به نام خدا

دانشگاه تهران

دانشکده مهندسی برق و کامپیوتر

Introduction to Data Science

Final Project

Phase 1

نام و نام خانوادگی	محمد عسکری
شماره دانشجویی	۸۱۰۱۹۸۴۴۱
تاریخ ارسال گزارش	۱۴۰۴/۰۱/۱۹

Introduction

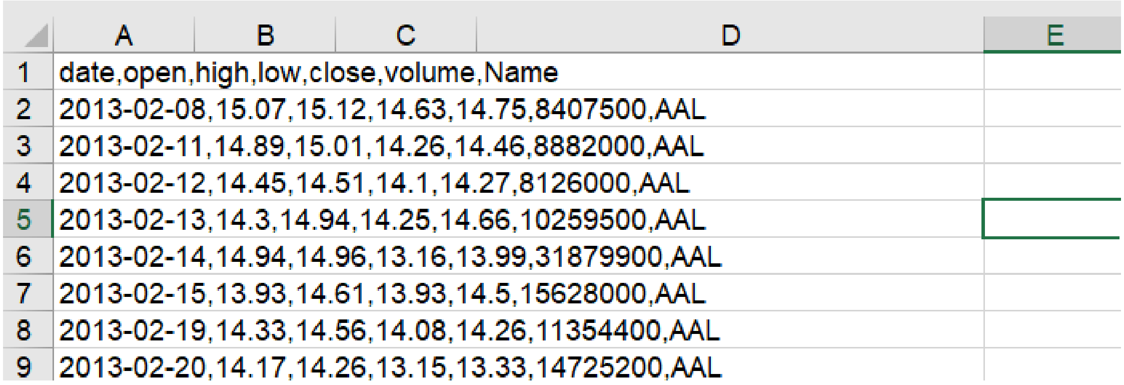
In financial markets, balancing risk and return is a fundamental challenge for investors. The **Markowitz Portfolio Optimization Model** provides a mathematical framework for constructing efficient investment portfolios based on asset returns and their covariances. This project applies Markowitz's model to historical data from the **S&P 500 index** to compare the **risk profiles of various 10-stock portfolios**, rather than advising on optimal investments.

Dataset Information

The dataset used in this project contains **historical stock price data for companies listed in the S&P 500 index**, spanning a period of **five years**. The data includes daily trading metrics for each company, providing a comprehensive basis for financial analysis and portfolio construction.

- **Title**: Historical Stock Data of S&P 500 Companies (2013–2018)
- **Source**: Kaggle – S&P 500 Stock Data
- **Data Type**: Time-series, tabular
- **Number of Companies**: 500+
- **Time Range**: Approximately 2013 to 2018 (5 years)
- **Frequency**: Daily trading data
- **Key Attributes**:
  - `date`: The date of the trading session
  - `open`: Opening price of the stock
  - `high`: Highest price of the day
  - `low`: Lowest price of the day
  - `close`: Closing price of the stock
  - `volume`: Total number of shares traded
  - `Name`: Ticker symbol representing the company

A sample of the dataset is shown below for reference:



	A	B	C	D	E
1	date,open,high,low,close,volume,Name				
2	2013-02-08,15.07,15.12,14.63,14.75,8407500,AAL				
3	2013-02-11,14.89,15.01,14.26,14.46,8882000,AAL				
4	2013-02-12,14.45,14.51,14.1,14.27,8126000,AAL				
5	2013-02-13,14.3,14.94,14.25,14.66,10259500,AAL				
6	2013-02-14,14.94,14.96,13.16,13.99,31879900,AAL				
7	2013-02-15,13.93,14.61,13.93,14.5,15628000,AAL				
8	2013-02-19,14.33,14.56,14.08,14.26,11354400,AAL				
9	2013-02-20,14.17,14.26,13.15,13.33,14725200,AAL				

. . .

The dataset was downloaded in CSV format and has been preprocessed using standard cleaning techniques including removal of missing values, date parsing, and normalization of numerical fields.

Project Objectives

1. **Data Exploration**
   - Clean and normalize the dataset
   - Analyze price trends, stock-wise behavior, and volatility
2. **Portfolio Construction**
   - Select multiple 10-stock portfolios
   - Calculate returns, covariance matrices, and volatility
   - Use the Markowitz Efficient Frontier for analysis
3. **Risk-Return Analysis**
   - Visualize risk vs. expected return
   - Compare risk metrics (e.g., standard deviation, Sharpe ratio) across portfolios

Expected Outcome

By the end of Phase 1:

- We will have prepared and explored the dataset
- Constructed and evaluated sample portfolios
- Defined the groundwork for advanced modeling in Phase 2
  In future phases, more refined optimization and interactive visualizations will be developed.

Objective 1 – Data Exploration

The first step of the project involved exploring and preparing the dataset for financial analysis. Using Python and libraries such as `pandas`, `numpy`, and `scikit-learn`, I performed a series of data cleaning and transformation steps to make the raw data suitable for portfolio modeling.

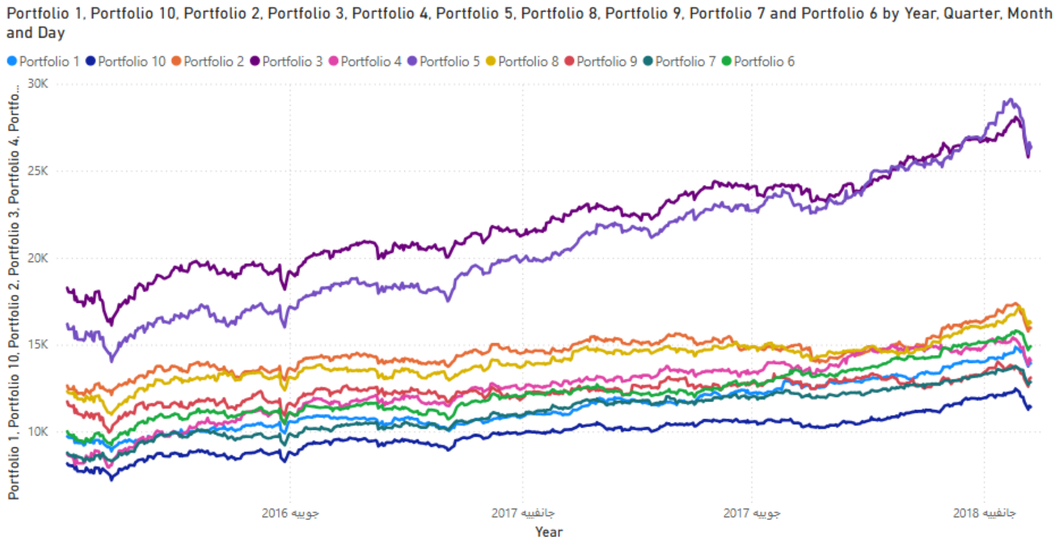
This included removing rows with missing values, converting date strings into datetime objects, and normalizing numerical columns (`open`, `high`, `low`, `close`, and `volume`) using **MinMaxScaler** to bring them into a comparable scale. These steps are crucial for eliminating biases caused by differing value ranges and ensuring the accuracy of correlation and covariance calculations later on.

All relevant code for this preprocessing pipeline is included in the Jupyter Notebook titled **`DataExploration.ipynb`**, submitted along with this report.

Objective 2 – Portfolio Construction

To evaluate the risk-return behavior of different investment strategies, I constructed **10 unique portfolios**, each consisting of **10 randomly selected stocks** from the S&P 500 dataset. For each portfolio, I calculated the **daily returns** based on the percentage change in closing prices and computed the **cumulative return** over time to represent overall profit.

All relevant implementation has been included in the Jupyter Notebook titled `PortfolioConstruction.ipynb`. The output of this process is a comparative plot showing how each portfolio's value evolves throughout the 5-year period. This visualization provides insight into the diversity of return profiles based on stock selection, laying the groundwork for deeper risk analysis in the following phases.



Objective 3 – Risk-Return Analysis

To evaluate the trade-off between risk and profitability, I performed a detailed analysis across the 10 randomly constructed portfolios. Each portfolio's **daily return** was calculated using the percentage change in its value over time, derived from the cumulative return data previously computed.

To quantify **risk**, I calculated both the **variance** and **standard deviation** of daily returns for each portfolio. The **final portfolio value** (based on an initial $10,000 investment) was used to represent return. These metrics were compiled into a KPI summary table and saved in the file `portfolio_avg_cov_vs_profit.csv`

Covariance and Portfolio Profitability

As shown in the visualization and KPI table, there appears to be a strong positive relationship between **average stock covariance** and **portfolio profitability**. Notably,

Portfolio 9, which had the **highest average covariance (0.087350)** among all portfolios, also achieved one of the **highest final values** at **$30,999.44**.

This trend suggests that portfolios composed of stocks that moved more closely together (i.e., higher covariance) tended to yield higher overall returns in this sample. While this behavior might seem counterintuitive to traditional diversification principles, it highlights how positively correlated assets can still drive strong portfolio performance — possibly due to exposure to strongly trending sectors or market conditions.

This report was written and developed with the assistance of ChatGPT, used as a supportive tool for analysis, explanation, and code generation.