5. Data Preparation and Exploratory Data Analysis

5.1 Preparation Approach

The data preparation stage transforms validated raw data into a **clean, structured, and machine-learning-ready dataset** suitable for feature engineering and recommendation modeling.

This stage focuses on:

- Cleaning and filtering interaction data
- Enriching user interactions with product metadata
- Encoding categorical attributes
- Normalizing numerical variables
- Generating reproducible exploratory analysis artifacts

All transformations are performed programmatically to ensure **consistency and reproducibility**.

5.2 Data Cleaning and Enrichment

The following cleaning steps are applied to the clickstream interaction data:

- Removal of records with missing `user_id` or `product_id`
- Filtering to retain valid interaction types (`view`, `click`, `add_to_cart`)
- Conversion of timestamps into standardized datetime format

User interaction data is then **enriched** by joining with product metadata using `product_id`, allowing interaction records to include product attributes such as category, price, and ratings.

5.3 Feature Encoding and Normalization

To prepare the dataset for downstream modeling, the following transformations are applied:

Categorical Encoding

- Interaction types are mapped to numerical interaction strength values:
  - `view = 1`
  - `click = 2`
  - `add_to_cart = 3`
- Product categories are label-encoded into numerical identifiers

Numerical Normalization

- Product prices are normalized using min–max scaling
- Temporal information is extracted from timestamps (hour of day) to capture time-based user behavior patterns

These steps ensure that all features are represented in a format suitable for machine learning algorithms.

5.4 Exploratory Data Analysis (EDA)

Exploratory analysis is conducted to understand the structure and characteristics of the prepared data. The following analyses are performed:

- **Interaction distribution** across event types

- **Item popularity analysis** based on user interactions
- **User–item sparsity analysis** to quantify the sparsity of the interaction matrix

All EDA visualizations are generated automatically and saved as reproducible artifacts.

5.5 Prepared Data and EDA Artifacts

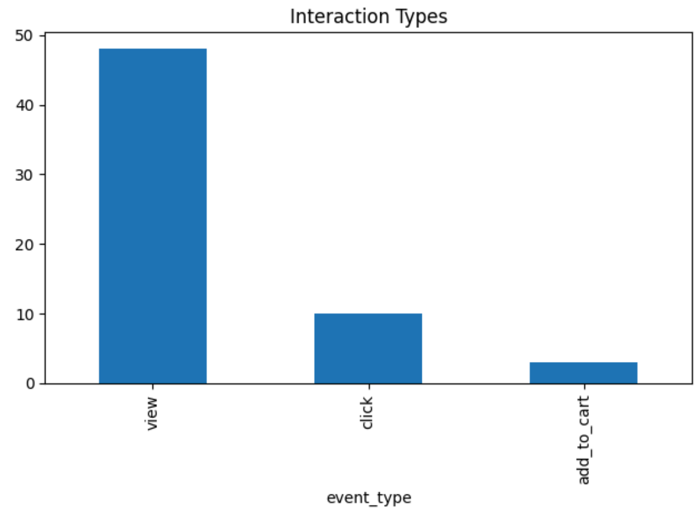
Processed Data Output

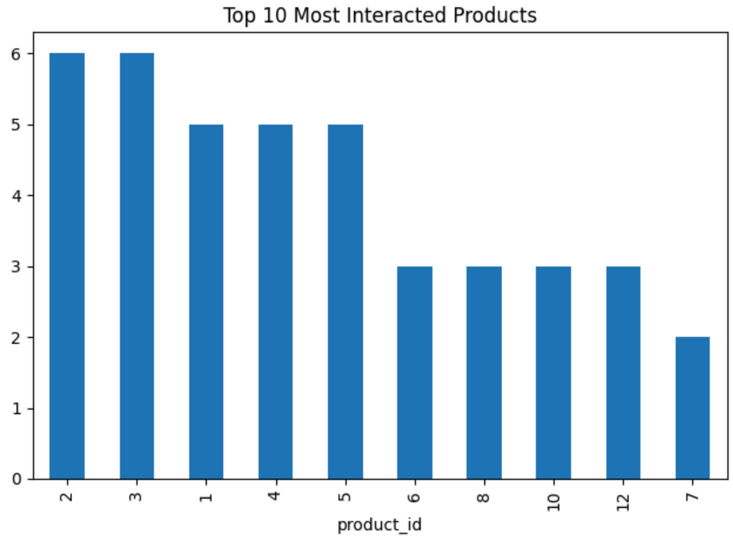
```
data/processed/prepared_interactions.csv
```

This dataset contains cleaned and enriched interaction records and is ready for feature engineering and model training.

EDA Artifacts

```
data/processed/eda/
├── interaction_distribution.png
└── top_products.png
```





These visualizations summarize key interaction patterns and are used for analysis and reporting.

5.6 Reproducibility and Pipeline Readiness

- All preparation steps are implemented as a standalone script
- Outputs are deterministic and reproducible
- Generated datasets and plots can be versioned and reused in downstream stages

This preparation layer provides a reliable foundation for **feature engineering and transformation** in the subsequent pipeline stage.