GRAMENER CASE STUDY

SUBMISSION: 1-JULY-2018

Team:

- **Manjula Raman**
- **Srinivas Panganamala**
- **Greenu Sharma**
- **Vishnu Das**

PROBLEM STATEMENT

The “company” is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

The “company” wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

DATA SOURCING

“loan.csv” - It contains the complete loan data for all loans issued through the time period 2007 t0 2011.

“Data_Dictionary.xlsx : The data dictionary which describes the meaning of these variables

DATA CLEANING

Fixing Rows & Columns

❑ 1. **Removed columns with only one values and NA**

1. mths_since_last_major_derog (NA)
2. policy_code(1)
3. application_type(INDIVIDUAL)
4. annual_inc_joint (NA)
5. dti_joint (NA)
6. verification_status_joint(NA)
7. acc_now_delinq (0)
8. tot_coll_amt(NA)
9. tot_cur_bal(NA)
10.open_acc_6m(NA)
11.open_il_6m (NA)
12.open_il_12m (NA)
13.open_il_24m (NA)
14.mths_since_rcnt_il (NA)
15.total_bal_il il_util (NA)
16.open_rv_12m (NA)
17.open_rv_24m (NA)
18.max_bal_bc all_util (NA)
19.total_rev_hi_lim (NA)
20.inq_fi total_cu_tl (NA)
21. inq_last_12m (NA)
22.acc_open_past_24mths (NA)
23.avg_cur_bal (NA)
24.bc_open_to_buy(NA)
25. bc_util (NA)
26. delinq_amnt(0)
27. mo_sin_old_il_acct(NA)
28. mo_sin_old_rev_tl_op (NA)
29. mo_sin_rcnt_rev_tl_op (NA)
30. mo_sin_rcnt_tl mort_acc (NA)
31. mths_since_recent_bc (NA)
32. mths_since_recent_bc_dlq (NA)
33. mths_since_recent_inq (NA)
34. mths_since_recent_revol_delin (NA)
35. q num_accts_ever_120_pd (NA)
36. num_actv_bc_tl (NA)
37. num_actv_rev_tl num_bc_sats (NA)
38. num_bc_tl num_il_tl (NA)
39. num_op_rev_tl num_rev_accts (NA)
40. num_rev_tl_bal_gt_0 (NA)
41. num_sats num_tl_120dpd_2m (NA)
42. num_tl_30dpd (NA)
43. num_tl_90g_dpd_24m (NA)
44. num_tl_op_past_12m (NA)
45. pct_tl_nvr_dlq (NA)
46. percent_bc_gt_75 (NA)
47. tot_hi_cred_lim(NA)
48. total_bal_ex_mort (NA)
49. total_bc_limit (NA)
50. total_il_high_credit_limit(NA)
51. pymnt_plan(n)
52. initial_list_status(f)

DATA CLEANING

Fixing Rows & Columns

□ 2. **Removed columns with more than 90% one values(0) and rest NA**

53. tax_liens(0,NA)
54. chargeoff_within_12_mths(0,NA)
55. collections_12_mths_ex_med(0,NA)

□ 3. **Removed columns that are auto generated (Id’s) & duplicated**

56. id

□ 4. **Removed columns that are not relevant for the analysis**

57. url
58. Desc
59. Zip_code

□ 5. **Removed columns that are duplicate or redundant**

60. int_rate
61. sub_grade
62. Funded_amnt

DATA CLEANING

Fixing Rows & Columns

❑ 6. **Removed columns with calculated (aggregate, summary) or not unique for analysis**

63. Installment
64. funded_amt
65. total_pymnt
66. total_pymnt_inv
67. total_rec_prncp
68. total_rec_int
69. total_rec_late_fee
70. recoveries
71. collection_recovery_fee
72. last_pymnt_amnt
73. next_pymnt_d
74. out_prncp
75. out_prncp_inv
76. last_credit_pull_d
77. title

DATA CLEANING

Fixing & standardising Values

1. **emp_length** – 0 means less than 1 year 10 means more than 10 years. This need to be formatted and converted as numeric stripping out the string portions . <1 year is set to 0 for analysis purpose and n/a values(6 records) are set to 0
2. **Title and purpose** fields are to describe loan need. Out of which purpose seems to be more categorical and unique. Hence eliminating title field from analysis and using purpose instead
3. **Emp_title** : This filed is the employer name as the data present is before 2013. Blanks records are present and is replaced with "UNKNOWN"
4. **Loan_amnt** : Set to numeric
5. **Annual_inc** : set to numeric and 2 decimal points
6. **Term**: Formatted to numeric for analysis purpose and made factor
7. **Grade** : Is made factor variable and later converted to numeric for correlation metrics
8. **Verification_Status**: Made as factor variable
9. **Revol_util**: 50 blank values set to 0 for analysis purpose and percentage symbol removed and made numeric
10. **pub_rec_bankruptcies**: 697 NA values set to 0 for correlation and analysis purpose
11. **home_ownership**: made as factor variable
12. **delinq_2yrs**: Made as Factor variable
13. **inq_last_6mths** : Made as Factor variable
14. **addr_state**: Made as Factor variable
15. **Loan_Status**: Factor variable

DATA CLEANING

Filtering Values

- ❑ **Loan_status : As we are interested in finding the trend in defaulters, the data is filtered with Charged off and Fully Paid customers**

DATA ANALYSIS

Selected Attributes

1. Customer Attributes

1. Member_id
2. Emp_title
3. Emp_length
4. Home_ownership
5. Annual_inc
6. Purpose
7. Addr_state
8. Dti
9. Delinq_2yrs
10. Earliest_cr_line
11. Inq_last_6mths
12. Mths_since_last_delinq
13. Mths_since_last_record
14. Open_cc
15. Pub_rec
16. Revol_bal
17. Revol_util
18. Total_acc
19. Pub_rec_bankruptcies

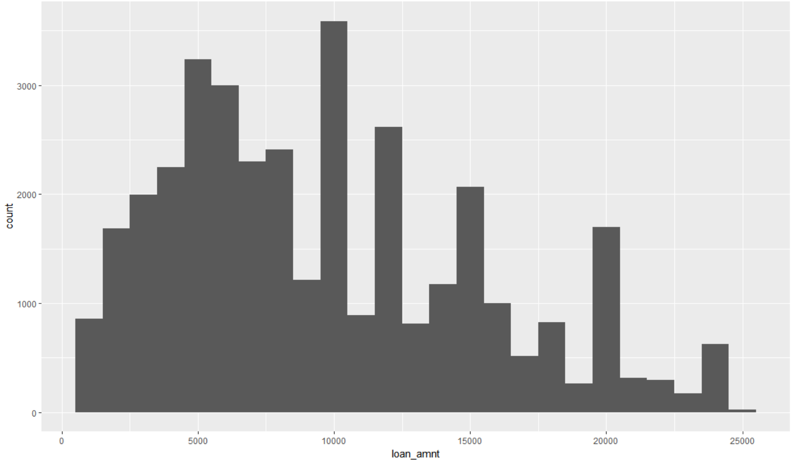
2. Loan Attributes

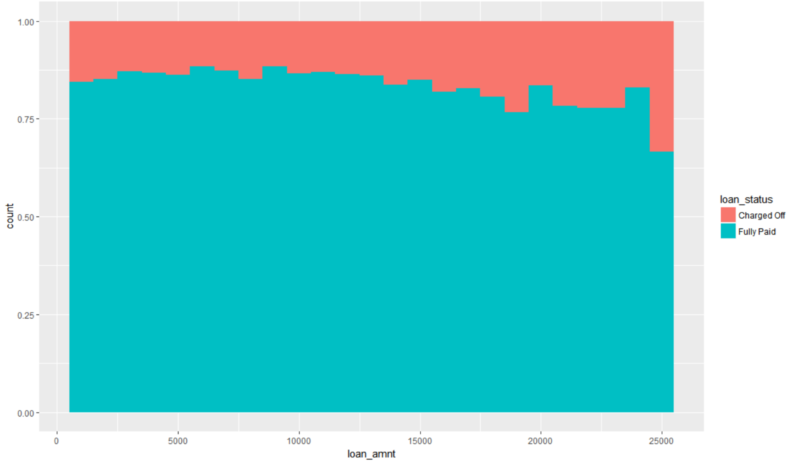
1. Loan_amnt
2. Term
3. installment
4. Grade
5. Verification_status
6. Loan_status

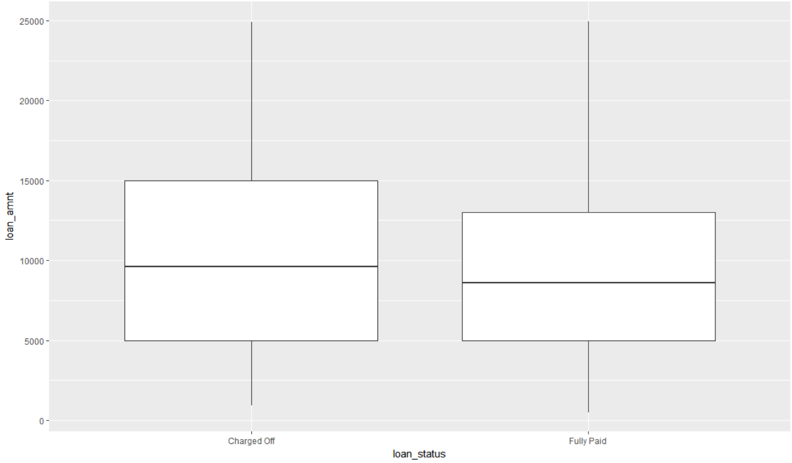
DATA ANALYSIS

Univariate

Loan Status vs Loan Amount





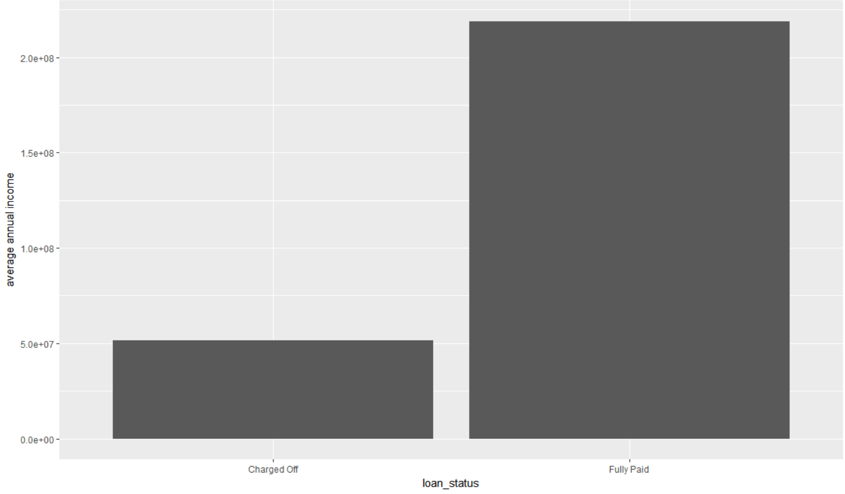


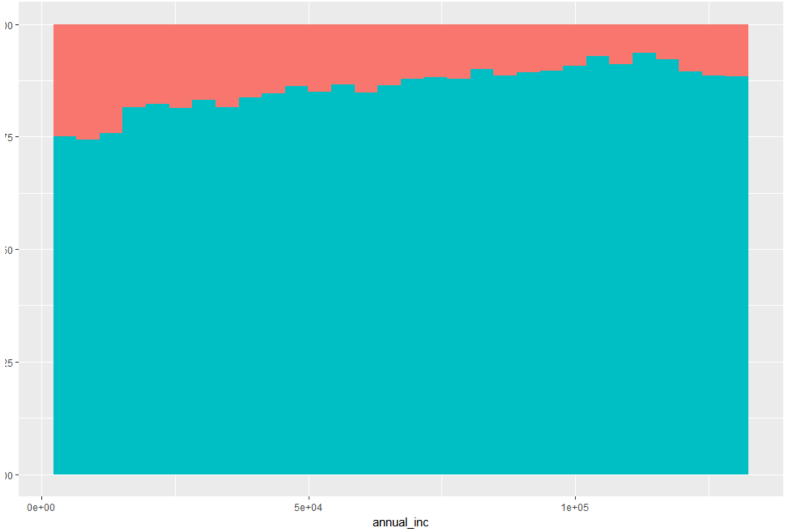
Loan frequency plot shows that most loans are taken for 5000,10000(peak),12000,15000,20000 and 25000 values
Box plot shows that the 75 percentile of charged off loans (~15000) is higher than the fully paid loans(~13000) indicating that **higher amount loans have a higher % charged off. The same is supported by the Fill plot**

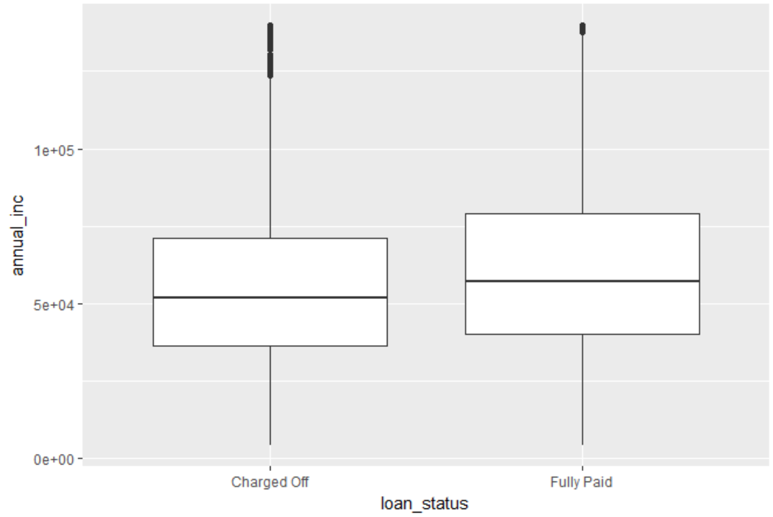
DATA ANALYSIS

Univariate

Loan Status vs Annual Income







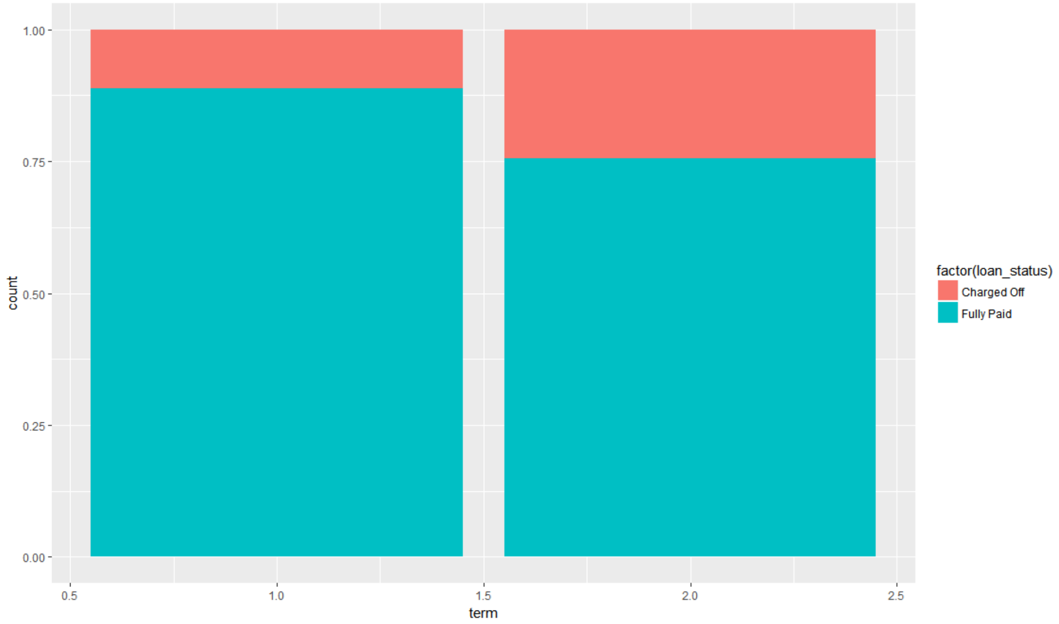
As Annual income increases, % of loans Fully paid increases from 75% to 87%

For Lower Annual Income the trend to get charged off is high and decreases with annual income

DATA ANALYSIS

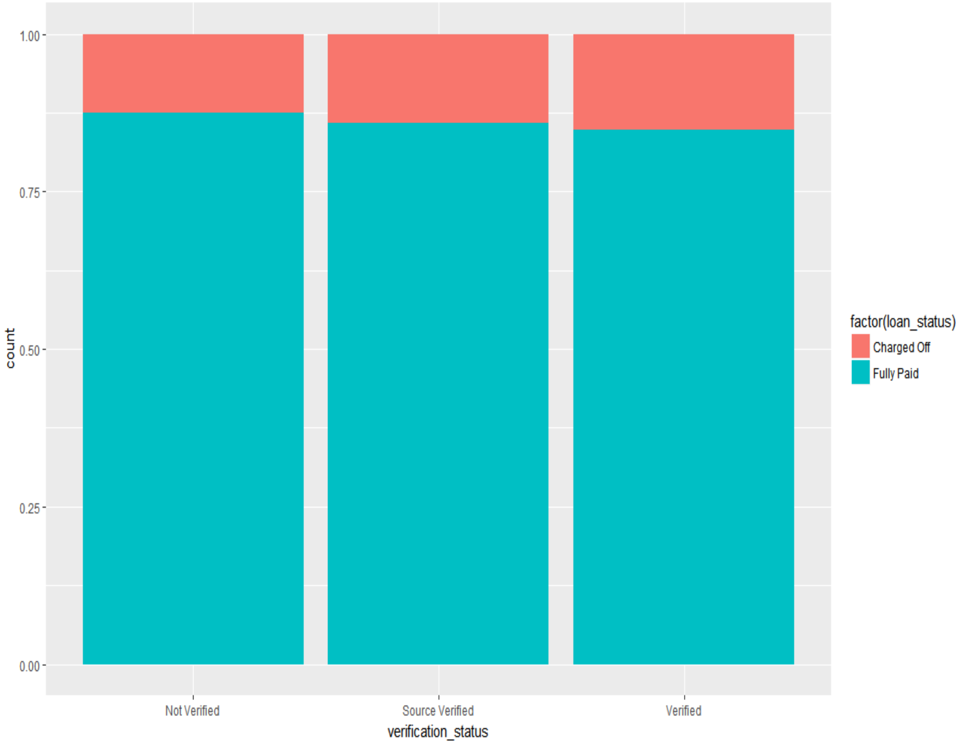
Univariate

Loan Status vs Term



87% loans get fully paid for 36 months loan
88% loans get fully paid for 36 months term
75% loans get fully paid for 60 months term
Shorter term loans have higher % fully paid and Longer term have higher % of charged off

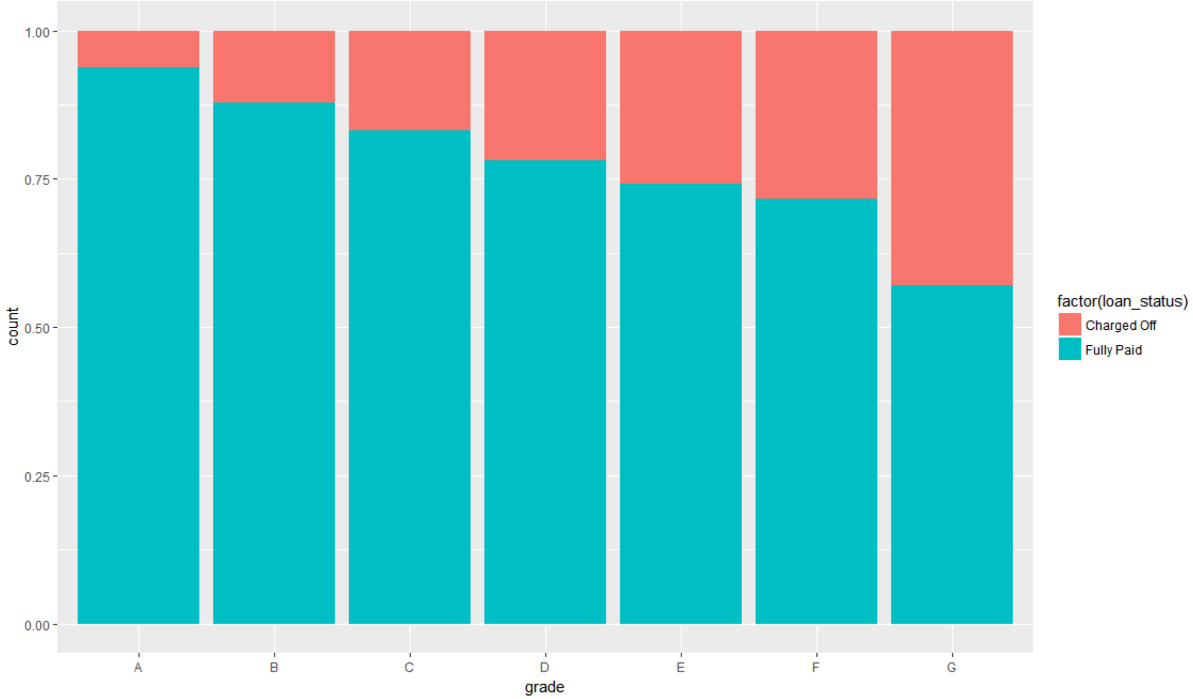
Loan Status vs Verification Status



The Verification status indicates whether the declared annual income, the income source has been verified or not.

There is no significant effect of verification on the % of loans being fully paid and charged off

Loan Status vs Grade

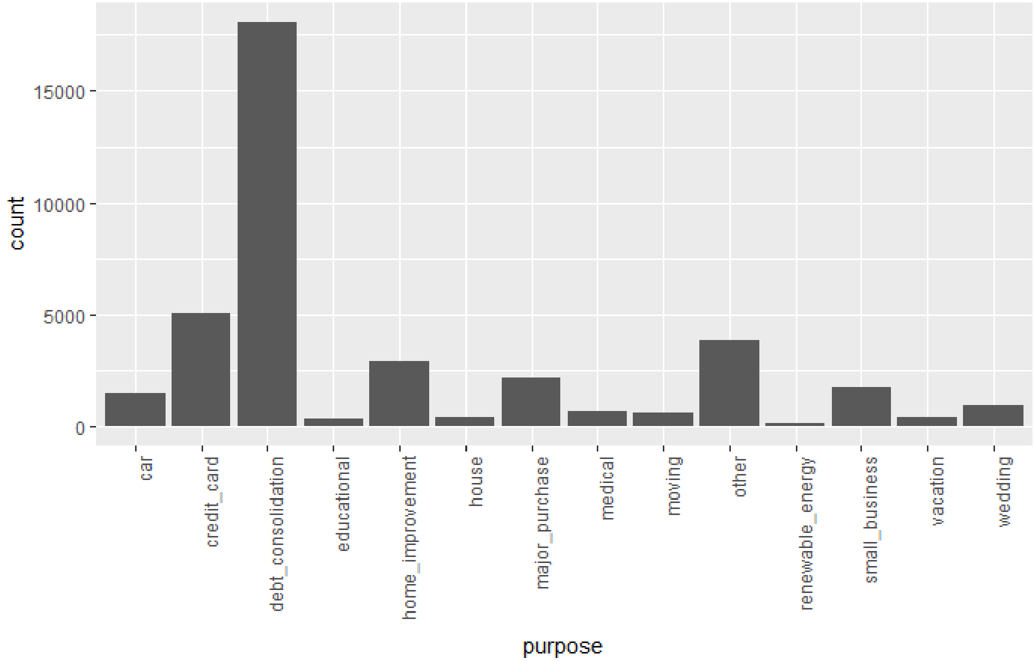


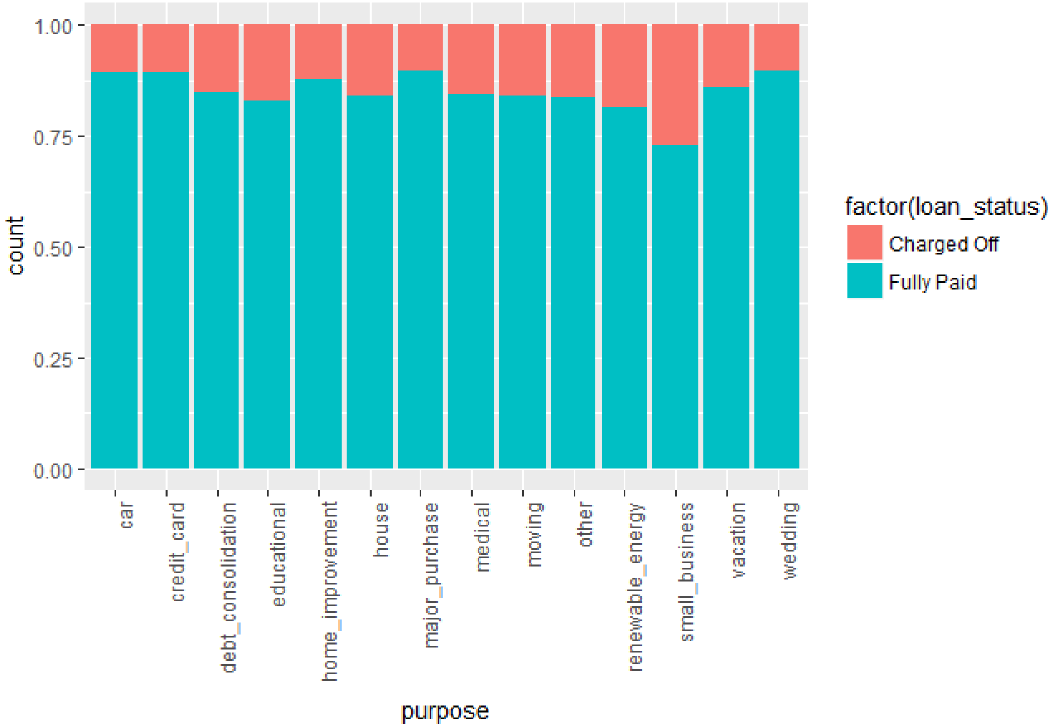
As the interest rate increases, the % of charged off loans increase. ‘A’ grade loans are 90% fully paid, ‘G’ grade loans are 67.5% fully paid. **Lower interest rate (grade) loans likely to be fully paid and higher interest rate (grade) loans have increased chances of charged off**

DATA ANALYSIS

Univariate

Loan Status vs Purpose





Majority of the loans are for **debt consolidation**, **credit card**, **home improvement (Other not considered)**

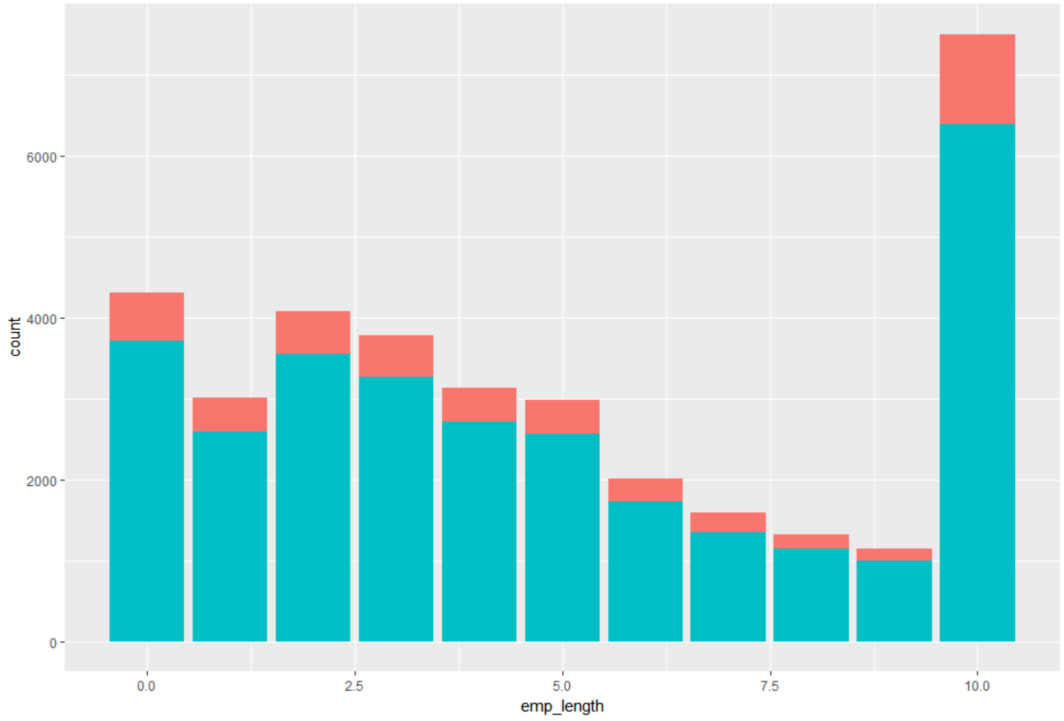
Loans that are for **small business** and **renewable energy** have the highest % of charged off loans 20-30%

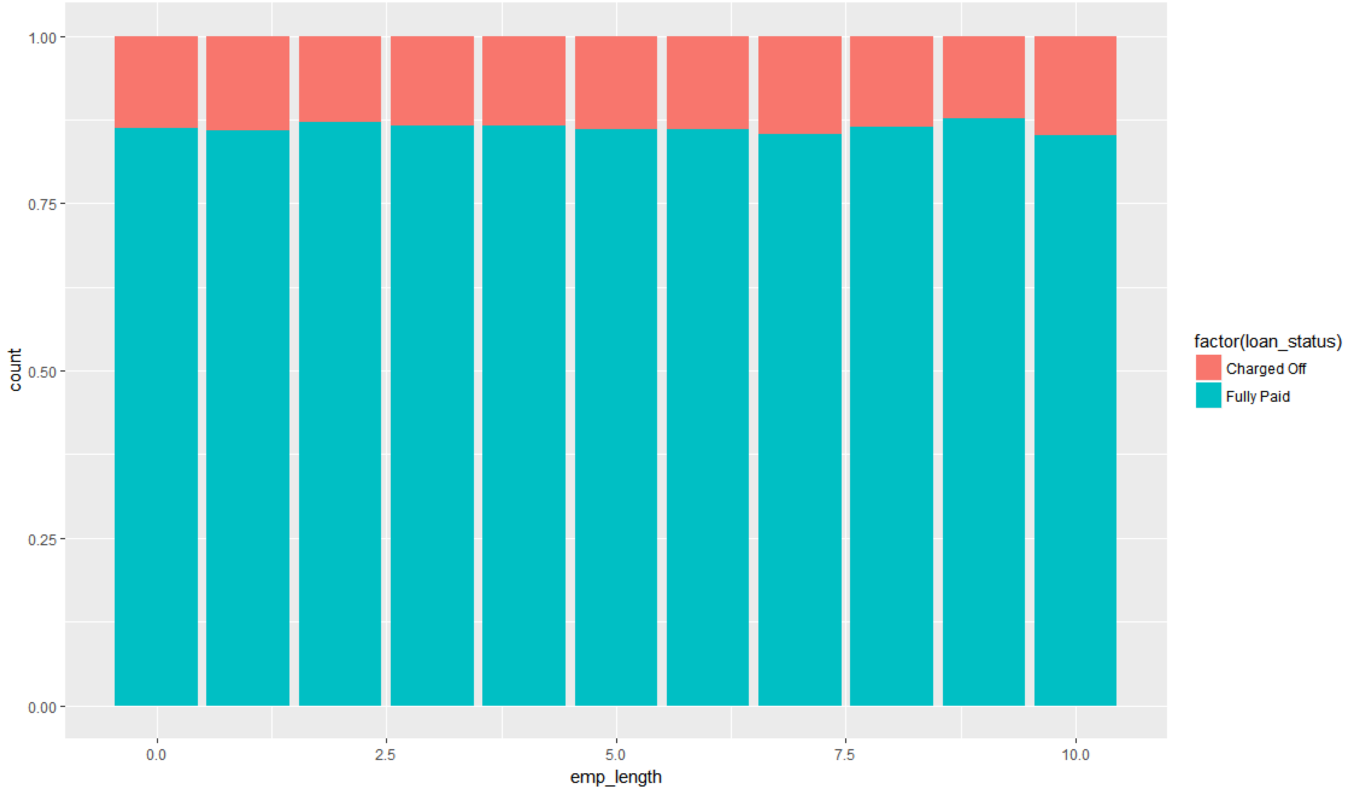
Loans for **wedding, car, credit card and major purchase** have 85% fully paid loans

DATA ANALYSIS

Univariate

Loan Status vs Employee Years of Experience

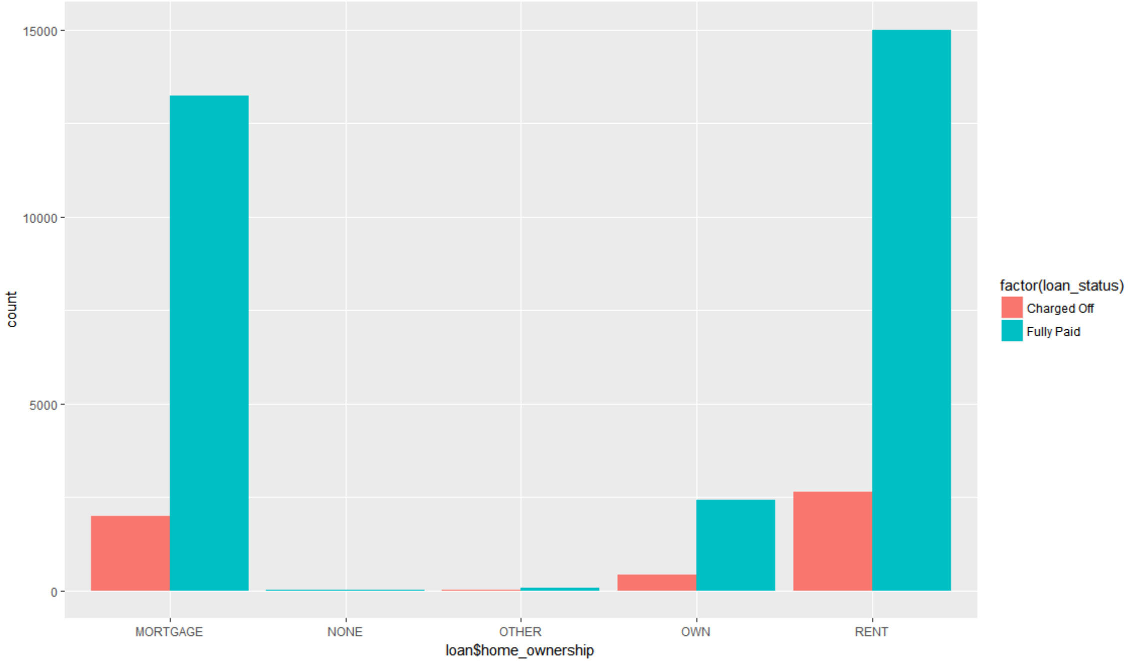


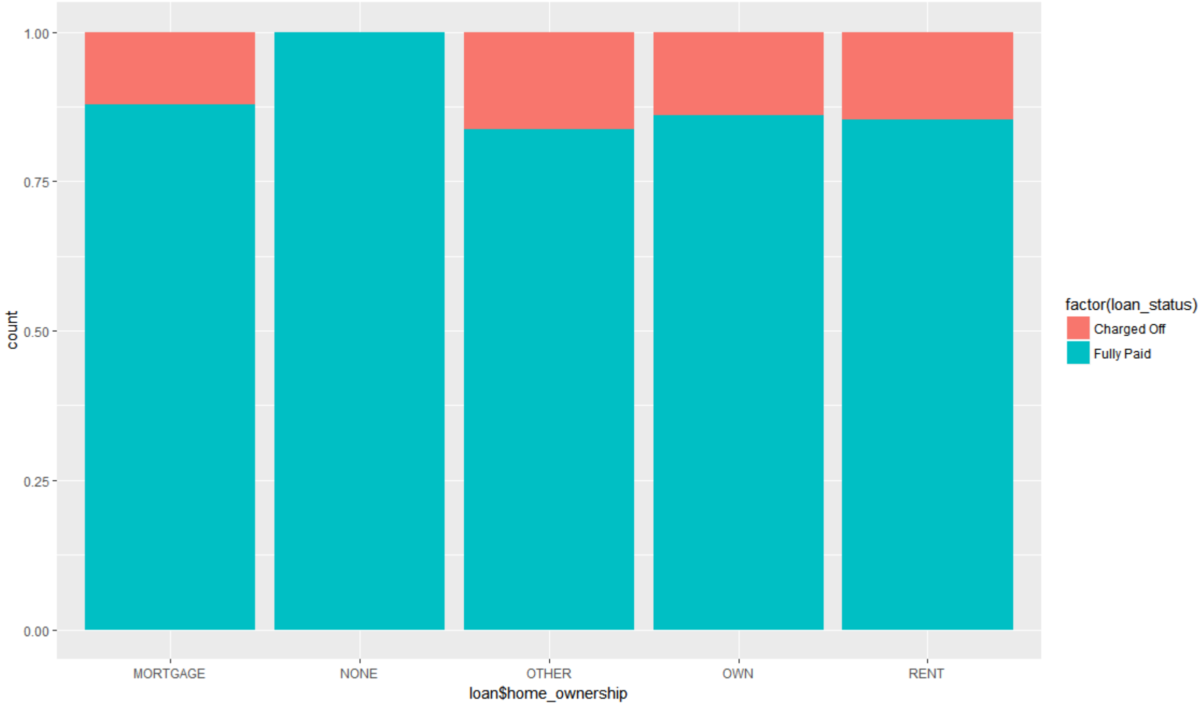


Most loans are taken at 10+ years of employment and < 1year, then 2 and 3 years

No significant trend found for Employment length on loan status

Loan Status vs Home Ownership



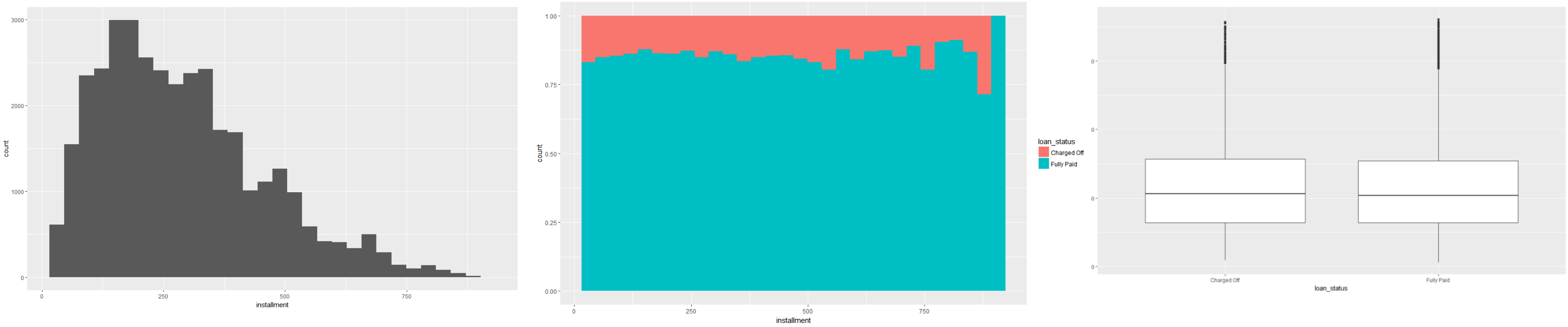


Most loans are taken by people on Rent and Mortgage. **No significant trend found on home ownership for loan status**

DATA ANALYSIS

Univariate

Loan Status vs Installment

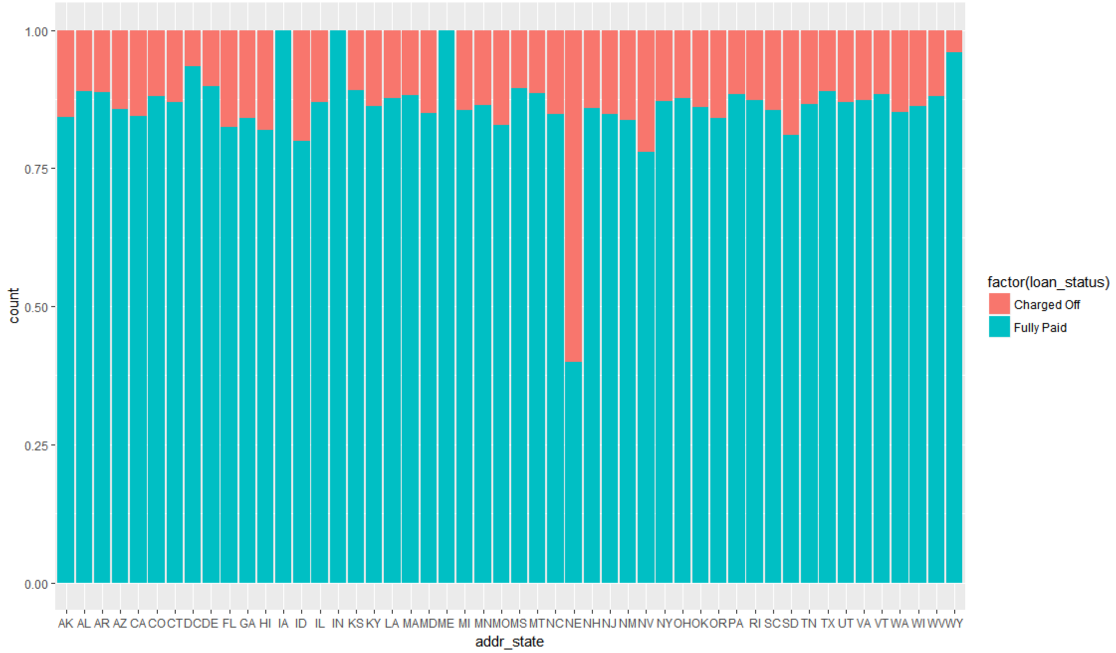


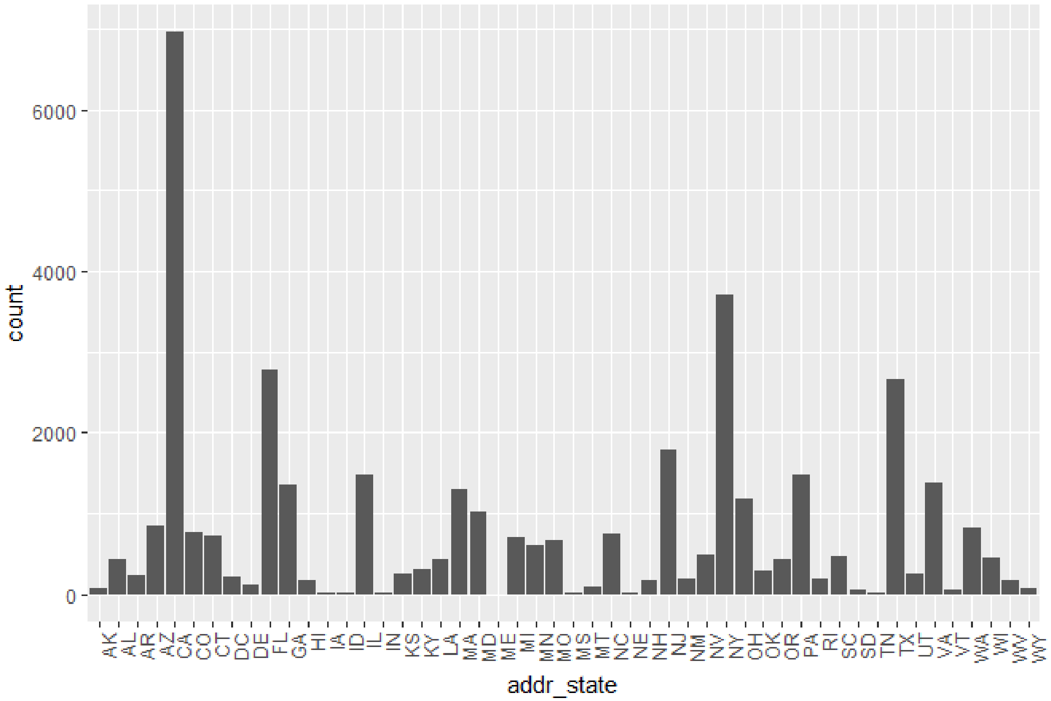
Instalment has a trend similar to loan amount. The 75% percentile of charged off loans is **slightly** higher than the Fully Paid loans , indicating that loans with **higher instalments have more % charged off than loans with lower instalments**

DATA ANALYSIS

Univariate

Loan Status vs State





observations:

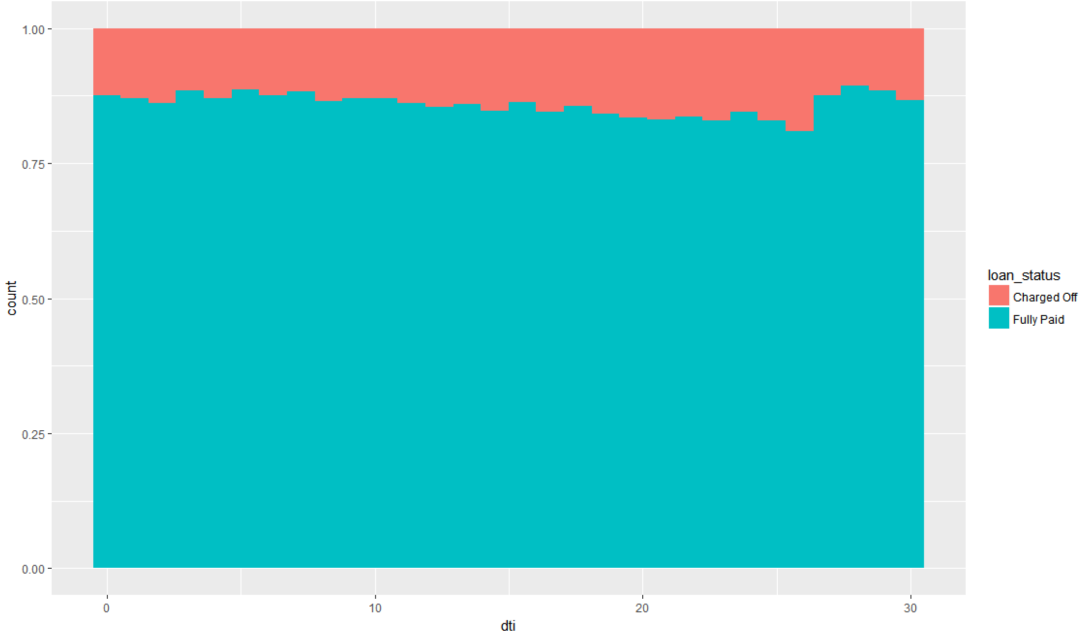
States MI,IN,IA have 100% fully paid off loans

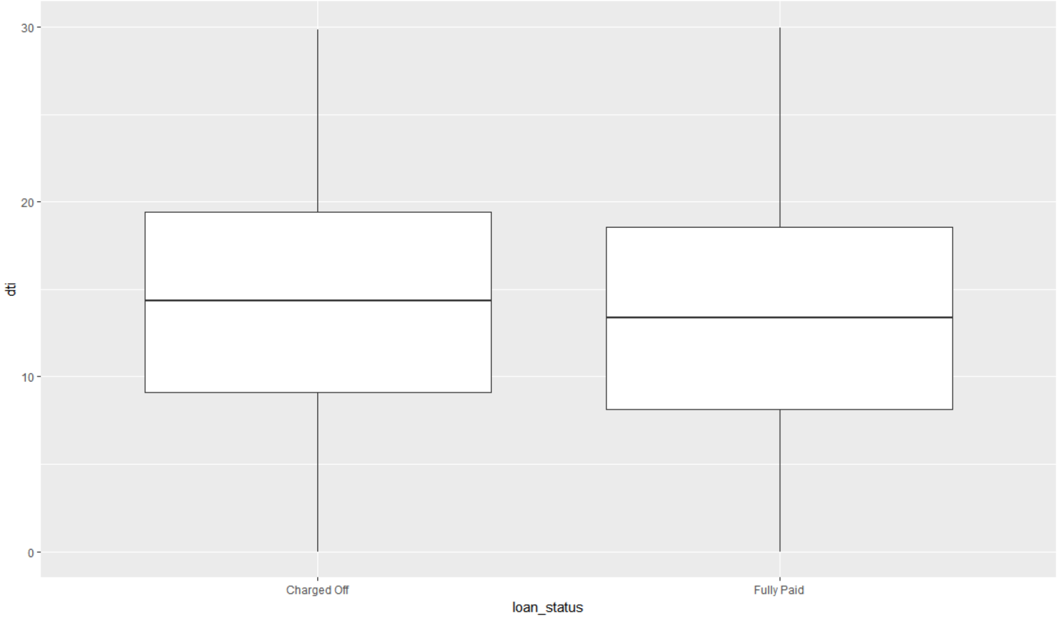
State NE (4 loans) has 60% charged off and second largest is NV (399 loans) 20% loans charged off

DATA ANALYSIS

Univariate

Loan Status vs Dti



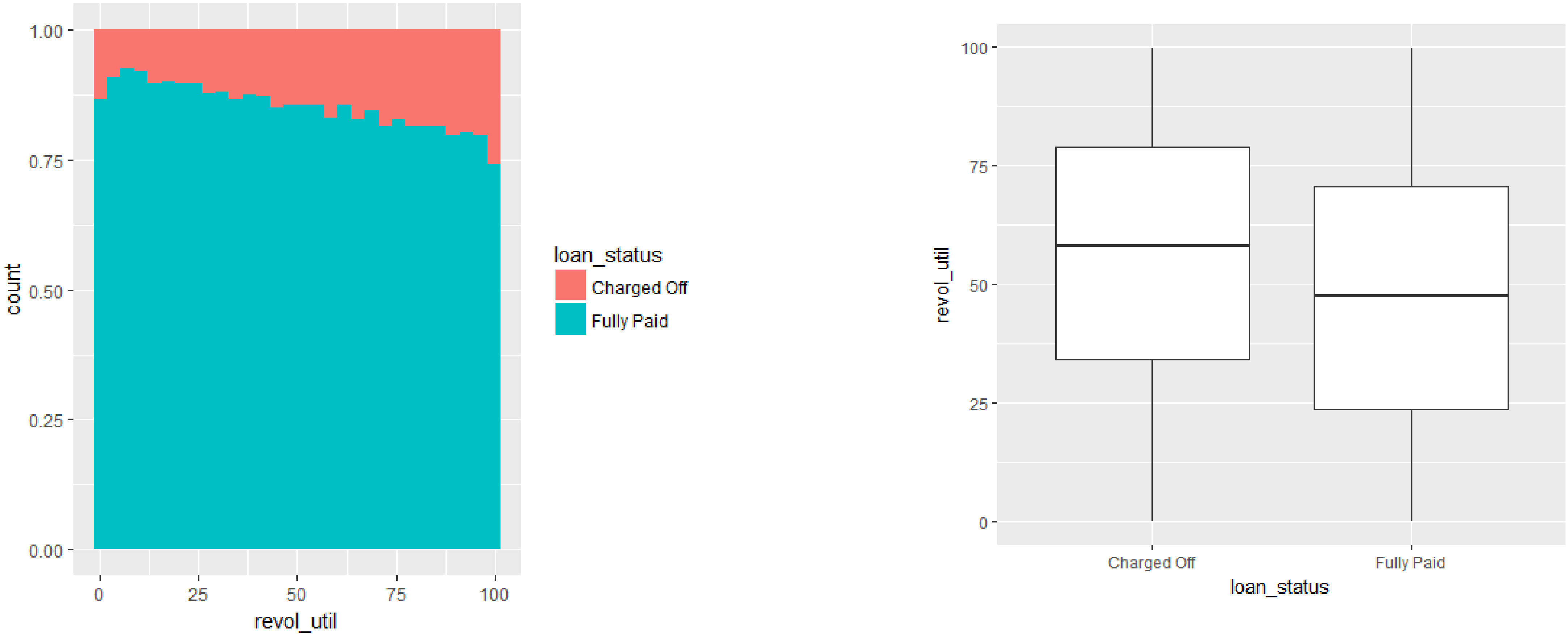


The box statistics shows that the DTI median,upper and lower quartile for fully paid loans are lower than the fully paid loans indicating **Lower dti means better loan repayment.A slight increasing tendency of charged off when the dti goes up.**

DATA ANALYSIS

Univariate

Loan Status vs Revolving Utilization

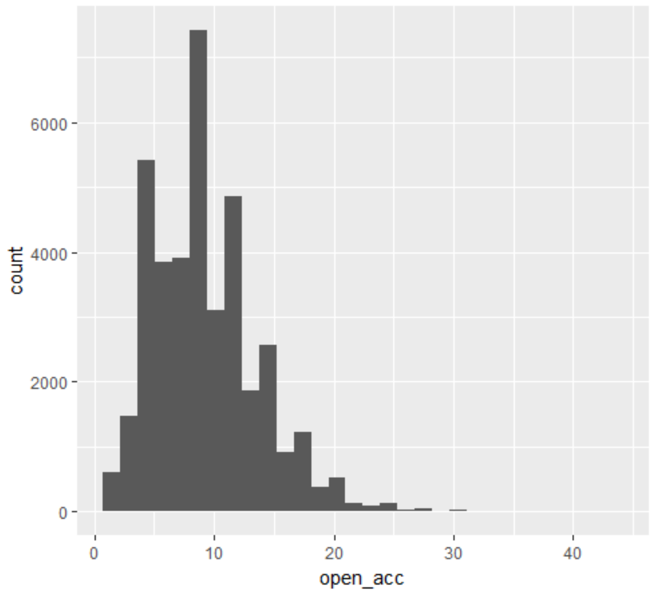


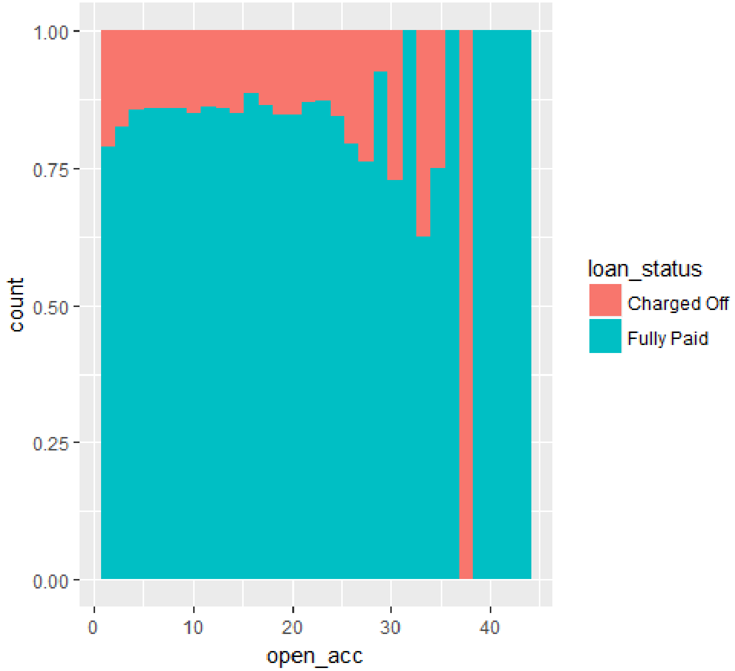
Higher revolving utilization tends to more charged off

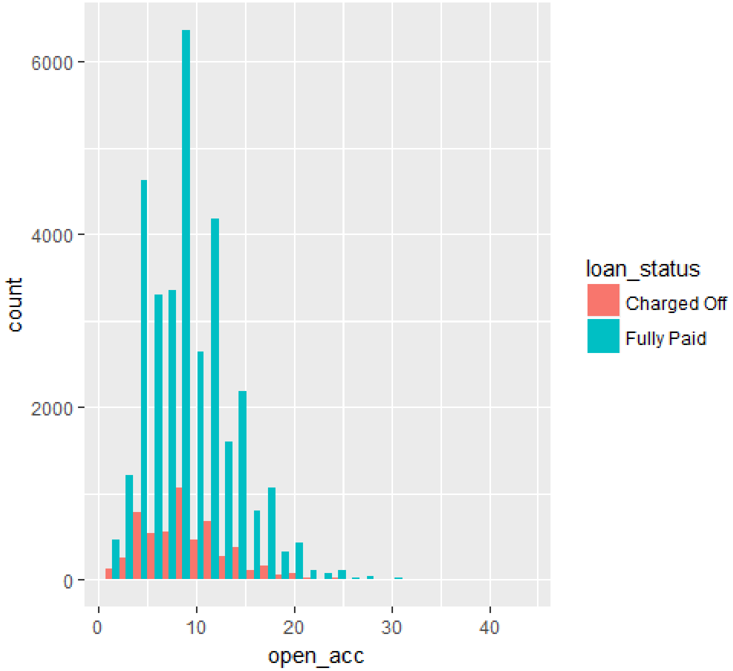
DATA ANALYSIS

Univariate

Loan Status vs Open Account





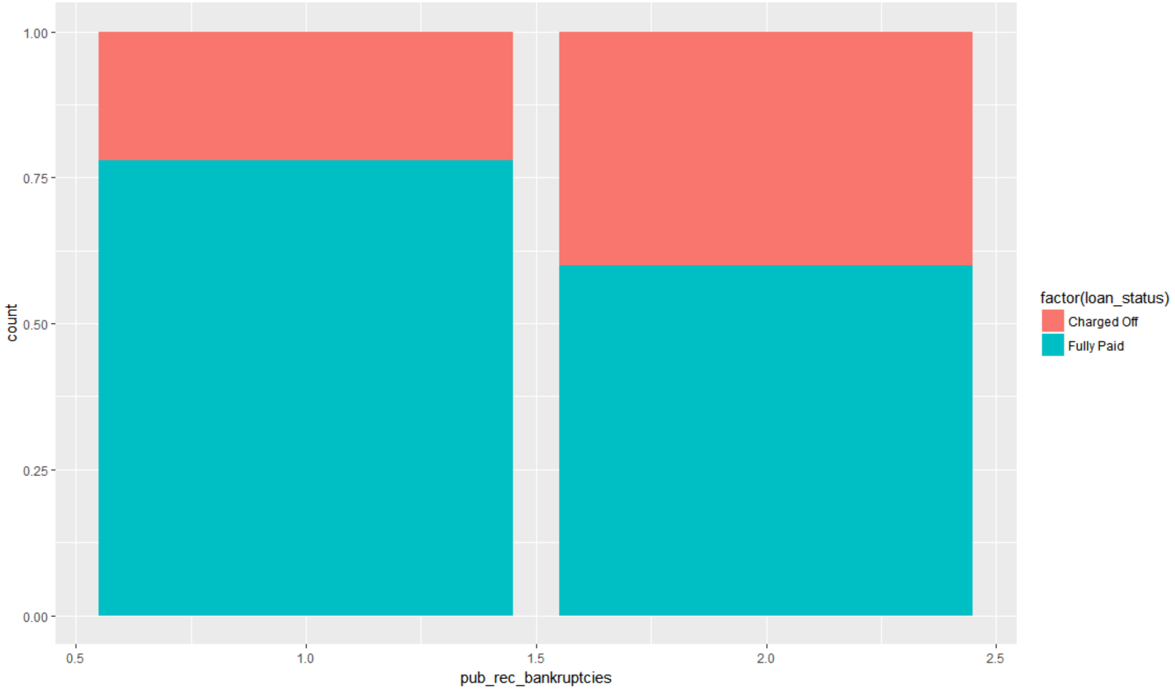


Lower credit lines leads to higher charged off . Interesting tendency. As the credit lines goes up, charged off goes down and once it crosses the median (20), the charged off goes up (credit seeking behaviour). Between 10- 20 the charged of tendency is low and below and above that range, the charged off shows increasing trend

DATA ANALYSIS

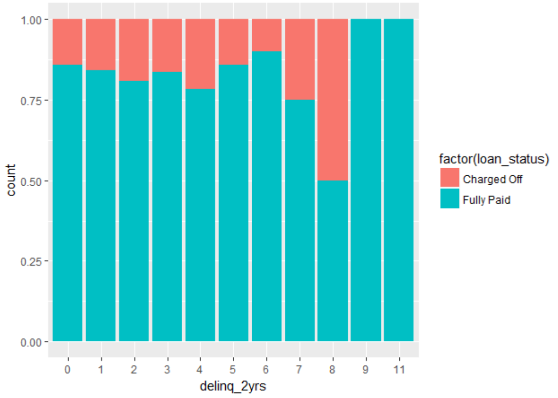
Univariate

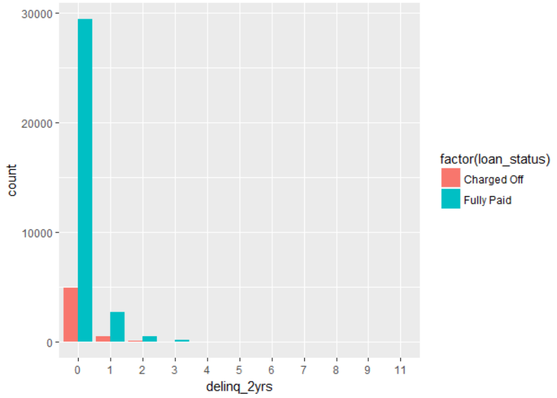
Loan Status vs Public record Bankruptcies



96% loans have 0 public recorded bankruptcies
For 1 and 2 public record bankruptcies, 22% ,35% of total loans get charged off

Loan Status vs Past due deliquency



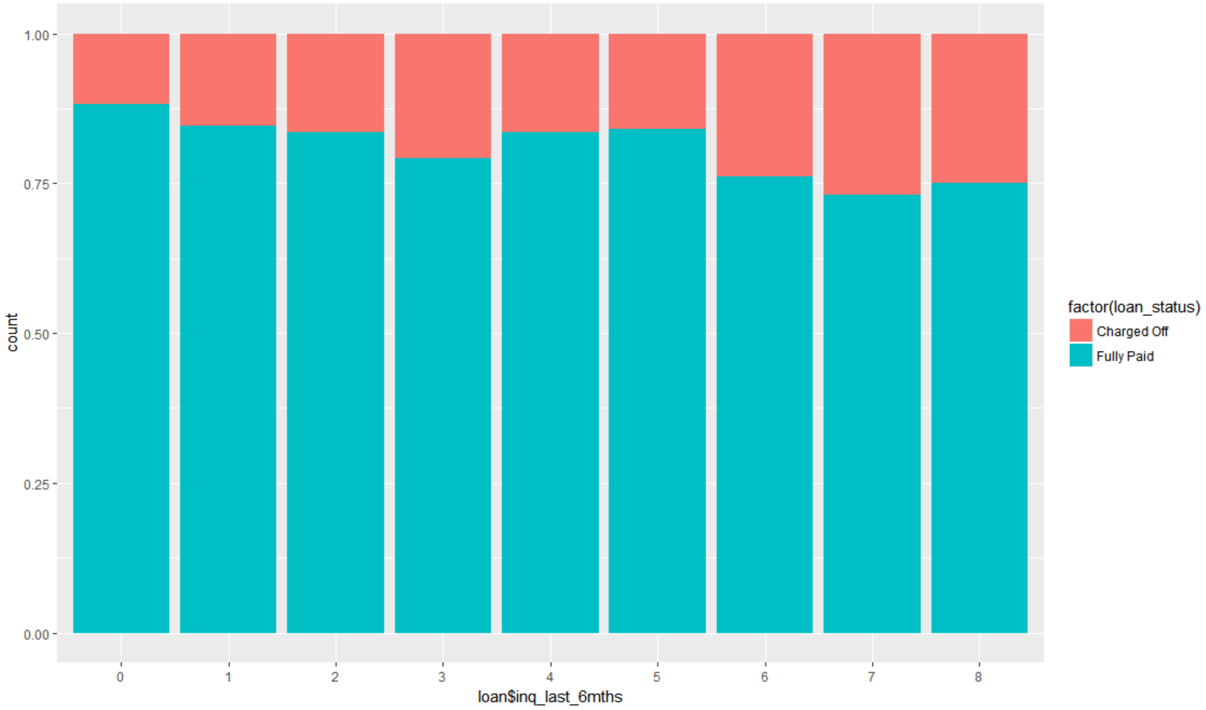


For 0 delinquencies 85% of the loans are fully paid off
As the number of 30+ past due delinquencies increased to 10, as very slight trend of increasing charged off observed. The data beyond 2 delinquencies are not so significant as the frequency is very low

DATA ANALYSIS

Univariate

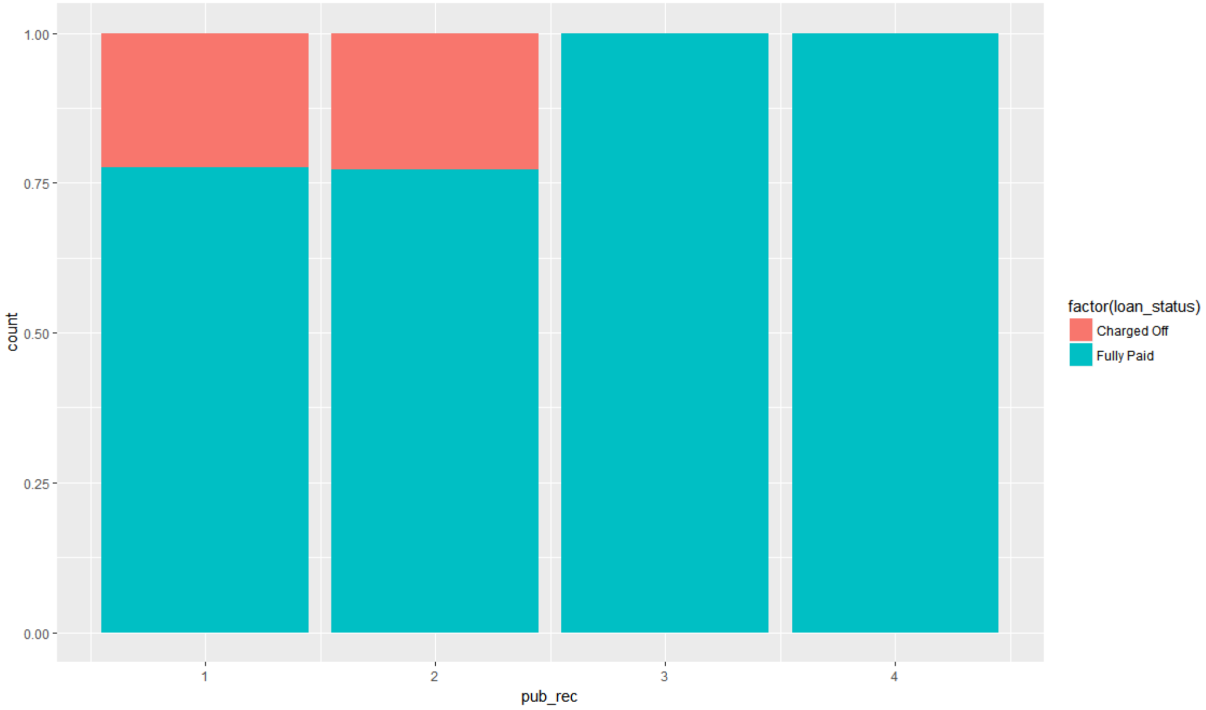
Loan Status vs Inquiries



Loans with 0 inquiries are 87% fully paid

As number of inquiries increases to 7, 75% are fully paid and 25% are charged_off

Loan Status vs Public Records



95% loans have 0 public record

25% loans with 1 and 2 public records get charged off

DATA ANALYSIS

Bivariate

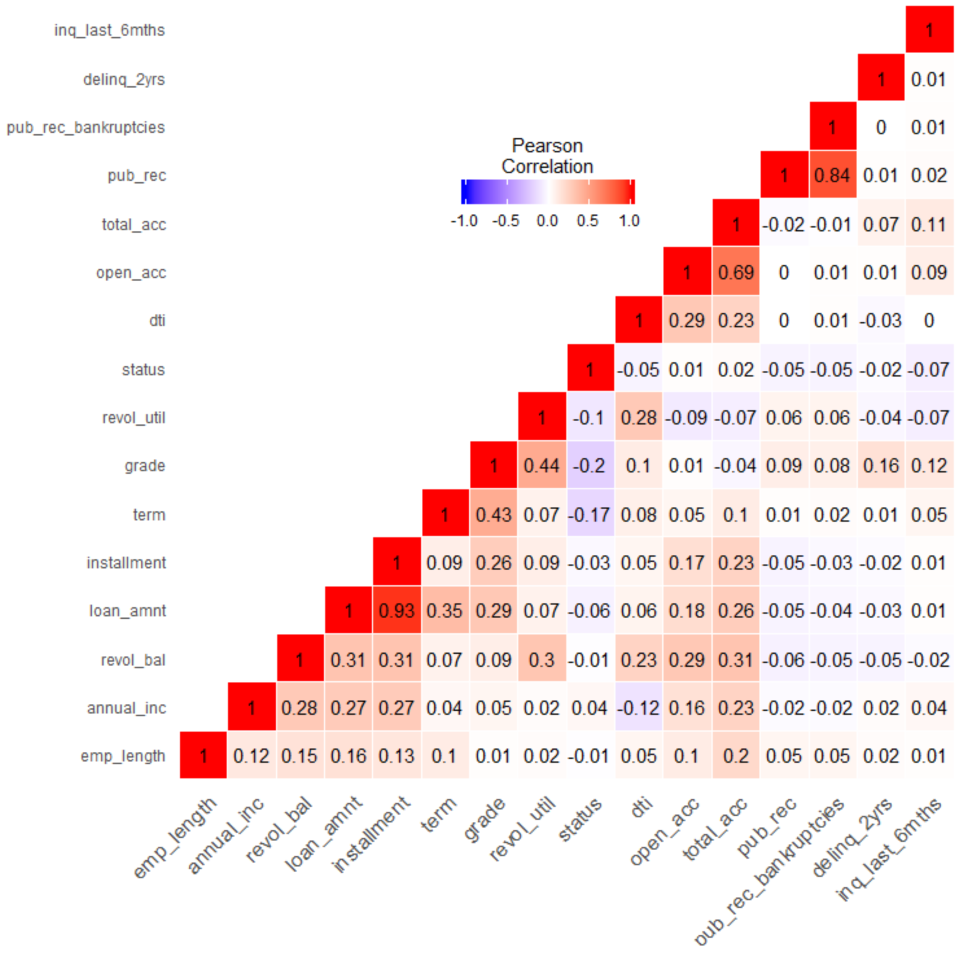
Correlation Metrics

Variables correlated positively with status

- Annual_inc 0.04, Greater income ; more likely to be fully paid

Variables correlated negatively with status

- Loan_amnt -0.06
- Instalment-0.03
- Grade -0.2
- **Term -0.17,** Higher term, more likely to be charged off
- Dti -0.05
- Revol_bal -0.01
- Revol_util -0.1
- Inq_last_6_mnths -0.07
- Public_records-0.05
- Public_recorded bankruptcies-0.05



	emp_length	annual_inc	revol_bal	loan_amnt	installment	term	grade	revol_util	status	dti	open_acc	total_acc	pub_rec	pub_rec_bankruptcies	delinq_2yrs	inq_last_6mths
inq_last_6mths																1
delinq_2yrs															1	0.01
pub_rec_bankruptcies														1	0	0.01
pub_rec													1	0.84	0.01	0.02
total_acc												1	-0.02	-0.01	0.07	0.11
open_acc											1	0.69	0	0.01	0.01	0.09
dti										1	0.29	0.23	0	0.01	-0.03	0
status									1	-0.05	0.01	0.02	-0.05	-0.05	-0.02	-0.07
revol_util								1	-0.1	0.28	-0.09	-0.07	0.06	0.06	-0.04	-0.07
grade							1	0.44	-0.2	0.1	0.01	-0.04	0.09	0.08	0.16	0.12
term						1	0.43	0.07	-0.17	0.08	0.05	0.1	0.01	0.02	0.01	0.05
installment					1	0.09	0.26	0.09	-0.03	0.05	0.17	0.23	-0.05	-0.03	-0.02	0.01
loan_amnt				1	0.93	0.35	0.29	0.07	-0.06	0.06	0.18	0.26	-0.05	-0.04	-0.03	0.01
revol_bal			1	0.31	0.31	0.07	0.09	0.3	-0.01	0.23	0.29	0.31	-0.06	-0.05	-0.05	-0.02
annual_inc		1	0.28	0.27	0.27	0.04	0.05	0.02	0.04	-0.12	0.16	0.23	-0.02	-0.02	0.02	0.04
emp_length	1	0.12	0.15	0.16	0.13	0.1	0.01	0.02	-0.01	0.05	0.1	0.2	0.05	0.05	0.02	0.01

DERIVED METRICS

Type-driven metrics

- **Loan status was categorical for plots and made numeric for correlation metric**
- **Term was categorical for plots and made numeric for correlation metric**
- **Grade treated as Ordinal**
- **Home Ownership was made as Nominal**
- **Verification status was made Nominal**
- **DTI was Ordinal ratio variable**

Business-driven metrics

- **Open Account, Bankruptcies, public rec , delinquencies were used for understanding the credit rank of the borrower**

Data-driven metrics

- **Grade was substituted for Interest Rate and Sub grade for our analysis purpose**
- **Open Account was considered more than total account**
- **Purpose was considered upon title for understanding loan details**

SUMMARY OF OBSERVATION

- ✓ The key drivers for loan default based on analysis are:
  1. A High Loan Amount
  2. Longer term loans (60 months)
  3. High Interest Rates are charged off (Loans Grade E, F and G)
  4. Higher Instalments are charged off
  5. Loans for purpose of small business and renewable energy are charged off
  6. For lower annual incomes loans shows higher trend to get charged off.
  7. Loans for the state of NE are 60% charged off;for NV are charged off 20%
  8. High DTI shows trend to get charged off
  9. High Revolving utilization shows trend for charged off.
  10. Inquiries 0-7 shows an increasing trend of charged off.
  11. For one or more public record loans (0-2) slight increasing trend of charged off

KEY DRIVER ATTRIBUTES

1. Customer Attributes

1. Member_id
2. Emp_title
3. Emp_length
4. Home_ownership
5. Annual_inc
6. Purpose
7. Addr_state
8. Dti
9. Delinq_2yrs
10. Earliest_cr_line
11. Inq_last_6mths
12. Mths_since_last_delinq
13. Mths_since_last_record
14. Open_cc
15. Pub_rec
16. Revol_bal
17. Revol_util
18. Total_acc
19. Pub_rec_bankruptcies

2. Loan Attributes

1. Loan_amnt
2. Term
3. installment
4. Grade
5. Verification_status
6. Loan_status

From our analysis , the key attributes which can help to make decision on loan approval is highlighted red from both customer and loan attributes lists available for analysis with proper data
Note: There were other critical fields which were present in the dictionary (eg: FICO range) which was not in the provided data csv file