Retail Sales Analysis: Superstore Dataset

G. Turgen

Introduction

This case study explores sales data from the Superstore dataset to uncover patterns in product performance, customer segments, and profitability trends. The goal is to derive actionable insights to improve revenue and optimize business strategies.

Business Task

Analyze Superstore sales data to identify high-performing products, profitable customer segments, and regional sales trends. Use these insights to inform marketing, inventory, and pricing strategies.

Stakeholders

- Regional Sales Directors
- Product Management Team
- Marketing & Strategy Department

About the Data

The dataset includes fictional sales transactions from a U.S.-based retail store, containing details such as order date, product category, region, customer segment, sales, and profit. This dataset is commonly used for data visualization and business analytics training.

ROCCC Assessment

- **Relevant**: Yes — closely aligned with the business questions.
- **Original**: No — publicly available, not proprietary.
- **Comprehensive**: Moderate — includes key variables (e.g., date, sales, region, segment).
- **Current**: Undated — no indication of real-world time frame.
- **Cited**: Yes — sourced from Tableau Public / Kaggle.

Data Cleaning

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.2     v tibble    3.2.1
## v lubridate 1.9.4     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(janitor)
```

```
##
## Attaching package: 'janitor'
##
## The following objects are masked from 'package:stats':
##
##     chisq.test, fisher.test
```

```
library(lubridate)
library(skimr)
library(ggplot2)
library(readr)
```

Load the data

```
superstore_df <- read_csv("/Users/dimitrid./Superstore Case Study/Sample - Superstore.csv") %>%
  clean_names()
```

```
## Rows: 9994 Columns: 21
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (16): Order ID, Order Date, Ship Date, Ship Mode, Customer ID, Customer ...
## dbl  (5): Row ID, Sales, Quantity, Discount, Profit
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Convert dates

```r
superstore_df <- superstore_df %>%
  mutate(order_date = mdy(order_date),
         ship_date = mdy(ship_date))
```

Exploratory Analysis

Product Performance

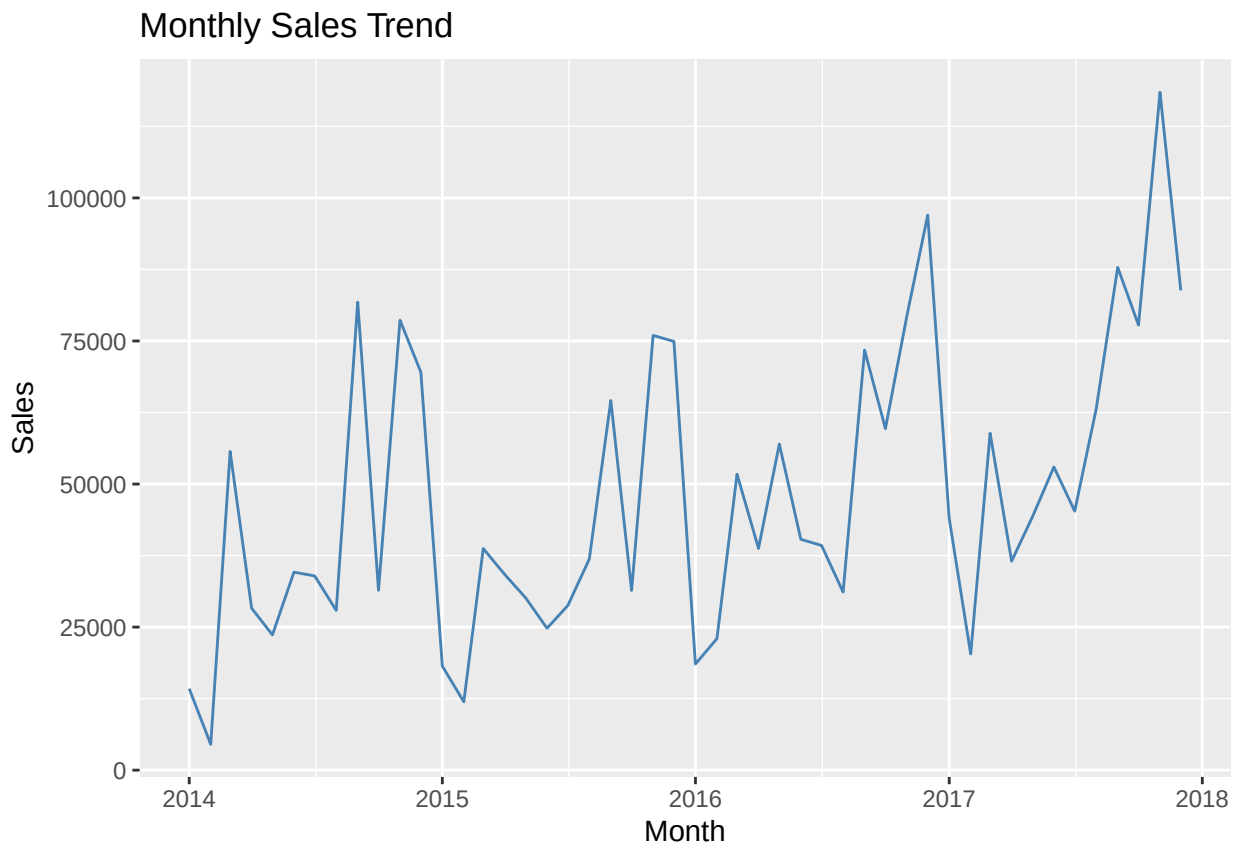
```r
superstore_df %>%
  group_by(product_name) %>%
  summarize(total_sales = sum(sales, na.rm = TRUE)) %>%
  arrange(desc(total_sales)) %>%
  head(10)
```

```
## # A tibble: 10 x 2
##    product_name                                                  total_sales
##    <chr>                                                               <dbl>
##  1 "Canon imageCLASS 2200 Advanced Copier"                           61600.
##  2 "Fellowes PB500 Electric Punch Plastic Comb Binding Machine with~ 27453.
##  3 "Cisco TelePresence System EX90 Videoconferencing Unit"           22638.
##  4 "HON 5400 Series Task Chairs for Big and Tall"                    21871.
##  5 "GBC DocuBind TL300 Electric Binding System"                      19823.
##  6 "GBC Ibimaster 500 Manual ProClick Binding System"                19024.
##  7 "Hewlett Packard LaserJet 3310 Copier"                            18840.
##  8 "HP Designjet T520 Inkjet Large Format Printer - 24\" Color"      18375.
##  9 "GBC DocuBind P400 Electric Binding System"                       17965.
## 10 "High Speed Automatic Electric Letter Opener"                     17030.
```

- The highest-grossing products are large-ticket items, such as copiers and videoconferencing units.
- The top product alone (Canon Copier) accounts for over $60,000 in sales.
- Sales are heavily concentrated in a few high-value products. This suggests that while the overall product range is broad, a small subset drives the majority of revenue.

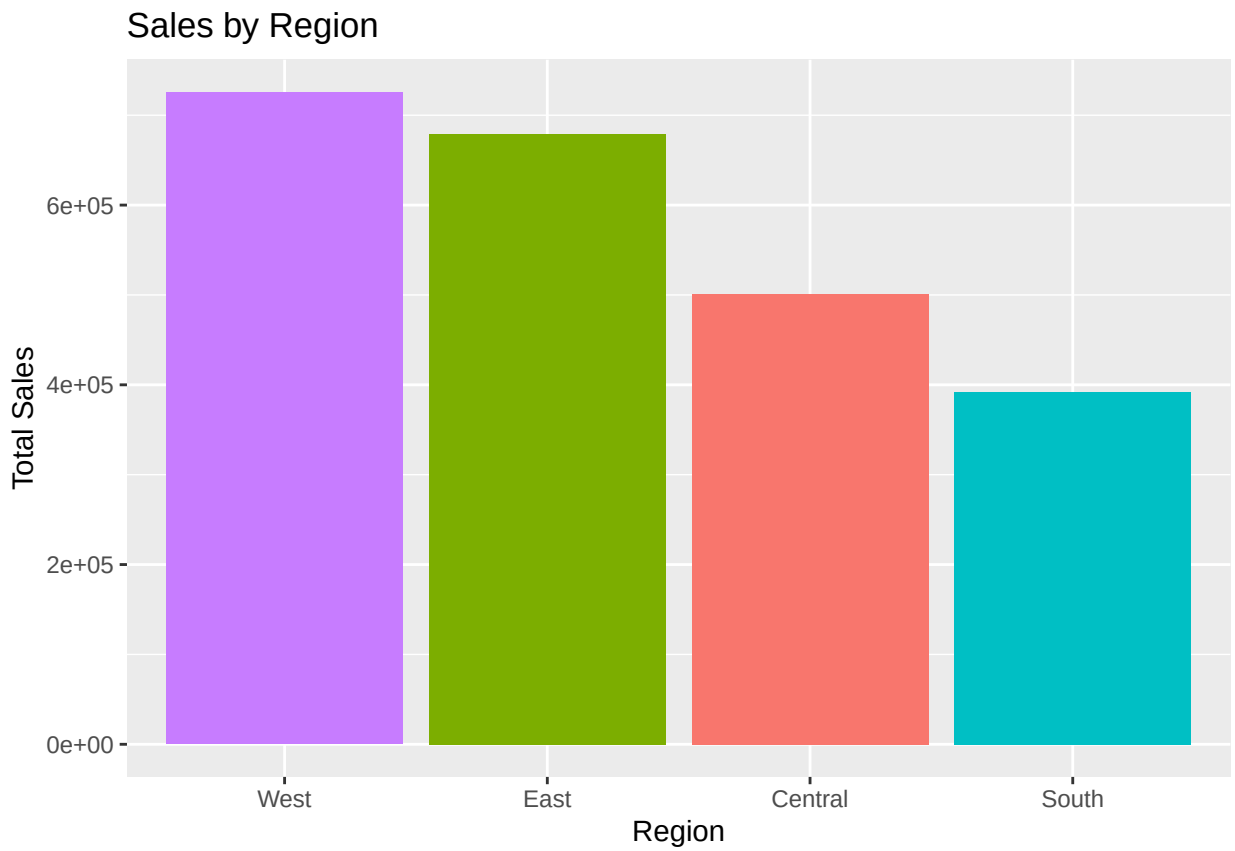
Sales Trends Over Time Visualization

```r
superstore_df %>%
  mutate(month = floor_date(order_date, "month")) %>%
  group_by(month) %>%
  summarize(monthly_sales = sum(sales, na.rm = TRUE)) %>%
  ggplot(aes(x = month, y = monthly_sales)) +
  geom_line(color = "steelblue") +
  labs(title = "Monthly Sales Trend", x = "Month", y = "Sales")
```



Regional Performance Visualization

```
superstore_df %>%
  group_by(region) %>%
  summarize(total_sales = sum(sales, na.rm = TRUE)) %>%
  ggplot(aes(x = reorder(region, -total_sales), y = total_sales, fill = region)) +
  geom_col() +
  labs(title = "Sales by Region", x = "Region", y = "Total Sales") +
  theme(legend.position = "none")
```



Customer Segment Analysis

Sales and Profit by Customer Segment:

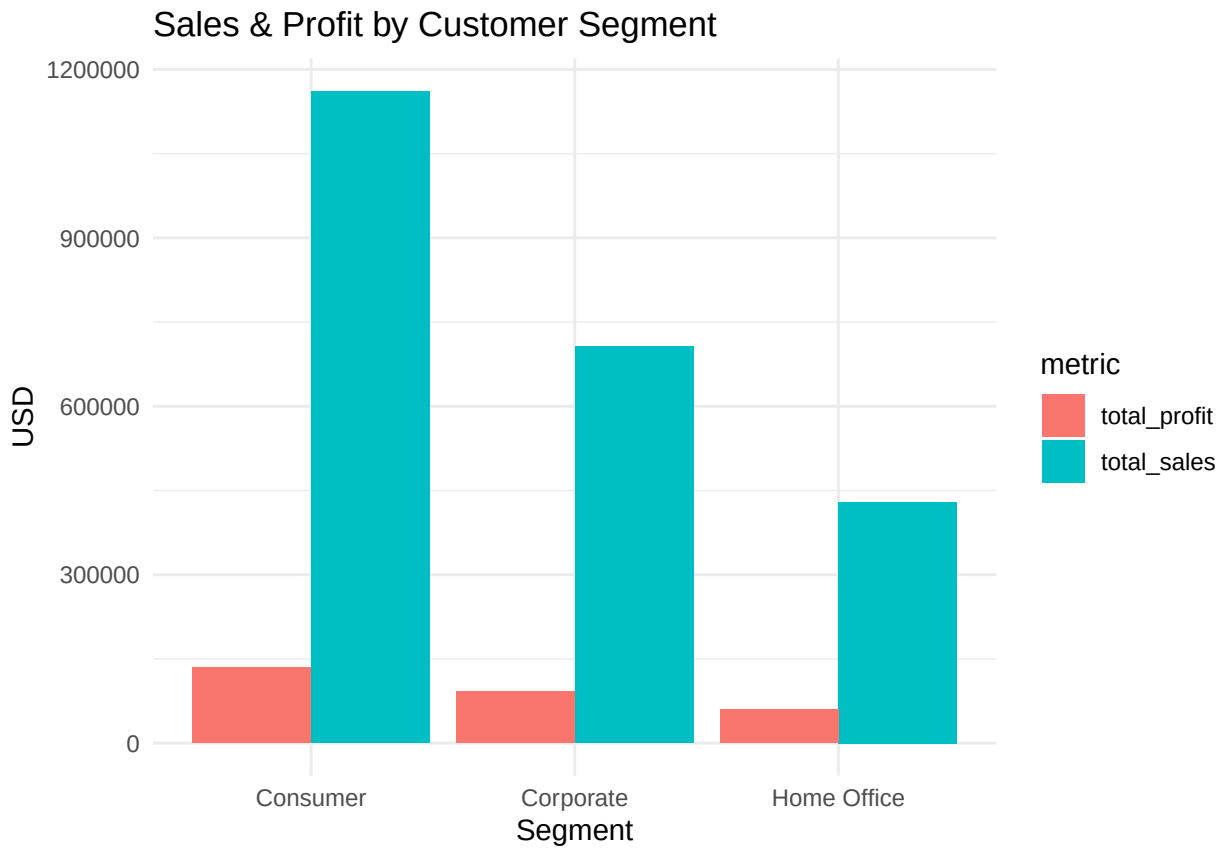
```
superstore_df %>%
  group_by(segment) %>%
  summarize(
    total_sales = sum(sales, na.rm = TRUE),
    total_profit = sum(profit, na.rm = TRUE),
    avg_profit_margin = mean(profit / sales, na.rm = TRUE),
    order_count = n()
  ) %>%
  arrange(desc(total_sales))
```

```
## # A tibble: 3 x 5
##   segment     total_sales total_profit avg_profit_margin order_count
##   <chr>             <dbl>        <dbl>             <dbl>       <int>
## 1 Consumer       1161401.      134119.             0.112        5191
## 2 Corporate       706146.       91979.             0.121        3020
## 3 Home Office     429653.       60299.             0.143        1783
```

While Consumers generate the most revenue, Home Office customers appear to be more profitable on average. This indicates that different customer types bring different types of value—volume vs. profitability—and may warrant distinct engagement strategies.

Visualize Sales and Profit:

```
superstore_df %>%
  group_by(segment) %>%
  summarize(
    total_sales = sum(sales, na.rm = TRUE),
    total_profit = sum(profit, na.rm = TRUE)
  ) %>%
  pivot_longer(cols = c(total_sales, total_profit), names_to = "metric", values_to = "value") %>%
  ggplot(aes(x = segment, y = value, fill = metric)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Sales & Profit by Customer Segment", x = "Segment", y = "USD") +
  theme_minimal()
```

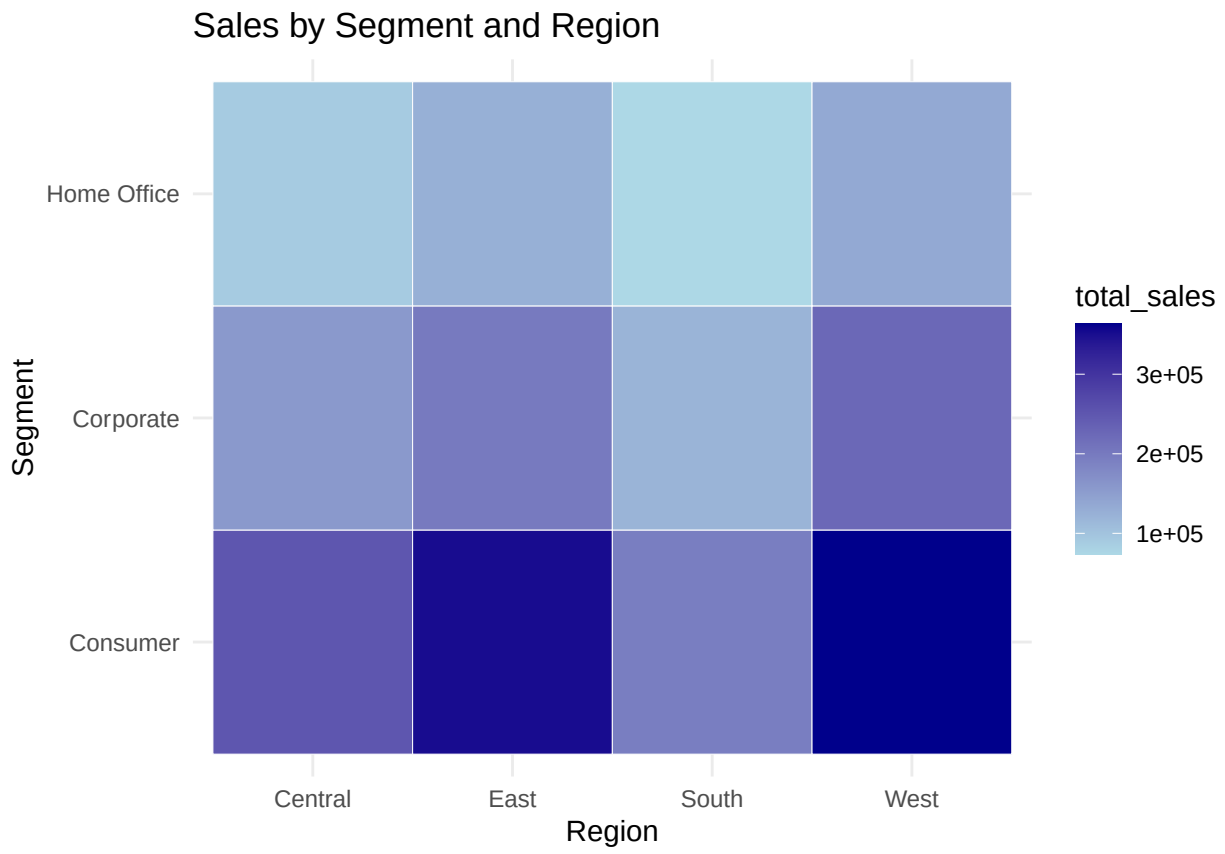


Visualize Sales by Segment + Region:

```
superstore_df %>%
  group_by(segment, region) %>%
  summarize(total_sales = sum(sales, na.rm = TRUE)) %>%
  ggplot(aes(x = region, y = segment, fill = total_sales)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightblue", high = "darkblue") +
  labs(title = "Sales by Segment and Region", x = "Region", y = "Segment") +
  theme_minimal()
```

```
## `summarise()` has grouped output by 'segment'. You can override using the
```

```
## `.groups` argument.
```



Profitability Breakdown

This analysis will help you understand where profit is actually coming from, regardless of sales volume. My goal is to identify which categories, sub-categories, or regions are driving or hurting profitability — even if their sales look good.

Profitability by Category and Sub-Category:

```
superstore_df %>%
  group_by(category, sub_category) %>%
  summarize(
    total_sales = sum(sales, na.rm = TRUE),
    total_profit = sum(profit, na.rm = TRUE),
    profit_margin = total_profit / total_sales
  ) %>%
  arrange(desc(profit_margin))
```

```
## `summarise()` has grouped output by 'category'. You can override using the
## `.groups` argument.
```

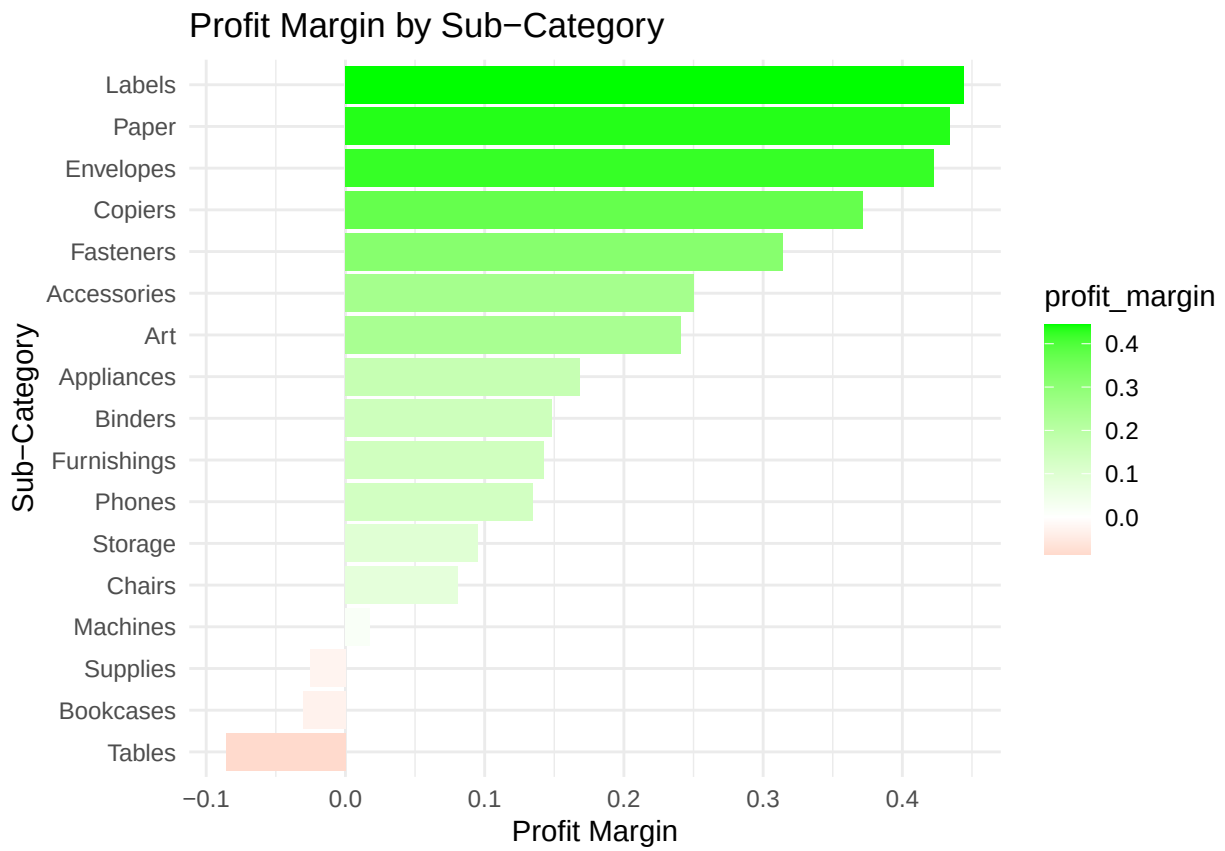
```
## # A tibble: 17 x 5
## # Groups:   category [3]
```

```
##    category        sub_category total_sales total_profit profit_margin
##    <chr>           <chr>              <dbl>        <dbl>         <dbl>
##  1 Office Supplies Labels             12486.        5546.        0.444
##  2 Office Supplies Paper              78479.       34054.        0.434
##  3 Office Supplies Envelopes          16476.        6964.        0.423
##  4 Technology      Copiers           149528.       55618.        0.372
##  5 Office Supplies Fasteners           3024.         950.        0.314
##  6 Technology      Accessories       167380.       41937.        0.251
##  7 Office Supplies Art                27119.        6528.        0.241
##  8 Office Supplies Appliances        107532.       18138.        0.169
##  9 Office Supplies Binders           203413.       30222.        0.149
## 10 Furniture       Furnishings        91705.       13059.        0.142
## 11 Technology      Phones            330007.       44516.        0.135
## 12 Office Supplies Storage           223844.       21279.        0.0951
## 13 Furniture       Chairs            328449.       26590.        0.0810
## 14 Technology      Machines          189239.        3385.        0.0179
## 15 Office Supplies Supplies           46674.       -1189.       -0.0255
## 16 Furniture       Bookcases         114880.       -3473.       -0.0302
## 17 Furniture       Tables            206966.      -17725.       -0.0856
```

- Office Supplies like Labels, Paper, and Envelopes have the highest profit margins ($> 40\%$).
- Some categories, such as Tables and Bookcases, show negative profitability.
- Profitability varies widely across sub-categories. Low-cost office supplies deliver strong margins, while some furniture items operate at a loss. This could be due to pricing issues, high return rates, or shipping costs. Product-level strategy adjustments may be needed.

Visualize Profit Margin by Sub-Category:

```r
superstore_df %>%
  group_by(sub_category) %>%
  summarize(
    total_sales = sum(sales, na.rm = TRUE),
    total_profit = sum(profit, na.rm = TRUE),
    profit_margin = total_profit / total_sales
  ) %>%
  ggplot(aes(x = reorder(sub_category, profit_margin), y = profit_margin, fill = profit_margin)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Profit Margin by Sub-Category", x = "Sub-Category", y = "Profit Margin") +
  scale_fill_gradient2(low = "red", high = "green", midpoint = 0) +
  theme_minimal()
```



Identify Loss-Making Categories:

```
superstore_df %>%
  group_by(sub_category) %>%
  summarize(total_profit = sum(profit, na.rm = TRUE)) %>%
  filter(total_profit < 0) %>%
  arrange(total_profit)
```

```
## # A tibble: 3 x 2
##   sub_category total_profit
##   <chr>               <dbl>
## 1 Tables            -17725.
## 2 Bookcases          -3473.
## 3 Supplies           -1189.
```

- Tables, Bookcases, and Supplies sub-categories have negative overall profit.
- These are clear loss-makers. Their continued inclusion in the catalog may require strategic review, particularly regarding pricing, supplier costs, or delivery/logistics challenges.

Profitability by Region:

```
superstore_df %>%
  group_by(region) %>%
  summarize(
    total_sales = sum(sales, na.rm = TRUE),
```

```r
    total_profit = sum(profit, na.rm = TRUE),
    profit_margin = total_profit / total_sales
  ) %>%
  arrange(desc(profit_margin))
```

```
## # A tibble: 4 x 4
##   region  total_sales total_profit profit_margin
##   <chr>         <dbl>        <dbl>         <dbl>
## 1 West        725458.      108418.        0.149
## 2 East        678781.       91523.        0.135
## 3 South       391722.       46749.        0.119
## 4 Central     501240.       39706.        0.0792
```

- The West region shows the highest profit margin (14.9%), while Central has the lowest (7.9%).
- Regional differences are significant. The West outperforms in both sales and margin, indicating stronger market dynamics or customer base. Central's lower margin may reflect operational inefficiencies or competitive pricing pressures.

Key Business Insights

1- Sales Are Concentrated in High-Value Items A small number of expensive products (e.g., Canon Copiers, Cisco Systems) generate a disproportionately high share of revenue. This highlights the importance of prioritizing inventory management, pricing strategies, and targeted marketing efforts around these "high-ticket" items.

2- Customer Segments Differ by Value Type While the Consumer segment brings the highest total sales, the Home Office segment has the highest average profit margin. This suggests segmentation-based strategies: scale-focused for consumers, margin-focused for home office users.

3- Sub-Category Profitability Is Highly Uneven Office Supplies like Labels and Paper offer exceptional margins (above 40%), while Furniture items like Tables and Bookcases result in financial losses. These insights point to opportunities for product rationalization and profit optimization.

4- Negative-Profit Sub-Categories Need Attention Tables and Bookcases consistently generate losses. Root causes (e.g., high shipping cost, overstock, discounting) should be investigated, and pricing or portfolio strategies revised.

5- West Region Leads in Profitability The West region achieves both high sales and the highest profit margin (14.9%), suggesting a robust customer base or effective operations. Conversely, the Central region shows underperformance and requires further analysis.

Limitations

- The dataset reflects a limited time period, and seasonal trends cannot be assessed reliably.
- Gender, age, and income levels are not provided, limiting the depth of customer behavior analysis.
- Product lifecycles, pricing changes, or promotions are not included—key factors that could affect profitability.
- The impact of returns, refunds, or discounts on profitability is not directly observable.

Conclusion

This analysis uncovered key areas of opportunity for the business to optimize profitability and streamline its product portfolio. High-value technology items and office supplies drive revenue and margins, whereas select furniture categories require attention. Regional and segment-level patterns suggest targeted strategies to unlock further value. With improved data granularity and longitudinal tracking, future analysis can better inform pricing, marketing, and inventory decisions.

Future Work

Future analyses could:

- analyze loss-making products by customer segment, region, and season to understand drivers,
- incorporate customer-level purchase patterns over time to assess CLV by segment,
- use time-series models to forecast demand and identify trends in top-performing categories,
- explore how discounts or promotions affect profit margins in different categories.