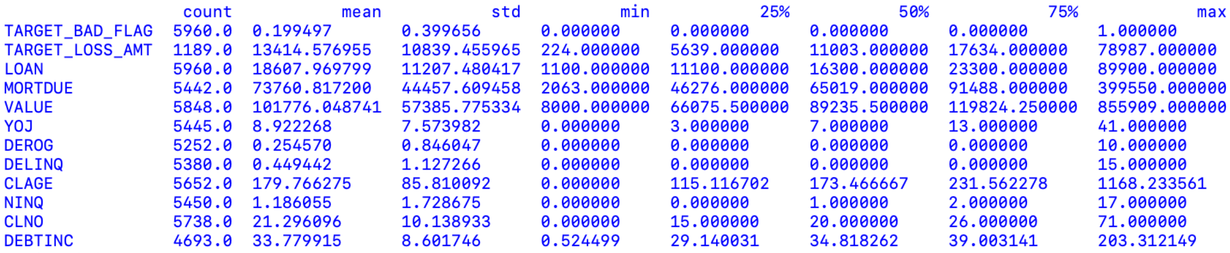
Vivian Xia

Assignment 1: Data Preparation – Graphs and Statistical Output

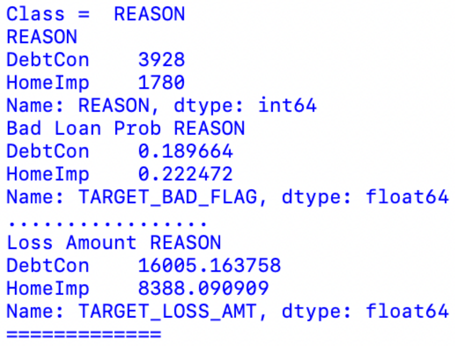
Descriptive Statistics:



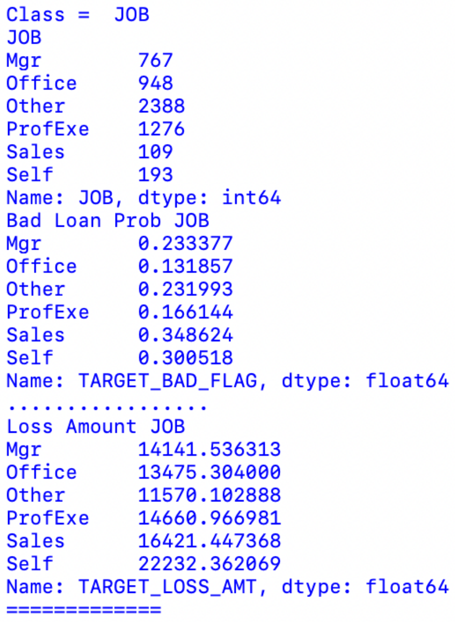
```
                 count           mean           std          min           25%           50%            75%            max
TARGET_BAD_FLAG 5960.0  0.199497       0.399656      0.000000     0.000000      0.000000      0.000000       1.000000
TARGET_LOSS_AMT 1189.0  13414.576955   10839.455965  224.000000   5639.000000   11003.000000  17634.000000   78987.000000
LOAN            5960.0  18607.969799   11207.480417  1100.000000  11100.000000  16300.000000  23300.000000   89900.000000
MORTDUE         5442.0  73760.817200   44457.609458  2063.000000  46276.000000  65019.000000  91488.000000   399550.000000
VALUE           5848.0  101776.048741  57385.775334  8000.000000  66075.500000  89235.500000  119824.250000  855909.000000
YOJ             5445.0  8.922268       7.573982      0.000000     3.000000      7.000000      13.000000      41.000000
DEROG           5252.0  0.254570       0.846047      0.000000     0.000000      0.000000      0.000000       10.000000
DELINQ          5380.0  0.449442       1.127266      0.000000     0.000000      0.000000      0.000000       15.000000
CLAGE           5652.0  179.766275     85.810092     0.000000     115.116702    173.466667    231.562278     1168.233561
NINQ            5450.0  1.186055       1.728675      0.000000     0.000000      1.000000      2.000000       17.000000
CLNO            5738.0  21.296096      10.138933     0.000000     15.000000     20.000000     26.000000      71.000000
DEBTINC         4693.0  33.779915      8.601746      0.524499     29.140031     34.818262     39.003141      203.312149
```

The TARGET_LOSS_AMT may have outliers considering the increase from the 75% to max value. Its outliers may be attributed to outliers in other variables including LOAN, MORTDUE, VALUE, CLAGE, CLNO, DEBTINC as their max value compared to its 75% is also significantly greater.

Probability of a loan being defaulted and the loss amount from the categorical variables:



```
Class =  REASON
REASON
DebtCon    3928
HomeImp    1780
Name: REASON, dtype: int64
Bad Loan Prob REASON
DebtCon    0.189664
HomeImp    0.222472
Name: TARGET_BAD_FLAG, dtype: float64
.................
Loss Amount REASON
DebtCon    16005.163758
HomeImp    8388.090909
Name: TARGET_LOSS_AMT, dtype: float64
=============
```

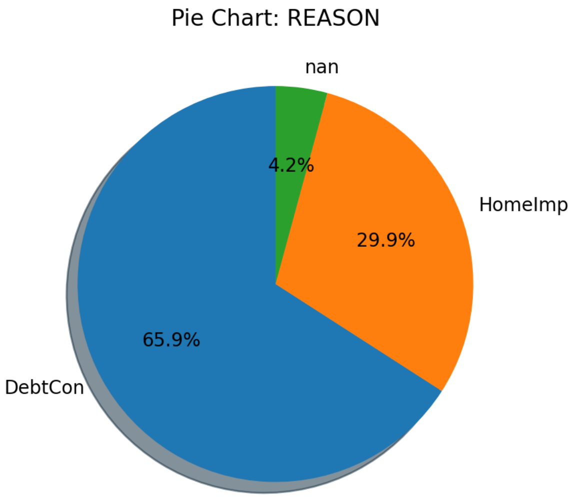


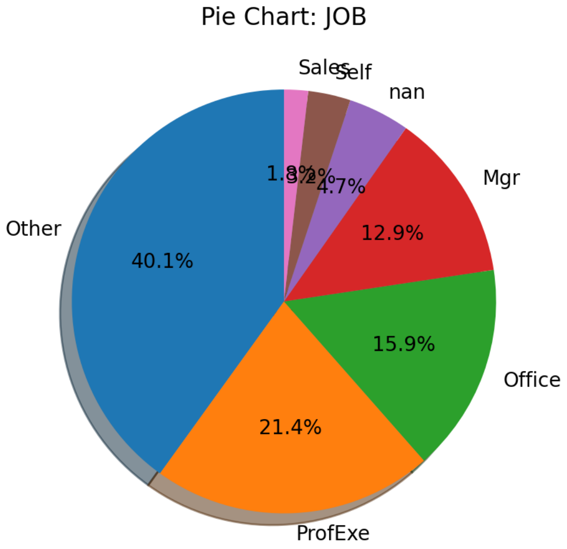
```
Class =  JOB
JOB
Mgr        767
Office     948
Other      2388
ProfExe    1276
Sales      109
Self       193
Name: JOB, dtype: int64
Bad Loan Prob JOB
Mgr        0.233377
Office     0.131857
Other      0.231993
ProfExe    0.166144
Sales      0.348624
Self       0.300518
Name: TARGET_BAD_FLAG, dtype: float64
.................
Loss Amount JOB
Mgr        14141.536313
Office     13475.304000
Other      11570.102888
ProfExe    14660.966981
Sales      16421.447368
Self       22232.362069
Name: TARGET_LOSS_AMT, dtype: float64
=============
```

For REASON, there is a higher probability that the loan was defaulted because of HomeImp (home improvement) compared to DebtCon (debt consolidaton). On the other hand, the loss amount for DebtCon is double that of HomeImp. DebtCon is also the most common reason compared to the HomeImp.

For JOB, there is a higher probability that the loan was defaulted to those who are in Sales or Self (self-employed). The loss amount of those in the category Self is significantly larger than that of those in other occupations. There is not much data in Sales and Self, 109 and 193 respectively, compared to the number in other categories. It would be helpful to see if the data is well-represented in their customers in Sales and Self occupation. There are a lot of customers in the Other category.

Pie Chart:

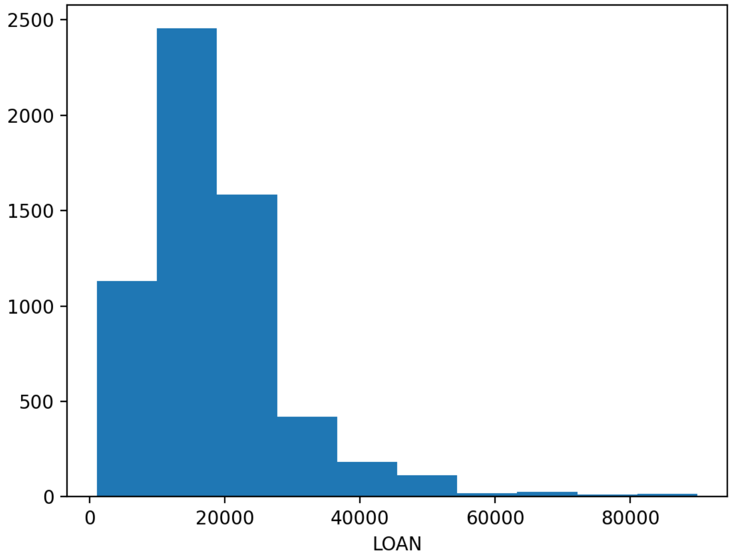


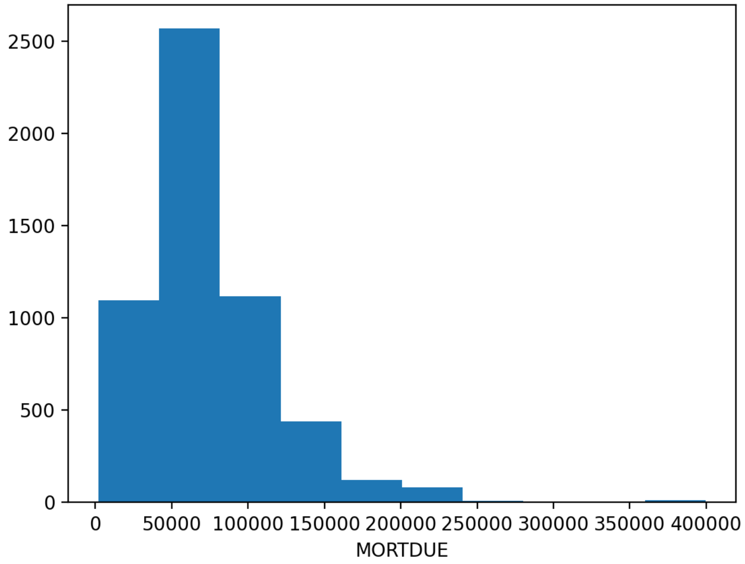


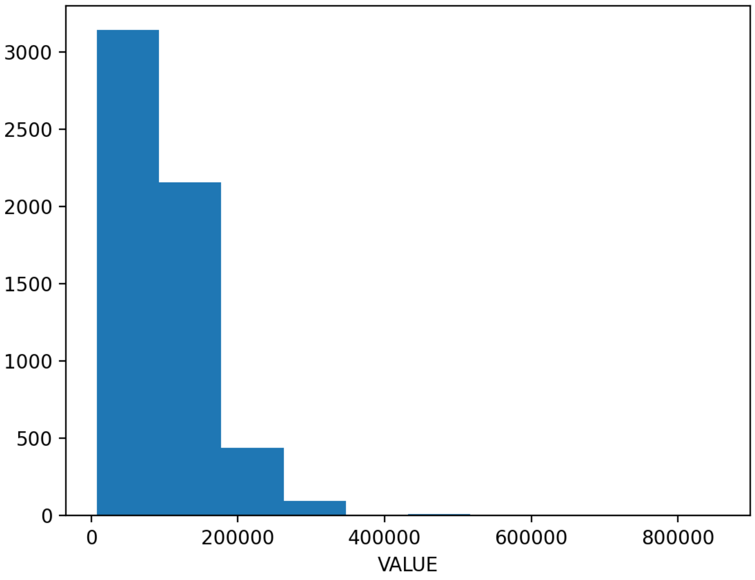
As mentioned above, the most common REASON for the loan is DebtCon. The most common JOB of the customer is Other. Sales and Self are the least common.

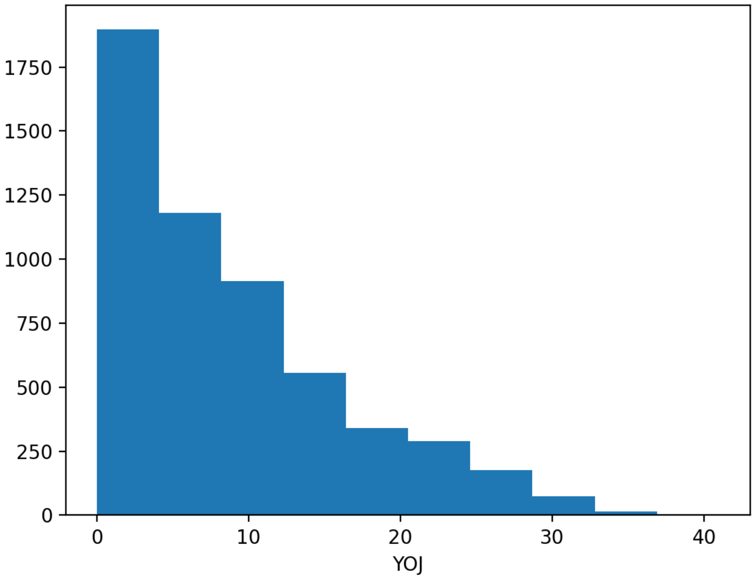
There are missing values in variables REASON and JOB.

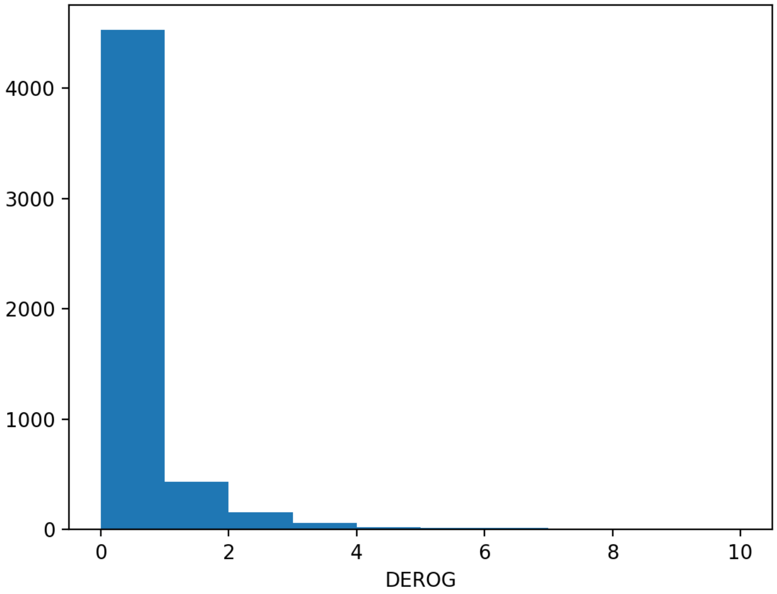
Histogram:

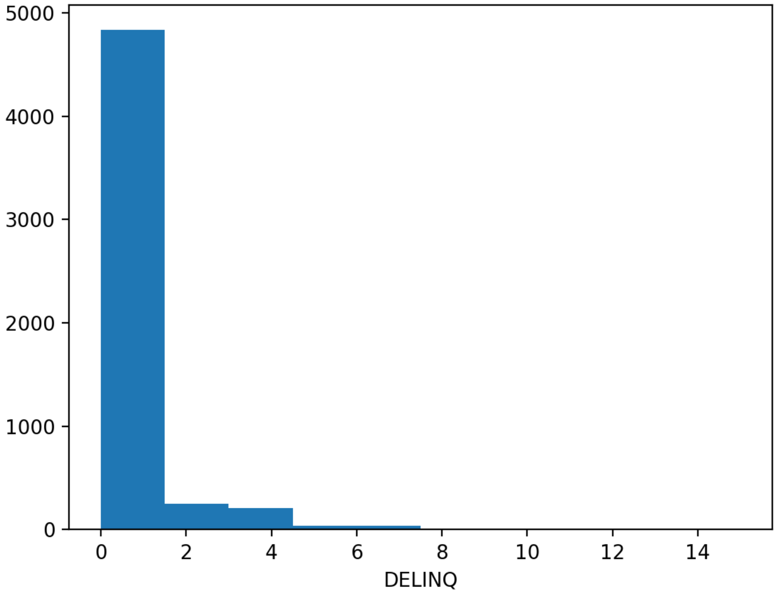

2500
2000
1500
1000
500
0
0
20000
40000
60000
80000
LOAN

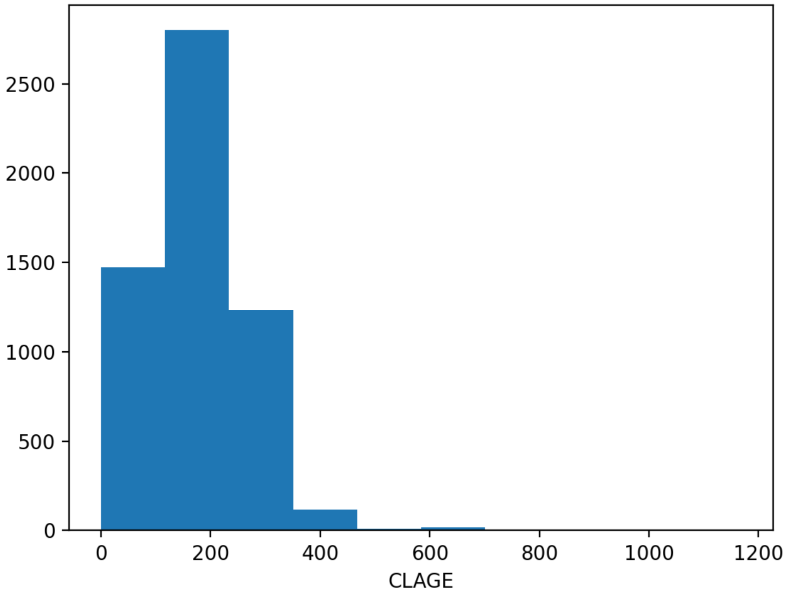

2500
2000
1500
1000
500
0
0
50000
100000
150000
200000
250000
300000
350000
400000
MORTDUE

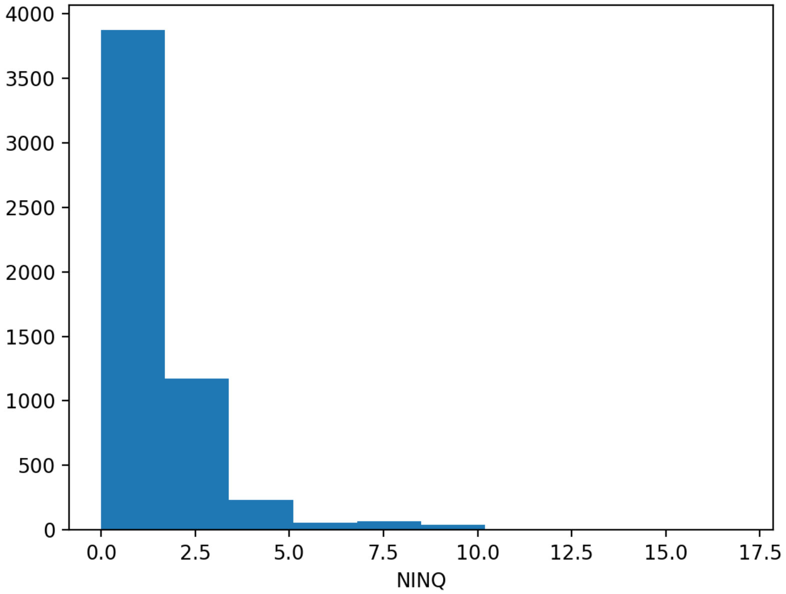

3000
2500
2000
1500
1000
500
0
0
200000
400000
600000
800000
VALUE

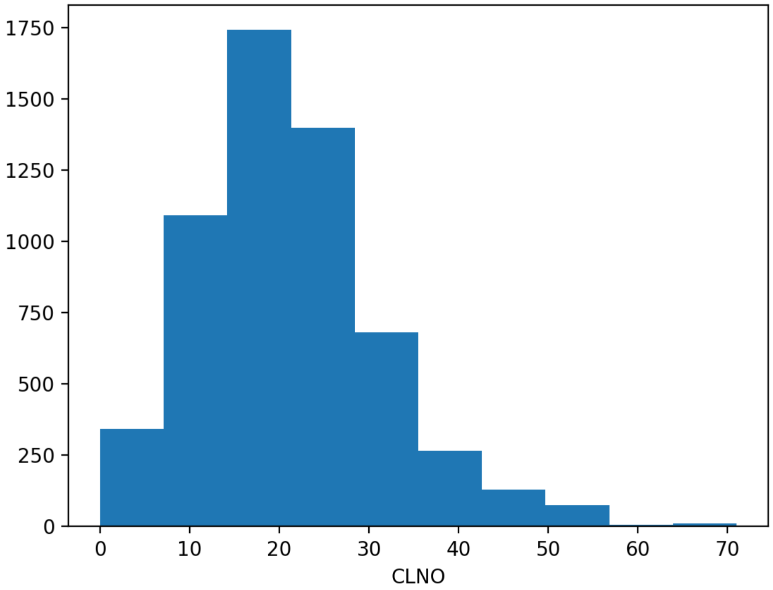

1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
YOJ

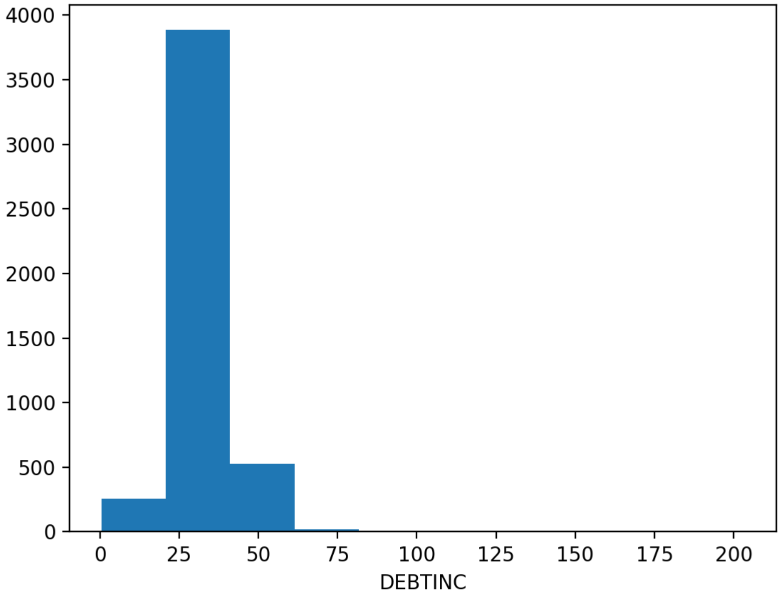

4000
3000
2000
1000
0
0
2
4
6
8
10
DEROG

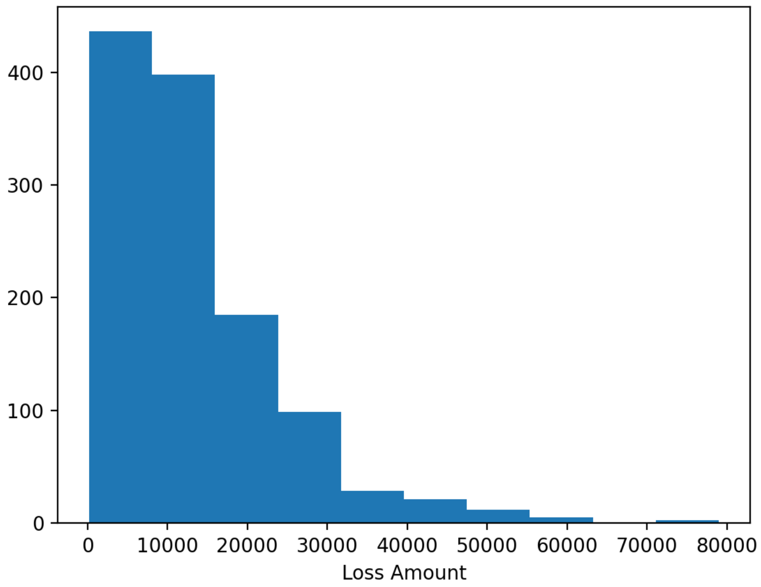

5000
4000
3000
2000
1000
0
0
2
4
6
8
10
12
14
DELINQ


2500
2000
1500
1000
500
0
0
200
400
600
800
1000
1200
CLAGE


4000
3500
3000
2500
2000
1500
1000
500
0
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
NINQ


1750
1500
1250
1000
750
500
250
0
0
10
20
30
40
50
60
70
CLNO


4000
3500
3000
2500
2000
1500
1000
500
0
0
25
50
75
100
125
150
175
200
DEBTINC



All the histogram distributions are right or positively skewed. The skewness is a result of outliers to the right that contribute to that long right tail. It may be helpful to transform the numerical data to mitigate the effect of outliers on the distribution.