1. **Dataset Summary**

This credit card transaction dataset has **96,753 records**, **9 independent fields**, and **one dependent field,** which represent whether the record is fraud or not. This dataset is gathered from state of Tennessee. The time period is from 01/01/2010 to 12/31/2010.

1. **Field Summary**

**Categorical Field**

There are 8 categorical fields in this data set. Summary information for each categorical fields is shown as Table 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field | Count | % Populated | Unique Value | Most Common |
| Recnum | 96,753 | 100.00% | 96,753 | Uniform Distributed |
| Cardnum | 96,753 | 100.00% | 1,645 | 5142148452 |
| Date | 96,753 | 100.00% | 365 | 2/28/10 |
| Merchnum | 93,378 | 96.51% | 13,092 | 9.3009E+11 |
| Merch description | 96,753 | 100.00% | 13,126 | GSA-FSS-ADV |
| Merch state | 95,558 | 98.76% | 228 | TN |
| Merch zip | 92,097 | 95.19% | 4,568 | 38118 |
| Transtype | 96,753 | 100.00% | 4 | P |

**Table 1.**

**Numerical Field**

In this dataset there is only 1 numerical field. Table 2 represents some basic statistic summary for this field.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| field | count | %Populated | mean | std | min | 25% | 50% | 75% | max |
| Amount | 96,753 | 100.00% | 427.9 | 10,006 | 0.01 | 33.48 | 137.98 | 428.2 | 3102046 |

**Table 2.**

**Response Field**

“Fraud” is a response field in this dataset. There are two types, “0” and “1”. Type “0” categorized as non-fraud transaction and Type “1” represents fraud transaction. Summary information is shown as Table 3.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field | records | % Populated | Non-Fraud | Fraud |
| Fraud | 96,753 | 100.00% | 95694 (98.9%) | 1059 (1.1%) |

**Table 3.**

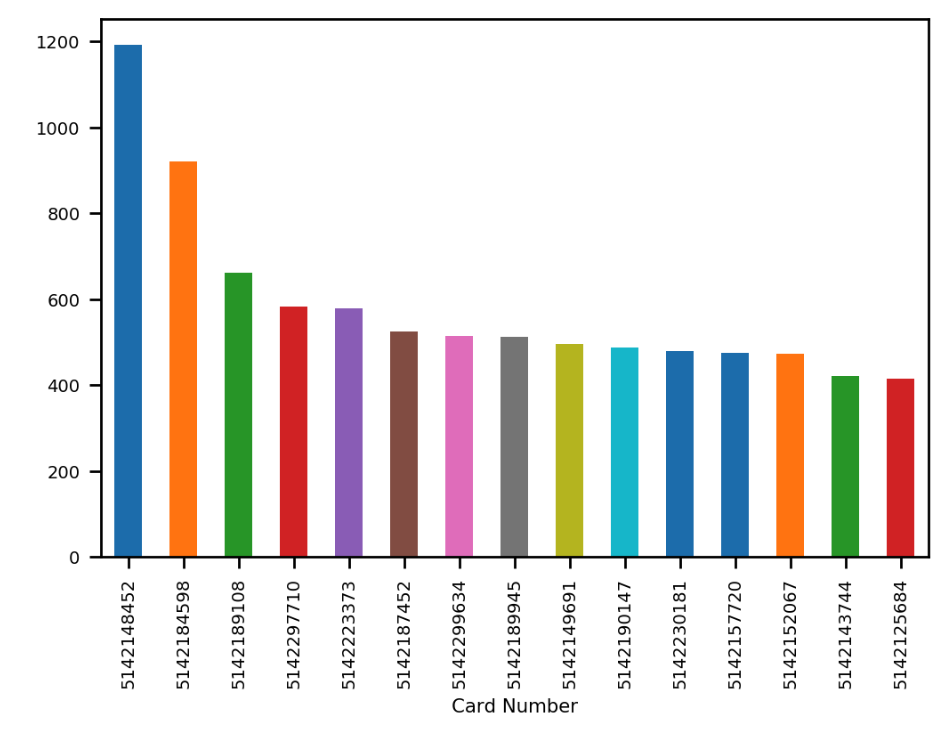
**III. Fields Description**

1. **Recnum**

**Recnum** represents the numbering for each record. There are 96753 unique values in the dataset which means each record has one corresponding **Recnum.**

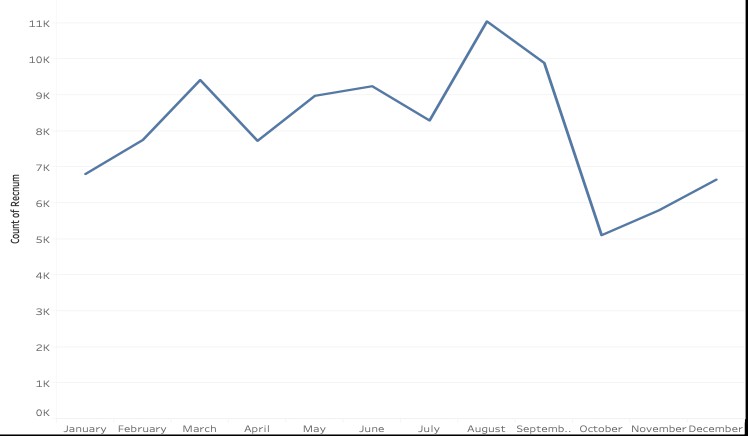
1. **Cardnum**

**Cardnum** stands for credit card number. The following graph shows the transaction count for the top 15 most common credit card numbers.

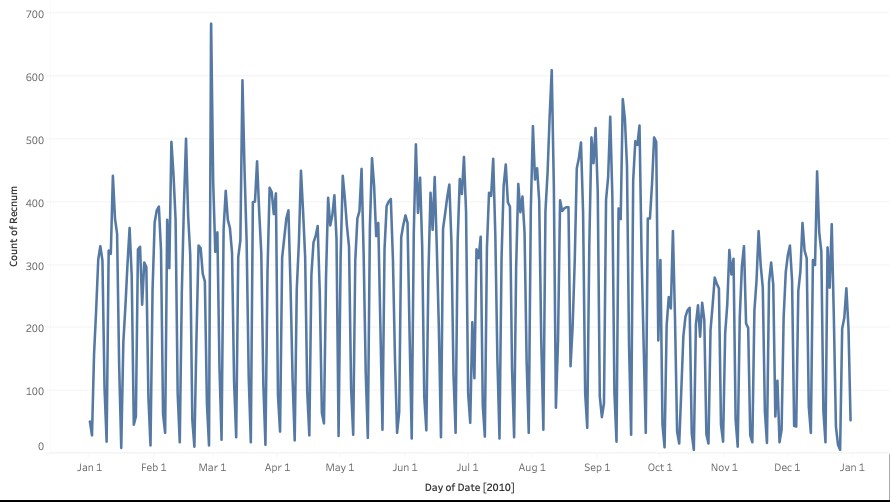


1. **Date**

The Date field contains 365 unique values, which are date from 1/1/2010 to 12/31/2010.



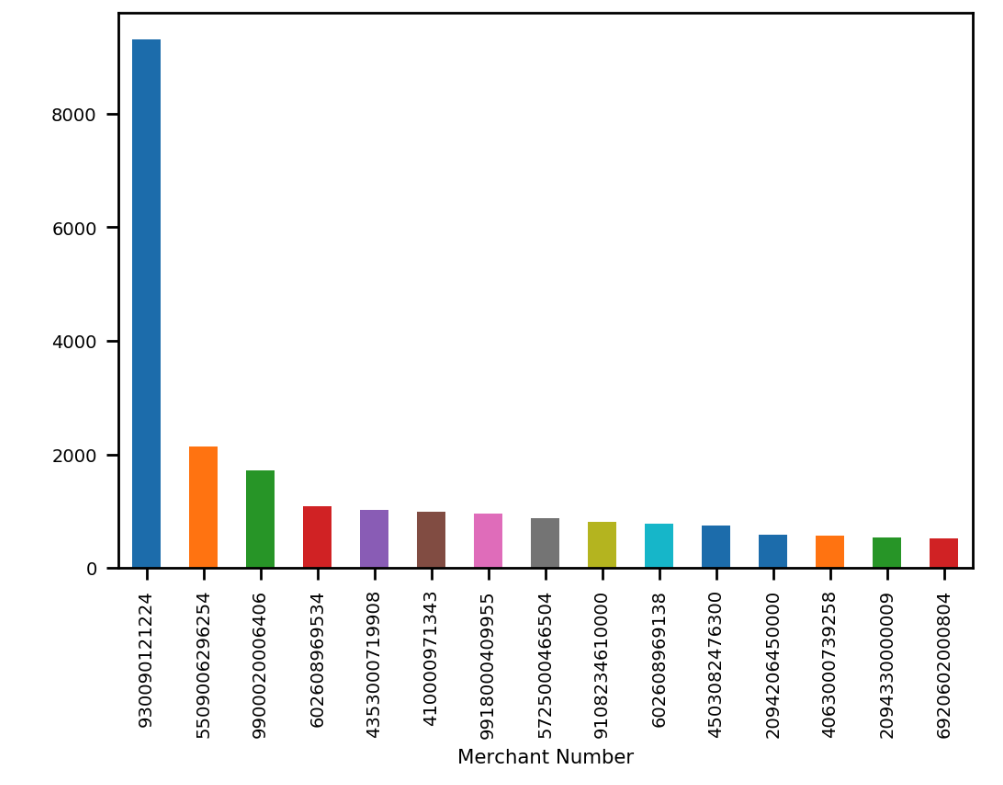
Count of Transactions by month



Count of Transactions by date

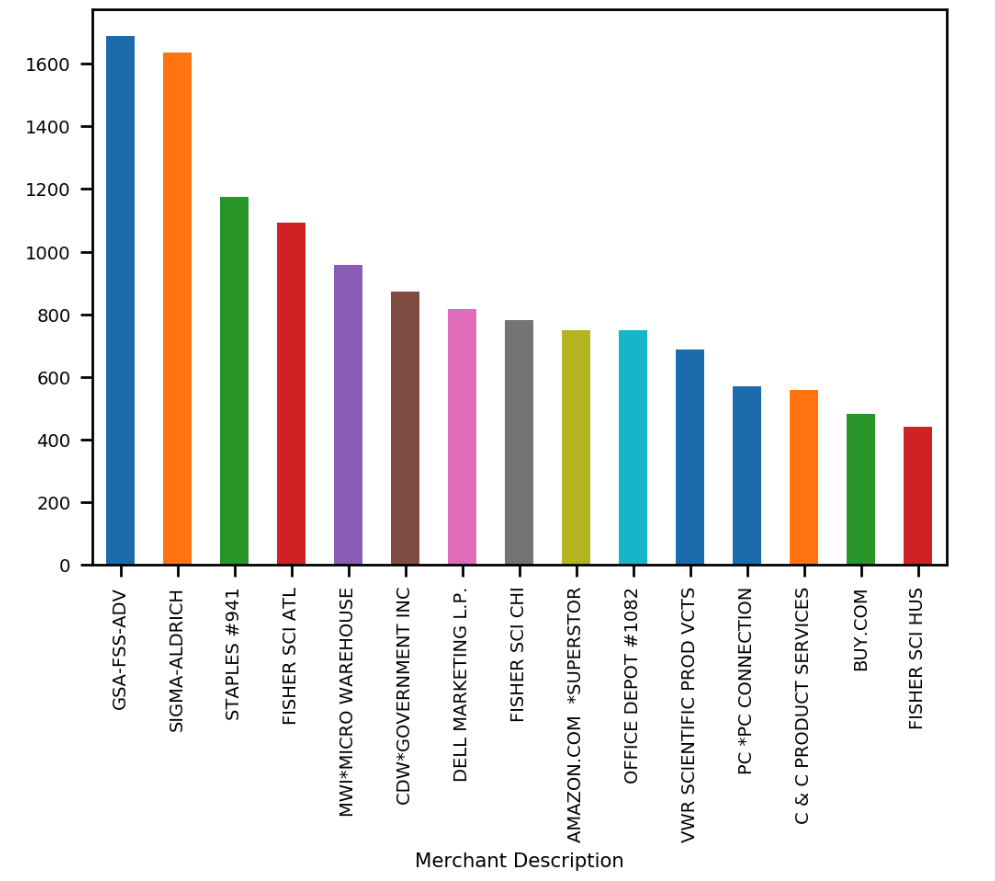
1. **Merchnum**

**Merchnum** represents merchant number. The graph below shows the count for the top 15 most common merchant numbers.



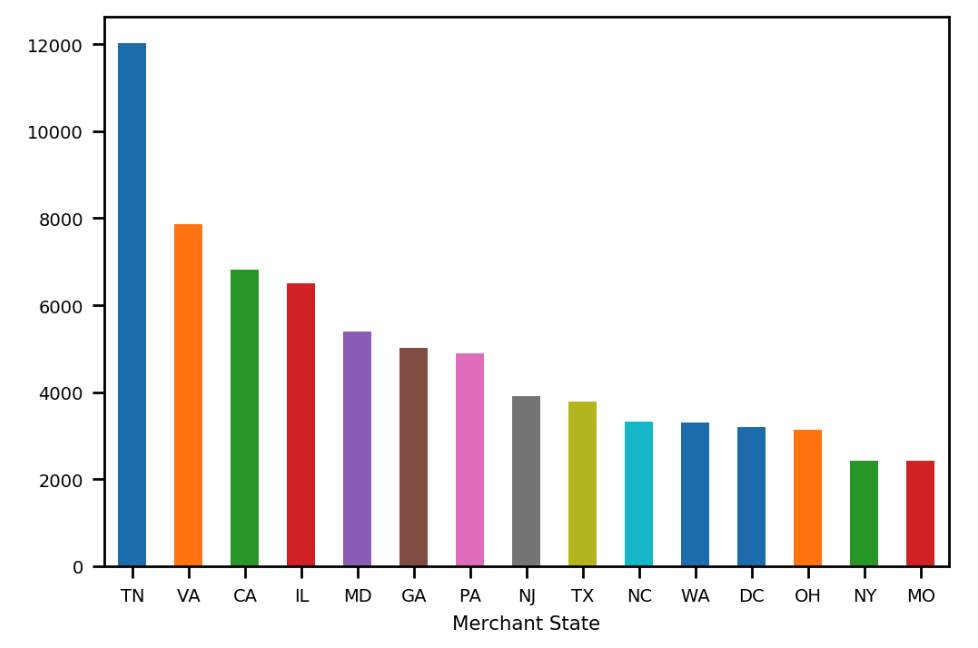
1. **Merch description**

**Merch description** is merchant description. Notice that there are 13,126 unique values for merchant description, which is more than that of merchant numbers. Below graph shows the count for the top 15 most common merchant description.



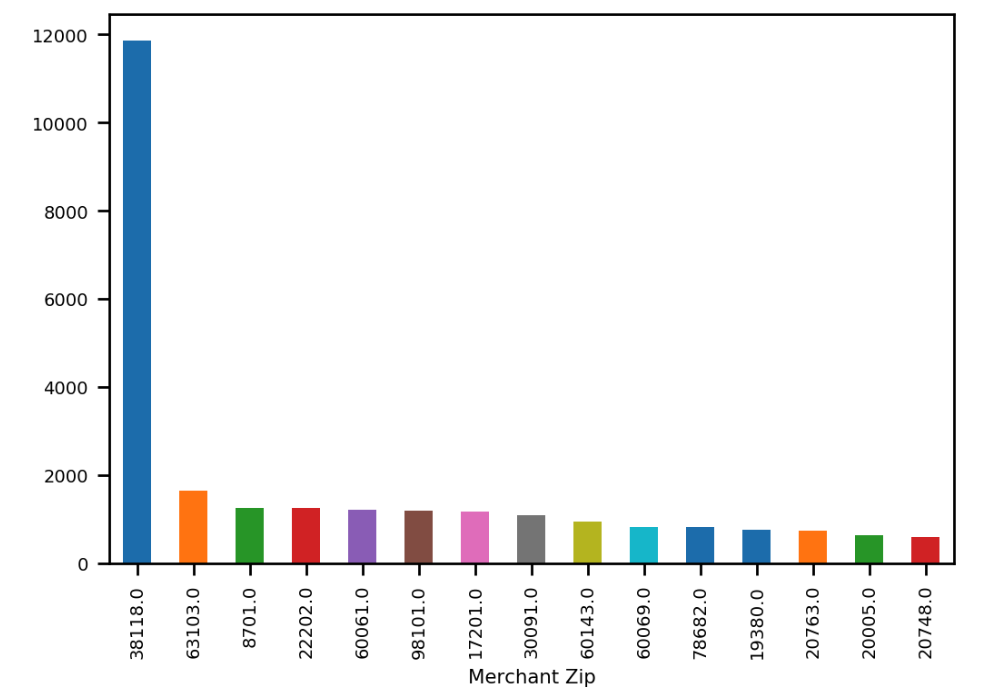
1. **Merch state**

**Merch state** represents the state where the merchant located. The following graph shows the count for the top 15 most common merchant states.



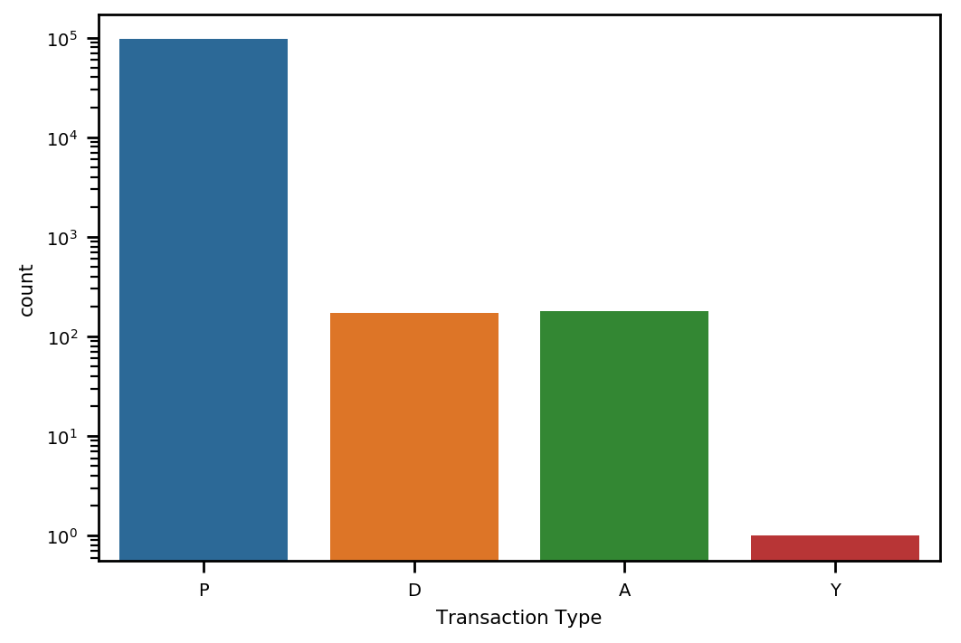
1. **Merch zip**

**Merch zip** represents the zip code where the merchant located. The following graph shows the count for the top 15 most common merchant zip code.



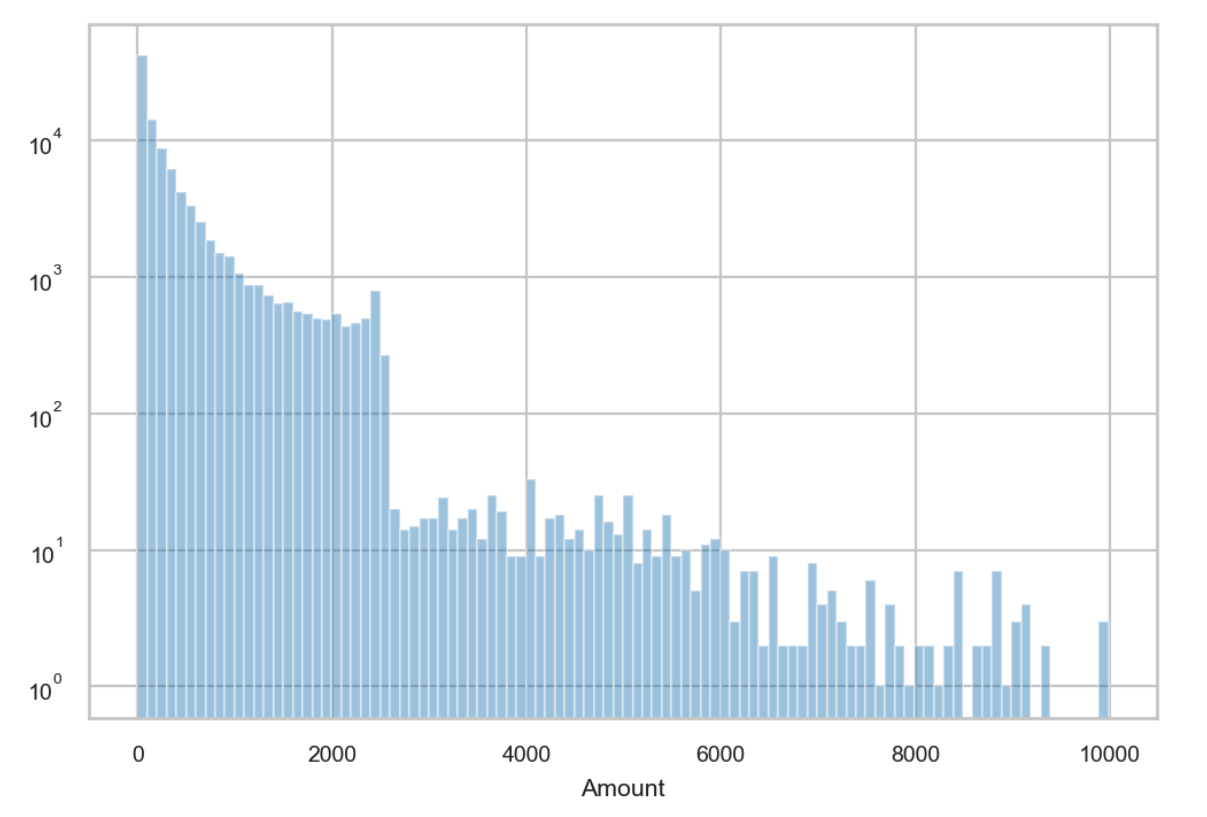
1. **Transtype**

**Transtype** stands for transaction types. There are 4 types of transaction in this dataset. “P”, which means “Purchase”, is the most common type in this filed. The following graph show the count for all types of transaction.



1. **Amount**

**Amount** is the transaction amount. Below graph shows the log scale distribution of transaction amount with amount limited to 10,000.



1. **Fraud**

**Fraud** is a response field in this dataset. There are two types, “0” and “1”. Type “0” categorized as non-fraud transaction and Type “1” represents fraud transaction. The fraud records account for about 1.1% of the whole dataset. The following is the log scale count of Fraud type.

