رواد مصر الرقمية

Analyzing Sales Performance

Team Members: Salma Gamal – Noha Ahmed – Farah Amr – Aya Hesham – Asmaa Khaled Mohamed Ashraf –

Led by : ENG : Aya El Ghaly

INTRO

Introduction

We are working with a dataset that consists of three different sources: Customer, Product, and Order data.

Our goal is to clean this data and prepare it for analysis by identifying and handling missing values, removing duplicates, and ensuring proper data types for effective analysis.

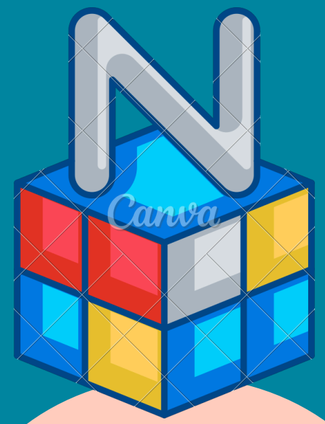
Key Steps

- Cleaning Process
- Analysis Process
- Tableau
- Conclusion

First Step
Cleaning Process

Cleaning

Importing Libraries



```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

PANDAS

For data manipulation and analysis

NUMPY

For numerical operations and array management

MATPLOTLIB .PYPLOT

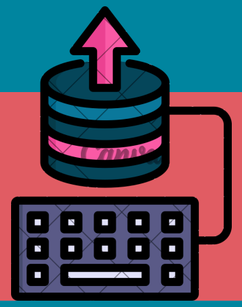
For visualizations

SEABORN

For statistical visualizations

Cleaning

Loading the Data



- To load the dataset, I used the pandas library's read_csv() function to import CSV files into dataframes.
- This allows me to easily work with the data and inspect its structure.
- Each CSV file contains data related to customers, orders, and products, and they were read using the appropriate file paths and encoding.

Input

```
# Load the datasets
df1 = pd.read_csv('/kaggle/input/bussiness1/Customer-Raw Data - Business Data Set (Copy-csv)2 - Copy.csv', encoding='ISO-8859-1')
df2 = pd.read_csv('/kaggle/input/bussiness1/Del Col- Raw Data - Business Data Set.csv', encoding='ISO-8859-1')
df3 = pd.read_csv('/kaggle/input/bussiness1/Product- Raw Data - Business Data Set (Copy-csv) - Copy.csv', encoding='ISO-8859-1')
```

```
# Display the first 10 rows
print(df2.head(10))
```

Output

Cleaning

Check for Missing Data

- We use the isnull().sum() function to check for missing values in each column.
- The dataset may have missing values in key columns like "Sales", "Quantity", and "Profit."

```
df1.isnull().sum()
```

```
df3.isnull().sum()
```

```
df2.isnull().sum()
```

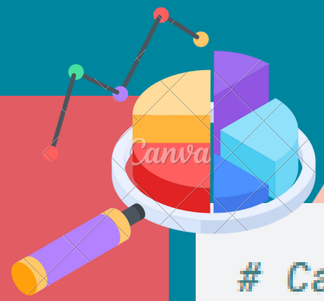
```
In [9]: df1.isnull().sum()

Out[9]: Customer ID      0
        Customer Name    0
        dtype: int64
```

```
In [10]: df3.isnull().sum()

Out[10]: Product ID      0
         Product Name    0
         dtype: int64
```

```
In [11]: df2.isnull().sum()

Out[11]: Order ID          0
         Order Date        0
         Ship Mode      1284
         Customer ID       0
         Product ID        0
         country           0
         city              0
         State             0
         Postal Code       0
         Region            0
         Category          0
         Sub-Category      0
         Sales          1596
         Quantity        558
         Discount          0
         profit         1921
         dtype: int64
```

Cleaning

Handling Missing Data



- We calculate the percentage of missing data in each column using a simple formula.
- Based on the percentage, we decide whether to drop the column or clean it (fill/drop rows).
- If missing values exceed a threshold (e.g., 60%), we drop the column.

```
# Calculating percentage of missing values
null_percentage = (df2.isnull().sum() / len(df2)) * 100
print(null_percentage)
```

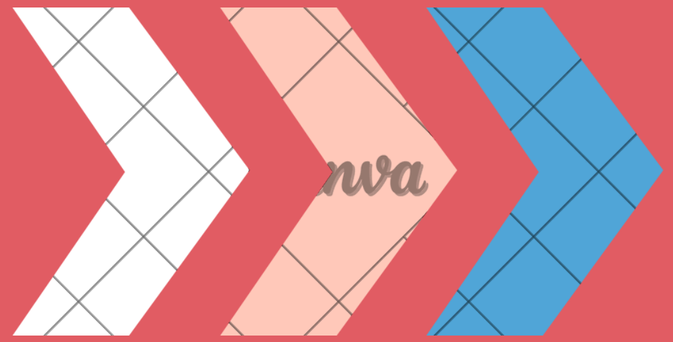
```
df_cleaned = df2.loc[:, null_percentage < 60]
```

```
# Dropping rows with missing values
df2_cleaned = df2.dropna()
```

```
# Checking if missing values are removed
print(df2_cleaned)
```

Cleaning

Output

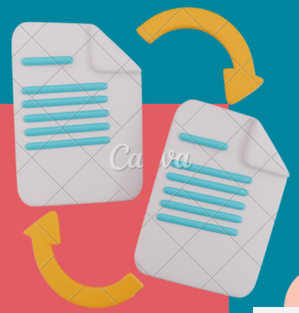


```
]: null_percentage = (df2.isnull().sum() / len(df2)) * 100
   print(null_percentage)
```

```
Order ID        0.000000
Order Date      0.000000
Ship Mode      12.847709
Customer ID     0.000000
Product ID      0.000000
country         0.000000
city            0.000000
State           0.000000
Postal Code     0.000000
Region          0.000000
Category        0.000000
Sub-Category    0.000000
Sales          15.969582
Quantity        5.583350
Discount        0.000000
profit         19.221533
dtype: float64
```

```
          Order ID       Order Date       Ship Mode Customer ID  \
2     CA-2016-138688   6/12/2016 0:00    Second Class    DV-13045
4     US-2015-108966  10/11/2015 0:00  Standard Class    SO-20335
7     CA-2014-115812   6/9/2014 0:00   Standard Class    BH-11710
14    US-2015-118983  11/22/2015 0:00  Standard Class    HP-14815
21    CA-2016-137330   12/9/2016 0:00  Standard Class    KB-16585
...              ...              ...             ...         ...
9989  CA-2014-110422   1/21/2014 0:00    Second Class    TB-21400
9990  CA-2017-121258   2/26/2017 0:00  Standard Class    DB-13060
9991  CA-2017-121258   2/26/2017 0:00  Standard Class    DB-13060
9992  CA-2017-121258   2/26/2017 0:00  Standard Class    DB-13060
9993  CA-2017-119914    5/4/2017 0:00    Second Class    CC-12220

          Product ID        country             city       State  \
2     OFF-LA-10000240  United States      Los Angeles  California
4     OFF-ST-10000760  United States  Fort Lauderdale     Florida
7     TEC-PH-10002275  United States      Los Angeles  California
...               ...            ...              ...         ...
21    OFF-AR-10000246  United States          Fremont    Nebraska
...               ...            ...              ...         ...
9989  FUR-FU-10001889  United States            Miami     Florida
9990  FUR-FU-10000747  United States        Costa Mesa  California
9991  TEC-PH-10003645  United States        Costa Mesa  California
9992  OFF-PA-10004041  United States        Costa Mesa  California
9993  OFF-AP-10002684  United States       Westminster  California

      Postal Code   Region         Category Sub-Category    Sales  Quantity  \
2           90036     West  Office Supplies       Labels   14.620       2.0
4           33311    South  Office Supplies      Storage   22.368       2.0
7           90032     West       Technology       Phones  907.152       6.0
14          76106  Central  Office Supplies   Appliances   68.810       5.0
21          68025  Central  Office Supplies          Art   19.460       7.0
...           ...      ...              ...          ...      ...       ...
9989        33180    South        Furniture  Furnishings   25.248       3.0
9990        92627     West        Furniture  Furnishings   91.960       2.0
9991        92627     West       Technology       Phones  258.576       2.0
9992        92627     West  Office Supplies        Paper   29.600       4.0
9993        92683     West  Office Supplies   Appliances  243.160       2.0

      Discount    profit
2          0.0    6.8714
4          0.2    2.5164
7          0.2   90.7152
14         0.8 -123.8580
21         0.0    5.0596
...        ...       ...
9989       0.2    4.1028
9990       0.0   15.6332
9991       0.2   19.3932
9992       0.0   13.3200
9993       0.0   72.9480
```

Cleaning



Checking for Duplicates

- Duplicates are checked and removed from the dataset to avoid redundant data affecting the analysis.

```
# Count the number of duplicate entries in the DataFrame df2
df2.duplicated().sum()
```

```
In [16]: df2.duplicated().sum()

Out[16]: 0
```

```
# Remove duplicate entries from the DataFrame df2_cleaned
df2_cleaned = df2_cleaned.drop_duplicates()
```

Cleaning

Data Formatting & Preprocessing

- Convert columns to appropriate data types (e.g., date format).
- Strip any whitespace from column names.

Input & Output

```
# Remove leading and trailing whitespace from column names
df2_cleaned.columns = df2_cleaned.columns.str.strip()
```

```
In [17]: df2_cleaned = df2_cleaned.drop_duplicates()

In [18]: df2_cleaned['Order Date'] = pd.to_datetime(df2_cleaned['Order Date'])

In [19]: print(df2_cleaned.head())
```

```
         Order ID Order Date       Ship Mode Customer ID       Product ID  \
2   CA-2016-138688 2016-06-12    Second Class    DV-13045  OFF-LA-10000240
4   US-2015-108966 2015-10-11  Standard Class    SO-20335  OFF-ST-10000760
7   CA-2014-115812 2014-06-09  Standard Class    BH-11710  TEC-PH-10002275
14  US-2015-118983 2015-11-22  Standard Class    HP-14815  OFF-AP-10002311
21  CA-2016-137330 2016-12-09  Standard Class    KB-16585  OFF-AR-10000246

          country             city       State  Postal Code   Region  \
2   United States      Los Angeles  California        90036     West
4   United States  Fort Lauderdale     Florida        33311    South
7   United States      Los Angeles  California        90032     West
14  United States       Fort Worth       Texas        76106  Central
21  United States          Fremont    Nebraska        68025  Central

            Category Sub-Category    Sales  Quantity  Discount    profit
2   Office Supplies       Labels   14.620       2.0       0.0    6.8714
4   Office Supplies      Storage   22.368       2.0       0.2    2.5164
7        Technology       Phones  907.152       6.0       0.2   90.7152
14  Office Supplies   Appliances   68.810       5.0       0.8 -123.8580
21  Office Supplies          Art   19.460       7.0       0.0    5.0596
```

Cleaning

Checking for Zeros in Key Columns

- The code checks for zero values in the specified columns and provides a simple report on their presence or absence.
- The code snippet you provided checks for missing (null) values in the DataFrame df2_cleaned and prints the columns that have missing values, along with the count of those missing values.

```
In [22]:
zeros_in_col_sales = (df_cleaned['Sales'] == 0).any()
zeros_in_col_ship_mode = (df_cleaned['Ship Mode'] == 0).any()

if not zeros_in_col_sales and not zeros_in_col_ship_mode:
    print("لا يوجد أصفار في العمودين Sales و Ship Mode.")
else:
    if zeros_in_col_sales:
        print("يوجد أصفار في العمود Sales.")
    if zeros_in_col_ship_mode:
        print("يوجد أصفار في العمود Ship Mode.")

لا يوجد أصفار في العمودين Sales و Ship Mode.

In [23]:
null_counts = df2_cleaned.isnull().sum()
print(null_counts[null_counts > 0])

Series([], dtype: int64)
```

Cleaning

Identifying Outliers

What is IQR?

The IQR represents the middle 50% of your data, which lies between the first quartile (Q1) and the third quartile (Q3). It helps in understanding the spread of the bulk of your data.

Q1 (First Quartile)

The value below which 25% of the data falls.

Q3 (Third Quartile)

The value below which 75% of the data falls.

IQR Calculation

IQR = Q3 - Q1

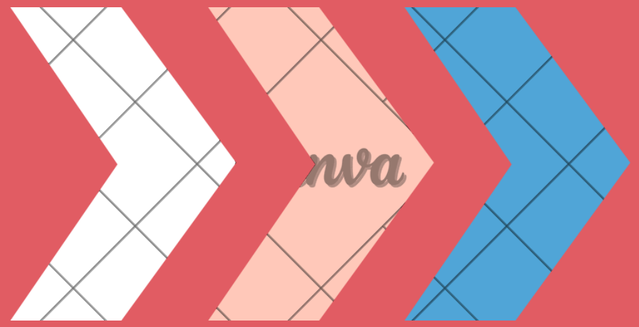
Identifying Outliers

Outliers are data points that fall outside the typical range of your dataset. Using the IQR method, we define outliers as values that are:

1. Less than Q1 - 1.5 * IQR (Lower bound)
2. Greater than Q3 + 1.5 * IQR (Upper bound)

Cleaning

Output



```
Q1 = df2['Sales'].quantile(0.25)
Q3 = df2['Sales'].quantile(0.75)
IQR = Q3 - Q1


lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR


outliers = df2[(df2['Sales'] < lower_bound) | (df2['Sales'] > upper_bound)]
print(outliers)
```

```
            Order ID       Order Date       Ship Mode Customer ID  \
3     US-2015-108966  10/11/2015 0:00  Standard Class    SO-20335
7     CA-2014-115812   6/9/2014 0:00  Standard Class    BH-11710
11    CA-2014-115812   6/9/2014 0:00             NaN    BH-11710
27    US-2015-150630  9/17/2015 0:00  Standard Class    TB-21520
35    CA-2016-117590  12/8/2016 0:00    First Class     GH-14485
...              ...             ...            ...         ...
9931  CA-2015-104948  11/13/2015 0:00  Standard Class    KH-16510
9942  CA-2014-143371  12/28/2014 0:00  Standard Class    MD-17350
9947  CA-2017-121559   6/1/2017 0:00    Second Class    HW-14935
9948  CA-2017-121559   6/1/2017 0:00    Second Class    HW-14935
9968  CA-2017-153871  12/11/2017 0:00  Standard Class    RB-19435

            Product ID        country             city       State  \
3     FUR-TA-10000577  United States  Fort Lauderdale     Florida
7     TEC-PH-10002275  United States      Los Angeles  California
11    TEC-PH-10002033  United States      Los Angeles  California
27    FUR-BO-10004834  United States     Philadelphia  Pennsylvania
35    TEC-PH-10004977  United States        Richardson       Texas
9931  FUR-BO-10004357  United States   San Bernardino  California
9942  OFF-ST-10001128  United States          Anaheim  California
9947  FUR-CH-10003746  United States     Indianapolis     Indiana
9948  OFF-AP-10002945  United States     Indianapolis     Indiana
9968  OFF-BI-10004600  United States       Plainfield  New Jersey

      Postal Code   Region         Category Sub-Category      Sales  Quantity  \
3           33311    South        Furniture        Tables   957.5775       NaN
7           90032     West       Technology        Phones   907.1520       6.0
11          90032     West       Technology        Phones   911.4240       4.0
27          19140     East        Furniture     Bookcases  3083.4300       7.0
35          75080  Central       Technology        Phones  1097.5440       7.0
...           ...      ...              ...           ...        ...       ...
9931        92404     West        Furniture     Bookcases   683.3320       4.0
9942        92804     West  Office Supplies       Storage   998.8200       9.0
9947        46203  Central        Furniture        Chairs  1925.8800       6.0
9948        46203  Central  Office Supplies    Appliances  2405.2000       8.0
9968         7060     East  Office Supplies       Binders   735.9800       2.0

      Discount     profit
3         0.45  -383.0310
7         0.20    90.7152
11        0.20    68.3568
27        0.50 -1665.0522
35        0.20   123.4737
...        ...        ...
9931      0.15   -40.1960
9942      0.00    29.9646
9947      0.00   539.2464
9948      0.00   793.7160
9968      0.00   331.1910

[805 rows x 16 columns]
```

Cleaning

Pivot Table Creation for Sales Summary

In this step, we created a pivot table to summarize sales data by region and category. This allows us to easily analyze the total sales performance of different product categories across various regions.

Code Overview:

We used the pivot_table function in Python to group sales by region and category.

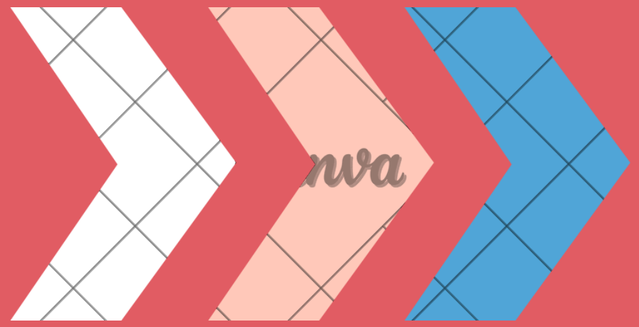
The table aggregates the total sales for each category in each region.

Purpose:

To provide a clear overview of how different regions perform in terms of sales, helping us identify trends and opportunities.

Cleaning

Output



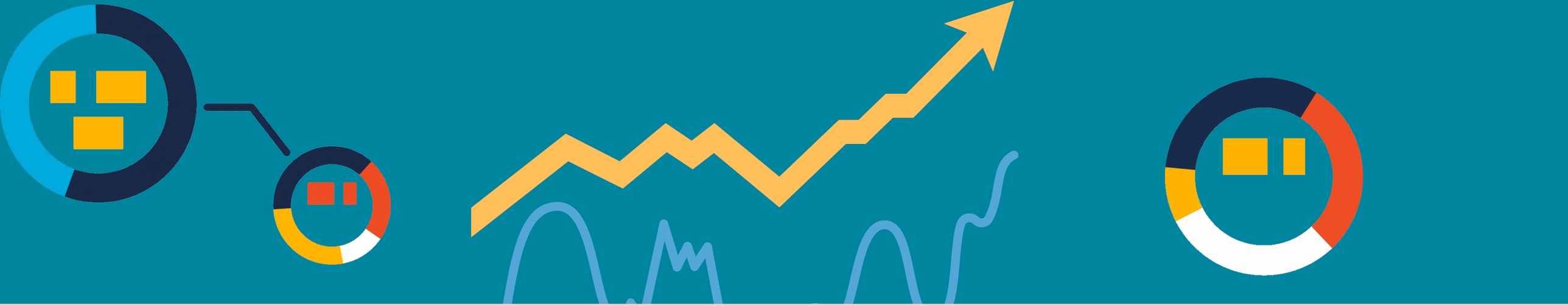
In [29]:

```
pivot_table = df2.pivot_table(
    values='Sales',
    index='Region',
    columns='Category',
    aggfunc='sum',
    fill_value=0
)
```

In [30]:

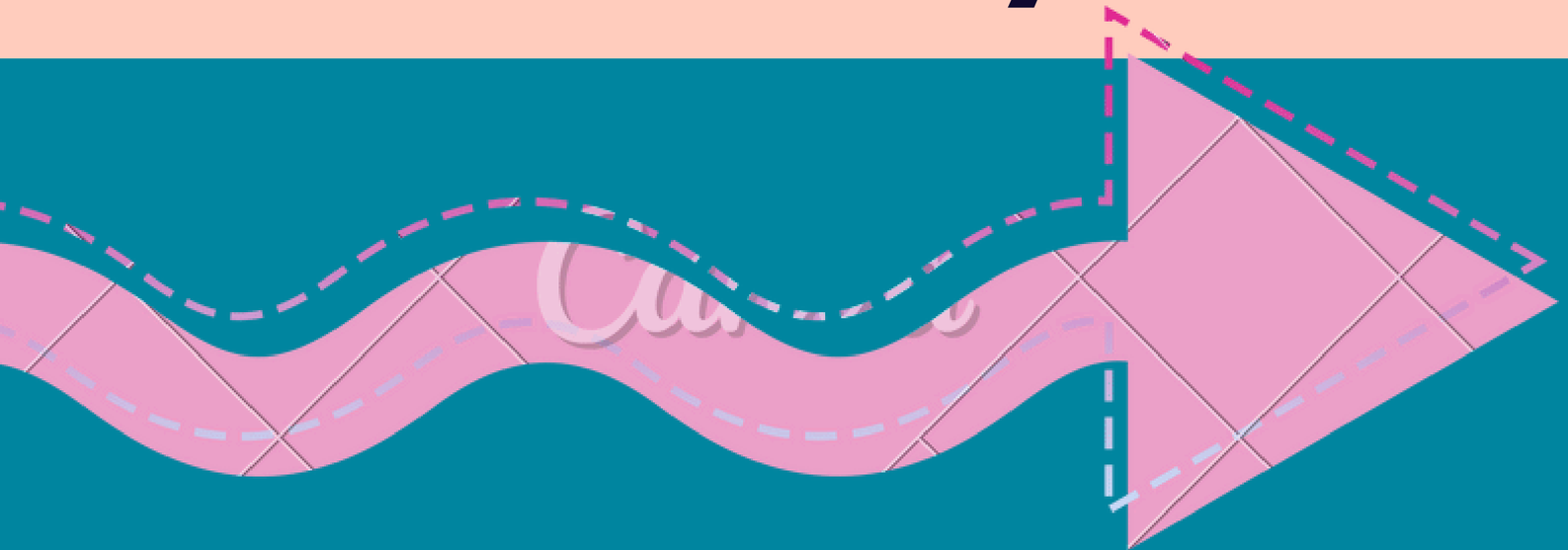
```
print(pivot_table)
```

```
Category     Furniture  Office Supplies  Technology
Region
Central    143028.5072       147617.447  141641.242
East       182563.8010       171074.916  223227.764
South       98250.8980       100359.534  130533.251
West       214402.4985       192492.751  215061.602
```

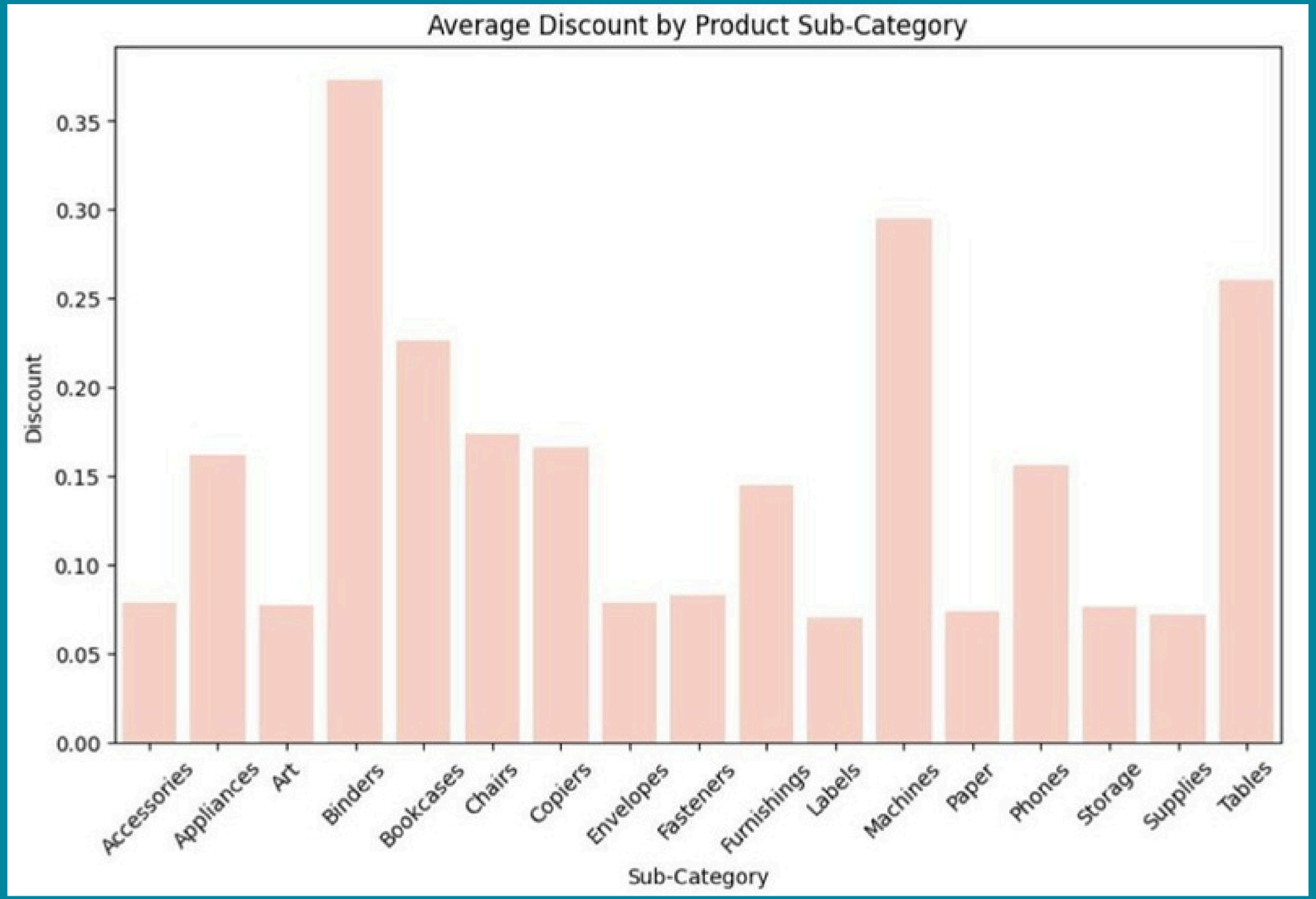


NEXT,

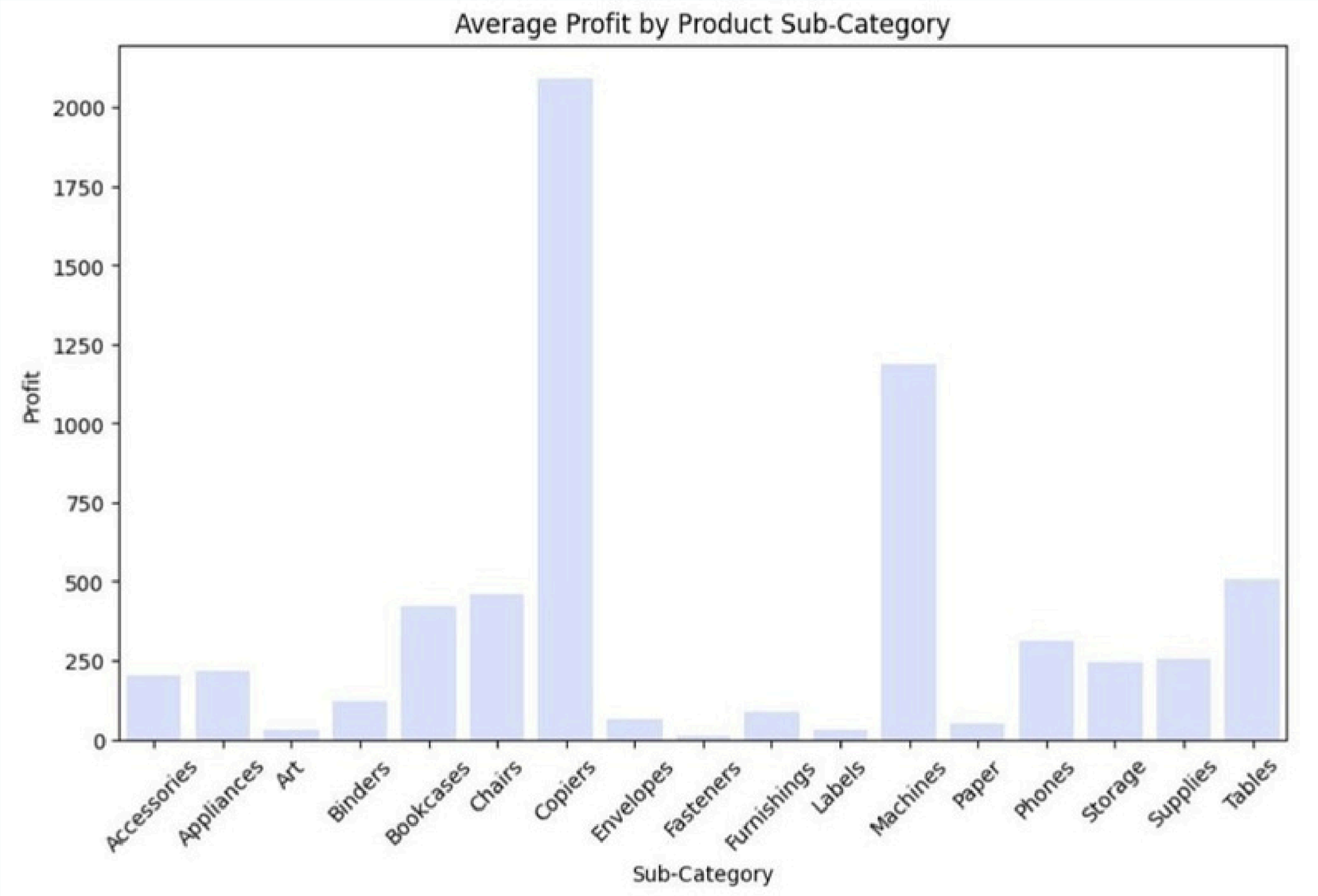
Analysis Process



Average Discount by product Sub-Category



Average Profit by Product Sub-Category



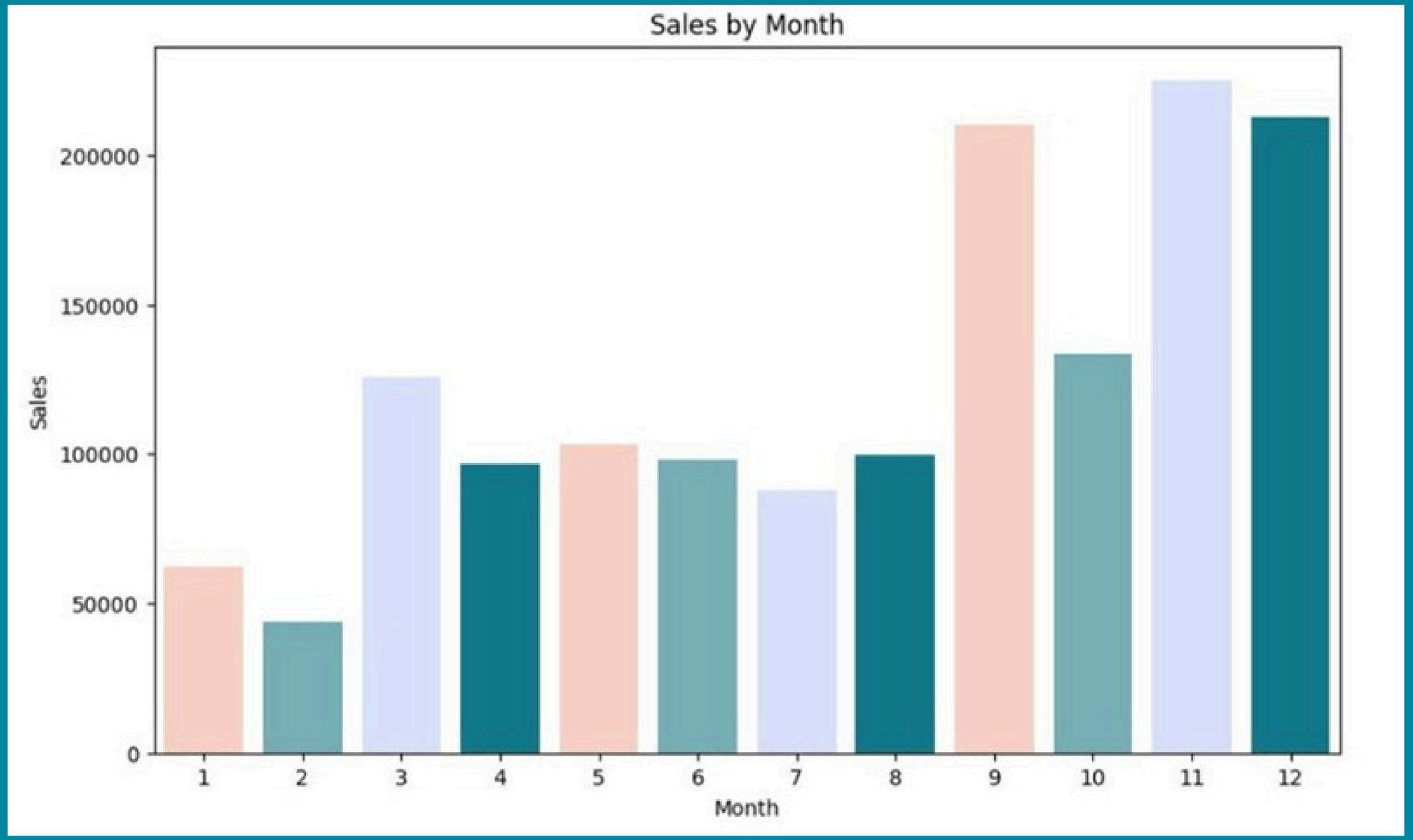
Sales by ship mode, region, category and sum category

	Category	Sub-Category	Sales
0	Furniture	Bookcases	75575.7316
1	Furniture	Chairs	220596.9850
2	Furniture	Furnishings	60520.3760
3	Furniture	Tables	134581.4235
4	Office Supplies	Appliances	71819.0080
5	Office Supplies	Art	16422.4040
6	Office Supplies	Binders	147734.7850
7	Office Supplies	Envelopes	10994.4200
8	Office Supplies	Fasteners	1786.0060
9	Office Supplies	Labels	7527.6700
10	Office Supplies	Paper	48728.9860
11	Office Supplies	Storage	137795.9540
12	Office Supplies	Supplies	34682.6460
13	Technology	Accessories	110385.7580
14	Technology	Copiers	110168.6400
15	Technology	Machines	105446.6030
16	Technology	Phones	204767.9860

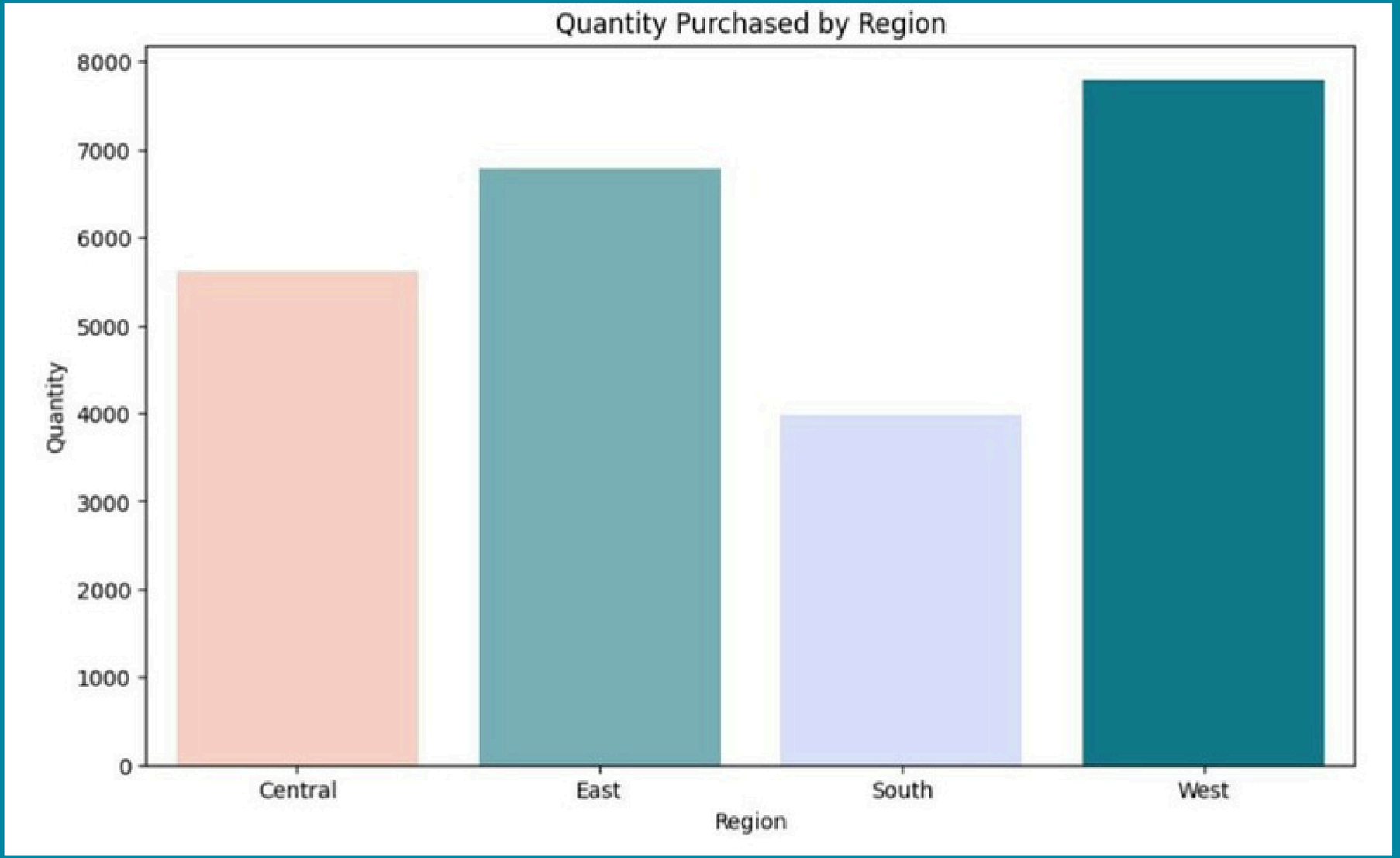
	Region	Sales
0	Central	342225.6916
1	East	463223.3870
2	South	230533.2635
3	West	463553.0400

	Ship Mode	Sales
0	First Class	224116.7439
1	Same Day	76555.2320
2	Second Class	299994.1230
3	Standard Class	898869.2832

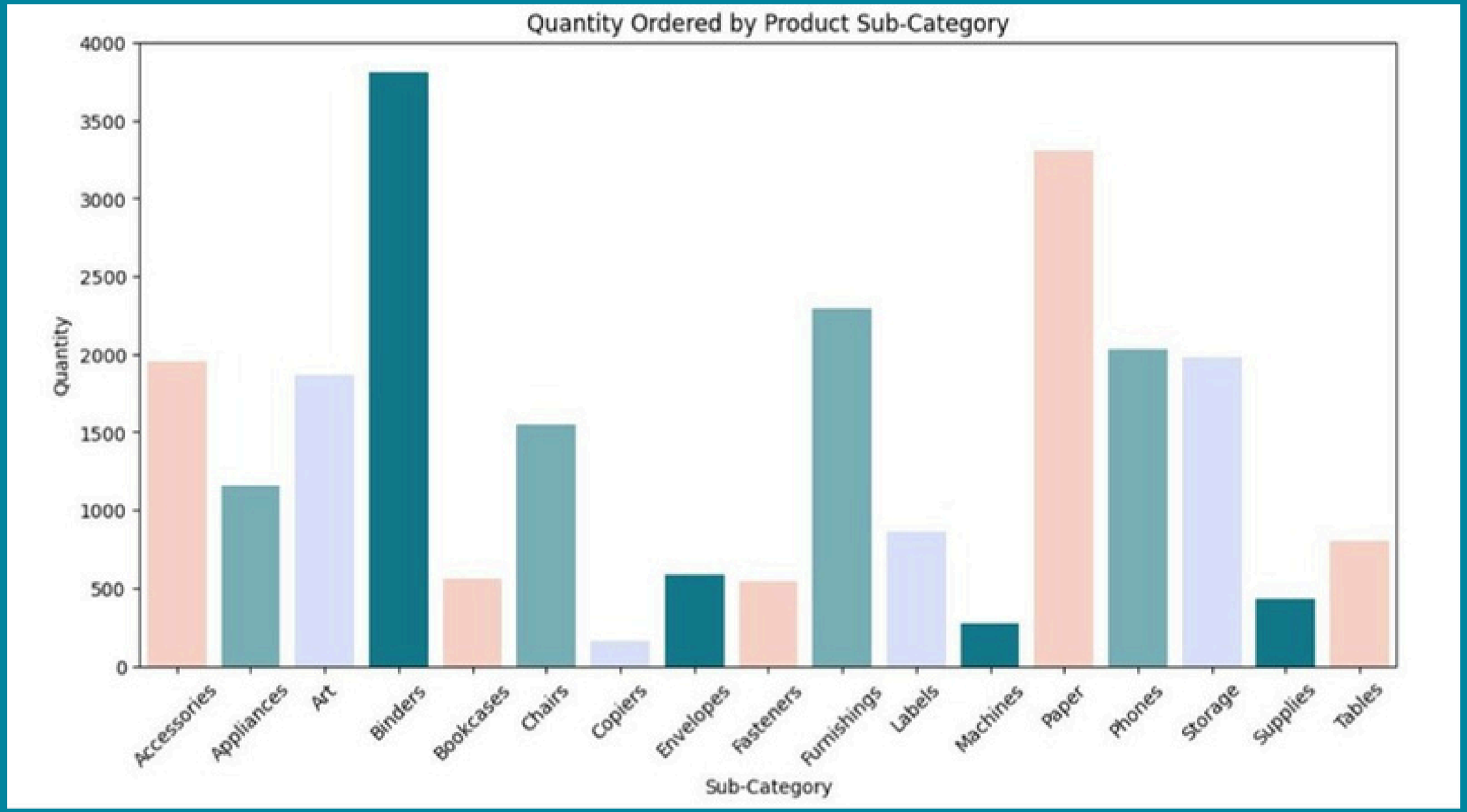
Sales by Month



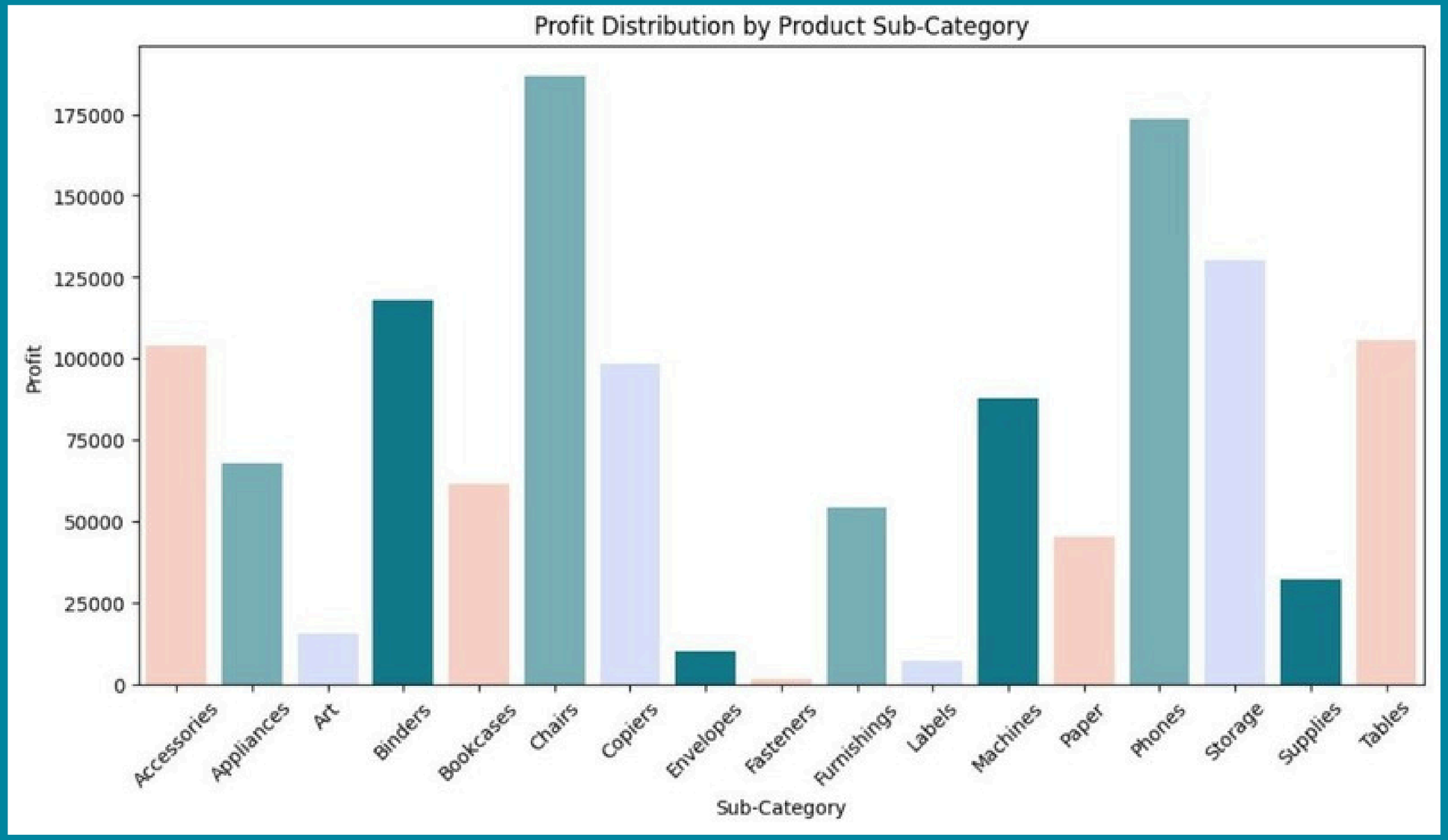
Quantity Purchased by Region



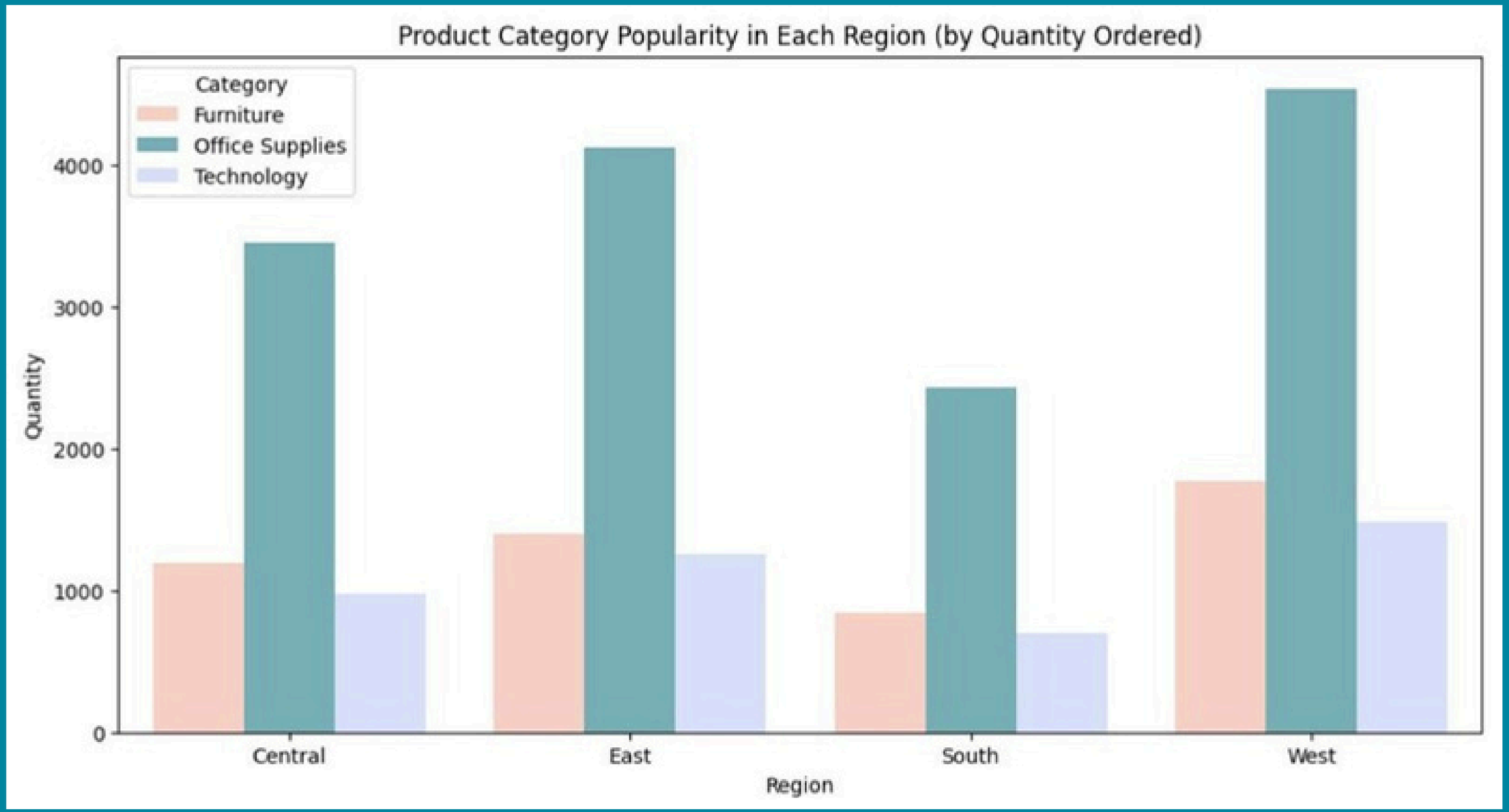
Quantity Ordered by Product Sub-Category



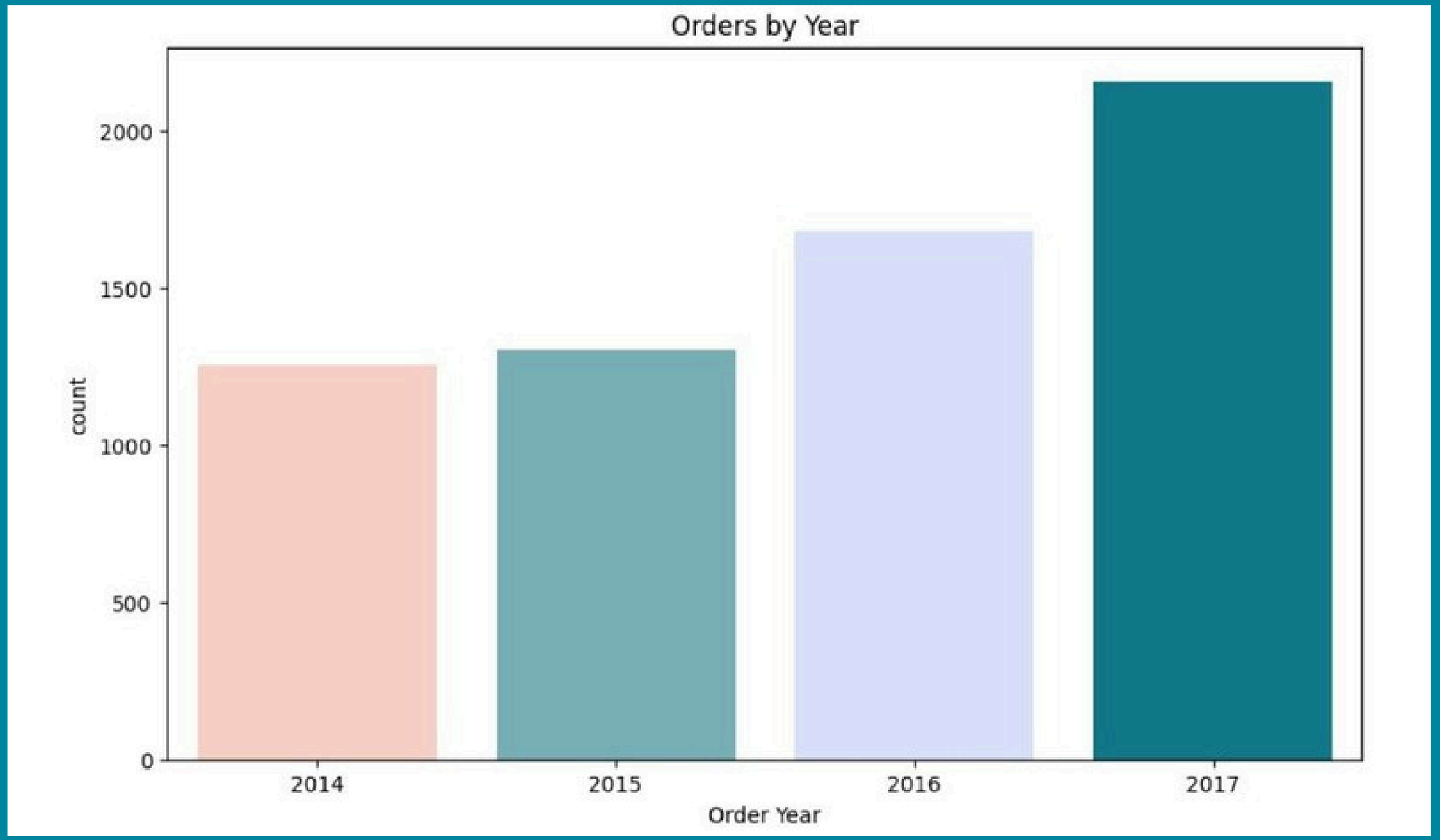
Profit Distribution by Product Sub-Category



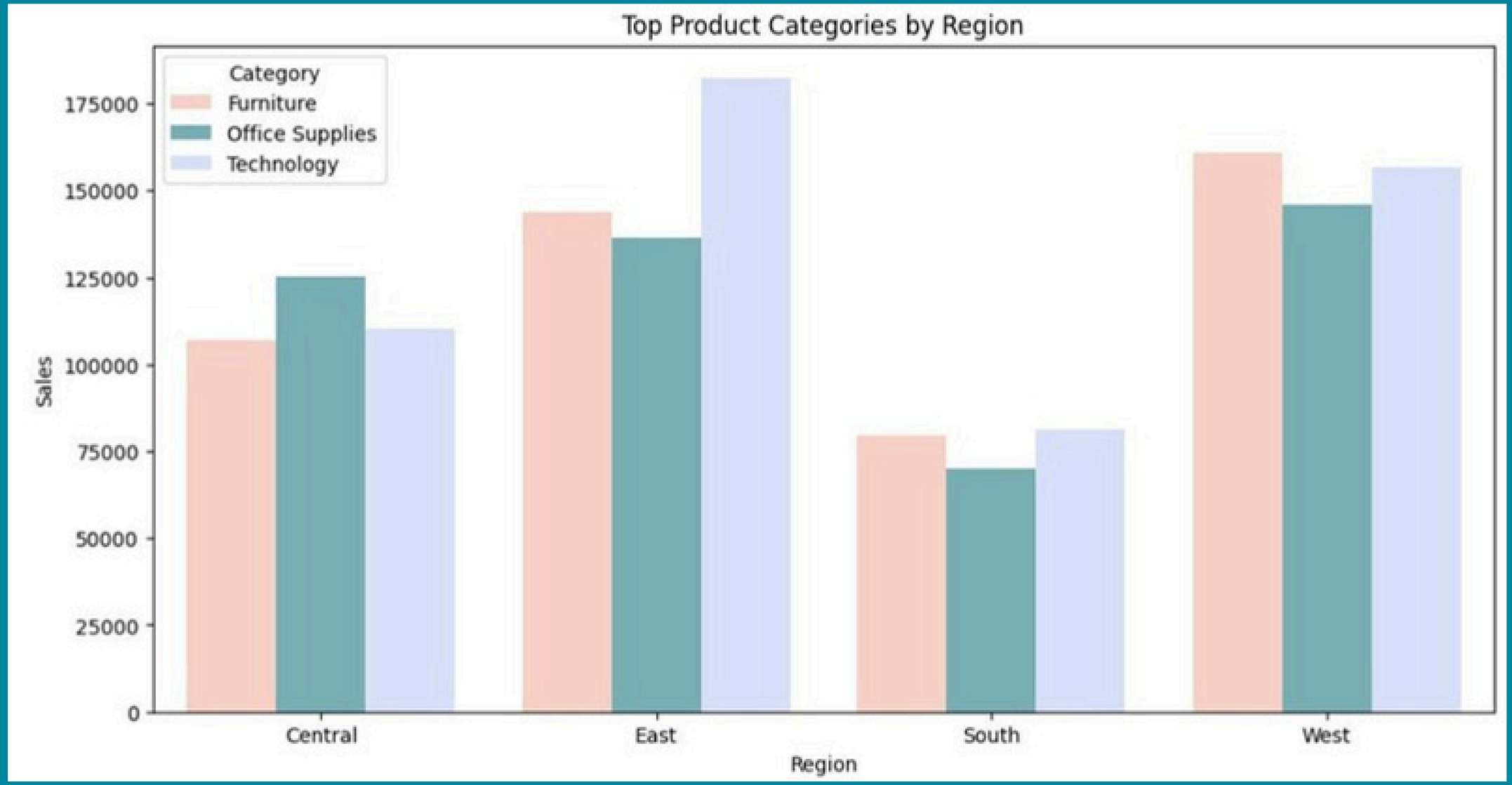
Product Category Popularity in each Region



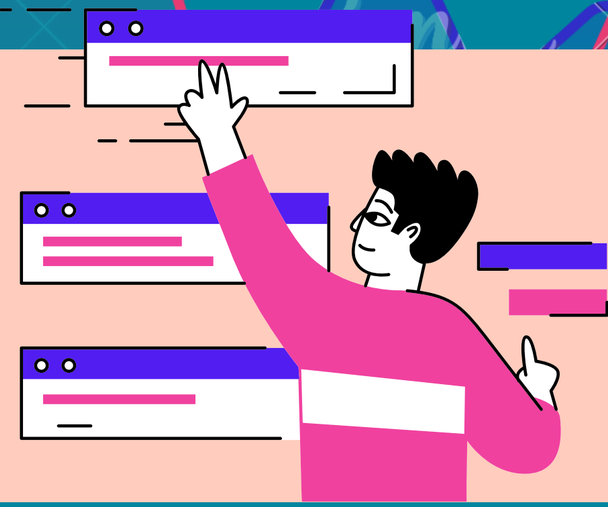
Orders by Year



Top Product Categories by Region



Finally,
Tableau



Tableau

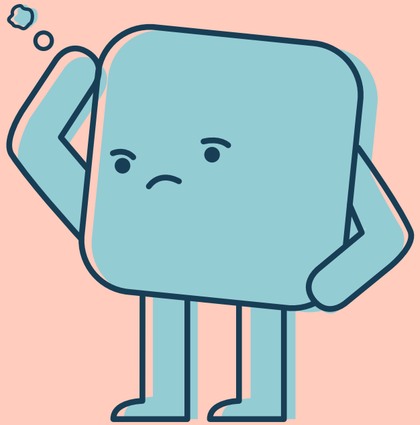
tableau public

INSIGHTS

DASHBOARD

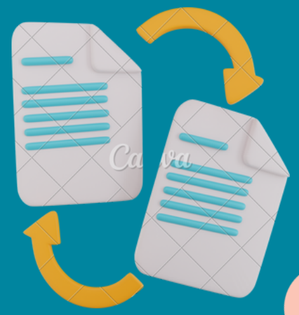
VISUALIZATION
TRENDS

ANALYSIS



Step 1	Step 2	Step 3
Requirements	Connect to data	Build charts
Choose the right charts	Build data model	Create calc fields
Decide on colours	Check data types	Check format

Requirements

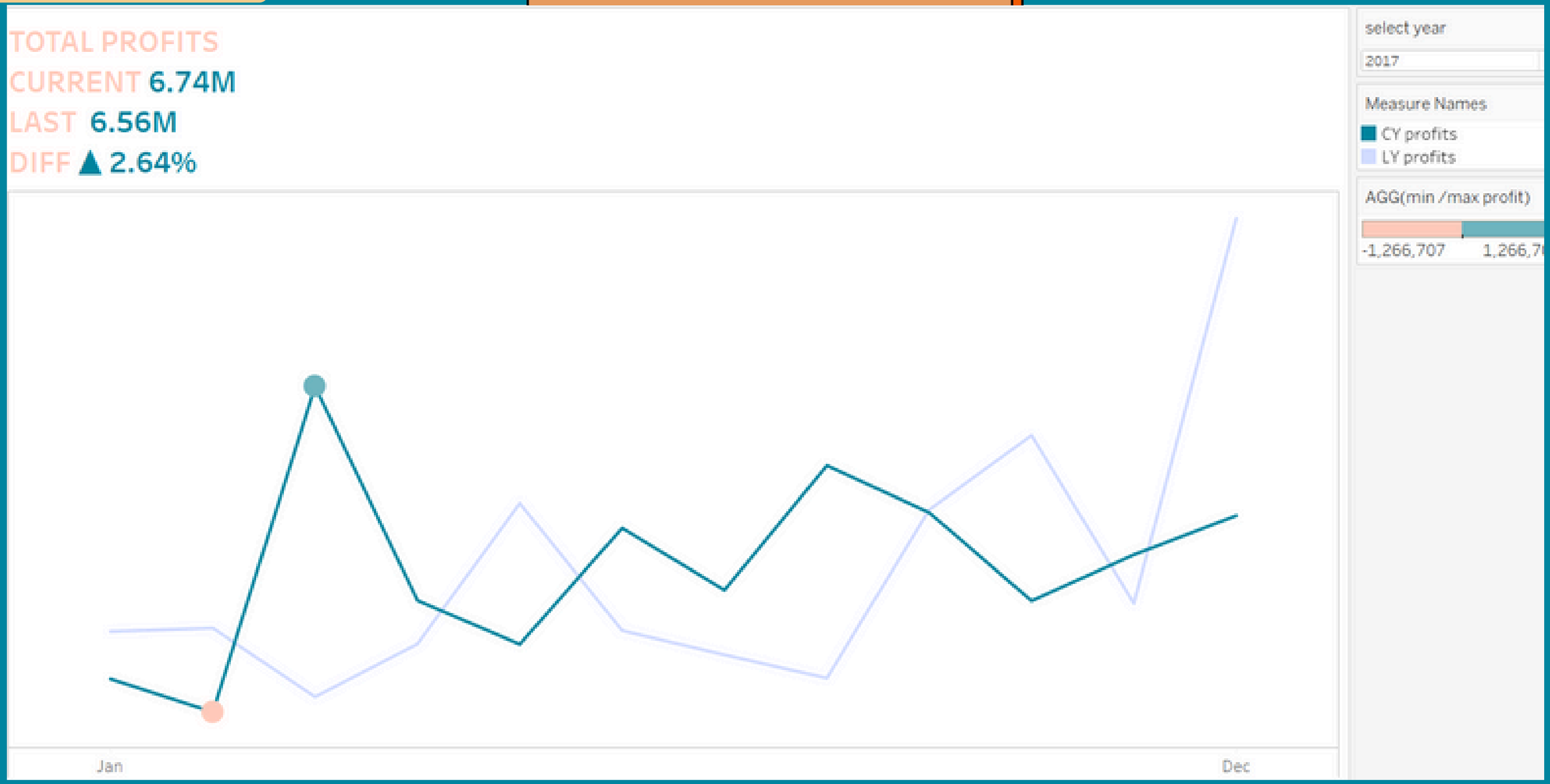


Present the data for each KPI on a monthly basis for both the current year and the previous year, identify months with highest and lowest sales, profits and quantity and make them easy to recognize.

- parameter
- Current and last year
- Max / Min for current year

Tableau

Profit



Tableau

Insights

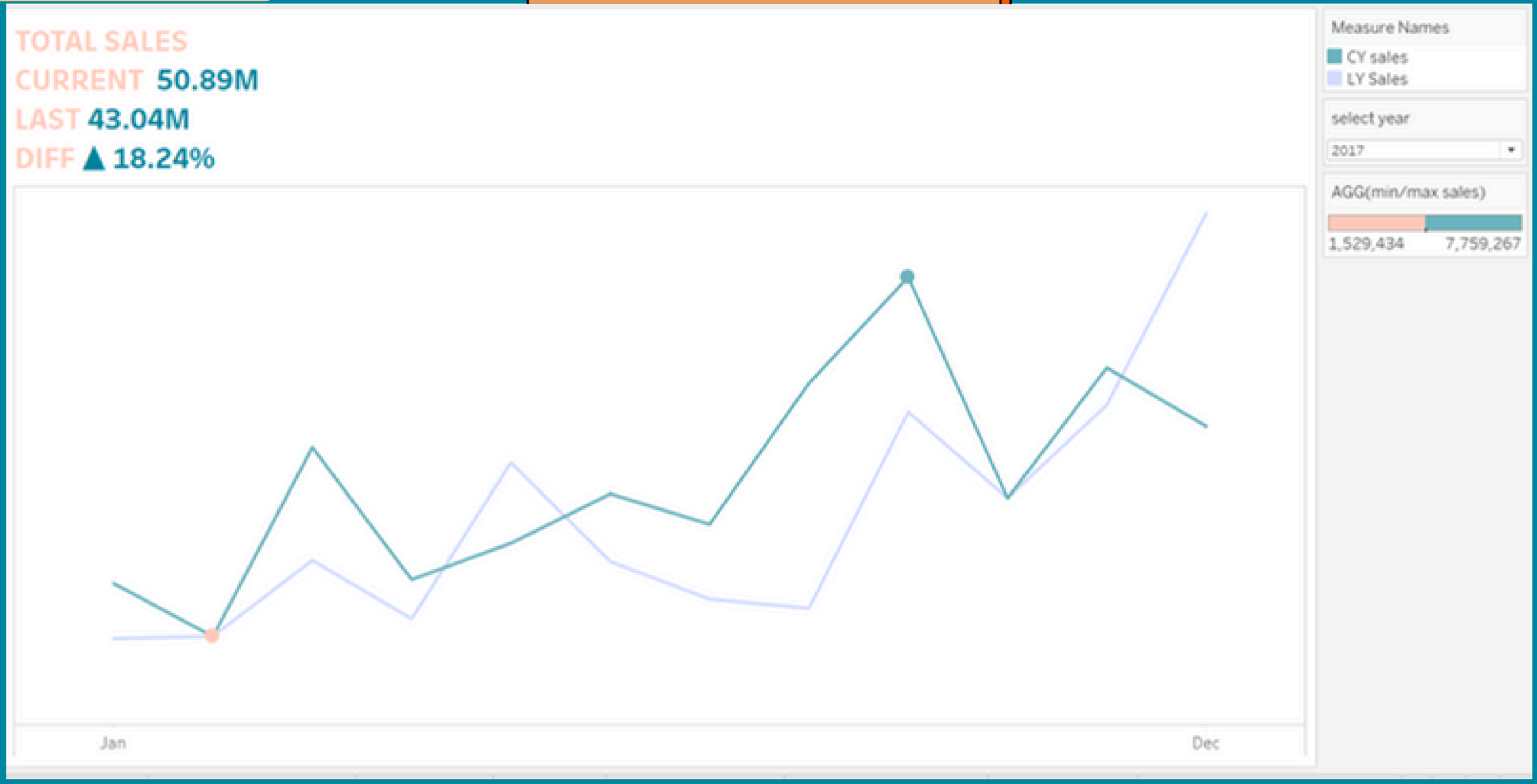
- There is a slight increase in total profits Compared to last year
- The highest profit was in March and the lowest in February. After that, there was a fluctuation in profits until they rose again slightly in December.
- In March, profits were the highest, unlike last year, when profits were the lowest in March.

Recommendations

- Focus on improving stability in the early months
- Capitalize on end-of-year growth
- Analyze the factors causing large fluctuations

Tableau

sales



Tableau

Insights

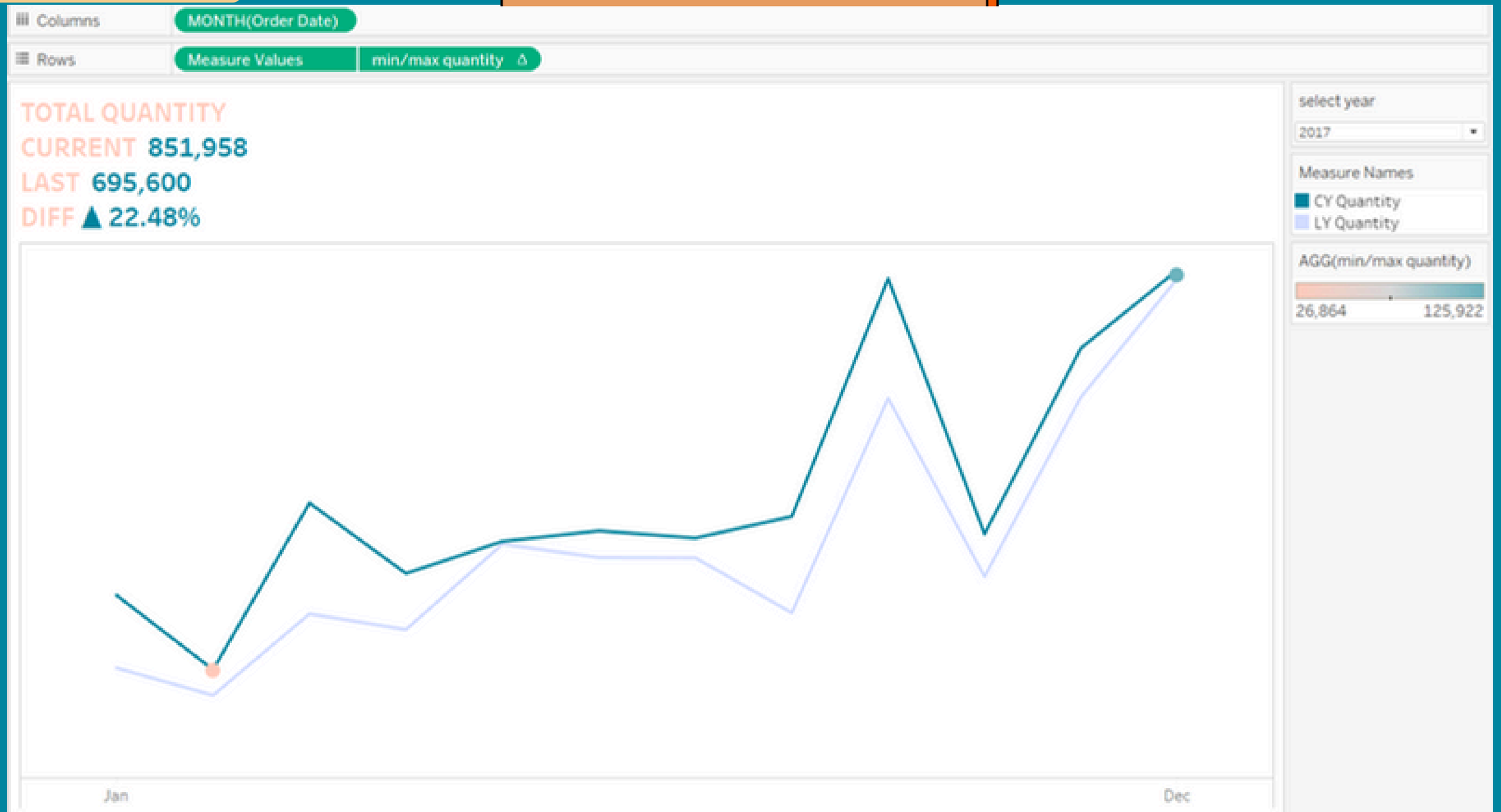
- There is a significant increase in total sales Compared to last year
- Highest sales in September and lowest in February
- Relative stability in other months: In the other months, despite some minor fluctuations, sales remained relatively stable without sharp drops, reflecting overall stability throughout the year.

Recommendations

- The sales strategy can be improved in the last and first months.
- Capitalize on end-of-year growth
- Improve performance in weaker periods and Analyze the factors causing large fluctuations

Tableau

Quantity



Tableau

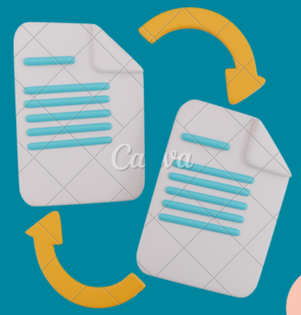
Insights

- There is a slight increase in quantities compared to last year
- The highest quantity was in December and the lowest in February
- There was stability in the middle of the year and the quantity increased at the end of the year by a large percentage

Recommendations

- Focus on maximizing production specially in the first and middle of the year
- Capitalize on end-of-year growth
- Investigate June’s High Demand:

Requirements

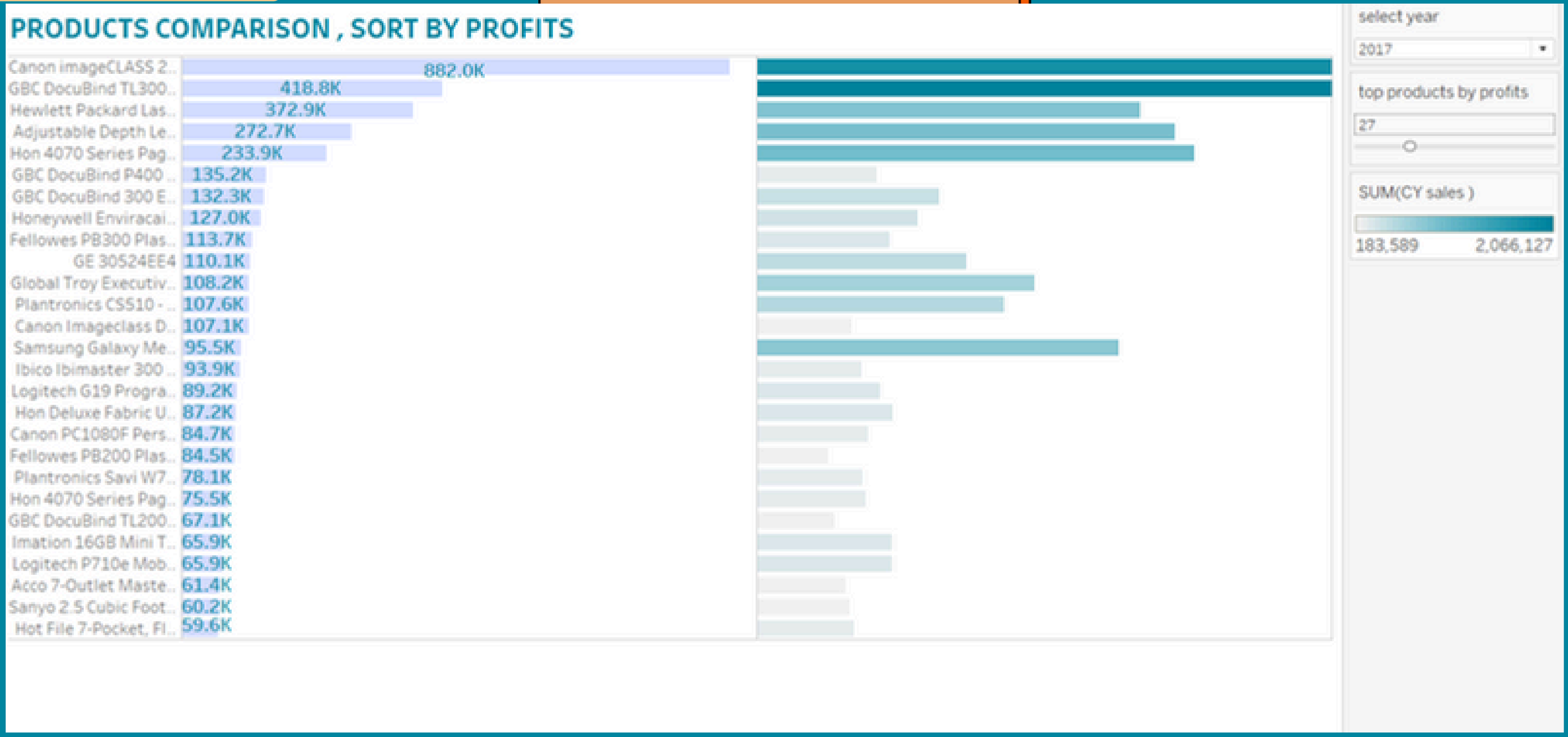


- Comparison between sales and profits of products monthly
- Comparison between sales and profits of sub category monthly
- Monthly trend, perdiction by sales and quantity

- Top sub category parameter
- Filter , forecast line
- Select year parameter

Tableau

products



Tableau

Insights

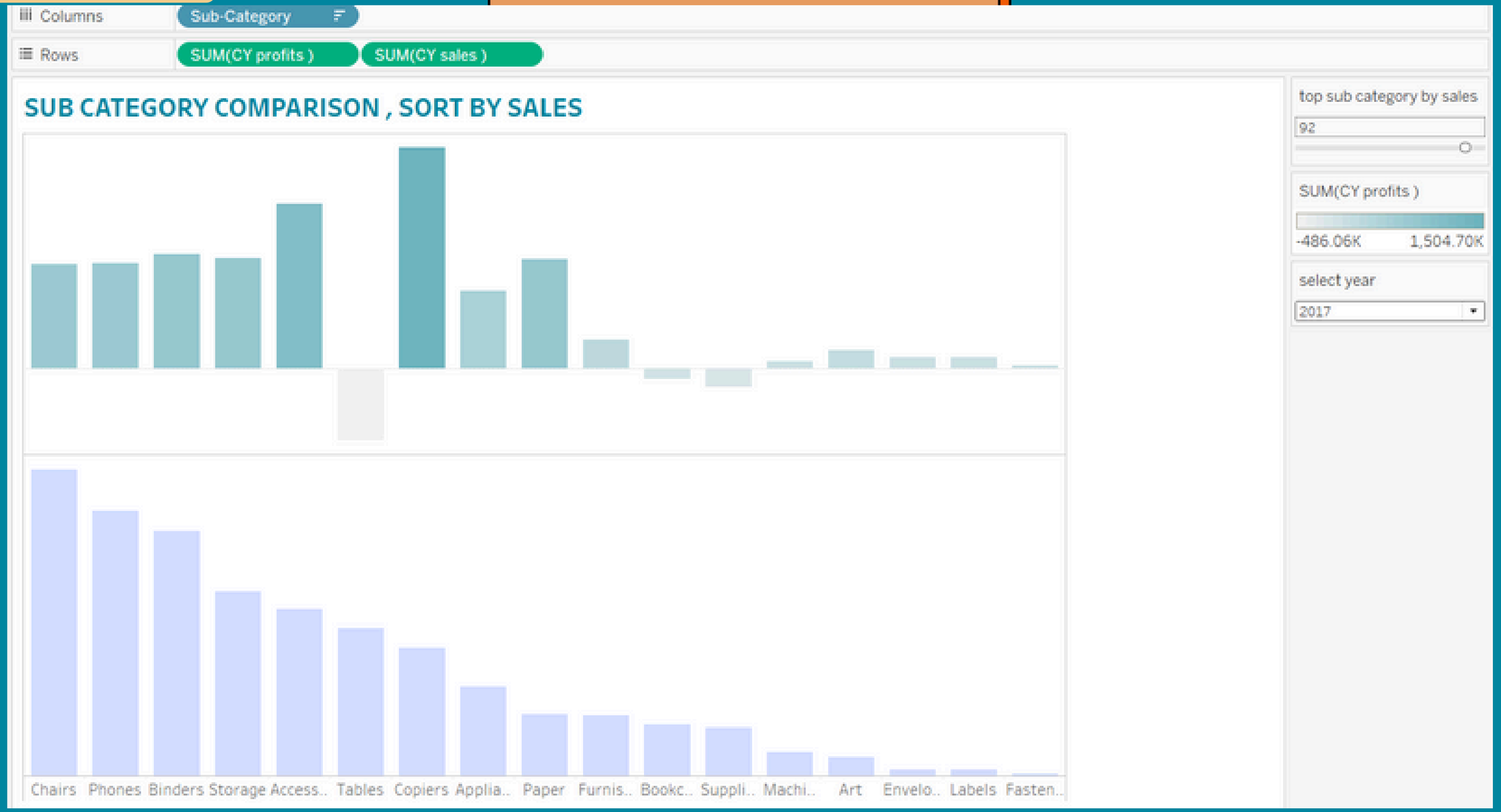
- High profit products ranked from high to low
- Products that are dark in color have a high percentage of sales, while products that are light in color have a low percentage of sales
- Top Profitable Products: Canon image CLASS 2200

Recommendations

- Focus on High-Profit Products: Invest more in marketing and promoting the top-profit products
- Reevaluate Low-Profit Products: For products with lower profits, it may be necessary to reassess pricing, distribution strategies, or even discontinue them if they do not add value.
- Diversify Product: Since profits are heavily concentrated in a few products, consider diversifying the product portfolio and improving the performance of mid-tier products to reduce risk and increase overall profitability.

Tableau

Sub category



Tableau

Insights

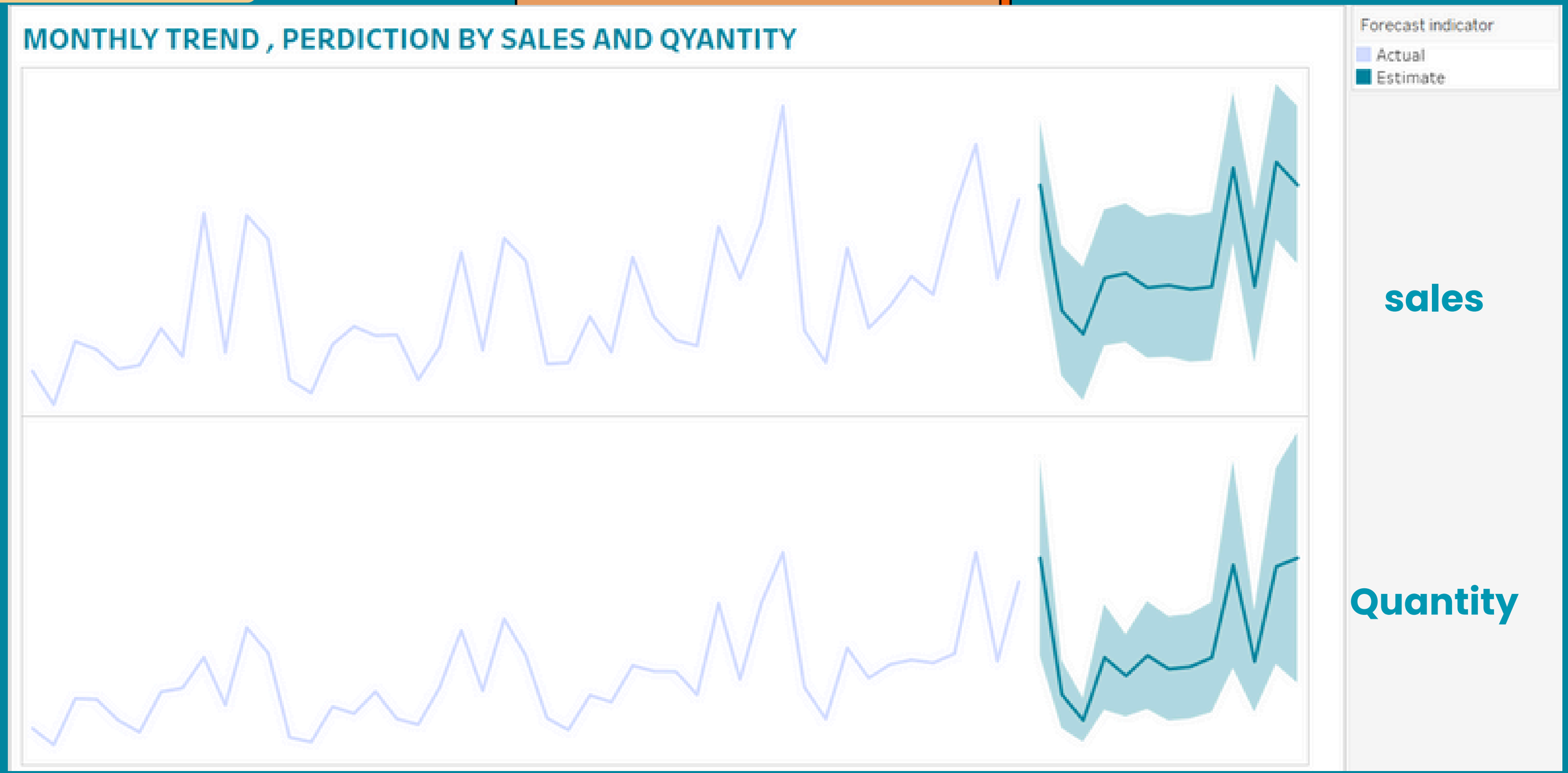
- High sales sub category ranked from high to low
- Sub category that are dark in color have a high percentage of profit , while products that are light in color have a low percentage of profit
- Top sales of Sub category: chairs

Recommendations

- Enhance Best-Selling Categories: Invest in promoting categories like Tables and Copiers through stronger marketing campaigns to boost sales and profits.
- Reduce Losses: Analyze the reasons for profit losses in categories like Binders and Storage Access, and work on improving efficiency or reducing costs.

Tableau

Monthly trend



Tableau

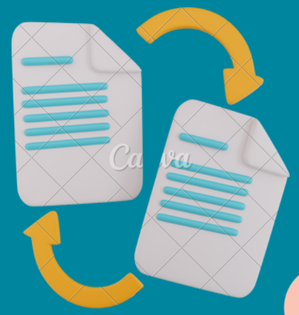
Insights

- During the years since 2014, fluctuations in production and profits
- According to the forecast line in Tableau, the first month of 2018 will have large profits and production,
- and after that, it will decrease in varying proportions and increase again at the end of the year.

Recommendations

- Follows up the reasons for the decline in production and profits
- Through the forecast line of the tableau, be careful of the middle months of the year because a decline is expected in them.

Requirements

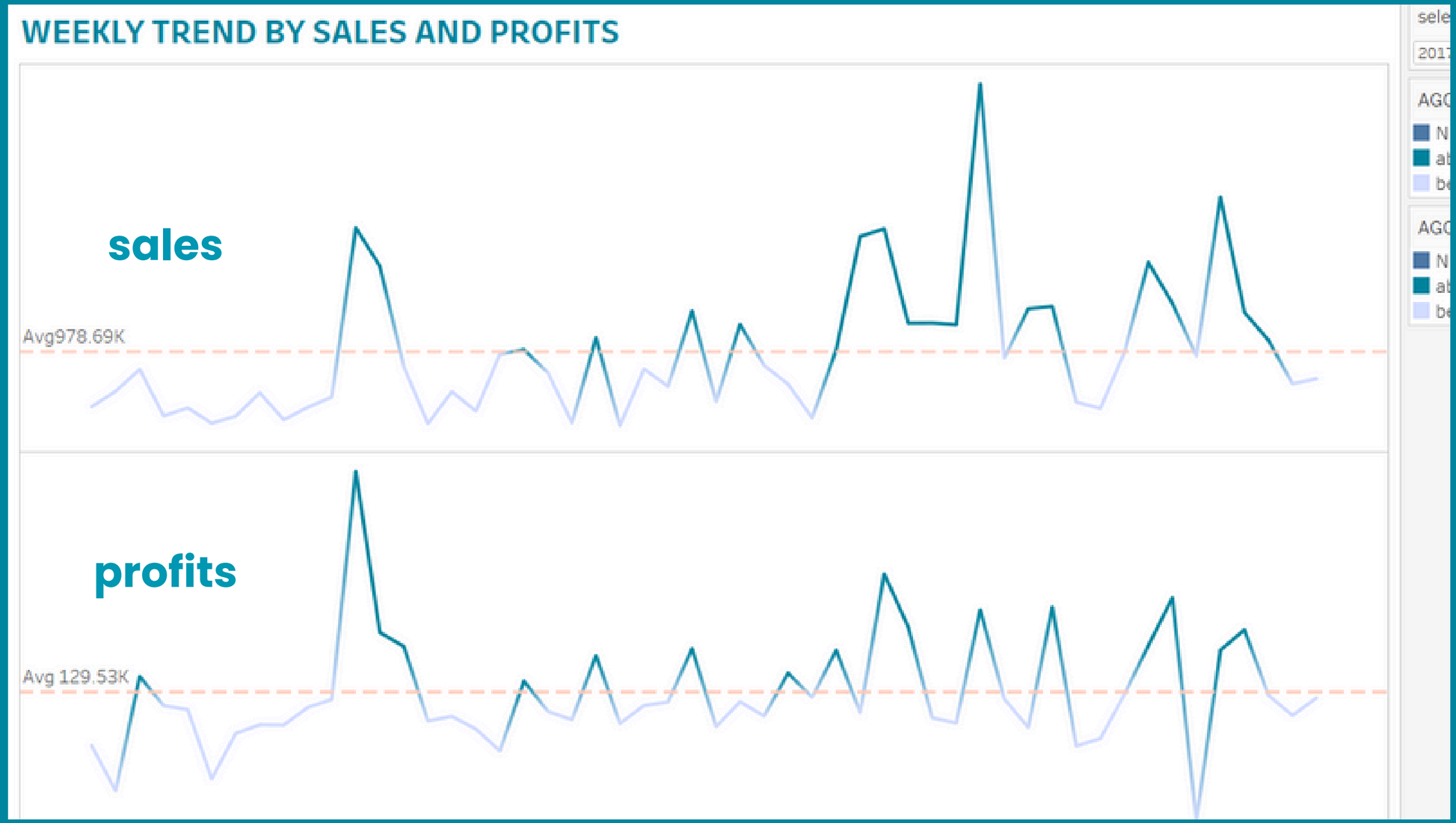


- Weekly trend by sales and profits
- Top products by state
- Top customers by profit
- Sales by category

- Maps
- Above or below avg profit
- Select year parameter

Tableau

weekly trend



Tableau

Insights

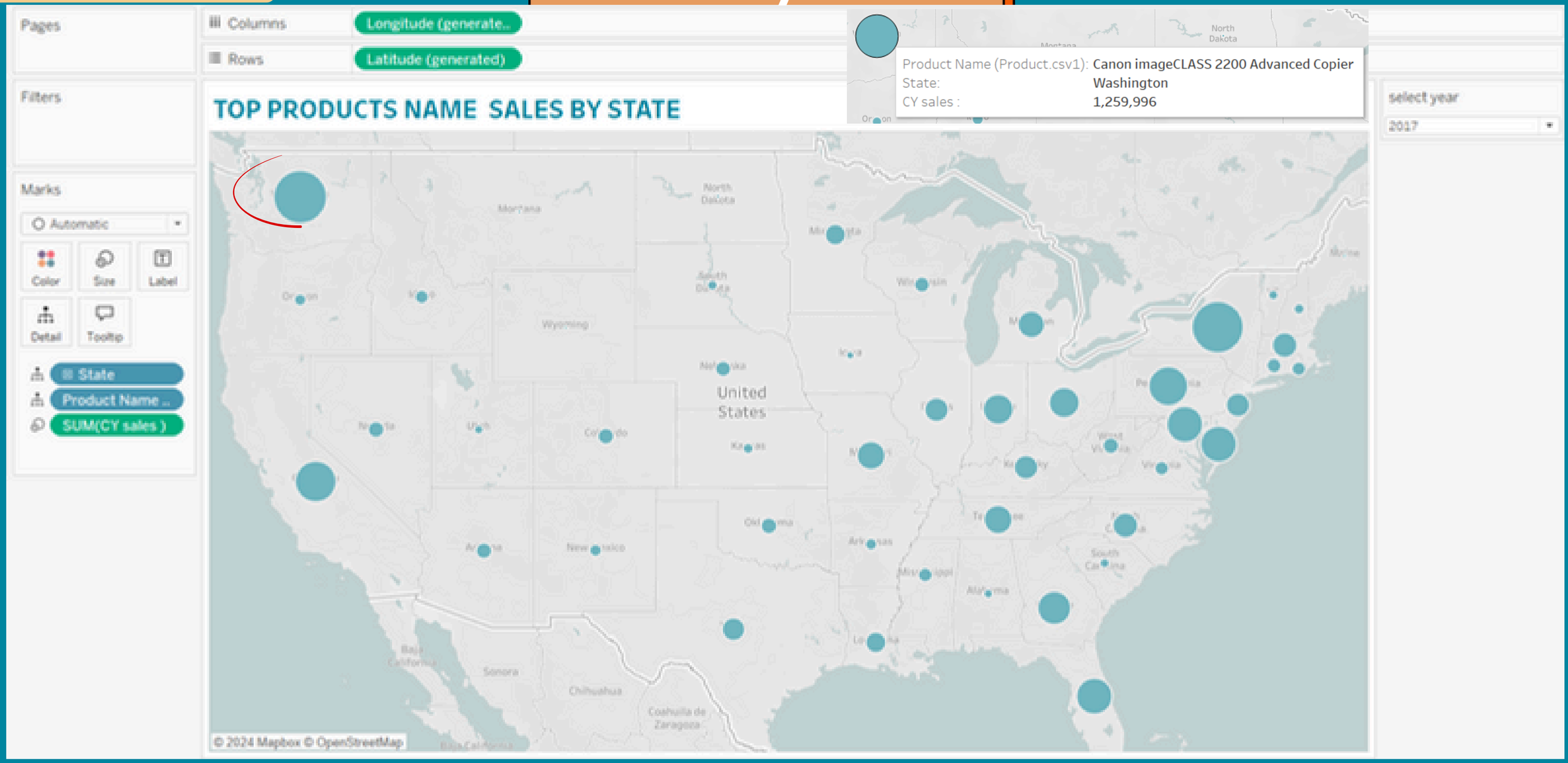
- Those above the average line are sales that are above the average and are considered to have made a good sales and profit
- Below the average line are sales and profit that are below the average and are considered to have made a small profit
- There are fluctuations in profits and sales during the weeks

Recommendations

- Follows up the reasons for the decline in sales and profits
- Try out a variety of sales and marketing strategies.

Tableau

Top products sales by state



Tableau

Insights

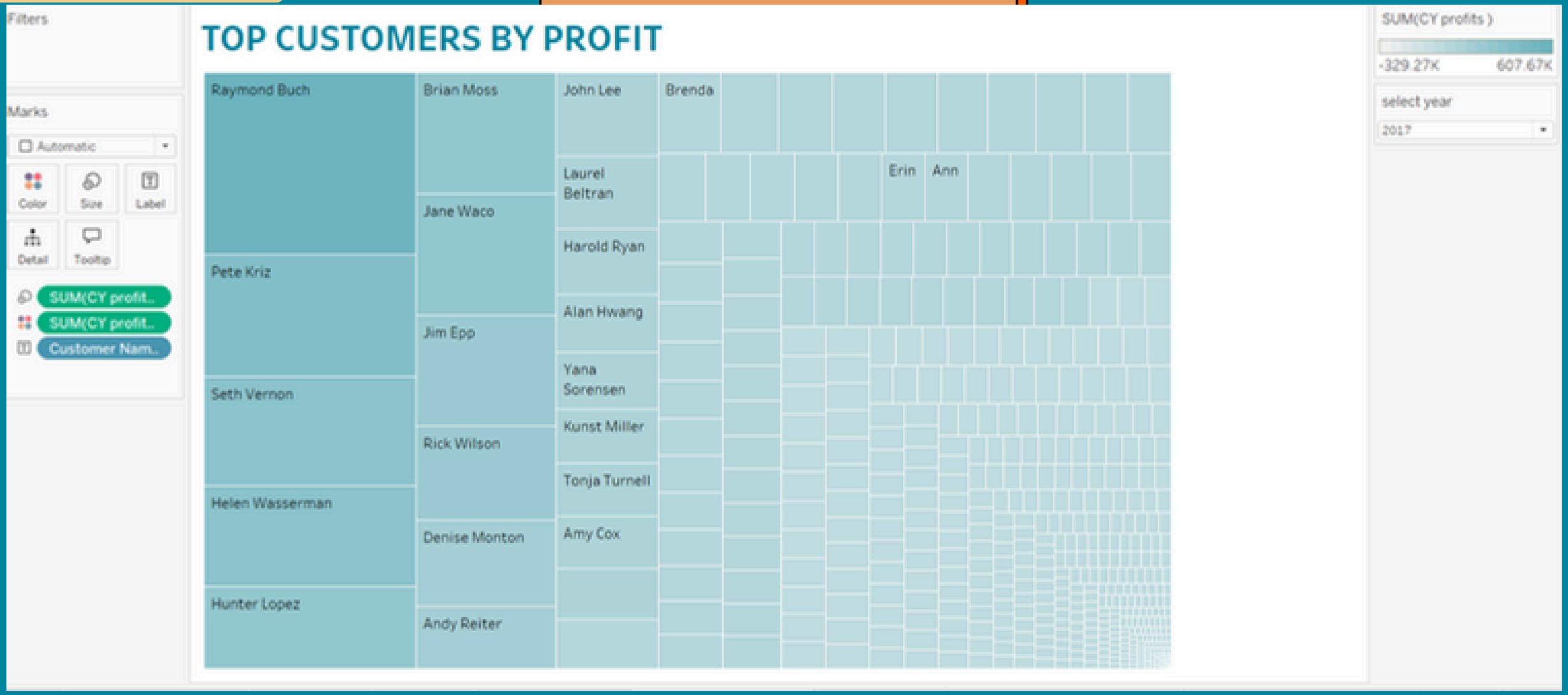
- Washington and New York top sales in canon image advanced copier and electric bending system
- state with large circles have high sales and vice versa.

Recommendations

- Increase production and change the marketing, sales and advertising strategy in areas with low sales.
- The 2 top products are considering the best seller in states Develop their profits to maintain the same level of profit.

Tableau

Top customers



Tableau

Insights

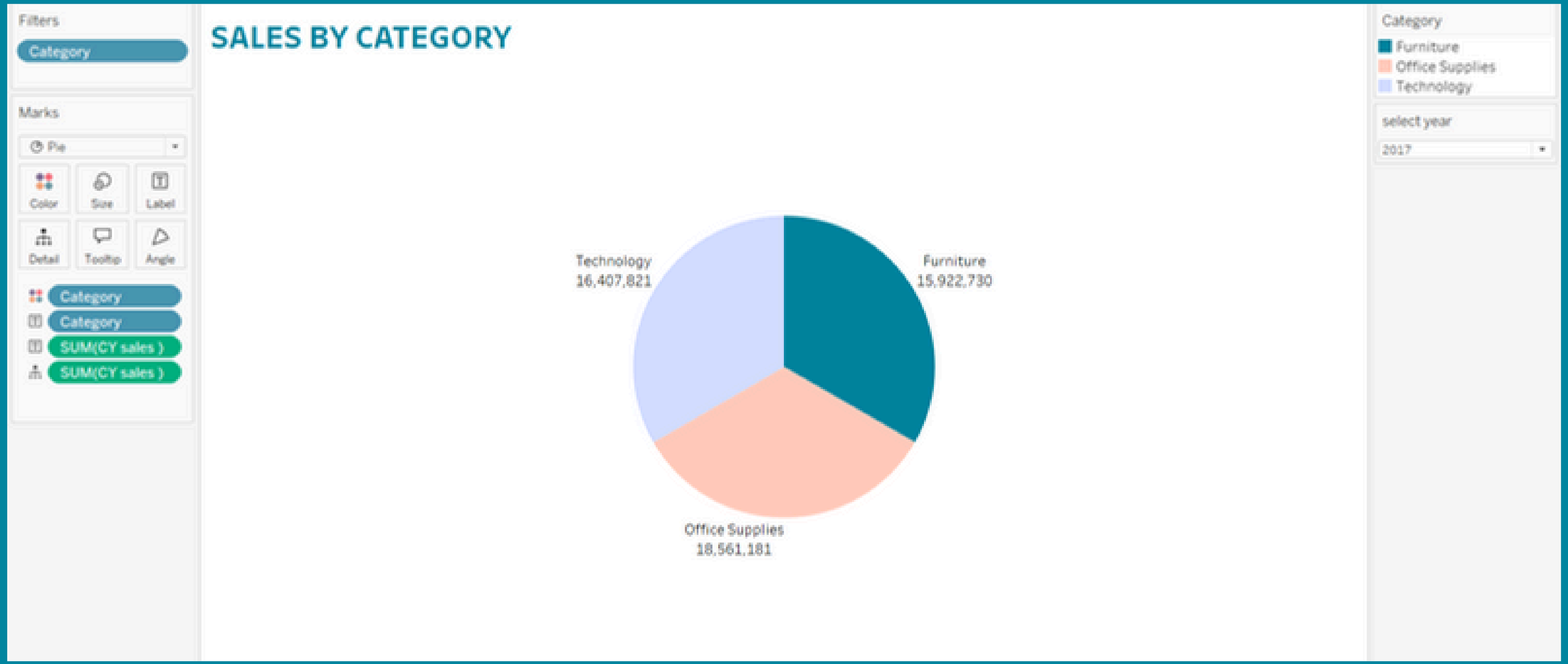
- Larger squares have higher customers' profits and vice versa
- The darker the color, the more that this customer make a big profit

Recommendations

- I keep customers who make a good profit and make a discount to ensure that they will always remain my customers
- I try to make a voucher, coupon, for whose profits are small

Tableau

Sales by category



Tableau

Insights

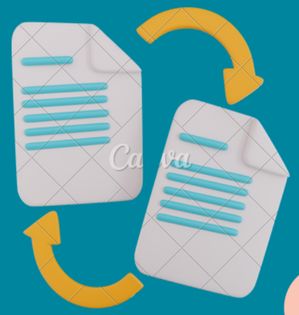
Top sales:

1. Office supplies then technology and furniture at the end
2. Technology
3. Furniture at the end

Recommendations

- Develop a furniture sales strategy, marketing method and advertising
- And keep the strategy of office supplies and technology

Requirements

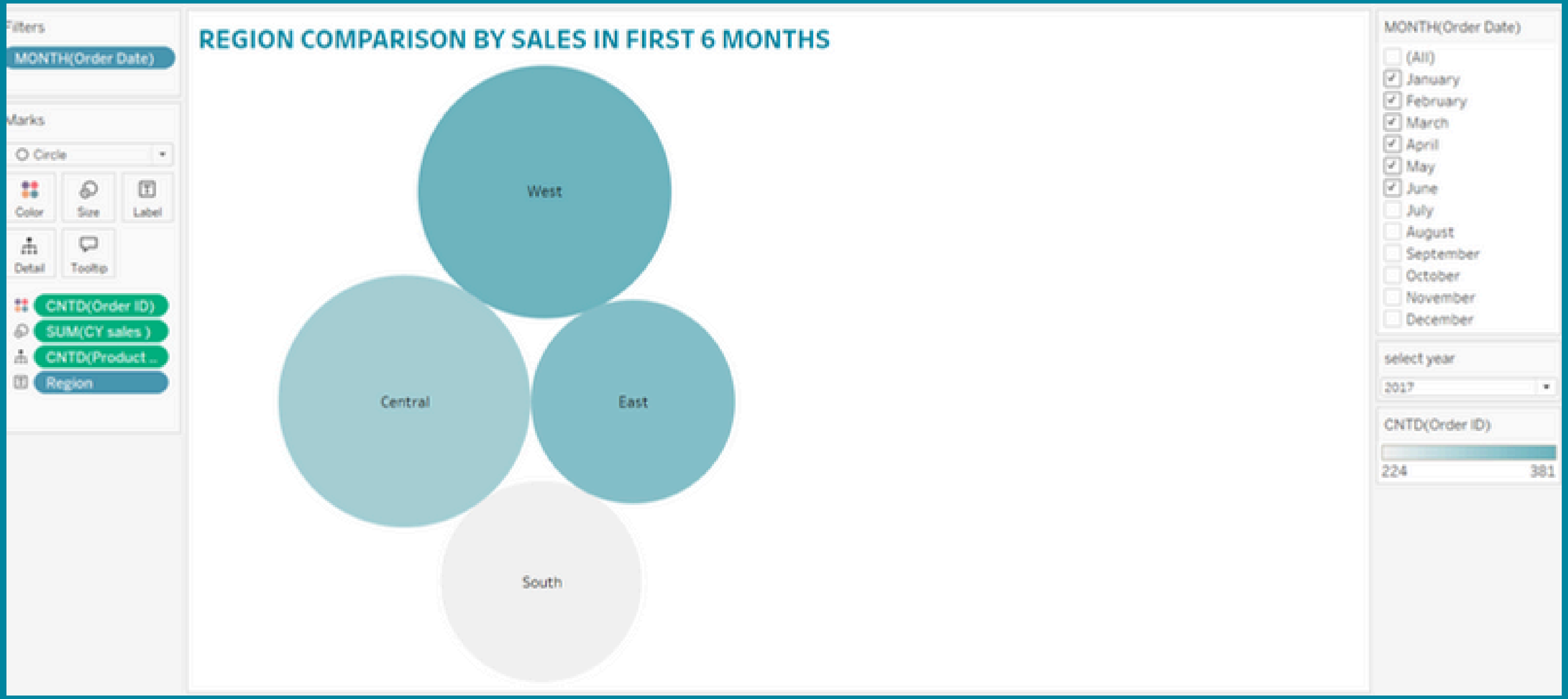


- Region comparison by sales in first 6 month , countd order id and product name
- COUNTD of orders by sub category in cities
- Profits based on discount
- Sales by ship mode

- Maps
- Month filters
- Select year parameter

Tableau

Region sales



Tableau

Insights

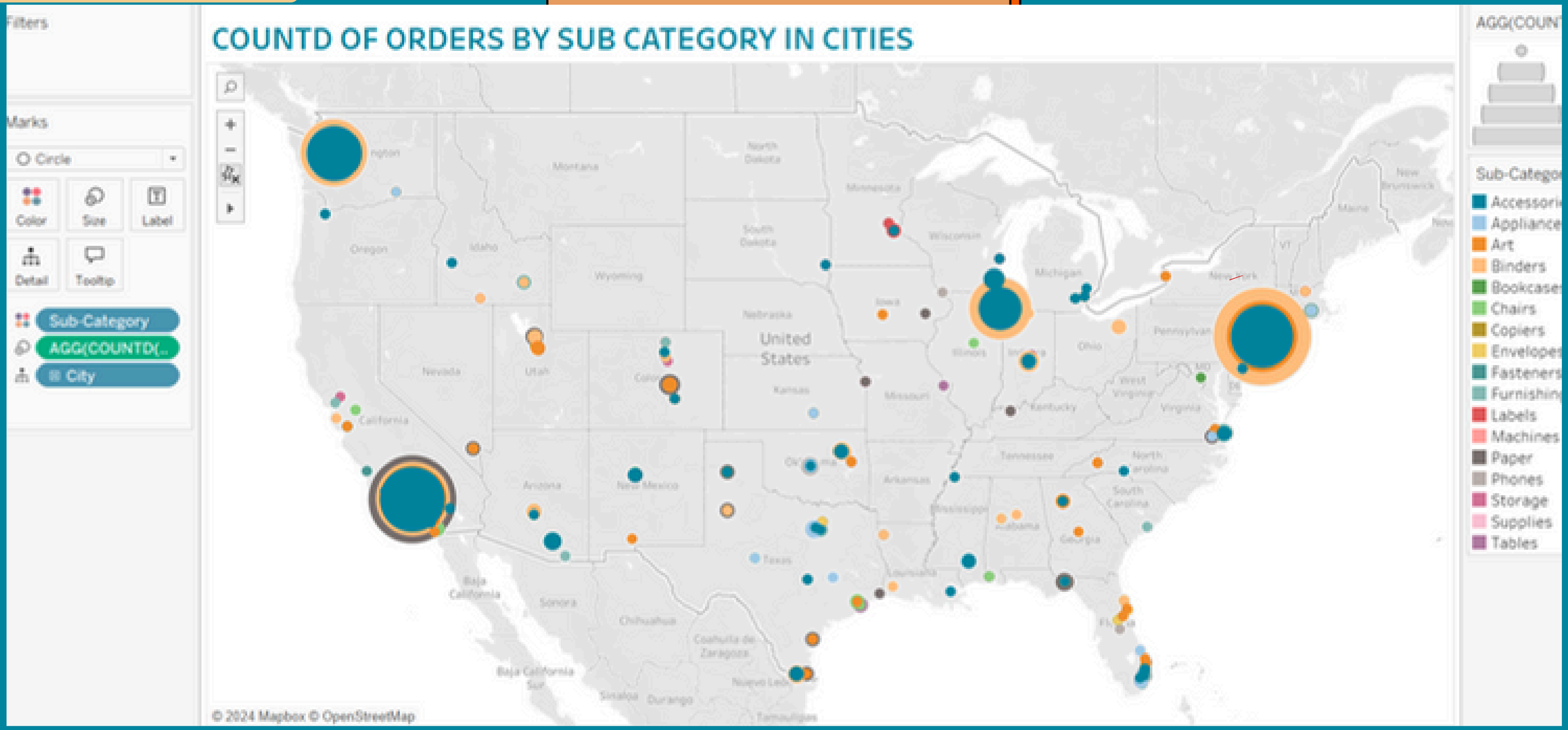
- The highest sales were in the West and central
- The lowest sales were in the east and central
- The West region has many orders, unlike the South

Recommendations

- Develop a south sales strategy, marketing method and advertising
- And keep the strategy of office west and central
- to ensure their top sales

Tableau

Orders



Tableau

Insights

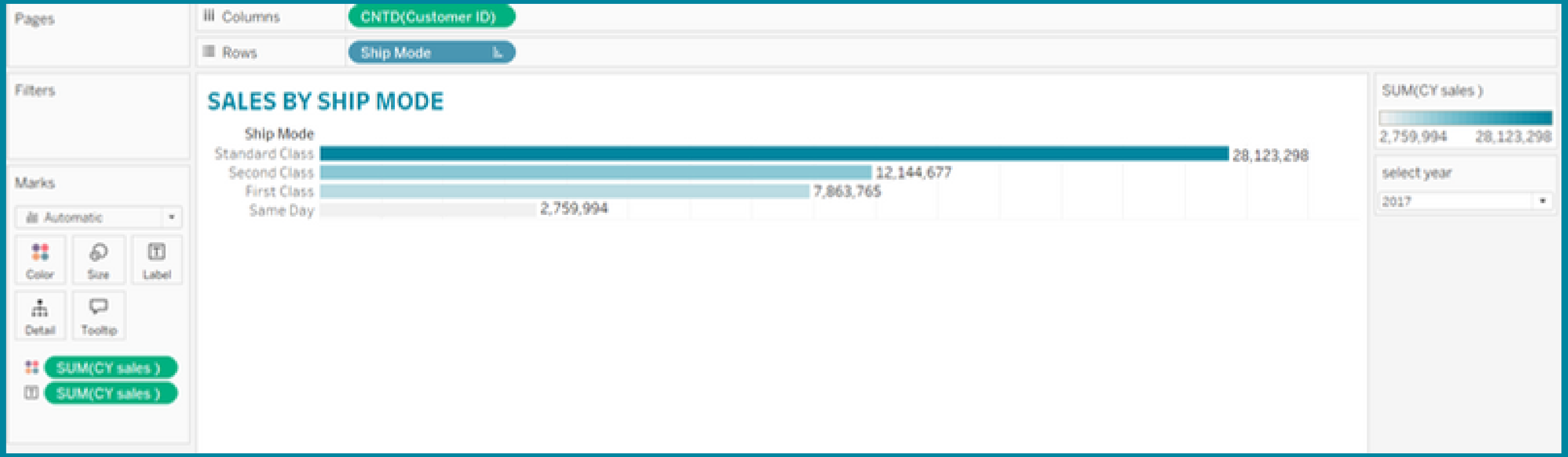
- The larger the circle, the more count of orders based on sub category
- New York has the largest number of orders : blinders

Recommendations

- Focus on areas with a small circles and on sales methods and strategies there
- And keep the strategy of New York l to ensure their top orders

Tableau

Sales by ship mode



Tableau

Insights

- Standard class is the highest sales then second class
- first class is the lowest sales then same day
- Also the Standard class is the highest count of customers then second class
- first class is the lowest sales then same day

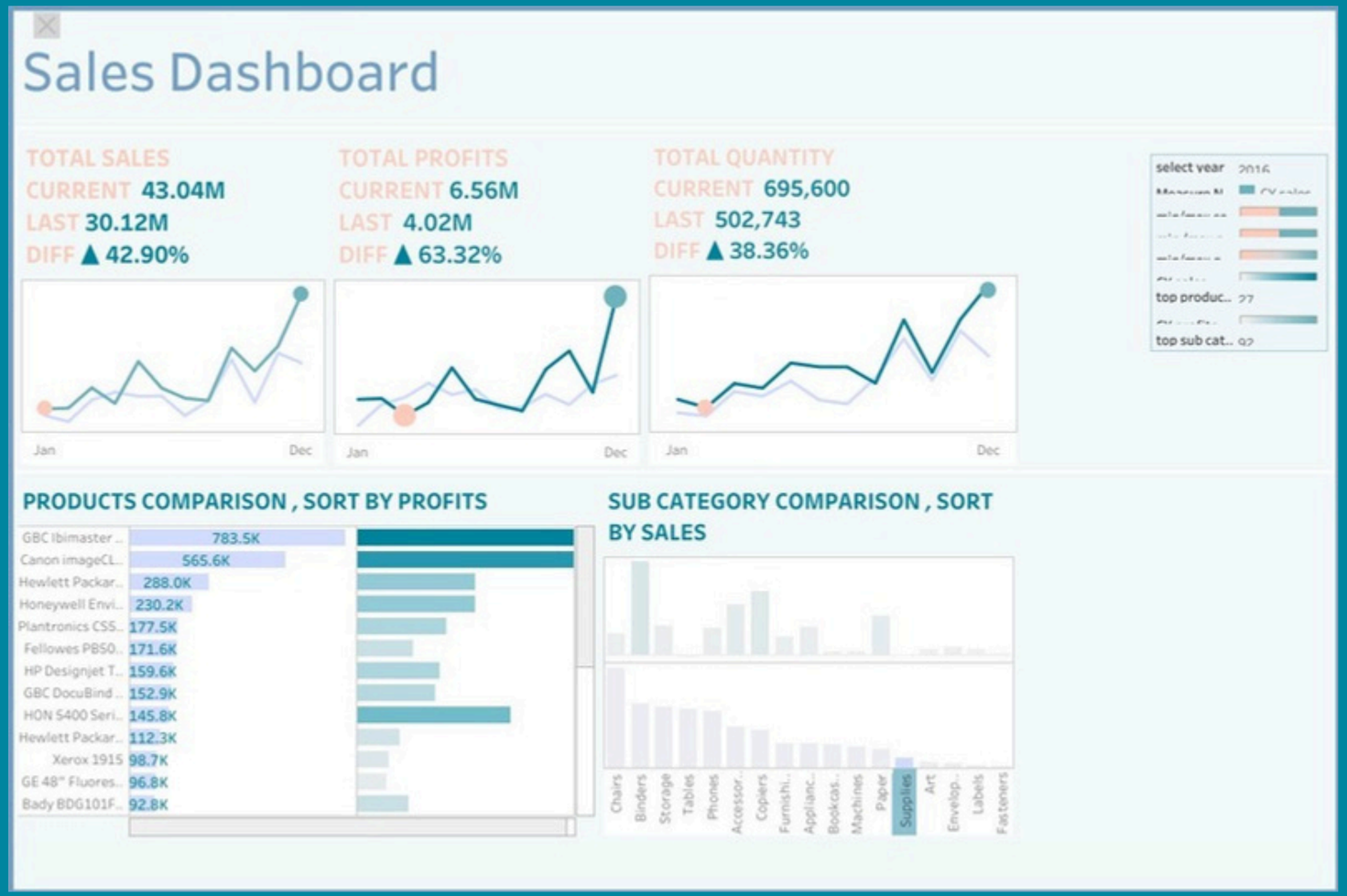
Recommendations

- Focus on count of customers in first class and same day, try to find new strategies for them
- And keep the strategy of standard class l and second class

Tableau Dashboards

Tableau

Sales Dashboard Analysis

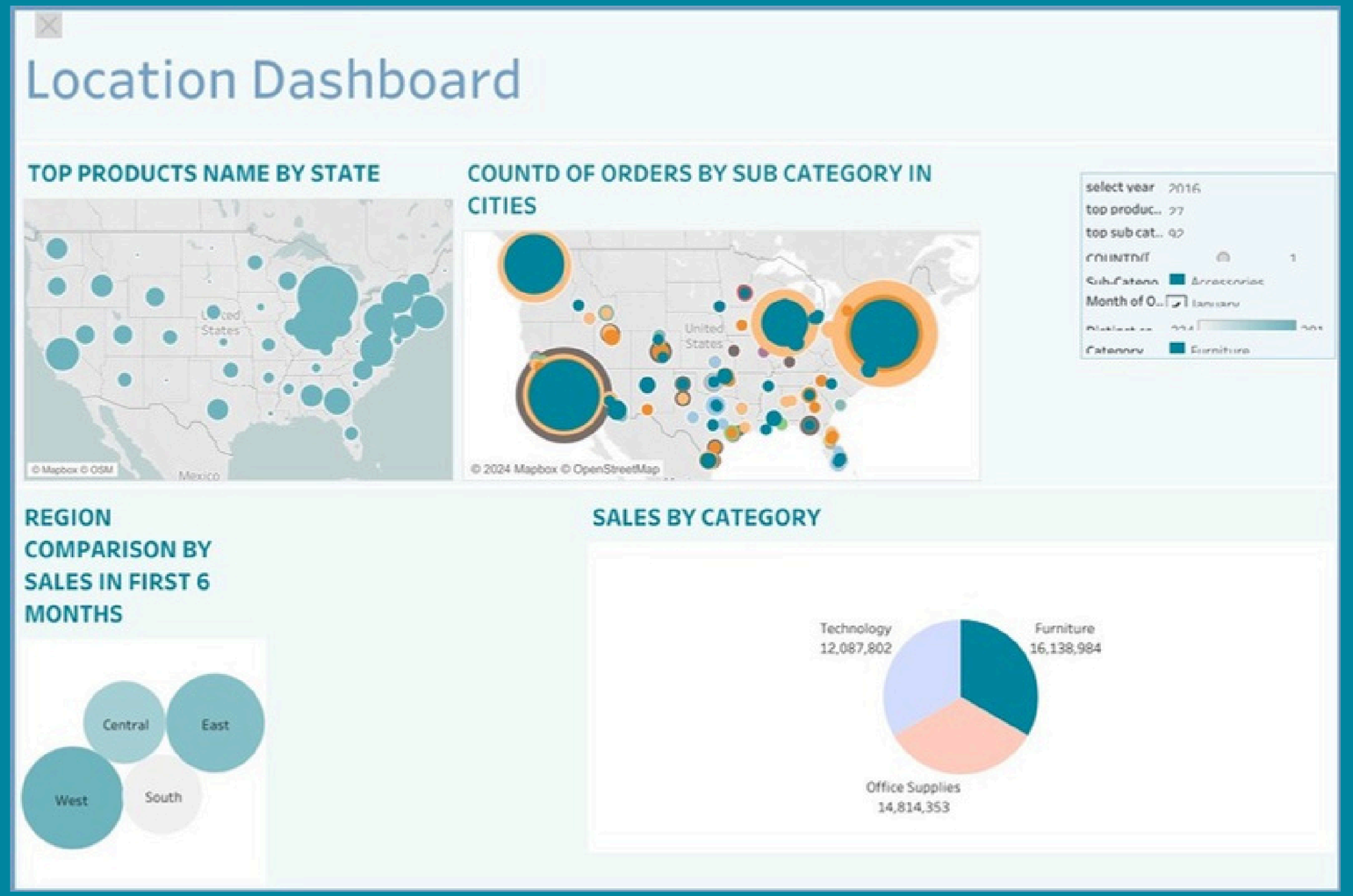


Tableau

Tops & Trends

Tableau

Location Dashboard



Tableau

Dashboard Details

- Horizontal Bars: Horizontal Containers used to group elements side by side.
- Vertical Bars: Vertical Containers used to group elements in a vertical layout.
- Blanks: Empty objects used to create space for better layout and spacing.
- Filters: Allow users to filter data on the dashboard based on specific criteria.

Conclusion

- In this project, we successfully analyzed the data to uncover key insights and trends. Our findings highlight significant patterns that can inform decision-making and strategy development. By leveraging advanced analytical techniques, we provided actionable recommendations that can enhance operational efficiency and drive growth. Future work should focus on continuous data monitoring and further refinement of our models to adapt to changing conditions.

Resources

https://www.assaal.com/file-share/fa740f32-9b0d-404d-b155-594c31c827a9


THANK
YOU!