Journal of Communication Sciences and Information Technology (JCSIT)
An International Journal

Using Machine Learning to Predict Missing Values in the Egyptian Stock Exchange

Ismail. M. Hagag

EL Madina Higher institute of Administration and Technology, Giza, Egypt.

Email: drismailhagag@gmail.com

Mohamed A. Amin

EL Madina Higher institute of Administration and Technology, Giza, Egypt.

Email: mohamedabdeilhamed@gmail.com

ABSTRACT

Financial markets are rich in information and encompass a wide range of variables. Contrary to the efficient market hypothesis, extensive research has been conducted to predict asset prices with promising accuracy. However, developing robust models requires the extraction of meaningful information from the available datasets. This study focuses on the main Egyptian stock exchange indices (EGX 30, EGX 50, EGX 70, and EGX 100) and constructs alternative portfolios by identifying significant linear combinations of the EGX components. This is achieved through the application of principal component analysis (PCA) followed by a missing data prediction technique. The results highlight the importance of the principal components derived from the analysis. Cross-validation (CV) of principal component regression (PCR) reveals that the most significant insights are obtained by analyzing trends in INDEXOPEN, INDEXHIGH, INDEXLOW, and INDEXCLOSE over time. These trends provide valuable insights into the overall performance of the index. The correlation coefficients between these indicators range from -1 to 1, where 1 indicates a perfect positive relationship, -1 indicates a perfect negative relationship, and 0 indicates no linear relationship. The analysis demonstrates very strong correlations (close to 1) between INDEXOPEN, INDEXHIGH, INDEXLOW, and INDEXCLOSE, indicating that these indicators move in tandem significantly.

Keywords*:* Machine Learning - Predict Missing Values -Egyptian Stock Exchange

Introduction

The Egyptian Stock Exchange (EGX) is one of the oldest financial markets in the Arab region and plays a pivotal role in the Egyptian economy. As the volume of financial data continues to grow, analyzing this data has become increasingly complex. A significant challenge faced by analysts and researchers is the presence of missing values in financial datasets, which can compromise the accuracy of analyses and predictions.

This study investigates the application of machine learning techniques to predict missing values in data from the Egyptian Stock

Exchange. By addressing this issue, the research aims to enhance data quality and improve the reliability of financial analyses. Machine learning, a branch of artificial intelligence, enables systems to learn patterns and relationships from data without explicit programming. This capability allows for the development of models that can adapt to new information and improve over time. The performance of machine learning depends on the selection and implementation of appropriate algorithms, which are designed to process large datasets and generate actionable insights. These insights can assist business owners and investors in making informed decisions.

Machine learning mimics human problem-solving by enabling systems to tackle complex challenges, even when encountering unfamiliar scenarios. When faced with a new problem, a machine learning system can explore potential solutions and adapt its approach based on the nature of the data. Key functions of machine learning include extracting relevant data, analyzing it to generate interpretable information, predicting future outcomes, and adapting to evolving conditions to produce innovative solutions. These capabilities make machine learning a powerful tool for addressing dynamic challenges in various fields.

In the context of financial markets, machine learning has emerged as a transformative technology for forecasting and decision-making. By leveraging artificial intelligence and advanced data analysis, machine learning can predict future trends and events with increasing accuracy. Its applications span diverse sectors, including e-commerce, marketing, healthcare, and finance. Recent advancements in machine learning algorithms have significantly improved predictive accuracy, enabling more reliable decision-making processes. However, forecasting stock performance remains a complex scientific challenge due to the vast and dynamic nature of trading data. This study highlights the importance of employing machine learning methods to predict stock performance, demonstrating how these techniques can assist investors in making data-driven decisions.

Machine learning models [9, 10]

Machine learning is a branch of artificial intelligence that uses statistical techniques to enable systems to learn and improve their performance on specific tasks over time. It focuses on identifying patterns within target datasets. Essentially, machine learning involves using algorithms to analyze data, derive insights, and make decisions or predictions based on that information. In summary, machine learning is a field dedicated to developing models that can learn from data and predict future outcomes without explicit programming.

requiring detailed programming for every task. It serves as a bridge between traditional artificial intelligence and deep learning. Deep learning, a subset of machine learning, aims to simulate the human brain by constructing neural networks capable of handling complex data with higher precision.

There are two types of Machine Learning models. On the one hand, Supervised Learning involves training a machine using a dataset with labeled examples. The machine learns from this data to identify patterns and make predictions based on new, unseen data. It is commonly used in social sciences for tasks such as classification and regression. On the other hand,

Unsupervised Learning, where the machine is given a dataset without any labelled responses. The machine independently seeks out patterns and relationships within the data. Unsupervised learning is employed when dealing

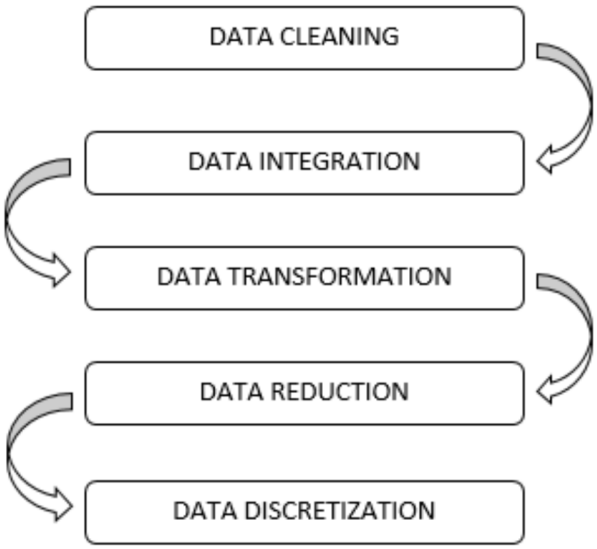

Fig. 1. Data Mining Task (7).

To use the prediction of missing values, the researcher used the prediction algorithms (Random Forests and Linear Regression), which can be exposed as follows:

Linear Regression [10, 11].

In simple linear regression, represented the combined effect upon the dependent variable of all the important variables not included in the model. It seems only natural, therefore, that we might wish to add variables we think are significant to the model to reduce the (unexplained) random variation in the dependent variable, i.e., producing a better-fitting model.

The model for multiple linear regression is given by

with large volumes of unclassified data, and it often involves iterative techniques such as deep learning to identify hidden structures and insights. In the following, we present the Four Machine Learning Models in detail.

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$, where

- k equals the number of independent variables in the model
- X_i is the i^{th} independent variable (out of k)
- Y and ε are random variables
- $\beta_0, \beta_1, \ldots, \beta_k$ are the parameters

II Assumptions

The Multiple Linear Regression model, $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$, makes two different kinds of assumptions.

A. Linearity

- The first of these, mentioned previously, postulates that the dependent variable Y is linearly related to the collection of independent variables *taken together*, i.e., $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$.

The Simple Linear Regression model $Y = \beta_0 + \beta_1 X$ hypothesizes that the relationship between Y and X can best be described by a line. Similarly, the Quadratic (Polynomial) model $Y = \beta_0 + \beta_1 X + \beta_2 X^2$ hypothesizes that the relationship follows a parabola. The Multiple Linear Regression model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, hypothesizes that the relationship between the dependent variable Y and the independent variables X_1 and X_2 can be pictured as a plane. The picture below right shows such a plane. This model may be suggested by experience, theoretical considerations, or exploratory data analysis.

Mean Squares [12, 13].

While a sum of squares measures the total variation, the ratio of a sum of squares to its degrees of freedom is a variance. The advantage of a variance is that it can be used to compare different data sets and different models (since it incorporates information about the size of the sample and the number of independent variables used in the model). These variances are called mean squares.

$S^2 = \frac{SST}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ = The (sample) variance of the y-values observed (see the notes "Review of Basic Statistical Concepts"). You probably never knew (or cared, perhaps) that the sample variance you computed in your introductory statistics course was an example of a mean square!

$MSE = \frac{SSE}{n-k-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$ = The (sample) variance of the dependent variable unexplained by the model. The Mean Square Error is the sample estimate of the variance of the error variable,

$MSR = \frac{SSR}{k} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}$ = the variance of the dependent variable explained by the model, the Mean Square for Regression.

Data description and preprocessing

This research deals with the EGX 30 index, which is a market value-weighted index of the 30 largest companies in terms of activity and liquidity. However, the Egyptian economy and political systems have witnessed a period of instability that has caused economic instability around the world for the past three years. After that, we studied the EGX 30 price index and its components for the past three years, where the total number of shares registered during this period was 42 shares. Description of individual stocks' names, symbols, and sectors is provided in Table 1.

Summary Statistics:

	INDEXDATE	INDEXOPEN	INDEXHIGH	INDEXLOW	INDEXCLOSE	TRADE_VOLUME	TRADE_VALUE	INDEXCODE_sentiment
count	2586	2586.000000	2586.000000	2586.000000	2586.000000	2.586000e+03	2.586000e+03	2586.0
mean	2014-04-21 07:16:33.967517184	8497.013592	8560.036241	8445.408445	8500.974629	1.350935e+08	5.150430e+08	0.0
min	2008-11-27 00:00:00	3389.310000	3484.100000	3380.420000	3389.310000	1.314341e+07	4.444680e+07	0.0
25%	2011-09-05 06:00:00	5585.255000	5627.977500	5541.917500	5586.492500	6.685154e+07	2.981348e+08	0.0
50%	2014-05-04 12:00:00	7133.885000	7171.340000	7072.530000	7139.125000	1.059057e+08	4.396475e+08	0.0
75%	2016-12-21 18:00:00	11983.457500	12138.727500	11914.462500	12017.627500	1.683166e+08	6.441167e+08	0.0
max	2019-08-27 00:00:00	18363.290000	18412.280000	18304.460000	18363.290000	8.092787e+08	3.563463e+09	0.0
std	NaN	3718.003556	3734.457734	3707.582283	3718.667454	1.045515e+08	3.296924e+08	0.0

This image depicts a table with summary statistics for a dataset. The dataset appears to contain information about some kind of index, potentially financial, over a period of time. Here's a breakdown of the information provided:

Columns:

INDEXDATE: The date of the record.

INDEXOPEN: The opening value of the index on that date.

INDEXHIGH: The highest value of the index on that date.

INDEXLOW: The lowest value of the index on that date.

INDEXCLOSE: The closing value of the index on that date.

TRADE_VOLUME: The volume of trades on that date.

TRADE_VALUE: The total value of trades on that date.

INDEXCODE _sentiment: A sentiment score associated with the index on that date.

Summary Statistics:

count: The number of records in the dataset (2586).

mean: The average value for each column. For example, the average INDEXOPEN is approximately 8497.

min: The minimum value for each column.

25%: The 25th percentile value for each column. This means that 25% of the data falls below this value.

50%: The median value for each column. This is the middle value when the data is sorted.

75%: The 75th percentile value for each column. This means that 75% of the data falls below this value.

max: The maximum value for each column.

std: The standard deviation for each column. This measures the spread or dispersion of the data around the mean.

Observations: The dataset spans a considerable period, as indicated by the count of 2586 records.

The INDEXCODE _sentiment has a mean of 0.0, suggesting that the sentiment associated with the index is generally neutral or unavailable.

The standard deviation for INDEXDATE is NaN, which is expected since it's a date column and not a numerical value.

Potential Insights: By analyzing the trends in INDEXOPEN, INDEXHIGH, INDEXLOW, and INDEXCLOSE over time, one could gain insights into the overall performance of the index.

The TRADE_VOLUME and TRADE_VALUE could provide information about the level of trading activity associated with the index.

Correlation Matrix:

	INDEXOPEN	INDEXHIGH	INDEXLOW	INDEXCLOSE	TRADE_VOLUME	TRADE_VALUE	INDEXCODE_sentiment
INDEXOPEN	1.000000	0.999817	0.999784	0.999506	0.443540	0.455188	NaN
INDEXHIGH	0.999817	1.000000	0.999791	0.999802	0.448643	0.461824	NaN
INDEXLOW	0.999784	0.999791	1.000000	0.999832	0.444309	0.455570	NaN
INDEXCLOSE	0.999506	0.999802	0.999832	1.000000	0.448739	0.461301	NaN
TRADE_VOLUME	0.443540	0.448643	0.444309	0.448739	1.000000	0.693599	NaN
TRADE_VALUE	0.455188	0.461824	0.455570	0.461301	0.693599	1.000000	NaN
INDEXCODE_sentiment	NaN	NaN	NaN	NaN	NaN	NaN	NaN

This image shows a table containing a correlation matrix for data. The data appears to include information about some type of index, perhaps financial, over a period of time. Here is a detailed breakdown of the information provided:

Columns:

INDEXOPEN: The index's opening value on that date.

INDEXHIGH: The highest value of the index on that date.

INDEXLOW: The lowest value of the index on that date.

INDEXCLOSE: The index's closing value on that date.

TRADE_VOLUME: The trading volume on that date.

TRADE_VALUE: The total trading value on that date.

INDEXCODE _sentiment: The degree of significance associated with the index on that date.

Correlation Matrix: This matrix displays the correlation coefficients between each pair of columns. The correlation coefficient measures the strength and direction of a linear relationship between two variables.

Correlation coefficient values range from -1 to 1. 1 means a perfect positive relationship, -1 means a perfect negative relationship, and 0 means no linear relationship.

There are very strong correlations (close to 1) between INDEXOPEN, INDEXHIGH, INDEXLOW, and INDEXCLOSE. This indicates that these indicator values move significantly together.

Research Methodology

The analysis reveals weaker correlations, ranging approximately from 0.44 to 0.46, between **TRADE_VOLUME** and **TRADE_VALUE** and the indicator values. While these correlations indicate a positive relationship, they are notably weaker compared to the strong relationships observed among the indicator values themselves. Furthermore, the variable **INDEXCODE_sentiment** contains only **NaN** values, suggesting that this data is either unavailable or irrelevant for the analysis.

The correlation matrix serves as a valuable tool for identifying variables that may contribute to modeling or analyzing the behavior of an indicator. For example, the observed correlations between **TRADE_VOLUME**, **TRADE_VALUE**, and the indicator values suggest that these trading metrics could be incorporated into predictive models to improve their accuracy. Visual representations, such as heatmaps, can further enhance the interpretation of the correlation matrix by providing an intuitive depiction of the strength and direction of relationships, thereby facilitating the identification of meaningful patterns.

In summary, the correlation matrix highlights strong interrelationships among the indicator values and weaker, yet positive, relationships between trading metrics and the indicator values. These findings provide valuable insights for selecting relevant variables in predictive modeling and analyzing the behavior of financial indicators.

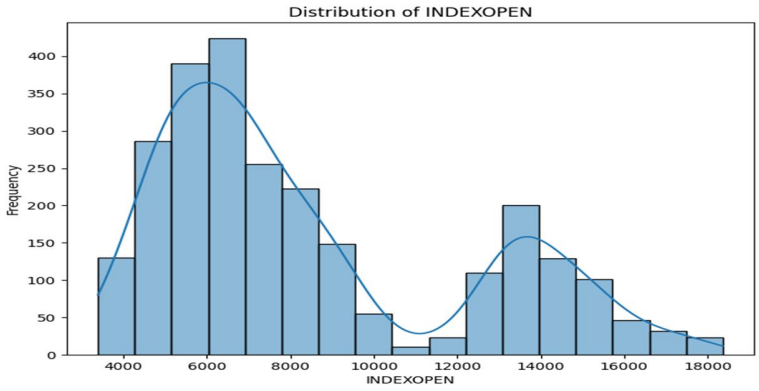

Fig. 2. Distribution of INDEXOPEN

This graph illustrates the distribution of **INDEXOPEN**, which represents the opening price of the index on the stock exchange. The horizontal axis (X-axis) corresponds to the values of **INDEXOPEN**, indicating the opening price of the index, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a specific opening price occurs in the dataset. Each bar in the graph corresponds to a range of **INDEXOPEN** values, with the height of the bar reflecting the frequency of values within that range. Additionally, a curve is overlaid on the graph, representing an estimate of the probability density distribution of the data. This curve provides insight into the overall shape and characteristics of the data distribution.

The graph reveals that the distribution of **INDEXOPEN** is multimodal, characterized by two or more distinct peaks. This suggests the presence of multiple datasets or conditions that result in different **INDEXOPEN** values. A primary peak is observed in the range of 6000–8000, while a secondary peak appears in the range of 14,000–16,000. This bimodal structure may indicate distinct time periods or varying market conditions that influenced the opening prices. Furthermore, the distribution deviates from a normal (bell-shaped) distribution, exhibiting some degree of skewness. This skewness implies the presence of outliers or anomalous values within the dataset.

These findings suggest that **INDEXOPEN** may be influenced by a variety of factors, leading to significant fluctuations in the opening prices. Further analysis is warranted to investigate the underlying causes of the multimodal distribution and to identify the factors contributing to the observed peaks. Understanding the distribution of **INDEXOPEN** can provide valuable insights for making informed investment decisions and enhancing analytical models.

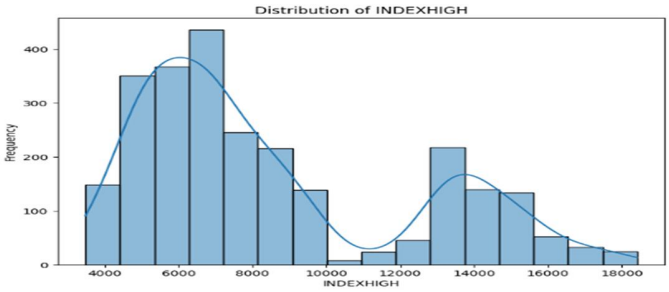

Fig. 3. Distribution of INDEXHIGH

This graph depicts the distribution of **INDEXHIGH**, which represents the highest price attained by the index on the stock exchange during a specific time period. The horizontal axis (X-axis) corresponds to the values of **INDEXHIGH**, indicating the highest price reached by the index, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a particular **INDEXHIGH** price occurs in the dataset. Each bar in the graph represents a range of **INDEXHIGH** values, with the height of the bar reflecting the frequency of values within that range. Additionally, a curve is overlaid on the graph, representing an estimate of the probability density distribution of the data. This curve provides insight into the overall shape and characteristics of the data distribution.

The graph reveals that the distribution of **INDEXHIGH** is multimodal, characterized by two or more distinct peaks. This suggests the presence of multiple datasets or conditions that result in different **INDEXHIGH** values. A primary peak is observed in the range of 6000–8000, while a secondary peak appears in the range of 14,000–16,000. This bimodal structure may indicate distinct time periods or varying market conditions that influenced the highest prices. Furthermore, the distribution deviates from a normal (bell-shaped) distribution, exhibiting some degree of skewness. This skewness implies the presence of outliers or anomalous values within the dataset.

These findings suggest that **INDEXHIGH** may be influenced by a variety of factors, leading to significant fluctuations in the highest prices. Further analysis is warranted to investigate the underlying causes of the multimodal distribution and to identify the factors contributing to the observed peaks. Understanding the distribution of **INDEXHIGH** can provide valuable insights for making informed investment decisions and enhancing analytical models.

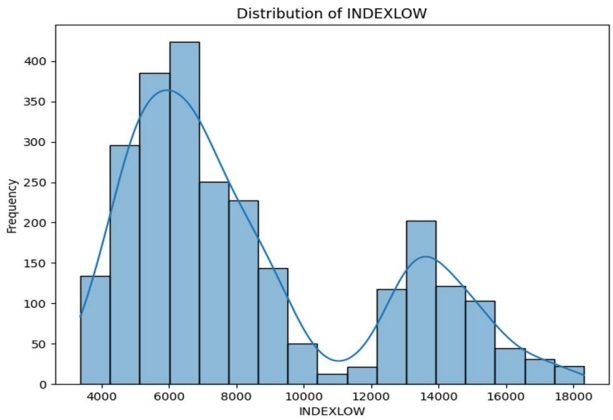

Fig. 4. Distribution of INDEXLOW

This chart illustrates the distribution of **INDEXLOW**, which represents the lowest price attained by the index on the stock exchange during a specific time period. The horizontal axis (X-axis) corresponds to the values of **INDEXLOW**, indicating the lowest price reached by the index, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a particular **INDEXLOW** price occurs in the dataset. Each bar in the chart represents a range of **INDEXLOW** values, with the height of the bar reflecting the frequency of values within that range. Additionally, a curve is overlaid on the chart, representing an estimate of the probability density distribution of the data. This curve provides insight into the overall shape and characteristics of the data distribution.

The chart reveals that the distribution of **INDEXLOW** is multimodal, characterized by two or more distinct peaks. This suggests the presence of multiple datasets or conditions that result in different **INDEXLOW** values. A primary peak is observed in the range of 6000–8000, while a secondary peak appears in the range of 14,000–16,000. This bimodal structure may indicate distinct time periods or varying market conditions that influenced the lowest prices. Furthermore, the distribution deviates from a normal (bell-shaped) distribution, exhibiting some degree of skewness. This skewness implies the presence of outliers or anomalous values within the dataset.

These findings suggest that **INDEXLOW** may be influenced by a variety of factors, leading to significant fluctuations in the lowest prices. Further analysis is warranted to investigate the underlying causes of the multimodal distribution and to identify the factors contributing to the observed peaks. Understanding the distribution of **INDEXLOW** can provide valuable insights for making informed investment decisions and enhancing analytical models.

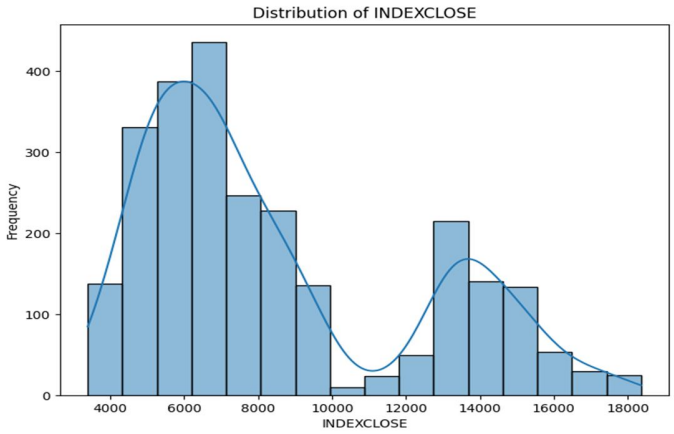

Fig. 5. Distribution of INDEXCLOSE

This graph illustrates the distribution of **INDEXCLOSE**, which represents the closing price of the index on the stock exchange during a specific time period. The horizontal axis (X-axis) corresponds to the values of **INDEXCLOSE**, indicating the closing price of the index, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a particular **INDEXCLOSE** price occurs in the dataset. Each bar in the graph represents a range of **INDEXCLOSE** values, with the height of the bar reflecting the frequency of values within that range. Additionally, a curve is overlaid on the graph, representing an estimate of the probability density distribution of the data. This curve provides insight into the overall shape and characteristics of the data distribution.

The graph reveals that the distribution of **INDEXCLOSE** is multimodal, characterized by two or more distinct peaks. This suggests the presence of multiple datasets or conditions that result in different **INDEXCLOSE** values. A primary peak is observed in the range of 6000–8000, while a secondary peak appears in the range of 14,000–16,000. This bimodal structure may indicate distinct time periods or varying market conditions that influenced the closing prices. Furthermore, the distribution deviates from a normal (bell-shaped) distribution, exhibiting some degree of skewness. This skewness implies the presence of outliers or anomalous values within the dataset.

These findings suggest that **INDEXCLOSE** may be influenced by a variety of factors, leading to significant fluctuations in the closing prices. Further analysis is warranted to investigate the underlying causes of the multimodal distribution and to identify the factors contributing to the observed peaks. Understanding the distribution of **INDEXCLOSE** can provide valuable insights for making informed investment decisions and enhancing analytical models.

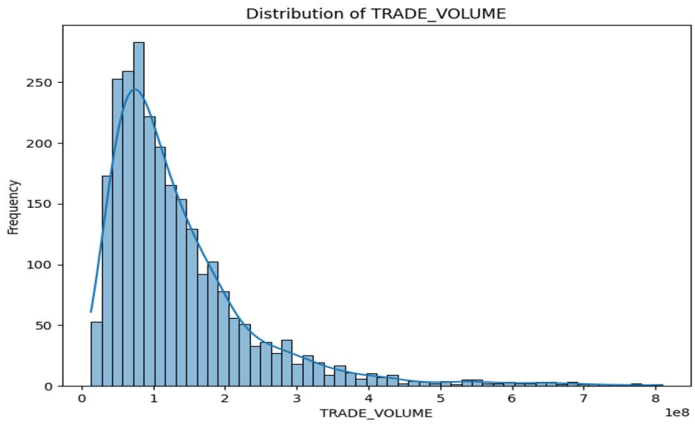

Fig. 6. Distribution of TRADE_VOLUME

This graph illustrates the distribution of **TRADE_VOLUME**, which represents the trading volume on the stock exchange over a specific time period. The horizontal axis (X-axis) corresponds to the values of **TRADE_VOLUME**, indicating the trading volume, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a particular trading volume occurs in the dataset. Each bar in the graph represents a range of **TRADE_VOLUME** values, with the height of the bar reflecting the frequency of values within that range. Additionally, a curve is overlaid on the graph, representing an estimate of the probability density distribution of the data. This curve provides insight into the overall shape and characteristics of the data distribution.

The graph reveals that the distribution of **TRADE_VOLUME** is right-skewed, indicating that a few instances of very high trading volume are present, while the majority of values are relatively low. A prominent peak is observed at the left end of the graph, corresponding to smaller trading volumes. As the trading volume increases, the frequency of values decreases significantly, suggesting that large trading volumes are less common.

These findings suggest that trading volume on the stock exchange is highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of large trading volume values may correspond to specific periods of intense trading activity, such as during significant price fluctuations or market events. Further analysis could help identify the underlying causes of these variations and their implications for market behavior.

Smaller trading volume values may indicate quieter time periods, where trading is less active.

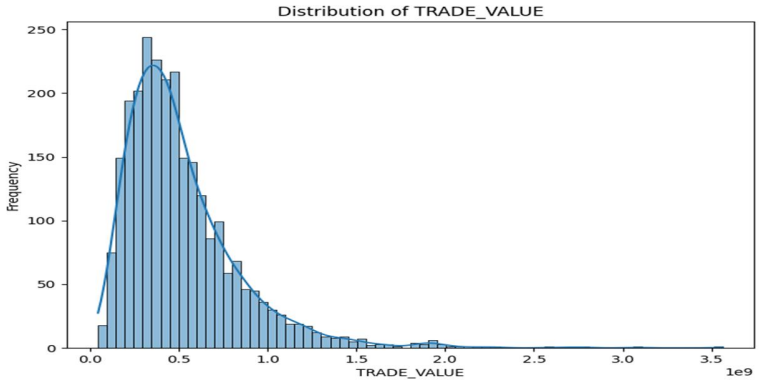

Fig. 7. Distribution of TRADE_VALUE

This graph illustrates the distribution of **TRADE_VALUE**, which represents the trading value on the stock exchange over a specific time period. The horizontal axis (X-axis) corresponds to the values of **TRADE_VALUE**, indicating the trading value, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a particular trading value occurs in the dataset. Each bar in the graph represents a range of **TRADE_VALUE** values, with the height of the bar reflecting the frequency of values within that range. Additionally, a curve is overlaid on the graph, representing an estimate of the probability density distribution of the data. This curve provides insight into the overall shape and characteristics of the data distribution.

The graph reveals that the distribution of **TRADE_VALUE** is right-skewed, indicating that a few instances of very high trading values are present, while the majority of values are relatively low. A prominent peak is observed at the left end of the graph, corresponding to smaller trading values. As the trading value increases, the frequency of values decreases significantly, suggesting that large trading values are less common.

These findings suggest that trading value on the stock exchange is highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of large trading values may correspond to specific periods of intense trading activity, such as during significant price fluctuations or market events. Conversely, smaller trading values may indicate quieter periods with less trading activity. Further analysis could help identify the underlying causes of these variations and their implications for market behavior.

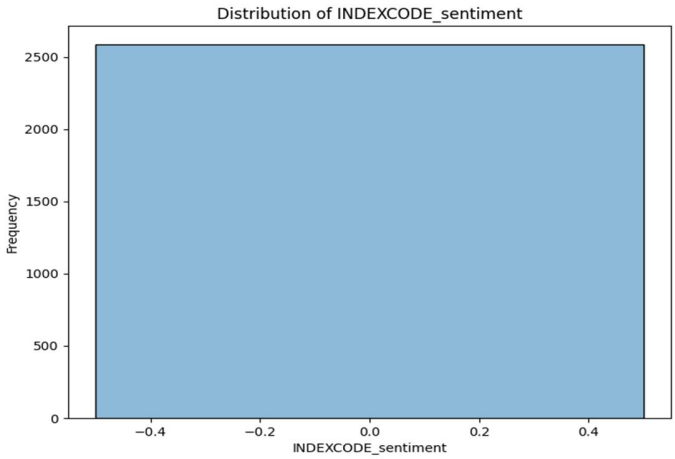

Fig. 8. Distribution of INDEXCODE _sentiment

This chart illustrates the distribution of **INDEXCODE_sentiment**, which represents traders' sentiment toward the performance of the index on the stock exchange. The horizontal axis (X-axis) corresponds to the values of **INDEXCODE_sentiment**, indicating the level of traders' sentiment, while the vertical axis (Y-axis) represents the frequency of these values, i.e., the number of times a particular sentiment value occurs in the dataset. Each bar in the chart represents a range of **INDEXCODE_sentiment** values, with the height of the bar reflecting the frequency of values within that range.

The chart reveals that the distribution of **INDEXCODE_sentiment** is nearly uniform, meaning that all sentiment values occur with similar frequency. There are no distinct peaks or trends in the distribution, indicating that traders' sentiment is evenly spread across the available range. This uniformity suggests a balanced sentiment among traders, with no clear bias toward positive or negative outlooks.

These findings imply that traders' sentiment on the stock exchange may be relatively neutral, without significant polarization. Further analysis could help uncover the reasons behind this uniform distribution, such as market stability or the absence of extreme economic or political events during the observed period. Understanding the distribution of **INDEXCODE_sentiment** can provide valuable insights for making informed investment decisions and enhancing analytical models.

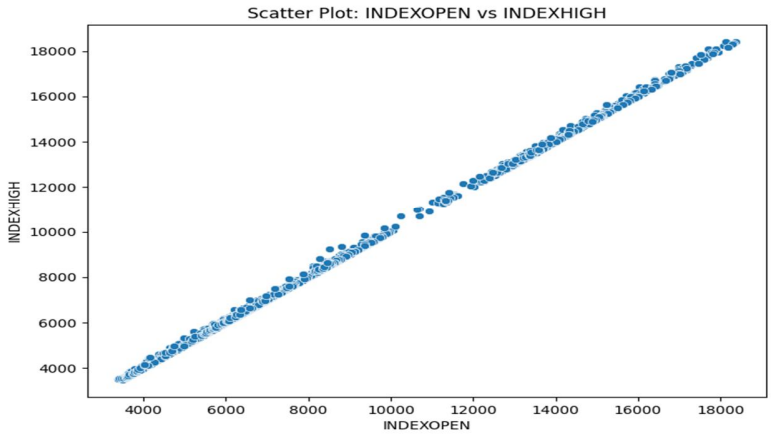

Fig. 9. Relationship between INDEXOPEN and INDEXHIGH

This chart illustrates the relationship between **INDEXOPEN**, which represents the opening price of the index, and **INDEXHIGH**, which represents the highest price the index reached during a specific time period. The horizontal axis (X-axis) corresponds to the values of **INDEXOPEN**, while the vertical axis (Y-axis) represents the values of **INDEXHIGH**. Each point on the chart represents a pair of **INDEXOPEN** and **INDEXHIGH** values for a given time period.

The chart reveals a strong positive relationship between **INDEXOPEN** and **INDEXHIGH**, indicating that higher opening prices are associated with higher maximum prices during the day. The points are distributed in a nearly linear pattern, suggesting a strong linear correlation between the two variables. This linearity implies that **INDEXOPEN** can serve as a reliable predictor of **INDEXHIGH**.

These findings suggest that the opening price of the index can be used to estimate the highest price it may reach during the trading day. Understanding this relationship can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore additional factors influencing this relationship and refine predictive accuracy.

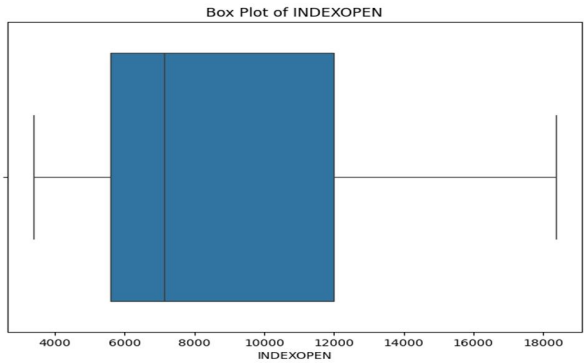

Fig. 10. Chart shows the distribution of the INDEXOPEN

This chart illustrates the distribution of **INDEXOPEN**, the opening price of the index on the stock exchange, using a **Box and Whisker Plot**. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **INDEXOPEN**, while the vertical axis (Y-axis) provides the scale for the distribution. The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **INDEXOPEN** is right-skewed, indicating that a few instances of very high opening prices are present, while the majority of values are relatively low. The extended length of the box suggests significant variability in **INDEXOPEN** values. Additionally, several outliers are observed at the right end of the chart, representing extreme high values for the opening price. These outliers may correspond to periods of significant market volatility or unusual trading activity.

These findings suggest that the opening price of an index on the stock exchange can be highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of extreme high values may indicate specific time periods characterized by heightened market activity or volatility. Understanding the distribution of **INDEXOPEN** can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore the underlying causes of these extreme values and their implications for market behavior.

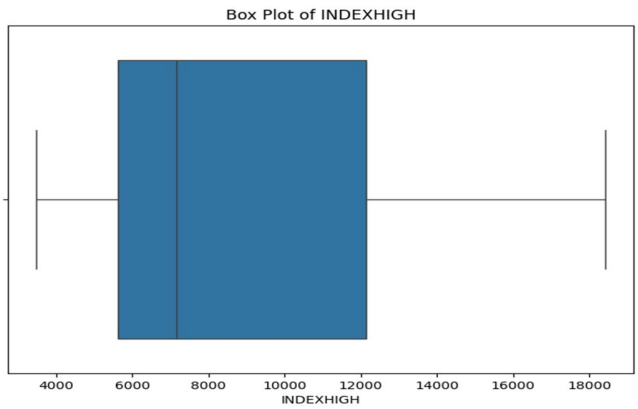

Fig. 11. Chart shows the distribution of INDEXHIGH

This chart illustrates the distribution of **INDEXHIGH**, the highest price reached by the index on the stock exchange during a specific time period, using a **Box and Whisker**

Plot. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **INDEXHIGH**, while the vertical axis (Y-axis) provides the scale for the distribution. The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **INDEXHIGH** is right-skewed, indicating that a few instances of very high maximum prices are present, while the majority of values are relatively low. The extended length of the box suggests significant variability in **INDEXHIGH** values. Additionally, several outliers are observed at the right end of the chart, representing extreme high values for the maximum price. These outliers may correspond to periods of significant market volatility or unusual trading activity.

These findings suggest that the highest price of an index on the stock exchange can be highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of extreme high values may indicate specific time periods characterized by heightened market activity or volatility. Understanding the distribution of **INDEXHIGH** can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore the underlying causes of these extreme values and their implications for market behavior.

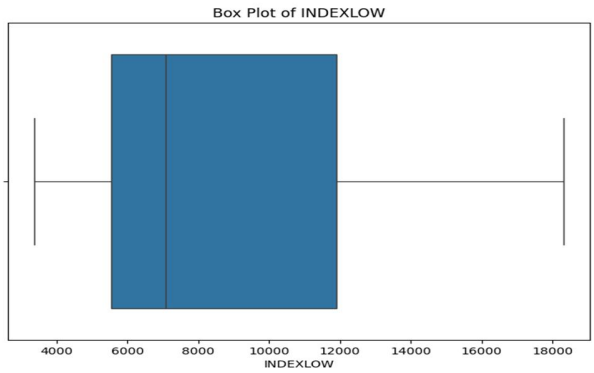

Fig. 12. Chart shows the distribution of INDEXLOW

This chart illustrates the distribution of **INDEXLOW**, the lowest price reached by the index during a specific time period, using a **Box and Whisker Plot**. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **INDEXLOW**, while the vertical axis (Y-axis) provides the scale for the distribution.

The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **INDEXLOW** is right-skewed, indicating that a few instances of very low prices are present, while the majority of values are relatively higher. The extended length of the box suggests significant variability in **INDEXLOW** values. Additionally, several outliers are observed at the right end of the chart, representing extreme low values for the index. These outliers may correspond to periods of significant market volatility or unusual trading activity.

These findings suggest that the lowest price of an index on the stock exchange can be highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of extreme low values may indicate specific time periods characterized by heightened market activity or volatility. Understanding the distribution of **INDEXLOW** can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore the underlying causes of these extreme values and their implications for market behavior.

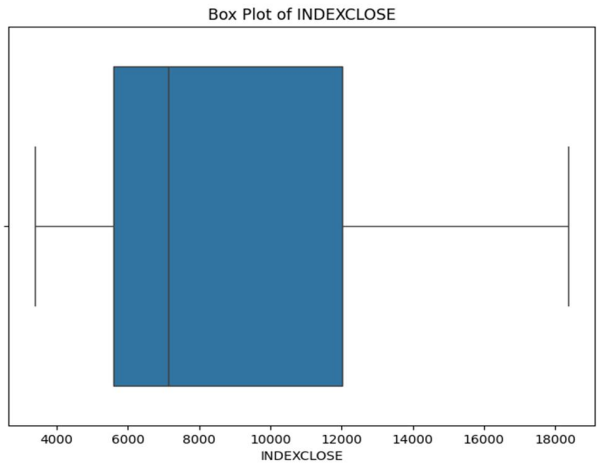

Fig. 13. Chart shows the distribution of the INDEXCLOSE

This chart illustrates the distribution of **INDEXCLOSE**, the closing price of the index on the stock exchange over a specific time period, using a **Box and Whisker Plot**. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **INDEXCLOSE**, while the vertical axis (Y-axis) provides the scale for the distribution. The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of

the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **INDEXCLOSE** is right-skewed, indicating that a few instances of very high closing prices are present, while the majority of values are relatively lower. The extended length of the box suggests significant variability in **INDEXCLOSE** values. Additionally, several outliers are observed at the right end of the chart, representing extreme high values for the closing price. These outliers may correspond to periods of significant market volatility or unusual trading activity.

These findings suggest that the closing price of an index on the stock exchange can be highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of extreme high values may indicate specific time periods characterized by heightened market activity or volatility. Understanding the distribution of **INDEXCLOSE** can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore the underlying causes of these extreme values and their implications for market behavior.

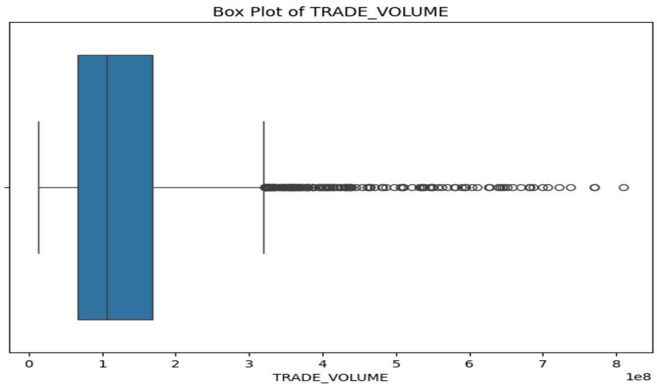

Fig. 14. Chart shows the distribution of TRADE_VOLUME

This chart illustrates the distribution of **TRADE_VOLUME**, the trading volume on the stock exchange over a specific time period, using a **Box and Whisker Plot**. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **TRADE_VOLUME**, while the vertical axis (Y-axis) provides the scale for the distribution.

The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **TRADE_VOLUME** is right-skewed, indicating that a few instances of very high trading volumes are present, while the majority of values are relatively lower. The extended length of the box suggests significant variability in **TRADE_VOLUME** values. Additionally, several outliers are observed at the right end of the chart, representing extreme high values for trading volume. These outliers may correspond to periods of significant market volatility or unusual trading activity.

These findings suggest that trading volume on the stock exchange can be highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of extreme high values may indicate specific time periods characterized by heightened market activity or volatility. Understanding the distribution of **TRADE_VOLUME** can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore the underlying causes of these extreme values and their implications for market behavior.

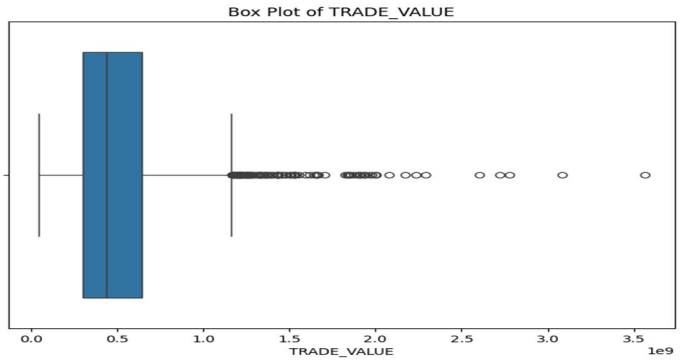

Fig. 15. Chart shows the distribution of TRADE_VALUE

This chart illustrates the distribution of **TRADE_VALUE**, the value of trading on the stock exchange over a specific time period, using a **Box and Whisker Plot**. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **TRADE_VALUE**, while the vertical axis (Y-axis) provides the scale for the distribution. The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **TRADE_VALUE** is right-skewed, indicating that a few instances of very high trading values are present, while the majority of values are relatively lower. The extended length of the box suggests significant variability in **TRADE_VALUE** values. Additionally,

several outliers are observed at the right end of the chart, representing extreme high values for trading value. These outliers may correspond to periods of significant market volatility or unusual trading activity.

These findings suggest that the trading value on the stock exchange can be highly variable and influenced by factors such as economic events, political developments, and investor sentiment. The presence of extreme high values may indicate specific time periods characterized by heightened market activity or volatility. Understanding the distribution of **TRADE_VALUE** can provide valuable insights for making informed investment decisions and improving analytical models. Further analysis could explore the underlying causes of these extreme values and their implications for market behavior.

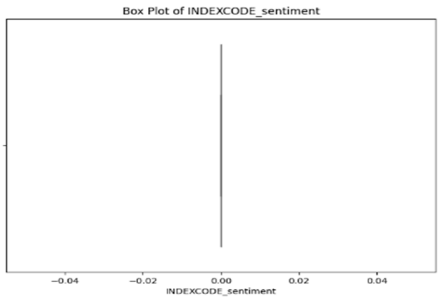

Fig. 16. Chart shows the distribution of INDEXCODE _sentiment

This chart illustrates the distribution of **INDEXCODE_sentiment**, which represents traders' sentiment toward the performance of the index on the stock exchange, using a **Box and Whisker Plot**. This type of plot provides a statistical summary of the data, highlighting key features such as central tendency, variability, and outliers. The horizontal axis (X-axis) represents the values of **INDEXCODE_sentiment**, while the vertical axis (Y-axis) provides the scale for the distribution. The **box** in the plot represents the interquartile range (IQR), which contains the middle 50% of the data, spanning from the first quartile (Q1) to the third quartile (Q3). The vertical line inside the box indicates the **median**, the middle value of the dataset. The **whiskers** extend from the box to the furthest data points within 1.5 times the IQR, and any data points beyond the whiskers are considered **outliers**, representing extreme values.

The chart reveals that the distribution of **INDEXCODE_sentiment** is approximately uniform, meaning that all sentiment values occur with similar frequency. The small size of the box indicates that the data points are closely clustered, suggesting low variability in traders' sentiment. Additionally, there are no outliers in the chart, indicating the absence of extreme sentiment values. This uniformity suggests a balanced sentiment among traders, with no strong bias toward positive or negative outlooks.

These findings imply that traders' sentiment on the stock exchange may be relatively neutral, without significant polarization. Further analysis could help uncover the reasons behind this uniform distribution, such as market stability or the absence of extreme economic or

political events during the observed period. Understanding the distribution of **INDEXCODE_sentiment** can provide valuable insights for making informed investment decisions and enhancing analytical models.

Results and Analyses

1. INDEXOPEN Distribution: The distribution of INDEXOPEN is multimodal, with two distinct peaks observed around 6000–8000 and 14,000–16,000. This suggests the presence of two or more datasets or distinct time periods/circumstances that led to different opening prices. The multimodal nature indicates significant variability in opening prices, potentially influenced by external factors such as market conditions or economic events.
2. INDEXHIGH Distribution: Similarly, the distribution of INDEXHIGH is multimodal, with peaks around 6000–8000 and 14,000–16,000. This indicates that the highest prices reached by the index vary significantly, likely due to different market conditions or time periods. The presence of multiple peaks suggests that the index experiences distinct phases of high-price activity.
3. INDEXLOW Distribution: The distribution of INDEXLOW is also multimodal, with peaks around 6000–8000 and 14,000–16,000. This implies that the lowest prices of the index vary across different time periods or circumstances. The distribution is skewed, indicating the presence of outliers or unusual values, which may reflect periods of extreme market volatility.
4. INDEXCLOSE Distribution: The distribution of INDEXCLOSE is multimodal, with peaks around 6000–8000 and 14,000–16,000. This suggests that the closing prices of the index vary significantly, potentially due to different market conditions or external influences. The multimodal nature highlights the variability in closing prices over time.
5. TRADE_VOLUME Distribution: The distribution of TRADE_VOLUME is right-skewed, with a clear peak at the lower end of the trading volume range. This indicates that most trading volumes are relatively small, with a few instances of very high trading volumes. The skewness suggests that periods of intense trading activity are less common but significant when they occur.
6. TRADE_VALUE Distribution: The distribution of TRADE_VALUE is also right-skewed, with a peak at the lower end of the trading value range. This indicates that most trading values are relatively small, with a few instances of very high trading values. The skewness suggests that high trading values are less frequent but may correspond to periods of significant market activity.
7. INDEXCODE_sentiment Distribution: The distribution of INDEXCODE_sentiment is approximately uniform, with no clear peaks or extreme values. This suggests that traders' sentiment is evenly distributed across the range, indicating a

balanced sentiment without strong positive or negative biases. The small interquartile range (IQR) indicates that the data points are closely clustered, further supporting the uniformity of sentiment.

8. Relationship Between INDEXOPEN and INDEXHIGH: There is a strong positive linear relationship between INDEXOPEN and INDEXHIGH, as evidenced by the nearly straight-line distribution of points in the scatter plot. This indicates that higher opening prices are strongly associated with higher maximum prices during the trading day. This relationship can be leveraged for predictive modeling and decision-making.
9. INDEXCODE_sentiment Uniformity: The uniform distribution of INDEXCODE_sentiment suggests that traders' sentiment is consistent and evenly spread across the observed range. The absence of outliers or extreme values indicates a stable sentiment pattern, which may reflect balanced market conditions or the absence of significant external shocks.

Key Conclusions:

- The multimodal distributions of INDEXOPEN, INDEXHIGH, INDEXLOW, and INDEXCLOSE suggest that the index experiences distinct phases of activity, likely influenced by varying market conditions or external factors.
- The right-skewed distributions of TRADE_VOLUME and TRADE_VALUE indicate that periods of intense trading activity are less common but significant when they occur.
- The uniform distribution of INDEXCODE_sentiment reflects balanced trader sentiment, with no strong positive or negative biases.
- The strong linear relationship between INDEXOPEN and INDEXHIGH highlights the predictability of maximum prices based on opening prices, which can be useful for modeling and decision-making.

These findings provide valuable insights into the behavior of the index and trading activity, which can inform investment strategies and analytical models. Further analysis could explore the underlying causes of the observed patterns and their implications for market behavior.

Conclusion

This research paper investigates the Egyptian stock market, focusing on forecasting missing data and analyzing the behavior of the EGX 30 index over the past three years to mitigate the impact of periods of instability. The study examines the dynamic composition of the index, which includes many stocks that are frequently added or removed. To address the high dimensionality of the data, Principal Component Analysis (PCA) is employed. The results demonstrate that the first three principal

components (PCs) account for 83% of the total variance in the data, effectively capturing the most significant patterns and trends.

A Principal Component Regression (PCR) model is developed to predict missing data for the EGX 30 index. PCR is a regression technique applied to a reduced set of variables derived from PCA, ensuring efficient and accurate modeling. The cross-validation (CV) results of the PCR model highlight the importance of analyzing trends in key indicators such as INDEXOPEN, INDEXHIGH, INDEX LOW, and INDEXCLOSE over time. These indicators provide critical insights into the overall performance of the index.

The correlation analysis reveals strong relationships between these indicators. The correlation coefficients, which range from -1 to 1, indicate the strength and direction of these relationships. A value of 1 signifies a perfect positive relationship, -1 signifies a perfect negative relationship, and 0 indicates no linear relationship. The analysis shows very strong positive correlations (close to 1) between INDEXOPEN, INDEXHIGH, I NDEXLOW, and INDEXCLOSE, suggesting that these indicators move in tandem significantly. This strong interdependence underscores the consistency in the behavior of these metrics and their collective influence on the index's performance.

In summary, this study demonstrates the effectiveness of PCA and PCR in reducing data dimensionality and predicting missing values in the EGX 30 index. The strong correlations between key indicators highlight their importance in understanding market trends and making informed investment decisions. These findings provide a robust foundation for further research and practical applications in financial analysis and decision-making.

Declarations

Author contribution statement

All authors listed have significantly contributed to the development and the writing of this article and all authors are participated equally in this research paper.

Data availability statement

Data will be available on request by contacting the corresponding author

Declaration of interest's statement

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1- Heba Elsegai1, Hanem S. H. M. El-Metwally & Hisham M. Almongy. (2024) Predicting the Trends of the Egyptian Stock Market Using Machine Learning and Deep Learning Methods, Computational Journal of Mathematical and Statistical Sciences, 4(1), 186–221.

2- Vadlamudi, S. (2017). Stock market prediction using machine learning: a systematic literature. review. American Journal of Trade and Policy, 4(3), 123-128.

3- Ghorbani, M., and E. Chong. 2020. "Stock price prediction using principal component." PloS one 15 (3): 1 - 20. doi:https://doi.org/10.1371/journal.pone.0230124.

4- Hargreaves, C. A. 2019. "An automated stock investment system using machine learning techniques: An application in Australia." World Academy of Science, Engineering and Technology International Journal of Mathematical and Computational Sciences 13 (10): 189 - 192.

5- Jolliffe, I. T. 2002. Principal Component Analysis, Second Edition. New York: Springer.

6- Zhong, X., and D. Enke. 2019. "Predicting the daily return direction of the stock market using hybrid machine learning algorithms." Financ Innov 24 (5): 1 - 20. doi:https://doi.org/10.1186/s40854-019-0138-0.

7- Susianto, Y and Notodiputro, KA and Kurnia, A and Wijayanto, H, ”A Comparative Study of Imputation Methods for Estimation of Missing Values of Per Capita Expenditure in Central Java” in

IOP Conference Series: Earth and Environmental Science, 58(1),012017, 2017.

8- Allassonniere, St´ephanie and Kuhn, Estelle, Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation, Computational Statistics & Data Analysis, 91, 4–19, 2015.

9- Vachhani, H., Obiadat, M. S., Thakkar, A., Shah, V., Sojitra, R., Bhatia, J., & Tanwar, S. (2020).Machine learning based stock market analysis: A short survey. In Innovative Data Communication Technologies and Application: ICIDCA 2019 (pp. 12-26). Springer International Publishing.

10- Soni, P., Tewari, Y., & Krishnan, D. (2022). Machine learning approaches in stock price prediction:a systematic review. In Journal of Physics: Conference Series (Vol. 2161, No. 1, p. 012065). IOP Publishing.

11- Mehtab, S., Sen, J., & Dutta, A. (2021). Stock price prediction using machine learning and LSTMbased deep learning models. In Machine Learning and Metaheuristics Algorithms, and Applications: Second Symposium, SoMMA 2020, Chennai, India, October 14–17, 2020, Revised Selected Papers 2 (pp. 88-106). Springer Singapore.

12- Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. Ieee Access, 8, 150199-150212.

13- Demirel, U., C, am, H., & U¨ nlu¨, R. (2021). Predicting stock prices using machine learning methods and deep learning algorithms: The sample of the Istanbul Stock Exchange. Gazi University Journal of Science, 34(1), 63-82.

14- Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685-695.

15- Nabipour, M., Nayyeri, P., Jabani, H., Shahab, S., & Mosavi, A. (2020). Predicting stock market trends using machine learning and deep learning algorithms via continuous and binary data; a comparative analysis. Ieee Access, 8, 150199-150212.

16- Gavankar, Sachin and Sawarkar, Sudhirkumar (2015), Decision Tree: Review of Techniques for

Missing Values at Training, Testing and Compatibility, Artificial Intelligence, Modelling and Simulation (AIMS), 2015 3rd International Conference on, 122–126,

17- Gerardnico (2017), Data Mining - K-Nearest Neigbors, CC Attribution-Noncommercial-Share Alike 4.0 International,.

18- Beretta, Lorenzo and Santaniello (2016), Alessandro, Nearest neighbor imputation algorithms: a critical evaluation, BMC medical informatics and decision making, 16(3):74,.

19- Kenward, M. G.(2013), The handling of missing data in clinical trials,Clinical Investigation, 3(3):241–250,.

20- Liu, Yuzhe and Gopalakrishnan (2017), Vanathi, An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data, Data, 2(1):8,