Data Analytics.

Section (3)

Assignment Title:

Technical Report.

Assignment Term:

Final Term.

Submitted By:

Osama Zamari.

Submitted To:

Spring 2023

Table Of Contents

Data:

Data is information that has a specific meaning that monitor something, it can be found everywhere because everything in life can be extracted and consider as data that can be benefited from to analyze it and make decisions to solve a problem based on the data we have. To be more specific, any activity that happens and it is recorded such as what customers buy in a supermarket. In addition. images and videos or voice records etc...

Data Analytics Activities:

It is the science of analyzing data that we have and visualizing it for better vision of what the data is containing by drawing charts that can be understood easily and give statistical measures, and according to it, we can make decisions that help us to solve a specific problem or to increase the success or to escape a potential risk. But data analytics job does not just about visualizing the data, the analyzer needs to examine raw data and be able to communicate with stakeholders of the organization to show them what his analyzing means and what to do based on them. We gain a lot advantages from data analytics such as building up business decisions and avoiding risks that might affect the organization. In addition, increasing on the revenue of the business and avoids as many losses as possible.

The science of data analytics can be everywhere where the data is produced and can be found in all industries even for the entertainment industry such as sports, data analytics is used to measure the performance of players and check on team's performance which might lead to the understanding of why the team lost or what is the reason of bad performance.

On the other hand, data analytics power is related more in businesses especially for organizations that run big businesses so they cannot neglect the satisfaction of customers to not lose profits. For example, data analytics is very valuable for clothing markets to understand how well the organization is performing, what months the profit is the most and the reasons of that and what cause setbacks of the business, in that way, the profits will increase and the risks will be avoided.

Data Analytics Techniques:

There are many techniques that the science of data analytics uses such as:

Exploratory Data Analysis: It is the science of analyzing data by discovering it by statistical measurement to understand the nature of it, or we can represent it in charts so we could understand the relationship between the data. It is a very effective technique to make the analyzer understand how to deal with the data. In addition it is good for communication with stakeholders by showing them what the data hides and how we can improve our business. For example, charts such as bar plots, line plots, pie charts and histogram and distribution plots.

Statistical Measures: It is very important for the analysis to understand what the data measures such as:

> Mean: We can understand what is the mean of the data which helps us to know what the average is of this data. For example in a market shop, if the mean is 100, it means that

you can buy multiple things with 100$ only because it gives the average which will allow customers to take an initial understanding of prices in the market.

Mode: we can understand what is the mode of the data which helps us to know what the most repeated value of this data is. For example, if the mode between cats and dogs in pet store is cats, we can understand that 1. The store has more cats than dogs, 2. The customers buy cats more than dogs.

Percentiles: We can understand what the data value in a specific percentile of the data is. For example, if the percentile (65%) of employees salary prices is equal to 6000$, that means 65% of employees has less or equal than 6000$.

Regression Analysis: It is a technique that helps us to understand the relationship between independent and dependent variables, which helps us to understand what the correlation between two variables is by knowing how much this variable affect the another.

Time series analysis: it is the science of analyzing data that is collected in determined period of time to see how the data values are changing over time to make predictions or forecasts values in the future. It helps organizations to understand the trends and why did it occur by using visualization to understand the data changing over time. For example, whether data and stock prices.

Data Analytics Tools:

Tools are helpful equipment for data analyzers which help them prepare the data and explore it and discover the hidden patterns of the data. Here are five of tools that are widely used:

Tool	Definition	Benefits	Limitation
Microsoft Excel	It is a spreadsheets software from Microsoft, it has many functions for calculations and graphing which are good for analyzing the data.	Good at calculations. Good at understanding the data and handling it due to its organization.	Prone to human errors. Cannot handle big data. Not designed for collaborative work.
Tableau	It is a software that does not require programming skills, it is simple that can be used by everyone for visualizing the data and analyzing it.	No programming skills needed. Simple yet strong for visualizations.	No pre-processing for the data.

Python	It is a programming language that is huge with libraries which make analyzing the data wider, as well as with machine learning and data manipulation.	Full procedure from data manipulation to machine learning to visualizing. Big community and python developers. Simple and close to English language. Rich Libraries	Heavy memory usage. Slow speed.
Power BI	It is Microsoft software which has strong features for visualizing the data and is easy for the users by making dashboards and reports.	Collaborating and sharing. Adding Ai systems.	Any mistake in connecting tables will affect the results
R	A programming language used for statistical data analysis and visualization.	Rich Libraries. Can handle heavy computing tasks.	Syntaxes are harder than python.

Data Analytics Methods:

There are many methods for the science of data analytics, I will cover the three methods which are:

Descriptive Analytics:

Definition: It is the science of using old historical recorded data to check out and analyze the data and explain its component and its results and figure out how the story of a business came to that way. In addition, we use descriptive analytics to check the correlations between the variables and what they depend on and prepare the data for model's building.

Advantage: understanding the weaknesses of a business by knowing the story of it, for example, if we have a clothing shop, using descriptive analytics helps us to know whether brand A has more sales than brand B, so we can increase our brand A products and lower brand B products.

Techniques: Frequency Measures such as:

Frequency: which shows how many times this value appeared in a dataset.

Relative Frequency: which shows the percentage of how many times this value happened compared to the rest of the dataset.

Central Tendency Measures: which means describe the value position in the dataset such as:

Mean: the sum of the whole data divided by its number which helps us to know what the average of the data is. for example, the salary of employees has a mean of 30000$ thousands, that tell us that the salary of employees is around the 30000$ thousands.

Median: The middle value of a dataset. For example, the salary of employees has a median of 50000$ thousands, that can help data scientist to identify the outliers, and we can understand that half of employees are less than 50000$ thousands and the other half is more.

Mode: The times of the frequency of a value. For example, the salary of employees has a mode of 6000$ thousands, that tells us that most employees salary is 6000$ thousands.

Dispersion Measures: measures that tells us the nature of the dataset such as:

Range: the difference between the biggest value and the lowest value which shows us the max and the minimum values of the dataset and tells us how the data spread.

Standard Deviation: uses to check whether the data are spread out from each other more or less, the highest deviation is, the more the data is spread.

Position Measures:

Percentile: it shows a specific percentile of the dataset to know it's what the value is in that percentile.

Decile: it divides the dataset to ten equal parts, and each decile contains 10% of the data, and it shows the value of 10th percentages which helps us understand how the data spread in each 10th percentage.

Contingency Table: it shows how much a specific observation is repeated according to another value, for example, we make a contingency table to how many times a stocking market trades happened in each month.

Predictive Analytics:

Definition: It is the science of future predictions by using the historical data that we have collected and prepare them to use machine learning and predict and make decisions based on it. Predictive analytics can be in every field that has data and intentions to make future predictions. For example, in healthcare, we can make predictions of whether a patient has heart disease or not, in sports, we can make predictions of whether a player in the team will perform well in the future, in business, we can make predictions of whether the stocks prices will increase or decrease.

Advantage: Predict the future to get the best result of actions.

Techniques: Machine learning: by using models to predict such as decision tree, linear regression, or K-Nearest-Neighbors etc… and regression analysis which shows us the relationship between two or more variables. In addition to decision making for the future.

Prescriptive Analytics:

Definition: It is the science of using statistical algorithms and machine learning techniques to analyze the data and give the best results for a business after optimizations to lower costs and efforts and increase productivity and profit.

Advantage: It provides the best solution to manage a specific task on how the organization can accomplish the task with less cost and best work.

Techniques: it requires data science, machine learning and optimization methods to generate the best action of a task.

Use For Each Analytics Method:

Descriptive Analytics:

Markets now generate data of their own business, such as clothing market by recording when the cloth got sold and how much did it cost and how much it was sold and the brand of the cloth and the type of the cloth whether it is a summer cloth or a winter cloth. The market was losing profit, so the science of descriptive analytics became obligatory to determine why the story became in that way, by analyzing the data through visualizing it, it shows that the sales on winter clothing is not big enough to give a good profit, and it shows multiple reasons for that such as unacceptable of the winter clothing of the market, or maybe because the timing was not good, and after more analyzing, it shows that the sales on all winter clothing is average but only one brand was making less than the average, so, it shows that the reason is that customers does not love this clothing brand, which now we can conclude and build up a decision to make the market release the contract with the brand and stop buying from it and starting to look for another winter clothing brand that will satisfy customers and gain profit, or maybe increase the number of products of the most successful winter brand that the market has.

Predictive Analytics:

In healthcare, there are many records of patients that stored to generate for us data which can be used in machine learning to predict something such as predicting whether a patient in different health measures has diabetes or not based on historical data that a machine learning model has trained on so he could predict the patient status of having diabetes or not.

Prescriptive Analytics:

When we have multiple shipping branches, and we need to split the products to ship in all branches will lowest cost possible without exceeding the limits of how many a branch can ship, we use prescriptive analytics by using optimization algorithms to give the best possible outcome with least cost and without exceeding the limits of the branches ability.

Descriptive Analytics (My Work):

Feature no.	Feature Name	Descriptive Measure / Technique	Explanation
1	HIGH	Decile (Measure Of Position).	I used decile to see what the value in each 10th percentage of the data is to realize in what percentage the outliers got detected in the data and what are the values under the detected decile.
		Percentiles (Measure Of Position).	I used percentiles to see in what exact percentile of the data the outliers started to being obvious.
		Mean (Measure Of Central Tendency).	I used the mean to see what the average prices of the Jordanian stocks market is, which shows us in what area the stocks prices are around and what is the amount of money that could be considered enough to start trading.
2	NO_OF_TRADES	Decile (Measure Of Position).	I used decile to see what the value in each 10th percentage of the data is to realize in what percentage the outliers got detected in the data what are the values under the detected decile.
		Percentiles (Measure Of Position).	I used percentiles to see in what exact percentile

			of the data the outliers started to being obvious.
		Coefficient Of Variation (Measure Of Dispersion).	I used coefficient of variation to see how much values in this feature are spread out comparing to the mean which shows the difference between the values positions. The high coefficient of variation is the more difference between values are compared to the mean.
3	MARKET	Mode (Measure Of Central Tendency).	I used mode to see what market the most trading in the Jordanian stocks is.
		Frequency (Measure Of Frequency).	I used frequency to understand how many trades happened in each market.
		Relative Frequency (Measure Of Frequency).	I used relative frequency to see what the percentage of each market is because when communicating with stakeholders, they would prefer a percentage of the data not an exact number. For example, instead of saying market 1 has 8064 trades, we can say that market 1 has trades equals to 28% of the total trades.

Extra Necessary Information:

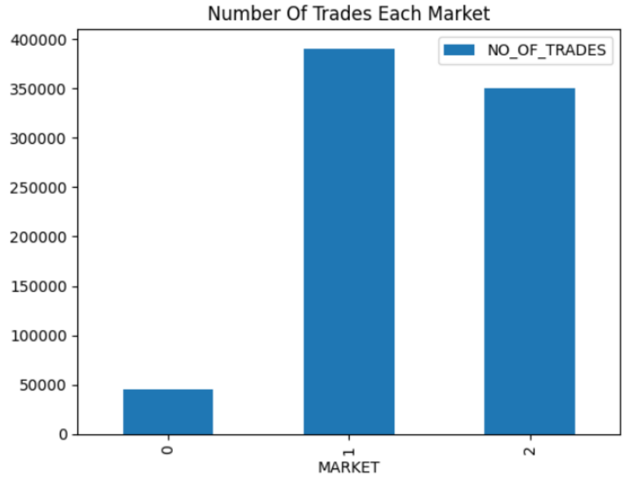
The minimum of high price is 0.02.

The maximum of high price is 44.5.

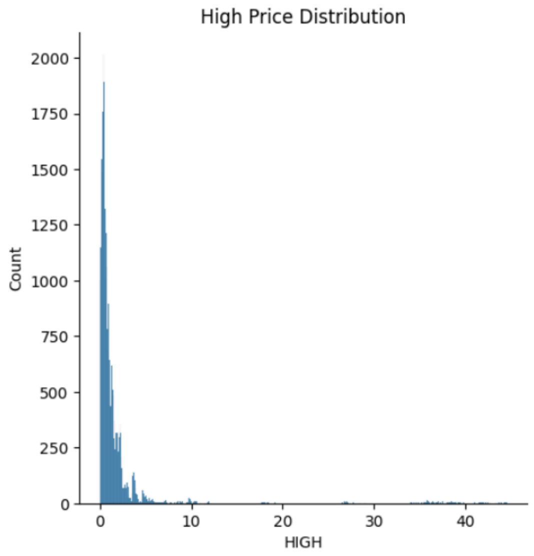
The minimum of number of trades is 1.

The maximum of number of trades is 1700.

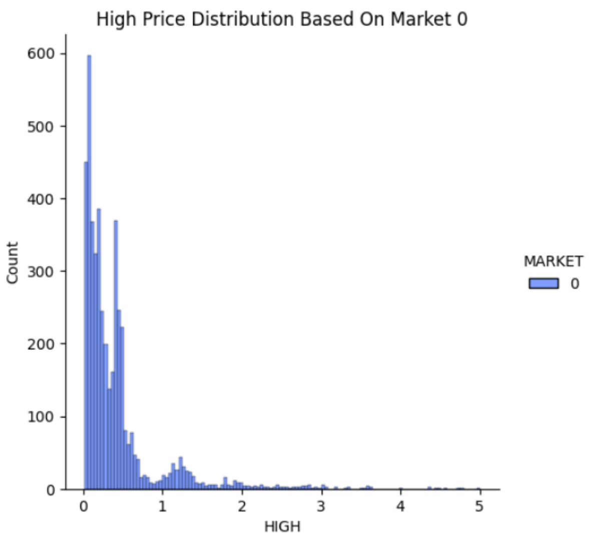
Visualizing Of The Features & Decision Making:

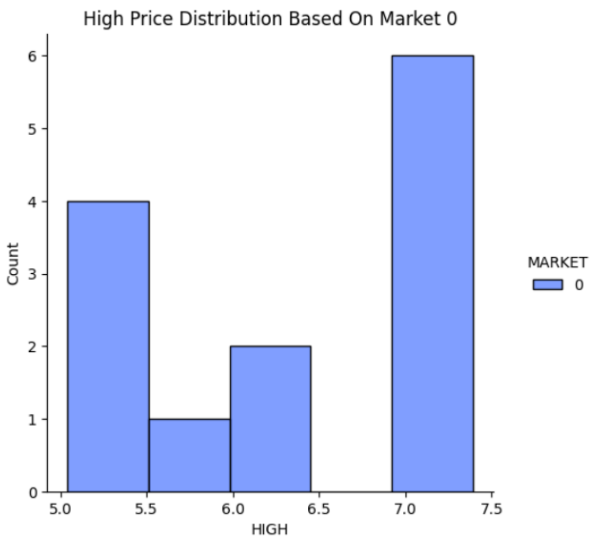


I used this figure to see the summation of the number of trades in each market. And as we can see, market 1 has many more trades than the other markets, which makes us conclude that people buy and sell stocks the most in market number 1 then market number 2. And about market number 0, there are much less trades compared to the other markets. We can build up a decision which avoids trading in market 0 most of the time because there is something on it for the lack of trades.



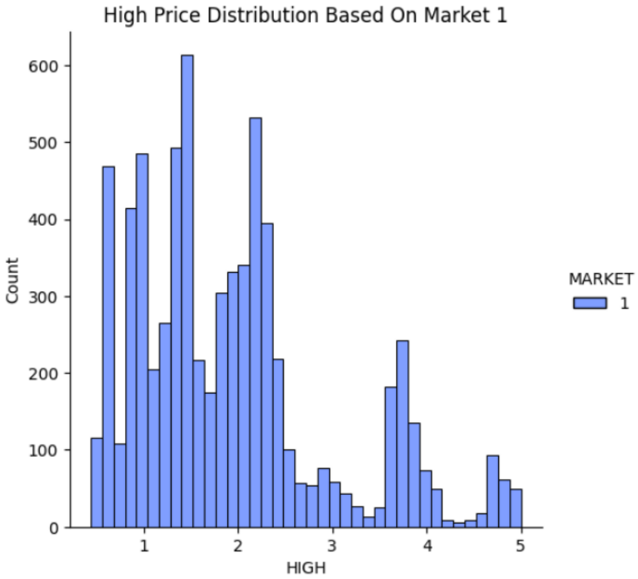
I used this figure to see the distribution of the high prices that a stock reached. And as we can see, approximately from after 0 to 5, there are many stocks that reached this prices, then the number of stocks that have larger price rapidly decreases.

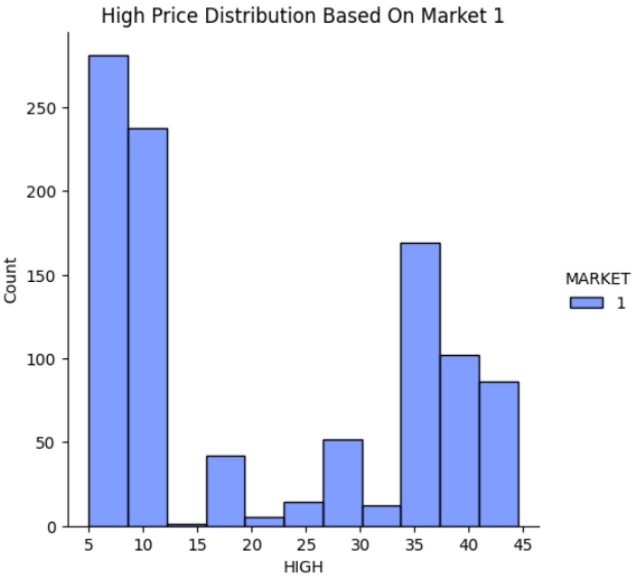

High Price Distribution Based On Market 0


High Price Distribution Based On Market 0

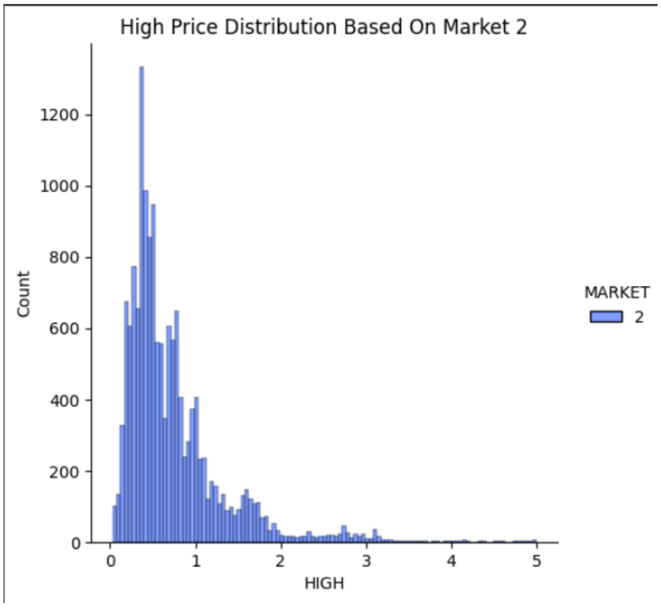
The graph in the left shows the high price of stocks from after 0 to and equal 5 for market number 0. Which shows that the distribution is nearly the same as the distribution of high prices. However, the graph on the right shows the high price of stocks from 5 to the maximum price which is 44.5 shows that the highest price in market 0 has reached approximately 7.4, which might be a reason that justify the low number of trades that happened in market number 0.

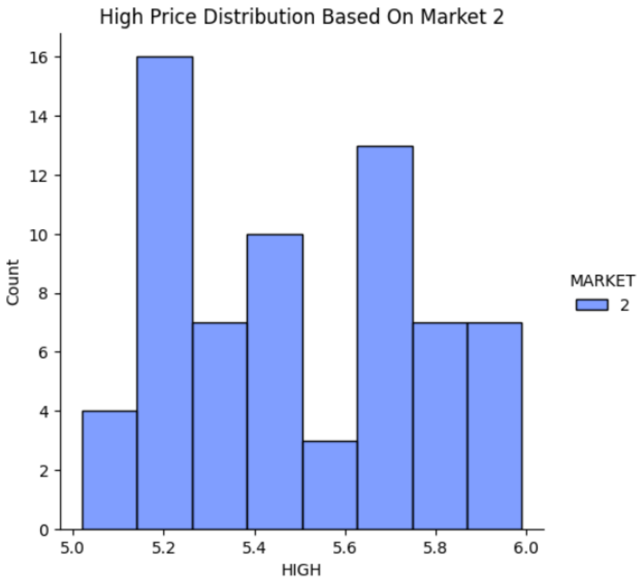
Theory 1: The low prices that stocks reach in market number 0 might be the reason of the low number of trades of it.


High Price Distribution Based On Market 1


High Price Distribution Based On Market 1

The graph in the left shows the high price of stocks from after 0 to and equal 5 for market number 1, and the right is after 5 to the maximum price, we can see that market number 1 has all the prices from the minimum price 0.02 to the maximum price 44.5.

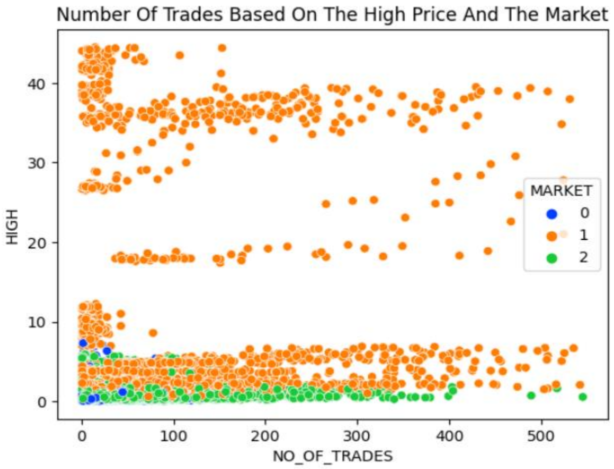


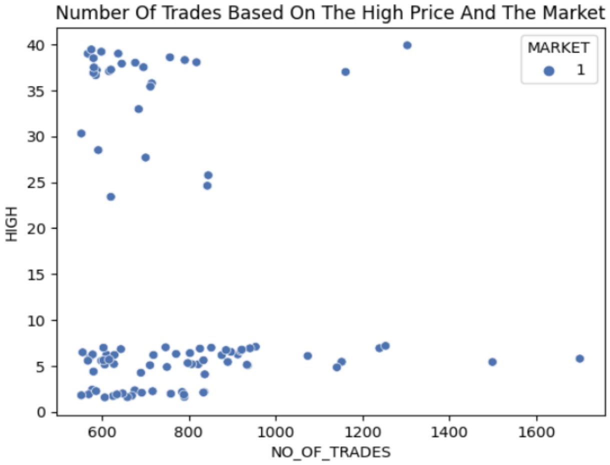


The graph in the left shows the high price of stocks from after 0 to and equal 5 for market number 2, and the right is after 5 to the maximum price. From theses graphs, we must change on theory 1 about market 0 because market number 2 trades are close to market number 1 even that the market number 2 stock prices reached to approximately 6 only.

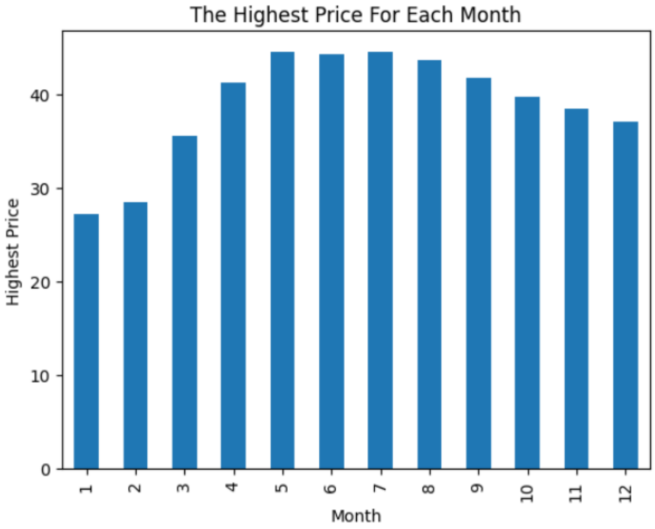
Conclusion + Theory 2: high prices of stocks are not a determiner of the number of stocks. Giving this, theory number 1 is disproven.

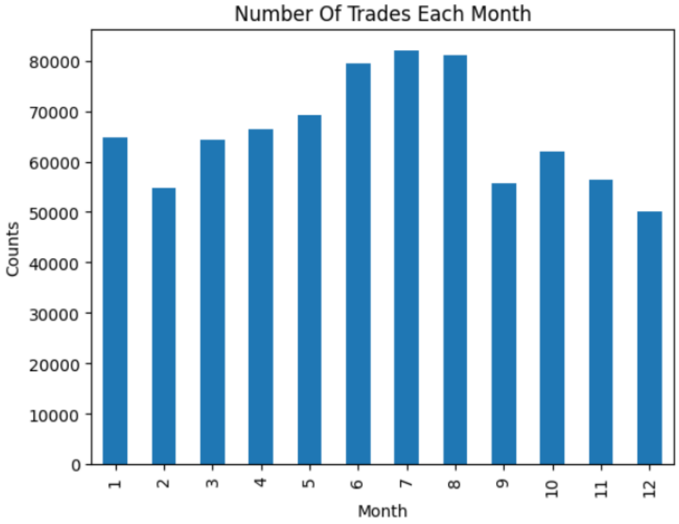
From these visualizing, We can build up a decision for traders, I recommend to buy stocks from market 0 and market 2 because the high prices it reaches is low so you won't lose from buying it, and you can sell from market 1 because the high price in market 2 are much bigger than market 0 and 1. In addition, you can begin trading in stocks in market 0 and 2 so you won't lose much if you have lost.





These two graphs show us the number of trades based on the high prices with colors to show which market it is. The left graph shows us the number of trades less than and equal to 545, which can tell us that market number 1 is dominating the other markets in terms of the number of trades and the prices of stocks. Furthermore, the right graph shows us the number of trades greater than 545, which all of them are from market 1. Which we can conclude that market 1 is the favorite market in terms of prices of stocks and it is the favorite for people to buy and sell on.

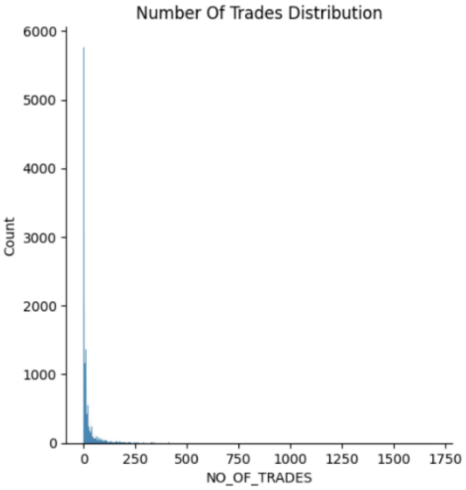




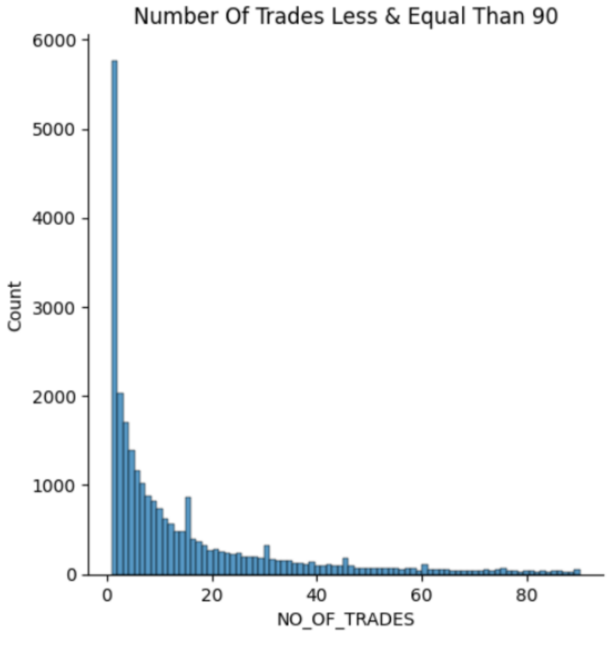
In the left graph, it shows the maximum price of all stocks in each month, and we can say that in the middle of the year, the stocks prices reach to the highest price. Furthermore, in the right graph, it shows the summation of total number of trades each month, and from the graph, we can say that month 6,7 and 8 are the largest number of trades compared to the rest of months.

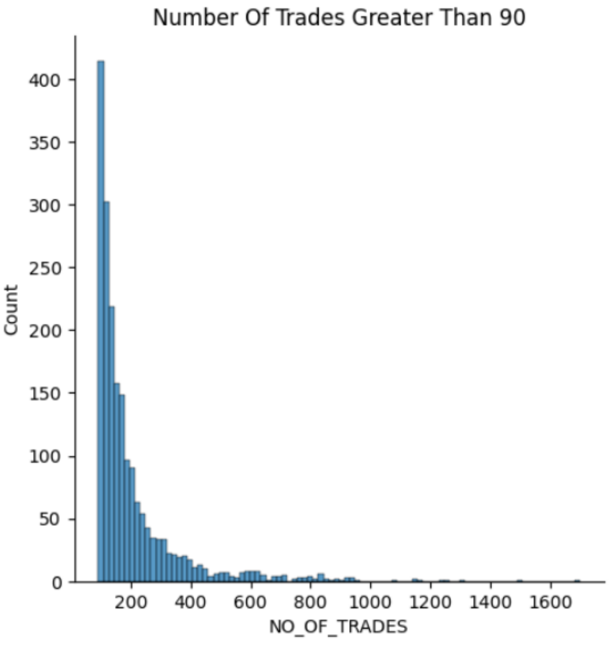
Theory 3: In June, July, and August, it is a middle easter vocation from schools and universities, and the official vocation for Gulf countries which might lead to increasing of the number of trades. However, it is a weak assumption.

We can build up a decision which is it is favorable to sell the stocks in middle months (June, July, August) because the prices reach the highest than any other month. In addition, increase the probability of selling the stock because these months have the most trades.



In this graph, we can’t really see the distribution of number of trades very well, yet it can show us that the number of trades has a big start then a big curve of decreasing.





The left graph shows the number of trades from after 0 to and equal 90, which shows that there is almost 6000 record of 1 trade only, then the trades rapidly decreasing. In the right graph, we see that about 90 and after, number of trades became about 500 trades.

Some extra information shown while analyzing:

ALFA is the lowest gain with (1) trade only and a volume equals to (37.4$).

JOPH is the highest gain with (68677) trades and a volume equals to (319533626.62$).

JOPT is the second highest gain with (88038) and a volume equals to (291098910.1$).

The difference between Rank 1 and Rank 2 in gains is equal to (28436716.52$)

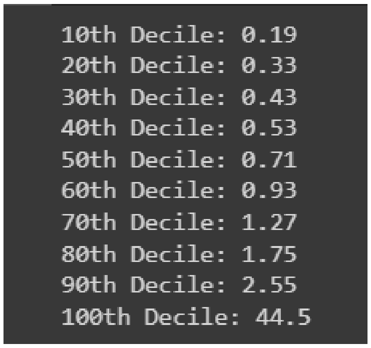
Decision Making For Measures:

Outliers:

First, I did not remove the outliers in the dataset because these observations represent the Jordanian stock market which each observation consider as a living soul that we need to take care of and analyze them just as any other observation. However, I considered that high stocks greater than 5$ is outlier which we will show why below.

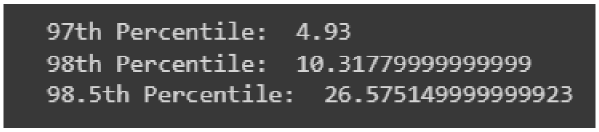
For Features Measures:

High:



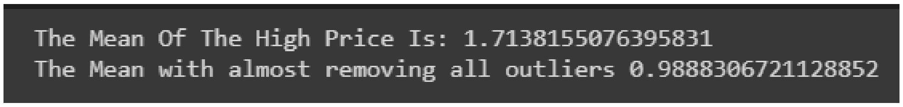
```
10th Decile: 0.19
20th Decile: 0.33
30th Decile: 0.43
40th Decile: 0.53
50th Decile: 0.71
60th Decile: 0.93
70th Decile: 1.27
80th Decile: 1.75
90th Decile: 2.55
100th Decile: 44.5
```

As we can see, the 9th decile shows that all the data until 90% percent of it are under 2.55$, which is nearly in the same age, however there is an increase in the prices between 90% to 100%, so, to understand where the outlier starts, I needed to implement percentiles measure to check in what percentile the outliers started.



```
97th Percentile:  4.93
98th Percentile:  10.31779999999999
98.5th Percentile:  26.575149999999923
```

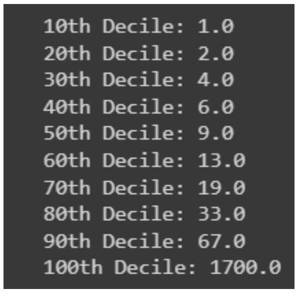
As we can see, nearly 97 percent of the data considered in the same domain of the data. However, in the 98% percentile, we could see the outliers started to appear. In another meaning, we can say that 2 percent of the data are considered as outliers due to the difference between the normal range and them.



```
The Mean Of The High Price Is: 1.7138155076395831
The Mean with almost removing all outliers 0.9888306721128852
```

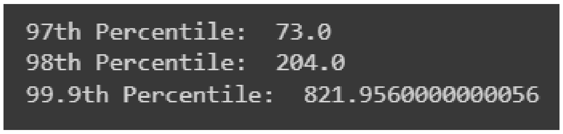
As we can see, the mean for all data is 1.7$, which we can understand from the fact that if a person wanted to start trading, 1.7$ consider a good value to cover most of the prices. Furthermore, the mean for all data expects the 2% percent that is considered to be as outliers is 0.98$ which means a person is able to get a good percentage of stocks, and that's why I measured with and without outliers because people in stock market want to pay less to get more.

Number Of Trades:



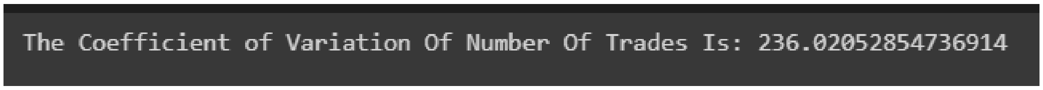
```
10th Decile: 1.0
20th Decile: 2.0
30th Decile: 4.0
40th Decile: 6.0
50th Decile: 9.0
60th Decile: 13.0
70th Decile: 19.0
80th Decile: 33.0
90th Decile: 67.0
100th Decile: 1700.0
```

As we can see, in the 9th decile, the data are nearly within the same range, and in the 10th decile, the trades 1700, but I could not consider any outliers in number of trades because it is a feature that is affected by so many factors and has no standards to limits despite the high feature which is affected by less factors than any other feature.



```
97th Percentile:  73.0
98th Percentile:  204.0
99.9th Percentile:  821.9560000000056
```

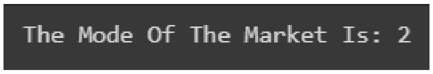
As we can see, nearly 97 percent of the data considered in the same domain of the data. However, in the 98% percentile, we could see the outliers started to appear. In another meaning, we can say that 2 percent of the data are considered as outliers due to the difference between the normal range and them.



```
The Coefficient of Variation Of Number Of Trades Is: 236.02052854736914
```

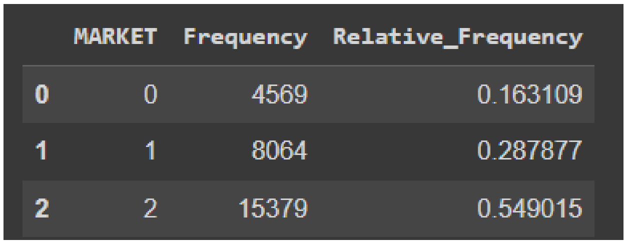
As we can see, the coefficient of variation for number of trades feature is 236% which means that there are spread out so much between values compared to the mean.

Market:



```
The Mode Of The Market Is: 2
```

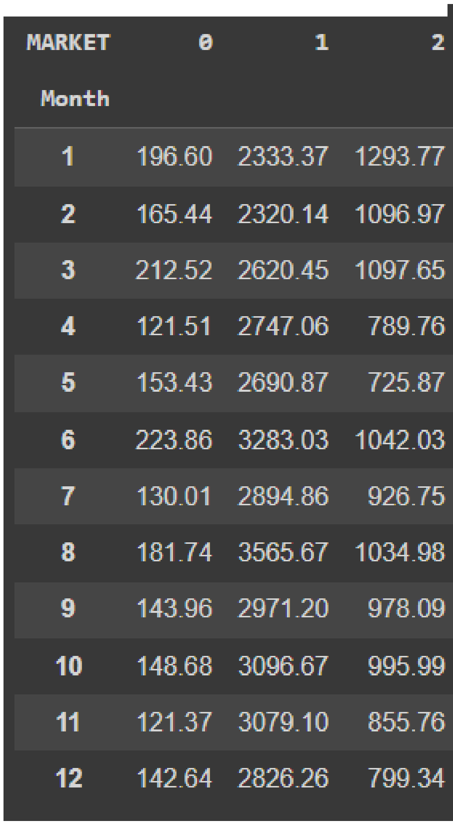
As we can see, the most recorded data was in market 2.



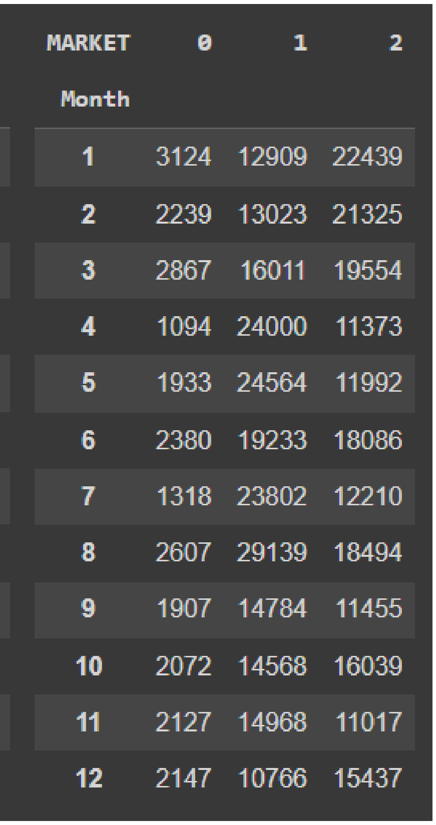
	MARKET	Frequency	Relative_Frequency
0	0	4569	0.163109
1	1	8064	0.287877
2	2	15379	0.549015

As we can see, the frequency for each market which tells us what the most recorded data is, and for relative frequency, it represents the percentage of each market in the data and how much it represents from the data.

Contingency Table: it shows how much a specific observation is repeated according to another value, for example, we make a contingency table to how many times a stocking market trades happened in each month. For my contingency table:



MARKET	0	1	2
Month			
1	196.60	2333.37	1293.77
2	165.44	2320.14	1096.97
3	212.52	2620.45	1097.65
4	121.51	2747.06	789.76
5	153.43	2690.87	725.87
6	223.86	3283.03	1042.03
7	130.01	2894.86	926.75
8	181.74	3565.67	1034.98
9	143.96	2971.20	978.09
10	148.68	3096.67	995.99
11	121.37	3079.10	855.76
12	142.64	2826.26	799.34



MARKET	0	1	2
Month			
1	3124	12909	22439
2	2239	13023	21325
3	2867	16011	19554
4	1094	24000	11373
5	1933	24564	11992
6	2380	19233	18086
7	1318	23802	12210
8	2607	29139	18494
9	1907	14784	11455
10	2072	14568	16039
11	2127	14968	11017
12	2147	10766	15437

The left table is the summation of prices for each month for each market.

The right table is the total number of trades for each month for each market.

From the right table, I can see how many trades happened in each month and in each market perfectly, it helped me to know that market 0 is active the most in January. And in august for market 1. And in January for market 2. And by the help of the summation of prices for each month in each market in the left table. I can conclude that:

Market 0:

People trade much in January with a total of 3124 trades and a total price equals to 196.60$. However, in June, it has the most prices between all months with a total price equals to 223.86$ with a total number of trades equals to 2380 which is less than January yet it gives more money. Also for November and December, the total prices are about 130$ with almost 2200 trades which is good months to buy stocks due to the low price of it, and selling these stocks in June because it gives the highest profit possible in market 0.

Market 1:

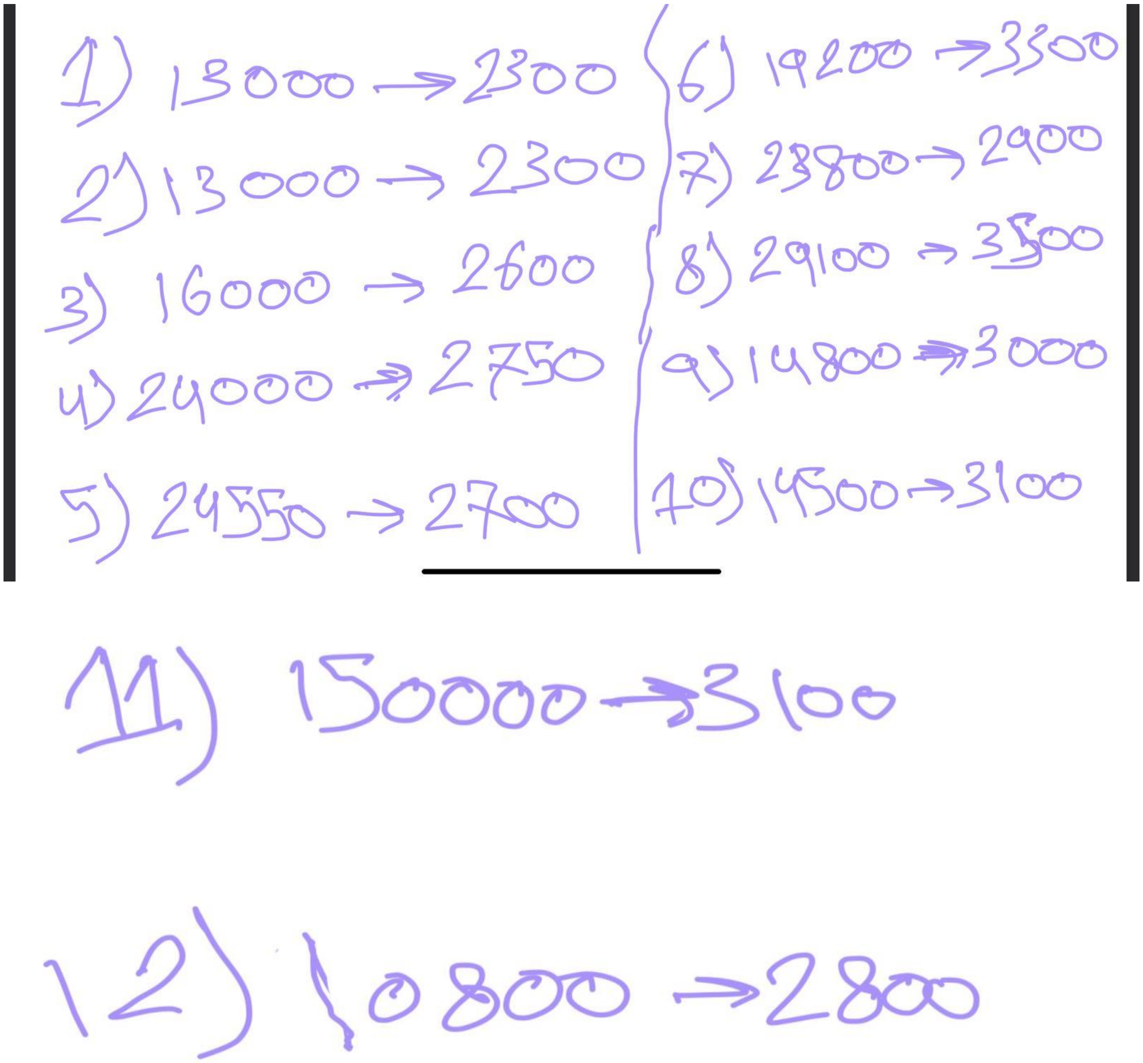
People trade much in August with a total of 29139 trades and a total price equals to 3565.67$. However, in April and May and July, the prices were almost 2750$ for both and trades almost 24000 for both April and May and 2900$ with 23800 trades in July, which can make us conclude that all of these trades gave low price comparing to other months. As a result, for that, I recommend buying stocks in April and May and July because it gives the lowest price and sell stocks in June and September and October and November and December because these months has less trades than other months and give high prices which means increasing the profit.

Market 2:

People trade much in January with a total of 22439 trades and a total price equal to 1293.77$. However, in July and September, the prices are higher than any other months according to the number of trades, July has 12210 trades with a price equals to 926$, and September has 11455 trades with a price equals to 978$. I recommend that whatever month you bought in (because numbers are approximately near each other so no need for false assumptions about when the best time to buy stocks because it is cheap), just keep in mind to sell it in July and especially in September because these two months has higher prices according to the number of trades.

Evaluation:

My analyzing was not enough because there are so many other features that I could use to prove my theories and strengthen up my decisions because they are considered weak assumptions and it needs to be proven by more evidence. In addition, I should have use more techniques to check what is the best feature selection technique even if I got high performance because sometimes it would be better to try more feature selections to whether check if there is better than the one you are using, or to prove that the one you used is the best. Furthermore, I should have been able to handle visualizing better, even though I'm convinced with the charts that I have made, but there is an inner feeling that no matter how good I have done, there are still better versions than what I have done, but it is all around the other features, because for surly thing that if I analyzed other features I would have seen hidden patterns. In addition, I tried my best to analyze the contingency table because it is so important and shows hidden patterns about the stocks prices and trades, I also analyze it using a simple pattern to check how many trades in each month gave the price. Lastly, I should have done more research on the Jordanian market and ask multiple traders and companies to analyze and explore how to build decisions and understand why these trends happened.



1) 13000 → 2300
2) 13000 → 2300
3) 16000 → 2600
4) 24000 → 2750
5) 24550 → 2700
6) 19200 → 3300
7) 23800 → 2900
8) 29100 → 3500
9) 14800 → 3000
10) 14500 → 3100

11) 150000 → 3100

12) 10800 → 2800

Month) Total Trades → Total Price.

These analysis is for market number 1.

Predictive Analytics (My Work):

Feature Selection:

It is an algorithm technique to choose the best fit features to make the predictions properly with high performance, and to reduce the computational cost that the model does. It helps by reducing the amount of time taken to choose the best features.

FS no.	Name	Description	Results (Selected Features)
1.	Sequential Feature Selection.	It is a method to select the features based on how much it affects the performance. It has two ways, forward which is having empty set and adding features and backward having all features and removing the unnecessary. The performance is being evaluated by a measure called cross-validation which is a testing measure to see how accurate the model is. picking features can be determined.	For KNN: (SEC_CODE, SYMBOL1, MARKET, BEST_ASK_PRICE, BEST_BID_PRICE). For DT (High): (SEC_CODE, SYMBOL1, MARKET, BEST_ASK_PRICE, NO_OF_TRADES). For DT(Low): ((SEC_CODE, SYMBOL1, MARKET, BEST_ASK_PRICE, BEST_BID_PRICE). LR: (SEC_CODE, SYMBOL1, MARKET, VOLUME, TRADE_QTY).
2.	Select K Best	It is a method to select the features by measure their score on how its effect will help the accuracy to be better, then eliminate any other unused feature and keep on the highest score feature. the score is measured by statistical measures.	For all regression techniques (k = 6): (SEC_CODE, MARKET, VOLUME, NO_OF_TRADES, BEST_ASK_PRICE, BEST_BID_PRICE)

Regression Techniques:

Tech. no.	**Name**	**Description**
1.	K-Nearest-Neighbor Regressor.	It is a machine learning model which works based on the nearest values for the desired input. We will determine the number of neighbors (K), if we set the k value to 3, the model will take the nearest 3 values and take the mean of it and give the prediction.
2.	Decision Tree Regressor.	It is a machine learning model which is represented as trees that works on if else if statements to get the equation that fits the desired output according to the input.
3.	Linear Regression.	It is a machine learning model which fits a line according to the linear regression equation to get the desired predictions.

Compare Techniques:

Low Predictions:

Explanation for each measure:

Mean Absolute Error: It is the average of errors that the model has done between the predicted result and the actual result.

Mean Squared Error: It is the average of the squared differences between the predicted result and the actual result.

Root Mean Squared Error: It is derived from mean squared error, and it is the average of the errors between the predicted results and the actual results.

The difference between mean squared error and mean absolute error and root mean squared error is that MAE is not as sensitive to outliers as MSE and RMSE

R^2: It is a regression evaluation measure that shows how the model performs on predicting the target efficiently.

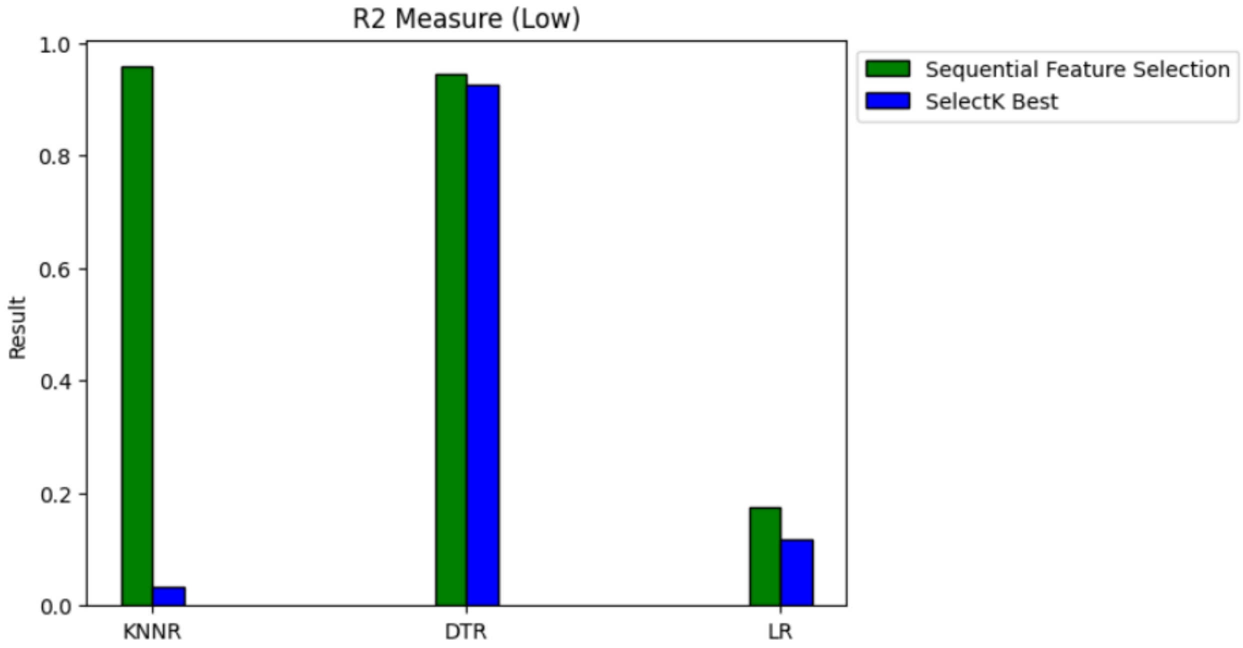
FS no.	Tech no.	MAE	MSE	RMSE	R^2
Sequential	**KNNR**	0.324	2.485	1.571	0.882
Sequential	**DTR**	0.213	1.182	1.081	0.944
Sequential	**LR**	1.603	17.284	4.154	0.174
SelectK	**KNNR**	0.89	9.439	3.07	0.548
SelectK	**DTR**	0.236	1.545	1.239	0.926
SelectK	**LR**	1.654	18.477	4.296	0.117

Visualizing:

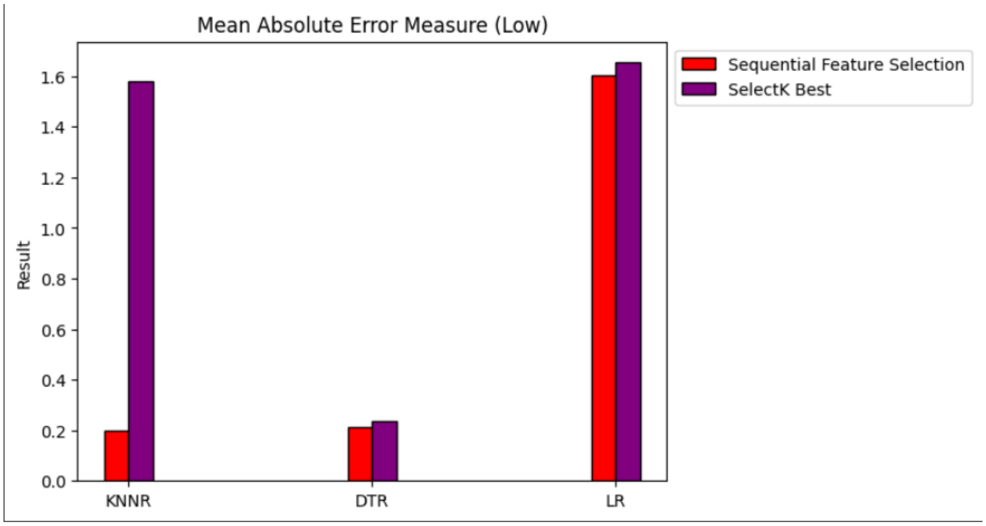
I used two charts which are:

Bar plot: a rectangular bar that expresses the value of specific object which is good for comparing and can help people to understand how much the difference between these values.

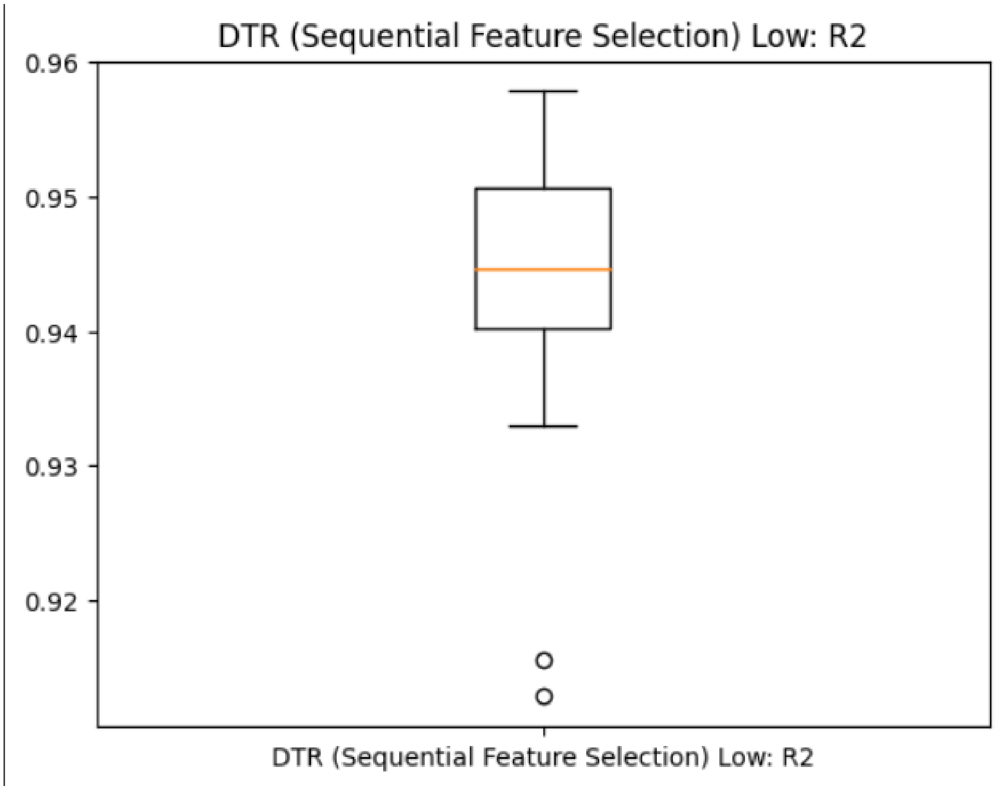
Box plot: a box that contains a list of values to be drawn within the box to see what the range of the data is and to see the outliers. In addition, box plots show the median of the data and preferred to be used only for data scientist because not all people can understand the box plot.



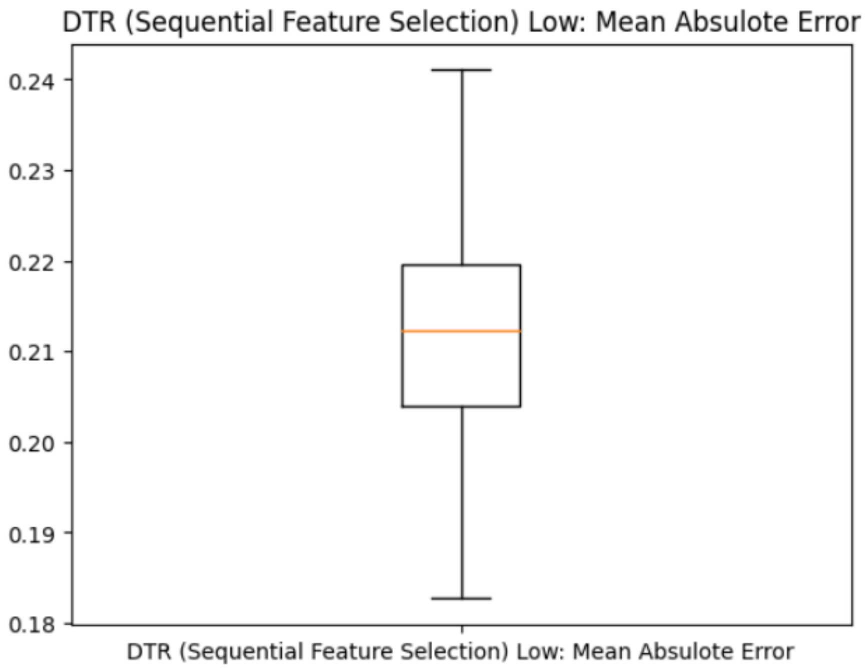
In this graph, it shows the R2 which shows how the model is good at predictions and from the graph we can see the results in the table and see that KNNR in sequential Feature Selection is the best.



This graph shows the mean absolute error which shows the absolute difference between the true value and the result and shows how many errors there are in the model.



In this figure, we can understand that the distribution of data, we can see that the R2 of decision tree model by using sequential feature selection are between the range from 94 to 95.5. and we can see some outliers that are under 92 percent.



This figure shows the mean absolute error box plot which shows that the distribution of the data is between the range of 0.20 to nearly 0.22.

KNN: in sequential feature selection, it was great performance which gave an R2 equals to 95 which is consider good model, and the select k best, the R2 equals to 0.033. however, increase K value would change the accuracy. So, it is obviously better to use sequential feature due to its better performance.

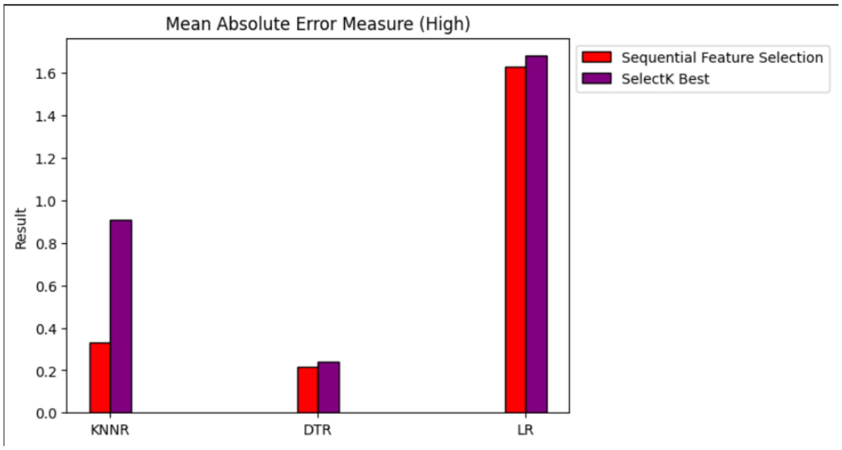
Decision Tree: in sequential feature selection, it was better R2 than knn which gave 94, and in seleck k best, the R2 was 92 which is way much better than KNN because decision tree is better at handling outliers and better in non-linear relations between features.

Linear Regression: in sequential feature selection, it gave R2 equals to 17 which is very low even in select K best it gave an R2 equals to 11, and that is due to the dataset not having a linear relationships between features.

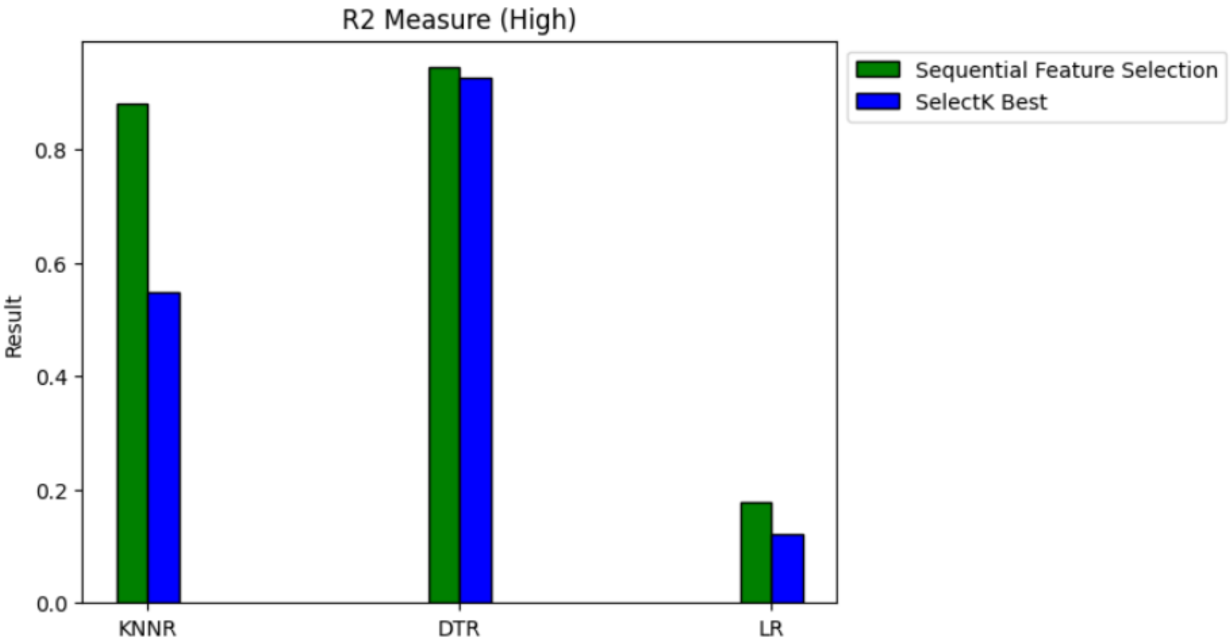
High Predictions:

FS no.	Tech no.	MAE	MSE	RMSE	R^2
Sequential	**KNNR**	0.332	2.629	1.616	0.879
Sequential	**DTR**	0.217	1.218	1.098	0.944
Sequential	**LR**	1.629	17.838	4.221	0.178
Select K	**KNNR**	0.908	9.784	3.125	0.549

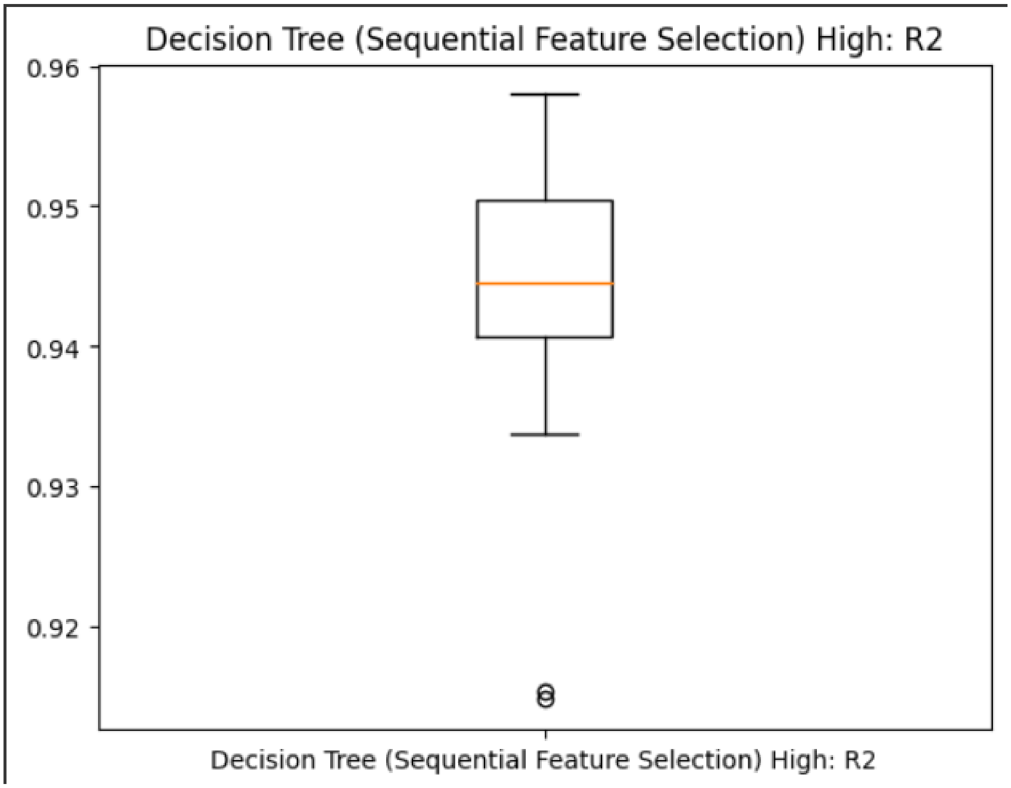
Select K	**DTR**	0.24	1.587	1.255	0.927
Select K	**LR**	1.681	19.101	4.368	0.12



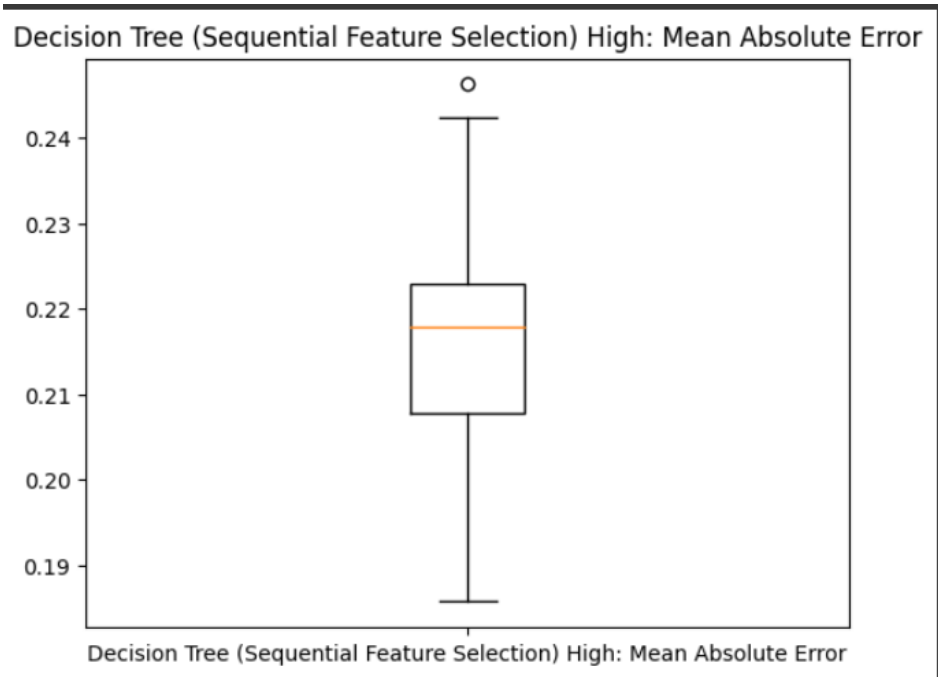
This graph shows the mean absolute error which shows the absolute difference between the true value and the result and shows how many errors there are in the model.



In this graph, it shows the R2 which shows how the model is good at predictions and from the graph we can see the results in the table and see that KNNR in sequential Feature Selection is the best.



This box plot shows the R2 of High feature, it shows that the distribution of the data which is between the range from 94 percent to 95 percent. in addition to the outliers that are under 92 percent.



This box plot shows the distribution of the data of high feature mean absolute error, which show as outliers that are above 0.24.

Note that box plots need multiple results to draw a box shows the distribution of the data, in these results, I have made multiple iterations to get multiple results and draw it in box plot. And for the bar plot, I have taken these results and took the average of them to draw it which is the right choice to make to try multiple iterations and take the mean of it.

KNN: in sequential feature selection, it was great performance which gave an R2 equals to 87 which is consider good model, and the select k best, the R2 equals to 54 which is obviously better to use sequential feature due to its better performance.

Decision Tree: in sequential feature selection, it was better R2 than knn which gave 94, and in seleck k best, the R2 was 92 which is way much better than KNN because decision tree is better at handling outliers and better in non-linear relations between features.

Linear Regression: in sequential feature selection, it gave R2 equals to 17 which is very low even in select K best it gave an R2 equals to 12, and that is due to the dataset not having a linear relationships between features.

The graphs are pretty good and show the difference between each model and each feature selection, which make people able to know the difference between all of them which make things easy to analyze and give the stakeholders better vision on the performances of the models and the features selection methods. However, I should have increased the performance of the low models that gave low measures, which could be increasing the number of K in the KNN model or increase the K of Select K best technique, but since I have got a good accuracy, I can count on it.

Tech. no.	Name	Description
1.	The Whale Optimization Algorithm	It is an optimization algorithm which mimics the behavior of food hunting of whales by exploring to see the high food concentration. This algorithm remembers where the best solution was from the past and try on improving it to give the best modified solution.
2.	The Particle Swarm Optimization.	It is an optimization algorithm which mimic a group of birds that are looking for food where each bird is helping the whole group to find the best food by using his expertise to help.
3.	The Genetic Algorithm.	It is an optimization algorithm which mimic the breeding of plants by taking the best genetics parents and breed them to produce a better genetics and repeat this process until getting the best optimal solution.

4.	The Firefly Algorithm.	It is an optimization algorithm which mimic the flashlight process of the fireflies bugs, it is considered as a computation technique that can search the feature space to give the best optimal solution.

Prescriptive Analytics:

Techniques For Finding The Best Course Of Action:

WOA: It is an algorithm that works by a group of whales (for example) that are in the ocean and have a desired goal to achieve such as eating, and each whale has his own position to search for in the space area. This algorithm has three phases, exploration phase is the first phase which assigning the whales to find new areas to cover all possible outcomes, then encircling which means that the algorithm has surrounded the optimal solution to a smaller area. Lastly the exploitation phase which means getting to the optimal solution or near optimal solution by using bubble net attacking model.

PSO: It is an algorithm that works by a group of population are working together to find an optimal solution for the problem, but the optimal solution would be hidden and not known, but all the population can communicate with each other, then, each element of the population is positioned in the space area and trying to work to see where the best solution is. Then while the population communicating with each other, they would be in track where the best solution is.

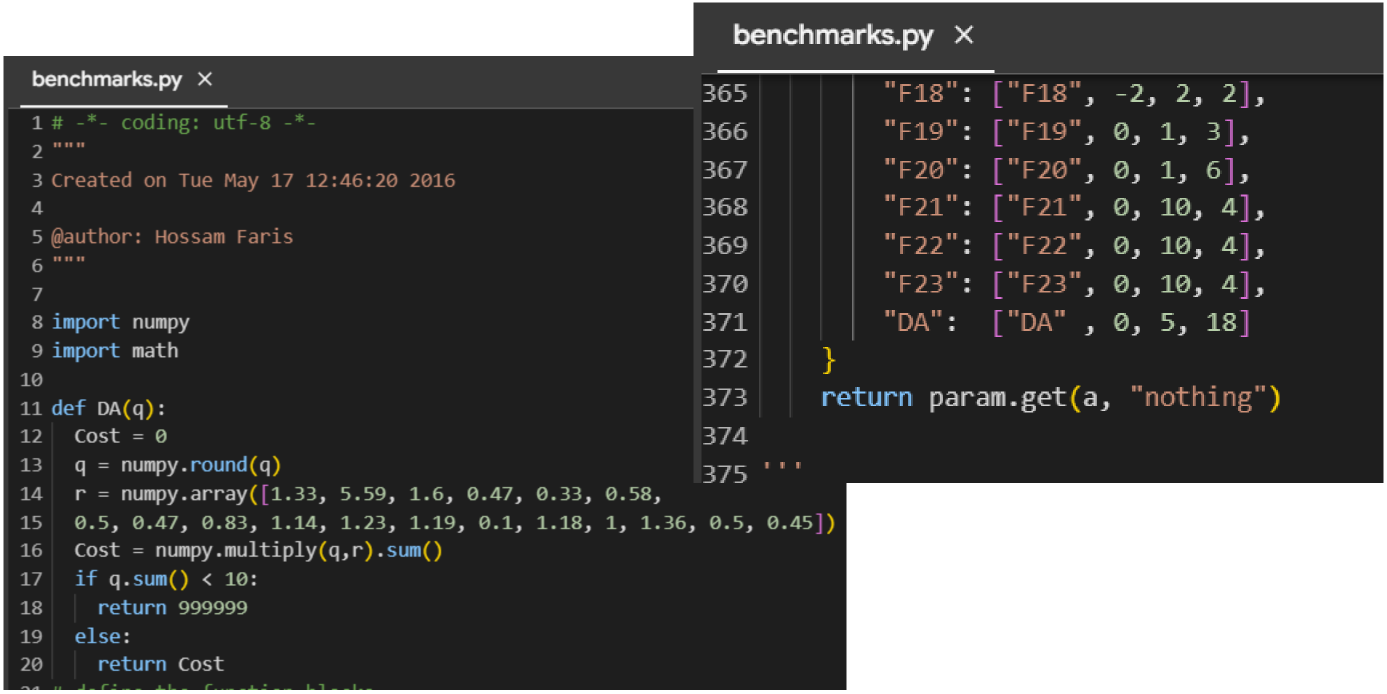
GA: It is an algorithm that works by a group of population that are a potential solutions for a problem, and each element of the group has special features that called genes, the good genes will transfer to the next generation which make the new population redesigned to solve the problem in a better way to give the optimal solution.

Objective Function:

```
def DA(q):
  Cost = 0
  q = numpy.round(q)
  r = 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])
  Cost = numpy.multiply(q,r).sum()
  if q.sum() < 10:
    return 999999
  else:
    return Cost
```

This is the function that will be using the three optimization algorithms to give us the minimum cost of buying 10 stocks.

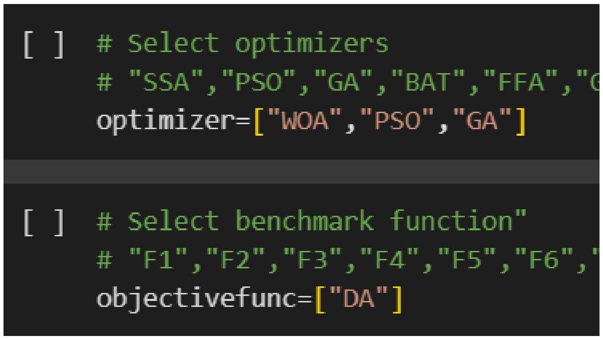
Applying The Techniques:



```
# -*- coding: utf-8 -*-
"""
Created on Tue May 17 12:46:20 2016

@author: Hossam Faris
"""

import numpy
import math

def DA(q):
    Cost = 0
    q = numpy.round(q)
    r = 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])
    Cost = numpy.multiply(q,r).sum()
    if q.sum() < 10:
        return 999999
    else:
        return Cost
# define the function blocks
```

```
            "F18": ["F18", -2, 2, 2],
            "F19": ["F19", 0, 1, 3],
            "F20": ["F20", 0, 1, 6],
            "F21": ["F21", 0, 10, 4],
            "F22": ["F22", 0, 10, 4],
            "F23": ["F23", 0, 10, 4],
            "DA":  ["DA" , 0, 5, 18]
        }
        return param.get(a, "nothing")

'''
```

In the screenshots above, in the benchmarks file, I added the objective function and added it to the function lists to use it.



```
[ ] # Select optimizers
    # "SSA","PSO","GA","BAT","FFA","G
    optimizer=["WOA","PSO","GA"]

[ ] # Select benchmark function"
    # "F1","F2","F3","F4","F5","F6","
    objectivefunc=["DA"]
```

In the screenshot above, in the main file, I changed to my preferences of optimization algorithms and changed the objective function to what I have just implemented.

Results And Explanation:

Desired plan is to buy 10 stocks from 3 banks in different days and has a limit quantity of 5.

There were different results for each optimization:

Whale Optimization Algorithm (WOA): in first iteration the cost was 26.6$ and in each iteration, the algorithm is trying to learn how to give the optimal price. At the second iteration which gave 9.44$, then couple of iterations more, the algorithm was learning how to get the optimal price. Then, it got the best possible price in the 16th iteration which was 4$ and stayed on that price till the end. We can conclude that the best of this algorithm could give with 50 iteration is 4$ only. I also tested twice more and the results were kind of close, in the second experiment it gave a cost equals to 7.97$, and the third experiment it gave a cost equals to 4.79$.

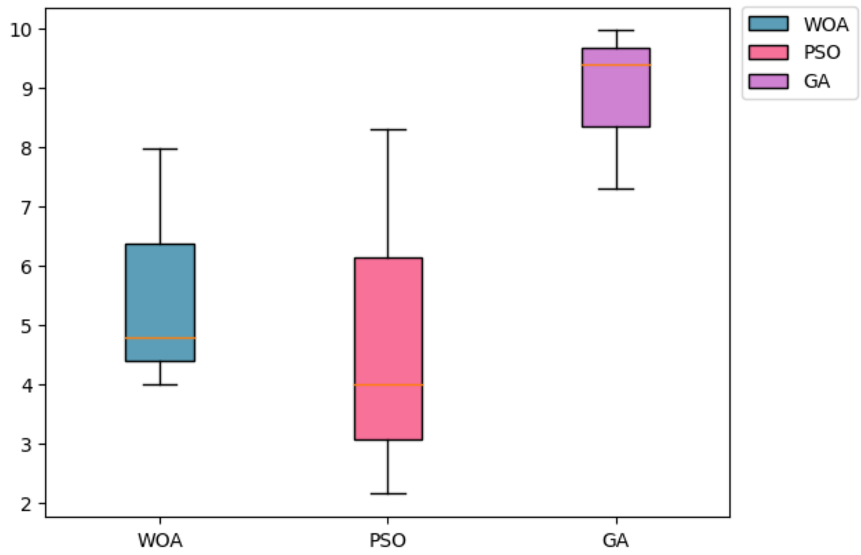
Genetic Algorithm (GA): In first iteration the cost was 31.9$, and in each iteration, the algorithm is trying to learn how to give the optimal price. Yet, it was slower than Whale Optimization Algorithm because in iteration 16, the genetic algorithm gave a cost equals to 18.06$ while the whale algorithm gave a cost equals to 4$, then by completing the 50 iterations, the genetic algorithm gave it's optimal best cost which was 7.3$. I also tested twice more and the results were close, in the second experiment it gave a cost equals to 9.4$, and the third experiment gave a cost equals to 9.97$.

Which make us conclude that in that it is required to do multiple experiments to see which one is better and try multiple algorithms because sometimes problems are designed for other optimization algorithms.

The Particle Swarm Optimization (PSO): In the first iteration, the cost was 28.14$, and in each iteration, the algorithm is trying to learn how to give the optimal price. However, in iteration 12, the algorithm already gave the best optimal price which was 2.15$. the algorithm gave said:

"Buy 5 stocks from Arab Bank in Wednesday where the stock is equal to 0.33$, the total cost will be 1.65$, and then, buy 5 stocks from Al Ahli Bank in Saturday where the stock is equal to 0.1$, the total cost will be 0.5$. And the total cost of these 10 stocks is 2.15$"

Visualization And Explanation:



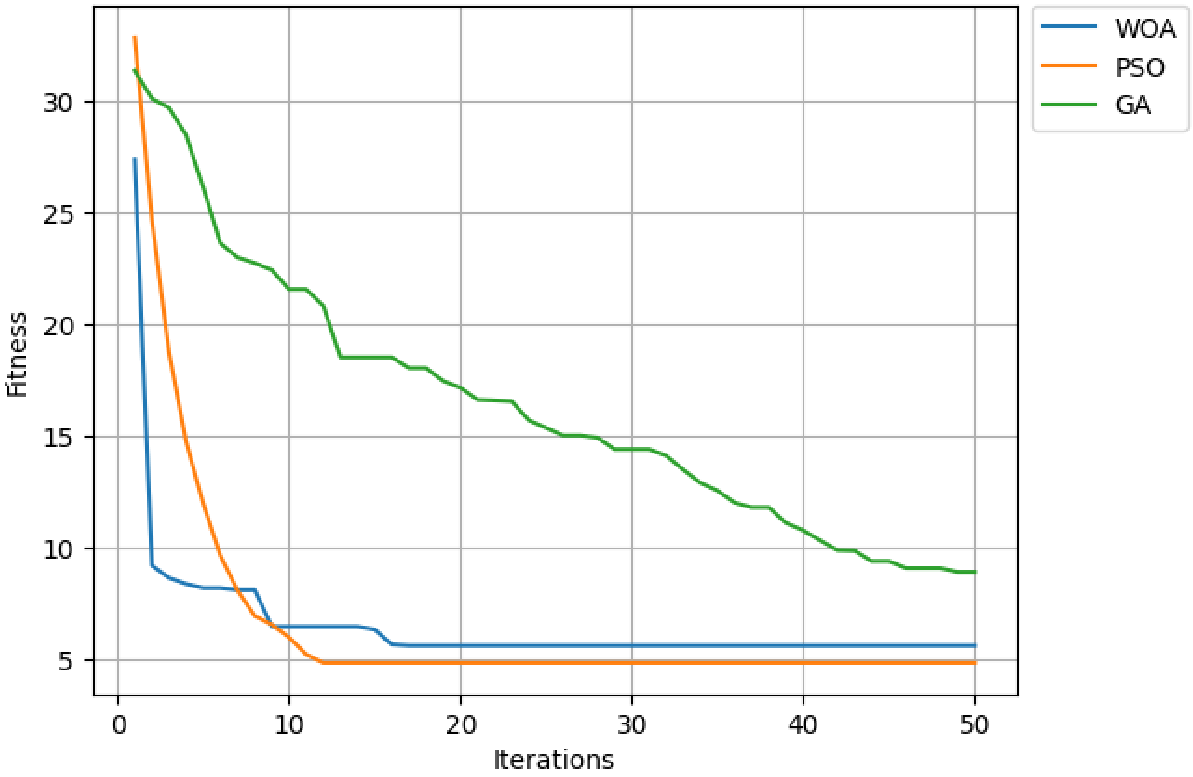
Box plot is a chart that shows how the data is distributed in a rectangular shape. And as we can:

WOA: it shows that the iterations values were approximately from 4 to 8 and most of the data were from 4.5 to 6.5.

PSO: it shows the iterations values were approximately from 2 to 8.5, and most of the data were from 3 to 6.

GA:it shows the iterations values were approximately from 7.5 to 10 and most of the data were from 8.5 to 9.5.

From the boxes, we can tell that the values in GA were close to each other yet the minimum it reached was about 7. And for PSO, we can see that values were widely distributed and the minimum of it is 2.15.



The line plot above is a comparison between the three algorithms which shows the number of iteration in the x axis and the price in y axis to show us what algorithm reached the optimal first and how was the learning curve for each algorithm.

The line graph shows us how each algorithm was doing, as we can see, the PSO learning curve was so high because it reached the optimal price at iteration 12, then it kept at the same cost. For the WOA algorithm, the learning curve was higher at the first of iterations then it almost reached the final optimal price, and in iteration 16, the cost was the final optimal solution that this algorithm could provide. For GA, the learning curve was slowly decreasing but did not reach to a good price comparing to the other two algorithms, it did reach the final optimal price it could give in iteration 46 which was too late comparing to PSO and WOA.

References:

Hillier, W. (2023). *The 9 Best Data Analytics Tools [For 2021 And Beyond]*.

Stitch. (n.d.). *5 benefits of data analytics for your business | Stitch resource*.

G, N. (2021). *5 Important Benefits of Excel for Students*. [online] career guide.

www.javatpoint.com. (n.d.). *10 Disadvantages of Microsoft Excel - javatpoint*.

www.nobledesktop.com. (n.d.). *Tableau Prerequisites: What is Required to Learn Tableau?*

kalyan (2017). *5 Core Activities of Data Analysis | Epicycles of Data Analysis*.

www.kancloud.cn. (n.d.). *3. Stating and Refining the Question · The Art of Data Science ·*

IBM (2020). *What is Exploratory Data Analysis? | IBM*.

Maryville Online. (2021). *Top 4 Data Analysis Techniques*.

Novotny, J. (2023). *A Programmers' Guide to Python: Advantages & Disadvantages*.

Scardina, J. (2022). *What is Microsoft Power BI? - Definition from WhatIs.com*.

Team, S. (2022). *10 benefits of Power BI*. [online] SysKit.

Google Cloud. (n.d.). *Looker Studio: Business Insights Visualizations*.

Investopedia. (n.d.). *Descriptive Analytics: What They Are and Related Terms*.

www.knowledgehut.com. (n.d.). *Descriptive Analytics: Steps, Techniques, Use Case, Examples*

Cote, C. (2021). *What Is Predictive Analytics? 5 Examples | HBS Online*. [online] Business Insights - Blog.

TIBCO Software. (2021). *What is Prescriptive Analytics?*

CIO. (n.d.). *What is Prescriptive Analytics? Definition from SearchCIO*.

Orbit Analytics. (n.d.). *Descriptive Analytics*.

scikit-learn. (n.d.). *sklearn.feature_selection.SequentialFeatureSelector*.

scikit-learn.org. (n.d.). *sklearn.feature_selection.SelectKBest — scikit-learn 0.23.0 documentation*.

DataTechNotes (n.d.). *SelectKBest Feature Selection Example in Python*.

Scikit-learn.org. (2019). *sklearn.linear_model.LinearRegression — scikit-learn 0.22 documentation*.

Scikit-learn.org. (2019). *sklearn.neighbors.KNeighborsRegressor — scikit-learn 0.22 documentation*.

Teixeira-Pinto, A. (n.d.). *2 K-nearest Neighbours Regression | Machine Learning for Biostatistics*.

scikit-learn. (n.d.). *Decision Tree Regression*.

GeeksforGeeks. (2018). *Python | Decision Tree Regression using sklearn*.

Tableau (2022). *Time Series Analysis: Definition, Types, Techniques, and When It's Used*.

resources.eumetrain.org. (n.d.). *Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE)*.

Mirjalili, S. and Lewis, A. (2016). The Whale Optimization Algorithm. *Advances in Engineering Software*,

GeeksforGeeks. (2021). *Whale Optimization Algorithm (WOA)*.

Tam, A. (2021). *A Gentle Introduction to Particle Swarm Optimization*.

Frankenfield, J. (2022). *Data Analytics.*

Emary, E., Zawbaa, H.M., Kamal, K., Aboul Ella Hassanien and B. Parv (2015). Firefly Optimization Algorithm for Feature Selection.

Gad, A.G. (2022). Particle Swarm Optimization Algorithm and Its Applications: A Systematic Review. *Archives of Computational Methods in Engineering*, 29(5), pp.2531–2561

Bouguezzi, W. by: S. (2023) *Whale optimization algorithm, Baeldung on Computer Science*.

Gad, A. (2018). *Introduction to Optimization with Genetic Algorithm*. [online] Towards Data Science.