Data Analytics for a Retail Store

Contents

1.Executive Summary

In today's world of increased competitiveness, fine-tuning processes and effectively utilization of data are an absolute must for success. Retail companies face multiple challenges, notably competitive pricing, meeting expectations of user experience, and profits for concerned stakeholders. This report provides a comprehensive data analysis pertaining to operations of retail store such as tracking popular products, analysis of customers based on region and product preferences etc. For the report, the software tableau was adopted to prepare the visualizations.

2.Introduction

Data analytics plays a critical role in retail by tracking customer behaviour, product trends, and sales distribution. Retailers use customer data for insights into purchase trends, average transaction value, customer engagement, and purchasing frequency. Effective data analysis provides these insights, empowering retail operations to optimize stock levels, pricing strategies, and marketing efforts. This project leverages **customer intelligence**—the practice of using data-driven insights from historical and predictive behaviour—to help the retail store understand customer needs and enhance sales strategy.

3.Problem Statement

The objective is to help a retail store analyze daily transactions and monitor customers across different locations and product categories. By integrating and analyzing transaction data, this report will answer key business questions, such as identifying popular products among different demographics, monitoring high-performing stores, and evaluating spending patterns.

4.Methodology

In line with the problem statement, the following questions from the reference document were chosen as base for identifying the metrics:

1. Merge the datasets Customers, Product Hierarchy and Transactions as Customer_Final. Ensure to keep all customers who have done transactions with us and select the join type accordingly.
2. Prepare a summary report for the merged data set
    a. Get the column names and their corresponding data types
    b. Top/Bottom 10 observations
    c. “Five-number summary” for continuous variables (min, Q1, median, Q3 and max)
    d. Frequency tables for all the categorical variables
3. Generate histograms for all continuous variables and frequency bars for categorical variables.

4. Calculate the following information using the merged dataset:
    a. Time period of the available transaction data
    b. Count of transactions where the total amount of transaction was negative
5. Analyze which product categories are more popular among females vs male customers.
6. Which City code has the maximum customers and what was the percentage of customers from that city?
7. Which store type sells the maximum products by value and by quantity?
8. What was the total amount earned from the "Electronics" and "Clothing" categories from Flagship Stores?
9. What was the total amount earned from "Male" customers under the "Electronics" category?
10. How many customers have more than 10 unique transactions, after removing all transactions which have any negative amounts?
11. For all customers aged between 25 - 35, find out:
    a. What was the total amount spent for “Electronics” and “Books” product categories?
    b. What was the total amount spent by these customers between 1st Jan, 2014 to 1st Mar, 2014?

4.1. Data Source

The sources of the data were csv as follows (with respective fields):

1. Customer(customer_Id,DOB,Gender,city_code)
2. Transactions(transaction_id,cust_id,tran_date,prod_subcat_code,prod_cat_code,Qty,Rate,Tax,total_amt,Store_type)
3. Prod_cat_info(prod_cat_code,prod_cat,prod_sub_cat_code,prod_subcat)

4.2. Data Pre-processing

No explicit data-preprocessing was carried out for this report generation

4.3. Data-processing

In tableau, to meet the requirements pertaining to the reference document, certain fields and attributes were created. They are as follows:

- total_amt_calculation: A filter to differentiate between positive and negative amount

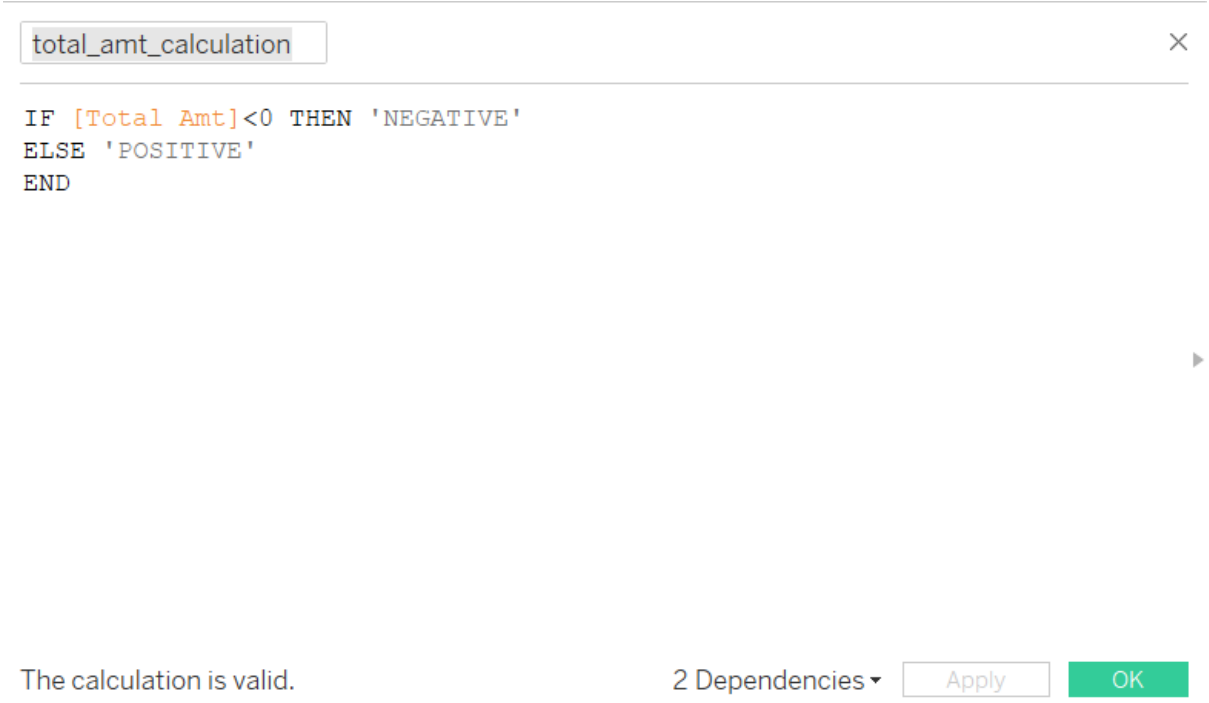


Figure 1:total_amt_calculation

- Range_tran _date: A filter to calculate to calculate the total number of days the transaction was conducted (as per the dataset provided)

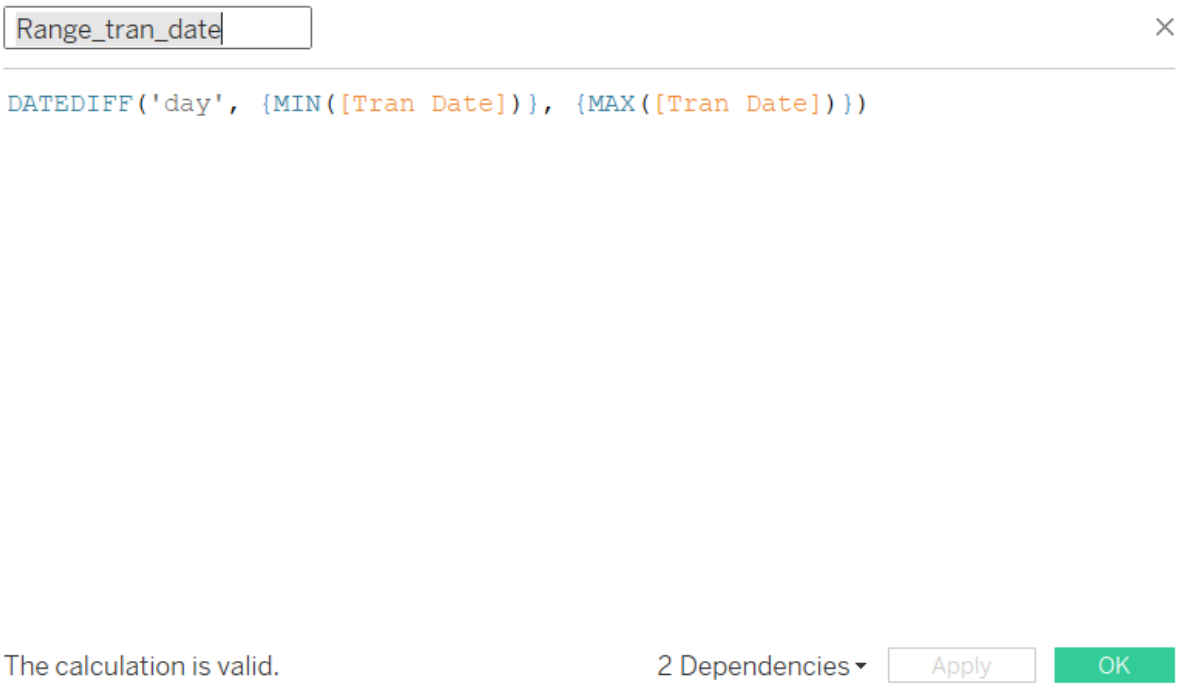


Figure 2:Range_tran_date

- Age: A filter that calculates age of the customers who have conducted a transaction

Figure 3:Age measure

5.Results and Discussion

5.1. Dashboards

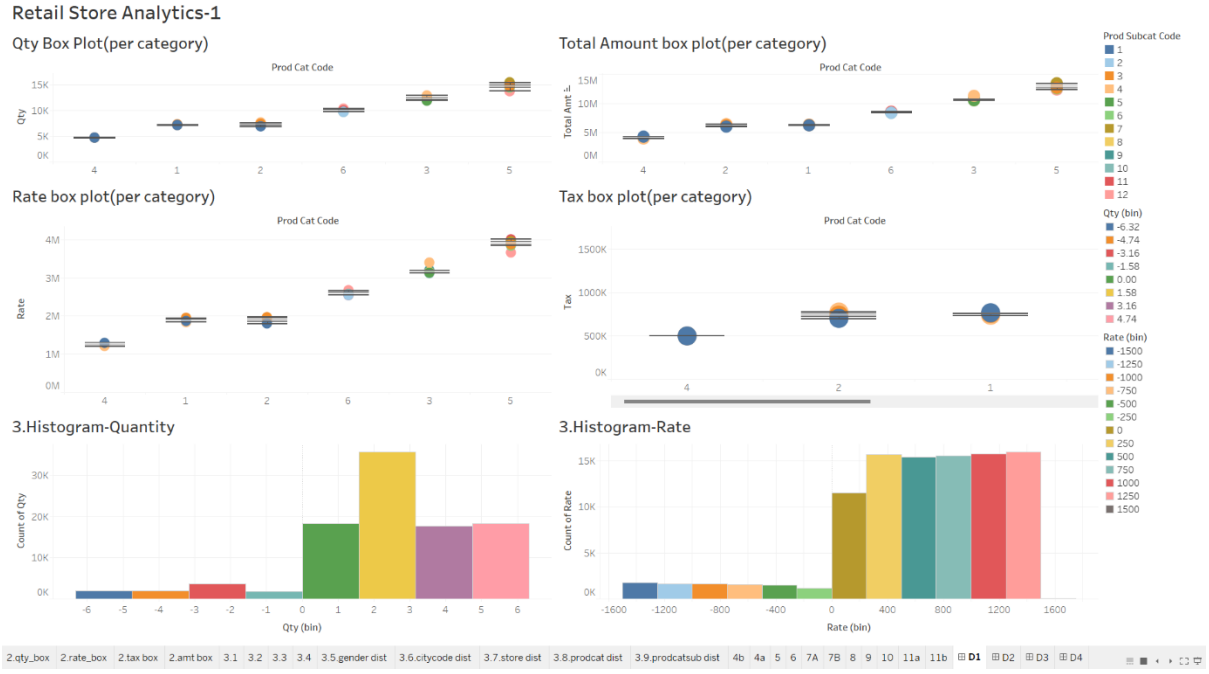


Figure 4:Dashboard1

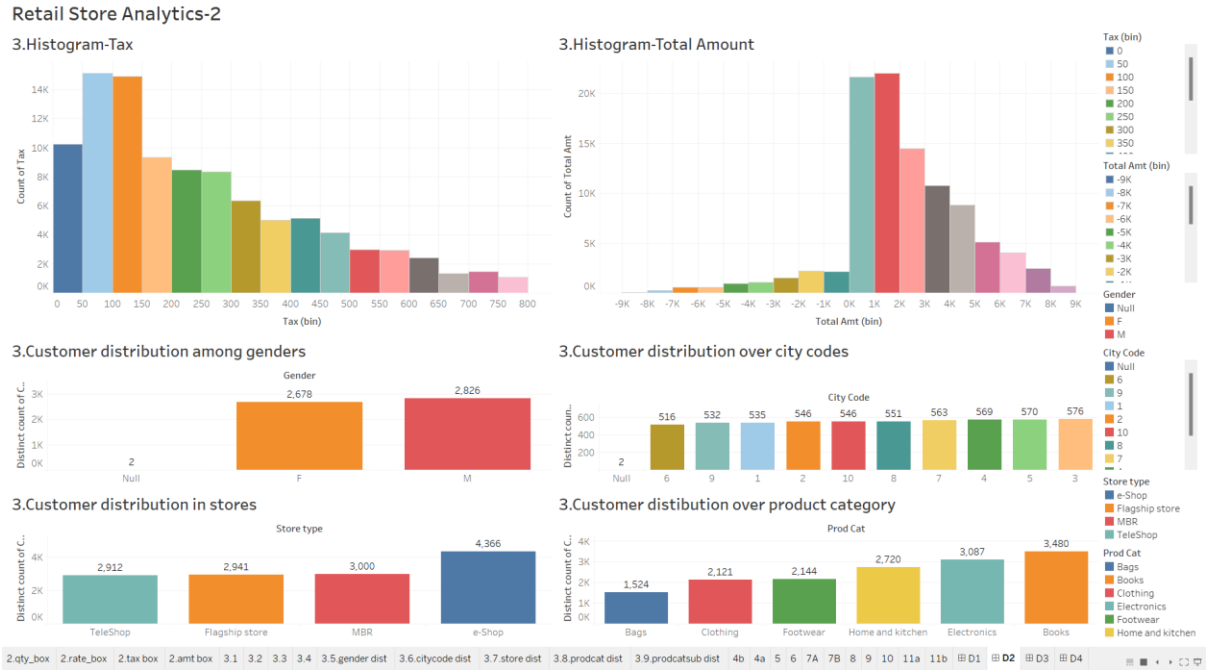


Figure 5:Dashboard 2

Figure 6:Dashboard 3

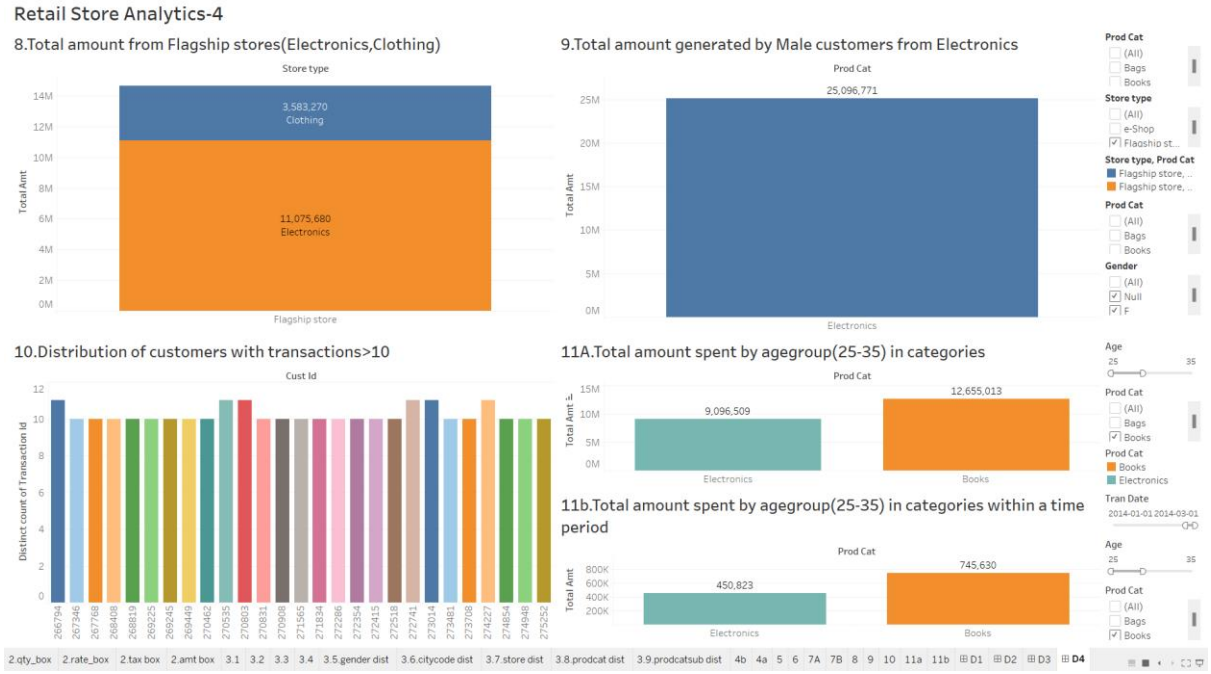


Figure 7:Dashboard-4

5.2. Merging of relevant datasets for appropriate data

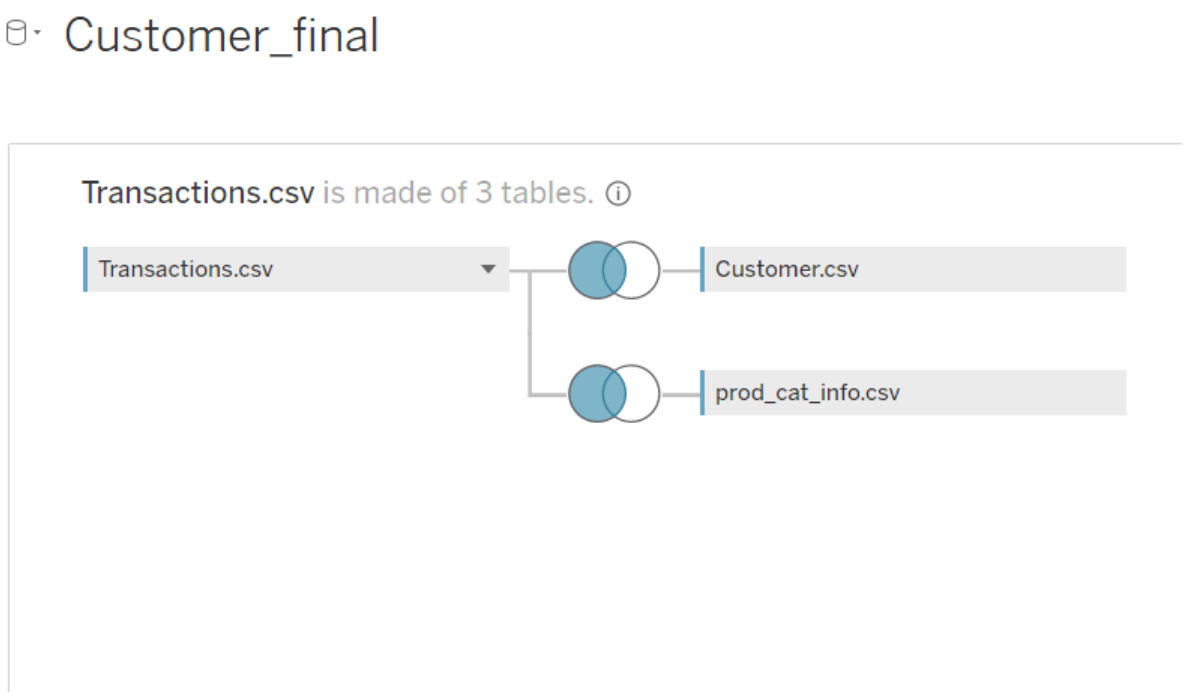


Figure 8:Table join

The Customer, Prod_cat_info, and Transaction csv's were merged using a left join to include all customer transactions. This resulted in the final dataset, Customer_Final, facilitating unified analysis.

Transactions.csv Transaction Id	Transactions.csv Cust Id	Calculation Cust Id (copy)	Calculation COUNTD(CUST ID COPY)	Transactions.csv Tran Date	Calculation max_tran_date	Calculation Min_tran_date	Transactions.csv Prod Subcat Code
46129128420	268627	268,627	1	2014-01-02	2014-01-02	2014-01-02	4
46129128420	268627	268,627	1	2014-01-02	2014-01-02	2014-01-02	4
46129128420	268627	268,627	1	2014-01-02	2014-01-02	2014-01-02	4
32125935023	272067	272,067	1	2014-01-02	2014-01-02	2014-01-02	10
32125935023	272067	272,067	1	2014-01-02	2014-01-02	2014-01-02	10
32125935023	272067	272,067	1	2014-01-02	2014-01-02	2014-01-02	10
32125935023	272067	272,067	1	2014-01-02	2014-01-02	2014-01-02	10
32125935023	272067	272,067	1	2014-01-02	2014-01-02	2014-01-02	10
50346649770	270616	270,616	1	2014-01-02	2014-01-02	2014-01-02	4
50346649770	270616	270,616	1	2014-01-02	2014-01-02	2014-01-02	4

Figure 9:Table generated post join

Transactions.csv Prod Subcat Code	Transactions.csv Prod Cat Code	Transactions.csv Qty	Transactions.csv Rate	Transactions.csv Tax	Transactions.csv Total Amt	Calculation total_amt_calculation	Transactions.csv Store type	Customer.csv customer Id	Customer.csv DOB
4	1	3	515	162.225	1,707.23	POSITIVE	e-Shop	268627	1970-06-01
4	1	3	515	162.225	1,707.23	POSITIVE	e-Shop	268627	1970-06-01
4	1	3	515	162.225	1,707.23	POSITIVE	e-Shop	268627	1970-06-01
10	3	5	1,211	635.775	6,690.78	POSITIVE	e-Shop	272067	1989-05-14
10	3	5	1,211	635.775	6,690.78	POSITIVE	e-Shop	272067	1989-05-14
10	3	5	1,211	635.775	6,690.78	POSITIVE	e-Shop	272067	1989-05-14
10	3	5	1,211	635.775	6,690.78	POSITIVE	e-Shop	272067	1989-05-14
10	3	5	1,211	635.775	6,690.78	POSITIVE	e-Shop	272067	1989-05-14
4	2	2	977	205.170	2,159.17	POSITIVE	MBR	270616	1988-08-01
4	2	2	977	205.170	2,159.17	POSITIVE	MBR	270616	1988-08-01

Figure 10:Table generated post join continued

5.3.Summary report of customer_final dataset

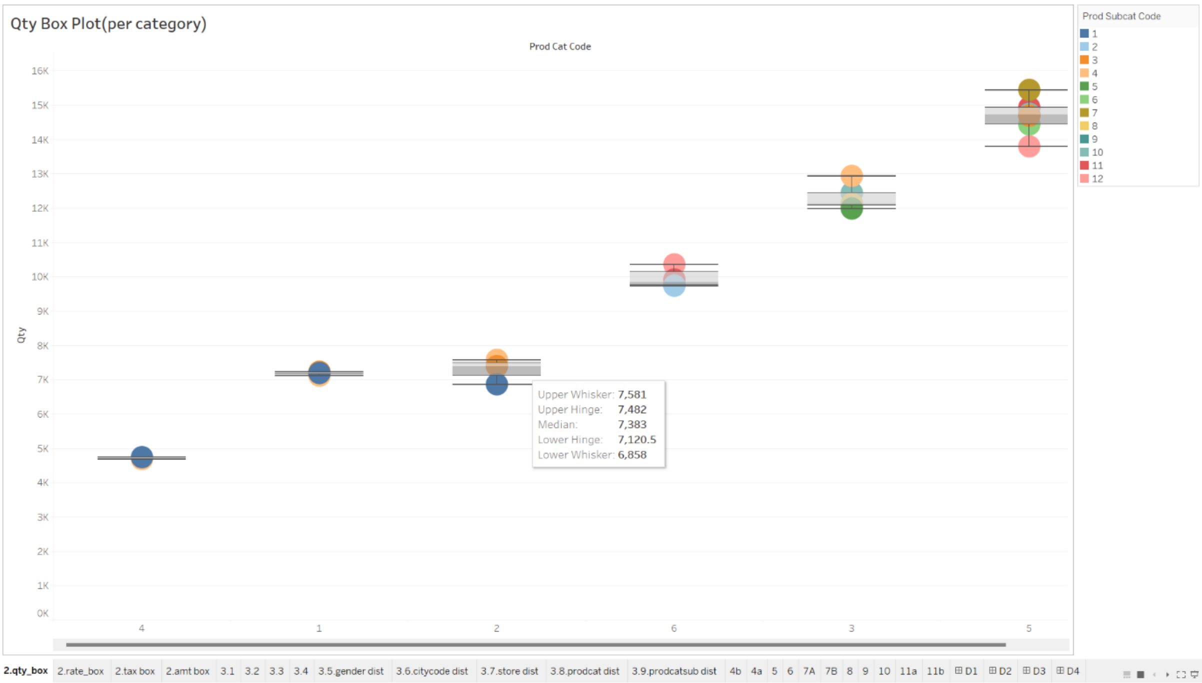


Figure 11:Qty box plot

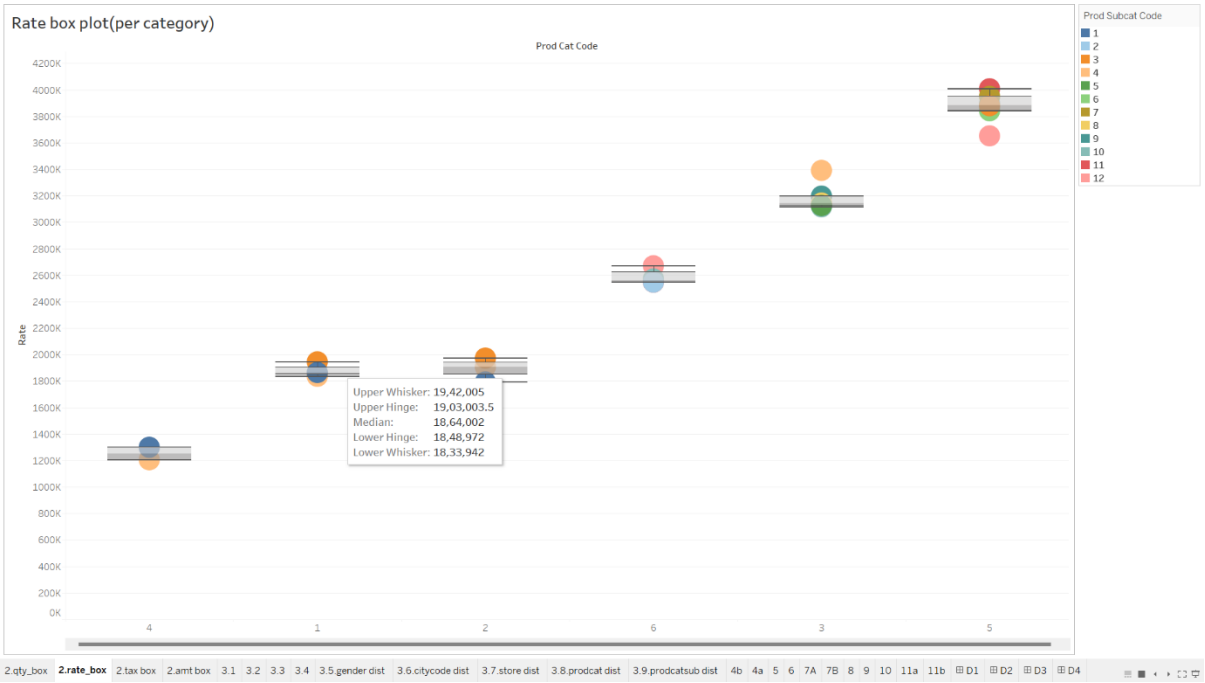


Figure 12:Rate box plot

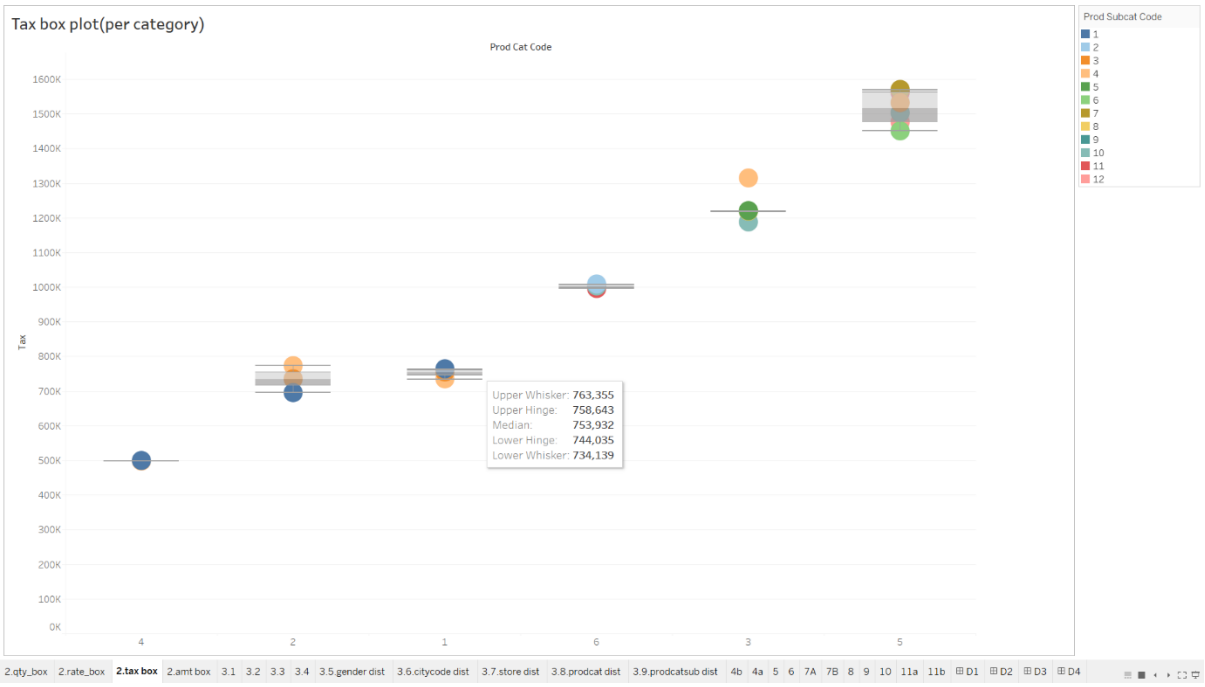


Figure 13: Tax box plot

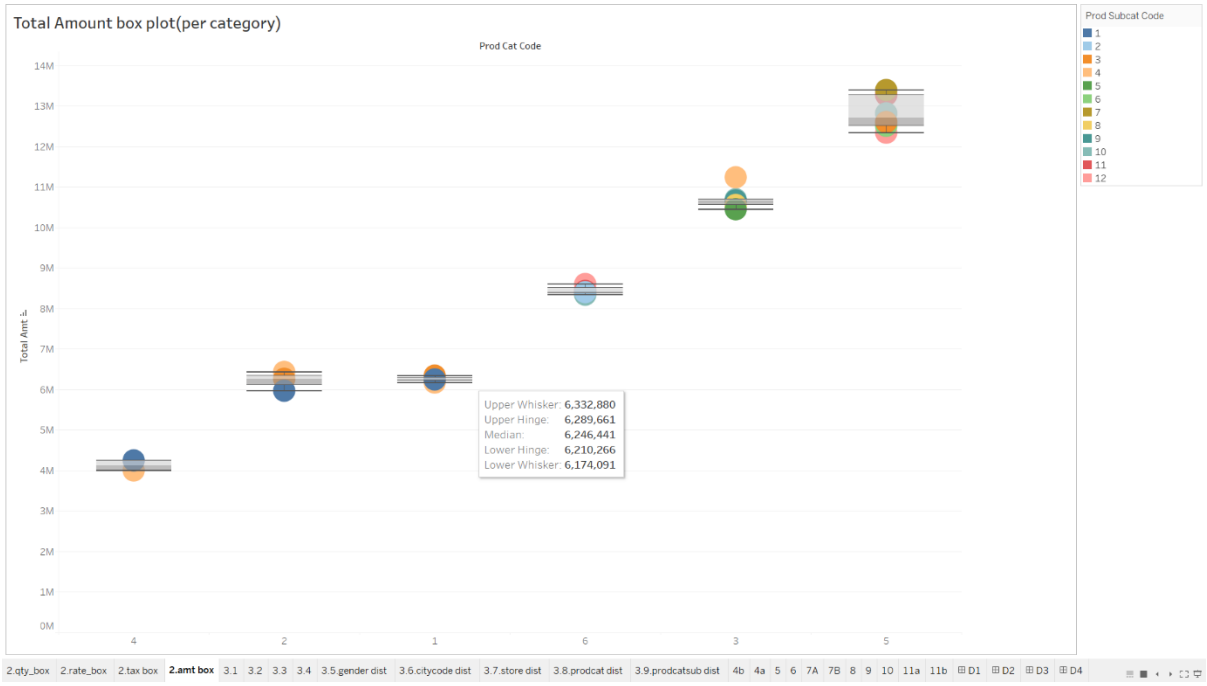


Figure 14: Amt box plot

The box plots for the continuous variables (quantity, rate, tax, total_amount) are shown above. Each of the box plots provides 5 values: Upper Whisker, Upper Hinge,Median,Lower Hinge,Lower Whisker where the whisker values are based on a value that is 1.5 times the standard deviation. Additionally, each of these measures was calculated for each of the product sub-categories within a category for a one to one comparison between the values of the measures between categories.

5.4.Visualizations for continuous and categorical variables

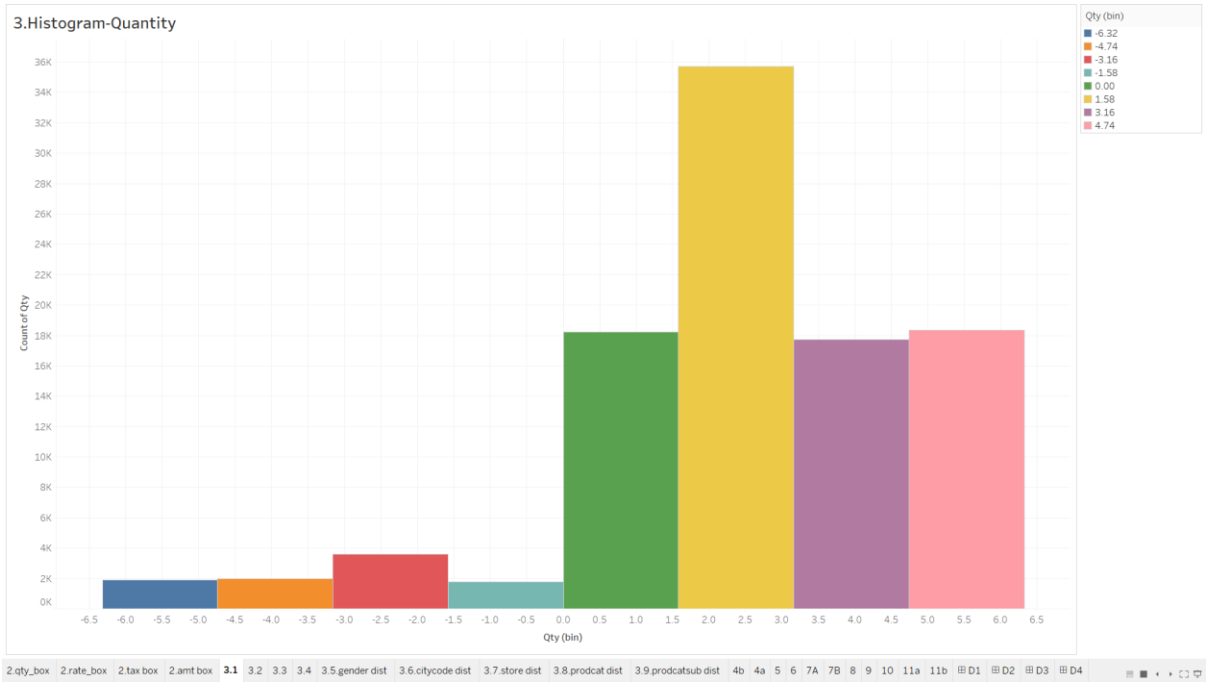


Figure 15: Histogram qty

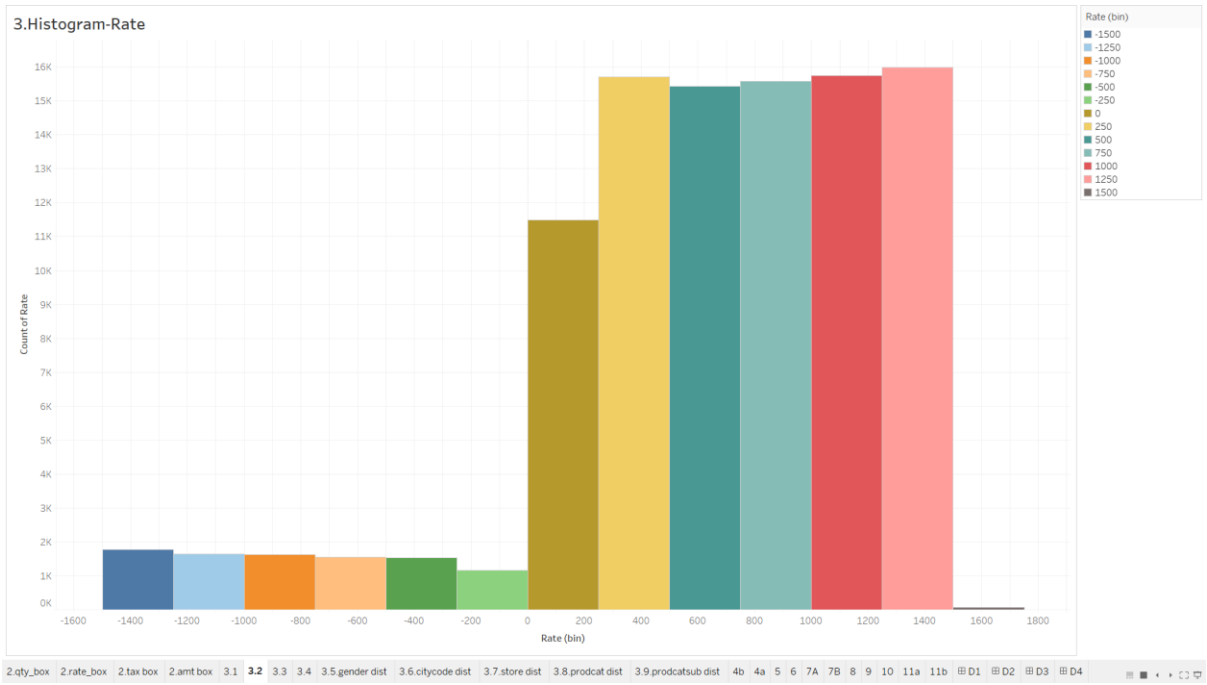


Figure 16: Histogram rate

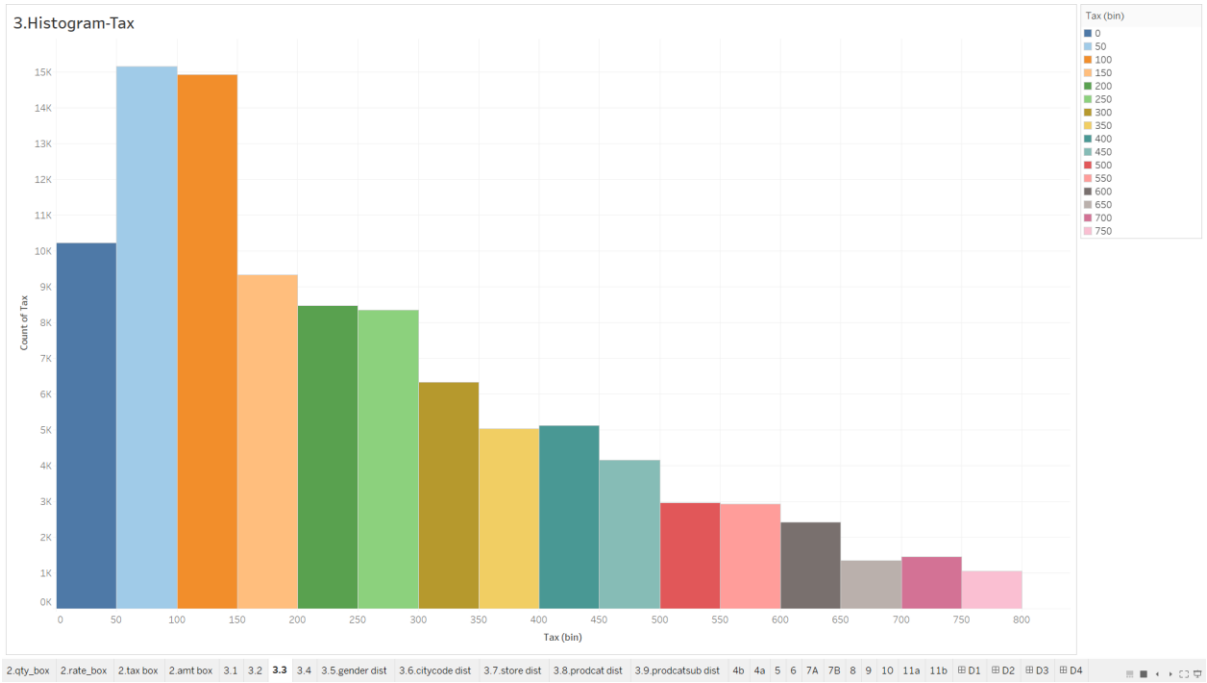


Figure 17: Histogram tax

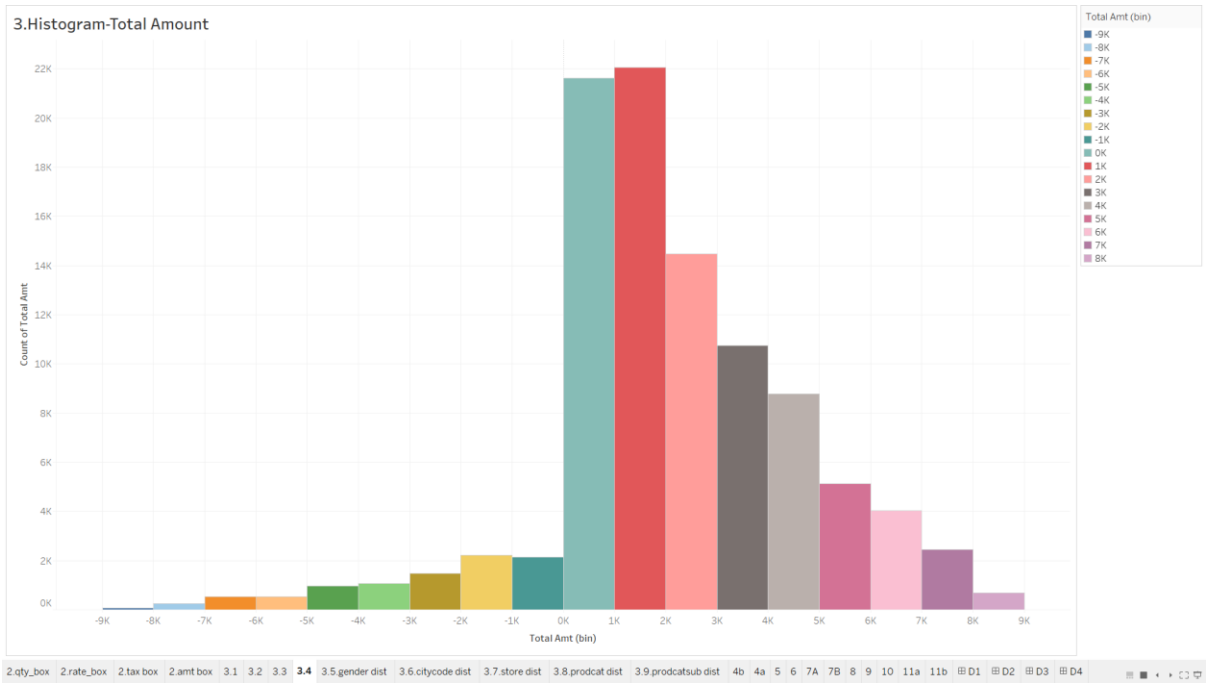
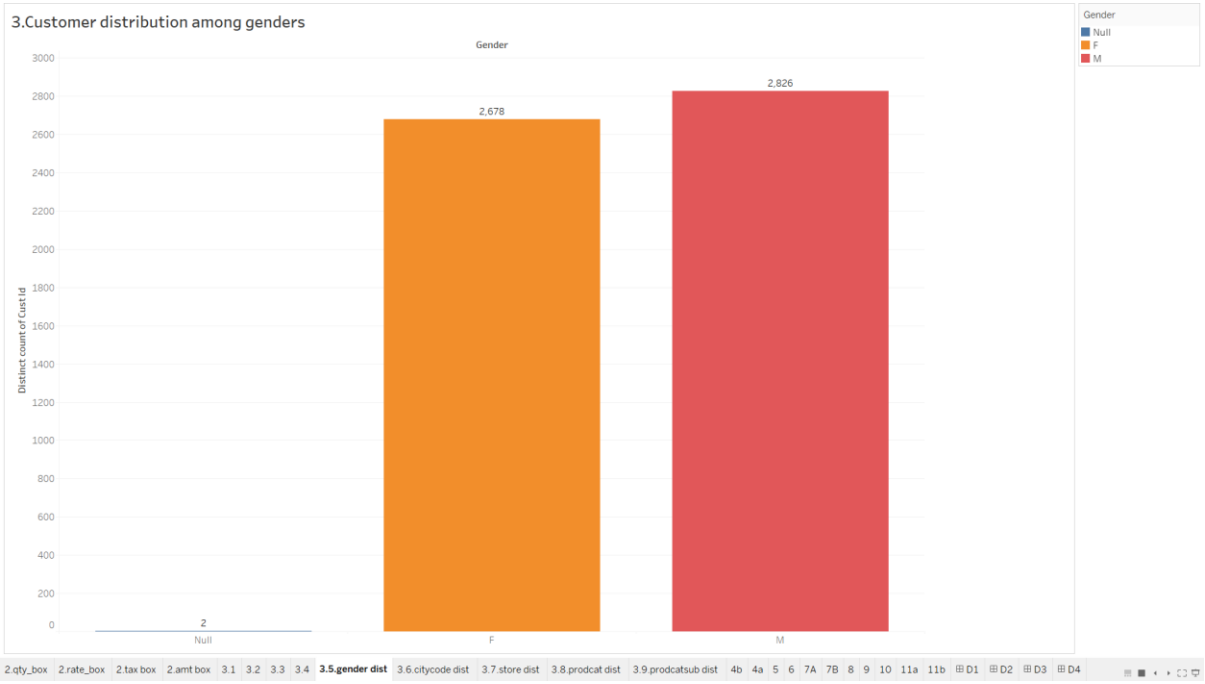
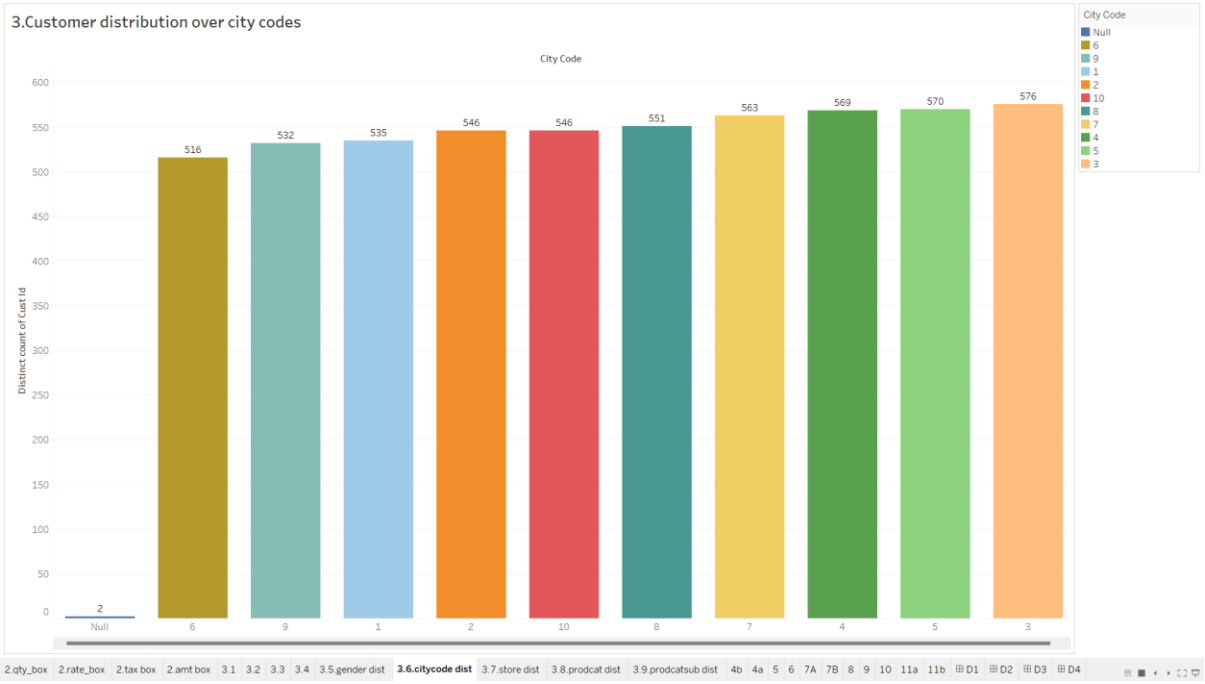


Figure 18: Histogram total_amount

The histograms for the continuous variables (quantity, rate, tax, total_amount) are shown above.

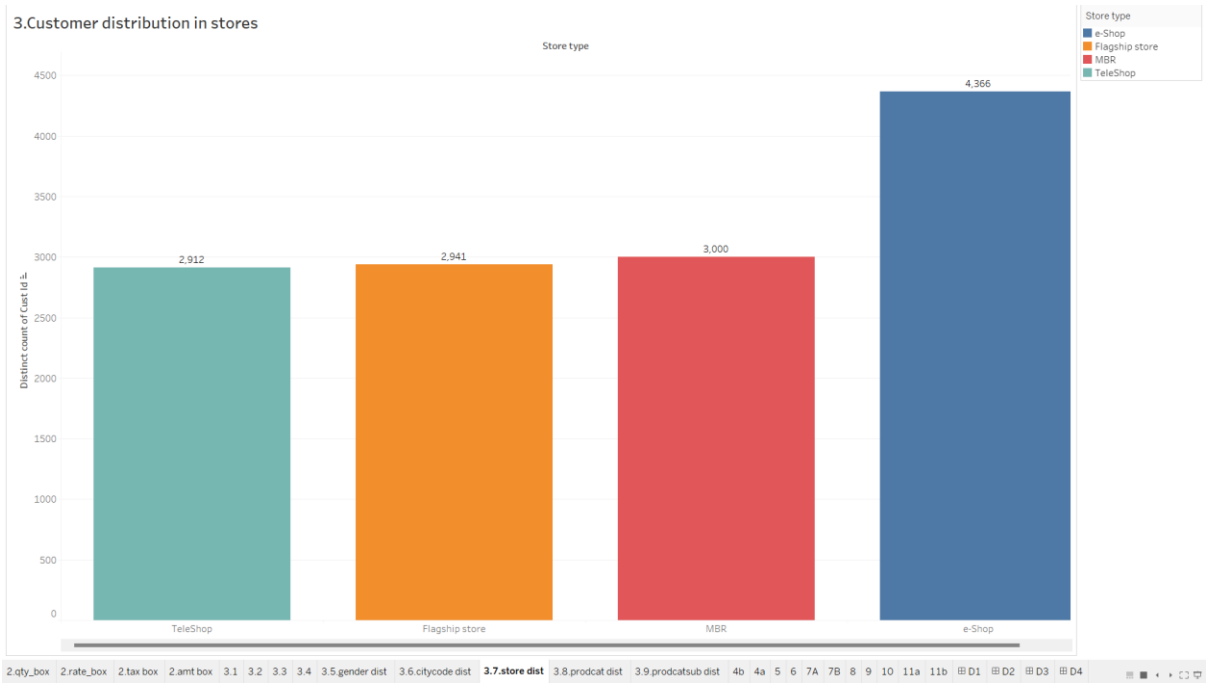


The above is a bar chart indicating customer distribution based on gender. It is observed that male customers are greater in number as compared to female customers.

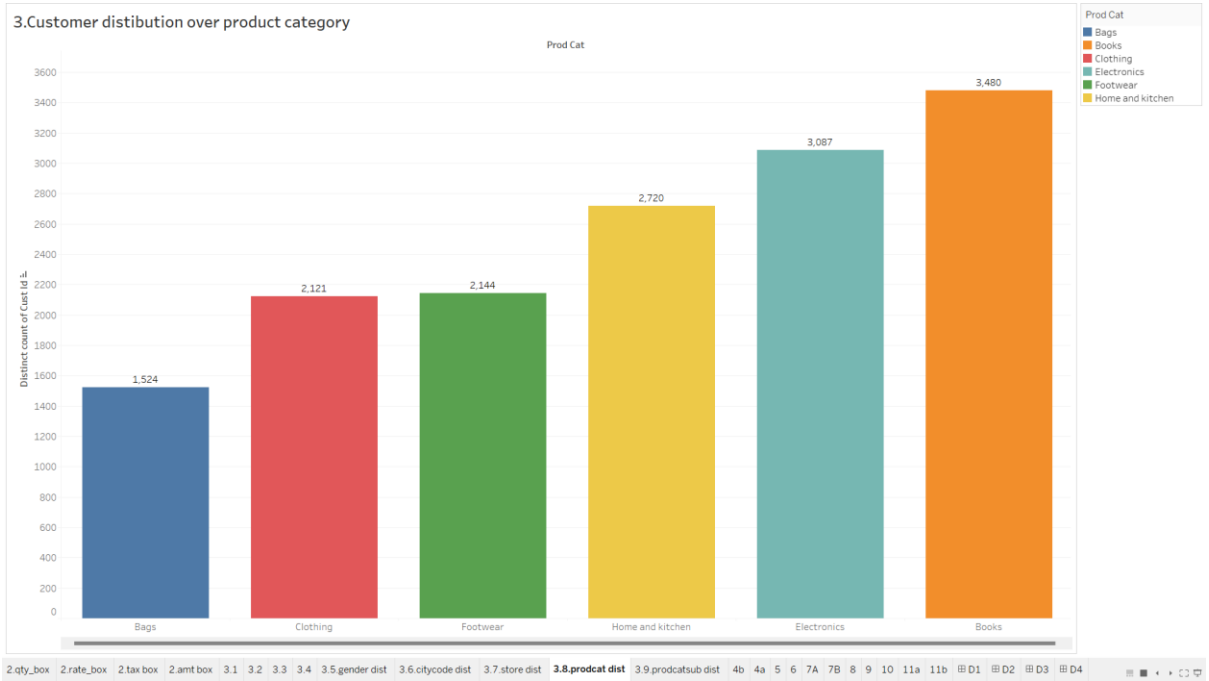


The above bar chart shows the distribution of customers over different city codes. City Code 6 has the least number of customers and City Code 3 has the maximum number of customers.

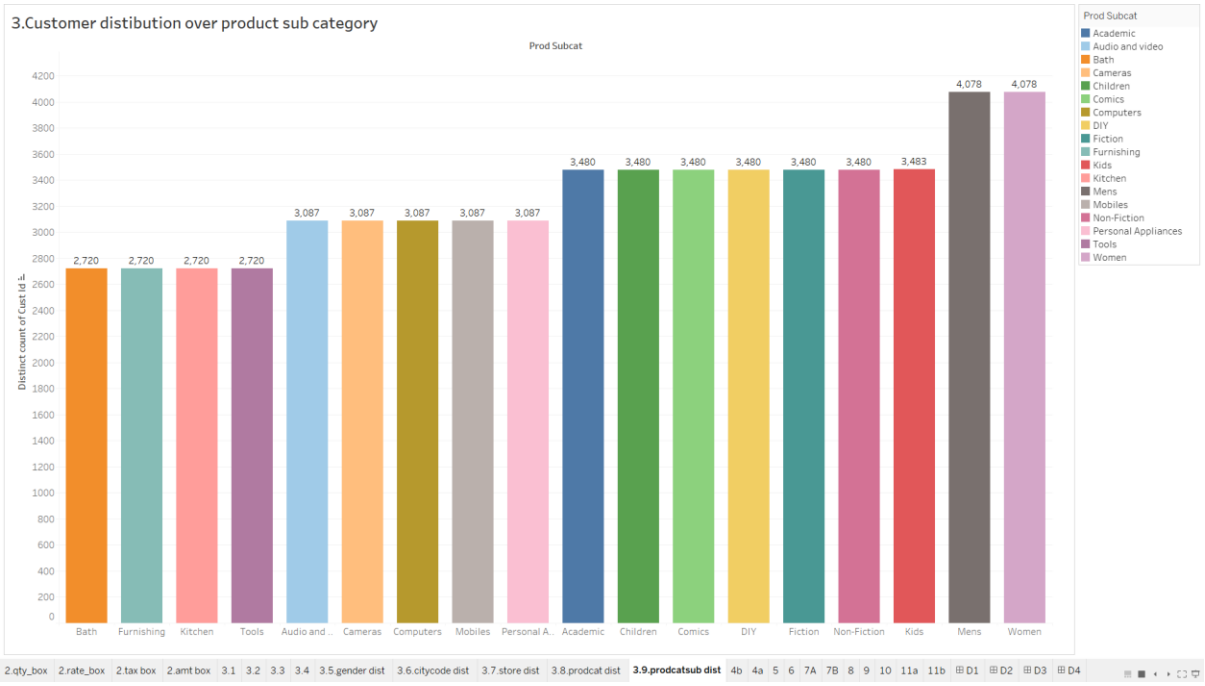
Additionally, City Codes 2 and 10 have the same number of customers.



Above is a bar chart describing the distribution of customer based on choice of stores adopted. It is observed that the largest number of customers frequented the e-shop store and that they preferred the teleshop the least.

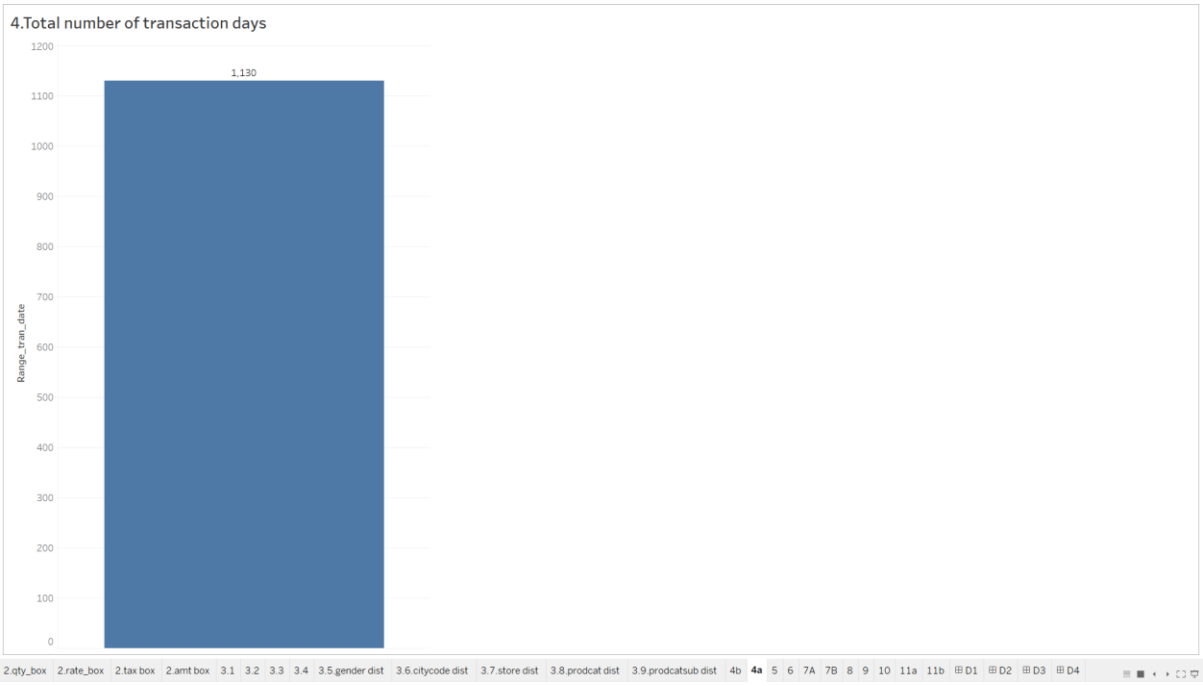


Above is a bar chart describing the distribution of customers based on choice of product categories purchased. It is observed that customers preferred books the most followed by electronics and home and kitchen items respectively. Bags were the least purchased followed with clothing items.

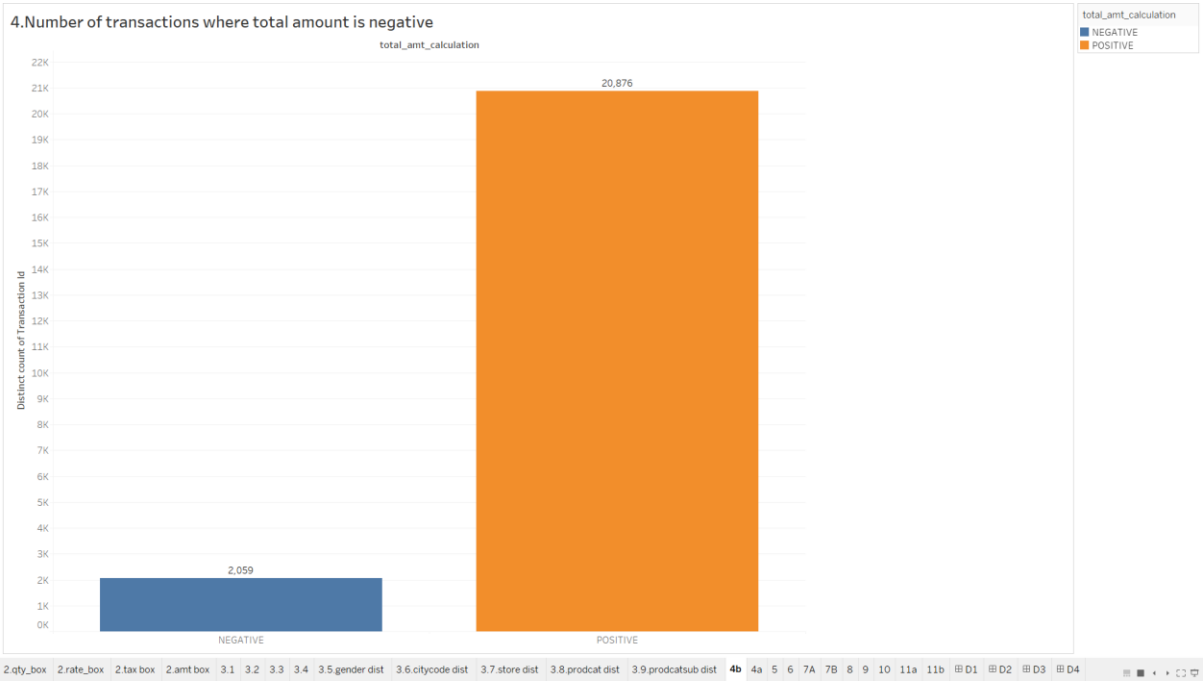


Above is a bar chart describing the distribution of customers based on their choice of product sub categories. It is observed that customers preferred customers preferred products related to the men and women sub categories the most followed by Academia, Comics, DIY and fiction showcasing similar levels of interest from customers. Bathing, Furnishing, Kitchen and Tools were preferred the least in comparison to other subcategories.

5.5. Calculations on the merged dataset merged dataset:



The total number of transaction days was calculated by means of a calculated field, range_tran_date for which the formula is given below:
DATEDIFF('day', {MIN([Tran Date])}, {MAX([Tran Date])})

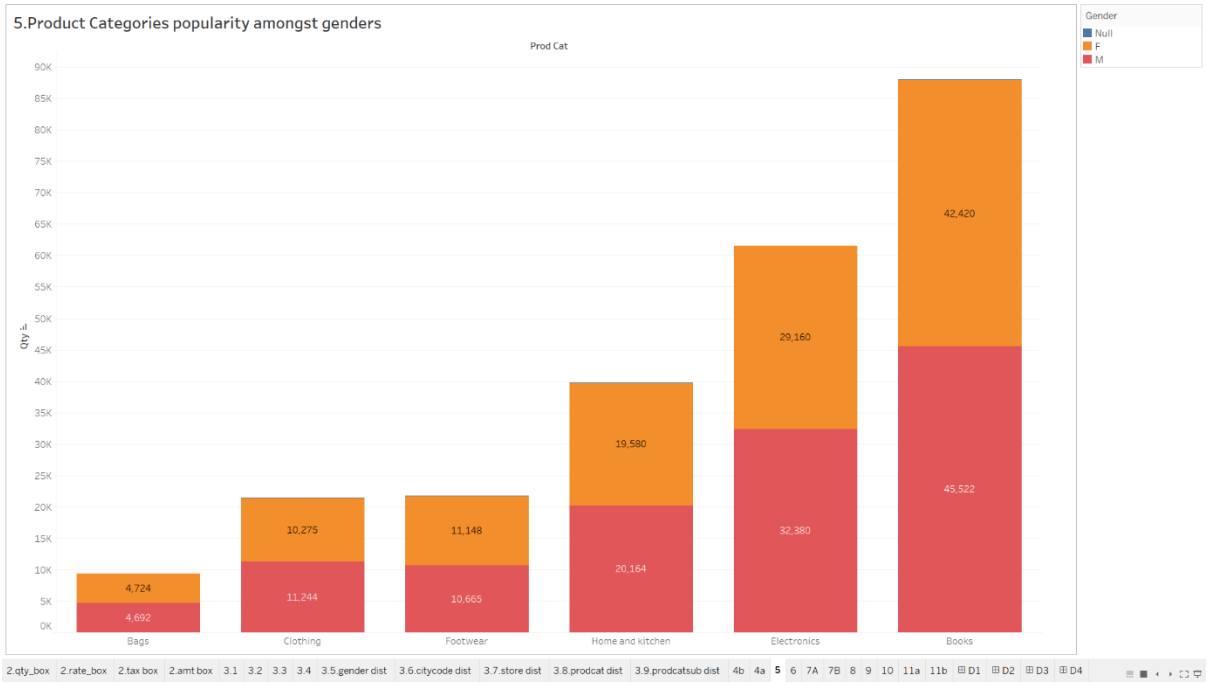


The total number of transactions where the total amount was negative was calculated using a calculated field total_amt_calculation which was adopted as a filter to

showcase positive and negative amounts respectively.The formula adopted for the same is as follows:

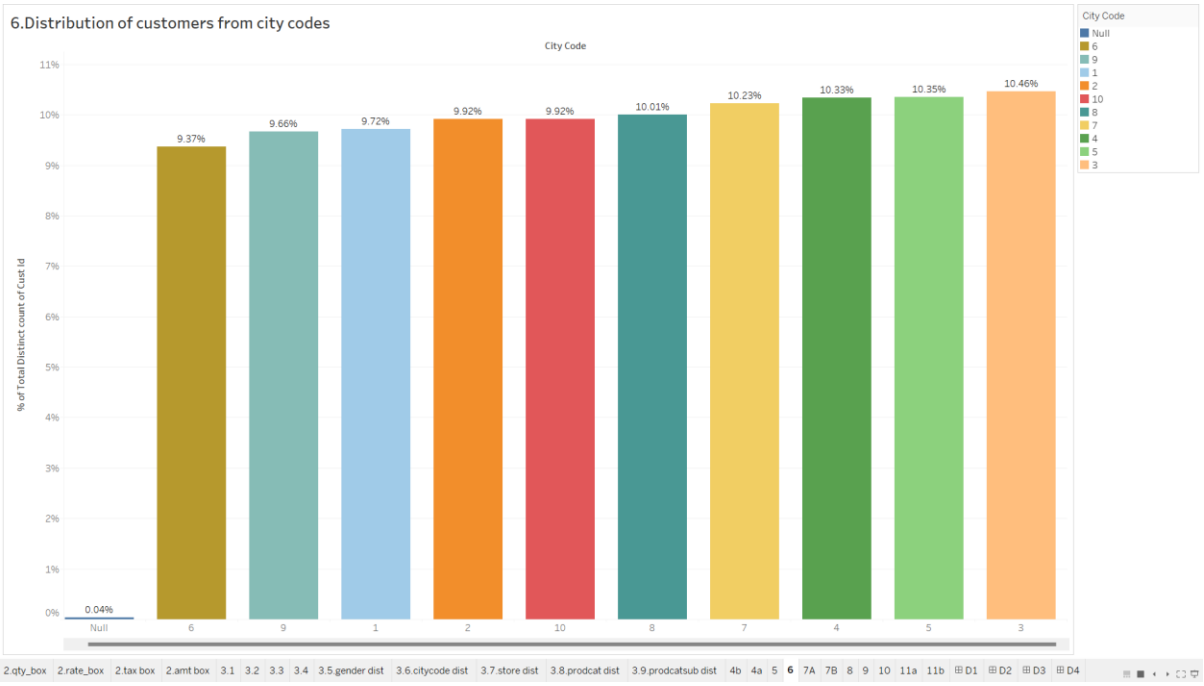
```
IF [Total Amt]<0 THEN 'NEGATIVE'
ELSE 'POSITIVE'
END
```

5.6.Product Categories popularity amongst genders



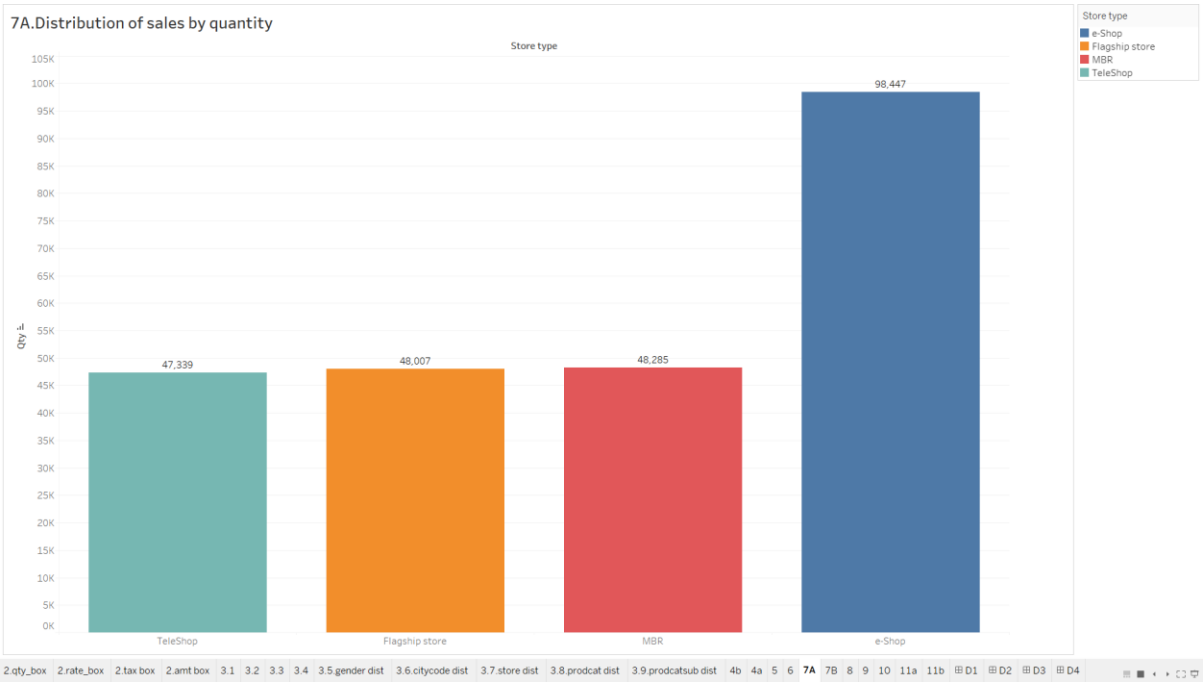
The above stacked bar chart showcases popularity of customer preferences based on gender. Both genders prefer books the most in contrast with bags which is preferred the least.In general , it is observed that men have a greater quantity of items processed as per categories with the exception of footwear where women have a greater number of items processed.

5.7.Distribution of customers from city codes

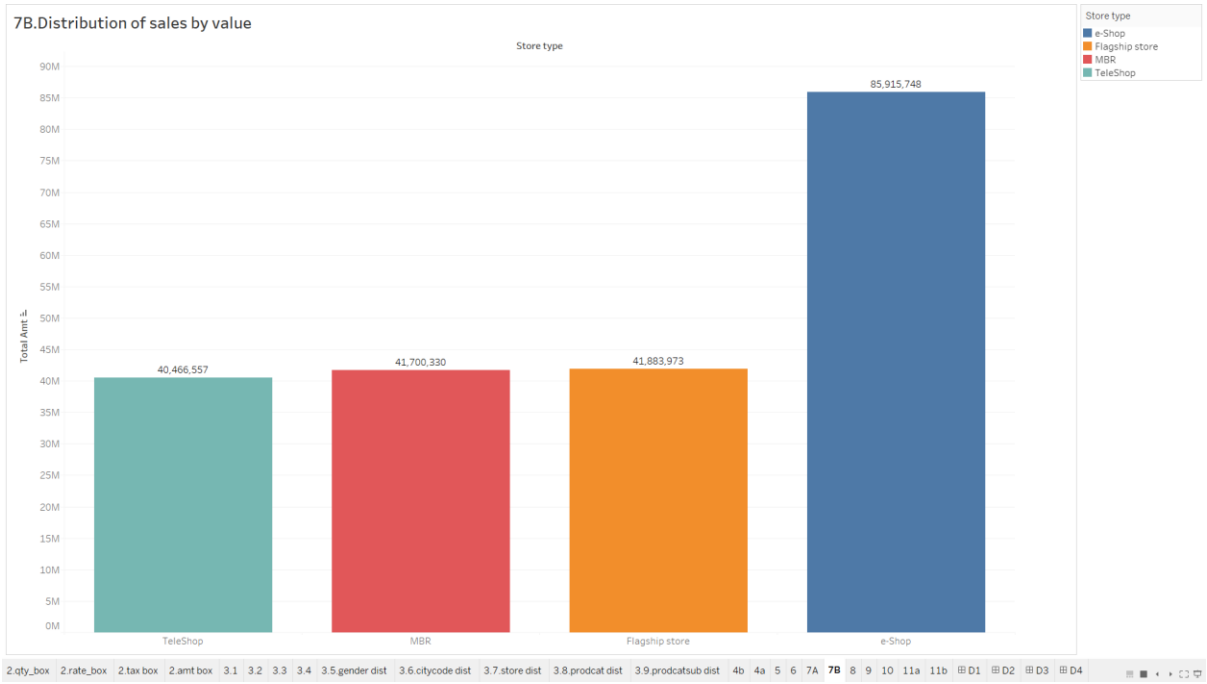


The above bar chart shows the contribution of customers in percentages from each of the city codes. It is found that the greatest contribution comes from city code 3 where the contribution is 10.46%.

5.8. Distribution of sales by quantity and value

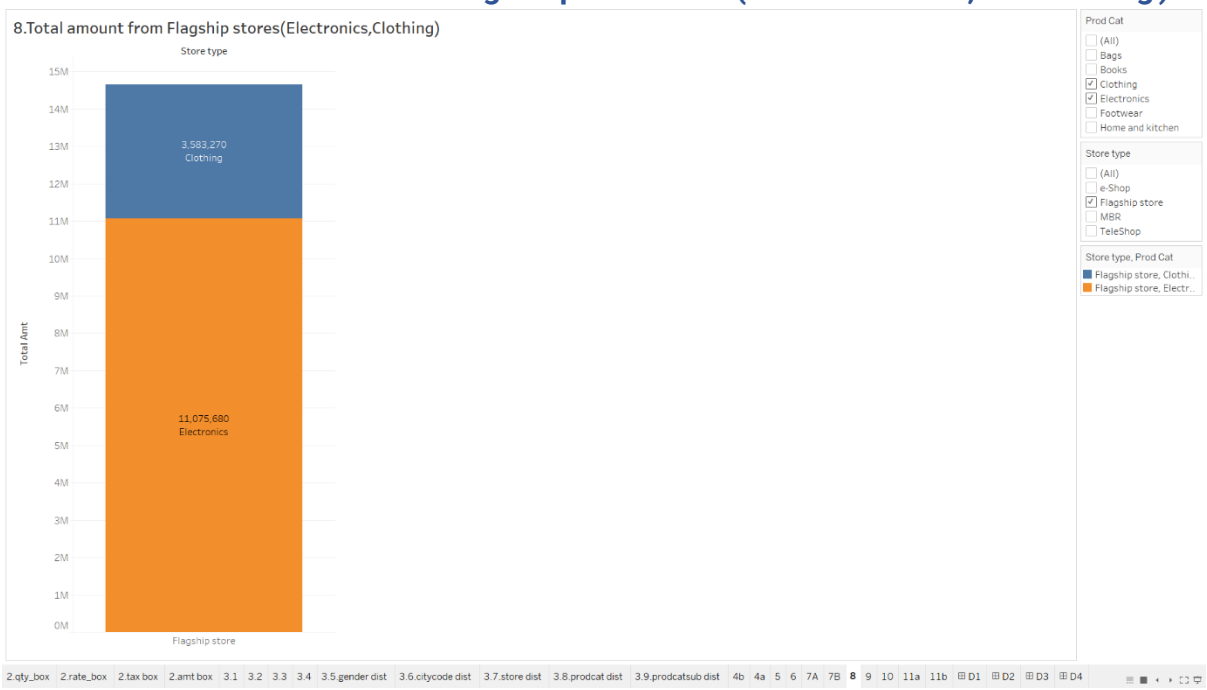


The above bar chart indicates that the e-Shop has the greatest sales going by quantity followed by MBR and flagship stores.



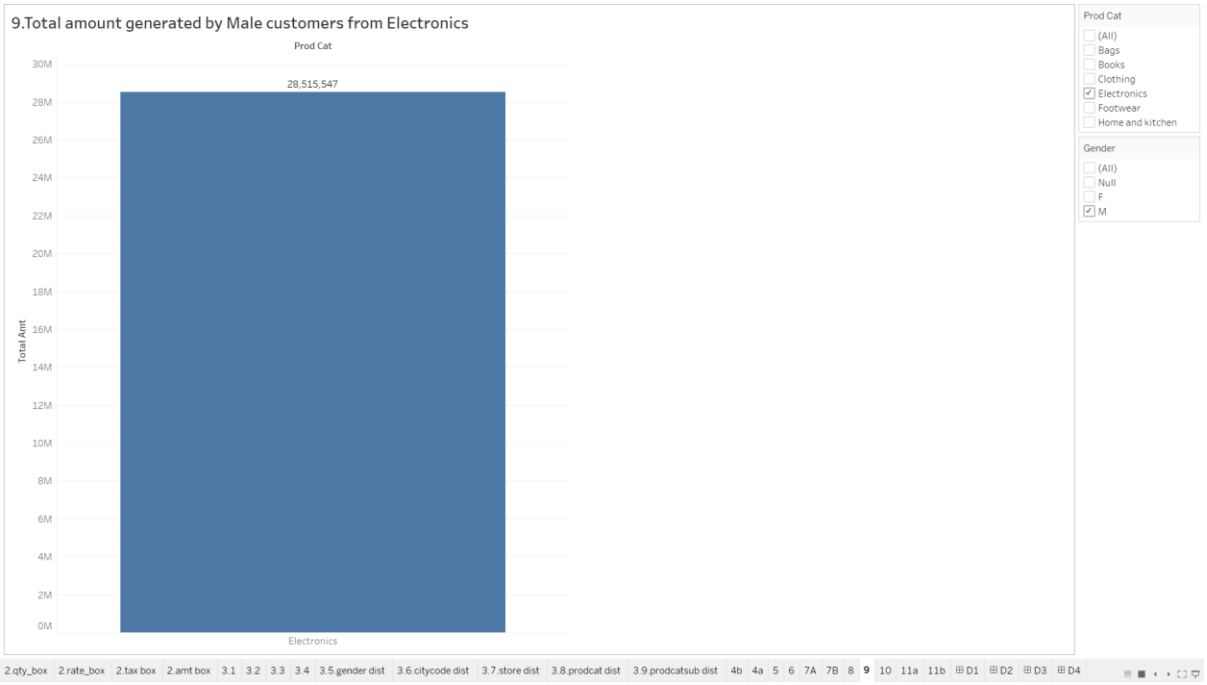
The above bar chart indicates that the e-Shop has the greatest sales going by the total amount generated followed by flagship stores and MBR.

5.9.Total amount from Flagship stores (Electronics, Clothing)



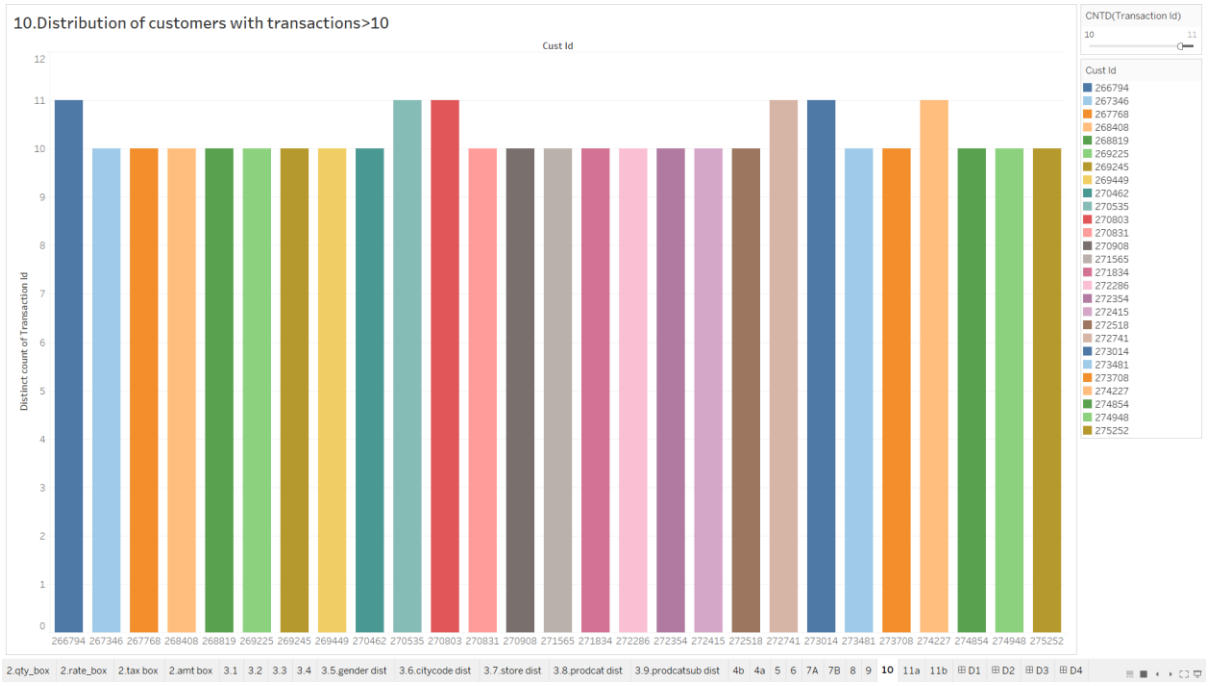
The above chart shows that the total amount earned from Flagship Stores in the Electronics category was 11,075,680 units and the total amount earned by the Clothing category was 3,583,270 units.

5.10.Total amount generated by Male customers from Electronics



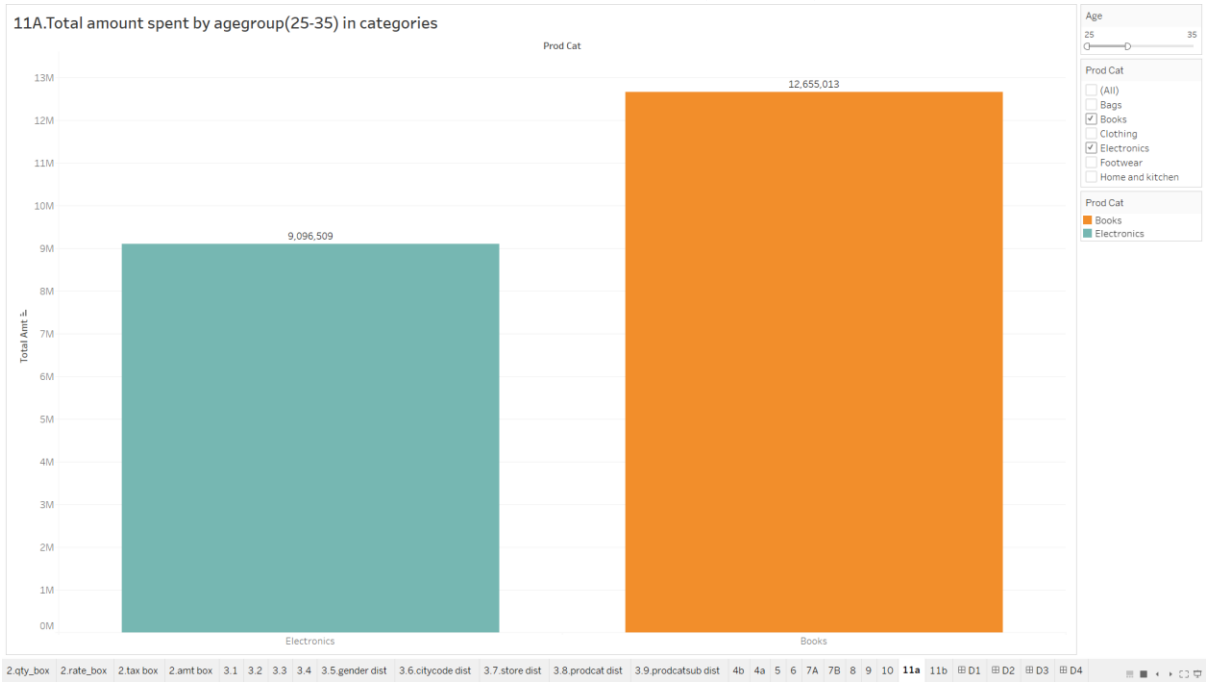
It is found that the total amount generated by males from electronics is 28,515,547 units.

5.11.Distribution of customers with transactions>10

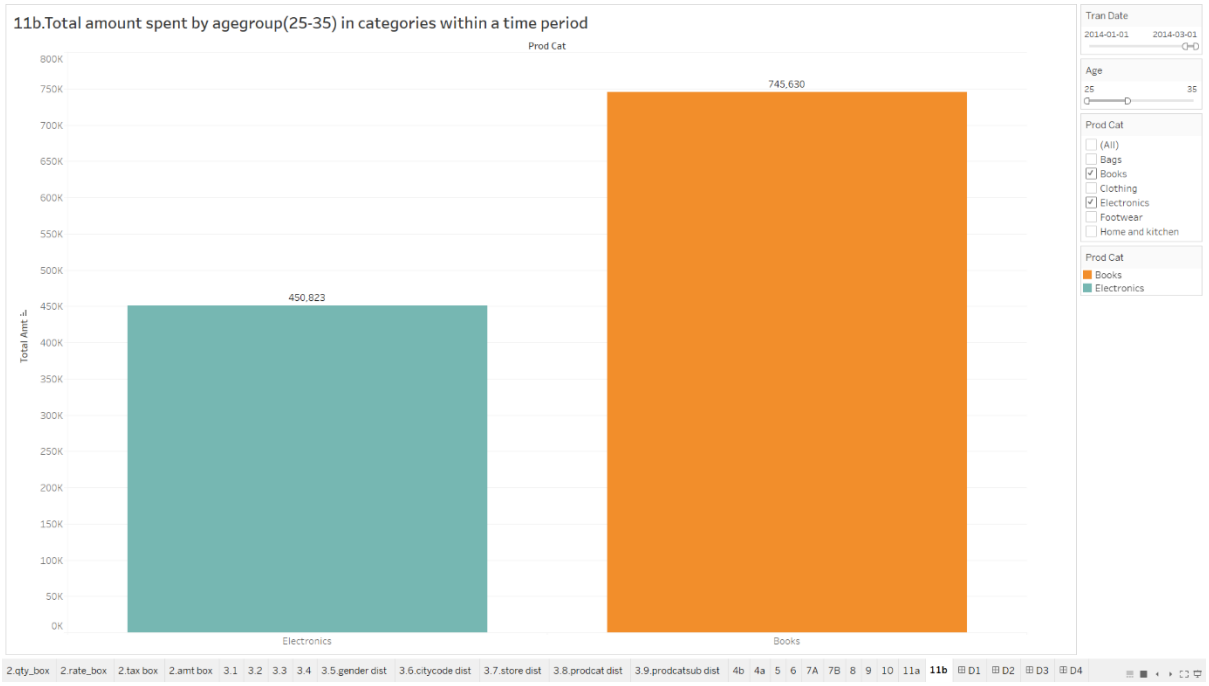


The above chart showcases the customers who have greater than 10 unique transactions.

5.12.Total amount spent by age group(25-35) in categories



The above chart shows the total amount spent by customers in the age group of 25-35. They spent 9,096,509 units on Electronics and a sum of 12,655,013 on Books.



The above chart shows the amount spent by customers of age group 25-35 in electronics and books over the course of two months from 1 Jan, 2014 to 1 March, 2014. The customers spent 450,823 on Electronics and 745,630 on books.

6.Conclusion

This analysis underscores the importance of data-driven insights in navigating the competitive landscape of retail. Through comprehensive examination of customer transactions, demographic patterns, and sales performance across categories, the report offers actionable insights that support strategic decision-making in several key areas.

Firstly, the findings on product popularity among different demographics provide essential information for inventory management and targeted marketing strategies. By understanding gender-based preferences and age-group spending patterns, the store can better align its offerings to customer demands, enhancing satisfaction and loyalty.

Secondly, the revenue analysis by store type and category highlights where the retail store can focus its efforts for maximum profitability. Identifying flagship stores as high performers, especially in electronics and clothing, reveals

potential areas for investment and promotional campaigns that can drive further sales.

Lastly, insights on customer behaviour, such as transaction frequency and city distribution, enable the retail store to personalize engagement efforts and optimize resource allocation based on customer concentration and shopping habits. By recognizing cities with higher customer densities and tailoring outreach programs, the store can strengthen its presence in key locations.

In summary, this data-driven approach not only helps the store respond to current trends but also enables proactive planning for future growth.