Superstore Sales Data Analysis Project

About Dataset

Context

Retail dataset of a global superstore for 4 years.

Content

Time series analysis deals with time series based data to extract patterns for predictions and other characteristics of the data. It uses a model for forecasting future values in a small time frame based on previous observations. It is widely used for non-stationary data, such as economic data, weather data, stock prices, and retail sales forecasting.

Key Steps : Data Preparation, Analysis & Reporting

1. Data Cleaning and Preparation

- Reviewed and standardized data types across all columns for consistency.
- Identified and resolved import issues in MySQL Workbench due to inconsistent encoding in the `Product Name` column (e.g., `\xa0` characters). This was successfully corrected using VS Code by replacing non-breaking spaces with standard whitespace.

2. Data Import and Transformation

- Successfully imported the cleaned CSV dataset into MySQL.
- Handled missing, null, and blank values to ensure data quality and integrity.
- Applied `STR_TO_DATE()` function to convert string-based date formats into standard date types.
- Modified date columns to appropriate SQL `DATE` datatype for time-based analysis.

3. Data Warehouse Creation

- Structured the cleaned and transformed dataset into a star-schema-based data warehouse model to enable efficient querying and reporting.

4. KPI Analysis and Metrics Derivation

- Performed comprehensive analysis to extract key performance indicators (KPIs) such as:
    - Total Sales
    - Average Delivery Time
    - Sales by Category, Region, and Segment
    - Year-over-Year Growth and Profitability Trends

5. Reporting and Visualization

- Developed dynamic, interactive dashboards in **Power BI**, including metrics and visual comparisons (e.g., current vs. last year sales).
- Published the reports to **Power BI Service** for stakeholder access and collaboration.

6. Business Insights and Recommendations

- Identified actionable insights to optimize profit margins.
- Proposed strategies to streamline inventory and enhance marketing effectiveness based on sales patterns and customer behavior trends.

🔢 Derived Metrics

📦 Sales & Revenue Metrics

- **Total Sales** over time (monthly, quarterly, yearly)

```
34 •  SELECT ROUND(SUM(Sales),2) AS Total_sales FROM sales_data;
35
36
37
38
39
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Total_sales
2252607.41

```
36      -- Total sales by Yearly, Quarterly, Monthly
37
38 •    SELECT
39          YEAR(`Order Date`) AS yearly,
40          QUARTER(`Order Date`) AS quarterly,
41          MONTH(`Order Date`) AS month_num,
42          DATE_FORMAT(`Order Date`, '%b') AS month_name,
43          ROUND(SUM(Sales), 2) AS total_sale
44      FROM sales_data
45      GROUP BY yearly, quarterly, month_num, month_name
46      ORDER BY yearly, quarterly, month_num;
47
```

yearly	quarterly	month_num	month_name	total_sale
2015	1	1	Jan	14205.71
2015	1	2	Feb	4519.89
2015	1	3	Mar	55205.8
2015	2	4	Apr	27906.85
2015	2	5	May	23644.3
2015	2	6	Jun	34322.94
2015	3	7	Jul	33781.54
2015	3	8	Aug	27117.54
2015	3	9	Sep	81623.53
2015	4	10	Oct	31453.39
2015	4	11	Nov	77907.66
2015	4	12	Dec	68167.06
2016	1	1	Jan	18066.96
2016	1	2	Feb	11951.41
2016	1	3	Mar	32339.32
2016	2	4	Apr	34154.47
2016	2	5	May	29959.53
2016	2	6	Jun	23599.37
2016	3	7	Jul	28608.26
2016	3	8	Aug	36818.34
2016	3	9	Sep	63133.61
2016	4	10	Oct	31011.74
2016	4	11	Nov	70129.3
2016	4	12	Dec	74543.6
2017	1	1	Jan	16870.18

Result 77

- **Sales by Product Category/Sub-Category**

```sql
WITH Category_wise AS(
SELECT
      category,
      `Sub-Category`,
      ROUND(SUM(Sales) OVER(partition by category,`Sub-Category` ),2) AS Category_total_sales,
      ROUND(sum(Sales) OVER(),2) AS OverallSales
      FROM sales_data)
      SELECT
        DISTINCT category,
      `Sub-Category`,
      Category_total_sales,
      OverallSales,
     CONCAT(ROUND(((Category_total_sales/ OverallSales)*100),2),'%') AS PerctOfOverallSales
      FROM Category_wise;
```

category	Sub-Category	Category_total_sales	OverallSales	PerctOfOverallSales
Furniture	Bookcases	109408.3	2252607.41	4.86%
Furniture	Chairs	322107.53	2252607.41	14.3%
Furniture	Furnishings	89212.02	2252607.41	3.96%
Furniture	Tables	202810.63	2252607.41	9%
Office Supplies	Appliances	104075.46	2252607.41	4.62%
Office Supplies	Art	26697.37	2252607.41	1.19%
Office Supplies	Binders	200028.79	2252607.41	8.88%
Office Supplies	Envelopes	16126.01	2252607.41	0.72%
Office Supplies	Fasteners	3001.96	2252607.41	0.13%
Office Supplies	Labels	12347.73	2252607.41	0.55%
Office Supplies	Paper	76736.1	2252607.41	3.41%
Office Supplies	Storage	217779.1	2252607.41	9.67%
Office Supplies	Supplies	46420.31	2252607.41	2.06%
Technology	Accessories	163881.69	2252607.41	7.28%
Technology	Copiers	146248.09	2252607.41	6.49%
Technology	Machines	189238.63	2252607.41	8.4%
Technology	Phones	326487.7	2252607.41	14.49%

- **Sales by Region/State/City**

```sql
WITH sales_by_region AS (
SELECT
    City,
    Region,
    State,
    ROUND(SUM(Sales), 2) AS Region_total_sales
FROM sales_data
GROUP BY City, Region, State
),
total_sales AS (
    SELECT ROUND(SUM(Sales), 2) AS OverallSales
    FROM sales_data
)
SELECT
    r.City,
    r.Region,
    r.State,
    r.Region_total_sales,
    t.OverallSales,
    CONCAT(ROUND((r.Region_total_sales / t.OverallSales) * 100, 2), '%') AS PerctOfOverallSales
FROM sales_by_region r
CROSS JOIN total_sales t
ORDER BY (r.Region_total_sales / t.OverallSales) DESC;
```

City	Region	State	Region_total_sales	OverallSales	PerctOfOverallSales
New York City	East	New York	252462.55	2252607.41	11.21%
Los Angeles	West	California	173420.18	2252607.41	7.7%
Seattle	West	Washington	116106.32	2252607.41	5.15%
San Francisco	West	California	109041.12	2252607.41	4.84%
Philadelphia	East	Pennsylvania	108841.75	2252607.41	4.83%
Houston	Central	Texas	63956.14	2252607.41	2.84%
Chicago	Central	Illinois	47820.13	2252607.41	2.12%
San Diego	West	California	47521.03	2252607.41	2.11%
Detroit	Central	Michigan	42446.94	2252607.41	1.88%
Jacksonville	South	Florida	39133.33	2252607.41	1.74%
San Antonio	Central	Texas	21843.53	2252607.41	0.97%
Newark	East	Delaware	20319.98	2252607.41	0.9%
Dallas	Central	Texas	20127.95	2252607.41	0.89%
Lafayette	Central	Indiana	19624.51	2252607.41	0.87%
Atlanta	South	Georgia	17197.84	2252607.41	0.76%
Minneapolis	Central	Minnesota	16870.54	2252607.41	0.75%
Springfield	South	Virginia	16628.53	2252607.41	0.74%
Providence	East	Rhode Island	15980.65	2252607.41	0.71%
Columbus	East	Ohio	15900.79	2252607.41	0.71%
Henderson	South	Kentucky	15661.01	2252607.41	0.7%
Jackson	Central	Michigan	15420.04	2252607.41	0.68%
Springfield	Central	Missouri	15051.38	2252607.41	0.67%
Lakewood	East	New Jersey	14843.1	2252607.41	0.66%
Columbus	South	Georgia	14801.37	2252607.41	0.66%
Arlington	South	Virginia	14366.76	2252607.41	0.64%
Charlotte	South	North Carolina	13450.99	2252607.41	0.6%
Burlington	South	North Carolina	12681.28	2252607.41	0.56%

- **Average Order Value (AOV)** = Total Sales / Number of Orders

```
94 •  SELECT
95        ROUND(SUM(Sales)/ COUNT(`Row ID`),2) AS Avg_Order_value
96        FROM sales_data;
97
98
99
```

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

Avg_Order_value
230.12

- **Top 10 Customer**

```
-- Top 10 Customer By Sales

SELECT
      `Customer ID`,
      `Customer Name`,
      Segment,
      City,
      State,
      Region,
      ROUND(SUM(Sales),2) AS Top10Customer
      FROM sales_data
      GROUP BY `Customer ID`,`Customer Name`,Segment, City, State, Region
      ORDER BY Top10Customer DESC LIMIT 10;
```

Customer ID	Customer Name	Segment	City	State	Region	Top10Customer
SM-20320	Sean Miller	Home Office	Jacksonville	Florida	South	23661.23
TC-20980	Tamara Chand	Corporate	Lafayette	Indiana	Central	18336.74
RB-19360	Raymond Buch	Consumer	Seattle	Washington	West	14052.48
TA-21385	Tom Ashbrook	Home Office	New York City	New York	East	13723.5
BM-11140	Becky Martin	Consumer	San Antonio	Texas	Central	10539.9
HL-15040	Hunter Lopez	Consumer	Newark	Delaware	East	10499.97
SC-20095	Sanjit Chand	Consumer	Minneapolis	Minnesota	Central	9900.19
AB-10105	Adrian Barton	Consumer	Detroit	Michigan	Central	9892.74
BS-11365	Bill Shonely	Corporate	Lakewood	New Jersey	East	9135.19
SE-20110	Sanjit Engle	Consumer	Arlington	Virginia	South	8805.04

📈 Operational Metrics

- **Shipping Time** = `Ship Date - Order Date`
- **Average Shipping Time by Ship Mode**
- **Number of Orders per Shipping Mode**

```sql
WITH ShippingTimes AS (
    SELECT
        `Order ID`,
        `Ship Mode`,
        DATEDIFF(`Ship Date`, `Order Date`) AS Shipping_time
    FROM sales_data
)
SELECT
    `Ship Mode`,
    count(`Order ID`) AS NoOfOrders,
    CONCAT(ROUND(AVG(Shipping_time), 0), " Days") AS AvgShippingTime
FROM ShippingTimes
GROUP BY `Ship Mode`
ORDER BY AvgShippingTime;
```

Ship Mode	NoOfOrders	AvgShippingTime
Same Day	538	0 Days
First Class	1501	2 Days
Second Class	1901	3 Days
Standard Class	5849	5 Days

🧍 Customer Metrics

- **Number of Unique Customers**

```sql
-- NUmber Of Total Unique Customers And Total Customers

SELECT
    COUNT(`Row ID`) AS Total_Order,
    COUNT(DISTINCT `Customer ID`) AS Unique_Customers
    FROM sales_data;
```

Total_Order	Unique_Customers
9789	793

Customer Segmentation Analysis (e.g., based on "Segment")

```sql
WITH SalesBySegment AS (
SELECT
      distinct Segment,
      COUNT(`Order ID`) OVER(partition by Segment) AS Total_Order_By_Segments,
      ROUND(SUM(Sales) OVER(partition by Segment),2) AS Sales_by_segment,
      ROUND(SUM(Sales) OVER(),2) AS OverallSales
      FROM sales_data
      )
      SELECT
            distinct Segment,
            Total_Order_By_Segments,
            Sales_by_segment,
            OverallSales,
            CONCAT(ROUND((Sales_by_segment/ OverallSales) *100 ,2),'%') AS Perct_of_sales_By_Segments
            FROM SalesBySegment;
```

- **Top 10 Customers by Revenue**

Segment	Total_Order_By_Segments	Sales_by_segment	OverallSales	Perct_of_sales_By_Segments
Consumer	5096	1146708.15	2252607.41	50.91%
Corporate	2948	682211.83	2252607.41	30.29%
Home Office	1745	423687.43	2252607.41	18.81%

🛍️ Product Metrics

- **Top-Selling Products**

```sql
-- top Selling Products

SELECT
      Category,
      ROUND(SUM(Sales),2) AS Total_Sales
      FROM sales_data
      GROUP BY Category
      ORDER BY Total_sales DESC;
```

Category	Total_Sales
Technology	825856.11
Furniture	723538.48
Office Supplies	703212.82

- **Most Profitable Product Categories**

```sql
-- Product Performance By Region

SELECT
   DISTINCT Region,
   Category,
   ROUND(SUM(Sales) OVER(partition by Region, Category ORDER BY Region),2) AS Total_sales,
   ROUND(SUM(Sales) OVER(),2) AS Overall_Sales,
   CONCAT(ROUND((SUM(Sales) OVER(partition by Region, Category ORDER BY Region)/SUM(Sales) OVER()*100),2),' %') AS perct_Overall_sales
   FROM sales_data;
```

- **Product Performance by Region**

Region	Category	Total_sales	Overall_Sales	perct_Overall_sales
Central	Furniture	160317.46	2252607.41	7.12 %
Central	Office Supplies	163590.24	2252607.41	7.26 %
Central	Technology	168739.21	2252607.41	7.49 %
East	Furniture	201341.29	2252607.41	8.94 %
East	Office Supplies	197731.3	2252607.41	8.78 %
East	Technology	261516.77	2252607.41	11.61 %
South	Furniture	116531.48	2252607.41	5.17 %
South	Office Supplies	124424.77	2252607.41	5.52 %
South	Technology	148195.21	2252607.41	6.58 %
West	Furniture	245348.25	2252607.41	10.89 %
West	Office Supplies	217466.51	2252607.41	9.65 %
West	Technology	247404.93	2252607.41	10.98 %

🧠 Business Problems

1. Sales Performance Analysis

Problem: Understand what's driving revenue.

- Which products/categories perform best?
- Which regions bring in the most revenue?
- Are there seasonal patterns in sales?

2. Customer Segmentation

Problem: Personalize marketing efforts.

- Which customer segments (Consumer, Corporate, Home Office) are most profitable?
- Who are your most loyal or high-value customers?

3. Inventory Management

Problem: Stock popular products, reduce dead stock.

- Identify high-demand and low-demand products.

```sql
-- Inventory Optimization
-- Identify high-demand and low-demand products.

  WITH CTE AS (
SELECT
    Category,
    `Sub-Category`,
    ROUND(SUM(Sales), 2) AS Total_sales_by_Sub_category
FROM sales_data
GROUP BY Category, `Sub-Category`
)
SELECT
    Category,
    `Sub-Category`,
    Total_sales_by_Sub_category,
    DENSE_RANK() OVER(PARTITION BY Category ORDER BY Total_sales_by_Sub_category DESC) AS rn
FROM CTE;
```

Category	Sub-Category	Total_sales_by_Sub_category	rn
Furniture	Chairs	322107.53	1
Furniture	Tables	202810.63	2
Furniture	Bookcases	109408.3	3
Furniture	Furnishings	89212.02	4
Office Supplies	Storage	217779.1	1
Office Supplies	Binders	200028.79	2
Office Supplies	Appliances	104075.46	3
Office Supplies	Paper	76736.1	4
Office Supplies	Supplies	46420.31	5
Office Supplies	Art	26697.37	6
Office Supplies	Envelopes	16126.01	7
Office Supplies	Labels	12347.73	8
Office Supplies	Fasteners	3001.96	9
Technology	Phones	326487.7	1
Technology	Machines	189238.63	2
Technology	Accessories	163881.69	3
Technology	Copiers	146248.09	4

```sql
-- Inventory Optimization
-- Identify high-demand and low-demand products.

  WITH CTE AS (
SELECT
    Category,
    `Sub-Category`,
    ROUND(SUM(Sales), 2) AS Total_sales_by_Sub_category
FROM sales_data
GROUP BY Category, `Sub-Category`
)
SELECT
    Category,
    `Sub-Category`,
    Total_sales_by_Sub_category,
    DENSE_RANK() OVER(PARTITION BY Category ORDER BY Total_sales_by_Sub_category DESC) AS rn
FROM CTE;
```

Category	Sub-Category	Total_sales_by_Sub_category	rn
Furniture	Chairs	322107.53	1
Furniture	Tables	202810.63	2
Furniture	Bookcases	109408.3	3
Furniture	Furnishings	89212.02	4
Office Supplies	Storage	217779.1	1
Office Supplies	Binders	200028.79	2
Office Supplies	Appliances	104075.46	3
Office Supplies	Paper	76736.1	4
Office Supplies	Supplies	46420.31	5
Office Supplies	Art	26697.37	6
Office Supplies	Envelopes	16126.01	7
Office Supplies	Labels	12347.73	8
Office Supplies	Fasteners	3001.96	9
Technology	Phones	326487.7	1
Technology	Machines	189238.63	2
Technology	Accessories	163881.69	3
Technology	Copiers	146248.09	4

- Predict future sales based on trends.

4. Shipping Optimization

Problem: Reduce delivery time and costs.

- Which shipping methods are fastest or most cost-effective?
- Are there regions with longer delivery times?

```
-- Are there regions with longer delivery times?


WITH RankedCities AS (
    SELECT
        State,
        City,
        CONCAT(ROUND(AVG(DATEDIFF(`Ship Date`, `Order Date`)), 0),' Days') AS AvgDelivery_Time
    FROM sales_data
    GROUP BY State, City
),
Ranked AS (
    SELECT
         State,
        City,
        AvgDelivery_Time,
        DENSE_RANK() OVER(partition by State ORDER BY AvgDelivery_Time  ) AS RankBasedOnTime
    FROM RankedCities
)
SELECT *
FROM Ranked
ORDER BY AvgDelivery_Time;
```

State	City	AvgDelivery_Time	RankBasedOnTime
Montana	Billings	0 Days	1
Texas	Bryan	0 Days	1
Louisiana	Kenner	0 Days	1
South Carolina	Rock Hill	0 Days	1
Michigan	Mount Pleasant	0 Days	1
Indiana	Portage	0 Days	1
Ohio	Mentor	0 Days	1
Illinois	Tinley Park	1 Days	1
Georgia	Smyrna	1 Days	1
Missouri	Kirkwood	1 Days	1
Missouri	Gladstone	1 Days	1
Pennsylvania	Bethlehem	1 Days	1
Texas	The Colony	1 Days	2
California	Antioch	1 Days	1
Pennsylvania	Altoona	1 Days	1
California	Redlands	1 Days	1
Illinois	Palatine	1 Days	1
Utah	Murray	1 Days	1
Colorado	Pueblo	1 Days	1
Michigan	Lansing	1 Days	2
Iowa	Cedar Rapids	1 Days	1
New York	Utica	1 Days	1
Oregon	Eugene	2 Days	1
Georgia	Marietta	2 Days	2
Missouri	Saint Charles	2 Days	2
Louisiana	Lake Charles	2 Days	2
Arkansas	Pine Bluff	2 Days	1
Louisiana	Bossier City	2 Days	2

5. Market Expansion Strategy

Problem: Decide where to expand or focus sales efforts.

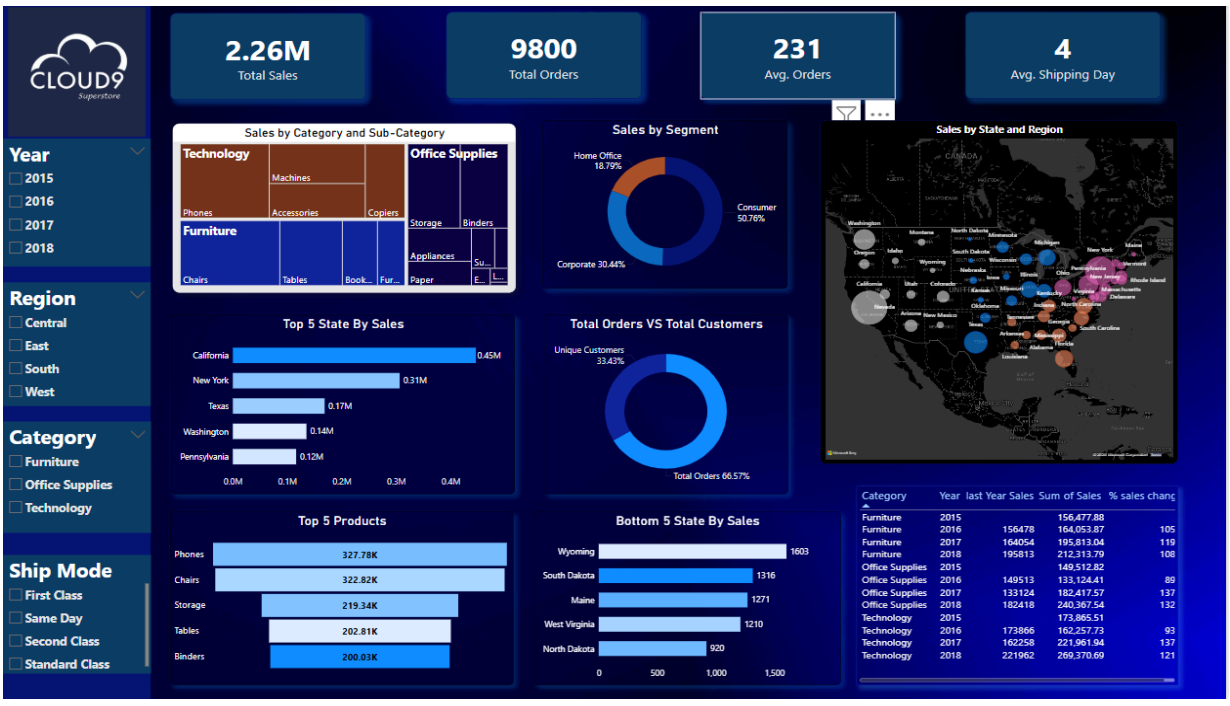
- Which regions have untapped potential?
- Where are sales underperforming?

6. Profitability and Cost Analysis

Problem: Maximize profits.

- Which categories have the highest margins? (if cost data is added)
- Are expensive shipping methods justified?

KEY FINDINGS



🟦 Overall Business Metrics

- **Total Sales**: **$2.26M**
- **Total Orders**: **9,800**
- **Average Orders** : **231**
- **Average Shipping Time**: **4 days**

🔍 This indicates a relatively efficient delivery process.

📊 Category and Sub-Category Performance

- **Top Performing Category**: **Technology** (by sales volume and sub-category spread)
- **Strongest Sub-Categories**:
  - **Chairs** – $327.78K
  - **Phones**
  - **Storage**
  - **Tables**
  - **Binders**
- **Underperforming Sub-Categories** (implied by size in treemap):
  - **Fasteners**, **Labels**, and **Supplies**

✅ **Opportunity**: Focus promotions on underperforming sub-categories or evaluate inventory reduction strategies.

🧭 Regional and State-Level Performance

📌 Top 5 States by Sales:

1. **California** – $450K+
2. **New York** – $310K+
3. Texas
4. Washington
5. Pennsylvania

🟥 Bottom 5 States by Sales:

- **North Dakota**, **West Virginia**, **Maine**, **South Dakota**, **Wyoming**

📉 Sales below 1.6K, indicating either low market penetration or demand.

🔍 **Recommendation**: Conduct market research in low-performing states to identify barriers and explore opportunities for localized campaigns.

🧭 Sales by Segment

- **Consumer Segment**: **50.76%** of total sales – dominant contributor.
- **Corporate**: 30.44%
- **Home Office**: 18.79%

📌 Insight: Focus on **consumer-specific marketing** and loyalty programs.

🧮 Order vs Customer Insights

- **Unique Customers**: 33.43%
- **Repeat Orders**: 66.57%

💡 **High repeat purchase rate** suggests strong customer retention.

🗓️ Sales Trend by Year & Category

- **Consistent Year-over-Year Growth** across all categories.
- **Highest % Sales Change**:
    - **Office Supplies** (2017–2018): +137%
    - **Furniture** (2017–2018): +132%
    - **Technology** (2017–2018): +121%

📈 **Conclusion**: Robust annual growth confirms effective sales strategies, though **Office Supplies** stands out for exponential improvement.

🌍 Sales by State and Region (Map View)

- Larger blue bubbles indicate **high-volume states like California and Texas**
- Pink clusters in the Southeast suggest **growing opportunities** in South/East regions.

💼 Regional sales teams could focus efforts on expanding high-engagement zones and nurturing emerging markets.

✅ Strategic Recommendations

1. **Expand** operations and marketing in high-growth regions (West and Northeast).
2. **Improve** engagement in low-sales states using regional promotions.
3. **Double down** on high-performing sub-categories (e.g., Chairs, Phones).
4. **Analyze underperforming products** to optimize inventory or bundle offerings.
5. **Enhance consumer targeting**, as they represent over 50% of revenue.