Financial Technologies – AI-Powered Robo-Advisory Simulation

22290617 – Onur Yiğit KOCATÜRK

2290042 – Ömerfaruk SARIBAL

22290230 – Ahmet ATAR

1. Data Source Selection and Data Description

In this project, we utilize publicly available financial market data from Investing. Investing is chosen because it provides:

- *Reliable and up-to-date historical price data*
- *Easy access without requiring an API key*
- *High-quality time series suitable for forecasting models such as ARIMA and LSTM*
- *Consistent data for equities, ETFs, bonds, and commodities*

These make it an ideal source for building an AI-powered robo-advisory simulation.

Number of Observations (N)

In total, the cleaned dataset contains k = 6 numeric variables (Close, Open, High, Low, Volume, Change) and 1 date variable (Date).

Each asset contains approximately 2.500 daily observations spanning the period from January 2015 to November 2024.

6 assets x 2.500 observations makes approximately 15.000 data rows.

Variable	Type	Unit	Description
Date	Date	–	Trading day…
Close	Numeric	USD	Closing price…
Open	Numeric	USD	Opening price…
High	Numeric	USD	Highest intraday price
Low	Numeric	USD	Lowest intraday price
Volume	Numeric	Shares	Normalized trading volume
Change	Numeric	Decimal	Percentage change in close

The dataset does not include categorical variables (such as sector or rating classes), therefore no frequency tables or bar/pie charts are reported.

2. Selected Asset Set

To construct a diversified and realistic portfolio universe, we include six widely-used financial instruments:

- *AAPL – Apple Inc. (technology stock)*
- *MSFT – Microsoft Corp. (technology stock)*
- *SPY – S&P 500 ETF (broad market benchmark)*
- *QQQ – NASDAQ-100 ETF (growth/tech heavy)*
- *TLT – 20+ Year Treasury Bond ETF (low-risk fixed income)*
- *GLD – Gold ETF (commodity hedge asset)*

This set provides representation across different risk levels, allowing the model to create meaningful conservative, balanced, and aggressive portfolios.

3. Time Range Selection

We selected the time period January 1, 2015 – November 1, 2024, covering nearly 10 years of market activity. This range was chosen because it includes multiple important market regimes:

- *Pre-pandemic bull market*
- *COVID-19 crash (2020)*
- *Post-pandemic recovery*
- *High-inflation and interest-rate cycle (2022–2023)*
- *Recent stabilization phase (2024)*

Using such a broad window ensures that the forecasting models observe different market conditions, improving robustness and making the risk/return analysis more realistic.

4. Why This Dataset Works for Our Project

This dataset supports all key components of the robo-advisor:

- *Forecasting:* Sufficient long-term historical data for ARIMA and LSTM
- *Risk Analysis:* High-quality daily returns for volatility computation
- *Portfolio Construction:* Assets that naturally map to different investor risk profiles
- *Scenario Testing:* Includes crisis and recovery periods
- *ESG Integration:* Selected assets have publicly available ESG scores

Overall, this combination provides a strong foundation for building and evaluating an AI-powered investment recommendation system.

Data Cleaning & Preprocessing

Investing.com datasets use European formatting (comma decimals, dot thousands).
The following steps were applied:

Cleaning numeric fields

- Replace "." (thousand separator)
- Replace "," with "." (decimal separator)
- Convert to numeric

Example: "250,42" → 250.42

Volume normalization

Volumes contain "M" or "B" units (millions, billions).
These are converted into numerical form:

- "39,48M" → 39.48×10^6
- "1,2B" → 1.2×10^9

Sorting and indexing

- Dates converted from string (dd.mm.yyyy)
- Sorted chronologically (oldest → newest)
- New index assigned

This ensures correct time-series behavior

Missing data and handling

After cleaning, there are essentially no missing values in prices or returns. The only missing entries are in the Volume column: **5 observations for SPY** and **26 observations for GLD**. All other fields (Date, Close, Open, High, Low, Change) have zero missing values for all assets. Since Volume is not used directly as a model input in this assignment and the proportion of missing rows is below 1%, we did not apply explicit imputation. In all return-based calculations, rows with missing values are naturally ignored via dropna().

Ticker	Date	Close	Open	High	Low	Volume	Change
AAPL	0	0	0	0	0	0	0
MSFT	0	0	0	0	0	0	0
SPY	0	0	0	0	0	5	0
QQQ	0	0	0	0	0	0	0
TLT	0	0	0	0	0	0	0
GLD	0	0	0	0	0	26	0

Closing Prices

The closing price trends of the selected assets reveal clear differences in market behavior across stocks, bonds, and commodities during the 2015–2024 period. Equity assets—AAPL, MSFT, SPY, and QQQ—show strong upward long-term trends, reflecting the post-2015 technology-driven bull market and the significant acceleration in growth after the COVID-19 market recovery. Among these, QQQ and AAPL exhibit the highest growth rates, consistent with the dominance of large-cap technology companies over the last decade.

TLT, the long-term U.S. Treasury bond ETF, behaves very differently. Its closing prices display a downward trend beginning in 2021, primarily due to rising interest rates and inflationary pressures. This makes TLT a counter-cyclical asset relative to equities, providing diversification benefits.

GLD, representing gold prices, shows a steady but moderate upward trend, with sharp increases during periods of heightened uncertainty—most notably during the 2020 COVID crisis and again during the inflation spike in 2022–2023. This confirms gold's role as a defensive hedge asset.

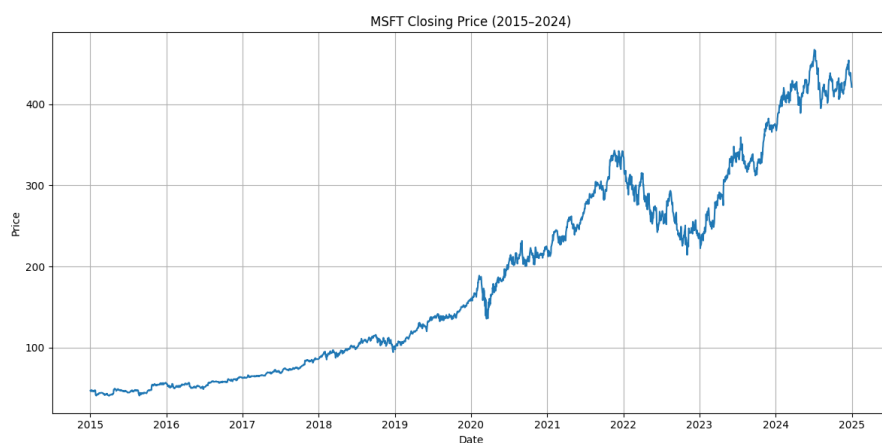Overall, the closing price analysis highlights three distinct behavior groups:

1. **High-growth equities:** *AAPL, MSFT, QQQ, SPY*
2. **Defensive commodity:** *GLD*
3. **Interest-rate-sensitive fixed income:** *TLT*

These patterns support the construction of diversified portfolios tailored to different risk levels, aligning with the objectives of the robo-advisory system.
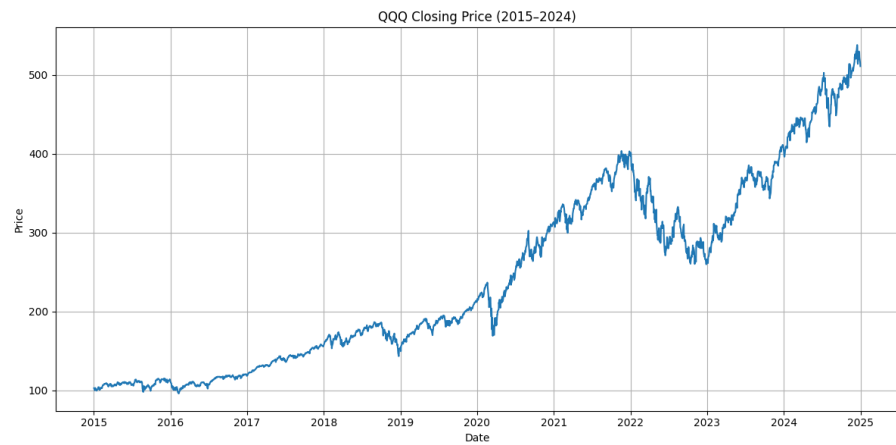
AAPL

MSFT

SPY

SPY Closing Price (2015–2024)

QQQ

QQQ Closing Price (2015–2024)

TLT

TLT Closing Price (2015–2024)

GLD

GLD Closing Price (2015–2024)

30-Days Moving Average

The 30-day moving average (MA30) provides a smoothed representation of short-term price dynamics and helps identify trend direction while reducing noise caused by daily volatility. Across the assets in our dataset, the MA30 line closely tracks the underlying price series, but with delayed reactions at turning points, which is characteristic of moving average indicators.

For equity assets (AAPL, MSFT, SPY, QQQ), the MA30 clearly highlights prolonged upward trends, especially during major growth periods such as 2017–2019 and the post-pandemic surge beginning in mid-2020. The MA30 consistently stays below the closing price during strong bullish momentum, confirming sustained trend strength. Periods where the MA30 and price converge or cross indicate short-term corrections or consolidations.

In contrast, TLT's MA30 demonstrates a persistent downward trend after 2021, reflecting rising interest rates and bond price pressure. The smooth decline in MA30 visually reinforces the long-term weakening of the bond market during this period.
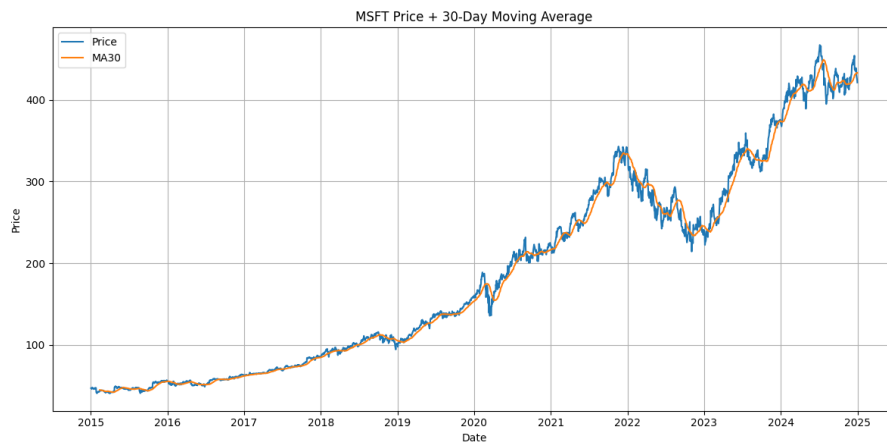
GLD's MA30 shows moderate but steady upward movement with noticeable upward shifts during global uncertainty—most prominently in 2020 and again during 2022–2023. The MA30 confirms gold's role as a safe-haven asset with cyclical surges aligned with macroeconomic stress.

Overall, the MA30 analysis helps distinguish stable long-term trends from short-term volatility, offering useful insight for the robo-advisor's risk-based portfolio recommendations. By smoothing price action, MA30 supports more reliable trend detection, which is valuable for both return forecasting and regime identification.
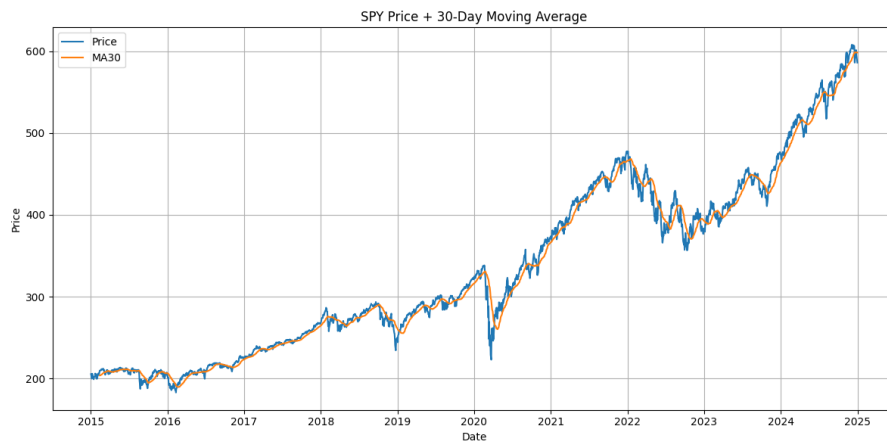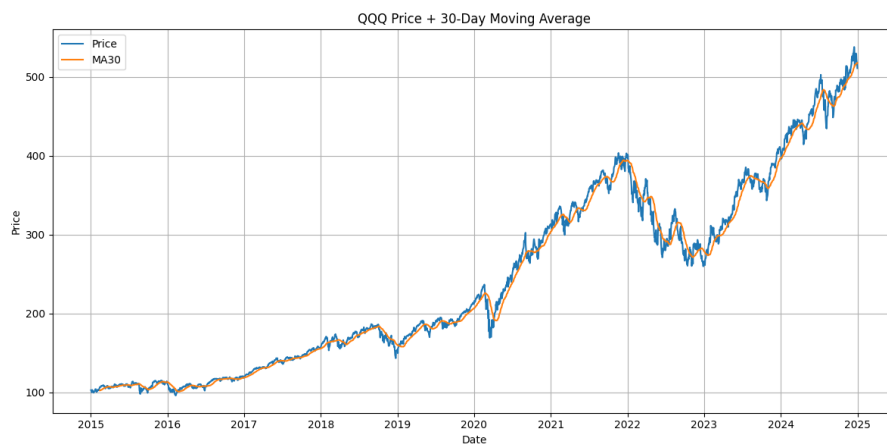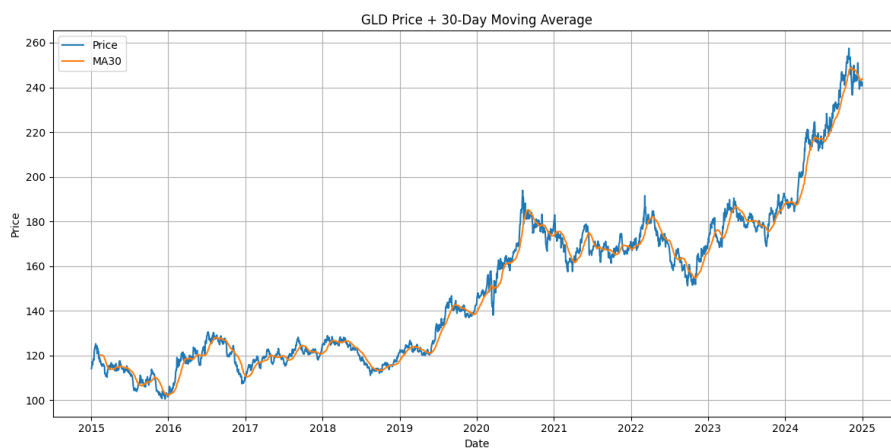
AAPL

AAPL Price + 30-Day Moving Average

MSFT

MSFT Price + 30-Day Moving Average

SPY

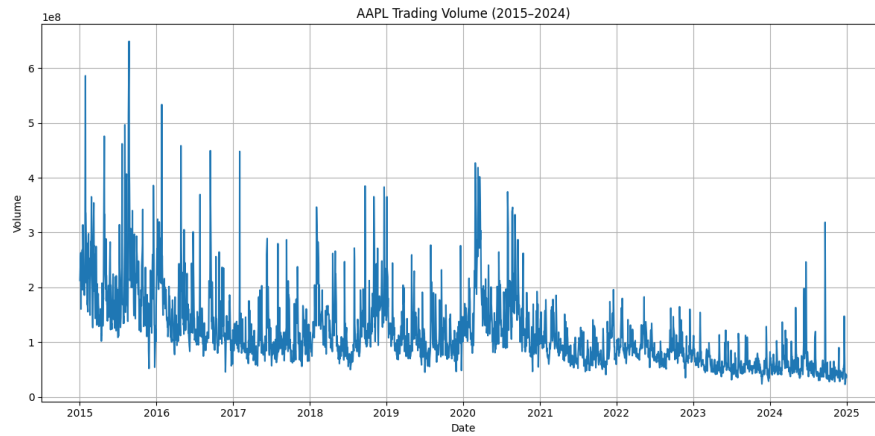SPY Price + 30-Day Moving Average

QQQ

GLD

Volume

The trading volume patterns across all assets reflect how investor attention, market stress, and macroeconomic conditions evolved over the last decade. Equity assets such as AAPL, MSFT, SPY, and QQQ show very high and volatile trading activity between 2015 and 2018, with multiple sharp spikes associated with earnings cycles, product announcements, and broader market events. Following the COVID-19 shock in 2020, all four equities experience another wave of extreme volume surges, indicating intensified investor reactions during periods of uncertainty. After 2021, however, volumes gradually stabilize and decline, suggesting a shift toward more mature, less speculative trading behavior in large-cap technology names.

TLT exhibits a different structure: its most significant volume increases occur after 2021, during the rapid rise in interest rates. These spikes reflect heightened investor repositioning within the bond market as long-term Treasury prices fell sharply. GLD, meanwhile, maintains moderate day-to-day volume but shows clear spikes during global stress periods such as the 2020 pandemic and the inflation-driven environment in 2022–2023, reinforcing gold's reputation as a defensive asset that attracts flows during uncertainty.
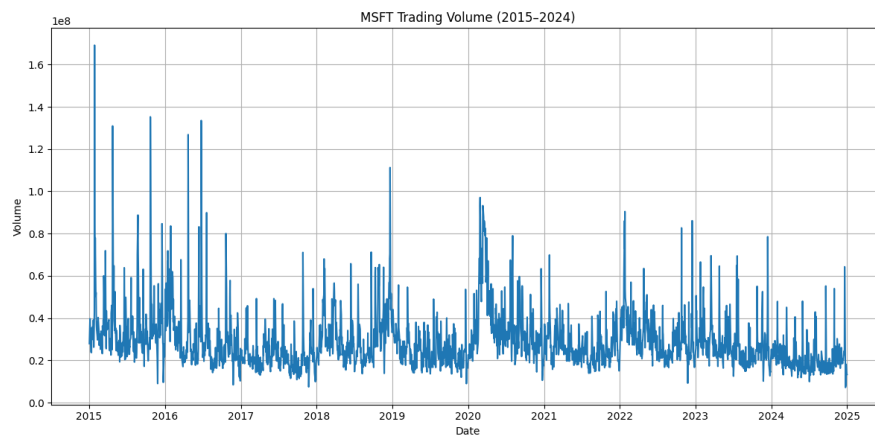
Taken together, the volume analysis reveals important differences in how these assets respond to market events. Equities experience broad market-driven cycles of intense trading, bonds

react primarily to monetary policy shifts, and gold attracts attention during risk-off phases. This overall pattern helps the robo-advisor interpret liquidity conditions and understand when price movements are accompanied by strong market participation, which is valuable for assessing trend reliability and investor sentiment.
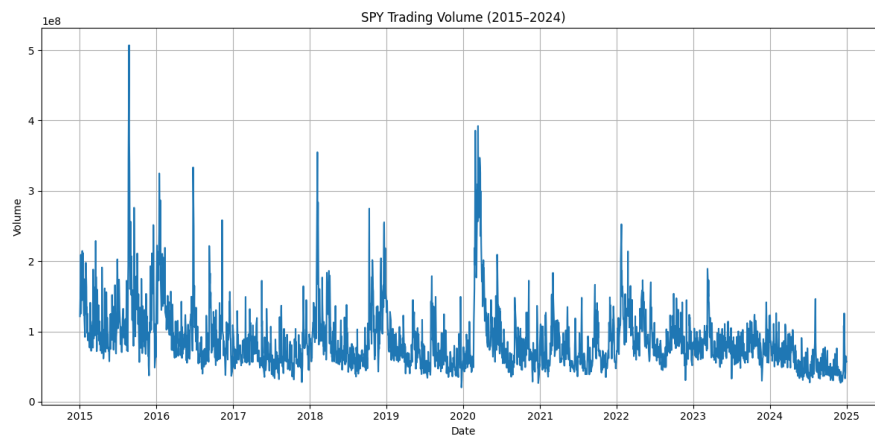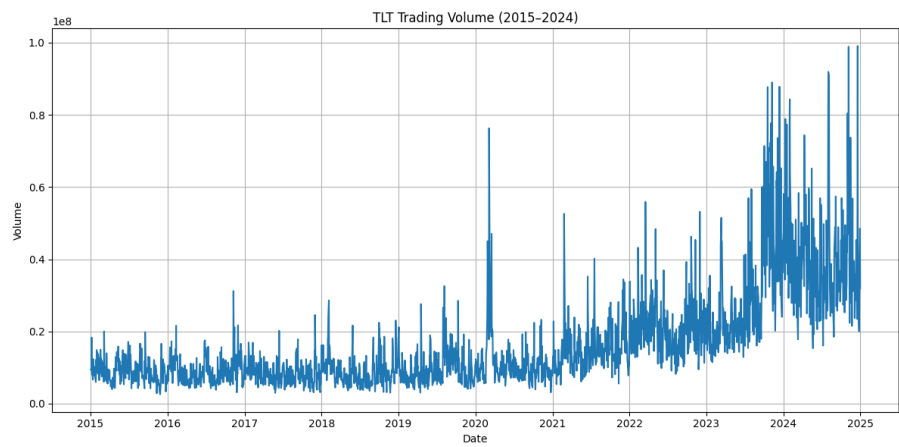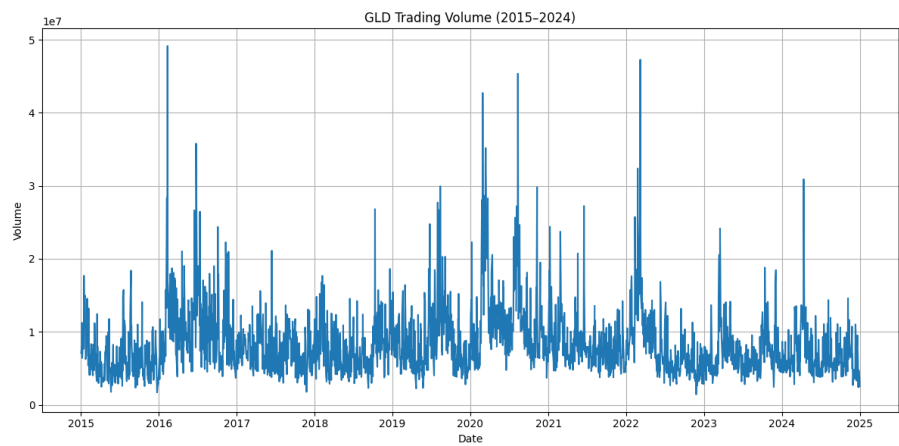
AAPL

MSFT

SPY

TLT

TLT Trading Volume (2015–2024)

GLD

GLD Trading Volume (2015–2024)

Daily Return

Descriptive statistics of daily returns

Ticker	mean	median	std	min	max	skewness	kurtosis
AAPL	0.00104	0.00086	0.01794	-0.12865	0.11972	0.00083	5.30210
MSFT	0.00102	0.00091	0.01710	-0.14739	0.14217	0.07967	7.79462
SPY	0.00048	0.00057	0.01114	-0.10942	0.09060	-0.55489	12.59089
QQQ	0.00073	0.00116	0.01375	-0.11980	0.08469	-0.37755	6.26875
TLT	-0.00010	0.00029	0.00965	-0.06668	0.07520	0.09718	4.51237
GLD	0.00034	0.00039	0.00885	-0.05369	0.04912	-0.10447	2.81175

Equity assets (AAPL, MSFT, SPY, QQQ) have positive mean daily returns and the highest standard deviations, confirming their high-risk–high-return profile. TLT shows a slightly negative mean with lower volatility, while GLD has modest positive returns and the smallest
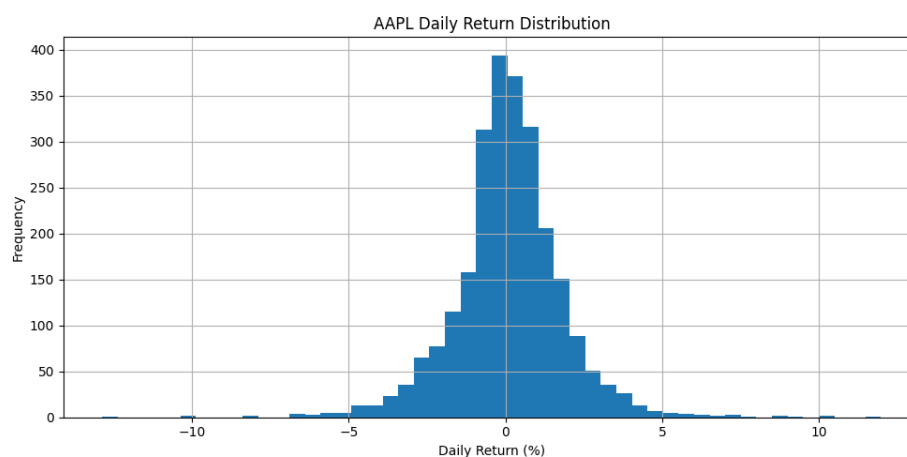
dispersion. Kurtosis values well above 3, especially for SPY, indicate heavy tails and the presence of extreme events compared to a normal distribution.

The daily return distributions provide a clear view of the short-term risk and volatility characteristics of each asset. Equity assets such as AAPL, MSFT, SPY, and QQQ display bell-shaped distributions centered near 0%, with relatively wider spreads compared to bonds and gold. This indicates that equities experience more frequent and larger day-to-day price movements, consistent with their higher risk–higher return profile. Occasional extreme observations on both the positive and negative sides correspond to earnings surprises, macroeconomic announcements, and large market shocks—most notably the COVID-19 period, where volatility temporarily expanded across all equities.
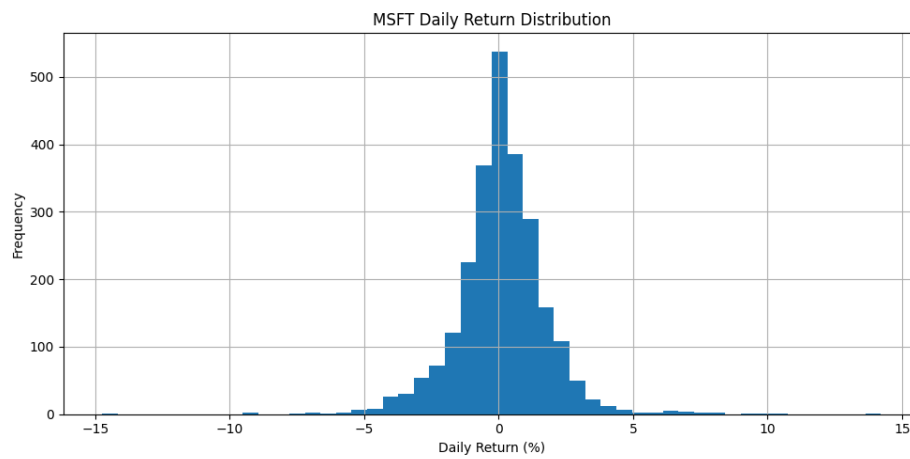
TLT shows a distribution that is narrower but slightly skewed, reflecting the bond market's sensitivity to interest rate changes. The presence of several larger negative daily returns after 2021 aligns with the rapid rise in interest rates that caused significant repricing of long-term Treasury bonds. GLD, in contrast, exhibits a tighter and more symmetric distribution with fewer extreme movements. This behavior is typical of gold, which tends to show lower short-term volatility and acts as a stabilizing component in diversified portfolios.

Overall, the return distributions highlight the clear distinction between high-volatility assets (equities), interest-rate-sensitive assets (TLT), and defensive assets (GLD). These patterns are important for understanding risk contributions within the portfolio and provide the robo-advisor with a quantitative basis for aligning asset weights with different investor risk profiles.
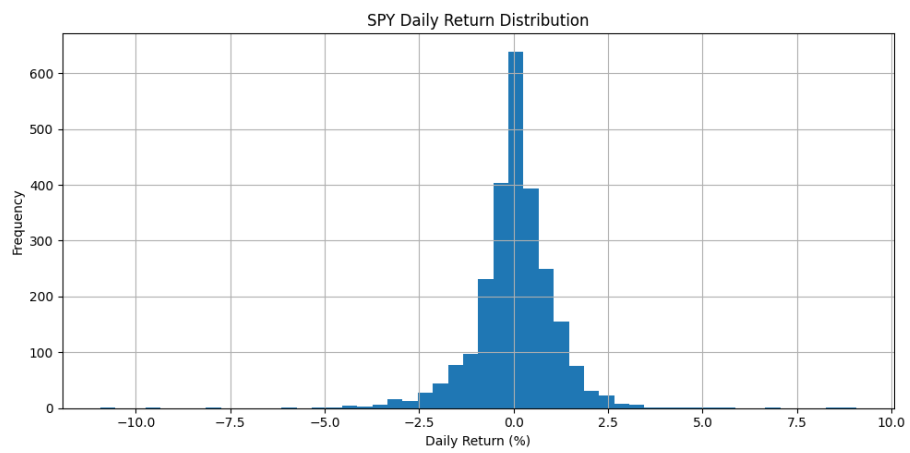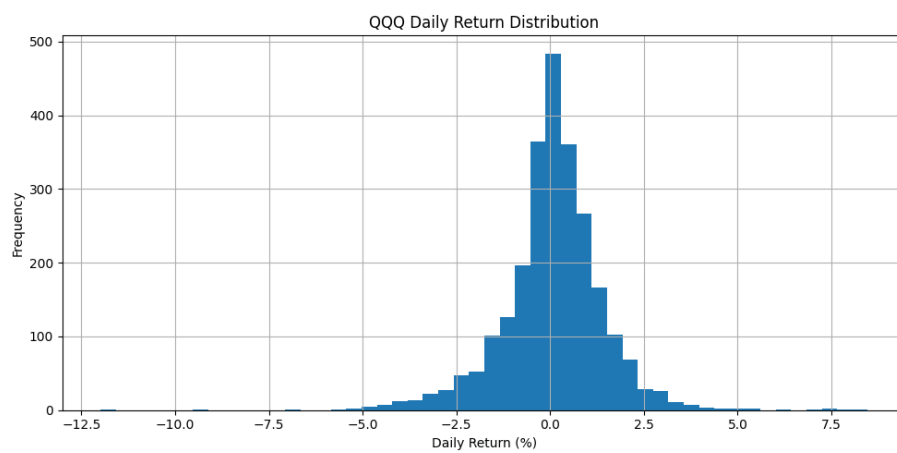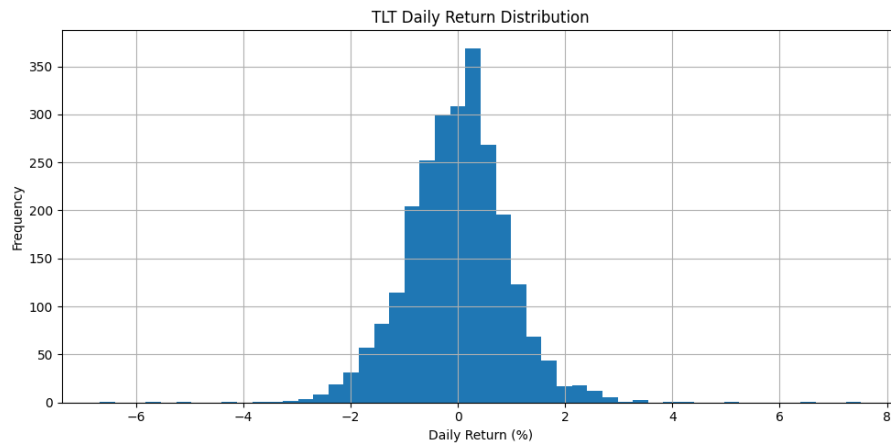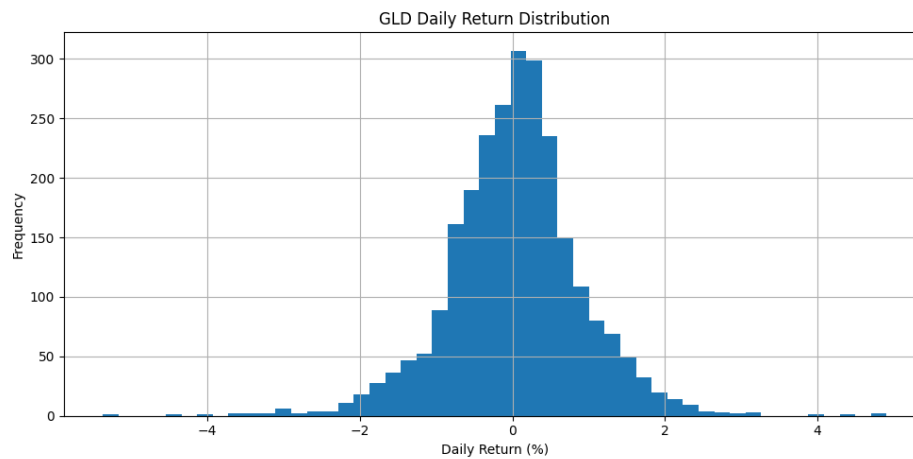
AAPL

MSFT

MSFT Daily Return Distribution

SPY

SPY Daily Return Distribution

QQQ

QQQ Daily Return Distribution

TLT

GLD

Outlier analysis

Ticker	outlier_count	total_count	outlier_ratio
AAPL	139	2515	0.0553
MSFT	132	2516	0.0525
SPY	168	2520	0.0667
QQQ	169	2516	0.0672
TLT	60	2515	0.0239
GLD	97	2541	0.0382

Outliers are detected using the standard 1.5×IQR rule. The outlier ratios for equity assets (SPY, QQQ, AAPL, MSFT) are between 5–7% of all observations, while TLT and GLD have much lower ratios (around 2–4%). These extreme points correspond to crisis and stress periods rather than data errors. Because such events are crucial for risk management, we decided to keep the outliers in the dataset instead of winsorising or deleting them.

Risk–Return Metrics (Annualised)

Ticker	annual_return	annual_volatility	Sharpe
AAPL	0.2626	0.2847	0.9222
MSFT	0.2572	0.2715	0.9471
SPY	0.1205	0.1768	0.6819
QQQ	0.1845	0.2183	0.8453
TLT	-0.0260	0.1532	-0.1699
GLD	0.0845	0.1405	0.6016

Annualised metrics confirm that AAPL and MSFT deliver the highest risk-adjusted performance, with Sharpe ratios around 0.9. QQQ and SPY also provide attractive risk–return trade-offs, while GLD offers lower but positive returns with relatively low volatility. TLT has a negative Sharpe ratio over the sample, meaning that its main contribution is diversification rather than standalone performance.
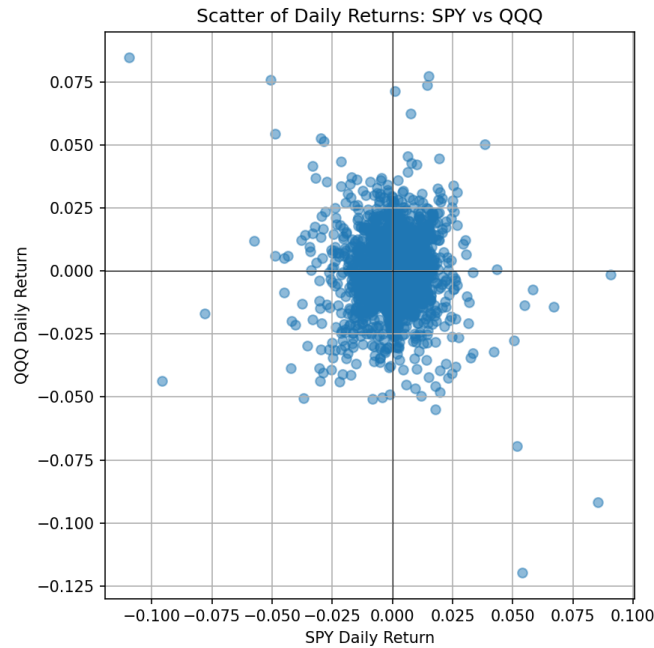
Corelation Heatmap

The return correlation matrix provides important insight into how different asset classes behave relative to each other, which is essential for constructing effective risk-managed portfolios. Overall, the heatmap reveals three distinct clusters: (1) highly correlated technology-driven equity assets (AAPL, MSFT, SPY, QQQ), (2) negatively correlated long-term government bonds (TLT), and (3) largely uncorrelated safe-haven assets such as gold (GLD).

1. High Equity–Equity Correlation

The strongest correlations appear among the four equity assets.

- AAPL–QQQ (0.81), MSFT–QQQ (0.86), and SPY–QQQ (0.93) reflect the close interconnectedness of large-cap growth stocks with the broader market and especially with the technology-heavy Nasdaq index.
- AAPL–MSFT (0.68) and AAPL–SPY (0.75) also show consistently strong positive relationships, indicating that large U.S. tech companies tend to move in the same direction in response to macroeconomic conditions, corporate earnings cycles, and market sentiment.

This cluster behaves cohesively during both expansions and downturns, meaning that although these assets offer high growth potential, they provide limited diversification when combined with one another.

Scatter of Daily Returns: SPY vs QQQ

*This scatter plot visually confirms the strong positive correla*tion (0.93) between the broad market (SPY) and the tech sector (QQQ). The linear clustering along the diagonal demonstrates that these two assets move almost identically, reinforcing the lack of diversification benefit between them.

2. Equity–Bond Negative Correlation

Long-term U.S. Treasury bonds (TLT) are negatively correlated with equity assets:

- SPY–TLT: –0.21
- AAPL–TLT: –0.13
- QQQ–TLT: –0.15

These negative correlations represent the classic "flight-to-safety" behavior: equity markets tend to fall during periods of economic fear, while long-maturity bonds typically rise as investors seek safe, interest-bearing assets. Although the magnitude is moderate rather than strong, the relationship is sufficiently negative to make TLT a valuable diversifier in a robo-advisor portfolio.

3. Gold as a Low-Correlation Diversifier

Gold (GLD) shows near-zero correlation with all major equity indices:

- GLD–AAPL: 0.03
- GLD–MSFT: 0.03
- GLD–SPY: 0.05
- GLD–QQQ: 0.06

and a modest positive correlation with long-term bonds:

- GLD–TLT: 0.31

This behavior is consistent with gold's role as a long-term hedge against macroeconomic uncertainty, inflation, and systemic risk. Because gold moves independently of equities, it significantly improves diversification efficiency even with a relatively small allocation.

Overall Interpretation for Portfolio Construction

The heatmap confirms the expected market dynamics:

- Equities move together → high growth, but high co-movement risk.
- Bonds move opposite of equities → strong stabilizer during downturns.
- Gold moves independently → hedge against inflation, volatility, and tail-risk events.

For a robo-advisor system, this structure is essential because it allows the model to:

- reduce portfolio volatility by mixing negatively correlated assets,
- maintain growth potential through highly correlated equity components, and
- improve long-term resilience with low-correlation hedges such as gold.

This correlation pattern creates a strong foundation for building risk-based portfolios that respond effectively to changes in market regimes.

Return Correlation Heatmap (2015–2024)