LENDING CLUB ASSIGNMENT

SUBMISSION

Name:

VINYAS SHUKLA

SRIKANTH PADMANABHUNI

Abstract :

- A consumer finance company provides loans to different kinds of people in urban. But company feels difficult in finding whether a person pays/defaults the loan
- If person doesn't pay the loan, then the investor faces the loss. If a person who pays complete loan doesn't get approval for loan then also investor and bank gets loss
- We as a Data scientist need to find the factors that are responsible for finding the person can pay loan or default based on the data provide by the company

Approach for Analysis:

- Load the data from the csv to start the analysis.
- **Cleaning Data:**
  - Find the percentage of null data available in the dataset. We can see that a large of data is null.
  - There are some columns which are having 100% null data in them. So, keeping them in data set is not much useful. So, lets drop those columns
  - There are some more columns which are having >30% of null data. So, lets drop those columns instead of imputing and making data biased towards the median
  - After removing those null data still, we can find data with <10% null for following columns **'emp_title, emp_length, title, revol_util, last_payment_d, laste_credit_pull_d, collections_12_mths_ex_med, chargeoff_within_12_mths, pub_rec_bankruptcies, tax_lines'.**
  - **pub_rec_bankruptcies** is a kind of categorical data and most of the columns are with data '**0**'. So, instead of imputing data and create bias over the 0 lets drop rows with null data
  - Most of the other columns are kind of categorical/non numeric data. So, instead of imputing them lets drop the rows and make data clean.
  - After cleaning all the data we can find total columns available are 53.
  - Among them, we can further drop the columns based on domain knowledge and the data dictionary provided. We can understand which are not required for our analysis.
  - After dropping columns which are not required for our analysis, we have the total remaining columns are 23.
  - Since we need to work on the data related to Charged Off and Fully Paid loan data, so lets remove the rows which are having loan status **'Current'** [Since, theya re currently paying and not marked as Default].

- There are columns like emp_length which are having data like <1 and >10+, so, lets clean them by removing <,>,+ symbols and lets make it as numerical categorical data
- There is one more column called term, which is having data like 36months, 60months. Lets remove months from it and make it as numerical categorical data

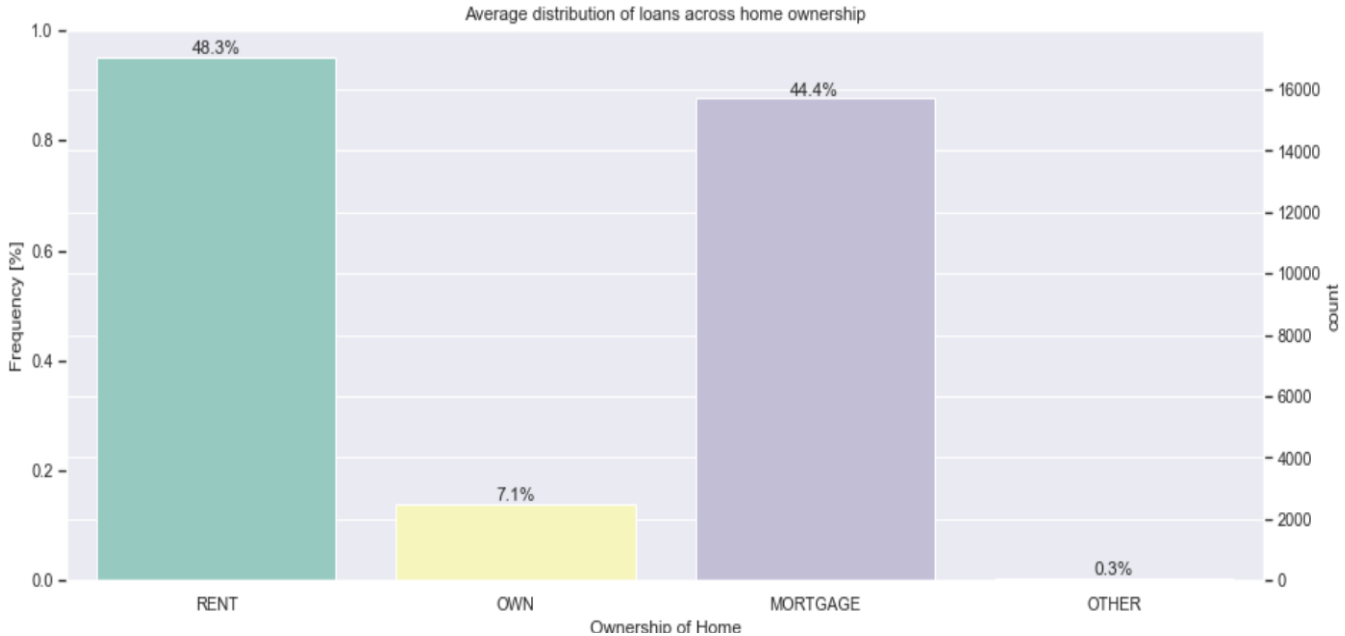
Deriving New Variables:

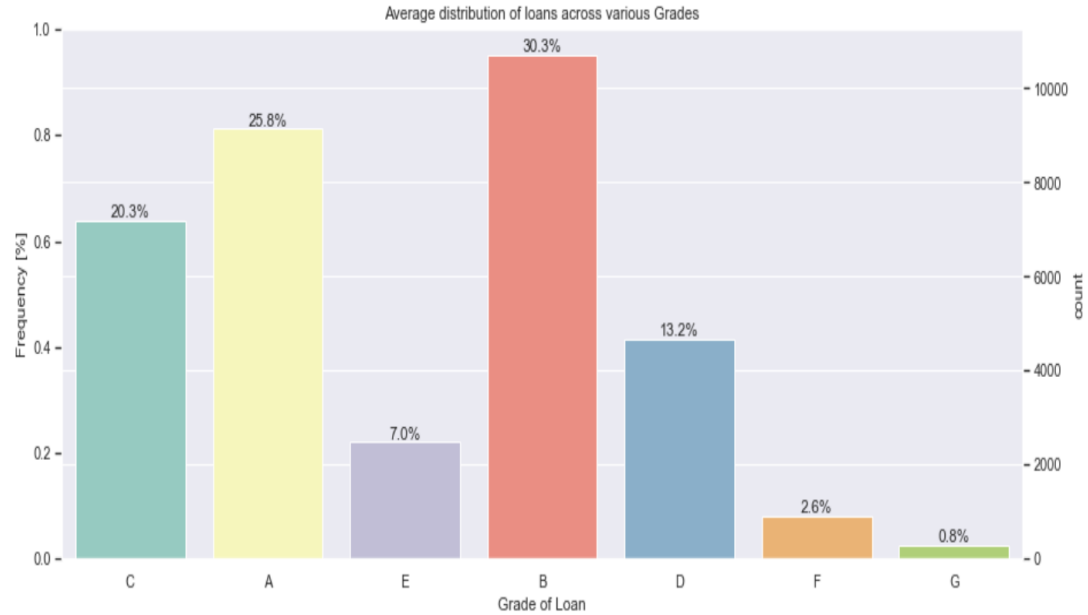
- Now, from the existing we can derive new column, which can help further in our analysis part.
- Deriving **'year'** from the 'issue_d' column, which helps in knowing in which year that loan is approved instead of a date.
- Deriving **'l_t_ai'** from the data of 'annual_inc' and 'loan_amnt'. This column says the ratio of the loan amount taken and the annual income, which helps in deriving further more analysis.
- Deriving column called **'income_bin'** from income. It is providing the category like 1,2,.. Based on the ranges of annual income. This helps in making the analysis on the persons based on the range of annual income instead of one particular annual income
- Deriving the columns **'dti_range', 'l_t_ai_range', 'exp_level'** from **dti, l_t_ai, emp_length** columns respectively. It provides the categories to those columns data and helps in making analysis based on thatv range instead of a particular value of the data.
- Now, based on the above derived columns and the existing columns we can start working on the Univarient and BiVarient analysis and Plot graphs forn them to get more insights.

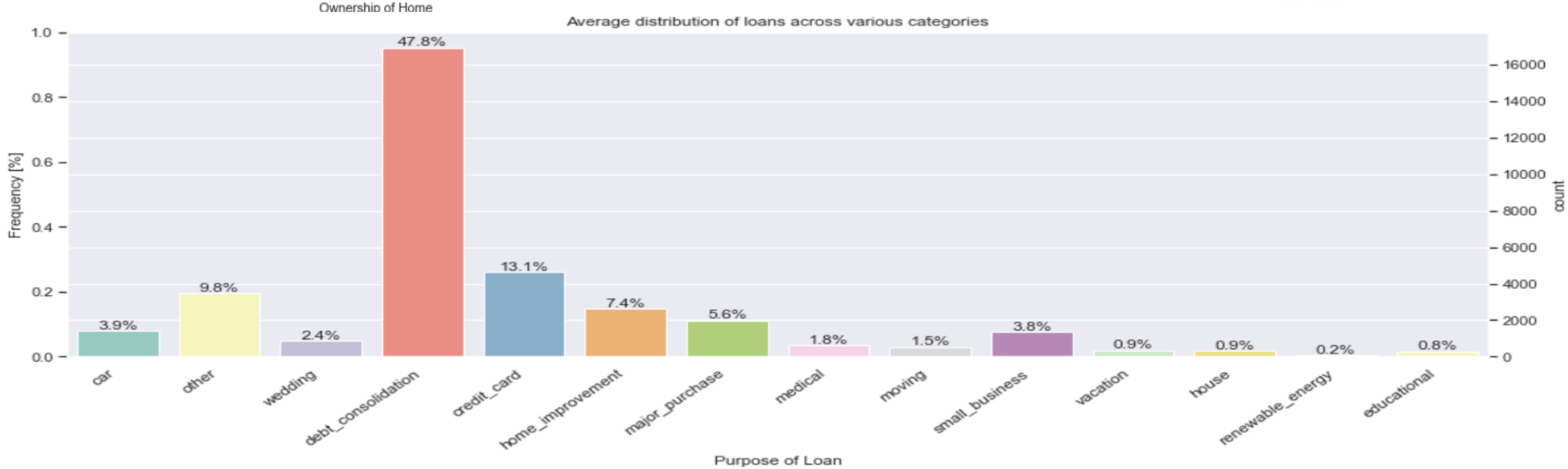
Univarient Analysis:

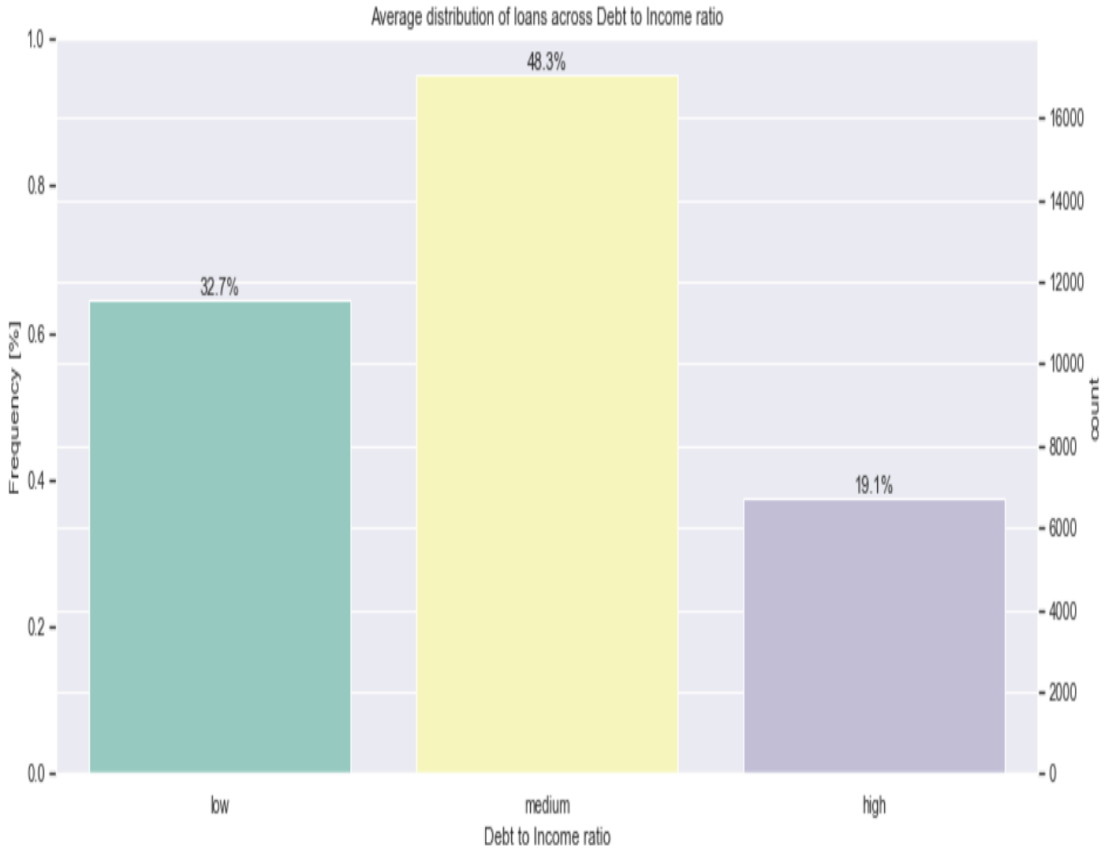
- Plot the graph for different columns to find its frequency among the data. Like, how such kind of data is spread among the dataset.
- **Loan Status:** Among the dataset 14% of people are having with default loan status, where as 86% of people paid loan Fully
- **Issue Year:** From the analysis we can find that people started taking loans from the year 2008, but it gradually increases while the year increase. From 2010 to 2011 people who are taking loans got increased enormously to 54%.
- **Purpose of Loan:** Most of the people are taking loan with the purpose of 'debt_consolidation' (Means paying outside debts with loan) and 'credit_card' (loans on credit card).
- **Grades and Sub grades:** Finance companies provided some grades and subgrades to people based on some conditions. Among them people with grades **B,A,C** and with subgrades **A4, B3, A2** are the people taking/getting approvals for loans more frequently
- **Years of Experience/Experience levels:** People having the10(>=10+) and 1(<=1) years of experience are taking/getting approvals for loans more frequently. When it comes to derived column of experience level Juniors are getting more approvals for loans frequently
- **Home ownership:** People who are rented/mortgage home are having more approvals for loans than own house people.
- **State:** People from states **'CA', 'NY', 'FL'** are having more approvals for loans than for other state people
- **Debt to loan and loan to income ratio:** People with medium DTI range and Medium, High loan to income ratio people are having more approvals for loans than other people.
- From the above analysis we can see that
  - people having less years of employment experience
  - People having rented/mortgaged house
  - People with the grades B,A,C and subgrades A4, B3, A2.
  - People from states CA, NY, FL.
  - People with medium DTI range and Medium, High loan to income ratio

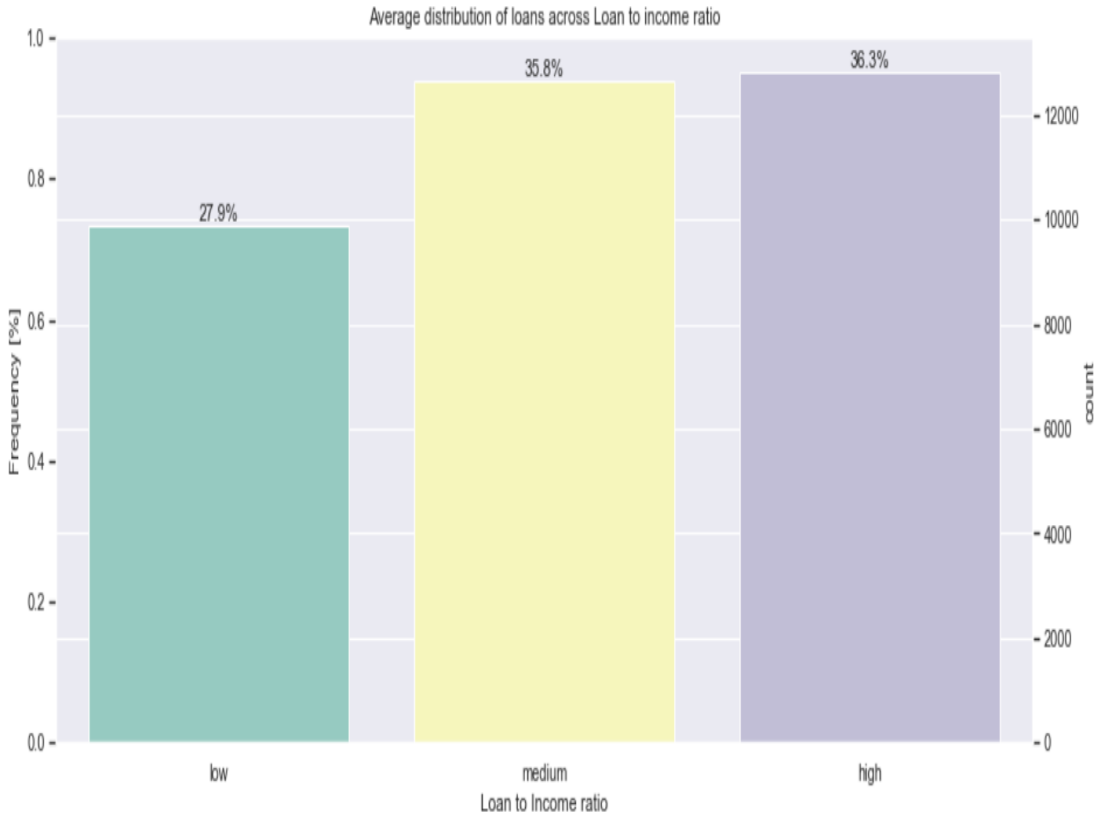
  are having more approvals for loans.
- Now, lets see the above data graphically.


Average distribution of loans across home ownership
Frequency [%]
count
48.3%
7.1%
44.4%
0.3%
RENT
OWN
MORTGAGE
OTHER
Ownership of Home


Average distribution of loans across various Grades
Frequency [%]
count
20.3%
25.8%
7.0%
30.3%
13.2%
2.6%
0.8%
C
A
E
B
D
F
G
Grade of Loan


Average distribution of loans across various categories
Frequency [%]
count
3.9%
9.8%
2.4%
47.8%
13.1%
7.4%
5.6%
1.8%
1.5%
3.8%
0.9%
0.9%
0.2%
0.8%
car
other
wedding
debt_consolidation
credit_card
home_improvement
major_purchase
medical
moving
small_business
vacation
house
renewable_energy
educational
Purpose of Loan


Average distribution of loans across Debt to Income ratio
Frequency [%]
count
32.7%
48.3%
19.1%
low
medium
high
Debt to Income ratio


Average distribution of loans across Loan to income ratio
Frequency [%]
count
27.9%
35.8%
36.3%
low
medium
high
Loan to Income ratio

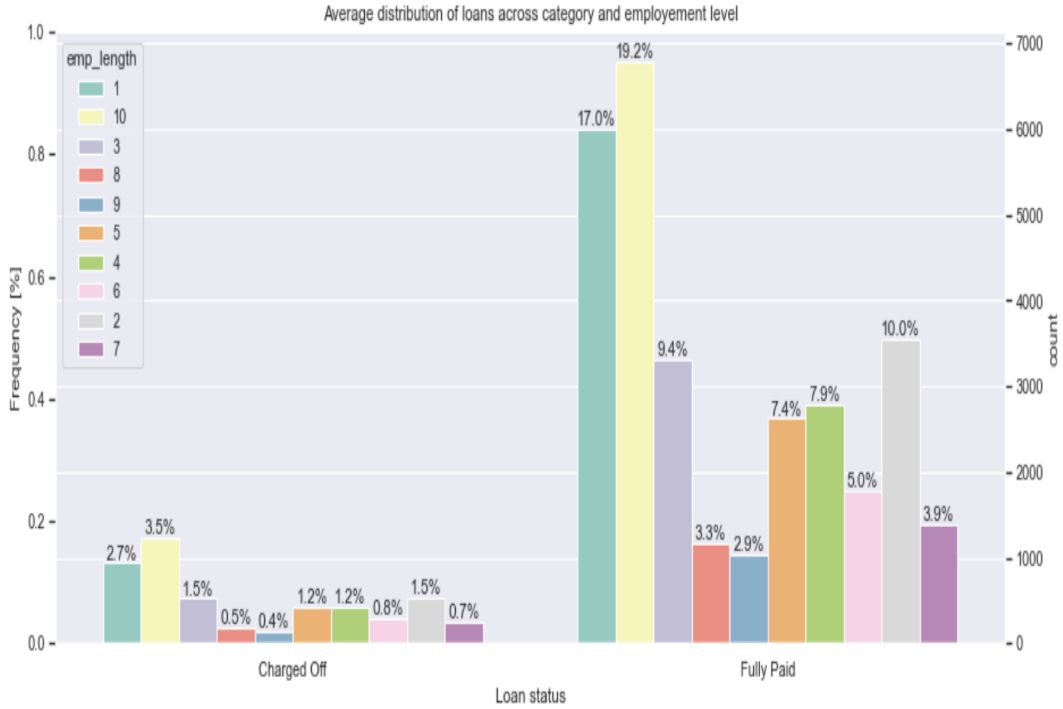
Segmented Univarient Analysis:

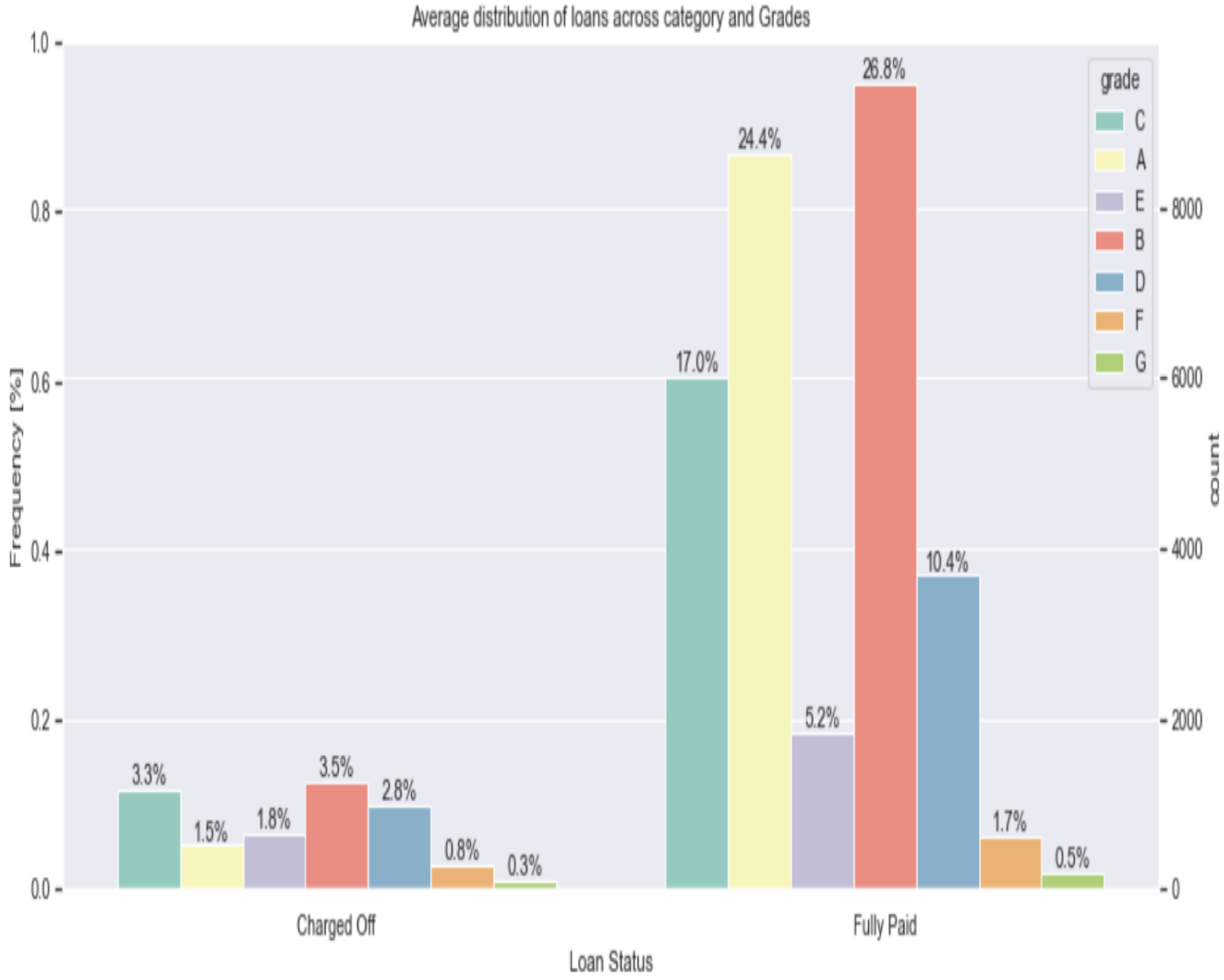
- On making an analysis for experience level of employee for different loan status we have found that, employees of junior level (<=3 Years) are having the more chances of default than other employees
- On making analysis on DTI(Debt to Income ratio) range, people with medium(**>10.00 and <= 20.00**) DTI range are having more chances of default than the other people. Similarly, people with High(**>0.20**) Loan to Income ratio range are having more chances of Default.

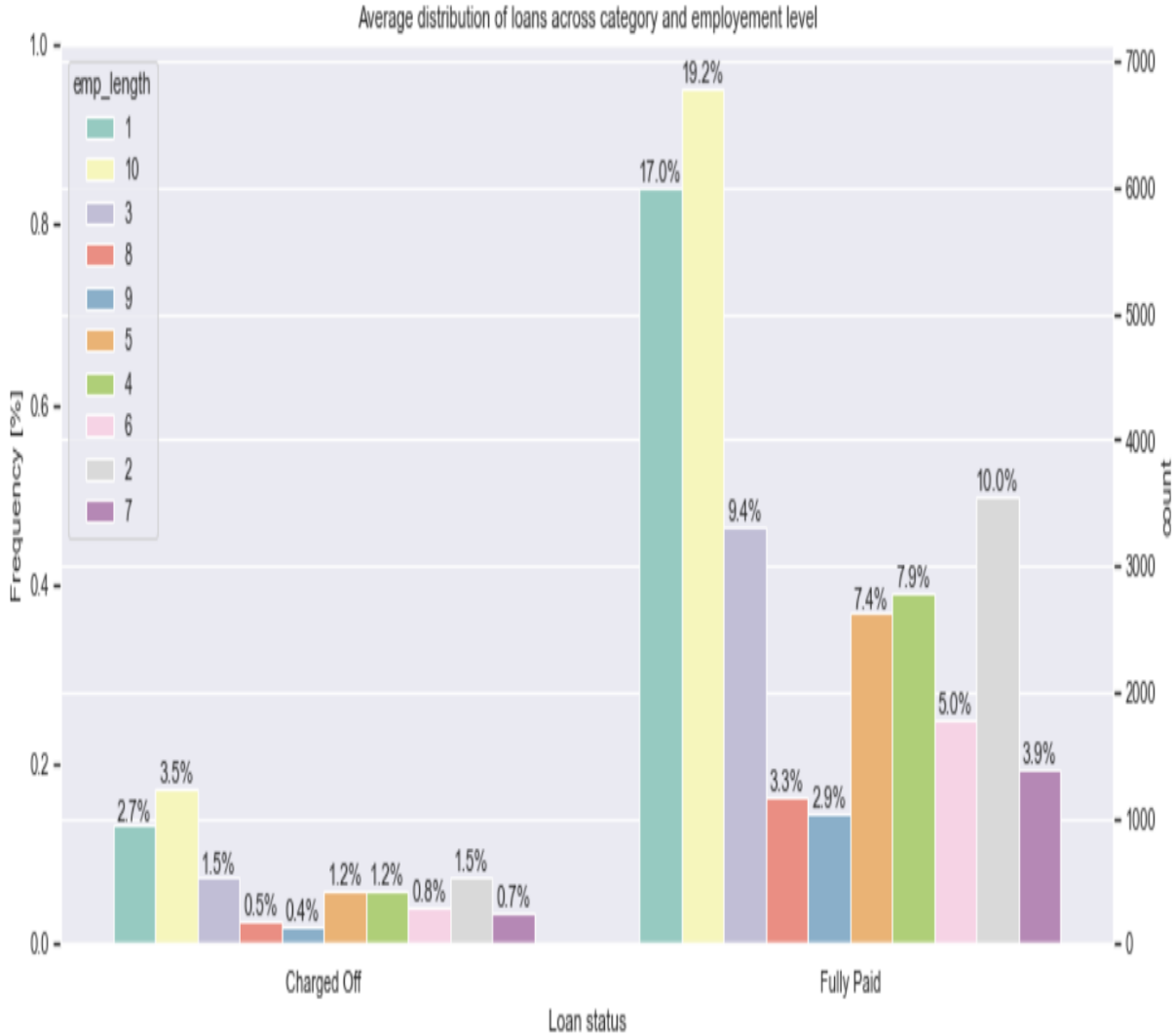
It means that people of having less employment experience will be having more DTI and LTI ratio since, there income will be less and expenditures including loan instalments will be high. Since, there will be more chance of defaulting the loan due to the burden of other expenditures.

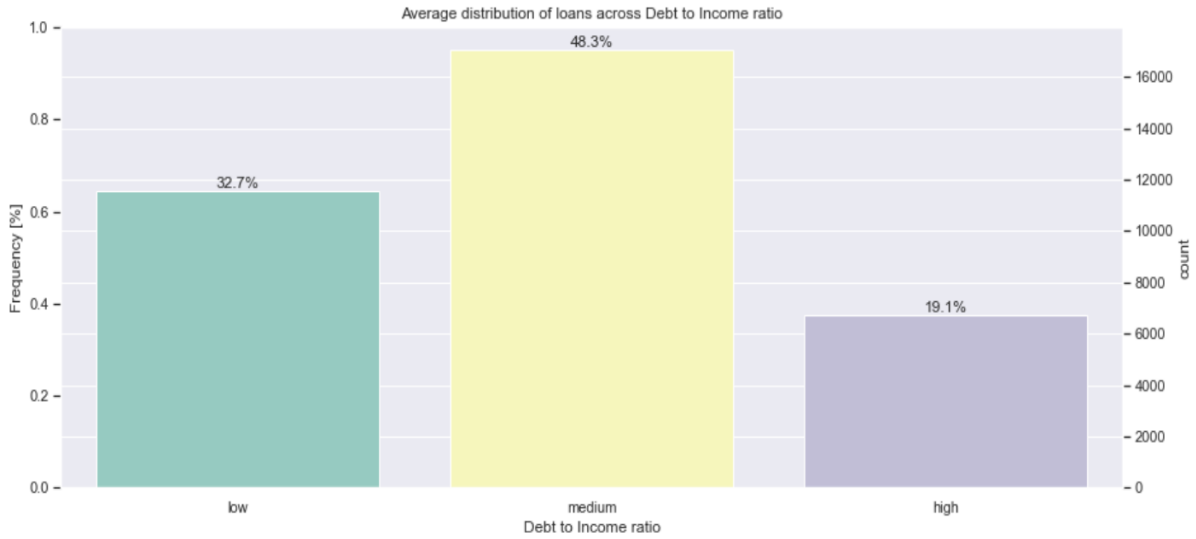
- On making analysis for different grades of people, we found that people with grades B,C are having more chances of default
- When it comes to purpose of loan, people who took loan for the purpose of '**debt_consolidation**' are having more chances of default than other purposes. It means, people who are taking loans to clear other debts are having more chances of default
- People who are having own houses are having less chances of default than people who are living in rented/mortgaged houses

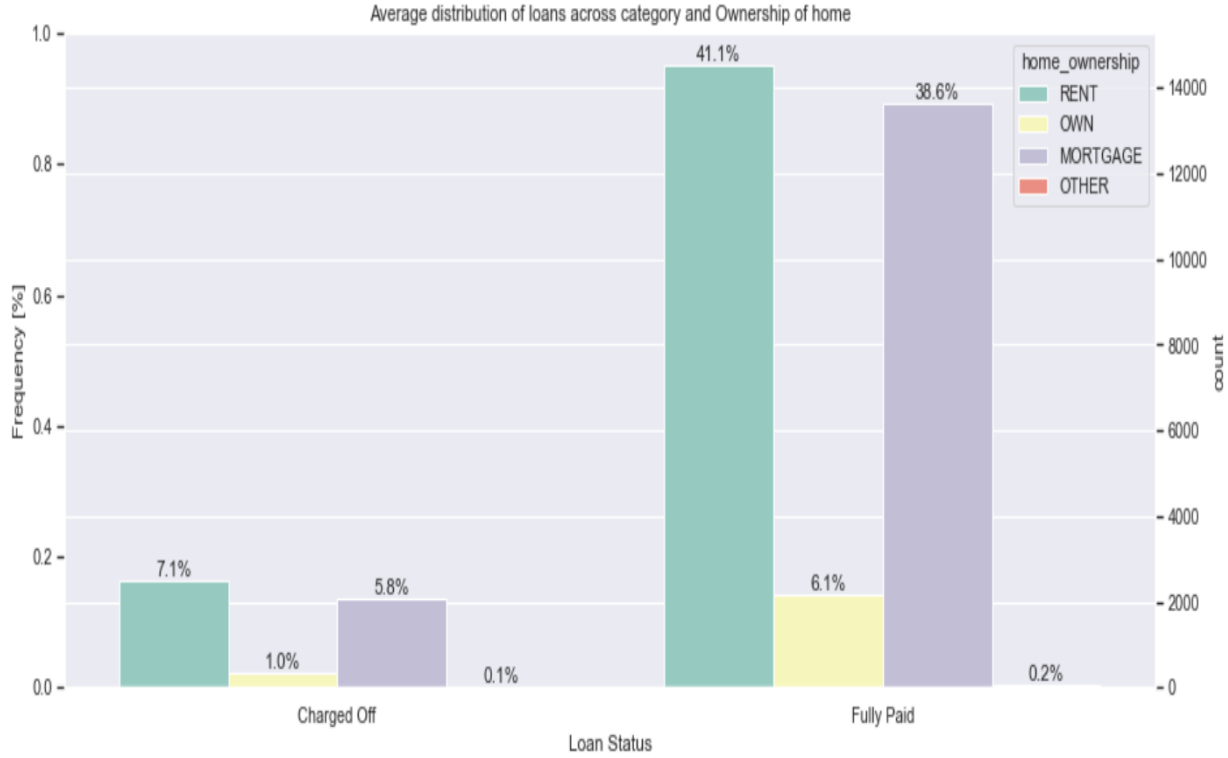
Following graphs helps in better understanding of above analysis easily

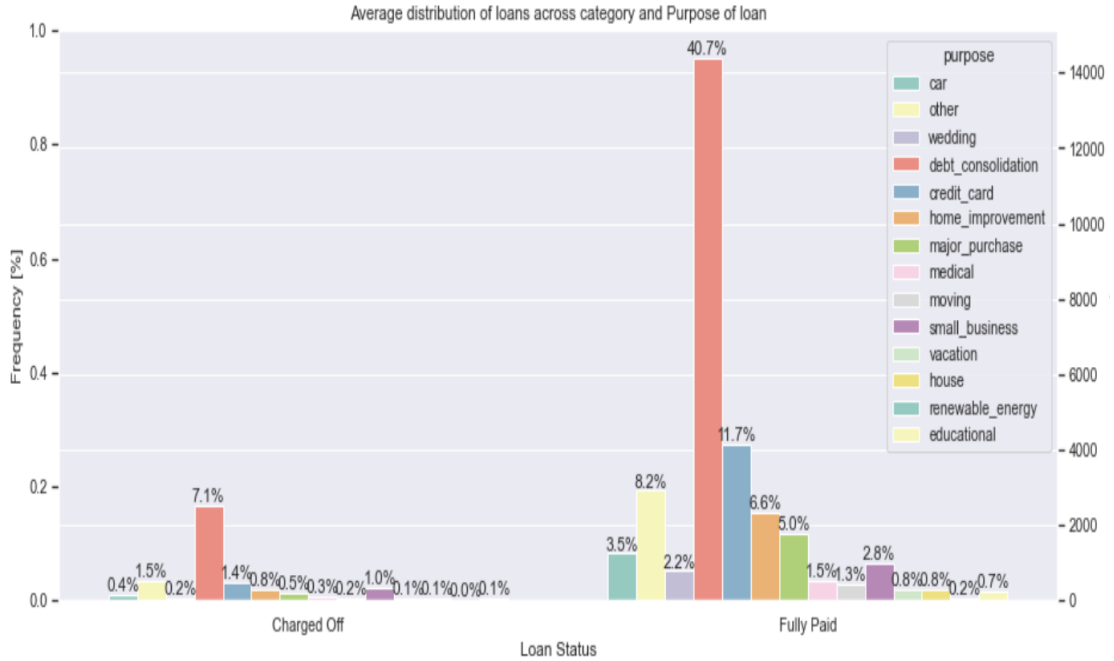


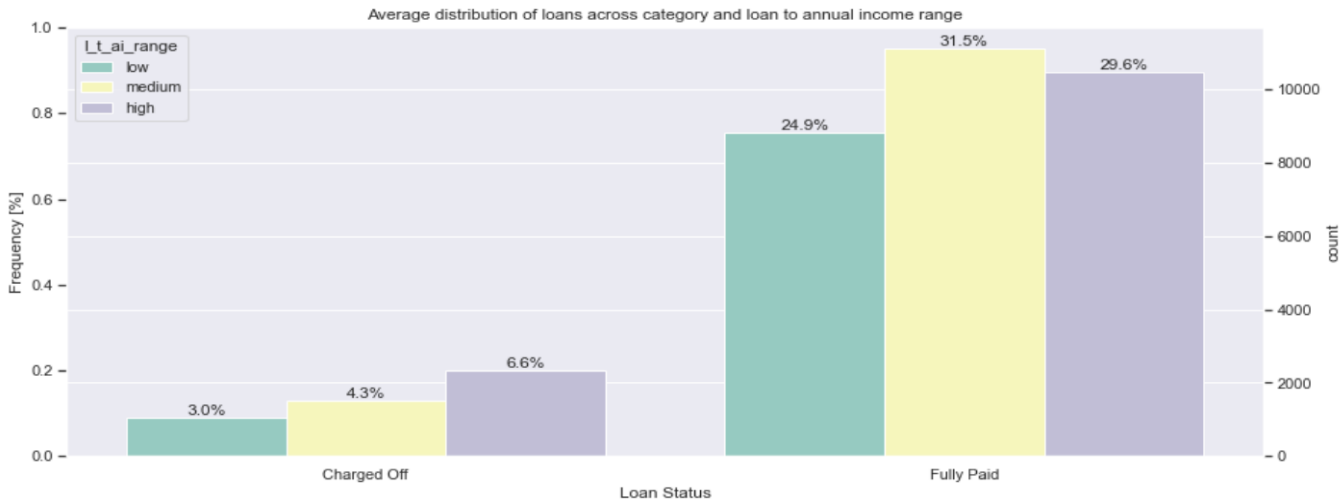

Average distribution of loans across category and Grades
Frequency [%]
count
grade
C
A
E
B
D
F
G
3.3%
1.5%
1.8%
3.5%
2.8%
0.8%
0.3%
17.0%
24.4%
5.2%
26.8%
10.4%
1.7%
0.5%
Charged Off
Fully Paid
Loan Status


Average distribution of loans across category and employement level
Frequency [%]
count
emp_length
1
10
3
8
9
5
4
6
2
7
2.7%
3.5%
1.5%
0.5%
0.4%
1.2%
1.2%
0.8%
1.5%
0.7%
17.0%
19.2%
9.4%
3.3%
2.9%
7.4%
7.9%
5.0%
10.0%
3.9%
Charged Off
Fully Paid
Loan status


Average distribution of loans across Debt to Income ratio
Frequency [%]
count
32.7%
48.3%
19.1%
low
medium
high
Debt to Income ratio


Average distribution of loans across category and Ownership of home
home_ownership
RENT
OWN
MORTGAGE
OTHER
Frequency [%]
count
7.1%
1.0%
5.8%
0.1%
41.1%
6.1%
38.6%
0.2%
Charged Off
Fully Paid
Loan Status


Average distribution of loans across category and Purpose of loan
purpose
car
other
wedding
debt_consolidation
credit_card
home_improvement
major_purchase
medical
moving
small_business
vacation
house
renewable_energy
educational
Frequency [%]
0.4%
1.5%
0.2%
7.1%
1.4%
0.8%
0.5%
0.3%
0.2%
1.0%
0.1%
0.1%
0.0%
0.1%
3.5%
8.2%
2.2%
40.7%
11.7%
6.6%
5.0%
1.5%
1.3%
2.8%
0.8%
0.8%
0.2%
0.7%
Charged Off
Fully Paid
Loan Status


Average distribution of loans across category and loan to annual income range
l_t_ai_range
low
medium
high
Frequency [%]
count
3.0%
4.3%
6.6%
24.9%
31.5%
29.6%
Charged Off
Fully Paid
Loan Status

Bi Varient Analysis:

- In bivariant analysis we made analysis how ownership of home and verification status effects the loan status and made following outcomes from it
- **Ownership of house:**

1. Rented House
   A. Debt to Income ratio : Medium DTI is having more chances of default
   B. Loan to Annual Income : High loan to Income ratio is having more chances of default
   C. Experience level : Juniors are having more chances of default
   D. Grades : B, C grade people are having more chances of default

2. Mortgaged House
   A. Debt to Income ratio : Medium DTI is having more chances of default
   B. Loan to Annual Income : High loan to Income ratio is having more chances of default
   C. Experience level : Specialist are having more chances of default
   D. Grades : B, C grade people are having more chances of default

3. Own House
   A. Debt to Income ratio : Medium DTI is having more chances of default
   B. Loan to Annual Income : High loan to Income ratio is having more chances of default
   C. Experience level : Junior are having more chances of default
   D. Grades : B, C grade people are having more chances of default

- **Verification Status**

1. Verified Status

   A. Debt to Income ratio : Medium DTI is having more chances of default

   B. Loan to Annual Income : High loan to Income ratio is having more chances of default

   C. Experience level : Juniors are having more chances of default

   D. Grades : B, C grade people are having more chances of default

   E. No.Of Years Experience: 1, 10 Years of experience are having more chances of default

2. Not Verified Status

   A. Debt to Income ratio : Medium DTI is having more chances of default

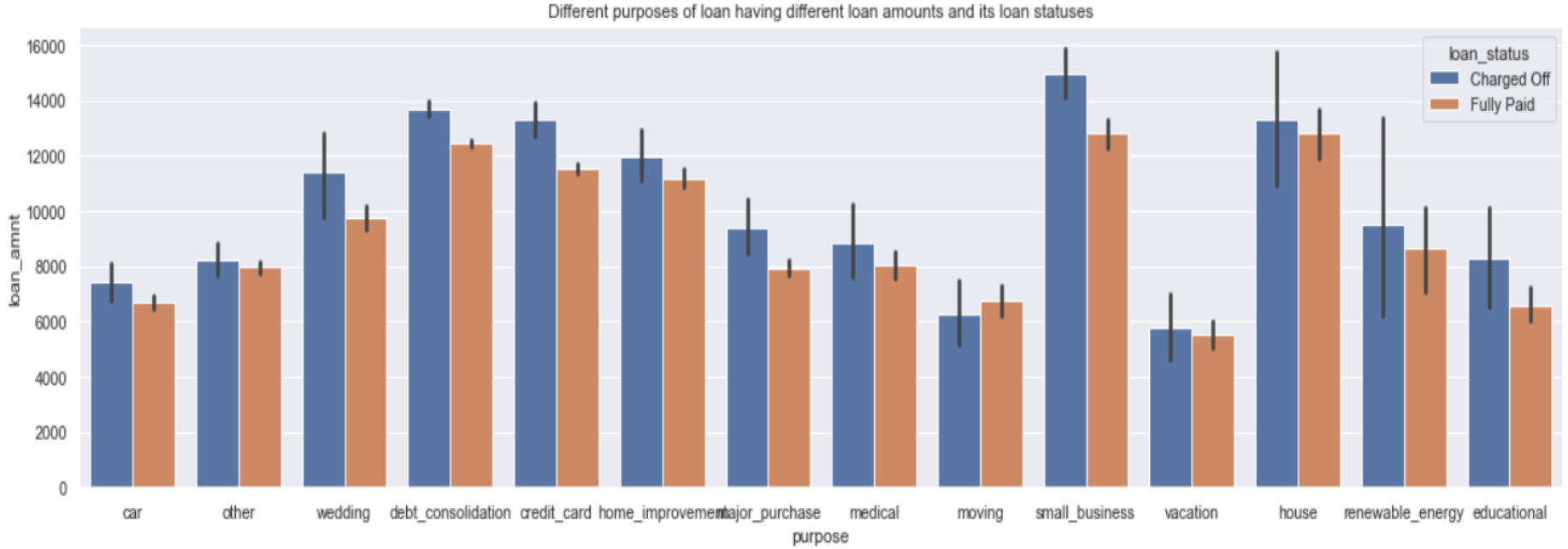
   B. Loan to Annual Income : Medium loan to Income ratio is having more chances of default

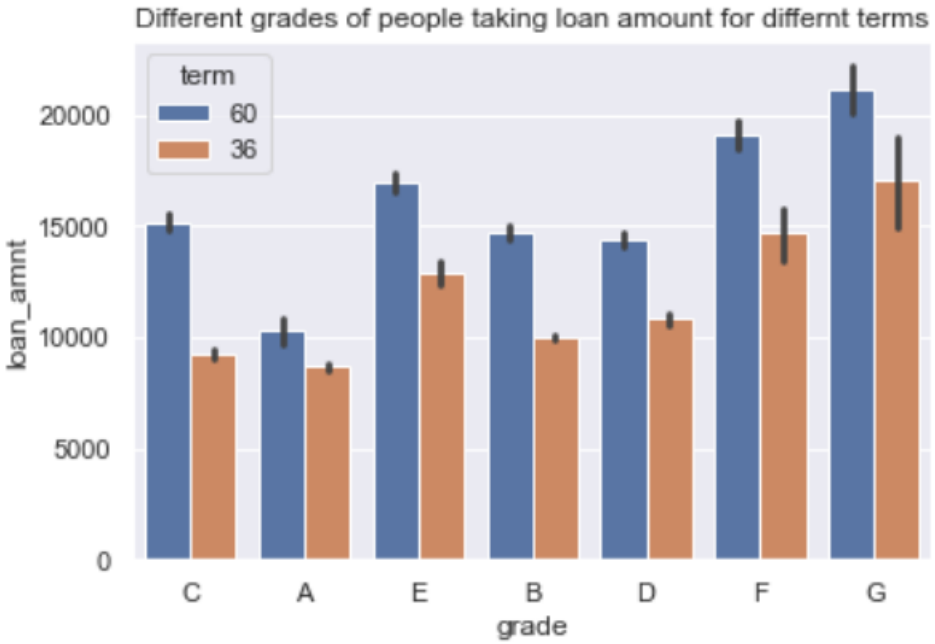
   C. Experience level : Juniors are having more chances of default

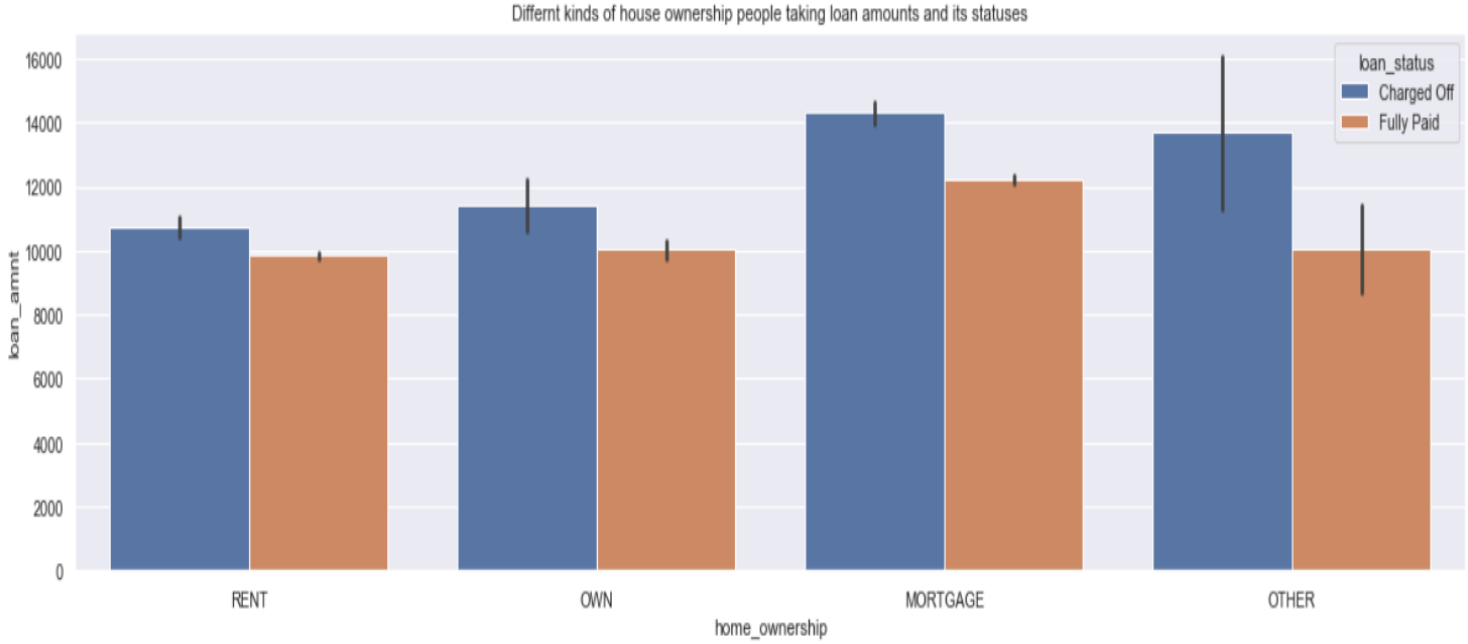
   D. Grades : B, C grade people are having more chances of default

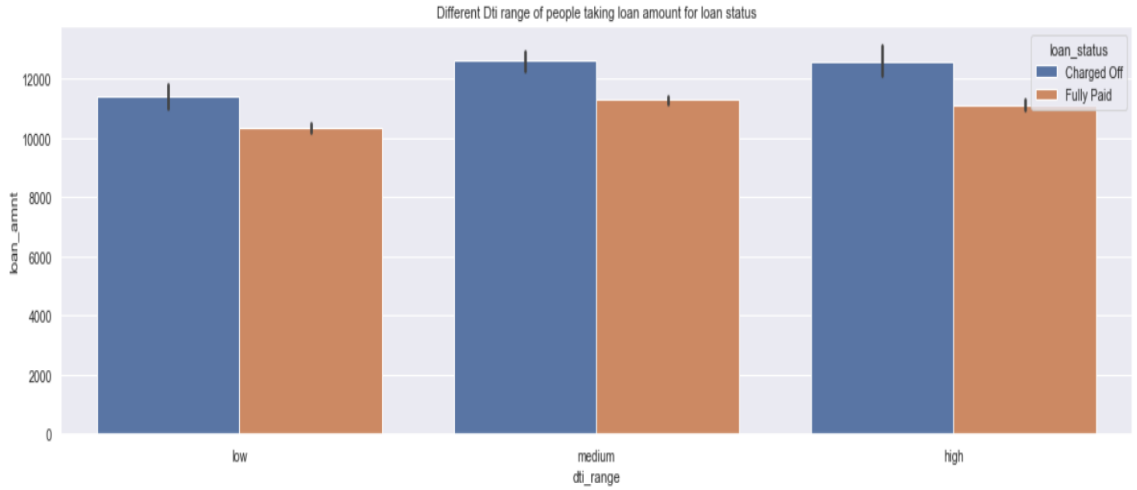
   E. No.Of Years Experience: 1, 10 Years of experience are having more chances of default

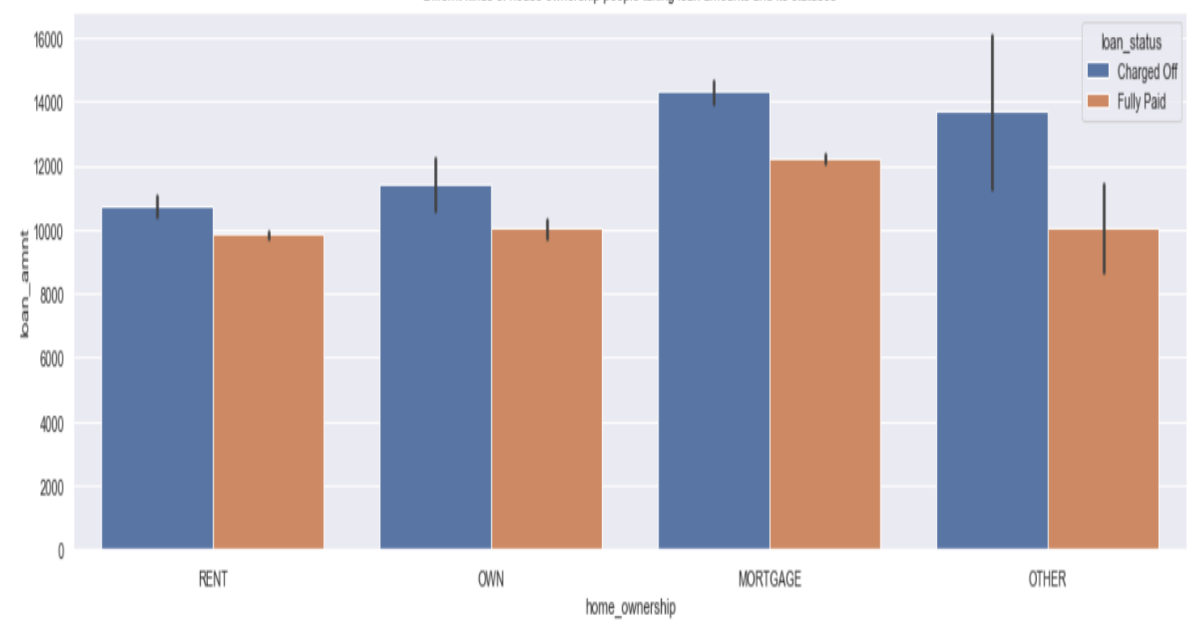
- Similarly, while making analysis between loan status, amount with different variables like as follows:
  - Verification Status: ***Loans with higher amount are verified but still they are more charged off***
  - Grades: **People with low grades takes more loan amount and are the people having more chances of defaults and people with low grades take higher amount of loans for longer duration**
  - Term: **People with high term are having more chances of default**
  - DTI Range: **People with medium & high debt to income ratio are having more chances of default. Since, they have more debts than income and fails ton pay loan**
  - LTI Range: **People with high income to loan ratio are having high chances of default by taking more amount of loan. Since there is less compared to their loan taken**
  - Employee Exp: **People with more experience level around 8-10 are having high chances of default by taking more amount of loan**
  - Purpose of loan: **People who took loan for the purpose of 'small_business', 'debt_conslidation', 'credit_card' are most likely to have chance of default by taking more amount of loan**
  - Ownership of home: **People with house of other and mortage took more amount of loan and likely to default more**
- Lets see the above insights in graphs for better understanding

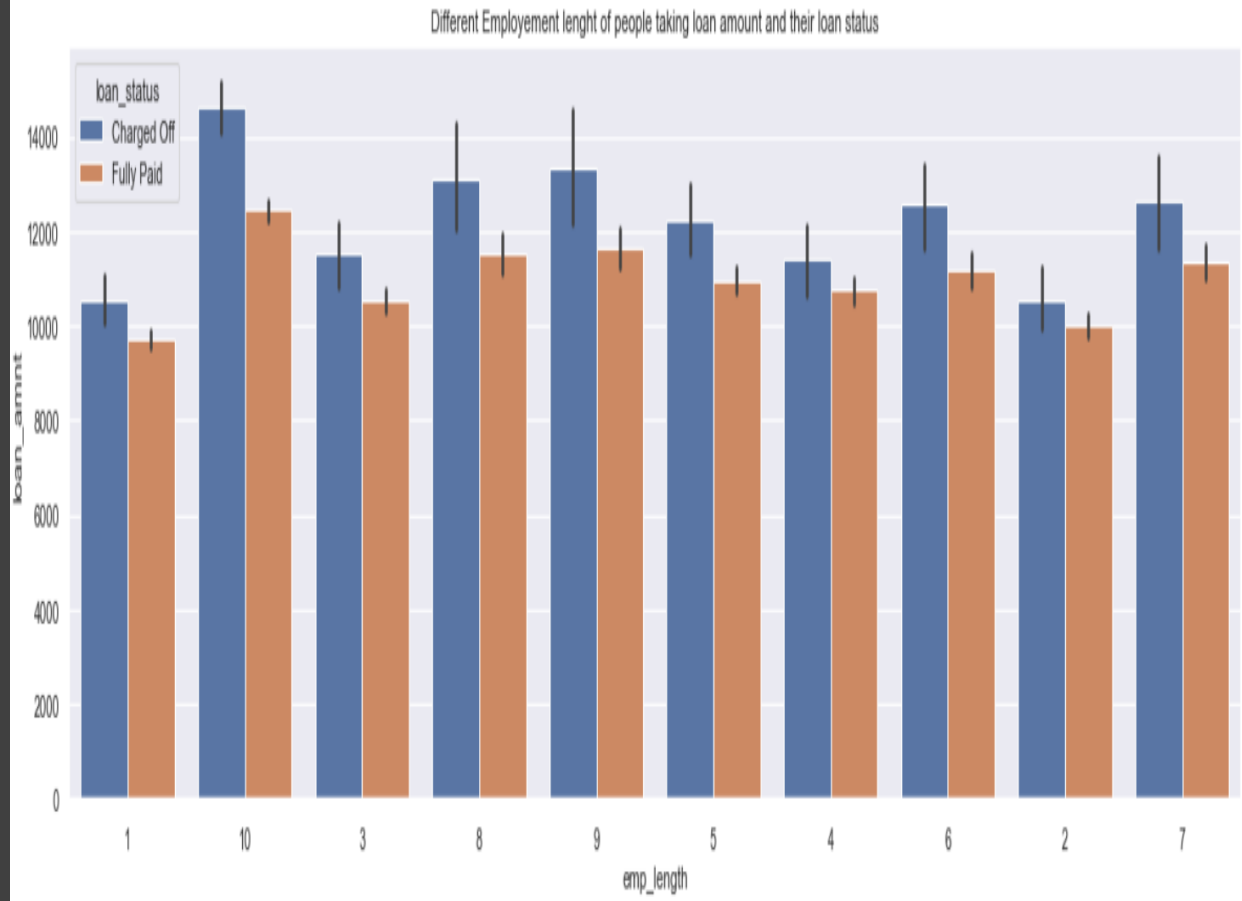

Different purposes of loan having different loan amounts and its loan statuses
loan_status
Charged Off
Fully Paid
loan_amnt
0
2000
4000
6000
8000
10000
12000
14000
16000
car
other
wedding
debt_consolidation
credit_card
home_improvement
major_purchase
medical
moving
small_business
vacation
house
renewable_energy
educational
purpose


Different grades of people taking loan amount for differnt terms
term
60
36
loan_amnt
0
5000
10000
15000
20000
C
A
E
B
D
F
G
grade

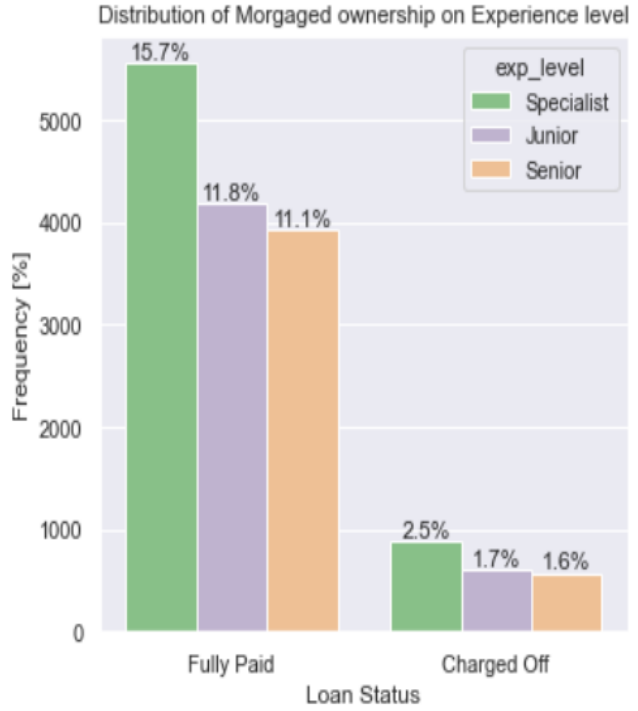

Differnt kinds of house ownership people taking loan amounts and its statuses
loan_status
Charged Off
Fully Paid
loan_amnt
0
2000
4000
6000
8000
10000
12000
14000
16000
RENT
OWN
MORTGAGE
OTHER
home_ownership

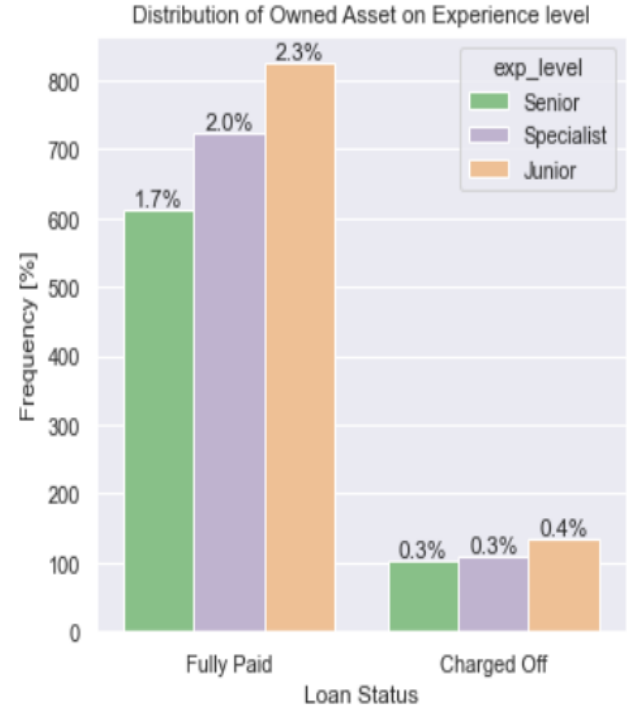

Different Dti range of people taking loan amount for loan status
loan_status
Charged Off
Fully Paid
loan_amnt
0
2000
4000
6000
8000
10000
12000
low
medium
high
dti_range

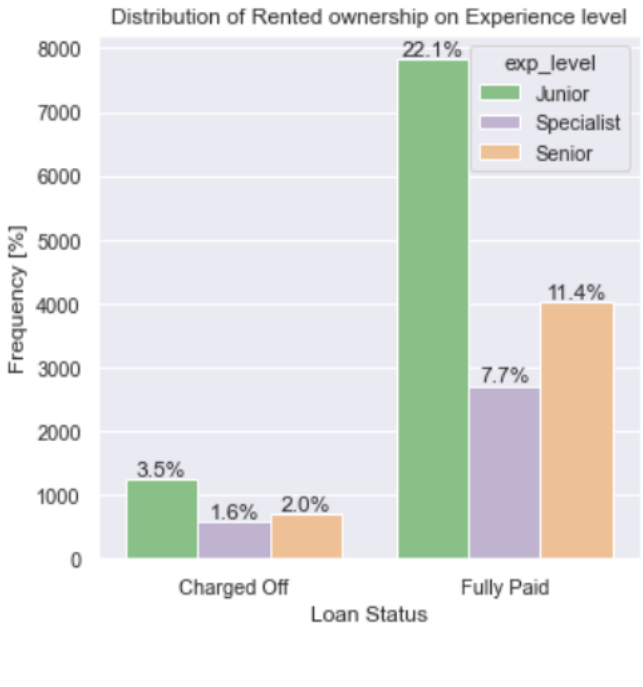

Differnt kinds of house ownership people taking loan amounts and its statuses
loan_status
Charged Off
Fully Paid
loan_amnt
0
2000
4000
6000
8000
10000
12000
14000
16000
RENT
OWN
MORTGAGE
OTHER
home_ownership

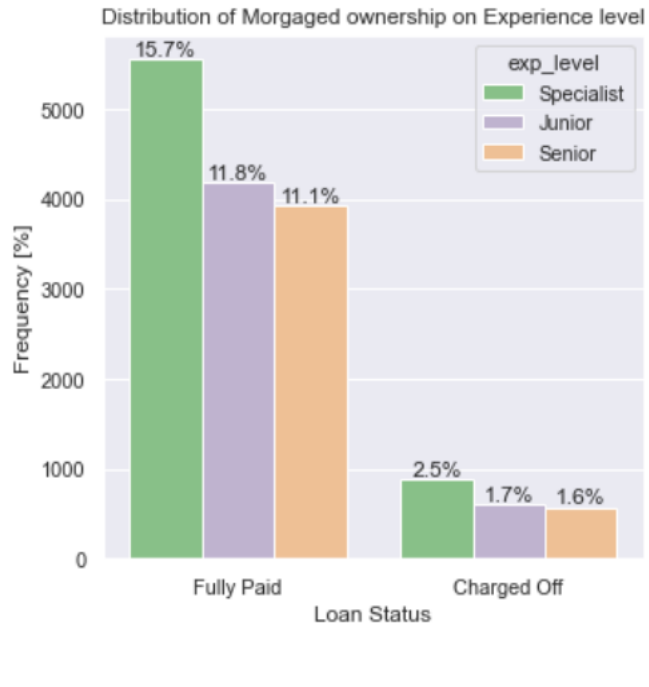

Different Employement lenght of people taking loan amount and their loan status
loan_status
Charged Off
Fully Paid
loan_amnt
0
2000
4000
6000
8000
10000
12000
14000
1
10
3
8
9
5
4
6
2
7
emp_length


Distribution of Rented ownership on Experience level
exp_level
Junior
Specialist
Senior
Frequency [%]
Loan Status
Charged Off
Fully Paid
3.5%
1.6%
2.0%
22.1%
7.7%
11.4%

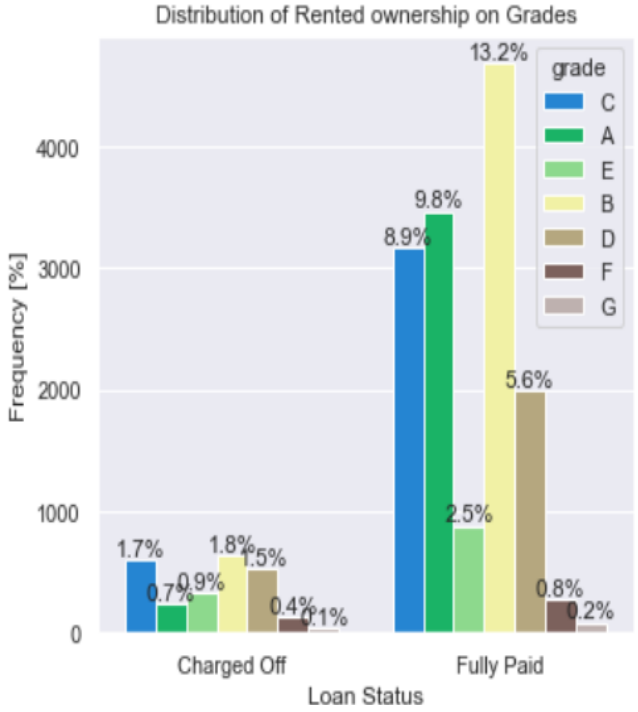

Distribution of Morgaged ownership on Experience level
exp_level
Specialist
Junior
Senior
Frequency [%]
Loan Status
Fully Paid
Charged Off
15.7%
11.8%
11.1%
2.5%
1.7%
1.6%

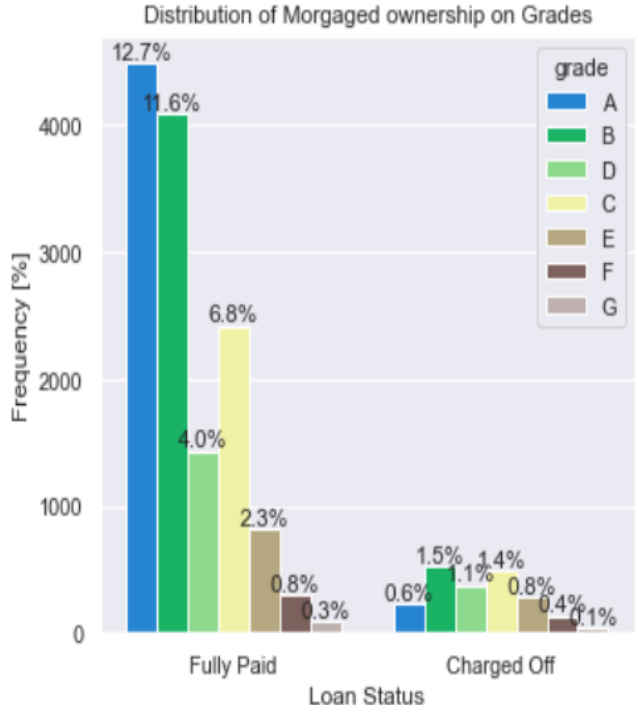

Distribution of Owned Asset on Experience level
exp_level
Senior
Specialist
Junior
Frequency [%]
Loan Status
Fully Paid
Charged Off
1.7%
2.0%
2.3%
0.3%
0.3%
0.4%

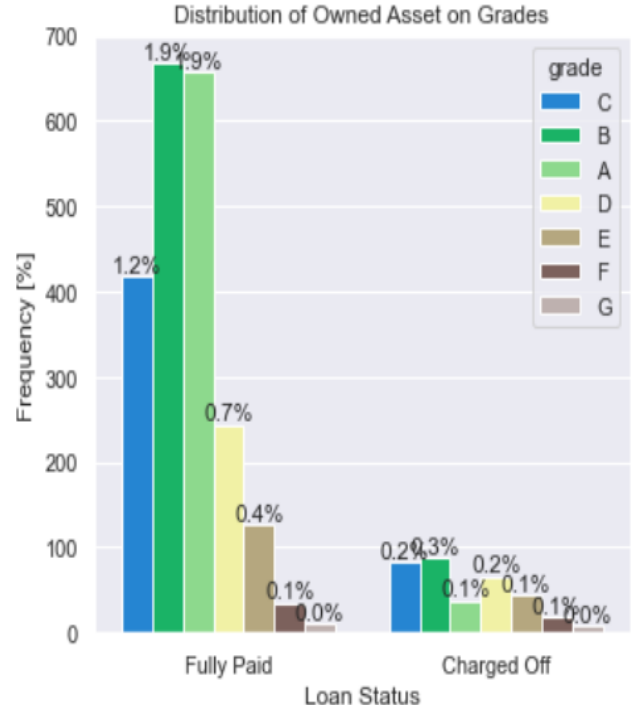

Distribution of Rented ownership on Experience level
exp_level
Junior
Specialist
Senior
Frequency [%]
Loan Status
Charged Off
Fully Paid
3.5%
1.6%
2.0%
22.1%
7.7%
11.4%

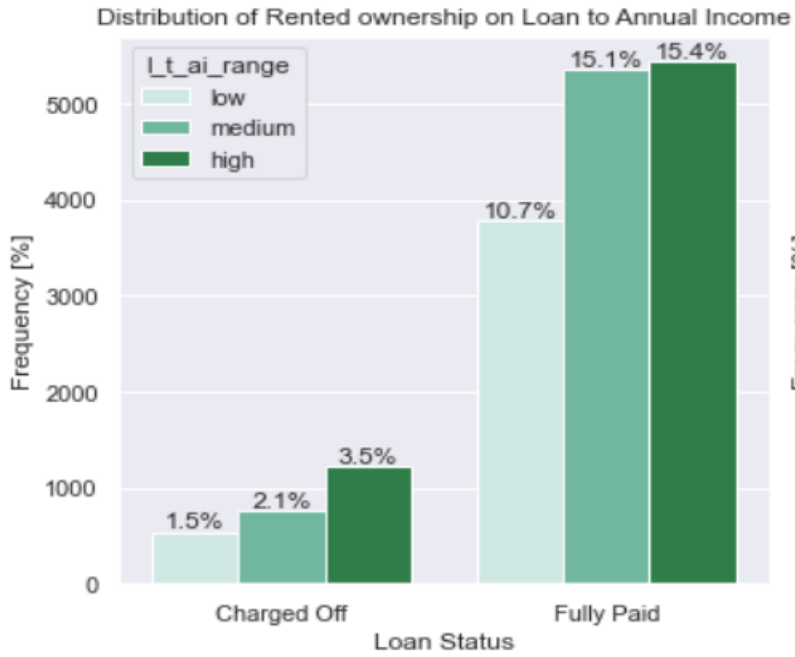

Distribution of Morgaged ownership on Experience level
exp_level
Specialist
Junior
Senior
Frequency [%]
Loan Status
Fully Paid
Charged Off
15.7%
11.8%
11.1%
2.5%
1.7%
1.6%

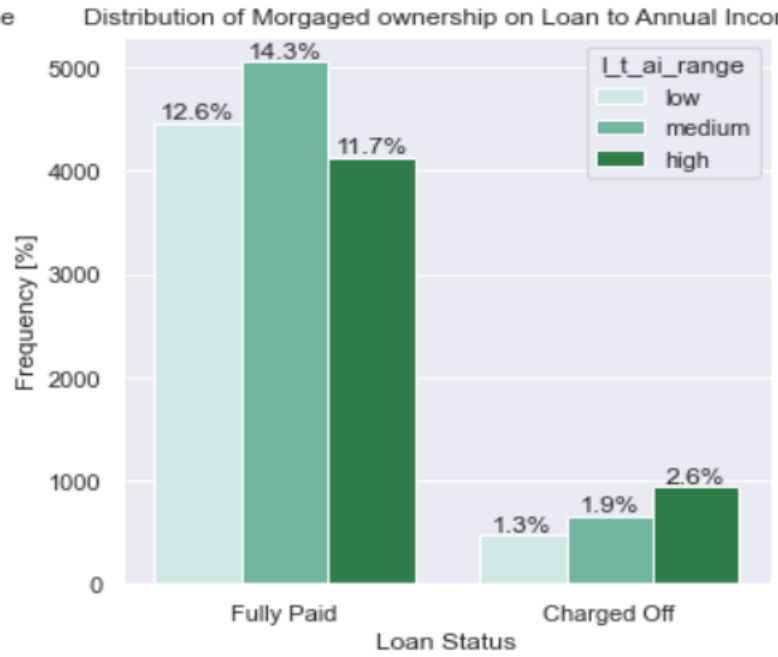

Distribution of Owned Asset on Experience level
exp_level
Senior
Specialist
Junior
Frequency [%]
Loan Status
Fully Paid
Charged Off
1.7%
2.0%
2.3%
0.3%
0.3%
0.4%

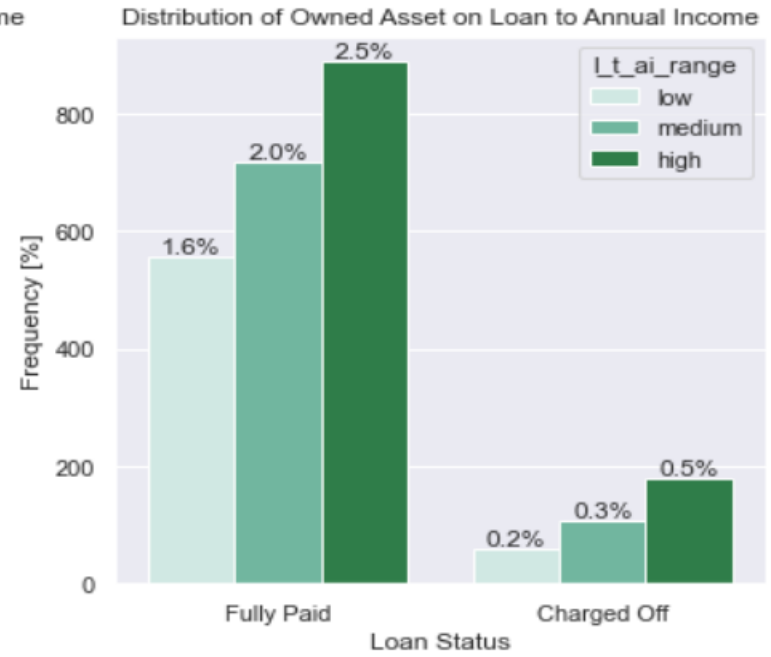

Distribution of Rented ownership on Grades
grade
C
A
E
B
D
F
G
Frequency [%]
Loan Status
Charged Off
Fully Paid
1.7%
0.7%
0.9%
1.8%
1.5%
0.4%
0.1%
8.9%
9.8%
2.5%
13.2%
5.6%
0.8%
0.2%

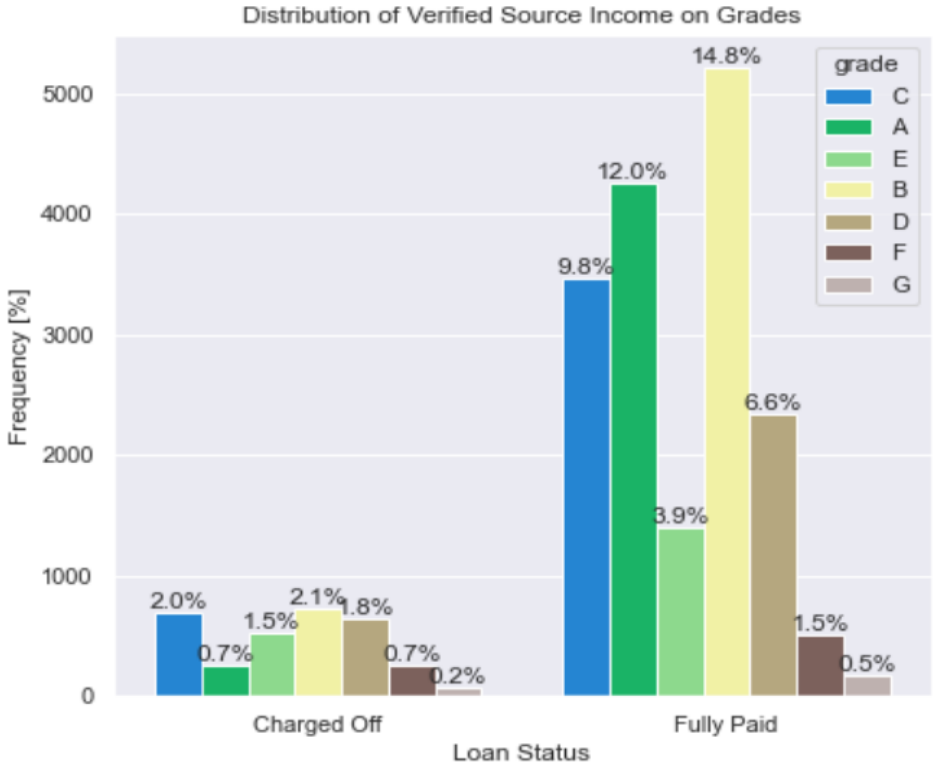

Distribution of Morgaged ownership on Grades
grade
A
B
D
C
E
F
G
Frequency [%]
Loan Status
Fully Paid
Charged Off
12.7%
11.6%
4.0%
6.8%
2.3%
0.8%
0.3%
0.6%
1.5%
1.1%
1.4%
0.8%
0.4%
0.1%

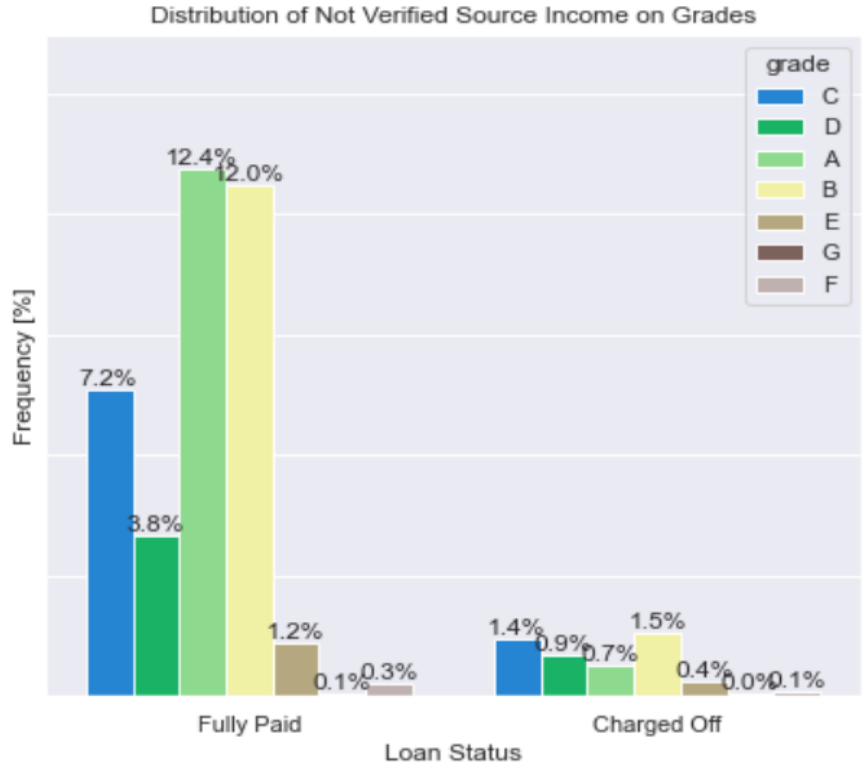

Distribution of Owned Asset on Grades
grade
C
B
A
D
E
F
G
Frequency [%]
Loan Status
Fully Paid
Charged Off
1.2%
1.9%
1.9%
0.7%
0.4%
0.1%
0.0%
0.2%
0.3%
0.1%
0.2%
0.1%
0.1%
0.0%

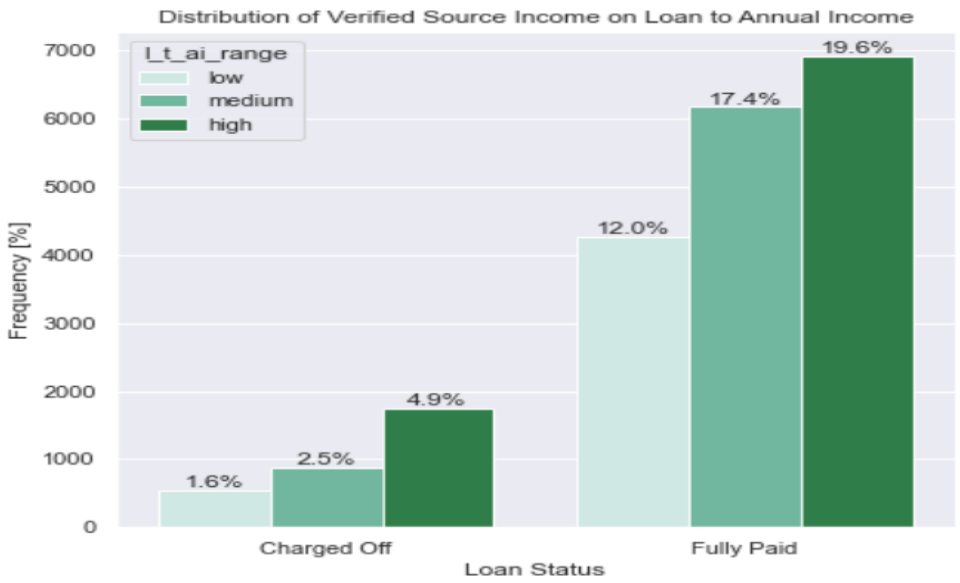

Distribution of Rented ownership on Loan to Annual Income
l_t_ai_range
low
medium
high
Frequency [%]
Loan Status
Charged Off
Fully Paid
1.5%
2.1%
3.5%
10.7%
15.1%
15.4%

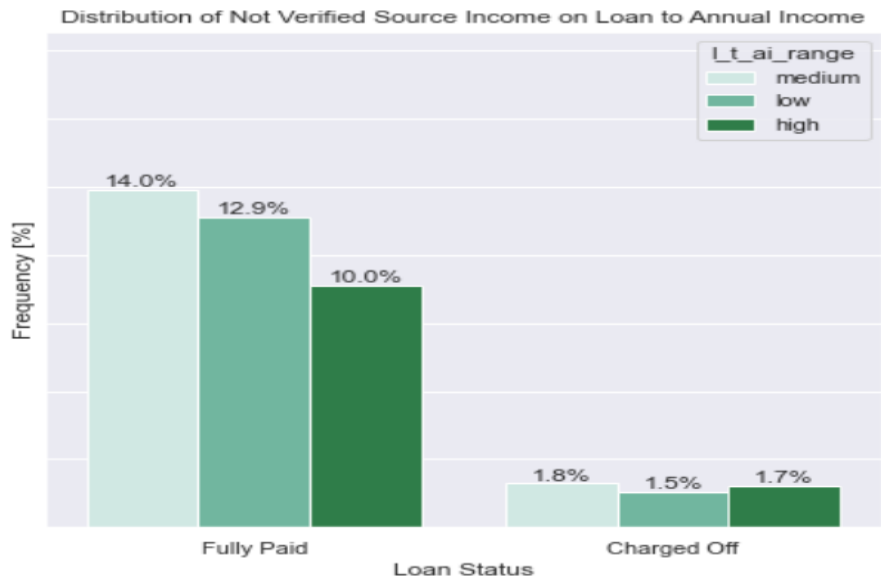

Distribution of Morgaged ownership on Loan to Annual Income
l_t_ai_range
low
medium
high
Frequency [%]
Loan Status
Fully Paid
Charged Off
12.6%
14.3%
11.7%
1.3%
1.9%
2.6%

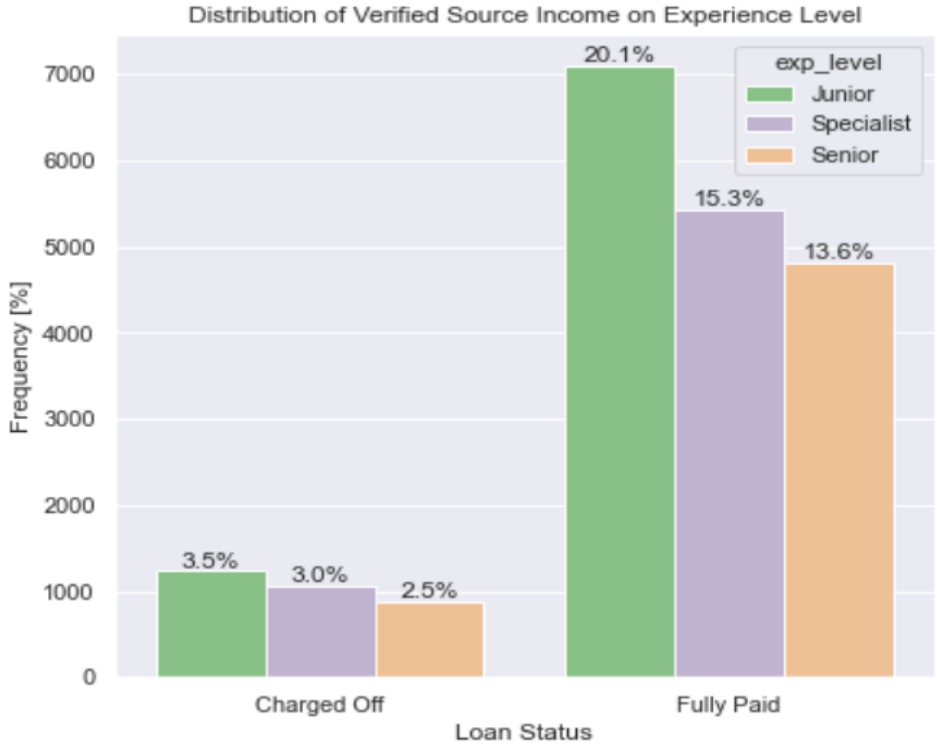

Distribution of Owned Asset on Loan to Annual Income
l_t_ai_range
low
medium
high
Frequency [%]
Loan Status
Fully Paid
Charged Off
1.6%
2.0%
2.5%
0.2%
0.3%
0.5%

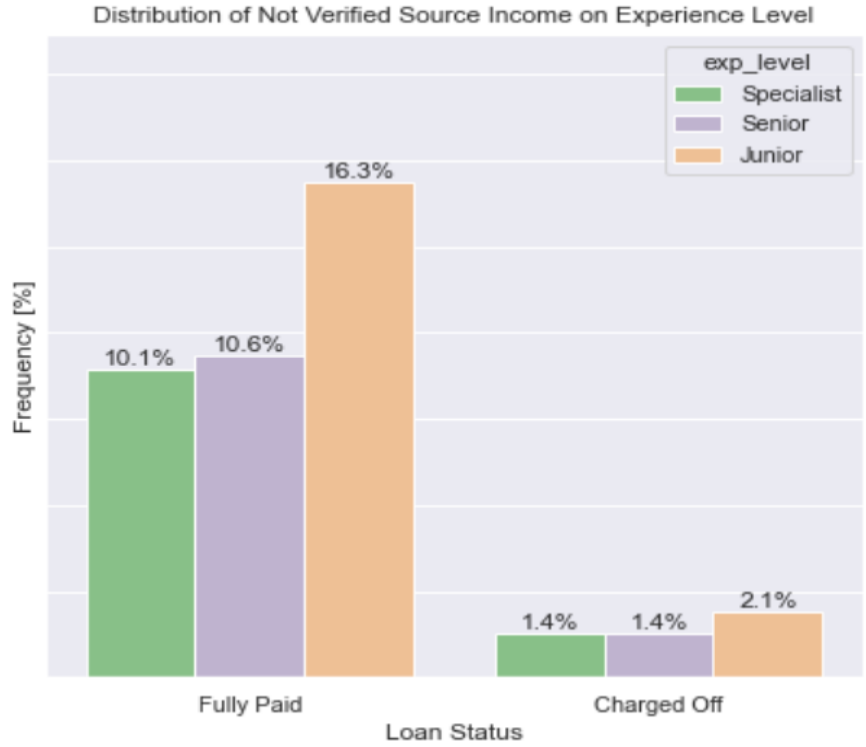

Distribution of Verified Source Income on Grades
Frequency [%]
Loan Status
Charged Off
Fully Paid
grade
C
A
E
B
D
F
G
2.0%
0.7%
1.5%
2.1%
1.8%
0.7%
0.2%
9.8%
12.0%
3.9%
14.8%
6.6%
1.5%
0.5%

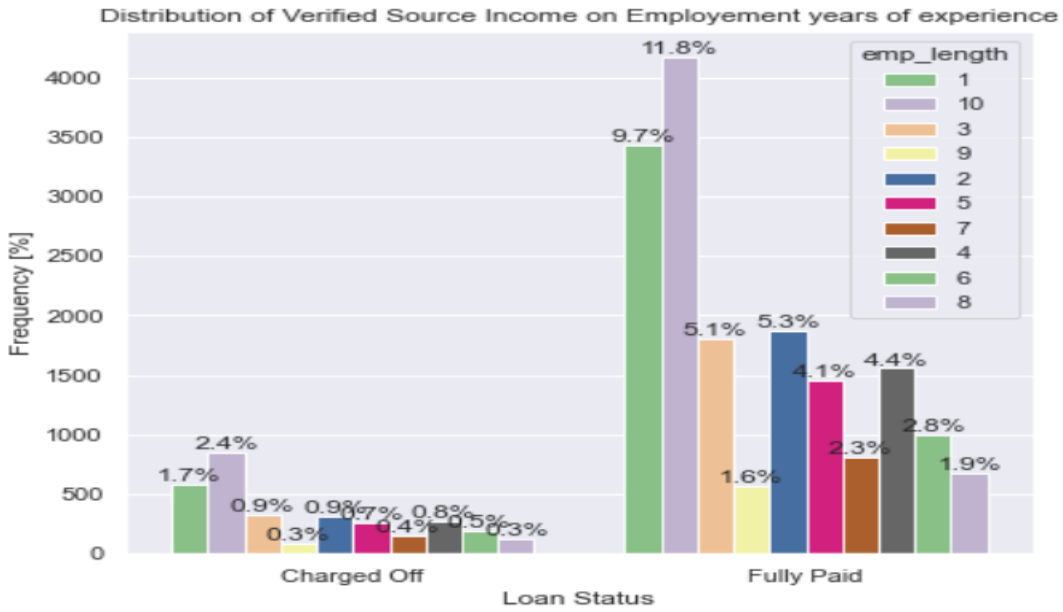

Distribution of Not Verified Source Income on Grades
Frequency [%]
Loan Status
Fully Paid
Charged Off
grade
C
D
A
B
E
G
F
7.2%
3.8%
12.4%
12.0%
1.2%
0.1%
0.3%
1.4%
0.9%
0.7%
1.5%
0.4%
0.0%
0.1%

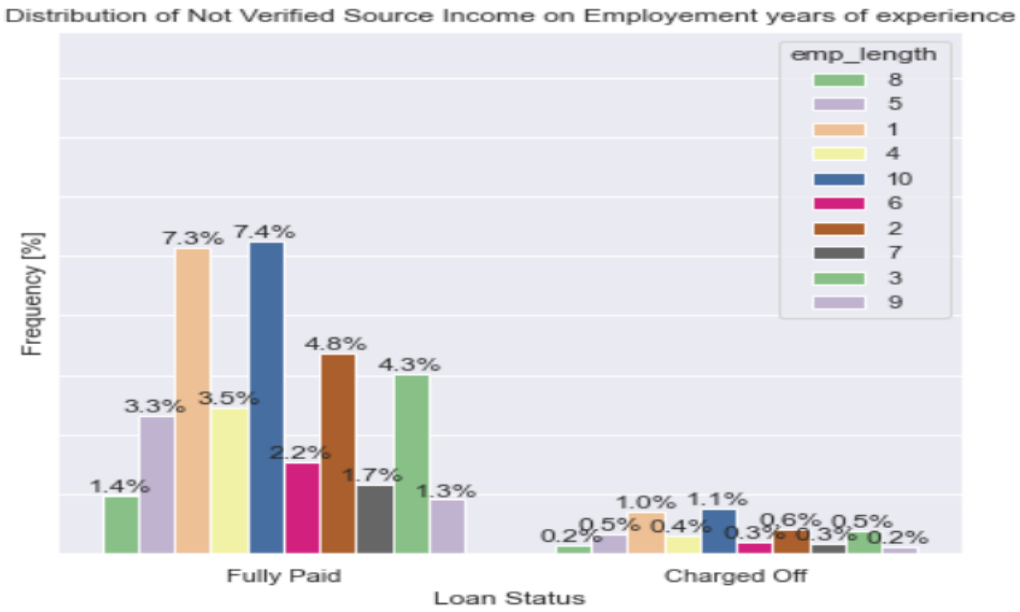

Distribution of Verified Source Income on Loan to Annual Income
Frequency [%]
Loan Status
Charged Off
Fully Paid
l_t_ai_range
low
medium
high
1.6%
2.5%
4.9%
12.0%
17.4%
19.6%


Distribution of Not Verified Source Income on Loan to Annual Income
Frequency [%]
Loan Status
Fully Paid
Charged Off
l_t_ai_range
medium
low
high
14.0%
12.9%
10.0%
1.8%
1.5%
1.7%


Distribution of Verified Source Income on Experience Level
exp_level
Junior
Specialist
Senior
Frequency [%]
Loan Status
Charged Off
Fully Paid
3.5%
3.0%
2.5%
20.1%
15.3%
13.6%


Distribution of Not Verified Source Income on Experience Level
exp_level
Specialist
Senior
Junior
Frequency [%]
Loan Status
Fully Paid
Charged Off
10.1%
10.6%
16.3%
1.4%
1.4%
2.1%


Distribution of Verified Source Income on Employement years of experience
emp_length
1
10
3
9
2
5
7
4
6
8
Frequency [%]
Loan Status
Charged Off
Fully Paid


Distribution of Not Verified Source Income on Employement years of experience
emp_length
8
5
1
4
10
6
2
7
3
9
Frequency [%]
Loan Status
Fully Paid
Charged Off

Correlation:

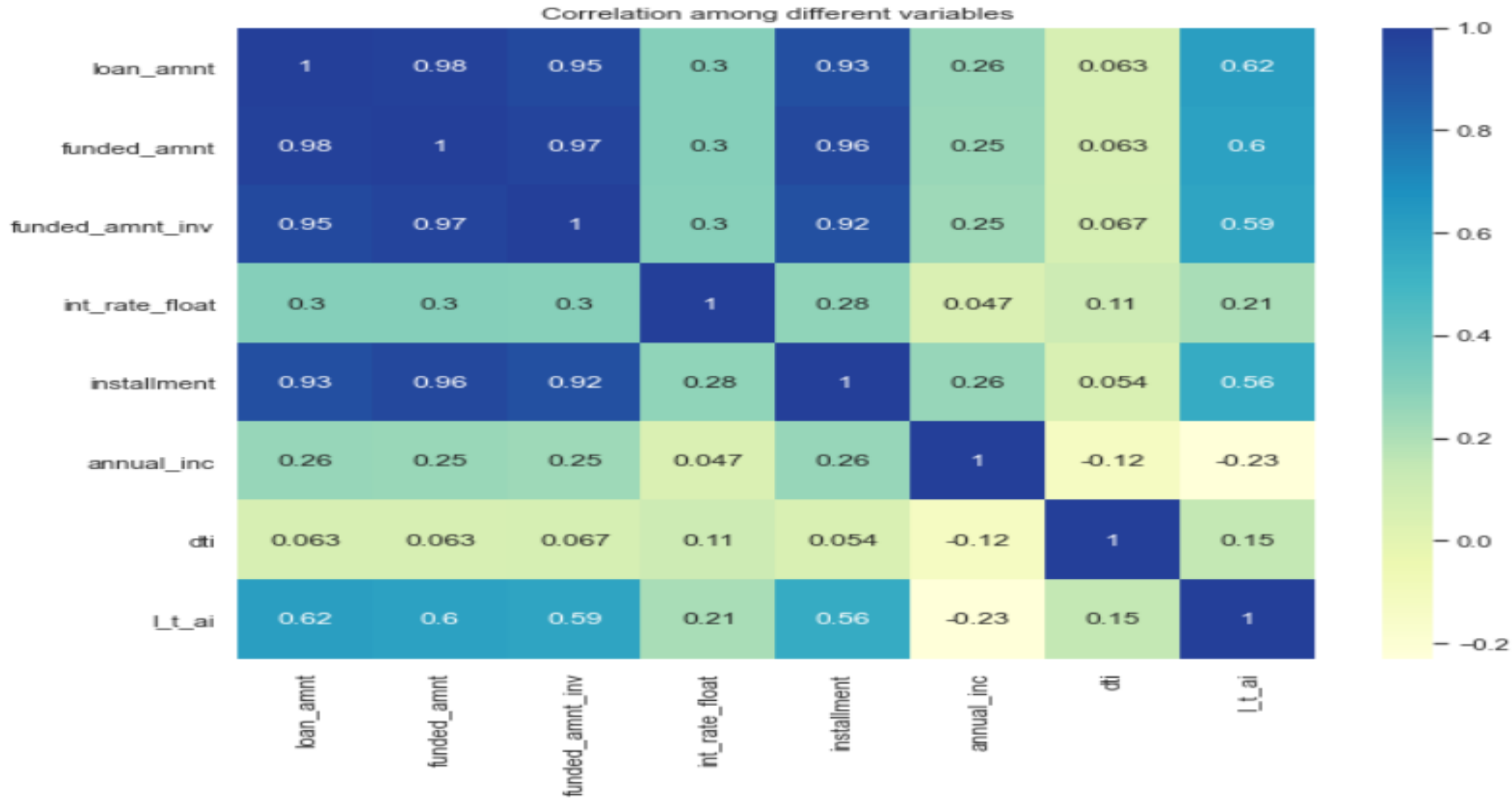
- While making analysis using correlation, the following are the outcomes observed
  - High corelated:

    1. loan amount is highly corelated with funded_amt, funded_aqmt_inv, installment, income to amount respectively from high to low
    2. funded_amt_inv is higly corelated with funded_amt, loan_amt, installement, income to loan ratio respectively from high to low
    3. installment is corelated with funded_amt, loan_amt, funded_amt_inv, income to loan ratio respectively from high to low
    4. income to loan ratio is high corelated with loan_amt, funded_amt_inv, installement respectively from high to low

  - Low Corelated:

    1. Annual income is negative or less corelated with loan to income ratio and debt to income

- Lets see the graph for better understanding of above analysis


Correlation among different variables
loan_amnt
funded_amnt
funded_amnt_inv
int_rate_float
installment
annual_inc
dti
l_t_ai
1 0.98 0.95 0.3 0.93 0.26 0.063 0.62
0.98 1 0.97 0.3 0.96 0.25 0.063 0.6
0.95 0.97 1 0.3 0.92 0.25 0.067 0.59
0.3 0.3 0.3 1 0.28 0.047 0.11 0.21
0.93 0.96 0.92 0.28 1 0.26 0.054 0.56
0.26 0.25 0.25 0.047 0.26 1 -0.12 -0.23
0.063 0.063 0.067 0.11 0.054 -0.12 1 0.15
0.62 0.6 0.59 0.21 0.56 -0.23 0.15 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2

Conclusion:

Lending needs to take care of following characters while considering loan:

- Debt, Loan to Income ratio
- Grade
- Purpose of loan
- Income

Following are conclusions that should be considered as per the above analysis while providing loan to the person:

- People with low grade are taking more amount of loans for long tenure are having more chances of default. So, better to give loan who are having the A grade.
- People who are taking loans with the purpose of debt_considellation, credit_card, others are having more chances of default
- People who are having own houses are having less chances of default
- People having less income and more debts are having high chances of default
- People with less delinques are having less chances of default