**STOCK MARKET PREDICTION USING BIG DATA ANALYTICS**

**Table of Contents**

[1. Introduction 3](#_Toc164267714)

[a. Project Scope 3](#_Toc164267715)

[b. Project Limitations & Constraints 3](#_Toc164267716)

[c. Feasibility Study [Technical | Operational] 3](#_Toc164267717)

[2. System Requirement Specifications 3](#_Toc164267718)

[3. System Design 4](#_Toc164267719)

[a. Use-Case Diagram 4](#_Toc164267720)

[b. Sequence Diagram 5](#_Toc164267721)

[4. Data Design 5](#_Toc164267722)

[a. ETL Process (Explanation) 5](#_Toc164267723)

[b. Data Management 5](#_Toc164267724)

[5. Data Analytics and Modeling 6](#_Toc164267725)

[6. Data Visualization 11](#_Toc164267726)

[7. SQL queries 16](#_Toc164267727)

[References 19](#_Toc164267728)

[Appendix 20](#_Toc164267729)

# 1. Introduction

## a. Project Scope

The project scope involves a crypto historical data analysis through PySpark with the aim of participating in the data processing and the engineering of machine learning models. The process implies achieving the same results by executing the EDA in order to get more information about the data and the trend (Jay *et al.* 2020). The MySQL Workbench as a data querying and understanding tool will make use of SQL queries mainly. This project has the goal of developing machine learning models which include Linear Regression, Random Forest Regression, and gradient-boosted trees to predict crypto prices. The visualization and plotting have been provided by libraries such as Matplotlib and Seaborn.

## b. Project Limitations & Constraints

The project has the aspects of data availability, computational resources, and the market dynamics complexity as restrictions and constraints. Acquiring the sufficient data that is historically & credibly cryptocurrency in nature with may be difficult. The processing of large datasets with PySpark is consumed by the memory resource leading to possible shortages (Khedr *et al.* 2021). Besides, this broad volatility and complexity of cryptocurrency market risk predicting prices precisely.

## c. Feasibility Study [Technical | Operational]

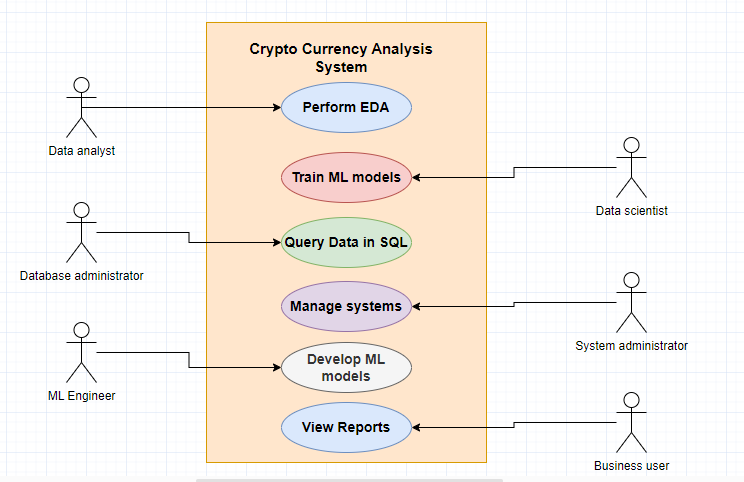
The technical aspects of the project are feasible which uses PySpark for data processing and machine learning to an extent. One of the strengths of PySpark is its ability to handle datasets of any sizes which makes it appropriate for analytics of crypto data. Furthermore, Python was used because it incorporates machine learning libraries that are used for the creation of prediction models (Patel *et al.* 2020). On the operational side, the project is feasible as it needs competent people with skills for data analysis, modeling, and visualization which are accessible. Nonetheless, considering the importance of computational resources and the complexity of cryptocurrency market, operational challenges may emerge.

# 2. System Requirement Specifications

The project involves the use of an operating system (Windows), Python for programming, PySpark for data processing, Google Colab for development, and MySQL Workbench for SQL queries. Moreover, accurate computation abilities are required for processing big data and training machine learning models (Chen, 2023). The project also needs historical cryptocurrency data for the research. As for the requirement of data storage to hold the datasets and model outcomes, it is necessary.

# 3. System Design

## a. Use-Case Diagram

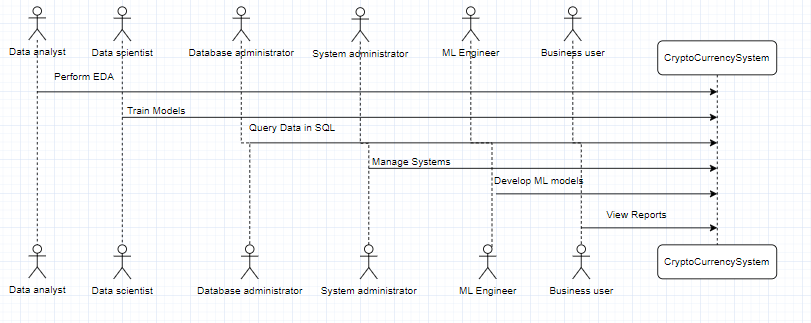


**Figure 1: Use case diagram**

(Source: Self-Created)

This use case illustrates the interaction between the actors and the Cryptocurrency Analysis System by depicting the use cases. The Data Analyst gets ahead with exploratory data analysis (EDA), the Data Scientist puts into practice machine learning models, the Database Administrator queries data in SQL, and the System Administrator takes care of the system. The duties of the Machine Learning Engineer are to build the machine learning model, and the Business User discovers the information through the reports.

## b. Sequence Diagram



**Figure 2: Sequence diagram**

(Source: Self-Created)

This diagram represents a sequence of actions that take place within the Cryptocurrency Analysis System between the actors. Every character initializes an independent action which includes descriptive data analysis, creation of machine learning models, querying the dataset with SQL, management of the system, implementation of machine learning models, or viewing the reports.

# 4. Data Design

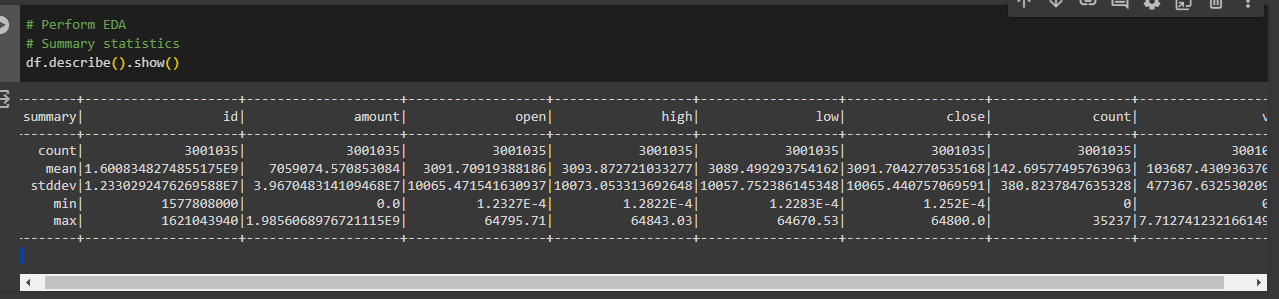
## a. ETL Process (Explanation)

This process of data transformation (Extract, Transform, Load), on the other hand, entails converting the CSV data into a form-fitting for analysis, moving it to a MySQL database, and finally storing it permanently in the database. These data are changed into some other forms and then this transformed data is ingested into Apache Spark for data processing, machine learning model development, and visualization (Hamayel and Owda, 2021). This procedure guarantees that the data is cleaned, aligned in a format that is structured for analysis, and rendered properly so that analysis and the resulting insights come in most efficiently. The data is retrieved and transformed on the basis of ETL and it is in this way that the data becomes more usable for further analysis and decision-making.

## b. Data Management

Data management implies well-structured data organizing and storage of cryptocurrency data. Data cleaning is conducted here that ensures data integrity, also the security of data is considered before being provided to users at ease. Data are stored in the database of MySQL which involves simple querying and retrieval. As for data management, the practices are used to help ensure data accuracy and consistency (Ho *et al.* 2021). It does not only cover data backup, data validation, and maintenance but also includes the prevention, detection and response to information and cyber threats. The project succeeds in managing the data well so that the data is accurate and accessible for the using for analysis, modelling, and visualization.

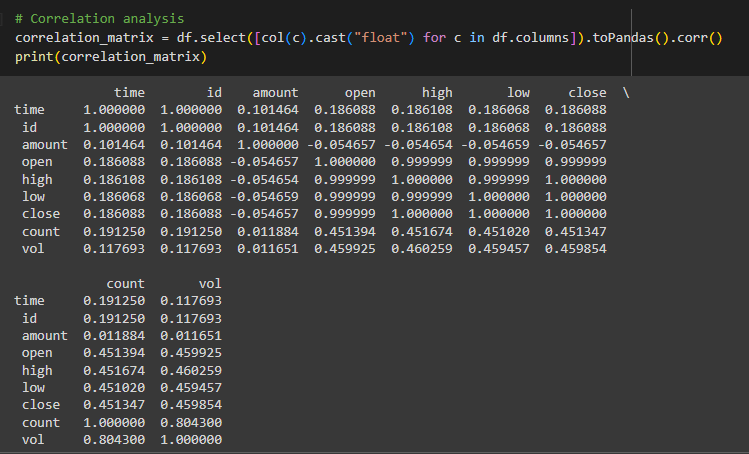
# 5. Data Analytics and Modeling



**Figure 3: Summary Statistics of the data**

(Source: Self-Created)

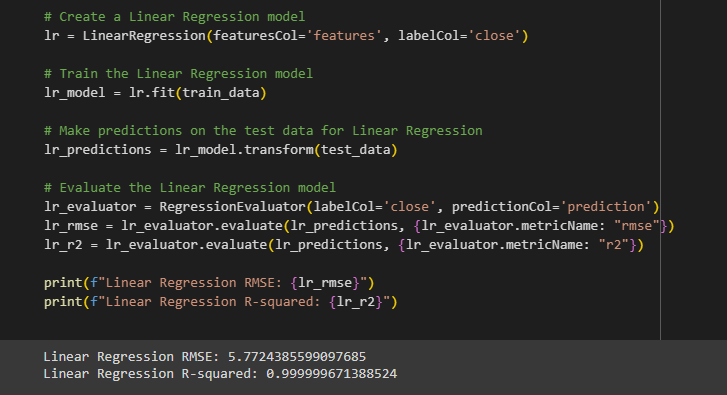
The `describe()` function gives summaries of the sample statistics of the modelling columns. It demonstrates the count, mean, standard deviation, min and max values of each column. This estimate is an important help to a better understanding of the information on the data, which is fundamental to EDA.



**Figure 4: Correlation matrix**

(Source: Self-Created)

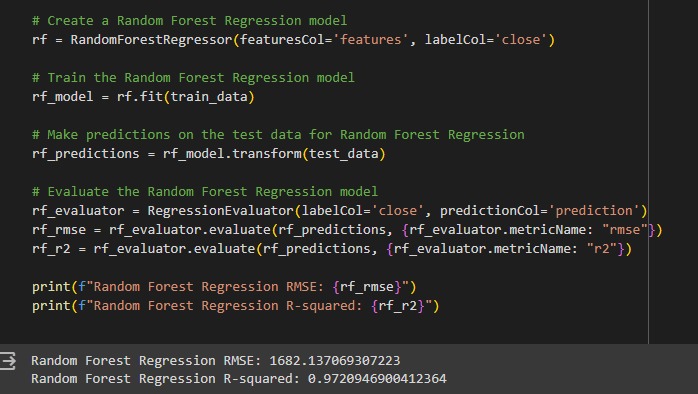
The correlation matrix demonstrates correlation coefficients between each two variables as they exist in the dataset. It provides a measurement of the strength and direction of the linear relationships that exist between variables. For instance, a powerful positive interrelation between 'open', 'high', 'close' and 'low' prices can be observed, meaning that they move parallel. There is a strong coherence in the point spread between 'count' and 'vol', this behaviour indicates the connection of deal number and volume.



**Figure 5: Linear Regression model**

(Source: Self-Created)

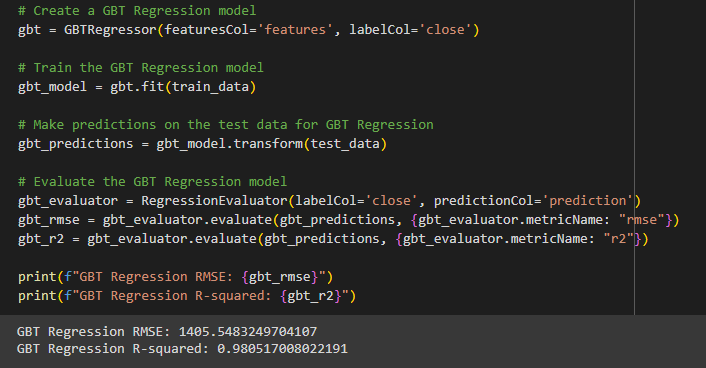
The Linear regression model had a Root mean square error (RMSE) of 5.77 and an R-squared value of 0.999, thus providing evidence that the model fits the data very well. This proves that a substantial linear connection between the trading variables such as count, amount, open, low, high, and volume and the closing price of the cryptocurrency is here to be seen.



**Figure 6: Random Forest Regression model**

(Source: Self-Created)

The Random Forest Regression model gave the Root Mean Squared Error (RMSE) of 1682.14 and the R-squared value of 0.972, that is the model is good in predicting cryptocurrency closing price, based on the performance metrics. The Random Forest technique, which involves the use of several decision trees, has achieved complex relations between the features and the dependent target by providing a good accuracy level of predictions.

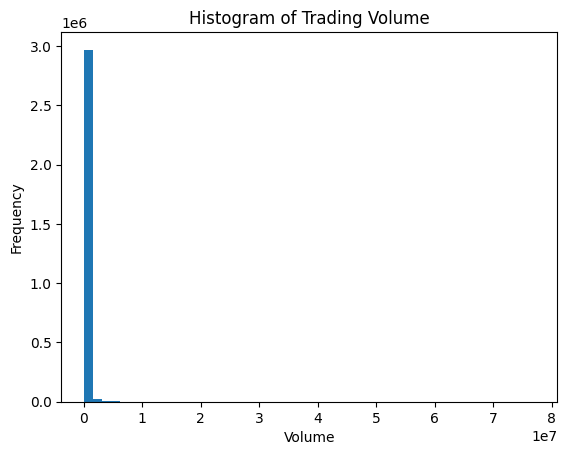


**Figure 7: GBT Regression model**

(Source: Self-Created)

The GBT Regression's RMSE score, which is 1405.55, and its R-squared value, which is 0.981, demonstrate outstanding prediction performance when forecasting the closing prices of cryptocurrencies. The GBT algorithm that uses the forward-phase approach has successfully explained the nature of data.

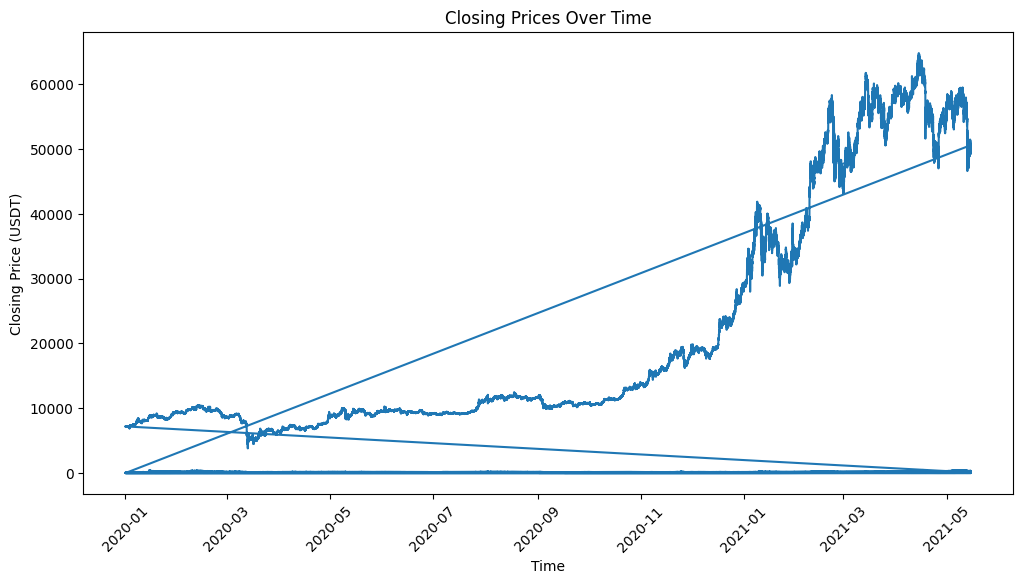
# 6. Data Visualization



**Figure 8: Histogram of trading volume**

(Source: Self-Created)

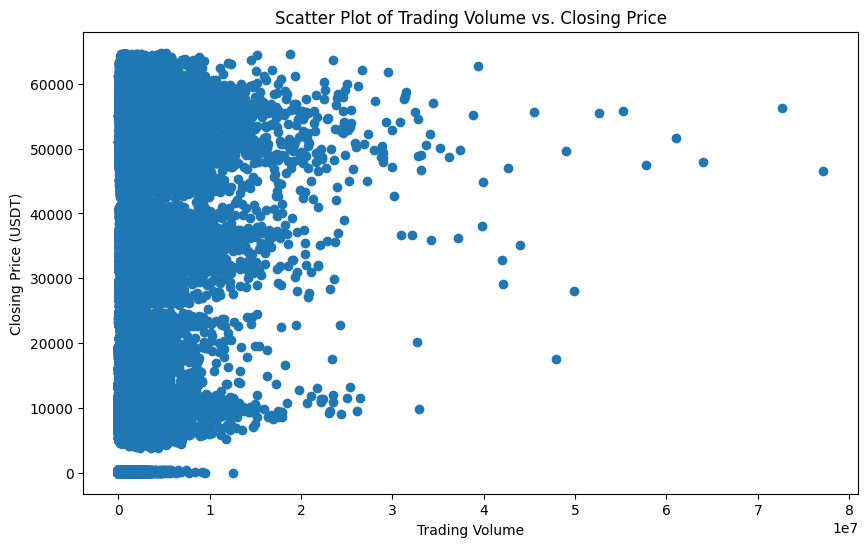
The distribution of the chart showing trading volume demonstrates that the most volume tends to occur at the lower levels, with a peak frequency between 0 and 3.0. That could mean there are a lot of small market deals with a remarkably low overall volume compared to high-volume transactions, which might have consequences on liquidity and market patterns of the cryptocurrency market.



**Figure 9: Line Plot of Closing Prices Over Time**

(Source: Self-Created)

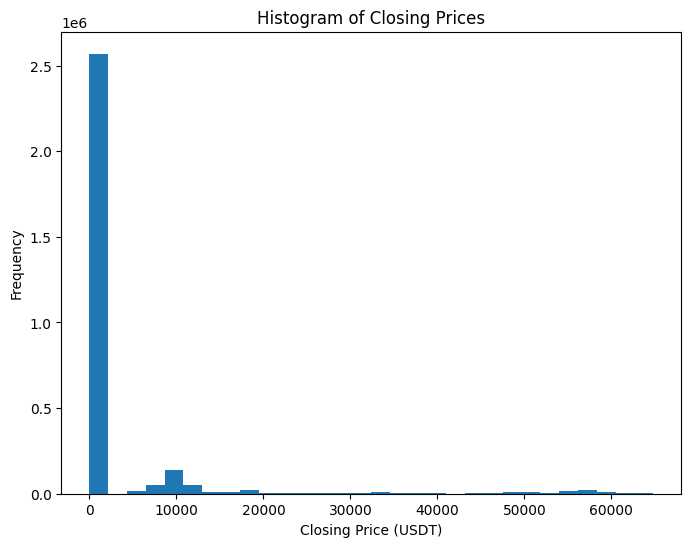
The line plot shows the price trend as time goes by, and prices went up from March 2021 to May 2021. This could very possibly reflect a major event or trend in the cryptocurrency market bearing high impact like an increased market interest or market fluctuation. Market depth could also be another angle for a more in-depth study to bring out the factors influencing the current price trend and their implications on trading strategies and market dynamics.



**Figure 10: Scatter Plot of Trading Volume vs. Closing Price**

(Source: Self-Created)

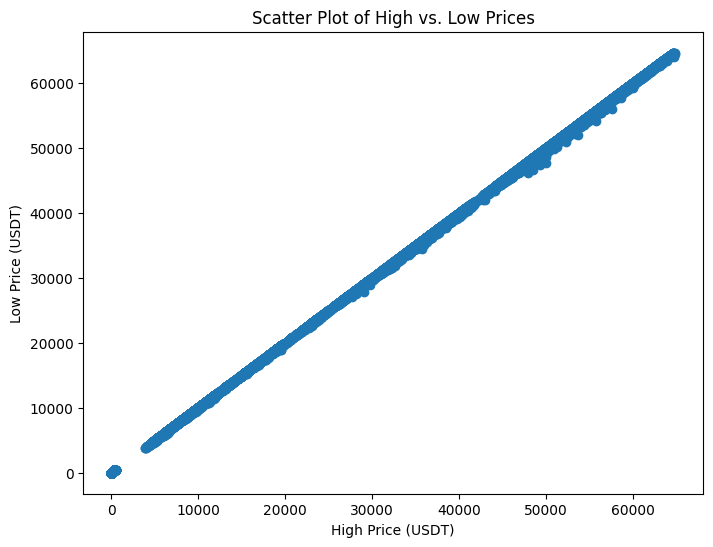
The scatter graph of trading volume against the closing price captures all the details that are important in this case. The close price presents a concentration of points on the left side, especially the trading volume from 0 to 1. Hence, more than half of the data are related to trading periods with very low trading volumes and zero difference between the opening and closing prices.



**Figure 11: Histogram of Closing Prices**

(Source: Self-Created)

The histogram of close prices shows the frequency of closing prices in the dataset. A histogram shows that the over prices lie at certain price ranges, and the most frequent price appears as a field in the histogram. These characteristics represent the prices of the cryptocurrency and can demonstrate price volatility within the selected period.

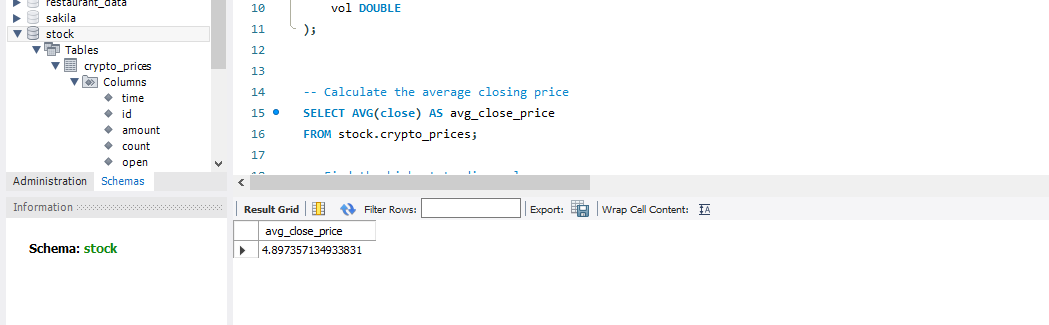


**Figure 12: Scatter Plot of High vs. Low Prices**

(Source: Self-Created)

The scatter plot of high vs low would be the main point of this section which illustrates a strong positive correlation between the quantity demanded and the low prices. These are represented by the line of the points connecting them. Therefore such a trend suggests that as the high price rises, the low price rises accordingly.

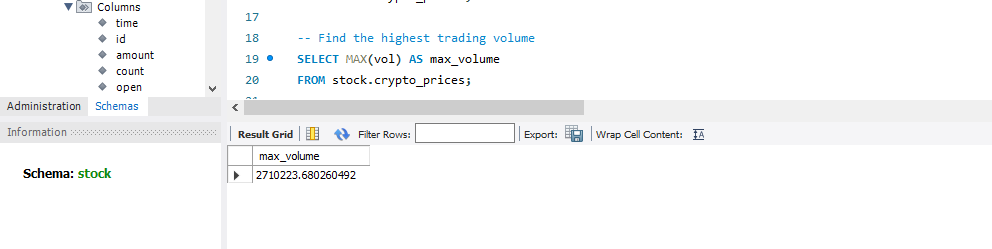
# 7. SQL queries



**Figure 13: Calculate the average closing price**

(Source: MySQL Workbench)

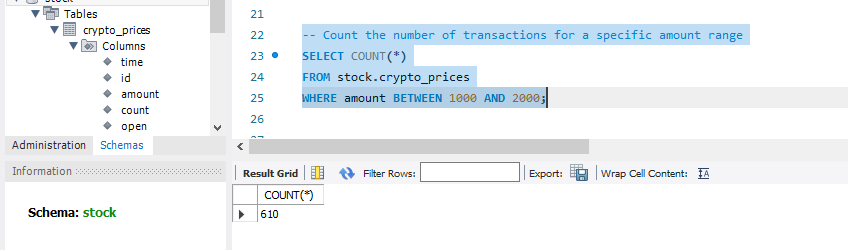
The average closing price of cryptocurrency can be seen in the data sample as 4.90 USDT. Through the closing price mean, an investor or a trader could understand the typical value of the closing prices, which is useful for figuring out the overall price trends that are indicative of informed decisions.



**Figure 14: Find the highest trading volume**

(Source: MySQL Workbench)

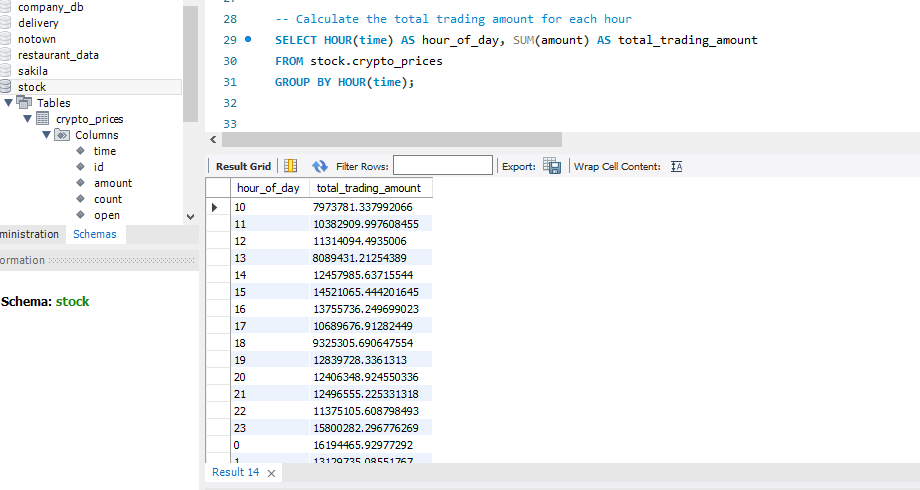
The total amount of cryptocurrency trading of 2,710,223.68 was observed with the highest volume. This figure indicates the highest trading volume of the period, which is an essential part of liquidity and the popularity rate of the market among investors.



**Figure 15: Count the number of transactions for a specific amount range**

(Source: MySQL Workbench)

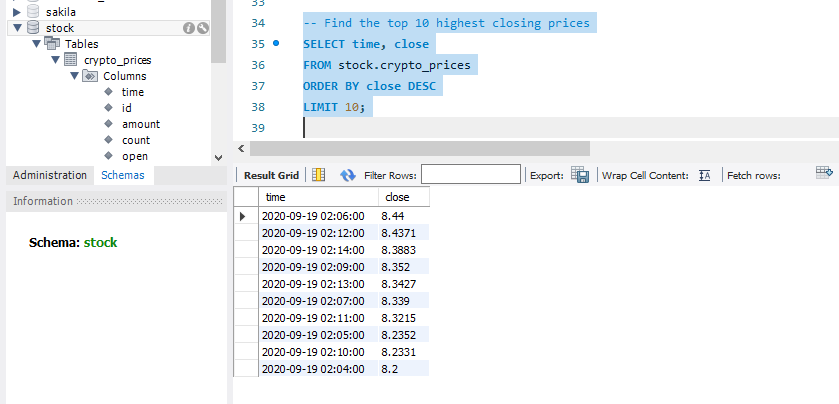
The query disclosed 610 deals with amounts between $1000 and $2000. This information assists in not only determining the pattern of these transactions but also aids in analysing the market and assessing risk.



**Figure 16: Calculate the total trading amount for each hour**

(Source: MySQL Workbench)

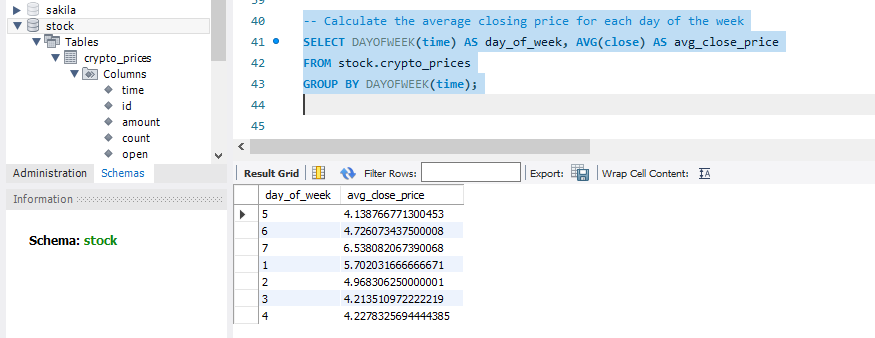
The inquiry supplies the total trading amount on an hourly basis, which enables one to get information regarding trading patterns and liquidity trends. Such data can be applied for the improvement of strategies and definition of trading hours in the peak.



**Figure 17: Find the top 10 highest closing prices**

(Source: MySQL Workbench)

The query is mentioned to be among the top 10 highest closing prices, which reveals the historic peak prices. The needed information is used to define the market trend, to find the possible outliers, and to understand pricing volatility in the cryptocurrency market.



**Figure 18: Calculate the average closing price for each day of the week**

(Source: MySQL Workbench)

The given query computes the average closing price of each day, had any patterns or trends based on the day can be figured out with that. It can be useful in discovering the day-to-day average prices thus helping in making trading decisions by allowing mid-term market analysis.

# References

Chen, J., 2023. Analysis of Bitcoin price prediction using machine learning. Journal of Risk and Financial Management, 16(1), p.51.

Hamayel, M.J. and Owda, A.Y., 2021. A novel cryptocurrency price prediction model using GRU, LSTM and bi-LSTM machine learning algorithms. Ai, 2(4), pp.477-496.

Ho, A., Vatambeti, R. and Ravichandran, S.K., 2021. Bitcoin price prediction using machine learning and artificial neural network model. Indian Journal of Science and Technology, 14(27), pp.2300-2308.

Jay, P., Kalariya, V., Parmar, P., Tanwar, S., Kumar, N. and Alazab, M., 2020. Stochastic neural networks for cryptocurrency price prediction. Ieee access, 8, pp.82804-82818.

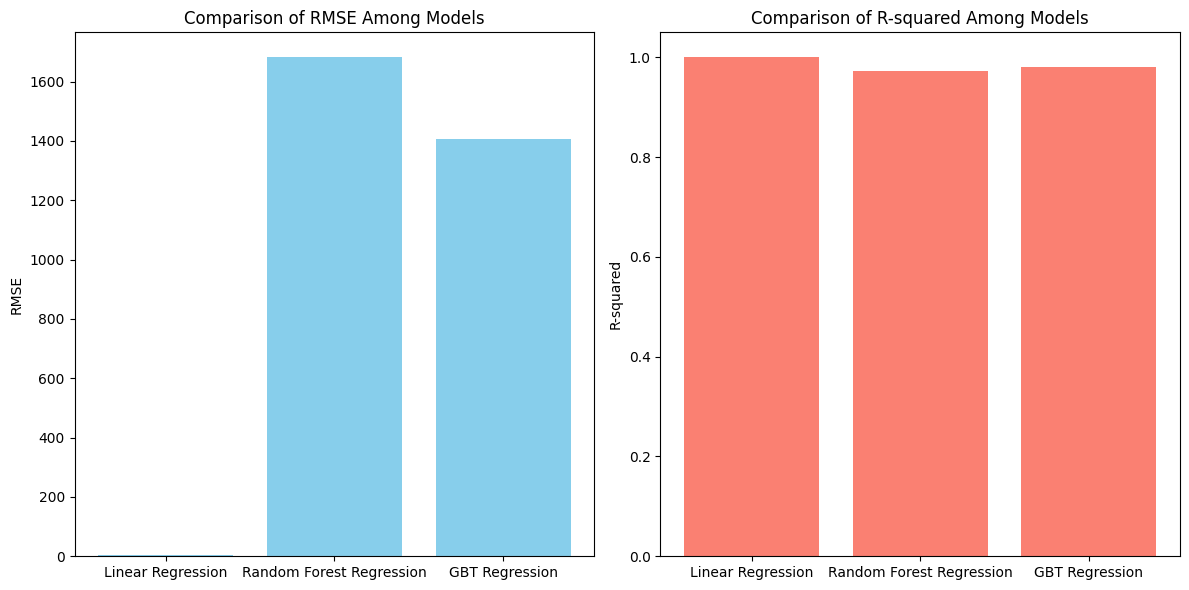
Khedr, A.M., Arif, I., El‐Bannany, M., Alhashmi, S.M. and Sreedharan, M., 2021. Cryptocurrency price prediction using traditional statistical and machine‐learning techniques: A survey. Intelligent Systems in Accounting, Finance and Management, 28(1), pp.3-34.

Patel, M.M., Tanwar, S., Gupta, R. and Kumar, N., 2020. A deep learning-based cryptocurrency price prediction scheme for financial institutions. Journal of information security and applications, 55, p.102583.

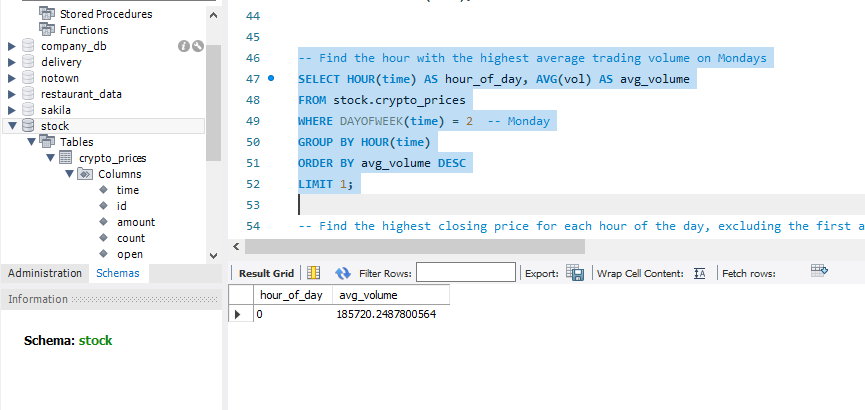
# 

# Appendix

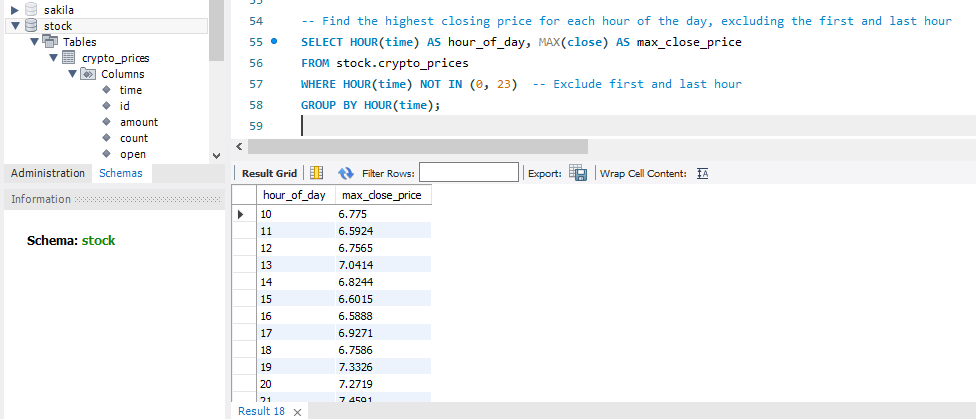
**ML Model comparison plot**



**Find the hour with the highest average trading volume on Mondays**



**Find the highest closing price for each hour of the day, excluding the first and last hour**



**Calculate the average trading amount for each hour of the day, for weekends (Saturday and Sunday) separately**

