# Lending club case study

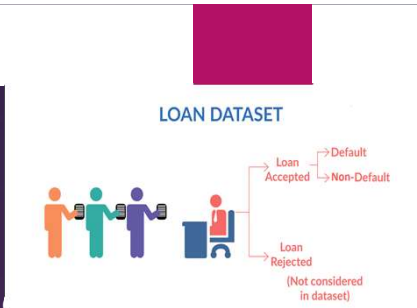
Krishna Pal

# Problem Statement

To analyze the loan application data for lendingclub.com to derive meaningful insights which can help lenders/LC to minimize the risk of losing business opportunity due to excessive rejection or financial loss due to bad loan (not likely to repay the loan)

The Lending club company wanted to understand all possible driving factors behind the loan defaults, i.e. which features are strong indication of loan defaults. So that they can update the rule engine to reject such application, or reduce the loan amount or offer loan at higher interest rate.

# Data understanding

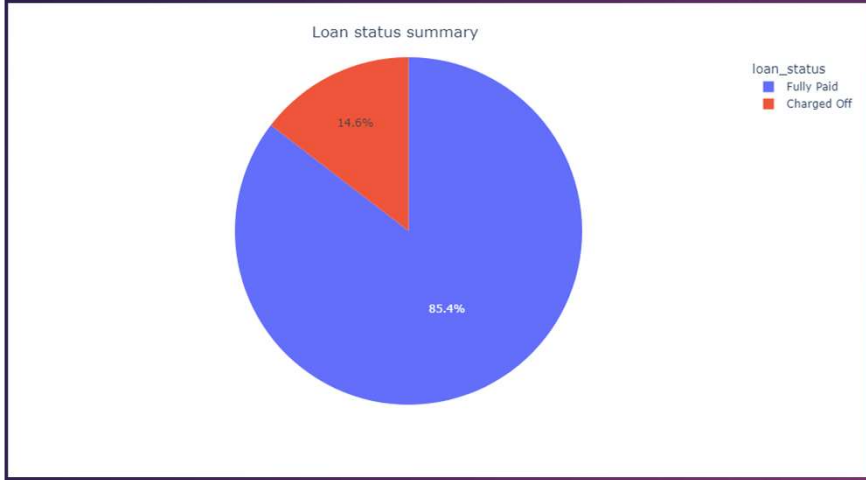



- There are total 39717 records with 111 feature variable present in dat
- The dataset contains data only for the accepted loan application from past.
- There are 54 columns which have fill rate 0%, means null values in all rows, hence we have remove these columns from analysis.
- There are 3 columns (mths_since_last_delinq, mths_since_last_record and next_pymnt_d) having 64, 92 and 97% missing values respectively, so drop them.
- There are 24 features variable(Appendix A) which we can say customer behavioral data and are not available during loan application time, so not suitable for our analysis, hence we can drop them from our analysis.

# Data understanding cont…

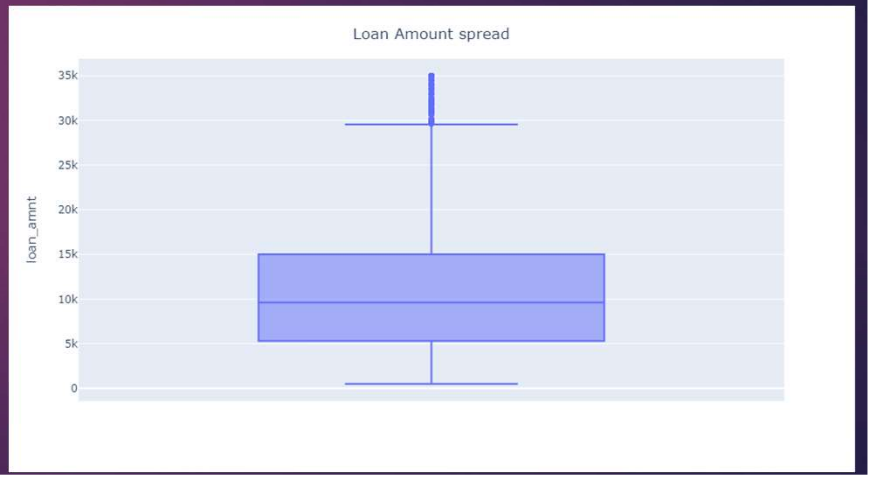
- Few features have unique values across all rows hence not suitable for analysis so we can remove them too. ( refer Appendix A )
- There are 3 types of loan status we have in dataset. Fully paid, charged off( defaulter) and current.
- There are 1140 rows with loan status = ‘current’ which can not be used to make any business decision as they are still ongoing. Hence we can drop them from our further analysis.

# Univariate Analysis

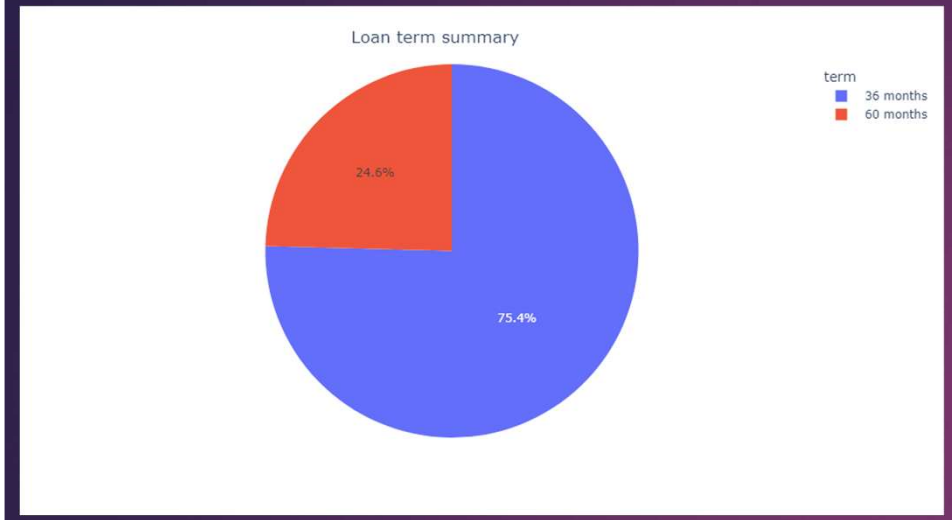



- ✓ There are 85% loan with status = fully paid
- ✓ Remaining 14.6% are with status = Charged off

- ✓ The loan amount is not normally distribution. The majority of population falls b/w 5k-15k
- ✓ The 1st quantile(q1) is 5.3K, 2nd quantile(q2) or median is 9.6k and the 3rd quantile(q3 is 15k.
- ✓ The box plot shows that there are some outliers present at upper whisker side.

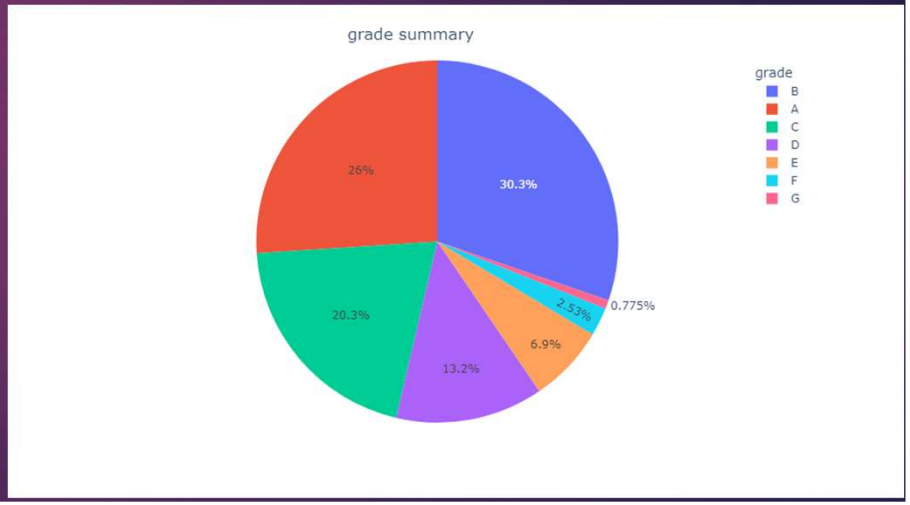


# Univariate Analysis cont..

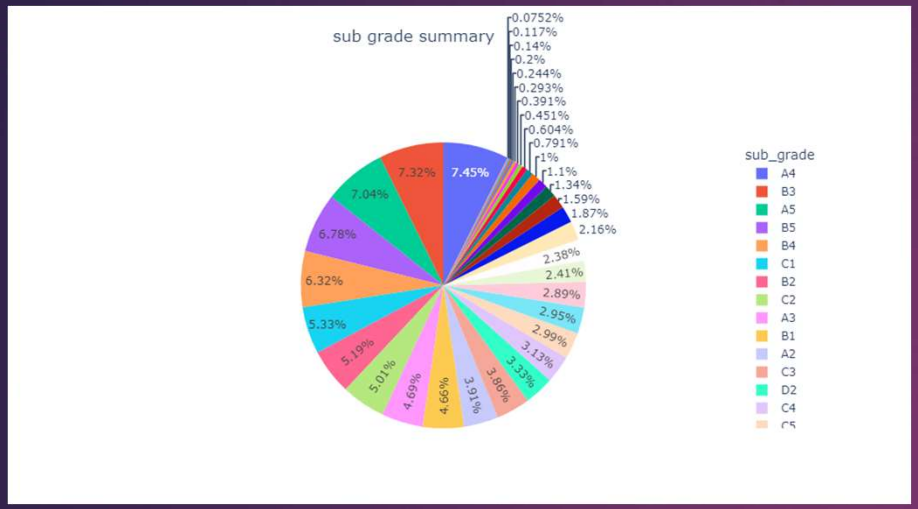



- ✓ Around 75% of borrowers opted for loan term 36 months.
- ✓ Remaining 25% of borrowers opted longer loan term i.e. 60 months.

- ✓ Most borrowers have assigned either A,B, C grade.
- ✓ Few of borrowers also been assigned E,F,G grade as well.

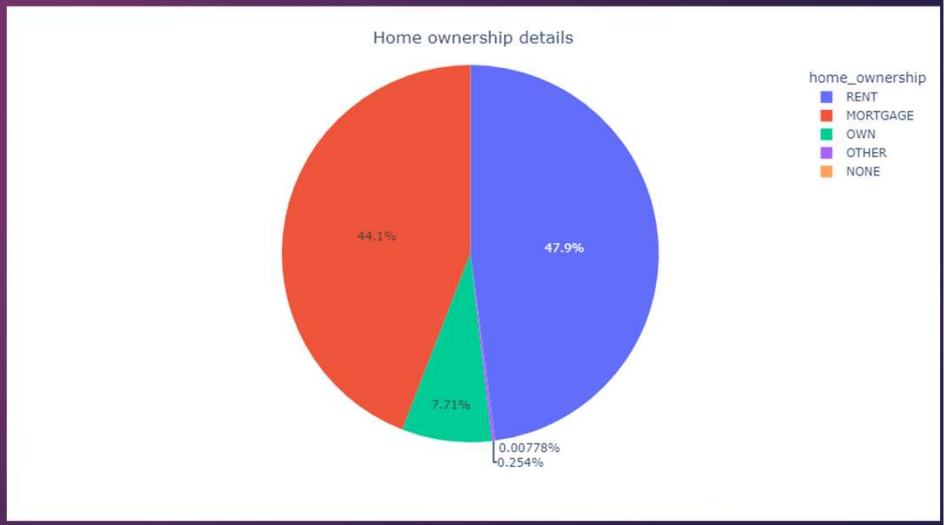


# Univariate Analysis cont..




- ✓ The grades have further extended to subgrades and it is based on risk score.

- ✓ 47% borrowers lives in rented property.
- ✓ 44% borrowers lives in mortgage property.
- ✓ Very few 7% borrowers have their own house.
- ✓ Looks people who lives in rented accommodation have higher tendency of taking loan.

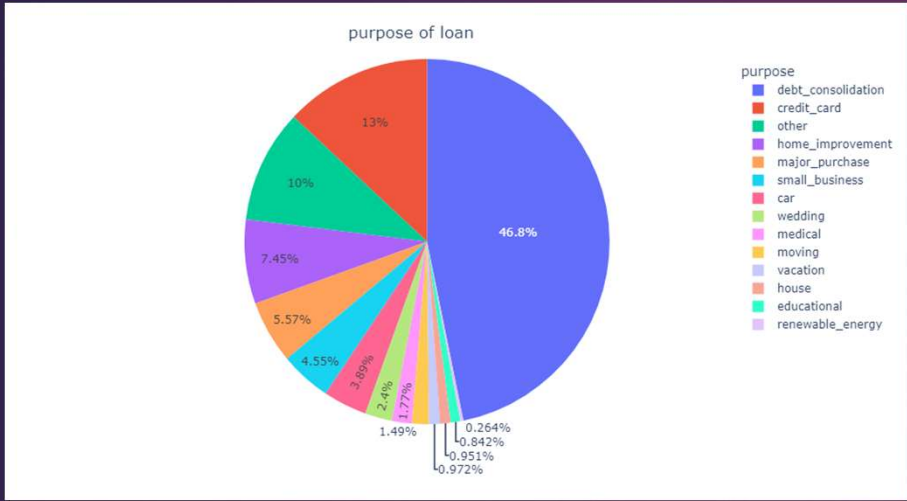


# Univariate Analysis cont..


- ✓ Median of income is 58K
- ✓ q1= 40K, q3 = 82k
- ✓ Upper whisker = 145K, lower whisker = 4k
- ✓ The income range seems having outliers at upper range side.

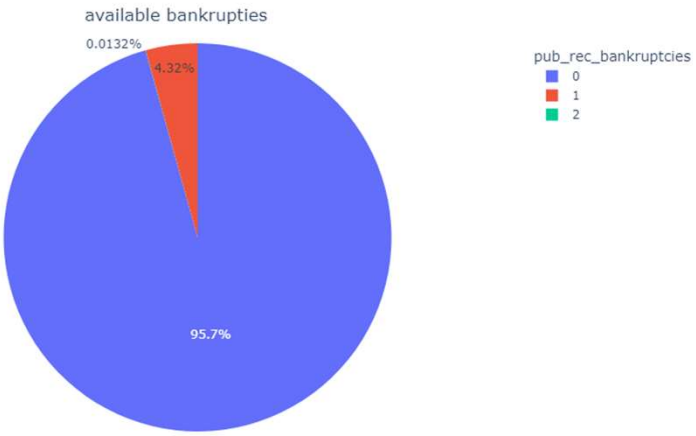
- ✓ Around 55% population have verified source of income
- ✓ Remaining population have unverified source of income.

# Univariate Analysis cont..



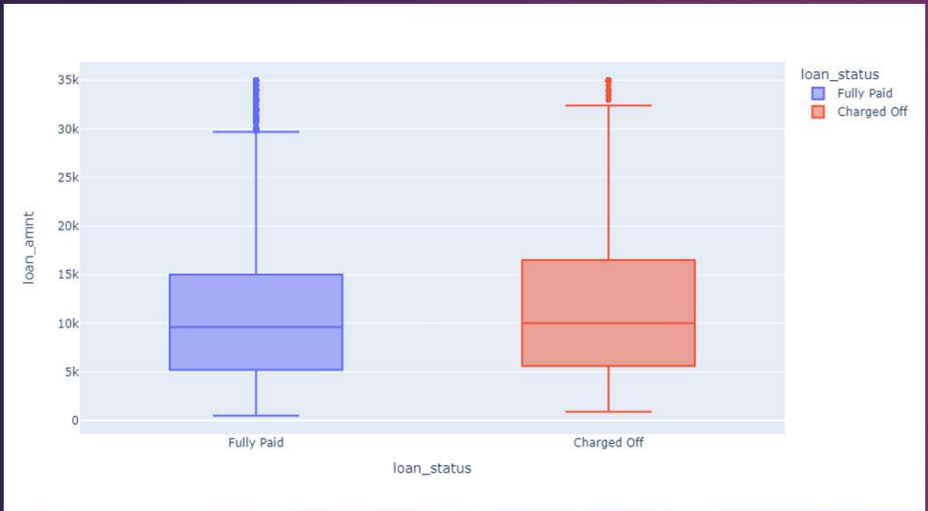

- ✓ Around 46.8% borrowers have taken loan for debt consolidation, while 13% have taken for credit card.
- ✓ Remaining have taken for other purpose.

- ✓ Around 95% borrowers have no public record for bankruptcies.
- ✓ 4% borrowers have 1 public record for bankruptcies.
- ✓ .01% borrowers have 2 public records for bankruptcies.



# Bivariate Analysis

## Loan Amount vs Loan_status





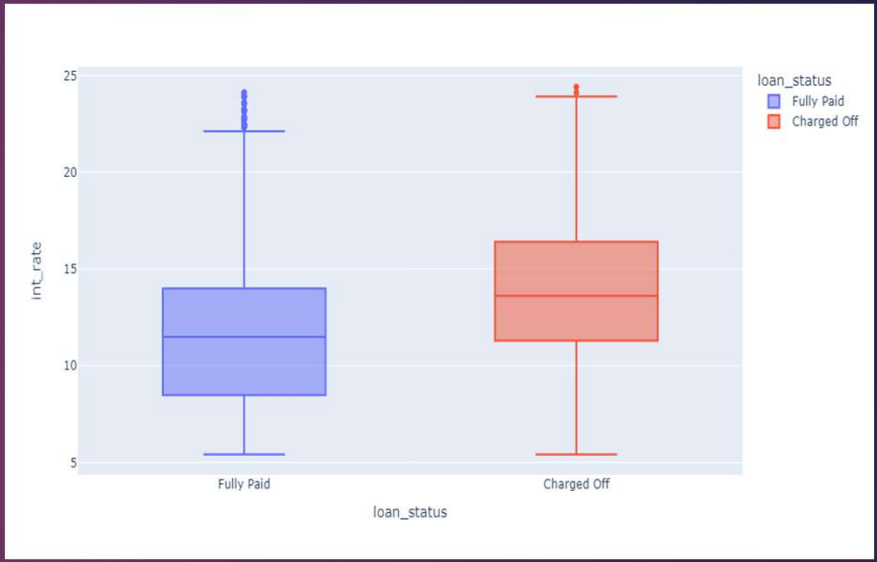
- ✓ The average loan amount for Charged off cases is slightly higher than fully paid cases

# Bivariate Analysis cont..

## Term vs Loan_status


Observation : The chances of getting default is 2 times higher for term 60 months as compare to 36 months.

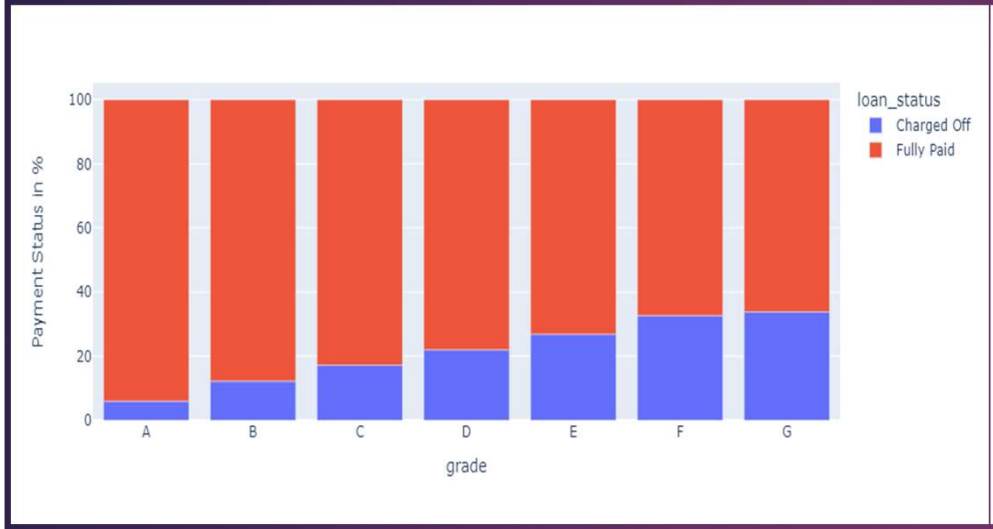
## Int_rate vs Loan_status



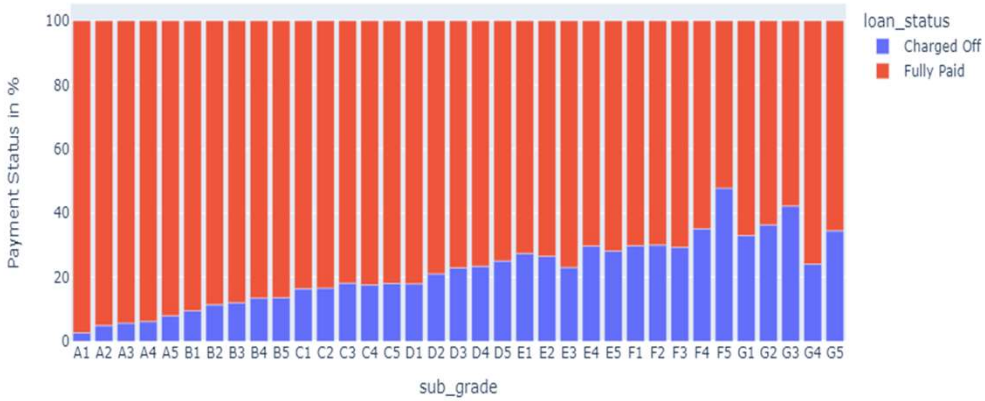

Observation: The loan with higher interest rates have more tendency to go charged off .

# Bivariate Analysis cont..

## Grade vs Loan_status




## Sub_Grade vs Loan_status

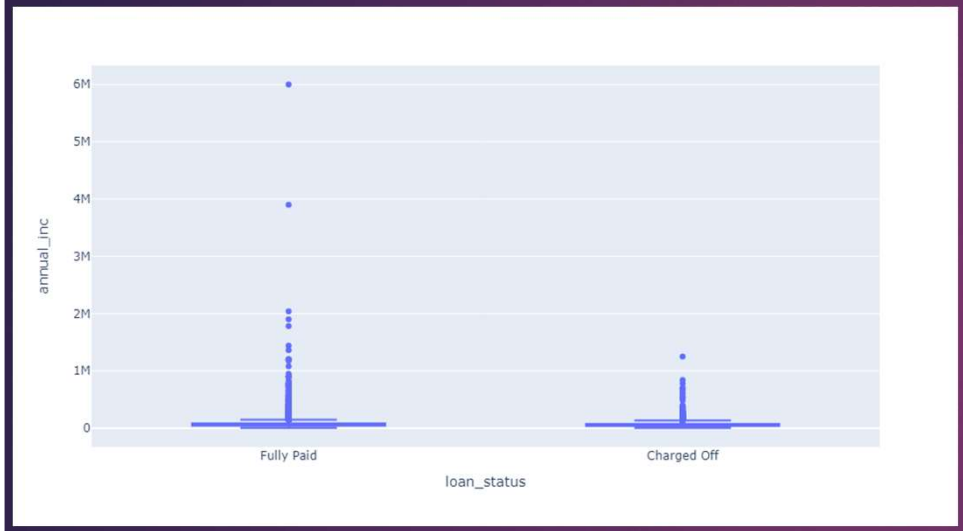



Observation : From above grade and sub_grade graph it is clear that the grade E,F,G and associated subgrade have higher chance to become Charged off ( defaulters)

Recommendation : if the loan applicants comes under grade E,F, G and associated subgrades more scrutiny is required.

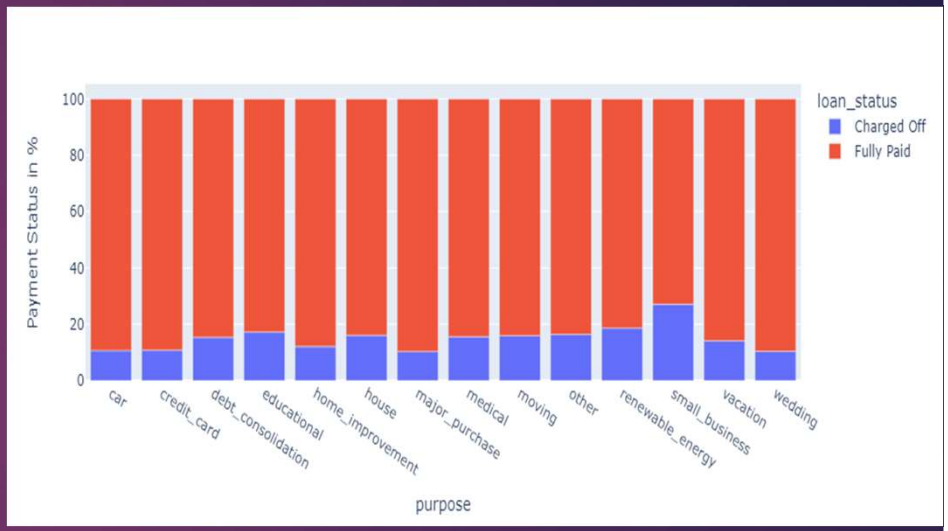
# Bivariate Analysis cont..

## Income vs Loan_status




Observation : The median annual for fully paid category is 60k while for charged off it is 53k. That means the people having low income have higher tendency of getting charged off.

## purpose vs Loan_status

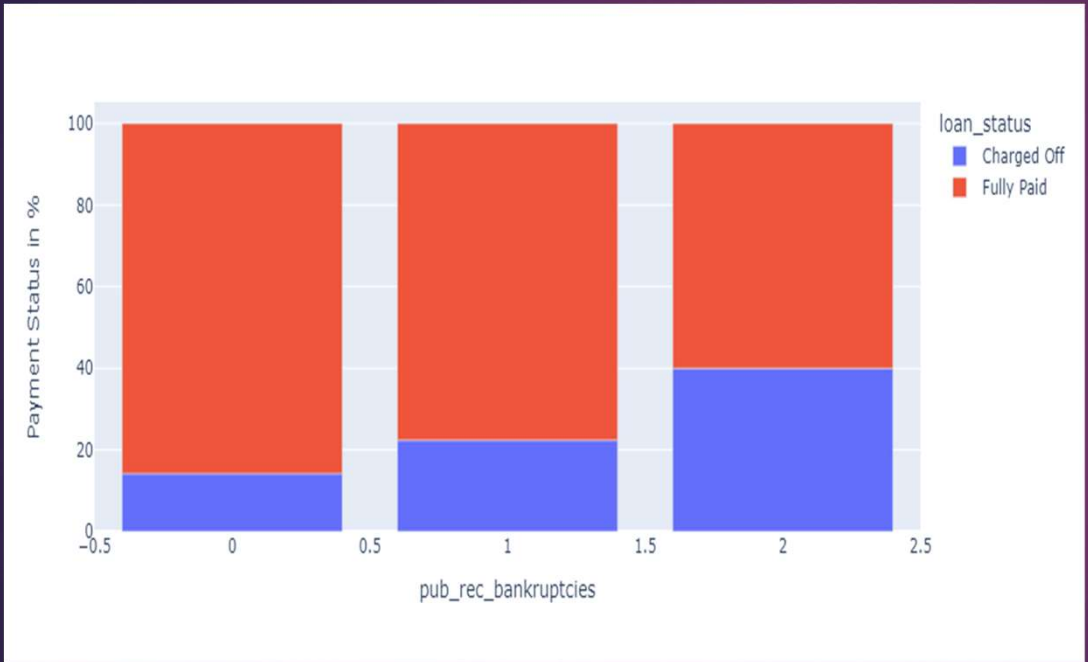



Observation: The chances of getting charged off is 27% for the purpose “small_business”.

Recommendation: more scrutiny is required for risky loan purpose.

# Bivariate Analysis cont..

## Income vs Loan_status




Observation : The chances of getting charged is very high for the applicants having at least one public bankruptcies records.

Recommendation: Do not offer loan to applicants having public bankruptcies record more than or equal to 2 as the % of getting charged off is around 40%.

# Conclusion & Recommendations:

After Analyzing the lending club dataset we can inferred below insights.

- Below features are clearly driving factors that impacts loan repayment certainly.
  - ✓ Loan term
  - ✓ Grade & sub-Grade
  - ✓ Rate of interest
  - ✓ Income & purpose of loan
  - ✓ Public bankruptcies records

- Below features have visible indication of associated risks so must be utilized for risk scoring, if already happening then logic to be reviewed and updated accordingly.
  - ✓ Grade & sub-grade
  - ✓ Public bankruptcies records

# Appendix

Appendix A.txt

Thank You!!