Homework 7: Fraud Analytics
Data Quality Report: Credit Card Transactions Data

1. Data Description

Credit Card Transactions Data contains details for credit card transactions for the period of one year from January 1, 2010 to December 31, 2010. We will use this data to create a supervised fraud model to detect credit card transactions fraud.

Following are some of the characteristics of this dataset:

- Dataset Name: Credit Card Transactions Data
- Number of Records/Rows: 96753
- Number of Fields/Columns: 10
- 9 fields in the dataset are categorical whereas only one is numerical (Amount).
- Each record in the dataset represents a credit card transaction and the characteristics of the transaction have been captured in fields such as Card number, Date, Transaction Type, Merchant number, Amount, Merchant Description, Zip code of merchant etc. The fraud label field indicates whether the record represents a fraudulent transaction or not. Here, 1 corresponds to a fraudulent activity whereas 0 does not.

2. Summary of Data Fields

9 fields are categorical whereas only one is numerical.

Table 1: Summary Characteristics of All Fields

Variable	# records with value	Type	% populated	# unique values	#records with zero as a value	Mean	SD	Min Value	Max Value	Most common value (MCV)	Freq. of MCV
Recnum	96753	Categorical	100%	96753	0	NA	NA	NA	NA	NA	NA
Cardnum	96753	Categorical	100%	1645	0	NA	NA	NA	NA	5142148452	1192
Date	96753	Categorical	100%	365	0	NA	NA	NA	NA	2010-02-28	684
Merchnum	93378	Categorical	96.5%	13092	0	NA	NA	NA	NA	930090121224	9310
Merch description	96753	Categorical	100%	13126	0	NA	NA	NA	NA	GSA-FSS-ADV	1688
Merch State	95558	Categorical	98.76%	227	0	NA	NA	NA	NA	TN	12035
Merch Zip	92097	Categorical	95.18%	4568	0	NA	NA	NA	NA	38118	11868
Transtype	96753	Categorical	100%	4	0	NA	NA	NA	NA	P	96398
Amount	96753	Numerical	100%	34909	0	4.27e+02	1.0e+04	1.0e-02	3.1e+06	NA	NA
Fraud	96753	Categorical	100%	2	95694	NA	NA	NA	NA	0	95694

3. Description and Exploration of Data Fields

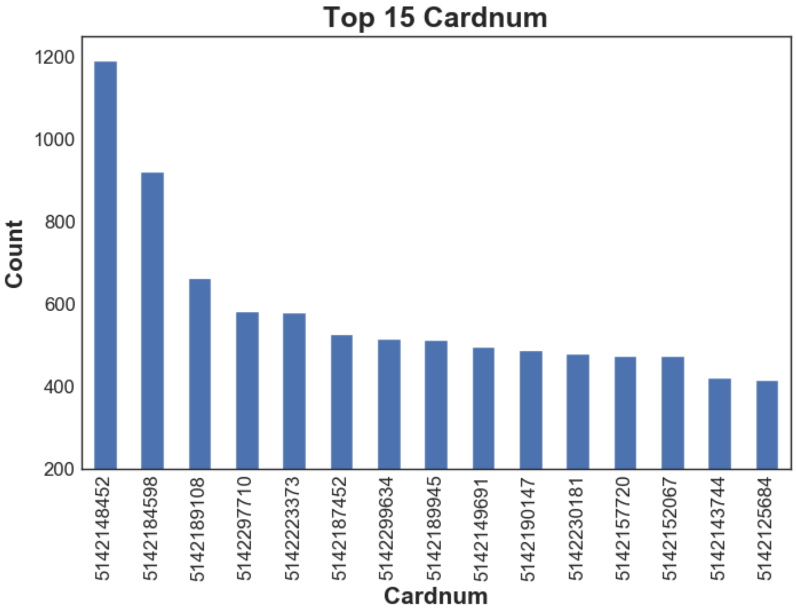
We now discuss in detail each of the above fields with a written description and a visual representation (wherever possible):

3.1 Field Name: Recnum

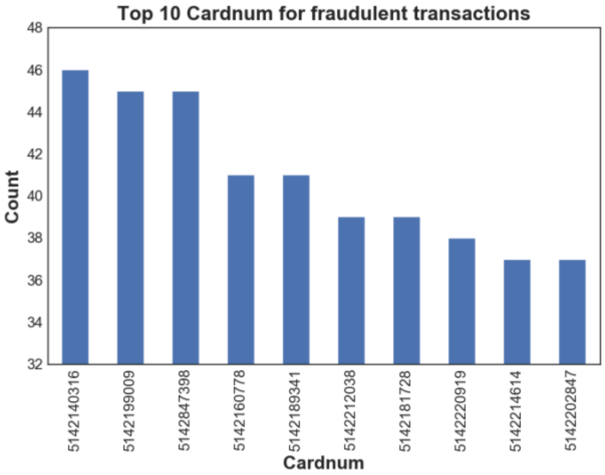
Description: This field represents a unique identifier for each record or row in the dataset. Values in this field have a unique number for each of the 96753 rows in the dataset.

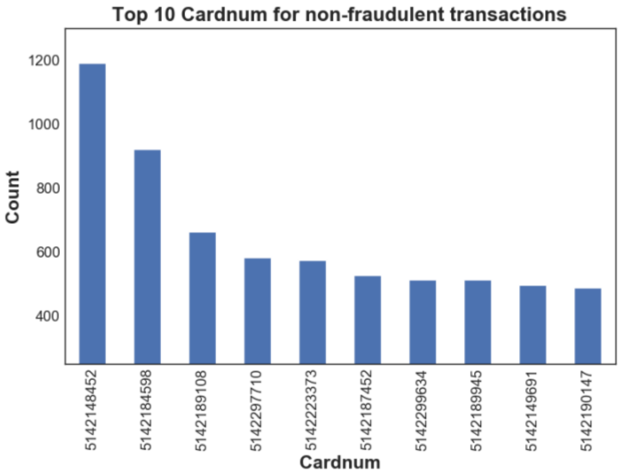
3.2 Field Name: Cardnum

Description: This is a categorical field and corresponds to the credit card number used for each transaction. There are 1645 unique values and there are no missing values for this field. We can see that the card number '5142148452' is repeated 1192 times in the dataset. The top 15 card numbers according to the count of transactions are shown as follows in the table below.



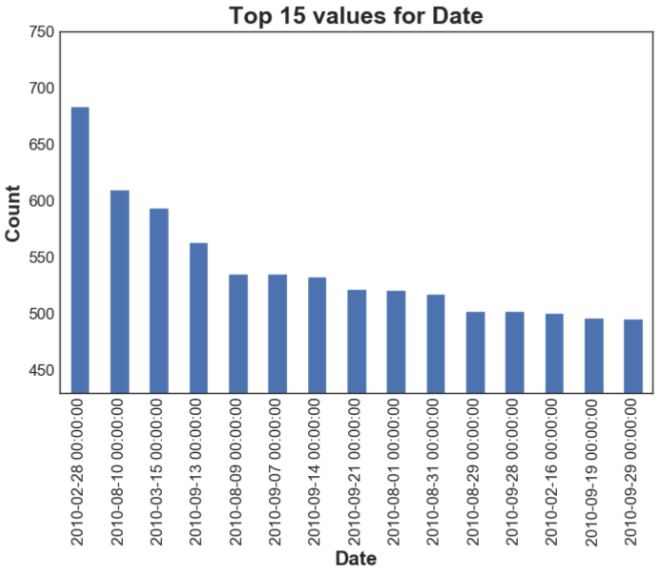
The following graphs show the count of top 10 credit card numbers used for fraudulent as well non-fraudulent transactions.



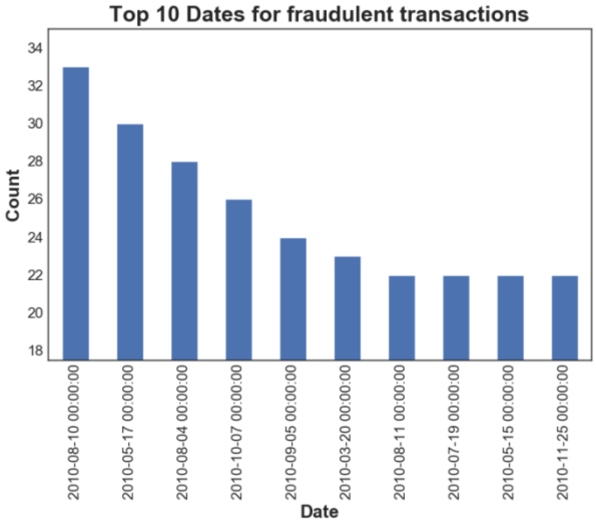


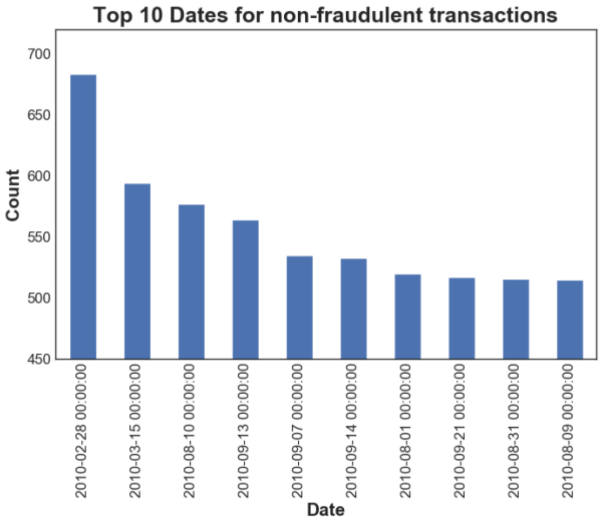
3.3 Field Name: Date

Description: This field contains the Date of the transaction and is of type datetime. There are 365 unique values corresponding to each day of the year 2010.



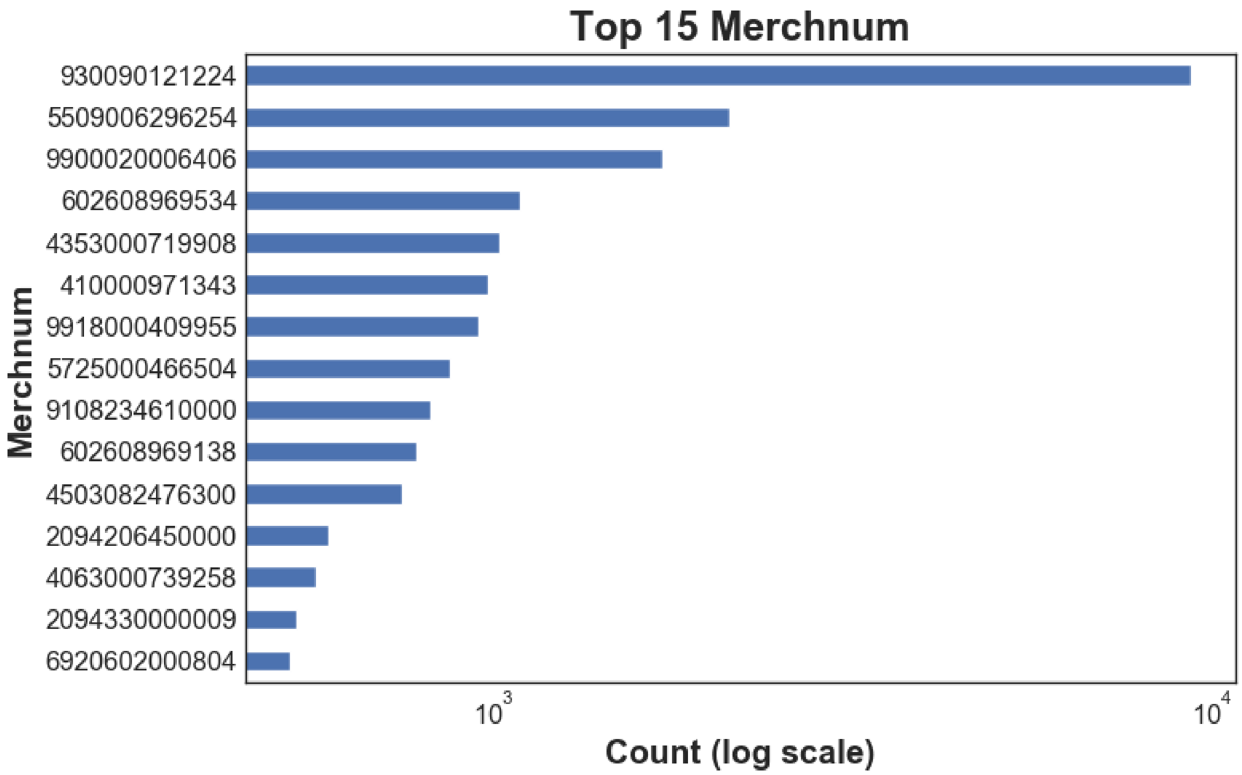
The following graphs show the count of top 10 dates for fraudulent as well non-fraudulent transactions.





3.4 Field Name: Merchnum

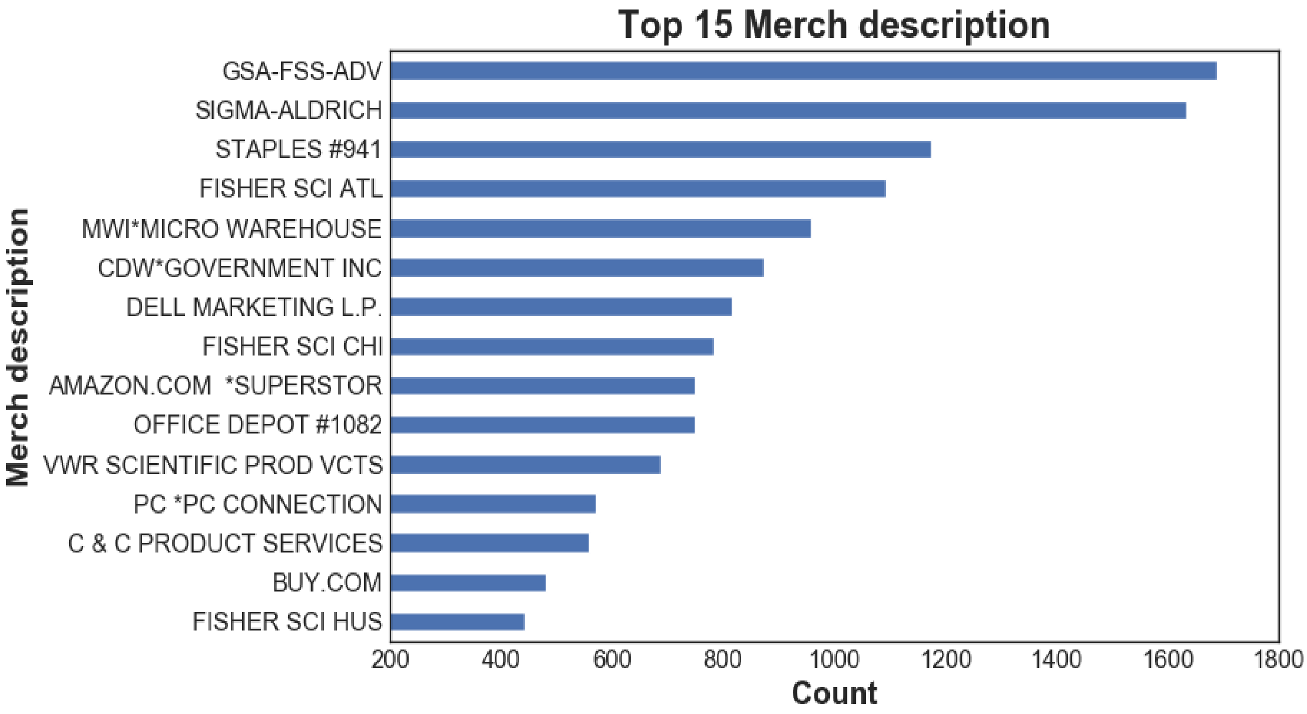
Description: This is a categorical field and around 96.5% values are filled. The bar graph and the table show top 15 Merchnum by count.



Merchnum	Count
930090121224	9310
5509006296254	2131
9900020006406	1714
602608969534	1092
4353000719908	1020
410000971343	982
9918000409955	956
5725000466504	872
9108234610000	817
602608969138	783
4503082476300	746
2094206450000	590
4063000739258	568
2094330000009	533
6920602000804	523

3.5 Field Name: Merch description

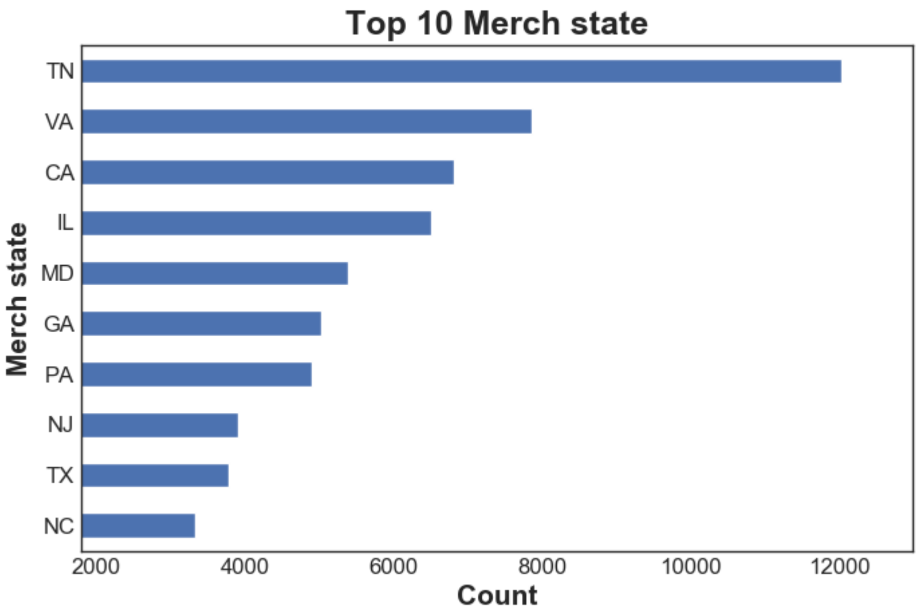
Description: This field contains the merchant description and is a categorical field. It gives information about the location of each merchant. There are 13126 unique values, and it is 100% populated. The bar graph shows the top 15 Merch description by Count.



Merch description	Count
GSA-FSS-ADV	1688
SIGMA-ALDRICH	1635
STAPLES #941	1174
FISHER SCI ATL	1093
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 Field Name: Merch state

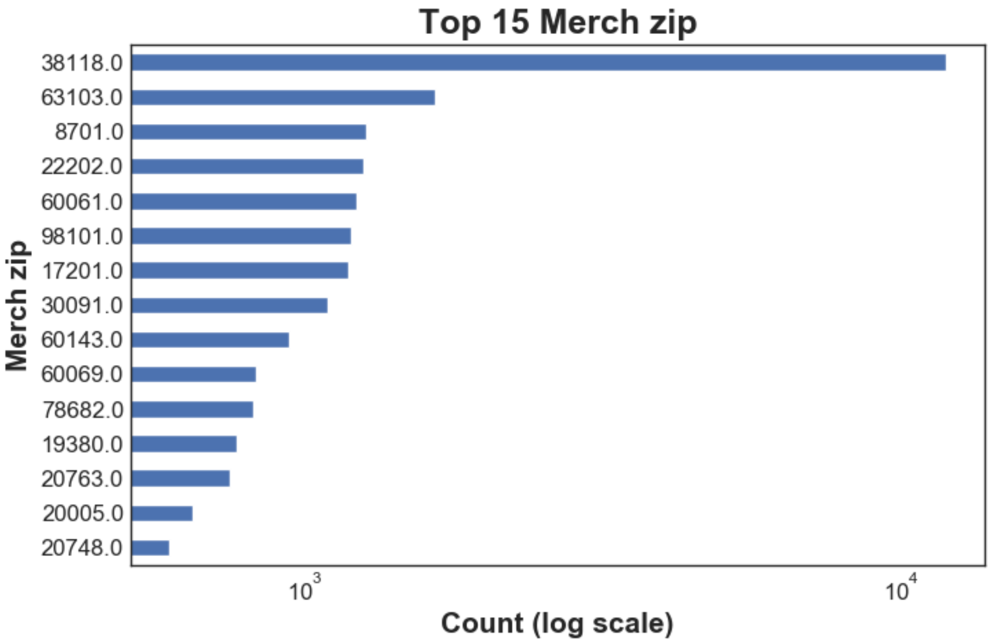
Description: This field contains the merchant state which indicates the location of each merchant. The bar graph shows the top 10 Merchant states by count.



Merchant state	Count
TN	12035
VA	7872
CA	6817
IL	6508
MD	5398
GA	5025
PA	4899
NJ	3912
TX	3790
NC	3322

3.7 Field Name: Merch zip

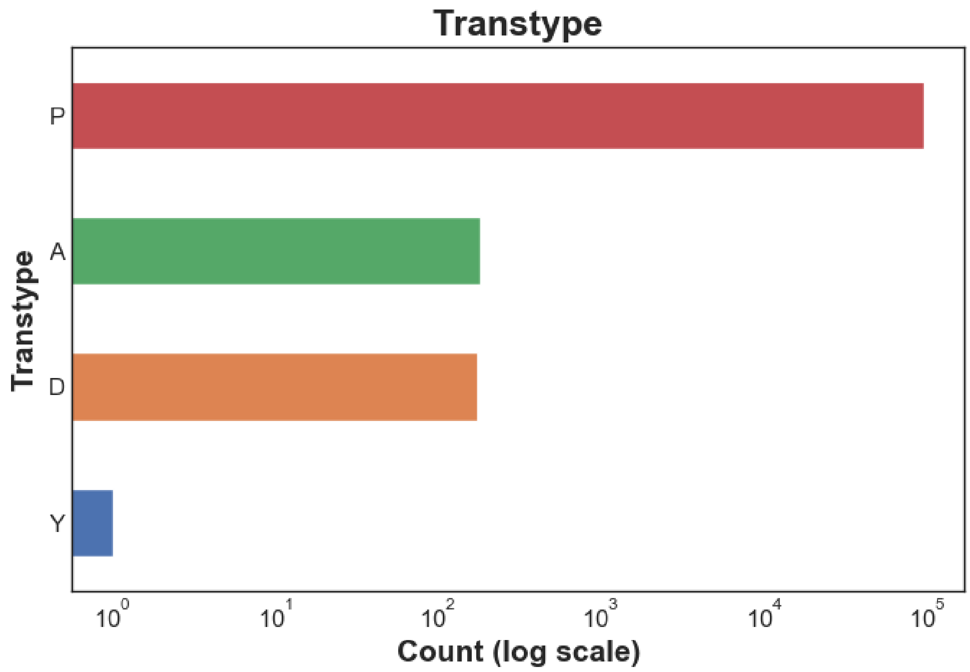
Description: This is a categorical field contains the zip codes of the merchants for each transaction. It has 4568 unique values in the dataset. The bar graph shows the top 15 values by count (log scale).



Merch zip	Count
38118	11868
63103	1650
8701	1267
22202	1250
60061	1221
98101	1197
17201	1180
30091	1092
60143	942
60069	826
78682	817
19380	769
20763	749
20005	648
20748	592

3.8 Field Name: Transtype

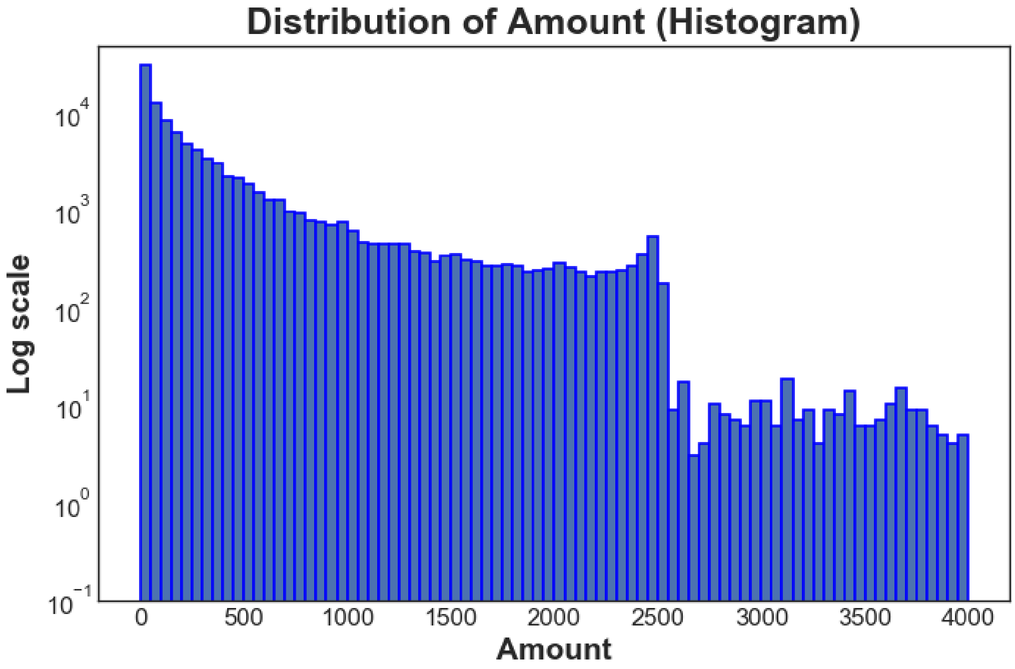
Description: This field is categorical and contains the type of each transaction corresponding to each record in the dataset. There are 4 types of transactions in the dataset. We note that transtype ‘P’ is the most common with a frequency of 96398. The bar graph shows count (log scale) of each transtype.



Transtype	Count
P	96398
A	181
D	173
Y	1

3.9 Field Name: Amount

Description: This is a numeric field and contains the amount of the particular transaction. The histogram shows the distribution of this field on a log y scale.

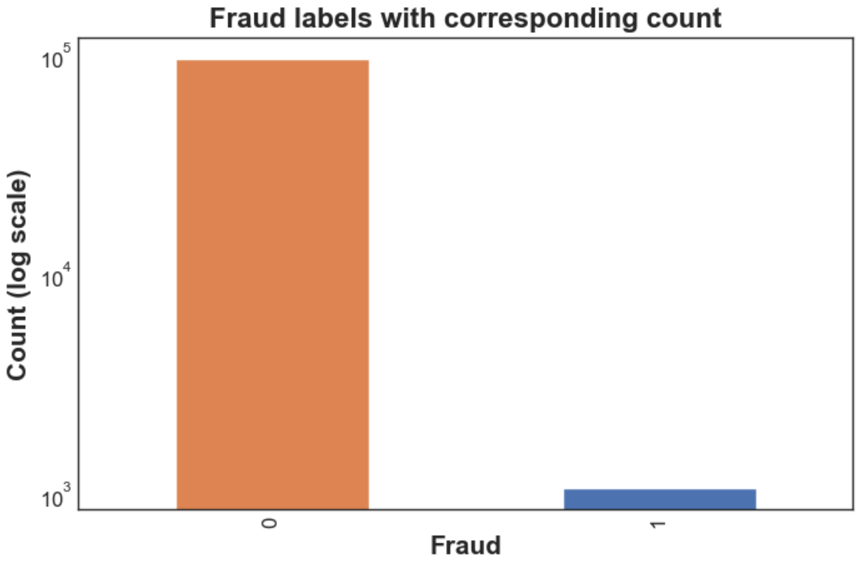


Count	9.675300e+04
Mean	4.278857e+02
Std	1.000614e+04
Min	1.000000e-02
25%	3.348000e+01
50%	1.379800e+02
75%	4.282000e+02
Max	3.102046e+06

3.10 Field Name: Fraud

Description: This field contains the label which indicates whether the particular record is identified as fraud or not. It has two unique values, 0 which indicates the particular record is not fraud whereas 1 indicates that the record is fraudulent. There are 95694 zeros and 1059 ones. This shows that around 1.1% of the dataset consists of fraudulent transactions. Below is the log distribution of these two categories.

Value	Count	Percent
0	95694	98.9%
1	1059	1.1%


Fraud labels with corresponding count
Count (log scale)
10^5
10^4
10^3
Fraud