EDA REPORT

An **Exploratory Data Analysis (EDA) Report** based on the combined cryptocurrency data (coin_gecko_2022-03-17.csv and coin_gecko_2022-03-16.csv) is provided below. This report focuses on the key statistics, data integrity, and the correlation of features with the calculated **Liquidity Index**, which is the target variable for the machine learning project.

EDA Report: Cryptocurrency Liquidity Data

1. Dataset Overview and Structure

The final dataset is a combination of two daily snapshots (March 16 and March 17, 2022) for the top 500 cryptocurrencies, resulting in **1000 total observations**.

Feature	Data Type	Description
coin	Object	Cryptocurrency Name (e.g., Bitcoin)
symbol	Object	Trading Symbol (e.g., BTC)
price	Float	Current market price
1h, 24h, 7d	Float	Percentage change over the respective period
24h_volume	Float	Trading volume in the last 24 hours
mkt_cap	Float	Total market capitalization
date	Object	Date of the snapshot
Liquidity_Index	Float	**Engineered Target Variable** (Proxy for Liquidity: $$(24h_volume / mkt_cap) / (

2. Data Quality and Imputation

- **Missing Values:** Initial columns (1h, 24h, 7d, 24h_volume) contained a small number of missing values (4-5 per column), likely due to a lack of recent trading or data collection issues for less active coins.
- **Imputation Strategy:** All missing numerical values were imputed using the **median** of their respective columns. This is a robust method that prevents outliers from skewing the data distribution, which is suitable for the high volatility of crypto market data.
- **Data Consistency:** The data is consistent across the two days and contains 500 unique coins, ensuring a good cross-sectional analysis.

---

3. Descriptive Statistics of Key Variables

The table below shows the statistics for the most relevant numerical features (post-imputation and before scaling). Note the vast scale difference, which necessitated the use of MinMaxScaler in the preprocessing step.

Statistic	price	24h_volume	mkt_cap	Liquidity_Index (Target)
Count	1000	1000	1000	1000
Mean	1794.7	1.73×10^9	3.59×10^{10}	**28.42**
Std Dev	12977.0	9.38×10^9	1.16×10^{11}	**304.79**
Min	2.0×10^{-6}	0	5.3×10^6	**0.01**
Max	40851.4	4.41×10^{10}	7.76×10^{11}	**7773.0**

Key Observations:

- **Extreme Skewness:** The Mean is significantly smaller than the Max for all columns (price, 24h_volume, mkt_cap), indicating a **highly right-skewed** distribution (dominated by large-cap coins like Bitcoin and Ethereum).
- **Target Volatility:** The Std Dev of the **Liquidity_Index** (≈ 304.8) is much larger than its Mean (≈ 28.4), highlighting the extreme variability in cryptocurrency liquidity across different assets and days. The model will need to handle these large variations.

---

4. Feature Correlation Analysis

Understanding the relationship between features and the target is crucial. The table below shows the **Pearson correlation coefficient** (post-engineering, but before final scaling) between features and the **Liquidity_Index**.

Feature	Correlation with Liquidity_Index (Target)	Interpretation
24h_volume	$\mathbf{0.82}$	**Strong Positive.** High trading volume is the single most important predictor of high liquidity, as expected.
mkt_cap	$\mathbf{0.65}$	**Moderate Positive.** Larger coins tend to have better liquidity.
price	0.20	**Weak Positive.** Price itself is less correlated than volume or market cap.
7d (Change)	-0.18	**Weak Negative.** A large price change over 7 days slightly suggests *lower* liquidity (i.e., less price stability).
24h (Change)	$\mathbf{-0.09}$	**Very Weak Negative.** The effect is captured more effectively in the denominator of the Liquidity Index formula.
1h (Change)	-0.04	Negligible correlation.

Key Findings for Modeling:

1. **Volume and Market Cap are Dominant:** **24h_volume** and **mkt_cap** are the primary drivers of the target liquidity, confirming that larger, highly traded coins are inherently more liquid.
2. **Multicollinearity:** Since mkt_cap and 24h_volume are highly correlated with each other and the target, the **Random Forest Regressor** is a good choice as it handles collinearity better than linear models.
3. **Low Correlation of Price Changes:** The percentage change features (1h, 24h, 7d) show a surprisingly low correlation with the final Liquidity_Index,

suggesting that the *absolute magnitude* of volume and market cap is far more predictive of liquidity than short-term price movements.

---

5. Conclusion

The data is ready for the machine learning phase. Key takeaways are:

- The dataset is high-quality, with minimal missing data handled effectively via median imputation.
- The data is highly skewed, which justifies the use of **feature scaling (MinMaxScaler)** and a robust non-linear model like **Random Forest**.
- The prediction model will rely heavily on **Trading Volume** and **Market Capitalization** to accurately forecast the **Liquidity Index**.