Lending Club Case Study

Group Members:

Shubhada Prabhakar Babhulgaonkar
Shravan Kumar Yadav

Problem Statement

When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile. Two types of risks are associated with the bank’s decision:

- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

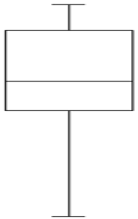
- If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company

Steps to solve the problem

Data Cleaning

Data Understanding

Data Filtering



Univariate Analysis

Bivariate Analysis

Recommendations

Data Cleaning

We have cleaned the data on the basis of below observations.

1) Remove columns having all NA values
2) Remove columns having single unique values as they do not have any impact on analysis.
3) Removing columns having values NA and blank.
4) Remove columns that are unique to each column. Example id, url
5) Removing the columns which are not know to bank at the time of loan application filled. These are created after loan is approved and running. For example revol_bal, out_prncp etc.
6) Removing columns in which we have null values more 90% of rows. example mths_since_last_record
7) Removing “%” character from interest rate column

Data Understanding

- No. of rows and columns in data
- Name of columns
- Not Null values for each column
- Data types of columns
- Number of numeric and string columns
- Check on mean, median,min,max for numeric columns
- View on sample of rows to see values stored in each column

Data Filtering

We observed that the data contains 3 different loan status.

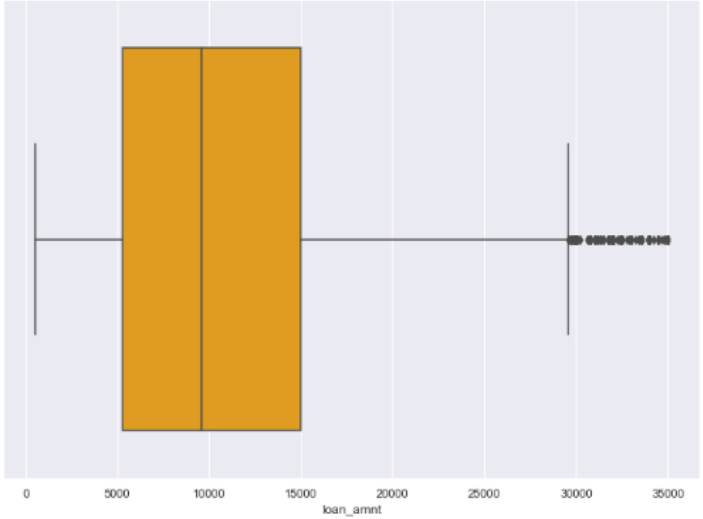
1) Fully Paid
2) Charged Off
3) Current

Fully paid are loans which are completely paid by borrowers. Charged off is where the borrower is unable to pay whole loan. Current is the current running loan which is neither complete nor charged off. For analysis the impact of attributes on loan status , we need to drop the loan with status “Current” because these are running loan. We are not including it in analysis because our main aim was to analyse the key factors which will impact the loan being fully paid or charged off. Therefore minimising the company loses.

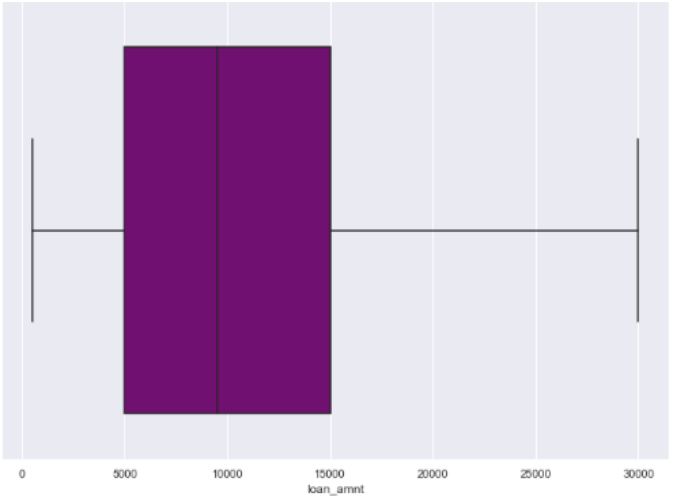
Univariate Analysis

Loan Amount

- In the given dataset, min loan amount is 500 while max loan amount is 35000. IQR is 9700. Value at 98th percentile is 30000. We are considering values beyond 98th percentile as outliers by looking at the plot.
- After removing outliers, Min loan amount is 500 while max loan amount is 30000. IQR is 13900.



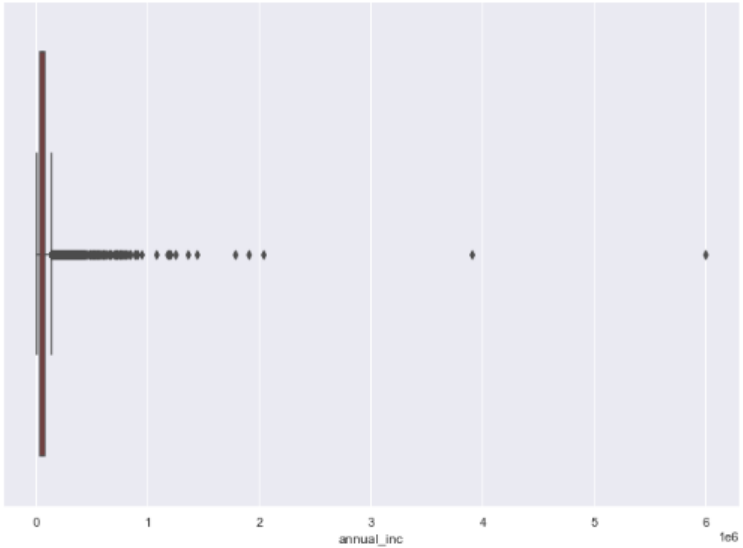
Before removing outliers



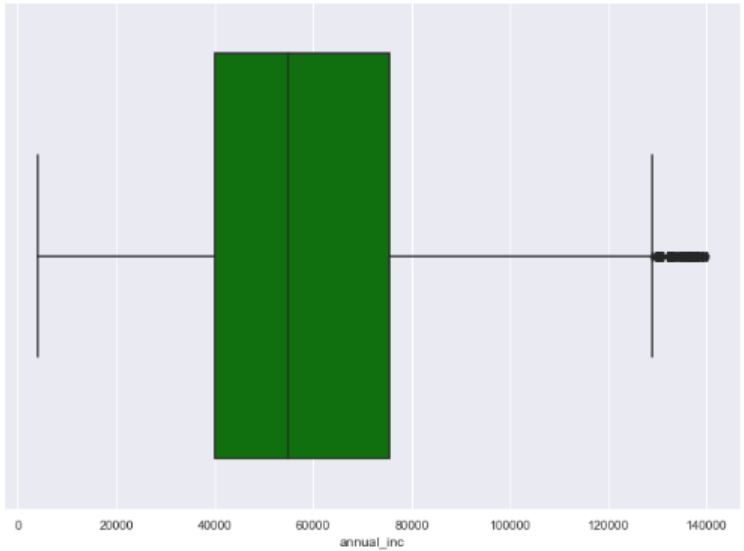
After removing outliers

Annual income

- In the given dataset, min annual income is 4000 while max annual income is 6000000. IQR is 9700. Value at 95th percentile is 140000. We are considering values beyond 95th percentile as outliers by looking at the plot.
- After removing outliers, Min loan amount is 4000 while max loan amount is 139992. IQR is 35600.



Before removing outliers



After removing outliers

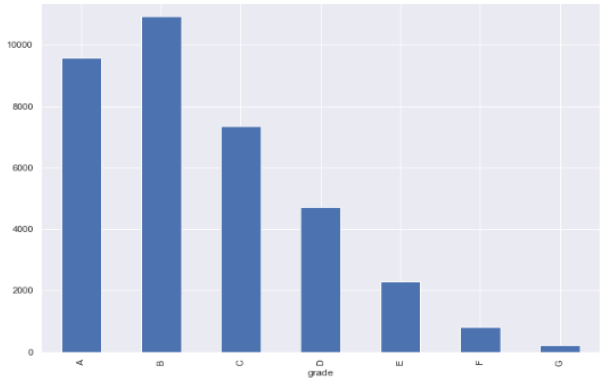
Interest Rate

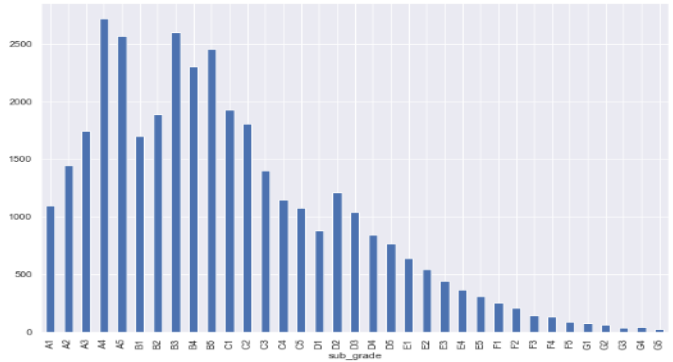
- Distribution for interest rate shows, median interest rate is 11.71%, after 75% percentile interest rate increased from 14.35% to 24.4%.

```
count    38191.000000
mean        11.917876
std          3.684108
min          5.420000
25%          8.940000
50%         11.710000
75%         14.350000
max         24.400000
Name: int_rate, dtype: float64
```

Loan Grade and Subgrade

- Bar plots below indicate that B is the most frequent loan grade followed by A and C while A4 is the most frequent loan sub grade followed by B3 and A5





Employment length

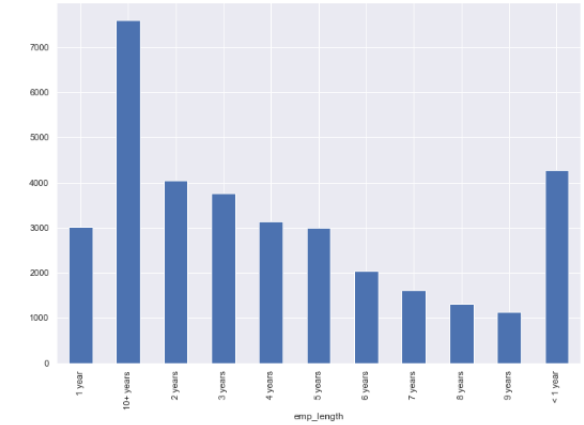
- Bar plot employees with 10+ years of work experience are the most frequent applicants followed by employees with <1 years works experience and 2 years of work experience

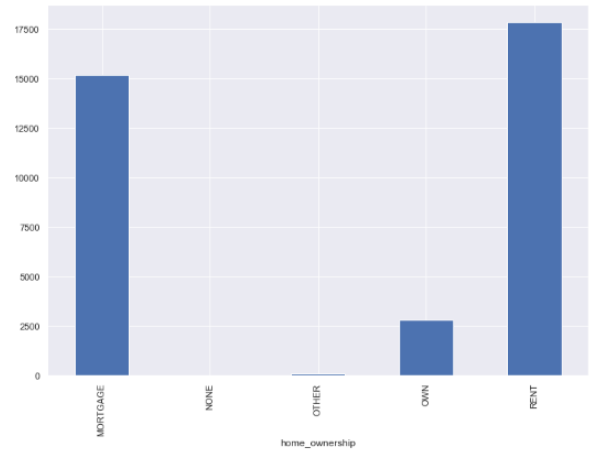
Home ownership

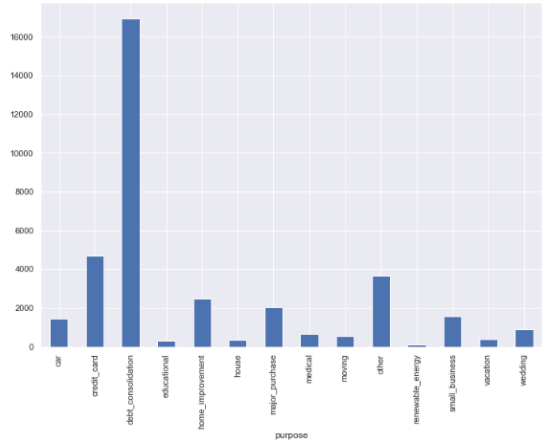
- Plot alongside shows most of the applicants have a rented home followed by mortgaged home and own home categories

Loan Purpose

- Most of the applicants have Debt consolidation as the reason for loan application followed by credit card, other, home improvement







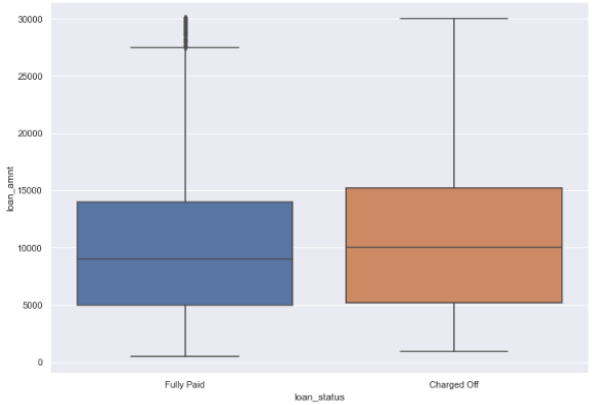
Bivariate Analysis

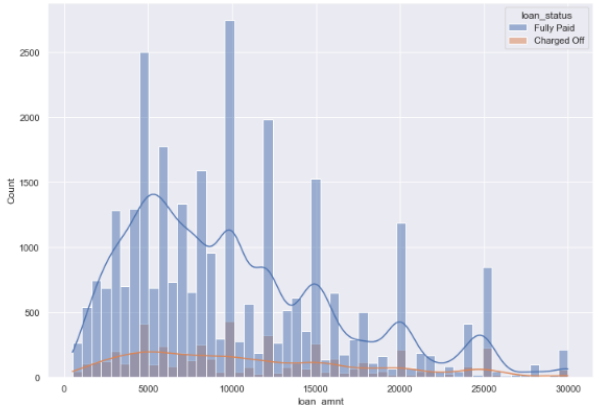
- Lets start the bivariate analysis by plotting a correlation plot between variables
- This plot indicates that Loan amount, investor amount, funding amount, instalment are strongly correlated
- Annual income and dti are negatively correlated
- Employment length and annual income are positively correlated

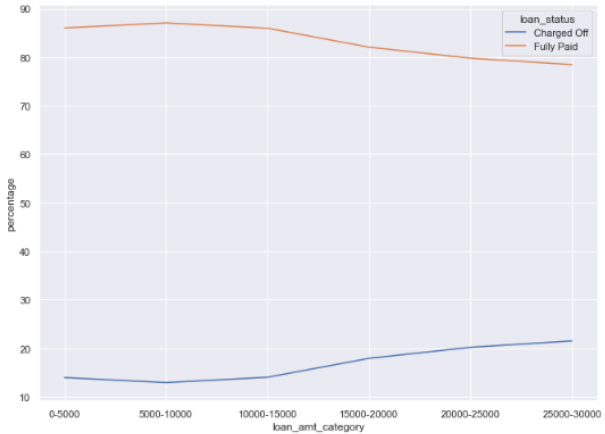
	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	pub_rec_bankruptcies	int_rate_flt	issue_year	issue_month
loan_amnt	1.00	0.98	0.93	0.93	0.38	0.09	-0.03	0.26	0.07	0.04
funded_amnt	0.98	1.00	0.95	0.96	0.38	0.09	-0.03	0.27	0.09	0.03
funded_amnt_inv	0.93	0.95	1.00	0.89	0.36	0.10	-0.03	0.27	0.23	0.06
installment	0.93	0.96	0.89	1.00	0.38	0.08	-0.02	0.25	0.02	0.02
annual_inc	0.38	0.38	0.36	0.38	1.00	-0.07	-0.00	0.03	0.02	0.01
dti	0.09	0.09	0.10	0.08	-0.07	1.00	0.00	0.11	0.09	0.02
pub_rec_bankruptcies	-0.03	-0.03	-0.03	-0.02	-0.00	0.00	1.00	0.09	0.00	-0.02
int_rate_flt	0.26	0.27	0.27	0.25	0.03	0.11	0.09	1.00	0.02	0.02
issue_year	0.07	0.09	0.23	0.02	0.02	0.09	0.00	0.02	1.00	-0.04
issue_month	0.04	0.03	0.06	0.02	0.01	0.02	-0.02	0.02	-0.04	1.00

Loan Amount vs Loan Status

- From the distribution shown below higher the loan amount more likely the borrower will default
- Lets try to analyse for different values how is the distribution of loan getting fully paid or defaulted. Distribution of the same is shown in figure 2
- For further analysis we have created 6 buckets of loan amount, 5000 each starting from 0 to 30000.
- Highest loan amount category i.e. 25000 to 30000 shows highest percentage of default 21.52% followed by second highest loan amount category which shows 20.20% default rate. Overall, default rate goes on increases with loan amount.

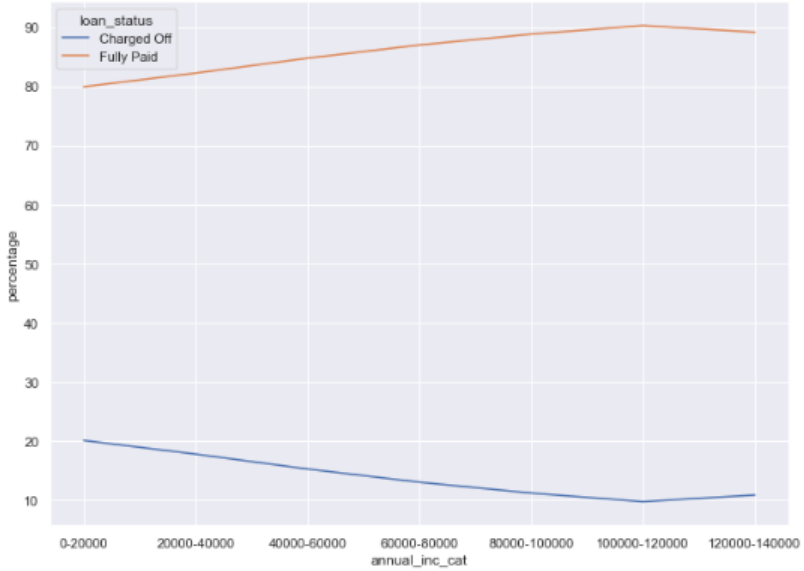






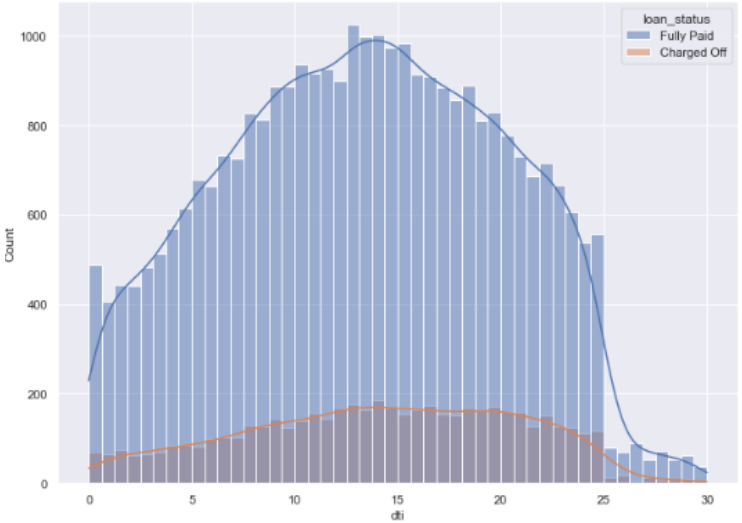
Annual income vs Loan Status

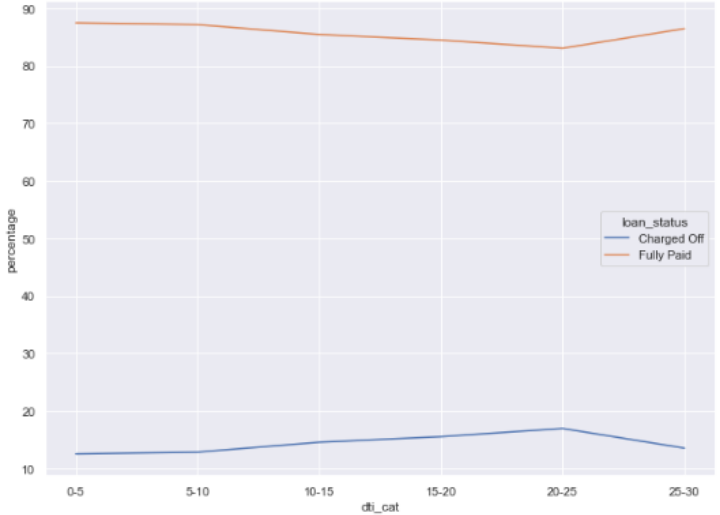
- For this analysis we have created 7 buckets for annual income, 20000 each starting from 0 to 140000.
- Here we observed a completely opposite trend as compared to loan amount vs loan status.
- Category having annual income 0-20000 has highest chances of being Charged off i.e. 20.08% followed by category with annual income 20000-40000 which shows 17.77%
- We can see that annual income between 100000-1200000 has least chances that it will be charged off.



Debt to income ratio vs Loan Status

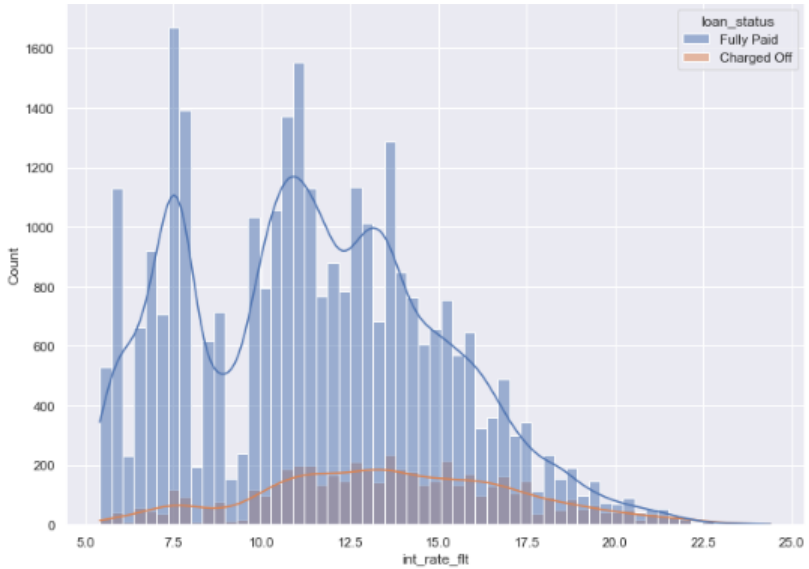
- From the histogram of dti below, when the dti value is low from 0 to 15, the gap between the charged off line and fully paid is increasing. This means that while moving from dti value 0 to 15 there are lower chances that loan will be charged off by borrower. From dti value 15-30 , we clearly see the gap between lines decreasing and borrower is more likely to default the loan.
- For further analysis, we have created 6 buckets for dti, 5 each starting from 0 to 30. Dti category 20-25 shows highest default percentage 16.94% followed by category 15-20 with 15.51% default, category 25-30 with 13.52% default.

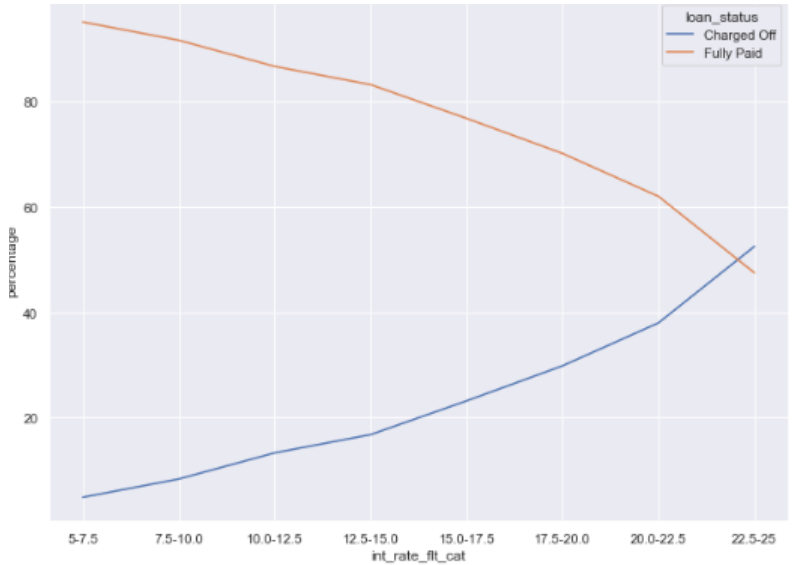




Interest rate vs Loan Status

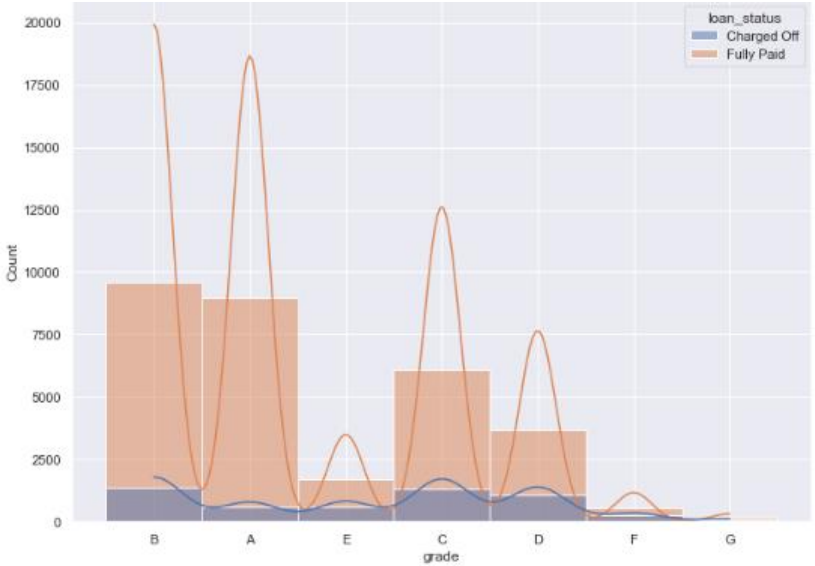
- We have created 8 buckets for interest rate, 2.5% each starting from 5 to 25.
- Percentage of default keeps on increasing with increase in rate of interest. Category 22.5-25 shows highest default percentage i.e. 52.5% followed by category 20.0-22.5 which show 38.01% default rate.

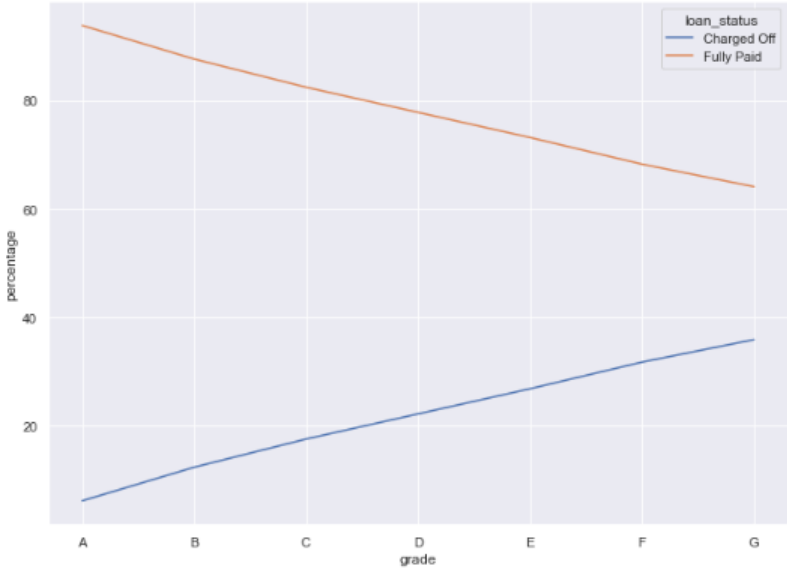




Grade vs loan status

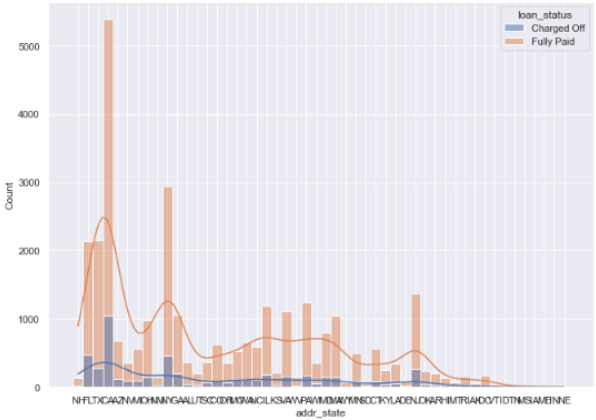
- We have plotted count of borrowers in each grade and percentage of charged off and fully paid borrowers in each grade.
- With each drop in grade we see 4%-5% drop in percentage of people paying the whole loan and 4%-5% increase in percentage of default.
- Grade is a key factor in approving the loans.

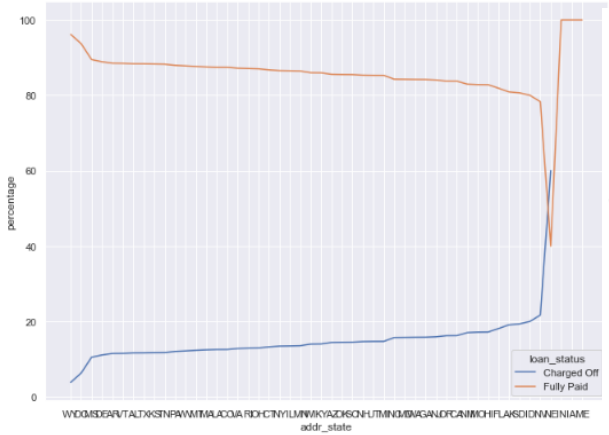


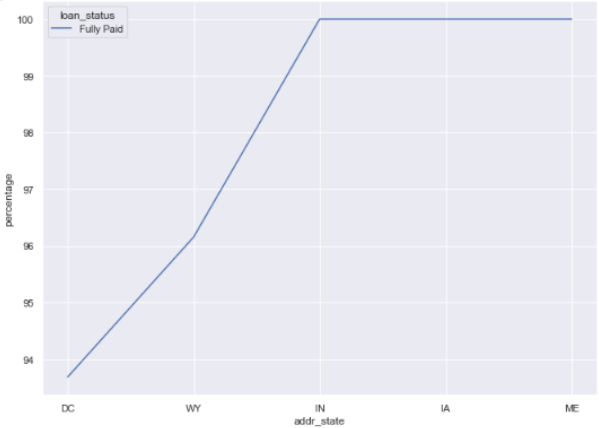


Address State vs loan status

- We have plotted count of loans for each address state. Also, we have plotted charged off and fully paid percentage distribution of loans for address states
- From plots, The loans in state IN, IA and ME are fully paid, there are no defaulters in these states, but the number of loans taken in these states is very less. (fig 2 below)
- If we plot the number of states have more than 90% of loan which was fully paid by borrowers, we see that only DC, WY, IN, IA and ME qualify this criteria (fig 3 below)

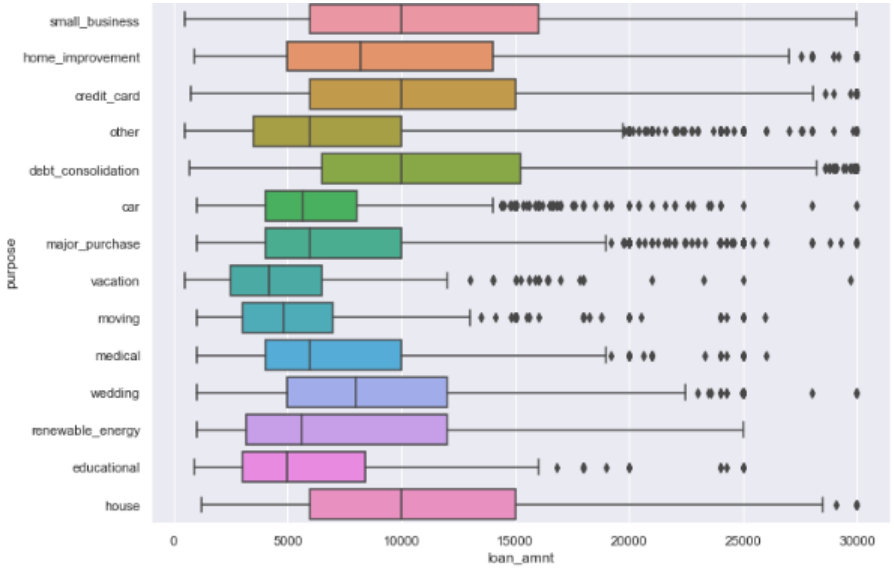






Loan amount vs Purpose

- From the plot below, Median,95th percentile,75th percentile of loan amount are highest for loan taken for small business purpose among all purposes. Debt consolidation is second and Credit card comes 3rd



Recommendations

1. LC should reduce the rate of interest for 60 months tenure to borrowers as the chance for them to default is more.
2. Grade is a good attribute in detecting the defaulters. LC should be more careful for borrowers having grade G to A.
3. LC should avoid giving loans to states ID, NV and NE.
4. Loan for the purpose of small businesses are likely to default the loan. LC should reduce approving loan to them.
5. LC should look into the annual income of the borrower. Borrowers with less annual income are likely to default the loan.

THE END