Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

Executive Summary

This report provides an analysis and evaluation of 20,000 sales transaction records for an electronics company over a one-year period, spanning from September 2023 to September 2024. This pooled cross-sectional data was found and pulled from Kaggle, with a total of 16 fields included in the data. The methods of analysis include exploratory analysis of the data by answering questions of customer demographics distribution, purchasing behaviors, and purchasing trends. Analysis of the data includes overseeing the data distribution, variable correlation, and reasoning for these discoveries using the ggplot2 package in R. Further, the analysis also includes solving data queries that provide further insight into some patterns that exist in the data, through the use of the sqldf package. All coding can be found in the R code uploaded to the group repository.

The key variables of interest to note are Gender, Purchase Date, Total Price, and Loyalty Member as they make up a large amount of the analysis within our project. The original data was organized into one big dataset, 'Electronic_sales_Sep2023-Sep2024,' shortened to 'final' in pre-processing for simplicity reasons. In preparation for analysis, the raw data was split into two datasets: 'customer_table' and 'transaction_table' joined at the primary key 'Customer.ID' in the 'customer_table' and the foreign key 'Customer.ID' in the 'transaction_table.' The 'customer_table' was created from the 'final' to include all the variables regarding customer information including 'Customer.ID,' 'Age,' 'Gender,' and 'Loyalty.Member.' The 'transaction_table' was also created from the 'final' to include all the variables regarding transactional information including 'Customer.ID,' 'Product.Type,' 'SKU,' 'Rating,' 'Order.Status,' 'Payment.Method,' 'Total.Price,' 'Unit.Price,' 'Quantity,' 'Purchase.Date,' 'Shipping.Type,' 'Add.ons.Purchased,' and 'Add.on.Total.' The results of the data analyzed show little to no correlation between Gender and Total Price, Total Price and Loyalty Member status, and Purchase Date and Shipping Type.

Cleaning The Data

The initial step in data cleaning involved handling missing values. Only one observation had a missing value in the 'Gender' field, which was removed. Removing just one observation of 20,000 would not have any significant impact on any future analysis, which is why the decision to completely delete the observation was made. In order to remove the singular record that included an N/A value in it, we used the code (final_filtered <- final %>% filter(Gender != "#N/A")). To differentiate the 'final' data from the data with now only 19,999, it was named 'cleaned.'

For the next step, the 'Purchase Date' column was converted to a date format using the lubridate package, and a new 'Month' column was created to facilitate date-based analyses, such as line graphs. Prior to this step, date-based analysis could not be conducted. The code used to

facilitate this step in the cleaning process was, cleaned$Purchase.Date <- ymd(cleaned$Purchase.Date). cleaned$Month <- floor_date(cleaned$Purchase.Date, "month").

Additionally, inconsistencies in the 'Payment Method' column were resolved by combining entries for 'PayPal' and 'Paypal' to be recognized as the same type of payment method. Prior to this part of the cleaning process, ‘PayPal’ and ‘Paypal’ were recognized by R as two different types of payment. To enact this part of the cleaning process, cleaned$Payment.Method <- gsub("Paypal", "PayPal", cleaned$Payment.Method) was used.

A similar issue arised within the “Add-Ons” field of our data. The three different types of add-ons offered, impulse item, accessory, and extended warranty, were not grouped together properly for analysis. The code to resolve this issue was transaction_table <- transaction_table %>% separate_rows(Add.ons.Purchased, sep = ",") %>% mutate(Add.ons.Purchased = trimws(Add.ons.Purchased)) %>% filter(Add.ons.Purchased != "").

Finally, the newly cleaned data was then split into two tables, one for customer information and another for transaction information, as mentioned in the Executive Summary section of this report.

Analysis - Queries

Query 1: What are the Demographics of the Customers?

```
sqldf(Query1)
Male_Count
     10164
sqldf(Query1.1)
Female_Count
        9835
```

```
sqldf(Query1.2)
Loyalty_Member
          4342
sqldf(Query1.3)
Nonloyalty_Member
           15657
```

The exploratory analysis began with a query to help understand the demographics of the electronics company's customers. We simply wanted to count the number of observations where the gender of the customer was male or female, as well as whether they were a loyalty member or not. We noticed an almost uniform distribution between the number of males and females among our records. We also see that there is a significant difference between the loyalty and non-loyalty membership. This information can be helpful for the electronics company as it signifies to them that they need to advertise their loyalty program to their customers to encourage better customer retention and loyalty to the company.

Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

Query 2: Which Months had the Most Total Sales?

	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 2 looks at Total Sales grouped by month and year in descending order. This query was able to be executed only after cleaning the raw data with the lubridate package. Instead of selecting the 'Purchase.Date' column of the raw data, 'Month' was selected from the 'cleaned' data. From our results, it was extremely surprising to see the total sales for Dec 2023 to be so low on the list considering that December should be a high sales month for retail stores due to the holiday season.

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

	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

Based on Query 2, we thought it would be interesting to rank loyalty members by greatest total spending across all purchases they may have made. Our results showed that a female aged 40 years old was the electronics company's top spending customer. Following her were four males in the age range of 25 to 52. Although we were looking to find some correlation between our variables, we see no correlation between age and gender when considering total spending in this query.

Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

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

```
sqldf(Query4)
Product.Type Total_Quantity
  Headphones           5656
      Laptop          10029
  Smartwatch          10158
      Tablet          10704
  Smartphone          14227
```

Loyalty Members

```
sqldf(Query4.1)
Product.Type Total_Quantity
  Headphones          18078
  Smartwatch          35136
      Laptop          35732
      Tablet          36346
  Smartphone          54182
```

Non-Loyalty Members

Continuing, we wanted to look into the most popular products among loyalty members versus non-loyalty members. For both loyalty and non-loyalty members, the most popular product was smartphones. We also found that tablets were the second most popular product for both. This can signify to the electronics company that it is important to keep extra stock of the most popular products to ensure customers are not being turned away due to stock outs, which ultimately would lead to loss of revenue.

Query 5: What is The Average Rating of Each Product Type Among Loyalty Versus Non-Loyalty Members?

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

Loyalty Members

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

Non-Loyalty Members

Query 5 provides insight into the average rating of each product type for both loyalty and non-loyalty members. We see that for both categories, smartphones have the highest rating. This can partially explain our findings in Query 4 with smartphones being the most popular product, regardless of membership status. Furthermore, since all these ratings are out of five stars, we can conclude that although smartphones have the highest ratings these ratings are all very low. The electronics store should consider looking into the quality of their products, which products they want to keep selling, and how to satisfy their customer's needs with the products they sell.

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

Add_on	Purchase_Count
Impulse Item	10234
Accessory	10048
Extended Warranty	9975

Moving onto Query 6, we wanted to see the exact purchase amount for each add-on type. To do this, we needed to remove extra spaces in the 'Add.Ons.Purchased' column, using 'trimws' to properly group the 3 types of add ons. Prior to this step of our data cleaning, the types of add-ons were not grouping together correctly, and our query results showed we had 6 types of add-ons. The results showed that impulse items, accessories, and extended warranties were all purchased in almost a uniform distribution. The electronics store should keep all of the add-on types they offer as they all sell well.

Query 7: How Many Times Were Each Shipping Type Chosen Among Loyalty Versus Non-Loyalty Members?

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

Loyalty Members

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

Non-Loyalty Members

Query 7 looks into how many times each shipping type was chosen among loyalty and non-loyalty members. The main pattern we see here is that standard shipping is most common for both. It also shows that for non-loyalty members overnight and same day was more common. Shipping varies per customer's preferences and needs, so it is important to have a few options so that they are satisfied with how fast they are receiving their products.

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

Query 8 ranks products that were cancelled, or returned the most. Here, we see that smartphones were the most cancelled, which can relate to the fact that they were purchased the most. With the total number of cancelled orders, grouped by product type, we also see the total monetary value of each cancelled product. Totalling these numbers, this is a significant amount of loss sales for the company. It could be helpful for the store to understand the customer's reasoning for cancelling their orders so they can uncover what needs to be changed about their products to satisfy the customer's expectations and wants.

Query 9: How Does Total Spending Differ By Age Group?

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 9 results show how total spending differs by age group. To run this query, we first categorized our data by grouping customers into age groups. We see that total spending is close to uniform across all age groups, despite 'Under 20' and '80 and above.' This is not a surprise as electronics are most geared towards these other age groups.

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

Loyalty.Member	Total_Sales
No	104744447
Yes	29201894

Following the total sales theme from Query 9, we wanted to also understand how total spending differs between loyalty and non-loyalty members. Results showed a $75,562,553 difference between non-loyalty and loyalty members. While this number appears to be significantly high, it must be taken into consideration that the distribution between loyalty and non-loyalty members is skewed. This means that it is very difficult to make any correlation between loyalty member status and total sales.

Query 11: What Are the Top Total 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

Query 11 began to look at which product generated the most amount of total sales, categorized by male and female. Our initial thought with this query was hoping to see some type of correlation between gender and sales. However, this was proven otherwise with this query. There seems to be zero correlation between these variables as total sales by product type among males and females are pretty close in amount.

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

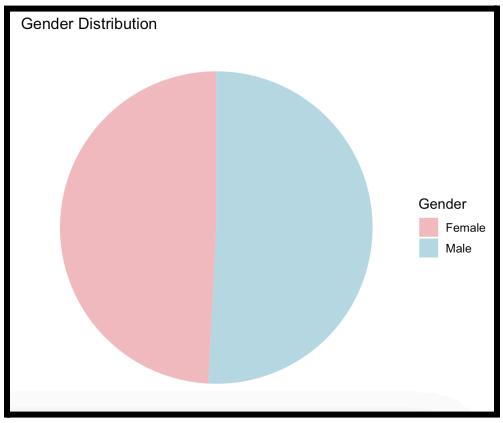
For the last Query 12, we wanted to examine customer retention. In this query, we looked at all records where a customer made more than one purchase. We then ranked the total spending of all purchases made in descending order. We limited the result to just the top 10, as there were 5,499 customers who made more than one purchase. Considering the 19,999 observations from our 'cleaned' data, this signifies that 27.5% of customers make more than one purchase at the electronics store. Customer loyalty is key to the success of the store. The main correlation takeaway from this query is that as the number of purchases increase, total spending also increases. This is logical since the more purchases someone makes, the most they spend.

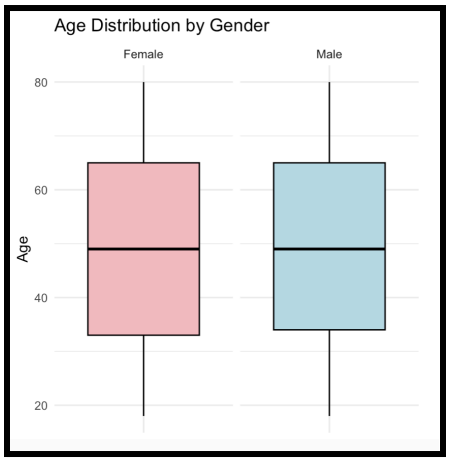
Analysis - Visualizations

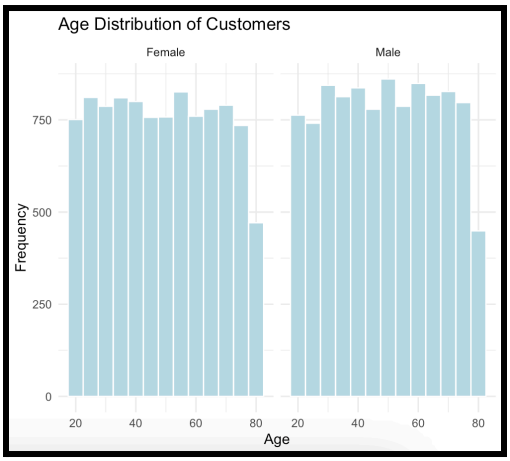
Several visualizations were created using ggplot2 to complement some of our queries from the above section.

Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

Visualization #1 - Who Are The Customers in Terms of Demographics?

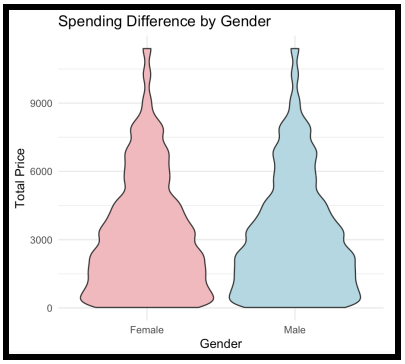






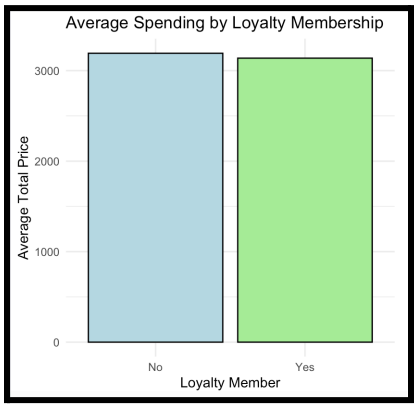
Relating to Query 1, we wanted to understand and analyze age, gender, and loyalty membership distribution of our cleaned data. To do this, we create a number of different visualizations including bar charts, box plots, and pie charts. The data was distributed pretty uniformly among males and females, as well as age. However, this was not the same for the distribution of loyalty membership status as results showed most people purchasing were not loyalty members.

Visualization #2 - How Does Total Spending Differ By Gender?



Using geom_violin(), we created a violin plot showing the distribution of total amount spent, facet wrapped by gender. Similar to the gender and age distribution of our data, the total amount spent categorized by gender also was pretty uniformly distributed. We found this result to be shocking as we originally hypothesized to see the correlation of males spending more on electronics than females, considering that the electronics industry is typically a male dominated one.

Visualization #3 - Do Loyalty Members Spend More On Average Than Non-Loyalty Members?

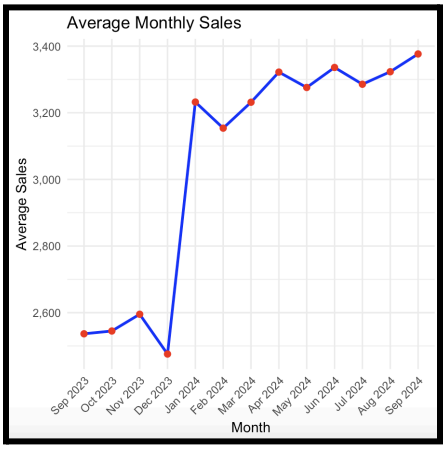


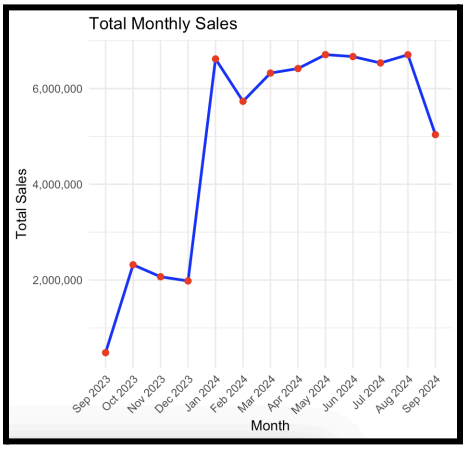

```
#Calculate average amount spent grouped by loyalty member or not
cleaned %>%
  select(Loyalty.Member, Total.Price) %>%
  filter(Loyalty.Member == "Yes") %>%
  summarize(member_average_purchases = mean(Total.Price))
#3138.011

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

Used our cleaned data, our 4th visualization reveals average spend by loyalty membership status. To our surprise, the distribution appears to be very uniform, with non-loyalty members spending just $53.96 more on average. It was an even bigger surprise to see non-loyalty members spending more on average as a loyalty program is supposed to incentivize members to spend more through rewards and exclusive deals. We were expecting to see a slight correlation between loyalty membership status and average spending, yet we were proven otherwise.

Visualization #4 - Are There Monthly Trends In Sales?

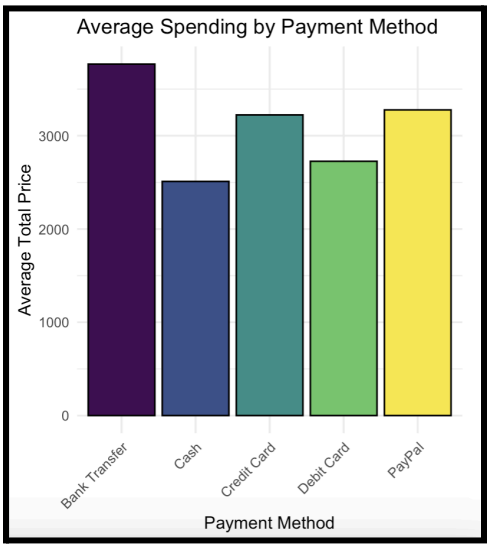




Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

Visualization 4 is closely related to Query 2. This visualization could not have been created without the use of the lubridate package in RStudio to clean our raw data. We wanted to show monthly trends in total sales by both average monthly sales and total monthly sales to offer a different understanding and perspective from both line plots. For example, between the months of Aug 2024 and Sept 2024, we see average sales increasing, yet total sales decreasing. This could signify to the store that customers are spending more on average, yet less customers are making purchases. This relates back to customer retention and the importance of strong customer loyalty programs to continue to draw customers back to make more purchases.

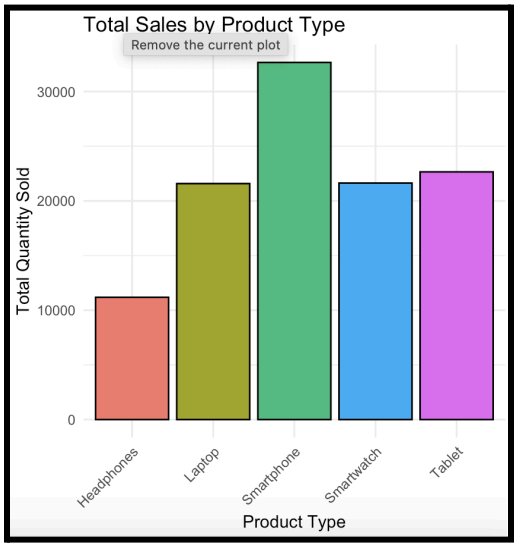
Visualization #5 - Which Payment Methods Are Associated With Higher Spending?



The bar chart visualization was created to visualize the average amount spent by different payment methods used by customers. Before step 3 of our cleaning process, the first bar chart created had 6 different types of payment methods on the x-axis, the 6th type being 'Paypal.' Due to the different spellings of 'PayPal' in the original data, R recognized this as two different types of payment methods, when it really should just be one. Thus, after this step in cleaning, we created this bar chart with just 5 payment methods on the x-axis. We want to create this visualization as different payment methods can influence how a customer perceives the cost of the purchase. Our visualization shows bank transfers relating to the highest average spending with cash being the lowest. This could potentially be explained by the fact that people tend to spend more using credit cards or bank transfers compared to cash. This is based on the fact that the act of handing over cash creates a stronger sense of money leaving a customer's pocket which can incentivize them to spend less on average.

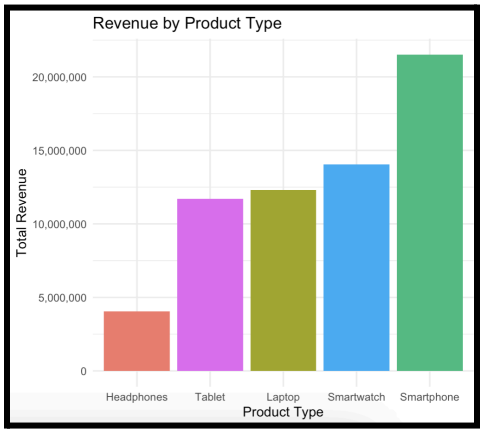
Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

Visualization #6 - Which Product Types Are Most Popular?



In this visualization, it shows which product type was the most popular based on total quantity sold. We see that the electronics store's customers' highest demand is for smartphones. We can also see from this visualization that the total quantity sold of headphones makes up a very small portion of the aggregate quantity sold across all product types.To maximize profits, the store may reconsider if having headphones for sale is worth the holding and selling costs associated with it.

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



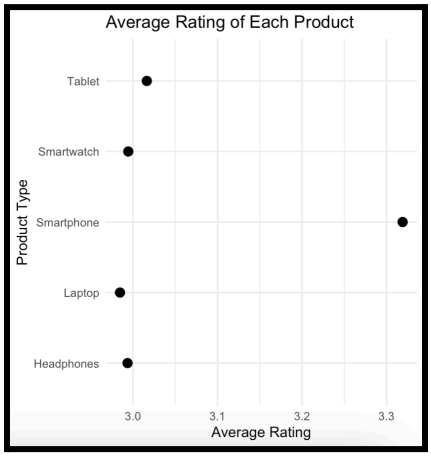

```
# Merge the dataset with the lookup table
data_with_product_type <- cleaned %>%
  left_join(sku_lookup, by = "SKU")

# 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))
```

Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

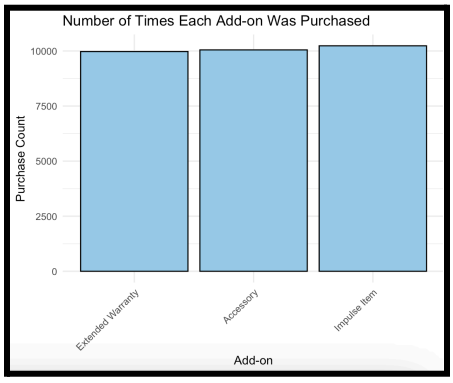
To find which product type generates the most revenue, we merged the dataset with the SKU lookup table, using the left_join() function. The SKU lookup table was created because the same product had multiple different SKU numbers. After merging these tables together, we grouped our results by product type and summed the total revenue. This visualization can be related back to visualization 6 since the demand for smartphones was highest, thus explaining the revenue for smartphones to also be the highest. Since headphones were lowest in demand we can also connect them to being the least amount of revenue considering all products.

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



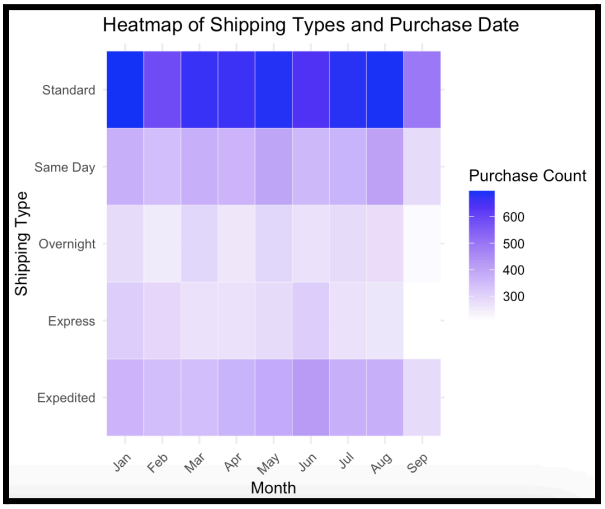
The dot plot visualization was created to visualize and compare the average rating for each product type. We clearly see smartphones having the highest average rating of 5 stars compared to the other products. Without considering the scale of the x-axis, this can make the rating for smartphones appear significantly higher compared to the other products. However, when you consider the scale of the x-axis, the rating for smartphones is only .3 greater than the rating for the 2nd top rated product, tablet. In general, all of the ratings for all products were extremely low. It would be helpful for the electronic store to understand customers' reasoning for how they rated the product to better understand what they are looking for in a product.

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



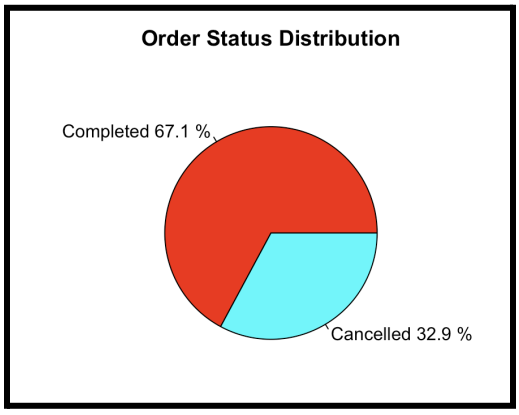
Directly related to Query 6, this bar chart visually shows the frequency of which different add-ons were purchased. While cleaning, we had to properly group the add-on options into three groups, extended warranty, accessories, and impulse item. Similar to the same issue that occurred in Query 6, the first bar chart we created had each add-on option twice on the x-axis. Results then pointed to a uniform distribution. Since all were purchased similarly, the electronic store should consider keeping all of these options.

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



Related to Query 7, we created a heatmap to show shipping type and purchase date for all products only in the month of 2024. Since 2023 had too much missing data, we decided to include data for just months in 2024. Results showed that most people opted for standard shipping for all months. Therefore, not much relationship can be seen between shipping type options and month of purchase.

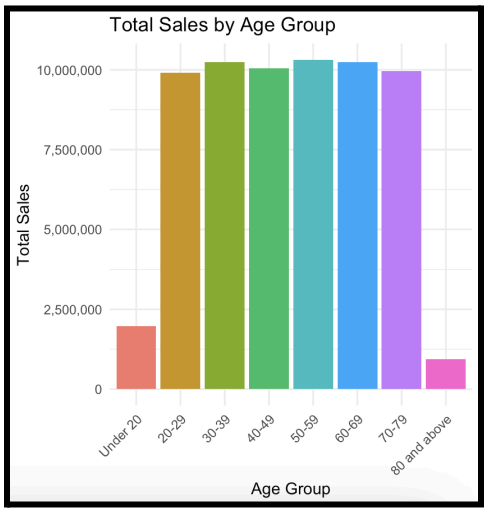
Visualization #11 - What is the Distribution of Completed and Cancelled Orders?



Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

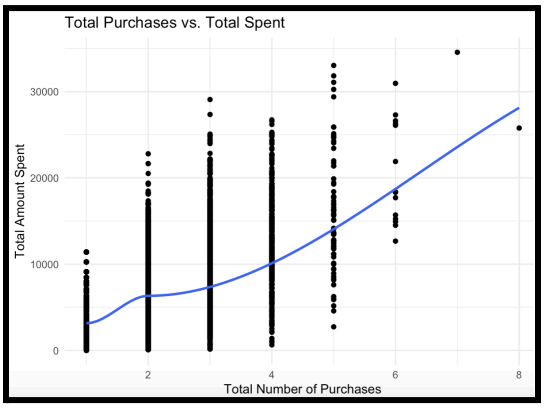
To visualize order status distribution between completed and cancelled orders, we opted for a pie chart. 67.1% of orders were completed whereas 32.9% were canceled. To better understand why the rate of cancellations is significantly high it would be beneficial for the electronic store to get consumer feedback to avoid further loss of revenue, especially since canceled orders make up almost ⅓ of all order statuses.

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



This bar chart visualization was created to visualize how total sales differed among different age groups. Relating to Query 9, peak spending for total sales was in the age group 30-39. Overall results show a uniform distribution, despite age groups ‘Under 20’ and ‘80 and above. This uniform distribution can be largely explained by the uniform distribution of age in our data set. Since the average amount spent among loyalty and non-loyalty customers and male and female is almost the same, it makes sense that total sales by age group is also primarily uniform.

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



Data-Mining Project
BUAN 314/370
Dec 15th, 2024
Sophia Saenger, Taylor Piecukonis, and Nicolas Corona

Our 13th and final visualization is a scatter plot with a smoother line of total purchases on the x-axis and total amount spent on the y-axis. We can see from this visualization a positive correlation between the number of purchases and the total amount spent by an individual customer through this positively sloping smoother line. This result was expected as the more purchases a customer makes, the more they are spending.

Conclusion

Throughout our analysis, we attempted to discover correlations between fields of our data. When running the code, cor(cleaned$Age, cleaned$Total.Price), the correlation value between the fields were 0.003, indicating no correlation. No other correlations were found between other variables when attempting other graphs despite our last visualization, number 13. Contrary to our hypothesis, expenditure habits were not higher for males compared to females, considering that the company specializes in electronics, where traditionally, male spending is presumed to be higher. We also found that the total spend of December 2023 was very low which was surprising as that month traditionally should record the highest sales in accordance with holiday purchases and after-Christmas sales. These findings highlight the importance of conducting thorough research into the store's marketing and sales strategies over the past month, especially considering the underperformance we have observed.

Customer retention could be positively affected if some more inducement towards loyalty registration were designed in the stores. One of the important things we learned from this project is that cleaning data involves a lot more than simply deleting N/As. The cleaning process of raw data is multi stepped, complex, and essential, as it guarantees the reliability and precision of our analysis.