# Report

10204210 Data Analytics

Section number:

Student number: 20120098

Student name: Hasan Alhwietat

Instructor Name: Raneem Qaddoura

# Introduction

## Data analytic activities

1. Data Collection: This process is done by collecting data from various sources such as survey and data set, etc. This process includes identifying important elements and ensuring their quality and accuracy to support the goals of analysis, addressing research questions, identifying the source of data, and helping in the decision-making process. This process must be planned from by defining the goal, the type of data to be collected, and what are the methods for processing and storing the data in order to comply with the objectives and ensure the reliability and validity of the data. [1]
2. Data Cleaning: This process is done by purifying and processing the data by removing duplicates, converting the data type, processing empty values, removing columns and rows, removing outliers, correcting wrong numbers, and treating inaccurate data types, so if the data is not cleaned, it will lead to inappropriate and inappropriate results. It is understandable, so this process helps for the data analysis stage to make the format appropriate and gives clear details and information, and it is easy to understand. This data that has been processed will help to find more useful and meaningful insights. [2]
3. Data Interpretation: This process is done by understanding the collected data, analysing the data results and translating them into useful and meaningful insights to help find solutions to problems and validating the results to make a critical evaluation of these results that will help in making decisions. [3]
4. Data Visualization: This process is done by creating visual representations and visualizations to communicate the results appropriately, including graphs and dashboards based on the types of data that require different visualizations that help in understanding and communicating ideas and help in making more effective decisions and attracting attention to take important insights from things that may It is not important and these visualizations help to communicate the story of each visualization without having to read the information. [4]

## Data analytic techniques

1. Time series: It is a technique of a sequence of data points that are collected or recorded over time to identify patterns and trends. Analysts record data over a specific period of time rather than data intermittently and randomly and show how the data changes over time to ensure data accuracy and reliability. These data can be visualized to understand patterns. And trends in data graphs over time, it also provides insights about patterns, helps in making predictions about future data, and understands why a company’s stocks are going down and up. Examples of time series analysis include weather data, precipitation, brain monitoring, stock prices, etc. [5]
2. Regression: It is a technique that shows the relationship between a dependent variable and one or more independent variables that helps in the process of predicting the future and determining the effect on independent variables. Finance and economics by understanding the relationships between commodities or stocks. [6]
3. Clustering: It is a technique that works to group similar points together based on similar characteristics into different groups. It is a form of data analysis that aims to discover the inter-relationships within the data in order to know the individual or few data clusters that affect the relationships. Experiments are considered to choose the appropriate grouping method It is important to interpret the results in an effective and understandable way. The types of grouping are hard and soft grouping. [7]
4. Classification: It is a technique that classifies data into categories based on common features and is used in statistical modelling and textual analytics and helps in better understanding the data set and more accurate predictions. Techniques such as decision trees, random forests, etc. are used. [8]

## Data analytic tools

1. Power BI: It is a powerful data visualization tool developed by Microsoft that provides an easy visual interface to perform operations for downloading data in all its forms, processing and cleaning it, creating data visualizations of all kinds, defining relationships between tables, creating columns and measures, and performing any of the calculations, and allows business owners and users to make decisions as a result of analysis meaningful insights and generate reports on results. [9]
2. Microsoft Excel: It is the most common and used data analysis tool developed by Microsoft and is equipped with built-in pivot tables. It is a multi-capacity data management program such as data retrieval, processing, cleaning, modification, analysis and visualization. It supports different types such as numbers, dates, texts, etc., and performs mathematical operations. It contains many functions that allow Users summarize and analyse large data and generate reports with ease. [10]
3. Python: It is a modern, high-level programming language that is widely used and used in data analysis. It is fast and easy to use. It processes data and includes visualization of all different types of data, performing mathematical equations in an easy way, facilitating understanding of ideas, accelerating data processing processes, and discovering problematic data and empty data. It provides many technologies, such as prediction, contains many large libraries, and provides multiple functions. [11]
4. R: It is an open-source programming language and is widely used in data analytics, graphics and statistics and is used to identify patterns and build practical models and is used to create programs and applications that perform analyses and provides a graphical user interface and provides a set of tools and libraries to perform analyses and supports many modelling techniques such as compilation and strings temporal, etc. It is considered one of the leading analytical languages for years, and it is a difficult language to learn, but it provides powerful and more accurate analyses. [12]

## Types of data analytic methods

• Descriptive analytics: It is a branch of data analytics that summarizes current and historical data and is the basis of all data insight. It is the simplest and most common use of data in business and data interpretation to understand patterns and trends. It aims to answer the question (what happened?) i.e., the main features and patterns of the data. It provides summary statistics to provide a brief representation of the data. The data can be divided into different meaningful categories, such as median, mean, standard deviation, and range. Percentages, etc. It is the basis for further analyzes, understanding historical data, summarizing and visualizing it, and making informed decisions based on descriptive analyses. [13]

Advantages: [14]

1. Helps summarize complex and large data in meaningful ways by performing statistical calculations and creating summary reports and visualizations
2. It provides a clear and concise representation of the data, facilitates understanding of the data, and helps in making decisions
3. It helps to understand historical data to see what happened in the past through relationships and patterns to help in future planning
4. It helps to identify problems in the data such as missing values, outliers, and errors by discovering and repairing them that affect the accuracy and reliability of the data.
5. It provides visualization of data in an attractive and intuitive way to understand patterns and relationships and easily describe and communicate ideas in an easy way
6. Monitoring, evaluating and measuring performance indicators through rapid analysis and providing insights at the appropriate times, allowing for rapid response

Disadvantages: [14]

1. It does not provide deeper insights into the causes and motives behind these patterns and is not sufficient to understand complex relationships or make predictions for future outcomes.
2. It does not provide a complete understanding of the factors influencing patterns. Descriptive analyzes must be combined with other techniques
3. It relies on historical data that may not accurately reflect the current or future situation
4. Data can be subject to bias and assumptions if the data is incomplete, which affects the analysis and results
5. Descriptive analyzes do not provide any performance improvement that requires additional techniques for implementation
6. When dealing with large and complex data that requires increasing the infrastructure to accommodate the large amount of data

• Predictive analytics: is a branch of data analytics used to make predictions about future outcomes that may be challenging. It answers the question (what could happen?) It relies on the data that we have summarized and on statistical modeling and machine learning techniques to make predictions about possible future outcomes and evaluate the likelihood of occurrence and requires additional forecasting technology The accuracy of the prediction depends on the quality of the detailed data and aims at insights and expectations that help in making decisions Informed and proactive actions. [15]

Advantages: [15]

1. Predictive analytics helps uncover hidden insights by identifying possible future events and behaviors using statistical algorithms
2. Improving the decision-making process by helping institutions to make more accurate decisions in view of the possible results and patterns
3. Predictive analytics helps patterns and relationships within complex data and variables that influence outcomes
4. It enables organizations to forecast sales, needs and key factors that support accurate forecasts
5. It helps in discovering and mitigating problems before they occur, which reduces material losses
6. Detecting and preventing fraud through predictive models and identifying suspicious activities

Disadvantages: [15]

1. It depends mainly on the quality and availability of data, so that if the data is incomplete, it greatly affects the accuracy and reliability of predictions and bad decision-making.
2. Making forecasts requires experience and knowledge in predictive data analysis and statistical modeling
3. The predictive analyzes are incorrect because they are based on assumptions and simplifications between variable relationships, which leads to inaccurate insights.
4. Ethical and privacy concerns when dealing with sensitive data such as predicting the lives of individuals
5. Predictive analyzes can produce incomprehensible and complex results that are difficult to explain and require an understanding of the reasons behind their prediction.

• Prescriptive analytics: It is a branch of data analytics that combines insights from all previous analytics to determine the appropriate decision to solve a problem. This analysis uses the latest technology and AI is an example of descriptive analytics. It answers (what should I do?) depends on actions and decisions to improve results to predict what is likely to happen in the future to achieve the expected goals and results and aims to provide actionable insights and achieve better results. [16]

Advantages: [16]

1. Descriptive analyzes help in making the optimal decision and determining the best course of action in different scenarios so that it provides the most effective and efficient solutions.
2. Descriptive analyzes help strategic planning and improve resources such as the budget in the most effective way and achieve the plan's goals
3. Mitigating risks and minimizing their severity through descriptive analysis and assessing the potential impact and addressing it before it occurs
4. Prescriptive analytics aids real-time decision making to enable rapid response to changing circumstances
5. Prescriptive analytics help improve processes to improve productivity, reduce costs and streamline operations
6. Forecasting future resources for which demand will occur and managing them effectively through historical analyzes, market trends and other external factors.

Disadvantages: [16]

1. Requires experience and knowledge of complex algorithms and modeling techniques that are used to improve decision-making
2. Descriptive analyzes depend on high-quality data, and this is difficult because the data is different and usually lacks accuracy and reliability
3. Assumptions about relationships and variables may be incorrect, leading to substandard insights
4. Implementing descriptive analytics solutions is difficult because it requires integration with existing systems, decision-making processes, and external factors
5. Decision makers need to understand the limitations of descriptive models and practice critical thinking in order to apply all knowledge and experience when making a decision.
6. It requires significant costs to implement descriptive analytics solutions and also requires strong infrastructure and continuous maintenance

## Uses of data analytic methods in real life

* Descriptive analytics: [17]

**Business Performance Analysis:** Descriptive analyses are used to evaluate and analyse business performance such as revenue, sales, and profits. Through analysis, they enable them to identify company trends, patterns, and outliers, and help transform these analyses into insights that contribute to business success and important decision-making.

**Financial Analysis:** Descriptive analyses are used to analyse historical financial data such as expenses, costs, profits, performance improvement, and financial profits to find out the economic effects, make decisions to improve the budget, predict the future, and improve investment.

**Healthcare Analytics:** Descriptive analyses are used to analyse patient data through patient trends, health analysis results, continuous analysis of treatment results, knowledge of the patient's improvement rate, and help in knowing the appropriate medication for each patient based on analyses to improve patient health.

**Social Media Analysis:** Descriptive analytics is used to understand user behaviour and measure interest in all videos by analysing the percentage of views he watches to know interests. These analyses help promote ads, evaluate success rates, and know potential effects.

* Predictive analytics: [18]

**Sales and Demand Forecasting**: Predictive analyses are used to predict sales and demand for products. This is done by analysing the market, its trends, and economic indicators for each country, and what are the factors that affect the increase in demand. It helps companies make accurate predictions about future sales and improve production processes.

**Risk Assessment:** Predictive analyses are used to assess risks and determine the possibility of these risks occurring through analysing current data and problems. Based on this data, future predictions can be made of what may happen in order to secure and solve problems before they occur. They help in making decisions to transform problems and risks and their effects on the company

**Healthcare Diagnosis and Treatment:** Predictive analyses are used to improve the health status of the patient by conducting analyses and examinations, conducting an analysis of this data, anticipating the symptoms that may occur to the patient in the future, and knowing the appropriate solutions for each patient by proposing treatments to mitigate and reduce the incidence.

**Fraud Detection:** Predictive analytics is used to detect and prevent attacks and intrusions by analysing patterns and user behaviour and predicting suspicious activities in real time by taking preventive measures to reduce the incidence of intrusions and reduce financial losses

* Prescriptive analytics: [19]

**Supply Chain Optimization:** Descriptive analytics are used by forecasting demand and determining production capabilities and transportation costs to optimize supply operations through the use of optimal distribution strategies for products and improve user satisfaction

**Resource Allocation and Planning:** Descriptive analyses are used by workforce management, project planning, budgeting, and deadlines for each project in order to allocate resources and distribute the budget in a manner commensurate with the plans specified for each project.

**Energy Management:** Descriptive analyses are used in energy management to improve energy consumption and reduce costs by developing patterns and analytical models to save energy and improve usage rates to reduce energy waste.

**Fraud Prevention:** Prescriptive analytics is used to prevent fraud, identify potential suspicious activity, and prevent it by analysing patterns, user behaviour, and potential impersonation cases, providing real-time alerts, providing security recommendations, and actions to mitigate fraud and protect data

# Descriptive Analytics

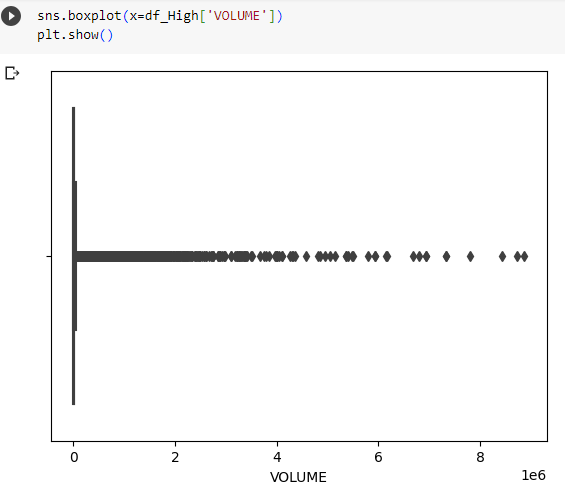
## Techniques & Examples (Your work)

### Features Analysis and Explanation

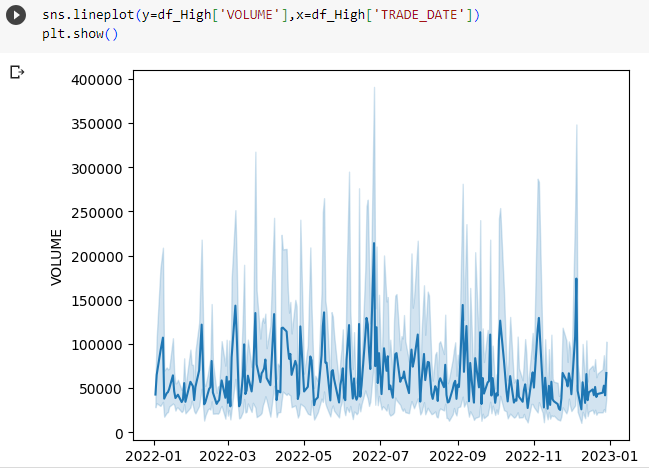
|  |  |  |  |
| --- | --- | --- | --- |
| **Feature no.** | **Feature Name** | **Descriptive Measure / Technique** | **Explanation** |
| **1**  **2**  **3** | VOLUME  TRADE\_QTY  BEST\_BID\_QTY | Mean | It was used to calculate the average of all values and divide it by the total number of values to help understand the typical or central value within the distribution |
| Median | A statistical measure that represents the median values of values in ascending or descending order. It is the value that separates the lower values from the upper values. Useful for understanding the typical value or the central trend in the distribution. |
| Stander deviation | A statistical measure that determines the amount of variance or dispersion within the data. It shows how the values are spread. The smaller standard deviation indicates the points closest to the mean. The larger one indicates greater variance and is used as a measure of the spread and dispersion of the data. |
| Describe | It is a process that describes the data in detail and helps to understand and visualize the data and shows many values such as count, mean, std, min, 25%, 50%, 75%, max |
| Minimum | It is a statistical process that shows the minimum value for a group of values and is useful in discovering the smallest value within a group of data |
| Maximum | It is a statistical process that shows the highest value for a group of values and is useful in discovering the largest value for a group of data |
| Z\_score | It is a standard score, which is a statistical measure that determines the number of standard deviations for a specific value that is far from the mean of the data set to allow a better understanding of their relative positions within the distribution and help identify outliers and analyse data when it relates to their distribution |
| Variance | It is a statistical measure that determines the spread or dispersion of a set of data and provides a numerical value that indicates how far each value is from the average of the data set. |
| Mode | It is a statistical measure that identifies the most frequent value in a data set and helps to find out which value is the most common |
| Coefficient | It is a statistical measure that refers to a numerical factor or multiplier applied to a variable and is used in various mathematical calculations, integration, statistics, etc. Coefficients provide valuable information about the relationship between variables and to make predictions and analyse data |
| Percentiles | It is a statistical measure used to measure equal parts of a set of data. It calculates percentages ranging from 0 to 100 and is useful for understanding the distribution of values within a data set. |
| Quartile | It is a statistical measure for a group of data into four parts, each containing 25% of the data. It is usually used to extract the value Q1 = 25%, Q2 = 50%, Q3 = 75%. It is useful for understanding the spread and distribution of data, as well as the interquartile range that shows the range between Q1, Q2 |
| Outlier | It is a statistical measure that shows data that deviates significantly from a set of other data, a value that is out of the pattern and usually occurs as a result of an error or rare events and affects the analysis significantly |
| **4** | Market | Frequency | It is a statistical measure that indicates the number of times a value or set of data is repeated and is used to understand the distribution and pattern of the data |
| Relative frequency | It is a statistical measure that indicates the number of times a value or data set is repeated as a percentage of the total number. It helps in making comparisons and understanding the spread of values. |
| Cumulative frequency | It is a statistical measure that represents the running total of frequencies up to a specific value or category of data. It represents the cumulative frequency in ascending order. The process is done by adding the frequency to the previous values. |
| Cumulative frequency relative | It is a statistical measure that represents the running total of the relative frequencies of a particular category of data in percentage terms by the relative frequency of the value and adds it to the relative frequencies of the value |

### Features Visualization and Explanation

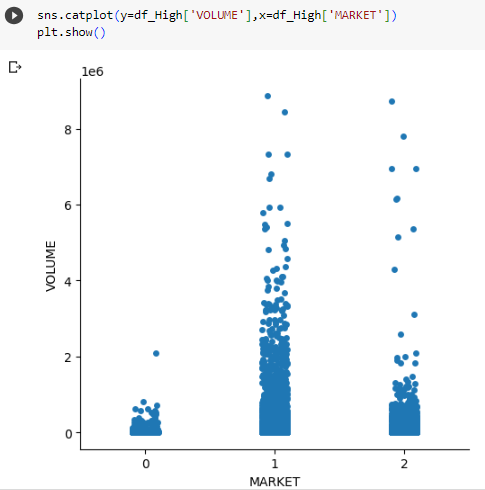
**Feature 1:** VOLUME



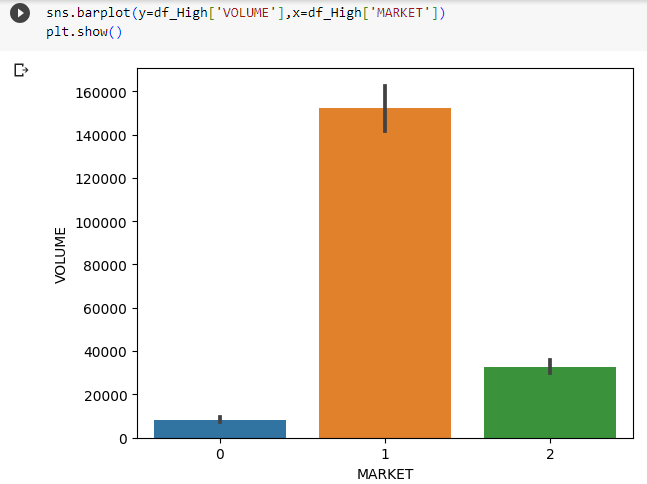
Box plot: It is a graphic representation of the VOLUME variable, which shows summary statistics such as the first quarter, the second quarter, the third quarter, the average, the outliers, the determination of the largest and the smallest value, the distribution and deviation of data, and helps in understanding the data and knowing the distribution, collection and dispersion of data to make decisions based on the results of the analysis.

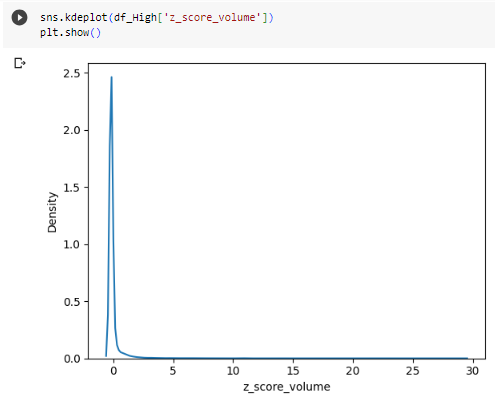


line plot: It is a graphic representation of the variable VOLUME and TRADE\_DATE, which shows the data points connected by a solid line that shows the relationship between the two variables, and here we used VOLUME and TRADE\_DATE, which shows the relationship of the variable with time and you can conclude at any time there was a greater amount and what times need to be improved.

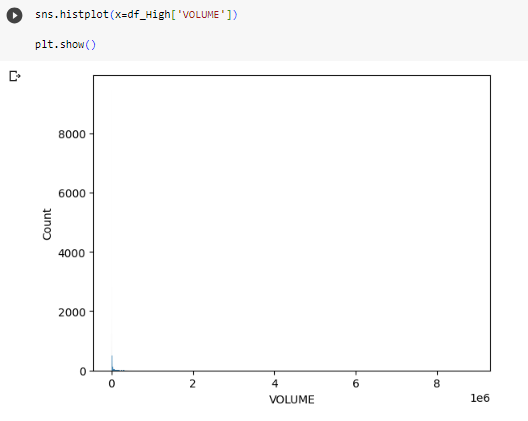


Cat plot: It is a graphic representation that displays the relationship between two variables, the VOLUME and MARKET. It shows the scattered points on the categorical axis, which shows the distribution of values in each category and shows the gathering of values and extreme values. Knowing the category that contains the largest value and knowing the category in which the most process occurs

  
Bar plot: It is a graphic representation used to display categorical data. It represents the frequencies of the number of VOLUME in each category of the MARKET. It shows the most frequent categories in the form of rectangles. It is useful for making comparisons and knowing the weak categories and trying to improve them to make informed decisions on the low categories.

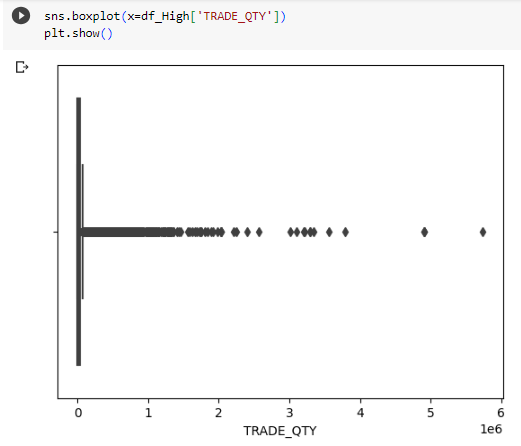


Kdeplot: It is a graphic representation that represents the basic distribution of the data. It was used for the variable that measures z\_score\_volume, which shows the number of standard deviations for the VOLUME variable far from the mean for the variable for a better understanding of the data distributions

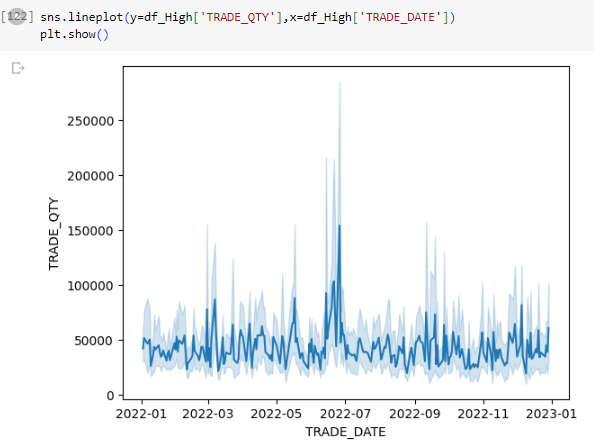


Hist plot: It is a graphic representation of the distribution of a continuous variable that displays the frequencies that fall within a period of time. It clarifies the distribution of the data, its shape, and detects spread and outliers. Each histogram shows the number of iterations of the values in the VOLUME variable.

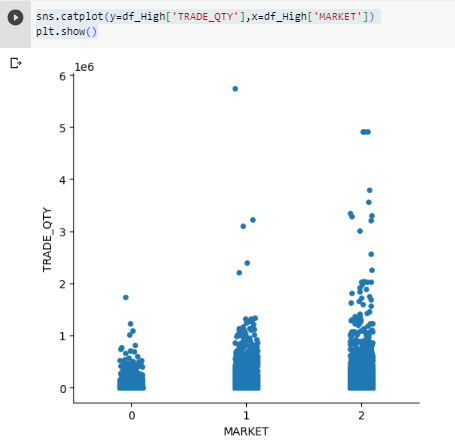
**Feature 2: TRADE\_QTY**



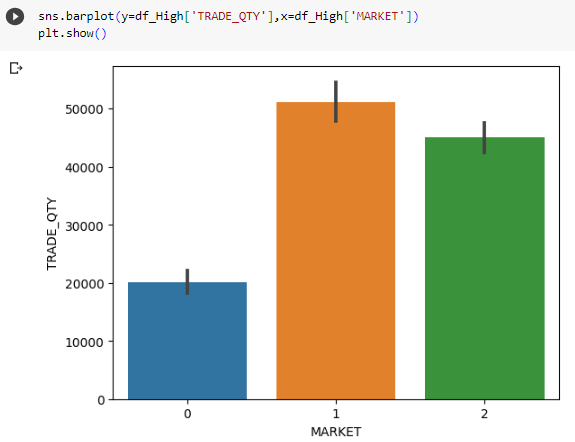
Box plot: It is a graphic representation of the **TRADE\_QTY** variable, which shows summary statistics such as the first quarter, the second quarter, the third quarter, the average, the outliers, the determination of the largest and the smallest value, the distribution and deviation of data, and helps in understanding the data and knowing the distribution, collection and dispersion of data to make decisions based on the results of the analysis.



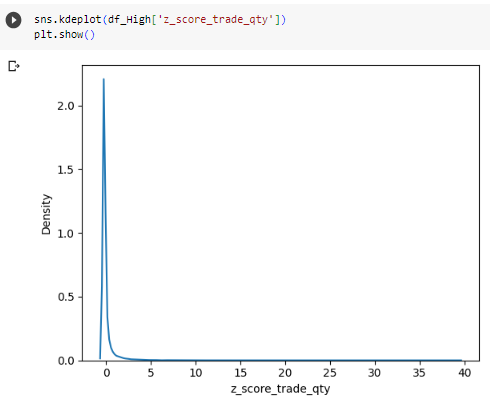
line plot: It is a graphic representation of the **TRADE\_QTY** and TRADE\_DATE, which shows the data points connected by a solid line that shows the relationship between the two variables, and here we used VOLUME and TRADE\_DATE, which shows the relationship of the variable with time and you can conclude at any time there was a greater amount and what times need to be improved.



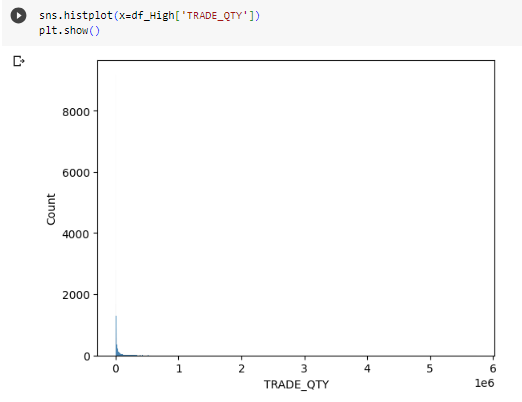
Cat plot: It is a graphic representation that displays the relationship between two variables, the **TRADE\_QTY** and MARKET. It shows the scattered points on the categorical axis, which shows the distribution of values in each category and shows the gathering of values and extreme values. Knowing the category that contains the largest value and knowing the category in which the most process occurs



Bar plot: It is a graphic representation used to display categorical data. It represents the frequencies of the number of **TRADE\_QTY** in each category of the MARKET. It shows the most frequent categories in the form of rectangles. It is useful for making comparisons and knowing the weak categories and trying to improve them to make informed decisions on the low categories.

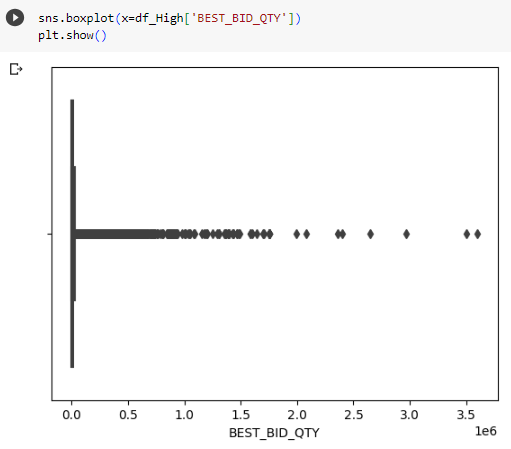


Kdeplot: It is a graphic representation that represents the basic distribution of the data. It was used for the variable that measures z\_score\_trade\_qty, which shows the number of standard deviations for the **TRADE\_QTY** variable far from the mean for the variable for a better understanding of the data distributions

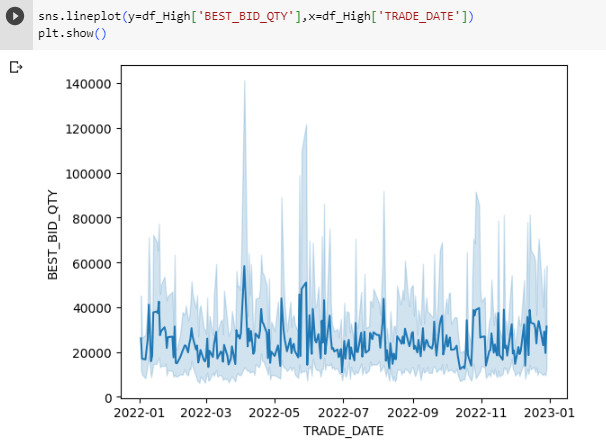


Hist plot: It is a graphic representation of the distribution of a continuous variable that displays the frequencies that fall within a period of time. It clarifies the distribution of the data, its shape, and detects spread and outliers. Each histogram shows the number of iterations of the values in the **TRADE\_QTY** variable.

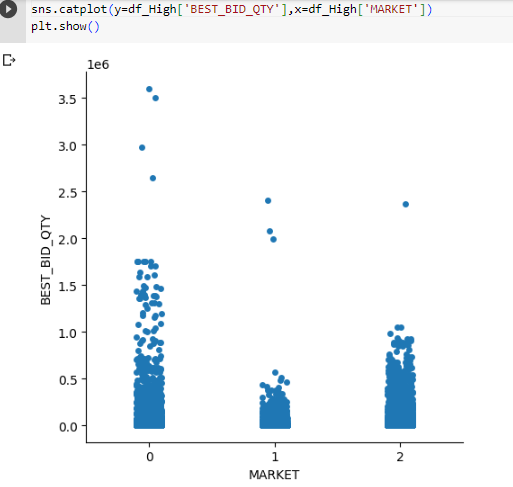
**Feature 3: Best\_BID\_QTY**

****

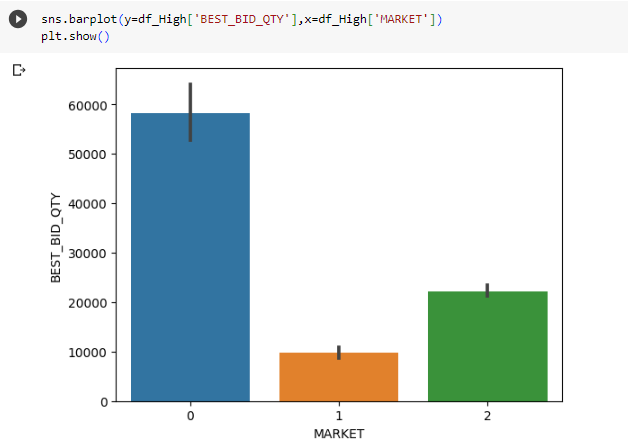
Box plot: It is a graphic representation of the **Best\_BID\_QTY** variable, which shows summary statistics such as the first quarter, the second quarter, the third quarter, the average, the outliers, the determination of the largest and the smallest value, the distribution and deviation of data, and helps in understanding the data and knowing the distribution, collection and dispersion of data to make decisions based on the results of the analysis.

****

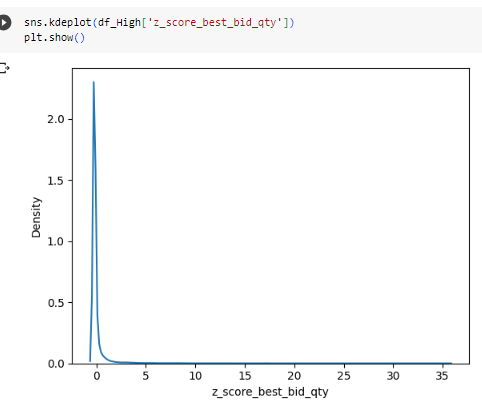
line plot: It is a graphic representation of the variable **Best\_BID\_QTY** and TRADE\_DATE, which shows the data points connected by a solid line that shows the relationship between the two variables, and here we used **Best\_BID\_QTY** and TRADE\_DATE, which shows the relationship of the variable with time and you can conclude at any time there was a greater amount and what times need to be improved.

****

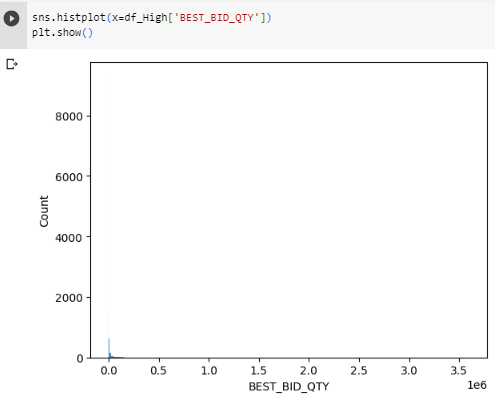
Cat plot: It is a graphic representation that displays the relationship between two variables, the **Best\_BID\_QTY** and MARKET. It shows the scattered points on the categorical axis, which shows the distribution of values in each category and shows the gathering of values and extreme values. Knowing the category that contains the largest value and knowing the category in which the most process occurs

****

Bar plot: It is a graphic representation used to display categorical data. It represents the frequencies of the number of **Best\_BID\_QTY** in each category of the MARKET. It shows the most frequent categories in the form of rectangles. It is useful for making comparisons and knowing the weak categories and trying to improve them to make informed decisions on the low categories.

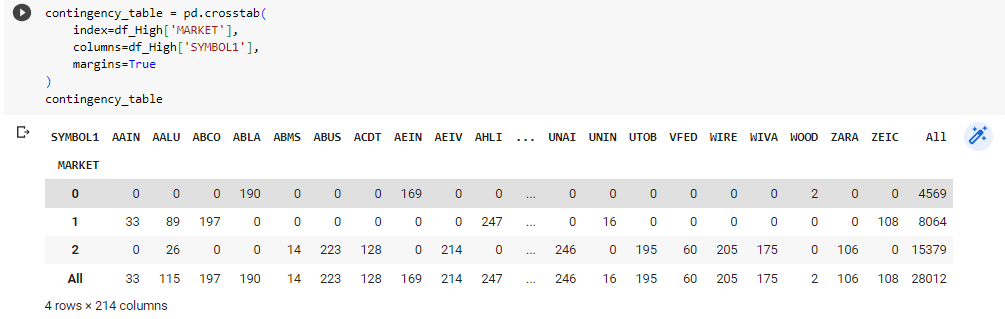
****

Kdeplot: It is a graphic representation that represents the basic distribution of the data. It was used for the variable that measures z\_score\_best\_bid\_qty, which shows the number of standard deviations for the **Best\_BID\_QTY** variable far from the mean for the variable for a better understanding of the data distributions

****

Hist plot: It is a graphic representation of the distribution of a continuous variable that displays the frequencies that fall within a period of time. It clarifies the distribution of the data, its shape, and detects spread and outliers. Each histogram shows the number of iterations of the values in the **Best\_BID\_QTY** variable.

### Contingency table and Explanation



It is a representation that summarizes the relationship between two categorical variables that illustrates each category with the number of frequencies that fall into each of the categories for both variables and helps provide insights on how to increase the frequencies in each category to improve the categories. The two categorical variables are MARKET and SYMBOL1, which shows the relationship between them and shows each bank. In a market that is the best in terms of iterations, as for the markets, it shows which market categories are best in terms of the number of iterations, and this makes us make offers on the few iterations in order to improve higher and create more iterations.

## Techniques for decision-making (Your work)

Many techniques, measurements, and graphic representations have been used in my project, and they have been greatly benefited from, so that these techniques that have been worked on build ideas and extract meaningful insights from the analysed data, so that they identify the problems that suffer from them and the changes that need to make some improvements Trying to improve the categories of markets in terms of frequencies in each category and knowing the problem behind this decline in each category of markets and what are the banks that have a few iterations in a category of markets and deducing the basic problems using these techniques Many measurements, patterns and perceptions were made that helped greatly To make effective decisions that represent an improvement to the needs of users and to find the best solutions for each company to compete in the markets by identifying strengths, weaknesses, opportunities and threats through the analysis processes that have been conducted, we conclude that there are solutions to eliminate threats, exploit opportunities, improve weaknesses and mitigate risks that can lead to decisions More informed and effective towards success.

## Evaluation (Your work)

Analysis techniques mainly help in decision-making processes as they provide valuable hidden insights that are discovered and interpreted so that decision-makers gain a deeper understanding of patterns and relationships that help organizations make decisions based on analysis and calculated value results instead of relying on expectations and assumptions that decision-makers benefit from by finding many solutions And the many options suitable for each specific category and provide reports that show each variable and its impact on other data to improve the quality of decisions made and increase the ability to predict through patterns and graphs for each variable that helps in proactive measures and effective plans to know customer behaviour and market trends, banks and stock prices if you are going to rise or not It will decrease by analysing each variable and understanding the data set. We can mitigate risks and find solutions to problems before they occur. We can also evaluate and measure performance by performance indicators. The techniques that have been used such as average, standard deviation, outliers, and the minimum and maximum value for each of the features that have been used have helped in understanding Data and finding relationships between them and also using various charts for all features to visualize the data and try to find every relationship for each variable. All this helped me to understand more data and know what actions I will take in order to improve the markets and know that feature VOLUME at any time is low and at any time It is low, so decisions will be taken to help increase the volume in the months that are low by developing plans to improve it and developing plans for continued improvement.

# Predictive Analysis

## Techniques & Examples (Your work)

### Feature Selection Techniques

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **FS no.** | **Name** | **Description** | **Results (High)** | **Results (LOW)** |
|  | Removing features with low variance method | Variance Threshold is a technique used in selecting features with low variance and removing them to improve prediction ability because it does not provide useful information for predictive modelling, improving the performance and efficiency of machine learning algorithms, and removing redundant features. [20] | MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY | MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY |
|  | SelectKBest method | It is a technique that selects the best K features from a data set based on its importance and its relationship to the target variable. The best features with the highest degrees are selected. Many experiments are conducted to find the best features by evaluating the impact on relationships and performance. It removes redundant data to increase optimal and effective decisions. [21] | MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY | MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_PRICE, BEST\_ASK\_QTY, BEST\_BID\_PRICE, BEST\_BID\_QTY |
|  | Sequential Feature Selection | It is a technique used to find an optimal set of features by adding or removing features iteratively based on a specified criterion and is done and involves the process of searching forward through empty features and adding them gradually or backward the features are present and removed gradually Time and cost constraints must be considered when applying this technology. [22] | MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_QTY, BEST\_BID\_PRICE | MARKET, VOLUME, TRADE\_QTY, NO\_OF\_TRADES, BEST\_ASK\_QTY, BEST\_BID\_PRICE |

### Regression Techniques

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | linear Regression | It is a statistical modelling technique used to find the relationship between a dependent variable and one or more independent estimators by having a linear relationship between the two variables. [23] |
|  | KNNR | K-Nearest Neighbors Regression It is a learning technique that is used for regression tasks. It predicts a continuous target variable. The target value is determined by looking at the nearest neighbors of the training data set. The target value is the nearest neighbor. It calculates the distance between data points. It does not provide insights between relationships and the target variable. The analyzer determines the size of the neighborhood to minimize the squared error. [24] |
|  | Decision Tree Regression | It is a learning technique that uses the knowledge of classification and regression, and it is a common algorithm because of the ability to deal with categorical and numerical data. It determines the best feature based on some criteria such as the mean square error and is used to make predictions and evaluate its accuracy by training it on the data and then evaluating the accuracy of the model. [25] |

## Compare Techniques (Your work)

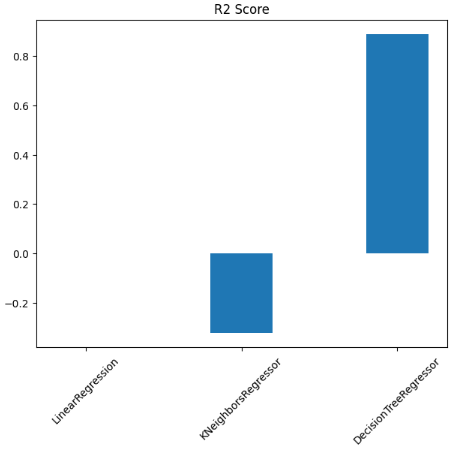
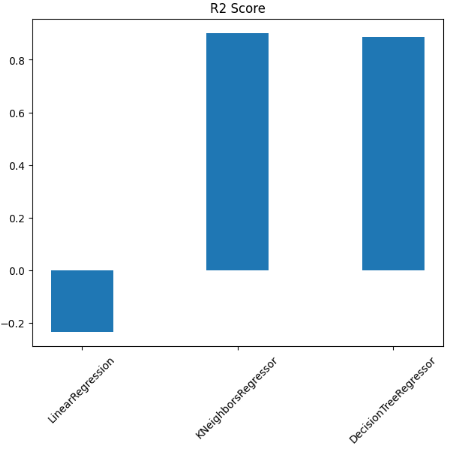
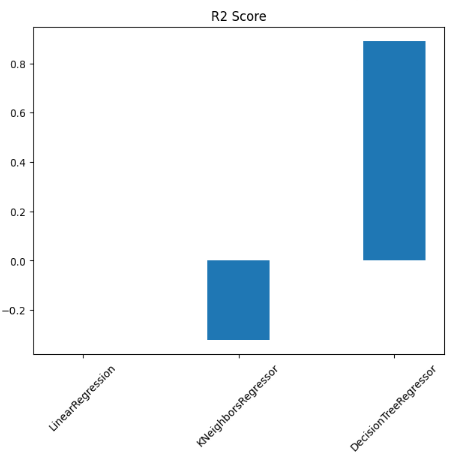
**“Low” Prediction**

**Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **1** | **1** | 1.934 | 26.029 | 5.099 | -0.233 |
| **1** | **2** | 0.297 | 2.111 | 1.450 | 0.900 |
| **1** | **3** | 0.327 | 2.361 | 1.528 | 0.888 |
| **2** | **1** | 1.677 | 21.113 | 4.592 | 0.001 |
| **2** | **2** | 2.028 | 27.906 | 5.266 | -0.320 |
| **2** | **3** | 0.326 | 2.357 | 1.527 | 0.889 |
| **3** | **1** | 1.676 | 21.111 | 4.592 | 0.001 |
| **3** | **2** | 2.027 | 27.933 | 5.268 | -0.322 |
| **3** | **3** | 0.327 | 2.367 | 1.529 | 0.888 |

**Visualization of results:**

Removing features with low variance method SelectKBest method Sequential Feature Selection

**** ****

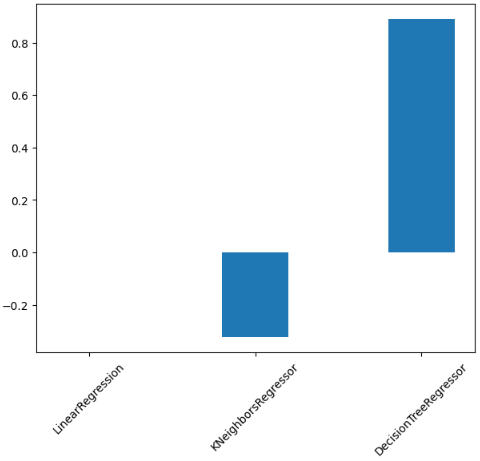
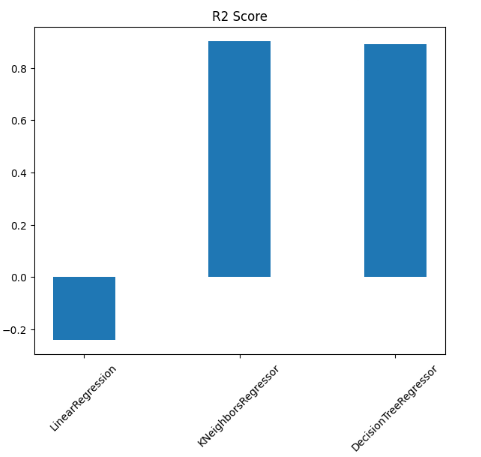
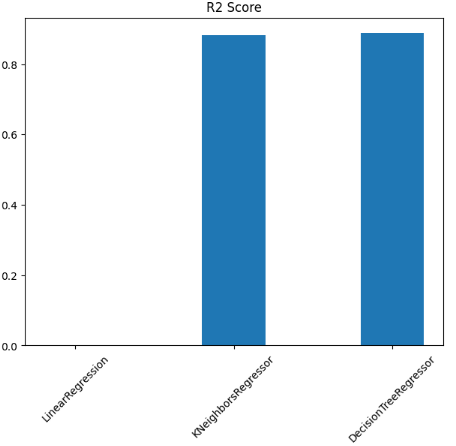
**“High” Prediction**

**Comparison:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **FS no.** | **Tech no.** | **MAE** | **MSE** | **RMSE** | **R2** |
| **1** | **1** | 1.971 | 27.161 | 5.209 | -0.239 |
| **1** | **2** | 0.301 | 2.172 | 1.470 | 0.901 |
| **1** | **3** | 0.329 | 2.419 | 1.546 | 0.890 |
| **2** | **1** | 1.706 | 21.915 | 4.679 | 0.001 |
| **2** | **2** | 0.338 | 2.598 | 1.608 | 0.882 |
| **2** | **3** | 0.330 | 2.446 | 1.555 | 0.888 |
| **3** | **1** | 1.706 | 21.914 | 4.679 | 0.001 |
| **3** | **2** | 2.058 | 28.931 | 5.361 | -0.319 |
| **3** | **3** | 0.329 | 2.441 | 1.552 | 0.889 |

**Visualization of results:**

Removing features with low variance method SelectKBest method Sequential Feature Selection

**** ****

## Evaluation (Your work)

Feature Selection Techniques were used, including Removing features with low variance method, SelectKBest method, and Sequential Feature Selection, and each of them was used on Regression Techniques, including linear Regression, KNNR, and Decision Tree Regression. Based on the results that appeared in the high data set, the best R2 score for Removing features with low variance method, especially in the KNNR, whose value was higher than the rest and equal to 0.901. As for the SelectKBest method, especially in the Decision Tree Regression, whose value was higher than the rest and equal to 0.888. As for the Sequential Feature Selection, especially in the Decision Tree Regression, its value was higher than the rest and equal to 0.889.

As for the LOW data set, and based on the results that appeared, the best R2 score is for Removing features with low variance method, especially in the KNNR whose value was higher than the rest and equal to 0.900. As for the SelectKBest method, especially in the Decision Tree Regression, whose value was higher than the rest and equal to 0.889 As for the Sequential Feature Selection, especially in the Decision Tree Regression, whose value was higher than the rest, equal to 0.888.

These techniques provide predictive analytics useful in predicting future results. Feature Selection Techniques, including Removing features with low variance method, SelectKBest method, and Sequential Feature Selection, helped to find the best predictive methods by reducing the percentage of MAE and increasing the R2. All of this helps in finding the best features for Performing a better and more accurate forecasting process and reaching near-perfect solutions. Many Regression Techniques that we used were used to clarify and show the methods used by each of them and help in making informed decisions and understanding more about relationships and patterns. All these features help organizations to progress and develop and increase customer satisfaction.

# Prescriptive Analysis

## Techniques with Examples (General)

|  |  |  |
| --- | --- | --- |
| **Tech. no.** | **Name** | **Description** |
|  | Optimization [26] | It is the technique of optimizing algorithms to find the best solution among a set of options to find the best solution to problems. The optimal solution can apply many methods to improve usage such as linear programming, integer programming, or nonlinear programming, solve problems of resource allocation, inventory management, etc. |
|  | Simulation [26] | It is a technique for creating models that simulate a system or process in the real world to understand behaviour and allow decision makers to assess the impact of changes before implementing them and are studied in order to obtain the greatest amount of detail. Simulated models such as supply chains, transportation networks and health care operations can be used to improve complex systems by analysing the results The different scenarios enable informed decisions to be made and to determine the most effective strategies. |
|  | Game Theory [27] | It is a technique for analysing and improving decisions in situations that involve conflicting interests. It is concerned with the interactions and strategies between different players to make decisions based on optimal results, such as the Nash equilibrium, cooperative games, Bayesian games, and the allocation of competitive price strategies. It focuses on understanding the behaviour of players in decision-making processes and developing strategies. |
|  | Decision-Trees [28] | It is a technique used for prescriptive analysis through a series of decisions and possible outcomes in a tree-like structure. Each node represents a decision and takes decisions such as cost, risk, and expected values to determine paths, etc. to reach the best decision and the best path. |
|  | Decision-Variable [29] | It is a technique used to make decisions based on variables, which are factors that affect the desired results. Variables are changed and modified to improve decision-making and affect goals such as a change in price to improve revenues and profits. Change in supply operations. A change in mitigation plans and strategies due to the discovery of risks that need to amend plans. |
|  | Meta Heuristic [30] | It is a technique used as an algorithm to improve near-ideal solutions to complex problems and to explore solutions efficiently and uses high-level strategies as an example of these algorithms: Genetic Algorithms (GA) and Particle Swarm Optimization (PSO) This technology provides flexibility and robustness to solve complex problems and helps provide valuable insights to provide optimal solutions. |

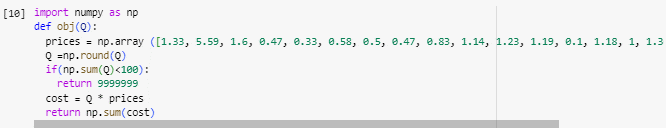
## Techniques for finding the best course of action (General)

Descriptive analysis techniques enable organizations to make informed decisions and achieve desired results. These technologies leverage advanced analytics, mathematical models, and optimization algorithms to process decision-making and support strategic planning in complex environments. Applying these techniques requires studying data quality, model assumptions, and computational requirements, so there are many algorithms that will help find the best course of action, including:

1. Genetic Algorithm (GA): It is an algorithm inspired by the process of natural evolution from the principles of genetics. It maintains a set of potential solutions to generate new solutions that take a lifetime to solve and is used to improve the allocation of multiple tasks based on their skills and project constraints. The algorithm develops potential tasks to generate solutions. New and choose the best task based on criteria There are many common applications in the field of resource allocation and project management. [31]
2. Particle Swarm optimization (PSO): It is an algorithm that relies on collective behaviour, which is a group of particles moving across the solution space and used for continuous improvement of problems such as improving the location of the facility. This is done by determining the optimal locations to open new facilities to serve customers based on factors such as costs the ability of the particles represents the potential sites of the facility and gives the optimal solution based on their converging movements. [32]
3. Gray Wolf Optimizer (GWO): It is an algorithm inspired by the hunting behaviour of Gray wolves that hunt large prey through groups in nature aiming to solve problems and improve through simulated hierarchy where each wolf represents a potential solution and then the wolves search for the optimal solution The thousand wolf represents the perfect solution Wolves update their sites in search of new sites that provide better solutions but meet the specified completion condition Examples of uses Engineering design and function optimization It is a simple and effective algorithm. [33]
4. Bat Algorithm (BAT): It is a medical-inspired algorithm based on the behaviour of bats by echolocation. It solves optimization problems. It simulates the characteristics of bats and adjusts their conditions based on environmental observations. Each bat represents a potential solution, then each bat is tuned by emission frequency and echolocation, then They search for current best solutions Bats make a random walk to explore new areas until the termination condition is met Examples of uses are data collection and parameter estimation. [34]
5. Firefly Algorithm (FFA): It is an algorithm inspired by the flashing behaviour of fireflies that aims to solve optimization problems. This algorithm uses flashing fireflies to attract friends to direct each other towards the optimal solution, where each fly represents a possible solution. The intensity of light represents the optimal solution. The more it increases, the closer it gets to the best solution. Examples of uses are image registration and feature selection. [35]

The use of any of these algorithms will help in finding the best course of action that helps in finding solutions to problems and improving them to reach a solution similar to the optimum.

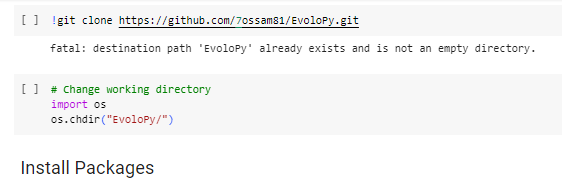
## Objective Function Code (Your Work)

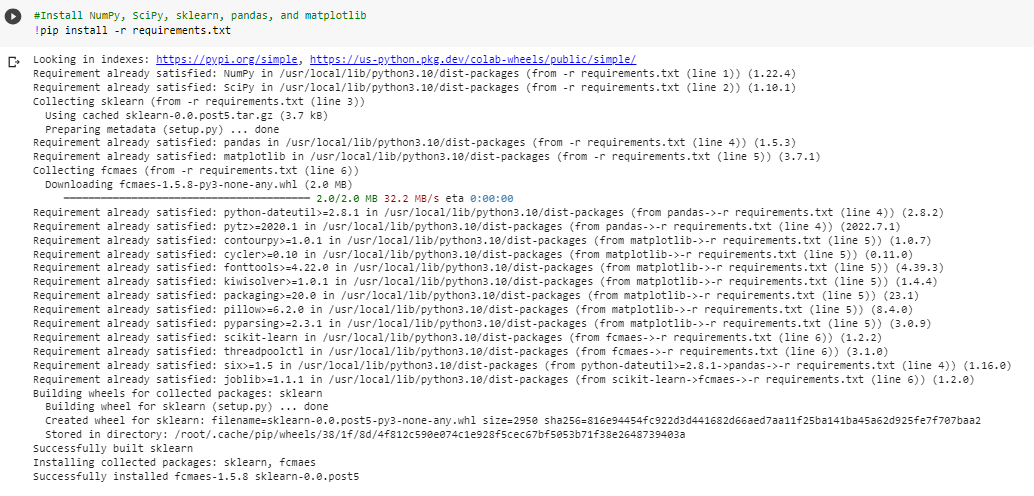


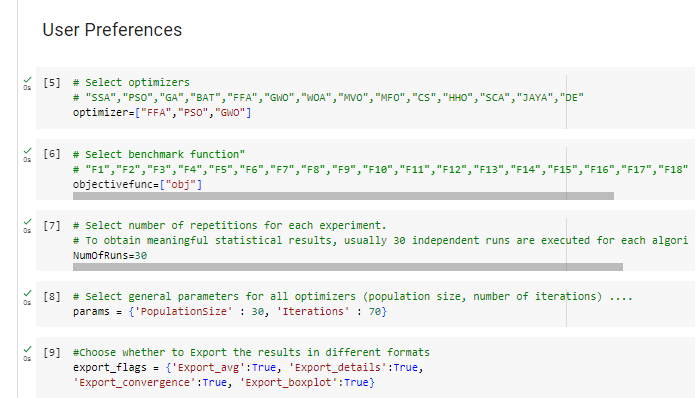
“obj”: [“obj”, 5, 10, 18],

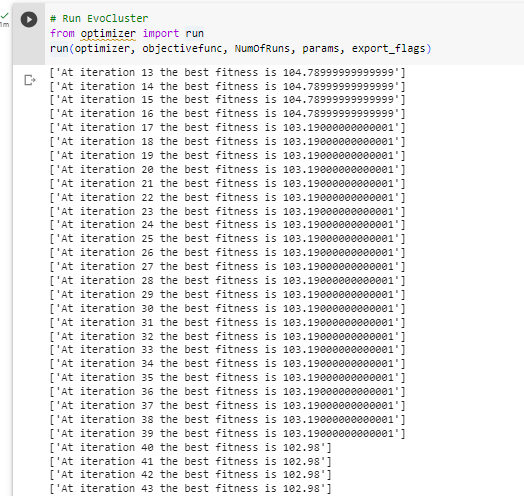
## Apply the techniques (Your Work)

### Code screenshots

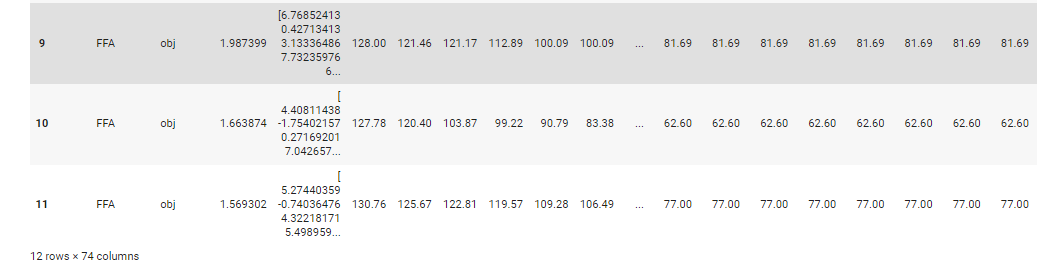
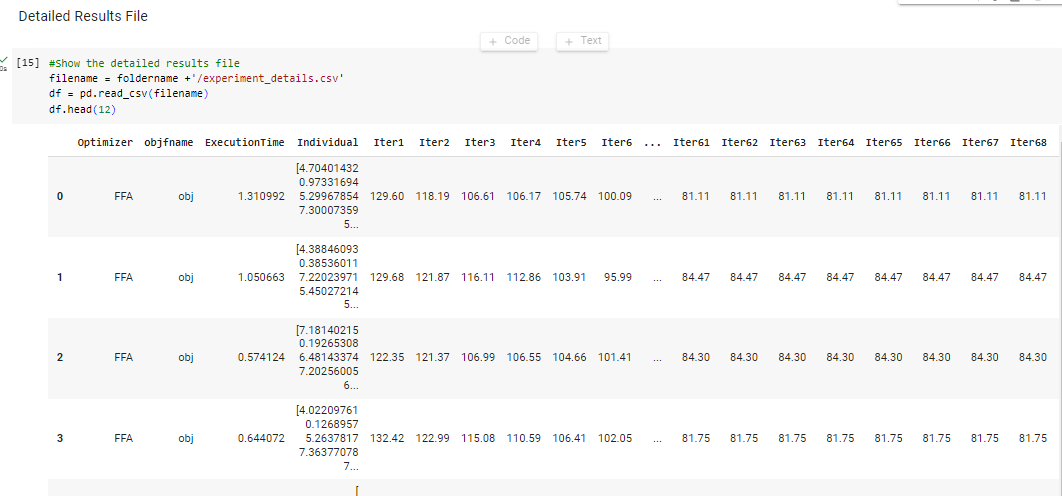
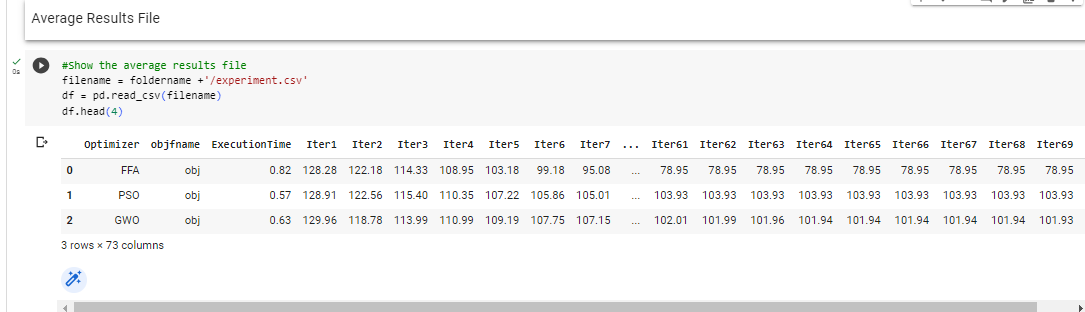
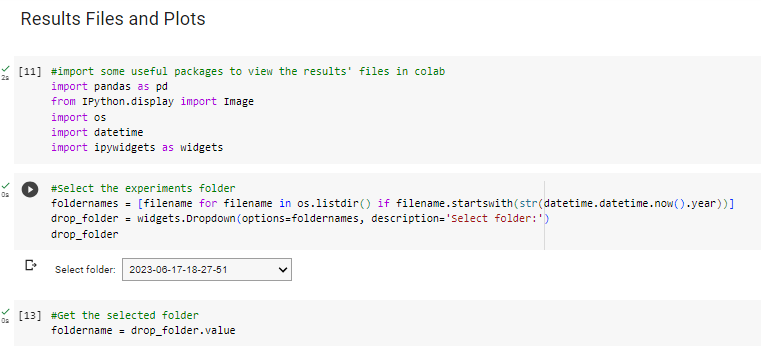






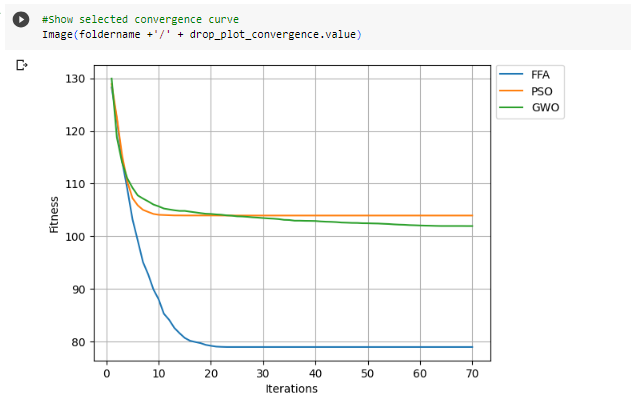
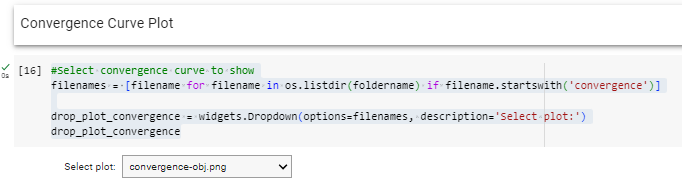


### Results and Explanation

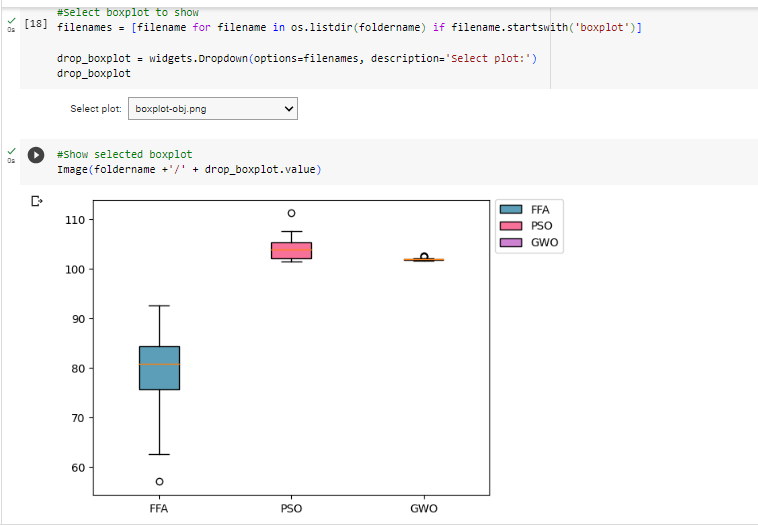


Based on the results, the algorithms that were chosen are FFA and PSO and GWO and based on iter1 the values were close and after reaching iter69 it was found that FFA is the best algorithm because the value was 78.95 and GWO is the second-best algorithm with a value of 101.93 and PSO is the last best algorithm with a value of 103.93.

### Visualization and Explanation



The graphic representation used for the chosen algorithms, namely FFA, PSO and GWO, shows this graph showing the iterations and fitness. Accordingly, at the beginning the iteration values were close, and after 10 iterations, it was found that the FFA algorithm is the best algorithm that shows the lowest value.



The graph shows the box plot for all the three algorithms that were chosen, and the graph shows what is the average, Q1, Q2, Q3, and the minimum and maximum value for each algorithm. We conclude from the graph that the FFA algorithm was the most change rate between the iteration and there were differences between the initial value and the final so it was the best algorithm.

# References

1. <https://www.simplilearn.com/what-is-data-collection-article>
2. <https://www.tableau.com/learn/articles/what-is-data-cleaning>
3. <https://www.questionpro.com/blog/data-interpretation/#:~:text=Data%20interpretation%20is%20the%20process,data%20to%20answer%20critical%20questions>
4. <https://www.tableau.com/learn/articles/data-visualization>
5. <https://www.tableau.com/learn/articles/time-series-analysis#definition>
6. <https://www.investopedia.com/terms/r/regression.asp#:~:text=A%20regression%20is%20a%20statistical,more%20of%20the%20explanatory%20variables>
7. <https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/>
8. <https://emeritus.org/in/learn/data-science-classification-analysis/>
9. <https://learn.microsoft.com/en-us/power-bi/fundamentals/power-bi-overview#:~:text=Power%20BI%20is%20a%20collection,on%2Dpremises%20hybrid%20data%20warehouses>.
10. <https://www.simplilearn.com/tutorials/excel-tutorial/data-analysis-excel#:~:text=Microsoft%20Excel%20is%20one%20of,analyze%2C%20and%20visualize%20your%20data>.
11. <https://www.futurelearn.com/info/courses/introduction-to-data-analytics-with-python/0/steps/262730#:~:text=Python%20has%20been%20around%20since,analysis%2C%20modelling%2C%20and%20visualisation>.
12. <https://www.tibco.com/reference-center/what-is-r-analytics#:~:text=R%20analytics%20is%20data%20analytics,patterns%20and%20build%20practical%20models>.
13. <https://online.hbs.edu/blog/post/descriptive-analytics>
14. <https://www.linkedin.com/advice/1/what-benefits-limitations-descriptive-analytics>
15. <https://www.ibm.com/topics/predictive-analytics#:~:text=Predictive%20analytics%20is%20a%20branch,to%20identify%20risks%20and%20opportunities>.
16. <https://www.talend.com/resources/what-is-prescriptive-analytics/>
17. <https://www.knowledgehut.com/blog/data-science/descriptive-analytics>
18. <https://marutitech.com/predictive-analytics-use-cases/>
19. <https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwjchqqftMr_AhWU0HcKHV8uBjYYABAAGgJlZg&ohost=www.google.com&cid=CAESauD2vESwhL7tRmZlLcsSs4Ecz3G3ffNELEG3PawuoJao0y_AJCLLZkkJPu4zweX-6jhVnjZZvl-LP0LlvpYIy8SDKuqdDhxe_MvU3jmapIYQ71qMqq056C6vfIPCrrNR9Jy2qKvHR0RhGXU&sig=AOD64_2USzQ6XXEq1nit8oULIhnLn_-UHA&q&adurl&ved=2ahUKEwiSoaGftMr_AhX3i_0HHUwbCgAQ0Qx6BAgHEAE>
20. <https://scikit-learn.org/stable/modules/feature_selection.html#:~:text=1.-,Removing%20features%20with%20low%20variance,same%20value%20in%20all%20samples>
21. <https://www.kaggle.com/code/jepsds/feature-selection-using-selectkbest>
22. <https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.SequentialFeatureSelector.html#:~:text=This%20Sequential%20Feature%20Selector%20adds,validation%20score%20of%20an%20estimator>
23. <https://www.ibm.com/topics/linear-regression>
24. <https://bookdown.org/tpinto_home/Regression-and-Classification/k-nearest-neighbours-regression.html#:~:text=KNN%20regression%20is%20a%20non,observations%20in%20the%20same%20neighbourhood>
25. <https://www.ibm.com/topics/decision-trees>
26. <https://baobabsoluciones.es/en/blog/2020/11/19/prescriptive-analytics-optimisation-and-simulation/>
27. <https://www.projectpro.io/article/is-game-theory-important-for-data-scientists/139>
28. <https://www.palisade.com/analytics-pyramid/#:~:text=Decision%20trees%20are%20another%20example%20of%20a%20prescriptive%20process>
29. <https://www.fivetran.com/learn/prescriptive-analytics>
30. <https://stackoverflow.com/questions/10445700/what-is-the-difference-between-heuristics-and-metaheuristics#:~:text=Metaheuristics%20are%20problem%2Dindependent%20techniques,treat%20functions%20as%20black%20boxes>.
31. <https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_introduction.htm#:~:text=Genetic%20Algorithm%20(GA)%20is%20a,take%20a%20lifetime%20to%20solve>.
32. <https://machinelearningmastery.com/a-gentle-introduction-to-particle-swarm-optimization/>
33. <https://www.baeldung.com/cs/grey-wolf-optimization>
34. <https://www.hindawi.com/journals/mpe/2020/9495281/#:~:text=The%20bat%20algorithm%20(BA)%20is,its%20low%20global%20exploration%20ability>
35. <https://www.hindawi.com/journals/jam/2012/467631/#:~:text=Firefly%20algorithm%20is%20one%20of,performance%20on%20the%20objective%20function>