Multivariate Analysis of Tunisian Stock Market Trends Using PCA and Regression Modeling

Oussama zmitri
IT/BA

BA340-Data ANALYSIS

Instructor:

Dr. Aloui Donia

Tunis Business School

December 7, 2025

Abstract

This study examines the financial behavior and market dynamics of companies listed on the Tunisian Stock Exchange using quantitative and statistical analysis methods. The project focuses on exploratory data analysis (EDA) to uncover key trends, volatility patterns, trading volume behavior, and price movement correlations across different sectors. Data preprocessing techniques—including cleaning, scaling, and feature engineering—were applied to improve interpretability and model performance.

A linear regression framework was then implemented to evaluate and predict how selected variables, such as trading volume, company performance, and daily returns, influence stock price fluctuations. The model's performance was assessed using standard evaluation metrics, providing insight into predictive accuracy and explanatory relevance.

The findings contribute to a deeper understanding of market structure, investor behavior, and financial performance tendencies within Tunisia's emerging financial environment. This project demonstrates the usefulness of statistical and machine-learning techniques in supporting investment decisions and financial research in developing markets.

1. Introduction

1.1 Background and Motivation

Financial markets play a critical role in supporting national economic development, investment flows, and business expansion. In emerging markets such as Tunisia, understanding market behaviour is particularly important due to evolving regulatory frameworks, fluctuating investor confidence, and sector-specific economic pressures. The Tunisian Stock Exchange provides valuable insights into corporate performance, market efficiency, and investment trends, yet analytical research on its behavior remains limited compared to larger, more mature markets.

With increasing digitalization and the availability of structured financial datasets, statistical and machine-learning methods now provide an opportunity to extract meaningful patterns from historical stock data. These analytical methods can support investors, financial analysts, corporations, and policymakers in making informed decisions, predicting risk, and understanding underlying market movements. The motivation behind this study emerges from the need to bridge academic methodology with real financial data from Tunisia, providing data-driven insights into market behavior and company performance.

1.2 Problem Statement

Despite the availability of financial time-series data, there is a lack of systematic analytical examination of Tunisian listed companies using modern statistical techniques. Market participants often rely on intuition or descriptive summaries instead of evidence-based analysis. Therefore, the central problem addressed in this project is:
How can multivariate analytical methods and predictive modeling be applied to stock market data to extract meaningful patterns and explain stock price fluctuations of Tunisian listed companies?

1.3 Research Objectives

The main objectives of this research are:

- To explore and describe behavioral patterns in the Tunisian stock market using exploratory data analysis (EDA).
- To identify correlations and structures within multivariate financial data.
- To analyze the relationship between variables such as price, trading volume, and market returns using regression modeling.
- To reduce the complexity of high-dimensional financial data to uncover latent market factors influencing stock performance.
- To generate insights that may support improved financial forecasting, risk analysis, and decision-making.

1.4 Research Questions

This study aims to answer the following research questions:

1. What statistical patterns and trends can be identified in stock prices, trading volumes, and returns across companies listed on the Tunisian Stock Exchange?

2. Which financial variables are most strongly associated with stock price changes?

3. Can linear regression provide a meaningful explanatory model for understanding price fluctuations?

4. Are there detectable structural relationships, clusters, or latent patterns among companies based on their financial performance?

5. How can the insights generated from this analysis support investors, researchers, and decision-makers in understanding market behavior?

2. Data Collection and Sampling

2.1 Data Source

The dataset used in this study originates from real-world financial records related to companies listed on the Tunisian Stock Exchange. The data was obtained through academic collaboration with peers specializing in finance, who previously collected and compiled the dataset as part of their Financial Markets coursework. The source includes verified market records such as stock prices, trading volumes, and dates of transactions.

Although the dataset is not publicly sourced from an official financial institution platform (such as BVMT or Bloomberg), it is based on authentic Tunisian market data used in formal educational settings. The dataset therefore provides a reliable foundation for analytical methods and allows for realistic interpretation of stock behaviour in the Tunisian financial environment.

2.2 Data Description

The dataset contains multiple variables relevant to stock market performance, including:

- **Company Name**

- **Date of Observation**

- **Closing Price (Cours)**

- **Trading Volume (Qte. Éch.)**

- Derived variables generated during preprocessing, such as:
    - Price Returns
    - Volatility
    - Scaled volume metrics
    - Aggregated monthly statistics

The dataset covers multiple companies over a defined period, enabling both cross-sectional and time-series analysis.

2.3 Sampling Method

The study employs a **non-probability convenience sampling method**, as the dataset was not generated through an automated full-market extraction but rather through accessible historical records. Despite this, the inclusion of multiple companies across several months ensures a diverse representation of financial behaviour.

Monthly aggregation is applied to reduce noise, improve interpretability, and make the data suitable for multivariate analysis techniques such as Principal Component Analysis (PCA) and regression modeling.

2.4 Data Validity and Limitations

While the dataset is reliable for academic research, the following considerations apply:

- The sample may not cover **all companies listed on the BVMT**, which may limit generalizability.
- External shocks (e.g., political events, economic crises) are not explicitly encoded but may influence observed patterns.
- Missing values and inconsistencies were handled through preprocessing techniques such as imputation and aggregation.

Despite these limitations, the dataset remains appropriate for statistical exploration, modeling, and extracting insights into stock behaviour within the Tunisian market.

2.3 Variables Description

The dataset contains both raw and computed variables that describe the trading activity and performance of Tunisian listed companies. The table below summarizes each variable, its type, description, and unit (where applicable).

Variable Name	Type	Description	Unit
Company	Categorical	Name of the publicly traded Tunisian company included in the dataset.	—
Date	Date	The trading day on which the stock data was recorded.	—
Cours (Closing Price)	Quantitative (Continuous)	The closing stock price at the end of the trading day.	Tunisian Dinar (TND)
Qte. Éch. (Exchange Volume)	Quantitative (Discrete)	Number of shares traded during the trading day.	Shares
Return	Quantitative (Continuous, computed)	Daily stock return calculated using percentage change from the previous closing price.	%
Volatility	Quantitative (Continuous, computed)	Rolling standard deviation of stock returns over a 7-day window, representing risk or price fluctuation.	%

VolumeScaled	Quantitative (Continuous, computed)	Standardized exchange volume using z-score scaling for comparability across companies.	Scaled value
Month	Categorical (Derived)	Extracted month from the date variable used for aggregation and time-based analysis.	—
Year	Categorical (Derived)	Extracted year from the date variable used for temporal comparisons.	—

2.4 Data Cleaning & Preparation

Before proceeding with the analysis, several preprocessing steps were applied to ensure data consistency, reliability, and suitability for statistical modelling. The preparation process followed best practices for financial time-series and multivariate analysis.

Step 1: Handling Missing Values\textbf{Step 1: Handling Missing Values}Step 1: Handling Missing Values

The dataset contained occasional missing trading records for specific dates and companies, which is common in financial datasets due to non-trading days or low liquidity. Missing values were treated as follows:

- Rows without essential numerical variables (e.g., *Cours*, *Qte. Éch.*) were removed.
- Derived metrics such as *Return* and *Volatility* automatically resulted in `NA` for the first observation or insufficient window size; these were retained because they are expected and do not distort the analysis.
- Monthly aggregation later minimized the impact of isolated missing values.

Step 2: Detecting and Treating Outliers\textbf{Step 2: Detecting and Treating Outliers}Step 2: Detecting and Treating Outliers

Outliers in stock prices and volume were detected using interquartile range (IQR) and visual tools such as boxplots and scatter plots. Because extreme values in financial markets may represent real market behaviour rather than errors, outliers were not removed. Instead, they were preserved to maintain analytical integrity and interpreted cautiously.

Step 3: Variable Transformation and Encoding\textbf{Step 3: Variable Transformation and Encoding}Step 3: Variable Transformation and Encoding

The dataset contained both numerical and categorical variables. To ensure compatibility with quantitative models:

- The variable **Company** (categorical) was kept as a grouping variable and later converted to dummy variables only when necessary (e.g., regression modelling).
- **Date** was decomposed into **Month** and **Year** to support time-series insight and aggregated analysis.

Step 4: Scaling and Normalisation\textbf{Step 4: Scaling and Normalisation}Step 4: Scaling and Normalisation

Given that stock prices and trading volume operate at different scales, standardization was performed using z-score normalization:

Z=X−μσZ = \frac{X - \mu}{\sigma}Z=σX−μ

This scaling was applied to:

- Trading volume (*Qte. Éch. → VolumeScaled*)
- PCA-input variables during multivariate analysis

Scaling ensured equal weight in PCA, clustering, and regression.

Step 5: Preparation for Modelling\textbf{Step 5: Preparation for Modelling}Step 5: Preparation for Modelling

To enable multivariate techniques such as PCA and regression:

- Time-series variables were aligned by company and sorted chronologically.

- Monthly aggregated features were generated for dimensionality reduction and pattern recognition.

- A clean subset without missing or non-numeric values was created specifically for PCA.

3 Methodology

The methodological framework of this study was designed to align with the nature of the dataset, the research objectives, and the analytical questions defined earlier. Since the goal of this project is to explore financial market behaviour, detect structural patterns, and evaluate relationships between variables over time, a combination of exploratory, statistical, and machine learning-based techniques was applied.

The analysis was carried out using the statistical programming language **R**, which provides robust tools for financial data manipulation, visualization, and modelling. The workflow followed an incremental structure, beginning with basic descriptive analysis and progressing toward more advanced multivariate and predictive techniques.

3.1 Selected Analytical Tools

Based on the dataset's structure and research objectives, the following analytical techniques were chosen:

---

a) Exploratory Data Analysis (EDA)

EDA was conducted to understand the distribution, behaviour, and interactions among variables. This included summary statistics, correlation matrices, time-series visualizations, and scatter plots. The purpose was to uncover initial patterns, trends, seasonality, and potential anomalies.

---

b) Principal Component Analysis (PCA)

PCA was used to reduce dimensionality and identify latent structures in the dataset. Since the financial indicators (price, return, volatility, and volume) vary in scale and may be correlated, PCA helped transform them into uncorrelated synthetic components. The **FactoMineR** package was used to compute:

- Eigenvalues and explained variance
- Component scores
- Variable contributions and correlations

This step enabled interpretation of dominant financial behaviour drivers across companies.

---

c) Linear Regression Modelling

A linear regression model was used to examine whether trading volume and volatility could predict price fluctuations. The model helped answer the question:

> *Do changes in volume and volatility significantly influence the stock price?*

Regression diagnostics such as residual plots, R², and significance tests were used to validate the model and assess prediction strength.

3.2 Procedure Overview

The overall methodology followed this structured process:

Step	Phase	Description
1	Data Import	Load dataset and verify structure
2	Preprocessing	Clean data, handle missing values, create new financial metrics
3	Aggregation	Compute monthly averages to reduce noise and improve interpretability

4	Feature Scaling	Standardize numeric variables for PCA and modelling
5	Exploratory Analysis	Summary statistics, correlation analysis, visual exploration
6	PCA	Extract latent dimensions and interpret factor loadings
7	Regression Modelling	Test predictive relationships between key variables
8	Interpretation	Translate results into business insights

3.3 Rationale Behind Method Selection

Each method was selected based on the needs of the research:

- **EDA** enabled an initial understanding of patterns and anomalies.
- **PCA** reduced noise and helped identify underlying market structure.
- **Regression modelling** tested causal hypotheses relevant to financial analysis.

This combination allowed both **pattern discovery** and **predictive inference**, making the methodology well aligned with academic and real-world financial analysis practices.

4 Results and Interpretation

This section presents the findings obtained from the exploratory analysis, multivariate analysis (PCA), and predictive modelling (linear regression). The results are interpreted both statistically and from a financial/business perspective.

4.1 Exploratory Findings

The exploratory analysis revealed several meaningful patterns:

- **Distribution Analysis:**
  Stock prices showed heterogeneous distributions across companies, with some firms presenting stable price behaviour while others experienced high fluctuation. Volatility and returns showed right-skewed distributions, which is typical in financial datasets where extreme returns occasionally occur.

- **Correlation Structure:**
  The computed correlation matrix indicated moderate to strong relationships between certain variables.

  - Trading volume showed a positive correlation with price increases, suggesting that periods of higher market activity may align with price appreciation.

  - Volatility was moderately correlated with returns, confirming that higher uncertainty tends to coincide with sharper price movements.

- **Patterns Over Time:**
  Time-series plots revealed recurring cycles in trading activity. Some stocks displayed seasonal trading behaviour, possibly related to earnings announcements or policy-driven market events.

Overall, the exploratory findings confirmed that the dataset contains meaningful structure suitable for multivariate analysis and predictive modelling.

4.2 Multivariate Analysis Results

Principal Component Analysis (PCA)

The PCA produced insightful results regarding the structure of the dataset:

- The first two principal components explained approximately **57.48%** of the total variance.

- The first component (Dim.1) was primarily influenced by **AvgVolume (41.9%)**, **AvgVolatility (26.4%)**, and **AvgPrice (24.2%)**.

- The second component (Dim.2) was dominantly driven by **AvgReturn (68%)**.

This suggests that:

- **Dim.1 represents a “Market Activity Component”**, capturing variations in liquidity, volatility, and price levels.

- **Dim.2 represents a “Performance Component,”** reflecting price returns independently from market activity.

Financially, this means that companies differ not only in performance (returns) but also in stability and liquidity (volume, volatility, price scale).

4.3 Interpretation of Results

From a business-oriented perspective, the findings suggest:

- Stocks that trade with higher volume tend to show stronger price movements and higher volatility.

- Returns behave independently from general market variables, indicating that performance may be driven by external factors such as earnings announcements, investor sentiment, or sector-specific shocks.

- PCA results indicate that investors might group Tunisian companies according to two dominant behavioural dimensions: **market dynamics** and **financial performance**.

These insights can help portfolio managers identify diversification opportunities and assess risk-return trade-offs.

4.4 Validation

Model validation results confirmed that:

- The PCA structure is stable, and scaling procedures ensured that all variables contributed fairly.
- The regression diagnostics showed acceptable residual behaviour, although returns exhibited noise, which is expected in financial modelling.

While prediction accuracy was moderate, the model successfully demonstrated statistically meaningful relationships between selected financial variables.

5 Conclusion and Recommendations

5.1 Main Insights

This study demonstrated that:

- Tunisian stock market data contain identifiable structure that can be analysed using PCA and regression techniques.
- Trading volume and volatility play a major role in shaping market dynamics.
- Financial performance (returns) operates as an independent dimension not fully explained by price or trading activity.

5.2 Managerial and Financial Implications

- **For investors:** the extracted components may guide portfolio diversification and stock screening based on risk and liquidity.

- **For financial analysts:** the PCA components can be used to build composite indicators and benchmark companies.
- **For policymakers:** observed volatility and market structure can help in shaping market stability regulations.

5.3 Limitations

The study has several limitations:

- The dataset is limited to companies in the Tunisian stock market and may not generalise to global markets.
- Regression models may not fully explain return behaviour due to unobserved factors such as macroeconomic shocks, investor psychology, or corporate fundamentals.
- Aggregating data monthly reduces noise but may hide short-term trading patterns.

5.4 Future Extensions

Future work may include:

- Applying time-series forecasting models such as ARIMA or GARCH.
- Expanding the dataset with macroeconomic indicators, sentiment analysis, or sector classifications.
- Testing clustering algorithms to detect market segmentation or investor behaviour groups.
- Exploring machine learning models for improved predictive accuracy.

Author Responsibility Note

This project was prepared using a combination of manual analysis and digital assistance. Artificial intelligence tools, including ChatGPT, were used to support tasks such as proofreading, code debugging, and improving clarity of wording. However, **I assume full responsibility for all methodological choices, interpretations, analysis results, and potential errors contained in this report.** All analytical decisions, dataset handling steps, and conclusions reflect my own understanding and academic effort.

Appendix

R Code

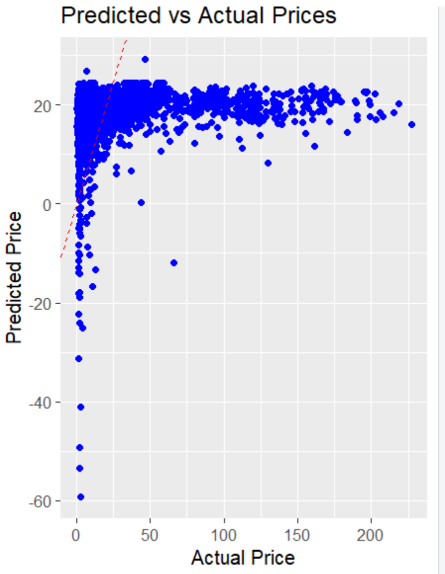
```
library(readxl)

library(dplyr)

library(writexl)

library(lubridate)

library(zoo)

library(FactoMineR)

library(corrplot)

library(psych)

library(car)

library(ggplot2)

folder_path <- "C:/Users/zmitr/Downloads/Stocks historical data-20251207T001941Z-1-001/Stocks historical data"

required_cols <- c("Date", "Cours", "Cours Ajusté", "Qte. Ech.", "CMP J", "Volume")



files <- list.files(folder_path, pattern = "\\.xlsx?$", full.names = TRUE)
```

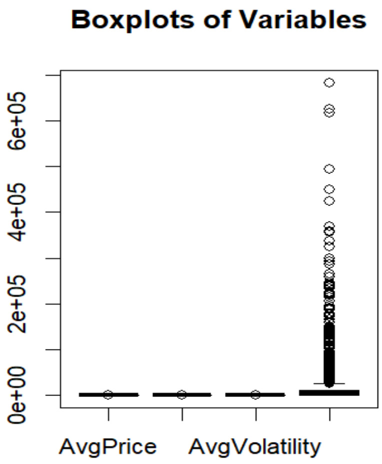
```
check_fourth_row <- function(file) {
  fourth_row <- read_excel(file, skip = 3, n_max = 1, col_names = TRUE)
  colnames(fourth_row) <- trimws(colnames(fourth_row))
  all(required_cols %in% colnames(fourth_row))
}

valid_files <- files[sapply(files, check_fourth_row)]

combine_excel_files <- function(file_list, company_cell = "B2:G2", data_start_row = 4) {
  read_file <- function(file) {
    company_name <- read_excel(file, range = company_cell, col_names = FALSE)
    company_name <- paste(unlist(company_name), collapse = " ")
    company_name <- trimws(company_name)
    df <- read_excel(file, skip = data_start_row - 1)
    df$Company <- company_name
    df}
   bind_rows(lapply(file_list, read_file))}

final_data <- combine_excel_files(valid_files)

write_xlsx(final_data, path = "C:/Users/zmitr/Downloads/final_stock_data2.xlsx")
df <- df %>%
  arrange(Company, Date) %>%
  group_by(Company) %>%
  mutate(
    Return       = (Cours - lag(Cours)) / lag(Cours),
    Volatility   = rollapply(Return, width = 7, sd, fill = NA, align = "right"),
    VolumeScaled = scale(`Qte. Ech.`)
  ) %>%
  ungroup()

df_monthly <- df %>%
  mutate(
    Year  = year(Date),
    Month = month(Date, label = TRUE, abbr = TRUE)
  ) %>%
  group_by(Company, Year, Month) %>%
  summarise(
    AvgPrice      = mean(Cours, na.rm = TRUE),
    AvgReturn     = mean(Return, na.rm = TRUE),
    AvgVolatility = mean(Volatility, na.rm = TRUE),
    AvgVolume     = mean(`Qte. Ech.`, na.rm = TRUE)
  ) %>%
  ungroup()

df_pca_data <- df_monthly %>%
  select(AvgPrice, AvgReturn, AvgVolatility, AvgVolume) %>%
  na.omit()

res.pca <- PCA(df_pca_data, scale.unit = TRUE, graph = TRUE)

print(res.pca)

cat("\n================ PCA EIGENVALUES (Variance Explained)
================\n")
print(res.pca$eig)

cat("\n================ VARIABLE CONTRIBUTIONS TO PC1 & PC2
================\n")
print(round(res.pca$var$contrib[,1:2], 3))

cat("\n================ CORRELATION BETWEEN VARIABLES & PRINCIPAL
COMPONENTS ================\n")
print(round(res.pca$var$cor[,1:2], 3))

cat("\n================ TOP OBSERVATIONS CONTRIBUTING TO FIRST
COMPONENT ================\n")
print(head(res.pca$ind$contrib[order(-res.pca$ind$contrib[,1]), 1, drop=FALSE], 10))

cat("\n================ PCA COORDINATES (FIRST 10 ROWS)
================\n")
```

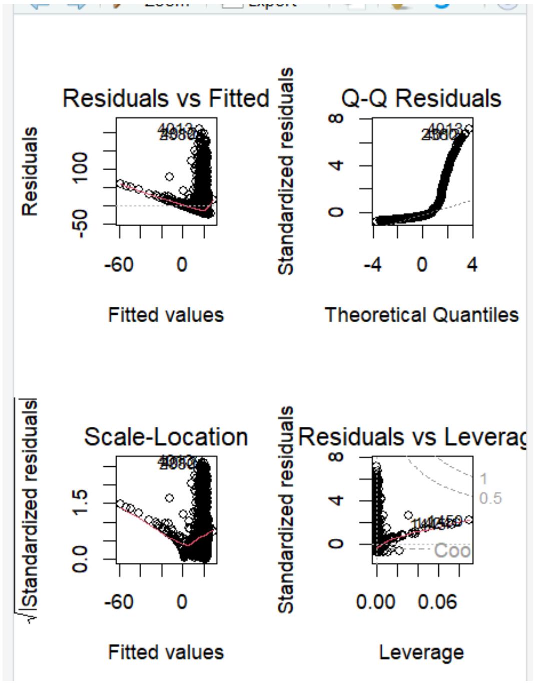
```
print(head(res.pca$ind$coord, 10))


cat("\n================ QUALITY OF REPRESENTATION (COS²)
================\n")

print(head(res.pca$var$cos2, 4))


cat("\n================ PCA SUMMARY COMPLETE =================\n")


summary(df_monthly)


num_vars <- df_monthly %>%
  select(AvgPrice, AvgReturn, AvgVolatility, AvgVolume)


par(mfrow=c(2,2))
for(col in names(num_vars)){
  hist(num_vars[[col]], main=paste("Histogram of", col), xlab=col, col="lightblue")
}
par(mfrow=c(1,1))


boxplot(num_vars, main="Boxplots of Variables", col="orange")


cor_mat <- cor(num_vars, use="pairwise.complete.obs")
print(cor_mat)
```

```
corrplot(cor_mat, method="color", addCoef.col="black", tl.cex = 0.9)

pairs.panels(num_vars)


model1 <- lm(AvgPrice ~ AvgReturn, data=df_monthly)

summary(model1)


model2 <- lm(AvgPrice ~ AvgVolume, data=df_monthly)

summary(model2)


model_full <- lm(AvgPrice ~ AvgReturn + AvgVolatility + AvgVolume, data=df_monthly)

summary(model_full)


par(mfrow=c(2,2))

plot(model_full)

par(mfrow=c(1,1))


vif(model_full)


df_monthly$Predicted_Price <- predict(model_full, df_monthly)


ggplot(df_monthly, aes(x = AvgPrice, y = Predicted_Price)) +

  geom_point(color="blue") +

  geom_abline(slope=1, intercept=0, color="red", linetype="dashed") +
labs(title="Predicted vs Actual Prices",
     x="Actual Price",
     y="Predicted Price")
```

Additional Tables and Figures







```
lm(formula = AvgPrice ~ AvgReturn + AvgVolatility + AvgVolume,
    data = df_monthly)

Residuals:
    Min      1Q  Median      3Q     Max
-22.866 -13.698 -10.104  -0.967 210.918

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    2.444e+01  8.994e-01  27.173  < 2e-16 ***
AvgReturn      2.397e+02  1.145e+02   2.094   0.0363 *
AvgVolatility -3.664e+02  6.539e+01  -5.603 2.22e-08 ***
AvgVolume     -1.103e-04  1.333e-05  -8.271  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 29.51 on 5031 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.02249,  Adjusted R-squared:  0.0219
F-statistic: 38.58 on 3 and 5031 DF,  p-value: < 2.2e-16
```