

# Loans Case Study

Submission by –

Maxim Rohit,

Amiyanshu Pratihari

Abhishek Ranjan

Karthikeyan Seetharaman





01 Executive Summary

02 Approach

03 Analysis and Inferences

04 Summary

05 Appendix

## Key Objectives

- To understand the **driving factors (or driver variables)** behind loan defaults.
- Utilize this knowledge for its portfolio and risk assessment to **minimize credit loss and business loss.**

## Business Understanding

- Consumer finance company- largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.
- 2 risks associated with banks decision to approve loans.
  - 1. loss of likely to repay the loan, then not approving the loan results in a loss of business to the company
  - 2. not likely to repay the loan, then approving the loan may lead to a financial loss for the company

## Approach

- Structured approach as prescribed iiit-b and S Anand

Analysis	Brief
<ul style="list-style-type: none"> <li>✓ Data cleaning</li> <li>✓ Univariate analysis</li> <li>✓ Bivariate analysis</li> </ul>	<ul style="list-style-type: none"> <li>✓ Segment Analysis</li> <li>✓ Derived metrics analysis.</li> <li>✓ <b>Correlation analysis</b></li> </ul>
Univariate analysis	<ul style="list-style-type: none"> <li>✓ Categorical variables are analyzed – group, sub group, loan status, state, verification status etc. Bar plots used</li> <li>✓ Continuous variables are analyzed – amount, DTI, interest rate, revol percent, etc. box plots and histograms used.</li> </ul>
Bivariate and segment analysis	<ul style="list-style-type: none"> <li>✓ Categorical variable are plotted against other categorical variables to gain more information on the composition of the categorical data.</li> <li>✓ Ex – state vs loan status, purpose vs loan status etc.</li> <li>✓ Continuous variables are plotted against Categorical variables to gain more insights. Ex total fund amount by grades</li> <li>✓ Continuous variable vs continuous variables – ex dti vs revol utilization, charge off vs revol util etc. scatter plots used</li> </ul>
Derived metrics analysis	<ul style="list-style-type: none"> <li>✓ charge off amount is derived from subtracting total principals received from funded amount.</li> <li>✓ Charge off amount percentage is derived.</li> </ul>
* Univariate analysis, Bivariate analysis and respective Derived metrics analysis	

## Key Take always

- ✓ Understand consumer attributes and loan attributes influence the tendency of default.
- ✓ Recommendations on 5 important driver variables
- ✓ Get the correlation matrix as soon as possible and avoid analysis paralysis with lot of graphs.

## Deliverables

- **One zip file containing**
  - ✓ R Code
  - ✓ Presentation in PDF format

## Data cleansing

Importing and Data cleansing

- ✓ NA's analysis
- ✓ Duplicate
- ✓ Changing the class of observation.
- ✓ Formatting and standardizing date time – (issue\_d), percentage etc.
- ✓ Creating derived metrics –
  - Charge off amount
  - Charge off as a percentage of funded amount

## Analysis

Exploratory data analysis – Question the data to perform

- ✓ **Univariate\* analysis of both categorical and continuous variables.**
- ✓ **Bivariate analysis\* –**
  - Categorical vs categorical
  - Categorical vs continuous
  - Continues vs continuous
- ✓ **Correlation matrix.**

\*\*Univariate analysis, Bivariate analysis includes Derived metrics as well

## Plotting

- ✓ Using Tableau and R to create graphs that aid in
  - Defining the issues
  - Analysis (univariate, bivariate)
  - Segmentation analysis
- ✓ Communicate inferences, understanding with supporting analysis and graphs to decision making audience and any larger audience.

## Tools used

- RStudio for Import, Data cleansing, Analysis & Plotting
- Tableau for Analysis and Plotting

	0%	25%	50%	75%	100%	variable_means	no_of_NAs	not_NAs	total	no_of_NAs_percentage
id	54734.00	516221.00	665665.00	837755.00	1077501.00	683131.91	0	39717	39717	0.00
member_id	70699.00	666780.00	850812.00	1047339.00	1314167.00	850463.56	0	39717	39717	0.00
loan_amnt	500.00	5500.00	10000.00	15000.00	35000.00	11219.44	0	39717	39717	0.00
funded_amnt	500.00	5400.00	9600.00	15000.00	35000.00	10947.71	0	39717	39717	0.00
funded_amnt_inv	0.00	5000.00	8975.00	14400.00	35000.00	10397.45	0	39717	39717	0.00
installment	15.69	167.02	280.22	430.78	1305.19	324.56	0	39717	39717	0.00
annual_inc	4000.00	40404.00	59000.00	82300.00	6000000.00	68968.93	0	39717	39717	0.00
dti	0.00	8.17	13.40	18.60	29.99	13.32	0	39717	39717	0.00
delinq_2yrs	0.00	0.00	0.00	0.00	11.00	0.15	0	39717	39717	0.00
inq_last_6mths	0.00	0.00	1.00	1.00	8.00	0.87	0	39717	39717	0.00
mths_since_last_delinq	0.00	18.00	34.00	52.00	120.00	NA	25682	14035	39717	64.66
mths_since_last_record	0.00	22.00	90.00	104.00	129.00	NA	36931	2786	39717	92.99
open_acc	2.00	6.00	9.00	12.00	44.00	9.29	0	39717	39717	0.00
pub_rec	0.00	0.00	0.00	0.00	4.00	0.06	0	39717	39717	0.00
revol_bal	0.00	3703.00	8850.00	17058.00	149588.00	13382.53	0	39717	39717	0.00
total_acc	2.00	13.00	20.00	29.00	90.00	22.09	0	39717	39717	0.00
out_prncp	0.00	0.00	0.00	0.00	6311.47	51.23	0	39717	39717	0.00
out_prncp_inv	0.00	0.00	0.00	0.00	6307.37	50.99	0	39717	39717	0.00
total_pymnt	0.00	5576.93	9899.64	16534.43	58563.68	12153.60	0	39717	39717	0.00
total_pymnt_inv	0.00	5112.31	9287.15	15798.81	58563.68	11567.15	0	39717	39717	0.00
total_rec_prncp	0.00	4600.00	8000.00	13653.26	35000.02	9793.35	0	39717	39717	0.00
total_rec_int	0.00	662.18	1348.91	2833.40	23563.68	2263.66	0	39717	39717	0.00

total_rec_late_fee	0.00	0.00	0.00	0.00	180.20	1.36	0	39717	39717	0.00
recoveries	0.00	0.00	0.00	0.00	29623.35	95.22	0	39717	39717	0.00
collection_recovery_fee	0.00	0.00	0.00	0.00	7002.19	12.41	0	39717	39717	0.00
last_pymnt_amnt	0.00	218.68	546.14	3293.16	36115.20	2678.83	0	39717	39717	0.00
collections_12_mths_ex_med	0.00	0.00	0.00	0.00	0.00	NA	56	39661	39717	0.14
policy_code	1.00	1.00	1.00	1.00	1.00	1.00	0	39717	39717	0.00
acc_now_delinq	0.00	0.00	0.00	0.00	0.00	0.00	0	39717	39717	0.00
chargeoff_within_12_mths	0.00	0.00	0.00	0.00	0.00	NA	56	39661	39717	0.14
delinq_amnt	0.00	0.00	0.00	0.00	0.00	0.00	0	39717	39717	0.00
pub_rec_bankruptcies	0.00	0.00	0.00	0.00	2.00	NA	697	39020	39717	1.75
tax_liens	0.00	0.00	0.00	0.00	0.00	NA	39	39678	39717	0.10
issue_d_conv_year	2007.00	2010.00	2011.00	2011.00	2011.00	2010.33	0	39717	39717	0.00
issue_d_conv_month_num	1.00	4.00	7.00	10.00	12.00	7.17	0	39717	39717	0.00
int_rate_conv	5.42	9.25	11.86	14.59	24.59	12.02	0	39717	39717	0.00
revol_util_conv	0.00	25.40	49.30	72.40	99.90	NA	50	39667	39717	0.13
derieved_chargedoff_amnt	-0.25	0.00	0.00	0.00	35000.00	1154.36	0	39717	39717	0.00
derieved_chargedoff_per	0.00	0.00	0.00	0.00	100.00	9.43	0	39717	39717	0.00
term_conv	36.00	36.00	36.00	60.00	60.00	42.42	0	39717	39717	0.00
emp_length_conv	1.00	2.00	4.00	9.00	10.00	NA	1075	38642	39717	2.71

- Charged of amount (for charged of account) = funded amount – total principle received
- Charged off percentage = charged off amount divided by funded amount \* 100

After dropping some empty and categorical variables we are left with 41 continuous variables in total.

- ✓ **Key variable** off interest for analysis – Funded amount, annual income, debt to income ratio, total received principal, interest rate, employment length,
- ✓ **Derived fields** –  
Charged of amount  
Charged off percentage

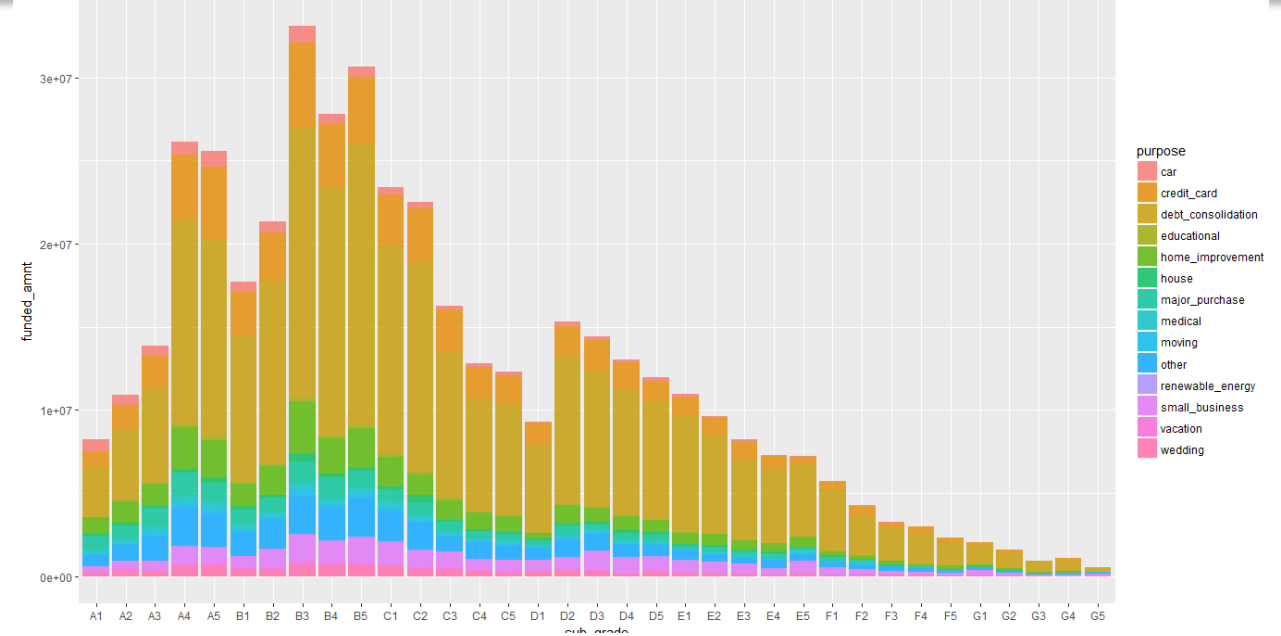
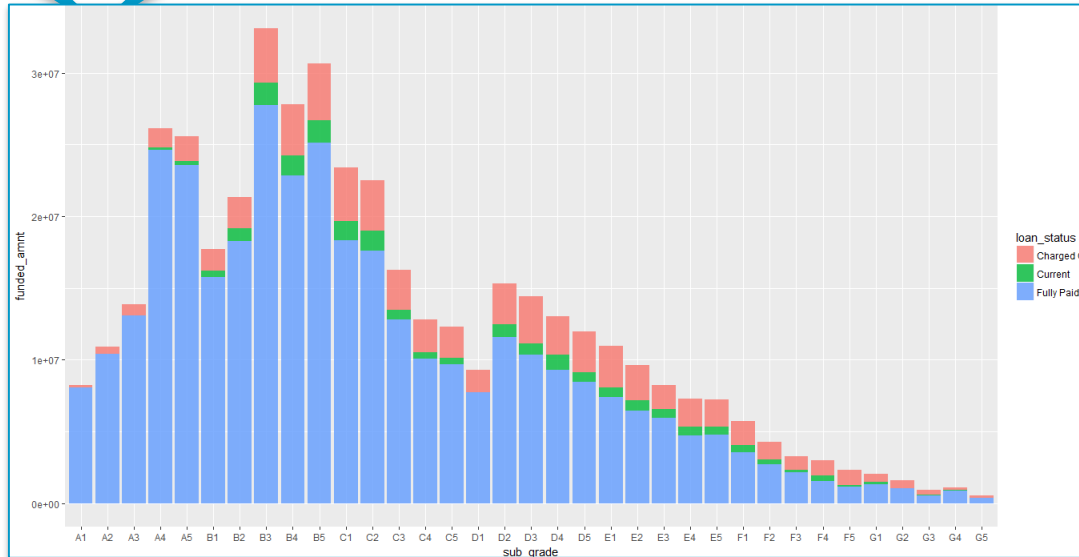


	Category	count	count_percentage	categorical_Variable_names
1	36 months	29096	73.26	term
2	60 months	10621	26.74	term
3	B	12020	30.26	grade
4	A	10085	25.39	grade
5	C	8098	20.39	grade
6	D	5307	13.36	grade
7	E	2842	7.16	grade
8	F	1049	2.64	grade
9	G	316	0.80	grade
10	B3	2917	7.34	sub_grade
11	A4	2886	7.27	sub_grade
12	A5	2742	6.90	sub_grade
13	B5	2704	6.81	sub_grade
14	B4	2512	6.32	sub_grade
15	C1	2136	5.38	sub_grade
16	B2	2057	5.18	sub_grade
17	C2	2011	5.06	sub_grade
18	B1	1830	4.61	sub_grade
19	A3	1810	4.56	sub_grade
20	C3	1529	3.85	sub_grade
21	A2	1508	3.80	sub_grade
22	D2	1348	3.39	sub_grade

22	D2	1348	3.39	sub_grade
23	C4	1236	3.11	sub_grade
24	C5	1186	2.99	sub_grade
25	D3	1173	2.95	sub_grade
26	A1	1139	2.87	sub_grade
27	D4	981	2.47	sub_grade
28	D1	931	2.34	sub_grade
29	D5	874	2.20	sub_grade
30	E1	763	1.92	sub_grade
31	E2	656	1.65	sub_grade
32	E3	553	1.39	sub_grade
33	E4	454	1.14	sub_grade
34	E5	416	1.05	sub_grade
35	F1	329	0.83	sub_grade
36	F2	249	0.63	sub_grade
37	F3	185	0.47	sub_grade
38	F4	168	0.42	sub_grade
39	F5	118	0.30	sub_grade
40	G1	104	0.26	sub_grade
41	G2	78	0.20	sub_grade
42	G4	56	0.14	sub_grade
43	G3	48	0.12	sub_grade

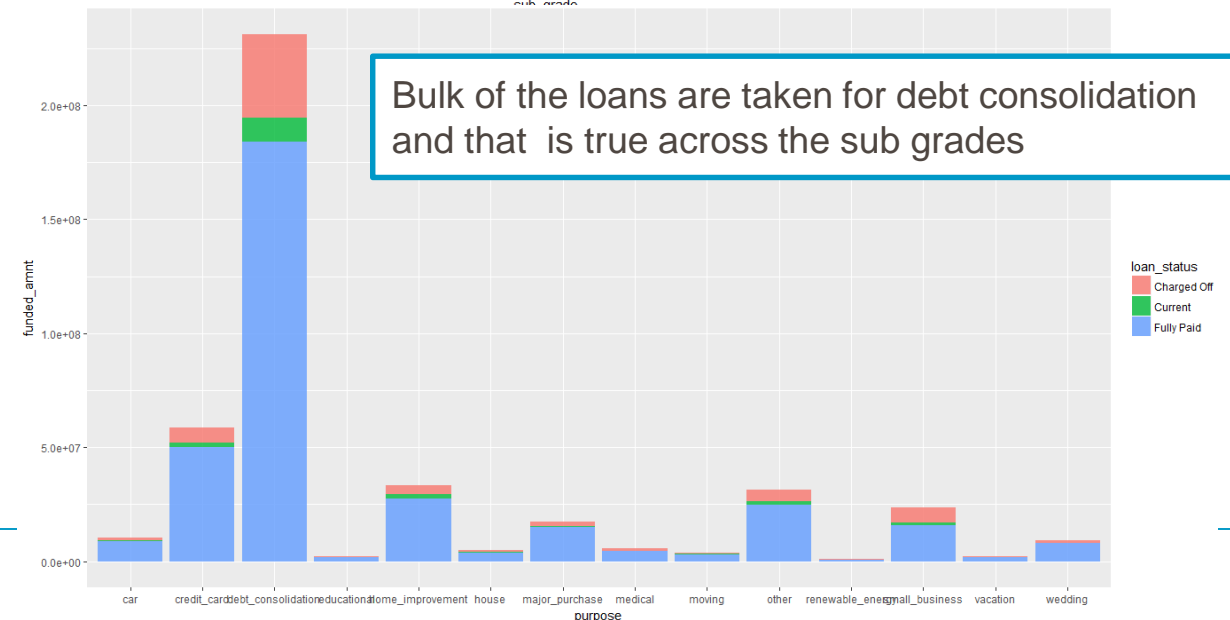
43	G3	48	0.12	sub_grade
44	G5	30	0.08	sub_grade
45	10+ years	8879	22.36	emp_length
46	< 1 year	4583	11.54	emp_length
47	2 years	4388	11.05	emp_length
48	3 years	4095	10.31	emp_length
49	4 years	3436	8.65	emp_length
50	5 years	3282	8.26	emp_length
51	1 year	3240	8.16	emp_length
52	6 years	2229	5.61	emp_length
53	7 years	1773	4.46	emp_length
54	8 years	1479	3.72	emp_length
55	9 years	1258	3.17	emp_length
56	n/a	1075	2.71	emp_length
57	RENT	18899	47.58	home_ownership
58	MORTGAGE	17659	44.46	home_ownership
59	OWN	3058	7.70	home_ownership
60	OTHER	98	0.25	home_ownership
61	NONE	3	0.01	home_ownership
62	Not Verified	16921	42.60	verification_status
63	Verified	12809	32.25	verification_status
64	Source Verified	9987	25.15	verification_status

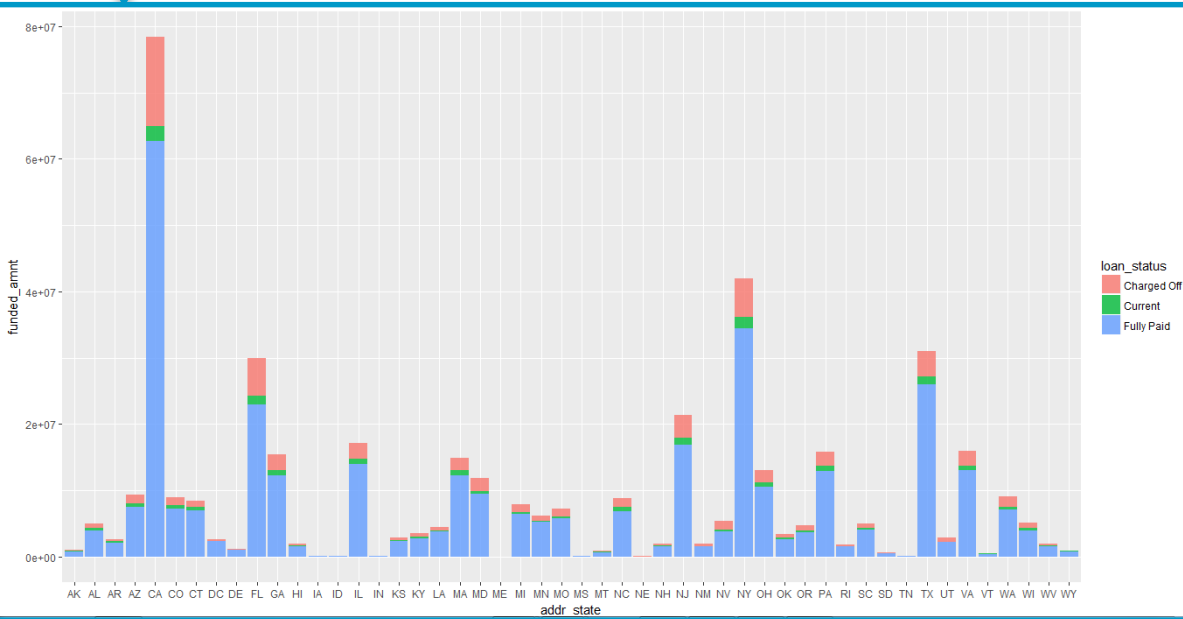
- ✓ After dropping some empty and continuous variables we are left with below key categorical variables in total. The above is just a few for reference
- ✓ Term, Grade, Sub Grade, employment length, home ownership, loan status, purpose of loan, state info.
- ✓ Few character fields like date, interest rate, revol utilization rate etc were converted to appropriate computational fields respectively.



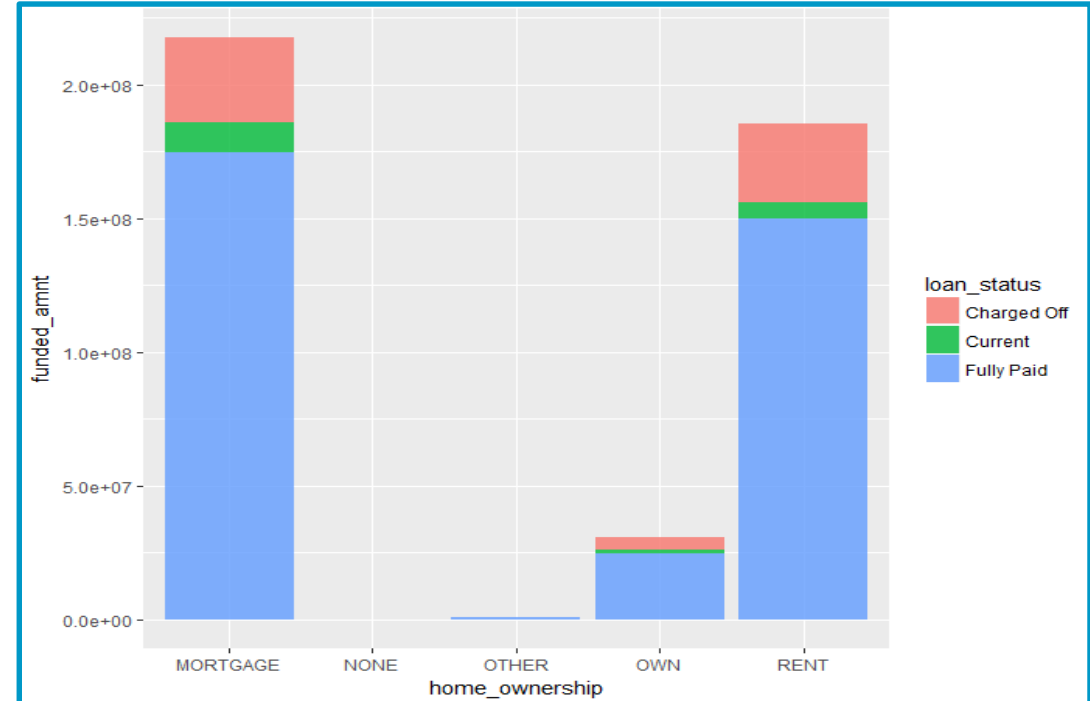
sub_grade	loan_status	total_funded_amnt	total_derieved_chargedoff_amnt	derieved_chargedoff_per	count
1 B5	Charged Off	3990200	2486595	37.68245	356
2 B3	Charged Off	3761700	2413693	35.83505	341
3 C1	Charged Off	3746550	2409662	35.68318	336
4 C2	Charged Off	3508125	2274662	35.16018	321
5 B4	Charged Off	3610875	2245824	37.80388	329
6 D3	Charged Off	3312075	2164186	34.65770	256
7 E1	Charged Off	2913175	1984624	31.87419	198
8 D5	Charged Off	2868450	1963501	31.54836	209
9 D2	Charged Off	2857725	1948400	31.81989	271
10 D4	Charged Off	2686400	1884805	29.83900	215
11 C3	Charged Off	2808350	1877886	33.13205	270

Top charge off's happening at end of B grade and beginning of C grade



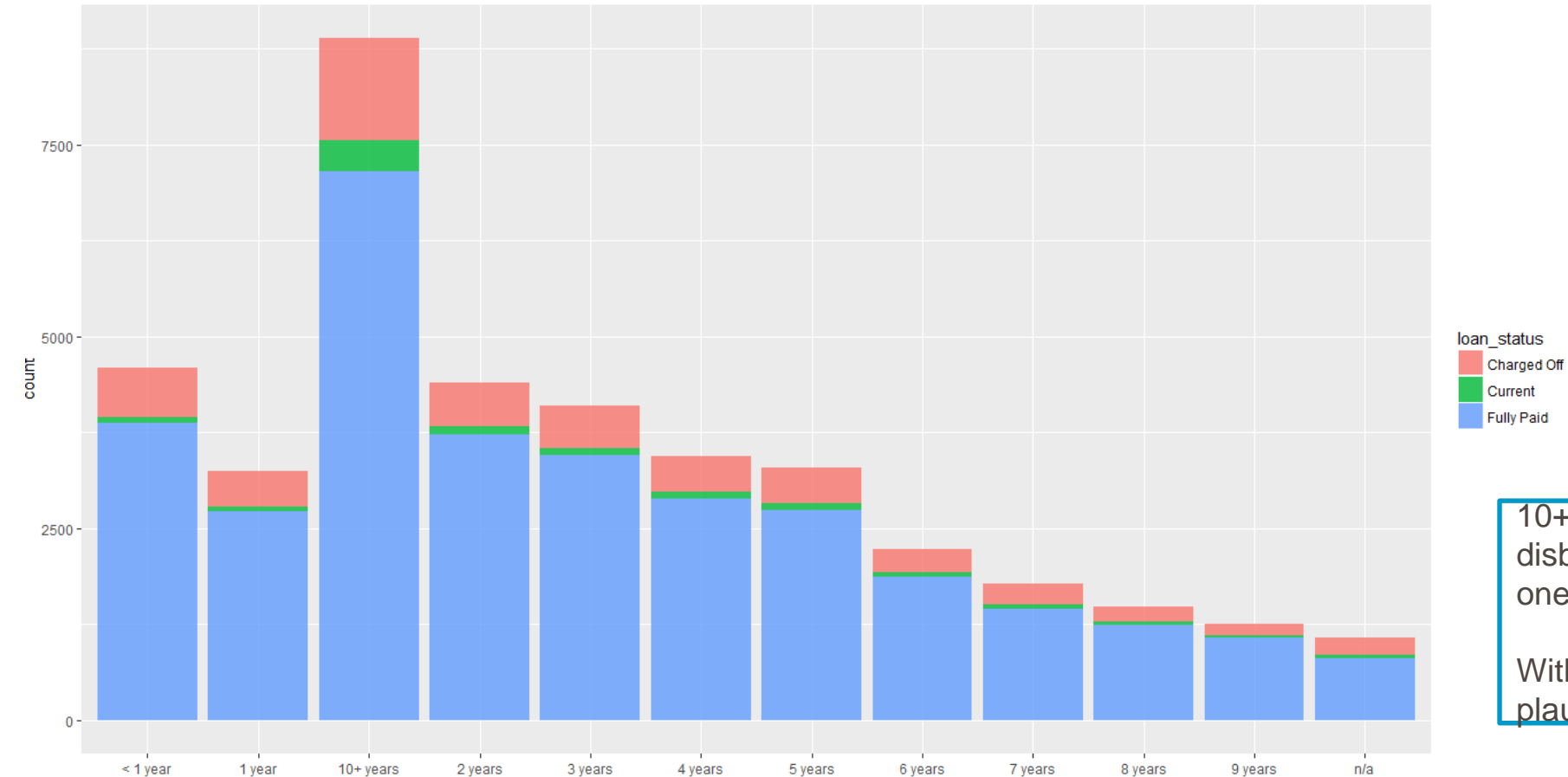


CA tops in the loans disbursed followed by NY (both in terms of fund amount and charged off amount)



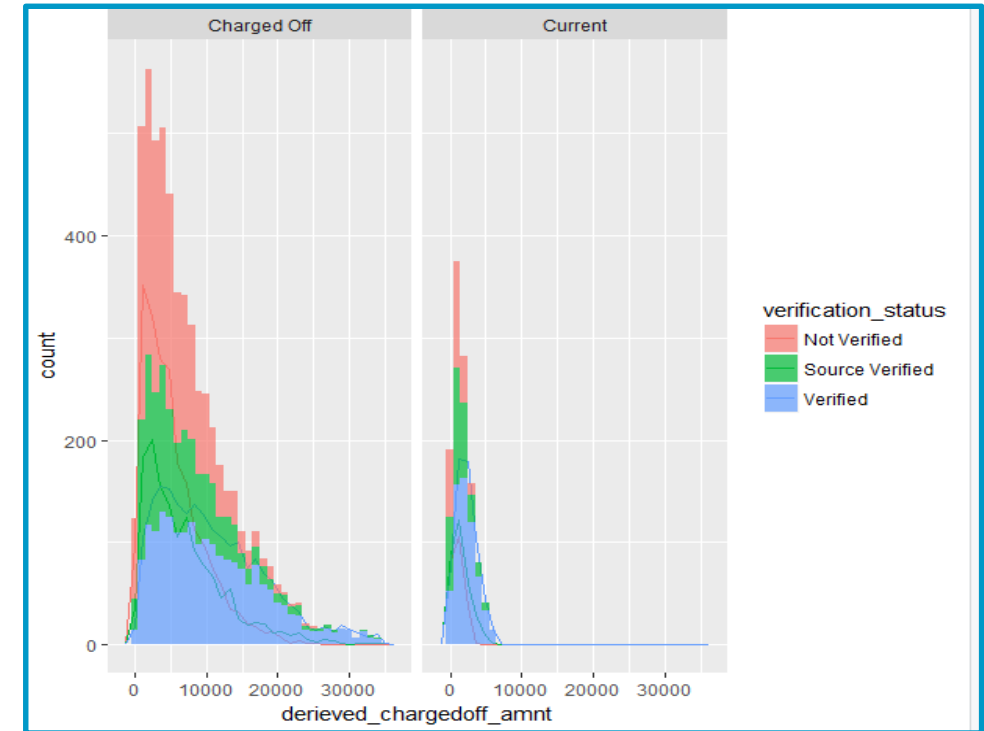
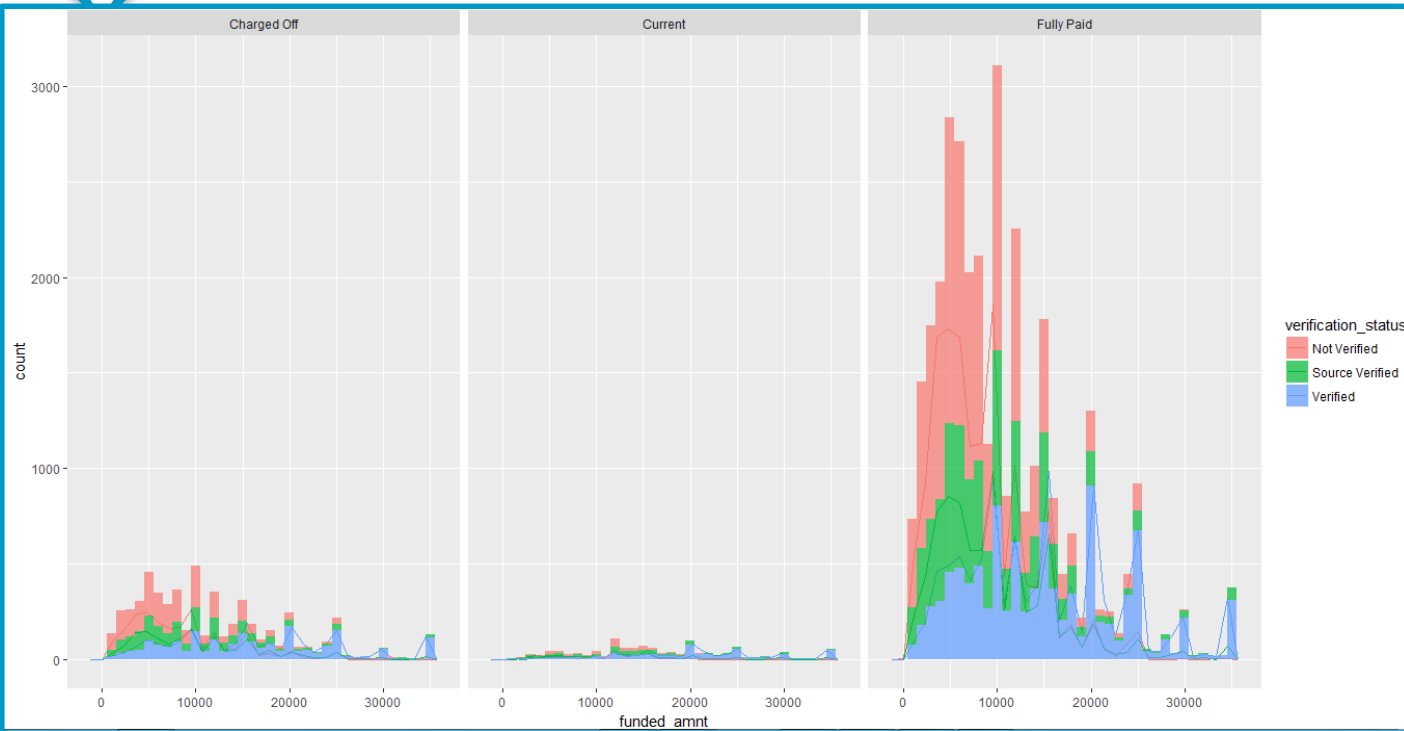
Bulk of the loans are taken by people living on rent or currently servicing a mortgage.



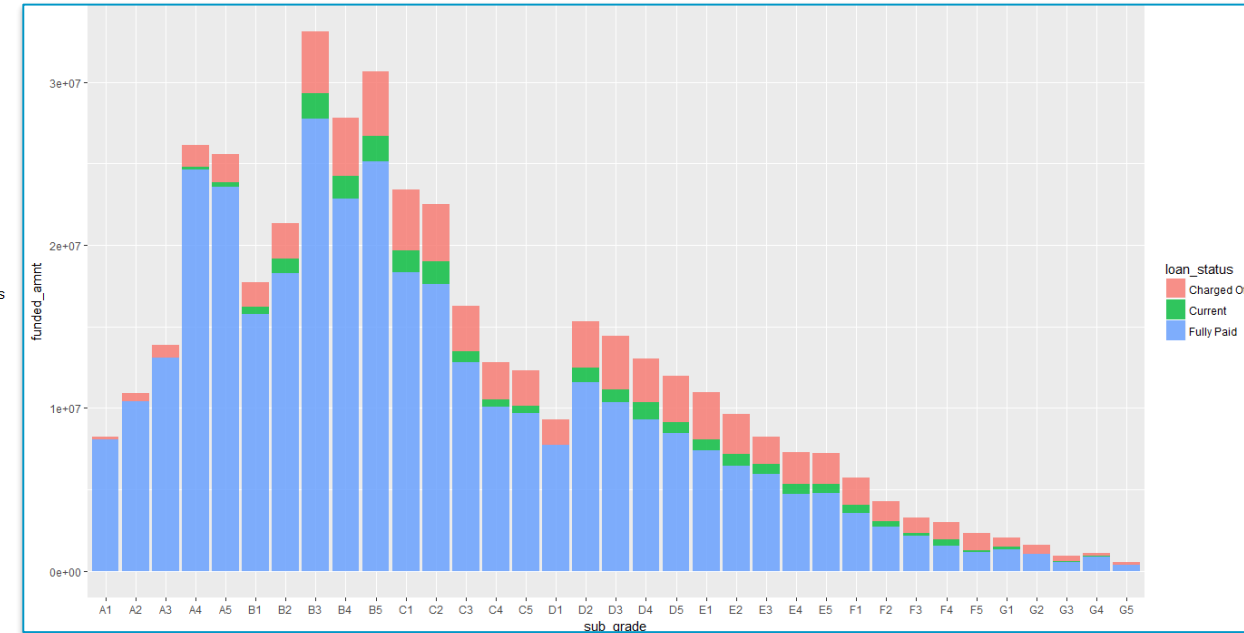
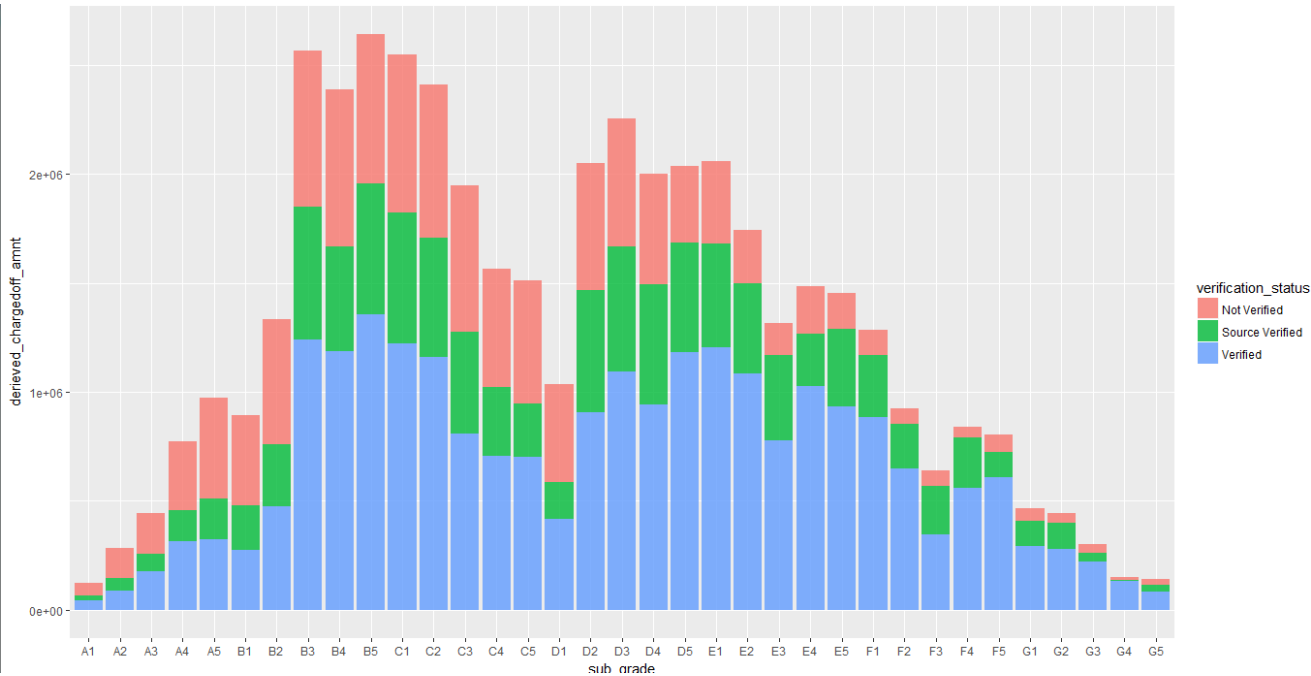


10+ bucket has the maximum loans disbursed, followed by less than one year. And two years.

Within 1-3 year of employment the plausible risk is high for default.

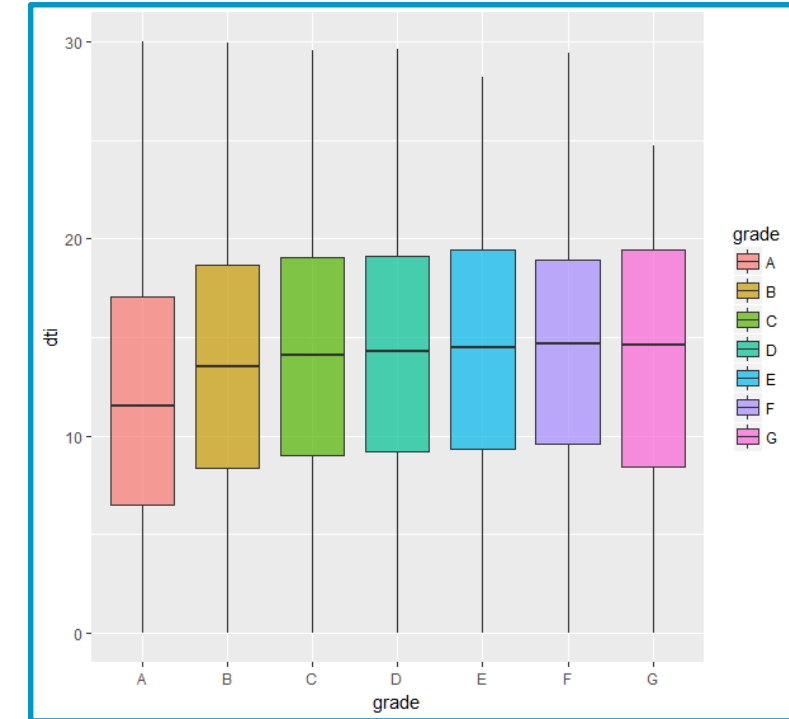
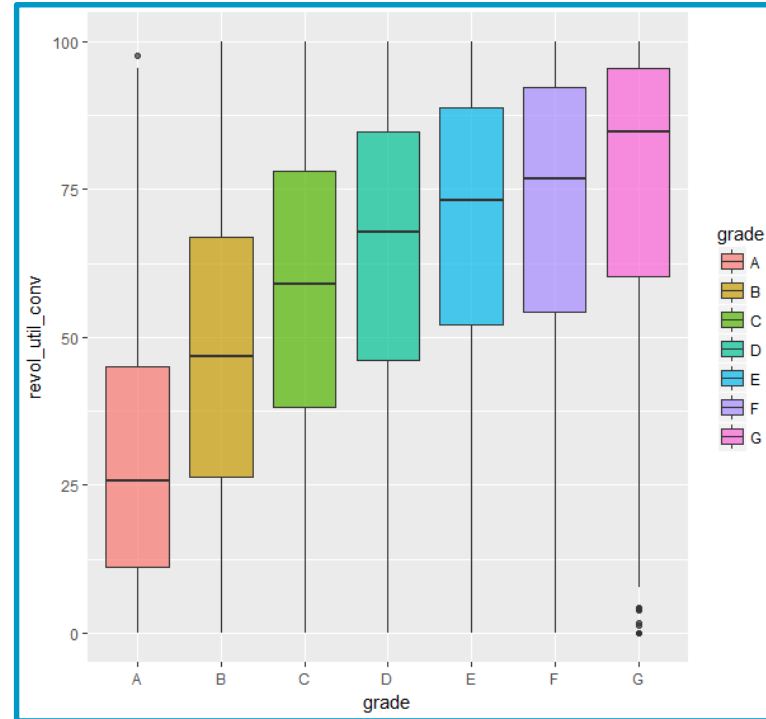
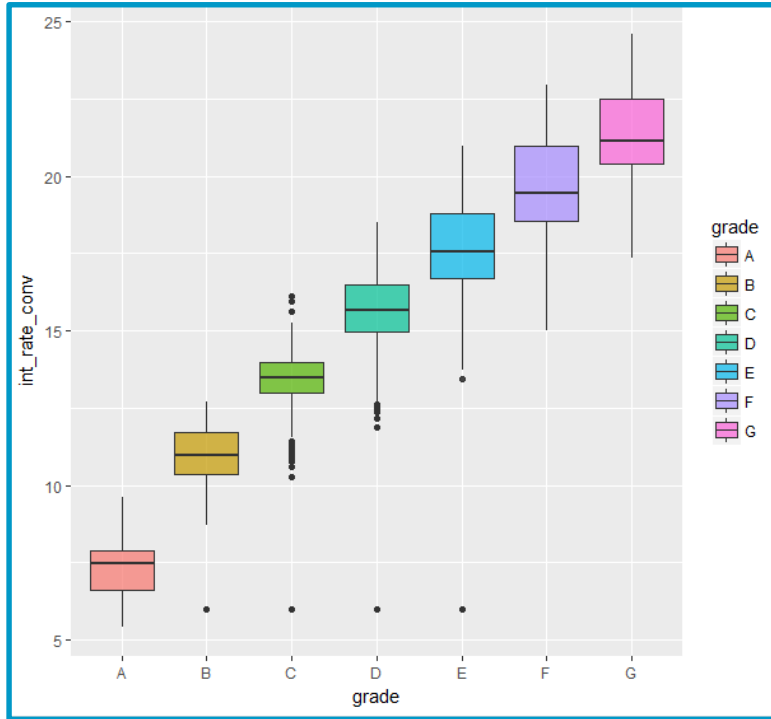


**High counts of amounts less than 10000 are issued with out verification.**



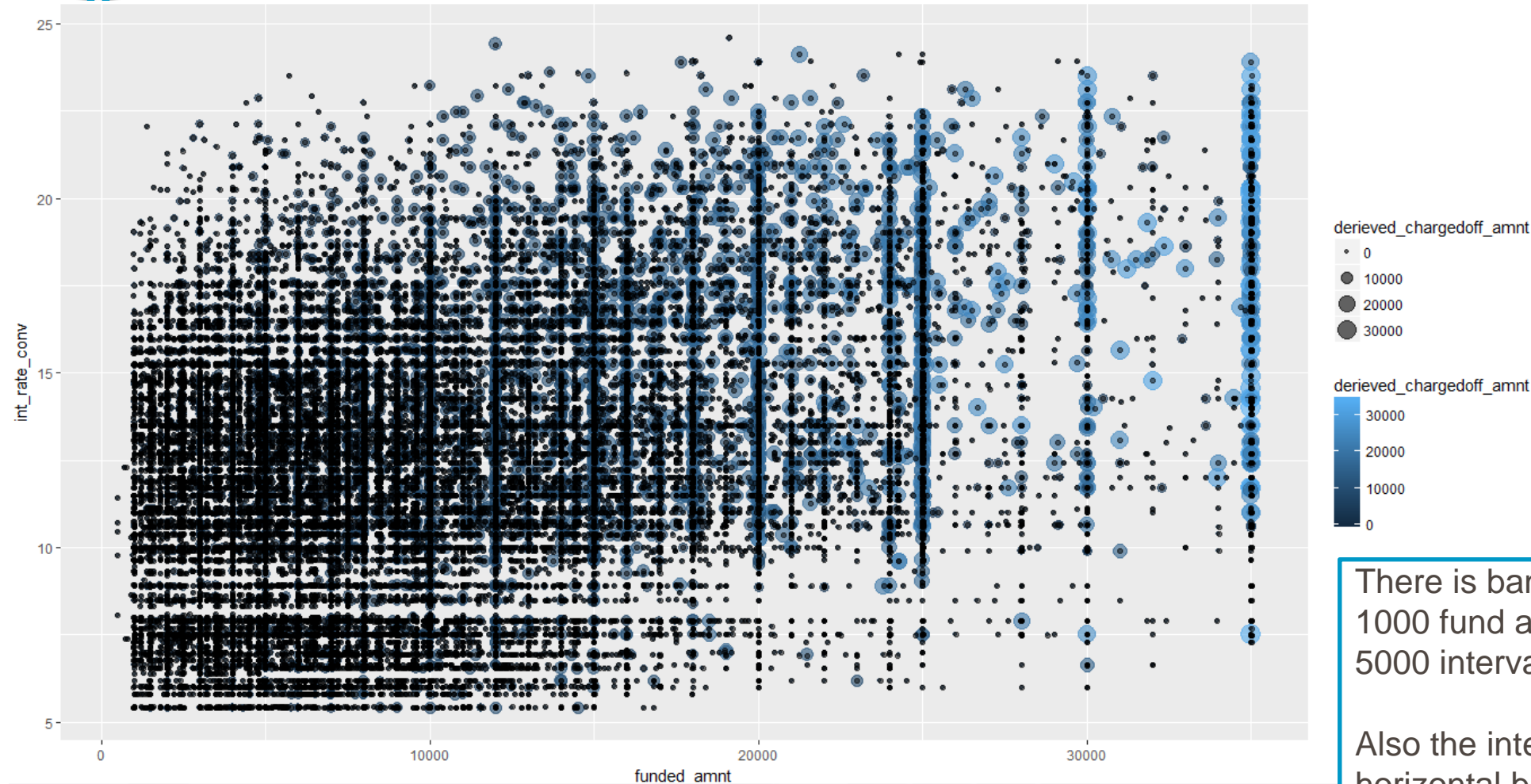
The plot above show verification across grades, on the right top loan status across the grades.

**Lack of Verification is resulting to charge offs???**



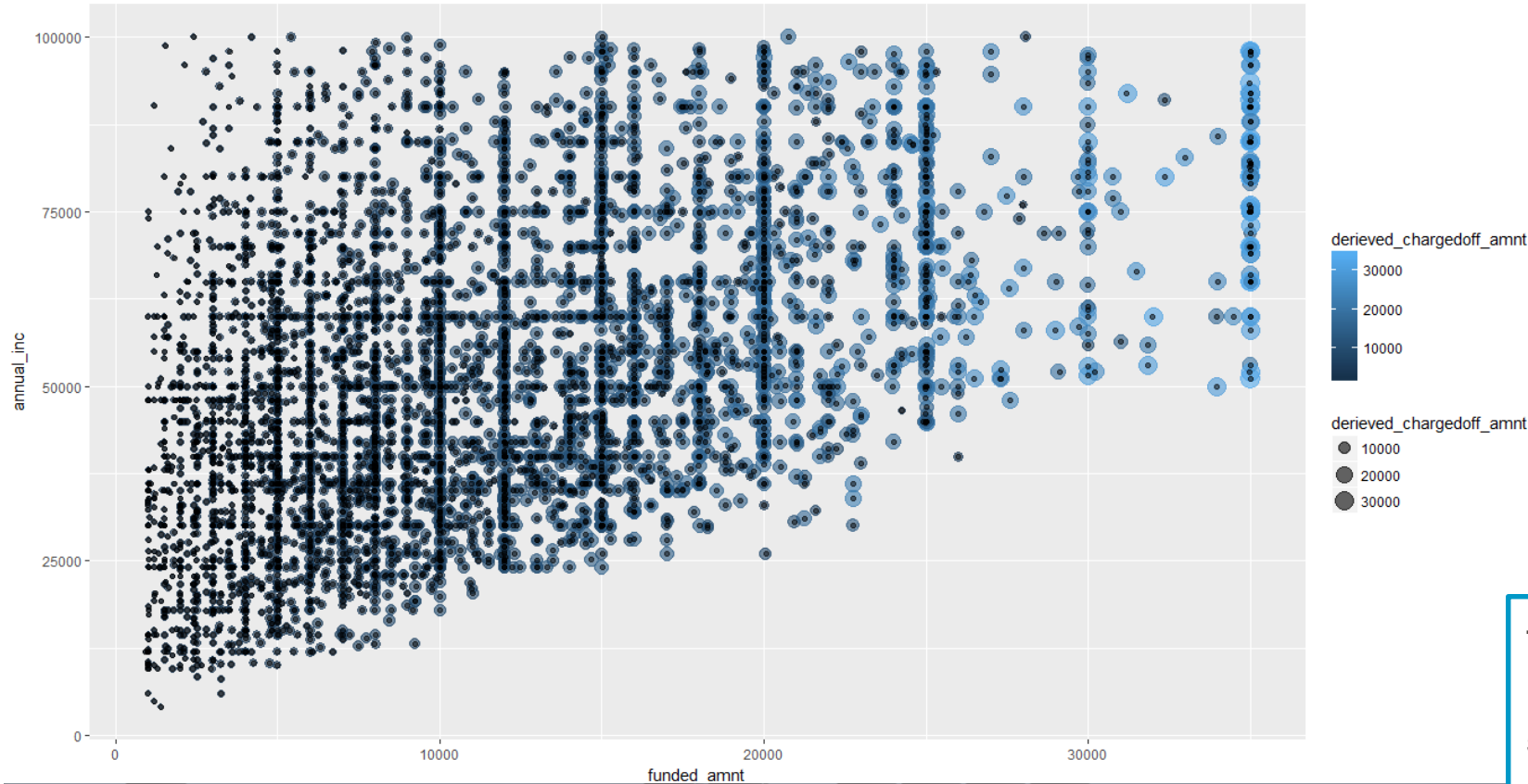
Loan Interest rate, revol utilization and deb to income ratio seem to correlate with grades.

- 'A' being the best grade the interest rate is low, they also seem to have the lowest dti and revolve utilization
- 'G' being the lowest grade the interest rate is high, revolve utilization and dti is also higher .



There is banding pattern at multiples of 1000 fund amount. More prominent in 5000 intervals.

Also the interest rate seem to have horizontal banding as well. seems like many professionals of the same type are qualifying for similar interest rates.

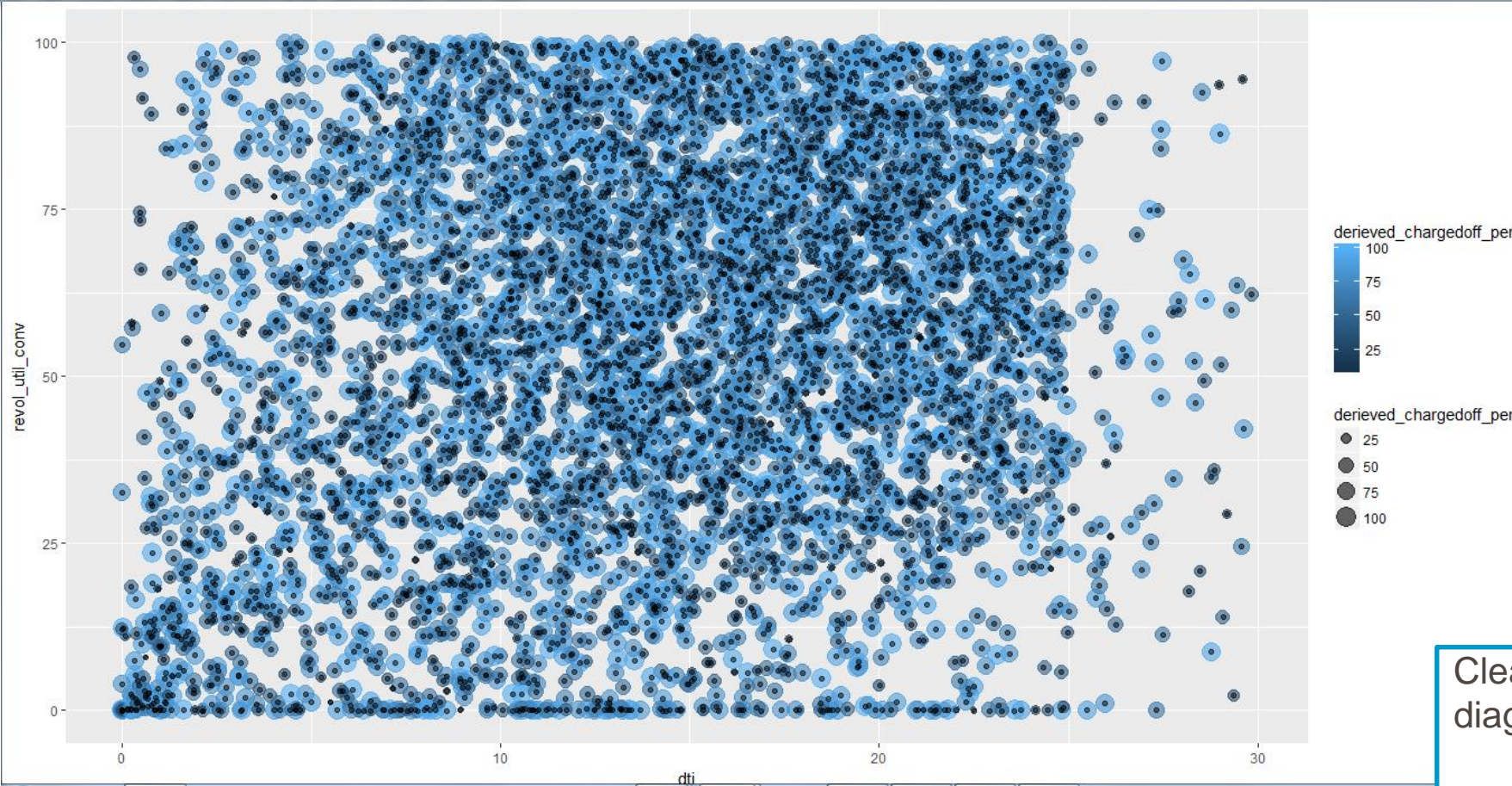


Plot has excluded income greater than 100,000. and only includes charged of status.

There is banding pattern at multiples of 1000 fund amount. More prominent in 5000 intervals.

Also the interest rate seem to have horizontal banding as well. professionals from similar fields earning same salary?

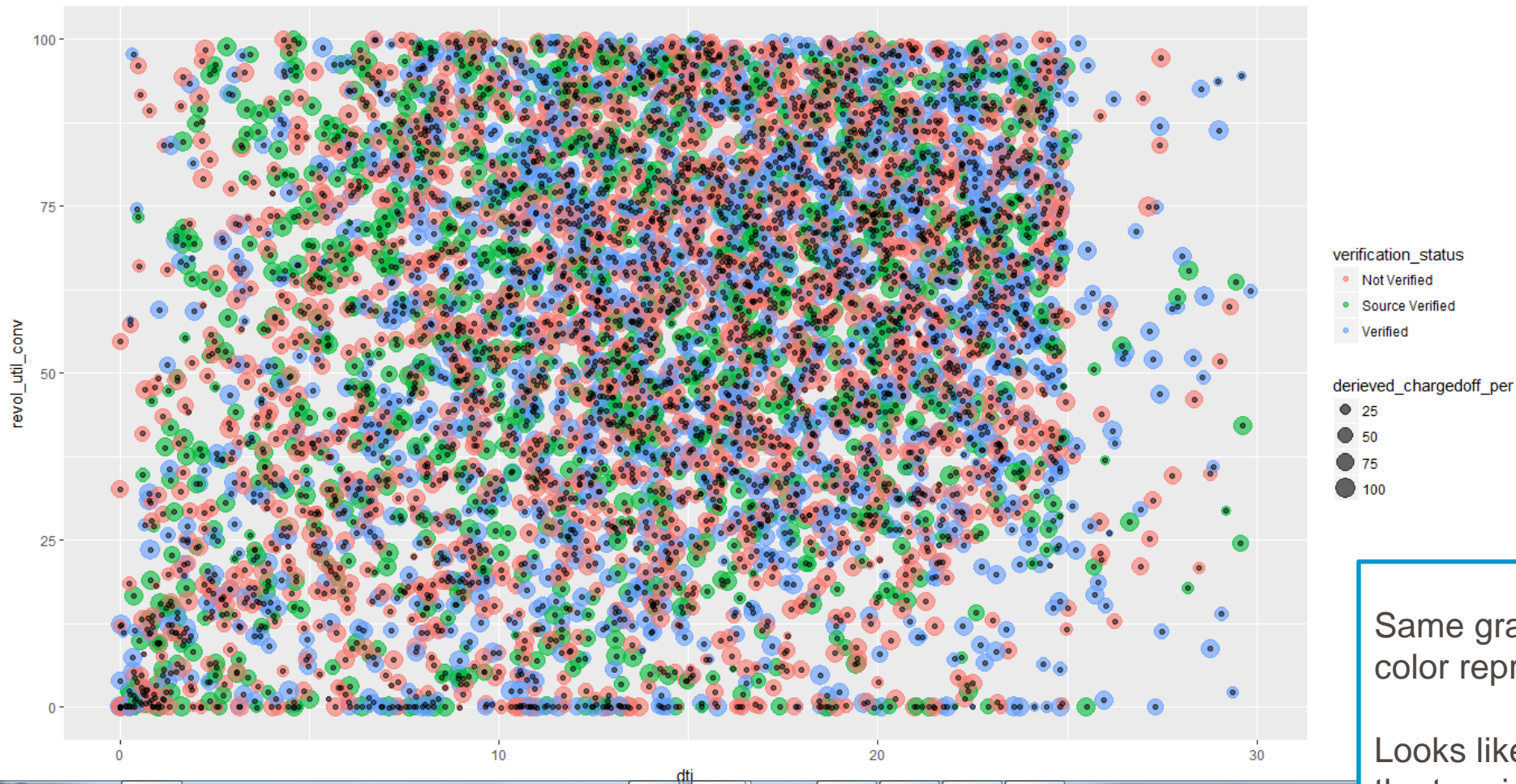




Dti vs revolving utilization - Plot has excluded income greater than 100,000. and only includes charged of status.

Clearly there is a clustering above the diagonals from top left to bottom right.

Top right clustering seems logical for charged off accounts. As both revolving utilization % increases and debt to income ratio increases the tendency to charge of increases.

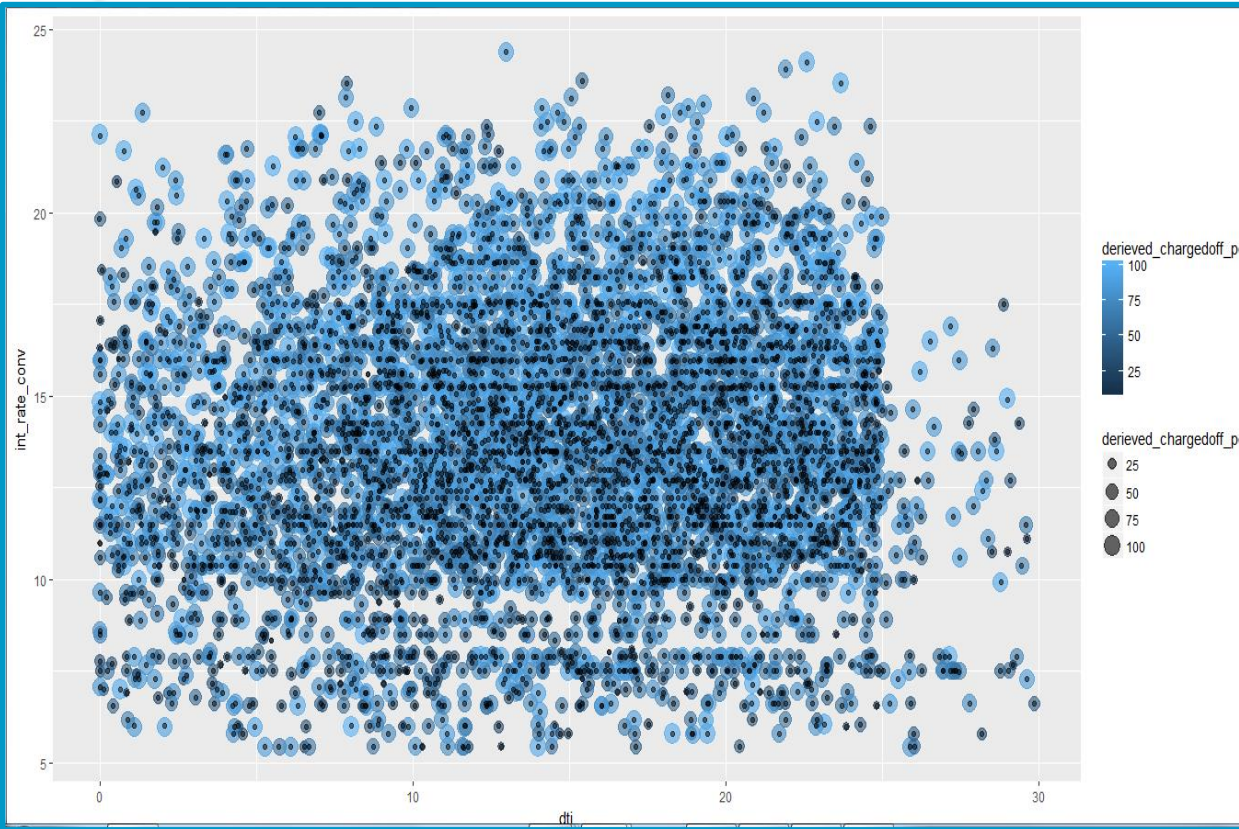


Plot has excluded income greater than 100,000. and only includes charged of status.

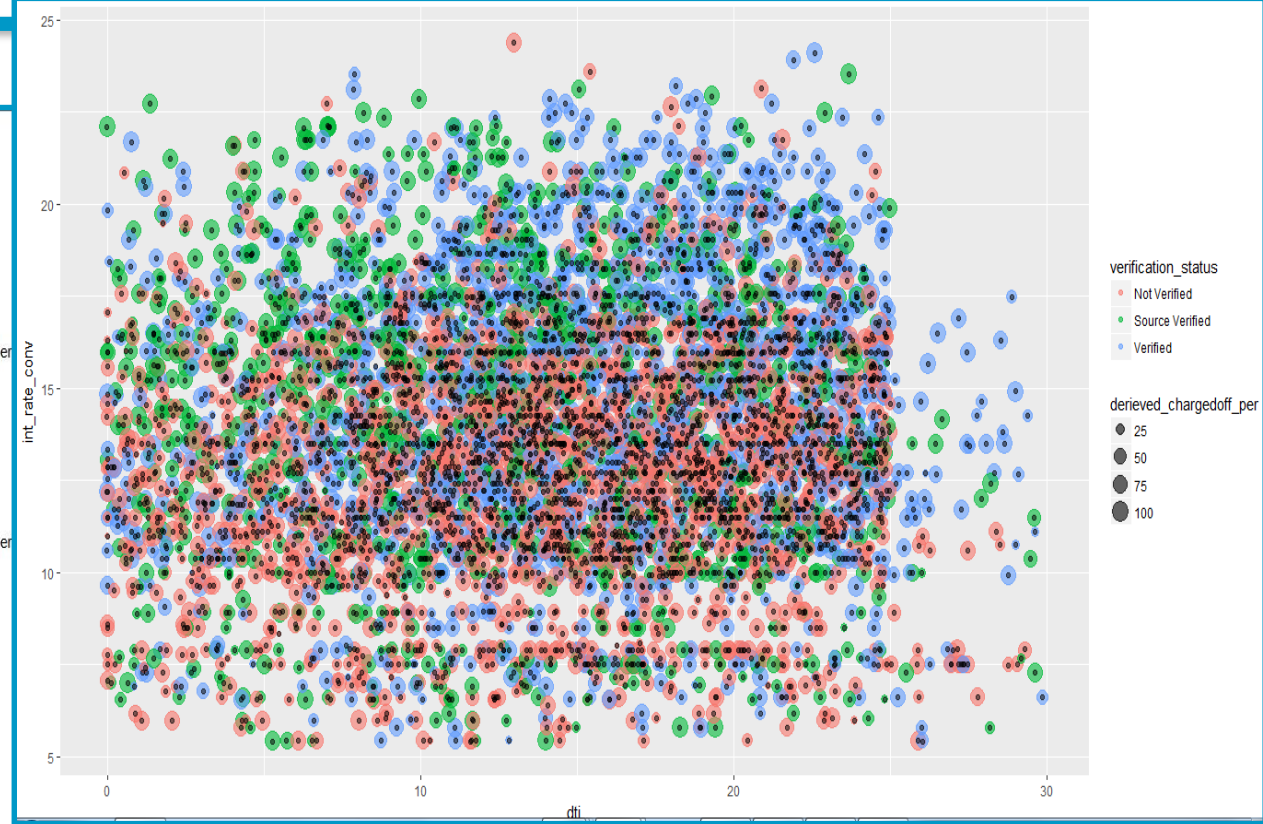
Same graph as the previous one. But the color represents verification status.

Looks like the reds are concentrating at the top right?? Meaning 'not verified' as potential reasons for charge offs??



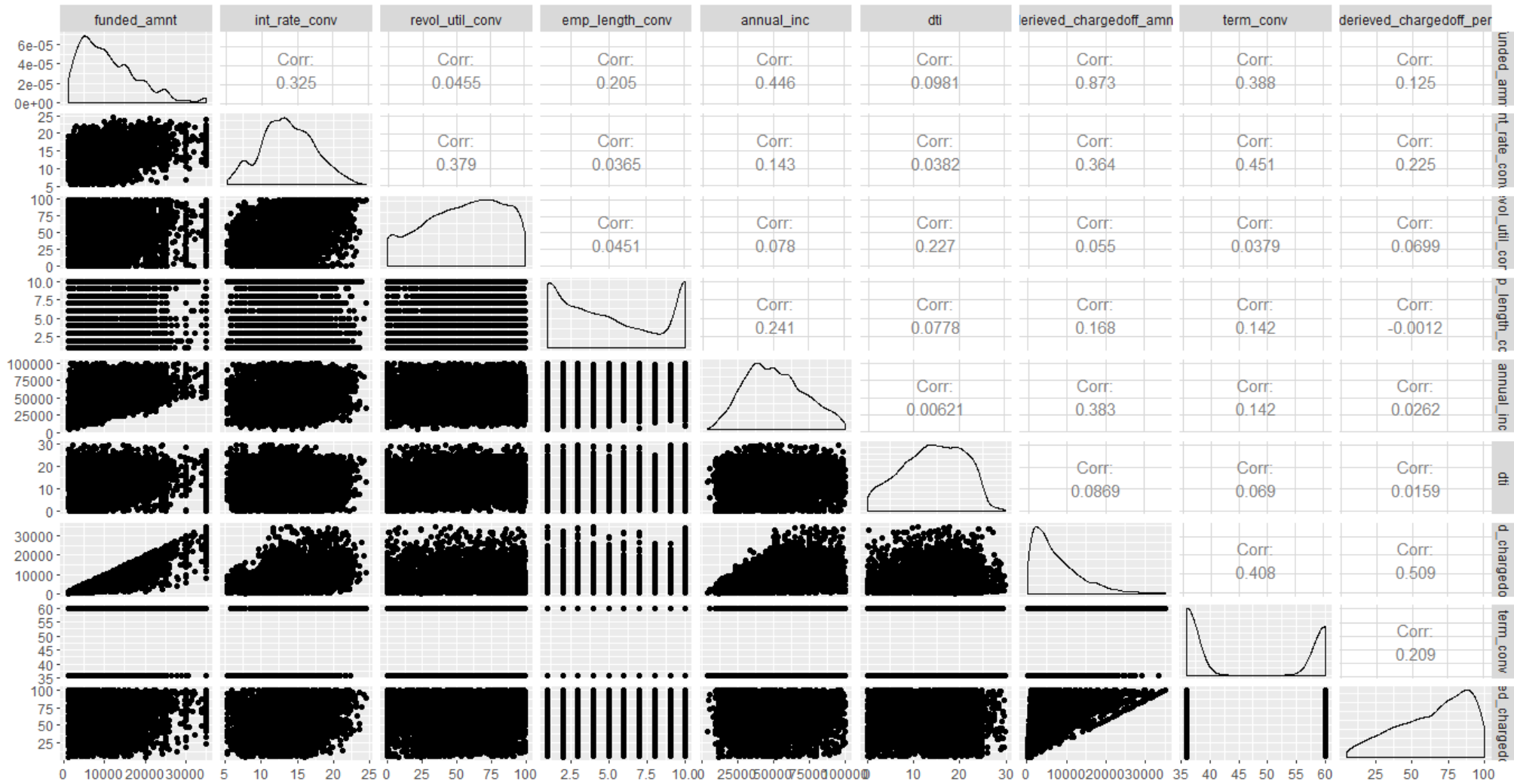


Dti vs interest rate - Plot has excluded income greater than 100,000. and only includes charged of status.

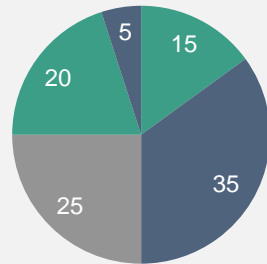


Between 5% and 10% there seem to be less charge of percentage.

While magically above 10% there is a jump in charge off percentage... and so is the red colors (not verified)???



- ✓ Top charge off's happening at end of B grade (B3) and beginning of C grade (C2).
- ✓ Bulk of the loans are taken for debt consolidation and that is true across the sub grades
- ✓ Bulk of the loans are taken by people living on rent or currently servicing a mortgage.
- ✓ 10+ years of employment has the maximum loans disbursed, followed by less than one year, two years. Within 1-3 year of employment the plausible risk is high for default.
- ✓ High counts of amounts less than 10000 are issued with out verification. And these are charging off.
- ✓ Banding patten is observed on funding amounts at 1000 interval and more prominent in 5000 interval.
- ✓ Dti vs revolving utilization shows- Top right clustering seems logical for charged off accounts. As both revolving utilization % increases and debt to income ratio increases the tendency to charge of increases.
- ✓ Fund amount vs interest rate - the interest rate seem to have horizontal banding as well. seems like many professionals of the same type are qualifying for similar interest rates. There seem to be correlation to professions.

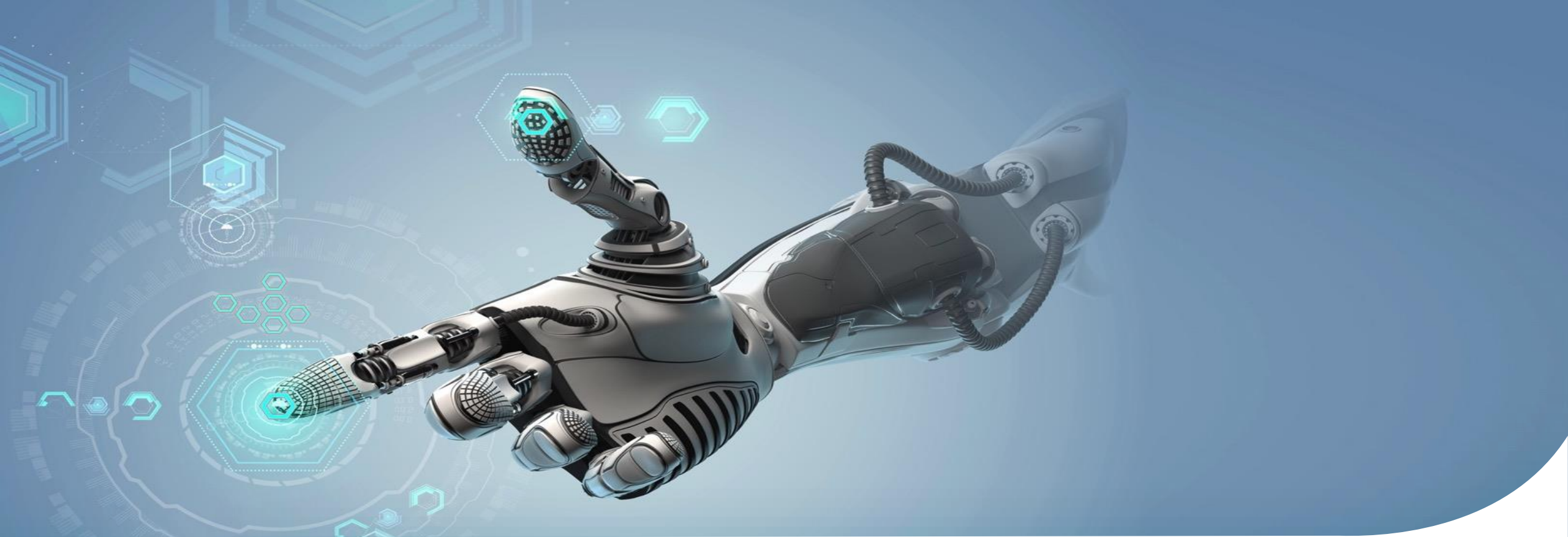


## 5 important driver variables


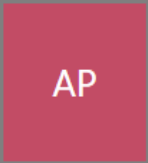


- DTI – debt to income ratio
- Interest rate
- Revolving utilization ratio
- Funding amount
- Sub grade
- Verification status

Note there seem to be evidence that suggest that profession of the candidate is also a strong driving variable.





# Thank You.

 <b>Abhishek Ranjan</b> Bangalore	 <b>Amiyanshu Pratihari</b> Pune
 <b>Maxim Rohit</b> Pune	 <b>Karthikeyan Seetharaman</b> Pune