Data Analytics using Python.

Data Analytics

June 18, 2023

Sama al-qaisi

Dr. Bassam Kassbeh

Table of Contents

[1 Introduction 2](#_Toc136556945)

[1.1 Data analytic activities 2](#_Toc136556946)

[1.2 Data analytic techniques 3](#_Toc136556947)

[1.3 Data analytic tools 4](#_Toc136556948)

[1.4 Types of data analytic methods 4](#_Toc136556949)

[1.5 Uses of data analytic methods in real life 4](#_Toc136556950)

[2 Descriptive Analytics 4](#_Toc136556951)

[2.1 Techniques & Examples (Your Work) 4](#_Toc136556952)

[2.1.1 Features Analysis and Explanation 4](#_Toc136556953)

[2.1.2 Features Visualization and Explanation 4](#_Toc136556954)

[2.1.3 Contingency Table and Explanation 5](#_Toc136556955)

[2.2 Techniques for decision-making (Your work) 5](#_Toc136556956)

[2.3 Evaluation (Your work) 5](#_Toc136556957)

[3 Predictive Analysis 5](#_Toc136556958)

[3.1 Techniques & Examples (Your Work) 5](#_Toc136556959)

[3.1.1 Feature Selection Techniques 5](#_Toc136556960)

[3.1.2 Regression Techniques 5](#_Toc136556961)

[3.2 Compare Techniques (Your Work) 5](#_Toc136556962)

[3.3 Evaluation (Your work) 6](#_Toc136556963)

[4 Prescriptive Analysis 6](#_Toc136556964)

[4.1 Techniques with Examples (General) 6](#_Toc136556965)

[4.2 Techniques for finding the best course of action (General) 7](#_Toc136556966)

[4.3 Objective Function Code (Your Work) 7](#_Toc136556967)

[4.4 Apply the Techniques (Your Work) 7](#_Toc136556968)

[4.4.1 Code Screenshots 7](#_Toc136556969)

[4.4.2 Results and Explanation 7](#_Toc136556970)

[4.4.3 Visualization and Explanation 7](#_Toc136556971)

[References 7](#_Toc136556972)

# **Introduction**

## Data analytic activities

* ***What is data analytics?***

Data Analytics is a broad term that focuses on the process of analyzing and understanding data to find useful information. It entails tasks and activities which I will go through next. It works on assisting organizations in making informed decisions, improving productivity, and gaining insights for success.(1, 2023)

* ***What do we do in data analytics?***

The process of analyzing data consists of 5 stages, as follows:

* **Gather the Data you want to analyze:**

In this activity, you must collect data about the problem that will help you understand it better. The data can be acquired from different sources, data analysts usually work on collecting data themselves from for example surveys, or sometimes they lean toward buying datasets from the owners or a data collection specialist, the collected data can be saved in different formats such as a comma-separated value format (CSV file) or (JSON file)(2, 2023).

* **Data cleaning & pre-processing:**

This activity will involve preparing the data by ensuring that the data is devoid of any missing values, feature scaling has been performed on the columns, avoiding any duplicate data, and removing outliers, and that can be done by a programming language(2, 2023).

* **Analyze the data:**

In this activity, we need to understand our collected data, and that can be done using data analytics techniques for example statistical measures, some data analysts work on visualizing some features that they believe would correlate, and these visuals work on identifying patterns as well as trends in the data, this activity will help the data analysts on making decisions regarding how to handle the data(2, 2023)**.**

* **Interpret the analysis of the results:**

This activity is to interpret your findings after visualizing, and that can be done by understanding the results you got and connecting them with the main objectives and goals of this analysis(2, 2023).

* **Inform the stakeholders of the findings:**

After finding the results of the analysis it is the core of the data analysts’ job to share these findings with the stakeholders, and you can apply that by presenting after that your answers and results using visuals or reports(2, 2023).

* ***Benefits of data analytics***
* **Informed Decision-Making:**

Based on the analysis of the data, there will be insights gained that will help the companies in making decisions based on evidence, some of the data analytics types indicate what would happen as a result of the changes generated from the analysis and the best way to react to these changes(3, 2023).

* **Mitigating Risk:**

Data analytics can assist companies to identify potential risks from the historical data that have been collected and the analysis can discover un-normal behaviors that can indicate having a risk before it happens and work on avoiding it(3, 2023).

* **Enhancing and Improving the Customer Experience:**

Analyzing customer data, organizations can gain significant insights into customer behavior, preferences, and requirements by analyzing customer data. This data enables customized marketing efforts, individualized products or services, and increased consumer pleasure and loyalty(4, 2023).

* **Increase the efficiency of work:**

Within a business, data analytics is critical for discovering inefficiencies, bottlenecks, and possibilities for process improvement. Organizations can streamline processes, optimize workflows, and improve overall efficiency by evaluating data, resulting in cost savings and higher productivity(4, 2023).

## Data analytic techniques

## To implement the process of data analytics, there are some techniques to do so, here are some of the techniques:

## ***Exploratory Data Analysis:***

The EDA will work on exploring the data by visualizing the features as well as identifying a trend and unusual behavior, visualizing depends on which type of EDA you are implementing, the types of EDA are Univariate analysis (to analyze all the characteristics of one feature), Bivariate Analysis (to analyze by comparing between two features), Multivariate Analysis (to analyze by comparing between more than 2 features)(5, 2023).

* ***Statistical Measures:***

Are the measures that will provide you with numeric results that will help in summarizing the features in the dataset, there are different measures like measures of frequency (frequency, relative frequency, cumulative frequency), measures of dispersion (range, variance, standard deviation), measures of central tendency (mean, mode, median), and measures that will detect outliers(6, 2023).

* ***Time Series Analysis:***

Time Series Analysis is the process of analyzing data collected over time to uncover patterns, trends, and relationships that allow for future predictions and insights. It entails analyzing patterns, seasonality, autocorrelation, and forecasting based on the findings(7, 2023)**.**

## Data analytic tools

* ***Power BI:***

Microsoft Power BI is a business intelligence application that allows users to build interactive dashboards, reports, and infographics. It may connect to various data sources, whether local or cloud-based, allowing users to explore and analyze data through a simple drag-and-drop interface. Data modeling tools are also included in Power BI for defining associations between datasets and performing calculations. It is frequently used in businesses of all sizes for data visualization and reporting(8, 2023).

* ***Excel:***

Excel is a popular spreadsheet tool from Microsoft used for data analysis and manipulation. It has several built-in functions and formulas for doing computations, creating charts, and analyzing data. Its pivot table functionality, in particular, simplifies the summary and analysis of massive datasets. Excel is adaptable, accommodating to users of varying technical levels, and is often used for activities such as data cleansing, filtering, and aggregation. Its advantages include its ease of use, rich learning resources, and ability to perform modeling, visualization, and report-generating activities(8, 2023).

* ***Python:***

Python is a language known for its simple syntax, ease of reading, and the variety of libraries that can be provided. Data scientists are interested in using it because it can be used for data analysis, machine learning, and data visualization. Python is considered a high-level language, but its simplicity makes it a great language for beginners(8, 2023)**.**

* ***R:***

R is a statistical computing and graphics-oriented programming language and environment. It's widely used in data analysis tasks like cleaning, visualization, statistical hypothesis testing, modeling, and report generation. Because of its full statistical capabilities, including data manipulation, exploratory data analysis, hypothesis testing, regression analysis, and sophisticated statistical modeling, R is highly valued by statisticians and academics. Its strength is in handling complex statistical analysis while producing visually pleasing results(8, 2023).

* ***SQL:***

SQL for data analysis entails interacting with many databases and doing data analysis activities using SQL, a database querying language. It is a versatile and extensively used programming language that enables the development of complex analytics tools and dashboards. SQL's popularity stems from its ease of use and capacity to perform sophisticated data analysis jobs. It can be used to create user-friendly dashboards, reporting tools, and data warehouses. SQL can also be combined with other frameworks such as Python to improve data science and large data management functionality. SQL, in general, acts as a link between end-users and sophisticated data storage systems, allowing for efficient data analysis and manipulation(9, 2023).

## Types of data analytic methods

* ***Descriptive analytics:***

Descriptive analytics is about analyzing historical data to gain insights into past events and trends. It summarizes and interprets the data using statistical measurements and data exploration tools. Without generating predictions or judgments, descriptive analytics provides a platform for knowledge. It presents the data in an intelligible manner by using simple computations and visual tools. Identifying essential indicators, gathering, and processing data, analyzing it for patterns, and presenting the results are all part of the process. Descriptive analytics is useful for finding areas for improvement and day-to-day operations. It does not, however, delve deeper into the data. Descriptive analytics can be used to summarize sales data, analyze social media participation, and report on general patterns(10, 2020).

* ***Predictive analysis:***

Predictive analytics forecasts future outcomes using historical data and statistical models. It recognizes patterns and generates predictions using machine learning methods. Predictive analytics improves efficiency, customer service, fraud detection, and risk management by enabling initiative-taking decision-making. It is based on historical data and has applications in a wide range of industries, including healthcare, e-commerce, sales, human resources, IT security, and healthcare resource planning(10, 2020)**.**

* ***Prescriptive analysis:***

Prescriptive analytics is the final stage of business analysis in which suggestions are made based on previous descriptive and predictive analyses. It entails complex methodologies and experience to provide the best choice possibilities and capitalize on future chances. Prescriptive analytics considers a variety of probable events and their effects, improving corporate strategy and performance. Nonetheless, it is based on a large amount of data and may not fully account for external circumstances. Prescriptive analytics examples include GPS technology for route optimization and its uses in industries such as manufacturing, healthcare, insurance, and pharmaceutical research(10, 2020).

## Uses of data analytic methods in real life

* ***Descriptive analytics:***

Case Study: Using Descriptive Analytics to Improve the User Experience on Netflix's Streaming Platform.

Netflix has used descriptive analytics to evaluate user data and improve its streaming product. This strategy has led to the implementation of numerous measures to improve the user experience:

* Personalized Content suggestions: Netflix uses descriptive analytics to provide tailored content suggestions based on viewer data. This guarantees that users receive ideas that are appropriate to their preferences and interests.
* Enhanced Viewer Engagement: Using descriptive analytics, Netflix can track and analyze user engagement metrics to better understand viewer behaviors and preferences. This insight aids in the optimization of the user interface, the introduction of binge-watching features, and the tailoring of content distribution methods for maximum viewer engagement.
* Data-Driven Content Evaluation: Netflix examines the performance of its original and licensed content using descriptive statistics. They use statistics such as viewership, ratings, and user comments to make informed decisions about content renewal, production spending, and prospective acquisitions.

The usage of descriptive analytics by Netflix has considerably improved the user experience on their streaming platform. The capacity to make data-driven content decisions, deliver personalized suggestions, and enhance user engagement has helped their success in the highly competitive streaming sector(11, 2021).

* ***Predictive analysis:***

Predictive Analytics in Finance: Enhancing Risk Management and Driving Growth in a Volatile Market.

This case study looks at how predictive analytics is used in the banking business, where complex statistical algorithms and machine learning approaches are used to forecast future events based on historical data. Financial institutions may make more informed decisions, improve risk management processes, and drive growth by leveraging the potential of predictive analytics. In this industry, key predictive analytics applications include:

Predictive analytics assists firms in anticipating prospective dangers and taking proactive efforts to mitigate them by examining prior data and detecting patterns. Predictive analytics, for example, is used by banks to detect and prevent fraudulent actions by analyzing client data and spotting unexpected patterns in account activity.

Investment Decisions: Firms such as BlackRock use predictive analytics to evaluate massive volumes of data on stocks, bonds, and other assets. This allows them to spot good investment opportunities and optimize their portfolios for higher returns.

client Segmentation: Financial firms such as American Express can evaluate client data and adapt financial goods and services to specific consumer categories using predictive analytics. This tailored approach aids in the retention and loyalty of customers.

Fraud Detection: FICO specializes in fraud detection through predictive analytics, and its solutions are used by financial institutions all over the world. Predictive analytics helps prevent financial fraud and protect consumer assets by evaluating transaction data and recognizing questionable patterns or behaviors.

These real-world examples show how predictive analytics may be successfully integrated into finance. Businesses may improve risk management processes, optimize investment strategies, provide personalized customer experiences, and protect themselves from fraudulent activity by employing data-driven insights(12, 2023).

* ***Prescriptive analysis:***

In this real-world case study, titled "Predict, then Schedule: Prescriptive Analytics Approach for Machine Learning-Enabled Sequential Clinical Scheduling," prescriptive analytics was used to improve the efficiency of a healthcare appointment system for sequential patient scheduling.

The main goal of the implementation was to reduce patient waiting time, doctor idle time, and overtime. Advanced machine learning techniques were integrated into the appointment system design using a predict-then-schedule architecture. Machine learning models were created to predict patient-specific parameters like the frequency of no-shows and the length of appointments. These forecasts were then used to make educated scheduling decisions such as resource allocation, appointment sequencing, and overbooking techniques.

This implementation's results demonstrated consistent and considerable gains over previous approaches, resulting in a 60% boost in efficiency. The appointment system was able to optimize scheduling decisions by taking into consideration individual patient characteristics and uncertainties, such as the likelihood of no-shows and projected consultation durations. This resulted in shorter wait times for patients.

Prescriptive analytics in real-world healthcare settings has various advantages. Healthcare firms may optimize their appointment systems, allocate resources more effectively, and improve the entire patient experience by leveraging the power of machine learning predictions. Personalized scheduling decisions based on individual patient characteristics help to increase resource usage and patient satisfaction.

This case study, "Predict, then Schedule: Prescriptive Analytics Approach for Machine Learning-Enabled Sequential Clinical Scheduling," demonstrates how prescriptive analytics may be successfully integrated into real-world healthcare systems. Healthcare facilities can use this strategy to enhance efficiency and provide better patient care. The case study demonstrates the possibility of improving healthcare scheduling systems and optimizing resource allocation in real-world healthcare settings(13, 2023).

# **Descriptive Analytics**

## Techniques & Examples (Your Work)

### Features Analysis and Explanation

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature no.** | **Feature Name** | **Descriptive Measure / Technique** | **Explanation** |
| **1** | **Market** | **Frequency (measures of frequency)** | I have applied the measures of frequency on the Market feature because the frequency will help me in knowing how many trades happened in each market and give me some information about the distribution of the trades in each market, on the other hand, the relative frequency will provide me with the proportion of the trades that have happened in each market relative to the total number of trades. also determining which market has a higher or lower trading activity and comparing the trading activity between the market’s sections. |
| **Relative frequency (measures of frequency)** |
| **2** | **VOLUME** | **Variance (measures of dispersion)** | I have applied the measures of dispersion on the Volume feature because the variance and the standard deviation will help in understanding the variability or volatility in the process of trading so for example if I have a higher variance or standard deviation will lead me to know that there is a huge turn in the volume of trading which indicates either huge variability or volatility. |
| **Standard deviation (measures of dispersion)** |
| **4** | **(TRADE\_QTY, VOLUME, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY)** | **Detect outliers (measures of dispersion)**  **Z-score method**  **IQR method** | I have applied the measures of dispersion to the following features, to detect if they have outliers or not, I have implemented the two methods to detect which are the Z-score and IQR methods. detecting the outliers will help me in Knowing if the data I have, contain some values that are unnormal, and also to identify unusual behavior that will be happening in the trading process or in the features I have implemented it on, such as discovering extreme trade volumes, unusual bids, or major deviations from normal trading patterns. These outliers can provide useful information for understanding market dynamics, finding unique trading opportunities, and spotting potential market irregularities. |

### Features Visualization and Explanation

**Feature 1: HIGH - Visualizing the highest closing price of a stock activity by months.**

A picture containing text, colorfulness, screenshot, plot

Description automatically generated

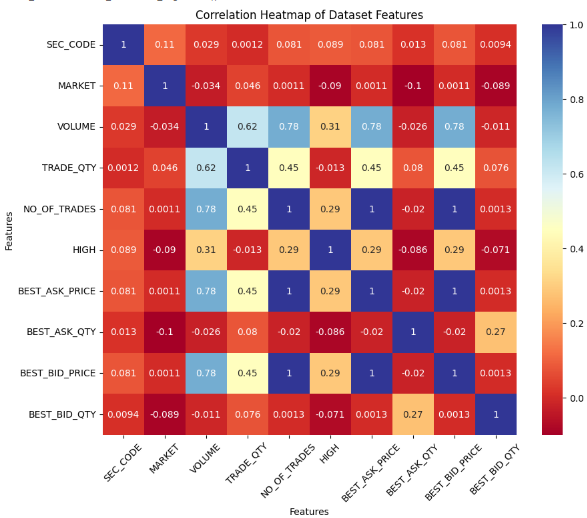
This plot will show the high stock closing prices by the month, which will help me in identifying the price levels or trends during different months, there will be months that have a higher activity during certain months and there will be months where the activity is the lowest, so I will be able to compare the bars through the months and observe if I have consistency for the price or not, so if there is an inconsistency I will work on finding a solution to gain the consistency that I want.

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

Description automatically generatedFeature 2: VOLUME - Visualizing the total number of shares bought or sold in a specific period or during the trading day in months.**

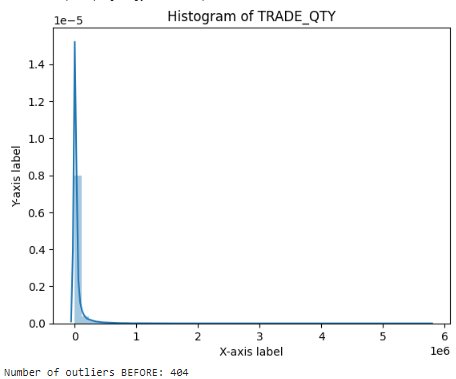
This plot will show the trading volume activity through the months, which will help me in identifying the month where the trading volume was at the highest and the months where the trading volume was at the lowest, which will guide me to implement a strategy to enhance the performance of the trading volume in certain months, also the plot will help me in identifying some patterns that will generate in trading volume over time, which will lead me to compare the bars through the months if they are consistent in the Hight of high and low trading volume over a specific month, I can benefit from this information in understanding the seasonal variation and the behavior of the buyers or the sellers.

**Feature 3: All Numeric columns (SEC\_CODE, MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, HIGH, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY) - Correlation Heatmap.**

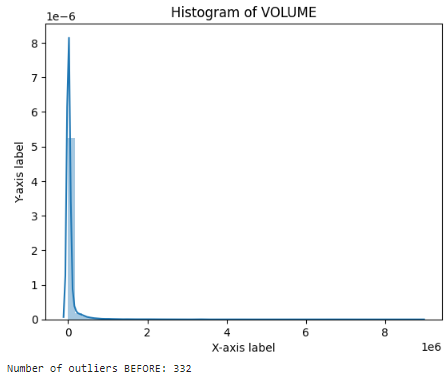


The reason behind implementing the heatmap is to find patterns and correlations as well as interactions between features that I couldn’t conclude just by understanding the data, also the correlation heatmap has helped in selecting features that I want and don’t want to implement the predictive analysis on.

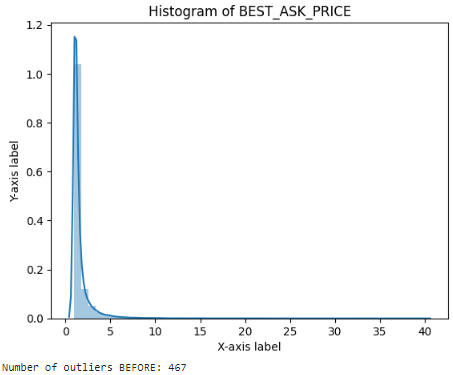
**Feature 4: (TRADE\_QTY, VOLUME, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY)**

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

Description automatically generatedA picture containing text, screenshot, plot, line

Description automatically generatedBefore removing the outliers:**

A picture containing text, screenshot, plot, line

Description automatically generated

A picture containing text, screenshot, diagram, plot

Description automatically generated**A picture containing text, screenshot, plot, diagram

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

Description automatically generatedAfter removing the outliers:**A picture containing text, screenshot, diagram, plot

Description automatically generated

A picture containing text, diagram, plot, screenshot

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

Description automatically generated

These visuals are to show the distribution of every feature that has outliers before removing them and then after removing them, which will show you how the distribution of the values(data) will be changed from a narrow distribution to a wide distribution which will return to its original shape and resemble the expected normal distribution type.

### Contingency Table and Explanation

I have implemented the contingency table on the SYMBOL1/MARKET features because I want to know how many trades happened in each company symbol within each market section, allowing me to identify companies that are actively participating in specific market sections and companies that have a significant presence in the market through multiple sections.

Also, it will guide me to know the frequency of trades for each market, which will lead me to identify which market section is active in trading and which is not. Additionally, the contingency table will show the frequency of trades for each company symbol, enabling me to understand the trading behavior for each company symbol within each market, which means I can identify which company is dominant in some market sections and which is not.

## Techniques for decision-making (Your work)

* ***Market Frequencies:***

Depending on the results I have gained from calculating the frequency and relative frequency lead me to make decisions based on the provided results, which is to focus all my attention on the market with the higher results, which will help me in optimizing the potential profits, by deploy more resources and investments to market sections with greater trade activity.

* ***Volume Volatility:***

The decision has made based on the results I have gained from calculating the variance and the standard deviation for the volume feature, which indicates high variability in the trading volume, which is to make a risk management process because high variance and standard deviation can risk the market, and work on strategies to diversify the investments to reduce potential losses during times of extreme volatility.

* ***Outlier Detection:***

As you have seen before I have calculated the outlier detection using the two methods which are the Z-score method and the IQR method, and have decided on implementing the Z-score, not the IQR because the z-score is sensitive to huge values because the base of the z-score is the mean and standard deviation, on the other hand, the IQR is not sensitive toward extreme values that is why in the z-score it detects smaller numbers than the IQR did, the decision was to handle the discovered outliers by removing them.

* ***Volume / High By Month:***

I have plotted the high and volume features to study the change in their activities during the months and that has led me to adjust trading strategies during certain months.

* ***Correlation Heatmap:***

After studying the heatmap I have concluded that there are some features the correlation between them was 1 which indicates that they have a strong positive relationship between them, so by implementing the feature selection methods they will be selected as the best features, but already know that they have strong relationships, so the decision was to remove them before the feature selection step so they won’t be involved the selection process and the predictive analytics method and left one of them which is the number of trades to predict with it.

## Evaluation (Your work)

After getting the descriptive work done, I need to evaluate my work, I am going to discuss the effectiveness of my results on my decision-making, and how I can enhance my work.

***Features Analysis:***

First, my objective of finding the frequency and relative frequency to the **market** feature is to focus all my attention on the market with the higher results based on the result I have gained market 2 held the higher results for frequency and the relative frequency by 15379, and 54% than the rest of the markets, which have helped me to meet my objective by deploy more resources and investments to market 2 that has greater trade activity, which was how I met my objective.

Second, my objective of finding the variance and the standard deviation for the **volume** feature is to make a risk management process, and based on the gained results the variance and standard variation were high by 91355285832.59985 and 302250.369449898 which lets me decide on working on strategies to diversify the investments to reduce potential losses, which was how I met my objective.

Third, my object of finding the outliers detection to the certain features was to handle them and make the data as outliers free, based on the results I got that the features **TRADE\_QTY, VOLUME, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, and BEST\_BID\_QTY,** by the order was 404, 332, 467, 542, 467, and 428 the total of them is approximately 2500 and the whole data set is approximately 28000, so removing them won’t take a huge amount of the data which led me to handle them by removing them, which was how I met my objective.

***Features Visualization:***

Fourth, my objective for plotting the **High** feature by month was to identify the price levels or trends during different months and observe if I have consistency for the price or not, so if there is an inconsistency I will work on finding a solution to gain the consistency that I want, and the results of my observation are that in the month May there was the high price level trends, on the other hand, month February held the lowest price level which has led me to meet my objective by me adjusting trading strategies during certain months.

Also, my object for plotting the **volume** feature by month was to identify the seasonal variation and the behavior of the buyers or the sellers, the result I gained from my observation is that the month where the trading volume was at the highest in May and the month where the trading volume was at the lowest in November, which led me to meet my objective by implementing a strategy to enhance the performance of the trading volume in certain months.

Fifth, my objective for the correlation heatmap was to study my data very closely and if I can observe something that will help in the performance of my project, the results I got from my observing that the **NO\_OF\_TRADES, BEST\_ASK\_PRICE, BEST\_BID\_PRICE** where the features that have a strong positive relation which led me to meet my objective was to remove them before the feature selection step so they won’t be involved the selection process and the predictive analytics method and left one of them which is the **NO\_OF\_TRADES** to predict with it, which will enhance the performance of my project.

Now, After talking about how my objectives were met, I will take about how I can enhance my work:

***Expand My Analysis:***

I can implement more descriptive measures on my features like the measures of central tendency to complement my measures of frequency, also to complement my data distribution I could calculate the percentiles.

***Include More Data Visualization & Techniques:***

I did already include a good amount of features visualization, but I could include more charts and plots to cover a wider range of data to gain more accurate insights that will influence my decision-making strategies.

***Include Multivariate Analysis:***

Multivariate analysis is to analyze by comparing more than 2 features, so to not limit our analysis to just one feature we can combine more than one feature to conduct a wide range of analyses in which we can discover relations that have been overlooked.

***Use Advanced Outlier Detection Techniques:***

As I have mentioned earlier, I have tried the two methods to identify outliers which are the Z-score method and the IQR method then I have settled with the z- score for the reasons I have clarified but there are more advanced techniques that I could implement other than the one I already did, which can give me more accurate identification and handling of outliers in my data.

# **Predictive Analysis**

## Techniques & Examples (Your Work)

### Feature Selection Techniques

|  |  |  |  |
| --- | --- | --- | --- |
| **FS no.** | **Name** | **Description** | **Results (Selected Features)** |
| **1** | **SelectKBest method** | SelectKBest is a feature selection method that is imported from the sickit-learn library, usually, the feature selection method will be executed before the implementation of the machine learning model it focuses on finding the important features in the dataset to do the training on them, in this method the number of features is decided based on the value of k that you initialize.  The difference between the KBest method and the other methods is that the SelectKBest rank the highest score that is evaluated by the statistical relation between the target variable and every feature individually, then it works on ranking the features based on their scores and selects the highest k features based on the highest score.(14, 2023) | **HIGH dataset:**  Linear Regression Model:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY']  K-Nearest Neighbors Regressor Model:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY']  Decision Tree Regression Model:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY']  **LOW dataset:**  Linear Regression Model:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY']  K-Nearest Neighbors Regressor Model:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY']  Decision Tree Regression Model:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY'] |
| **2** | **Sequential method** | Sequential is a feature selection method that is imported from the sickit-learn library, usually, the feature selection method will be executed before the implementation of the machine learning model it focuses on finding the important features in the dataset to do the training on them.  Sequential specialization is reaching the best feature subset that increases the machine learning model’s performance. It is divided into two approaches: forward selection and backward selection. (15, 2023) | **HIGH dataset:**  Linear Regression Model:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY']  K-Nearest Neighbors Regressor Model:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'NO\_OF\_TRADES', 'z\_score']  Decision Tree Regression Model:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'NO\_OF\_TRADES', 'BEST\_ASK\_QTY']  **LOW dataset:**  Linear Regression Model:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY']  K-Nearest Neighbors Regressor Model:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY']  Decision Tree Regression Model:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY'] |

### Regression Techniques

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
| **1** | **Linear Regression (LR)** | It is a supervised machine learning model that works on finding the best fit of the linear relationship between the dependent and independent variables, which is done by decreasing the sum of squared differences between the observed and predicted values by calculating their coefficients.  There is simple linear regression that has one independent variable on the other hand there is multi-linear regression which has two and more independent variables. (16, 2023) |
| **2** | **K-Nearest Neighbors Regressor (KNNR)** | Is a supervised machine learning model that works on finding the nearest neighbors when we train the data and using the target feature values to make a regression prediction.  The model focuses on identifying the K nearest neighbors based on the Euclidean distance, which measures the distance between the data points. and choose the similar in terms of their feature values. after that, the regression prediction will be after determining the k nearest neighbors. (17, 2023) |
| **3** | **Decision Tree Regression (DTR)** | Is a supervised machine learning model that works on regression tasks that focuses on building a tree structure, each node of every tree is identified as a decision node, and every leaf nod for the tree represents a predicted value. (18, 2023) |

## Compare Techniques (Your Work)

**“Low” Prediction**

**Comparison:  
Table1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **1** | **1** | **1.61** | **18.69** | **4.32** | **0.11** |
| **1** | **2** | **0.75** | **7.00** | **2.64** | **0.66** |
| **1** | **3** | **0.28** | **2.07** | **1.43** | **0.90** |
| **2** | **1** | **1.60** | **17.50** | **4.18** | **0.17** |
| **2** | **2** | **0.17** | **0.84** | **0.91** | **0.96** |
| **2** | **3** | **0.14** | **0.53** | **0.72** | **0.97** |

**Visualization of results:**

**Mean Squared Error for all the models in each feature selection:**

A picture containing text, screenshot, line, diagram

Description automatically generated

**R2 Score for all the models in each feature selection:**

A picture containing text, screenshot, line, diagram

Description automatically generated

**R2 Score for the Decision Tree Regression Model using the Sequential as a feature selection (the best R2 Score out of all the models in each feature selection):**

A picture containing diagram, screenshot, text, line

Description automatically generated

**Mean Squared Error for the Decision Tree Regression Model using the Sequential as a feature selection (the best MSE out of all the models in each feature selection):**

A picture containing diagram, screenshot, line, rectangle

Description automatically generated

**“High” Prediction**

**Comparison:**

**Table 2:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **1** | **1** | **1.64** | **19.32** | **4.39** | **0.11** |
| **1** | **2** | **0.76** | **7.26** | **2.69** | **0.66** |
| **1** | **3** | **0.29** | **2.14** | **1.45** | **0.90** |
| **2** | **1** | **1.63** | **18.05** | **4.24** | **0.17** |
| **2** | **2** | **0.31** | **2.26** | **1.5** | **0.89** |
| **2** | **3** | **0.24** | **1.65** | **1.28** | **0.92** |

**Visualization of results:**

**Mean Squared Error for all the models in each feature selection:**

A picture containing text, screenshot, line, diagram

Description automatically generated

**R2 Score for all the models in each feature selection:**

A picture containing screenshot, text, line, diagram

Description automatically generated

**R2 Score for the Decision Tree Regression Model using the Sequential as a feature selection (the best R2 Score out of all the models in each feature selection):**

A picture containing diagram, screenshot, line, rectangle

Description automatically generated

**Mean Squared Error for the Decision Tree Regression Model using the Sequential as a feature selection (the best MSE out of all the models in each feature selection):**

A picture containing diagram, line, rectangle, technical drawing

Description automatically generated

## Evaluation (Your work)

**Following the machine learning models and visualization stages, the result analysis step is required to gain insights into how the model performed, whether it did well or underperformed in certain areas.**

**The outcomes of my experiment were produced by the three machine learning models that I applied to my dataset, which are the Linear Regression (LR), K-Nearest Neighbors Regressor (KNNR), and Decision Tree Regression (DTR). Each one of these models was implemented on two datasets (HIGH, LOW), and each data set applied two feature selection methods (SelectKBest, Sequential), so to summarize there was the high dataset that have the high as a target feature and the low data set that have the low as a target feature, and each dataset have applied two feature selections to each model.**

**I will go through the tables filled with the results above:**

**First, let us start with the HIGH dataset, I will go through viewing the average that I have found for each model, going back to Table 2, it was shown that the Decision Tree Regression model using the Sequential method held the best results of the four measures (MAE, MSE, RMSE, R2) out of the rest of the models by 0.24, 1.65, 1.28, 0.92. However, the Linear Regression model held the best results using the Sequential method than the SelectKBest by 1.63, 18.05, 4.24, 0.17, Moreover, the K-Nearest Neighbors Regressor model held the best result using the Sequential method than the SelectKBest by 0.31, 2.26, 1.5, 0.89.**

**Second the LOW dataset, I will go through viewing the average that I have found for each method, going back to Table 1, it was shown that the Decision Tree Regression model using the Sequential method held the best results of the four measures (MAE, MSE, RMSE, R2) out of the rest of the models by 0.14, 0.53, 0.72, 0.97. However, the Linear Regression model held the best results using the Sequential method than the SelectKBest by 1.60, 17.50, 4.18, 0.17, Moreover, the K-Nearest Neighbors Regressor model held the best result using the Sequential method than the SelectKBest by 0.17, 0.84, 0.91, 0.96.**

**Now, I will go through the plots I have implemented:**

**First, for the HIGH dataset, I implemented two boxplots the first plot had the best R2 score out of all the models with the two different feature selection methods which were for the Decision Tree Regression model using the Sequential method, and the second boxplot was the best Mean Squared Error out of all the models with the two different feature selection methods which was the Decision Tree Regression model using the Sequential method. The box plot combined the R2 score and the MSE of each iteration for the machine-learning model, as is shown in the graph above the box plot provides us with information on the maximum, minimum value, first and third quartile, as well as the median. As you can see the two box plots display normal distributed values and if you see there are not any outliers which conclude in not having any unusual observations.**

**There were also two count plots one for the MSE and one for the R2 score for all the models with the two different feature selection methods, which can show you the variation of the bar’s height to determine the highest result, as you can see in the R2 score plot, the Decision Tree Regression model using the Sequential method was the higher bar than the rest, and the lower bar than the rest in the MSE plot.**

**Second, for the LOW dataset, I implemented two boxplots the first plot had the best R2 score out of all the models with the two different feature selection methods which were for the Decision Tree Regression model using the Sequential method, and the second boxplot was the best Mean Squared Error out of all the models with the two different feature selection methods which was the Decision Tree Regression model using the Sequential method. The box plot combined the R2 score and the MSE of each iteration for the machine-learning model, as is shown in the graph above the box plot provides us with information on the maximum, minimum value, first and third quartile, as well as the median. As you can see the two box plots display normal distributed values and if you see there are not any outliers which conclude in not having any unusual observations.**

**There were also two count plots one for the MSE and one for the R2 score for all the models with the two different feature selection methods, which can show you the variation of the bar’s height to determine the highest result, as you can see in the R2 score plot, the Decision Tree Regression model using the Sequential method was the higher bar than the rest, and the lower bar than the rest in the MSE plot.**

# **Prescriptive Analysis**

## Techniques with Examples (General)

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | **Decisions trees** | Is a model that represents the shape of a tree. Each node is the feature, each branch is the decision rule, and the leaf node is the result.  Example: predicting if a loan applicant will be approved. (19, 2023) |
|  | **Linear programming** | It is an optimization technique based on math that will work on finding the best outcome by minimizing and maximizing the defined linear objective function based on a given restriction.  Example: Estimate the greatest amount of cake that can be baked using 5kg of wheat and 1kg of fat, because one cake requires 200g of flour and 25g of fat, while another requires 100g of flour and 50g of fat. (20, 2023) |
|  | **Sensitivity analysis** | Is a technique used to understand if the parameters change and how they will affect the results. which will allow us to assess the response of the model regarding different factors and identify the essential business components.  Example: A sensitivity study can include evaluating how much a company's prices will rise before they begin to negatively impact customer sales. In this case, it's a tool to evaluate your customers’ sensitivity to price changes so that the corporation can choose how much to raise them. (21, 2023) |

## Techniques for finding the best course of action (General)

* **Gradient descent**:

Is a machine learning optimization and mathematical modeling algorithm. This technique aims to find the minimum function by changing the model's parameters to achieve the largest reduction in the objective function. The algorithm finds the gradient that reflects the direction of the largest dip and then executes it. It will continue to update the parameters in small stages proportional to the gradient until the gradient gradually converges toward the optimal solution. It works well when the objective function is differentiable. Gradient, in general, is a useful tool for determining optimal model parameter values that minimize a given function and is widely used in many fields, including machine learning. (22, 2023)

* **Genetic Algorithms:**

Is an optimization algorithm influenced by natural genetics, these algorithms can run a set of possible solutions, represented as parameter groups of chromosomes, and the algorithm progressively improves the population, generating superior solutions, these algorithms excel at solving complex problems with vast search areas and multiple optimal solutions. Genetic algorithms effectively explore and uncover optimal or near-optimal solutions by emulating natural selection and genetic variation. Overall, genetic algorithms provide a trustworthy and resilient approach to optimization in a variety of disciplines. (23, 2023)

* **Ant Colony Optimization (ACO):**

Ant colony optimization (ACO) is a technique that draws inspiration from the foraging behavior of ants to solve combinatorial optimization problems, by simulating the behavior of ants depositing pheromone trails, and directs the ACO algorithms to search for better solutions. The pheromone pathways act as a communication mechanism, directing the other ants to follow the paths that lead to higher-quality solutions. The ACO explores the solution space iteratively, adjusting the pheromone pathways according to the quality of the selected solutions. The algorithm ultimately converges on optimal or near-optimal solutions. ACOs are useful in a variety of problem areas, including the itinerant salesman's dilemma and staffing schedules. ACO provides an efficient and powerful technology for optimization by replicating the collective intelligence of ants. (24, 2023)

## Objective Function Code (Your Work)

# define the function blocks

def obj\_fun(q):

  for i in range(len(q)):

    q[i]=round(q[i])

  min\_cost=sum(numpy.array(q)\*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]))

  if sum(q)<10:

      return 9999

  return min\_cost

## Apply the Techniques (Your Work)

### Code screenshots

A screenshot of a computer

Description automatically generated with medium confidence



A screenshot of a computer code

Description automatically generated with low confidence

### Results and Explanation

I need to use prescriptive technics to find the optimal stock quantity for each bank so that traders can minimize the amount they spend on the stock.

And I will explain the results I got as follow:

As you have seen from the snippet of code, I have provided I have used an equation in the objective function, which will calculate the min cost I can get, following some constraints, which are that each quantity is limited to 5 stocks and total quantity is not less than 10 stocks, I have used the three optimizers which are SSA, PSO, GA, each one of them will have 15 number of run the number of iteration will be 80 and the population size is 100.

Talking briefly about the optimizations:

**SSA:** Simulated annealing (SSA) is a fancy term for an optimization approach that aids in the creation of truly excellent solutions. It's similar to pretending to heat and cool a substance to explore multiple choices and avoid local deadlocks. We can gradually approach the best feasible solution to our problem via SSA.

**PSO:** article Swarm Optimization (PSO) is a sophisticated algorithm inspired by how birds and fish move together. It helps us find the optimal solution by simulating the movement of particles in our problem space. The position of each particle is adjusted based on its own experience and what it learns from neighboring particles. The PSO guides us toward the solution we want by repeating this procedure.

**GA**: A genetic algorithm (GA) is an optimization technique that simulates how nature works. It is more like simulating evolution to meet tough challenges. GA starts with a set of possible answers, similar to a large family. GA breeds new generations that improve over time through a series of selection, mixing, and transformation procedures. By repeating this cycle, GA hopes to find the best, or at least very good, solutions to our complex challenges.

A screenshot of a table

Description automatically generated with low confidenceDepending on running the code, the experiment sheet shows the following:

A screenshot of a table

Description automatically generated with low confidence 🡪



That the optimizer PSO has the best results which is the minimum cost (3.33) at the last iteration, on the other hand, the GA had the higher cost (5.55).

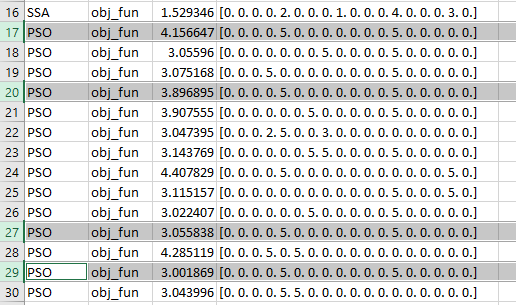
Now going to the experiment details sheet:

A screenshot of a graph

Description automatically generated with low confidence 🡪

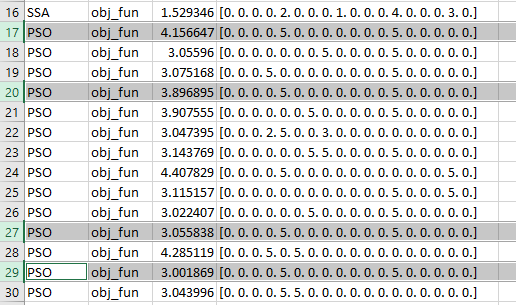


As you can see the PSO held the lower results (2.15) 4 times, now I will go through the individuals and see who is better and who is not, and if they have the same results:





They have the same results, now after comparing them I will go through one of them and clarify on it what is the next move for the trader:

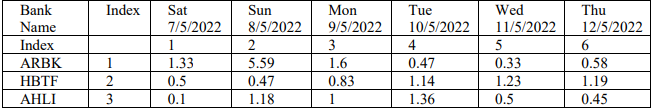




The first 6 elements represent the Bank Name which is ARBK.

The second 6 elements represent the Bank Name which is HBTF.

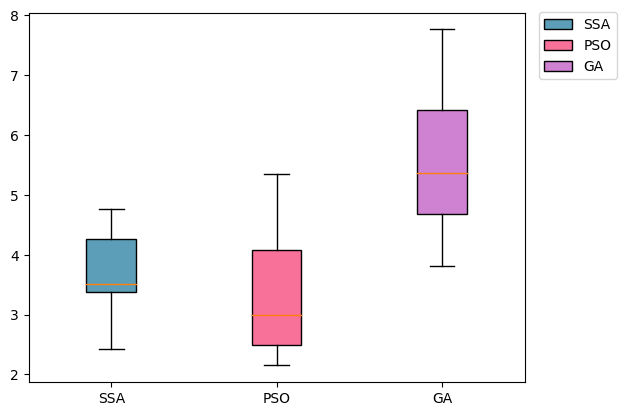
The third 6 elements represent the Bank Name which is AHLI.



Based on the results, the trader should Buy 5 stocks from ARBK bank on Wednesday 11/5/2022, and buys 5 stocks from AHLI bank on Saturday 7/5/2022, but the on the rest of the days the trader should buy zero stocks from the rest of the banks, this will ensure that the trader will spend the minimum amount on the stock which will result on spending 0.33×5=1.65 from the first bank and 0.1×5=0.5 from the third bank.

### Visualization and Explanation

As you can see, this is a chart that shows the three optimization models PSO, SSA, and GA. As you can observe, the PSO was the quickest to reach the minimum value, which it did in 7 iterations, while the SSA took 34 iterations. This minimum value shows the best result, indicating that this model is the most efficient and effective among the other models for finding the minimum value, also it is considered in having some traits that enable it to find the optimal solution more effectively.



As we know, A box plot uses a box and whisker diagram to summarize data distribution. It displays the data’s median, quartiles, and range, highlighting probable outliers. It aids in the assessment of central tendency and variability, as well as the comparison of different datasets or groups. Consider the context and goals of interpretation.

We concluded earlier that the PSO has the best results, and then the SSA then the GA, and that is shown also in the figure above, the GA is the least good box plot because of how it is spread and the min-max values. the PSO has the characteristics that made it have the results it has but if we want to improve, the SSA box plot may be better as you can see the max value for the SSA 4.7 is smaller than the PSO 5.5 which can indicate better performance also in finding the minimum cost it is better to have a box plot with a lower spread and less variance is better for determining the minimum cost since smaller spreads suggest a tighter range between the minimum and maximum values. It is proposed that the cost values obtained by the optimization model are clustered more closely around the minimum, which is desired when aiming to minimize costs since less variance equals less dispersion or variances in cost values. It shows that the optimization model routinely produces outcomes that are near the least cost. Lower variation increases confidence and predictability in attaining the needed cost savings.

So, in conclusion, the PSO held the best result, but to gain improved results the SSA is better because of its characteristics.

# **References**

* 10, 2020. *Descriptive, Predictive & Prescriptive Analytics: What are the differences?* [Online]   
  Available at: https://studyonline.unsw.edu.au/blog/descriptive-predictive-prescriptive-analytics  
  [Accessed 3 6 2023].
* 11, 2021. *How Netflix uses Data Analytics: A Case Study.* [Online]   
  Available at: https://plainenglish.io/blog/how-netflix-uses-data-analytics-data-science-general-research-case-study-4d525b881038  
  [Accessed 6 6 2023].
* 1, 2023. *What Is Data Analytics?* [Online]   
  Available at: https://www.comptia.org/content/guides/what-is-data-analytics  
  [Accessed 2 6 2023].
* 12, 2023. *Predictive Analytics in Finance: Enhancing Risk Management and Driving Growth in a Volatile Market.* [Online]   
  Available at: https://www.linkedin.com/pulse/predictive-analytics-finance-enhancing-risk-driving-ankleshwariya/  
  [Accessed 5 6 2023].
* 13, 2023. *Predict, then schedule: Prescriptive analytics approach for machine learning-enabled sequential clinical scheduling.* [Online]   
  Available at: https://www.sciencedirect.com/science/article/abs/pii/S0360835222003357  
  [Accessed 5 6 2023].
* 14, 2023. *Feature Selection with SelectKBest in Scikit Learn.* [Online]   
  Available at: https://lifewithdata.com/2022/03/19/feature-selection-with-selectkbest-in-scikit-learn/  
  [Accessed 14 6 2023].
* 15, 2023. *Sequential Feature Selection.* [Online]   
  Available at: https://www.mathworks.com/help/stats/sequential-feature-selection.html  
  [Accessed 14 6 2023].
* 16, 2023. *What is linear regression?* [Online]   
  Available at: https://www.ibm.com/topics/linear-regression#:~:text=Related%20solutions-,What%20is%20linear%20regression%3F,is%20called%20the%20independent%20variable  
  [Accessed 14 6 2023].
* 17, 2023. *K-nearest Neighbours Regression.* [Online]   
  Available at: https://bookdown.org/tpinto\_home/Regression-and-Classification/k-nearest-neighbours-regression.html  
  [Accessed 14 6 2023].
* 18, 2023. *Decision Tree - Regression.* [Online]   
  Available at: https://www.saedsayad.com/decision\_tree\_reg.htm#:~:text=Decision%20tree%20builds%20regression%20or,decision%20tree%20is%20incrementally%20developed  
  [Accessed 14 6 2023].
* 19, 2023. *What Is a Decision Tree? Algorithms, Template, Examples, and Best Practices.* [Online]   
  Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-decision-tree/  
  [Accessed 14 6 2023].
* 20, 2023. *What Is Linear Programming?* [Online]   
  Available at: https://www.spiceworks.com/tech/it-strategy/articles/linear-programming/#:~:text=Linear%20programming%20(LP)%20uses%20many,restrictions%20of%20supplies%20and%20personnel  
  [Accessed 14 6 2023].
* 21, 2023. *Understanding Sensitivity Analysis.* [Online]   
  Available at: https://study.com/learn/lesson/sensitivity-analysis-overview-examples.html#:~:text=An%20example%20of%20a%20sensitivity,much%20to%20increase%20them%20by  
  [Accessed 14 6 2023].
* 2, 2023. *What Does a Data Analyst Do?* [Online]   
  Available at: https://www.coursera.org/articles/what-does-a-data-analyst-do-a-career-guide  
  [Accessed 2 6 2023].
* 22, 2023. *Gradient Descent in Machine Learning: A Basic Introduction.* [Online]   
  Available at: https://builtin.com/data-science/gradient-descent  
  [Accessed 14 6 2023].
* 23, 2023. *Introduction to Genetic Algorithms — Including Example Code.* [Online]   
  Available at: https://towardsdatascience.com/introduction-to-genetic-algorithms-including-example-code-e396e98d8bf3#:~:text=Crossover%20is%20the%20most%20significant,random%20from%20within%20the%20genes  
  [Accessed 14 6 2023].
* 24, 2023. *Introduction to Ant Colony Optimization.* [Online]   
  Available at: https://www.geeksforgeeks.org/introduction-to-ant-colony-optimization/  
  [Accessed 14 6 2023].
* 3, 2023. *The benefits of data analytics for your business.* [Online]   
  Available at: https://shout.com/business/benefits-of-data-analytics/  
  [Accessed 2 6 2023].
* 4, 2023. *5 benefits of data analytics for your business.* [Online]   
  Available at: https://www.stitchdata.com/resources/benefits-of-data-analytics/  
  [Accessed 2 6 2023].
* 5, 2023. *Exploratory Data Analysis (EDA): Types, Tools, Process.* [Online]   
  Available at: https://www.knowledgehut.com/blog/data-science/eda-data-science  
  [Accessed 3 6 2023].
* 6, 2023. *Statistical Measures.* [Online]   
  Available at: https://www.studysmarter.co.uk/explanations/math/statistics/statistical-measures/  
  [Accessed 3 6 2023].
* 7, 2023. *Time Series Analysis: Definition, Types, Techniques, and When It's Used.* [Online]   
  Available at: https://www.tableau.com/learn/articles/time-series-analysis#:~:text=What%20is%20time%20series%20analysis,data%20points%20intermittently%20or%20randomly.  
  [Accessed 3 6 2023].
* 8, 2023. *Comparison of Data Analysis Tools: Excel, R, Python, and BI Tools.* [Online]   
  Available at: https://towardsdatascience.com/comparison-of-data-analysis-tools-excel-r-python-and-bi-tools-6c4685a8ea6f  
  [Accessed 3 6 2023].
* 9, 2023. *SQL for Data Analysis.* [Online]   
  Available at: https://www.sisense.com/glossary/sql-for-data-analysis/  
  [Accessed 3 6 2023].