

**Data Quality Report**

**(DQR)**

# Summary

This dataset contains 96,708 records on credit card transactions from January 1st, 2010 to December 31st, 2010. There are total ten variables including card number, transaction date, merchant number, merchant description, location or state of the merchants, zip code of the merchants, transaction type, transaction amount, and whether a record has been labeled as fraud. Only amount is the only numeric variable, and the rest of the variables are categorical or date.

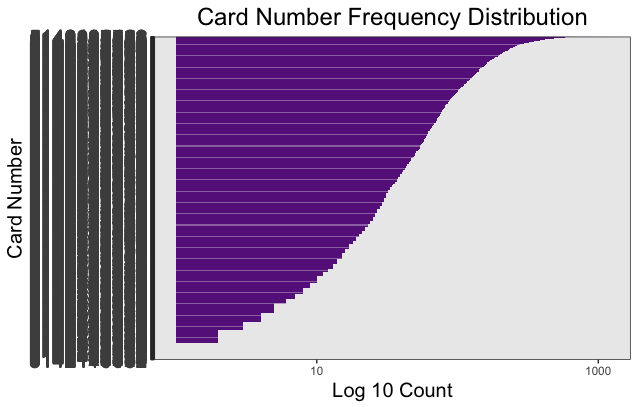
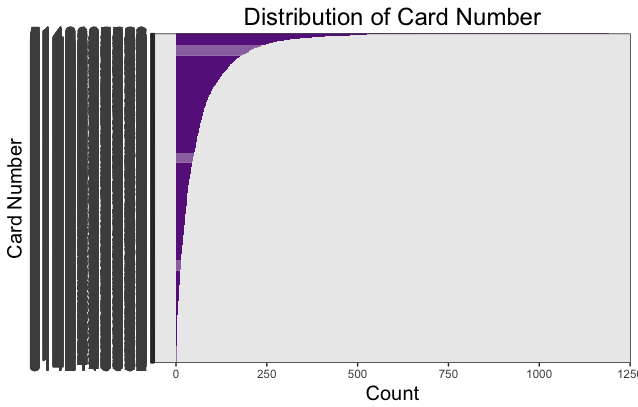
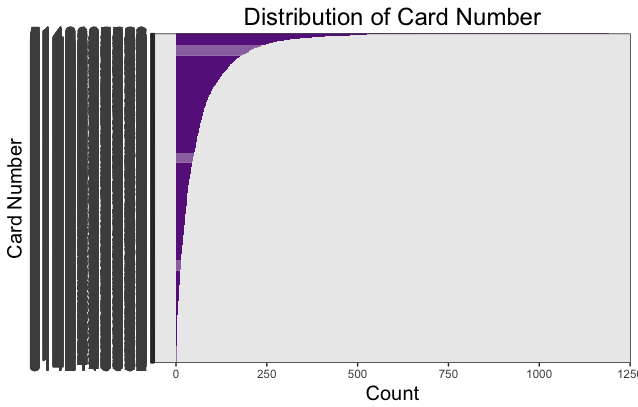
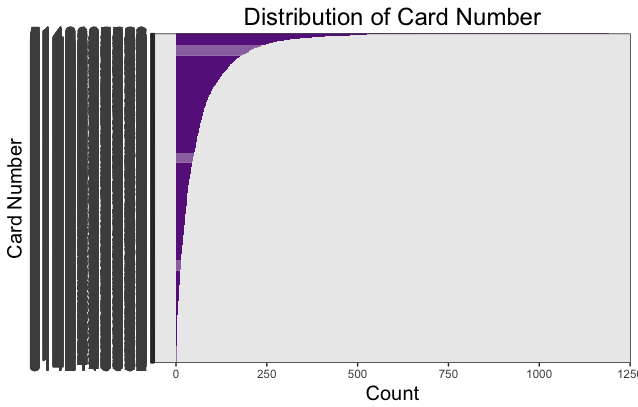
The table below shows summary statistics of each variable, including the percentage of record populated and the number of unique values.

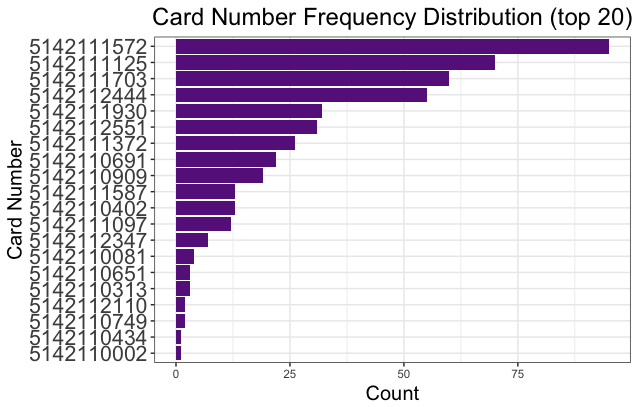
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | Type | Percentage Populated | Unique Count | Min | Max | Mean | SD |
| Recordnum | Categorical | 100% | 96708 | - | - | - | - |
| Cardnum | Categorical | 100% | 1644 | - | - | - | - |
| Date | Date | 100% | 365 | - | - | - | - |
| Merchantnum | Categorical | 96.51% | 13091 | - | - | - | - |
| Merch.Description | Categorical | 100% | 13125 | - | - | - | - |
| Merchant.State | Categorical | 98.76% | 228 | - | - | - | - |
| Merchant.Zip | Categorical | 95.19% | 4568 | - | - | - | - |
| Transtype | Categorical | 100% | 4 | - | - | - | - |
| Amount ($) | Numerical | 100% | 34876 | 0.01 | 3102045.53 | 199.36 | 230.74 |
| Fraud (Y/N) | Boolean | 100% | 2 | - | - | - | - |

The section starting next page shows descriptions and visualizations of each variable.

# Descriptions of Each Variable

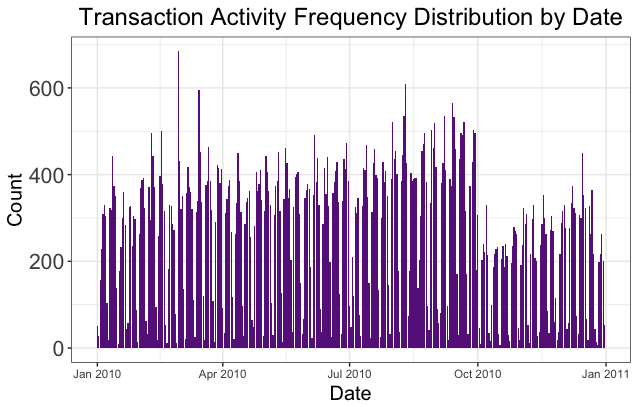
|  |  |
| --- | --- |
| Cardnum | |
| Description: | The credit card number of the card transaction |
| Class: | Categorical |
| Unique Values: | 1644 |
| Populated %: | 100% |

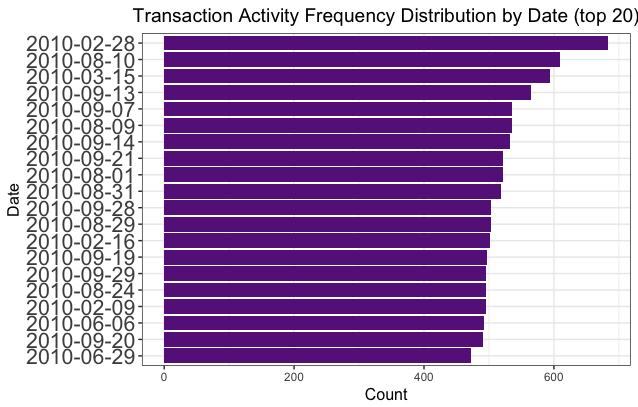


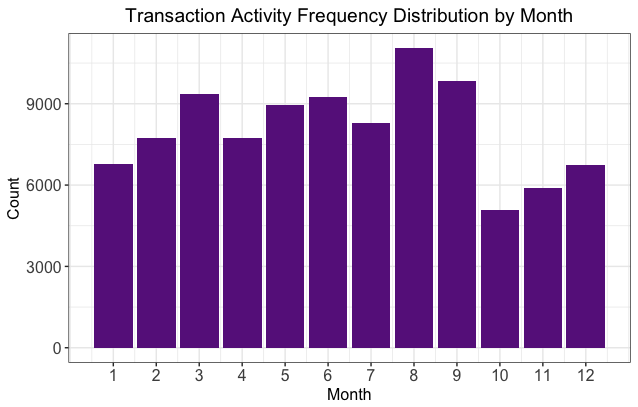


A few cards have extremely large number of transactions during the year. Majority cards have transactions below 200 times in the year. The most frequently used card is 5142111572.

|  |  |
| --- | --- |
| Date | |
| Description: | The date of the card transaction. |
| Class: | Date |
| Unique Values: | 365 |
| Populated %: | 100% |

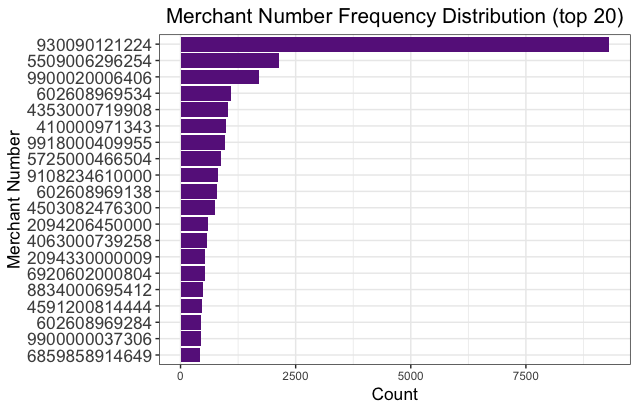


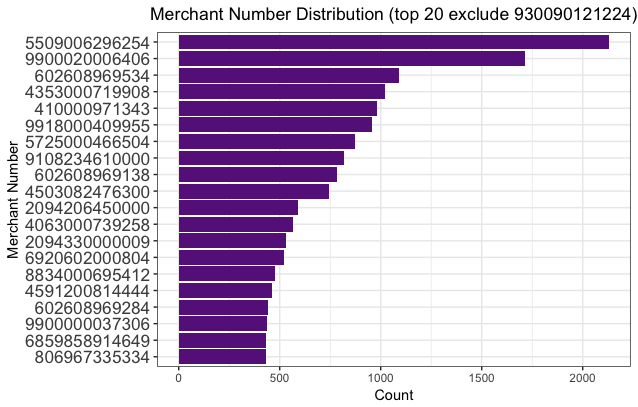




The single day with the most number of transactions is in March. When examining transactions by month, we found that August has the most number of transactions. October and November are the lowest activity months. This is a little strange because usually, people shop for the holiday season in these two months.

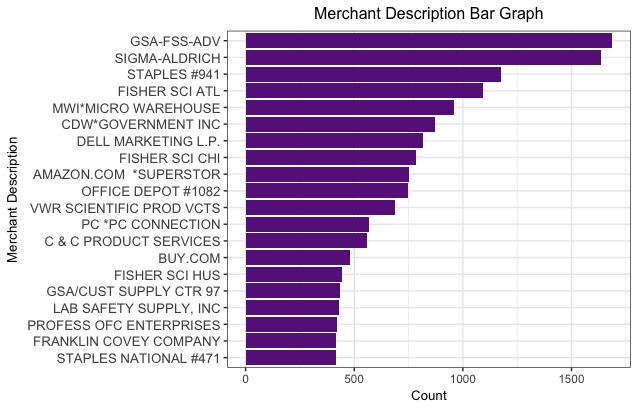
|  |  |
| --- | --- |
| Merchantnum | |
| Description: | The merchant account number for each bank account. |
| Class: | Categorical |
| Unique Values: | 365 |
| Populated %: | 100% |





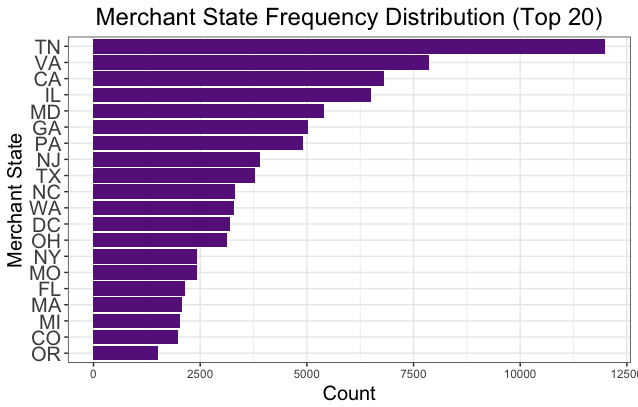
The first graph shows top 20 most frequently appeared merchant number, and 930090121224 has an extremely high number of transactions compared to the rest of merchant. After removing this merchant number, we see that only a few merchants have the number of transactions greater than 1,000. Since there is a total of 365 unique merchant numbers, the majority of the merchant numbers appear in less than 500 records.

|  |  |
| --- | --- |
| Merch.Description | |
| Description: | The description of each credit card transaction |
| Class: | Categorical |
| Unique Values: | 13125 |
| Populated %: | 100% |



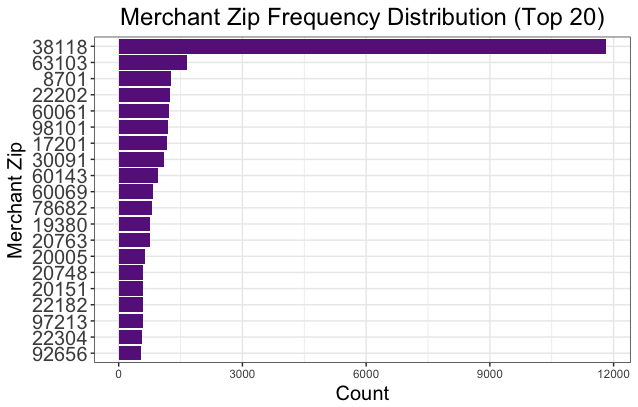
GSA-FSS-ADV and SIGMA-ALDRICH are the two most frequently appeared merchant description. GSA could be General Service Administration, and interestingly SIGMA-ALDRICH is an American chemical, life science, and biotechnology company. It could be because many of the transactions are purchases for the products or services of the company.

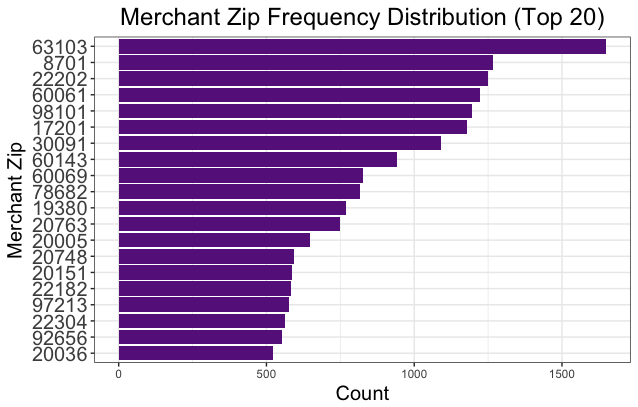
|  |  |
| --- | --- |
| Merch.State | |
| Description: | The state location for each credit card transaction |
| Class: | Categorical |
| Unique Values: | 228 |
| Populated %: | 98.76% |



There are 228 unique values for Merchant state. In addition to the US states abbreviations, the majority of them are three-digit integers, which could indicate that part of the transaction merchant in the records is overseas. The top 20 most frequently occurred merchant states are still US states. TN (Tennessee) ranks the top.

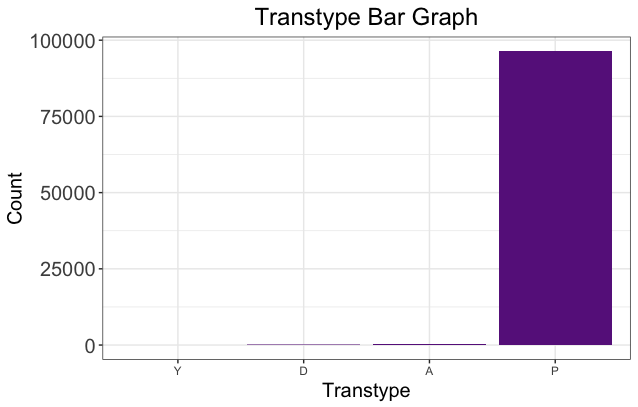
|  |  |
| --- | --- |
| Merch.Zip | |
| Description: | The zip code location for each credit card transaction |
| Class: | Categorical |
| Unique Values: | 4568 |
| Populated %: | 95.19% |

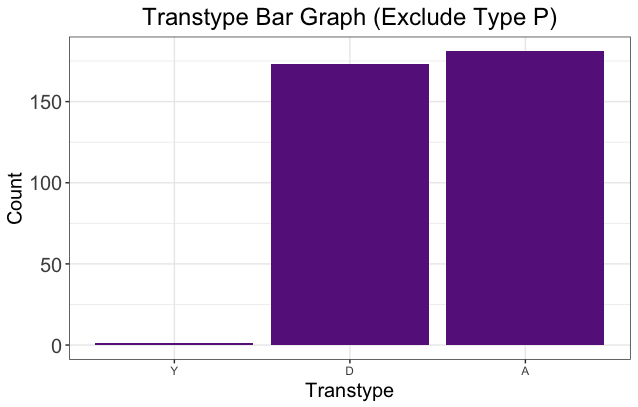




The merchant with zip code 38118 has the most number of transactions in the year. It is more than four times as any of other zip code, which is quite abnormal. After removing this zip code, we have a much better distribution of the merchant zip frequency, with the most frequently appeared zip code 63103.

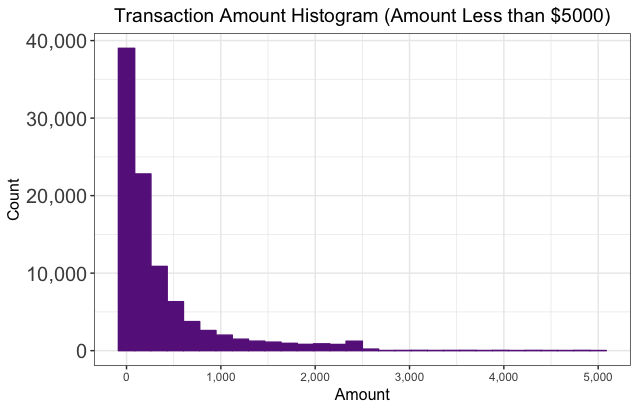
|  |  |
| --- | --- |
| Transtype | |
| Description: | The transaction type of each credit card transaction |
| Class: | Categorical |
| Unique Values: | 4 |
| Populated %: | 100% |

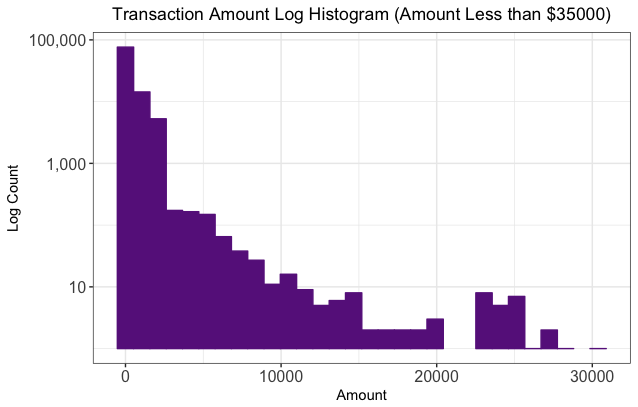




Almost 99% of the transactions in this dataset are type p. Type A and type D has the close number of occurrences. Type Y transactions are minimal.

|  |  |
| --- | --- |
| Amount | |
| Description: | The transaction amount of each credit card transaction |
| Class: | Numerical |
| Unique Values: | 34876 |
| Populated %: | 100% |





Majority of the transactions amount in this dataset is less than $500. Many of them are less than $200. While the largest transaction amount is up to $3,102,045.53, the second largest transaction amount drops to $47,900. The merchant description of the highest record is "INTERMEXICO". It is possible that the amount is recorded in Peso, instead of USD. When we conduct further analysis, we will either exclude this record or convert it to USD.

|  |  |
| --- | --- |
| Fraud | |
| Description: | Whether the credit card translation is fraudulent (“1” for Yes, “0” for No) |
| Class: | Boolean |
| Unique Values: | 2 |
| Populated %: | 100% |



There is only about 1% fraud record in this dataset.