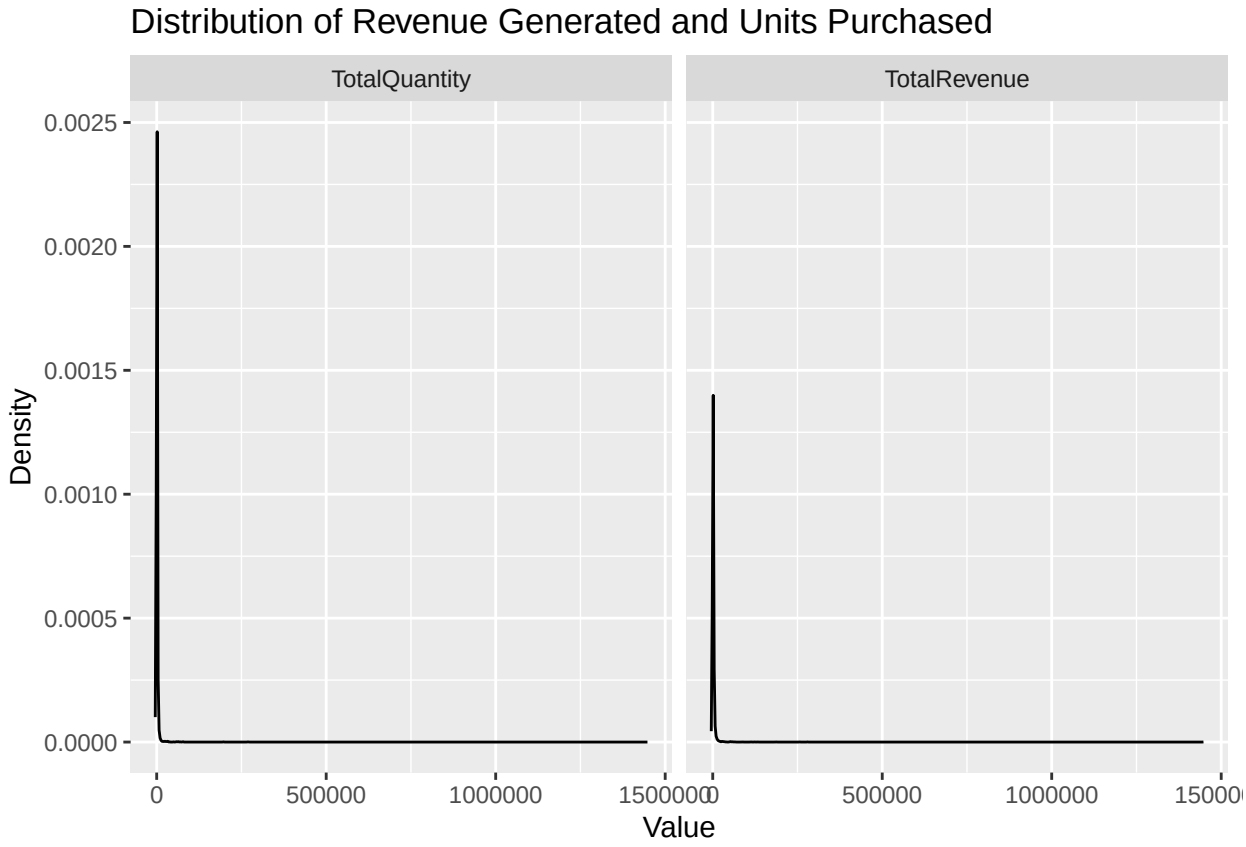
It looks like the average spend and number of products bought by consumers is roughly the same. But are the underlying distributions different? We can use histograms to find out.

First, let's plot a combined histogram of all the Countries' data. What is the distribution of # of products bought and revenue generated by customers?

```
dataset %>%
  #filter(Country %in% c("United Kingdom", "Germany", "France")) %>%
  mutate(TotalCost = Quantity * UnitPrice) %>%
  group_by(CustomerID) %>%
  summarise(TotalQuantity = sum(Quantity), TotalRevenue = sum(TotalCost)) %>%
  pivot_longer(cols = c(TotalQuantity, TotalRevenue), names_to = "Metric", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_freqpoly(stat = "density", fill = "pink") +
  labs(x = "Value",
       y = "Density",
       title = "Distribution of Revenue Generated and Units Purchased") +
```

```
  facet_wrap(~Metric)
```

```
## Warning in geom_freqpoly(stat = "density", fill = "pink"): Ignoring unknown
## parameters: `fill`
```

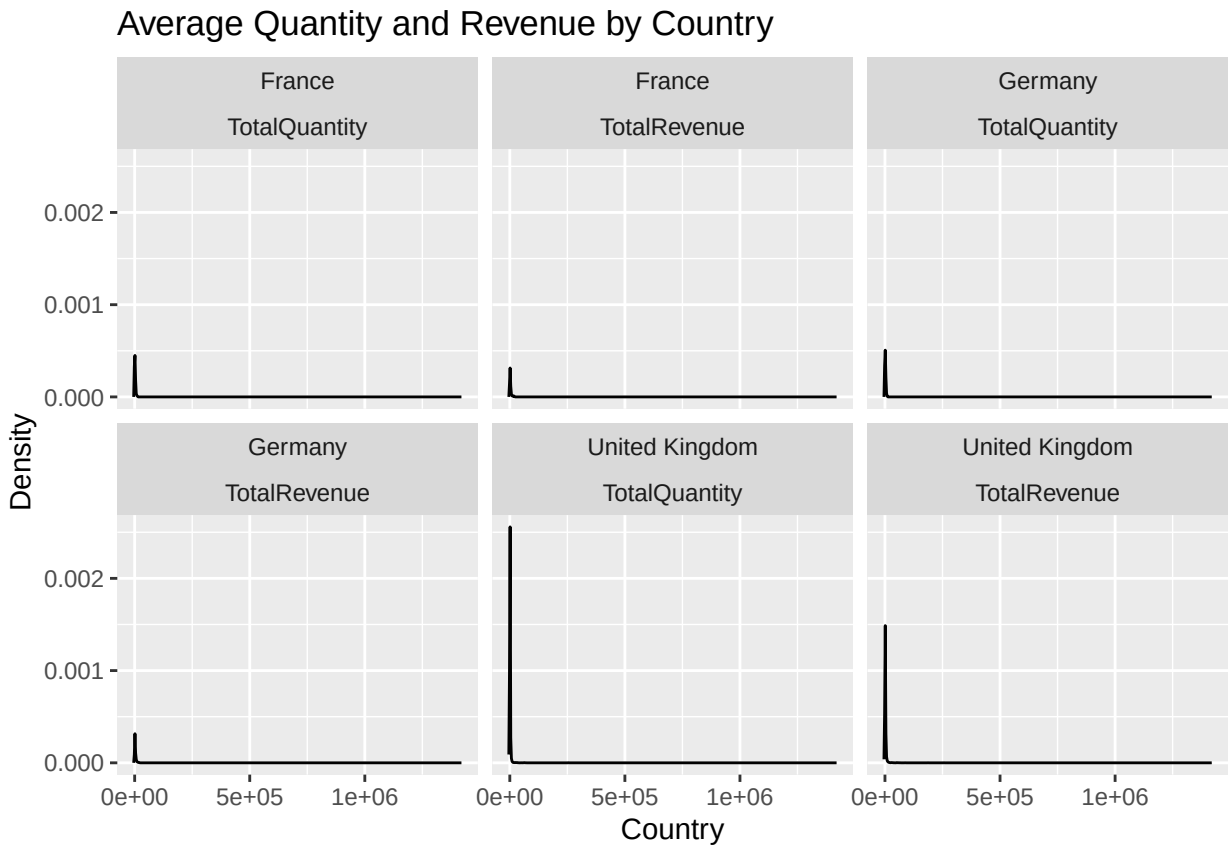


It would seem that on the whole, a lot of people spend a little and buy few products, and a few people buy many products/spend a lot of money. This is consistent with underlying assumptions about consumers. But are the distributions different for different countries?

```
dataset %>%
  filter(Country %in% c("United Kingdom", "Germany", "France")) %>%
  mutate(TotalCost = Quantity * UnitPrice) %>%
  group_by(Country, CustomerID) %>%
  summarise(TotalQuantity = sum(Quantity), TotalRevenue = sum(TotalCost)) %>%
  pivot_longer(cols = c(TotalQuantity, TotalRevenue), names_to = "Metric", values_to = "Value") %>%
  ggplot(aes(x = Value)) +
  geom_freqpoly(stat = "density", fill = "pink") +
  labs(x = "Country",
       y = "Density",
       title = "Average Quantity and Revenue by Country") +
  facet_wrap(Country~Metric)
```

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

```
## Warning in geom_freqpoly(stat = "density", fill = "pink"): Ignoring unknown
## parameters: `fill`
```

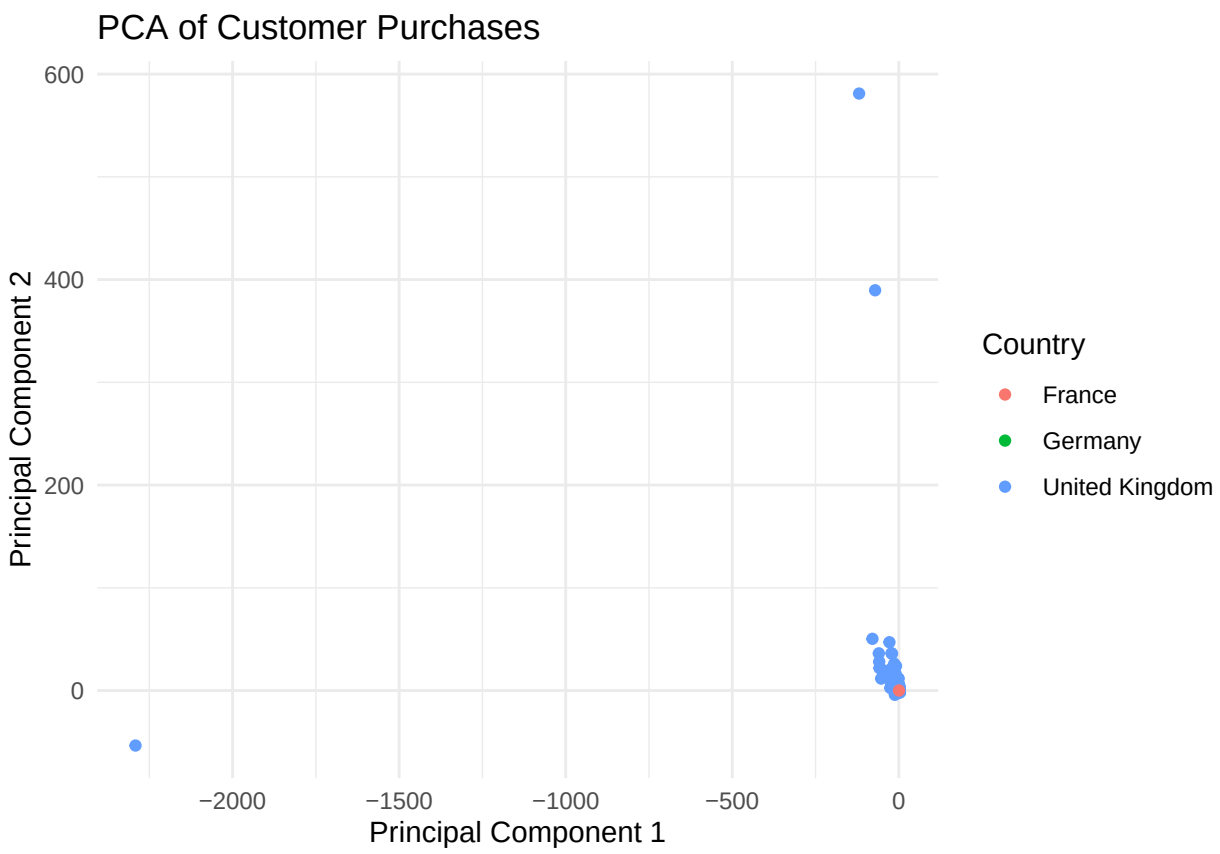


The distributions are similar for the different countries.

However, are different consumers from different countries interested in different products? We can create a matrix of consumers and the products they purchased and use PCA to reduce the dimensionality of the data. We can then plot the first two principal components to see if there are any patterns. We can use a biplot to represent the data, since the data is numeric, and use position to represent the two principal components. We can also color the points by country to see if there are any patterns.

```
# Summarize data to have one row per customer with their product quantities as a fraction of total purc
pca_data <- dataset %>%
  filter(Country %in% c("United Kingdom", "Germany", "France")) %>%
  group_by(CustomerID, Country, Description) %>%
  summarise(TotalQuantity = sum(Quantity), .groups = "drop") %>%
  #group_by(Country) %>%
  #mutate(Fraction = TotalQuantity / sum(TotalQuantity)) %>%
  #ungroup() %>%
  pivot_wider(names_from = Description, values_from = TotalQuantity, values_fill = 0)

# Remove constant or zero-variance columns
pca_data_filtered <- pca_data %>%
  select(-CustomerID, -Country) %>%
  select(where(~ var(.) != 0))

# Perform PCA on the filtered data
pca_result <- prcomp(pca_data_filtered, center = TRUE, scale. = TRUE)

# Extract PCA scores (first two principal components) and add back CustomerID and Country
pca_scores <- as.data.frame(pca_result$x) %>%
```

```
  select(PC1, PC2) %>%
  bind_cols(pca_data %>% select(CustomerID, Country))

# Plot the PCA results with ggplot2, coloring by country
ggplot(pca_scores, aes(x = PC1, y = PC2, color = Country)) +
  geom_point() +
  labs(title = "PCA of Customer Purchases", x = "Principal Component 1", y = "Principal Component 2") +
  theme_minimal() +
  theme(legend.position = "right")
```



It looks like there are outliers in the dataset. Maybe removing them would help? For the below plot, I played around with setting alpha and finally decided to set it to 0.2 for the UK data points. This way, the UK data points are still visible, but the other data points are not completely obscured by the massive number of UK data points. I look forward to receiving feedback on this particular point, since I actually wasn't quite sure what the best "visualization practice" would be.

```
pca_scores %>%
  filter(PC1 > -20) %>%
  filter(PC2 < 10) %>%
  mutate(transparency = if_else(Country == "United Kingdom", 0.2, 1)) %>%
# Plot the PCA results with ggplot2, coloring by country
ggplot(aes(x = PC1, y = PC2, color = Country, alpha = transparency)) +
  geom_point() +
  labs(title = "PCA of Customer Purchases", x = "Principal Component 1", y = "Principal Component 2") +
  theme_minimal() +
  theme(legend.position = "right")
```

From the PCA plot, it looks like there are no clear patterns in the data. This suggests that consumers from different countries are not interested in different products.

Conclusion

In this assignment, we used the Customer Segmentation dataset from Kaggle to answer some basic questions about the data. We found that different products were popular in different countries, and that the average spend and number of products bought by consumers was roughly the same across countries. We also found that the underlying distributions of revenue generated and units purchased were similar across countries. Finally, we found that consumers from different countries could not really be clustered on the basis of which products they bought.

Overall, this dataset provides a lot of interesting insights into consumer behavior, and could be used to drive business decisions and strategies in different countries in the future.