Lending Club Data Case Study

Pitch Deck By

Durga Srinivasan

Abbas Bagwala

Table of Contents

Problem Statement

Lending Club is a website/service that operate as an interface between borrowers and lenders. Their business model requires them to provide verification data to lenders on the borrowers to help determine the action required on a loan application - Either acceptance or rejection.

The problem at hand is to analyse the dataset which holds information about the borrower's data who have applied for a loan and determine various factors/variables that will lead to borrower defaulting the loan.

This key information that comes out of the data analysis task will help the lenders take an well informed decision to either accept the loan application (so borrower would pay in full - leading to profit) or reject the application (again leading to prevention of loss)

Data Understanding

Data for our analysis was provided in .csv file. The first step was to observe the data and understand the information that is available.

From initial look the following variables from the dataset seemed too important for our analysis

- Loan Status – Provides status if the loan (Target Variable)
- Loan amount – That borrower applied for
- Term – Tenure for Repayment
- Interest Rate – Suggested interest rate
- Grade/subgrade – Grade the borrower falls into
- Annual income
- Verification status – Verified (Income verified by tax returns/pay slips),
  Source Verified (Verified by 3rd party
  Not Verified
- DTI – Debt to Income ratio
- Employment length
- Home Ownership – Rent, Mortgage or own
- Purpose of loan

Data Cleaning

The entire analysis is to determine the borrowers who will default based on various factors.

Following details were considered and performed to clean data –

- The loans that are currently in payment do not fall into this category as they have been already granted the loans. Any loans in this status are removed.
- All columns with behavioral data of customers were removed as they correspond to loans in 'Current' status
- All columns with missing values over 30% was removed. This percentage was set on looking at other columns where missing percentages were less than 1%.
- Drop duplicates if any

Data Manipulation/Imputation

- Manipulation

  Date columns – Year was extracted separately into another column

- Fill missing Values/Impute

  emp_title – Missing values ignored in this column as it differ, and we can't impute it with a modal value

  title – Only 11 missing values which can be imputed from the column purpose

- Fix Data

  1. Convert interest rate column into Float
  2. Extract only the number of years from emp_length column

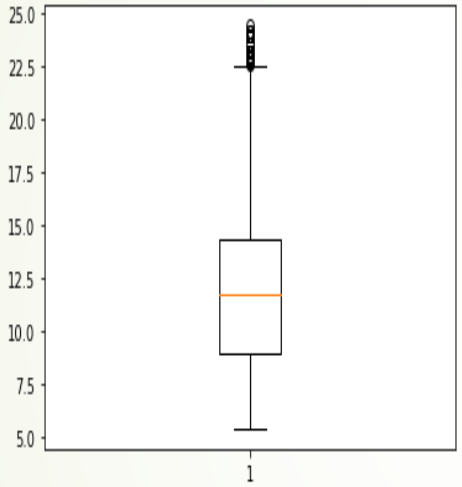
- Sanity Checks

  Check if there are any other loan status apart from Fully Paid or Charged off

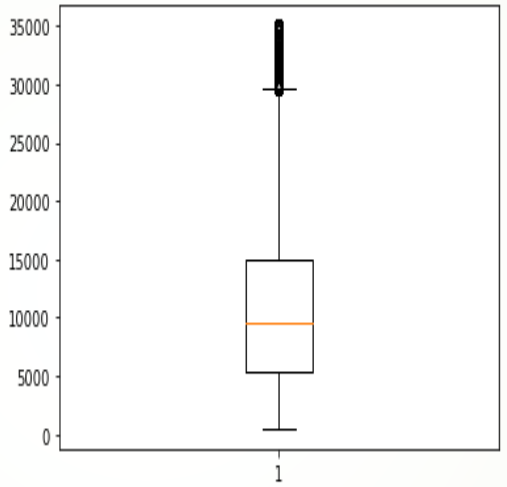
  Make sure that loan amount requested by borrower is less than agency approved funded amount and funded amount invested by the lender.

Outlier Analysis

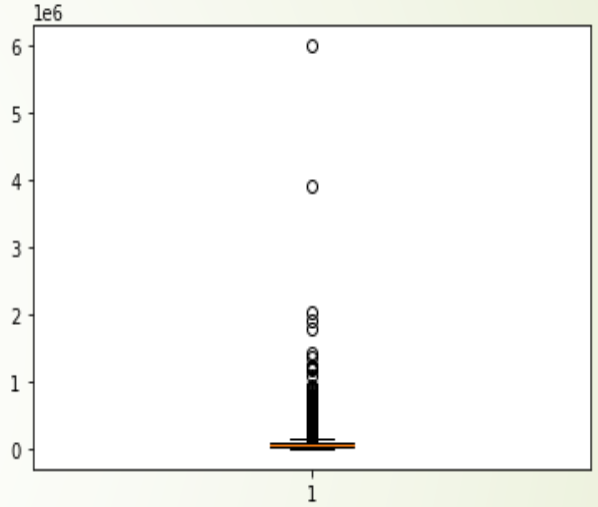
Outlier check for Interest Rates



Outlier check for Loan Amount



Outlier check for Annual Income



Insights:

1. The 75th Percentile for int_rate falls between 12.5% and 15.0%. There are loan applications which have int_rate beyond the 75th percentile
2. Loan amount has 75th percentile around 15K. Max goes around 35K which is not a huge amount. We cannot consider this loan amounts having outliers for our analysis
3. Annual income overall is less than 100K which is very common in US

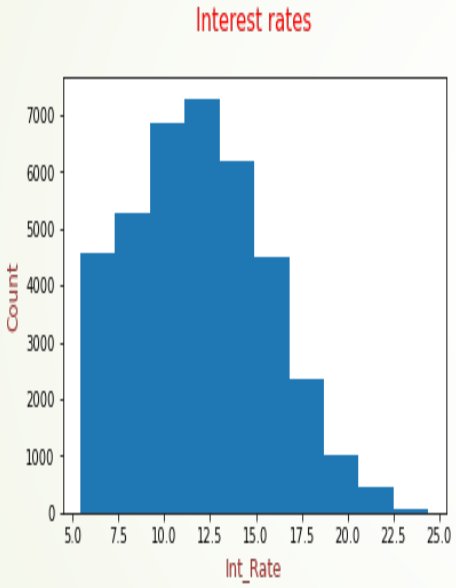
Recommendations:

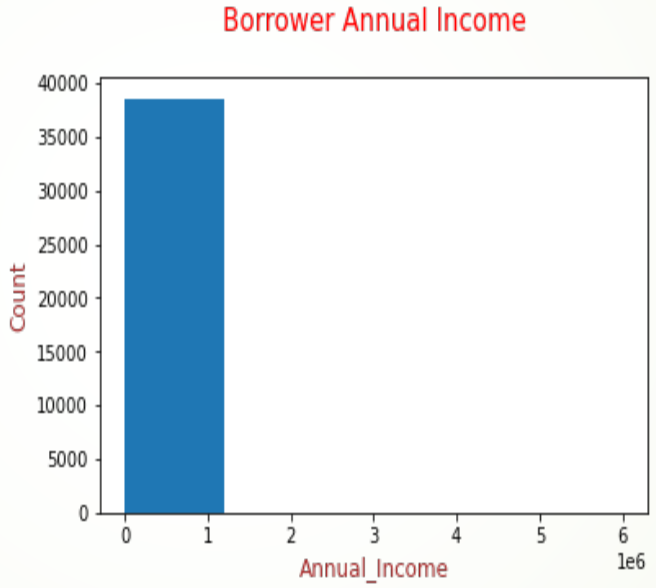
We will retain the data frame as the number of records is minimal and will not impact analysis.

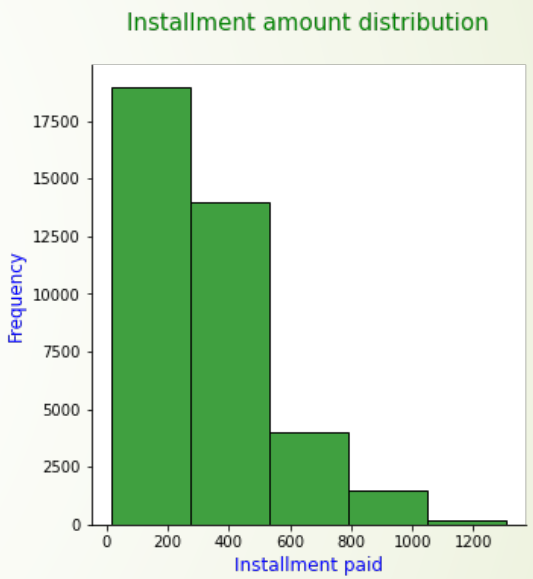
Data Analysis - Univariate

Univariate Analysis: Numerical variables

Under this section numerical variables were analyzed to determine the frequency of occurrence, distribution and composition. Next few slides will list those analysis done.







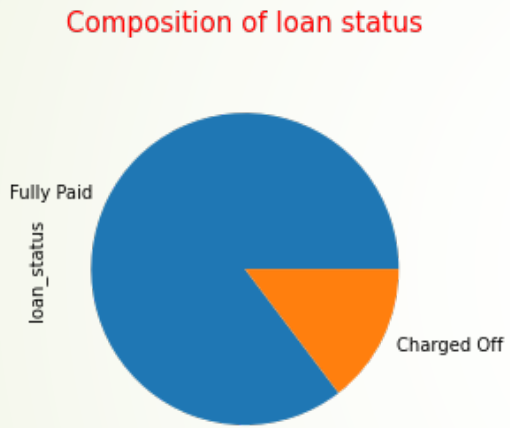
Insights –

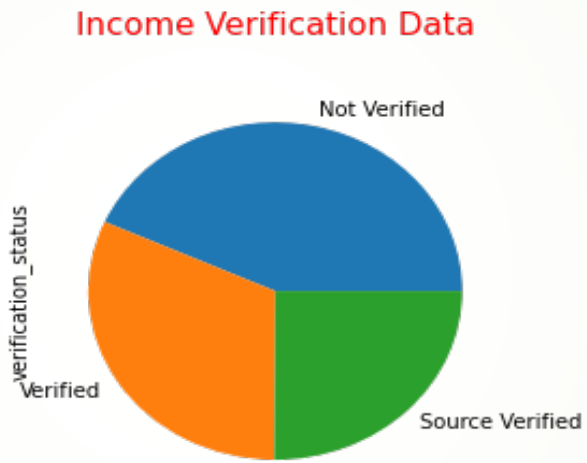
Interest rates are generally around 10 to 15.5 % for many borrowers.
Annual income on average is low for all applicants. Around 15K
Installments paid are in higher count when it is less than $500

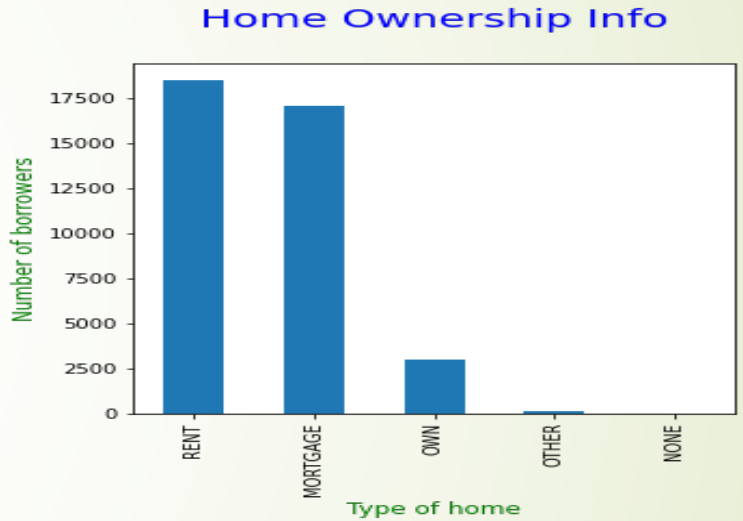
Univariate Contd...

Segmented Univariate –

Under this section segmented univariate analysis done to determine the distribution and composition. Refer to insights gained from these –







Insights –

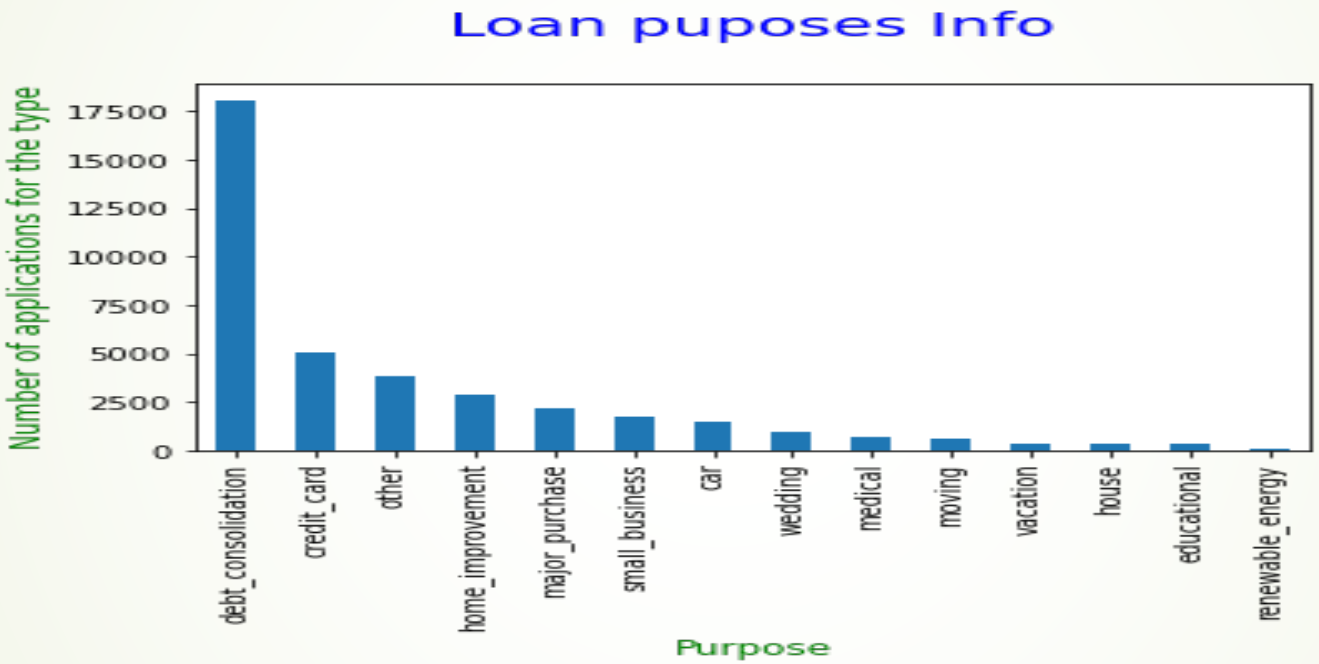
1. 85 % of loans have their status as 'Fully Paid' and 14% Charged Off
2. Annual income – 43% of applications not verified, 25% verified by 3rd party and 31% verified.
3. Equal amount of people either RENT a home or have a MORTGAGE and very few OWN home without any mortgage.

Recommendations –

1. Being on RENT/MORTGAGE and annual income verified by 3rd party/not verified might result in more defaulters

Univariate Contd...

Under this analysis done to determine the distribution of purpose of the loan . Refer to insights gained from these –



Insights –

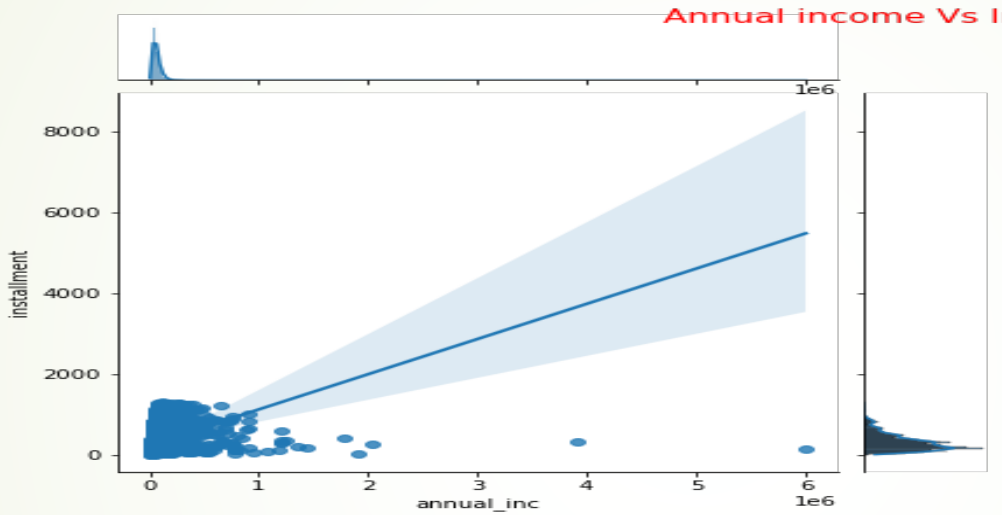
Most of the loans had an applied for debt consolidation followed by credit cards

Recommendations –

Determine the loan status when the purpose is these two values. It would give an idea if there is any tilt towards defaulting

Data Analysis – Bivariate

Let's see how two variables behave/vary based on one another –



Insights:

Annual income against installment is linear. Higher annual income, the installment amount paid is higher.

Recommendations:

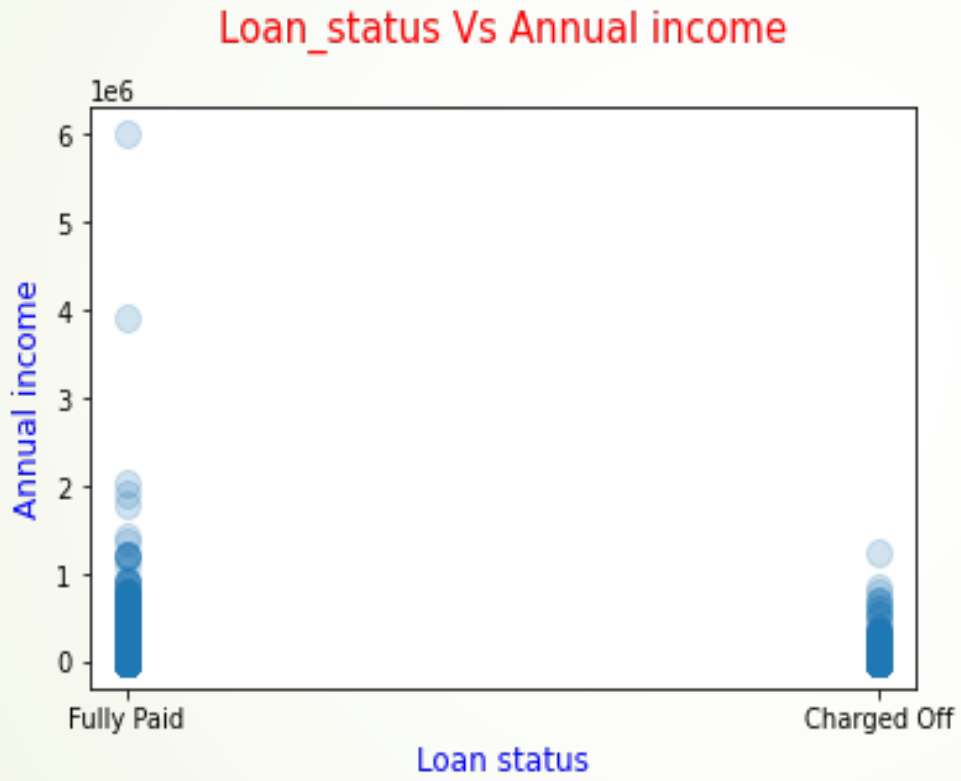
Expected behavior. We can determine how annual income behaves with other variables

Annual income Bucket:

To simplify analysis as we go down, we split the annual income into Very Low(VL), Low(L), Medium (M), High(H), Very High(VH) buckets

Bivariate Contd...

Annual income Vs Loan Status

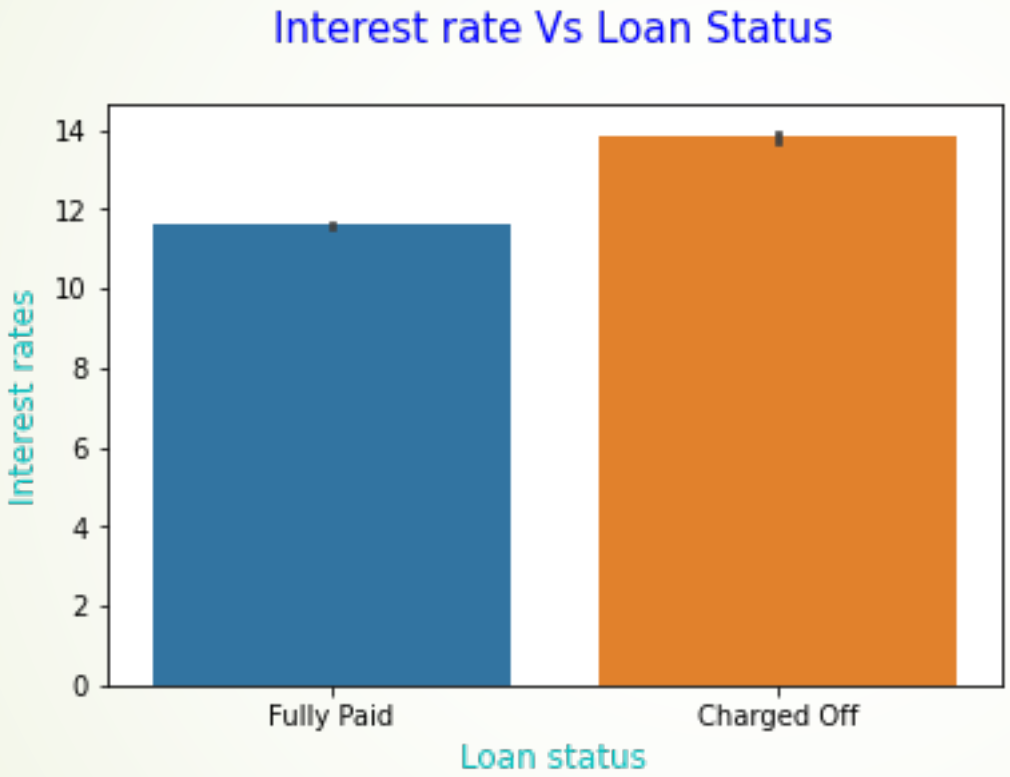


Insights:
Charge offs are higher when annual income is lesser.

Recommendation:
Charge offs are higher when annual income is lesser. But let's analyze further to get an understanding of how other variables impact the loan status

Bivariate Contd...

Loan Status Vs Interest Rate



Insights:
Charge offs happens at all kinds of interest rates

Recommendation:
Let's proceed further and analyze other variables.

Multivariate analysis

Let's compare more than two variables to understand how it impacts the target variable (Loan status).

How annual income and loan amount affect the loan status ?

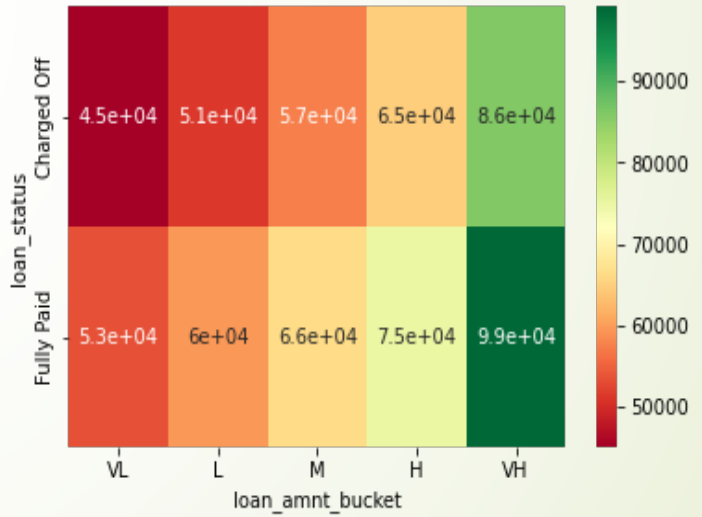
Loan amount bucket:
To simplify analysis as we go down, we split the loan amount requested into Very Low(VL), Low(L), Medium (M), High(H), Very High(VH) buckets

Pivot table –
We will create a pivot table with loan status as index, loan amount bucket as columns and values of annual income

Insights:
When loan amount is very low, low, medium and annual income is < 60K then more chances of Default

Recommendations –
Let's drill down further to see behavior of other variables

Correlation b/w loan status, loan amount and Annual income

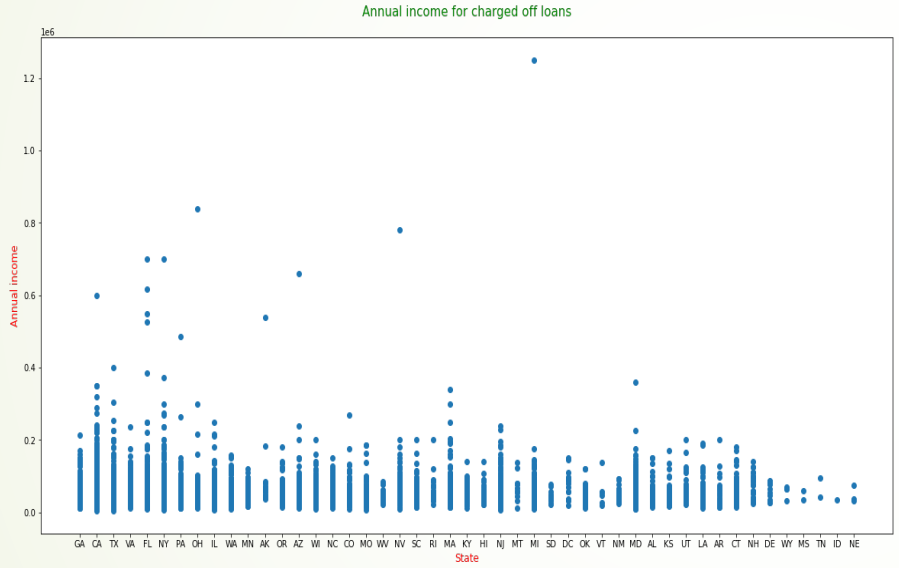


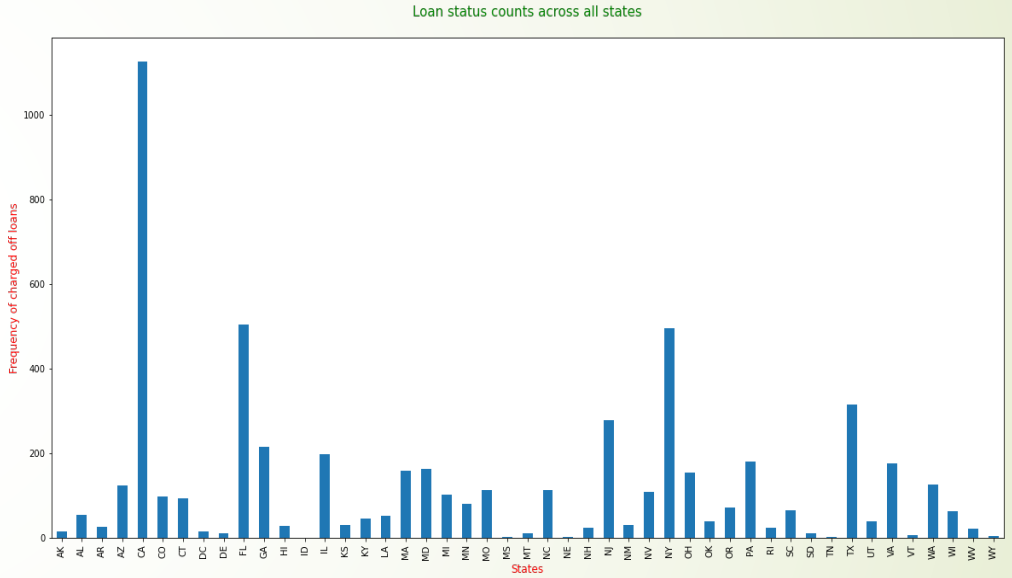
We create a heat map to see the relation these three variables have

Multivariate Contd...

How state of residence & annual income on these places impacts loan status.

To determine this accurately, we will create two data frames. One with only Charged-off loan status and other with Fully Paid status.





Insights:

From the above plots, more loans are charged off in CA, FL, NY and TX. These are states where annual income is comparably high to other states. But the annual income of the borrower is low, and cost of living is high

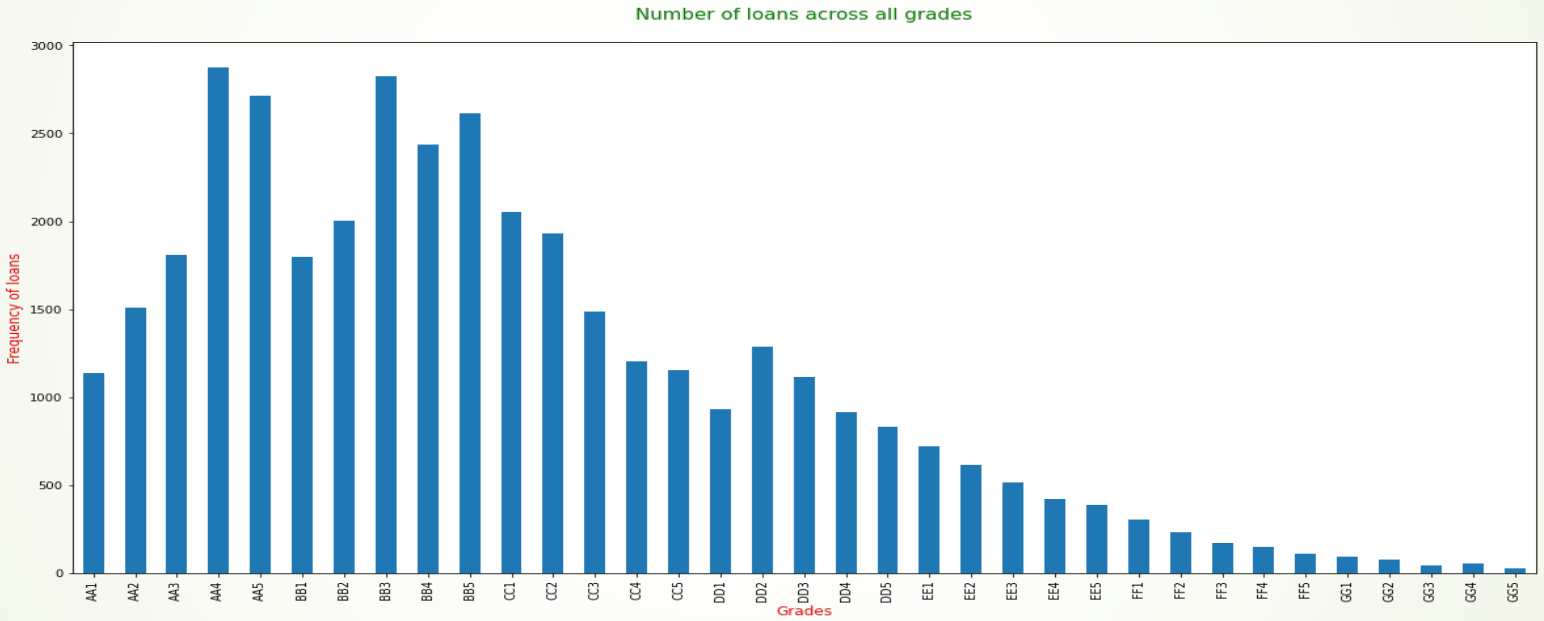
Recommendations:

These type of borrowers who live in states where cost of living is high and have a low/medium annual income have higher probability of defaulting. Hence make sure the annual income is comparably on the higher end

Multivariate Contd...

How the Grades of the borrowers impacts loan status

The two columns grade, sub-grade and we will merge into Grades for easier visualization



Insights:

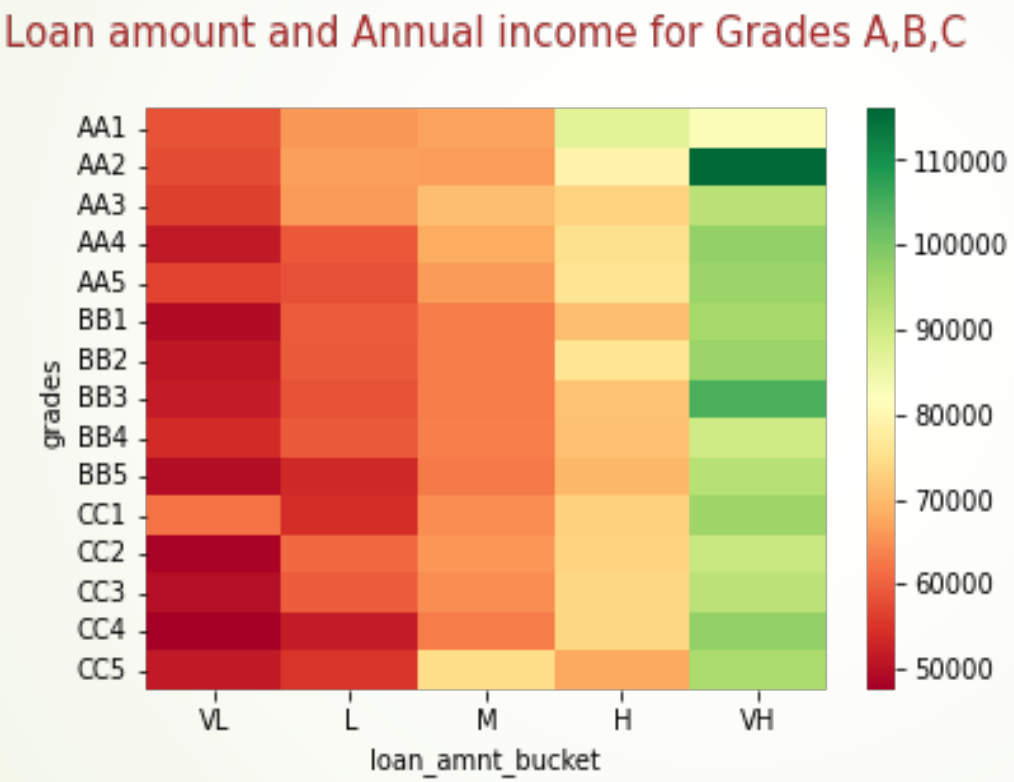
More loans are under grades A, B and C

Recommendation:

Let's see how annual income is distributed for these grades.

Multivariate Contd...

Loan amount distribution for loans in grades A, B, C



Insights:

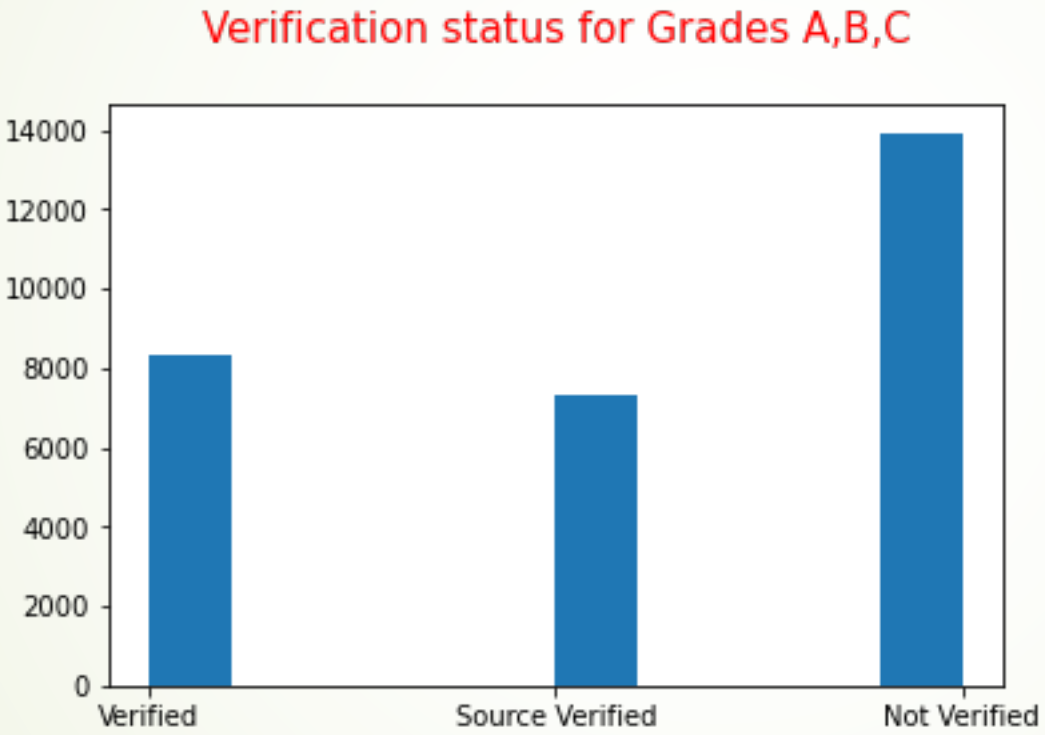
Grades AA4, CC4 - CC2 and BB5, BB3 - BB1 have low income when compared to other grades. These might impact borrowers paying the loan back

Recommendation:

Determine the verification status so we can see if there are any relation to it

Multivariate Contd...

Plot a histogram on verification status for the borrowers whose grades are A, B, C



Insights:

The number of applicants whose is income is Not verified is higher. Percentage of actual verification done is low compared to combined percentage of Source verified and Not verified

Recommendation:

When the borrowers fall under Grades AA4, CC4 - CC2 and BB5, BB3 - BB1, the probability of defaulting is higher when the income is not verified.

It is crucial to verify the income of these applicants.

Multivariate Contd...

Determine the relation between Loan status and DTI and annual income

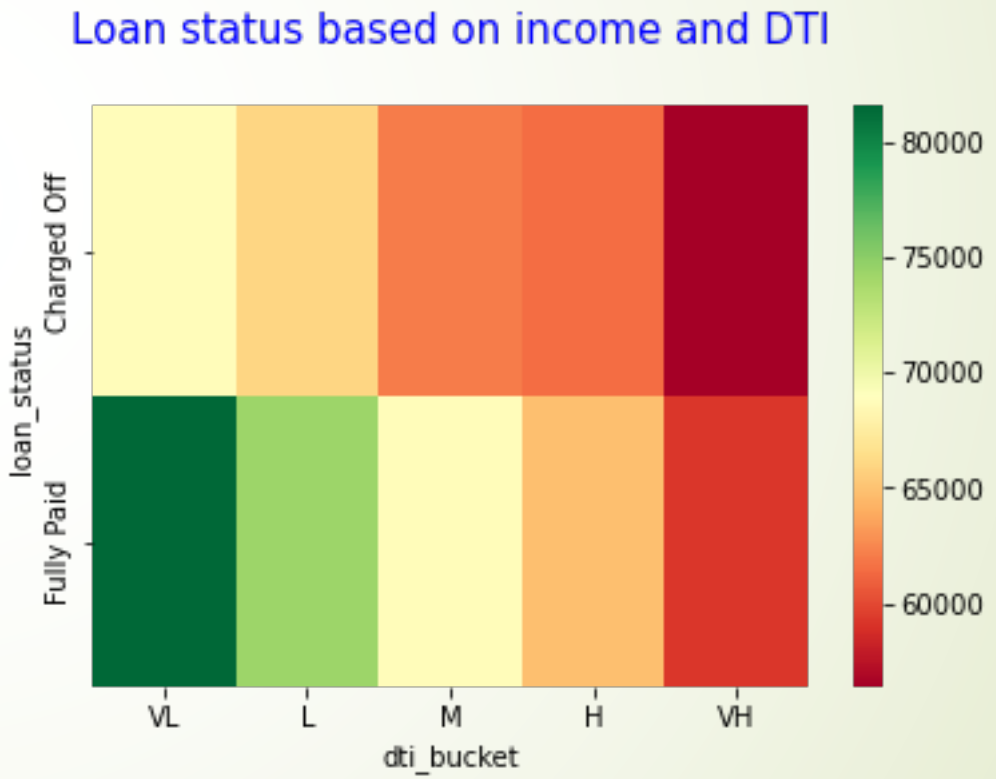
1. Create buckets for the dti value to plot it easier
2. Create a pivot table with data for loan status, dti bucket and annual income
3. Plot heat map for pivot table

Insights:

When DTI is too high and annual income is too low, charge offs are higher.
A low DTI and higher/medium income has higher probability of being fully repaying the loan

Recommendation:

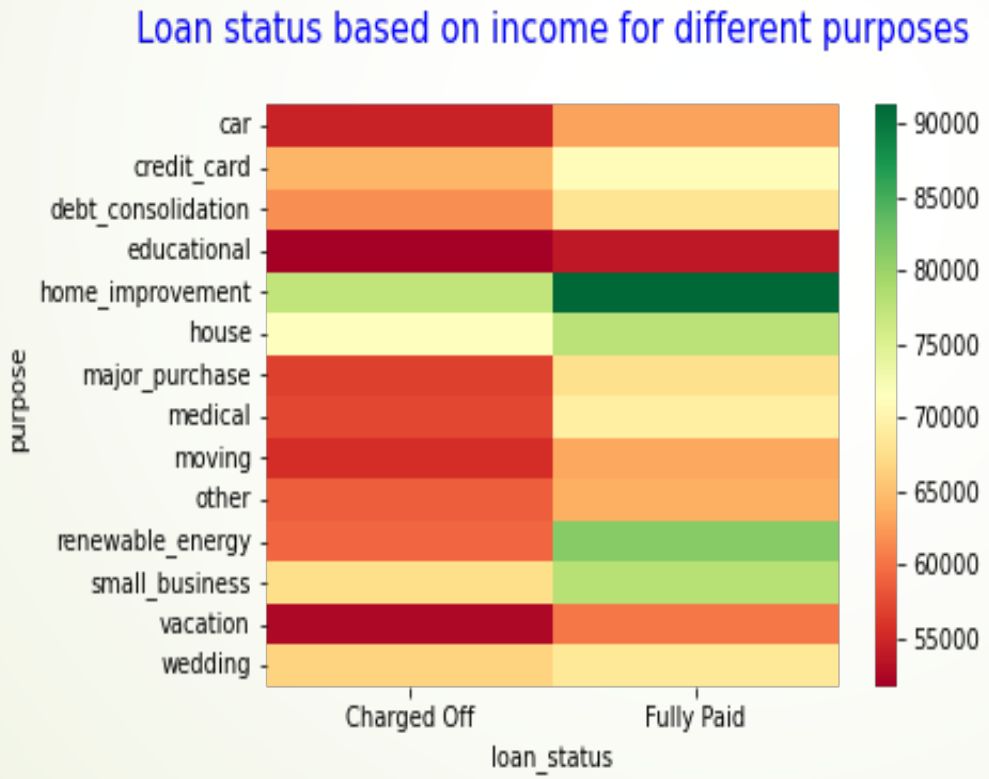
For a borrower when annual income is low or medium, the DTI value is expected to be very low or low to avoid defaulting



Multivariate Contd...

Determine if the purpose of loan has impact on loan status

1. Create a pivot table with data for loan status, dti bucket and annual income
2. Plot heat map for pivot table

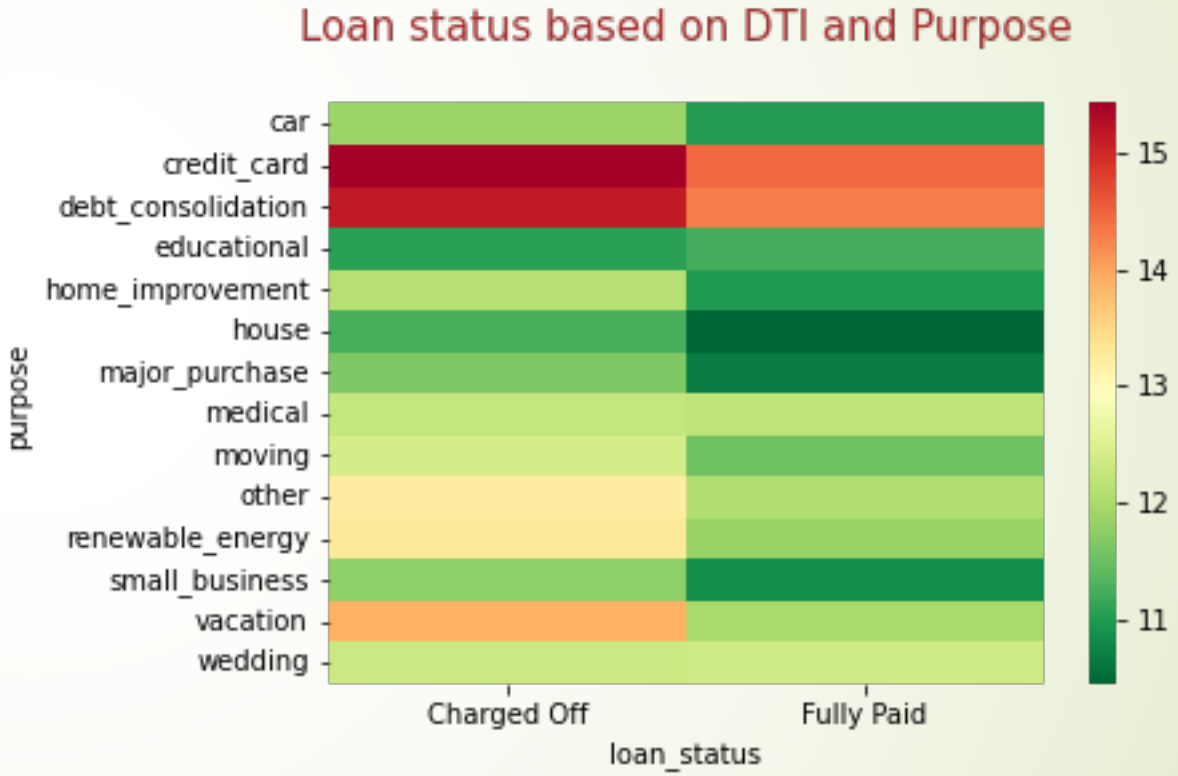


Insights:
When annual income is low, most of the purposes has higher risk of default

Recommendation:
Let's try how the same varies with DTI

Multivariate Contd...

1. Create a pivot table with data for purpose, loan status and dti bucket
2. Plot heat map for pivot table



Insights:

When DTI is higher and the most common purpose of debt consolidation, credit card are high risk factors.

Recommendation:

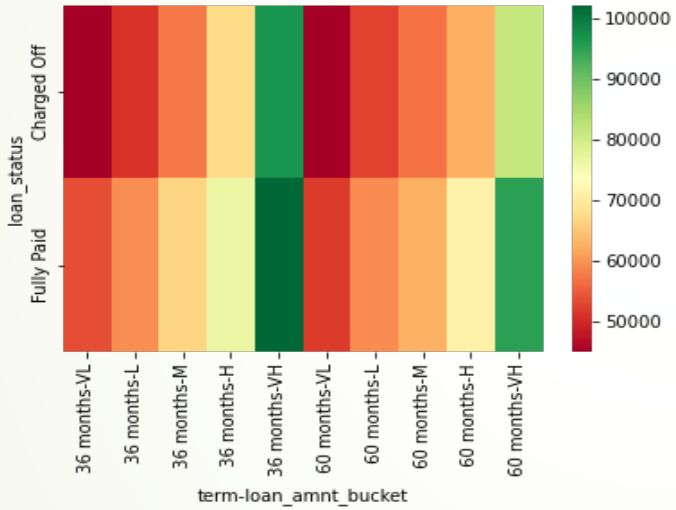
When a borrower has higher DTI and comes for a loan with purposes 'Debt Consolidation, credit card', they are highly likely to default

Multivariate Contd...

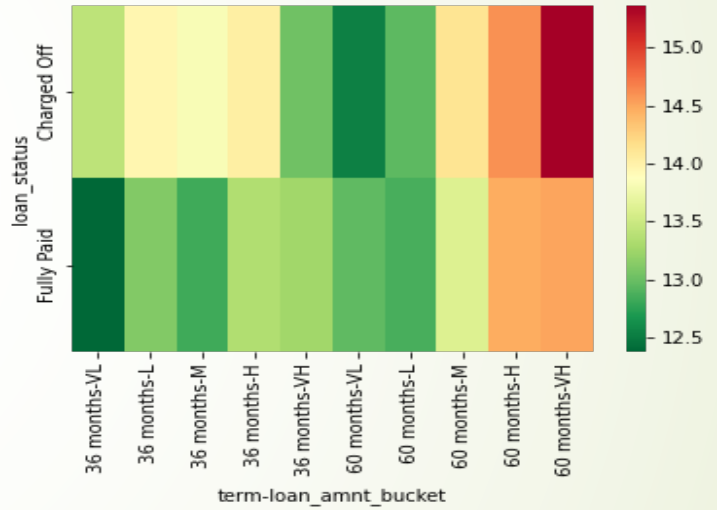
Determine how the number of terms influences loan status based on annual income and dti

1. Create a pivot tables with data for term, loan status, loan amount bucket and annual income/dti
2. Plot heat map for pivot tables

Loan status based on terms,loan amount and income



Loan status based on terms,loan amount and dti



Insights:

From the heatmaps above,

1. Low annual income irrespective of the number of terms - higher chances of getting defaulted.
2. Though number of terms are higher, a higher DTI and medium to high loan amount - higher chances of default.

Recommendations:

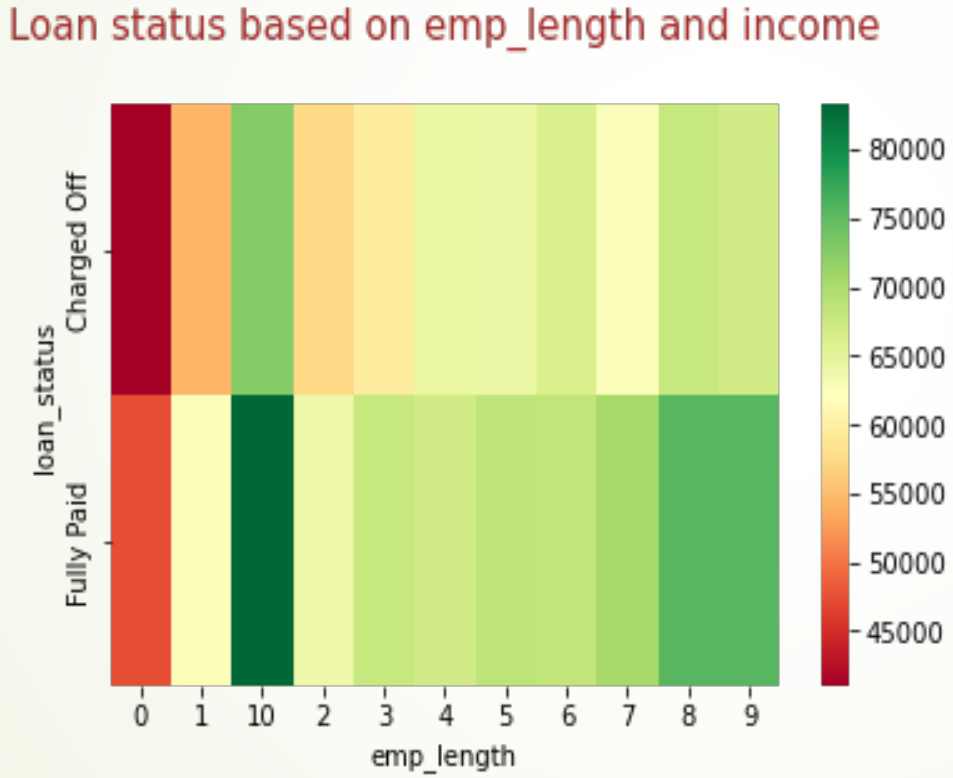
When DTI is higher and number of terms opted by borrower is higher, the loan amount approved should be lesser to avoid these borrowers defaulting.

Lower income borrowers are always high risk of defaulting loans.

Lower annual income seems to be a high-risk factor from above visualizations. Let's try to understand how the employment length affects this

1.Create a pivot table with data for emp_length, loan status, and annual income

2. Plot for the pivot table



Insights:

When the length of the employment is less than 3 years, the annual income falls under VL, L, and medium category.

Recommendation:

When employment length is low, the annual income will be low. Hence risk of defaulting is higher.

Conclusions

1. Loan status is **impacted by the state the borrower resides in and home ownership**. States with higher cost of living and low income will result in borrowers defaulting

2. Loan status is **impacted by the grade the borrower belongs** to. When the borrowers are in grades A, B and C and low income, then if annual income is not verified properly chances of defaulting is higher

3. Loan status is **impacted by the DTI of the borrower**. When DTI is higher and borrower has a low annual income, they are more likely to default

4. Loan status is **impacted by loan amount when borrower has a low income** and lesser loan amount, defaulting chances are higher. This could be due to human factor that we can somehow repay as the loan amount is lower and keep skipping installments

5. Loan status is **impacted by the purpose when the borrower has higher DTI and low, medium annual income**.

6. Loan **amount approved should be lower when a borrower has a higher DTI** and number of **terms is higher**.

7. When annual **income is low, irrespective of the loan amount and number of terms, chances for default is higher. These are huge risk borrowers.**

8. When the **length of employment is lesser (less than 3 years),** the annual income is lesser. These are high risk borrowers