LENDING CLUB CASE STUDY

Submitted by

Sreekumar N.P.
sreekumarnp4u@gmail.com

Group Facilitator :
Naseemuddin Mohammed

Objective

Lending Club 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.

Two **types of risks** are associated with the bank's decision while approving the loans :

- 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

If the company is able to identify these risky loan applicants then such loans can be reduced In order to cut down the amount of credit loss. Identification of such driving factors using EDA is the objective of this case study.

- ✓ From this case study , the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Problem solving methodology

The exploratory data analysis (EDA) methodology used in this case study is grouped in to four steps :

1. Data Understanding

2. Data Cleaning

3. Data Analysis

4. Recommendations

Step 1 : Data understanding

For this case study, an excel sheet which contains loan data, Applicants demographic data and behavioral data is provided. The data dictionary to understand the definition of the variables is also provided additionally.

The data understanding is performed by following the steps as follows :

1. Understanding the Shape of the data (Rows, Columns)
   There are 39717 rows and 111 columns before data cleaning.

2. Check if the data in the data frame is consistent such that there is no misplaced data between columns.

3. Check the data types of the data to see if the quantitative values are int/float and categorical is object / String

4. Understanding the Outliers

5. Understand the presence of Null/ NaN values in columns and in rows in general

6. Understand the variables against data dictionary

7. Understand the variables from the column names

8. Understand the variables importance to gain domain knowledge by browsing though the internet

Step 2 : Data Cleaning

1. Remove the Columns which contain 90% or more Null values
2. Check individual variables including Target variable and perform data cleaning

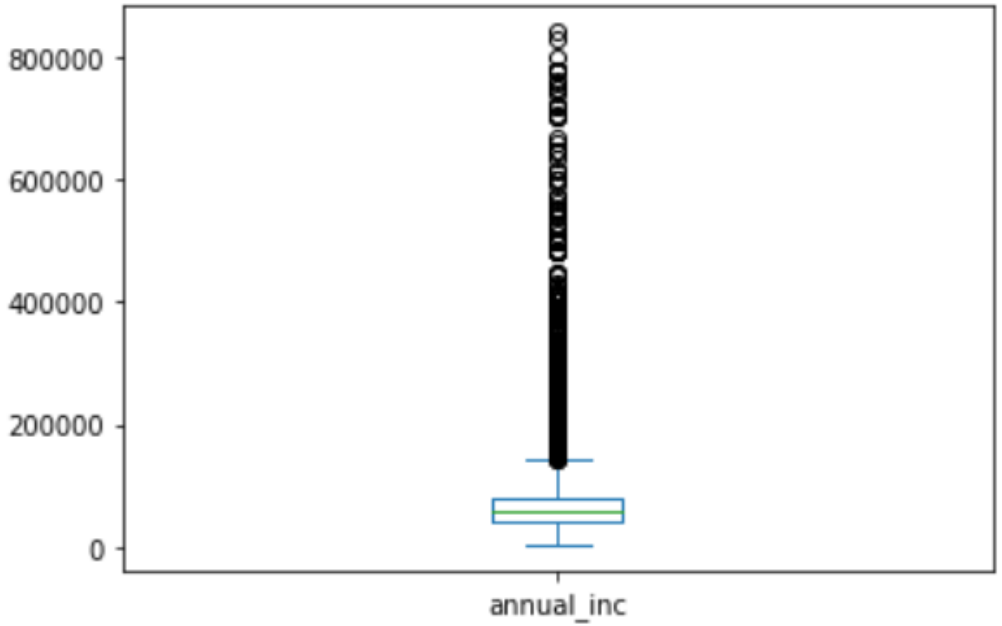
- ❑ Remove outliers
  - ➢ The charged Off maximum annual income is 840000, hence the outliers of annual income above it can be removed
- ❑ Correct their data type
  - ➢ The % suffix has been removed to convert the Percentage from Object data type to float
- ❑ Impute the data where Null/NaN is present or to remove them.
  - ➢ Employment length (emp_length) consisted of null values. As it is not meaningful to impute them, those rows has been deleted.

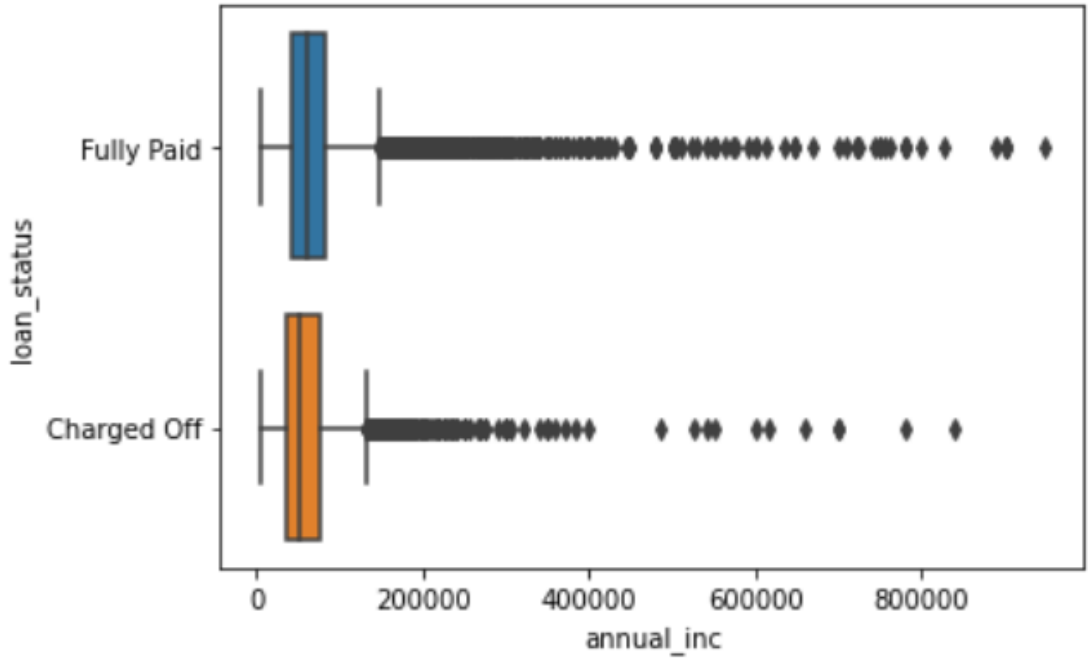
3. Those columns which doesn't have any business relevance can be removed or ignored from the analysis.

4. New columns are derived from the existing as a part of segmented univariate analysis which is then also used in bivariate analysis (eg: grouping based on (L/M/H) for Annual income, DTI, Interest rate)

e.g. Removing outliers

After removing extreme outliers from the income group, there are still more values above the upper whiskers. As those contains representative higher income group useful in the analysis, I have decided not to remove them. Then I analyzed the Loan status against Annual income and removed income range above 840000 as there are no Charged Off cases above that range.





Step 3 : Data Analysis

There are broadly three types of variables :

1. Demographic variables (Those related to the applicants information)
2. Loan Characteristic information (Amount , Interest rate purpose etc.)
3. Customer behavioral variables (Generated after the loan is granted)

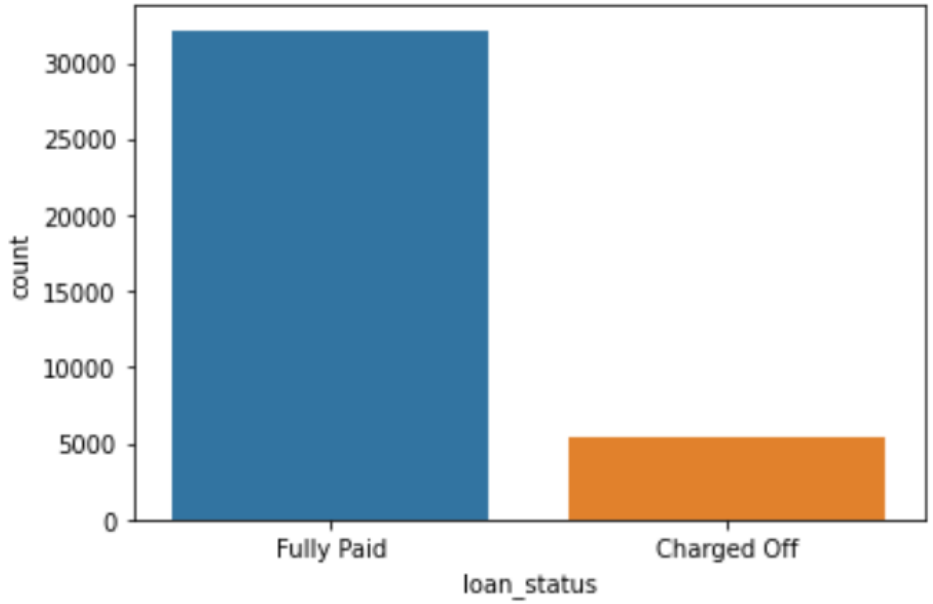
The customer behavioral variables can be ignored in this analysis as they are generally not available during loan application.

The Loan status (loan_status) is the **‘Target variable’** for the analysis. The sub category ‘Current’ can be ignored from the analysis as they are neither fully paid nor defaulted.

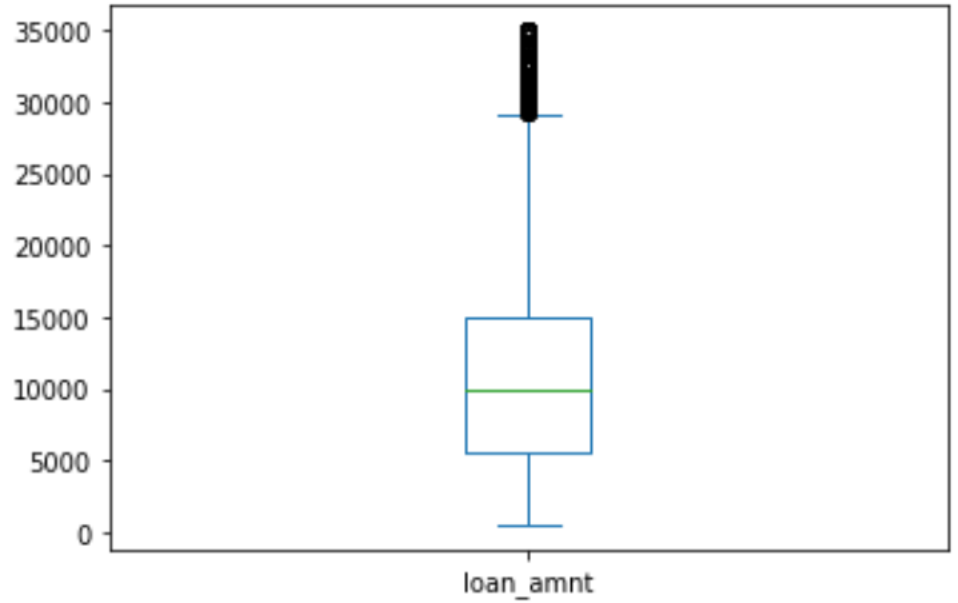
The important demographic and loan variables used in this analysis are :

1. Annual Income (annual_inc)
2. Loan Amount (loan_anmt)
3. Term (term)
4. Grade (grade)
5. Sub grade (sub_grade)
6. Loan Amount (loan_amnt)
7. Purpose of loan (purpose)
8. DTI (dti)
9. Employment years (emp_length)
10. Home Ownership (home_ownership)
11. Verification Status (verification_status)

Univariate Analysis

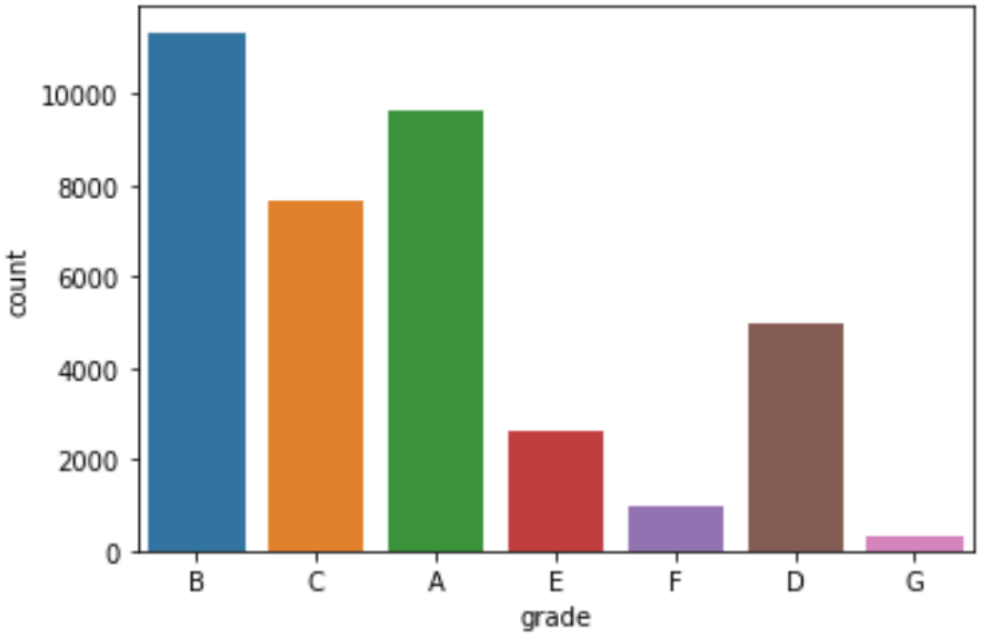


The number of fully paid loans is much higher than charged Off

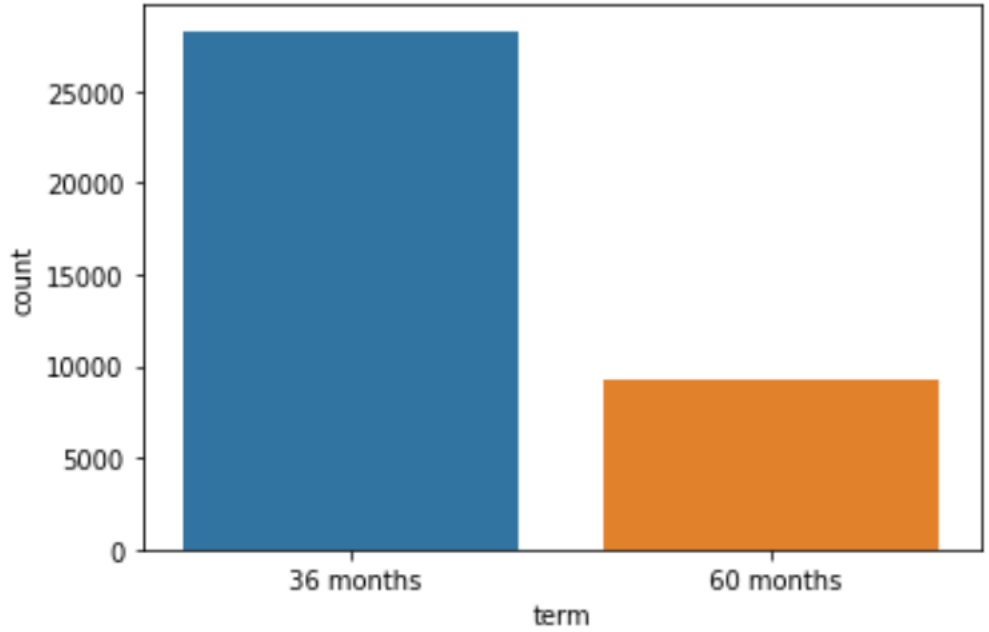


Loan amount has wide range of values max being 35000 , min being 500 and 75th percentile lies at 15000.

Univariate Analysis

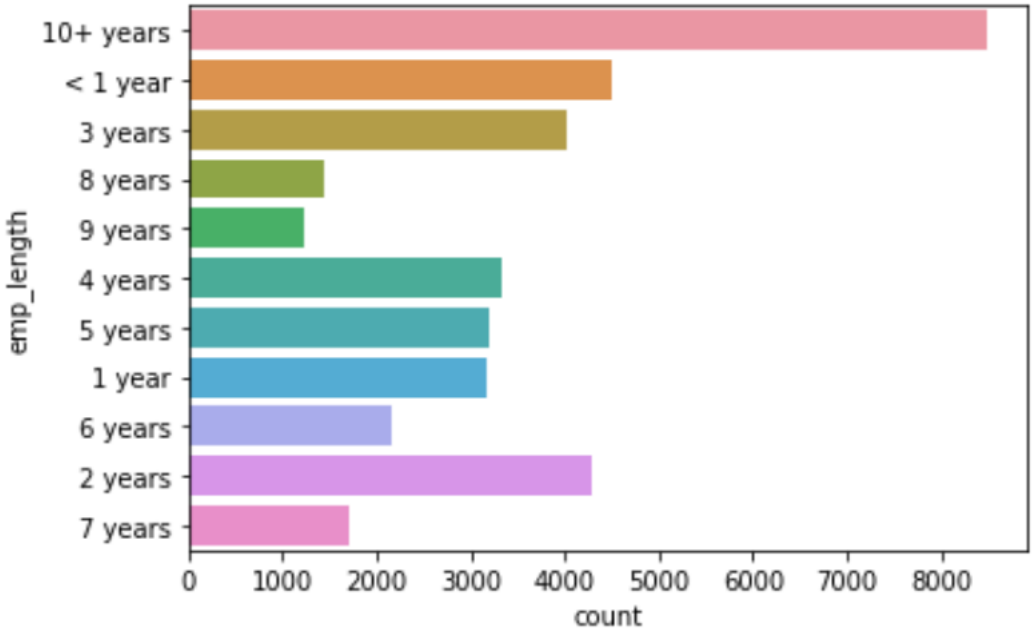


People in high risk group finds difficult to get loans.



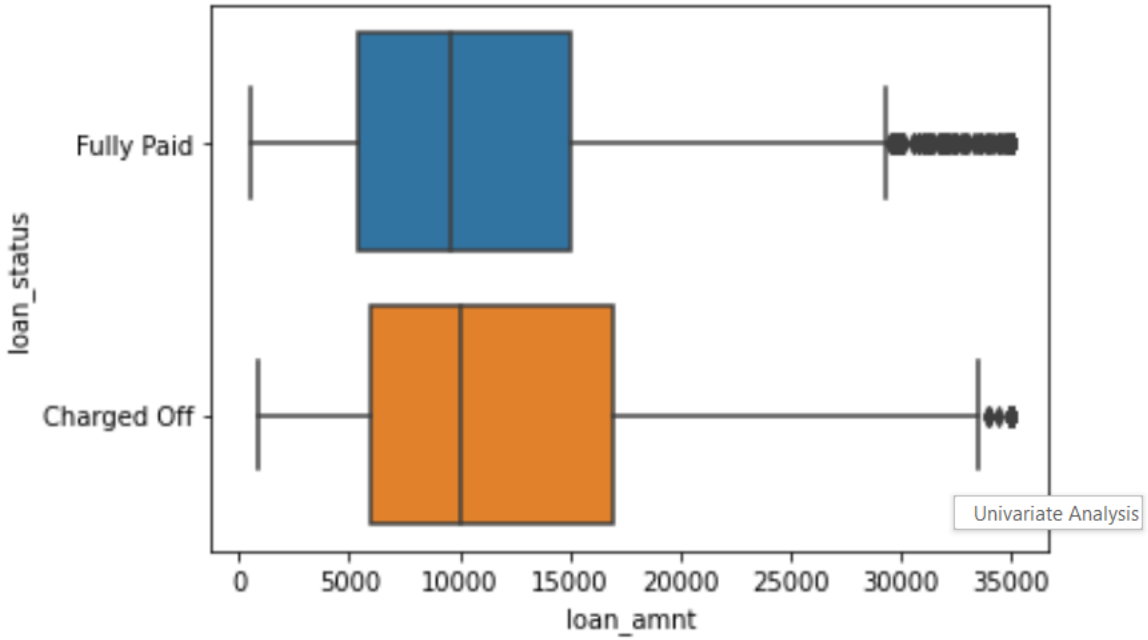
Short terms loans are more compare to long term.

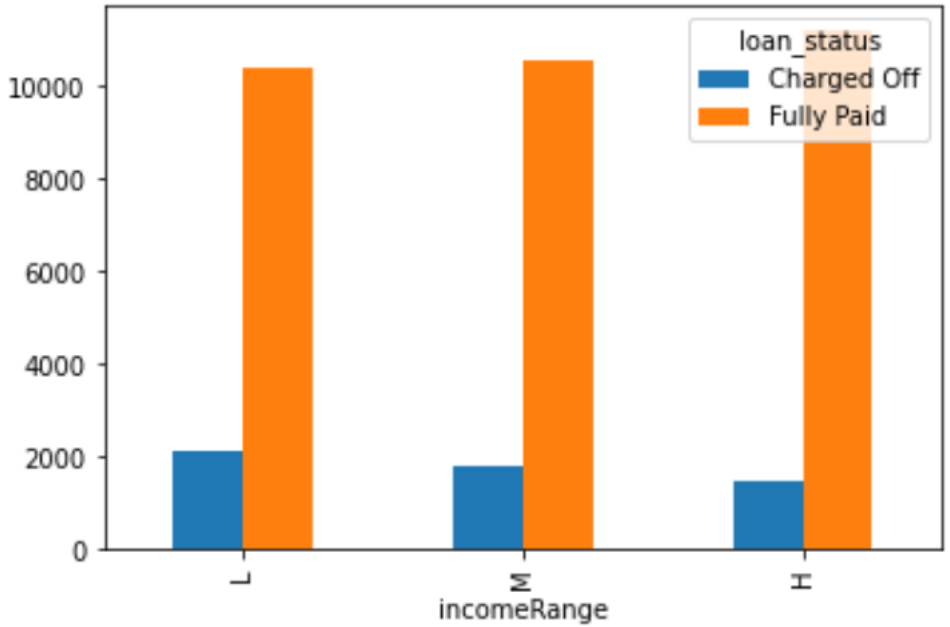
Univariate Analysis



People with 10+ years experience has taken more loans.

Segmented Univariate Analysis

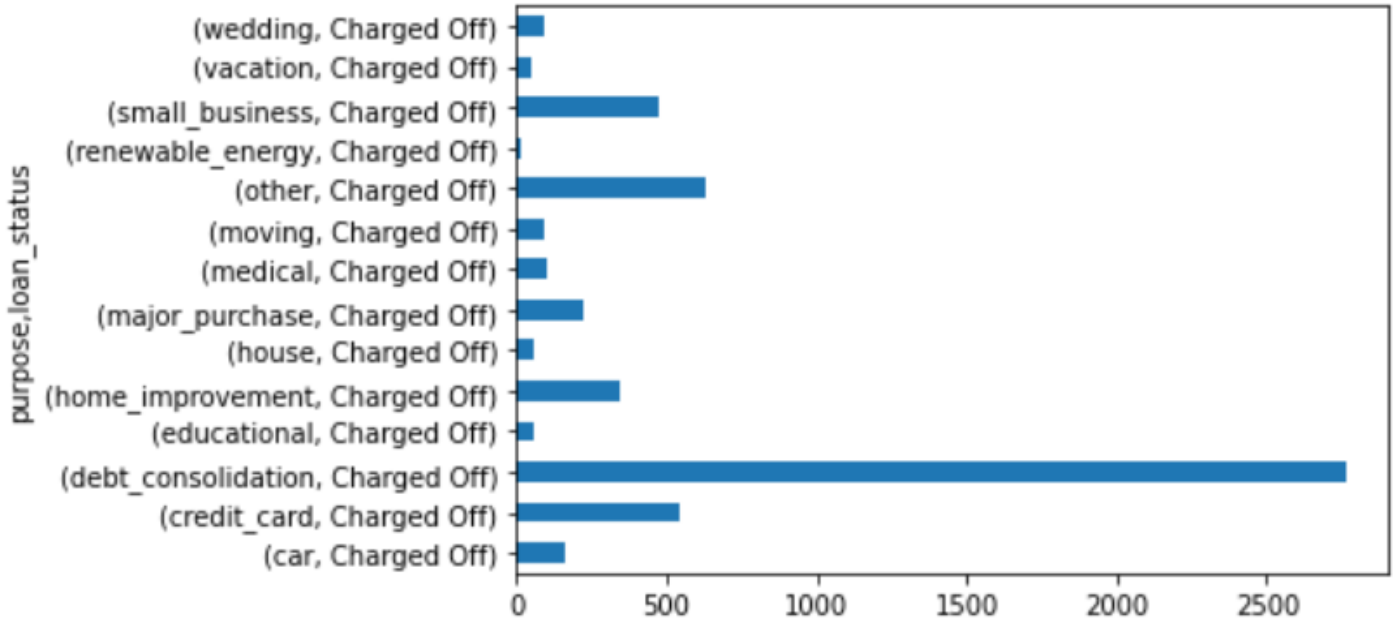




a) The median of the loan amount charged off is almost equal to the median of the Fully paid.
b) The upper quartile indicates that charged off happens on higher loan amounts.

When the income range increases the charged Off cases decreases.

Bivariate Analysis

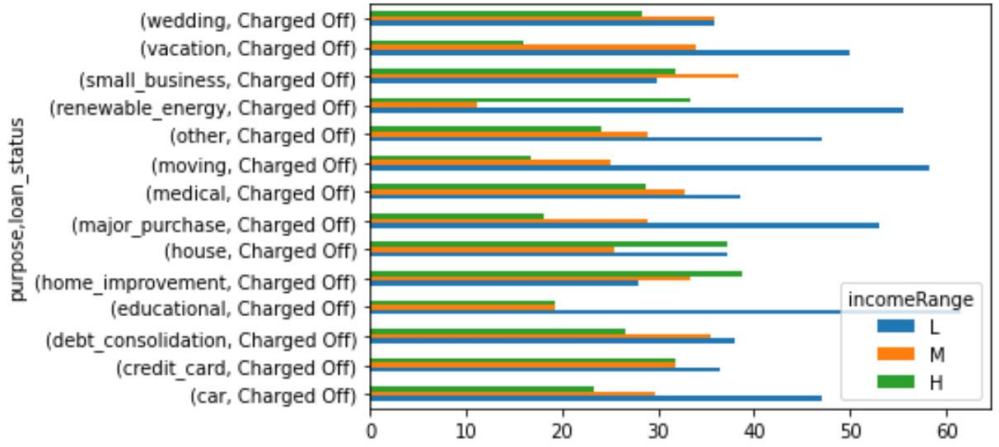


Purpose vs Loan status

Purpose vs Loan status vs Loan amount

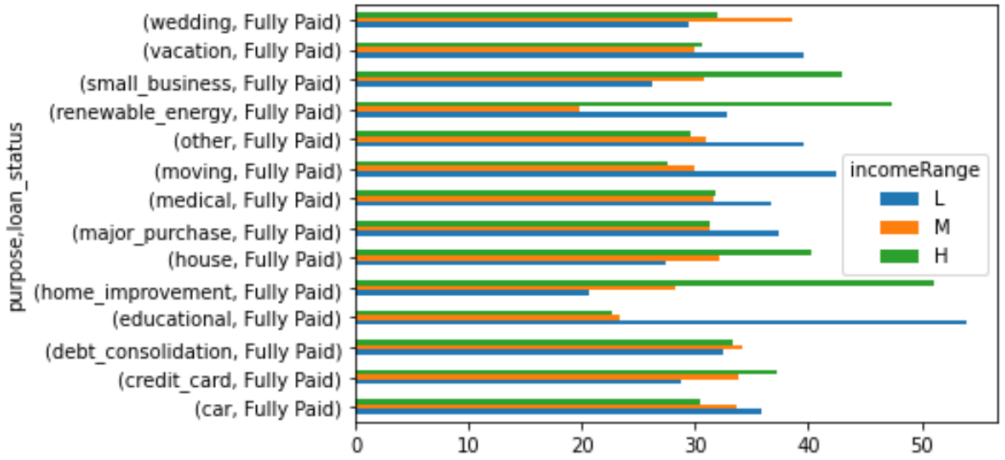
The loan amount taken for credit-card, debt consolidation, small business are among the top 3 Charged Off

Analysis of Percentage of Purpose among all Income range against loan status



	incomeRange	L	M	H	All
purpose	loan_status				
car	Charged Off	47.096774	29.677419	23.225806	100.0
credit_card	Charged Off	36.470588	31.764706	31.764706	100.0
debt_consolidation	Charged Off	37.938760	35.548917	26.512323	100.0
educational	Charged Off	61.538462	19.230769	19.230769	100.0
home_improvement	Charged Off	27.878788	33.333333	38.787879	100.0

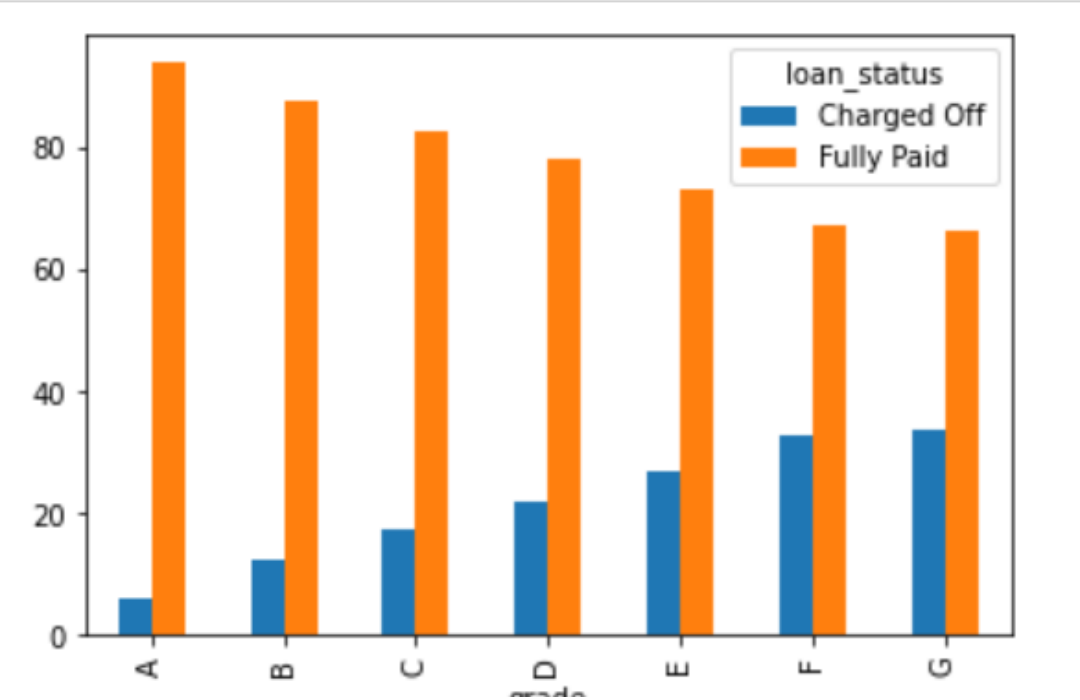
Low income group creates more charged offs for educational, moving, renewable energy loans.



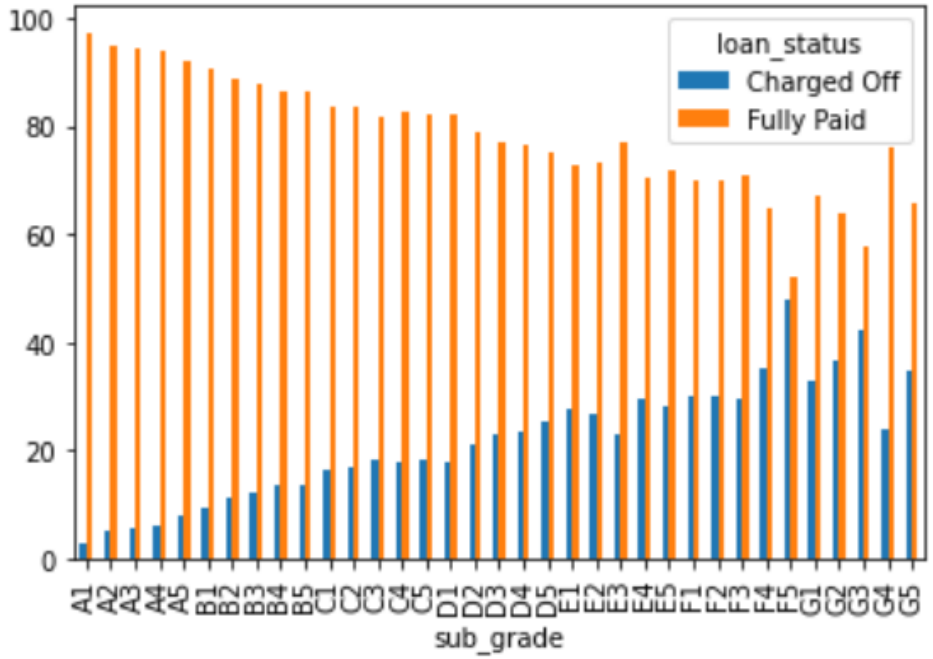
	incomeRange	L	M	H	All
purpose	loan_status				
car	Fully Paid	35.885538	33.720031	30.394432	100.0
credit_card	Fully Paid	28.828624	33.910665	37.260711	100.0
debt_consolidation	Fully Paid	32.519349	34.174006	33.306645	100.0
educational	Fully Paid	53.962264	23.396226	22.641509	100.0
home_improvement	Fully Paid	20.595432	28.262643	51.141925	100.0

Low income group is highest in paying off educational loans though they are highest charged off in the same category.

Analysis of Percentage of Loan applicants charged off among Loan status against grades



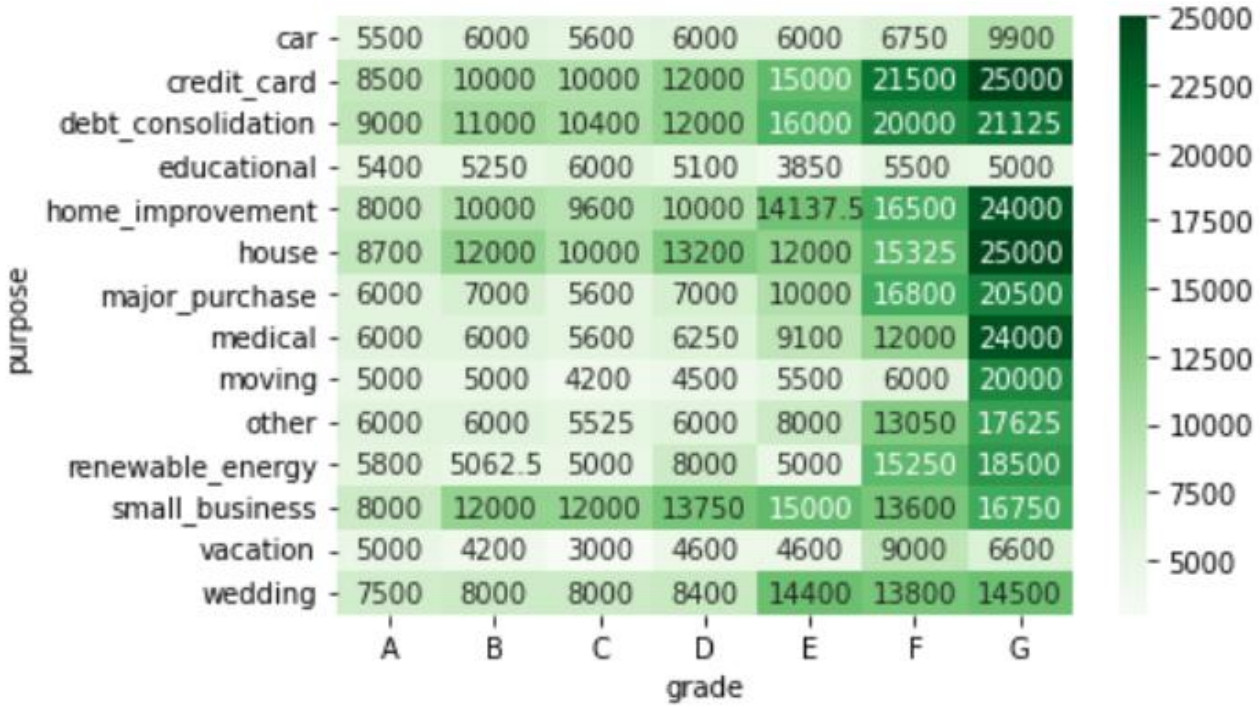
The percentage of charged off increases with risk represented by higher Grades



The higher the subgrade, the higher the percentage of Charged Off

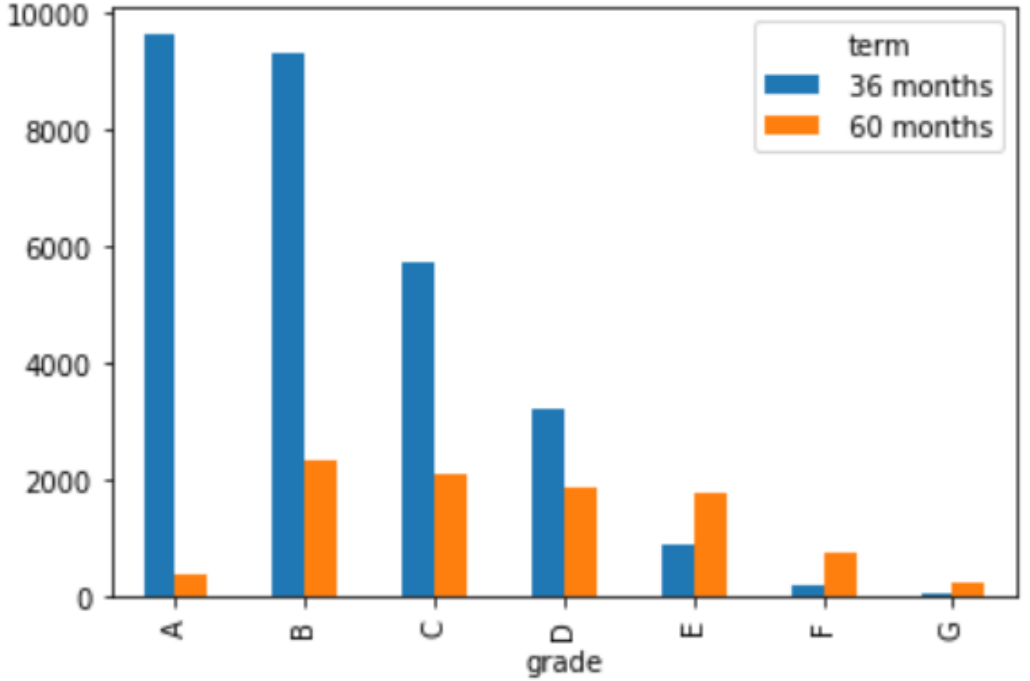
Additional analysis of Purpose , Grade , Loan amount

Analysis of Grade vs Purpose, Loan amount



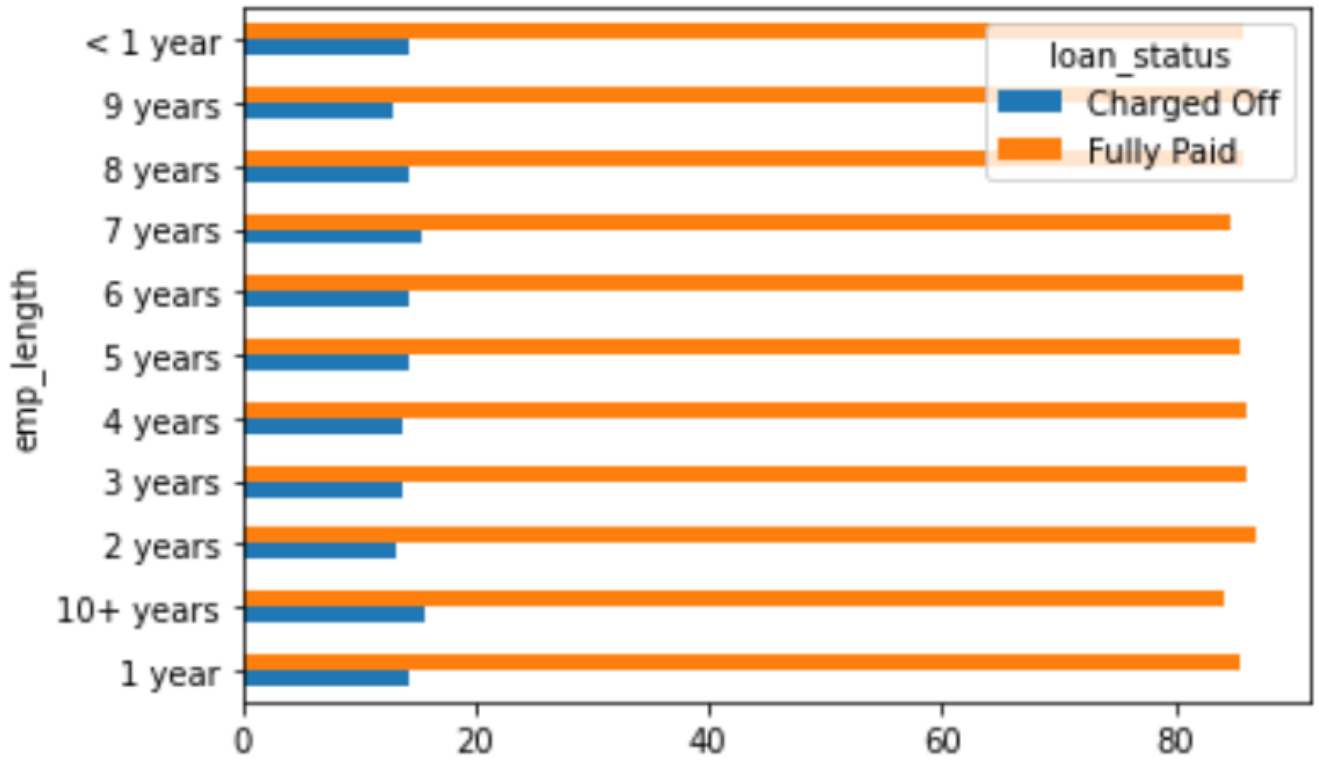
High risk group needs more loan for credit Card, medical, home improvement purposes

Analysis of Grade , Loan Term



While the credit score deteriorates, people tend to go for long term loans.

Analysis of Employment length vs Loan Status



Charged Off vs Fully paid in numbers

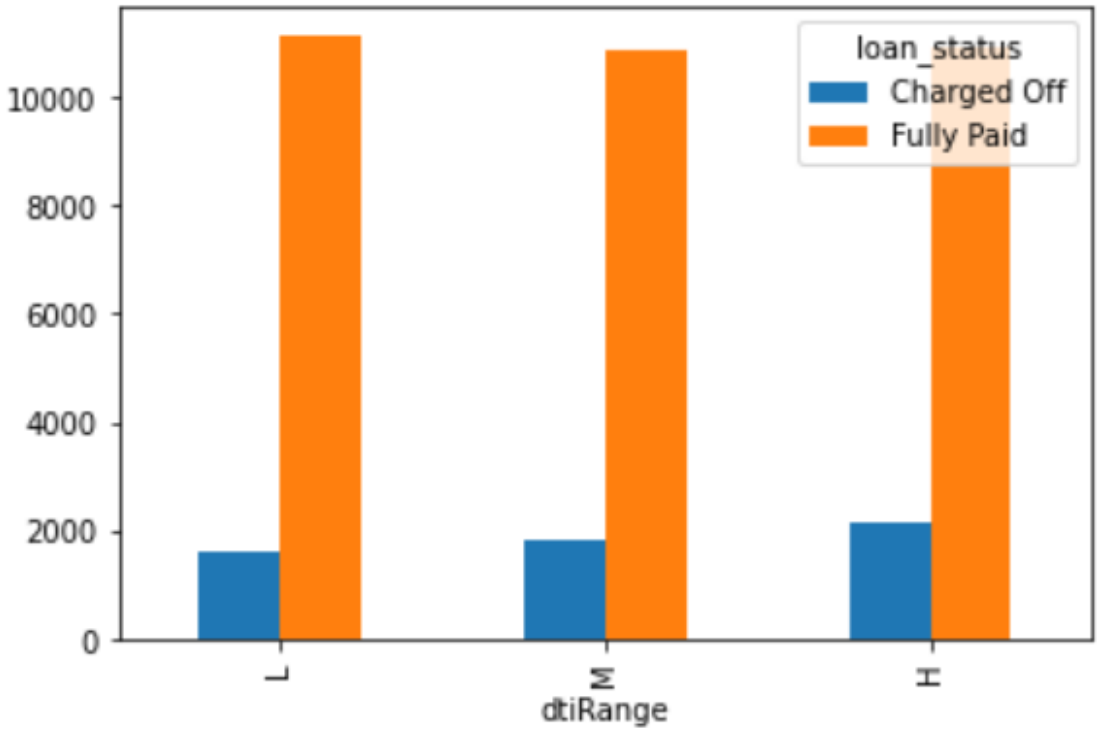
loan_status	Charged Off	Fully Paid	All
emp_length			
1 year	456	2711	3167
10+ years	1331	7148	8479
2 years	566	3724	4290
3 years	555	3456	4011
4 years	462	2880	3342

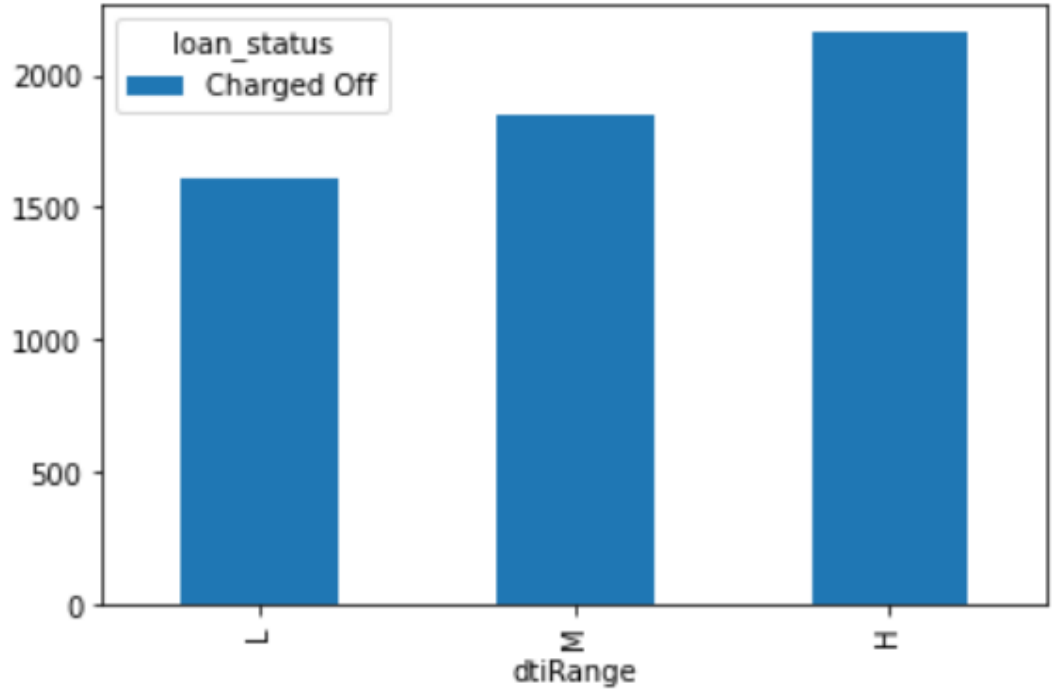
Charged Off vs Fully paid in Percentage

loan_status	Charged Off	Fully Paid	All
emp_length			
1 year	14.398484	85.601516	100.0
10+ years	15.697606	84.302394	100.0
2 years	13.193473	86.806527	100.0
3 years	13.836948	86.163052	100.0
4 years	13.824057	86.175943	100.0

From the total number of charged off cases, it appears that 10+ years are among the most charged off cases. But when we consider the percentage of charged off cases within the Loan status, it is understood that they are within +/-2% difference among all employment years.

Analysis DTI against Loan status

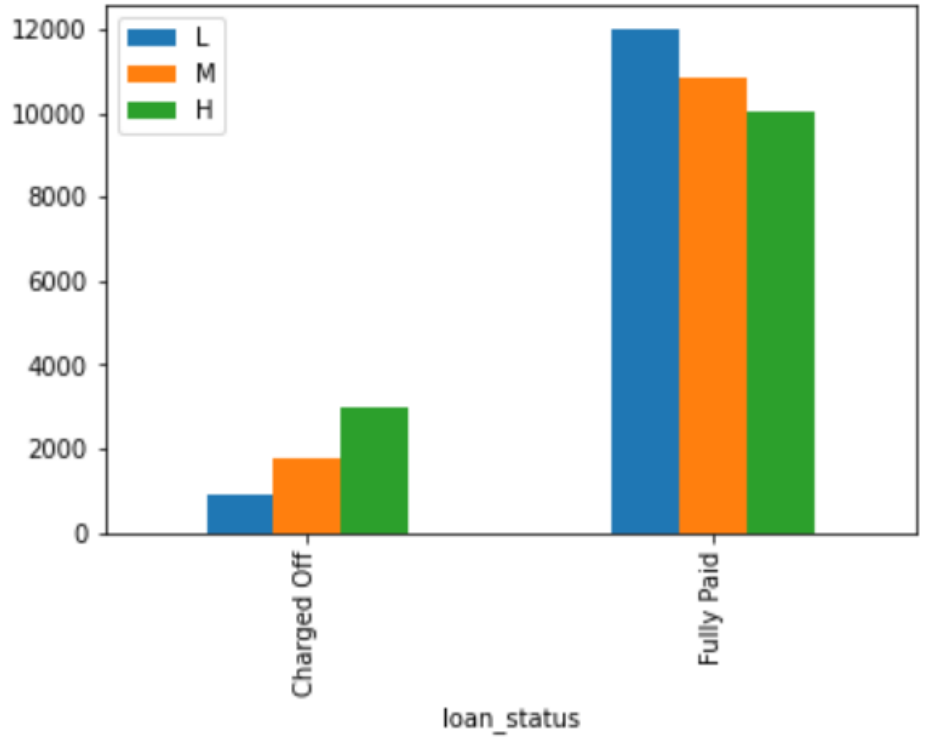




Higher the DTI then higher the charged Off

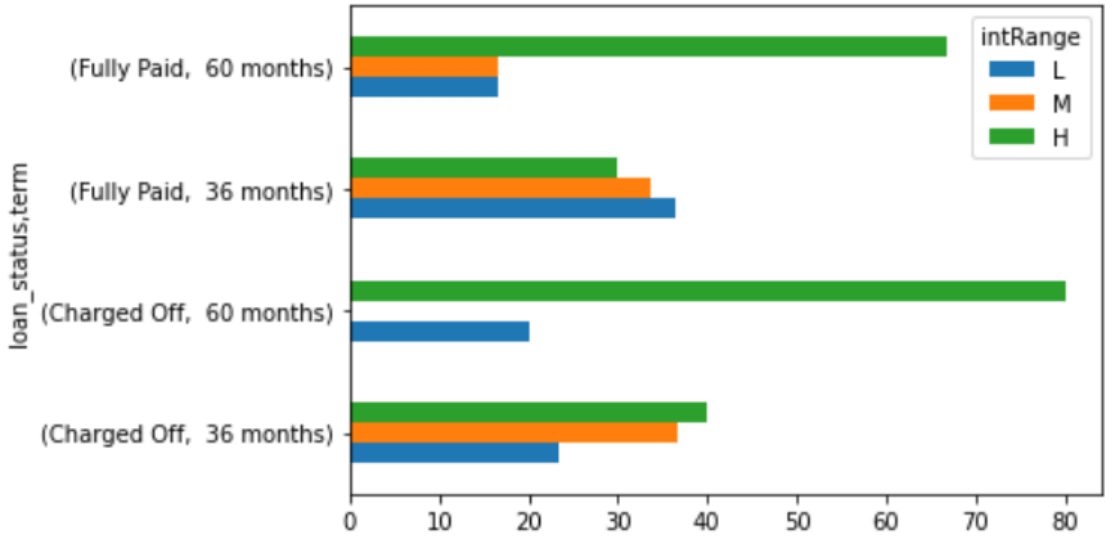
Analysis of Interest rate dependencies

Analysis of Loan status, income range and Interest range



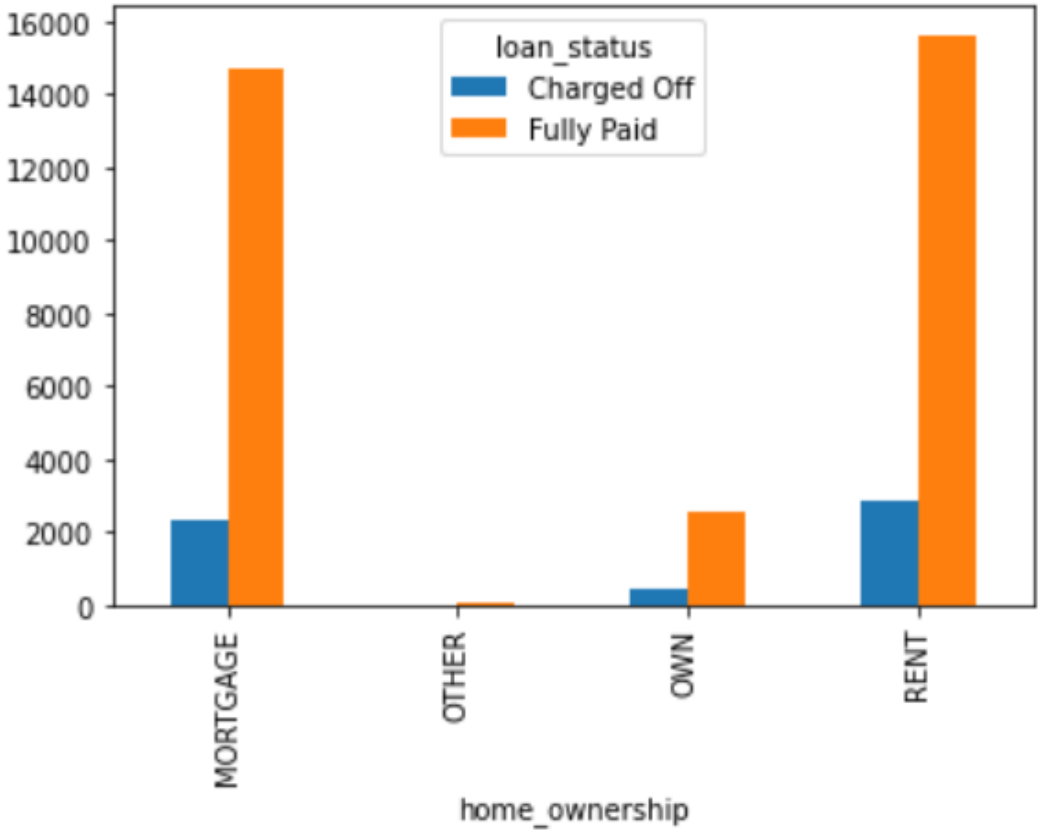
High interest rates get more charged Off while Low interest rates gets fully Paid

Analysis of Loan status, Term and Interest range specifically for educational loans



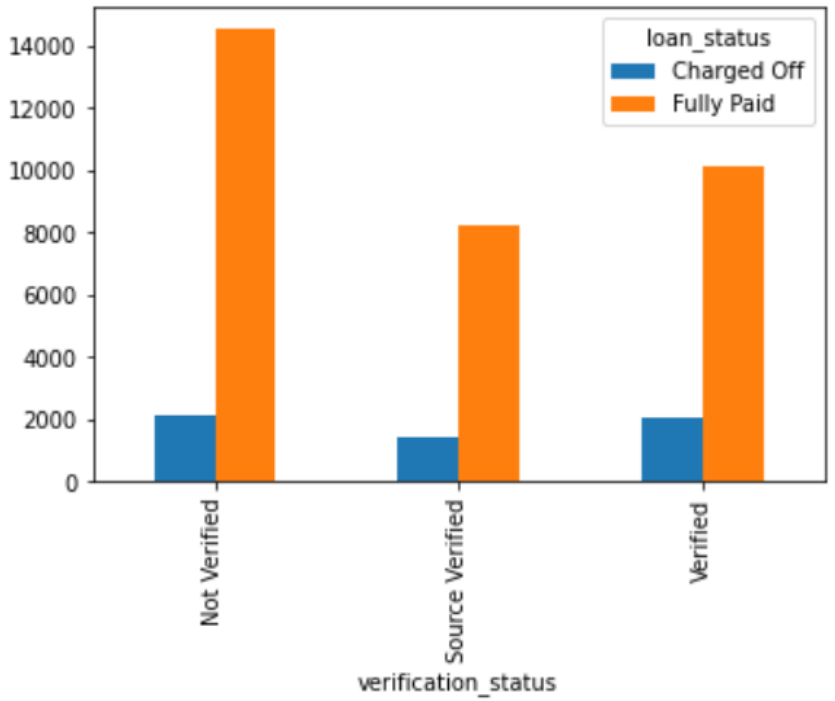
For educational loans, when the interest rate is high, Low income group will get more charged Off for short term and long term loans.

Analysis of Home ownership against Loan status



People who lives in Rented accommodation and those who has mortgage has more charged off than people who has Home ownership.

Analysis of Verification status against Loan status



loan_status	Charged Off	Fully Paid
verification_status		
Not Verified	2142	14540
Source Verified	1434	8236
Verified	2050	10152

The charged off cases of people with 'Not Verified' verification status is slightly high compared to others.

Step 4 : Recommendations :

Please see the driving factors/ dependencies identified to be considered while loan approval

1. The percentage of charged off increases with risk represented by higher **'Grades' and 'Sub grades'.**
2. While the credit score deteriorates (increase in grade / sub grade) , people tend to go for **long term** loans.
3. High risk group needs more loan for **purposes** such as credit Card, medical, home improvement purposes
4. Low **annual income** group creates more charged offs for purposes : educational, moving, renewable energy loans.
5. Higher the **DTI** then higher the charged Off
6. High **interest rates** get more charged Off while Low interest rates gets fully Paid
7. People who lives in Rented accommodation and those who has mortgage has more charged off than people who has **Home ownership**.