EDA Case Study

Risk Analytics in Banking & Financial Services

Strategy and Business Objectives

Business Objective

- Identify patterns which shows up if a person is likely to default, this will help in taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate etc.

Strategy

- Analyse the different factors/variables of the dataset to identify some dependency.
- Focus on some variables to validate if they are deciding factors in risk analysis for this bank.

Approaching the problem

The analysis is divided into following parts:

- Cleaning the Loan dataset to minimize the research area and have only meaningful attributes.
- Analysing different attributes to find the one’s which can be a deciding factor.
- Creating derived metrics in the process if needed to identify the problem.
- Finally how summarising the variables affecting the loan grant decision.

Understanding the different variables/attributes of the dataset

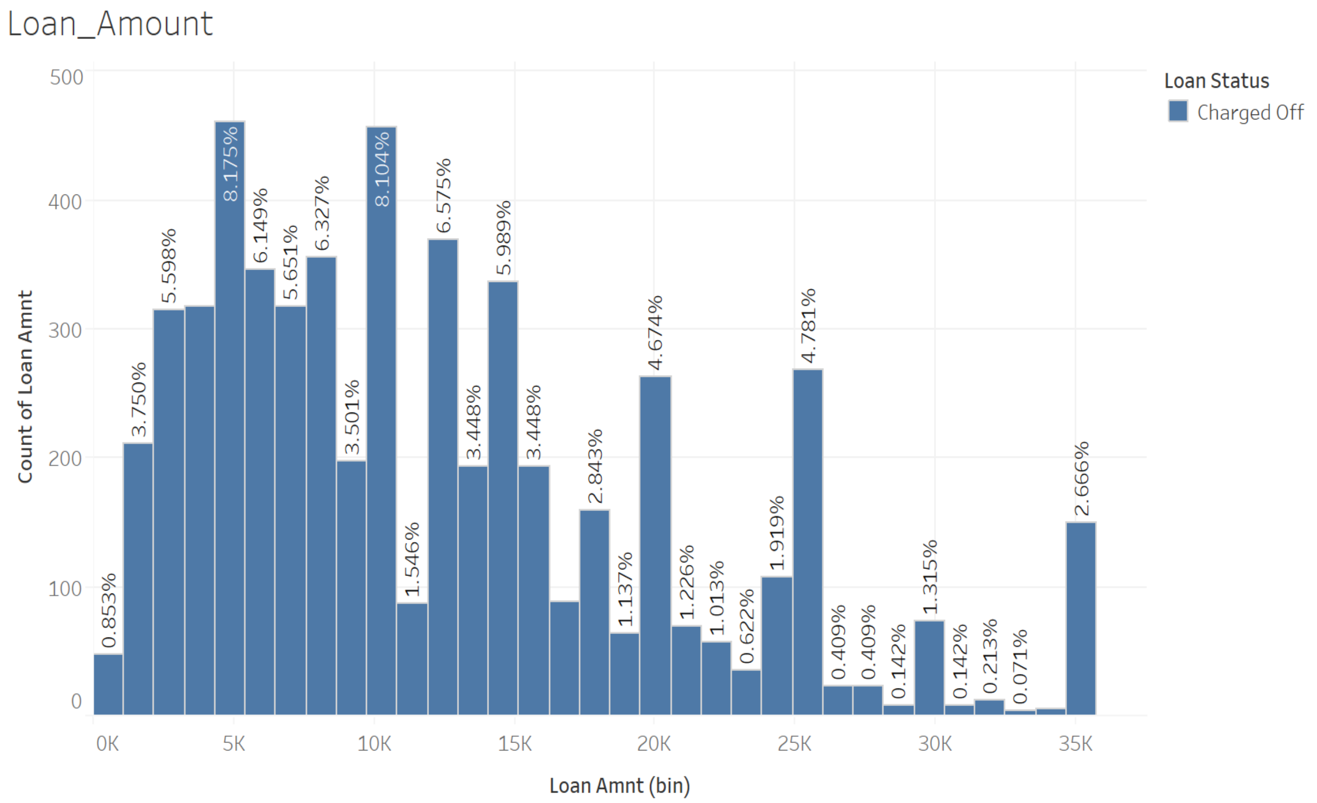
Summary of some variables are as follows:

- **“id”:** Unique Id for the loan listing.
- **“member_id”:** Unique Id assigned for the borrower member.
- **“loan_amnt”:** The amount of loan for which borrower has applied for the loan.
- **“emp_title”:** The job title applied by borrower while applying for loan.
- **“emp_length”:** The no. of years of job experience of borrower when applied for loan.
- **“home_ownership”:** The type of home borrower resides when applied for loan.
- **“revol_bal”:** The amount of credit card spending that remains unpaid at the end of a monthly billing cycle.
- **Similarly there are a lot of attributes of which some are useful and some needs to be cleaned.**

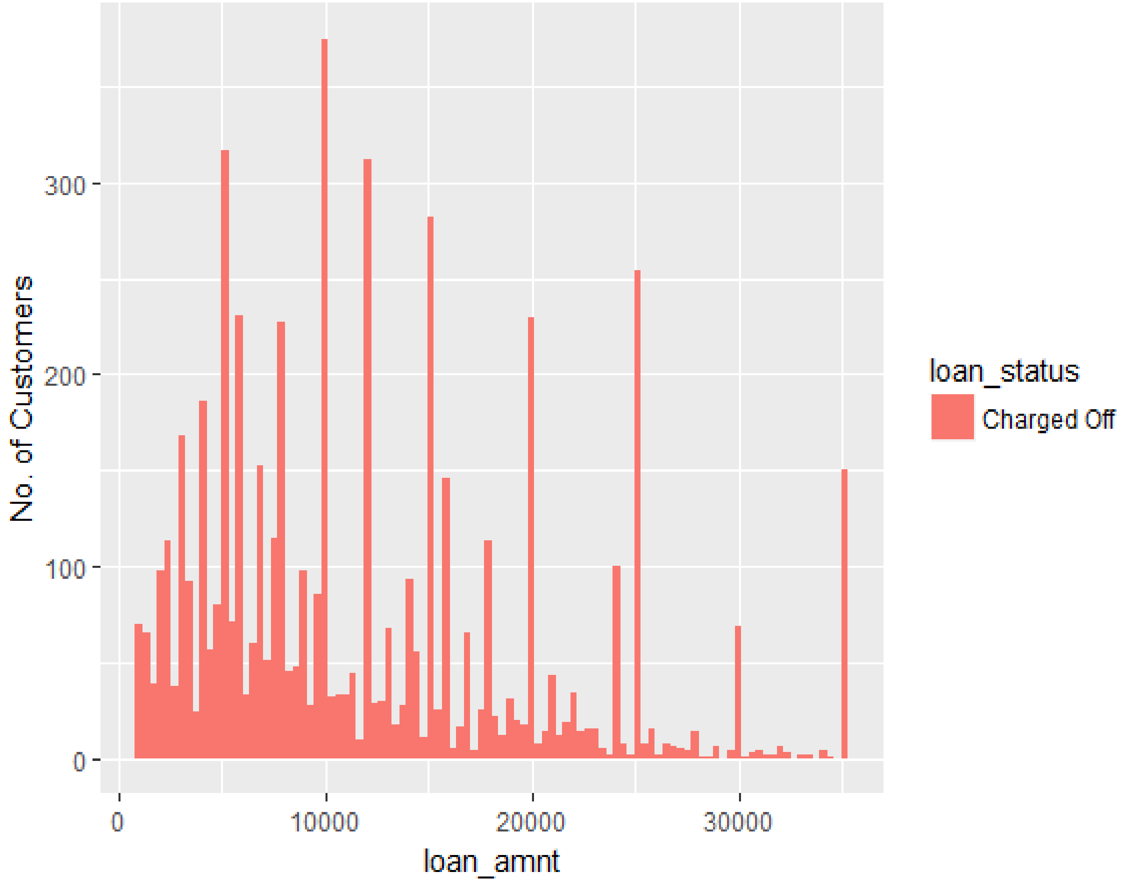
Cleaning the dataset

1. The dataset has lot of redundant and stale data like NA values, repeating values and missing values.
2. The Summary of cleaned data is as follows:
   1. **Checking for duplicate values:** *There are no duplicate values in the dataset.*
   2. **Checking missing values:** *While loading the dataset used “na.strings=c("","NA")” as an argument to replace the missing values with NA’s.*
   3. **Checking for NA Values:** *There are lot of NA values in the dataset. Therefore for the columns having all the rows equal to “NA” were removed. More than 50 columns have been removed by this. Along with this, 3 columns names “mths_since_last_delinq”, “mths_since_last_record” and “next_pymnt_d” have been removed since they majority of data as NA’s.*
   4. **Checking for Outliers in Annual Income Attributes:** *There were some outliers and they were replaced by 5% and 95% of values depending upon where they were lying in 1.5 times Inter Quartile Range (IQR).*
   5. **Handling Strings:** *There were some characters in “int_rate” and “emp_length” which were removed subsequently to make it all integer columns.*
   6. **Removing values where there are many "0", "1" or any other entry which is same for all the rows:** *Removed as many as 15 columns as the values for all the rows were same and they were not helping in analysis, "collections_12_mths_ex_med", "pub_rec", "delinq_2yrs", "out_prncp", "out_prncp_inv", "total_rec_late_fee", "acc_now_delinq", "policy_code", "initial_list_status", "chargeoff_within_12_mths", "application_type", "delinq_amnt", "pub_rec_bankruptcies", "tax_liens", "recoveries", "collection_recovery_fee“.*
   7. **Handling Date formats:** *Dates were converted in suitable standard formats for “issue_d”, “earliest_cr_line”, “last_pymnt_d” and “last_credit_pull_d”.*

Univariate & Segmented Analysis



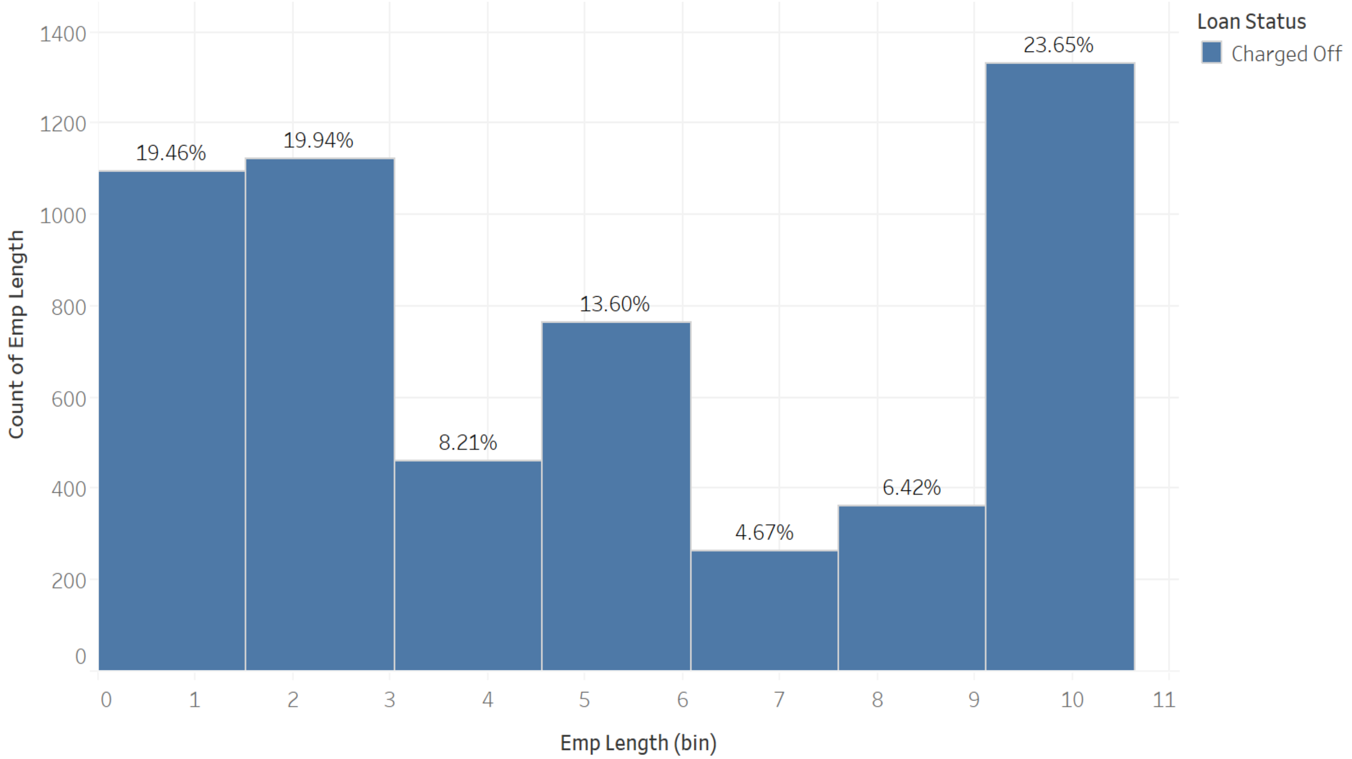
The trend of count of Loan Amnt for Loan Amnt (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



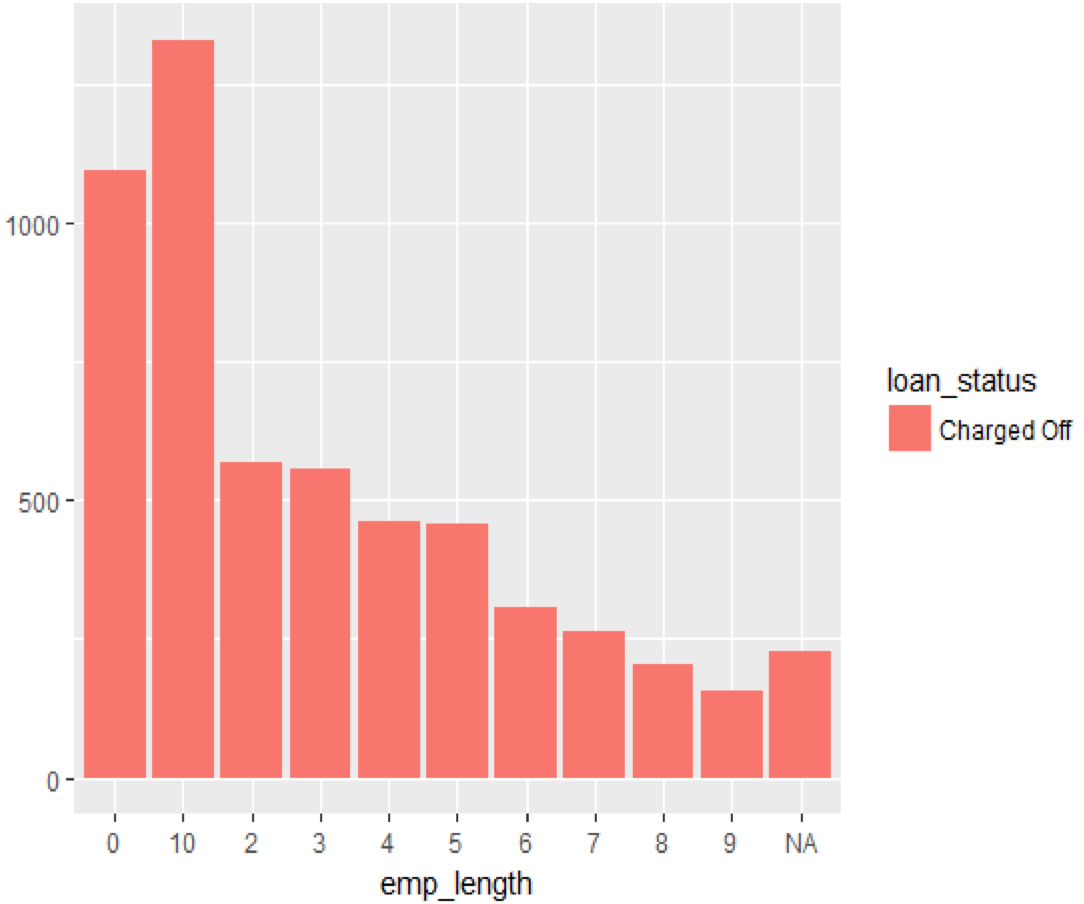
As can be seen here, most of "Charged Off" loan are for "loan_amnt" ranging from 2k to 12K.

Univariate & Segmented Analysis

Emp_Length



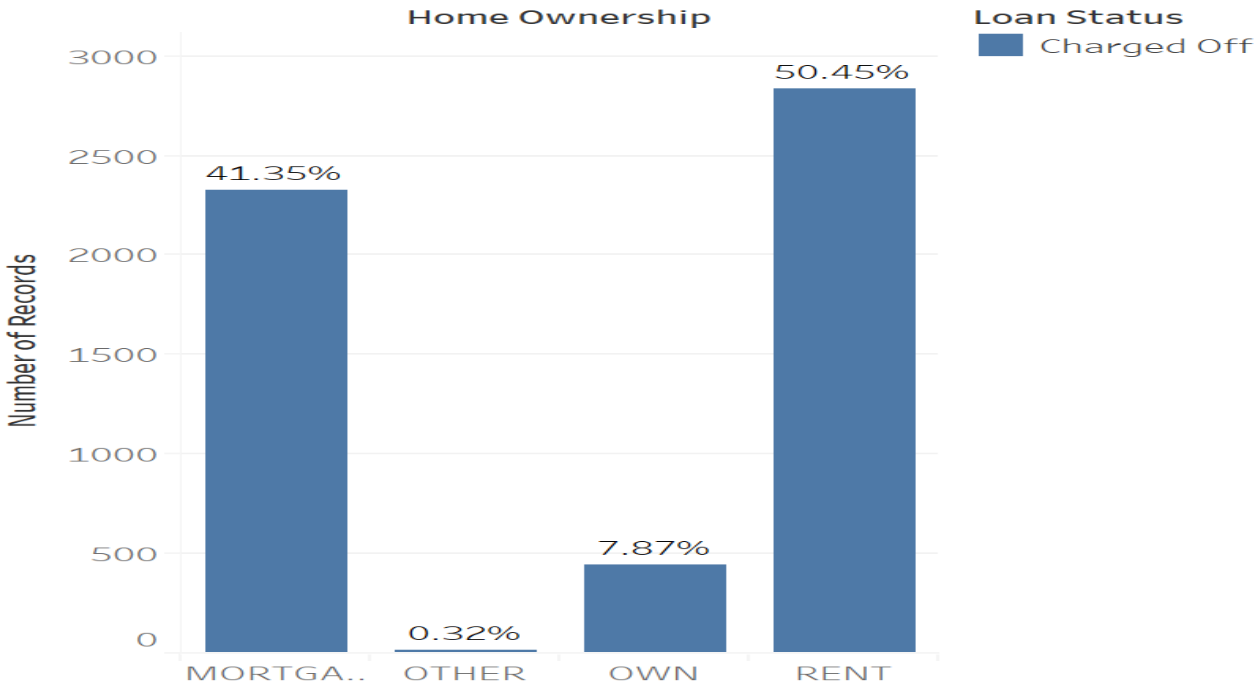
The trend of count of Emp Length for Emp Length (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



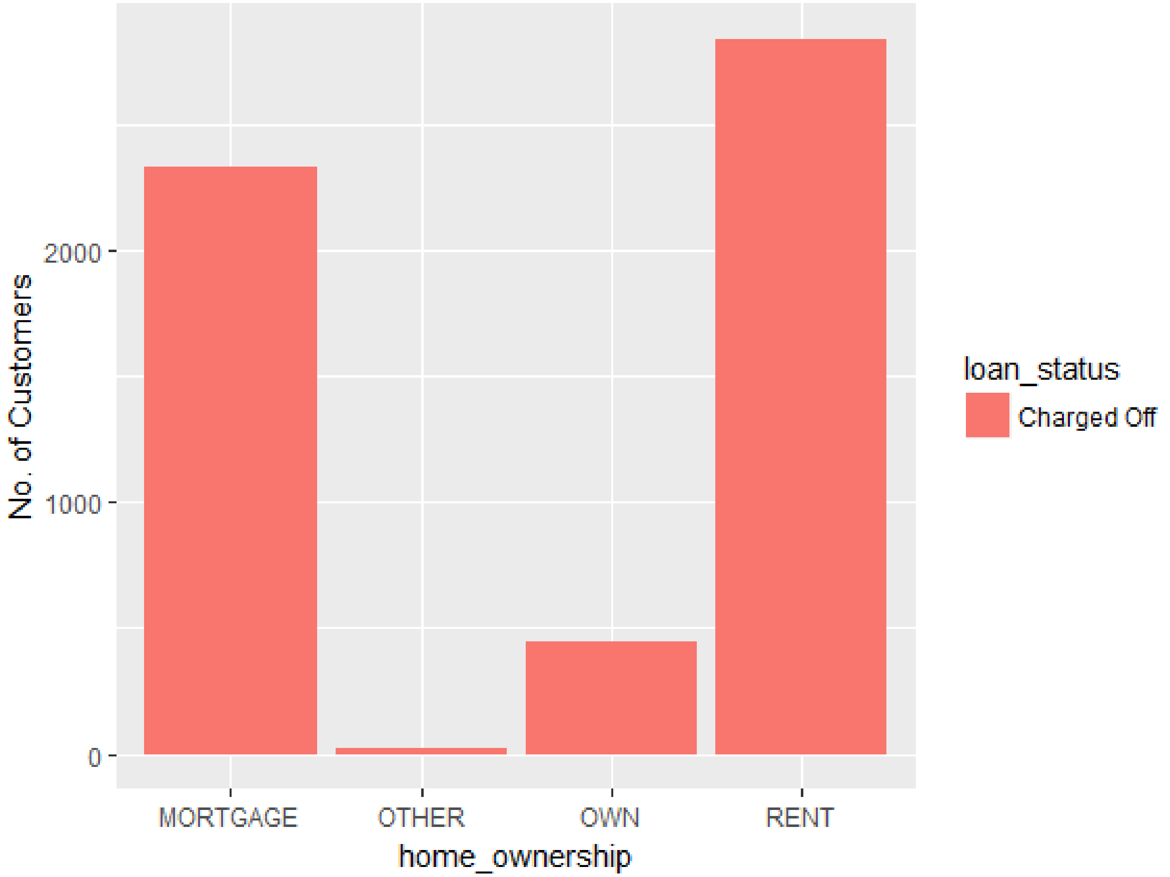
As can be seen here, most of "Charged Off" loan are for "emp_length" < 1 year and for 10+ years.

Univariate & Segmented Analysis

Home_Ownership



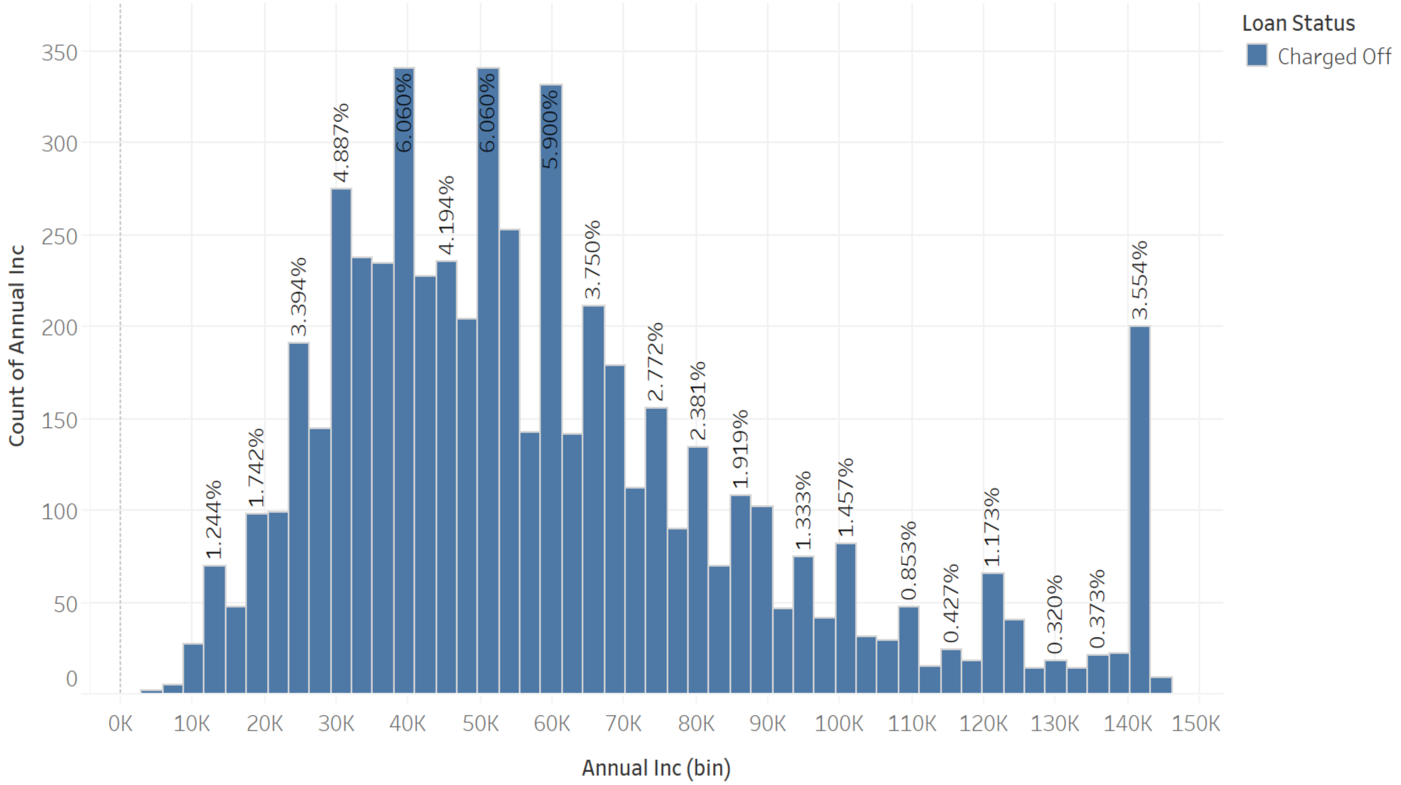
Sum of Number of Records for each Home Ownership. Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



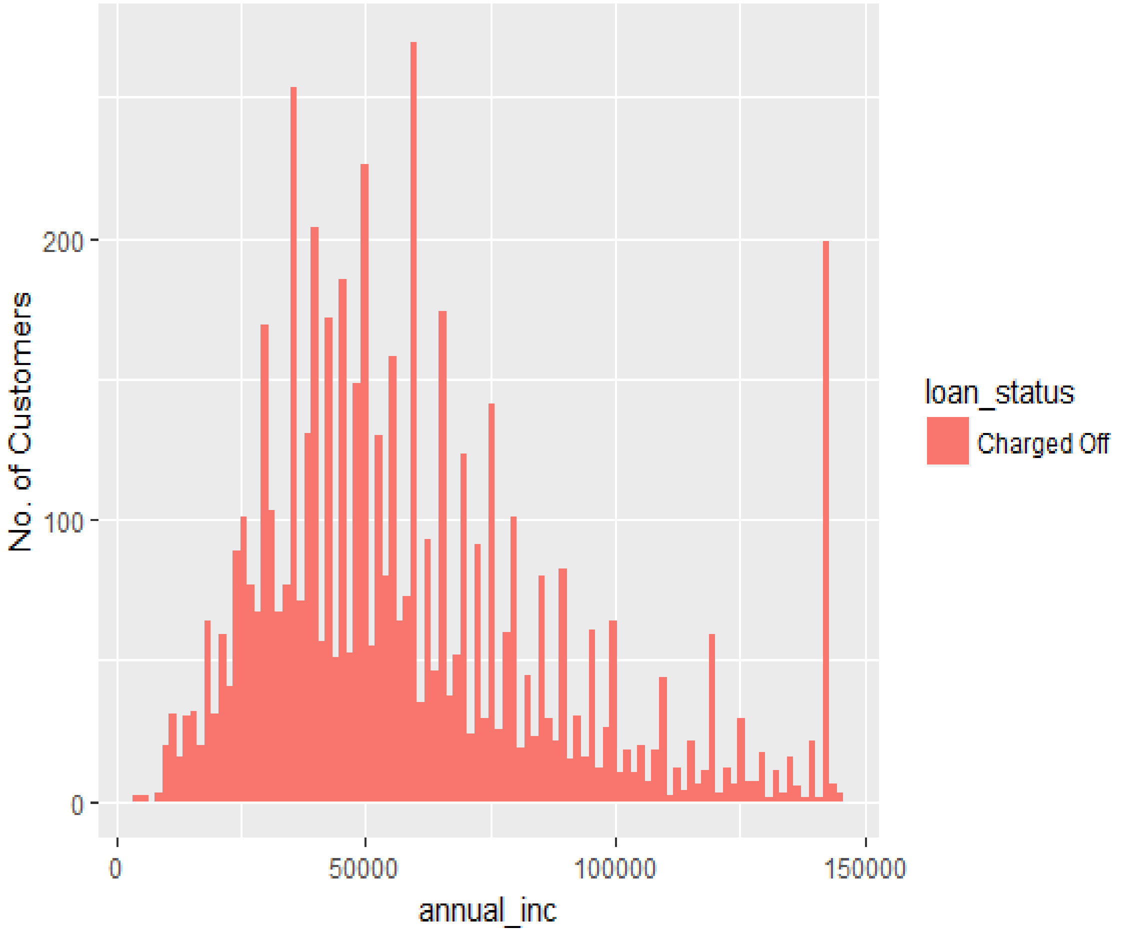
As can be seen here, most of "Charged Off" loan are for "home_ownership" type of "MORTGAGE" and "RENT".

Univariate & Segmented Analysis

Annual_Income



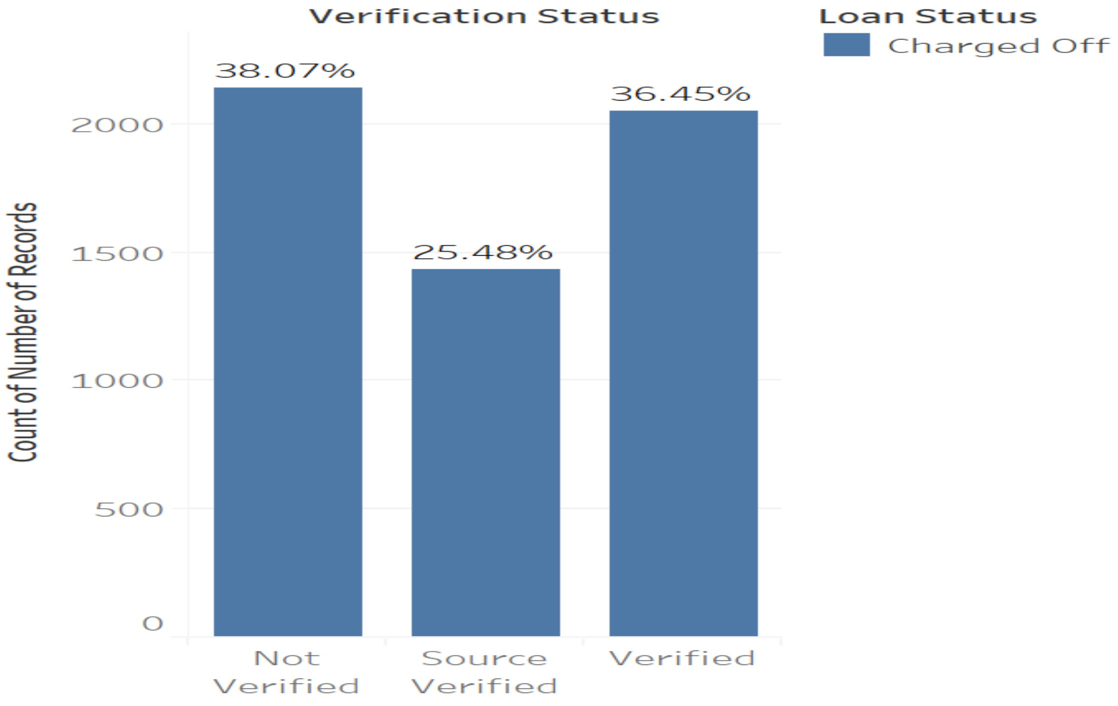
The trend of count of Annual Inc for Annual Inc (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



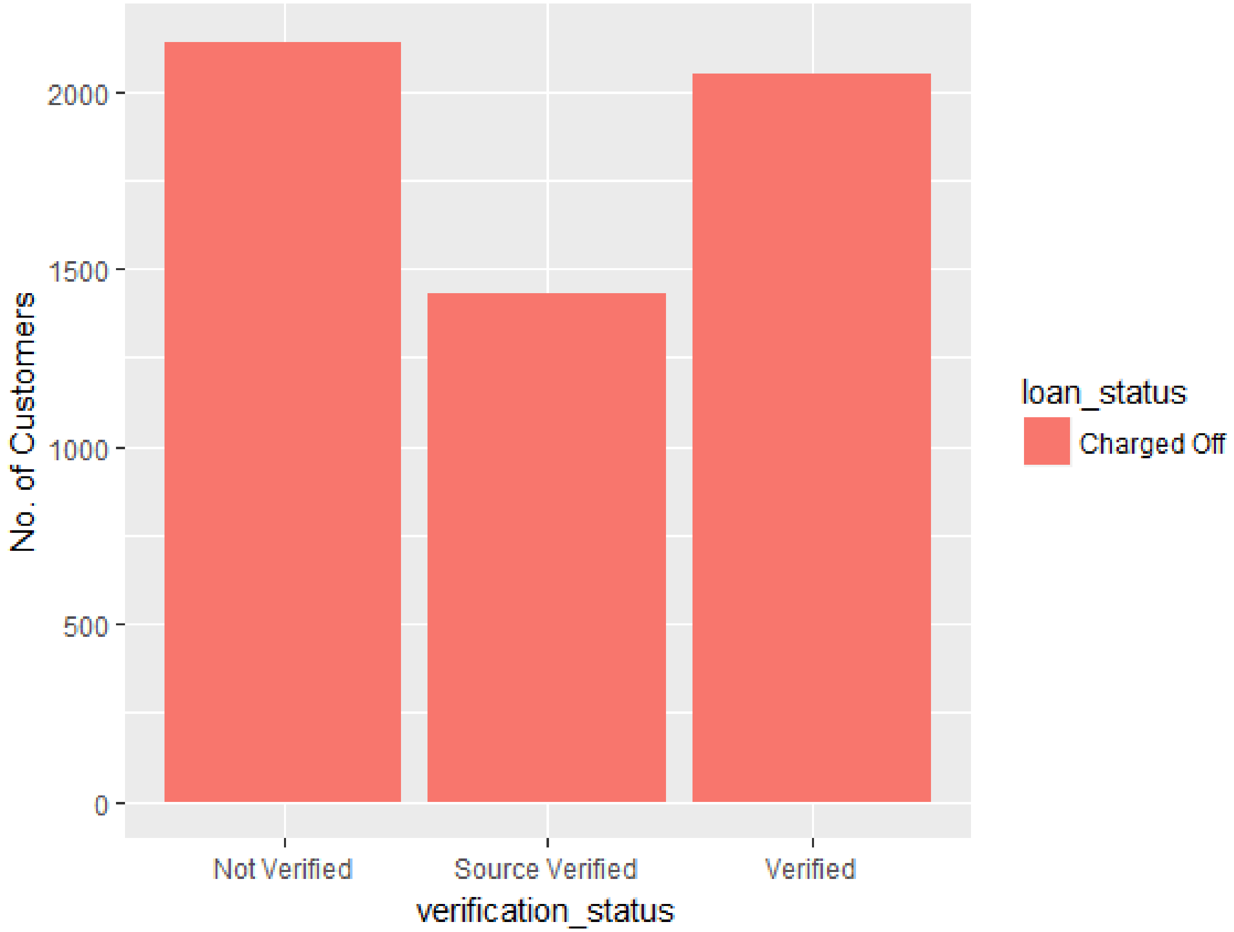
As can be seen here, most of "Charged Off" loan are for "annual_inc" in between 30K $ 60K. There is a spike at the end for around 140K this can be due the outlier treatment we did earlier.

Univariate & Segmented Analysis

Verification_Status



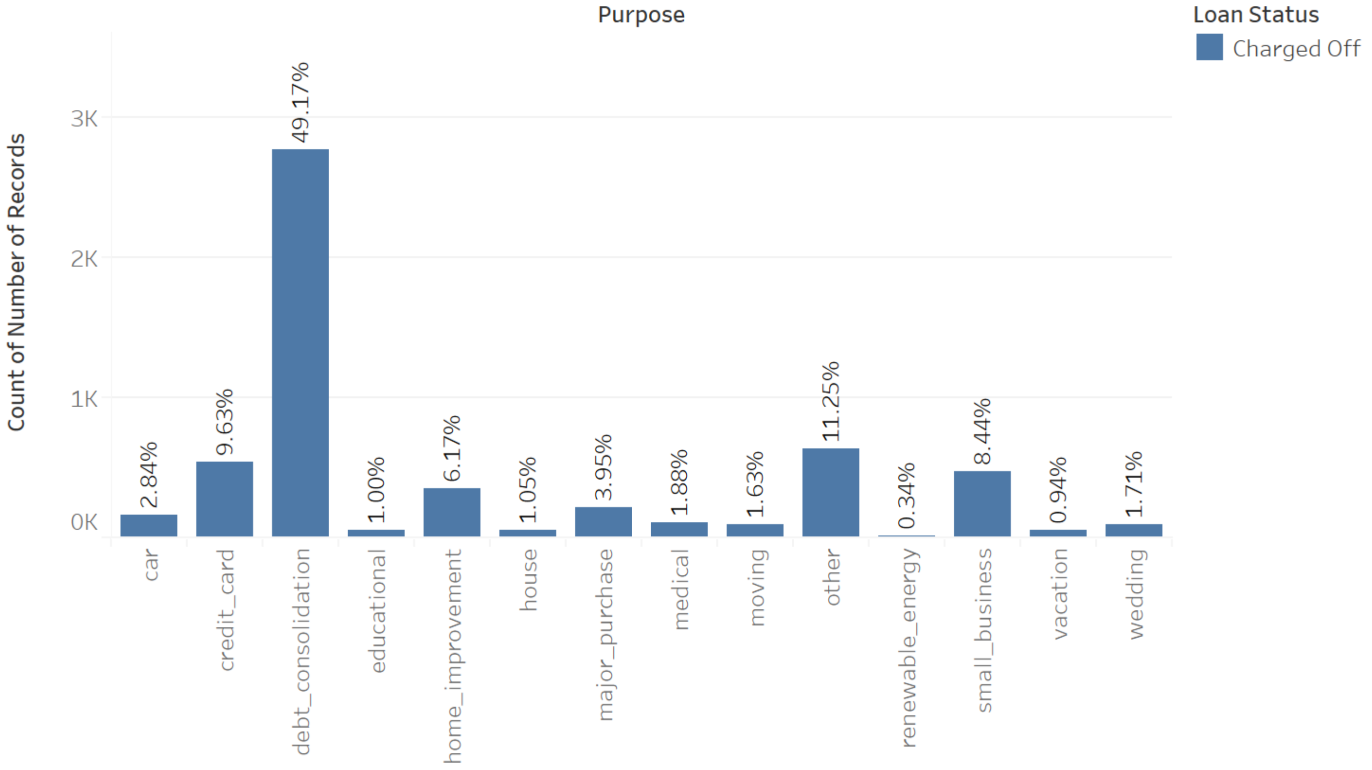
Count of Number of Records for each Verification Status. Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



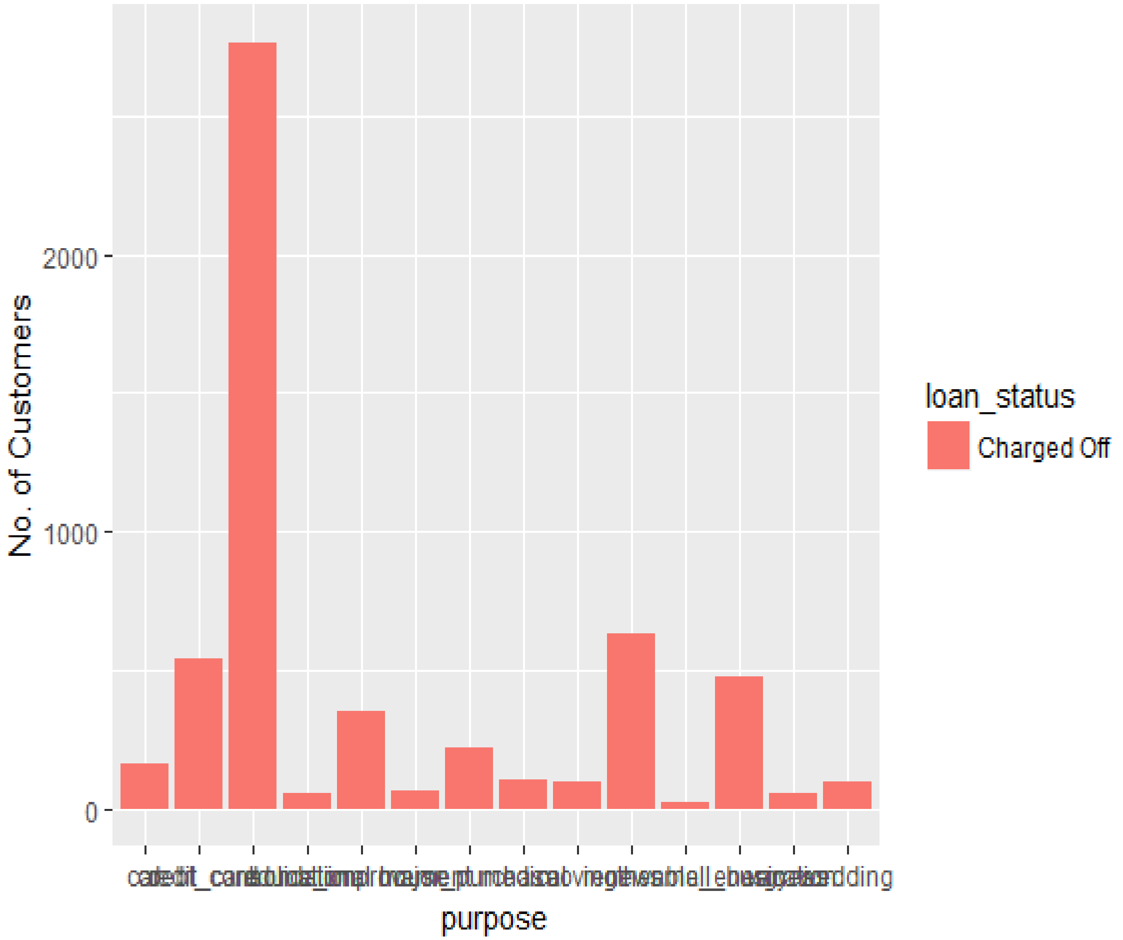
As can be seen here, most of "Charged Off" loan are for "verification_status" set as "Not Verified" and "Verified".

Univariate & Segmented Analysis

Purpose



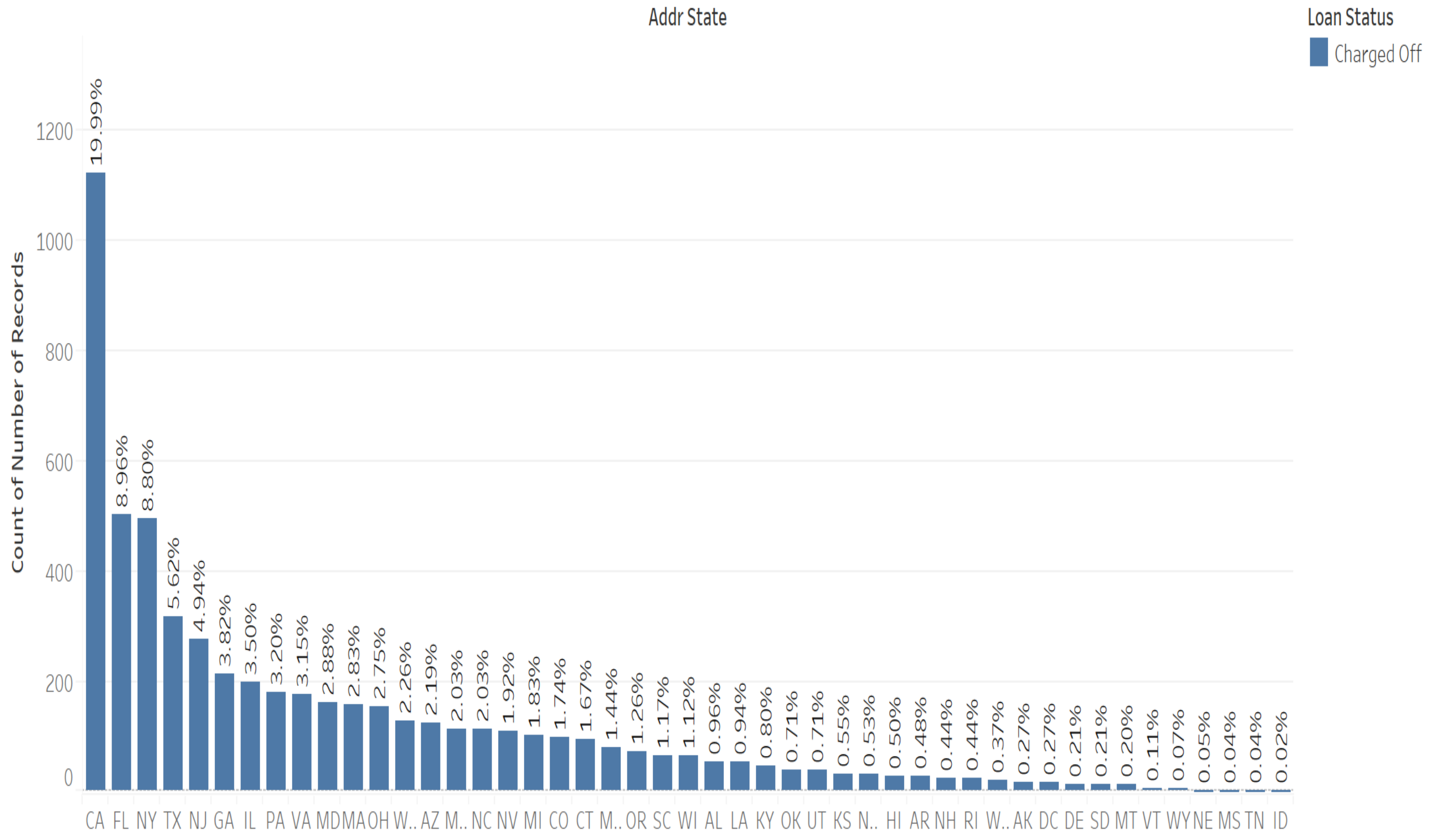
Count of Number of Records for each Purpose. Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



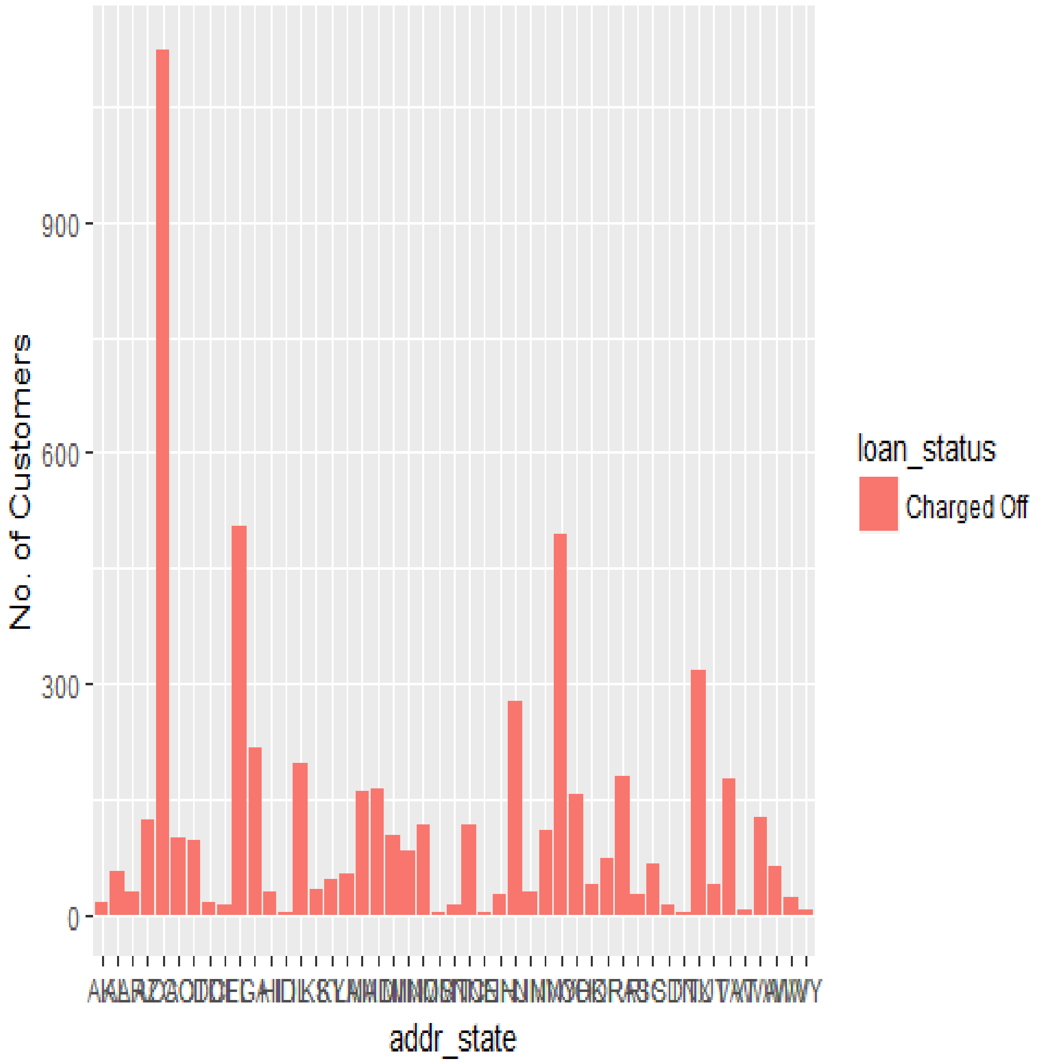
As can be seen here, most of "Charged Off" loan are for "purpose" equal "debt_consolidation".

Univariate & Segmented Analysis

Addr_State

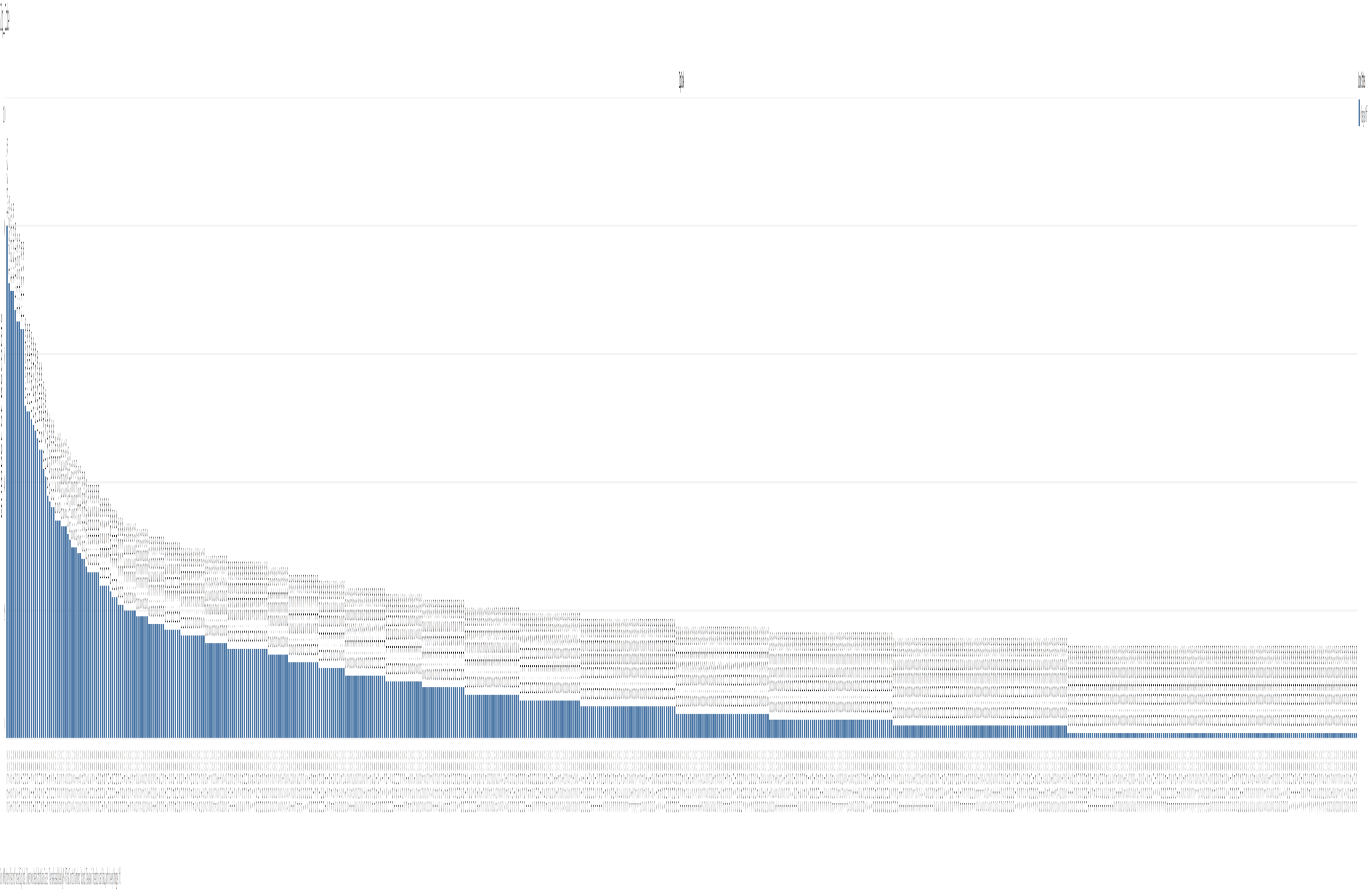


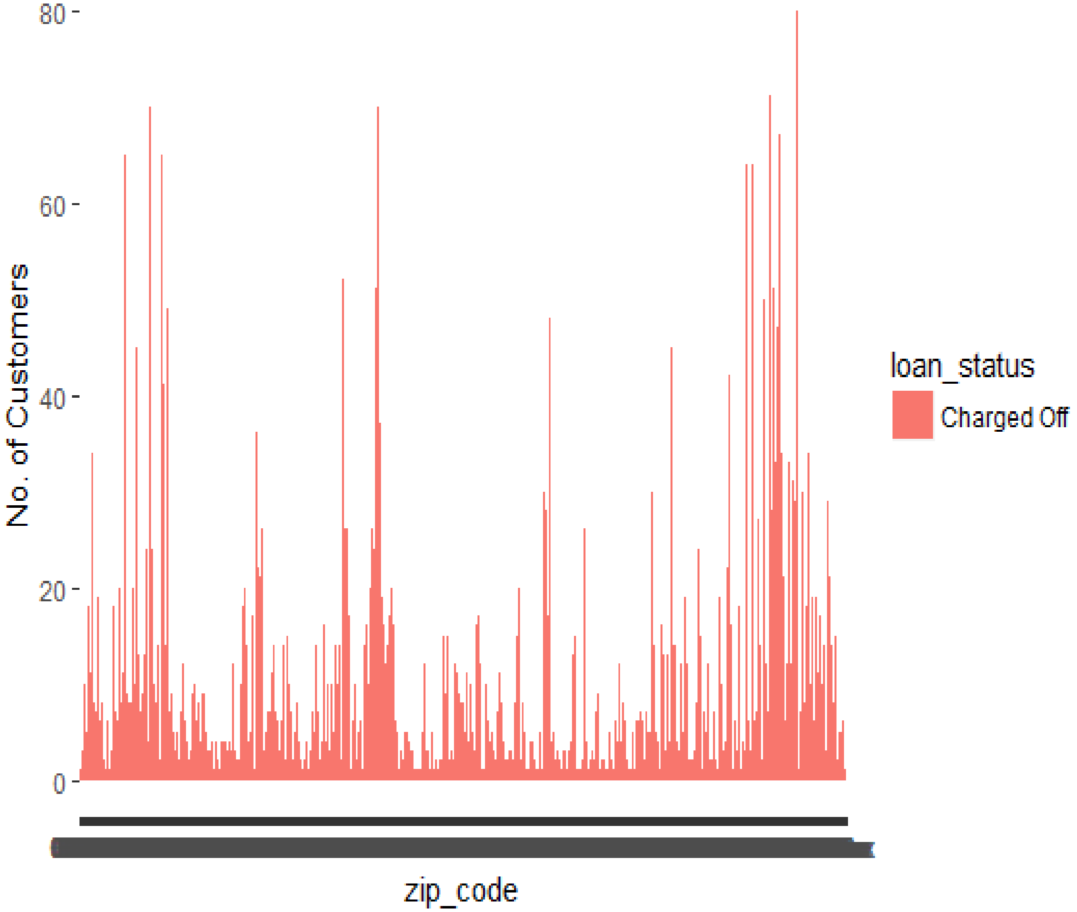
Count of Number of Records for each Addr State. Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



As can be seen here, most of "Charged Off" loan are for "addr_state" equal to "CA“ at around 20%.

Univariate & Segmented Analysis

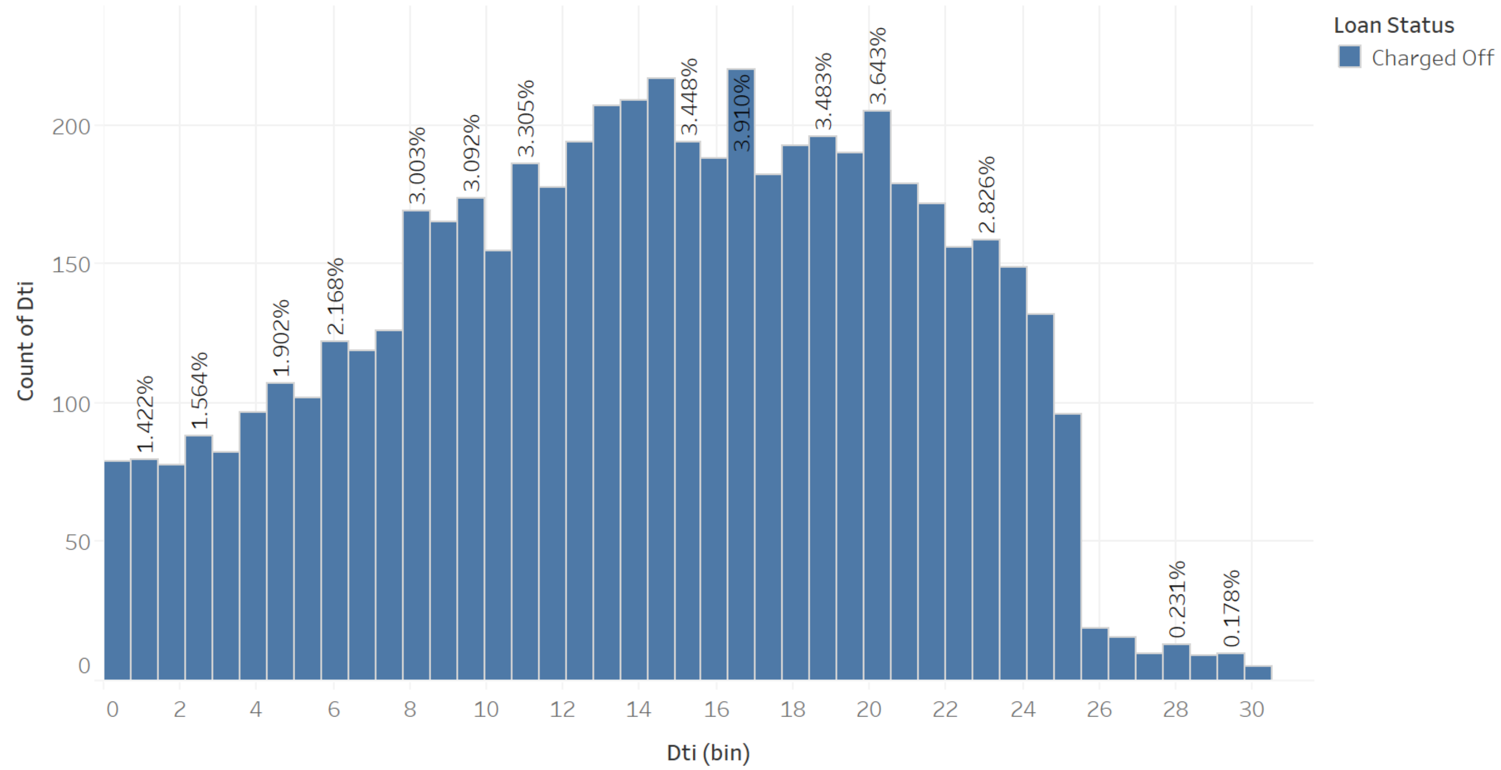




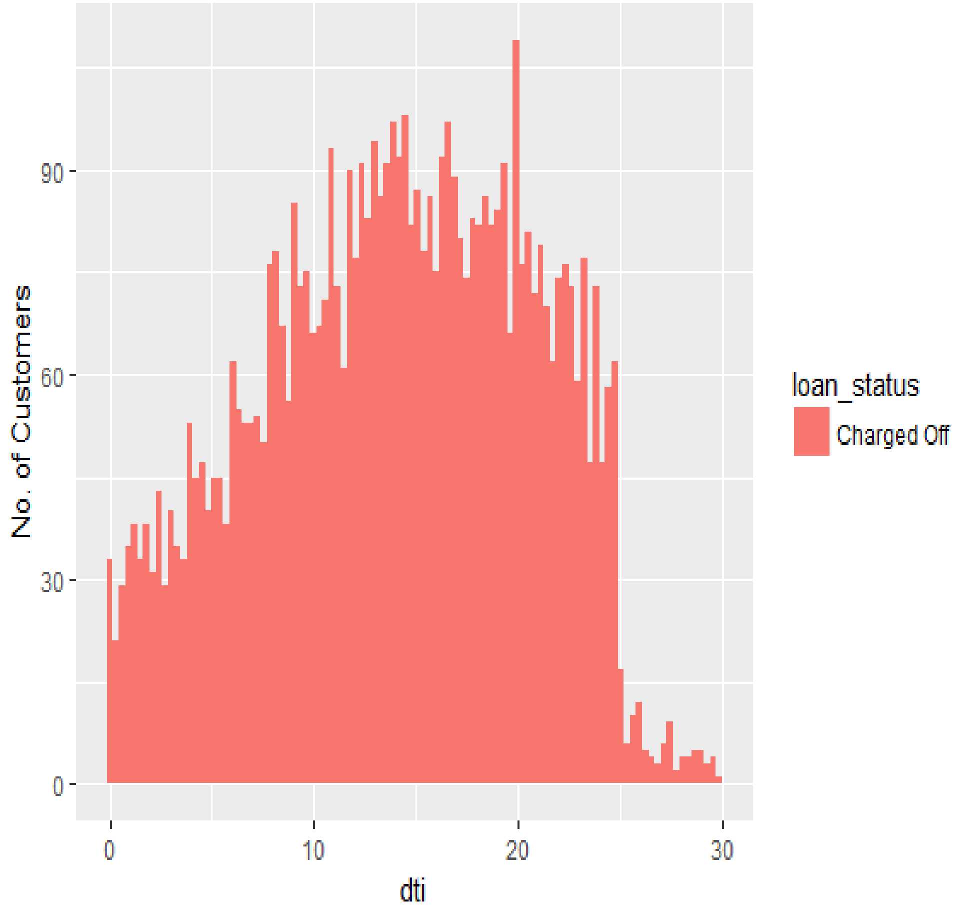
As can be seen here, most of "Charged Off" loan are for "zip_code" mainly of 8 zipcodes like 945XXX, 917XXX etc.

Univariate & Segmented Analysis

dti

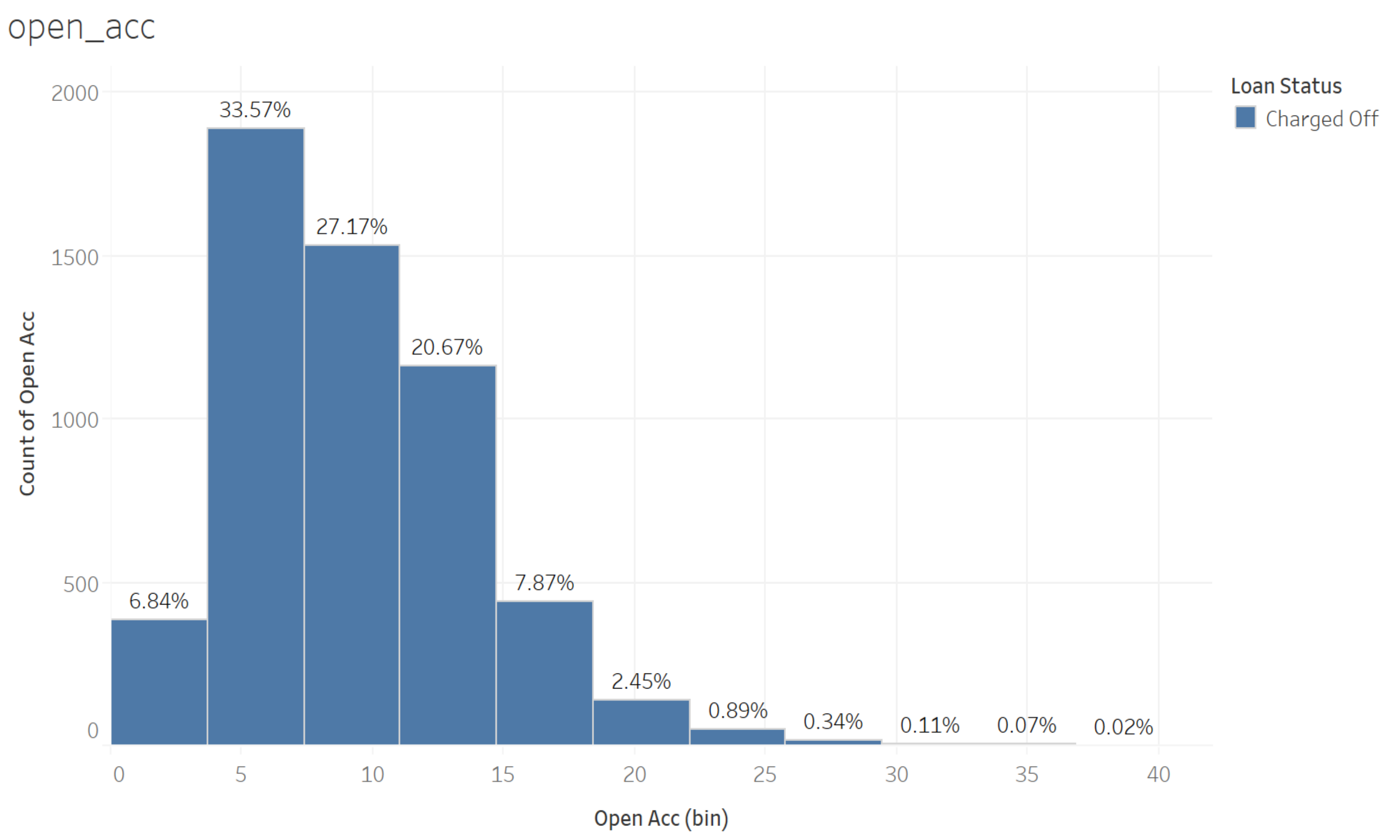


The trend of count of Dti for Dti (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.

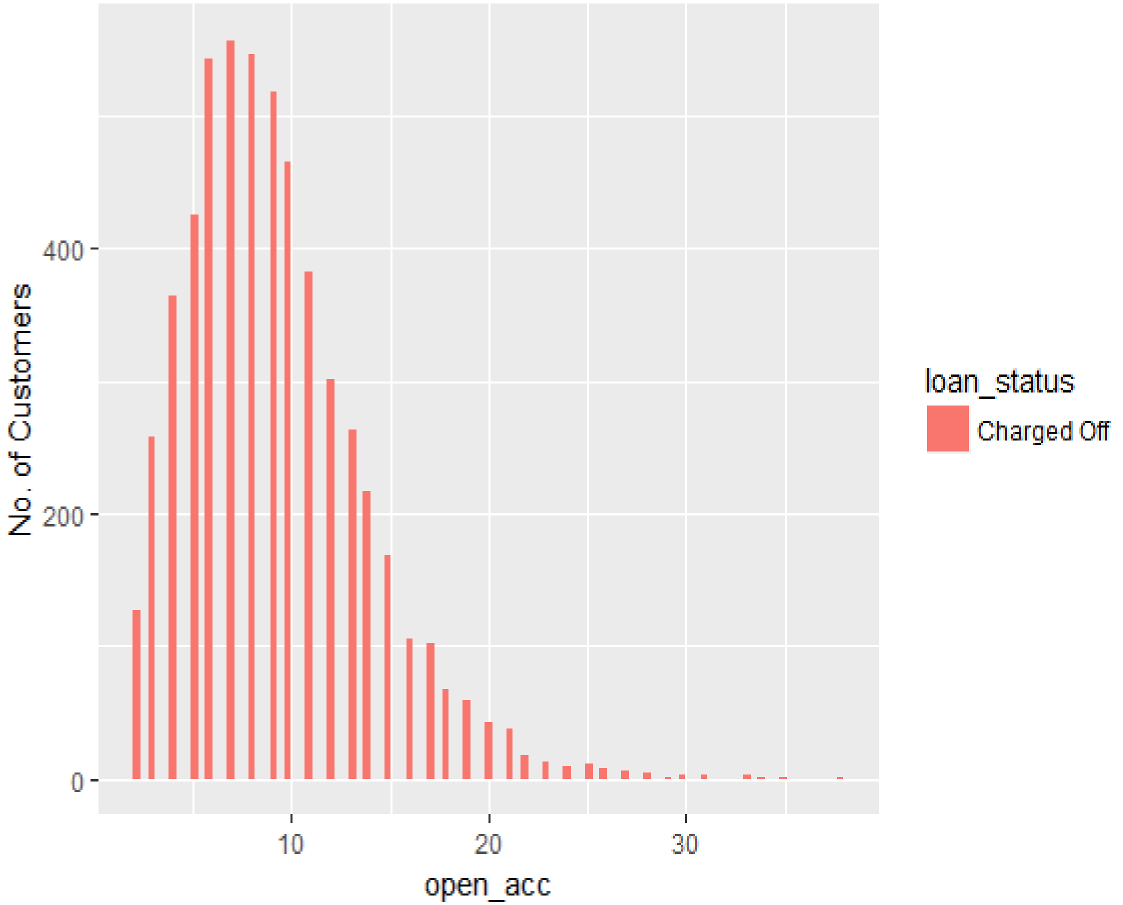


As can be seen here, most of "Charged Off" loan are for "dti" ranges from 12 to 20. A low “dti” has less loan defaults.

Univariate & Segmented Analysis

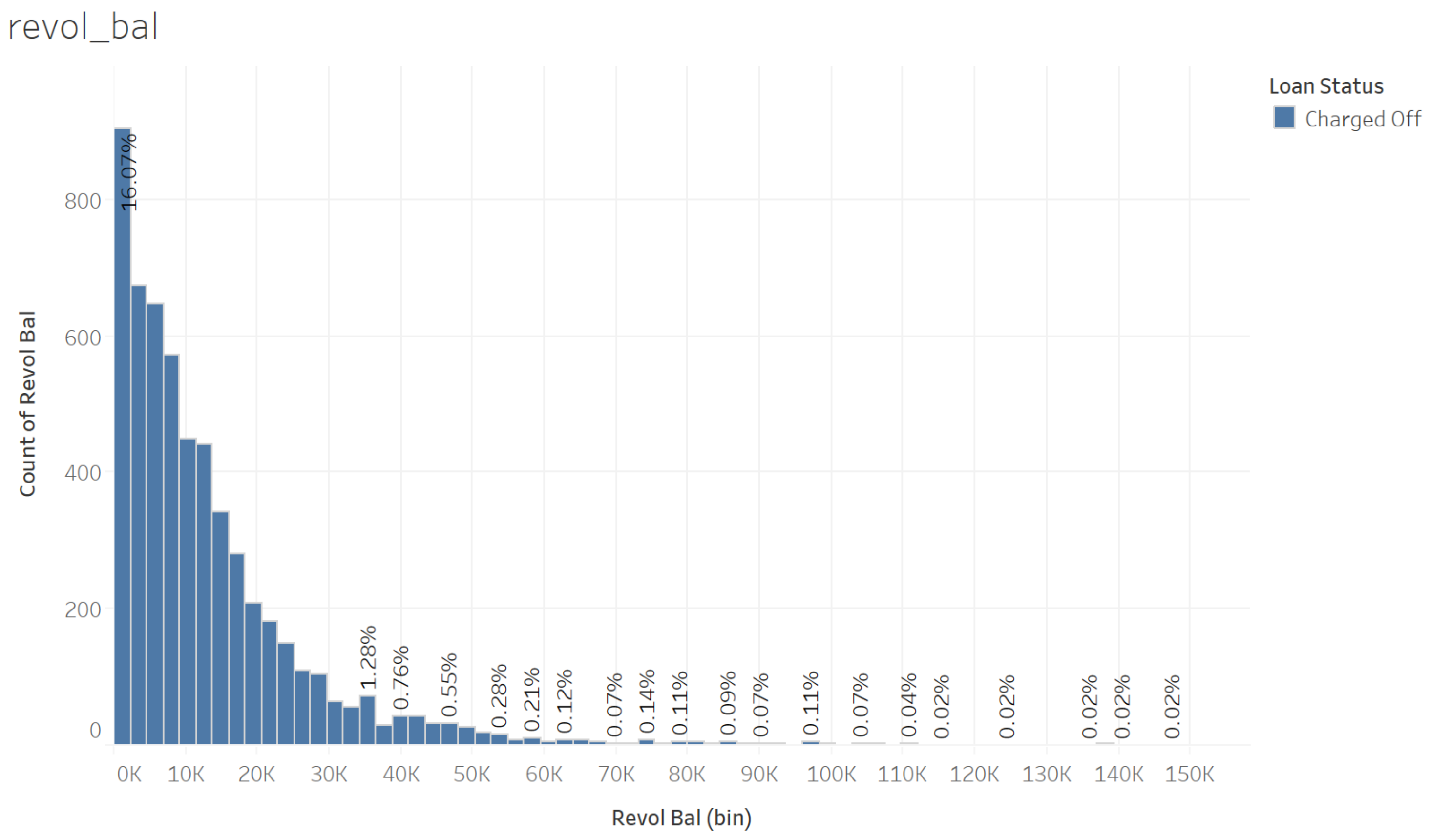


The trend of count of Open Acc for Open Acc (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.

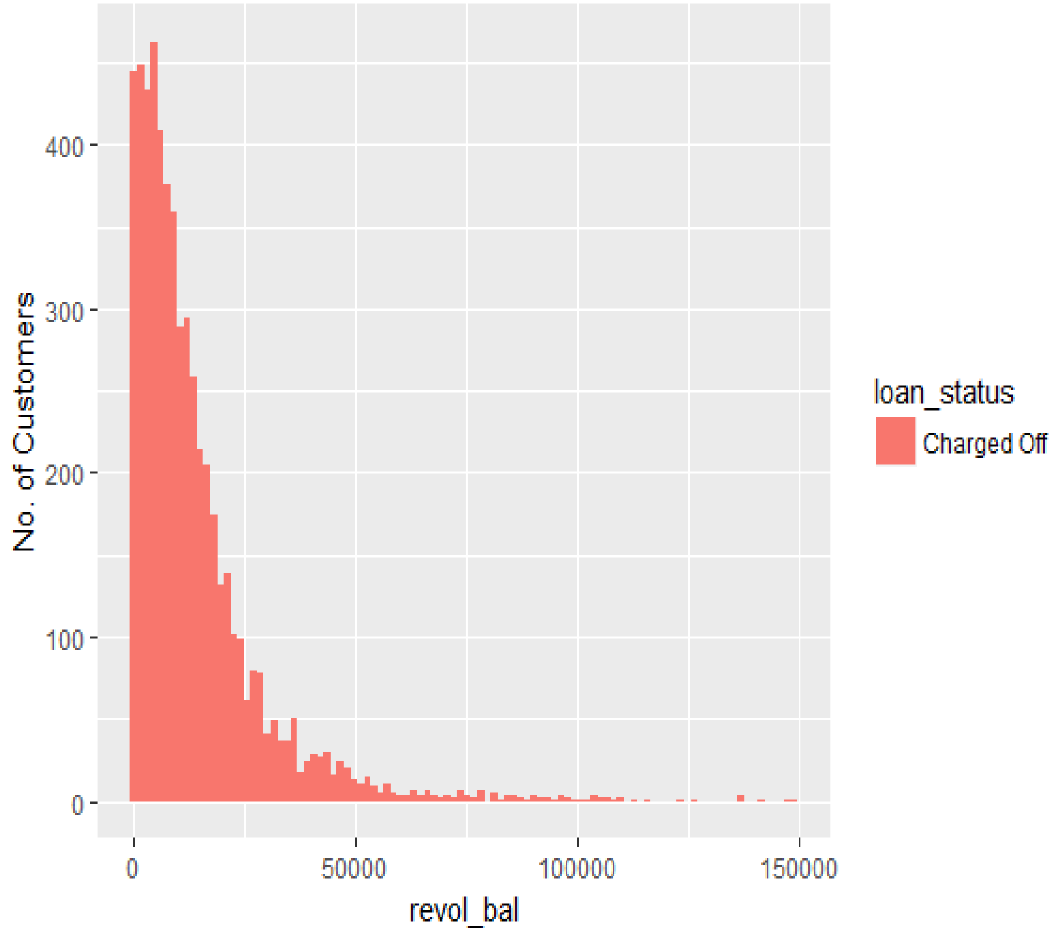


As can be seen here, most of "Charged Off" loan are for "open_acc" i.e. no. of credit lines ranging from 5 to 17.

Univariate & Segmented Analysis

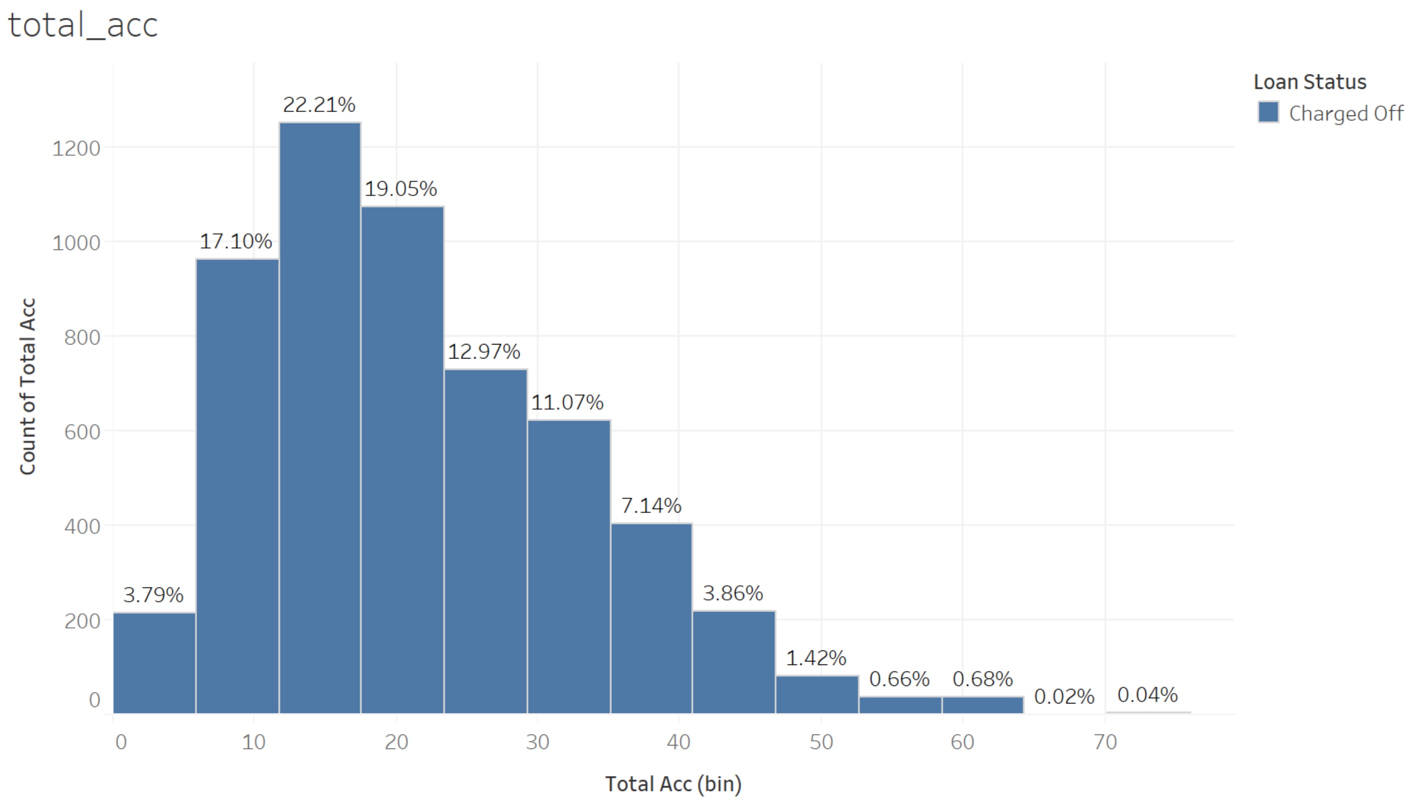

revol_bal

The trend of count of Revol Bal for Revol Bal (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.

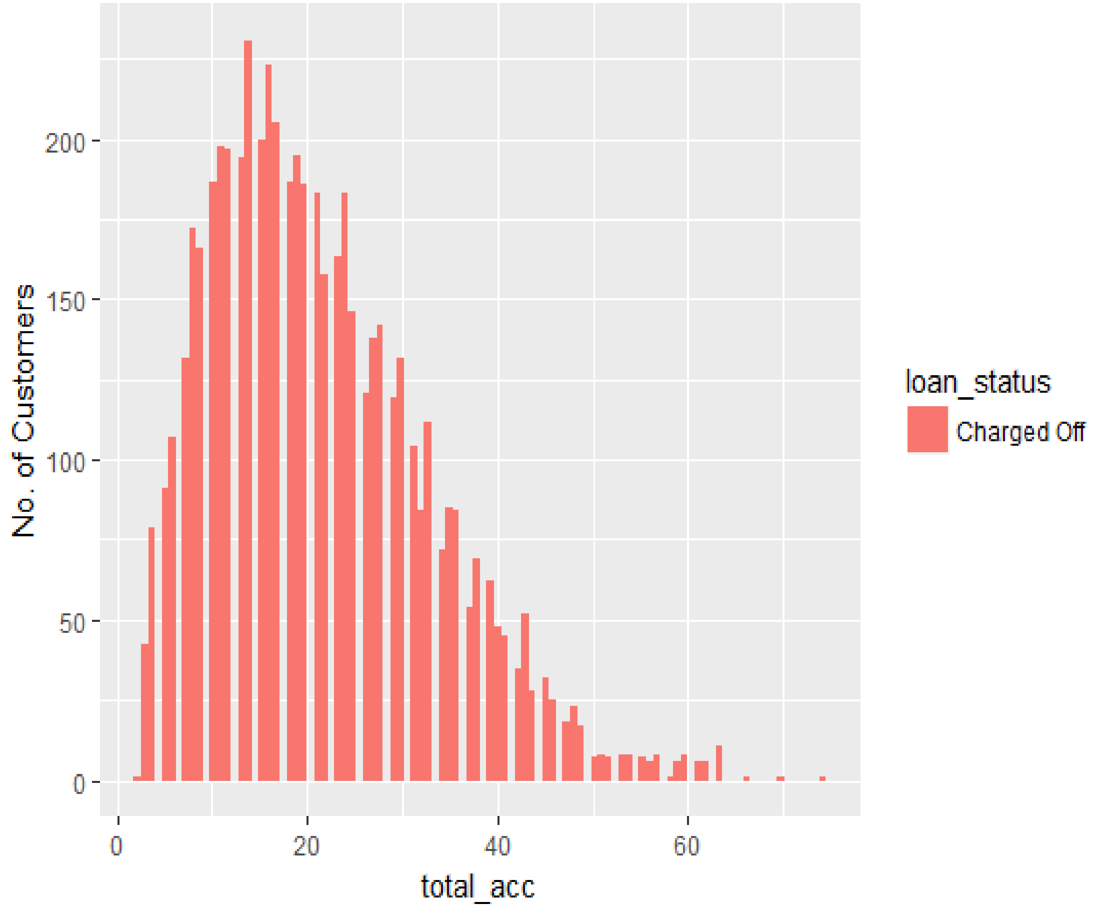


As can be seen here, most of "Charged Off" loan are for "revol_bal" ranging from 0 to 16K.

Univariate & Segmented Analysis



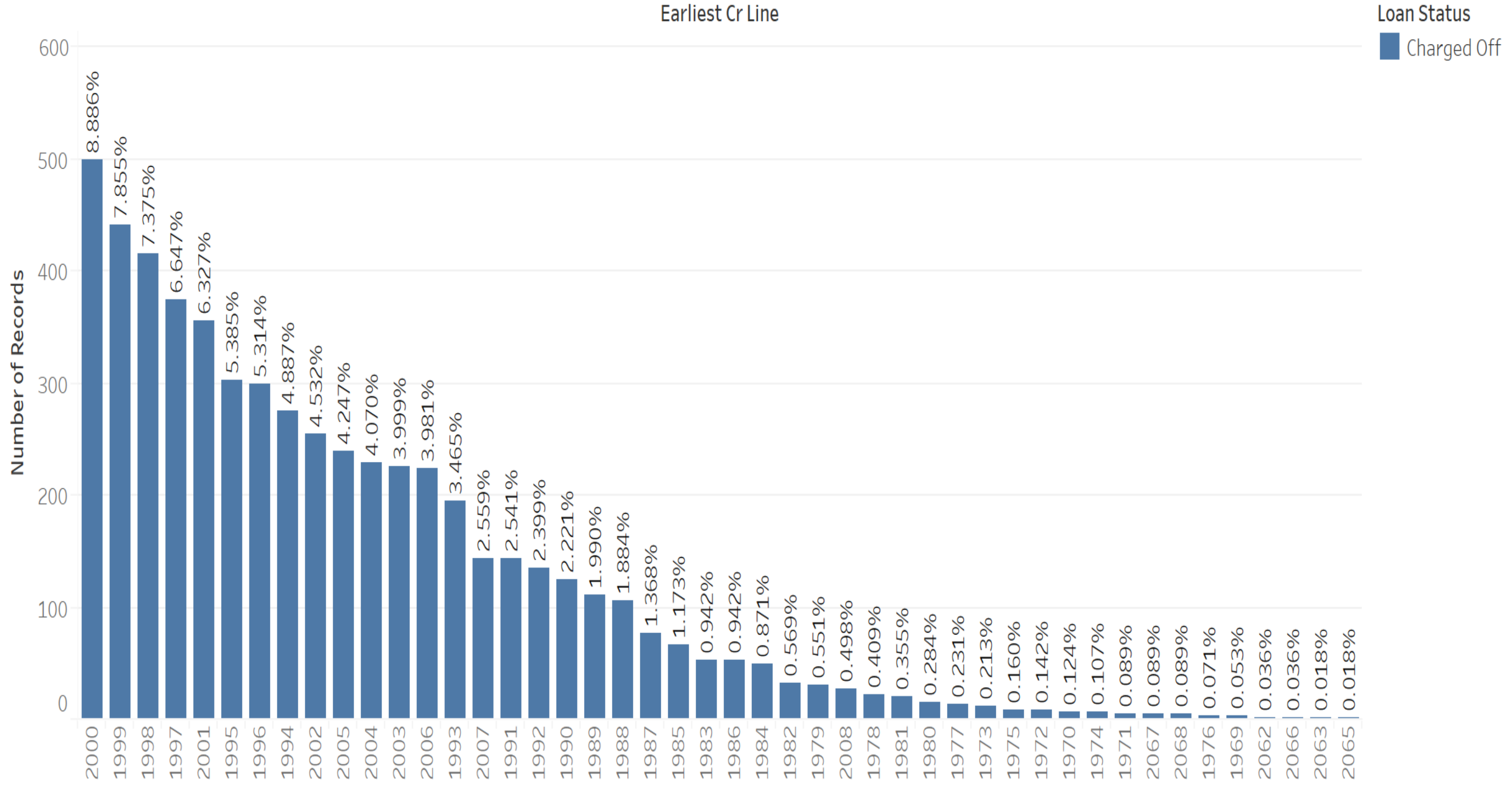
The trend of count of Total Acc for Total Acc (bin). Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.



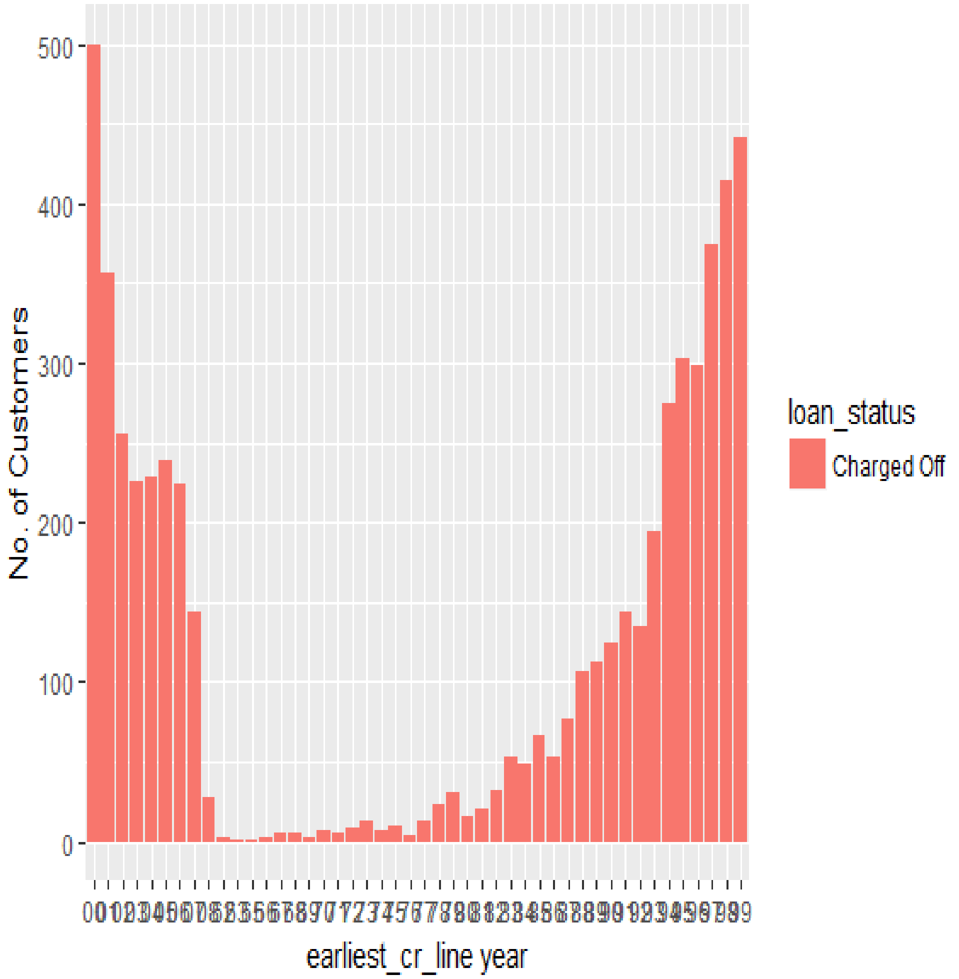
As can be seen here, most of "Charged Off" loan are for "total_acc“ i.e. no. of credit lines ranging from 6 to 23.

Univariate & Segmented Analysis

ecl_year



Sum of Number of Records for each Earliest Cr Line Year. Color shows details about Loan Status. The marks are labeled by % of Total Count of Number of Records. The view is filtered on Loan Status, which keeps Charged Off.

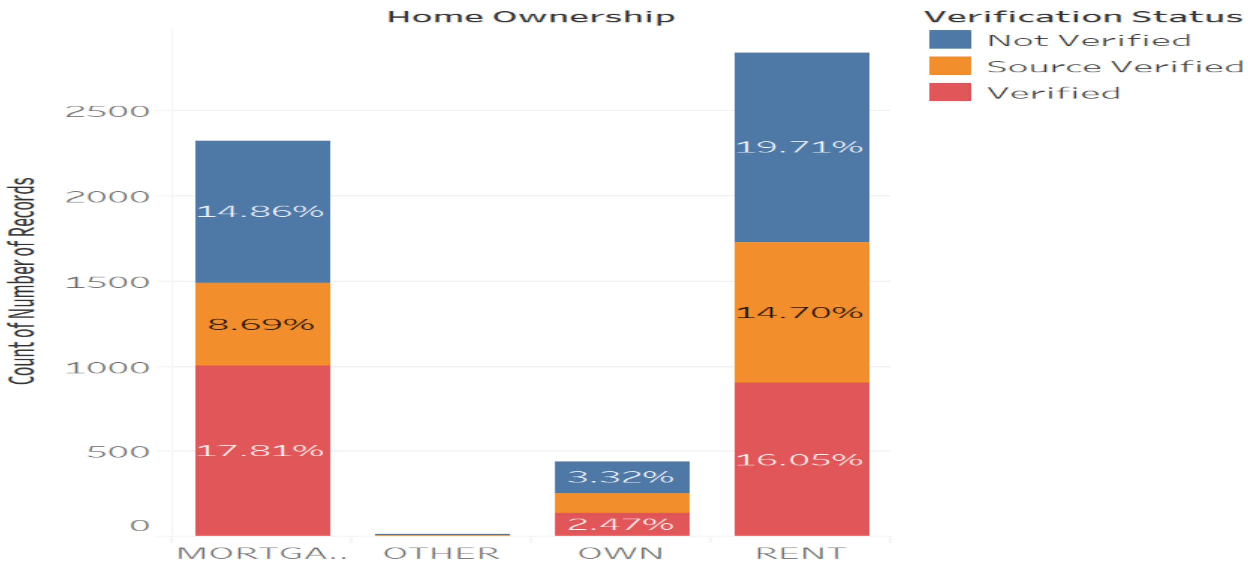


As can be seen here, most of "Charged Off" loan are for "earliest_cr_line" where it was opened in between 1994 to 2002.

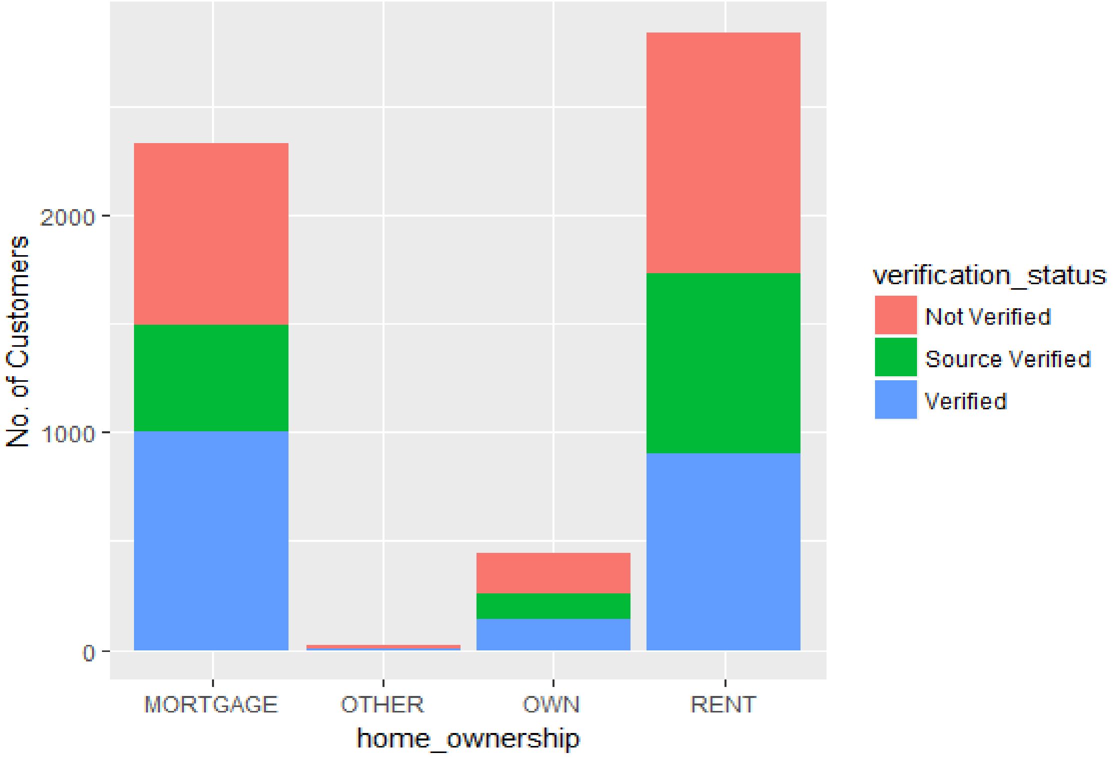
Bivariate Analysis

1. For Continuous Variables:
   - "loan_amnt" and "annual_inc" with around 43.5% correlation.
   - "open_acc" and "total_acc" with around 68.7% correlation.
2. Categorical Variables: They are as follows. For Effect of "home_ownership" and "verification_status" on "loan_status"

home_ownership_verification_status



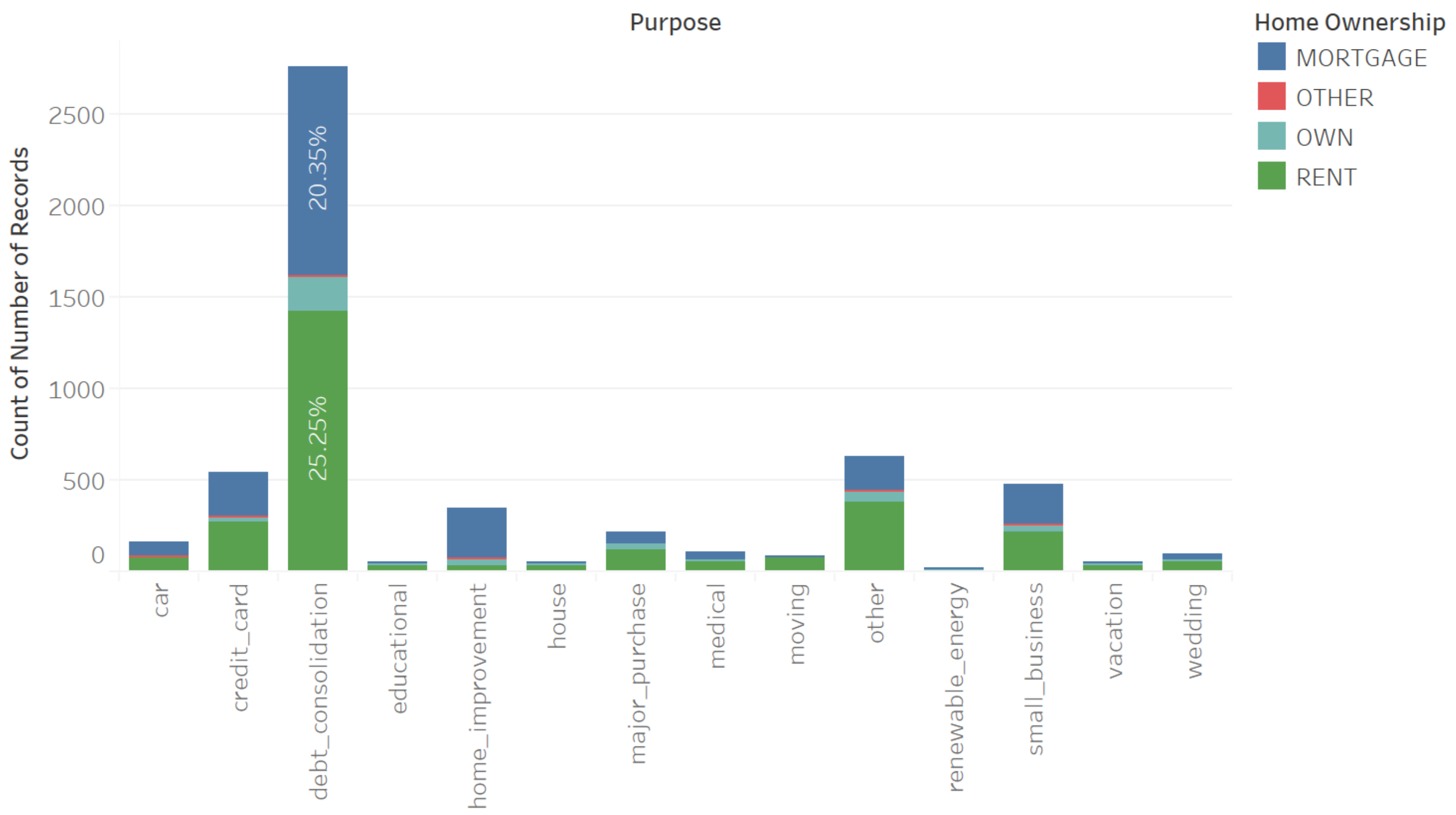
Count of Number of Records for each Home Ownership. Color shows details about Verification Status. The marks are labeled by % of Total Count of Number of Records. The data is filtered on Loan Status, which keeps Charged Off.



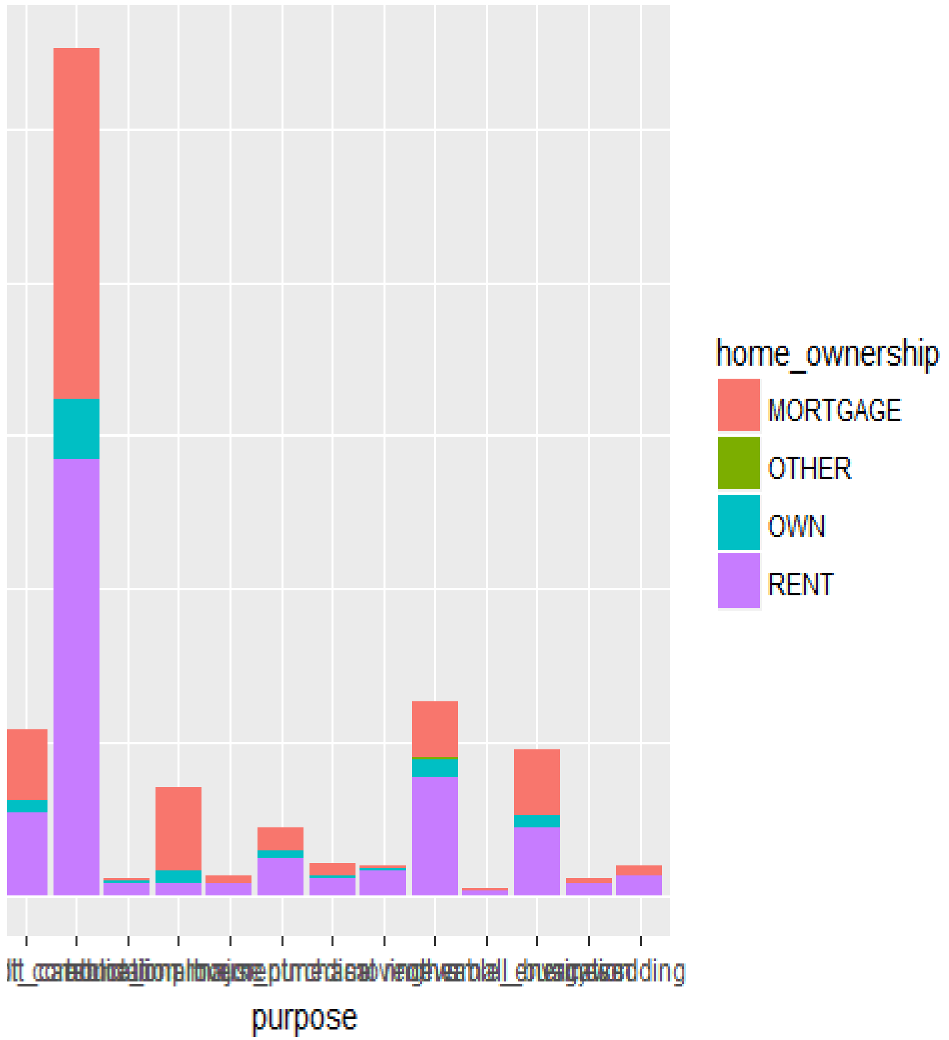
As can be seen here, most of "Charged Off" loan are for "RENT" & "NOT VERIFIED" around 19.7%, "MORTGAGE" & "VERIFIED" around 17.8%.

Bivariate Analysis

purpose_home_ownership



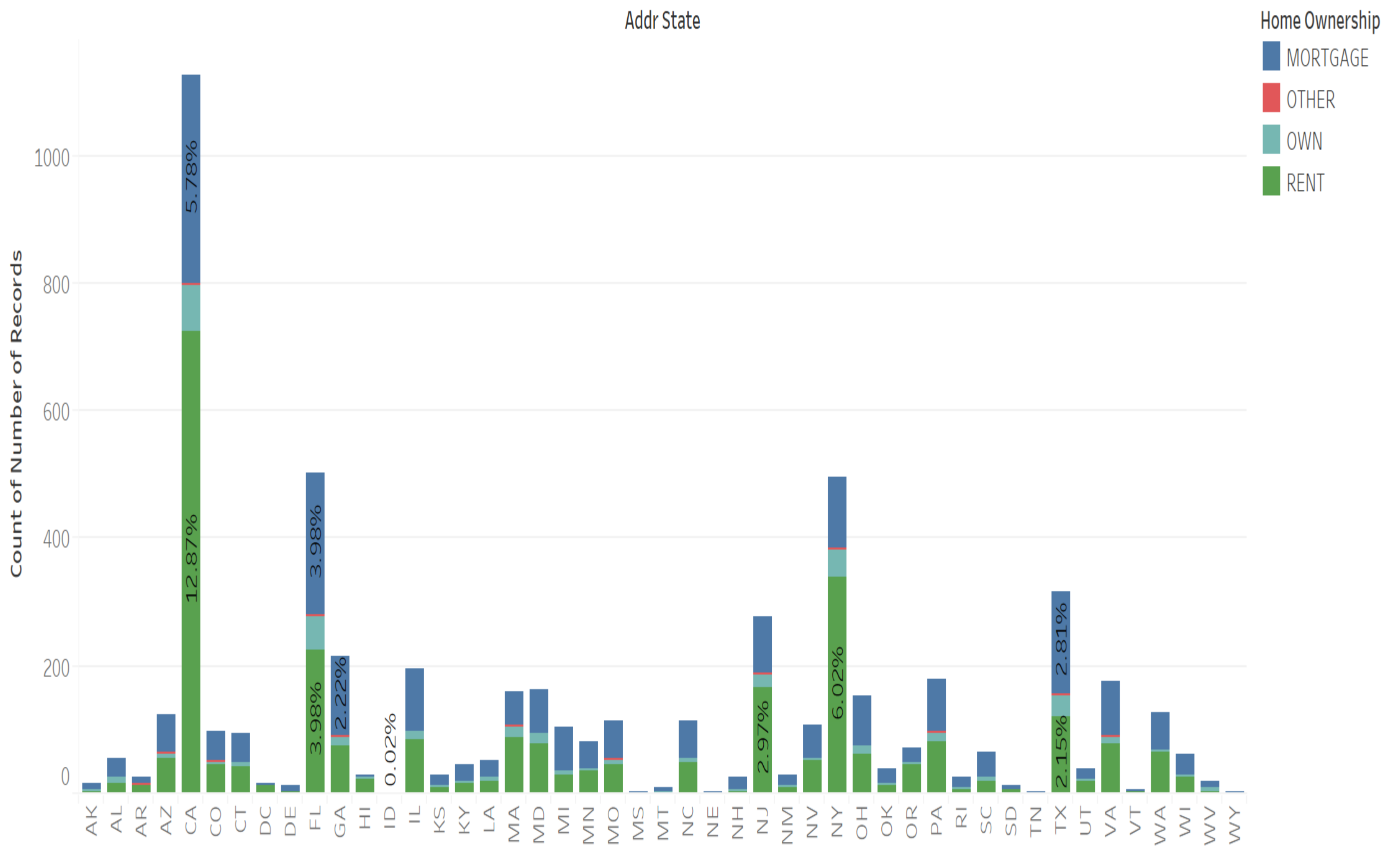
Count of Number of Records for each Purpose. Color shows details about Home Ownership. The marks are labeled by % of Total Count of Number of Records. The data is filtered on Loan Status, which keeps Charged Off.



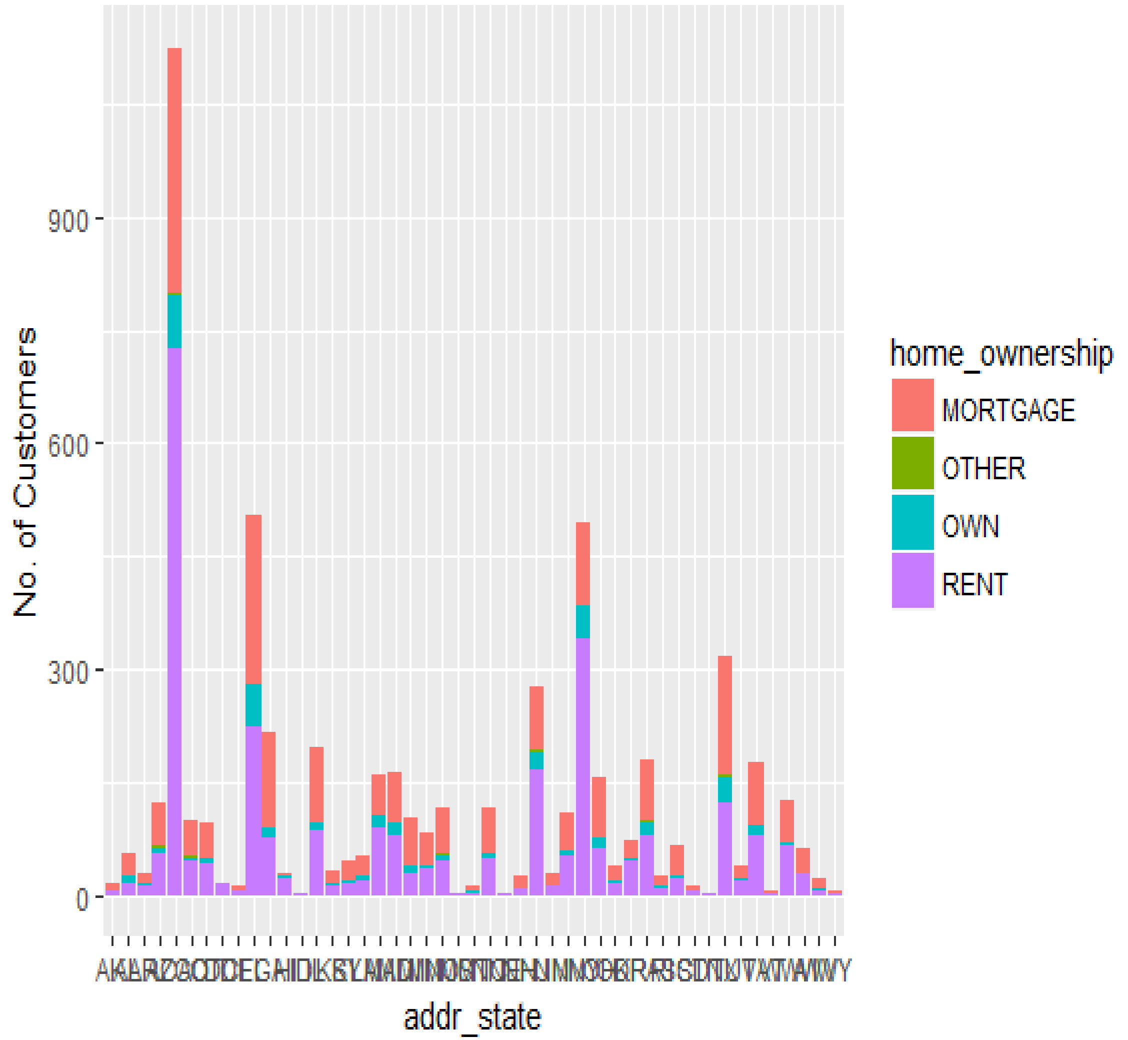
As can be seen here, most of "Charged Off" loan are for "debt_consolidation" with "RENT" & "MORTGAGE" with around 25% and 20%.

Bivariate Analysis

addr_state_home_ownership



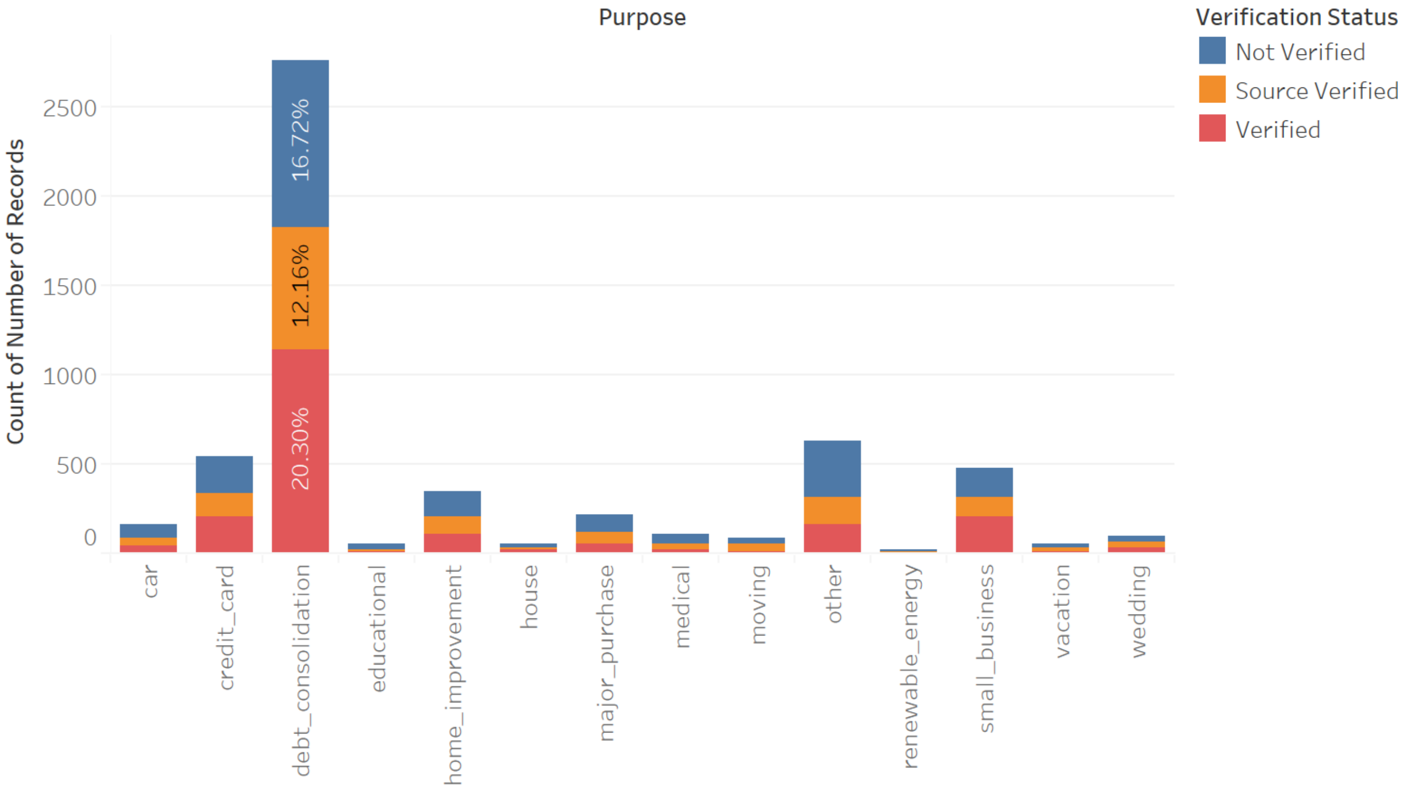
Count of Number of Records for each Addr State. Color shows details about Home Ownership. The marks are labeled by % of Total Count of Number of Records. The data is filtered on Loan Status, which keeps Charged Off.



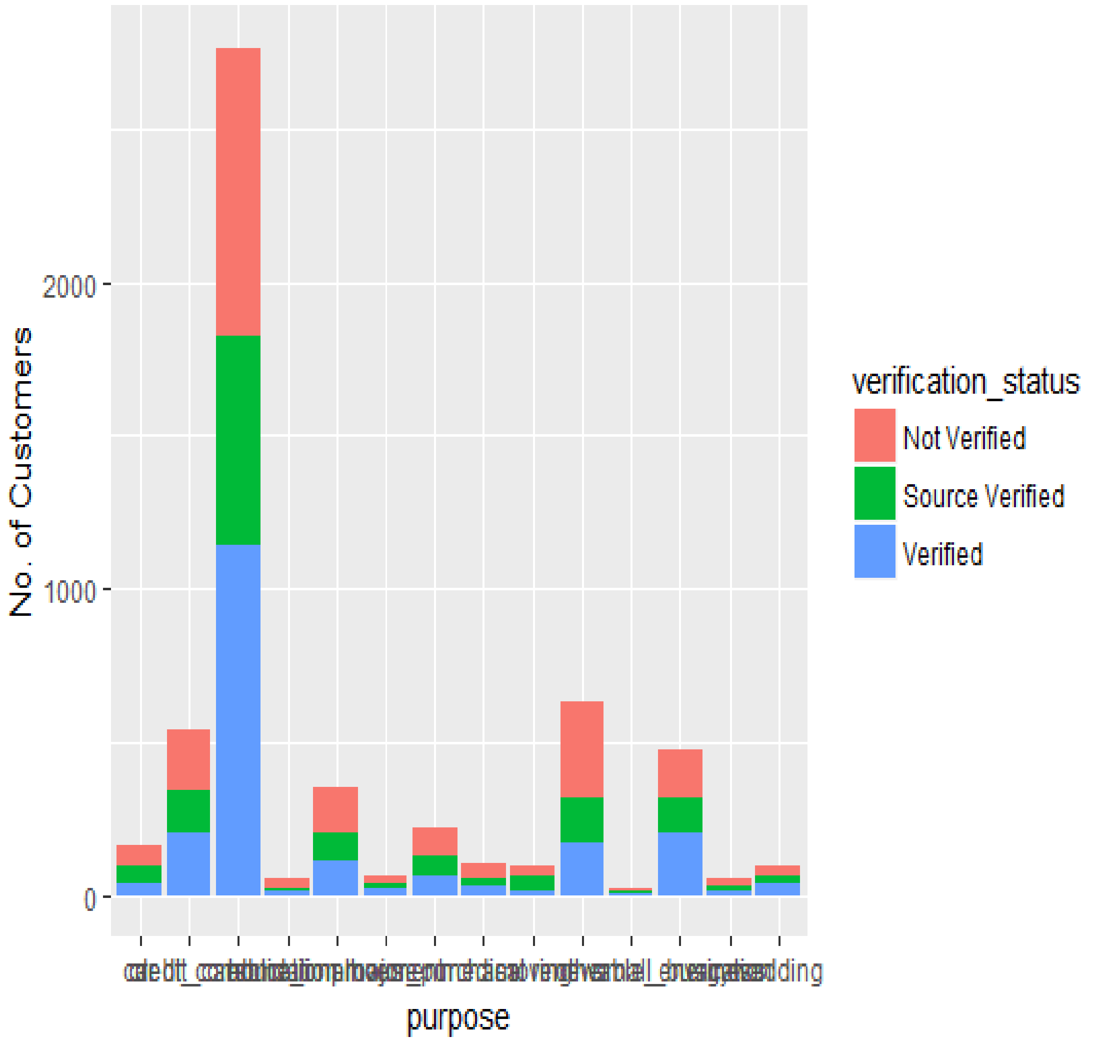
As can be seen here, most of "Charged Off" loan are for for "CA" with "RENT" and "MORTGAGE" with around 13% & 5.7%.

Bivariate Analysis

purpose_verification_status



Count of Number of Records for each Purpose. Color shows details about Verification Status. The marks are labeled by % of Total Count of Number of Records. The data is filtered on Loan Status, which keeps Charged Off.



As can be seen here, most of "Charged Off" loan are "debt_consolidation" with "VERIFIED" & "NOT VERIFIED" with around 20% & 16.7%.

Derived Metrics

We have derived few new columns/metrics as needed in the case study process:

1. We have derived a new column "ecl_year" which separates "year" from "earliest_cr_line" to analyse year wise credit line as mentioned in Slide No. 18.
2. Right now we don't see any requirement of creating any other Business or Data driven new metrics for this analyses. If needed new metrics can be derived with some constraints.

Summary of Analysis

Some of the observations are as follows:

1. From Univariate and Segmented univariate analysis we can see that, there are many variables which seems to be an indication of the possible prediction of defaulting on loan*. Some of them are 'loan_amnt','emp_length','annual_inc','dti','open_acc', 'revol_bal', 'total_acc', 'ecl_year', "home_ownership", "verification_status", "purpose", "addr_state“ and "zip_code“.*
2. *As mentioned in previous slides they do show the possible relation between the customers being on loan default and these variables.*
3. *From them, 'emp_length', "home_ownership", "verification_status", "purpose", "addr_state“, 'open_acc' and 'total_acc' affect the "charged off" state of loan most.*
4. Also from bivariate analysis we saw that, the effect of various variables on each other also. We saw the effects of:
   - "home_ownership" and "verification_status" on "loan_status“
   - "addr_state" and "home_ownership" on "loan_status“
   - "purpose" and "verification_status" on "loan_status“
   - "purpose" and "home_ownership" on "loan_status“.
5. *All the related data have been mentioned in the respective slides.*

Recommendation to lessen the loan default rate

Recommendations to lessen the loan default rate can be that the banks should refrain or give loan at higher rate to customers with *:*

1. *The customers 'emp_length' being either 0 or 10+ years.*
2. The customers having '*home_ownership*' as "RENT" or "MORTGAGED".
3. The customers having '*verification_status*' as "VERIFIED" or "NOT VERIFIED". They should be "VERIFIED"
4. The customers having *'purpose' of taking loan as "debt_consolidation" to pay off existing dues.*
5. The customers residing in *'addr_state' as "CA". They are not paying of loans either because of high expenditures in "CA".*
6. The customers having no. of **'open_***acc' in range of 5 to 17. They might be trying to get money from different sources which indicates instablilty.*
7. The customers having *'total_acc' in range of 6 to 23. Again this shows the instability.*