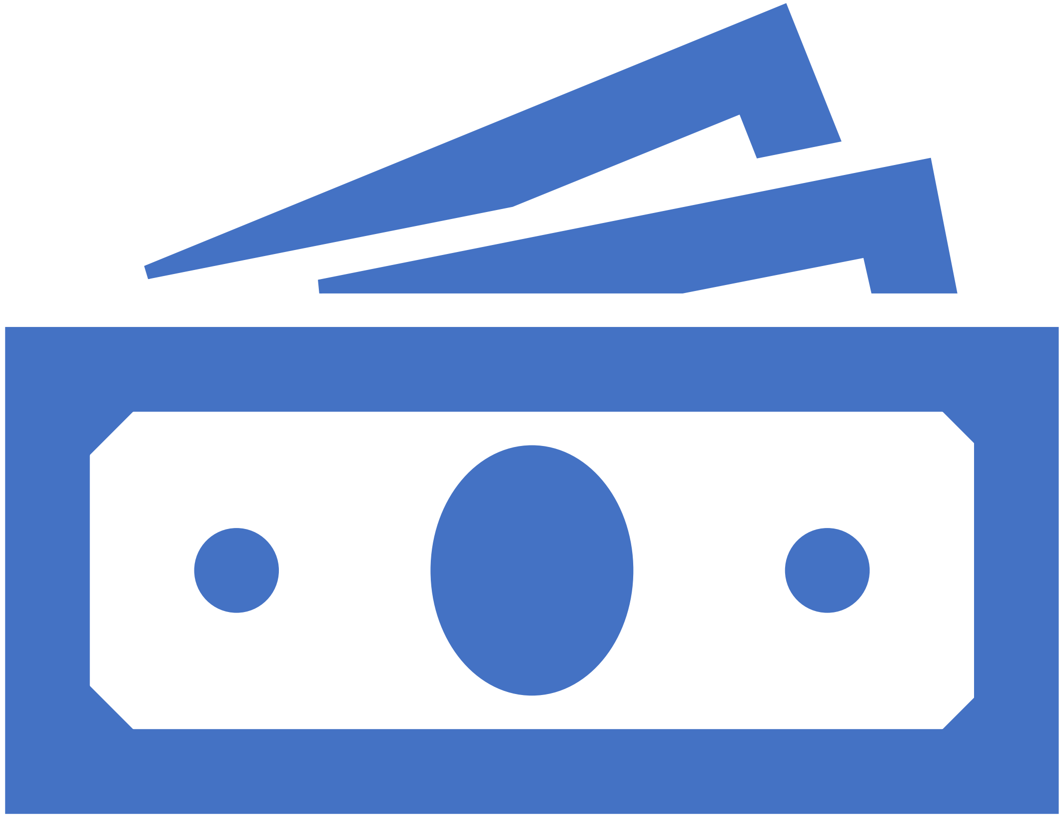
ML C65

Lending Club Case Study



Jyoti Panda
Prajwal K R

Business Problem and Objectives

What is the Lending Club?

- This 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.

What is the business problem?

Company lending loans to ‘risky’ applicants is the largest source of financial loss (called credit loss).
Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

→

When the company receives a loan application, the company must decide for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision.

→

If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

↓

If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.

→

The data contains information about past loan applicants and whether they ‘defaulted’ or not. The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate of default.

→

Since the company operated online, making customers happy and being clear is important. So, we need to be able to explain why we decide to approve or deny a loan.

What is the Business Objective?

Identify the factors associated with credit risk in the form of business insights.

To be able to identify risky loan applicants, so that such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

We want to showcase the driving factors (or driver variables) behind loan default, i.e. (Loan_status = 'charged off'), the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Data Cleaning & Manipulation

Treating null values

Dropping Columns unnecessary for our analysis

Original dataset rows and columns

- Rows - 39717, Columns – 111

After Cleansing Rows and Columns

- Rows - 39717, Columns – 28

Data Preparation

loan status = "Current" doesn't help our analysis for approving or rejecting application, so filtering this data.

Similarly looking at other columns for our analysis to concur how it can influence our decision making.

Removing Outliers from columns & creating new columns wherever necessary

Correcting Date-Time formats (Y2K corrected, Formatting & datatype changes) , Splitting Month and Year to separate columns if needed, calculating months since last credit line to date.

EDA – Exploratory Data Analysis

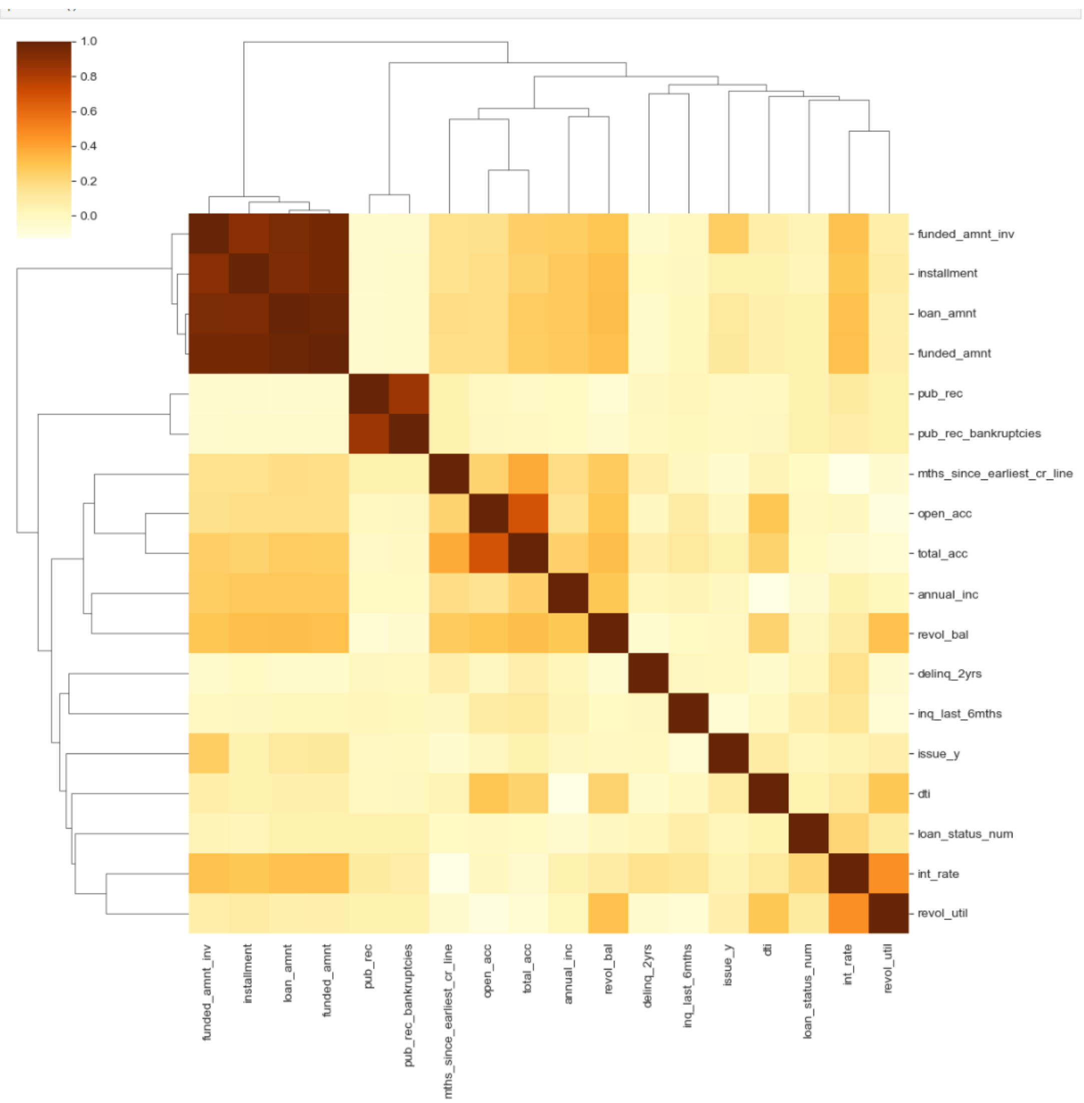
Univariate Analysis

As data is now cleaned , we then proceed with EDA Univariate& Bi variate Analysis.

Let us start by understanding the correlation between the different numeric fields and see if they are related (high correlation values)

Numerical-Correlations

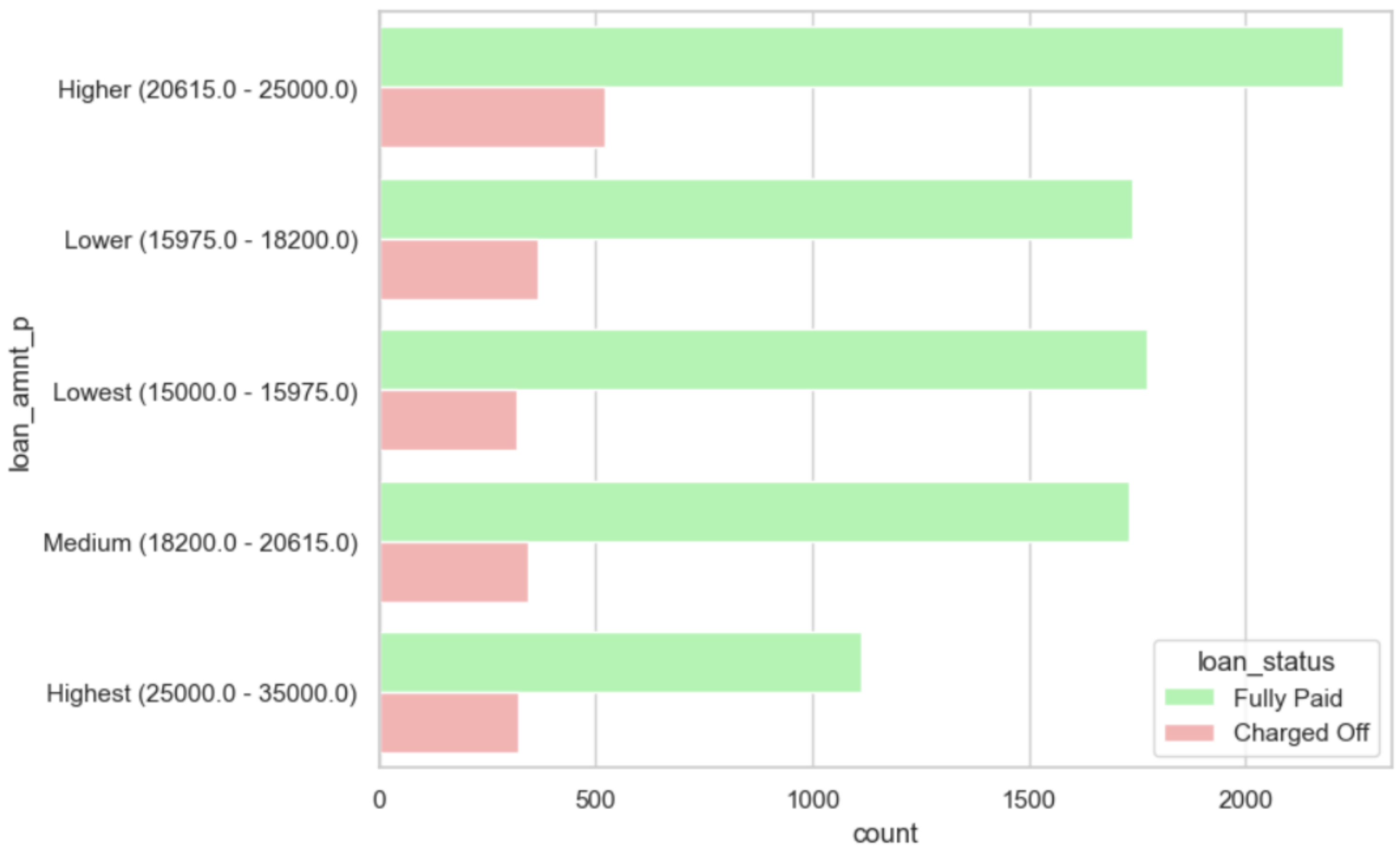
Since we know darker the value higher the correlation , we can clearly see ‘loan_amnt’, ‘funded_amnt’, ‘funded_amnt_inv’ and ‘installment’ have huge correlation. These fields are proportional to each other. Next, the public records related fields ‘pub_rec’ & ‘pub_rec_bankruptcies’ and number of accounts related fields ‘open_acc’ & ‘total_acc’ are correlated.



Columns Histogram

Loan status vs Numerical continuous variables

Let us now compare the loan_status fields with all the numerical variable. Since the analysis for this category will be similar, we will construct a common function which can be used for all.



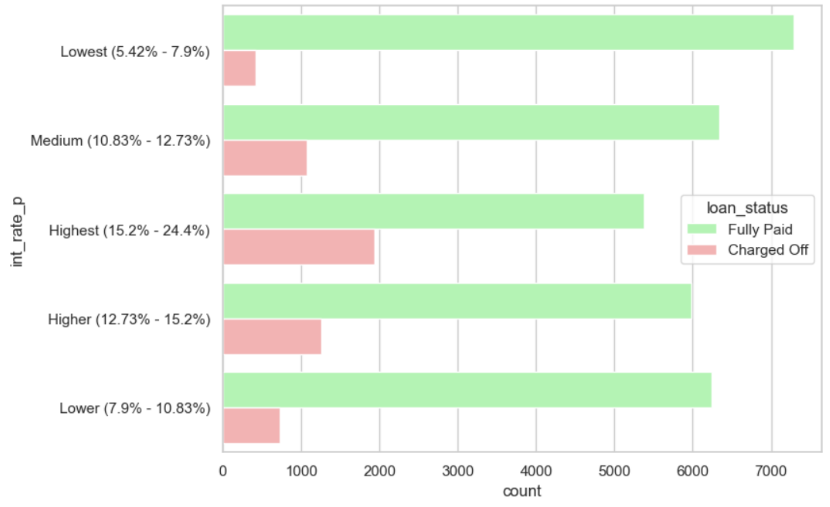
Loan Status v/s Loan amount

From the plot we can conclude that higher the loan amount, greater the chance of the loan getting default.

Loan Status v/s Interest Rate

int_rate_p	Charged off %	Record count
Highest (15.2% - 24.4%)	0.265092	7322
Higher (12.73% - 15.2%)	0.173529	7238
Medium (10.83% - 12.73%)	0.144591	7414
Lower (7.9% - 10.83%)	0.103993	6962
Lowest (5.42% - 7.9%)	0.054892	7706

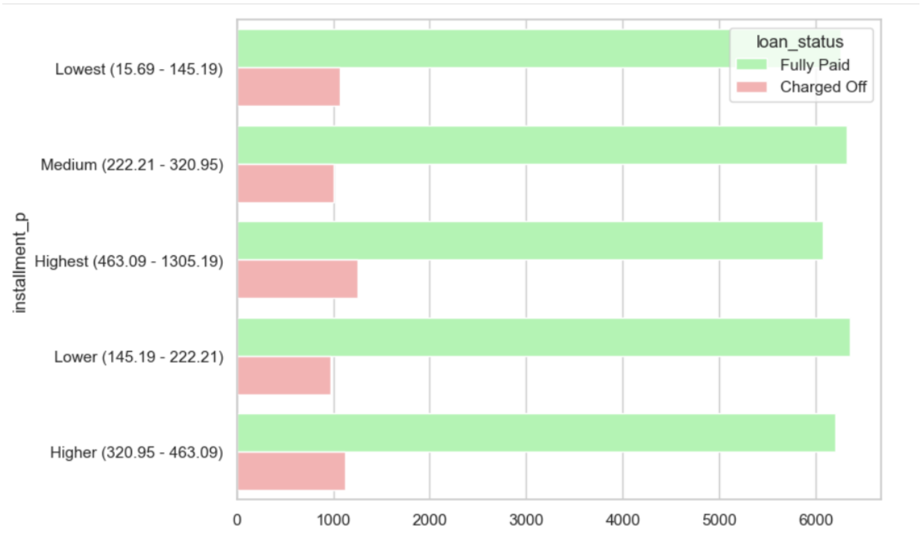
Similarly , Higher the interest rate leads to higher 'Charged off %'



Loan Status v/s Installment

The data and plot shows that higher installment amounts shows slightly higher default percentages.

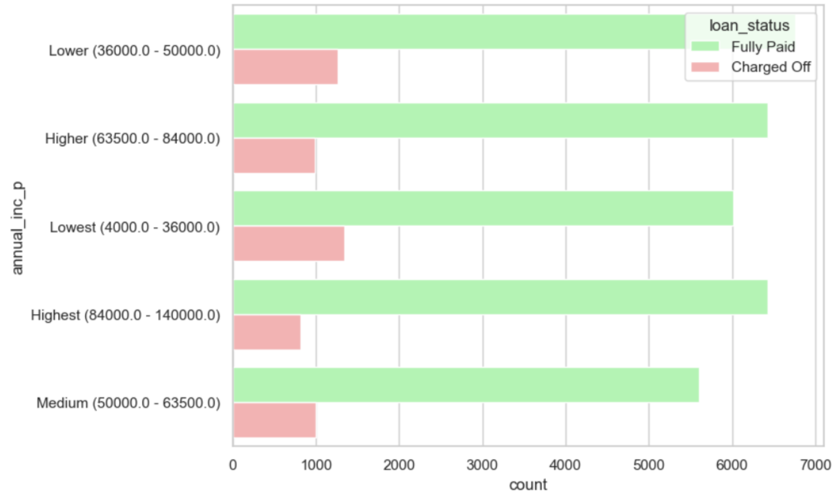
installment_p	Charged off %	Record count
Highest (463.09 - 1305.19)	0.170965	7329
Higher (320.95 - 463.09)	0.152744	7326
Lowest (15.69 - 145.19)	0.146092	7331
Medium (222.21 - 320.95)	0.136717	7329
Lower (145.19 - 222.21)	0.132524	7327



Loan Status v/s Annual Income

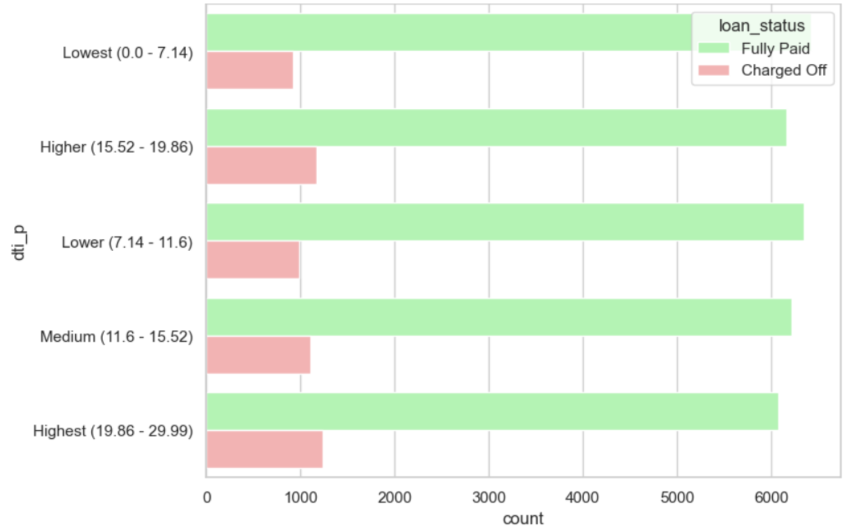
annual_inc_p	Charged off %	Record count
Lowest (4000.0 - 36000.0)	0.182967	7362
Lower (36000.0 - 50000.0)	0.157718	8027
Medium (50000.0 - 63500.0)	0.151515	6600
Higher (63500.0 - 84000.0)	0.132946	7409
Highest (84000.0 - 140000.0)	0.112921	7244

Higher the income higher the repayment percentile, lower income gets more loan charged off



Loan Status v/s DTI

dti_p	Charged off %	Record count
Highest (19.86 - 29.99)	0.168854	7314
Higher (15.52 - 19.86)	0.159667	7334
Medium (11.6 - 15.52)	0.150505	7322
Lower (7.14 - 11.6)	0.134497	7331
Lowest (0.0 - 7.14)	0.125596	7341

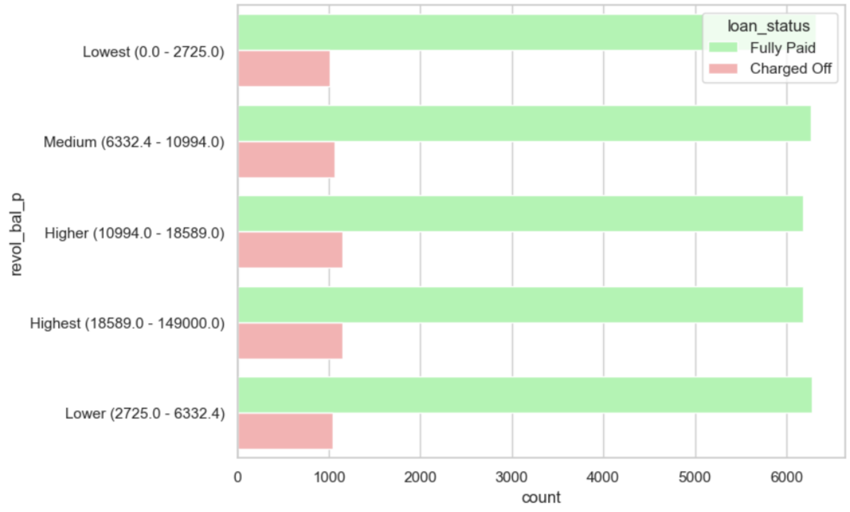


Higher DTI (debt to income ratio) will lead to higher charged off %

Loan Status v/s Revolving Balance

revol_bal_p	Charged off %	Record count
Highest (18589.0 - 149000.0)	0.156659	7328
Higher (10994.0 - 18589.0)	0.156250	7328
Medium (6332.4 - 10994.0)	0.145313	7329
Lower (2725.0 - 6332.4)	0.142467	7321
Lowest (0.0 - 2725.0)	0.138359	7336

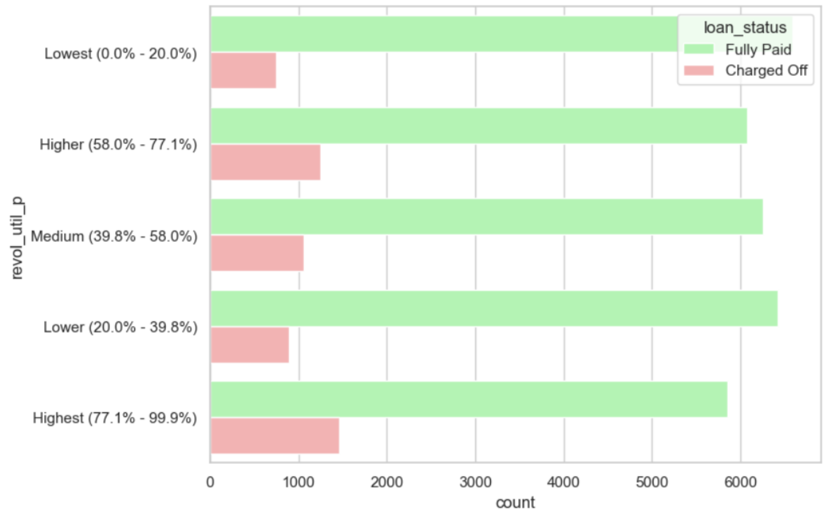
This shows the total credit revolving balances slightly influence the default percentage. Higher the revolving balance, bigger the chance of the loan getting defaulted.



Loan Status v/s Revolving Line Utilization

This data shows that the revolving line utilization rate has a large impact to the default percentage. When this increases, the charged off percentage rises.

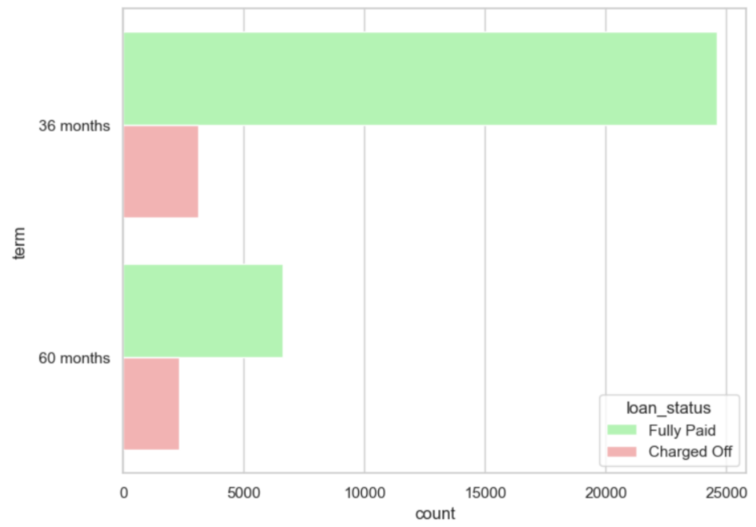
revol_util_p	Charged off %	Record count
Highest (77.1% - 99.9%)	0.199206	7304
Higher (58.0% - 77.1%)	0.170329	7327
Medium (39.8% - 58.0%)	0.145122	7318
Lower (20.0% - 39.8%)	0.121771	7317
Lowest (0.0% - 20.0%)	0.101528	7328



Loan Status v/s Term

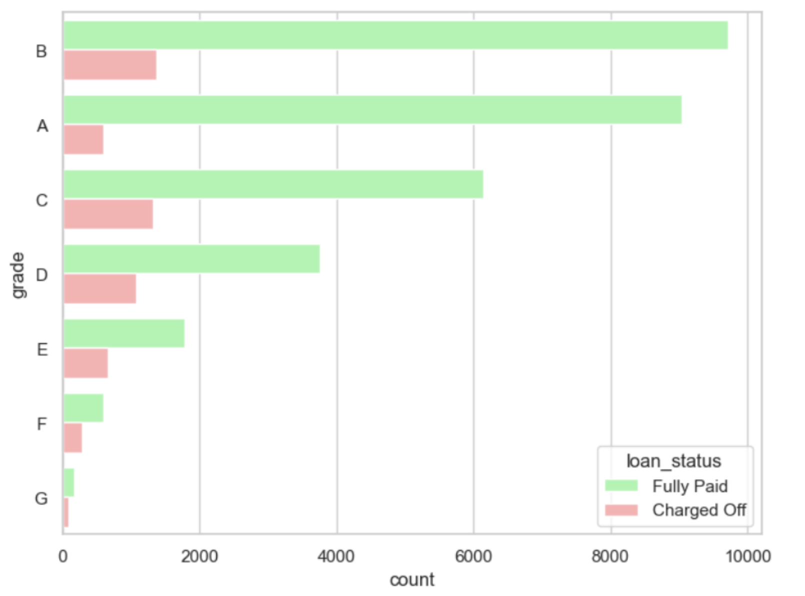
For loans with 5-year repayment term, the default percent is 25%. And for 3-year loan repayment term, the default is only for 11% of the cases. Therefore, loan repayment term plays a factor in judging the default rate lower loan terms default less.

term	Charged off %	Record count
60 months	0.257953	8928
36 months	0.112326	27714

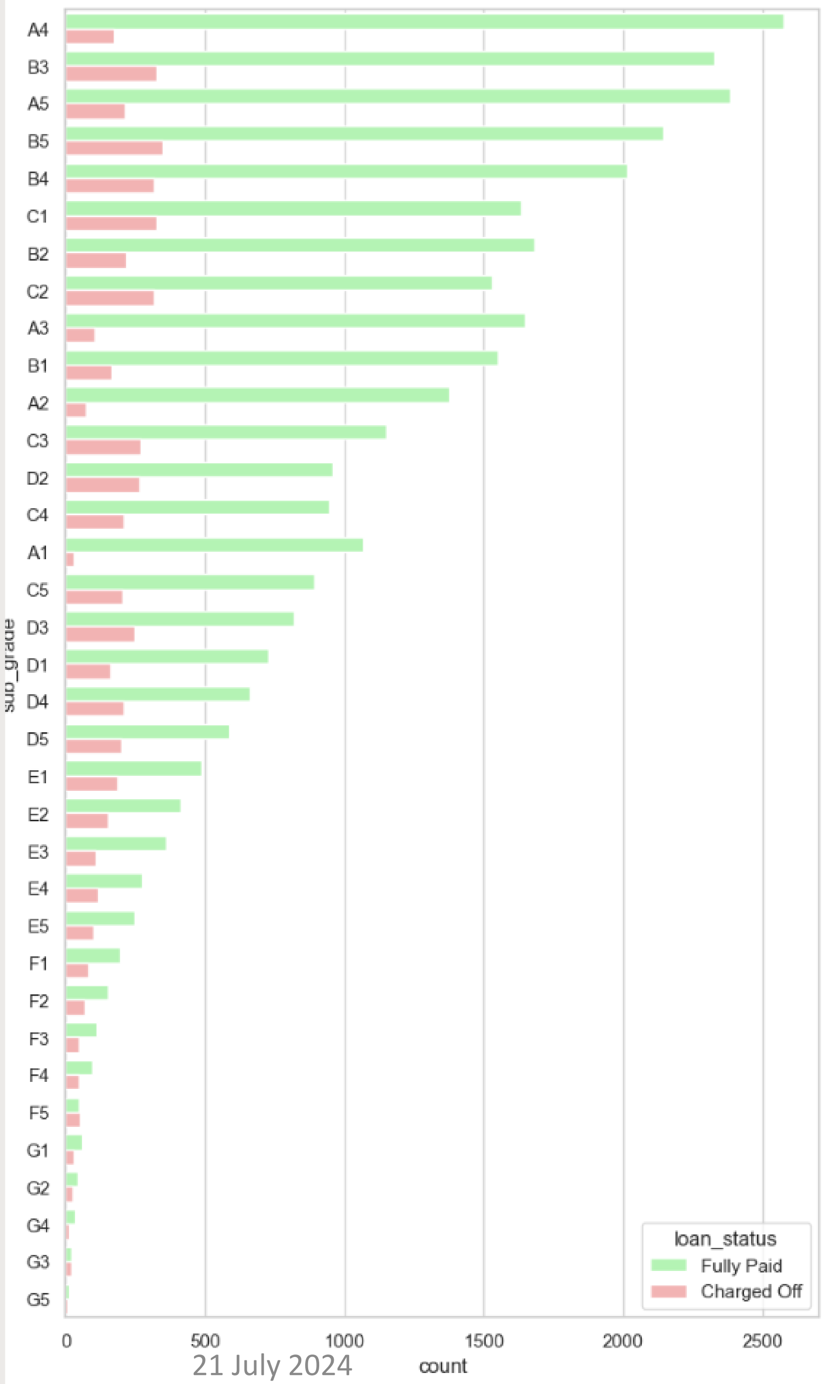


Loan Status v/s Grade

grade	Charged off %	Record count
G	0.363985	261
F	0.326185	886
E	0.269530	2445
D	0.223417	4834
C	0.177076	7477
B	0.124009	11096
A	0.061495	9643



We can clearly see that loan grades having highest default percentages. G, F, E and D form grades where default rate is much higher than others.



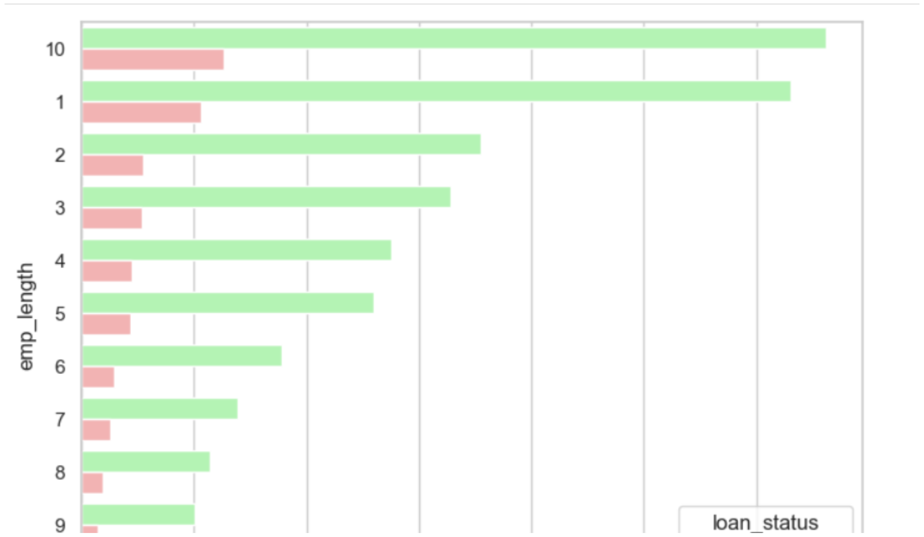
Loan Status v/s Sub Grade

The table shows the loan sub-grade versus the default percentage. The G3 and F5 subcategories have above 40% default rate, with F5 being more than 50% default. This field is a clear indicator of the default percent.

sub_grade	Charged off %	Record count
F5	0.510204	98
G3	0.475000	40
G5	0.409091	22
G2	0.393939	66
G1	0.340909	88
F4	0.330935	139
F2	0.308756	217
E4	0.298469	392
F3	0.294872	156
F1	0.289855	276
E5	0.286957	345
E1	0.273810	672
E2	0.265487	565
D5	0.253485	789
G4	0.244444	45
D4	0.239954	871
E3	0.231423	471
D3	0.231421	1063
D2	0.215863	1223
C3	0.188293	1418
C5	0.187044	1096
D1	0.181306	888
C4	0.180952	1155
C2	0.171088	1847
C1	0.166752	1961
B5	0.139647	2492
B4	0.136734	2333
B3	0.122833	2654
B2	0.114271	1899
B1	0.096624	1718
A5	0.081955	2599
A4	0.062955	2748
A3	0.058924	1748
A2	0.050999	1451
A1	0.027347	1097

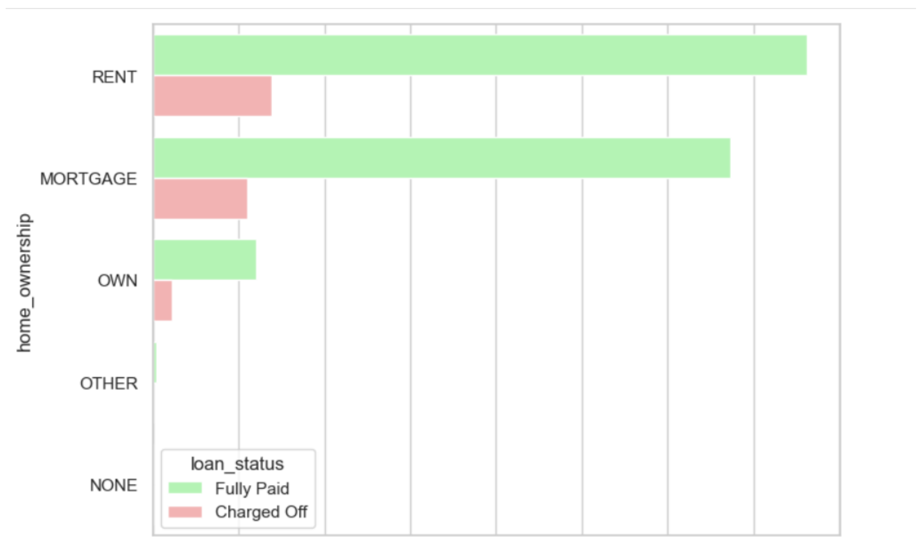
Loan Status v/s Employment Length

0	0.221893	1014
10	0.160959	7884
7	0.153378	1643
5	0.144875	3044
8	0.144668	1341
1	0.144640	7370



Although there is nothing much to conclude from the data, we can see having no employment will lead to Defaulted payment

Loan Status v/s Home Ownership

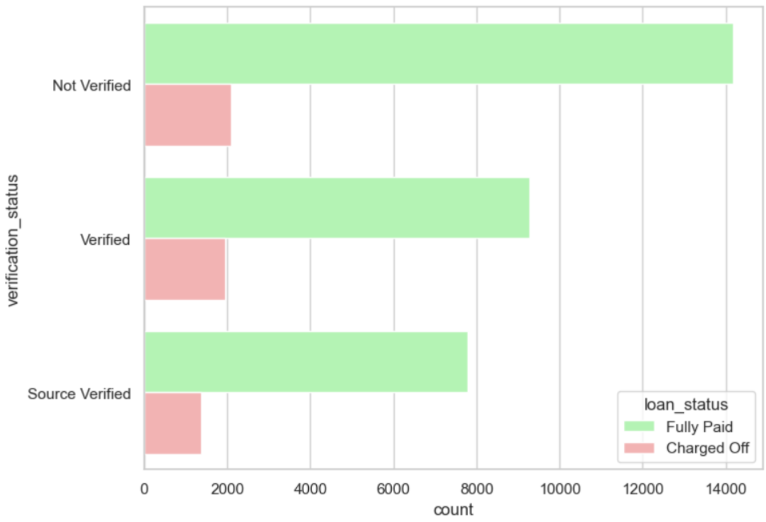


home_ownership	Charged off %	Record count
OTHER	0.184783	92
RENT	0.153808	18029
OWN	0.152174	2852
MORTGAGE	0.139921	15666

Although there is not much to conclude from the data, as it says “Other”. We can see most people either have mortgage or rent their home rather than owning it outright.

Loan Status v/s Verification Status

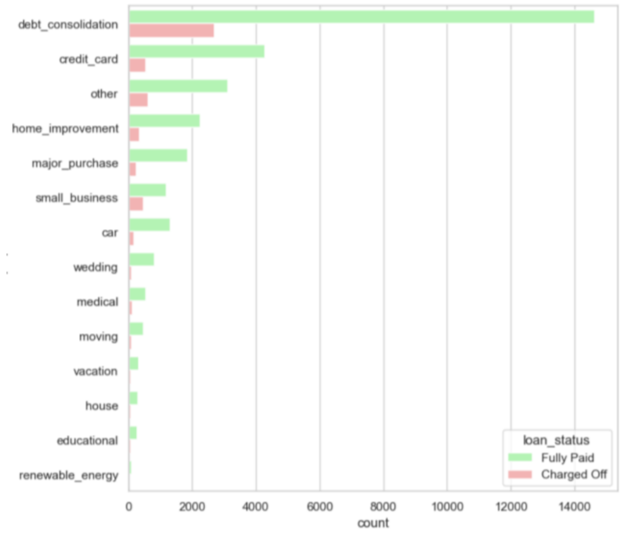
verification_status	Charged off %	Record count
Verified	0.173711	11214
Source Verified	0.150055	9150
Not Verified	0.128701	16278



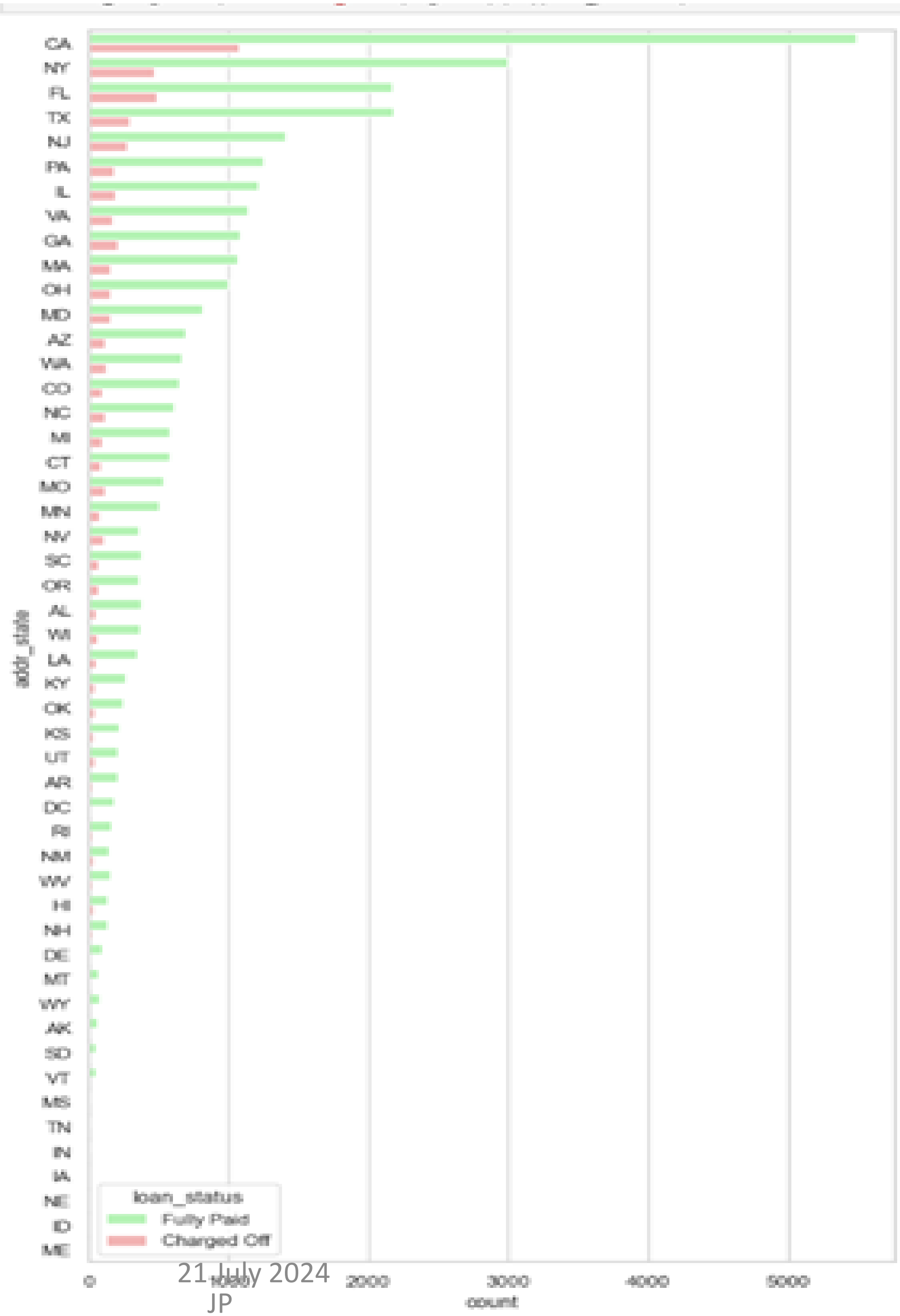
Interestingly verified applicants have more defaulted % which means the verification process is not being done in a proper/correct manner.

Loan Status v/s Purpose

From the analysis it is evident that the loans taken for small businesses, renewable energy sector and education are the riskier ones. Also, we see maximum number of people take loan to repay credit card or for the purpose of Debt consolidation.



purpose	Charged off %	Record count
small_business	0.279729	1623
renewable_energy	0.189474	95
educational	0.173077	312
other	0.163690	3696
moving	0.163636	550
house	0.162722	338
medical	0.162500	640
debt_consolidation	0.154615	17301
vacation	0.142466	365
home_improvement	0.126027	2555
car	0.109807	1448
credit_card	0.108532	4782
major_purchase	0.106112	2045
wedding	0.103139	892



Loan Status v/s Address State

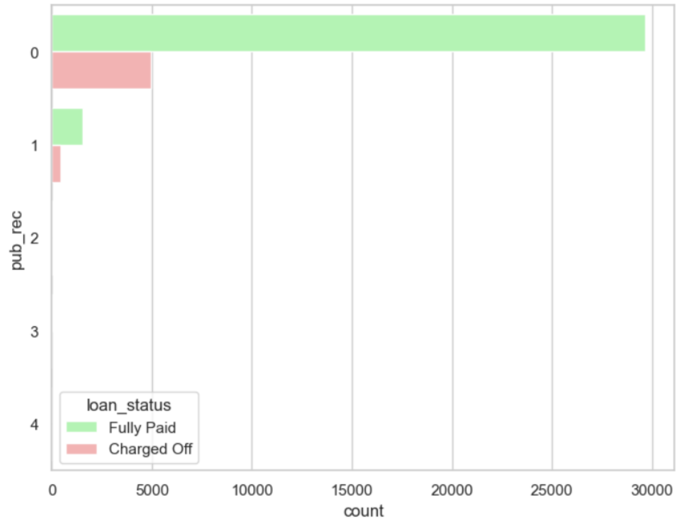
From the graph & table we can see that the some of the risky states are NE, NV , SD , AK , FL , MO

addr_state	Charged off %	Record count
NE	0.600000	5
NV	0.225383	457
ID	0.200000	5
SD	0.193548	62
AK	0.183099	71
FL	0.182881	2652
HI	0.175000	160
MO	0.171561	647
NM	0.169492	177
CA	0.163567	6550
OR	0.163551	428
GA	0.161742	1286
MD	0.161491	966
NJ	0.160981	1671
NH	0.159236	157
WA	0.158163	784
NC	0.157163	719
UT	0.157025	242
KY	0.148026	304
MI	0.146628	682
SC	0.145125	441
AZ	0.144444	810
WI	0.142518	421
OK	0.142349	281
MN	0.137990	587
IL	0.136655	1405
NY	0.135143	3463
CT	0.134128	671
VA	0.133028	1308
MT	0.132530	83
OH	0.131648	1147
CO	0.129555	741
RI	0.129032	186
LA	0.127182	401
PA	0.124475	1422
WV	0.124260	169
MA	0.124069	1209
AL	0.122931	423
KS	0.122449	245
TX	0.119144	2476
TN	0.117647	17
AR	0.113537	229
VT	0.113208	53
DE	0.110092	109
MS	0.105263	19
DC	0.066667	195
WY	0.050633	79

Loan Status v/s Public Record

pub_rec	Charged off %	Record count
1	0.228499	1965
2	0.217391	46
0	0.143171	34623

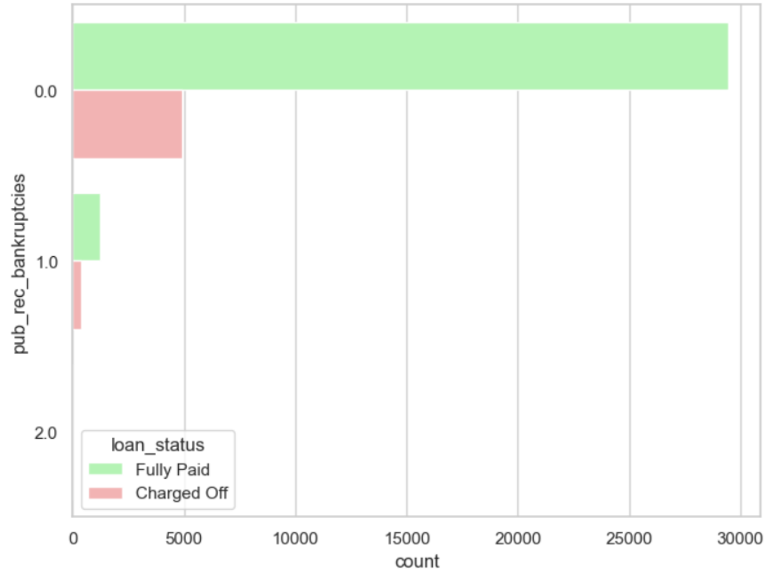
The data and graph clearly shows that individuals with non-zero derogatory public records have higher chances of charged off.



Loan Status v/s Public Bankruptcy Records

pub_rec_bankruptcies	Charged off %	Record count
2.0	0.400000	5
1.0	0.225766	1599
0.0	0.143717	34380

Higher the number of public bankruptcy records, bigger the chance of defaulting the loan.

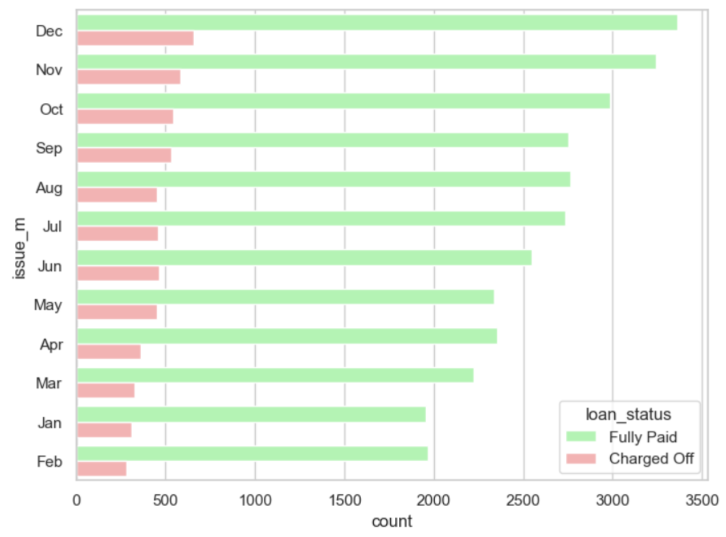


Loan Status v/s Loan Issued Month

We can clearly see December is the month which has the highest number of loan applications per year and have the biggest default ratio.

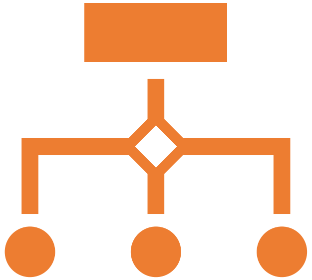
People might be taking loan to travel/party during year end holidays (Christmas and New Year) and then are not able to pay back.

Month of May is also another one, which is during the summer break and right before the Memorial day and Independence day holidays in US when people plan their vacations.

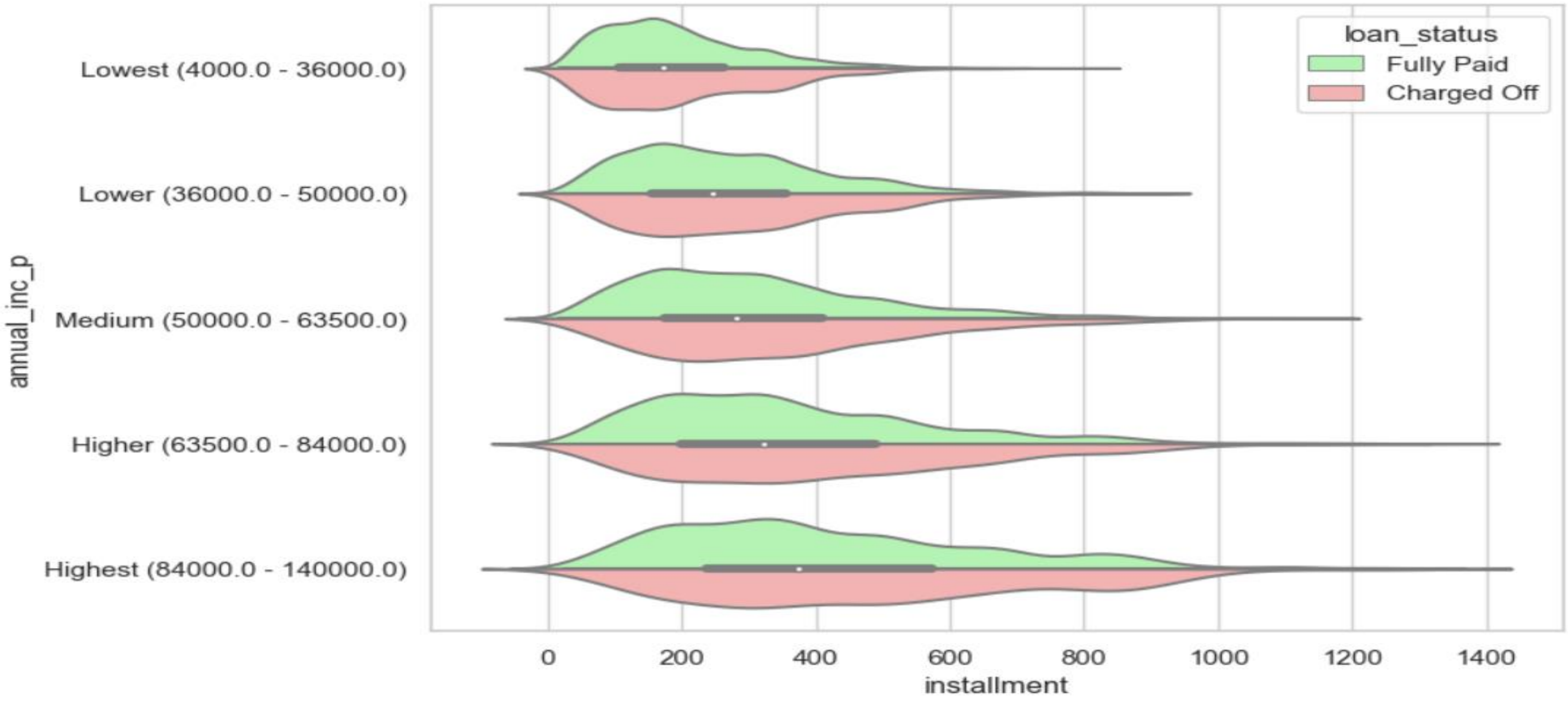


issue_m	Charged off %	Record count
Dec	0.163225	4019
May	0.162482	2788
Sep	0.162154	3287
Oct	0.154304	3532
Jun	0.153258	3008

EDA – Cont.

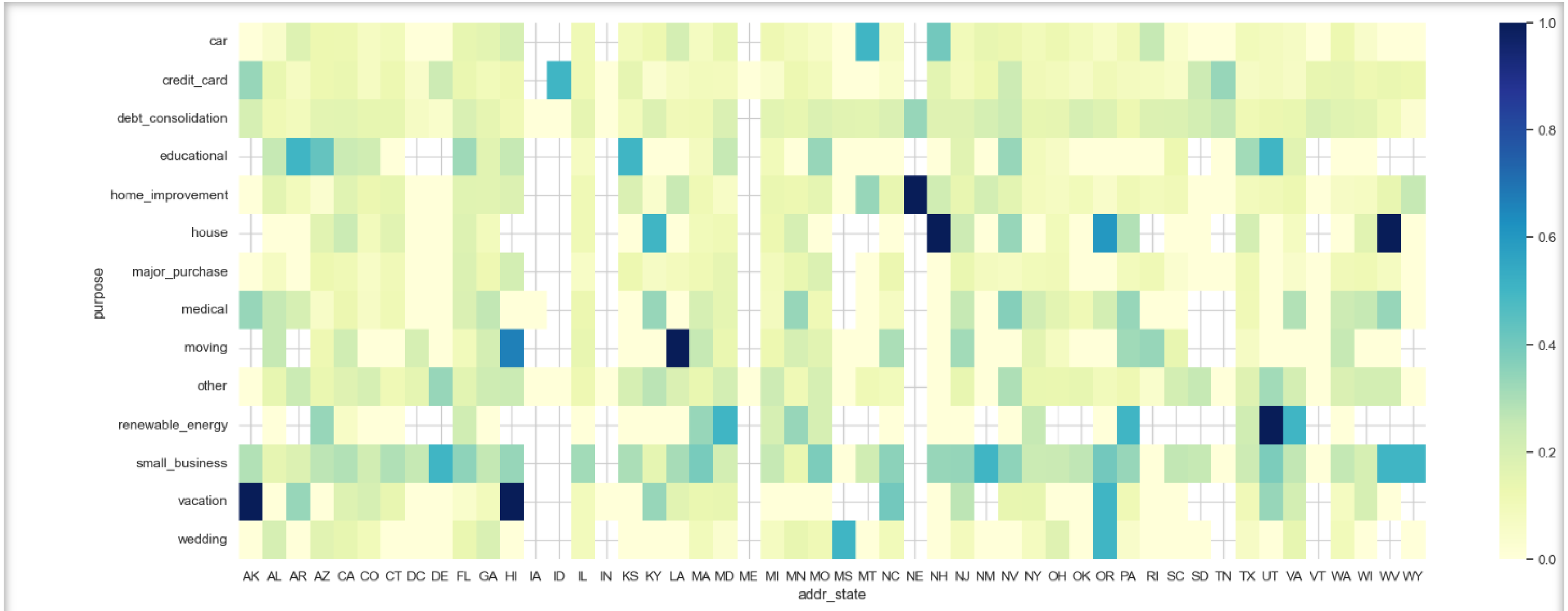


Now that we have analyzed each of the variables and its impact on the loan-status, let us take group of variables together and analyze their combined effect on the loan-status. These categories are based on our business understanding. The original distribution column shows the average trend in all the data and we compare that with the data after applying our conditions.



Loan Status v/s Installment & Annual Income

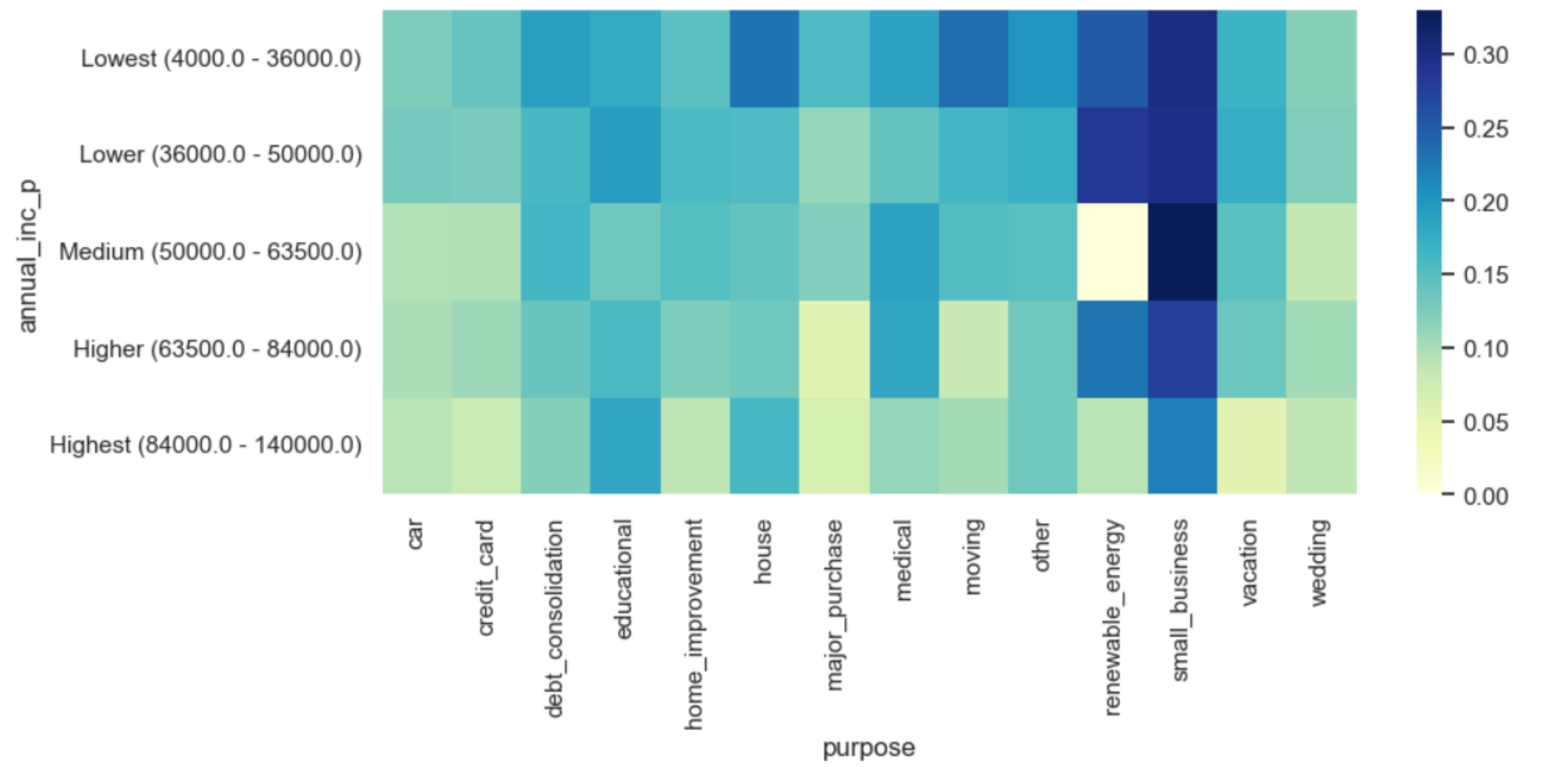
The figure shows that for higher installments for any income group have a greater number of defaults.



Home Ownership & Purpose

As per the above plot, the darker the intersection of 'addr_state' has with the 'purpose' of the loan, the risker the loan application is. Some of the examples are below:

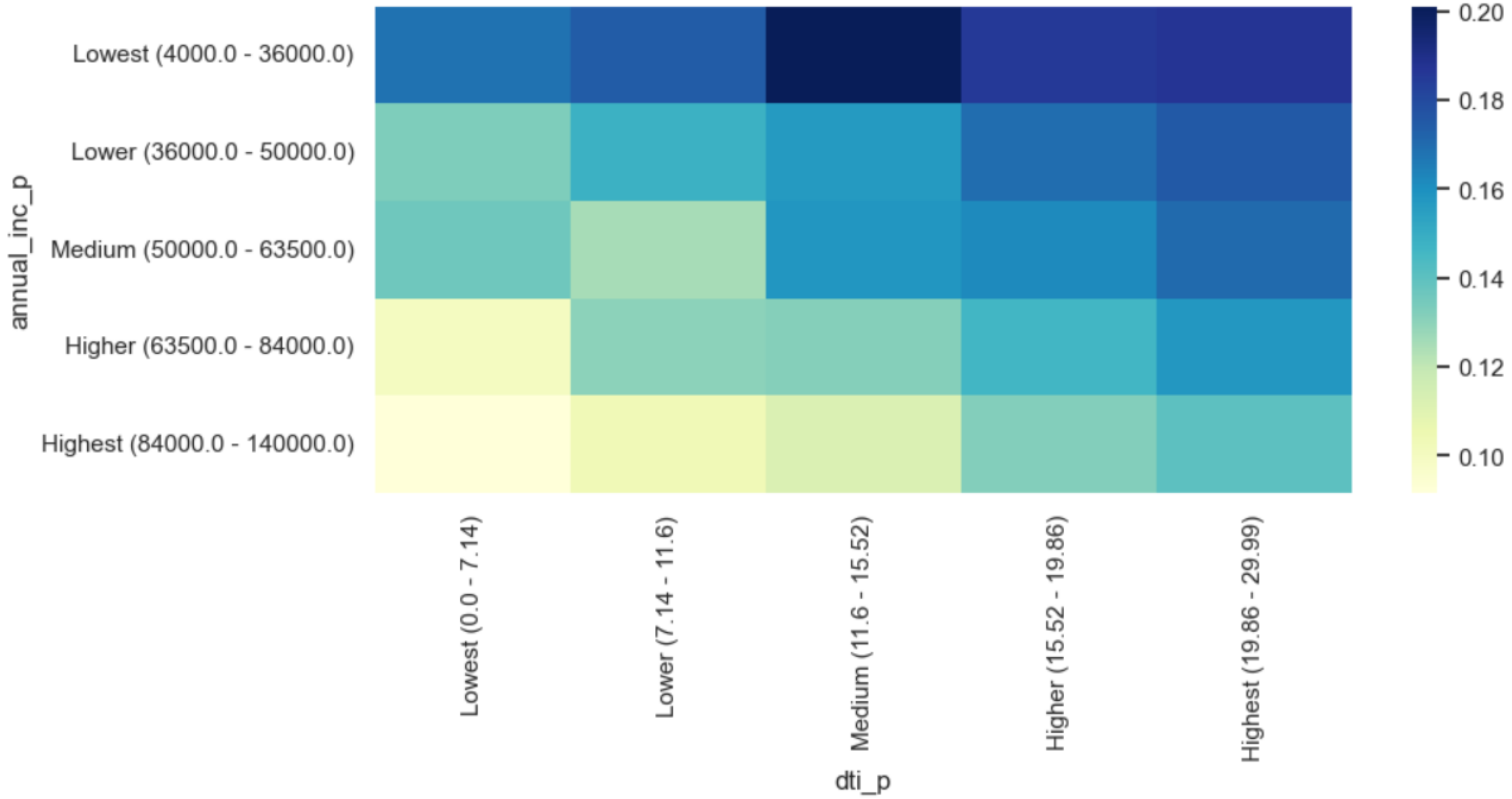
- Vacation loans in AK, HI, OR
- House Loans in NH , WV , OR
- Education loans in AR, KS, UT
- Small business loans in DE, NM, WV, WY
- Wedding loans in MS, OR



Annual Income & Purpose

Plot of various income groups versus the risky purposes of loans for them. Some examples are:

- In general, small business loans are riskier than other types of loans , especially for lowest and medium income groups
- Small business, educational and renewable energy are the top 3 loans purposes with highest credit risk. Credit card, car and 'major_purchase' are the top 3 with lowest credit risk.



Debt to Income ratio & Annual Income

Medium debt-to-income group in the lowest income range is the riskiest when it comes to loan repayment.

EDA Insight

Drivers for Loan Default

These are some of the most impactful drivers for Loan Default as per our analysis.

Minor Impact

- Higher loan amount (above 16K)
- Higher installment amount (above 327)
- Lower annual income (below 37K)
- Higher debt to income ratio (above 15%)
- Applicant’s address state (NV, SD, AK, FL, etc.)
- Loan issue month (May, Sep)

Drivers for Loan Default(cont.)

Heavy impact

- Higher interest rate (above 13%)
- Higher revolving line utilization rate (above 58%)
- Repayment term (5 years)
- Loan grade & sub-grade (D to G)
- Missing employment record
- Loan purpose (small business, renewable energy, educational)
- Derogatory public records (1 or 2)
- Public bankruptcy records (1 or 2)

Drivers for Loan Default(cont.)

- **Combined impact**
    - High loan amount & interest rate for lower income group
    - High installment and longer repayment term
    - Home ownership (other) and loan purpose (car, moving or small business)
    - Residential state and loan purpose
    - Income group and loan purpose

End Of Analysis

Thank You