

Final Report: Customer Data Curation and Analysis

Overview

Retail companies generate large volumes of customer transaction data, loyalty card information, and demographic attributes. However, this data is often inconsistently formatted and difficult to analyze. The objective of this project is to perform a comprehensive data curation and customer segmentation analysis using the Dunnhumby “The Complete Journey” dataset (Kaggle, 2025). It focuses on cleaning, integrating, provenance tracking, reproducibility and documenting customer and household-level data. The **use case** focuses on understanding household and customer purchasing behaviors based on demographics available.

Context

Datasource

The datasets were obtained from Kaggle, specifically the Dunnhumby “The Complete Journey” dataset, which includes detailed transactional, household demographic, product, campaign, and coupon redemption data. The datasets used include below and has been documented in [metadata.md](#) file.

Data use and ethics

Data was obtained from Kaggle (Dunnhumby - The Complete Journey). All household identifiers in the dataset are **fully anonymized**, and the data contains **no personally identifiable information (PII)**. Our analysis adheres to ethical data practices by avoiding any

attempt at re-identification, limiting use to aggregate statistics, and ensuring that outputs cannot be traced back to individual households. All processing and reporting steps are transparent and reproducible, and the project complies with responsible data-handling standards.

Artifact repository

- <https://github.com/PremnishaBalakumar/customer-analysis/>.
- Due to the large size of raw and cleaned data folder, uploaded it as [zip file](#)

Quality Assessment

Initial assessment (from `01_data_exploration.ipynb`) found:

- Missing or ambiguous demographic categories (e.g., "Unknown", "Not Available")
- Inconsistencies in casing/whitespace
- Rare invalid dates
- Household records missing multiple mandatory demographic fields
- Transaction entries with zero/negative quantities

These findings were addressed during the cleaning phase.

Workflow Overview

The curated workflow consists of **five main components**:

Input Acquisition

- Raw CSVs downloaded from Kaggle

- Stored in `data/raw/`
- Documented in `data/metadata.md`

Data Cleaning and Standardization

Implemented in `src/data_cleaning.py`

- Duplicate Handling: Removed duplicate rows from all datasets and recorded % duplicates removed in provenance logs
- Missing & Unknown Values:
 - Identified columns with NaN or placeholder values like "Unknown", "", "NaN".
 - Recorded counts and percentages of missing/unknown values per column for all datasets
- Standardization: Trimmed leading/trailing spaces in string columns.
- Invalid Transaction Filtering: Removed transactions with non-positive QUANTITY or SALES_VALUE
- Provenance Logging: All cleaning steps, counts, and percentages recorded in `logs/workflow.txt`
- Cleaned Data Folder: `data/cleaned/`

Integration and Aggregation

`src/customer_analysis.py` integrates cleaned household demographics with transaction aggregates

- Integrate Datasets
 - Merge transactions with demographics using `household_key`.

- Merge the result with product data using `PRODUCT_ID`.
- Produces a fully enriched transaction-level dataset.
- Aggregate household metrics like coupons redeemed.
- Merge aggregates with demographics.

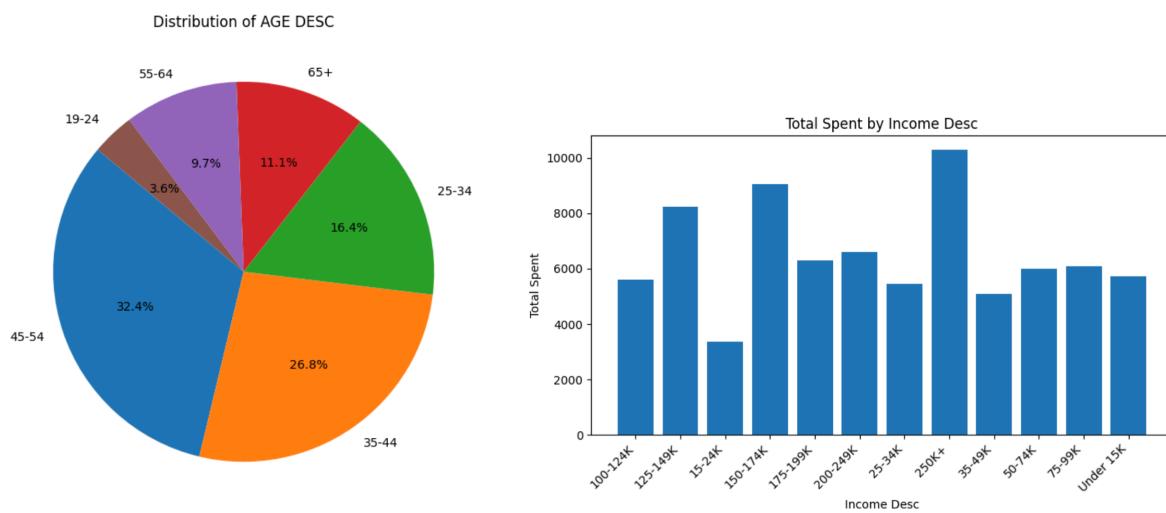
Results saved as: `data/customer_segmentation/transactions_merged.csv`

Analysis & Visualization

`generate_demographic_piecharts()` in `customer_analysis.py` produces pie charts for key demographics like marital status, income range, age etc.

`generate_histogram()` in `customer_analysis.py` produces histograms for coupon code redeemed, basket income ratio and spend income ratio.

Figures saved under: `data/customer_segmentation/figures/`. Some example figures generated can be seen below



Automated Workflow Execution

A full pipeline runner:

```
python3 -m src.customer_workflow
```

This performs below steps making the entire workflow reproducible and re-executable.

- Loading and verification of raw data
- Cleaning (duplicates, invalid rows, normalization)
- Demographic mapping & heuristics
- Merge & aggregation of household-level metrics
- Figures generation saved to `data/customer_segmentation/figures/`
- Writes logs to `logs/workflow.txt`

DCC Lifecycle Mapping

DCC Stage	Project Activities
Ingest & Acquire	Downloaded raw CSVs; stored in <code>data/raw</code> ; created metadata entry documenting source & schema.
Appraise & Select	Identified incomplete household demographics; excluded those missing required fields.
Transform & Clean	Cleaning pipeline

Store & Preserve	Stored cleaned CSVs in <code>data/cleaned</code> ; versioned via Git; logs saved to <code>logs/workflow.txt</code> .
Create & Use	Aggregated dataset produced (<code>transactions_merged.csv</code>). Output of analysis created as figures
Reuse	Reproducible python module, runbook in readme to reproduce the workflow, logs for provenance tracking with timestamp.
Disseminate	Public GitHub repo , project report, well-documented folders.

Learning from Course Project

Metadata Strategy

Located in `data/metadata.md`, includes dataset source, variables, description and misaligned values.

Identifier System

- **Household keys:** anonymized IDs provided by dataset
- **Provenance tracking:** sequential logs capture transformations

Documentation

- **README:** overview + instructions
- **RUNBOOK:** step-by-step to execute pipeline
- **Code comments:** especially in merge and cleaning functions
- **Environment file:** `requirements.txt` for reproducibility

Provenance & Transparency

Logged via `logs/workflow.txt`. Automated timestamps and details of steps executed

Workflow & Reproducibility

- Pipeline re-doable end-to-end
- Outputs deterministic given same input
- Notebooks reference static snapshots in `data/cleaned`

Findings and Challenges

- Large transaction files made some steps slow and required efficiency fixes.
- Useful learnings wrt provenance, reproducible workflow was great.
- A substantial set of records were dropped due to missing and invalid demographic fields.
Filtering ensured fair and unbiased comparisons across key metrics.
Although representativeness decreased, data quality and analytic validity were prioritized.
- Few aggregated metrics like coupons redeemed could not be accurately measured across due to missing demographic information and invalid baskets.

References

- [Dunnhumby Retail Dataset, Kaggle](#)
- Github repo - <https://github.com/PremnishaBalakumar/customer-analysis/>
- DCC Curation Lifecycle Model