Exploratory Data Analysis (EDA) Report

1. Introduction

This Exploratory Data Analysis (EDA) is performed to understand the historical behavior of cryptocurrency market data and to identify patterns related to price movement, volatility, and liquidity. The goal of this analysis is to gain insights from the data before building a machine learning model for volatility prediction. All analysis steps strictly follow the preprocessing and feature engineering implemented in the project code.

---

2. Dataset Overview

The dataset is loaded from a CSV file and contains daily historical data for a cryptocurrency. Initially, an unnamed index column was present in the dataset, which was removed as it did not add any analytical value.

Dataset attributes include: - timestamp: Exact time of record generation - date: Trading date - crypto_name: Name of the cryptocurrency - open: Opening price of the day - high: Highest price of the day - low: Lowest price of the day - close: Closing price of the day - volume: Trading volume - marketCap: Market capitalization

The dataset represents time-series data where each row corresponds to a single trading day.

---

3. Data Cleaning and Preprocessing

Initially, the dataset was checked for missing values. No null values were found in the raw dataset. The data types of all columns were verified to ensure correctness. The `timestamp` and `date` columns were converted into datetime format to support proper time-based analysis. After conversion, the data was sorted according to the `date` column to maintain chronological order.

Numerical features such as open, high, low, close, volume, and market capitalization were scaled using StandardScaler. Feature scaling was applied to bring all numerical values onto a similar scale, which helps improve model performance and stability.

---

4. Feature Engineering

Several new features were engineered to better capture market behavior:

- **Volatility:** Calculated using the formula (high − low) / close, which represents daily price fluctuation relative to closing price.
- **Rolling Volatility (7-day):** Standard deviation of closing price over a 7-day rolling window.

- **Rolling Volatility (14-day):** Standard deviation of closing price over a 14-day rolling window.
- **Moving Averages:** 7-day and 14-day moving averages of closing price to identify short-term and medium-term trends.
- **Liquidity Ratio:** Calculated as volume divided by market capitalization to measure market activity and liquidity.

After feature creation, rolling window operations introduced missing values at the beginning of the dataset. These rows were removed to ensure a clean dataset for analysis.

5. Statistical and Distribution Analysis

The statistical summary of numerical features shows that price-related variables exhibit significant variation, confirming the volatile nature of cryptocurrency markets. Trading volume and market capitalization show uneven distributions, with noticeable spikes during certain periods, indicating high market participation during volatile phases.

6. Correlation Analysis

Correlation analysis was performed on all numerical features to understand their relationships. A correlation matrix and heatmap were generated for visualization. Strong positive correlation was observed among price-based features such as open, high, low, and close prices. Volatility showed noticeable correlation with price fluctuations and moderate association with volume and liquidity ratio. This confirms that trading activity and price movements jointly influence market volatility.

7. Key Observations

- Cryptocurrency prices show frequent and sharp fluctuations over time.
- Volatility increases during periods of high price movement.
- Volume spikes often align with increased volatility.
- Rolling statistics and moving averages help smooth short-term noise and capture underlying trends.
- Liquidity ratio provides additional insight into market participation.

8. Conclusion

The EDA highlights the highly volatile nature of the cryptocurrency market and justifies the use of engineered features such as volatility measures, moving averages, and liquidity ratio. These insights directly influenced feature selection for the machine learning model and improved the model's ability to predict future volatility.