Data Quality Report

Sheetal Srivastava

---

1. Data Description

The dataset is **Card Transactions Data**, which contains **business transaction** information of credit cards for an organization. The data consists of real card transactions coming from a U.S. government organization. The data spans over the time period of a year, from **1st January, 2010** to **31st December, 2010**. There are **10 fields** and **96753 records**.

2. Summary Tables

Numeric Fields Table

Field Name	# Records With Values	% Populated	# Zeros	Min	Max	Mean	Most Common	Stdev
Date	96753	100.00%	0	2010-01-01	2010-12-31	2010-06-25	2010-02-28	98 days
Amount	96753	100.00%	0	0.01	3102045.53	427.89	3.62	10006.14

Categorical Fields Tables

Field Name	# Records With Values	% Populated	# Zeros	# Unique Values	Most Common
Recnum	96753	100.00%	0	96753	1
Cardnum	96753	100.00%	0	1645	5142148452
Merchnum	93378	96.50%	0	13091	930090121224
Merch description	96753	100.00%	0	13126	GSA-FSS-ADV
Merch state	95558	98.80%	0	227	TN
Merch zip	92097	95.20%	0	4567	38118
Transtype	96753	100.00%	0	4	P
Fraud	96753	100.00%	95694	2	0

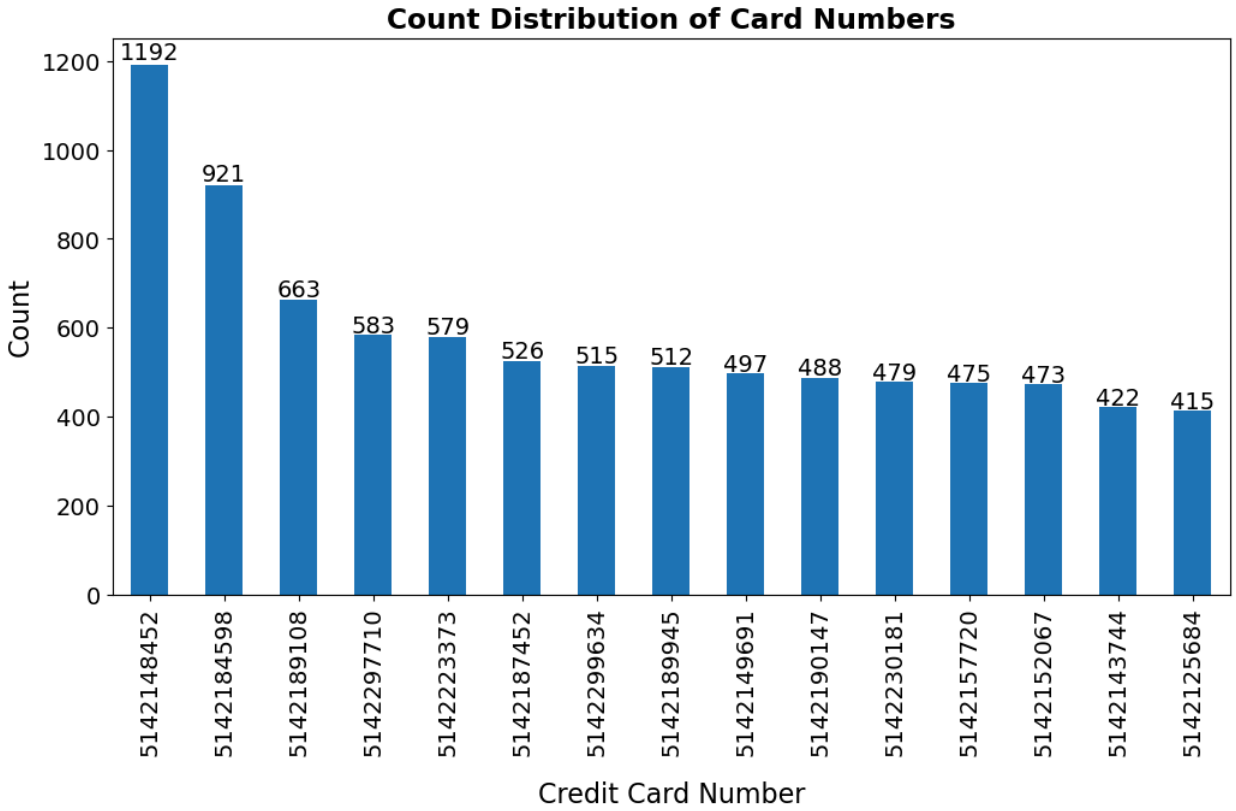
3. Visualization of Each Field

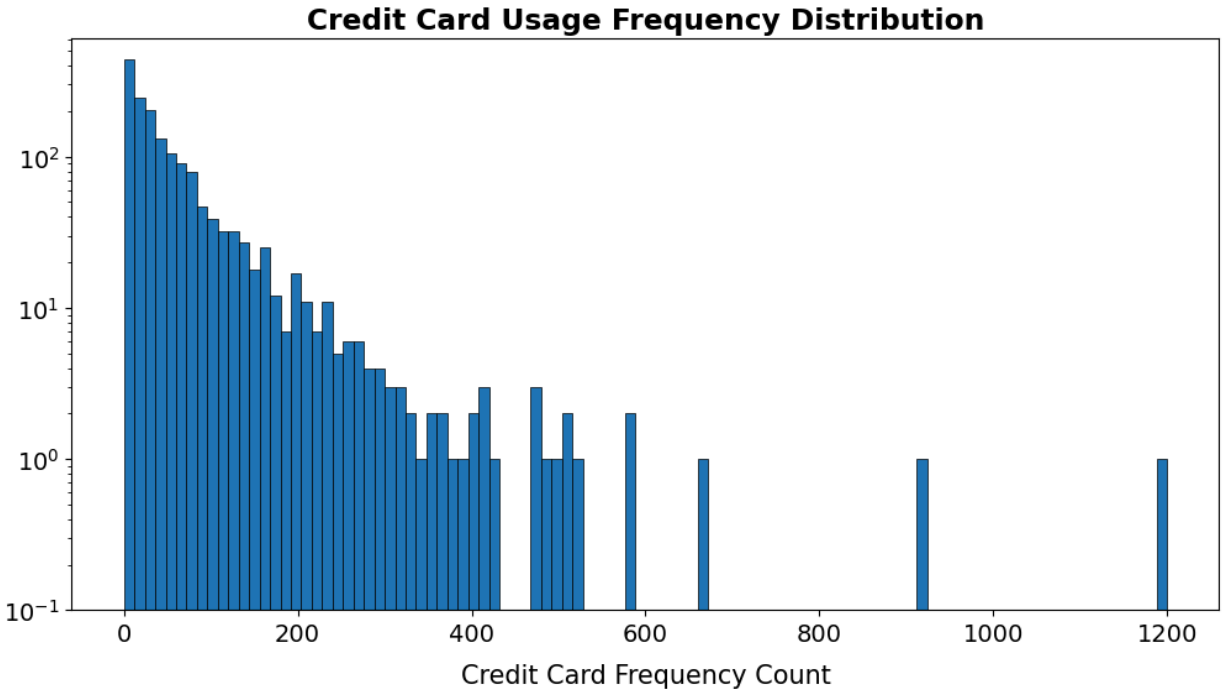
a. Field Name : **Recnum**

Description : Ordinal unique positive integer for each credit card transaction record, from 1 to 96753.

b. Field Name : **Cardnum**

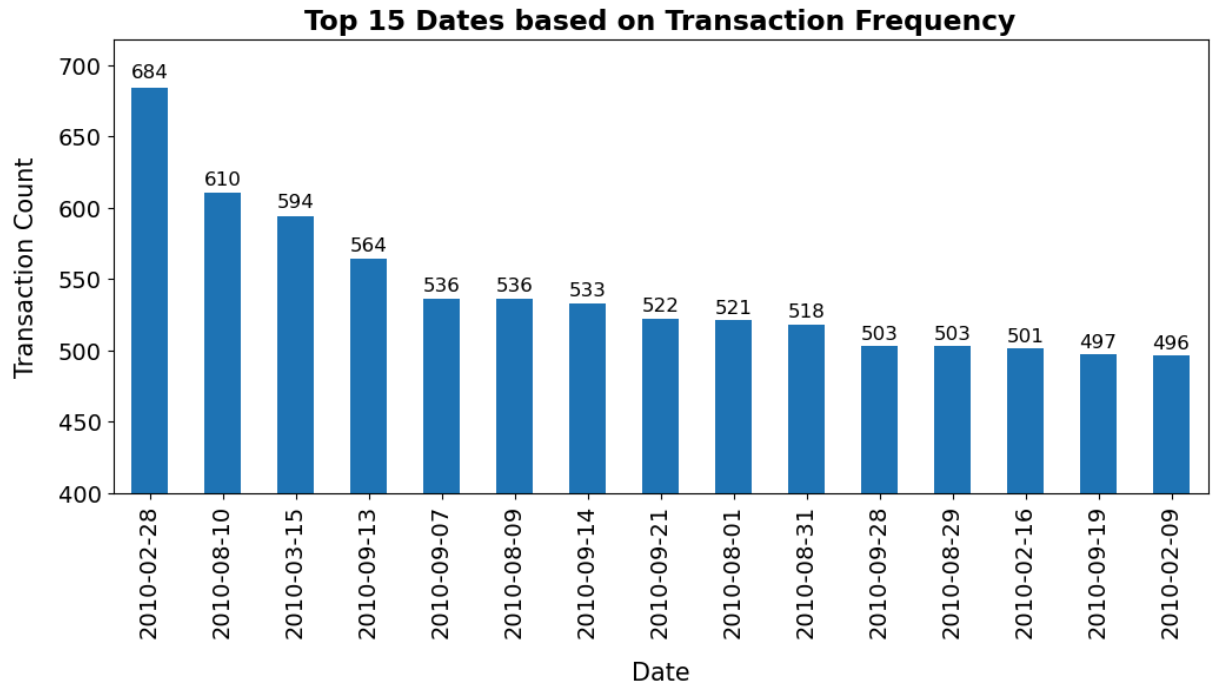
Description : Credit card number. This field has 1645 unique values.
The first graph shows the top 15 credit cards with maximum occurrences in the dataset. The credit card with the card number “5142148452” is the most common card used, with 1192 total transactions.
Observing the frequency distribution graph (second figure) we can see that only 1 card has been used more than 1000 times (1192 times, to be precise). Majority of the credit cards have been used 1 to 200 times only.

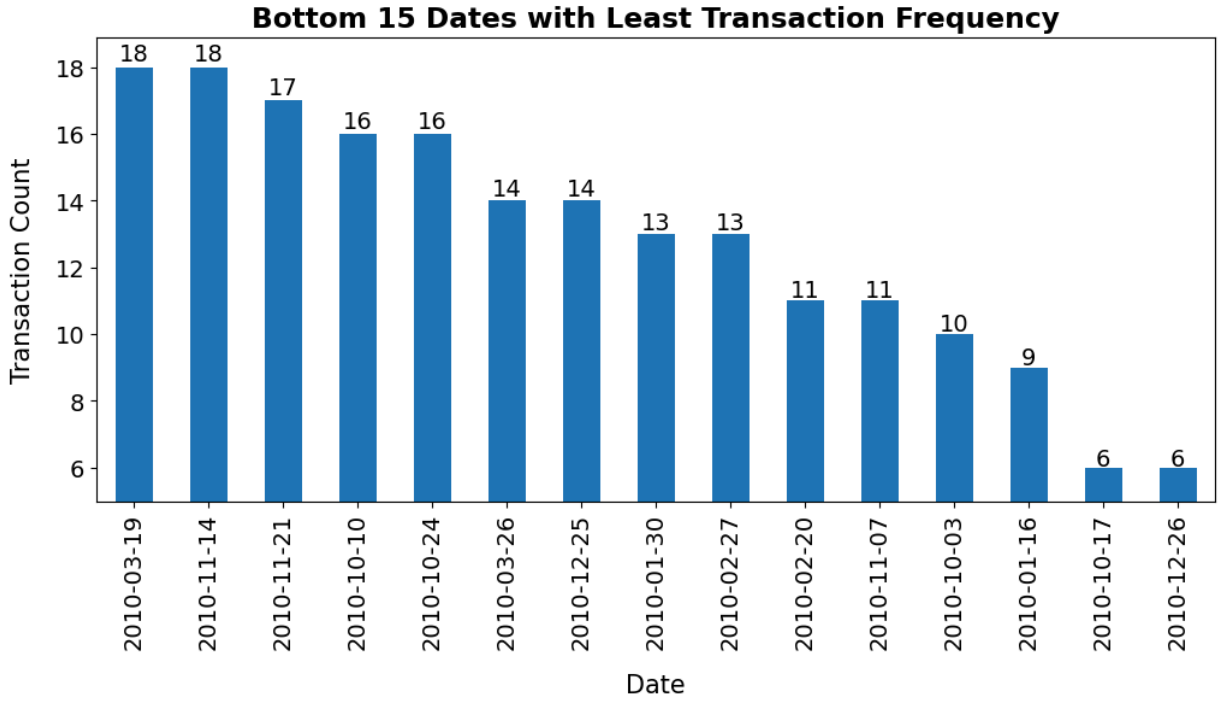


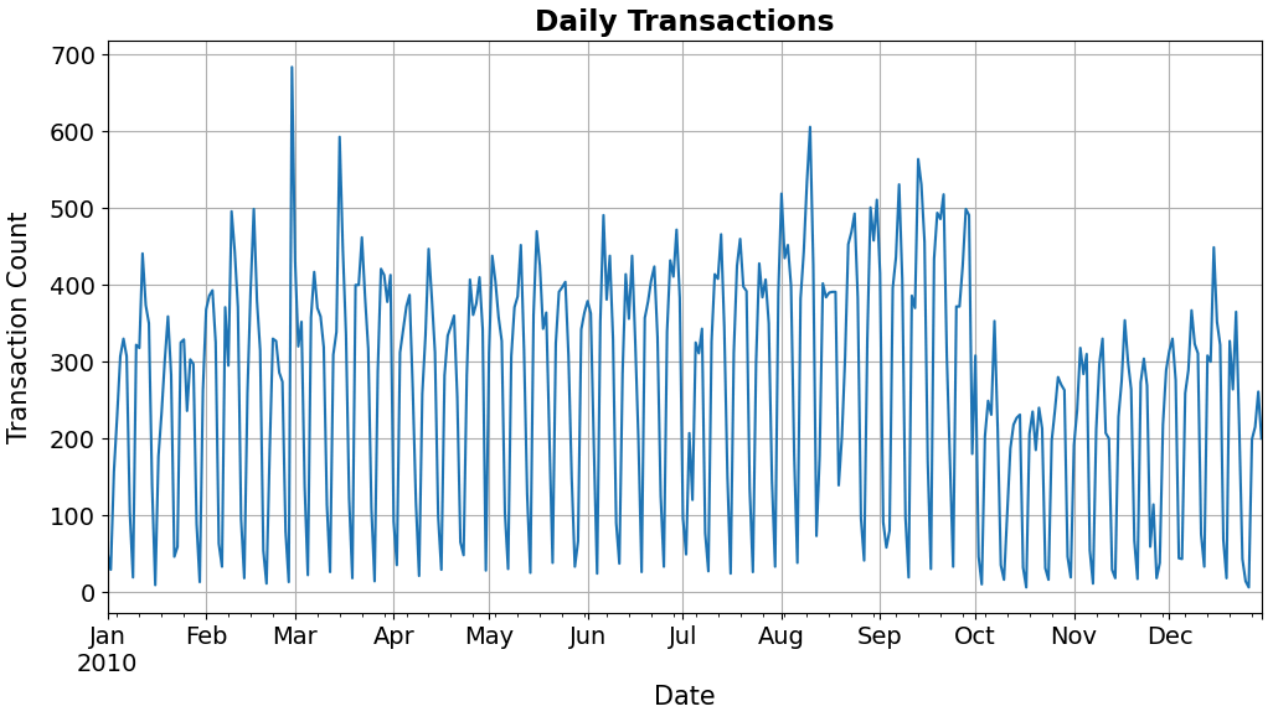


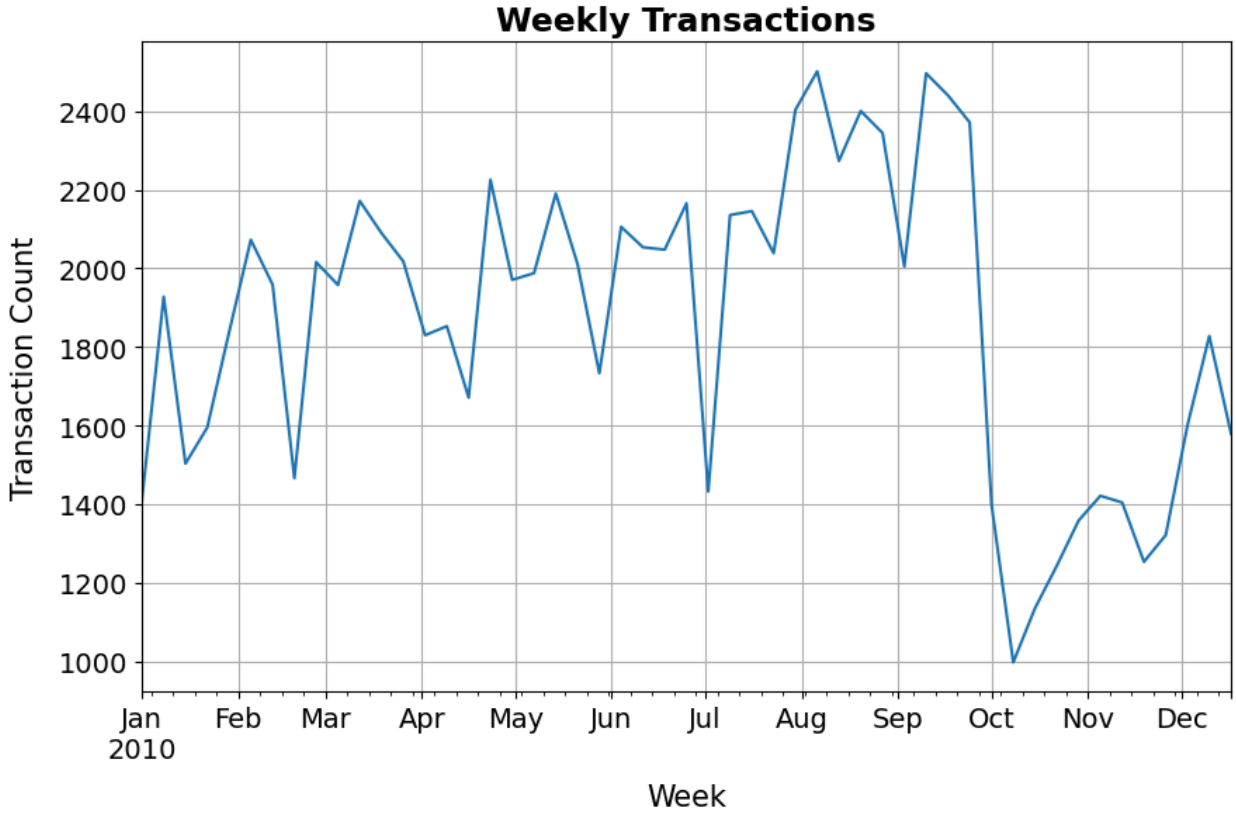
c. Field Name : **Date**

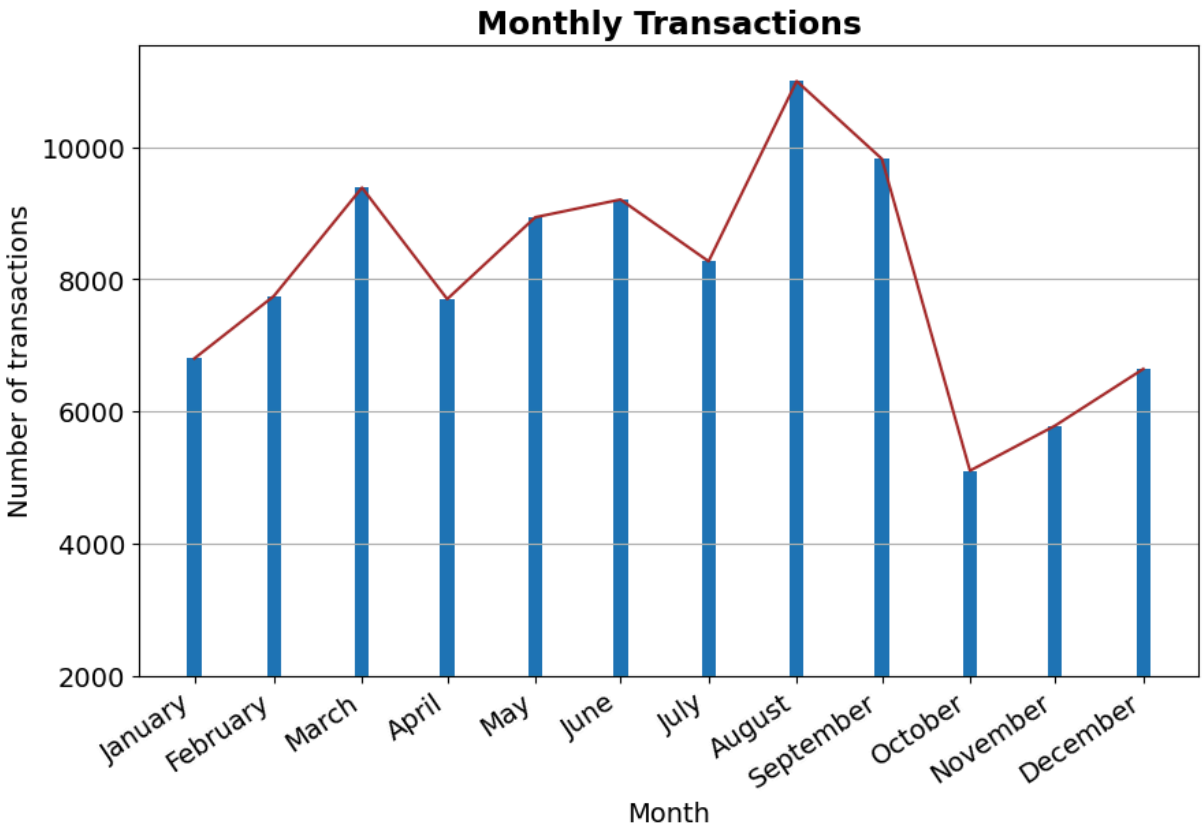
Description : Transaction date. This field has 365 unique values. We have exactly a year's worth of data, for the year 2010 (Jan 1st to Dec 31st). Based on the two Transaction Frequency graphs, we can see that the minimum number of transactions on any given date were 6, and the maximum number of transactions made on a single day were 684 (on 28th February, 2010).
The remaining three plots show the daily, weekly and monthly transaction frequency distributions over time. We can see that the month of August has recorded the highest number of transactions, in 2010.


Top 15 Dates based on Transaction Frequency
Transaction Count
Date
700
650
600
550
500
450
400
684
610
594
564
536
536
533
522
521
518
503
503
501
497
496
2010-02-28
2010-08-10
2010-03-15
2010-09-13
2010-09-07
2010-08-09
2010-09-14
2010-09-21
2010-08-01
2010-08-31
2010-09-28
2010-08-29
2010-02-16
2010-09-19
2010-02-09


Bottom 15 Dates with Least Transaction Frequency
Transaction Count
Date
18
16
14
12
10
8
6
18
18
17
16
16
14
14
13
13
11
11
10
9
6
6
2010-03-19
2010-11-14
2010-11-21
2010-10-10
2010-10-24
2010-03-26
2010-12-25
2010-01-30
2010-02-27
2010-02-20
2010-11-07
2010-10-03
2010-01-16
2010-10-17
2010-12-26

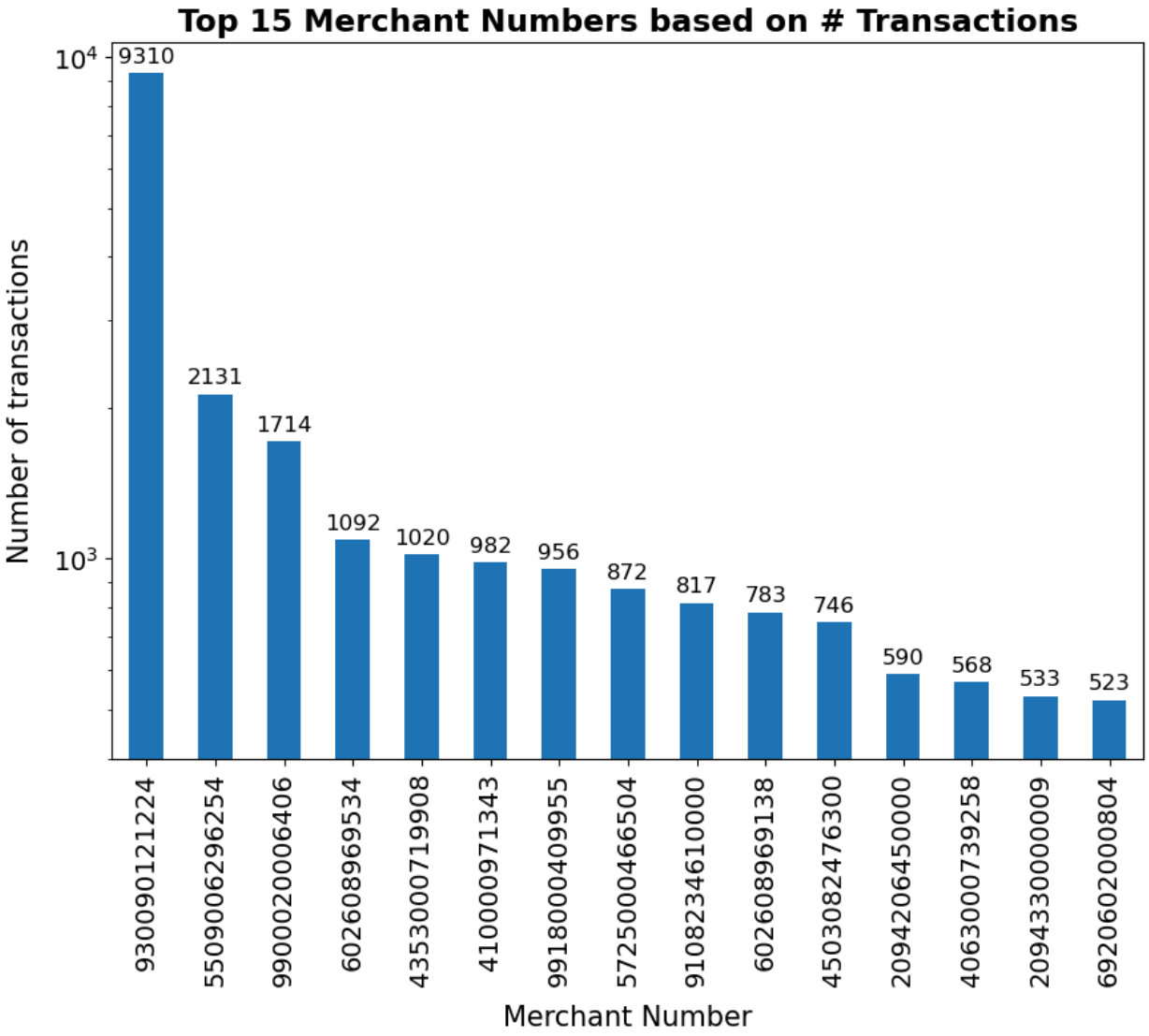

Daily Transactions
Transaction Count
700
600
500
400
300
200
100
0
Jan
2010
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Date


Weekly Transactions
Transaction Count
2400
2200
2000
1800
1600
1400
1200
1000
Jan
2010
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Week


Monthly Transactions
Number of transactions
10000
8000
6000
4000
2000
January
February
March
April
May
June
July
August
September
October
November
December
Month

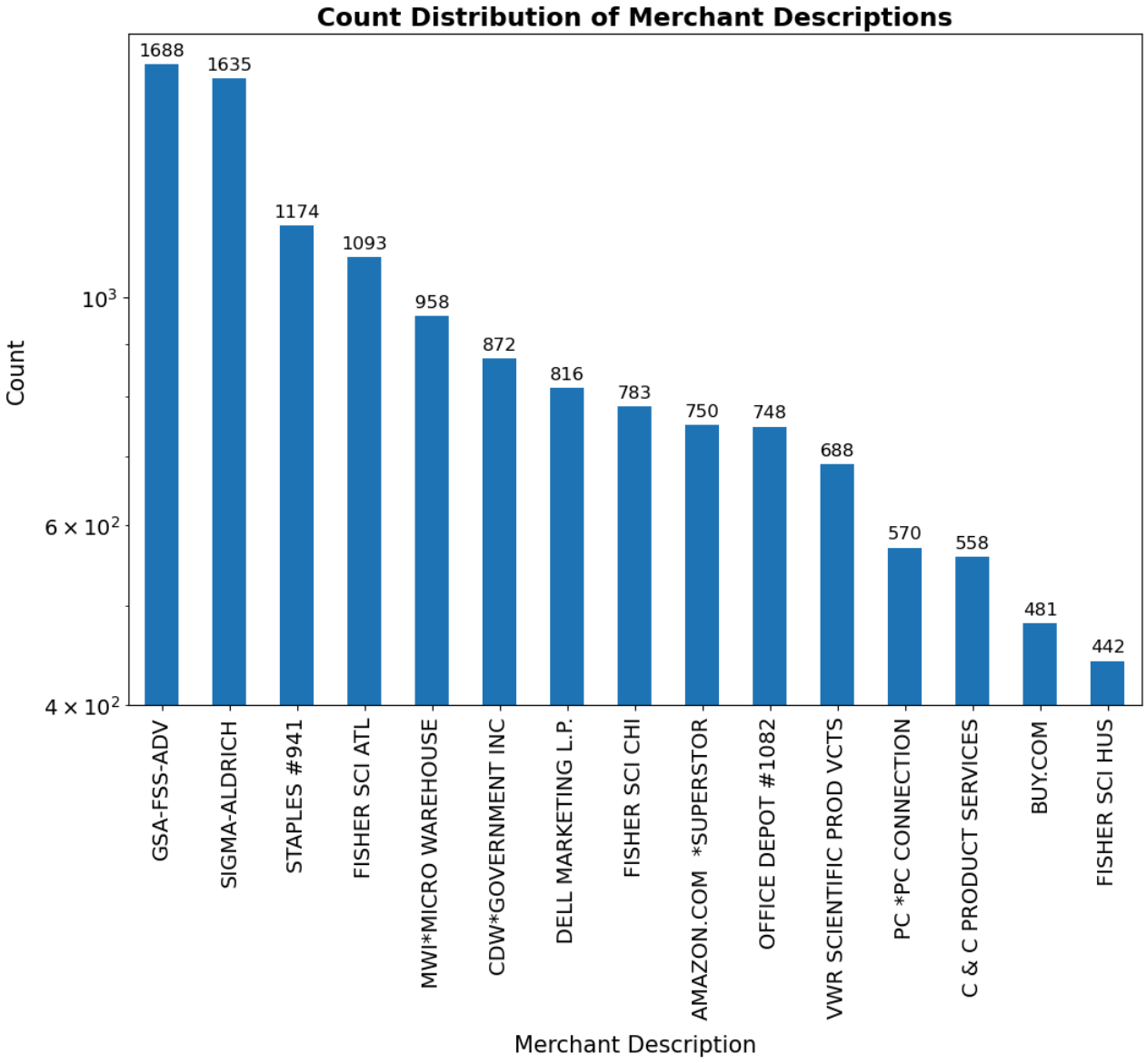
d. Field Name : **Merchnum**

Description : Merchant Number. This is a categorical field with 13092 distinct values. The most common merchant number is “930090121224”, with a total transaction count of 9310.



e. Field Name : **Merch description**

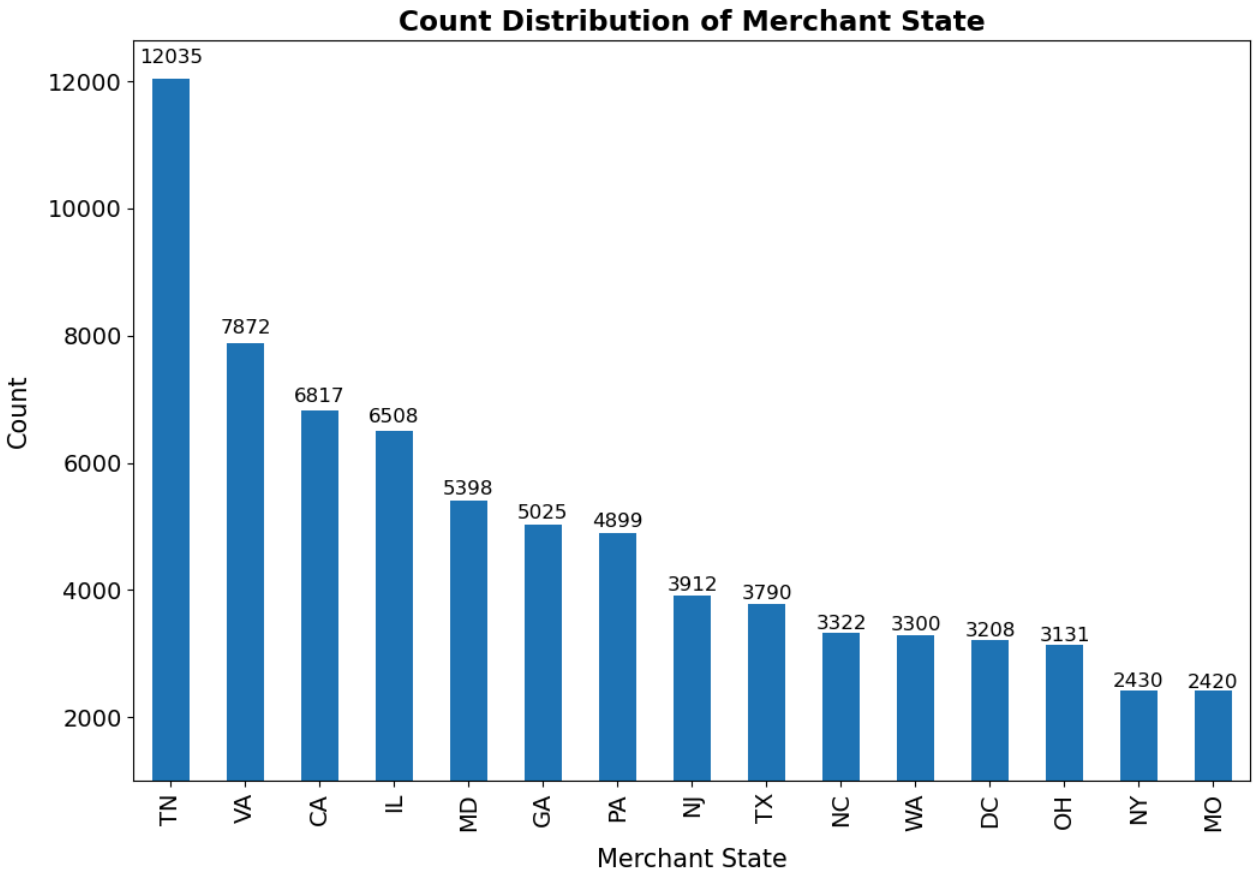
Description : Merchant Description. This a text field, giving a brief description of the merchant. This field has 13126 distinct values. The most common merchant description is “GSA-FSS-ADV” with a total count of 1688.



f. Field Name : **Merch state**

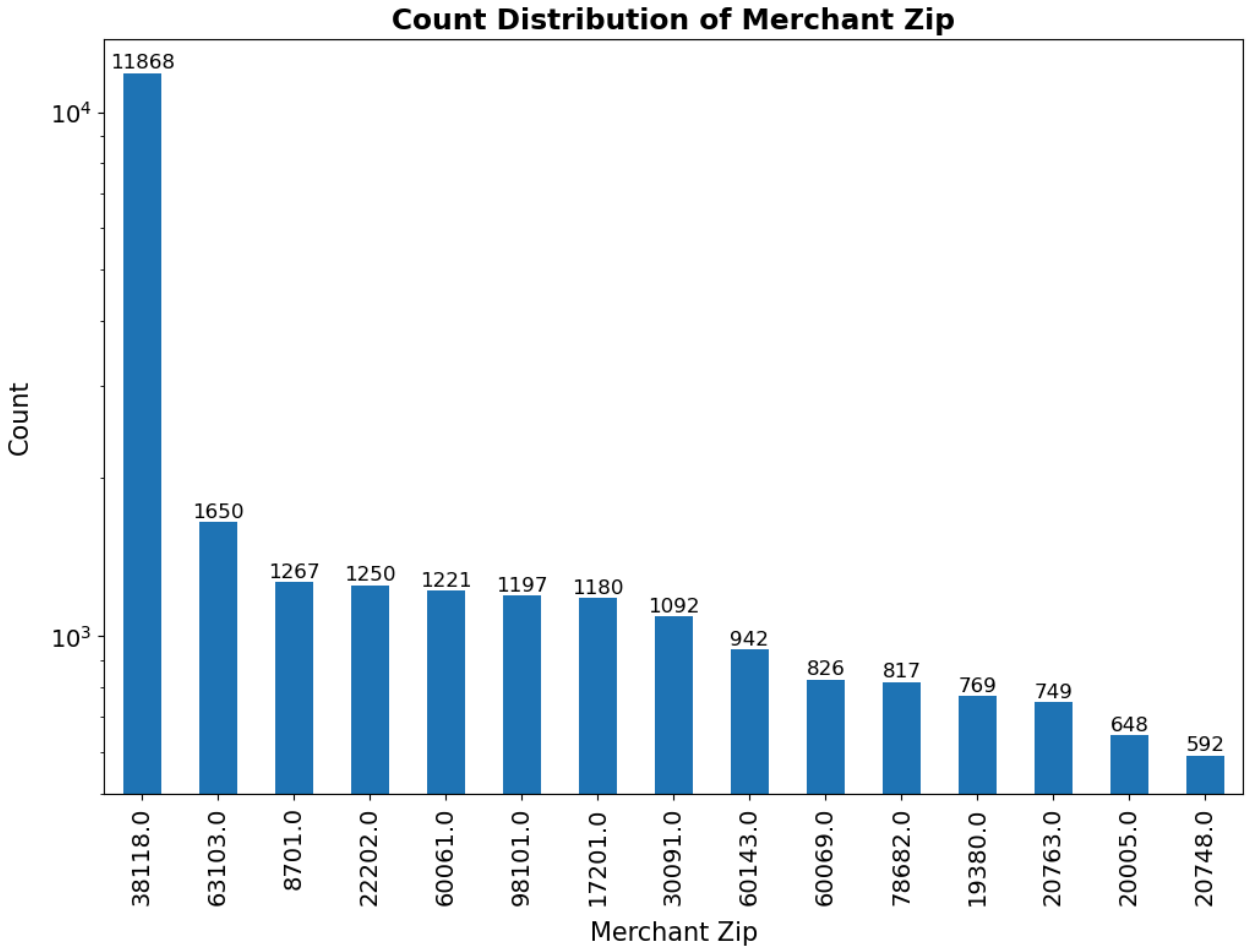
Description : Merchant State. This is a categorical field that contains the two letter state code of the state to which the merchant belongs. This field has 228 unique values.

The figure below shows the top 15 merchant states based on the number of occurrences. The most common merchant state is “TN” with a total count of 12035 transactions.



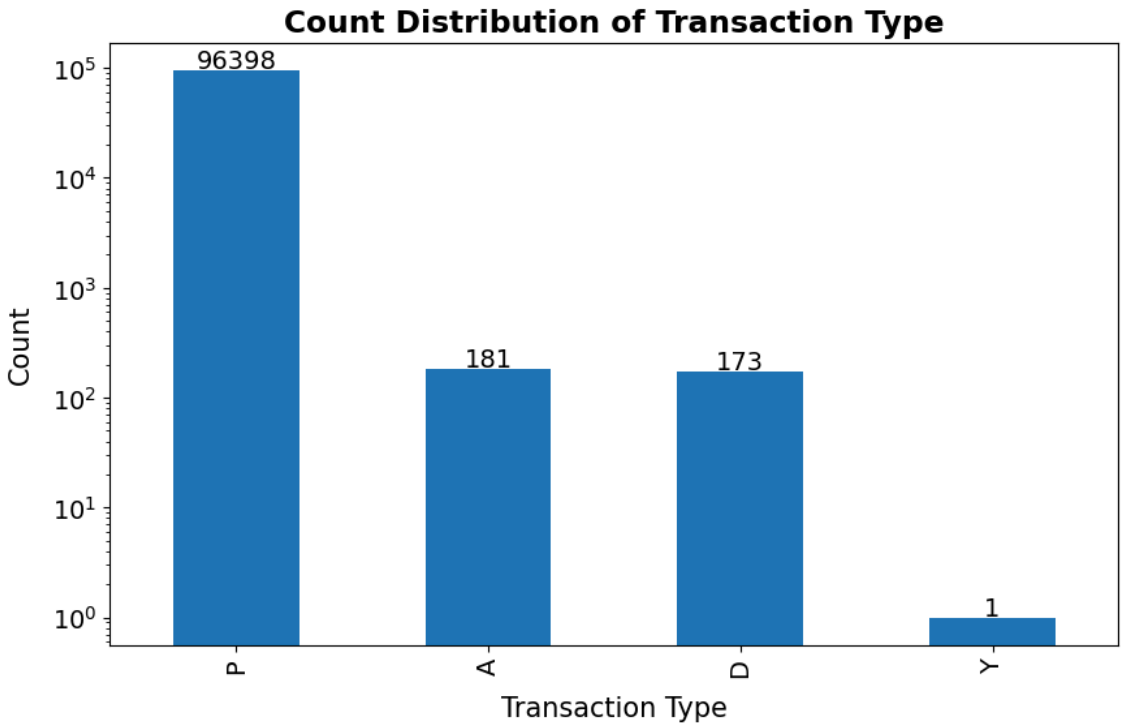
g. Field Name : **Merch zip**

Description : Merchant zip code. This is a categorical field consisting of the zip code values of the merchant. This field has 4568 distinct values.
The most common Merchant Zip is “38118” with 11868 total occurrences.



h. Field Name : **Transtype**

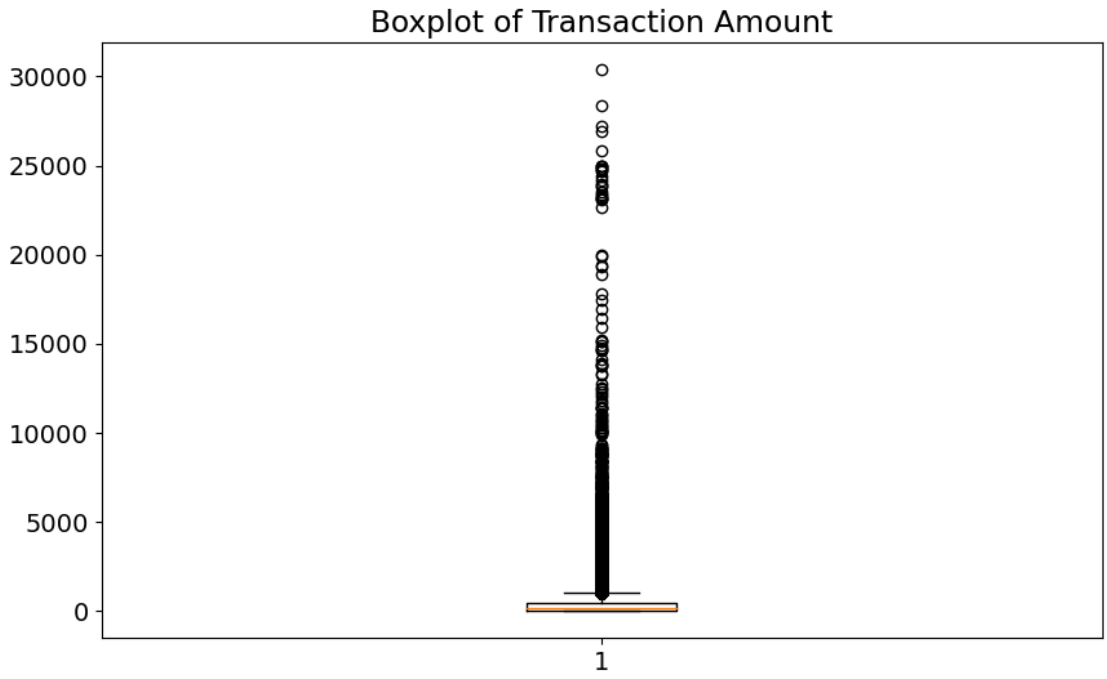
Description : Transaction type. This is a categorical field classifying the type of transaction. This field has 3 distinct values which are - P, A, D, and Y.
The most common transaction type is “P” (purchase) with a total of 96398 occurrences out of a total of 96753 records.

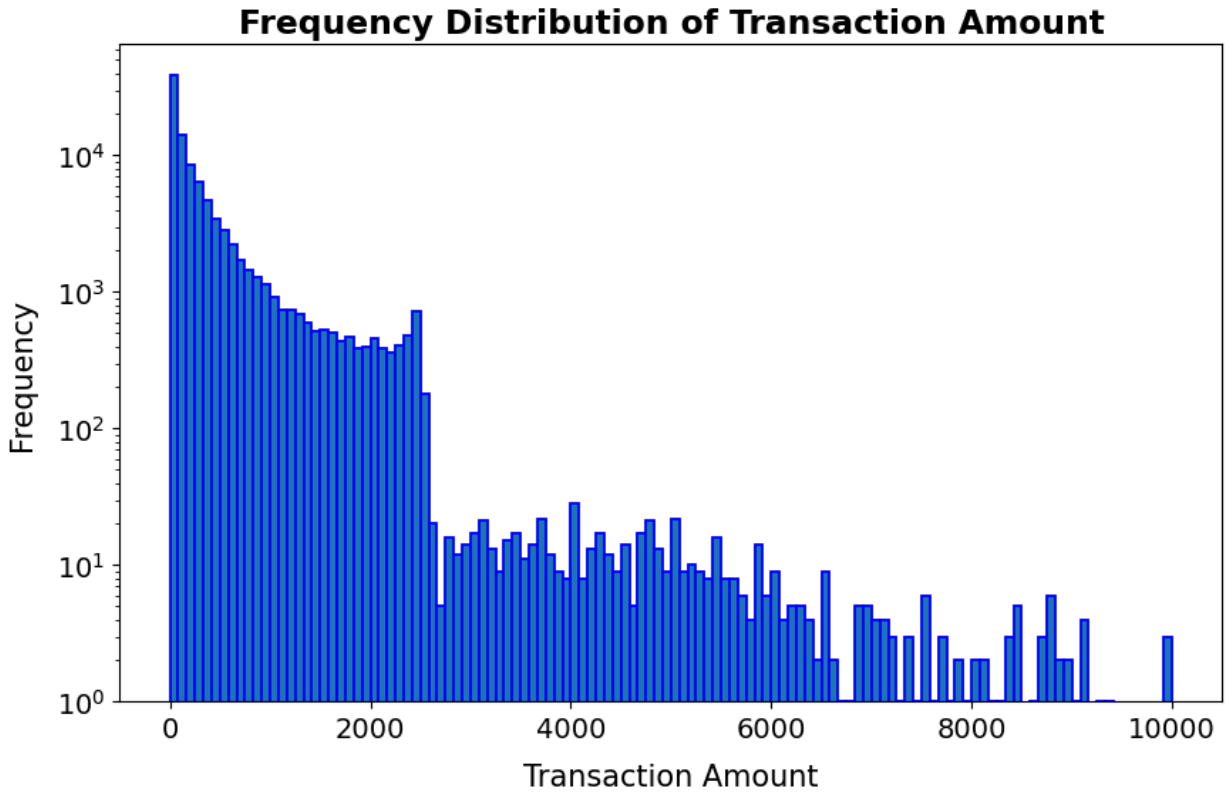


i. Field Name : **Amount**

Description : This numerical field contains the value of the transaction amount. The amount value ranges from 0.01 to 3102045.53 with a mean value of 427.89 and mode value of 3.62.
Examination of this data shows that the max value (3102045.53) is an outlier. Removing this outlier value, we get the first figure, a boxplot, showing that the majority of the transactions fall within the range of 0.01 to 10,000.

Therefore, we plot the second graph, which gives us the distribution of the transaction amount, keeping the amount limit as 10,000. This plot shows us that the bulk of the transactions are of an amount ranging from 0.01 to 2700.


Boxplot of Transaction Amount
30000
25000
20000
15000
10000
5000
0
1


Frequency Distribution of Transaction Amount
Frequency
10^4
10^3
10^2
10^1
10^0
0
2000
4000
6000
8000
10000
Transaction Amount

j. Field Name : **Fraud**

Description : This is a binary categorical field that specifies the label of the records, classifying it as fraud if the value = 1, and not a fraud transaction if value = 0. Since it's a binary field, it only takes 2 values : 0 and 1.
The total count of fraudulent transactions = 1,059.
The total count non-fraudulent transactions = 95,694.