Data Mining Presentation

Taylor Piecukonis, Sophia Saenger, Nicolas Corona

Executive Summary

Electronic Sales

- Electronics company’s sales transactions over a one year period (September 2023 - September 2024) with 16 fields
- Includes information about each customer’s demographics, purchasing behaviors, and product types
- 5 product types: smartphones, laptops, tablets, smartwatches, and headphones
- The methods of exploratory analysis include answering questions of data distribution using the ggplot2, sqldf, and dplyr package in R

Cleaning the N/A From the Data Set

- Only one observations included a N/A value in one of the fields (Gender)
- Allowed for simple cleaning of the data
- Deleted the entire observation as we had 20,000 total observations
- Deleting one observation wouldn’t significantly impact our analysis of the data

```r
#Cleaning data; filter out one record where gender = N/A
cleaned <- final %>% filter(Gender != "#N/A")
```

cleaned	19999 obs. of 16 variables
Electronic_sa…	20000 obs. of 16 variables

Cleaning the Purchase Date column

- Need to convert the purchase date column to be in date format in order to perform date-based analyzations such as creating a line chart to view total sales over time
- Use of lubridate package
- Creates a new column titled ‘Month’ making the dataset now 17 variables

```r
#Cleaning data; transform purchase data column in time format
library(lubridate)

# Ensure the column is of type character
cleaned$Purchase.Date <- as.character(cleaned$Purchase.Date)

# Convert the Purchase Date column to Date format
cleaned$Purchase.Date <- ymd(cleaned$Purchase.Date)

# Extract month and year from Purchase Date
cleaned$Month <- floor_date(cleaned$Purchase.Date, "month")
```

cleaned	19999 obs. of 17 variables

Separating the Dataset Into Two Tables

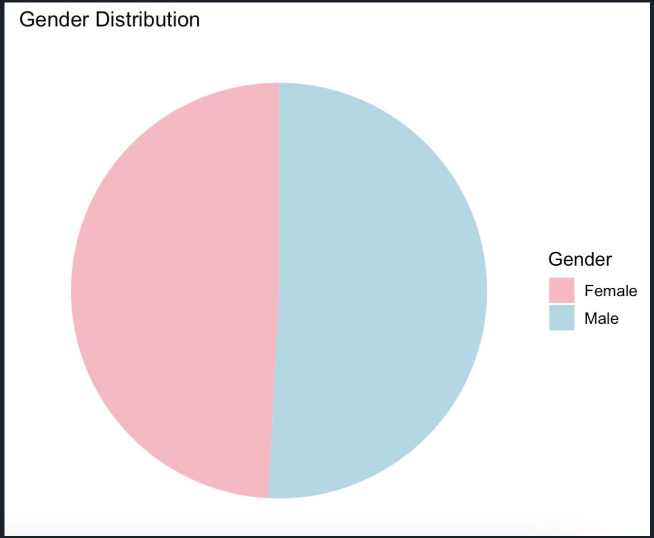
- From our dataset, we created one table dedicated to customer information and another table dedicated to transaction information

```r
# Create the Customer Table
customer_table <- cleaned %>%
  select(Customer.ID, Age, Gender, Loyalty.Member)


# Create the Transaction Table
transaction_table <- cleaned %>%
  select(Customer.ID, Product.Type, SKU, Rating, Order.Status, Payment.Method,
         Total.Price, Unit.Price, Quantity, Purchase.Date, Shipping.Type,
         Add.ons.Purchased, Add.on.Total)
```

customer_table	19999 obs. of 4 variables
transaction_ta...	19999 obs. of 13 variables

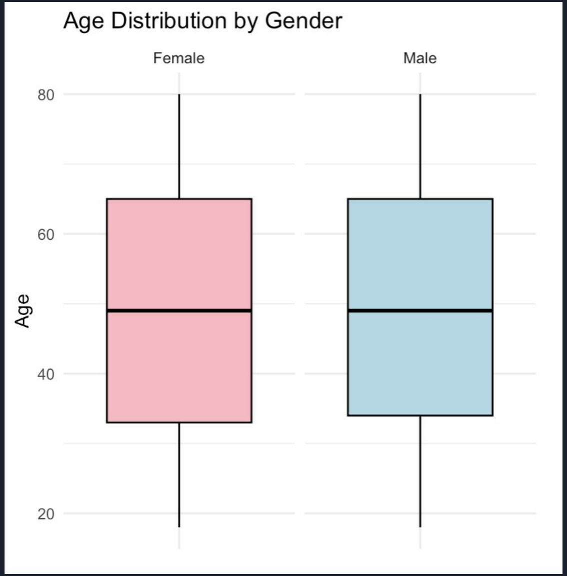
Visualization #1 - Who are the customers in terms of demographics?

- We want to understand the age, gender, and loyalty membership distribution of our data set



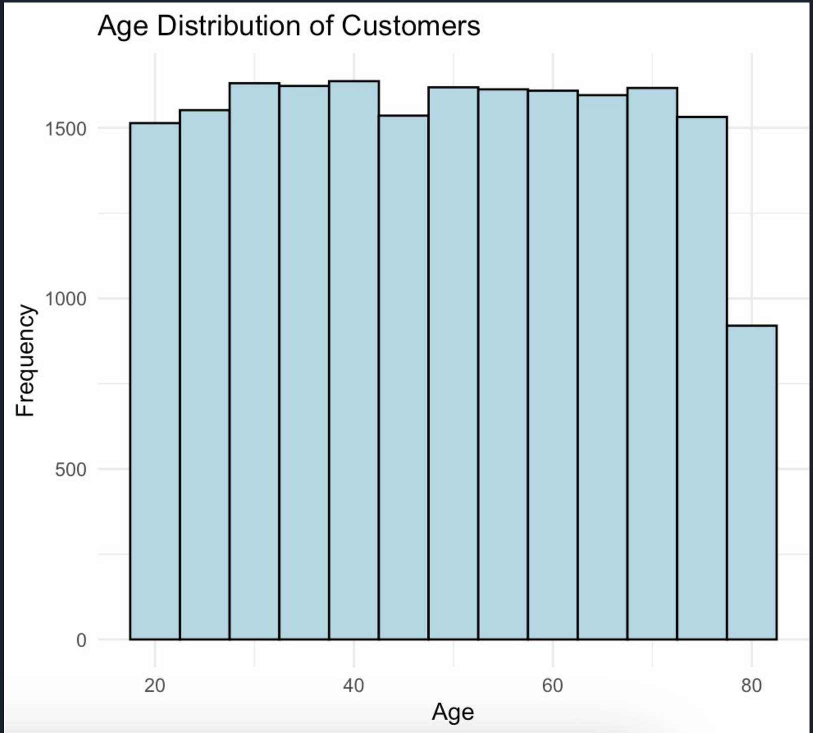
Geom_pie

Geom_bar

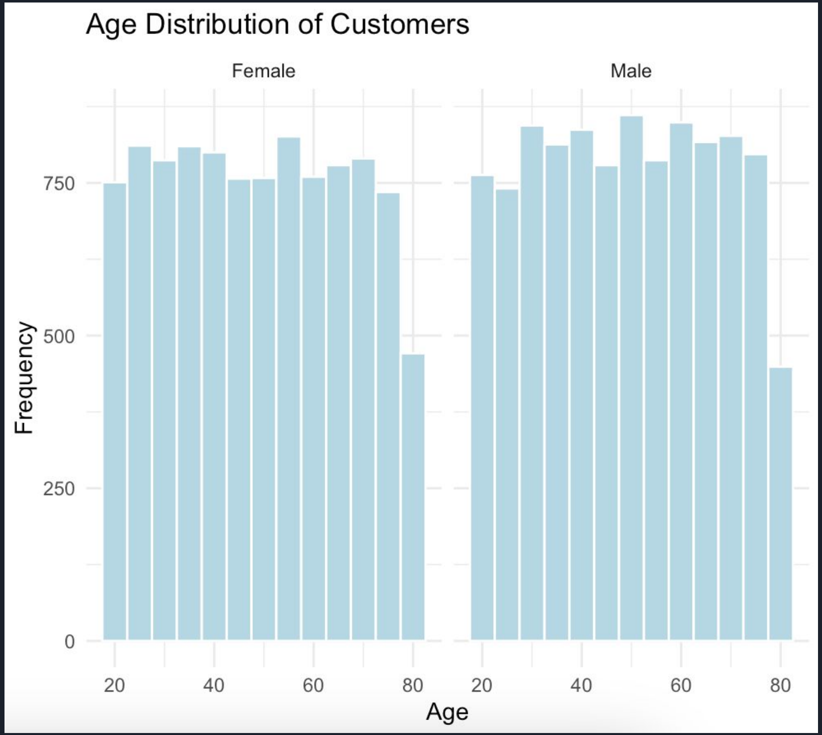


Geom_boxplot

Visualization #1 - Who are the customers in terms of demographics?



Facet_wrap



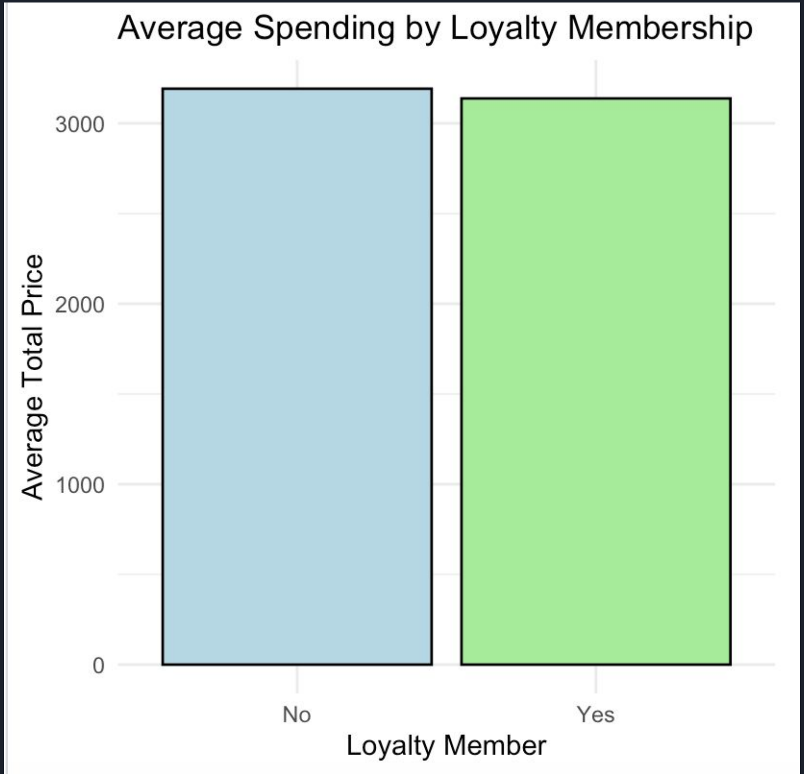
Visualization #2 - How does total spending differ by gender?

- Want to see if total spending differs by gender
- Overall, average spending by gender seems to be uniform
- Geom_violin

Visualization #3 - Do loyalty members spend more on average than non-loyalty members?

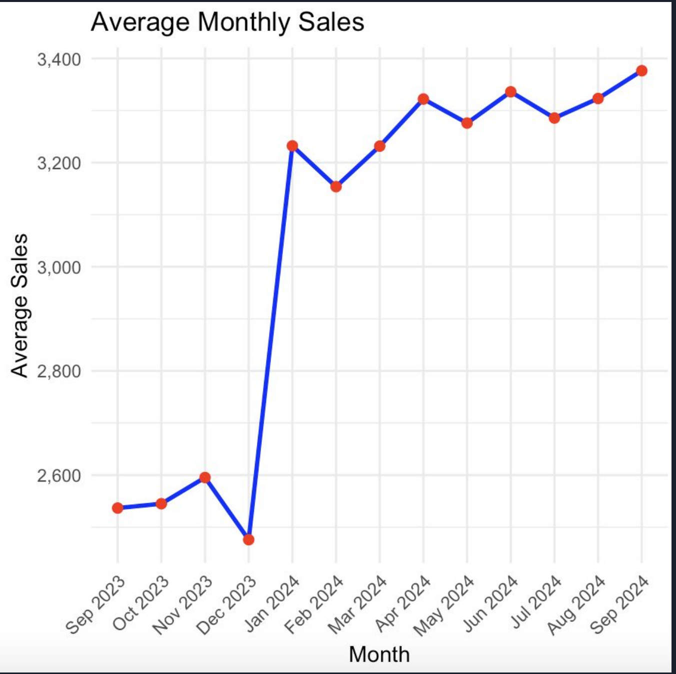
- You would expect loyalty members to spend more as they are often targeted with special promotions, discount, and rewards to incentivize more spending

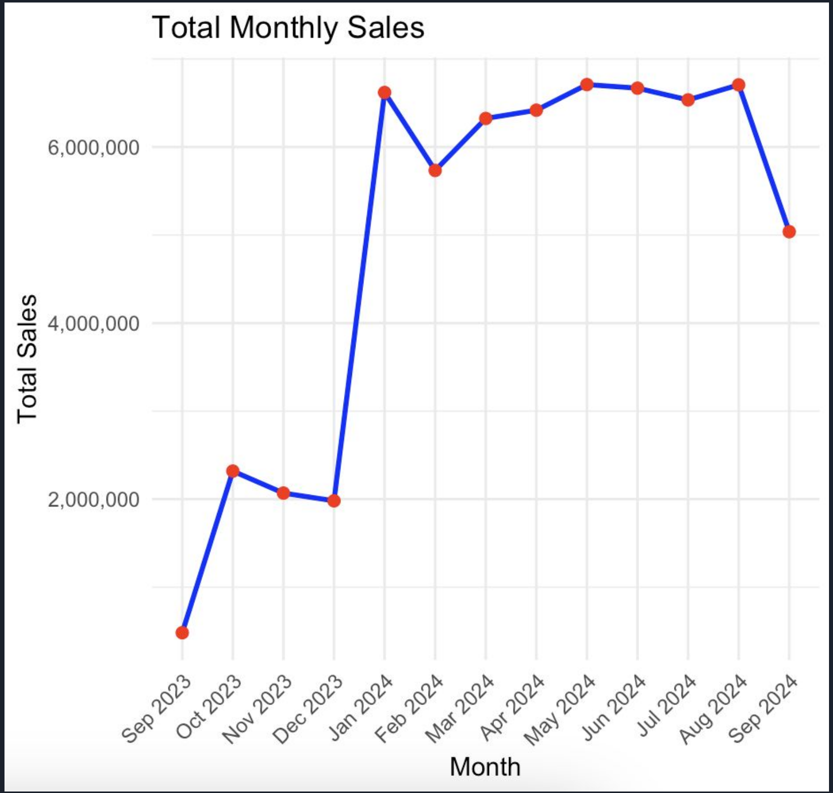
```
#Calculate average amount spent grouped by loyalty member or not
cleaned %>%
  select(Loyalty.Member, Total.Price) %>%
  filter(Loyalty.Member == "Yes") %>%
  summarize(member_average_purchases = mean(Total.Price))
#3138.011

cleaned %>%
  select(Loyalty.Member, Total.Price) %>%
  filter(Loyalty.Member == "No") %>%
  summarize(nonmember_average_purchases = mean(Total.Price))
#3191.975
```



Visualization #4 - Are there monthly trends in sales?

- Use of lubridate package when cleaning data



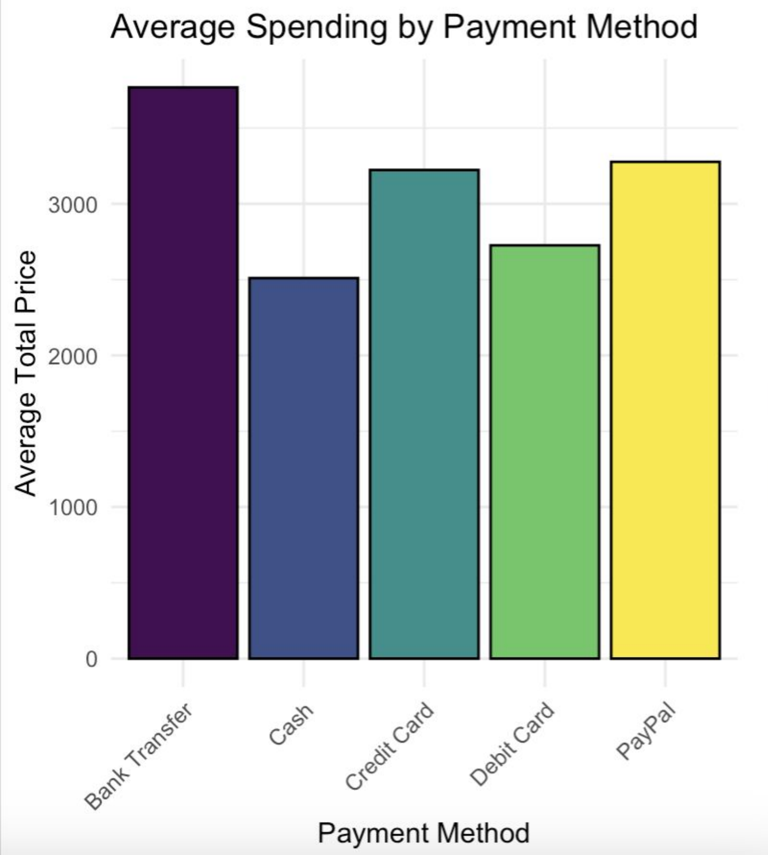


Visualization #5 - Which payment methods are associated with higher spending?

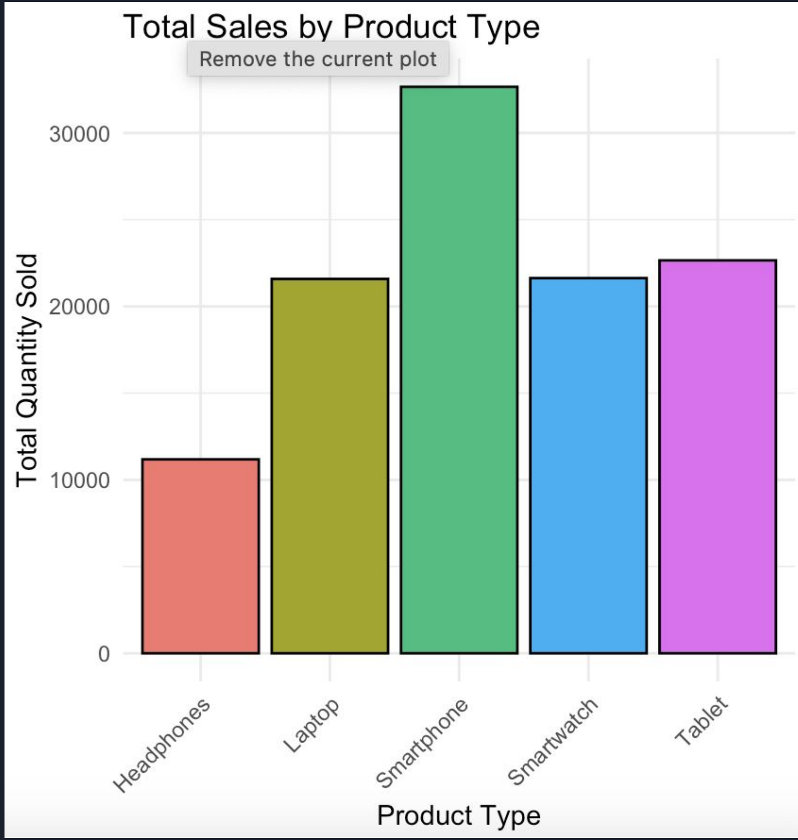
- First trial of making this bar chart, PayPal was two columns (PayPal and Paypal)
- Had to combine the two PayPal columns together (4th step of cleaning)

```
# Combine the two PayPal columns together
cleaned$Payment.Method <- gsub("Paypal", "PayPal", cleaned$Payment.Method)

# Verify the changes
table(cleaned$Payment.Method)
```

```
# Calculate the average spending amount by payment method
avg_spending_by_payment <- aggregate(Total.Price ~ Payment.Method, cleaned, mean)
```



Visualization #6 - Which product types are most popular?

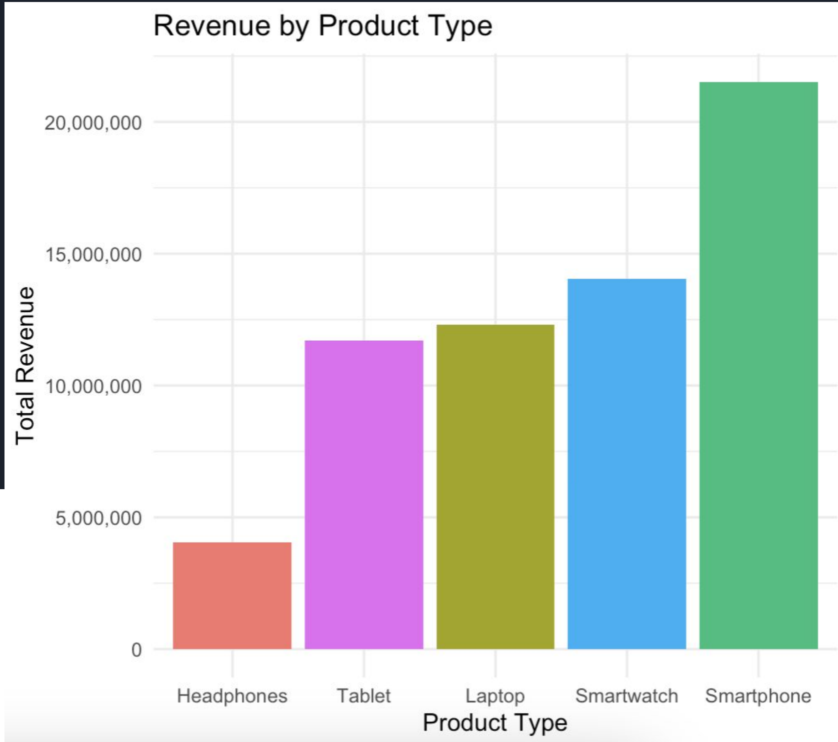


- The electronics store's customers has the highest demand for smartphones
- Demand for headphones makes up very little of the electronics store's overall demand (aggregate demand of all 5 products)
    - Is it worth keeping headphones as a product type when considering holding and selling costs associated with it?

Visualization #7 - Which Product Type Generates the Most Revenue?

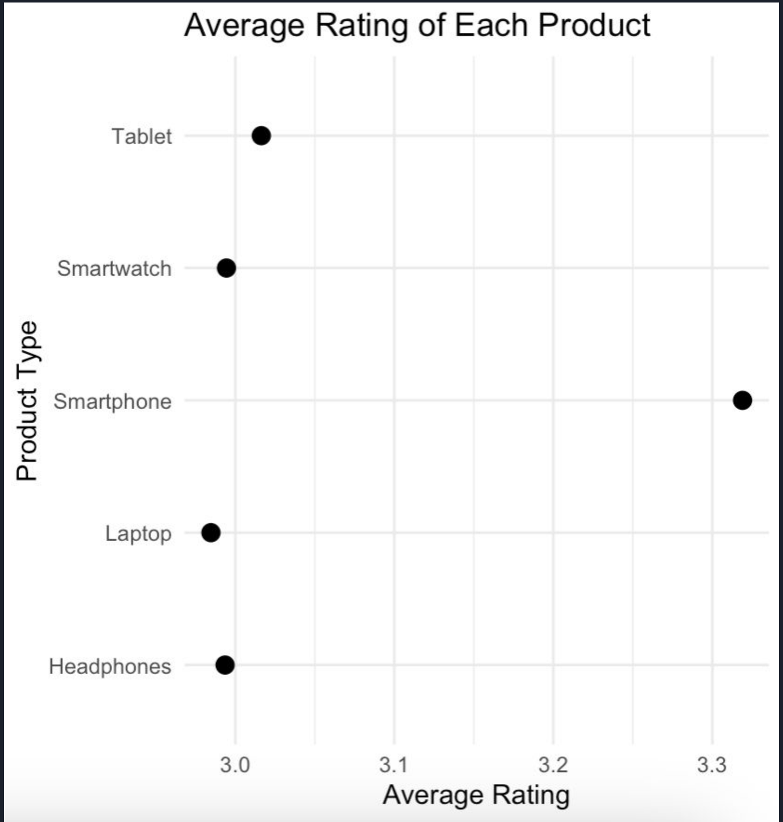
- Smartphone: SKU1001, SKU1004, SMP234
- Tablet: SKU1002, TBL345
- Smartwatch: SKU1003, SWT567
- Laptop: SKU1005, LTP123
- Headphones: HDP456

- Same product types have different SKUs
- Needed to group product type by SKU
- Most revenue derived from Smartphone sales; reflects Smartphones being in highest demand




```r
# Merge the dataset with the lookup table
data_with_product_type <- cleaned %>%
  left_join(sku_lookup, by = "SKU")

# Group by Product_Type and sum total revenue
revenue_by_product_type <- data_with_product_type %>%
  group_by(Product_Type) %>%
  summarise(Total_Revenue = sum(Total.Price, na.rm = TRUE)) %>%
  arrange(desc(Total_Revenue))
```

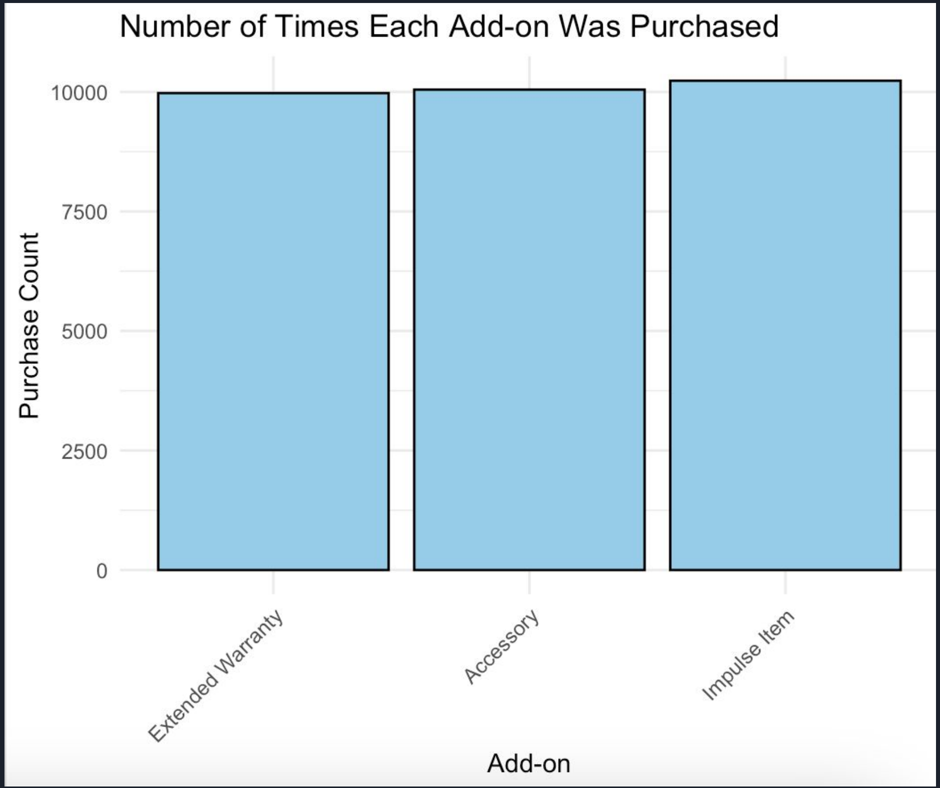
Visualization #8 - What is The Average Rating of Each Product?



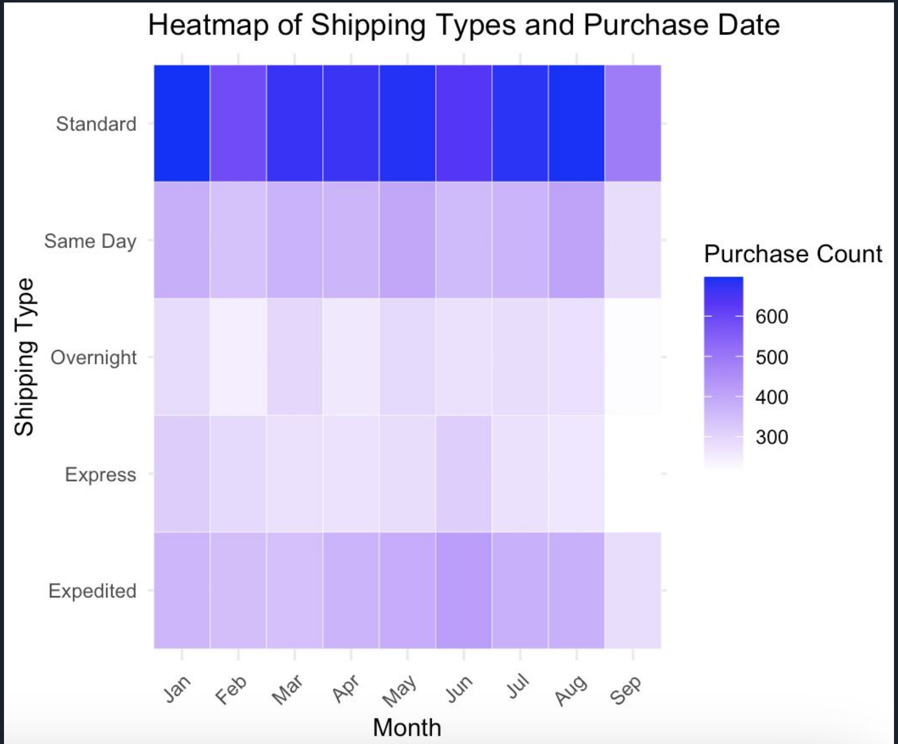
- The scale of the x-axis makes the rating for Smartphone look higher than it really is
- In general, rating for all products offered by the electronics store is extremely low
- It would be helpful to understand the customer’s reasoning for their ratings to better understand what their customers want in a product

Visualization #9 - What Add-Ons Were Most Frequently Purchased?

- Some add-on were purchase multiple times in one transaction
- While cleaning, had to properly group the add-on options
    - First bar chart had each add-on option twice on the x-axis
- Seems that each items sells the same amount of times (uniform distribution)
- It is worth it for the company to keep these three add on options as they all sell well

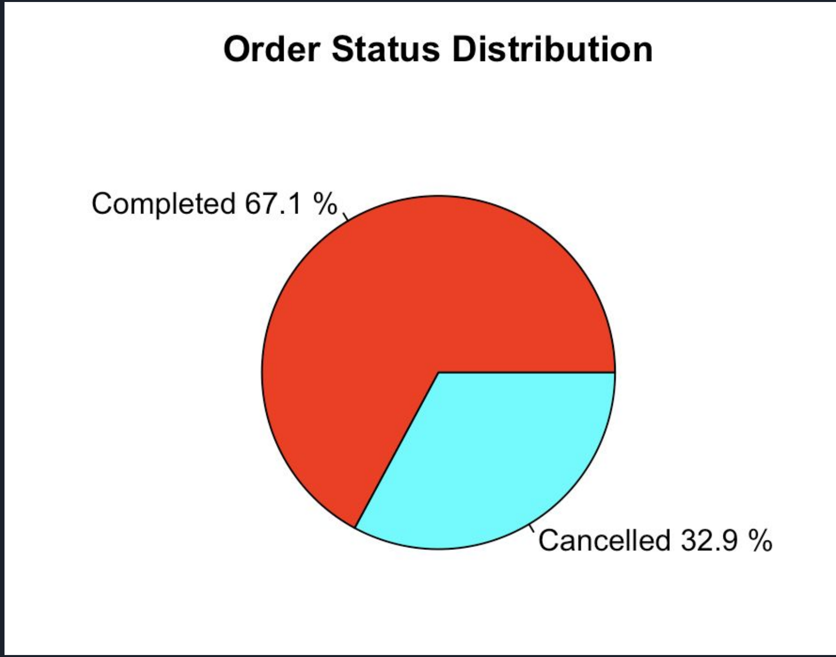


Visualization #10 - Is There a Correlation Between Different Shipping Types and The Month of Purchase?



- Only shows data for months in 2024
- Most people opt for Standard shipping
- We see expedited shipping peak in the month of June
- Not much relationship can be seen between shipping type options and month of purchase

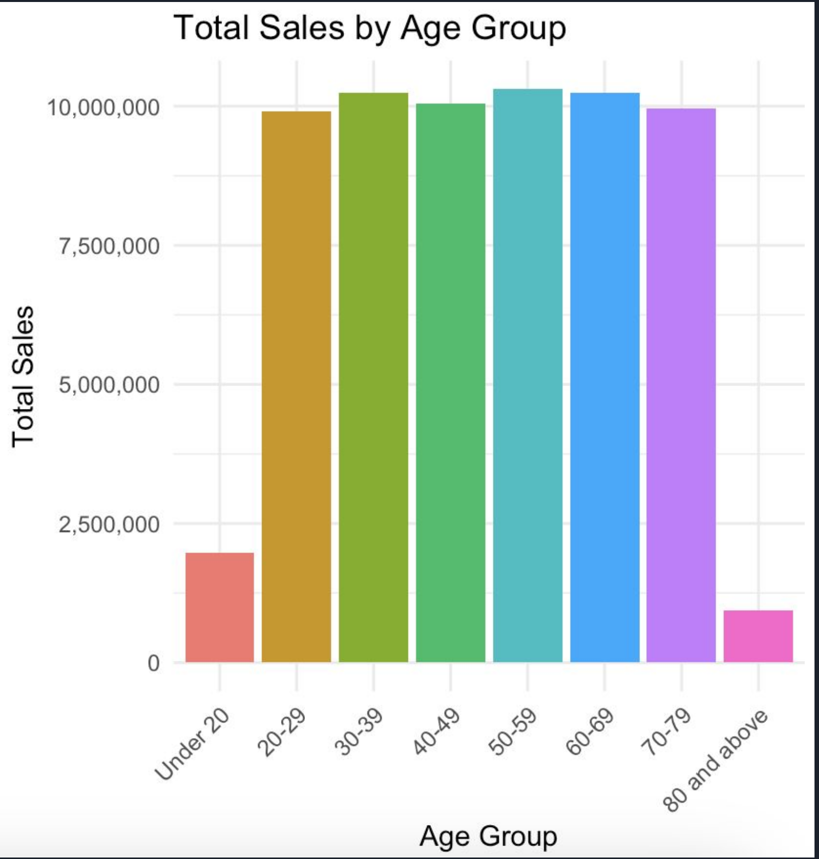
Visualization #11 - What is the Distribution of Completed and Cancelled Orders?



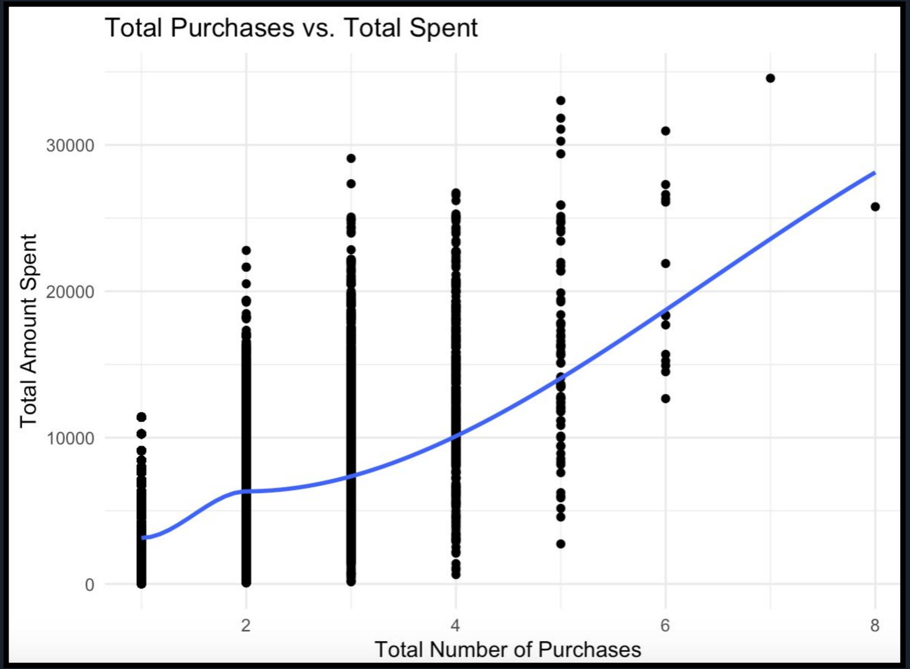
- Lots of loss revenue from cancelled orders
- Would be helpful for the store to understand the reasoning for cancellation in order to improve

Visualization #12 - How Does Total Sales Differentiate Among Age Groups?

- Spending peaks in the age group 30 -39
- Overall, pretty uniform spending across age groups (leaving out 'Under 20' and '80 and Above' - this is not surprising)



Visualization #13 - What is the Correlation Between Total Number of Purchases and Total Amount Spent?



- We see a positive correlation between total number of purchases and total amount spent by an individual customer
- We would expected this as the more purchases someone makes, the more they spend

Query #1 - What are the Demographics of the Customers?

- 10,164 Males
- 9,835 Females
- 4,342 Loyalty Members
- 15,657 Non-Members

*Know your demographic for possible correlations

```
##Query1##
#Count number of males and females from Gender in the dataset
Query1 <- "SELECT COUNT(*) AS Male_Count
          FROM cleaned
          WHERE Gender = 'Male';"
sqldf(Query1)
#Male_Count 10,164

Query1.1 <- "SELECT COUNT(*) AS Female_Count
          FROM cleaned
          WHERE Gender = 'Female';"
sqldf(Query1.1)
#Female_Count 9,835

Query1.2 <- "SELECT COUNT(*) AS Loyalty_Member
            FROM cleaned
            WHERE `Loyalty.Member` = 'Yes';"
sqldf(Query1.2)
#Loyalty_Member 4,342

Query1.3 <- "SELECT COUNT(*) AS Nonloyalty_Member
            FROM cleaned
            WHERE `Loyalty.Member` = 'No';"
sqldf(Query1.3)
#Nonloyalty_Member 15,657
```

Query #2 - Which Months had the Most Total Sales?

```
#Total monthly sales sorted from highest to lowest
Query2 <- "SELECT Month, SUM(`Total.Price`) AS Total_Sales
          FROM cleaned
          GROUP BY Month
          ORDER BY Total_Sales DESC;"
```

- May 2024 recorded the highest sales for the month
- September 2023 recorded the lowest
- Surprising to see sales so low in Dec 2023

*What months to have promotions.

	Month	Total_Sales
1	May 2024	6709042.9
2	Aug 2024	6706118.6
3	Jun 2024	6668633.6
4	Jan 2024	6619498.2
5	Jul 2024	6535129.5
6	Apr 2024	6418253.6
7	Mar 2024	6324367.8
8	Feb 2024	5733696.1
9	Sep 2024	5037691.1
10	Oct 2023	2318466.4
11	Nov 2023	2068434.1
12	Dec 2023	1980700.3
13	Sep 2023	481961.8

Query #3 - What Was the Top 5 Total Spending Amounts Among Loyalty Members?

```
# Find top 5 transactions for loyalty members only
Query3 <- "SELECT c.`Customer.ID`, c.Age, c.Gender, SUM(t.`Total.Price`) AS Total_Spending
FROM customer_table c
JOIN transaction_table t
ON c.`Customer.ID` = t.`Customer.ID`
WHERE c.`Loyalty.Member` = 'Yes'
GROUP BY c.`Customer.ID`, c.Age, c.Gender
ORDER BY Total_Spending DESC
LIMIT 5;"
```

- We were hoping to find an age pattern
- No age or gender correlation with total spending

	Customer.ID	Age	Gender	Total_Spending
1	2447	40	Female	106464.52
2	12276	52	Male	92883.54
3	11101	25	Male	72068.16
4	13823	33	Male	67851.84
5	12616	25	Male	65698.74

Query #4 - What Is The Most Popular Product Among Loyalty Versus Non-Loyalty Members?

```
Query4 <- "SELECT t.`Product.Type`, SUM(t.'Quantity') AS Total_Quantity
          FROM customer_table c
          JOIN transaction_table t
          ON c.`Customer.ID` = t.`Customer.ID`
          WHERE c.`Loyalty.Member` = 'Yes'
          GROUP BY t.`Product.Type`
          ORDER BY Total_Quantity;"

sqldf(Query4)
```

```
sqldf(Query4)
```

Product.Type	Total_Quantity
Headphones	5656
Laptop	10029
Smartwatch	10158
Tablet	10704
Smartphone	14227

Loyalty Members

- First 2 items were the same until smartwatch/laptop

*See what loyalty members want promotions on, sweepstakes.

```
sqldf(Query4.1)
```

Product.Type	Total_Quantity
Headphones	18078
Smartwatch	35136
Laptop	35732
Tablet	36346
Smartphone	54182

Non-Loyalty Members

Query #5 - What is The Average Rating of Each Product Type?

```
Query5 <- "SELECT t.`Product.Type`, AVG(t.`Rating`) AS Average_Rating
          FROM customer_table c
          JOIN transaction_table t
          ON c.`Customer.ID` = t.`Customer.ID`
          WHERE c.`Loyalty.Member` = 'Yes'
          GROUP BY t.`Product.Type`
          ORDER BY Average_Rating DESC;"
sqldf(Query5)
```

Product.Type	Average_Rating
Smartphone	3.347612
Smartwatch	3.023497
Tablet	3.015682
Headphones	2.953629
Laptop	2.953261

Loyalty Member

- Smartphones were the highest rated item.

*What products to keep on shelves.

Product.Type	Average_Rating
Smartphone	3.301124
Tablet	3.000302
Smartwatch	2.985719
Laptop	2.978610
Headphones	2.978402

Non-Loyalty Member

Query #6 - How Many Times Was Each Add-On Purchased?

Number of types an ‘Add-On’ was purchased by a customer

- Had to cleaned our data an additional time as when we first ran the query, the types of add-ons were not grouping together correctly

```
transaction_table <- transaction_table %>%
  separate_rows(Add.ons.Purchased, sep = ",") %>%
  mutate(Add.ons.Purchased = trimws(Add.ons.Purchased)) %>%
  filter(Add.ons.Purchased != "")
```

Add_on	Purchase_Count
Impulse Item	10234
Accessory	10048
Extended Warranty	9975

Query #7 - How Many Times Were Each Shipping Type Chosen?

```
Query7 <- "SELECT t.`Shipping.Type`, COUNT(*) AS Shipping_Count
           FROM transaction_table t
           JOIN customer_table c
           ON t.`Customer.ID` = c.`Customer.ID`
           WHERE c.`Loyalty.Member` = 'Yes'
           GROUP BY t.`Shipping.Type`
           ORDER BY Shipping_Count DESC;"
```

Shipping.Type	Shipping_Count
Standard	3178
Express	1527
Expedited	1517
Same Day	1511
Overnight	1480

Loyalty Member

- Standard was the highest average for both
- Overnight/Same Day was more common with Non-members

*Determine shipping availability

Shipping.Type	Shipping_Count
Standard	10927
Overnight	5541
Same Day	5445
Express	5378
Expedited	5371

Non-Loyalty Member

Query #8 - What Product Purchased Was Cancelled the Most?

Product.Type	Number_of_Cancelled_Orders	Total_Cancelled_Value
Smartphone	1974	7108919
Smartwatch	1298	4637682
Tablet	1359	3989368
Laptop	1287	3930335
Headphones	650	1306749

- Smartphone, most cancelled, most purchased(68,409).
- Tablet's (47,050) second most purchased second most cancelled
- Smartwatches (45,294) fourth most purchased, fourth cancelled
- Laptop(45,761) third most purchased, third cancelled
- Headphones(23,734)5th most purchased, least cancelled

*Important to determine QA issues or unpopular products.

Query #9 - How Does Total Spending Differ By Age Group?

1. **Categorizing Data**: Groups people into age groups
2. **Summarizing Data**: It adds total sales for each age group
3. **Connecting Tables**: It combines two tables
4. **Grouping Results**: It organizes the data by age group
5. **Sorting Results**: It orders the groups Highest to lowest sales

*Important to know your age demographic for marketing purposes.

```
Query9 <- "SELECT CASE WHEN Age < 20 THEN 'Under 20'
          WHEN Age BETWEEN 20 AND 29 THEN '20-29'
          WHEN Age BETWEEN 30 AND 39 THEN '30-39'
          WHEN Age BETWEEN 40 AND 49 THEN '40-49'
          WHEN Age BETWEEN 50 AND 59 THEN '50-59'
          WHEN Age BETWEEN 60 AND 69 THEN '60-69'
          WHEN Age BETWEEN 70 AND 79 THEN '70-79'
          ELSE '80 and above'
          END AS Age_Group, SUM(t.`Total.Price`) AS Total_Sales
          FROM customer_table c
          JOIN transaction_table t
          ON c.`Customer.ID` = t.`Customer.ID`
          GROUP BY Age_Group
          ORDER BY Total_Sales DESC;"
```

Age_Group	Total_Sales
30-39	21921256
60-69	21785938
50-59	21443460
40-49	21418777
20-29	21315935
70-79	20220023
Under 20	4135229
80 and above	1705723

Query #10 - How Does Total Spending Differ Between Loyalty and Non-Loyalty Members?

- $75,562,553 difference

Loyalty.Member	Total_Sales
No	104744447
Yes	29201894

*This query can be important to incentivise workers to offer Loyalty Memberships.

Query #11 - What Are the Top Sales Divided By Gender and Product Type?

Gender	Product.Type	Total_Sales
Female	Smartphone	22701107
Male	Smartphone	22685504
Male	Smartwatch	14761039
Female	Smartwatch	14541530
Female	Laptop	13175207
Male	Laptop	12998128
Male	Tablet	12555808
Female	Tablet	11951124
Male	Headphones	4502423
Female	Headphones	4074472

Males bought more
- Smartwatches (219,509)
- Tablets (604,684)
- Headphones (427,951)

Females bought more
- Smartphones (15,603)
- Laptops (177,079)

* This can be important if you want to find out what gender to target in marketing strategy.

Query #12 - Who Were the Top 10 Spending Customers Who Made More Than One Purchase?

Customer.ID	Number_of_Purchases	Total_Spending
16357	7	34563.70
16863	5	33035.92
13813	5	31830.16
11476	5	31077.61
12276	6	30961.18
13635	5	30260.36
12749	5	29394.56
15399	3	29084.88
12319	3	27352.32
19996	6	27296.78

- VIP customers.
- Expected high purchase totals

*This query can be important to decide who to offer discounts or rewards to

Conclusions

- No correlation found between other variables when attempting other graphs due to data used for project(despite Visualization #13)
- We were expecting for spending to be greater among men over women as its an electronic company
- The store could create more incentives to get customers to join loyalty program; could help with customer retention
- Lessons learned: cleaning data is a lot more complex than just deleting N/As

```
> cor(cleaned$Age, cleaned$Total.Price)
[1] 0.003036134
```