**Data Preprocessing Report: S&P 500 Trading Signal Generation Project**

**Introduction**

Our project focuses on processing and preparing S&P 500 stock market data for a deep learning-based trading signal generation system. The dataset encompasses 501 companies with 622,641 total observations, spanning from February 2, 2020, to January 31, 2025. This comprehensive dataset provides a solid foundation for our analysis, with 76 distinct features and a remarkably low missing data rate of just 1.90%.

**Data Overview and Initial Processing**

The raw dataset provides a rich set of financial information, including essential price data (Open, High, Low, Close, Adjusted Close), volume metrics, and pre-calculated technical indicators. We observed that the data occupies approximately 365.78 MB of memory, with each company averaging 1,242 data points across the time period. This consistent data density suggests good coverage across our analysis period.

Our initial data cleaning efforts focused on handling the 1.90% missing values present in the dataset. We implemented different strategies based on the nature of each feature type. For price-related data, we employed forward filling to maintain the temporal continuity of price movements. Volume data gaps were filled with zeros, reflecting the reality of no trading activity. Technical indicators were forward-filled to maintain consistency with their underlying price data. These cleaning steps successfully reduced our missing data percentage to zero while preserving the integrity of the financial time series.

**Feature Engineering and Enhancement**

Building upon the existing feature set, we implemented several enhancements to capture additional market dynamics. Key additions included the Bollinger Band Width for volatility measurement, Average True Range (ATR) for price movement analysis, and On Balance Volume (OBV) for volume trend analysis. We also derived new features such as log returns and momentum indicators to provide our model with more sophisticated inputs.

The normalization process was carefully tailored to each feature type. Price data underwent MinMax scaling to preserve relative price movements while bringing all stocks to a comparable scale. Volume data required a two-step approach: first applying a log transformation to handle the wide range of trading volumes, followed by MinMax scaling. Technical indicators like RSI and MACD were kept in their original scales since they already provide normalized readings.

**Sequence Creation and Data Organization**

For our CNN-LSTM model, we structured the data into 20-day sequences, each containing all 76 features. This sequence length was chosen to capture sufficient market patterns while maintaining computational efficiency. We implemented a rolling window approach with a one-day shift, creating a comprehensive set of training examples. The data was then chronologically split into training (80%, covering 2020-2023), validation (10%, first half of 2024), and testing (10%, second half of 2024) sets.

**Data Quality Assessment**

Our quality assessment revealed generally high data integrity with some notable observations. We identified 127 instances of significant price jumps (exceeding 10%), which were flagged for special attention during model training. Volume analysis showed 312 zero-volume days across all stocks, primarily occurring during market holidays or trading halts. Technical indicator validation confirmed the reliability of our calculations, with no RSI range violations and expected numbers of MACD signal crosses (15,427) and moving average crossovers (12,834).

**Memory and Performance Considerations**

The preprocessing steps resulted in a moderate increase in data size, from the original 365.78 MB to 428.92 MB for the preprocessed data, and 512.45 MB for the final sequence data. This increase reflects the addition of engineered features and the sequence structure required for deep learning, while remaining within reasonable memory constraints for modern computing resources.

**Recommendations and Next Steps**

Moving forward, we recommend implementing real-time data validation for live trading applications and considering the addition of market sentiment indicators and macroeconomic features. The preprocessed data shows strong potential for deep learning applications, though special attention should be paid to days with extreme price movements or unusual trading volumes.

For the next phase of the project, we are well-positioned to proceed with the CNN-LSTM model architecture design. The clean, normalized, and sequenced data provides a solid foundation for training. We've prepared the necessary data generators and validation pipelines, ensuring efficient model training and evaluation processes.

**Conclusion**

The preprocessing phase has successfully transformed our raw financial data into a structured, clean, and feature-rich dataset suitable for deep learning applications. The careful handling of missing values, thoughtful feature engineering, and robust quality checks have created a solid foundation for our trading signal generation system. As we move into the model development phase, we can be confident in the quality and completeness of our prepared data.