BITCOIN Price Analysis and Predictive Model Building

Phân tích giá BITCOIN và xây dựng mô hình dự đoán

Nhóm 2 - DA63
Mentor: Trần Danh Ngân

Members

—Nhan Nhu Ngoc

—Ta Vo Quang Huy

Menu

Discovery
What is Bitcoin?

The importance of Bitcoin price prediction

01

02

Analyze
Exploratory data analysis

Summarize the insights and review

Model prediction
Build a machine learning model for prediction

03

04

Suggestion
Suggest solution for short or long term investment

01

Discovery

Business understanding
What is Bitcoin?
The importance of Bitcoin price prediction
Dataset introduction

Business Understanding

Objectives:

- Understanding Bitcoin and importance of Price Prediction
- Utilizing Machine Learning
- Insight Analysis and Review

Expected results:

- Price Prediction Models
- Trend Analysis
- Impact of Variables
- Actionable Insights

Introduction

What is Bitcoin?

Bitcoin is the very first cryptocurrency and blockchain technology introduced to the world.

BTC on Binance

Breaking news!

Vietnam ranks second in the world in cryptocurrency ownership rate

```
Ngày có giá trị Low thấp nhất là: 2018-12-15 với giá trị là 3156.26
Ngày có giá trị High cao nhất là: 2024-03-14 với giá trị là 73777.0
```

70,621 USD

Is the Profit if you buy BTC at Lowest cost 3,156 USD and sell at Highest cost 73,777 USD

According to statistics from Coin98 Insights, 64% of cryptocurrency investors in Vietnam are not profitable, with nearly 44% incurring losses in the past year.

This indicates that investment decisions were made at the wrong time, due to the difficulty investors face in capturing the trends of BTC as well as the events that occur.

The importance of Bitcoin price prediction

It helps investors **manage risks and optimize profits** in the context of strong Bitcoin price fluctuations.

With its decentralized nature and the influence of psychological factors and legal regulations, Bitcoin price prediction helps investors better **understand the market and protect assets**.

In addition, it **promotes technological** development, especially in AI and machine learning, and creates opportunities for **research and application to other digital assets**, directly affecting the global financial market.

Dataset

The dataset have **12 columns** and total **2441 records** (stand for 2441 days from 2018 to 2024)

The main features we use to analyze are:
Open time: Timestamp when the trading period begins.
Open: Price of Bitcoin at the start of the trading period.
High: Highest price during the trading period.
Low: Lowest price during the trading period.
Close: Price of Bitcoin at the end of the trading period.
Volume: Total number of Bitcoin traded during the period.
Number of Trades: The number of transactions is an important indicator to assess the market's vitality.

```
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Open time                   2441 non-null   object
 1   Open                        2441 non-null   float64
 2   High                        2441 non-null   float64
 3   Low                         2441 non-null   float64
 4   Close                       2441 non-null   float64
 5   Volume                      2441 non-null   float64
 6   Close time                  2441 non-null   object
 7   Quote asset volume          2441 non-null   float64
 8   Number of trades            2441 non-null   int64
 9   Taker buy base asset volume 2441 non-null   float64
 10  Taker buy quote asset volume 2441 non-null  float64
 11  Ignore                      2441 non-null   int64
dtypes: float64(8), int64(2), object(2)
memory usage: 229.0+ KB
```

02

Analyze

Data preparation
Exploratory data analysis
Summarize the insights and review

Data Preparation

Data Preparation

Check for Null and duplicate Values:
First, we checked for any missing (null) and duplicate values in the dataset to ensure data completeness. The result indicated that there were no null values and no duplicate, so the dataset was perfect.

```
#check null & duplicate
df.isnull().sum()
```

	0
Open time	0
Open	0
High	0
Low	0
Close	0
Volume	0
Close time	0
Quote asset volume	0
Number of trades	0
Taker buy base asset volume	0
Taker buy quote asset volume	0
Ignore	0

```
[ ] df.duplicated().sum()
```

```
0
```

Data Preparation

Drop ‘Ignore’ and ‘CLose time’ column:

Because 'Ignore' column have just one unique value so we need to drop it. Also, we use ‘Open time’ column already so we drop ‘Close time’.

	dtypes	missing#	missing%	uniques	count	min	max	mean
Open time	object	0	0.000000	2441	2441	nan	nan	nan
Open	float64	0	0.000000	2440	2441	3211.710000	73072.400000	25638.038853
High	float64	0	0.000000	2382	2441	3276.500000	73777.000000	26260.692028
Low	float64	0	0.000000	2383	2441	3156.260000	71333.310000	24956.096841
Close	float64	0	0.000000	2440	2441	3211.720000	73072.410000	25654.639590
Volume	float64	0	0.000000	2441	2441	1521.537318	760705.362783	73408.989955
Close time	object	0	0.000000	2441	2441	nan	nan	nan
Quote asset volume	float64	0	0.000000	2441	2441	11770168.043866	17465307097.884071	1772412409.582372
Number of trades	int64	0	0.000000	2439	2441	12417.000000	15223589.000000	1553324.748464
Taker buy base asset volume	float64	0	0.000000	2441	2441	844.258813	374775.574085	36497.579067
Taker buy quote asset volume	float64	0	0.000000	2441	2441	6532638.637519	8783916247.676138	877987275.806545
Ignore	int64	0	0.000000	1	2441	0.000000	0.000000	0.000000

Data Preparation

```
[ ]  #check if the column ['Open time'] is a continuous, unbroken series of days

     # Convert 'Open time' to datetime objects
     df['Open time'] = pd.to_datetime(df['Open time'])

     # Sort the DataFrame by 'Open time'
     df = df.sort_values('Open time')

     # Check if the 'Open time' column is a continuous series of days
     expected_dates = pd.date_range(start=df['Open time'].min(), end=df['Open time'].max(), freq='D')
     actual_dates = df['Open time'].unique()

     if set(expected_dates) == set(actual_dates):
       print("The 'Open time' column is a continuous series of days from 2018 to 2024.")
     else:
       print("The 'Open time' column is not a continuous series of days.")
       missing_dates = set(expected_dates) - set(actual_dates)
       print("Missing dates:", missing_dates)
```

```
The 'Open time' column is a continuous series of days from 2018 to 2024.
```

Check for Missing Days in Time Series:

Since the dataset involves time-series data, we verified the continuity of the 'Open time' column to ensure there were no missing days. The result confirmed that no days were missing, ensuring a consistent time series.

Data Preparation

Check and Correct Data Types:

The 'Open time' column, initially stored as an object type, was converted to the correct datetime format for accurate time-based analysis and plotting.

```
Data columns (total 12 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Open time                   2441 non-null   object
 1   Open                        2441 non-null   float64
 2   High                        2441 non-null   float64
 3   Low                         2441 non-null   float64
 4   Close                       2441 non-null   float64
 5   Volume                      2441 non-null   float64
 6   Close time                  2441 non-null   object
 7   Quote asset volume          2441 non-null   float64
 8   Number of trades            2441 non-null   int64
 9   Taker buy base asset volume 2441 non-null   float64
 10  Taker buy quote asset volume 2441 non-null  float64
 11  Ignore                      2441 non-null   int64
dtypes: float64(8), int64(2), object(2)
memory usage: 229.0+ KB
```

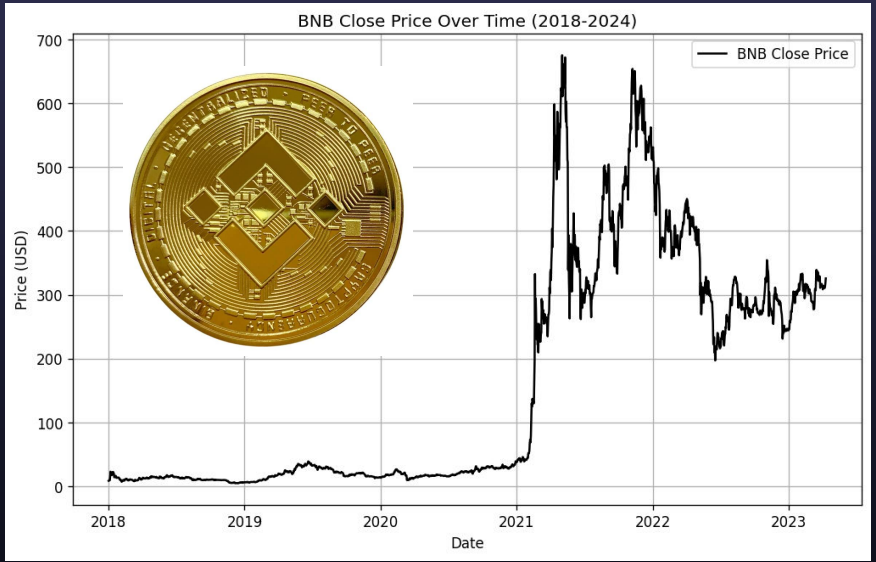
Data Preparation

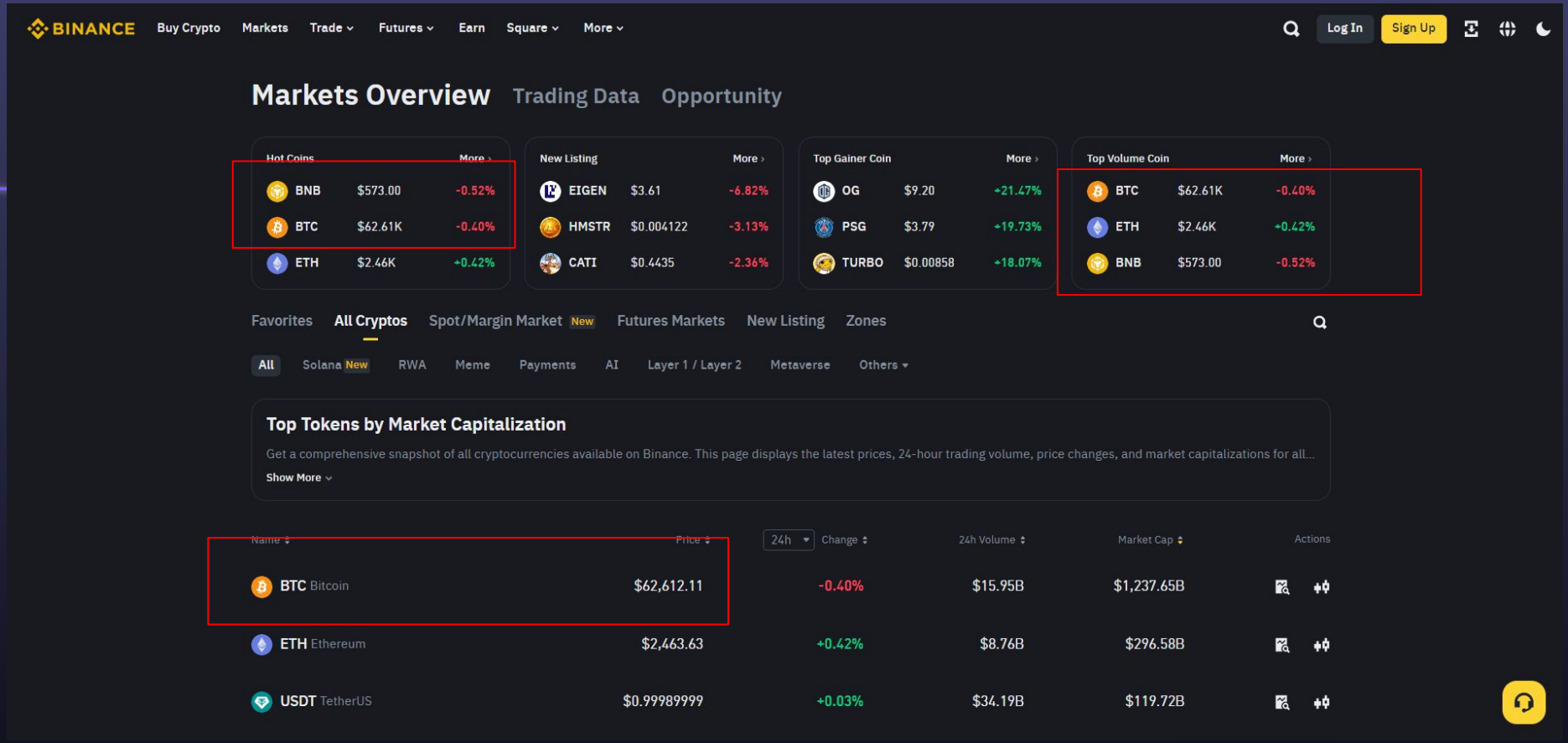
Data Extension:

To demonstrate that Bitcoin (BTC) significantly influences the entire cryptocurrency market, we prepared a second dataset for BNB (Binance Coin), a popular cryptocurrency in Vietnam. In the exploratory data analysis (EDA), we will compare the 'Close' prices of both BTC and BNB to see their trends and prove that BTC often sets the tone for market movements in the crypto space.

```
df1.head()
```

	Date	Open	High	Low	Close	Adj Close	Volume
0	2017-11-10	2.00773	2.06947	1.64478	1.79684	1.79684	11155000
1	2017-11-11	1.78628	1.91775	1.61429	1.67047	1.67047	8178150
2	2017-11-12	1.66889	1.67280	1.46256	1.51969	1.51969	15298700
3	2017-11-13	1.52601	1.73502	1.51760	1.68662	1.68662	12238800
4	2017-11-14	1.68928	1.73537	1.56827	1.59258	1.59258	7829600



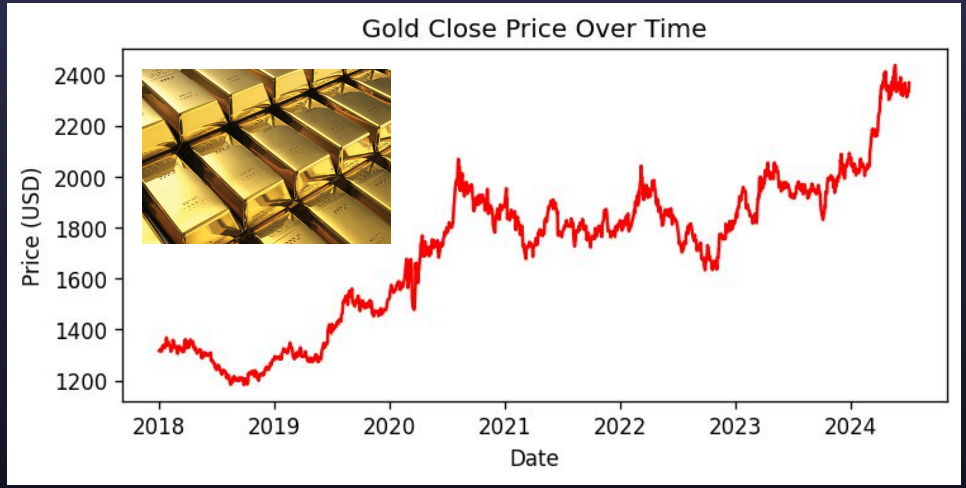

BINANCE
Buy Crypto
Markets
Trade
Futures
Earn
Square
More
Log In
Sign Up
Markets Overview
Trading Data
Opportunity
Hot Coins
More
BNB $573.00 -0.52%
BTC $62.61K -0.40%
ETH $2.46K +0.42%
New Listing
More
EIGEN $3.61 -6.82%
HMSTR $0.004122 -3.13%
CATI $0.4435 -2.36%
Top Gainer Coin
More
OG $9.20 +21.47%
PSG $3.79 +19.73%
TURBO $0.00858 +18.07%
Top Volume Coin
More
BTC $62.61K -0.40%
ETH $2.46K +0.42%
BNB $573.00 -0.52%
Favorites
All Cryptos
Spot/Margin Market New
Futures Markets
New Listing
Zones
All
Solana New
RWA
Meme
Payments
AI
Layer 1 / Layer 2
Metaverse
Others
Top Tokens by Market Capitalization
Get a comprehensive snapshot of all cryptocurrencies available on Binance. This page displays the latest prices, 24-hour trading volume, price changes, and market capitalizations for all...
Show More
Name
Price
24h
Change
24h Volume
Market Cap
Actions
BTC Bitcoin $62,612.11 -0.40% $15.95B $1,237.65B
ETH Ethereum $2,463.63 +0.42% $8.76B $296.58B
USDT TetherUS $0.99989999 +0.03% $34.19B $119.72B

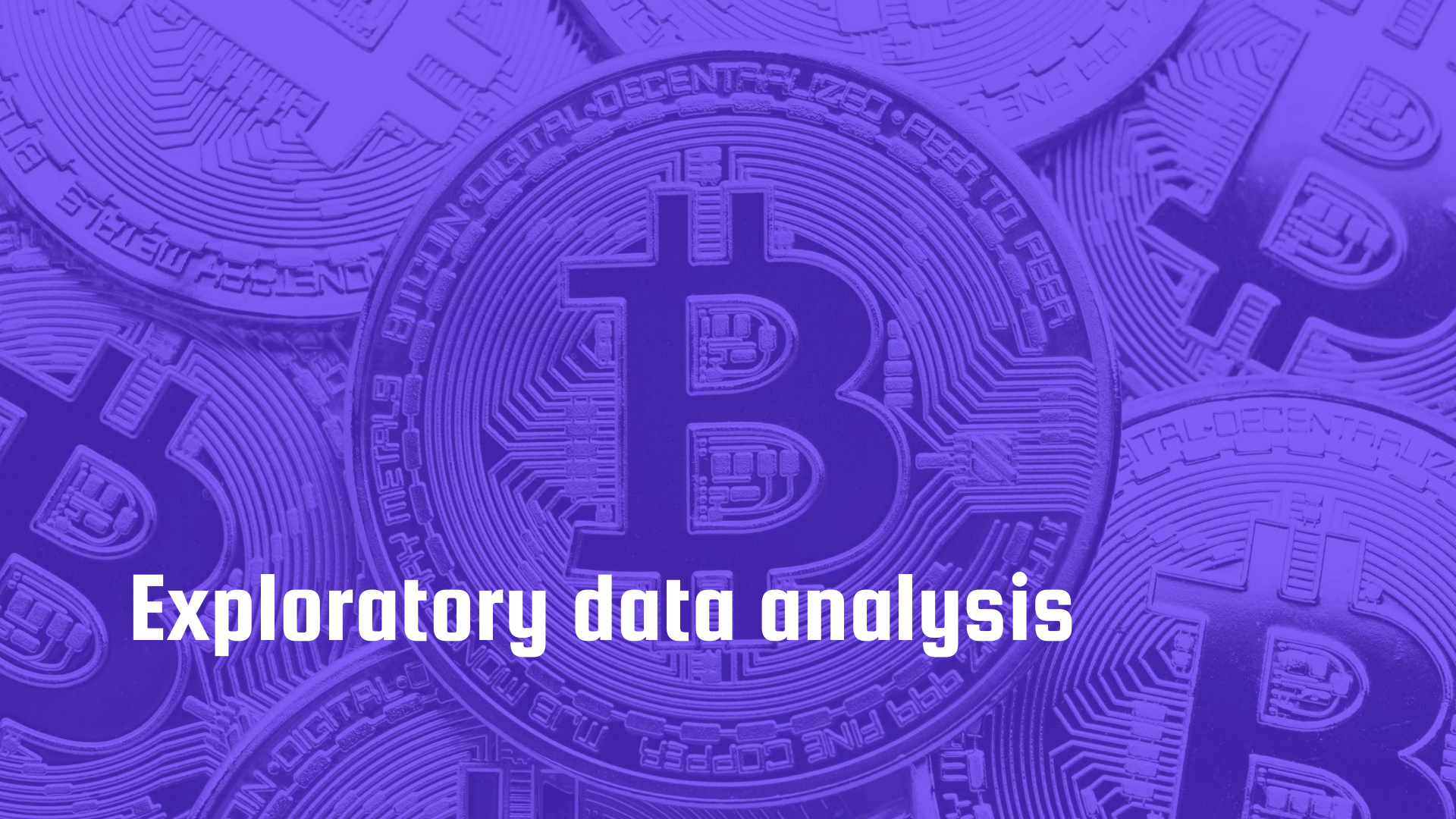
Data Preparation

Date	Close/Last	Volume	Open	High	Low
2018-01-02	1316.1	269072.0	1305.3	1320.4	1304.6
2018-01-03	1318.5	342866.0	1319.0	1323.0	1308.9
2018-01-04	1321.6	350803.0	1315.5	1327.3	1307.1
2018-01-05	1322.3	322422.0	1324.4	1324.7	1314.6
2018-01-08	1320.4	238332.0	1321.8	1323.0	1315.7

Data Extension:

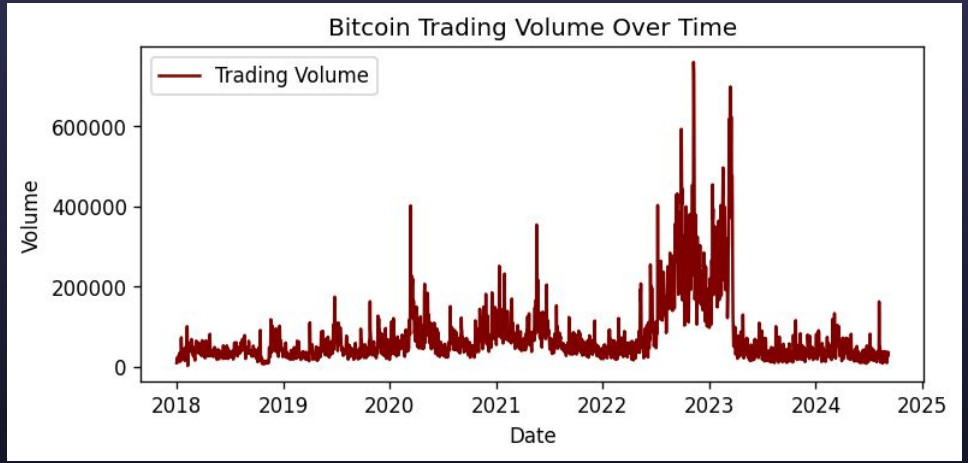
To broaden the comparison of BTC prices with the general market, we have prepared an additional dataset on global gold price fluctuations (in USD) to compare Close prices. This will help investors determine whether to invest in gold (a traditional investment channel) or BTC (a new investment trend).




Exploratory data analysis

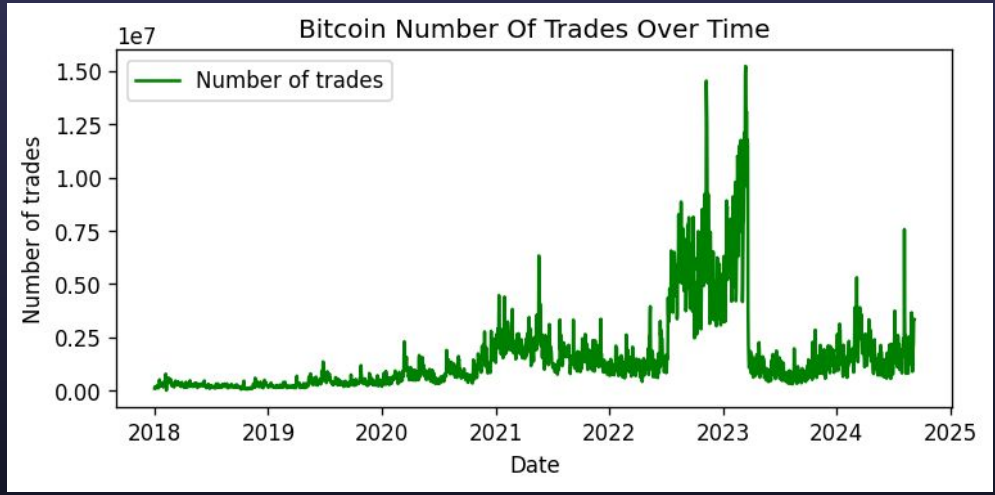
Univariate analysis

The data shows that there is a large variation in the number of transactions over time. Periods with high volumes of transactions are often accompanied by higher price volatility, indicating strong investor participation. The distribution of the number of transactions tends to be uneven, with some trading sessions having high levels of activity outpacing the rest.

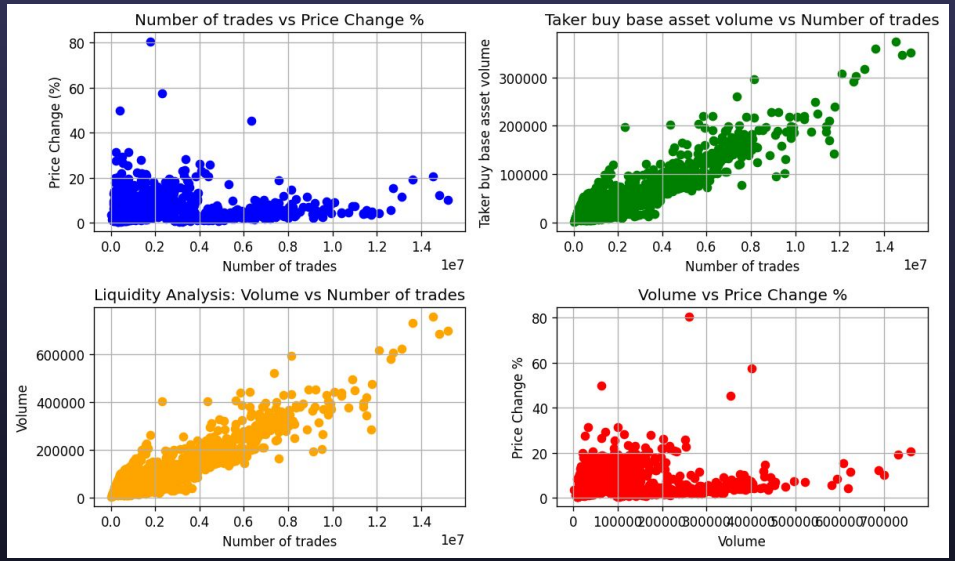


Univariate analysis

Number of Trades reflects the amount of Bitcoin that is bought and sold during each trading session. When number is high, the market tends to be more liquid, meaning there are more buyers and sellers and the price is less affected by large orders.



Multivariate Analysis



Volume vs Price Change: Investors should watch for rising volume with price increases for potential opportunities, but be cautious of volume spikes with price drops as they may signal a future downtrend.

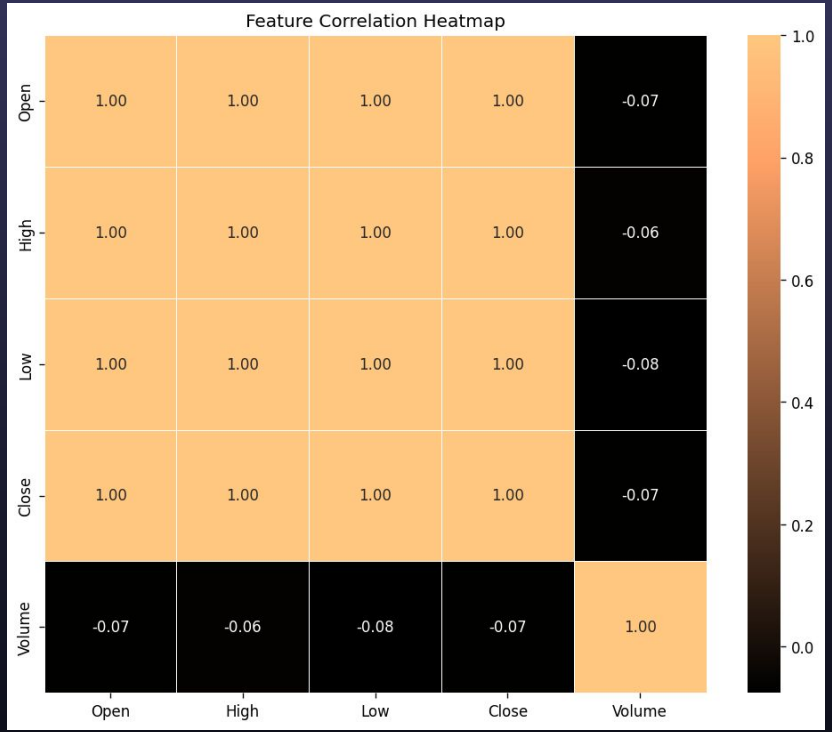
Taker buy base asset volume vs Number of trades: As the number of taker buy base asset volume increases, buy volume also tends to increase, reflecting higher buying momentum.

Volume vs Number of Trades: High liquidity periods are preferable for smoother trading and reduced risk from price volatility.

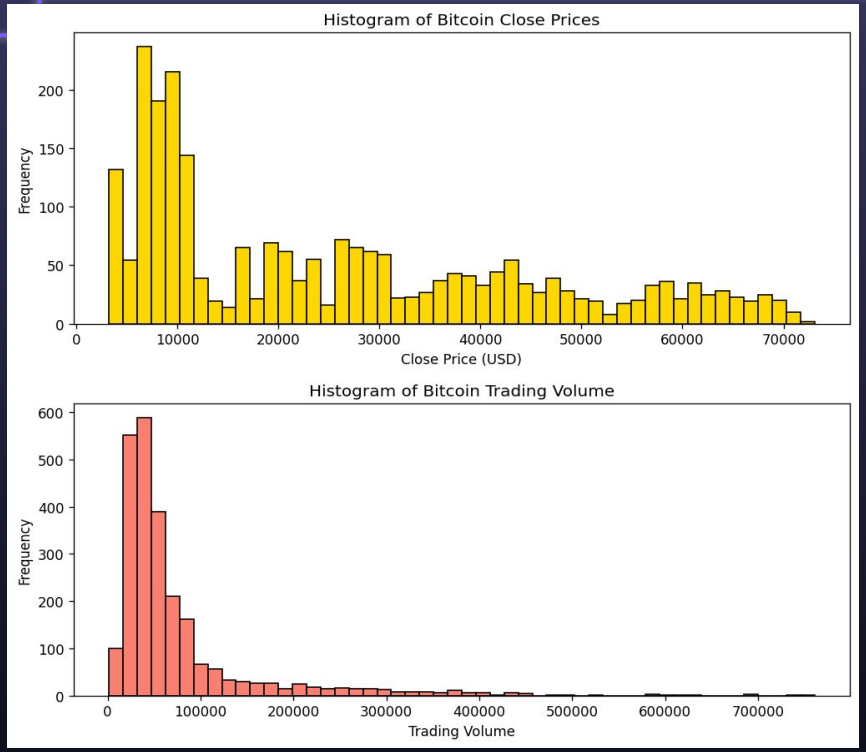
Number of Trades vs Price Change: The number of trades can help predict market fluctuations and identify investment opportunities or risks.

Feature Correlation Heatmap

The heatmap shows a perfect positive correlation (1.00) between Open, High, Low, and Close prices, indicating these features move in sync. However, Volume has a very weak negative correlation (around -0.07) with the price-related features, suggesting that changes in trading volume have little influence on price movements.



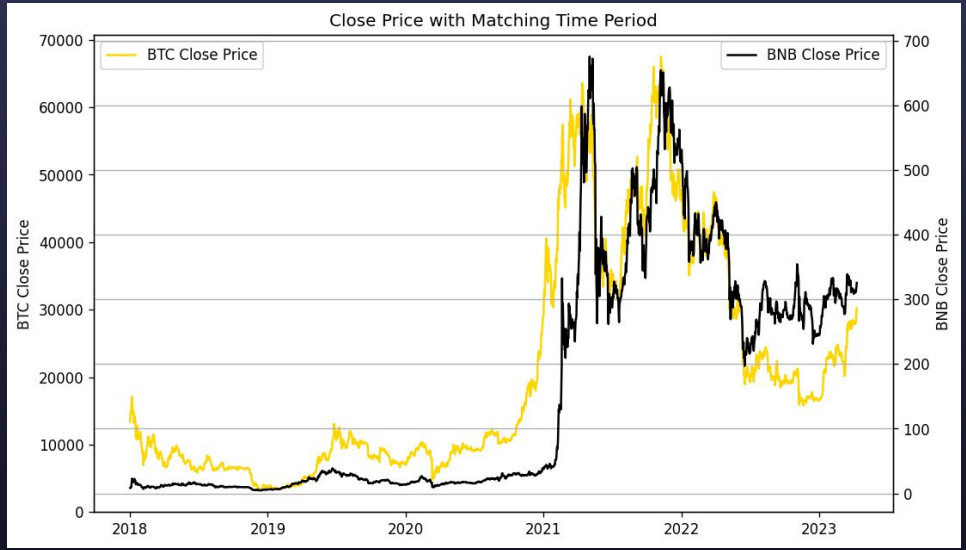
Histogram of Bitcoin Close Price vs. Trading Volume



The histograms for Bitcoin Close Price and Trading Volume display similar distributions, indicating that periods of high trading activity tend to correspond with higher closing prices. This suggests that Bitcoin's price and trading volume often rise and fall together, reflecting increased market participation during price surges.

BNB vs BTC Close Price

Both BTC and BNB rose significantly in late 2020, peaking in early to mid-2021. BTC showed more volatility, reaching nearly $70,000 before declining, while BNB followed with less drastic changes. After 2021, both trended downward, but BNB stabilized around $300 post-2022. The overall correlation suggests BNB's price is influenced by BTC's movements during major market events.



Gold vs BTC Close Price

For risk-averse investors: Gold seems to be the safer choice, as it is less prone to drastic price fluctuations and offers long-term security.

For risk-tolerant investors: BTC could potentially yield higher returns, especially if there is another significant surge in price. However, its high volatility suggests that it is a riskier investment that could lead to significant losses if not timed well.

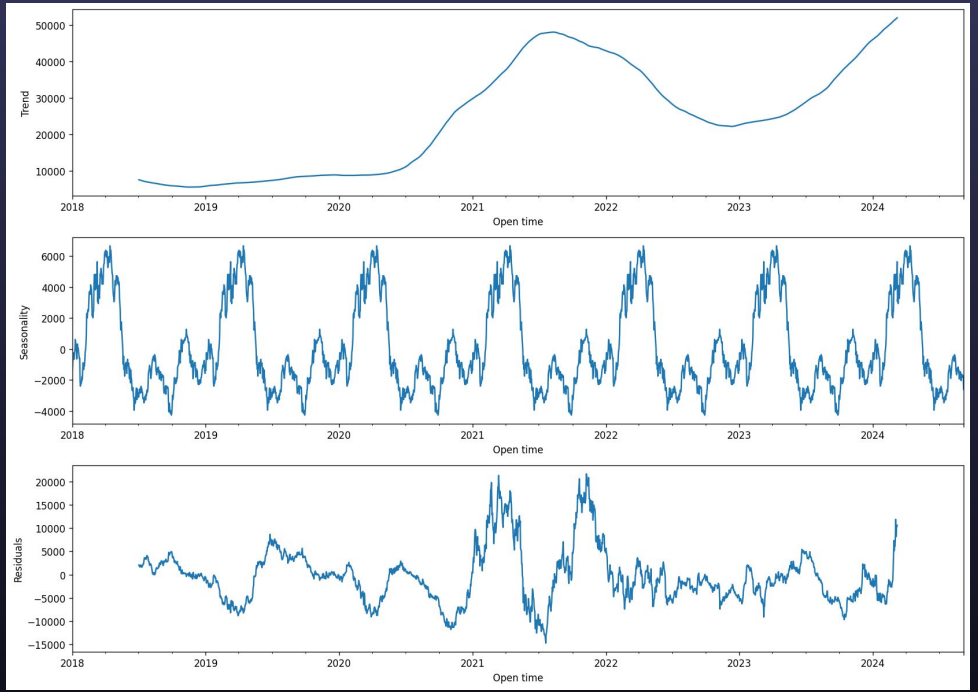
Ultimately, the decision should be based on the investor's risk tolerance.

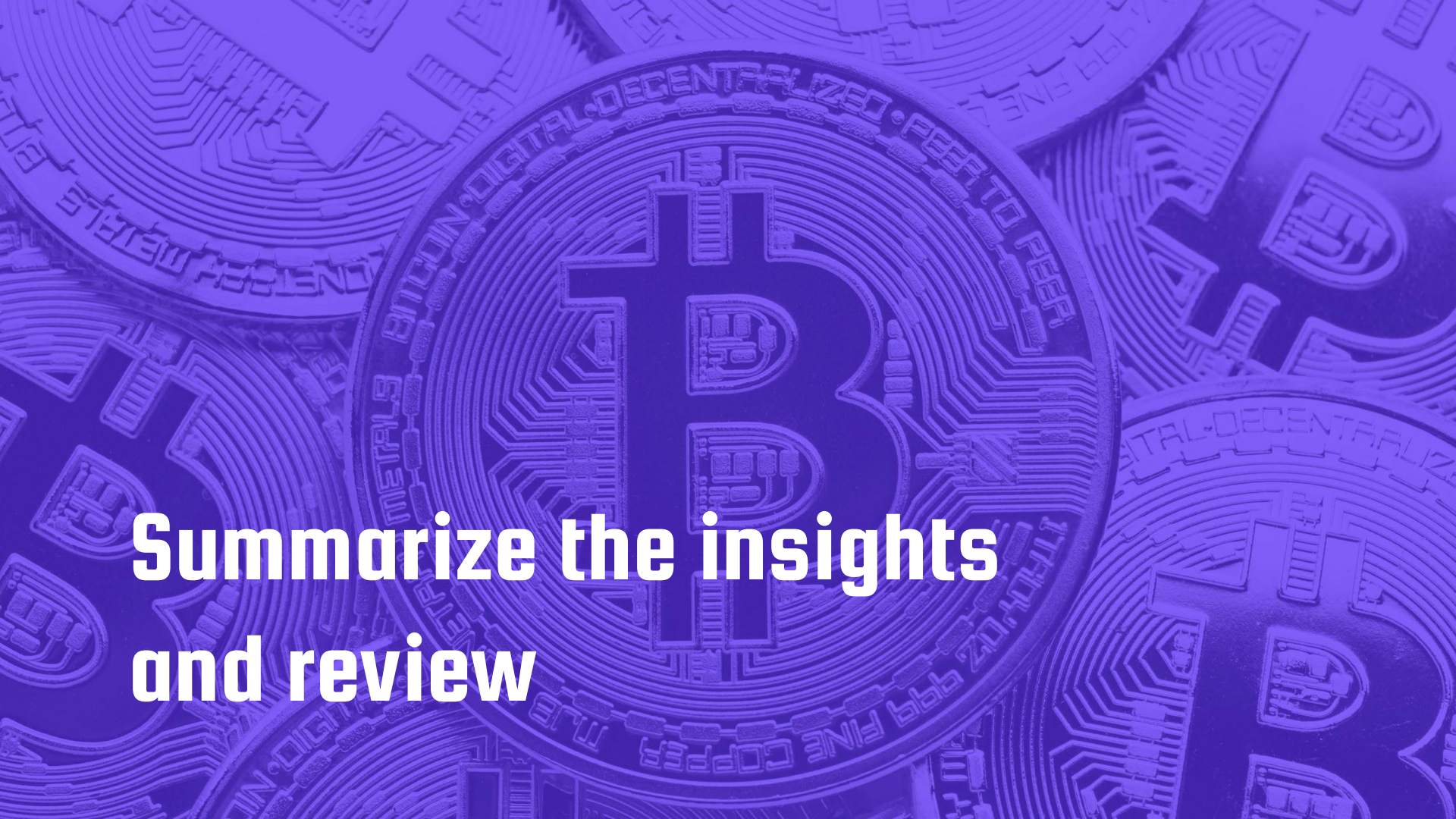
Trend, Seasonality and Residuals of BTC

Trend: A rise peaking in early 2022, followed by a decline and a slight uptick in 2024.

Seasonality: Regular, cyclical patterns repeat annually.

Residuals: Unexplained noise, with higher volatility around 2021-2022 and more stable fluctuations afterward.




Summarize the insights
and review

Number of trades and Volume

Transaction Volatility: High transaction volumes often lead to greater price volatility, reflecting strong market activity. Some trading sessions see significantly more transactions, driving market liquidity.

Price-Volume Correlation: Bitcoin's Open, High, Low, and Close prices move together, while trading volume has little effect on price. However, high trading volumes often coincide with higher closing prices, indicating increased participation during price surges.

In crypto market - BTC vs. BNB: Both BTC and BNB surged in late 2020, peaking in 2021. BTC was more volatile, while BNB stabilized post-2022, with its price influenced by BTC during major market shifts.

In general market - BTC vs. Gold:

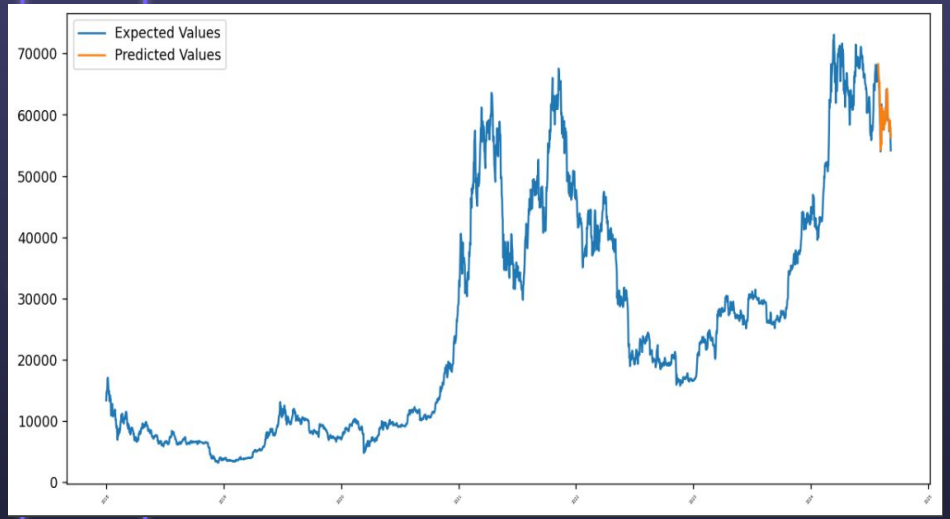
- **Risk-averse:** Gold is a safer, stable investment.
- **Risk-tolerant:** BTC offers higher potential returns but with higher volatility.

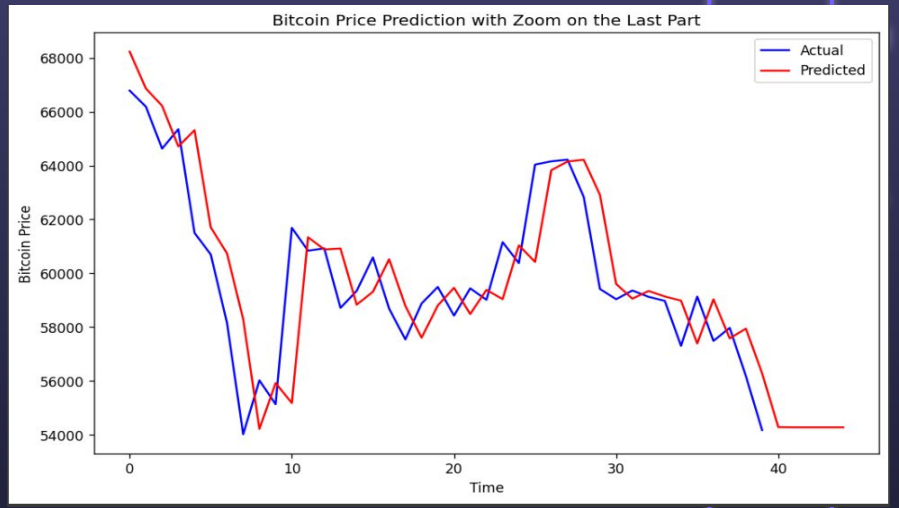
03

Model prediction

Build a machine learning model for prediction

ARIMA





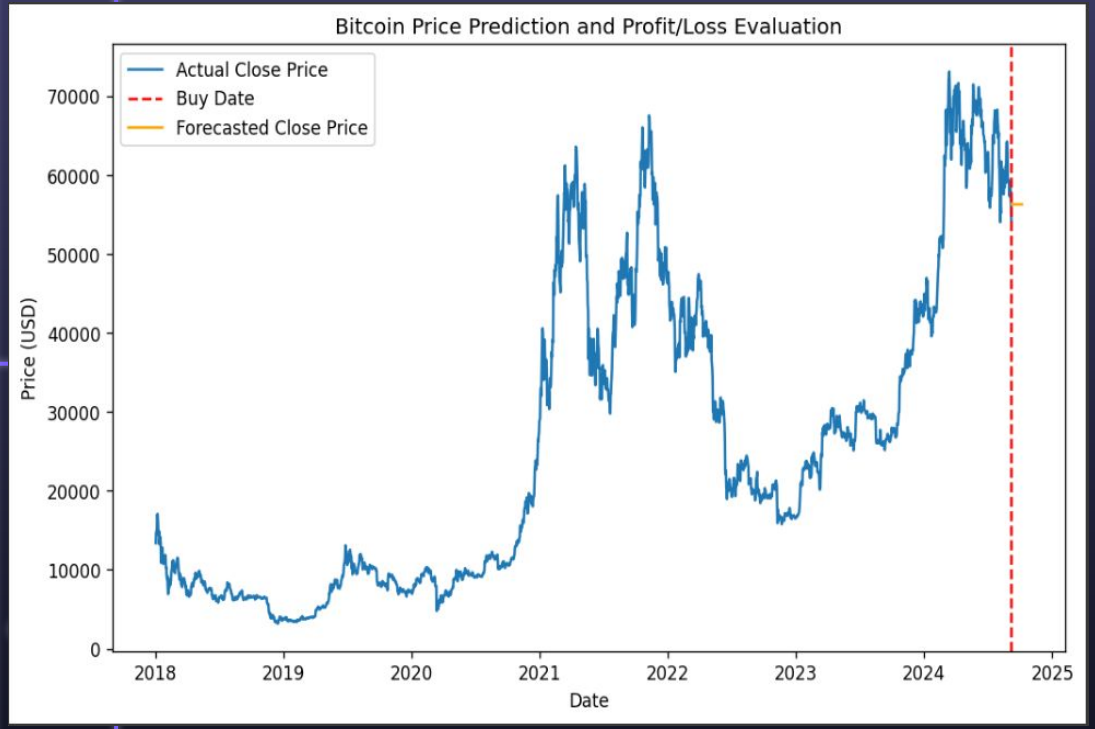

```
R2 score: 0.57
Mean squared error score: 3926232.97
Test RMSE: 1981.47
```

The results of the 1-step prediction show that the R2 score is not too bad, the large MSE may be due to the volatile data, which Bitcoin price often encounters.

```
R2 score: -3.48
Mean squared error score: 41051662.30
Test RMSE: 6407.16
```

It can be seen that when trying to run the prediction 5 more steps, the indicators all get much worse.

ARIMA



Applying the ARIMA model to the process of buying and predicting the price of Bitcoin in the next few days after buying. The prediction results will help to have an overview of the future price trend, based on historical data. In addition, calculating Profit/Loss is also an important part for any investor, as it helps to understand the performance of their investment decision.

LINEAR REGRESSION

Create additional target columns for further forecast dates, for example the closing price of the next session (shift -2, -3, -4, -5). The goal of the model is to predict the closing price of the next sessions based on current data.

```
Training model to predict Close2
Results for Close2:
MSE: 139056140.8476264
RMSE: 11792.206784466865
MAE: 7904.344609541559
R2: 0.6172691677002228
```

```
Training model to predict Close3
Results for Close3:
MSE: 125278512.02376951
RMSE: 11192.788393593864
MAE: 7843.369927554638
R2: 0.6350768929827351
```

```
Training model to predict Close4
Results for Close4:
MSE: 124797022.5754394
RMSE: 11171.258773094436
MAE: 7731.591603343411
R2: 0.6490680062705974
```

```
Training model to predict Close5
Results for Close5:
MSE: 129430064.07979965
RMSE: 11376.733453843402
MAE: 8057.653391168241
R2: 0.6549630504417597
```

```
Training model to predict Close6
Results for Close6:
MSE: 123024757.59209046
RMSE: 11091.652608700404
MAE: 7748.4994086976385
R2: 0.67167473143702
```

Increasing R^2: From Close2 to Close6, the R^2 value gradually increases from 0.6173 to 0.6717, indicating that the model has better predictive ability when predicting further in time.

Decreasing error: In general, RMSE and MAE tend to decrease, especially clearly at Close6, indicating that the model is more stable in predicting values as time goes on.

Good model performance: Although there is still a significant amount of error (RMSE around 11,000), the model did quite well in predicting values with about 60-67% explained by the model.

RANDOM FOREST

```
Evaluating Random Forest model for Close2
Training Results for Close2:
Train MSE: 516675.20247103606, Train RMSE: 718.8012259804765, Train R2: 0.9986219909407291
Test MSE: 4505961.821371706, Test RMSE: 2122.7250932166667, Test R2: 0.9875980268998379

Evaluating Random Forest model for Close3
Training Results for Close3:
Train MSE: 709189.674988655, Train RMSE: 842.134000613118, Train R2: 0.998132368880584
Test MSE: 4629139.014365406, Test RMSE: 2151.5434028541945, Test R2: 0.9865158057463473

Evaluating Random Forest model for Close4
Training Results for Close4:
Train MSE: 834517.9118087238, Train RMSE: 913.5195191175303, Train R2: 0.997783237087666
Test MSE: 6869158.509689896, Test RMSE: 2620.9079552113035, Test R2: 0.980683774009179

Evaluating Random Forest model for Close5
Training Results for Close5:
Train MSE: 997159.0571029887, Train RMSE: 998.5785182463063, Train R2: 0.9973189305218275
Test MSE: 6352546.657615914, Test RMSE: 2520.4258881419055, Test R2: 0.983065268983265

Evaluating Random Forest model for Close6
Training Results for Close6:
Train MSE: 1267540.7209730577, Train RMSE: 1125.8511095935633, Train R2: 0.9965953603189659
Test MSE: 8514049.494782101, Test RMSE: 2917.884421080126, Test R2: 0.97727792647882
```

1. Training Performance:
   - The Train R^2 values for all variables are very high (from 0.9965 to 0.9986), indicating that the model has learned very well on the training data.
   - The Train MSE (Mean Squared Error) and Train RMSE (Roo Mean Squared Error) of variables Close2 to Close6 are quite low, reflecting that the prediction error on the training data is small.

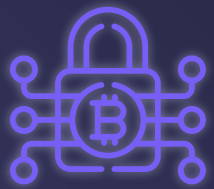
2. Test Performance:
   - Test R^2 for variables Close2 to Close6 are all greater than 0.97, showing that the model has the ability to predict quite accurately on the test data.
   - However, Test MSE and Test RMSE for the variables are much larger than the training data, especially Close4 to Close6 with RMSE over 2500. This shows that there is a quite large deviation when predicting on the test data compared to the training data.

04

Suggestion

Suggest solution for short or long term investment

Suggestion



I. Short-Term

Active trading during high volatility sessions, leveraging momentum for quick gains.

Short-Term Investment Strategy (Active Trading)

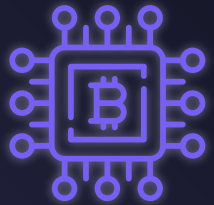
Trend Following: Based on results from linear or logistic regression models, investors can track trends and trade according to momentum. If the forecast indicates that Bitcoin's price will continue to rise due to strong increases in transaction volume and the number of transactions, this could be a signal to buy in the short term.

Buy the Dip Strategy: Using predictive models like ARIMA, investors can identify times when the price is likely to decrease in the short term. This strategy can optimize profits by buying when prices are low and selling when prices recover.

Suggestion

2. Long-term Investment (Holding)
Holding a Long-term Position (Holding): Based on analyses from artificial neural network models, investors can take advantage of long-term strategies, especially when market trends indicate strong growth. If data shows that Taker Buy Volume is continuously increasing and the price maintains an upward trend, this could be a sign that large investors are accumulating Bitcoin, creating favorable conditions for a long-term bull cycle.

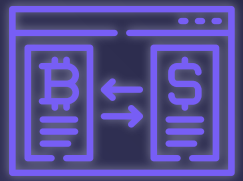
Portfolio Diversification: In addition to holding Bitcoin, investors can consider diversifying into other cryptocurrency assets like Ethereum or Binance Coin (BNB). This helps mitigate risk in case Bitcoin experiences significant volatility, while other cryptocurrencies remain stable or continue to grow.



2. Long-Term

Hold BTC for growth potential, use dollar-cost averaging, and diversify with assets like Gold and BNB to manage risk.

Suggestion



3. Risk Management

Leverage management and the use of hedging tools.

Risk Management

Leverage Management: When using predictive models, investors need to be cautious with leverage. While leverage can quickly increase profits during upward trends, it also carries significant risks if the market reverses.

Using Hedging Tools: Investors can use tools such as futures contracts or options to hedge against risks when the market shows signs of high volatility. This helps protect the investment portfolio from unexpected events like sharp sell-offs.

Conclusion

Opportunity

For short-term traders, capitalizing on transaction volume spikes and momentum can lead to quick gains

Long-term investors, on the other hand, can benefit from Bitcoin's potential for significant growth, especially when balanced with more stable assets like Gold or BNB to mitigate market downturns.

Risk

Ultimately, successful BTC investment strategies depend on understanding market dynamics, managing leverage wisely, and diversifying to hedge against risk.

Thank you!

Do you have any questions?

Data source:
https://www.kaggle.com/code/ahmedaboraida/bitcoin-price-eda-and-prediction-r2-score-99/input