# Data Analysis Wallpapers - Wallpaper Cave

Section number: 3

Student number: 21110325

Student name: Sara Aljamal

Instructor Name: Bassam Alkasabeh

**Data analytics report**

Contents

[1](#_Toc138014849)

[1 Introduction 3](#_Toc138014850)

[1.1 Data analytic activities 3](#_Toc138014851)

[1.2 Data analytic techniques 3](#_Toc138014852)

[1.3 Data analytic tools 4](#_Toc138014853)

[1.4 Types of data analytic methods 4](#_Toc138014854)

[1.5 Uses of data analytic methods in real life 5](#_Toc138014855)

[2 Descriptive Analysis 5](#_Toc138014856)

[2.1 Techniques & Examples (Your work) 5](#_Toc138014857)

[2.1.1 Features Analysis and Explanation 5](#_Toc138014858)

[2.1.2 Features Visualization and Explanation 6](#_Toc138014859)

[2.1.3 Contingency table and Explanation 9](#_Toc138014860)

[2.2 Techniques for decision-making (Your work) 9](#_Toc138014861)

[2.3 Evaluation (Your work) 10](#_Toc138014862)

[3 Predictive Analysis 11](#_Toc138014863)

[3.1 Techniques & Examples (Your work) 11](#_Toc138014864)

[3.1.1 Feature Selection Techniques 11](#_Toc138014865)

[3.1.2 Regression Techniques 12](#_Toc138014866)

[3.2 Compare Techniques (Your work) 12](#_Toc138014867)

[3.3 Evaluation (Your work) 16](#_Toc138014868)

[4 Prescriptive Analysis 17](#_Toc138014869)

[4.1 Techniques with Examples (General) 17](#_Toc138014870)

[4.2 Techniques for finding the best course of action (General) 18](#_Toc138014871)

[4.3 Objective Function Code (Your Work) 19](#_Toc138014872)

[4.4 Apply the techniques (Your Work) 19](#_Toc138014873)

[4.4.1 Code screenshots 19](#_Toc138014874)

[4.4.2 Results and Explanation 20](#_Toc138014875)

[4.4.3 Visualization and Explanation 21](#_Toc138014876)

[5 References 22](#_Toc138014877)

# Introduction

## Data analytic activities

To gain insights and achieve the aim of data analytics, we have some activities to do:

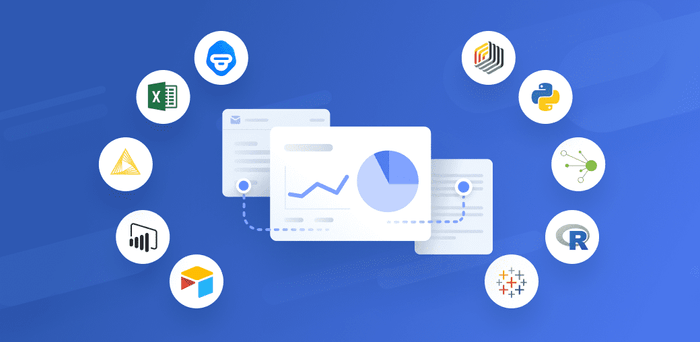
1. **Defining the problem:** In this activity, we set the problem statement and our objective from the analysis we will do. It might be a problem related to any business or industry.
2. **Collecting data:** After determining the goal, we need the data we want to analyse, if we already have this data in an organized way then this activity is done, otherwise, we must collect and organize the data.
3. **Cleaning and preparing data:** Here, we clean the data and prepare it to be analysed, for example, we handle the null values, duplicated observations, check for outliers, check if the data is balanced or not, and much more.
4. **Analysing data:** In this activity, we analyse the data to achieve our goal and solve our problem, analysing data includes multiple methods such as descriptive, predictive, and perspective that will be explained later.
5. **Visualization and results:** In this step, we evaluate the results we get from the analysis, document, and report our findings, and share them with the decision makers in order to make better decisions based on a deep analysis and understanding for the data.

## Data analytic techniques

Data analytic techniques are some processes that we apply to our data to find insights, patterns or trends, and reach conclusions that affects the decision-making process. Such as:

1. **EDA** (**Exploratory data analysis):** Analysing and understanding data using different measures and visualizations to reach conclusions, in addition to find patterns and relationships. In the EDA, it’s important to choose the right statistical measure, for example, it’s not efficient to visualize a continuous variable using a count plot, so choosing the right measure leads to more readable and explainable results.
2. **Regression analysis:** To understand and analyse the relationship between independent variables and the dependent variable, we can use a regression analysis to see the correlation between them.
3. **Time series analysis:** It’s a statistical technique that we can use to discover patterns or cycles over a period of time, it shows how a variable change throughout a period of time. Using this technique, we can predict future changes or make decisions depending on it, for example, using the time series forecasting technique with a dataset related to ice cream shops in Jordan, we can conclude the demand on ice cream increases on summer, according to this conclusion, we may decide to increase prices, or add new flavors to increase sales in summer.

## Data analytic tools

Data analytic tools refer to the software, programs, platforms, and libraries that help in performing data analytic activities, such as, Python, R, excel, and tableau.

1. **Python:** It’s an open-source programming language that is widely used in data science in data analytics, it contains thousands of free libraries that can be used for data analytics such as NumPy and pandas, as well as matplotlib which is used for visualization.
2. **Microsoft excel:** It’s a spreadsheet software that contains a lot of f functions that can be used in data manipulation and visualization Excel is an important data analytic tool, but it also has limitations such as it runs very slowly with large datasets.
3. **Tableau:** It’s a commercial data visualization tool that doesn’t require advanced coding expertise, can handle large datasets, and has a drag and drop interface. But it can’t be used for manipulating data or preprocessing.

## Types of data analytic methods

Depending on the goal of the data analysis, we can choose one of these methods:

1. **Descriptive analytics:** Analysing and understanding historical data to describe and summarize “what happed in the past and why”, by analysing data, finding patterns and discovering hidden insights. In descriptive statistics mainly use statistical measures and plots.
2. **Predictive analytics:** In predictive analytics we use the results found in descriptive analytics to predict the future “what is likely to happen”, so we use past trends to predict future trend using predictive models such as ML models or neural networks.
3. **Prescriptive analytics:** This type focuses on “what action to take”, in prescriptive analytics, we provide recommendations to decision makers, in order to minimize a problem or maximize an opportunity.

## Uses of data analytic methods in real life

**Case study:** XYZ is a furniture store, they are experiencing a drop in revenues and seeking to increase profits. To deal with this problem, they decided to use analytics.

1. **Descriptive analytics:** They started applying descriptive statistics to examine and analyse their past sales, by analysing the trends, customers behaviour, best seller products, underperforming products, seasonality for some products.
2. **Predictive analytics:** They used past sales data to predict future sales. Using predictive models such as time series forecasting or ML models, they can predict sales based on many factors such as price changes or offers and seasonality. So, they can predict whether their sales will increase, decrease, or remain the same.
3. **Prescriptive analytics:** here, XYZ store uses prescriptive analytics todecide on the best action, prescriptive analytics recommend actions and best decisions for each product to maximize the revenue and the sales. for example, they apply prescriptive analytics to determine the best offer price for an underperforming product to increase the demand and maximize the profit.

# Descriptive Analysis

## Techniques & Examples (Your work)

### Features Analysis and Explanation

|  |  |  |  |
| --- | --- | --- | --- |
| Feature no. | Feature Name | Descriptive Measure / Technique | Explanation |
| 1 | Trade date | To date time function | After transforming the trade date column into date time type instead of string, I made a new column called ‘Month’ which includes only the month instead of the whole date and dropped the trade date column. Because working with the month will be much easier and we can discover more insights in the data. The reason for choosing the month instead year or day is that this data is for 1 year which is 2022, so it’s impossible to take the year. |
| 2 | Sec code, Symbol | Frequency (Groupby(), mean) | Sec code and symbol refer to each company’s ID and symbol, in order to compare companies, find out which company has more trades, I used groupby function, to group the sec code and the symbol according to the mean of the trade quantity (TRADE\_QTY), so I calculated the trade quantity average for each company and then sorted them. VFED company with the sec code 131011 was the company with the highest trading quantity average (547930), while ALFA with the sec code 131083 was the company with the least trading quantity. |
| 3 | Best Ask Price | Min, max, and quartiles | The minimum ask price =1, while the max ask price= 40, Q1 = 1.05, Q2= 1.18, and the Q3= 1.55. These values indicate a lot of things:  1- Most of the values are below 1.55  2- 50% of the value are less than 1.18  3- We can detect outliers easily by looking at the difference between Q3 and the max ask price, which is very high, but the difference between the min ask price and Q1 is normal so we can say that most of the outliers bigger than the normal range and not smaller. |
| Mean | The mean =1.62 which is bigger than Q3, this indicated we have a lot of outliers in this column because the mean is affected by these values so that’s why it’s bigger than Q3. |
| 4 | Best Bid QTY | Standard deviation | The standard deviation= 100818, which is a very big number, this indicates:  1- That the data is spread out, so the values aren’t close to the mean, this also indicates that the data has a wide range. So, the number of individuals willing to buy differs and it hard to predict the number from the mean.  2- A big standard deviation also means we have outliers or extreme values in our data. |
| Q1, Q2, and Q3 | The range between Q1 and Q3 is very big, this is also considered as an indicator for having a wide range.  In addition, the difference between Q1 and Q2 is less than the difference Q1=400, Q2=1912, Q3=10286  between Q2 and Q3 in a clear way, this also tells us that 50% of the total bid quantities are below 1912, and also indicates a wide range for the data. | |

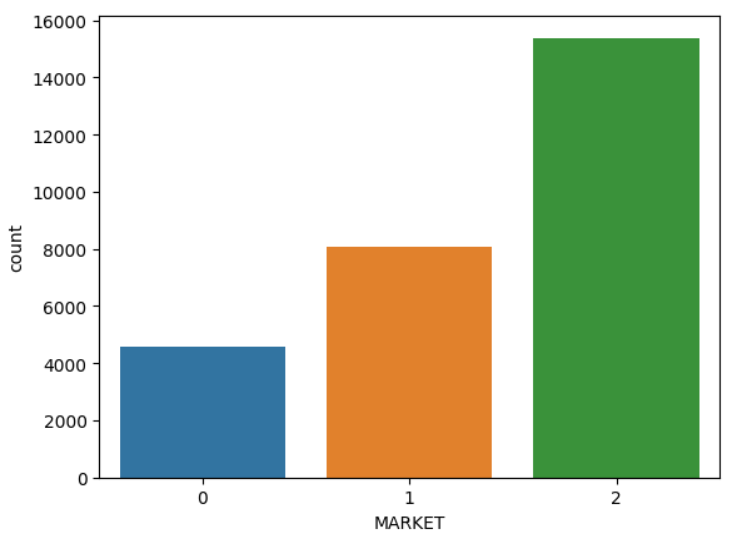
### Features Visualization and Explanation

**Month:**

In order to determine whether there is a month where trades rise, I used a count plot to visualize the ‘month' column. Most months had 2000–2500 trading sessions, while June had about 2760, making it the month with the most trading sessions overall. We can also say that in July, there are more than 2500 sessions. This helps us if we want to determine when the trading market has its busiest periods.

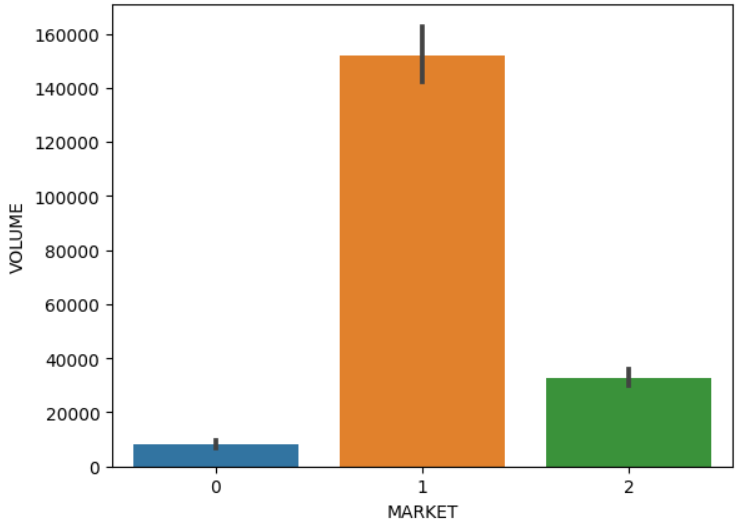
**A picture containing text, colorfulness, screenshot, plot

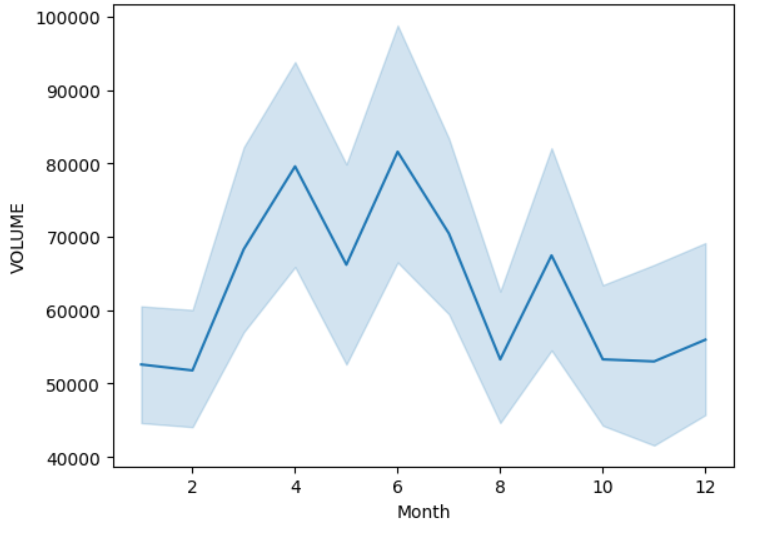
Description automatically generated**

**MARKET:**

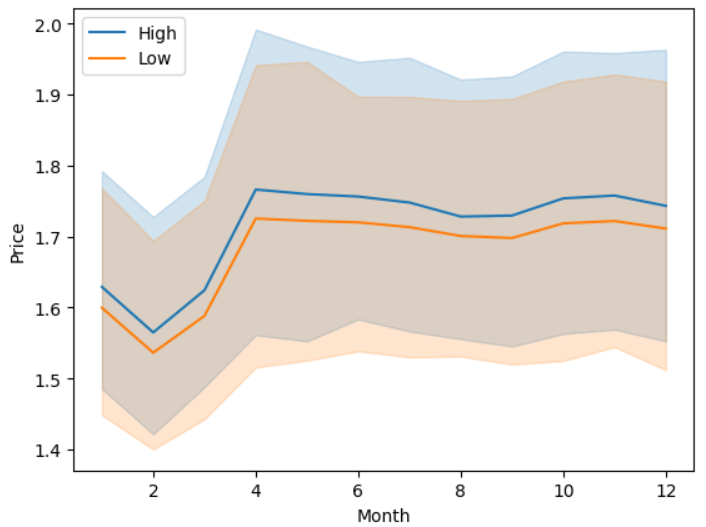
In our dataset, we have 3 trading markets, I made a count plot to see where most trading sessions occurs, we have 28000 rows in the dataset which means 28000 trading session, around 16000 from these sessions happened in market 2, while 8000 of them happened in market 1, and around 4500 sessions happened in market 0. We can benefit from this information to tell investors where they should invest and trade, so it's clear that market 2 has more trading session for those who care about the number of trading sessions in each market.

But for those who care for the volume of trades in every trading session, they should invest in market 1. From the plot that shows the volume’s mean in each market we notice that the trading sessions that occurs in market 1 has more volume than those in market 2 and 0, even though the number of sessions is higher in market 2.

****

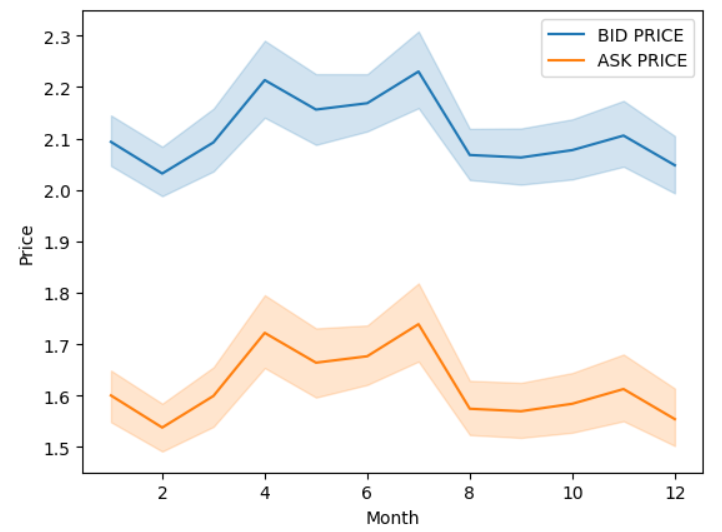
**VOLUME:**

Since the volume is a continuous feature, a line plot can be utilized to show how it has changed over time. We now have a line plot that demonstrates how the volume changed during the entire year using the month column. The plot shows that the two months where the trade volume is at its peak are April and June. We can also observe the months where the volume increases, like March, and where it falls, like July, we can also notice the fact that the year begins and ends with quiet trading with low volumes.

**HIGH and LOW:**

In the data frame I worked on, I added both high and low columns and not only one of them (Only in descriptive statistics). The high column represents the highest price of a stock, while the low column represents the lowest price of a stock. These variables are also continuous, so I wanted to look at how they changed over time as well as how they differed from one another and if there was a certain time of the year when that difference increased or decreased. Therefore, this line plot displays two lines that symbolize their change over time. I realized that changes occur on both high and low at the same time (they rise and fall together), for example. the lowest price for both was in February, while the months of April through September may be described as having stable pricing with little change over time.

**BEST BID PRICE and BEST ASK PRICE:**

****

The best bid price is the highest price buyers will pay to buy a stock or a security, while the best ask price (also called the offer price), is the lowest price sellers are willing to sell their security or stock. To visualize these columns a used a line plot, I also included the two lines in one plot to compare their changing. It’s clear that the blue line represents the best bid price because its values are bigger that the orange line’s values which is the best ask price. It’s noticeable that they change together over time. From the plot, we can recognize that the least prices were in February, where the maximum prices for both were in July.

### Contingency table and Explanation

A picture containing text, font, screenshot, number

Description automatically generated

The columns in the table represents the months, where the rows in the table represent the markets. The table shows the frequency for each combination, in other words, it shows how many trade sessions happened every month for each market. Margins in the table calculate the sum of sessions for each row or column, for example, the last column in the first row (4569), is the sum of all sessions that happened in market 0 during the whole year. The aim of this table is to observe the months where the trading activity in the markets gets higher or lower, in addition to analysing the distribution of the trading sessions over the year. March was the month where the highest number of sessions occurred in market 0, while June was the month where the highest number of sessions occurred in market 1, and market 2. From the table, we can also reach the conclusion during the year, most trading sessions occurred in market 2.

## Techniques for decision-making (Your work)

Descriptive statistics are performed to describe and discover what happened in the past, as well as benefiting form past data to make good decisions in the future, understanding how the trading market changes and discovering opportunities will help traders and investors a lot in maximizing their profit from this field. From the figures and statistics explained above, we can reach many conclusions that helps traders in decision making.

Knowing in which period of the year the market is more active assist investors and traders in determining when to concentrate more on this market and when they should be more active, from the figure that shows how trading sessions are distributed among the months (month count plot), we can conclude that summer can be considered the season for trading, specially June as it’s the month with the largest number of trading sessions. Another evidence that may convince traders is the volume of trades in each session (from the volume line plot over the year), we can reach the conclusion that June is the month with the biggest number of shares in each session, as well as April.

In our data, there were 3 markets where traders can invest and trade in, traders can trade in all markets, but knowing which is the best market for them and their type of trading help them a lot instead of trying all markets, when looking at the distribution of trading session among the 3 markets (market count plot), and the mean volume for trades in every market (market count plot with volume), we can figure out that market 2 is the market where most trading session happens, but market 1 is the market where the transactions volume is much bigger. So, if traders are looking for more session, they should invest in market 2, but if the trading quantity is what they are looking for, then they should concentrate their investments in market 1.

Monitoring and understanding how high and low prices change over time (high and low line plot) can be very beneficial for traders, as they will know when to sell to get the maximum profit from knowing when the high price reaches its peak, which is in April, so they increase the price for their stocks. And they should take advantage of February as it’s the month where the “low” reaches its lowest levels, so buying more stocks in February will minimize the cost.

Being aware of the change in best bid and best ask prices (best ask price and best price line plot) is very important, exploiting the opportunity and increasing the ask price in July, will increase the profit. In addition, knowing when to increase the bid price and when to reduce it will make traders experts and help them in taking decisions.

Lastly, more understanding for the trading market means better decisions, maximized profit, and reduced cost. More trading data, for example more than one year will aid in making better decisions because decision will be built based on more than one year.

## Evaluation (Your work)

In my work in descriptive statistics, I made sure to understand and analyse features in an appropriate way to get readable and better visualizations, so I did the right visualization for each column based on its data type and other factors, I found a lot of insight that helps in decision making as mentioned in the question above. I realized that analysing features using appropriate measures leads us to better understandings and conclusions. I found it much better when I deal with months instead of dates after extracting the month feature from the trade date column, without doing this, I won’t be able to visualize the volume, high and low, in addition to the best bid and ask prices in the way I did. Understanding the distribution of the data, and that it’s not normally distributed helped me in determining which measure to use in detecting outliers. Combining “high” and “low” columns in one dataset allowed me to visualize both and compare how they changed over time. Overall, I think the descriptive statistics I made on the provided dataset was great and led us to many conclusions that benefit traders and investors, maybe performing more statistics on column such as best bid quantity and trade quantity will lead me to more results and conclusions.

# Predictive Analysis

## Techniques & Examples (Your work)

### Feature Selection Techniques

|  |  |  |  |
| --- | --- | --- | --- |
| FS no. | Name | Description | Results (Selected Features) |
|  | Select K best | It’s a feature selection technique that selects most relevant K features from the whole dataset, this technique works by ranking all feature according to their relationship with the target variable, the ranking process depends on the parameter that we select, for example chi squared, ANOVA, or f\_ regression. We select the parameter in addition to K (the number of features to select) depending on our problem and the aim of the model. After the features are ranked, top K features are selected and only these features will be in the input for our model. | Because K best feature selection doesn’t rely on the ML model, all features selected in before all models will be the same.  Selected features:  ['SEC\_CODE', 'MARKET', 'VOLUME', 'NO\_OF\_TRADES', 'BEST\_ASK\_PRICE', 'BEST\_ASK\_QTY', 'BEST\_BID\_PRICE'] |
|  | Sequential feature selection | Sequential feature selection evaluates different combinations of features in the selection process, these combinations are evaluated by their importance in the model’s performance, unlike the K best technique, in sequential feature selection we specify the ML model because the feature selection process depends on the model’s performance. Another parameter is the number of features where it runs until the we reach the given number of features. Sequential feature selection has two strategies, either forward selection (starting from empty set and adding features), or backward elimination (starting from all features and removing unwanted features) | Selected features in linear regression:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY', 'BEST\_ASK\_PRICE', 'BEST\_ASK\_QTY']  Selected features in decision tree:  ['SEC\_CODE', 'SYMBOL1', 'MARKET', 'NO\_OF\_TRADES', 'BEST\_ASK\_PRICE', 'BEST\_BID\_PRICE', 'Month']  Selected features in KNN:  ['SYMBOL1', 'MARKET', 'VOLUME', 'TRADE\_QTY', 'BEST\_ASK\_PRICE',  'BEST\_ASK\_QTY', 'BEST\_BID\_QTY'] |

### Regression Techniques

|  |  |  |
| --- | --- | --- |
| Tech. no. | Name | Description |
|  | Linear regression | Linear regression is a supervised regression model that assumes a linear relationship between the independent variable (feature) and the dependent variable (target), this technique aims to find the best fit line for the relationship between variables, finding this best line is determined by finding the minimum squared difference between the predicted and the actual value, which is the MSE (Mean Squared Error). |
|  | Decision tree regression | Decision tree regression technique builds a tree where we start from the root node, keep going through multiple decision nodes, until we arrive to the leaf node which is the decision or the predicted value. So, we split the data according to multiple conditions, for example, if this condition is true, move to the left node, if it’s false then move to the right node. After multiple splitting iterations, every leaf node will contain several data points, so when we want to predict a new value, we take the mean of all values in that leaf node. |
|  | KNN regression | Unlike other models, KNN model doesn’t require any training time, instead, the running time for this technique is the time where it calculates the distance between the new data point and all other data points, after that, depending on the K value (number of neighbours) which is a parameter we specify, the model takes the mean value for these neighbours and predict the value for the new data point. |

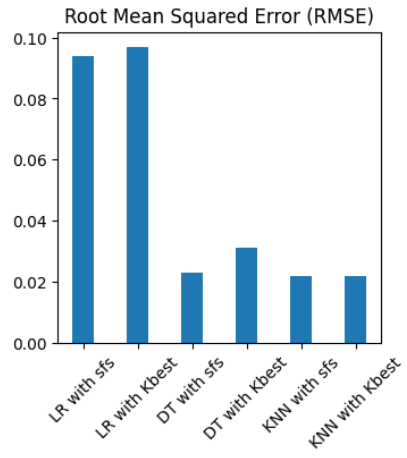
## Compare Techniques (Your work)

**“Low” Prediction**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| FS no. | Tech no. | MAE | MSE | RMSE | R2 |
| Sequential feature selection | Linear regression | 0.036 | 0.009 | 0.094 | 0.18 |
| Sequential feature selection | Decision tree regression | 0.004 | 0.001 | 0.023 | 0.95 |
| Sequential feature selection | KNN regression | 0.004 | 0.0 | 0.022 | 0.956 |
| Select K best | Linear regression | 0.037 | 0.009 | 0.097 | 0.124 |
| Select K best | Decision tree regression | 0.006 | 0.001 | 0.031 | 0.909 |
| Select K best | KNN regression | 0.004 | 0.0 | 0.022 | 0.955 |

**Comparison:**

**Visualization of results:**

* R2 score and RMSE average values for 30 iterations for every model:

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

Description automatically generated**

* Box plot that visualizes the 30 R2 values for decision tree model with sequential feature selection:

A picture containing screenshot, text, rectangle, diagram

Description automatically generated

* Box plot that visualizes the 30 RMSE values for the KNN with sequential feature selection:

A picture containing text, screenshot, line, diagram

Description automatically generated

**“High” Prediction**

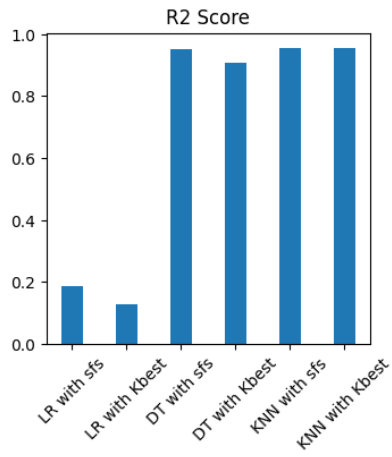
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| FS no. | Tech no. | MAE | MSE | RMSE | R2 |
| Sequential feature selection | Linear regression | 0.036 | 0.009 | 0.095 | 0.185 |
| Sequential feature selection | Decision tree regression | 0.004 | 0.001 | 0.023 | 0.95 |
| Sequential feature selection | KNN regression | 0.004 | 0.0 | 0.022 | 0.956 |
| Select K best | Linear regression | 0.038 | 0.01 | 0.098 | 0.128 |
| Select K best | Decision tree regression | 0.006 | 0.001 | 0.032 | 0.908 |
| Select K best | KNN regression | 0.004 | 0.0 | 0.022 | 0.955 |

**Comparison:**

**Visualization of results:**

* R2 score and RMSE average values for 30 iterations for every model:

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

Description automatically generated**

* Box plot that visualizes the 30 R2 values for decision tree model with sequential feature selection:

A picture containing screenshot, text, line, diagram

Description automatically generated

* Box plot that visualizes the 30 RMSE values for the KNN with sequential feature selection:

A picture containing screenshot, text, line, diagram

Description automatically generated

## Evaluation (Your work)

To predict the “low” column, which represents the lowest price for a stoke, and the “high” column, which represents the highest price for a stoke, I used 3 techniques, linear regression, decision tree regression, and KNN regression, with every model I used two feature selection techniques, select K best technique, and the sequential feature selection technique.

First of all, from the results in the tables above, we can notice that the models performed the same on the “high” and on the “low” columns, and this is very expected and normal, due to fact that these two features are affected by the same features, from the figure above that demonstrates how both of them changed over time, it was clear that they rise and fall together.

Secondly, when comparing the models with each other, it’s clear that linear regression was the worst of all, from the low table, R2 measure for linear regression was very low regardless the feature selection technique, it was 0.18 using sequential feature selection, and 0.124 using K best, another way to notice that is from the bar plots that shows the distribution for the R2 measure, where linear regression has the shortest bars, and RMSE measure where linear regression had the longest bars. The reason behind these values is because linear regression assumes linear relationship between variables, while they may not be linear. In addition to the reason that linear regression gets affected a lot by outliers in the data, and out data contains lots of outliers. Decision tree regression and KNN regression performed better than linear regression, the difference between them is very low, but we can say that both of them are less sensitive to outliers than linear regression and they are more appropriate to be used with our dataset.

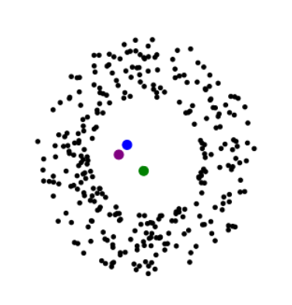
Finally, the results for both feature selection techniques with all models were very close, bar plots, are easier to see the difference between the two techniques than the tables, sequential feature selection was better than K best will all models, as it gave us higher R2 and less RMSE.

# Prescriptive Analysis

|  |  |  |
| --- | --- | --- |
| Tech. no. | Name | Description |
|  | GWO (Grey Wolf Optimization) | It’s an optimization algorithm inspired by the hierarchy within the grey wolves’ family and their unique hunting techniques. This technique is built on the hierarchy in the wolves’ family, starting with **alpha** at the top of them, which represented the best fitted and optimal solution found in this pack, followed by **beta** the second optimal solution, moving to **delta** which refers to the third optimal solution, and the remaining wolves in the pack are described as **omega**, all other solutions. |
|  | MVO (Multi verse optimization) | This optimization algorithm is based on the idea of multi verse, which means that we have more than one universe. Each universe in the MVO represented a solution, where particles try to find the best solution. MVO riles on three things, white holes, black holes, and wormholes, white holes represent the exploration phase, where particles explore new universes in the search space, black holes represent exploitation phase, where particles are attracted toward better solutions based on their fitness which is calculated through the objective function, finally, particles communicate through wormholes to exchange information and lead each other. |
|  | SCA (Sin Cosine Algorithm) | It’s a new population-based metaheuristic algorithm that is inspired by the periodic behaviour of the sin and cosine functions, in this algorithm, search agents’ positions are updated according to a trigonometric function which tends to find the optimal solution and position for the search agents in the search space. The sin function allows the search agents to explore different solutions and regions in the search space, by that, new positions are generated to be visited and explored by the search agents, so we can say that sin function is used for exploration. On the other hand, cosine function, which is used for exploitation, guides search agents toward better and more optimal solutions in the search space, so their values are updated until reaching the optimal or near optimal solution. |

## Techniques with Examples (General)

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

**GWO:** In grey wolf optimization algorithm, Alpha, the guide, or the leader, aims to find the optimal solution, while Beta helps Alpha in making decisions and keeping the pack disciplined, this makes Beta the second optimal solution if Alpha wasn’t there, Delta assist Beta and provides updates to Alpha. Finally, Omega, the remaining wolves in the pack, in our algorithm, Omegas are all other possible solutions, they are guided by Alpha, Beta, and Delta in the search space to get closer to the top 3 optimal solutions which is the positions of Alpha, Beta, and Delta. In the figure, Alpha, is the green point which represents the optimal solution found, Beta and Delta, the blue and purple points are very close to each other, and all other black points are considered Omegas, the possible solutions. At the end, we can see that all Omegas points are located around the optimal solution. Increasing The number of iterations may lead to better results, because with more iterations, all other possible solutions get closer to the optimal or near optimal one.

**MVO:** The MVO algorithm starts by generating a random set of possible solutions, where each solution is considered a universe, then particles move around these universes in the exploration (white holes) , and exploitation (black holes), through the wormholes (the communication channel between universes), each universe is evaluated by calculating the fitness of the particles using the objective function, through the iterations, universes are updated, and after all iterations, the universe with the best fitness is considered the optimal solution for the optimization problem. MVO is affected by the number of iterations as well as the population and the complexity of the problem.

**SCA:** The algorithm based on trigonometric function, where we explore the search space using the sin function, and we exploit the solutions to get the optimal solution using the cosine function. Similar to MVO, SCI starts by generating a set of possible solutions, that are then explored and exploited using sin, cosine, as well the objective function to test the fitness of possible solutions. As other optimization algorithms, SCA gets affected by the population, the problem complexity, number of iterations, and other factors.

## Objective Function Code (Your Work)

def cost(Q):

  prices =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])

  Q=numpy.round(Q)

  if numpy.sum(Q)<10:

    return 9999999

  else:

    cost= numpy.sum(Q\*prices)

    return cost

## Apply the techniques (Your Work)

### Code screenshots

* **Changes on the benchmarks file:**

**A picture containing text, font, screenshot, algebra

Description automatically generated**

****

* **Changes on the code:**

1- Selecting optimizers:

2- Selecting the objective function from the benchmarks file:

3- Changing the number of runs:

### Results and Explanation

I tried 3 different techniques, GWO, MVO, SCA. GWO and SCA results were very close to each other and finally both reached the same least cost, which was 2.15, while MVO’s least cost was 3.19.

A picture containing text, screenshot, font, number

Description automatically generated

GWO best individual:

[0, 0, 0, 0, 5, 0, 0, 0,0

0, 0, 0, 5, 0, 0, 0, 0, 0]

Using the table above, the GWO suggests buying 5 stocks from ARBK bank on Wednesday, in addition to 5 stocks from AHLI bank on Saturday, with total stocks =10.

SCA best individual:

[0, 0, 0, 0, 5, 0, 0, 0,0

0, 0, 0, 5, 0, 0, 0, 0, 0]

Using the table above, SCA suggests buying 5 stocks from ARBK bank on Wednesday, in addition to 5 stocks from AHLI bank on Saturday, which is the same result as GWO, so we can conclude that both reached the exact same result.

Analysing the result: A stock from ARBK bank on Wednesday costs 0.33, while a stock from AHLI bank on Saturday costs 0.1, (0.33\* 5) + (0.1 \*5) = 2.15

MVO best individual:

[0, 0, 0, 3, 1, 1, 0, 1, 0,

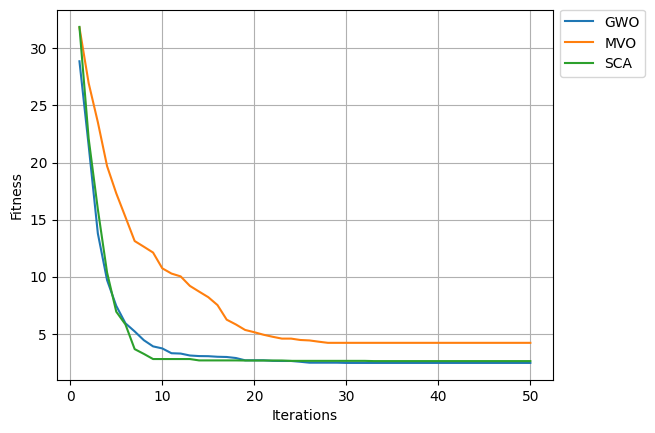
0, 0, 0, 4, 0, 0, 0, 0, 0]

MVO suggested buying 5 stocks from ARBK bank, 3 on Tuesday, 1 on Wednesday, and 1 on Thursday. In addition to 1 stock from HBTF on Sunday. And 4 stokes from AHLI bank on Saturday.

Analysing the result: (3 \* 0.47) + (1\* 0.33) + (1\*0.58) + (1 \* 0.47) + (4 \* 0.1) = 3.19

### Visualization and Explanation

To visualize, compare, and analyse the performance of the 3 optimization techniques:

**1- Line plot:**

From the plot above, we can clearly notice that the GWO, and SCA are very close to each other from the beginning of the iterations, which means they have the same performance, but on the 9th iteration SCA arrived to the least cost, which was 2.15, and GWO arrived later on the 18th iteration. Although at the end both techniques showed the same exact results, but SCA was faster than GWO, so we can say that SCA performs better than GWA, because if the number of iterations was smaller, SCA may show better results than GWO, for example, if we had 15 iterations only, SCA least cost will be 2.15, while GWO least cost will be 2.46. MVO technique was the worst because the least cost after 50 iterations was 3.19.

**A picture containing text, screenshot, diagram

Description automatically generated2- Box plot:**

From the 3 box plots, we can see the maximum and minimum cost for each technique, we can also notice the range for each one of them, MVO for example, had the longest range between all of them, a long range mean this technique took the longest time to arrive to the least cost for it or to be stable. Boxplots are used to show how every technique performed in the 50 iterations, in addition to comparing all technique with each other’s.

# References

Stevens, E. *et al.* (2023) *The 7 most useful data analysis techniques [2023 guide]*, *CareerFoundry*. Available at: https://careerfoundry.com/en/blog/data-analytics/data-analysis-techniques/#time-series (Accessed: 16 June 2023).

Hillier, W. *et al.* (2023) *The 9 best data analytics tools for data analysts in 2023*, *CareerFoundry*. Available at: https://careerfoundry.com/en/blog/data-analytics/data-analytics-tools/ (Accessed: 16 June 2023).

Peixeiro, M. (2022) *A practical guide to feature selection using Sklearn*, *Medium*. Available at: https://towardsdatascience.com/a-practical-guide-to-feature-selection-using-sklearn-b3efa176bd96 (Accessed: 12 June 2023).

*Sklearn.feature\_selection.Selectkbest* (no date) *scikit*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SelectKBest.html (Accessed: 12 June 2023).

*Sklearn.feature\_selection.Sequentialfeatureselector* (no date) *scikit*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.feature\_selection.SequentialFeatureSelector.html (Accessed: 12 June 2023).

Mali, K. (2023) *Everything you need to know about linear regression!*, *Analytics Vidhya*. Available at: https://www.analyticsvidhya.com/blog/2021/10/everything-you-need-to-know-about-linear-regression/ (Accessed: 12 June 2023).

*Decision tree regression clearly explained!* (2021) *YouTube*. Available at: https://www.youtube.com/watch?v=UhY5vPfQIrA (Accessed: 12 June 2023).

M, D. (2022) *All you need to know about grey wolf optimization*, *Analytics India Magazine*. Available at: https://analyticsindiamag.com/a-complete-overview-of-grey-wolf-optimization-technique/#:~:text=The%20grey%20wolf%20optimization%20tries,converge%20to%20a%20bad%20solution. (Accessed: 15 June 2023).

Zvornicanin, W. by: E. (2022) *Grey Wolf Optimization algorithm*, *Baeldung on Computer Science*. Available at: https://www.baeldung.com/cs/grey-wolf-optimization (Accessed: 16 June 2023).

Bouguezzi, W. by: S. (2023) *Multi-verse optimizer*, *Baeldung on Computer Science*. Available at: https://www.baeldung.com/cs/multi-verse-optimizer#:~:text=MVO%20is%20a%20population%2Dbased,numerical%20optimization%20issues%20in%202015. (Accessed: 15 June 2023).

Bansal, J.C. *et al.* (1970) *Sine cosine algorithm*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-981-19-9722-8\_2#:~:text=SCA%20is%20a%20population%2Dbased,the%20sine%20and%20cosine%20functions. (Accessed: 16 June 2023).