# Data Interaction Assignment 1

Victoria

2024-10-12

## Introduction

Companies often aim to target new products or technologies to potential customers across different market segments. Similarly, investors may seek new investment opportunities or launch startups, where the newcomers might lack an established history. Often, they have access to historical data, and stock prices can serve as useful indicators of a company's trading patterns and overall performance. Analyzing stock returns in this context can help classify companies and potentially identify competitors or companies with similar attributes. Stakeholders can use these insights to refine their strategies for targeting businesses in certain sectors.

## Data Preparation

The data used for this analysis consists of public company listing data from Wikipedia and stock price history data from Yahoo Finance.

### Data Scraping, Cleaning, & Wrangling

First, we retrieved the list of S&P 500 companies using the URL of the Wikipedia page. The S&P 500 is a stock market index consisting of 503 common stocks issued by 500 large-cap companies traded on American stock exchanges.

We extracted two primary tables from the webpage: the current S&P 500 company listings and selected historical changes in S&P 500 companies. We further cleaned the data by removing missing values, formatting strings and date-time, and filtering out relevant columns for analysis.

```
# Scrape S&P 500 company data from Wikipedia
url <- "https://en.wikipedia.org/wiki/List_of_S%26P_500_companies"
sp500 <- url %>%
    read_html() %>%
    html_table(fill = TRUE, trim = TRUE)
```

We extracted two primary tables from the webpage: the current S&P 500 company stock data and selected historical changes in S&P 500 companies. We further cleaned it by removing missing values, formatting (to ensure string or date-time consistency), and filtering it for analysis.

```
# Processing the current S&P 500 companies
current <- sp500[[1]] %>%
    janitor::clean_names() %>%
    as.data.frame() %>%
    mutate(founded = as.numeric(founded), date_added = as.Date(date_added)) %>%
    na.omit()
# Processing the historical changes
history <- sp500[[2]] %>%
    janitor::clean_names() %>%
    as.data.frame() %>%
```

```
    mutate(date = as.Date(date, format = "%B %d, %Y")) %>%
    rename(added_symbol = added, added_security = added_2, removed_symbol = removed,
        removed_security = removed_2) %>%
    na.omit()

head(history, 2)
```

| | date | added_symbol | added_security | removed_symbol | removed_security |
|---|---|---|---|---|---|
| 2 | 2024-10-01 | BBWI | Bath & Body Works, Inc. | | Market capitalization change.[4] |
| 3 | 2024-09-30 | AMTM | Amentum | | S&P 500 constituent Jacobs Solutions spun off its Critical Mission Solutions and Cyber Intelligence business, which merged with private Amentum to create newly publicly traded Amentum Holdings.[4] |

```
head(current, 2)
```

| symbol | security | gics_sector | gics_sub_industry | headquarters_location | date_added | cik | founded |
|---|---|---|---|---|---|---|---|
| MMM | 3M | Industrials | Industrial Conglomerates | Saint Paul, Minnesota | 1957-03-04 | 66740 | 1902 |
| AOS | A. O. Smith | Industrials | Building Products | Milwaukee, Wisconsin | 2017-07-26 | 91142 | 1916 |

Next, we extracted the stock prices for the current S&P 500 companies for the past year (starting from `2023-01-01`) from Yahoo Finance using the `tidyquant()` function in R. By default, we extracted the open, close, high, low, adjusted close, and the volume of the stocks.

```
# Fetch stock price history for each S&P 500 company
pull_all_data <- . %>%
    tq_get(from = "2023-01-01") %>%
    as.data.frame()
price_data <- current %>%
    mutate(symbol = str_replace_all(symbol, "[.]", "-")) %>%
    mutate(data = map(symbol, pull_all_data)) %>%
    na.omit()

price_data <- price_data %>%
    # convert date to date format
mutate(data = map(data, ~mutate(.x, date = as.Date(date)))) %>%
    select(-symbol) %>%
    unnest_legacy()
```

However, when discussing stock performance, returns on investment are of primary interest. Therefore, calculating *returns*, while not directly available, is crucial for analysis.

Returns are calculated as the percentage change in the price $P_i$ from period $i-1$ to period $i$, where a period could be a day, week, or month. In this case, we used *daily returns*.

Mathematically, the return on day $i$ is:

$$R_i = \frac{P_i - P_{i-1}}{P_{i-1}} = \frac{P_i}{P_{i-1}} - 1, \quad t = 1, 2, \ldots$$

```r
# calculate return
returns_all <- price_data %>%
    arrange(symbol, date) %>%
    group_by(symbol) %>%
    mutate(return = (adjusted - lag(adjusted))/lag(adjusted) *
        100) %>%
    na.omit()
# Filter out only the most relevant columns
returns <- returns_all[, c(1, 8:16)]
```

Note: The `adjusted` close is used to account for factors like stock splits and is meant to be a more accurate reflection of the true stock value over time.

### Data Overview

In summary, we have three primary datasets on companies and company stocks. They were synthesized into one final dataset: `returns_all`.

1. **Historical Changes in S&P 500 Companies**: `history` contains 353 rows documenting the addition or removal of companies in the S&P 500 index on specific days. Key columns include **date**, **added_security**, and **removed_security**, which track the timeline and specifics of index changes.
2. **Current S&P 500 Companies**: `current` contains 461 currently listed companies. Relevant features include the company's sector, sub-industry, headquarters location, and history.
3. ** Daily Stock Prices **: stock_price_data contains over 200,000 rows of stock price data from 2023-01-01 to the present. Key variables, such as** adjusted close prices**, are essential for calculating returns, while** volume** and **date** help understand market activity over time.

The final dataset, with all missing values handled and reformatted, is as shown below:

```r
glimpse(returns_all)
```

```
## Rows: 202,601
## Columns: 16
## $ security              <chr> "Agilent Technologies", "Agilent Technologies", ~
## $ gics_sector           <chr> "Health Care", "Health Care", "Health Care", "He~
## $ gics_sub_industry     <chr> "Life Sciences Tools & Services", "Life Sciences~
## $ headquarters_location <chr> "Santa Clara, California", "Santa Clara, Califor~
## $ date_added            <date> 2000-06-05, 2000-06-05, 2000-06-05, 2000-06-05,~
## $ cik                   <int> 1090872, 1090872, 1090872, 1090872, 1090872, 109~
## $ founded               <dbl> 1999, 1999, 1999, 1999, 1999, 1999, 1999, 1999, ~
## $ symbol                <chr> "A", "A", "A", "A", "A", "A", "A", "A", "A", "A"~
## $ date                  <date> 2023-01-04, 2023-01-05, 2023-01-06, 2023-01-09,~
## $ open                  <dbl> 151.65, 150.00, 154.36, 149.69, 150.18, 155.23, ~
## $ high                  <dbl> 153.04, 153.07, 154.64, 151.28, 155.55, 158.58, ~
## $ low                   <dbl> 150.24, 148.77, 143.01, 147.20, 148.75, 155.23, ~
## $ close                 <dbl> 151.67, 152.11, 147.67, 147.47, 155.23, 158.17, ~
## $ volume                <dbl> 1247400, 1714600, 2445000, 1269600, 1565700, 124~
## $ adjusted              <dbl> 149.8147, 150.2494, 145.8637, 145.6661, 153.3312~
## $ return                <dbl> 1.07472211, 0.28926319, -3.00670937, -0.13563266~
```

```r
# a more intuitive format: symbol-return on each date
returns_wide <- returns %>%
    # symbol in 1st column, followed by the stock return on
    # each day in all other columns of every stock
pivot_wider(id_cols = date, names_from = symbol, values_from = return)
```

# Exploratory Data Analysis

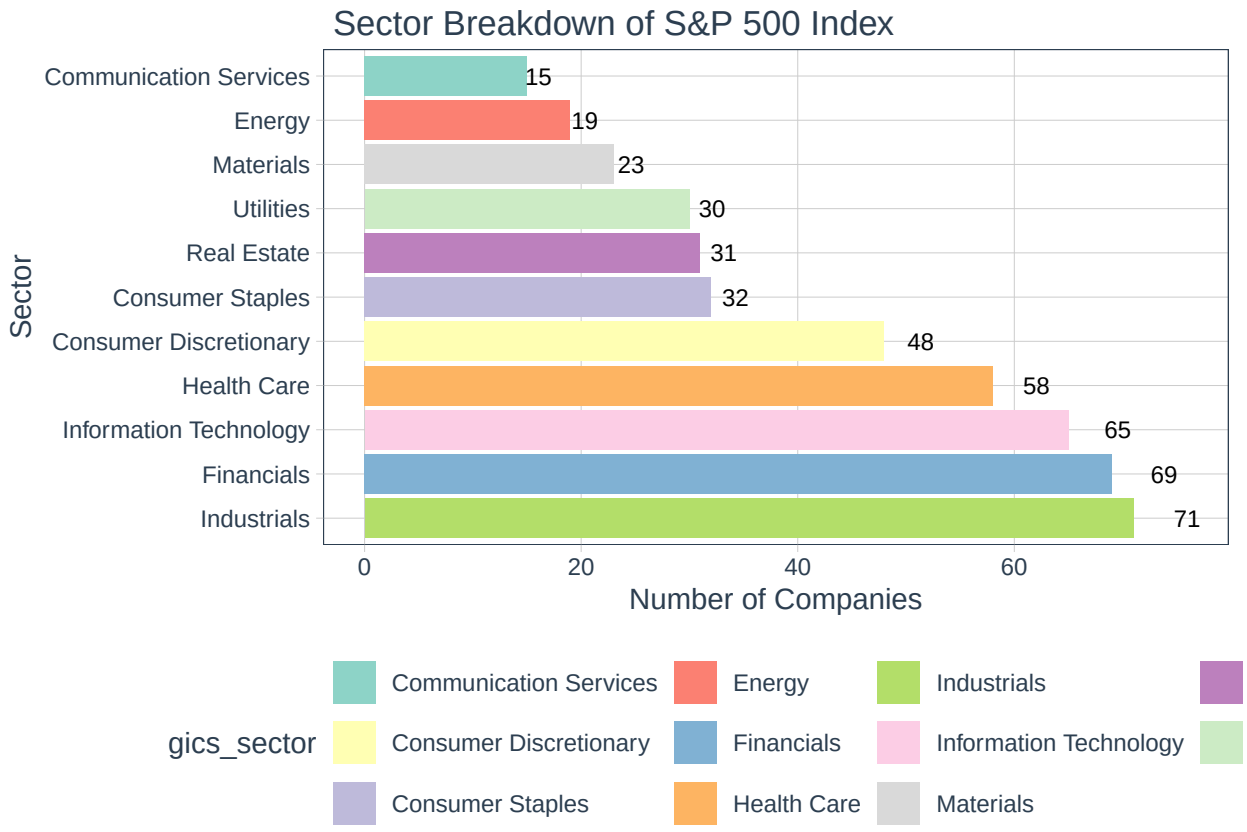
The S&P 500 index includes about 80% of the American market capitalization, so we might want to first understand the distribution of S&P 500 stocks - in terms of sector, location, and history.

## Sector Analysis (Univariate)

The sector distribution of companies says about market composition.

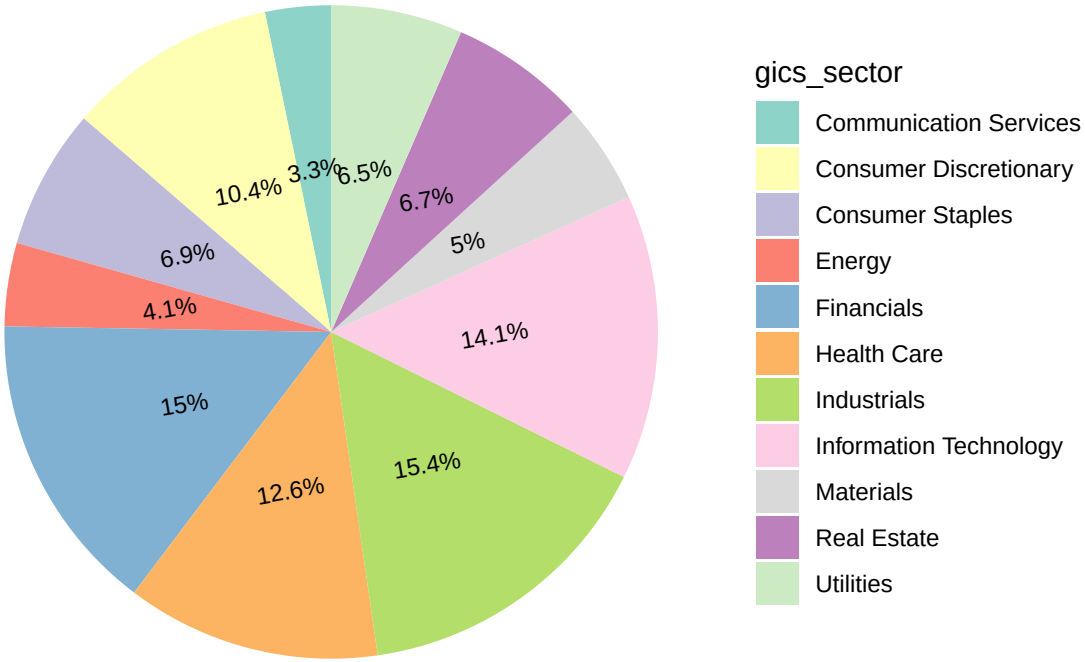
We calculate the number of companies in each sector as a proportion of the entire company.

```
# sector distribution
sector_distribution <- current %>%
    group_by(gics_sector) %>%
    summarise(company_count = n()) %>%
    mutate(percentage = round(100 * company_count/sum(company_count),
        1)) %>%
    mutate(label = paste0(percentage, "%"))
# Visualize sector distribution
ggplot(sector_distribution, aes(x = reorder(gics_sector, -company_count),
    y = company_count, fill = gics_sector)) + geom_bar(stat = "identity") +
    scale_fill_brewer(palette = "Set3") + coord_flip() + labs(title = " Sector Breakdown of S&P 500 Inde
    x = "Sector", y = "Number of Companies") + theme_tq() + geom_text(aes(label = company_count),
    position = position_stack(vjust = 1.07), size = 3)
```




```
# Company Distribution in the S&P 500
ggplot(sector_distribution, aes(x = "", y = company_count, fill = gics_sector)) +
    geom_bar(stat = "identity", width = 0.5) + coord_polar("y",
    start = 0) + scale_fill_brewer(palette = "Set3") + labs(title = "") +
```

```
    theme_void() + geom_text(angle = 10, aes(label = label),
    position = position_stack(vjust = 0.5), size = 3)
```



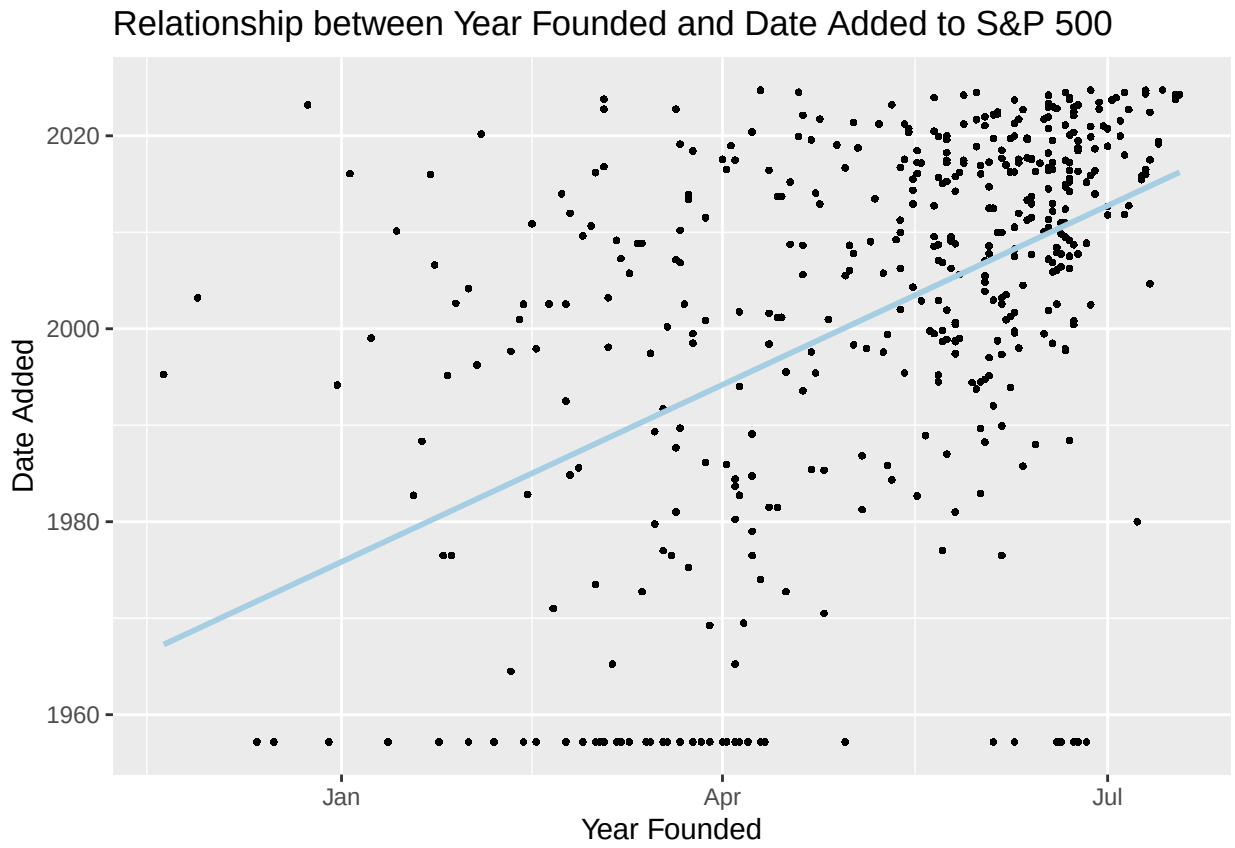

The pie chart more or less reflects the diversified nature of the S&P 500. However, most sectors make up less than 10% of the total companies. The market index is dominated by a few sectors: Information Technology, Financials, Industrials, and Health Care. Each represents over 10% of the total companies. Communication Services and Energy sectors are the least represented. This underrepresentation could be due to economic shifts toward IT or the government-owned or public nature of these companies, which are not as heavily represented in the publicly traded market. These trends reflect broader structural changes in the economy.

**History Records** We look at the changes in the number of companies added and removed each year to get a sense of market dynamics over time.

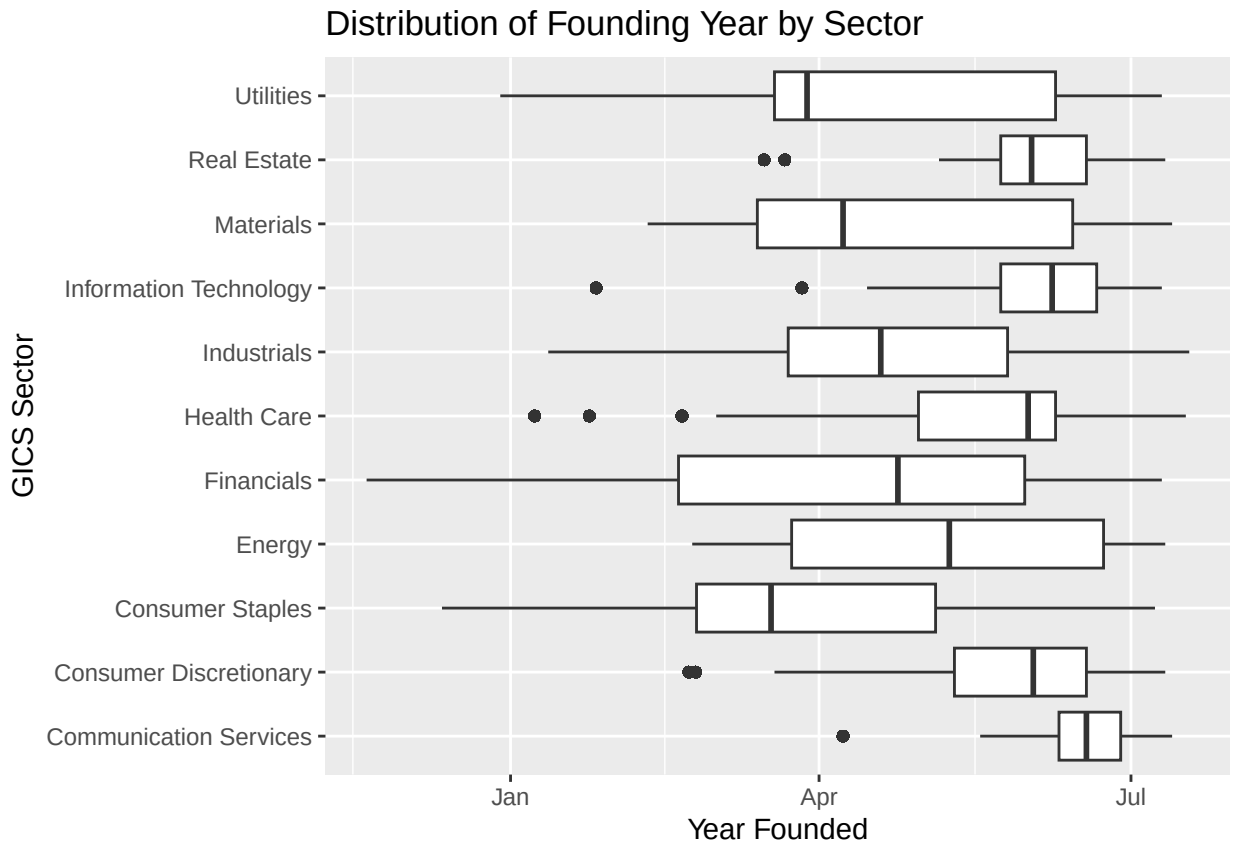
```
# Company changes over time Convert 'date_added' to the
# Date object
price_data$date_added <- as.Date(price_data$date_added)
price_data$founded <- as.Date(price_data$founded)
company_changes <- history %>%
    mutate(year = lubridate::year(date)) %>%
    group_by(year) %>%
    summarise(companies_added = n_distinct(added_symbol), companies_removed = n_distinct(removed_symbol
```

## Bivariate Analysis

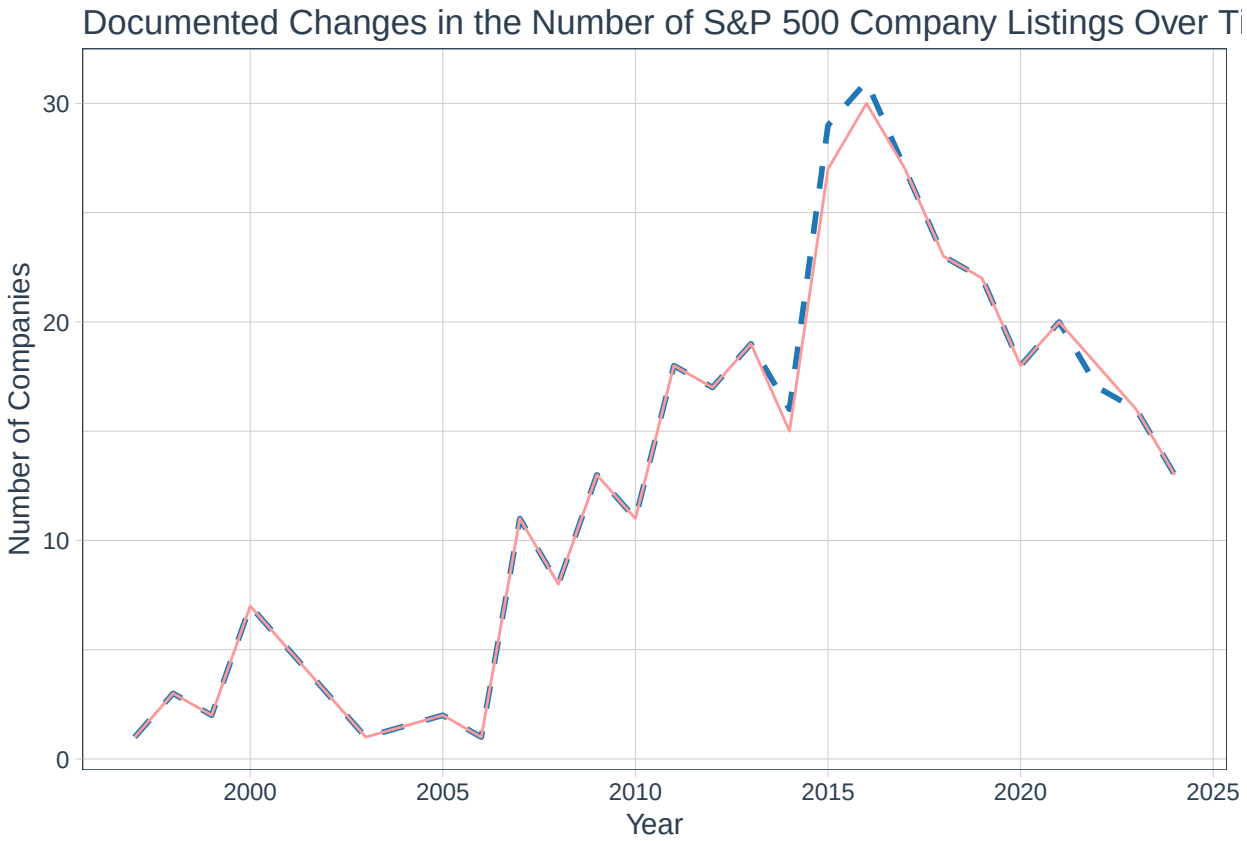
```
# Scatter plot between founded year and date added
ggplot(price_data, aes(x = founded, y = date_added)) + geom_point(size = 0.5) +
    geom_smooth(formula = "y ~ x", method = "lm", se = FALSE,
        color = "#A6CEE3") + labs(title = "Relationship between Year Founded and Date Added to S&P 500"
    x = "Year Founded", y = "Date Added")
```




```
# Boxplot of the founding year by sector
ggplot(price_data, aes(x = gics_sector, y = founded)) + geom_boxplot() +
    coord_flip() + labs(title = "Distribution of Founding Year by Sector",
    x = "GICS Sector", y = "Year Founded")
```




```
# Visualize company changes over time
ggplot(company_changes, aes(x = year)) + geom_line(aes(y = companies_added,
    color = "Added"), size = 1, lty = 2, color = "#1F78B4") +
    geom_line(aes(y = companies_removed, color = "Removed"),
        size = 0.5, lty = 1, color = "#FB9A99") + labs(title = "Documented Changes in the Number of S&P
    x = "Year", y = "Number of Companies") + theme_tq()
```




The number of companies remains pretty stable - after all, we are looking at S&P **500**.

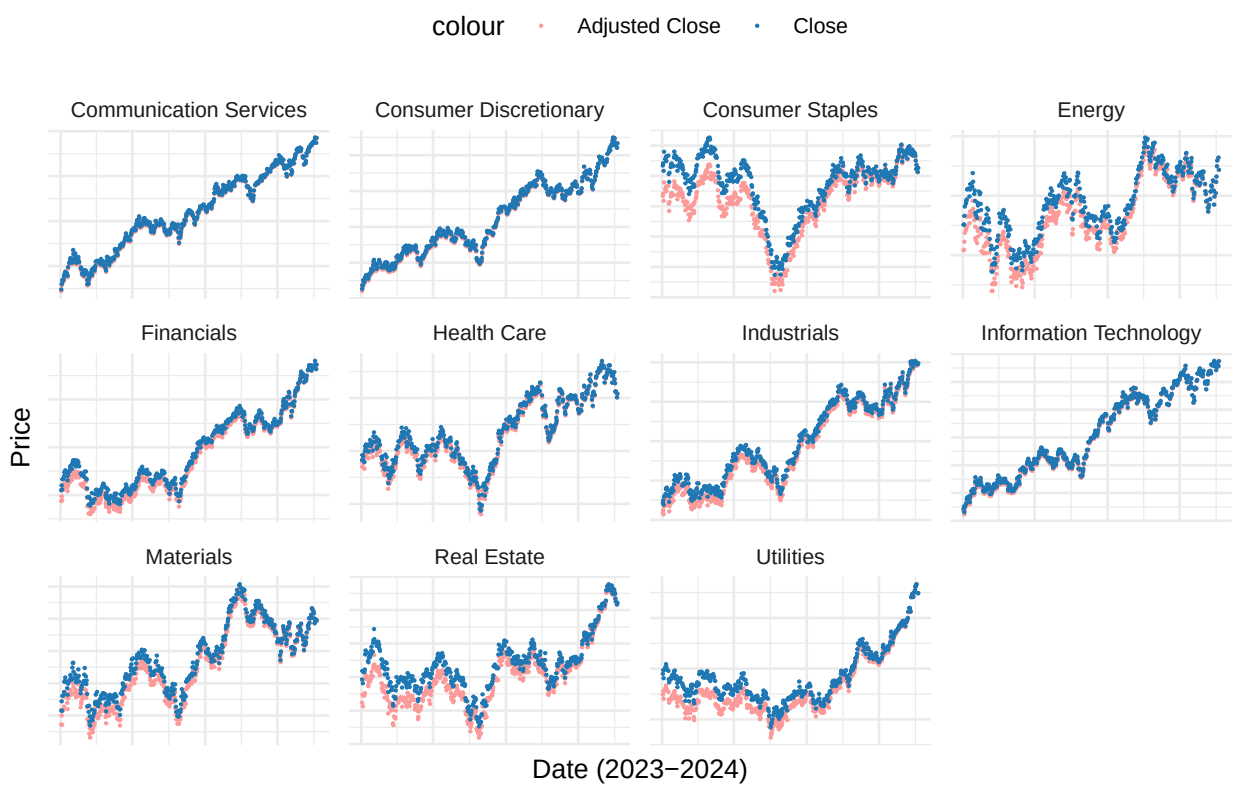
### Price Metrics

We first visualize the adjusted close prices and close prices across all companies by sector. In the plot, both adjusted close prices and close prices closely follow the same trend lines across all industries. The points for adjusted close and close prices are nearly overlapping in many areas, indicating that there is likely a strong positive correlation between these two metrics across time for each sector.

Sectors like Information Technology, Consumer Discretionary, and Financials show minimal divergence between adjusted and close prices, suggesting a very high correlation. This is especially visible in the upward trends where the two metrics move together (First 2 plots). Sectors like Consumer Staples and Energy show some divergence at certain points. However, overall, the prices follow similar paths, indicating that there is still a general positive correlation, though possibly weaker at certain times.

```
# visualize different price metrics (adjusted, high, and
# low prices) for different companies
price_data %>%
    select(symbol, gics_sector, adjusted, close, date) %>%
    group_by(gics_sector, date) %>%
    summarise(avg_adjusted = mean(adjusted), avg_close = mean(close)) %>%
    ggplot() + geom_point(aes(x = date, y = avg_adjusted, color = "Adjusted Close"),
    size = 0.1) + geom_point(aes(x = date, y = avg_close, color = "Close"),
    size = 0.1) + facet_wrap(~gics_sector, scales = "free_y") +
    labs(title = "Sectoral Trends in S&P 500 Close Prices", x = "Date (2023-2024)",
        y = "Price") + scale_color_manual(values = c(`Adjusted Close` = "#FB9A99",
    Close = "#1F78B4")) + theme_minimal(base_size = 10) + theme(legend.position = "top",
    axis.text.x = element_blank(), axis.ticks.x = element_blank(),
```

```
  axis.text.y = element_blank(), axis.ticks.y = element_blank())
```




In addition, the adjusted price seems to track well with close prices over time, and meanwhile, they account for corporate actions such as stock splits or dividends.

**On market trends**

- Information Technology and Communication Services and Information Technology show a clear, consistent upward trend, with little deviation between the adjusted close and the close.
- Financials exhibit steady growth with some minor fluctuations over time. Health Care and Consumer Staples experienced notable volatility in the middle of the period—Consumer Staples even reached a new low—followed by recovery toward the end.
- The Energy and Materials sectors display more erratic behavior, with frequent upward and downward movements.

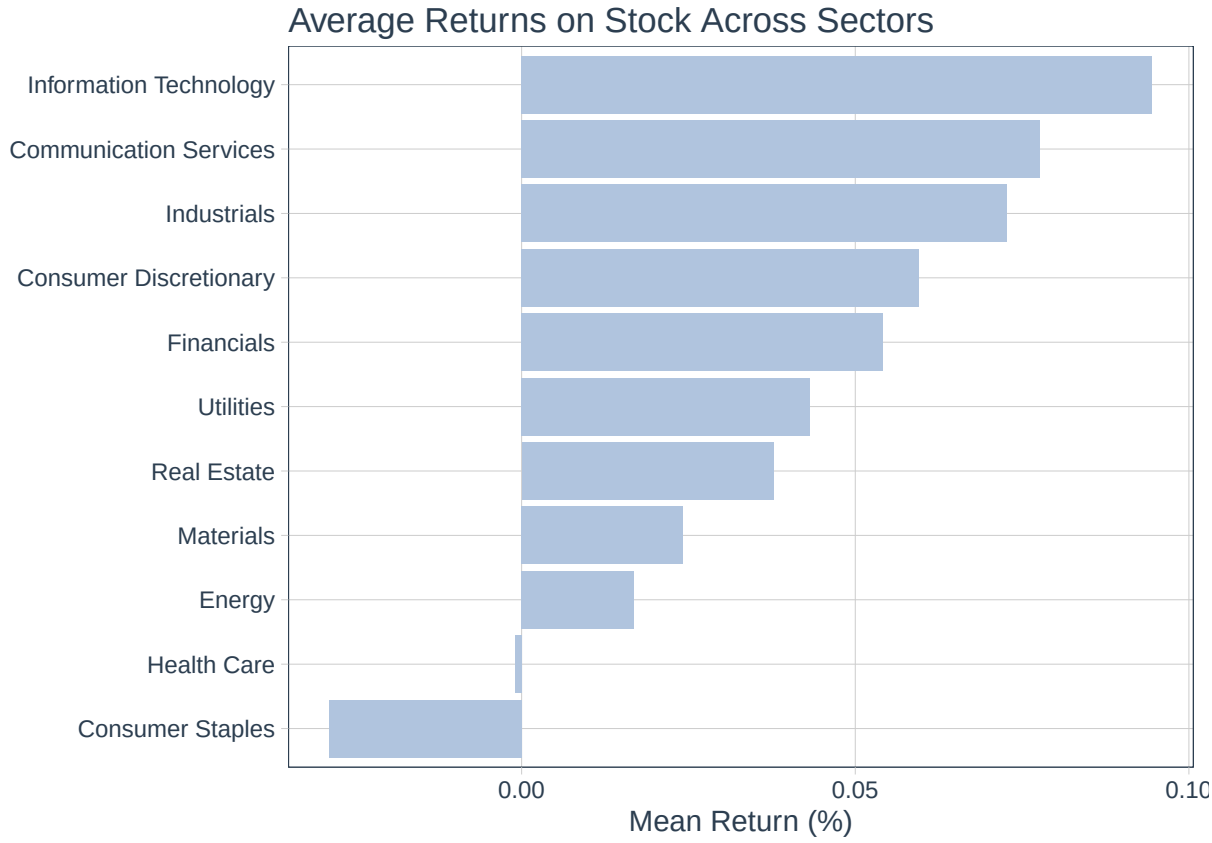
The chart presents the average daily return of S&P 500 companies across different sectors from 2023 to 2024.

We can see that

- Top Performing Sectors include Information Technology, Communication Services.
- Sectors such as Financials and Real Estate fall in the middle range.
- Sectors like Consumer Staples and Health Care have negative average daily returns.

```
# Mean returns by sector
mean_returns <- returns_all %>%
    group_by(gics_sector) %>%
    summarise(mean_return = mean(return, na.rm = TRUE))
# Plot mean returns by sector
ggplot(mean_returns, aes(x = reorder(gics_sector, mean_return),
```

```
    y = mean_return)) + geom_bar(stat = "identity", fill = "lightsteelblue") +
    coord_flip() + labs(title = "Average Returns on Stock Across Sectors",
    x = "", y = "Mean Return (%)") + theme_tq()
```

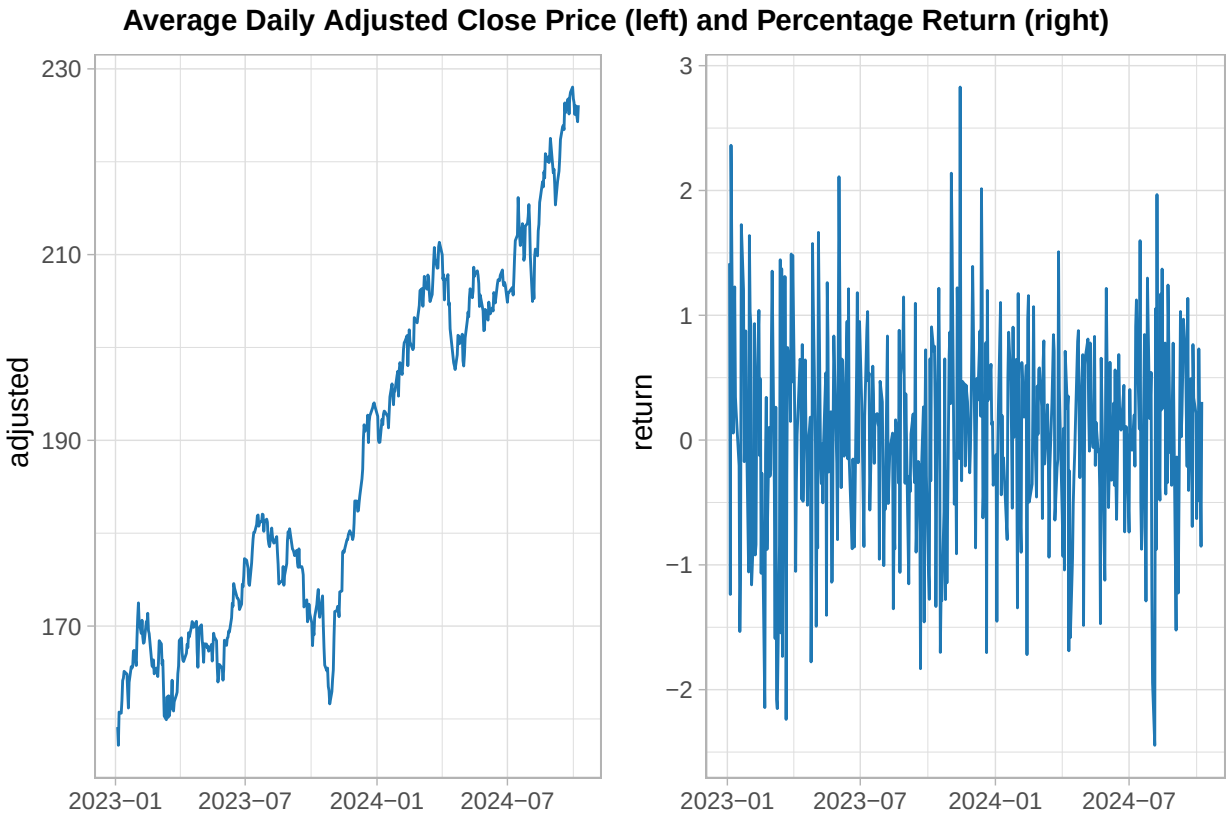



### Percentage Returns or Absolute Price Changes

Both Price metrics and percentage returns are associated with the market but behave differently. The Left Panel shows the mean adjusted close price across all companies from 2023 to 2024, and the right panel shows the mean return across all companies in the same period.

```
plot_list <- list()
# Loop to calculate the mean of 'adjusted' and 'return'
# across all companies over time
for (metric in c("adjusted", "return")) {
    # Create the mean calculation for the chosen metric
    p <- returns_all %>%
        # Pivot the data so each symbol's 'metric' becomes
        # a column
    pivot_wider(id_cols = date, names_from = symbol, values_from = all_of(metric)) %>%
        # Calculate the row-wise mean for all companies,
        # excluding the date column
    mutate(mean_all = rowMeans(select(., -date), na.rm = TRUE)) %>%
        as.data.frame() %>%
        ggplot(aes(x = date, y = mean_all)) + geom_line(color = "#1F78B4") +
        labs(x = "", y = paste("", metric)) + theme_light()
    plot_list[[metric]] <- p
}
```

```r
grid.arrange(plot_list[["adjusted"]], plot_list[["return"]],
    ncol = 2, top = textGrob("Average Daily Adjusted Close Price (left) and Percentage Return (right)",
        gp = gpar(fontsize = 11, fontface = "bold")))
```




**A few comparisons**

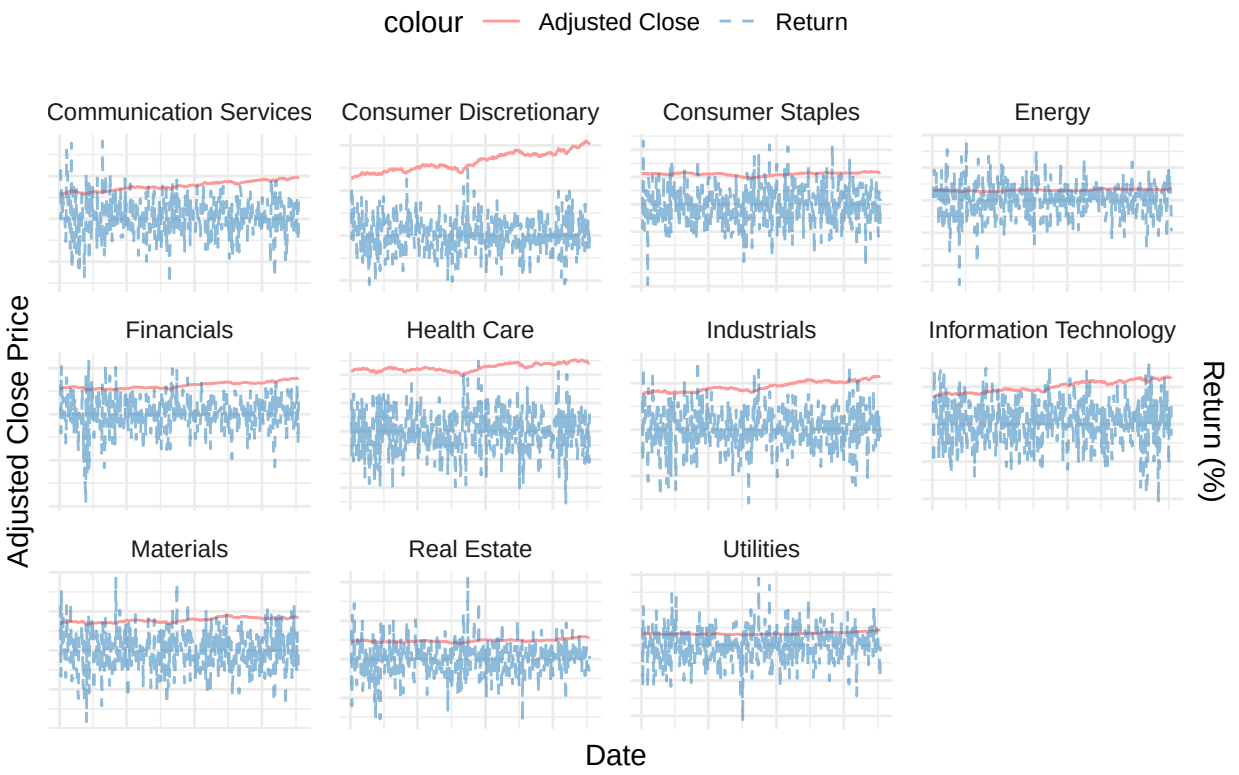
1. The plot of adjusted stock prices (left) displays a steady yet overall pronounced upward trend. By contrast, the returns (right) drift upwards or downwards very drastically from day to day.
2. Adjusted stock prices can vary substantially over time. As seen in the left panel of the chart, the adjusted price rose from approximately 165 to over 220 between 2023 and 2024. Even though returns exhibit high volatility on a daily basis, the range tends to be constrained from the beginning to the end of the period—between -3% and 3% in the plot. After all, we use the percentage return.

**Remark**

Although it may initially seem counterintuitive that the price should go upward so much, this is, in fact, the case in a growing economy or during periods of market expansion or high inflation. Furthermore, since the S&P 500 is composed of large, established companies, many of which have demonstrated long-term growth, the mean adjusted price naturally rises over time.

```r
# Dual-axis plot with adjusted close price on the primary
# y-axis and returns on the secondary y-axis
returns_all %>%
    # filter(gics_sector %in% c('Information Technology',
    # 'Financials', 'Health Care')) %>%
group_by(date, gics_sector) %>%
    summarise(avg_adjusted = mean(adjusted, na.rm = TRUE), avg_return = mean(return,
        na.rm = TRUE)) %>%
```

```
ggplot(aes(x = date)) + geom_line(aes(y = avg_adjusted, color = "Adjusted Close"),
size = 0.5) + geom_line(aes(y = avg_return * 100, color = "Return"),
size = 0.5, alpha = 0.5, linetype = "dashed") + facet_wrap(~gics_sector,
scales = "free_y") + scale_y_continuous(name = "Adjusted Close Price",
sec.axis = sec_axis(~./100, name = "Return (%)")) + labs(title = "",
x = "Date") + scale_color_manual(values = c(`Adjusted Close` = "#FB9A99",
Return = "#1F78B4")) + theme_minimal() + theme(legend.position = "top",
axis.text.x = element_blank(), axis.text.y = element_blank())
```




**Interesting Findings & Takeaways**

Looking at Sector Breakdown, We can find that Information Technology consistently has the highest average daily return. Indeed, Tech stocks have outperformed other sectors.

At the same time, IT and Financials display the most fluctuation in returns - risk is more concentrated in growth sectors. This suggests a *trade-off between risk and return.*

The largest representation in the index doesn't necessarily equate to profitability - as seen with Health Care. This sector is in the top 4 sectors by company count yet seems to yield negative returns (using the data in the analysis; limited time frame).

It could be that it leans towards more public service-oriented industries, which tend to be less volatile and offer steady but lower returns.

Interestingly, I expected Financials to be one of the highest returns. However, it may be underperformed due to economic inflation, recent trade issues, and other macroeconomic factors affecting the financial industry.

- While the overall market grows, day-to-day price changes can be erratic. In other words, despite short-term fluctuations, the market's overall trajectory is positive.
- *In relation to the following analysis*, the percentage returns naturally normalize the data and seem more suitable for more meaningful comparisons across companies, regardless of their price levels. Thus, I did the clustering based on returns, hoping to identify companies with similar performance patterns.

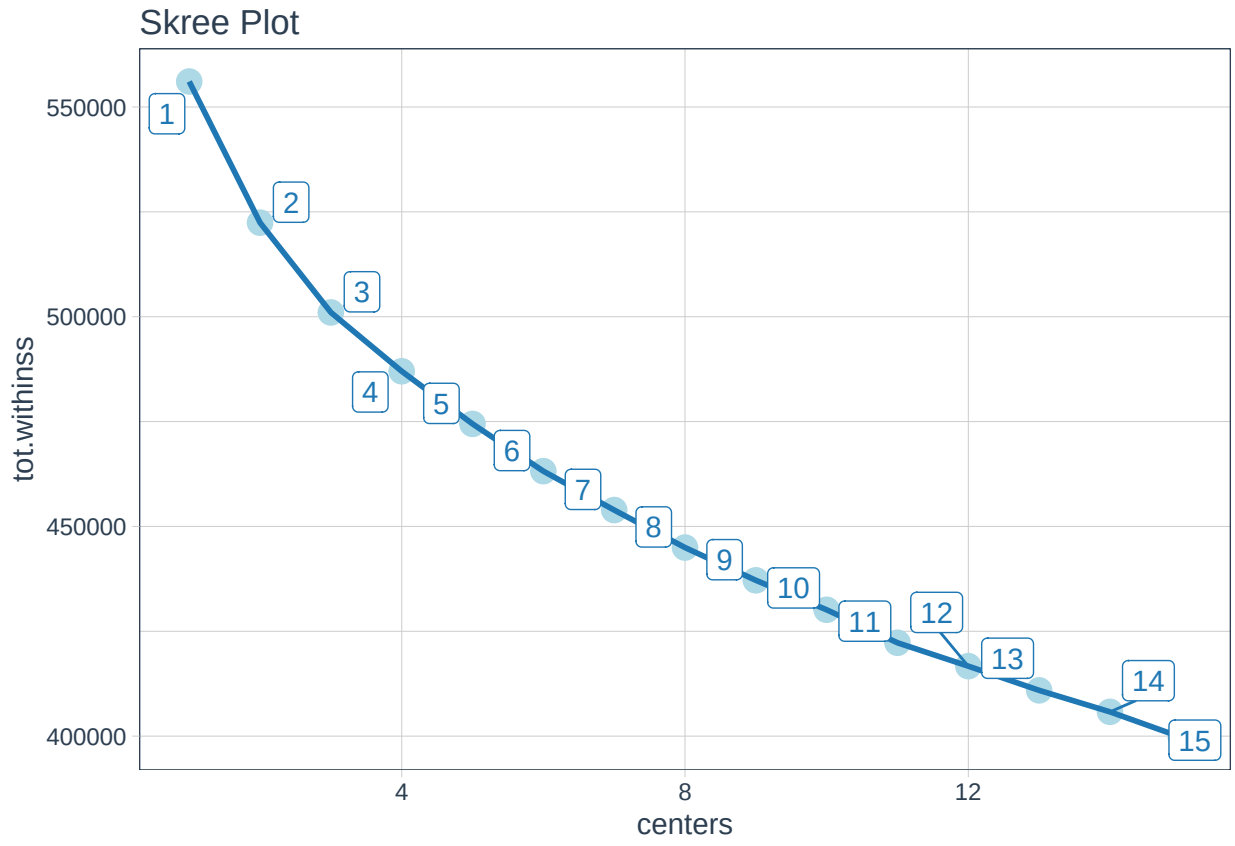
# Multivariate Analysis

### S&P 500 Index Segmentation via Clustering

```
library(ggrepel)
returns_matrix <- returns_all %>%
    # symbol in 1st column, follow by the stock return on
    # each day in all other columns of every stock
pivot_wider(id_cols = symbol, names_from = date, values_from = return,
    values_fill = 0)
kmeans_mapper <- function(centers = 3) {
    returns_matrix %>%
        select(-symbol) %>%
        na.omit() %>%
        kmeans(centers = centers, nstart = 40)
}
cluster_map <- tibble(centers = 1:15) %>%
    mutate(k_means = centers %>%
        purrr::map(kmeans_mapper)) %>%
    mutate(glance = k_means %>%
        purrr::map(glance))
cluster_map %>%
    unnest(glance) %>%
    select(centers, tot.withinss)
```

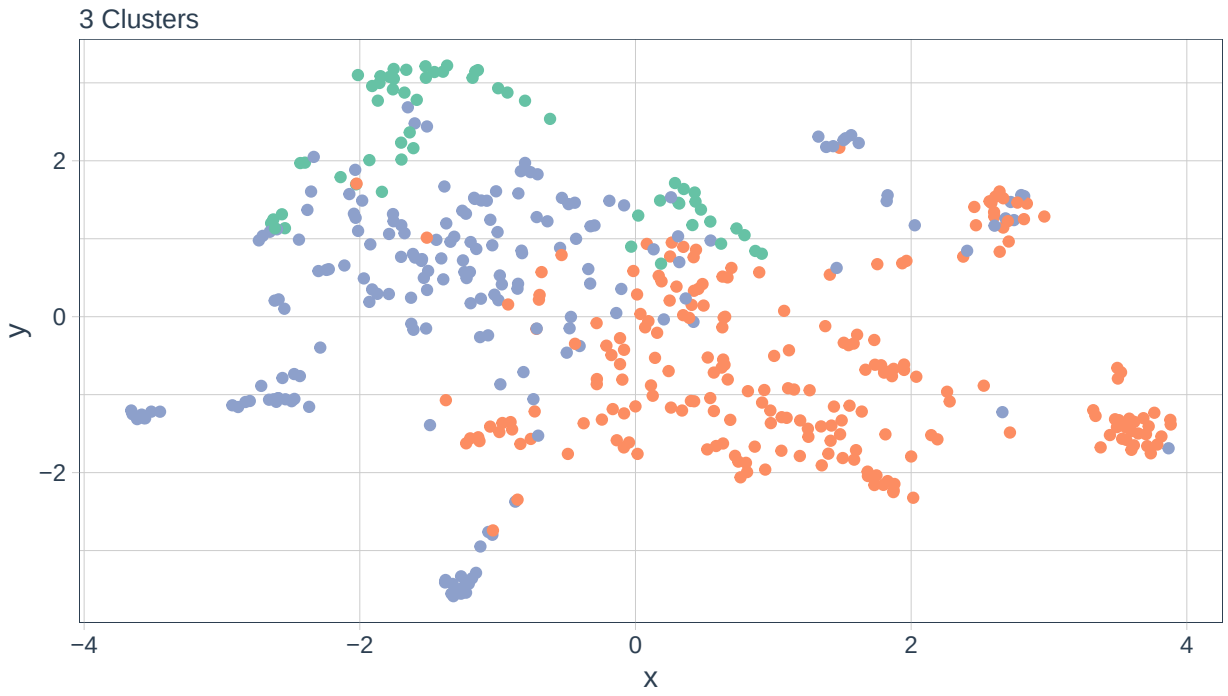
| centers | tot.withinss |
|---|---|
| 1 | 556072.5 |
| 2 | 522420.1 |
| 3 | 501018.7 |
| 4 | 486995.8 |
| 5 | 474443.9 |
| 6 | 463175.7 |
| 7 | 453905.1 |
| 8 | 444962.4 |
| 9 | 437152.6 |
| 10 | 430127.4 |
| 11 | 422262.5 |
| 12 | 416683.3 |
| 13 | 410917.7 |
| 14 | 405794.6 |
| 15 | 399768.6 |

```
cluster_map %>%
    unnest(glance) %>%
    select(centers, tot.withinss) %>%
    ggplot(aes(centers, tot.withinss)) + geom_point(color = "lightblue",
    size = 4) + geom_line(color = "#1F78B4", size = 1) + ggrepel::geom_label_repel(aes(label = centers)
    color = "#1F78B4") + theme_tq() + labs(title = "Skree Plot")
```




```
umap_obj <- returns_matrix %>%
    select(-symbol) %>%
    umap()
mapped <- umap_obj$layout %>%
    as_tibble() %>%
    set_names(c("x", "y")) %>%
    cbind(symbol = returns_matrix$symbol)
three_clusters <- cluster_map %>%
    pull(k_means) %>%
    pluck(3) %>%
    augment(returns_matrix) %>%
    select(symbol, .cluster)
mapped %>%
    left_join(three_clusters, by = "symbol") %>%
    mutate(label_text = str_glue("Company: {symbol}
                                  Cluster: {.cluster}")) %>%
    ggplot(aes(x, y, color = .cluster)) + geom_point() + theme_tq() +
    scale_color_brewer(palette = "Set2") + labs(title = "", subtitle = "3 Clusters")
```

**3 Clusters**



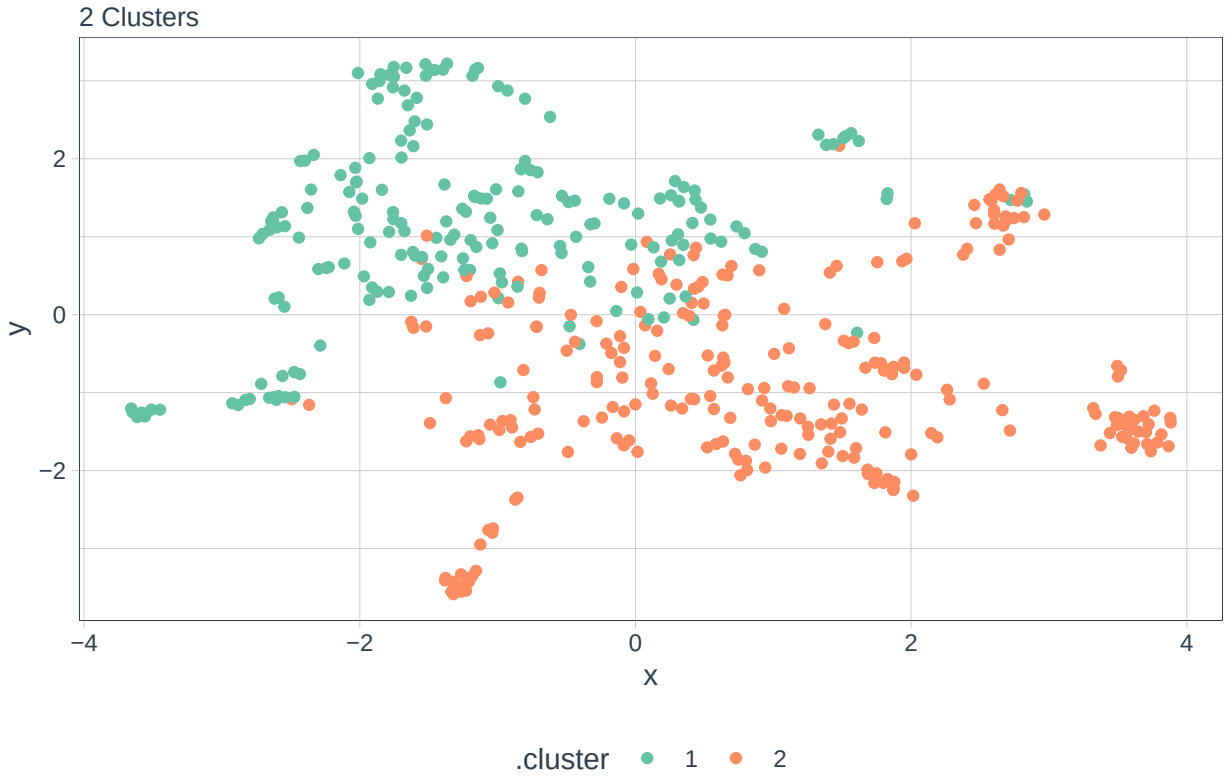

```
three_clusers_res <- three_clusters %>%
    left_join(returns_all, by = "symbol") %>%
    select(symbol, security, gics_sector, .cluster) %>%
    unique()
three_clusers_res %>%
    group_by(.cluster, gics_sector) %>%
    summarise(count = n()) %>%
    mutate(percentage = paste(round(100 * count/sum(count), 1),
        "%"))
```

| .cluster | gics_sector | count | percentage |
|---|---|---|---|
| 1 | Communication Services | 4 | 7 % |
| 1 | Consumer Discretionary | 7 | 12.3 % |
| 1 | Health Care | 1 | 1.8 % |
| 1 | Industrials | 3 | 5.3 % |
| 1 | Information Technology | 40 | 70.2 % |
| 1 | Utilities | 2 | 3.5 % |
| 2 | Communication Services | 7 | 3.2 % |
| 2 | Consumer Discretionary | 15 | 6.9 % |
| 2 | Consumer Staples | 28 | 12.8 % |
| 2 | Energy | 2 | 0.9 % |
| 2 | Financials | 31 | 14.2 % |
| 2 | Health Care | 45 | 20.6 % |
| 2 | Industrials | 21 | 9.6 % |
| 2 | Information Technology | 14 | 6.4 % |
| 2 | Materials | 7 | 3.2 % |
| 2 | Real Estate | 22 | 10.1 % |
| 2 | Utilities | 26 | 11.9 % |
| 3 | Communication Services | 4 | 2.2 % |

| .cluster | gics_sector | count | percentage |
|---|---|---|---|
| 3 | Consumer Discretionary | 26 | 14 % |
| 3 | Consumer Staples | 4 | 2.2 % |
| 3 | Energy | 17 | 9.1 % |
| 3 | Financials | 38 | 20.4 % |
| 3 | Health Care | 12 | 6.5 % |
| 3 | Industrials | 47 | 25.3 % |
| 3 | Information Technology | 11 | 5.9 % |
| 3 | Materials | 16 | 8.6 % |
| 3 | Real Estate | 9 | 4.8 % |
| 3 | Utilities | 2 | 1.1 % |

```
two_clusters <- cluster_map %>%
    pull(k_means) %>%
    pluck(2) %>%
    augment(returns_matrix) %>%
    select(symbol, .cluster)
mapped %>%
    left_join(two_clusters, by = "symbol") %>%
    mutate(label_text = str_glue("Company: {symbol}
                                 Cluster: {.cluster}")) %>%
    ggplot(aes(x, y, color = .cluster)) + geom_point() + theme_tq() +
    scale_color_brewer(palette = "Set2") + labs(title = "", subtitle = "2 Clusters")
```

**2 Clusters**

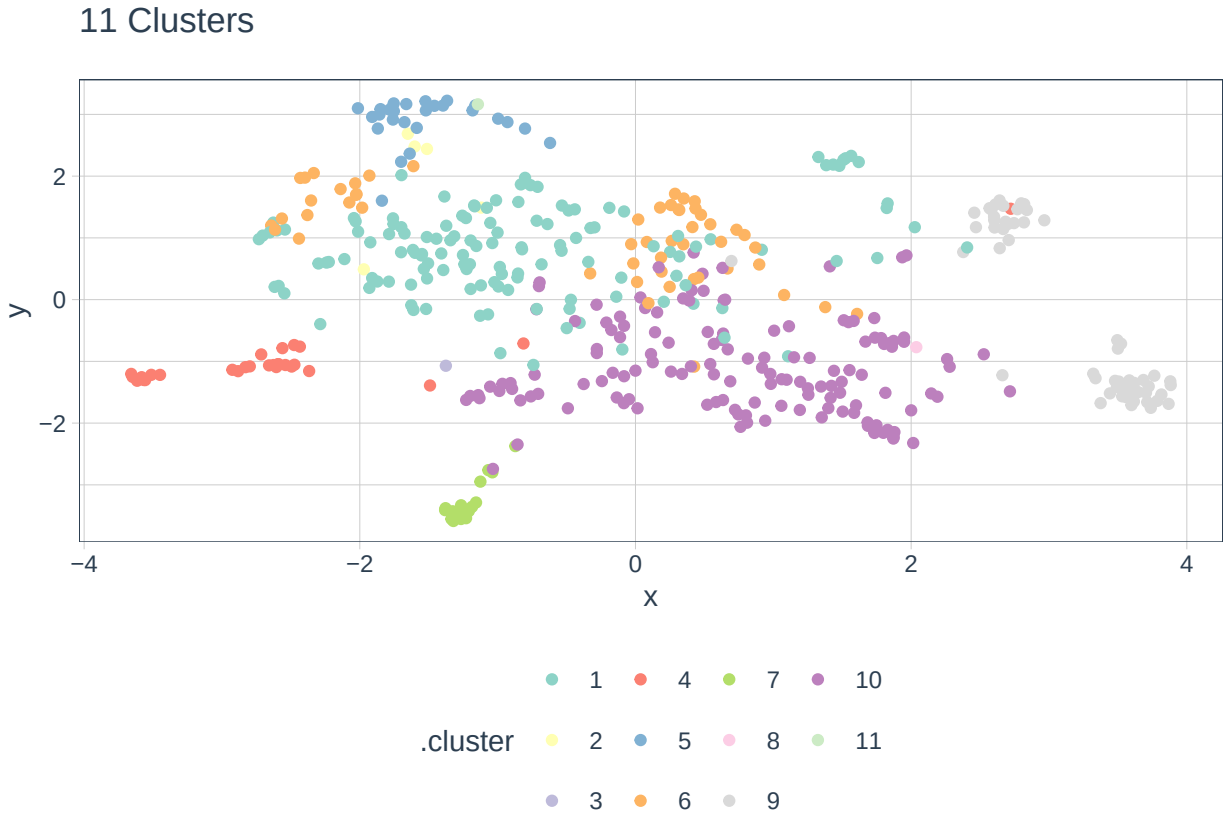


```
two_clusters_res <- two_clusters %>%
    left_join(returns_all, by = "symbol") %>%
    select(symbol, security, gics_sector, .cluster) %>%
    unique()
two_clusters_res %>%
    group_by(.cluster, gics_sector) %>%
    summarise(count = n()) %>%
    mutate(percentage = paste(round(100 * count/sum(count), 1),
        "%"))
```

| .cluster | gics_sector | count | percentage |
|---|---|---|---|
| 1 | Communication Services | 9 | 4.4 % |
| 1 | Consumer Discretionary | 30 | 14.7 % |
| 1 | Consumer Staples | 1 | 0.5 % |
| 1 | Financials | 34 | 16.7 % |
| 1 | Health Care | 12 | 5.9 % |
| 1 | Industrials | 46 | 22.5 % |
| 1 | Information Technology | 54 | 26.5 % |
| 1 | Materials | 10 | 4.9 % |
| 1 | Real Estate | 5 | 2.5 % |
| 1 | Utilities | 3 | 1.5 % |
| 2 | Communication Services | 6 | 2.3 % |
| 2 | Consumer Discretionary | 18 | 7 % |
| 2 | Consumer Staples | 31 | 12.1 % |
| 2 | Energy | 19 | 7.4 % |
| 2 | Financials | 35 | 13.6 % |
| 2 | Health Care | 46 | 17.9 % |
| 2 | Industrials | 25 | 9.7 % |
| 2 | Information Technology | 11 | 4.3 % |
| 2 | Materials | 13 | 5.1 % |
| 2 | Real Estate | 26 | 10.1 % |
| 2 | Utilities | 27 | 10.5 % |

```
eleven_clusters <- cluster_map %>%
    pull(k_means) %>%
    pluck(11) %>%
    augment(returns_matrix) %>%
    select(symbol, .cluster)
mapped %>%
    left_join(eleven_clusters, by = "symbol") %>%
    mutate(label_text = str_glue("Company: {symbol}
                                 Cluster: {.cluster}")) %>%
    ggplot(aes(x, y, color = .cluster)) + geom_point() + theme_tq() +
    scale_color_brewer(palette = "Set3") + labs(title = "11 Clusters",
    subtitle = "")
```

**11 Clusters**




```
eleven_clusters_res <- eleven_clusters %>%
    left_join(returns_all, by = "symbol") %>%
    select(symbol, security, gics_sector, .cluster) %>%
    unique()
eleven_clusters_res %>%
    group_by(.cluster, gics_sector) %>%
    summarise(count = n()) %>%
    mutate(percentage = paste(round(100 * count/sum(count), 1),
        "%"))
```

| .cluster | gics_sector | count | percentage |
|---|---|---|---|
| 1 | Communication Services | 5 | 3.7 % |
| 1 | Consumer Discretionary | 32 | 23.9 % |
| 1 | Consumer Staples | 3 | 2.2 % |
| 1 | Financials | 10 | 7.5 % |
| 1 | Health Care | 13 | 9.7 % |
| 1 | Industrials | 41 | 30.6 % |
| 1 | Information Technology | 10 | 7.5 % |
| 1 | Materials | 16 | 11.9 % |
| 1 | Real Estate | 4 | 3 % |
| 2 | Health Care | 1 | 20 % |
| 2 | Industrials | 1 | 20 % |
| 2 | Information Technology | 2 | 40 % |
| 2 | Materials | 1 | 20 % |
| 3 | Financials | 1 | 100 % |
| 4 | Financials | 28 | 96.6 % |
| 4 | Real Estate | 1 | 3.4 % |

| .cluster | gics_sector | count | percentage |
|---|---|---|---|
| 5 | Consumer Discretionary | 1 | 3.8 % |
| 5 | Information Technology | 25 | 96.2 % |
| 6 | Communication Services | 5 | 9.6 % |
| 6 | Consumer Discretionary | 5 | 9.6 % |
| 6 | Consumer Staples | 1 | 1.9 % |
| 6 | Financials | 1 | 1.9 % |
| 6 | Health Care | 4 | 7.7 % |
| 6 | Industrials | 11 | 21.2 % |
| 6 | Information Technology | 22 | 42.3 % |
| 6 | Utilities | 3 | 5.8 % |
| 7 | Energy | 17 | 94.4 % |
| 7 | Materials | 1 | 5.6 % |
| 8 | Health Care | 1 | 100 % |
| 9 | Financials | 1 | 1.8 % |
| 9 | Health Care | 1 | 1.8 % |
| 9 | Materials | 1 | 1.8 % |
| 9 | Real Estate | 26 | 46.4 % |
| 9 | Utilities | 27 | 48.2 % |
| 10 | Communication Services | 5 | 3.6 % |
| 10 | Consumer Discretionary | 10 | 7.2 % |
| 10 | Consumer Staples | 28 | 20.3 % |
| 10 | Energy | 2 | 1.4 % |
| 10 | Financials | 28 | 20.3 % |
| 10 | Health Care | 38 | 27.5 % |
| 10 | Industrials | 18 | 13 % |
| 10 | Information Technology | 5 | 3.6 % |
| 10 | Materials | 4 | 2.9 % |
| 11 | Information Technology | 1 | 100 % |

# Possible Extensions

**Location-Based Analysis** is an area of interest and potential exciting findings. It would be interesting to explore the geographical distribution of company headquarters in more detail. Some key questions include: Do headquarters tend to cluster in specific cities or regions? Additionally, is there a temporal effect where older headquarters locations attract more companies over time, or do we observe a dispersal effect where companies move away from traditional hubs? This could resemble historical patterns like the decline of the Rust Belt and the rise of the Sun Belt in the U.S., where economic and industrial shifts influenced corporate decisions on where to establish headquarters. These trends could provide valuable insights into regional economic growth and the changing corporate landscape.

Another compelling question is whether specific sectors tend to cluster in certain regions. For example, Silicon Valley is well-known for its concentration of technology companies. Exploring whether the location of a company's headquarters correlates with stock performance could reveal important regional trends. Tech-oriented regions, like Silicon Valley, might show higher returns due to their industry concentration, making location a potential factor in predicting company performance.

Furthermore, incorporating location and historical company changes into the clustering analysis could enhance our understanding of how geographic factors influence company behavior and performance. The location could be a strong predictor in clustering companies with similar stock performance, providing deeper insights into how regional and industry-specific trends shape financial outcomes.

It would also be worthwhile to explore **how ownership structure (public vs. private) affects stock performance**, particularly in industries with significant government influence, such as Energy and Utilities.

Public and private companies in these sectors may exhibit different stock behaviors due to regulatory factors, government policies, or market conditions. Investigating these differences could provide further insights into how ownership structure impacts risk, returns, and overall company performance.

Additionally, exploring the **risk-return trade-offs** across sectors could help identify companies or industries that balance growth potential with stability. Incorporating factors like ownership structure (public vs. private) into the analysis, especially in government-influenced sectors like Energy and Utilities, could uncover deeper connections between company control and stock performance.

## Conclusion

The performance of S&P 500 sectors in 2023 and 2024 shows significant differences in growth patterns. Growth sectors, such as Information Technology, continue to outperform, while more traditional sectors, like Utilities, maintain their stability. Volatility in sectors like Energy reflects their sensitivity to external economic factors. These sectoral trends are valuable for understanding market dynamics and guiding investment strategies based on sector-specific risk and growth potential.