MAIB CSC 101

Exploratory Data Analysis

Supply Chain Data

Abbiegael Klara Go Chu

3-31-2023

Table of Contents

Table of Figures

1. Problem Statement

1.1.Shipping Delays vs Order Destination and Categories

During the data exploration, it was seen that more than half of deliveries were considered late. This analysis aims to see the association of shipping delays by Order Destination and by Categories.

1.2.Customer Profile

Customer Profiles are descriptions of a company's customers, detailing what the customers' preferences are. A Customer Profile sheet could assist in the company's marketing campaigns and locations targeting, as well as became a benchmark for Key Performance Indicators during analyses.

1.3.Recency, Frequency, and Monetary (RFM) Score Analysis

According to (anuragnayak, 2021), RFM means Recency, Frequency, and Monetary value that a customer provides. An RFM Analysis ranks and splits customers into different segments. The analysis will be able to provide insight on which segments are customers in and select which customers are the best customers.

1.4.Basket Analysis

Basket Analysis can identify which products are being purchased together. The company can base its marketing offers on this analysis and base product recommendations with it. Moreover, a basket analysis can aid in product placement in physical stores (Selvaraj, 2022).

1.5.Simple Linear Regression Analysis

A simple linear regression analysis finds the relationship between an independent and a dependent variable that is defined as a straight line (Jidge, 2020). A rough sales forecast can be created based on a simple linear regression analysis to aid in the company's early stages of inventory planning.

2. Dataset

The supply chain dataset was retrieved from this link:

https://www.kaggle.com/datasets/shashwatwork/dataco-smart-supply-chain-for-big-data-analysis

The dataset contains 54 features and 180,519 rows. The features of the dataset are:

'Type', 'Days for shipping (real)', 'Days for shipment (scheduled)', 'Benefit per order', 'Sales per customer', 'Delivery Status', 'Late_delivery_risk', 'Category Id', 'Category Name', 'Customer City', 'Customer Country', 'Customer Email', 'Customer Fname', 'Customer Id', 'Customer Lname', 'Customer Password', 'Customer Segment', 'Customer State', 'Customer Street', 'Customer Zipcode', 'Department Id', 'Department Name', 'Latitude', 'Longitude', 'Market', 'Order City', 'Order Country', 'Order Customer Id', 'Order Date', 'Time', 'Order Id', 'Order Item Cardprod Id', 'Order Item Discount', 'Order Item Discount Rate', 'Order Item Id', 'Order Item Product Price', 'Order Item Profit Ratio', 'Order Item Quantity', 'Sales', 'Order Item Total', 'Order Profit Per Order', 'Order Region', 'Order State', 'Order Status', 'Order Zipcode', 'Product Card Id', 'Product Category Id', 'Product Description', 'Product Image', 'Product Name', 'Product Price', 'Product Status', 'shipping date (DateOrders)', 'Shipping Mode'.

This dataset was assigned as an academic project for an Exploratory Data Analysis.

3. Data Cleaning and Feature Engineering

The following features were discarded as the data did not serve any purpose to the problem statements stated above:

- Days for shipment (scheduled)
- Benefit per order
- Sales per customer
- Late_delivery_risk
- Category Id
- Customer Email
- Customer Fname
- Customer Lname
- Customer Password
- Customer State
- Customer Street
- Customer Zipcode
- Department Id
- Department Name
- Latitude
- Longitude
- Market
- Order Customer Id
- Order Id
- Order Item Cardprod Id
- Order Item Discount
- Order Item Id
- Order Item Product Price
- Order Item Profit Ratio
- Order Region
- Order State
- Order Zipcode
- Product Card Id

- Product Category Id
- Product Description
- Product Image
- Product Price
- Product Status
- Shipping date (Date Orders)

After data cleaning, only 20 features were left. No rows were null value after data cleaning.

The top 5 category names were identified: Cleats, Men’s Footwear, Women’s Apparel, Indoor/ Outdoor Games, and Fishing. The apostrophes in Men’s Footwear and Women’s Apparel were taken out for easier data processing.

4. Shipping Delays vs Categories and Order Destination

4.1.Shipping Delays vs Categories

Based on Figure 1, Golf Bags & Carts have the highest percentage of late deliveries by 69%. The category Lacrosse follows with 60% of orders being late. No association of shipping delays was determined by Category.

4.2.Shipping Delays vs Order Destination

In the data supply chain description, order country is defined as the order's destination. Referring to Figure 2, 10 countries appear to have a hundred percent late deliveries, while 7 countries have zero percent late orders. The number of orders does not appear to have any bearing to the late delivery rate since the 10 countries and 7 countries have 10 or less orders each.

5. Customer Profile

5.1.General Customer Profile Summary

- **Customer Segmentation**

Based on Figure 3, consumers make up 51.8% of total customers, followed by Corporate at 30.4% and Home Office at 17.9%.

- **Payment Type**

According to Figure 4, most popular payment type among the company's customers is Debit, consisting of 38.4%.

- **Category Names**

Referring to Figure 5, the top 5 categories, namely Cleats, Men's Footwear, Women's Apparel, Indoor/ Outdoor Games, and Fishing, make up almost 60% of orders. Others category consists of all other product categories not included in the top 5.

- **Customer Country and City**

Customer Country and City are defined as which country and city the customer made the order in according to the Data Supply Chain Set Description file. The insights identified in this analysis can assist with geolocation marketing efforts.

Based on Figure 6, the EE. UU. makes up 62% of orders, while the other 38% comes from Puerto Rico. The orders from EE. UU. are more equally distributed compared to Puerto Rico. The top 3 customer cities in EE. UU. are Chicago, Los Angeles, and Brooklyn, each is 3% of the total EE. UU. orders. On the other hand, Caguas is the main customer city in Puerto Rico, consisting 96% of Puerto Rico's orders.

- **Order Country and City**

Order Country and City are defined as which country and city the orders' destinations are according to the Data Supply Chain Set Description file. The insights identified in this analysis can assist with warehouse location planning to decrease delivery delays.

Looking at Figure 7, the top 5 order destination countries are Estados Unidos (14%), Francia (7%), Mexico (7%), Alemania (5%), and Australia (5%). The top 3 cities in Estados Unidos are New York City, Los Angeles, and Philadelphia, making up 9%, 7%, and 5% of Estados Unidos orders respectively.

- **Shipping Mode**

Figure 8 shows that 59.7% of customers use Standard Class shipping, and only 5.4% avail the Same Day Shipping Option.

The average delivery time of the company's Standard Class shipping is 3.99 days, while Same Day shipping is 0.48 days.

- **AOV**

  The Average Order Value of the company is 203.77, excluding discounts.

- **Basket Size**

  The Average Basket Size of the company is 2.13. A Basket Analysis is appropriate to determine which products customers are buying together.

- **Average Order Item Discount Rate**

  The Average Order Item Discount Rate of the company is 10.17%.

- **Average Profit per Order**

  The Average Profit per Order of the company is 21.97.

- **Profit Margin**

  The Average Profit Margin of the company is 10.78%.

5.2. Customer Segment Comparisons

After exploring the customer data, it appears that the data has a Customer Segment feature made up of three types of customers, Consumer, Corporate, and Home Office. Further data exploration was conducted to determine if the different customer segments have unique preferences and purchasing patterns.

The dataset classified 3 types of customers in its Customer Segment column, which are Consumer, Corporate, and Home Office. Under normal circumstance, Customer Segments show different purchasing metrics, such as basket size, average order value, product categories, etc.

After investigating the dataset based on the customer segments given, it was discovered that the purchasing behavior of all three customer segments are the same, or with minimal differences in decimal figures. Figure 9, Figure 10, Figure 11, Figure 12, and Figure 13 are figures that compare customer segments by Payment Type, Shipping Mode, Categories, Customer Country, and Order Country respectively.

Moving forward, the marketing team of the company can refer to the customer profile analysis above or determine more ways to segment their customers. For example, customer demographics, like age, salary, etc., and behavioral attributes could also be collected by the company.

6. Recency, Frequency, and Monetary (RFM) Score Analysis

6.1. RFM Customer Value

Order statuses marked as 'Canceled' and 'Suspected Fraud' were removed from this analysis as an RFM Analysis is focused on determining a company's best customers and assigning segments.

The RFM Score determines and ranks the company's best customers based on their Recency, Frequency, and Monetary scores. Recency, Frequency, and Monetary do not share equal weights.

In the analysis, Monetary was given the largest weight, 57%, while Frequency has 28% and Recency makes up 15%. The weights should be dependent on the company's nature of business. However, in this scenario, since the company's nature of business is unclear, the weights were based from Geeks for Geeks's "RFM Analysis Analysis Using Python." (anuragnayak, 2021)

Once the RFM Scores have been computed, customers are assigned customer value labels for easier categorization. Figure 14 shows that Top Customers only make up 1.7% of total customers, while Lost Customers, the lowest ranking, make up 25.4%. Customer Value labels were also referenced from Geeks for Geeks's "RFM Analysis Analysis Using Python." (anuragnayak, 2021)

6.2.RFM Ranks Segmentations

Another RFM analysis was conducted, and rather than basing the scores on weights, the RFM Ranks were based on each metric's quantile value, ranging from 1 to 5, with 5 being the highest and 1 being the lowest.

Customers were segmented based on Connectif's article, "What Are RFM Scores and How To Calculate Them." The article provided guidance on which RFM Ranks corresponded to a specific client segment. The top segments were called Champion and Loyal Customers. Champions have recently purchased from the company, and they frequently make purchases and spend more money than normal customers. Loyal Customers frequently buy from the company and spend more than normal customers. And, on the other end of the spectrum, Lost Customers had the lowest scores in the recency, frequency, and monetary criteria. (Connectif, 2022)

Based on Figure 15, there are 499 Champion customers and no Lost customers. The company's customer base largely consists of Potential Loyalists, making up 10,037 of its customers. According to (Connectif, 2022), a strategy to convert Potential Loyalists into Loyal Customers is by inviting Potential Loyalists to membership/ loyalty programs.

7. Basket Analysis

7.1.Terms

According to (Selvaraj, 2022), the following Basket Analysis terms have these definitions:

- **Antecedents**
  The items added to the customer's cart first.
- **Consequents**
  The items that may be added based on what the antecedent is. It has an if-then relationship with antecedent, if antecedent was added to the cart, then there is a likelihood that the consequent will be added.
- **Support**
  The proportion where category or item X is present in all transactions.
- **Antecedents Support**

The proportion of how many transactions contain the category as the antecedent.

- **Consequents Support**

  The support for the itemset with the consequent.
- **Confidence**

  The probability of the consequent appearing if the basket contains the antecedent.
- **Lift**

  The ratio of the consequent's sale when an antecedent sale. It can be interpreted as customers are (lift's value) times more likely to purchase the consequent if the company sold the antecedent.
- **Leverage**

  The comparison between frequency of the antecedent and consequent occurring together vs the frequency of the antecedent and consequent occurring separately. A leverage closer to zero indicates higher independence.
- **Conviction**

  The numerical representation of how dependent the consequent is to the antecedent. A higher conviction means the consequent is highly dependent on the antecedent.

7.2.Basket Analysis by Category

A 0.007946 for minimum support was selected to narrow down the selection. This number is the 75th percentile of a minimum support of 0.001. The 5 most popular category combinations are:

- Men's Footwear and Cleats
- Women's Apparel and Cleats
- Inddor/Outdoor Games and Cleats
- Women's Apparel and Men's Footwear
- Fishing and Cleats

The company may use these combinations to design marketing campaigns. Figure 16 heatmap, ‘Basket Analysis by Product Category', also shows which Product Categories have an antecedent and consequent relationship.

7.3.Basket Analysis by Product

A 0.005378 for minimum support was selected to narrow down the selection. This number is the 75th percentile of a minimum support of 0.001. The 5 most popular category combinations are:

- Nike Men's CJ Elite 2 TD Football Cleat and Perfect Fitness Perfect Rip Deck
- Nike Men's Dri-FIT Victory Golf Polo and Perfect Fitness Perfect Rip Deck
- O'Brien Men's Neoprene Life Vest and Perfect Fitness Perfect Rip Deck
- Nike Men's Dri-FIT Victory Golf Polo and Nike Men's CJ Elite 2 TD Football Cleat
- O'Brien Men's Neoprene Life Vest and Nike Men's CJ Elite 2 TD Football Cleat

The company may use these combinations to design marketing campaigns. Figure 17, ‘Basket Analysis by Product Name', also shows which Product Categories have an antecedent and consequent relationship.

8. Simple Linear Regression Analysis

8.1.Sales Data

Order Statuses with Suspected Fraud and Cancelled values were not considered for this analysis as these two order statuses did not contribute to future sales.

In this analysis, the sales data of the top 5 categories and their total were grouped into monthly sales to lessen the congestion of data. Figure 18 shows the monthly sales drastically drop off towards the end of the period.

At the last 4 points of the graphs, October 2017 until January 2018, the sales of all categories dropped significantly. Cause of drop is unknown, but it might be due to the company’s sales tracking system having a glitch.

8.2. Linear Regression Analysis

The first linear regression analysis was conducted on the total sales of the top 5 categories, which included all months. Figure 20 shows that the line is in a downward trend, mainly due to the 4 outlying points at the end of the graph.

Another linear regression analysis was performed with the same sales data but excluding the sales data from October 2017 until January 2018, and the results appear more normal. See Figure 21. Based on the 2 linear regressions conducted, the linear regression analyses for the top 5 categories will exclude sales data from October 2017 until January 2018.

The following are the linear regression formulas for future sales calculations:

- Top 5 Categories = 7.2865 * x + 524800
- Cleats = -0.2831 * x + 115600
- Men's Footwear = 4.3434 * x + 73050
- Women's Apparel = 0.1576 * x + 81840
- Indoor/ Outdoor Games = -2.6961 * x + 76530
- Fishing = 5.7648 * x + 177800

The linear regression models for Cleats (Figure 22), Men’s Footwear (Figure 23), Women’s Apparel (Figure 24), Indoor/Outdoor Games (Figure 25), and Fishing (Figure 26) can be found in the Table of Figures.

9. References

anuragnayak. (2021, November 8). *RFM Analysis Analysis Using Python*. Retrieved from GeeksforGeeks: https://www.geeksforgeeks.org/rfm-analysis-analysis-using-python/

Connectif. (2022, July 18). *What Are RFM Scores and How To Calculate Them*. Retrieved from Connectif: https://connectif.ai/en/what-are-rfm-scores-and-how-to-calculate-them/

Jidge, A. (2020, May 25). *The Complete Guide to Linear Regression Analysis*. Retrieved from Towards Data Science: https://towardsdatascience.com/the-complete-guide-to-linear-regression-analysis-38a421a89dc2

Selvaraj, N. (2022, November 4). *How to Perform Market Basket Analysis in Python*. Retrieved from 365 Data Science: https://365datascience.com/tutorials/python-tutorials/market-basket-analysis/

Figure 1 Percentage of Late Deliveries by Category

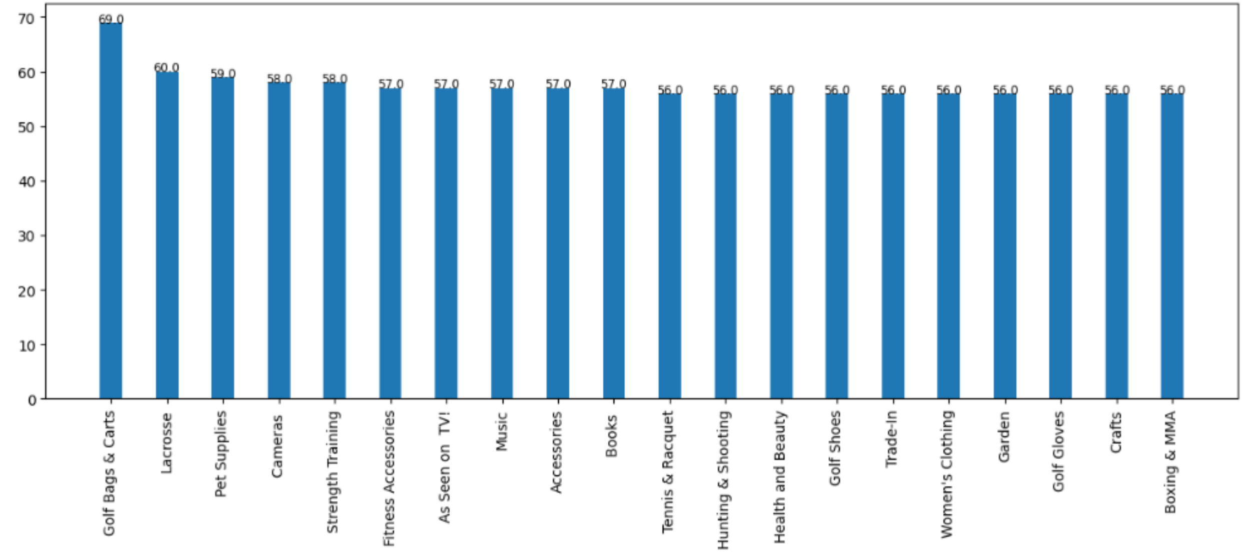


Figure 2 Percentage of Late Deliveries by Destination Country

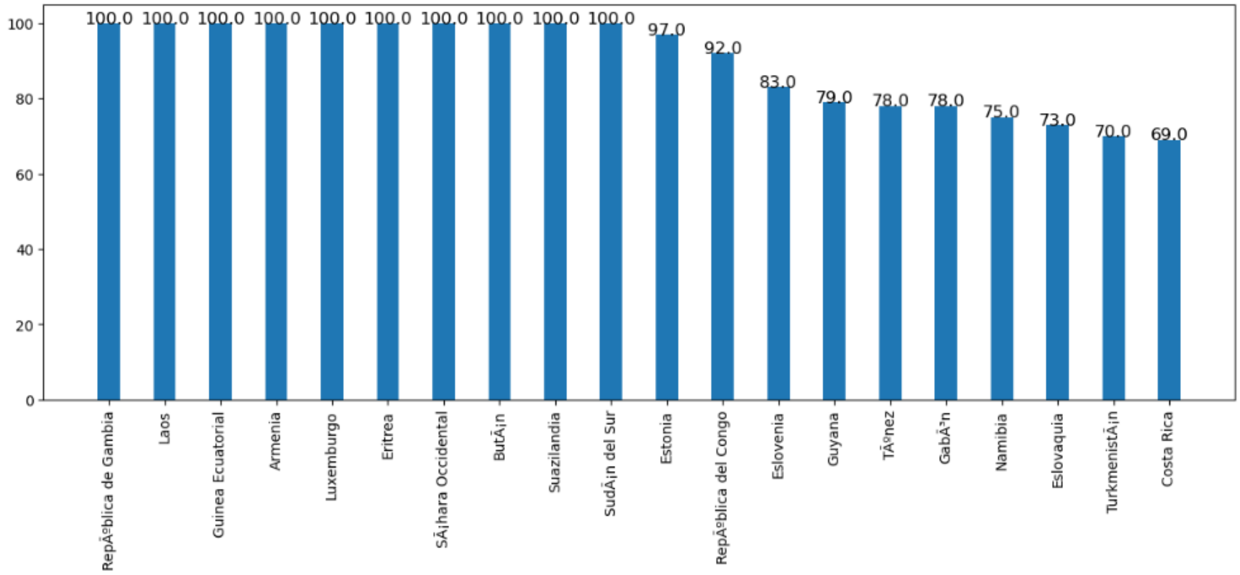


Figure 3 Customer Segmentation Distribution

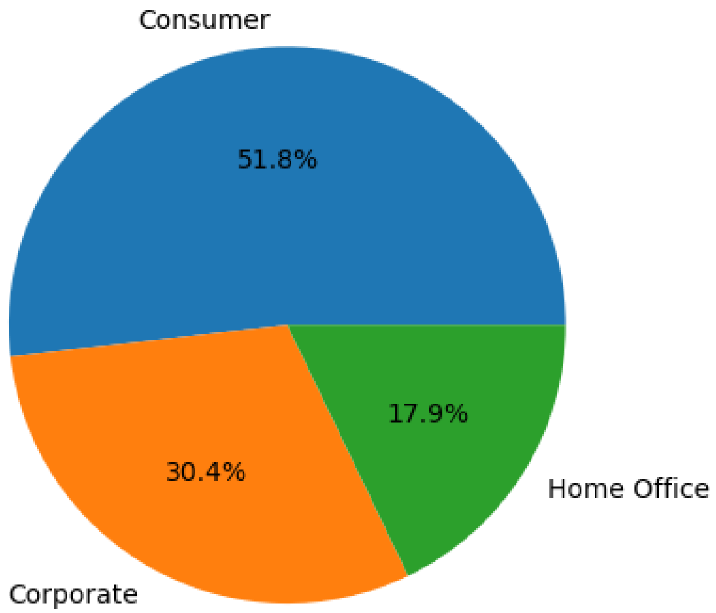


Figure 4 Payment Type Distribution

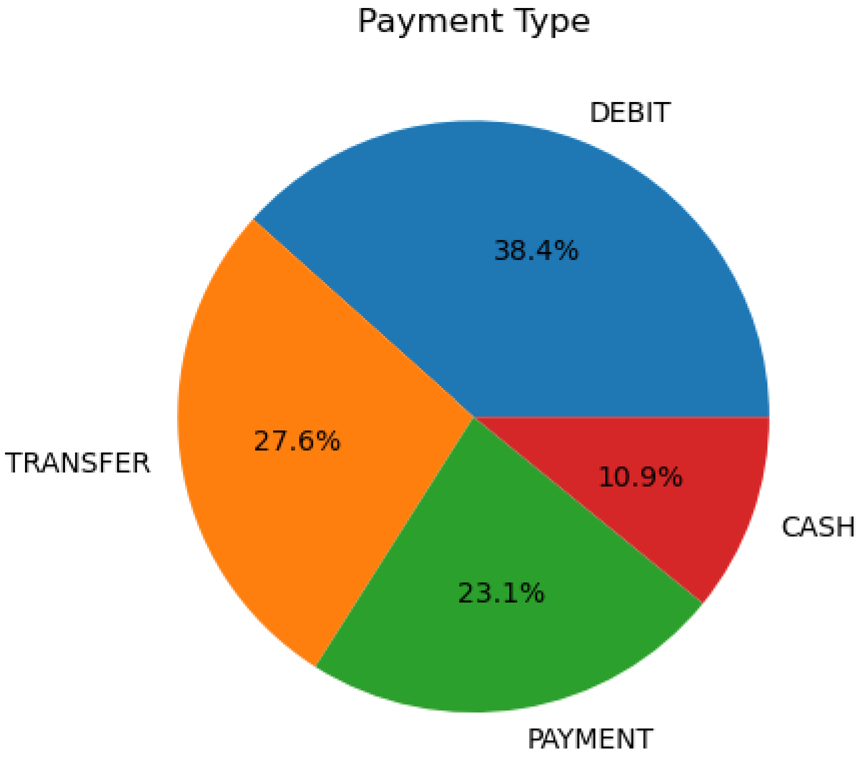


Figure 5 Category Name Distribution

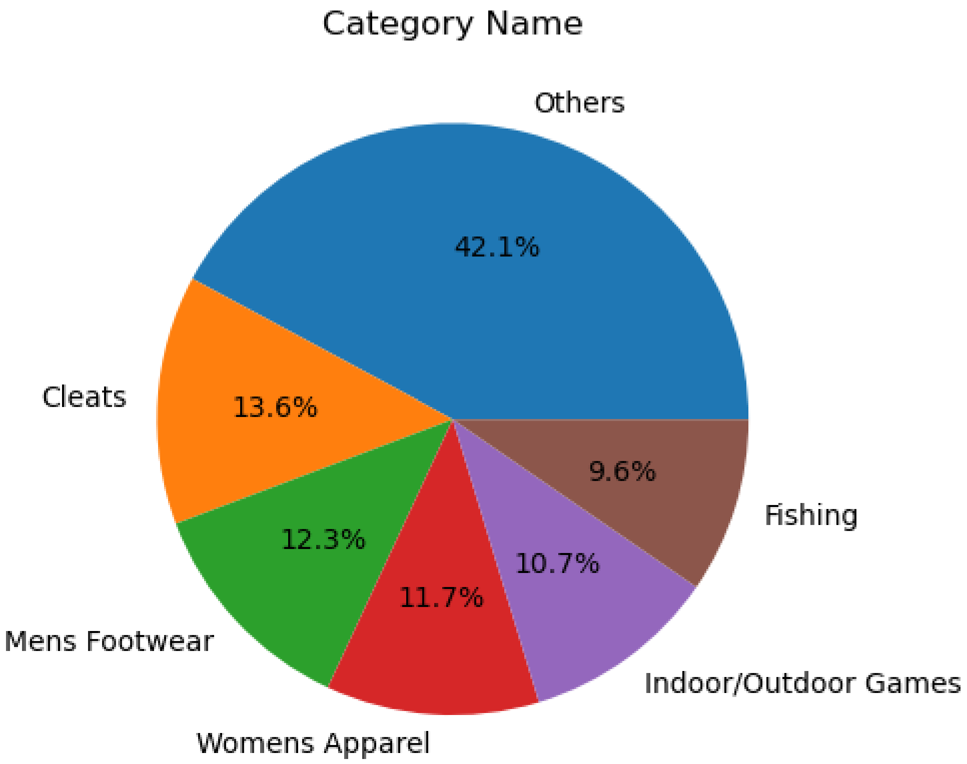


Figure 6 Customer Country and City Distribution

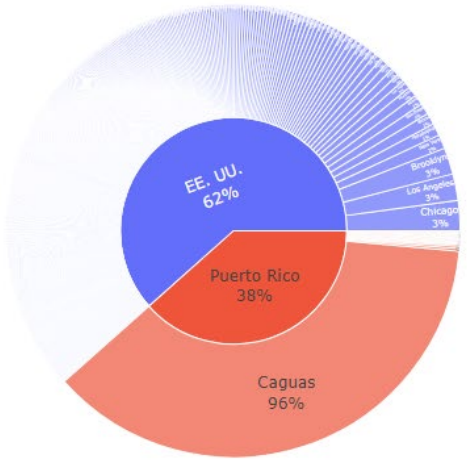


Figure 7 Order Country and City Distribution

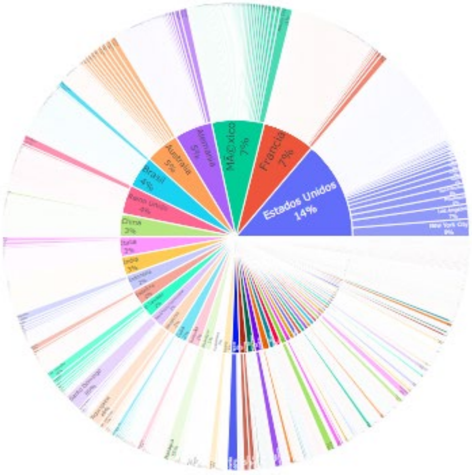


Figure 8 Shipping Mode Distribution

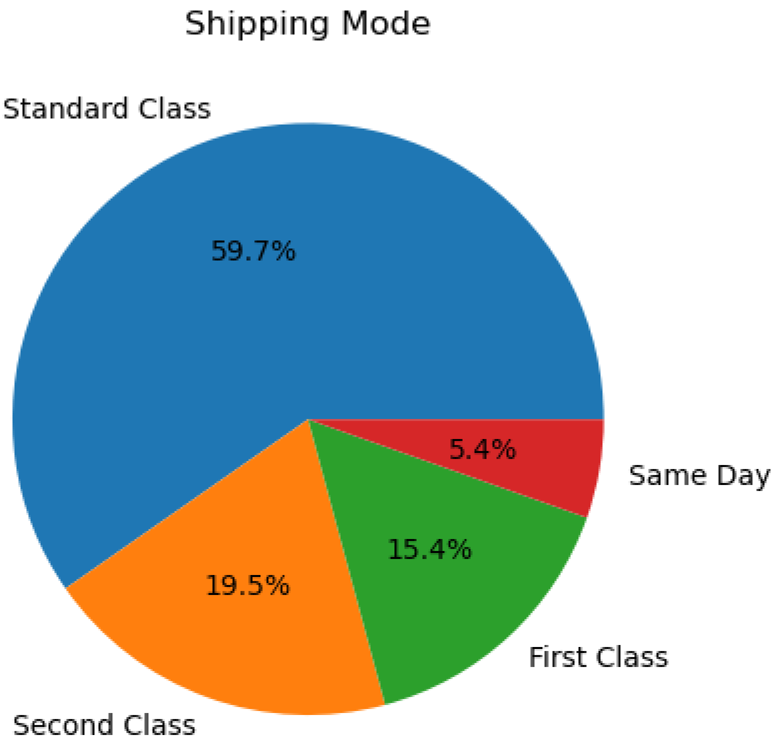


Figure 9 Payment Type by Customer Segment

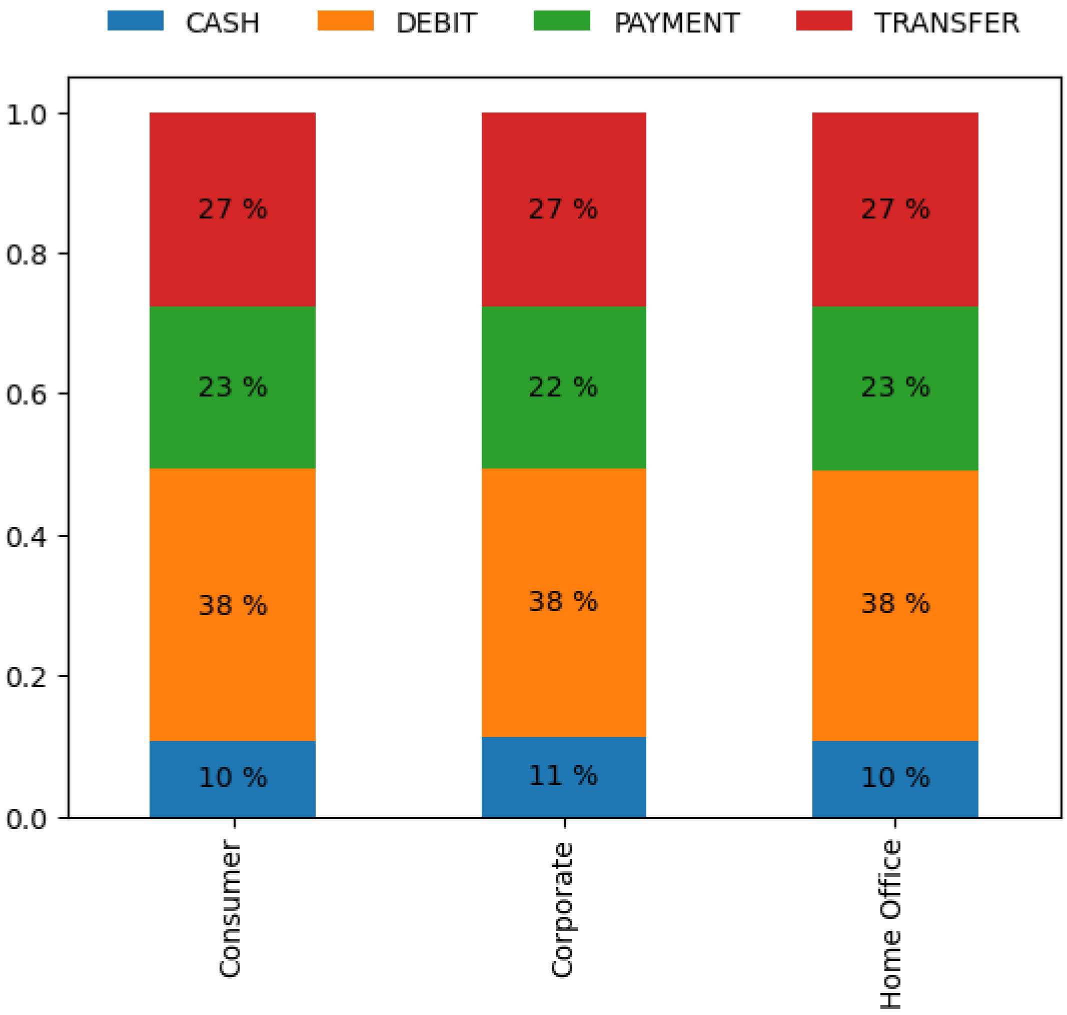


Figure 10 Shipping Mode by Customer Segment

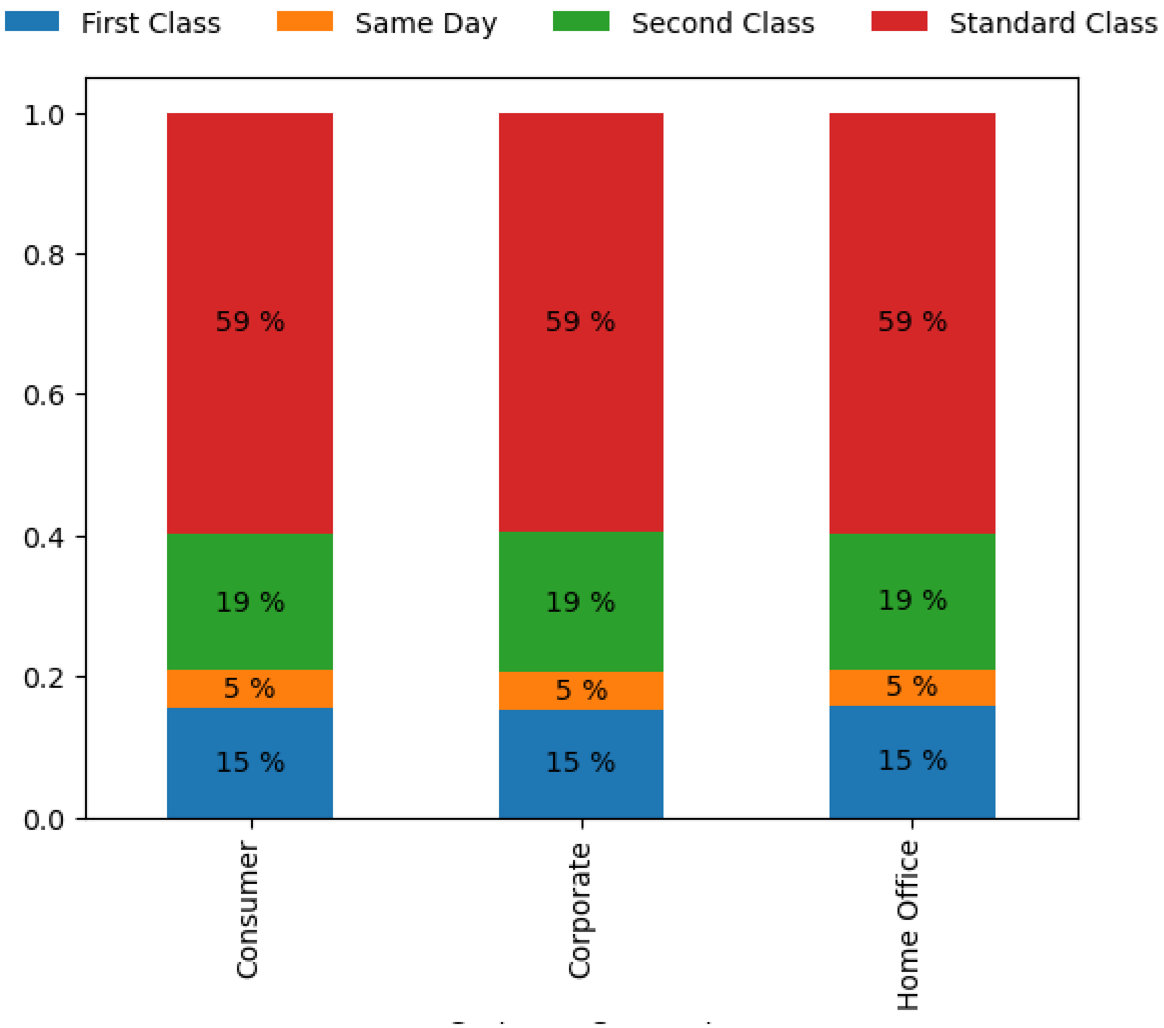


Figure 11 Categories by Customer Segment

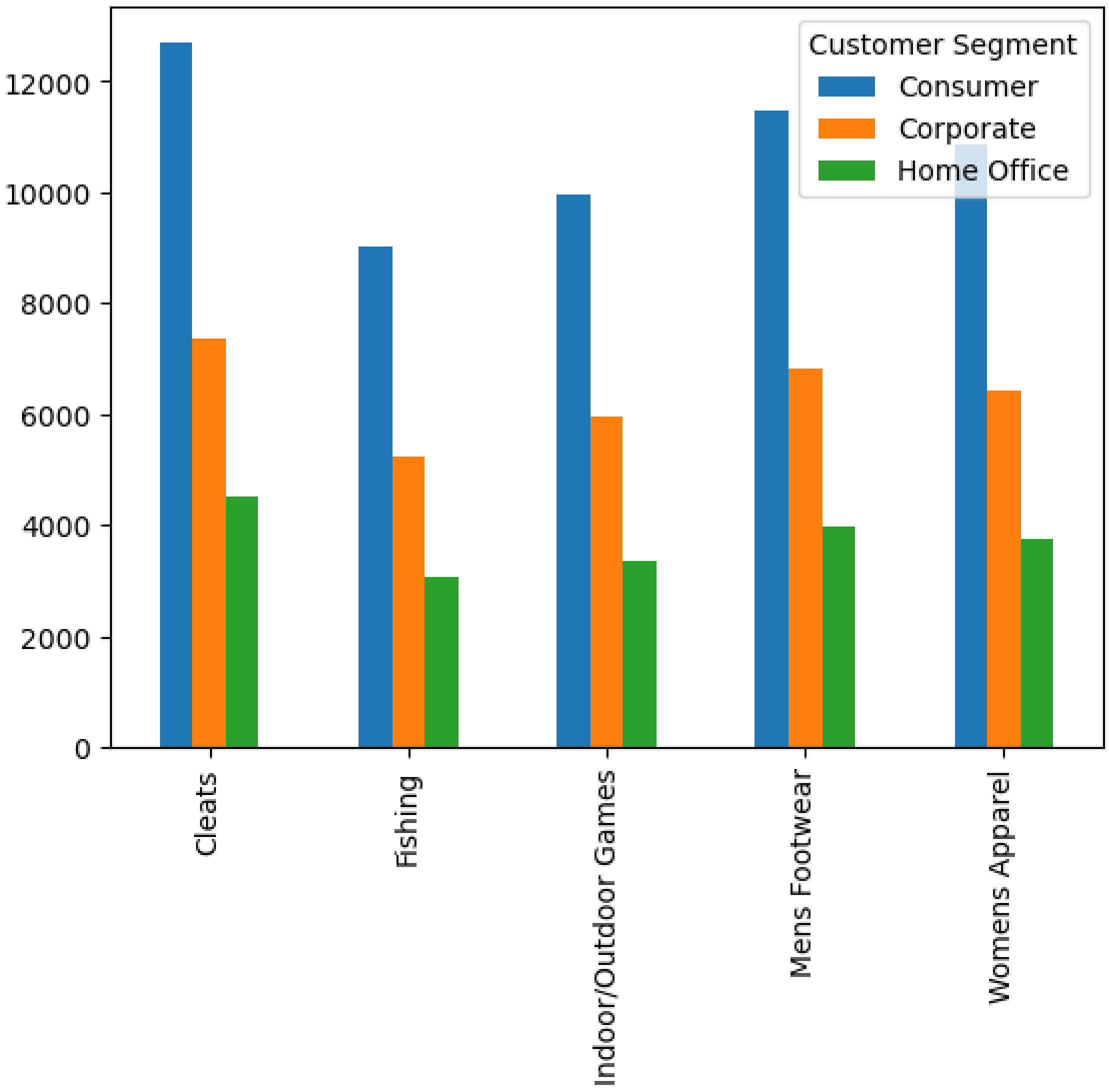


Figure 12 Customer Country by Customer Segment

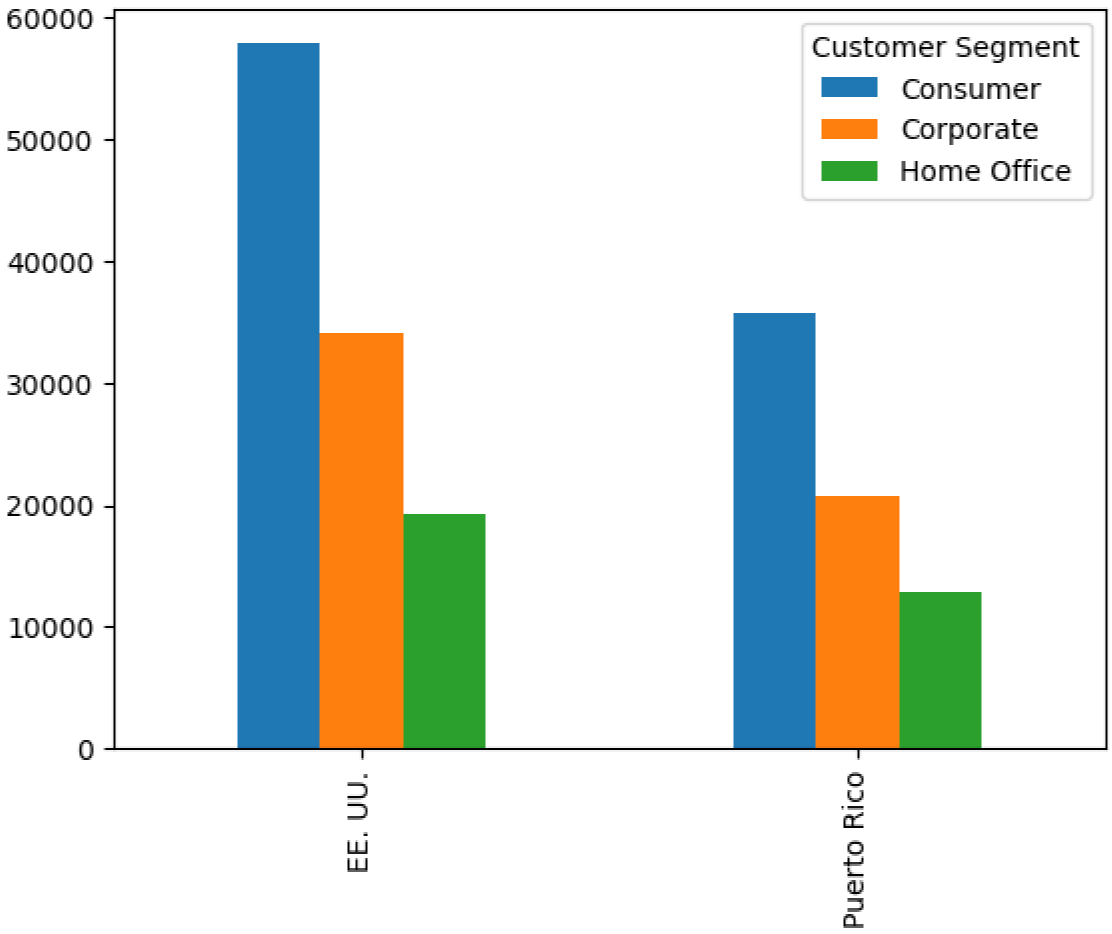


Figure 13 Order Country by Customer Segment

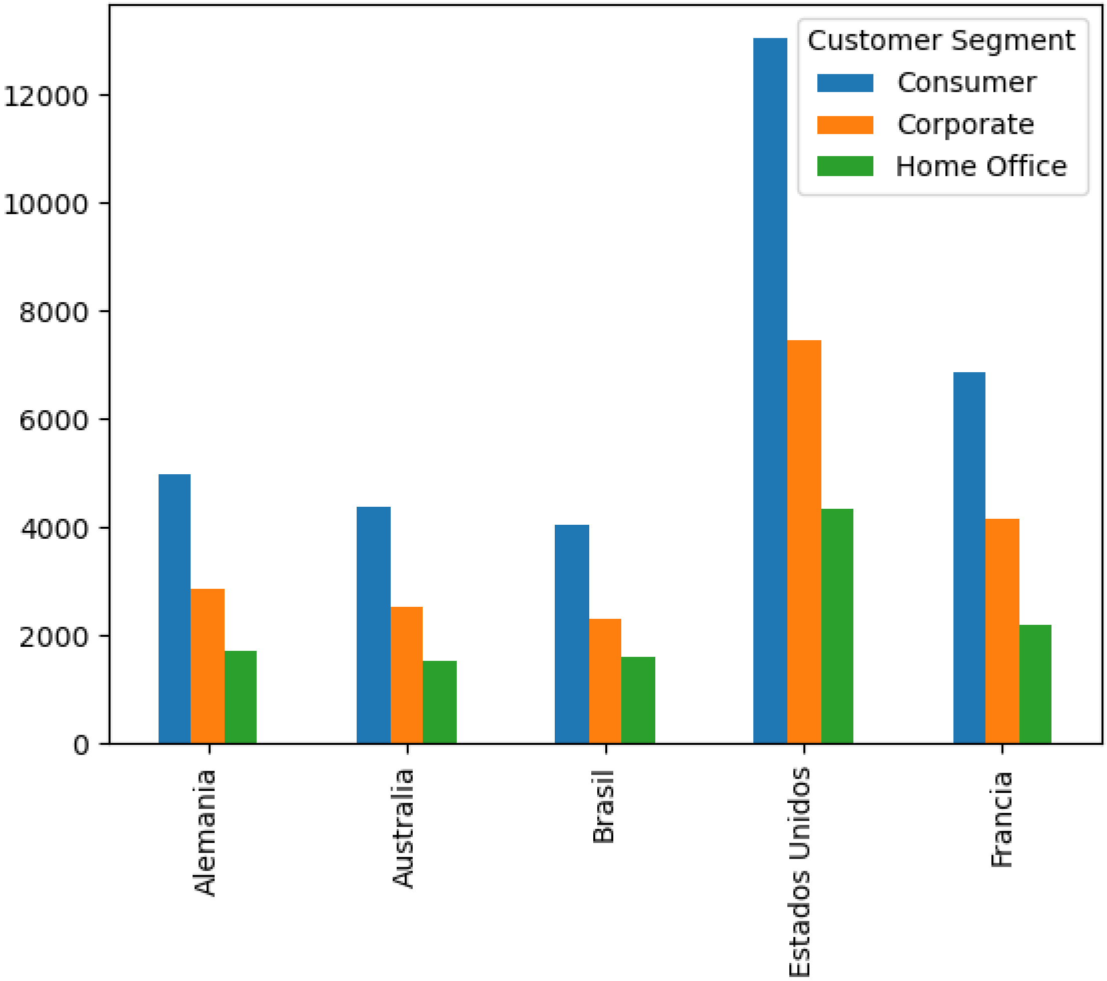


Figure 14 Customer Value Distribution

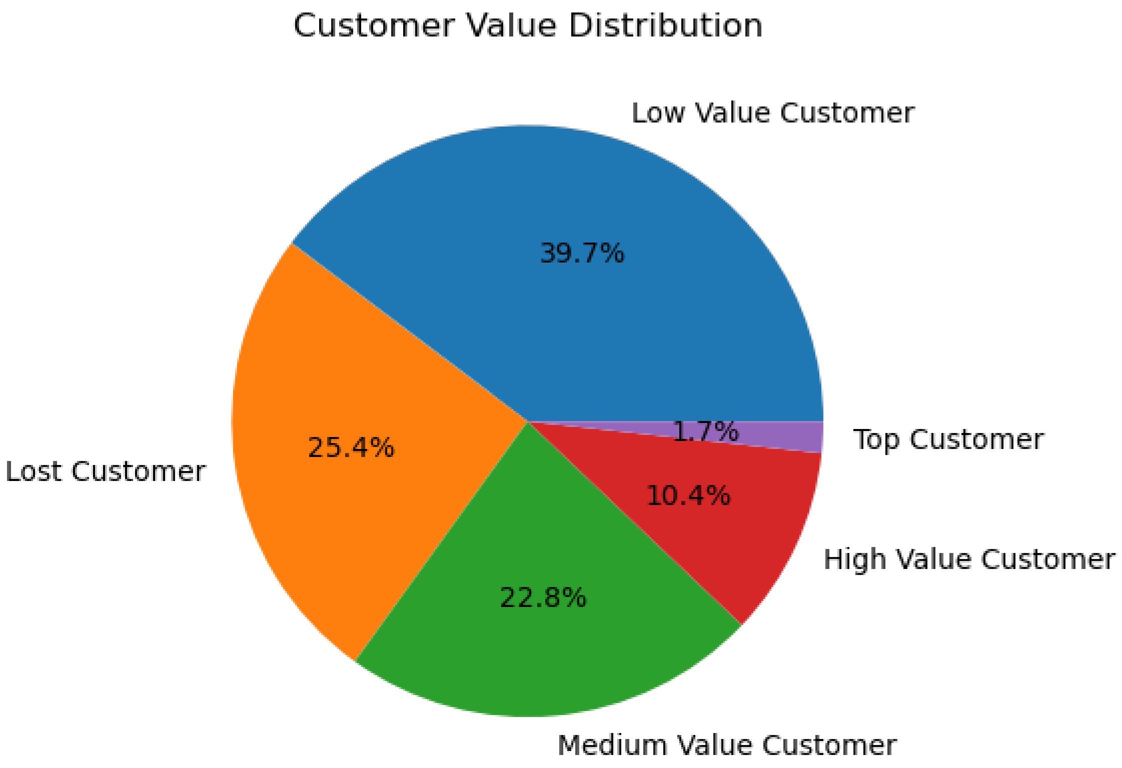


Figure 15 RFM Segmentation Distribution

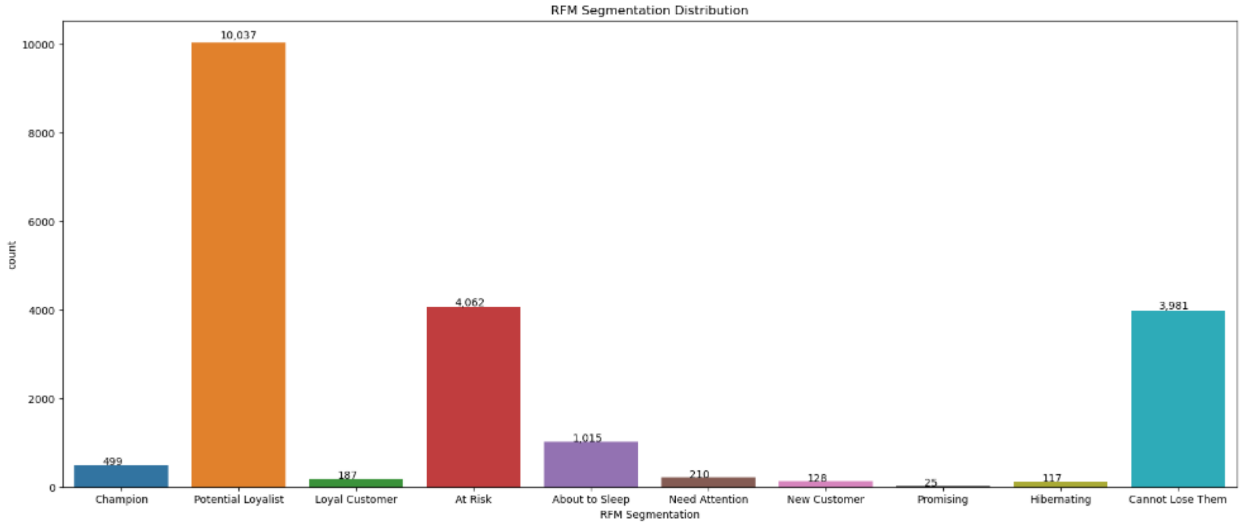


Figure 16 Basket Analysis by Product Category

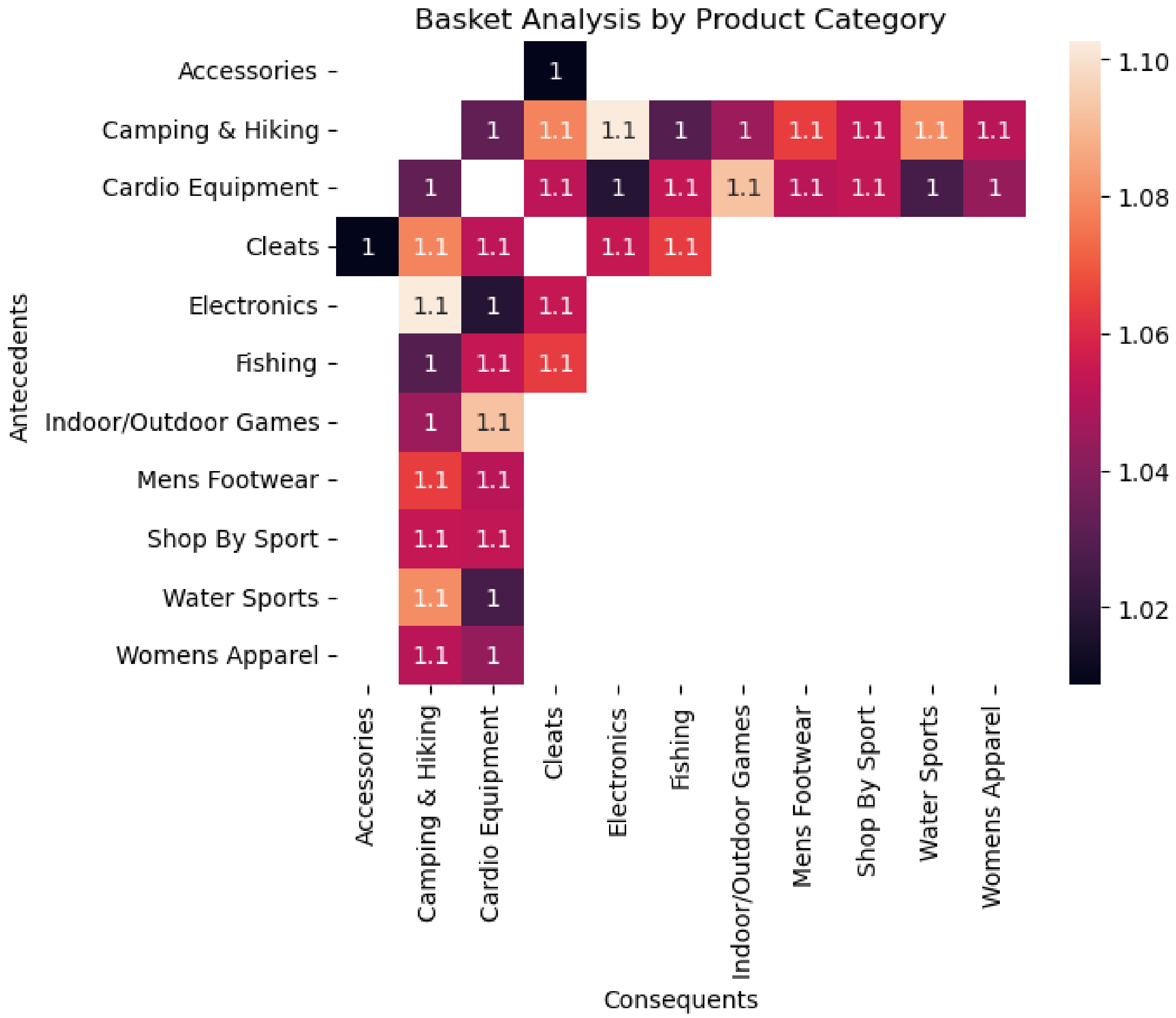


Figure 17 Basket Analysis by Product Name

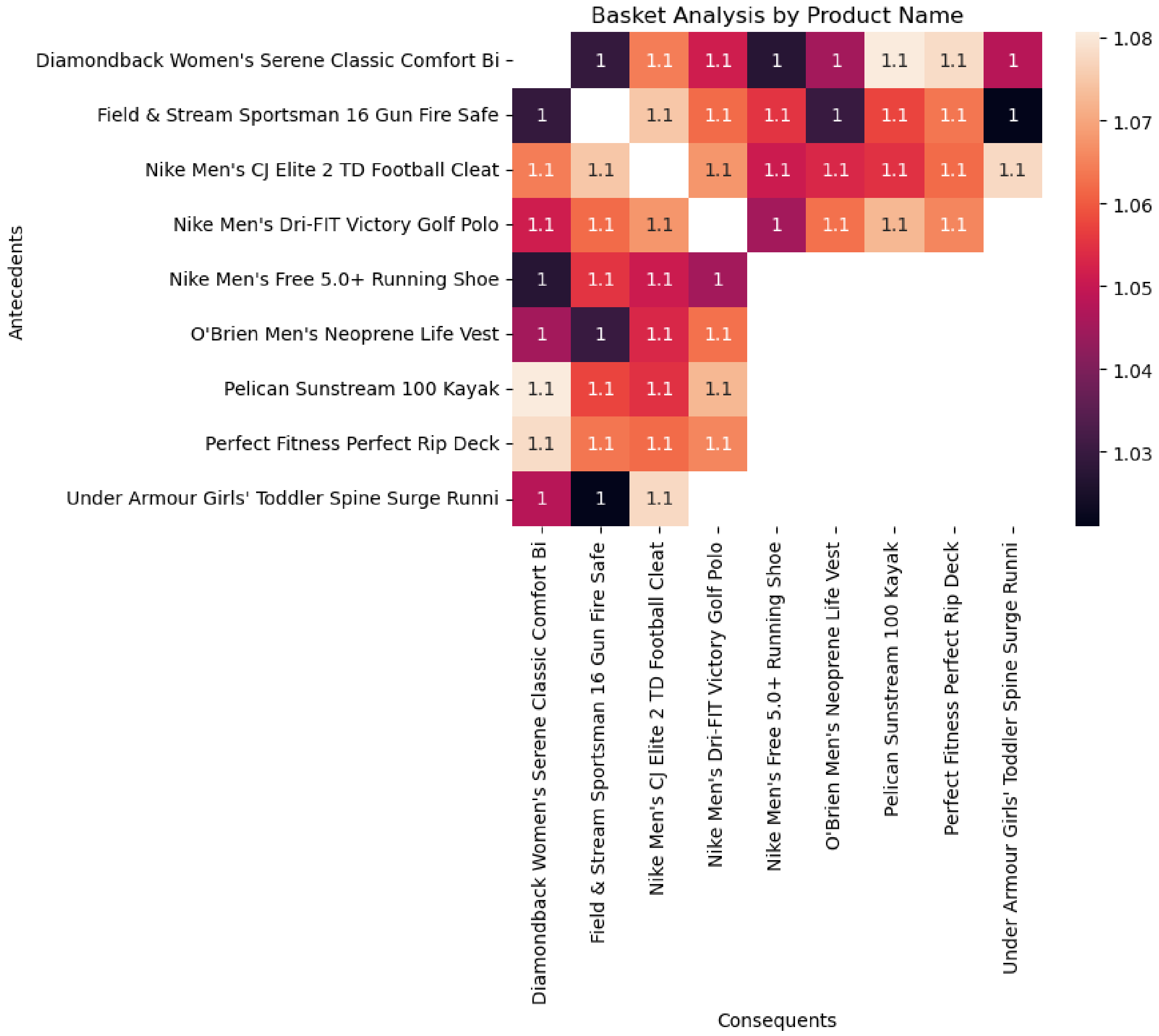


Figure 18 Monthly Sales of Top 5 Categories

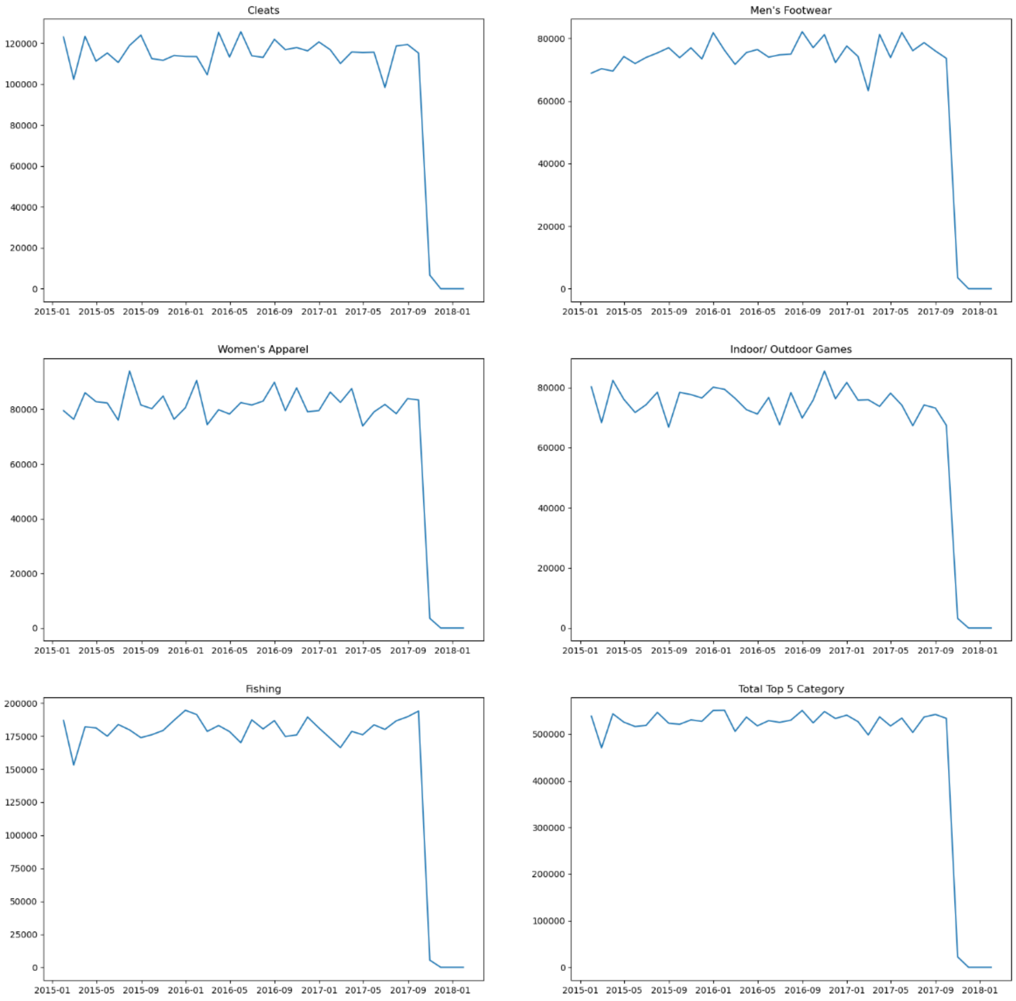


Figure 19 Total Top 5 Scales by Order Date

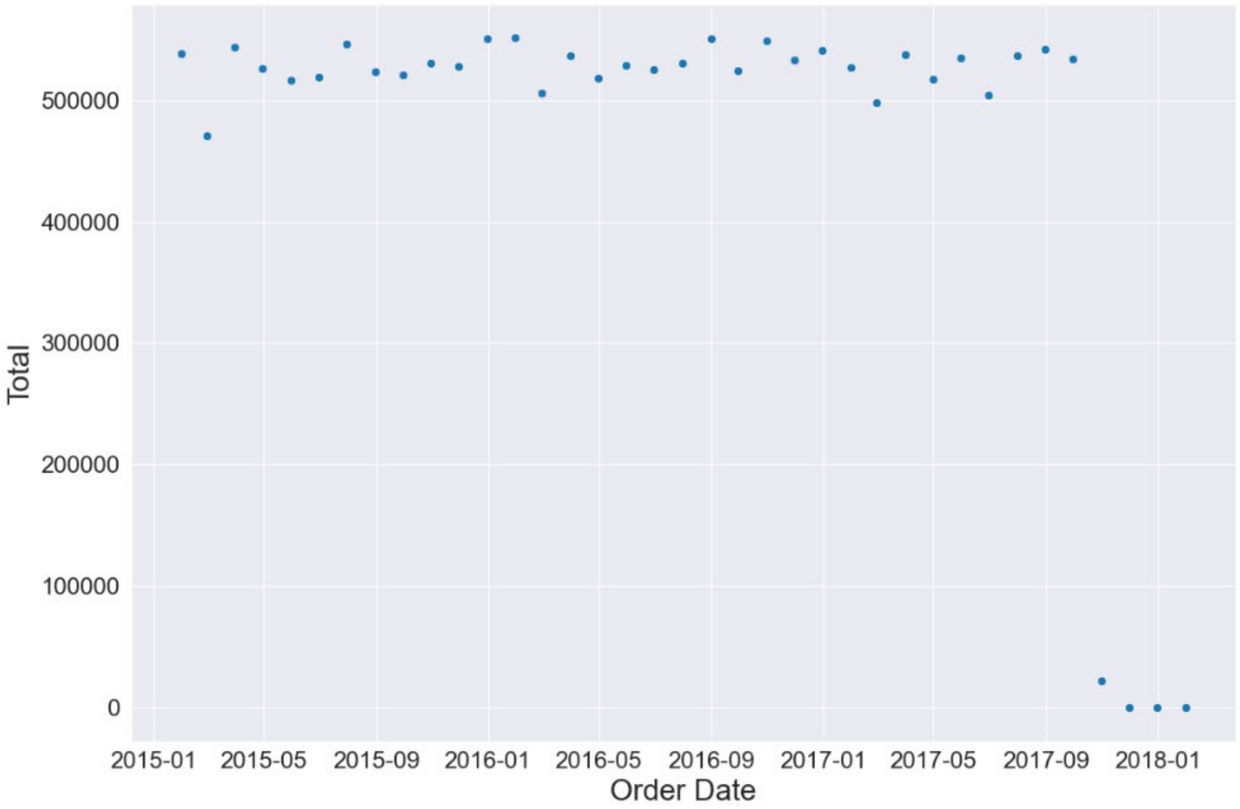


Figure 20 Total Top 5 Sales by Order Date with Linear Regression

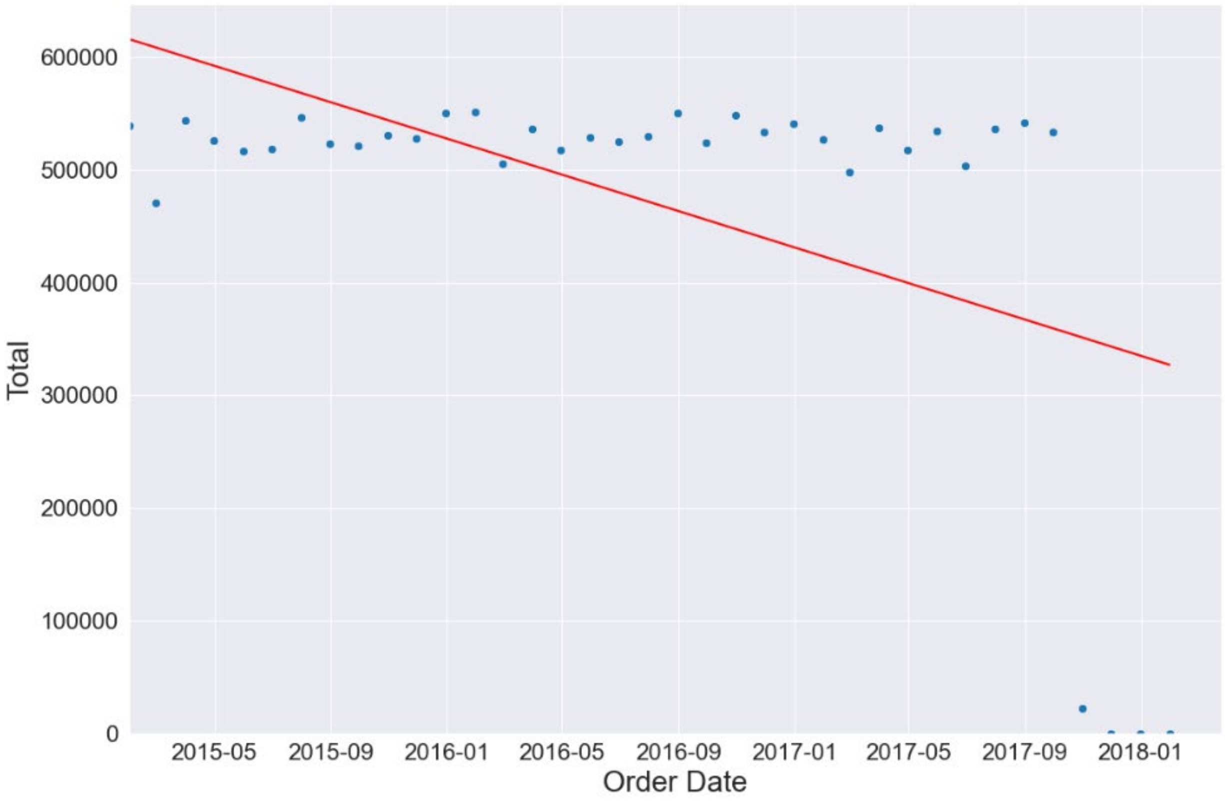


Figure 21 Total Top 5 Sales by Order Data Linear Regression excluding outliers

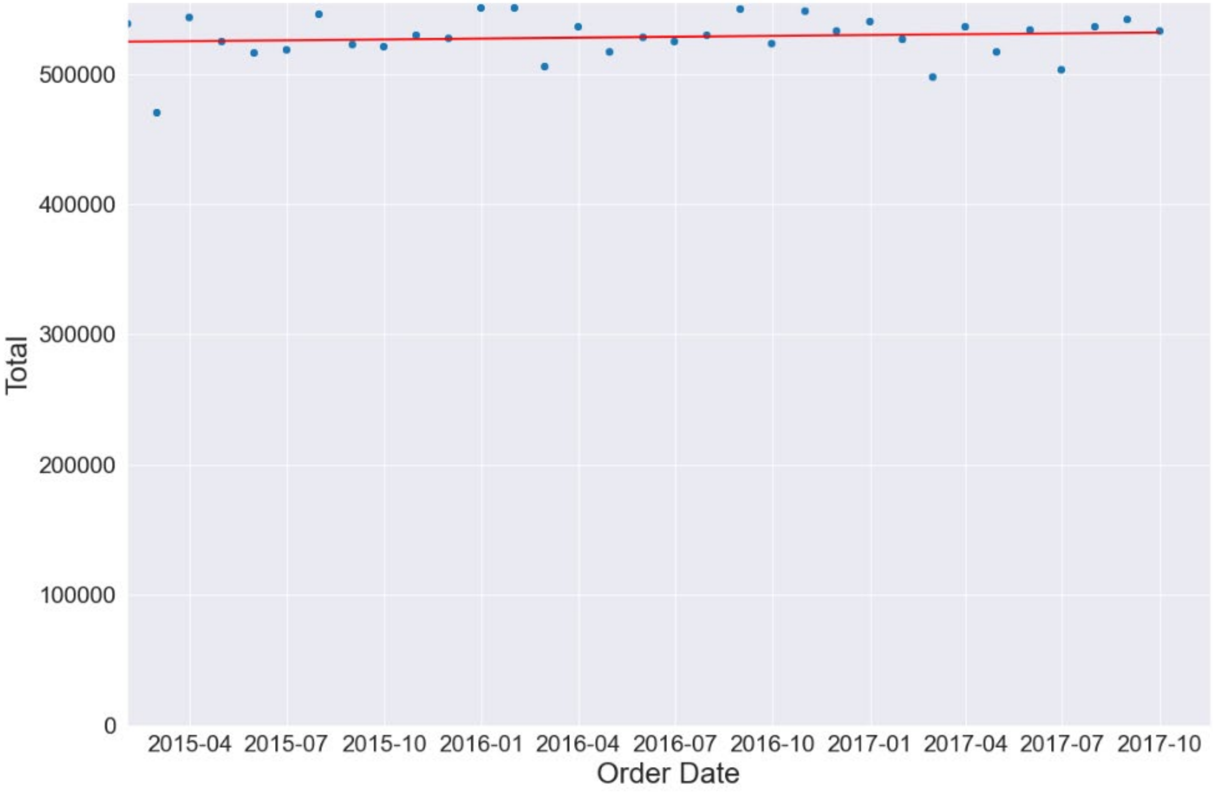


Figure 22 Linear Regression: Cleats

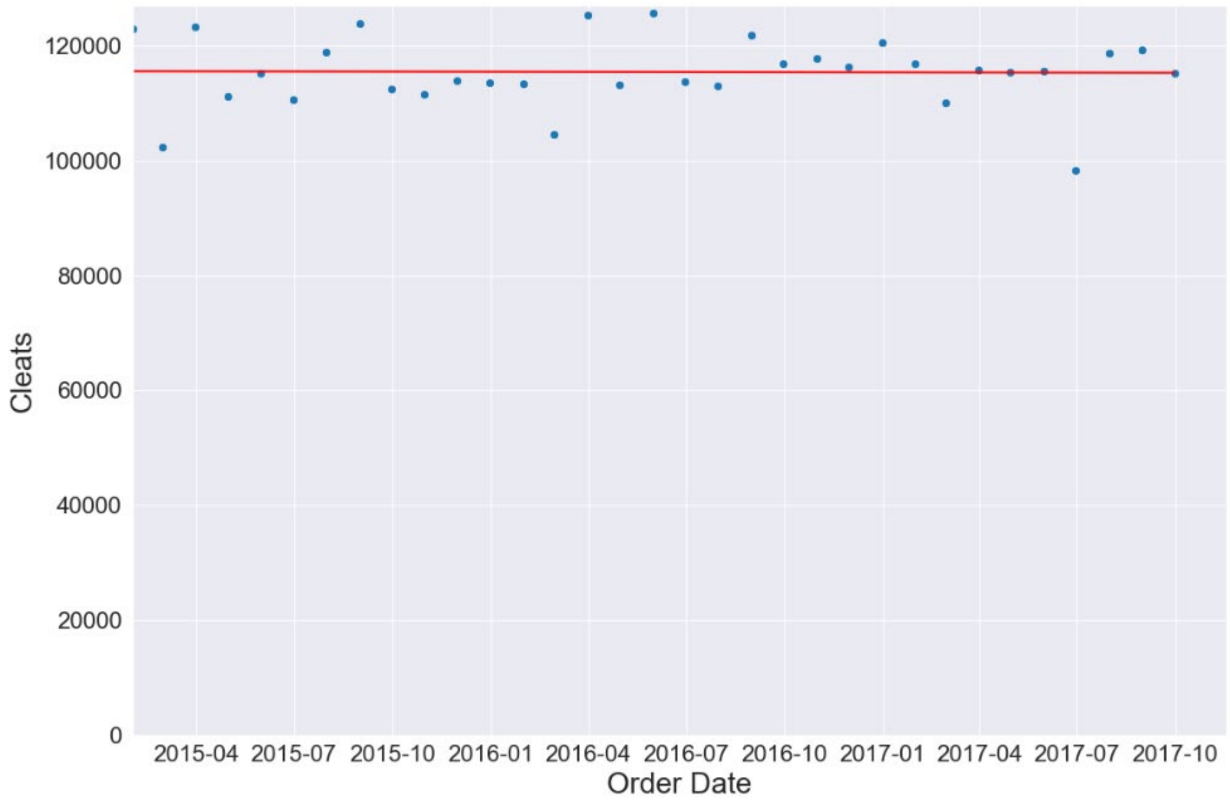


Figure 23 Linear Regression: Men's Footwear

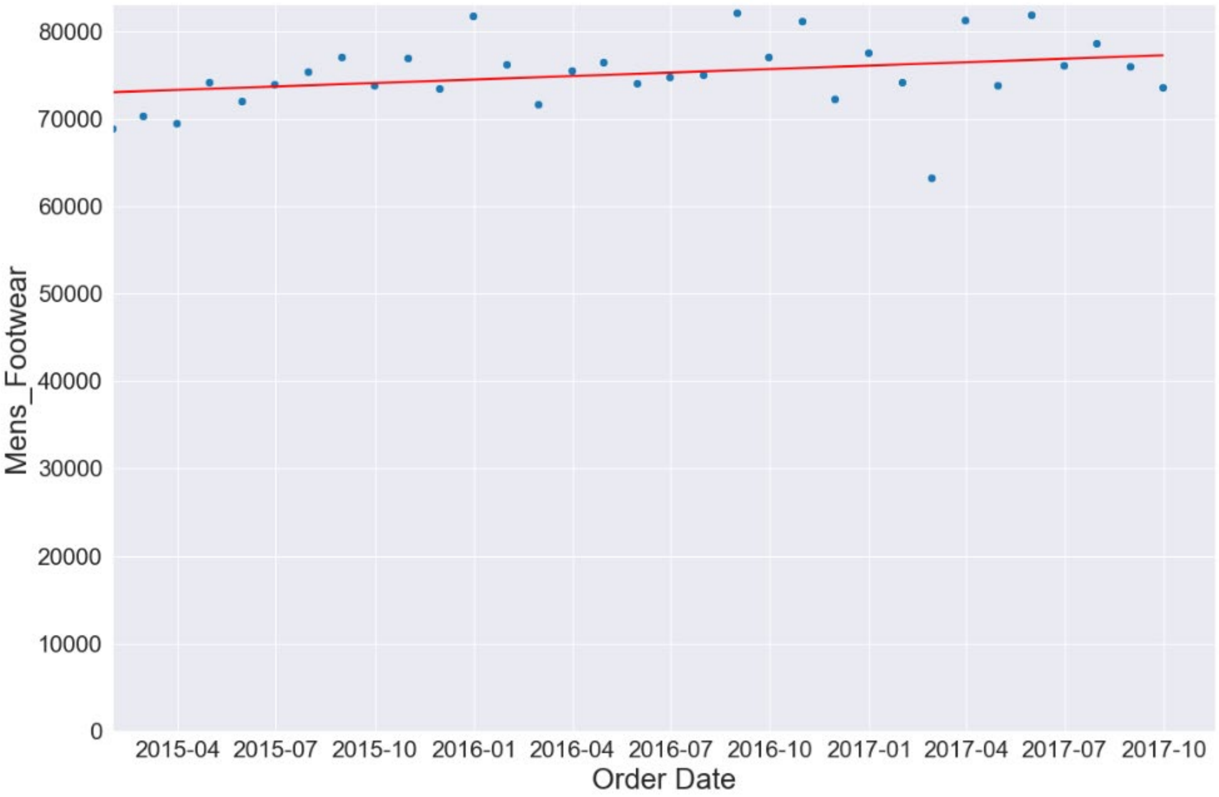


Figure 24 Linear Regression: Women's Apparel

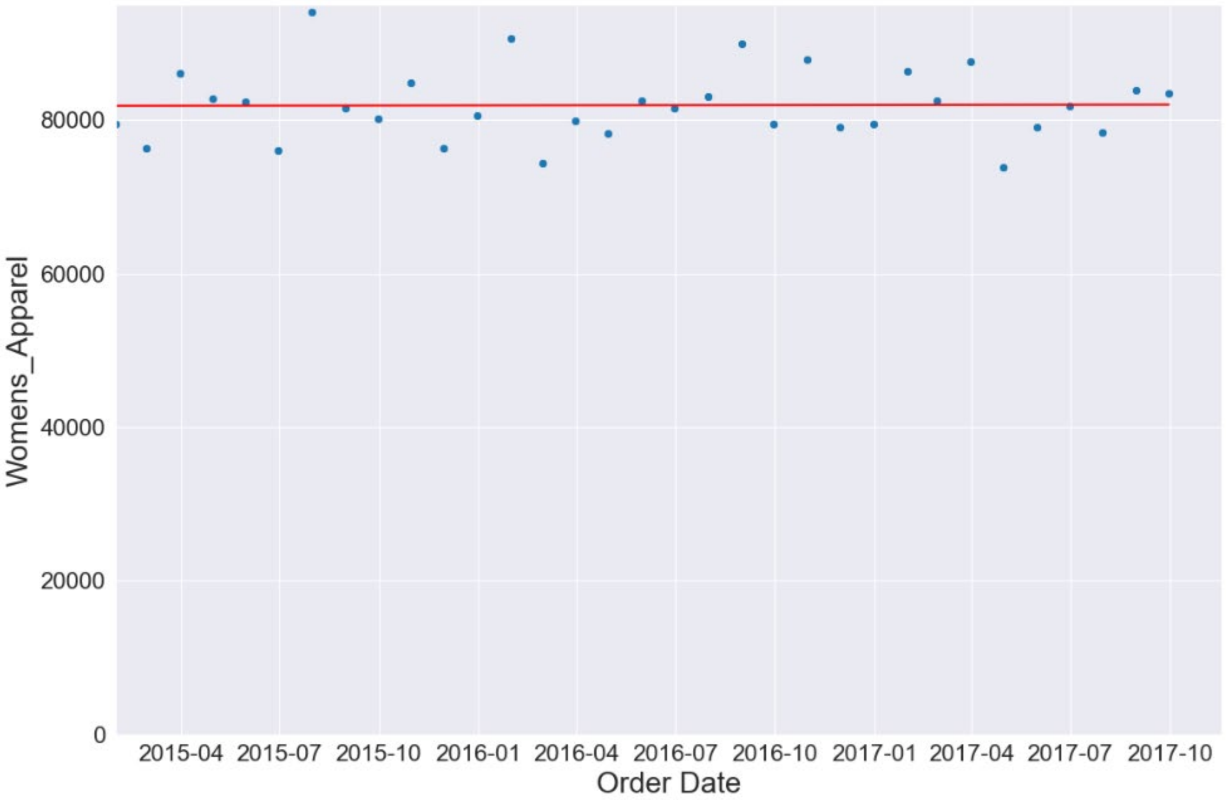


Figure 25 Linear Regression: Indoor/ Outdoor Games

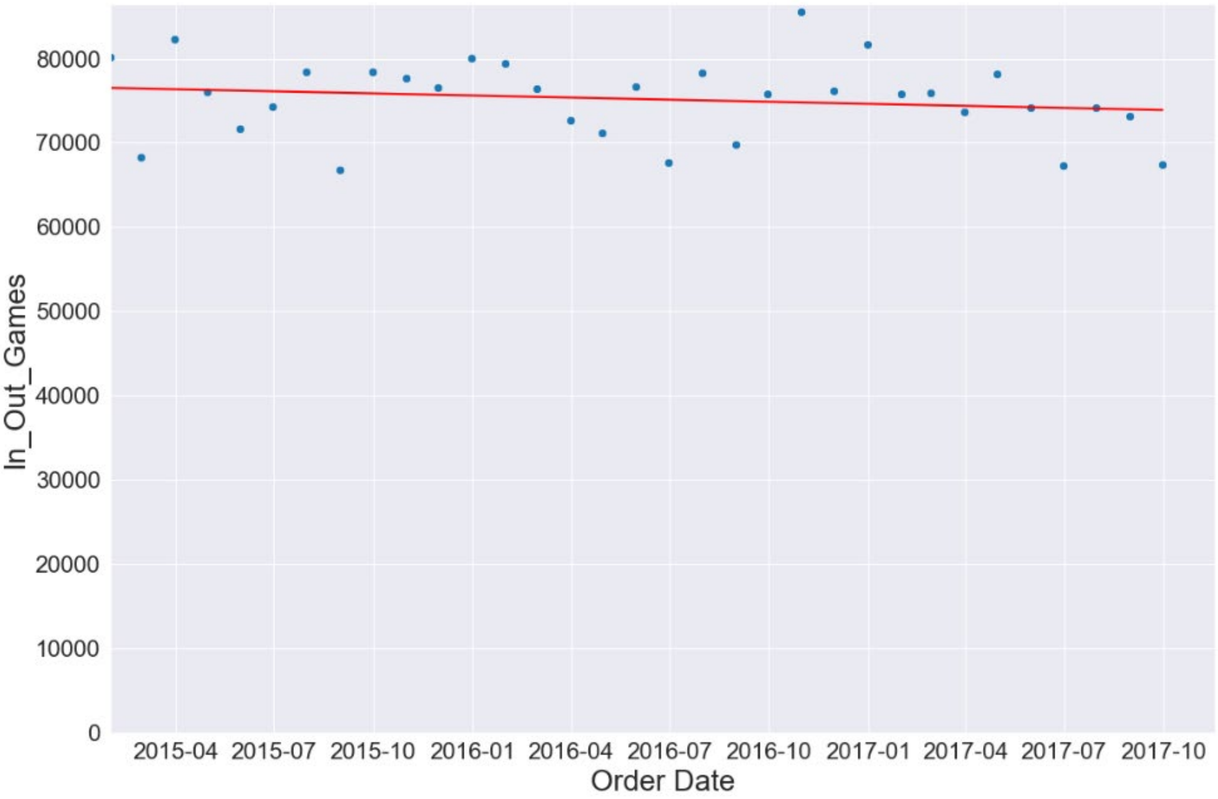
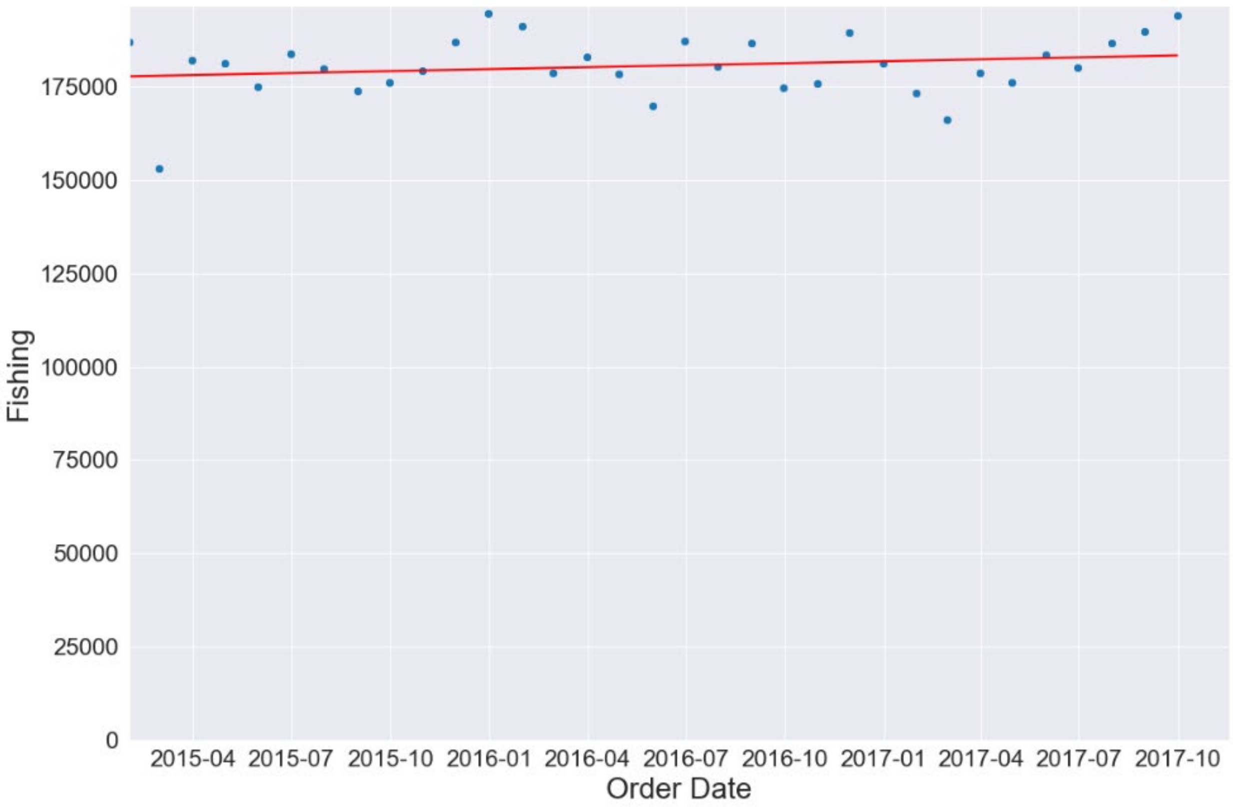


Figure 26 Linear Regression: Fishing


Fishing
Order Date