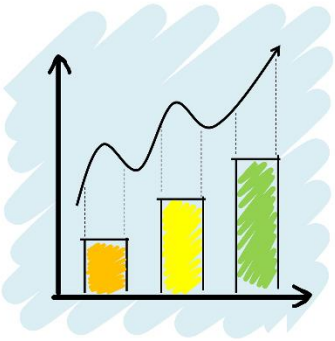
Superstore Dataset Analysis

INFO8075 SQL Final Project

Bo(Barry) Huang



Data Source: www.kaggle.com

Project Introduction

One of the superstore company wants to understand how they can improve their business by understanding their products, regions, categories and customer segments from their own sales data.

Dataset

This dataset is from 2014 to 2017. There are 3 tables included:

1. Customers:
    - Customer ID : Unique ID to identify each Customer.
    - Customer Name : Name of the Customer.
    - Segment : The segment where the Customer belongs.
    - Country : Country of residence of the Customer.
    - City : City of residence of the Customer.
    - State : State of residence of the Customer.
    - Postal Code : Postal Code of every Customer.
    - Region : Region where the Customer belong.
2. Orders:
    - Order ID : Unique Order ID for each Customer.
    - Order Date : Order Date of the product.
    - Ship Date : Shipping Date of the Product.
    - Ship Mode: Shipping Mode specified by the Customer.
    - Sales : Sales of the Product.
    - Quantity : Quantity of the Product.
    - Discount : Discount provided.
    - Profit : Profit/Loss incurred.
3. Products:
    - Product ID : Unique ID of the Product.
    - Category : Category of the product ordered.
    - Sub-Category : Sub-Category of the product ordered.
    - Product Name : Name of the Product

Objective

- Understand which products, regions, categories and customer segments the company should target or avoid.
- Try to provide some business suggestions to superstore on how to improve the business.

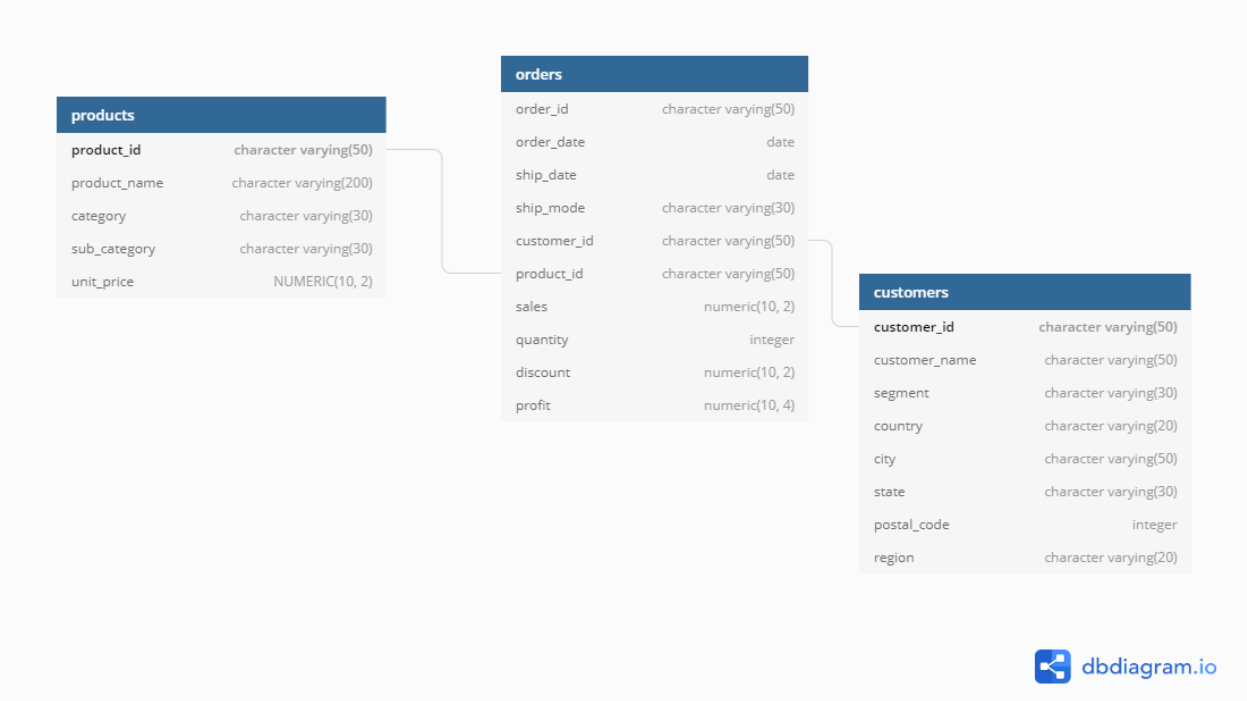
Data Preparation

-Create tables and copy data in the database

```
CREATE TABLE products (
product_id character varying(50) NOT NULL PRIMARY KEY,
product_name character varying(200)  NOT NULL,
category character varying(30) NOT NULL,
 sub_category character varying(30)  NOT NULL,
 unit_price NUMERIC(10,2)
        );
COPY products
FROM 'D:\Business Analytics\8075 SQL\Final Project\SuperStore Dataset\products.csv' WITH CSV HEADER;


CREATE TABLE customers (
customer_id character varying(50) NOT NULL PRIMARY KEY,
customer_name character varying(50)  NOT NULL,
segment character varying(30)  NOT NULL,
country character varying(20) NOT NULL,
city character varying(50)  NOT NULL,
state character varying(30) NOT NULL,
 postal_code integer NOT NULL,
 region character varying(20) NOT NULL
        );
```

```
COPY customers

FROM 'D:\Business Analytics\8075 SQL\Final Project\SuperStore Dataset\customers.csv' WITH CSV HEADER;


CREATE TABLE orders (

order_id character varying(50)  NOT NULL,

order_date date NOT NULL,

ship_date date NOT NULL,

ship_mode character varying(30) NOT NULL,

customer_id character varying(50) NOT NULL,

product_id character varying(50) NOT NULL,

sales numeric(10,2) NOT NULL,

quantity integer NOT NULL,

discount numeric(10,2) NOT NULL,

 profit numeric(10,4) NOT NULL,

FOREIGN KEY (product_id) REFERENCES products (product_id) ON UPDATE CASCADE ON DELETE CASCADE,

FOREIGN KEY (customer_id) REFERENCES customers (customer_id) ON UPDATE CASCADE ON DELETE CASCADE

        );

COPY orders

FROM 'D:\Business Analytics\8075 SQL\Final Project\SuperStore Dataset\orders.csv' WITH CSV HEADER;
```

-Diagram



-Data clean

1. In order to analysis easily, I will extract the year and month from the shipping date. The shipping date would be treated as the order completed day. Aand I add one column named "delivery_days" to show the number of days the customers need to wait for the orders.

```
ALTER TABLE orders

ADD COLUMN shipping_year INTEGER,

ADD COLUMN shipping_month INTEGER,

ADD COLUMN delivery_days INTEGER;


UPDATE orders

SET shipping_year = EXTRACT (year FROM ship_date);

UPDATE orders

SET shipping_month =EXTRACT(month FROM ship_date);
```

UPDATE orders

SET delivery_days= ship_date-order_date;

order_id character varying (50)	order_date date	ship_date date	ship_mode character varying (30)	customer_id character varying (50)	product_id character varying (50)	sales numeric (10,2)	quantity integer	discount numeric (10,2)	profit numeric (10,4)	shipping_year integer	shipping_month integer	delivery_days integer
CA-2015-117415	2015-12-27	2015-12-31	Standard Class	SN-20710	OFF-EN-10002986	113.33	9	0.20	35.4150	2015	12	
CA-2015-117415	2015-12-27	2015-12-31	Standard Class	SN-20710	FUR-BO-10002545	532.40	3	0.32	-46.9764	2015	12	
CA-2015-117415	2015-12-27	2015-12-31	Standard Class	SN-20710	FUR-CH-10004218	212.06	3	0.30	-15.1470	2015	12	
CA-2015-117415	2015-12-27	2015-12-31	Standard Class	SN-20710	TEC-PH-10000486	371.17	4	0.20	41.7564	2015	12	
CA-2017-120999	2017-09-10	2017-09-15	Standard Class	LC-16930	TEC-PH-10004093	147.17	4	0.20	16.5564	2017	9	
CA-2016-101343	2016-07-17	2016-07-22	Standard Class	RA-19885	OFF-ST-10003479	77.88	2	0.00	3.8940	2016	7	
CA-2017-139619	2017-09-19	2017-09-23	Standard Class	ES-14080	OFF-ST-10003282	95.62	2	0.20	9.5616	2017	9	
CA-2016-118255	2016-03-11	2016-03-13	First Class	ON-18715	TEC-AC-10000171	45.98	2	0.00	19.7714	2016	3	
CA-2016-118255	2016-03-11	2016-03-13	First Class	ON-18715	OFF-BI-10003291	17.46	2	0.00	8.2062	2016	3	
CA-2014-146703	2014-10-20	2014-10-25	Second Class	PO-18865	OFF-ST-10001713	211.96	4	0.00	8.4784	2014	10	

2. There are only 42 rows of data belongs to 2018, but it will affect the analysis easily. Here I will delete all data in 2018.

DELETE FROM orders

WHERE shipping_year=2018;

Analysis

--Check total sales, profit by year

SELECT shipping_year, SUM(sales) AS total_sales, SUM(profit)AS total_profit

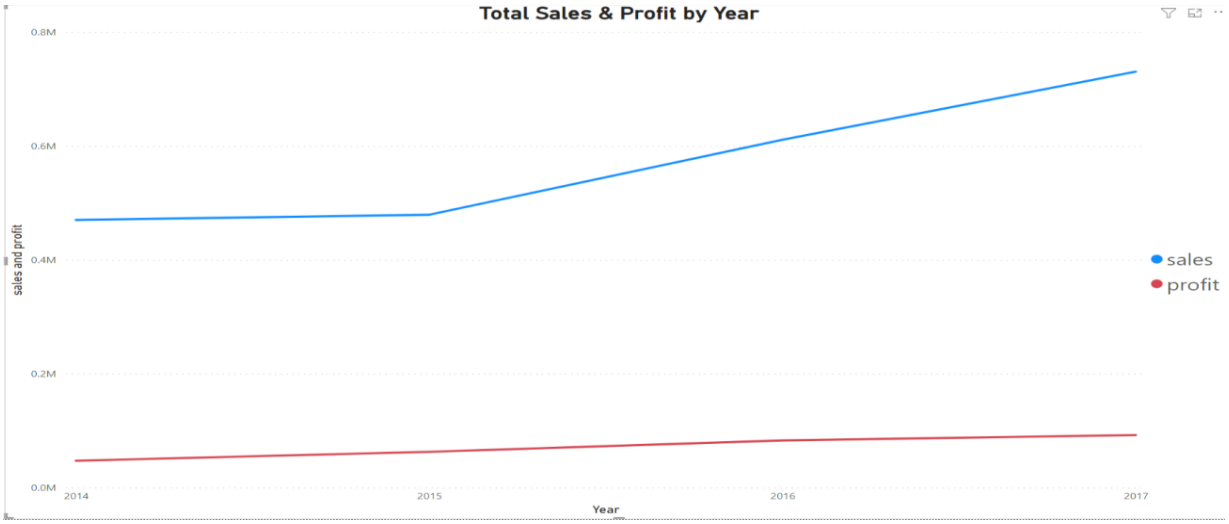
FROM orders

GROUP BY shipping_year

ORDER BY shipping_year;

	shipping_year integer	total_sales numeric	total_profit numeric
1	2014	470383.24	47292.7255
2	2015	479442.46	62881.5299
3	2016	611326.01	82941.1017
4	2017	730889.67	92346.8752

Here is the line chart for the table above:



By the chart, it is easily to find that from 2014 to 2017, the sales increased a lot, but the profit almost keep the same level. There should be something the superstore could do to increase the profit.

--Check sales, profit,discount by region and segment

SELECT shipping_year,region, segment, SUM(sales) AS total_sales, SUM(profit)AS total_profit, SUM(discount) AS total_discount

FROM customers C

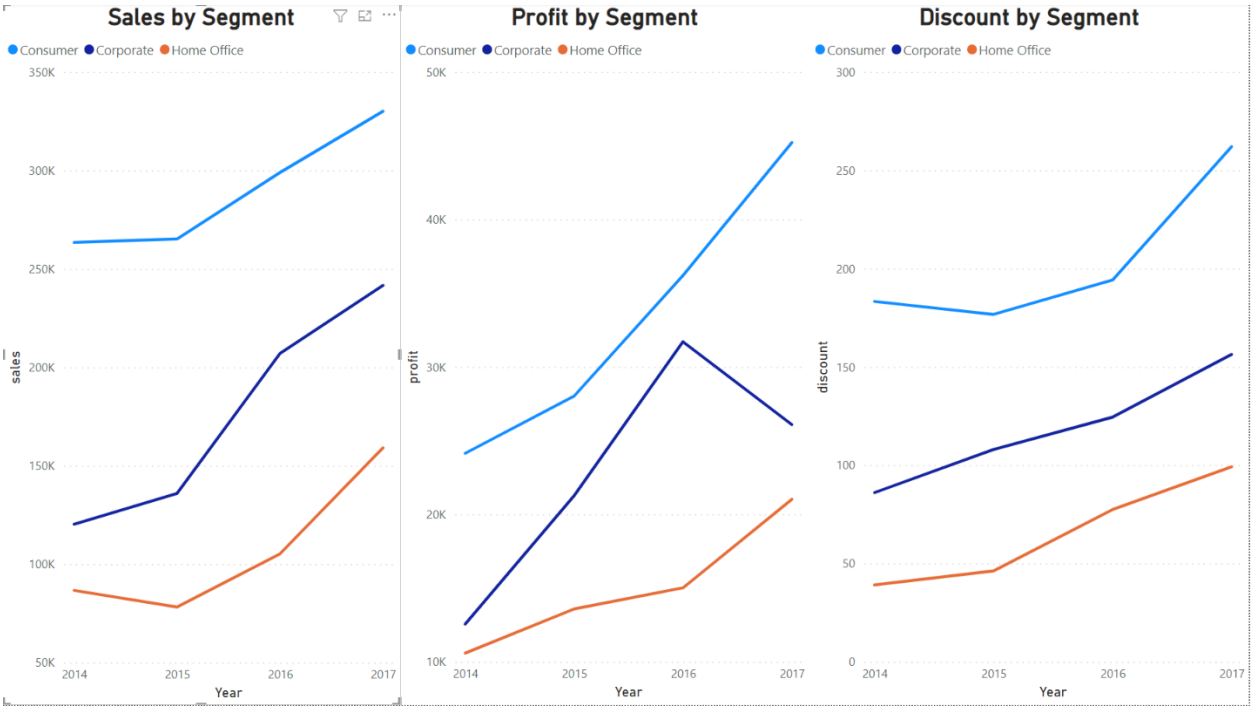
INNER JOIN orders O ON c.customer_id = o.customer_id

GROUP BY shipping_year,region, segment

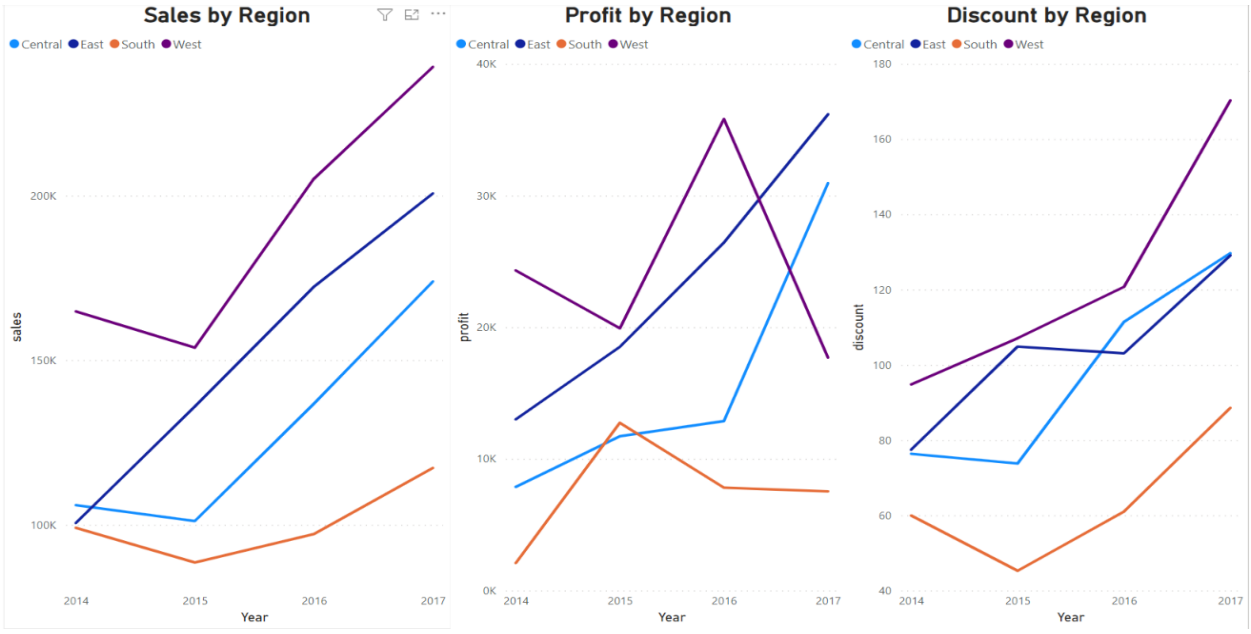
ORDER BY shipping_year,region;

shipping_year integer	region character varying (20)	segment character varying (30)	total_sales numeric	total_profit numeric	total_discount numeric
2014	Central	Corporate	29260.75	1928.5931	23.35
2014	Central	Home Office	14264.19	3057.4214	10.60
2014	Central	Consumer	62445.20	2888.1161	42.42
2014	East	Consumer	51454.30	5822.0948	44.60
2014	East	Home Office	18313.54	4685.1235	10.05
2014	East	Corporate	30759.98	2502.6602	22.80
2014	South	Consumer	35479.55	-1968.2795	37.60
2014	South	Home Office	37293.66	1024.8023	5.20
2014	South	Corporate	26303.43	3031.8210	17.15
2014	West	Home Office	16812.27	1825.8422	13.32

Here are charts based on the table above:

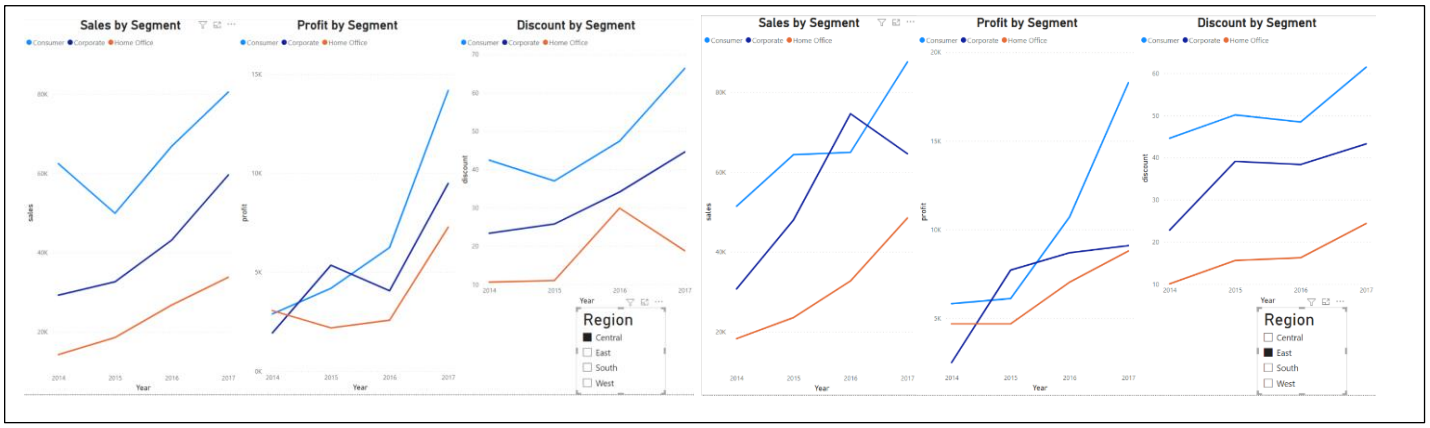


The charts above shows overall sales and profit by segment, consumer segment is always good, though the company spent more cost on promotion (discount), but the sales and profit of consumer segment also increase a lot, same as the home office segment. Though the corporate segment's sales increased following the promotion, the profit of corporate segment's profit decreased a lot in 2017. That might be the promotions in corporate segment spent too much cost. The strategy for corporate segment should be adjusted.

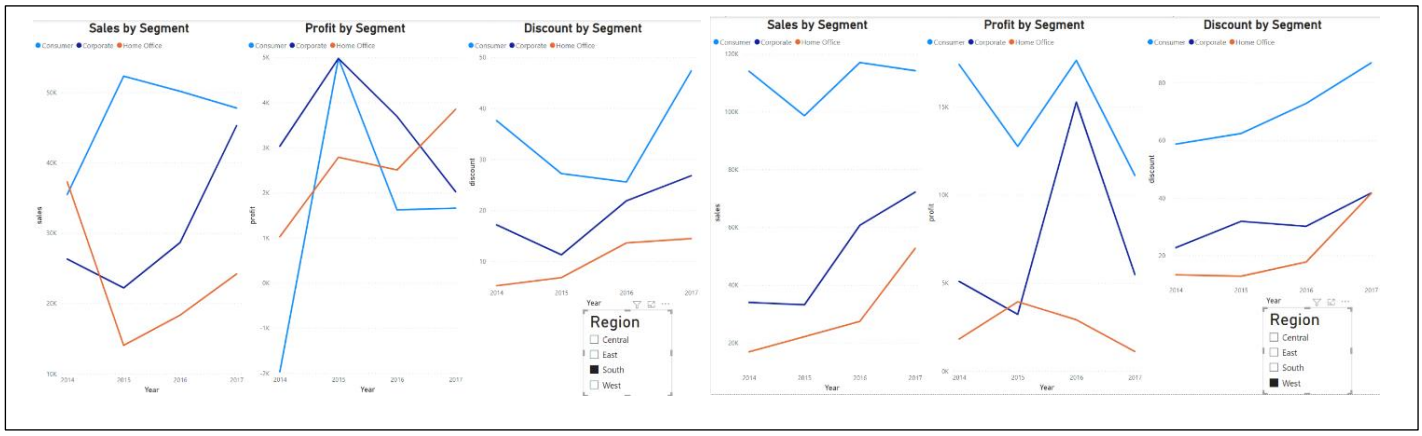


The charts above shows the overall sales and profit based on regions. Sales in all regions had a good trend from 2014 to 2017, But both west and south regions had problems on profits, especially the west region. The profit in west region had a dramatic drop in 2017, and the promotions in west region increased too much.

Then I checked sales and profit in segments by regions, here are charts:



Central Region East Region



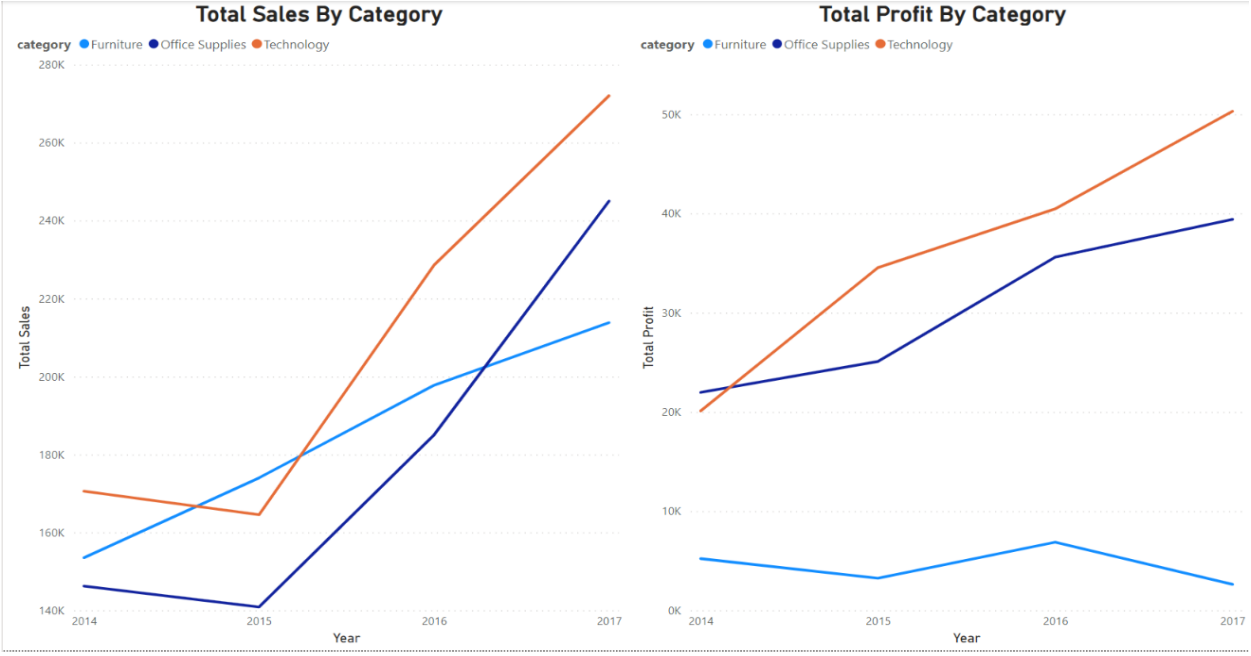
South Region West Region

When I checked the charts, obviously the west region had big problem on marketing and sales strategy. The profit of all three segments in west region decreased from 2014 to 2017 though sales increased at the same period. The company spent too much cost on promotions in west region. The south region also had the same problem, but the home office segment in south region increased at least. The other two regions, central and east, were doing well.

--Check sales, profit, and discount by product, category and sub_category in each year

```
SELECT shipping_year,  category,  SUM(sales) AS total_sales, SUM(profit) AS total_profit

FROM orders O

INNER JOIN products P ON O.product_id = P.product_id

GROUP BY shipping_year,  category

ORDER BY shipping_year,  category;
```

	shipping_year integer	category character varying (30)	total_sales numeric	total_profit numeric
1	2014	Furniture	153528.79	5218.0549
2	2014	Office Supplies	146256.21	21967.9198
3	2014	Technology	170598.24	20106.7508
4	2015	Furniture	174006.62	3245.3851
5	2015	Office Supplies	140885.46	25088.8687
6	2015	Technology	164550.38	34547.2761
7	2016	Furniture	197764.01	6872.9005
8	2016	Office Supplies	185005.61	35600.7818
9	2016	Technology	228556.39	40467.4194
10	2017	Furniture	213836.15	2625.7887



From the chart above, office supplies and technology products were doing well in both sales and profit since 2014, and the technology products were the best category. Though the sales of furniture increased all the time, the profit was not exciting. It even decreased in 2017.

--Check the top 3 sub_categories by sales and prfit

```
SELECT category, sub_category,
sub_category_sales_rank,sub_category_profit_rank

FROM (

        WITH order_product AS (

                SELECT*

                FROM orders O

                INNER JOIN products P ON O.product_id = P.product_id)

               South Region                                                West Region

        SELECT sub_category, category, SUM(sales), SUM(profit),

        RANK() OVER (PARTITION BY category ORDER BY SUM(sales) DESC) AS
sub_category_sales_rank,

        RANK() OVER (PARTITION BY category ORDER BY SUM(profit) DESC)
AS sub_category_profit_rank

        FROM order_product

        GROUP BY category, sub_category) C

WHERE sub_category_sales_rank =1 OR sub_category_profit_rank=1
```

	category character varying (30)	sub_category character varying (30)	sub_category_sales_rank bigint	sub_category_profit_rank bigint
1	Furniture	Chairs	1	1
2	Office Supplies	Storage	1	3
3	Office Supplies	Paper	4	1
4	Technology	Phones	1	2
5	Technology	Copiers	4	1

There are quite some sub categories in each category. To make it easy, I checked the No. 1 sub category in sales and profit. Based on the sales and profit on both category and sub category, the superstore should pay more attention on promoting products in sub categories like copiers, phones , storage and paper.

--Check top 3 products by sales and profit

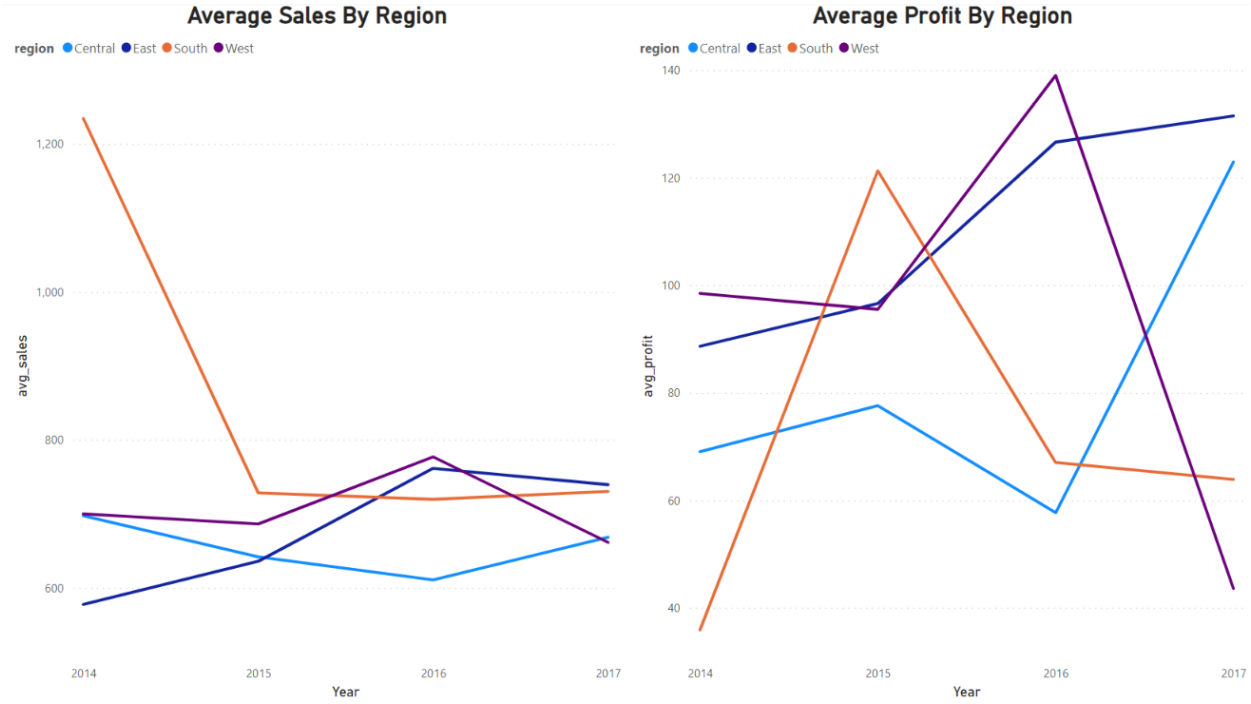
```
SELECT category, product_name, product_sales_rank,product_profit_rank

FROM (

    WITH order_product AS (

        SELECT*

        FROM orders O

        INNER JOIN products P ON O.product_id = P.product_id)

    SELECT product_name, category, SUM(sales), SUM(profit),

    RANK() OVER (PARTITION BY category ORDER BY SUM(sales) DESC) AS product_sales_rank,

    RANK() OVER (PARTITION BY category ORDER BY SUM(profit) DESC) AS product_profit_rank

    FROM order_product

    GROUP BY category, product_name) C

WHERE product_sales_rank <4 OR product_profit_rank<4
```

	category character varying (30)	product_name character varying (200)	product_sales_rank bigint	product_profit_rank bigint
1	Furniture	HON 5400 Series Task Chairs for Big and Tall	1	251
2	Furniture	Riverside Palais Royal Lawyers Bookcase, Royale Cherry Finish	2	359
3	Furniture	Bretford Rectangular Conference Table Tops	3	333
4	Furniture	Hon Deluxe Fabric Upholstered Stacking Chairs, Rounded Back	7	1
5	Furniture	Global Deluxe High-Back Manager's Chair	12	2
6	Furniture	Hon Pagoda Stacking Chairs	26	3
7	Office Supplies	Fellowes PB500 Electric Punch Plastic Comb Binding Machine with Manual Bind	1	1
8	Office Supplies	GBC DocuBind TL300 Electric Binding System	2	6
9	Office Supplies	GBC Ibimaster 500 Manual ProClick Binding System	3	29
10	Office Supplies	Ibico EPK-21 Electric Binding System	7	2

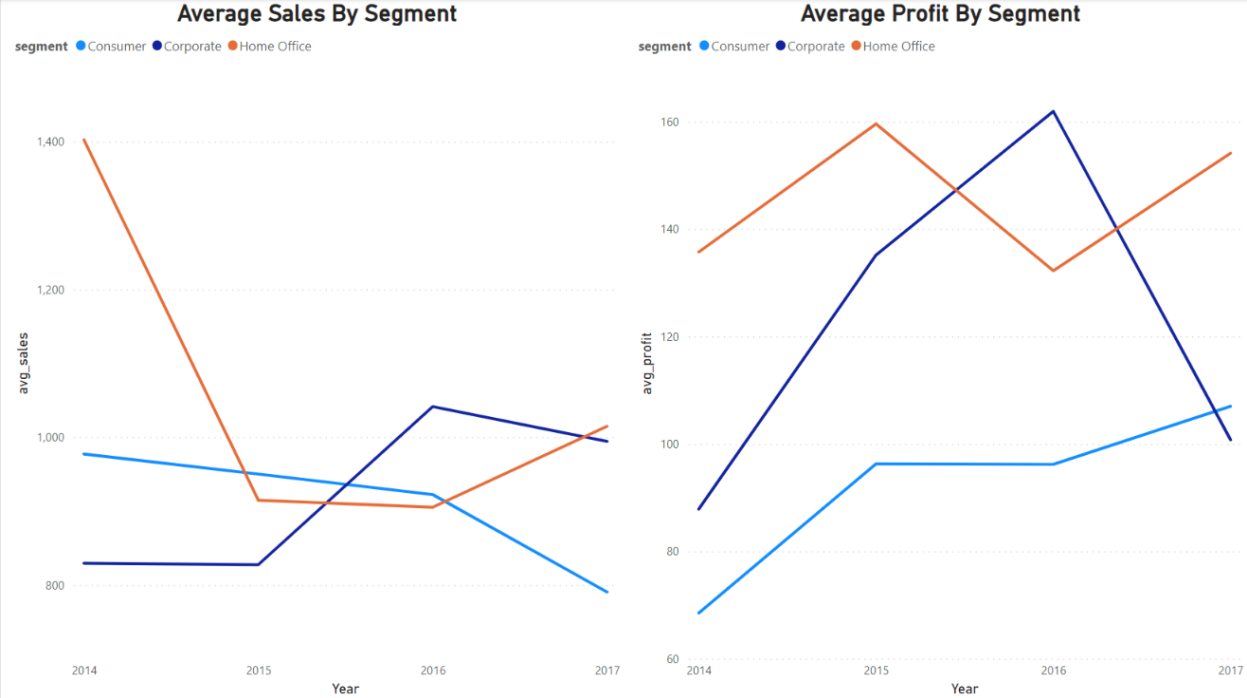
Same as sub category, I checked top 3 products by sales and profit in each category. There are 15 products showed in the table. Those products should be promoting more in the future.

--Check average sales and profit by region and segment, create the view

```
CREATE VIEW average_sales_profit AS

SELECT shipping_year, region, segment, AVG(sales) AS avg_sales, AVG(profit) AS avg_profit

FROM orders O

INNER JOIN customers C ON O.customer_id = C.customer_id

GROUP BY shipping_year, region, segment

ORDER BY shipping_year, region;
```



By the table and charts, we could find the average order size dropped a lot in west region, but the other 3 regions almost kept the same level. Regarding the average profit, south region and west regions decreased quite a lot, especially west region, though central and east regions increased .



From the average order size by segment perspective, the home office segment would be the most potential segment for superstore.

--Check average delivery days for each ship mode

SELECT ship_mode, ROUND(AVG(delivery_days),0)

FROM orders

GROUP BY ship_mode;

	ship_mode character varying (30)	round numeric
1	Second Class	3
2	Standard Class	5
3	Same Day	0
4	First Class	2

From the right table, it shows that the standard class takes the longest time to deliver products to customer, and the customer could receive the product by selecting the same day ship mode.

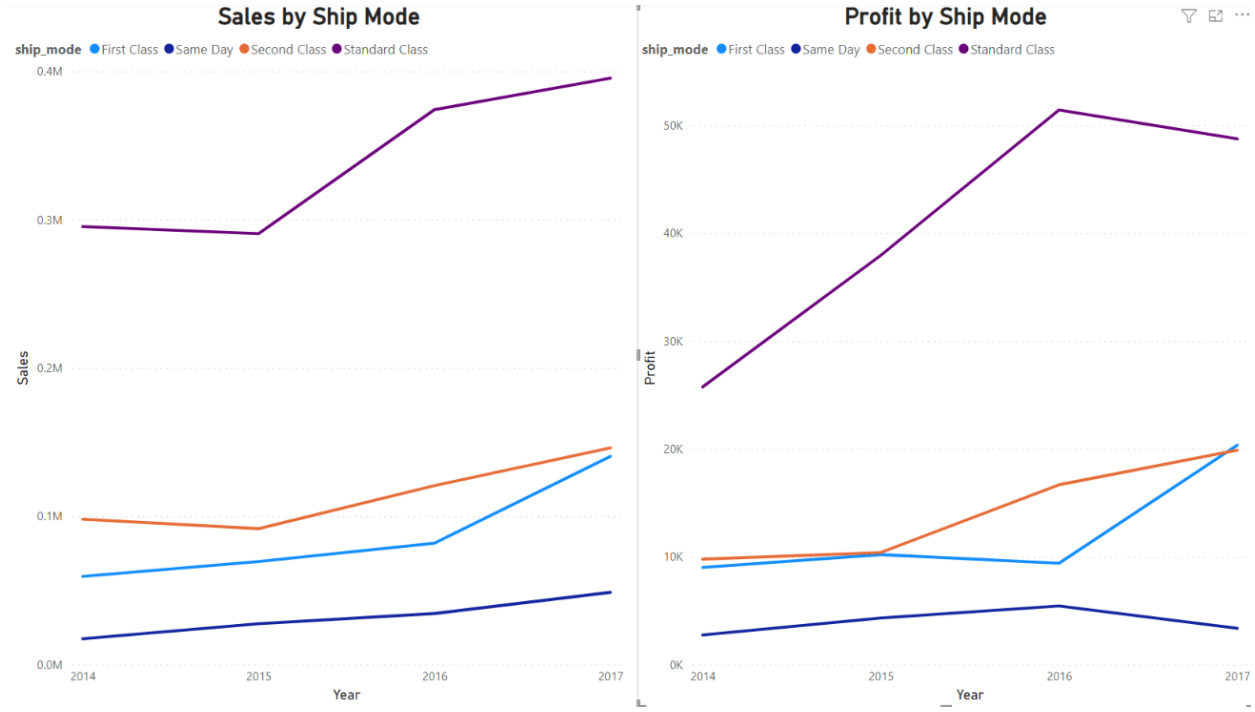
--Check sales and profit by ship mode

SELECT shipping_year, ship_mode, SUM(sales) AS sales_shipmode, SUM(profit) AS profit_shipmode

FROM orders

GROUP BY shipping_year, ship_mode

	shipping_year integer	ship_mode character varying (30)	sales_shipmode numeric	profit_shipmode numeric
1	2017	Same Day	48775.75	3368.3751
2	2015	Same Day	27611.46	4327.4275
3	2016	Same Day	34505.80	5440.4492
4	2015	Second Class	91706.37	10396.8400
5	2014	First Class	59499.28	9011.4321
6	2016	Second Class	120715.36	16677.5172
7	2014	Standard Class	295343.70	25742.7776
8	2014	Second Class	98070.15	9783.0087
9	2014	Same Day	17470.11	2755.5071
10	2017	First Class	140485.10	20350.0686



Obviously, most customers chose standard ship mode because of the shipping cost, but it seemed more and more people chose fist class and second class ship mode in order to receive orders faster. Superstore might pay more attention on those two ship modes to increase the customer's satisfaction level with increasing company's profit.

--Check ship modes used in regions

CREATE VIEW order_count_region AS

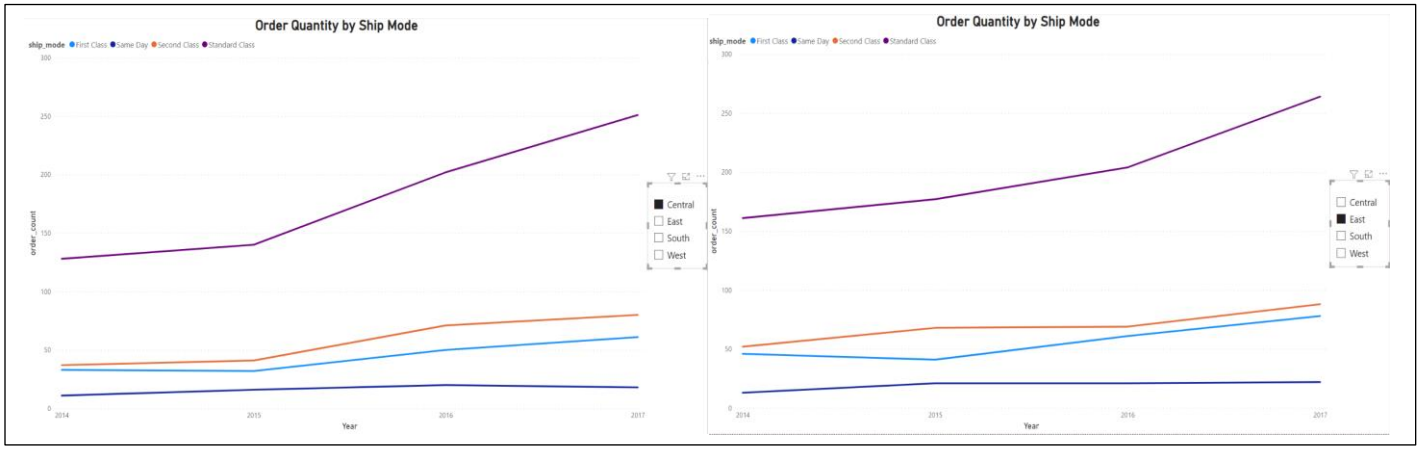
SELECT shipping_year, region, COUNT(DISTINCT(order_id)) AS order_count, ship_mode

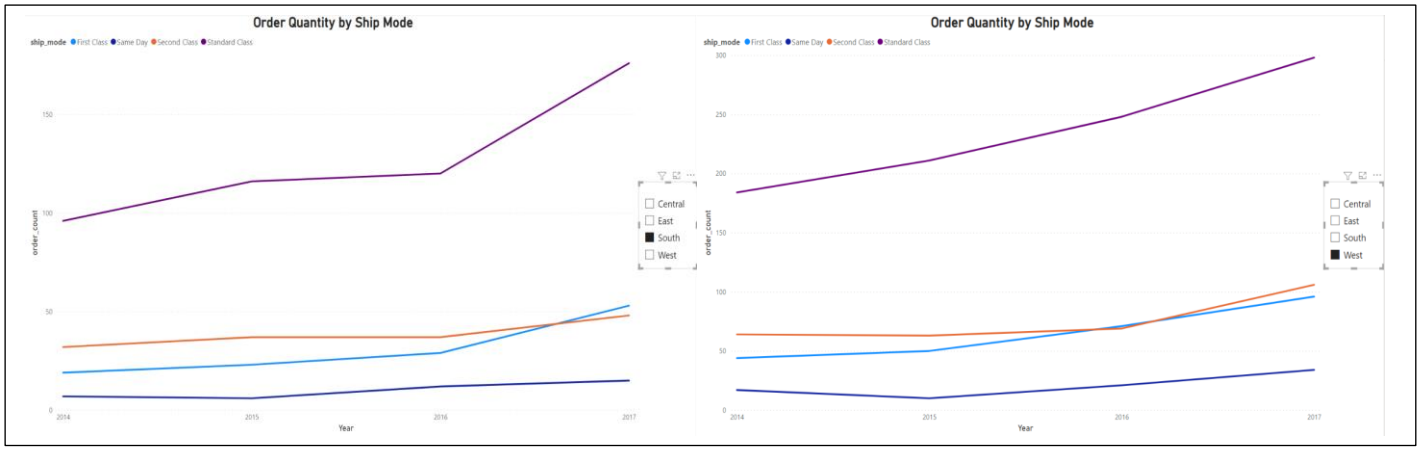
FROM orders O

INNER JOIN customers C ON O.customer_id = C.customer_id

GROUP BY shipping_year, region, ship_mode;

	shipping_year integer	region character varying (20)	order_count bigint	ship_mode character varying (30)
1	2014	Central	33	First Class
2	2014	Central	11	Same Day
3	2014	Central	37	Second Class
4	2014	Central	128	Standard Class
5	2014	East	46	First Class
6	2014	East	13	Same Day
7	2014	East	52	Second Class
8	2014	East	161	Standard Class
9	2014	South	19	First Class
10	2014	South	7	Same Day





Regarding the order quantities by regions, all regions looked almost same situation. The west region people seemed to use second and first class ship mode a little bit more than other regions.

--Check customers quantity by state and region

SELECT region, state, COUNT(DISTINCT (customer_id)) AS customer_number

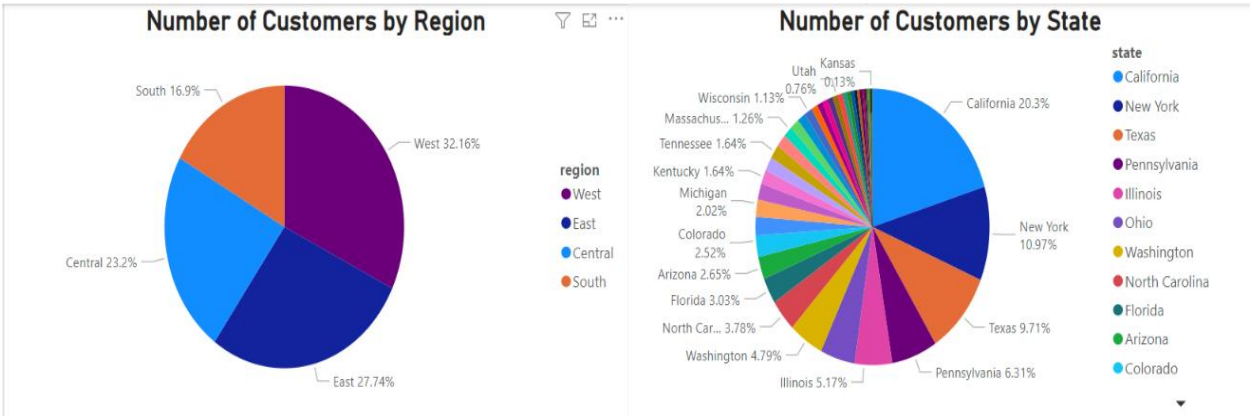
FROM customers

GROUP BY region, state, segment

ORDER BY region, state, segment;

--Check customers by segment by region;

	region character varying (20)	state character varying (30)	customer_number bigint
1	Central	Illinois	14
2	Central	Illinois	11
3	Central	Illinois	16
4	Central	Indiana	7
5	Central	Indiana	2
6	Central	Indiana	3
7	Central	Iowa	1
8	Central	Iowa	1
9	Central	Iowa	1
10	Central	Kansas	1

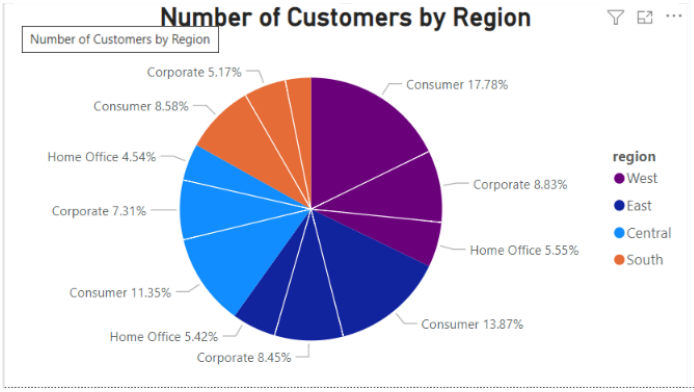


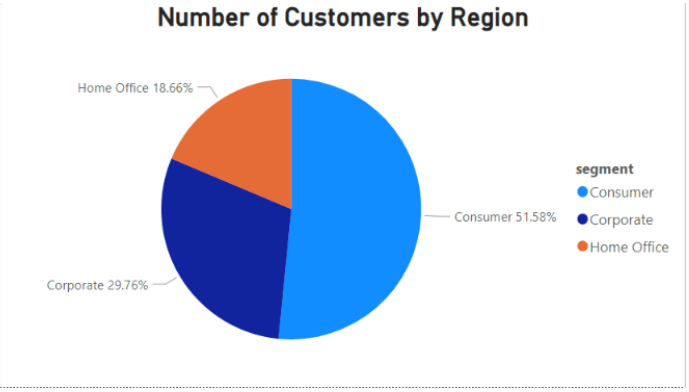
Currently, the biggest number of customers were from California. Based on the previous analysis, the superstore should pay less attention on this state, but pay more attention in New York state which is in east region.

--Check customers by segment by region

```
SELECT region, segment, COUNT(DISTINCT (customer_id)) AS
customer_number

FROM customers

GROUP BY region, segment

ORDER BY region,segment;
```

	region character varying (20)	segment character varying (30)	customer_number bigint
1	Central	Consumer	90
2	Central	Corporate	58
3	Central	Home Office	36
4	East	Consumer	110
5	East	Corporate	67
6	East	Home Office	43
7	South	Consumer	68
8	South	Corporate	41
9	South	Home Office	25
10	West	Consumer	141





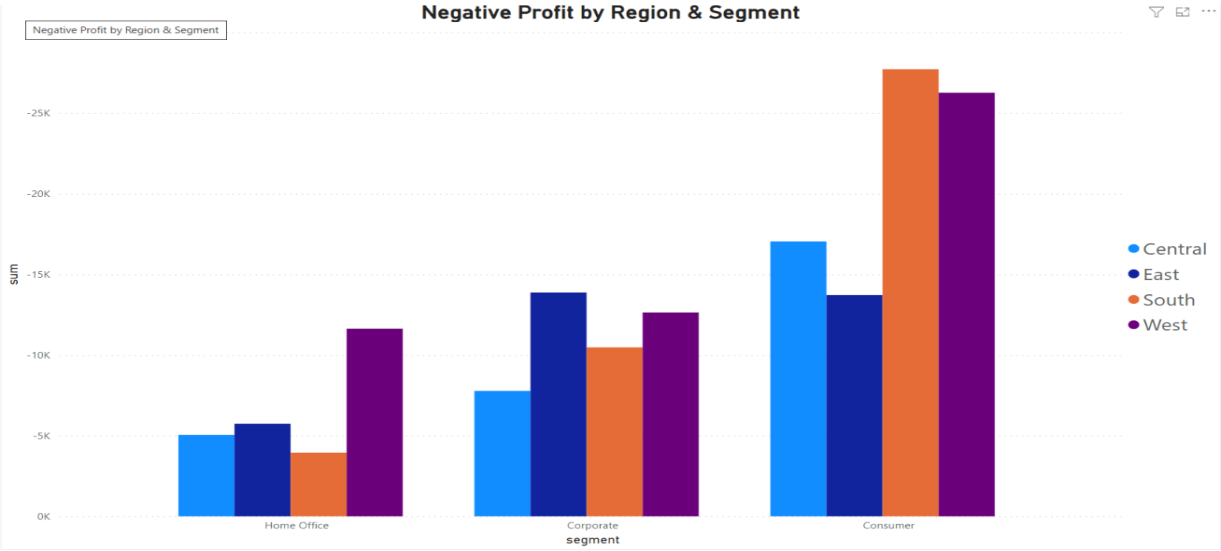
Regarding the segment customers, home office would be very potential since the number of customers from home office was only 18.66% in the end of 2017.

--Check negative profit product, orders

```
CREATE VIEW negative_profit_segment_region AS

        WITH  negative_profit AS
```

```
(SELECT *
 FROM orders
 WHERE profit<0)
SELECT product_name, SUM(profit), segment, region
FROM negative_profit N
INNER JOIN products P ON N.product_id = P.product_id
INNER JOIN customers C ON N.customer_id = C.customer_id
GROUP BY region, segment, product_name
```

	product_name character varying (200)	sum numeric	segment character varying (30)	region character varying (20)
1	Mitel MiVoice 5330e IP Phone	-192.4930	Consumer	Central
2	Avery Heavy-Duty EZD View Binder with Locking Rings	-19.5228	Consumer	West
3	Surelock Post Binders	-12.2240	Consumer	West
4	KI Adjustable-Height Table	-13.7568	Corporate	East
5	Logitech Media Keyboard K200	-1.0497	Consumer	South
6	High-Back Leather Manager's Chair	-28.5978	Corporate	West
7	DXL Angle-View Binders with Locking Rings by Samsill	-3.8550	Corporate	South
8	Heavy-Duty E-Z-D Binders	-10.0372	Consumer	South
9	Eldon Image Series Desk Accessories, Burgundy	-14.9739	Consumer	West
10	Deflect-o EconoMat Studded, No Bevel Mat for Low Pile Carpeting	-18.1808	Corporate	West



Finally, I checked all products that had negative profit by segments and region. All those products in the table should be handled very carefully since the company lost money on them. We could also find that the west and

south region sold a lot of those products, that should be one of the reasons why south and west region didn't have good performance on profit.

Conclusion

After reviewing all aspects of the dataset, here are some insights for Superstore:

- From 2014 to 2017, the Superstore's sales kept increasing but the profit almost kept the same level.
- The central and east region were doing well for both sales and profit, but the west and south region didn't have good profit increasement. The west region even decreased the profit in 2017.
- The corporate customers had profit dropped in 2017, but home office and consumer customers were growing in both sales and profit.
- Most customers chose standard ship mode in that period, and at the same time, first and second ship mode was growing slowly.
- The superstore did well on offering discounts in east and central regions, but not in west and south regions.
- In the period of 2014 to 2017, customers were almost even from 4 regions, but half of them were consumers.

Here are some recommendations:

1. Superstore should stop offering discounts for all the products which had negative profits in all regions.
2. In west and south regions, adjust the marketing strategy since the profit were dropping though the sales increased. The company might do more marketing campaigns other than offering discounts.
3. The company should put more efforts on promoting the products in the sub categories or categories with top profits.
4. Superstore needs to put more resources on developing more customers in home office and consumer segments.
5. Superstore could work with the shipping company to get better deal for the first and second class ship mode. Then Superstore could offer the customers better shipping package to improve the customers' satisfaction level. At the same time, Superstore would expect to attract more sales for more profits.