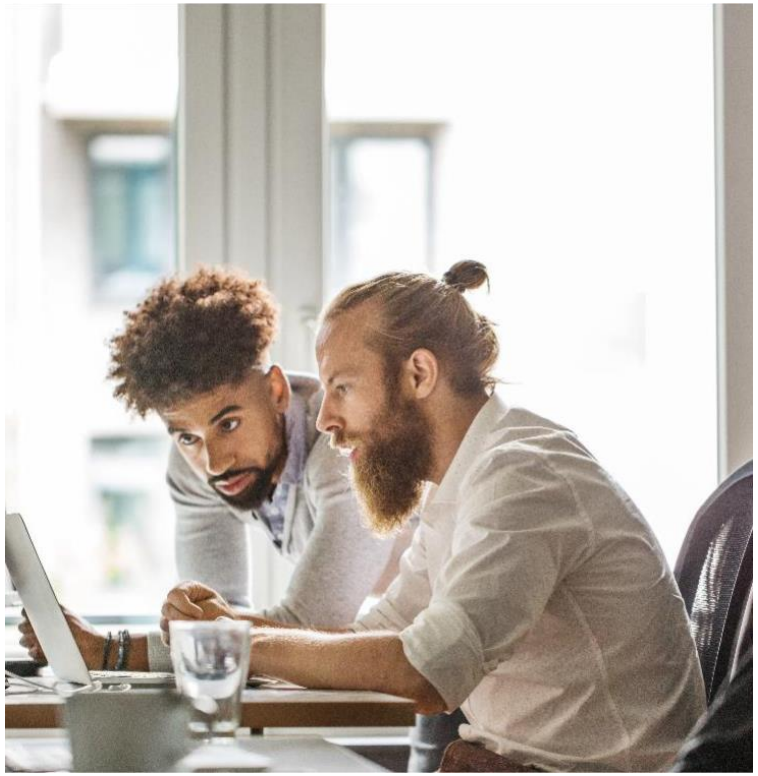
Lending Club Case study

Course 2 Module 4

May 2022

Pooja Surve-Mahadik & Nikhil Nan

Agenda



Project Overview

Analysis Approach

Analysis and Findings

Results Dsicussion

Recommendations and Conclusions

Project Overview

Objective

Lending Club is an online market place company facilitating personal loans, business loans, and financing of medical procedures. Objective of the project is to **analyse** the organization's **historic loan data** and **identify patterns and key drivers that will enable the organization to take actions** such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc. to a new applicant.

Analysis Approach

The project will progress as follows:

- Understanding of the dataset
- Data Cleansing and Preparation
- Univariate Analysis
- Bivariate Analysis
- Discussion of Observations
- Recommendations

Understanding of the dataset

1

Data Familiarization and Data Preparation

- **Data Description**

The data set contains 5 years loan data of the company for the period 2007 to 2011. It contained over 110 attributes of the data such as the loan amount details, applicant details, sanction status, repayment status etc.

- **Null value & Unique value fields**

After the initial screening, it has been found that the over 50 attributes were empty attributes with null values. These 54 columns were dropped from the data set.

Further there were around 10 columns which had some portion of null values. These attributes were prepared by techniques such as mean/median/mode value replacement.

Additionally, the columns that contained unique values (same values in all the cells of the column) were dropped as it is not adding any value to the analysis.

- **Data type conversion**

By analyzing the data, appropriate data conversions were done such as converting interest rate column from text to float.

- **Derived Columns**

Columns were derived based on the existing columns such as bucketizing loan amount & income, deriving date, month and year from date columns etc. to add more dimensions to data for deriving insights.

- **Summary**

After data cleaning and preparation, the dataset was ready with about 50 attributes to start with for the analysis

Observations

Understanding the data set

- There are multiple columns with only null values
- Some columns have only 1 unique value and null
- Few columns have more than 70% null values
- Columns like int_rate, revol_bal and emp_length which are numeric are counted as string because of additional text
- The data set has loan types as Fully Paid, Charged Off and Current. Current type of loans won’t be needed for the analysis.

Data Cleansing and Preparation

2

Data Cleaning

Deleting columns with only null or single values

Observation	Dropped columns
There are few columns with only null or single values. These columns have been given first priority for deletion.	tot_hi_cred_lim total_bal_ex_mort total_bc_limit total_il_high_credit_limit
Columns with too many null values	next_pymnt_d mths_since_last_record mths_since_last_delinq
Columns with only zero or null values	collections_12_mths_ex_med chargeoff_within_12_mths tax_liens
Not useful	last_pymnt_d last_credit_pull_d
Columns with only 1 unique value	pymnt_plan initial_list_status policy_code application_type acc_now_delinq delinq_amnt
Since current types of load won't be useful for analysis, the same shall be removed.	
There are some outliers in annual_inc. Therefore, values above 95 percentile shall be removed.	

Data Cleaning

Missing values

emp_length	•The column has texts which needs to be removed to create buckets. •<1 years shall be replaced by 0 and 10+ years with 10. All other years shall be replaced by corresponding number. •The null values shall be replaced by median of the data set. It is also observed that the mean of the column is almost 4.
revol_util	•The column has % sign after every value. This sign needs to be removed to perform analysis. •Null values in the column shall be replaced by the median of the data set. •Also it is observed that there is not much difference in mean and median.
int_rate	•The column has % sign after every value. This sign needs to be removed to perform analysis.
pub_rec_bankruptcies	•Null values shall be replaced by 0 i.e. the median of the column
home_ownership	•The column has only 3 NONE as values, which can be considered as null. •Since the mode of the column is RENT, replacing all NONE values with RENT

Deriving values

New features for analysis

issue_d
Issue date of the loan

The column first needs to be converted to correct date format.

Post that month and year of the issue date can be derived.

annual_inc
Annual income of the borrower

Annual income of the borrower can be used to compute the monthly income.

By subtracting installment amount from above value, approximate savings value can be derived

Univariate Analysis

3

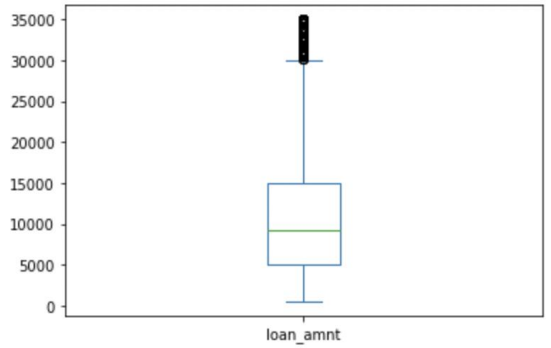
Univariate Analysis Discussion

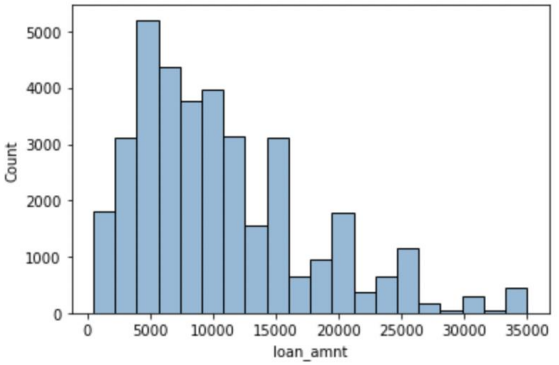
Loan Amount Distribution

Univariate analysis helped to look at the distribution of the loan amounts and the key observations were:

- Average loan amount was ~$10k, while the median was around ~$9.2k.
- 75% of the loan amounts were under $15k.
- Maximum loan amount sanctioned was $35k

Loan amount distribution box plot and histogram



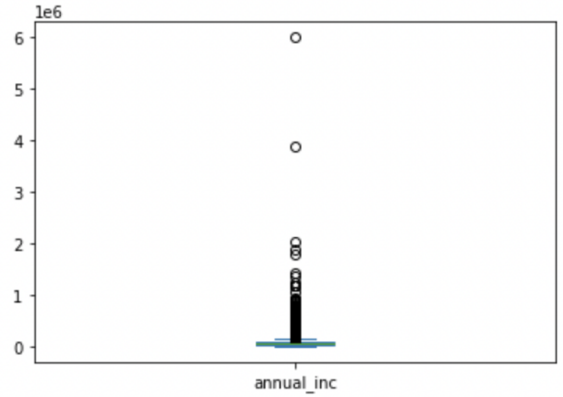


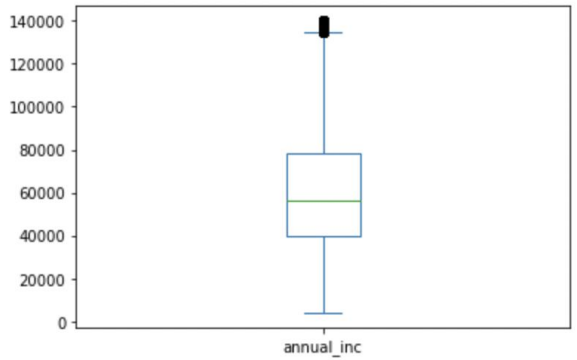
Annual Income Distribution

Univariate analysis helped to look at the distribution of the annual income of the applicants and the key observations were:

- The annual amount had outliers revealed by the box plot.
- Outliers were removed by removing values >95 percentile.
- Annual income varied from $4k to $117k
- Mean was ~$57k and median ~$55k
- 75% of the applicant's income were under $74k

Annual Income Distribution plot before and after outlier removal





Bivariate Analysis Discussion

4

Bivariate Analysis Discussion

Laon Amount vs. Delinquency

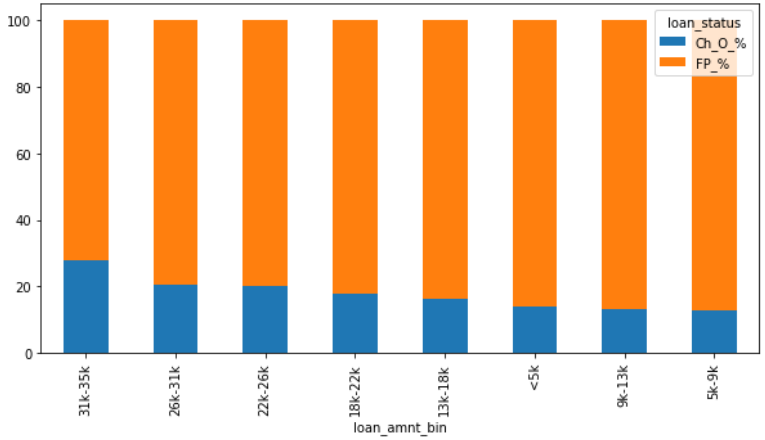
- There's an increasing trend in delinquency: delinquencies are higher in higher loan amounts; highest 28% in $31k-$35k category and lowest - ~13% in $5k-10k category

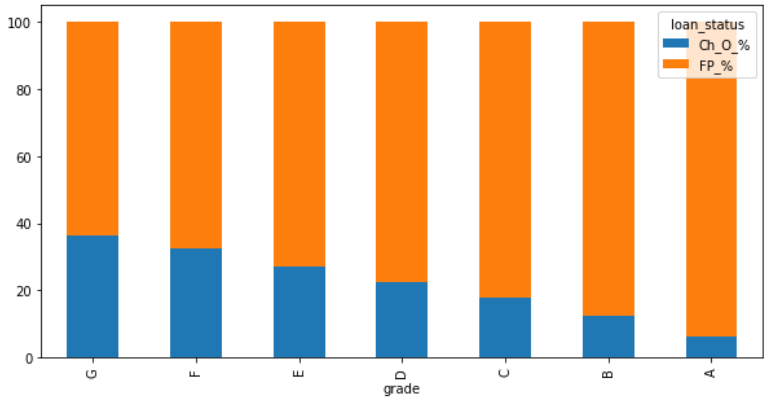
Grade vs. Delinquency

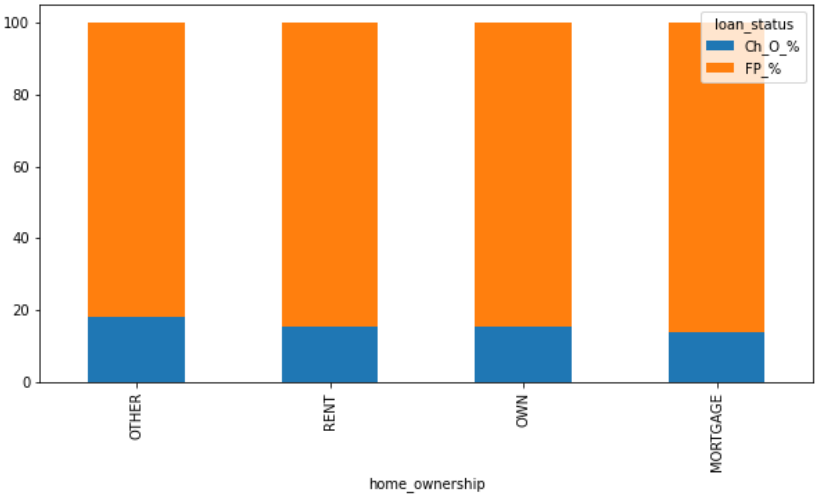
- High delinquencies are observed as Grade decreases
- There is a difference of ~30% delinquency rate in case of grade G and A. People with grade G are highly likely to default.

Interest Rate vs. Delinquency

- Relatively similar delinquency rates across home ownership categories
- Highest observed in OTHER category with 18%



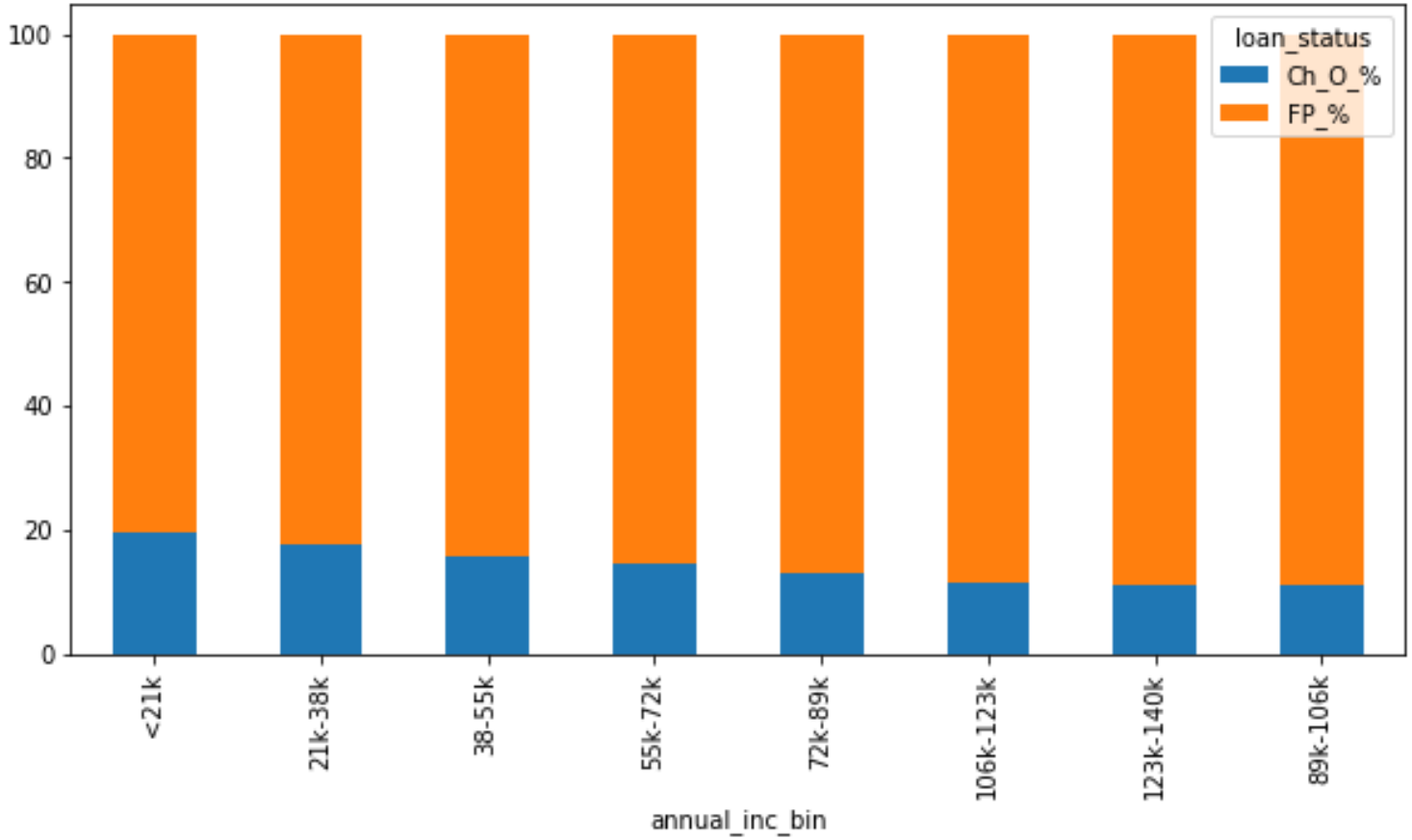




Bivariate Analysis Discussion

Annual Income vs. Delinquency

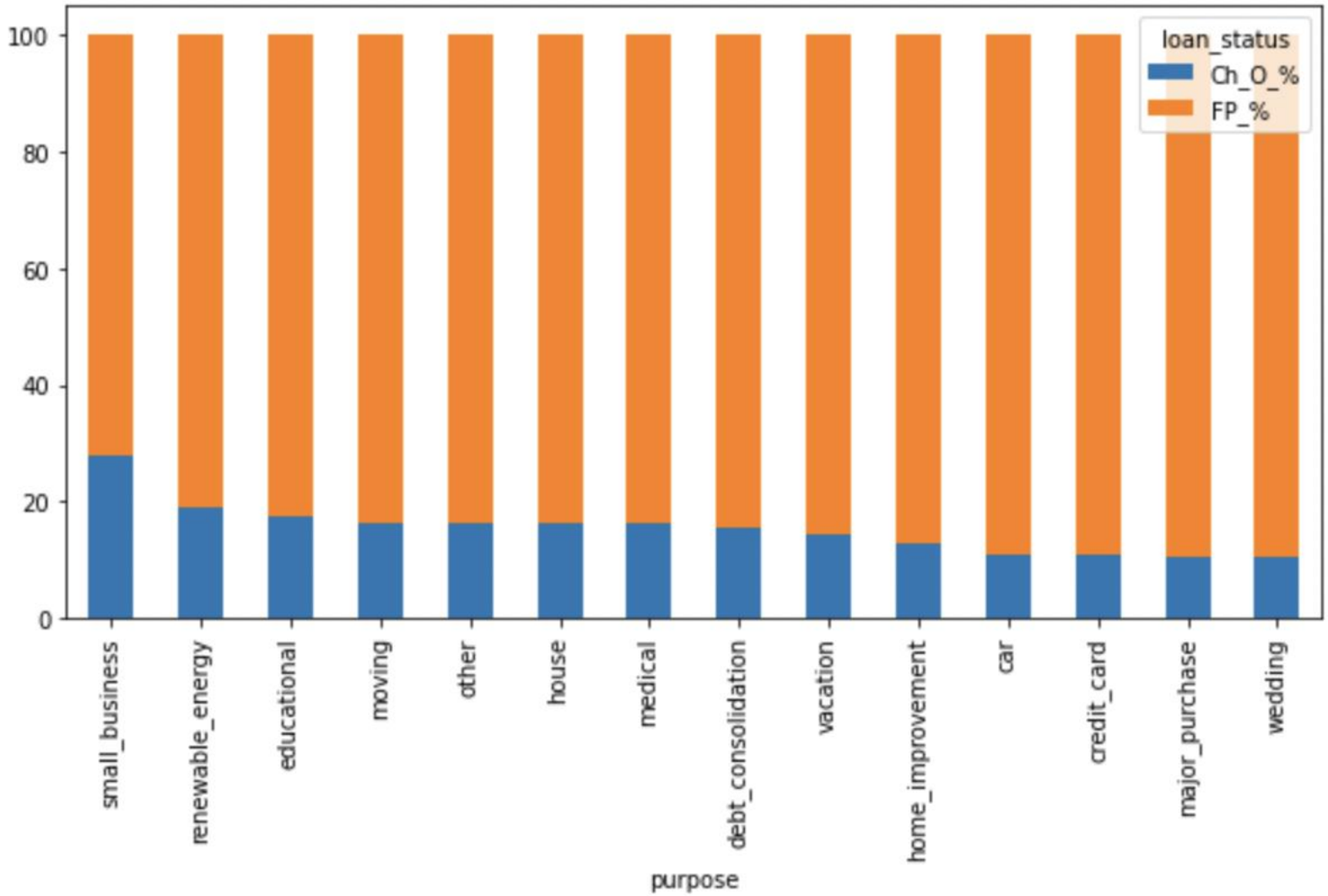
- Highest delinquency in the lower income group
- Highest delinquency at ~20% in <$21k annual income



Bivariate Analysis Discussion

Loan Purpose vs. Delinquency

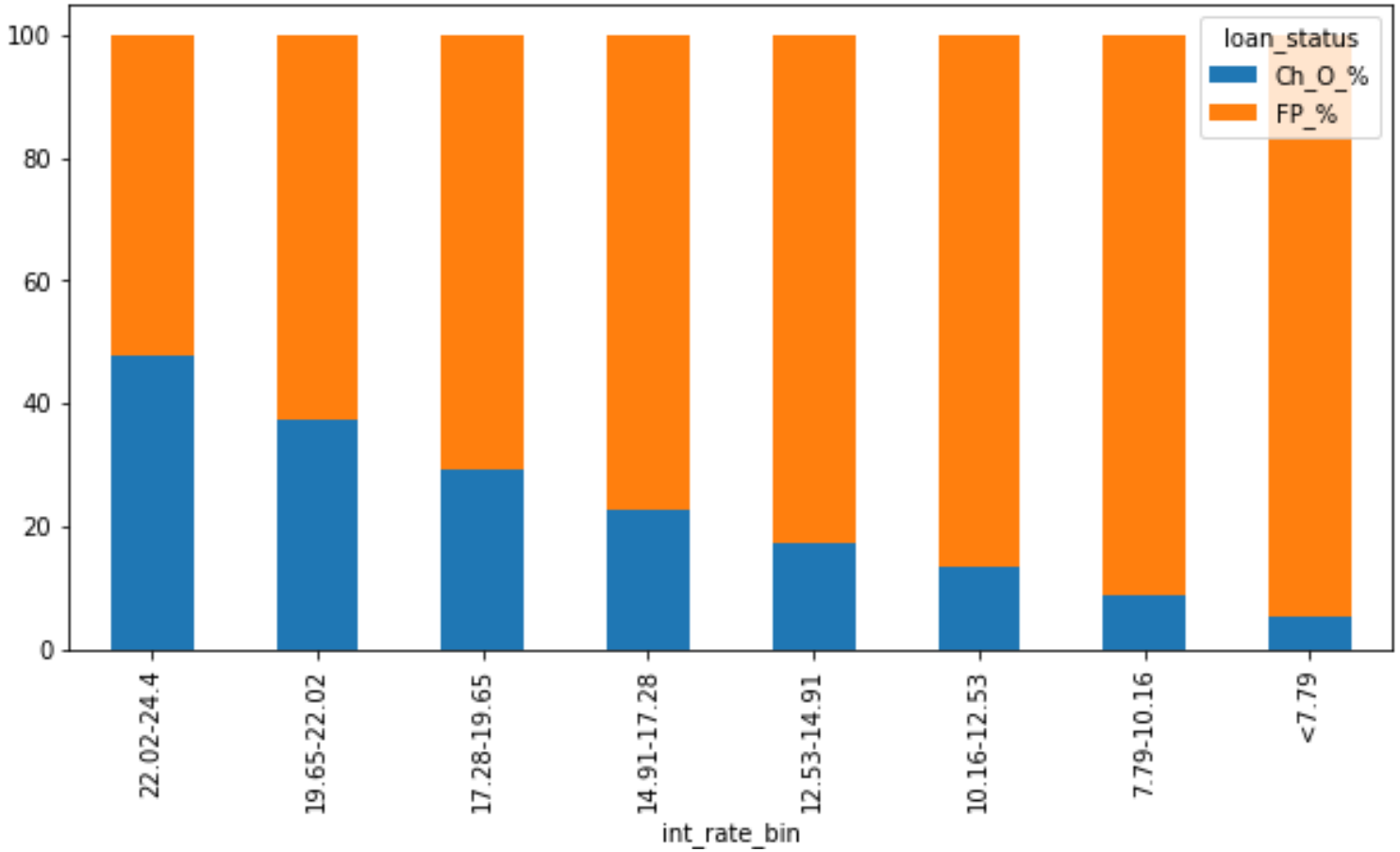
- Purpose SMALL BUSINESS has the highest delinquency with 28%
- Wedding has the lowest at 10.3%



Bivariate Analysis Discussion

Interest rate vs. Delinquency

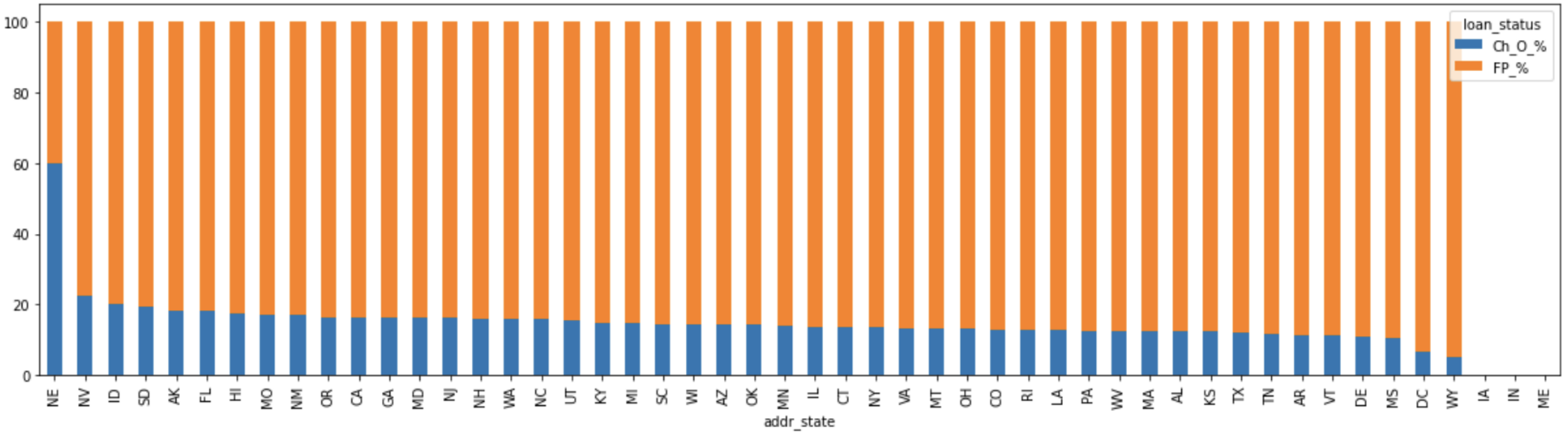
- As expected, the rate of delinquency increases with increase in interest rate
- Delinquency increases from 5% to ~48% as the interest rate goes from <7.79 to 24.4



Bivariate Analysis Discussion

State vs. Delinquency

- Highest delinquency of 60% observed in Nevada
- Lowest in Wyoming at 5%

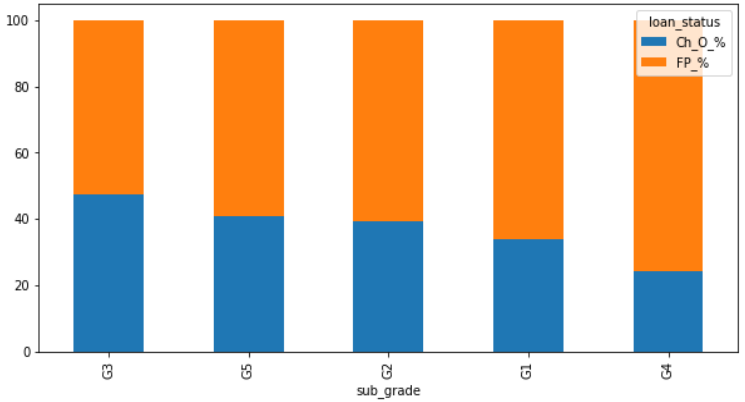


Multivariate Analysis Discussion

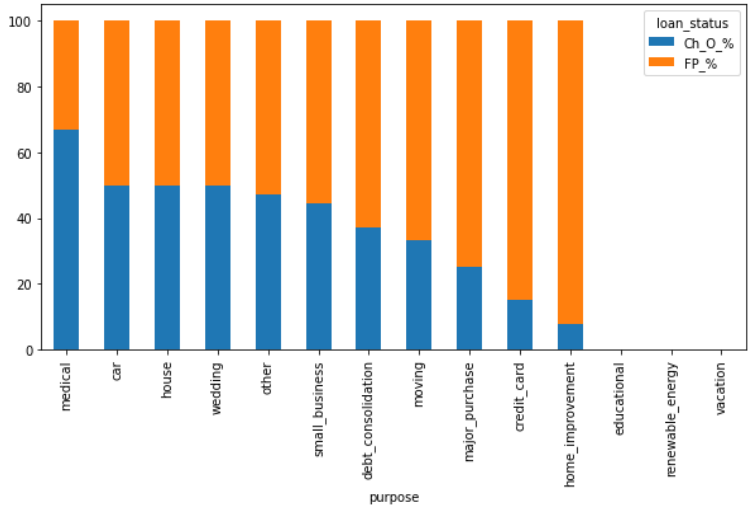
5

Based on key influencers

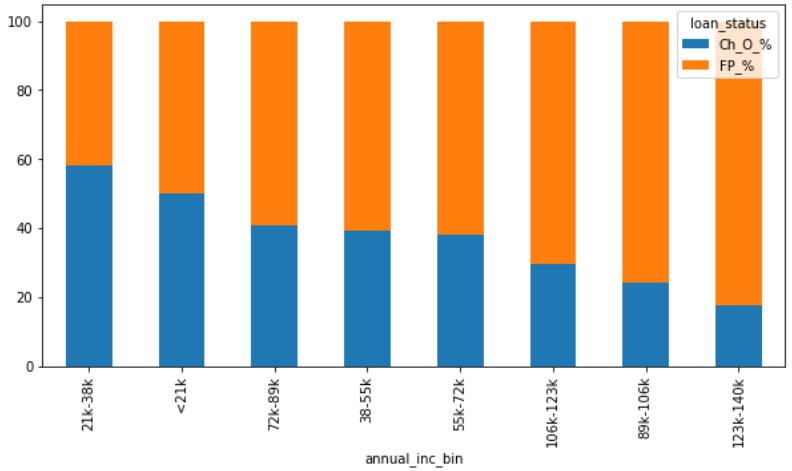
Considering grade G has highest rate of delinquency



Sub grade G3 has highest rate of delinquency with 47.5%



- Earlier it was observed that small business have highest delinquency rate.
- However if focussed on grade G, loans taken with medical **purposes** have highest delinquency rate with 66.67%



Annual income group between **21k and 38k** and group G have 3/5 chances of being defaulters

Conclusion

6

Discussion of Observations & Recommendations

Key Influencers

- Key influencer attributes are:
  - Purpose
  - Annual Income
  - State
  - Grade
  - Interest rate

Recommendation

- A risk model to be developed based on these attributes with a base interest rate.
- Based on the risk model, add on interest rates are added to the base interest rate to compensate for delinquency risk.