CUSTOMER ANALYTICS - RETAIL PURCHASE DATA

In this report, I have applied various techniques to analyse the retail purchase dataset. The objective of the project is to understand customer behaviours. Five major parts of the report are::

1. **Exploratory Analysis:** to understand the data
2. **Customer Metrics Understanding**: AOV and CLV
3. **Customer Retention (Cohort Analysis)**: to investigate how the company retained customers
4. **Market Basket Analysis**: to identify pairs of products / subcategories that often were bought together
5. **Customer Clustering with K-Means and PCA (Principal Component Analysis)**: to figure out customer clusters that share similar characteristics and behaviours

Data Description

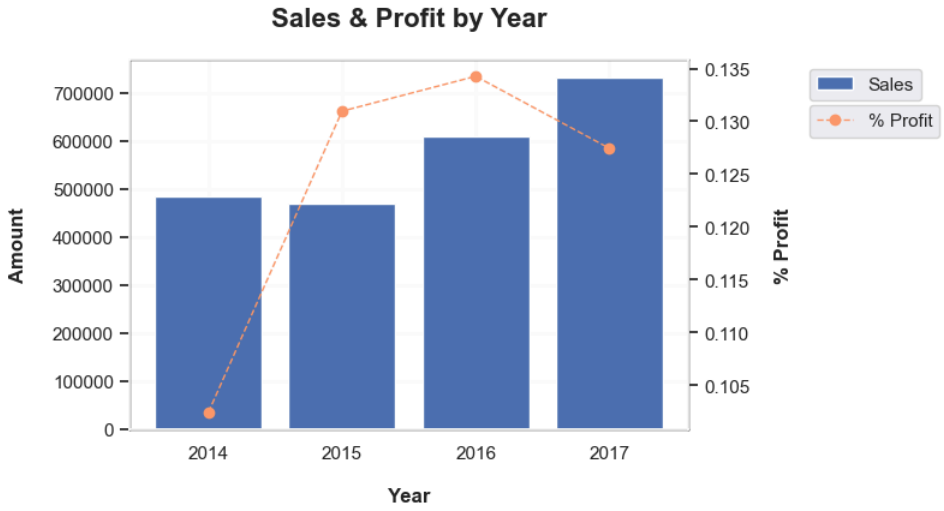
The data has 9994 rows x 20 columns

The order details were captured, with information as below:

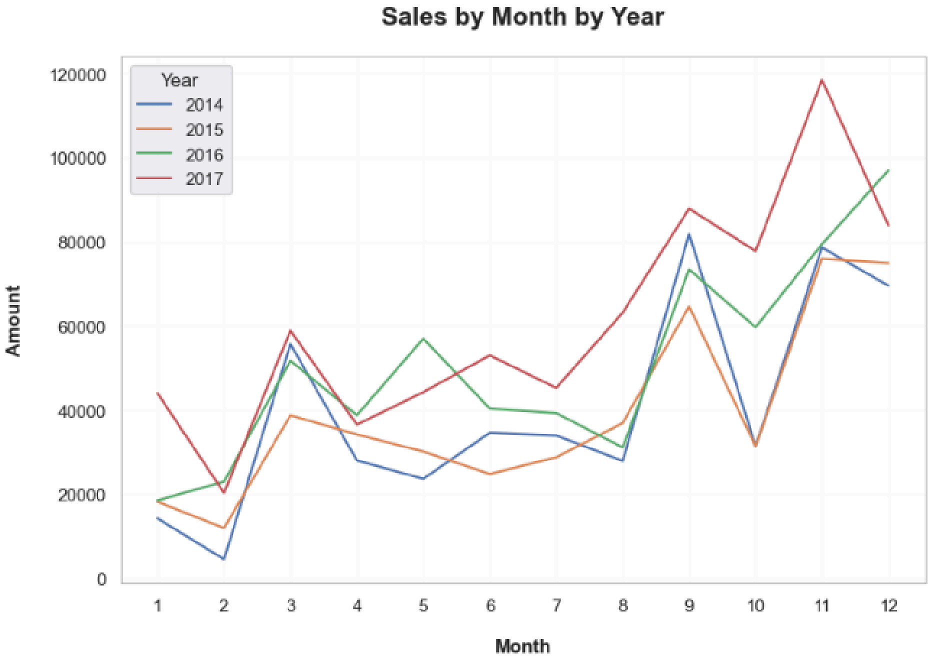
- **Row ID**: the unique value of record
- **Order ID**: the unique value of the order
- **Order Date**: the date when the order was made
- **Ship Date**: the date when the order was shipped
- **Ship Mode**: the type of shipping
- **Customer ID**: the unique value of customer
- **Customer Name**: the name of customer
- **Segment**: the customer segment (Consumer / Home Office / Corporate)
- **Country**
- **City**
- **State**
- **Postal Code**
- **Region**
- **Product ID**: the unique value of the product
- **Category**: product category
- **Sub-Category**: product subcategory
- **Product Name**
- **Sales**
- **Quantity**
- **Discount**
- **Profit**

1. Exploratory Analysis

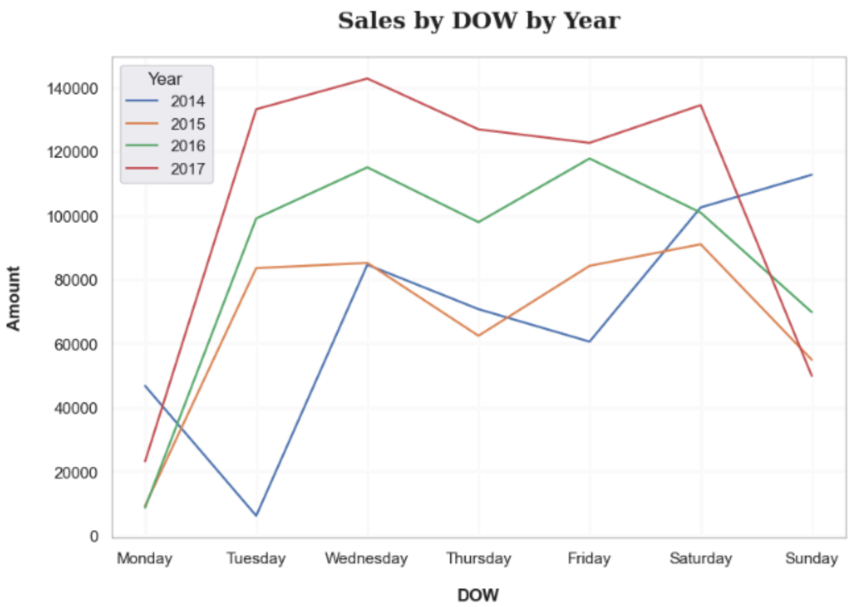
- The Sales has been increased over years



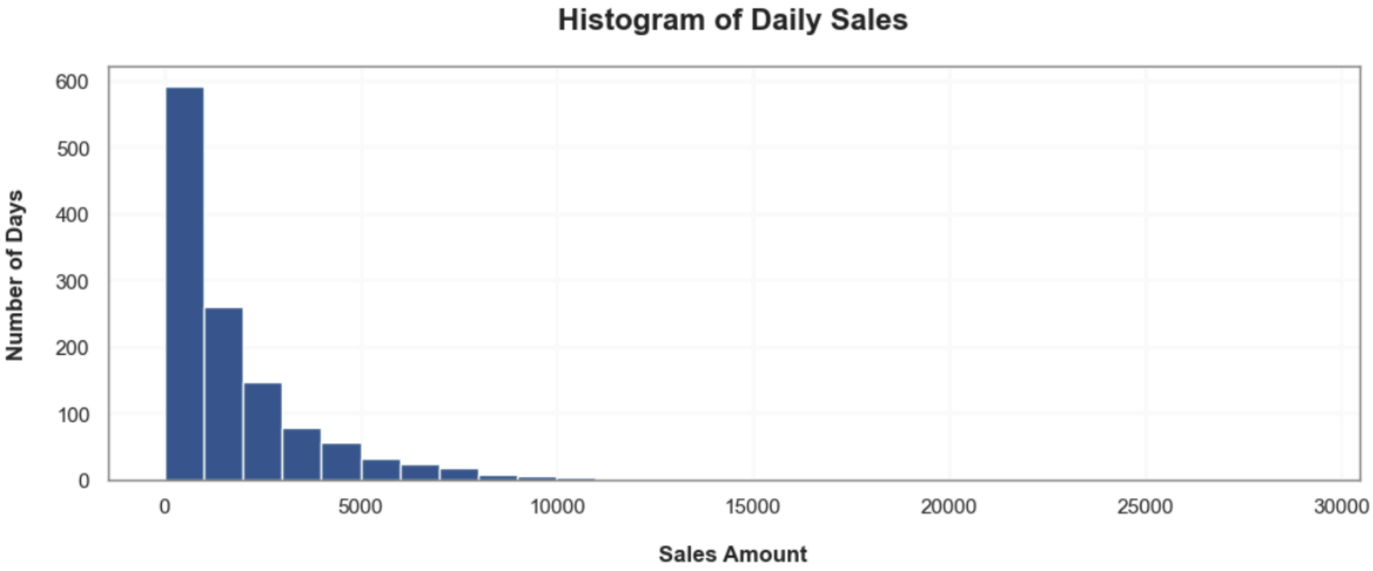
- There is a seasonality in Sales: customers bought more in September and November, contrasting to February



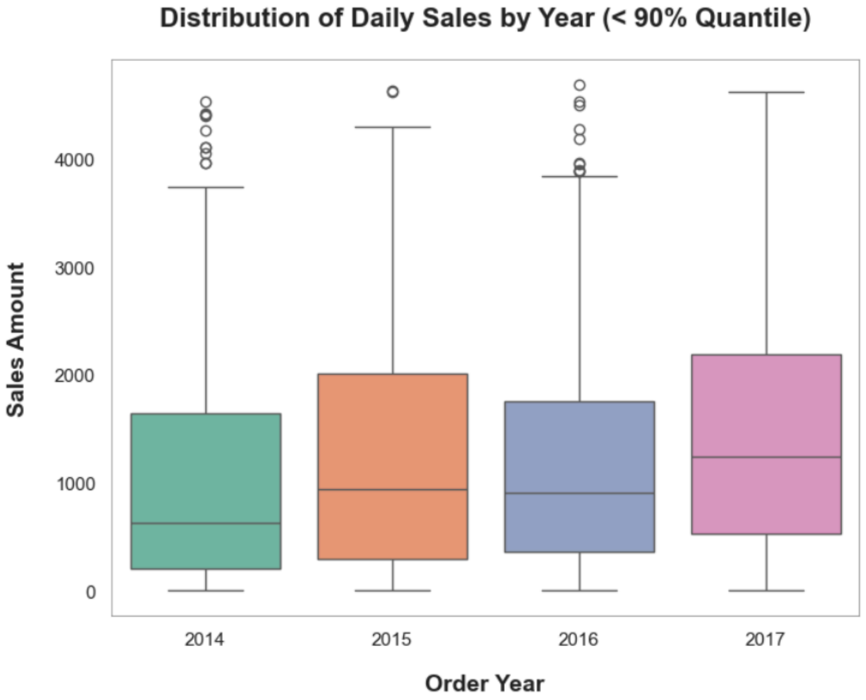
- They didn't prefer to visit superstore and spend on Monday and Sunday - the first and last day of week



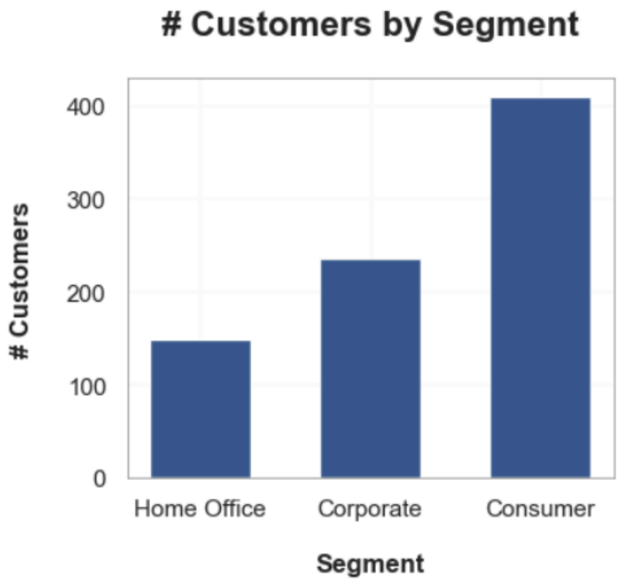
- Majority of daily sales are under 1000

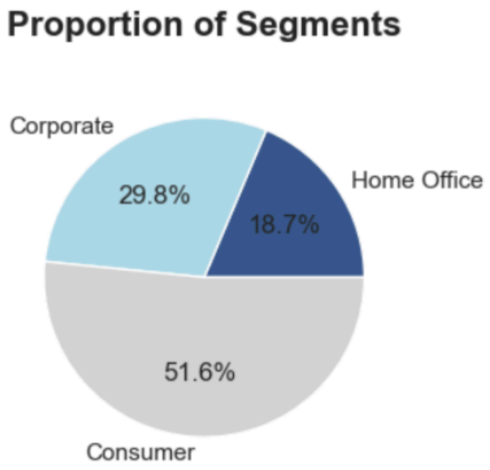


- Although there was a slowdown in Sales in 2016, median of daily sales in 2017 surpassed 1000

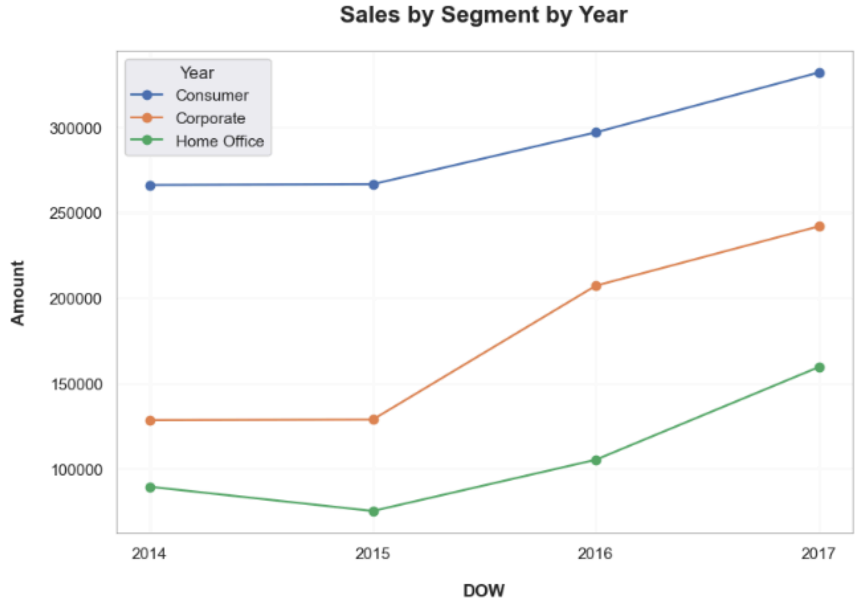


- Half of customers are individual consumers, ⅔ of the rest are Corporate

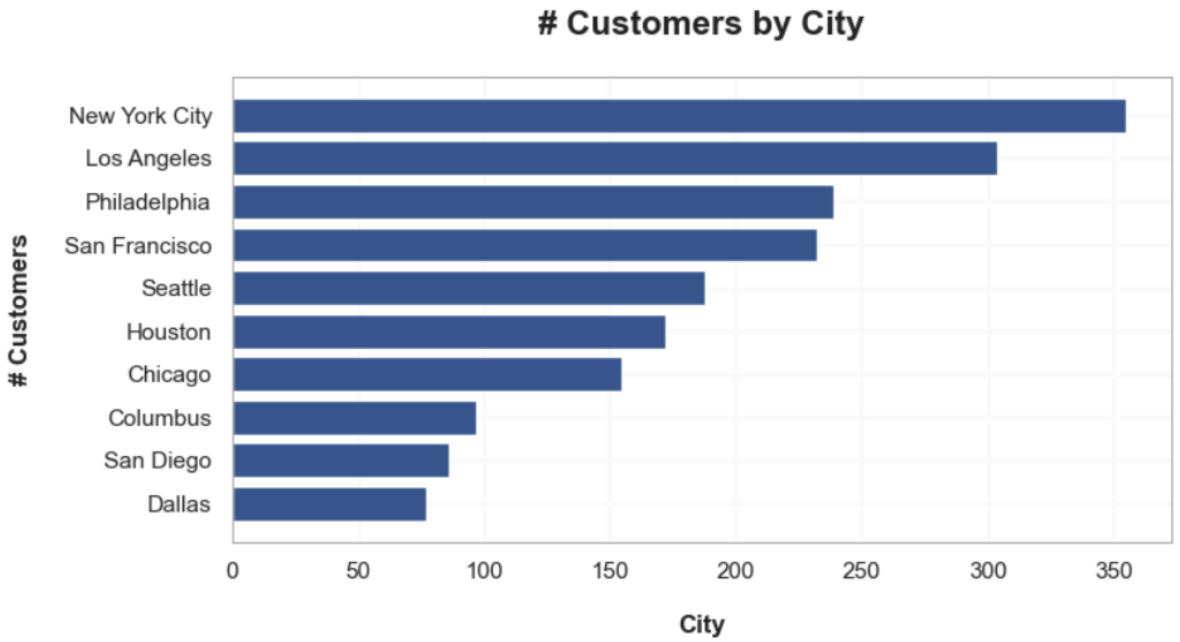




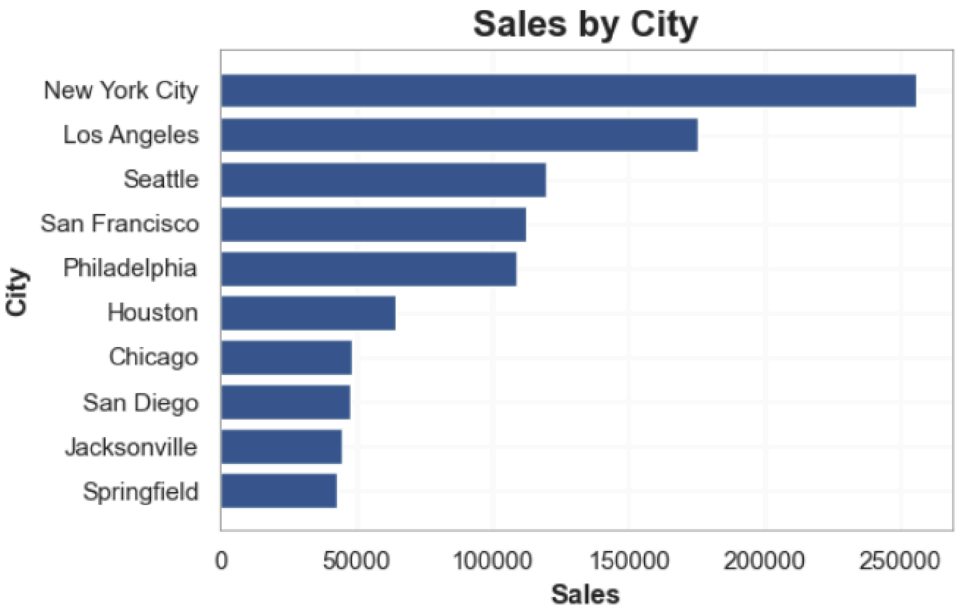
- The customer base witnessed an increasing trend over years among segments



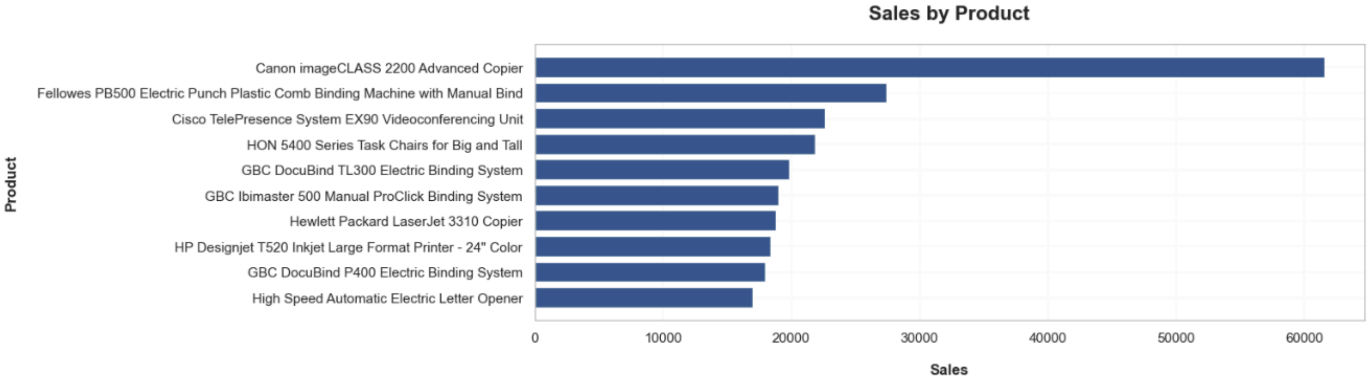
- New York City and Los Angeles have the biggest customer base



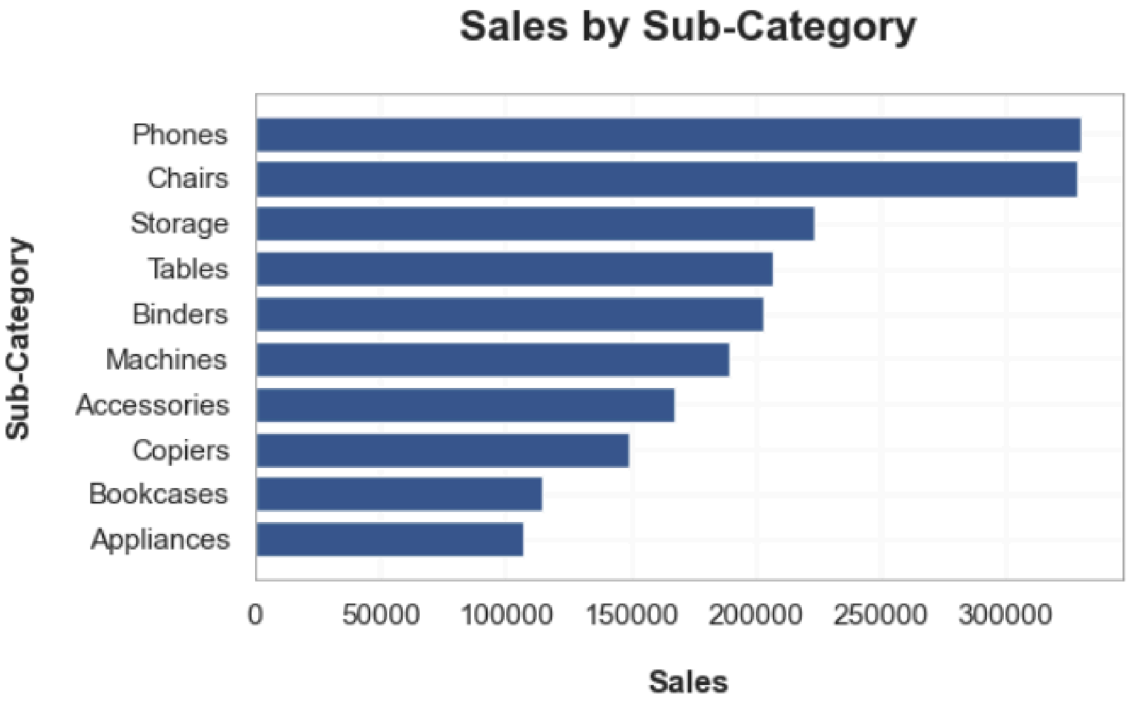
- Customers in these 2 cities also contributed the highest Sales



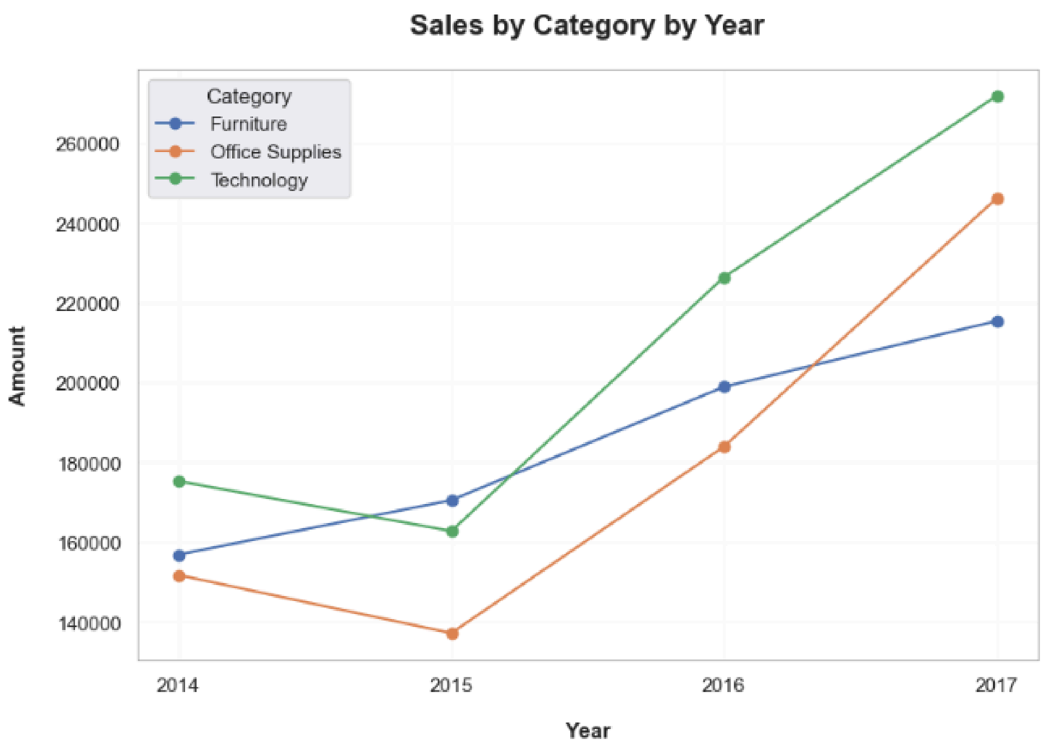
- Product: Canon Copier dominated others



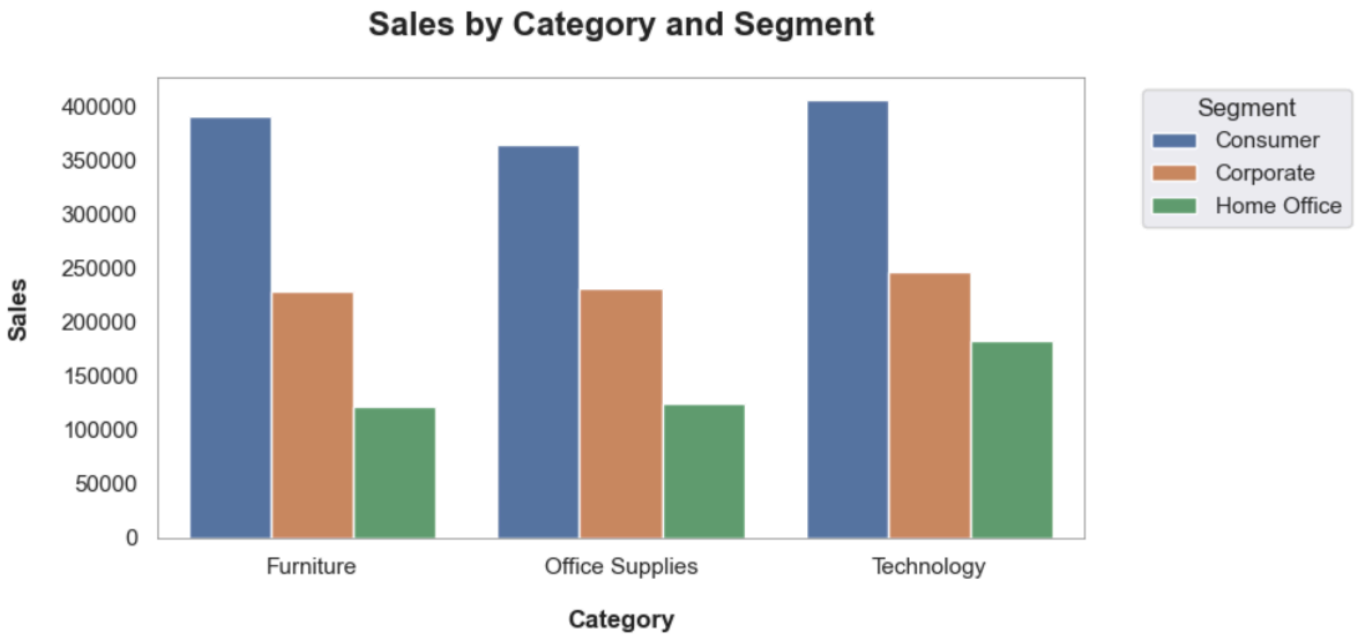
- Subcategory: Phones and Chairs were in top 2 with similar Sales



- Towards the end of report period, Office Supplies were increasingly preferred than Furniture



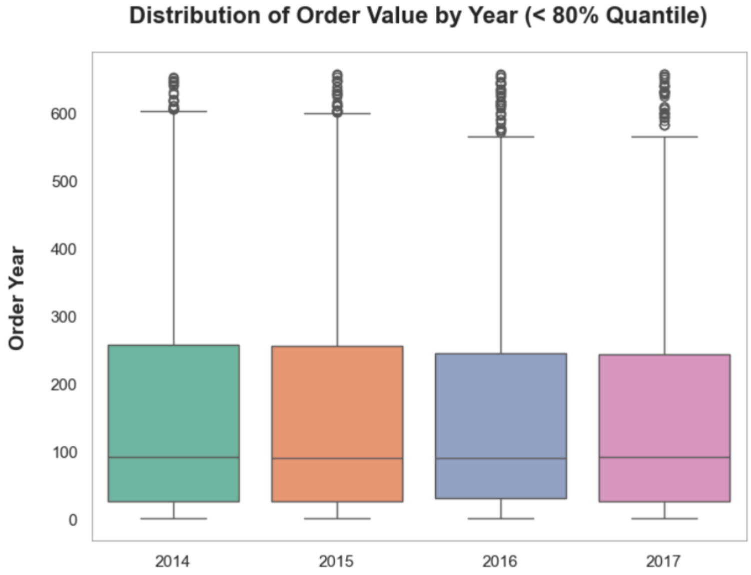
- There was no significant preference for particular categories among segments



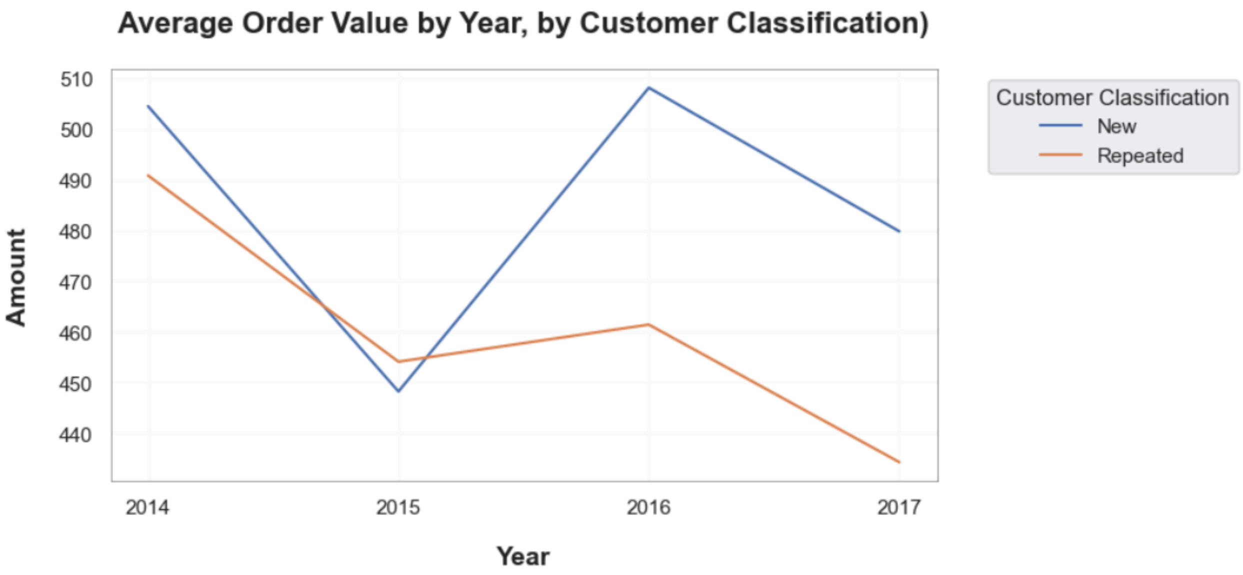
2. Customer Metrics Understanding

a. Average Order Value

- There was no significant difference in AOV over years

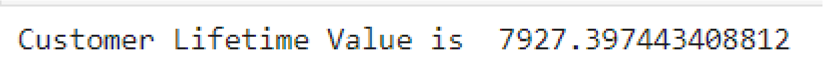

Distribution of Order Value by Year (< 80% Quantile)

- New customers spent more on each order in the last 2 years of the period

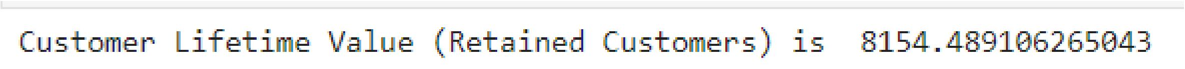

Average Order Value by Year, by Customer Classification)

b. Customer Lifetime Value

- The CLV of retained customers was higher than average
- Customer retention could be considered one of the most important objectives



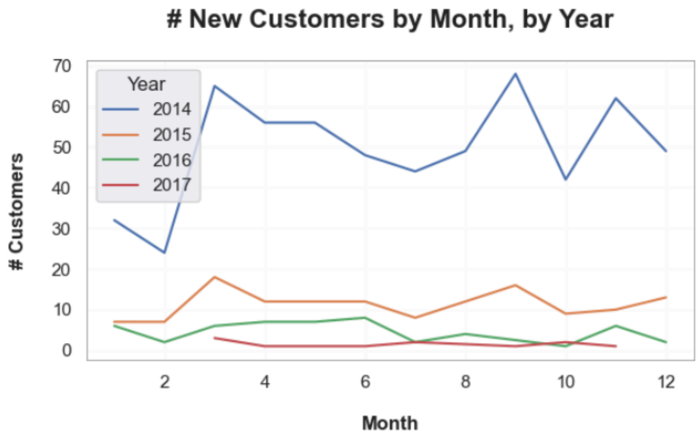
```
Customer Lifetime Value is  7927.397443408812
```



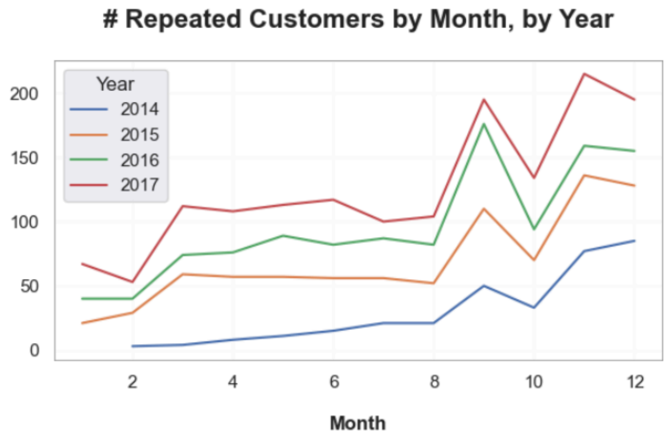
```
Customer Lifetime Value (Retained Customers) is  8154.489106265043
```

3. CUSTOMER RETENTION - COHORT ANALYSIS

- Indeed, the company acquired fewer new customers over years
- They focused on retaining customers

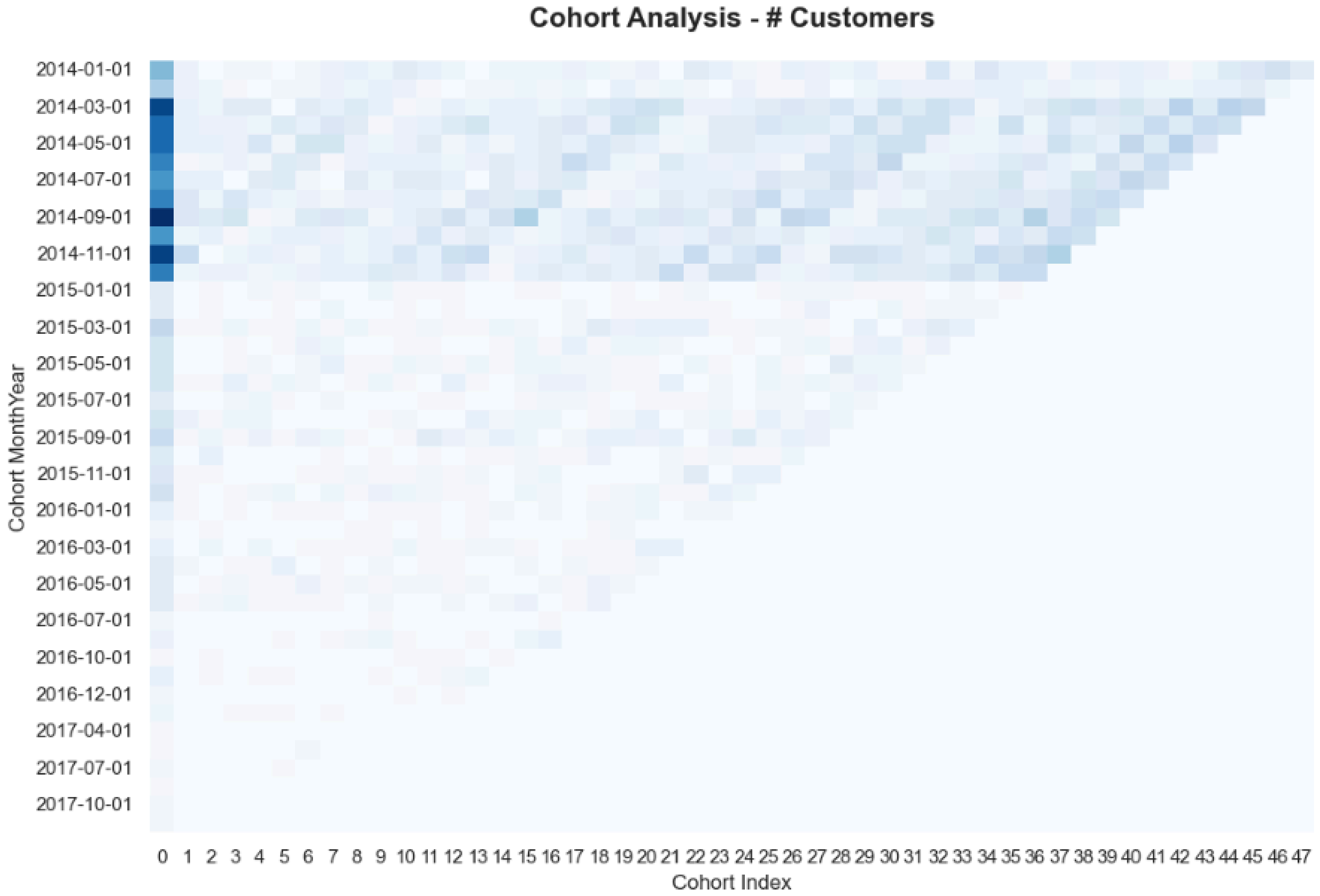


New Customers by Month, by Year



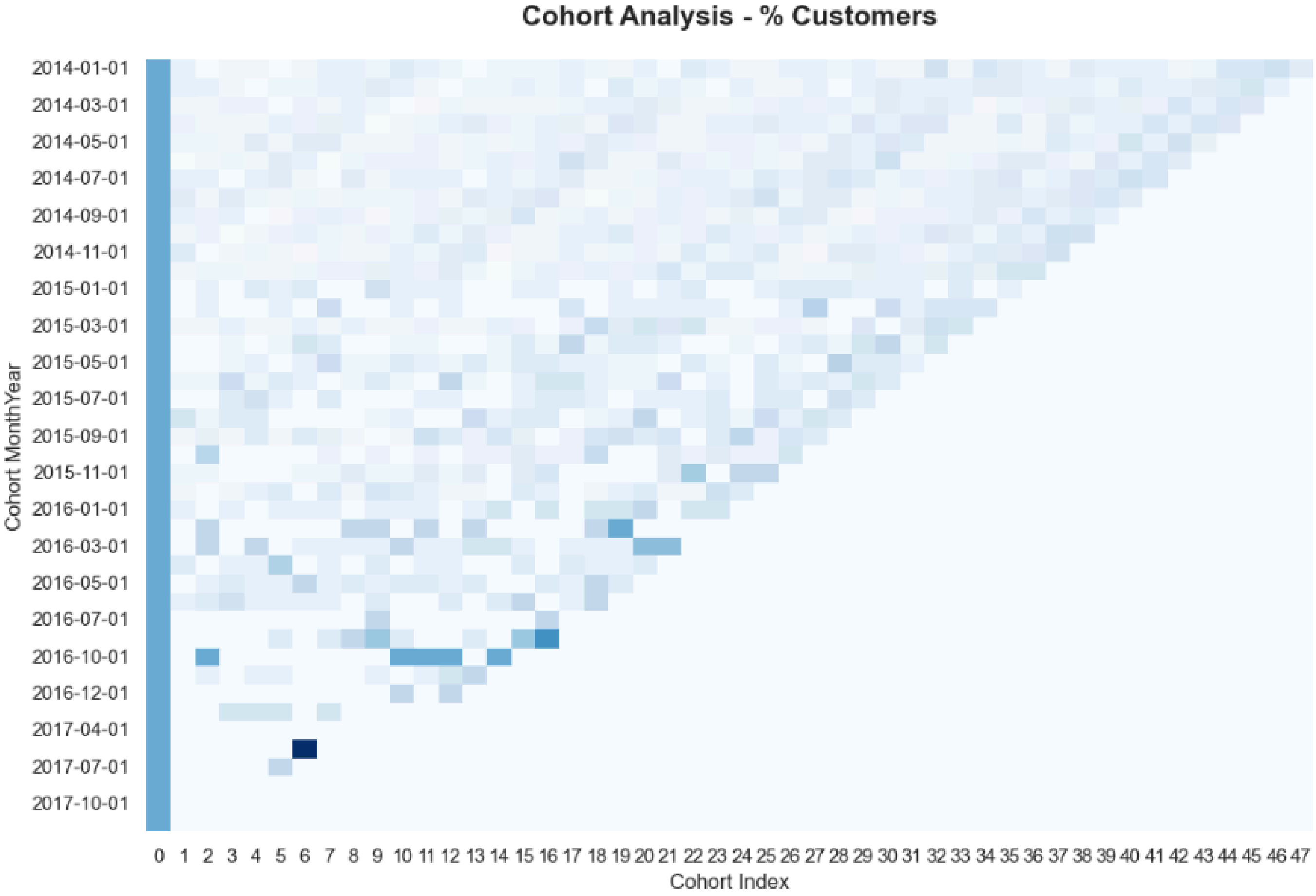
Repeated Customers by Month, by Year

- Cohorts of the first year were bigger than those of the following years
- These cohorts were most loyal until the end of the period



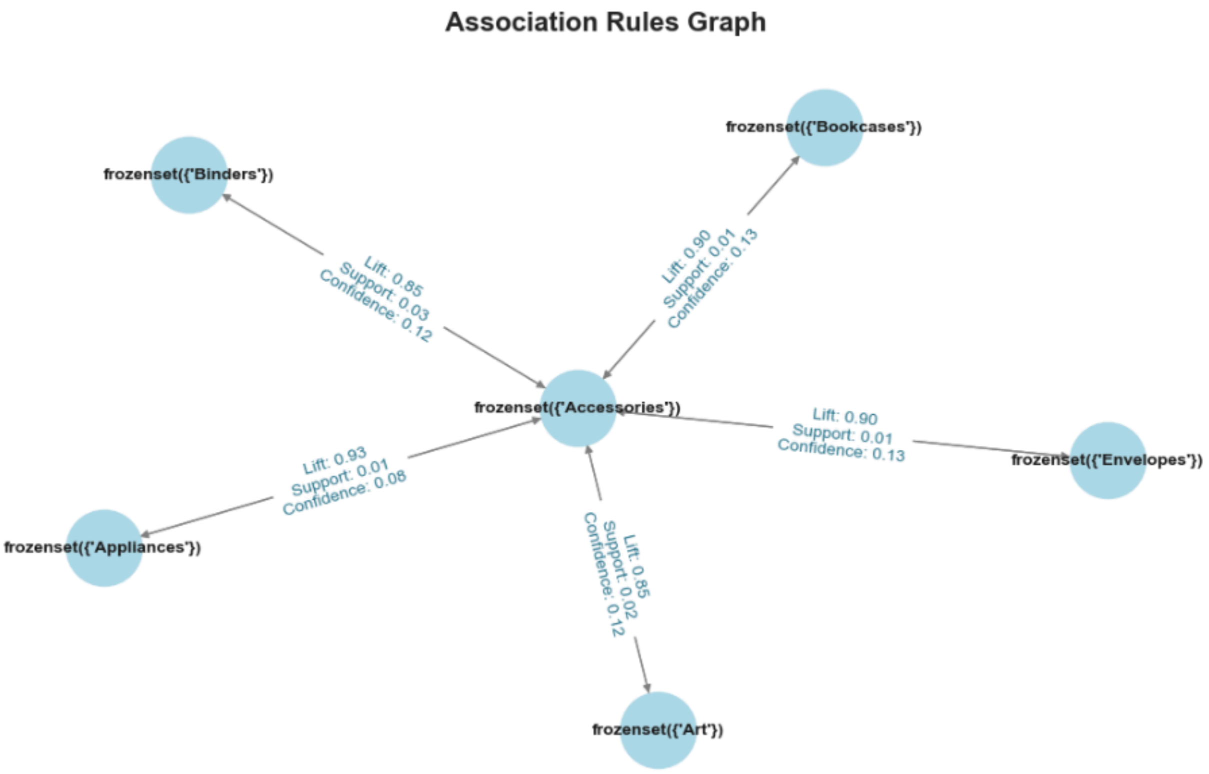
Cohort Analysis - # Customers

- Although cohorts of 2016 and 2017 have a smaller size, their retention rate within first 2 years were significantly higher than that of earlier cohorts
- These cohorts should be treated with particular policies, as retaining customers is a key objective, as mentioned above.

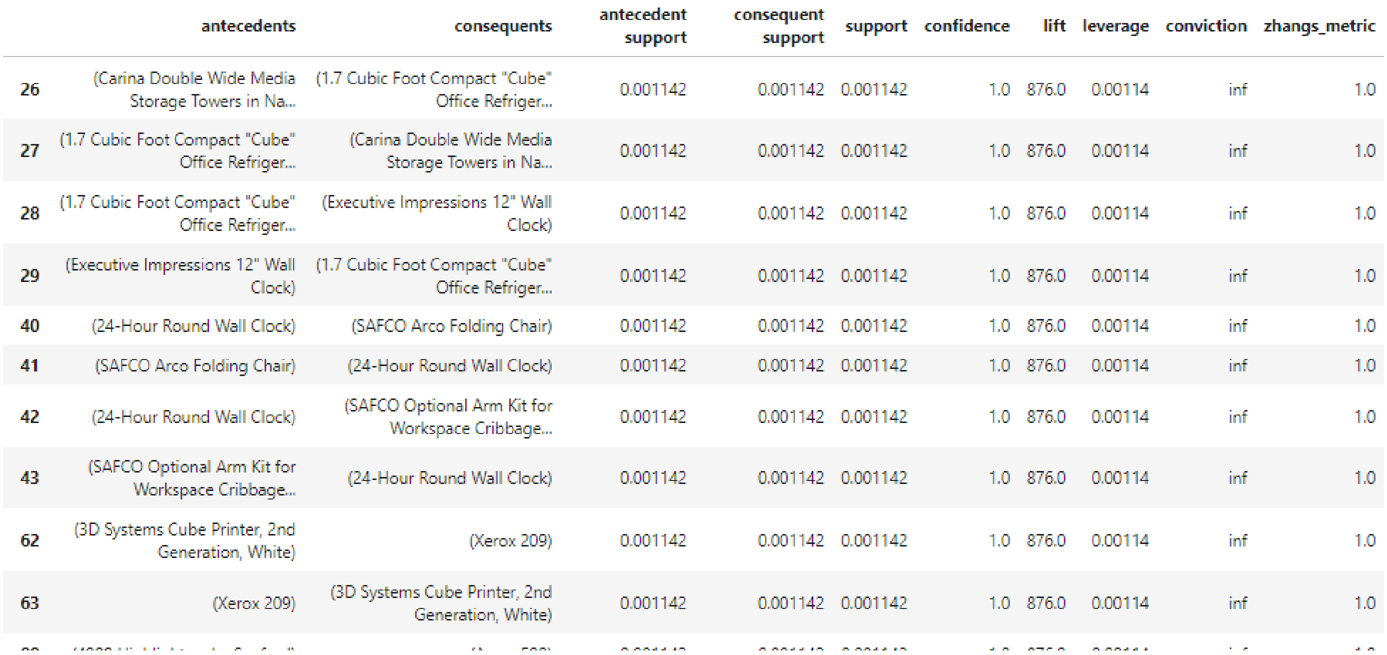


4. BASKET ANALYSIS

- Below are pairs of subcategories that were frequently bought together



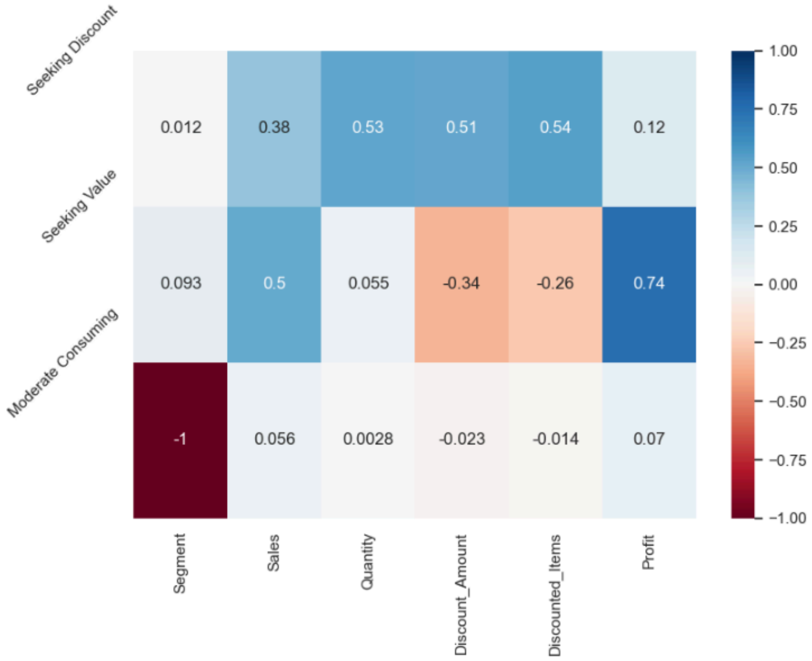
- **The basket of Consumer segment in 2017**



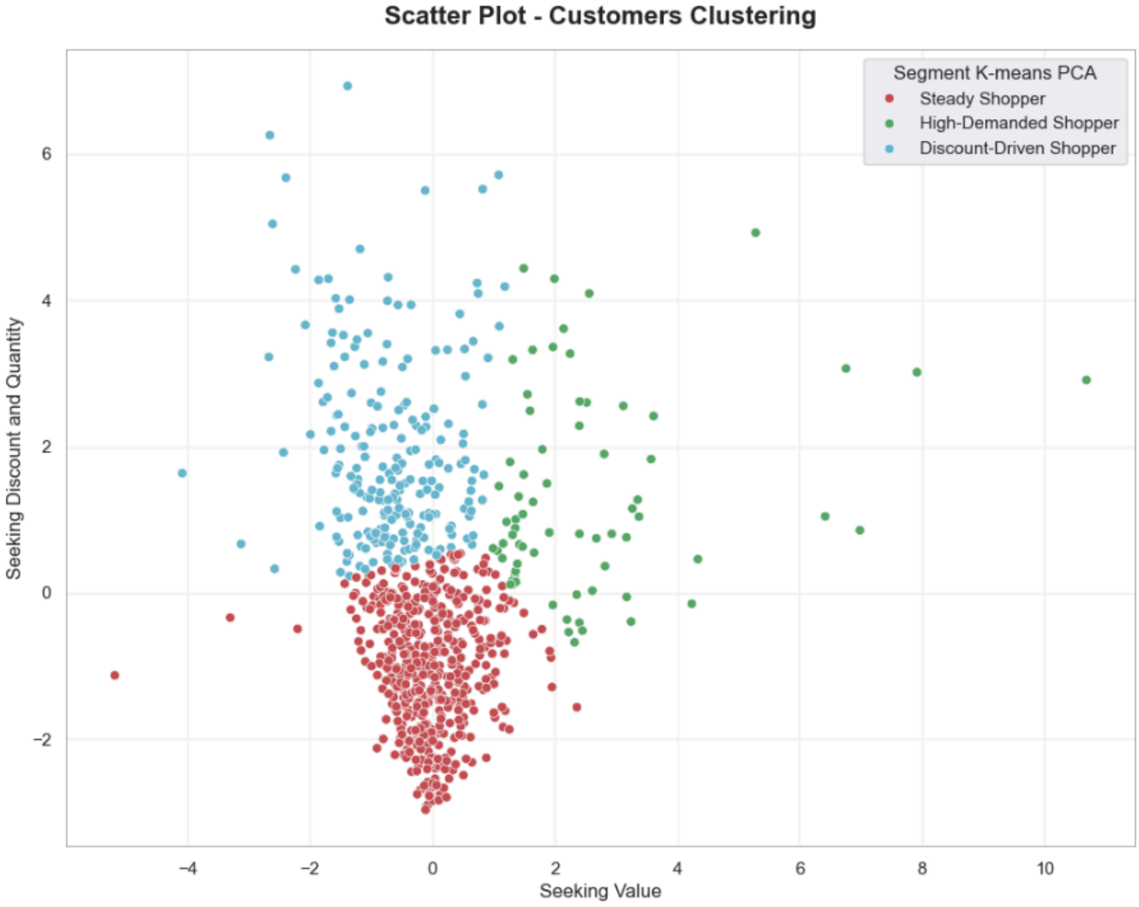
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
26	(Carina Double Wide Media Storage Towers in Na...	(1.7 Cubic Foot Compact "Cube" Office Refriger...	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
27	(1.7 Cubic Foot Compact "Cube" Office Refriger...	(Carina Double Wide Media Storage Towers in Na...	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
28	(1.7 Cubic Foot Compact "Cube" Office Refriger...	(Executive Impressions 12" Wall Clock)	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
29	(Executive Impressions 12" Wall Clock)	(1.7 Cubic Foot Compact "Cube" Office Refriger...	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
40	(24-Hour Round Wall Clock)	(SAFCO Arco Folding Chair)	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
41	(SAFCO Arco Folding Chair)	(24-Hour Round Wall Clock)	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
42	(24-Hour Round Wall Clock)	(SAFCO Optional Arm Kit for Workspace Cribbage...	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
43	(SAFCO Optional Arm Kit for Workspace Cribbage...	(24-Hour Round Wall Clock)	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
62	(3D Systems Cube Printer, 2nd Generation, White)	(Xerox 209)	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0
63	(Xerox 209)	(3D Systems Cube Printer, 2nd Generation, White)	0.001142	0.001142	0.001142	1.0	876.0	0.00114	inf	1.0

5. CUSTOMER CLUSTERING WITH K-MEANS AND PCA

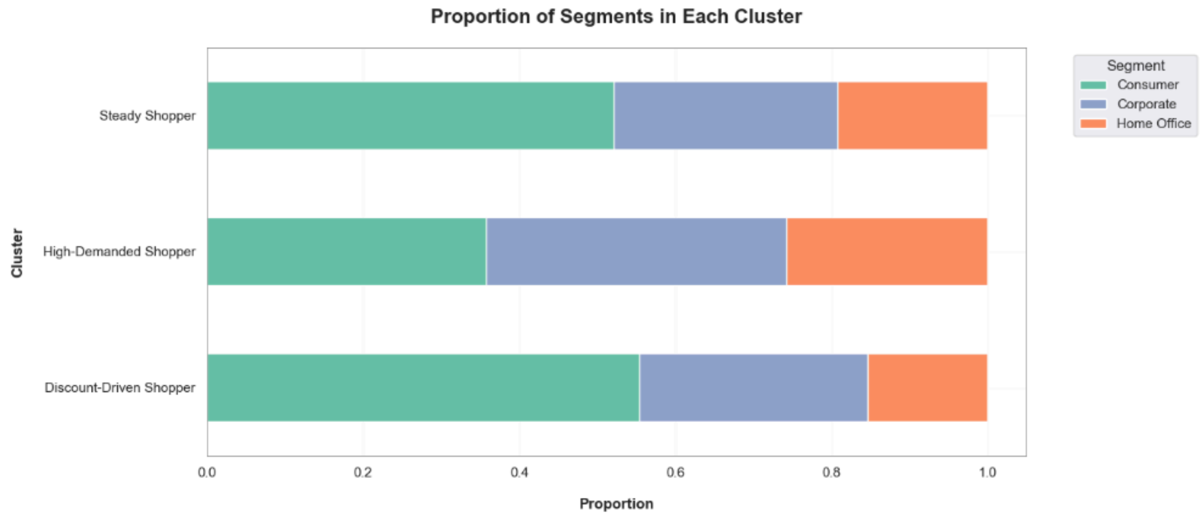
As there are multiple columns reflecting customer behaviours and characteristics, I applied PCA to reduce them to 3 components: Seeking Discount, Seeking Value, and Moderate Consuming

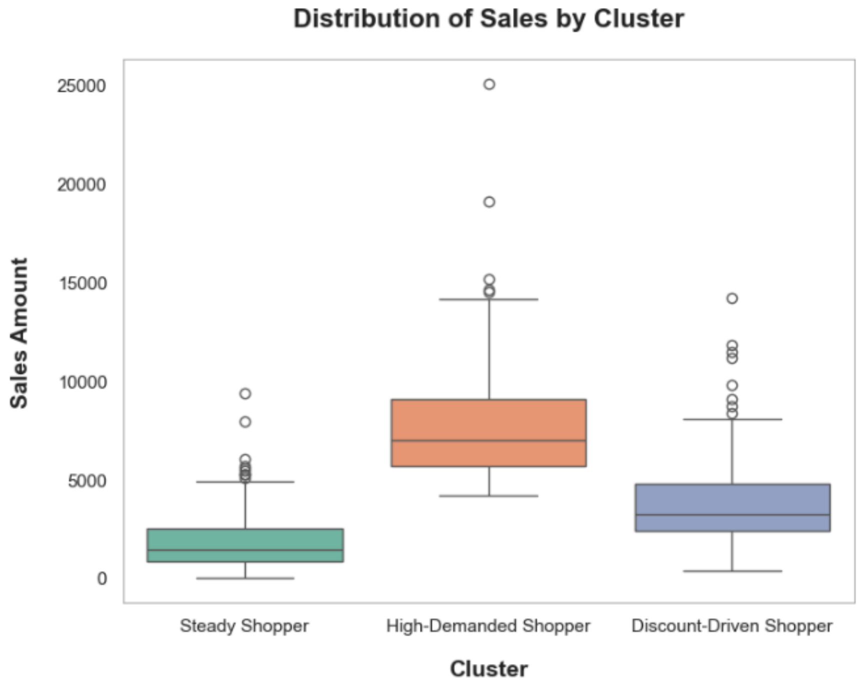


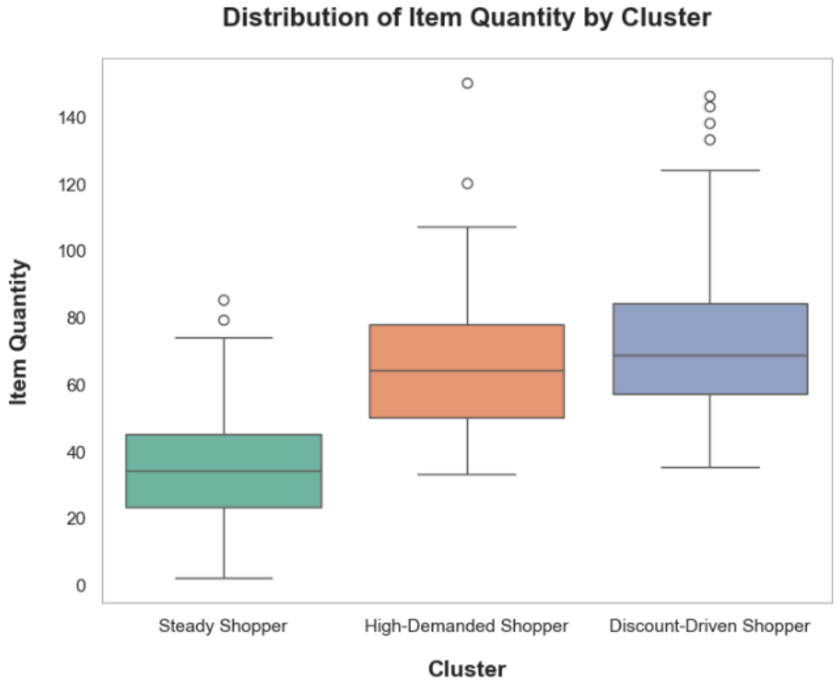
- Elbow and Silhouette methods were leveraged to determine the number of customer clusters: customers would be segmented to 3 clusters
- Based on their characteristics, I named them
  - Steady Shopper
  - High-Demanded SHopper
  - Discount-Driven Shopper

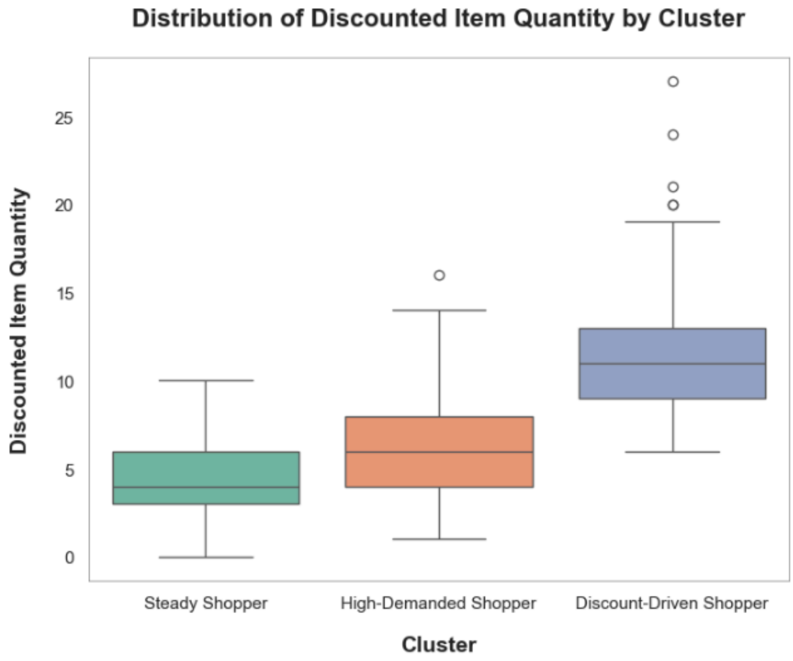


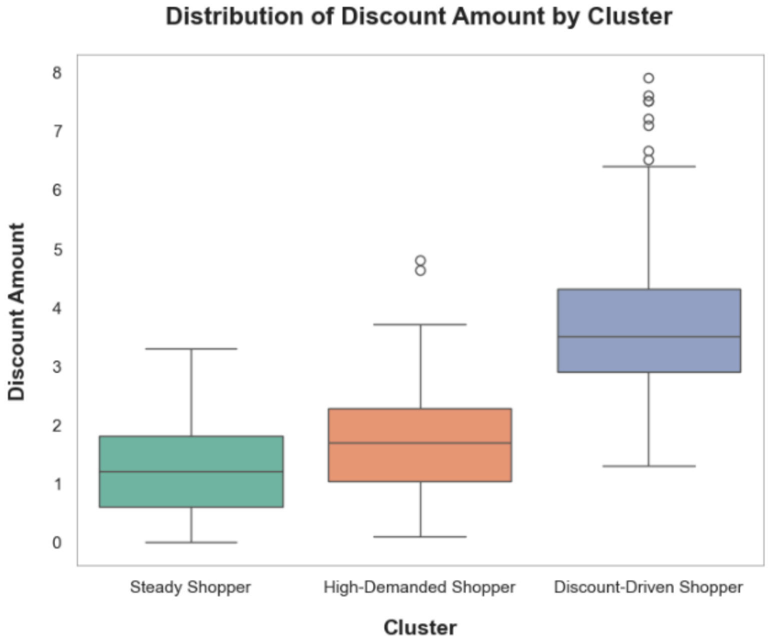
- Examining these clusters:

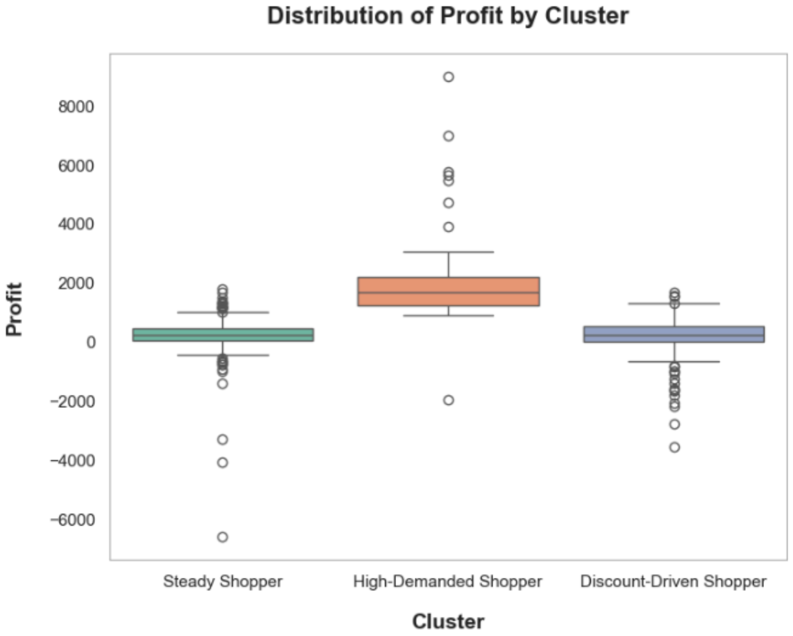







Distribution of Discounted Item Quantity by Cluster
Discounted Item Quantity
0
5
10
15
20
25
Steady Shopper
High-Demanded Shopper
Discount-Driven Shopper
Cluster


Distribution of Discount Amount by Cluster
Discount Amount
0
1
2
3
4
5
6
7
8
Steady Shopper
High-Demanded Shopper
Discount-Driven Shopper
Cluster


Distribution of Profit by Cluster
Profit
−6000
−4000
−2000
0
2000
4000
6000
8000
Steady Shopper
High-Demanded Shopper
Discount-Driven Shopper
Cluster

Steady Shopper	Discount-Driven Shopper	High-Demanded Shopper
Accounting for 63% of customer base	Accounting for 28% of customer base	Only a minority ~ 9% of the total
Half of them were Consumers	More than 50% are Consumers	A balance among 3 segments
Lowest Sales contributed	Moderate Sales contributed	Highest Sales contributed
Lowest Item Quantity bought	Highest Item Quantity bought	High Item Quantity bought
Rarely buying items with promotion	Discount lovers	Not preferring buying discounted items
Low profit brought to company, some loss	Low profit brought to company, many loss	High profit resulted
They buy small quantities and don’t prefer discounts. Due to their majority in customer base, they help maintain good consumption and cash flow, even though they don’t bring much profit to the company	They buy a lot of products, due to discounts offered. That’s why the profit from them is relatively low, although they contribute moderate level of sales	They only buy demanded products and have an aversion to discounts, that’s why they contribute a large amount of sales and profit.