Data Understanding & Exploration - Self-Checkout Fraud Dataset

Philipp Altenbach, Ronny Grieder, Omar Rahiel, Emre Yelögrü

2025-02-28

This document presents an initial exploratory analysis of the data set related to self-checkout fraud detection. The focus is on understanding the structure of the `fraud.csv` data set before proceeding with further analysis.

Loading and Inspection of the Data

```
##| echo: true #This can be added to selectively show specific code chunks.
# Load dataset
data_path <- file.path(dirname(dirname(here())), "Data", "fraud.csv")
df <- fread(data_path)
```

Dimensions of the Data Set

```
# Dimensions of the dataset
cat("The dataset contains", num_rows, "rows and", num_cols, "columns.\n")
```

```
The dataset contains 498121 rows and 10 columns.
```

Underlying Data Types

```r
# Display the data type of each attribute
column_info <- sapply(df, class)

# Print column names with their data types
for (col_name in names(column_info)) {
  print(paste("Column:", col_name, "- Data type:", column_info[col_name]))
}
```

```
[1] "Column: trustLevel - Data type: integer"
[1] "Column: totalScanTimeInSeconds - Data type: integer"
[1] "Column: grandTotal - Data type: numeric"
[1] "Column: lineItemVoids - Data type: integer"
[1] "Column: scansWithoutRegistration - Data type: integer"
[1] "Column: quantityModifications - Data type: integer"
[1] "Column: scannedLineItemsPerSecond - Data type: numeric"
[1] "Column: valuePerSecond - Data type: numeric"
[1] "Column: lineItemVoidsPerPosition - Data type: numeric"
[1] "Column: fraud - Data type: integer"
```

Summary of the whole data set

```r
summary(df)
```

```
   trustLevel    totalScanTimeInSeconds   grandTotal    lineItemVoids
 Min.   :1.000   Min.   :   1.0         Min.   : 0.00   Min.   : 0.000
 1st Qu.:2.000   1st Qu.: 458.0         1st Qu.:24.93   1st Qu.: 3.000
 Median :4.000   Median : 916.0         Median :50.03   Median : 5.000
 Mean   :3.503   Mean   : 915.6         Mean   :49.99   Mean   : 5.496
 3rd Qu.:5.000   3rd Qu.:1374.0         3rd Qu.:75.02   3rd Qu.: 8.000
 Max.   :6.000   Max.   :1831.0         Max.   :99.99   Max.   :11.000
 scansWithoutRegistration quantityModifications scannedLineItemsPerSecond
 Min.   : 0.000           Min.   :0.000         Min.   : 0.000546
 1st Qu.: 2.000           1st Qu.:1.000         1st Qu.: 0.008682
 Median : 5.000           Median :2.000         Median : 0.016940
 Mean   : 5.001           Mean   :2.499         Mean   : 0.068054
 3rd Qu.: 8.000           3rd Qu.:4.000         3rd Qu.: 0.033929
 Max.   :10.000           Max.   :5.000         Max.   :30.000000
 valuePerSecond     lineItemVoidsPerPosition     fraud
 Min.   : 0.00000   Min.   : 0.0000          Min.   :0.00000
 1st Qu.: 0.02735   1st Qu.: 0.1600          1st Qu.:0.00000
 Median : 0.05455   Median : 0.3529          Median :0.00000
 Mean   : 0.22218   Mean   : 0.7352          Mean   :0.04763
 3rd Qu.: 0.10909   3rd Qu.: 0.6923          3rd Qu.:0.00000
 Max.   :99.71000   Max.   :11.0000          Max.   :1.00000
```

Checking for missing values and duplicates

```
cat("The dataset contains", sum(missing_values), "missing values and ", duplicate_count, " du
```

```
The dataset contains 0 missing values and  0  duplicates.
```