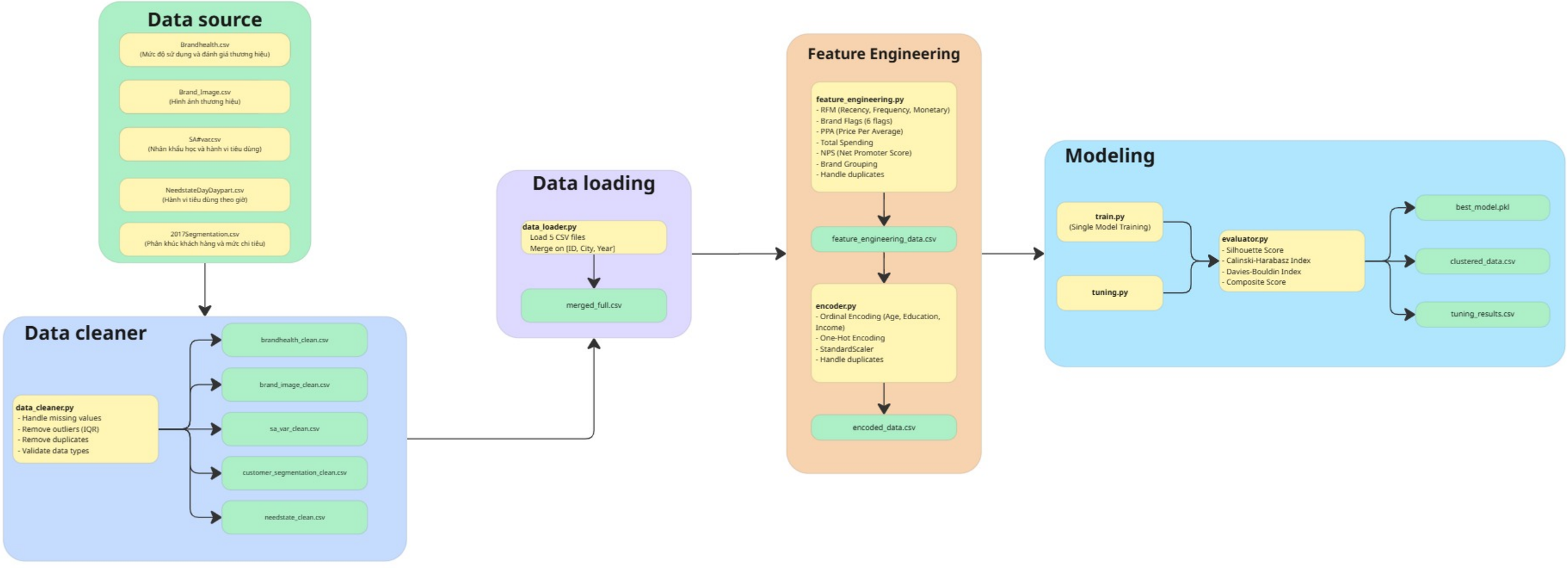
Data source
Brandhealth.csv
(Mức độ sử dụng và đánh giá thương hiệu)
Brand_Image.csv
(Hình ảnh thương hiệu)
SAtfvar.csv
(Nhân khẩu học và hành vi tiêu dùng)
NeedstateDayDaypart.csv
(Hành vi tiêu dùng theo giờ)
2017Segmentation.csv
(Phân khúc khách hàng và mức chi tiêu)
Data cleaner
data_cleaner.py
- Handle missing values
- Remove outliers (IQR)
- Remove duplicates
- Validate data types
brandhealth_clean.csv
brand_image_clean.csv
sa_var_clean.csv
customer_segmentation_clean.csv
needstate_clean.csv
Data loading
data_loader.py
Load 5 CSV files
Merge on [ID, City, Year]
merged_full.csv
Feature Engineering
feature_engineering.py
- RFM (Recency, Frequency, Monetary)
- Brand Flags (6 flags)
- PPA (Price Per Average)
- Total Spending
- NPS (Net Promoter Score)
- Brand Grouping
- Handle duplicates
feature_engineering_data.csv
encoder.py
- Ordinal Encoding (Age, Education, Income)
- One-Hot Encoding
- StandardScaler
- Handle duplicates
encoded_data.csv
Modeling
train.py
(Single Model Training)
tuning.py
evaluator.py
- Silhouette Score
- Calinski-Harabasz Index
- Davies-Bouldin Index
- Composite Score
best_model.pkl
clustered_data.csv
tuning_results.csv