Data Cleaning Report

Raw Data Information

Column Name	Non-Null Counts
Recnum	97,852
Cardnum	97,852
Date	97,852
Merchnum	**94,455**
Merch description	97,852
Merch state	**96,649**
Merch zip	**93,149**
Transtype	97,852
Amount	97,852
Fraud (Label)	97,852

Shape: (97852, 10)

Merchnum, Merch description, Merch zip are the 3 fields containing null values.

Outlier Removal

Field: Amount

The outlier threshold of the dataset is $30275 (3 std away from the mean). Recnum 53179 has transaction amount $3,102,045. After discussing with our business manager, we decided to remove this anomaly from the model.

Total transaction amount: 97852 - 1 = **97851**

Data Exclusions

Field Name: Transtype

There are 4 categories in Transtype: P(97,497), A(181), D(173), Y(1). Since the business manager can't identify the meaning of A, D, Y, we decided to remove data that are Transtype A, D, Y.

Total transaction amount: 97851 - 355 = **97496**

Discovering potential null values

Field: Merchnum

In the field Merchnum, there are 59 transactions that have a 0 as the Merchandise number. We should consider them null values as well.

Fields to be filled:

Column Name	Null Counts
Merchnum	3,279
Merch state	1,028
Merch zip	4,347

Null Values Imputation for the fields

Field: Merchnum

1) There are **3,279** null values for Merchnum. The following steps set forth the imputation process for Merchnum.
2) Recover Merchnums that have Merch description with corresponding Merchnums that are displayed on other non-null data. This action took care of 1,164 records, leaving **2,115** null values remaining.
3) Merch description 'RETAIL CREDIT ADJUSTMENT' and 'RETAIL DEBIT ADJUSTMENT' do not have corresponding Merchnums. Replace these null values to 'unknown'. This action took care of 694 records, leaving **1,421** null values remaining.
4) For the remaining 1,421 records without Merchnum, there are 515 unique Merch descriptions, indicating that the remaining transactions are mostly evenly distributed. We decided to give each of the 515 unique Merch descriptions a new Merch num, winding down the number of missing values in Merchnum to **0**.

Field: Merch state

1) There are **1,028** null values for Merch state. The following steps set forth the imputation process for Merch state.
2) Recover Merch states that have Merch zip information. We can align back the Merch state information if the zip codes are provided. There're 22 zip codes that are not in the not-null data but can be manually searched. With 4,567 unique Merch zip that have Merch state information. We took care of **74** records, leaving **954** null values remaining.
3) We replicated the same strategy again with Merchnum and Merch description. We took care of 2 records, leaving 952 null values remaining.
4) Same as the third step in the imputation process for Merchnum, there are several transactions involving 'RETAIL CREDIT ADJUSTMENT' and 'RETAIL DEBIT ADJUSTMENT'. We decided to set their Merch state as 'unknown'. We took care of **655** records, leaving **297** null values remaining.
5) For the remaining 297 null values, we fill the Merch state value with 'unknown'.
6) (Additional verification) Since determining whether it is a foreign transaction can be pivotal to identify fraud, we replaced the value in Merch state with 'foreign' if it is not a US State, null value, nor 'unknown'.

Field: Merch zip

1) There are **4,347** null values for Merch zip. The following steps set forth the imputation process for Merch zip.
2) Recover Merch zips that have Merch num or Merch description information, tracing back the Merch zips from non-null values. We took care of **1,722** records, leaving **2,625** null values remaining.
3) For records with Merch state but not Merch zip, we placed the most populated zip codes in each state for null values. We took care of **1,437** records, leaving **1,188** null values remaining.
4) For the remaining null values in Merch zip, we decided to place 'unknown'.