INVESTMENT CASE STUDY

SUBMISSION

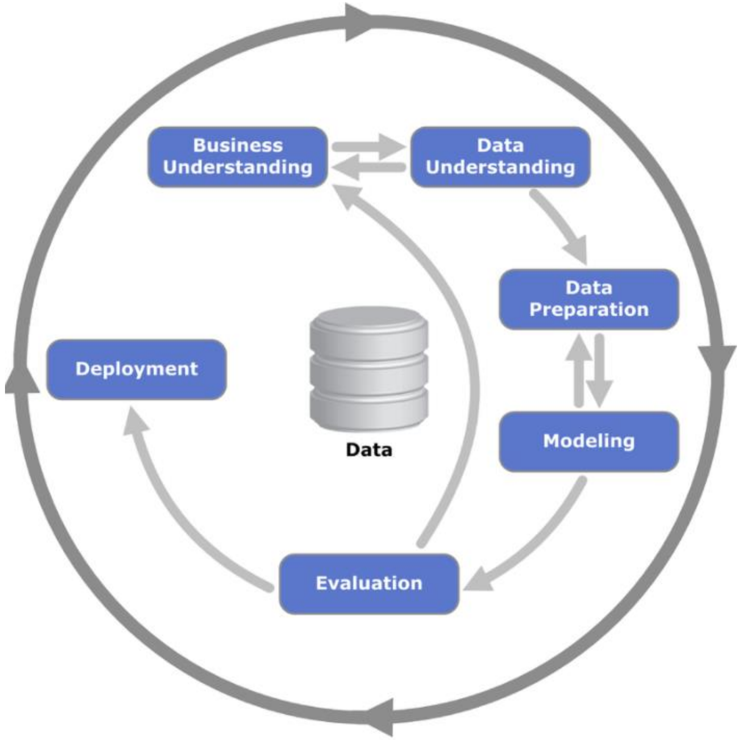
Group Members:

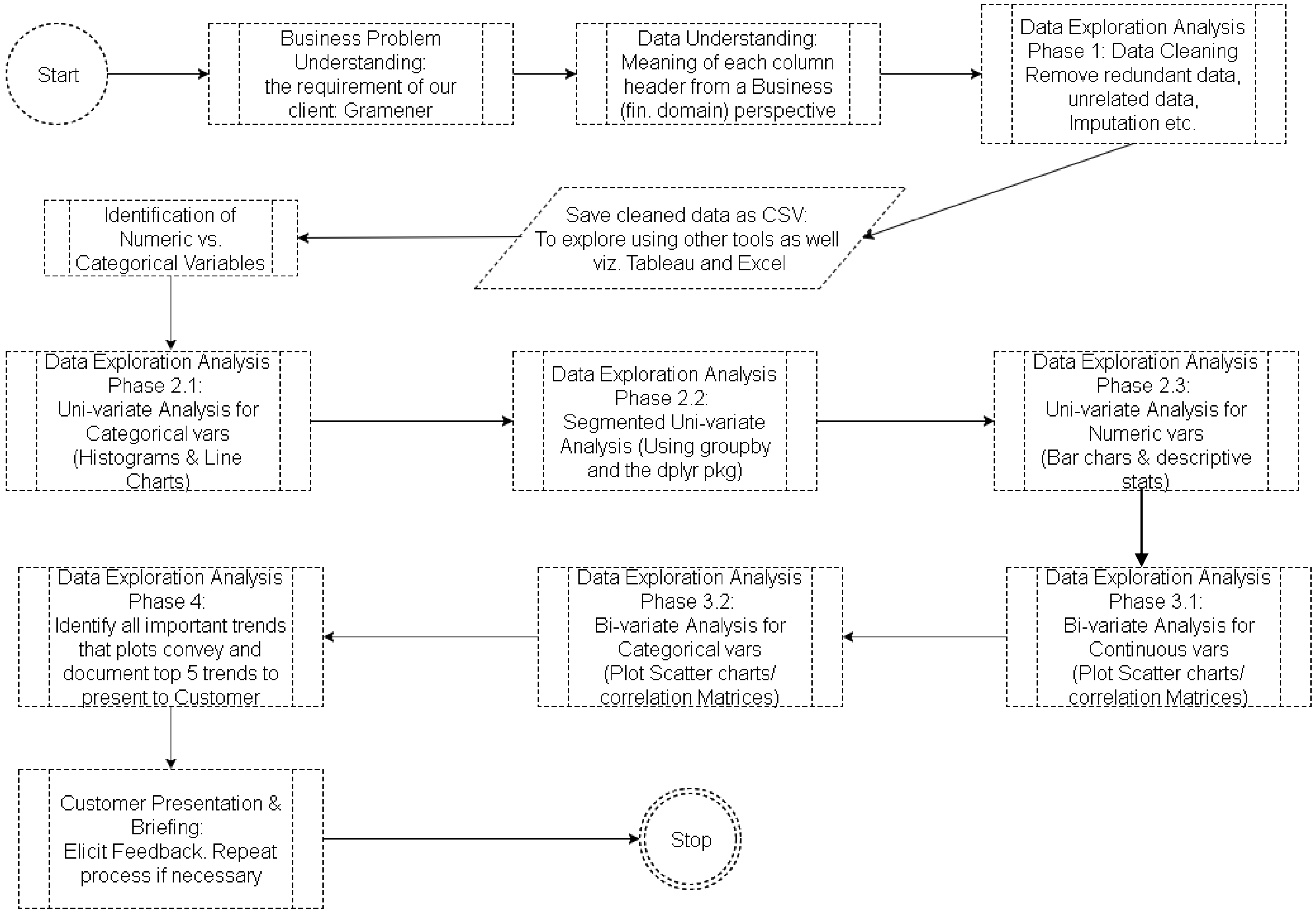
1. Sudeep Upadhya
2. Rishabh
3. Vikash Prasad
4. Amit Mankikar

Case Study Objectives

- Identification of Loan Applicant traits that tend to ‘default’ paying back
- Understand the ‘Driving Factors’ or ‘Driver Variables’ behind Loan Default phenomena
- Gramener may choose to utilize this knowledge for its portfolio and risk assessment of new loan applicants

Problem solving methodology using CRISP-DM





Data Cleaning Steps

1. Verify the dates are imported correctly. If not convert using yearmon/ POSIXlt
2. Remove all columns that don't change. Justification: There is no variance, it cannot help us to determine the reason for default. We can save memory and analysis, plotting and data frame transformations are faster
3. Identify all columns that don't provide any value. E.g. the url column: It provides us a link to access more data, but we don't have a username/ passwd so it is useless
4. Remove all columns that need Text Processing. We are not going to perform any NLP/ NLU and therefore such columns are not very useful
5. Remove redundant columns. E.g. The purpose of loan is a drop down which is already a categorical variable. We don't need the title column as it becomes redundant. We can identify all such columns and remove them.
6. Identify all columns that don't have any other value other than NA and 0. Remove such columns.
7. Since we are going to identify defaulter status based on Employer Name, we better scan through the emp_title column and consolidate all employer names. E.g. ARMY and US ARMY are same. Walmart, WALMART, Walmart and WAL-MART are same. We make use of gsub to convert all the upper case and remove all punctuation characters and spaces before we analyse this column.
8. Few columns such as Rate of Interest have been read as characters due to the presence of the "%" symbol. Convert all such instances to numeric
9. Since we are dealing with the aggregate, we may not need the primary keys in this analysis such as id and member_id. We can remove these as well.
10. Imputation: Identify all the NA values and replace them with appropriate value. We don't do this in the master frame but instead as and when that particular column is getting analysed.

Analysis

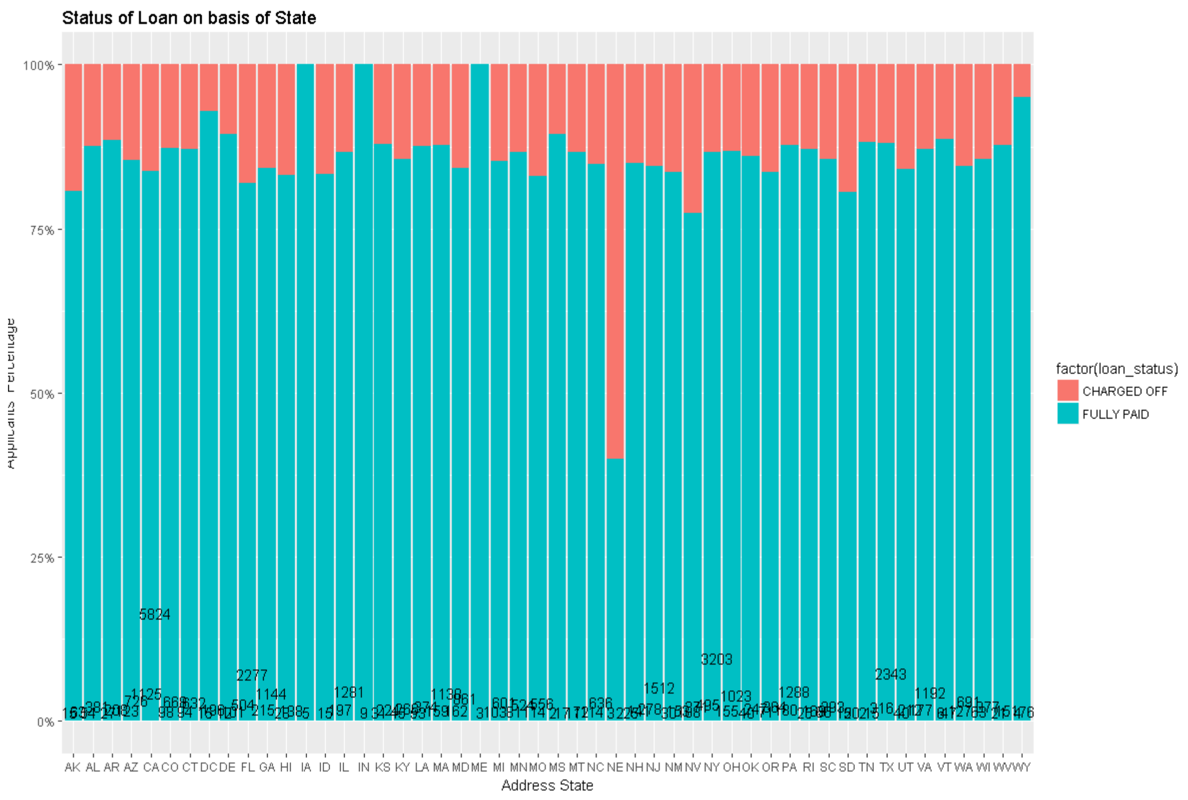
Uni-varate Analysis

- For Categorical Variables
- For Numeric Variables
- For Segmented Variables

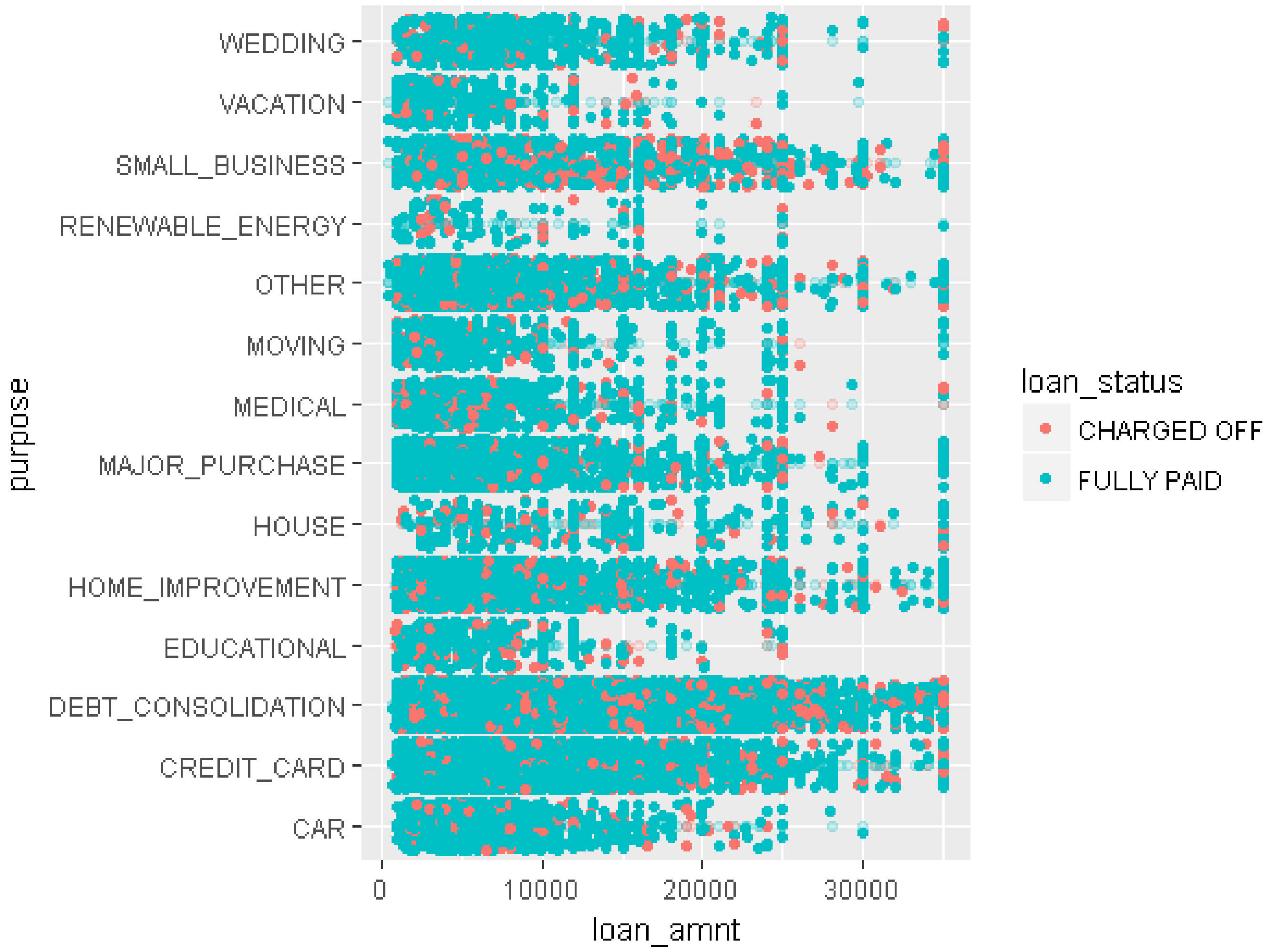
Bi-variate Analysis

- Keeping loan_status fixed in one of the columns
- Scatter plots

Conclusion: Top 5 Observations

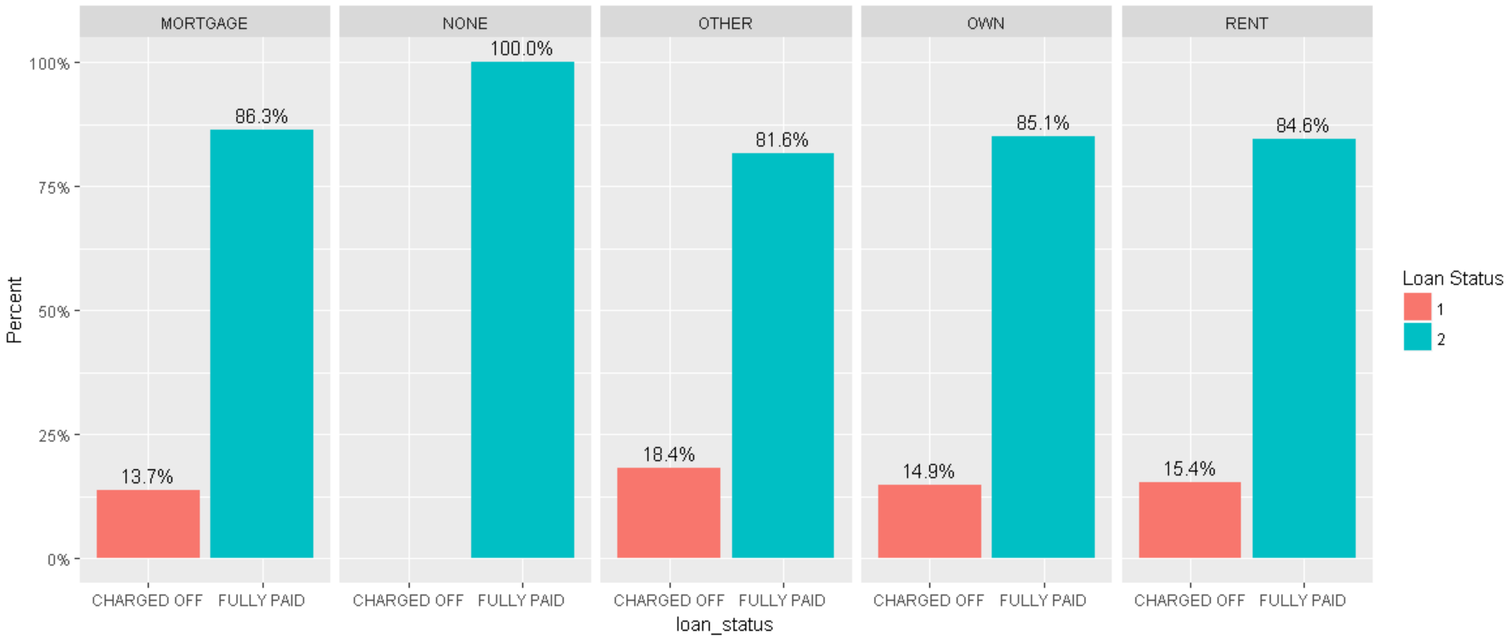


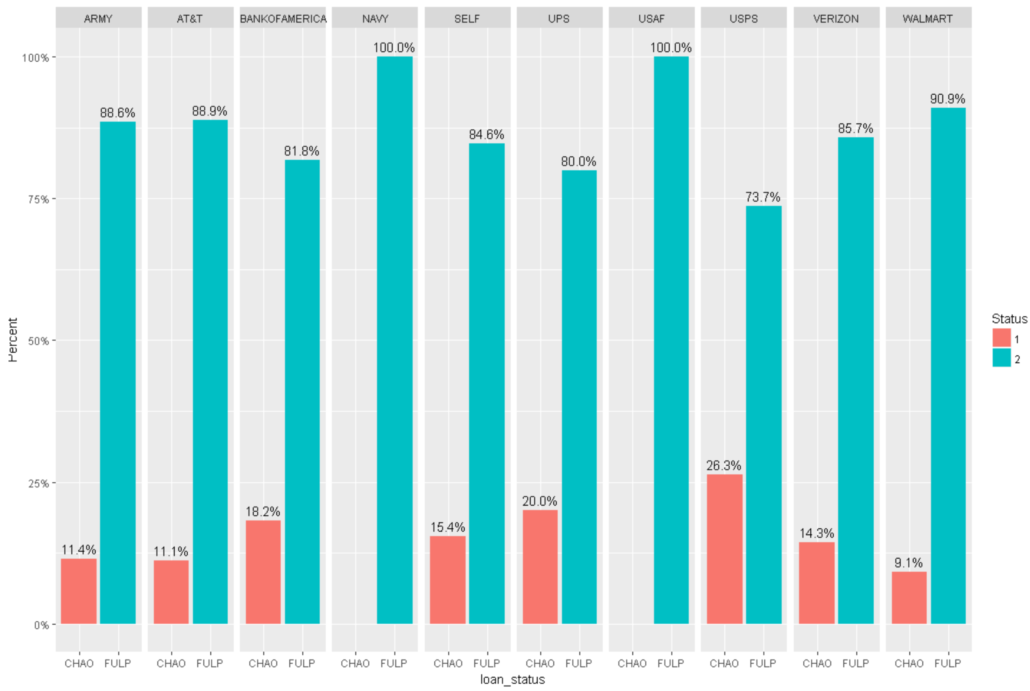
Insight 1: Most defaulting borrowers' are from **NEVADA**



Insight 2: Most defaulting borrows mention purpose as **SMALL BUSINESS**

Conclusion: Top 5 Observations

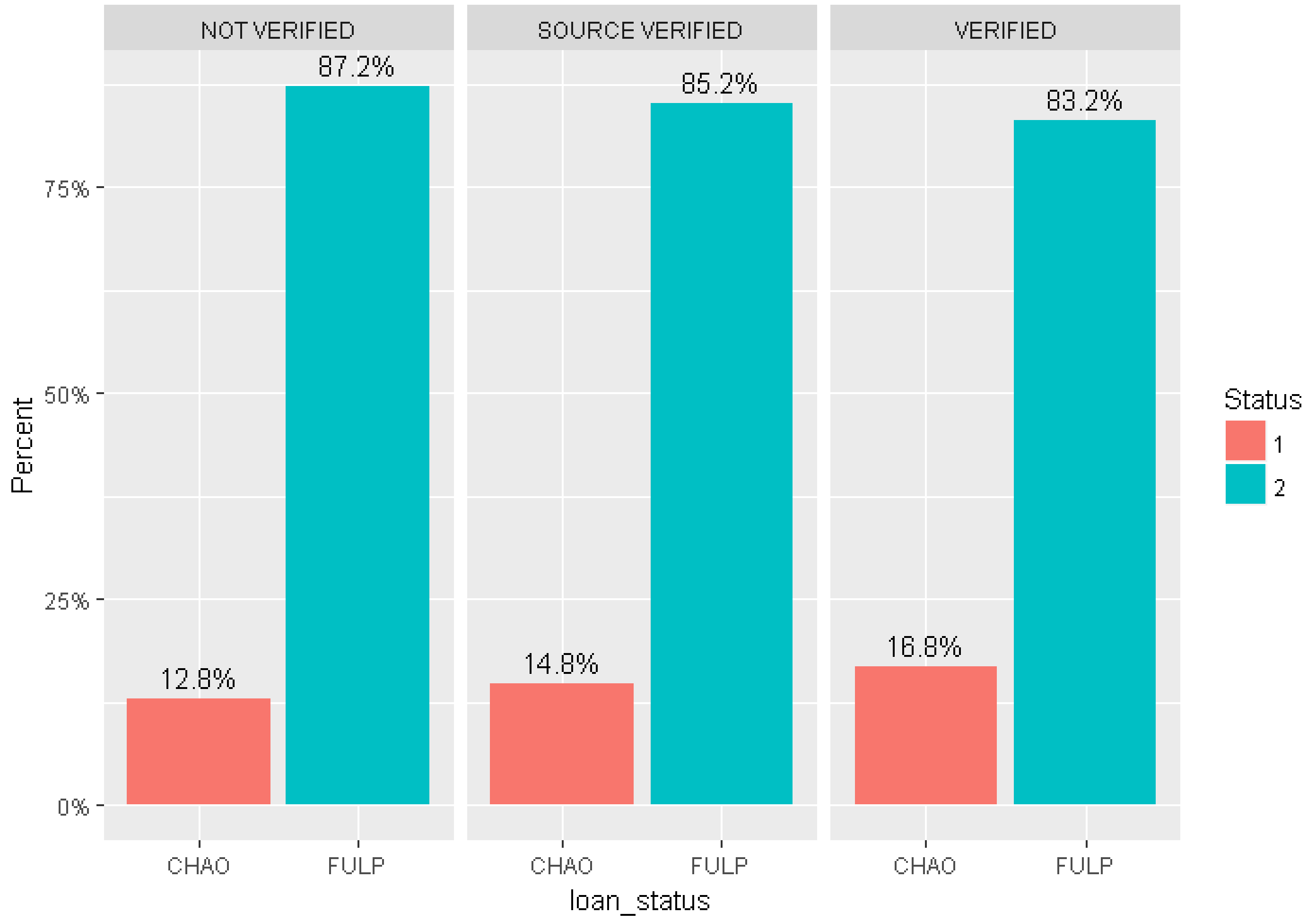




<u>Insight 3.</u> Most defaulting borrowers have **<u>"OTHERS"</u>** as ownership status

<u>Insight 4.</u> Most defaulting borrowers' are employed with <u>USPS</u>

Conclusion: Top 5 Observations



Insight 5 Even though, LC has verified borrowers and borrowers sources, the percentage of defaulters is higher than "Not Verified" cases.

Suggestion 1: Verification Process needs to be reviewed and improved accordingly

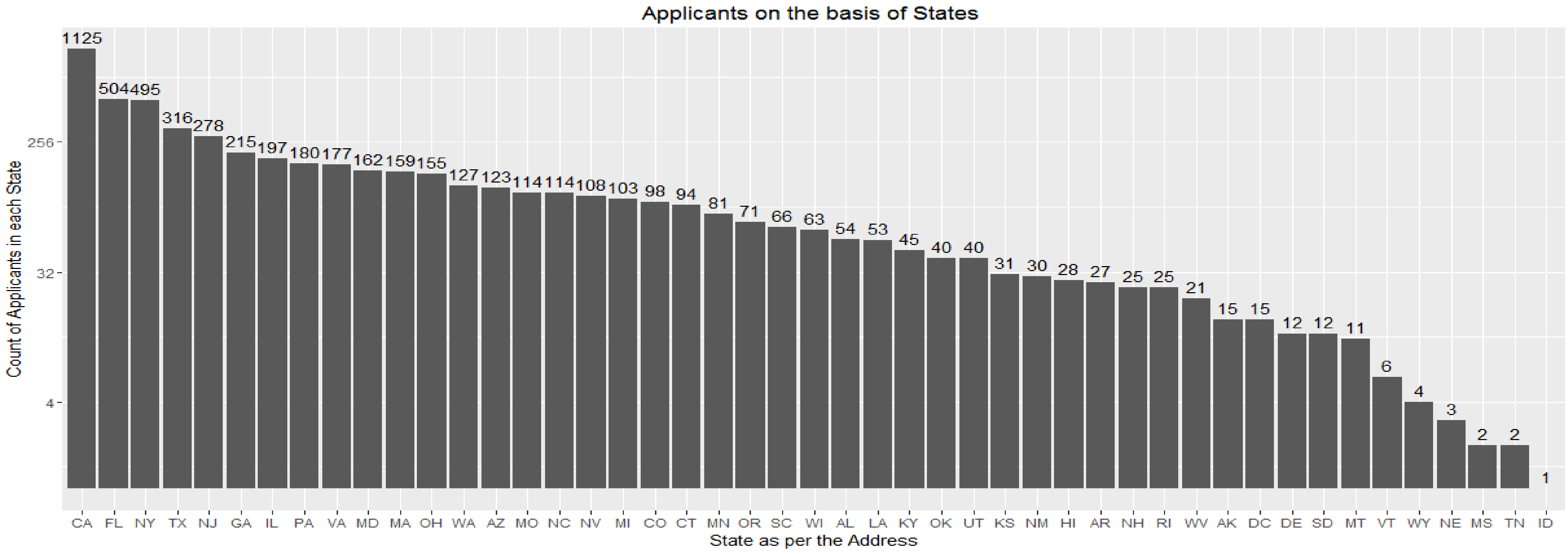
Suggestion 2: LC must target loan customers who are employed for lesser duration compared to customers who are employed for 10+ Years

Thank you!

All plots follow..

Univariate analysis 1/30

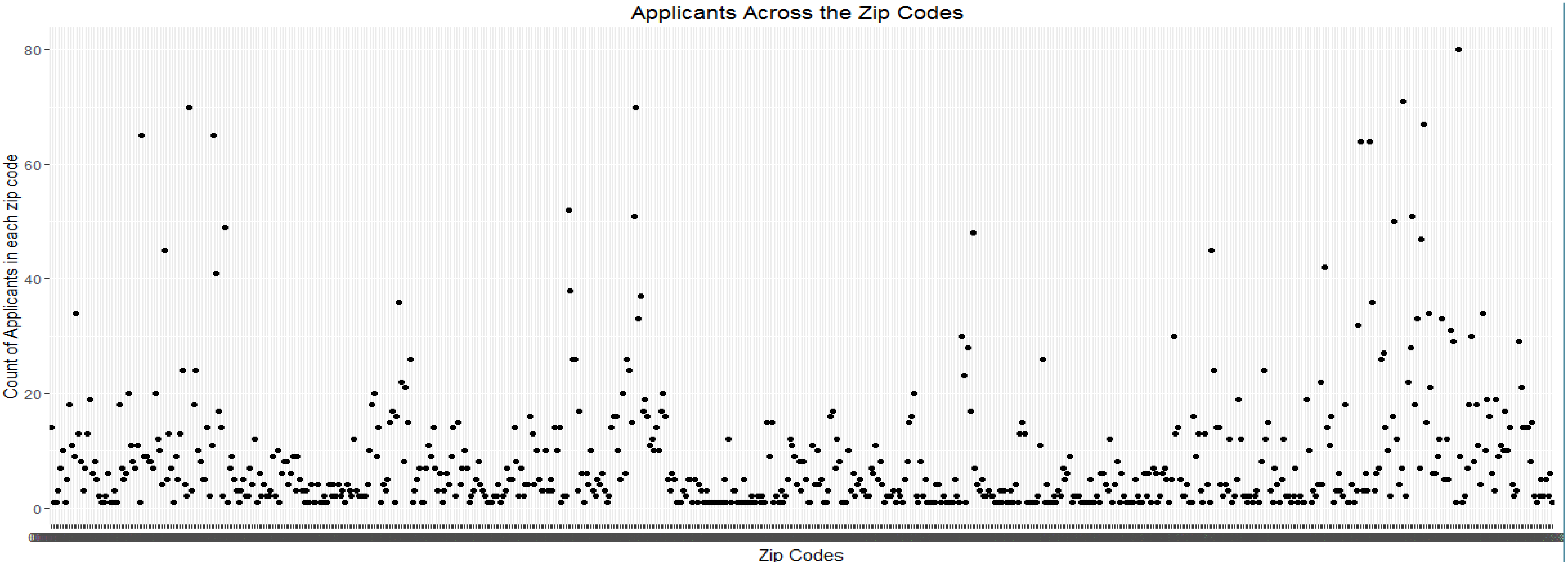
1. Log 2 Power Scale Histogram of CHARGED OFF loans per State



Observations: California, Florida, New York, Texas, New Jersey are top 5 affected states with loan status: “Charged Off”

Univariate analysis Contd. 2/30

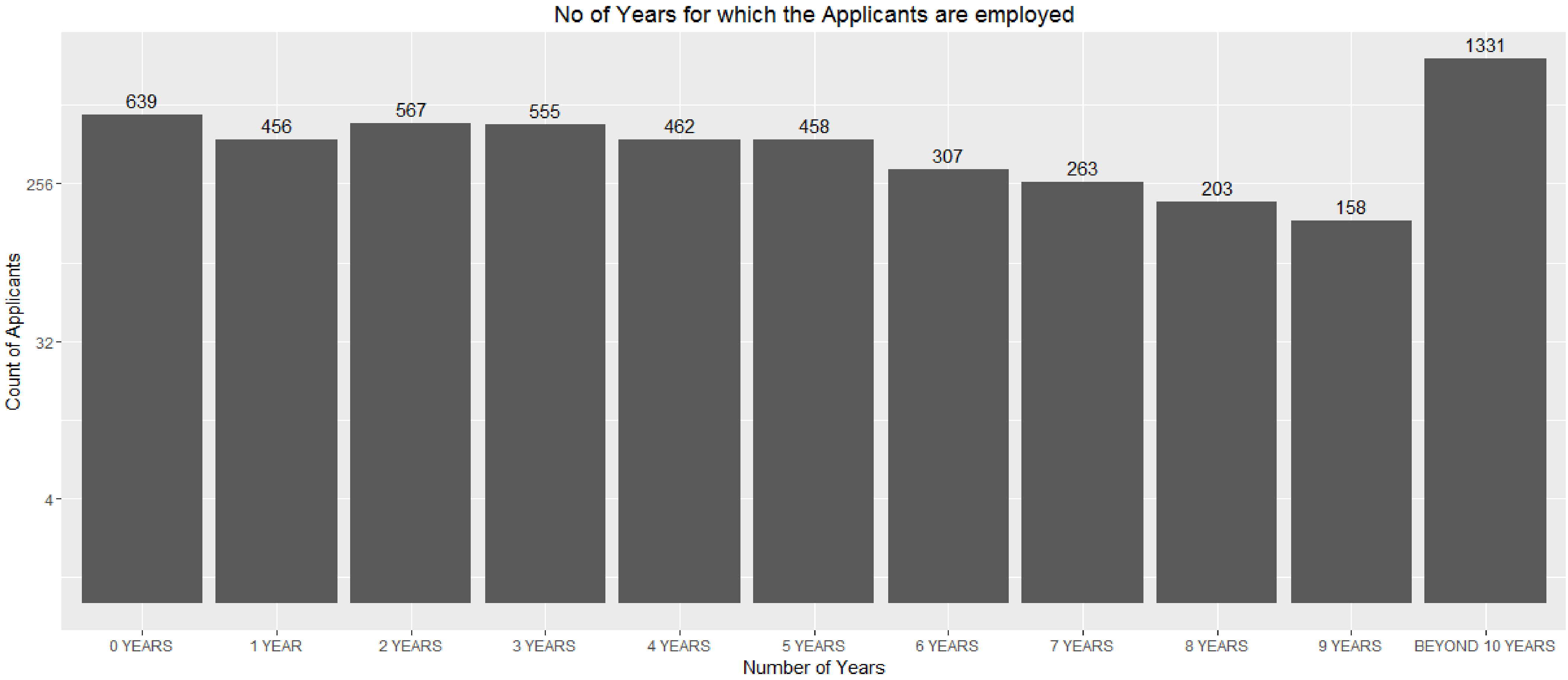
2. Scatter plot of Loan Charged Off zip code



Observations: Since zip codes are in ascending order in X axis we can see that no of "CHARGED OFF" loans are particularly high in 9xxxx. This is the state of California.

Univariate analysis Contd. 3/30

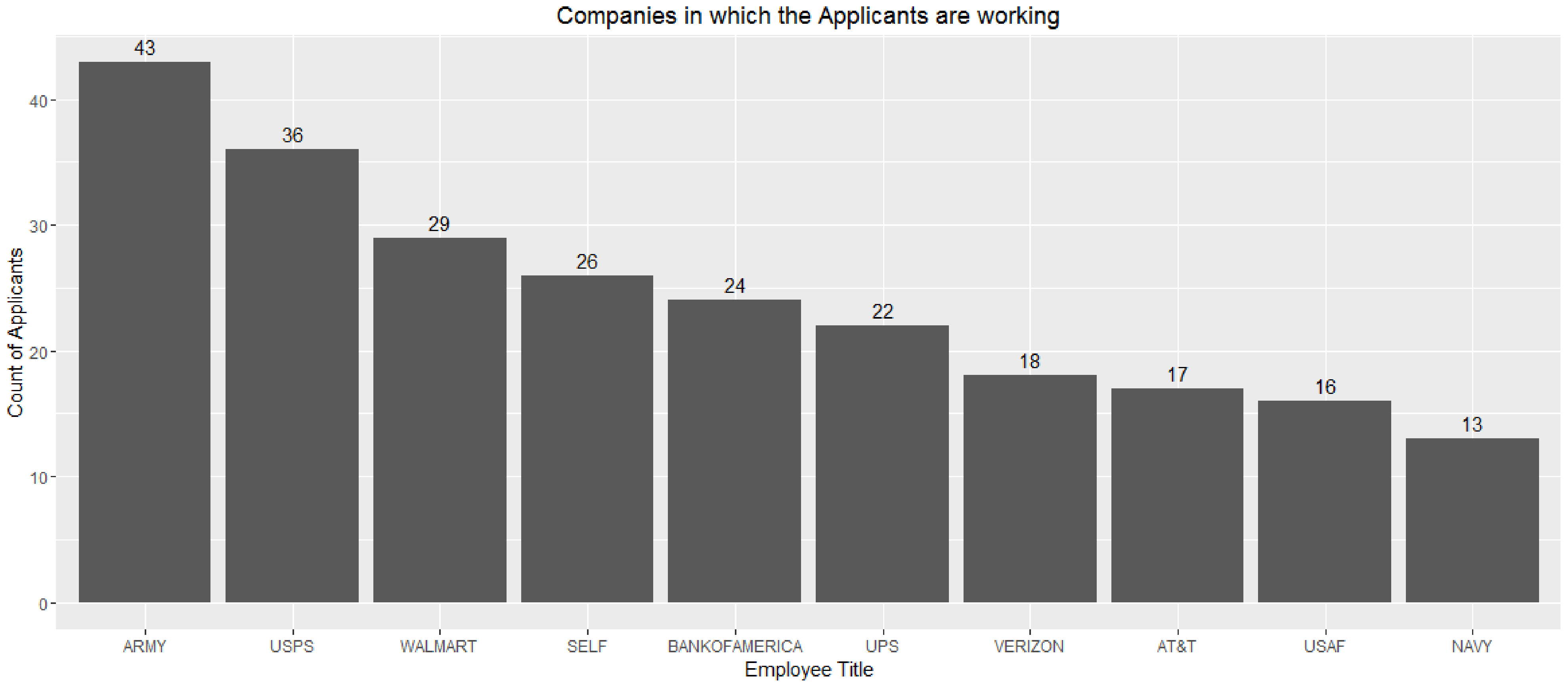
3. Histogram of Employment length in Charged Off Loans



Observations: Charged off loans has a decreasing trend w.r.t tenure. However the number of Charged Off loans within the first year is too high.

Univariate analysis Contd. 4/30

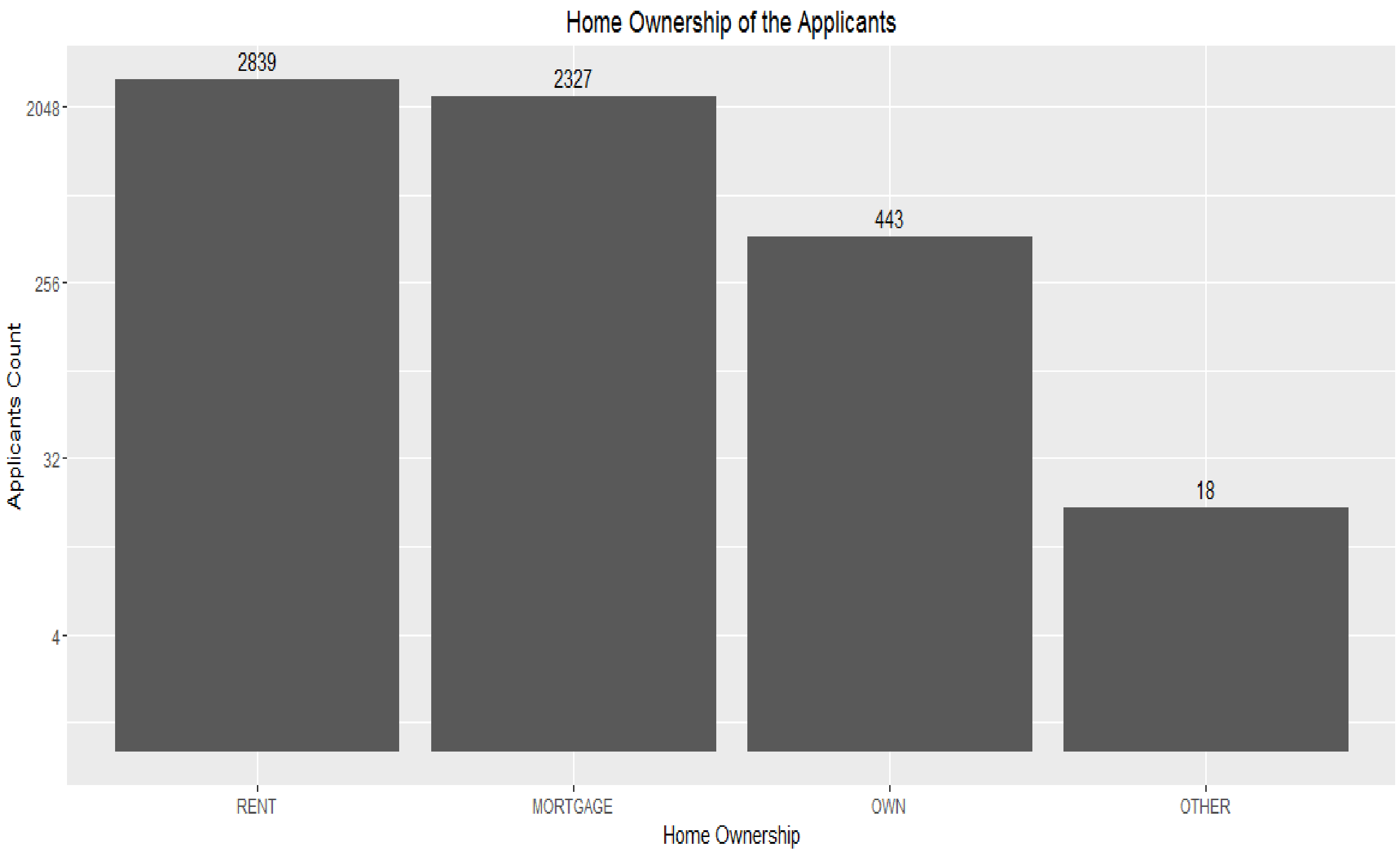
4. Histogram of Employer Name in Charged Off Loans



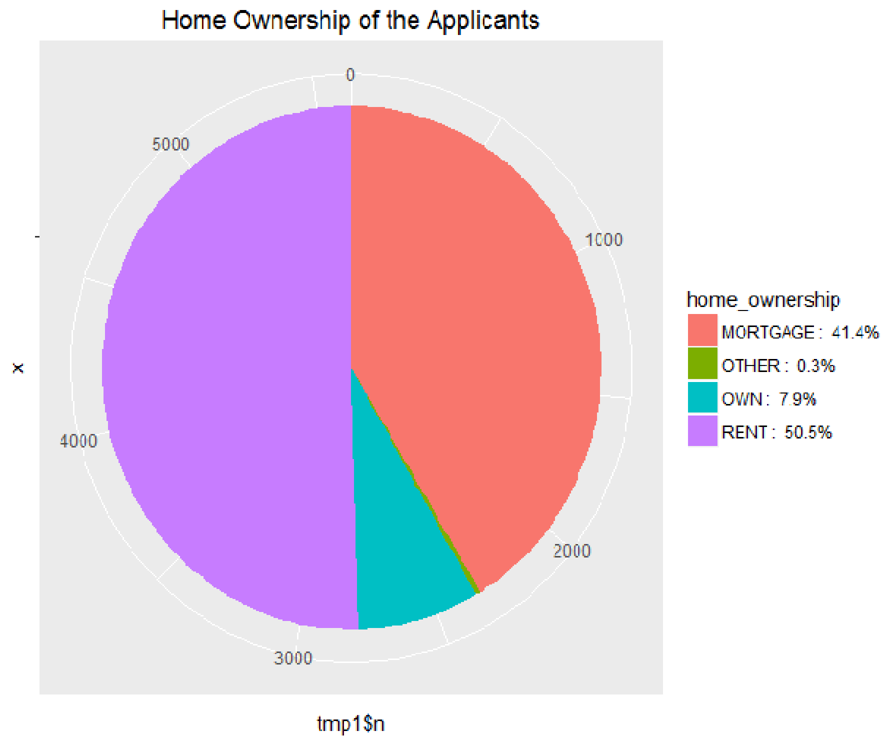
Observations: It is evident from above plot employees of which organizations are defaulting

Univariate analysis Contd. 5/30

5.1 Histogram of Home ownership status in Charged Off Loans



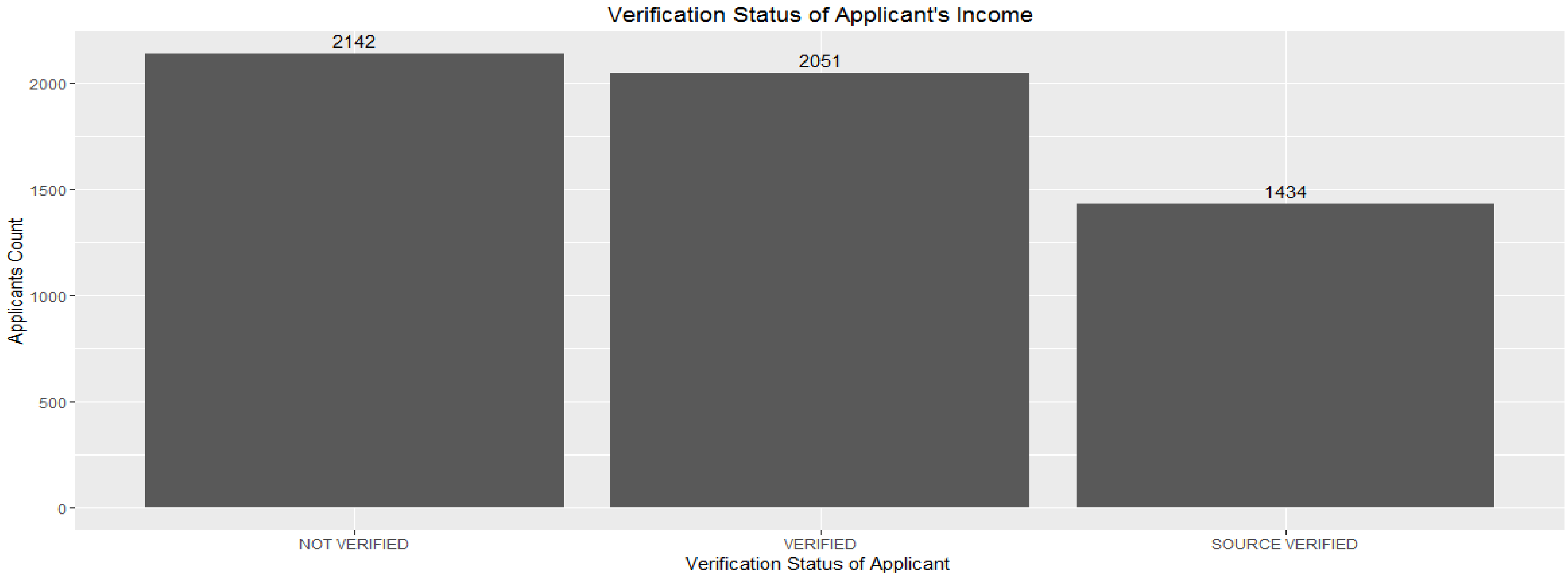
5.2 Pie Chart for comparison within category



Observations: People who live in rented accommodation constitute of 50.5% of the population

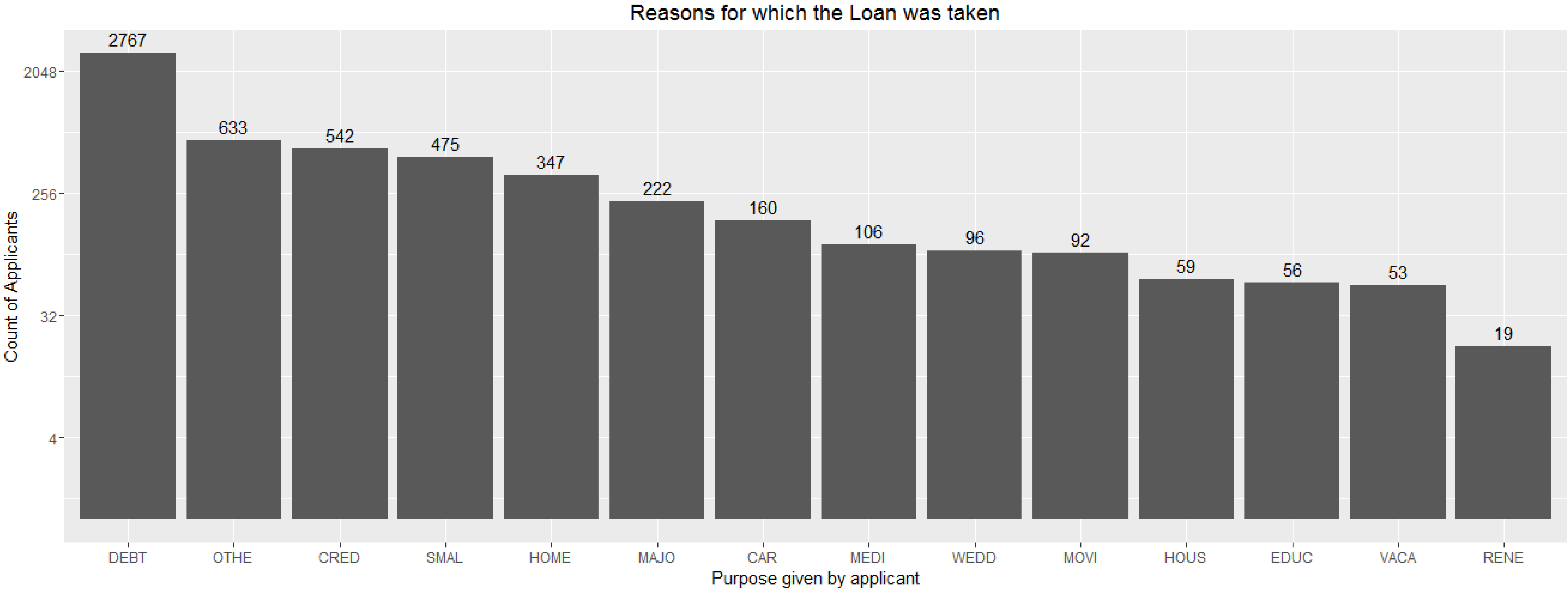
Univariate analysis Contd. 6/30

6. Histogram of Income Source Verification in Charged Off Loans



Observations: When the income source is verified, the Charged Off phenomenon seems to be marginally lower

Univariate analysis Contd. 7/30

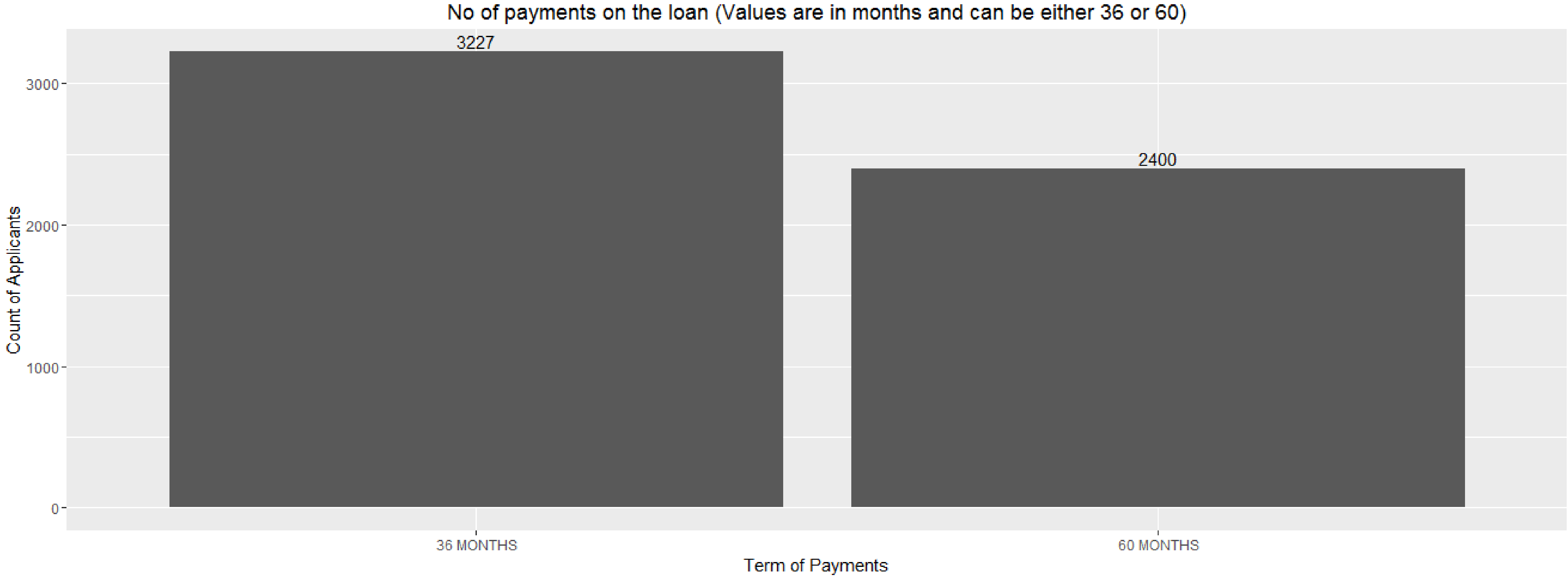


Observations: Note that this is Log 2 Power plot, and yet the purpose of "Debt Consolidation" is prominently high.

A high number of "Other" category also tells us that our data collection method is inadequate. We must add more categories to the selection drop down menu which is used at the time of applying for loan

Univariate analysis Contd. 8/30

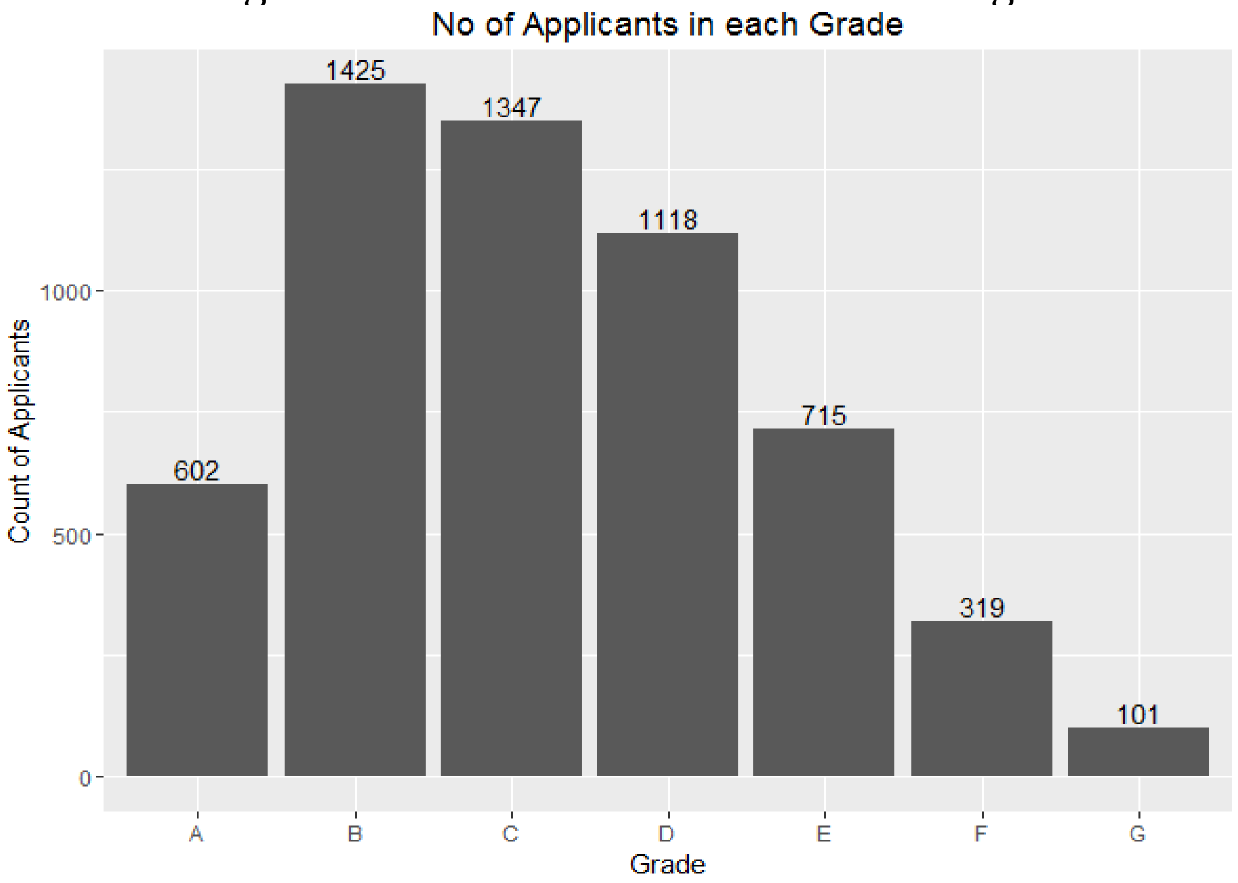
8. Histogram of Loan term in Charged Off Loans

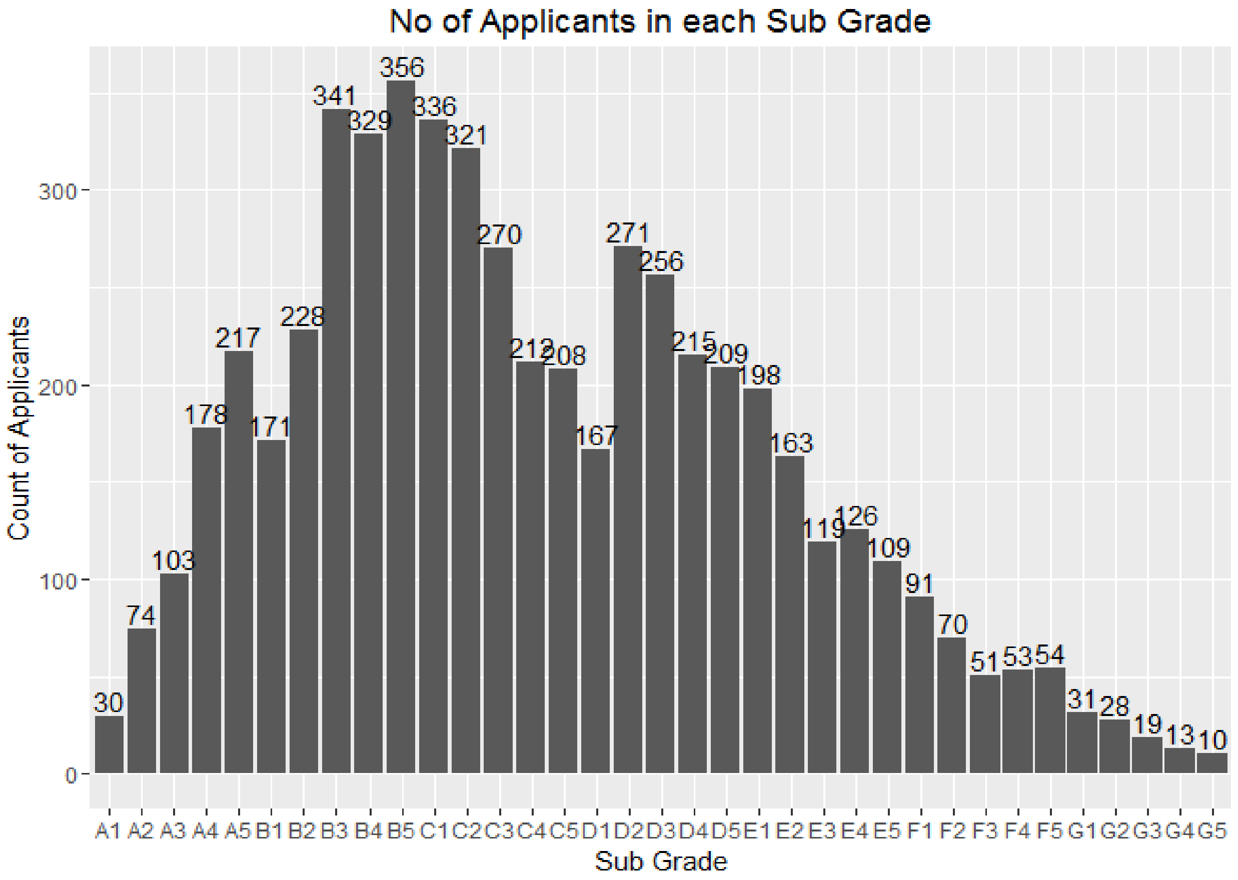


Observations: The number of CHARGED OFF loans for 3 years tenure is higher than 5 year tenure loans

Univariate analysis Contd. 9/30

9. Histogram of Grade and Sub-Grade in Charged Off Loans

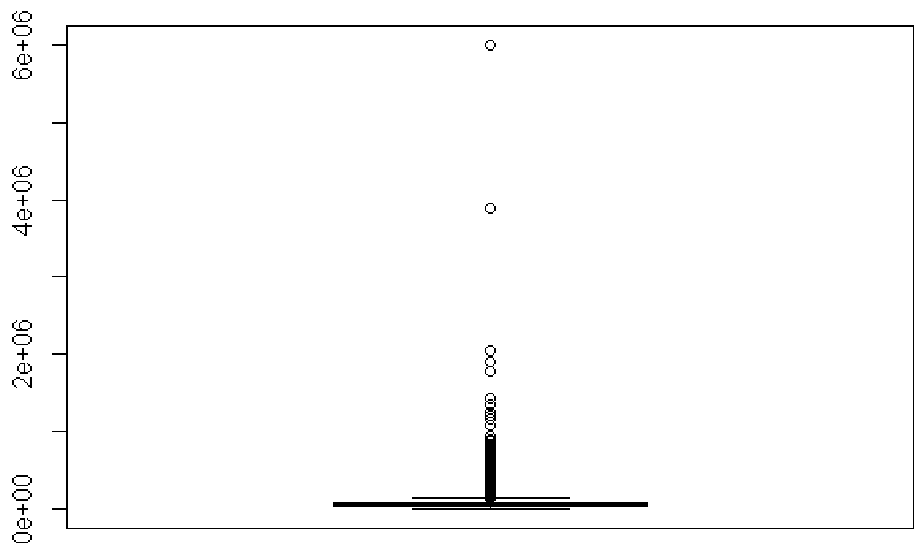




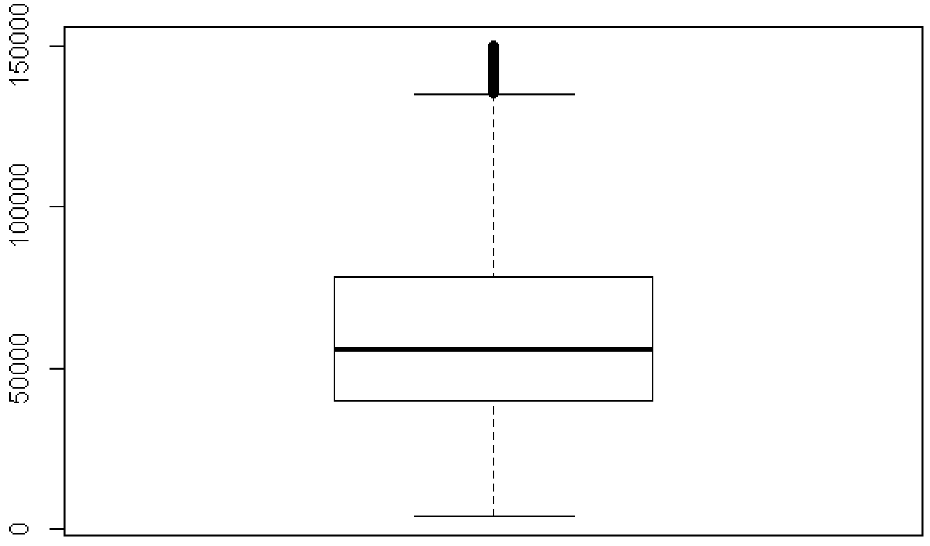
Observations: There is an evident downtrend in Grade to Charged Off loan status. An assigned Loan grade depends on Credit Report. Loan Grades A ~G are from least risky to most risky. (https://www.lendingclub.com/public/rates-and-fees.action) . It appears that most CHARGED OFF cases are in B3 ~ C3 and also D2~E1

Univariate analysis Contd. 10/30

10. Analysis of Annual Income




```
> summary(loan$annual_inc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4000   40000   58870   68780   82000 6000000
```

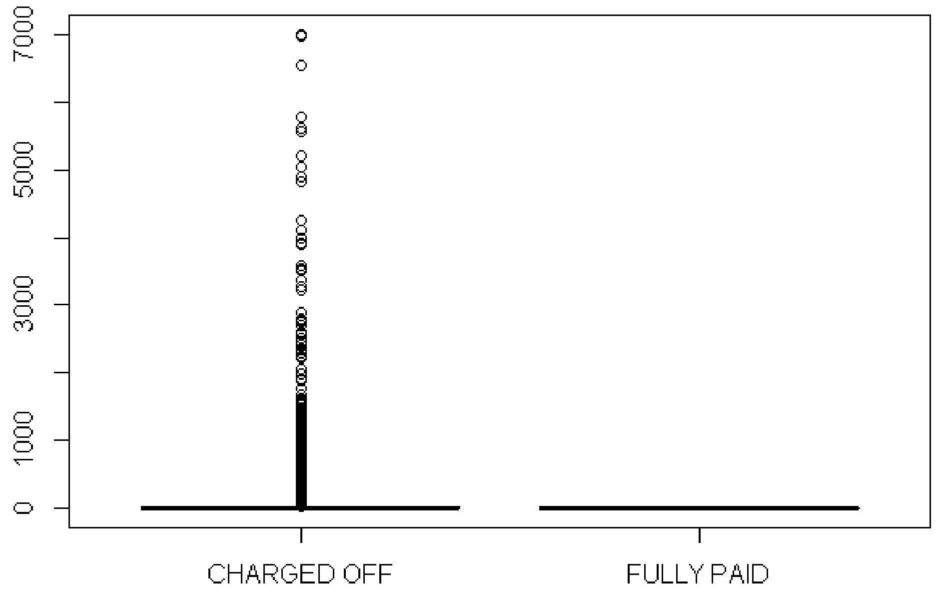



```
> summary(loan$annual_inc[loan$annual_inc<150000])
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4000   40000   56000   61340   78000  150000
```

Observations: Before cleaning, annual income contained so many outliers on the top. For any analysis related to annual income, we must remove these outliers. With an annual income upper limit as 150000, the outliers in the box plot look to be within tolerable limits.

Univariate analysis Contd. 11/30

11. Analysis of Collection Recovery Fee

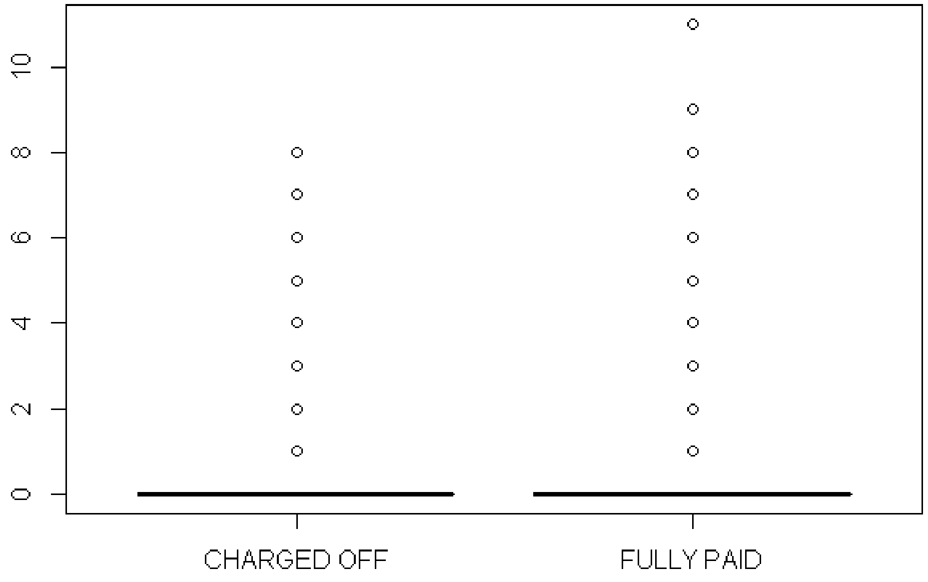



```
> nrow(loan[loan$collection_recovery_fee != 0, ])
[1] 3782
> nrow(loan)
[1] 38577
> nrow(chargedoffloan[chargedoffloan$collection_recovery_fee != 0, ])
[1] 3782
> nrow(chargedoffloan)
[1] 5627
> nrow(fullypaidloan[fullypaidloan$collection_recovery_fee != 0, ])
[1] 0
> nrow(fullypaidloan)
[1] 32950
```

Observations: It appears that there is not a single instance when collection recovery fee was levied for a FULLY PAID loan. Which means this column is not useful for identifying its impact on future loan status of CURRENT loans. Therefore we can remove this column as it doesn't add any value to our analysis.

Univariate analysis Contd. 12/30

12. Analysis of delinquency incidents for last 2 years

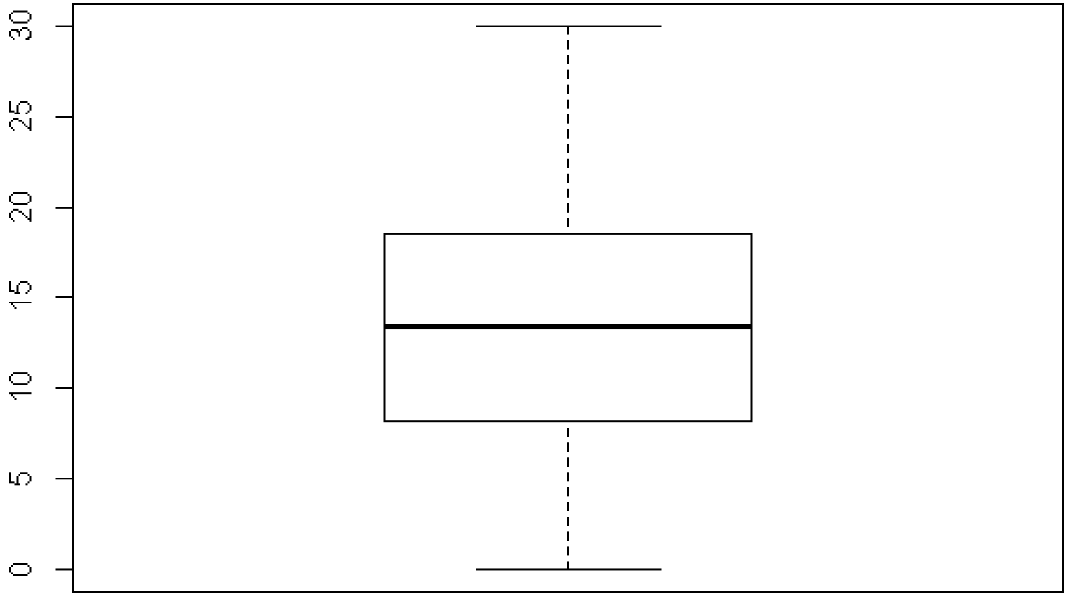



```
> summary(loan$delinq_2yrs)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
 0.0000  0.0000  0.0000  0.1467  0.0000 11.0000 
> nrow(loan[loan$delinq_2yrs != 0, ])
[1] 4191
> nrow(loan)
[1] 38577
> nrow(chargedoffloan[chargedoffloan$delinq_2yrs != 0, ])
[1] 691
> nrow(chargedoffloan)
[1] 5627
> nrow(fullypaidloan[fullypaidloan$delinq_2yrs != 0, ])
[1] 3500
> nrow(fullypaidloan)
[1] 32950
```

Observations: Number of zero values are too high in this column. Also they are present for both CHARGED OFF as well as FULLY PAID cases.

Univariate analysis Contd. 13/30

13. Analysis of DTI column



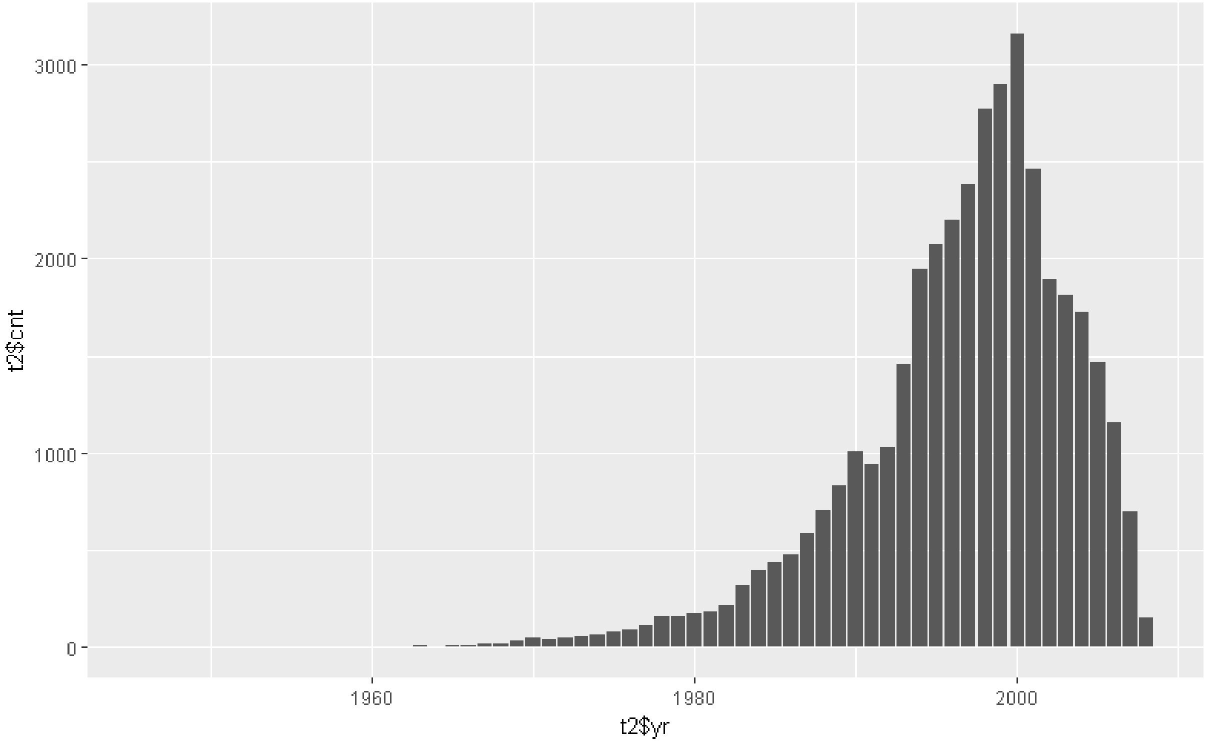

```
> summary(loan$dti)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    8.13   13.37   13.27   18.56   29.99
```

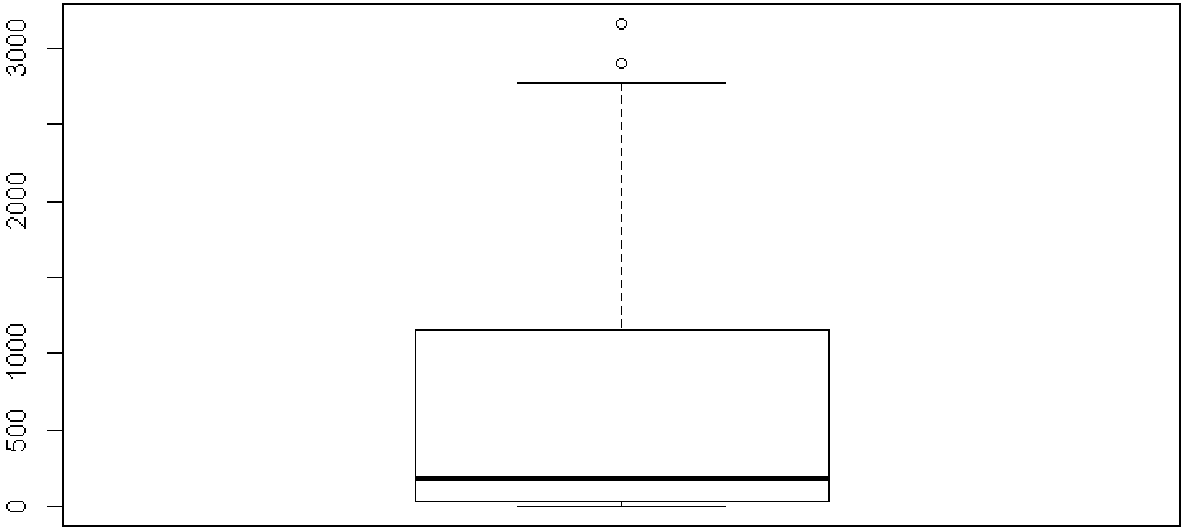
Observations: The DTI column seems to be well distributed and will be very helpful when we are doing the bivariate analysis

Univariate analysis Contd. 14/30

14. Analysis of Earliest Credit Line

```
> summary(t2)
       yr              cnt        
 Min.   :1946   Min.   :   1.0  
 1st Qu.:1969   1st Qu.:  37.0  
 Median :1982   Median : 181.0  
 Mean   :1982   Mean   : 727.9  
 3rd Qu.:1995   3rd Qu.:1154.0  
 Max.   :2008   Max.   :3160.0  
```

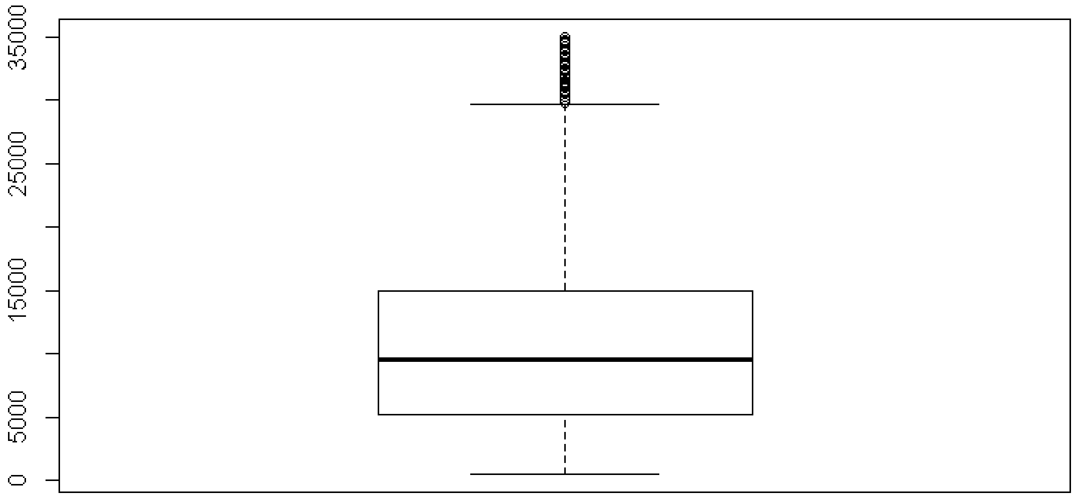


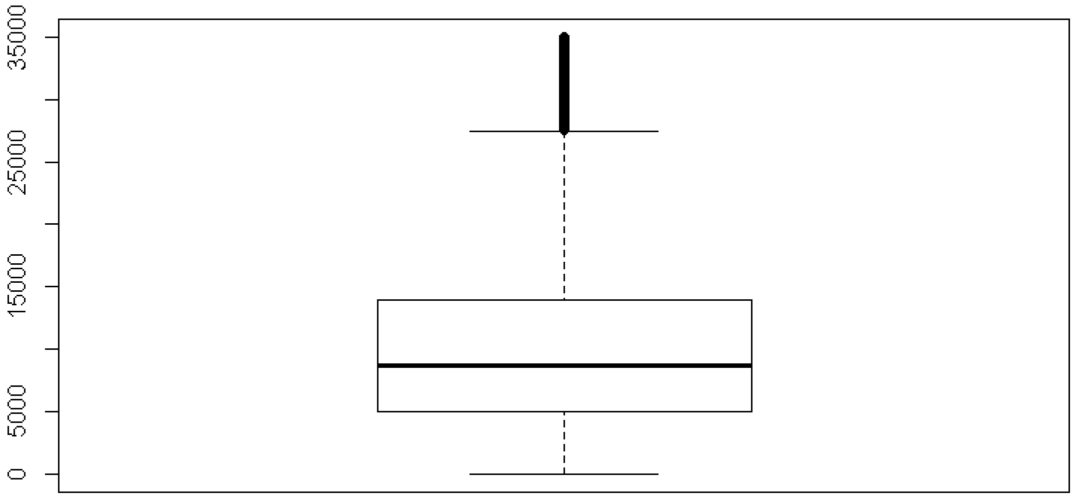


Observations: The number of Early Credit Lines seem to be increasing until 2000 where it peaked. After that it appears that fewer Credit Lines were opened.

Univariate analysis Contd. 15/30

15. Analysis of Funded amount and Funded amount (investor)





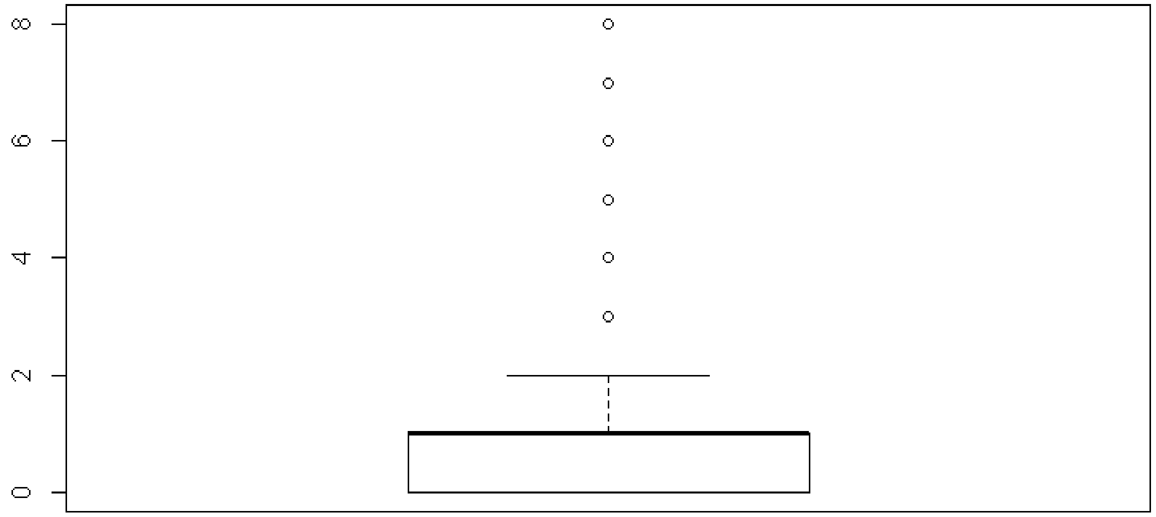

```
> summary(loan$funded_amnt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    500    5200    9550   10780   15000   35000 
```

```
> summary(loan$funded_amnt_inv)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0    5000    8733   10220   14000   35000 
```

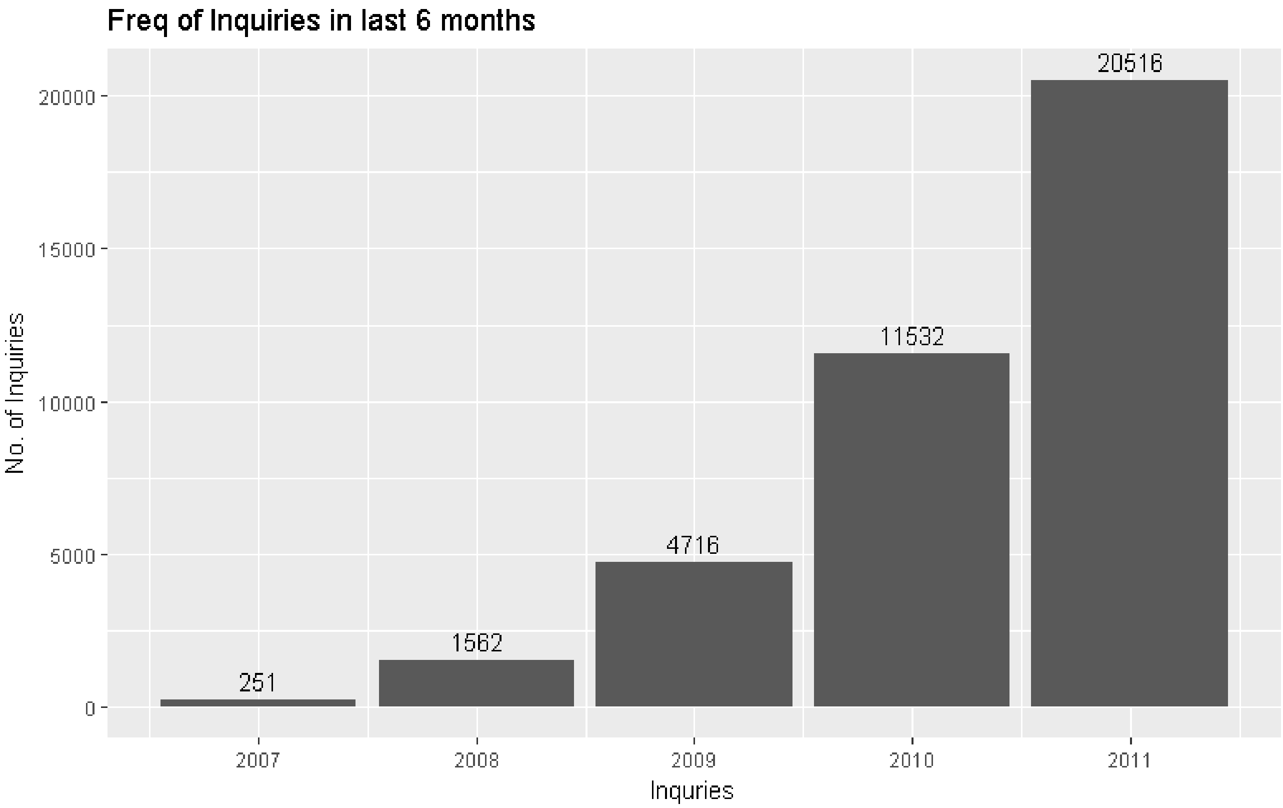
Observations: There are few outliers on top end of the distribution which may need to be excluded during analysis

Univariate analysis Contd. 16/30

16. Analysis of No of inquiries in last 6 months



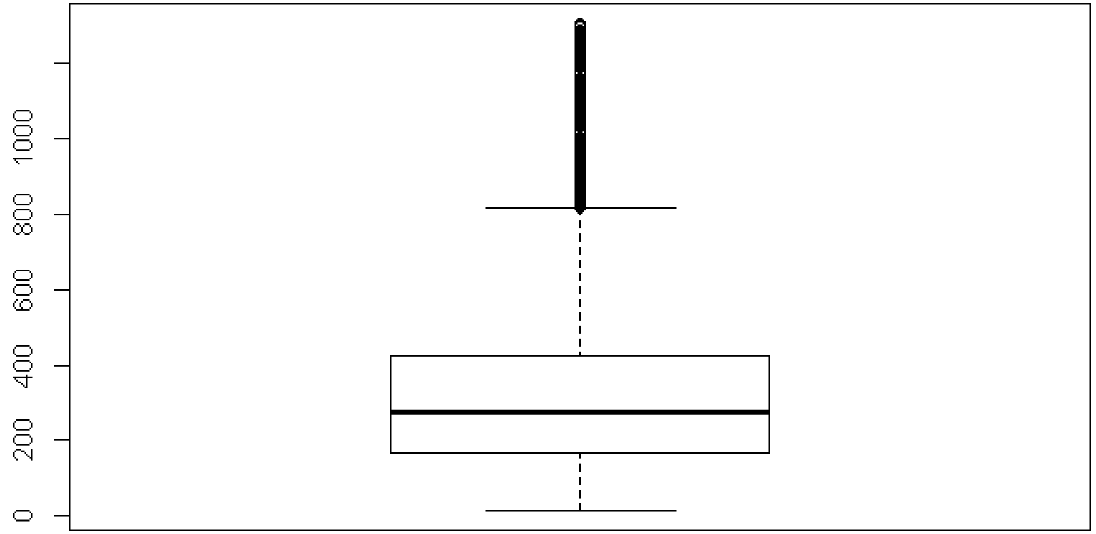

```
> summary(loan$inq_last_6mths)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  1.0000  0.8717  1.0000  8.0000
```



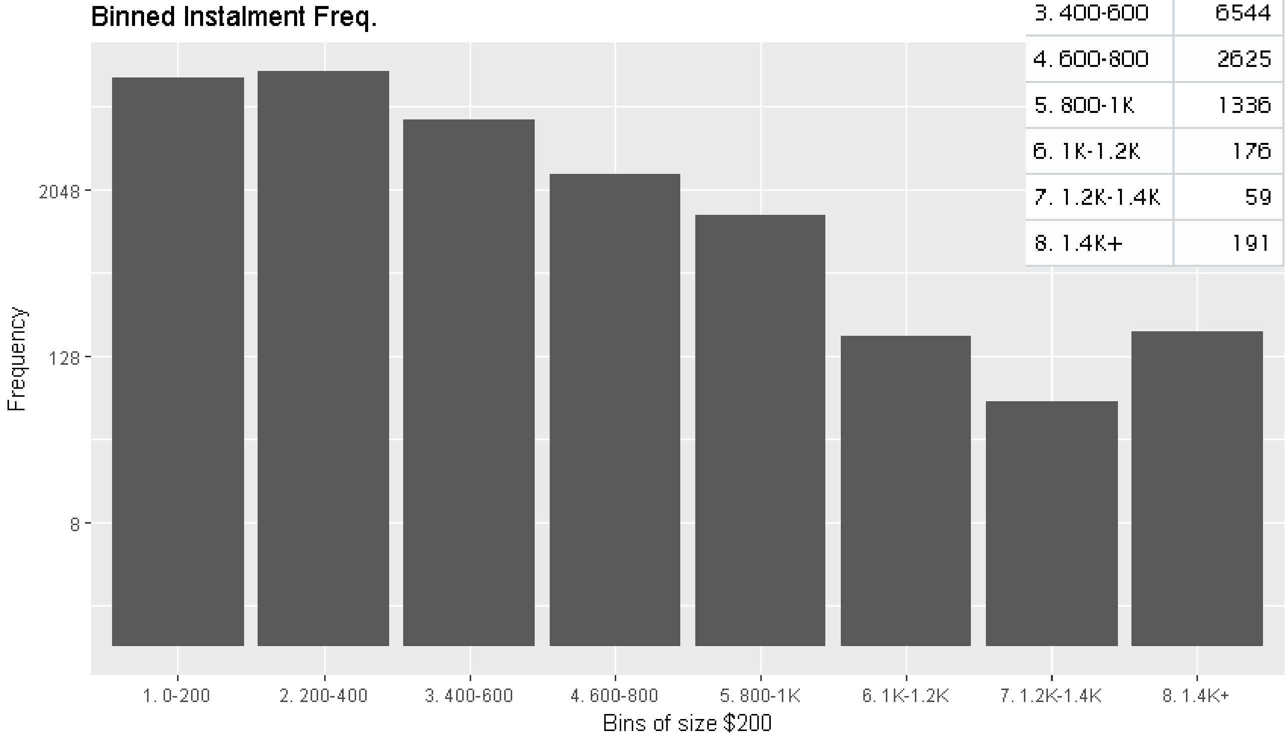
Observations: From the plots, it appears that there are too many 0 values (no inquiries in last 6 months)

Univariate analysis Contd. 17/30

17. Analysis of Instalment




```
> summary(loan$installment)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  15.69  165.70  277.90  322.50  425.60 1305.00 
```



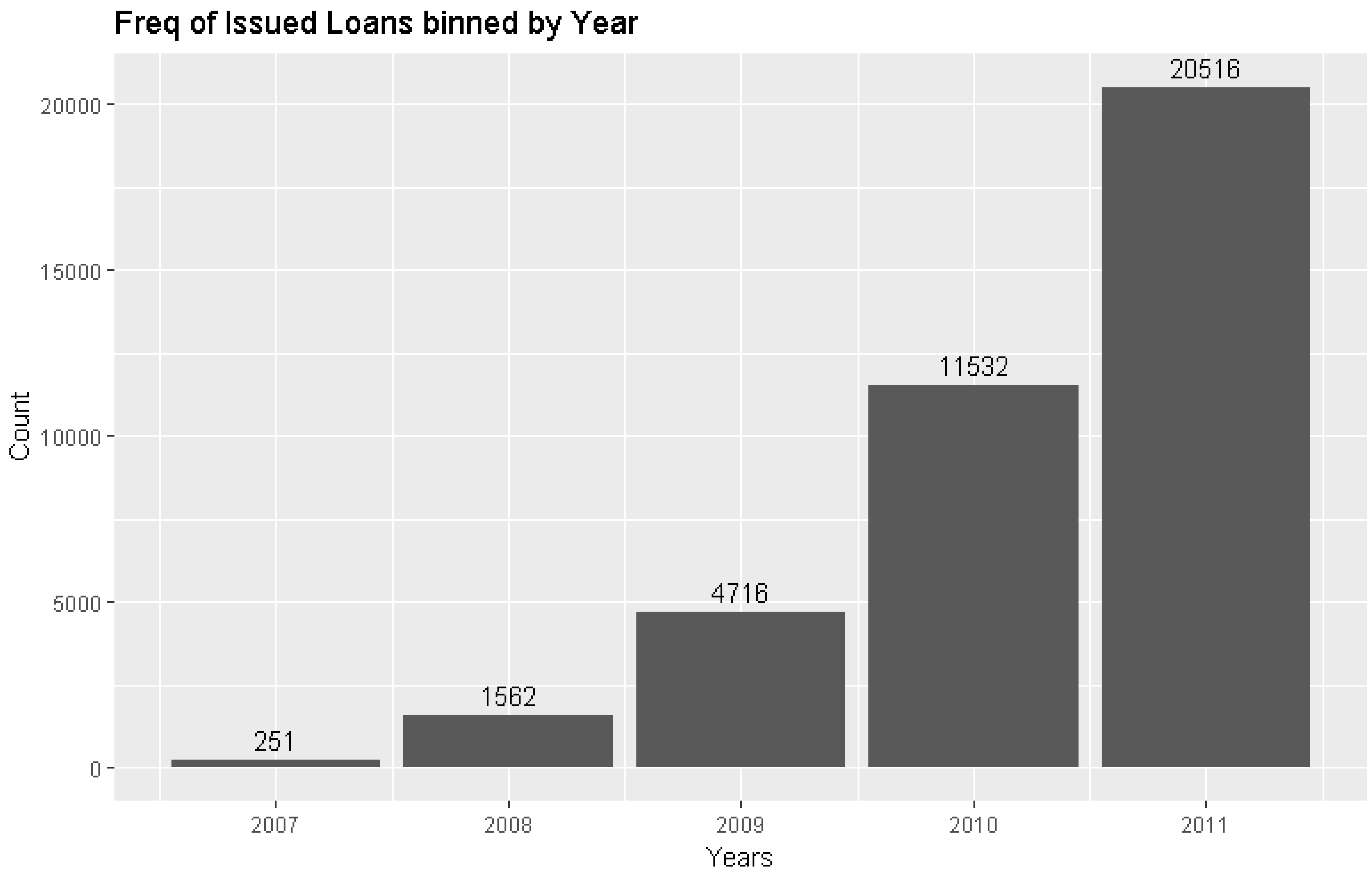
bins	total
1. 0-200	12974
2. 200-400	14672
3. 400-600	6544
4. 600-800	2625
5. 800-1K	1336
6. 1K-1.2K	176
7. 1.2K-1.4K	59
8. 1.4K+	191

Observations: From the box plot, it appears that there are a considerable number of outliers that we may need to remove. The distribution seems to be well skewed

Univariate analysis Contd. 18/30

18. Analysis of Issue Date

```
> summary(t2)
       yr            cnt        
 Min.   :2007   Min.   :  251  
 1st Qu.:2008   1st Qu.: 1562  
 Median :2009   Median : 4716  
 Mean   :2009   Mean   : 7715  
 3rd Qu.:2010   3rd Qu.:11532  
 Max.   :2011   Max.   :20516  
```

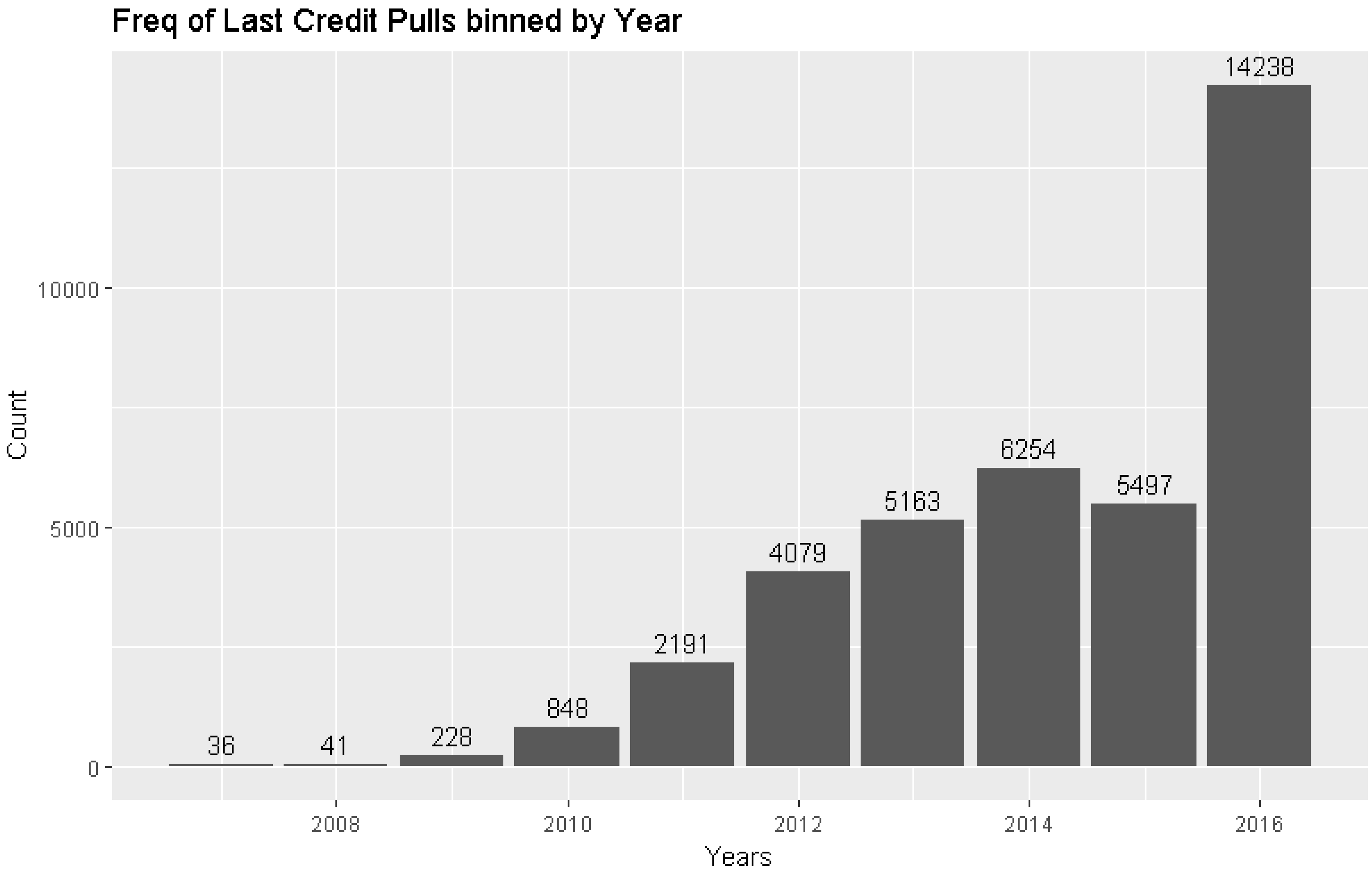


Observations: There is an increasing trend in the no. of issued loans every year

Univariate analysis Contd. 19/30

19. Analysis of Last Credit Pull Date

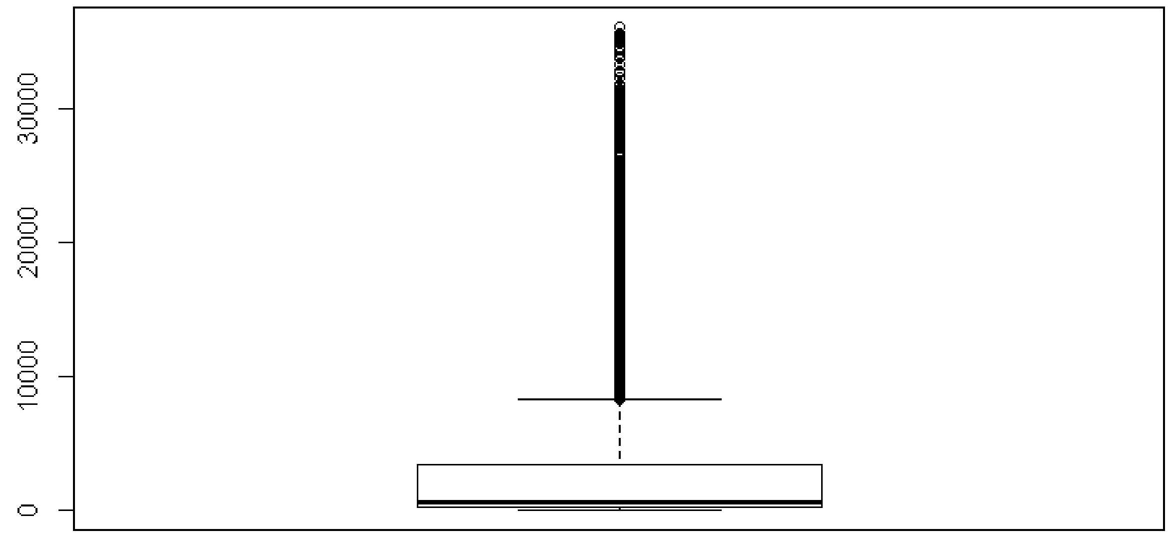
```
> summary(t2)
       yr             cnt         
 Min.   :2007   Min.   :    2.0  
 1st Qu.:2009   1st Qu.:  134.5  
 Median :2012   Median : 2191.0  
 Mean   :2012   Mean   : 3507.0  
 3rd Qu.:2014   3rd Qu.: 5330.0  
 Max.   :2016   Max.   :14238.0  
 NA's   :1
```



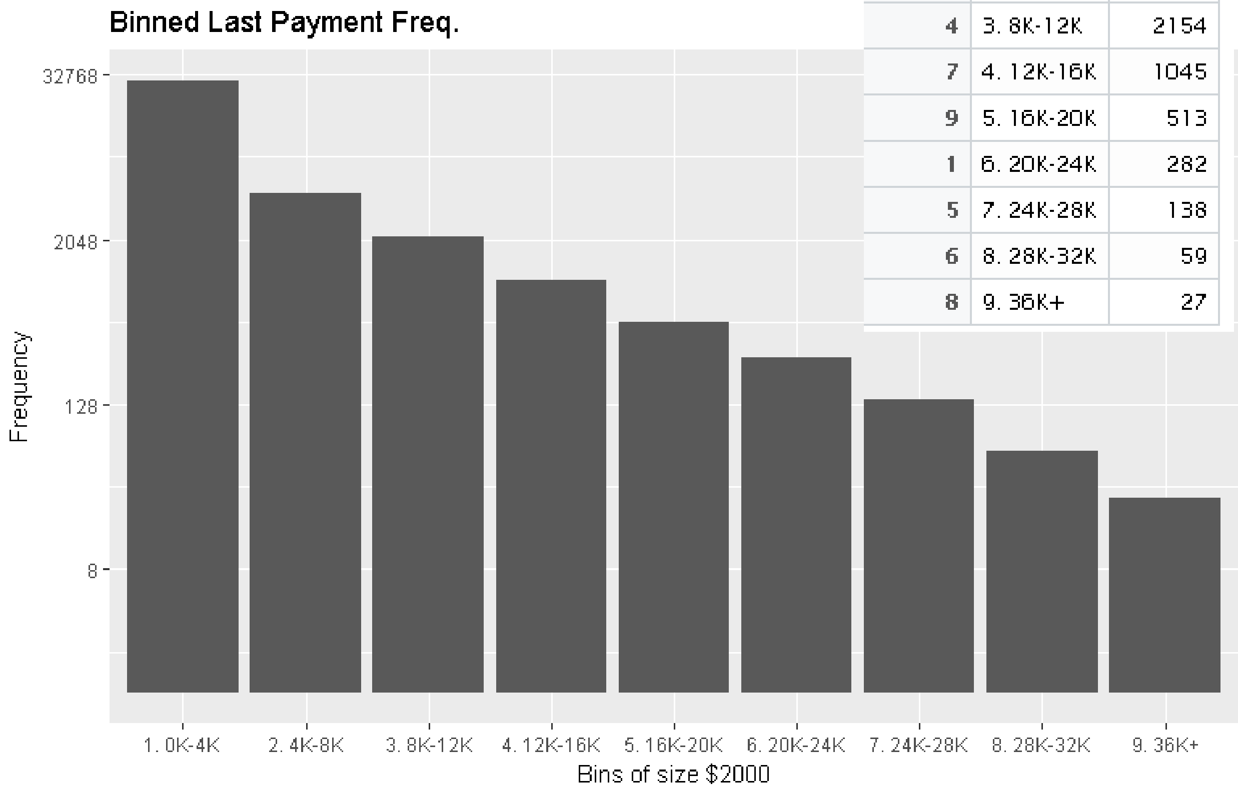
Observations: There is an increasing trend in the no. of credit pulls by creditors every year

Univariate analysis Contd. 20/30

20. Analysis of Last Payment Amount




```
> summary(loan$last_pymnt_amnt)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.0   217.4   568.3  2746.0  3447.0 36120.0
```



	bins	total
3	1. 0K-4K	29915
2	2. 4K-8K	4444
4	3. 8K-12K	2154
7	4. 12K-16K	1045
9	5. 16K-20K	513
1	6. 20K-24K	282
5	7. 24K-28K	138
6	8. 28K-32K	59
8	9. 36K+	27

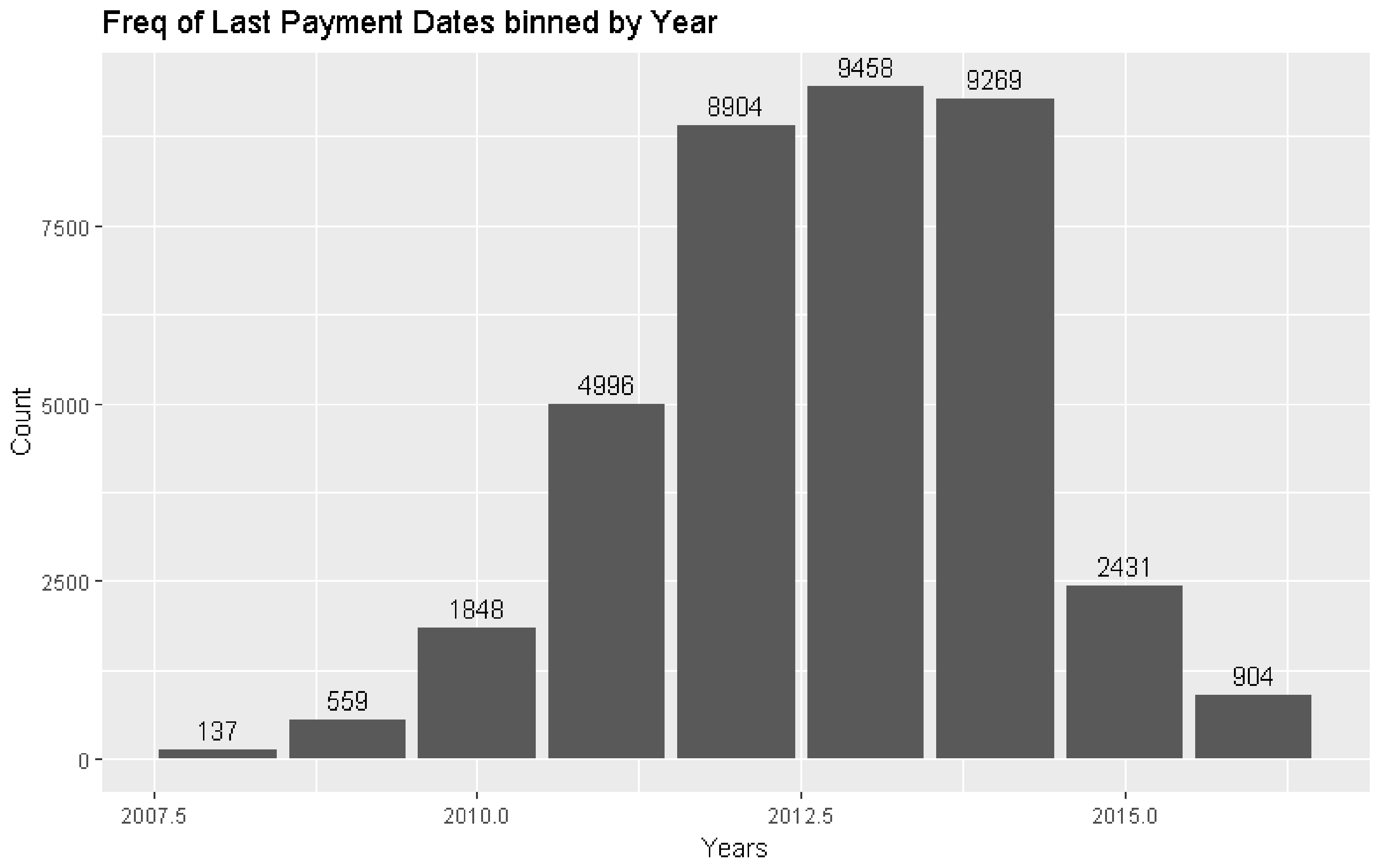
Observations: From the box plot & the distribution, there are so many outliers at the upper end of the distribution. We could consider removing these at the time of analysis.

Univariate analysis Contd. 21/30

21. Analysis of Last Payment Date

```
> summary(t2)
       yr             cnt      
 Min.   :2008   Min.   : 137  
 1st Qu.:2010   1st Qu.: 904  
 Median :2012   Median :2431  
 Mean   :2012   Mean   :4278  
 3rd Qu.:2014   3rd Qu.:8904  
 Max.   :2016   Max.   :9458  
```

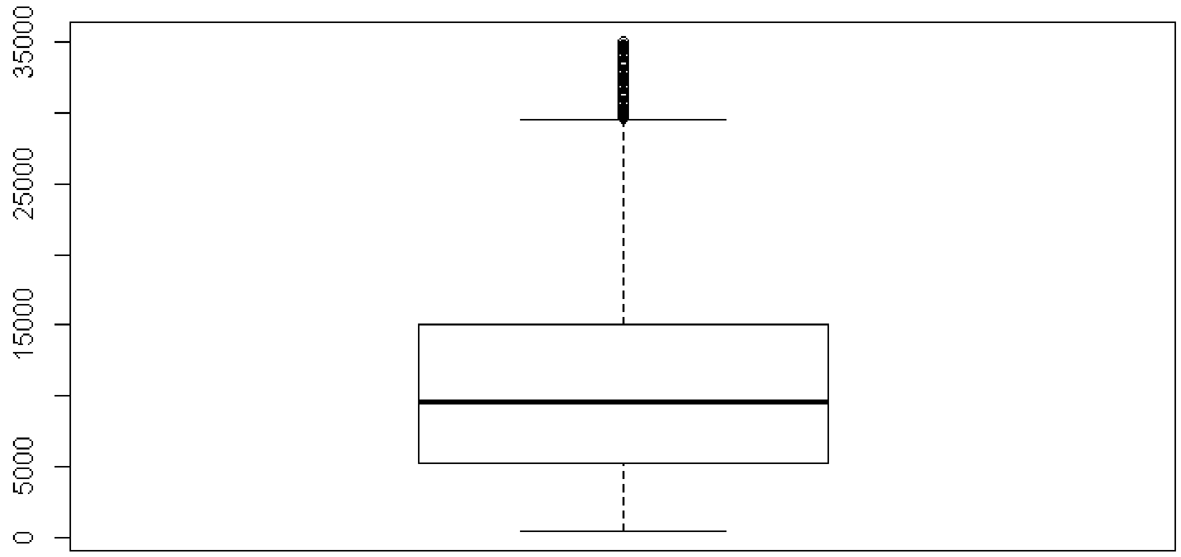
yr	cnt
2008	137
2009	559
2010	1848
2011	4996
2012	8904
2013	9458
2014	9269
2015	2431
2016	904



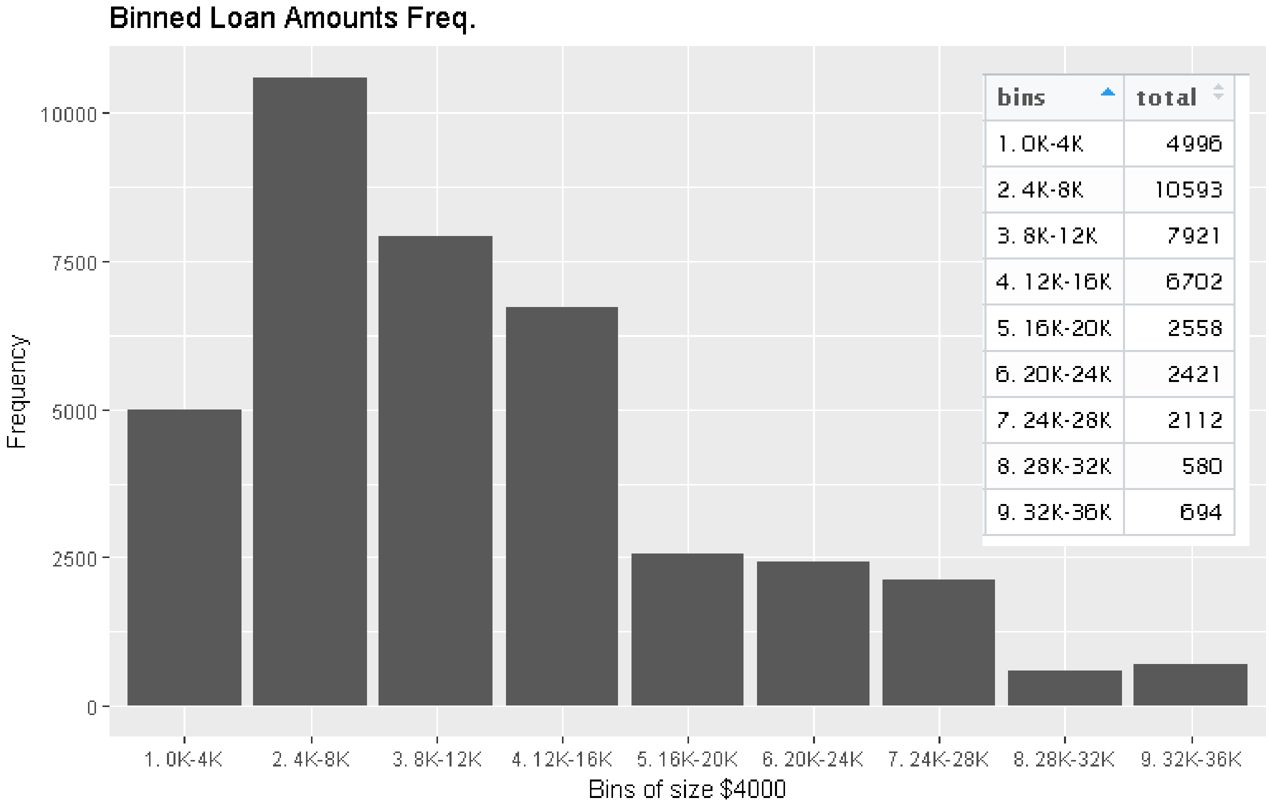
Observations: Last Payment Dates were at its peak in 2014, after that they seem to dwindle

Univariate analysis Contd. 22/30

22. Analysis of Loan Amount



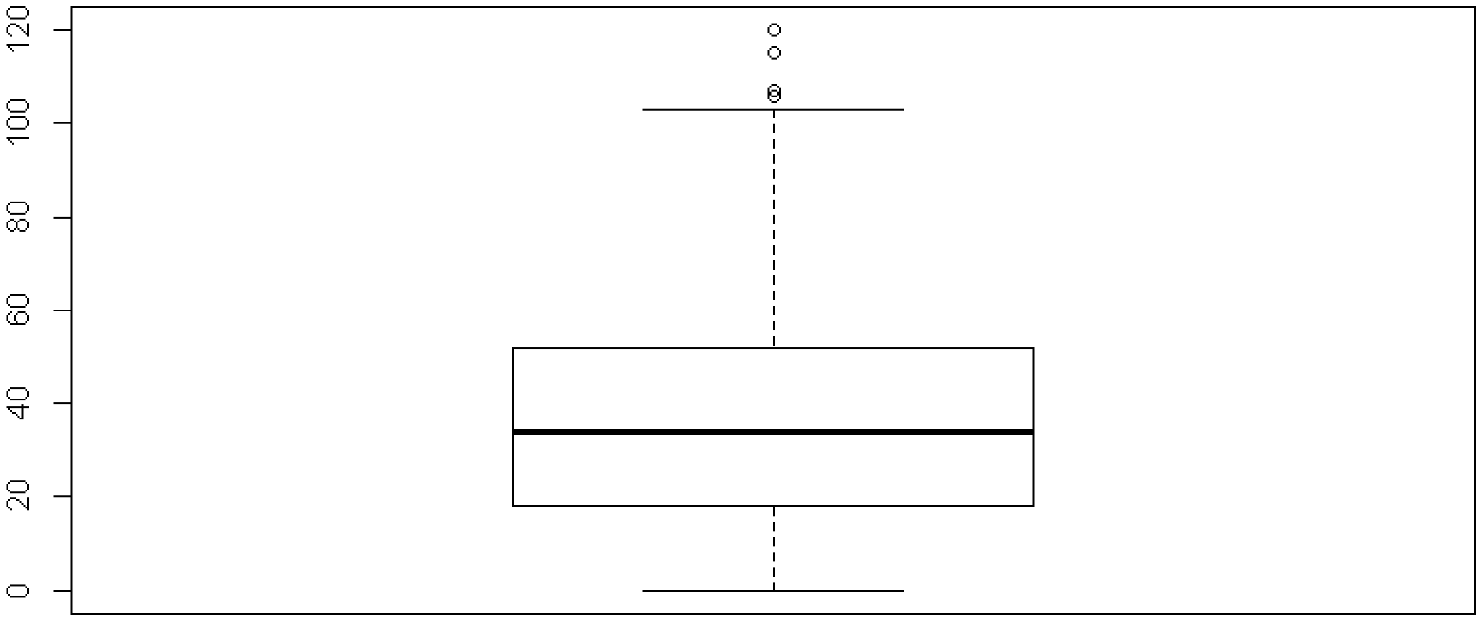

```
> summary(loan$loan_amnt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    500    5300    9600   11050   15000   35000
```



Observations: From the box plot & the distribution, we can clearly see that it is a beautiful Gaussian but there are outliers at the far end. We could remove these using the 95% variance rule.

Univariate analysis Contd. 23/30

23. Analysis of Months since last delinquency

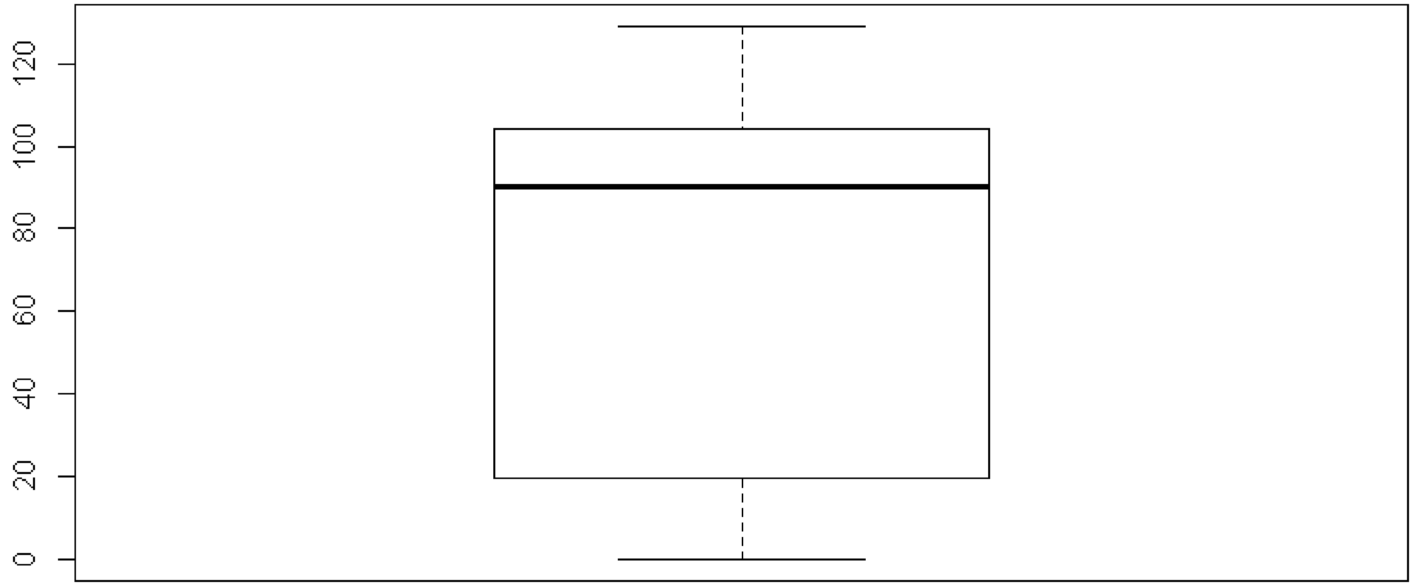



```
> summary(loan$mths_since_last_delinq)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   18.00   34.00   35.88   52.00  120.00   24905 
```

Observations: From the box plot we can see there are some outliers at the top. Also from the descriptive statistics, we notice there are too many NA values. These must be removed at the time of analysis

Univariate analysis Contd. 24/30

24. Analysis of Months since last record

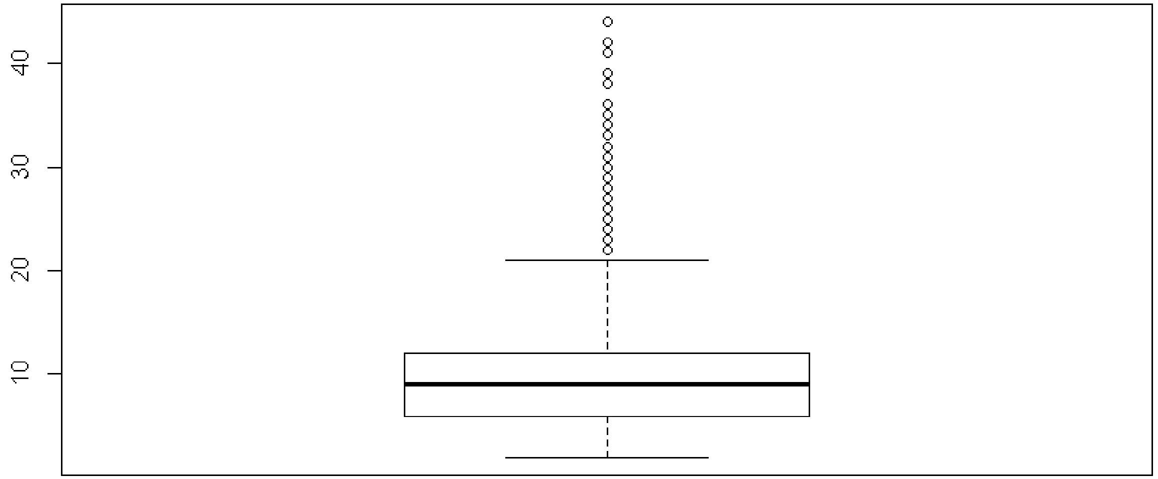



```
> summary(loan$mths_since_last_record)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
   0.00   19.75   90.00   69.26  104.00  129.00   35837 
```

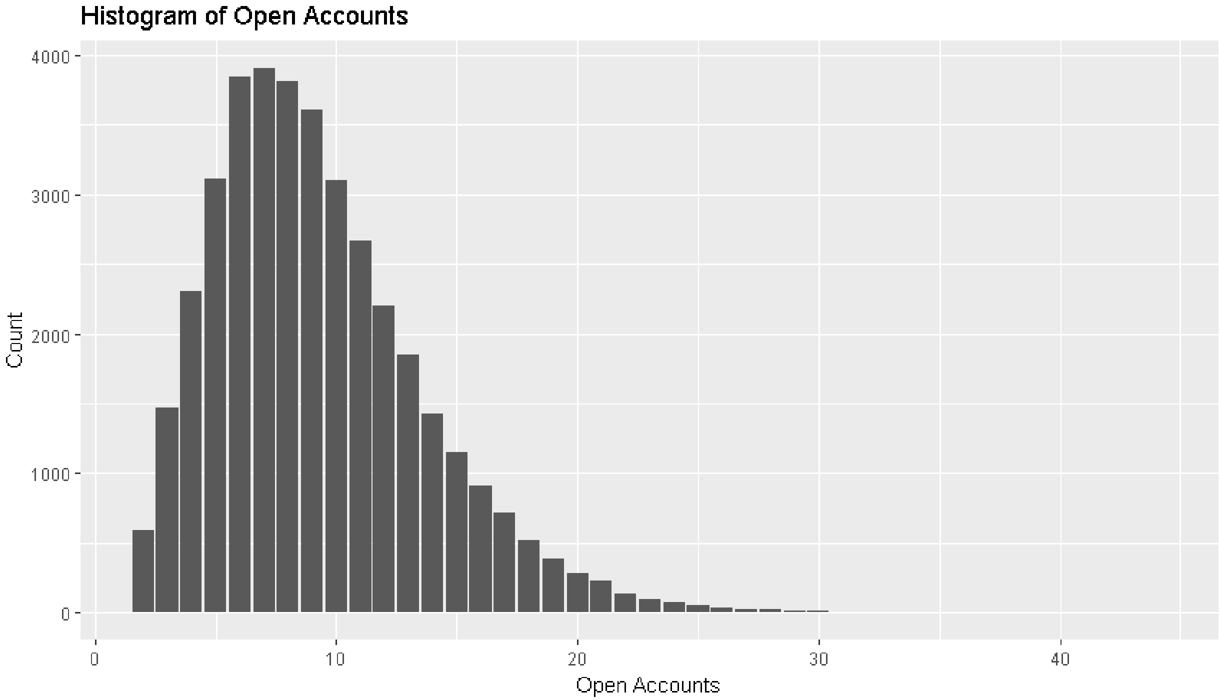
Observations: From the descriptive statistics, we notice there are too many NA values. These must be removed at the time of analysis

Univariate analysis Contd. 25/30

25. Analysis of Open Account



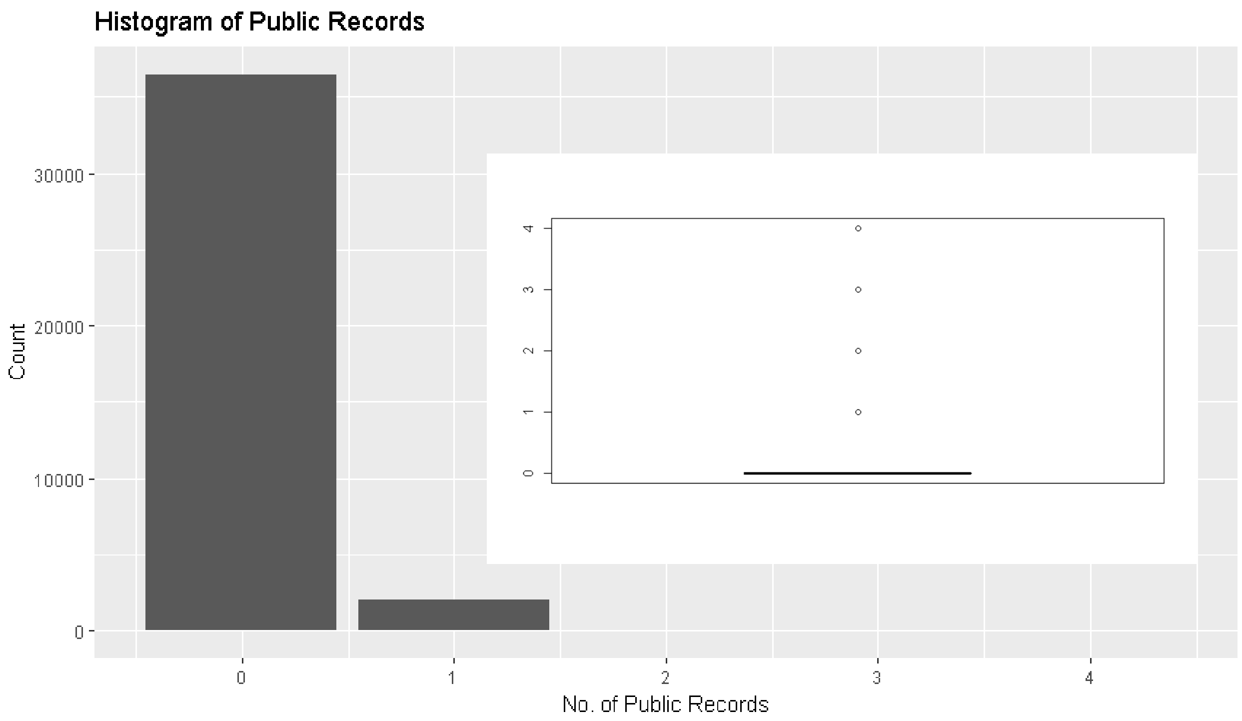

```
> summary(loan$open_acc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   6.000   9.000   9.275  12.000  44.000
```

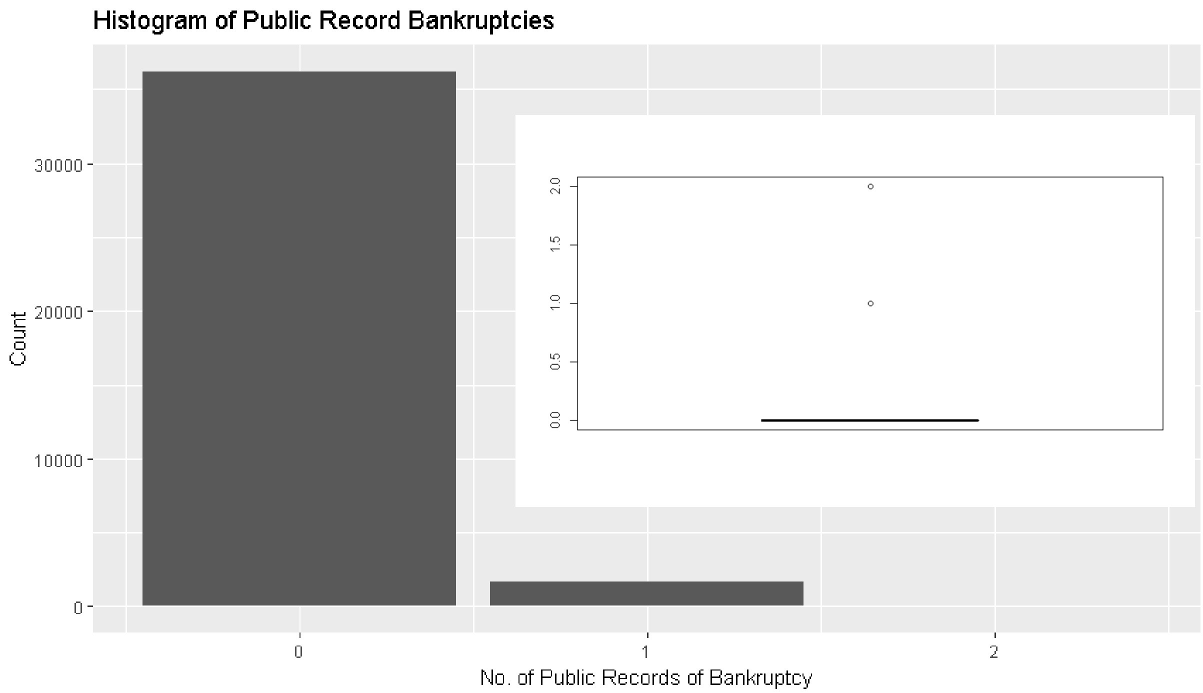


Observations: From the box plot we can see there are some outliers at the top. These must be removed at the time of analysis

Univariate analysis Contd. 26/30

26. Analysis of Derogatory Public Records & Public Record Bankruptcies






```
> summary(loan$pub_rec)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
0.00000 0.00000 0.00000 0.05542 0.00000 4.00000 
```

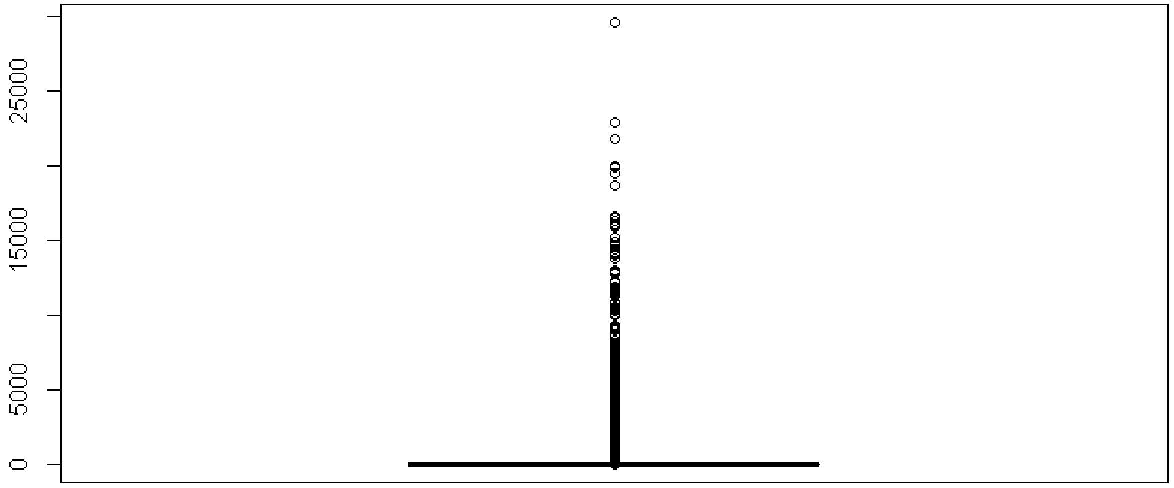
```
> summary(loan$pub_rec_bankruptcies)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 0.0000  0.0000  0.0000  0.0435  0.0000  2.0000     697 
```

Observations: There are hardly any derogatory public records. Fortunately we don’t have any NA in that column.

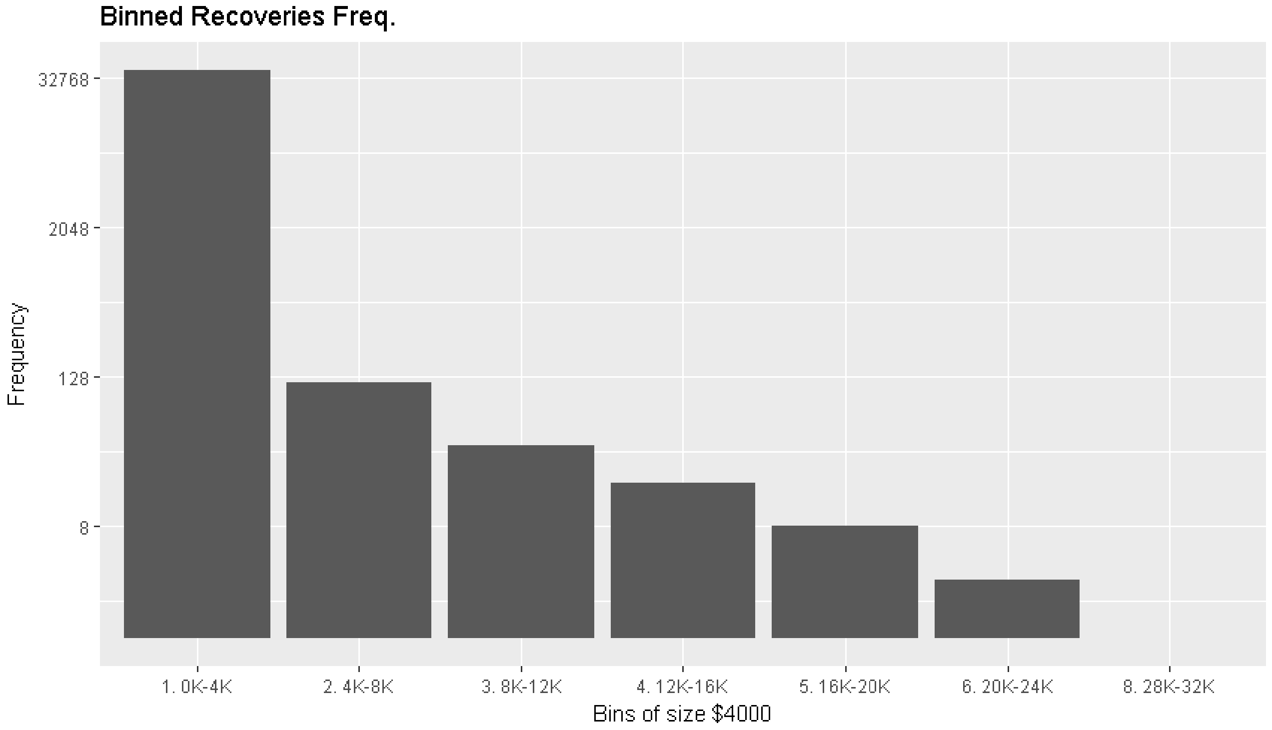
Public Record of Bankruptcies column is sparsely populated too

Univariate analysis Contd. 27/30

27. Analysis of Recoveries



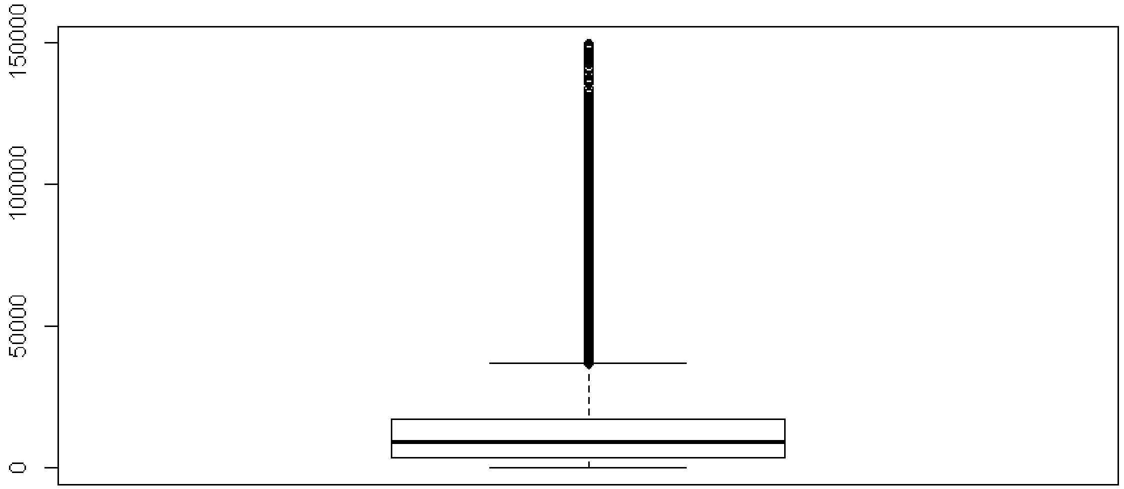

```
> summary(loan$recoveries)
   Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
   0.00     0.00     0.00    98.04     0.00 29620.00
```



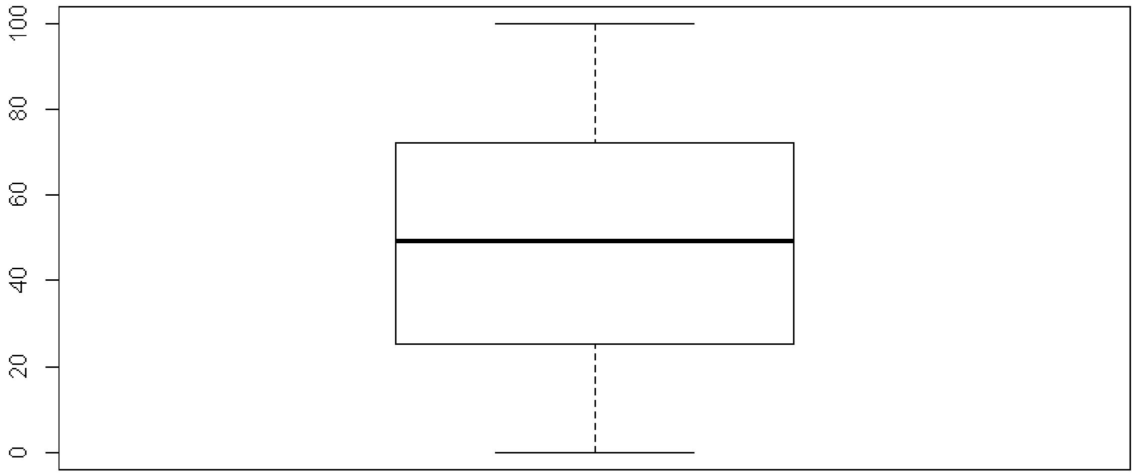
Observations: From the box plot, it appears that there are a considerable number of outliers that we may need to remove. Otherwise the no of recoveries seem to reduce as the amount increases

Univariate analysis Contd. 27/30

27. Analysis of Revolving Balance & Revolving Credit




```
> summary(loan$revol_bal)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    3650    8762   13290   16910  149600
```

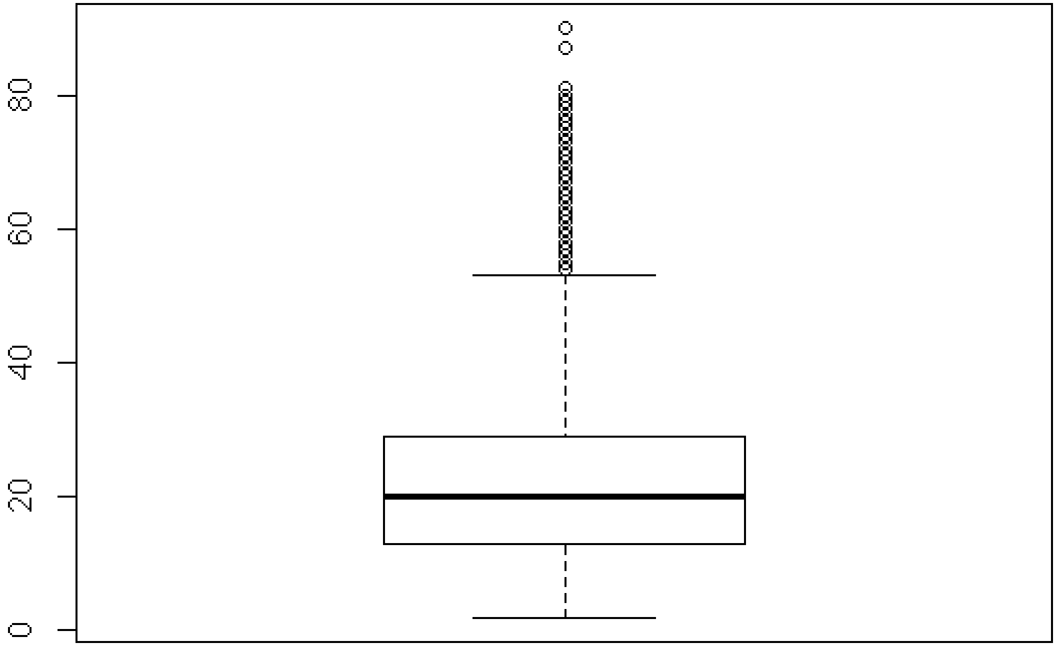



```
> summary(loan$revol_util)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
    0.0    25.2    49.1    48.7    72.3    99.9      50
```

Observations: From the box plot for Revolving Balance, it appears that there are a considerable number of outliers that we may need to remove. On the other hand Revolving Utilization (%age) looks normally distributed.

Univariate analysis Contd. 28/30

28. Analysis of Total number of Credit Lines

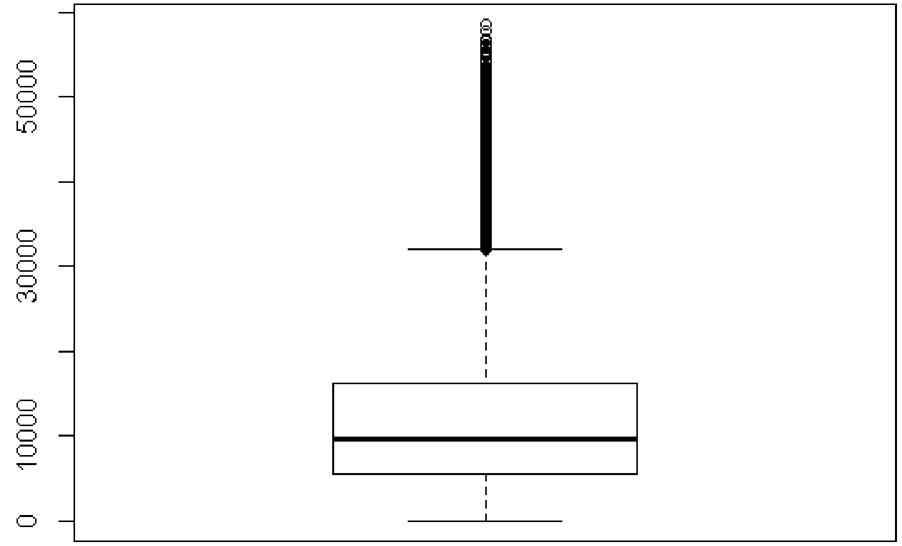



```
> summary(loan$total_acc)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00   13.00   20.00   22.05   29.00   90.00
```

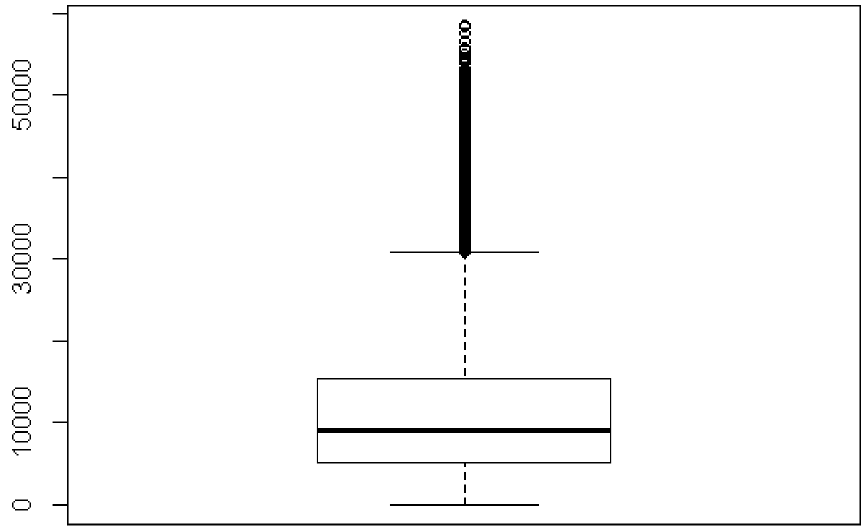
Observations: From the box plot for Total open credit lines, it appears that there are a considerable number of outliers on top that we may need to remove.

Univariate analysis Contd. 29/30

29. Analysis of Total Payment & Total Payment Inv.




```
> summary(loan$total_pymnt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    5513    9674   11870   16140   58560
```

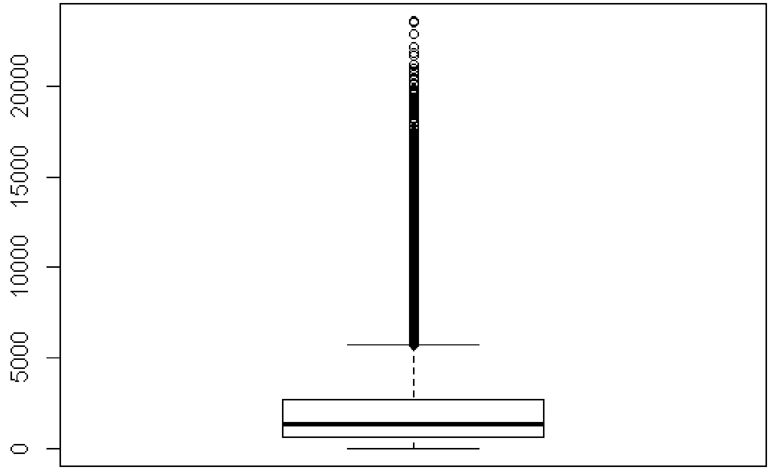



```
> summary(loan$total_pymnt_inv)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
      0    5020    9067   11270   15330   58560
```

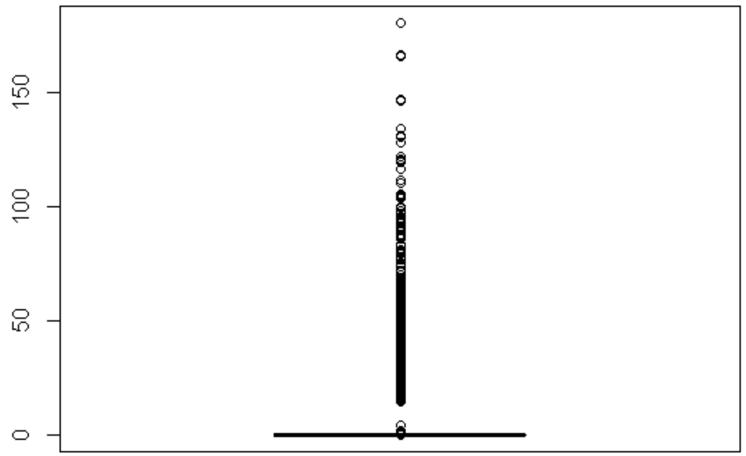
Observations: From the box plot for both columns, it appears that there are a considerable number of outliers on top that we may need to remove. However both of them seem to have a similar distribution and high correlation

Univariate analysis Contd. 30/30

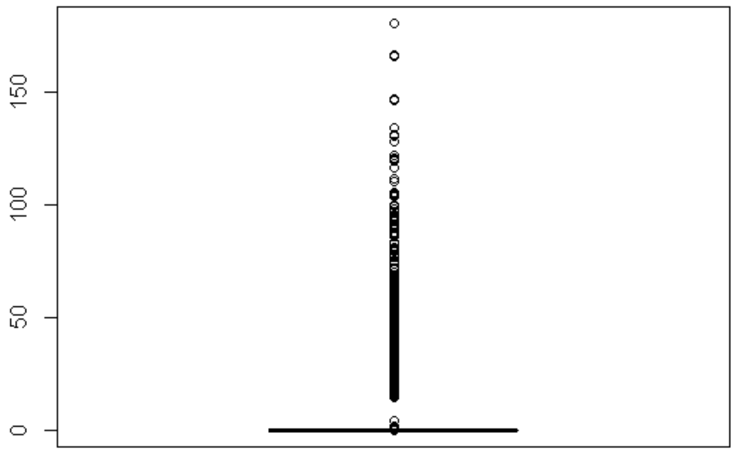
30. Analysis of Total Interest Recd., Total Late Fee Recd., and Total Principal Recd.




```
> summary(loan$total_rec_int)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
      0     644    1300    2119    2672   2356C
> 
```




```
> summary(loan$total_rec_late_fee)
   Min. 1st Qu.  Median    Mean 3rd Qu.    M
  0.000   0.000   0.000   1.369   0.000 180.
> 
```

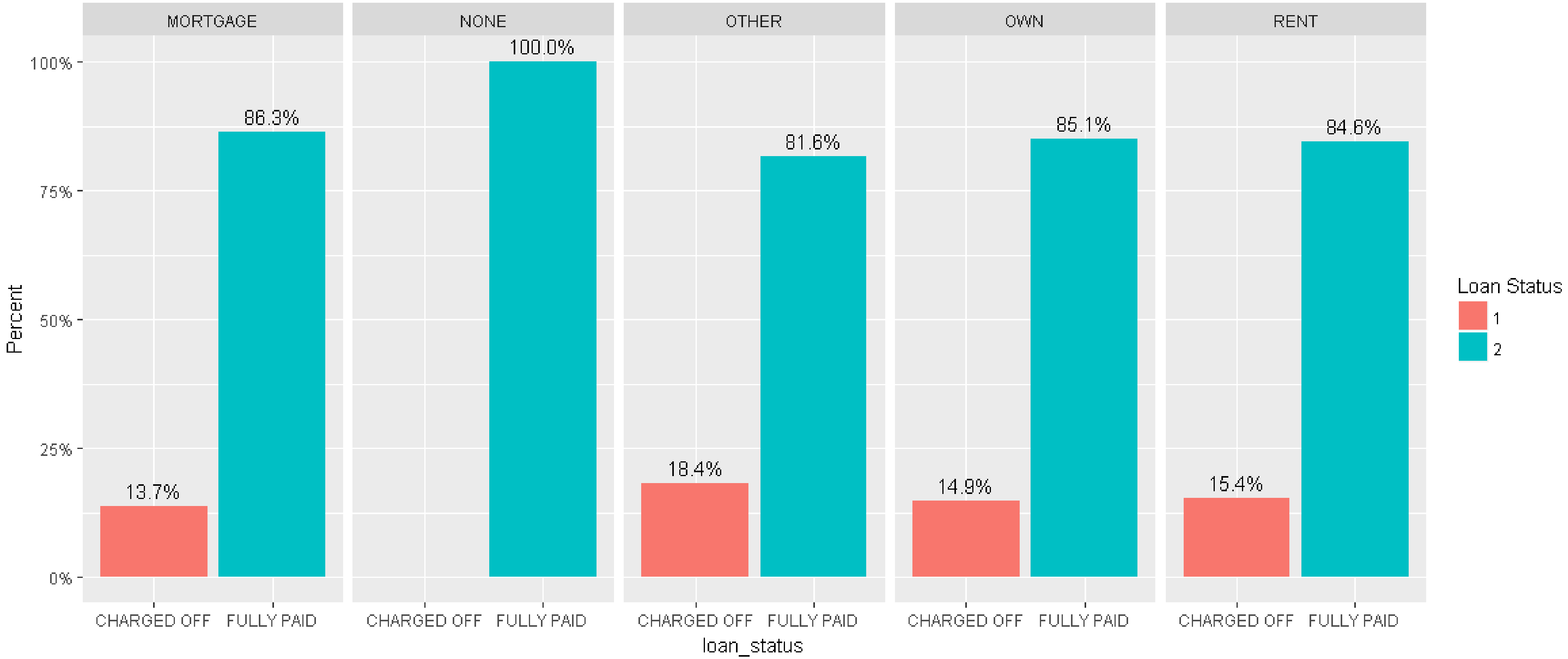



```
> summary(loan$total_rec_late_fee)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.000   0.000   1.369   0.000 180.200
```

Observations: From the box plot for all 3 columns, it appears that there are a considerable number of outliers on top that we may need to remove

Bivariate analysis (1/7)

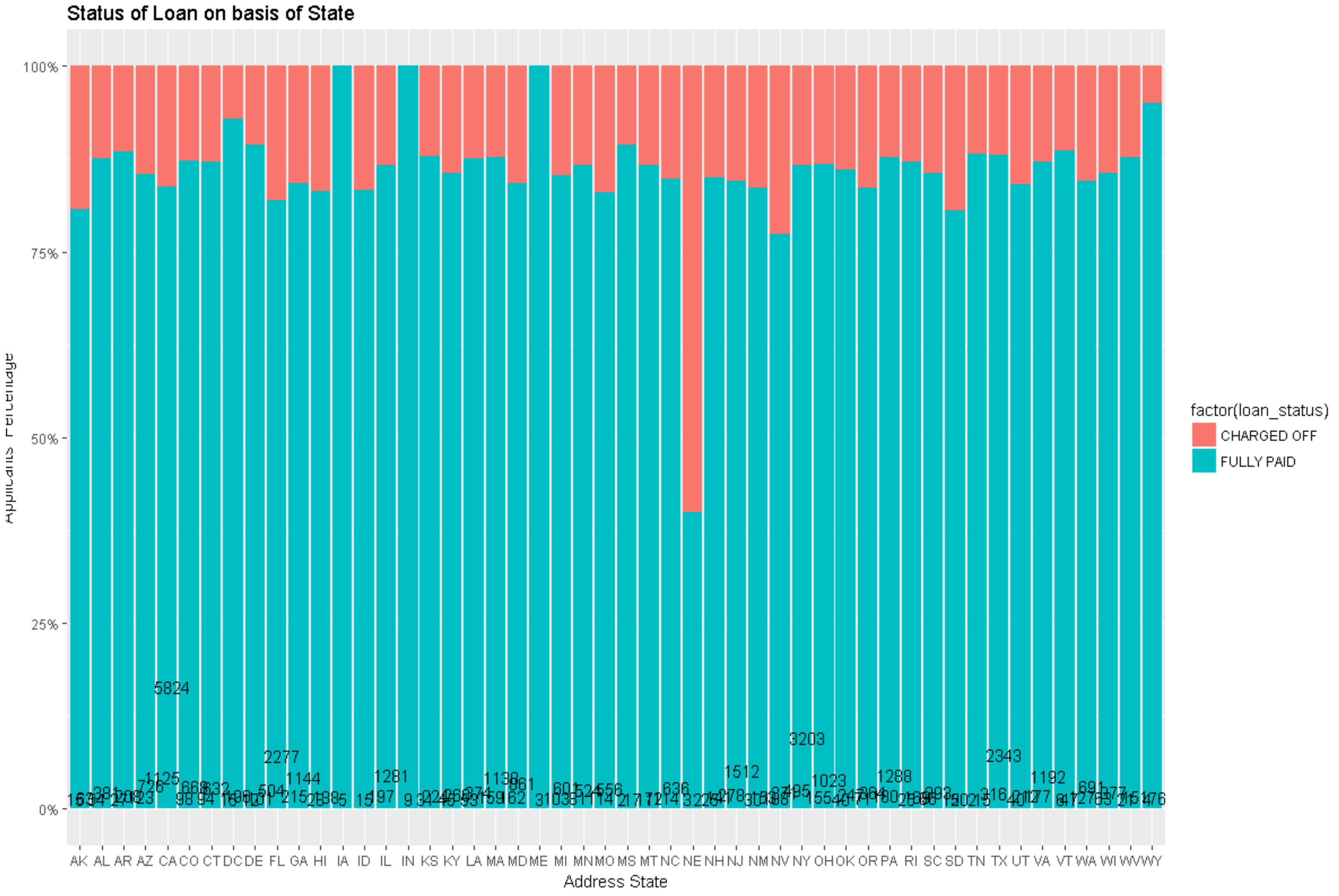
1. Homeowner Attribute vs. Loan Status



Observation: If the borrower's home ownership status is "OTHERS", he is 18% likely to be a defaulter compared to other categories

Bivariate analysis (2/7)

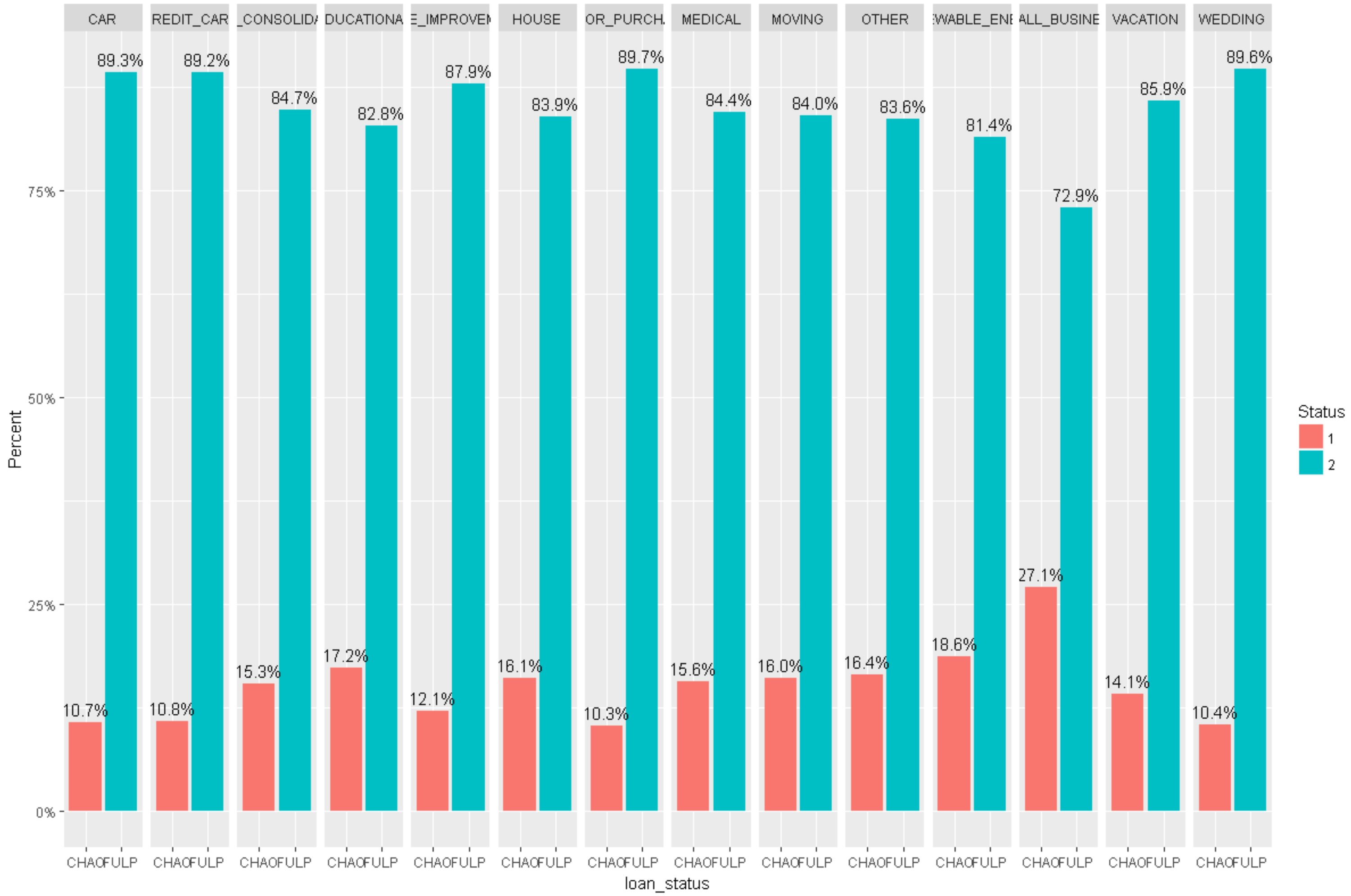
2. State vs. Loan Status



Observation: Nevada is the most risky state to lend a loan compared to other states

Bivariate analysis (3/7)

3. Purpose Attribute vs. Loan Status



Observation: Borrowers for ‘Small Business’ are most likely to default for the loan compared to other purposes

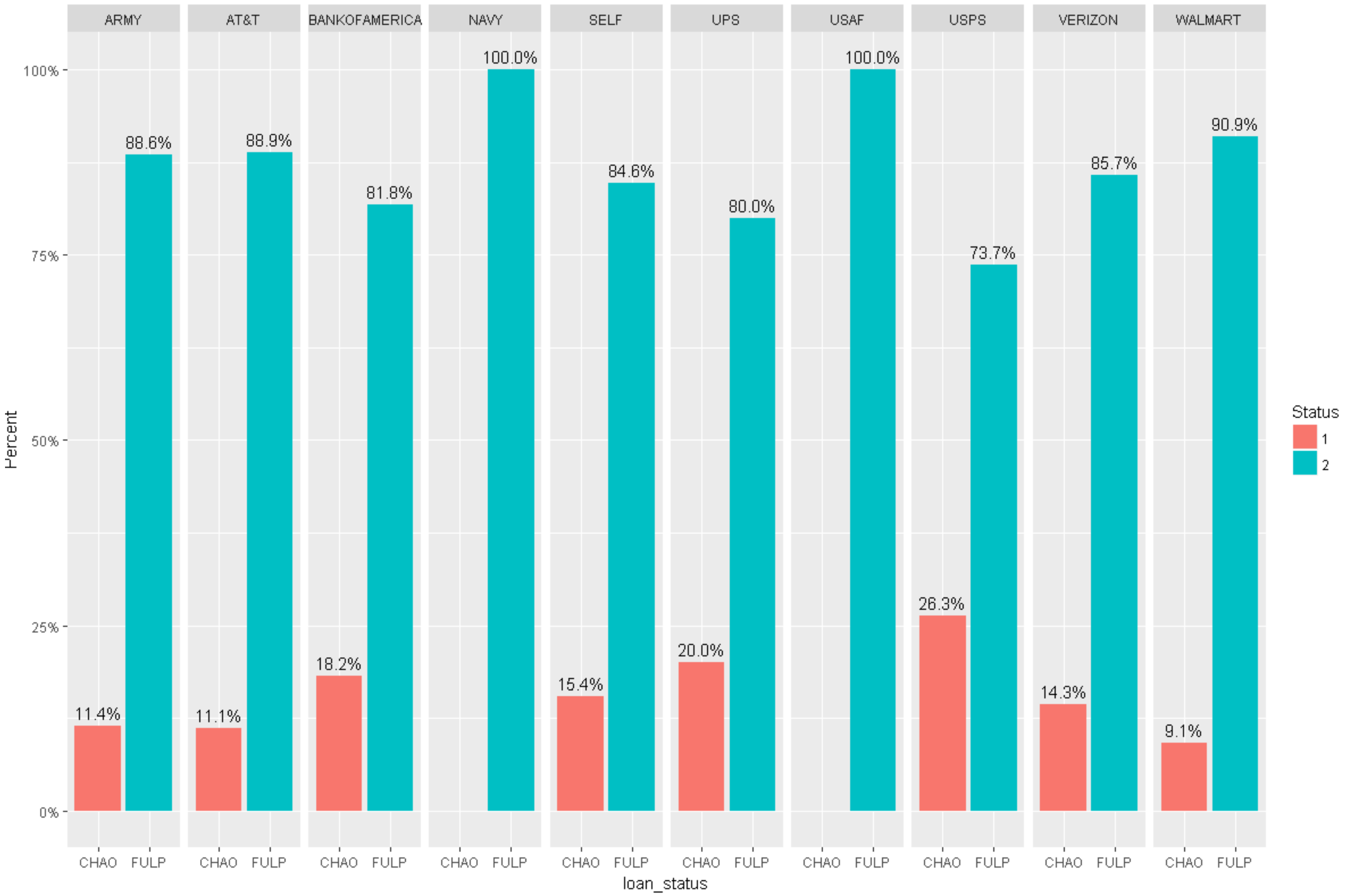
Bivariate analysis (4/7)

4. Employment Duration vs. Loan Status

Observation: Borrowers employed for 10+ years are most likely to default the loan compared to other durations

Bivariate analysis (5/7)

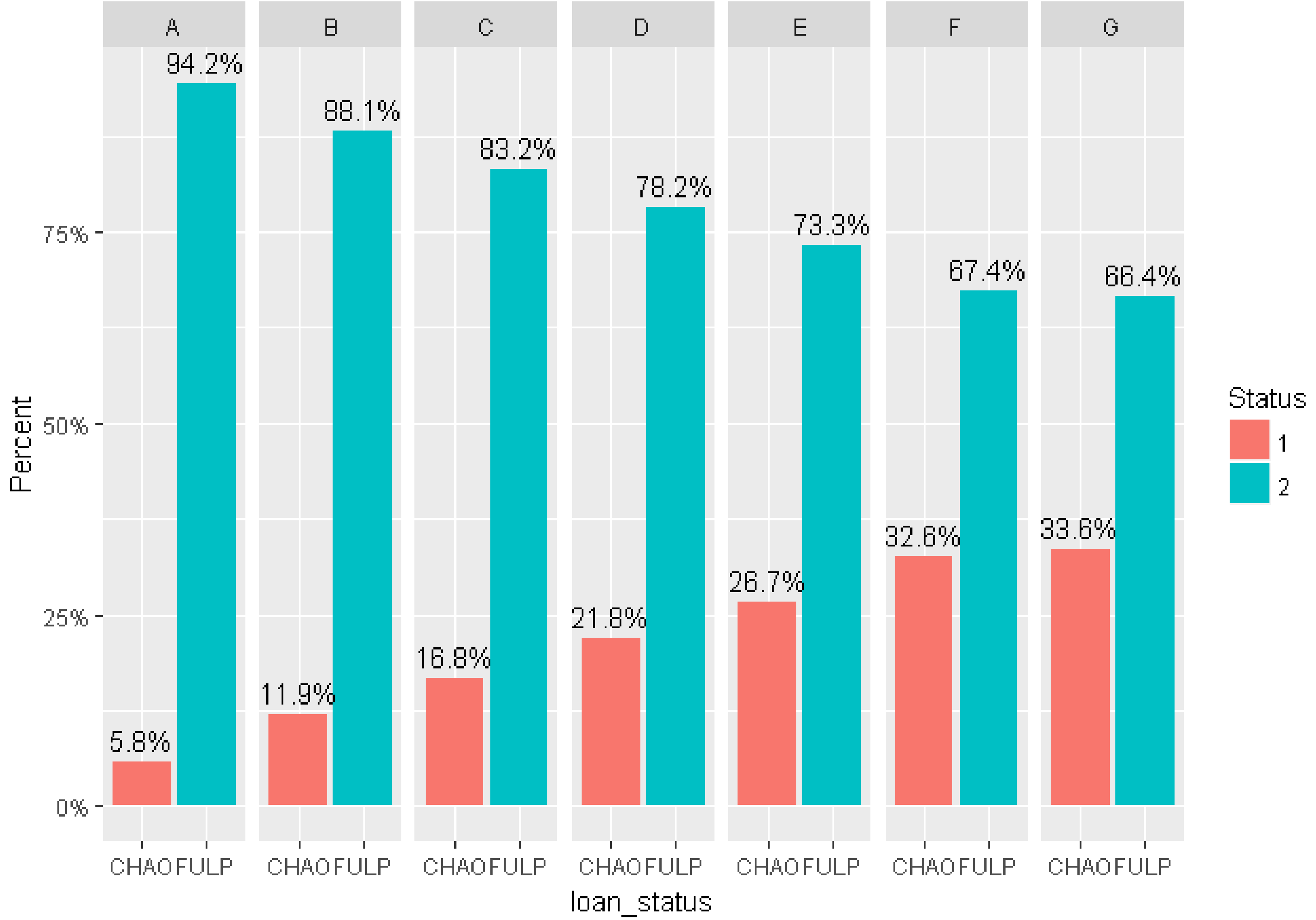
5. Top 10 Employers vs. Loan Status



Observation: Employers working in USPS are more likely to default

Bivariate analysis (6/7)

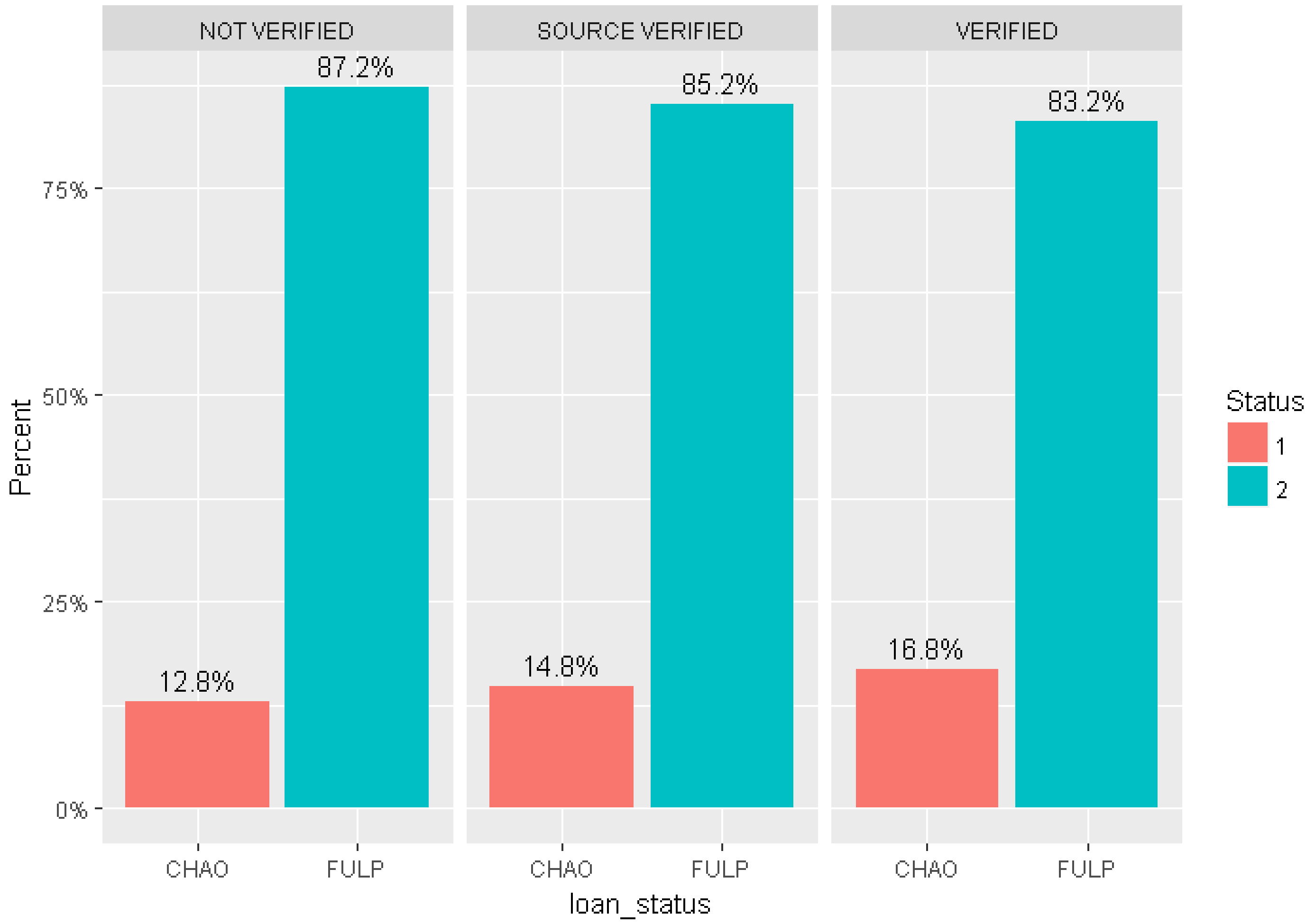
6. Grade vs. Loan Status



Observation: Loans with Grade as G are most likely to be defaulted

Bivariate analysis (7/7)

7. Verification Status vs. Loan Status



Observation: In-spite of verified status, the likelihood of a loan being defaulted is in the “Verified” Category