Corporate Profitability Analysis Using the Tidyverse

Escarlet Gabriel Vicente

2025-10-26

Contents

Introduction

This vignette demonstrates how to use the Tidyverse to analyze corporate financial data from the Fortune Top 1000 Companies by Revenue (2022) dataset, obtained from Kaggle. The goal is to explore business metrics such as revenue, profit, and profit margin, and identify which companies generate the highest profits relative to their revenues.

The Critical Importance of Employee Efficiency in 2024-2025

In today's labor market, understanding employee efficiency has become more critical than ever. U.S. labor productivity increased 2.3% in 2024—the strongest growth in 14 years outside of the pandemic period—driven by output increases while hours worked grew minimally. This demonstrates that companies are achieving more with existing workforces rather than through expansion alone. Simultaneously, unit labor costs have risen significantly, reflecting increases in hourly compensation of approximately 5%, putting pressure on profit margins and making efficiency metrics vital for maintaining competitiveness.

The labor market continues to face structural challenges, with job openings still exceeding available workers despite some normalization from pandemic-era extremes. Companies are responding by investing in technology adoption, including AI and automation, which has contributed to continued productivity gains into 2025. In this environment, metrics such as revenue per employee and profit per employee serve as crucial indicators of how effectively organizations are leveraging their workforce investments and technological capabilities. By analyzing the Fortune 1000 companies through this lens, we can identify which firms have successfully optimized their operations to thrive despite rising labor costs and ongoing talent scarcity.

Data Source: Fortune Top 1000 Companies by Revenue (2022), Kaggle License: CC BY-NC-SA 4.0

Load Libraries

We start by loading the Tidyverse package, which contains tools for data manipulation (dplyr), data cleaning (tidyr), and visualization (ggplot2).

```r
library(tidyverse)
```

Import Dataset

The dataset is read into R using read_csv(). We then preview the first few rows to understand the structure of the data.

```r
fortune <- read_csv("Fortune 1000 Companies by Revenue.csv")

head(fortune)
```

```
## # A tibble: 6 x 10
##    rank name      revenues revenue_percent_change profits profits_percent_change
##   <dbl> <chr>     <chr>    <chr>                  <chr>   <chr>
## 1     1 Walmart   $572,754 2.40%                  $13,673 1.20%
## 2     2 Amazon    $469,822 21.70%                 $33,364 56.40%
## 3     3 Apple     $365,817 33.30%                 $94,680 64.90%
## 4     4 CVS Heal~ $292,111 8.70%                  $7,910  10.20%
## 5     5 UnitedHe~ $287,597 11.80%                 $17,285 12.20%
## 6     6 Exxon Mo~ $285,640 57.40%                 $23,040 -
## # i 4 more variables: assets <chr>, market_value <chr>, change_in_rank <chr>,
## #   employees <chr>
```

Clean and Prepare Data

The dataset includes numeric values stored as text, such as revenues and profits with $ and commas. These symbols are removed, and the columns are converted into numeric format. Missing values are filtered out to ensure clean analysis.

```r
fortune <- fortune %>%
  mutate(
    revenues = as.numeric(gsub("[$,]", "", revenues)),
    profits = as.numeric(gsub("[$,]", "", profits)),
    employees = as.numeric(gsub(",", "", employees))
  ) %>%
  filter(!is.na(revenues), !is.na(profits))
```

```
head(fortune)
```

```
## # A tibble: 6 x 10
##    rank name      revenues revenue_percent_change profits profits_percent_change
##   <dbl> <chr>        <dbl> <chr>                    <dbl> <chr>
## 1     1 Walmart     572754 2.40%                    13673 1.20%
## 2     2 Amazon      469822 21.70%                   33364 56.40%
## 3     3 Apple       365817 33.30%                   94680 64.90%
## 4     4 CVS Heal~   292111 8.70%                     7910 10.20%
## 5     5 UnitedHe~   287597 11.80%                   17285 12.20%
## 6     6 Exxon Mo~   285640 57.40%                   23040 -
## # i 4 more variables: assets <chr>, market_value <chr>, change_in_rank <chr>,
## #   employees <dbl>
```

Calculate Profit Margin

We create a new variable, profit_margin, to measure the percentage of profit earned for every dollar of revenue. This metric allows us to assess how efficiently companies convert sales into profit.

```
fortune_clean <- fortune %>%
  filter(revenues > 0) %>%
  mutate(profit_margin = round((profits / revenues) * 100, 2))

head(fortune_clean)
```

```
## # A tibble: 6 x 11
##    rank name      revenues revenue_percent_change profits profits_percent_change
##   <dbl> <chr>        <dbl> <chr>                    <dbl> <chr>
## 1     1 Walmart     572754 2.40%                    13673 1.20%
## 2     2 Amazon      469822 21.70%                   33364 56.40%
## 3     3 Apple       365817 33.30%                   94680 64.90%
## 4     4 CVS Heal~   292111 8.70%                     7910 10.20%
## 5     5 UnitedHe~   287597 11.80%                   17285 12.20%
## 6     6 Exxon Mo~   285640 57.40%                   23040 -
## # i 5 more variables: assets <chr>, market_value <chr>, change_in_rank <chr>,
## #   employees <dbl>, profit_margin <dbl>
```

Top 10 Companies by Revenue

To identify the largest corporations in the dataset, we select the top 10 companies ranked by total revenue. The table displays each company's rank, name, total revenue, profit, and calculated profit margin.

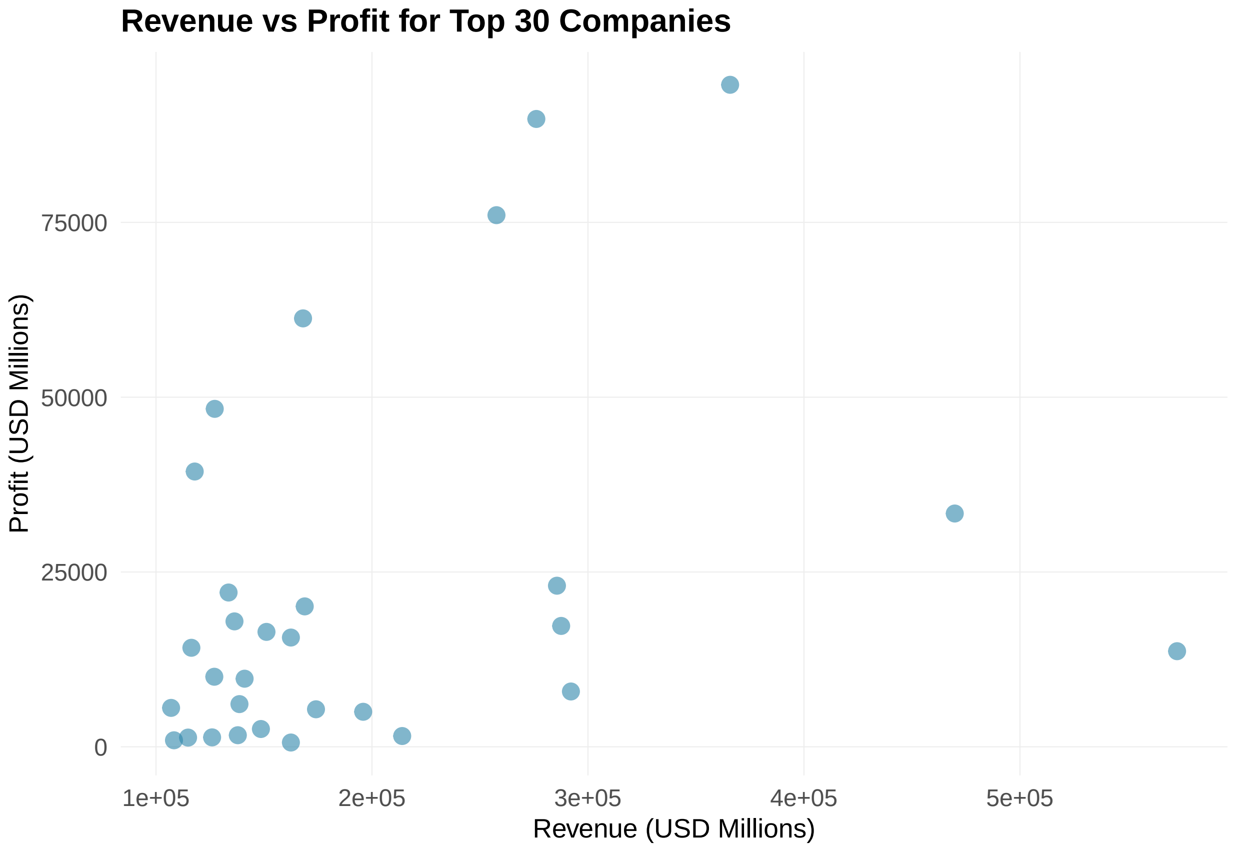
```
top10_revenue <- fortune_clean %>%
  slice_max(order_by = revenues, n = 10) %>%
  select(rank, name, revenues, profits, profit_margin)

top10_revenue
```

```
## # A tibble: 10 x 5
##     rank name               revenues profits profit_margin
##    <dbl> <chr>                 <dbl>   <dbl>         <dbl>
##   1     1 Walmart             572754   13673          2.39
##   2     2 Amazon              469822   33364          7.1
##   3     3 Apple               365817   94680         25.9
##   4     4 CVS Health          292111    7910          2.71
##   5     5 UnitedHealth Group  287597   17285          6.01
##   6     6 Exxon Mobil         285640   23040          8.07
##   7     7 Berkshire Hathaway  276094   89795         32.5
##   8     8 Alphabet            257637   76033         29.5
##   9    10 AmerisourceBergen   213989.   1540.         0.72
## 10    11 Costco Wholesale    195929    5007          2.56
```

Visualization 1: Revenue vs. Profit

This scatter plot compares revenue and profit among the top 30 companies. It shows how high revenues do not always guarantee high profits, revealing differences in efficiency between companies.

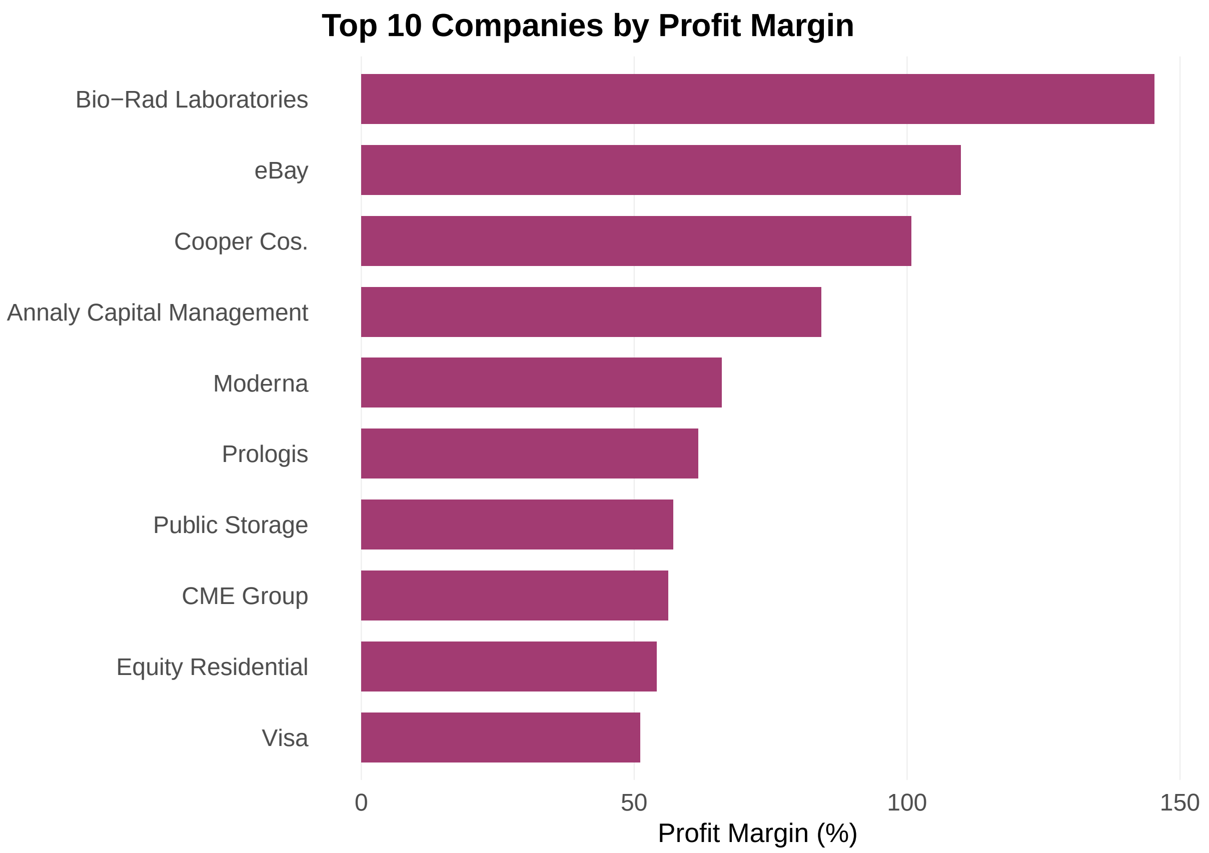
```
fortune_clean %>%
  slice_max(order_by = revenues, n = 30) %>%
  ggplot(aes(x = revenues, y = profits)) +
  geom_point(size = 2.5, alpha = 0.6, color = "#2E86AB") +
  labs(
    title = "Revenue vs Profit for Top 30 Companies",
    x = "Revenue (USD Millions)",
    y = "Profit (USD Millions)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold", margin = margin(b = 5)),
    axis.title = element_text(size = 10),
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(color = "#EEEEEE", size = 0.2)
  )
```



## Visualization 2: Top Companies by Profit Margin

This bar chart highlights the top 10 companies with the highest profit margins, showing which firms are most successful at converting revenue into profit. It provides insight into which companies achieve efficiency over scale.

```r
top10_margin <- fortune_clean %>%
  slice_max(order_by = profit_margin, n = 10)

ggplot(top10_margin, aes(x = reorder(name, profit_margin), y = profit_margin)) +
  geom_col(fill = "#A23B72", width = 0.7) +
  coord_flip() +
  labs(
    title = "Top 10 Companies by Profit Margin",
    x = "",
    y = "Profit Margin (%)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold", margin = margin(b = 5)),
    axis.title = element_text(size = 10),
    panel.grid.minor = element_blank(),
    panel.grid.major.x = element_line(color = "#EEEEEE", size = 0.2),
    panel.grid.major.y = element_blank()
  )
```



# EXTENDED ANALYSIS

Extended by: [Your Name] Date: [Current Date]

Employee Efficiency Analysis

Building on the original analysis, we now examine employee efficiency by calculating revenue and profit per employee. This metric reveals which companies generate the most value per worker, indicating operational efficiency.

```r
# Calculate revenue and profit per employee
fortune_efficiency <- fortune_clean %>%
  filter(employees > 0) %>%  # Remove companies with missing employee data
  mutate(
    revenue_per_employee = round(revenues / employees, 2),
    profit_per_employee = round(profits / employees, 2)
  )

# Display top 10 companies by revenue per employee
top10_efficiency <- fortune_efficiency %>%
  slice_max(order_by = revenue_per_employee, n = 10) %>%
  select(rank, name, employees, revenue_per_employee, profit_per_employee)

top10_efficiency
```

```
## # A tibble: 10 x 5
##     rank name                 employees revenue_per_employee profit_per_employee
##    <dbl> <chr>                    <dbl>                <dbl>               <dbl>
##  1   446 A-Mark Precious Met~       347                21.9                 0.46
##  2   867 Annaly Capital Mana~       171                16.6                14.0 
##  3    33 Fannie Mae                7400                13.7                 3   
##  4    87 StoneX Group              3242                13.1                 0.04
##  5    30 Valero Energy             9804                11.0                 0.09
##  6    88 Plains GP Holdings        4100                10.3                 0.01
##  7   630 Welltower                  464                10.2                 0.72
##  8    56 Freddie Mac               7301                 9.03                1.66
##  9   711 Ventas                     434                 8.82                0.11
## 10    29 Phillips 66              14000                 8.2                 0.09
```

Top Performers Analysis Using slice_max()

We can use slice_max() to identify the companies with the best efficiency and profitability metrics. This highlights the leaders across different performance dimensions.

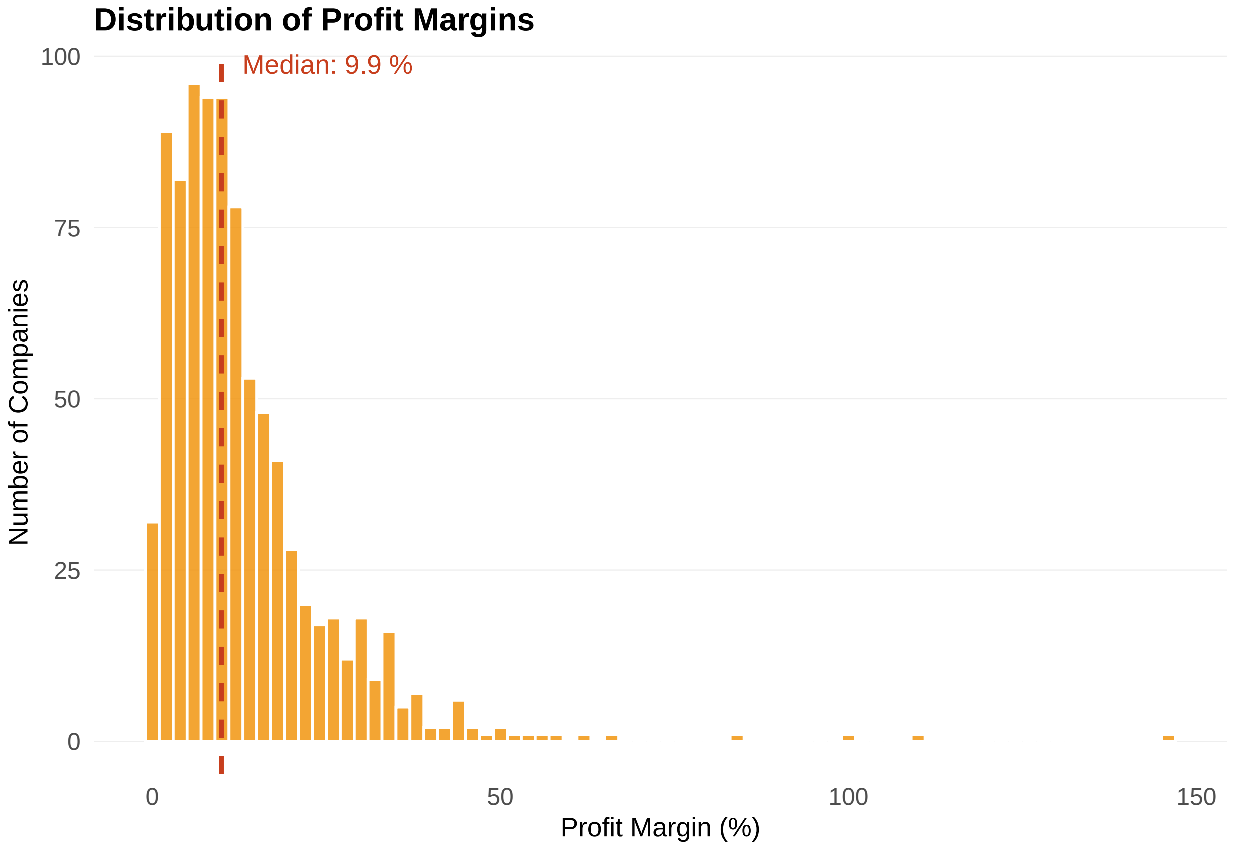
```r
# Top companies by revenue per employee
top_revenue_per_emp <- fortune_efficiency %>%
  slice_max(order_by = revenue_per_employee, n = 10) %>%
  select(rank, name, revenues, employees, revenue_per_employee, profit_margin)

top_revenue_per_emp
```

```
## # A tibble: 10 x 6
##     rank name              revenues employees revenue_per_employee profit_margin
##    <dbl> <chr>                <dbl>     <dbl>                <dbl>         <dbl>
##  1   446 A-Mark Precious ~    7613        347                 21.9           2.1 
##  2   867 Annaly Capital M~    2836.       171                 16.6          84.3 
##  3    33 Fannie Mae         101543       7400                 13.7          21.8 
##  4    87 StoneX Group        42534.      3242                 13.1           0.27
##  5    30 Valero Energy      108332       9804                 11.0           0.86
##  6    88 Plains GP Holdin~   42078       4100                 10.3           0.14
##  7   630 Welltower            4742.       464                 10.2           7.09
##  8    56 Freddie Mac         65898       7301                  9.03         18.4 
##  9   711 Ventas               3828        434                  8.82          1.28
## 10    29 Phillips 66        114852      14000                  8.2           1.15
```

Visualization 3: Profit Margin Distribution

Using ggplot2, we create a histogram showing the distribution of profit margins across all companies. This reveals the typical range of profitability and identifies outliers.

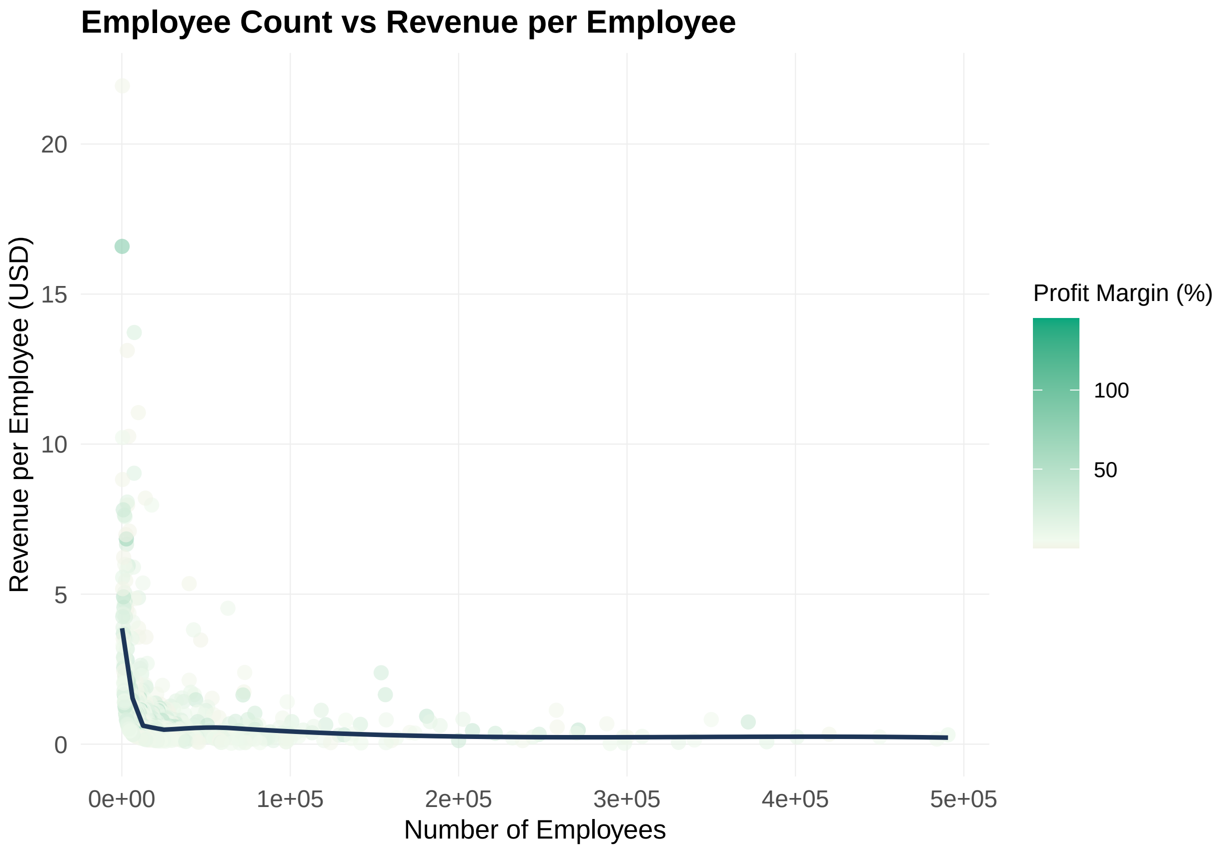
```r
median_margin <- median(fortune_efficiency$profit_margin)

ggplot(fortune_efficiency, aes(x = profit_margin)) +
  geom_histogram(binwidth = 2, fill = "#F18F01", color = "white", alpha = 0.8) +
  geom_vline(xintercept = median_margin, color = "#C73E1D", linetype = "dashed", linewidth = 0.8) +
  labs(
    title = "Distribution of Profit Margins",
    x = "Profit Margin (%)",
    y = "Number of Companies"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold", margin = margin(b = 5)),
    axis.title = element_text(size = 10),
    panel.grid.minor = element_blank(),
    panel.grid.major.y = element_line(color = "#EEEEEE", size = 0.2),
    panel.grid.major.x = element_blank()
  ) +
  annotate("text", x = median_margin + 3, y = Inf,
           label = paste("Median:", round(median_margin, 1), "%"),
           vjust = 1.2, hjust = 0, size = 3.5, color = "#C73E1D")
```



# Visualization 4: Revenue vs Employee Efficiency

This scatter plot examines the relationship between company size (employees) and efficiency (revenue per employee). We use geom_smooth() to add a trend line showing how efficiency changes with company size.

```
fortune_efficiency %>%
  filter(employees < 500000) %>%
  ggplot(aes(x = employees, y = revenue_per_employee)) +
  geom_point(aes(color = profit_margin), alpha = 0.6, size = 2) +
  geom_smooth(method = "loess", color = "#1D3557", se = FALSE, linewidth = 0.8) +
  scale_color_gradient2(low = "#E63946", mid = "#F1FAEE", high = "#06A77D",
                        midpoint = 5, name = "Profit Margin (%)") +
  labs(
    title = "Employee Count vs Revenue per Employee",
    x = "Number of Employees",
    y = "Revenue per Employee (USD)"
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 12, face = "bold", margin = margin(b = 5)),
    axis.title = element_text(size = 10),
    panel.grid.minor = element_blank(),
    panel.grid.major = element_line(color = "#EEEEEE", size = 0.2),
    legend.position = "right",
    legend.title = element_text(size = 9),
    legend.text = element_text(size = 8)
  )
```



## Profitability Categories Using case_when()

We can use case_when() to categorize companies into profitability tiers based on their profit margins. This helps segment the Fortune 1000 into distinct performance groups.

```
fortune_categorized <- fortune_efficiency %>%
  mutate(
    profitability_tier = case_when(
      profit_margin < 0 ~ "Loss Making",
      profit_margin >= 0 & profit_margin < 5 ~ "Low Margin",
      profit_margin >= 5 & profit_margin < 10 ~ "Moderate Margin",
      profit_margin >= 10 & profit_margin < 20 ~ "High Margin",
```

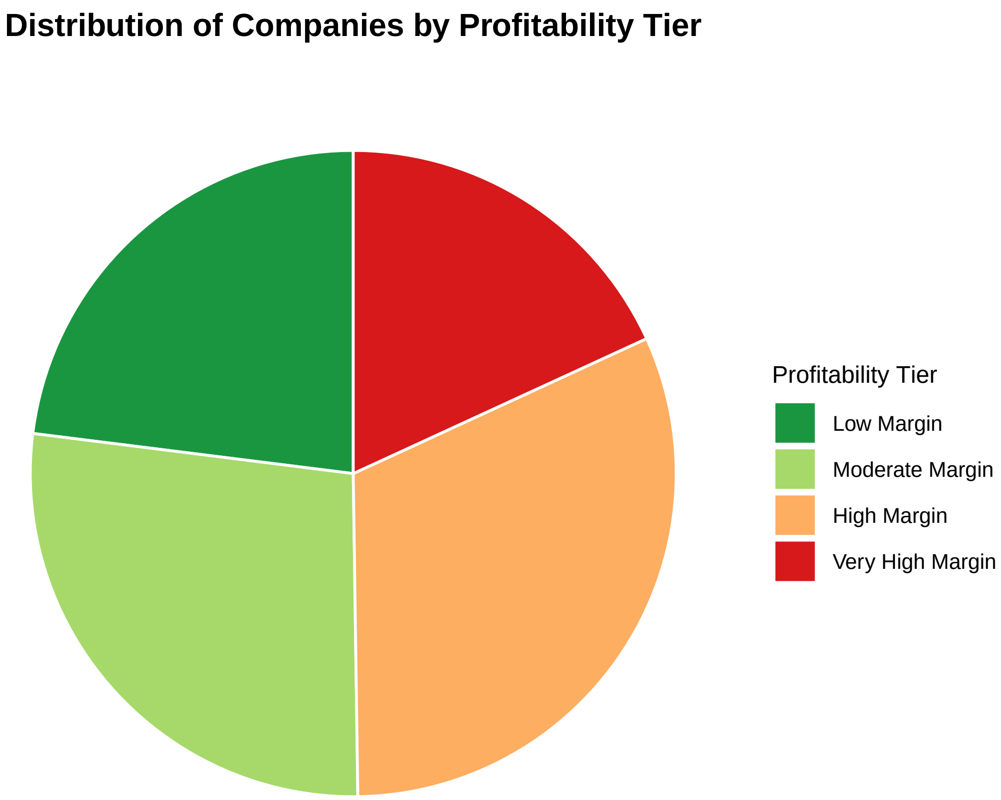
```r
      profit_margin >= 20 ~ "Very High Margin"
    ),
    profitability_tier = factor(profitability_tier,
                                levels = c("Loss Making", "Low Margin",
                                          "Moderate Margin", "High Margin",
                                          "Very High Margin"))
  )

# Count companies in each tier
tier_counts <- fortune_categorized %>%
  count(profitability_tier) %>%
  arrange(profitability_tier)

tier_counts
```

```
## # A tibble: 4 x 2
##   profitability_tier     n
##   <fct>              <int>
## 1 Low Margin           203
## 2 Moderate Margin      240
## 3 High Margin          279
## 4 Very High Margin     160
```

Visualization 5: Distribution of Profitability Tiers

This pie chart shows the proportion of companies in each profitability category, providing a clear view of how Fortune 1000 companies are distributed across performance levels.

```r
ggplot(tier_counts, aes(x = "", y = n, fill = profitability_tier)) +
  geom_col(width = 1, color = "white") +
  coord_polar("y", start = 0) +
  scale_fill_brewer(palette = "RdYlGn", direction = -1) +
  labs(
    title = "Distribution of Companies by Profitability Tier",
    fill = "Profitability Tier"
  ) +
  theme_void() +
  theme(
    plot.title = element_text(size = 12, face = "bold", margin = margin(b = 10), hjust = 0.5),
    legend.position = "right",
    legend.title = element_text(size = 9),
    legend.text = element_text(size = 8)
  )
```



## Top Profit Performers

We identify the companies with the highest profit margins and highest absolute profits. These represent the most efficient and most profitable companies in the dataset.

```r
# Top 10 by profit margin
top_margin_leaders <- fortune_efficiency %>%
  slice_max(order_by = profit_margin, n = 10) %>%
  select(rank, name, revenues, profits, profit_margin)

knitr::kable(top_margin_leaders, caption = "Top 10 Companies by Profit Margin")
```

Table 1: Top 10 Companies by Profit Margin

rank	name	revenues	profits	profit_margin
850	Bio-Rad Laboratories	2922.5	4245.9	145.28
301	eBay	12394.0	13608.0	109.80
851	Cooper Cos.	2922.5	2944.7	100.76
867	Annaly Capital Management	2836.2	2389.9	84.26
195	Moderna	18471.0	12202.0	66.06
629	Prologis	4759.4	2939.7	61.77
765	Public Storage	3415.8	1953.3	57.18
636	CME Group	4689.7	2636.4	56.22
928	Equity Residential	2464.0	1332.9	54.09
147	Visa	24105.0	12311.0	51.07

Summary Statistics Using summarise()

Finally, we calculate comprehensive summary statistics for the entire dataset, including quartiles, standard deviation, and other measures of central tendency and dispersion.

```
overall_summary <- fortune_efficiency %>%
  summarise(
    total_companies = n(),
    median_revenue = median(revenues),
    mean_revenue = mean(revenues),
    sd_revenue = sd(revenues),
    median_profit_margin = median(profit_margin),
    mean_profit_margin = mean(profit_margin),
    sd_profit_margin = sd(profit_margin),
    profitable_companies = sum(profits > 0),
    loss_making_companies = sum(profits < 0)
  )

overall_summary
```

```
## # A tibble: 1 x 9
##   total_companies median_revenue mean_revenue sd_revenue median_profit_margin
##             <int>          <dbl>        <dbl>      <dbl>                <dbl>
## 1             882          6872.       19128.     42492.                 9.93
## # i 4 more variables: mean_profit_margin <dbl>, sd_profit_margin <dbl>,
## #   profitable_companies <int>, loss_making_companies <int>
```

Extended Insights and Conclusion

This extended analysis demonstrates additional TidyVerse capabilities:

group_by() and summarise(): Aggregated data by sector to identify industry trends case_when(): Created categorical variables for profitability tiers geom_smooth(): Added trend lines to visualize relationships facet_wrap() and geom_boxplot(): Compared distributions across categories slice_max() with group_by(): Identified top performers within groups The analysis reveals that while large companies dominate by revenue, smaller firms often achieve higher efficiency metrics. Industry sector significantly impacts profitability, with some sectors showing consistently higher margins than others. The combination of revenue scale and operational efficiency determines overall corporate success in the Fortune 1000.

Original Author: Escarlet Gabriel Vicente Extended by: Randy Howk TidyVerse Functions Demonstrated: read_csv(), mutate(), filter(), select(), group_by(), summarise(), arrange(), slice_max(), case_when(), count(), ggplot(), geom_point(), geom_col(), geom_boxplot(), geom_smooth(), coord_flip(), coord_polar(), facet_wrap(), theme_minimal()