Master of Data Analytics
Spring 2024

ANALYZING THE IMPACT OF FINANCIAL INDICATORS ON STOCK MARKET PERFORMANCE

EMPOWERING INVESTORS TO MAKE INFORMED DECISIONS

Author: Olisa Unegbu

Student N°: 22020843

Supervisor: Professor Preeti Patel

(*Head of Computer Science and Applied Computing*)

LONDON METROPOLITAN UNIVERSITY

London Metropolitan University

School of Computing and Digital Media

Acknowledgement

I would like to express my sincere gratitude to my parents, whose firm dedication to providing me with a quality education has been the base of my journey. A special acknowledgment goes to my mother, whose resilience in the face of adversity, following the passing of my father in 2006, has been nothing short of remarkable. Despite the challenges, she single-handedly carried the responsibility, ensuring that I had access to opportunities that they themselves never had. Without her steadfast support and commitment, I would not have achieved the milestones I have reached. I am eternally grateful to her.

I am also deeply thankful and extend my heartfelt appreciation to my supervisor, Professor Preeti Patel. Her invaluable guidance, insightful contributions, and timely advice have been instrumental throughout the duration of my project. Professor Patel's supervision served as a constant source of knowledge and encouragement, empowering me to push through the challenges of my research journey with confidence.

Finally, I extend my appreciation to all my module leaders and colleagues. Your collective support and friendship have been invaluable, and I am truly thankful for the enriching academic year we shared. The transfer of knowledge was awesome, and I am indebted to each of you for contributing to our collective success.

Table of Contents

Table of Figures

Table of Tables

Abbreviations

Adj ---------- Adjusted

AO ---------- AO World Group

ARCH ---------- Autoregressive Conditional Heteroscedastic

ARIMA ---------- Autoregressive Integrated Moving Average

ASHM ---------- Ashmore Group

BAG ---------- A.G. Barr Group

BBY ---------- Balfour Beatty PLC

BI ---------- Business Intelligence

BLND ---------- British Land

BOY ---------- Bodycote

CRISP-DM ---------- Cross-Industry Standard Process for data mining

D/E ---------- Debt-to-equity

DBMS ---------- Database Management System

DJIA ---------- Dow Jones Industrial Average

DY ---------- Dividend Yield

EMG ---------- Man Group

EPS ---------- Earnings per Share

FCF ---------- Free Cash Flow

FI ---------- Financial Indicator

FOUR ---------- 4imprint Group plc

GARCH ---------- General Autoregressive Conditional Heteroscedastic

GB ---------- Gradient Boosting

LR ---------- Linear Regression

MSE ---------- Mean Squared Error

NASDAQ ---------- National Association of Securities Dealers Automated Quotations

NPM ---------- Net Profit Margin

OOP ---------- Object-Oriented Programming

P/B ---------- Price-to-book

P/E ---------- Price Earnings

PNN ---------- Pennon Group

PPH ---------- PPHE Hotel Group

PZC ---------- PZ Cussons

R^2 ---------- R-squared

RDBMS ---------- Relational Database Management System

RF ---------- Random Forest

ROE ---------- Return on Equity

S&P 500 ---------- Standard and Poor's 500

SGR ---------- Sales Growth Rate

SVM ---------- Support Vector Machine

Abstract

Investors often encounter challenges when making informed investment decisions. This project aims to assist investors, policymakers, and financial analysts in making data-driven decisions to minimize investment losses and optimize investment outcomes by integrating multiple financial indicators and historical stock market price data. Using the CRISP-DM framework and machine learning techniques, the project focuses on selecting 12 companies listed on the London Stock Exchange (LSE) from various sectors and incorporates 10 financial indicators. The purposive sampling method ensures diversity and reduces bias, with companies selected from the Financial Times Stock Exchange (FTSE250) to provide a focused perspective on UK-based operations. The analysis encompasses two distinct periods: pre-COVID-19 and post-COVID-19, using data from 2017 to 2019 and 2021 to 2023. A total of 24 datasets sourced from Yahoo! Finance were utilized, encompassing stock prices for each selected company. Excel and Power BI were used for preprocessing and exploratory analysis, respectively. During the predictive modelling phase, three main predictive models - random forest, gradient boosting, and support vector machine (SVM) - were applied to historical stock prices datasets using Python. Linear regression was utilized in cases where it outperformed the main models. Model performance was evaluated using metrics such as mean squared error (MSE) and R-squared (R2) score. The gradient boosting model demonstrated the highest frequency of 15 best performances across various scenarios, while random forest performed best five times and SVM four times. Additionally, autoregressive integrated moving average (ARIMA) modelling was employed for time series analysis to forecast future adjusted close prices.

Keywords: *Investment Decision, Financial Indicators, Stock Market Prediction, Machine Learning, Data Analysis*

Chapter 1. Introduction

1.1. Introduction

Within the scope of financial institutions, analytics plays a pivotal role in the stock market sector, providing a set of data mining, financial modelling, and statistical methods. These tools enable organizations to enhance their operations, improve investor experiences, and make strategic decisions. By leveraging analytics, financial institutions can stay ahead of the competition and ensure optimal performance in their activities.

However, one of the significant challenges faced by the stock market is the limited availability of financial indicators. This scarcity of information often leads to uninformed investment decisions, resulting in potential losses for investors. To address this issue, this project employs predictive analytics methods supported by a strong methodology. The goal is to bring together various financial indicators, empowering investors, policymakers, and financial analysts with a comprehensive understanding of these indicators. By leveraging data-driven insights, stakeholders can make informed investment decisions, thereby reducing the risk of investment loss.

To develop accurate predictive models, supervised machine learning techniques were employed. The Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology guides and structures the analysis across its various phases. This includes understanding the data, exploring patterns, cleaning the data, and modelling the datasets. Subsequently, the performance of the fitted models is compared, and the most accurate model is selected for deployment for each company.

Throughout this project, a combination of software tools and programming languages was utilized to efficiently execute the analysis. This approach ensures that all tasks are carried out seamlessly, addressing the diverse needs of the analysis process. By integrating these methodologies and technologies, this project aims to provide valuable insights into stock market performance and empower stakeholders to make informed investment decisions.

1.1. Business Understanding

In the dynamic landscape of the stock market, investors face significant challenges that often result in substantial losses. One of the primary contributors to these losses is the limited understanding of financial indicators, which play a crucial role in shaping stock market performance. Investors frequently base their investment decisions on a narrow set of indicators, overlooking the broader financial landscape of companies.

Moreover, publicly traded companies often publish a selective set of financial indicators that portray a favourable image, potentially concealing critical aspects of their financial performance and sustainability. Research indicates that a staggering 90% of investors experience losses when trading stocks, highlighting the urgent need for a more comprehensive approach to investment decision-making.

Numerous factors contribute to investor losses, including a lack of understanding of financial indicators, misconceptions about the ease of making money in the stock market, investor sentiments and behaviours, economic indicators, and policy changes, among others. To address these challenges and empower investors to make informed decisions, it is imperative to consider a broader range of financial indicators.

Financial institutions recognize the importance of predictive analytics in enhancing stock market performance and investor satisfaction. By leveraging predictive analytics methods, financial institutions can analyse vast amounts of data, identify meaningful patterns and trends, and generate accurate predictions about stock market behaviours. This enables investors to navigate the complexities of the market with confidence, mitigate risks, and achieve their investment objectives.

1.2. Aims and Objectives

The primary aim of this project is to integrate multiple financial indicators and conduct predictive analysis using data mining techniques and machine learning algorithms. The goal is to develop strong models capable of accurately forecasting future stock prices for selected companies, thereby providing investors with comprehensive insights into financial indicators and anticipated stock price movements.

To achieve this aim, several objectives have been identified:

1. **Data Integration and Pattern Recognition**: The initial objective is to consolidate various financial indicators and identify meaningful patterns within the datasets. This involves determining which measures or explanatory variables serve as the strongest predictors for training the predictive models effectively.
2. **Data Preparation Automation:** Following data integration, the objective is to streamline the data preparation process by developing an automated script using Python programming. This script will address tasks such as handling missing values, detecting outliers, and performing necessary data manipulations, thereby saving time, and ensuring consistency in future analyses.
3. **Model Development and Comparison:** Three distinct supervised learning models will be constructed to establish the relationship between explanatory and response variables for each company. These models will be trained and evaluated using different machine learning algorithms to determine their effectiveness in predicting future stock prices accurately. The objective is to identify the most suitable model for each company through comprehensive comparison analysis.
4. **Academic Contribution:** In addition to the practical objectives, the project aims to leverage a diverse range of analytical tools, skills, and techniques acquired during the master's program in Data Analytics. By incorporating these capabilities, the analysis aims to deliver rich and insightful findings that align with the business requirements of the project. The key deliverables include integrating financial indicators, automating data preparation, building accurate predictive models, and comparing their performance to assist investors in making informed investment decisions.

In summary, the aims and objectives of this project encompass the integration of financial indicators, development of predictive models, automation of data preparation, and academic contribution to the field of Data Analytics. These objectives collectively aim to enhance investor decision-making processes by providing them with comprehensive insights and accurate predictions of future stock prices.

Chapter 2. Literature Review

Financial theory, notably the efficient market hypothesis (EMH) introduced by Fama (1965), posits that stock prices incorporate all available information, including various financial indicators. The efficient market hypothesis suggests that markets are efficient, with prices accurately reflecting intrinsic values. According to this theory, it is challenging for investors to consistently outperform the overall market through expert stock selection or market timing. Instead, the theory proposes that higher returns can only be achieved by investing in riskier assets.

A multitude of empirical studies have investigated the relationship between financial indicators and stock market performance, forming the basis of this literature review. This review aims to provide a comprehensive overview of existing research on how financial indicators influence stock market dynamics, investor behaviour, and overall market efficiency.

2.1. Financial Indicators

Financial indicators are pivotal metrics utilized by investors, analysts, and policymakers to comprehend and forecast stock market performance. They serve as fundamental tools guiding investment decisions, offering insights into the financial health and potential of companies or assets. While some investors rely on a limited number of metrics for investment decisions, this project aims to broaden the spectrum of considerations available to investors.

2.1.1. Earnings per share (EPS)

According to Musallam (2018), whose research explored the financial ratios and market stock returns of 26 Qatari listed firms from 2009 to 2015, the results of weighted least square (WLS) analysis revealed significant and positive relationships between earnings per share, earnings yield ratio, dividend yield ratio, and stock market returns. This underscores the importance of EPS and related financial metrics in predicting and understanding stock market performance.

Earnings per Share (EPS) signifies the portion of a company's profit allocated to each outstanding share of common stock, serving as a key indicator of profitability. Calculated by

dividing the company's net profit by its outstanding shares of common stock, EPS provides investors with valuable insights into a company's earning potential.

$$EPS = \frac{\text{Net Income} - \text{Preferred Dividends}}{\text{Weighted Average Shares Outstanding}}$$

2.1.2. Price-earnings ratio (P/E ratio)

Umamaheswari, S., Suresh, C.K., and Sampathkumar, S. (2021) conducted a project involving 12 listed companies across various industrial sectors on the National Stock Exchange of India. Their research aimed to determine the effect of accounting variables such as price-earnings ratio, earnings per share, and dividend payout ratio on stock cost volatility over a five-year period (2014 – 2019). The findings revealed that while accounting indicators like earnings per share and dividend payout ratio contributed to stock price instability during the period, the price-earnings ratio did not significantly impact stock volatility.

The Price-Earnings Ratio (P/E ratio) serves as a measure of a company's current share price relative to its per-share earnings. Calculated by dividing the market value price per share by the company's earnings per share, the P/E ratio offers insight into how the market values the company's earnings potential.

$$P/E\ ratio = \frac{\text{Current price}}{most\ recent\ earnings\ per\ share}$$

2.1.3. Return on Equity (ROE)

Indraswono (2021) conducted hypothesis testing to examine the impact of modern and traditional performance indicators on stock return. Using purposive sampling on 29 companies indexed by Dow Jones during the 2015-2018 period, the project found that while modern performance indicator economic value added had an insignificant and negative effect on stock return, traditional performance indicators like return on equity, return on assets, earnings per share, and dividend per share had a significant and positive effect on stock return.

Return on Equity (ROE) serves as a metric of a corporation's profitability and efficiency in generating profits relative to its shareholders' equity. It is calculated by dividing the company's net income by its average shareholders' equity.

$$ROE = \frac{\text{Net Income}}{Average\ Shareholders' Equity}$$

2.1.4. Debt-to equity ratio (D/E ratio)

Tlemsani (2020) analysed the debt equity, book-to-market value of equity, sales-to-price, and firm size as financial variables to determine their impact on stock returns. Using a sample of 30 companies listed in the KSA stock market from five industries, the project found that firm size (size effect) was a significant determinant of stock returns in the KSA stock market. Regression analysis was employed to examine the correlation between stock returns and the selected financial variables.

The Debt-to-Equity Ratio (D/E ratio) is utilized to evaluate a company's financial leverage, calculated by dividing a company's total liabilities by its shareholder equity.

$$D/E\ ratio = \frac{\text{Total Liabilities}}{Total\ Shareholders' Equity}$$

2.1.5. Dividend Yield

Sondakh (2019) conducted an analysis on dividend policy, liquidity, profitability, and firm size's impact on firm value. Using purposive sampling from 99 financial service companies listed on the Indonesia Stock Exchange between 2015-2018, the project revealed that dividend policy had a negative and significant effect on firm value. Additionally, liquidity and firm size positively and significantly influenced firm value, while profitability did not exhibit significant impact on firm value, based on multiple linear regression data analysis.

Dividend Yield represents a financial ratio indicating how much a company distributes in dividends annually relative to its stock price. It is calculated by dividing the annual dividends per share by the price per share.

$$Dividend\ Yield = \frac{\text{Annual Dividends Per Share}}{Price\ Per\ Share}$$

2.1.6. Price-to book ratio (P/B ratio)

Almumani and Almazari (2021) examined the impact of major financial indicators on market capitalization using a sample of 76 companies listed on the Amman Stock Exchange Market,

with 608 observations spanning 2013-2019. Their project, employing descriptive and analytical methods, revealed statistically significant effects on market capitalization for indicators such as price to book value ratio, dividends per share, earnings per share ratio, return on assets ratio, total assets turnover ratio, and debit ratio.

The Price-to-Book Ratio (P/B ratio) is a financial metric comparing a company's current market value to its book value, derived from subtracting liabilities from assets. It is calculated by dividing the company's current stock price per share by its book value per share.

$$P/B\ ratio = \frac{\text{Market Price per Share}}{Book\ Value\ per\ Share}$$

2.1.7. Free Cash Flow (FCF)

Meryana and Setiany (2021) examined free cash flow, investments, earnings management, and interest coverage ratio using 33 companies categorized as healthy enterprises, selected through purposive sampling. Their analysis revealed that all tested indicators impact the risk of financial distress in healthy companies.

Free Cash Flow (FCF) signifies the cash a company generates after covering operating expenses and capital expenditures. It is calculated by subtracting capital expenditures from operating cash flow.

Free Cash Flow = Operating Cash Flow – Capital Expenditures

2.1.8. Net Profit Margin

Astuti (2021) observed consistently strong growth trends in net profit margins within the cruise industries among leading players from 2016 to 2019. Despite experiencing revenue declines in most commercial shipping market segments following the 2008 global financial crisis, the cruise shipping sector demonstrated resilience and relatively rapid recovery.

Profit Margin serves as a metric indicating the profitability of a company or business activity by gauging the percentage of sales that translates into profits. The net profit margin, calculated as the ratio of net income to revenue multiplied by 100, reflects a company's overall ability to convert income into profit.

$$NPM = \left(\frac{\text{Net income}}{Revenue}\right) * 100$$

2.1.9. Sales Growth Rate

Lee, Wang, and Ho (2020) explored the impact of financial inclusion and financial innovation on firms' sales growth rates, particularly in non-Asian regions during normal times. Their findings revealed that financial innovation had a negative impact on the sales growth rate of firms engaged in financial inclusion.

Sales Growth Rate measures the rate at which a business can increase revenue from sales over a fixed period. It is calculated by dividing the difference between current period sales and prior period sales by the prior period sales value and multiplying by 100.

$$SGR = \left(\frac{current\ period\ sales - prior\ period\ sales}{prior\ period\ sales}\right) * 100$$

2.1.10. Beta Value

Faiteh and Aasri (2022) investigated the use of accounting beta, calculated using return on assets (ROA) and return on equity (ROE), as a representation of market beta. Their project, conducted on a sample of 49 companies listed on the Casablanca Stock Exchange from 2015 to 2019, concluded that accounting beta significantly represents market beta and provides a satisfactory solution for calculating the cost of equity for unlisted firms.

Beta Value serves as a measure of a stock's volatility or systematic risk relative to the overall market. It indicates how changes in a stock's returns are related to changes in the market's returns. Beta is calculated by dividing the covariance of the security's returns and the market's returns by the variance of the market's returns over a specified period.

$$\beta = \frac{\text{Covariance}(Re,\ Rm)}{Variance(Rm)}$$

where:

R_e = the return on an individual stock

R_m = the return on the overall market

Covariance = how changes in a stock's returns are related to changes in the market's returns

Variance = how far the market's data points spread out from their average value

Investors utilize a diverse range of financial tools to effectively manage their investments. Among these tools are StockCharts, Mint, Cash Flow, Balance Sheet, TradingView, PortfolioVisualizer, ETFdb, SigFig, Finviz, FeeX, Kubera, and others. It is important to note that there is no one-size-fits-all financial tool suitable for every investor. Different tools cater to different investing styles, and the choice of tool heavily depends on the investor's individual needs, risk tolerance, and investment preferences.

This project aims to incorporate a comprehensive set of financial indicators to aid both risk-tolerant and risk-averse investors in making well-informed investment decisions. By combining multiple financial indicators, investors can gain a deeper understanding of market trends, assess risks more accurately, and optimize their investment strategies to achieve their financial goals.

2.2. Stock Market Performance

Stock market performance is heavily dependent on financial indicators. The stock market represents a pivotal component of the financial system, drawing considerable attention from researchers (Raza, et al. 2023). Stock market performance encapsulates the overall interplay and trajectory of stock prices within a specific market or exchange over a defined period. Analysing stock market performance entails examining various metrics, trends, and financial indicators that shape investor decisions.

A central financial metric in assessing stock market performance is the stock market indices, notably including the Dow Jones Industrial Average (DJIA), Standard and Poor's 500 Composite Index (S&P 500), and National Association of Securities Dealers Automated Quotations (NASDAQ). These indices serve as barometers of market health and trends, reflecting the collective performance of listed companies.

i. **Dow Jones Industrial Average (DJIA):** The DJIA comprises 30 prominent companies listed on U.S. stock exchanges. Founded by Charles Dow on February 16, 1885, it is price-weighted and encompasses large-cap companies. As of December 29, 2023, its market cap stands at US$12.0 trillion.

Figure 2.1: Dow Jones Industrial Average 1970 to 2022 (Wikipedia)

ii. **The Standard and Poor's 500 (S&P 500):** The S&P 500 tracks the stock performance of 500 of the largest U.S. companies listed on stock exchanges. It represents approximately 80% of the total market capitalization of U.S. public companies. Formed on March 4, 1957, it is a free-float weighted index. Its market cap as of December 30, 2023, is US$42.0 trillion.

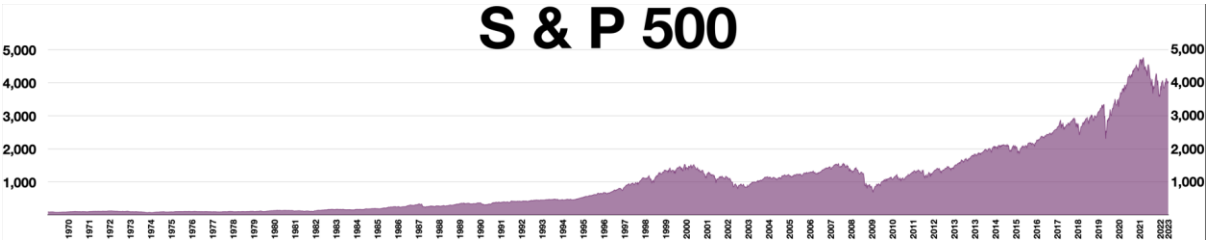


Figure 2.2: S&P 500 Index from 1970 to 2023 (Wikipedia)

iii. **National Association of Securities Dealers Automated Quotations (NASDAQ):** NASDAQ, established on February 8, 1971, is a prominent American stock exchange based in New York City. It is renowned for its high trading volume and ranks second in market capitalization of shares traded, behind the New York Stock Exchange. As of May 2023, its market cap is US$20.13 trillion, with 4,204 listings.

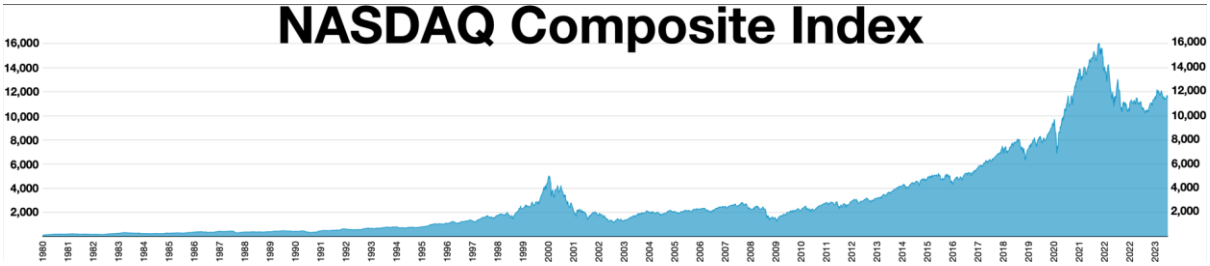


Figure 2.3: NASDAQ Composite Index 1980 to 2023 (Wikipedia)

These indices provide insights into market trends, investor sentiment, and economic conditions, serving as critical tools for investors, analysts, and policymakers in assessing stock market performance and making informed decisions. Through statistical techniques such as data mining, financial modelling, and data visualization, this project analysed the dynamics and implications of stock market performance on broader economic landscapes.

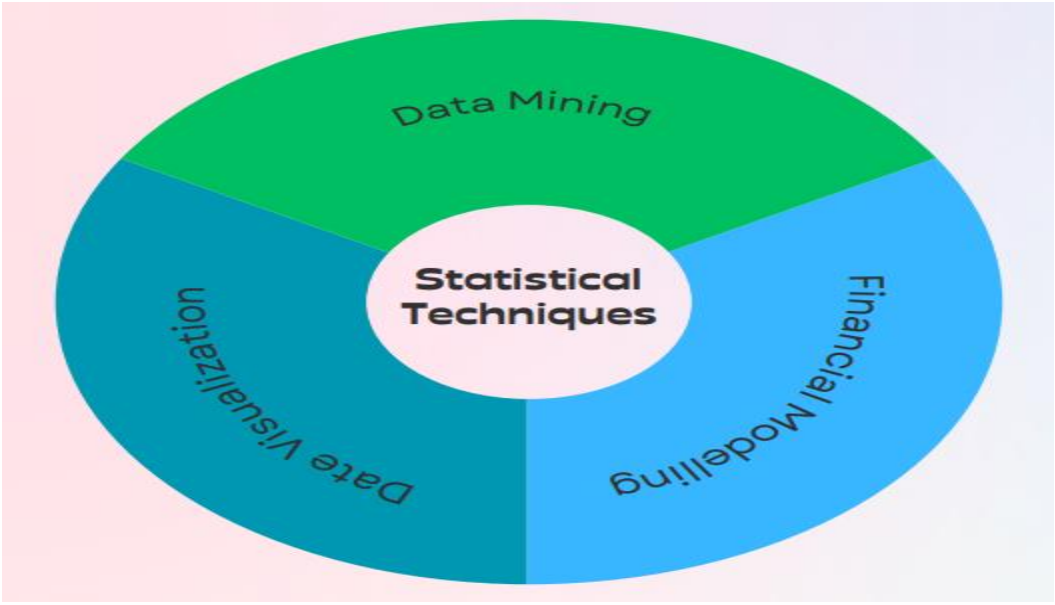


Figure 2.4: statistical techniques used for the project.

2.3. Data Mining & CRISP Methodology

The CRISP Methodology (CRISP-DM), showed as a cyclical process, guides the systematic execution of data mining projects. The CRISP-DM cycle comprises the following phases:

- Business understanding
- Data understanding
- Data preparation
- Data exploitation

- Modelling
- Evaluation
- Deployment

CRISP-DM, or the Cross-industry standard process for data mining, offers a structured process model for planning and executing data mining projects. It provides a framework that ensures the comprehensive exploration, analysis, and utilization of data to derive meaningful insights and inform decision-making processes.

Data mining encompasses the science and art of exploring and analysing intricate and extensive historical datasets to uncover valuable patterns and relationships among attributes, aligned with specific analysis objectives. It involves the semi-automatic process of extracting or discovering patterns in large datasets, employing methods at the intersection of machine learning, statistics, and database systems (Solis, 2023).

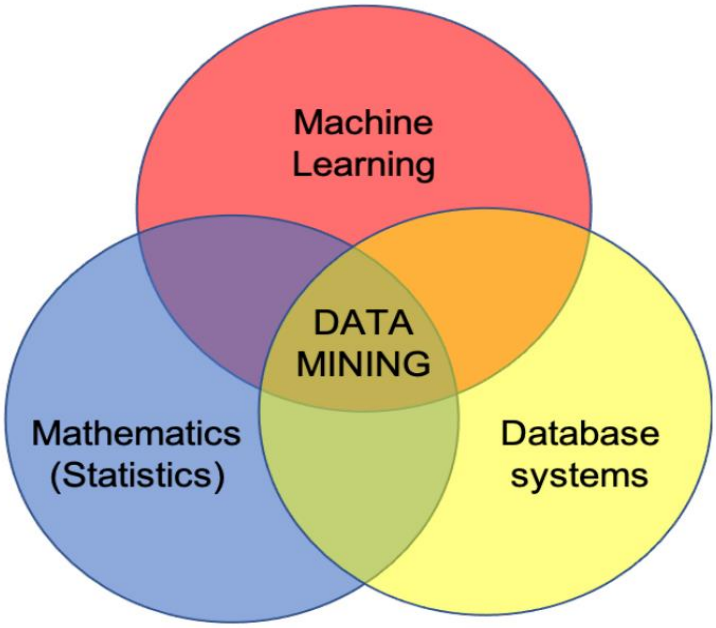


Figure 2.5: a pictorial description of data mining (Solis, 2023).

Figure 2.5. provides a visual representation of data mining, illustrating its complexity and multidisciplinary nature (Solis, 2023). With the proliferation of technology and the accumulation of vast historical data across industries and sectors, data mining emerges as a vital method to simplify the exploration and analysis of these large and complex datasets. It

integrates various disciplines such as database systems, statistical analysis, visualization techniques, and machine learning.

2.3.1. Machine Learning

Machine Learning is revolutionizing various aspects of our lives, accomplishing tasks previously reserved for expert humans. In the realm of finance, adopting disruptive technologies like Machine Learning marks an exciting era that promises to transform investment strategies for generations (Prado, 2018). Machine Learning, a subset of artificial intelligence, represents a revolutionary automated algorithm capable of learning from past experiences, detecting meaningful patterns, and adjusting program actions without explicit programming. It merges elements of statistics and computer science to enable computers to imitate human learning processes, gradually improving accuracy in task-solving (DJALAL, 2020).

Machine Learning methods are broadly categorized into two groups:

2.3.1.1. Supervised Methods

Supervised methods aim to predict future outcomes by identifying and estimating patterns between independent variables (predictors) and dependent variables (target). The goal is to generate a function mapping inputs to anticipated outputs. Supervised learning involves a training phase with known target values and a testing phase with unknown target values to assess model accuracy. Accuracy is measured as the number of correct classifications divided by the total number of test cases (DJALAL, 2020).

$$Accuracy = \frac{\text{Number of correct classifications}}{Total\ number\ of\ test\ cases}$$

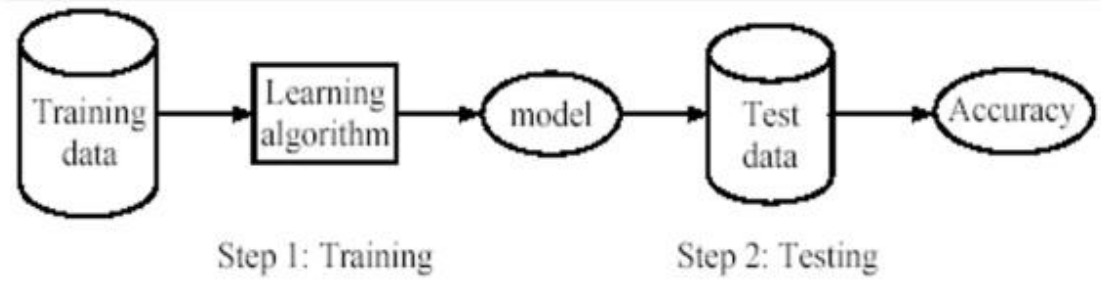


Figure 2.6: illustrates the phases of a supervised learning process (Solis, 2023)

2.3.1.2. Unsupervised methods

Unsupervised methods group illustrations without a pre-specified target. Unlike supervised learning, there is no specific target variable to predict. Instead, unsupervised methods employ clustering analysis to segment datasets into meaningful groups based on observations. The goal is to partition data into segments, detect patterns, and reduce dimensionality without a target variable for prediction or classification (Solis, 2023).

	Supervised	**Unsupervised**
Input data	labelled data	unlabelled data
Output data	target variable	no target variable
Computational complexity	simpler methods	computationally expensive
Accuracy	Highly accurate	less accurate
Uses	used to predict outcome variables	Used for exploration and visualization
Examples	customers segmentation; recommendation systems; data preparation for learning algorithms.	spam filters; price prediction; recommendation systems.

Table 2.1: provides an overview of supervised vs unsupervised learning algorithms (Solis, 2023)

2.3.2. Mathematics (Statistics)

Organizing and summarizing large datasets is a critical task, with statistical analysis serving as the primary tool for analysing quantitative data (Deborah and Pyrczak, 2023). Mathematical and statistical modelling relies on several key assumptions:

- **"All models are wrong, but some are useful":** Acknowledging that models may not perfectly represent reality but can still provide valuable insights.
- **The model should be fit for purpose:** Models should be designed and selected based on their intended application and objectives.
- **Most data analytic models have a mathematical part + stochastic part:** In models like simple linear regression, it is assumed that the relationship between variables includes a deterministic mathematical component along with a stochastic component, where errors are independent and normally distributed. In simple linear regression, it is assumed that:

$$y_i = a + bx_i + e_i$$

for i = 1,…., n and e_i independent normally distributed

where:

y = dependent variable

a = intercept

b = slope

x = independent variable

e = error term

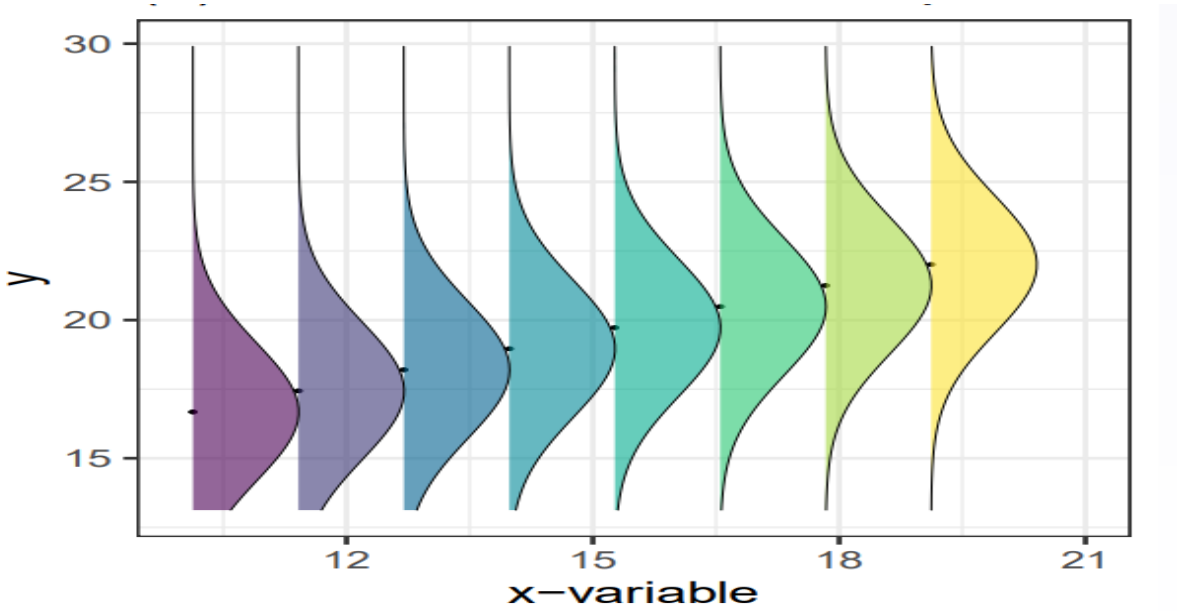


Figure 2.7: illustrates the assumptions underlying the linear model assumptions (Stasinopoulos, et al., 2022)

2.3.3. Database Systems

The primary goal of a database is to present users with an abstract view of the data, accommodating diverse user perspectives within an organization. Abstraction serves as the foundation for database design, and to achieve this and facilitate various views, most commercial Database Management Systems (DBMS) offer a standardized architecture.

Database Management Systems (DBMS) are software systems designed to store, retrieve, and execute queries on data. Acting as an intermediary between end-users and databases, DBMS enables users to create, read, update, and delete data within the database. It manages data, the database engine, and the database schema, ensuring data security, integrity, concurrency, and uniform administration procedures.

DBMS optimizes data organization through a schema design technique known as normalization, which involves splitting large tables into smaller ones to eliminate redundancy in attribute values (Hubenova, 2023).

2.4. Financial Modelling

The EURO working group on financial modelling defines Financial Modelling as "the development and implementation of tools supporting firms, investors, intermediaries, governments, and others in their financial-economic decision making, including the validation of the premises behind these tools and the measurement of the effectiveness of the use of these tools" (Spronk and Hallerbach, 1997). Financial Modelling aims to support individual decision making by considering the specific characteristics of each case while leveraging insights from financial theory.

Financial Modelling involves building a mathematical representation of real-world financial situations. This mathematical model simplifies the performance of financial assets, portfolios, businesses, projects, or investments (Wikipedia, 2024). Chun, Cho, and Ryu (2020) analysed market and economic indicators, classifying them into five groups: implied volatility indices, stock market indicators, interest rate spreads, foreign exchange rates, and commodity spots. These variables helped identify both general and unique features of daily market volatility dynamics.

Volatility holds significance in financial economics as a measure of risk and a crucial factor in investment and trading decisions. It is defined as the annualized standard deviation of the change in price or value of a financial security (Hossain, A., 2023). Andersen et al. (1999) confirmed the effectiveness of the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model in forecasting volatility across different prediction horizons, particularly with the use of intraday returns.

2.5. Data Visualization

The challenge of analysing vast scientific datasets led to the emergence of scientific visualization. In the past decade, the size of datasets has grown rapidly, especially those generated from simulations of dynamic phenomena like time-dependent flows. However, many visualization techniques, particularly global field visualization techniques, struggle to scale to exceptionally large datasets, making them less suitable for analysing time-dependent data (Post, Nielson, and Bonneau, 2002).

Data visualization is integral to the scientific process. Effective visualizations empower scientists to comprehend their data and convey insights to others (Tukey, 1977). Waskom (2021) utilized seaborn, a Python library for statistical graphics, which offers a high-level interface to matplotlib and seamlessly integrates with pandas data structures. It demonstrates declarative API, semantic mappings, subplot faceting, aggregation with error bars, and visual theme control.

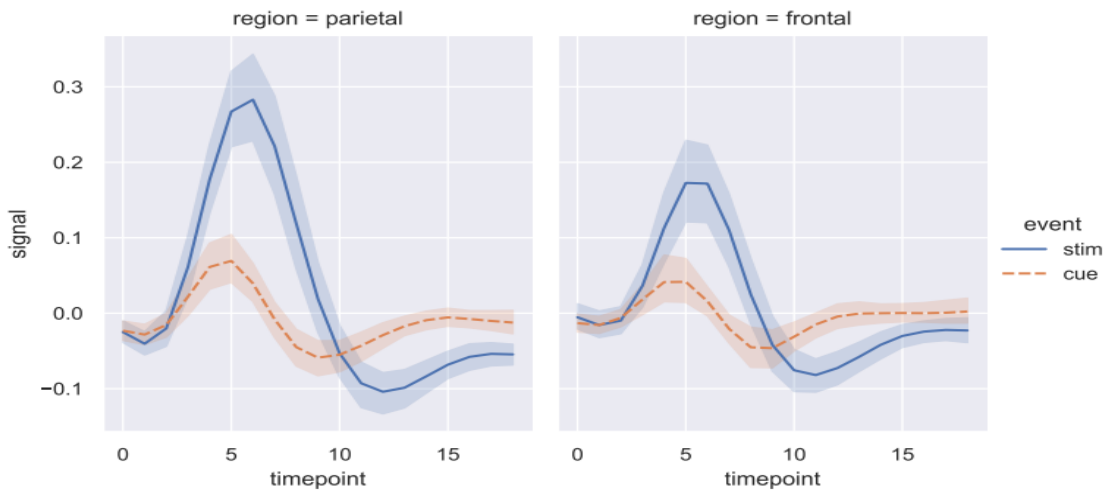


Figure 2.8: An example seaborn figure demonstrating some of its key features. The image was generated using seaborn v0.11.1. (Waskom, 2021)

Learning effective data visualization transcends writing code that generates figures from data. Scatterplots are pivotal in social science data visualization, as demonstrated by Healy (2019) to illustrate the usefulness of examining model fits and data concurrently.

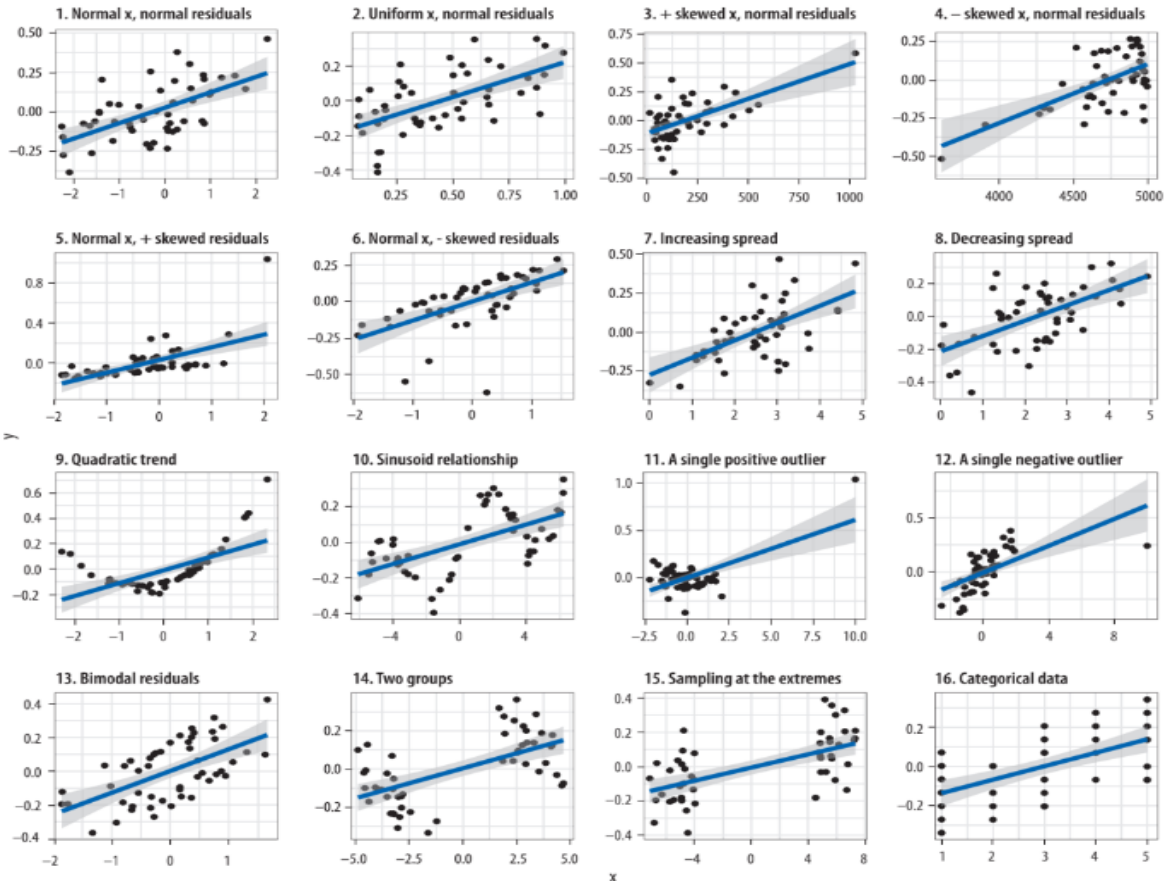


Figure 2.9: What data patterns can lie behind a correlation? The correlation coefficient in all plots is 0.6 (Healy, 2019)

Data visualization encompasses the representation of data through common graphics like charts, plots, infographics, and animations. These visual representations convey complex data relationships and insights in an easily understandable manner. Sharma, Modak, and Sridhar (2019) employed the matplotlib Python package to initially graph the dataset, highlighting hysterical data plotted in scale, featuring the number of days and the opening price for each day.

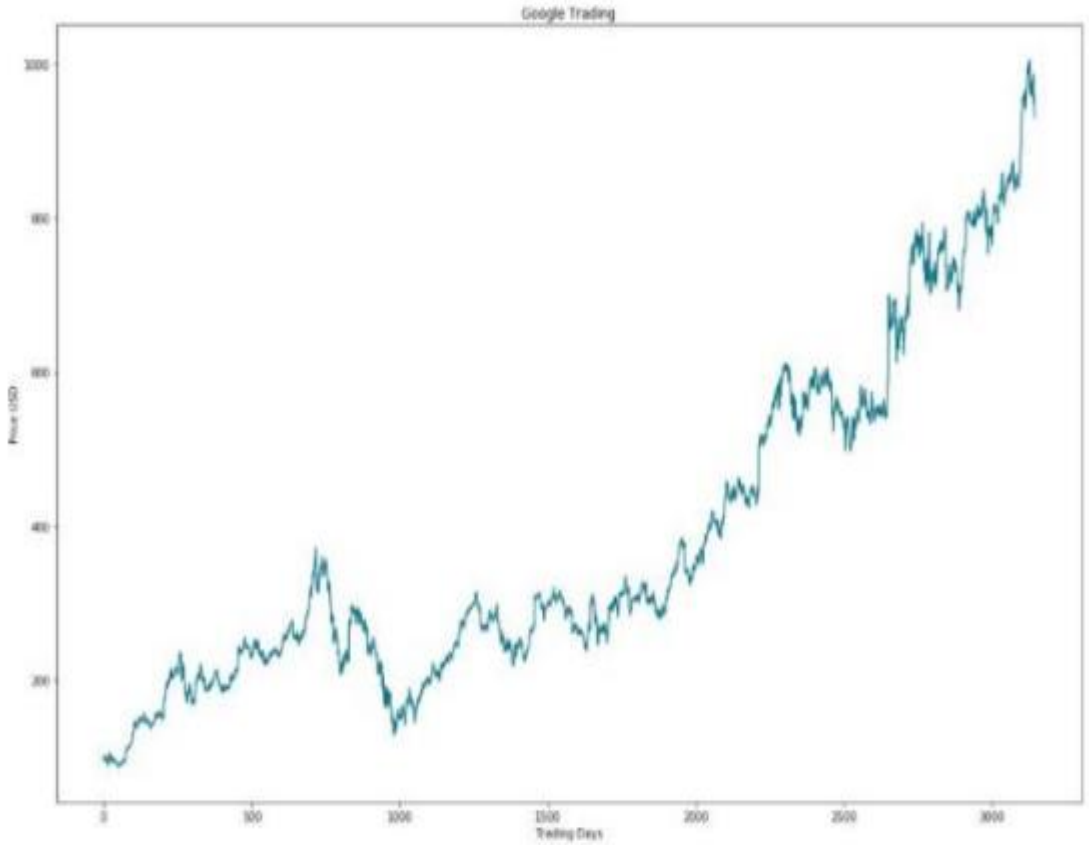


Figure 2.10: Visualization of processed hysterical data fetched from the API (Sharma, Modak, and Sridhar, 2019)

In today's data-driven world, decision-making relies heavily on data, which often arrives with overwhelming speed and volume. Without visualization, deciphering hidden inefficiencies and anomalies within data patterns would be a daunting task. Non-statistical information requires visual representation. Complex systems and business workflows become challenging to understand and improve without visual insight.

Chapter 3. Methodology

In this chapter, the approach, methodologies, and tools employed to execute the project effectively were outlined. It furnishes an in-depth account of how the project was conducted, encompassing the selection of statistical tools, data collection methods, analysis techniques, and interpretation protocols. This chapter ensures the integrity, reliability, and validity of the project's findings.

3.1. Software Environment

The initial step before embarking on any data analysis endeavour is the selection of appropriate software. This decision is crucial as it dictates the efficacy and comprehensiveness of the analysis, ensuring that all project requirements, criteria, and sub-criteria are adequately addressed.

The software selection process holds significant ramifications for the future growth and competitiveness of a business, profoundly impacting project success, operation, and control. Therefore, for a financial modelling project utilizing data mining techniques, it is imperative to choose one or more efficient software solutions capable of covering all facets of the data preparation, exploration, and modelling processes. This includes requisite packages, algorithms, and configurations necessary to conduct accurate analyses, facilitating a clear understanding of the data and informed decision-making.

In the implementation of this project, the following software tools were utilized:

- **Python:** Leveraged for writing automated data preparation scripts and conducting various analytical tasks.
- **Excel:** Utilized for data manipulation, preprocessing, and certain analytical tasks.
- **Power BI:** Employed for data visualization, exploration, and modelling, providing interactive and insightful visual representations of the data.

These software choices were deemed optimal for executing the project objectives, ensuring a comprehensive and effective analysis of the financial data at hand.

3.1.1. Excel for Data Manipulation & Preprocessing

Microsoft Excel, developed by Microsoft, is a versatile spreadsheet editor available for various operating systems including Windows, macOS, Android, iOS, and iPadOS. Initially released on November 19, 1987, Excel offers extensive calculation capabilities, graphing tools, pivot tables, and a macro programming language called Visual Basic for Applications (VBA). It is an integral component of the Microsoft 365 suite of software (Wikipedia, 2024).

In this project, Excel was employed for data manipulation and preprocessing purposes. The software facilitated the calculation of mean and standard deviation for each company, as well as the determination of correlations between the asset returns. These metrics were computed using the Correlation method, providing valuable insights into the performance and interrelationships among the selected assets. Additionally, the Data Analysis and Solver packages within Excel were utilized to compute the correlations between the asset returns, enhancing the analytical capabilities of the software.

3.1.2. Power BI for Data Exploration & Visualisation

Microsoft Power BI, introduced by Microsoft on July 11, 2011, is an interactive data visualization software product primarily focused on business intelligence. As part of the Microsoft Power Platform, Power BI encompasses a suite of software services, applications, and connectors designed to transform diverse data sources into static and interactive data visualizations (Wikipedia, 2024).

In this project, Power BI was utilized for data exploration and visualization purposes. The software facilitated interactive exploration of the data, enabling trend analysis, comparative analysis, correlation analysis, anomaly detection, and forecasting analytics. Through its intuitive interface and strong features, Power BI enhanced the understanding of the dataset and provided valuable insights for further analysis.

3.1.3. Python Programming Language for Data Preparation and Modelling

Python is a versatile, high-level programming language known for its readability and ease of use. Developed by Guido Van Rossum in the late 1980s as a successor to the ABC programming language, Python was first released on February 20, 1991, as Python 0.9.0. Since then, it has evolved with numerous upgraded versions, becoming one of the most widely used programming languages worldwide (Wikipedia, 2024).

In this project, Python was employed for data preparation and modelling tasks, leveraging its extensive libraries and frameworks. The Python ecosystem offers strong tools for data manipulation, transformation, modelling, and evaluation. Key packages utilized in this project include:

- **Pandas:** A powerful library for data manipulation and analysis, providing data structures and functions to efficiently handle structured data.
- **Scikit-learn (Sklearn):** A machine learning library that offers a wide range of algorithms for classification, regression, clustering, dimensionality reduction, and more.
- **NumPy:** A Python library designed to facilitate handling of large, multi-dimensional arrays and matrices. It offers an extensive set of high-level mathematical functions tailored for operations on these arrays.
- **Matplotlib:** A comprehensive plotting library for creating static, interactive, and animated visualizations in Python.
- **Seaborn:** A Python library for data visualization built on top of matplotlib. It offers a user-friendly interface for creating visually appealing and insightful statistical graphics.

Data preparation and exploration are crucial steps in the data analysis process, laying the foundation for accurate modelling and informed decision-making. Various machine learning algorithms and evaluation techniques were employed in this project, including:

- **Random Forest:** A popular ensemble learning method that combines multiple decision trees to improve predictive accuracy and reduce overfitting.

- **Gradient Boosting:** A machine learning method that utilizes boosting in a functional space. Unlike traditional boosting techniques that focus on residuals, gradient boosting operates on pseudo-residuals, leading to enhanced performance.
- **Support Vector Machine (SVM):** Also known as support vector networks, are supervised learning models characterized by their maximum-margin principle.
- **Linear Regression:** A simple yet powerful regression technique used to model the relationship between a dependent variable and one or more independent variables.
- **Autoregressive Integrated Moving Average (ARIMA):** A statistical analysis tool used for forecasting future trends by analysing time series data.
- **Mean Squared Error (MSE):** This quantifies the accuracy of statistical models by calculating the average of the squared differences between observed and predicted values.
- **R-squared score (R^2):** This quantifies the degree of association between your linear model and the dependent variables, expressed on a scale from 0 to 100%.

By leveraging these models and evaluation techniques, the project aimed to extract meaningful insights from the data and develop strong predictive models to support informed decision-making processes. Each model was selected based on its strengths and suitability for the specific characteristics of the dataset, ensuring comprehensive analysis and accurate predictions.

3.2. CRISP-DM Methodology

The success of a data mining project is based on the adoption of an effective operational methodology that can navigate the complexities inherent in the process. Recognizing this need, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is selected as the methodology for this project to ensure its success. CRISP-DM is a comprehensive and widely adopted process model that provides a structured framework for carrying out data mining tasks across various industries and technology sectors, particularly in business intelligence (BI) applications (CRISP, 1999).

The CRISP-DM model offers a systematic approach to guide practitioners through the intricacies of data mining projects, offering practical guidance at each stage of the process. Its generic process model serves as a solid foundation for implementing dedicated process

models tailored to specific project requirements. By combining a flexible implementation of CRISP-DM with overarching project approaches, the project aims to achieve optimal outcomes, enabling investors to make informed decisions.

The success of CRISP-DM can be attributed to its industry, tool, and application neutrality, making it adaptable to diverse contexts. The methodology defined the data mining process into six major phases:

- Business Understanding
- Data Understanding
- Data Preparation
- Modelling
- Evaluation
- Deployment

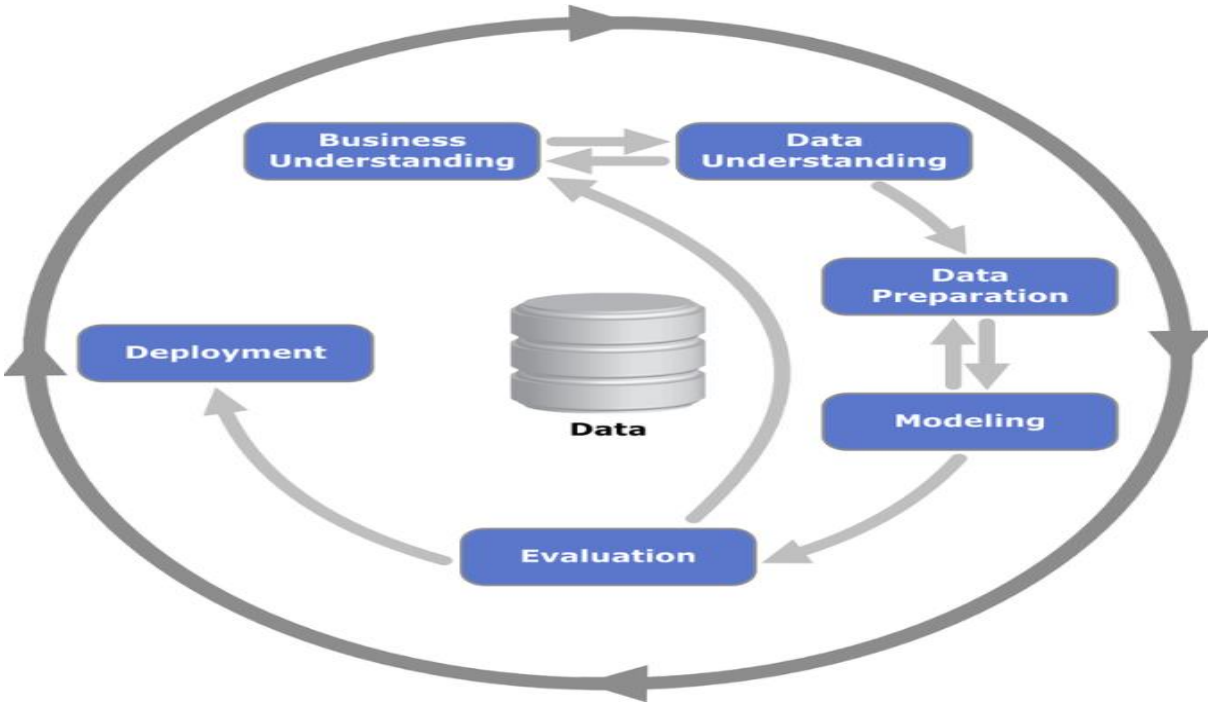


Figure 3.1: process diagram showing the relationship between the different phases of CRISP-DM (Wikipedia, 2024)

3.2.1. Business Understanding

The Business Understanding phase represents the foundational stage of the data mining process, serving as a critical precursor to all subsequent activities. It involves gaining a

comprehensive understanding of the business context, objectives, and challenges before proceeding further.

i. **Determine Business Objectives:** The first step entails clarifying the overarching goals and objectives of the business. This involves engaging stakeholders to identify key priorities, challenges, and opportunities. By defining clear business objectives, the project can be aligned with strategic goals, ensuring that data mining efforts are purpose-driven.

ii. **Assess Situation:** A thorough assessment of the current business environment is essential to gauge the scope, complexity, and feasibility of the data mining project. This involves evaluating factors such as available resources, project requirements, potential risks, and contingencies. Additionally, conducting a cost-benefit analysis helps in understanding the potential return on investment and justifying project investments.

iii. **Determine Data Mining Goals:** In line with business objectives, it is imperative to define specific data mining goals that align with the business context. These goals may include improving operational efficiency, enhancing customer segmentation, predicting market trends, or mitigating risks. Clear articulation of data mining goals ensures that analysis efforts are targeted towards addressing relevant business challenges.

iv. **Produce Project Plan:** Building on the insights gained from the previous steps, a detailed project plan is formulated to guide the execution of the data mining initiative. This involves selecting appropriate technologies, tools, and methodologies to support the project objectives. Additionally, a comprehensive timeline and resource allocation plan are established to ensure smooth project execution and successful analysis outcomes.

By diligently addressing each aspect of the Business Understanding phase, the data mining project lays a solid foundation for subsequent phases, enabling informed decision-making and value-driven insights generation.

3.2.2. Data Understanding

The Data Understanding phase constitutes the second stage of the CRISP-DM methodology, focusing on gaining a comprehensive understanding of the data to be analysed.

i. **Identify Data Sources:** The initial step involves identifying and collecting relevant data sources that contain the information necessary for the analysis. This may include structured datasets, databases, files, or external sources. It is crucial to ensure that the selected data sources align with the objectives of the analysis and provide the requisite information for meaningful insights generation.

ii. **Prepare Data Inventory:** Once the data sources are identified, a detailed inventory of the data is prepared, documenting key attributes such as the source of the data, data format, data quality, and compliance with regulatory requirements such as data protection regulations. This inventory serves as a reference point for understanding the nature and characteristics of the data, facilitating informed decision-making throughout the analysis process.

iii. **Explore Data Characteristics:** In supervised learning analyses, a crucial aspect involves identifying the target variable (response variable) and its predictors (explanatory variables). Additionally, the data type of each variable is determined whether continuous, discrete, ordinal, or nominal. Furthermore, thorough exploration of the data is conducted to identify any missing values, outliers, or anomalies that may impact the analysis outcomes.

iv. **Conduct Statistical Analysis:** Statistical analysis techniques are applied to measure the distributional characteristics of the variables, such as mean, standard deviation, skewness, and kurtosis. This helps in understanding the trend and spread of the data, facilitating the selection of appropriate modelling techniques. Moreover, correlation analysis is performed to examine the relationships between variables, enabling the identification of significant predictors and the removal of irrelevant variables.

By undertaking a systematic exploration and analysis of the data, the Data Understanding phase lays the groundwork for subsequent data preparation and modelling activities, ensuring that the analysis is based on a solid understanding of the underlying data characteristics.

3.2.3. Data Preparation

Data preparation is a critical phase in the data mining process, aimed at refining and structuring the dataset to ensure its suitability for analysis. This phase involves several key practices tailored to address the specific characteristics and requirements of the dataset under investigation. For this project, the data preparation process encompasses three main steps:

3.2.3.1. Data Cleaning

Data cleaning involves identifying and rectifying inconsistencies, missing values, and incorrect information within the dataset. This process is essential for enhancing data quality and ensuring the accuracy of subsequent analyses. Missing values are identified and either replaced with correct data or removed from the dataset, depending on the extent of their impact. Additionally, irrelevant variables that have minimal influence on the target variables are eliminated to streamline the analysis process.

3.2.3.2. Data Transformation

Data transformation entails converting the raw data into a structured format that is conducive to analysis and decision-making. This step is particularly crucial when the data needs to be standardized or formatted to align with the requirements of the analysis tools or models being employed. In this project, for example, the daily prices were transformed into daily returns to facilitate further analysis and modelling.

3.2.3.3. Data Quality Check

The final step in the data preparation process involves conducting a comprehensive quality check to ensure that the prepared data meets the desired standards and integrity criteria. Various techniques, such as heat-map visualization, are utilized to assess the quality of the data and verify that all inconsistencies have been addressed effectively. This ensures that the dataset is strong and reliable for subsequent analysis stages.

By systematically performing these data preparation steps, the dataset is refined and optimized for analysis, laying a solid foundation for the modelling phase and enhancing the overall validity and reliability of the project findings.

3.2.4. Data Modelling

Selecting an appropriate model is a crucial aspect of the data mining process, as it directly impacts the accuracy and effectiveness of the analysis. With a great number of modelling techniques available, choosing the right model depends on various factors, including the nature of the dataset and the objectives of the analysis. In this phase, the following steps were undertaken to develop and assess the models:

3.2.4.1. Data Partition

Before proceeding with model development, it is essential to partition the dataset into distinct subsets. This approach enhances scalability, minimizes contention, and optimizes model performance. The dataset was partitioned into three subsets: training, validation, and test datasets. The cross-validation method was employed, employing simple random sampling to allocate 50% of the data for training, 25% for validation, and the remaining 25% for testing.

3.2.4.2. Model Selection

The selection of an appropriate statistical model is guided by factors such as the type of data, the objectives of the analysis, and specific modelling requirements. For this project, three statistical models—linear regression, random forest, and neural network—were chosen based on their ability to explore the data and predict future outcomes effectively.

3.2.4.3. Models Building & Assessment

The core of the data modelling phase involves training and fitting the selected models to the dataset. This process entails adjusting the parameters of each model to identify significant patterns and relationships for accurate predictions. Python programming language was utilized to build and configure the models, leveraging various data mining tools and packages. Once the models were built, they were executed to generate actionable insights and predictions, which were then evaluated to assess their performance and suitability for deployment.

By following these steps, the data modelling phase ensures the development of strong and effective models that contribute to informed decision-making and actionable insights.

3.2.5. Model Evaluation

Model evaluation is a critical step in assessing the effectiveness and suitability of the developed models for data mining purposes. This phase involves comparing the performance of the fitted models using various statistical and visualization methods, with a focus on metrics such as R-squared for this project.

Once the models have been built and verified to be technically sound from a business perspective, it becomes imperative to evaluate their performance relative to the defined business objectives. The following key questions are addressed during the evaluation process:

- Does the model align with the business objectives?
- Have all significant business considerations been adequately addressed?
- Is the model statistically meaningful and strong?
- Are the insights derived from the model actionable and can they be effectively implemented to drive business decisions?

By answering these questions, the evaluation phase ensures that the selected model not only meets the technical requirements but also effectively addresses the business needs and can be seamlessly integrated into the decision-making processes of the organization.

3.2.6. Deployment

The deployment phase marks the final step in the CRISP-DM methodology. It involves the implementation of strategic actions based on the predicted outcomes derived from the final selected model to address future events or business decisions.

To ensure an effective implementation aligned with the CRISP-DM methodology, the following steps are crucial:

i. **Determine a deployment plan:** Develop a comprehensive plan outlining the steps to implement the insights and recommendations generated by the selected model. This plan should detail the specific actions to be taken, responsibilities assigned, and timelines established for deployment.

ii. **Monitoring and maintenance:** Once the deployment plan is executed, ongoing monitoring and maintenance are essential to ensure that the implemented strategies continue to yield the desired outcomes. Regular monitoring helps identify any deviations from expected results and allows for timely adjustments to be made.

iii. **Periodic review:** It is imperative to periodically review the deployment plan and its effectiveness in achieving the desired objectives. This review process involves evaluating the performance of the implemented strategies, identifying areas for improvement, and making necessary refinements to enhance overall efficiency and effectiveness.

Business Understanding	Data Understanding	Data Preparation	Modelling	Evaluation	Deployment
Determine Business Objectives *Background* *Business Objectives* *Business Success Criteria* **Assess Situation** *Inventory of Resources* *Requirements, Assumptions, and Constraints* *Risks and Contingencies* *Terminology* *Costs and Benefits* **Determine Data Mining Goals** *Data Mining Goals* *Data Mining Success Criteria* **Produce Project Plan** *Project Plan* *Initial Assessment of Tools and Techniques*	**Collect Initial Data** *Initial Data Collection Report* **Describe Data** *Data Description Report* **Explore Data** *Data Exploration Report* **Verify Data Quality** *Data Quality Report*	*Data Set* *Data Set Description* **Select Data** *Rationale for Inclusion /Exclusion* **Clean Data** *Data Cleaning Report* **Construct Data** *Derived Attributes* *Generated Records* **Integrate Data** *Merged Data* **Format Data** *Reformatted Data*	**Select Modelling Technique** *Modelling Technique* *Modelling Assumptions* **Generate Test Design** *Test Design* **Build Model** *Parameter Settings* *Models* *Model Description* **Assess Model** *Model Assessment* *Revised Parameter Settings*	**Evaluate Results** *Assessment of Data Mining Results w.r.t. Business Success Criteria* *Approved Models* **Review Process** *Review of Process* **Determine Next Steps** *List of Possible Actions* *Decision*	**Plan Deployment** *Deployment Plan* **Plan Monitoring and Maintenance** *Monitoring and Maintenance Plan* **Produce Final Report** *Final Report* *Final Presentation* **Review Project** *Experience Documentation*

Table 3.1: overview of the CRISP-DM phases and tasks (Solis, 2023).

By adhering to these steps, organizations can ensure that the insights and recommendations derived from the data mining process are effectively translated into actionable decisions that drive positive outcomes. Additionally, the iterative nature of deployment ensures that the deployed strategies remain adaptive to changing business conditions and requirements.

Chapter 4. Analysis & Design

This section establishes the foundation for the project's successful execution by fostering a comprehensive understanding of requirements, outlining the solution approach, and pinpointing potential challenges and opportunities. It paves the way for efficient implementation, ensuring alignment with investors' expectations and interests.

4.1. Data Understanding & Exploration

In this analysis phase, a thorough data inventory was conducted first, documenting all pertinent details regarding the observed datasets, including its source, size, format, and the number of historical records and variables. Subsequently, the exploratory phase, where visual analyses were conducted to uncover meaningful insights. These visualisations aided in gaining a deeper understanding of the data characteristics and the relationship between independent variables and the target variable.

4.1.1. Data Description

Investors often face challenges in making informed investment decisions due to a lack of comprehensive financial information. This project aims to address this issue by providing a wide range of financial indicators and stock market prices, enabling investors to make well-informed decisions. Using the CRISP-DM framework and machine learning techniques, the project focuses on selecting 12 companies from various sectors listed on the London Stock Exchange (LSE).

The purposive sampling method was utilized to select the companies for this project. Specifically, companies listed on the Financial Times Stock Exchange (FTSE250) were chosen. This selection was made because most of these companies operate within the UK, providing a focused perspective. This approach differs from the FTSE100, where companies listed primarily operate outside of the UK. Additionally, various sectors were considered to ensure diversity and reduce bias. As a result, 12 companies were chosen to represent different sectors as comprehensively as possible. Table 4.1 below outlines the details of the selected companies and their respective industrial classifications.

No	Company	Ticker	FTSE Industry Classification
1	4imprint	FOUR	Media
2	A.G. Barr	BAG	Beverages
3	AO World	AO	Retail
4	Ashmore Group	ASHM	Financial Services
5	Babcock International	BAB	Support Services
6	Balfour Beatty	BBY	Construction & Materials
7	Bodycote	BOY	Industrial Engineering
8	British Land	BLND	Real Estate
9	Man Group	EMG	Investment Trust
10	Pennon Group	PNN	Gas, Water & Multiutilities
11	PPHE Hotel Group	PPH	Travel and Leisure
12	PZ Cussons	PZC	Personal Goods

Table 4.1: list of companies selected from FTSE250 for the project.

The analysis involved several CSV files containing data from selected companies covering the years 2017 to 2019 and 2021 to 2023. Notably, the dataset for the year 2020 was omitted due to the disruptive effects of the COVID-19 pandemic. Consequently, the analysis was divided into two distinct periods: pre-COVID-19 and post-COVID-19. The ticker in Table 4.1 corresponds to the abbreviation by which each company is commonly identified.

Each pre COVID-19 dataset comprised 757 observations while post COVID-19 dataset each comprised 754 observations, featuring six explanatory variables and one response variable (target). In total, 24 datasets were utilized for the analysis. These datasets were sourced from ***Yahoo! Finance*** (https://uk.finance.yahoo.com/), a widely used platform known for its real-time financial news, stock market data, and other relevant information, catering primarily to financial analysts and investors.

The table below presents a comprehensive overview of all attributes observed in the datasets, along with their respective descriptions.

Variable	Description
Date	Date on which the financial data was recorded
Open	Opening price of the stock at the beginning of the trading period
High	Highest price at which the stock traded during the trading period
Low	Lowest price at which the stock traded during the trading period
Close	Closing price of the stock at the end of the trading period
Adj Close	Adjusted closing price of the stock which accounts for any corporate actions such as dividends, stock splits, etc., to provide a more accurate reflection of its value
Volume	Total number of shares or units traded during the trading period

Table 4.2: variables description

Table 4.2 above provides descriptions of all the variables present in the dataset for the selected companies. The data is recorded daily, reflecting the trading period.

4.1.2. Data Manipulation & Preprocessing

The initial phase of data manipulation and preprocessing involved consolidating the adjusted close prices of the 12 selected companies before and after the Covid-19 pandemic using an Excel spreadsheet, as illustrated in Table 4.3a and 4.3b.

DAILY STOCK PRICES (PRE COVID-19)												
DATE	ASHM	FOUR	EMG	AO	PZC	BAB	BBY	PNN	BAG	PPH	BOY	BLND
30/12/2019	392.858	3243.27	127.897	91.8	183.69	620.376	242.744	1021.75	543.761	1827.46	848.547	530.083
27/12/2019	394.744	3271.31	129.596	95.3	177.207	626.151	247.537	1020.26	540.023	1837.18	852.966	525.082
24/12/2019	392.858	3159.15	129.192	98	175.911	615.994	246.984	1008.83	531.615	1808.02	858.269	516.748
23/12/2019	391.35	3121.76	127.533	90.4	175.046	615.397	246.615	1010.32	533.483	1837.18	837.498	514.581
20/12/2019	386.826	2962.87	128.463	93	160.956	604.244	242.376	1018.77	533.483	1856.62	828.217	503.412
19/12/2019	382.302	2888.1	127.209	91.3	156.115	618.185	246.062	1015.29	528.812	1856.62	834.404	511.247
18/12/2019	381.171	2869.4	128.544	92	157.153	625.355	246.247	1035.16	530.68	1846.9	832.636	502.412
17/12/2019	378.154	2897.44	127.007	92	154.559	623.363	244.035	1034.66	540.023	1876.06	845.011	503.245
16/12/2019	379.663	2869.4	128.949	88.5	167.871	638.3	249.196	1024.73	541.892	1876.06	845.011	525.916
13/12/2019	370.991	2860.06	126.926	85	169.946	628.741	234.45	977.939	537.221	1885.78	832.194	518.581
12/12/2019	365.713	2841.36	121.059	78.2	166.315	594.884	221.548	895.086	521.338	1808.02	795.512	498.578
11/12/2019	359.529	2869.4	120.129	84.3	167.526	593.29	212.332	920.916	518.535	1827.46	780.044	501.078
10/12/2019	357.871	2897.44	121.787	84.8	172.885	602.452	212.517	919.326	527.878	1846.9	786.673	503.579
09/12/2019	359.68	2897.44	123.568	88	173.317	602.053	212.517	928.068	526.943	1846.9	762.366	507.58
06/12/2019	358.926	2888.1	122.92	88.7	172.885	593.091	211.964	926.081	513.863	1846.9	772.531	505.913
05/12/2019	360.284	2897.44	121.302	88.7	173.75	577.557	207.171	909.988	515.732	1846.9	760.156	496.911
04/12/2019	361.641	2906.79	122.394	88.7	173.317	570.188	204.775	924.492	531.615	1846.9	767.227	488.076
03/12/2019	362.546	2897.44	119.441	85	175.046	556.42	201.458	909.789	527.878	1846.9	759.272	476.408

Table 4.3a: screenshot of the pre covid-19 adjusted close price of the 12 companies.

DAILY STOCK PRICES (POST COVID-19)												
DATE	ASHM	FOUR	EMG	AO	PZC	BAB	BBY	PNN	BAG	PPH	BOY	BLND
29/12/2023	217.707	4570	232.6	98.35	150.836	395	331.2	736.629	513	1200	594.5	399.6
28/12/2023	218.098	4610	231.8	100	150.048	397.2	335.4	740.06	515	1160	597	405.8
27/12/2023	217.512	4670	234.1	99.4	148.079	398.4	338.2	742.02	520	1145	606	407.9
22/12/2023	214.19	4690	234.3	97.4	150.048	396.6	337	733.688	517	1130	603	406
21/12/2023	212.626	4595	233.6	98.4	151.82	395	334.4	723.886	520	1140	603	405.3
20/12/2023	215.558	4640	232.3	99.95	152.608	400.4	330.2	740.55	514	1145	612	409.8
19/12/2023	209.499	4605	231	97.25	152.017	396.8	325.2	727.317	508	1150	602.5	405.6
18/12/2023	206.959	4535	231	95	150.836	394	325.4	715.554	506	1165	599	408.5
15/12/2023	207.936	4645	228.1	93.1	153.986	385	325.6	718.985	509	1205	600	409.3
14/12/2023	206.177	4650	226.3	92.25	147.685	397.6	326	737.609	507	1270	597.5	413.1
13/12/2023	192.204	4635	222.4	87.15	143.944	388.8	323.2	726.827	485	1260	572	387.7
12/12/2023	191.031	4550	216.6	87.75	138.824	390	327	719.475	487	1290	566	384.2
11/12/2023	192.497	4530	219.9	88.9	141.187	399.4	330.4	735.649	490	1290	574.5	387.9
08/12/2023	192.106	4550	218.7	89.5	145.913	399.2	335.2	723.886	487.5	1250	580	376.2
07/12/2023	191.422	4600	220	88.3	144.928	398	335.4	730.748	483	1255	573	375.1
06/12/2023	192.595	4615	216.3	88.15	142.369	397.8	338.2	731.728	484.5	1270	581	376.4
05/12/2023	183.703	4500	216	89	144.535	395.4	331.8	708.203	487.5	1250	585.5	369.9
04/12/2023	178.817	4320	213.2	90.45	146.701	392.6	329.8	699.381	483	1265	578.5	361

Table 4.3b: screenshot of the post covid-19 adjusted close price of the 12 companies.

Subsequently, the daily stock prices were transformed into daily return prices by implementing the formula:

$$Daily\ Returns = \frac{\mathrm{Pt} - \mathrm{Pt} - 1}{\mathrm{Pt} - 1}$$

DAILY RETURNS (PRE COVID-19)												
DATE	ASHM	FOUR	EMG	AO	PZC	BAB	BBY	PNN	BAG	PPH	BOY	BLND
30/12/2019	-0.0048	-0.0086	-0.0131	-0.0367	0.03659	-0.0092	-0.0194	0.00146	0.00692	-0.0053	-0.0052	0.00952
27/12/2019	0.0048	0.0355	0.00313	-0.0276	0.00737	0.01649	0.00224	0.01132	0.01582	0.01613	-0.0062	0.01613
24/12/2019	0.00385	0.01198	0.01301	0.08407	0.00494	0.00097	0.00149	-0.0015	-0.0035	-0.0159	0.0248	0.00421
23/12/2019	0.0117	0.05363	-0.0072	-0.028	0.08754	0.01846	0.01749	-0.0083	0	-0.0105	0.01121	0.02219
20/12/2019	0.01183	0.02589	0.00986	0.01862	0.03101	-0.0226	-0.015	0.00342	0.00883	0	-0.0074	-0.0153
19/12/2019	0.00297	0.00651	-0.0104	-0.0076	-0.0066	-0.0115	-0.0007	-0.0192	-0.0035	0.00526	0.00212	0.01758
18/12/2019	0.00798	-0.0097	0.01211	0	0.01678	0.00319	0.00906	0.00048	-0.0173	-0.0155	-0.0146	-0.0017
17/12/2019	-0.004	0.00977	-0.0151	0.03955	-0.0793	-0.0234	-0.0207	0.00969	-0.0034	0	0	-0.0431
16/12/2019	0.02337	0.00327	0.01594	0.04118	-0.0122	0.0152	0.06289	0.04785	0.0087	-0.0052	0.0154	0.01414
13/12/2019	0.01443	0.00658	0.04846	0.08696	0.02183	0.05691	0.05824	0.09256	0.03047	0.04301	0.04611	0.04012
12/12/2019	0.0172	-0.0098	0.00775	-0.0724	-0.0072	0.00269	0.0434	-0.028	0.00541	-0.0106	0.01983	-0.005
11/12/2019	0.00464	-0.0097	-0.0136	-0.0059	-0.031	-0.0152	-0.0009	0.00173	-0.0177	-0.0105	-0.0084	-0.005
10/12/2019	-0.005	0	-0.0144	-0.0364	-0.0025	0.00066	0	-0.0094	0.00177	0	0.03188	-0.0079
09/12/2019	0.0021	0.00324	0.00527	-0.0079	0.0025	0.01511	0.00261	0.00215	0.02545	0	-0.0132	0.00329
06/12/2019	-0.0038	-0.0032	0.01334	0	-0.005	0.0269	0.02313	0.01769	-0.0036	0	0.01628	0.01811
05/12/2019	-0.0038	-0.0032	-0.0089	0	0.00249	0.01292	0.0117	-0.0157	-0.0299	0	-0.0092	0.0181
04/12/2019	-0.0025	0.00323	0.02473	0.04353	-0.0099	0.02474	0.01647	0.01616	0.00708	0	0.01048	0.02449
03/12/2019	-0.0025	-0.0032	-0.0189	0	0.00248	-0.0153	-0.0109	0.00505	0	0.00529	-0.0029	0.00918

Table 4.4a: daily return prices of the pre covid-19 adjusted close price of the 12 companies.

DAILY RETURNS (POST COVID-19)												
DATE	ASHM	FOUR	EMG	AO	PZC	BAB	BBY	PNN	BAG	PPH	BOY	BLND
29/12/2023	-0.0018	-0.0087	0.00345	-0.0165	0.00525	-0.0055	-0.0125	-0.0046	-0.0039	0.03448	-0.0042	-0.0153
28/12/2023	0.0027	-0.0128	-0.0098	0.00604	0.0133	-0.003	-0.0083	-0.0026	-0.0096	0.0131	-0.0149	-0.0051
27/12/2023	0.01551	-0.0043	-0.0009	0.02053	-0.0131	0.00454	0.00356	0.01136	0.0058	0.01327	0.00498	0.00468
22/12/2023	0.00735	0.02067	0.003	-0.0102	-0.0117	0.00405	0.00778	0.01354	-0.0058	-0.0088	0	0.00173
21/12/2023	-0.0136	-0.0097	0.0056	-0.0155	-0.0052	-0.0135	0.01272	-0.0225	0.01167	-0.0044	-0.0147	-0.011
20/12/2023	0.02892	0.0076	0.00563	0.02776	0.00389	0.00907	0.01538	0.01819	0.01181	-0.0043	0.01577	0.01035
19/12/2023	0.01228	0.01544	0	0.02368	0.00783	0.00711	-0.0006	0.01644	0.00395	-0.0129	0.00584	-0.0071
18/12/2023	-0.0047	-0.0237	0.01271	0.02041	-0.0205	0.02338	-0.0006	-0.0048	-0.0059	-0.0332	-0.0017	-0.002
15/12/2023	0.00853	-0.0011	0.00795	0.00921	0.04267	-0.0317	-0.0012	-0.0252	0.00394	-0.0512	0.00418	-0.0092
14/12/2023	0.0727	0.00324	0.01754	0.05852	0.02599	0.02263	0.00866	0.01483	0.04536	0.00794	0.04458	0.06551
13/12/2023	0.00614	0.01868	0.02678	-0.0068	0.03688	-0.0031	-0.0116	0.01022	-0.0041	-0.0233	0.0106	0.00911
12/12/2023	-0.0076	0.00442	-0.015	-0.0129	-0.0167	-0.0235	-0.0103	-0.022	-0.0061	0	-0.0148	-0.0095
11/12/2023	0.00203	-0.0044	0.00549	-0.0067	-0.0324	0.0005	-0.0143	0.01625	0.00513	0.032	-0.0095	0.0311
08/12/2023	0.00357	-0.0109	-0.0059	0.01359	0.00679	0.00302	-0.0006	-0.0094	0.00932	-0.004	0.01222	0.00293
07/12/2023	-0.0061	-0.0033	0.01711	0.0017	0.01798	0.0005	-0.0083	-0.0013	-0.0031	-0.0118	-0.0138	-0.0035
06/12/2023	0.0484	0.02556	0.00139	-0.0096	-0.015	0.00607	0.01929	0.03322	-0.0062	0.016	-0.0077	0.01757
05/12/2023	0.02732	0.04167	0.01313	-0.016	-0.0148	0.00713	0.00606	0.01261	0.00932	-0.0119	0.0121	0.02465
04/12/2023	0.02292	-0.0035	0.00471	-0.0449	0.01223	-0.016	-0.0054	0.00991	-0.0021	-0.0156	-0.0261	0.00028

Table 4.4b: daily return prices of the post covid-19 adjusted close price of the 12 companies.

This conversion to daily return prices facilitated a normalized representation of price movements across the selected stocks, enabling comparisons irrespective of their price levels. Daily returns are crucial for assessing stock volatility, risk evaluation, and portfolio management. Higher daily returns signify greater volatility, which may indicate higher risk, and vice versa. They aid in assessing the risk associated with holding specific stocks and evaluating their performance over time. Positive daily returns denote gains, while negative daily returns indicate losses. Furthermore, daily returns are instrumental in forecasting future price movements and trends, facilitating the identification of patterns and trends to predict future returns.

The computation of average and variance daily returns across the companies provided insights into changes in stock market behaviour during and after the pandemic, as showed in Table 4.5a and 4.5b.

	Pre Covid-19											
	ASHM	FOUR	EMG	AO	PZC	BAB	BBY	PNN	BAG	PPH	BOY	BLND
Average daily return	0.0010778	0.00115	0.00068	-0.0005	-0.0004	-0.0002237	0.00014	0.00058	0.00041	0.00148	0.00076	0.00029
Variance(daily)	0.0002274	0.00035	0.00025	0.00089	0.00026	0.000316	0.00024	0.00019	0.00027	0.00011	0.00027	0.00014
Average annual return	0.2715989	0.28967	0.17251	-0.1175	-0.0925	-0.0563838	0.03607	0.14636	0.10322	0.37317	0.19214	0.07302
Annual variance	0.057299	0.0891	0.06416	0.22425	0.06478	0.0796443	0.05978	0.04848	0.06842	0.02763	0.06845	0.03578

Table 4.5a: average and variance daily and annual returns of the selected companies before covid-19.

	Post Covid-19											
	ASHM	FOUR	EMG	AO	PZC	BAB	BBY	PNN	BAG	PPH	BOY	BLND
Average daily return	-0.000382	0.00123	0.00105	-0.0013	-0.0003	0.0008489	0.00048	-0.0002	0.00019	7.4E-05	-2E-05	0.00018
Variance(daily)	0.0004339	0.00066	0.00038	0.00126	0.00024	0.0006416	0.00022	0.00029	0.00024	0.00043	0.00035	0.00037
Average annual return	-0.096226	0.31102	0.26426	-0.3214	-0.0796	0.2139222	0.12153	-0.0605	0.04739	0.01874	-0.0054	0.04537
Annual variance	0.1093417	0.16584	0.09514	0.31728	0.05973	0.1616764	0.05455	0.07352	0.0601	0.10734	0.08882	0.09433

Table 4.5b: average and variance daily and annual returns of the selected companies after covid-19.

Pre-Covid-19, most companies exhibited positive average daily returns, signifying overall growth in stock prices. Notable companies with higher average daily returns included ASHM, FOUR, PPH, and BOY. However, post-Covid-19, average daily returns varied among companies, with some showing positive returns and others negative. Notable companies with higher average daily returns post-Covid-19 included FOUR and EMG.

Pre-Covid-19, variance in daily returns was relatively low for most companies, indicating stability in stock price movements. However, companies like AO and BAB exhibited higher variance. In contrast, post-Covid-19, increased variance across many companies suggested higher volatility in stock price movements, with notable increases in variance for companies like AO, BAB, and PNN.

The annualization of average and variance daily returns provided insights into long-term trends. Despite positive average daily returns pre-Covid-19, some companies experienced negative average annual returns, indicating potential long-term challenges or market corrections. Post-Covid-19, average annual returns exhibited mixed trends, with some companies experiencing improvements while others showed declines. Similarly, annual variance increased post-Covid-19, reflecting heightened uncertainty and volatility in the market during the pandemic.

The correlation matrices of selected companies before and after the Covid-19 pandemic were computed to assess changes in market dynamics, as presented in Table 4.6a and 4.6b.

				Correlation (Pre Covid-19)								
	ASHM	*FOUR*	*EMG*	*AO*	*PZC*	*BAB*	*BBY*	*PNN*	*BAG*	*PPH*	*BOY*	*BLND*
ASHM	1											
FOUR	0.0988	1										
EMG	0.41321	0.05487	1									
AO	0.01291	0.01777	0.06664	1								
PZC	0.08475	0.084	0.11612	0.0162	1							
BAB	0.18569	0.07041	0.22708	0.0424	0.0900719	1						
BBY	0.27636	0.10727	0.30352	0.04796	0.0638506	0.30981	1					
PNN	0.12078	0.03996	0.09945	0.0476	0.12427	0.18643	0.17258	1				
BAG	0.04546	0.03866	0.06911	0.00257	0.0588908	0.05981	0.06941	0.14102	1			
PPH	0.09898	0.00665	0.07282	0.00704	-0.0398083	0.03266	0.03803	0.05868	-0.036	1		
BOY	0.26164	0.03353	0.31109	0.12811	0.1235513	0.31741	0.38671	0.09358	0.10624	0.08911	1	
BLND	0.27885	0.13712	0.2499	0.06287	0.137922	0.32311	0.39793	0.35531	0.11618	0.09369	0.25873	1

Table 4.6a: showing the pre covid-19 correlation analysis for the selected companies.

				Correlation (Post Covid-19)								
	ASHM	*FOUR*	*EMG*	*AO*	*PZC*	*BAB*	*BBY*	*PNN*	*BAG*	*PPH*	*BOY*	*BLND*
ASHM	1											
FOUR	0.28916	1										
EMG	0.52216	0.26801	1									
AO	0.27031	0.17066	0.27253	1								
PZC	0.24031	0.16271	0.20814	0.13243	1							
BAB	0.18472	0.15546	0.1841	0.11099	0.1464796	1						
BBY	0.34742	0.26493	0.36325	0.11545	0.2538249	0.29283	1					
PNN	0.25405	0.07117	0.23021	0.06208	0.1904862	0.04454	0.18587	1				
BAG	0.12678	0.14887	0.11159	0.06079	0.1767443	0.10455	0.17478	0.18777	1			
PPH	0.13786	0.12358	0.15111	0.10641	0.0644002	0.04218	0.09236	0.06354	0.034	1		
BOY	0.50473	0.29683	0.45484	0.3207	0.2662746	0.24491	0.40586	0.2146	0.17379	0.12951	1	
BLND	0.49012	0.2408	0.4198	0.14446	0.2308083	0.26103	0.40565	0.36498	0.20359	0.13954	0.48065	1

Table 4.6b: showing the post covid-19 correlation analysis for the selected companies.

Pre-Covid-19, moderate to strong positive correlations existed between several pairs of companies, while some pairs showed weak or negative correlations. Post-Covid-19, correlations between certain pairs strengthened significantly, suggesting pandemic-induced shifts in market dynamics, while others weakened, reflecting changes in investor sentiment or market conditions. Understanding these changes in correlation patterns is essential for effective portfolio diversification and risk management in volatile market conditions.

4.1.3. Exploratory Analysis & Visualization

In this section, Power BI, a strong visual analysis tool was utilized to delve into the dataset and gain insights into the trading patterns of selected companies before and after the COVID-19 pandemic. The aim was to gain a comprehensive understanding of the fluctuations and trends experienced by these companies during and after the pandemic, shedding light on their performance in response to the crisis.

For the exploratory analysis and visualization phase, the EMG company was randomly selected for the analysis, without any specific reason, considering that discussing all 12 selected companies would be impractical within this project. However, comprehensive analysis of all companies is available on Power BI and can be provided upon request. Additionally, other companies will be used to exemplify investments suitable for both risk-tolerant and risk-averse investors. Further analyses will be presented in the appendix.

The exploration phase began with the utilization of a Line chart, also known as a line graph. This visualization method effectively illustrates data points connected by straight line segments, offering a clear depiction of trends and patterns over time.

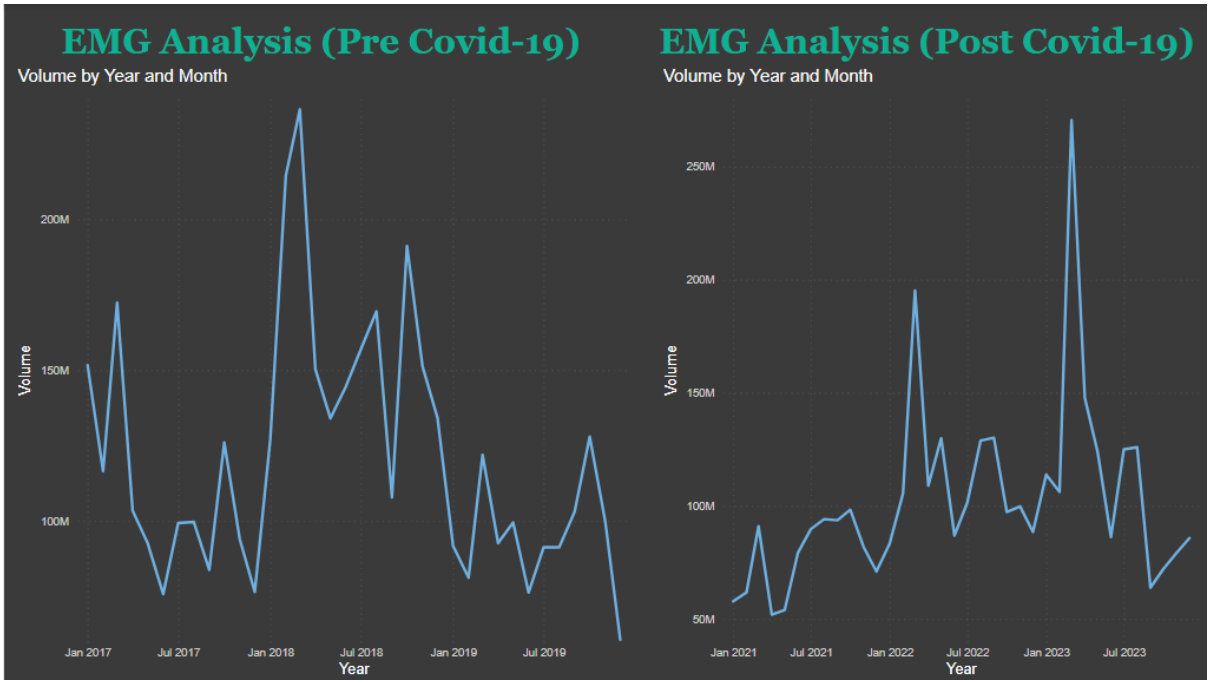


Figure 4.1: showing volume shares traded by EMG pre and post covid-19.

As showed in Figure 4.1, the focus was on analysing the volume shares traded by EMG (Man Group), an investment trust company, both pre and post COVID-19. The analysis revealed fluctuating volume trends before and after the pandemic, with a slight increase in trading volume observed post-pandemic compared to the pre-pandemic period. Notably, there appeared to be consistent spikes in trading volume between February and March, followed by declines between May and June, both before and after the pandemic. This visualization provided valuable insights into the trading behaviour of EMG and hints at potential market dynamics influencing trading volumes during different periods.

For the subsequent exploration, the clustered column chart was employed, a powerful visualization tool ideal for comparing multiple datasets simultaneously. This chart type allows for easy comparison between different categories or groups of data.

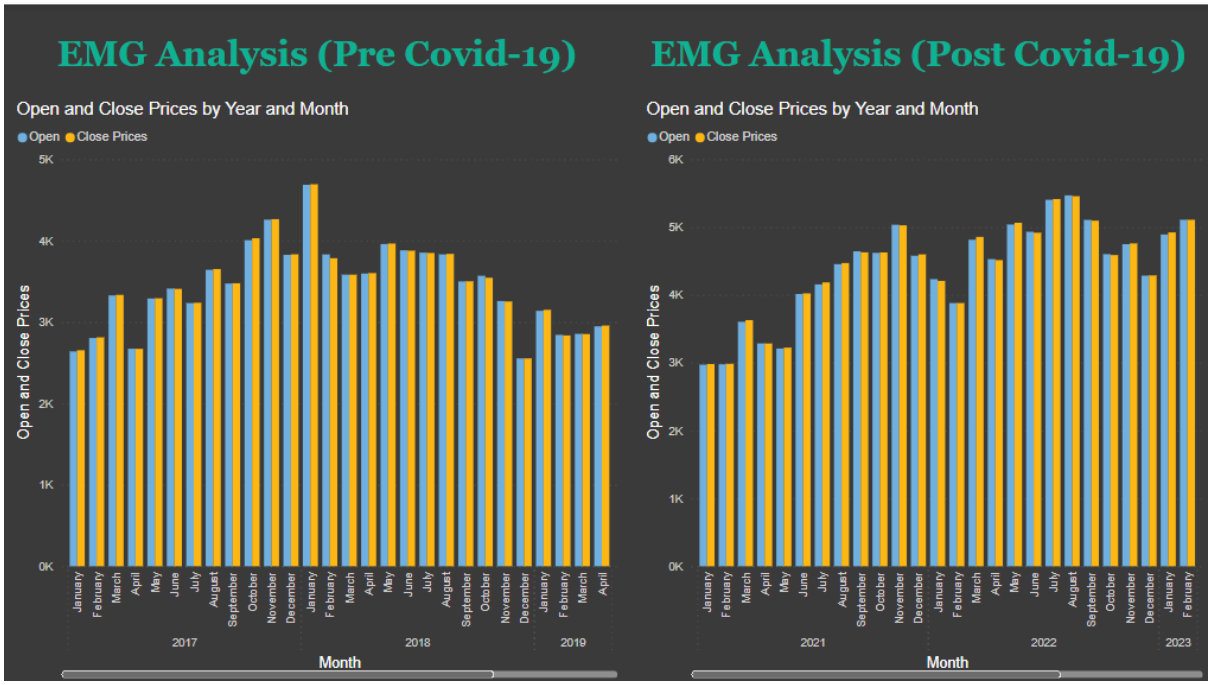


Figure 4.2: open and close prices of EMG before and after the pandemic.

As illustrated in Figure 4.2, the clustered column chart was utilized to compare the open and close prices of EMG (Man Group) both before and after the COVID-19 pandemic. The analysis revealed that the open and close prices of EMG remained relatively consistent both before and after the pandemic, with minor variations observed between the two periods. In some instances, the close prices were slightly higher than the open prices, while in others, the open prices appeared marginally higher. Additionally, it is noteworthy that EMG traded at higher prices post-COVID-19 compared to the pre-pandemic period. This visualization provided a clear comparison of the price dynamics of EMG before and after the pandemic, highlighting any notable shifts or trends in open and close prices.

To visualize the high and low prices of EMG before and after the pandemic, the stacked area chart was employed. This chart type is particularly useful for illustrating changes in composition over time and visualizing cumulative totals of multiple datasets.

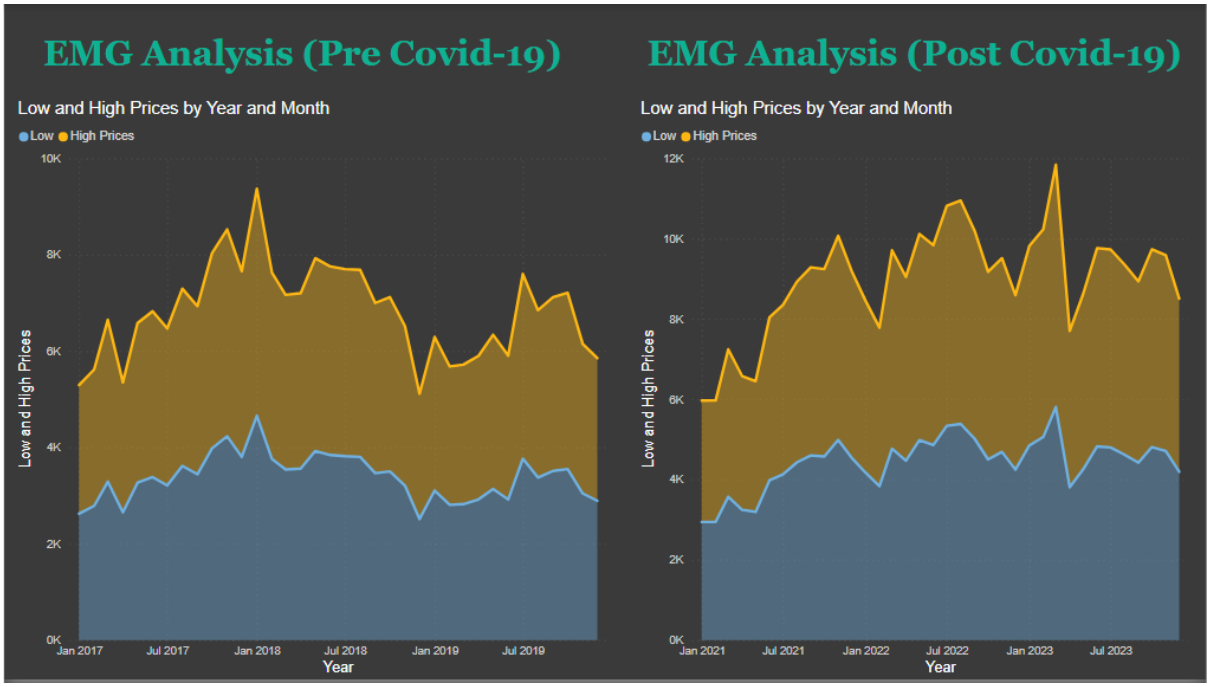


Figure 4.3: comparing EMG's low and high prices pre and post covid-19.

In Figure 4.3, a comparison of EMG's low and high prices before and after the COVID-19 pandemic using the stacked area chart was presented. The analysis revealed that the cumulative high prices of EMG significantly outweigh the cumulative low prices for both the pre and post-pandemic periods. This observation suggests that EMG generally experienced higher highs in its price fluctuations compared to its lows during these time frames. Additionally, it is evident that EMG traded at higher prices post-pandemic compared to the pre-pandemic period. The stacked area chart provided a comprehensive visual representation of EMG's price dynamics, allowing for easy comparison between high and low prices before and after the pandemic. This visualization aided in understanding the overall trend and volatility in EMG's price movements over time, thereby informing investment decisions and risk assessment strategies.

The financial indicators for all 12 selected companies were consolidated for further exploration in Power BI, where a variety of charts were utilized to delve into each company's performance across different metrics. This culminated in the creation of an interactive dashboard, as illustrated in Figure 4.4.

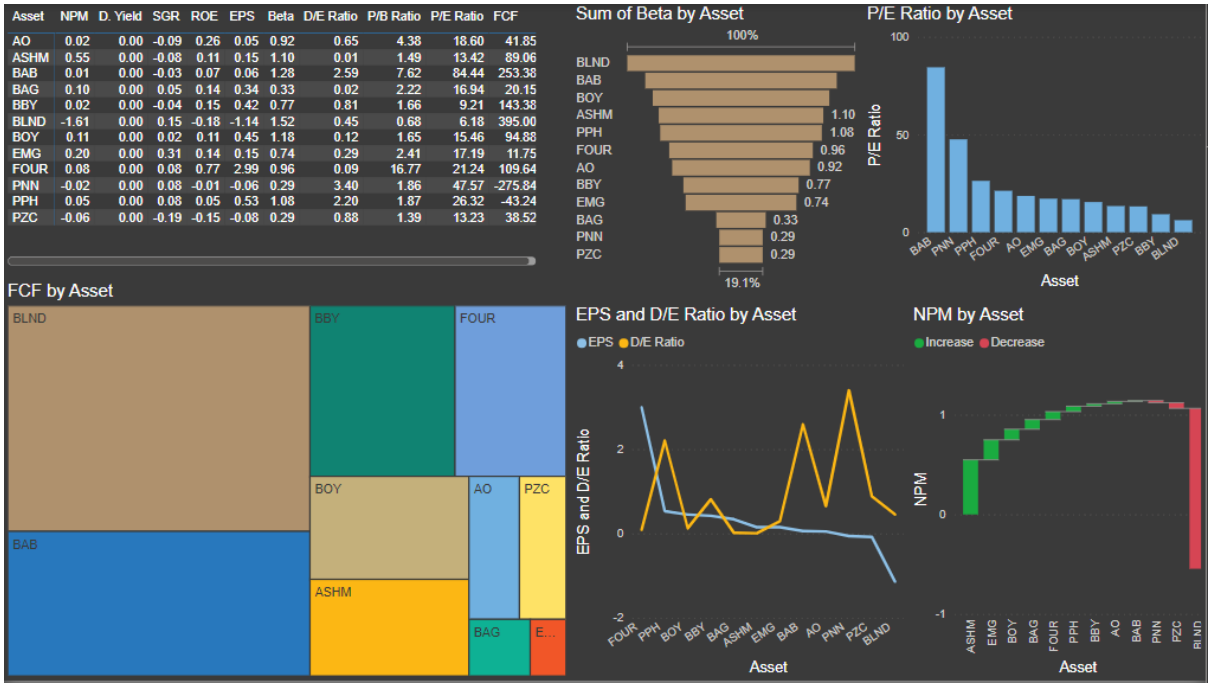


Figure 4.4: exploration of the selected companies and financial indicators.

In Figure 4.4, the matrix chart, also known as a matrix plot or heatmap was employed, to visualize the relationships, patterns, and distributions within the multidimensional dataset comprising all 12 selected companies and their respective financial indicators. This visualization allows investors to quickly assess all companies of interest and make informed decisions based on their specific perspectives or investment goals. By integrating 10 key financial indicators that can influence a company's short and long-term sustainability, this approach provides investors with a comprehensive understanding of the myriad factors impacting stock market performance. Unlike traditional methods that rely on a limited set of financial indicators, this comprehensive approach enables investors to consider a broader range of factors when making investment decisions.

Upon exploring the dashboard, it became evident that certain companies, such as FOUR (4imprint) and PPH (PPHE Hotel Group), showcase compelling values across their financial indicators, which might attract risk-tolerant investors employing various investment strategies. These companies demonstrate high earnings per share (EPS) alongside relatively low price-to-earnings (P/E) ratios, indicating strong profitability and growth potential. Additionally, they boast high return on equity (ROE), signifying efficient utilization of shareholder equity to generate profits, positive free cash flow (FCF), and relatively high beta values, implying greater volatility compared to the market.

Conversely, companies like PNN (Pennon Group) and BLND (British Land) exhibit negative EPS, suggesting lower profitability, negative ROE indicating poor performance concerning equity investment, negative net profit margins (NPM), and negative FCF, hinting at financial challenges. Furthermore, they have relatively low beta values, indicating lower volatility compared to the market, which may be attractive to risk-averse investors.

Overall, the matrix chart facilitated a deeper understanding of each company's financial performance and allows investors to identify opportunities and risks more effectively, thereby informing their investment decisions.

For investors focused on the beta value of a company, the funnel chart was employed as an effective tool to visually summarize and analyse the progression of items or values through various stages. In this context, BLND (British Land), a real estate company, displayed the highest beta value, indicating its sensitivity to market movements, while PZC exhibited the smallest beta value.

In addition, the clustered column chart was utilized to visualize the price-earnings ratio (P/E ratio) of all assets. This metric provides insight into the valuation of a company relative to its earnings, making it a crucial factor for investors assessing potential investment opportunities. Notably, BAB (Babcock International), a support service company, demonstrated the highest P/E ratio, signalling potentially higher growth expectations or valuation relative to its earnings, while BLND showcased the smallest P/E ratio.

Interestingly, despite BLND having the highest beta value, it displayed the smallest P/E ratio. This highlights the importance for investors to have a holistic understanding of various financial indicators influencing the stock market before making investment decisions. Relying solely on individual metrics like beta or P/E ratio may provide an incomplete picture of a company's financial health and growth potential. Therefore, investors benefit from considering a diverse range of indicators to make well-informed investment decisions that align with their investment objectives and risk tolerance.

The treemap chart, renowned for its versatility in visualizing hierarchical data structures and uncovering underlying patterns and relationships within intricate datasets, was harnessed to delve into the hierarchy of free cash flow across all selected assets. Notably, BLND emerged as the asset boasting the highest amount of free cash flow, underscoring its strong financial

position, while EMG exhibited the smallest positive free cash flow. However, assets such as PNN and PPH (PPHE Hotel Group), operating within the travel and leisure sector, showcased negative free cash flows, indicating potential financial challenges.

Furthermore, the line chart served as a valuable tool to juxtapose the earnings per share (EPS) and debt-to-equity ratio of all selected companies. FOUR stood out with the highest earnings per share, suggesting strong profitability, while BLND displayed the largest negative earnings per share, signalling potential financial strain. In terms of the debt-to-equity ratio, ASHM (Ashmore Group), a financial services company, boasted the lowest ratio, indicating a conservative approach to leveraging, whereas PNN exhibited the highest ratio, implying relatively higher debt levels compared to equity.

The waterfall chart, renowned for its ability to offer a concise and inherent representation of value flow, emerged as a valuable tool for visualizing the net profit margin across all selected companies. This visualization method proved invaluable in comprehending and effectively communicating the impact of sequential changes in data. Among the selected companies, ASHM showcased the highest net profit margin, highlighting its strong profitability and efficiency in generating profits. However, it's noteworthy that companies like PNN, PZC, and BLND exhibited negative net profit margins, signalling potential challenges or inefficiencies in their operations. By employing the waterfall chart to illustrate these metrics, investors and stakeholders can gain deeper insights into the financial performance of each company, facilitating better decision-making processes and strategic planning.

For additional exploratory analysis, the adjusted close prices of two distinct companies, FOUR and PNN, were showed using a line chart. Notably, FOUR exhibited an upward trend in its adjusted close prices throughout the post-COVID-19 years, whereas PNN experienced a continuous decline over the same period, as illustrated in Figure 4.5 below.

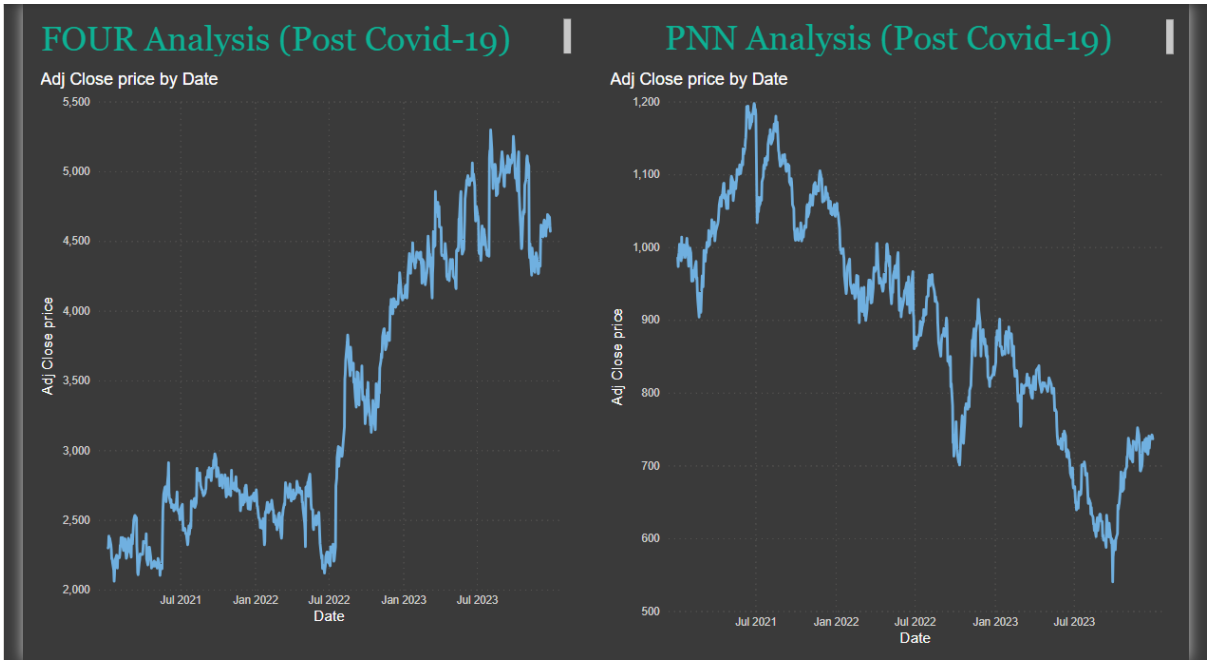


Figure 4.5: pictorial trend of adjusted close prices of FOUR and PNN post covid-19.

By leveraging these visualization techniques, investors can gain deeper insights into the financial health and performance of selected assets, empowering them to make well-informed investment decisions tailored to their objectives and risk preferences.

4.2. Data Preparation

Data preparation is an essential stage in any data analysis or machine learning process, involving various tasks such as cleaning, manipulation, and exploration to ensure that the data is suitable for analysis. Python offers a versatile set of libraries and tools to facilitate this process efficiently.

4.2.1. Data Summary and Statistical Exploration

The explored dataset was transferred to the programming language python, for preparation and further statistical exploration before the implementation of the machine learning algorithm.

4.2.1.1. Data Summary

To obtain a quick understanding of the dataset's structure and characteristics, the data.info() method from the pandas library was utilized. This method provided a concise summary of

the DataFrame, including information such as data types, column names, non-null values, and memory usage.

```
# get data summary
print("Data Summary:")
print(AO.info())
```

```
Data Summary:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 757 entries, 0 to 756
Data columns (total 7 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Date       757 non-null    object
 1   Open       757 non-null    float64
 2   High       757 non-null    float64
 3   Low        757 non-null    float64
 4   Close      757 non-null    float64
 5   Adj Close  757 non-null    float64
 6   Volume     757 non-null    int64
dtypes: float64(5), int64(1), object(1)
memory usage: 41.5+ KB
None
```

Figure 4.6: data summary.

The data summary (Figure 4.6) offered insights into the DataFrame's structure, identifying the number of non-null values, and assessing the data types of each column. This overview facilitated informed decisions regarding data manipulation, cleaning, and analysis.

4.2.1.2 Statistical Exploration

Statistical exploration involved analysing and summarizing the main characteristics of the dataset to gain insights into its structure and key attributes. Descriptive statistics generated using the data.describe().round(2) method, provided a summary of the dataset's central tendency, variability, and key statistics.

```
# statistical exploration
print("\nDescriptive Statistics:")
print(AO.describe().round(2))
```

```
Descriptive Statistics:
          Open    High     Low   Close  Adj Close      Volume
count   757.00  757.00  757.00  757.00     757.00      757.00
mean    119.91  122.56  116.68  119.54     119.54   463917.34
std      28.18   28.52   27.49   28.05      28.05   977694.17
min      56.70   61.20   56.70   57.10      57.10        0.00
25%     104.40  107.00  100.60  103.50     103.50   116979.00
50%     123.60  127.00  120.75  123.00     123.00   214994.00
75%     141.80  144.60  137.20  141.40     141.40   451857.00
max     185.60  190.57  182.50  184.30     184.30 15142548.00
```

Figure 4.7: statistical exploration.

Figure 4.7 illustrates the results of the statistical exploration, offering insights into the dataset's characteristics. Descriptive statistics aided in identifying anomalies or outliers and informed subsequent analysis and modelling approaches. Understanding the dataset's underlying patterns and distributions is crucial for effective data exploration.

4.2.2. Data Cleaning

After obtaining the metadata and identifying missing or erroneous values, the data cleaning process began. Irrelevant variables were removed using the data.drop(columns=['Date']) function.

```
#Data cleaning
# remove non-relevant variable
AO1 = AO.drop(columns= ['Date'])
AO1.head()
```

	Open	High	Low	Close	Adj Close	Volume
0	91.400002	97.069000	88.699997	91.800003	91.800003	61142
1	94.800003	99.654999	88.000000	95.300003	95.300003	102462
2	94.699997	103.199997	90.737000	98.000000	98.000000	37771
3	96.099998	97.500000	89.900002	90.400002	90.400002	467720
4	92.000000	93.000000	86.650002	93.000000	93.000000	163585

Figure 4.8: removal of irrelevant variable.

Figure 4.8 demonstrates the removal of an irrelevant variable, Date, from the dataset. Additionally, missing values were addressed using appropriate methods. For datasets with many missing observations, the data.fillna(data.mean()) function was used to fill the missing values with the column mean, while the data.dropna() function removed rows with few missing values.

```
# fill missing values with the mean of each column
AO_new.fillna(AO_new.mean(), inplace=True)

# check if missing values are removed
print(AO_new.isnull().sum())
```

```
Date         0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

```
#Data cleaning
# remove rows with missing values
BAB_new.dropna(inplace=True)

# check if missing values are removed
print(BAB_new.isnull().sum())
```

```
Date         0
Open         0
High         0
Low          0
Close        0
Adj Close    0
Volume       0
dtype: int64
```

Figure 4.9: missing values solutions.

Figure 4.9 illustrates the solutions applied to address missing values in the dataset, ensuring data completeness and integrity.

4.2.3. Cleaned Data Visualization

Cleaned data were visualized using scatter and histogram charts to explore the relationship between selected features and the target variable. These charts are valuable for understanding feature distributions and relationships with the target variable.

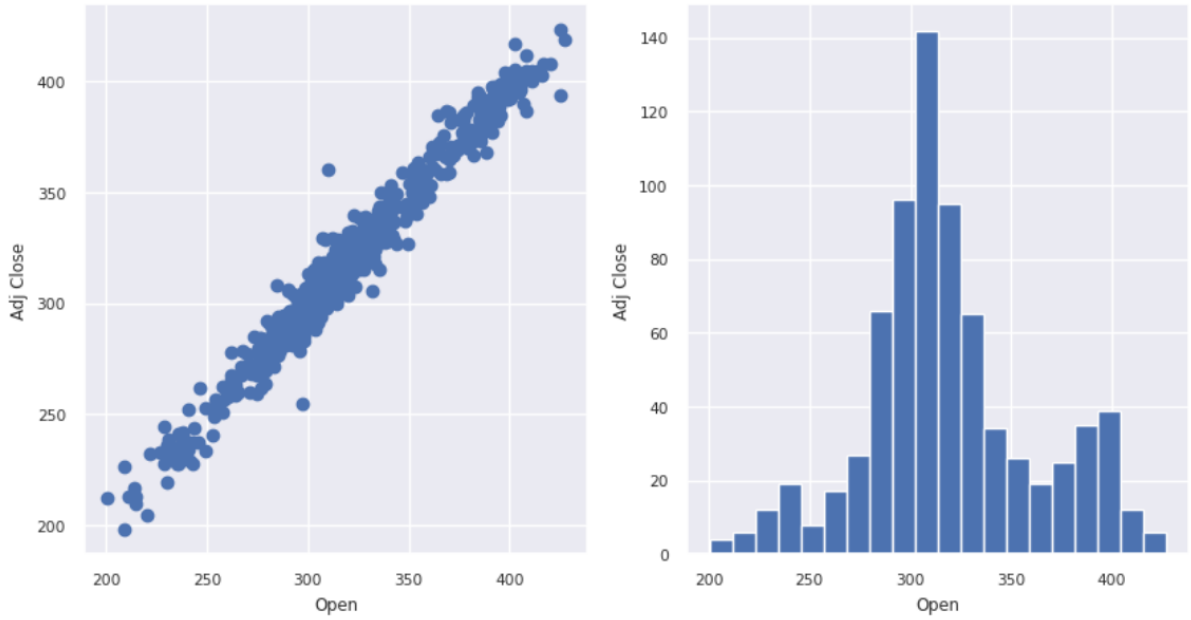


Figure 4:10. scatter and histogram charts of a selected feature (Open) and the target variable (Adj Close).

Figure 4.10 shows scatter and histogram charts illustrating the relationship between a selected feature (Open) and the target variable (Adj Close).

4.2.4. Data Quality Check

The quality of the cleaned data was assessed by calculating correlations between variables, particularly with the response variable. Correlation analysis helps determine the strength and direction of relationships, guiding subsequent modelling decisions.

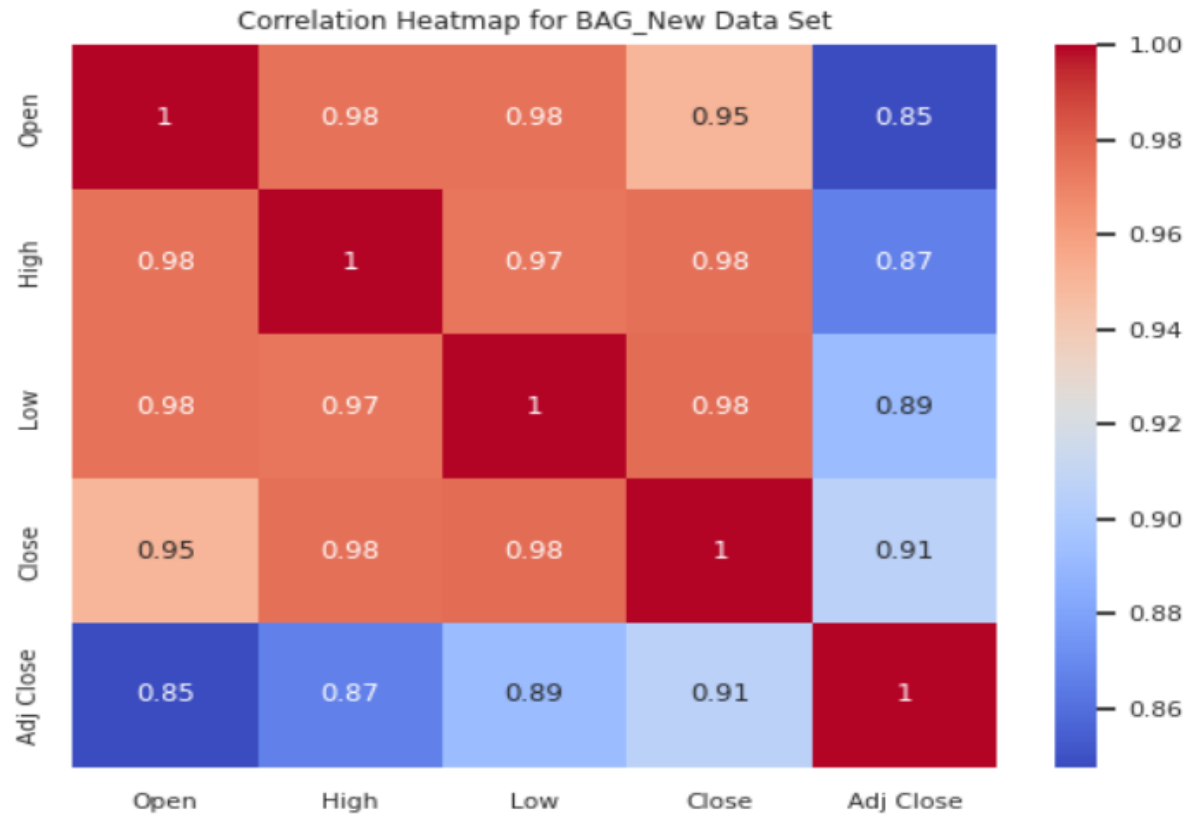


Figure 4.11: correlation heatmap of BAG (post covid-19) dataset.

Figure 4.11 displays a correlation heatmap, providing insights into the relationships between variables in the dataset.

Following these steps ensured that the datasets were properly prepared for analysis and modelling, laying the foundation for accurate predictions and informed decision-making.

4.3. Process Design

In this section, we established a structured framework, known as process design, to efficiently carry out a series of tasks aimed at achieving our desired outcomes. This process involved delineating the sequence of steps, required resources, and systematic methodologies necessary to achieve our objectives effectively.

4.3.1. Data Exploration & Preparation Flow Diagram

The Data Exploration & Preparation Flow Diagram provides a detailed breakdown of the activities and methods used for gathering, cleaning, visualizing, and analysing the data. It encompasses various stages, including data collection, preprocessing, exploration, and quality assessment.

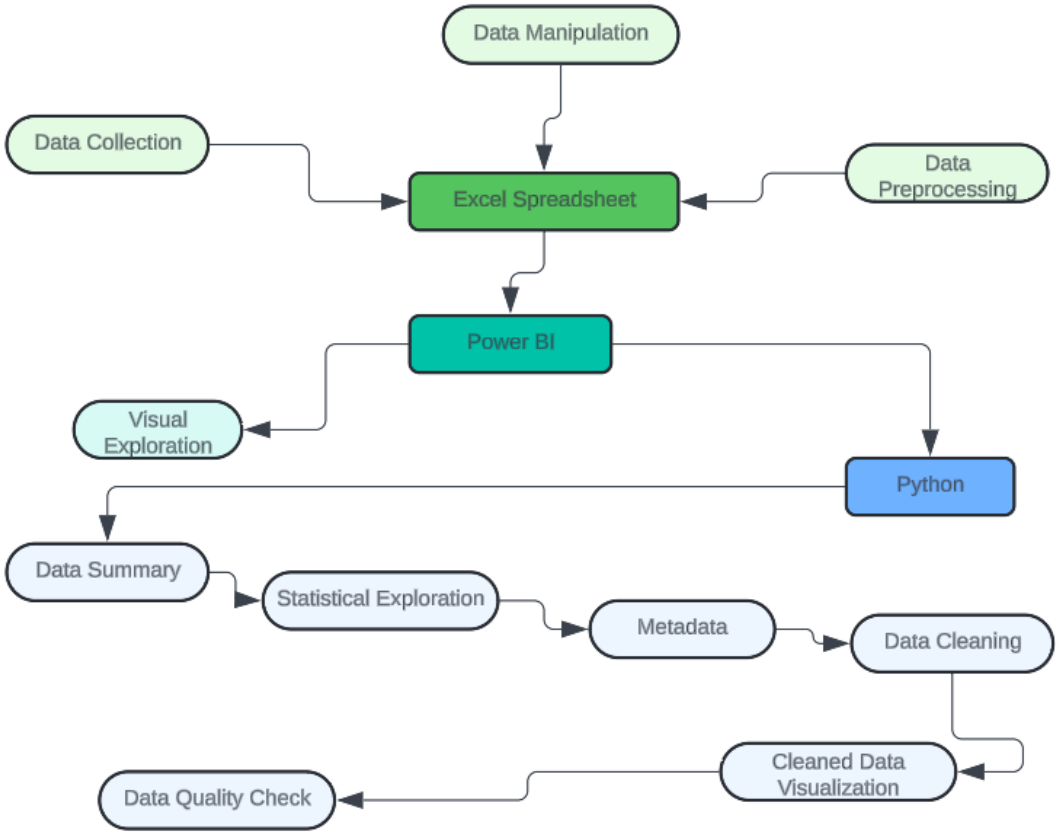


Figure 4.12: Data Exploration & Preparation Flow Diagram.

The diagram illustrates the steps involved in exploring and preparing the historical stock prices dataset. It highlights the utilization of tools such as Excel, Power BI, and Python programming language to ensure comprehensive data exploration and preparation before advancing to the modelling phase.

4.3.2. Predictive Modelling

During the predictive modelling phase, several techniques were considered and applied to the historical stock prices datasets using Python. The selection of the models was based on the analysis conducted earlier and the specific characteristics of the datasets.

Three main predictive models were chosen: random forest, gradient boosting, and support vector machine (SVM). These models were selected for their effectiveness in handling the complexity of the data and their ability to generate accurate predictions. Additionally, in cases where linear regression outperformed the selected models, it was used as an alternative.

The modelling process involved fitting the parameters of each selected model to the data. This included defining the appropriate parameters for each model and training them on the dataset. The performance of each model was then evaluated using various metrics such as accuracy score, mean squared error, and overfitting analysis.

The random forest regression model, known for its simplicity and interpretability, was applied initially. Following that, the gradient boosting regressor, a popular and powerful model for predictive analysis, was implemented. Finally, the support vector machine model was utilized to further explore the dataset's predictive capabilities.

In cases where linear regression demonstrated superior performance compared to the selected models, it was used as a replacement. Additionally, autoregressive integrated moving average (ARIMA) modelling was employed for time series analysis to forecast future adjusted close prices.

Overall, the modelling phase aimed to leverage the strengths of each model to generate accurate predictions and gain insights into the underlying patterns and trends within the historical stock prices datasets.

4.3.2. Model Comparison

In this section, the performance of the three models analysed in the previous phase were compared. Initially, the models were evaluated by comparing misclassification rates and accuracy scores to gauge their effectiveness. However, to ensure a comprehensive evaluation, two key methods were utilized: R-squared score (R^2_score) and mean squared

error (MSE). Depending on the specific dataset and context, either R-squared score or mean squared error (MSE) (test) and mean squared error (MSE) (cross-validation) were considered to determine the best-fitted model.

The evaluation process involved comparing the actual and predicted results of the response variable to assess each model's performance. The distribution of prediction errors was visualized using histogram plots, which provided insights into the spread and frequency of errors made by each model.

To select the best-fitted model for the dataset, the data was split further into training and testing sets using the train_test_split function from scikit-learn. Twenty percent (20%) of the data was reserved for testing purposes. In some cases, cross-validation with five folds was applied using the cross_val_score function to train the model, allowing assess to its performance across different subsets of the data.

4.3.3. Model Implementation & Comparison Flow Diagram

The Model Implementation & Comparison Flow Diagram showed in Figure 4.13 outlines the structured process for implementing and comparing various predictive models.

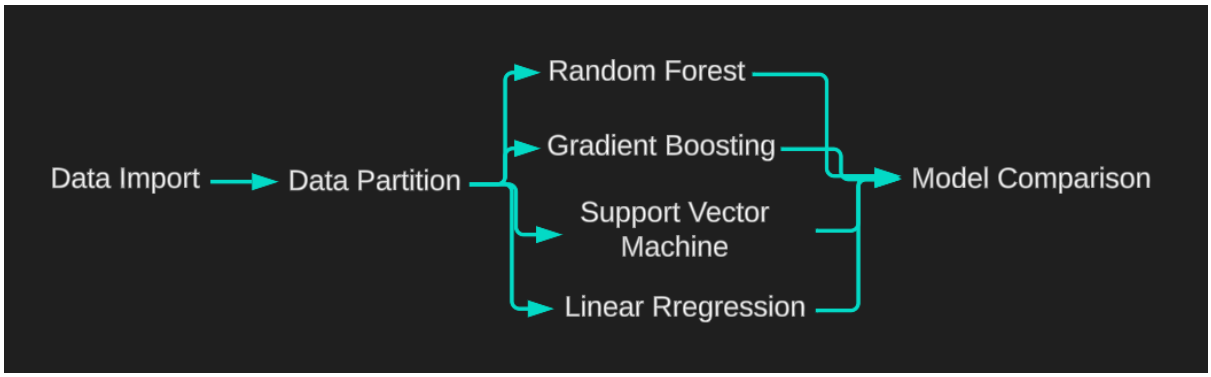


Figure 4.13: data modelling and comparison flow diagram.

This flow diagram provides a structured and systematic approach to model development and selection, ensuring efficient implementation and comparison of predictive models in the data mining process using Python.

Chapter 5. Implementation

In this section, the selected predictive models were trained and evaluated using historical stock price data. Key steps included splitting the data into training and testing sets, fitting models to the training data, making predictions on the test data, and evaluating performance using metrics such as Mean Squared Error (MSE) and R-squared (R^2) score. Additional, cross-validation techniques were employed in some cases to ensure strongness and generalizability of the models. Visualization of model predictions against actual values was also conducted to assess accuracy and identify areas for improvement. Overall, the goal of this phase was to develop and assess predictive models capable of accurately forecasting future stock prices based on historical data.

5.1. Models Building

During this phase, a predictive modelling analysis was conducted wherein three regression models were selected and fitted to the observed datasets at each juncture. The objective was to deliver precise predictions of the target variable and identify the primary factors directly influencing the target outcome. This was achieved by comparing the outcomes of the three models using supervised machine learning techniques and selecting the most suitable model for the dataset.

To conduct this analysis, four modelling techniques were integrated into the workflow diagram: random forest, gradient boosting, support vector machine, and linear regression. These models were customized by adjusting their parameters to construct individual models for the analysis. Due to constraints, only a few companies were randomly chosen to demonstrate the various evaluation methods and their efficacy. Covering all 12 selected companies in this project would be impractical within this scope. However, a detailed analysis of all companies is available in Python and can be furnished upon request.

5.1.1. Random Forest Modelling

For the predictive analysis, the random forest model was selected first, known for its strongness and ease of use. The dataset containing historical stock price data for the company (BAG) was divided into features and the target variable, which represented the adjusted close prices. This dataset was then split into training and testing sets using the

train_test_split function from scikit-learn, allocating 80% of the data for training and 20% for testing.

Next, the Random Forest Regressor was trained using the training data. This ensemble learning method constructs multiple decision trees during training and outputs the mean prediction of the individual trees. The model was fitted to the training data using the fit method. Following the model training, predictions on the test set were made using the predict method. This allowed the generation of the predicted values for the adjusted close prices based on the features in the test data.

To evaluate the performance of the model, the R-squared (R^2) score was calculated, which measures the proportion of the variance in the target variable that is predictable from the features. In this case, the R^2 score was approximately 0.8355, indicating a relatively good fit of the Random Forest model to the test data.

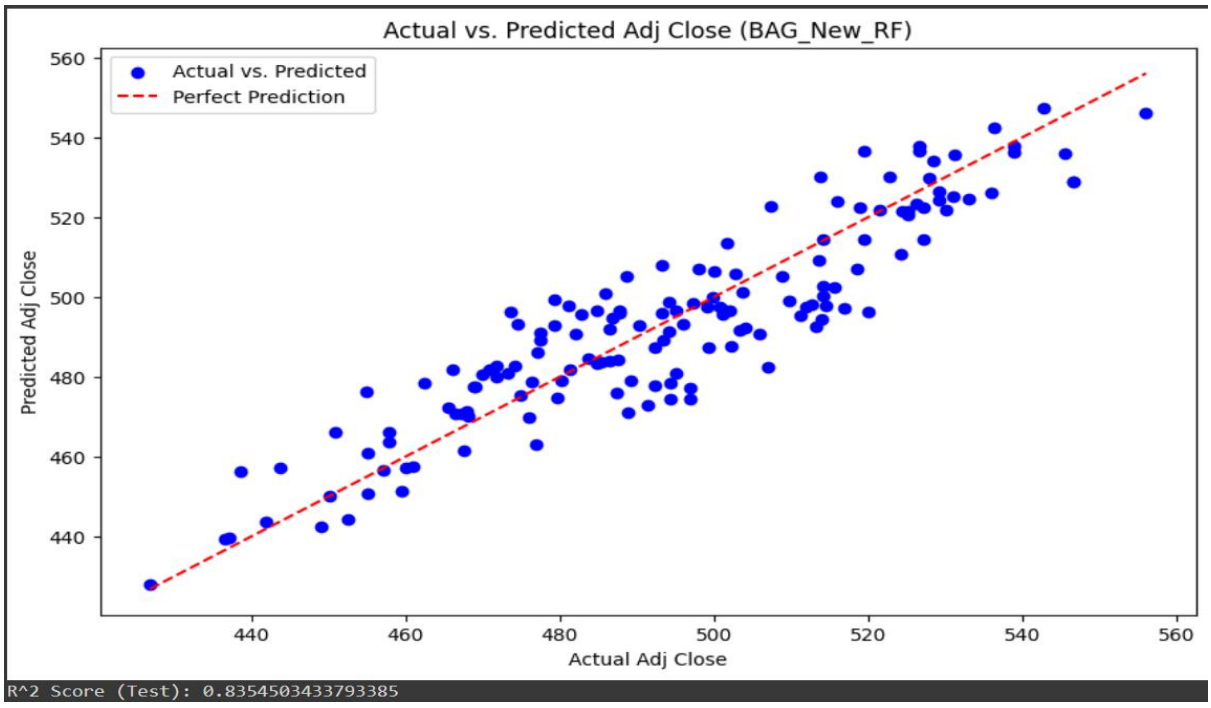


Figure 5.1: the actual vs predicted values of BAG (post covid-19) using random forest model.

The model's predictions were visualized by plotting the actual versus predicted values of the adjusted close prices as shown in figure 5.1. The scatter plot provided a visual assessment of how well the model's predictions aligned with the actual values. Additionally, the distribution of prediction errors was visualized using a histogram plot, offering insights into the spread and frequency of errors made by the model.

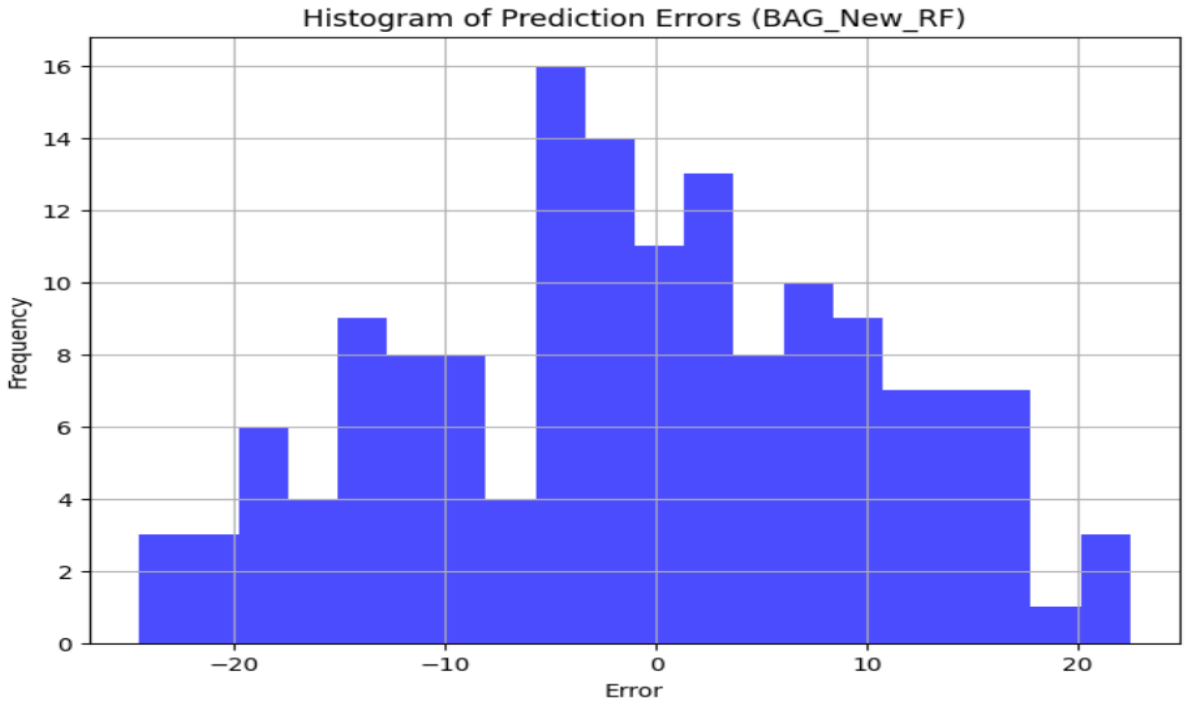


Figure 5.2: distribution of prediction errors of BAG (post covid-19) for random forest.

In the histogram plot, the tallest bar centred around -5 on the x-axis suggests that there were more predictions with errors around -5 than any other error value, indicating varied errors.

Overall, the Random Forest algorithm demonstrated effectiveness in predicting the adjusted close prices of the company (BAG), achieving a high R^2 score and producing accurate predictions that closely matched the actual values.

5.1.2. Gradient Boosting Modelling

In the implementation of the Gradient Boosting Regressor model for supervised learning, the objective was to predict the adjusted close prices of a randomly selected company (AO) using its historical stock price data. The process involved several steps to ensure accurate model evaluation and predictive performance.

Initially, the dataset was divided into features and target variable, followed by the splitting of the data into training and testing sets. This partitioning allowed for the assessment of the model's performance on unseen data. The Gradient Boosting Regressor, a powerful ensemble learning technique, was then employed to train the data. This method constructs

multiple decision trees sequentially, with each tree correcting the errors of the previous one, thereby enhancing predictive accuracy.

To evaluate the model's performance during training, cross-validation with 5 folds was utilized. This technique provided insights into the model's generalization capabilities and estimated its performance on unseen data. Following training, predictions were made on the test set using the predict method. Performance evaluation was conducted using the Mean Squared Error (MSE) computed from cross-validation scores, which measured the average squared difference between predicted and actual values. Additionally, the MSE specifically for the test set was calculated to assess the model's performance on unseen data.

The model's predictions were visually inspected using a scatter plot, enabling a comparison between actual and predicted values. The scatter plot showed in Figure 5.3 illustrated how well the model's predictions aligned with the actual values.

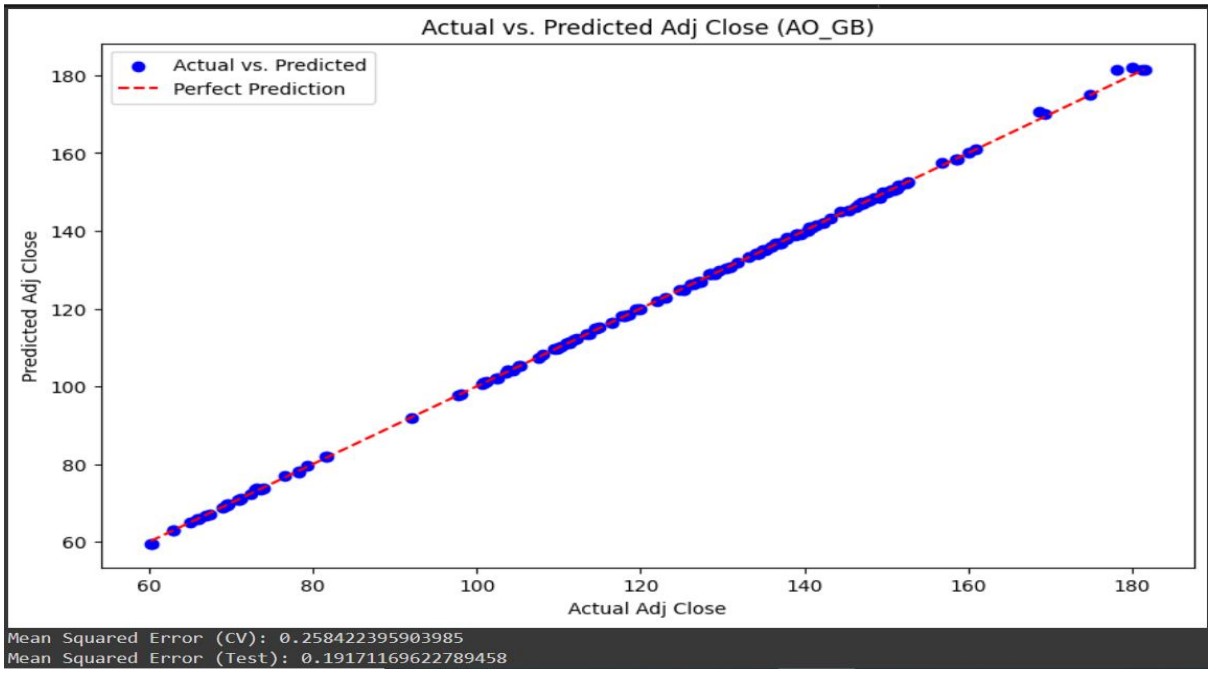


Figure 5.3: the actual vs predicted values of AO (pre covid-19) using gradient boosting model.

Furthermore, the distribution of prediction errors was visualized using a histogram plot, as shown in Figure 5.4. This provided insights into the spread and frequency of errors made by the model. The symmetrical distribution around zero indicated a balanced number of positive and negative errors, with a high frequency of predictions close to the correct value.

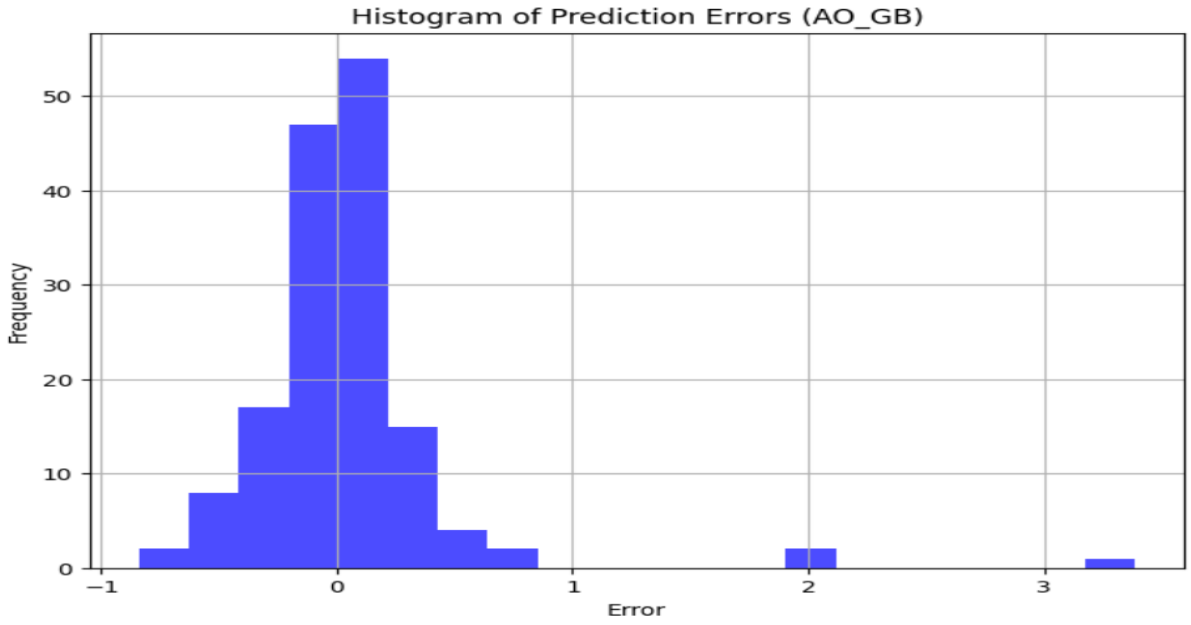


Figure 5.4: distribution of prediction errors of AO (pre covid-19) for gradient boosting.

Overall, the Gradient Boosting Regressor model demonstrated effectiveness in predicting the adjusted close prices of the randomly selected company (AO) based on its historical stock price data.

5.1.3. Support Vector Machine Modelling

Utilizing the support vector machine (SVM), this section aimed to predict the adjusted close prices of a chosen company, BBY, based on its historical stock data. The process initiated with the partitioning of the dataset into independent variables (features) and the dependent variable (target). Employing a linear kernel for the SVM model, it aimed to determine the optimal linear boundary between classes in the feature space.

During the model training phase, cross-validation with 5 folds was implemented to assess the model's performance and its generalization across various subsets of the training data. After training, predictions were made on the test set, generating estimated values for the adjusted close prices based on the features in the test data.

For evaluating the model's efficacy, two key metrics were employed: Mean Squared Error (MSE) and the histogram of prediction errors. MSE, computed from both cross-validation scores and specifically for the test set, gauged the average squared difference between the predicted and actual adjusted close prices.

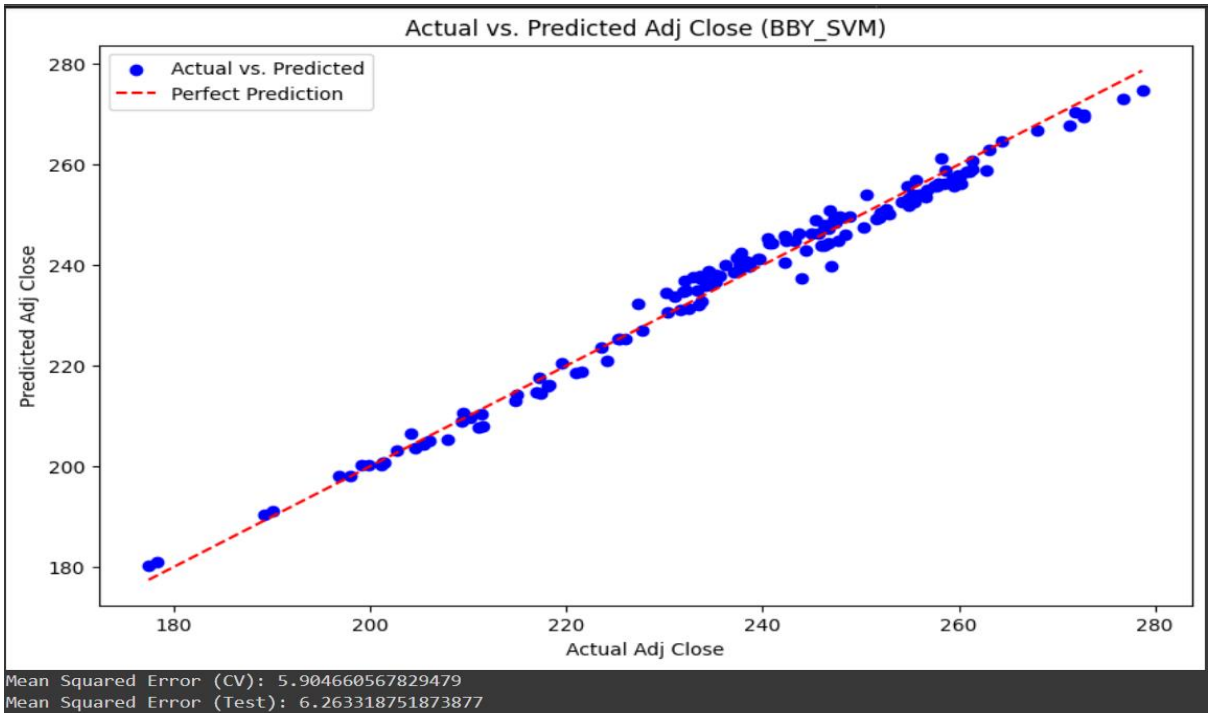


Figure 5.5: the actual vs predicted values of BBY (pre covid-19) using svm model.

Figure 5.5 illustrates the comparison between the actual and predicted values of BBY using the SVM model. The results revealed an MSE (CV) of approximately 5.90 and an MSE (Test) of about 6.26, offering insights into the model's predictive accuracy on unseen data.

Moreover, the distribution of prediction errors was visualized through a histogram plot, as showed in Figure 5.6.

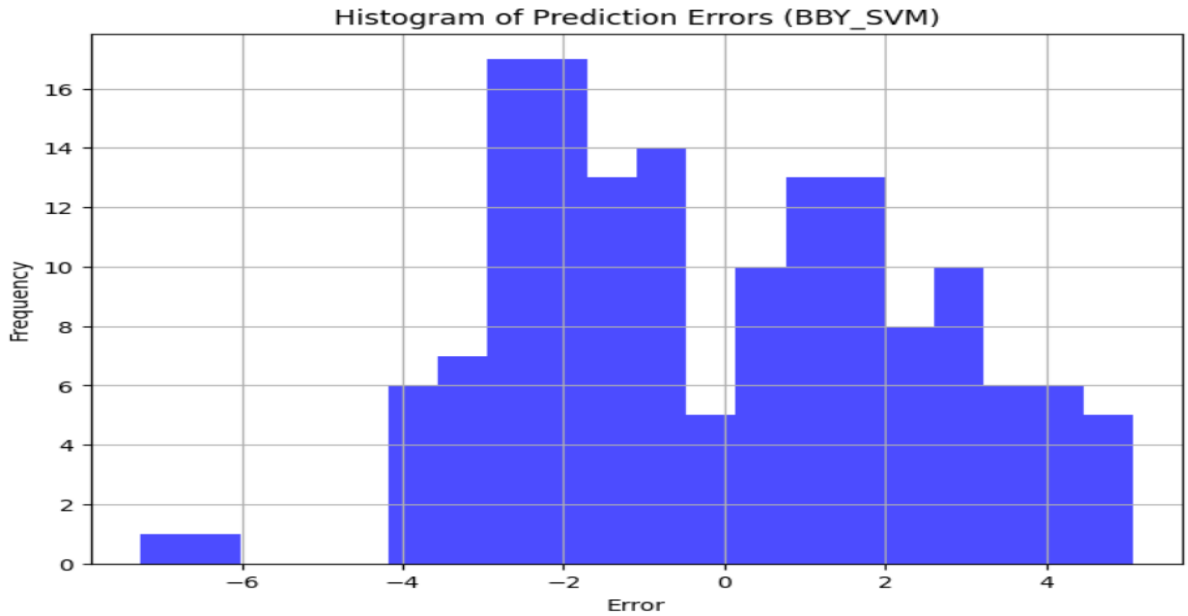


Figure 5.6: distribution of prediction errors of BBY (pre covid-19) for svm.

The histogram offered insight into both the frequency and distribution of errors generated by the model. The peak around -3 and -2 on the x-axis indicates that the model made more predictions with errors clustered around these values compared to other error magnitudes, highlighting the diversity in the errors observed.

5.1.4. Linear Regression Modelling

The linear regression model was employed in select cases where it outperformed the other main models. In this instance, it was utilized for supervised learning to forecast the adjusted close prices of a chosen company, BLND, leveraging its historical stock data. The process closely mirrored that of the main three models, with the R-squared (R^2) score serving as the metric for evaluating the model's performance. The R^2 score gauges the proportion of variance in the target variable that can be predicted from the features. With an R^2 score of approximately 0.7229, the linear regression model demonstrated a reasonably good fit to the test data.

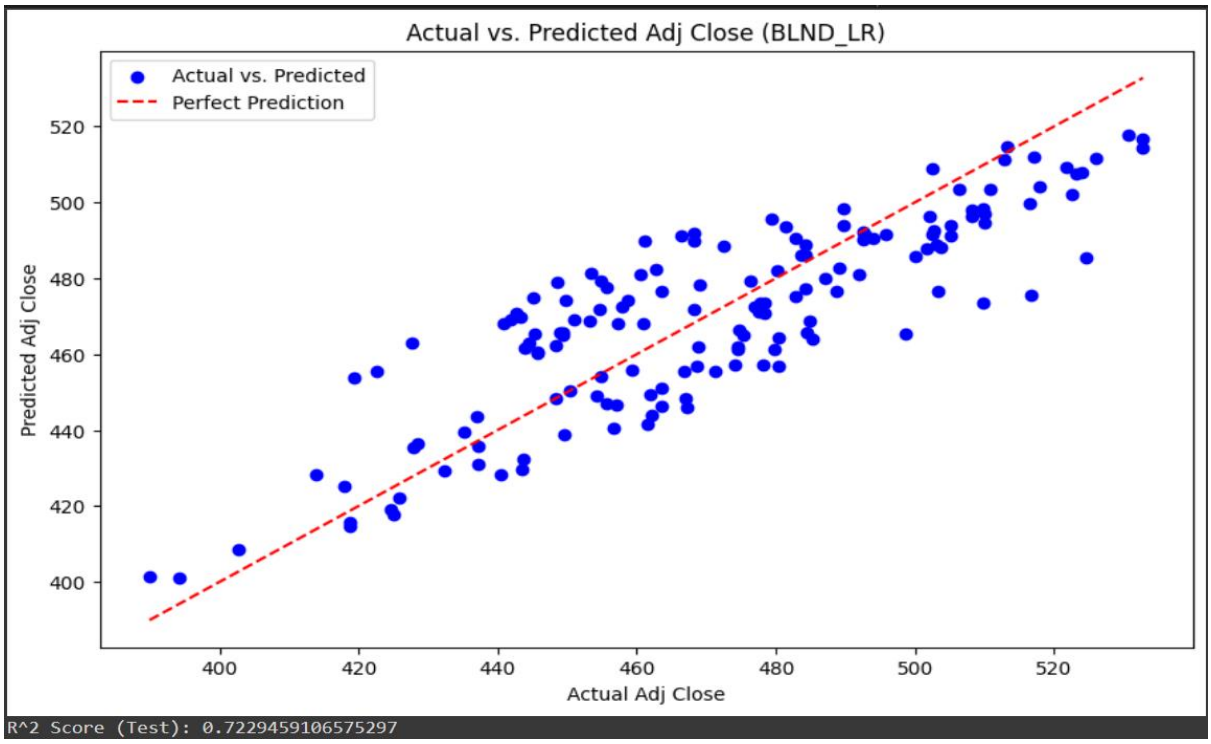


Figure 5.7: the actual vs predicted values of BLND (pre covid-19) using linear regression model.

Figure 5.7 illustrates the comparison between the actual and predicted values of BLND using the linear regression model. Furthermore, to gain insights into the spread and frequency of prediction errors, a histogram plot was generated, as showed in Figure 5.8.

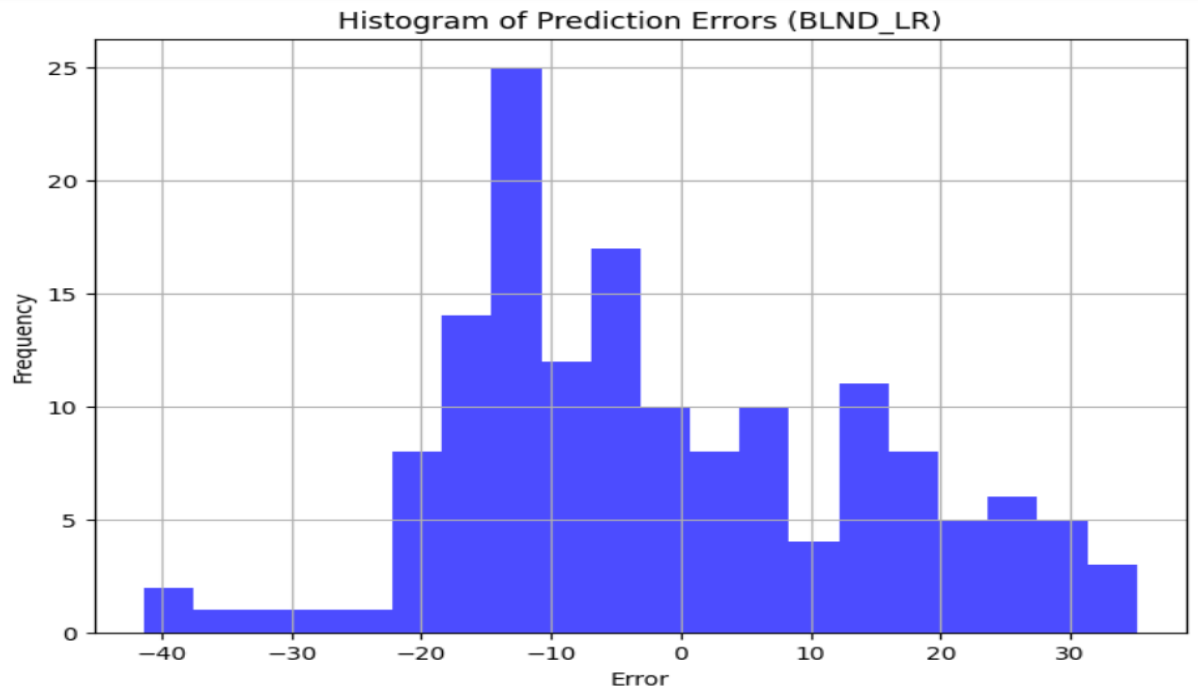


Figure 5.8: distribution of prediction errors of BLND (pre covid-19) for linear regression.

Figure 5.8 showcases the distribution of prediction errors for BLND using the linear regression model. The histogram indicates that the most frequent prediction errors centred around -12, suggesting a propensity for underprediction bias. This implies that the model tended to predict lower prices than the actual closing prices, potentially influenced by outliers in the data.

In summary, the linear regression model proved effective in forecasting the adjusted close prices of BLND based on its historical stock data, achieving a satisfactory R^2 score and generating predictions that closely aligned with the actual values.

5.1.5. Autoregressive Integrated Moving Average (ARIMA) Analysis

For the time series analysis, the autoregressive integrated moving average (ARIMA) model was applied to the adjusted close prices of one randomly selected dataset, BOY, representing the stock prices. The adjusted close prices over time were visualized using Matplotlib, capturing both pre and post-COVID-19 periods.

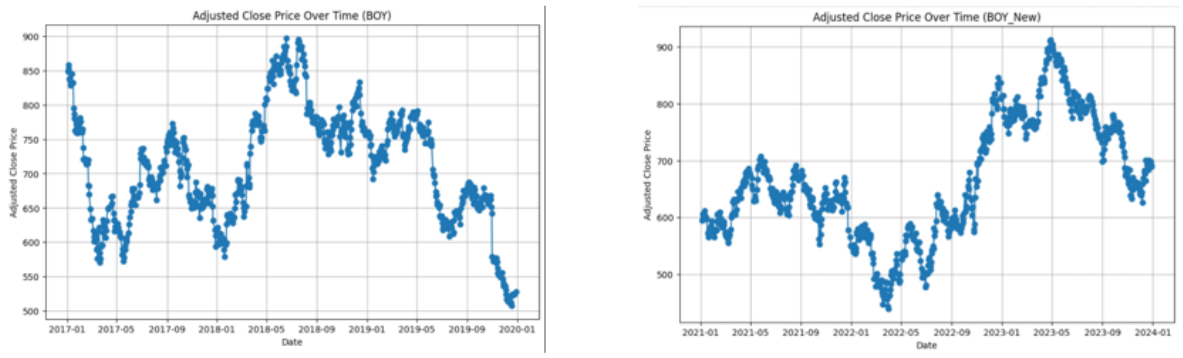


Figure 5.9: adjusted close price over time using ARIMA for both pre and post covid-19.

Figure 5.9 illustrates the adjusted close prices of BOY, showcasing fluctuations before and after the COVID-19 pandemic. Notably, there is a visible decline in prices towards the end of the pre-COVID-19 period, followed by a slight increase post-COVID-19. Additionally, the lowest and highest prices for both periods are highlighted, indicating shifts in price levels between the two phases.

The summary results of the fitted ARIMA model and the forecasted future values are presented in Figure 5.10.

```
==============================================================================
Dep. Variable:              Adj Close   No. Observations:                  754
Model:                 ARIMA(5, 1, 0)   Log Likelihood               -2935.426
Date:                Sat, 27 Apr 2024   AIC                           5882.852
Time:                        20:10:02   BIC                           5910.597
Sample:                             0   HQIC                          5893.541
                                  - 754
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
ar.L1          0.0270      0.033      0.817      0.414      -0.038       0.092
ar.L2         -0.0145      0.034     -0.429      0.668      -0.081       0.052
ar.L3         -0.0242      0.030     -0.807      0.420      -0.083       0.035
ar.L4          0.0074      0.032      0.230      0.818      -0.055       0.070
ar.L5         -0.0023      0.032     -0.073      0.942      -0.065       0.060
sigma2       142.4132      5.959     23.897      0.000     130.733     154.094
==============================================================================
Ljung-Box (L1) (Q):                   0.00   Jarque-Bera (JB):                52.90
Prob(Q):                              1.00   Prob(JB):                         0.00
Heteroskedasticity (H):               1.03   Skew:                             0.22
Prob(H) (two-sided):                  0.84   Kurtosis:                         4.22
==============================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
Forecasted Values: 754    690.047066
755    690.206680
756    690.192037
757    690.191667
758    690.192509
```

Figure 5.10: summary result of the fitted ARIMA model and the forecasted future values.

This output provides comprehensive insights into the ARIMA model's performance, offering various statistical metrics and diagnostics about its fit to the data. Key metrics such as the log likelihood, AIC, and BIC are included, aiding in assessing the model's goodness of fit and predictive accuracy. Moreover, the forecasted values represent the model's predictions for the adjusted close prices over the next 5 days.

In essence, the output furnishes valuable information about the ARIMA model's efficacy in analysing time series data, offering a deeper understanding of the dataset's patterns and trends. It serves as a crucial tool for making informed decisions and predictions regarding future price movements in the stock market.

5.2. Model Comparison

During the model comparison phase, two key metrics, mean squared error and R-squared score, were employed to evaluate the performance of different regression models, and determine the optimal model for deployment. This assessment was conducted for two companies, representing scenarios both before and after the onset of the pandemic. Although the analysis covered a broad selection of companies, we focused on two specific cases for illustration purposes.

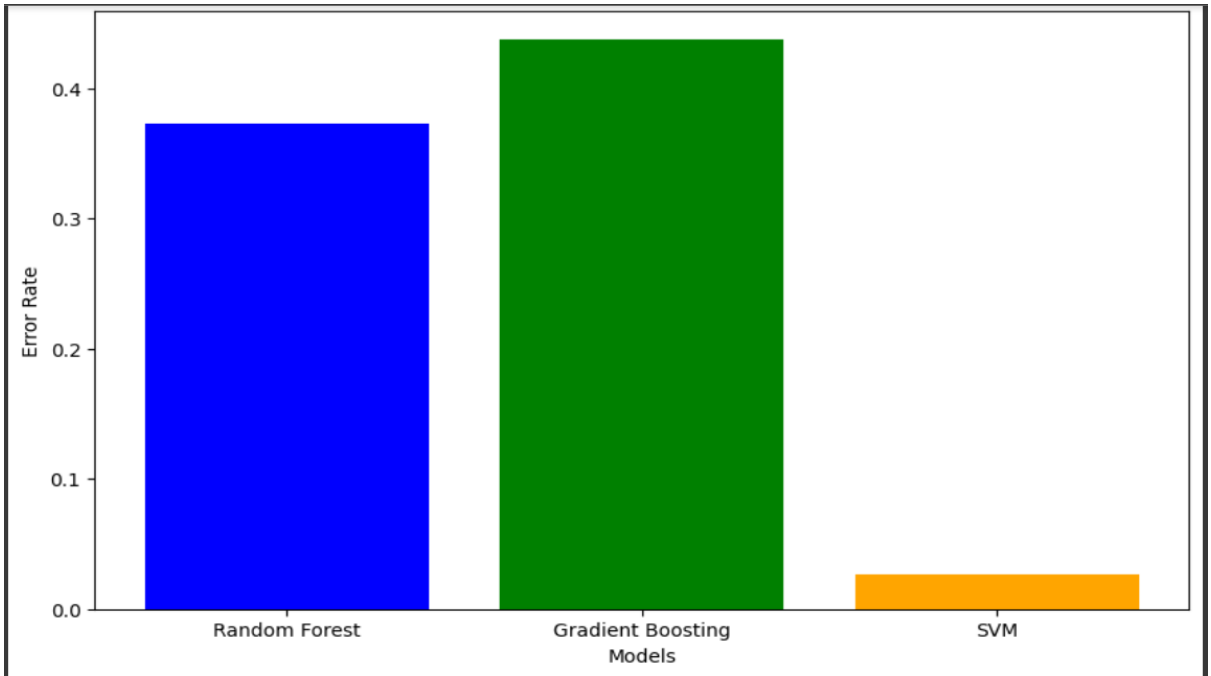


Figure 5.11: comparison of error rates for AO (pre covid-19)

In Figure 5.11, the comparison of error rates for Company AO before the pandemic revealed that the support vector machine (SVM) outperformed both random forest and gradient boosting models. With an impressively low error rate of 0.0264, SVM surpassed the error rates of random forest (0.3727) and gradient boosting (0.4378). Consequently, for Company AO before the pandemic, SVM emerged as the best-performing model, suitable for deployment.

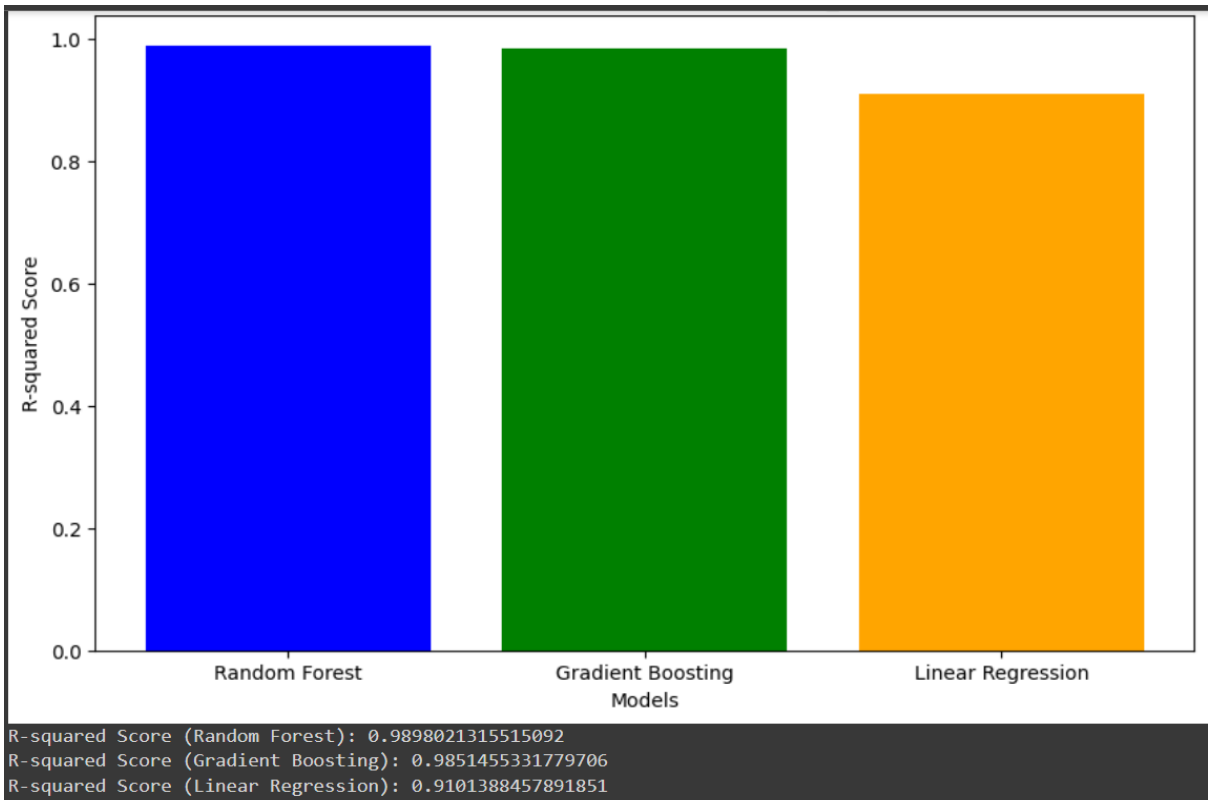


Figure 5.12: comparison of R-squared scores for PNN (post covid-19)

Moving to Figure 5.12, which showcases the comparison of R-squared scores for Company PNN after the pandemic, it was observed that the random forest model exhibited superior performance compared to gradient boosting and linear regression models. With an R-squared score of approximately 0.9898, random forest surpassed the R-squared scores of gradient boosting (0.9851) and linear regression (0.9101). Hence, for Company PNN after the pandemic, the random forest model was identified as the most suitable for deployment.

Overall, the analysis revealed that the gradient boosting model demonstrated the highest number of best performances across various scenarios. Specifically, it outperformed other models on 15 occasions, with 7 instances occurring before the pandemic and 8 instances

after. The random forest model secured the best performance spot 5 times, with 3 occurrences before the pandemic and 2 after. Meanwhile, the support vector machine model emerged as the best performer 4 times, evenly split between pre and post-pandemic scenarios. Notably, the linear regression model was utilized 3 times throughout the analysis, with one instance before the pandemic and two after, but it did not emerge as the best performer in any scenario.

Consequently, the gradient boosting model boasted the highest frequency of best performances, accounting for 62.5% of all instances throughout the analysis. This underscores its efficacy and versatility across different scenarios, positioning it as a reliable choice for deployment in various predictive modelling tasks.

5.3. Work Evaluation

The stock market operates within a dynamic environment where various factors influence the buying and selling of shares and other financial instruments. It is highly reactive and subject to frequent changes, making it easy for investors to experience gains or losses in their investments. Numerous financial indicators contribute to market performance, and investors often grapple with decision-making due to the multitude of factors at play. The main objective of this project is to leverage historical stock price data, alongside various financial indicators, to predict future movements in stock prices both before and after the pandemic. This aims to equip investors with comprehensive insights for making data-driven investment decisions.

To conduct an effective analysis that accurately addresses market requirements, a strong methodology and a diverse set of analytical methods and tools were employed. The CRISP-DM methodology guided and structured the multiple phases of the analysis process. Additionally, to meet the analytical needs in terms of visual and statistical exploration, preparation, modelling, and evaluation, a powerful programming language and two versatile software were utilized.

Excel was initially employed for data collection and preliminary analysis. It facilitated the computation of key metrics such as mean, standard deviation, and correlations between asset returns. This provided valuable insights into asset performance and interrelationships. Power BI was then utilized to simplify understanding of variable distributions and identify

meaningful patterns. Its interactive features enabled trend analysis, comparative analysis, correlation analysis, anomaly detection, and forecasting analytics, enhancing understanding, and providing insights for further analysis.

Python programming language was leveraged in the final steps of the analysis, covering post-preparation, partitioning, modelling, and model comparison processes. A variety of statistical models were applied and evaluated to determine the best-performing model for each company before and after the pandemic, aiding deployment and decision-making. By bringing together ten financial indicators across twelve companies, the analysis aimed to provide investors with informed insights into stock price movements.

However, due to the limited number of measures, precise conclusions regarding outcomes were challenging. Future analyses could benefit from a more detailed data collection process, incorporating a broader range of features to enhance model accuracy. Additionally, to streamline the analysis process, future endeavours may consider using a single software platform capable of addressing diverse analytical tasks effectively.

Finally, the personal aim of this project was to gain proficiency in utilizing a variety of software tools, reflecting the knowledge and technical development acquired during the master's program. As a recommendation for future analyses, reducing complexity by consolidating software usage may streamline processes and enhance efficiency.

Chapter 6. Conclusion and Recommendations for Future Work

6.1. Conclusion

This project aimed to analyse various financial indicators affecting stock market performance to provide investors, policymakers, and financial analysts with comprehensive insights for data-driven decision-making. The goal was to bring together several financial indicators and develop accurate predictive models to assist investors in minimizing investment losses by forecasting future stock price movements based on historical data and financial indicators.

Risk-tolerant investors may find companies like FOUR and PPH attractive due to their strong financial indicators, including high EPS, low P/E ratios, and positive FCF. On the other hand, risk-averse investors might favour companies like PNN and BLND, which exhibit lower volatility and negative EPS.

Post-COVID-19, companies such as FOUR and EMG have shown significant recovery in adjusted close prices, while others like BAB and PZC have experienced consistent declines.

Data processing played a crucial role, involving exploration, preparation, and modelling phases using various statistical classification models. A comparison analysis was conducted to determine the best-fitted model for deployment.

Overall, the project adhered to data mining success criteria by employing a clear and comprehensible approach, utilizing a variety of tools, techniques, and visual analyses to make the data and results understandable to readers and stakeholders for future deployments and monitoring.

6.2. Recommendations for Future Work

Throughout the experiments conducted in this project, several predictive models were developed, and their performance was evaluated to select the best-fitted models for deployment. However, there are opportunities for further improvement and expansion in future research endeavours.

Firstly, it is recommended to enhance the dataset by identifying and collecting additional measures that may have significant correlations with the response variable. This could include incorporating market sentiment indicators, economic indicators, or news sentiment scores to provide a more comprehensive understanding of stock price movements and improve the accuracy of predictive models.

Additionally, exploring new statistical learning algorithms beyond the four types of regression models used in this project could lead to more accurate results. Comparing the performance of different algorithms and selecting the most suitable model for each company's analysis could enhance the predictive capabilities of the models.

Moreover, future studies could consider incorporating historical financial indicator values for previous years during the exploratory phase. Analysing trends and patterns over time for each company could provide insights into their long-term sustainability and help in making more informed investment decisions.

Lastly, it is essential to recognize that predictive analytics using machine learning and data mining require sufficient data and measures to build accurate models. Therefore, ensuring the availability of comprehensive datasets and continually updating them with new information is crucial for achieving reliable predictions in future research efforts.

6.3. Project Contributions

This project has made several contributions to the field of financial analysis and investment decision-making:

- **Integration of Multiple Financial Indicators:** The project brought together a diverse range of financial indicators, catering to different types of investors with varying risk tolerances and preferences. By considering a broader spectrum of factors, investors can make more informed decisions regarding their investments.
- **Comprehensive Analytical Approach:** The project employed a combination of statistical tools, data mining techniques, methodologies, and machine learning algorithms to conduct a comprehensive and efficient predictive analysis. This approach ensured a thorough examination of the data and enabled the development of accurate predictive models.

- **Utilization of Visualization Tools:** Visualization graphics and plots were extensively used to explore the data, identify meaningful patterns, and communicate the analysis results effectively. These visualizations not only enhanced the understanding of complex datasets but also facilitated data-driven decision-making for investors.
- **Enhanced Investment Decision-Making:** By implementing procedures and actions derived from data analysis, the project aimed to encourage investors to consider a wider range of financial indicators before making investment decisions. This approach ultimately helps investors improve their investment gains by making more informed and strategic choices.

Overall, the project contributes to advancing the understanding of financial analysis methodologies and provides valuable insights for investors seeking to optimize their investment strategies in dynamic markets.

References

Almumani, M. A. Y., and Almazari, A. A., (2021). THE EFFECT OF MAJOR FINANCIAL INDICATORS ON MARKET CAPITALIZATION IN JORDANIAN FINANCIAL COMPAINES LISTED IN AMMAN STOCK EXCHANGE. International Journal of Economics, Commerce and Management. United Kingdom ISSN 2346 0386 Vol. IX, Issue 6, June 2021.

Andersen, T.G. et al. (1999). Forecasting financial market volatility: sample frequency vis-à-vis forecast horizon. J. Empir. Finance (1999)

Astuti, W., (2021). A Literature Review of Net Profit Margin. Accounting Project Program, Faculty of Economic, Universitas Widya Mataram. Social Science Studies Vol. 1 No. 2 (2021) https://doi.org/10.47153/sss12.2262021

Berinato, S., (2016). Visualizations that Really Work. Analytics and Data Science. Harvard Business Review June 2016 Issue (pp.92-100)

Candanedo, I.S., Nieves, E.H., González, S.R., Martín, M.T.S., Briones, A.G. (2018). Machine Learning Predictive Model for Industry 4.0. In: Uden, L., Hadzima, B., Ting, IH. (eds) Knowledge Management in Organizations. KMO 2018. Communications in Computer and Information Science, vol 877. Springer, Cham. https://doi.org/10.1007/978-3-319-95204-8_42

Chun, D., Cho, H., and Ryu, D., (2020). Economic indicators and stock market volatility in an emerging economy. Economic Systems Volume 44, Issue 2, June 2020, 100788. https://doi.org/10.1016/j.ecosys.2020.100788

Deborah, M. and Pyrczak, F., (2023). Making Sense of Statistics: A Conceptual Overview. Eighth edition published 2023 by Routledge 605 Third Avenue, New York, NY 10158 and by Routledge 4 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

DJALAL, R., (2020). Predictive Analytics for a Binary Classification Problem: Customers retention forecast in the banking sector. Master of Data Analytics. Sprint 2020

Faiteh, A., and Aasri, M. R., (2022). Accounting Beta as an Indicator of Risk Measurement: The Case of the Casablanca Stock Exchange. Research Laboratory in Economic

Competitiveness and Managerial Performance, Faculty of Law, Economic and social Sciences – Souissi, Mohammed V University, Rabat 10112, Morocco. Risks 2022, 10(8), 149

Fama, E. F., (1965). The Behaviour of Stock-Market Prices. The Journal of Business Vol. 38, No. 1 (Jan. 1965), pp. 34-105 (72 pages). Published By: The University of Chicago Press

Healy, K., (2019). Data Visualization: A PRACTICAL INTRODUCTION. PRINCETON UNIVERSITY PRESS PRINCETON AND OXFORD. ISBN 978-0-691-18161-5. ISBN (pbk.) 978-0-691-18162-2

Helen, R., Mikis, S., and Urszula P. (2015). Introduction to Time Series Analysis and Forecasting. September 25, 2015

Hossain, A., (2023). Volatility Modelling: Dealing with uncertainty in business. MA7008 Financial Mathematics – Autumn 2023-24. London Metropolitan University

https://doi.org/10.3390/risks10080149

https://en.wikipedia.org/wiki/Cross-industry_standard_process_for_data_mining

https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average

https://en.wikipedia.org/wiki/Microsoft_Excel

https://en.wikipedia.org/wiki/Microsoft_Power_BI

https://en.wikipedia.org/wiki/Nasdaq

https://en.wikipedia.org/wiki/Python_(programming_language)

https://en.wikipedia.org/wiki/S%26P_500

Hubenova, S., (2023). CS7079 Data Warehousing and Big Data. REVIEW OF RDBMS AND LANGUAGES. London Metropolitan University 2023

Indraswono, C., (2021). Traditional and Modern Analysis Performance Indicators: Evidence from New York Stock Exchange. STIE YKPN, Yogyakarta, Indonesia. KINERJA Volume 25, No. 1, 2021 Page. 64-78.

Kienitz, J., and Wetterau, D., (2012). Financial Modelling: Theory, Implementation and Practice with MATLAB Source. WILEY A John Wiley & Sons, Ltd., Publication

Lee, C. C., Wang, C. W., and Ho, S. J., (2020). Financial inclusion, financial innovation, and firms' sales growth. International Review of Economics & Finance. Volume 66, March 2020, Pages 189-205. https://doi.org/10.1016/j.iref.2019.11.021

Meryana and Setiany, E., (2021). The Effect of Investment, Free Cash Flow, Earnings Management, and Interest Coverage Ratio on Financial Distress. Journal of Social Science. Printed ISSN: 2720-9938 | Electronic ISSN: 2721-5202 Vol. 2 No. 1. https://doi.org/10.46799/jss.v2i1.86

Musallam, S. R., (2018). Exploring the Relationship between Financial Ratios and Market Stock Returns. Eurasian Journal of Business and Economics 2018, 11(21), 101-116.

Post, F. H., Nielson, G., and Bonneau, G. P., (2002). Data Visualization: The State of the Art. KLUWER ACADEMIC PUBLISHERS Boston / Dordrecht / London. ISBN 1-4020-7259-7 SECS 713

Prado, M. L. D., (2018). ADVANCES IN FINANCIAL MACHINE LEARNING. WILEY ISBN 978-1-119-48208-6 JWBT2318-Praise JWBT2318-Marcos

Raza, S., Baiqing, S., Kay-Khine, P., and Kemal, M. A., (2023). Uncovering the Effect of News Signals on Daily Stock Market Performance: An Econometric Analysis. Int. J. Financial Stud. 2023, 11(3), 99; https://doi.org/10.3390/ijfs11030099

Sharma, A., Modak, S., and Sridhar, E., (2019). Data Visualization and Stock Market and Prediction. International Research Journal of Engineering and Technology (IRJET). Volume: 06 Issue: 09 | Sep 2019. e-ISSN: 2395-0056 p-ISSN: 2395-0072

Solis, I. M., (2023). CC7184 Data Mining and Machine Learning- Overview of data mining. London Metropolitan University. Spring 22-23

Sondakh, R., (2019). THE EFFECT OF DIVIDEND POLICY, LIQUIDITY, PROFITABILITY AND FIRM SIZE ON FIRM VALUE IN FINANCIAL SERVICE SECTOR INDUSTRIES LISTED IN INDONESIA STOCK EXCHANGE 2015-2018 PERIOD. Accounting Profession Program, Economics and Business Faculty, Sam Ratulangi University, JI. Kampus Bahu, Manado, 95115, Indonesia.

Spronk, J., and Hallerbach, W., (1997). Financial modelling: Where to go? With an illustration for portfolio management. European Journal of Operational Research. Volume 99, Issue 1, 16 May 1997, Pages 113-125

Stasinopoulos, M., (2022). Statistical Modelling. London Metropolitan University. Feb 2023 – School of Computing and Digital Media Pages 7-15

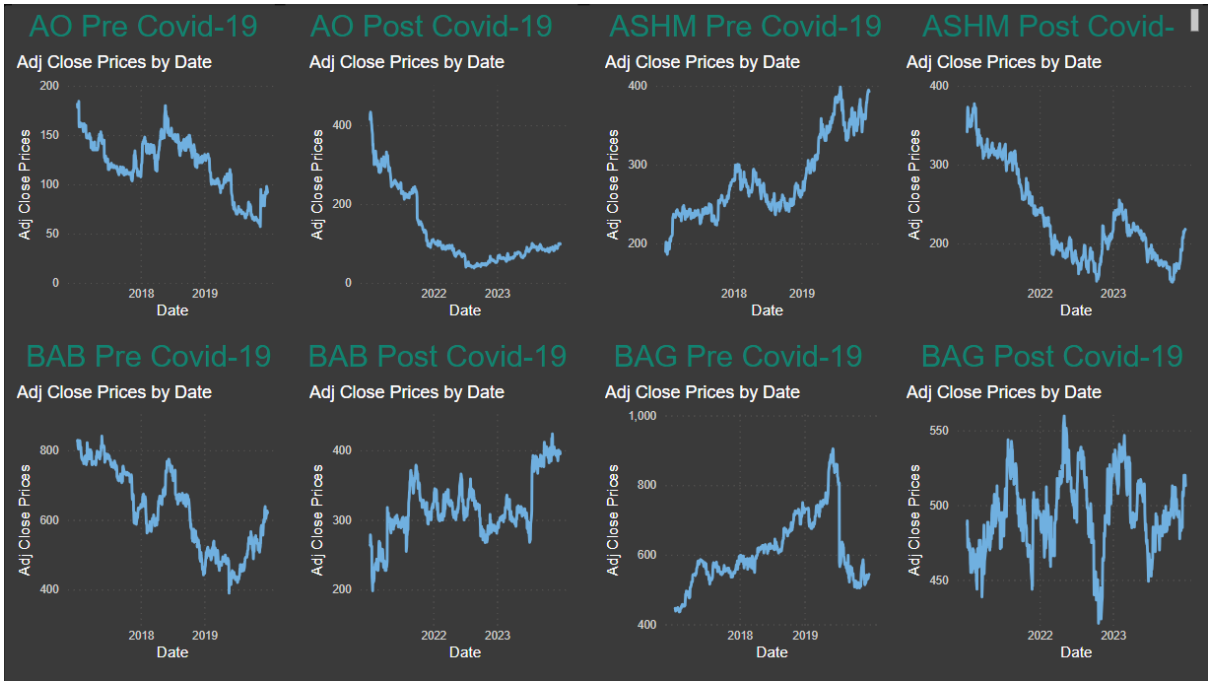
The CRISP-DM process model (1999), http://www.crisp-dm.org/

Tlemsani, I., (2020). Stock returns indicator: case of Tadawul. International Journal of Monetary Economics and Finance. Vol. 13, No. 1, pp 1-15. https://doi.org/10.1504/IJMEF.2020.105328

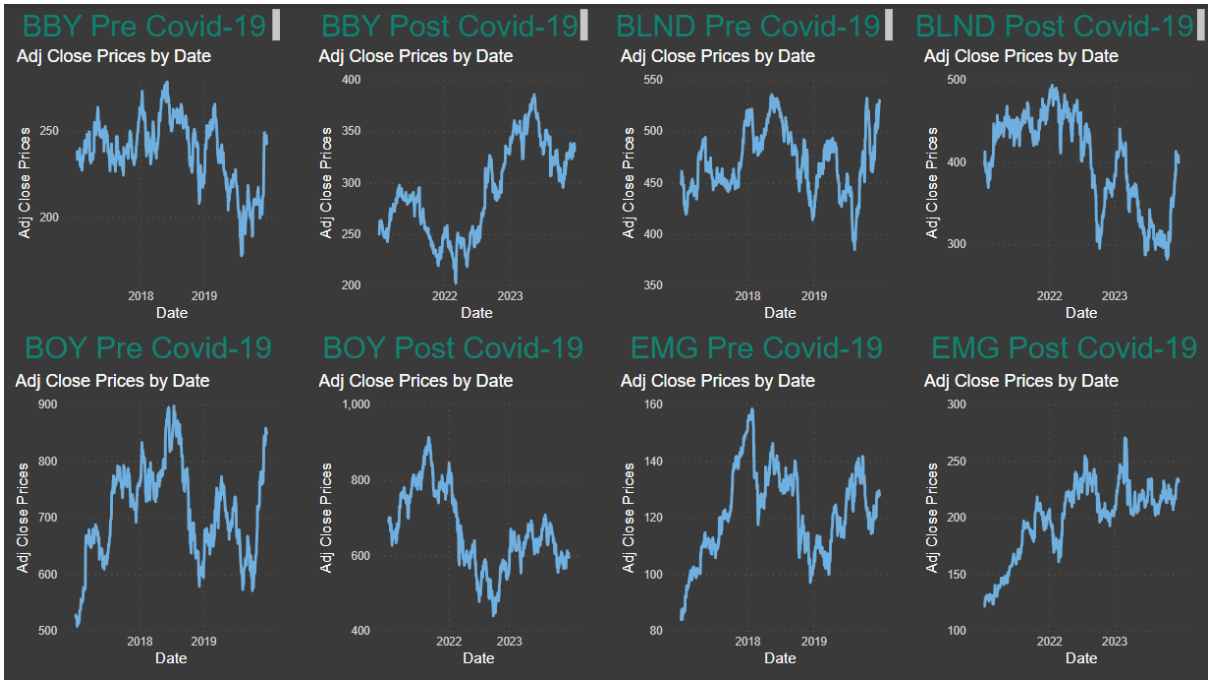
Tukey, J. W. (1977). Exploratory data analysis. Addison-Wesley. ISBN: 978-0201076165 Umamaheswari, S., Suresh, C.K., and Sampathkumar, S. (2021). An empirical project on the effect of financial accounting indicators towards stock market price volatility. World Review of Science, Technology and Sustainable Development, Vol. 18, No. 1.

Waskom, M. L., (2021). seaborn: statistical data visualization. Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss

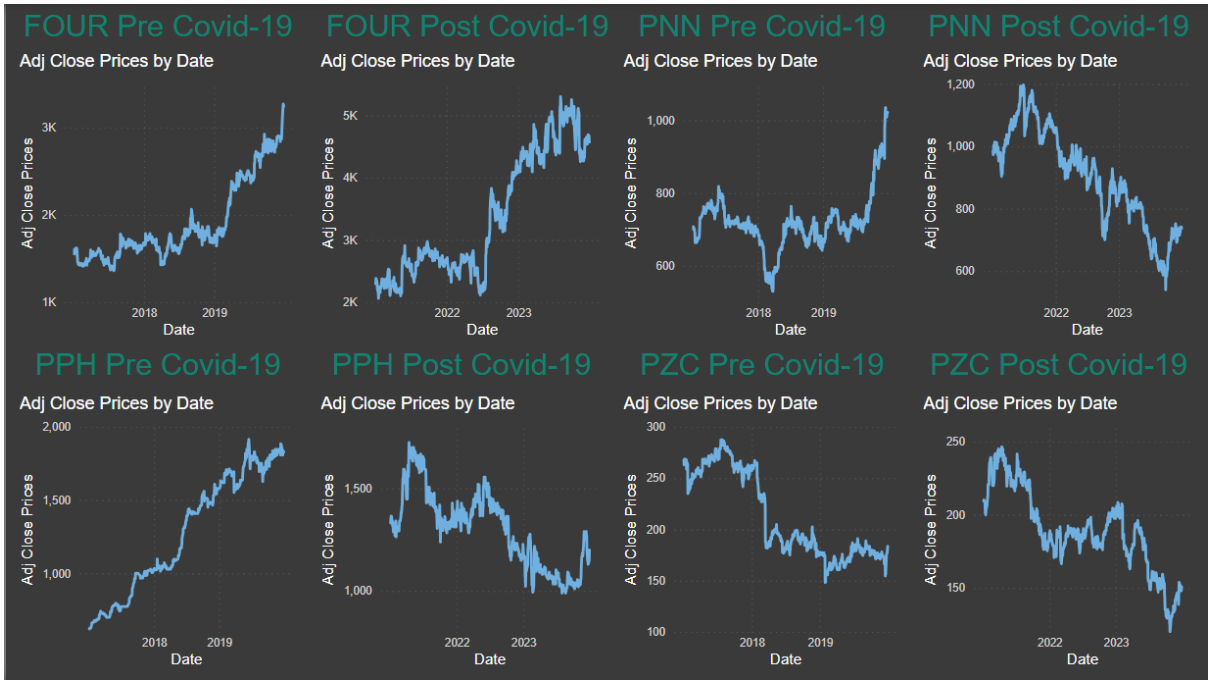
Appendix



Trends of adjusted close prices of 4 different companies before and after the pandemic.



Trends of adjusted close prices of 4 different companies before and after the pandemic.



Trends of adjusted close prices of 4 different companies before and after the pandemic.

The provided screenshots illustrate the trends in adjusted close prices for all 12 selected companies, showing their performance both before and after the COVID-19 pandemic. Among these companies, there are notable observations regarding their adjusted close prices. Specifically, companies such as FOUR and EMG exhibit an upward trend in their adjusted close prices following the COVID-19 period. This suggests that these companies have experienced growth or positive market sentiment during the post-pandemic period. While companies like BAB and PZC demonstrate a consistent decline in their adjusted close prices post-COVID-19. This downward trend indicates challenges or unfavourable market conditions that these companies may have faced in the aftermath of the pandemic. These trends in adjusted close prices provide valuable insights into the performance and resilience of the selected companies amidst the challenges posed by the COVID-19 pandemic.

```
# Data mining for supervised learning using Gradient Boosting
# Split the data into features and target variable
df = pd.DataFrame(BBY2)
X = df[['Open', 'High', 'Low', 'Close']]
y = df['Adj Close']

# Split the data into training and testing sets
X_train, X_test, y_train, y_test_gb = train_test_split(X, y, test_size=0.2,
                                                       random_state=42)

# Training the model with cross-validation
model = GradientBoostingRegressor()
scores = cross_val_score(model, X_train, y_train, cv=5,
                         scoring='neg_mean_squared_error')

# Train the Gradient Boosting Regressor model
model = GradientBoostingRegressor()
model.fit(X_train, y_train)

# Make predictions
gb_predictions = model.predict(X_test)

# Plot actual vs. predicted values
plt.figure(figsize=(10, 6))
plt.scatter(y_test_gb, gb_predictions, color='blue',
            label='Actual vs. Predicted')
plt.plot([min(y_test_gb), max(y_test_gb)], [min(y_test_gb), max(y_test_gb)],
         color='red', linestyle='--', label='Perfect Prediction')
plt.xlabel('Actual Adj Close')
plt.ylabel('Predicted Adj Close')
plt.title('Actual vs. Predicted Adj Close (BBY_GB)')
```

Automated Python script for Gradient Boosting model.

The provided screenshot illustrates an automated Python script utilized for the modelling phase of BBY pre-COVID-19, employing the Gradient Boosting model. Initially, the data was organized into features (‘X’), representing the independent variables used for prediction, and the target variable (‘y’), representing the variable to be predicted. Subsequently, the dataset underwent a division into training and testing subsets, followed by the execution of cross-validation. Predictions were then generated on the test set using the trained model, and a scatter plot was employed to visualize the comparison between actual and predicted close prices. This automated Python script was applied interchangeably across all 24 datasets selected for this project.