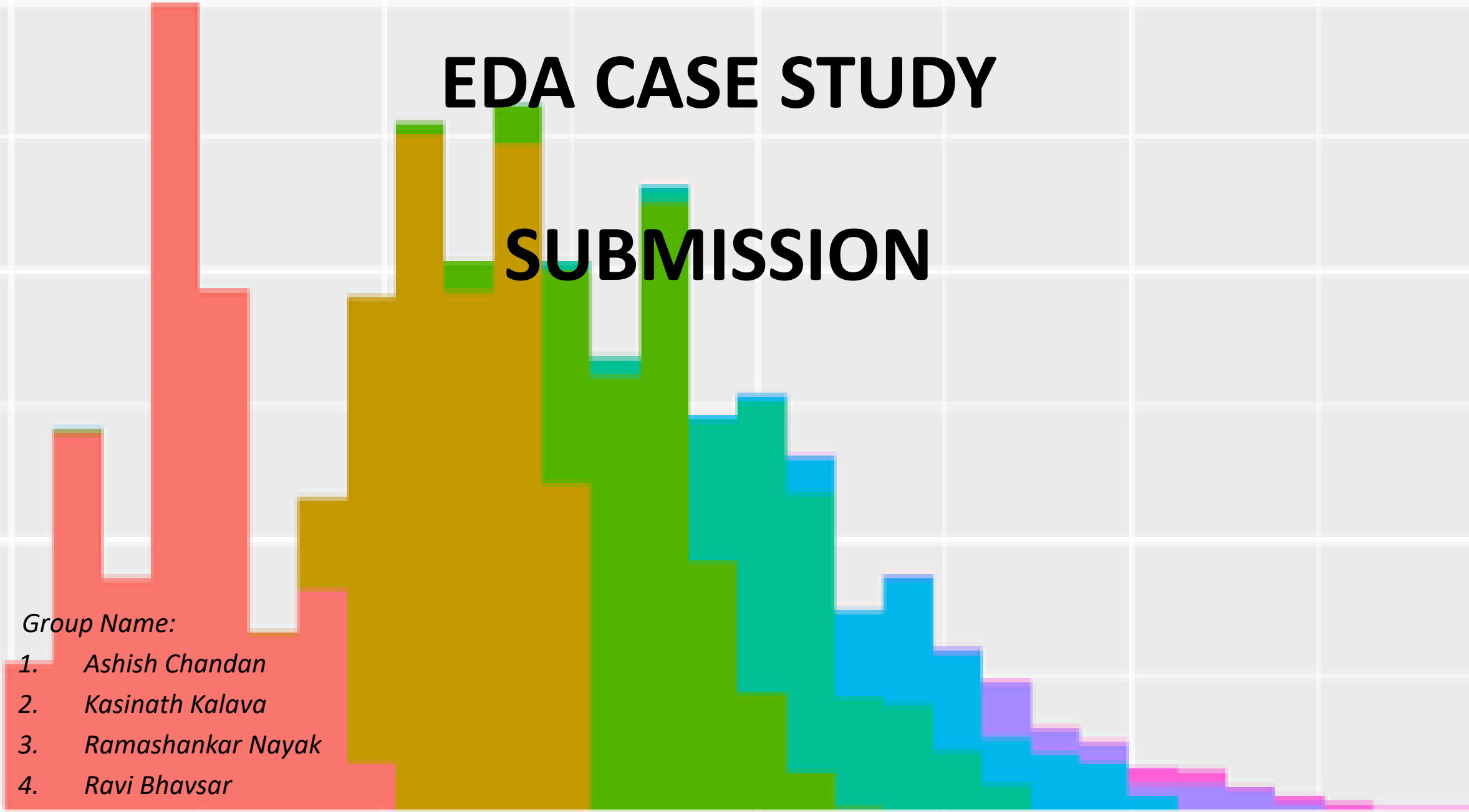
EDA CASE STUDY

SUBMISSION



Group Name:

1. *Ashish Chandan*
2. *Kasinath Kalava*
3. *Ramashankar Nayak*
4. *Ravi Bhavsar*

Objective & Methodology

Business Objective: The objective is to understand the driving factors behind loan default, i.e. variables which are strong indicators of default.

Methodology:

• Understanding the provided loan dataset with the objective to identify various variables.

• Perform EDA on the loan dataset.

• Determine the driving factors and attributes that are strong indicators for loan defaults.

• Finding trends on the basis of Count and Ratio of Charged Off loans using CrossTable in the below format and corresponding plot:

Each cell contains Count of combination and Ratio of Count with Row Total

to determine %age of Charged off/Fully Paid

Count
Ratio

Count refers Count of combination

Ratio refers Ratio of Count with row total

Data Exploration, Data Cleaning and Data Manipulation

Data Exploration:

1. Data set contains 111 columns, however, we considered following for our analysis :

id	loan_amnt	term	int_rate	installment	grade
sub_grade	emp_length	home_ownership	annual_inc	verification_status	issue_d
loan_status	purpose	addr_state	dti	delinq_2yrs	inq_last_6mths
open_acc	pub_rec	revol_bal	revol_util	revol_util	

• We have eliminated any fields that would not have been known at the time of issuance as we'll be trying to make decisions on loan investments using available pre-issuance data.
• We have also eliminate a few indicative data fields that are repetitive or too granular to be analyzed.

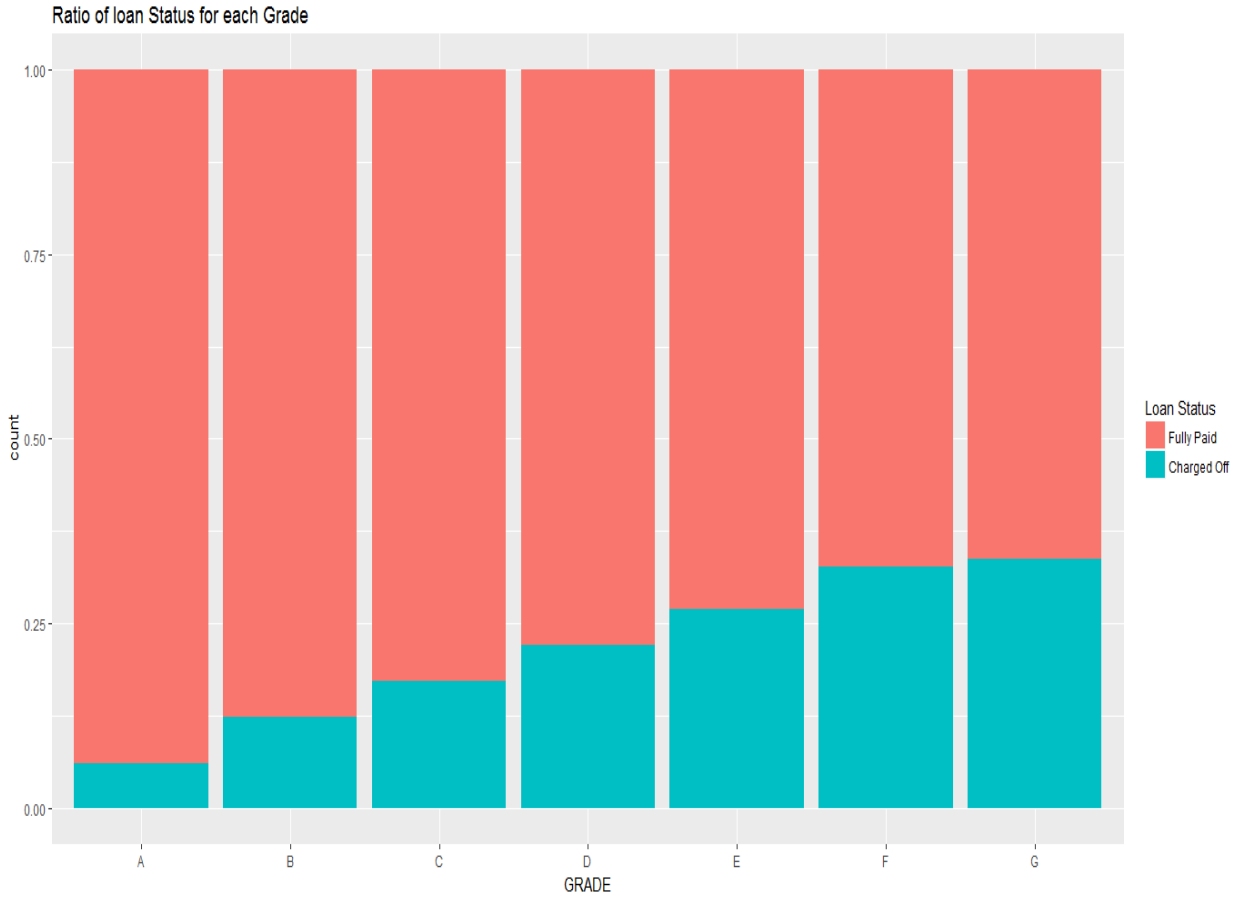
2. We have to find driving variables to determine defaulters, so just keeping "Charged Off" and "Fully Paid" loan_status and have removed "Current" as we cannot categorize the record into defaulter/non-defaulter.

Data Cleaning and Data Manipulation: Possible data inconsistencies are as follows:

- Checking for any duplicate ids
- Check for missing values
- Removing extra space
- Extracting year and month from issue date
- Convert a required variables to factors
- Convert int_rate and revol_util into numeric

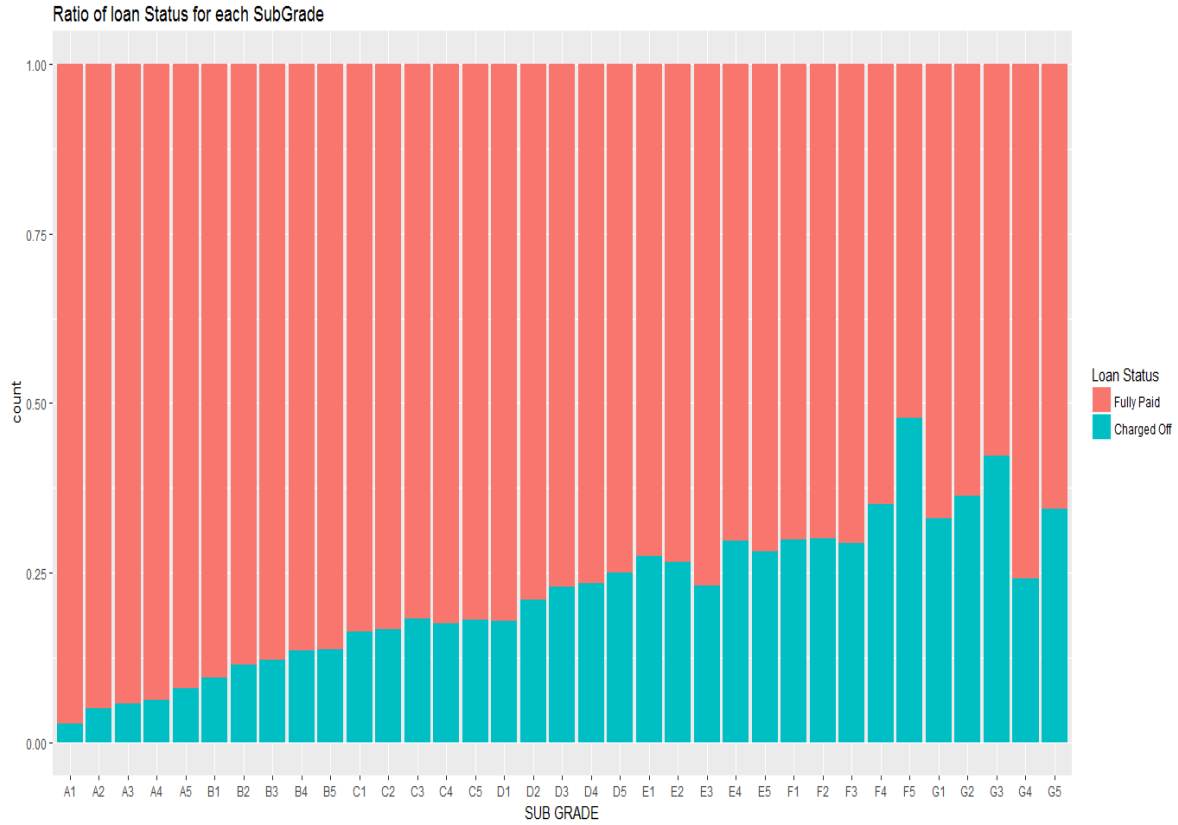
Analysis: Grades

loan$grade	loan$loan_status Charged Off	Fully Paid	Row Total
A	602	9443	10045
	0.060	0.940	0.260
B	1425	10250	11675
	0.122	0.878	0.303
C	1347	6487	7834
	0.172	0.828	0.203
D	1118	3967	5085
	0.220	0.780	0.132
E	715	1948	2663
	0.268	0.732	0.069
F	319	657	976
	0.327	0.673	0.025
G	101	198	299
	0.338	0.662	0.008
Column Total	5627	32950	38577



Default rate steadily increases as the loan grades worsen from A to G.

Analysis: Sub-Grades



Plot shows similar pattern to Grades, although the trend begins to weaken across the G1-G5.
In contrast to data points for the A1 to F5, there are fewer data points for G1-G5 and differences are not large enough to be significant. Hence, G1-G5 sub-grades can be ignored.

Frequency

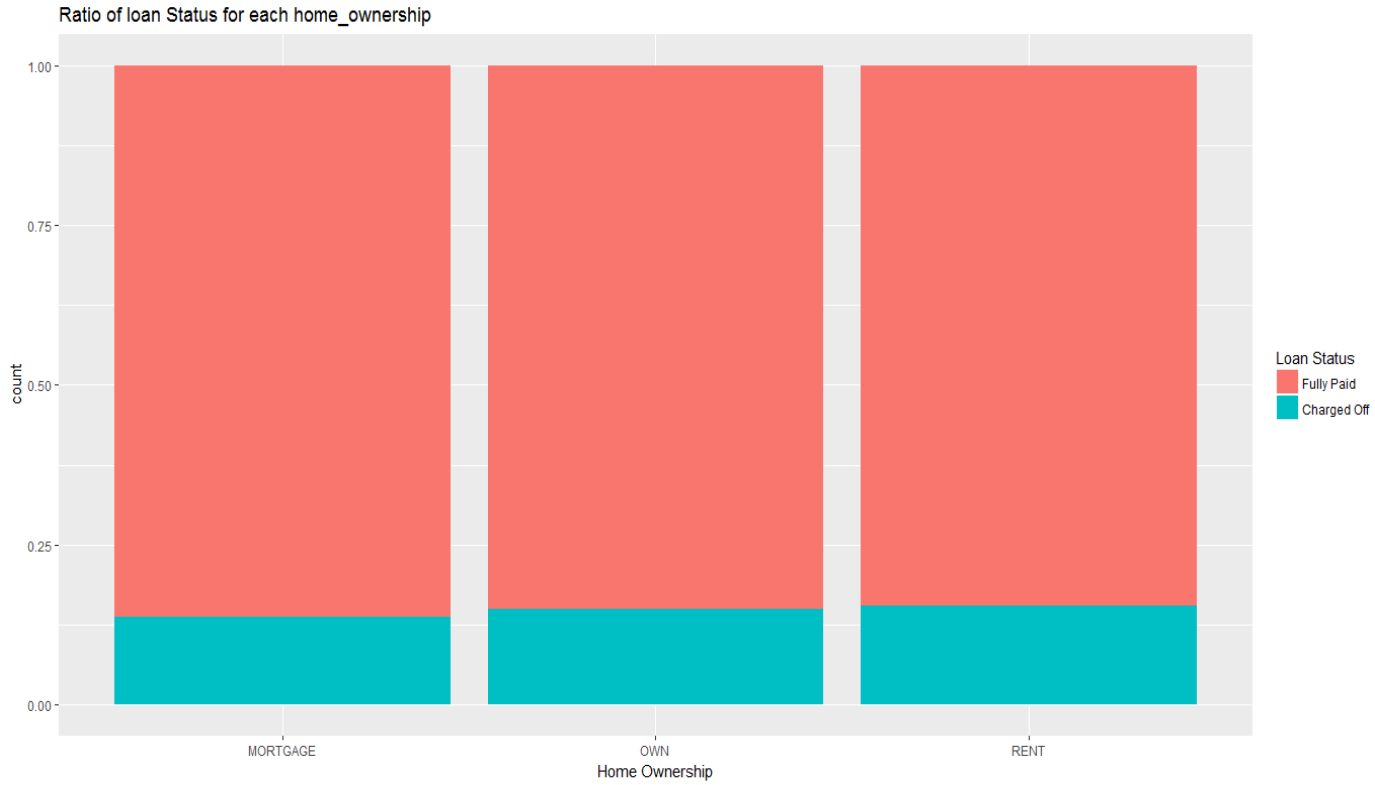
	Charged Off	Fully Paid
A1	30	1109
A2	74	1434
A3	103	1707
A4	178	2695
A5	217	2498
B1	171	1626
B2	228	1773
B3	341	2484
B4	329	2108
B5	356	2259
C1	336	1719
C2	321	1610
C3	270	1218
C4	212	994
C5	208	946
D1	167	764
D2	271	1015
D3	256	860
D4	215	703
D5	209	625
E1	198	524
E2	163	451
E3	119	397
E4	126	298
E5	109	278
F1	91	214
F2	70	163
F3	51	123
F4	53	98
F5	54	59
G1	31	63
G2	28	49
G3	19	26
G4	13	41
G5	10	19

Ratio

	Charged Off	Fully Paid
A1	0.0053	0.0337
A2	0.0132	0.0435
A3	0.0183	0.0518
A4	0.0316	0.0818
A5	0.0386	0.0758
B1	0.0304	0.0493
B2	0.0405	0.0538
B3	0.0606	0.0754
B4	0.0585	0.0640
B5	0.0633	0.0686
C1	0.0597	0.0522
C2	0.0570	0.0489
C3	0.0480	0.0370
C4	0.0377	0.0302
C5	0.0370	0.0287
D1	0.0297	0.0232
D2	0.0482	0.0308
D3	0.0455	0.0261
D4	0.0382	0.0213
D5	0.0371	0.0190
E1	0.0352	0.0159
E2	0.0290	0.0137
E3	0.0211	0.0120
E4	0.0224	0.0090
E5	0.0194	0.0084
F1	0.0162	0.0065
F2	0.0124	0.0049
F3	0.0091	0.0037
F4	0.0094	0.0030
F5	0.0096	0.0018
G1	0.0055	0.0019
G2	0.0050	0.0015
G3	0.0034	0.0008
G4	0.0023	0.0012
G5	0.0018	0.0006

Analysis: Home Ownership

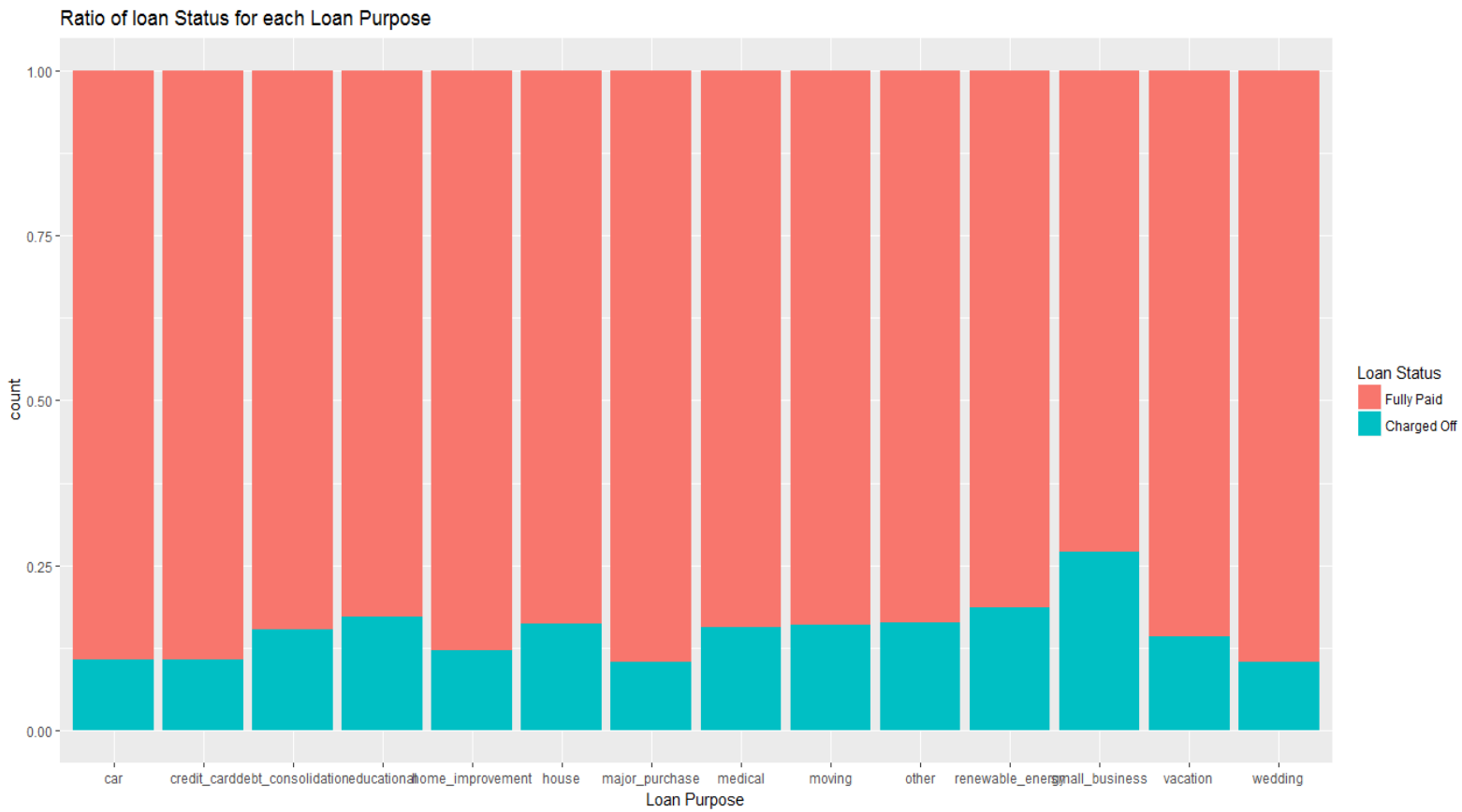
loan1$home_ownership	loan1$loan_status Charged Off	Fully Paid	Row Total
MORTGAGE	2327	14694	17021
	0.137	0.863	0.442
OWN	443	2532	2975
	0.149	0.851	0.077
RENT	2839	15641	18480
	0.154	0.846	0.480
Column Total	5609	32867	38476



Those with Mortgages default (Charged off) the least, followed by those who own their own homes.

Analysis: Loan Purpose

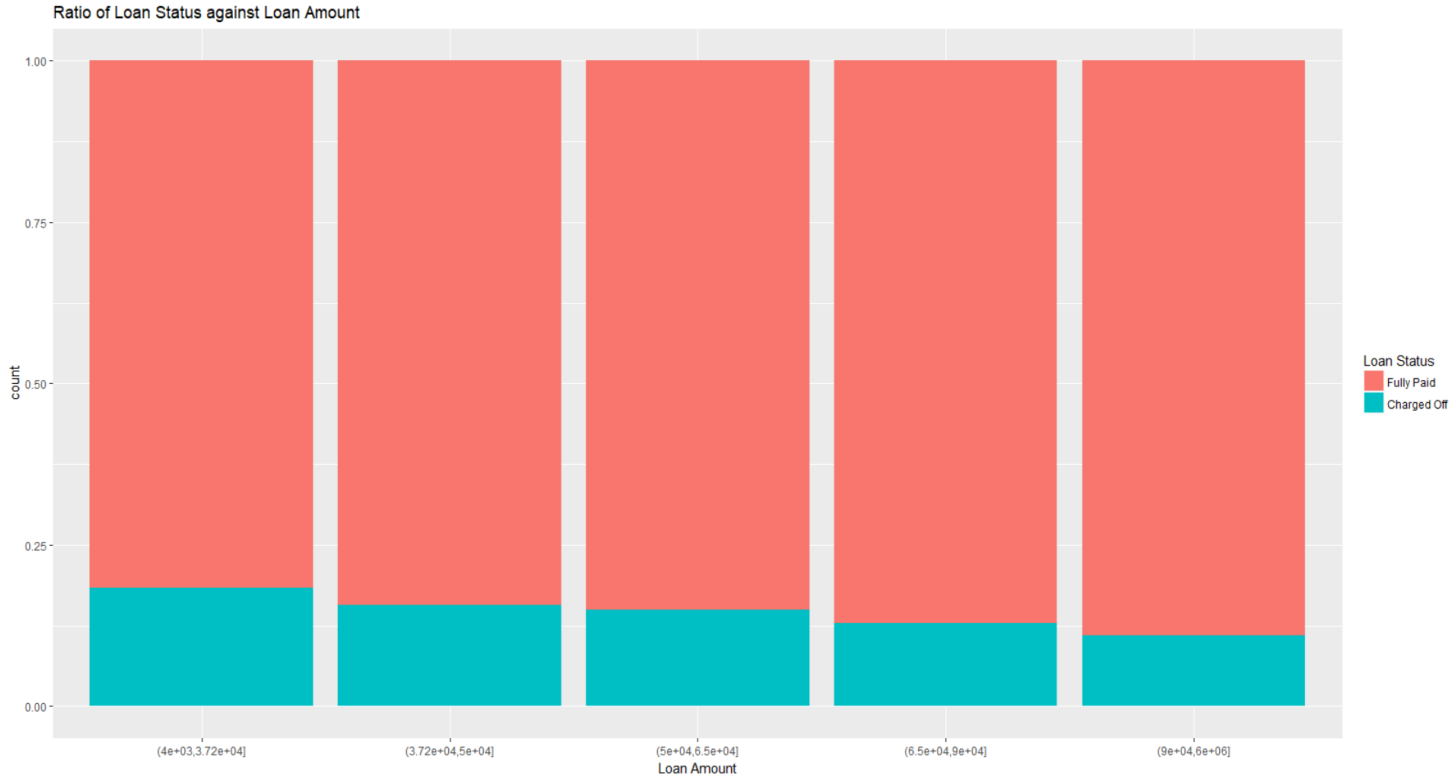
loan$purpose	loan$loan_status Charged Off	Fully Paid	Row Total
car	160 0.107	1339 0.893	1499 0.039
credit_card	542 0.108	4485 0.892	5027 0.130
debt_consolidation	2767 0.153	15288 0.847	18055 0.468
educational	56 0.172	269 0.828	325 0.008
home_improvement	347 0.121	2528 0.879	2875 0.075
house	59 0.161	308 0.839	367 0.010
major_purchase	222 0.103	1928 0.897	2150 0.056
medical	106 0.156	575 0.844	681 0.018
moving	92 0.160	484 0.840	576 0.015
other	633 0.164	3232 0.836	3865 0.100
renewable_energy	19 0.186	83 0.814	102 0.003
small_business	475 0.271	1279 0.729	1754 0.045
vacation	53 0.141	322 0.859	375 0.010
wedding	96 0.104	830 0.896	926 0.024
Column Total	5627	32950	38577



Loan purpose refers to the borrower's stated reason for taking out the loan.
Here, loan status for Small business is very poor.

Analysis: Loan Amount

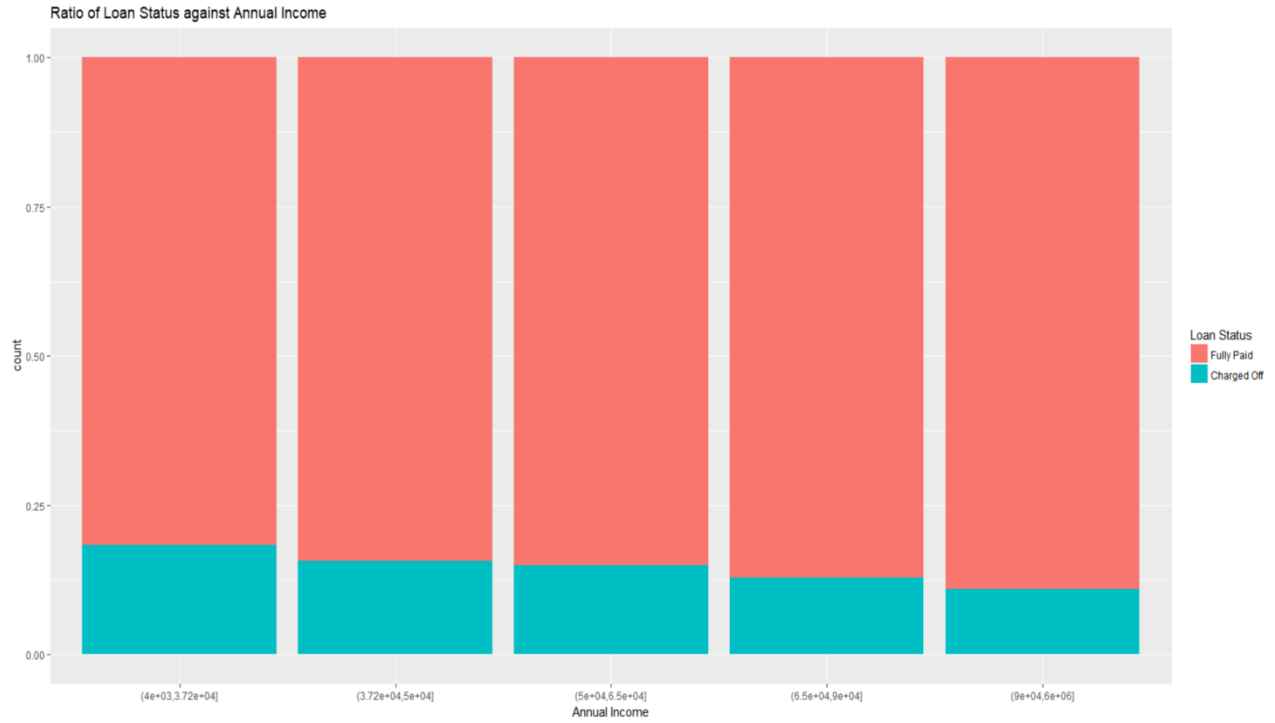
loan$loan_amnt_bkt	loan$loan_status: Charged Off	Fully Paid	Row Total
(0,15000]	4011	25946	29957
	0.134	0.866	0.777
(15000,30000]	1436	6449	7885
	0.182	0.818	0.204
(30000,35000]	180	555	735
	0.245	0.755	0.019
Column Total	5627	32950	38577



As the amount borrowed increases, we see increasing rates of defaulting loans

Analysis: Annual Income

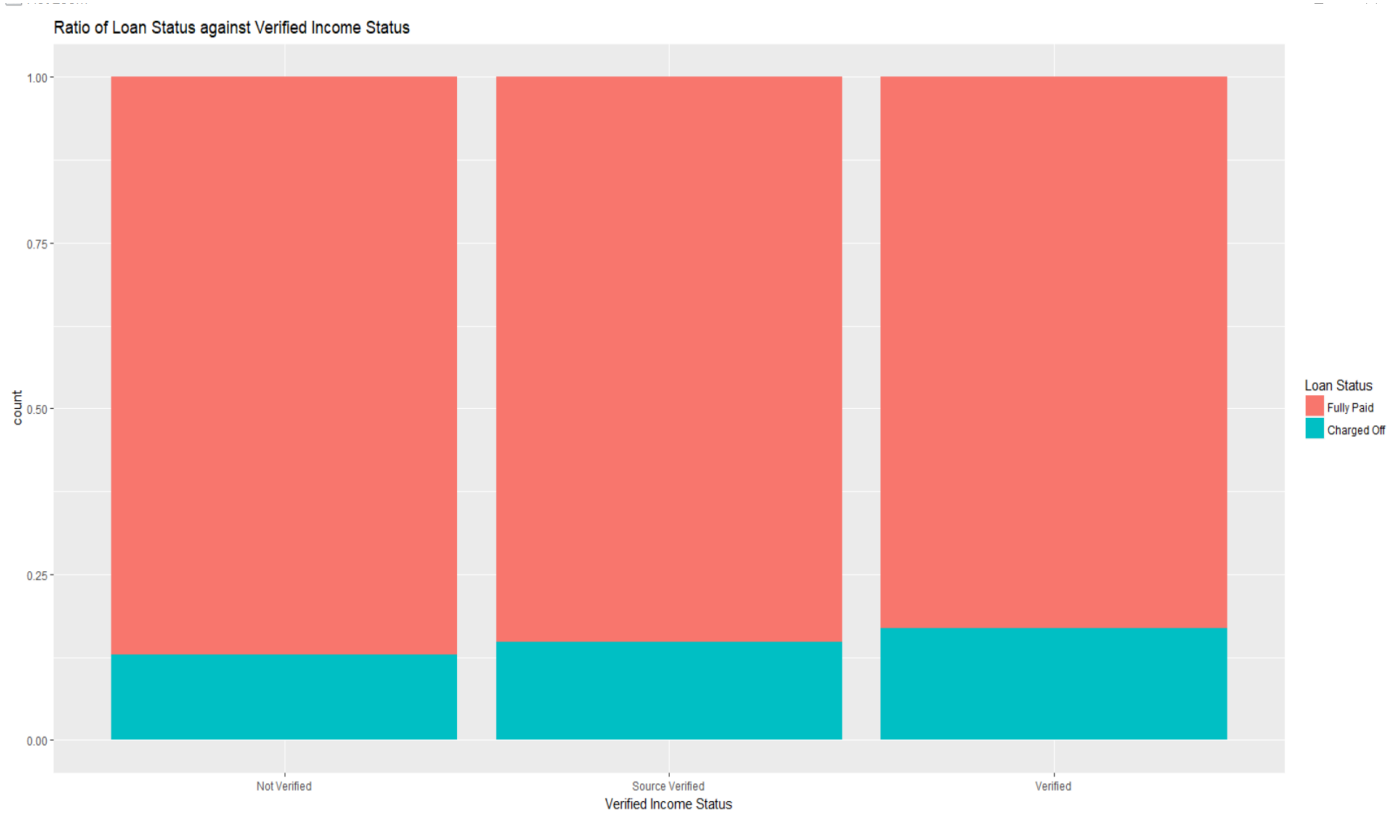
loan$annual_inc_bkt	loan$loan_status Charged Off	Fully Paid	Row Total
(4e+03,3.72e+04]	1413	6302	7715
	0.183	0.817	0.200
(3.72e+04,5e+04]	1219	6549	7768
	0.157	0.843	0.201
(5e+04,6.5e+04]	1149	6530	7679
	0.150	0.850	0.199
(6.5e+04,9e+04]	1028	6901	7929
	0.130	0.870	0.206
(9e+04,6e+06]	818	6667	7485
	0.109	0.891	0.194
Column Total	5627	32949	38576



The higher a borrower's annual income the less likely they are to default to repay their loans.

Analysis: Verified Income Status

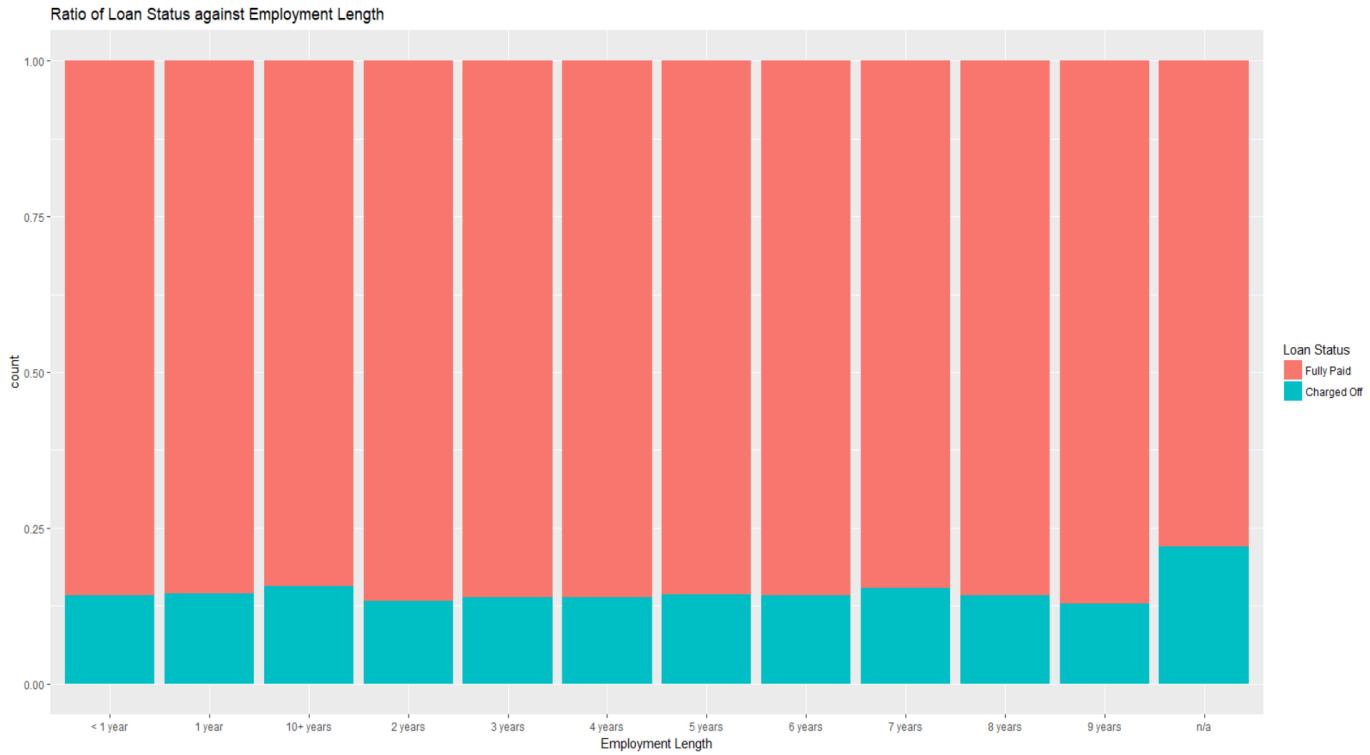
loan$verification_status	loan$loan_status Charged Off	Fully Paid	Row Total
Not Verified	2142	14552	16694
	0.128	0.872	0.433
Source Verified	1434	8243	9677
	0.148	0.852	0.251
Verified	2051	10155	12206
	0.168	0.832	0.316
Column Total	5627	32950	38577



Interestingly, we see that as income verification increases, the loan performance actually worsens.

Analysis: Employment Length

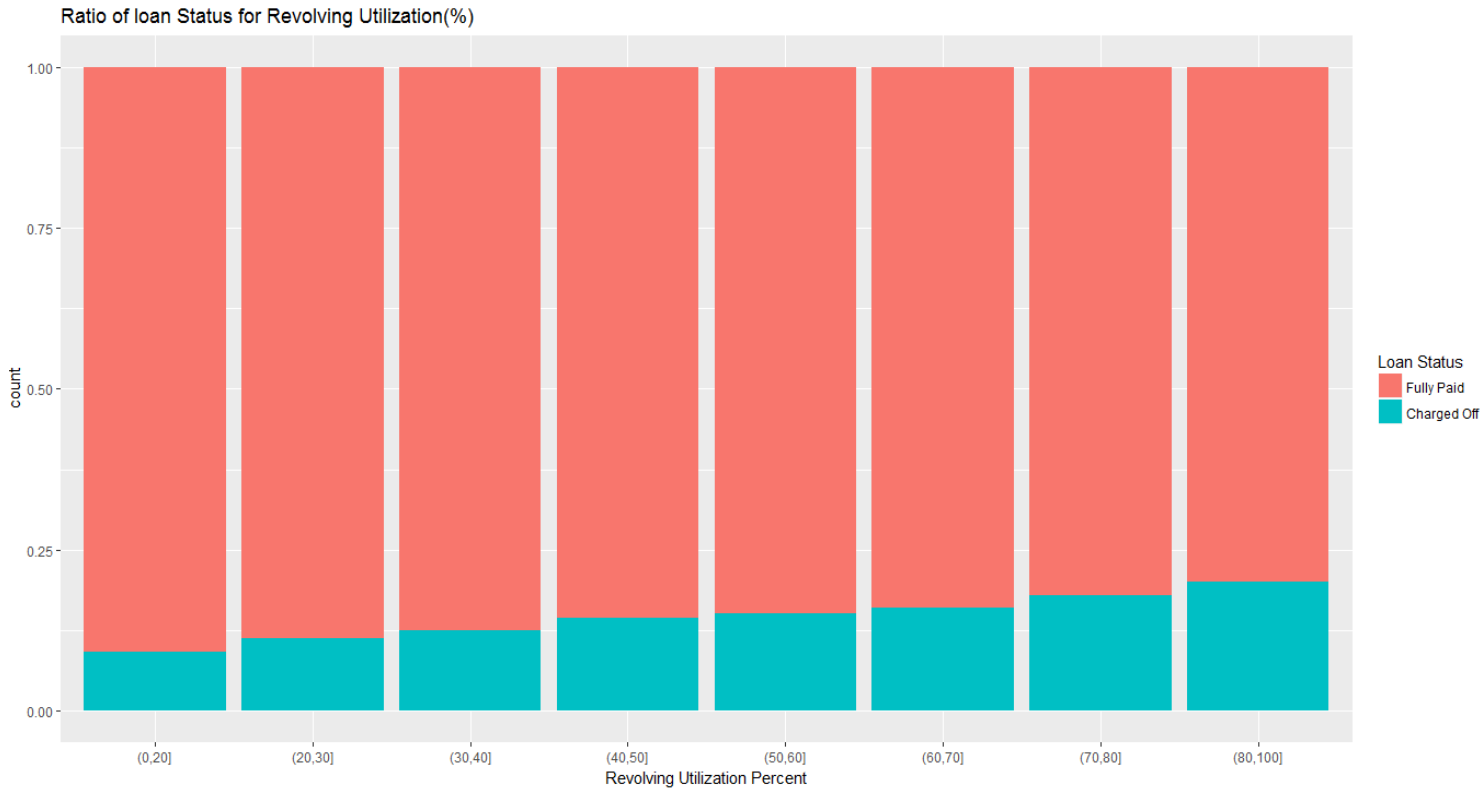
loan$emp_length	loan$loan_status Charged Off	Fully Paid	Row Total
< 1 year	639 0.142	3869 0.858	4508 0.117
1 year	456 0.144	2713 0.856	3169 0.082
10+ years	1331 0.157	7157 0.843	8488 0.220
2 years	567 0.132	3724 0.868	4291 0.111
3 years	555 0.138	3457 0.862	4012 0.104
4 years	462 0.138	2880 0.862	3342 0.087
5 years	458 0.143	2736 0.857	3194 0.083
6 years	307 0.142	1861 0.858	2168 0.056
7 years	263 0.154	1448 0.846	1711 0.044
8 years	203 0.141	1232 0.859	1435 0.037
9 years	158 0.129	1068 0.871	1226 0.032
n/a	228 0.221	805 0.779	1033 0.027
Column Total	5627	32950	38577



Assuming NA as Unemployed/Self-employed/Not mentioned, it clearly depicts that the unemployed are most likely to default.

Analysis: Revolving Utilization Percent

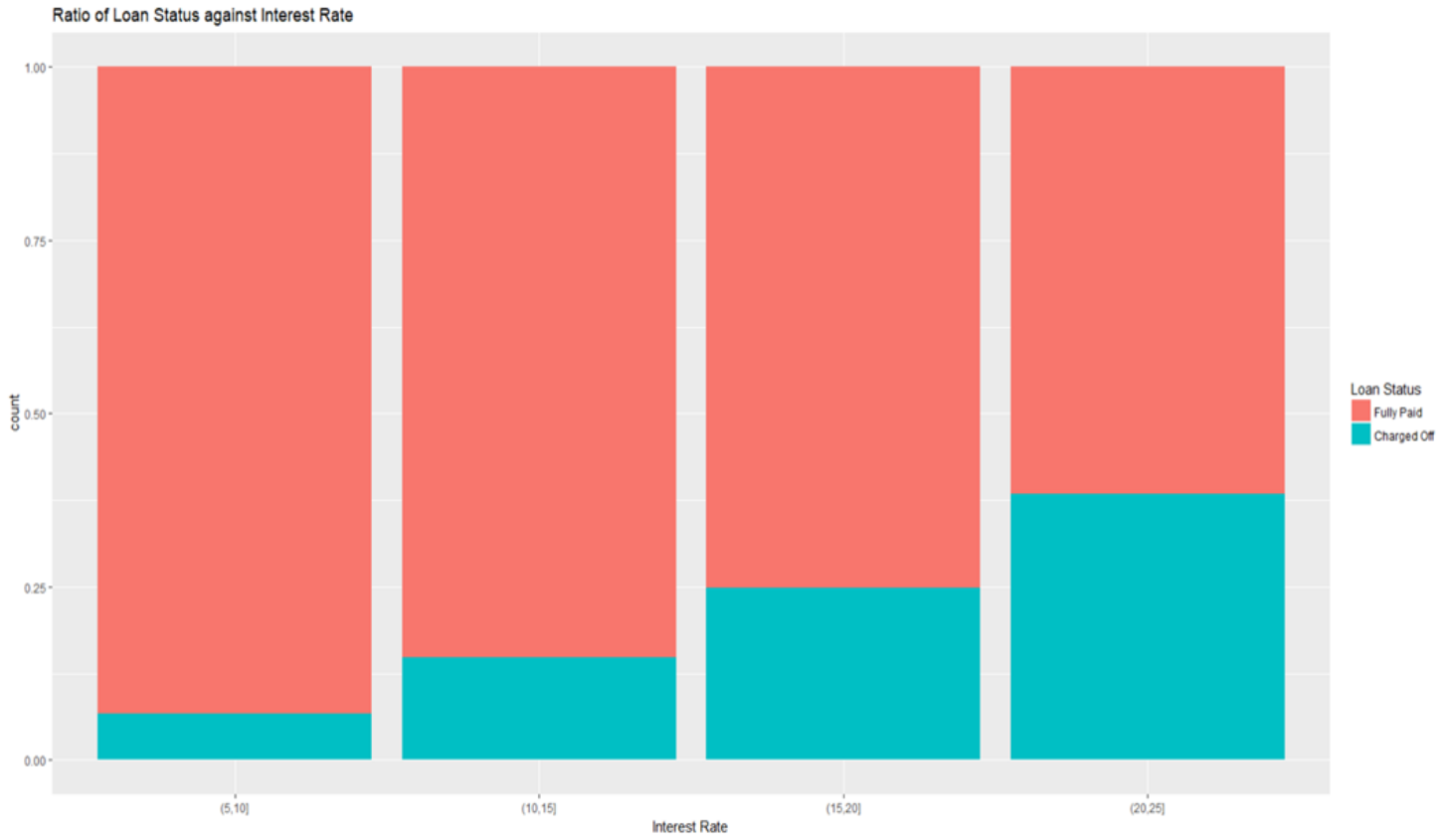
loan$revol_util_bkt	loan$loan_status Charged Off	Fully Paid	Row Total
(0,20]	626	6180	6806
	0.092	0.908	0.181
(20,30]	414	3290	3704
	0.112	0.888	0.099
(30,40]	501	3500	4001
	0.125	0.875	0.106
(40,50]	606	3618	4224
	0.143	0.857	0.112
(50,60]	630	3554	4184
	0.151	0.849	0.111
(60,70]	659	3458	4117
	0.160	0.840	0.110
(70,80]	697	3200	3897
	0.179	0.821	0.104
(80,100]	1334	5306	6640
	0.201	0.799	0.177
Column Total	5467	32106	37573



The percentage of defaulting loans steadily increases with utilization rate. Borrowers with high utilization rates are more likely to have high fixed credit card payments which might affect their ability to repay their loans.

Analysis: Interest Rate

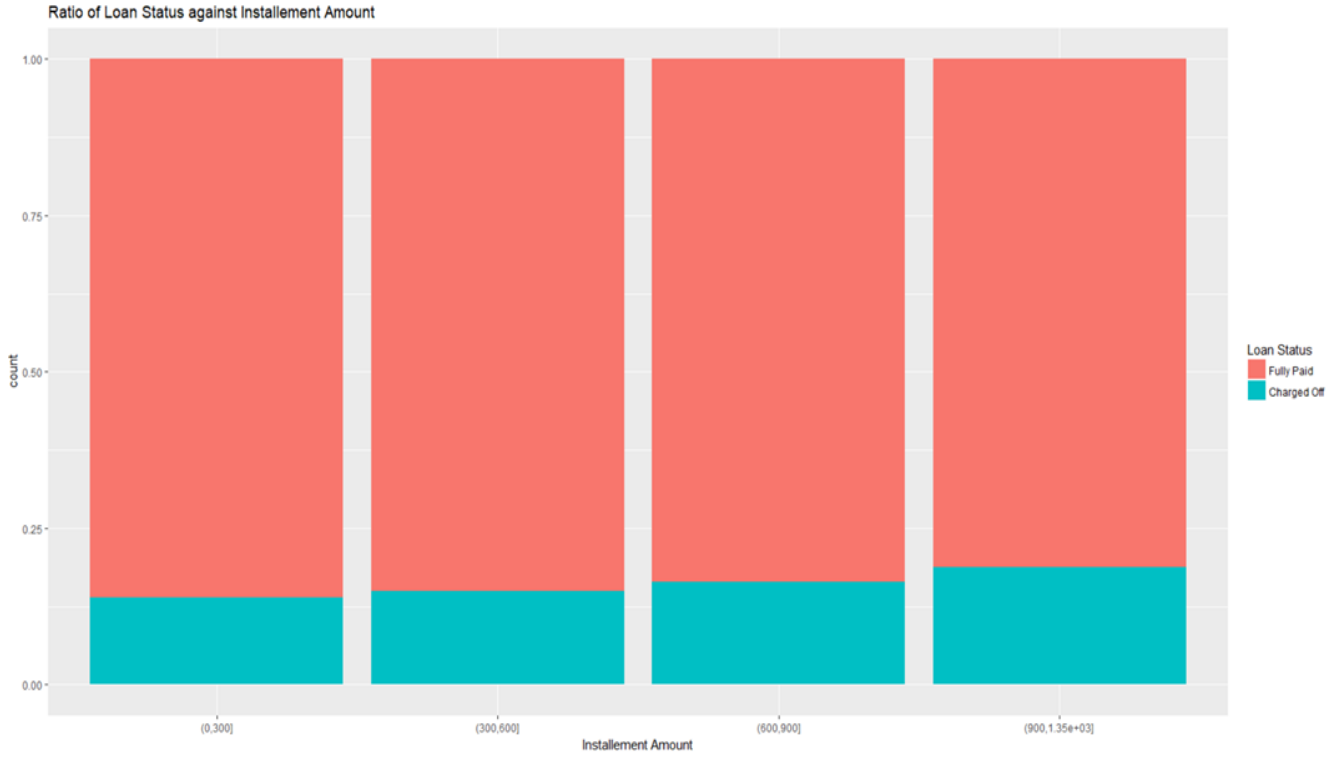
loan$int_rate_bkt	loan$loan_status Charged Off	Fully Paid	Row Total
(5,10]	830	11486	12316
	0.067	0.933	0.319
(10,15]	2707	15558	18265
	0.148	0.852	0.473
(15,20]	1794	5432	7226
	0.248	0.752	0.187
(20,25]	296	474	770
	0.384	0.616	0.020
Column Total	5627	32950	38577



Number of Defaulters increase with increase in interest rates.

Analysis: Instalment Amount

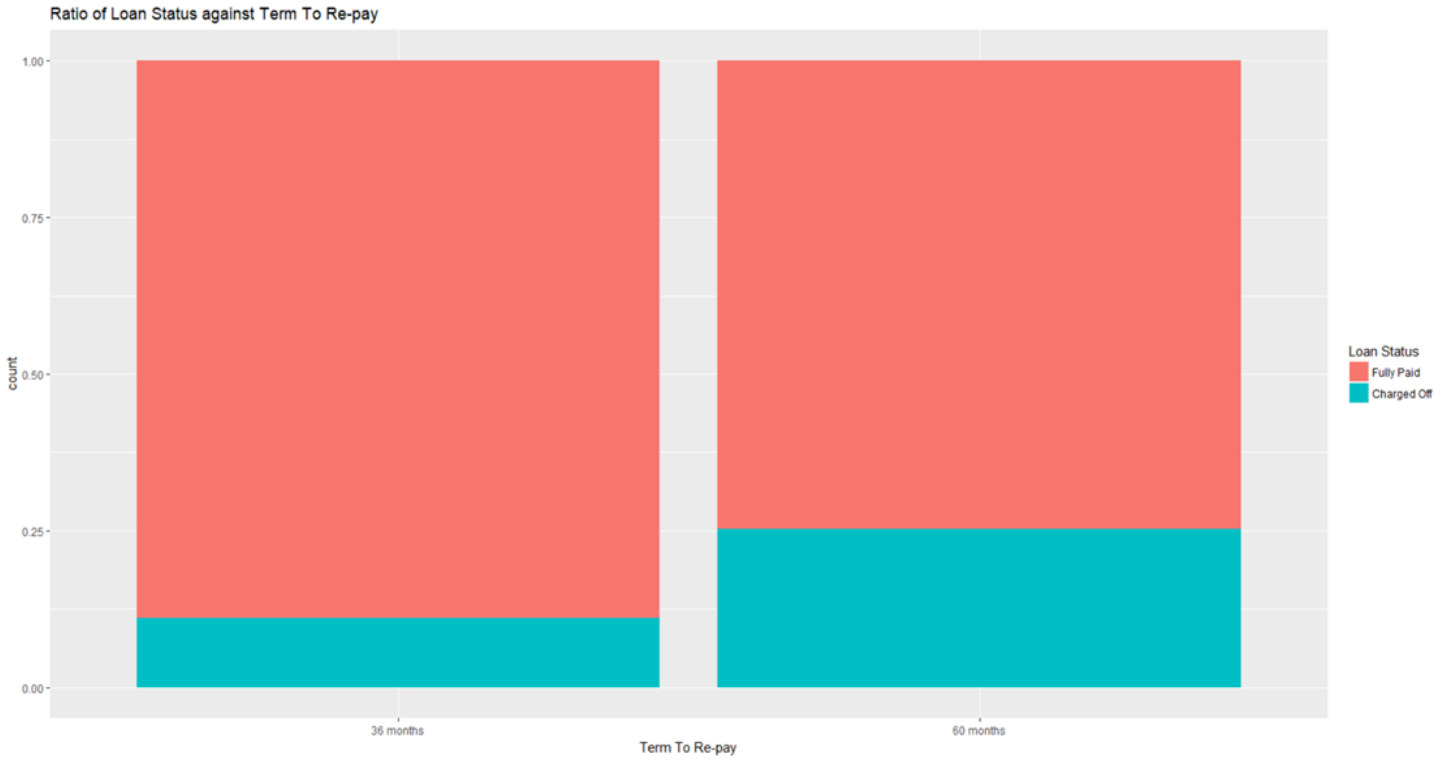
loan$installment_bkt	loan$loan_status: Charged Off	Fully Paid	Row Total
(0,300]	2881	17801	20682
	0.139	0.861	0.536
(300,600]	2042	11645	13687
	0.149	0.851	0.355
(600,900]	606	3082	3688
	0.164	0.836	0.096
(900,1350]	98	422	520
	0.188	0.812	0.013
Column Total	5627	32950	38577



As the instalment amount increases the defaulter's percentage increases as well.

Analysis: Term To Re-pay

	loan$loan_status		
loan$term	Charged Off	Fully Paid	Row Total
36 months	3227	25869	29096
	0.111	0.889	0.754
60 months	2400	7081	9481
	0.253	0.747	0.246
Column Total	5627	32950	38577



The plot clearly shows that 36 month term has more probability of getting Charged Off.

EDA Case Study Conclusion

Grade and **Sub-grade** variables provide the most predictive power for determining expected loan performance.

A large number of the other variables also provide strong indications of expected performance, most descriptive among them are:

- Grades/ Sub-grades /Interest Rates
- Home Ownership Status
- Loan Purpose
- Loan Amount
- Annual Income
- Verified Income Status
- Employment Length
- Revolving Utilization Percent
- Instalment Amount
- Term to Re-pay

Verified income status and show results opposite from what we would expect. This is likely due to increased standards on borrowers with poorer credit history, so all else equal we see outperformance in these loans.