Data Quality Report, Project 3

Chengjun Liu, 7116099552

# Data Description

Dataset Name: Card Transaction

Release Date: N/A

Number of Columns: 10

Number of Records: 96753

Summary

Overview:

|  |  |  |  |
| --- | --- | --- | --- |
| Variable Name | Data Type | Non-null Records | Description |
| Recnum | Integer | 96753 | Number of records |
| Cardnum | Integer | 96753 | Credit card number |
| Date | Datetime | 96753 | Transaction date |
| Merchnum | String | 93378 | Identity of merchants |
| Merch description | String | 96753 | Description of merchants and transactions |
| Merch state | String | 95558 | Name of state where transaction machines are located |
| Merch zip | Float | 92097 | Zip code of where transaction machines are located |
| Transtype | String | 96753 | Type of transactions |
| Amount | Float | 96753 | Amount of money for transactions |
| Fraud | Integer(Binary) | 96753 | Whether transactions are fraud(1) or not(0) |

Because all variables in integers are regarded and analyzed as categorical variables, we include integer variables in the categorical variable table:

Summary of Numeric Variables:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Variable Name | Non-null Record | Non-null Percentage | Min | Max | Mean |
| Amount | 96753 | 100% | 0.01 | 3102046 | 427.8857 |

Summary of Categorical Variables:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Variable Name | Non-null Record | Non-null Percentage | Number of Unique Values | Most frequent category |
| Recnum | 96753 | 100% | 96753 | None |
| Cardnum | 96753 | 100% | 1645 | 5142148452 |
| Date | 96753 | 100% | 365 | 2010-02-28 |
| Merchnum | 93378 | 96.5117% | 13091 | 930090121224 |
| Merch description | 96753 | 100% | 13126 | GSA-FSS-ADV |
| Merch state | 95558 | 98.7649% | 12035 | TN |
| Merch zip | 92097 | 95.1877% | 4567 | 38118 |
| Transtype | 96753 | 100% | 4 | P |
| Fraud | 96753 | 100% | 2 | 0 |

Data Field Exploration

Variable 1

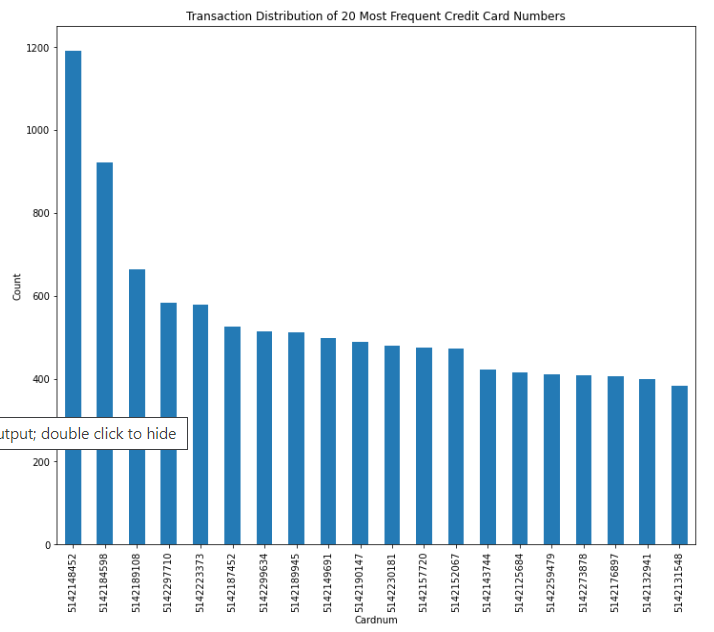
Name: Recnum

Description: Number of the unique mark of each transaction.

Variable 2

Name: Cardnum

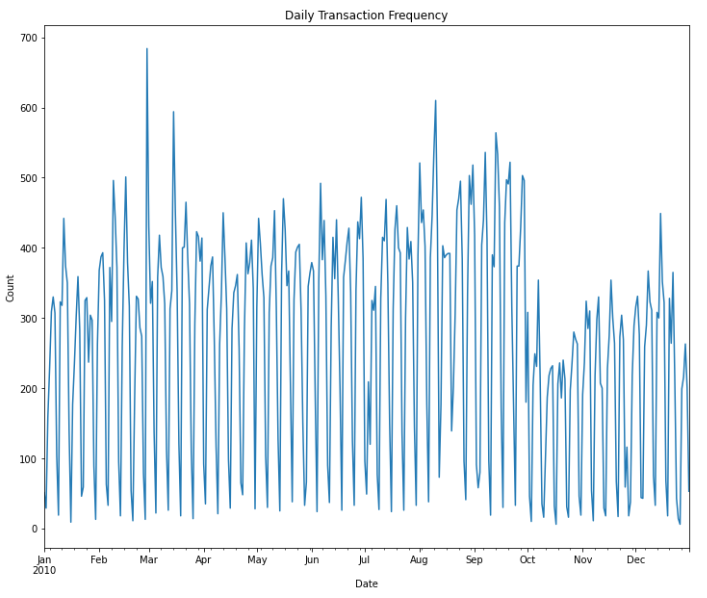
Description: Identity of credit card number in the transaction.

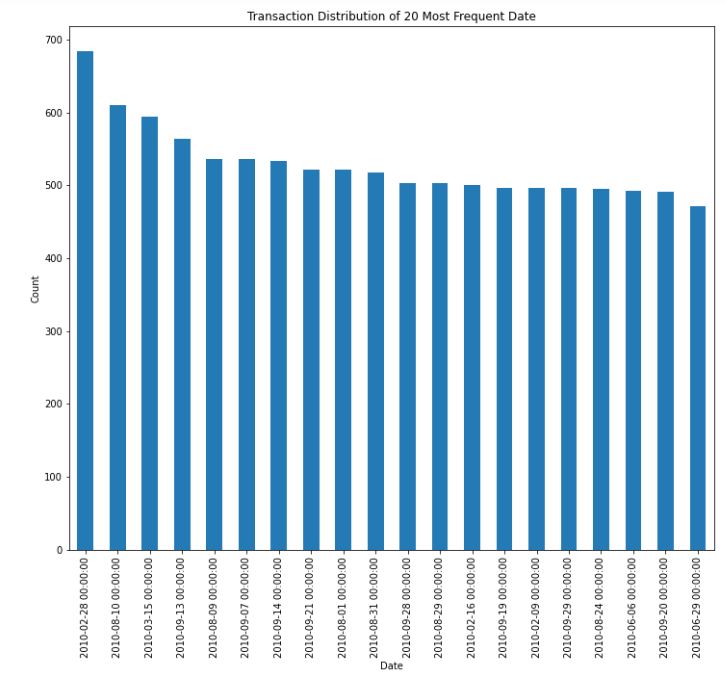


Variable 3

Name: Date

Description: Date when the transaction happens.

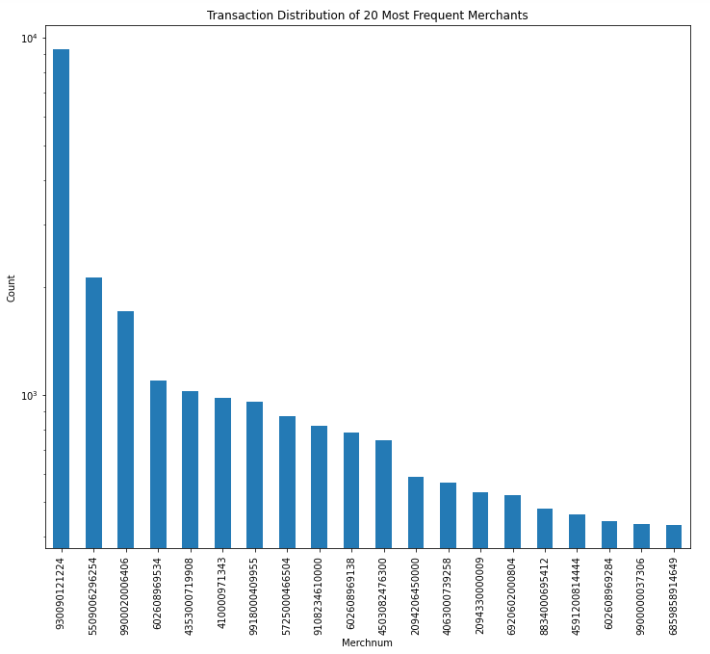
To show the distribution of the whole time and the most frequent dates, corresponding plots are here.



Variable 4

Name: Merchnum

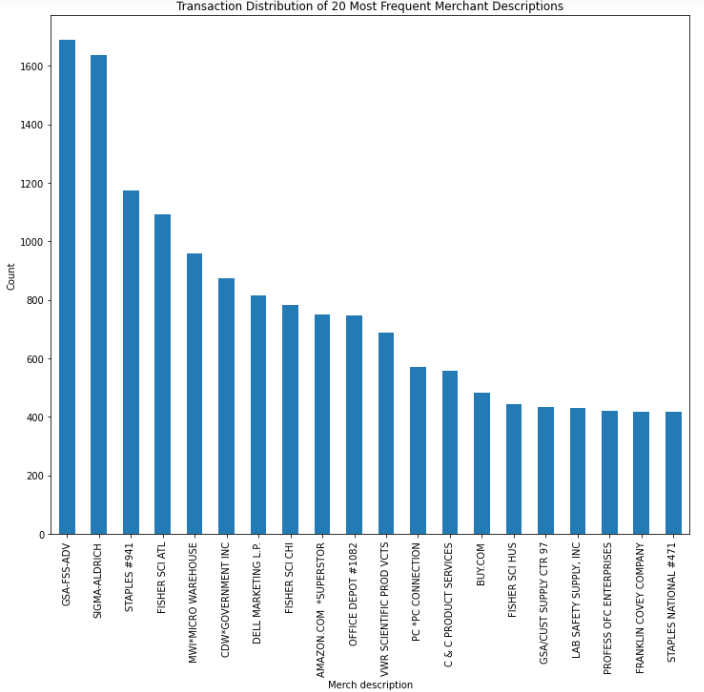
Description: Identity number of the merchant in the transaction.



Variable 5:

Name: Merch description

Description: Description of merchant, machine and transaction information

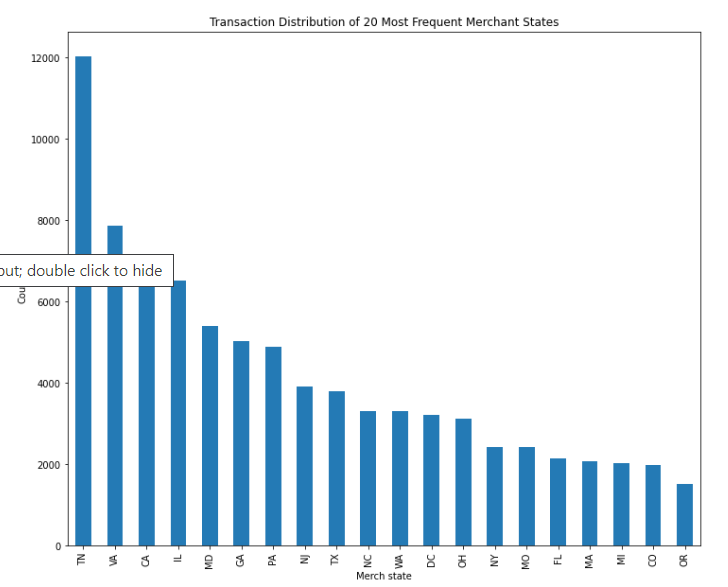


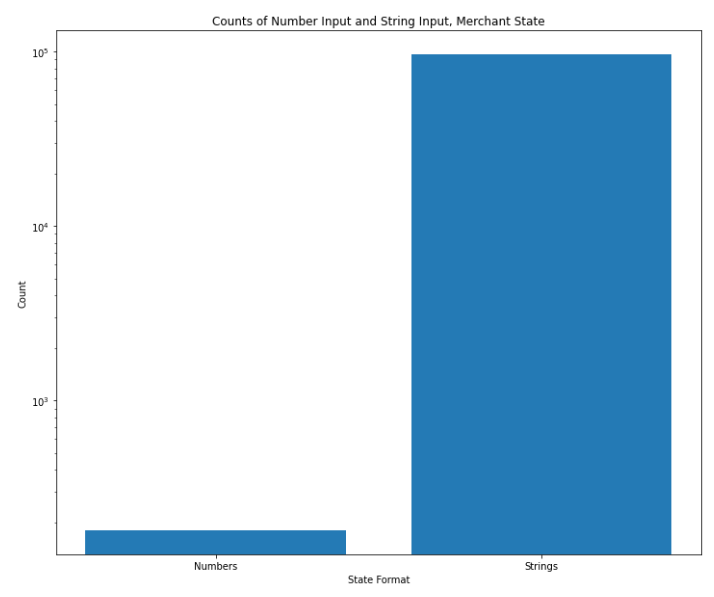
Variable 6

Name: Merch state

Description: Name of the state where merchant is located, and transaction happens.

Because apparent numbers are mixed in the records, a comparison of numbers and strings is below.

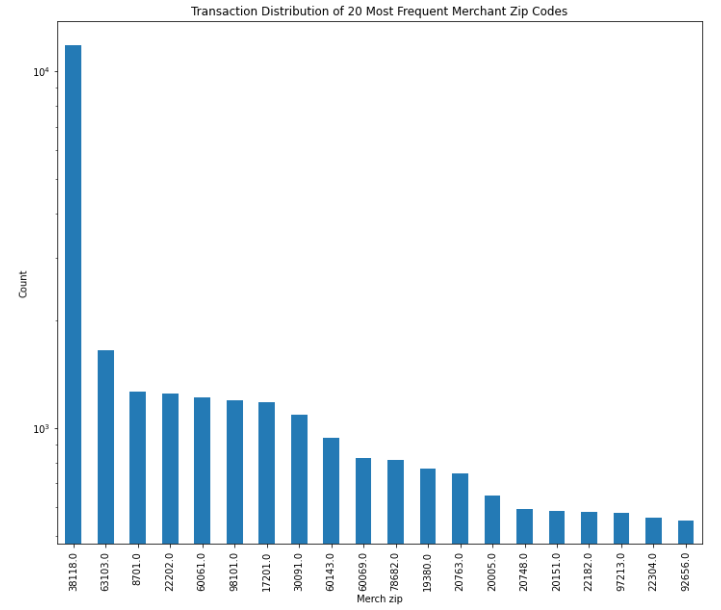




Variable 7

Name: Merch zip

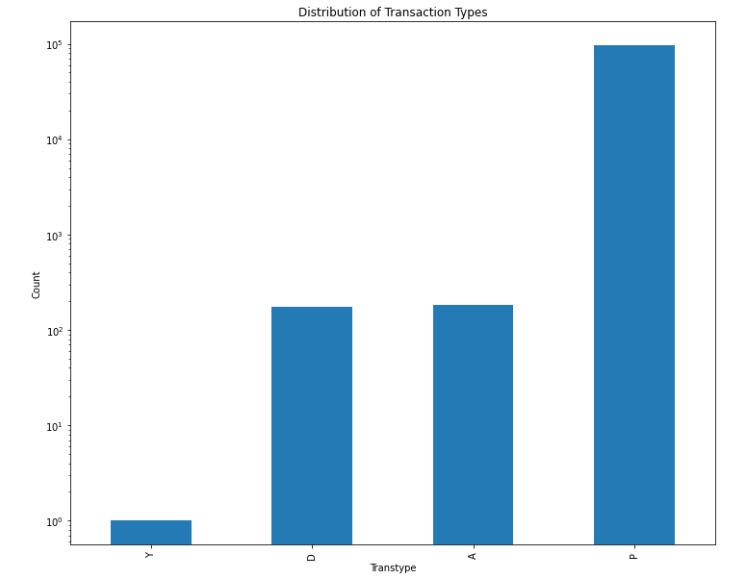
Description: Zip code of the address where merchant is located, and transaction happens.



Variable 8

Name: Transtype

Description: Type of the transaction.

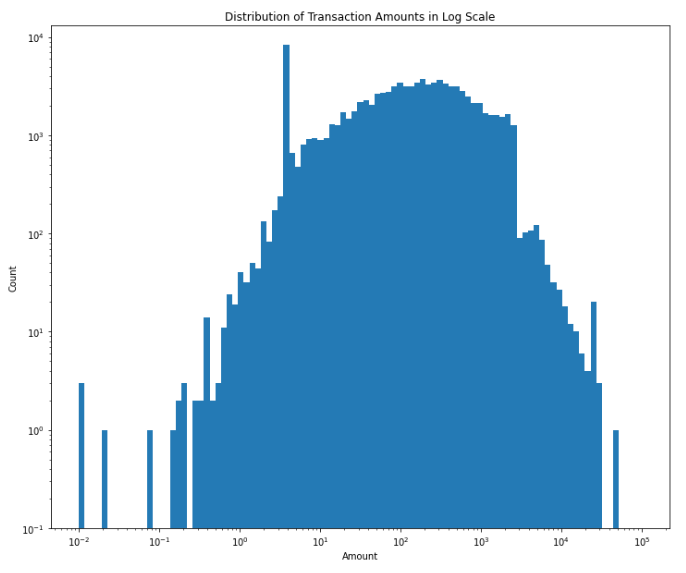


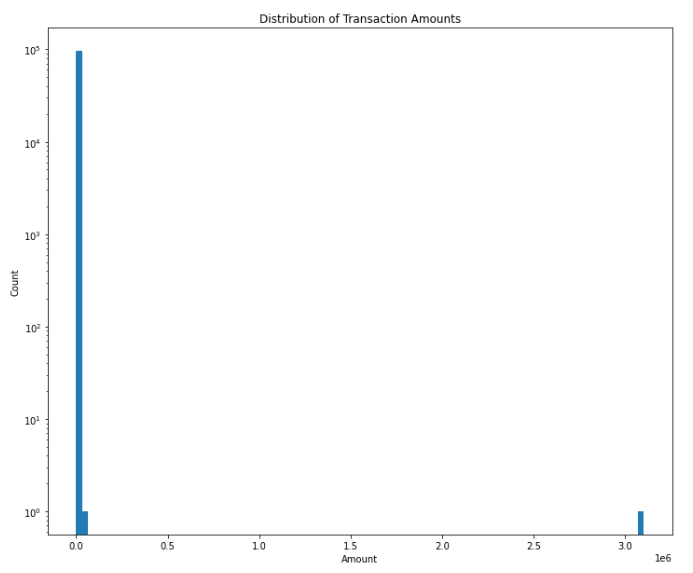
Variable 9

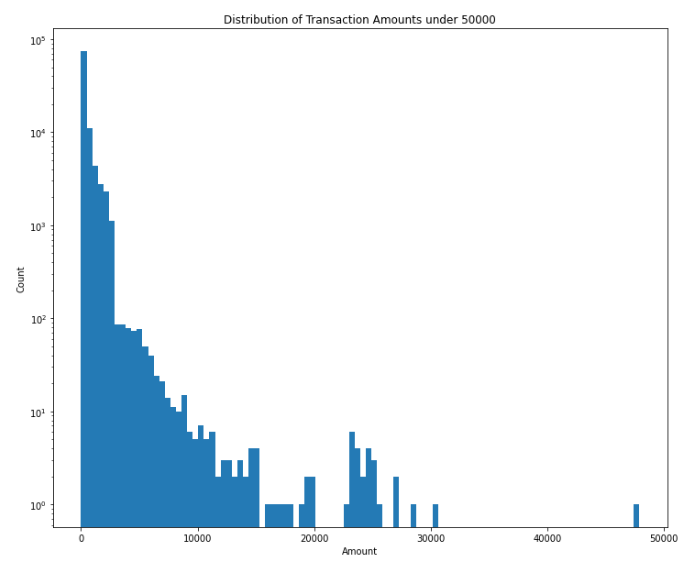
Name: Amount

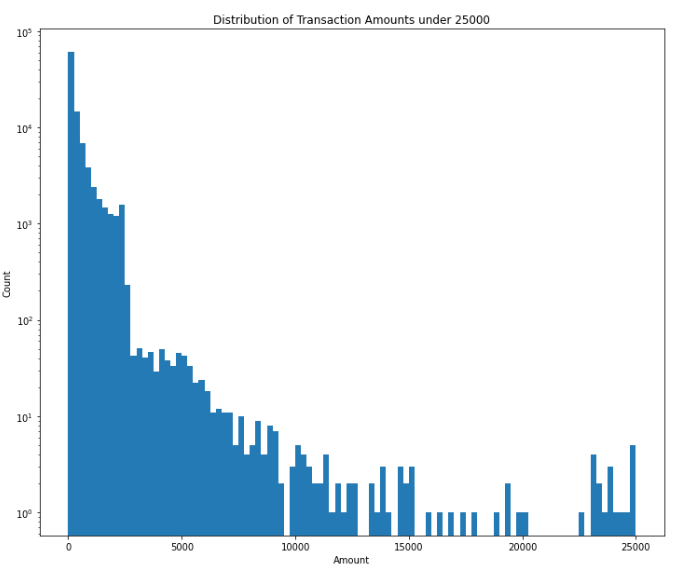
Description: Amount of payment in the transaction.

Because there are abnormal amounts in the records, different scales are applied in graphs.





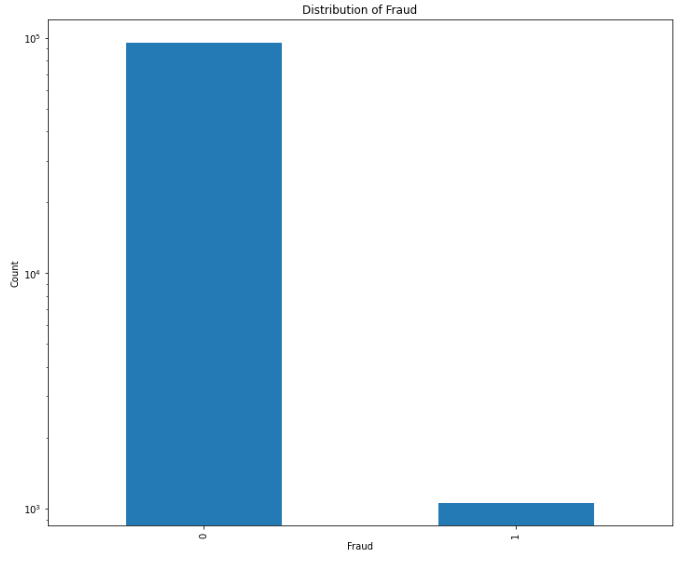




Variable 20

Name: Fraud

Description: Whether the transaction is a fraud (0 means no, 1 means yes).



In addition, to explore the relationship of fraud and date, we plot the distributions of fraud and no fraud and highlight dates with most fraud and no fraud.

