IJHSR
ijhighschoolresearch.org
■ RESEARCH ARTICLE

Analyzing Macro-Sector Yield and Technical Indicator Correlation: An Exploratory Data Analysis in R

Ely Hahami

The Lawrenceville School, 2500 Main Street, Lawrenceville, NJ, 08648, USA; elyhahami18@gmail.com

ABSTRACT: Investing in the stock market is one of the world's best ways to generate wealth. This study uses the R programming language and statistical methods to analyze and visualize a dataset publicly available on Kaggle that includes >3,800 stocks, their associated financial indicators, and the sector that classifies that stock in the macro-area. After data preprocessing, we evaluated 95% confidence intervals and a chi-squared test to determine that the financial services sector had the most significant proportion of profitable stocks and, on average, was the most net profitable sector in magnitude. This study then transitions from profitability to technical indicators and establishes a moderate positive correlation between the price-to-earnings ratio and payout ratio. Finally, this study analyzes these financial indicators on a per-sector basis using probability density functions. These inferences allow mathematicians and investors alike to analyze and make appropriate risk-optimization conclusions when diversifying stock portfolios by sector.

KEYWORDS: Mathematics; Probability and Statistics; R; Stock Market; Economics.

■ Introduction

The Kaggle dataset includes more than 200 financial indicators commonly found in the 10-K filings of each publicly traded U.S company released yearly.[1] Additionally, each of the 3809 listed companies within the dataset is part of a sector that classifies it in a macro-area. For example, Microsoft would be classified as a technology stock, and Bank of America Corp. and JPMorgan Chase & Co. would be classified as Financial Services stocks. The data is from 2014, giving stocks ample time to recover from the macroeconomic shock during the 2008 financial crisis that disrupted pricing and trading across the U.S. This study will shed light on the most profitable sector for investors using rigorous statistical analysis.

■ Methods

Data Preprocessing in R:

In our data set, some financial indicator values are missing (specific cells are n/a) or assume non-finite values. To remedy this, we clean the dataset using R's dropna() and fillna() functions. Moreover, to get rid of the cells that falsely assume values of "0" or "0.0," we use alternate data frames that take a subset of the overall data frame using the as.data.frame(subset(variable!=0)) function in R. For this study, variables of interest are all quantitative and continuous, except "sector," which is categorical. Once our data preprocessing is completed, we are ready to analyze the stock data adequately.

■ Results and Discussion

Profitability on a Per-Sector Basis:

Let us define $\hat{p}$ as the sample proportion of net profitable stocks (in the data set, this represents the values where 'PRICE VAR (%)' is positive) for 2014. We can thus construct 95% confidence intervals using the z^* critical value to estimate p, the true proportion of stocks that yield profit, by sector. Our conditions are met: across all sectors, $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$ (see values in Table 1), thus yielding sampling distribution normality. Additionally, the statistic is an unbiased estimator, so the randomness condition is met. However, we must proceed cautiously because the number of stocks (n) in each sector is not less than 10% of the total stocks per sector, thus failing to meet the 10% condition. Table 1 gives the 95% CIs:

Table 1: 95% confidence intervals for the true proportion of stocks that would yield profit.

Sector	Number of stocks (n)	Probability of Success (Investor Profit = $\hat{p}$)	$n\hat{p}$, $n(1-\hat{p})$	95% Confidence Interval (lower bound, upper bound)
Basic Materials	242	0.2190083	53, 189	(0.1685678 0.2764725)
Communication Services	83	0.4337349	36, 47	(0.3252588, 0.5470797)
Consumer Cyclical	457	0.404814	185, 272	(0.3594602, 0.4514079)
Consumer Defensive	165	0.5393939	89, 76	(0.4601951, 0.6171522)
Energy	221	0.1312217	29, 192	(0.08967307, 0.18299164)
Financial Services	660	0.6242424	412, 248	(0.5860459, 0.6613225)
Healthcare	582	0.4484536	261, 321	(0.4075399, 0.4898930)
Industrials	501	0.3293413	165, 336	(0.2883024, 0.3724087)
Real Estate	224	0.4553571	102, 122	(0.3888856, 0.5230259)
Technology	576	0.4513889	260, 316	(0.4102214, 0.4930574)
Utilities	97	0.4329897	42, 55	(0.3326922, 0.5374903)

We thus see that among all the sectors, Financial Services and Consumer Defensive seem to have a high proportion of profitable stocks. We are 95% confident that the true population proportion of stocks that are profitable within the Financial Services sector is between 0.586 and 0.661 and 95% confident that the true population proportion of stocks that are profitable within the Consumer Defensive industry is between

DOI: 10.36838/v5i2.8

0.460 and 0.617. To see if there is a true difference between these sectors' proportions, and since all the conditions are met, we run a chi-squared test in Table 2.

Table 2: Chi-Squared ($\chi 2$) Test.

	Profitable	Not Profitable	Marginal Row Totals
Consumer Defensive	89 (100.2) [1.25]	76 (64.8) [1.94]	165
Financial Services	412 (400.8) [0.31]	248 (259.2) [0.48]	660
Marginal Column Totals	501	324	825 (grand total)

Using R, we calculate the chi-squared statistic ($\chi 2$) to be 3.9846, and the p-value is .045917. Thus, since our p-value is less than α (0.045917 < 0.05), we reject H_0 at the $\alpha = 0.05$ significance level. There is a true difference between these sectors' proportions. Thus, on average, an investor that, in 2014, invested a random Financial Services stock would have the highest statistical probability of profiting than if that investor had chosen to invest in a different sector.

The Extent of Investability on a Per-Sector Basis:

While the above confidence intervals and chi-squared tests provide important information regarding the proportion of profitability among the sectors, one shortcoming is that they fail to highlight the magnitude, or the extent, to which this profitability occurs. This is because a certain sector may have a high proportion of profitable stocks, but said stocks only yield a small profit. Similarly, another sector may have a comparably lower proportion of profitable stocks but said stocks are profitable to a relatively larger extent. Therefore, below we relay a comparative boxplot on a per-sector basis in Figure 1. A high price variance indicates profitability to a large extent; similarly, a low-price variance indicates a dramatic loss in value.

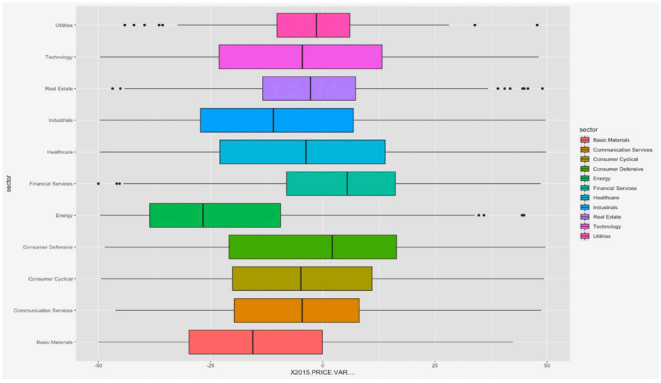

Figure 1: Comparative boxplot of 2015 price variance (%) by sector.

Figure 1 informs us of the price variability distribution. For instance, Q1 and Q3 are relatively equidistant for most sectors, thus yielding a symmetric distribution. However, the first quartile is closer to the median than the third quartile is for the energy sector, thus yielding a right-skewed distribution. Moreover, the boxplot informs us of the spread. For example, the Utilities boxplot has less variability (a low interquartile range), while the distribution for the Healthcare sector is comparatively more spread out. These measures of spread are particularly important in analyzing risk: since, for instance, the Utilities sector is less spread out, it is perhaps a better option for a conservative investor looking to make a secure, small profit.

On the other hand, the highly spread-out data for the Industrials distribution would yield a less appealing option for the said investor because high variability implies that extreme data values (those that yield either a substantial profit or a substantial loss) are more likely to occur. This makes the sectors that are considerably spread out (with high IQRs) more appealing to riskier investors that could afford to incur losses for the sake of a few stocks of their portfolio booming. In addition, the high number of outliers on the high end of price variation highlights that the Real Estate sector, for instance, had a few stocks that yielded an extremely high profit across the year. The opposite phenomenon is valid for the Utilities sector, with extreme outliers on the lower end of the price variability scale — relaying stocks that would lose an investor a large amount of money. Additionally, only two sectors have a positive median Price Variability: Consumer Defensive and Financial Services. This finding, coupled with the aforementioned CIs and sig test, implies that these two sectors have the highest proportion of profitable stocks and are, on the median, net positive in terms of "investability." It is important to note that the Financial Services median price variability is more positive than that of the Consumer Defensive sector, thus deeming Financial Services the most profitable sector. Communication Services, Consumer Cyclical, Healthcare, Real Estate, Technology, and Utilities all have a median stock price variability of approximately 0% (or a price variability that is slightly negative). Basic Materials, Industrials, and Energy all yield, on median, significantly more negative price variabilities.

Technical Indicator Correlation:

A key part of investing is using technical indicators to gauge a stock's characteristics. We explore two key technical indicators — price-to-earnings (P/E) ratio and payout ratio. P/E ratio is one of the most widely used tools by which investors and analysts determine a stock's relative valuation — a high P/E ratio suggests a stock's price is high relative to its earnings and thus potentially signals that the stock is overvalued;[2] similarly, a low P/E ratio might indicate that a stock is undervalued. The payout ratio, another technical indicator, shows the percentage of a company's earnings paid out as dividends (corporate profits) to shareholders.[3] Thus, a low payout ratio suggests a company reinvesting most of its earnings into expanding operations. In contrast, a payout ratio of over 100% indicates that the company is paying out more in dividends than it is earning can support, which some view as an unsustainable practice.

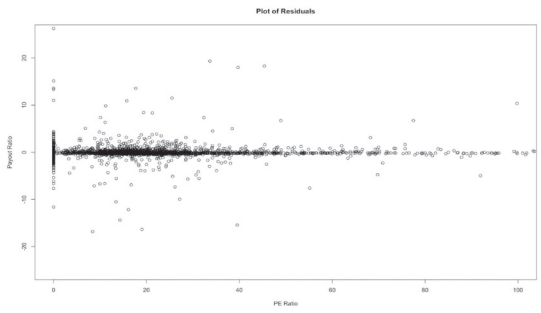

Figure 2: Plot of residuals for P/E ratio and payout ratio for Kendall tau (τ) regression.

DOI: 10.36838/v5i2.8

Accordingly, we ran a Kendall rank correlation tau test — an alternative to Pearson's correlation test — to gauge the relationship between P/E Ratio and the Payout Ratio. We determine the robust Kendall test to be the most representative because our two variables are continuous with a considerable amount of high and low extreme outliers. We also consider the Kendall test (with test statistic τ) the most appropriate over the Pearson correlation or Spearman's rank test (with test statistic ρ) because our data has failed one assumption — the 10% condition — of regression inference.[4] Since the Kendall test is more representative of this scenario, and since our data does contain two continuous variables with a monotonic relationship (thus establishing monotonicity), we use the cor.test() function in R and obtain an output relayed in Figure 3.

```
        Kendall's rank correlation tau

data:  mydata2$PE.ratio and mydata2$Payout.Ratio
z = 31.235, p-value < 2.2e-16
alternative hypothesis: true tau is not equal to 0
sample estimates:
      tau
0.3890106
```

Figure 3: Output from Kendall correlation test.

Since the test's p-value (p-value = $2.2*10^{-16}$) is less than the α = 0.05 significance level, we reject Ho and conclude that there is a moderately positive correlation (true tau = 0.389) between P/E ratio and payout ratio. High payout ratios, as expected, are positively correlated with P/E ratios.

Assessing Stock Valuation through Per-Sector P/E Ratio:

Figure 4 is a kernel density plot that estimates the probability density function (pdf) of sectors' P/E ratio. The P/E ratio distribution is approximately normal for some sectors (i.e., Industrials, Consumer Defensive) and right-skewed for other sectors (Real Estate, Technology, among others). Since the P/E ratio needs to be considered against the backdrop of the P/E for the company's industry (which is why, for instance, comparing mean or median P/E ratio across sectors only would be misguided), these discrepancies in shape inform the appropriate measure of central tendency for a sector's P/E ratio. For the right skewed sector distributions, companies within the said sector should be considered against the backdrop of the median sector P/E ratio because the median is "resistant" to extreme data values; for approximately symmetric distributions, the mean should be used.

In Figure 4, there is a considerable concentration (or a high density) of stocks within the Communication Services sector with high P/E ratios. This finding suggests that a significant number of stocks within the Communication Services sector may be overvalued. Conversely, there seems to be a high density of lower P/E ratio stocks (relative to the sector as a whole) within the Basic Materials sector, potentially implying that said stocks are undervalued because their stock prices trade lower relative to their fundamentals; these stocks might provide an investor a great bargain and will prompt investors to buy the stock before the market corrects the stocks' valuation.

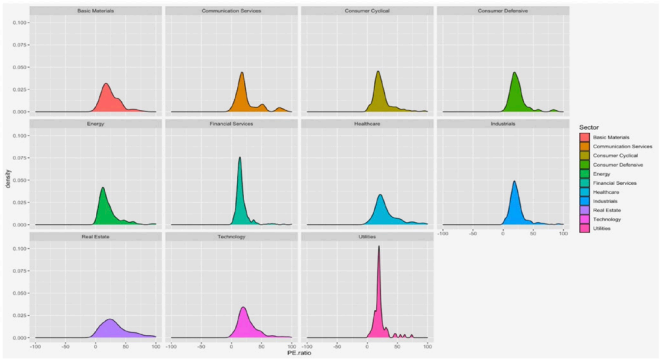

Figure 4: Density plot of P/E ratio distribution by sector.

Assessing Stock Sustainability through Per-Sector Payout Ratio Plots:

We can draw various conclusions from Figure 5. Since a payout ratio over 100% (1) might be viewed as an unsustainable practice, we see a considerable number of stocks within the Real Estate sector that investors might consider unsustainable and thus might short or sell if they own said stocks. Another general practice of investing is that a higher payout ratio is a sign of a strong balance sheet, and thus companies with a 35% to 55% payout ratio signal stability.[2] Thus, investors might be drawn to the technology sector, where there is a high density of stocks (a peak) within the technology sector with a payout ratio between 0 and 1. There also appear to be a few Financial services stocks with desirable payout ratios compared to the sector as a whole. Lastly, there seem to be a few stocks with a negative payout ratio — in which the company had to use existing cash or raise additional money to pay the dividend when it incurred net loss — within the energy sector. This may be a major repellent for potential investors.

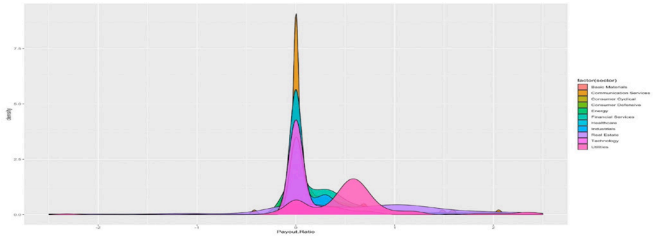

Figure 5: Overlayed kernel density plot of sectors' payout ratio distribution.

■ Limitations

It is important to note that the limitations of this study are broader generalizations. While all conclusions are entirely representative of the year 2014, these conclusions would be accurate for future years only if conditions are similar to those that existed in 2014.

■ Conclusion

In this analytical study, we have representatively analyzed profitability and "investability" on a per-sector basis and highlighted Financial Services to be the most profitable stock market sector. After discovering a moderate positive correlation between P/E ratio and payout ratio, we established appropriate measures of center based on skew for P/E ratio distribution by sector. We specified sectors that might be overvalued or undervalued based on P/E ratio and evaluated stock sustainability

through payout ratio. These inferences and risk-optimization methods are applicable for mathematicians and investors alike to analyze and make appropriate conclusions when diversifying stock portfolios by sector.

Acknowledgments

I want to extend gratitude to Mr. Doug Piper, the instructor of my 'Calculus-Based Probability and Statistics course at the Lawrenceville School, for his guidance and support. I would also like to thank SUNY Maritime Professor Dr. Daniel An, Dr. Robert Teseo of the Wheatley School, and Mr. Michael Gnozzio of the Groton School for their incredibly valuable feedback.

References

1. *Carbone, N. 200+ financial indicators of US stocks (2014–2018). https://www.kaggle.com/cnic92/200-financial-indicators-of-us-stocks-20142018 (accessed May 21, 2022).*
2. Chase, J. P. M. What is a P/E ratio? https://www.chase.com/personal/investments/learning-and-insights/article/what-is-a-p-e-ratio (accessed May 21, 2022).
3. Bloomenthal, A. What is payout ratio? https://www.investopedia.com/terms/p/payoutratio.asp (accessed May 21, 2022).
4. Kendall's tau and Spearman's rank correlation coefficient. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/kendalls-tau-and-spearmans-rank-correlation-coefficient/ (accessed May 21, 2022).
5. Kendall's tau and Spearman's rank correlation coefficient. https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/kendalls-tau-and-spearmans-rank-correlation-coefficient/ (accessed May 21, 2022).

Authors

Ely Hahami is a senior at the Lawrenceville School in NJ with interests in economics and statistics.

DOI: 10.36838/v5i2.8