Superstore Sales Analysis using SQL, Power BI, and Python

Prepared by: Alina Khan

Date: 14-09-2025

Tools Used: SQL, Power BI, Python (Pandas, Matplotlib, Statsmodels)

Table of Contents

1. Project Introduction

This project analyzes Superstore sales data (9994 rows) to uncover meaningful insights for business decision-making. The dataset includes information on sales, profit, discount, region, category, customers, and products. The objective is to find valuable insights that can improve profitability, customer retention, and sales strategy.

2. Data Overview

Dataset contains 9994 rows and 21 columns such as:

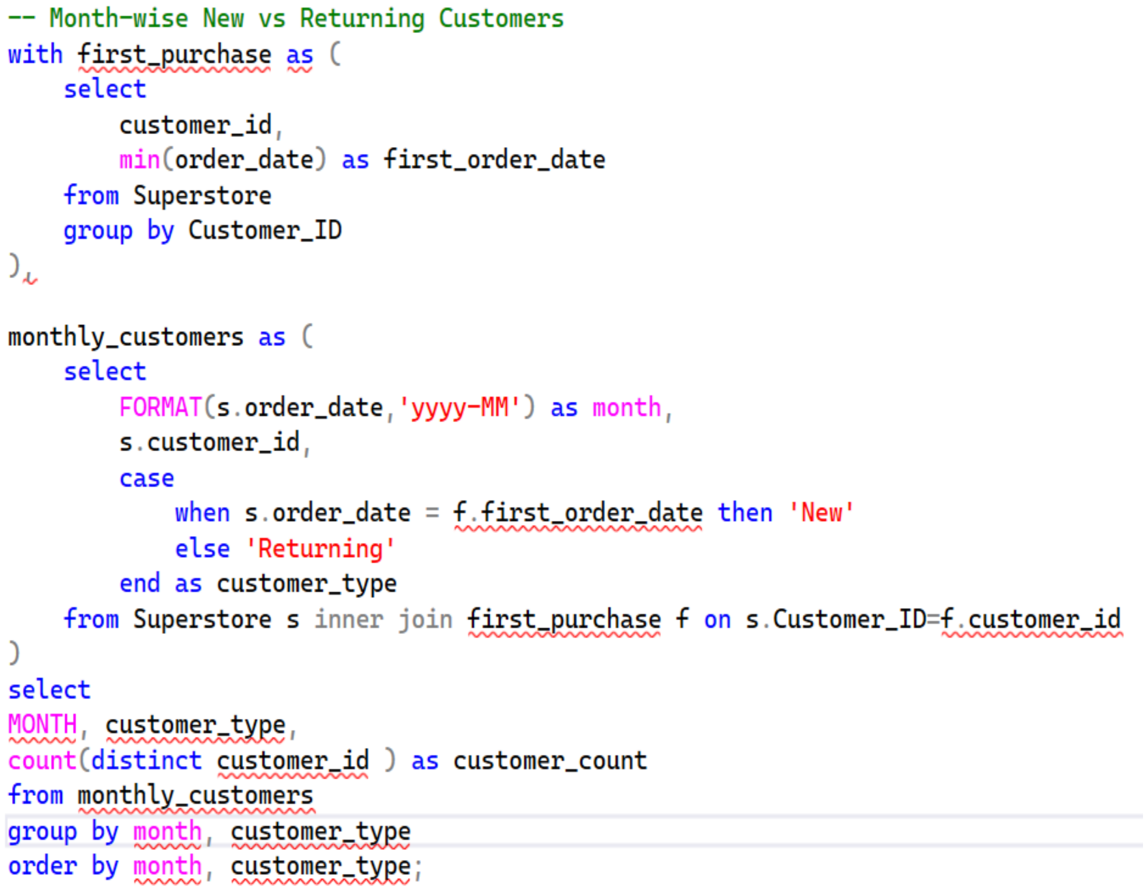
- Order Date, Ship Date, Customer ID, Region, Category, Sub-Category, Sales, Profit, Discount, Quantity.

Insert sample data screenshot here.

3. SQL Analysis & Insights

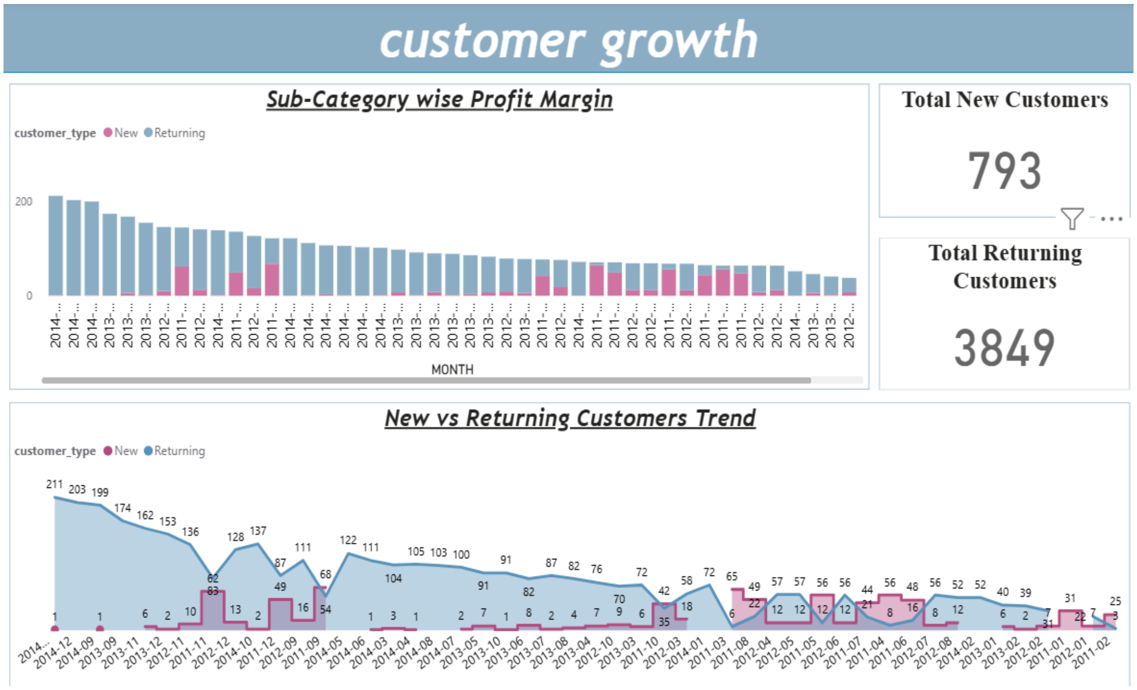
Below queries were written in SQL Server to analyze the dataset:

3.1 New vs Returning Customers



```
-- Month-wise New vs Returning Customers
with first_purchase as (
    select
        customer_id,
        min(order_date) as first_order_date
    from Superstore
    group by Customer_ID
),

monthly_customers as (
    select
        FORMAT(s.order_date,'yyyy-MM') as month,
        s.customer_id,
        case
            when s.order_date = f.first_order_date then 'New'
            else 'Returning'
        end as customer_type
    from Superstore s inner join first_purchase f on s.Customer_ID=f.customer_id
)
select
MONTH, customer_type,
count(distinct customer_id ) as customer_count
from monthly_customers
group by month, customer_type
order by month, customer_type;
```



customer growth

Sub-Category wise Profit Margin

Total New Customers

793

Total Returning Customers

3849

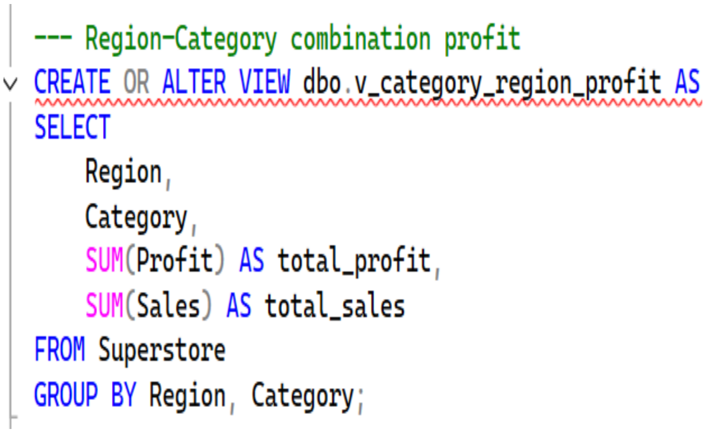
New vs Returning Customers Trend

Insights:
Looking at the trend, I noticed that the number of **new customers was higher in the beginning**, but gradually **returning customers started dominating** month by month. From late 2011 onwards, returning customers consistently outnumber new ones, which clearly shows that once customers purchase, they tend to come back.

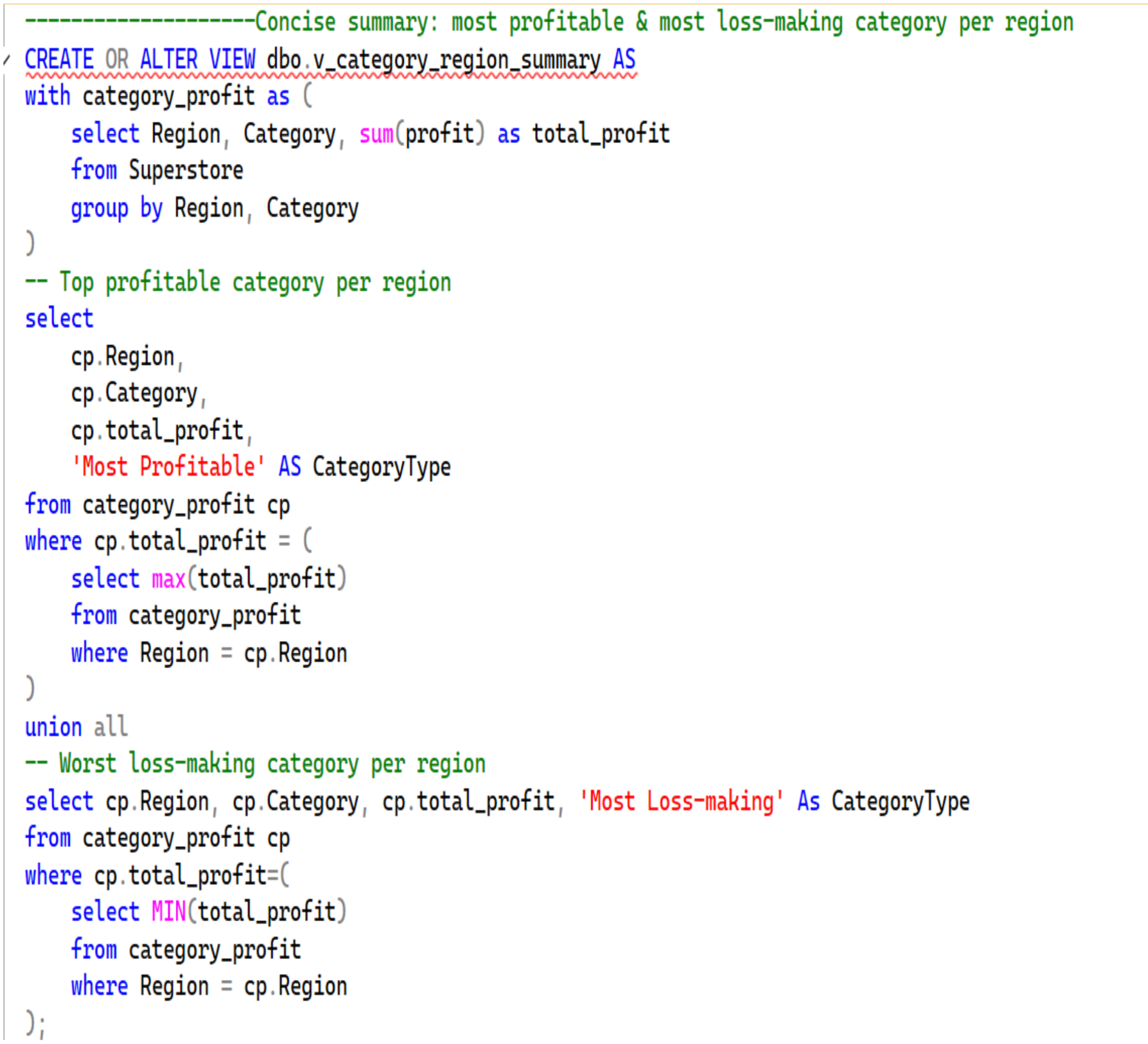
Another observation is that **new customer acquisition slows down after 2012**, but retention becomes stronger. For example, in 2013–2014, even with very few new customers joining, the business still maintained a good customer base because of loyal, repeat buyers.

Overall, the cards also confirm this — the **total number of returning customers is significantly higher than new customers**, highlighting that customer retention is one of the strongest aspects of this business.

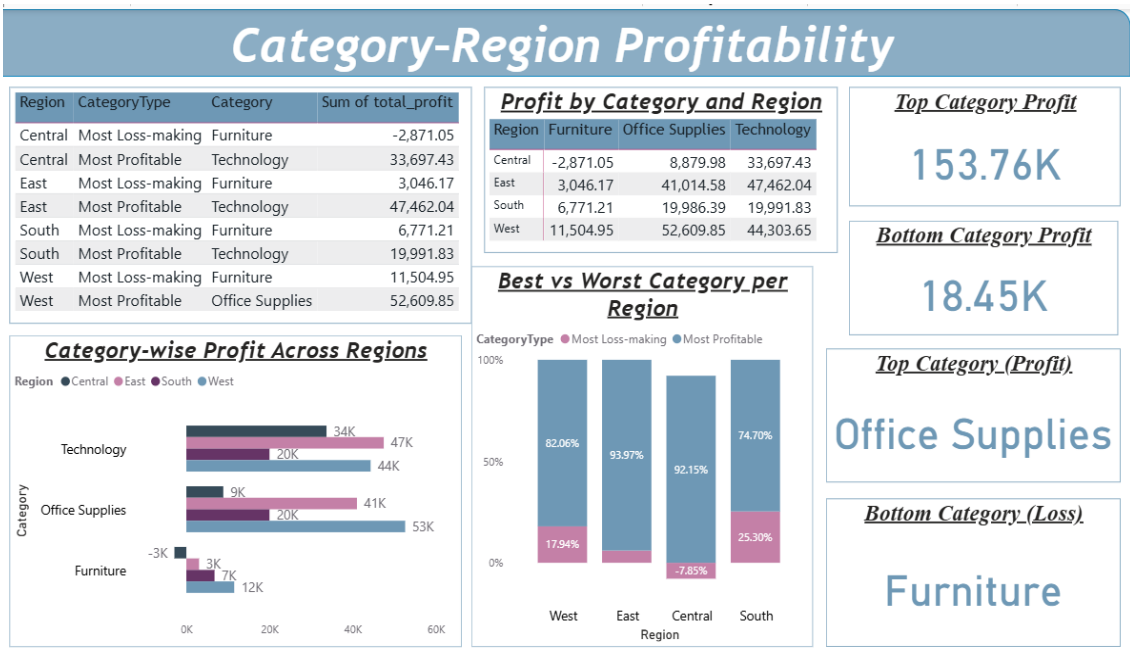
3.2 Category–Region Profitability



```
--- Region-Category combination profit
CREATE OR ALTER VIEW dbo.v_category_region_profit AS
SELECT
    Region,
    Category,
    SUM(Profit) AS total_profit,
    SUM(Sales) AS total_sales
FROM Superstore
GROUP BY Region, Category;
```



```
-------------------Concise summary: most profitable & most loss-making category per region
CREATE OR ALTER VIEW dbo.v_category_region_summary AS
with category_profit as (
    select Region, Category, sum(profit) as total_profit
    from Superstore
    group by Region, Category
)
-- Top profitable category per region
select
    cp.Region,
    cp.Category,
    cp.total_profit,
    'Most Profitable' AS CategoryType
from category_profit cp
where cp.total_profit = (
    select max(total_profit)
    from category_profit
    where Region = cp.Region
)
union all
-- Worst loss-making category per region
select cp.Region, cp.Category, cp.total_profit, 'Most Loss-making' As CategoryType
from category_profit cp
where cp.total_profit=(
    select MIN(total_profit)
    from category_profit
    where Region = cp.Region
);
```



Category-Region Profitability

Region	CategoryType	Category	Sum of total_profit
Central	Most Loss-making	Furniture	-2,871.05
Central	Most Profitable	Technology	33,697.43
East	Most Loss-making	Furniture	3,046.17
East	Most Profitable	Technology	47,462.04
South	Most Loss-making	Furniture	6,771.21
South	Most Profitable	Technology	19,991.83
West	Most Loss-making	Furniture	11,504.95
West	Most Profitable	Office Supplies	52,609.85

Profit by Category and Region

Region	Furniture	Office Supplies	Technology
Central	-2,871.05	8,879.98	33,697.43
East	3,046.17	41,014.58	47,462.04
South	6,771.21	19,986.39	19,991.83
West	11,504.95	52,609.85	44,303.65

Top Category Profit

153.76K

Bottom Category Profit

18.45K

Top Category (Profit)

Office Supplies

Bottom Category (Loss)

Furniture

Category-wise Profit Across Regions

Best vs Worst Category per Region

Insights:

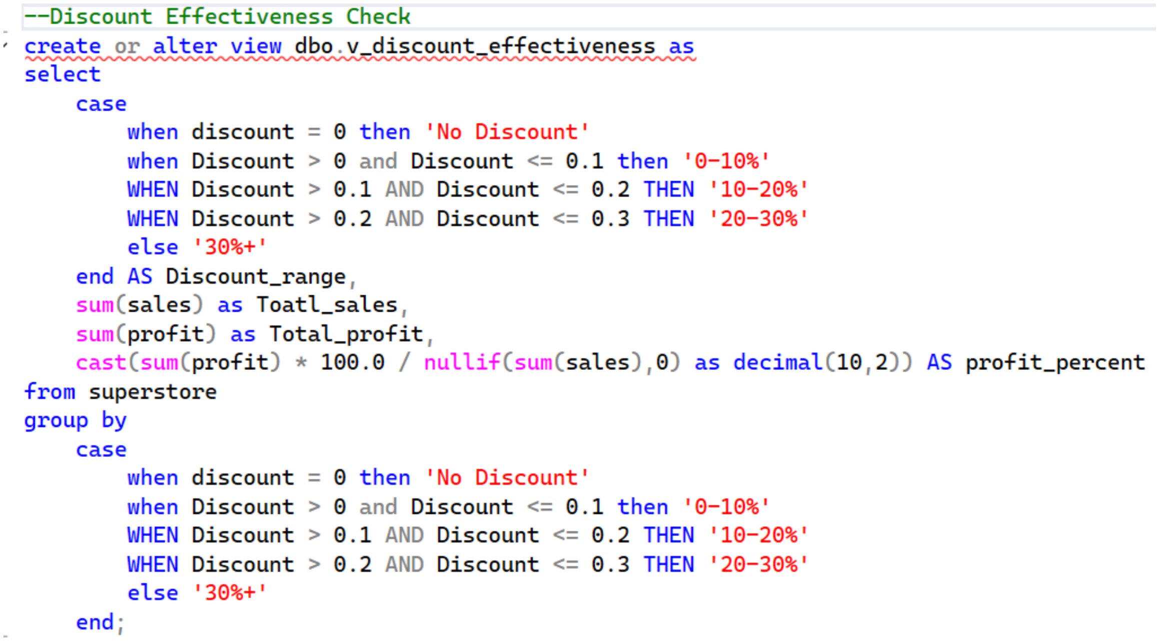
When I compared profit across regions and categories, I found that **Technology consistently appears as the most profitable category** in every region. For example, in the West, Technology generated the highest profit of ~44K, and similarly, in the East, it crossed ~47K profit. This indicates that Technology is a strong growth driver across the business.

On the other hand, **Furniture seems to be the weak spot**. In every region, Furniture shows the lowest profit, and in some cases, even losses. For example, in the Central region, Furniture recorded a **loss of ~2.8K**, making it the worst-performing segment there.

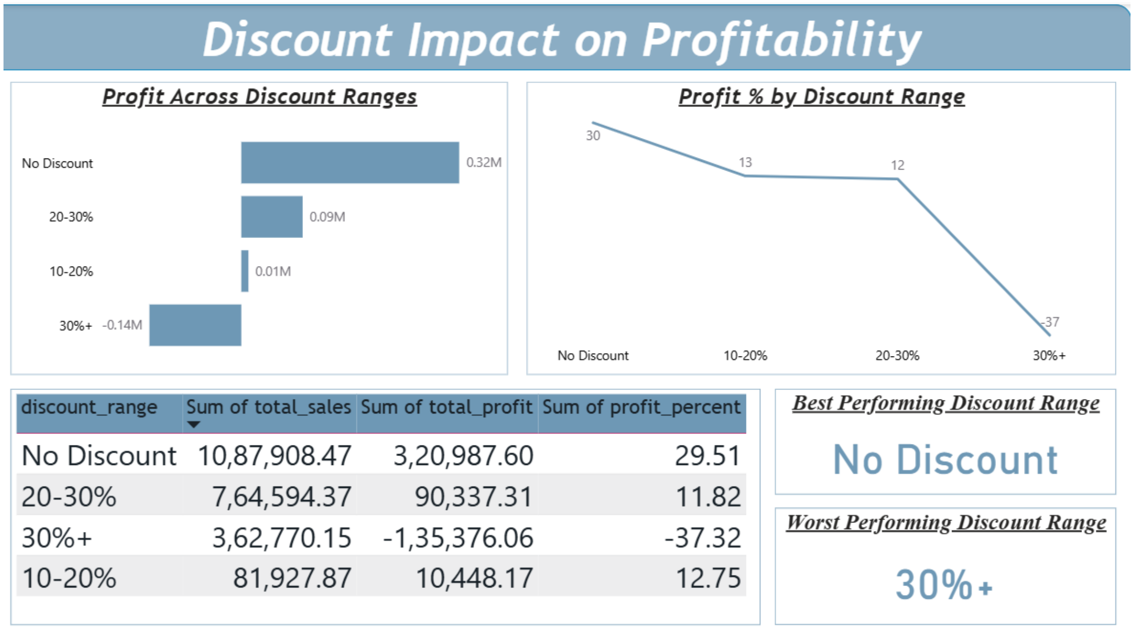
Office Supplies shows a mixed performance — while it is the most profitable in the West (~52K), in other regions its profit margins are not as strong compared to Technology.

Overall, the insight here is very clear: **Technology drives profitability across all regions**, while **Furniture needs immediate attention**, as it is consistently underperforming and dragging down overall margins.

3.3 Discount Effectiveness



```
--Discount Effectiveness Check
create or alter view dbo.v_discount_effectiveness as
select
    case
        when discount = 0 then 'No Discount'
        when Discount > 0 and Discount <= 0.1 then '0-10%'
        WHEN Discount > 0.1 AND Discount <= 0.2 THEN '10-20%'
        WHEN Discount > 0.2 AND Discount <= 0.3 THEN '20-30%'
        else '30%+'
    end AS Discount_range,
    sum(sales) as Toatl_sales,
    sum(profit) as Total_profit,
    cast(sum(profit) * 100.0 / nullif(sum(sales),0) as decimal(10,2)) AS profit_percent
from superstore
group by
    case
        when discount = 0 then 'No Discount'
        when Discount > 0 and Discount <= 0.1 then '0-10%'
        WHEN Discount > 0.1 AND Discount <= 0.2 THEN '10-20%'
        WHEN Discount > 0.2 AND Discount <= 0.3 THEN '20-30%'
        else '30%+'
    end;
```



Discount Impact on Profitability

Profit Across Discount Ranges

Profit % by Discount Range

discount_range	Sum of total_sales	Sum of total_profit	Sum of profit_percent
No Discount	10,87,908.47	3,20,987.60	29.51
20-30%	7,64,594.37	90,337.31	11.82
30%+	3,62,770.15	-1,35,376.06	-37.32
10-20%	81,927.87	10,448.17	12.75

Best Performing Discount Range

No Discount

Worst Performing Discount Range

30%+

Insights:

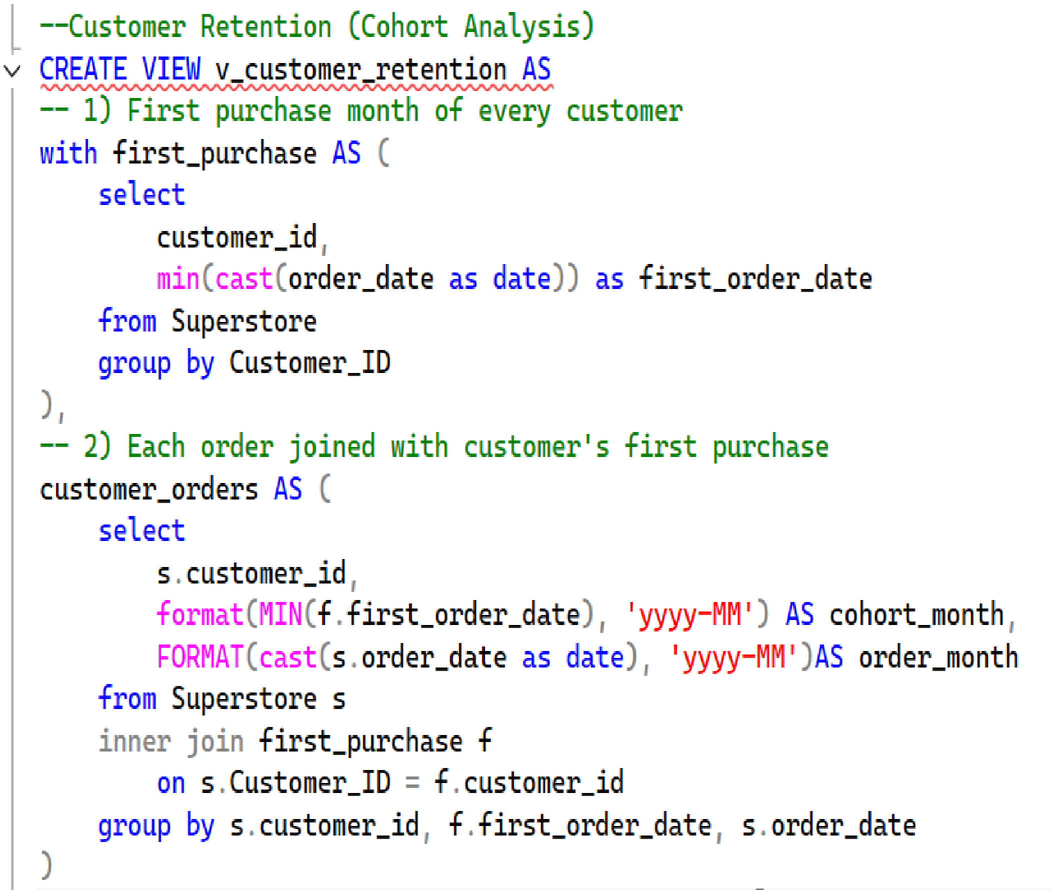
The analysis clearly shows that **“No Discount” orders deliver the highest profitability**. With over **1M in sales and ~321K profit**, the profit margin here is almost **29.5%**, making it the most effective range for the business.

Discounts in the **10–20%** and **20–30%** range still generate some profit, but the margins drop significantly to around **12–13%**. This suggests that while moderate discounts can attract sales, they reduce the overall profitability.

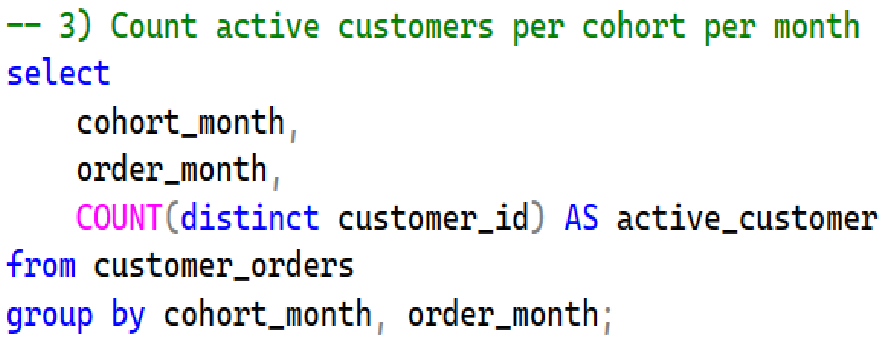
The biggest concern is the **30%+ discount range**. Even though it generated the second-highest sales (~363K), it ended up with a **loss of ~135K**, giving a **negative margin of –37%**. This means that heavy discounting is actually destroying value rather than driving growth.

Overall, the insight is straightforward: **No Discount and low discount ranges are profitable, but aggressive discounting (30%+) severely hurts the business**. If discounts are to be offered, they should be kept under 20% to maintain profitability.

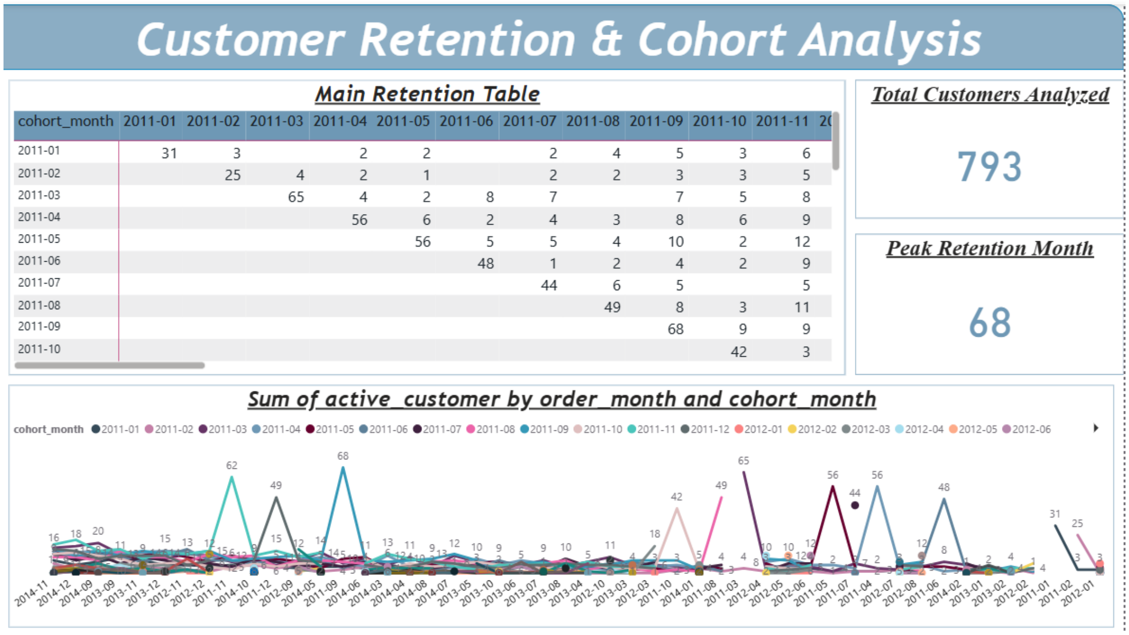
3.4 Customer Retention (Cohort Analysis)



```
--Customer Retention (Cohort Analysis)
CREATE VIEW v_customer_retention AS
-- 1) First purchase month of every customer
with first_purchase AS (
    select
        customer_id,
        min(cast(order_date as date)) as first_order_date
    from Superstore
    group by Customer_ID
),
-- 2) Each order joined with customer's first purchase
customer_orders AS (
    select
        s.customer_id,
        format(MIN(f.first_order_date), 'yyyy-MM') AS cohort_month,
        FORMAT(cast(s.order_date as date), 'yyyy-MM')AS order_month
    from Superstore s
    inner join first_purchase f
        on s.Customer_ID = f.customer_id
    group by s.customer_id, f.first_order_date, s.order_date
)
```



```
-- 3) Count active customers per cohort per month
select
    cohort_month,
    order_month,
    COUNT(distinct customer_id) AS active_customer
from customer_orders
group by cohort_month, order_month;
```



Customer Retention & Cohort Analysis

Main Retention Table

cohort_month	2011-01	2011-02	2011-03	2011-04	2011-05	2011-06	2011-07	2011-08	2011-09	2011-10	2011-11	20
2011-01	31	3		2	2		2	4	5	3	6	
2011-02		25	4	2	1		2	2	3	3	5	
2011-03			65	4	2	8	7		7	5	8	
2011-04				56	6	2	4	3	8	6	9	
2011-05					56	5	5	4	10	2	12	
2011-06						48	1	2	4	2	9	
2011-07							44	6	5		5	
2011-08								49	8	3	11	
2011-09									68	9	9	
2011-10										42	3	

Total Customers Analyzed

793

Peak Retention Month

68

Sum of active_customer by order_month and cohort_month

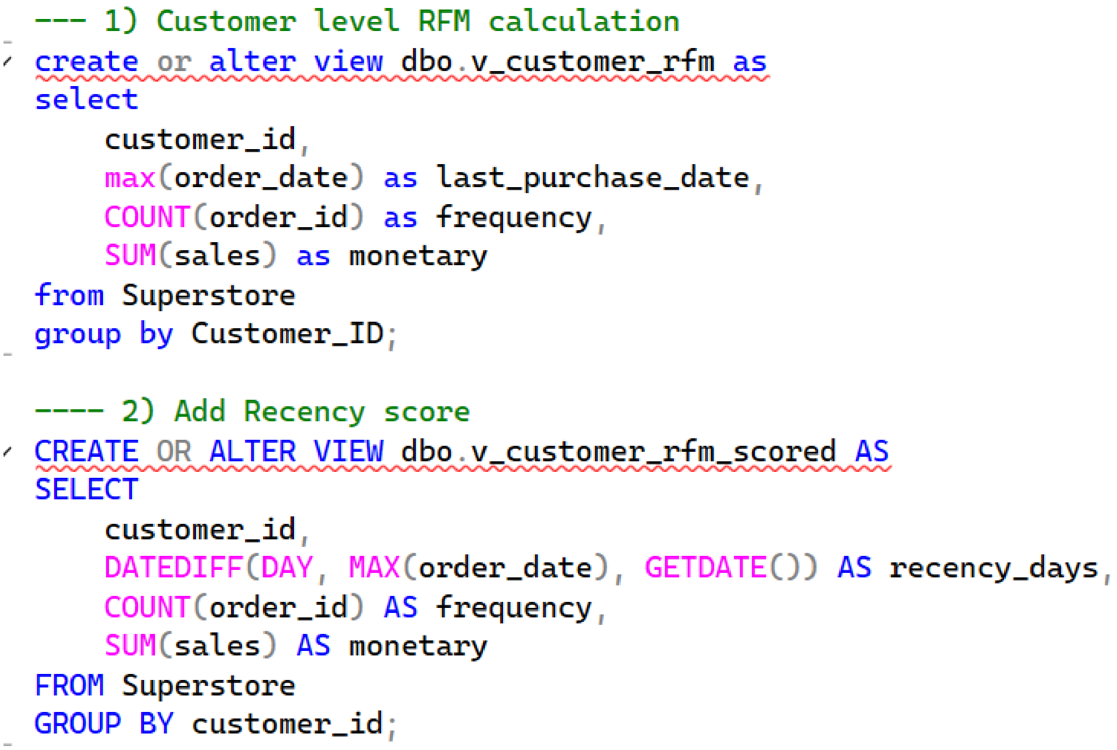
Insights:

The cohort analysis clearly highlights how customer retention behaves over time. Most customers tend to make their first purchase and a significant portion drop off in the following months. However, we see that a small but consistent share of customers continues to purchase regularly, creating a loyal base.

For example, customers acquired in early months like Jan–Mar 2011 show a steep decline after the first month, but those who stayed engaged continued purchasing for multiple months ahead. This trend repeats across cohorts, suggesting that while acquisition is effective, retention strategies need improvement to keep customers active beyond their initial purchase period.

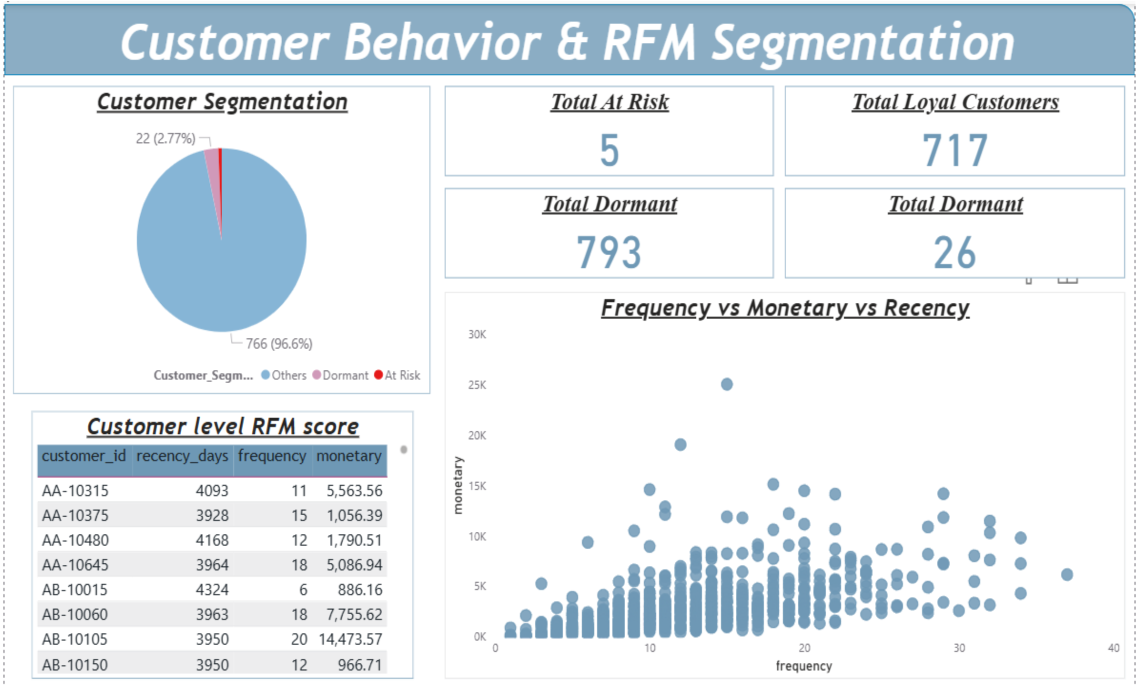
Overall, the analysis shows that **retention rates are strong in the short term (first 1–2 months)** but weaken considerably afterwards. Building engagement strategies like personalized offers, loyalty programs, or targeted campaigns can help extend the customer lifecycle.

3.5 RFM Segmentation



```sql
--- 1) Customer level RFM calculation
create or alter view dbo.v_customer_rfm as
select
    customer_id,
    max(order_date) as last_purchase_date,
    COUNT(order_id) as frequency,
    SUM(sales) as monetary
from Superstore
group by Customer_ID;

---- 2) Add Recency score
CREATE OR ALTER VIEW dbo.v_customer_rfm_scored AS
SELECT
    customer_id,
    DATEDIFF(DAY, MAX(order_date), GETDATE()) AS recency_days,
    COUNT(order_id) AS frequency,
    SUM(sales) AS monetary
FROM Superstore
GROUP BY customer_id;
```



Customer Behavior & RFM Segmentation

Customer Segmentation

Total At Risk
5

Total Loyal Customers
717

Total Dormant
793

Total Dormant
26

Customer level RFM score

customer_id	recency_days	frequency	monetary
AA-10315	4093	11	5,563.56
AA-10375	3928	15	1,056.39
AA-10480	4168	12	1,790.51
AA-10645	3964	18	5,086.94
AB-10015	4324	6	886.16
AB-10060	3963	18	7,755.62
AB-10105	3950	20	14,473.57
AB-10150	3950	12	966.71

Frequency vs Monetary vs Recency

Insights:

- Customers with **recent purchases, higher frequency, and higher monetary value** represent the **most valuable group**. These customers should be nurtured through loyalty programs, exclusive offers, and early access to new products.

- A segment of customers shows **moderate frequency and monetary value** but still purchases relatively often. These are **potential loyalists**, and with the right engagement (personalized discounts or product bundles), they can be converted into long-term loyal customers.

- There is also a group of **high spenders with low recency** — they used to spend significantly but haven't purchased recently. These are **"At Risk" customers**, and they require re-engagement campaigns such as win-back emails or targeted promotions.

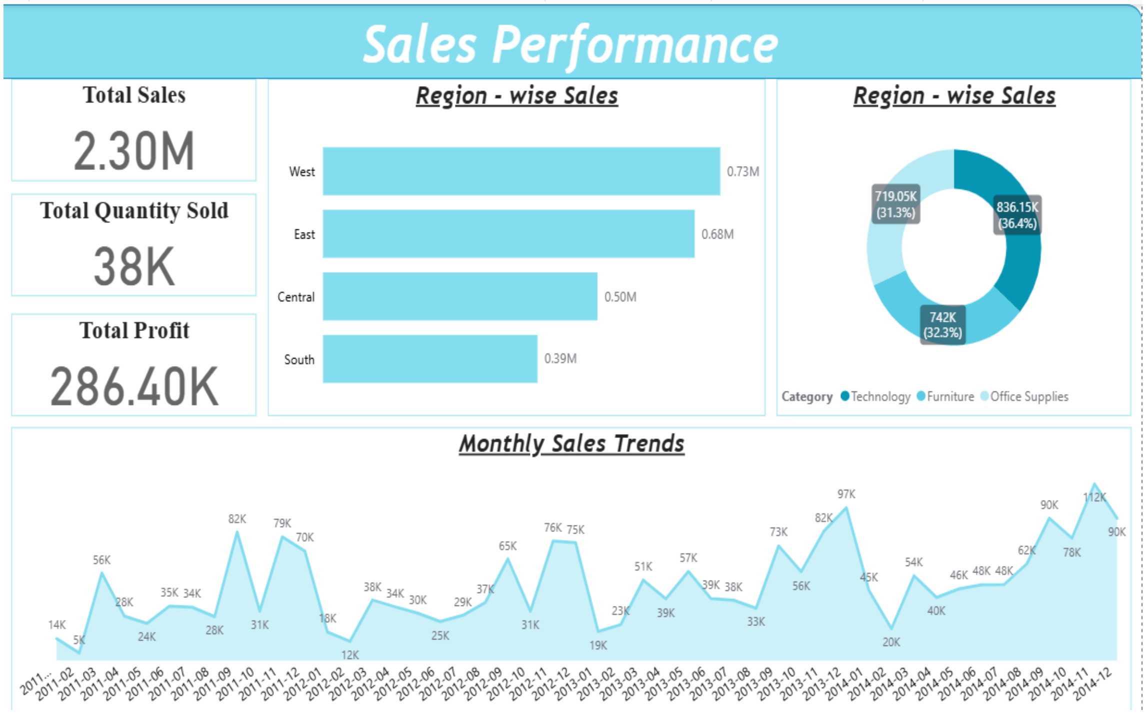
- Customers with **low frequency and low spend** but recent activity, fall into the **new customers** category. This group requires onboarding, education, and incentives to increase their purchasing frequency.

- Finally, customers with **low recency, frequency, and monetary values** represent the **lost customers segment**. These accounts provide minimal value currently, but selective reactivation offers may still bring some of them back.

4. Power BI Dashboards

Screenshots of dashboards built in Power BI:

- Sales Overview Dashboard



Sales Performance

Total Sales

2.30M

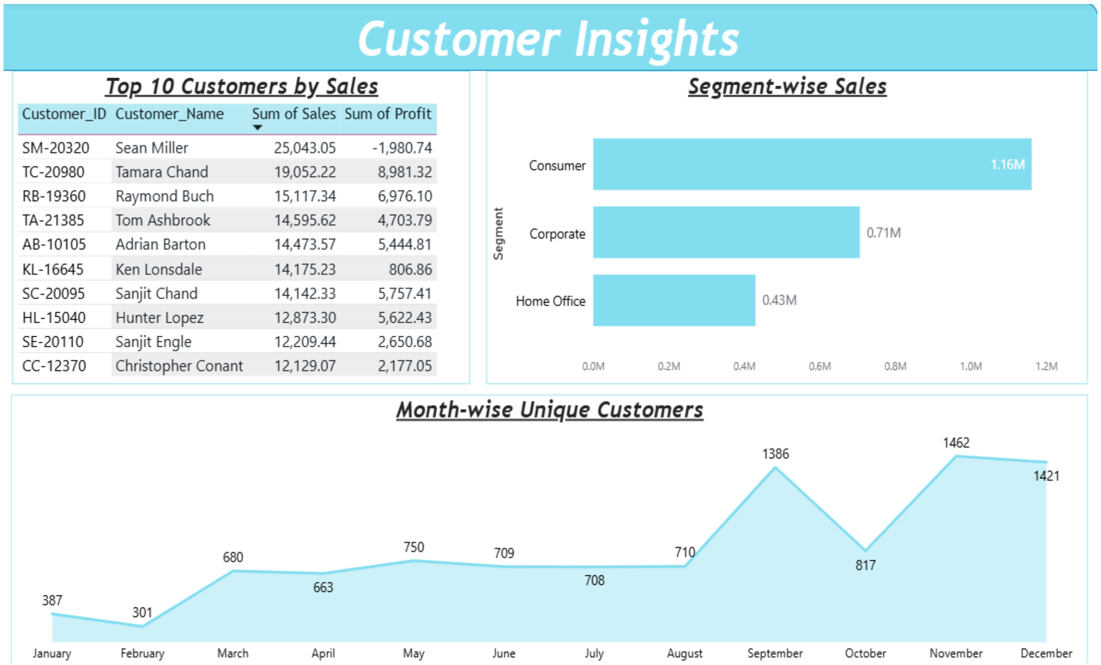
Total Quantity Sold

38K

Total Profit

286.40K

- Customer Analysis Dashboard

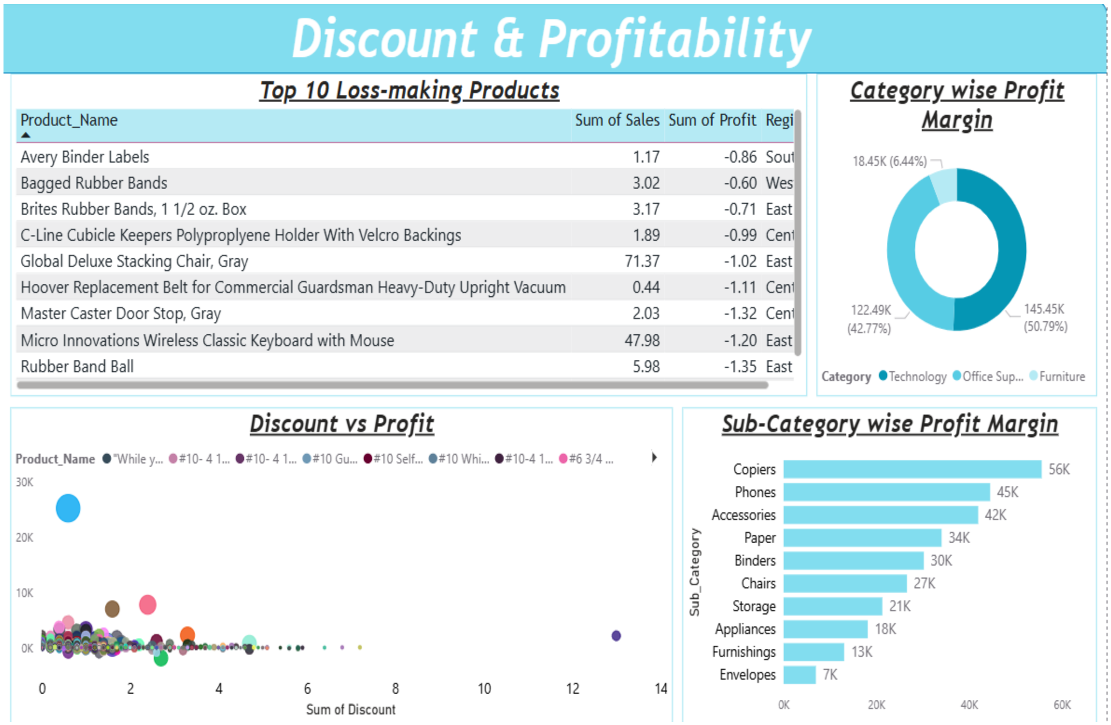


Customer Insights

Top 10 Customers by Sales

Customer_ID	Customer_Name	Sum of Sales	Sum of Profit
SM-20320	Sean Miller	25,043.05	-1,980.74
TC-20980	Tamara Chand	19,052.22	8,981.32
RB-19360	Raymond Buch	15,117.34	6,976.10
TA-21385	Tom Ashbrook	14,595.62	4,703.79
AB-10105	Adrian Barton	14,473.57	5,444.81
KL-16645	Ken Lonsdale	14,175.23	806.86
SC-20095	Sanjit Chand	14,142.33	5,757.41
HL-15040	Hunter Lopez	12,873.30	5,622.43
SE-20110	Sanjit Engle	12,209.44	2,650.68
CC-12370	Christopher Conant	12,129.07	2,177.05

- Discount Effectiveness Dashboard



Discount & Profitability

Top 10 Loss-making Products

Product_Name	Sum of Sales	Sum of Profit	Regi
Avery Binder Labels	1.17	-0.86	Sout
Bagged Rubber Bands	3.02	-0.60	Wes
Brites Rubber Bands, 1 1/2 oz. Box	3.17	-0.71	East
C-Line Cubicle Keepers Polyproplyene Holder With Velcro Backings	1.89	-0.99	Cent
Global Deluxe Stacking Chair, Gray	71.37	-1.02	East
Hoover Replacement Belt for Commercial Guardsman Heavy-Duty Upright Vacuum	0.44	-1.11	Cent
Master Caster Door Stop, Gray	2.03	-1.32	Cent
Micro Innovations Wireless Classic Keyboard with Mouse	47.98	-1.20	East
Rubber Band Ball	5.98	-1.35	East

Category wise Profit Margin

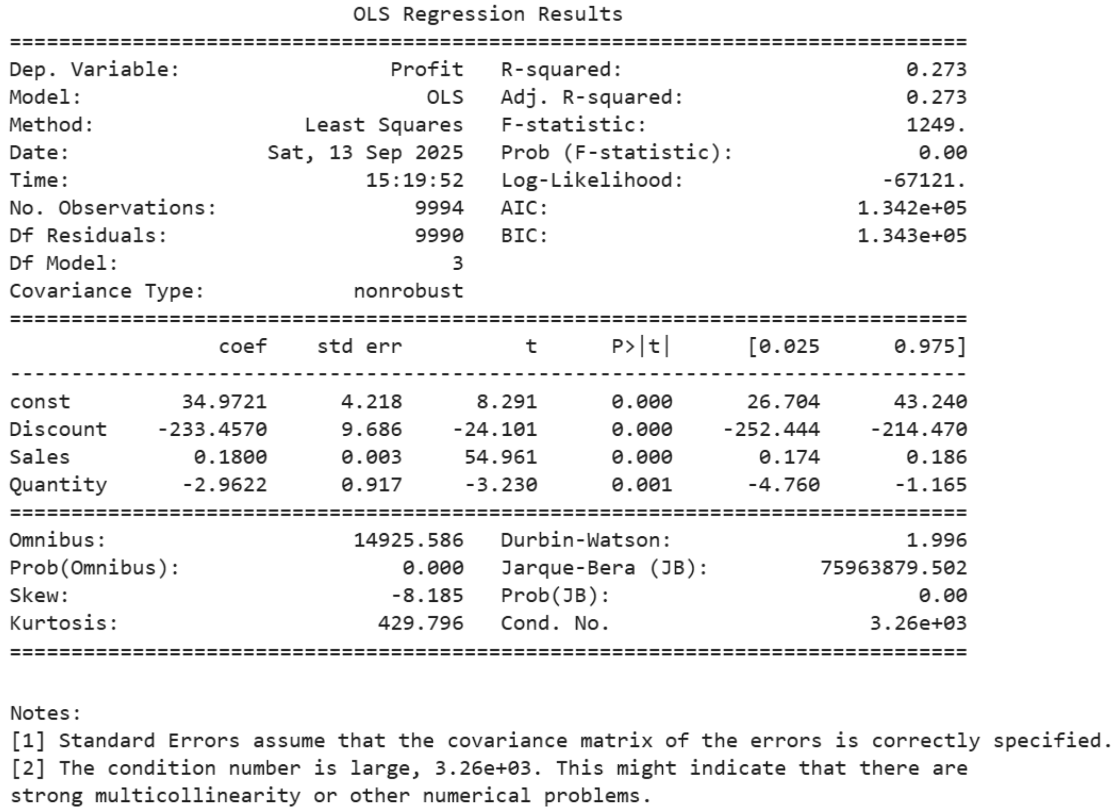
Discount vs Profit

Product_Name ● "While y... ● #10- 4 1... ● #10- 4 1... ● #10 Gu... ● #10 Self... ● #10 Whi... ● #10-4 1... ● #6 3/4 ...
30K
20K
10K
0K
0
2
4
6
8
10
12
14
Sum of Discount

Sub-Category wise Profit Margin

5. Python Analysis (Regression & Forecasting)

5.1 Regression Analysis



```
                            OLS Regression Results
==============================================================================
Dep. Variable:                 Profit   R-squared:                       0.273
Model:                            OLS   Adj. R-squared:                  0.273
Method:                 Least Squares   F-statistic:                     1249.
Date:                Sat, 13 Sep 2025   Prob (F-statistic):               0.00
Time:                        15:19:52   Log-Likelihood:                -67121.
No. Observations:                9994   AIC:                         1.342e+05
Df Residuals:                    9990   BIC:                         1.343e+05
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         34.9721      4.218      8.291      0.000      26.704      43.240
Discount    -233.4570      9.686    -24.101      0.000    -252.444    -214.470
Sales          0.1800      0.003     54.961      0.000       0.174       0.186
Quantity      -2.9622      0.917     -3.230      0.001      -4.760      -1.165
==============================================================================
Omnibus:                    14925.586   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         75963879.502
Skew:                          -8.185   Prob(JB):                         0.00
Kurtosis:                     429.796   Cond. No.                     3.26e+03
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 3.26e+03. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Insights:

- The model explains **27.3% of the variation in Profit (R-squared = 0.273)**, meaning while the selected factors (Discount, Sales, and Quantity) do impact profit, other external factors may also play a significant role.

- **Discount has a strong negative effect on Profit** (Coefficient = -233.45, p < 0.001). This indicates that as discounts increase, profitability decreases significantly. Offering high discounts directly erodes profit margins.

- **Sales has a positive relationship with Profit** (Coefficient = 0.18, p < 0.001). Higher sales generally lead to higher profits, validating sales growth as a key driver of profitability.

- **Quantity has a slight negative impact on Profit** (Coefficient = -2.96, p = 0.001). Selling more units at lower margins may not always translate into higher profitability.

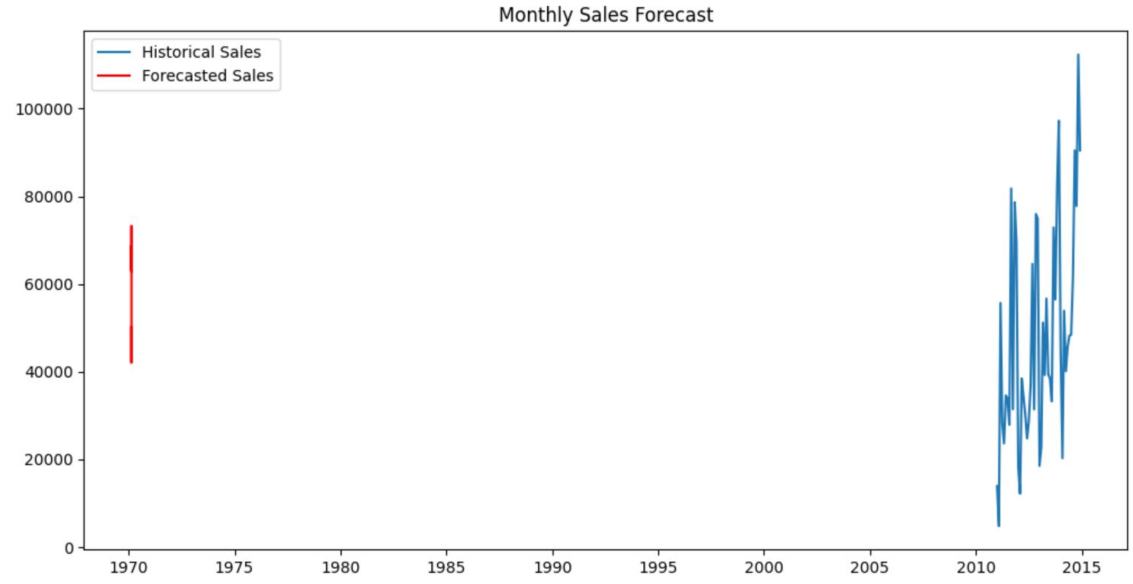
- The **constant value (34.97)** represents the baseline profit when all independent variables are zero.

- The high statistical significance of all predictors ($p < 0.05$) confirms that Discount, Sales, and Quantity are meaningful factors in determining Profit.

Key takeaway:
The regression clearly highlights that **discounting strategies need to be carefully controlled**, as they directly harm profitability, while driving higher sales volumes with better margins can significantly improve profits.

5.2 Forecasting

Screenshot of sales forecasting chart (next 6 months).



Insights:

- The sales forecasting model provides an outlook for the next six months.
- The forecast indicates that monthly sales are expected to remain within the **₹45,000 – ₹75,000 range**, showing a moderate but consistent sales trend.
- Compared to historical fluctuations, the forecast suggests **stability rather than sharp peaks or declines**, which can help in better inventory and resource planning.
- This projection allows stakeholders to anticipate demand more accurately and align **supply chain, marketing campaigns, and budget allocation** with the expected sales performance.
- Overall, the forecast reflects **steady business momentum** in the upcoming months, with no significant risk of sales downturn.

6. Key Insights & Business Recommendations

- **Sales Growth:** Overall sales have shown a consistent upward trend over time, indicating strong business momentum.
- **Regional Performance:** Certain regions and states contribute disproportionately to total revenue, while underperforming regions highlight potential areas for expansion.
- **Category & Sub-Category Analysis:** Technology and Office Supplies drive significant revenue, while categories like Furniture show lower profit margins, requiring careful inventory management.
- **Profitability:** Although sales are increasing, profit margins vary widely across products and regions, indicating the need for better pricing and cost optimization.
- **Customer Segment Analysis:** Corporate and Consumer segments generate higher revenue, while Home Office lags behind, suggesting opportunities to boost engagement in this segment.
- **Forecasting:** The next six months' forecast shows stable sales within the ₹45,000–₹75,000 range, enabling stakeholders to plan budgets and operations with greater confidence.

7. Executive Summary

This report analyzes sales, profit, and customer trends using SQL, Power BI, and forecasting models. Key findings show that while sales are growing, profitability is highly impacted by discounting strategies and regional variations. The six-month sales forecast provides visibility into upcoming demand, enabling data-driven decision-making. Strategic actions around pricing, product focus, and regional optimization are recommended to drive sustainable growth and improved profitability.

8. Conclusion

This analysis provided a comprehensive view of sales, profit, and customer behavior across multiple dimensions such as region, category, and segment. The insights derived from SQL queries, Power BI dashboards, and Python forecasting models highlight both the strengths of the business and areas that require attention.

While overall sales growth remains strong, profit margins and regional disparities indicate the need for a more targeted strategy. The six-month sales forecast further equips stakeholders with data-driven visibility into future performance, supporting better operational and financial planning.

In conclusion, by acting on the recommendations outlined, the business can enhance profitability, optimize resources, and ensure sustainable growth in the coming months.