CUSTOMER SEGMENTATION

Final Project Presented by
Mayugari
Coach By
Anwar Sanusi

TABLE OF CONTENT

BUSINESS & DATA UNDERSTANDING

BUSINESS UNDERSTANDING

Background

The dataset is from an E-commerce company which contain the sales data from 2015 to 2018. There are 9800 rows and 18 columns that shows details about the customer behaviour such as customer segment, customer name, ship mode, location, and product ordered.

Objective

How to utilize the customer's behavioral data to improve Customer Relationship Management (CRM) and drive opportunity sales growth.

ANALYTICAL APPROACH

ANALYTICAL APPROACH

To analyze the data, we used two way of analytical approach, which are RFM Analysis and Unsupervised Learning.

- RFM Analysis is a method that segmenting the customer into categories based on their Recency (the amount of time since the customer's most recent transaction), Frequency (the total number of transactions made by the customer), and Monetary (the total amount that the customer has spent across all transactions).
- While the Unsupervised Learning is a machine learning techniques that required training an algorithm on the dataset devoid of any labeled data. In this case, we use this method to clustered the data into segments of the customer.

Exploratory Data Analysis & Visualization

EDA & DATA PRE-PROCESSING

The dataset contains information of :

- Row ID: The number of rows
- Order ID: The ID of the order
- Order Date: The date the product ordered
- Ship Date: The date the product shipped
- Ship mode: The type of shipping mode
- Customer ID: The ID of the customer
- Customer Name: The name of the customer
- Segment: The segment type of customer
- Country: The country of the customer
- State: The state of the customer
- City: The city of the customer
- Region: The region of the customer
- Postal Code: The postal code of customer
- Product ID: The ID of the product
- Product Name: The name of the product
- Category: The category of the product
- Sub-Category: The sub-category of the product
- Sales: The value of the sales

EDA & DATA PRE-PROCESSING

Missing Value :

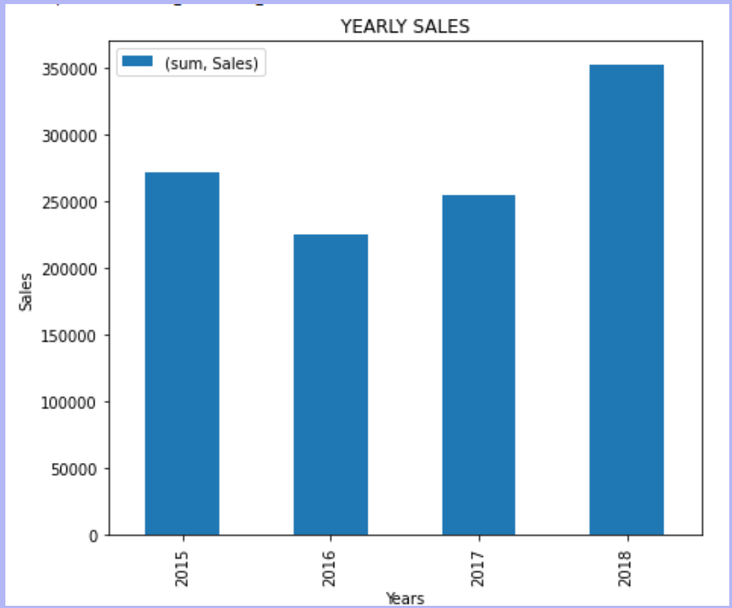
- There are 11 missing values in the Postal Code column. All from the same city (Burlington, Vermont)
- The missing values are solved by filled the postal code found from Google

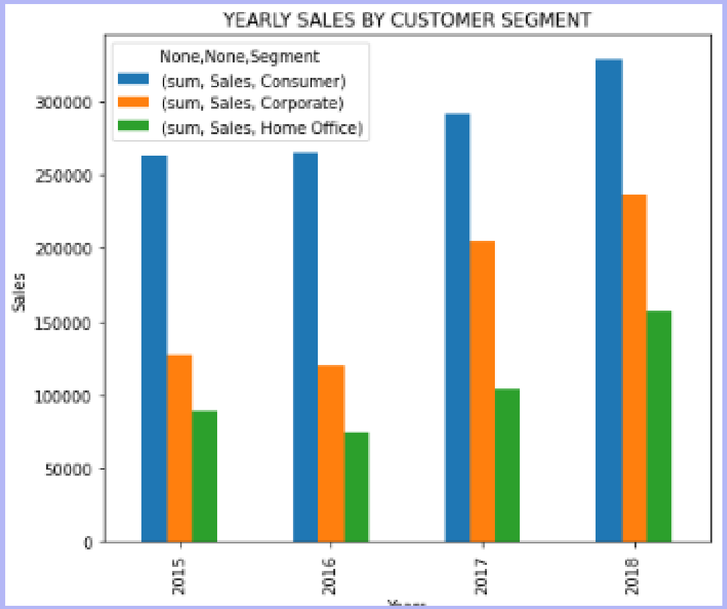
Duplicate Value :

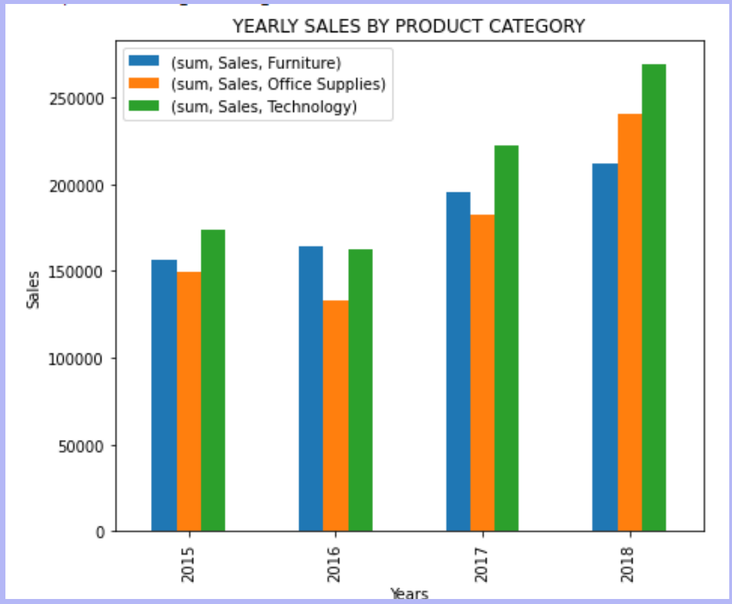
- There are no duplicate value in this dataset

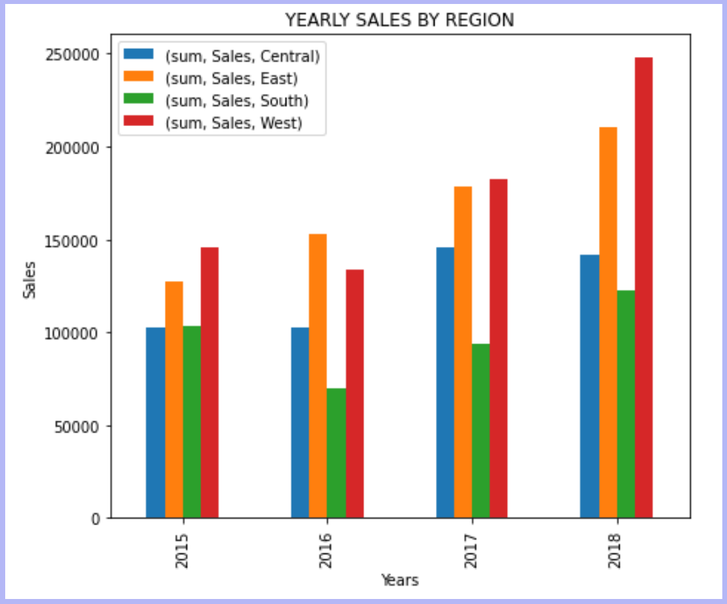
Change the datatypes :

- The Order Date and Ship Date columns have string data type, which needs to be changed to DateTime
- The Postal Code column has float data type, which needs to be changed to integer
- Removing Row ID, as it's not necessary



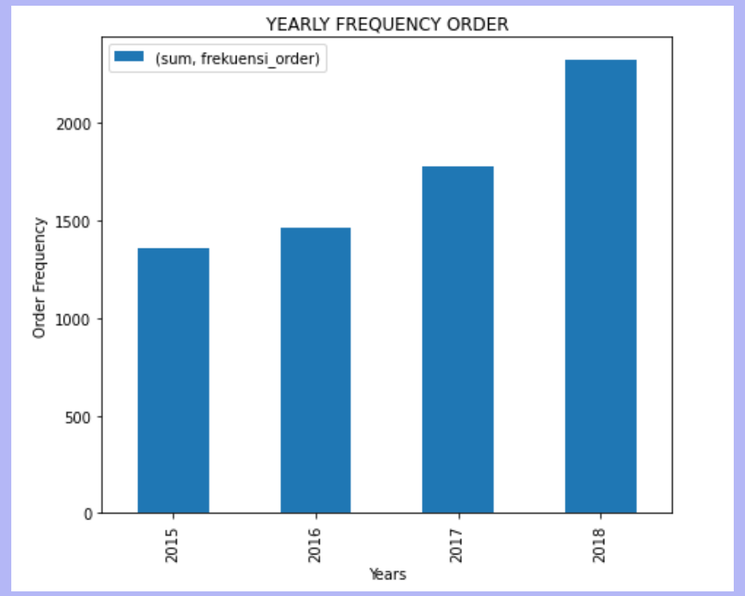


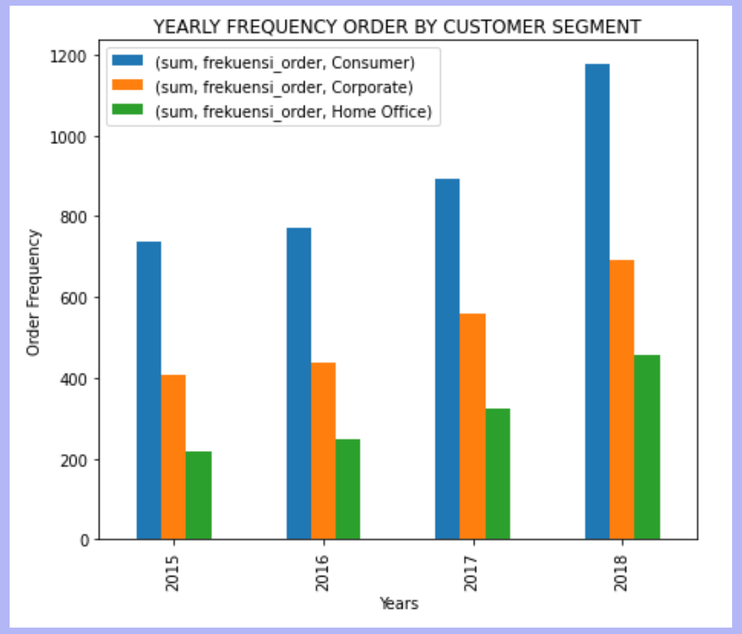


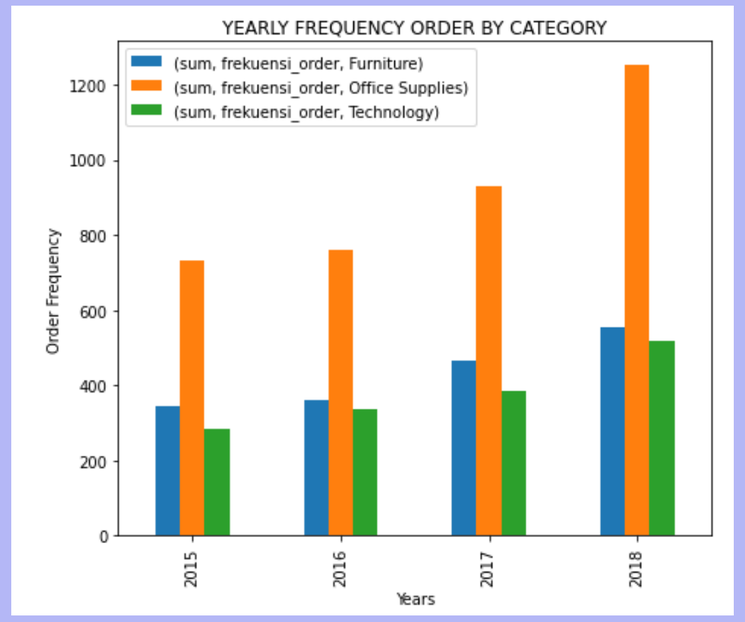


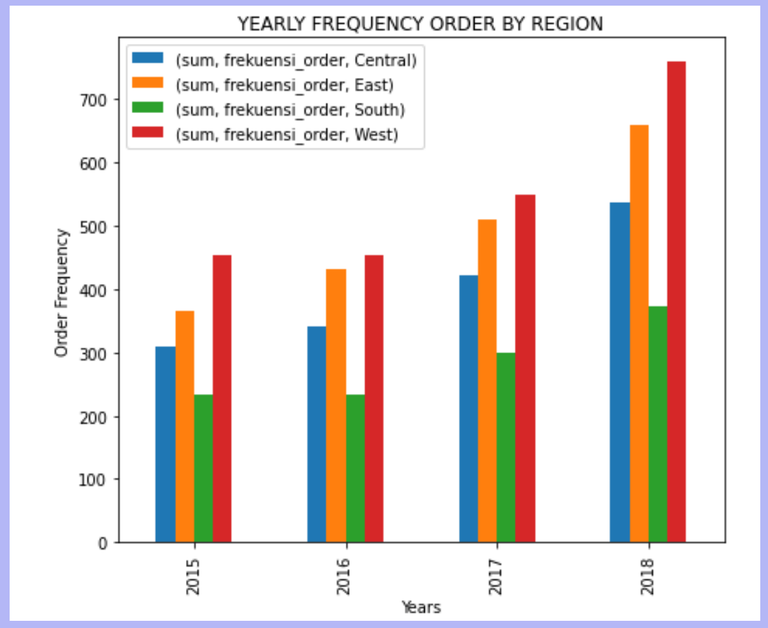
BIVARIATE ANALYSIS TO SALES

- Sales decerease by 2016 but consistently growth since 2017-2018.
- Top contributor sales drive by consumer segment, but corporate segment has biggest growth by 2017.
- There's an opportunity to drive growth of sales on home office category.
- Technology is top contributor sales of category product.
- Furniture category consistently growth year by year.
- West is top contributor sales of region area.
- East region consistently growth year by year.
- There's an opportunity to drive growth of sales in South Region.



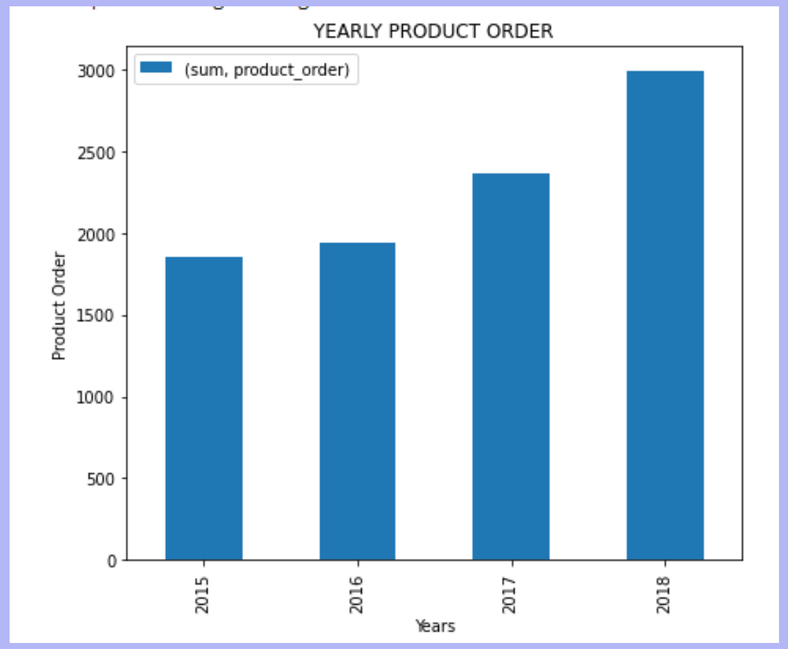


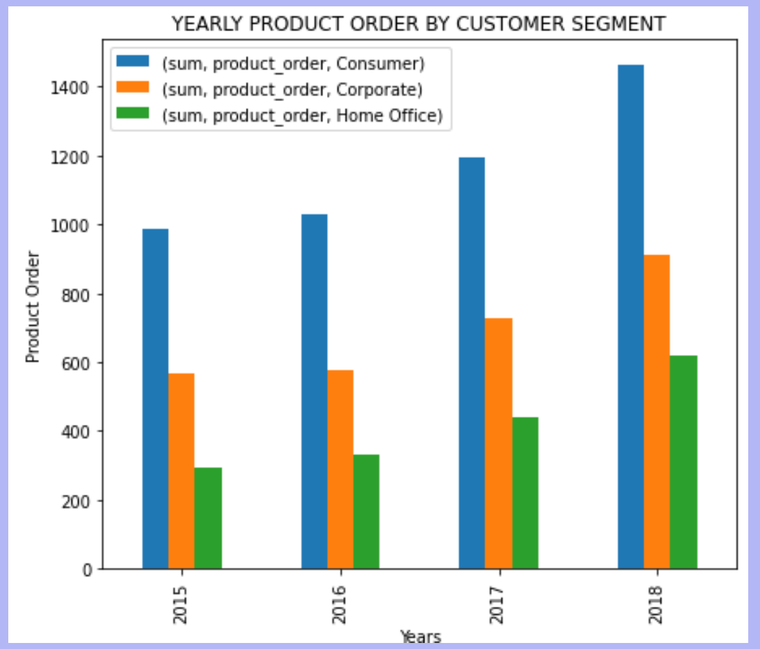


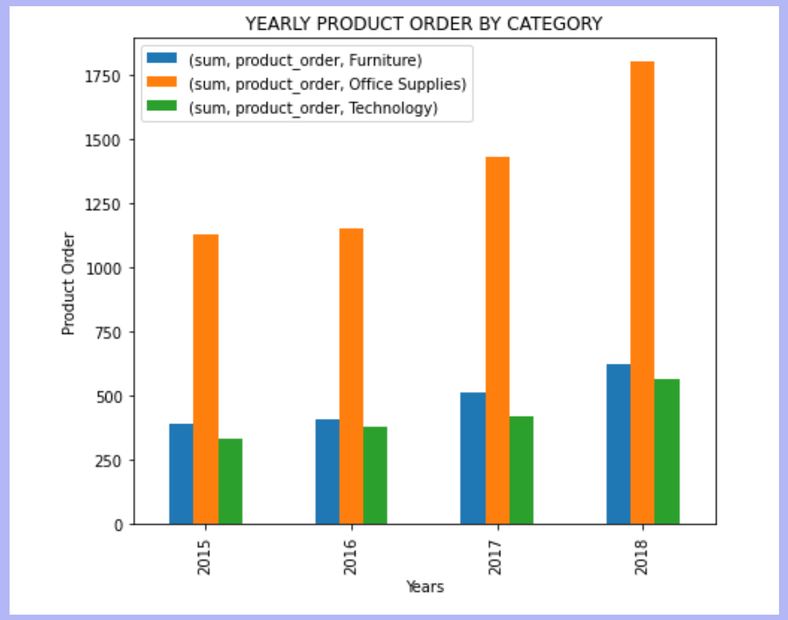


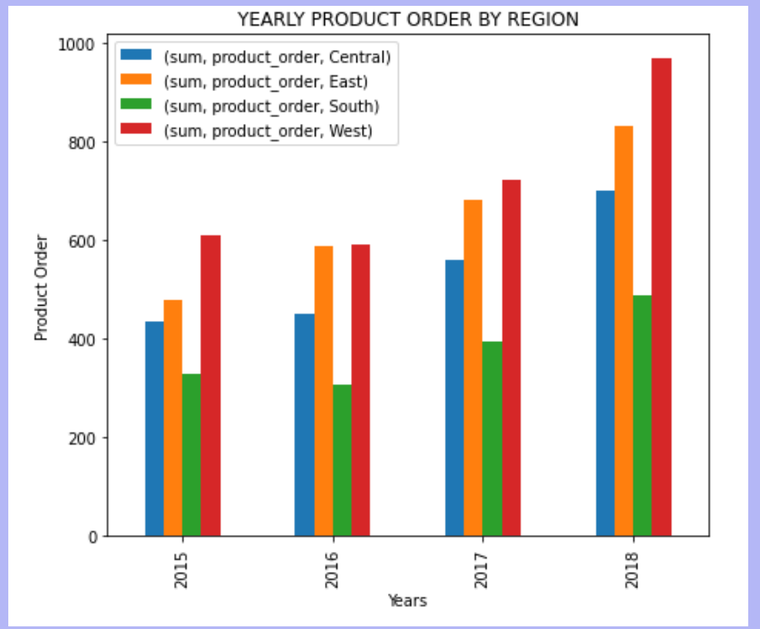
BIVARIATE ANALYSIS TO FREQUENT ORDER

- Frequency order customer consistently increase year by year. Compare to sales 2016 customer more frequent order but less spent.
- Each segment of customer consistently increase frequency order year by year.
- Each category product consistently increase frequency order year by year.
- There're opportunity on furniture and technology category to drive more frequent of order.
- South and West Region looks decrease frequent order by 2016.
- There's an opportunity to drive more frequent of order in South Region.









BIVARIATE ANALYSIS OF PRODUCT ORDER

- Count of product variance customer consistently increase year by year.
- Segment of customer relatively increase product variance order year by year.
- There's opportunity on home office segment to drive more product variance order.
- Each category product consistently increase product variance order year by year.
- There're opportunity on furniture and technology category to drive more product variance order.
- West Region has biggest growth on product variance order.
- There's an opportunity to drive more product variance order in South Region.

EDA

Explore frequency order, count of product order and total sales by time series, customer segment, category product and region customer.

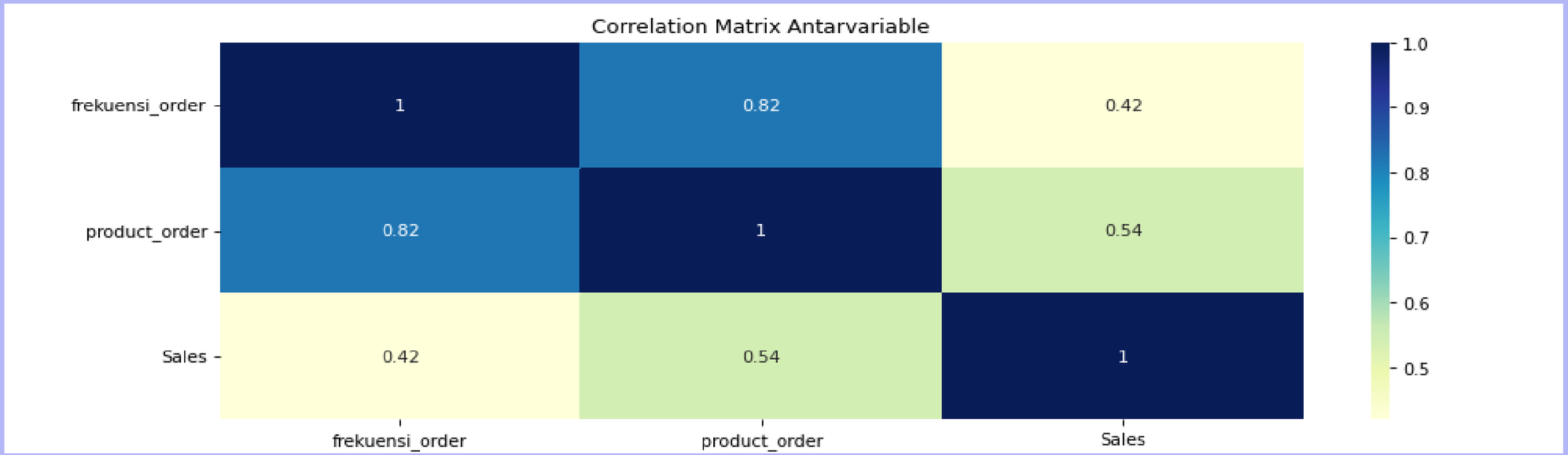
Order Year	Segment	Category	Region	frekuensi_order	product_order	Sales
2015	Consumer	Furniture	Central	45	52	20028.0196
2015	Consumer	Furniture	East	50	57	26679.3480
2015	Consumer	Furniture	South	29	32	12045.9945
2015	Consumer	Furniture	West	66	71	26991.9315
2015	Consumer	Office Supplies	Central	94	158	26707.0080

Explore frequency order, count of product order and total sales for each customer.

Customer ID	frekuensi_order	product_order	Sales
AA-10315	5	11	5563.560
AA-10375	9	15	1056.390
AA-10480	4	12	1790.512
AA-10645	6	18	5086.935
AB-10015	3	6	886.156

- Frequency order based on unique order ID.
- Product order based on unique product ID.
- Total sales sum of sales for each customer on the data set

CORRELATION MATRIX



- Sales , product variance order and frequent order have strong positive correlation.
  - More variance product and frequent of customer order, customer will spent more purchased in this E-commerce.
- The strongest correlation between variance product order and frequent order customer.

MODELLING & RECOMMENDATION

RFM Analysis

- RFM analysis is a data-driven customer behavior segmentation technique.
- RFM stands for recency, frequency, and monetary value.
- The idea is to segment customers based on when their last purchase was, how often they've purchased in the past, and how much they've spent overall. All three of these measures have proven to be effective predictors of a customer's willingness to engage in marketing messages and offers.
- Benefit of RFM Analysis including
  - Personalization: By creating effective customer segments, you can create relevant, personalized offers
  - Improve Conversion Rates: Personalized offers will yield higher conversion rates because your customers are engaging with products they care about.
  - Improve unit economics
  - Increase revenue and profits

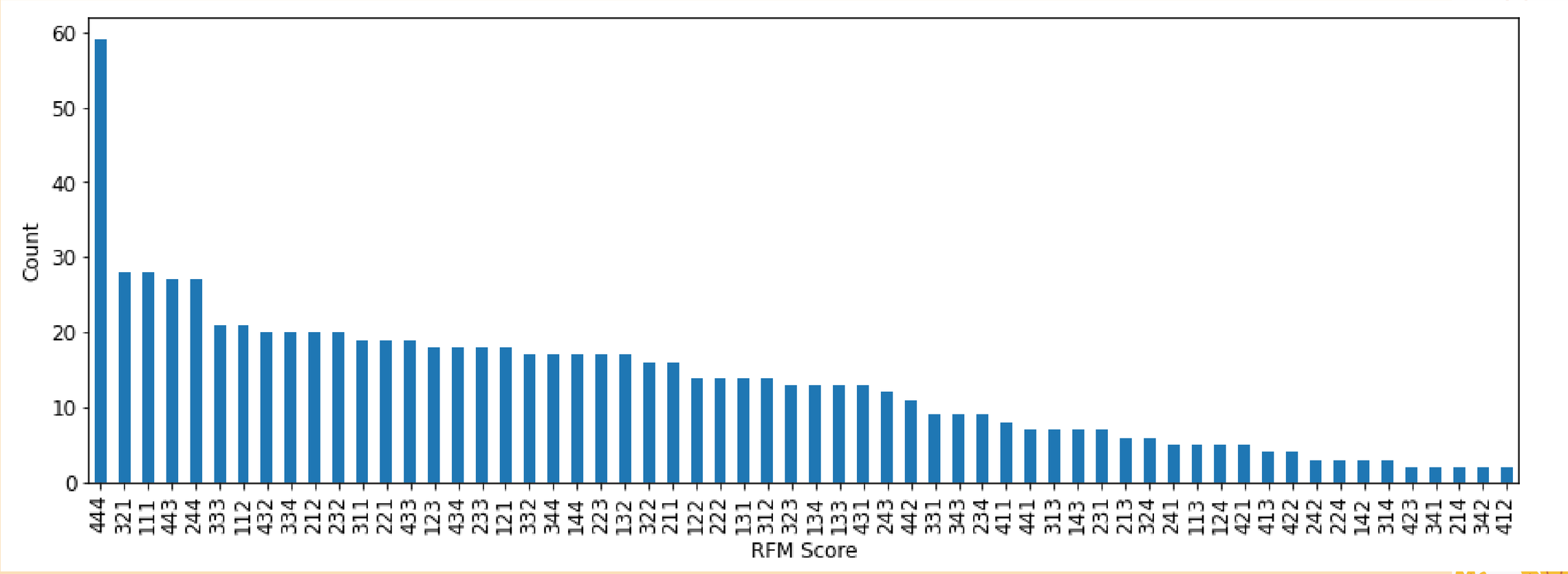
RFM Analysis

- Every customer will be rated for each metric into four categories, with one as the highest and 4 as the lowest
- The rating will be determined based on how each data stand between the quartile range of each metrics
- For example, if the customer has RFM score of 231, it means he/she ranked 2 in recency (last purchasing between 39-98 days) , 3 in frequency (purchasing 6-8 times, and 1 in monetary (money spent between USD 3670.258 - USD 25043.050

	Q_Recency	Q_Frequency	Q_Monetary
min	2.0	1.0	4.833
q1	39.0	4.0	1081.466
q2	98.0	6.0	2215.002
q3	223.0	8.0	3670.258
q4	1167.0	17.0	25043.050

RFM Analysis

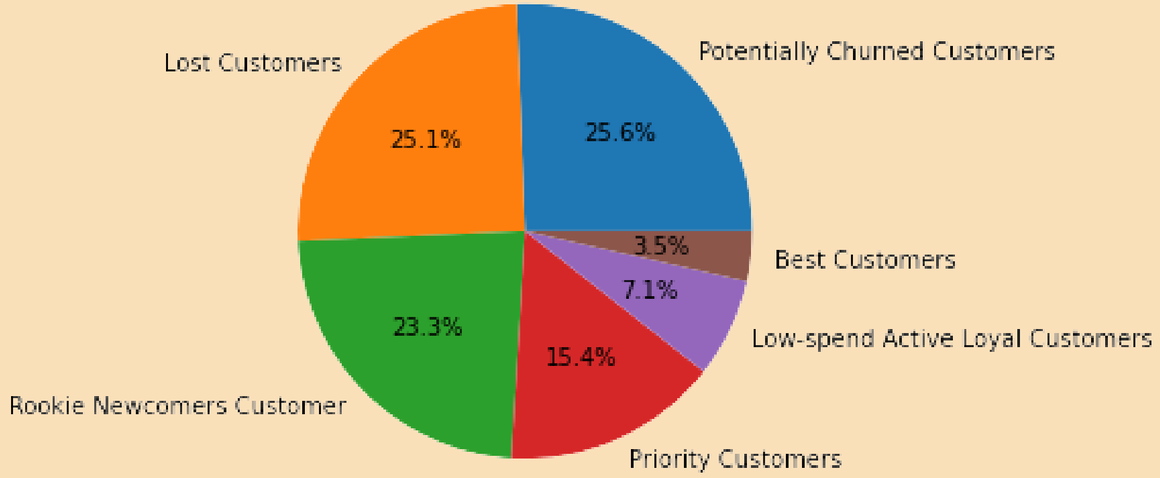
The results can be seen as follows:



RFM Analysis

Based on the RFM score, we divide the customers into segments which are:

- Best Customer
- Priority Customer
- Low Spend but Loyal & Active Customer
- Rookie Newcomers Customer
- Potentially Churned Customer
- Lost Customer



RFM Analysis

- Best Customer:
  - The company’s ace. They are active, loyal, and willing to spend their money
  - RFM Score: 111
  - Action plan for this segment:
    - Focused on giving them the best customer service possible
    - Paid VIP loyalty program, which includes things such as loyalty card with points that can be redeemed for free products, customized referral codes, early access to new products, and customized UI UX experience

RFM Analysis

- Priority Customer:
  - Not as important as the Best segment, yet still a valuable asset for the company
  - RFM Score 112, 121, 211, 122, 212, 221, 222
  - Action plan for this segment:
    - Focused on generating more customer referrals from them, since this segment is proven had a good experience with us
    - Referral code that could be redeemed for a discount on their next purchase, discount on special holidays that are usually associated with presents (friends' birthdays, valentine, Christmas, etc)

RFM Analysis

- Low Spend but Loyal & Active Customer:
  - High engagement, but not generating much money.
  - RFM Score: 113, 114, 123, 124, 213, 214, 223, 224
  - Action plan for this segment:
    - Focused on making them increase their shopping value since this segment is proven already loyal to the company despite not spending much money
    - More discounts that will be given with minimum purchasing, slightly more expensive product recommendations similar to their previous purchasing

RFM Analysis

- Rookie Newcomers Customer:
  - Newcomers. It’s an important phase to make good first impressions
  - RFM Score: 131, 132, 133, 134, 141, 142, 143, 144, 231, 232, 233, 234, 241, 242, 243, 244
  - Action plan for this segment:
    - Focused on built trust, since this segment is starting to get to know us we need to make sure they will stay with us
    - More discounts that will be given for their next purchase, create a strong onboarding experience

RFM Analysis

- Potentially Churned Customer:
  - Customers whose last purchase has been a while ago. Have the potential to left for good unless we do something
  - RFM Score: 311, 312, 313, 314, 321, 322, 323, 324, 331, 332, 333, 334, 341, 342, 343, 344
  - Action plan for this segment:
    - Focus on something to make them kickstart their comeback. For some reason they stop shopping in us, find out why
    - Research to what makes them left us via (email, survey, etc), engage with something that reminds them of their last purchase (similar product), discount with time-limit (offers end at 48 hours)

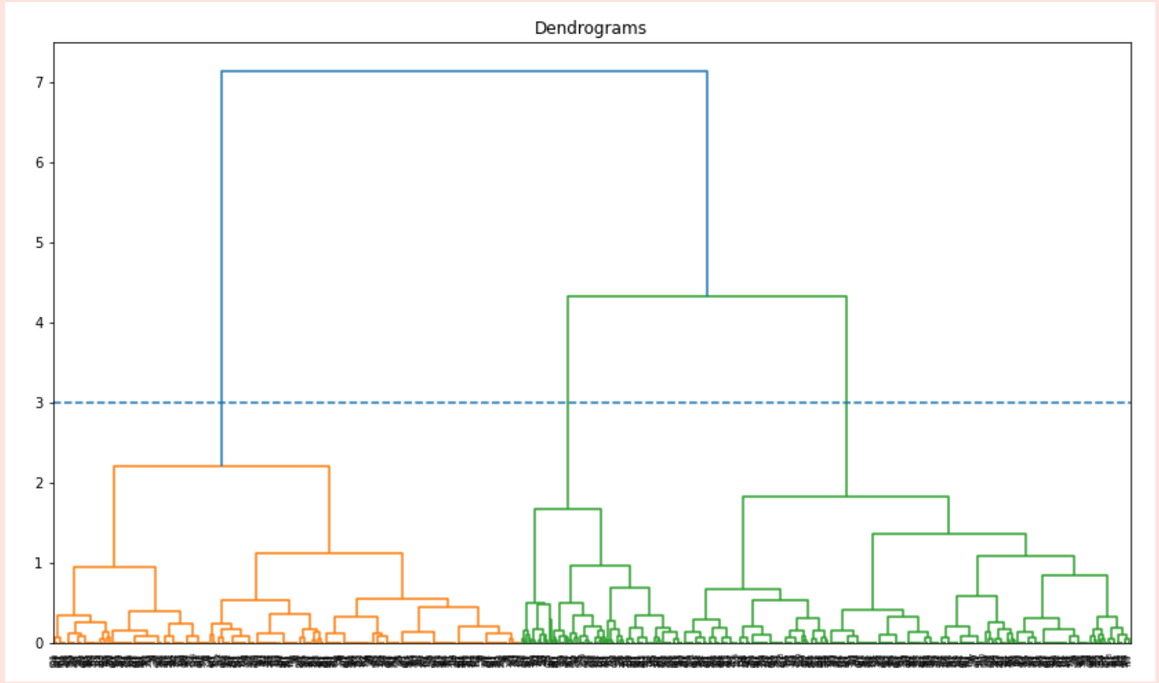
RFM Analysis

- Lost Customer:
  - Customers who are considered already lost and will not come back
  - RFM Score: 411, 412, 413, 414, 421, 422, 423, 424, 431, 432, 433, 434, 441, 442, 443, 444
  - Action plan for this segment:
    - Focus on finding what specific characteristic/reason made them leave us, and see if we can eliminate it and where exactly should we improve our services.
    - Find new customers to replace this segment

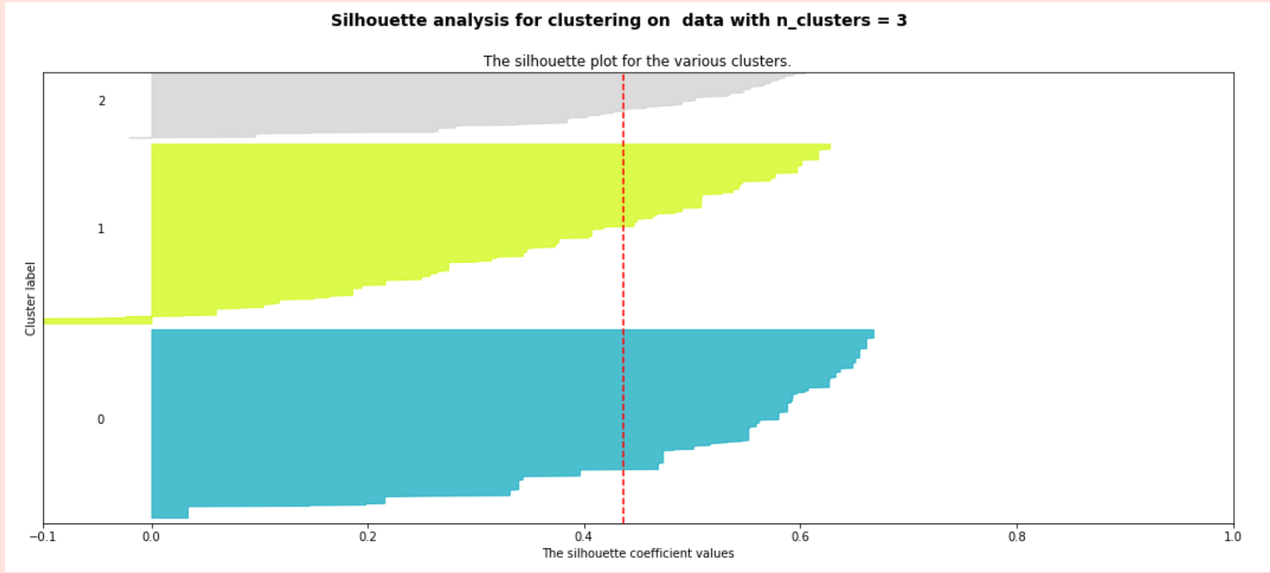
Hierarchical clustering

Product Order vs Frequency Order

- Objective drive variance product order and repeat order customer.
- According to the process we are using Min-maxscaler as for the data and also end up using Hierarchical Ward Linkage.

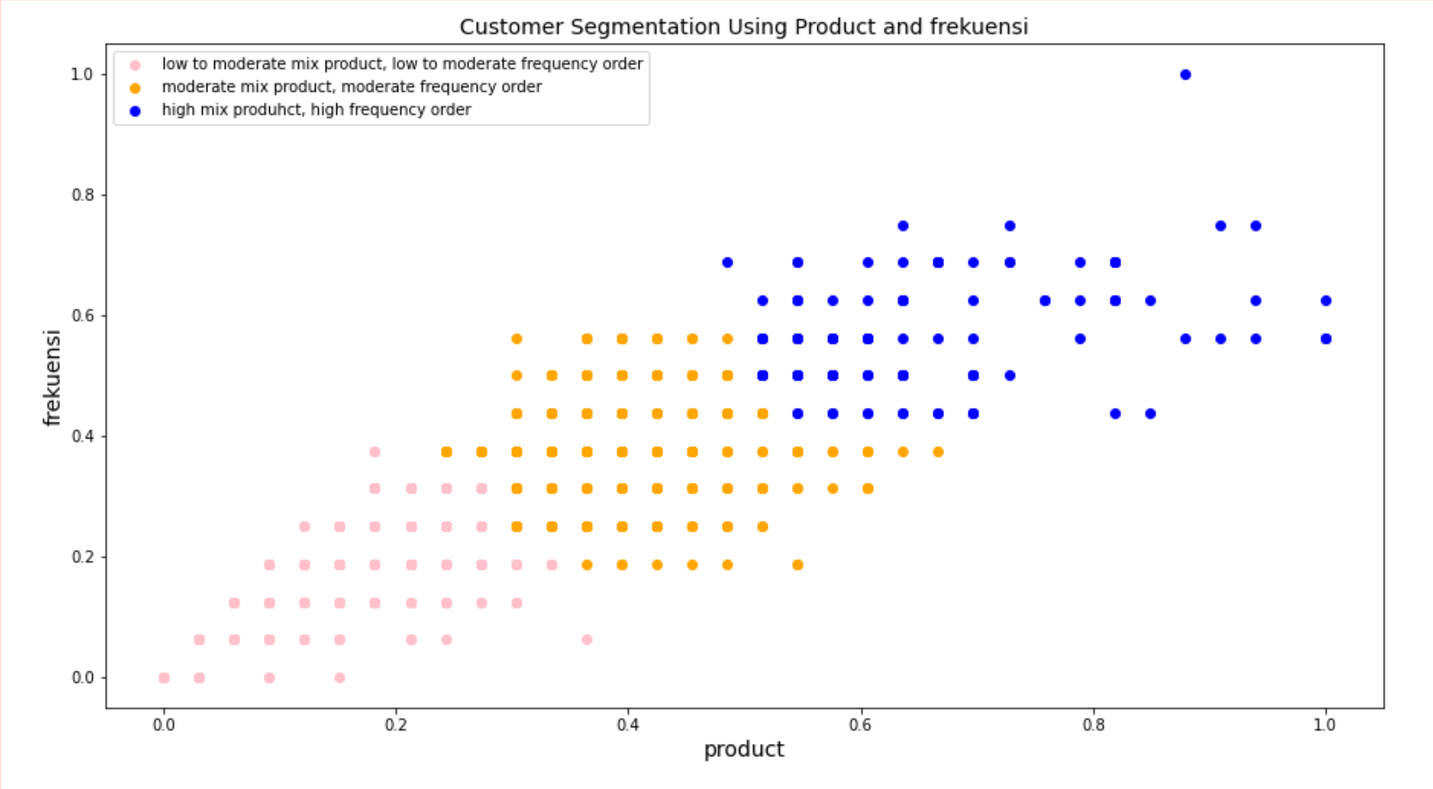


Refers to Dendograms the results of n-Clusters = 3



Hierarchical clustering

Product Order vs Frequency Order



CUSTOMER SEGMENT & RECOMENDATION

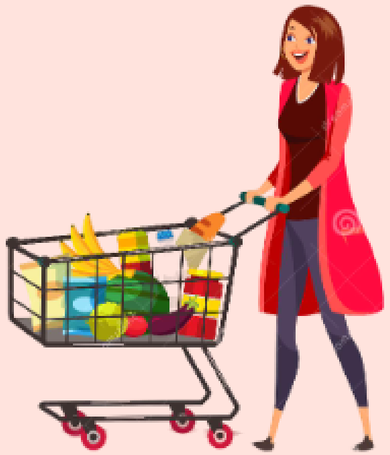
First Customer Segment

Low to moderate mix product, low to moderate frequency order.

Promo bundling product furniture and technology

Give voucher discout product for next purchase with minimum order minimal order,

Additional free shipping for South region



Second Customer Segment

Moderate mix product, moderate frequency order

Promo buy 3 get 1 cross category furniture and technology.

Give voucher discout product and free shipping for next purchase with minimum order minimal order.

Third Customer Segment
Monthly voucher discout product and voucher free shipping.

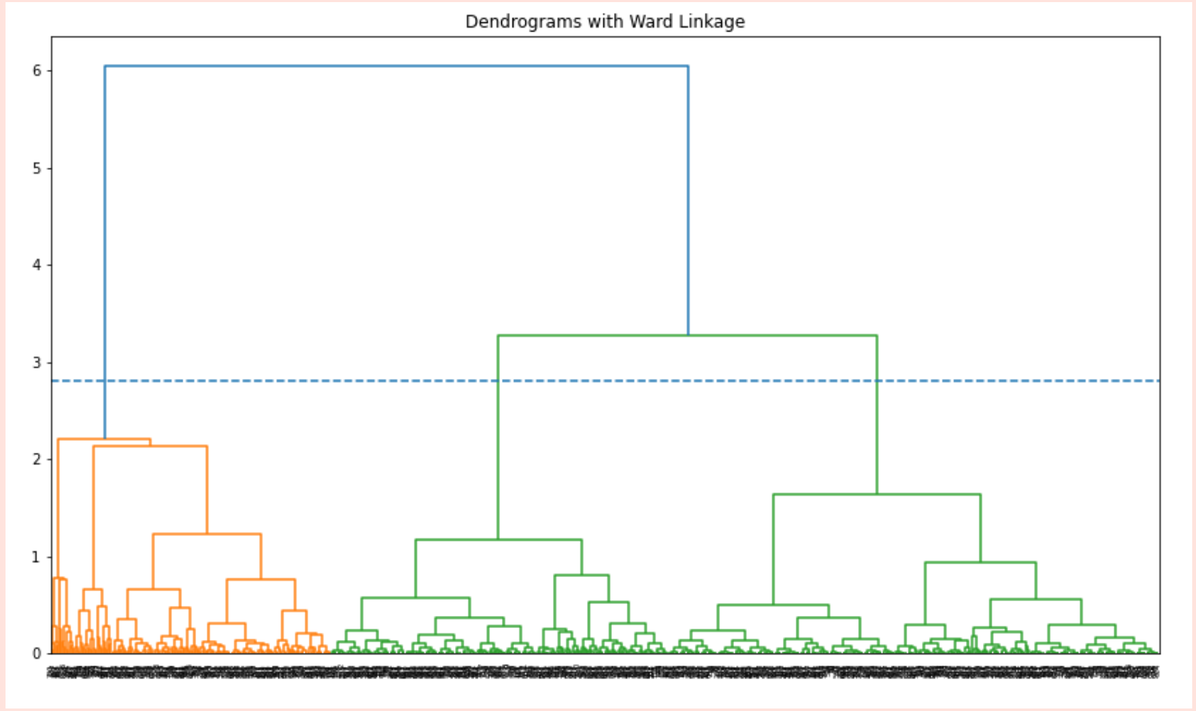
Promo buy 2 disc 10%, buy 3 disc 20% and more disc 30% cross category product furniture and technology

Voucher free shipping without minimum order with minimum spent for certain period.

Hierarchical clustering

Product Order vs Sales

- Objective drive growth of sales by additional frequent order customer.
- According to the process we are using Minmaxscaler as for the data and the output using Hierarchical Ward Linkage.

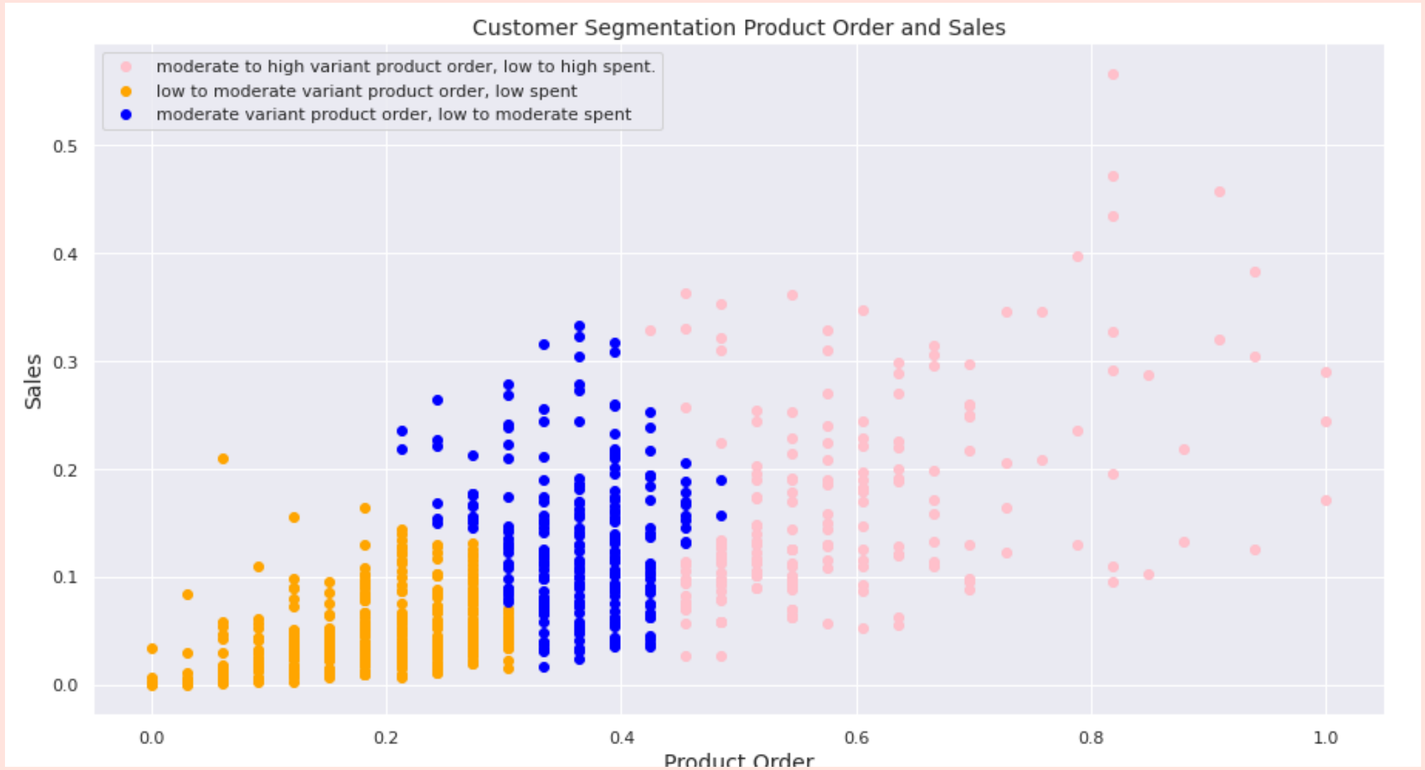


Refers to Dendograms the results of n-Clusters = 3

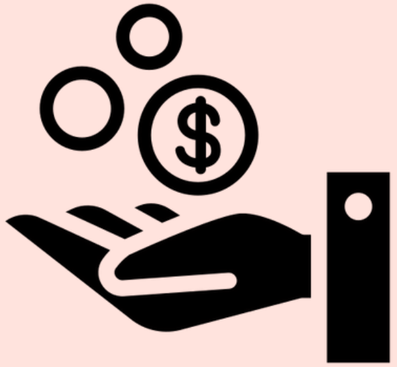
Hierarchical clustering

Product Order vs Sales

The Output.



CUSTOMER SEGMENT & RECOMENDATION



Passive Customer Segment

Low to moderate variant product order, low spent.

Promo buy 1 get 1 free product.

Voucher discount for next purchased without minimum order.

Additional free shipping for South region

Active Customer Segment

Moderate variant product order, low to moderate spent

Discount product with ordinal purchased (buy 2 kind of product disc 10%, buy 3 or more kind of product with min 3 pcs per product disc 30%)

Free shipping with minimum order.

Consumtive Customer Segment

Moderate to high variant product order, low to high spent.

Monthly voucher discount with min 2 kind of product without minimum order purchased.

Monthly voucher discout product and voucher free shipping.

CONCLUSION

Customer Segment by RFM Analysis to maintain Customer Relationship Management (CRM)

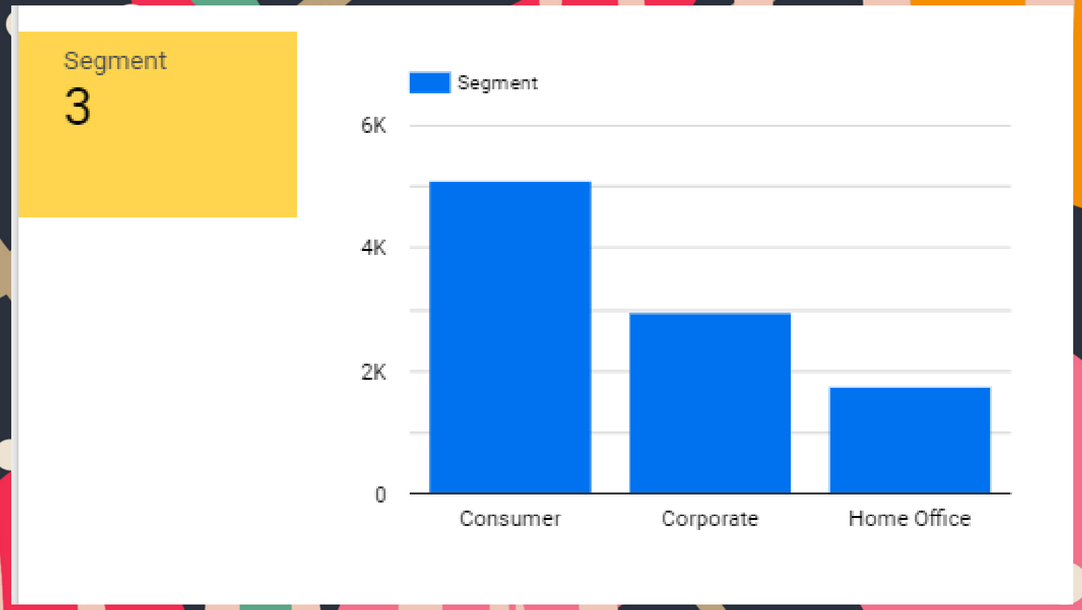
Customer segment by hierarchical ward linkage to drive variance product order and repeat order customer.

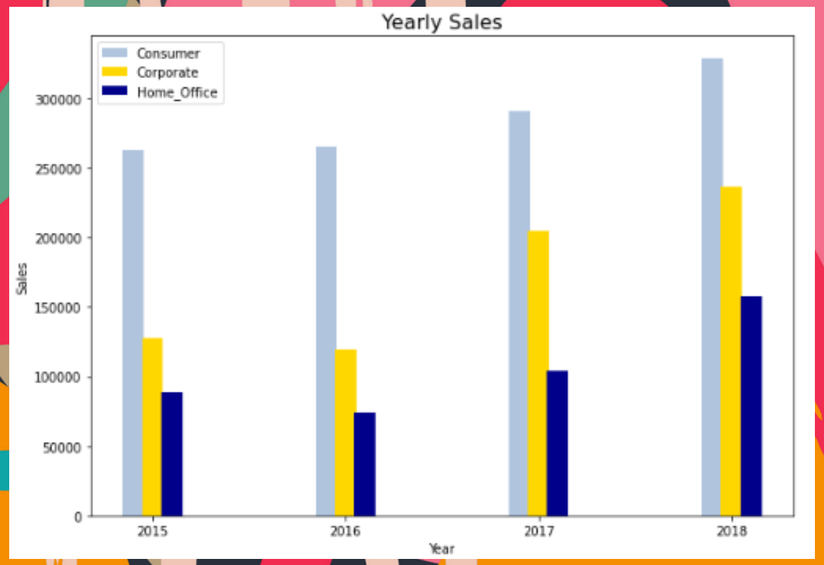
Customer segment by hierarchical ward linkage to drive growth of sales by additional frequent order customer.

THANK YOU

APPENDIX

Univariate Analysis





Insight berdasarkan Segment

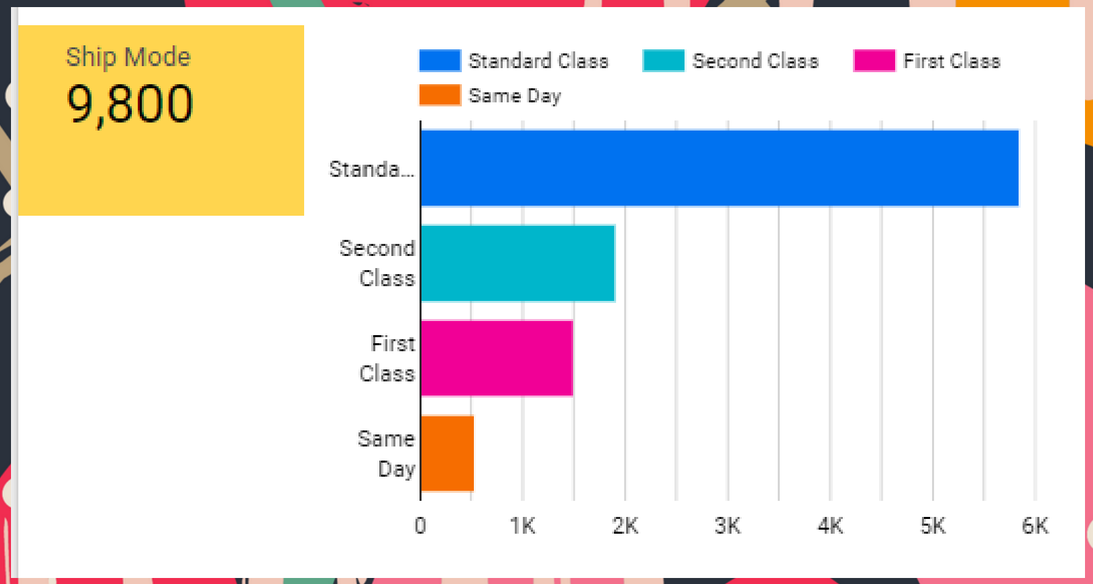
- Sales 2015 ke 2018 menunjukan positive trend untuk setiap segment, except Corporate & Corporate segment menunjukkan negative trend dari 2015 ke 2016.
- Consumer segment menunjukkan konstan growth year by year.
- Corporate segment menunjukkan high growth in 2017.
- Home office segment menunjukkan high growth in 2018.

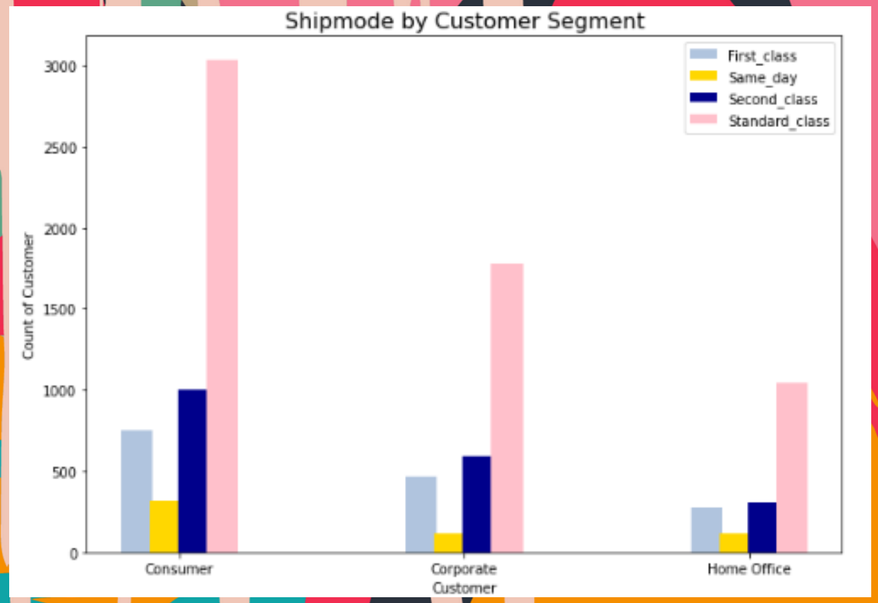
Notes: Segment terdiri atas 3, dimana terbanyak berasal dari Consumer dengan 5101, Corporate 2953, dan HO 1746.

Univariate Analysis

Insight berdasarkan Shipmode

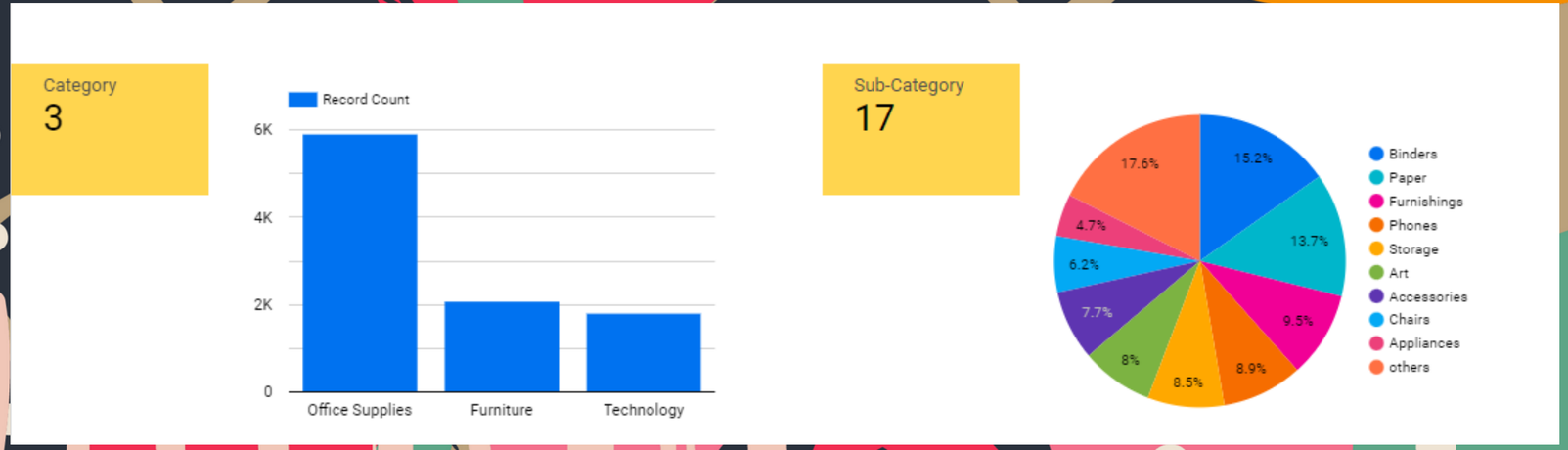
Notes: Shipmode terdiri atas 4, dimana terbanyak berasal dari Standard 5859, dan tedikir same day dengan 538





Univariate Analysis

Insight of category & sub-category



Notes: Berdasarkan Category, Chart Office Supplies mempunyai jumlah terbanyak yakni 5909, diikuti oleh Furniture 2078, dan Technology 1813

Notes: Berdasarkan Sub-category, Binders mempunyai pesanan terbanyak dengan 1492, sedangkan paling sedikit Copiers dengan 66

Univariate Analysis

Insight berdasarkan Product

Product Name
1,849

	Product Name	Record Count
1.	Staple envelope	47
2.	Staples	46
3.	Easy-staple paper	44
4.	Avery Non-Stick Binders	20
5.	Staple remover	18
6.	Staples in misc. colors	18
7.	Storex Dura Pro Binders	17
	Grand total	9,800

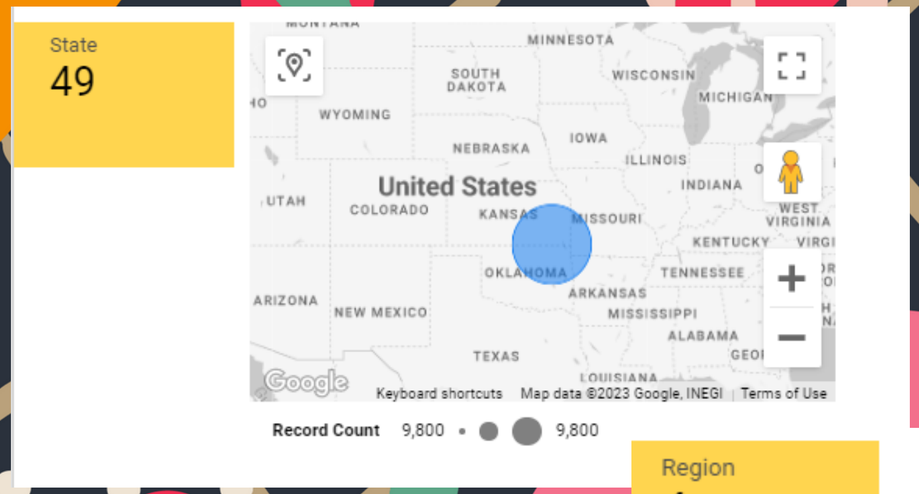
1 - 100 / 1849

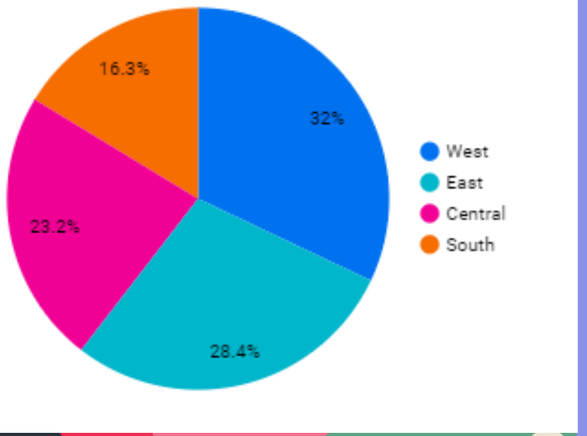
Univariate Analysis

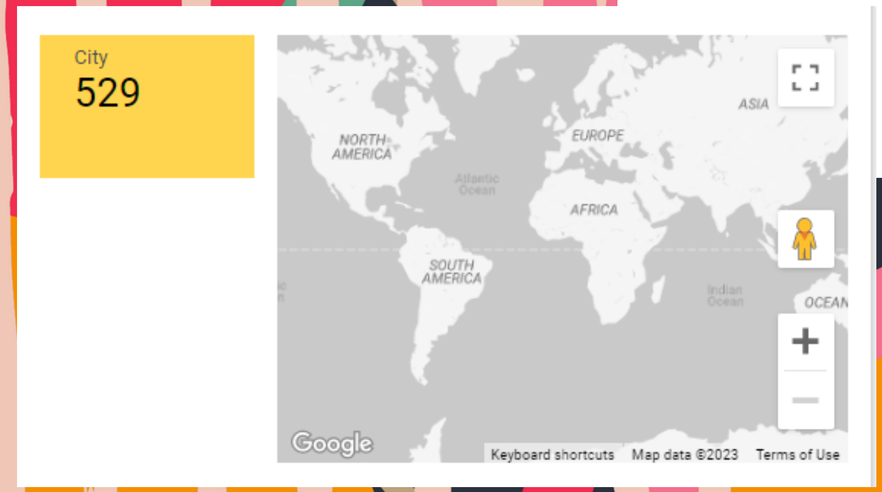
Insight berdasarkan Location

Berdasarkan State seluruhnya berasal dari USA dan dari berbagai City di dalamnya.

Dimana terbagi menjadi 4 Daerah , dimana West mendiami peringkat pertama dan South menjadi yang terakhir

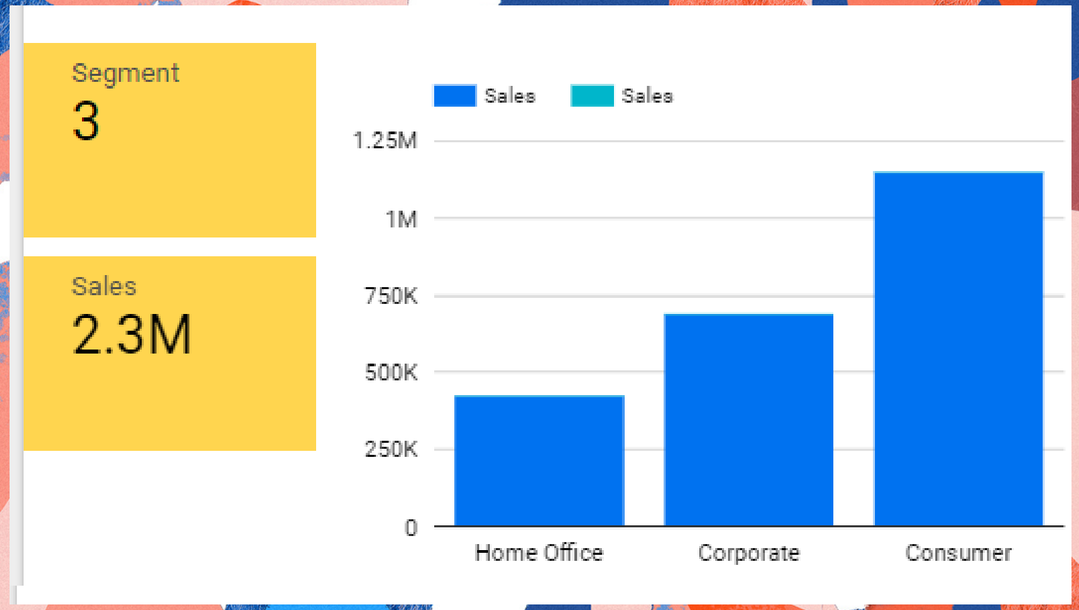






Bivariate Analysis

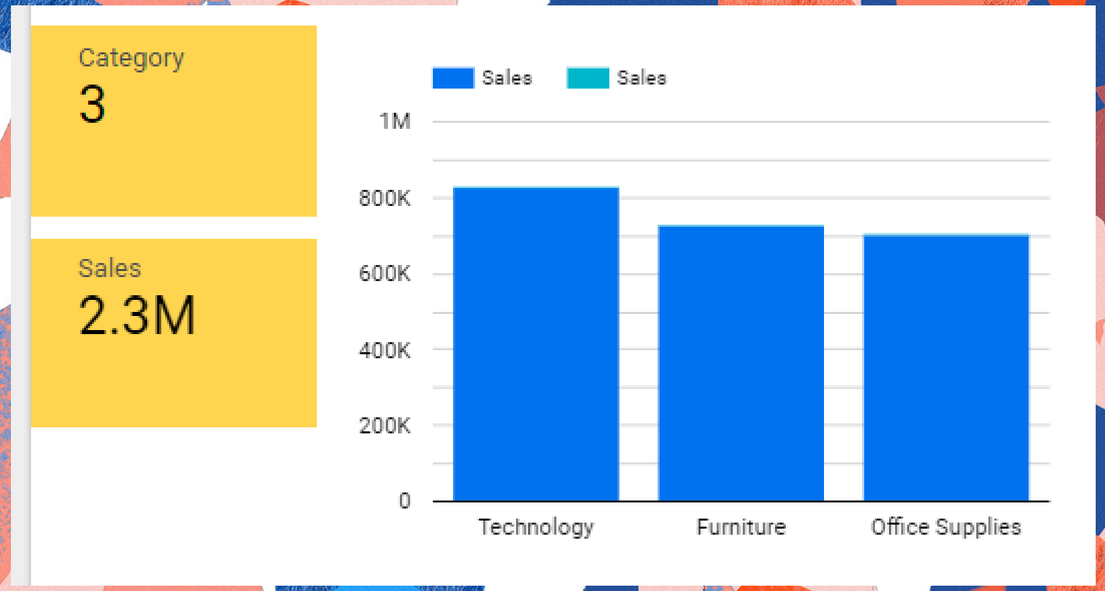
Insight berdasarkan Segment & Sales

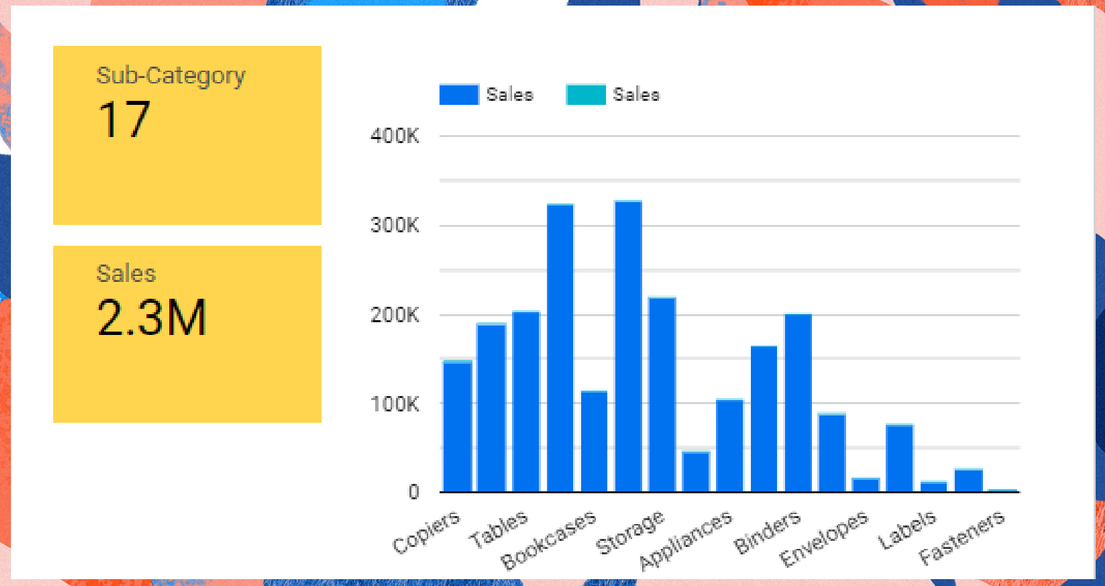


Insight berdasarkan Shipmode & Sales

Bivariate Analysis

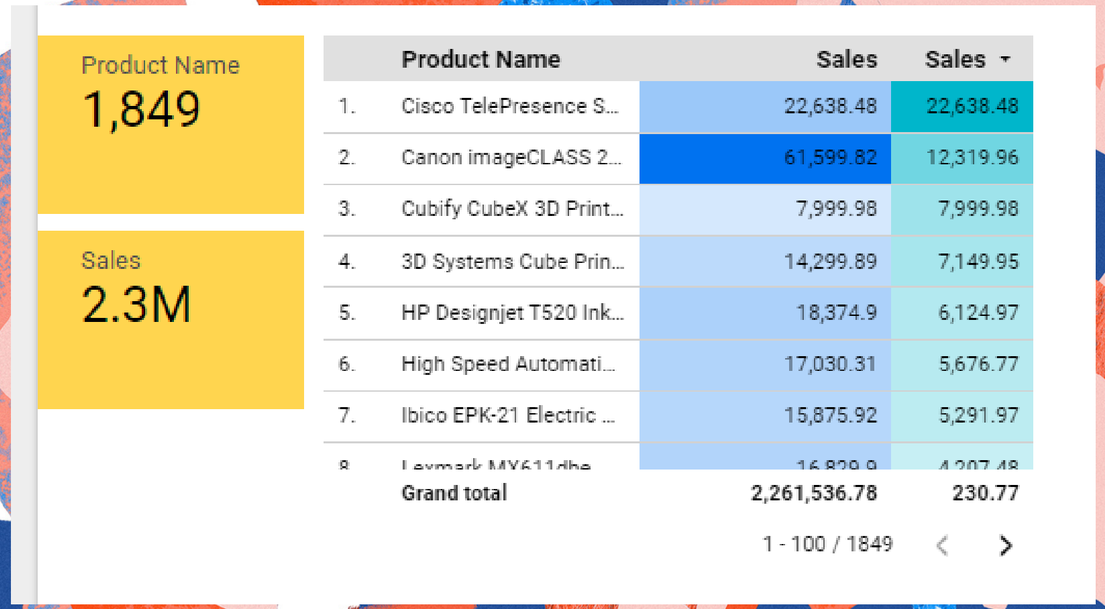
Insight of category-sales & sub-category -sales





Bivariate Analysis

Insight of Product-sales

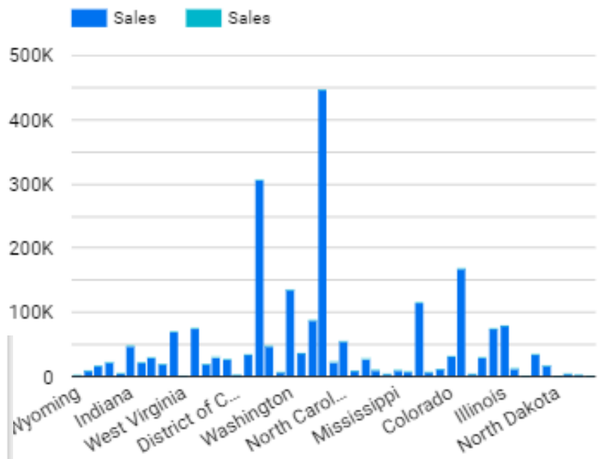


Bivariate Analysis

Insight of Location-sales

State
49

Sales
2.3M



City
529

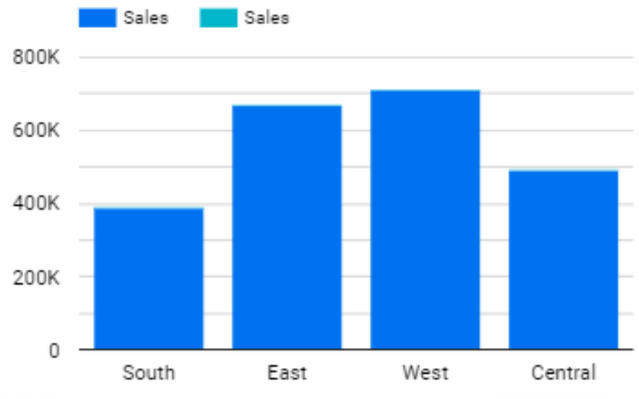
Sales
2.3M

	City	Sales	Sales
1.	Jamestown	4,708.79	2,354.4
2.	Cheyenne	1,603.14	1,603.14
3.	Bellingham	3,790.24	1,263.41
4.	Independence	2,417.37	1,208.69
5.	Burbank	3,247.16	1,082.39
6.	Buffalo	9,063.5	906.35
7.	Burlington	21,668.08	866.72
8.	Beverly	2,583.19	861.06
	Grand total	**2,261,536.78**	**230.77**

1 - 100 / 529

Region
4

Sales
2.3M

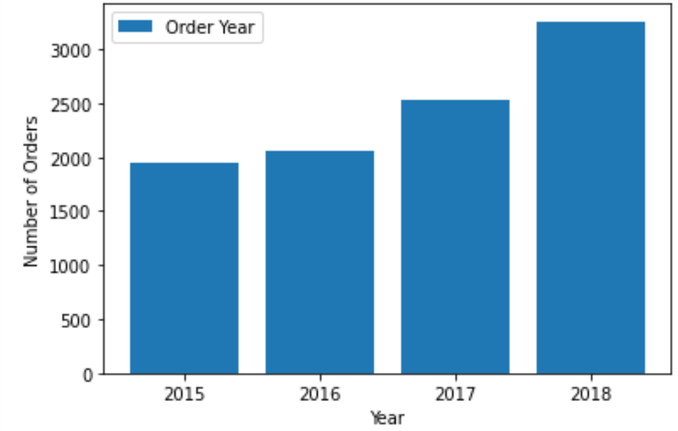


Exploratory Data Analysis

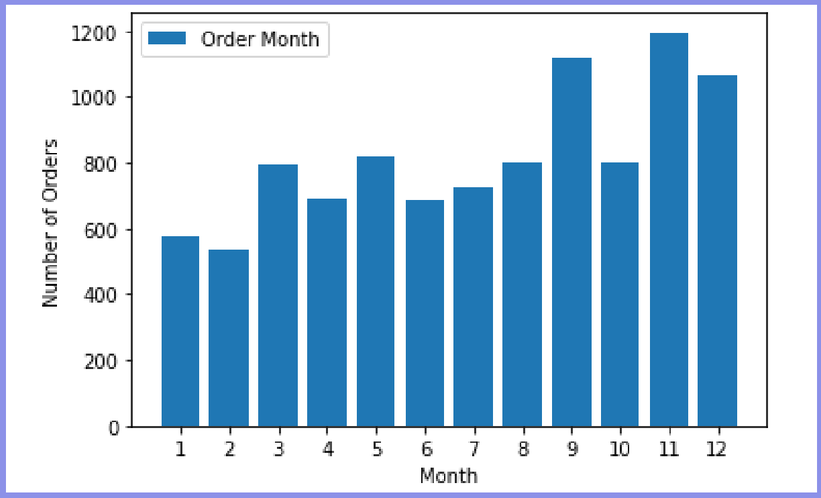
Time Series Analysis

Depends on Order Per Year and Month

	index	order_year
0	2018	3258
1	2017	2534
2	2016	2055
3	2015	1953



	index	order_month
0	11	1194
1	9	1116
2	12	1065
3	5	818
4	10	801
5	8	799
6	3	797
7	7	724
8	4	691
9	6	684
10	1	575
11	2	536



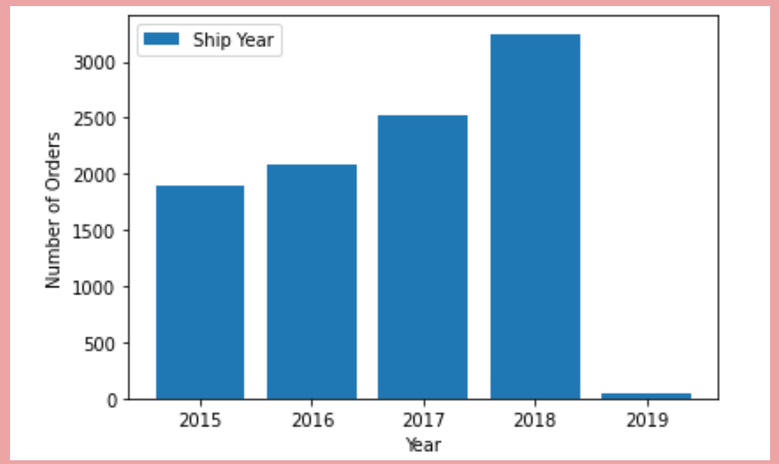
Notes : Secara Graphic, ada kenaikan secara tahun in case of order, tapi apabila kita trajectory per month bisa dilihat pada high season order meningkat secara drastis terutama pada bulan September, November dan Desember

Exploratory Data Analysis

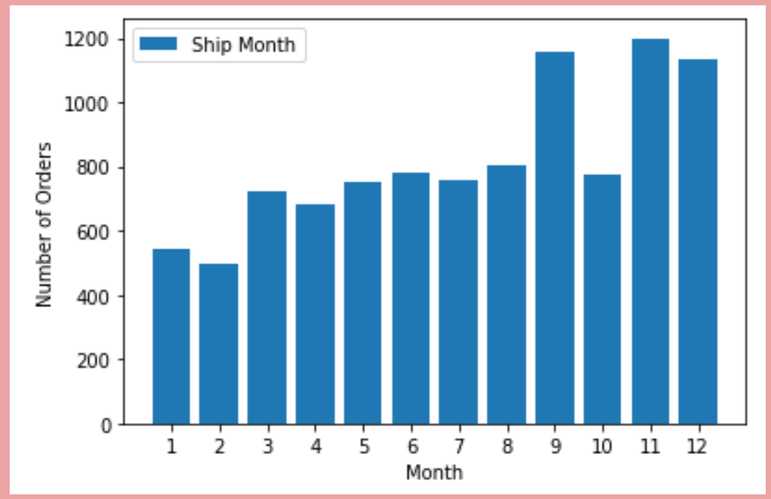
Time Series Analysis

Depends on Ship Per Year and Month

	index	ship_year
0	2018	3249
1	2017	2524
2	2016	2083
3	2015	1902
4	2019	42



	index	ship_month
0	11	1199
1	9	1159
2	12	1131
3	8	801
4	6	780
5	10	775
6	7	757
7	5	754
8	3	721
9	4	684
10	1	545
11	2	494



Keytake : Secara Graphic, ada kenaikan secara tahun in case of shipping kecuali pada tahun 2019 bisa jadi data yang diambil kurang lengkap,
Tapi apabila kita trajectory per month bisa dilihat pada high season order meningkat secara drastis terutama pada bulan September, November dan Desember

Exploratory Data Analysis

Transaction in Contrary

Sales counts for different sub-categories of products for each year

Dalam hal ini, dengan menganalisis hitungan untuk tahun tertentu dalam hal ini kami ambil tahun 2018, kami mengidentifikasi jumlah transaksi tertinggi dan terendah dimana Penyalin adalah yang terendah. Kertas adalah yang tertinggi.

```
[31] # Extract rows where ship_year is 2018
     df_2018 = symcount[symcount['ship_year'] == 2018]

     # Extract sub-category with the lowest count
     sub_category_lowest_count = df_2018.loc[df_2018['counts'].idxmin()]['Sub-Category']

     print(sub_category_lowest_count)
```

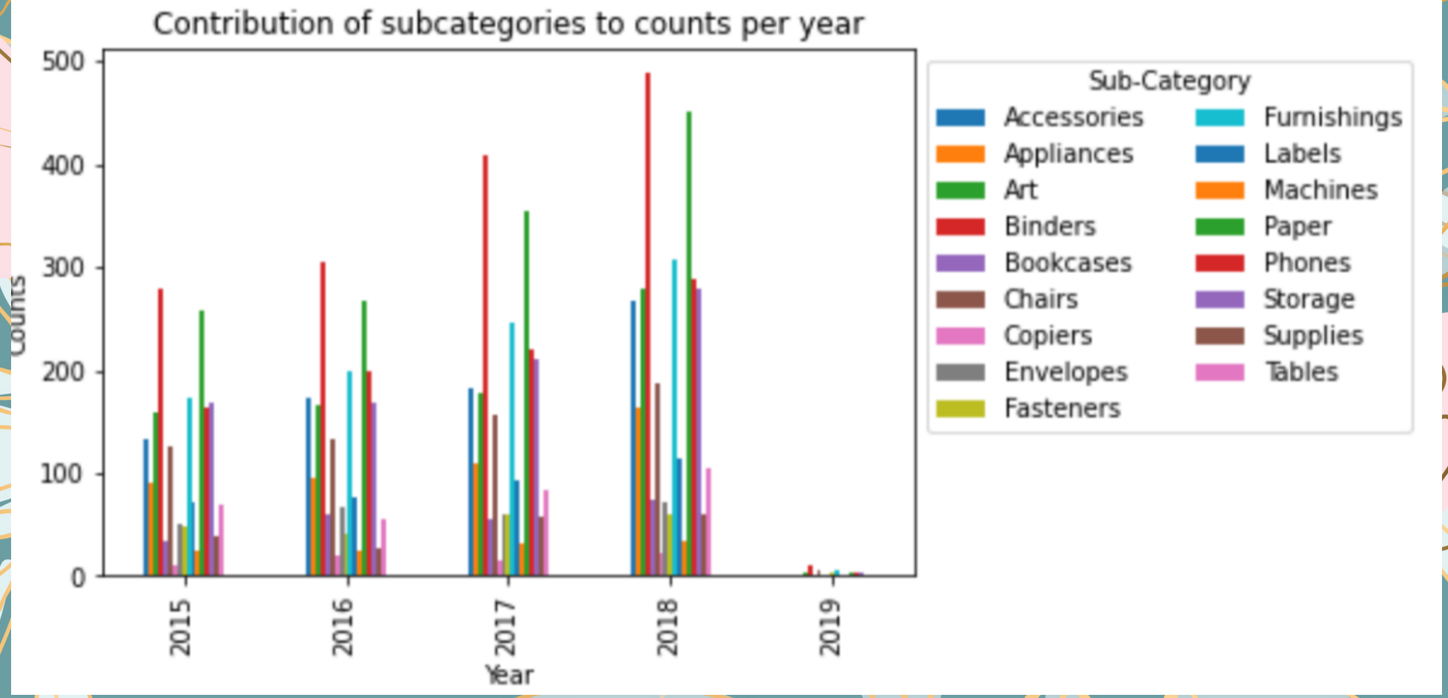
```
Copiers
```

```
[34] # Extract rows where ship_year is 2018
     df_2018 = symcount[symcount['ship_year'] == 2018]

     # Extract sub-category with the highest count
     sub_category_highest_count = df_2018.loc[df_2018['counts'].idxmax()]['Sub-Category']

     print(sub_category_highest_count)
```

```
Paper
```



Exploratory Data Analysis

The highest Contribution

We want to know who is most likely that have the largest contribution to sales in a consecutive year

Untuk Teknologi di AS, Kota New York memiliki jumlah penjualan yang besar dengan $41361.558

```
[ ]  # To determine the country and city that have the largest contribution to sales in a specific year you can
     sales_by_year_country_city =df_fe.groupby(['ship_year','Category','Country','City']).agg({'Sales': 'sum'}).
     sales_by_year_country_city.head()
```

				Sales
ship_year	Category	Country	City	
2018	Technology	United States	New York City	41361.558
2016	Technology	United States	New York City	32217.626
2018	Technology	United States	Seattle	27221.642
2017	Technology	United States	Los Angeles	25833.530
2015	Technology	United States	Jacksonville	23803.548

```
[ ]  sales_by_year_country_city.tail()
```

				Sales
ship_year	Category	Country	City	
2018	Office Supplies	United States	Elyria	1.824
			Brownsville	1.744
2019	Office Supplies	United States	Peoria	1.680
2018	Technology	United States	Memphis	1.584
	Office Supplies	United States	Abilene	1.392

Outlier & Missing Data

We also check the Outlier since there's no missing data in Sales Column.
Outlier:

Quantile of Sales

- Q1 (25%)= 17.248
- Q2 (50%)= 54.590
- Q3 (75%)= 210.605

IQR / Interquartile Range

- Q3-Q1 = 193.357

Lower Bound & Upper Bound
lower bound sales = q1_sales - (1.5 * iqr_Sales) = -272.7875
upper bound_sales = q3_sales + (1.5 * iqr_Sales) = 500.640

Turns out 1145 data found in the Outlier