USC Marshall
School of Business

Data Quality Report

Card Transaction Data

DSO 562 Fraud Analytics
Chutong Yan
March 24, 2020

Table of Contents

1. High-level File Description

The “card transaction.csv” file stores information of 96,753 transactions made by a Tennessee government organization in year 2010 (2010-01-01 to 2010-12-31). It covers 10 fields including record number, card number, date, merchant number, merchant description, merchant state, merchant zip code, transaction type, amount, and a computer-generated label specifying if the card transaction is a fraud. The dataset is shared by Professor Stephen Coggeshall in March 2020.

2. Summary Statistics Table

Besides *Recnum**, The dataset contains 8 categorial fields and 1 numerical field. The summary statistics of these fields are provided in Table 1 and 2.

Table 1 Summary Table of Categorical Fields

Name	Type	# Not null	% Populated	# Unique	# zero	Most Common Field Value
Cardnum	Categorical	96,753	100.00%	1,645	0	5142148452
Date	Datetime	96,753	100.00%	365	0	2010-02-28
Merchnum	Categorical	93,378	96.51%	13,091	0	930090121224
Merch description	Categorical	96,753	100.00%	13,126	0	GSA-FSS-ADV
Merch state	Categorical	95,558	98.76%	227	0	TN
Merch zip	Categorical	92,097	95.19%	4,567	0	38118
Transtype	Categorical	96,753	100.00%	4	0	P
Fraud	Categorical	96,753	100.00%	2	95,694	0

* Field *Recnum* is not specified in the table above because summary statistics of unique record key is not meaningful.

Table 2 Summary Table of Numerical Fields

Name	Type	# Not null	% Populated	# Unique	# Zero	Mean	St dev	Min	Max
Amount	Numerical	96,753	100.00%	34,909	0	427.89	10,006.14	0.01	3,102,045.53

3. Explorations of Fields

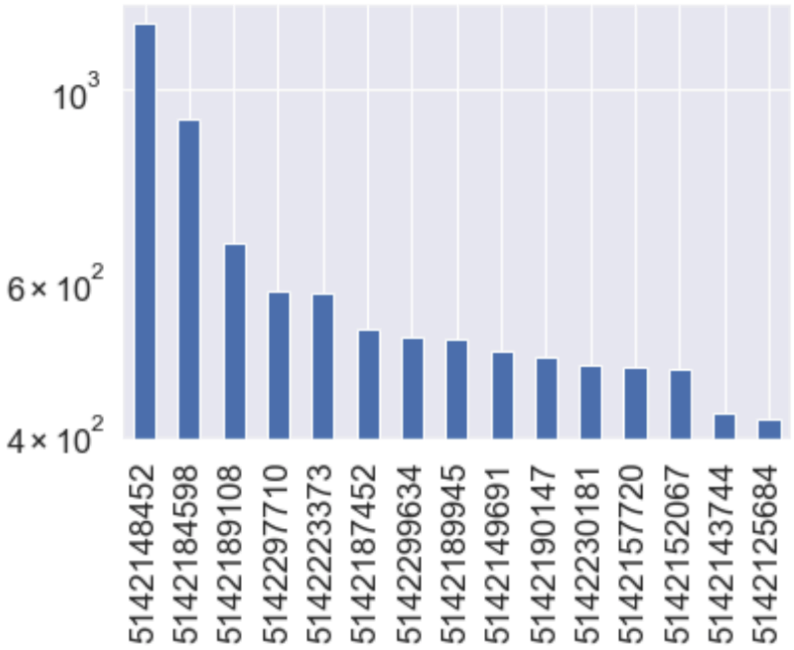
3.1 Recnum

Recnum field is the unique integer label for each record, ranging from 1 to 96,753.

3.2 Cardnum

Cardnum field represents the card numbers used in transactions. The bar plot below shows the frequency of the top 15 values in the *Cardnum* field.

Figure 1 Frequency of Top 15 values of field *Cardnum*



3.3 Date

Date field represents the date of card transaction. The bar plot (Figure 2) below shows the frequency of the top 20 values in the *Date* field. The line plot (Figure 3) indicates a significant decrease in number of transactions between September and October. This pattern is in line with the characteristics of government organizations. Staffs tends to be conservative upon receiving their budgets on October 1st (Start of Fiscal Year) and usually spend all of their left-over budget at the end of fiscal year (September 30th).

Figure 2 Frequency of Top 20 values of field *Date*

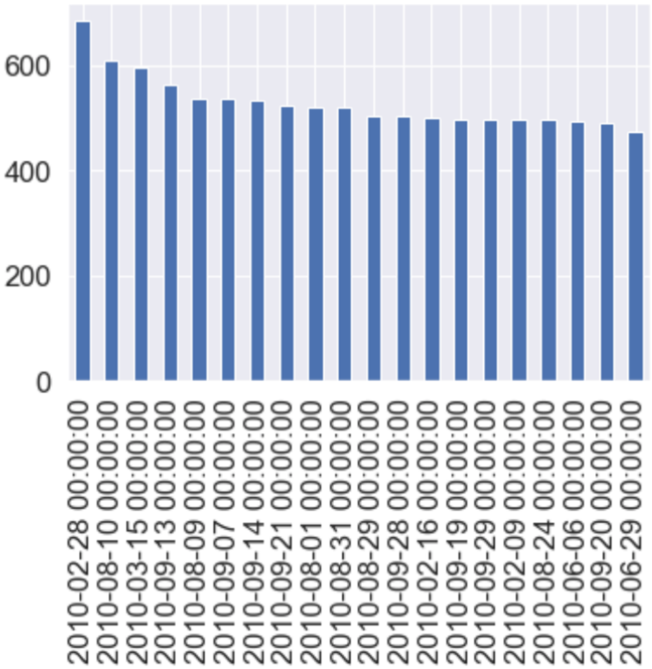
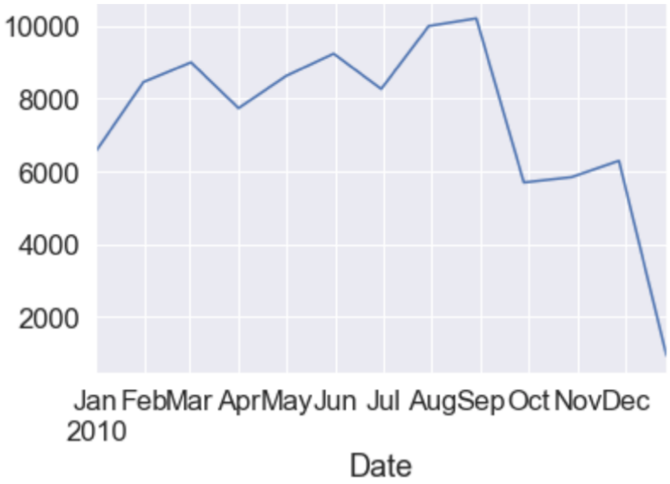


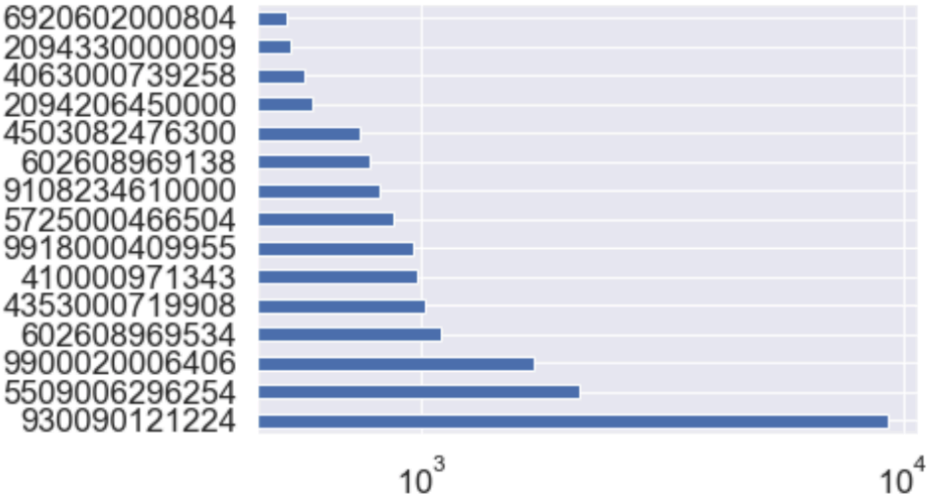
Figure 3 Number of Approximately Monthly Transactions



3.4 Merchnum

Merchnum field represents the unique id of merchants. The bar plot below shows the frequency of the top 15 values in the *Merchnum* field.

Figure 4 Frequency of Top 15 values of field *Merchnum*



3.5 Merch description

Merch description field represents the text description of merchants. The table below shows the frequency of the top 15 values in the *Merch description* field.

Table 3 Frequency of Top 15 values of *Merch description* field

Merchant Description	Frequency
GSA-FSS-ADV	1,688
SIGMA-ALDRICH	1,635
STAPLES #941	1,174
FISHER SCI ATL	1,093
MWI*MICRO WAREHOUSE	958
CDW*GOVERNMENT INC	872
DELL MARKETING L.P.	816
FISHER SCI CHI	783
AMAZON.COM *SUPERSTOR	750
OFFICE DEPOT #1082	748
VWR SCIENTIFIC PROD VCTS	688
PC *PC CONNECTION	570
C & C PRODUCT SERVICES	558
BUY.COM	481
FISHER SCI HUS	442

3.6 Merch state

Merch state field represents the states where the merchants are located in. The table below shows the frequency of the top 15 values in the *Merch state* field.

Table 4 Frequency of Top 15 values of *Merch state* field

Merchant State	Frequency
TN	12,035
VA	7,872
CA	6,817
IL	6,508
MD	5,398
GA	5,025
PA	4,899
NJ	3,912
TX	3,790
NC	3,322
WA	3,300
DC	3,208
OH	3,131
NY	2,430
MO	2,420

3.7 Merch zip

Merch zip field represents the zip codes of merchants. The table below shows the frequency of the top 15 values in the *Merch zip* field.

Table 5 Frequency of Top 15 values of *Merch zip* field

Merchant Zip Code	Frequency
38118	11,868
63103	1,650
8701	1,267
22202	1,250
60061	1,221
98101	1,197
17201	1,180
30091	1,092
60143	942
60069	826
78682	817
19380	769
20763	749
20005	648
20748	592

3.8 Transtype

Transtype field represents the type of transactions, where P = purchase. The table below shows the frequency of values in the *Transtype* field.

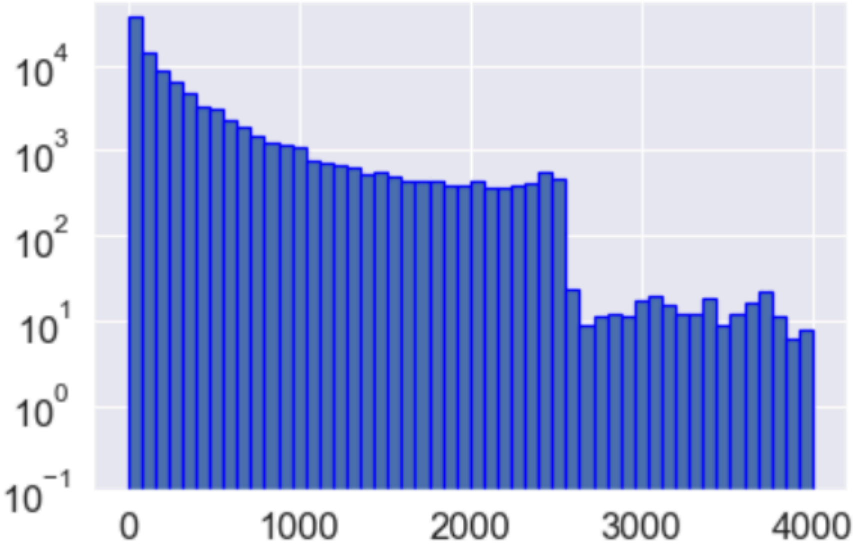
Table 6 Frequency of values in field *Transtype*

Transaction Type	Frequency
P	96,398
A	181
D	173
Y	1

3.9 Amount

Amount field represents the amount of transactions. The plot below shows the distribution of values in *Amount* field (omitted values greater than 4,000, that is 486 records and 0.5% of total valid records in this field).

Figure 5 Distribution of field *Amount*



3.10 Fraud

Fraud field represents whether or not a card transaction is identified as fraud, where 1 = fraud and 0 = not fraud. The table below shows the frequency of values in the *Fraud* field.

Table 7 Frequency of values in field *Fraud*

Fraud Label	Frequency
0	95,694
1	1,059