

Retail Store Sales Dataset Documentation

2025-12-02

Team

Melisa Cihan, Hrusheekesh Sawarkar, Kezia Fernandes, Raju Ahmed

1. Introduction

This document provides a short documentation of the dataset used for the project. The dataset was obtained from Kaggle:

Source: *Retail Store Sales (Dirty) – For Data Cleaning* URL: <https://www.kaggle.com/datasets/ahmedmohamed2003/retail-store-sales-dirty-for-data-cleaning/data>

The goal of this documentation is to summarize the dataset, describe its variables and types, and provide a reproducible R script used to load the data.

2. Dataset Description

The dataset contains transactional retail sales data including product category, pricing, customer information, and purchase behavior. It is intended for data cleaning and preprocessing tasks.

```
# required packages
library(readr)
library(dplyr)

data <- read_csv("retail_store_sales.csv")

## Rows: 12575 Columns: 11
## -- Column specification -----
## Delimiter: ","
## chr (6): Transaction ID, Customer ID, Category, Item, Payment Method, Location
## dbl (3): Price Per Unit, Quantity, Total Spent
## lgl (1): Discount Applied
## date (1): Transaction Date
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

3. Summary of the Dataset

3.1 Number of Rows and Columns

```
dim(data)
```

3.2 Summary of Variables

```
summary(data)

## Transaction ID          Customer ID          Category           Item
## Length:12575            Length:12575            Length:12575        Length:12575
## Class :character        Class :character        Class :character    Class :character
## Mode  :character        Mode  :character        Mode  :character    Mode  :character
##
## Price Per Unit          Quantity          Total Spent       Payment Method
## Min.   : 5.00            Min.   : 1.000         Min.   : 5.0        Length:12575
## 1st Qu.:14.00           1st Qu.: 3.000        1st Qu.: 51.0      Class :character
## Median :23.00           Median : 6.000        Median :108.5      Mode  :character
## Mean   :23.37           Mean   : 5.536        Mean   :129.7
## 3rd Qu.:33.50           3rd Qu.: 8.000        3rd Qu.:192.0
## Max.   :41.00           Max.   :10.000        Max.   :410.0
## NA's   :609              NA's   :604           NA's   :604
## Location                Transaction Date     Discount Applied
## Length:12575            Min.   :2022-01-01    Mode :logical
## Class :character         1st Qu.:2022-09-30   FALSE:4157
## Mode  :character         Median :2023-07-13   TRUE :4219
##                           Mean   :2023-07-12   NA's :4199
##                           3rd Qu.:2024-04-24
##                           Max.   :2025-01-18
##
```

3.3 Data Types of Each Variable

```
glimpse(data)

## # Rows: 12,575
## # Columns: 11
## # $ 'Transaction ID'      <chr> "TXN_6867343", "TXN_3731986", "TXN_9303719", "TXN_9-
## # $ 'Customer ID'        <chr> "CUST_09", "CUST_22", "CUST_02", "CUST_06", "CUST_0-
## # $ Category              <chr> "Patisserie", "Milk Products", "Butchers", "Beverag-
## # $ Item                   <chr> "Item 10 PAT", "Item 17 MILK", "Item 12 BUT", "Item-
```

```

## $ 'Price Per Unit'      <dbl> 18.5, 29.0, 21.5, 27.5, 12.5, NA, 5.0, 33.5, 27.5, ~
## $ Quantity              <dbl> 10, 9, 2, 9, 7, 10, 8, NA, 1, 3, 9, 8, 7, 6, 2, NA, ~
## $ 'Total Spent'         <dbl> 185.0, 261.0, 43.0, 247.5, 87.5, 200.0, 40.0, NA, 2~
## $ 'Payment Method'       <chr> "Digital Wallet", "Digital Wallet", "Credit Card", ~
## $ Location               <chr> "Online", "Online", "Online", "Online", "Online", "~-~
## $ 'Transaction Date'     <date> 2024-04-08, 2023-07-23, 2022-10-05, 2022-05-07, 20~
## $ 'Discount Applied'      <lgl> TRUE, TRUE, FALSE, NA, FALSE, NA, TRUE, TRUE, FALSE~

```

4. Variable Description

Variable Name	Description	Data Type
Transaction ID	Unique identifier for each transaction	Nominal
Customer ID	Unique identifier for each customer	Nominal
Category	Product category (e.g., Patisserie, Butchers)	Nominal
Item	Specific purchased item	Nominal
Price Per Unit	Price of a single item unit	Numeric
Quantity	Quantity purchased	Numeric
Total Spent	Price \times Quantity	Numeric
Payment Method	Payment mode (Digital Wallet, Credit Card, etc.)	Nominal
Location	Purchase location (Online or physical store)	Nominal
Transaction Date	Date of the transaction	Ordinal
Discount Applied	Indicates if a discount was used (True/False)	Boolean

5. Example Data (First Few Observations)

```
head(data, 5)
```

```

## # A tibble: 5 x 11
##   `Transaction ID` `Customer ID` Category     Item    `Price Per Unit` Quantity
##   <chr>           <chr>        <chr>       <chr>      <dbl>        <dbl>
## 1 TXN_6867343    CUST_09      Patisserie   Item_1~      18.5         10
## 2 TXN_3731986    CUST_22      Milk Products Item_1~      29            9
## 3 TXN_9303719    CUST_02      Butchers    Item_1~      21.5         2
## 4 TXN_9458126    CUST_06      Beverages   Item_1~      27.5         9
## 5 TXN_4575373    CUST_05      Food        Item_6~      12.5         7
## # i 5 more variables: `Total Spent` <dbl>, `Payment Method` <chr>,
## #   Location <chr>, `Transaction Date` <date>, `Discount Applied` <lgl>

```

6. Conclusion

This dataset provides a structured but intentionally “dirty” retail transaction record intended for data cleaning exercises. The documentation summarizes its structure and provides reproducible R code for loading and exploring the data. Furthermore, a in-depth domain expertise is not necessary for this dataset, since most people already have some everyday knowledge in this retail context.