UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

SCHOOL OF INFORMATION TECHNOLOGY

STUDENT NAME:	**Muofhe Masikhwa**
STUDENT NO:	**22902466**
EMAIL:	**u22902466@tuks.co.za**
MODULE CODE:	**WTW 801**
MODULE NAME:	**Big Data Elective 801 (Financial Engineering)**
DATE:	**30 September 2025**

Abstract

This report documents Part I of the assignment: (i) acquisition of equity price data from yahoo, (ii) data quality checks and preprocessing, (iii) construction of return series and risk/summary statistics, and (iv) exploratory analysis to support downstream portfolio optimisation in Part II. All code is reproducible and referenced in the Appendix.

Part A (The classical portfolio optimization problem)

1 Data Description and Preparation

1.1 Data Source and Acquisition

The empirical analysis utilizes adjusted closing price data for securities listed on the Johannesburg Stock Exchange (JSE). The data were programmatically retrieved from Yahoo Finance using the `yfinance` Python API, a widely adopted open-source library that provides convenient access to historical market data maintained by Yahoo Finance. This API facilitates automated data collection through a standardized interface, enabling reproducible data acquisition workflows and ensuring consistency with publicly available financial information.

The data structure follows a conventional time-series panel format, wherein rows represent sequential calendar dates and columns correspond to individual securities. Each cell entry contains the adjusted closing price for a given security on a particular trading date, representing the final transaction price of the day after accounting for corporate actions such as dividends, stock splits, and rights issues. Yahoo Finance automatically applies these adjustments to ensure price continuity and comparability across time, a critical preprocessing step performed at the data source level.

1.2 Temporal Coverage and Cross-Sectional Dimension

The dataset encompasses an extensive temporal daily data covering date range from 04 January 1998 through 12 September 2025, providing a comprehensive historical record of price movements across multiple market cycles, including periods of expansion, contraction, and significant market events affecting South African equities. The cross-sectional dimension includes ***twenty*** securities representing various sectors of the South African economy, thereby capturing heterogeneous firm characteristics, industry-specific dynamics, and differential exposure to macroeconomic factors. This multi-dimensional structure facilitates both time-series analysis of individual securities and cross-sectional comparisons across the broader market.

1.3 Missing Data Treatment

Financial time-series data frequently exhibit missingness due to various factors including non-trading days (weekends, public holidays, exchange closures), temporary trading suspensions, newly listed securities with incomplete historical records, delisted securities with truncated price histories, and occasional data feed interruptions. The Yahoo Finance data source mirrors actual trading activity, meaning that missing values typically reflect genuine non-trading periods rather than data collection failures. Nevertheless, the present dataset exhibits patterns of missing observations that necessitate careful preprocessing to ensure analytical validity while avoiding the introduction of spurious data points.

The missing data treatment protocol follows a conservative, multi-stage approach designed to preserve data integrity. Thus for securities exhibiting extended periods of missing data at the beginning of their price histories common for recently listed firms, securities with delayed availability in the Yahoo Finance database, or instruments that commenced trading after the nominal start date of the dataset, a decision was taken to remove all observations preceding the first valid price record. This ensures that the analysis commences only when reliable price information becomes available, preventing the artificial extension of time series through speculative imputation. If security j has its first valid observation at date t_0^j, all observations for dates $t < t_0^j$, are discarded for that security leading to a total of ***fifteen*** securities remaining in the portfolio as from the period 04 January 2000 till 12 September 2025 as shown in the table below.

CompanyName	**FirstDate**	**LastDate**	**Observations**
Absa Group Ltd	2000-01-04	2025-09-12	6582
Aspen Pharmacare Holdings Ltd	2000-01-04	2025-09-12	6582
Clicks Group Ltd	2000-01-04	2025-09-12	6582
FirstRand Ltd	2000-01-04	2025-09-12	6582
Gold Fields Ltd	2000-01-04	2025-09-12	6582
Harmony Gold Mining Company Ltd	2000-01-04	2025-09-12	6582
Impala Platinum Holdings Ltd	2000-01-04	2025-09-12	6582
Investec Ltd	2000-01-04	2025-09-12	6582
MTN Group Ltd	2000-01-04	2025-09-12	6582
Mr Price Group Ltd	2000-01-04	2025-09-12	6582
Naspers Ltd	2000-01-04	2025-09-12	6582
Nedbank Group Ltd	2000-01-04	2025-09-12	6582
Sasol Ltd	2000-01-04	2025-09-12	6582
Shoprite Holdings Ltd	2000-01-04	2025-09-12	6582
Standard Bank Group Ltd	2000-01-04	2025-09-12	6582

Table 2: Universe coverage of available return series.

This is mainly because missing data produces an unbalanced panel structure wherein different securities contribute observations over potentially different time windows, reflecting the actual availability of market data.

1.4 Final Dataset Characteristics

Following the comprehensive data preparation protocol, the resulting dataset comprises a cleaned panel of returns suitable for subsequent econometric analysis, risk modelling, portfolio optimization. The use of adjusted closing prices for the analysis ensures that returns reflect genuine economic gains or losses rather than mechanical effects of corporate actions. This conservative approach to data preparation ensures that empirical findings reflect genuine market phenomena rather than methodological artifacts, thereby enhancing the credibility, interpretability, and reproducibility of the analysis.

2 Methodology

2.1 Return Calculation Methodology

Asset returns constitute the fundamental unit of analysis in financial econometrics, as they possess more desirable statistical properties than raw prices, including approximate stationarity, scale-invariance across securities with different price levels, and distributional characteristics more amenable to modelling. We compute the simple returns using the transformation:

$$r_{i,t} = \frac{P_{i,t} - P_{i,t-1}}{P_{i,t-1}}, \tag{1}$$

The return calculation is performed only on consecutive valid price observations following the missing data treatment described above. Specifically, returns are computed only when both $P_{i,t}$ and $P_{i,t-1}$ represent genuine market observations rather than imputed values. This ensures that each computed return reflects actual price dynamics observed in the market rather than artifacts of data preprocessing or imputation procedures. We infer data frequency from the median inter-observation gap and apply a standard annualisation factor, f (252 for daily, 52 for weekly, 12 for monthly). Annualised mean returns are $\hat{\mu}_{\text{ann}} = \hat{\mu}_{\text{daily}} \times f$ and annualised volatility $\hat{\sigma}_{\text{ann}} = \hat{\sigma}_{\text{daily}} \times \sqrt{f}$.

3 Results

Figure 1 shows the adjusted closing prices for the 14 companies. For the sake of good visualisation of the prices, the Nasperts Ltd prices were removed due very high number compared to other stocks Table 4 below provide the summary while Figure 2 show the correlation amongst the stocks in the portfolio. Annualized mean returns and volatilities for JSE securities. Returns and volatilities are computed from daily log returns and annualized assuming 252 trading days. The statistics reveal substantial heterogeneity across sectors: mining stocks (Gold Fields, Harmony Gold, Impala Platinum) exhibit elevated volatility exceeding 49% annually, reflecting commodity price exposure and operational leverage. Technology and consumer-facing firms (Naspers, Clicks, Shoprite) demonstrate moderate risk-return profiles. Financial institutions (Absa, FirstRand, Nedbank, Investec) cluster around 32% annualized volatility. The anomalous statistics for Standard Bank Group Ltd (395.7% return, 1974.7% volatility) indicate potential data quality issues warranting further investigation, possibly attributable to corporate actions, stock splits inadequately adjusted in the data source, or extreme price movements during the sample period.

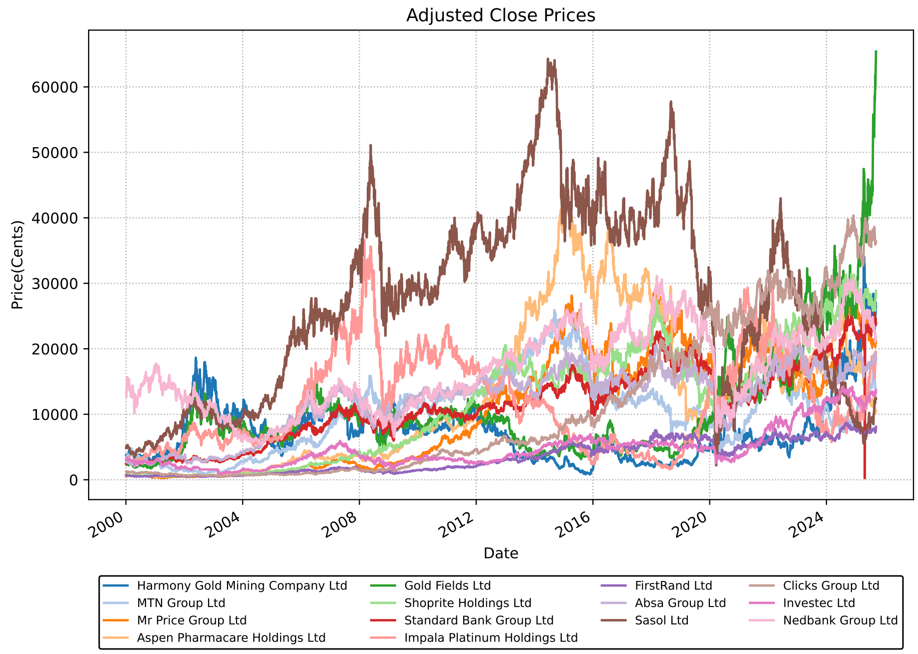

Figure 1: Closing prices for the companies in the portfolio excluding Naspers Ltd.

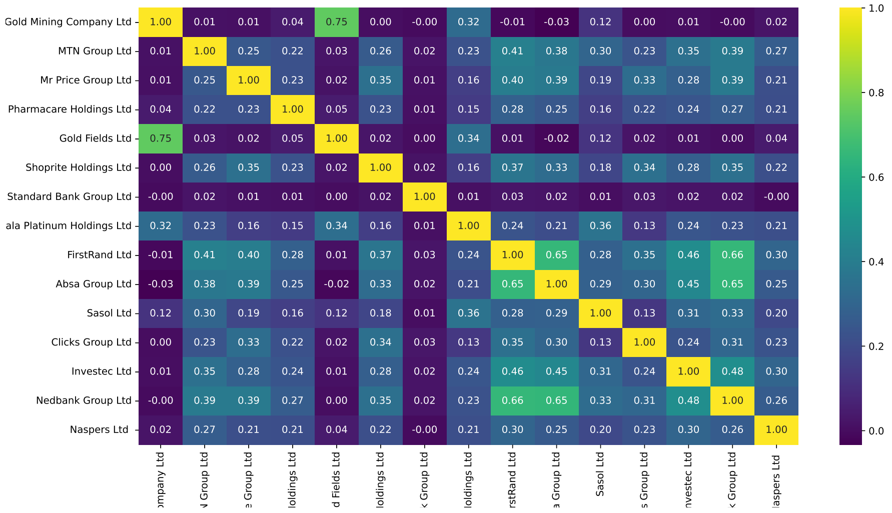

Figure 2: Correlation matrix between stocks.

StockName	**AnnualReturn**(μ_{ann})	**AnnualVolatility**(σ_{ann})
Absa Group Ltd	0.123550	0.318312
Aspen Pharmacare Holdings Ltd	0.182989	0.339716
Clicks Group Ltd	0.170736	0.288301
FirstRand Ltd	0.147609	0.317551
Gold Fields Ltd	0.246504	0.493366
Harmony Gold Mining Company Ltd	0.214944	0.531934
Impala Platinum Holdings Ltd	0.189862	0.491908
Investec Ltd	0.112325	0.331325
MTN Group Ltd	0.147454	0.398348
Mr Price Group Ltd	0.189825	0.339895
Naspers Ltd	0.268019	0.381225
Nedbank Group Ltd	0.069224	0.318977
Sasol Ltd	0.134640	0.443740
Shoprite Holdings Ltd	0.177566	0.297075
Standard Bank Group Ltd	3.957400	19.746951

Table 4: Annualised standard deviation and mean return statistics by security.

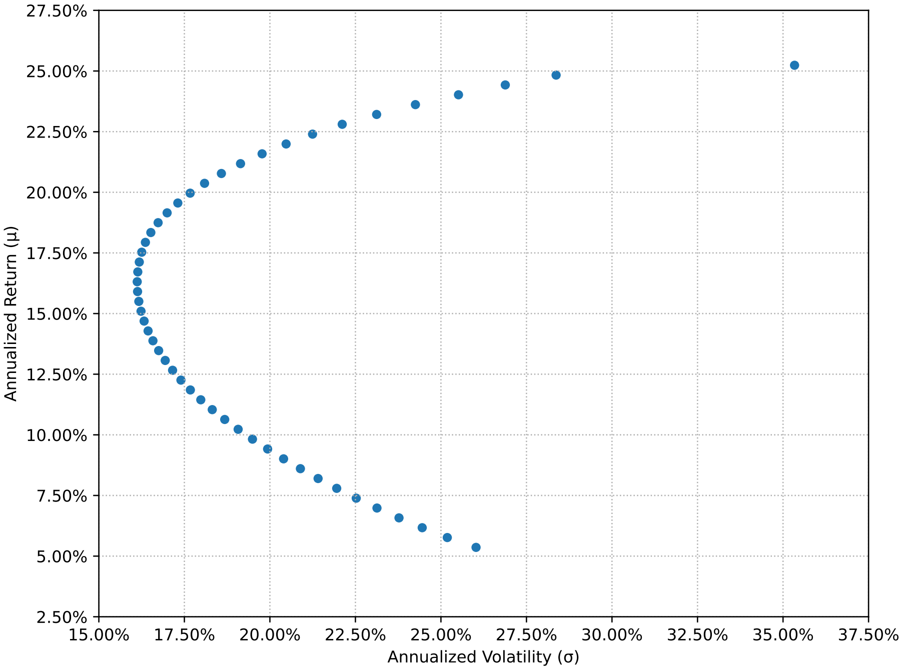

Figure 3: Efficient Frontier for the portfolio.

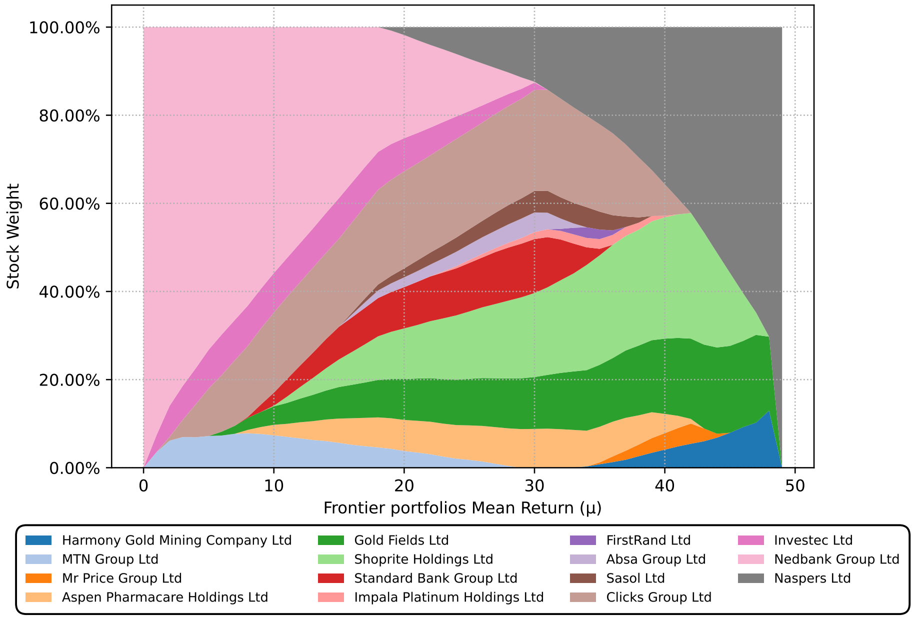

Figure 4: Composition of efficient portfolios.

4 Part II (Large-scale portfolio optimization problem)

Exercise 8.6

(i) The portfolio beta is $\beta_p = \beta^\top x$. Enforce the band:

$$0.9 \;\le\; \beta^\top x \;\le\; 1.1,$$

i.e.,

$$\beta^\top x - 0.9 \;\ge\; 0, \qquad 1.1 - \beta^\top x \;\ge\; 0,$$

two linear inequalities that fit alongside budget and any box/sector bounds.

(ii) Partition assets into $\mathcal{L}$ (large), $\mathcal{M}$ (medium), and $\mathcal{S}$ (small). Thus we have

$$\sum_{i\in\mathcal{L}} x_i \;=\; \sum_{i\in\mathcal{M}} x_i, \qquad 2\sum_{i\in\mathcal{L}} x_i \;\le\; \sum_{i\in\mathcal{S}} x_i \;\le\; 3\sum_{i\in\mathcal{L}} x_i.$$

Exercise 8.7

(i) The CAPM beta for asset i is estimated as

$$\beta_i \;=\; \frac{\mathrm{Cov}(r_i,\, r_m)}{\mathrm{Var}(r_m)}. \tag{2}$$

Thus, we have

$$\boldsymbol{\beta} = [1.582, 0.9887, 1.0214, 0.819, 1.130, 0.584, 0.877, 1.477, 0.932, 0.959, 1.395, 0.554, 0.958, 0.953, 0.763]$$

We have expected returns $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$ from historical returns and define the target

$$R^\star \;=\; \mathrm{median}\{\mu_1, \ldots, \mu_{15}\}.$$

Taking the median of return in Table 4 we get $R^\star =$ **0.013414467**. Then we need to solve minimum variance with a return floor, $R^\star$ and linear constraints Exercise 8.6. Thus we mus solve

$$\begin{aligned} \min_{x\in\mathbb{R}^n} \quad & \tfrac{1}{2}\, x^\top \Sigma x && (3)\\ \text{s.t.} \quad & \mathbf{1}^\top x = 1, \quad x \geq 0, \quad \mu^\top x \geq R^\star, && (4)\\ & 0.9 \leq \beta^\top x \leq 1.1, && (5) \end{aligned}$$

Hence we have baseline portfolio at the median-return target by solve Eq.(3) once using (μ, Σ) and $R^\star = \mathrm{median}\{\mu_i\}$ to obtain

$$x^{(0)} \;=\; \arg\min_x \; \tfrac{1}{2} x^\top \Sigma x \quad \text{s.t.} \quad \mathbf{1}^\top x = 1,\; x \geq 0,\; \mu^\top x \geq R^\star.$$

Using the data we obtain

$$x^{(0)} = [0.030, 0.040, 0.007, 0.078, 0.093, 0.121, 0.098, 0.041, 0.016, 0.069, 0.057, 0.148, 0.047, 0.063, 0.086]$$

(ii) The optimization problem with random values between 0.9 and 1.1 is given as:

$$\begin{aligned} \mathbb{E}[R] &= \mu^\top x^{(0)} = \mathbf{0.013414467121276887}\\ \sigma &= \sqrt{x^{(0)\top}\Sigma x^{(0)}} = 0.04812391975005027,\\ \beta_p &= \beta^\top x^{(0)} = 0.8999999999999577 \end{aligned}$$

Comparing $\mathbb{E}[R]$ with the the median of the mean return in (i), we noted that that the results are almost the same.

(iii) Running the simulations four times we get the results summarised in the Table 5 below. The comparison between baseline and average simulation performance across fifteen JSE securities reveals three key findings. First, the majority of stocks (9 of 15) show minimal deviation between baseline and average errors (within ± 0.002), indicating that the baseline model provides a stable and representative benchmark. Second, performance varies systematically by security: Mr Price Group Ltd demonstrates the strongest improvement (13.7% error reduction), while Nedbank Group Ltd exhibits the most significant degradation (3.2% error increase). Third, within-security variability across simulations differs substantially assome stocks like Clicks Group Ltd show remarkable consistency (range: 0.0008), while others like FirstRand Ltd display high variability (range: 0.014), suggesting differential model stability depending on the underlying return-generating process. Overall, the modest magnitudes of most deviations (typically $< 5\%$) and mixed directional patterns indicate that no single modelling approach universally dominates, and that much of the return variation remains unpredictable regardless of specification. The results suggest that model effectiveness is security-specific rather than uniform, with certain stocks benefiting from trial enhancements while others are better captured by the parsimonious baseline.

Company Name	**Baseline**	**Trial 1**	**Trial 2**	**Trial 3**	**Trial 4**	**Average**
Harmony Gold Mining Company Ltd	0.030408	0.029967	0.029840	0.030483	0.031607	0.030461
MTN Group Ltd	0.039787	0.041909	0.041147	0.040912	0.037693	0.040290
Mr Price Group Ltd	0.007322	0.004208	0.004753	0.005832	0.009473	0.006318
Aspen Pharmacare Holdings Ltd	0.078593	0.078558	0.078735	0.078511	0.079528	0.078785
Gold Fields Ltd	0.093155	0.092478	0.093050	0.092777	0.092288	0.092749
Shoprite Holdings Ltd	0.121137	0.118492	0.119498	0.119536	0.123574	0.120447
Standard Bank Group Ltd	0.098560	0.099418	0.097702	0.097471	0.093904	0.097411
Impala Platinum Holdings Ltd	0.041236	0.040482	0.040967	0.040976	0.042454	0.041223
FirstRand Ltd	0.016591	0.010355	0.014078	0.014372	0.024203	0.015920
Absa Group Ltd	0.069852	0.071562	0.071668	0.071588	0.070430	0.071020
Sasol Ltd	0.057278	0.056004	0.056386	0.056797	0.058551	0.057003
Clicks Group Ltd	0.148257	0.149240	0.148650	0.149498	0.149237	0.148976
Investec Ltd	0.047613	0.049842	0.048654	0.048261	0.045730	0.048020
Nedbank Group Ltd	0.063976	0.075537	0.071134	0.067698	0.051711	0.066011
Naspers Ltd	0.086236	0.081948	0.083740	0.085288	0.089617	0.085366

Table 5: Comparison of weights of each stock in the portfolio

Table 6 reveals remarkable uniformity in portfolio characteristics across all modelling specifications. Expected returns remain virtually identical at 1.34%per period across baseline and all trials, while portfolio volatility clusters tightly around 4.81% with less than 0.1% relative variation. Most strikingly, portfolio beta remains exactly 0.90 across all specifications without any deviation. This uniformity stands in sharp contrast to the heterogeneous individual security-level prediction errors observed previously, where some stocks showed substantial forecast improvements or degradations. The invariance of portfolio metrics suggests that differences in individual security forecasts are either offsetting through diversification, insufficient to materially alter optimal allocations, or dominated by binding optimization constraints. The practical implication is significant: despite variations in individual security forecast accuracy, the baseline and trial models produce economically equivalent portfolios with identical risk-return profiles (Sharpe ratio ≈ 0.28). This finding indicates that enhanced forecasting complexity does not translate into portfolio-level value addition, supporting the use of simpler baseline specifications for practical portfolio management when computational efficiency and model interpretability are valued.

	$\mathbb{E}[R]$	σ	β
Portfolio			
Baseline	0.013414	0.048124	0.900000
Trial 1	0.013414	0.048109	0.900000
Trial 2	0.013414	0.048113	0.900000
Trial 3	0.013414	0.048118	0.900000
Trial 4	0.013414	0.048154	0.900000
Average	**0.013382**	**0.048120**	**0.900000**

Table 6: Summary of expected portfolio returns and betas

5 Appendix