LENDING CLUB CASE STUDY

A Group Study by

Jonathan DSilva and Pratik Kaviya

INTRODUCTION

The Lending Club is consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company must decide the loan approval based on the applicant's profile.

There are two types of risks, 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

Aim

Identify patterns in data which indicate if a loan applicant is likely to default on loan repayment.

Objective

Based on the data available from the consumer finance company, understand the driving variables available at the time of the application and historical data to identify strong indictors of default.

Approach

Utilize the Exploratory Data Analysis concepts learnt to undertake portfolio and risk assessment in the banking and finance sector.

PROBLEM STATEMENT

- Data Sourcing
- Data Cleaning
- Univariate Analysis
- Bivariate Analysis
- Observation and Conclusions

EXPLORATORY DATA ANALYSIS

DATA SOURCING

- We have received the historical data in a CSV format, loan.csv and the data dictionary, Data_Dictionary.xlsx

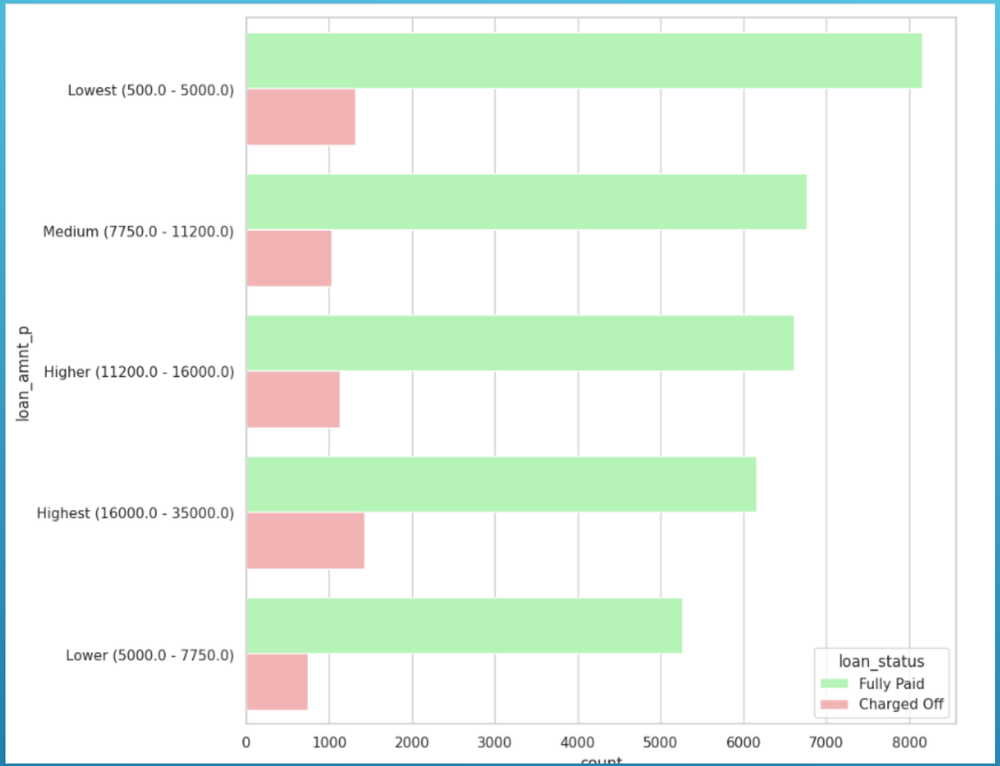
DATA CLEANING

- Identify and drop columns missing more than 30% of data.
- Identify and drop columns with single value as they will be of no use for our analysis.
- Drop the columns that were created after the loan application was approved and do not apply to our analysis.
  "id","member_id","url","zip_code","out_prncp","out_prncp_inv","total_pymnt","total_pymnt_inv", "total_rec_prncp","total_rec_int","total_rec_late_fee","recoveries","collection_recovery_fee", "last_pymnt_d","last_pymnt_amnt","last_credit_pull_d“
- Delete the rows from loan_status column, where the status of the loan is Current. Since it does not help with our analysis.
- Calculate the percentage of loans paid vs charged off for all the records for further analysis later
- Convert the loan_status to a numeric variable, assign 1 for defaulted loans and 0 for paid off ones

- Convert 'term' to numeric variable "tern_in_months“ and drop “term”.
- Remove the % symbol from 'int_rate'
- Retrieve the year and month from 'issue_d' and save the values to issue_year and issue_month. Drop 'issue_d' as it is no longer required.
- For the 'earliest_cr_line' column data, the year is a 2 digit and not compliant with Y2K, we can derive a new value "days_from_earliest_cr_line" to extract meaningful data and drop 'earliest_cr_line'.
- Convert the 'revol_util' values in numeric by removing %

DATA CLEANING

Loan Status vs Loan Amount

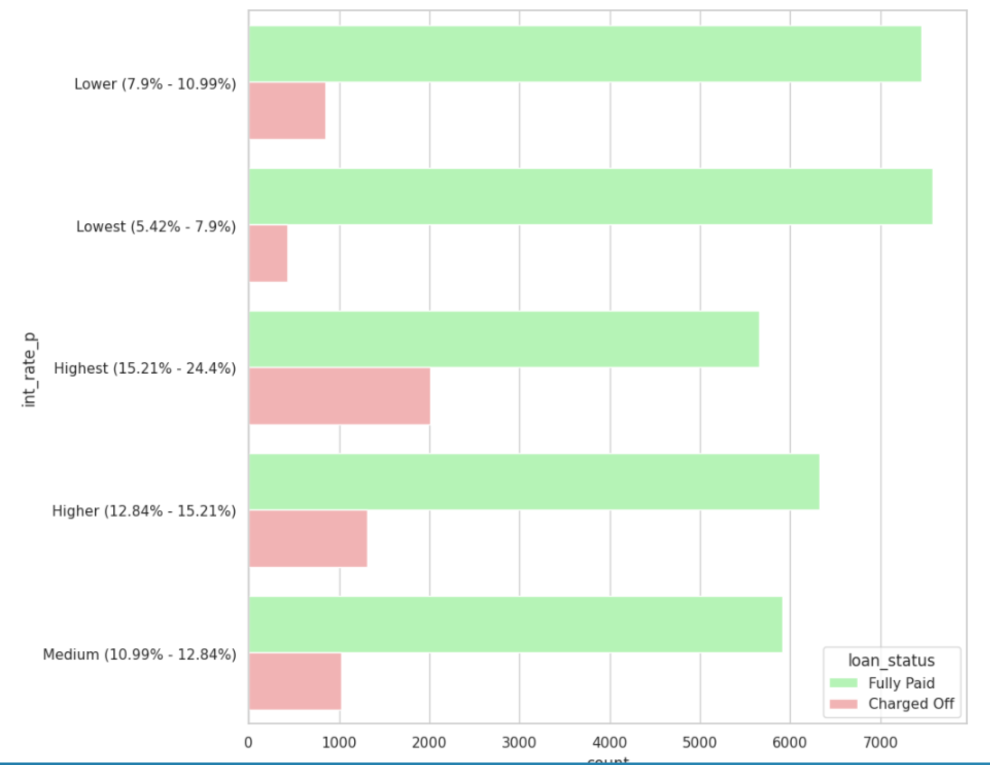


loan_amnt_p	Charged off %	Record count
Highest (16000.0 - 35000.0)	0.187624	7579
Higher (11200.0 - 16000.0)	0.145368	7739
Lowest (500.0 - 5000.0)	0.138725	9472
Medium (7750.0 - 11200.0)	0.131613	7788
Lower (5000.0 - 7750.0)	0.123521	5999

Observation: Higher the loan amount, greater the chance of the loan being defaulted.

UNIVARIATE ANALYSIS

Loan Status vs Interest Rate

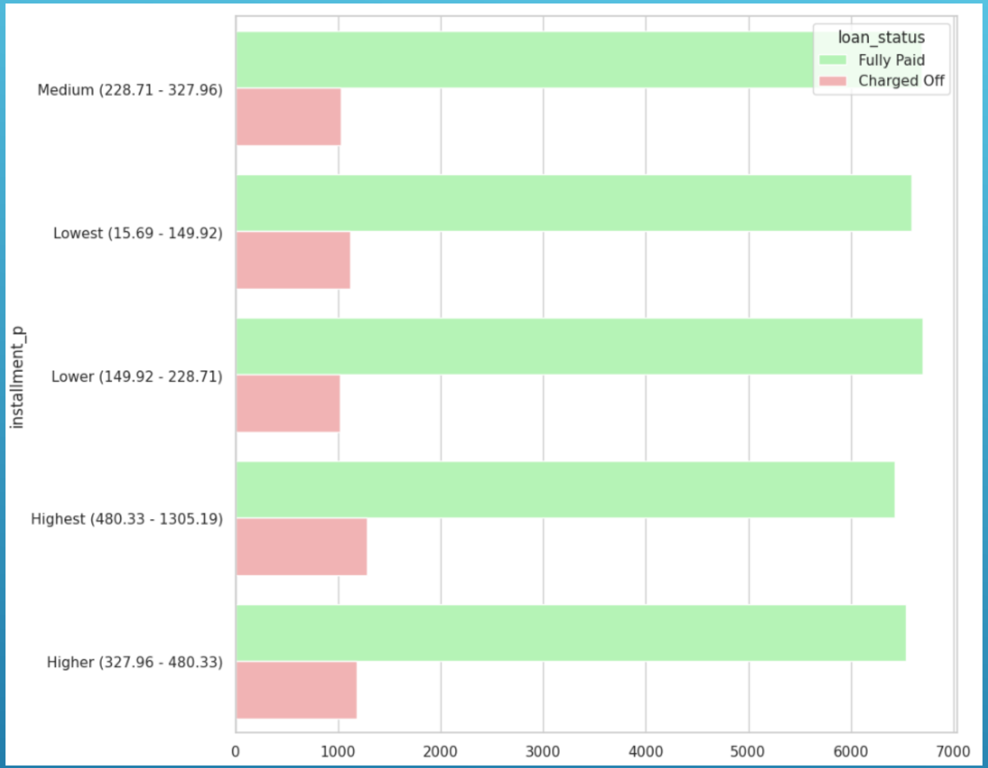


int_rate_p	Charged off %	Record count
Highest (15.21% - 24.4%)	0.262150	7675
Higher (12.84% - 15.21%)	0.171619	7639
Medium (10.99% - 12.84%)	0.146995	6939
Lower (7.9% - 10.99%)	0.102876	8311
Lowest (5.42% - 7.9%)	0.053538	8013

Observation: Higher the interest rate leads to higher charged off percentage

UNIVARIATE ANALYSIS

Loan Status vs Installment Amount

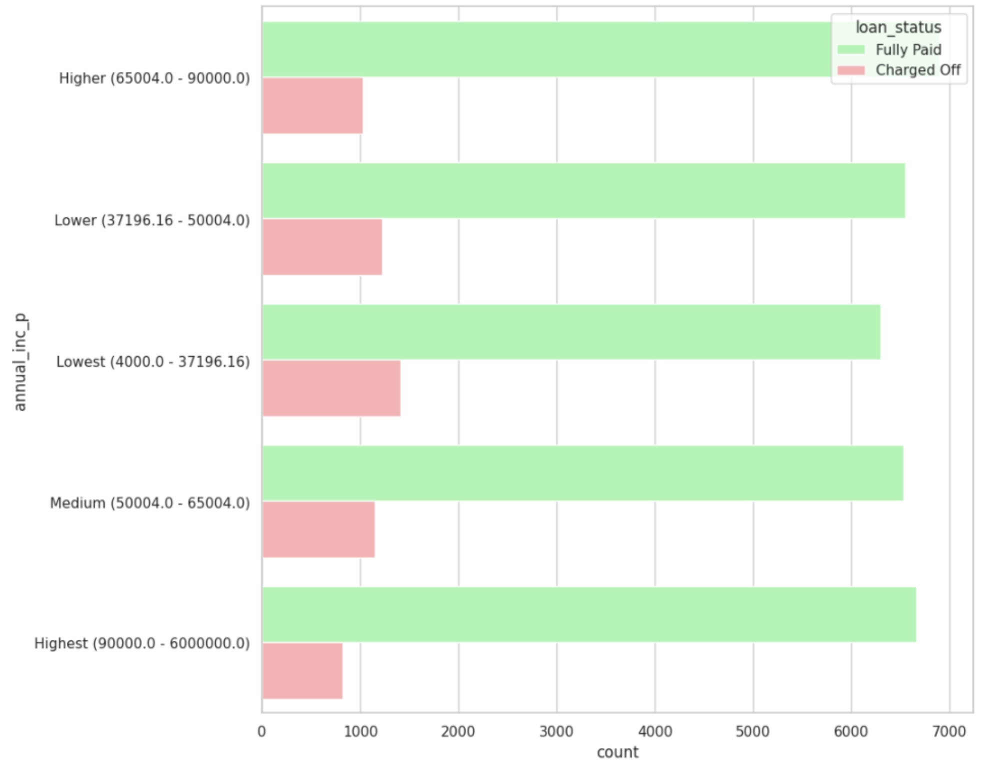


installment_p	Charged off %	Record count
Highest (480.33 - 1305.19)	0.166321	7714
Higher (327.96 - 480.33)	0.152988	7713
Lowest (15.69 - 149.92)	0.145153	7716
Medium (228.71 - 327.96)	0.132936	7718
Lower (149.92 - 228.71)	0.131934	7716

Observation: Higher installment amounts, the higher is the charged off percentage.

UNIVARIATE ANALYSIS

Loan Status vs Annual Income

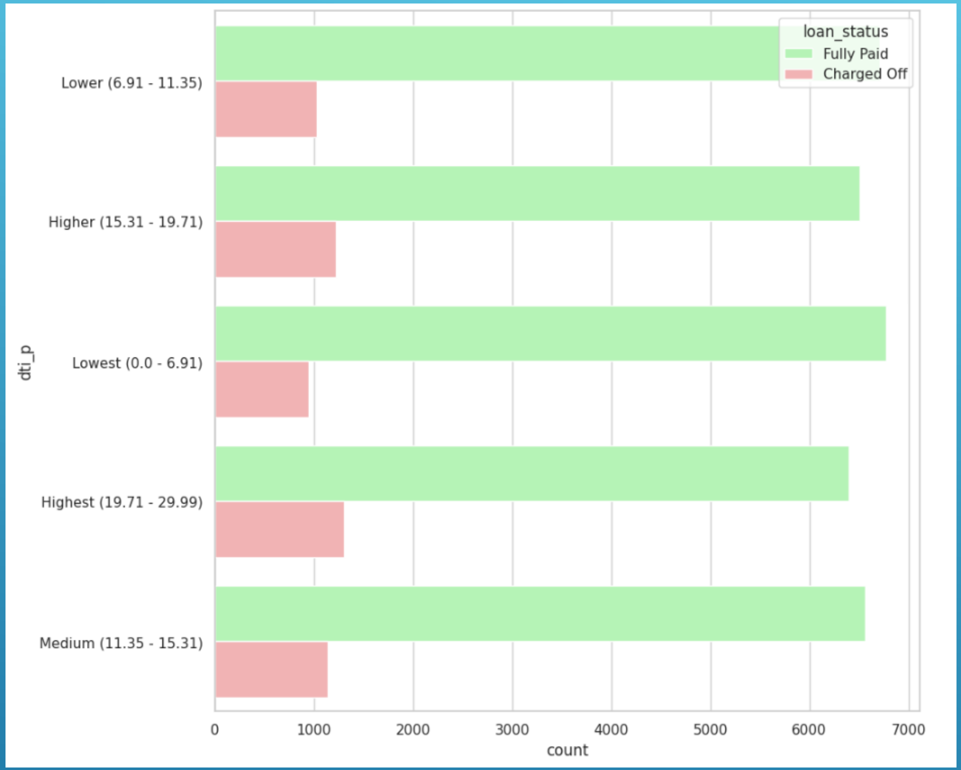


annual_inc_p	Charged off %	Record count
Lowest (4000.0 - 37196.16)	0.183126	7716
Lower (37196.16 - 50004.0)	0.156926	7768
Medium (50004.0 - 65004.0)	0.149629	7679
Higher (65004.0 - 90000.0)	0.129651	7929
Highest (90000.0 - 6000000.0)	0.109285	7485

Observation: Higher the annual income, lower is the Charged Off percentage

UNIVARIATE ANALYSIS

Loan Status vs Debt to Income ratio

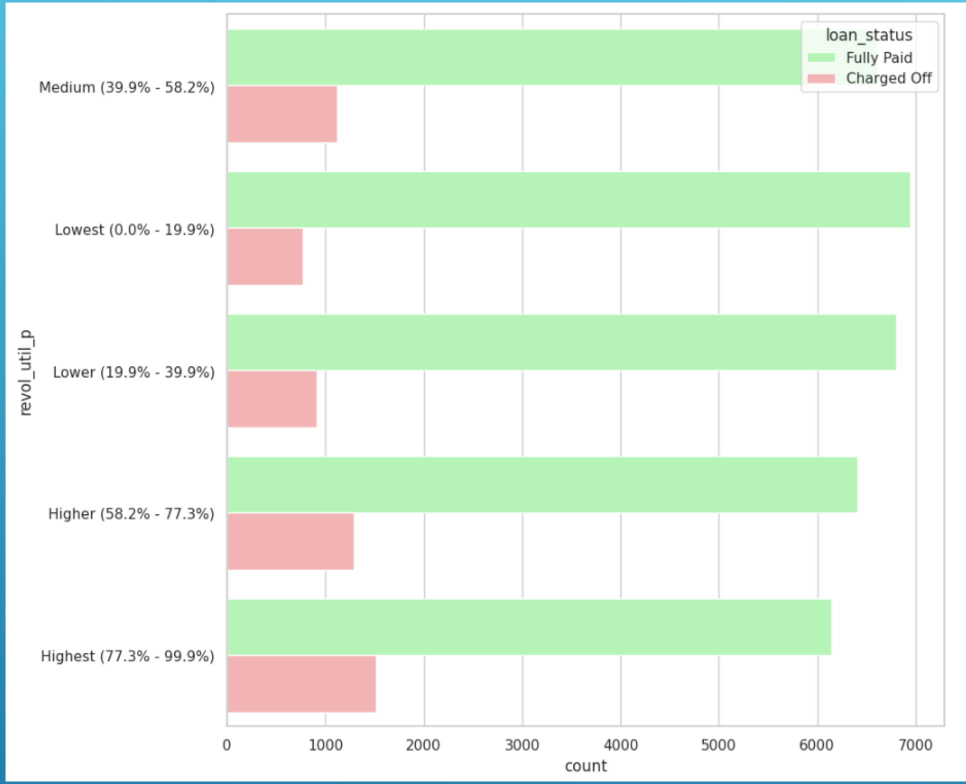


dti_p	Charged off %	Record count
Highest (19.71 - 29.99)	0.168853	7699
Higher (15.31 - 19.71)	0.157908	7726
Medium (11.35 - 15.31)	0.147609	7696
Lower (6.91 - 11.35)	0.132627	7736
Lowest (0.0 - 6.91)	0.122409	7720

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

UNIVARIATE ANALYSIS

Loan Status vs Revolving line utilization rate

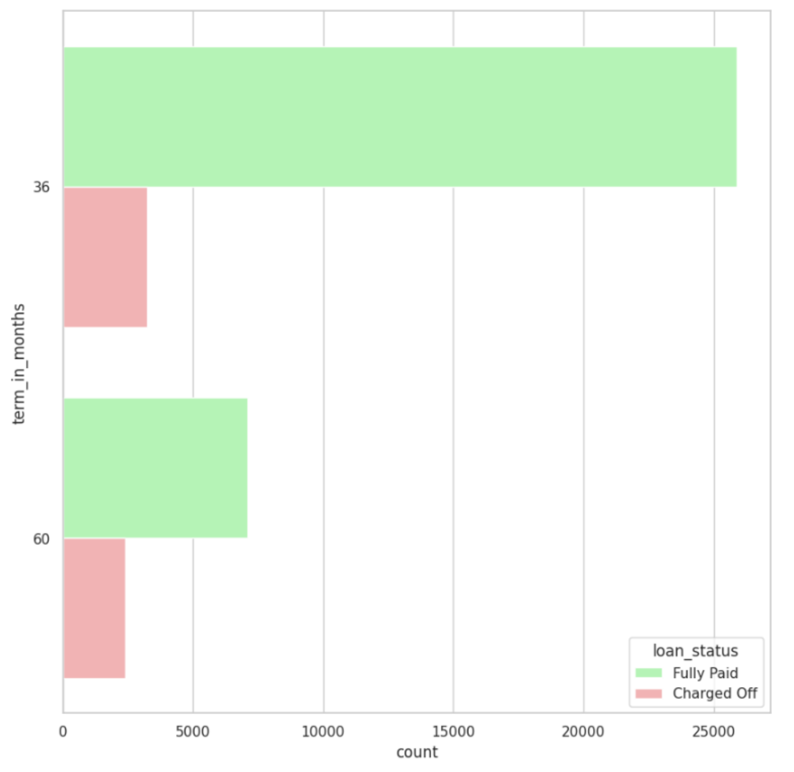


revol_util_p	Charged off %	Record count
Highest (77.3% - 99.9%)	0.198069	7664
Higher (58.2% - 77.3%)	0.167727	7703
Medium (39.9% - 58.2%)	0.145001	7731
Lower (19.9% - 39.9%)	0.118356	7714
Lowest (0.0% - 19.9%)	0.099417	7715

Observation: When the revolving line utilization rate increases, the charged off percentage rises.

UNIVARIATE ANALYSIS

Loan Status vs Loan Term

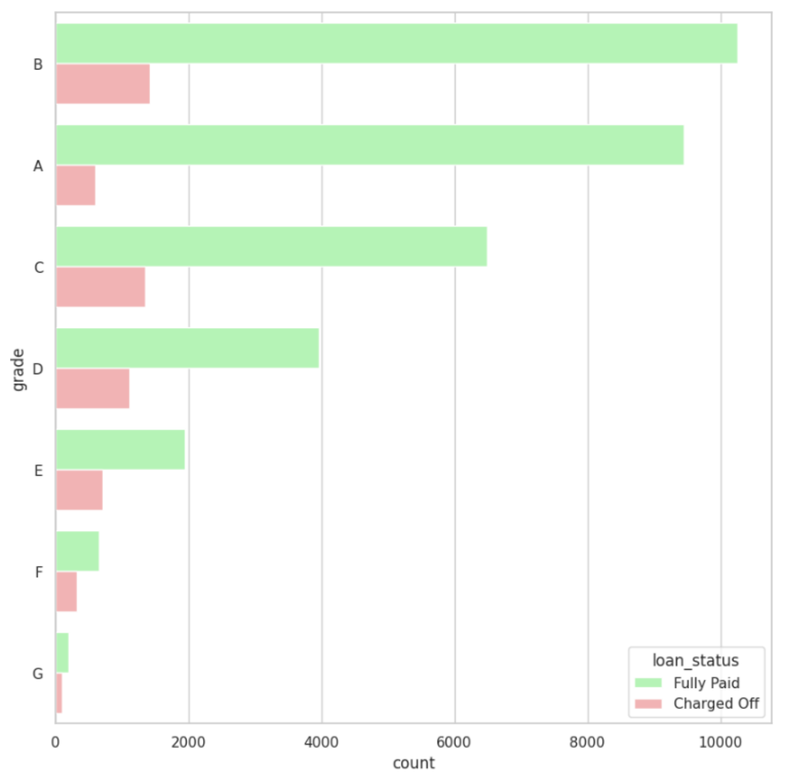


term_in_months	Charged off %	Record count
60	0.253138	9481
36	0.110909	29096

Observation: 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

UNIVARIATE ANALYSIS

Loan Status vs Loan Grade

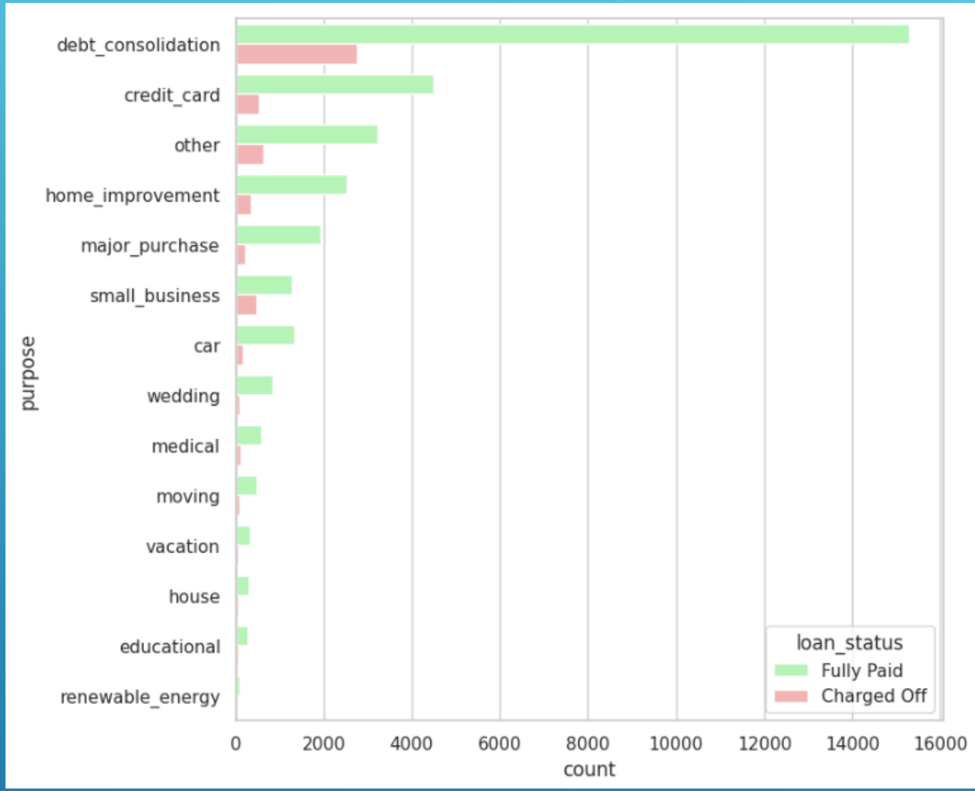


grade	Charged off %	Record count
G	0.337793	299
F	0.326844	976
E	0.268494	2663
D	0.219862	5085
C	0.171943	7834
B	0.122056	11675
A	0.059930	10045

Observation: 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.

UNIVARIATE ANALYSIS

Loan Status vs Loan Purpose

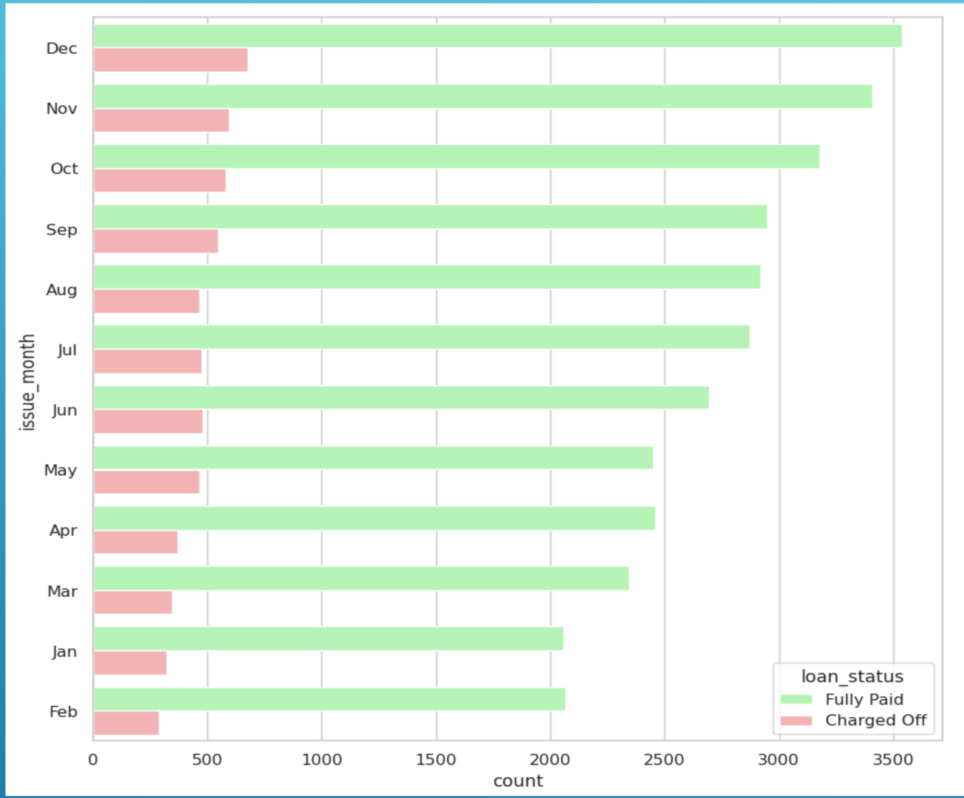


purpose	Charged off %	Record count
small_business	0.270810	1754
renewable_energy	0.186275	102
educational	0.172308	325
other	0.163777	3865
house	0.160763	367
moving	0.159722	576
medical	0.155653	681
debt_consolidation	0.153254	18055
vacation	0.141333	375
home_improvement	0.120696	2875
credit_card	0.107818	5027
car	0.106738	1499
wedding	0.103672	926
major_purchase	0.103256	2150

Observation: it is evident that the loans taken for small_business, renewable_energy and educational are the riskier ones.

UNIVARIATE ANALYSIS

Loan Status vs Month Loan Issued



issue_month	Charged off %	Record count
Dec	0.160854	4215
May	0.159644	2919
Sep	0.156375	3498
Oct	0.154214	3761
Jun	0.151887	3180
Nov	0.149276	4006
Jul	0.142942	3351
Aug	0.138135	3388
Jan	0.134931	2379
Apr	0.130696	2831
Mar	0.128948	2691
Feb	0.122986	2358

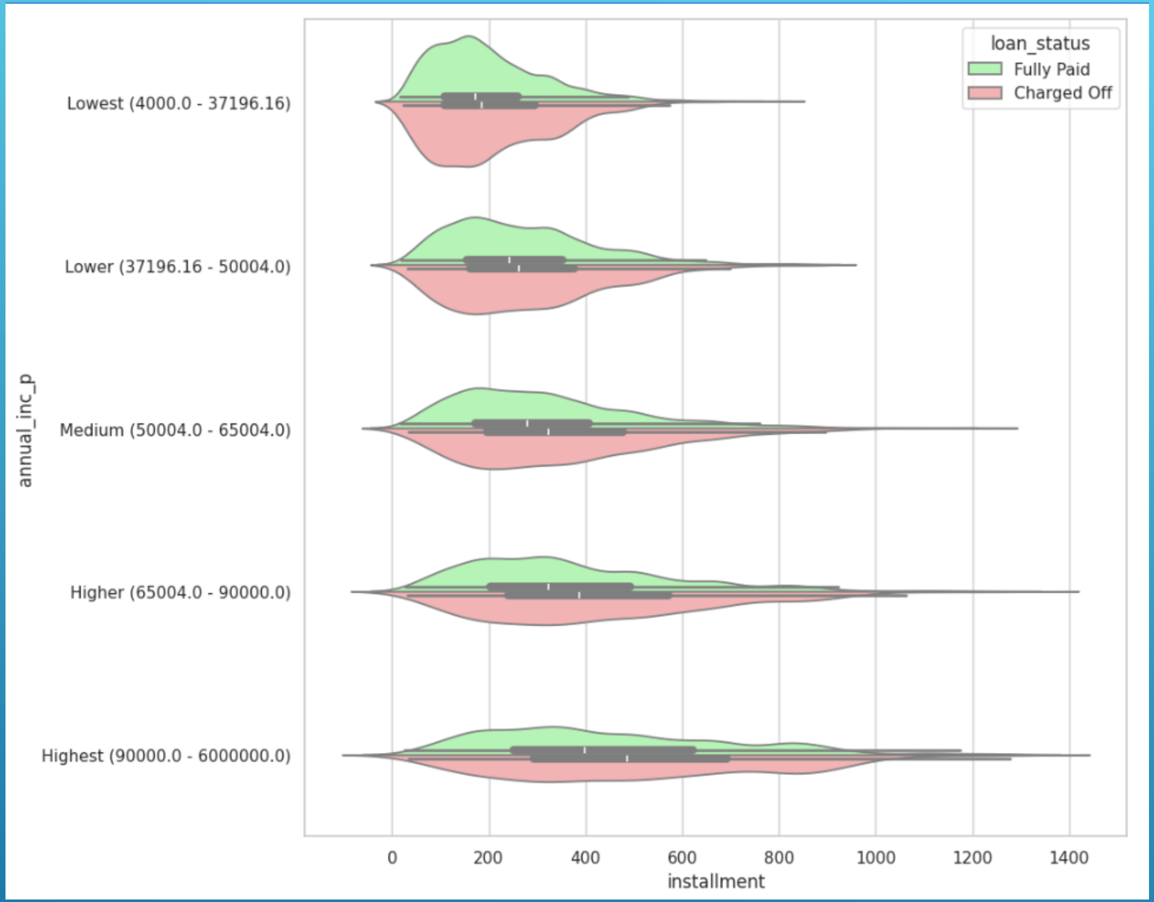
Observation: We can clearly see December is the month which has the highest number of loan applications per year and also have the biggest default ratio. It is likely that people take a loan for travelling or socializing during Christmas and then are not able to pay back. Month of May is also another one, which is during the summer break and right before the Memorial day and Independence day breaks in US where people love to travel.

UNIVARIATE ANALYSIS

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.

BIVARIATE ANALYSIS

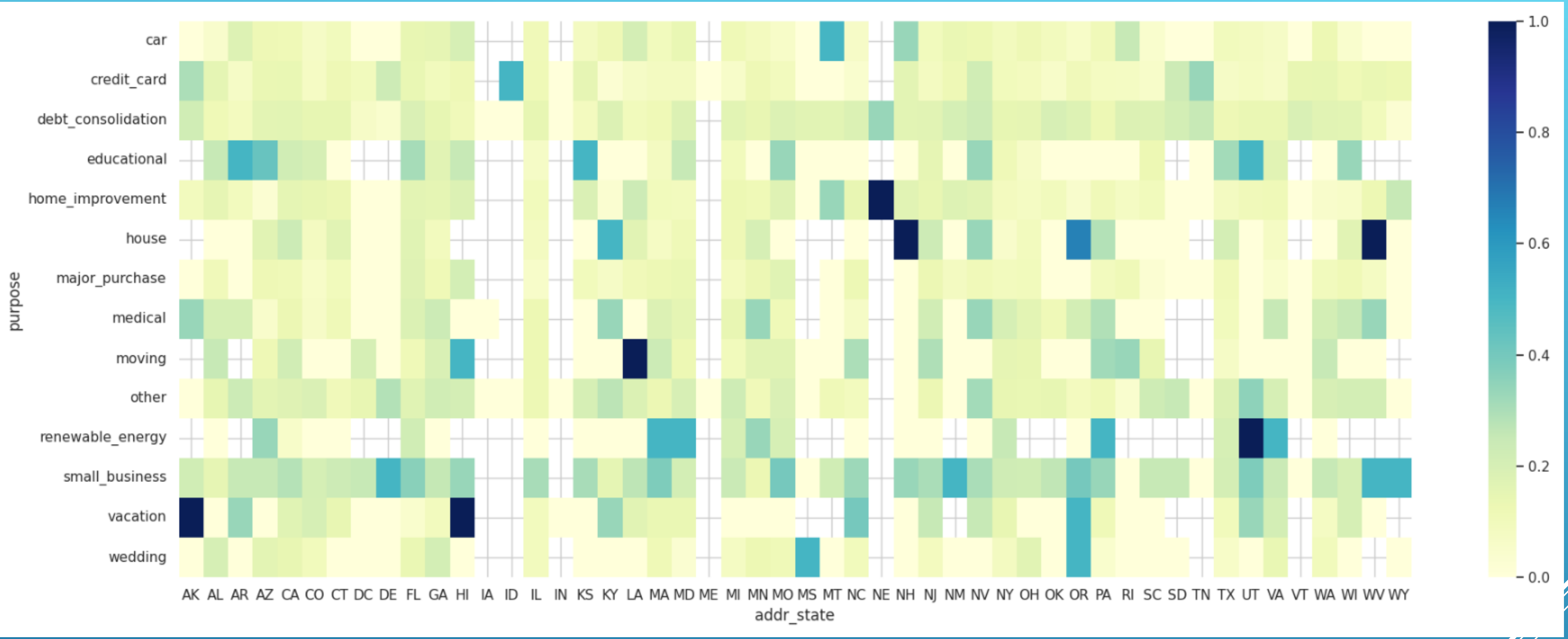
Loan installment vs Annual Income vs Loan Status



Observation: Higher installments for any income group have a high number of defaults.

BIVARIATE ANALYSIS

State vs Purpose vs Loan Status

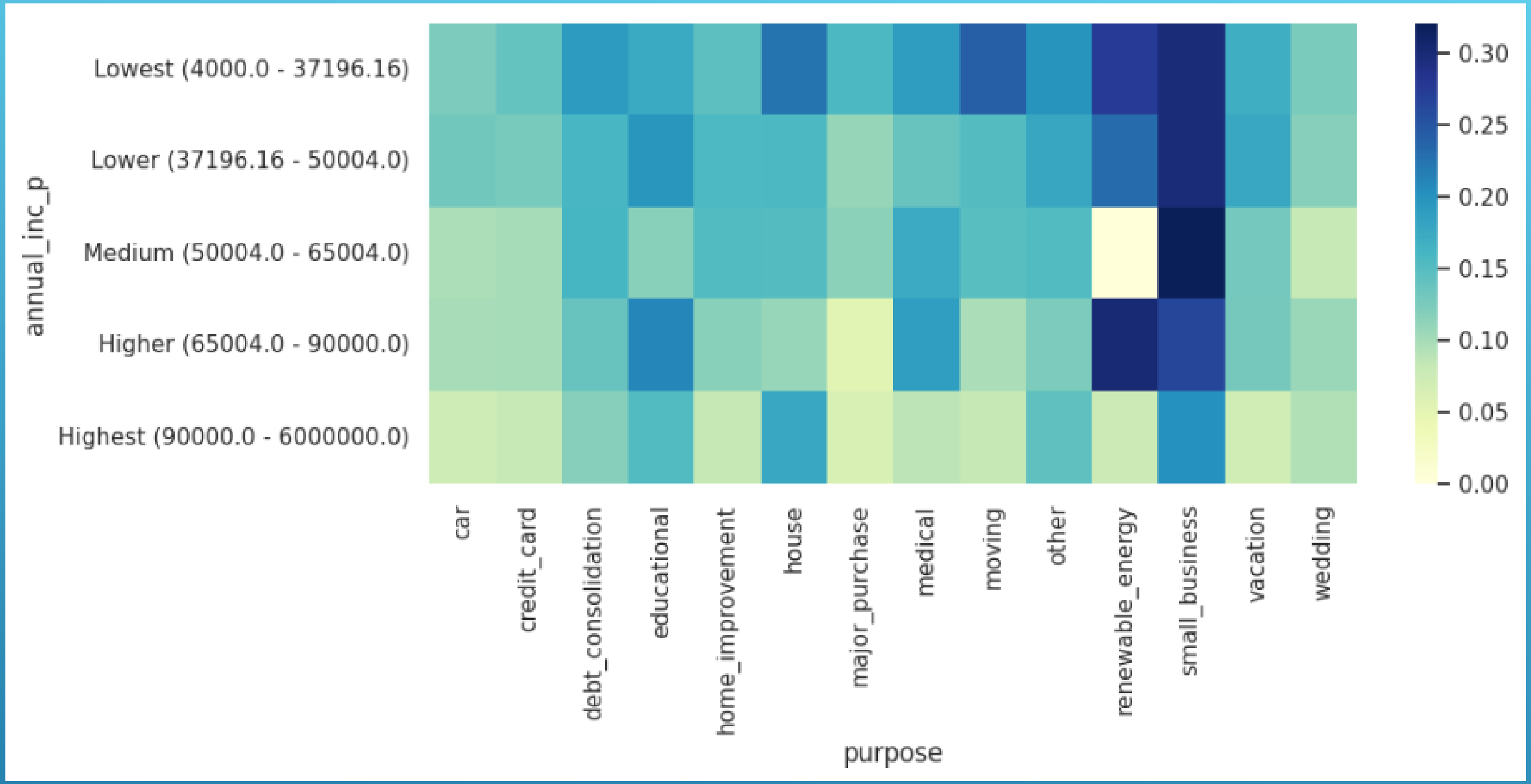


Observation: the darker the intersection of addr_state has with the purpose of the loan, the risker the loan application is. Some of the examples are below:

- vacation loans in AK, HI, OR
- small business loans in DE, NM, WV, WY

BIVARIATE ANALYSIS

Annual Income group vs Purpose vs Loan Status

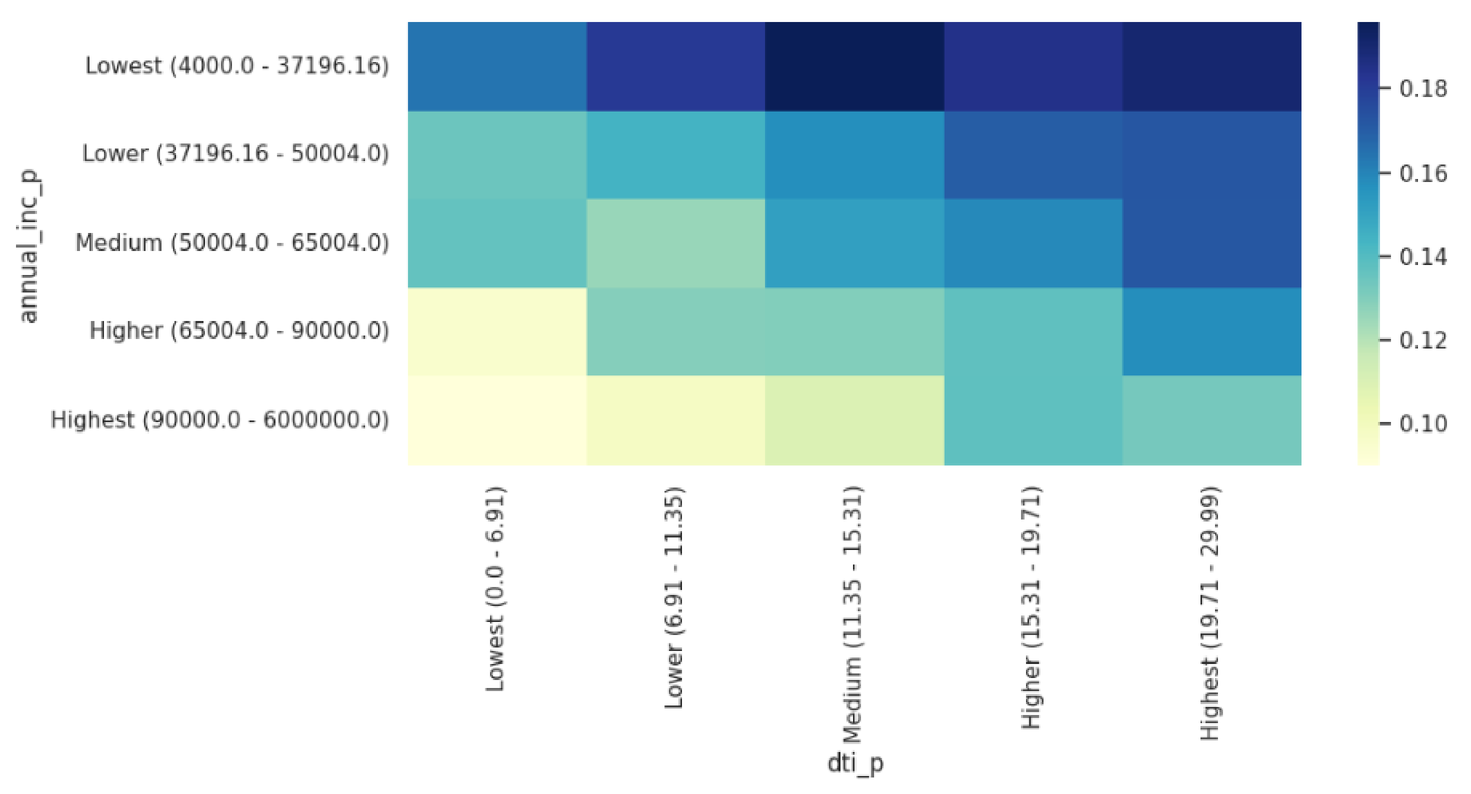


Observation: Plot of various income groups versus purposes of loans shows the following risky loans;

- small business loans for lowest and medium income groups
- renewable energy loans for higher income group

BIVARIATE ANALYSIS

Debt-to-Income ratio vs Annual Income vs Loan Status



Observation: Medium debt-to-income group in the lowest income range is the most risky when it comes to loan repayment.

BIVARIATE ANALYSIS

From the analysis performed above we can conclude the impact of the following driving factors;

- Low to Moderate Impact
  - Higher loan amount (above 16K)
  - Higher installment amount (above 327)
  - The state in which the applicant resides (NV, SD, AK, FL, etc.)
- High Impact
  - Higher interest rate (above 13%)
  - Repayment term (5 years)
  - Loan purpose (small business, renewable energy, educational)
- Combined Impact
  - High loan amount and interest rate for lower income group
  - High installment and longer repayment term
  - Home ownership (other) and loan purpose (car, moving or small business)
  - Income group and loan purpose

OBSERVATION SUMMARY

It is recommended that a portfolio and risk assessment be done using the following key influencers;

- Loan purpose
- Annual Income
- State
- Loan grade and Sub-grade
- Interest rate

CONCLUSION / RECOMMENDATION