🔧 Pipeline Architecture

📄 Pipeline Documentation

📌 Title: Cryptocurrency Market Cap Analysis & Prediction Pipeline

🎯 Purpose

To predict the 24h_mkt_cap_ratio for cryptocurrencies using historical market data and engineered features, aiming to support decision-making in trading or investment.

📁 Data Sources

File Name	Description
coin_gecko_2022-03-16.csv	Snapshot of cryptocurrency data on 2022-03-16
coin_gecko_2022-03-17.csv	Snapshot on 2022-03-17

📊 Processing Steps

1. Data Ingestion

- Read both CSVs using pandas.
- Merge using np.vstack() and reinitialize DataFrame.

2. Data Cleaning

- Drop irrelevant columns: symbol, date, coin
- Convert all data to float

3. Outlier Detection

- IQR-based function computes limits per column.
- Outliers are clipped within [Q1 - 1.5*IQR, Q3 + 1.5*IQR].

4. Missing Value Imputation

- Columns 1h, 24h, 7d, and 24h_volume are filled using median.

5. Feature Engineering

- variability_score: Std dev across 1h, 24h, 7d
- 24h_mkt_cap_ratio: 24h change divided by mkt_cap
- coin_number: mkt_cap / price (later dropped)

6. Visualization

- Histograms with KDE for all features using seaborn.histplot

7. Model Preparation

- Target: 24h_mkt_cap_ratio
- Feature Matrix: Remaining columns (after dropping target and price, coin_number)
- Scaling: StandardScaler
- Split: 70% training / 30% testing

8. Modeling

- Model: Assumed RandomForestRegressor or similar ensemble
- Evaluation: Performed on test set using suitable metrics (MAE, RMSE)

🔍 **Assumptions**

- No categorical variables
- Model will regress a continuous variable (24h_mkt_cap_ratio)
- Evaluation metrics not explicitly coded but implied

📈 **Outputs**

- Cleaned and transformed dataset
- Trained ML model
- Feature importance (if model supports it)
- Performance metrics (assumed in final steps)

Would you like me to deliver this in a downloadable .docx or .pdf format?