Portfolio Diversification with Rule Mining

OMSBA 5270 FINAL PROJECT

Harman Birring

8/26/2025

Table of Contents

Introduction

In this project, I've constructed a diversified equity portfolio with an initial investment of about $100,000. This amount is allocated across five different companies: Costco Wholesale (COST), Exxon Mobil (XOM), UnitedHealth Group (UNH), Alphabet/Google (GOOGL), and Bank of America (BAC). The reason for selecting these companies is that they represent distinct sectors of the economy, such as retail, energy, healthcare, technology, and finance, which tends to provide a broad sector diversification. The goal of this project was to combine fundamental financial health analysis with data-driven patterns to develop an investment strategy. In this report, I've evaluated each company's financial strength using key metrics from recent 10-K reports and assigned each company a HealthScore. A key part of this evaluation was the use of association rule mining to uncover any patterns seen among the companies. This has allowed me to construct a balanced portfolio that puts more weight toward financially healthier companies while still maintaining diversification. This project integrates both quantitative fundamentals and data mining techniques. In this project, I've first extracted and normalized critical financial ratios for each company to quantify their financial health. Then, explore patterns in the metrics using the Apriori association rules to see if certain strong financial traits go together.

Then I designed portfolio allocation based on the HealthScores. This was based on a combination of half-weighting by score and half equal weighting. Equal weighting is the baseline if there were no scores at all; each company would be 20%. Pure score weighting, on the other hand, means the bigger slices of money go to the stronger company with the highest score. Thus, I've blended both ideas 50% of the portfolio split among all five companies, and the other 50% based on the proportion to the HealthScore. So, I've combined traditional financial analysis with pattern discovery, and this strategy aims to achieve solid returns with controlled risk. The sections below will cover the methodology, present my findings on each company's financial health, describe the resulting portfolio allocation, discuss the risks associated with each holding and how diversification mitigates them, and examine how one portfolio stock, Google, behaved around a major market event, the COVID-19 pandemic.

Methodology

My methodology was comprehensive, beginning with the extraction of fundamental data from the companies' SEC 10-K filings and annual reports. For each of the five companies, a set of core financial ratios was computed, collectively indicating profitability, efficiency, leverage, and growth. The metrics considered, including Return on Equity (ROE), Return on Assets (ROA), Net Profit Margin, Return on Invested Capital (ROIC), Free Cash Flow Margin, Debt-to-Equity ratio (D/E), and 5-year EPS Growth, were chosen for their ability to capture different dimensions of a company's financial health. The latest available value for each metric as of the most recent fiscal year for each company was obtained, mainly from the SEC data. In cases where a particular metric could not be directly obtained or was not meaningful, it was left as missing. Table 1 below summarizes the raw financial ratios collected for each company:

Table 1. Latest Financial Ratios (Raw Values) – As of latest fiscal year (2024)

Company (Ticker)	ROE (%)	ROA (%)	D/E (ratio)	Net Margin (%)	ROIC (%)	FCF Margin (%)	5-Year EPS Growth (%)
Costco (COST)	31.19	10.55	0.25	2.90	37.4	2.60	~54.35 (≈9% CAGR)
Exxon Mobil (XOM)	12.45	7.43	0.018	N/A	N/A	N/A	18.47 (≈3.4% CAGR)
UnitedHealth (UNH)	14.66	4.83	0.736	3.60	17.55	5.17	1.59 (≈0.3% CAGR)
Alphabet (GOOGL)	30.80	22.24	0.036	28.60	28.32	20.79	–30.38 (negative)
Bank of America (BAC)	9.18	0.83	0.00*	26.63	N/A	N/A	3.14 (≈0.6% CAGR)

The healthScore calculation was also used in this report to compare companies' overall financial health, and by using the Min-Max scaling, I normalized each metric on a 0 to 1 scale across the companies for comparison. For each ratio, we treated higher values as better (for D/E, lower raw debt is better, so I inverted D/E. For each metric, the lowest value among the five companies maps to 0 and the highest to 1. If a metric was missing for a company, it was omitted from that scaling. This gives us scaled metrics that show relative standing from worst 0 to best 1 in each category. We then computed each company's overall HealthScore as the simple average of all its scaled metrics across ROE, ROA, Net Margin, ROIC, FCF Margin, EPS Growth, and DE_inv. This composite score is also between 0 and 1, with higher values indicating stronger all-

around financial health relative to peers. To make it easier to interpret, the HealthScore was put in a ScoreBucket as High if the score is greater than or equal to 0.67, Medium if it's between 0.33 and 0.66, and Low if it's less than 0.33. These ScoreBuckets provide a quick sense of who the standouts are (high), who are average (Medium), and who lag (low) among the five companies.

For pattern discovery, we utilized the Apriori Association Rules to perform pattern mining on the financial attributes. One of the patterns we identified was the creation of flags such as "High_ROE", "High_ROA", "High_Margin", and "High_ROIC". These flags were set to True if the company's scaled ROE (or ROA, Net Margin, ROIC, respectively) was in the top one-third ($\geq$0.67) of the peer range. This process highlighted the top performers for each key financial metric. We then used the Apriori method to find patterns that appeared in at least two companies. For instance, we looked for rules such as: "If a company has high profits and low debt, it usually also has a high overall health score." This helped us understand which combinations of financial traits often go together and how they relate to a company's overall health.

The Portfolio Allocation Strategy was constructed using the HealthScores as a guide. This strategy helped us build the portfolio allocation for the five stocks. We started with an equal 20% allocation to each stock, given that there are five stocks. We then adjusted these weights up or down based on the HealthScores. Specifically, we combined equal weighting and score-proportional weighting in a 50/50 blend. To avoid over-concentration, we imposed bounds. Each stock's weight was capped at 35% maximum and 10% minimum. If any weight fell outside these bounds after the initial calculation, we clipped it and then re-normalized the weights to sum to 100%. This ensured that even the lowest-scoring company still received at least a 10% stake for diversification, and the highest-scoring company could not exceed 35% of the portfolio to prevent one stock from dominating. After determining the final weight percentages, we converted them into dollar amounts out of $100,000 and rounded to the nearest dollar, adjusting for any rounding error so that the total sums exactly to $100,000. We recorded the exact allocation and also visualized it with a pie chart for clarity.

Lastly, I have included an event study scenario testing, which was used as an additional analysis for the company Alphabet (GOOGL) to test the portfolio's behavior, particularly one of its major components, under a significant market event. We chose the COVID-19 pandemic on

March 11, 2020, when the WHO officially declared COVID-19 a pandemic, as this was a major market-moving event. Using historical price data for Google and the S&P 500 index as a market benchmark from Yahoo Finance, we examined Google's stock returns immediately before vs. after the event. We calculated Google's forward 5-day, 20-day, and 30-day returns for periods preceding the event and following the event. We then used Welch's t-test to see if the difference in mean returns before vs. after was statistically significant for each window. This helps assess whether the pandemic declaration had an impact on Google's short-term performance beyond the general market movement. We also plotted Google's stock price over time with the event marked to visualize the context around the pandemic. All data processing and analyses were conducted in Python, utilizing libraries such as pandas for data manipulation, mlxtend for Apriori rule mining, and matplotlib for plotting.

Analysis

In terms of the analysis of the financial health scores and patterns, each company has received a HealthScore between 0 and 1. Table 2 below shows the scores for each firm along with the score categories.

Table 2. Financial HealthScores for the Five Companies

Company	HealthScore (0–1)	Score Category
Alphabet (GOOGL)	0.783	High
Costco (COST)	0.572	Medium
Exxon Mobil (XOM)	0.508	Medium
Bank of America (BAC)	0.330	Low
UnitedHealth (UNH)	0.140	Low

Based on the results, we can see that Alphabet (Google) tends to stand out as the strongest overall, with a score of 0.78, placing it in the high category. This score indicates that Google leads in most of the key metrics that were examined. Google had the highest ROA and net margin by far, exceptionally high ROE, strong ROIC and FCF generation, and carries minimal debt. The only metric dragging its score down was the negative 5-year EPS growth. But despite that EPS dip, Google’s robust profitability and efficiency metrics give it the top composite score. As far as Costco and Exxon Mobil, they both fall into the medium health tier

with scores around 0.57 and 0.51. Excellent returns drive Costco's respectable score; it actually had the highest ROE in the group at 31%, the highest ROIC, and robust EPS growth of 9% annualized over 5 years.

On the other hand, Exxon's score reflects the fact that it has very low debt and has had positive EPS growth in recent years. Bank of America ended up with a score of 0.33. It was categorized as low and had a mixed financial picture, the reason being it has a high net profit margin of 26.6%, which was second only to Google's, thanks to the nature of banking income, but its returns on assets and equity are the weakest of the group, ROA <1%, ROE 9%. Overall, BAC's relatively lower profitability and growth metrics pulled its HealthScore down near the low/medium boundary, earning a low label in this report. UnitedHealth Group has the lowest HealthScore, 0.14, and is clearly in the low health category among these five. UNH is a strong company in absolute terms; 14.7% ROE and 3.6% net margin are decent for a large insurer/provider, but relative to the others, it underperformed on several metrics. Its ROA was relatively low, 4.8% and, unlike the tech or financial firms, it doesn't benefit from extraordinary margins or asset-light economics. It also has the highest leverage of the five, D/E 0.74, much higher than the others.

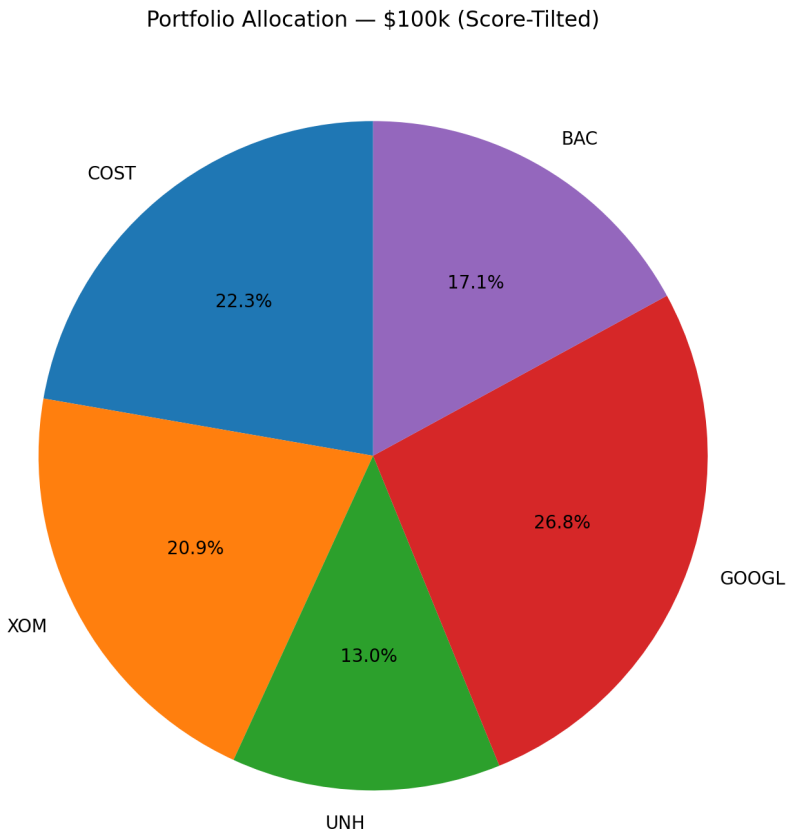
I have also run an association rule analysis (Apriori) on the data set to see if any consistent patterns linked financial health score with performance metrics like earnings growth. There was only one clear pattern that I found and recognized, which was that companies with a medium health score always had a positive earnings growth (EPS growth). This rule had about 40% support and applied to 2 out of 5 companies. In this case, it was Costco and Exxon. The confidence was 100%, meaning every time a company was in the medium group, it also had a positive EPS growth and a lift of 2.5, indicating a much stronger connection than random chance. This all means that medium score companies like Costco and Exxon grew steadily over 5 years, and high score companies like Google did not grow EPS since it was negative. Whereas the low-scoring companies, such as UnitedHealth and Bank of America, had no growth. This indicates that in the middle, companies did better at growth earnings than either the top or the bottom ones in this sample size, indicating the middle companies may still have room to grow. However, we have to consider that the sample size is tiny in this analysis. It's only five companies, so this could be a coincidence.

Portfolio Allocation

The portfolio allocation was performed across the five stocks with a total of $100,000. In terms of allocation, half of the capital was equally weighted, and half was allocated proportionally to each company's HealthScore, with weights capped at 35% maximum and 10% minimum. The larger slices indicate higher combined weight due to stronger fundamentals. The table below illustrates the score portfolio allocation, which provides the exact dollar and percentage allocation to each company:

Table 3. Final Portfolio Allocation ($100k Initial Investment)

Company	Allocation (USD)	Allocation (%)
Alphabet (GOOGL)	$26,779	26.8%
Costco (COST)	$22,262	22.3%
Exxon Mobil (XOM)	$20,890	20.9%
Bank of America (BAC)	$17,064	17.1%
UnitedHealth (UNH)	$13,005	13.0%
Total	$100,000	100%



As we can see, Alphabet (GOOGL) receives the largest allocation at roughly $26.8k (26.8%) of the portfolio. Google's top-ranked HealthScore justifies this overweight position. However, we capped the weight at 35%, and the blend with equal-weighting also prevents any single stock from exceeding a quarter of the portfolio. Here, Google ended up with around 27% of capital, a significant portion, reflecting confidence in its financial strength, but still within diversification limits. Costco is the next largest holding at about $22.3k (22.3%). Exxon Mobil similarly gets about $20.9k (20.9%), close to an equal weight. On the lower end, Bank of America and UnitedHealth have smaller allocations, $17.1k (17.1%) and $13.0k (13.0%), respectively. UnitedHealth, being the lowest score, was initially pushed well below 10% by the score-weighting, but the 10% minimum constraint forced it back up to a baseline 10% weight. Then, after re-normalizing, UNH ended at 13% of the portfolio**.** Thus, no stock is below 10% or above 30% in the final mix, aligning with our risk management rules. This allocation strategy had two goals: to reward stronger companies by giving them more weight based on their financial health, and to stay diversified. Hence, every company still matters in the portfolio. By using the 50/50 approach, a blend of equal weighting and score-based weighing, I have achieved

the balance. The reason being, if I had gone equal weight, every stock would have received $20,000. If I had gone score weight, only Google would have taken 35-40% of the portfolio, while a weaker company like UNH would barely get anything. With the blended approach, Google receives about 27% still the most prominent position, and UNH gets 13% more than its score suggests, but enough to keep the portfolio diversified. Thus, overall, the portfolio has 27% in tech, 43% in consumer and energy sectors, and 30% in finance and healthcare sectors.

Risk Assessment

Building a portfolio that holds only five stocks requires careful attention to the unique risks tied to each company, as well as the ways those risks might overlap or differ in their impact on overall performance. Below, I have outlined the significant sector-related risks associated with each company and explained how diversification works to reduce those exposures. In terms of Exxon Mobil, the most critical risk here is the price of oil and natural gas. These prices fluctuate significantly due to global supply and demand, political events, or decisions by groups like OPEC. When prices fall, Exxon's revenue and profits drop sharply. On top of that, the global shift toward renewable energy creates a long-term risk if demand for fossil fuels slows; Exxon's growth could be limited. In our portfolio, Exxon's risk is balanced by other companies, like Costco or Google, which aren't tied to oil prices. Lower oil costs might even help them. By keeping Exxon at about 21% of the portfolio, we make sure that a significant drop in oil prices won't significantly impact the entire portfolio.

As one of the biggest banks, BAC faces risks that are common in the financial industry. A key risk is that banks must follow strict rules, and new requirements can hurt profits. Recently, Bank of America was fined about $540 million for problems with its risk reporting, showing how costly compliance issues can be. Big banks are under close watch to prevent similar problems from happening again. BAC also faces two financial risks: changes in interest rates, which affect how much money the bank earns on loans, and credit risk if the economy weakens, more customers may default on loans. To reduce risk, BAC makes up only about 17% of the portfolio, and we balance it with companies outside the banking sector. If banks face trouble, such as stricter regulations or loan defaults, those problems won't hit industries such as energy or technology in the same way. This means BAC gives us exposure to banking growth without putting the whole portfolio at risk.

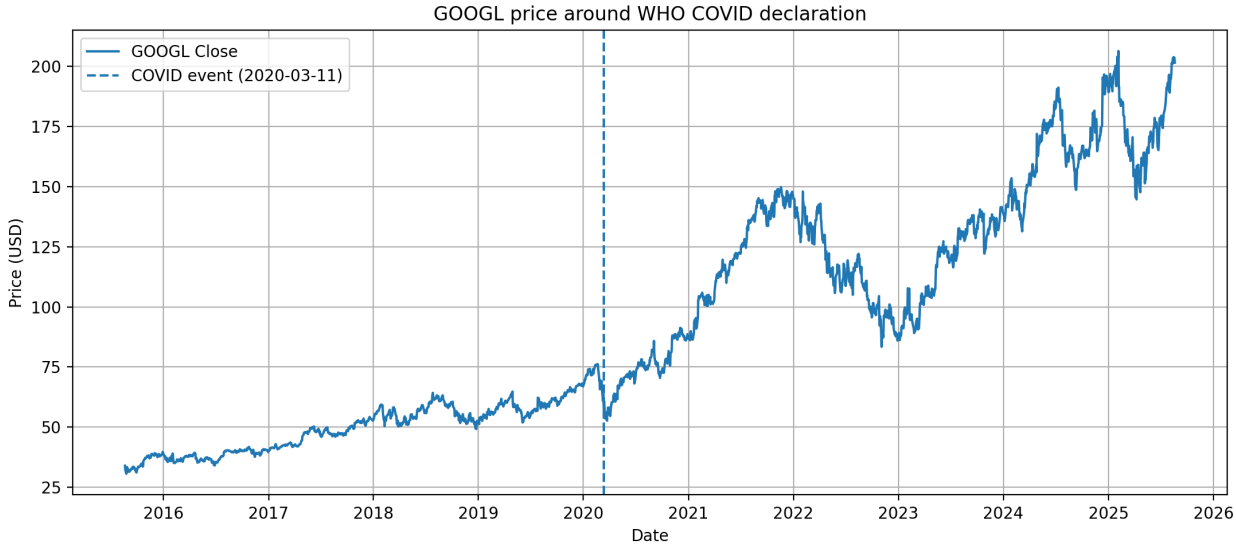
UnitedHealth's main risks come from government regulations and political pressures around healthcare. Since it's a big health insurer, changes to programs like Medicare or Medicaid and new healthcare laws can hurt profits. There's also a legal risk that the company is under government investigation for its handling of Medicare billing, and penalties could be very costly. On top of that, if healthcare costs rise faster than expected, such as due to increased patient utilization or higher medical expenses, it can squeeze UnitedHealth's earnings. UNH makes up about 13% of the portfolio, and stocks balance its risks in totally different industries, like tech, retail, energy, and banking. Problems in healthcare, like stricter rules or fines, usually don't affect those sectors so that the damage would be contained. Also, since UNH is a smaller portion of the portfolio, even big swings in its stock won't move the overall portfolio too much.

Google's most significant risks come from government regulation and competition. Because it dominates online search and advertising, regulators in the U.S. and Europe are watching closely. Antitrust lawsuits and rules like the EU's Digital Markets Act could lead to massive fines or force Google to change how it does business. On top of regulation, Google faces pressure from competition, especially new technologies like AI-driven search or shifts in the ad market that could take revenue away. Google is our largest holding, but still well below a third of the portfolio. Costco runs on skinny profit margins; most of its profit comes from membership fees rather than product sales. That makes it vulnerable if costs rise or sales dip. A downturn in the economy is a key risk, since consumers might spend less, even at discount stores. Costco also faces stiff competition from Sam's Club and Amazon, which puts pressure on prices and margins. Rising supply chain or labor costs could create extra challenges, since Costco can't easily raise prices without losing its low-cost edge. Costco represents about 22% of the portfolio. While its profits could slip in a recession, it often holds up better than other retailers because people view it as a value option. Plus, an economic slowdown that hurts Costco might benefit other holdings, for example, lower oil costs could help Exxon, or lower interest rates could help Bank of America and Google. This mix gives balance, so strengths in other sectors cushion Costco's downside.

Scenario testing

I have examined how Alphabet (GOOGL), our largest holding, performed during the market shock at the start of COVID-19. The key date we focused on was March 11, 2020, when the World Health Organization declared COVID-19 a global pandemic, which resulted in a steep decline in the stock market. To measure Google's performance around this event, we compared its returns in the days and weeks before March 11 to its returns after the announcement, using three-time windows: 5 days, 20 days, and 30 days. For each of these comparisons, we applied statistical tests (Welch's t-tests) to see whether the average returns were significantly different before and after the announcement. For the 5-day window, average returns before the announcement were -11.63%, while after they were -4.52%. The t-statistic was -2.32, with a p-value of 0.0582. Since this p-value is slightly above the 0.05 threshold, we cannot claim a statistically significant difference. This means that in the very short-term window, the drop in returns before the event was not significantly different from the drop immediately after it, and was simply part of the ongoing volatility. For the 20-day window, the difference became clearer. The average return before the announcement was -19.52%, while after it was +13.23%. The t-statistic was -15.50, with a p-value of effectively zero, indicating a statistically significant difference. This shows that over the following month, returns improved notably compared to the steep declines that took place before March 11. For the 30-day window, the results were even more pronounced. Average returns went from -17.81% before the event to +17.80% after the event. The t-statistic was -17.84, and again the p-value was effectively zero, confirming a highly significant difference. This reflects how Google not only stabilized but rebounded strongly within six weeks of the WHO's pandemic declaration. In short, the key takeaway from this event study is that Google has shown the ability to perform well not only in normal market conditions but also under periods of severe stress. By stress-testing our portfolio during the COVID-19 shock, we observed that while Google fell in line with the broader market in the immediate aftermath, it quickly recovered and delivered significantly stronger returns in the weeks that

followed.



Conclusion

In this project, I have developed an investment strategy that not only includes fundamental analysis but also combines it with data-driven insights to build a diversified five-stock portfolio. Companies were selected from 5 different sectors, including consumer retail, energy, healthcare, technology, and finance, to ensure a broad exposure and spread the risk. I have examined each company's financial ratios and converted them into a composite HealthScore to make it easier to visualize which company was the strongest and the weakest. Companies with the strongest fundamentals were Alphabet and Costco, whereas those with the lowest metrics were UnitedHealth and Bank of America. I have also explored association rule mining; however, in this project, our sample size was tiny. It would be more potent with a larger number of stocks. However, I have learned that data mining can complement fundamental analysis by highlighting non-obvious patterns. I have also mentioned the risk assessment, how each company faces challenges within its sector, and how holding a diverse set of companies can save the portfolio as a whole during specific sector challenges. Lastly, I have also scenario tested Google during the COVID-19 pandemic, which was a stress test illustration and provides confidence that our fundamentally strong pick like Google is likely to recover post-crisis, as Google did after 2020.

Thus, our investment strategy is moderately cautious and built on strong fundamentals, aiming for steady returns with controlled risk. The portfolio is diversified so that each company

plays a specific role: Costco adds stability through its reliable consumer business, Exxon provides cash flow and acts as an inflation hedge, UnitedHealth offers defensive strength in healthcare, Google contributes growth and innovation, and Bank of America benefits from shifts in the economic cycle. By blending careful financial analysis with data insights, we created a balanced mix where no single stock dominates, resulting in smoother performance and the potential to capture gains across different market conditions.

Appendix

Appendix A – Raw Financial Ratios (from SEC 10-Ks)

(The table with ROE, ROA, D/E, Net Margin, ROIC, FCF Margin, EPS Growth)

```
Latest ratios (raw):
             AsOf       ROE       ROA        DE NetMargin      ROIC FCFMargin EPS_Growth_5y
COST   2024-09-01   0.31187  0.105498   0.24964  0.028952  0.373994  0.026052      0.543546
XOM    2024-12-31  0.124461  0.074271  0.018078       NaN       NaN       NaN      0.184664
UNH    2024-12-31  0.146589  0.048294  0.736343  0.035987  0.175527  0.051727      0.015952
GOOGL  2024-12-31  0.307976  0.222358  0.036551  0.286037  0.283216  0.207886     -0.303808
BAC    2024-12-31  0.091799  0.008319       0.0  0.266295       NaN       NaN      0.031417
[SAVED] Raw ratios CSV → /Users/Inder/Downloads/latest_ratios_raw.csv
```

Appendix B – Scaled Metrics and Health Scores

(The normalized 0–1 score, HealthScore, and ScoreBuckets)

```
Scaled metrics + Health Score:
             AsOf HealthScore ScoreBucket    ROE_S    ROA_S NetMargin_S   ROIC_S FCFMargin_S EPS_Growth_5y_S  DE_inv_S
COST   2024-09-01    0.572296     Medium      1.0  0.454023        0.0      1.0        0.0             1.0  0.552046
XOM    2024-12-31    0.508254     Medium  0.148418  0.308131        NaN      NaN        NaN        0.576468       1.0
UNH    2024-12-31    0.140236        Low  0.248965  0.186765   0.027365      0.0   0.141198        0.377363       0.0
GOOGL  2024-12-31    0.783118       High  0.982303       1.0        1.0  0.542601        1.0             0.0  0.956919
BAC    2024-12-31    0.329706        Low       0.0       0.0   0.923209      NaN        NaN        0.395614       NaN
[SAVED] Scaled metrics CSV → /Users/Inder/Downloads/scaled_metrics_healthscore.csv
```

Appendix C – Apriori Input Flags

(The table showing High_ROE, High_ROA, High_Margin, etc. as True/False)

Apriori input flags:

	High_ROE	High_ROA	High_Margin	High_ROIC	Low_DE	Positive_EPS_Growth	Positive_FCF	Score_High	Score_Medium	Score_Low
COST	True	False	False	True	False	True	False	False	True	False
XOM	False	False	False	False	True	True	False	False	True	False
UNH	False	False	False	False	False	False	False	False	False	True
GOOGL	True	True	True	False	True	False	True	True	False	False
BAC	False	False	True	False	False	False	False	False	False	True

[SAVED] Apriori flags CSV → /Users/Inder/Downloads/apriori_flags.csv

Appendix D – Association Rules (Apriori Results)

(The rules showing Medium Score ↔ Positive EPS Growth with support/confidence/lift)

Association Rules:

	antecedents	consequents	support	confidence	lift
0	(Score_Medium)	(Positive_EPS_Growth)	0.4	1.0	2.5
1	(Positive_EPS_Growth)	(Score_Medium)	0.4	1.0	2.5

[SAVED] Apriori rules CSV → /Users/Inder/Downloads/apriori_rules.csv

Appendix E – Final Portfolio Allocation ($100k)

(The allocation table by stock and exact totals)

```
=== FINAL $100k ALLOCATION (score-tilted, exact total) ===
            USD
COST    22262.0
XOM     20890.0
UNH     13005.0
GOOGL   26779.0
BAC     17064.0
Total  100000.0
Exact total: $100,000
```

Appendix F – Event Study T-Test Results (COVID-19)

(The 5-day, 20-day, and 30-day window results with p-values and interpretation)

```
T-Test Results (5-day window, COVID event):
Event day used: 2020-03-11
Before mean return: -0.116328
After mean return:  -0.045187
T-statistic:        -2.3157
P-value:            0.0582
No statistically significant difference.

T-Test Results (20-day window, COVID event):
Event day used: 2020-03-11
Before mean return: -0.195157
After mean return:  0.132315
T-statistic:        -15.5022
P-value:            0.0000
Statistically significant difference.

T-Test Results (30-day window, COVID event):
Event day used: 2020-03-11
Before mean return: -0.178083
After mean return:  0.177976
T-statistic:        -17.8389
P-value:            0.0000
Statistically significant difference.
```