Lending Club Case Study

Nikita & Vaibhav

Abstract

Lending club 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.

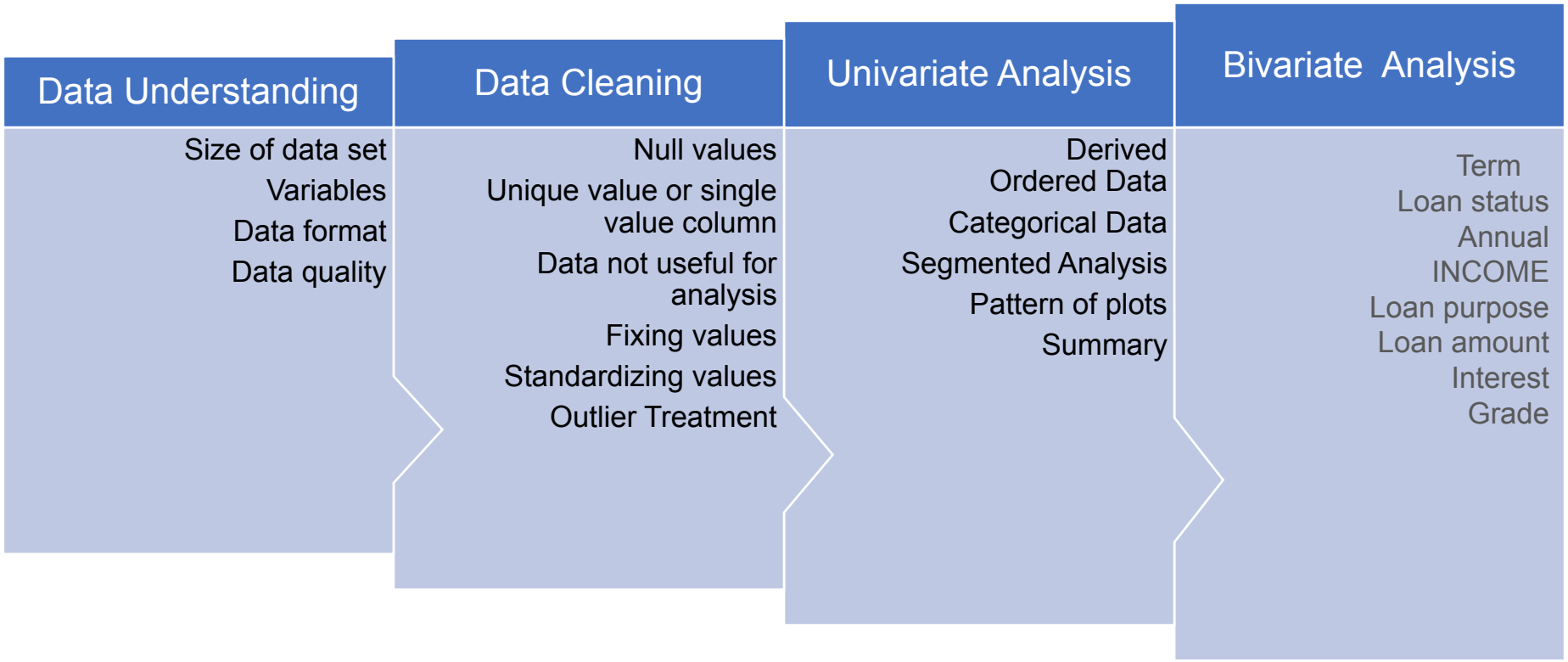
The objective of analysis is to use the information about past loan applicants and find whether they ‘defaulted’ or not.

The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default

Methodology

1. Data Understanding: Working with the Data Dictionary and getting knowledge of all the columns and their domain specific uses
2. Data Cleaning: Removing the null valued columns, unnecessary variables and checking the null value percentage and removing the respective rows.
3. Univariate Analysis: Analysing each column, plotting the distributions of each column.
4. Segmented Univariate Analysis: Analysing the continuous data columns with respect to the categorical column
5. Bivariate Analysis: Analysing the two variable behaviour like term and loan status with respect to loan amount
6. Recommendations: Analysing all plots and recommendations for reducing the loss of business by detecting columns best which contribute to loan defaulters.

Methodology



Data Understanding

- Description for the columns used in dataset

LoanStatNew	Description
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
acc_open_past_24mths	Number of trades opened in past 24 months.
addr_state	The state provided by the borrower in the loan application
all_util	Balance to credit limit on all trades
annual_inc	The self-reported annual income provided by the borrower during registration.
annual_inc_joint	The combined self-reported annual income provided by the co-borrowers during registration
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers
avg_cur_bal	Average current balance of all accounts
bc_open_to_buy	Total open to buy on revolving bankcards.
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
chargeoff_within_12_mths	Number of charge-offs within 12 months
collection_recovery_fee	post charge off collection fee
collections_12_mths_ex_med	Number of collections in 12 months excluding medical collections
delinq_2yrs	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
desc	Loan description provided by the borrower
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
dti_joint	A ratio calculated using the co-borrowers' total monthly payments on the total debt obligations, excluding mortgages and the requested LC loan, divided by the co-borrowers' combined self-reported monthly income
earliest_cr_line	The month the borrower's earliest reported credit line was opened
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
emp_title	The job title supplied by the Borrower when applying for the loan.*
fico_range_high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico_range_low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade
home_ownership	The home ownership status provided by the borrower during registration. Our values are: RENT, OWN, MORTGAGE, OTHER.
id	A unique LC assigned ID for the loan listing.
il_util	Ratio of total current balance to high credit/credit limit on all install acct
initial_list_status	The initial listing status of the loan. Possible values are – W, F
inq_fi	Number of personal finance inquiries
inq_last_12m	Number of credit inquiries in past 12 months

Data Understanding

- Data size

```
data.shape
```

```
(39717, 111)
```

- Data quality

```
[9]  # Finding columns with more than 100% missing values
     columns_with_missing_values = null_percentage[null_percentage == 100]
     columns_with_missing_values.count()
```

```
54
```

Data Cleaning

It is observed that there are a lot of columns with all null values. Let's first remove them

Removing the null columns which are 100%

```
[10] data.dropna(axis = 1, how = 'all', inplace = True)
```

```
[11] data.shape
```

```
(39717, 57)
```

Data Cleaning

There are many single valued column which are not useful for analysis, so removing them. The dataset size reduced to 48 columns

We have identified columns in the dataset with characteristics that make them less valuable for analysis:

Unique Value Columns: The following columns have only one unique value in all the rows, which means they do not provide useful information for analysis and can be removed: 'tax_liens' 'delinq_amnt' 'chargeoff_within_12_mths' 'acc_now_delinq' 'application_type' 'policy_code' 'collections_12_mths_ex_med' 'initial_list_status' 'pymnt_plan' These columns contain the same value throughout the dataset, making them redundant for any meaningful analysis. Removing them will simplify the dataset and reduce unnecessary computational overhead.

Removing single valued columns form the dataset

```
[13] single_value_col = ['pymnt_plan','initial_list_status','collections_12_mths_ex_med','policy_code','application_type', 'acc_now_delinq','chargeoff_within_12_mths', 'delinq_amnt','tax_l
     data.drop(columns = single_value_col, axis=1, inplace=True)
     data.shape
```

```
(39717, 48)
```

Data Cleaning

Further reducing size of dataset as removing columns which are not useful for analysis.

Removing columns which have more than 60% null values

```
[15] data.drop(['mths_since_last_delinq', 'mths_since_last_record', 'next_pymnt_d'], axis=1, inplace=True)
```

```
[16] data.shape

     (39717, 45)
```

Removing categorical data from dataset, string and fields which won't contribute to default prediction analysis

```
[17] unique_value_counts2 = data.nunique()
     print(unique_value_counts2)

     id                          39717
     member_id                   39717
     loan_amnt                     885
     funded_amnt                  1041
     funded_amnt_inv              8205
```

Data Cleaning

Removing string data

```
[19] data.drop(["id", "member_id", "title", "url", "emp_title", "zip_code","addr_state","desc"], axis = 1, inplace = True)
```

Removing data based on business logic

```
[20] data.drop(['delinq_2yrs', 'revol_bal', 'total_pymnt', 'total_rec_prncp','total_rec_int', 'total_rec_late_fee', 'last_credit_pull_d', 'recoveries', 'collection_recovery_fee'
```

```
[21] data.drop(['out_prncp', 'out_prncp_inv'], axis=1, inplace=True)
```

```
[22] data.shape

     (39717, 24)
```

Data Cleaning

List of columns after cleaning data set

Now we have 24 columns which are relevant for our analysis

1. loan_amnt 885
2. funded_amnt 1041
3. funded_amnt_inv 8205
4. term 2
5. int_rate 371
6. installment 15383
7. grade 7
8. sub_grade 35
9. emp_length 11
10. home_ownership 5
11. annual_inc 5318
12. verification_status 3
13. issue_d 55
14. loan_status 3
15. purpose 14
16. dti 2868
17. earliest_cr_line 526
18. inq_last_6mths 9
19. open_acc 40
20. pub_rec 5
21. revol_util 1089
22. total_acc 82
23. total_pymnt_inv 37518
24. pub_rec_bankruptcies 3

Data Cleaning

Handling missing data

- Removed the rows as % of null value is around 1%, further size reduced to rows = 37898

```
[24] #Finding percentage of null or missing values
     null_perc = round(100*(data.isnull().sum()/len(data.index)), 2)
     null_perc[ null_perc > 0 ]
```

```
emp_length              2.71
revol_util              0.13
pub_rec_bankruptcies    1.75
dtype: float64
```

Removing the the null valued rows in the above columns.

Data Cleaning

Standardizing the data for analysis

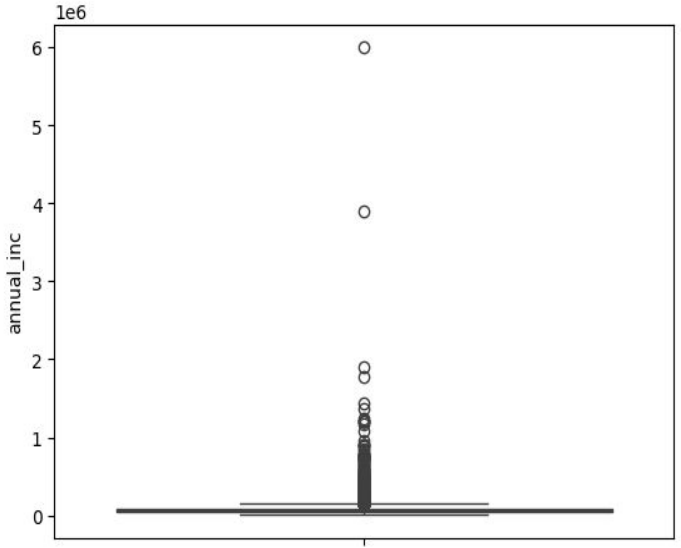
- "revol_util" column, has continuous values, but described as an object column. So we need to standardize the data in this column
- int_rate and revol_util** are having '%' symbol values and having data type of object. Let's remove % at the end and convert to float
- "emp_length" can be converted into numerical values with following assumption --> { (< 1 year) is taken as 0 and 10+ years is taken as 10 }

Converting data type

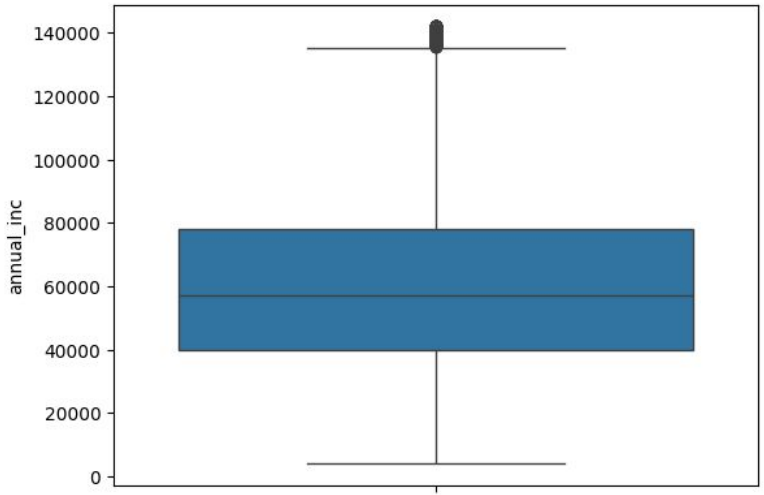
- issue_d, earliest_cr_line are having date values, let convert column data type to date.

Data Cleaning

Outlier Treatment - checked for three columns out of which ‘annual_inc’ was treated for analysis, removing outliers at 95 percentile



Before



After

Univariate Analysis

Derived variable

Created columns for further analysis

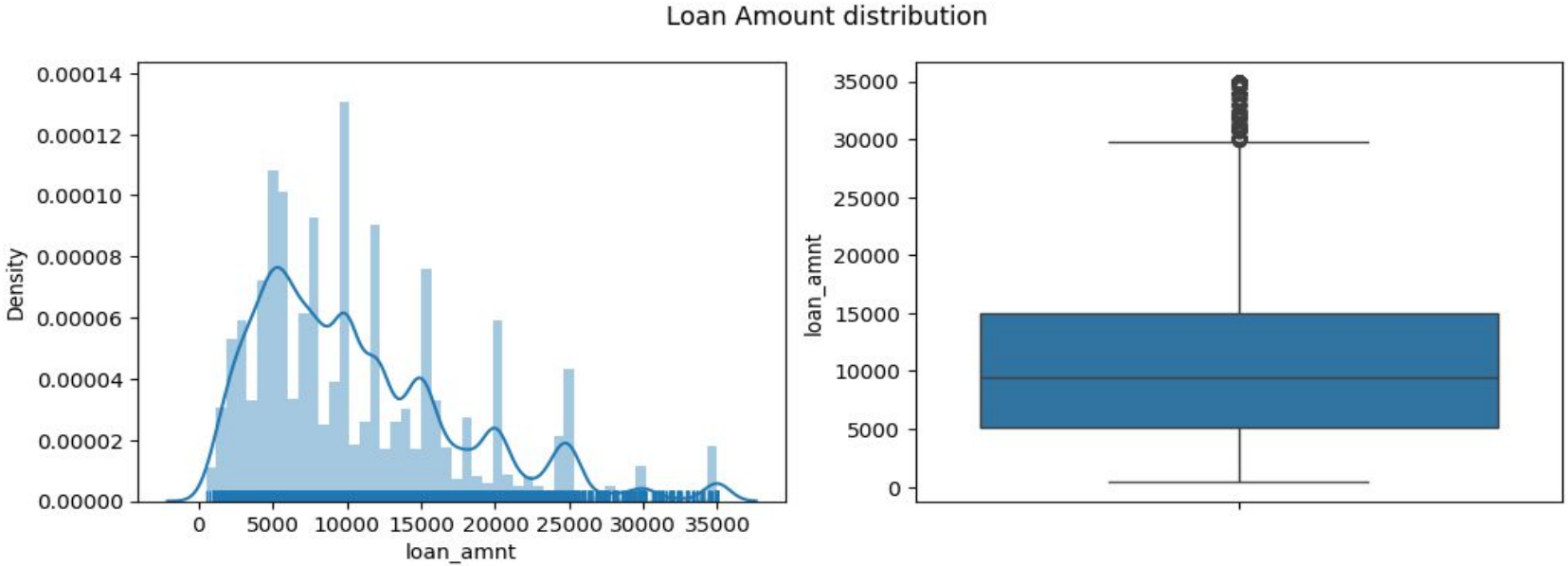
```
[33] #issue_d column
     data['issue_d_year'] = data.issue_d.dt.year
     data['issue_d_month'] = data.issue_d.dt.strftime('%b')
     data['issue_d_weekday'] = data.issue_d.dt.weekday
     #data type conversion of year and weekday
     data['issue_d_year'] = data['issue_d_year'].astype(object)
     data['issue_d_weekday'] = data['issue_d_weekday'].astype(object)

     #earliest_cr_line
     data['earliest_cr_line_year'] = data.earliest_cr_line.dt.year
     data['earliest_cr_line_month'] = data.earliest_cr_line.dt.strftime('%b')
     #data type conversion of year and weekday
     data['earliest_cr_line_year'] = data['earliest_cr_line_year'].astype(object)
```

Univariate Analysis

Loan amount

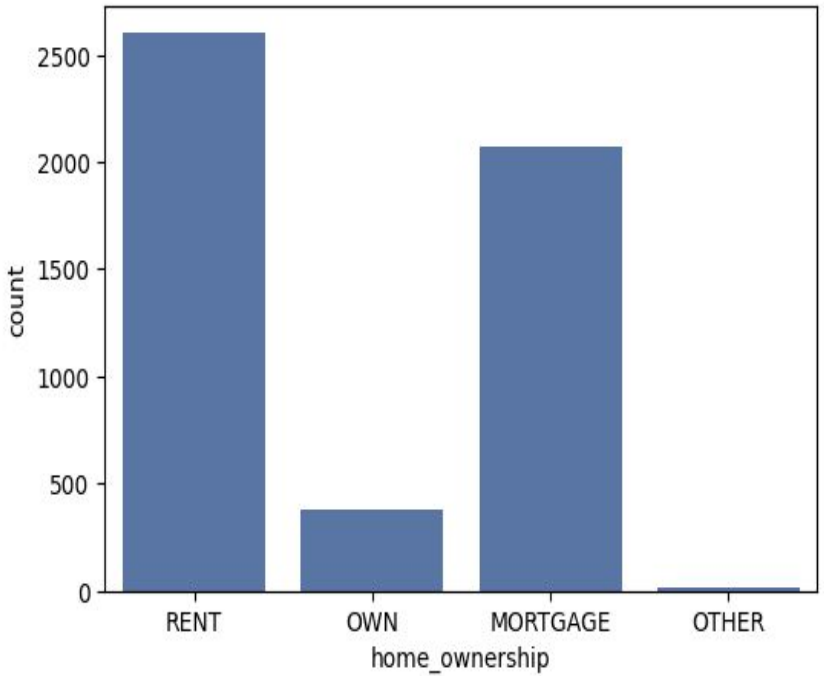
From the fig. below we can see loan amount 10000 is most taken



Univariate Analysis

Home ownership

From the fig. it indicates most of the defaulter are persons living in rented house



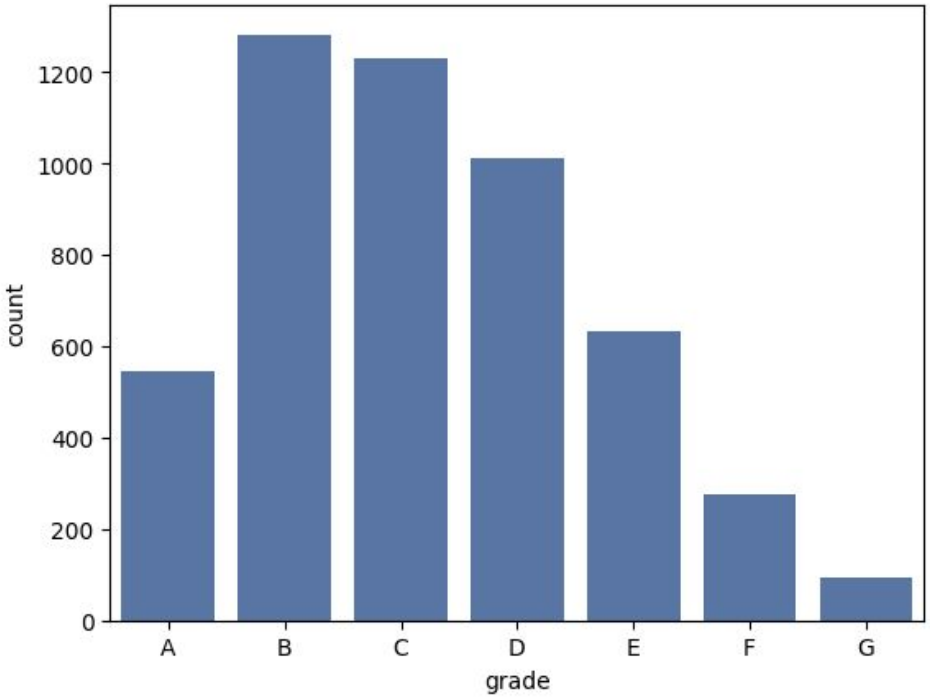
Home ownership

Univariate Analysis

-Loans graded as 'B' have an increased probability of default.

-Specifically, loans assigned a total grade of 'B5' have a higher likelihood of defaulting.

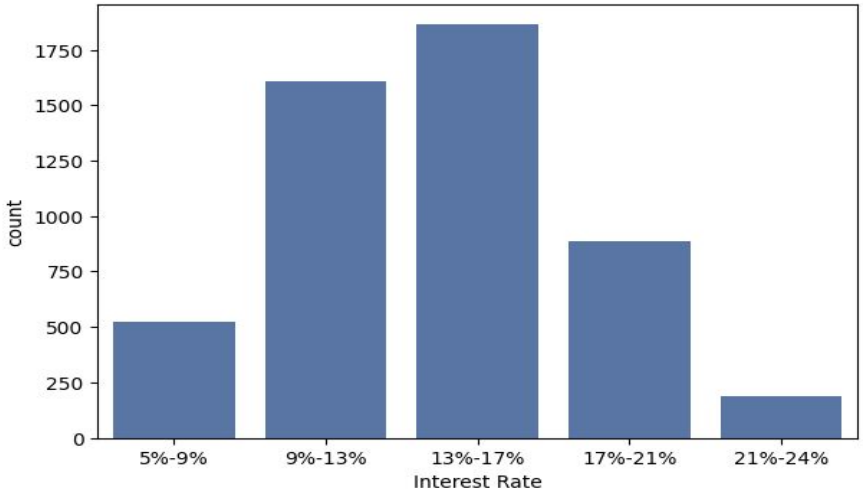
The fig. besides indicates the same

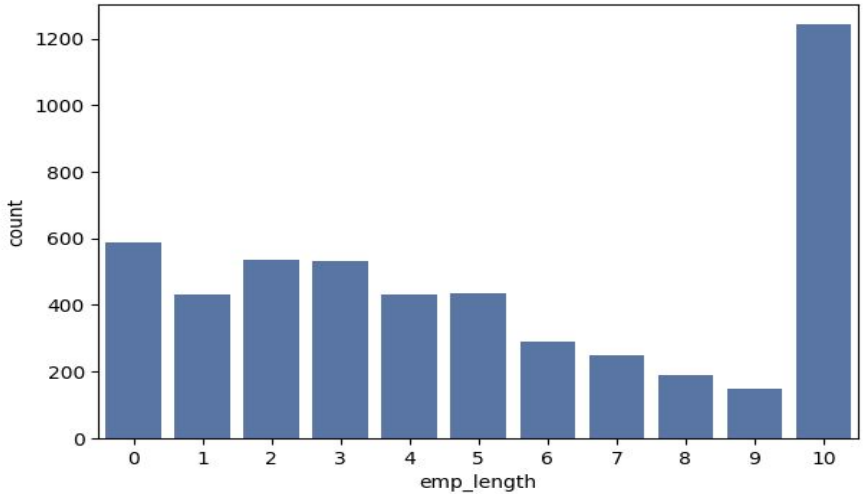


Univariate Analysis

Segmented Analysis

Creating bins for numerical value to make them categorical, the fig below indicates the interest rate 13-17% and employment year 10 as maximum

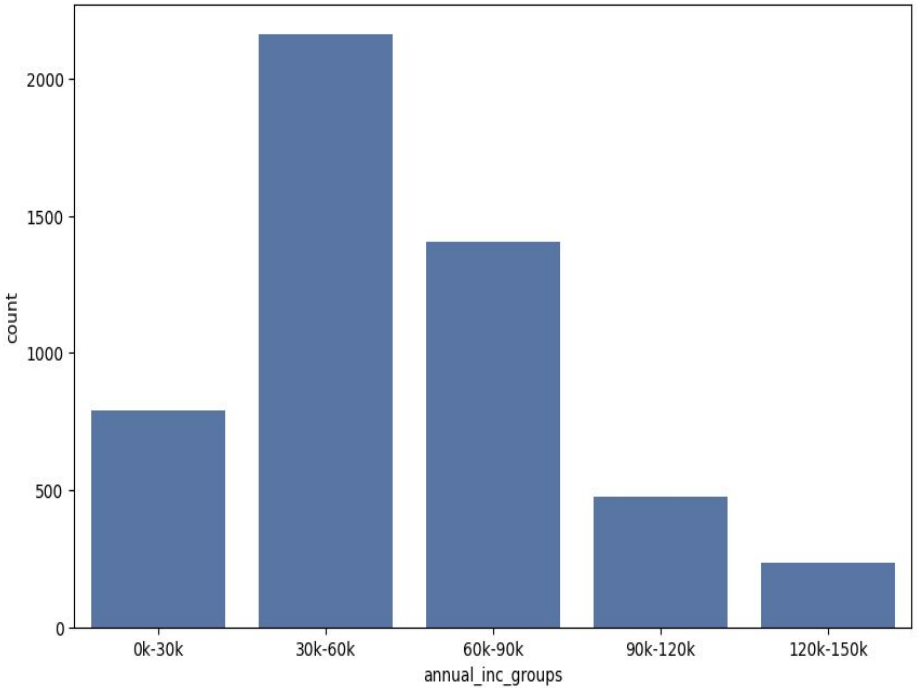




Interest rate and employment tenure

Univariate Analysis

It is observed from the fig. that persons with an annual income ranging from 30000 to 60000 are more prone to default.



Univariate Analysis

Based on the analysis with respect to the charged off loans for each variable suggests the following. There is a more probability of defaulting when :

- -Applicants with 'RENT' as their house ownership status have a higher probability of defaulting.
- -Borrowers who use the loan for debt consolidation purposes are more likely to default.
- -Default probability is elevated when the interest rate falls within the range of 13-17%.
- -Borrowers with an annual income ranging from USD 30000 to USD 60000 are more prone to default.
- -Individuals with 20-37 open accounts (open_acc) exhibit a higher likelihood of default.
- -Borrowers with a employment length of 10 years have an increased chance of defaulting.
- -Loans funded by investors within the range of 5,000 to 10,000 are associated with higher default rates.

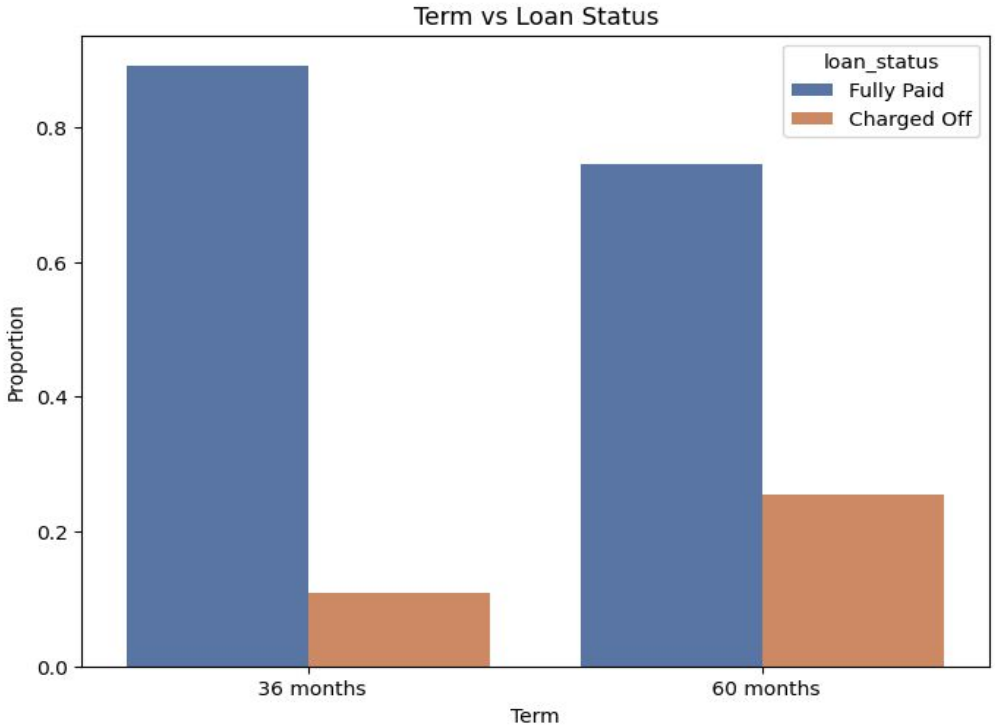
Univariate Analysis

- -Loan amounts falling between 5000 and 10000 are linked to an elevated likelihood of default.
- -A debt-to-income ratio (dti) in the range of 12-18 increases the probability of default.
- -Loans with monthly installments between 150 and 300 are more likely to default.
- -Loans with a term of 36 months have a higher probability of defaulting.
- -Loans without verified status are associated with a greater likelihood of default.
- -A borrower's default probability tends to be higher when they have zero recent inquiries in the last 6 months.
- -Borrowers with zero derogatory public records are more prone to default.
- -Loans intended for the purpose of debt consolidation exhibit a higher likelihood of default.
- -Loans graded as 'B' have an increased probability of default.
- -Specifically, loans assigned a total grade of 'B5' have a higher likelihood of defaulting.

Bivariate Analysis

Term Vs Loan Status

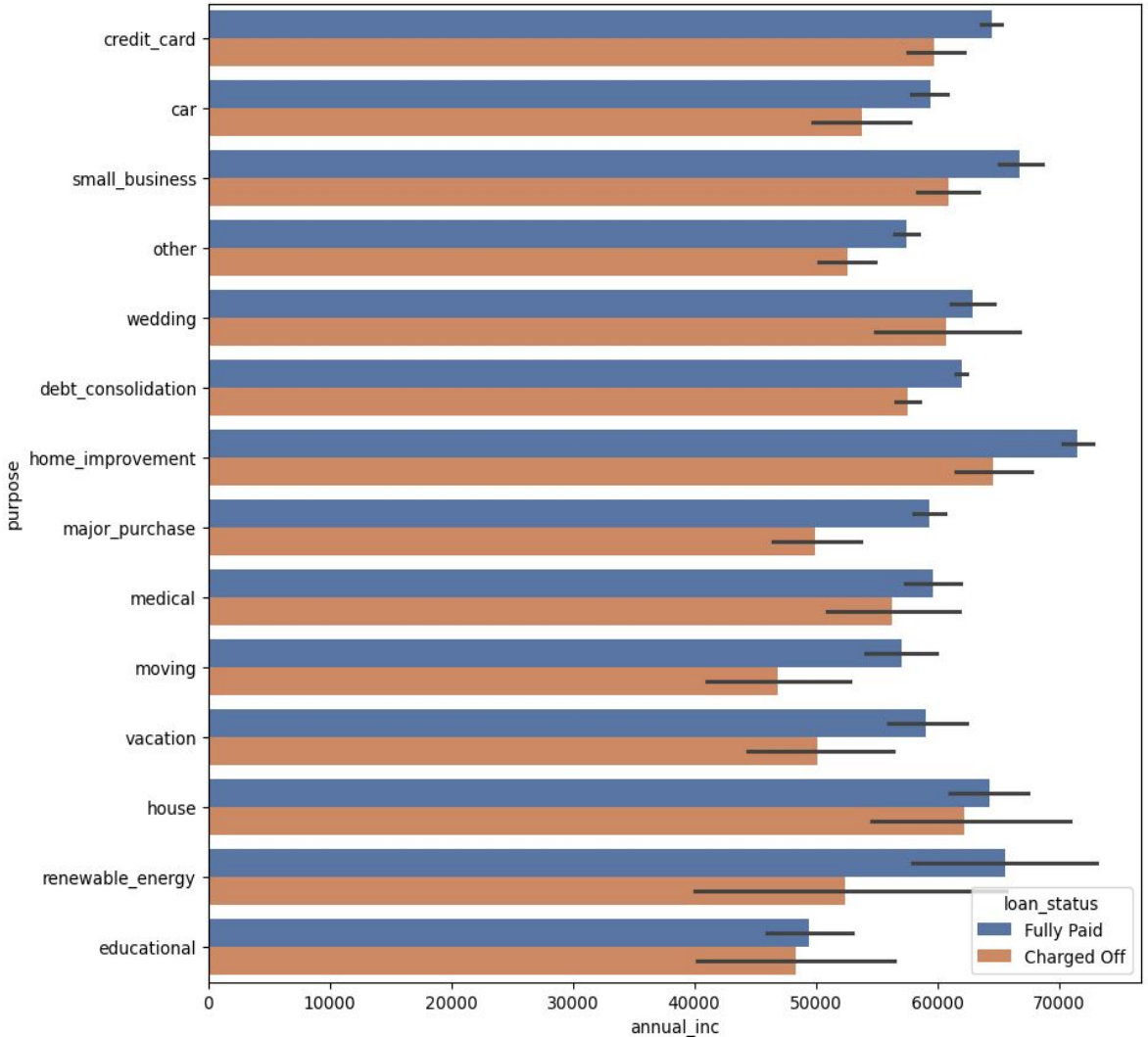
There are more proportion of borrowers defaulted loan in 60 months term then 36 months. Also the Fully Paid rate is higher in 36 months tenure



Bivariate Analysis

Annual income vs loan purpose

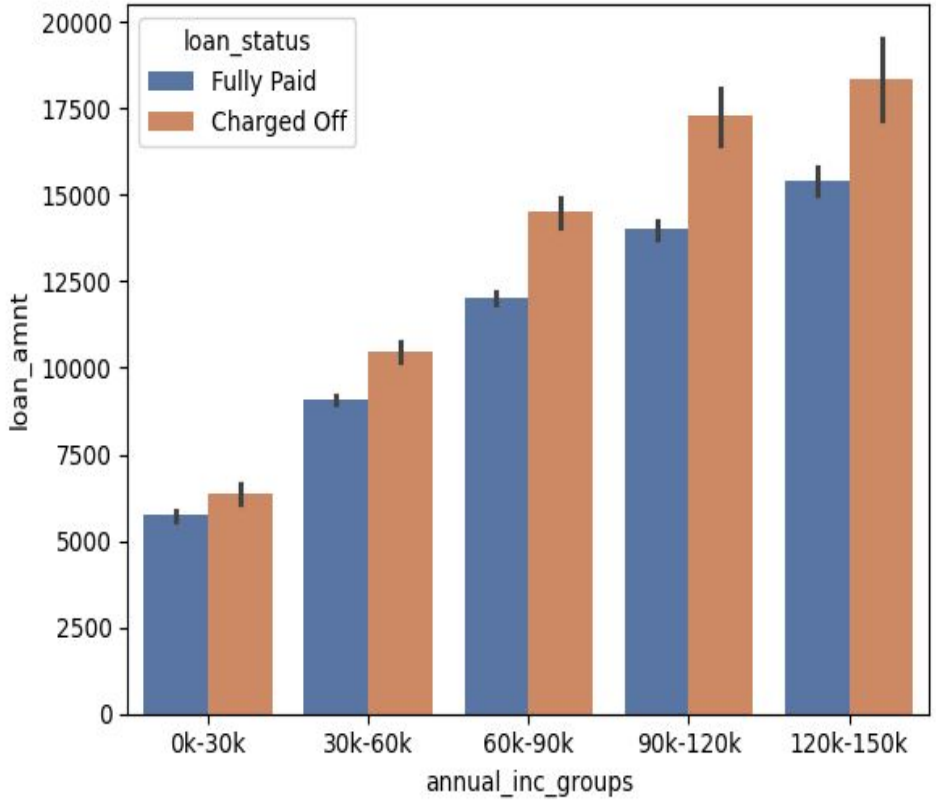
While the largest number of loan applications and defaults are associated with "debt consolidation," but the annual income of these applicants is not the highest. Borrowers with higher incomes tend to apply for loans primarily for purposes related to "home improvement," "house," "renewable energy," and "small businesses."



Bivariate Analysis

Annual income vs loan amount

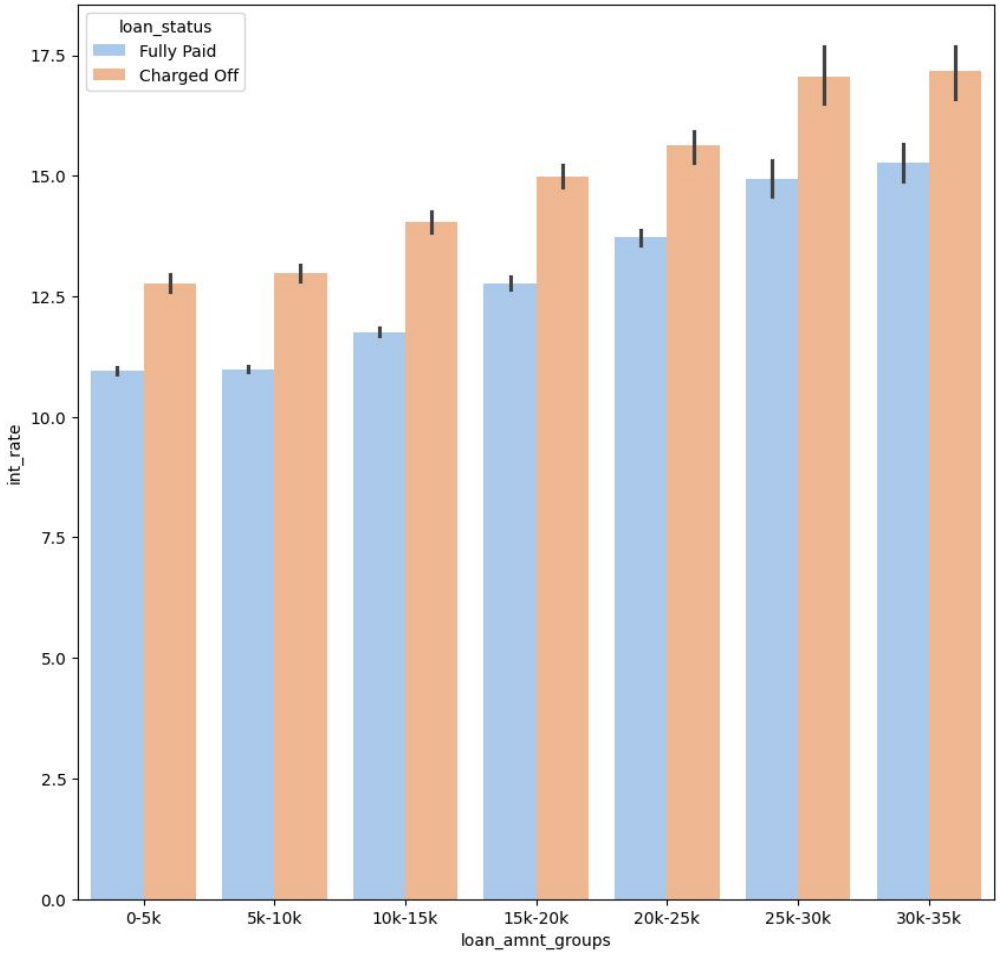
The loan amount is higher for high annual income group



Bivariate Analysis

Interest rate vs loan amount

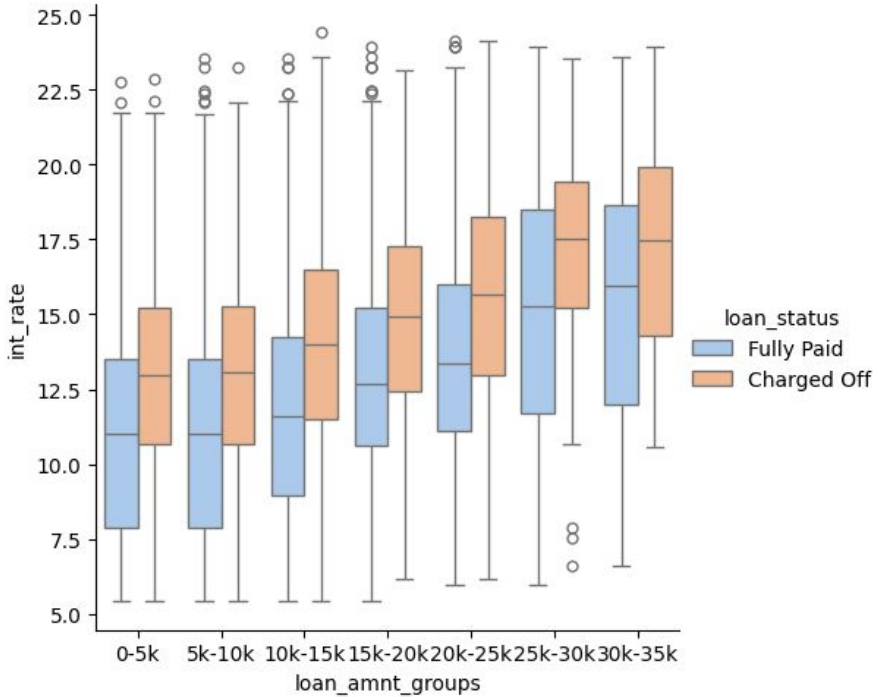
Interest rate is maximum for maximum charged off loan amount



Bivariate Analysis

Grade Vs Interest Rate

The interest rate for charged off loans is pretty high than that of fully paid loans in all the loan_amount groups



Bivariate Analysis

- There is a higher probability of defaulting among applicants who take out loans for 'home improvement' and have an annual income ranging from USD 65,000 to UD 70,000.
- Borrowers who own homes under 'MORTGAGE' status and earn between USD 60,000 and USD 70,000 annually are more likely to default.
- A heightened risk of default is associated with applicants who receive interest rates falling in the range of 21-24% and have an annual income between USD 70,000 and USD 80,000.
- Borrowers who take out loans in the range of USD 30,000 to USD 35,000 and are charged an interest rate between 15% and 17.5% exhibit an increased likelihood of default.
- Applicants who secure loans for small business purposes and have a loan amount exceeding USD 14,000 are more prone to default.

Bivariate Analysis Observation

- Individuals with 'MORTGAGE' home ownership status and loans ranging from USD 14,000 to USD 16,000 have an elevated probability of default.
- When loans are graded as 'F' and fall within the loan amount range of USD 15,000 to USD 20,000, there is a higher likelihood of default.
- Borrowers with an employment length of 10 years who obtain loans between USD 12,000 and USD 14,000 are at an increased risk of default.
- Loans that are verified and have a loan amount exceeding $16,000 are associated with a greater probability of default.
- For loans graded as 'G' with interest rates surpassing 20%, there is an elevated likelihood of default