**📊 Exploratory Data Analysis (EDA) Report**

**1. 📁 Data Overview**

* Two CSV files (coin\_gecko\_2022-03-16.csv, coin\_gecko\_2022-03-17.csv) are merged.
* Initial Shape: (500, 9) + (500, 9) → Combined using np.vstack then rewrapped as DataFrame (1000,9).
* Dropped Columns: ['symbol', 'date', 'coin'] – likely identifiers.
* All columns cast to float.

**2. 🧼 Missing Value Treatment**

* Filled missing values with **median** for:
  + 1h, 24h, 7d, 24h\_volume
  + Key Observations

1. \*\*Price Distribution\*\*:

- Prices range from nearly 0 to ~17.79

- Right-skewed distribution (mean > median)

- Most coins clustered at lower prices (75th percentile at 7.23)

2. \*\*Price Changes\*\*:

- Average 1h change: +1.01%

- Average 24h change: +2.00%

- Average 7d change: +0.12%

- Shows generally positive but modest short-term performance

3. \*\*Volatility\*\*:

- The `variablity\_score` shows most coins have low volatility (mean 0.034)

- Some coins show high volatility (max 0.131)

4. \*\*Market Cap Relationships\*\*:

- The `24h\_mkt\_cap\_ratio` is extremely small (mean 1.08e-10)

- `coin\_number` shows extreme range (3.85e+06 to 1.59e+17)

**3. 📏 Outlier Treatment**

* IQR-based clipping applied to each numeric column:
  + outlier() function calculated 25th, 75th percentiles and capped values outside the [Q1 - 1.5\*IQR, Q3 + 1.5\*IQR] range.
  + This prevents extreme values from skewing model learning.
  + 'price', '1h', '24h', '7d', '24h\_volume', 'mkt\_cap' columns had outlier which had been replaced with upper value and lower value of respective columns.

**4. 🛠️ Feature Engineering**

You created new features:

* variablity\_score → Standard deviation of 1h, 24h, 7d
* 24h\_mkt\_cap\_ratio → Target variable (24h return as % of market cap)
* coin\_number → Proxy for number of coins (mkt\_cap / price)

Dropped:

* 24h\_volume, mkt\_cap, price, coin\_number (final version)

Key Features:

* - `price`: Current price of the cryptocurrency
* - `1h`: 1-hour price change percentage
* - `24h`: 24-hour price change percentage
* - `7d`: 7-day price change percentage
* - `24h\_volume`: 24-hour trading volume
* - `mkt\_cap`: Market capitalization
* - `variablity\_score`: Standard deviation of price changes (1h, 24h, 7d)
* - `24h\_mkt\_cap\_ratio`: Ratio of 24h price change to market cap
* - `coin\_number`: Derived metric (market cap divided by price)

**5. 📉 Feature Distribution**

For each column (including engineered ones), you generated:

* **Histogram**
* **KDE Plot**

These are useful for:

* Checking skewness
* Spotting multimodal distributions
* Deciding transformations (e.g., log scale for highly skewed data)

**6. 🔍 Data Splitting and Scaling**

* Split into train/test using 70:30 ratio
* Applied **StandardScaler** to standardize numeric features (mean=0, std=1)

✅ This is good for linear models, but not required for tree-based models (like Random Forest or XGBoost).

**✅ Summary Table**

| **Step** | **Description** |
| --- | --- |
| Data Merge | Two CSVs merged vertically |
| Cleaning | Dropped non-numeric columns, filled missing |
| Outlier Handling | IQR-based clipping per column |
| Feature Engineering | 3 new columns derived, some dropped later |
| Visualization | Histograms + KDE for all features |
| Splitting & Scaling | Train-test split (70:30) + StandardScaler |