LENDING CLUB CASE STUDY

Group Facilitator: Abhinav Singh

Team Member: Bharathraj Saravanan

Lending Club Case Study - Flow Chart

The main aim is 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.

01

Data Cleaning

Dropped rows which have loan status = Current

Dropped columns which are -
1. Redundant,
2. Having more null values,
3. Customer behaviour variables

02

Data Analysis

Done some -
1. Univariate Analysis
2. Segmented Univariate Analysis
3. Some Bivariate Analysis and added few plots in the upcoming slides

03

Recommendations

Added some driving factors (or driving variables) and how they can reduce the chances of funding a likely defaulter.

Data Cleaning:

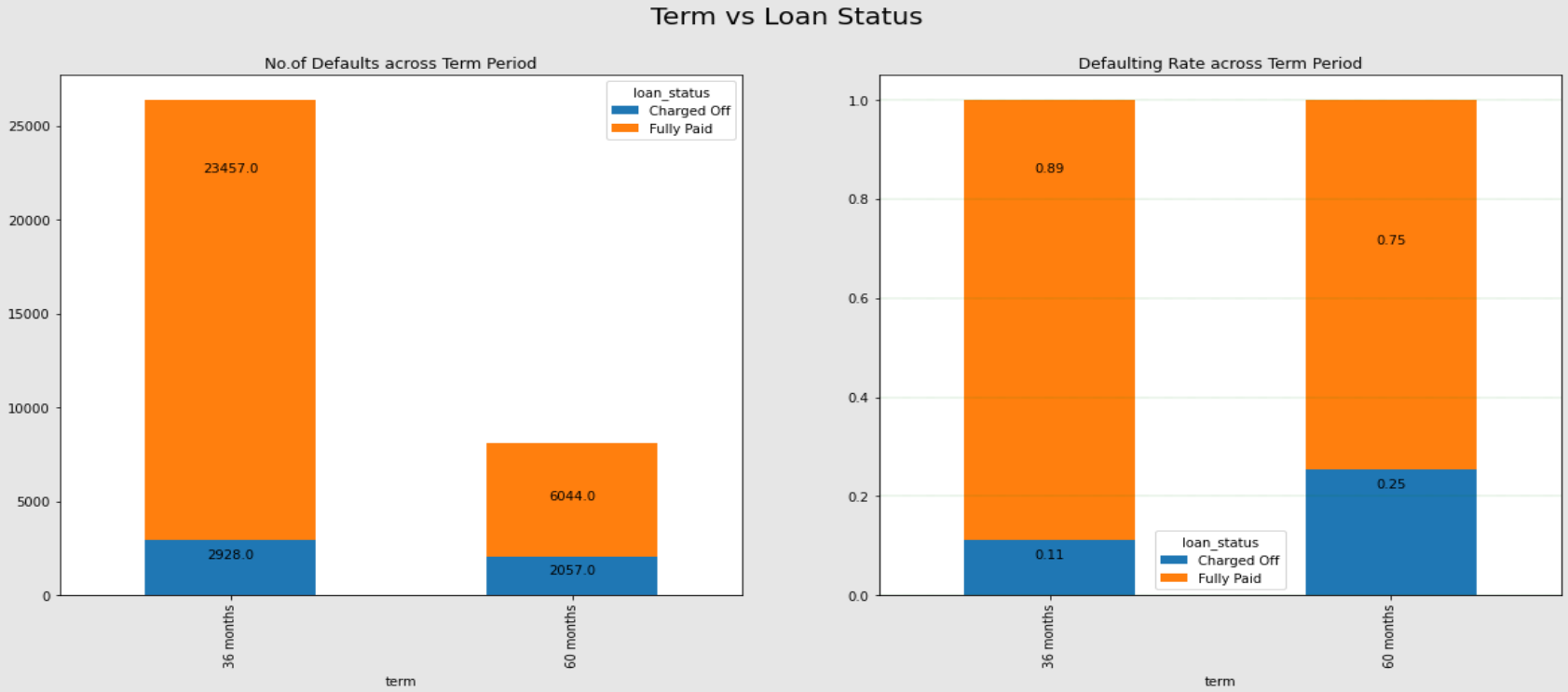
- Loaded given loan data.
- Dropped rows which have loan status = 'Current', as it is neither fully paid nor defaulted, so getting rid of the current loans.
- Dropped columns which are,
  - Redundant - columns which are not needed for defaulting analysis.
  - Having more null values – columns which have more than 60% null values.
  - Customer behaviour variables - are not available at the time of loan application, and thus they cannot be used as predictors for credit approval. So dropped it.

Data Analysis:

- After data cleaning, we got 33 columns which need to be analyzed.
- Started with Univariate and Segmented Univariate analysis simultaneously with the target variable loan_status.
- Then done some Bivariate analysis and given our final recommendations as conclusion

Plots: Added some plots for the deriving factors we found.

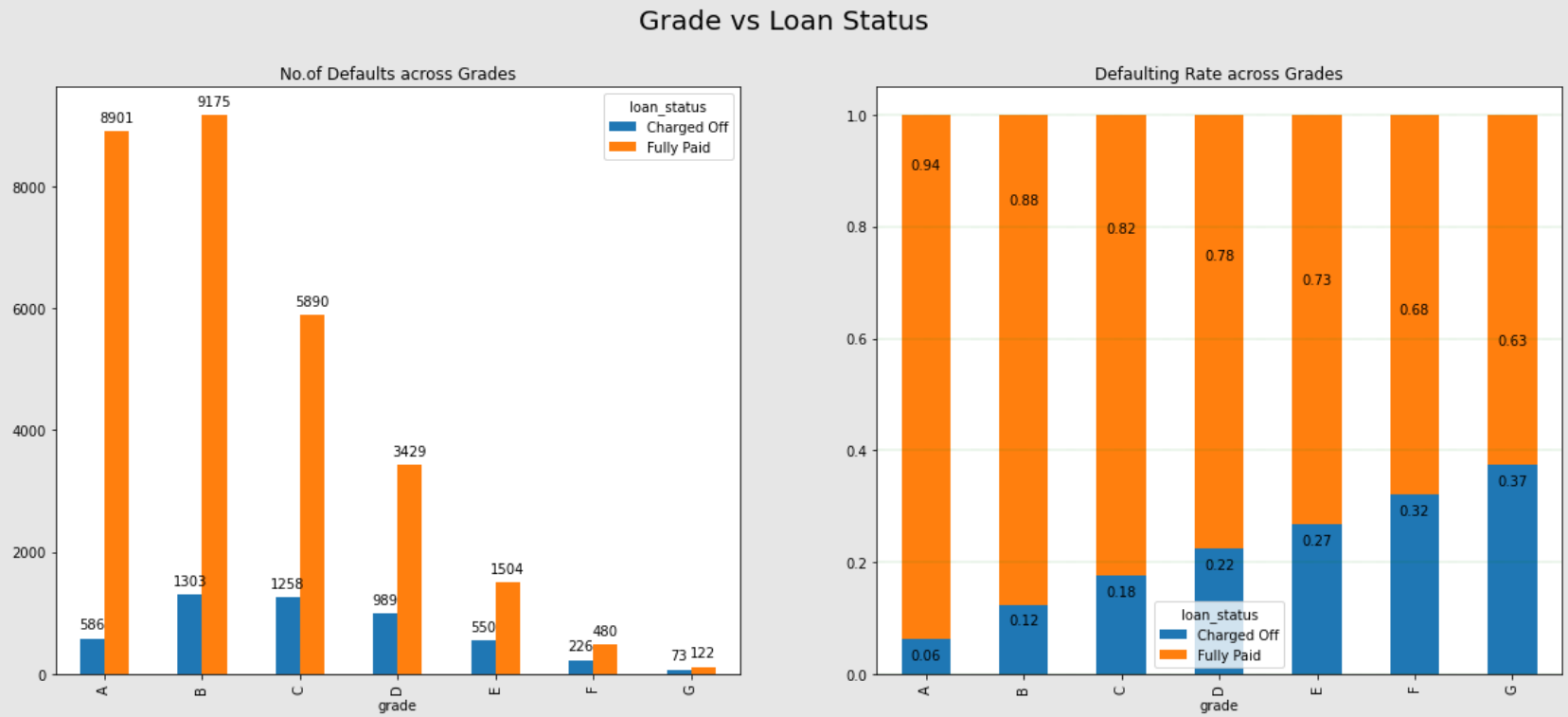
1. Term



- When comparing default rate across the terms, we can see more no .of loans are defaulted in 36 months term while comparing with 60 months term loans.
- But when seeing in percentage we can clearly see that, 25% of 60 months term loans are defaulting. On the other hand, 11% of 36 months term loans are defaulting.

Recommendation: So 36 months term loans are preferrable.

2. Grade - LC assigned loan grade. Possible values are – A - G

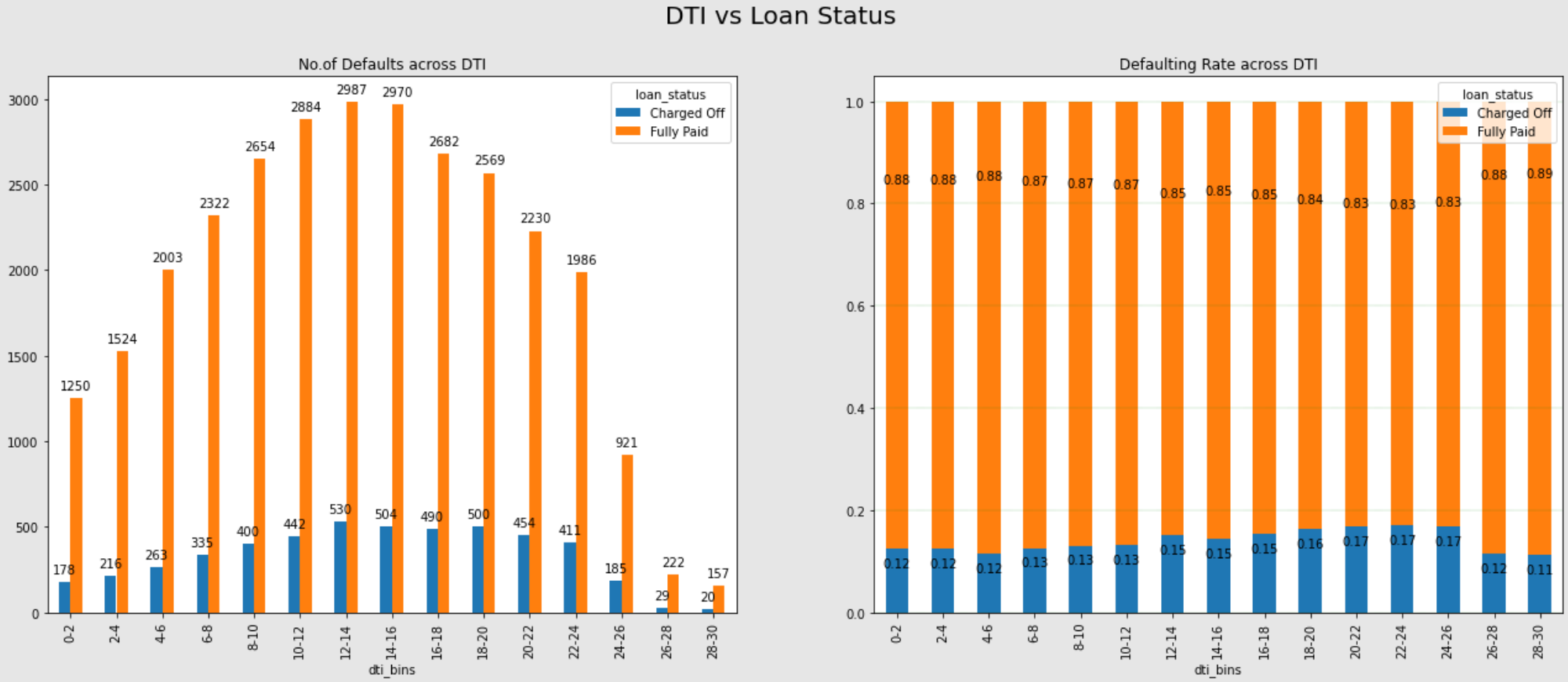


- The amount of data is decreasing as the grade increases alphabetically. And even though the data is smaller in higher alphabets, the defaulting rate is higher.
- So we can clearly say that, Higher the grade, Higher the risk of defaulting.

Recommendation: Lower the grade, Lower the risk of defaulting. Grade A is preferrable.

We are not going to show subgrade as both of them are correlated and without more classifications grade is giving us the same results as subgrade. So ignored it.

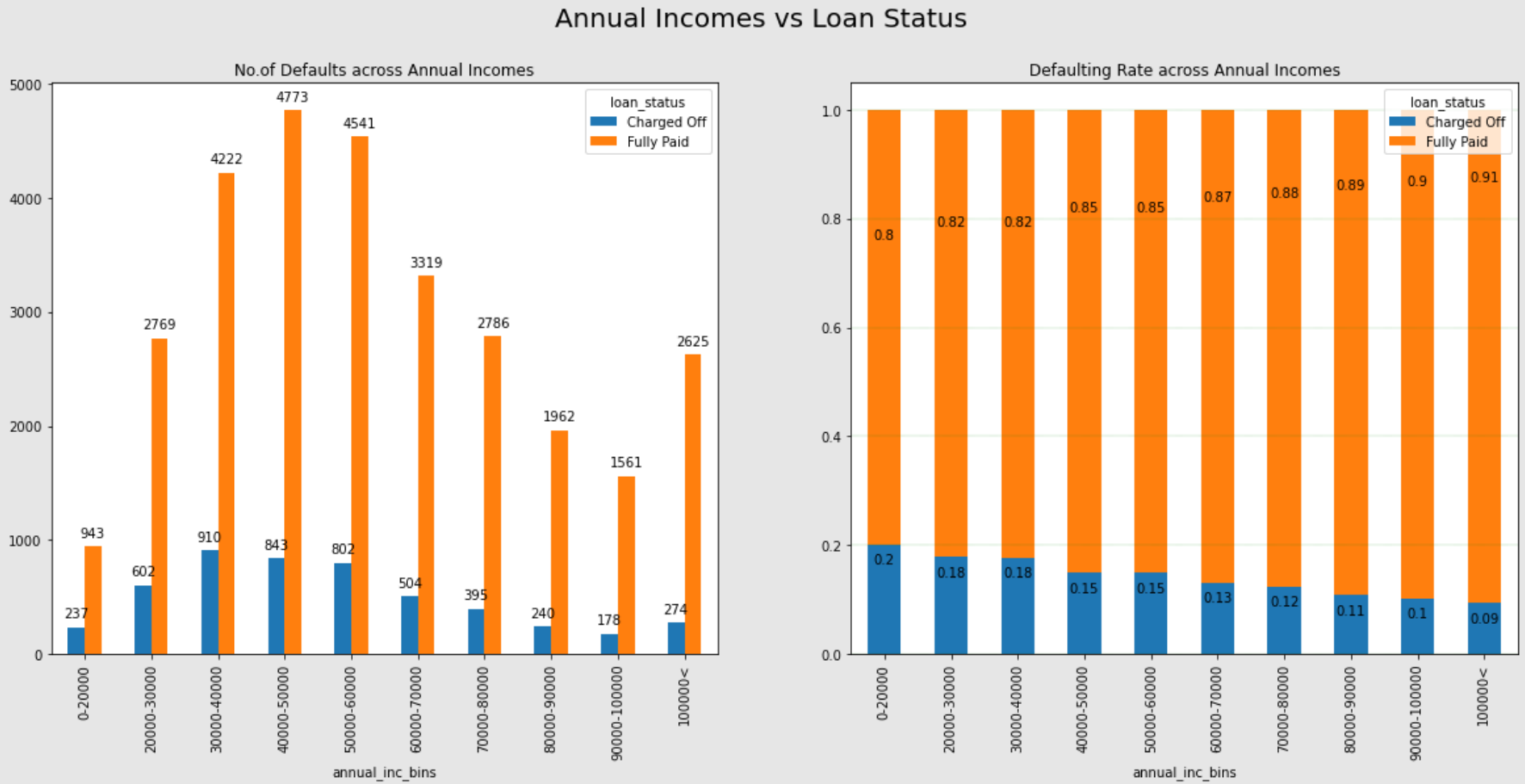
3. DTI - A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.



- From the two plots, we can see the default rate increases as the dti increases.
- The last two bins not following the pattern since we don't have enough data in it.

Recommendation: Higher the dti, Higher the risk of defaulting. So better to go with lower dti.

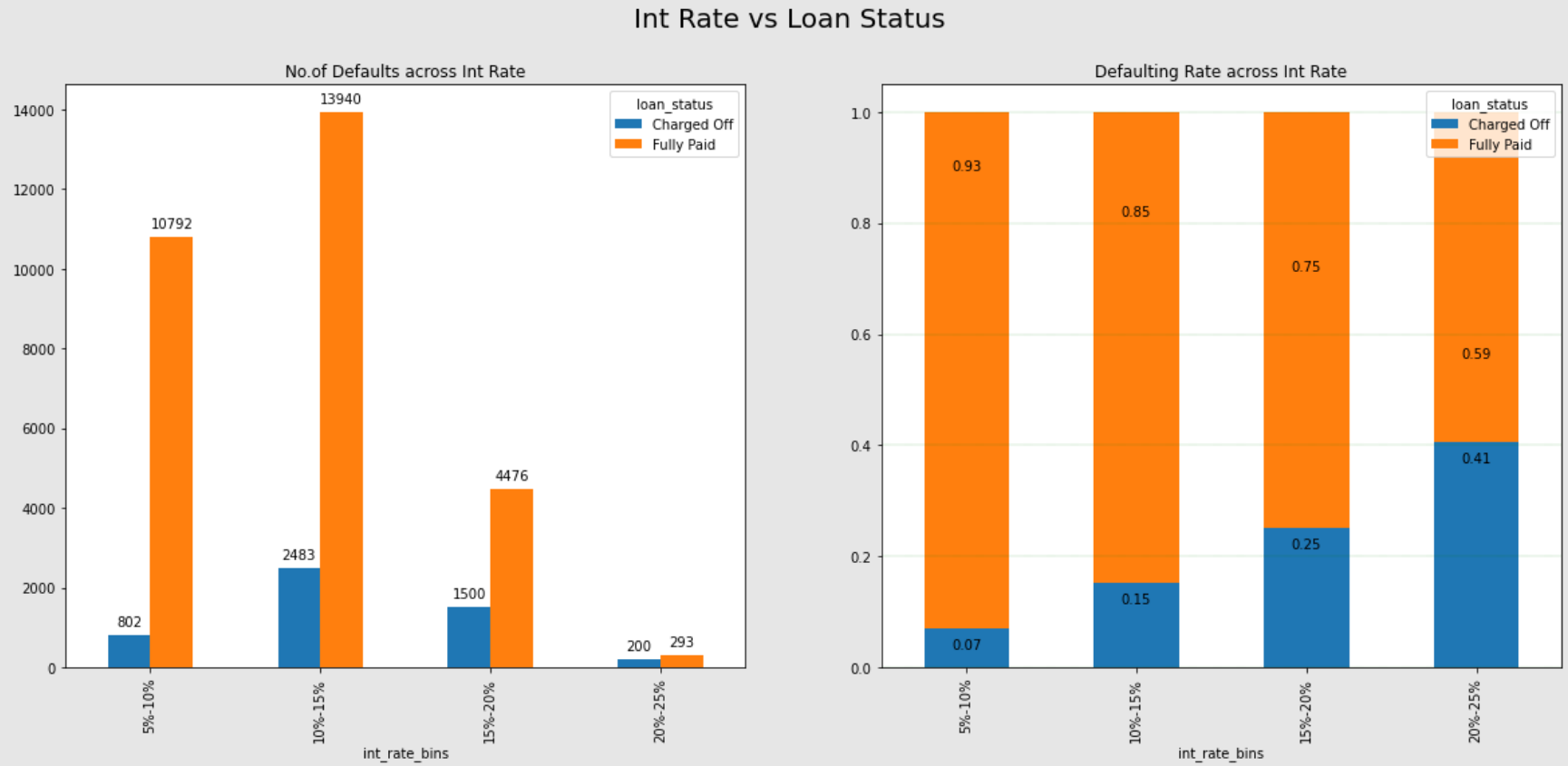
4. Annual Income - The self-reported annual income provided by the borrower during registration.



❑ We can clearly see the pattern, As the income increases, the default rate decreases.

Recommendation: Higher the annual income, lower the risk of defaulting.

5. Int Rate - Interest Rate on the loan.

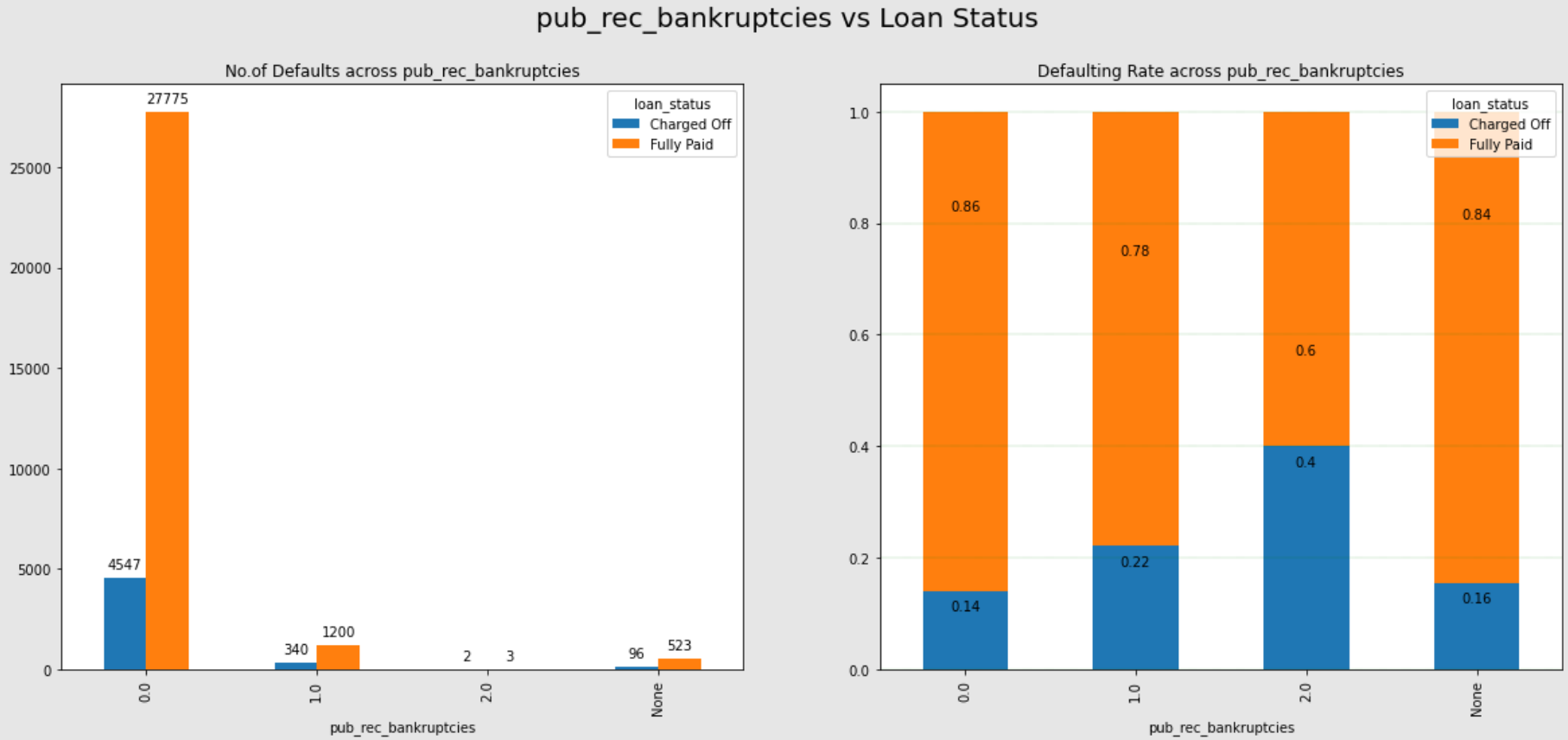


❑ The amount of data is decreasing as the int rate percentage increases. And even though the data is smaller in higher percentages, the defaulting rate is higher.

❑ So we can clearly say that, Higher the int_rate, Higher the risk of defaulting.

Recommendation: Lower the int rate, Lower the risk of defaulting. **5 to 10% int rate is preferrable.**

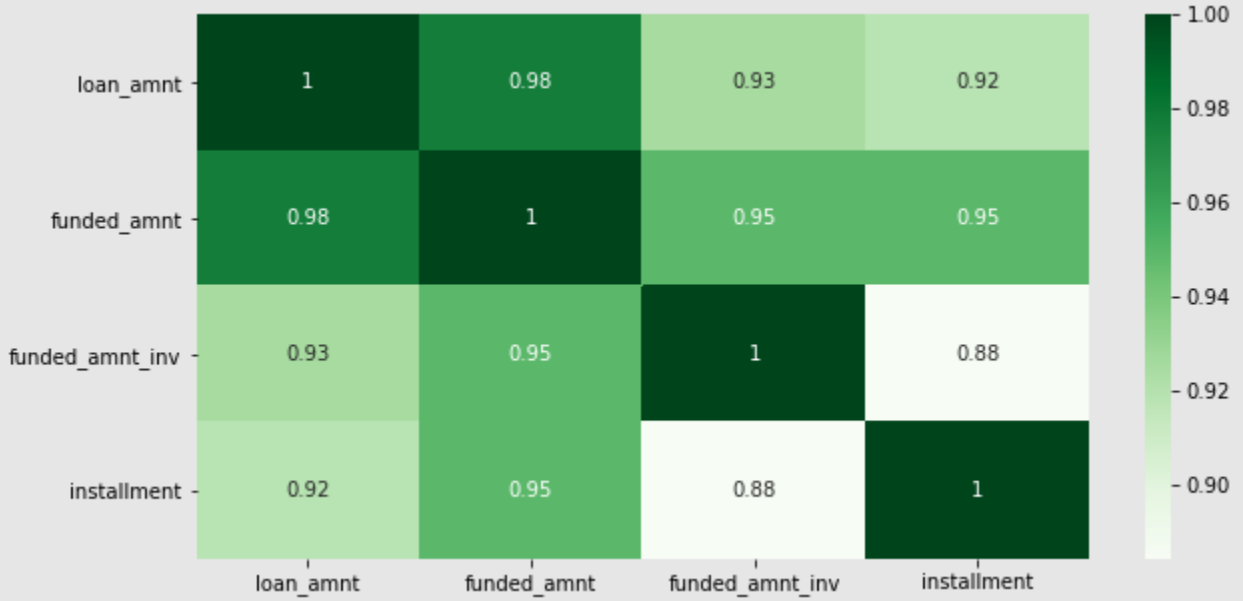
6. pub_rec_bankruptcies - Number of public record bankruptcies.



- We don't have enough data in 2.0 category but even though, out of 5; 2 of them is defaulted.
- So we can clearly see when bankruptcies increases, default rate increases. It's better to get data in some way if it's unknown.

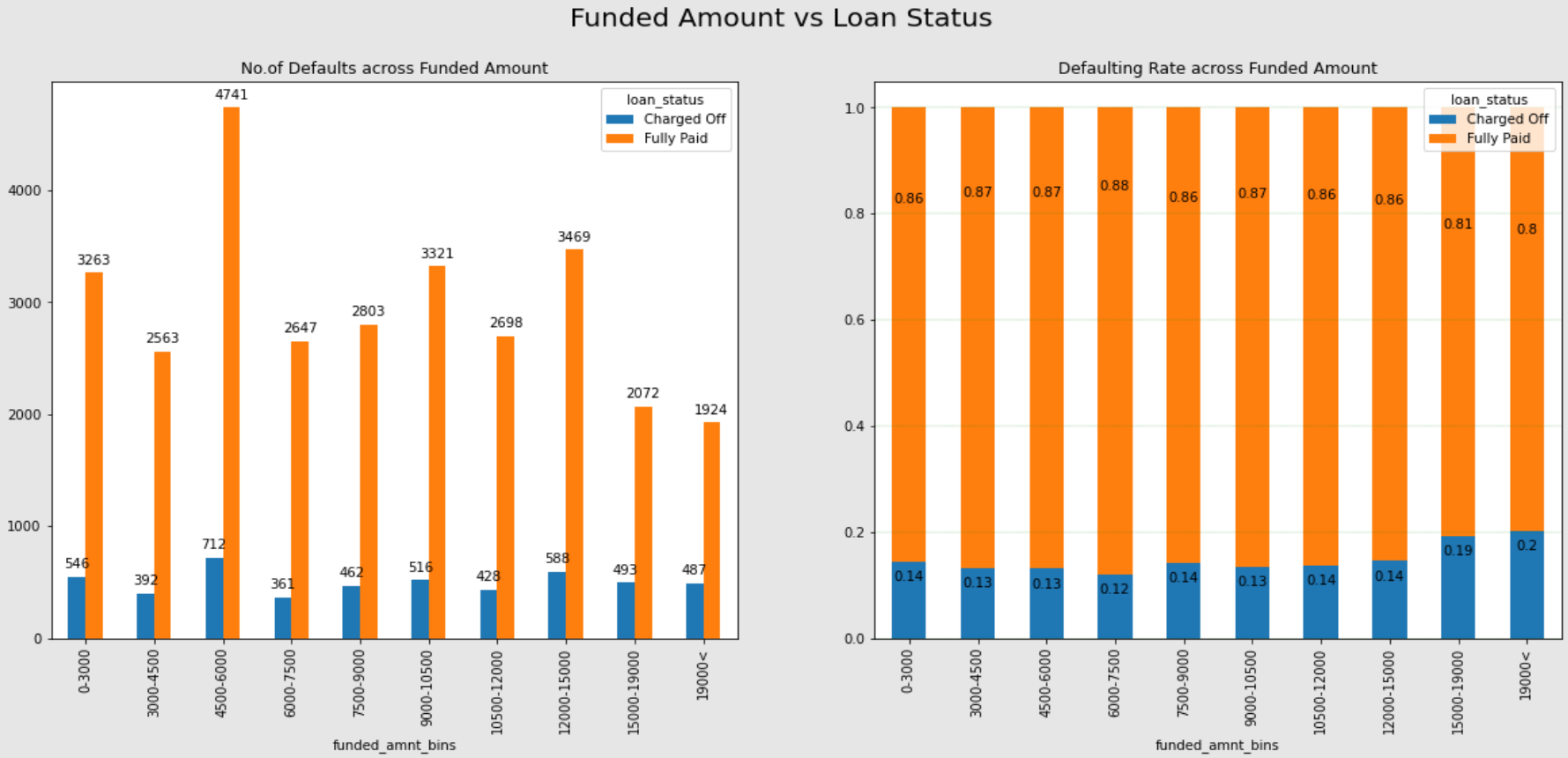
Recommendation: Higher the No. of bankruptcies, Higher the risk of defaulting.

7-10. loan_amt, funded_amnt, funded_amnt_inv, installment are highly correlated.



- From the Heat Map, We can see that **funded_amnt** was mostly correlated with all the other columns. So taking funded_amnt and analyzing.

10. funded_amnt – The total amount committed to that loan at that point in time.



- When we see the final plots, we can see that either the lower range or the upper range are having higher default rate. The in-between values are having lower default rate.

Recommendation : So, its better to ignore the two tails, like <4000 and >15000. As they have higher default rate.

Conclusion:

As a final results a quick wrap up, The driving variables behind loan default are

❖First-level driving factors:

- ❑ **Term** - 36 months term loans are preferrable than 60 months term loans.
- ❑ **Grade** - Lower the grade, Lower the risk of defaulting. Grade A is preferrable.
- ❑ **DTI** - Higher the dti, Higher the risk of defaulting. So better to go with lower dti.
- ❑ **Annual Income** - Higher the annual income, lower the risk of defaulting.

❖Second-level driving factors:

- ❑ **Int Rate** - Lower the int rate, Lower the risk of defaulting. 5 to 10% int rate is preferrable.
- ❑ **pub_rec_bankruptcies** - Higher the No of bankruptcies, Higher the risk of defaulting.
- ❑ **loan_amt, funded_amnt, funded_amnt_inv, installment** are highly correlated. As funded_amnt is mostly correlated with others, while considering funded_amnt and analyzing,
  - ❑ its better to ignore the two tails, like <4000 and >15000. As they have higher default rate.