Lending Club Case Study

July 5, 2023

Problem Statement

- Analyse the data obtained from a consumer finance company and understand how Consumer Attributes and Loan Attributes influence the tendency to default

Understanding the problem

Cleaning Data

- Extract Data from csv
- Identify and clean Null / NaN values
- Remove Columns to analysis
- Format Data
- Obtained Derived Data

Analysing Data

We have taken multiple approaches to analyse the data

- Identify relevant columns
- Univariate analysis
- Bivariate Analysis
- Plotted relevant charts

Drawing Conclusions

Conclusions were drawn by comparing relevant consumer attributes identified and loan attributes, plotting relevant charts and cross checking observations.

Data Cleaning:

Multiple steps were taken to clean the data set obtained :

Checking current Dataset after import

```
In [2]: data = pd.read_csv("loan.csv", low_memory=False)
```

```
In [3]: #check the shape of the data
        data.shape
```

```
Out[3]: (39717, 111)
```

```
In [4]: #preview the data in the csv
        data.head()
```

Out[4]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	...	num_tl_90g_dpd_24m	num_tl_op_
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	...	NaN	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	...	NaN	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	...	NaN	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	...	NaN	
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	...	NaN	

5 rows × 111 columns

Observed that the dataset contains 39717 rows and 111 columns.

Check for column with all Nulls

Clearing columns with missing values in all rows

In [9]:
```
# Cleaning up cloumns with all NaN values,as they will not contribute to our analysis
# first lets check if such columns exist from out initial hypothesis by getting sum of all null rows for every column
data.isnull().sum()
```

Out[9]:
```
loan_amnt                       0
funded_amnt                     0
funded_amnt_inv                 0
term                            0
int_rate                        0
                              ...
pub_rec_bankruptcies          697
tot_hi_cred_lim             39717
total_bal_ex_mort           39717
total_bc_limit              39717
total_il_high_credit_limit  39717
Length: 96, dtype: int64
```

In [10]:
```
# As seen ablove we have lot of columns with NaN values, hence removing them with following command
data = data.dropna(axis=1,how="all")
```

In [11]:
```
data.shape
```

Out[11]:
```
(39717, 42)
```

Found multiple columns with No data (NaN) hence removed those columns

Identified and removed additional columns

There are columns that does not help in analysis of defaulters according to our assumption.

The following columns are relevant when the loan is already provided to the customers thus they can't be used to predict if a new applicant is going to be a defaulter. Hence removing them

In [17]:
```
irrelevant_to_new_applicants = ['collection_recovery_fee','delinq_2yrs','earliest_cr_line','last_pymnt_amnt',
                                'last_pymnt_d','mths_since_last_delinq','out_prncp','out_prncp_inv',
                               'recoveries','total_pymnt','total_pymnt_inv','total_rec_int','total_rec_late_fee',
                                'total_rec_prncp','collections_12_mths_ex_med','last_credit_pull_d',
                                'pub_rec_bankruptcies','funded_amnt_inv','revol_bal']
```

In [18]:
```
data = data.drop(irrelevant_to_new_applicants,axis=1)
```

In [19]:
```
data.shape
```

Out[19]: (39717, 21)

We identified additional columns that do not contribute to our analysis in this phase and decided to remove them

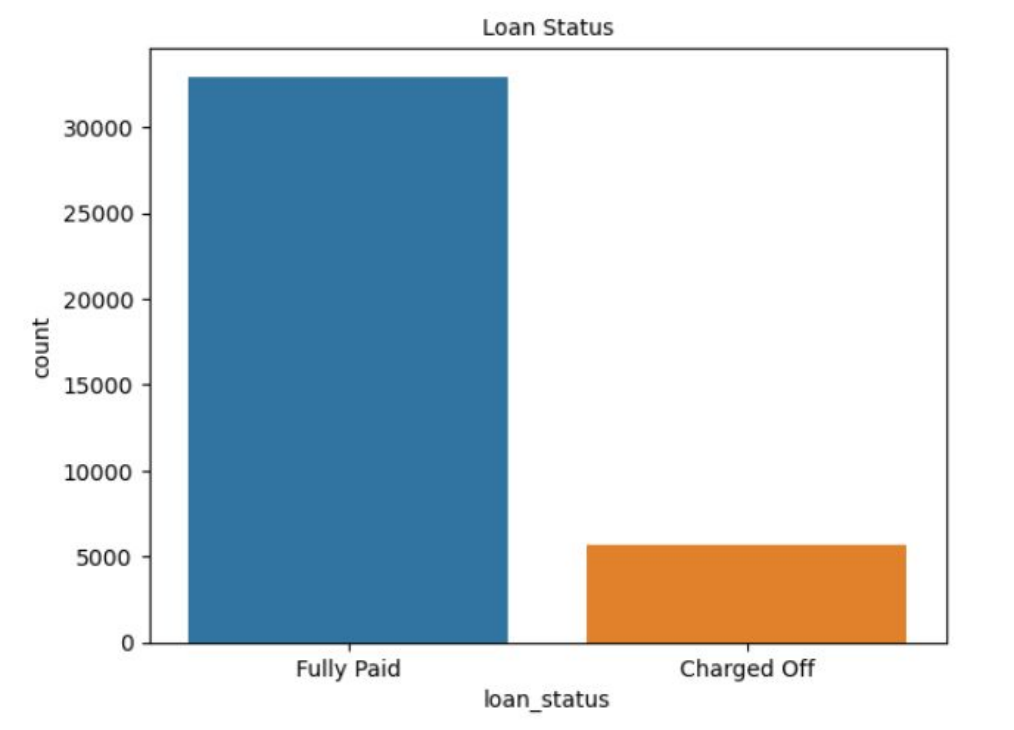
Converting Columns to relevant Data types

- Most columns were of type float64 or int64 , but a lot of columns were also object type
- For the column “term” we striped the string at the end and converted to int32
- For int_rate we removed “%” and converted to float64
- Formatted date columns to Date time format and extracted year and month for future analysis
- Excluded all entries with loan_status as Current for the analysis as we deemed it would be irrelevant for the situation
- Segmented some columns according to requirements
- Identified Outliers in columns and updated the dataset accordingly

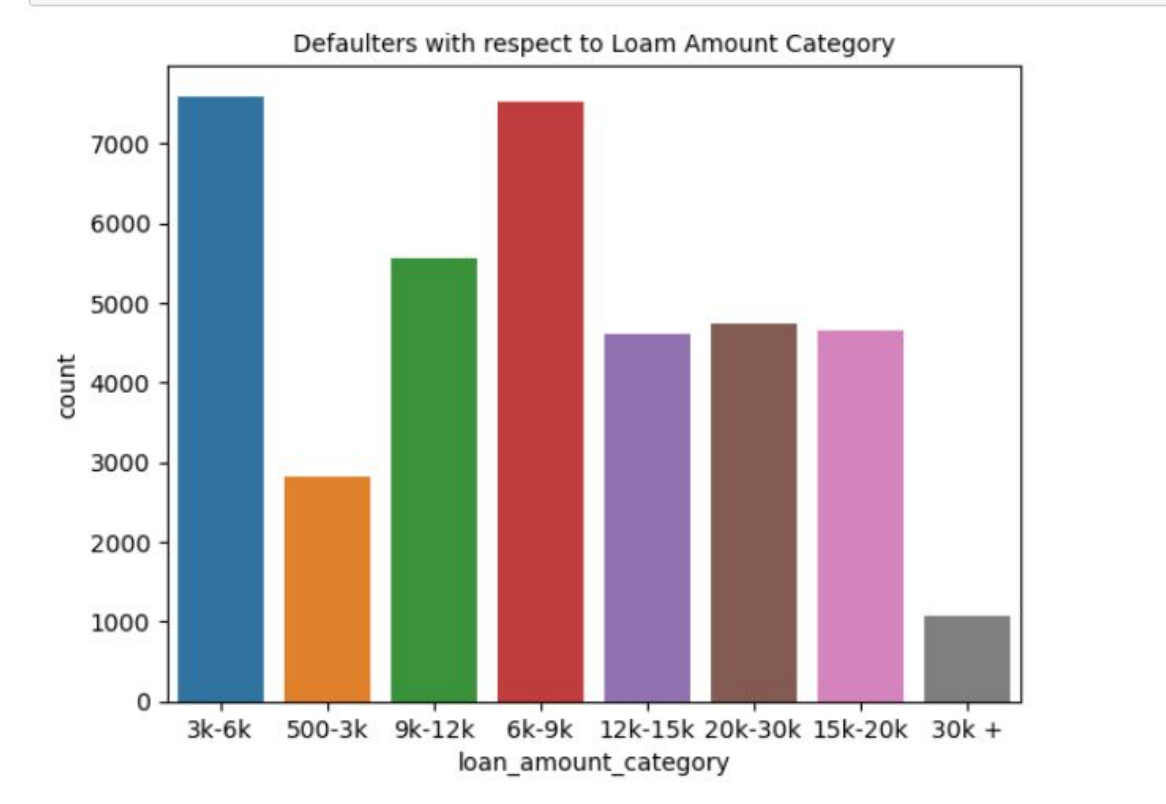
Data Analysis:

Univariate Analysis

We Identified Loan Status as the target column in the data set as it tells us who fully paid the loans and those who haven't

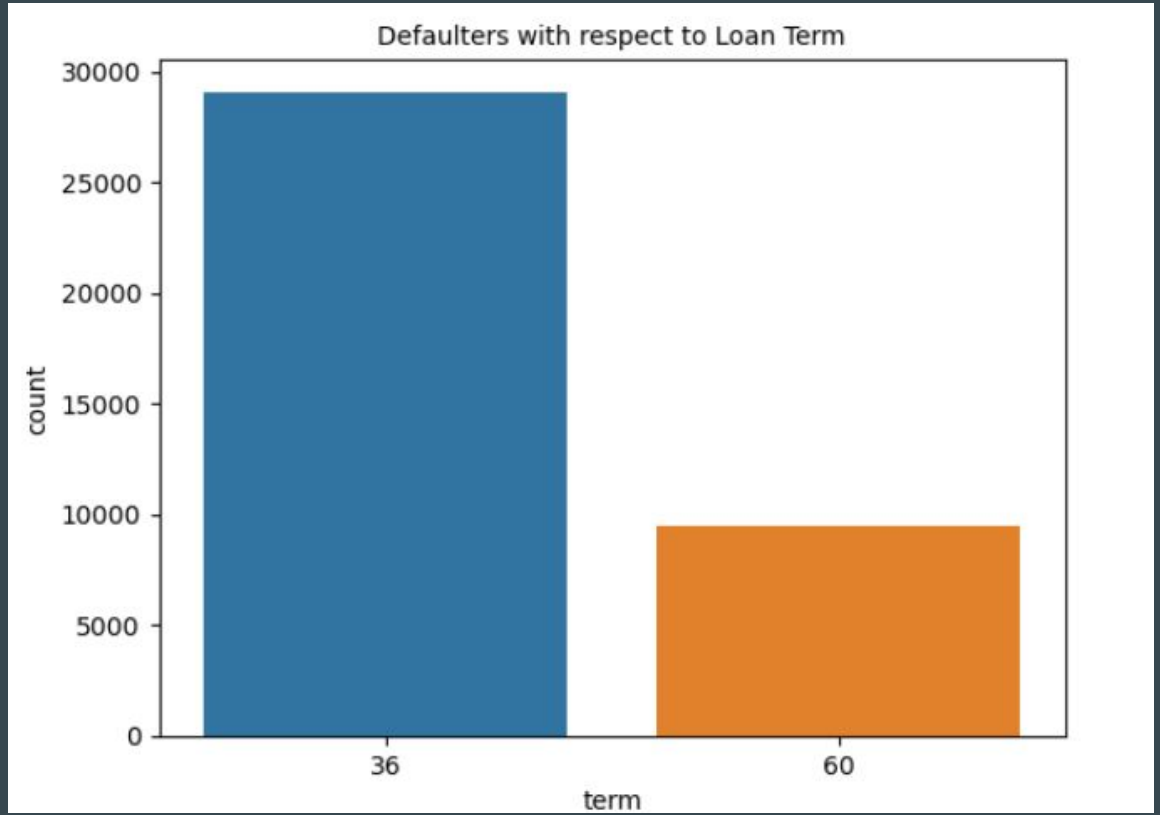


We plotted a bar chart to see the defaulters in each loan category



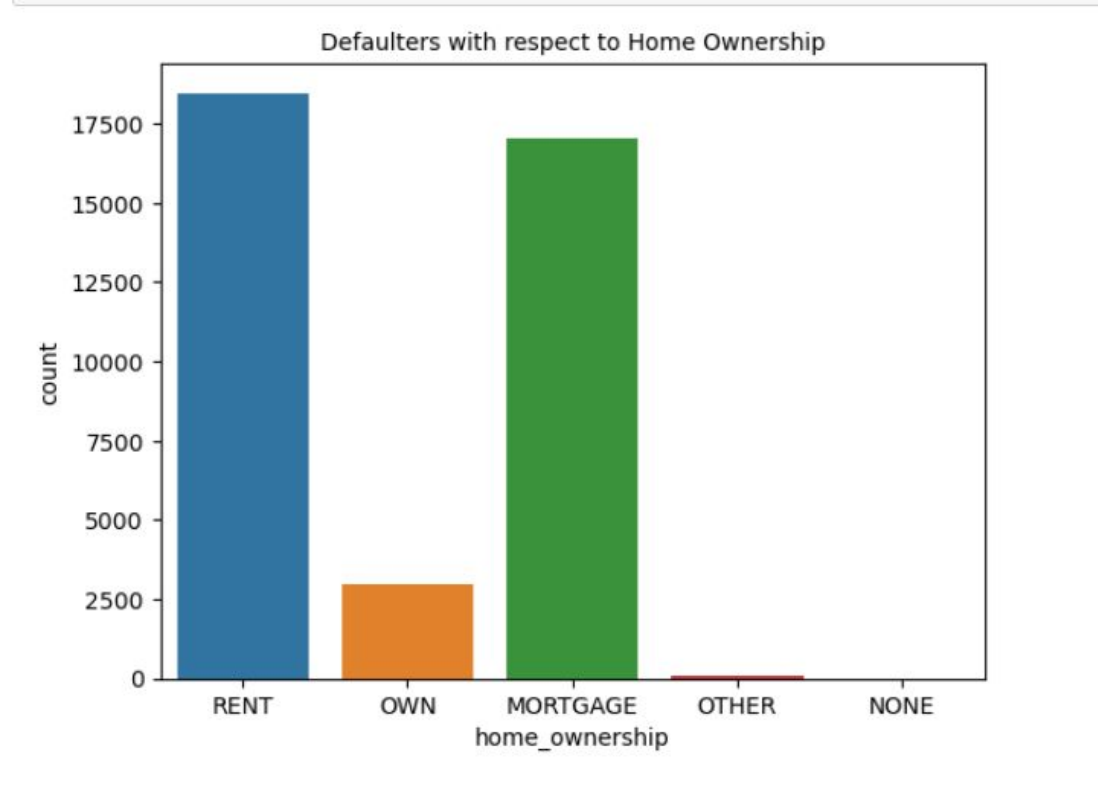
Around 7000 Defaulters when loan amount is between 3k-9k

Defaulters with respect to loan term



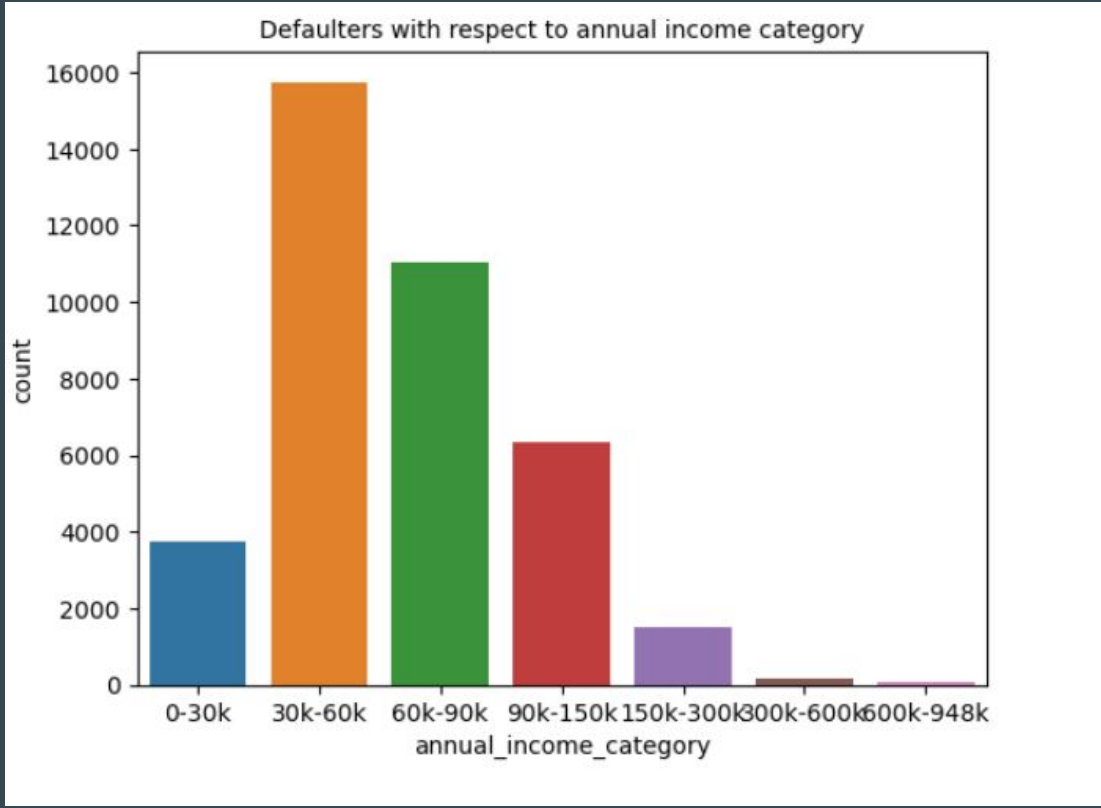
Overall we see that defaulters count for 36 months term is almost three times to 60 months term

Defaulters with respect to home ownership



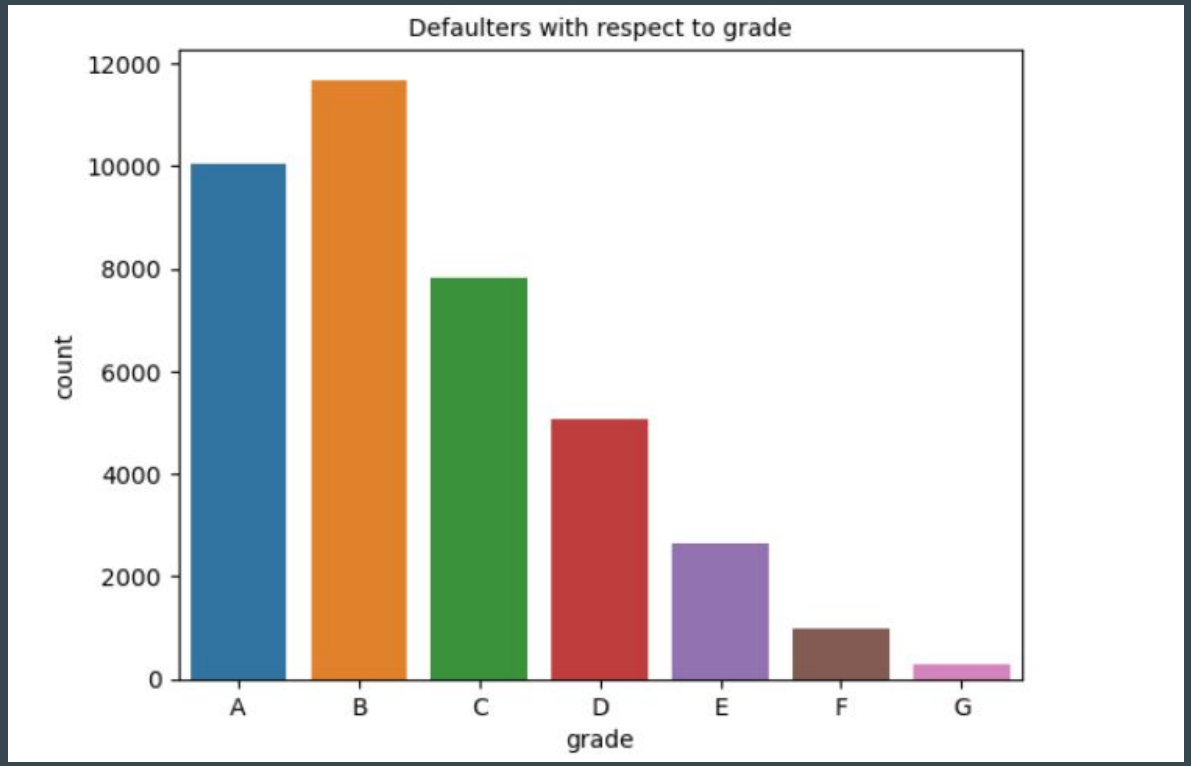
Customers who have home ownership as Rent and Mortgage tend to default more

Defaulters with respect to annual income



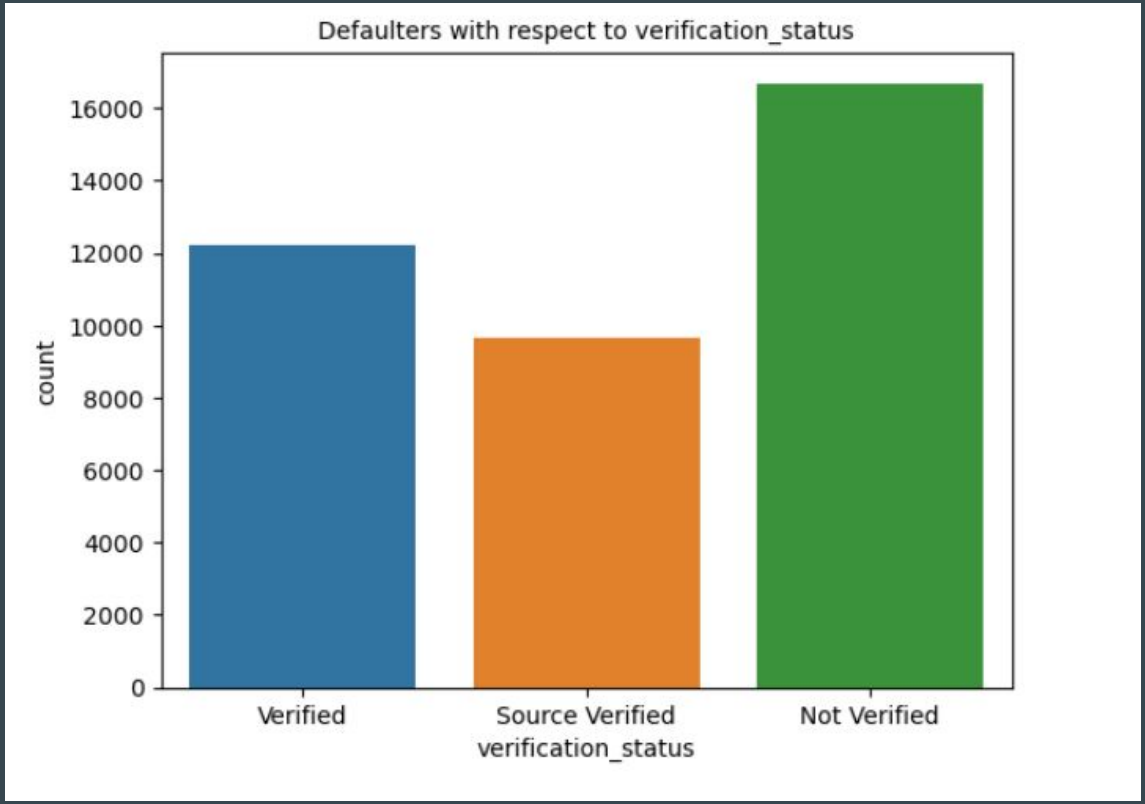
While looking at the loan status with only Annual income range of 30k to 60k tend to default on their loan more

Defaulters with respect to Grade



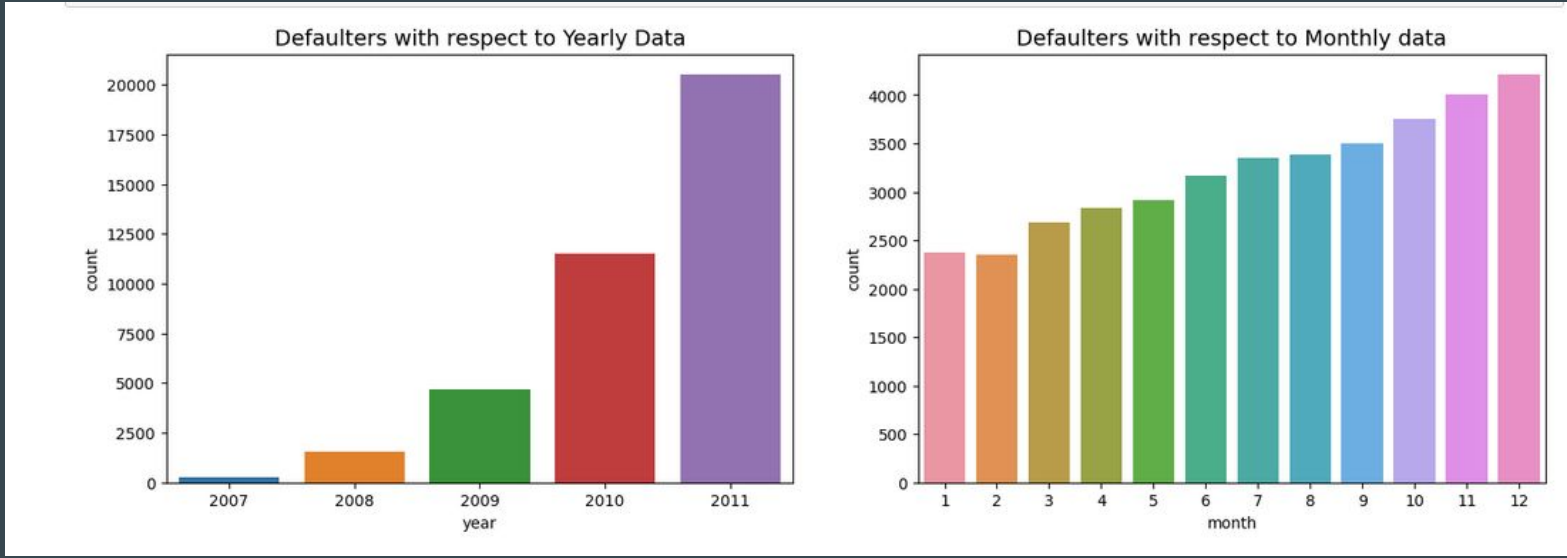
Observed A, B and C have high number of defaulters, B being the highest

Defaulters with respect to verification status



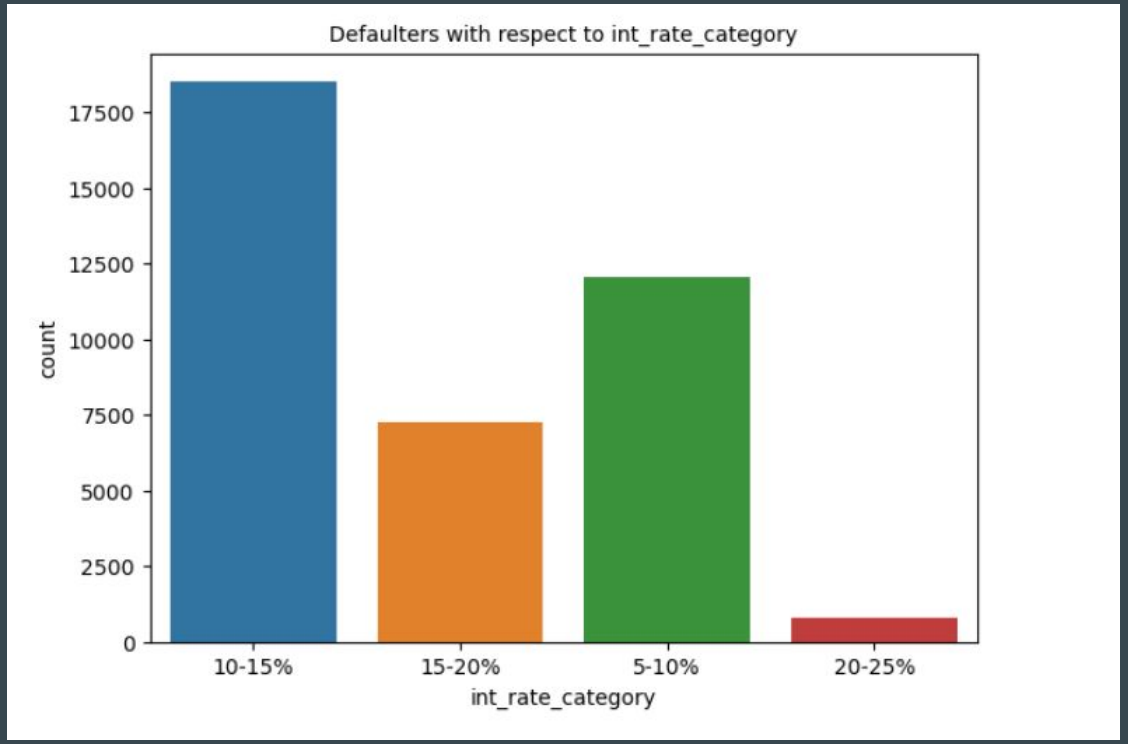
Customers whose income was unverified seem to have high probability of defaulters.

Defaulters with respect to Year and Month



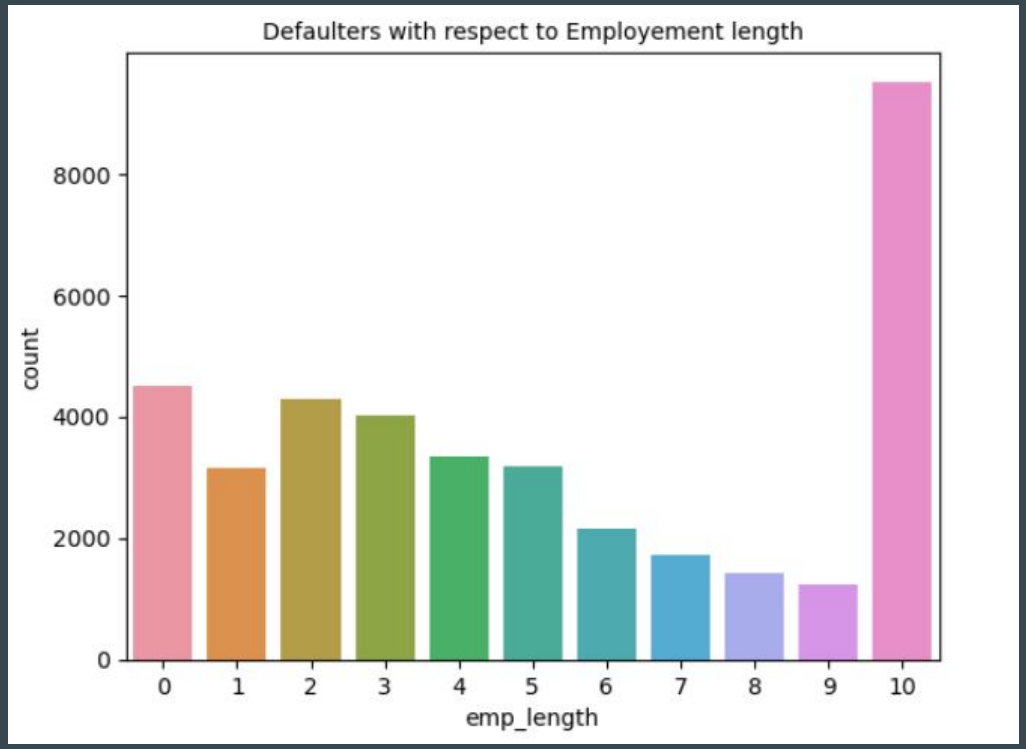
Loans sanctioned in the year 2011 have highest number of defaulters. And the trend for months shows that loans sanctioned at backend of the year have relatively more defaulters, December being the highest

Defaulters with respect to interest rate category



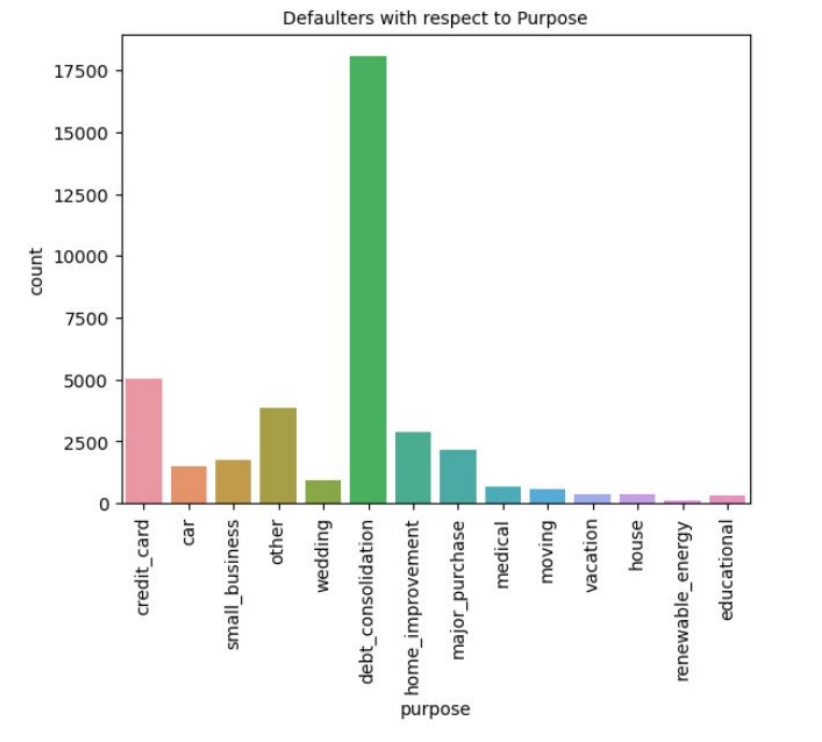
Interest rate category of 10-15% have high chance of defaulting

Defaulters with respect to Employment length



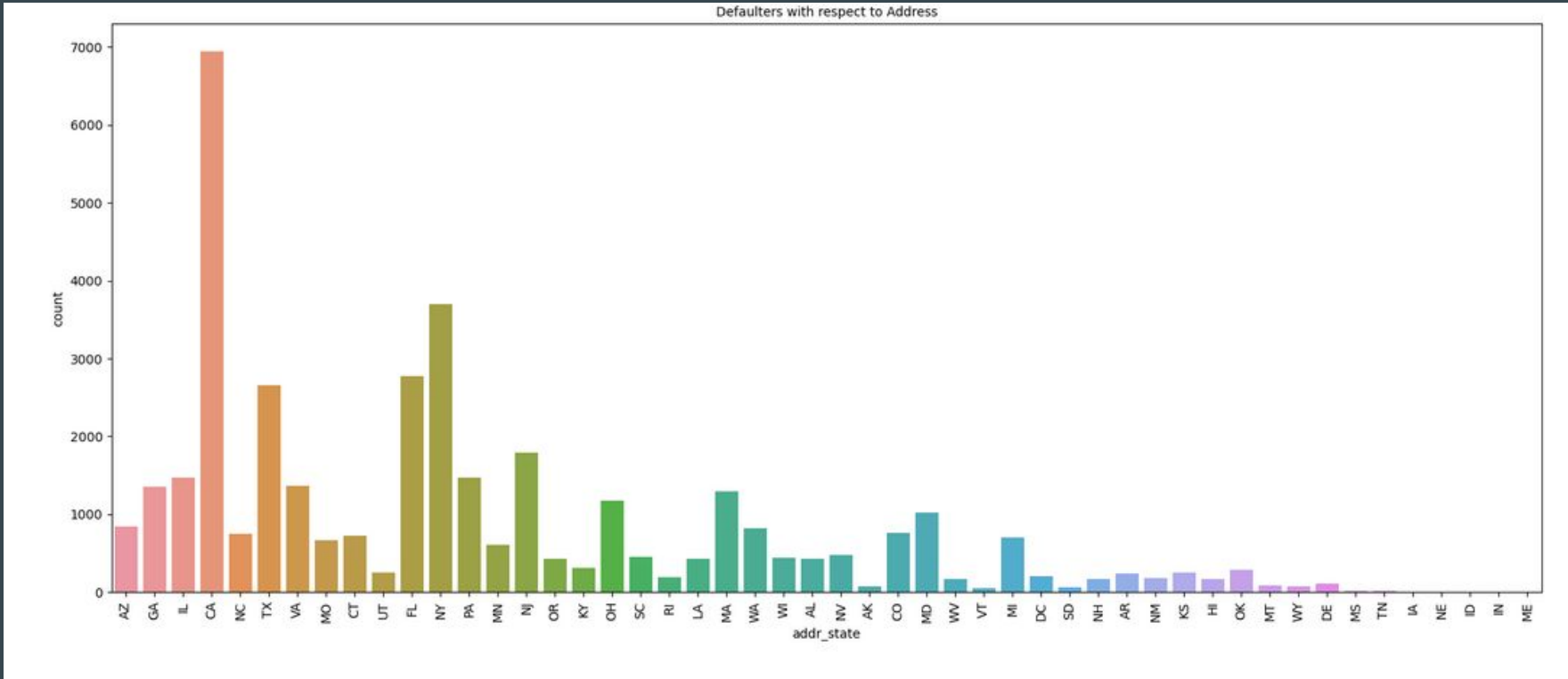
Customers with 10 years of experience tend to default more

Defaulters with respect to Purpose



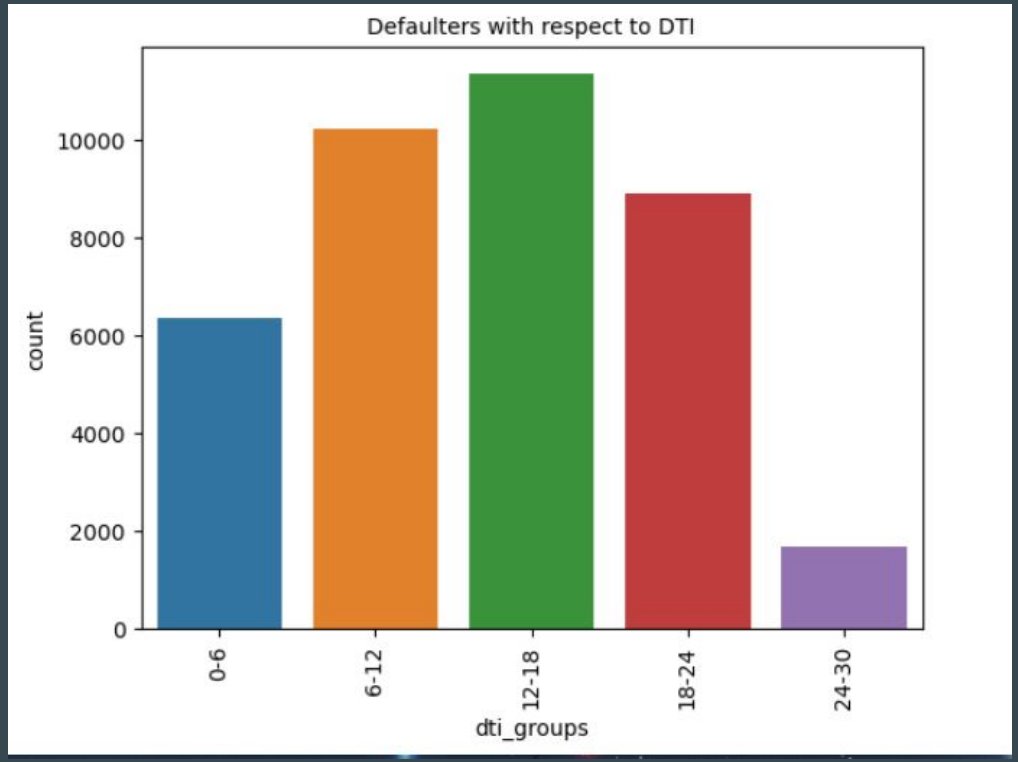
Loans taken for Debt consolidation, have high number of defaulters

Defaulters with respect to state



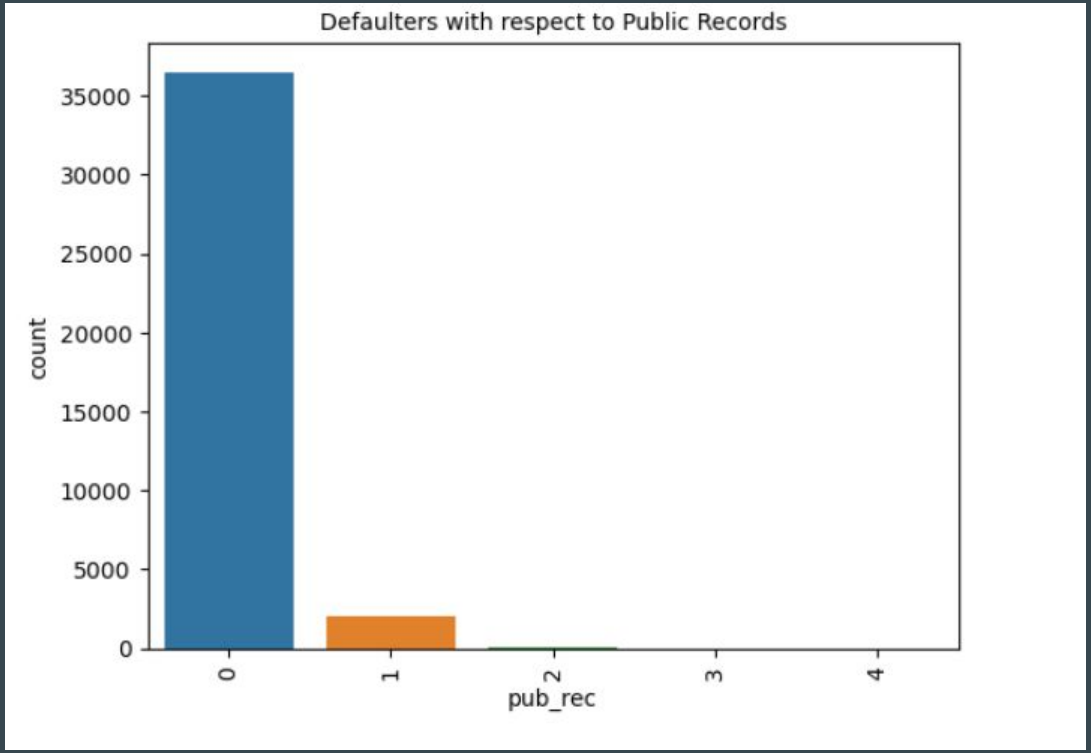
States of CA, FL and NY seem to have greater than 3000 defaulters

Defaulters with respect to DTI



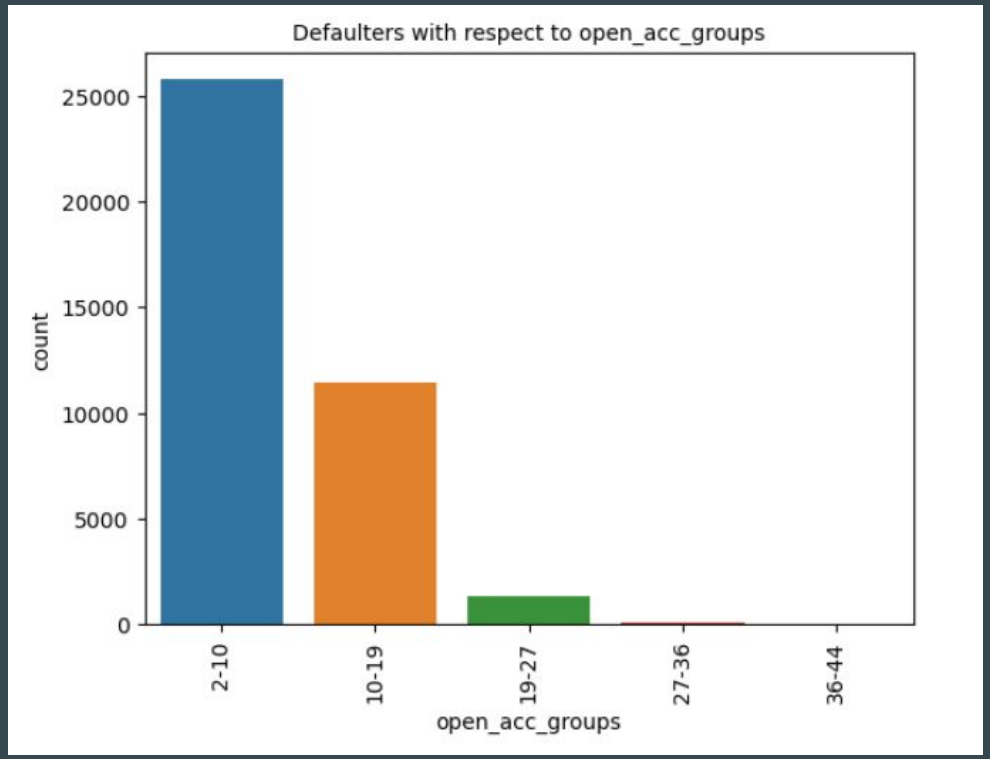
DTI range of 12-18 have high probability of defaulters

Defaulters with respect to public records



We can see that banks issue a lot of loans to people with 0 public records

Defaulters with respect to open account groups



We can see that banks issue a lot of loans to people with 2-10 open credit line,

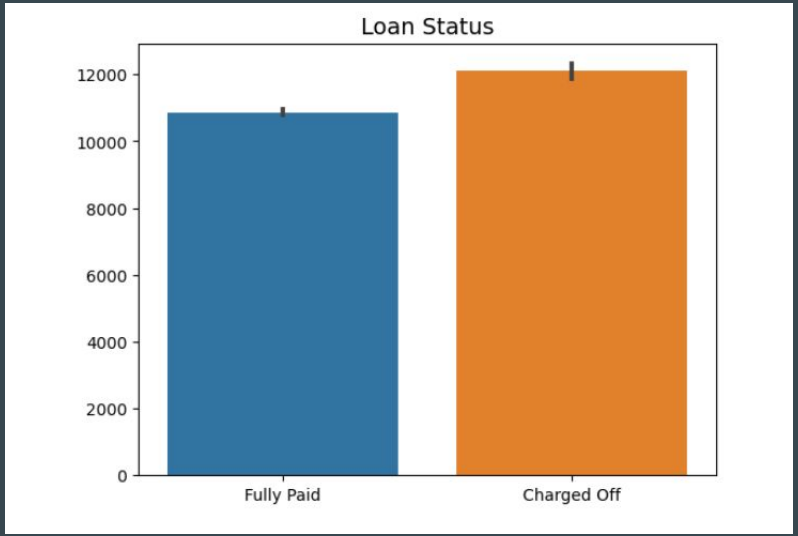
Univariate Observations

Default Indicators identified from univariate Analysis :

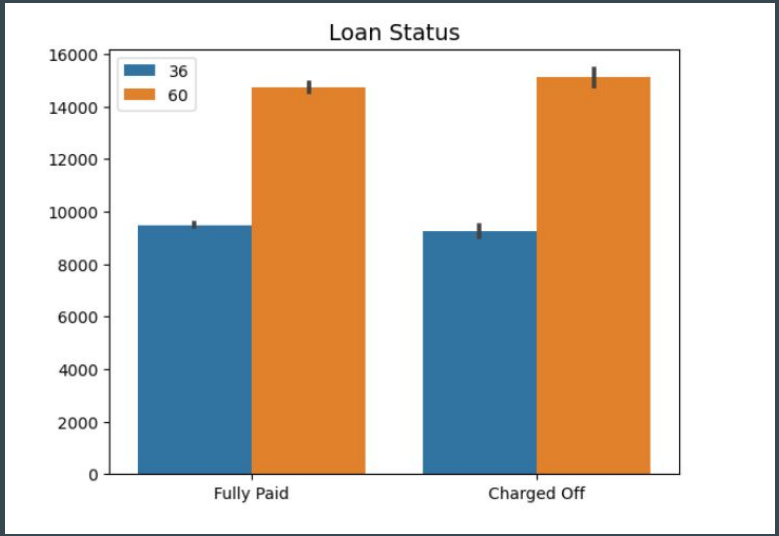
- **Loan_amount** - 3-9K loan has highest defaulters
- **Term** - Customers having 36 months have highest number of defaulters
- **Home Ownership** - mortgage and rent are the highest defaulters
- **Annual income** - 30-60k annual income range has high defaulters
- **Grade** - Grade A, B and C have high number of defaulters
- **Verification status** - Not verified customers had highest number of defaulters
- **Month** - seen a trend where customers taking loans on December defaults more
- **Interest rate** - Interest rate category of 10-15% have high chance of defaulting
- **Employment Length** - Customers with 10 years of experience tend to default more
- **Purpose** - Loans taken for Debt consolidation, have high number of defaulters
- **DTI** - DTI range of 12-18 have high probability of defaulters
- **Address State** - States of CA, FL and NY seem to have greater than 3000 defaulters
- **Public records** - banks issue a lot of loans to people with 0 public records
- **Open account groups** - banks issue a lot of loans to people with 2-10 open credit line

Bivariate Analysis

Loan status vs loan amount

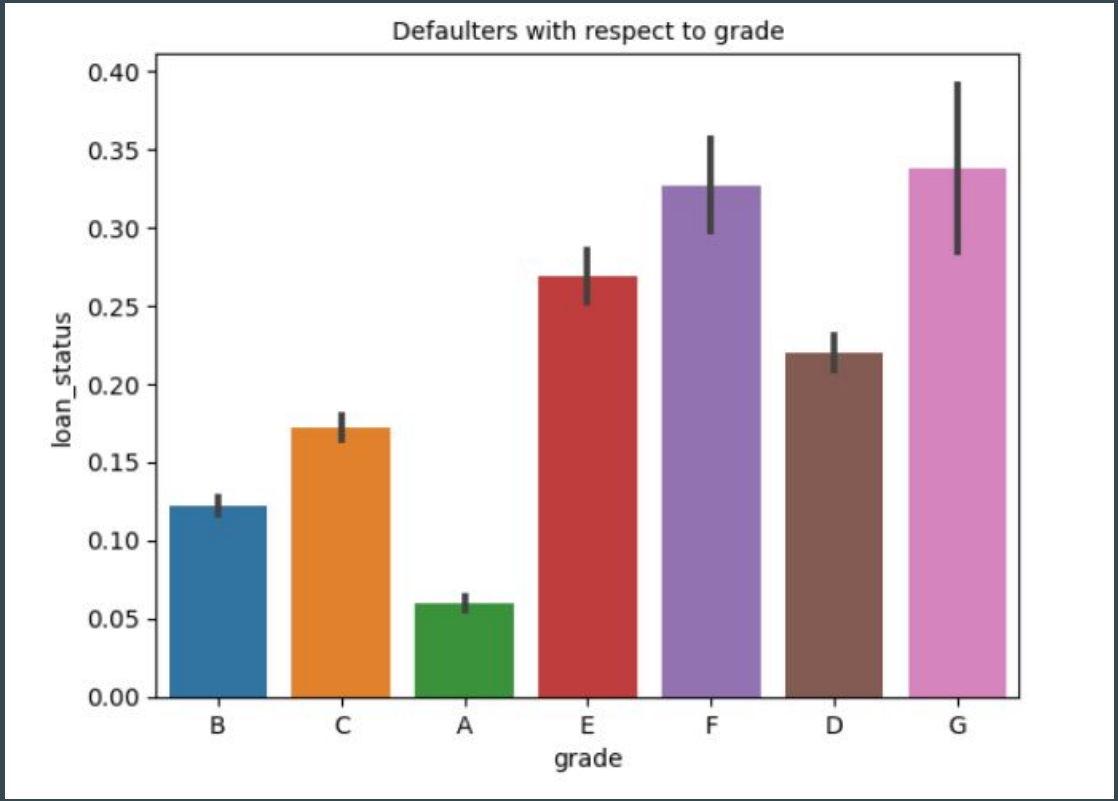


Loan status vs funded amount



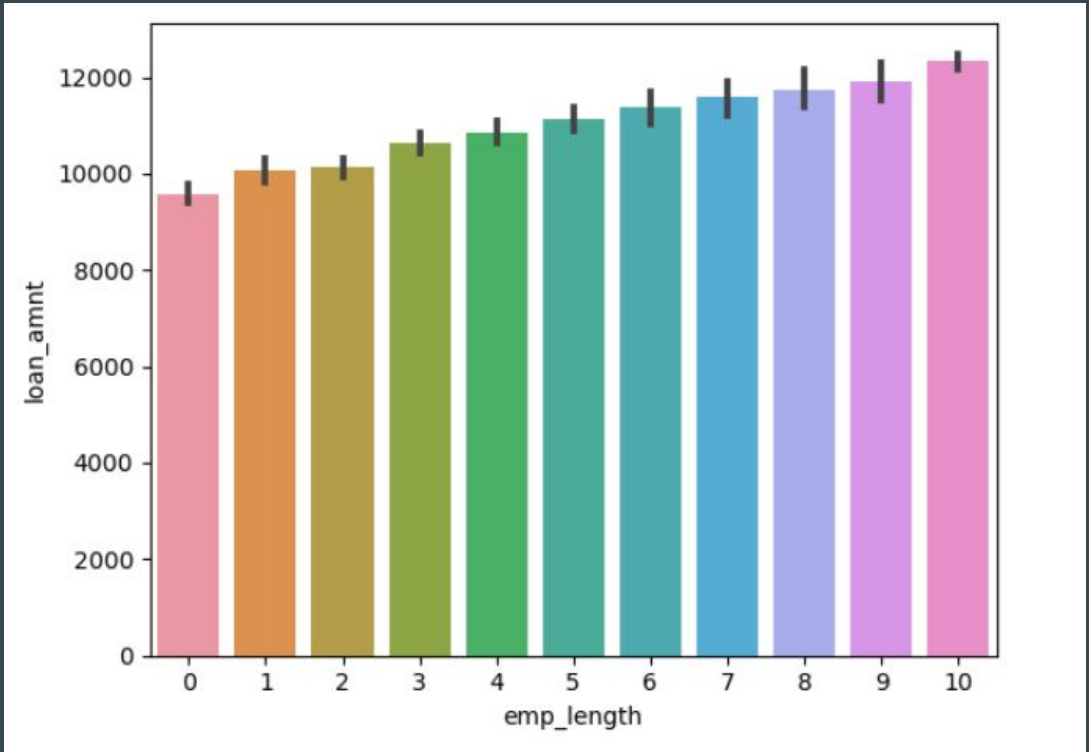
There is a high likelihood that if the loan amount is high the probability of defaulting is high for term of 60 months

Defaulters with respect to grade



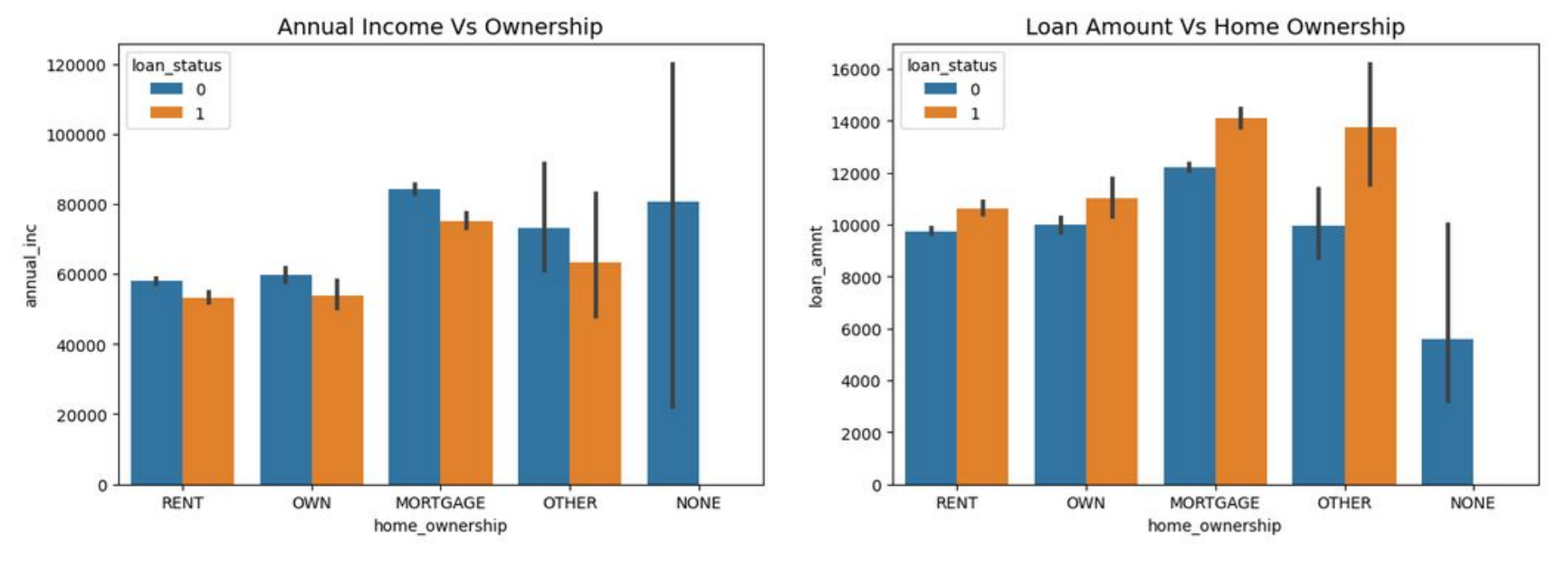
Loans with grade E,F,G seem to have high probability of defaulters.

Loan amount vs employment length

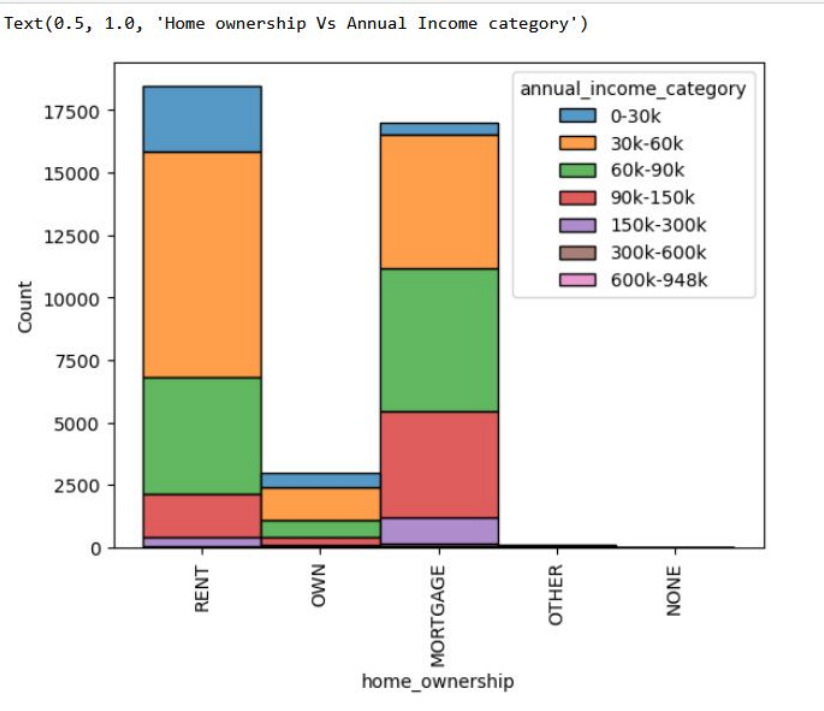


We see that people with higher experience are sanctioned higher loan amount and there is a higher probability of them not paying it back

Home ownership vs Annual income category pt.1



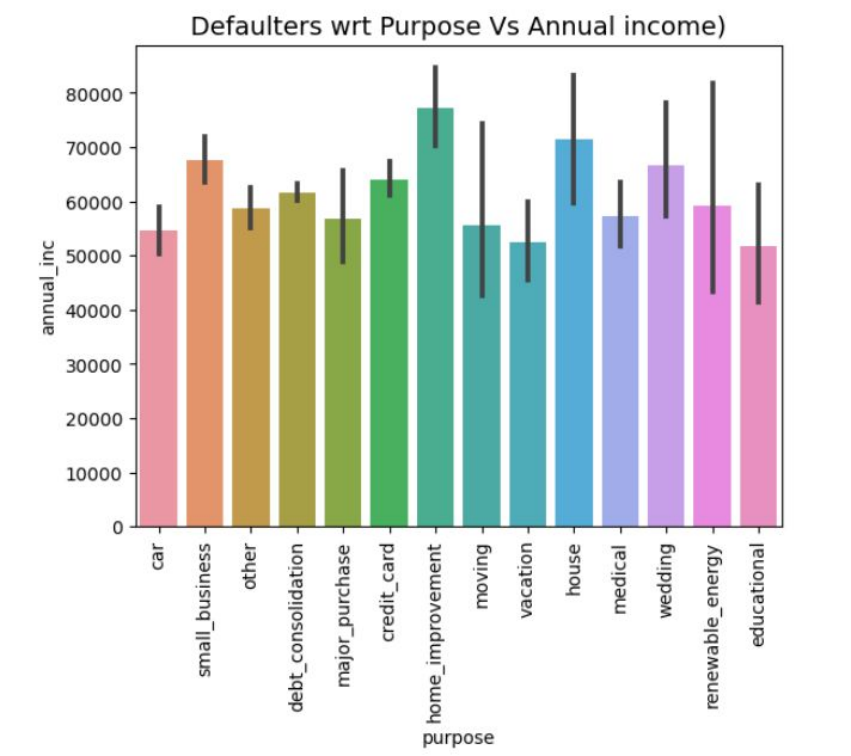
Home ownership vs Annual income category pt.2



- Customers with home ownership as mortgage are defaulting more even if their annual income is high
- Annual income category of customers with mortgage and rent have high number of defaulters is 30k-60k range.

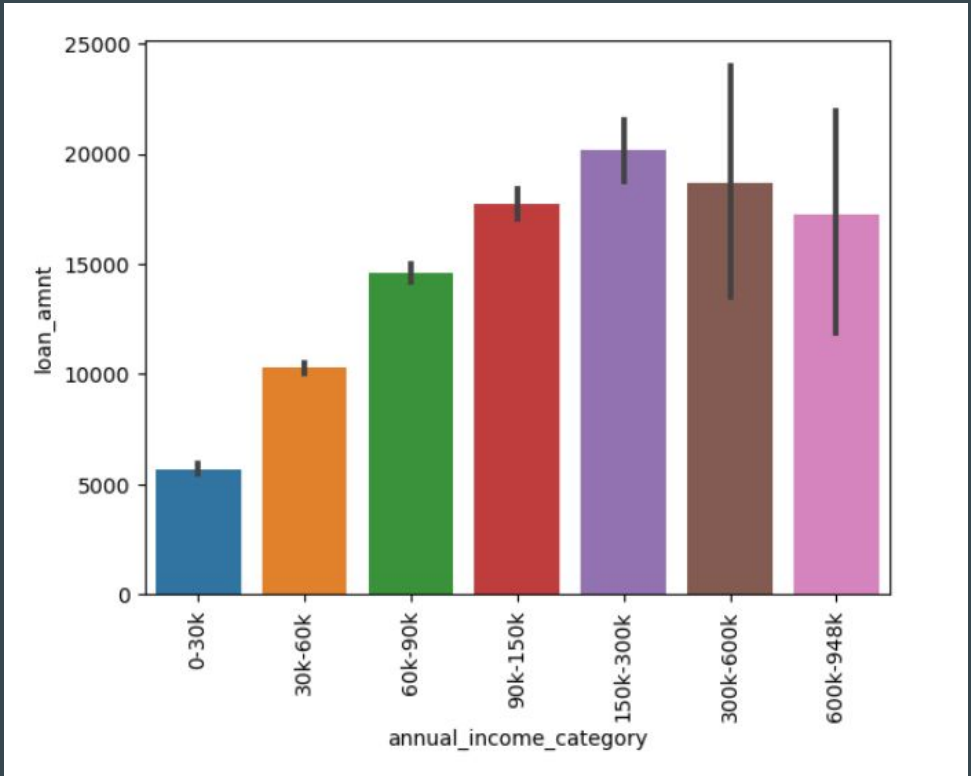
Default Indicator : Home ownership type as Mortgage and Rent with Annual income range of 30k-60k

Defaulters with respect to Purpose and annual income



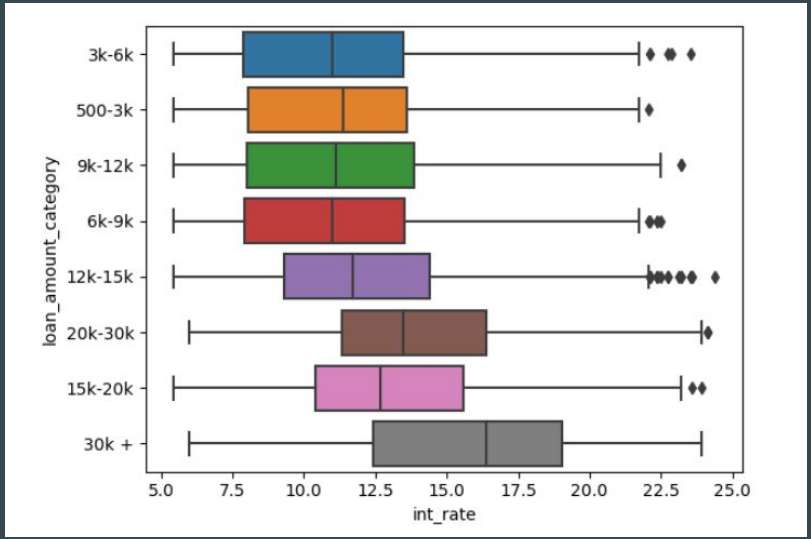
Default Indicator :Customer with annual income close to 80k and provide loan purpose as home improvement

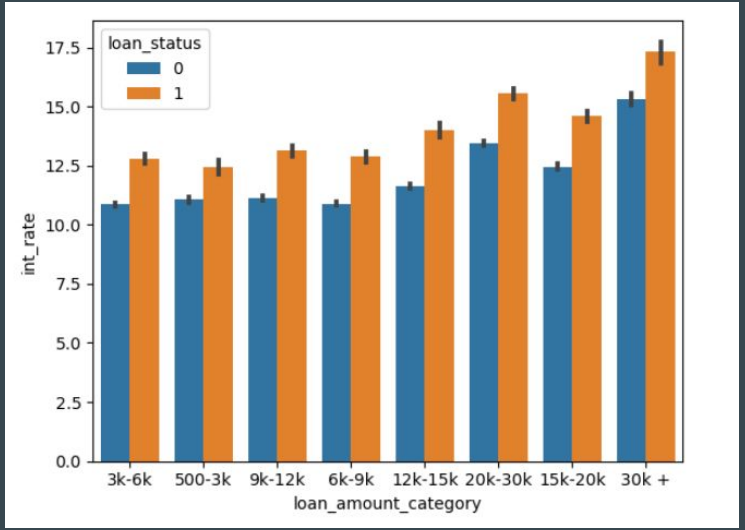
Loan amount Vs Annual Income Category



Default Indicator : Loan amount higher than 15k and annual income range between 150k-300k

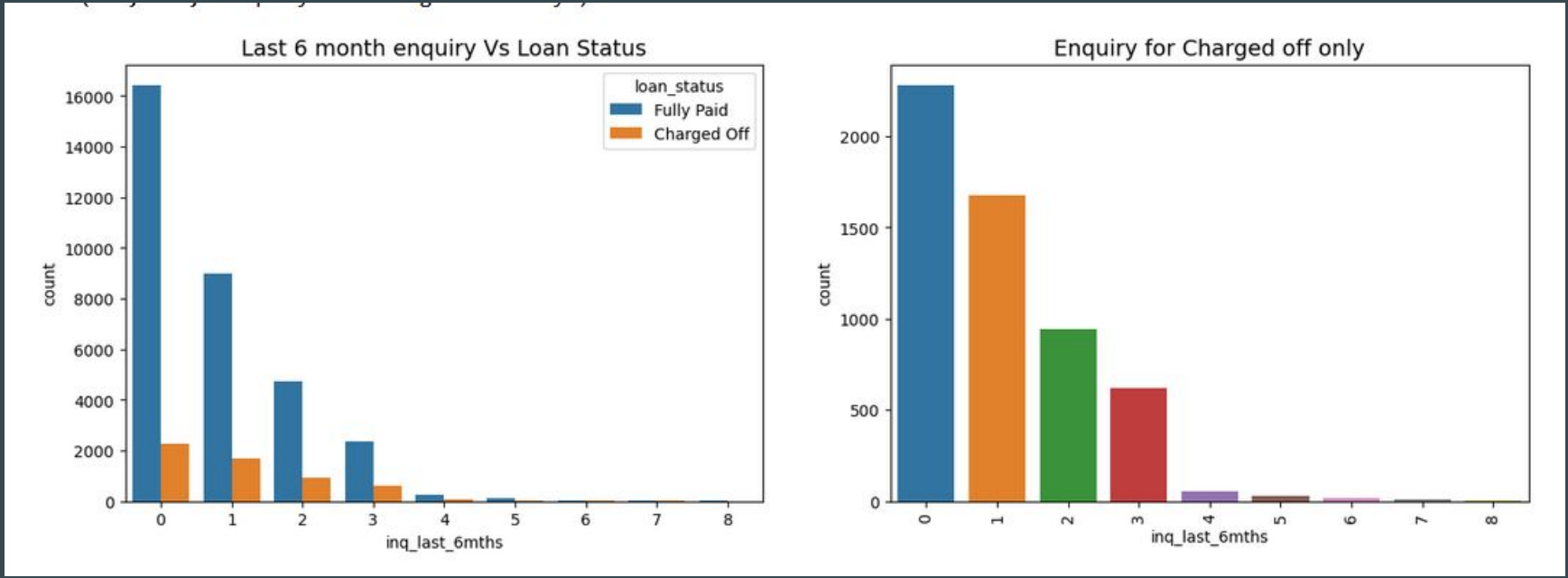
Loan amount category vs Interest rates





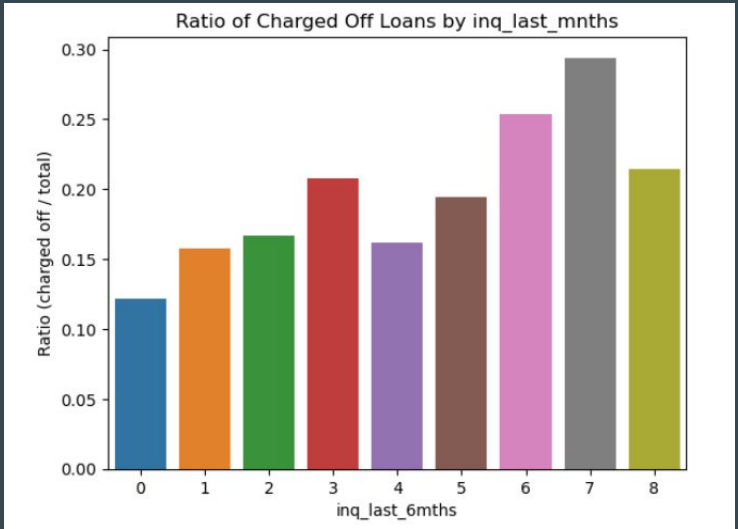
Default Indicator : Customers with loan amount more than 30k with rate of interest between 16-19%

Default rate wrt inq_last_6mths



When enquiries in last 6 months were 6 and more then the chance of defaulting is around 25% or more. Hence inq_last_6_months is good indicator for defaulters

Charged off loans wrt inq_last_6mths

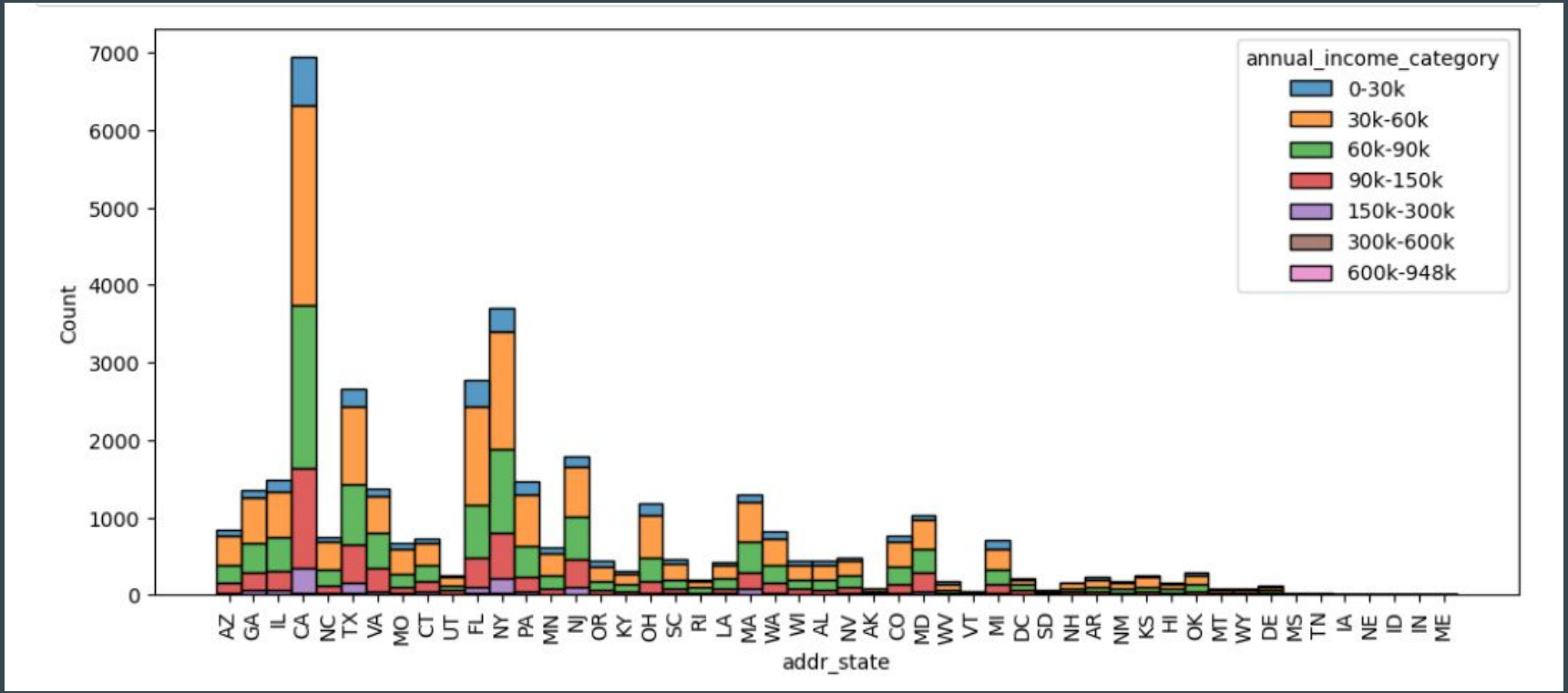


Observations :

When enquiries in last 6 months were 6 and more then the chance of defaulting is around 25% or more. Hence inq_last_6_months is good indicator for defaulters

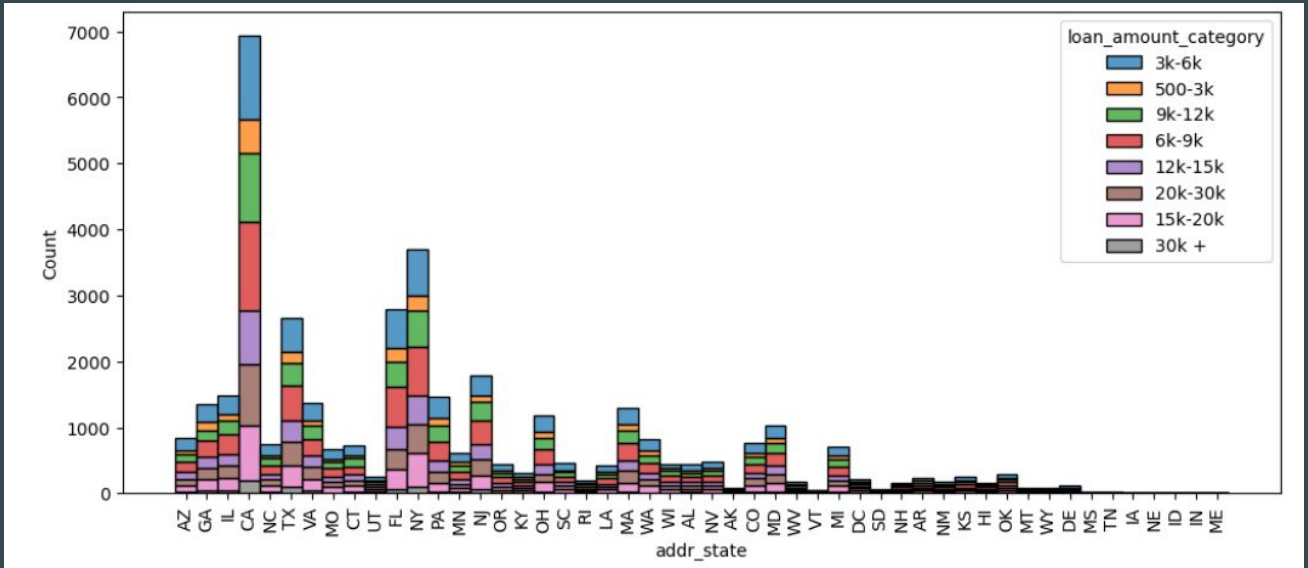
Default Indicator : inq_last_6_months greater or equal to 6

Address state with Annual income category



Bivariate analysis with annual_income_category show that 30k-60k is more in CA,FL and NA, which holds true wrt our previous analysis of defaulters in annual_income_category.

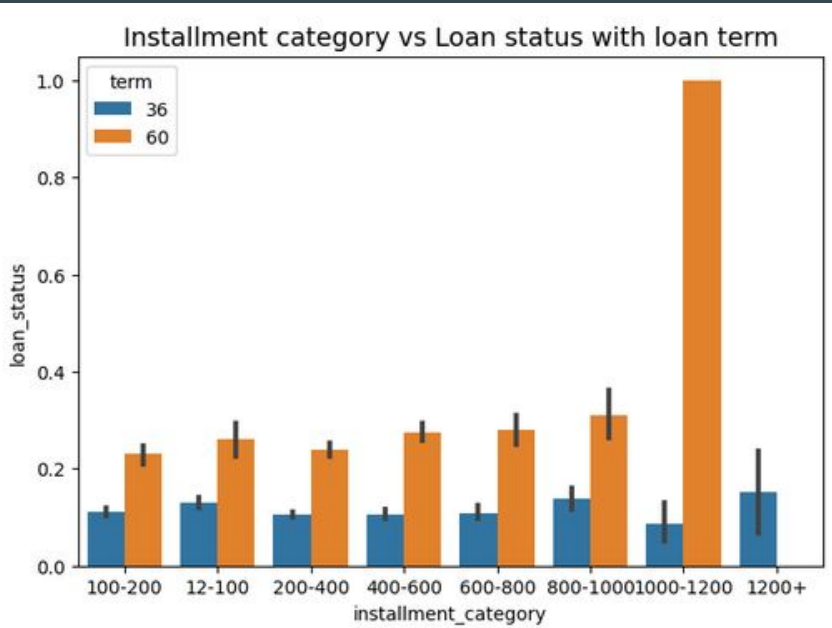
Address state with Loan amount category

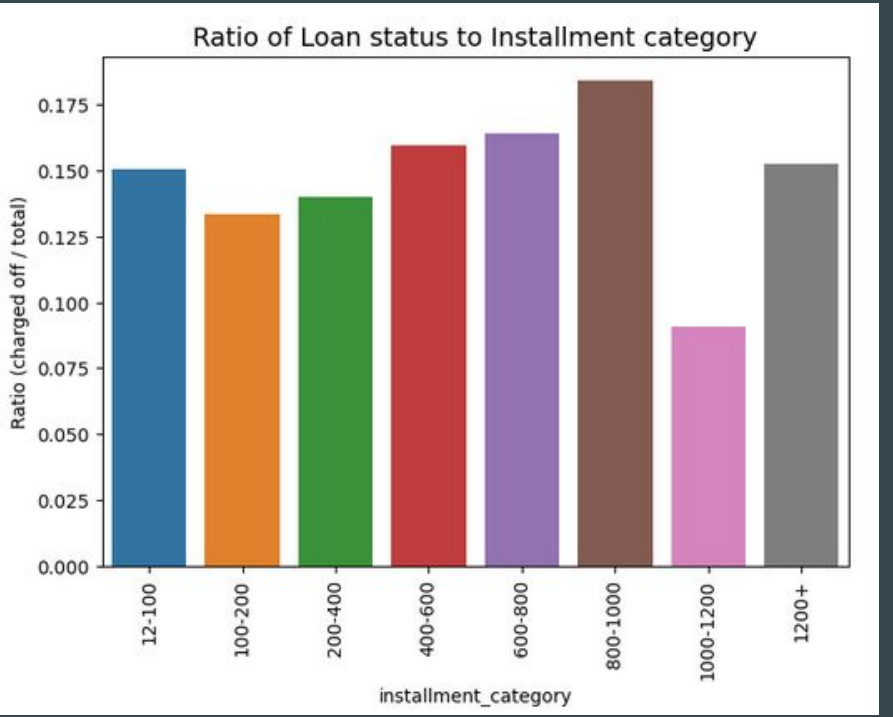


Bivariate analysis with loan_amount_category show that 6k-9k is more in CA,FL and NA, which holds true wrt our previous analysis of defaulters in loan_amount_category

Default indicators: Applicants from States of CA, FL and NY with annual income between 30k-90k and Loan amount between 6K - 9K

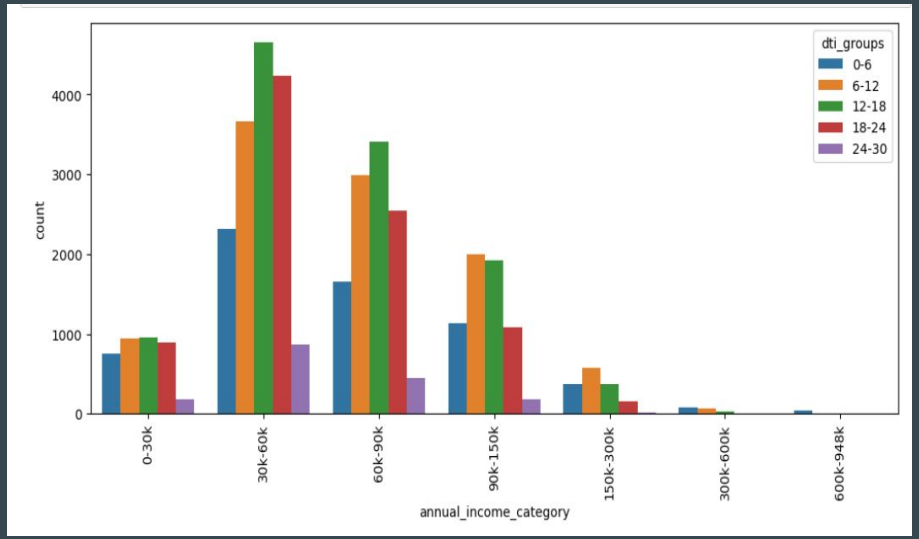
Analysis of Installment category , loan status , loan term

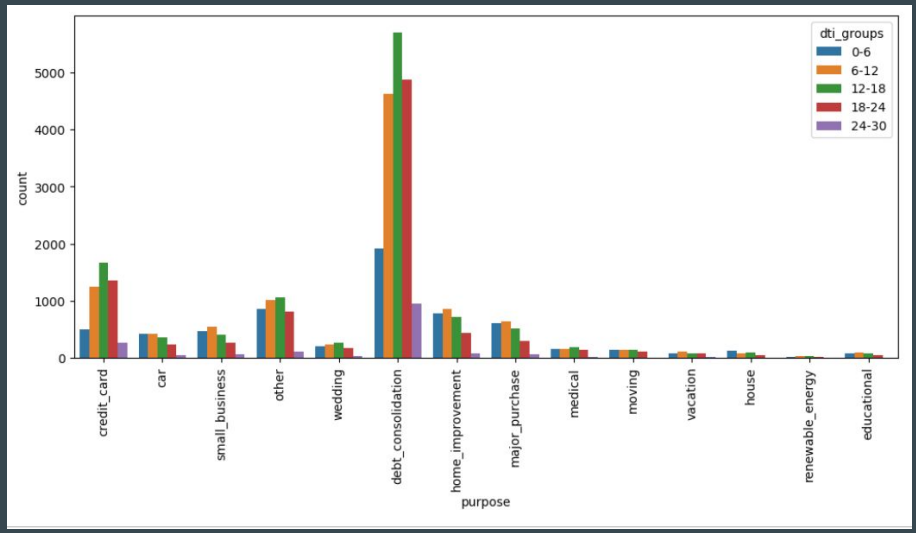




Default Indicator : Loan term=60 months and installment range 1000-1200

DTI analysis





Annual income between 30-60K and having DTI group between 12-18 have high defaulters

Default indicators : Purpose of debt consolidation and dti range of 12-18%

Bivariate Observations

Default Indicators identified from Bivariate Analysis :

- **Funded Amount :** when the funded amount is more than 10k and the term is greater than 60 months
- **Grade :** Plays a big factor, Loans with grade E,F,G are high risk
- **Employment length :** when emp Length 10 years, loan amount granted 10k+
- **Home Ownership :** type as Mortgage and Rent with Annual income range of 30k-60k
- **Purpose :** annual income close to 80k and provide loan purpose as home improvement
- **Loan Amount & Annual Income :** Loan amount higher than 15k and annual income range between 150k-300k
- **Interest Rate :** Loan amount more than 30k with rate of interest between 16-19%
- **Inquiry for Loan :** greater or equal to 6
- **Address State :** Applicants from States of CA, FL and NY with annual income between 30k-90k and Loan amount between 6K - 9K
- **Installments :** Loan term=60 months and installment range 1000-1200
- **Ownership and DTI :** Installment category of 200-400 , home ownership type as rent and mortgage with dti range of 12-18%

CONCLUSION

On observing the trends from the data set obtained the following factors should be considered by the bank before approving loans :

1. If customer is having home ownership as renting or mortgage , and with the annual income in the range 30-60k , it can be a high risk profile
2. If customers purpose for loan is home improvement and annual income is around 80k, huge loans should not be approved
3. Longer term loans with installments of 1000-2000 is a high risk
4. Bank should reduce providing loans between 6K - 9K for the states for CA, NY and FL
5. Bank should discourage providing loans for lower grades
6. Banks issue a lot of loans to people with 0 public records and 2-10 open credit line , and which ideally indicate low risk, but according to the data have a high number of defaulters
7. Bank should discourage sanctioning Loan amount more than 30k with rate of interest between 16-19% as it shows high number of defaulters
8. Banks should discourage approval of loans if Inquiry for Loan in the last 6 months greater or equal to 6
9. If the DTI range of 12-24% we found that there is a high likelihood they are taking loan for debt consolidation , which increases risk

The Team

Mahalakshmi Totad

Vivek P Rajeev