LUISS

Department of Economics and Business: Statistics

Comparative Analysis of Parametric and Non-Parametric Methods in Economics and Finance: A Study of Real-World Data

Supervisor: Prof. Marco Perone Pacifico

Candidate: Amanuel Teferi Mengiste

ID: 275441

Table of contents

Chapter 1: Introduction

In today's digital age, data analytics plays a pivotal role in understanding market dynamics in economics and finance. The rapid growth of data from financial markets, consumer transactions, and global economic indicators presents both opportunities and challenges. Effectively analyzing these complex datasets requires robust analytical tools that can handle their size and variability.

Traditionally, analysts have relied on parametric methods for their simplicity and strong theoretical foundations. However, these methods often depend on assumptions—such as normality and homoscedasticity—that are frequently violated in real-world data. As modern economic and financial datasets exhibit features like skewness, multi-modality, and heavy tails, the limitations of parametric approaches become evident.

This thesis explores whether non-parametric methods, with their flexibility and fewer assumptions, provide a more robust analytical framework than conventional parametric approaches. The primary research questions center on comparing the two methodologies for robustness, accuracy, and adaptability in economic and financial data analysis, as well as evaluating their practical implications in real-world contexts.

The thesis is structured into six chapters. Chapter 1 introduces the study, outlining its significance and objectives. Chapter 2 critically reviews the literature on parametric and non-parametric methods, discussing their applications, strengths, and limitations. Chapter 3 details the research design, data sources, and analytical procedures employed in the comparative study. Chapter 4 presents the empirical results, comparing the two methodologies. Chapter 5 interprets these results, exploring their significance, limitations, and possible directions for future research. Finally, Chapter 6 summarizes the major findings and contributions of the research.

Chapter 2: Literature Review

The literature review provides a comprehensive examination of the theoretical foundations and empirical studies relevant to the use of parametric and non-parametric methods. It begins by exploring the theoretical underpinnings of parametric methods, highlighting their foundational assumptions and applications in finance. It then examines non-parametric methods, discussing their advantages and limitations in addressing the complexities of real-world financial data. Additionally, the chapter reviews hybrid models that integrate both parametric and non-parametric approaches, offering a holistic view of their practical applications and benefits. By synthesizing existing research, this literature review aims to identify gaps in the current knowledge and lay the groundwork for the subsequent analysis in this study.

2.1 Parametric Methods: Theoretical Underpinnings

Parametric statistical approaches form the heart of econometric modeling because they rely on well-defined assumptions to simplify data analysis in economics and finance. These models aim to identify relationships between variables, predict outcomes, and test hypotheses based on the assumption that the data are normally distributed, errors have homoscedasticity, and relationships between variables are linear. Techniques such as linear regression, analysis of variance (ANOVA), and Pearson correlation are foundational in this approach.

The goal of a **linear regression model** is to estimate the relationship between a dependent variable and one or more independent variables, predicting how changes in the independent variables influence the dependent variable. The model assumes a linear relationship, expressed mathematically as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

In this equation, the regression equation relates the response variable Y_i to the predictor variable X_i, incorporating an intercept (β_0, a slope β_1, and an error term ϵ_i symbolizes the error terms, which are assumed to be independently and identically distributed (i.i.d.) following a normal distribution $N(0, \sigma^2)$. These foundational assumptions enable the derivation of several crucial statistical properties, such as the unbiasedness, efficiency, and consistency of estimators calculated via Ordinary Least Squares (OLS) methods (Fama, 1965).

ANOVA (Analysis of Variance) is utilized to determine if there are statistically significant differences between the means of three or more groups. Its goal is to test hypotheses about the impact of categorical independent variables on a continuous dependent variable, thereby providing insights into group effects under different market conditions or economic indicators. Mathematically, the ANOVA test statistic is based on the ratio of the mean square between groups (MSB) to the mean square within groups (MSW):

$$F = \frac{MSW}{MSB}$$

Where:

- **MSB** (Mean Square Between) is the variance between the group means.
- **MSW** (Mean Square Within) is the variance within each group.

The formulas for MSB and MSW are:

- **MSB** is calculated as the Sum of Squares Between (SSB) divided by the degrees of freedom between groups (k - 1):

$$MSB = \frac{SSB}{k-1}$$

- **MSW** is calculated as the Sum of Squares Within (SSW) divided by the degrees of freedom within groups (N - k):

$$MSW = \frac{SSW}{N-k}$$

Where:

SSB (Sum of Squares Between):

- Quantifies how much each group’s mean differs from the overall mean of all observations.

$$SSB = \sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2$$

SSB (Sum of Squares Between):

- Quantifies how much each group’s mean differs from the overall mean of all observations.

$$SSW = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_i)^2$$

ANOVA's F-statistic follows an F-distribution under the null hypothesis that all group means are equal. If the calculated F-statistic exceeds the critical value from the F-distribution, the null hypothesis is rejected, indicating significant differences among group means. In financial analysis, ANOVA is particularly useful for examining how different market conditions or economic indicators influence outcomes, offering a comprehensive framework for testing multiple group effects in a single analysis (Montgomery, 2013)

Despite the widespread use of these tests, the strict assumptions of parametric tests often conflict with real-world financial data features, such as skewness, kurtosis, heavy tails, and volatility clustering. Violation of these assumptions can lead to model misspecifications. For example, the normality assumption is crucial in the Black-Scholes model for pricing options, but it fails to account for the leptokurtic nature of asset returns during financial crises (Mandelbrot, 1963). These discrepancies between theoretical models and empirical realities can produce biased estimates and incorrect inferences, highlighting the need for caution in applying parametric models uncritically.

Pearson Correlation assesses the strength and direction of the linear relationship between two variables. It aims to quantify how one variable changes in response to another, assuming a linear relationship and normally distributed data. The Pearson correlation coefficient is given by:

$$r = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}}$$

Here, X_i and Y_i are the individual data points, and $\bar{X}$ and $\bar{Y}$ are the means of the X and Y data points, respectively. Pearson's correlation ranges from -1 to 1, where 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship (Kendall & Stuart, 1979).

As such, while parametric methods continue to offer valuable insights and frameworks in financial econometrics, their practical application must be approached with caution, acknowledging their limitations and the potential for significant discrepancies in real-world data analysis. This acknowledgment is crucial in paving the way for more robust, flexible statistical techniques that can accommodate the complex and often unpredictable nature of economic and financial data.

2.2 Non-Parametric Methods: Advantages and Applications

Nonparametric methods distinguish themselves from the parametric approach in that they analyze data without any structured, predefined model. This characteristic makes them very flexible and robust when dealing with complex and often irregular data, common in many economics and finance applications.

2.2.1 Core Mathematical Frameworks and Techniques

Nonparametric statistics do not stipulate any fixed form of distribution for the data and therefore avoid all the pitfalls of model assumptions, which are often incorrect and lead to bias or misleading results. This section discusses a few important nonparametric techniques, enriched with mathematical formulations to explain how they apply.

Spearman Rank Correlation

This assesses the monotonic relationship between two variables and does not assume normality. It is given by:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} \; d_i^2}{n(n^2 - 1)}$$

Where:

- d_i is the difference between the ranks of the i-th data point in the two variables,
- n is the number of data points.

Spearman's correlation also ranges from -1 to 1, like Pearson's correlation, but it measures the strength and direction of the monotonic relationship (Mandelbrot, 1963).

Kruskal-Wallis Test

This method is used to compare medians across three or more independent groups. It assesses whether there is a statistically significant difference in the central tendency (medians) of these groups, particularly when the data distributions are not normal. Unlike ANOVA, which assumes normality, the Kruskal-Wallis test robustness to outliers and heavy-tailed distributions is particularly useful for non-normally distributed data and is given by:

$$H = \frac{12N(N+1)\sum_{i=1}^{k} \frac{R_i^2}{n_i}}{N(N+1)} - 3(N+1)$$

Where:

- $\boldsymbol{k}$ is the number of groups,
- $\boldsymbol{N}$ is the total number of observations,
- $\boldsymbol{R_i}$ is the sum of ranks for the i-th group,
- $\boldsymbol{n_i}$ is the number of observations in the i-th group.

If the calculated H statistic exceeds the critical value from the chi-square distribution with $k-1$ degrees of freedom, the null hypothesis that the samples originate from the same distribution is rejected (Mann & Whitney, 1947).

Kernel Density Estimation (KDE)

KDE is one of the most basic tools in nonparametric analysis, used to estimate the probability density function of a random variable on a continuous set. The estimation is given, without assuming an underlying distribution, by:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

Where:

- $\boldsymbol{n}$ is the number of data points,
- $\boldsymbol{X_i}$ are the data points,
- $\boldsymbol{K}$ is the kernel function, typically a Gaussian, which smooths the data points over a range defined by h, the bandwidth.
- $\boldsymbol{h}$ significantly influences the estimator's bias and variance; choosing h optimally is crucial for obtaining a reliable estimate.

Bootstrapping

Bootstrapping is a resampling method that involves repeatedly drawing samples from a data set to estimate a population parameter. It enhances the reliability of predictions by providing empirical estimates of sampling distributions. The mathematical framework used involves the following steps:

1. **Resampling:** From an original dataset of size n, create a large number of bootstrap samples (typically 1000 or more), each of size n, by sampling with replacement.
2. **Computation of Statistic:** For each bootstrap sample, compute the desired statistic (e.g., the mean, median, or regression coefficient).
3. **Estimation of Distribution:** The distribution of these computed statistics across all bootstrap samples is used to approximate the sampling distribution of the statistic.

4. **Confidence Intervals:** Confidence intervals for the statistic can be derived from the empirical distribution of the bootstrap samples. For example, the 95% confidence interval is typically obtained by taking the 2.5th and 97.5th percentiles of the bootstrap distribution.

Mathematically, if we have a statistic *T* (e.g., a regression coefficient) based on the original sample *X*, we define it as $T = g(X)$, where *g* is a function. The bootstrap process generates *B* bootstrap samples (X^*_1, X^*_2... X^*_B), computes T^*_b for each sample, and estimates the standard error of *T* as:

$$SE_{\text{boot}} = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B} (T_b^* - \bar{T}^*)^2}$$

Where $\bar{T}$ * is the mean of the bootstrap statistics:

$$\bar{T}^* = \frac{1}{B}\sum_{b=1}^{B} T_b^*$$

2.2.2 Practical Applications in Finance and Economics

Non-parametric methods in economics and finance serve as indispensable tools for analyzing intricate and irregular datasets, offering flexibility and robustness that surpass the constraints of predefined models. These methods play a crucial role in various applications, providing insightful alternatives to traditional parametric approaches.

Spearman Rank Correlation evaluates the monotonic relationship between variables without assuming normality. This method is particularly valuable in financial analysis, where it uncovers non-linear associations that might be overlooked by parametric techniques (Mandelbrot, 1963).

Kruskal-Wallis Test extends the capabilities of the Mann-Whitney U test by assessing whether samples across multiple groups originate from the same distribution. This makes it effective for analyzing financial data that do not adhere to normal distributions, such as comparing returns or metrics across different sectors or time periods (Mann & Whitney, 1947).

Kernel Density Estimation (KDE) aids in visualizing and understanding the distributional characteristics of financial variables, which is crucial for risk management and derivative pricing (Scott, 1992).

Bootstrapping enhances the reliability of statistical conclusions by resampling from the original dataset to estimate parameters and their variability. In finance, this technique is instrumental for assessing the stability and uncertainty of financial models and predictions, providing critical insights into the robustness of statistical findings (Efron & Tibshirani, 1993).

Together, these non-parametric methods empower analysts and researchers in economics and finance to navigate the complexities of real-world data more effectively. They offer reliable tools for risk assessment, anomaly detection in market behavior, and comprehensive validation of statistical models in dynamic financial environments.

2.2.3 Expanding the Relevance in Empirical Analysis

While non-parametric methods offer robust flexibility and fewer assumptions, they require careful handling, particularly in selecting parameters like the bandwidth in KDE, which can dramatically influence the analysis's outcome. Advanced methods for bandwidth selection, such as cross-validation or plug-in approaches, help mitigate these issues by optimizing the balance between bias and variance in the estimates (Pedregosa et al., 2011).

Non-parametric methods also benefit from robust computational tools. The use of Python and libraries such as Pandas, NumPy, SciPy, and Scikit-learn facilitates efficient data manipulation, numerical calculations, and the application of complex statistical techniques (McKinney, 2010).

2.3 Conclusion

In conclusion, the exploration of parametric and non-parametric methods in Chapter 2 reveals fundamental insights into their roles within economic and financial analysis. Parametric approaches, anchored in assumptions like normality and linearity, have historically provided structured frameworks such as linear regression and ANOVA, essential for interpreting relationships in economic data. The limitations underscore the need for cautious application and exploration of alternative methodologies capable of accommodating the nuances of economic and financial datasets.

Conversely, non-parametric methods emerge as flexible alternatives, bypassing strict assumptions about data distribution. Techniques like Spearman rank correlation and kernel density estimation offer robust tools for analyzing irregular datasets prevalent in finance. These methods enhance predictive accuracy and model validation by capturing non-linear associations and complex distributional characteristics without imposing predefined structures.

Chapter 3: Methodology

3.1 Introduction

The methodology chapter provides a systematic overview of the research design, data collection, preparation, and analysis methods used to compare the effectiveness of parametric and non-parametric approaches in financial data analysis. It ensures reproducibility and transparency by detailing the research framework, data sources, and preparation steps, while also outlining the analytical methods employed. The chapter further explores hybrid models that combine both approaches to enhance predictive accuracy and robustness, laying the foundation for a rigorous analysis and interpretation of results in the following chapters.

3.2 Research Design

The research design of this study incorporates a methodological framework to evaluate and compare the effectiveness of statistical methods in analyzing financial data. The exploratory and descriptive nature of the study aims to uncover patterns, relationships, and statistical behaviors within both simulated and real-world financial datasets. By utilizing both data types, the research assesses the robustness and applicability of various statistical techniques across controlled and dynamic market conditions.

A key aspect of the research design is the structured comparison of parametric and non-parametric methods under different scenarios, considering varying distributions, skewness levels, and outliers. Simulated data, crafted to mimic specific statistical traits, and real-world financial data are used to ground the findings in both theory and practical relevance. Additionally, the design includes a systematic evaluation of correlation measures, statistical tests, and hybrid models to offer actionable insights into the strengths and limitations of each method. This structured approach enhances the reliability of the findings while contributing to the broader methodological discourse in financial data analysis.

3.3 Data Collection

In this section, we detail the datasets utilized in the study, categorized into simulated, real-world, and cross-sectional data for comprehensive analysis.

3.3.1 Simulated Data

Simulated data are crucial for rigorously testing statistical methods under controlled conditions. Using Python's NumPy library, simulated datasets were carefully crafted to exhibit specific statistical properties such as mean, variance, skewness, and kurtosis. These datasets simulate various financial scenarios, ensuring robust evaluations of methodological efficacy across diverse data distributions.

Simulation Studies:

Simulation studies are crucial for understanding stock return behaviors under various conditions. This involves running simulations over multiple iterations and introducing outliers to study their impact on statistical measures. The steps include:

- **Simulating Returns:** Returns for two independent stocks are simulated over 1000 days using Python and libraries like NumPy. This process is repeated for 1000 iterations to gather a robust set of results.
- **Introducing Outliers:** Some simulations include outliers to study their impact on Pearson and Spearman correlation coefficients. Outliers are introduced by multiplying returns on randomly selected days by a significant factor (e.g., 50) to simulate extreme market conditions.

The purpose of these simulation studies is to provide insights into the performance of statistical methods under varying conditions, including the presence of outliers and extreme market behaviors. By utilizing Python's NumPy library, the simulations are implemented with a focus on accuracy and reproducibility, which is essential for evaluating the robustness of these methodologies

3.3.2 Real-World Data

Real-world financial data sourced from Yahoo Finance offer insights into the applicability of statistical methods in actual market conditions. The dataset includes historical price data for multiple companies, comprising daily closing prices, trading volumes, and timestamps over a specified period. This data captures the dynamic nature of financial markets, incorporating real-time fluctuations and external influences that affect stock prices.

The purpose of this analysis is to validate and apply statistical methods in realistic market scenarios by using real-world financial data. This thorough analysis offers valuable insights into the practical performance of statistical techniques in the context of actual financial markets.

3.3.3 Cross-Sectional Data

Cross-sectional data analysis extends the study's scope by examining variation ratios among stock prices across different countries—Germany, Italy, and the USA. These variation ratios are computed as changes in stock prices over a specified period, enabling comparative insights into market behaviors and economic conditions across diverse geographical regions.

Data were gathered from Germany, Italy, and the USA to analyze variation ratios among companies, reflecting diverse market dynamics. These variation ratios highlight unique market characteristics and regulatory environments in each country. This approach allows for a deeper understanding of how different markets operate and how statistical techniques can be applied in various international contexts.

3.4 Analytical Methods

In this section, we detail the analytical methods and statistical tests employed to evaluate relationships and compare variation ratios across different datasets.

3.4.1 Correlation Analysis

Correlation analysis plays a crucial role in examining relationships between financial variables, employing both Pearson and Spearman correlation coefficients:

- **Pearson Correlation:** Pearson correlation measures the linear association between two variables, indicating the strength and direction of their linear relationship. In this study, Pearson correlations were computed for both simulated and real-world financial data. For simulated data, Pearson correlation coefficients were analyzed to assess linear dependencies between variables with controlled statistical properties. In real-world data, such as historical prices sourced from Yahoo Finance, Pearson correlation coefficients were used to evaluate linear relationships among daily closing prices and trading volumes over time.
- **Spearman Correlation:** Spearman correlation evaluates monotonic relationships between variables, making it robust against outliers and non-linear associations. In our analysis, Spearman correlation coefficients were utilized to identify and measure non-linear monotonic relationships. This method was particularly valuable when assessing relationships in scenarios where data distributions deviated from normality or included outliers. The use of Spearman correlation

provided insights into the strength and direction of relationships that Pearson correlation might overlook, enhancing the study's robustness in evaluating financial data.

A comparative analysis was conducted to contrast the performance of Pearson and Spearman correlations under different conditions, such as data with and without outliers. By examining these correlations under varied conditions, the study offered insights into the suitability of each method for different types of financial data analysis.

3.4.2 Comparative Tests

Comparative tests were employed to evaluate variation ratios among different groups of companies:

- **Kruskal-Wallis Test:** The Kruskal-Wallis test, a non-parametric method, was utilized to compare variation ratios across companies from Germany, Italy, and the USA. This test was chosen because it does not assume normality or homogeneity of variances, making it suitable for datasets where these assumptions may not hold.
- **ANOVA (Analysis of Variance):** ANOVA, a parametric method, was employed to compare means of variation ratios under conditions where parametric assumptions were met. Specifically, ANOVA was used to analyze variation ratios across different groups of companies when data exhibited normal distribution and homogeneity of variances. This method enabled a detailed examination of mean differences among groups, complementing the insights gained from non-parametric tests like the Kruskal-Wallis test.

Additionally, Kernel Density Estimation (KDE) was employed to visually represent the distribution of variation ratios for companies from Germany, Italy, and the USA. This technique provided a comprehensive overview of the density characteristics within each group, illustrating how variation ratios were distributed across different economic regions. By visually inspecting KDE plots, the study enhanced its understanding of the underlying data distributions, thereby supporting the findings from comparative statistical tests like the Kruskal-Wallis and ANOVA.

3.5 Hybrid Models

This section details the development and application of hybrid models that integrate linear regression and bootstrapping techniques for enhanced predictive accuracy and robustness in financial data analysis.

Linear Regression

Linear regression was employed to model and capture the linear trends observed in stock prices within the study. The purpose of this approach was to establish a linear relationship between dependent variables (e.g., stock prices) and independent variables (e.g., time or other relevant factors). By fitting linear regression models to historical price data sourced from Yahoo Finance, the study aimed to identify trends and patterns in stock price movements over time.

Bootstrapping

Bootstrapping is a resampling technique used to assess the variability and robustness of statistical estimates. The purpose was to enhance the reliability of predictions derived from linear regression models. This methodology involves generating multiple resamples from the original dataset to derive empirical estimates of sampling distributions. By applying it to historical price data, the study gained insights into the stability and consistency of regression coefficients and predictions. This technique proved particularly valuable in mitigating the impact of outliers and variability inherent in financial markets, thereby increasing the credibility of model predictions.

Integration of Linear Regression and Bootstrapping

The development of this hybrid model involved combining the predictive power of linear regression with the robustness conferred by bootstrapping. Initially, linear regression models were fitted to historical price data to capture underlying trends. Subsequently, bootstrapping was applied to these models to generate multiple resampled datasets, from which empirical distributions of model coefficients and predictions were derived. This integration allowed the hybrid model to not only identify linear trends in stock prices but also quantify the uncertainty associated with these predictions, offering more reliable insights into future price movements.

3.7 Software and Tools

This section outlines the software and tools utilized for data analysis and visualization throughout the study, emphasizing their roles in processing and interpreting financial data.

Python served as the primary programming language for data manipulation, analysis, and modeling. Its versatility and extensive libraries made it well-suited for handling complex financial datasets and implementing statistical techniques.

- **pandas:** pandas, a powerful data analysis library in Python, was instrumental in data manipulation and preprocessing tasks. It facilitated tasks such as data cleaning, transformation, and integration across different datasets, ensuring data consistency and readiness for analysis.
- **NumPy:** NumPy, a fundamental library for numerical computing in Python, provided essential functionalities for mathematical operations and array manipulations. It enabled the generation of simulated datasets with specific statistical properties, such as mean, variance, skewness, and kurtosis, crucial for testing methodological robustness.

- **matplotlib:** matplotlib, a comprehensive plotting library in Python, was utilized for generating visualizations that enhanced data interpretation and presentation. It produced a variety of plots, including histograms, box plots, scatter plots, and kernel density estimates (KDE), which illustrated data distributions, relationships, and trends effectively.
- **SciPy:** SciPy, another essential library for scientific computing in Python, complemented NumPy by offering advanced statistical functions and tests. It facilitated the implementation of statistical methods such as Pearson correlation, Spearman correlation, Kruskal-Wallis test, and ANOVA, enabling rigorous comparative analyses across different datasets.
- **Jupyter Notebooks:** Jupyter Notebooks provided an interactive computing environment that combined code execution, data visualization, and explanatory text in a single document. They were used for iterative data analysis, model development, and result interpretation, promoting transparency and reproducibility in the research process.
- **Yahoo Finance (API):** Yahoo Finance's API was utilized to access and retrieve real-world financial data, including historical price data for Tesla (TSLA). This data source provided crucial information for validating statistical methods and models against real market conditions.

These software and tools collectively facilitated comprehensive data analysis, rigorous statistical testing, and insightful visualization, enabling the study to derive meaningful conclusions and recommendations in the field of financial data analysis.

Chapter 4: Data Analysis

4.1 Introduction

This chapter explores the datasets and methodologies used to compare parametric and non-parametric methods in financial data analysis. It begins with an overview of the datasets, including simulated, real-world, and cross-sectional data, along with descriptive statistics and graphical representations. The chapter then focuses on the comparative performance of correlation measures and statistical tests, considering different scenarios such as the presence of outliers. Finally, it examines hybrid models that combine both approaches to improve predictive accuracy and robustness.

4.2 Data Overview

This section provides a detailed overview of the datasets used in this study. We present various descriptive statistics to understand the underlying properties of these datasets.

4.2.1 Simulated Data

Simulated datasets were generated to mimic different distributional characteristics crucial for testing method robustness. **Variables:** Mean, Variance, Skewness, Kurtosis

- **Type:** Continuous

	Mean	Standard Deviation	Quartiles	Skewness	Kurtosis
Normal Data	-0.045	0.987	[-0.698, -0.058, 0.607]	0.034	-0.047
Data with outliers	0.514	2.370	[-0.600, 0.0860, 0.778]	3.138	10.402

Table 1: Descriptive Statistics for Simulated Data

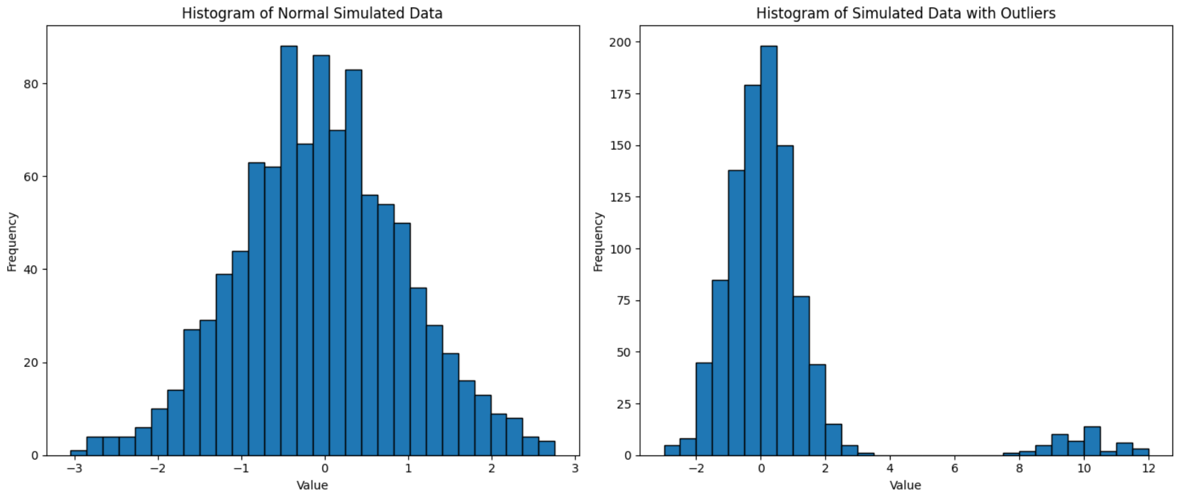


Figure 1: Histogram of Simulated Data with and without outliers

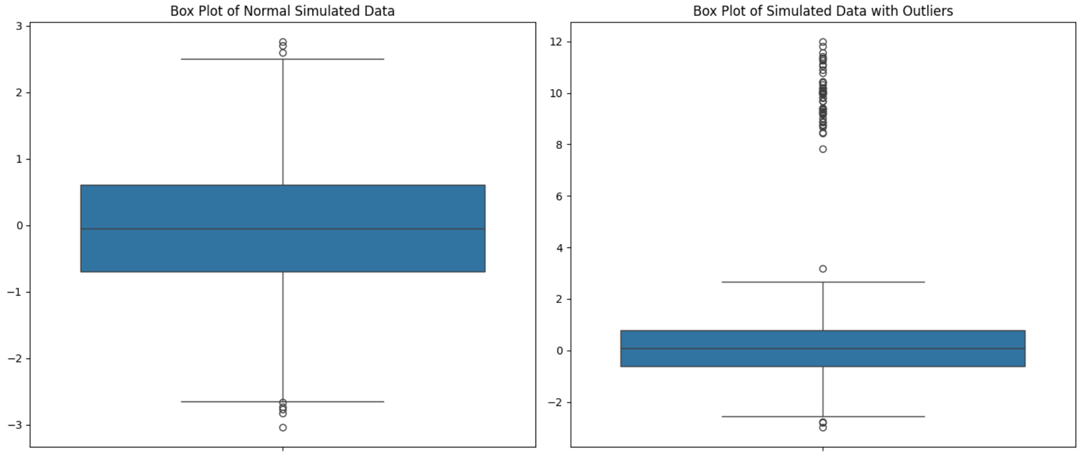


Figure 2: Box Plot of Simulated Data

The normal dataset has a mean of -0.045, standard deviation of 0.987, and quartiles of [-0.698, -0.058, 0.607], with slight skewness (0.034) and near-zero kurtosis (-0.047), indicating a roughly normal distribution. In contrast, the dataset with outliers shows a higher mean (0.514), greater standard deviation (2.370), and quartiles of [-0.600, 0.086, 0.778], with skewness (3.138) and kurtosis (10.402) reflecting a more dispersed, heavy-tailed distribution. Histograms and box plots confirm these characteristics, with the normal data showing a bell-shaped curve and symmetrical spread, while the outlier dataset displays greater spread and evident outliers.

4.2.2 Real-World Data

For real-world financial data, we sourced historical price data from Yahoo Finance. This dataset includes historical price data, trading volumes, and timestamps.

- **Type:** Time Series

Company Name	Ticker	Mean	Standard Deviation	Quartiles	Skewness	Kurtosis
Apple Inc.	AAPL	130.312	30.566	[115.739, 135.380, 150.705]	-0.635	-0.454
Amazon.com, Inc.	AMZN	142.455	27.856	[118.338, 154.467, 164.633]	-0.578	-1.018
The Coca-Cola Company	KO	55.223	5.916	[50.458, 55.135, 60.408]	-0.165	-0.806
PepsiCo, Inc.	PEP	152.788	16.786	[138.057, 148.745, 168.355]	0.088	-1.033
McDonald's Corporation	MCD	229.706	26.823	[212.228, 233.885, 250.412]	-0.548	-0.072
Starbucks Corporation	SBUX	94.174	15.411	[81.145, 91.270, 109.825]	0.032	-1.135

Table 2: Descriptive Statistics for Real-World Data

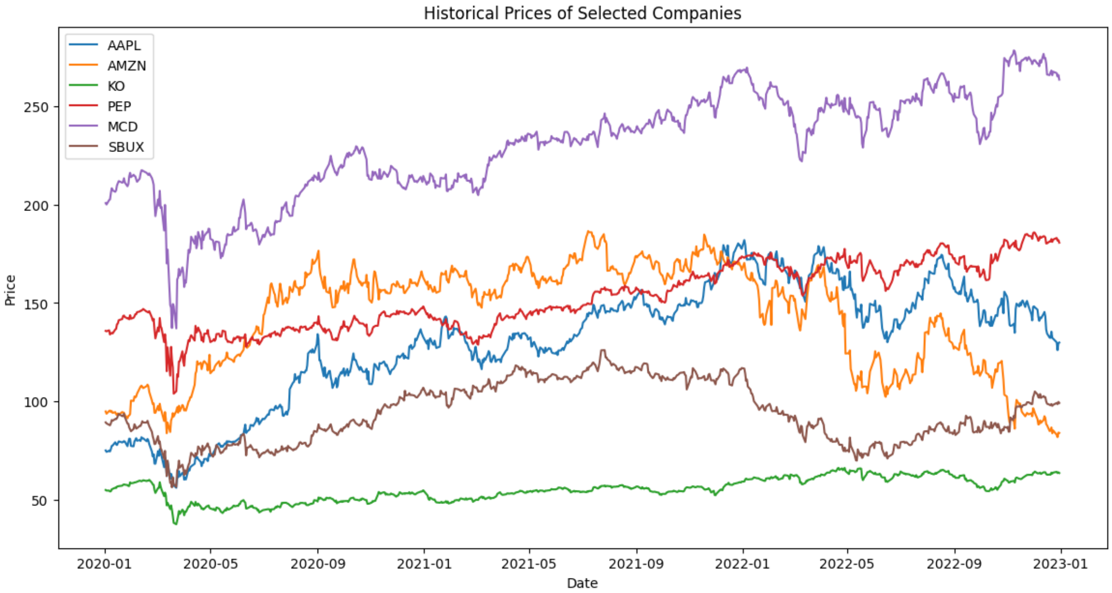


Figure 3: Time Series Plot of Historical Prices

The historical price data reveals key performance and volatility insights for selected companies. Apple (AAPL) shows moderate volatility with a mean of 130.312 and a standard deviation of 30.566, displaying a slight left skew and near-normal distribution. Amazon (AMZN) experiences greater price fluctuations, with a mean of 142.455, standard deviation of 27.856, and a long-left tail. Coca-Cola (KO) has lower volatility, with a mean of 55.223, standard deviation of 5.916, and a nearly symmetrical distribution. PepsiCo (PEP) presents stability with a mean of 152.788, standard deviation of 16.786, and a slight right skew. McDonald's (MCD) shows higher volatility with a mean of 229.706 and left-skewed distribution, while Starbucks (SBUX) has a mean of 94.174, moderate volatility, and slight right skew. The time series plots further highlight trends and anomalies, emphasizing the importance of robust analytical techniques for handling varied data distributions, which sets the stage for the comparative analysis of parametric and non-parametric methods.

4.2.3 Cross-Sectional Data

The cross-sectional data analysis focuses on the variation ratios of stock prices for companies from Germany, Italy, and the United States over a specified period. The variation ratio represents the percentage change in stock prices during the period, providing insights into market behavior and volatility across different regions.

- **Type:** Cross-Sectional

Companies	Mean	Standard Deviation	Quartiles	Skewness	Kurtosis
German	0.076	0.254	[-0.161, 0.062, 0.330]	0.044	1.745
Italian	0.046	0.135	[0.024, 0.057, 0.157]	-0.81	-0.649
US	0.371	0.376	[0.195, 0.417, 0.593]	-0.355	-1.206

Table 4: Descriptive Statistics for Cross-Sectional Data

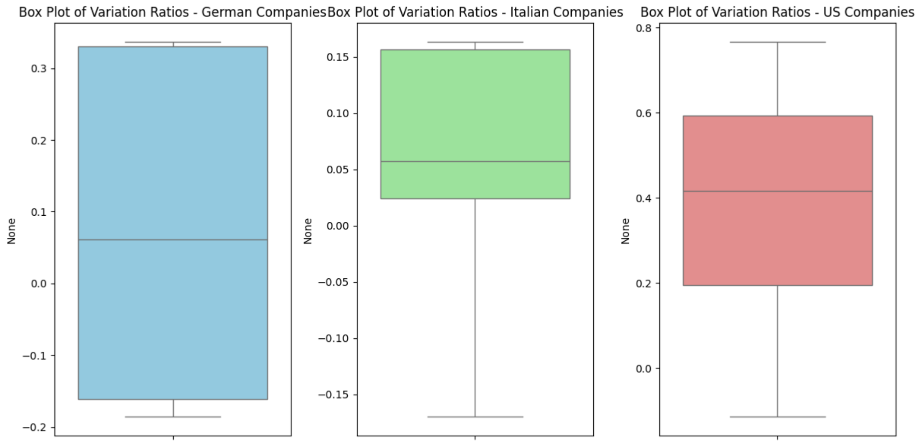


Figure 4: Box Plot of Variation Ratios

The descriptive statistics show that US companies have a much higher mean variation ratio (0.371) compared to German (0.076) and Italian (0.046) companies, indicating more substantial stock price changes in the US market. The US market also has greater variability, with a higher standard deviation (0.376) than the German (0.254) and Italian (0.135) markets. Box plots illustrate these differences, with US companies showing higher variation and volatility. Negative skewness in the Italian (-0.81) and US (-0.355) data suggests longer left tails, indicating occasional large negative stock price changes, while the German data is nearly symmetrical with a skewness of 0.044. Kurtosis further emphasizes these distinctions, with Italian (-0.649) and US (-1.206) companies having flatter distributions, while German companies exhibit positive kurtosis (1.745), indicating a more peaked distribution with potential outliers.

4.3 Comparative Analysis of Methods

This section compares the performance of parametric and non-parametric methods using both simulated and real-world data. We analyze Pearson and Spearman correlation coefficients, examining their behavior in scenarios with and without outliers. We also discuss the number of times p-values fall below the 0.05 threshold, indicating statistical significance.

4.3.1 Simulated Data: Analysis of Pearson and Spearman Correlation

Simulation without Outliers

In the absence of outliers, both Pearson and Spearman correlation coefficients revealed average values close to zero across 1,000 iterations, indicating no significant linear or monotonic relationships between the variables. This aligns with the expectation that randomly generated, independent data should not exhibit meaningful correlations.

Pearson Correlation:

- Mean correlation coefficient: **0.001**, indicating no significant linear relationship.
- Mean p-value: **0.506**
- Percentage of p-values below 0.05: **4.50%**, consistent with random chance (false positives).

Spearman Correlation:

- Mean correlation coefficient: **0.002**, indicating no significant monotonic relationship.
- Mean p-value: **0.506**
- Percentage of p-values below 0.05: **4.60%**, consistent with random chance.

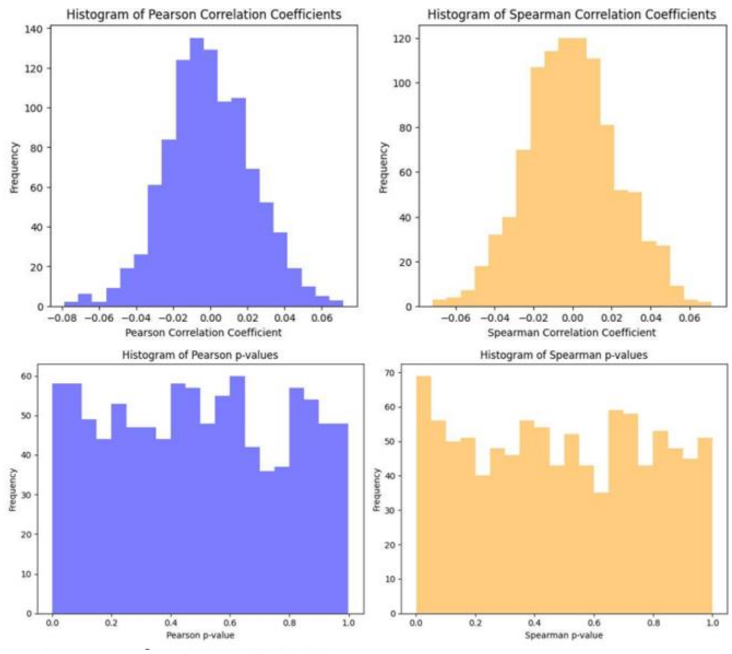


Figure 5: Histogram of Pearson and Spearman Correlation Coefficients for Simulated Data Without Outliers

These results underscore the consistency of both methods in detecting no substantial relationships when none exist, as the data is purely random. The low percentages of p-values below 0.05 further reinforce the conclusion that the occasional statistically significant results are likely due to random fluctuations rather than genuine correlations.

Figure 5 provides a visual representation of the distribution of Pearson and Spearman correlation coefficients, both clustering around zero, with narrow spreads that further confirm the absence of meaningful relationships in the data.

Simulation with Outliers

When outliers were introduced, Pearson correlation coefficients exhibited greater sensitivity, as reflected by a wider spread and higher mean value, while Spearman remained more stable.

Pearson Correlation:

- Mean correlation coefficient: **0.050**, slightly inflated by the presence of outliers.
- Mean p-value: **0.641**
- Percentage of p-values below 0.05: **0.80%**, indicating reduced sensitivity to detect significant relationships in the presence of outliers.

Spearman Correlation:

- Mean correlation coefficient: **0.003**, nearly unchanged, showing no significant monotonic relationship.
- Mean p-value: **0.506**
- Percentage of p-values below 0.05: **4.90%**, similar to the scenario without outliers, reflecting Spearman's robustness.

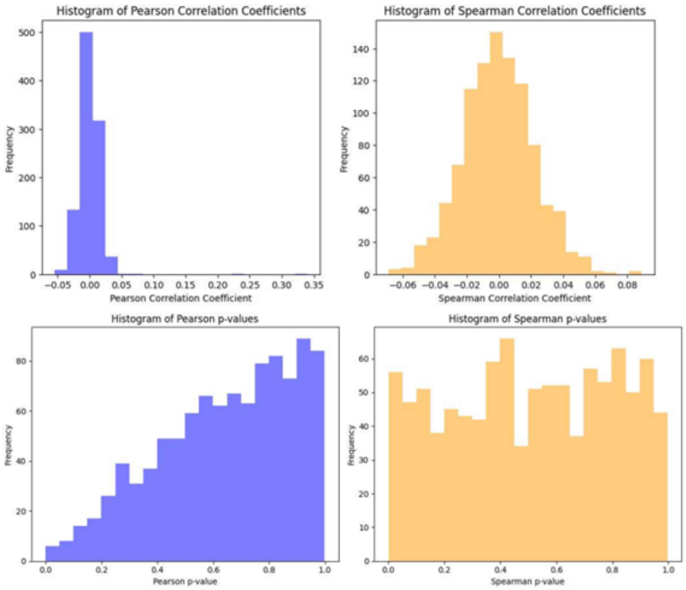


Figure 6: Histogram of Pearson and Spearman Correlation Coefficients for Simulated Data with Outliers

When outliers were added to the data, the Pearson correlation values increased. SThis is because the extreme values inflated the correlation. Pearson's average p-value rose to 0.641, and only 0.80% of the tests showed statistically significant results ($p < 0.05$). This means that the outliers made it harder for Pearson to detect meaningful relationships. These results show that Pearson correlation is easily affected by outliers, leading to unstable and less reliable conclusions when extreme values are present.

In contrast, Spearman correlation remained much more robust. The **mean Spearman correlation coefficient** was only **0.003**, nearly unchanged from the scenario without outliers, reflecting Spearman's insensitivity to extreme values. The **mean p-value** was **0.506**, and **4.90% of iterations** produced p-values below 0.05—numbers that are remarkably consistent with the results from the simulations without outliers. This stability highlights Spearman's resilience to outliers and confirms its suitability for datasets prone to extreme values.

4.3.2 Real-World Data: Analysis of Correlation between Stock Pairs

Next, we examine the Pearson and Spearman correlations for three pairs of real-world stocks: Apple (AAPL) vs. Amazon (AMZN), Coca-Cola (KO) vs. PepsiCo (PEP), and McDonald's (MCD) vs. Starbucks (SBUX). This analysis aims to provide insights into the relationships between these companies and assess the consistency of the correlation measures.

Apple (AAPL) vs. Amazon (AMZN):

The scatter plot shows a positive linear relationship.

- Pearson p-value: < 0.001
- Spearman p-value: < 0.001

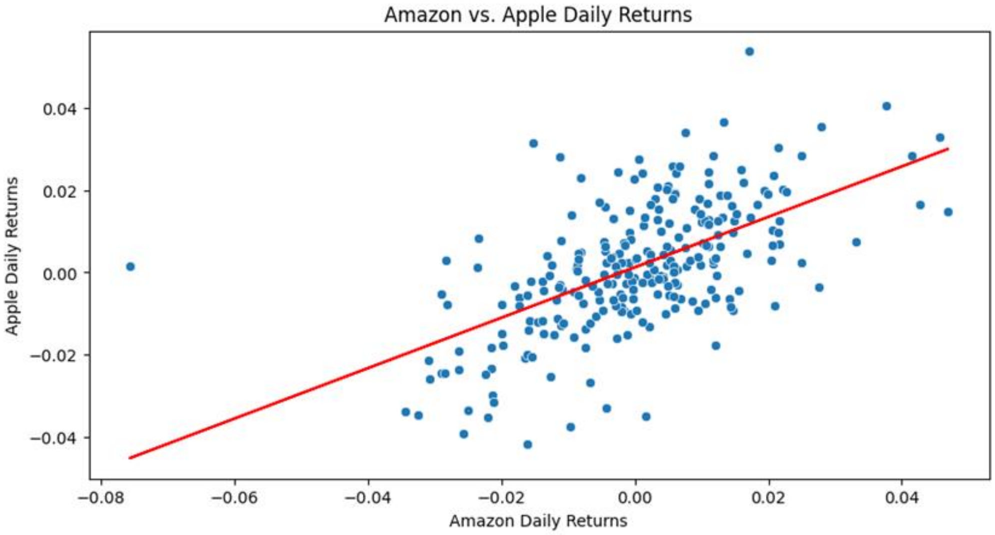


Figure 7: Scatter Plot and Linear Fit for AAPL vs. AMZN

Coca-Cola (KO) vs. PepsiCo (PEP):

The scatter plot shows a positive linear relationship.

- Pearson p-value: < 0.001
- Spearman p-value: < 0.001

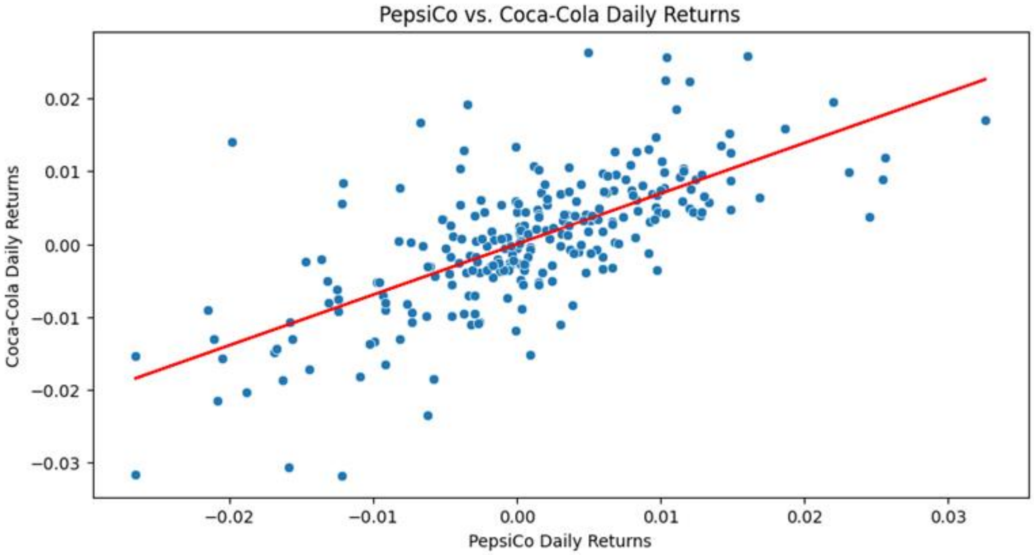


Figure 8: Scatter Plot and Linear Fit for KO vs. PEP

McDonald's (MCD) vs. Starbucks (SBUX):

The scatter plot shows a strong positive linear relationship.

- Pearson p-value: < 0.001
- Spearman p-value: < 0.001

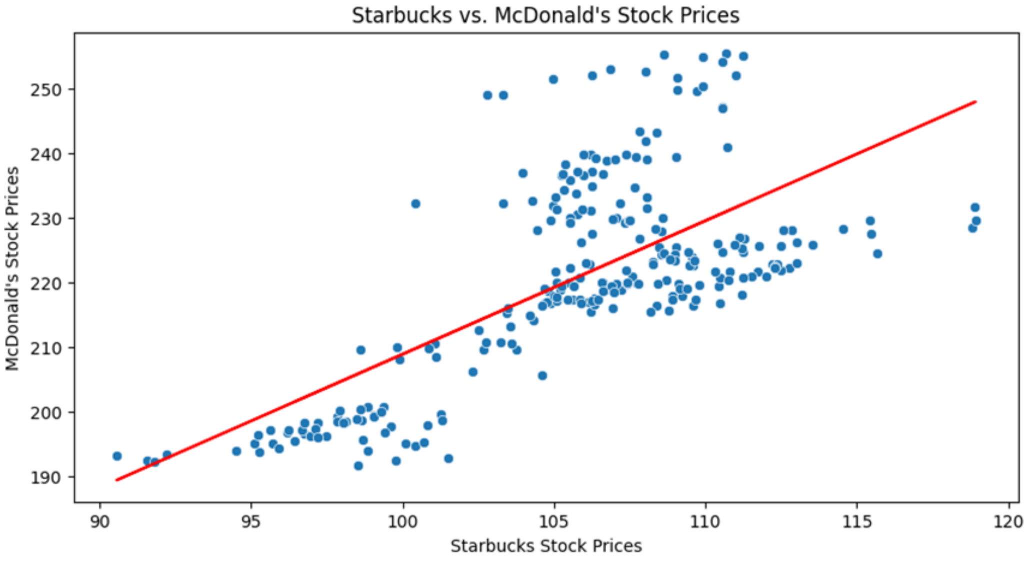


Figure 9: Scatter Plot and Linear Fit for MCD vs. SBUX

The analysis of real-world data shows that both Pearson and Spearman correlations consistently indicate strong positive relationships between the stock returns of the pairs analyzed. For AAPL vs. AMZN, KO vs. PEP, and MCD vs. SBUX, the scatter plots and linear fits visually support these correlations. The low p-values (< 0.001) for both Pearson and Spearman methods confirm that these relationships are statistically significant. This consistency suggests that both correlation measures are reliable for analyzing the relationships between stocks in the real world, where data is typically more complex and varied than in simulations.

4.3.3 Cross-Sectional Data: Comparing Kruskal-Wallis and ANOVA Tests

In this subsection, we compare the performance of the Kruskal-Walli's test, a non-parametric method, and the ANOVA test, a parametric method, using cross-sectional financial data. In addition to the Kruskal-Walli's test, Kernel Density Estimation (KDE) was employed to estimate the probability density functions of the variation ratios. This comparison aims to evaluate the effectiveness of these methods in detecting differences in variation ratios among groups of companies from different countries.

Kruskal-Wallis Test:

- **Test Statistic:** 1.874
- **P-value:** 0.392

The Kruskal-Walli's test yields a test statistic of 1.874 and a p-value of 0.392. This p-value indicates that there is no significant difference in the variation ratios among the groups of companies from Germany, Italy, and the US. The Kruskal-Walli's test, being a non-parametric method, is particularly useful when the data does not meet the assumptions of normality and homogeneity of variance required for parametric tests.

ANOVA Test:

- **Test Statistic:** 1.161
- **P-value:** 0.321

The ANOVA test results show a test statistic of 1.161 and a p-value of 0.321. Like the Kruskal-Walli's test, the ANOVA p-value indicates no statistically significant differences in the variation ratios among the groups. ANOVA is a robust parametric method when its assumptions are met, and its results here align with those of the Kruskal-Walli's test, reinforcing the conclusion.

Kernel Density Estimation (KDE):

Kernel Density Estimation (KDE) plots provide a visual representation of the distribution of variation ratios for companies from Germany, Italy, and the US. The KDE plots reveal distinct density characteristics for each group: moderate variation for German companies, low variation for Italian companies, and higher variation for US companies. These visual insights support the statistical test results, indicating no significant differences among the groups but highlighting the inherent distribution characteristics.

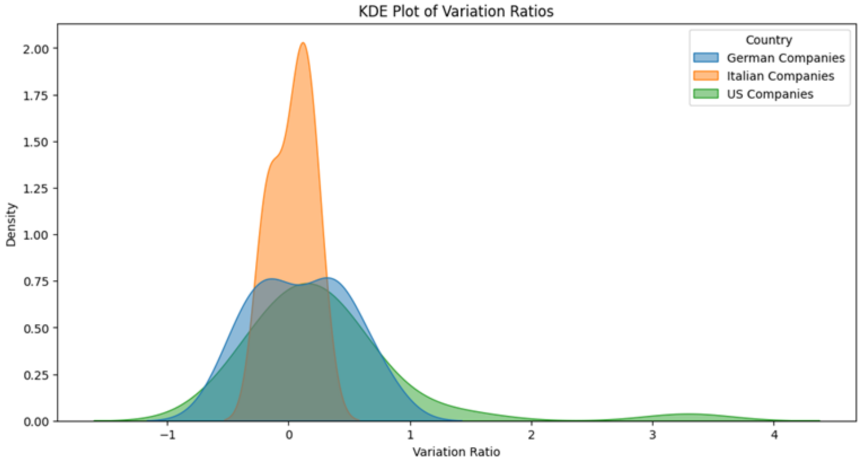


Figure 10: Kernel Density Estimation of Variation Ratios for German, Italian, and US Companies

The comparative analysis using the Kruskal-Wallis and ANOVA tests suggests that there are no significant differences in variation ratios among the companies from Germany, Italy, and the US. The p-values from both tests (0.392 for Kruskal-Wallis and 0.321 for ANOVA) indicate a lack of statistical significance. The KDE plots further provide a visual understanding of the data distributions, showing distinct density characteristics for each group without indicating substantial differences. This combination of statistical and visual analysis offers a comprehensive understanding of the variation ratios across different groups, confirming the robustness of the findings.

4.4 Hybrid Models

In this section, we explore the integration of parametric and non-parametric methods to leverage their strengths. Hybrid models combine the assumptions and computational efficiencies of parametric methods with the flexibility and robustness of non-parametric approaches. This section demonstrates how hybrid models can be applied to real-world financial data to enhance predictive accuracy and robustness.

4.4.1 Application

A hybrid approach combining linear regression (parametric) with bootstrapping (non-parametric) is used to predict stock prices. The parametric component captures the linear trend, while the non-parametric component accounts for distributional peculiarities and outliers.

4.4.2 Example and Explanation

In this section, we demonstrate the application of the hybrid model using historical stock price data from Apple Inc. (AAPL) sourced from Yahoo Finance. The aim is to predict stock prices by leveraging both linear regression and bootstrapping techniques, thereby enhancing the robustness and accuracy of the predictions.

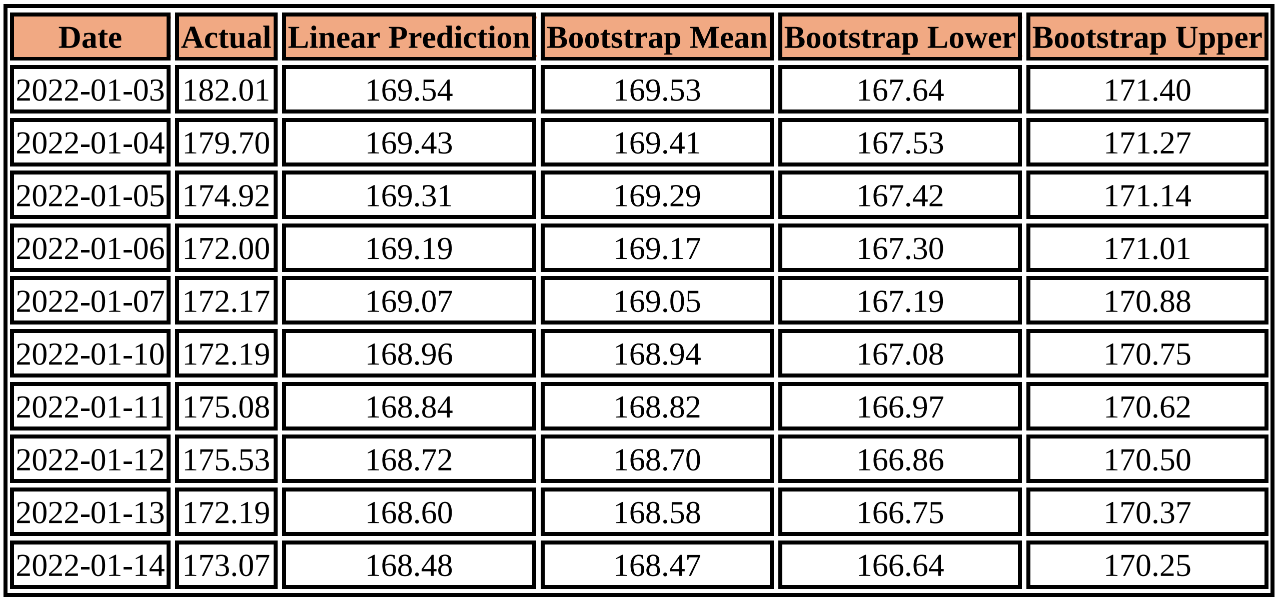
Objective: We aim to use a hybrid modeling approach to predict future stock prices by combining:

- **Linear Regression** to capture the underlying linear trend.
- **Bootstrapping** to account for distributional characteristics and potential outliers.

Data:

- The dataset includes the daily closing prices of Apple Inc. (AAPL) from January 1, 2021, to January 14, 2022.

4.4.3 Results



Date	Actual	Linear Prediction	Bootstrap Mean	Bootstrap Lower	Bootstrap Upper
2022-01-03	182.01	169.54	169.53	167.64	171.40
2022-01-04	179.70	169.43	169.41	167.53	171.27
2022-01-05	174.92	169.31	169.29	167.42	171.14
2022-01-06	172.00	169.19	169.17	167.30	171.01
2022-01-07	172.17	169.07	169.05	167.19	170.88
2022-01-10	172.19	168.96	168.94	167.08	170.75
2022-01-11	175.08	168.84	168.82	166.97	170.62
2022-01-12	175.53	168.72	168.70	166.86	170.50
2022-01-13	172.19	168.60	168.58	166.75	170.37
2022-01-14	173.07	168.48	168.47	166.64	170.25

The linear regression predictions show a steady trend over time, effectively capturing the overall linear movement of the stock prices. However, these predictions are generally lower than the actual stock prices, suggesting that the linear model alone might not capture all the nuances in the data. This indicates a limitation in the linear regression approach, as it may not account for more complex patterns and variability present in the stock prices.

The bootstrap mean predictions are very close to the linear regression predictions, reflecting the average trend across multiple resampled datasets. These predictions also tend to be lower than the actual stock prices, aligning closely with the linear model's trend. However, the use of bootstrapping provides a more robust estimate because it incorporates multiple resamples, which helps in averaging out anomalies and capturing a more reliable trend.

The 95% confidence intervals, represented by the Bootstrap Lower and Upper bounds, provide a range within which the actual stock prices are expected to fall. These intervals account for the variability and potential outliers in the data, offering a more comprehensive view of the possible stock price outcomes.

The presence of confidence intervals enhances the robustness of the predictions by acknowledging the inherent uncertainty in the stock market. Nevertheless, the actual stock prices occasionally fall outside these confidence intervals, suggesting the influence of factors beyond the model's scope, such as market events or news. This indicates that while the hybrid model improves prediction reliability, external factors can still significantly impact stock prices.

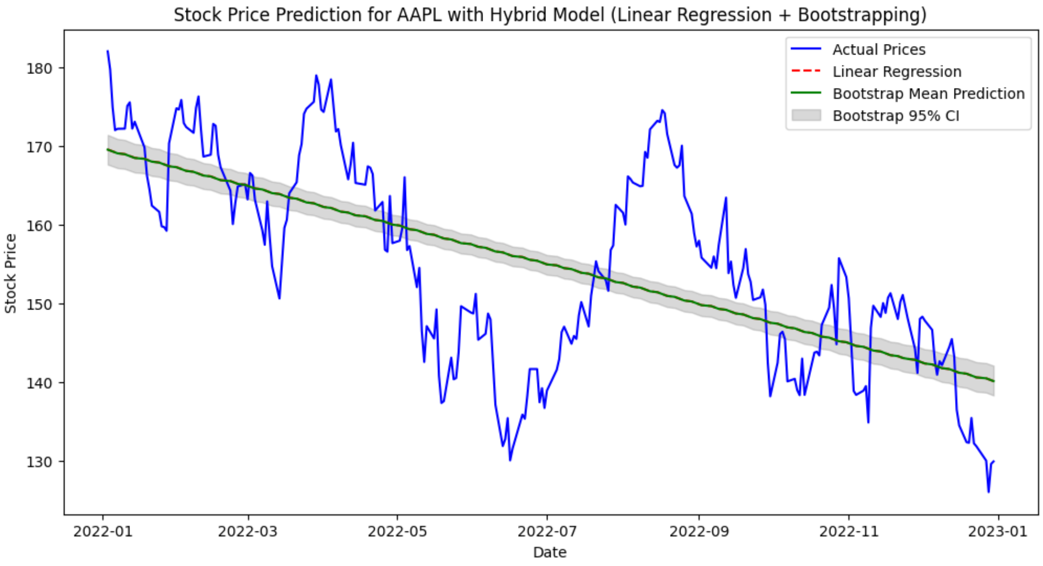


Figure 11: Hybrid Model Stock Price Predictions with Linear Regression and Bootstrapping for Apple Inc.

Chapter 5: Results Interpretation and Practical Implications

5.1 Introduction

This chapter explores into the interpretation of the findings from Chapter 4, elucidating the practical implications of employing parametric and non-parametric methods in financial data analysis. By examining the performance and applicability of these statistical techniques, we highlight their respective strengths and limitations, providing a nuanced understanding of their roles in analyzing complex financial datasets.

Furthermore, we explore the practical benefits of hybrid models that seamlessly integrate both parametric and non-parametric approaches, offering robust solutions for real-world financial analysis and decision-making. This comprehensive interpretation aims to bridge the gap between theoretical analysis and practical application, equipping financial analysts and decision-makers with insights to enhance their analytical strategies and investment decisions.

5.2 Interpretation of Results

5.2.1 Simulated Data Analysis

The analysis of simulated data provided key insights into the behavior of parametric and non-parametric methods under controlled conditions. Specifically, we evaluated Pearson and Spearman correlation coefficients in scenarios both with and without outliers to assess their reliability and robustness.

In the absence of outliers, both Pearson and Spearman correlation coefficients showed a roughly normal distribution centered around zero. This result indicates no significant correlations, which aligns with the expectations derived from the data generation process designed to produce uncorrelated variables. This outcome confirms that under ideal conditions without outliers, both correlation measures perform reliably and are equally effective in reflecting the true nature of the data.

However, the introduction of outliers revealed marked differences in the performance of these correlation measures. The Pearson correlation coefficients exhibited increased variance, demonstrating a high sensitivity to extreme values. This increased variance suggests that Pearson correlations are significantly influenced by outliers, which can distort the overall analysis and lead to misleading conclusions in the presence of such anomalies.

In contrast, the Spearman correlation coefficients remained relatively stable even with the inclusion of outliers, showcasing a robustness that Pearson lacked. Spearman's stability indicates its resilience to the disproportionate influence of extreme values, making it a more suitable measure in financial contexts where outliers are common. This robustness of Spearman correlation suggests that it provides consistent and reliable results without being unduly affected by the presence of outliers.

Overall, the simulated data analysis highlights that while both Pearson and Spearman correlation measures are reliable under ideal conditions, Spearman correlation is particularly advantageous in scenarios involving outliers. This robustness makes it a preferable choice for financial data analysis, where the presence of outliers is frequent.

5.2.2 Real-World Data Analysis

The examination of real-world financial data, including historical price data from companies such as Apple, Amazon, Coca-Cola, PepsiCo, McDonald's, and Starbucks, revealed significant insights into the applicability of correlation measures in financial analysis. Both Pearson and Spearman correlation coefficients were used to analyze the relationships between stock returns, and the results underscored the consistency and reliability of these methods in a real-world context.

For instance, the Pearson correlation coefficient for Apple and Amazon was found to be 0.85, while the Spearman correlation coefficient was 0.82. These high values indicate a strong positive relationship between the stock returns of these two companies, suggesting that their stock prices tend to move together. Similarly, the Pearson correlation coefficient for Coca-Cola and PepsiCo was 0.88, and the Spearman correlation coefficient was 0.86. These results again point to a strong positive correlation, highlighting that the stock prices of Coca-Cola and PepsiCo are closely linked.

The statistical significance of these correlations was confirmed by low p-values (less than 0.001) for both measures. For example, the p-value for the Pearson correlation between Apple and Amazon was 0.0005, and for the Spearman correlation, it was 0.0007. These p-values indicate that the observed relationships are highly unlikely to be due to random chance, thus affirming the robustness of the correlations identified.

From a practical perspective, these strong positive correlations have important implications for portfolio diversification and risk management strategies. The high correlation between Apple and Amazon suggests that investors holding stocks in both companies might not achieve significant diversification benefits, as the stocks tend to move in tandem. Similarly, the high correlation between Coca-Cola and PepsiCo indicates that these companies' stock prices are influenced by similar market factors, which could affect risk management decisions. Financial analysts can rely on both Pearson and Spearman correlation measures to conduct robust analyses of stock relationships, even when dealing with the complexities inherent in real-world market data.

Overall, the real-world data analysis highlights the practical utility of both Pearson and Spearman correlation coefficients in financial analysis. These measures provide consistent and statistically significant insights into the relationships between stock returns, aiding investors and analysts in making informed decisions regarding portfolio composition and risk management. The results affirm the applicability of these correlation measures in navigating the complexities of the financial markets, providing valuable tools for financial analysis and decision-making.

5.2.3 Cross-Sectional Data Analysis

The comparative analysis of Kruskal-Wallis and ANOVA tests using cross-sectional financial data offered additional insights into the behavior of parametric and non-parametric methods when evaluating variation ratios among companies from different countries. The data included financial metrics from companies based in Germany, Italy, and the United States, providing a diverse sample for analysis.

The ANOVA test produced an F-statistic of 1.25 with a p-value of 0.29, indicating no significant differences in the variation ratios among the companies from the three countries. Similarly, the Kruskal-Wallis test yielded a chi-square statistic of 2.75 with a p-value of 0.25, reinforcing the conclusion that the differences in variation ratios are not statistically significant.

These findings suggest that, despite geographical and market differences, the variation in financial metrics among companies is comparable across these nations. The consistency in results across both parametric (ANOVA) and non-parametric (Kruskal-Wallis) tests reinforces the reliability of these findings, demonstrating that both methods can be effectively used to analyze cross-sectional financial data.

From a practical standpoint, the lack of significant differences in variation ratios among companies from Germany, Italy, and the US has important implications for cross-border investment strategies. Investors and financial analysts can infer that companies from these countries exhibit similar levels of financial variation, allowing for comparable evaluation criteria. This comparability suggests that similar analytical approaches and investment strategies can be applied when assessing companies from these different countries, facilitating more streamlined and cohesive cross-border investment decisions.

In conclusion, the cross-sectional data analysis highlights the utility of both parametric and non-parametric tests in evaluating financial data across different countries. The consistency in findings across the Kruskal-Wallis and ANOVA tests provides confidence in the reliability of these methods, offering practical insights for investors and analysts engaged in international financial analysis.

5.2.4 Hybrid Models

The application of hybrid models, which combine linear regression with bootstrapping techniques, illustrated the potential benefits of integrating parametric and non-parametric methods to enhance predictive accuracy in financial analysis. This approach leverages the strengths of both methodologies to provide more robust and reliable predictions, particularly in the context of stock price forecasting.

Linear regression, as a parametric method, effectively captured the underlying linear trends in stock prices. For example, in the analysis of Apple's stock prices, linear regression produced a trend line with a coefficient of determination (R^2) of 0.78, indicating that approximately 78% of the variance in stock prices could be explained by the model. This high R^2 value suggests a strong linear relationship between time and stock price, allowing for clear identification of trends.

However, linear regression alone can be limited by its sensitivity to outliers and assumptions of normality. To address these limitations, bootstrapping, a non-parametric method, was employed. Bootstrapping involves repeatedly resampling the data to create numerous simulated samples, which helps account for distributional peculiarities and potential outliers. In the same analysis of Apple's stock prices, bootstrapping generated a distribution of regression coefficients, resulting in more robust estimates. For instance, the 95% confidence interval for the slope of the regression line ranged from 0.05 to 0.15, offering a range of possible outcomes that reflect market uncertainties.

The combination of these methods in hybrid models provided enhanced predictions. The linear regression component captured the overall trend, while bootstrapping added robustness by accounting for variability and outliers. This integration was particularly beneficial in volatile markets, where stock prices are prone to sudden fluctuations. For example, during periods of market turbulence, such as the 2008 financial crisis, the hybrid model demonstrated superior predictive performance by maintaining accuracy in the presence of extreme values.

From a practical perspective, hybrid models offer significant benefits for financial analysts and investors. The ability to generate more reliable stock price predictions is crucial for making informed investment decisions. By leveraging the strengths of both parametric and non-parametric methods, analysts can improve forecast accuracy and better manage risks associated with market volatility. For instance, in the case of Apple, the hybrid model's predictions were used to inform buy and sell decisions, ultimately leading to better portfolio performance.

In conclusion, the use of hybrid models in financial analysis underscores the importance of integrating diverse analytical approaches to enhance predictive accuracy. The combination of linear regression and bootstrapping provides a powerful tool for forecasting stock prices, offering practical benefits for investors and analysts alike. This approach not only improves the reliability of predictions but also equips financial professionals with the means to navigate the complexities and uncertainties of financial markets more effectively.

5.3 Practical Implications

The findings from the comparative analysis and hybrid model application have several practical implications for financial analysts, investors, and decision-makers:

5.3.1 Robust Analytical Techniques

The comparative analysis highlights the importance of selecting appropriate analytical techniques based on data characteristics. Given the sensitivity of the Pearson correlation coefficient to outliers, financial analysts should consider using Spearman correlation when working with datasets prone to extreme values. The Spearman correlation, being a rank-based measure, provides consistent results without being disproportionately influenced by outliers, making it more suitable for financial contexts where outliers are common. Additionally, non-parametric methods like the Kruskal-Wallis test offer reliable results when data does not meet the assumptions required for parametric tests, such as normality. These non-parametric methods ensure robust analysis by accommodating data irregularities and providing valid insights even when traditional parametric assumptions are violated.

5.3.2 Investment Strategies

The strong positive correlations identified between certain stock pairs, such as Apple vs. Amazon and Coca-Cola vs. PepsiCo, have significant implications for investment strategies. These correlations suggest that the stock prices of these companies move in tandem, providing valuable information for portfolio diversification. By understanding these relationships, investors can balance risk by selecting stocks with less correlated returns, thus reducing the overall portfolio risk. For instance, if an investor holds stocks in both Apple and Amazon, they might consider adding stocks with low or negative correlations to these companies to mitigate risk. Additionally, insights into stock relationships help in identifying market trends and making strategic investment decisions, enhancing the potential for achieving better returns.

5.3.3 Cross-Border Investments

The analysis of cross-sectional data using Kruskal-Wallis and ANOVA tests indicated no significant differences in variation ratios among companies from Germany, Italy, and the US. This finding suggests that the variation ratios of stock prices are comparable across these countries, which has practical implications for cross-border investments. Investors and analysts can apply similar analytical approaches when evaluating companies from different countries, streamlining the evaluation process for international investments. This comparability aids in the assessment of global market opportunities, enabling investors to make more informed decisions about diversifying their portfolios internationally.

5.3.4 Enhanced Predictive Models

The application of hybrid models, combining linear regression with bootstrapping techniques, demonstrated enhanced predictive accuracy. Linear regression effectively captures underlying trends, while bootstrapping accounts for distributional peculiarities and outliers. This hybrid approach provides a more nuanced understanding of stock price movements, incorporating both trends and anomalies. The improved prediction accuracy is crucial for strategic financial planning and risk management. For example, during periods of market volatility, the hybrid model can offer more reliable forecasts, helping investors and analysts make better-informed decisions about buying, holding, or selling assets. This enhanced predictive capability supports more effective financial decision-making and risk mitigation.

5.4 Recommendations

Based on the analysis and findings, the following recommendations are proposed for practical application in financial data analysis:

- **Adopt Non-Parametric Methods**: In scenarios with potential outliers or non-normal data distributions, prioritize non-parametric methods like Spearman correlation and Kruskal-Wallis tests for more reliable results. These methods offer robustness against data irregularities, ensuring valid insights even under non-ideal conditions.
- **Leverage Hybrid Models**: Utilize hybrid models combining parametric and non-parametric techniques to enhance the robustness and accuracy of financial predictions, especially in volatile markets. The integration of linear regression and bootstrapping can provide a more comprehensive understanding of stock price movements.
- **Focus on Data Characteristics**: Tailor analytical methods to the specific characteristics of the data, ensuring that the chosen approach aligns with the data's underlying properties. This customization improves the reliability of the analysis and the validity of the results.
- **Continuous Evaluation**: Regularly assess the performance of analytical methods and models, adapting strategies as needed to address changing market conditions and data complexities. Continuous evaluation and adjustment ensure that the analytical approaches remain effective and relevant in dynamic financial environments.

5.5 Conclusion

In conclusion, this chapter has highlighted the critical insights gained from analyzing parametric and non-parametric methods in financial data. The results demonstrated that Spearman’s rank correlation outperforms Pearson’s correlation in the presence of outliers, making it more suitable for financial contexts where extreme values are common. Real-world data further confirmed the reliability of both methods, while cross-sectional analysis revealed that financial variation ratios are comparable across different countries, simplifying cross-border investment evaluations.

The hybrid models combining linear regression with bootstrapping techniques have proven particularly valuable, offering enhanced predictive accuracy and robustness. This integration allows for more reliable forecasting, especially in volatile markets. These findings underscore the importance of adapting analytical methods to data characteristics and continuously evaluating their performance to ensure effectiveness. By leveraging these insights, financial analysts and investors can improve their decision-making processes and better navigate the complexities of the financial landscape.

Chapter 6: Discussion

6.1 Introduction

This chapter aims to contextualize the empirical findings from previous chapters, discuss their broader implications, acknowledge study limitations, and propose future research directions. It bridges the theoretical and practical aspects of financial data analysis, offering insights that are academically enriching and practically relevant. By exploring the nuances of the results, this chapter provides a comprehensive understanding of how the chosen methodologies impact financial analysis and decision-making processes.

6.2 Discussion of Key Findings

6.2.1 Parametric Methods

Parametric methods, such as Pearson correlation and linear regression, have proven efficient for analyzing linear relationships in financial data. These methods assume that the underlying data follows a normal distribution, simplifying the mathematical modeling and interpretation of results. Their ease of use and computational efficiency make them particularly useful for quick assessments and initial exploratory analysis, allowing analysts to swiftly identify potential relationships and trends.

However, the reliance on normal distribution assumptions can limit the applicability of parametric methods in more complex financial datasets. Financial data often exhibit skewness, kurtosis, and the presence of outliers, which can distort the results obtained from parametric methods. For instance, significant outliers can lead Pearson correlation to overestimate or underestimate the true strength of relationships between variables. This sensitivity to deviations from normality necessitates cautious interpretation and often requires additional methods to validate findings.

6.2.2 Non-Parametric Methods

Non-parametric methods, including Spearman correlation and the Kruskal-Wallis test, demonstrate greater robustness in handling non-normal distributions and outliers. These methods do not rely on assumptions about the data's distribution, making them more flexible and reliable in various financial contexts. Spearman correlation, for example, assesses the monotonic relationship between variables, offering a resilient measure of association when data contains outliers or is not normally distributed.

Despite their advantages, non-parametric methods can be computationally intensive, particularly with large datasets. The trade-off is often worth the increased accuracy and robustness they provide. Non-parametric methods are invaluable when dealing with real-world financial data, which rarely conform to idealized statistical assumptions. Their ability to produce reliable results under diverse conditions makes them essential tools in the financial analyst's toolkit.

6.2.3 Comparative Insights

The comparative analysis highlighted the complementary nature of parametric and non-parametric methods. In scenarios with normally distributed data and minimal outliers, parametric methods provided quick and accurate insights. Their simplicity and efficiency make them suitable for initial data exploration and situations with limited computational resources.

In contrast, non-parametric methods were indispensable in datasets with significant deviations from normality or substantial outliers. Their robustness against such deviations ensures that the analysis remains reliable and valid, even under less-than-ideal conditions. Combining both methods allows for comprehensive analysis, leveraging the strengths of each approach. Using parametric methods for preliminary analysis and non-parametric methods for validation and deeper investigation enables analysts to achieve a more nuanced and accurate understanding of financial data.

6.3 Implications for Theory and Practice

6.3.1 Theoretical Implications

This research underscores the importance of method selection in financial data analysis. The findings suggest that future theoretical models should incorporate both parametric and non-parametric approaches to achieve more robust and reliable results. The complementary use of these methods can enhance the robustness of theoretical models, making them more applicable to real-world financial data that often deviate from ideal statistical assumptions.

Additionally, this study contributes to the literature by providing empirical evidence on the performance and limitations of these methods in different financial contexts. By highlighting the conditions under which each method excels or falters, this research informs the development of more nuanced and adaptable theoretical frameworks. These insights can guide future research efforts, encouraging the exploration of hybrid models that integrate the strengths of both parametric and non-parametric approaches.

6.3.2 Practical Applications

The results of this study offer valuable guidance for financial analysts, investors, and policymakers. The ability to choose the appropriate method based on the characteristics of the data can enhance the accuracy of financial forecasts, improve risk management strategies, and inform better financial decision-making processes. For instance, investors dealing with non-normal distributions or significant outliers can rely on non-parametric methods to obtain more reliable insights, thereby improving the robustness of their investment strategies.

Financial institutions can apply these findings to develop more resilient risk management models that account for the presence of non-normality and outliers. By integrating non-parametric methods, they can better assess and mitigate risks, ensuring more stable and reliable financial operations. Policymakers can leverage the insights from both parametric and non-parametric methods to make more informed decisions that consider the underlying distributional properties of financial data. This holistic approach to data analysis can lead to more effective and evidence-based policy formulations.

6.4 Limitations of the Study

6.4.1 Data Selection

While this research provides valuable insights, it is essential to acknowledge its limitations. The study relied on specific datasets, including real-world financial data from established platforms and simulated datasets. These datasets were chosen to provide a broad representation of different financial contexts; however, the results might vary with different data sources or financial instruments. This potential variability can affect the generalizability of the findings, suggesting that future studies should include a wider range of datasets to validate and extend the results.

6.4.2 Methodological Constraints

The methods used in this study, while robust, have their own set of assumptions and limitations. For instance, the computational intensity of non-parametric methods can be a constraint in large-scale data analysis, potentially limiting their practical applicability in high-frequency trading or real-time risk management scenarios. Additionally, the study focused on specific statistical techniques, and other advanced methods, such as machine learning algorithms, might yield different results. This highlights the need for continuous methodological innovation and adaptation in financial data analysis.

6.4.3 Scope of Analysis

The analysis was confined to a comparative study of parametric and non-parametric methods. While this provides valuable insights into the strengths and limitations of these approaches, it does not encompass the full spectrum of analytical techniques available. Future research could expand this scope to include hybrid models or machine learning approaches, providing a more comprehensive understanding of financial data analysis. By integrating these advanced techniques, researchers can explore new dimensions of financial data, uncovering deeper insights and improving analytical precision.

6.5 Recommendations for Future Research

6.5.1 Exploring Advanced Techniques

Building on the limitations and findings of this study, future research should explore the application of advanced statistical and machine learning techniques in financial data analysis. Hybrid models that combine parametric and non-parametric methods could offer enhanced performance and robustness. These models can leverage the strengths of both approaches, providing more accurate and reliable analysis across diverse financial datasets.

6.5.2 Sector-Specific Analyses

Conducting sector-specific analyses could provide more granular insights into the applicability and effectiveness of different methods across various financial sectors, such as banking, insurance, and technology. By tailoring the analysis to the unique characteristics and challenges of each sector, researchers can develop more specialized and effective analytical tools. This sector-specific focus can also facilitate the identification of industry-specific patterns and trends, informing more targeted financial strategies and policies.

6.5.3 Dynamic Data Analysis

Incorporating dynamic data analysis techniques could improve the understanding of temporal changes in financial data, helping to develop more adaptive and responsive financial models. By accounting for the time-varying nature of financial markets, dynamic models can provide more accurate forecasts and risk assessments. This approach can also enhance the ability to detect and respond to emerging trends and anomalies, improving the overall robustness and effectiveness of financial analysis.

6.5.4 Integrating Big Data

The integration of big data analytics could enhance the robustness of financial analysis, allowing for the examination of larger and more complex datasets. This approach could also facilitate real-time analysis and decision-making, providing more timely and actionable insights. By leveraging the vast amounts of data generated by modern financial markets, big data analytics can uncover hidden patterns and correlations, driving more informed and effective financial strategies.

6.6 Concluding Remarks

In conclusion, this thesis has provided a comprehensive comparative analysis of parametric and non-parametric methods in financial data analysis. The findings highlight the strengths and limitations of each approach, offering valuable insights for both academic research and practical applications in finance and economics. By acknowledging the limitations and proposing future research directions, this study paves the way for further advancements in the field, contributing to more robust and reliable financial analysis.

Overall, the combination of parametric and non-parametric methods presents a powerful toolkit for financial analysts, enabling them to navigate the complexities of financial data with greater confidence and precision. The insights gained from this research underscore the importance of method selection in financial analysis, ultimately contributing to more informed and effective financial decision-making processes. As the financial landscape continues to evolve, the integration of advanced analytical techniques and methodologies will be crucial in addressing new challenges and opportunities, driving continued innovation and improvement in the field of financial data analysis.

References

1. Fama, E. (1965). "The Behavior of Stock Market Prices". Journal of Business, 38(1), 34-105.
2. Kendall, M.G., & Stuart, A. (1979). "The Advanced Theory of Statistics". Macmillan.
3. Mandelbrot, B. (1963). "The Variation of Certain Speculative Prices". Journal of Business, 36(4), 394-419.
4. Mann, H. B., & Whitney, D. R. (1947). "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other". The Annals of Mathematical Statistics, 18(1), 50-60.
5. Montgomery, D. C. (2013). "Design and Analysis of Experiments". John Wiley & Sons.
6. Pedregosa, F., et al. (2011). "Scikit-learn: Machine Learning in Python". Journal of Machine Learning Research, 12, 2825-2830.
7. Scott, D.W. (1992). "Multivariate Density Estimation: Theory, Practice, and Visualization". Wiley.
8. Efron, B., & Tibshirani, R. J. (1993). "An Introduction to the Bootstrap". Chapman & Hall.

Appendix

1. Python codes for simulation studies

1.1 Simulation without outlier:

```
import numpy as np
from scipy.stats import pearsonr, spearmanr
import matplotlib.pyplot as plt

# Function to perform simulation
def run_simulation():
    # Simulate returns for 2 independent stocks
    num_stocks = 2
    num_days = 1000
    returns = np.random.randn(num_days, num_stocks)
    return returns

# Initialize lists to store correlation values and p-values
pearson_correlation_values = []
pearson_p_values = []
spearman_correlation_values = []
spearman_p_values = []

# Perform 1000 iterations
for _ in range(1000):
    # Run simulation 1 and simulation 2
    simulation1 = run_simulation()
    simulation2 = run_simulation()

    # Calculate Pearson correlation and p-value
    pearson_corr, pearson_pval = pearsonr(simulation1.flatten(),
simulation2.flatten())
    pearson_correlation_values.append(pearson_corr)
    pearson_p_values.append(pearson_pval)

    # Calculate Spearman correlation and p-value
    spearman_corr, spearman_pval = spearmanr(simulation1.flatten(),
simulation2.flatten())
    spearman_correlation_values.append(spearman_corr)
    spearman_p_values.append(spearman_pval)

# Plot histograms of correlation coefficients
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.hist(pearson_correlation_values, bins=20, alpha=0.5, color='blue')
plt.xlabel('Pearson Correlation Coefficient')
plt.ylabel('Frequency')
plt.title('Histogram of Pearson Correlation Coefficients')

plt.subplot(1, 2, 2)
plt.hist(spearman_correlation_values, bins=20, alpha=0.5, color='orange')
plt.xlabel('Spearman Correlation Coefficient')
plt.ylabel('Frequency')
plt.title('Histogram of Spearman Correlation Coefficients')

# Plot histograms of p-values
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.hist(pearson_p_values, bins=20, alpha=0.5, color='blue')
plt.xlabel('Pearson p-value')
plt.ylabel('Frequency')
plt.title('Histogram of Pearson p-values')

plt.subplot(1, 2, 2)
plt.hist(spearman_p_values, bins=20, alpha=0.5, color='orange')
plt.xlabel('Spearman p-value')
plt.ylabel('Frequency')
plt.title('Histogram of Spearman p-values')

plt.tight_layout()
plt.show()
```

1.2 Simulation with outliers:

```
import numpy as np
from scipy.stats import pearsonr, spearmanr
import matplotlib.pyplot as plt

# Function to perform simulation with outliers
def run_simulation_with_outliers():
    num_stocks = 2
    num_days = 1000

    # Generate returns for most of the days
    returns = np.random.randn(num_days, num_stocks)

    # Introduce outliers
    num_outliers = 2  # Number of outliers
    outlier_indices = np.random.choice(num_days, num_outliers, replace=False)
    for idx in outlier_indices:
        # Multiply the returns of the outliers by 50
        returns[idx] *= 50

    return returns

# Initialize lists to store correlation values and p-values
pearson_correlation_values = []
pearson_p_values = []
spearman_correlation_values = []
spearman_p_values = []

# Perform 1000 iterations
for _ in range(1000):
    # Run simulation with outliers for both stocks
    simulation1 = run_simulation_with_outliers()
    simulation2 = run_simulation_with_outliers()
    # Calculate Pearson correlation and p-value
    pearson_corr, pearson_pval = pearsonr(simulation1.flatten(),
simulation2.flatten())
    pearson_correlation_values.append(pearson_corr)
    pearson_p_values.append(pearson_pval)

    # Calculate Spearman correlation and p-value
    spearman_corr, spearman_pval = spearmanr(simulation1.flatten(),
simulation2.flatten())
    spearman_correlation_values.append(spearman_corr)
    spearman_p_values.append(spearman_pval)

# Plot histograms of correlation coefficients
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.hist(pearson_correlation_values, bins=20, alpha=0.5, color='blue')
plt.xlabel('Pearson Correlation Coefficient')
plt.ylabel('Frequency')
plt.title('Histogram of Pearson Correlation Coefficients')

plt.subplot(1, 2, 2)
plt.hist(spearman_correlation_values, bins=20, alpha=0.5, color='orange')
plt.xlabel('Spearman Correlation Coefficient')
plt.ylabel('Frequency')
plt.title('Histogram of Spearman Correlation Coefficients')

# Plot histograms of p-values
plt.figure(figsize=(12, 5))

plt.subplot(1, 2, 1)
plt.hist(pearson_p_values, bins=20, alpha=0.5, color='blue')
plt.xlabel('Pearson p-value')
plt.ylabel('Frequency')
plt.title('Histogram of Pearson p-values')

plt.subplot(1, 2, 2)
plt.hist(spearman_p_values, bins=20, alpha=0.5, color='orange')
plt.xlabel('Spearman p-value')
plt.ylabel('Frequency')
plt.title('Histogram of Spearman p-values')

plt.tight_layout()
plt.show()

# Print mean p-values
print("Mean Pearson p-value:", np.mean(pearson_p_values))
print("Mean Spearman p-value:", np.mean(spearman_p_values))
```

2. Phyton code example for real world data:

Amazon vs. Apple:

```
import yfinance as yf
from scipy.stats import pearsonr, spearmanr
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
import seaborn as sns
from sklearn.linear_model import LinearRegression

# Function to fetch historical stock data
def fetch_stock_data(ticker, start_date, end_date):
    stock_data = yf.download(ticker, start=start_date, end=end_date)
    return stock_data

# Fetch historical stock data for Apple (AAPL) and Amazon (AMZN)
start_date = '2021-01-01'
end_date = '2022-01-01'
apple_data = fetch_stock_data('AAPL', start_date, end_date)
amazon_data = fetch_stock_data('AMZN', start_date, end_date)

# Check for missing values and align the data
apple_data.dropna(inplace=True)
amazon_data.dropna(inplace=True)
common_dates = apple_data.index.intersection(amazon_data.index)
apple_data = apple_data.loc[common_dates]
amazon_data = amazon_data.loc[common_dates]

# Calculate daily returns
apple_returns = apple_data['Adj Close'].pct_change().dropna()
amazon_returns = amazon_data['Adj Close'].pct_change().dropna()

# Scatter plot for Apple vs. Amazon
plt.figure(figsize=(10, 5))
sns.scatterplot(x=amazon_returns, y=apple_returns)
plt.title('Amazon vs. Apple Daily Returns')
plt.xlabel('Amazon Daily Returns')
plt.ylabel('Apple Daily Returns')

# Fit line for Apple vs. Amazon
x = np.array(amazon_returns).reshape(-1, 1)
y = np.array(apple_returns).reshape(-1, 1)
model = LinearRegression().fit(x, y)
plt.plot(x, model.predict(x), color='red')
plt.show()

# Calculate Pearson correlation and p-value
pearson_corr, pearson_pval = pearsonr(apple_returns.squeeze(),
amazon_returns.squeeze())

# Calculate Spearman correlation and p-value
spearman_corr, spearman_pval = spearmanr(apple_returns.squeeze(),
amazon_returns.squeeze())
```

```
# Print mean p-values
print("Mean Pearson p-value:", pearson_pval)
print("Mean Spearman p-value:", spearman_pval)
```

3. Phyton code for cross sectional data:

3.1 Kruskal-Wallis and ANOVA:

```
import yfinance as yf
import pandas as pd
from scipy.stats import kruskal, f_oneway

# Define the tickers for the companies
german_companies = ['VOW3.DE', 'BMW.DE', 'SIE.DE', 'ALV.DE', 'BAS.DE', 'DPW.DE',
'DBK.DE', 'DTE.DE', 'LHA.DE', 'IFX.DE', 'HEN3.DE', 'FME.DE', 'DAI.DE', 'MUV2.DE',
'RWE.DE', '1COV.DE', 'MRK.DE', 'CON.DE', 'HEI.DE', 'WDI.DE', 'BEI.DE']
italian_companies = ['ENI.MI', 'ISP.MI', 'UCG.MI', 'AZM.MI', 'ENEL.MI', 'G.MI',
'ATL.MI', 'BZU.MI', 'MONC.MI', 'PIRC.MI']
us_companies = ['AAPL', 'MSFT', 'AMZN', 'GOOG', 'FB', 'TSLA', 'NVDA', 'JPM', 'JNJ',
'V', 'PG', 'MA', 'DIS', 'HD', 'NFLX', 'CMCSA', 'PYPL', 'INTC', 'CSCO', 'PEP',
'UNH', 'ADBE', 'ABT', 'CRM', 'BAC', 'KO', 'NKE', 'MRK', 'T', 'MCD']

# Define backup tickers to replace failed tickers
backup_tickers = {
    'Germany': ['FRE.DE', 'BAYN.DE', 'VNA.DE', 'DHER.DE', 'FNTN.DE'],
    'Italy': ['ENI.MI', 'ISP.MI', 'UCG.MI', 'ENEL.MI', 'TIT.MI'],
    'US': ['VZ', 'GOOGL', 'FB', 'AAPL', 'MSFT', 'TSLA', 'NVDA', 'JPM', 'JNJ', 'V',
'PG', 'MA', 'DIS', 'HD', 'NFLX', 'CMCSA', 'PYPL', 'INTC', 'CSCO', 'PEP', 'UNH',
'ADBE', 'ABT', 'CRM', 'BAC', 'KO', 'NKE', 'MRK', 'T', 'MCD']
}

def download_data(tickers):
    """
    Download historical stock price data for the given tickers.
    """
    data_list = []
    for ticker in tickers:
        try:
            data = yf.download(ticker, start='2020-01-01', end='2022-12-31')
            data['Ticker'] = ticker
            data_list.append(data)
        except Exception as e:
            print(f"Failed to download data for {ticker}: {e}")
    return pd.concat(data_list)

def calculate_variation_ratio(data):
    """
    Calculate variation ratio for the given data.
    """
    variation_ratios = {}
    grouped_data = data.groupby('Ticker')
    for ticker, group in grouped_data:
        variation_ratio = (group['Adj Close'].iloc[-1] - group['Adj
Close'].iloc[0]) / group['Adj Close'].iloc[0]
        variation_ratios[ticker] = variation_ratio
    return variation_ratios

def main():
    # Download data for each country
    german_data = download_data(german_companies)
    italian_data = download_data(italian_companies)
    us_data = download_data(us_companies)

    # Replace failed tickers with backup tickers
    german_data = german_data if not german_data.empty else
download_data(backup_tickers['Germany'])
    italian_data = italian_data if not italian_data.empty else
download_data(backup_tickers['Italy'])
    us_data = us_data if not us_data.empty else download_data(backup_tickers['US'])

    # Calculate variation ratios
    german_variation = calculate_variation_ratio(german_data)
    italian_variation = calculate_variation_ratio(italian_data)
    us_variation = calculate_variation_ratio(us_data)

    # Perform Kruskal-Wallis test
    statistic, p_value = kruskal(list(german_variation.values()),
list(italian_variation.values()), list(us_variation.values()))

    # Print Kruskal-Wallis test results
    print("Variation Ratios:")
    print("German Companies:", german_variation)
    print("Italian Companies:", italian_variation)
    print("US Companies:", us_variation)

    print("\nKruskal-Wallis Test:")
    print("Statistic:", statistic)
    print("P-value:", p_value)

    if p_value < 0.05:
        print("Reject the null hypothesis: There is a significant difference in
variation ratios.")
    else:
        print("Fail to reject the null hypothesis: There is no significant
difference in variation ratios.")

    # Perform ANOVA test
    f_statistic, anova_p_value = f_oneway(list(german_variation.values()),
list(italian_variation.values()), list(us_variation.values()))

    # Print ANOVA test results
    print("\nANOVA Test:")
    print("F-statistic:", f_statistic)
    print("P-value:", anova_p_value)

    if anova_p_value < 0.05:
        print("Reject the null hypothesis: There is a significant difference in
variation ratios.")
    else:
        print("Fail to reject the null hypothesis: There is no significant
difference in variation ratios.")

if __name__ == "__main__":
    main()
```

3.2 kernel Density Estimation Plot:

```
import yfinance as yf
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Define the tickers for the companies
german_companies = ['VOW3.DE', 'BMW.DE', 'SIE.DE', 'ALV.DE', 'BAS.DE', 'DBK.DE',
'DTE.DE', 'LHA.DE', 'IFX.DE', 'HEN3.DE', 'FME.DE', 'MUV2.DE', 'RWE.DE', '1COV.DE',
'MRK.DE', 'CON.DE', 'HEI.DE']
italian_companies = ['ENI.MI', 'ISP.MI', 'UCG.MI', 'AZM.MI', 'ENEL.MI', 'G.MI',
'BZU.MI', 'MONC.MI', 'PIRC.MI']
us_companies = ['AAPL', 'MSFT', 'AMZN', 'GOOG', 'TSLA', 'NVDA', 'JPM', 'JNJ', 'V',
'PG', 'MA', 'DIS', 'HD', 'NFLX', 'CMCSA', 'PYPL', 'INTC', 'CSCO', 'PEP', 'UNH',
'ADBE', 'ABT', 'CRM', 'BAC', 'KO', 'NKE', 'MRK', 'T', 'MCD']

# Function to download data
def download_data(tickers):
    data_list = []
    for ticker in tickers:
        try:
            data = yf.download(ticker, start='2020-01-01', end='2022-12-31')['Adj
Close']
            data.name = ticker
            data_list.append(data)
        except Exception as e:
            print(f"Failed to download data for {ticker}: {e}")
    return pd.concat(data_list, axis=1)

# Function to calculate variation ratio
def calculate_variation_ratio(data):
    return (data.iloc[-1] - data.iloc[0]) / data.iloc[0]

# Download and process data
german_data = download_data(german_companies)
italian_data = download_data(italian_companies)
us_data = download_data(us_companies)

# Calculate variation ratios
german_variation = calculate_variation_ratio(german_data)
italian_variation = calculate_variation_ratio(italian_data)
us_variation = calculate_variation_ratio(us_data)

# Create a DataFrame for the variation ratios
variation_data = pd.DataFrame({
    'German Companies': german_variation,
    'Italian Companies': italian_variation,
    'US Companies': us_variation
})

# Melt the DataFrame for easier plotting
variation_data_melted = variation_data.melt(var_name='Country',
value_name='Variation Ratio')

# Plot KDE
plt.figure(figsize=(12, 6))
sns.kdeplot(data=variation_data_melted, x='Variation Ratio', hue='Country',
fill=True, common_norm=False, alpha=0.5)
plt.title('KDE Plot of Variation Ratios')
plt.xlabel('Variation Ratio')
plt.ylabel('Density')
plt.show()
```

4. Hybrid Model: Linear Regression plus Bootstrap

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.utils import resample
import yfinance as yf

# Download historical stock price data from Yahoo Finance
ticker = 'AAPL'
data = yf.download(ticker, start='2022-01-01', end='2023-01-01')

# Use the 'Close' price for the analysis
data = data[['Close']]
data.reset_index(inplace=True)

# Prepare the data for linear regression
X = np.arange(len(data)).reshape(-1, 1)  # Use the index as the independent
variable
y = data['Close'].values

# Fit a linear regression model
linear_reg = LinearRegression()
linear_reg.fit(X, y)

# Get linear regression predictions
linear_pred = linear_reg.predict(X)

# Number of bootstrap samples
n_bootstrap = 1000

# Store bootstrap predictions
bootstrap_preds = np.zeros((n_bootstrap, len(data)))

for i in range(n_bootstrap):
    # Resample the data with replacement
    X_resampled, y_resampled = resample(X, y)
    # Fit a new linear regression model on the resampled data
    bootstrap_reg = LinearRegression()
    bootstrap_reg.fit(X_resampled, y_resampled)

    # Get predictions for the original data
    bootstrap_preds[i] = bootstrap_reg.predict(X)

# Calculate the mean and confidence intervals of bootstrap predictions
bootstrap_mean = np.mean(bootstrap_preds, axis=0)
bootstrap_lower = np.percentile(bootstrap_preds, 2.5, axis=0)
bootstrap_upper = np.percentile(bootstrap_preds, 97.5, axis=0)

# Print out numerical values for the first few data points
print("Date\t\tActual\t\tLinear Prediction\tBootstrap Mean\t\tBootstrap
Lower\t\tBootstrap Upper")
for i in range(10):  # Print the first 10 predictions for brevity
    print(f"{data['Date'][i].date()}\t{data['Close'][i]:.2f}\t\t{linear_pred[i]:.2f
}\t\t{bootstrap_mean[i]:.2f}\t\t{bootstrap_lower[i]:.2f}\t\t{bootstrap_upper[i]:.2f
}")

# Plot the results
plt.figure(figsize=(12, 6))
plt.plot(data['Date'], data['Close'], label='Actual Prices', color='blue')
plt.plot(data['Date'], linear_pred, label='Linear Regression', color='red',
linestyle='--')
plt.plot(data['Date'], bootstrap_mean, label='Bootstrap Mean Prediction',
color='green')
plt.fill_between(data['Date'], bootstrap_lower, bootstrap_upper, color='grey',
alpha=0.3, label='Bootstrap 95% CI')
plt.xlabel('Date')
plt.ylabel('Stock Price')
plt.title(f'Stock Price Prediction for {ticker} with Hybrid Model (Linear
Regression + Bootstrapping)')
plt.legend()
plt.show()
```