Data Quality Report

1. Data Description

The data is a collection of real card transaction records from a US government organization. The goal is to find transaction fraud. There are 10 fields and 96735 records, including 2 numeric fields and 8 categorical fields. The data covers the time of 2010, from January 1, 2010 to December 31, 2010.

2. Summary Tables

- Numerical Table

Field name	% Populated	Min	Max	Mean	Stdev	% Zero
Date	100%	2010-01-01	2010-12-31	/	/	0.00
Amount	100%	0.01	3,102,045.53	427.885677	10,006.140302	0.00

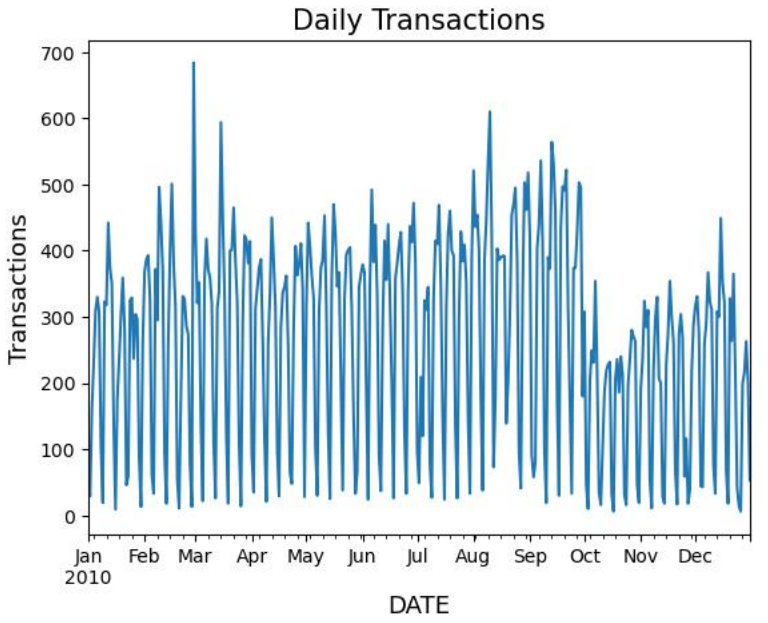
- Categorial Table

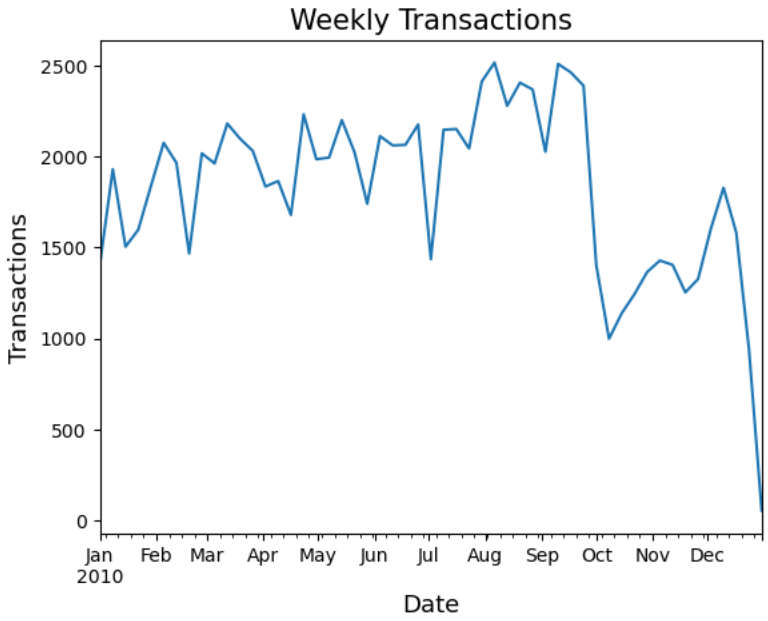
Field name	% Populated	# Unique Values	Most Common Value	#zeros	#blanks
Recnum	100%	96,753	1	0	0
Cardnum	100%	1,645	5142148452	0	0
Merchnum	96.5%	13,091	930090121224	231	3,375
Merch description	100%	13,126	GSA-FSS-ADV	0	0
Merch state	98.8%	227	TN	0	1,195
Merch zip	95.2%	4,567	38118.0	0	4,656
Transtype	100%	4	P	0	0
Fraud	100%	2	0	95,694	0

3. Field Exploration & Visualization

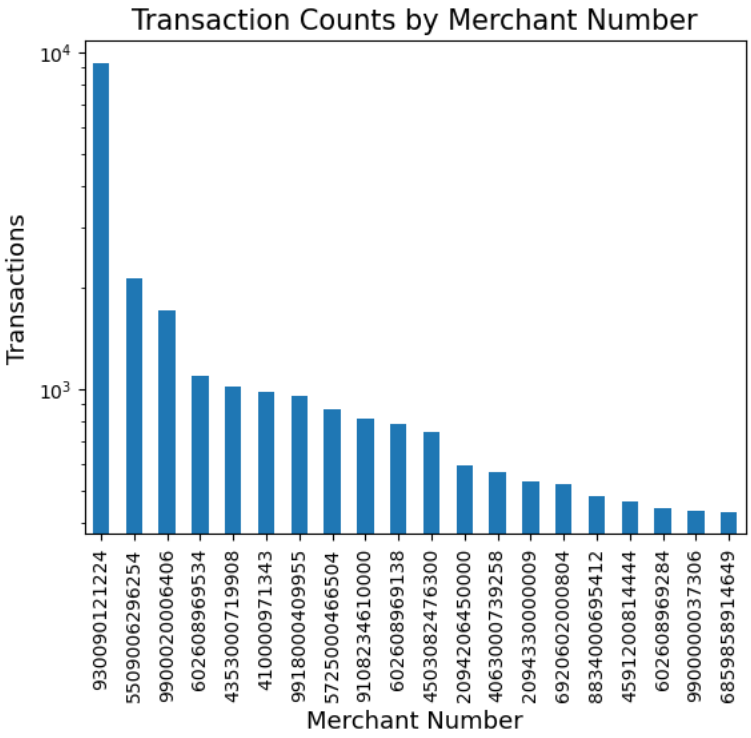
(1) Field Name: Recnum
Description: Record number; Ordinal unique positive integer for each record, from 1 to 96753.

(2) Field Name: Date
Description: Datetime field. The daily and weekly transaction distribution across time

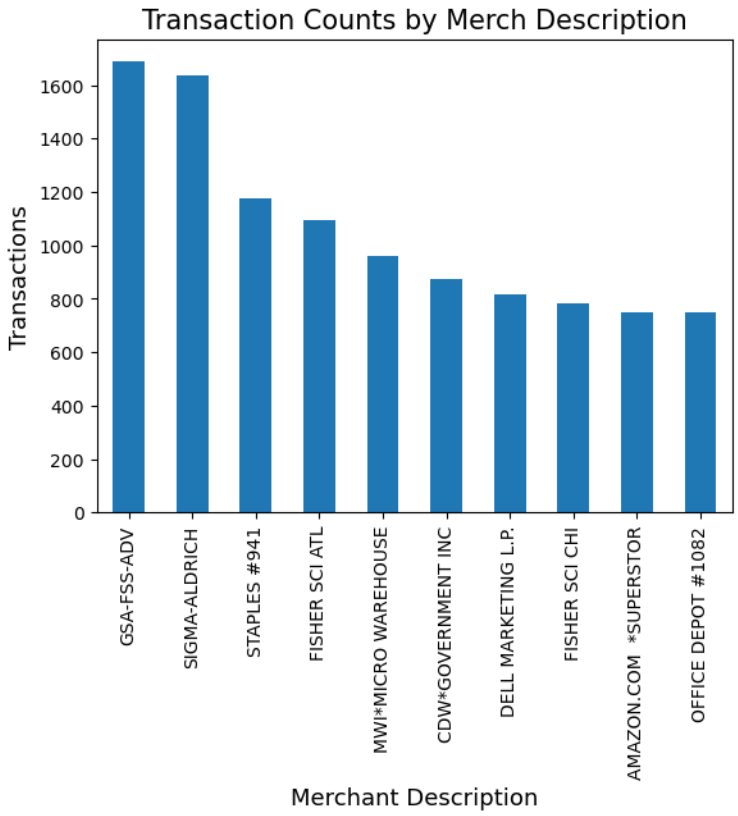




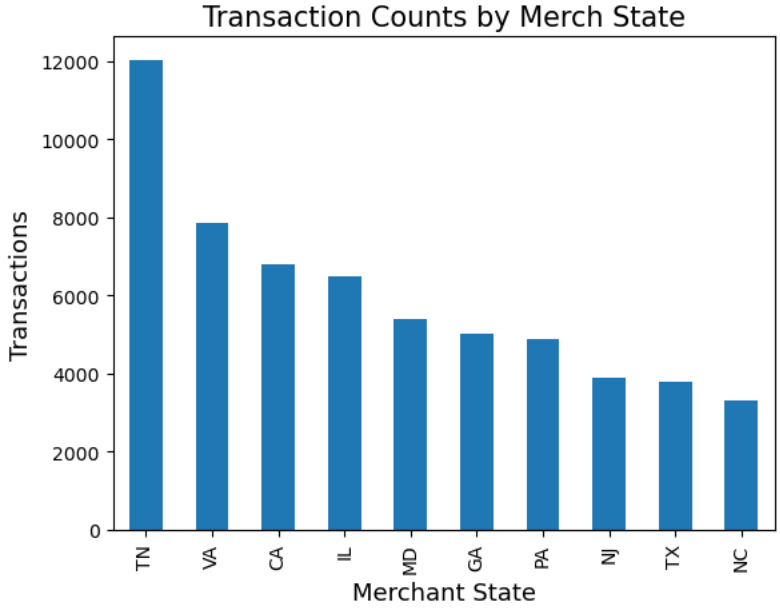
(3) **Field Name:** Merchnum
Description: The top 20 field values of the Merchnum
'Merchnum' is a categorical variable that represents the merchant number associated with each transaction in the dataset. There are missing values in this column.
The most commonly appearing merchant number is 930090121224, and the count is 9310.



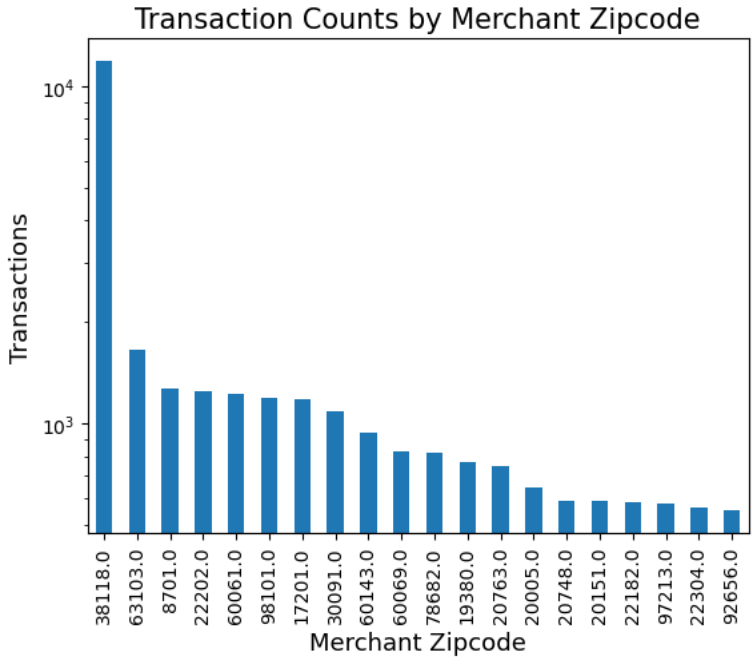
(4) Field Name: Merch description
Description: Top 10 field values of the Merch Description
'Merch descriprion' is a categorical variable that describes the merchant.
Most common Merch Description is 'GSA-FSS-ADV', and the count is 1688



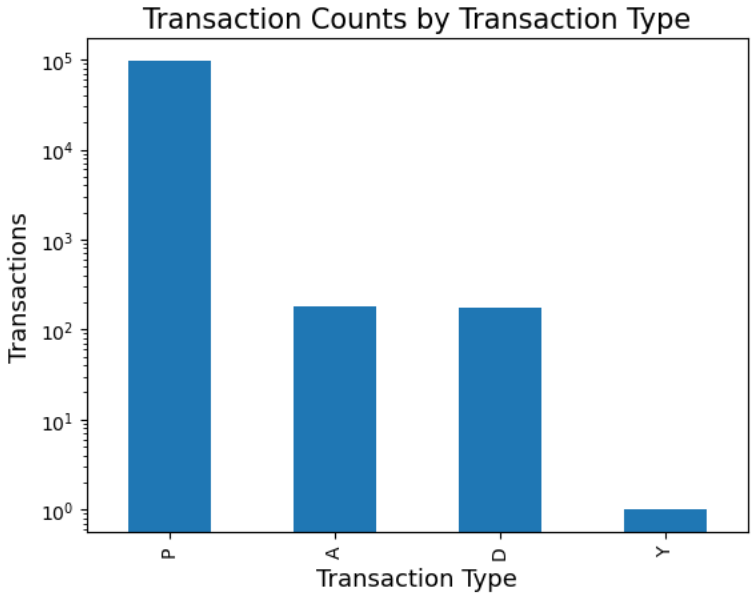
(5) **Field Name:** Merch state
Description: Top 10 field values of the Merch State
'Merch state' is a categorical variable that describes the state where the merchant from. Most common Merch State is TN, and the count is 12035



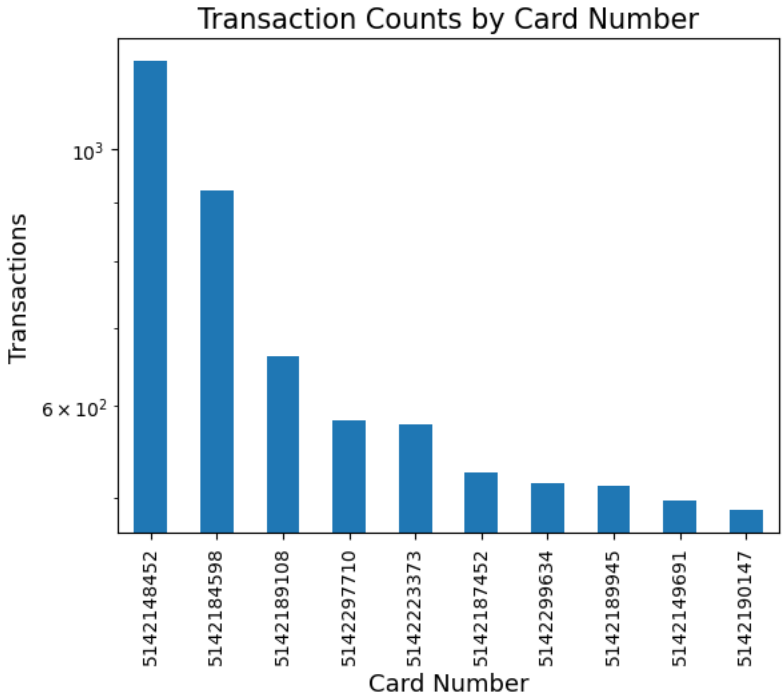
(6) **Field Name**: Merch zip
Description: Top 20 field values of the Merch Zip
'Merch zip' is a categorical variable that shows the zip code of the merchant.
Most common Merch Zip is 38188, and the count is 11868.



(7) **Field Name**: Transtype
Description: There are four types of transactions (P, A,D,Y), where P stands for 'purchase', A for 'authorization', D for 'debit', Y for 'year-end'.
Most common Transaction Type is P, and the count is 96398. The transaction type is extremely imbalanced.

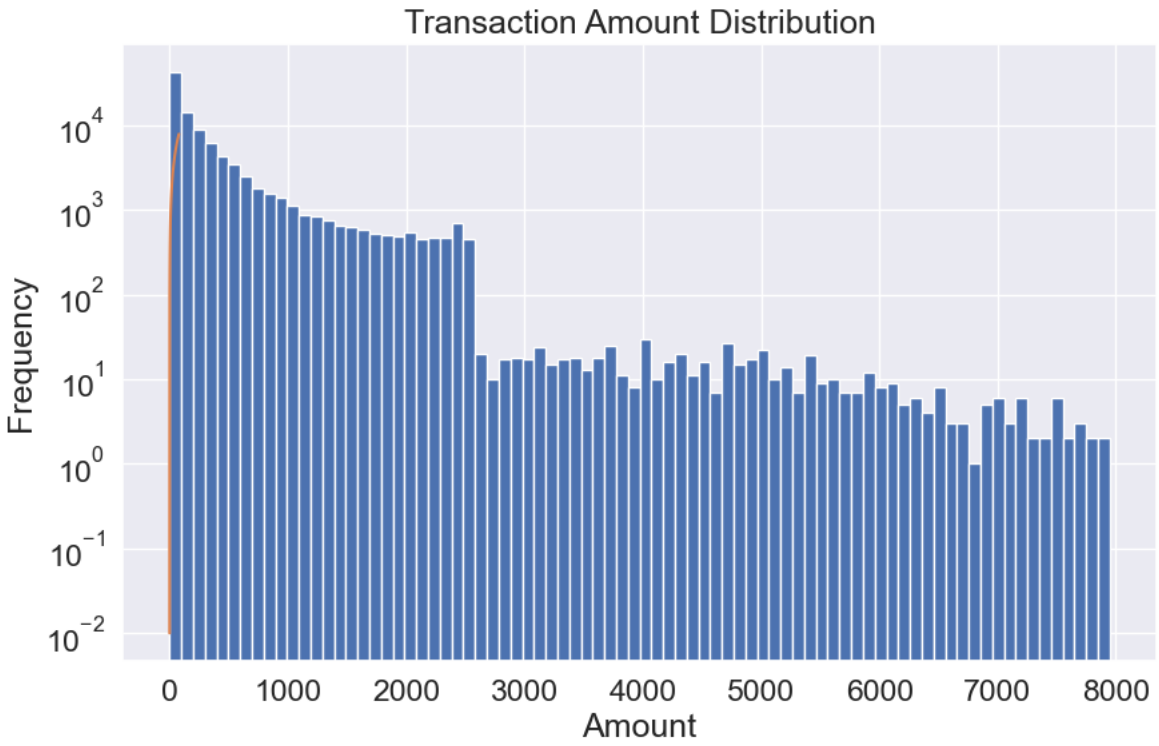


(8) **Field Name:** Cardnum
Description: The Top 10 field values of the Cardnum
'Cardnum' is a categorical variable that shows the card number of each transaction. Most common card number is 5142148452, and the count is 1192,

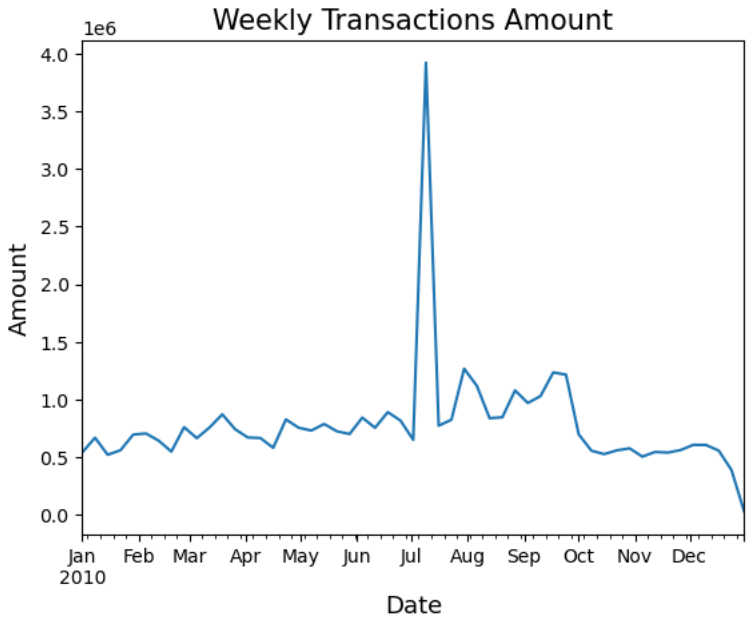


(9) **Field Name:** Amount
Description:
1: The distribution of amount ranges from 0 to 8000.

The min amount is 0.01, the mean amount for all the transactions is 427.89, and the highest amount recorded in the dataset is 3,102,045.53, which is considered an outlier in the field "transaction" and it is excluded in the graph below. Also, the amount tremendously decrease when approach 2500.



2: The weekly transaction amount distribution across time.
A numerical variable that shows the amount of each transaction. There outlier is shown during Jul and Aug.



- **Field Name:** Fraud

  Description: Fraud = 0 (non-fraud label), Fraud = 1 (fraud label)
  The count of non-fraud is 95694, fraud is 1059