Project 1

Kyle Tadokoro

2024-02-10

Introduction to Dataset and Relevance

This document will examine and interpret what conclusions can be made using the “larger_sales_dataset.csv” obtained from Hassane Skikri through kaggle.com. For convenience sake, we will be referring to the business this dataset comes from as BusA Sales. The set consists of 10000 observations of 10 variables. Though the data in this set is fictional, it represents a realistic example of data that a business may collect during standard operation. Being able to clean, sort, and make sense of such a dataset is a skill that would be beneficial for the vast majority of businesses to employ.

```
df = read.csv("larger_sales_dataset.csv", sep = ",", header = T)
```

Summary of BusA Sales Data

There are a total of 10 variables in the BusA data set. The fields that we will be focusing on in this analysis will be: `Product Category`, `Total Price`, `Order Date`, `Payment Type`, and `Order Status`.

```
summary(df)
```

```
  Order.ID          Product.ID        Product.Category      Quantity    
 Length:10000       Length:10000       Length:10000       Min.   :1.00  
 Class :character   Class :character   Class :character   1st Qu.:2.00  
 Mode  :character   Mode  :character   Mode  :character   Median :3.00  
                                                          Mean   :3.01  
                                                          3rd Qu.:4.00  
                                                          Max.   :5.00  
  Unit.Price      Total.Price       Order.Date        Customer.ID       
 Min.   : 10.07   Min.   :  10.09   Length:10000       Length:10000       
 1st Qu.:129.21   1st Qu.: 284.39   Class :character   Class :character   
 Median :251.67   Median : 602.82   Mode  :character   Mode  :character   
 Mean   :253.28   Mean   : 762.72                                         
 3rd Qu.:378.26   3rd Qu.:1129.88                                         
 Max.   :499.96   Max.   :2499.78                                         
 Payment.Type       Order.Status      
 Length:10000       Length:10000       
 Class :character   Class :character   
 Mode  :character   Mode  :character   
```

Through the following commands, we are able to see that the data was taken over the span of a year from January 1st, 2023 until December 31st, 2023 as well as having 10000 unique customers (no repeated customers throughout dataset).

```
range(df$Order.Date)
```

```
[1] "2023-01-01" "2023-12-31"
```

```
length(unique(df$Customer.ID))
```

```
[1] 10000
```

Questions to answer about this set

- In which category the most money is being made?
- What time of year are sales peaking for each category? Overall?
- What is the average total for orders? What does this mean regarding optimizing resource distribution?
- What percentage of sales are refunded or cancelled?

Where is the Money Being Made?

When taking a look at the set, we can see that BusA Sales sells items from any of 6 categories.

- Sports & Outdoors
- Home & Kitchen
- Beauty & Health
- Books
- Electronics
- Clothing

Sales Generated by Each Category

```
Yearly.Total = c(sum(df$Total.Price[df$Product.Category == "Sports & Outdoors"]), sum(df
$Total.Price[df$Product.Category == "Home & Kitchen"]), sum(df$Total.Price[df$Product.Ca
tegory == "Beauty & Health"]), sum(df$Total.Price[df$Product.Category == "Books"]), sum
(df$Total.Price[df$Product.Category == "Electronics"]), sum(df$Total.Price[df$Product.Ca
tegory == "Clothing"]))

Yearly.Total <- currency(Yearly.Total, digits = 0L)

totals = data.frame(Product.Category = c("Sports & Outdoors", "Home & Kitchen", "Beauty
& Health", "Books", "Electronics", "Clothing"), Yearly.Total)
arrange(totals, desc(Yearly.Total))
```

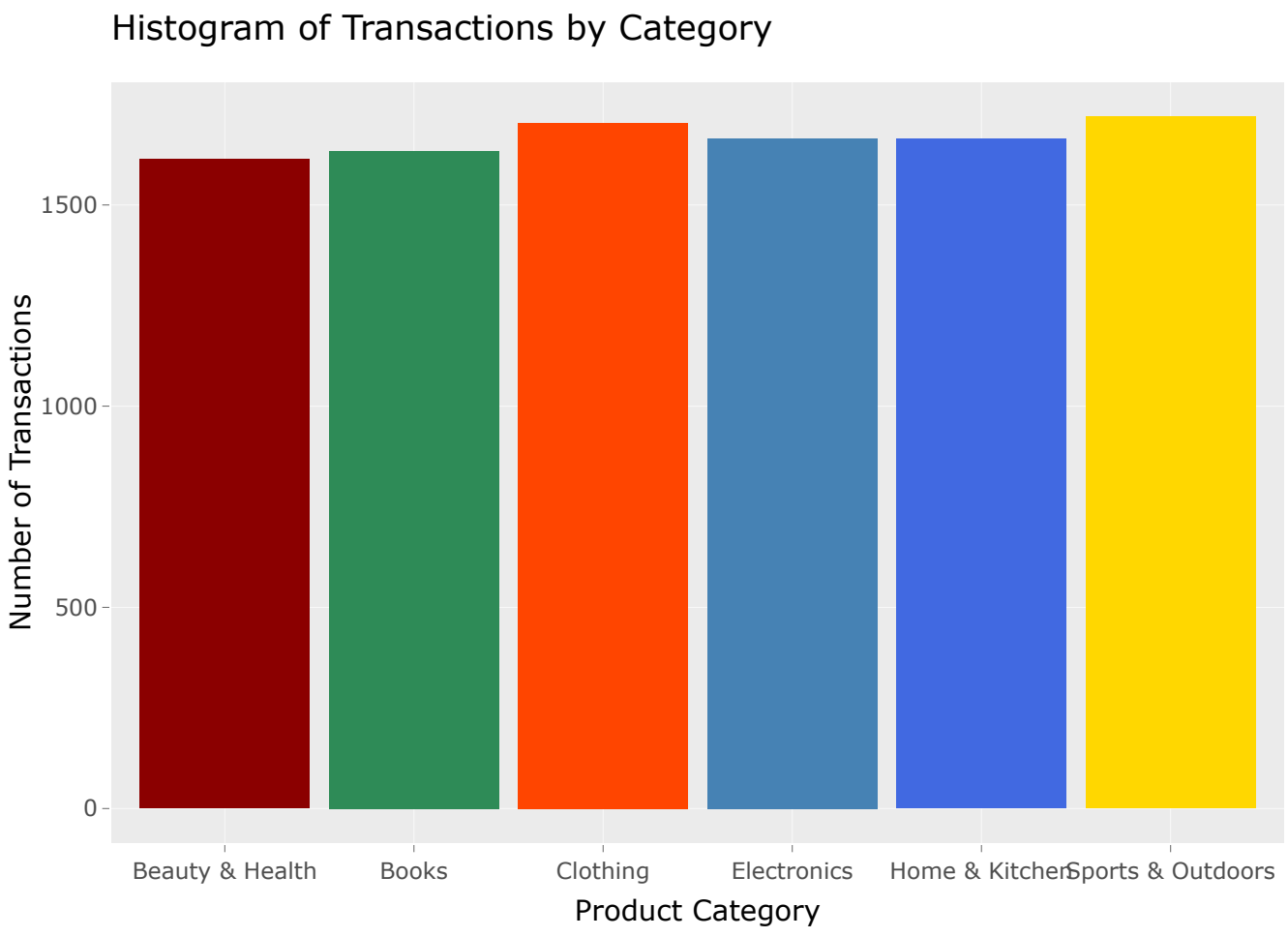
```
   Product.Category Yearly.Total
1 Sports & Outdoors   $1,313,735
2          Clothing   $1,303,679
3       Electronics   $1,290,283
4    Home & Kitchen   $1,268,472
5             Books   $1,249,307
6   Beauty & Health   $1,201,765
```

```
max(Yearly.Total) - min(Yearly.Total)
```

```
[1] $111,970
```

Number of Transactions

```
salesGG = ggplot(df, aes(x=Product.Category))+geom_bar(fill = c("darkred", "seagreen",
"orangered", "steelblue", "royalblue", "gold"))+ labs(x = "Product Category", y = "Numbe
r of Transactions", title = "Histogram of Transactions by Category")
ggplotly(salesGG)
```



```
       Category             Data
1  Most Popular Sports & Outdoors
2 Least Popular   Beauty & Health
3         Range  105 Transactions
```

10 Biggest & Smallest Orders

We are also able to see that the top 10 Biggest orders of the 2023 year averaged at Total Price of $2491.64. It is also worth noting that in this top ten, we see the category of Books 3 times, Beauty & Health and Electronics 2 times, and Home & Kitchen, Sports & Outdoors, and Clothing only once.

```
topTen = arrange(select(df, Total.Price, Product.Category), desc(Total.Price))
head(topTen, n = 10)
```

```
   Total.Price  Product.Category
1     2499.785    Home & Kitchen
2     2498.582   Beauty & Health
3     2494.533             Books
4     2494.450       Electronics
5     2492.415   Beauty & Health
6     2490.066 Sports & Outdoors
7     2488.129             Books
8     2487.236       Electronics
9     2485.949          Clothing
10    2485.286             Books
```

```
tenAvg = round(sum(topTen$Total.Price[1:10]) / 10, digits = 2)
tenAvg
```

```
[1] 2491.64
```

```
tail(topTen, n = 10)
```

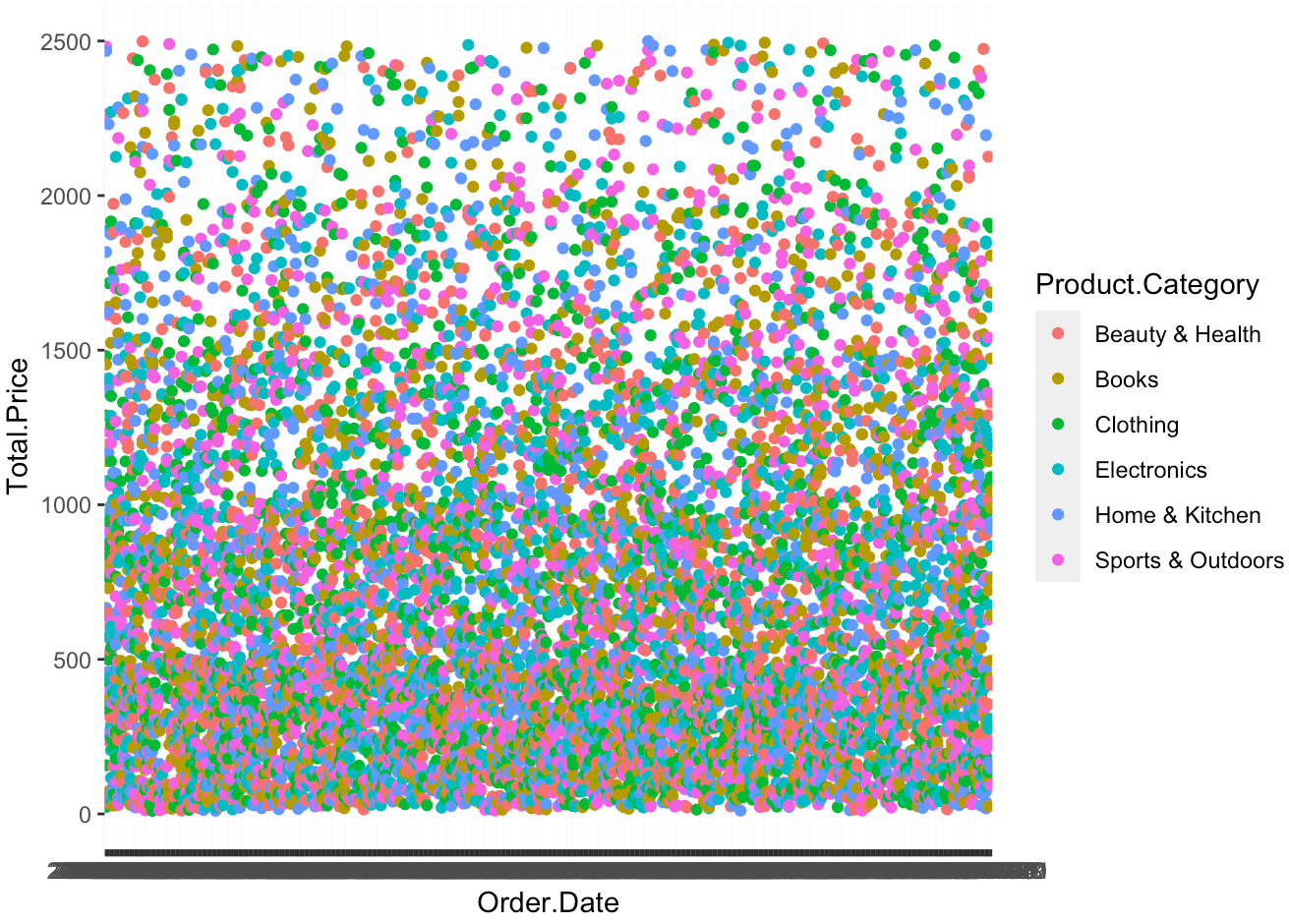
```
      Total.Price  Product.Category
9991     13.00103    Home & Kitchen
9992     12.18653    Home & Kitchen
9993     12.15613 Sports & Outdoors
9994     12.11102 Sports & Outdoors
9995     11.27107   Beauty & Health
9996     11.11300 Sports & Outdoors
9997     11.05853 Sports & Outdoors
9998     10.80243    Home & Kitchen
9999     10.59864    Home & Kitchen
10000    10.09248          Clothing
```

```
tenAvg = round(sum(topTen$Total.Price[9991:10000]) / 10, digits = 2)
tenAvg
```

```
[1] 11.44
```

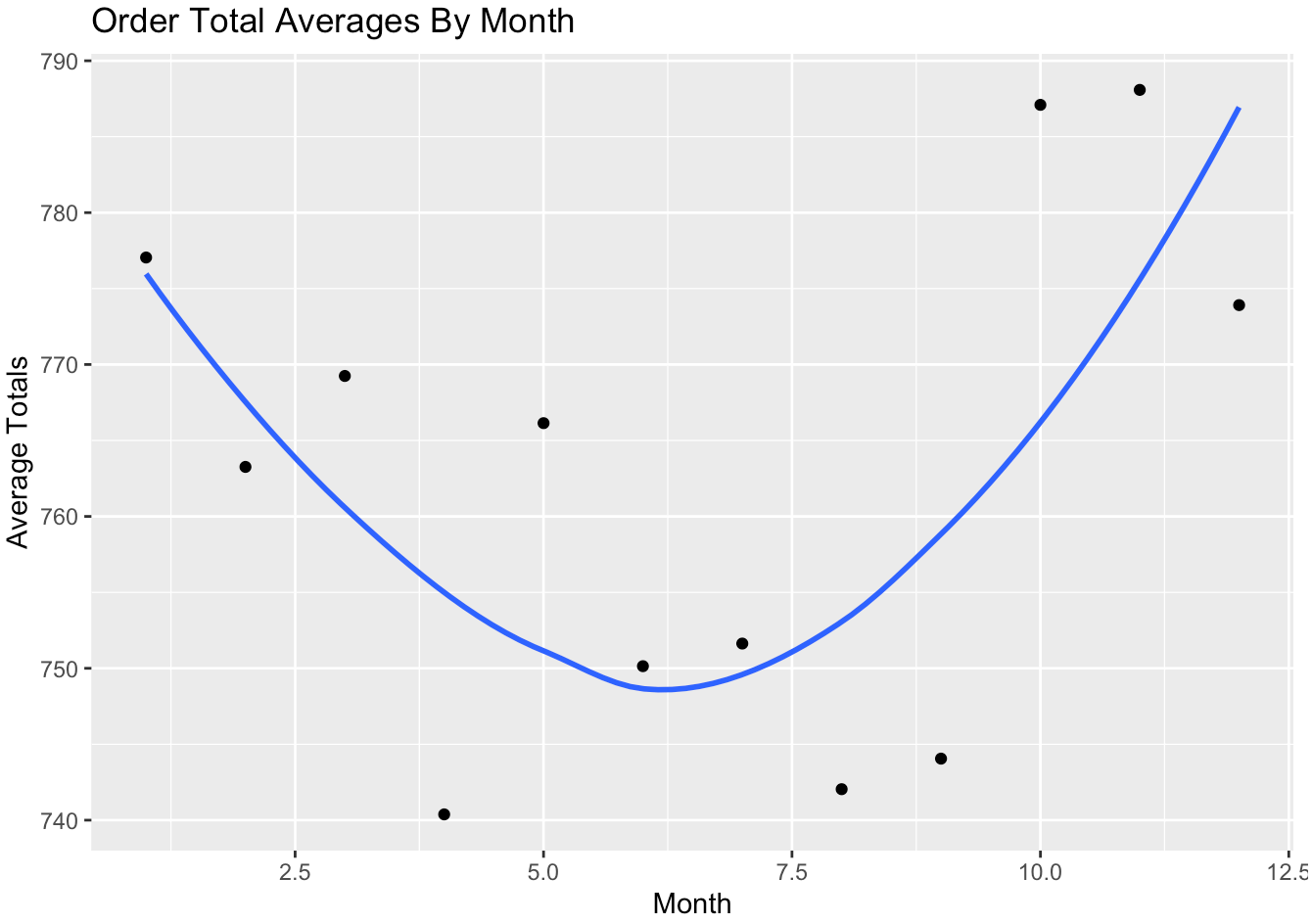
What Time of Year are Sales the Best?

When breaking down the distribution of sales throughout the year, it immediately becomes clear that the data in this file is fictional. With a normal set of data with variables such as ours, it is typical to expect patterns and trends to appear at different times of the year that correlate with seasons, holidays, and other factors. However, when plotting the total price of each observation against the order date, we can see an extraordinarily consistent layout develop through the 10,000 cases. In order to make the plot more legible, we can sort by category. Doing so does not reduce the amount of clutter, however it does show that the distribution for each of the categories is remarkably similar to eachother with no real outliers becoming apparant.

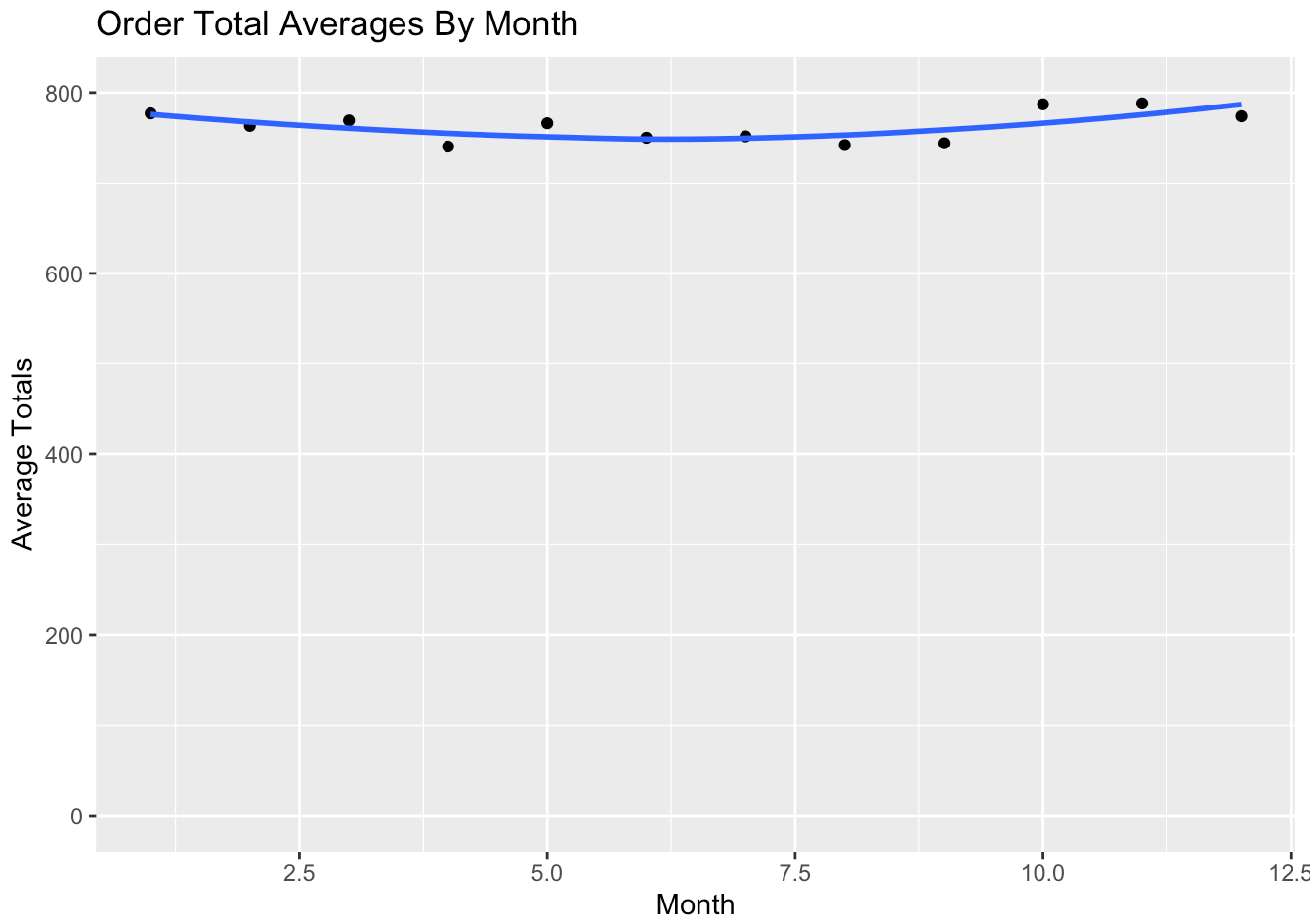


In order to summarize the plot above in a way that is easier understood, I took the average total for orders for each month and plotted them. Doing so, we are able to see that there is a slight deviation in order totals with the smallest amount made per order during the middle of the year before recovering back into the later months. When looking at the graph, it is important that we keep the y-axis in mind. Though the difference in average order totals seems dramatic, looking at the same plot with the bottom limit of our y-axis dropped to 0, we can see that the difference between the months is quite insignificant.

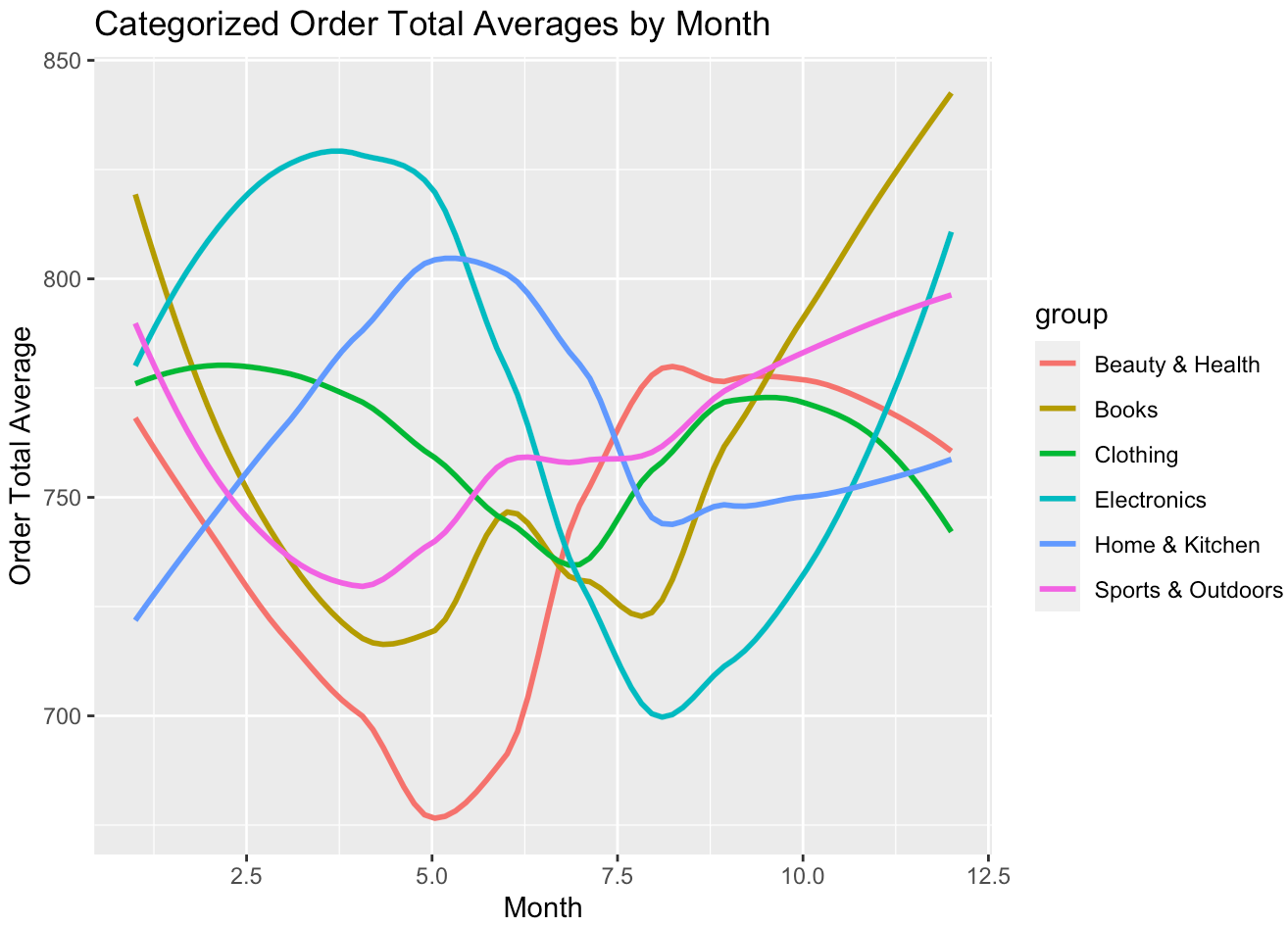
```
##Subset with sorted by Order.Date
dff = select(df, Order.Date, Total.Price, Product.Category)
dff$Order.Date = sort(format(as.Date(df$Order.Date), "%m"), decreasing = F)
dff$Order.Date = as.numeric(dff$Order.Date)

##Create vector of month totals
geeba = c()
for (i in 1:12){
  geeba = c(geeba, sum(dff$Total.Price[dff$Order.Date > (i - 1) & dff$Order.Date < (i + 1) ]))
  geeba[i] = geeba[i] / length(dff$Total.Price[dff$Order.Date > (i - 1) & dff$Order.Date < (i + 1) ])
}
monthDf = data.frame(c(1:12), geeba)
g = ggplot(monthDf, aes(c(1:12), geeba)) + geom_point() + geom_smooth(method = "loess", se = F, formula = y~x, span = 1) + labs(title = "Order Total Averages By Month", x = "Month", y = "Average Totals")
g
```

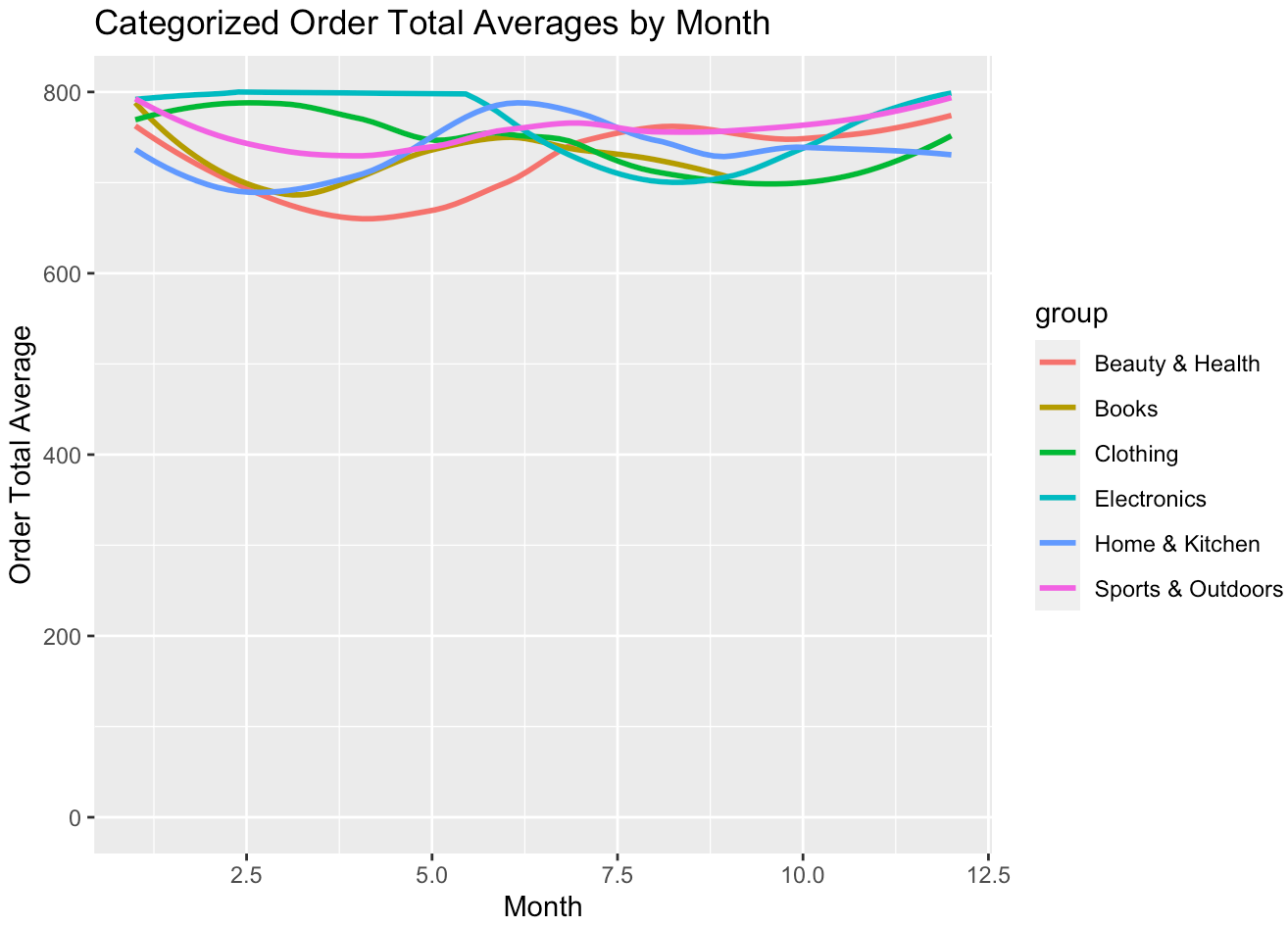



```
g + ylim(0,800)
```



Due to BusA Sales a variety of categories from which they sell their products from, it would also be beneficial for executives to understand how each of their categories are trending as well. In the plot below, I have used the same format as the one above only this time splitting the data into six individual curves.

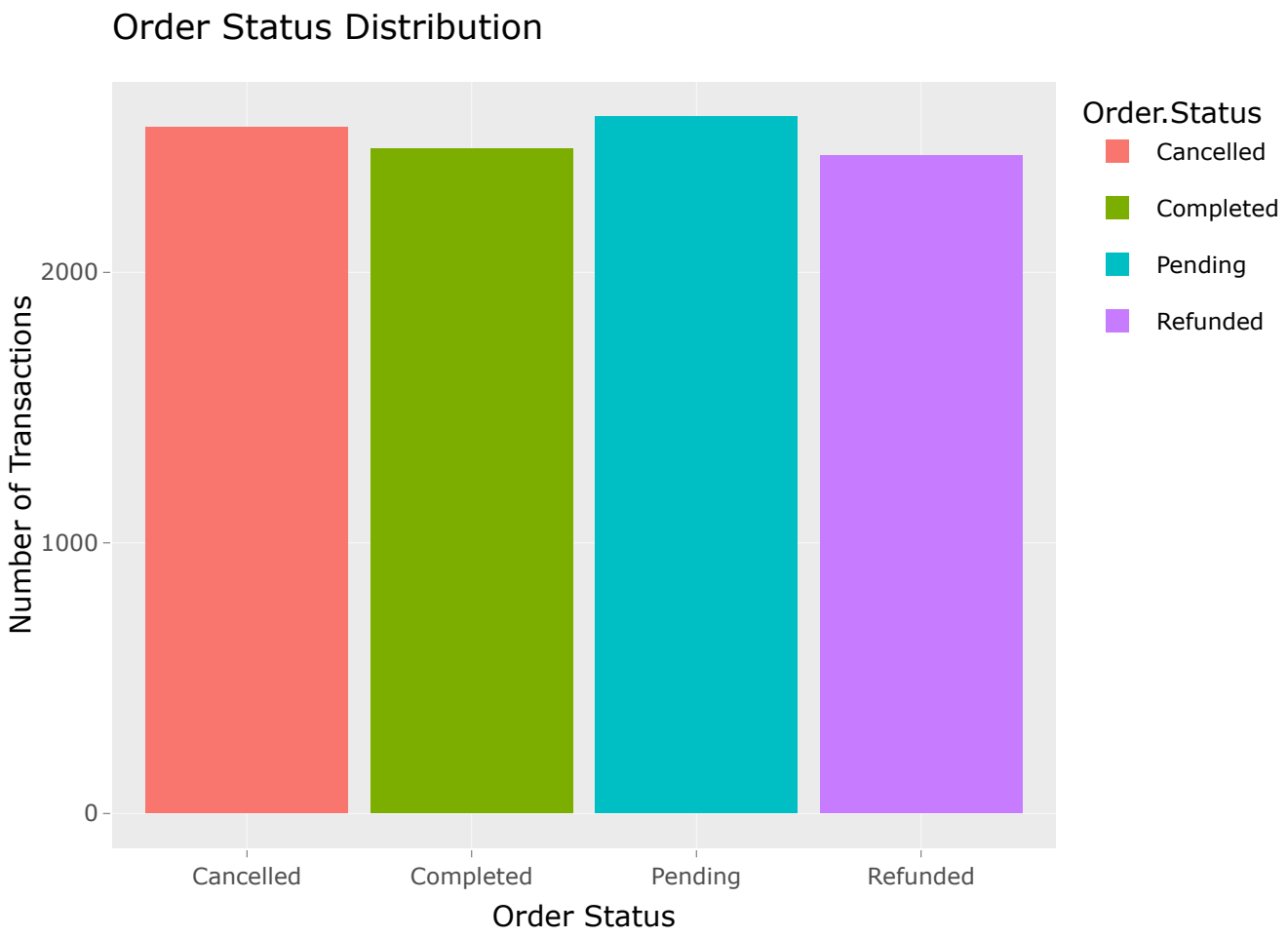

Categorized Order Total Averages by Month
Order Total Average
850
800
750
700
2.5
5.0
7.5
10.0
12.5
Month
group
Beauty & Health
Books
Clothing
Electronics
Home & Kitchen
Sports & Outdoors


Categorized Order Total Averages by Month
Order Total Average
800
600
400
200
0
2.5
5.0
7.5
10.0
12.5
Month
group
Beauty & Health
Books
Clothing
Electronics
Home & Kitchen
Sports & Outdoors

Percentage of sales refunded or Cancelled

Another important part of this data set is the amount of refunded and cancelled orders. Should this be a real business, having a refund and cancel rate of nearly 25% each would be unimaginable. In the plot below, we can see that the refunded and cancelled orders are nearly level with those of the completed and pending orders.

```
methodGG = ggplot(df, aes(x=Order.Status, fill = Order.Status))+geom_bar() + labs(title
= "Order Status Distribution", x = "Order Status", y = "Number of Transactions")
ggplotly(methodGG)
```

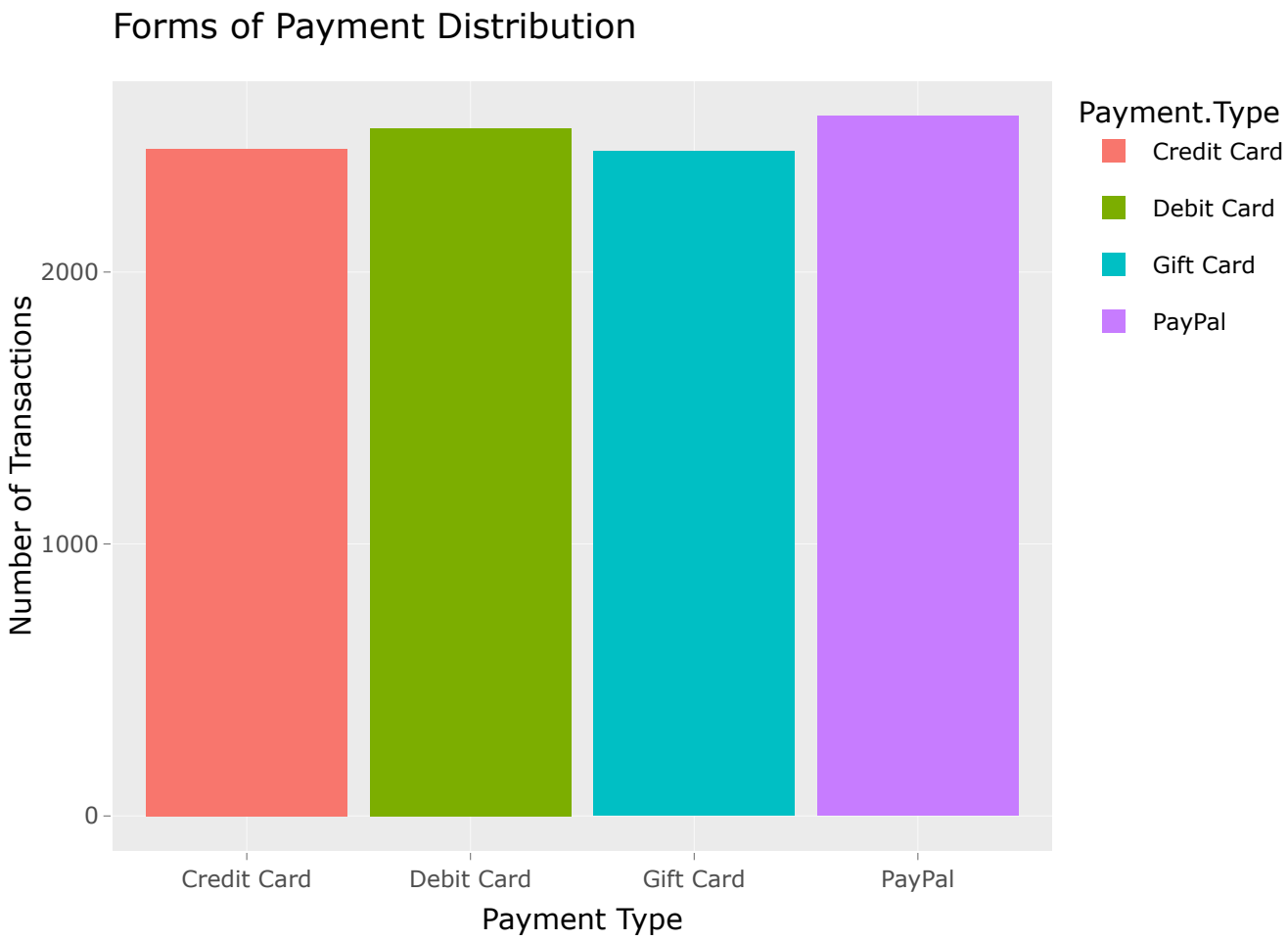



```
                  Stat       Data
1  Most Frequent Status    Pending
2 Least Frequent Status   Refunded
3                Range 144 Orders
4 Refunded or Cancelled     49.68%
```

Most used forms of Payment

Finally, below we have the distribution of payment types that customers used during the 2023 fiscal year. Similar to the Order Status plot, we observe a similar amount of transactions between all 4 Payment Types with the most used being Paypal.

```
payGG = ggplot(df, aes(x=Payment.Type, fill = Payment.Type))+geom_bar() + labs(title =
"Forms of Payment Distribution", x = "Payment Type", y = "Number of Transactions")
ggplotly(payGG)
```




```
        Type      Data
1  Most Used    Paypal
2 Least Used Gift Card
3      Range 128 Users
```

Conclusion

In conclusion, the “larger_sales_dataset” by Hassane Skikri produced some interesting results upon data cleaning and wrangling. We were able to determine that during the fiscal year of 2023, BusA Sales had a total of 10,000 customers. These customers ordered the most from the categories of Sports & Outdoors and Clothing and the least from Beauty & Health. They had the highest order totals in the months of October and November, used Paypal the most, and had an almost 50% rate of either returning or cancelling an order.

While the fields that we observed were realistic, the data seen in this file is highly unlikely to come about during an actual business’ operations. In reality, it is likely that you would expect a higher variance in customer preferences (both categorical and chronological), distribution of sales, and a (hopefully) lower margin of refunded/cancelled orders. In addition to what we were given in the set, another field that could prove useful towards BusA’s endeavors could be a customer satisfaction survey of some sort.