

**Data Analytics**

**10204210  
F/618/7415**

**Section (3)**

**Data Analytics using Python**

**Submitted to**

Eng. Bassam Kassasbeh

**Submitted on**

June 18th, 2023

**Submitted by**

Marwan Tareq Shafiq Al Farah

**Student ID**

21110011

**Spring 2022 - 2023**

**Table of Content**

[***Introduction*** 3](#_Toc137835714)

[**Data Analytic Activities** 3](#_Toc137835715)

[**Data Analytic Techniques** 4](#_Toc137835716)

[**Data Analytic Tools** 5](#_Toc137835717)

[**Types of Data Analytic Methods** 6](#_Toc137835718)

[**Uses of Data Analytic Methods in Real Life** 8](#_Toc137835719)

[***Descriptive Analytics*** 10](#_Toc137835720)

[**Techniques & Examples** 11](#_Toc137835721)

[ **Feature Analysis & Explanation** 11](#_Toc137835722)

[ **Features Visualization & Explanation** 13](#_Toc137835723)

[ **Contingency Table & Explanation** 20](#_Toc137835724)

[**Techniques for Decision-Making** 21](#_Toc137835725)

[**Evaluation** 22](#_Toc137835726)

[***Predictive Analytics*** 26](#_Toc137835727)

[**Techniques & Examples** 26](#_Toc137835728)

[ **Feature Selection Techniques** 27](#_Toc137835729)

[ **Regression Technique** 27](#_Toc137835730)

[**Compare Techniques** 28](#_Toc137835731)

[**Evaluation** 32](#_Toc137835732)

[***Prescriptive Analytics*** 36](#_Toc137835733)

[**Techniques with Examples** 36](#_Toc137835734)

[**Techniques for Finding the Best Course of Action** 38](#_Toc137835735)

[**Objective Function Code** 39](#_Toc137835736)

[**Apply the techniques** 41](#_Toc137835737)

[ **Code screenshots** 41](#_Toc137835738)

[ **Results and Explanation** 42](#_Toc137835739)

[ **Visualization and Explanation** 45](#_Toc137835740)

[***References*** 47](#_Toc137835741)

# ***Introduction***

**Data Analytic Activities:** (KALYAN, 2017; Coursera, 2023)

Data analytics encompasses a wide range of activities that are required in the process of examining, cleaning, transforming, and modeling data. The primary goal of these activities is to generate meaningful insights, draw sound conclusions, and help as much as possible in strategic decision-making. In the information age, these activities are critical pillars in driving a company’s value proposition.

**Data Collection:** The crucial task of gathering data, which serves as the raw material for generating insights, kicks off the data analytics process. This activity could include conducting surveys or experiments, extracting data from databases, leveraging APIs for data procurement, web scraping, or even obtaining data from reliable third-party sources. The most important aspect of data collection is ensuring that the collected data is relevant, accurate, and aligned with the analysis’s objectives. Furthermore, while carrying out this activity, it is critical to adhere to data privacy regulations.

**Data Cleaning and Preprocessing:** Following data collection, the gathered data is frequently unstructured and inconsistent, necessitating thorough cleaning and preprocessing. Handling missing values, dealing with outliers, correcting incorrect entries, removing duplicate records, and other tasks fall under this category. At this point, decisions are made about how to handle these anomalies, should they be excluded, corrected, or should missing or problematic data be imputed? This stage also includes feature selection and feature engineering, which involve selecting the most relevant variables for the analytical models and potentially creating new variables from existing ones to improve the model’s performance.

**Data Transformation:** After cleaning, data is frequently converted or restructured into a more analyzable format. This step may include activities such as normalization (scaling data to a small, specified range), or categorical variable encoding (converting categorical values into numeric values) methods. Additionally, data transformation entails creating derived variables, re-categorizing variables, and adjusting imbalances in the dataset to ensure that it is optimally primed for detailed analysis.

**Data Modeling:** After preprocessing the data, the next step is to build models that efficiently discover patterns within the data which first includes selecting the appropriate model based on the problem statement, such as a regression model, decision tree, clustering algorithm, neural network, or other options, then using the machine learning algorithms, as well as the selection of appropriate parameters and the validation of models to ensure their accuracy and reliability.

**Data Interpretation and Visualization:** Following the modeling stage, the derived insights must be meaningfully interpreted and effectively communicated. This process entails drawing meaningful conclusions from the data, explaining the patterns and trends discovered, justifying the models selected based on their performance, and communicating these findings to stakeholders in a clear and understandable manner. Data visualization tools play an important role at this stage by transforming complex data results into comprehensible graphs and charts, allowing for a more intuitive understanding of the findings.

**Data Analytic Techniques:** (Amadebai, 2022; Stevens, 2022; Calzon, 2023)

Data analytics techniques are extremely beneficial for extracting relevant insights from data and making sound judgments. They comprise a wide range of statistical, mathematical, and computational approaches for analyzing, interpreting, and drawing conclusions from data. Let’s look at some common data analytics techniques:

**Regression Analysis:** A statistical approach used to analyze and quantify the connection between one or more independent variables and a dependent variable. It assists in identifying patterns, trends, and correlations within data and enables for the prediction of future events based on previous data. Regression analysis establishes a solid framework for modeling complicated connections and generating accurate predictions, enabling organizations to obtain important insights and make data-driven choices.

**Classification:** Classification is a machine learning technique that uses available features to categorize data into predefined classes or categories. It entails using labeled data to train a model to learn the patterns and relationships between features and their corresponding classes. The trained model can then be used to accurately classify new, unlabeled data. With the help of classification algorithms, organizations may automate decision-making processes and boost operational effectiveness as they have are applied in many fields including image recognition, fraud detection, customer segmentation, and recommendation systems.

**Clustering:** Clustering is an unsupervised learning technique used to identify natural groupings or clusters within a dataset. Its goal is to group similar data points together while keeping dissimilar data points in separate clusters. Clustering algorithms enable exploratory data analysis, customer segmentation, anomaly detection, and other applications. Clustering enables targeted marketing strategies by revealing hidden patterns and structures in data.

**Time Series Analysis:** Time series analysis is concerned with analyzing and forecasting data gathered at regular intervals of time. It entails looking for patterns, trends, seasonality, and other temporal dependencies in data. ARIMA (Autoregressive Integrated Moving Average) and exponential smoothing are time series analysis techniques that are widely used in forecasting and trend analysis. Businesses may predict market needs, improve inventory management, and make strategic decisions based on anticipated future results thanks to accurate time series forecasting.

**Sentiment Analysis:** Sentiment analysis, also known as opinion mining, entails analyzing text or speech data to determine the sentiment or emotional tone expressed by individuals. It is frequently used in social media monitoring, customer feedback analysis, and brand reputation management. Businesses may assess public opinion, spot patterns, and base data-driven choices on consumer sentiment using sentiment analysis. Businesses may raise total customer satisfaction by understanding consumer sentiment and making improvements to their goods, services, and marketing initiatives.

These are only a handful of the countless data analytics methods that are accessible. The kind of data, the goals of the analysis, and the current business issue all have an impact on the selection and use of particular approaches. To gain comprehensive insights and maximize the value of data, data analysts and data scientists frequently employ a combination of techniques.

**Data Analytic Tools:** (Haan, 2022; Hillier, 2023; Sahiti Kappagantula, 2023)

A wide variety of data analytics tools have been created to assist data scientists and analysts in carrying out the different data analytics activities and techniques, from data collecting and cleaning through modeling and visualization. Some of the most helpful tools are listed below:

**Excel:** Because of its versatility, Microsoft Excel is a well-liked application for small-scale data analysis and visualization. This allows analysts to deal with data in a user-friendly setting as Excel offers a wide range of functionality for data processing, statistical calculations, data modeling, and charting despite its simple user interface.

**Python/R:** The open-source programming languages Python and R have become more popular in the world of data analytics because to their flexibility and strength as they offer comprehensive libraries with features for data handling, analysis, and visualization, including Pandas, NumPy, Matplotlib (for Python), and dplyr, ggplot2 (for R). They are favored by both data scientists and analysts because to their usability and adaptability.

**SQL:** A domain-specific language used for programming and managing data contained in RDBMS or RDSMS is called Structured Query Language (SQL). SQL queries are an essential tool for anybody dealing with data contained in relational databases since they may be used to extract particular data, carry out aggregations, and carry out sophisticated joins between tables.

**Tableau:** With its user-friendly interface and large choice of static, animated, and interactive visualizations, Tableau is a leading tool for data visualization. Tableau turns complicated data sets into understandable, shareable, and actionable insights by empowering people of all skill levels to comprehend, create, and share visual storytelling with data.

**PowerBI:** PowerBI is a Microsoft business analytics tool suite that combines interactive visualizations with self-service business intelligence capabilities. It helps you to create powerful dashboards, reports, and visualizations by connecting to a number of data sources. Users of PowerBI may build their own reports and dashboards without having to have a lot of technical skills.

**Hadoop/Spark:** Apache Hadoop and Spark are frameworks designed to handle and process large datasets across distributed systems. Hadoop’s distributed file system allows big data to be stored across multiple nodes, while its MapReduce component allows complex computations to be performed across these nodes. Meanwhile, Spark performs computations in memory and is thus significantly faster for certain applications.

**Machine Learning Tools:** A wide range of algorithms for applying machine learning in data analytics are provided by libraries like scikit-learn, TensorFlow, and Keras (for Python) as these tools make it possible to complete prediction and classification tasks as well as model creation, training, and validation efficiently on very large datasets.

**Types of Data Analytic Methods:** (Insight Software, 2021; University of Bath Online, 2022; Project Pro, 2023b)

The strength of data analytics comes from its capacity to turn unstructured or structured data into insightful knowledge that helps firms make well-informed and strategic decisions. Descriptive analytics, predictive analytics, and prescriptive analytics are the three main classifications of data analytics methods as each approach has a particular usage, makes use of various statistical and analytical methods, and offers unique advantages to firms who adopt it.

**Descriptive Analytics:** Descriptive analytics, often regarded as the most basic or fundamental type of data analytics, focuses primarily on analyzing past business events to gain insights into what occurred. By doing so, it provides a retrospective view of business operations, similar to a rear-view mirror reflecting past events.

Its main objective is to reduce massive volumes of frequently complex data into manageable, easily understood pieces of knowledge. In practice, descriptive analytics offers stakeholders a clear, succinct overview of numerous business characteristics, simplifying how they should be understood.

To accomplish this, descriptive analytics employs a variety of techniques, such as data aggregation and data mining. Furthermore, it derives meaning from data using fundamental mathematical and statistical computations. Measures of central tendency, such as the mean, median, and mode, are used to understand typical or average data points, for example.

In addition, dispersion measures such as the range, variance, and standard deviation are used to gain insight into the spread and variability of data. These measures aid in understanding the extent to which data points deviate from the average, providing a more complete picture of data distribution.

Descriptive analytics employs techniques such as frequency analysis and cross-tabulation to delve deeper into data patterns and relationships. To examine the data and identify correlations between specific features, univariant, bivariant, and multivariant charts are used. Descriptive analytics, in essence, lays the groundwork for more sophisticated and complex analytical procedures to follow.

**Predictive Analytics:** Predictive analytics, as the name implies, is all about forecasting future events. It goes beyond simply comprehending past events and ventures boldly into the realm of future scenarios. Statistical methods, machine learning algorithms, and data mining are used effectively in predictive analytics to provide an informed prediction about what the future may contain.

However, predictive analytics does not ensure the occurrence of forecasted events, instead, it is a powerful tool that provides probability-based predictions of future events and this predictive ability, that is supported by rigorous data analysis, enables businesses to be proactive, plan ahead, and make decisions today that are informed by potential future developments.

Due to the development, improvement, and use of prediction models, predictive analytics is far more sophisticated and complicated than descriptive analytics. These models are mathematical constructs that forecast future outcomes by combining historical and current data. The techniques used to create these models range from relatively simple statistical methods like regression models to more complex machine learning models like decision trees, random forests, and neural networks.

The type of the data and the particular business issue being addressed are typically determining factors in determining the approach to be employed. Regression models, for example, may be an appropriate choice if the problem involves predicting a numerical value, such as sales revenue. Decision tree or neural network models could be more suited if the objective is to categorize data points into different categories. Predictive analytics thus demands the careful selection and deployment of methodologies that are most appropriate for the current circumstance.

Predictive analytics also includes determining the validity and dependability of the predictive models developed. Cross-validation and out-of-time validation approaches are used to validate models to make sure they are reliable and can generalize well to new data. By creating, experimenting with, and evaluating predictive models, organizations may provide accurate projections that facilitate proactive decision-making.

**Prescriptive Analytics:** Prescriptive analytics, which is frequently regarded as the highest level of data analytics, goes beyond simply predicting future scenarios. It offers a proactive and forward-thinking perspective by making strong recommendations on the best actions to take to achieve specific business goals or to avoid potential risks. Prescriptive analytics incorporates a wide range of computational modeling procedures, such as optimization algorithms, machine learning, and artificial intelligence, to provide this advanced level of analysis.

Prescriptive analytics uses a number of techniques to facilitate decision-making. Among these methods, optimization, simulation, and game theory are common. For instance, simulation makes use of mathematical models to simulate real-world situations and predict possible outcomes. Organizations may assess risks and make wise decisions by simulating various scenarios to acquire insights into the range of potential outcomes and related probability.

Another effective approach in prescriptive analytics is game theory, which examines the tactical interactions between various decision-makers. It helps businesses to assess the dynamics of competitive environments and choose the best course of action. Organizations may strategically position themselves for successful results by comprehending the motivations and actions of various stakeholders.

Optimization is an additional method used in prescriptive analytics. The goal of optimization methods is to select the best option from a range of feasible options. Organizations can systematically evaluate trade-offs, restrictions, and goals to determine the best course of action by defining decision issues as optimization models. This enables them to accomplish desired results or to maximize productivity while lowering expenses.

Prescriptive analytics offers firms strong frameworks to direct decision-making by utilizing these simulation, game theory, and optimization methodologies. These strategies provide businesses the ability to foresee future outcomes, comprehend the effects of various choices, and successfully traverse complicated settings.

**Uses of Data Analytic Methods in Real Life:** (Naveen, 2020; Calzon, 2022; Karan, 2022; Vadapalli, 2022; Project Pro, 2023b)

Data analytic methods, such as descriptive, predictive, and prescriptive analytics, have numerous applications in a variety of industries. These methods provide valuable insights and aid in decision-making. Let’s look at how these methods are applied in real-world industry scenarios:

1. **Descriptive analytics:** it is commonly used to understand past events and gain insights into business operations. In practice, businesses use descriptive analytics in a variety of ways:

* **Retail:** Retailers look at historical sales data to determine popular products, peak shopping times, and consumer purchasing trends. Forecasting demand and enhancing pricing strategies are all made easier with the aid of this information.
* **Marketing:** Descriptive analytics enables marketers to better comprehend consumer behavior, divide their clientele into segments based on their demographics and purchasing patterns, and monitor the success of their marketing initiatives. Decisions about targeted advertising, customer retention tactics, and product development are guided by the information provided.
* **Supply Chain:** Descriptive analytics makes it possible to manage the supply chain by giving insight into inventory levels, monitoring shipping paths, and streamlining logistical processes. Businesses can locate bottlenecks, boost productivity, and guarantee timely delivery.

1. **Predictive analytics:** Predictive analytics concentrates on predicting future outcomes rather than just understanding the past. It is used in a variety of business sectors, including:

* **Finance:** To assess the creditworthiness of loan applicants, financial institutions use predictive analytics for credit scoring. Predictive models can predict the likelihood of loan defaults by analyzing historical data, allowing lenders to make informed decisions.
* **Healthcare:** Predictive analytics is important in healthcare because it aids in disease outbreak prediction, early diagnosis, and treatment planning. Predictive models can identify individuals at high risk for certain conditions and recommend appropriate interventions by analyzing patient data.
* **Manufacturing:** Predictive analytics is used to forecast machine failures, allowing proactive maintenance to minimize downtime. Businesses can optimize maintenance schedules and reduce costs by analyzing sensor data, historical maintenance records, and environmental factors.

1. **Prescriptive Analytics:** Prescriptive analytics extends predictive analytics by making recommendations for optimal actions. Here are a couple of examples:

* **Transportation:** Prescriptive analytics is used in the transportation industry to optimize routes, vehicle scheduling, and load balancing. Prescriptive models recommend the most efficient routes and resource allocation based on factors such as traffic, fuel costs, and delivery deadlines.
* **Energy Management:** Prescriptive analytics assists utilities in optimizing energy generation and distribution. Prescriptive models can recommend strategies for efficient energy generation, storage, and load balancing based on weather data, demand patterns, and grid conditions.
* **Fraud Detection:** Prescriptive analytics is critical in the detection and prevention of fraud. Prescriptive models can identify suspicious activities and trigger alerts or block fraudulent transactions in real time by analyzing patterns and anomalies in transactional data.

In conclusion, data analytic methods are widely used across industries to gain insights, predict outcomes, and optimize decision-making. Predictive analytics helps predict future events, while descriptive analytics provides historical perspectives. Prescriptive analytics recommends optimal actions. These methods’ specific applications and benefits vary depending on industry requirements, but they all contribute to improving operational efficiency, identifying growth opportunities, and mitigating risks.

# ***Descriptive Analytics***

Before we engage on the descriptive analytics path, we must first become acquainted with the type of data we will be dealing with. Our dataset includes several elements of trading operations. These qualities, which are represented in the dataset as features, supply us with an abundance of information regarding the nature and specificity of these activities.

This dataset primarily contains trading data, which includes elements such as dates, business identities, market classifications, and numerous trade-related factors such as volumes, quantities, and prices. Each element provides a distinct viewpoint on trade activity and, when examined combined, can show fascinating patterns, trends, or insights, giving a comprehensive picture of the trading environment.

This study is very useful when making business decisions since it tells us about the current state of the trade environment. It may assist in determining which stocks are more active, when trade volumes peak, which markets are more active, and even how a certain stock’s price swings. These insights are priceless for traders, investors, and market analysts seeking to comprehend market behavior and forecast future trends.

Before we go into descriptive analytics, we must first prepare the dataset for analysis. We begin by importing data from two independent CSV files into individual DataFrames, one with HIGH prices and the other with LOW prices. To provide a consolidated picture of the trade data, these two DataFrames are combined into a single DataFrame.

To assist simpler analysis across time, the TRADE\_DATE column in the generated DataFrame is presented in a consistent date format. Proper date formatting is essential since it allows us to monitor the sequence of trading actions and investigate temporal patterns or trends. Finally, we sort the dataset by TRADE\_DATE and SYMBOL1 to bring it into logical order and make it easier to interpret. By doing so, we ensure that the trade data for each firm is in chronological order, allowing for a more coherent study of individual companies’ trading data across time. We also converted the MARKET and the SEC\_CODE from integers to categorical values as they hold categorical values that cannot be treated the same as numerical values.

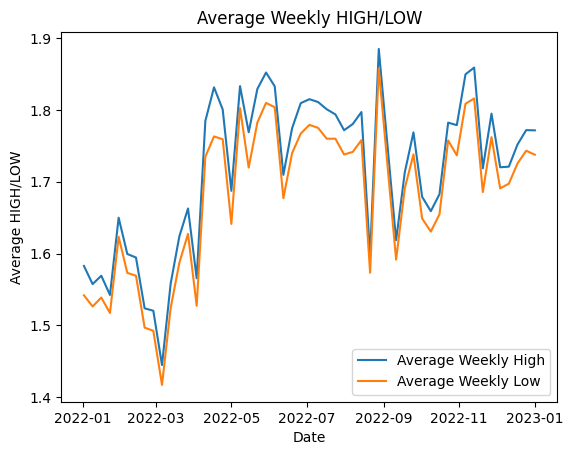
With the data well-structured and the characteristics known, we can now move on to descriptive analytics. This technique will extract valuable insights from trade data, allowing us to gain a better knowledge of the trading landscape and make more educated business choices.

**Techniques & Examples:**

* **Feature Analysis & Explanation:**

|  |  |  |
| --- | --- | --- |
| Feature Name | Descriptive Measure / Technique | Explanation |
| TRADE\_DATE | Mode, Frequency | The day with the most trades, March 13, 2022, has a total of 135 deals. This data is useful for determining the busiest trade days, assisting firms in preparing for greater demand on such days, and adjusting their purchasing and selling tactics appropriately. |
| VOLUME | Mean, Standard Deviation, Range | The average volume of shares traded every transaction is roughly 63,127, with a standard deviation of 302,255.8 and a range from 0.02 to 8,868,824. This substantial variety in the quantity of shares being exchanged might hint to very fluctuating market activity. This can indicate the liquidity of specific securities or markets for businesses, affecting how quickly they can buy or sell securities without significantly affecting the price. |
| TRADE\_QTY | Mean, Standard Deviation, Range | Each trade involves approximately 42,703 units on average, with a standard deviation of 145,178.5 and a range of 1 to 5,741,963. This variation demonstrates that the amount of units transferred in each transaction might vary greatly. Understanding the normal trading amount may influence order sizes when buying or selling, ensuring that their orders do not significantly alter market prices. |
| NO\_OF\_TRADES | Mean, Standard Deviation, Range | The average number of trades is on average 28 trades, with a standard deviation of 66.23 and a range from 1 to 1,700. This variation in the number of trades suggests that market activity can be quite volatile. Businesses can use this information to monitor market activity and alter their trading strategy accordingly. For example, on days with a higher-than-average number of trades, prices may be more volatile, needing a different strategy than on quieter days. |
| HIGH | Mean, Standard Deviation, Range | The average high price is about 1.71 units, with a standard deviation of 4.67 and a price range of 0.02 to 44.5. This price range may be indicative of the volatility of securities prices throughout trading periods. This is vital information for firms in controlling risk and establishing the ideal moment for performing deals. This volatility may bring possibilities for great rewards but also comes with significant risk. |
| BEST\_ASK\_PRICE | Mean, Standard Deviation, Range | The average best ask price for all transactions is around 1.62 units, with a standard deviation of 1.52 and a range of 1 to 40. This might imply that, on average, sellers are willing to sell at lower prices than buyers are ready to pay. Understanding the best ask price can help businesses anticipate the necessary funds for executing trading strategies by providing insight into potential purchasing costs for securities. |
| BEST\_ASK\_QTY | Mean, Standard Deviation, Range | The average best offer quantity across all transactions is around 19,800 units, with a standard deviation of 67,427.18 and a range of 0 to 2,323,450. This indicates the range of securities that sellers are ready to sell at the best ask price. Knowing the average ask quantity can assist businesses in estimating how much of a specific security they can expect to buy in the market without significantly affecting its price. |
| BEST\_BID\_PRICE | Mean, Standard Deviation, Range | With a standard deviation of 1.50 and a range of 0.10 to 40, the average best bid price is about 2.11 units. This suggests that buyers are willing to pay more than sellers’ asking prices on average, which is a common pattern in most markets. Understanding the optimal offer price may give firms insight into prospective selling prices for assets, allowing them to forecast the potential return on investment. |
| BEST\_BID\_QTY | Mean, Standard Deviation, Range | The average number of best bids is around 24,523 units, with a standard deviation of 100,818.4 and a range of 1 to 3,950,000. This chart depicts the fluctuation in the number of shares that buyers are willing to purchase at the best bid price. Understanding the average bid quantity can assist firms in estimating the amount of a certain security that can be sold in the market without materially altering its price. |
| LOW | Mean, Standard Deviation, Range | With a standard deviation of 4.58 and a range of 0.01 to 40, the average low price is roughly 1.68 units. Understanding the average low price can assist firms in anticipating prospective drops in security prices, which may provide buying opportunities. Furthermore, a large spread between high and low prices may indicate considerable volatility in the market, telling firms about the amount of risk involved with trading in that market. |

* **Features Visualization & Explanation:**

****

**Average Weekly High and Low Stock Prices Correlation**

We looked at the patterns in average weekly high and low stock prices in order to better understand market behavior and gain insights that could help us make strategic decisions.

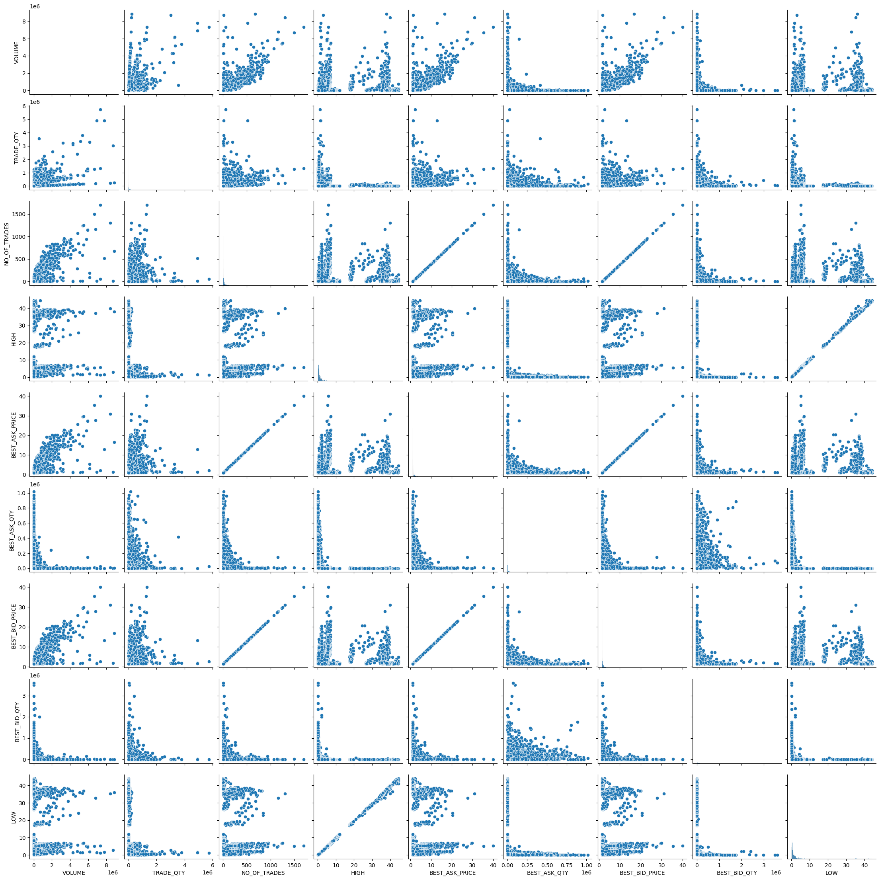
The line chart visually represents the average weekly high and low prices over time. Both of these values are clearly related, exhibiting a very similar trend across the timeframe examined. The graph shows one line for average weekly high prices and the another one for average weekly low prices.

As would be expected, the line representing low prices is always below the line representing high prices. The striking similarity in the two lines’ temporal movements also suggests a strong relationship between the two, so that when high prices rise or fall, low prices do the same, and vice versa. This finding implies that the market conditions that influence high stock prices also influence low stock prices.

In terms of business, this trend provides useful insights for risk management and decision-making processes. Because high and low prices are so closely related, if we see a rise or fall in the week’s highest stock prices, we can expect a similar trend in the lowest prices. This knowledge can assist us in better managing our portfolios, anticipating market trends, and strategizing our trading accordingly.

The observation also emphasizes the significance of constantly monitoring market trends and conditions. This allows us to react quickly to changes and make the best decisions possible based on the most recent data.

This pattern analysis is a fundamental step in our broader analytical approach, assisting us in understanding market dynamics and guiding our stock trading decisions.

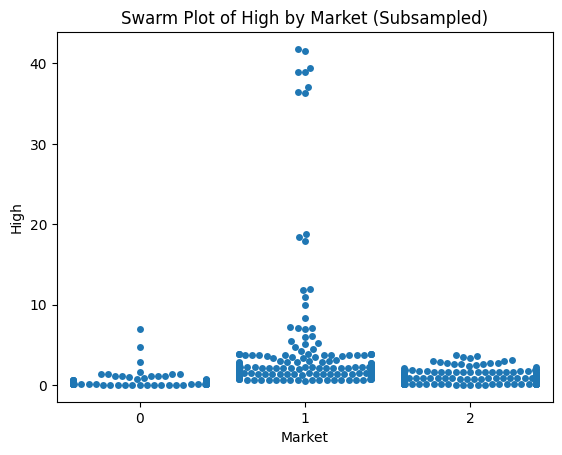


**Understanding the Interdependence of Market Features**

We used a pairplot to illustrate the relationships between different featuresof the trade data in order to understand how distinct market variables relate to one another. The final graph is composed of a matrix of scatterplots, where each plot shows the connection between the two features.

Some important findings we made from this study include the correlation between the highest and lowest stock prices (high and low), the number of trades (number of trades), and the best prices for selling and purchasing (best ask and best bid prices).

The graph shows a nearly linear relationship between these feature pairs. This implies, in terms of business, that as one of these features improves, so does the other. For instance, if the highest stock prices increase, we can expect the lowest stock prices to do the same, and the same is true for the other pairs of features. This understanding can be instrumental in anticipating market movements and strategizing our trades accordingly.



**Investigating the Connection Between Market Sections and High Stock Prices**

We used a swarm plot to improve our knowledge of the dynamics between different market segments and their impact on stock prices. This graph displayed the highest stock prices ‘HIGH’ for each market section, providing a snapshot of their distribution across markets.

Since the analysis of both the highest and lowest stock prices can be time-consuming, we focused on the highest stock prices. This decision was influenced by our earlier discovery that these two variables tend to move in tandem. Therefore, understanding one provides significant insights about the other.

An important observation from the feature analysis section is the relationship between market segments ‘MARKET’ and high and low stock prices. It appears that Market 1 tends to have higher values for both high and low prices as compared to Markets 0 and 2. This suggests that stocks traded in Market 1 may generally carry a higher financial risk, but also potentially higher returns. Understanding this relationship can be extremely useful when deciding which market segments to focus the trading efforts on.

In order to validate this observation, we examined the minimum and maximum high and low prices for each market. The results were as follows:

**For Market 0:**

* The lowest and highest low prices were 0.01 and 7.2, respectively.
* The lowest and highest high prices were 0.02 and 7.39, respectively.

**For Market 1:**

* The lowest and highest low prices were 0.44 and 44.04 respectively.
* The lowest and highest high prices were 0.44 and 44.5, respectively.

**For Market 2:**

* The lowest and highest low prices were 0.02 and 5.95, respectively.
* The lowest and highest high prices were 0.03 and 5.99 respectively.

These results reinforced our observation from the swarm plot, confirming that Market 1 indeed exhibited the highest range of stock prices.

This analysis provides useful insights from a business standpoint. Market 1, with its higher price range, could offer greater potential for profit. Higher prices, on the other hand, imply the possibility of larger losses, suggesting a higher level of risk involved with trading in Market 1.

We are better prepared to make strategic decisions now that we have these insights. Depending on our risk tolerance, we can decide whether to focus trading activities in the higher risk-reward environment of Market 1 or opt for the relatively stable conditions of Markets 0 and 2.

While this provides a significant piece of the puzzle, we know that the mechanics of stock markets are very complicated, with multiple interrelated forces at play. As a result, we continue to use descriptive analysis in order to uncover comprehensive insights.

A picture containing text, screenshot, display, diagram

Description automatically generatedA picture containing text, screenshot, display, diagram

Description automatically generated

**Insight Into Stock Price Fluctuations: HIGH and LOW**

The variation in stock prices is one of the most important factors in stock trading analysis. These changes are represented by ‘HIGH’ and ‘LOW’ prices. Understanding these elements can give useful insight into a stock’s volatility and prospective return.

We used two histogram plots in our study to graphically illustrate the distribution of ‘HIGH’ and ‘LOW’ prices. A histogram chart can help you visualize and understand the underlying distribution of data points.

The charts show that both ‘HIGH’ and ‘LOW’ prices have a noticeable skewness to the right, indicating that most companies’ stock prices are concentrated in the lower range. However, there have been a few cases where stock values have had considerably higher points.

For the ‘HIGH’ pricing, most stocks ranged from 0.02 to 44.5. This fluctuation may also be seen in the ‘LOW’ pricing, where equities have been traded within a range of 0.01 to 44.04.

The skewness to the right, which is defined by stocks with noticeably higher stock prices, may first be mistaken for an outlier. These data points aren’t just statistical anomalies to be ignored, though. Instead, they represent the top tier of the market, which includes huge, well-established businesses with higher stock values because of their steady growth and perceived worth. These outliers provide a more comprehensive perspective of the market’s potential and highlight chances for substantial profits. They stand for high-quality investment prospects that are appealing for long-term investments and symbolize market success and stability.

Understanding these distributions benefits both businesses and investors in a variety of ways. It offers a sense of the general pricing patterns in the market, and the skewness towards the lower price range might imply more inexpensive investment prospects for small and medium investors. Furthermore, the range of these prices illustrates the diversity in the stock market, catering to both risk-averse and risk-tolerant investors.

In essence, understanding the distributions of ‘HIGH’ and ‘LOW’ stock prices can help stakeholders in the market make informed and strategic decisions, enhancing their chances for profitable investments.

A picture containing text, screenshot, square, rectangle

Description automatically generated

**Using the Heatmap to Reveal Data Relationships**

We shifted our focus to a correlation heatmap, a visual tool that highlights the strength and direction of relationships between distinct data points. This approach not only simplifies the correlation matrix but also highlights complicated patterns that are simpler to visualize.

At first look, the ‘HIGH’ and ‘LOW’ closing prices have the strongest connection (nearly 1). This confirms our prior results that these two variables tend to move together, giving us a good foundation for further analysis.

Another finding is the high association between the number of trades ‘NO\_OF\_TRADES’ and the best ask ‘BEST\_ASK\_PRICE’ and bid ‘BEST\_BID\_PRICE’. These links are clear, with a correlation close to one, indicating that an increase or reduction in one might potentially impact the others.

Given the strong relationship between ‘NO\_OF\_TRADES,’ ‘BEST\_ASK\_PRICE,’ and ‘BEST\_BID\_PRICE,’ we opted to keep ‘NO\_OF\_TRADES’ while removing the other two features from our dataset during the preprocessing step. This will help us maintain our data clean, eliminating the inclusion of redundant information which may add noise to our predictions.

Finally, the heatmap not only supported the theories we had formed based on our previous charts, but it also provided us with actionable information. These insights will be critical as we proceed with our analysis, focusing on making accurate predictions and driving strategic business decisions. Finally, recognizing these relationships allows us to predict market developments, making this information an essential element of our toolset.

A picture containing text, plot, screenshot, diagram

Description automatically generated

**Unpacking Biweekly Trends in Average Stock Prices**

We investigated the changes in ‘HIGH’ stock prices for various firms, averaged over bi-weekly intervals. Because ‘HIGH’ and ‘LOW’ prices are strongly associated, as previously established, we picked ‘HIGH’ prices as a representation for both of them to simplify the representation.

The stock market is recognized for its volatility, with prices changing in response to a variety of reasons ranging from company-specific events to worldwide economic developments. By monitoring ‘HIGH’ prices on a bi-weekly timeframe, we can reduce the influence of daily price volatility and instead concentrate on longer-term trends and patterns.

We focused our examination on a group of organizations to make this examination feasible and aesthetically cohesive. The ‘SYMBOL1’ variable in our dataset represents these businesses. Each ‘SYMBOL1’ represents a distinct company, and each company is depicted as a distinct line in the graph in our analysis.

The graph clearly shows that each firm has its own unique trend of ‘HIGH’ stock values throughout time. These trajectories reflect the distinctive rhythm of each company’s stock price response to market events, investor sentiment, and overall performance.

The significance of the ‘SYMBOL1’ variable in stock price development is a critical discovery from this graph. Essentially, it emphasizes that the company’s identity is a major factor influencing its stock price behavior.

Insights gained from this type of chart can be quite beneficial from a business standpoint. Understanding these temporal trends and the influence of firm identity may assist businesses and investors forecast future price trajectories, refine investment strategies, and improve market knowledge.

A picture containing text, screenshot, plot, diagram

Description automatically generated

**Understanding Company Strength Through Stock Volume**

The volume of stocks traded in the stock market can be a good indicator of a company’s strength and market popularity. The term ‘volume’ refers to the total number of shares purchased and sold during a given time period. A large trade volume sometimes shows a lot of interest in a company’s shares, which might represent market confidence and contribute to price changes.

We focused on the ‘VOLUME’ variable in our dataset to gain insights into this aspect. This variable reflects the total number of shares traded for each ‘SYMBOL1’, or firm, in our investigation.

In order to zero in on firms with high trading volumes, we’ve set a requirement that the volume must surpass 20% of 100 million shares. This allows us to eliminate companies with low trading volumes and concentrate our efforts on those with higher volumes.

The chart depicts the total volume of stocks exchanged for each firm that meets our criteria in a straightforward and comprehensive visual format. Each bar in the graphic represents a distinct firm, and its height represents the total volume of stocks traded for that company.

As we can see, there is a significant difference in trading volumes between companies. ‘JOPH’ and ‘JOPT’ had the largest trading volumes, showing strong market interest. Companies like ‘JNTH’, on the other hand, have considerably smaller trade volumes, but they are nonetheless significant enough to be present in our study.

For firms and investors, such information is vital. It can assist in determining which firms are generating a lot of market activity and where substantial amounts of shares are being purchased and sold. It may also be used to evaluate market sentiment for individual firms, which can help with strategic choices such as portfolio development and investment diversification.

Finally, knowing these volumetric dynamics and how they differ between firms broadens our understanding of the market landscape and gives useful insights for strategic investment decisions.

* **Contingency Table & Explanation:**

In our investigation, we looked into the relationships between securities codes ‘SEC\_CODE’, company symbols ‘SYMBOL1’, and market distribution. Contingency tables were employed for this investigation because they give an ordered and clearly interpretable framework for understanding how different factors interact.

The initial step was to evaluate the relationship between securities codes ‘SEC\_CODE’ and company symbols ‘SYMBOL1’. The essential question was whether each securities code related to a unique business symbol, and similarly, if each company symbol was connected with a unique securities code.

The results of creating a contingency table for these two variables revealed that four securities codes were connected to more than one firm symbol. This finding suggests that a single security code can represent different types of securities or financial instruments from the same company or issuer, each represented by a distinct symbol in the market. This knowledge is critical for investors since the value and risk associated with different instruments issued by the same issuer might differ dramatically. As a result, while deciding on investing strategies, one must examine the distinct nature of each instrument, even if they are issued by the same company.

The contingency table, on the other hand, indicated that each firm symbol ‘SYMBOL1’ was paired with a distinct securities code. This implies that, as it should be, each unique corporate symbol is linked to a distinct identity in the financial market. The symbol often symbolizes a specific financial instrument issued by a corporation, and the securities code is a financial market identity for the issuer.

Following our investigation of the link between ‘SEC\_CODE’ and ‘SYMBOL1’, we went on to investigate the relationship between various markets ‘MARKET’ and business symbols ‘SYMBOL1’. The contingency table was utilized in this case to offer information on the distribution of various corporate symbols across various markets.

The resulting table revealed a number of interesting findings. First, it displayed the most commonly traded corporate symbol in each market. ‘MECE’ is the most traded symbol in Market 0, ‘AHLI’ is the most traded in Market 1, and ‘ATCO’ is the most traded in Market 2.

We also counted the number of distinct corporate symbols traded in each market. Market 0 trades 47 different symbols, Market 1 trades 57, and Market 2 trades 123.

A notable conclusion from the investigation is that no corporate symbol is traded in all three marketplaces, implying a degree of specialization or emphasis in each market.

These insights are extremely helpful for strategic decision-making, whether for a firm considering entering a market or for investors. For example, a corporation whose symbol is heavily traded in a particular area may signal significant recognition or popularity in that market, which may impact decisions about where to focus marketing or sales efforts.

Understanding which symbols are most traded in a market can offer investors with insight into where considerable financial activity is concentrated, perhaps identifying areas of interest for investment. Furthermore, the variety of symbols in a market might offer an indication of the market’s breadth and potential for diversification.

Finally, knowing the linkages between ‘SEC\_CODE’, ‘SYMBOL1’, and ‘MARKET; is crucial for both businesses and investors. This information may help drive strategic decisions, give useful market insights, and identify prospective investment possibilities.

**Techniques for Decision-Making:**

Descriptive analytics techniques contribute significantly to stock trading decision-making. Businesses may make educated decisions, manage risk effectively, and enhance their trading strategies by utilizing feature analysis, visualization, and contingency table analysis.

Feature analysis is important in decision-making since it provides meaningful assessments and evaluations of numerous trade aspects. Businesses may determine the busiest trading days by analyzing trade dates, allowing them to plan for higher demand and change their strategy appropriately. Understanding the most regularly traded securities based on security codes and market symbols gives insights into market trends and preferences, allowing organizations to more effectively focus their attention and resources on specifics symbols and security codes. Businesses can evaluate liquidity and order sizes by evaluating transaction volume and quantity, which is critical for executing deals without significantly impacting prices. Analyzing high and low prices, as well as optimum ask and bid prices, aids in risk management, calculating optimal buying and selling prices, and estimating possible investment returns. These analyses together provide vital information to organizations, allowing them to make educated decisions regarding when, what, and how to trade.

By providing clear and complete visual representations of data linkages, trends, and patterns, visualization techniques greatly contribute to decision-making. For example, visualizing average weekly high and low stock prices lets organizations to discover patterns and trends over time, allowing them to make strategic decisions based on market behavior. Pairplots illustrate correlations and linkages between distinct components of trade data by visualizing the interdependence of market variables. Businesses can predict market fluctuations and alter their trading strategy by studying these linkages. Furthermore, swarm plots aid in the identification of links between market sectors and high stock prices, assisting firms in identifying profit potential and risk levels. Businesses may use visualizations to swiftly digest complicated data, get insights, and make timely choices based on the most up-to-date information.

Analysis of contingency tables provides useful insights into the links between securities codes, corporate symbols, and market dispersion. Businesses may make better educated judgments regarding investment possibilities, risk assessment, and market segmentation by investigating these links. Businesses may analyze the diversity of financial instruments issued by a firm by analyzing the originality of securities codes and the distribution of corporate symbols across different securities codes. This knowledge enables them to appraise the value and risk associated with various instruments, influencing investing decisions. Furthermore, examining the distribution of corporate symbols across markets reveals market trends and preferences. This data helps organizations focus their efforts and resources, target specific markets for investment, and modify their plans to market emotions.

By giving a multidimensional perspective of the market and reinforcing the findings from each approach, the integration of feature analysis, visualization, and contingency table analysis improves decision-making. Feature analysis, for example, may indicate that certain securities codes are related with larger trade volumes, indicating market interest in those specific instruments. Visualization tools like as scatterplots and line charts can help to validate this discovery by visually demonstrating the link between trade volumes and securities codes. Furthermore, contingency table analysis may validate these findings by studying the distribution of corporate symbols across different markets, revealing which symbols are regularly traded and emphasizing the relevance of market segmentation. Businesses may gain confidence in their decision-making process and make more educated choices based on a thorough understanding of market trends, risk levels, and investment possibilities by examining these insights collectively.

To summarize, these descriptive analytics approaches considerably contribute to stock trading decision-making. Feature analysis gives critical metrics and evaluations of trading characteristics to firms, allowing them to make educated decisions about when and how to trade. Businesses may use visualization tools to acquire a thorough knowledge of trends, and patterns in the data, which can help them foresee market moves and make timely choices. Contingency table reveals links between securities codes, company symbols, and market distribution, allowing firms to analyze investment possibilities, assess risks, and target specific markets. These tactics provide firms with the resources they need to make data-driven choices, efficiently manage risk, and improve their trading strategies for success in the volatile world of stock trading.

**Evaluation:**

Data analytics is critical in decision-making processes in a variety of businesses, including stock trading. We will examine the findings of descriptive analytics to assess the relevance of data analytical approaches in the decision-making process as it provides important insights into trade data, enabling organizations to better understand market dynamics, discover patterns and trends, and make more educated decisions. We will also emphasize the importance of these strategies in stock trading decision-making by assessing the descriptive analysis findings.

Feature analysis was a critical step in comprehending trade data as it helped us acquire significant insights into numerous aspects of trade operations by employing various descriptive measurements and approaches. Let’s look at the findings of the feature analysis:

**TRADE\_DATE:** Understanding the busiest trade days allows businesses to plan for increased demand and change their strategy appropriately. Based on the descriptive analysis results, we discovered that March 13, 2022, had the most trades, suggesting a busy day in the market. This information enables firms to properly allocate resources and organize their trade activity to fulfill demand on such days.

**VOLUME, TRADE\_QUANTITY, and NO\_OF\_TRADES:** Examining the descriptive measurements of these characteristics reveals information on market activity, liquidity, and volatility. For example, the average volume of shares traded each transaction, as well as the standard deviation and range, assist firms in understanding market volatility. This data informs decisions about how quickly to acquire or sell assets without materially affecting prices. Similar insights may be gleaned by analyzing trade amount and trade volume, allowing enterprises to adjust their trading strategy based on market activity.

**BEST\_ASK\_PRICE, BEST\_BID\_PRICE, HIGH, and LOW:** Analyzing these prices’ descriptive parameters aids in risk management, finding ideal buying/selling prices, and comprehending price volatility. Businesses can examine the possible benefits and dangers involved with securities trading by assessing average values, standard deviations, and ranges. This data aids in determining the best time to execute transactions and successfully manage price variations.

Visualization approaches improve comprehension of data patterns and linkages. Let’s look at the outcomes of visualizing trade data:

**Average Weekly High and Low Stock Prices Correlation:** Visualizing the average weekly high and low prices over time gives insights into market behavior and patterns. The line chart depiction of these prices revealed a substantial connection between them, demonstrating that changes in high prices are followed by comparable movements in low prices and vice versa. This discovery helps companies to forecast market patterns and arrange trading activity accordingly.

**Understanding Market Interdependence:** Pairplots show the correlations between different market factors. Businesses may spot connections and understand market changes by visualizing these interdependencies. For example, the pairplot demonstrated a virtually linear link between the highest and lowest stock prices, showing that a change in one price causes a comparable trend in the other. This knowledge helps foresee market volatility and modify trading methods accordingly.

**Investigating the Relationship Between Market Sections and High Stock Prices:** Swarm plots show the distribution of high stock prices across market sections. Businesses may determine profit possibilities and risk levels associated with various market sectors by visualizing this connection. The swarm plot, for example, revealed that Market 1 had greater high and low prices than Markets 0 and 2. This knowledge enables organizations to analyze the possible profits and risks involved with trading in various market segments, assisting in decision-making processes.

**Observations on Stock Price Fluctuations: HIGH and LOW:** The distribution of high and low stock prices is depicted using histogram plots. These distributions can be used to get insight into price volatility and prospective investment opportunities. For example, the histogram skewness to the right suggests that most stock prices are concentrated in the lower range. This data assists businesses in determining the affordability and risk levels connected with various equities, allowing them to make risk-adjusted investing decisions.

**Using a Heatmap to Show Data Relationships:** Heatmaps visually depict relationships between various variables. We discovered considerable connections between high and low closing prices, as well as the number of trades and the best ask/bid prices, by examining the heatmap. These links enable organizations to predict price changes based on trade volumes and best pricing, so helping investment strategy and market forecasting decision-making processes.

Contingency table research reveals correlations between factors such as security codes, corporate symbols, and market dispersion. Let’s look at the outcomes of the contingency table analysis:

**Examining Securities Codes and Company Symbols:** The contingency table indicated that each securities code matches to a distinct company symbol. This provides correct identification of financial instruments issued by a corporation and facilitates successful market representation. However, it was discovered that numerous securities codes can be connected with a single corporate symbol. This implies the presence of many forms of securities or financial instruments issued by the same corporation, each with its own unique symbol. Understanding this link is critical for investors evaluating the value and risk associated with different securities from the same issuer.

**Market Segments and Symbols:** The contingency table analysis revealed the most commonly traded symbols in each market category. Businesses can use this data to discover market trends, popular securities, and market segmentation. Different symbols, for example, were discovered to be regularly traded in different marketplaces, implying specialization or emphasis in each market. Understanding market preferences helps organizations focus their efforts, target certain markets, and adjust their trade tactics accordingly.

The descriptive analysis results highlight the significance of data analytical methodologies in stock trading decision-making. The examination of various aspects, visualization approaches, and contingency table analysis all give useful insights for making educated decisions. Businesses can use these approaches to:

* Determine the busiest trading days and alter your techniques to fit market demand.
* Understand market trends, preferences, and liquidity in order to allocate resources effectively.
* For risk management, evaluate price volatility, optimal buying/selling prices, and prospective profits.
* Predict market changes and adapt trading methods depending on market interdependence.
* Determine profitable market sectors and evaluate the risks associated with certain securities.
* To make educated investing decisions, examine stock price distributions and correlations.
* For correct market representation, understand the links between securities codes, corporate symbols, and market distribution.

Businesses may use these tactics to make data-driven choices, efficiently manage risks, and improve their trading strategies for successful stock trading.

Data analytical techniques such as feature analysis, visualization, and contingency table analysis all play an important role in the stock trading decision-making process. The examination of descriptive analysis findings emphasizes the significance of these approaches in comprehending market dynamics, recognizing patterns and trends, and making sound judgments. Businesses may obtain significant insights into market behavior, liquidity, volatility, and investment possibilities by studying various aspects, displaying transaction data, and investigating correlations between factors. These insights allow organizations to modify their trading strategy, reduce risks, and optimize profits. Integrating data analytics into decision-making processes is critical for navigating the complicated and ever-changing world of stock trading, resulting in better informed and effective trading outcomes.

# ***Predictive Analytics***

Following the detailed insights gained from the descriptive analytics phase, we are now ready to dive into the report’s predictive analytics part. This following phase allows us to make predictions about future occurrences, broadening our study beyond the description and interpretation of previous data.

The predictive analytics phase began with the segmentation of our dataset into separate components, namely ‘X’, ‘y\_high’, and ‘y\_low’. This division allowed us to seperate between dependent variables and independent variables, laying the groundwork for subsequent investigation.

The elimination of the ‘BEST\_ASK\_PRICE’ and ‘BEST\_BID\_PRICE’ features became essential as we progressed. The correlation heatmap created during the descriptive analytics stage informed this conclusion, indicating that these attributes had a perfect correlation of 1 with the ‘NO\_OF\_TRADES’. As a result, these traits were judged redundant, and they were removed to provide more accurate predictions.

Next, we focused on data transformation, especially the conversion of object features to numeric feature by employing the Label Encoder, a tool capable of translating category labels into a format suited for machine learning algorithms. Based on the categorical variables determined during the descriptive analytics phase, this technique was carried out.

Finally, the data was normalized using the standard scaler, which scales the values in the dataset without affecting the range or distribution by normalizing it using the z-score. We reduced the danger of bias towards variables with broader ranges by doing so, opening the way for more accurate and successful predictive modeling.

The preceding steps have prepared our data for the predictive analytics that will follow, bringing us one step closer to deriving valuable and actionable insights from our dataset.

**Techniques & Examples:**

Our research methodology was divided into two major stages. The first stage was feature selection, in which we selected a subset of relevant features from the initial dataset to use in building our predictive models. We used these selected features in three different regression models in the second stage. The goal was to build models capable of accurately predicting both the highest ‘HIGH’ and lowest ‘LOW’ closing prices of stocks. We used two distinct techniques for feature selection: Sequential Feature Selector and Select K Best. We used the most relevant features to create three different regression models: Linear Regression, Random Forest Regression, and K-Nearest Neighbors Regression.

* **Feature Selection Techniques:**

|  |  |  |
| --- | --- | --- |
| Name | Description | Results |
| Sequential Feature Selector | The Sequential Feature Selector is a feature selection technique that builds the model one predictor at a time to find the best subset of features. The selection procedure can be carried out either forward or backward. Forward selection begins with no features and gradually adds them to improve the model’s performance. Backward selection begins with all features and removes one at a time that has the least impact on the model’s performance. In our project we employed forward selection. (Raschka, 2014; GIANLUCA MALATO, 2023) | **High:**  **Linear Regression:** SYMBOL1, MARKET, VOLUME, TRADE\_QTY  **KNN Regressor:** SEC\_CODE, MARKET, BEST\_ASK\_QTY, BEST\_BID\_QTY  **Random Forest Regressor:** TRADE\_DATE, SEC\_CODE, SYMBOL1, TRADE\_QTY  **Low:**  **Linear Regression:** SYMBOL1, MARKET, VOLUME, TRADE\_QTY  **KNN Regressor:** SEC\_CODE, SYMBOL1, MARKET, BEST\_BID\_QTY  **Random Forest Regressor:** TRADE\_DATE, SEC\_CODE, SYMBOL1, TRADE\_QTY |
| Select K Best | Select K Best is a method for selecting univariate features. This method chooses the best ‘k’ features based on a scoring function (f\_regrssion), which is typically a statistical test that assesses the relationship between each feature and the target variable. Each feature is evaluated separately. It’s especially useful if you have a lot of features and want to reduce the dimensionality of your dataset. (Bhatt, 2019; DataTechNotes, 2021) | **High:** SYMBOL1, MARKET, VOLUME, NO\_OF\_TRADES, BEST\_ASK\_QTY  **LOW:** SYMBOL1, MARKET, VOLUME, NO\_OF\_TRADES, BEST\_ASK\_QTY |

* **Regression Technique:**

|  |  |
| --- | --- |
| Name | Description |
| Linear Regression | Linear Regression is a statistical model that predicts the relationship between one or more independent variables (often denoted as X) and a dependent variable (often denoted as Y). The equation for the relationship predicts a response variable as a linear function of the parameters. It is frequently used in forecasting, time series modeling, and determining the relationship between variables. (IBM Documentation, 2022; MALI, 2023) |
| Random Forest Regressor | Random Forest Regressor is a regression ensemble learning method that constructs multiple decision trees during training and outputs the mean prediction of the individual trees. It performs well in high-dimensional spaces, and is less likely to overfit than a decision tree, and often performs well even without tuning parameters. (Bakshi, 2020; Dutta, 2020; Beheshti, 2022) |
| K-Nearest Neighbors Regressor | The K-Nearest Neighbors (KNN) Regressor is an instance-based learning algorithm that makes no assumptions about the underlying data distribution but is based on the assumption that similar input objects will produce similar outputs. When presented with a new, unknown observation, it searches its database to see which ones have the closest features and assigns the average value to the new observation. It’s highly adaptable and simple to use, but it can be computationally expensive as the dataset grows. (The University of Sidney, 2021; Singh, 2023) |

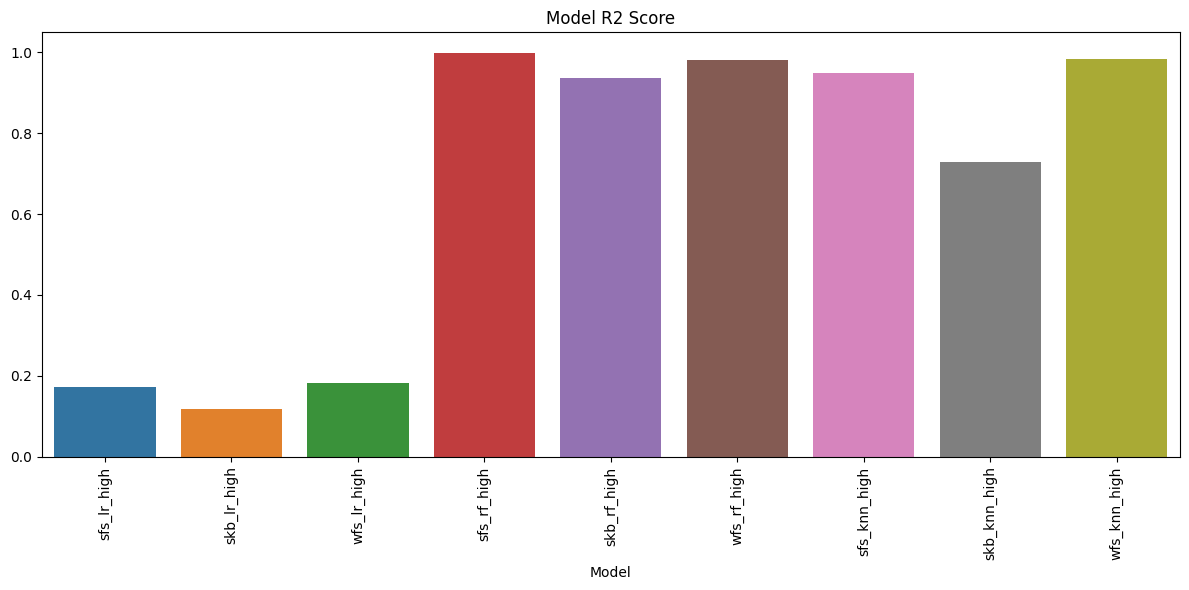
**Compare Techniques:**

Our evaluation metrics for the results include R2 Score, Mean Squared Error, Root Mean Squared Error, and Mean Absolute Error. These metrics provide information about the accuracy and performance of each model.

**Predictions for ‘HIGH’:**

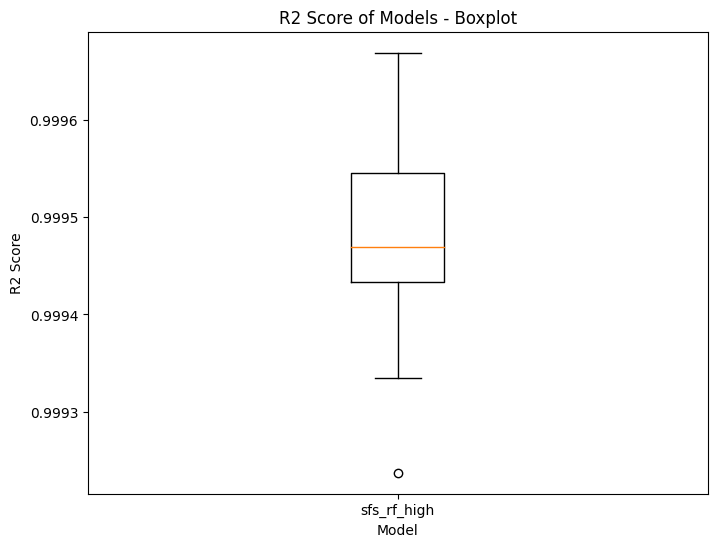
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Selection | Machine Learning | R2 Score | Mean Squared Error | Root Mean Squared Error | Mean Absolute Error |
| Sequential Feature Selection | **Linear Regression** | 0.1768 | 17.9049 | 4.2300 | 1.6427 |
| **Random Forest Regressor** | 0.9995 | 0.0113 | 0.1058 | 0.2963 |
| **K-Nearest Neighbors Regressor** | 0.9556 | 0.9643 | 0.9806 | 0.2082 |
| Select K Best | **Linear Regression** | 0.1207 | 19.1280 | 4.3722 | 1.6408 |
| **Random Forest Regressor** | 0.9473 | 1.1454 | 1.0667 | 0.2558 |
| **K-Nearest Neighbors Regressor** | 0.7286 | 5.8969 | 2.4270 | 0.5348 |
| None | **Linear Regression** | 0.1872 | 17.6803 | 4.2034 | 1.6155 |
| **Random Forest Regressor** | 0.9820 | 0.3924 | 0.6193 | 0.1339 |
| **K-Nearest Neighbors Regressor** | 0.9829 | 0.3727 | 0.6059 | 0.1468 |

**Visualization of the Results:**

** A picture containing text, screenshot, diagram, plot

Description automatically generated**

All Models’ Bar Plots (R2 Score & Mean Squared Error)

 A picture containing text, diagram, screenshot, line

Description automatically generated

Model with the highest R2 Score’s Box Plots (R2 Score & Mean Squared Error)

For ‘HIGH’ closing prices, the Sequential Feature Selection model in conjunction with the Random Forest Regressor model delivered exceptional performance across all evaluation metrics: R2 Score (0.9995), Mean Squared Error (0.0113), Root Mean Squared Error (0.1058), and Mean Absolute Error (0.2963). These values indicate the model’s ability to explain nearly all of the variability in the response data around its mean while also predicting with high precision.

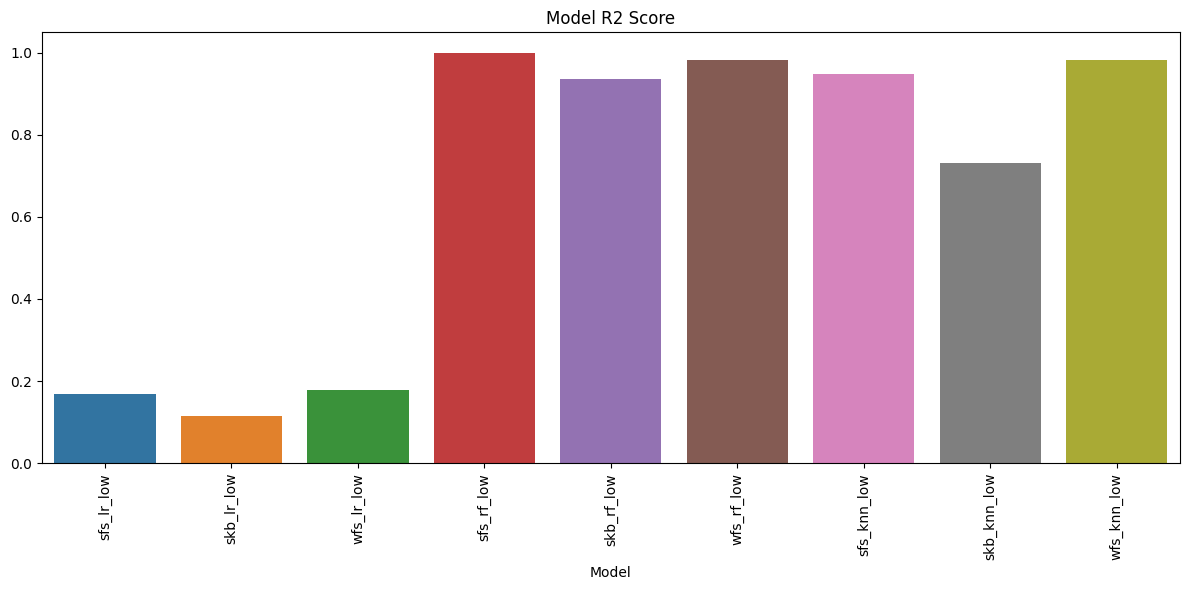
The bar plots clearly demonstrated this superior performance. The Random Forest model had the highest R2 Score and the lowest error metrics, indicating that its predictions were accurate and precise. The box plots emphasized this performance even more, displaying very closely distribution around high R2 Scores and low error metrics, highlighting consistent and reliable performance across different runs.

In contrast, the Linear Regression model performed the least well, with R2 Scores ranging from 0.1207 to 0.1872 and error metrics significantly higher. This performance was visible in the bar plots, with the Linear Regression model consistently achieving lower R2 Scores and higher error metrics.

**Predictions for ‘LOW’:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Feature Selection | Machine Learning | R2 Score | Mean Squared Error | Root Mean Squared Error | Mean Absolute Error |
| Sequential Feature Selection | **Linear Regression** | 0.1722 | 17.3456 | 4.1634 | 1.6167 |
| **Random Forest Regressor** | 0.9994 | 0.0128 | 0.1127 | 0.0306 |
| **K-Nearest Neighbors Regressor** | 0.9557 | 0.9276 | 0.9617 | 0.2028 |
| Select K Best | **Linear Regression** | 0.1172 | 18.4997 | 4.2997 | 1.6149 |
| **Random Forest Regressor** | 0.9470 | 1.1110 | 1.0508 | 0.2503 |
| **K-Nearest Neighbors Regressor** | 0.7313 | 5.6247 | 2.3702 | 0.5223 |
| None | **Linear Regression** | 0.1826 | 17.1285 | 4.1373 | 1.5904 |
| **Random Forest Regressor** | 0.9827 | 0.3673 | 0.5964 | 0.1311 |
| **K-Nearest Neighbors Regressor** | 0.9831 | 0.3537 | 0.5906 | 0.1452 |

**Visualization of the Results:**

** A picture containing screenshot, text, diagram, rectangle

Description automatically generated**

All Models’ Bar Plots (R2 Score & Mean Squared Error)

A diagram of a box plot

Description automatically generated with low confidence A picture containing text, diagram, screenshot, line

Description automatically generated

Model with the highest R2 Score’s Box Plots (R2 Score & Mean Squared Error)

For ‘LOW’ closing prices, the combination of Sequential Feature Selector and Random Forest Regressor outperformed other models once more, with an R2 Score of 0.9994, Mean Squared Error of 0.0128, Root Mean Squared Error of 0.1127, and Mean Absolute Error of 0.0306. These values indicate the model’s high accuracy and precision in predicting ‘LOW’ closing prices.

The bar plots reflected this dominance, with the Random Forest model achieving the highest R2 Score and the lowest error measurements. The box plots provided additional confirmation, displaying a tight score distribution and a small interquartile range, indicating consistent performance across different runs.

The Linear Regression model, on the other hand, lagged behind, with R2 Scores ranging from 0.1172 to 0.1826 and higher error measurements. The bar and box plots consistently highlighted this poor performance, indicating that the Linear Regression model was the least accurate and precise for predicting ‘LOW’ closing prices.

**Evaluation:**

The incorporation of machine learning and feature selection techniques into predictive analytics has opened up new avenues for improving the performance of these predictive models, allowing businesses to more effectively strategize their investments and mitigate risks.

Predictive analytics were used in our project to create models capable of forecasting the highest ‘HIGH’ and lowest ‘LOW’ closing prices of stocks. We took a comprehensive and systematic approach, combining feature selection techniques, regression models, multiple iterations, and evaluation metrics. The sections that follow will break down each component of this predictive analytics project, explaining how each step contributed to improving the models’ effectiveness and, ultimately, yielding reliable results.

Our project began with two critical stages: feature selection and regression model implementation. Feature selection is a critical step in predictive modeling, as it improves model performance by extracting the most relevant features and discarding the redundant ones. This strategy reduces overfitting, improves model accuracy, and speeds up the training process.

We used two feature selection techniques in this project: the Sequential Feature Selector (SFS) and the Select K Best (SKB) method. Each made a distinct contribution to model performance. In our project, the SFS works by adding features that improve model performance iteratively, using a forward selection approach. In contrast, the SKB method selects the top ‘k’ features based on a statistical relationship test between each feature and the target variable. The combination of these two techniques produced a balanced approach that effectively reduced the dimensionality of the dataset while retaining meaningful features that influenced prediction performance.

Following the feature selection phase, three regression models were implemented: Linear Regression (LR), Random Forest Regressor (RF), and K-Nearest Neighbors Regressor (KNN). The LR served as a baseline because it is a simple and interpretable model. However, its limitation of assuming linearity between variables prompted us to try RF and KNN. RF, an ensemble learning method, is resistant to overfitting and can handle high-dimensional data and multicollinearity. We reduced the number of estimators (n\_estimators) for the RF model for this project to 20, which allowed for faster model training without sacrificing performance. While computationally expensive, KNN is extremely flexible and can detect complex patterns in data. We embraced a comprehensive exploration of the data by using these models, capturing diverse data patterns and thus improving prediction performance.

The regression models were implemented in our project over the course of 30 iterations. Multiple iterations helped the model learn the underlying structure of the data more efficiently and guarantee that the results are a true reflection of the model’s predictive power and not just the result of chance by giving the models, with each iteration, a unique set of training and testing data by constantly changing the value of the random\_state in the train\_test\_split function that splits the data into training and testing data.

R2 Score, Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE) were the four evaluation metrics we used. Each metric offers a distinctive perspective on model performance, and taken together, they offer a complete picture of the model’s accuracy and precision. The selection of various evaluation metrics ensured the validity of our analysis and enabled us to identify and address any model flaws.

In this project, different models were assessed for their ability to forecast both “HIGH” and “LOW” closing prices of stocks. Sequential Feature Selector (SFS) and Random Forest Regressor (RF) were discovered to work best together in both prediction scenarios, demonstrating the effectiveness of ensemble learning and feature selection in stock price forecasting.

The SFS-RF combination produced an exceptional R2 Score of 0.9995 for the predicted ‘HIGH’ closing prices, indicating that this model can explain roughly 99.95% of the variability in the predicted ‘HIGH’ stock prices, indicating nearly perfect predictive ability. The model’s accuracy and precision were further demonstrated by the extremely low error measurements for this combination (MSE of 0.0113, RMSE of 0.1058, and MAE of 0.2963).

The SFS-RF model performed similarly well when applied to predictions of ‘LOW’ closing prices, with an R2 Score of 0.9994 and low error measurements (MSE of 0.0128, RMSE of 0.1127, and MAE of 0.0306). These results confirm the model’s exceptional ability to capture the intricate dynamics of stock prices.

In contrast, the Linear Regression (LR) models performed less well in capturing the complexity of stock price movements, regardless of the feature selection method used. In the case of SFS, the LR model, in contrast to the RF model, had a significantly lower R2 Score (0.1768 for “HIGH” and 0.1722 for “LOW”) and higher error measurements, indicating less precise and accurate predictions.

The RF model still performed well without any feature selection, showing an R2 Score of 0.9820 for “HIGH” and 0.9827 for “LOW.” This strong performance, even in the absence of explicit feature selection, highlights the RF model’s inherent ability to forecast stock prices.

When used in with no feature selection algorithm, the K-Nearest Neighbors Regressor (KNN) model performed admirably, albeit slightly less well than the RF model. The KNN’s R2 Scores of 0.9829 for ‘HIGH’ and 0.9831 for ‘LOW’ show that it was able to capture a sizable portion of the complexity of price movement, albeit not as well as the RF model.

These findings highlight the overall potency of combining various models, feature selection strategies, and multiple iterations to produce reliable and precise predictive models. The exceptional performance of the RF model, particularly when coupled with SFS, highlights the importance of using sophisticated, non-linear models and efficient feature selection strategies when forecasting stock prices.

In terms of Business Implications and Future Directions, the high accuracy attained by these predictive models holds significant promise for promoting business success. A business can maximize profits and minimize risks by making strategic investment decisions that are based on accurate stock price predictions. Predictive analytics can improve financial planning, resource allocation, and risk management by offering a trustworthy forecasting tool, which will ultimately benefit the success of the company as a whole.

Although the results of our study were encouraging, there is room for improvement. For instance, experimenting with additional feature selection strategies and regression models may result in predictive models that are even more accurate. Similar to this, investigating more complex methods like deep learning models could improve prediction accuracy even more.

It’s also important to keep in mind that a variety of outside factors, such as economic data and geopolitical developments, which were excluded from our analysis, affect stock prices. These elements should be taken into account in future studies to build even more reliable predictive models.

In conclusion, the development of highly accurate predictive models for predicting stock prices was successfully accomplished through the integration of feature selection techniques and multiple regression models, along with numerous iterations and the use of multiple evaluation metrics. These models can help businesses make wise investment decisions and boost their bottom line. However, ongoing efforts to improve these predictive models’ performance will greatly increase their value in assisting with business success.

# ***Prescriptive Analytics***

**Techniques with Examples:**

Prescriptive analytics, the highest level of data analytics, uses a complex collection of technologies to deliver actionable insights and decision suggestions based on data and in recent years this field has experienced an increase in attention and research. To properly comprehend the complexity and breadth of prescriptive analytics, one must first understand the approaches it employs. In this section, we will look at three fundamental approaches in prescriptive analytics: optimization, simulation, and game theory. (Rouse, 2012; Analytics, 2022)

1. **Optimization:**(Baobab, 2020; Kuttappa, 2020; Dilmegani, 2023)

Optimization is a key approach used in prescriptive analytics that aims to find the best (optimal) solution to a certain issue given certain limitations. It solves complicated problems using mathematical models and algorithms. The goal is to optimize desired aspects (such as earnings or efficiency) while minimizing undesirable ones (such as expenses or waste).

Supply chain management is one of the most common uses of optimization. Companies employ optimization in this context to determine the most effective allocation and use of resources in order to minimize costs, enhance service delivery, and maximize profitability. Algorithms can aid in the optimization of inventory management, manufacturing schedules, transportation routes, and a variety of other supply chain aspects.

Logistics businesses such as UPS and FedEx, for example, have used optimization algorithms to improve their delivery procedures. These organizations have achieved considerable cost savings and better service delivery times by designing the most effective routes for their delivery vehicles, taking into account elements such as road conditions, weather, traffic, and delivery deadlines.

In order to address challenging optimization issues, computer methods called Nature-Inspired Optimization Algorithms (NIOA) mimic natural processes. NIOA effectively explores the space of potential solutions, drawing inspiration from evolutionary processes, swarm intelligence, and natural behaviors. To come up with ideal or close to ideal solutions, they adapt, collaborate, and self-organize. Wide-ranging applications of NIOA may be found in engineering, finance, and data science, where they are used to solve difficult optimization issues when more conventional approaches may be insufficient. SCA, PSO, and GA are examples of these algorithms, each provide distinctive and complicated techniques to address challenging optimization issues.

The sine and cosine principles of math and physics serve as the basis for the Sine Cosine Algorithm (SCA). Through a well-balanced combination of exploration and exploitation strategies, it looks for global answers. Sine and cosine components are used by SCA to change the search’s direction while methodically controlling the size of the search space across iterations. This technique is very effective in locating the best answers in a multidimensional search space, such as when minimizing energy use in smart grids. Companies in the energy industry, for instance, have used SCA to specify a series of activities, such as balancing the power load and controlling the energy usage among various units, resulting in greatly better energy efficiency and cost savings. (Tam, 2021)

On the other hand, Particle Swarm Optimization (PSO) draws its inspiration from the social behavior of swarms. In PSO, each ‘particle’ in the swarm stands in for a potential solution, advancing through the problem space by emulating the particle with the highest fitness value. These particles modify their course based on their own knowledge and that of other particles, enabling a thorough investigation and utilization of the search region. For instance, PSO is used in the telecommunications sector to optimize network topologies, such as the positioning of routers and towers, to increase signal strength and coverage. (Mirjalili, 2016)

The process of natural evolution, which incorporates ideas like selection, crossover (recombination), and mutation, is the source of inspiration for Genetic Algorithms (GA). Each potential solution is represented as a “chromosome” and these chromosomes are evaluated using a fitness function, and those with greater fitness values have a higher likelihood of passing on their genes to the following generation. Combining this selection process with crossover and mutation processes causes populations to evolve over many generations in the direction of superior solutions. In the financial sector, GA has been used to choose a group of companies, decide how much money should be invested in each one, and then use crossover and mutation operations to develop these portfolios over several generations, producing the best possible investment strategy. (Gad, 2018)

1. **Simulation:** (Adra, 2014; Rossetti, 2018; Baobab, 2020)

Another effective tool in prescriptive analytics is simulation. It entails developing an artificial representation (model) of a system in order to forecast its future behavior. Typically, the model is run numerous times under varied settings to simulate a number of outcomes, assisting decision-makers in understanding the potential ramifications of their actions.

Simulation is typically used in conjunction with other prescriptive analytic approaches to examine multiple situations and identify the optimal course of action. Simulation models, for example, can play an important part in treatment planning in the healthcare business. Healthcare providers can make educated judgments regarding the best course of action by modeling the probable impact of several treatment options for a patient. This is especially useful in critical care settings where judgments must be made fast and properly.

1. **Game Theory:** (Wijesooriya.K, 2017; Kingston, 2022; Project Pro, 2023a)

Game Theory is a strong prescriptive analytics technique that works with strategic scenarios in which an individual’s decision depends not only on their own choices, but also on the decisions of others. This strategy entails developing mathematical models to comprehend the intricate interactions of decision makers, with the ultimate objective of determining the best tactics.

Game theory may be utilized to guide competitive strategy decisions in the commercial setting. It can, for example, assist organizations in anticipating competition responses to new product releases or pricing tactics, allowing them to build a best-response strategy. Furthermore, game theory may be applied in negotiations when each side’s strategy is contingent on the actions of the other party.

The well-known “Prisoner’s Dilemma” is an example utilized by many organizations in competitive markets. This game theory scenario investigates the repercussions of cooperative vs non-cooperative action and can assist organizations in making decisions in a competitive environment.

**Techniques for Finding the Best Course of Action:**

Prescriptive analytics, like predictive analytics, strives to not only foresee future events but also identify precise actions that may be performed to attain the intended conclusion. It achieves this goal by combining the power of several approaches, including those outlined above.

1. **Using Optimization for Actionable Insights:** (Baobab, 2020; Kuttappa, 2020; Dilmegani, 2023)

Optimization is used in the context of prescriptive analytics to find the optimum course of action to achieve a certain objective. For example, a manufacturing business may employ optimization to establish the best production plan in terms of cost and output. The end result would be a collection of specified actions for each manufacturing line.

Optimization may also be used in logistics to discover the most effective delivery routes, taking into account factors such as distance, traffic, and delivery time. With optimization, the organization is actively molding the outcome by adopting the optimal course of action rather than simply forecasting what will happen.

By modeling the decision variables, objective functions, and constraints, SCA, PSO, and GA may also be employed in action-based optimization situations to provide actionable insights.

For instance, SCA may be used to design a series of activities to optimize energy use in the context of energy management. The SCA algorithm can determine the ideal balance between energy load and usage by simulating the energy consumption patterns and restrictions of various units within smart grids. The solutions produced by SCA can provide certain energy management techniques, such as how to distribute the power load or when to switch on/off specific equipment, resulting in maximized energy efficiency and cost savings.

PSO can help businesses in the telecommunications sector optimize their network settings. The antennas’ positions and orientations, the transmission power levels, and other factors might be among the choice variables. The PSO method may identify the best configuration by simulating the signal quality and communication limitations, hence maximizing network coverage and capacity. PSO’s solutions include specific instructions on how to do things like where to put additional antennas or how to change the power levels, which leads to better service delivery.

GA may be used to the financial sector to enhance investing choices. The choice of stocks and the distribution of investments may be among the decision variables. GA may determine the ideal portfolio that maximizes return and minimizes risk by modeling the return rates and risk levels. Specific investing strategies, such as which stocks to buy/sell and how much to invest in each, are provided by the solutions created by GA, which results in maximized investment returns.

1. **Making Informed Decisions Using Simulation:** (Adra, 2014; Rossetti, 2018; Baobab, 2020)

Simulations give insights into the possible consequences of various actions, guiding the decision-making process. For example, an oil and gas corporation may utilize simulation to anticipate the probable yield and environmental effect of drilling in a certain region. They can then decide if drilling at that place is the best line of action based on the results.

In healthcare, simulation may be used to understand the prospective results of various treatment approaches. This information might then be used by the medical team to determine the optimum treatment strategy for a patient.

1. **Using Game Theory to Adopt a Strategic Approach:** (Wijesooriya.K, 2017; Kingston, 2022; Project Pro, 2023a)

Game theory is a useful tool for determining how to behave in strategic scenarios where the result is dependent on the actions of others. A telecoms corporation, for example, may utilize game theory to predict how their competitors would respond to their new pricing approach, and then change their own plan to maximize profit.

In the arena of politics, game theory is used to forecast the behavior of various countries. These insights are used by policymakers to decide the best strategic movements in diplomacy, commerce, or combat. The acts in game theory are not simply about forecasting future events, but also about strategically influencing them.

Finally, prescriptive analytics analyzes data and recommends particular actions using a variety of advanced methodologies. Organizations may make educated, data-driven decisions that correspond with their objectives by knowing the various consequences and risks associated with each action. In an increasingly complicated and unpredictable environment, these qualities are critical for making the best judgments possible.

**Objective Function Code:**

def obj\_func(q):

  q = numpy.round(q)

  values = numpy.array([1.33, 5.59, 1.6, 0.47, 0.33, 0.58,

                        0.5, 0.47, 0.83, 1.14, 1.23, 1.19,

                        0.1, 1.18, 1, 1.36, 0.5, 0.45])

  minCost = sum(q \* values)

  if sum(q) < 10:

    return 99999999

  return minCost

This section of the report explains the problem’s objective function. The purpose is to identify the ideal stock quantity for each bank so that traders may spend the least amount of money buying stock. This is fundamentally an optimization problem that may be solved through the use of various prescriptive analytic tools.

To discover the best solution, an objective function has been created. The purpose of this function is to minimize the cost associated with the number of stocks (q) acquired from each bank on any given day.

The objective function’s equation represents the overall cost of stock acquisition. The terms in the calculation, such as 1.33 \* Q11, 5.59 \* Q12, and so on, indicate the cost of each stock acquired. Q11 represents the number of stocks purchased from the first bank ARBK on the first day (Sat 7/5/2022), Q12 represents the number of stocks purchased from ARBK on the second day (Sun 8/5/2022), and so on. The amounts for the other banks follow a similar trend. The coefficients (1.33, 5.59, etc.) denote the stock price for the relevant bank on the respective day.

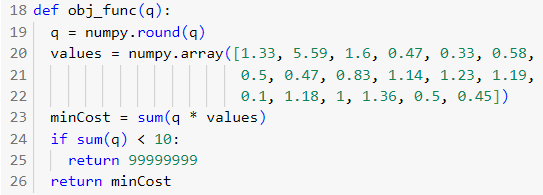
The obj\_func(q) function computes the cheapest way to buy stocks. To verify that the number of stocks is an integer, each quantity of stock is rounded to the nearest integer using numpy.round(q). The function also determines whether the total number of stocks acquired is fewer than ten. If it is, the function returns a very big number (99999999), indicating that this solution is not practicable since it violates the requirement that the total quantity must be less than ten stocks. If the total amount is equal to or more than 10, the function returns the total cost of stock acquisition (minCost).

Finally, the objective function is subject to the constraint that each quantity is limited to 5 stocks. This implies that no more than 5 equities can be purchased from any bank on any given day. This limitation guarantees that traders do not overinvest in a single bank or on a single day, allowing them to distribute their investments more evenly and reduce their risk.

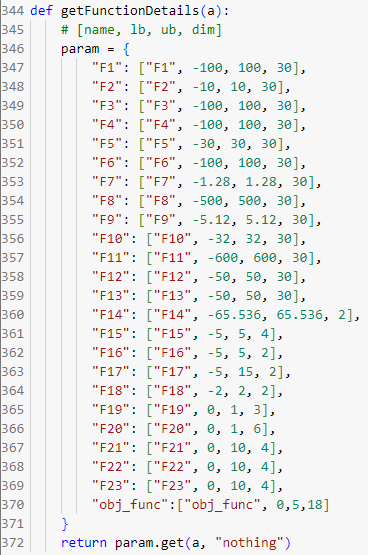
**Apply the techniques:**

* **Code screenshots:**

**benchmarks.py Modifications:**

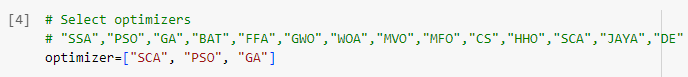
****

The Objective Function That is Explained in the Previous Section

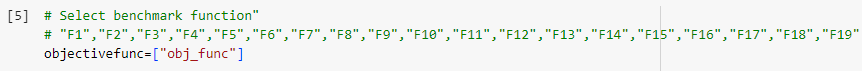
****

Adding the “obj\_func” to the getFunctionDetails function to determine the Minimum and Maximum value of each Q and the number of Qs

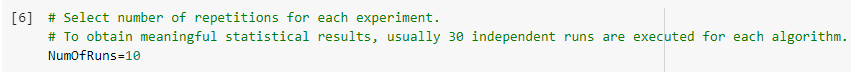
**EvoloPy.ipynb Modifications**

****

Determining the Optimizer Algorithms that will be used

****

Determining that we will use “obj\_func” as the objective function

****

Determining the Number of Runs

* **Results and Explanation:**

The complexities of stock trading optimization difficulties need the employment of advanced computing. approaches, notably prescriptive analytics. The foundation of this prescriptive analysis is the optimum selection of algorithms, which has a significant influence on the precision, computational efficiency, and practical relevance of the created solutions.

We looked at a variety of algorithms, including the Salp Swarm Algorithm (SSA), Bat Algorithm (BAT), Firefly Algorithm (FFA), Grey Wolf Optimizer (GWO), Whale Optimization Algorithm (WOA), Multi-Verse Optimizer (MVO), Moth Flame Optimization (MFO), Cuckoo Search (CS), Harris Hawks Optimization (HHO), JAYA, and Differential Evolution (DE). These algorithms are mostly nature-inspired, as seen by their distinct problem-solving methodologies and features. They have showed success in a variety of fields of study, solving difficult optimization challenges, by simulating numerous natural occurrences.

However, for this optimization issue, we choose to employ three optimization algorithms: Particle Swarm Optimization (PSO), Sine Cosine Algorithm (SCA), and Genetic Algorithm (GA). These algorithms were chosen largely for their adaptability in tackling a wide range of optimization issues.

The social behavior patterns of bird flocks inspire Particle Swarm Optimization (PSO), a population-based optimization algorithm. Its distinguishing characteristics, such as simplicity, the capacity to deliver excellent answers, and the rapid convergence to optimal solutions, proved its suitability for our situation. (Tam, 2021)

The Sine Cosine Algorithm (SCA), on the other hand, is a relatively newer algorithm in the field of optimization algorithms. It has performed admirably across a wide range of optimization jobs. It uses sine and cosine laws to establish a good balance between exploration and exploitation in the search space, making it a feasible option for our objective. (Mirjalili, 2016)

The Genetic Algorithm (GA) is based on the principles of natural development. Its resilience and adaptability, which allow it to tackle a wide range of issue kinds, explain for its continued prominence in the area. (Gad, 2018)

While the other algorithms studied had useful qualities as well, the decision to choose PSO, SCA, and GA was influenced by particular criteria. These included our problem’s unique peculiarities, the requirement for computing efficiency, and the need for a balanced approach to exploration (investigating new regions) and exploitation (refining promising areas). These factors aided our decision-making process, resulting in the selection of these three algorithms to provide important insights from our stock market information and assist traders in decreasing their stock trading expenses.

1. **Use of Prescriptive Analytics Techniques:**

We used a prescriptive analytics technique based on mathematical optimization concepts to achieve the goal of determining the best stock quantity for each bank and so helping traders to decrease the amount spent on stock. This technique was chosen because of its capacity to foresee the impacts of future actions, allowing for more effective resource management and strategy development.

As previously stated, the chosen algorithms include Particle Swarm Optimization (PSO), Sine Cosine Algorithm (SCA), and Genetic Algorithm (GA).

1. **Analytic Techniques Insights and Results Comparison:**

The algorithms were executed with the limitations of a maximum of five stocks per quantity and a total quantity of at least ten stocks. The ideal answer was 2.15, which was the lowest feasible cost after using the optimization techniques.

PSO was the first algorithm that succeed in convergent to this low value. The PSO algorithm accomplished this on its 10th run, during the 11th iteration. This accomplishment demonstrates PSO’s ability in tackling optimization issues, as it was able to quickly determine the lowest cost of stocks.

SCA’s performance, on the other hand, while excellent, required a few more iterations to reach the smallest achievable value, as it was the second, third, and fourth to converge, doing so in its 9th, 4th, and 6th runs, respectively, during the 13th, 18th, and 21st iterations. The sine and cosine laws motivate SCA’s method to generating and updating solutions, allowing it to effectively explore and utilize the search area. The findings show that it is more resilient and consistent in identifying optimal solutions than PSO, but with more iterations.

The Genetic Algorithm (GA), on the other hand, was unable to obtain the minimal feasible value within the stated restriction of 50 iterations over 10 cycles. GA’s performance mirrors its evolutionary character, since it generates solutions using strategies inspired by natural evolution such as selection, mutation, and crossover. However, in this particular situation, it was unable to discover the best solution in the allotted amount of iterations. This implies that, while GA is a robust method that performs well for many optimization problems, its applicability and performance might vary depending on the problem’s characteristics.

However, when the average values at the 50th iteration of each algorithm are examined, a different picture emerges. SCA comes out on top with the lowest average of 2.69, followed by PSO with a higher average of 5.19 and GA with a maximum average of 10.04. This suggests that SCA was more consistent in finding good solutions over all runs, rather than merely achieving the lowest value initially. This fact emphasizes the importance of selecting an optimization method that takes into account not only the speed of convergence but also the consistency of the approach throughout all iterations.

1. **Practical Stock Purchase Recommendations:**

As a result of the optimization process, some intriguing facts were revealed, which might be very useful for traders. When the algorithms arrived at the best option at a cost of 2.15, they all directed to the same purchase method. The recommendation was to acquire 5 shares of ARBK on Wednesday, 11/5/2022, and 5 shares of AHLI on Saturday, 7/5/2022. This is an informative piece of information that illustrates the algorithms’ capacity to not only uncover optimal solutions but also give useful decision-making advice.

This result is especially convincing since, although using distinct techniques, different optimization algorithms came to the same conclusion. It highlights the power of these algorithms in giving actionable information that can be utilized to influence stock market decisions. Furthermore, the fact that this buy advice appeared consistently throughout several runs lends weight to this idea.

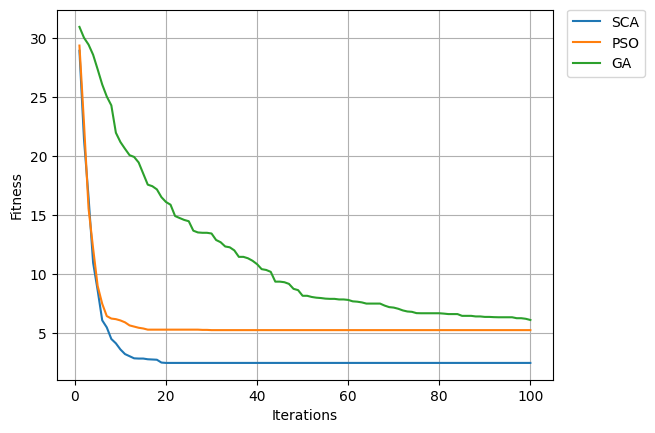
Regardless of the merits of any method, it is important noting the larger implications of these discoveries. They demonstrate the ability of prescriptive analytics in providing practical solutions that successfully maximize desired objectives, such as cost reduction in this example. Such insights are extremely valuable in the complicated and frequently unpredictable world of stock trading.

While the precise amounts and dates will undoubtedly change depending on the conditions, traders can use these insights to influence their trading methods. Furthermore, with regular data updates and market dynamics tweaks, the algorithms may be repeated to provide new suggestions. Another advantage of employing these advanced prescriptive analytics tools is their versatility.

Our findings highlight the necessity of selecting the correct optimization technique for the right problem (in this case, the SCA optimization algorithm), taking not just speed but also consistency between iterations into account. The knowledge obtained here may be used to make better educated judgments in the complicated world of stock trading. These optimization algorithms may be re-run when market dynamics change, giving updated and relevant answers for traders.

* **Visualization and Explanation:**

Using the iteration results offered in our data set, we generated a line chart to compare the performance of the Sine Cosine Algorithm (SCA), Particle Swarm Optimization (PSO), and Genetic Algorithm (GA). The convergence patterns and rates of the three algorithms may be intuitively compared thanks to this graphical depiction.

****

**Convergence Plot of the Optimization Algorithms**

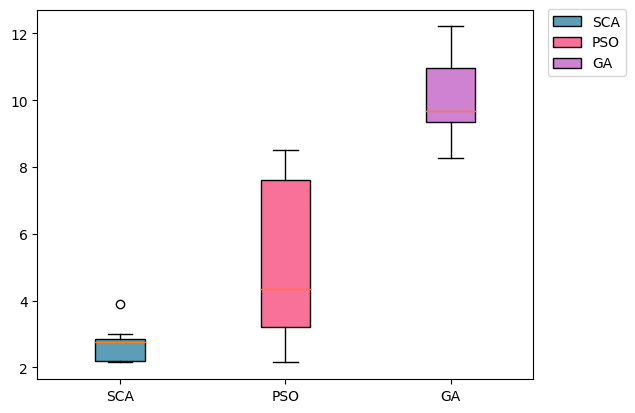
In the earliest iterations of the SCA, we see a dramatic decline in the values of the objective function, which reflects the SCA’s quick convergence to the best solution. In the first iteration, the numbers begin at 28.59 and quickly decrease until they reach 4.67 by the ninth iteration. After that, we see a slow but constant decline that finally reaches a plateau at 2.69 starting with the 39th iteration. SCA’s execution time of 1.09 units indicates that it could take a little bit longer than the other two methods.

The objective function values begin the PSO line chart with a similar quick reduction, going from 26.46 at the first iteration to 5.39 by the ninth. In contrast to SCA, PSO appears to settle significantly earlier, around the 15th iteration, at a value of 5.19, and then stays stable moving forward. PSO’s execution time, 0.96 units, is only a little bit shorter than SCA’s, suggesting that it could be the quicker choice.

The objective function values, however, reveal a more progressive fall in the GA, starting at 29.6 in the first iteration and slowly declining over the course of 50 iterations. It drops to 10.04 by the 50th repetition. GA, on the other hand, takes far less time to execute, at just 0.16 units. According to this, GA may require more iterations to get at the best answer, but in terms of raw execution time, it accomplishes it faster.

In conclusion, all three algorithms show a decreasing trend in the values of the objective function, demonstrating their success in moving closer to the ideal outcome. Nevertheless, they show variations in terms of execution durations, stability, and convergence speed. When selecting an optimization method for a particular issue, careful evaluation of these criteria is necessary in order to strike a balance between the demands of computing efficiency, accuracy, and speed.

The box plot under examination shows, graphically, how three distinct optimization algorithms performed at the run’s 50th iteration. The box plot enables for a more thorough comparison examination of these algorithms’ performance at this point of the process by graphically portraying this particular iteration data.

****

**Box Plot of the Optimization Algorithms**

The box plot for the SCA optimizer shows the range of outcomes, from around 2.15 to 3.9. This demonstrates that, despite some performance dispersion caused by the optimizer, the optimizer’s performance is generally stable, demonstrating a considerable amount of stability in optimization performance through the 50th iteration of the various runs.

The PSO optimizer, on the other hand, produces a larger range of outputs, ranging from 2.15 to 8.5. This wider gap shows that performance at the 50th iteration varies significantly among runs. This might signal that its optimization process is more inconsistent.

Finally, the GA optimizer produces an even wider range of outcomes, with values ranging from 8.28 to 12.21. This large range shows that the optimization performance at the 50th iteration varies significantly between runs, demonstrating significant variability in the GA’s optimization process.

In conclusion, the box plot provides a useful comparison of the performance of the SCA, PSO, and GA optimizers at the 50th iteration. It emphasizes the dispersion and variability in each optimizer’s performance, revealing their relative stability and consistency in optimization at this iteration.

# ***References***

Adra, H. (2014) *Realtime Predictive and Prescriptive Analytics With Real-Time Data and Simulation, Proceedings of the 2016 Winter Simulation Conference.* Available at: <https://www.createasoft.com/Simulation-White-Paper/Realtime-Predictive-and-Prescriptive-Analytics-with-Real-time-Data-and-Simulation>

Amadebai, E. (2022) *12 Data Analysis Techniques You Need To Know, Analytics for Decisions.* Available at: <https://www.analyticsfordecisions.com/data-analysis-techniques/>

Analytics, O. (2022) *What is Prescriptive Analytics - How does Prescriptive Analytics help, Orbit Analytics.* Available at: <https://www.orbitanalytics.com/prescriptive-analytics/>

Bakshi, C. (2020) *Random Forest Regression. Random Forest Regression is a…, Level Up Coding.* Available at: <https://levelup.gitconnected.com/random-forest-regression-209c0f354c84>

Baobab (2020) *Prescriptive analytics: optimisation and simulation, baobab.* Available at: <https://baobabsoluciones.es/en/blog/2020/11/19/prescriptive-analytics-optimisation-and-simulation/>

Beheshti, N. (2022) *Random Forest Regression. A basic explanation and use case in 7…, Towards Data Science.* Available at: <https://towardsdatascience.com/random-forest-regression-5f605132d19d>

Bhatt, B. (2019) *How does SelectKBest work in Feature Selection?, Bhavesh Bhatt.* Available at: <https://www.youtube.com/watch?v=UW9U0bYJ-Ys>

Calzon, B. (2022) *See 10 Big Data Analytics Examples & Applications In Real Life, Data Pine.* Available at: <https://www.datapine.com/blog/big-data-examples-in-real-life/>

Calzon, B. (2023) *What Is Data Analysis? Methods, Techniques, Types & How-To, Data Pine.* Available at: <https://www.datapine.com/blog/data-analysis-methods-and-techniques/>

Coursera (2023) *5 Data Analytics Projects for Beginners, Coursera.* Available at: <https://www.coursera.org/articles/data-analytics-projects-for-beginners>

DataTechNotes (2021) *SelectKBest Feature Selection Example in Python, DataTechNotes.* Available at: <https://www.datatechnotes.com/2021/02/seleckbest-feature-selection-example-in-python.html>

Dilmegani, C. (2023) *Prescriptive Analytics: Optimize Business Decisions in 2023, AI Multiple.* Available at: <https://research.aimultiple.com/prescriptive-analytics/>

Dutta, A. (2020) *Random Forest Regression in Python - GeeksforGeeks, GeeksforGeeks.* Available at: <https://www.geeksforgeeks.org/random-forest-regression-in-python/>

Gad, A. (2018) *Introduction to Optimization with Genetic Algorithm | by Ahmed Gad | Towards Data Science, Towardsdatascience.* Available at: <https://towardsdatascience.com/introduction-to-optimization-with-genetic-algorithm-2f5001d9964b>

GIANLUCA MALATO (2023) *A practical introduction to sequential feature selection, Your Data Teacher.* Available at: <https://www.yourdatateacher.com/2023/02/15/a-practical-introduction-to-sequential-feature-selection/>

Haan, K. (2022) *Best Data Analytics Tools & Software (2022), Forbes Advisor.* Available at: <https://www.forbes.com/advisor/business/software/best-data-analytics-tools/>

Hillier, W. (2023) *The 9 Best Data Analytics Tools for Data Analysts in 2023, Career Foundry.* Available at: <https://careerfoundry.com/en/blog/data-analytics/data-analytics-tools/>

IBM Documentation (2022) *About Linear Regression , Ibm.* Available at: <https://www.ibm.com/topics/linear-regression>

Insight Software (2021) *Comparing Descriptive, Predictive, Prescriptive, and Diagnostic Analytics - insightsoftware, Insight Software.* Available at: <https://insightsoftware.com/blog/comparing-descriptive-predictive-prescriptive-and-diagnostic-analytics/>

KALYAN (2017) *5 Core Activities of Data Analysis | Epicycles of Data Analysis, MAKE ME ANALYST.* Available at: <https://makemeanalyst.com/5-core-activities-data-analysis-epicycles-data-analysis/>

Karan, R. (2022) *Types of Data Analytics and their Applications in Real World, Shiksha.* Available at: <https://www.shiksha.com/online-courses/articles/types-of-data-analytics-and-their-applications-in-real-world/>

Kingston, M. (2022) *Game Theory for Data Scientists. The Basics: Simultaneous Move Games, Towards Data Science.* Available at: [https://towardsdatascience.com/game-theory-666abe63215e](https://towardsdatascience.com/game-theory-666abe63215e%20)

Kuttappa, S. (2020) *Why prescriptive analytics and decision optimization are crucial, Ibm.* Available at: <https://www.ibm.com/blogs/journey-to-ai/2020/04/why-prescriptive-analytics-and-decision-optimization-are-crucial/>

MALI, K. (2023) *Linear Regression | Everything you need to Know about Linear Regression, Analytics Vidhya.* Available at: <https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/>

Mirjalili, S. (2016) *SCA: A Sine Cosine Algorithm for solving optimization problems, Knowledge-Based Systems.* Available at: <https://doi.org/10.1016/j.knosys.2015.12.022>

Naveen (2020) *Top 10 Data Analytics Applications In Real Life, Intellipaat.* Available at: <https://intellipaat.com/blog/tutorial/data-analytics-tutorial/data-analytics-applications/?US>

Project Pro (2023a) *Is Game Theory important for Data Scientists?, Project Pro.* Available at: <https://www.projectpro.io/article/is-game-theory-important-for-data-scientists/139>

Project Pro (2023b) *Unleashing the Types of Analytics: Categories and Applications, Project Pro.* Available at: <https://www.projectpro.io/article/types-of-analytics-descriptive-predictive-prescriptive-analytics/209>

Raschka, S. (2014) *SequentialFeatureSelector: The popular forward and backward feature selection approaches incl. floating variants, Github.* Available at: [http://rasbt.github.io/mlxtend/user\_guide/feature\_selection/SequentialFeatureSelector/#:~:text=RFE is computationally less complex,defined classifier%2Fregression performance metric.](http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/%23:~:text=RFE%20is%20computationally%20less%20complex,defined%20classifier%2Fregression%20performance%20metric.)

Rossetti, M.D. (2018) *Simulation Modeling and Arena, Journal of Materials Processing Technology.* Available at: [http://dx.doi.org/10.1016/j.cirp.2016.06.001%0Ahttp://dx.doi.org/10.1016/j.powtec.2016.12.055%0Ahttps://doi.org/10.1016/j.ijfatigue.2019.02.006%0Ahttps://doi.org/10.1016/j.matlet.2019.04.024%0Ahttps://doi.org/10.1016/j.matlet.2019.127252%0Ahttp://dx.doi.o](http://dx.doi.org/10.1016/j.cirp.2016.06.001%0Ahttp:/dx.doi.org/10.1016/j.powtec.2016.12.055%0Ahttps:/doi.org/10.1016/j.ijfatigue.2019.02.006%0Ahttps:/doi.org/10.1016/j.matlet.2019.04.024%0Ahttps:/doi.org/10.1016/j.matlet.2019.127252%0Ahttp:/dx.doi.o)

Rouse, M. (2012) *What is prescriptive analytics? - Definition from WhatIs.com, TechTarget.* Available at: [https://www.techtarget.com/searchcio/definition/Prescriptive-analytics%0Ahttps://searchcio.techtarget.com/definition/Prescriptive-analytics](https://www.techtarget.com/searchcio/definition/Prescriptive-analytics%0Ahttps:/searchcio.techtarget.com/definition/Prescriptive-analytics)

Sahiti Kappagantula (2023) *Top 10 Data Analytics Tools in 2023 | Data Analytics Tools, Edureka.* Available at: <https://www.edureka.co/blog/top-10-data-analytics-tools/>

Singh, A. (2023) *KNN Algorithm: Guide to Using K-Nearest Neighbor for Regression, Analytics Vidhya.* Available at: <https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/>

Stevens, E. (2022) *The 7 Most Useful Data Analysis Techniques [2023 Guide], Career Foundry.* Available at: <https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/>

Tam, A. (2021) *A Gentle Introduction to Particle Swarm Optimization - MachineLearningMastery.com, Machine Learning Mastery.* Available at: <https://machinelearningmastery.com/a-gentle-introduction-to-particle-swarm-optimization/>

The University of Sidney (2021) *2 K-nearest Neighbours Regression | Machine Learning for Biostatistics, The University of Sidney.* Available at: <https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html>

University of Bath Online (2022) *Three Types of Business Analytics, University of Bath Online.* Available at: <https://online.bath.ac.uk/content/descriptive-predictive-and-prescriptive-three-types-business-analytics>

Vadapalli, P. (2022) *14 Fascinating Data Analytics Real Life Applications in 2023, Upgrad.* Available at: <https://www.upgrad.com/blog/top-challenges-in-artificial-intelligence/>

Wijesooriya.K (2017) *‘Naval postgraduate’, Security, (June), pp. 1–55.* Available at: <https://apps.dtic.mil/sti/pdfs/AD1046101.pdf>