

Low-Level Design (LLD)

1. Data Ingestion

- **Files:** coin_gecko_2022-03-16.csv, coin_gecko_2022-03-17.csv
 - **Implementation:**
 - Load both datasets using pandas
 - Stack them using `numpy.vstack()`
 - Reconstruct as a DataFrame
-

2. Data Cleaning

- **Remove:** Columns not useful for modeling (symbol, date, coin)
 - **Convert:** All columns to float type
-

3. Outlier Handling

- Function `outlier()` computes IQR for each numeric column
 - Clips values outside the $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$ range
 - Stores limits in dictionaries (Upper_limit, Lower_limit)
 - Applies clipping to outlier columns only
-

4. Missing Value Imputation

- Impute median for columns with missing values:
 - 1h, 24h, 7d, 24h_volume
-

5. Feature Engineering

- **variability_score:** Standard deviation across 1h, 24h, 7d
- **24h_mkt_cap_ratio:** Ratio of 24h change to market cap
- **coin_number:** Derived as `mkt_cap / price` (then dropped later)
- Drop 24h_volume, mkt_cap, price, and coin_number after feature creation

6. Exploratory Data Analysis

- Use seaborn.histplot to visualize distribution of each feature
- Automatic loop creates plots for every column

7. Model Preparation

- **Target variable:** 24h_mkt_cap_ratio
- **Features:** Remaining numeric columns
- **Split:** 70% training, 30% test
- **Scaling:** StandardScaler applied to both training and test features

8. Modeling & Evaluation

- Likely algorithm (based on structure): RandomForestRegressor or similar ensemble model
 - Train model on X_train, Y_train
 - Evaluate predictions on X_test, Y_test
 - Metrics expected: MAE, RMSE, or R^2 (not visible in the current preview but assumed)
-