Investment Task 1

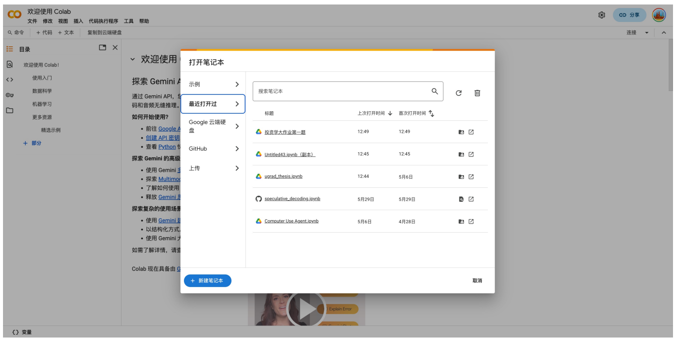
Siqi Zhu

May 2025

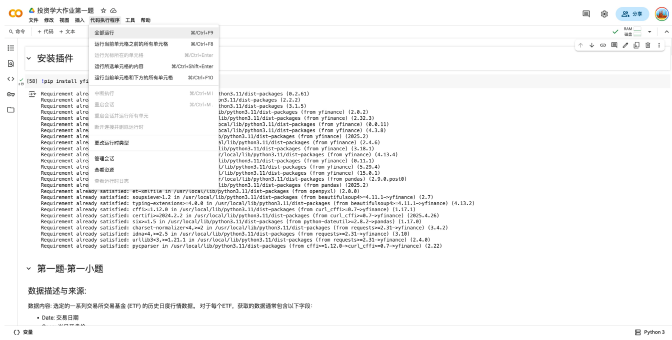
Contents

1 Introduction

I used Python to collect and analyze data for this task. All the results shown in this report can be reproduced by running the ipynb file from the link. You need to upload the file to Google Colab [see Image 1a], and then click "Runtime" / "Run all" in the upper left corner [see Image 1b].



(a) upload ipynb to colab



(b) run file

Figure 1: intro

2 Data Description and Source

2.1 Data Content

The core dataset comprises historical daily market data for a curated list of Exchange Traded Funds (ETFs). For each ETF, the acquired data includes the following fields:

Date: The trading date; **Open**: The opening price for the day; **High**: The highest price reached during the day; **Low**: The lowest price reached during the day; **Close**: The closing price for the day; **Volume**: The total number of shares traded during the day.

2.2 Data Source

All financial data was programmatically retrieved from Yahoo Finance using the `yfinance` Python library. Yahoo Finance is a widely recognized and extensively used provider of financial market data, news, analytical tools, and related information. The `yfinance` library provides a convenient and robust interface for accessing this publicly available data.

2.3 Industry Selection

The industries included in this study were chosen to represent a diverse range of sectors within the U.S. stock market. The selected sectors and their corresponding tickers are as follows:

XLV (Health Care); **DBA** (Agriculture); **XME** (Mining); **XLI** (Industrials); **XLU** (Utilities); **XLRE** (Real Estate); **XRT** (Retail); **IYT** (Transportation); **PEJ** (Consumer Discretionary); **XLF** (Financials)

2.4 Data Time Frame

The historical data for the selected ETFs covers the period from January 1, 2015, to December 31, 2024.

2.5 Calculated Statistical Indicators

Several statistical indicators [see Image 2] were computed from the raw data to facilitate analysis: average returns, standard deviation, and correlations.

ETF	行业	平均每日回报率	每日回报率标准差	年化平均回报率	年化标准差
XLV	医疗保健 (Health Care)	0.000396	0.010556	0.099892	0.167567
DBA	农林牧渔 (Agriculture)	0.000105	0.008223	0.026563	0.130535
XME	采矿业 (Mining)	0.000519	0.021056	0.130895	0.334252
XLI	制造业 (Industrials/Manufacturing)	0.000488	0.012376	0.122875	0.196462
XLU	水电煤气 (Utilities)	0.000390	0.012138	0.098295	0.192677
XLRE	房地产 (Real Estate)	0.000350	0.013158	0.088256	0.208879
XRT	批发和零售业 (Retail)	0.000406	0.016768	0.102314	0.266183
IYT	交通运输 (Transportation)	0.000347	0.014119	0.087458	0.224139
PEJ	住宿和餐饮业 (Consumer Discretionary/Travel)	0.000294	0.015204	0.074130	0.241353
XLF	金融业 (Financials)	0.000523	0.013982	0.131822	0.221956

(a) avg. returns and standard deviations

```
2. 每日回报率相关性矩阵:
        XLV   DBA   XME   XLI   XLU  XLRE   XRT   IYT   PEJ   XLF
XLV   1.000 0.128 0.461 0.718 0.563 0.639 0.519 0.638 0.551 0.686
DBA   0.128 1.000 0.260 0.208 0.087 0.146 0.166 0.188 0.210 0.196
XME   0.461 0.260 1.000 0.687 0.327 0.437 0.571 0.657 0.595 0.638
XLI   0.718 0.208 0.687 1.000 0.543 0.675 0.692 0.905 0.783 0.876
XLU   0.563 0.087 0.327 0.543 1.000 0.738 0.341 0.444 0.386 0.481
XLRE  0.639 0.146 0.437 0.675 0.738 1.000 0.535 0.608 0.588 0.625
XRT   0.519 0.166 0.571 0.692 0.341 0.535 1.000 0.705 0.719 0.650
IYT   0.638 0.188 0.657 0.905 0.444 0.608 0.705 1.000 0.778 0.816
PEJ   0.551 0.210 0.595 0.783 0.386 0.588 0.719 0.778 1.000 0.757
XLF   0.686 0.196 0.638 0.876 0.481 0.625 0.650 0.816 0.757 1.000
```

(b) correlations

Figure 2: statistics

3 Collect risk-free rate on daily basis

While no asset is perfectly risk-free (e.g., unexpected inflation can erode real returns), the 13-week U.S. Treasury bill yield is a widely accepted and conventional proxy in financial analysis and modeling for the short-term, nominal risk-free rate in U.S. dollars due to its minimal default risk, short tenor, high liquidity, and appropriate currency denomination. It is worth noting that historical and current data for this Treasury bill yield (^IRX) can be programmatically accessed using common financial data APIs, such as the `yfinance` library in Python.

```
测试无风险利率数据获取成功。
平均每日无风险利率: Ticker              ^IRX
Date
2015-01-02  5.952381e-07
2015-01-05  1.190476e-07
2015-01-06  7.936508e-07
2015-01-07  7.936508e-07
2015-01-08  7.142857e-07
...                  ...
2024-12-23  1.672619e-04
2024-12-24  1.666667e-04
2024-12-26  1.672619e-04
2024-12-27  1.657936e-04
2024-12-30  1.659524e-04

[2514 rows x 1 columns]
```

(a) table

```
已成功获取每日无风险利率数据。
数据点数量: 2514
平均每日无风险利率: 0.000070 (即 0.0070%)
平均年化无风险利率 (估算): 0.017559 (即 1.7559%)
```

(b) mean

Figure 3: risk-free rate

4 Assumptions

We assume there are 252 trading days in a year. The investor's utility function is specified as

$$U = \mathbb{E}(R) - \frac{1}{2}A\sigma^2,$$

where $\mathbb{E}(R)$ denotes the expected return, σ^2 represents the return variance, and A is the coefficient of risk aversion. In this study, we set $A = 5.0$. The expected return $\mathbb{E}(R)$ is estimated using the historical average return. Additionally, short selling is allowed throughout the analysis.

5 "Three-Step" approach

We follow the classical "Three-Step" approach in portfolio theory to derive the structure of the optimal risky portfolio and the optimal complete portfolio. The three steps are:

1. **Determine the Efficient Frontier of Risky Assets**: Identify the set of portfolios that provide the highest expected return for a given level of risk or the lowest risk for a given expected return.
2. **Find the Optimal Risky Portfolio**: On the efficient frontier, identify the portfolio with the highest Sharpe Ratio. This is known as the *Tangency Portfolio*.
3. **Determine the Optimal Complete Portfolio**: Combine the tangency portfolio with a risk-free asset based on the investor's risk aversion to construct the final portfolio.

5.1 Step 1: Efficient Frontier of Risky Assets

Suppose there are N risky assets. We estimate the following:

- Expected return vector: $E(\mathbf{R}) = [E(R_1), E(R_2), \ldots, E(R_N)]^T$
- Covariance matrix: $\mathbf{\Sigma}$, an $N \times N$ matrix where σ_{ij} denotes the covariance between assets i and j.

Let $\mathbf{w} = [w_1, w_2, \ldots, w_N]^T$ be the portfolio weights, where $\sum_{i=1}^{N} w_i = 1$.
Then the portfolio's expected return and variance are:

$$E(R_P) = \mathbf{w}^T E(\mathbf{R}), \quad \sigma_P^2 = \mathbf{w}^T \mathbf{\Sigma} \mathbf{w}$$

To find the portfolio with minimum variance for a given expected return E_0, we solve:

$$\min_{\mathbf{w}} \quad \frac{1}{2}\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}$$

subject to:

$$\mathbf{w}^T E(\mathbf{R}) = E_0, \quad \mathbf{w}^T \mathbf{1} = 1$$

Solving the system using the constraints yields the Minimum Variance Frontier.

The upper part of this frontier, beginning from the Global Minimum Variance Portfolio (GMVP), is called the *Efficient Frontier*.

The GMVP solution is:

$$\mathbf{w}_{GMVP} = \frac{\mathbf{\Sigma}^{-1}\mathbf{1}}{\mathbf{1}^T \mathbf{\Sigma}^{-1}\mathbf{1}}$$

I use Python to solve the equation, and visualize a few data points [see Image 4]

有效边界上的投资组合点（部分示例）：
（显示前10和后10个点）

Volatility (σ)	Return E(R)	Sharpe Ratio	W_XLV	W_DBA	W_XME	W_XLI	W_XLU	W_XLRE	W_XRT	W_IYT	W_PEJ	W_XLF
10.6674%	7.0175%	0.4932	24.5667%	59.3749%	0.0000%	0.0000%	16.0584%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
10.6903%	6.7292%	0.4652	22.0215%	63.8776%	0.0000%	0.0000%	14.1009%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
10.7100%	7.3058%	0.5182	28.4875%	55.0750%	0.0000%	0.7204%	15.7172%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
10.7815%	6.4409%	0.4345	20.1036%	68.3422%	0.0000%	0.0000%	11.5543%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
10.8089%	7.5941%	0.5401	26.3551%	52.1978%	0.0000%	4.0944%	17.3527%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%
10.9396%	6.1526%	0.4019	17.2798%	72.2260%	0.0000%	0.0000%	8.8096%	1.3397%	0.0000%	0.0000%	0.3449%	0.0000%
10.9447%	7.8823%	0.5598	27.9267%	48.9411%	0.0000%	6.3533%	16.2413%	0.0000%	0.0000%	0.0000%	0.0000%	0.5376%
11.1193%	8.1706%	0.5769	27.0006%	46.0878%	0.0000%	8.2127%	16.8602%	0.0000%	0.0000%	0.0000%	0.0000%	1.8387%
11.1523%	5.8643%	0.3684	13.7554%	75.7708%	0.0000%	0.0000%	6.0582%	3.5776%	0.0000%	0.0000%	0.8380%	0.0000%
11.3314%	8.4589%	0.5915	25.0731%	43.3022%	0.1878%	8.5190%	19.0114%	0.0000%	0.0000%	0.0000%	0.0000%	3.9064%
23.5813%	15.9544%	0.6021	0.0000%	0.0000%	42.8985%	2.4956%	9.1451%	0.0000%	0.0000%	0.0000%	0.0000%	45.4608%
24.4189%	16.2427%	0.5933	0.0000%	0.0000%	46.5293%	0.0000%	5.7111%	0.0000%	0.0000%	0.0000%	0.0000%	47.7596%
25.2889%	16.5310%	0.5843	0.0000%	0.0000%	49.9173%	0.0000%	1.3596%	0.0000%	0.0000%	0.0000%	0.0000%	48.7231%
26.2078%	16.8193%	0.5748	0.0000%	0.0000%	56.0161%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	43.9839%
27.2236%	17.1075%	0.5639	0.0000%	0.0000%	63.3468%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	36.6532%
28.3295%	17.3958%	0.5521	0.0000%	0.0000%	70.6774%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	29.3226%
29.5153%	17.6841%	0.5397	0.0000%	0.0000%	78.0081%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	21.9919%
30.7717%	17.9724%	0.5270	0.0000%	0.0000%	85.3387%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	14.6613%
32.0906%	18.2607%	0.5143	0.0000%	0.0000%	92.6694%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	7.3306%
33.4644%	18.5490%	0.5018	0.0000%	0.0000%	100.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%	0.0000%

Figure 4: Step1: efficient frontier

5.2 Step 2: Optimal Risky Portfolio

Introduce a risk-free asset with return R_f and zero standard deviation. Combining this with any risky portfolio P yields:

$$E(R_C) = (1 - w_P)R_f + w_P E(R_P), \quad \sigma_C = w_P \sigma_P$$

The line connecting R_f to any risky portfolio P is called the *Capital Allocation Line (CAL)*, with slope equal to the *Sharpe Ratio*:

$$S_P = \frac{E(R_P) - R_f}{\sigma_P}$$

To find the tangency portfolio P^* with maximum Sharpe Ratio:

$$\max_{\mathbf{w}} \quad \frac{\mathbf{w}^T E(\mathbf{R}) - R_f}{\sqrt{\mathbf{w}^T \mathbf{\Sigma} \mathbf{w}}} \quad \text{subject to } \mathbf{w}^T \mathbf{1} = 1$$

The solution is:

$$\mathbf{w}^* = \frac{\mathbf{\Sigma}^{-1}(E(\mathbf{R}) - R_f \mathbf{1})}{\mathbf{1}^T \mathbf{\Sigma}^{-1}(E(\mathbf{R}) - R_f \mathbf{1})}$$

From this we can compute:

$$E(R_{P^*}) = (\mathbf{w}^*)^T E(\mathbf{R}), \quad \sigma^2_{P^*} = (\mathbf{w}^*)^T \mathbf{\Sigma} \mathbf{w}^*$$

The line through $(0, R_f)$ tangent to the efficient frontier at P^* is the *Capital Market Line (CML)*:

$$E(R_C) = R_f + \frac{E(R_{P^*}) - R_f}{\sigma_{P^*}} \sigma_C$$

```
最优风险资产组合 (ORP) 特征:
  预期年化收益率 E(R_orp): 12.0050%
  年化波动率 (标准差) σ_orp: 15.6075%
  夏普比率 Sharpe_orp: 0.6567
  资产权重:
    XLV: 21.1790%
    DBA: 12.2091%
    XME: 12.2300%
    XLI: 17.9513%
    XLU: 23.6106%
    XLRE: 0.0000%
    XRT: 0.0000%
    IYT: 0.0000%
    PEJ: 0.0000%
    XLF: 12.8200%
```

Figure 5: Step2: Optimal Risky Portfolio

5.3 Step 3: Optimal Complete Portfolio (O)

With P^* known, investors allocate between it and the risk-free asset based on their risk aversion A. Assume utility:

$$U = E(R_C) - \frac{1}{2} A \sigma_C^2$$

Let y be the fraction invested in P^*:

$$E(R_C) = (1 - y)R_f + yE(R_{P^*}), \quad \sigma_C = y\sigma_{P^*}$$

Then utility becomes:

$$U(y) = R_f + y(E(R_{P^*}) - R_f) - \frac{1}{2} A y^2 \sigma^2_{P^*}$$

Maximizing $U(y)$:

$$\frac{dU}{dy} = E(R_{P^*}) - R_f - Ay\sigma^2_{P^*} = 0 \Rightarrow y^* = \frac{E(R_{P^*}) - R_f}{A\sigma^2_{P^*}}$$

Then:

$$E(R_O) = (1 - y^*)R_f + y^* E(R_{P^*}), \quad \sigma_O = y^* \sigma_{P^*}$$

The optimal complete portfolio O lies on the CML and corresponds to the point where the investor's indifference curve is tangent to the CML.

```
假设的投资者风险厌恶系数 (A): 5.0

投资于最优风险资产组合 (ORP) 的资金比例 (y*): 84.1491%
投资于无风险资产的资金比例 (1 - y*): 15.8509%

最优完整组合特征:
  预期年化收益率 E(R_complete): 10.3805%
  年化波动率 (标准差) σ_complete: 13.1336%

最终各ETF在总投资组合中的配置比例 (占总资产的百分比):
    XLV: 17.8220%
    DBA: 10.2738%
    XME: 10.2914%
    XLI: 15.1058%
    XLU: 19.8681%
    XLRE: 0.0000%
    XRT: 0.0000%
    IYT: 0.0000%
    PEJ: 0.0000%
    XLF: 10.7879%
```

Figure 6: Step3: Optimal Portfolio