***DQR Report***

1. **High-level description:**

The name of the data set is ‘card transactions’, it is an actual credit card purchases record from a US government organization. The data set describes people’s purchase recording including where, when, and types. The professor made up the fraud labels based on his experience with card transaction fraud. There are 10 fields and 96,753 records in the data set.

1. **Two field summary tables:**

*Numerical Field Table:*

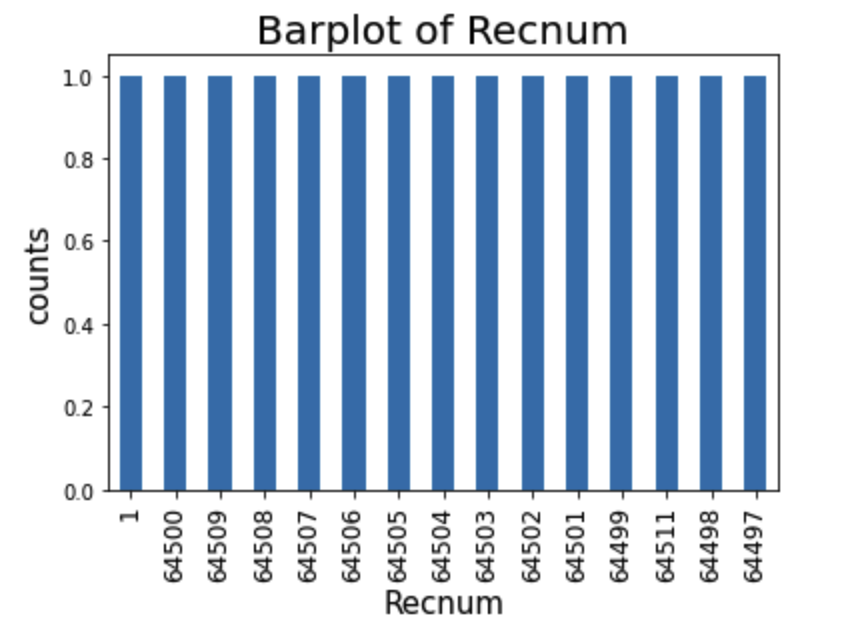
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Field Name** | **%**  **Populated** | **Min** | **Max** | **Mean** | **Stdev** | **%Zero** |
| date | 100 | 2006/01/01 | 2006/12/31 | N/A | N/A | 0 |
| amount | 100 | 0.01 | 3102045.53 | 427.886 | 10006.140 | 0 |

*Categorical Field Table:*

|  |  |  |  |
| --- | --- | --- | --- |
| **Field Name** | **%**  **Populated** | **#Unique Values** | ***Most Common Value*** |
| recnum | 100 | 96,753 | N/A |
| Cardnum | 100 | 1,645 | 5142148452 |
| merchnum | 100 | 13092 | 930090121224 |
| merch description | 100 | 13126 | GSA-FSS-ADV |
| merch state | 98.765 | 228 | TN |
| merch zip | 95.188 | 4568 | 38118.0 |
| transtype | 100 | 4 | P |
| fraud | 100 | 2 | 0 |

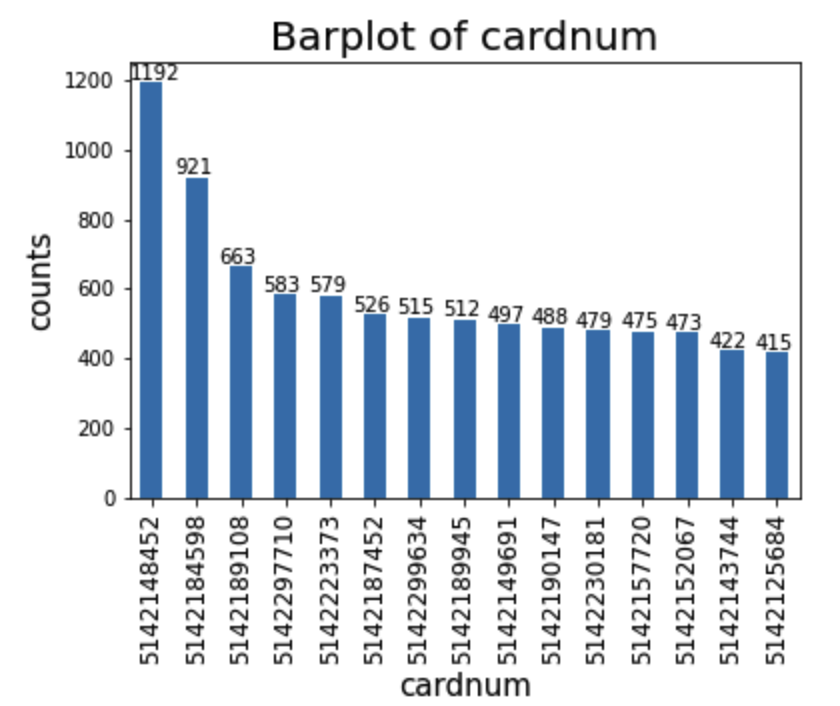
1. **Distribution for each field**
2. **recnum**

The graph below is the first 15 most repeated observations in the ‘recnum’ field. However, we can observe that they all have a count of 1. It is because field ‘recnum’ has unique values for each rows (represent the record of each observations).



1. **Cardnum**

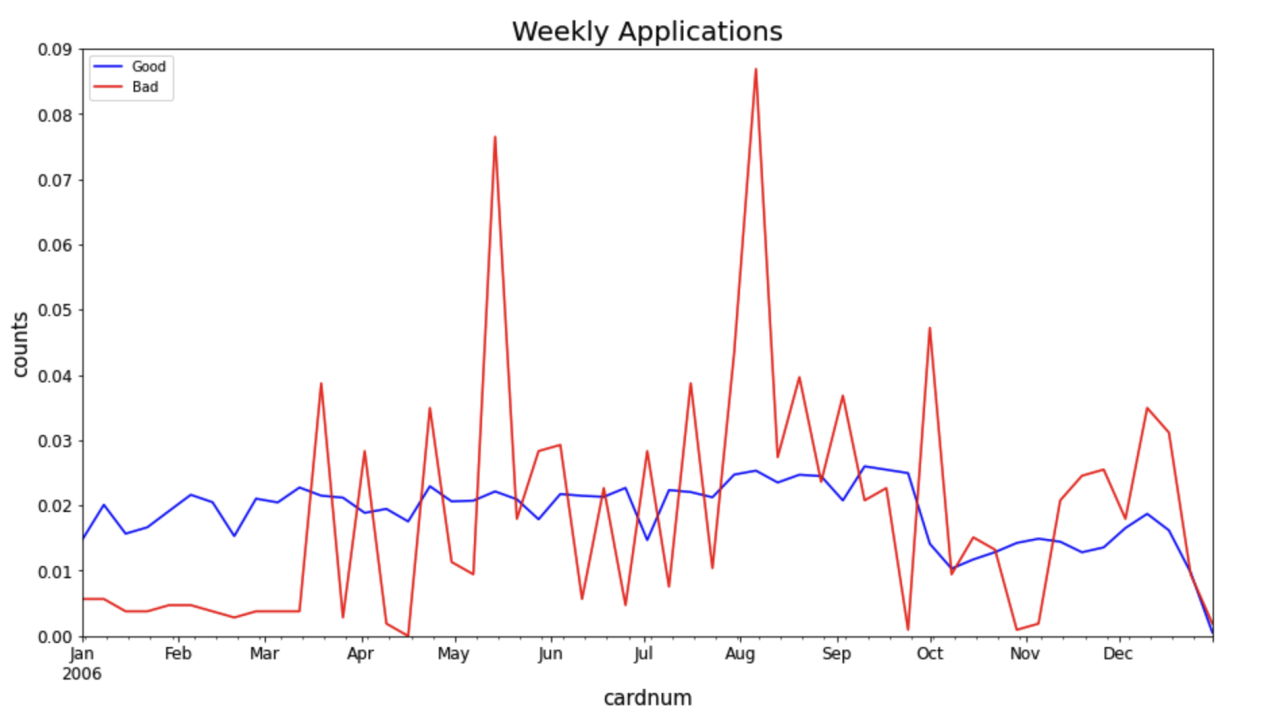
The graph below is the top 15 most repeated observations in the ‘cardnum’.



1. **Date**

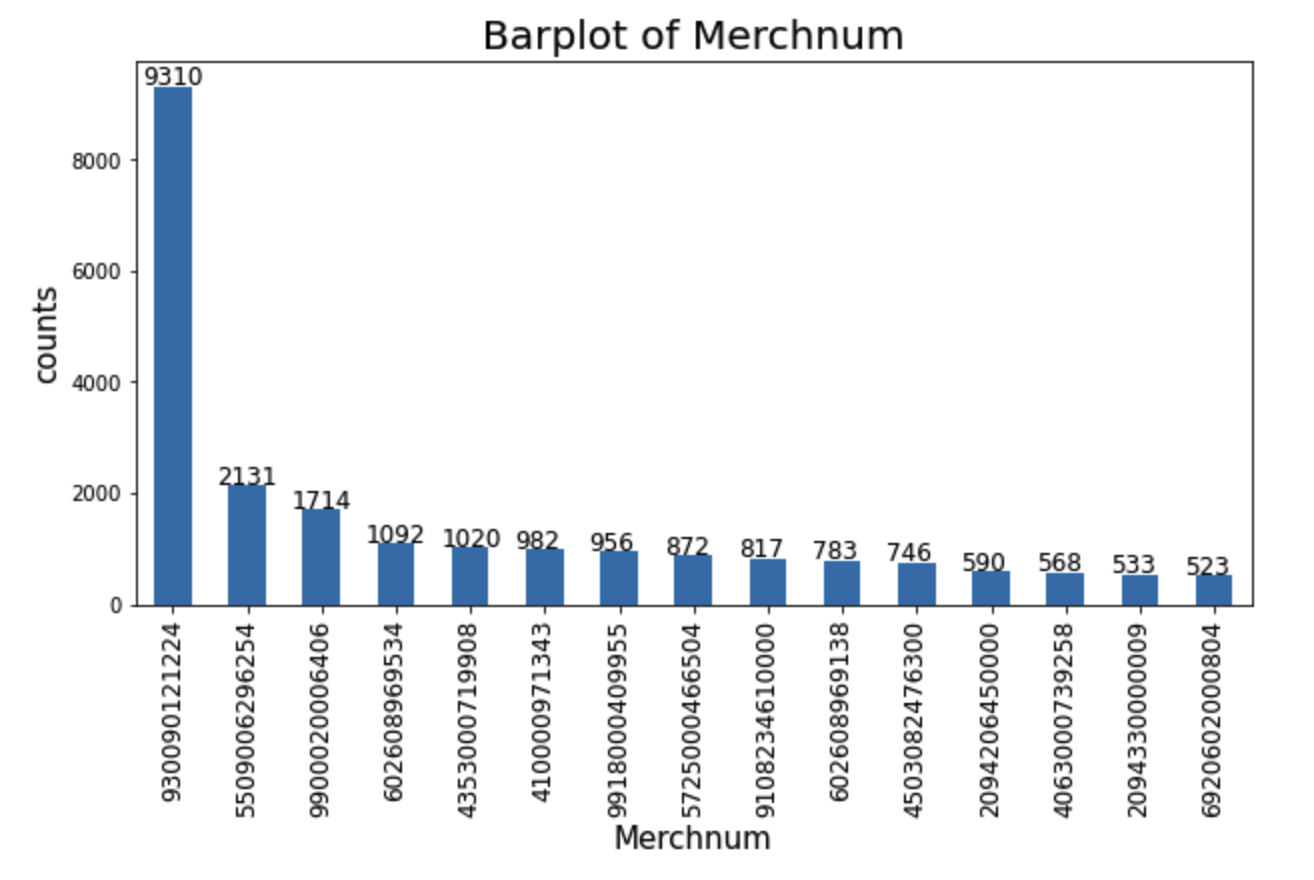
The graph below is the line plot of the good rate and the bad rate within a week (Weekly bad/good rate).

Noted: Please ignore the last point on the below graph

****

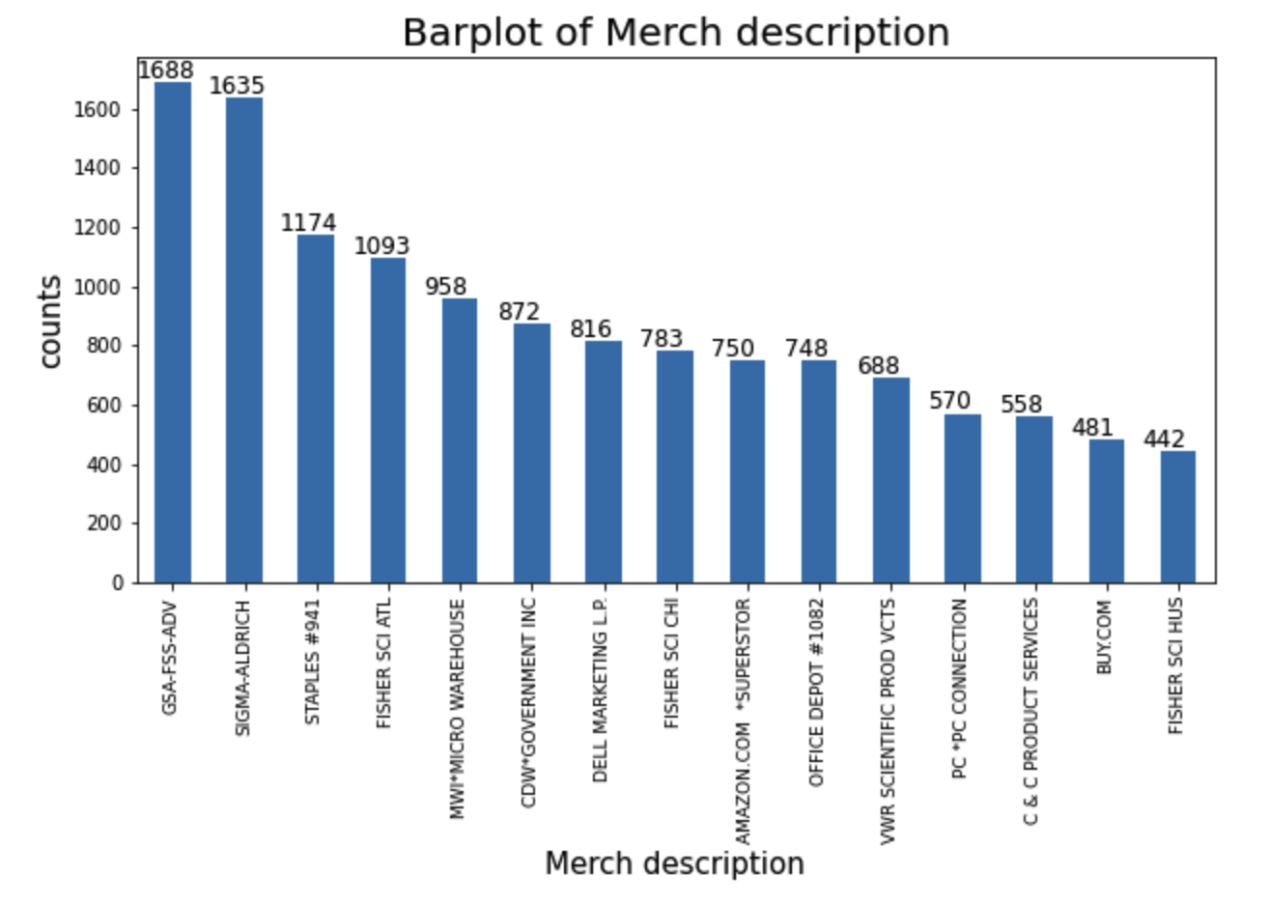
1. **Merchnum**

The graph below is the top 15 most repeated observations in the ‘Merchnum’ field.

****

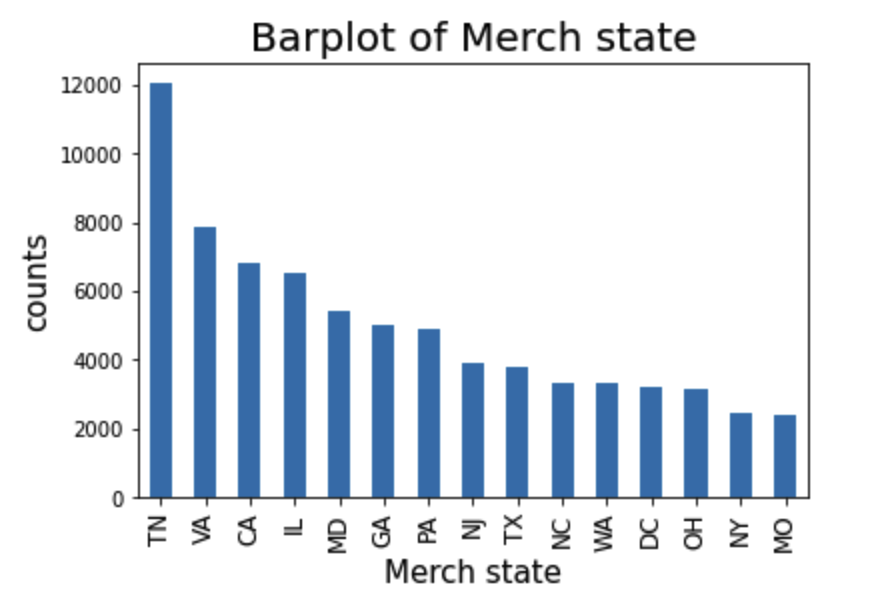
1. **Merch description**

The graph below is the top 15 most repeated observations in the ‘Merch description’ field.

****

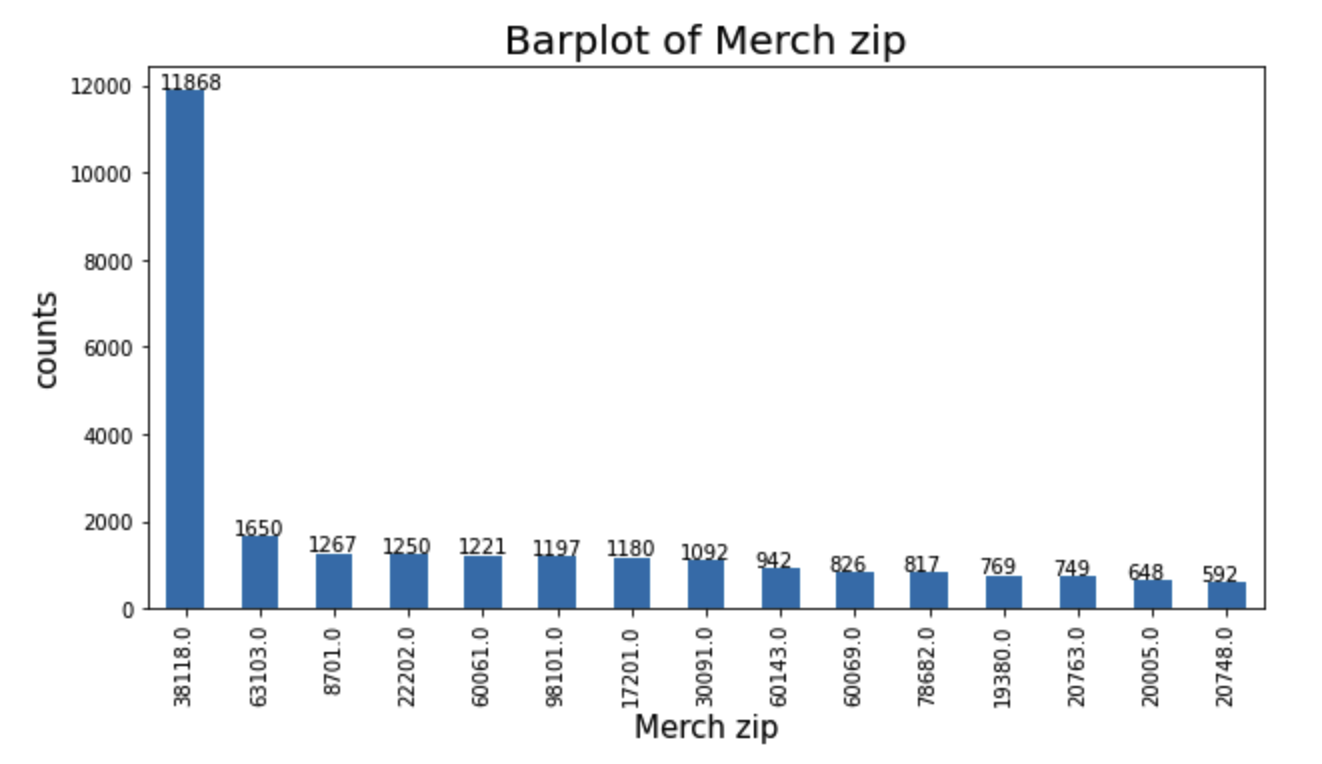
1. **Merch state**

The graph below is the top 15 most repeated observations in the ‘Merch state’ field.

****

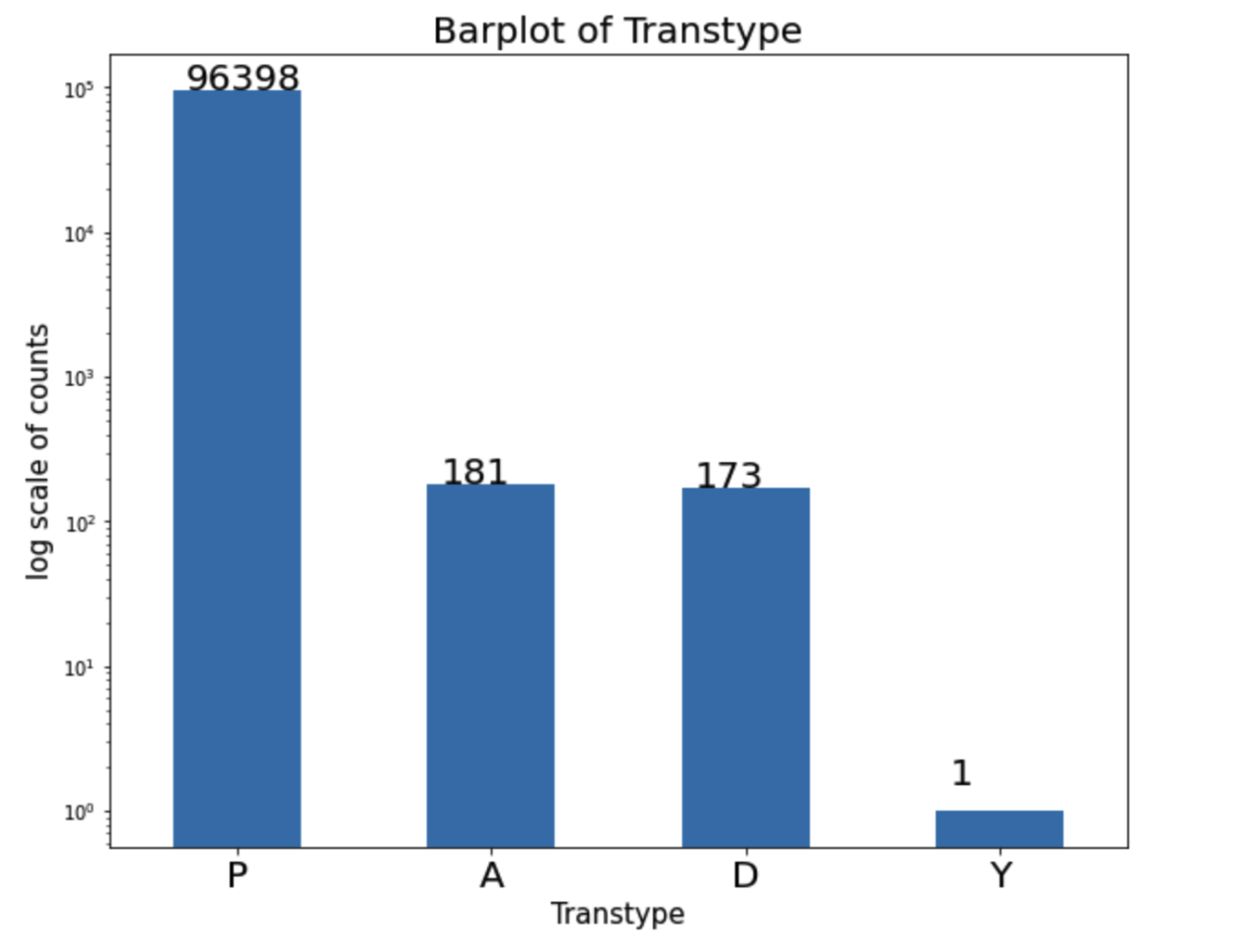
1. **Merch zip**

The graph below is the top 15 most repeated observations in the ‘Merch zip’ field.

****

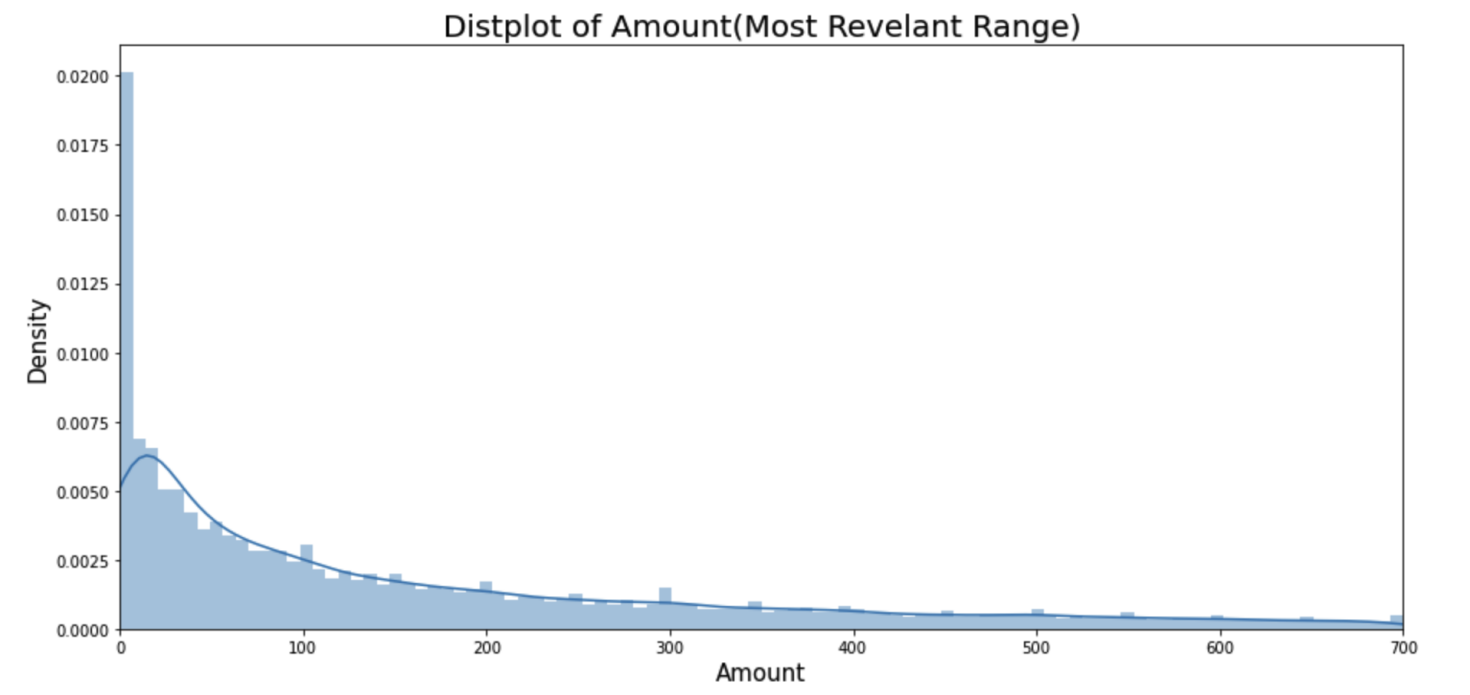
1. **Transtype**

The graph below is the top 15 most repeated observations in the ‘transtype’ field with log scaled y-axis for better visualizations.

****

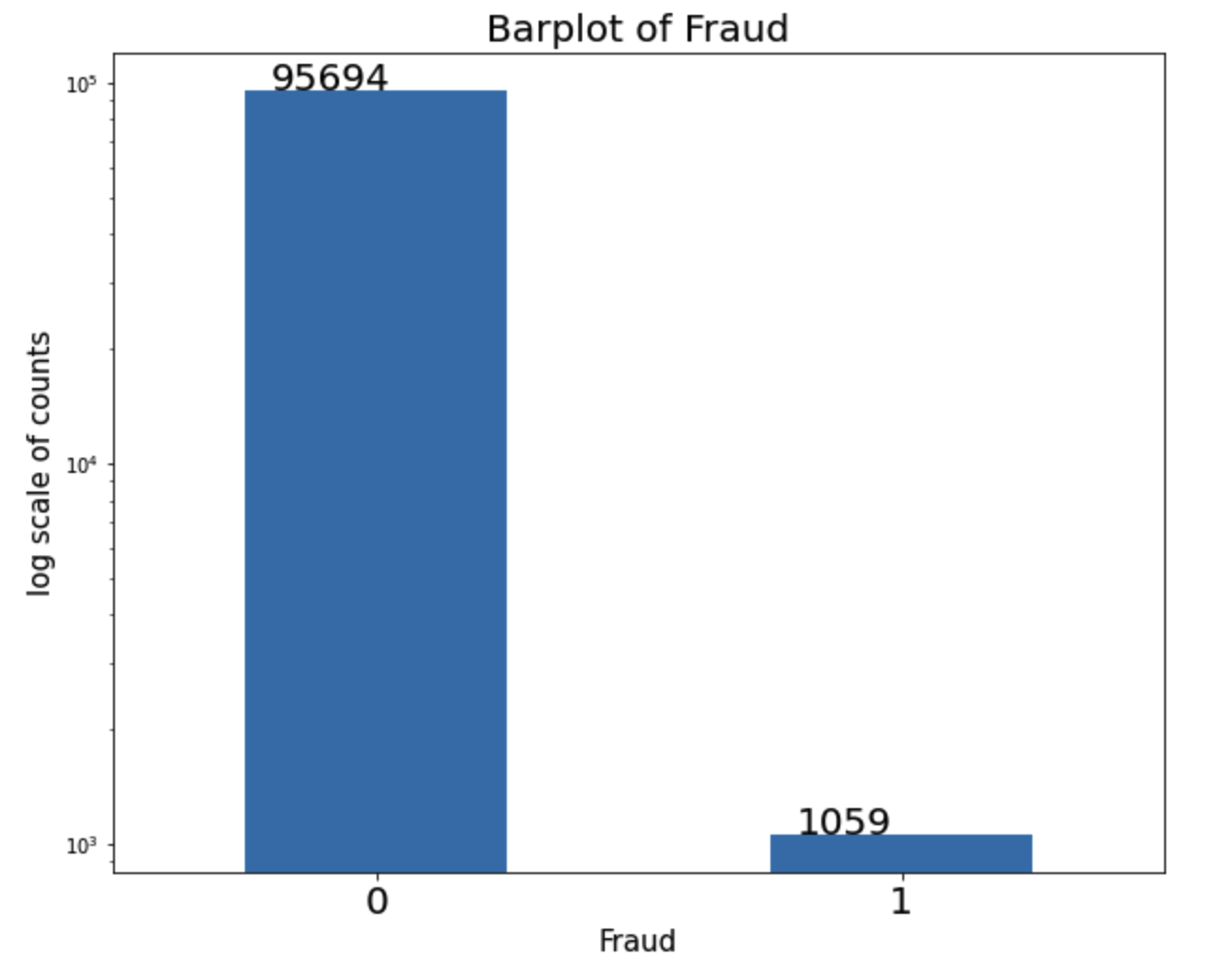
1. **Amount**

After the observation of the most relevant range for the 'Amount’ field. I choose 0-700 for the x range. The graph below is the distribution of the most relevant Amount observations.

****

1. **Fraud**

The graph below is the bar plot of the ‘Fraud’ field with log scaled y-axis for better visualizations.

****