E-commerce Product Range Analysis

By: Dorothy Kunth

Project Overview

Project Goal

The goal of this project is to identify the top-performing products, product categories and customers.

Project Scope

The project's objective is to carry out the following tasks and communicate data findings and recommendations:

- Data Preprocessing
- Exploratory Analysis
- Customer Segmentation using RFM
- Product Segmentation using RFM
- Product Categorization
- Product Category Analysis
- Statistical Hypotheses

Dataset Overview

The dataset is an open online retail dataset from https://archive.ics.uci.edu/ml/datasets/online+retail# which is a sales transactions history of an online store that sells household goods. It contains 541,909 transaction records from 2018-11-29 to 2019-12-07. The data has 7 attributes:

InvoiceNo: *reference number uniquely assigned to each transaction;*

If the number starts with C, it indicates a cancellation

StockCode: *item code uniquely assigned to each distinct product*

Description: *product name*

Quantity: *quantities for each product per transaction*

InvoiceDate: *transaction date and time*

UnitPrice: *product price per unit*

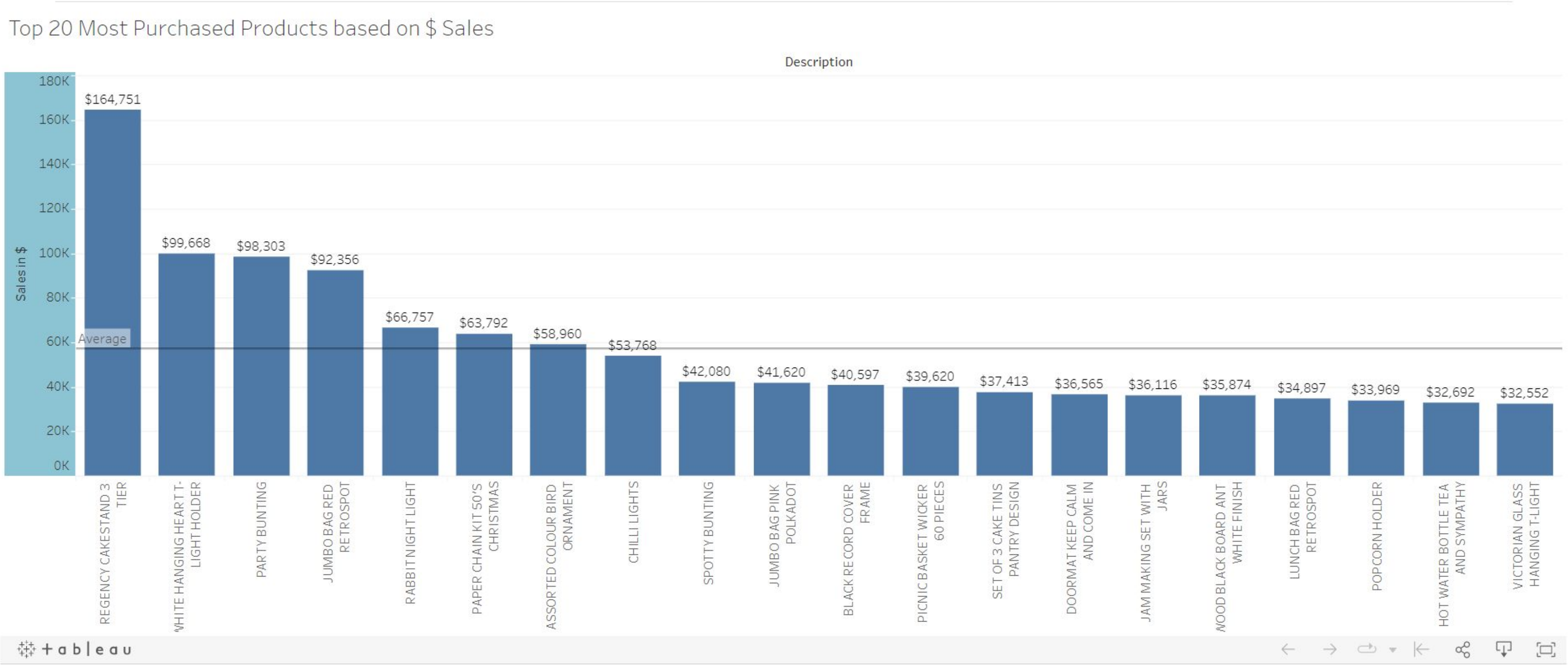
CustomerID: *reference number uniquely assigned to each customer*

Data Preprocessing

The first thing to manage before performing any analysis is to prepare the data. The data has been explored and preprocessed. And the following have been cleaned accordingly:

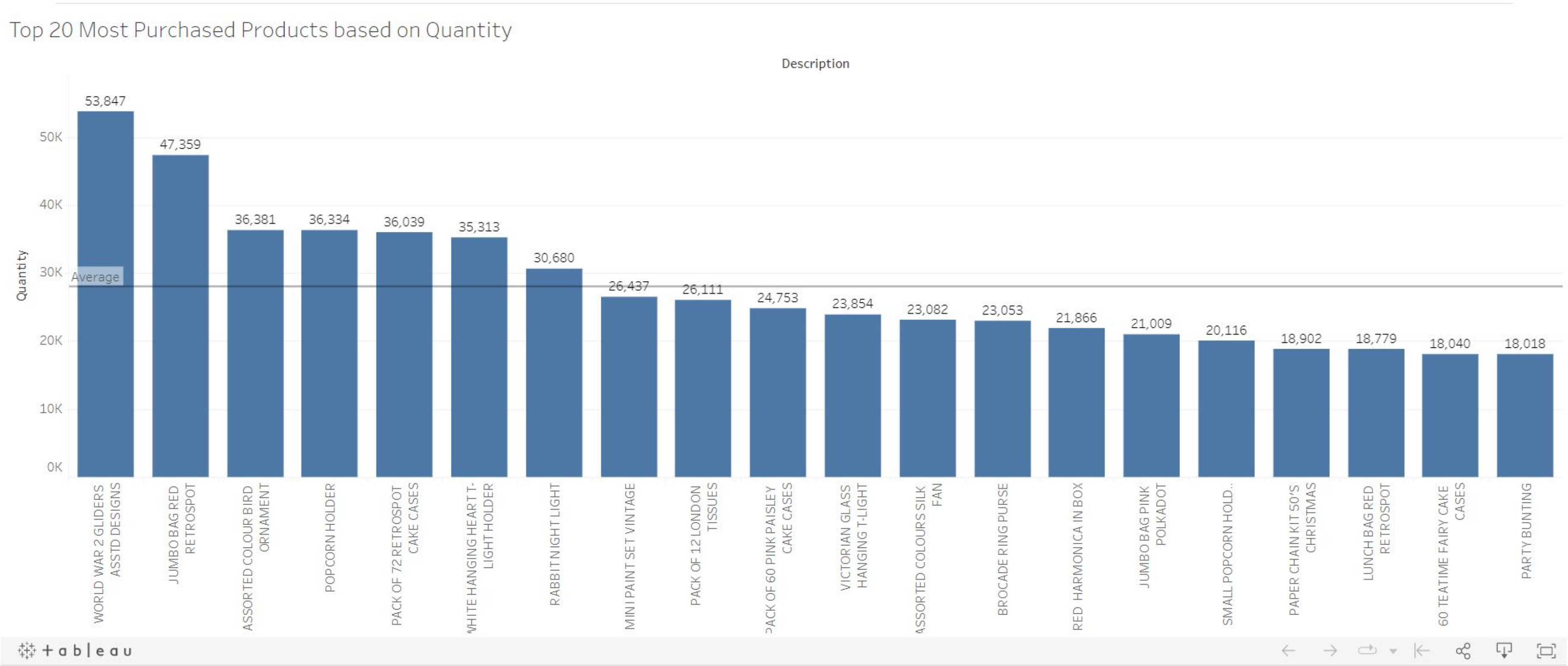
- Records where *InvoiceNo* starts with A were removed as these are related to bad debts adjustment.
- Records where *StockCode* is non-numeric were removed as these are not related to sales of products.
- Records where UnitPrice is 0 were removed as these are not valid sales transactions and mostly related to inventory adjustments, damages, etc.
- A *TotalSales* column was created. The value is the product of *Quantity* and *UnitPrice*.
- Invoices that starts with C are cancellations and most of the cancellations have a matching original invoice. For every customer where there is one cancellation and one matching invoice, the cancellation pairs (1:1) were removed.
- There are cases where for each customer, there is one cancellation and multiple matching invoices. The search was done for each customer for negative quantities below -100. All identified pairs were also removed.
- Individual columns were created for the extracted month, day, weekday and period. The date-only was also extracted from the InvoiceDate.

What are the top-performing products and product categories?

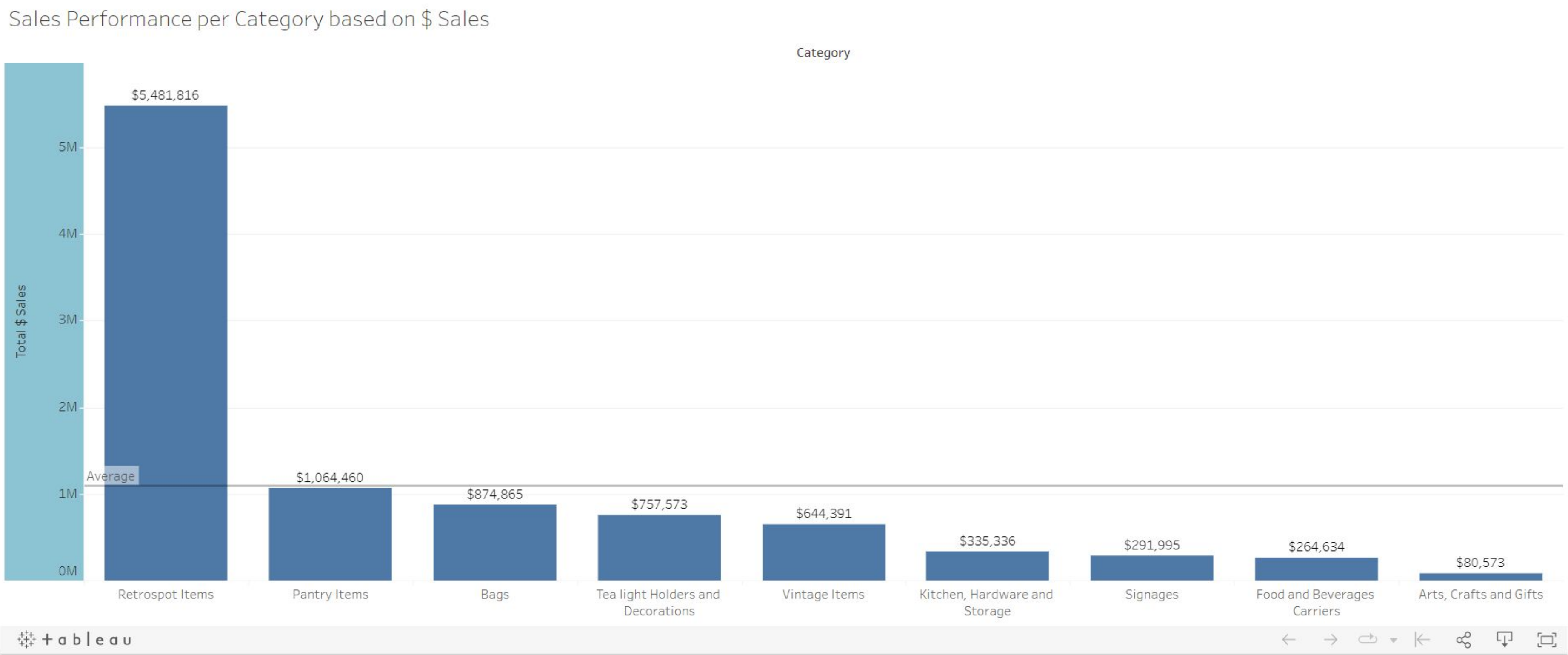


Top 20 Most Purchased Products based on $ Sales

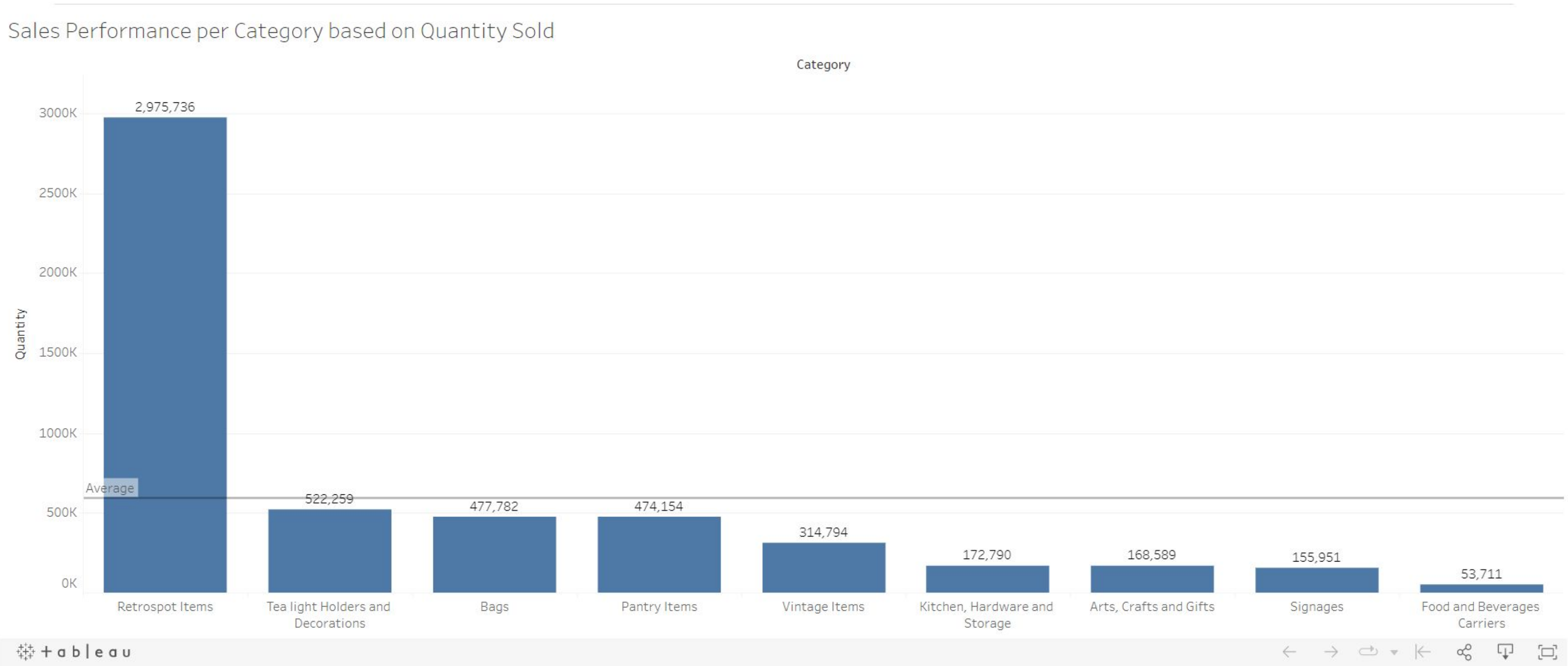
Link:https://public.tableau.com/views/Top20MostPurchasedProductsbasedonSales/Top20MostPurchasedProductsbasedonSales?:language=en-US&:display_count=n&:origin=viz_share_link



Link:
https://public.tableau.com/views/Top20MostPurchasedProductsbasedonQuantity/Top20MostPurchasedProductsbasedonQuantity?:language=en-US&:display_count=n&:origin=viz_share_link



Link: https://public.tableau.com/views/SalesPerformanceperCategorybasedonSales/SalesPerformanceperCategorybasedonSales?:language=en-US&:display_count=n&:origin=viz_share_link



Link:
https://public.tableau.com/views/SalesPerformanceperCategorybasedonQuantitySold/SalesPerformanceperCategorybasedonQuantitySold?:language=en-US&:display_count=n&:origin=viz_share_link

Top-performing Products and Product Categories

The top-performing **products** based on **Total $ Sales** are:

1. REGENCY CAKESTAND 3 TIER
2. WHITE HANGING HEART T-LIGHT HOLDER
3. PARTY BUNTING
4. JUMBO BAG RED RETROSPOT

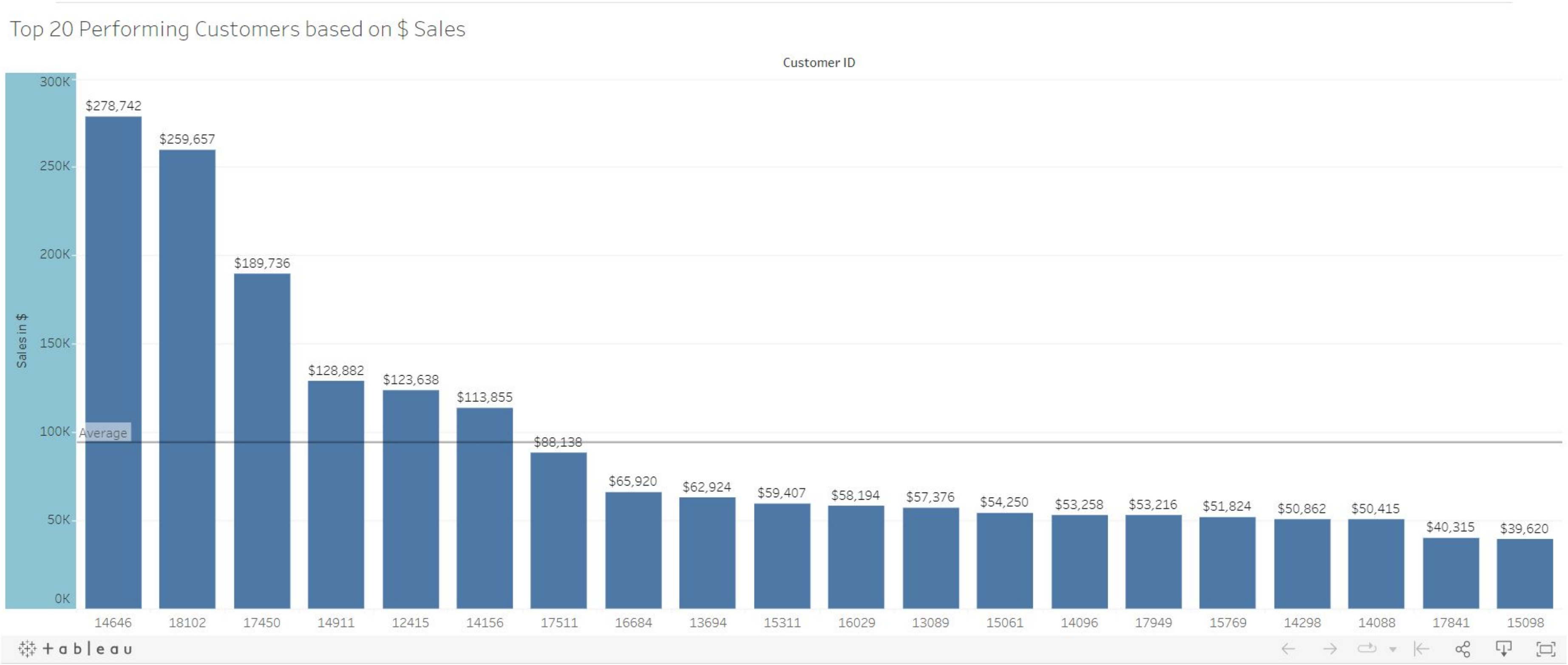
The top-performing **products** based on **Quantity** are:

1. WORLD WAR 2 GLIDERS ASSTD DESIGNS
2. JUMBO BAG RED RETROSPOT.
3. ASSORTED COLOUR BIRD ORNAMENT
4. POPCORN HOLDER
5. PACK OF 72 RETROSPOT CAKE CASES
6. WHITE HANGING HEART T-LIGHT HOLDER

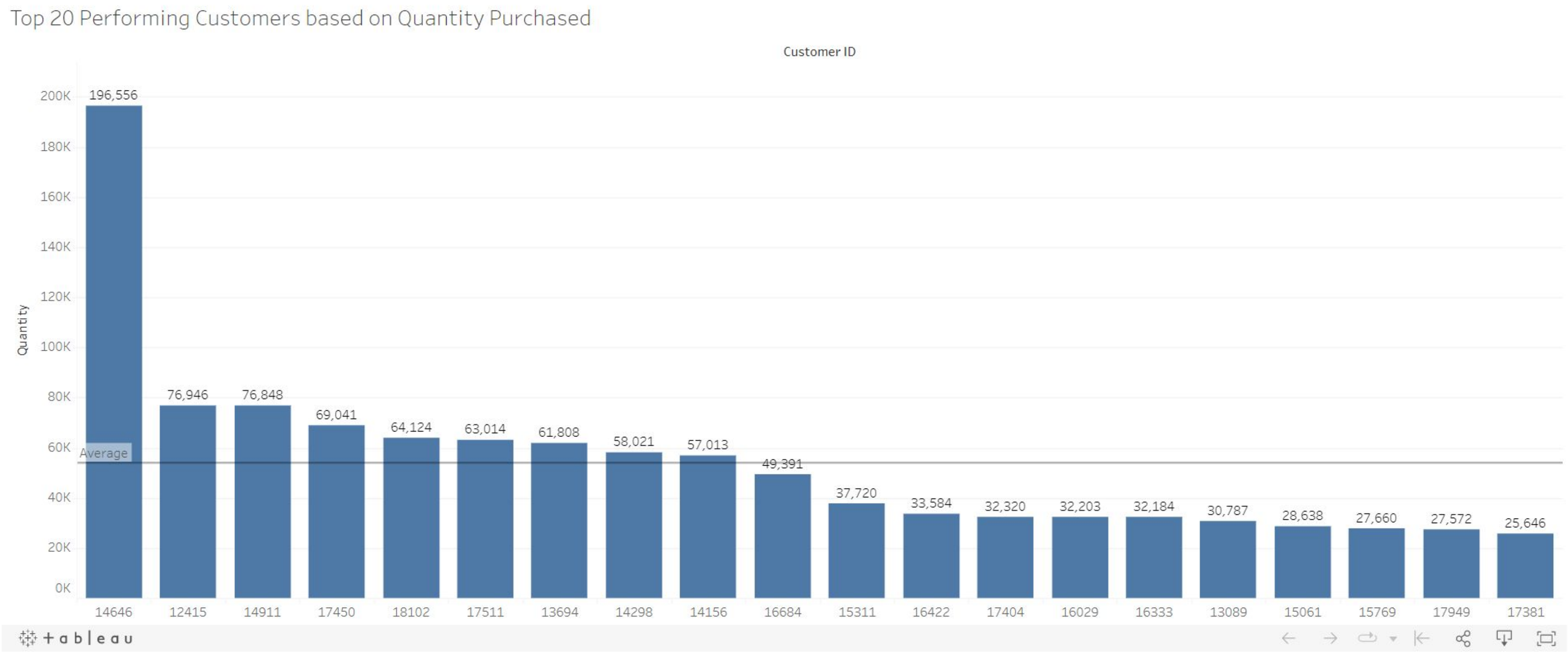
The top-performing **product category** based on both **Total $ Sales** and **Quantity** is:

- Retrospot Items

Who are the top-performing customers?



Link: https://public.tableau.com/views/Top20PerformingCustomersbasedonSales/Top20PerformingCustomersbasedonSales?:language=en-US&:display_count=n&:origin=viz_share_link



Link:
https://public.tableau.com/views/Top20PerformingCustomersbasedonQuantityPurchased/Top20PerformingCustomersbasedonQuantityPurchased?:language=en-US&:display_count=n&:origin=viz_share_link

Top-performing Customers

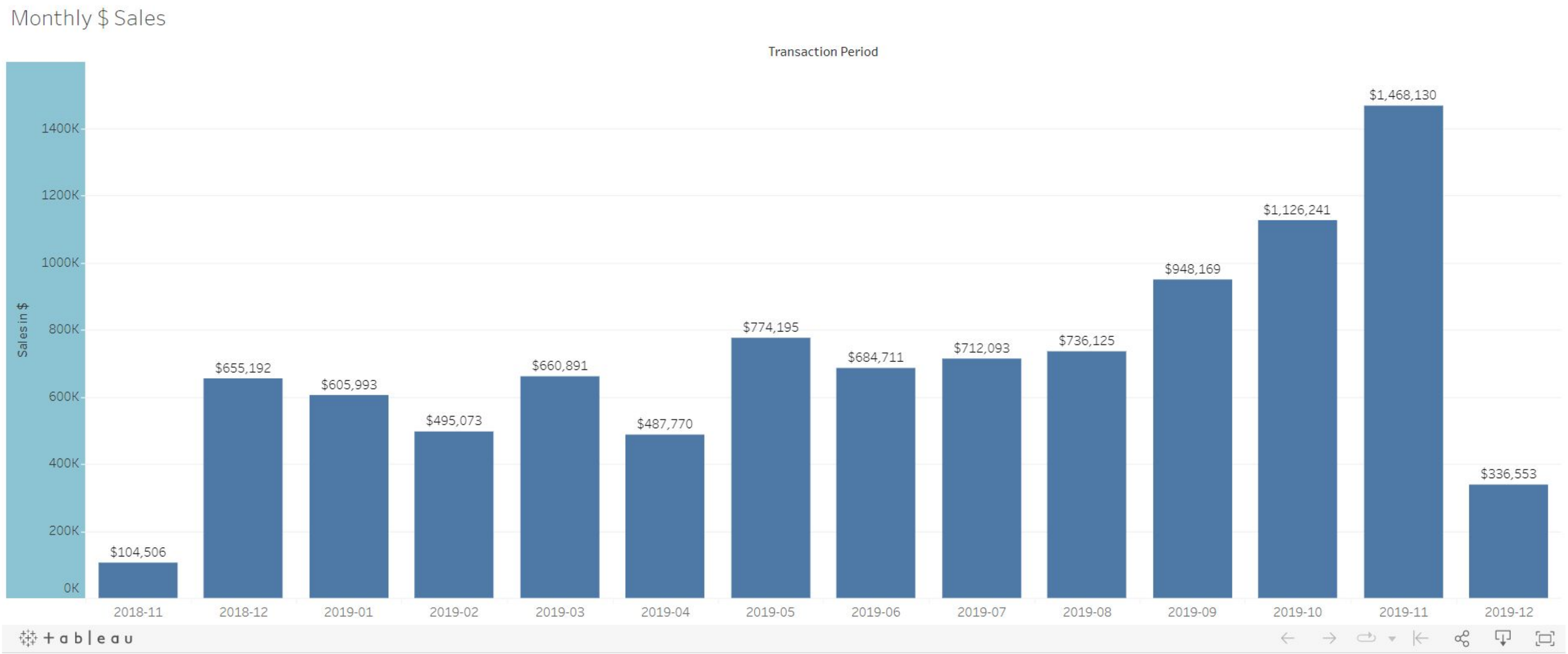
Based on **Total $ Sales**, the **top 5 customers** are:

1. 14646
2. 18102
3. 17450
4. 14911
5. 12415

Based on total **Quantity** purchased, the **top 5 customers** are:

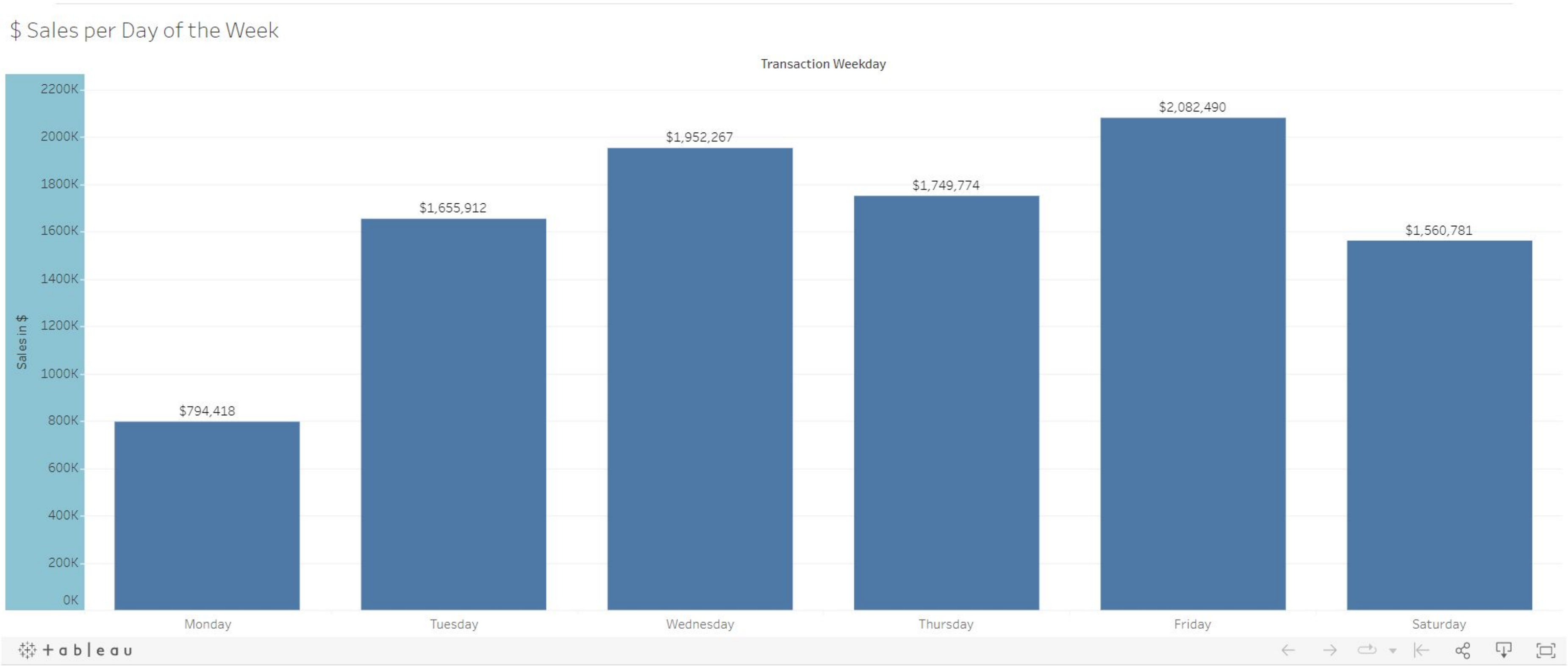
1. 14646
2. 12415
3. 14911
4. 17450
5. 18102

Noticeably, they are the same set of customers.



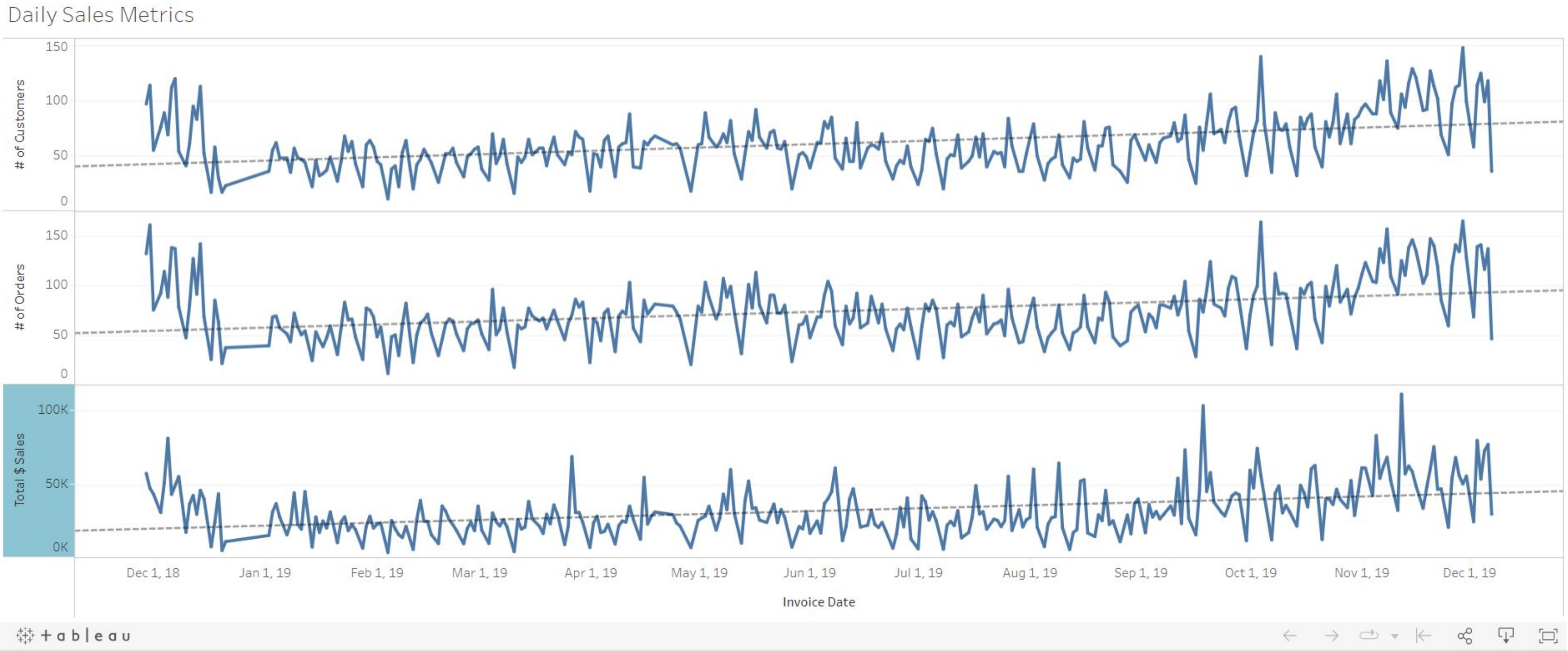
Link: https://public.tableau.com/views/MonthlySales_16598862693950/MonthlySales?:language=en-US&:display_count=n&:origin=viz_share_link

The dataset coverage is from 2018-11-29 to 2019-12-07. It is expected to have lower sales figures for 11-2018 and 12-2019 periods since these cover only a few days and not the full month. Surprisingly, the sales figures for the period of 12-2018 is relatively low as compared to sales figures in 09-2019, 10-2019 and 11-2019. 11-2019 is the best month.



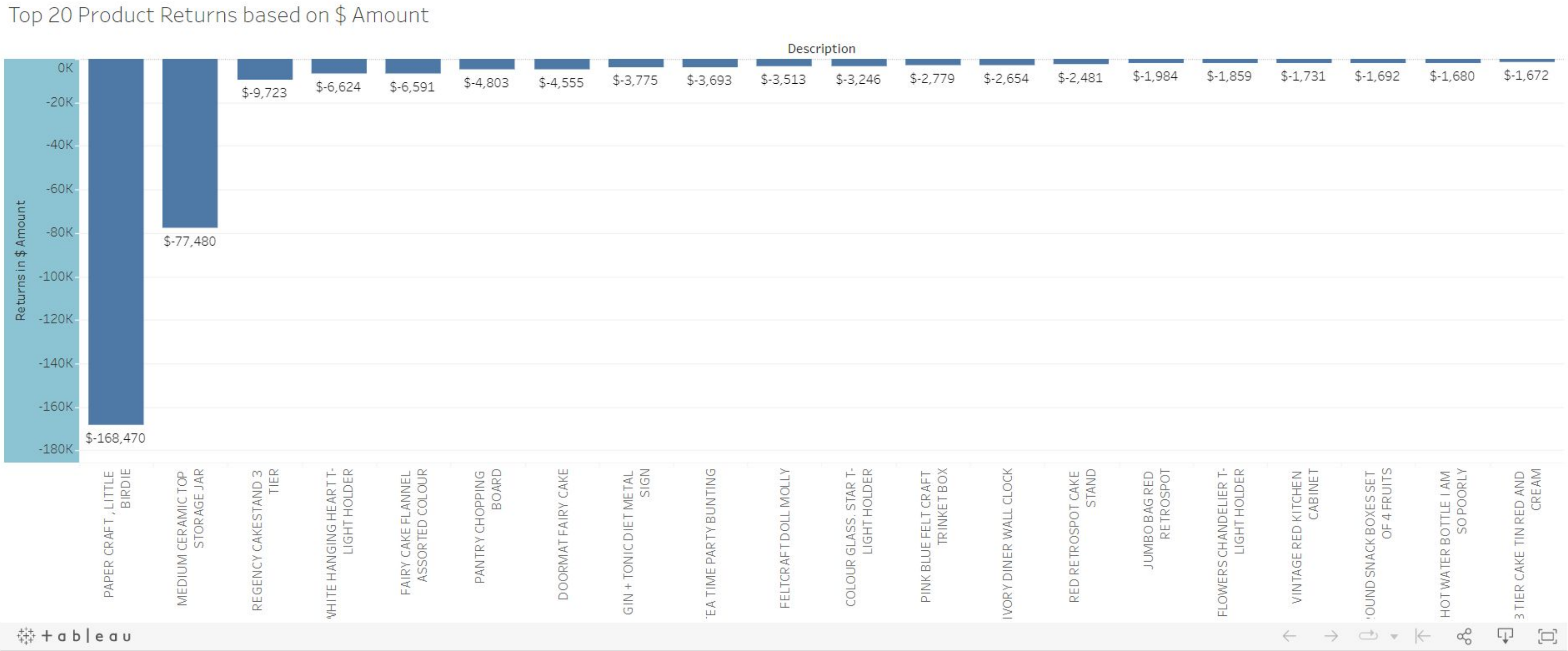
Link: https://public.tableau.com/views/SalesperDayoftheWeek/SalesperDayoftheWeek?:language=en-US&:display_count=n&:origin=viz_share_link

Most of the customers made their purchases during Fridays and Wednesdays but never on Sunday

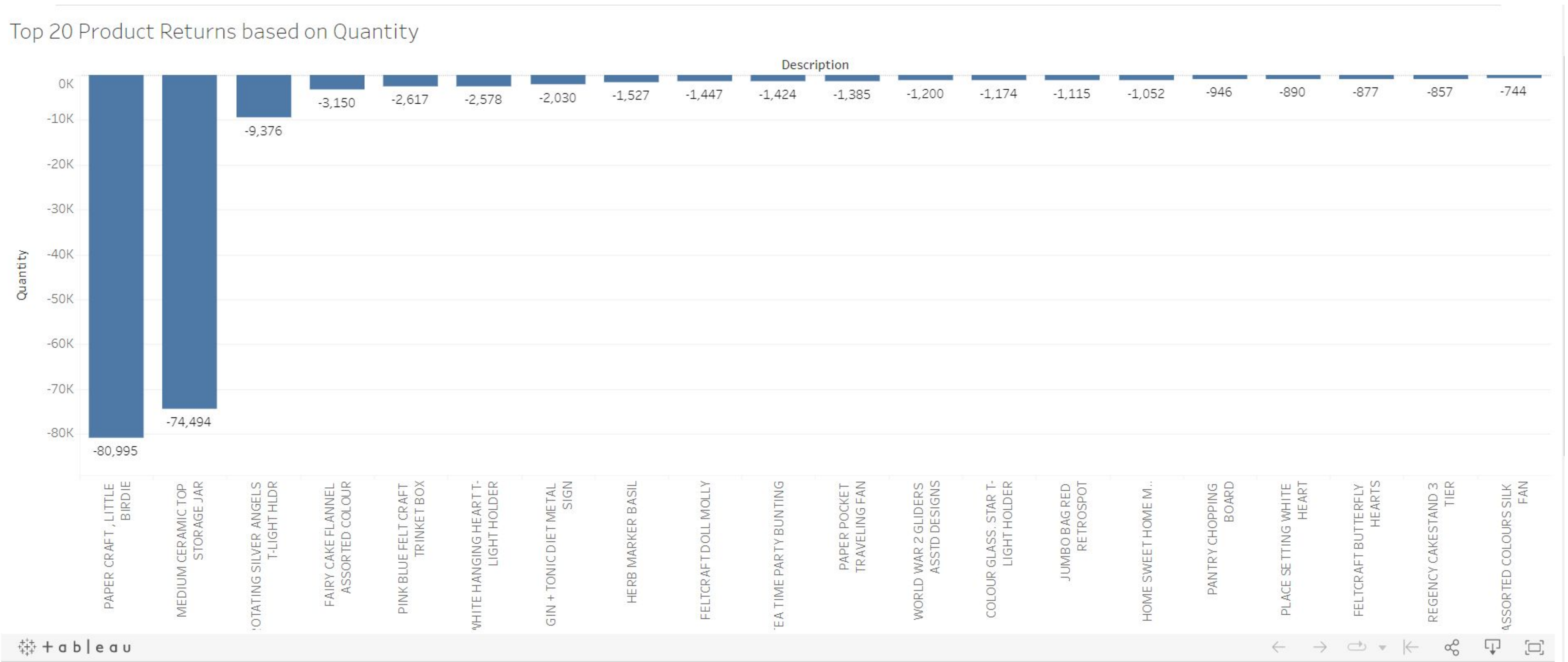


The line charts display the changes for the number of customers, number of orders and total sales over the period of one year. The uptrend line is a good indication that the company's sales performance is increasing over time.

Which products have the highest value and largest number of returns, and which items are most frequently returned?

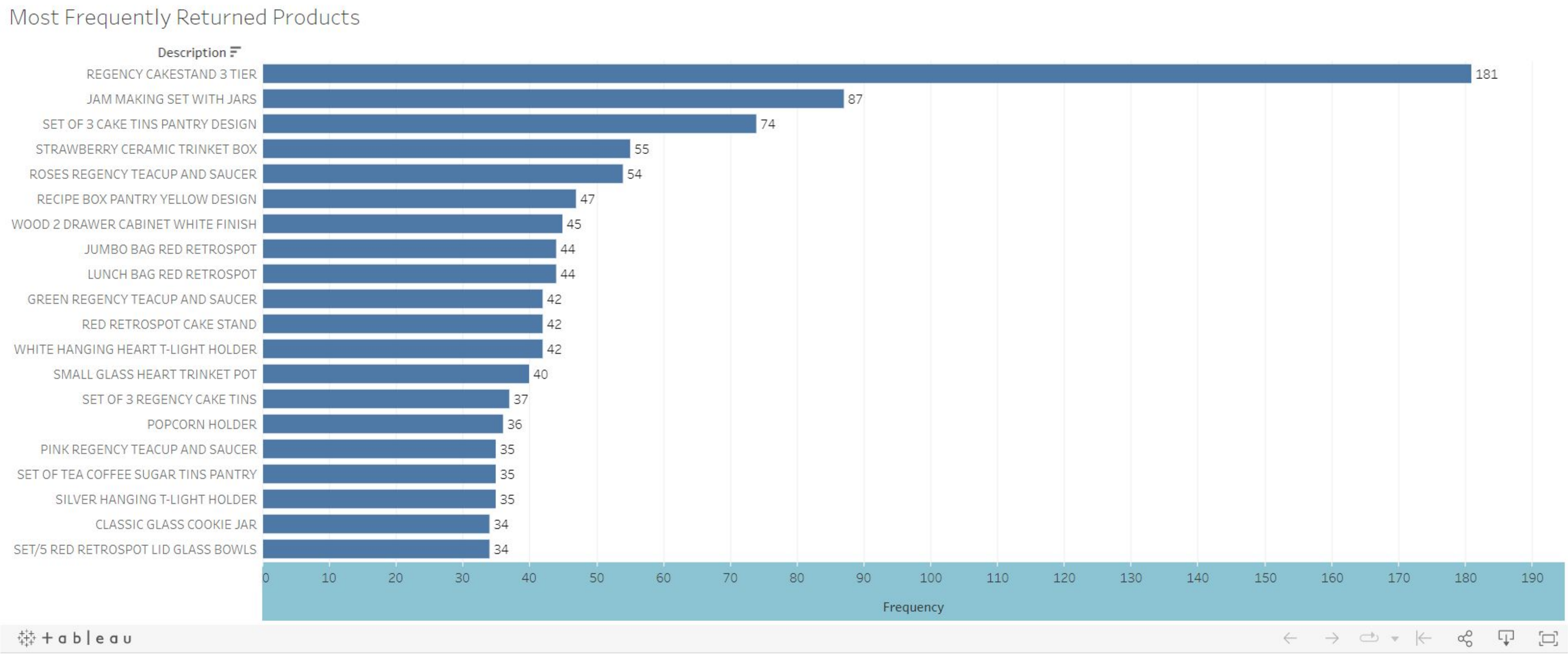


Link: https://public.tableau.com/views/Top20ProductReturnsbasedonAmount/Top20ProductReturnsbasedonAmount?:language=en-US&:display_count=n&:origin=viz_share_link



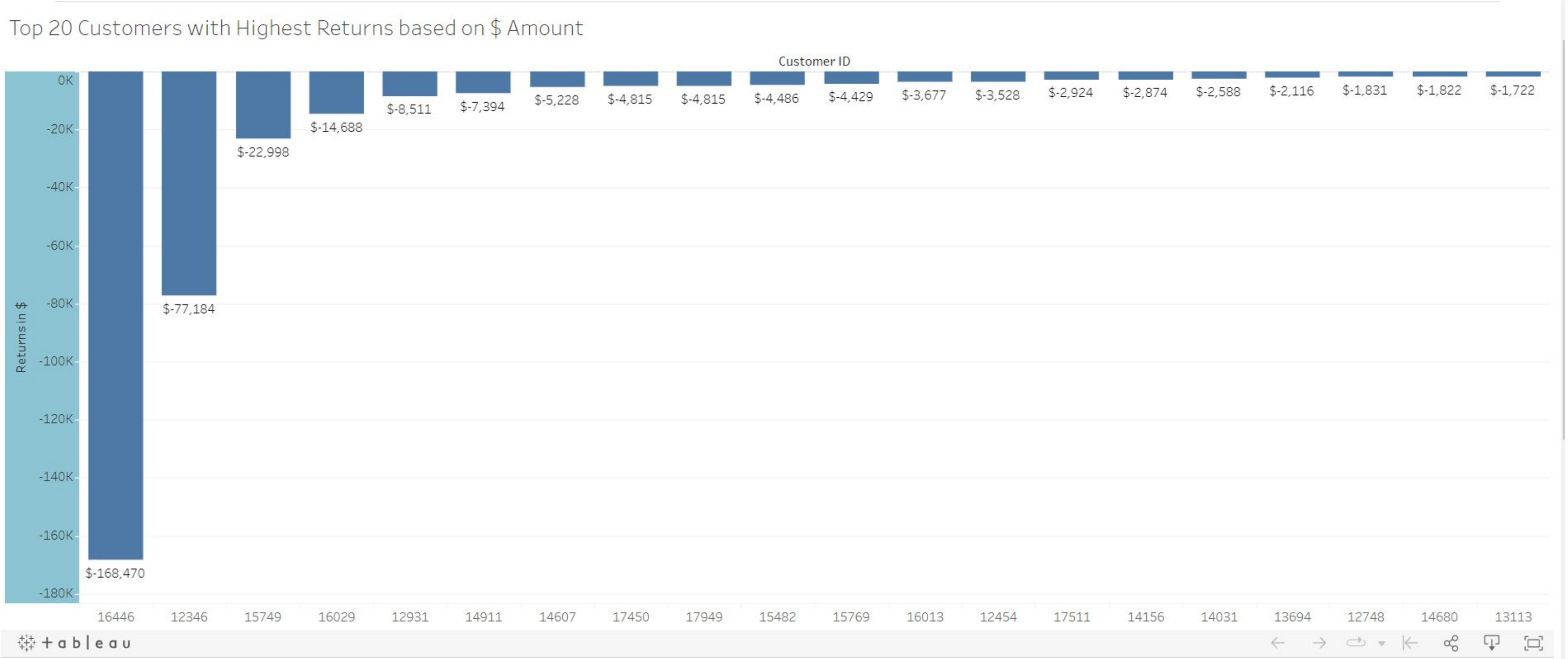
Top 20 Product Returns based on Quantity

Link: https://public.tableau.com/views/Top20ProductReturnsbasedonQuantity/Top20ProductReturnsbasedonQuantity?:language=en-US&:display_count=n&:origin=viz_share_link

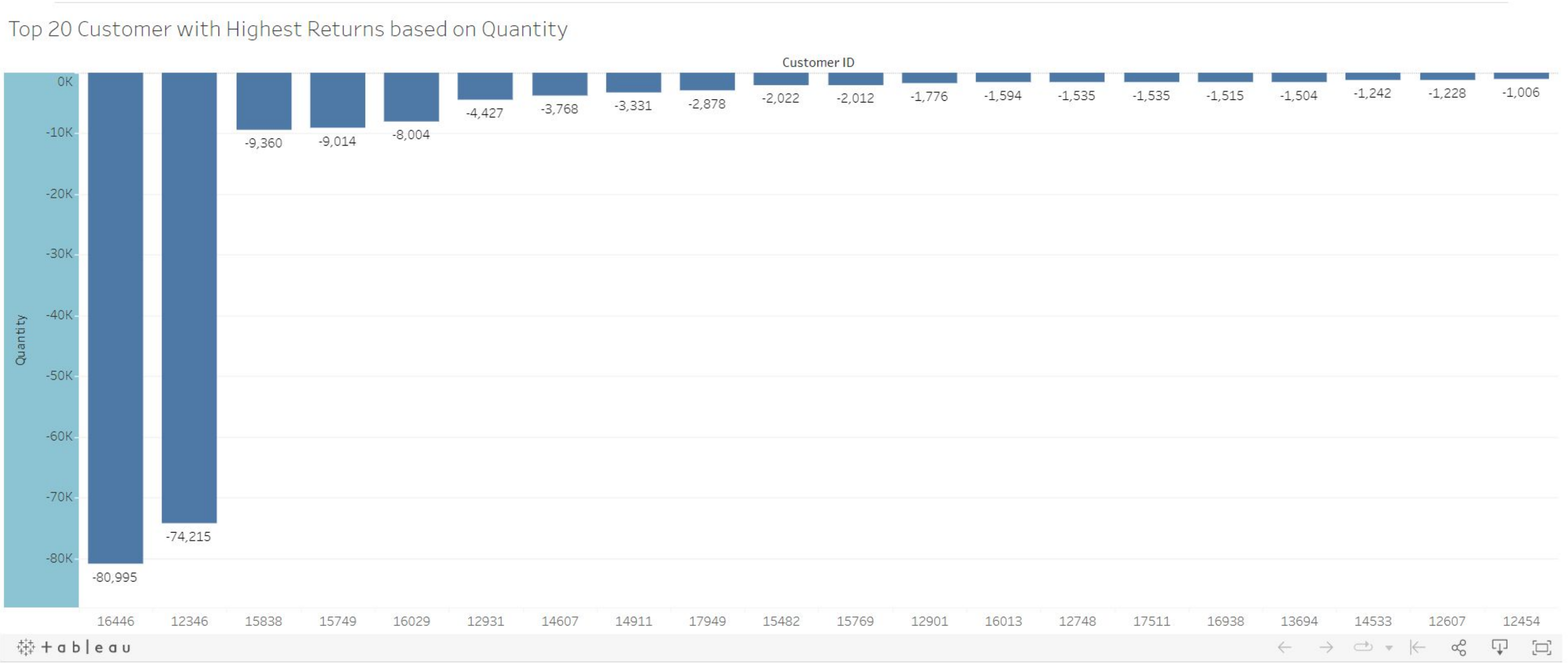


Link: https://public.tableau.com/views/MostFrequentlyReturnedProducts/MostFrequentlyReturnedProducts?:language=en-US&:display_count=n&:origin=viz_share_link

Who are the customers with the highest value and largest number of returns?



Link:
https://public.tableau.com/views/Top20CustomerswithHighestReturnsbasedonAmount/Top20CustomerswithHighestReturnsbasedonAmount?:language=en-US&:display_count=n&:origin=viz_share_link

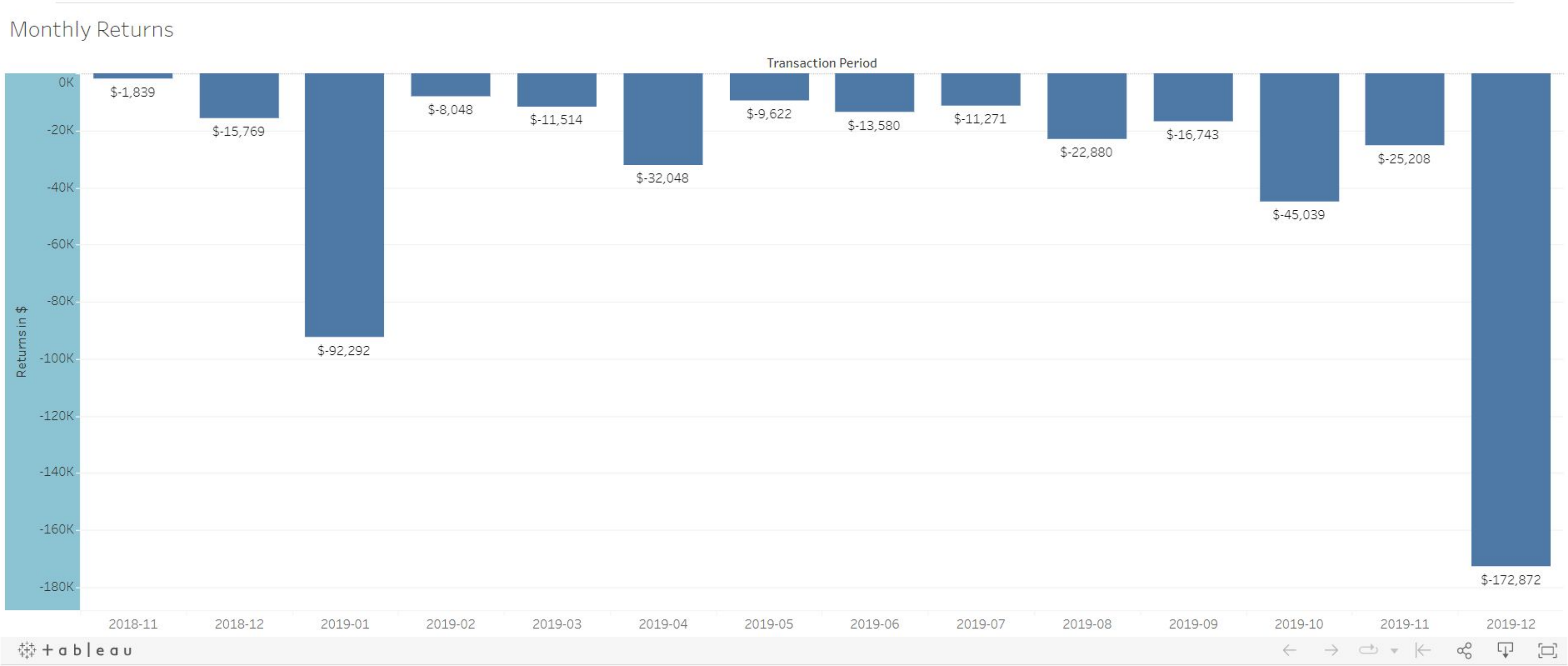


Top 20 Customer with Highest Returns based on Quantity

Link:
https://public.tableau.com/views/Top20CustomerwithHighestReturnsbasedonQuantity/Top20CustomerwithHighestReturnsbasedonQuantity?:language=en-US&:display_count=n&:origin=viz_share_link

What are the products and who are the customers who have significant impact on returns?

- **PAPER CRAFT; LITTLE BIRDIE** is the product with the **highest value** and **largest number** of returns. The total amount returned is **-$168,470** and the total quantity returned is **80,995**
- **REGENCY CAKESTAND 3 TIER** is the top product based on total sales. However, it is also the **most frequently returned product**. It was returned **181 times**.
- Customer **16446** is the customer with the **highest value** and **largest number** of returns. This is the customer who returned **80,995** pieces of the product **PAPER CRAFT; LITTLE BIRDIE with total returns amount of -$168,470**.



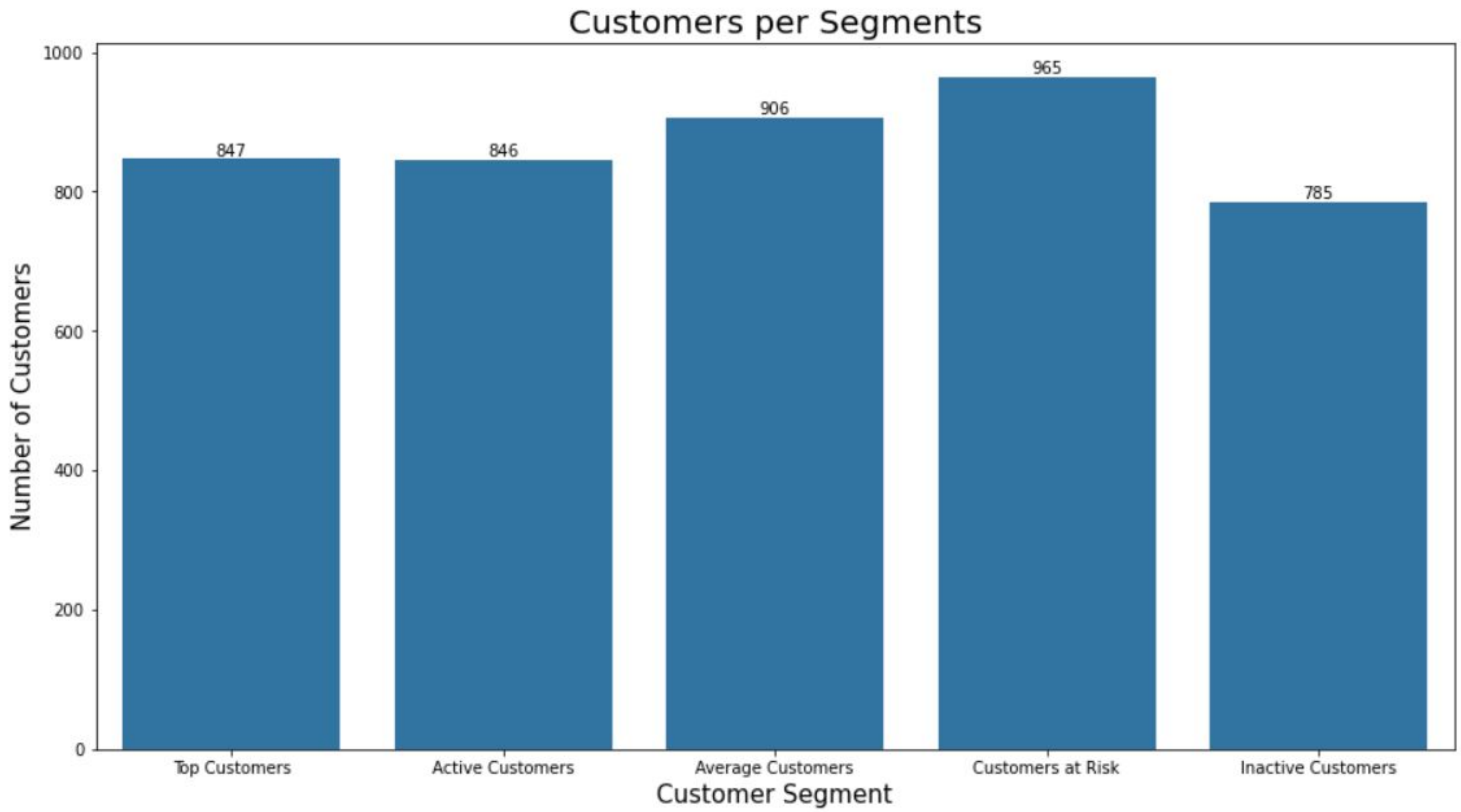
Link: https://public.tableau.com/views/MonthlyReturns/MonthlyReturns?:language=en-US&:display_count=n&:origin=viz_share_link

Surprisingly, the period 12-2019 has the highest amount of returns considering that this period has only 7 days. The total amount of returns is -$172,872.

Customer Segmentation using RFM

Based on RFM scores, the customers are grouped into the following segments and here are the recommendations:

- **Top Customers** - Send birthday and anniversary (as being customer) cards with discount vouchers. Create a referral program where they can get a discount upon the first purchase of their referrals.
- **Active Customers** - Create loyalty rewards program for this group where they can earn point for every purchase and convert these points into discount or voucher.
- **Average Customers** - This group of customers are not consistently inactive or active. Send them emails with discounted items or promotional items for every occasion like their birthdays, mother's day, father's day, Christmas etc.
- **Customers at Risk** - This group of customers need attention as they made some purchases but it's been a long time since their last purchase. Send them personalized emails as well containing promotional items or free samples of products to try to encourage them to be more active.
- **Inactive Customers** - Send them personalized emails containing discounted items to encourage them to order and be active.



The combined proportion of top, active and average customers is almost 60%. This is not bad as 60% of the customer base are making their purchases in a manner that can sustain the online store business.

However, the 40% requires attention. These customers need to be reactivated and encouraged to make more purchases and frequently.

Product Segmentation using RFM

The products were grouped into 3 clusters based on RFM and K-Means clustering:

- **Cluster 0** - is the **worst-performing product group** with average **recency of 160 days**, average **frequency of 10 times** and average **sales of $167**.
- **Cluster 1** - is the **best-performing product group** with average **recency of 2.9 days**, average **frequency of 306 times**, and average **sales of $6,231.**
- **Cluster 2** - is quite good and not bad with average **recency of 15 days**, average **frequency of 58 times** and average **sales of $655**.

Recommendations for each cluster:

- **Cluster 0** - Review the products in this cluster, it probably needs to have more assortment or variety, current stocks are maybe in incomplete sizes or colors and maybe the products are out of season or outdated. Recency of 160 days is like more than 5 months since the last time the products were purchased. Frequency of only 10 times is like the products are being purchased not even once a month. Sales of 167 throughout the year is very low.

- **Cluster 1** - For Cluster 1 products, probably add more variety and increase the quantities. Monitor the inventories so these products won't go out of stock. Continue whatever the existing advertising or marketing campaigns for these products.

- **Cluster 2** - For Cluster 2, with recency of 15, it is good. It means the products are marketable or saleable currently. However, the frequency of 58 times is quite rare and sales of 655 is low, it's like the products are being purchased about 5 times a month. This can be increased by creating more advertising and promotional strategies to target specific customer demographics.

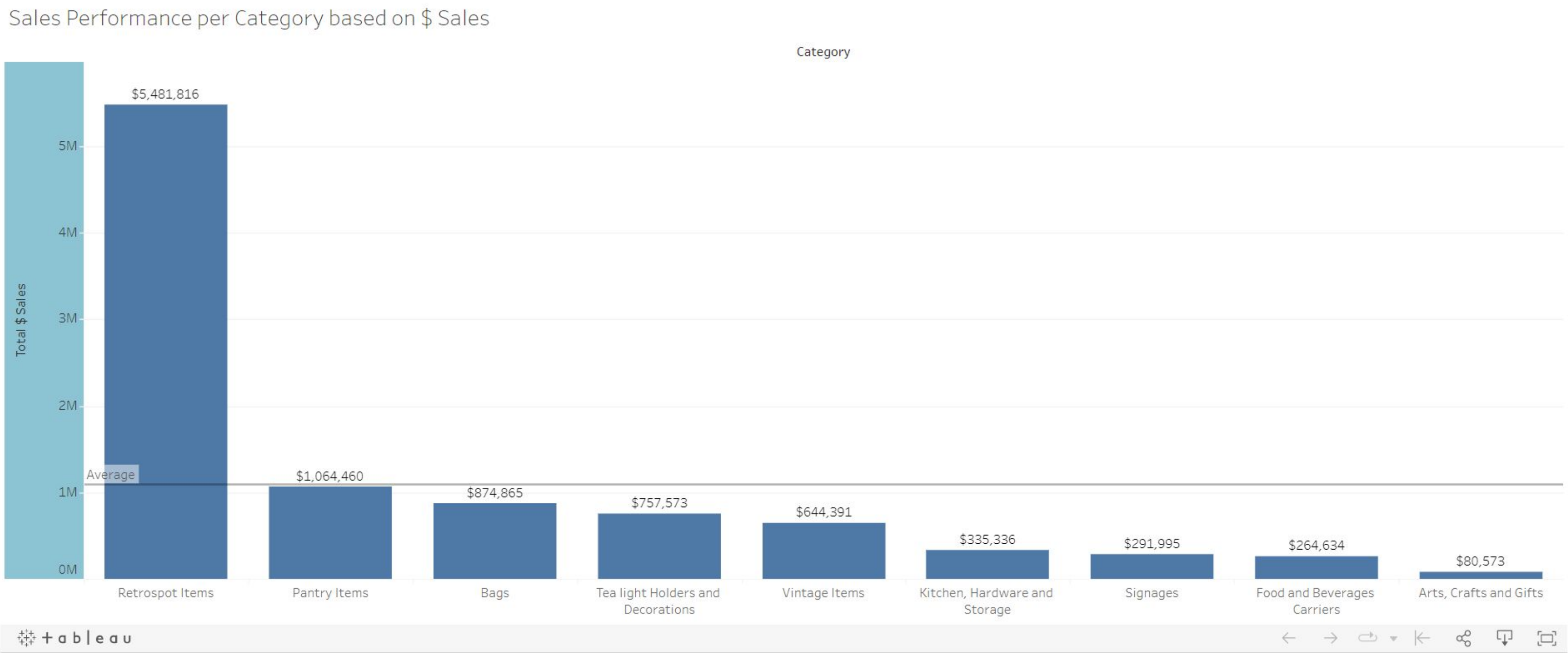
Product Categorization | Product Category Analysis

Product categorization is important as the quality of product analysis depends heavily on the ability to accurately cluster similar products. To create categories, the following were performed on the texts in the product description.

- Text-preprocessing - lowercasing all letters and removing stop words.
- Text vectorization algorithm - NLP TF-IDF Vectorizer to transform text into vectors.
- Clustering - K-Means (Elbow method) to identify the number of clusters and to cluster the texts.
- Most important words - WordCloud to identify the most important/common words in each cluster.

Here are the category names created:

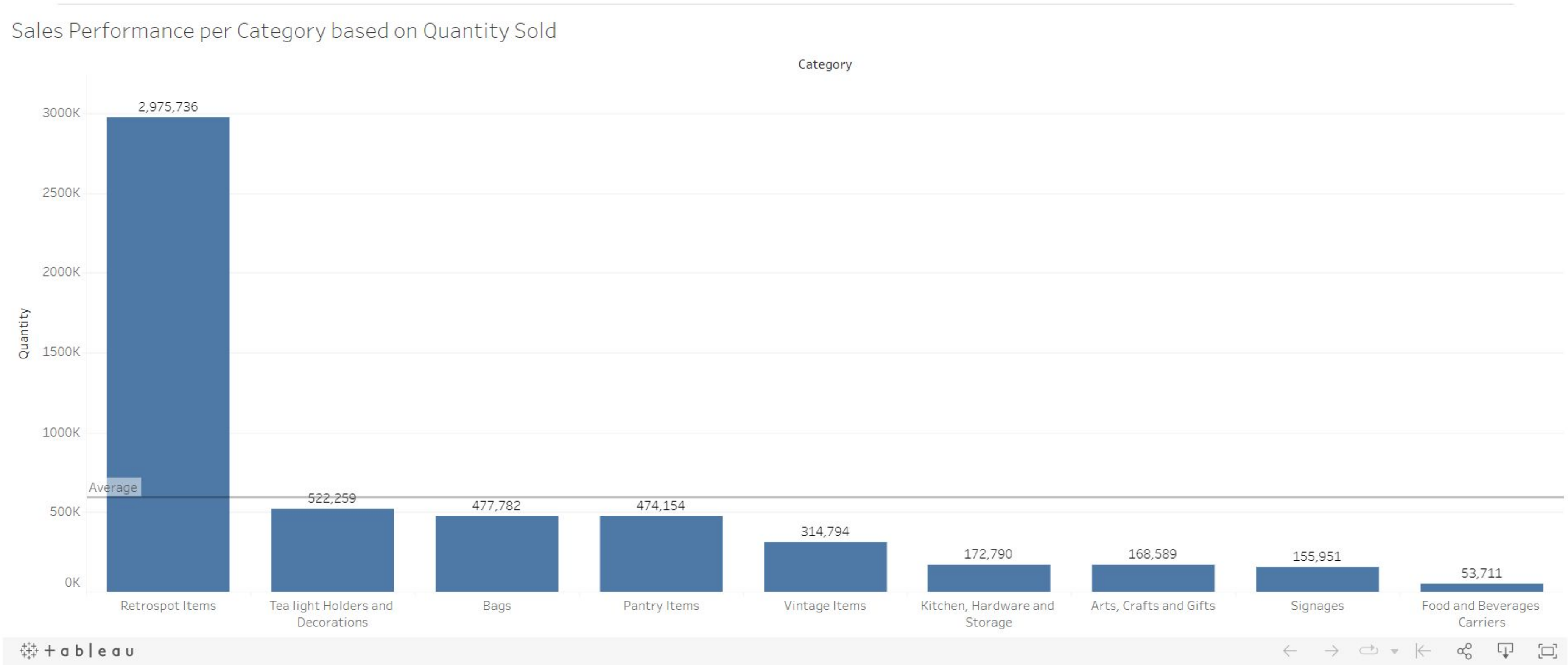
- Cluster 0 - Retrospot Items
- Cluster 1 - Tea light Holder and Decorations
- Cluster 2 - Arts, Crafts and Gifts
- Cluster 3 - Kitchen, Hardware and Storage
- Cluster 4 - Signages
- Cluster 5 - Pantry Items
- Cluster 6 - Vintage Items
- Cluster 7 - Bags
- Cluster 8 - Food and Beverage Carriers



Sales Performance per Category based on $ Sales

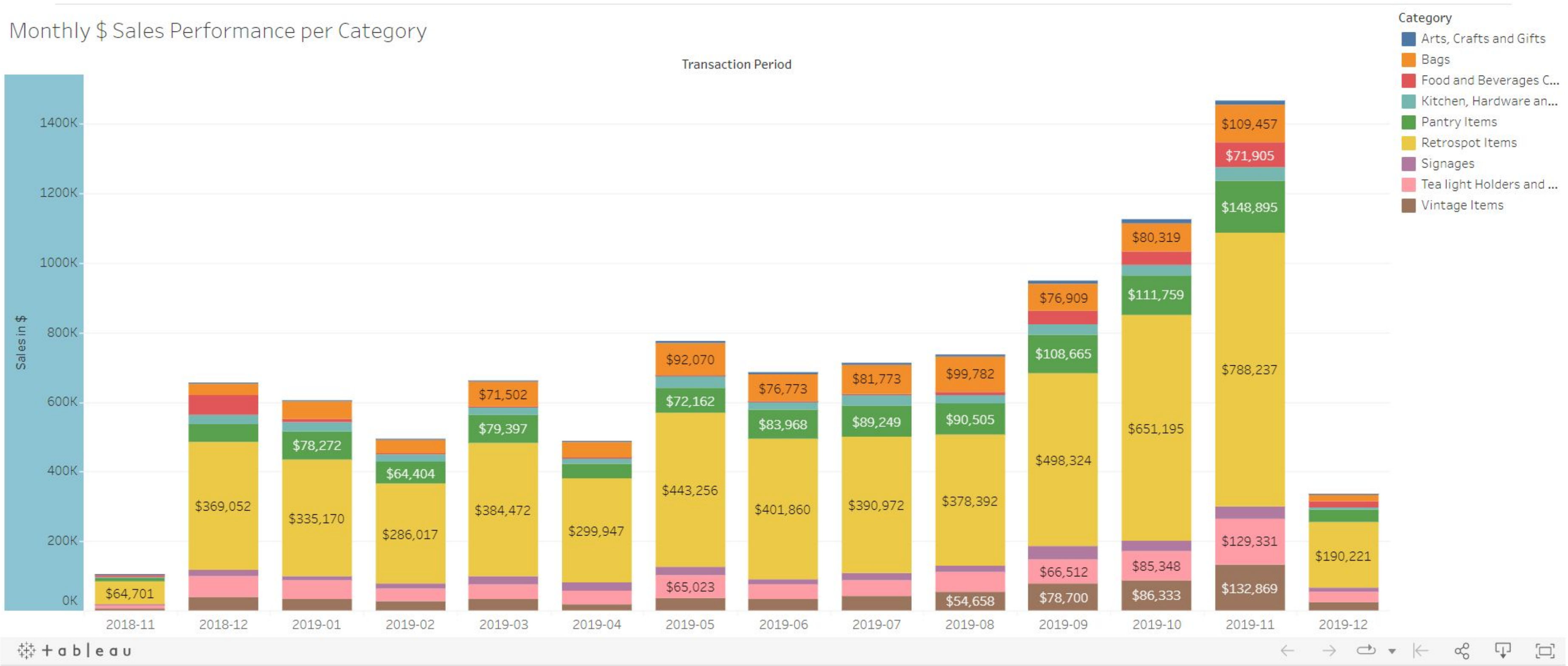
Link: https://public.tableau.com/shared/QH4RG84HT?:display_count=n&:origin=viz_share_link

Retrospot items is the category that contributes significantly to total sales. More than half of the total sales are from Retrospot items. The worst-performing is Arts, Craft and Gifts category



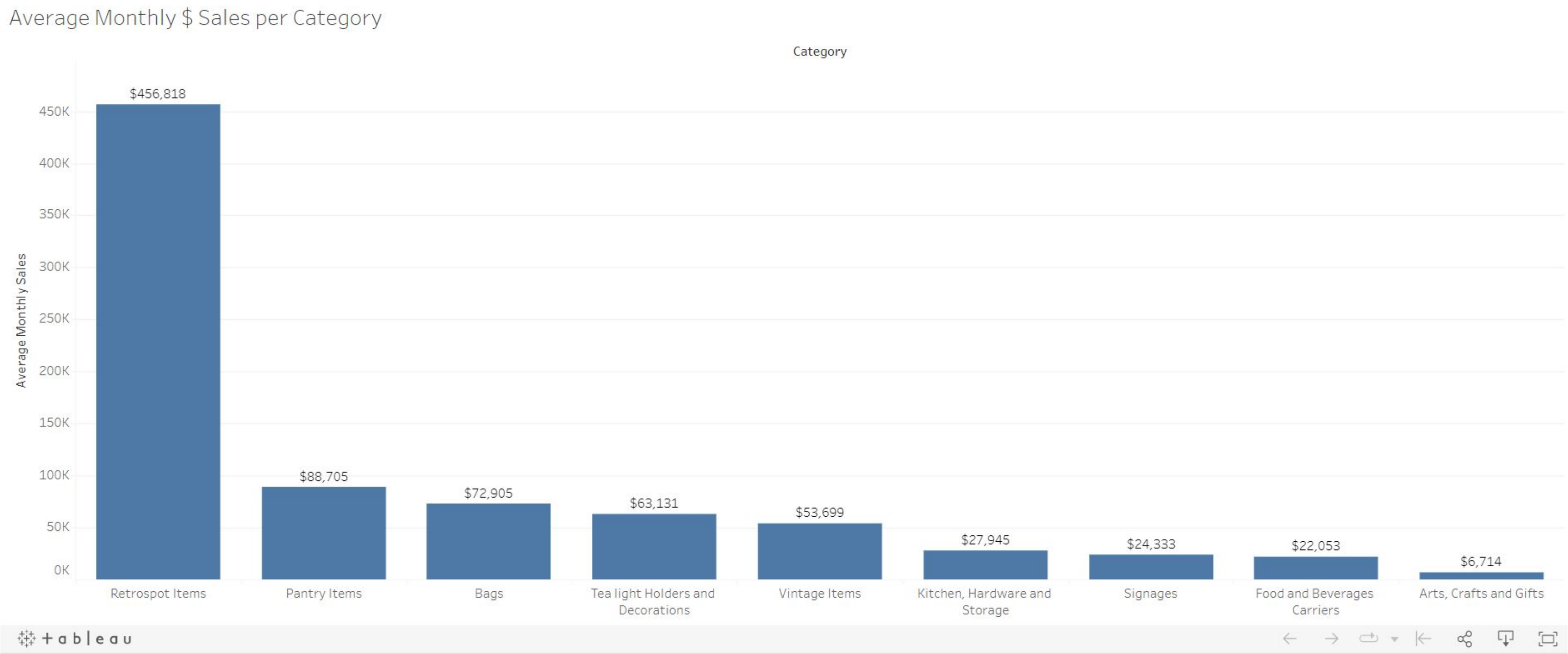
Link:
https://public.tableau.com/views/SalesPerformanceperCategorybasedonQuantitySold/SalesPerformanceperCategorybasedonQuantitySold?:language=en-US&:display_count=n&:origin=viz_share_link

The top-performing category based on quantity sold is Retrospot items which is almost 56% of the total quantity sold. The worst-performing is Food and Beverages category.



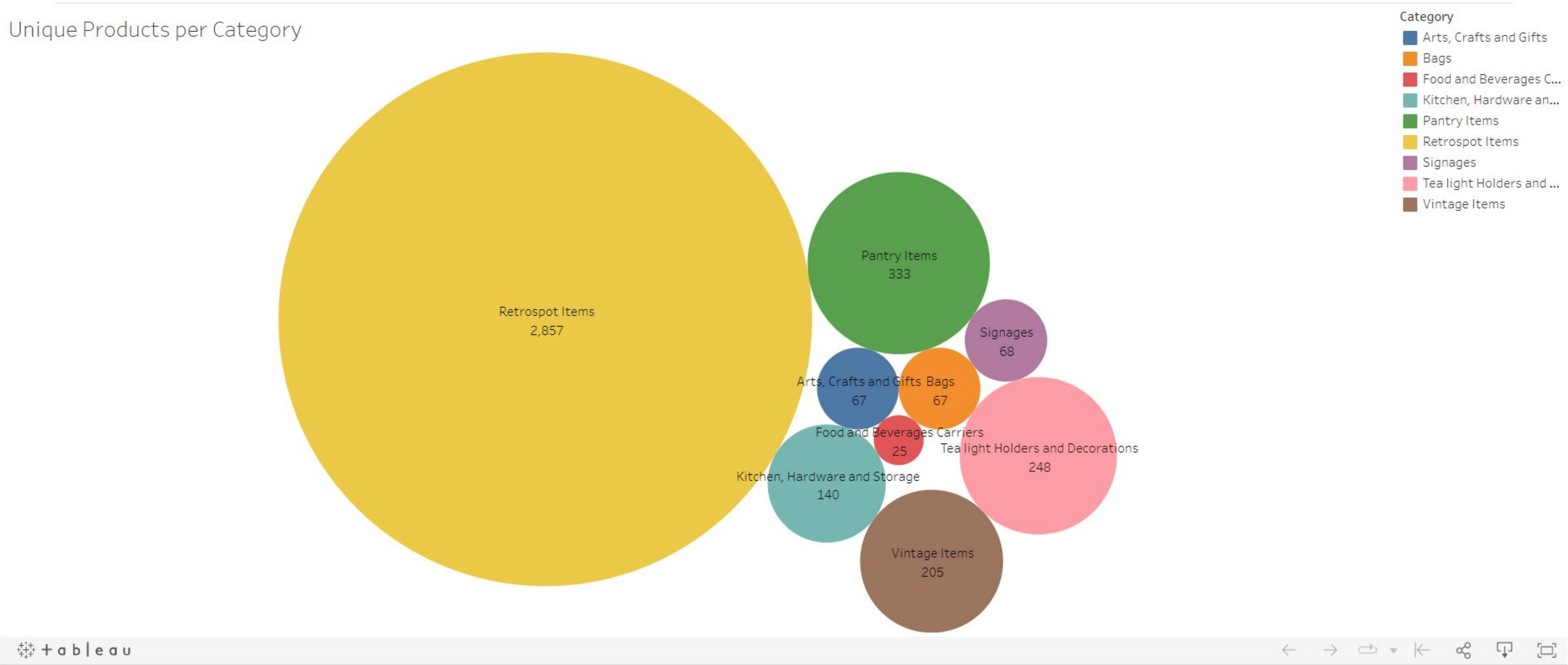
Link: https://public.tableau.com/views/MonthlySalesPerformanceperCategory/MonthlySalesPerformanceperCategory?:language=en-US&:display_count=n&:origin=viz_share_link

Retrospot items is the leading category every month. Retrospot's highest sales is for the period of 11-2019



Link: https://public.tableau.com/views/AverageMonthlySalesperCategory/AverageMonthlySalesperCategory?:language=en-US&:display_count=n&:origin=viz_share_link

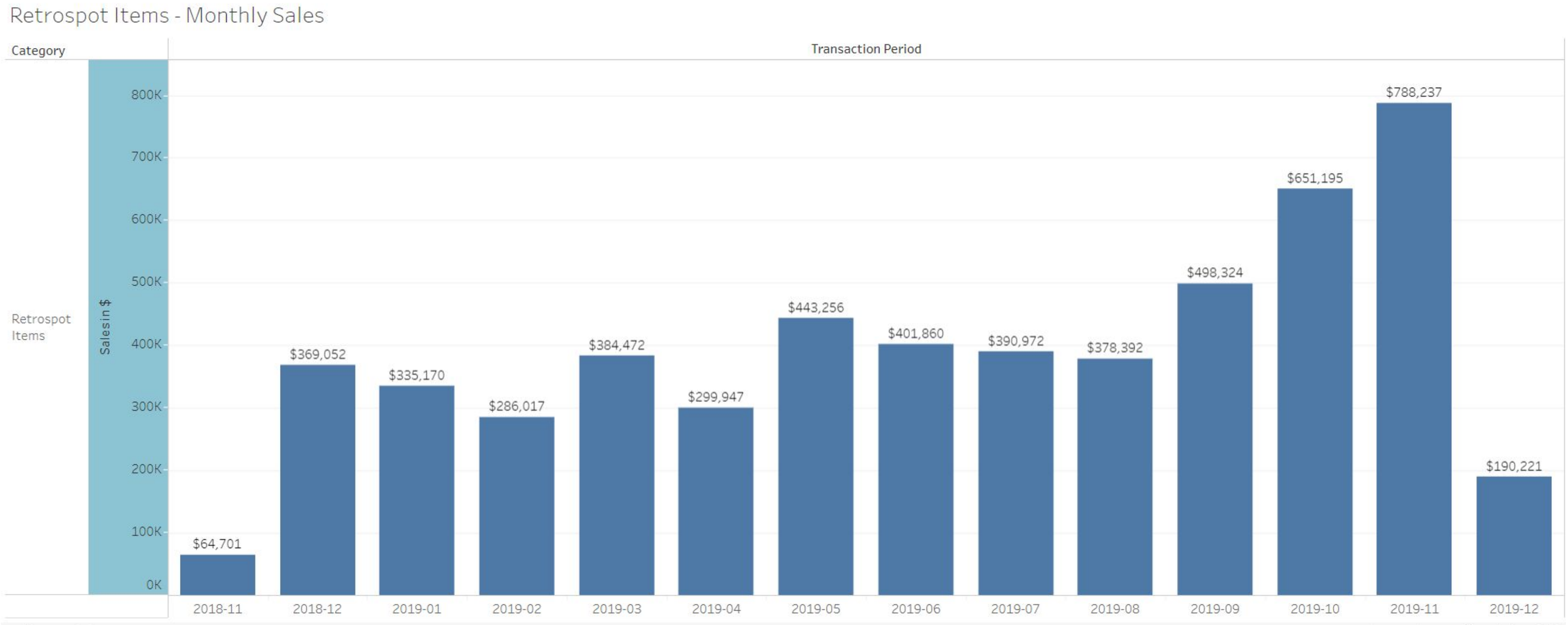
Retrospot items has the highest average monthly sales and Arts, Crafts and Gifts category has the lowest.



Link: https://public.tableau.com/views/UniqueProductsperCategory/UniqueProductsperCategory?:language=en-US&:display_count=n&:origin=viz_share_link

And the reason for having Retrospot items as the top-performing category, is that it has the largest variety of products with 2857 unique items.

Retrospot Items



Retrospot Items - Monthly Sales

Link: https://public.tableau.com/views/RetrospotItems-MonthlySales/RetrospotItems-MonthlySales?:language=en-US&:display_count=n&:origin=viz_share_link

Statistical Hypothesis 1:
High-priced products contribute to higher sales than the low-priced products

Null Hypothesis: There is no difference between the mean sales of low-priced products and high-priced products
Alternative Hypothesis: Mean sales of low-priced products is different than the mean sales of high-priced products

Criteria for Decision: alpha = 0.05

- Accept null hypothesis if p-value > alpha
- Reject null hypothesis if p-value < alpha

Based on the data, we reject the null hypothesis:
Based on the result of statistical test and excluding the outliers, we found out that there is a difference between the average sales of low-priced products and average sales of high-priced products. Low price values are unit prices below 20 and high price values are unit prices 20 and above. This could also mean that the average sales of low-priced products can be higher than the average sales of high-priced products or vice-versa.

Statistical Hypothesis 2:
Sales are higher during Christmas season (December) as compared to other months

Null Hypothesis: There is no difference between the average sales in December and the average sales in other months
Alternative Hypothesis: The average sales in December is different than the average sales in other months

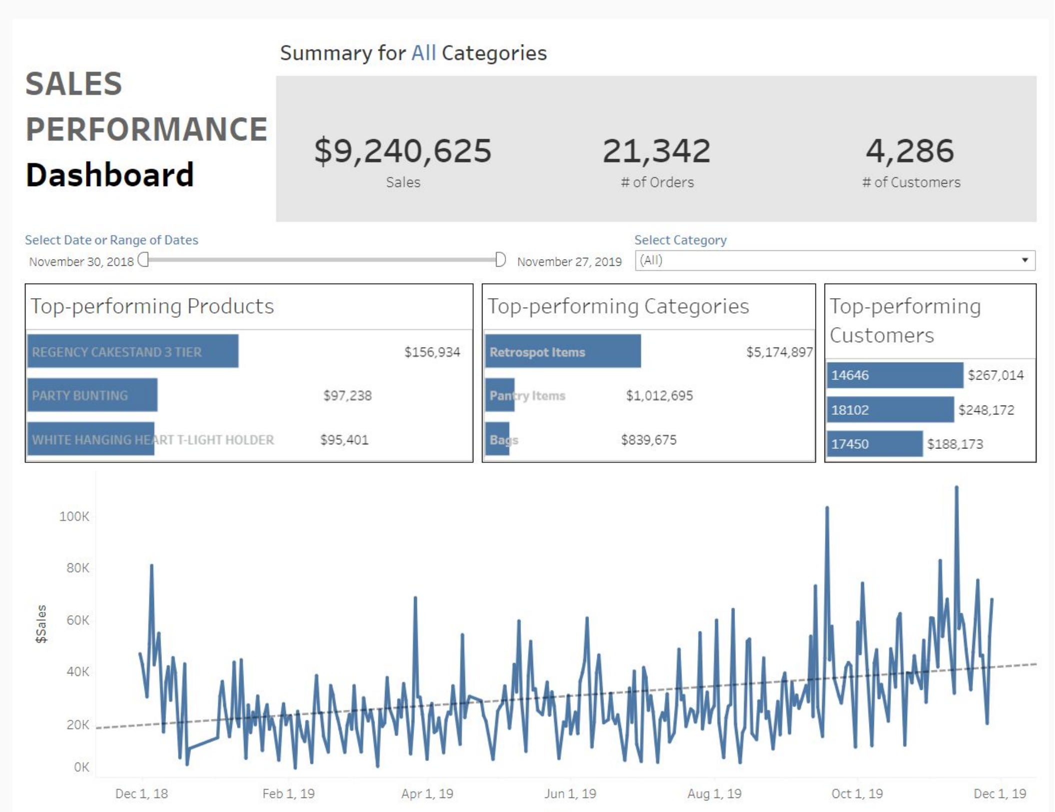
Criteria for Decision: alpha = 0.05

- Accept null hypothesis if p-value > alpha
- Reject null hypothesis if p-value < alpha

Based on the data, we reject the null hypothesis:
Based on the result of statistical test and excluding the outliers, we found out that there is a difference between the average sales generated in December and average sales in other months. This could also mean that the average sales in December can be higher than the average sales in other months or vice-versa.

Dashboard

Link: https://public.tableau.com/views/E-commerceDashboard_16599549421270/Dashboard2?:language=en-US&:display_count=n&:origin=viz_share_link



SALES PERFORMANCE **Dashboard**

Summary for All Categories

$9,240,625
Sales

21,342
# of Orders

4,286
# of Customers

Select Date or Range of Dates
November 30, 2018 – November 27, 2019

Select Category
(All)

Top-performing Products

REGENCY CAKESTAND 3 TIER	$156,934
PARTY BUNTING	$97,238
WHITE HANGING HEART T-LIGHT HOLDER	$95,401

Top-performing Categories

Retrospot Items	$5,174,897
Pantry Items	$1,012,695
Bags	$839,675

Top-performing Customers

14646	$267,014
18102	$248,172
17450	$188,173

Recommendations

Returns

For e-commerce business, customer returns are unavoidable. That’s why it is important for any e-commerce companies to have a good and fair return policy. Returns are a cost factor. It negatively affects revenues and profit. For this reason, it is recommended for the company to perform monthly returns analysis. Analysis of returns is a collection of data about the return of a product. It helps to understand how big or small the impact of returns, find out the cause and manage it.

Customer Segmentation

It is recommended to perform customer segmentation using RFM scores. It helps the business to analyze their existing customers in terms of their spending habits. When customers are segmented, it is easier to target specific customers with communications or marketing strategies that are relevant based on data about a particular set of customer behaviors.

Product Segmentation

It is also recommended to segment the products into clusters based on the product’s performance. It will help to analyze which products are currently marketable and which are generating high or low revenues.

Product Categorization

Product categorization plays an important role specifically for e-commerce. The quality of product analysis depends heavily on the ability to accurately cluster similar products. It also helps customers intuitively find what they want from the website. If products are not categorized properly, it may result to loss of sales opportunity or loss of customers as they couldn’t find the items that they want in the category they’re looking at.

Limitations

- If I would be the future owner of the business or if I am the current owner of the business, I would be more interested on the profit analysis. The top performing products or top performing customers could be different based on profit. Profit is computed by subtracting the costs from the revenue. As the data doesn't have information about costs, the analysis is limited only on the unit price, quantity and sales amount.

- The dataset is only a year of sales transactions history. It would be better for comparison analysis if it is at least 3 years sales transaction history. Like we can compare the performance of the store or the products every year.

- There is no information about the cancellation or return policy. The dataset has a significant number of cancellations/returns and it would be easier to understand how to manage those if there were additional information.

- There is no information about the store's pricing policy. As I have observed, some stockcodes have multiple prices which I assume, it's either a price increase or a distinction between a wholesale price and retail price. Some prices are higher if there is no customer id.

- I wanted to use BERTopic or spacy but for some reason, I had issues with installing bertopic and spacy took a very long time to run and it caused my computer to freeze. I would probably explore more on other NLP or ML tools had I better computer resources and longer time.

Resources

1. To get ideas on how to perform product categorization
https://techblog.commercetools.com/boosting-product-categorization-with-machine-learning-ad4dbd30b0e8
2. How to deal with negative values in the logarithmic transformation to unskew the data
https://campus.datacamp.com/courses/customer-segmentation-in-python/data-pre-processing-for-clustering?ex=4
3. How to interpret the p-value of 0.0
https://www.statology.org/here-is-how-to-interpret-a-p-value-of-0-000/
4. How to find cancellation pairs
https://stackoverflow.com/questions/38831088/remove-cancelling-rows-from-pandas-dataframe
5. To understand what is DOTCOM POSTAGE, DCGS (dotcomgiftshop) and Retrospot in the dataset
https://www.rexlondon.com/blog/dotcomgiftshop-changing-to-rex-london
6. To decide whether to use or not to use lemmatization
https://www.datacamp.com/tutorial/stemming-lemmatization-python
7. How to prepare data for hypothesis testing
https://towardsdatascience.com/hypothesis-testing-the-discount-bump-4b8b6b4c4fec