James Mortensen

Data Curation and Preprocessing for Value Investing Analysis

Milestone 2: One Pager

**Problem Domain**: Value investing is an investment strategy focused on identifying undervalued stocks with strong financial fundamentals. Analyzing financial metrics and market data is crucial to making informed investment decisions. However, financial data is often messy, incomplete, and sourced from multiple providers, making data preprocessing and curation a critical step in value investing analysis.

Data curation and preprocessing ensure that the collected data is accurate, consistent, and suitable for analysis, enabling investors to make better investment decisions and improve their investment strategies.

**Data Source:** Financial data for this project will be collected from at least two sources Yahoo Finance, Alpha Vantage. The dataset will include the following information for publicly traded companies:

1. Fundamental financial data: Key financial ratios such as price-to-earnings (P/E), price-to-book (P/B), price-to-sales (P/S), dividend yield, return on equity (ROE), return on assets (ROA), and debt-to-equity (D/E) ratio.
2. Historical stock prices: Daily or monthly stock prices, including adjusted closing prices that account for dividends, stock splits, and other corporate actions.
3. Market and sector data: Index levels, sector classifications, and other macroeconomic indicators to analyze market trends and compare stocks within their industry.

**Data Curation and Preprocessing Steps:** The main objective of the project is to curate and preprocess the collected financial data to make it suitable for value investing analysis. The following steps will be carried out:

1. Data collection: Gather financial data from multiple sources through APIs, web scraping, or downloading historical datasets.
2. Data integration: Combine the collected data into a unified dataset, ensuring the data is consistent and aligned by date, company ticker, and other relevant attributes.
3. Handling missing values: Investigate and address missing values in the dataset, using methods such as data imputation, interpolation, or removal, depending on the nature and extent of missing data.
4. Outlier detection and treatment: Identify and handle outliers in the financial metrics, which may be caused by data errors, corporate actions, or extreme market events. Techniques such as the interquartile range (IQR) or Z-score can be used to detect outliers.
5. Data transformation: Apply necessary transformations to the data, such as calculating financial ratios, adjusting stock prices for corporate actions, or standardizing the data for analysis.
6. Data validation: Ensure the curated dataset's accuracy and consistency by cross-referencing with other data sources, checking for discrepancies, and addressing any data quality issues.

**Expected Outcomes:** The primary outcome of this project is a clean, consistent, and accurate dataset that can be used for value investing analysis. By curating and preprocessing the financial data, investors will have access to reliable information for making informed investment decisions. This project will lay a solid foundation for further research and analysis in value investing, such as identifying undervalued stocks, developing investment strategies, or applying machine learning techniques for stock prediction.