Process Book – Data Processing

My motivation for undertaking this project arose from my interest in combining technology and finance. As someone who will be interning at a hedge fund next January, I saw this as an ideal opportunity to strengthen my data processing skills while deepening my understanding of financial markets. At the start, I had limited experience working with time-series data, and I quickly learned that financial datasets often come in wide, complex formats. The raw dataset I worked with contained 503 companies across 368 columns of daily price and volume information. Understanding how to transform this into a usable structure became my first major challenge.

Initially, I explored the dataset to identify patterns within the column names, such as the recurring suffixes "_opening", "_closing", and "_volume". This led me to separate the dataset into three corresponding tables. Through this process, I began to appreciate how deeply the structure of a dataset influences what analyses are possible. A key breakthrough came when I learned how to use the melt function, which allowed me to convert the dataset from a wide format into a tidy, long-format table suitable for time-series operations. This step, combined with the use of regex to extract dates and metric types, significantly improved my confidence in manipulating complex datasets.

Feature engineering was another important learning phase. I computed daily returns, cumulative returns, volatility, and dollar volume—metrics that are central in quantitative finance. Developing a weighted contribution metric by multiplying cumulative return with average dollar volume helped me quantify which companies meaningfully influenced the S&P 500. This taught me that raw performance alone does not tell the full story; liquidity and market participation amplify a stock's impact on an index

During the exploratory data analysis stage, I experimented with various visualizations, including sector distributions and risk–return scatterplots. This taught me not only the technical skills required to build meaningful visuals, but also how visual design influences interpretation. For example, visualizing cumulative returns against volatility with bubble sizes representing dollar volume instantly highlighted patterns that were not obvious in raw tables. I learned the importance of structuring a narrative through visuals, from sector-level overview to company-level attribution.

At the company level, I identified top contributors to index performance and observed a heavy concentration of gains in the Information Technology sector. This motivated me to investigate the underlying drivers of cumulative returns. Implementing a linear regression model required additional learning, particularly around preprocessing. I came to understand why scaling numeric variables and one-hot encoding categorical variables is essential, and why pipelines are considered best practice. Although the regression model produced a negative R^2 value, this was a valuable learning point. I

realized that short-horizon equity returns are inherently noisy and difficult to predict. Even so, the model provided interpretability into how risk, liquidity, and sector membership influence stock performance.

Aggregating results at the sector level reinforced the earlier findings: Information Technology overwhelmingly led the S&P 500 in the first half of 2025, while sectors such as Energy and Real Estate lagged. Conducting analysis at both the company and sector level taught me how different levels of granularity reveal different aspects of market behaviour.

Overall, this project taught me far more than technical data processing. It helped me understand how to approach a problem systematically, from data cleaning to feature engineering, visualization, modelling, and narrative construction. I learned to think like a data scientist: to question what the data represents, determine what story it can realistically tell, and design analyses that extract meaningful insight. Completing this project has strengthened my confidence in working with real-world financial datasets and prepared me for both my upcoming hedge fund internship and future data science roles.