DATA QUALITY REPORT

[SHUANG LIANG] [6790038156]

1. Summary of Data:

File description:
The data is a collection for card transactions in year 2010. This data file includes 96753 records of credit card transactions containing 10 fields: Record number, Card number, Merchant number, Date, Merchant description, Merchant state, Merchant zip, Transaction type, Transaction amount and Fraud.

File Name: Card transactions

Number of Records: 96,753 records

Number of Fields: 10 variables in total –except for record number, there are 8 categorical variables, 1 numeric variable. Detailed information of variables is in table 'categorical' and table 'numerical'.

Time period of the records: Year 2010

Numerical:

Field Name	Type	Missing values	Percentage Populated	Unique Values	# of 0	Mean	Standard Deviation	Max	Min
Amount	numerical	0	1	34,909	0	428	10,006	3,102,046	0

Categorical:

Field	Type	Missing values	Percentage Populated	Unique Values	No. of 0	Most Common Field Value
Cardnum	Categorical	0	100.00%	1645	0	5142148452
Date	Categorical	0	100.00%	365	0	2/28/2010
Merchnum	Categorical	3,375	96.51%	13,091	0	9.3009E+11
Merch description	Categorical	0	100.00%	13,126	0	GSA-FSS-ADV
Merch state	Categorical	1,195	98.76%	227	0	TN
Merch zip	Categorical	4,656	95.19%	4,567	0	38118
Transtype	Categorical	0	100.00%	4	0	P
Fraud	Categorical	0	100.00%	2	0	0

2. Detailed analysis of individual field

Field 1

Field Name: Cardnum

Description:

Cardnum is a categorical variable representing the credit card number. It has 10 digits.

Unique Values:

Cardnum has 1645 unique values. No missing values exist. The distribution is shown below, the most frequent value is marked with a *.

5142148452*	1192
5142184598	921
5142189108	663
5142297710	583
5142223373	579
5142187452	526
5142299634	515
5142189945	512

Field 2

Field Name: Date

Description:

Date is a categorical variable representing the credit card transaction date.

Unique Values:

Date has 365 unique values ranging from 1 to 5. No missing values exist. The distribution is shown below, the most frequent value is marked with a *.

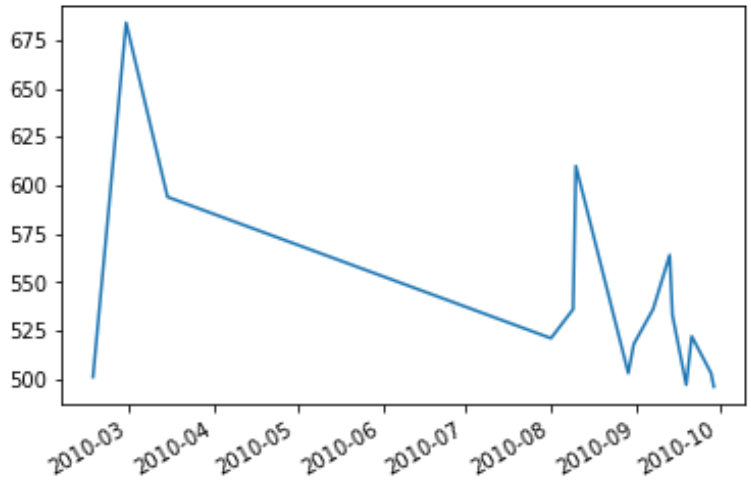
Distribution on the date level:

2010-02-28	684
2010-08-10	610
2010-03-15	594
2010-09-13	564
2010-08-09	536
2010-09-07	536
2010-09-14	533
2010-09-21	522

Distribution on the month level:

Month	Count
8	11050

9	9895
3	9421
6	9249
5	8982
7	8296
2	7756
4	7731
1	6810
12	6653
11	5801
10	5109



Field 3

Field Name: Merchnum

Description:

Merchnum is a categorical variable representing merchant number.

Unique Values:

Merchnum has 13,091 unique values and 3,375 missing values.

930090121224	9310
5509006296254	2131
9900020006406	1714
602608969534	1092
4353000719908	1020
410000971343	982
9918000409955	956
5725000466504	872

For records that are fraud:

4353000719908	107
930090121224	50
8834000695423	46
4503738417400	45
4620009957157	39
900009045549	36
618901687330	36
253052983001	33

Field 4

Field Name: Merch description

Description:

Merch description is a categorical variable, which describes the merchant information.

Unique Values:

Merch description has 13126 unique values. No missing values exist.

Field 5

Field Name: Merch State

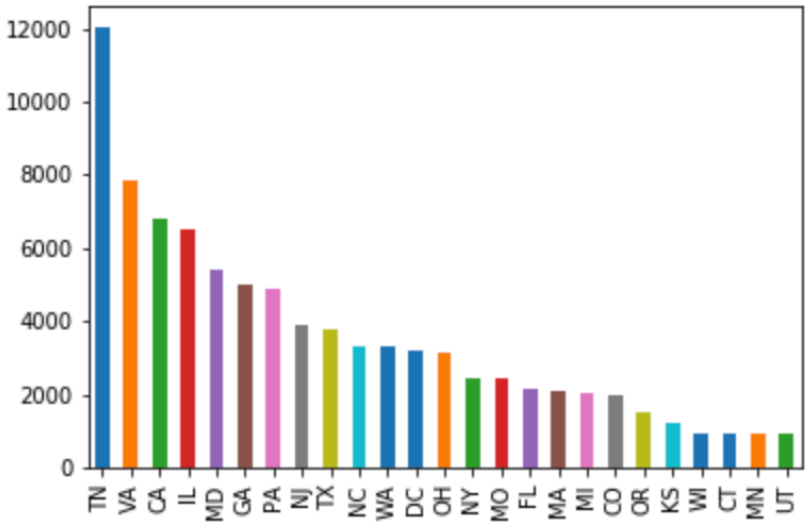
Description:

Merch State is a categorical variable describing which state merchants are in.

Unique Values:

Merch State has 227 unique values. No zeros exist in this field. There are 1195 missing values. The statistics and distribution are shown below: * is most common one.

State	Count	State	Count
TN*	12035	WA	3300
VA	7872	DC	3208
CA	6817	OH	3131
IL	6508	NY	2430
MD	5398	MO	2420
GA	5025	FL	2143
PA	4899	MA	2081
NJ	3912	MI	2033
TX	3790	CO	1987
NC	3322	OR	1510



Field 6

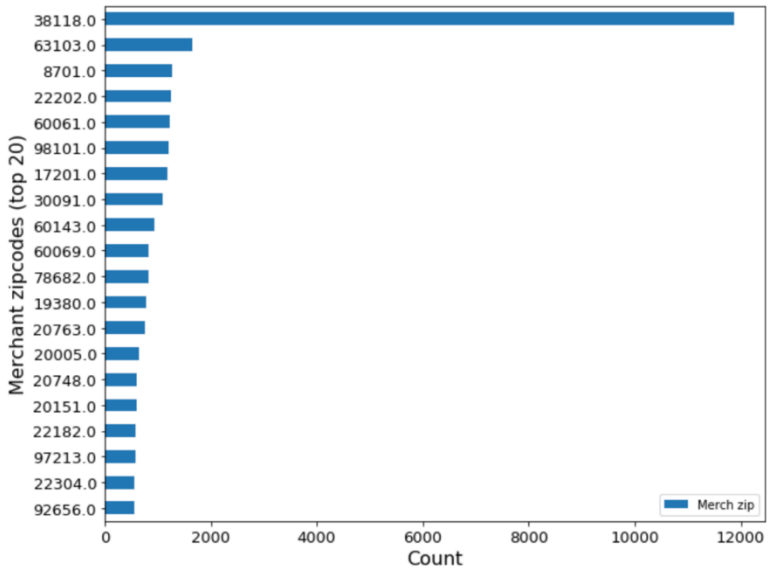
Field Name: Merch zip

Description:

Merch zip is a categorical variable describing zip code of the place where the transaction took place.

Unique Values:

Merch zip has 863348 unique values. There are 4656 missing values. The top 20 zips and times they appear are:



Field 7

Field Name: Transtype

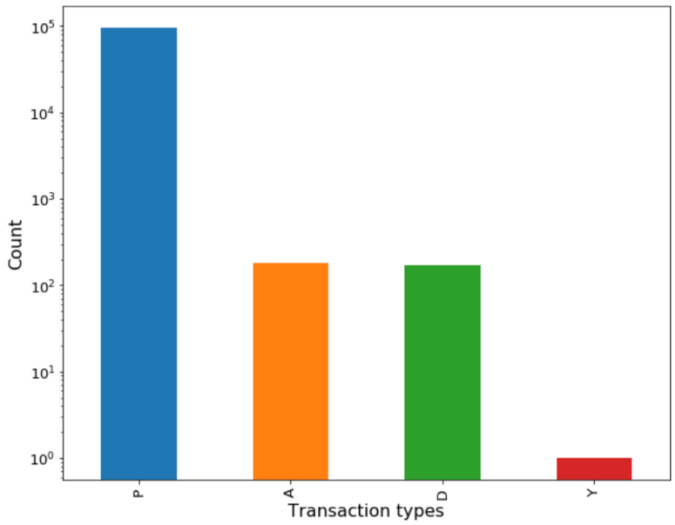
Description:

Transtype is a categorical variable describing types of transactions.

Unique Values:

Transtype has 4 unique values. There are no missing values. The categories are:

P*	96398
A	181
D	173
Y	1



Field 8

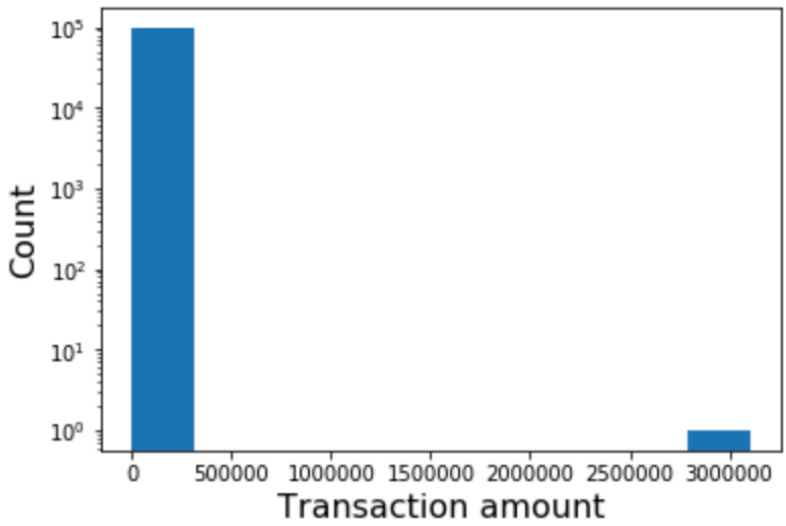
Field Name: Amount

Description:

Amount is a numerical variable describing transaction amount.

Unique Values:

Amount has 34909 unique values. No zeros in this field. No missing values exist. The statistics and distribution are shown below:



Field 9

Field Name: Fraud

Description:

Fraud is a categorical variable describing whether the transaction is a fraud.

Unique Values:

Fraud has 2 unique values (0 and 1). There are 95,692 transactions with 0 in the dataset. No missing values exist.

The statistics and distribution are shown as below:

0	95694
1	1059