Superstore Sales Analysis with R

Camila Pattaro

2025-02-23

Introduction

Superstore Sales contains information about products, sales, and profits, and it can be found here.

The goal of this project is to answer a number of business questions related to the company's sales.

1. How have sales evolved over time?

2. What percentage of sales and profit is represented by each category?

3. Which product sub-categories have the highest sales?

4. Which product sub-category generates the highest sales and which one generates the most profit? What is the relationship between them?

5. What are the TOP 10 products according to their sales?

6. Which states have the highest sales?

Data Preparation

```
# Loading libraries
library(readxl)
library(janitor)
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
library(treemapify)
library(tidyr)
library(sf)
library(scales)
library(maps)
```

```
# Loading the data
ss <-read_excel("C:\\Users\\camil\\OneDrive\\Documents\\KAGGLE\\sample_-_superstore.xls")
```

```
# Getting a overview of the data
glimpse(ss)
```

```
## Rows: 9,994
## Columns: 21
## $ `Row ID`       <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16,~
## $ `Order ID`     <chr> "CA-2016-152156", "CA-2016-152156", "CA-2016-138688", ~
## $ `Order Date`   <dttm> 2016-11-08, 2016-11-08, 2016-06-12, 2015-10-11, 2015-~
## $ `Ship Date`    <dttm> 2016-11-11, 2016-11-11, 2016-06-16, 2015-10-18, 2015-~
## $ `Ship Mode`    <chr> "Second Class", "Second Class", "Second Class", "Stand~
```

```
## $ `Customer ID`   <chr> "CG-12520", "CG-12520", "DV-13045", "SO-20335", "SO-20~
## $ `Customer Name` <chr> "Claire Gute", "Claire Gute", "Darrin Van Huff", "Sean~
## $ Segment         <chr> "Consumer", "Consumer", "Corporate", "Consumer", "Cons~
## $ Country         <chr> "United States", "United States", "United States", "Un~
## $ City            <chr> "Henderson", "Henderson", "Los Angeles", "Fort Lauderd~
## $ State           <chr> "Kentucky", "Kentucky", "California", "Florida", "Flor~
## $ `Postal Code`   <dbl> 42420, 42420, 90036, 33311, 33311, 90032, 90032, 90032~
## $ Region          <chr> "South", "South", "West", "South", "South", "West", "W~
## $ `Product ID`    <chr> "FUR-BO-10001798", "FUR-CH-10000454", "OFF-LA-10000240~
## $ Category        <chr> "Furniture", "Furniture", "Office Supplies", "Furnitur~
## $ `Sub-Category`  <chr> "Bookcases", "Chairs", "Labels", "Tables", "Storage", ~
## $ `Product Name`  <chr> "Bush Somerset Collection Bookcase", "Hon Deluxe Fabri~
## $ Sales           <dbl> 261.9600, 731.9400, 14.6200, 957.5775, 22.3680, 48.860~
## $ Quantity        <dbl> 2, 3, 2, 5, 2, 7, 4, 6, 3, 5, 9, 4, 3, 3, 5, 3, 6, 2, ~
## $ Discount        <dbl> 0.00, 0.00, 0.00, 0.45, 0.20, 0.00, 0.00, 0.20, 0.20, ~
## $ Profit          <dbl> 41.9136, 219.5820, 6.8714, -383.0310, 2.5164, 14.1694,~
```

```r
# Cleaning and standardizing column names
ss <- clean_names(ss)
colnames(ss)
```

```
##  [1] "row_id"        "order_id"      "order_date"    "ship_date"
##  [5] "ship_mode"     "customer_id"   "customer_name" "segment"
##  [9] "country"       "city"          "state"         "postal_code"
## [13] "region"        "product_id"    "category"      "sub_category"
## [17] "product_name"  "sales"         "quantity"      "discount"
## [21] "profit"
```

```r
# Checking missing values
colSums(is.na(ss))
```

```
##        row_id      order_id    order_date     ship_date     ship_mode
##             0             0             0             0             0
##   customer_id customer_name       segment       country          city
##             0             0             0             0             0
##         state   postal_code        region    product_id      category
##             0             0             0             0             0
##  sub_category  product_name         sales      quantity      discount
##             0             0             0             0             0
##        profit
##             0
```

```r
# Removing duplicate rows
ss <- ss[!duplicated(ss),]
```

Data Analysis

1. How have sales evolved over time?

```r
# Extracting month and year from order_date column
ss$year_month <- format(ss$order_date, "%Y-%m")
head(ss$year_month)
```

```
## [1] "2016-11" "2016-11" "2016-06" "2015-10" "2015-10" "2014-06"
```

```r
# Aggregating total sales by year and month
monthly_sales <- aggregate(sales ~ year_month, data = ss, sum)
```

```r
head(monthly_sales)
```

```
##   year_month     sales
## 1    2014-01 14236.895
## 2    2014-02  4519.892
## 3    2014-03 55691.009
## 4    2014-04 28295.345
## 5    2014-05 23648.287
## 6    2014-06 34595.128
```

```r
class(ss$year_month)
```

```
## [1] "character"
```

We can see the year_month collumn have the wrong data type.

```r
# Converting year_month to Date by appending "-01" to represent the first day of the month
ss$year_month <- as.Date(paste0(ss$year_month, "-01"), format="%Y-%m-%d")
head(ss$year_month)
```

```
## [1] "2016-11-01" "2016-11-01" "2016-06-01" "2015-10-01" "2015-10-01"
## [6] "2014-06-01"
```

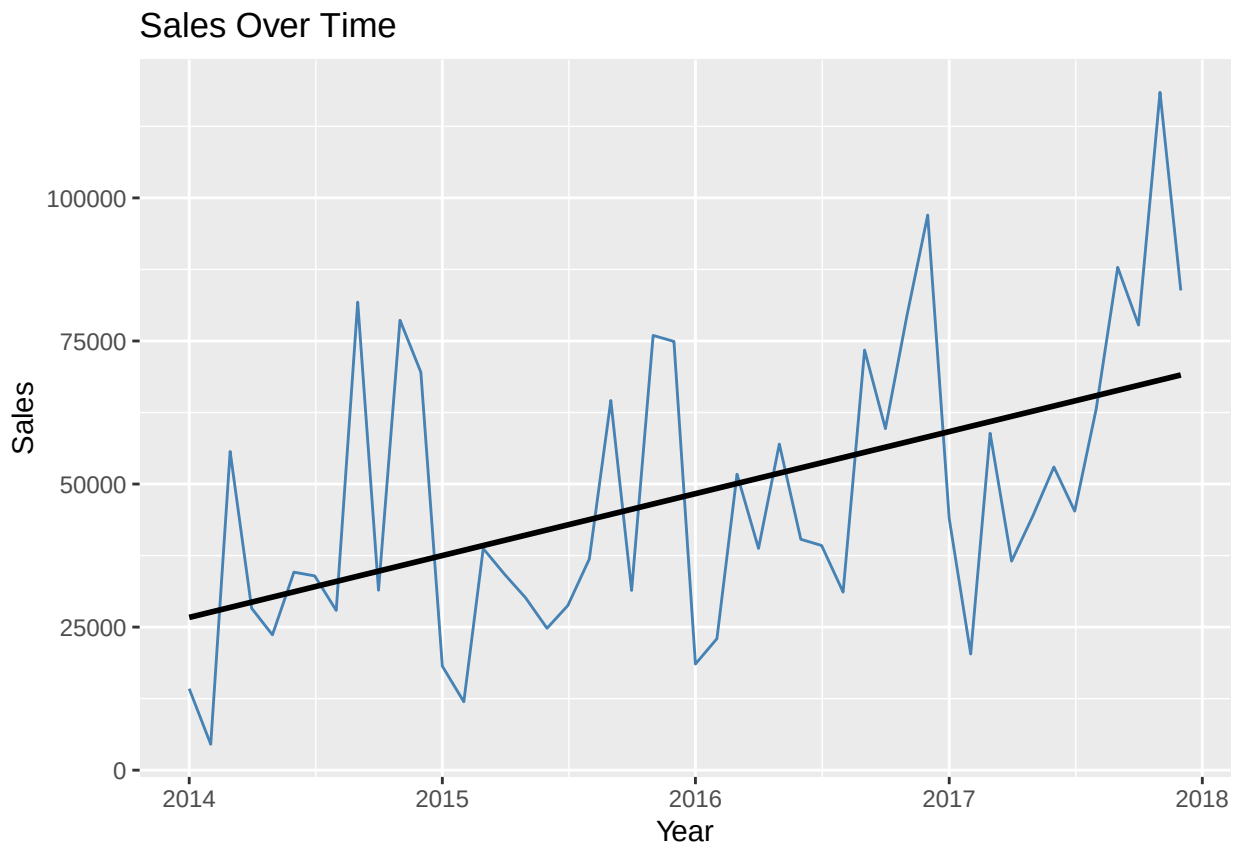
```r
# Convert 'year_month' to Date format
monthly_sales$year_month <- as.Date(paste0(monthly_sales$year_month, "-01"), format="%Y-%m-%d")
head(monthly_sales)
```

```
##   year_month     sales
## 1 2014-01-01 14236.895
## 2 2014-02-01  4519.892
## 3 2014-03-01 55691.009
## 4 2014-04-01 28295.345
## 5 2014-05-01 23648.287
## 6 2014-06-01 34595.128
```

Now we have the right data type.

```r
# Creating a line plot with a trend line
ggplot(monthly_sales, aes(x = year_month, y = sales)) +
  geom_line(color = "steelblue") +
  geom_smooth(method = "lm", color = "black", se = FALSE) +
  labs(title = "Sales Over Time", x = "Year", y = "Sales")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



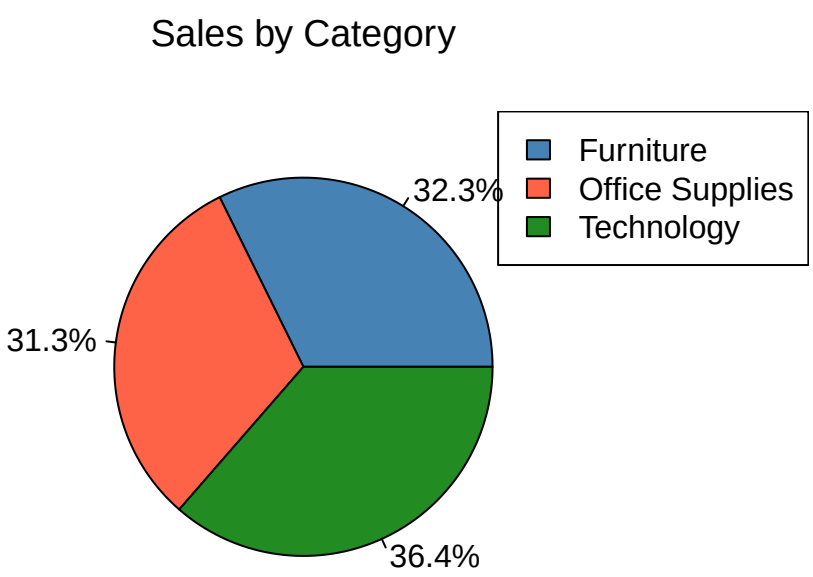
We observe a **growth** in sales over the years.

2. What percentage of sales and profit is represented by each category?

First let's analyse Sales:

```
# Grouping sales by category and calculate percentages
category_sales <- ss %>%
  group_by(category) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE)) %>%
  mutate(percentage = (total_sales / sum(total_sales)) * 100)
```

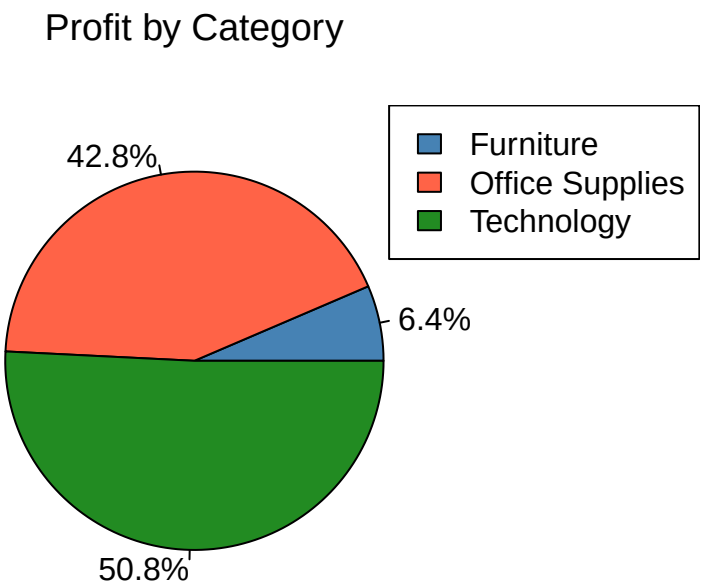
```
# Creating a pie chart
# Define custom colors (SteelBlue and Tomato)
custom_colors <- c("steelblue", "tomato", "forestgreen")

pie(category_sales$percentage,
    labels = paste0(round(category_sales$percentage, 1), "%"),
    main = "Sales by Category",
    col = custom_colors,
    font.main = 1)
legend("topright", legend = category_sales$category, fill = custom_colors)
```



Then Profits:

```r
# Grouping profit by category and calculate percentages
category_profit <- ss %>%
  group_by(category) %>%
  summarise(total_profit = sum(profit)) %>%
  mutate(percentage = (total_profit / sum(total_profit)) * 100)
```

```r
# Creating a pie chart
pie(category_profit$percentage,
    labels = paste0(round(category_profit$percentage, 1), "%"),
    main = "Profit by Category",
    col = custom_colors,
    font.main = 1)
legend("topright", legend = category_sales$category, fill = custom_colors)
```

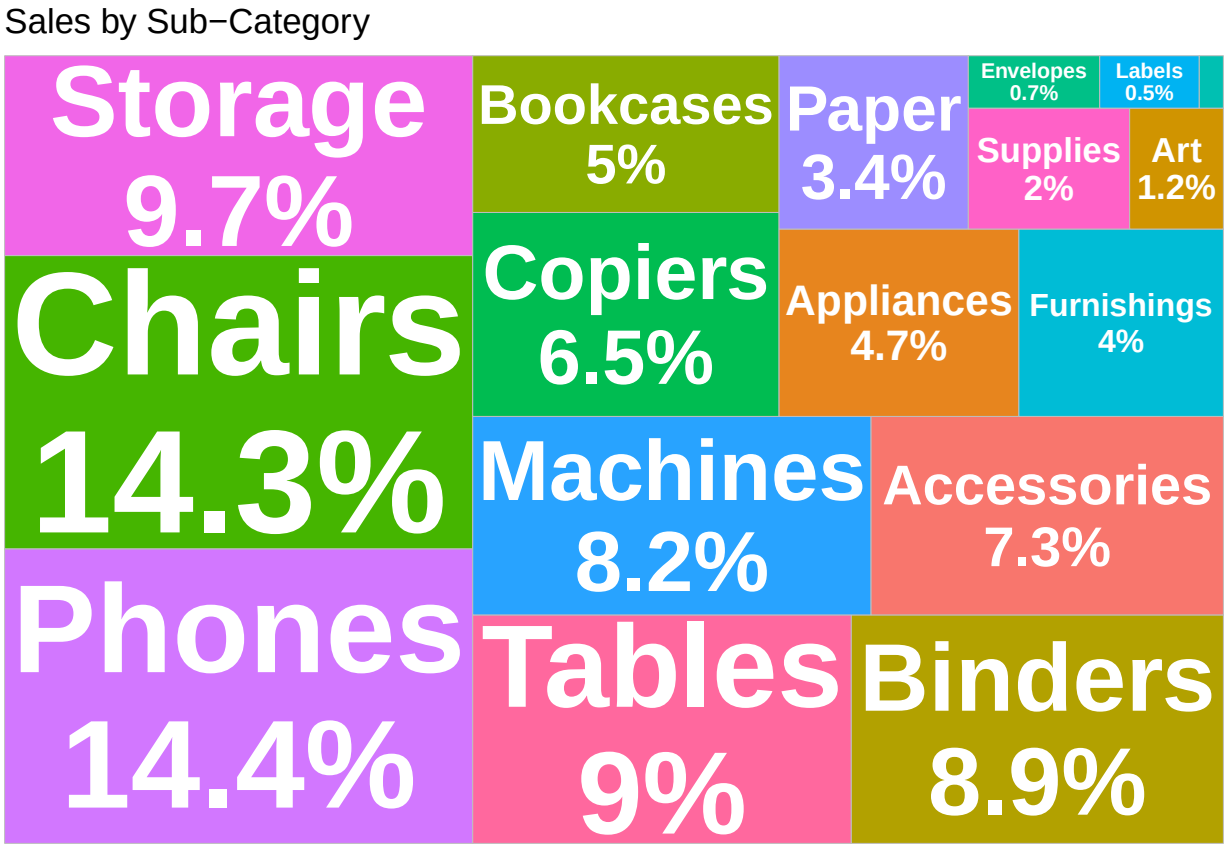


We can see that sales are distributed almost equally among the categories; however, profits are not. *Technology* accounts for more than half of the total profit, followed by *Office Supplies* with 43%, while *Furniture* represents only 6% of the total profit.

3. Which product sub-categories have the highest sales?

```
# Summarizing sales by sub_category
sub_category_sales <- ss %>%
  group_by(sub_category) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE)) %>%
  mutate(percentage = (total_sales / sum(total_sales))*100)
```

```
# Creating a treemap
ggplot(sub_category_sales, aes(area = total_sales, fill = sub_category,
                               label = paste0(sub_category, "\n", round(percentage, 1), "%"))) +
  geom_treemap() +
  geom_treemap_text(fontface = "bold", colour = "white", place = "centre", grow = TRUE) +
  labs(title = "Sales by Sub-Category") +
  theme_minimal() +
  theme(legend.position = "none")
```

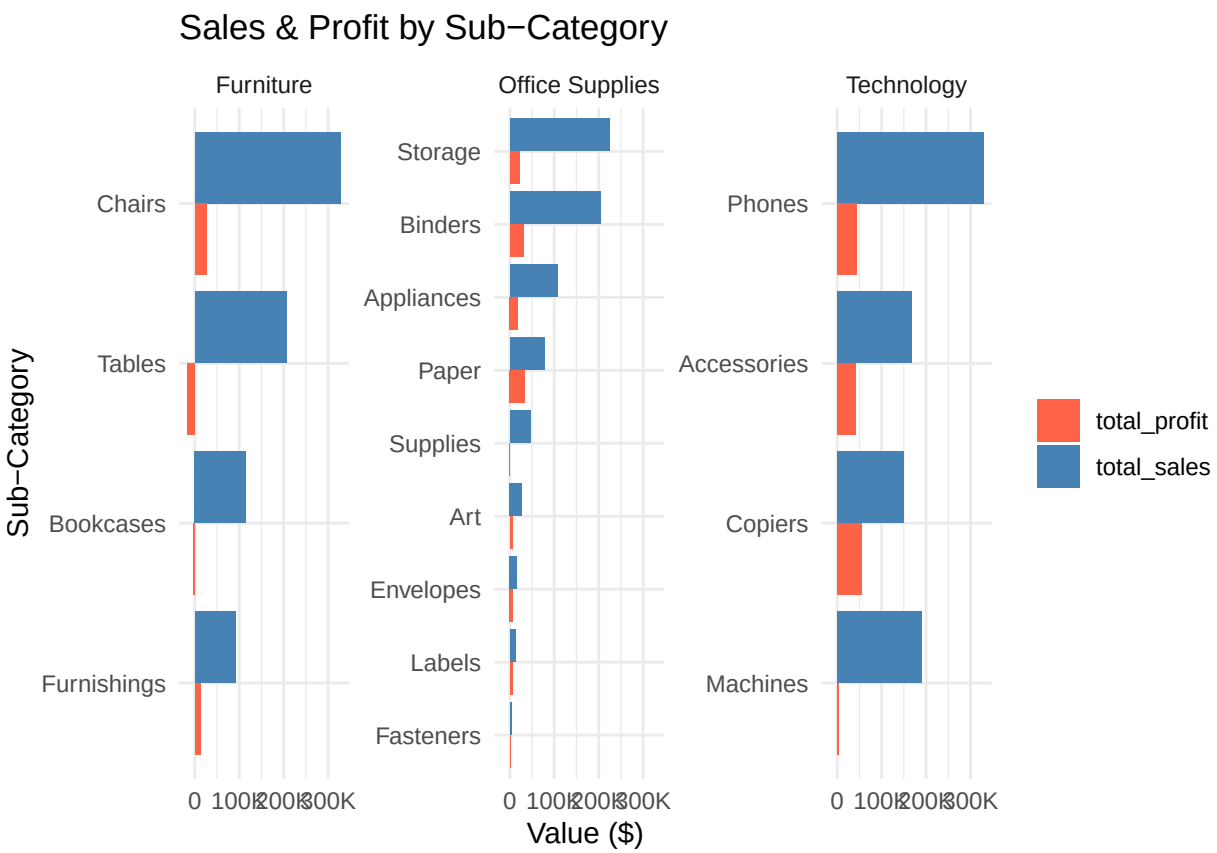


Phones and *Chairs* are the sub-categories with the highest sales, each representing 14% of the total sales.

4. Which product sub-category generates the highest sales and which one generates the most profit? What is the relationship between them?

```
# Summarizing sales and profit by category & sub-category
category_sales_profit <- ss %>%
  group_by(category, sub_category) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE),
            total_profit = sum(profit, na.rm = TRUE),
            .groups = "drop"  # Drop the grouping after summarizing
  ) %>%
  pivot_longer(cols = c(total_sales, total_profit),
               names_to = "metric",
               values_to = "value")  # Convert to long format
```

```
# Creating a side-by-side bar chart
ggplot(category_sales_profit, aes(x = value, y = reorder(sub_category, value), fill = metric)) +
  geom_col(position = "dodge") +
  facet_wrap(~category, scales = "free_y") +
  labs(title = "Sales & Profit by Sub-Category", x = "Value ($)", y = "Sub-Category") +
  theme_minimal() +
  scale_fill_manual(values = c("total_sales" = "steelblue", "total_profit" = "tomato")) +
  scale_x_continuous(labels = label_number(scale_cut = cut_short_scale())) +
  theme(legend.title = element_blank())
```

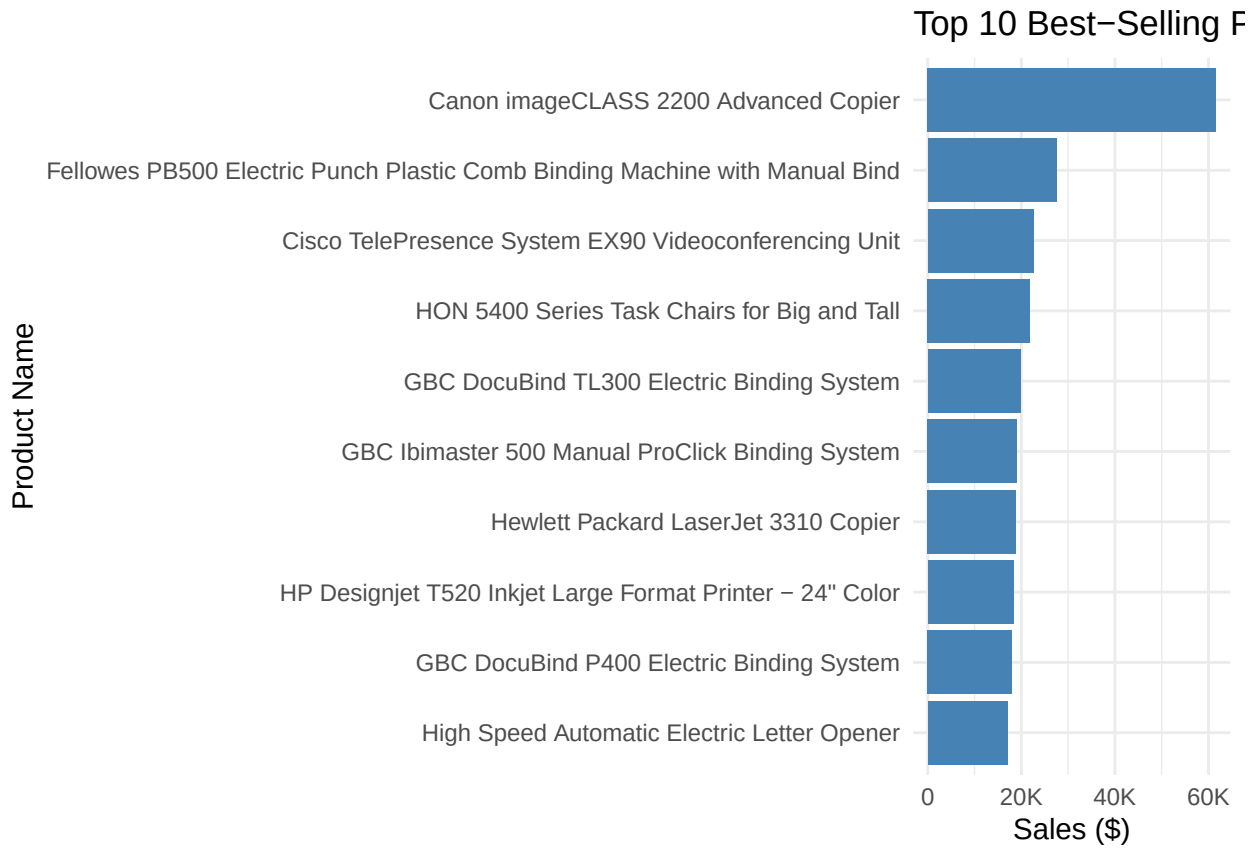


As we saw earlier, *Phones* and *Chairs* have the same revenue, but now we can observe that Phones generate much more profit than Chairs. *Tables* represent a high sales value but without profits, resulting in a significant loss. *Bookcases*, *Supplies*, and *Fasteners* also show losses. The only sub-category within Technology that does not present significant profit is *Machines*.

5. What are the TOP 10 products according to their sales?

```
# Summarizing the top 10 products by total sales
top_products <- ss %>%
  group_by(product_name) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE)) %>%
  arrange(desc(total_sales)) %>%
  slice_head(n = 10)
```

```
# Creating the horizontal bar chart
ggplot(top_products, aes(x = total_sales, y = reorder(product_name, total_sales))) +
  geom_col(fill = "steelblue") +
  scale_x_continuous(labels = label_number(scale_cut = cut_short_scale())) +
  labs(title = "Top 10 Best-Selling Products", x = "Sales ($)", y = "Product Name") +
  theme_minimal()
```



The Canon imageCLASS 2200 Advanced Copier sells twice as much as the second-best product.

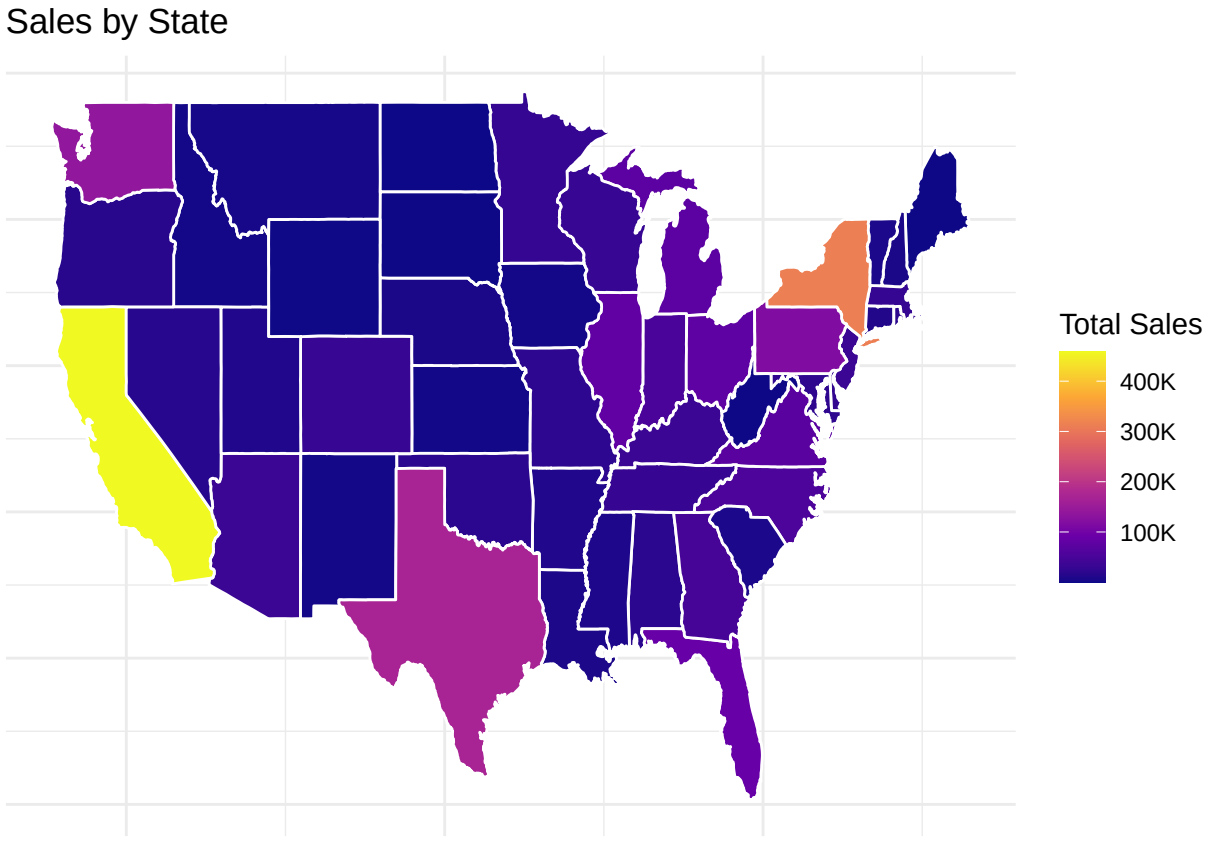
6. Which states have the highest sales?

```r
# Summarizing sales by state
state_sales <- ss %>%
  group_by(state) %>%
  summarise(total_sales = sum(sales, na.rm = TRUE))
```

```r
# Using the 'maps' package to get a map of US states
us_states_map <- map_data("state")
```

```r
# Merging the sales data with the map data (by state)
state_sales$region <- tolower(state_sales$state)  # Ensure state names match
merged_data <- left_join(us_states_map, state_sales, by = "region")
```

```r
# Creating a choropleth map
ggplot(merged_data, aes(x = long, y = lat, group = group, fill = total_sales)) +
  geom_polygon(color = "white") +
  scale_fill_viridis_c(option = "C", labels = label_number(scale_cut = cut_short_scale())) +
  theme_minimal() +
  labs(title = "Sales by State", fill = "Total Sales") +
  theme(axis.text = element_blank(), axis.title = element_blank())
```



California is the state with the highest sales, surpassing $400,000. New York ranks second, with approximately $300,000 in sales.

Recommendations

- Focus on promoting high-profit items like *Phones* while reassessing pricing and cost structures for low-profit sub-categories.
- Reduce investments in unprofitable sub-categories such as Tables, Bookcases, Supplies, and Fasteners or explore cost-cutting measures.
- Make the most of the success of the *Canon imageCLASS 2200 Advanced Copier* by expanding its availability, offering bundle deals, or increasing targeted marketing efforts.
- California and New York are the primaries source of revenue; consider reinforcing marketing campaigns and promotions in these states to further boost sales.
- Identify underperforming states and explore adapted strategies such as localized marketing.
- Investigate why *Tables* generate high sales but result in losses; being it a result of pricing, elevated costs, or substantial discounts.