Written Summary: Data Sources and Exploratory Data Analysis (EDA)

1. Data Sources Used

For this project, two primary categories of financial data were utilized to support the development of an AI-driven trading assistant.

a) Historical Market Data (OHLCV)

Type: Open, High, Low, Close, Volume (OHLCV)

Time Period: January 2022 – December 2024 (business days)

Companies Covered: Top companies by market capitalization across technology, finance, retail, and energy sectors

Intended Real-World Sources:

- Yahoo Finance
- Alpha Vantage
- Tiingo

Use Case: This dataset captures price movements, trading volume, and market volatility. It forms the foundation for technical indicators and machine learning-based trading signal generation.

b) Fundamental Financial Data

Type: Company-level financial metrics

Features Included:

- Revenue
- Net Income
- Price-to-Earnings (P/E) Ratio
- Debt-to-Equity Ratio
- Return on Equity (ROE)

Intended Real-World Sources:

- SEC EDGAR filings
- Yahoo Finance
- Alpha Vantage

Use Case: Fundamental data complements price-based indicators by providing insights into financial health, valuation, and profitability, enabling more robust trading strategies.

2. Data Cleaning and Preprocessing

The following preprocessing steps were applied to ensure high-quality data for machine learning:

- Converted all date fields to a standardized `datetime` format.
- Removed missing values generated by rolling window calculations.
- Ensured logical OHLC consistency: Low $\leq$ Open/Close $\leq$ High.

- Normalized numerical features where appropriate.
- Engineered technical indicators including:
  - Daily returns
  - 20-day moving average (MA-20)
  - 50-day moving average (MA-50)

These steps ensured that the datasets were clean, structured, and suitable for downstream model training.

3. Exploratory Data Analysis (EDA) Insights

a) Price Trends and Volatility

- Stock prices exhibited clear bullish and bearish phases over time.
- Technology stocks showed higher volatility compared to defensive sectors.
- Trading volume spiked during sharp price movements, indicating increased market participation.

b) Moving Average Behavior

- MA-20 responded quickly to short-term price changes, capturing momentum.
- MA-50 provided smoother long-term trend signals.
- Crossovers between MA-20 and MA-50 often preceded trend reversals.

c) Return Distribution

- Daily returns were centered around zero, as expected.
- Return distributions exhibited fat tails, indicating extreme price movements.

d) Fundamental Data Insights

- Higher ROE and consistent revenue growth correlated with stronger long-term performance.
- High P/E ratios were linked to growth stocks with higher volatility.
- High debt-to-equity ratios increased downside sensitivity during market downturns.

4. Notable Patterns and Anomalies

- Sudden volume spikes without proportional price changes suggested institutional activity or news anticipation.
- Short periods of abnormal volatility aligned with earnings announcements and macroeconomic events.
- Outliers in daily returns emphasized the importance of robust risk management.

Conclusion

The exploratory analysis confirms that combining historical price data with fundamental financial indicators provides a strong foundation for building adaptive, machine learning-driven trading strategies.