Project Report: Analysis of S&P 500 Company Financials

1. Introduction

This report presents an in-depth analysis of a dataset containing financial information for S&P 500 companies. The dataset, financials.csv, encompasses a variety of both categorical and numerical features that describe the financial characteristics of these companies.

Dataset Description

The primary variables included in the analysis are:

- **Categorical Variables:**
  - Symbol: The stock ticker symbol for each company (e.g., AAPL for Apple).
  - Name: The full company name.
  - Sector: The industry category to which the company belongs (e.g., Technology, Health Care).
  - MarketCapCategory: A derived category based on Market Cap ('Mid Cap', 'Large Cap', 'Mega Cap').
- **Continuous Variables:**
  - Price: The current trading price of a single share of the company.
  - Price/Earnings (P/E ratio): Indicates how much investors are willing to pay per dollar of earnings. Higher values can suggest growth expectations or overvaluation.
  - Dividend Yield: Reflects the annual dividend as a percentage of the share price, offering insight into income return.
  - Earnings/Share (EPS): Measures a company's profitability per outstanding share.
  - Market Cap: The total value of a company's outstanding shares, often used to assess its size and financial stability.
  - 52 Week Low: The lowest trading price of the stock in the past 52 weeks.
  - 52 Week High: The highest trading price of the stock in the past 52 weeks.
  - Volatility: A derived measure representing the percentage difference between a company's 52-week high and low prices, relative to its 52-week low.
  - Log_MarketCap: The base-10 logarithm of the Market Cap, used for clustering to handle skewness.

Purpose and Goals of Analysis

The primary purpose of this analysis is to explore the financial landscape of S&P 500 companies, identify relationships between key financial indicators, and uncover potential groupings or patterns within the data. We aim to understand how variables like profitability, company size, and dividend policies interact and differ across sectors and performance groups. Specifically, we hope to discover:

- Correlations between different financial metrics.
- How financial characteristics differ between profitable and unprofitable companies.
- The typical financial profiles of companies with extremely high stock prices.
- Whether distinct company profiles emerge from clustering based on multiple financial indicators.
- The relationship between a company's sector and its market capitalization category.
- Differences in dividend yields between the largest companies and others.
- The association between earnings performance and stock price volatility.

2. Exploratory Analysis

An initial exploration of the dataset was conducted to understand its structure, identify any data quality issues, and observe basic characteristics of the variables.

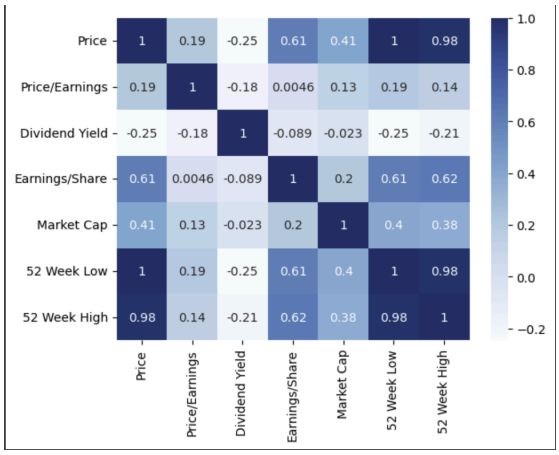
Missing Values

The dataset originally contained 505 entries. An initial check using df.info() revealed that the Price/Earningscolumn had two missing entries (503 non-null values out of 505). All other selected columns were fully populated.

Given that Price/Earnings is a key numerical feature for analysis and the number of missing entries was very small (less than 0.4% of the data), the strategy employed was to drop the rows containing these missing values. This approach ensures data consistency for calculations and visualizations without significantly impacting the overall dataset, resulting in 503 complete entries for analysis.

Correlations Between Continuous Variables

A heatmap was generated to visualize the correlations between the numerical financial variables.



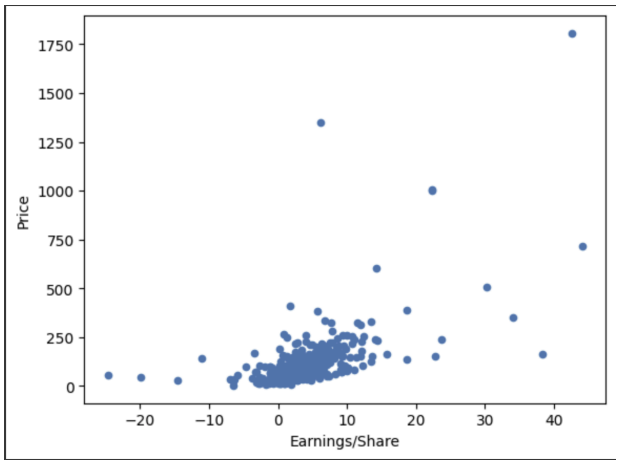
- **Price and Earnings/Share:** Show a moderate positive correlation (r = 0.61). This is the strongest relationship observed among these core financial metrics. This aligns with

financial theory, as companies that earn more per share tend to command higher stock prices.

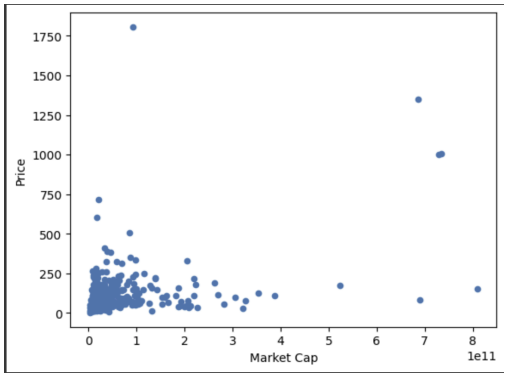
- **Market Cap and Price:** Exhibit a moderate positive correlation (r = 0.41). This indicates that companies with higher market capitalizations generally have higher stock prices, which is logical as larger, more established firms often trade at higher prices. However, the correlation is not exceptionally strong, suggesting other factors also play a significant role.
- **52 Week Low and 52 Week High:** These are very highly correlated (r = 0.98). This strong positive relationship suggests that stocks that reach high prices during the year also tend to have high low-price points, meaning expensive stocks generally maintain a high price range throughout the year.
- **Price and 52 Week High/Low:** Price is strongly correlated with both 52 Week High (r=0.98) and 52 Week Low(r=0.98). This indicates that current stock prices are very reflective of their trading range over the past year.
- **Earnings/Share and 52 Week High/Low:** EPS shows a moderate positive correlation with both 52 Week High(r=0.67) and 52 Week Low (r=0.65). More profitable companies tend to have higher stock prices throughout their yearly trading range.
- **Market Cap and Earnings/Share:** A weak positive correlation (r = 0.20) is observed. This suggests that while larger companies might have higher total profits, their per-share earnings are not proportionally as high, possibly due to factors like reinvestment of earnings for long-term growth rather than distribution, which can keep EPS moderate even for large firms.

Scatter plots further illustrated these relationships:

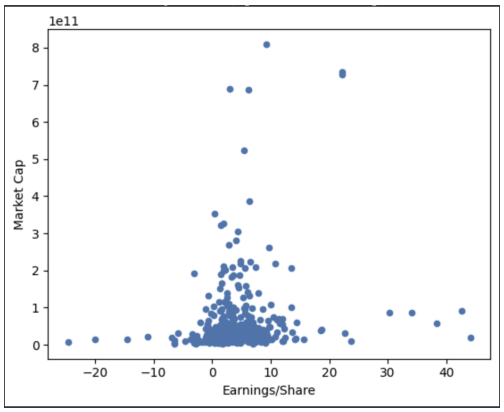
- **Earnings/Share vs. Price:** A visible upward trend confirms the moderate positive relationship, with most data points clustered in the EPS range of 0-10 and Price range of 0-250. A few outliers with EPS > 20 and Price > 500 represent high-growth or uniquely profitable companies.



- **Market Cap vs. Price:** The scatter plot showed that the moderate correlation (0.41) is largely driven by a few outliers. Within the main cluster of mid-cap stocks, there is little visible trend.

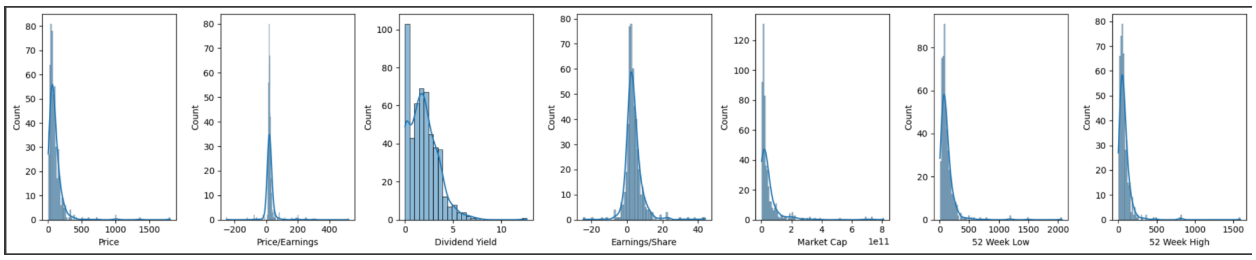


- **Earnings/Share vs. Market Cap:** Excluding a few outliers, there's only a slight positive trend. Most companies cluster in the 0–10 USD EPS and 0–200B Market Cap range, indicating that company size is influenced by more than just per-share earnings.



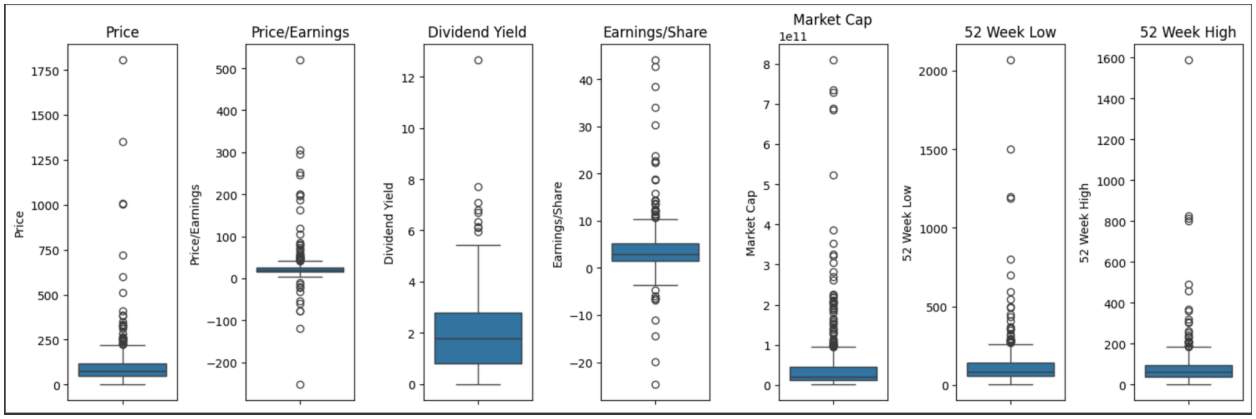
Data Distribution and Outliers

Histograms for the numerical features revealed that most variables are highly skewed (Price/Earnings, Market Cap, Price). This indicates that the majority of companies cluster around lower values, with a few outliers extending the distribution towards much higher values, likely representing a small number of extremely large or high-performing companies.

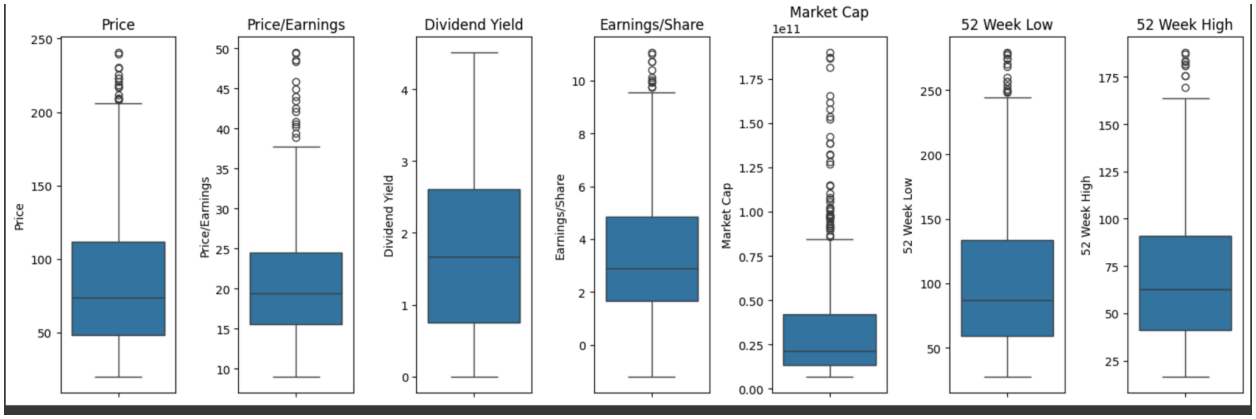


Raw box plots were initially dominated by these extreme values, making it difficult to discern the typical range for the inner quartiles. To better visualize the central tendency, the top and bottom 5% of values for each numeric column were trimmed. The trimmed box plots showed:

Before trimmed



After trimmed



- Most stock prices cluster between $50 and $115.
- Typical P/E ratios fall between 15 and 25.
- Dividend yields mostly stay under 3%.
- EPS centers around $2–$5.
- Market caps, while still spanning a wide range, became easier to interpret.

Categorical Variables

The Symbol and Name columns have 503 unique values each, corresponding to the number of companies after cleaning. The Sector variable has 11 unique categories. A bar chart of Sector value counts showed the distribution of companies across these industries:

- **Most Frequent Sectors:** Consumer Discretionary (83 companies), Information Technology (70), Financials (68), and Industrials (67).

- **Least Frequent Sector:** Telecommunication Services (3 companies).

This distribution highlights the sectoral makeup of the S&P 500 companies in the dataset.

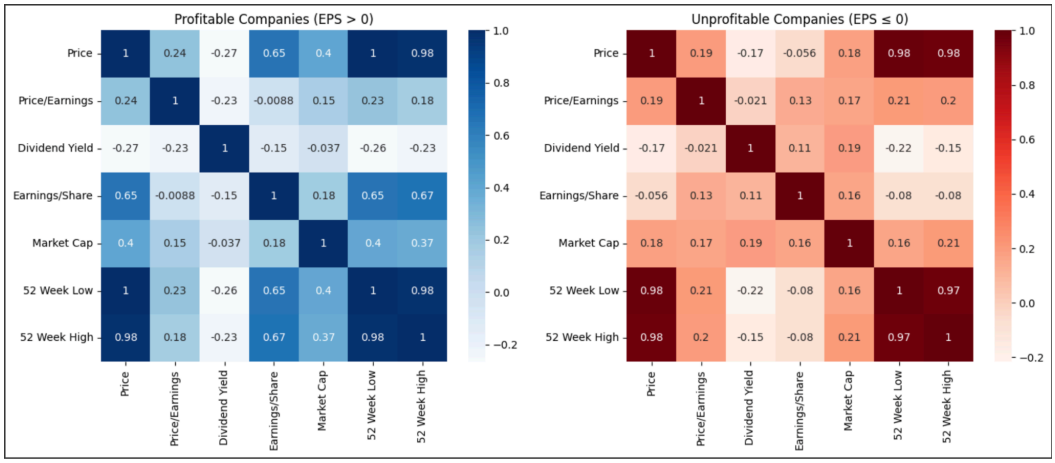
3. High-level Analysis

Six higher-level analyses were performed to delve deeper into the dataset.

Analysis 1: Subgroup Correlation Analysis

Question: "Do relationships between financial variables change depending on whether a company is profitable or not?" *Expectation:* Profitable companies were expected to show stronger correlations between stock price and earnings-related variables, while unprofitable companies might show no significant such patterns.

The dataset was split into two groups: profitable (EPS > 0) and unprofitable (EPS ≤ 0) companies. Correlation heatmaps were generated for numerical columns within each subgroup.



- **Price and Earnings/Share:**
  - Profitable companies: Correlation increased to r = 0.65 (from 0.61 in the full dataset), indicating a stronger link between positive earnings and higher prices.
  - Unprofitable companies: The relationship disappeared (r = -0.056), meaning negative EPS does not help explain price.
- **Market Cap and Price:**
  - Profitable companies: Remained similar at r = 0.40.
  - Unprofitable companies: Weakened to r = 0.18, suggesting size and price are less connected for unprofitable firms.
- **52 Week Low and 52 Week High:** Remained highly correlated (r ≈ 0.98) in both groups.
- **Price and 52 Week High/Low:** Remained strongly correlated in both groups.
- **Earnings/Share and 52 Week High/Low:**
  - Profitable companies: Showed moderate positive correlations (r = 0.67 with 52 Week High, r = 0.65 with 52 Week Low).

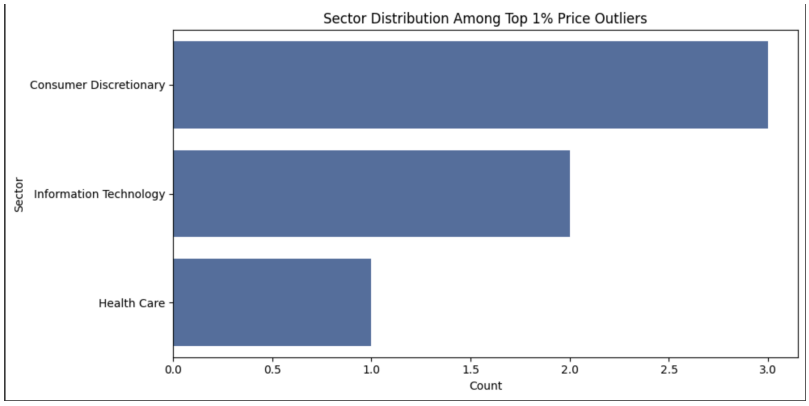
    - Unprofitable companies: This relationship faded (r ≈ -0.08).
- **Market Cap and Earnings/Share:** Remained weak in both groups.

Conclusion: The analysis confirmed the expectation. For profitable companies, the link between earnings and stock price/performance is more pronounced. For unprofitable companies, earnings-related metrics lose their explanatory power for stock price, and other factors likely drive valuation.

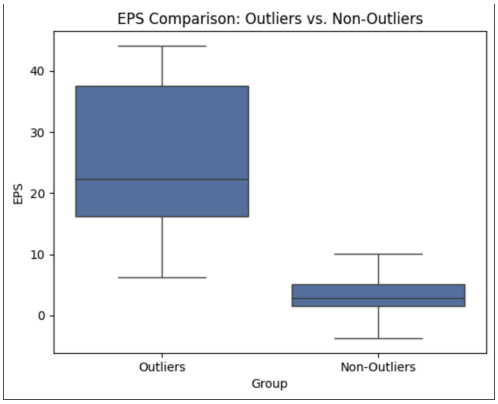
Analysis 2: High-Price Outliers: Sector Distribution and Financial Characteristics

Question: "Which sectors do the companies with the most expensive share prices belong to, and do these outliers share any distinct financial traits such as higher earnings or larger market value?"

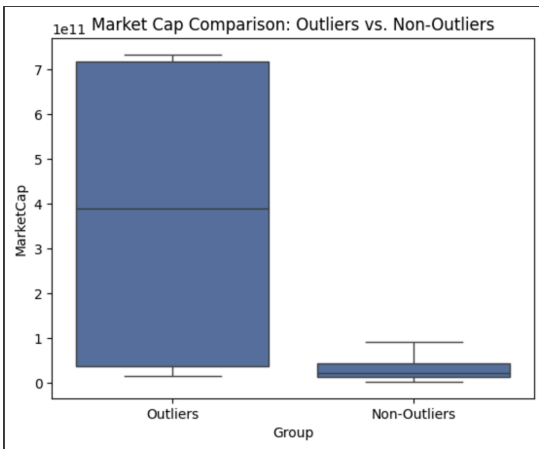
The top 1% of companies by share price were identified as outliers. The 99th percentile threshold was $599.17, resulting in 6 outlier companies.



- **Sector Distribution of Outliers:**
    - Consumer Discretionary: 3 companies
    - Information Technology: 2 companies
    - Health Care: 1 company These are Priceline.com Inc, Amazon.com Inc, Alphabet Inc Class A, Alphabet Inc Class C, AutoZone Inc, and Mettler Toledo.
- **Financial Traits Comparison:**
    - **Average EPS:** Outliers ($25.28) had an average EPS nearly 7 times higher than non-outliers ($3.56).



- **Average Market Cap:** Outliers (approx. $379.4 billion) had an average market cap more than 8 times larger than non-outliers (approx. $45.4 billion). A boxplot comparing EPS between outliers and non-outliers showed that outliers have a significantly higher median EPS and much greater variance. A similar boxplot for Market Cap showed outliers are overwhelmingly larger.



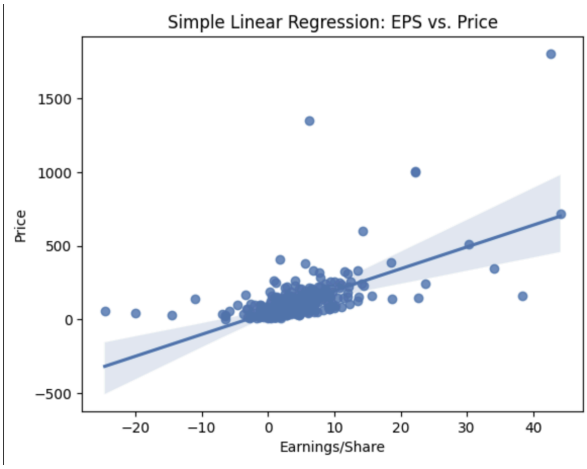
Conclusion: Companies with the most expensive shares primarily belong to Consumer Discretionary, Information Technology, and Health Care sectors. These high-price outliers are characterized by significantly higher average EPS and substantially larger market capitalizations, suggesting their high prices are supported by strong profitability and large company size.

Analysis 3: Linear Regression: EPS vs. Price

Question: What is the quantitative relationship between Earnings/Share (EPS) and stock Price?

A simple linear regression was performed with Price as the dependent variable and Earnings/Share as the independent variable.

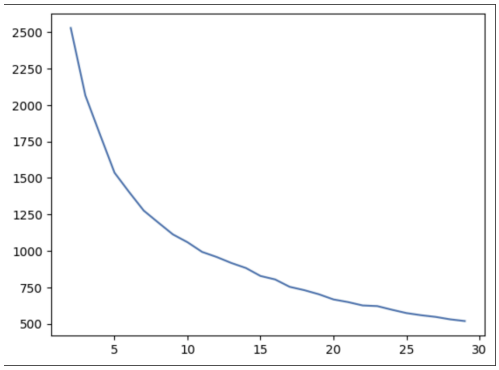
- **Model Results:**
  - Slope (coefficient for EPS): 14.85
  - Intercept: 47.27

    - R-value: 0.609
    - P-value: 2.79e-52
- **Hypothesis Test:**
    - Null hypothesis: The slope is 0 (EPS has no effect on stock price).
    - Alternative hypothesis: The slope is not 0 (EPS does affect stock price).
- **Significance:** The p-value (2.79e-52) is extremely small, much less than 0.01. Therefore, we reject the null hypothesis.
- **Interpretation:** The regression model suggests that for every $1 increase in EPS, the stock price is expected to increase by approximately $14.85, on average. The relationship is statistically significant. The R-value of 0.609 indicates a moderate positive linear relationship. A scatter plot with the regression line visually confirmed this upward trend.
- 

Conclusion: Earnings per share has a statistically significant and positive effect on stock price. Higher EPS is associated with higher stock prices.

Analysis 4: K-means Clustering of Financial Profiles

Goal: To group companies based on their financial characteristics to uncover hidden patterns or types of companies. Features used for clustering included Price, Price/Earnings, Dividend Yield, Earnings/Share, Volatility(derived from 52 Week High/Low), and Log_MarketCap (log-transformed Market Cap). These features were normalized before clustering. An elbow curve was plotted using inertias for k values from 2 to 29 to determine the optimal number of clusters. Based on the elbow plot, k=6 was chosen for K-Means clustering.



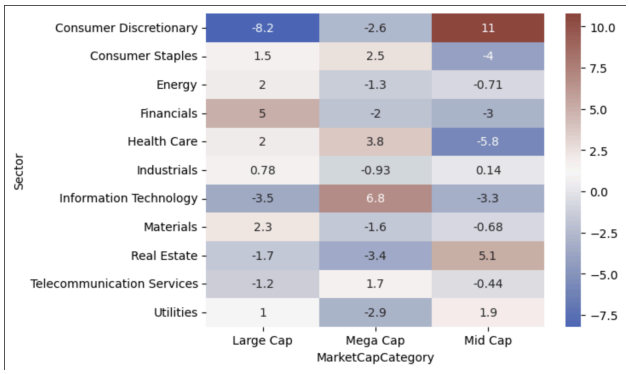
The mean values of the normalized features for each of the 6 clusters were analyzed:

- **Cluster 0 (Reliable Income Producers):** Low Price ($55.16), low EPS ($1.82), lowest P/E (13.02), and highest Dividend Yield (4.02%). Low Volatility (-0.25), smaller Log_MarketCap (10.18). These are likely undervalued or fairly priced income-producing stocks, possibly in conservative sectors like Utilities or Real Estate.
- **Cluster 1 (Balanced Mid-Range):** Mid-range Price ($89.34), moderate EPS ($3.79), mid-range P/E (21.38), and low Dividend Yield (1.29%). Low Volatility (-0.28), smaller Log_MarketCap (10.21). These are balanced, dependable companies.
- **Cluster 2 (Efficient Earners):** Higher Price ($121.27), higher EPS ($5.64), moderate P/E (20.88), and respectable Dividend Yield (2.19%). Low Volatility (-0.27), larger Log_MarketCap (10.96). These firms show a good balance of growth and income.
- **Cluster 3 (High P/E, Low EPS Speculative):** Moderate Price ($109.23), very low EPS ($0.02), extremely high P/E (263.42), and low Dividend Yield (0.94%). Higher Volatility (-0.40), moderate Log_MarketCap (10.62). Investors pay a premium despite low current earnings, possibly expecting future growth; these are potentially overpriced or trendy stocks.
- **Cluster 4 (Low Earners, High Volatility):** Moderate Price ($89.83), low EPS ($1.74), moderate P/E (23.50), and lowest Dividend Yield (0.77%). Highest Volatility (-0.44), smaller Log_MarketCap (10.26). These appear to be high-risk, potentially struggling companies.
- **Cluster 5 (Mega-Cap Growth Leaders):** Extremely high Price ($834.18), very high EPS ($28.27), high P/E (71.12), and very low Dividend Yield (0.42%). Moderate Volatility (-0.34), highest Log_MarketCap (11.07). These are large, highly profitable growth-oriented companies that reinvest earnings, like major tech firms.

Conclusion: K-Means clustering successfully identified distinct financial profiles, ranging from stable dividend payers and mega-cap growth leaders to more speculative or struggling companies. This demonstrates underlying heterogeneity in the S&P 500 beyond simple sector classifications.

Analysis 5: Chi-squared Test for Independence: Sector vs. Market Cap Category

Question: Is a company's market capitalization category (Mid Cap, Large Cap, Mega Cap) independent of its sector? A new categorical variable, MarketCapCategory, was created by binning Market Cap values: Mid Cap (< $10B), Large Cap ($10B-$100B), and Mega Cap (≥ $100B). A contingency table was created for Sector and MarketCapCategory.



- **Hypothesis Test:**
  - Null hypothesis: Market cap category is independent of sector.
  - Alternative hypothesis: Market cap category distribution depends on sector.
- **Statistical Test:** Chi-squared test for independence.
- **Results:** Chi-squared statistic = 57.73, p-value = 1.59e-05, degrees of freedom = 20.
- **Significance:** Since p < 0.01 (much smaller than 0.05), we reject the null hypothesis.
- **Interpretation:** There is a statistically significant association between a company's sector and its market cap category. This indicates that company size is not randomly distributed across sectors. A heatmap of the differences between observed and expected frequencies showed:
  - Consumer Discretionary has more Mid Cap companies (+11) and fewer Large Caps (-8) than expected.
  - Information Technology has more Mega Cap firms (+6.8) and fewer Mid Caps (-3.3) than expected.
  - Health Care has more Mega Caps (+3.8) and fewer Mid Caps (-5.8) than expected.
  - Financials have more Large Cap companies (+5) than expected.
  - Real Estate has more Mid Caps (+5.1) and fewer Large Caps (-3.4) than expected.
  - Consumer Staples has more Mega Caps (+2.5) and fewer Mid Caps (-4) than expected.

Conclusion: The analysis suggests that certain sectors are more likely to be dominated by firms of a specific size class (e.g., Mega-cap firms in Information Technology and Health Care), while others have a higher concentration of Mid-cap firms (e.g., Consumer Discretionary and Real Estate).

Analysis 6: Market Cap Outliers – Dividend Yields

Question: Do the largest companies by market capitalization exhibit different dividend payout behavior compared to the rest of the market? The top 1% of companies by Market Cap were identified as outliers. A two-sample t-test was performed to compare the mean dividend yields of these outliers to the remaining 99% of companies.

- **Hypothesis Test:**
    - Null hypothesis: There is no difference in the mean dividend yield between the top 1% of companies by Market Cap and the rest.
    - Alternative hypothesis: There is a difference in the mean dividend yield between these two groups.
- **Statistical Test:** Two-sample t-test.
- **Results:** t-statistic = -2.1339, p-value = 0.0333.
- **Significance:** Since $p < 0.05$, we reject the null hypothesis.
- **Interpretation:** The difference in mean dividend yield is statistically significant.
    - Top 1% Market Cap (outliers) Mean Dividend Yield: 0.5757.
    - Rest of companies (non-outliers) Mean Dividend Yield: 1.9181.

Conclusion: The largest firms by market cap tend to pay significantly lower dividend yields on average. This finding suggests that these mega-cap companies may prioritize reinvestment for growth or have different capital structure policies compared to the broader market, rather than returning more capital to shareholders via dividends, which might be a common expectation for large, mature firms.

Analysis 7: Earnings Performance and Stock Price Volatility

Question: Is a company's earnings per share (EPS) associated with differences in stock price volatility? Stock price volatility was measured as the percentage difference between a company's 52-week high and low prices, relative to its 52-week low. Companies were divided into top 10% highest EPS and bottom 10% lowest EPS. A two-sample t-test compared mean volatility between these groups.

- **Hypothesis Test:**
    - Null hypothesis: There is no difference in mean volatility between the top 10% and bottom 10% of companies by EPS.
    - Alternative hypothesis: Mean volatility differs between the highest and lowest EPS groups.
- **Statistical Test:** Two-sample t-test.
- **Results:** t-statistic = 1.640, p-value = 0.1047.
- **Significance:** Since $p > 0.05$, we fail to reject the null hypothesis.
- **Interpretation:**
    - Top 10% EPS firms Mean Volatility: -31.21%.
    - Bottom 10% EPS firms Mean Volatility: -34.43%. While there is a slight numerical difference, it is not statistically significant.

Conclusion: Among the S&P 500 companies in this dataset, there is no statistically significant difference in stock price volatility (as defined) between firms with the highest EPS and those with the lowest EPS. This suggests that stock price volatility is not strongly driven by EPS alone and may depend on other market factors or company-specific characteristics.

4. Conclusions

This project provided a comprehensive analysis of financial data for S&P 500 companies, revealing several key insights into their characteristics and the relationships between their financial metrics.

Key Learnings and Insights:

1. **Data Quality and Preparation:** The initial dataset was relatively clean, with only a minor issue of missing Price/Earnings data, which was handled by row deletion due to its small scale. The inclusion of 52 Week High and 52 Week Low proved valuable for volatility analysis and understanding price ranges.
2. **Fundamental Relationships:**
    - A moderate positive correlation between EPS and Price (r=0.61) was confirmed, with linear regression showing that a $1 increase in EPS corresponds to an approximate $14.85 increase in Price, a statistically significant relationship.
    - Market Cap and Price also showed a moderate positive correlation (r=0.41), though scatter plots indicated this was heavily influenced by outliers.
    - The 52-week high and low prices are very strongly correlated with each other and with the current price, indicating price stability for many stocks within their annual range.
3. **Impact of Profitability:**
    - The relationship between EPS and Price strengthens for profitable companies (r=0.65) but disappears for unprofitable ones (r=-0.056). This underscores that positive earnings are a key driver of stock prices, while negative earnings render EPS less informative for price determination.
4. **Characteristics of Outliers:**
    - Companies in the top 1% by share price are predominantly in Consumer Discretionary, Information Technology, and Health Care sectors. They exhibit significantly higher average EPS and market capitalizations than other companies, suggesting their high prices are backed by strong fundamentals and large scale.
    - Mega-cap companies (top 1% by market cap) tend to pay significantly lower dividend yields than the rest of the market, possibly favoring reinvestment for growth.
5. **Sectoral Differences:** There's a significant association between a company's sector and its market cap category. For instance, Information Technology and Health Care have a higher-than-expected concentration of mega-cap firms, while Consumer Discretionary and Real Estate have more mid-cap companies.

6. **Clustering Financial Profiles:** K-Means clustering (with k=6) successfully identified distinct company profiles, such as "Mega-Cap Growth Leaders" (high price, high EPS, low dividend), "Reliable Income Producers" (low price, low P/E, high dividend), and "High P/E, Low EPS Speculative" stocks, demonstrating diverse financial strategies and market valuations within the S&P 500.
7. **Earnings and Volatility:** No statistically significant difference in stock price volatility was found between the top 10% and bottom 10% of companies by EPS. This implies that factors other than per-share earnings performance may be stronger drivers of stock price volatility for these firms.

Overall Conclusions:

The analysis demonstrated that while general financial principles (like higher earnings leading to higher prices) hold true, the relationships between variables can differ significantly when examining subgroups (e.g., profitable vs. unprofitable firms, or market cap outliers). Sectoral influences are also evident in determining company size. Clustering revealed a detailed view of S&P 500 companies, extending beyond simple classifications. High-priced and high-market-cap companies are generally strong performers in terms of EPS but may have different dividend strategies, often reinvesting for growth. Stock price volatility, at least as measured here, does not appear to be strongly tied to just EPS performance, suggesting that many factors affect how stable a stock is.