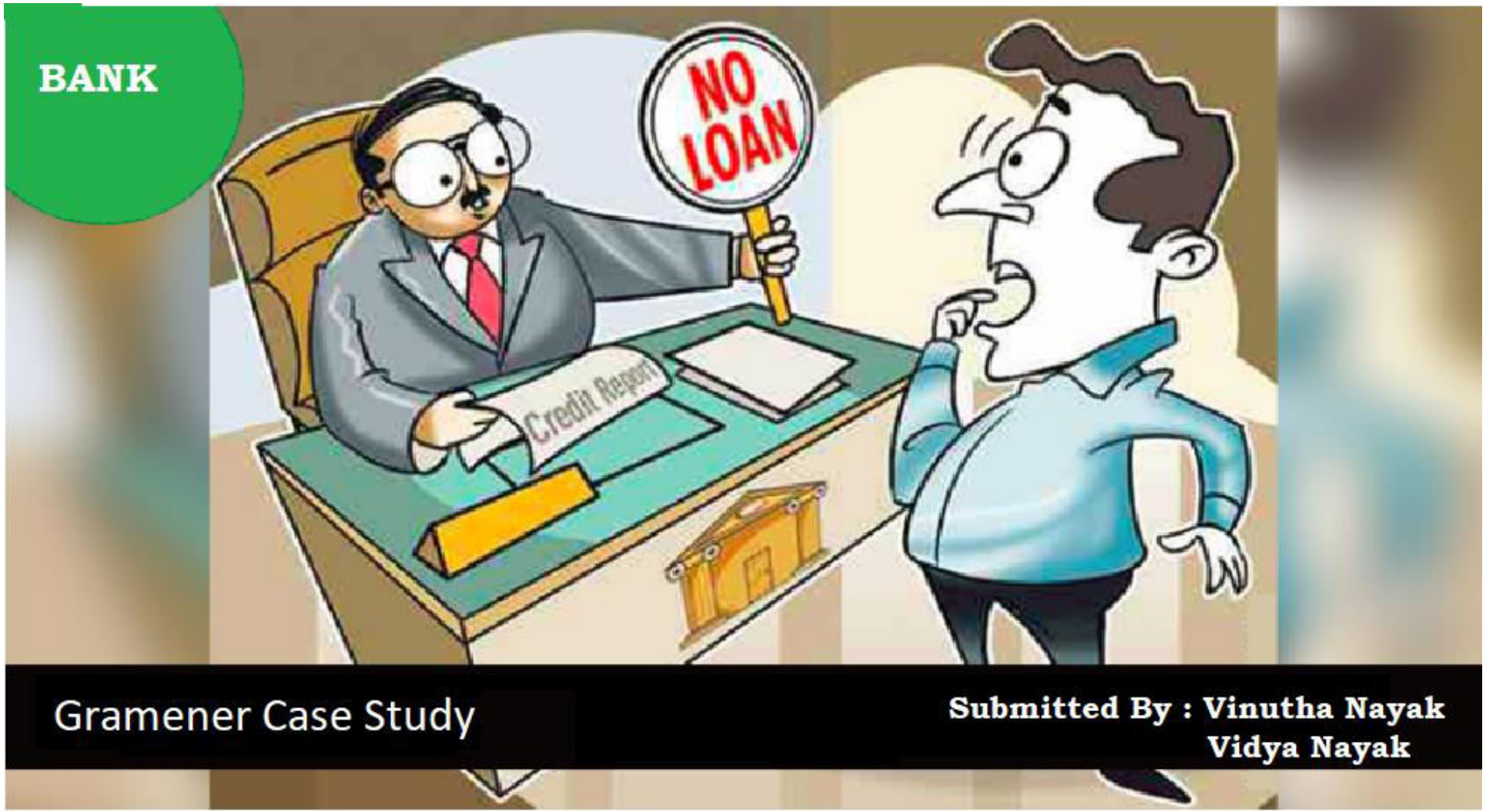


Gramener Case Study

Submitted By : Vinutha Nayak
Vidya Nayak

Loan Data Analysis

Abstract:

This study explored the pattern in the loan data set to find out the driving factors that can help in key decision to approve or reject the loan.

The study was conducted on historic data recorded for four years from 2007-2011 which includes all the variables that reveal essential consumer attributes. This data was then analyzed to gather various insights

- Percentage of loans that default
- Loan purposes that cost hefty amount to the bank in terms of “Charge-offs”
- Grade wise impact on the charged off loans
- Historical records like credit information revolving history etc which reveal consumer nature and risk

Our Analysis :-

Based on the study ,top three loans that cause maximum credit loss to the bank are that are taken for the purpose of debt_consolidation , credit card or small business. Of the top three areas, grades B,C and D have the maximum count of loans defaults. However borrowers under Grades E,F, are more likely to default, although the number of loans borrowed is less. Hence, there is a trade-off . Home ownership has no major impact on defaulting.However,finding suggest that higher the income rate , and lower the salary there are high chances of default.

EDA Strategy

Data Cleaning
- Clean the data, so as to only keep the data relevant for analysis

Derived Metrics
- Derive new data from the existing variables, that can help in data analysis

Plotting
- Plot the different variables and perform univariate/bivariate analysis

Conclusion
- Based on the insights derived, suggest the bank best way to reduce “credit loss”

Data Cleaning

Data Source/ Loan.csv
Input columns: 109

Fix rows and columns

1.Unique values or keys (Eg: id,member_id)
2. Future variables (Eg:last_paymnt_date)
3.Single valued variables(Ex:application_type)
4.Missing Values (Ex:num_actv_bc_tl)
5.High % of NA values (Ex : total_rec_late_fee - around 0.05% have values)
6.Delete duplicate columns(Ex:total_pymnt,total_pymnt_inv)
7.Irrelevant columns (Ex: url)

Standardize Format

1.Removing months in term
2.Removing % symbol in int_rate, revol_util
3.Remove “years*”, “<“ and “+” symbol in emp length

Cleaned data source
Output columns: 25

Derived Metrics

✓ Type Driven Metric

- Split Dates to month and year for columns Issue_d , earliest_cr_line

✓ Business Driven Metric

- Find the percentage monthly income that the consumer will pay as loan installment

✓ Data Driven Metric

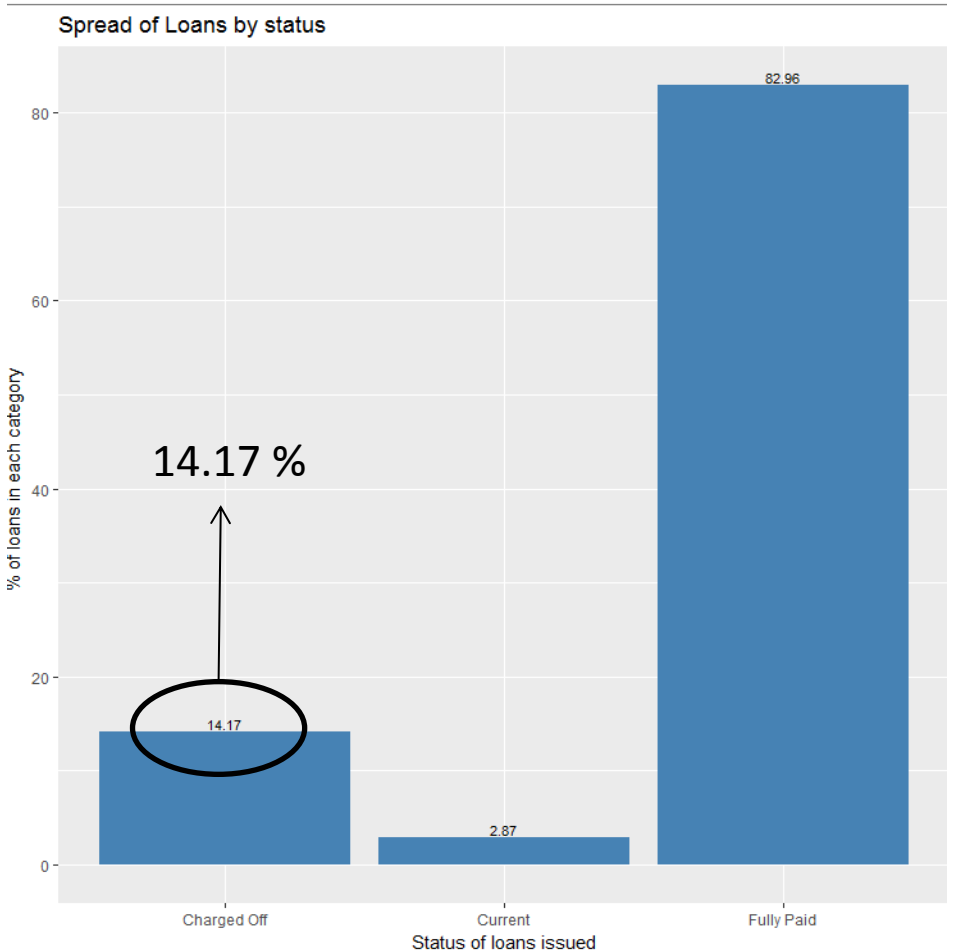
- % of loan paid as total_rec_prncp/loan_amt
- % of loan defaulted as (number of loans charged off)/Total number of loans in different categories

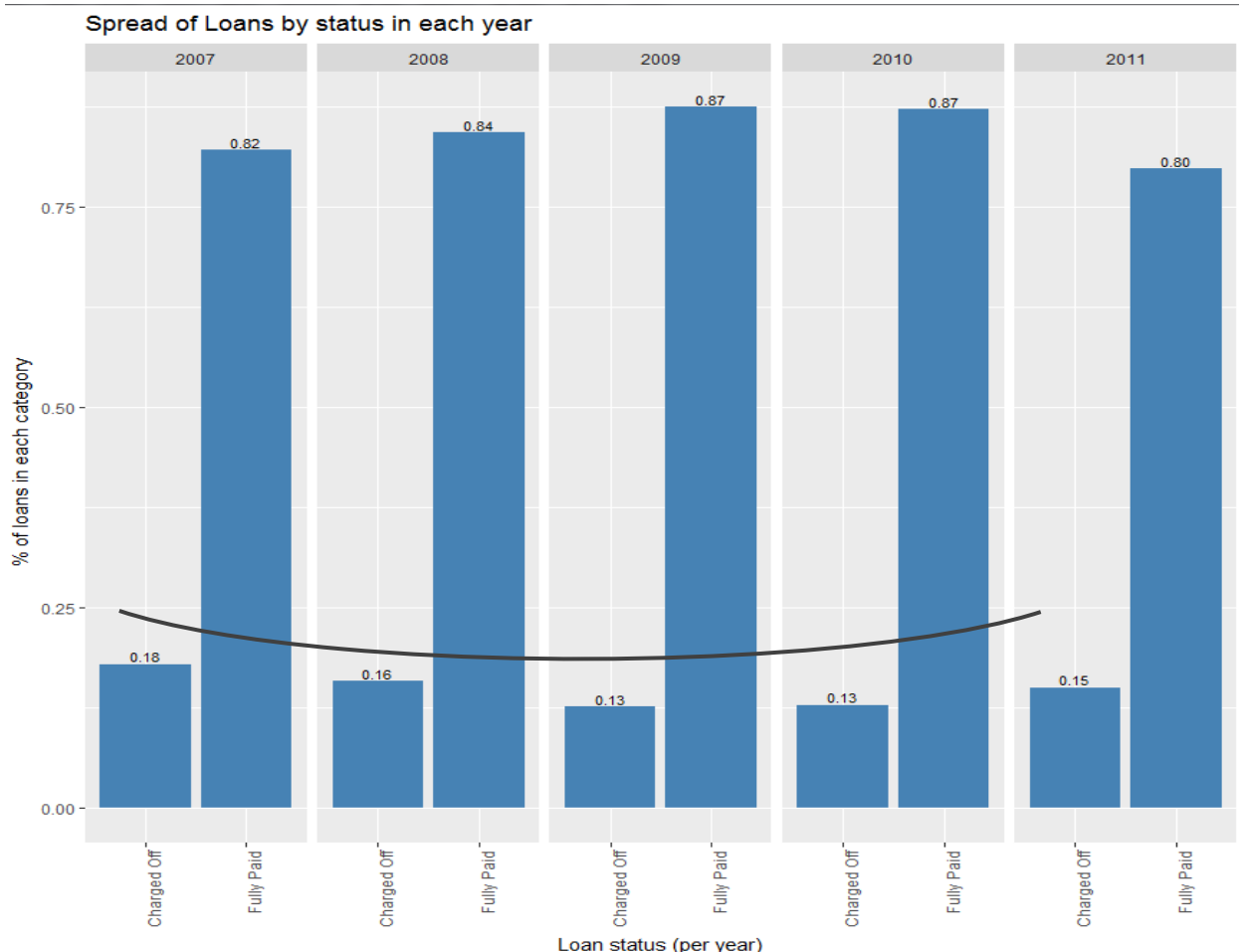
Univariate Analysis

Analysis 1 :

1.1 How much percent of the total loan gets defaulted ? **14.17%**

1.2 How is the trend of the loans across years ? Does number of defaulting loans increase ? No





Analysis 2 (Based on Purpose) :

2.1 Of the total loans defaulted, loans lent for which purpose have defaulted the most ?

2.2 Of the total loans defaulted, in which "Purpose" borrowers are more likely to default ?

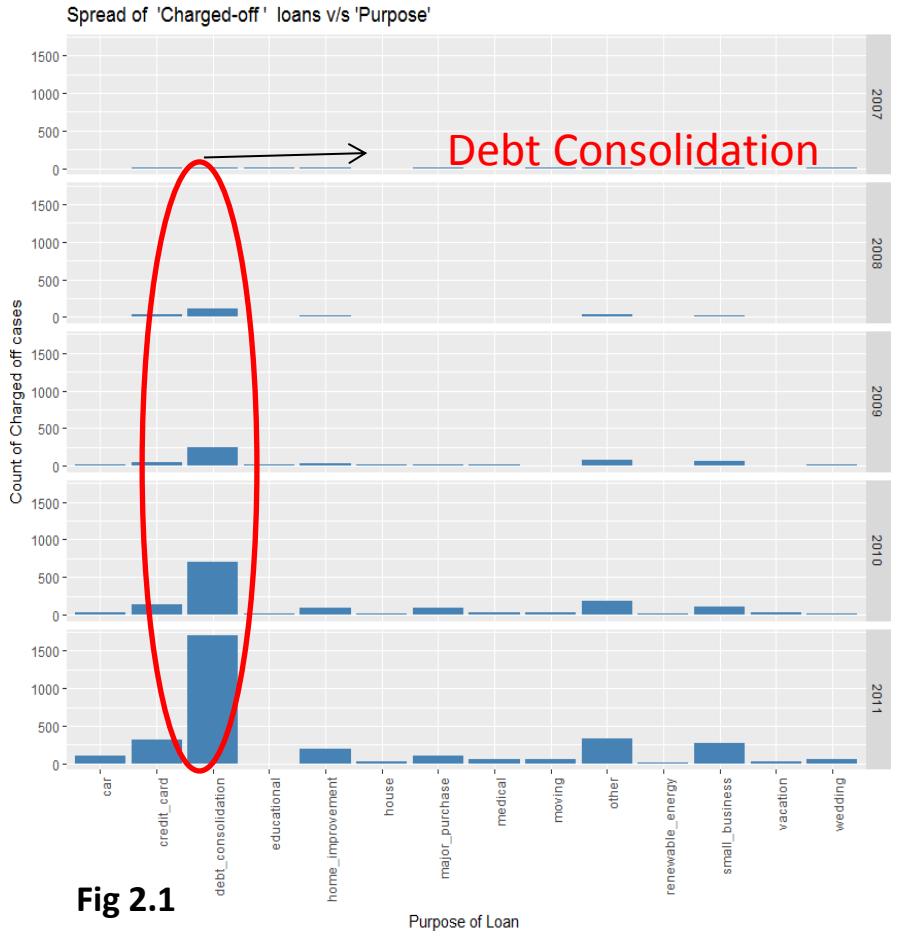


Fig 2.1

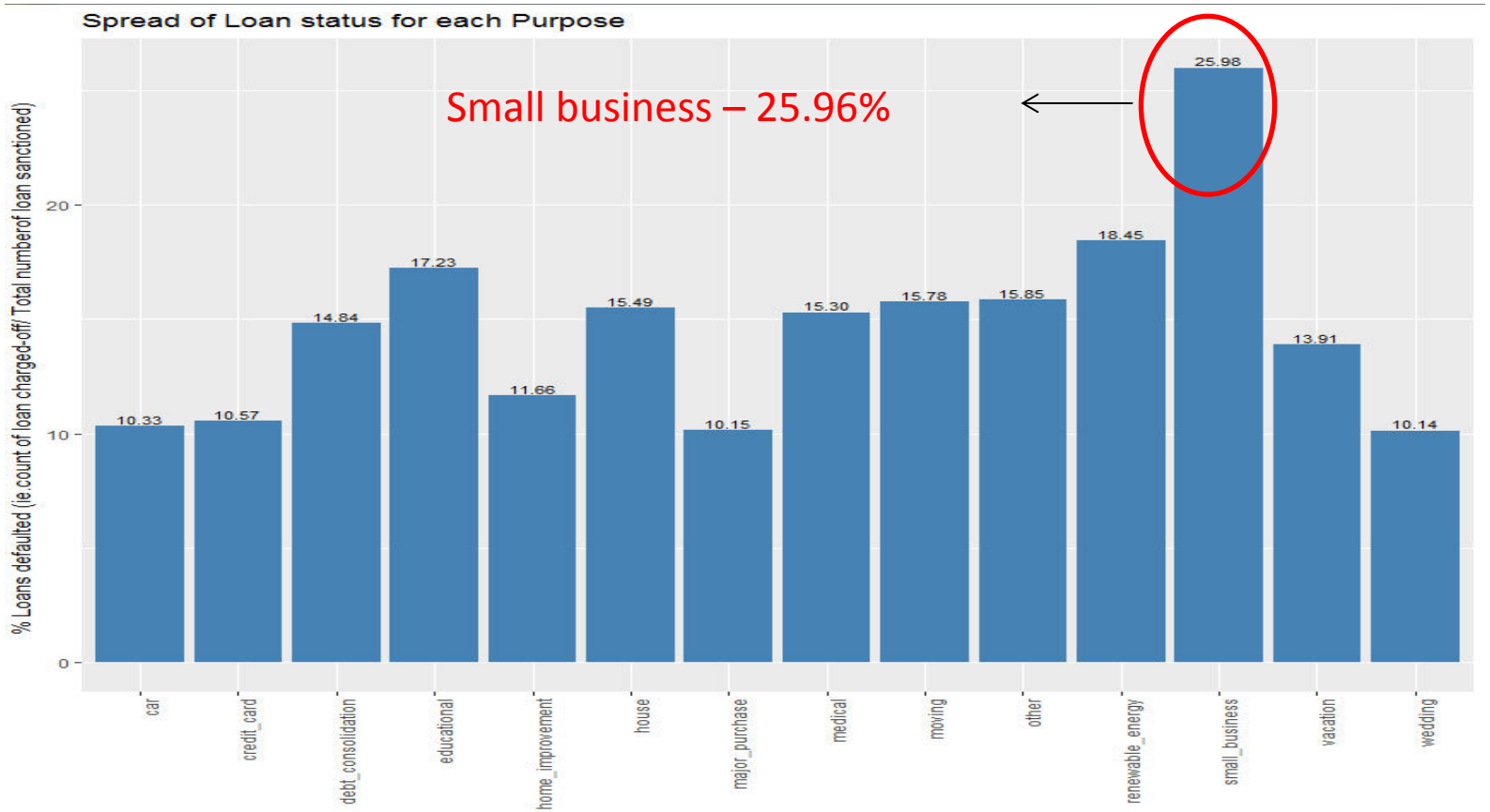
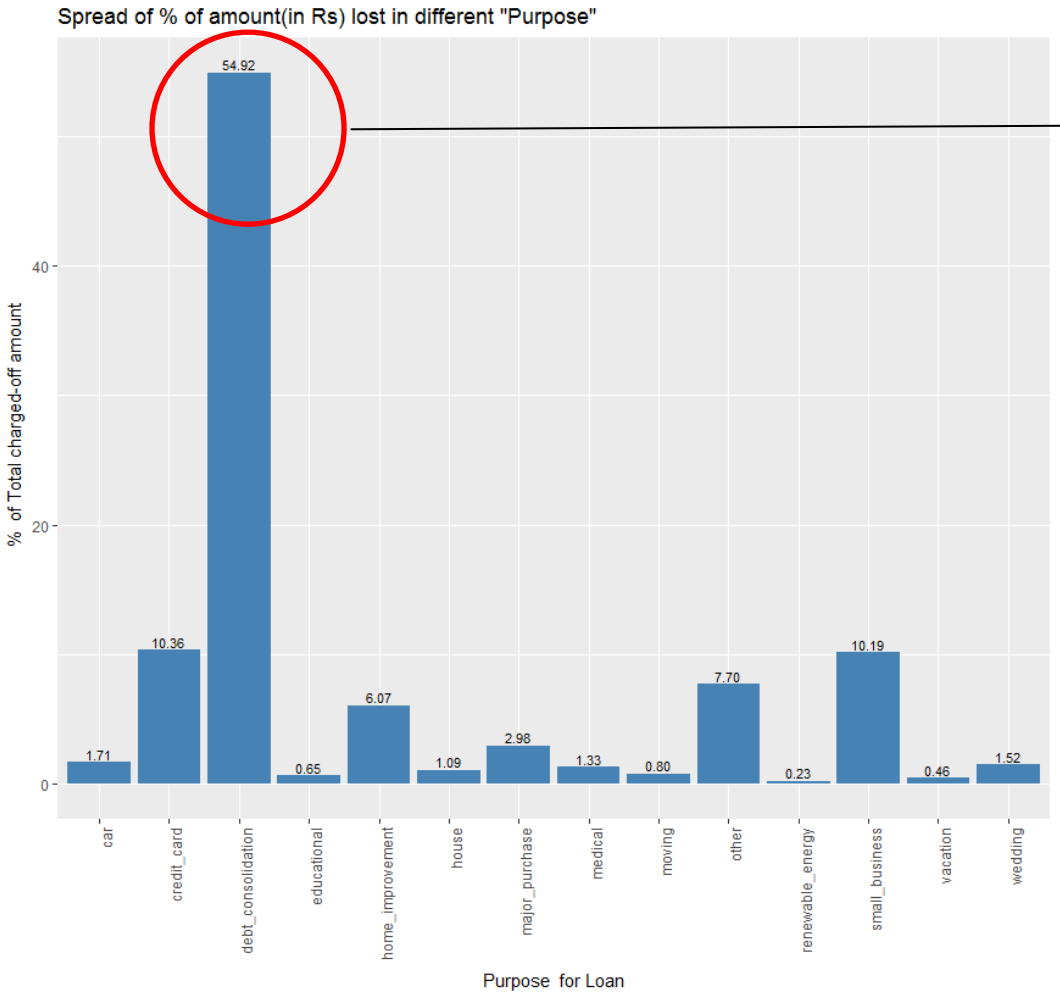


Fig 2.2

Analysis 2 :

2.3 Of the total loans defaulted , money lent for which "Purpose" is leading to major loss of money for bank ?



54.92% of the total charged-off amount of is constituted by loans lent for Debt-consolidation

Summary :

- Number of charged-off loans, that were taken for "debt_consolidation" purpose are consistently and significantly increasingly over the years
- Loans lent for "small buisness" has high (Toal number of loans charged-off/Total number of loans sanctioned ratio)
- Top three categories where money lost is maximum due to defaults are orderly :

1. **Debt_consolidation**
2. **Credit_card**
3. **Small_business**

Analysis 3 (Based on Grade) :

3.1 For the Top 3 defaulting categories , what is the spread of Grades ?

3.2 Does the grade of a consumer bear high significance ?If so , which are the risky ones?

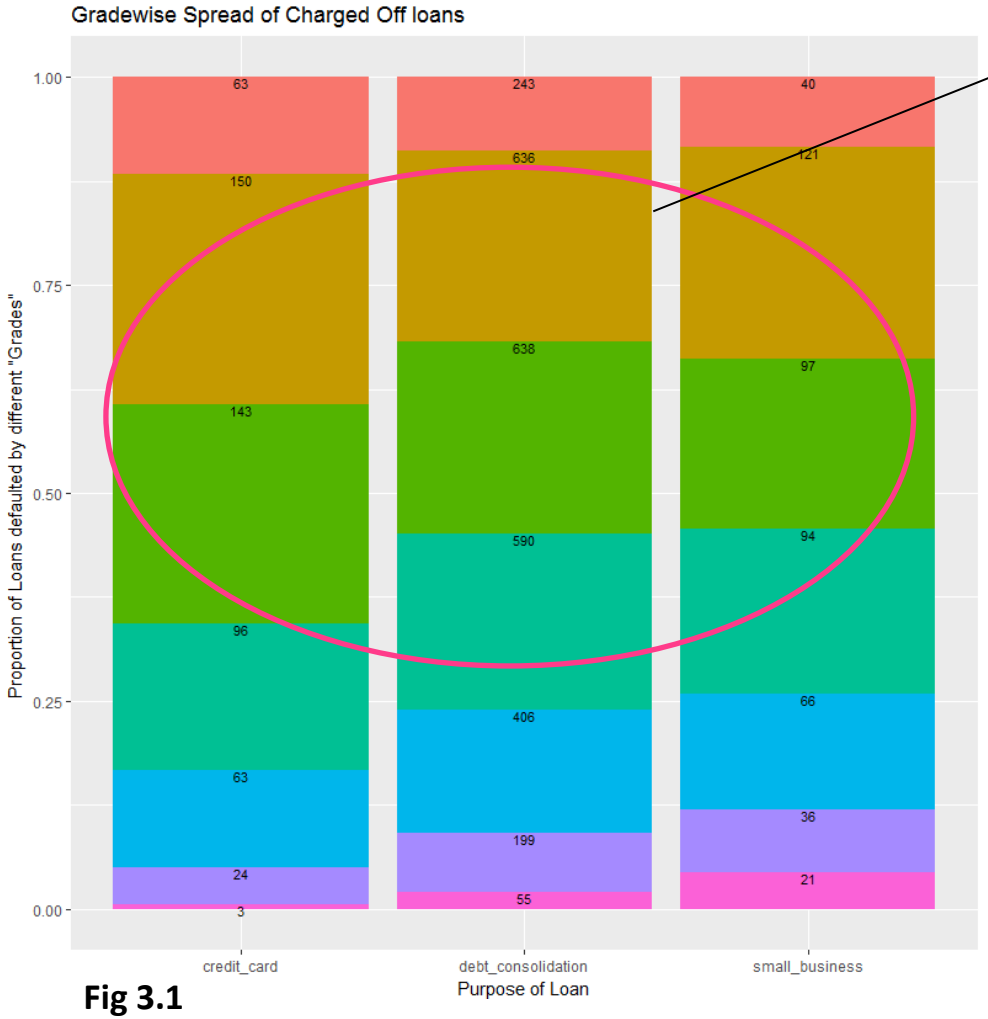


Fig 3.1

Grades B,C,D has maximum defaults

Grades E,F,G has maximum

percentage of charged off loans for that purpose

Total number of loans taken

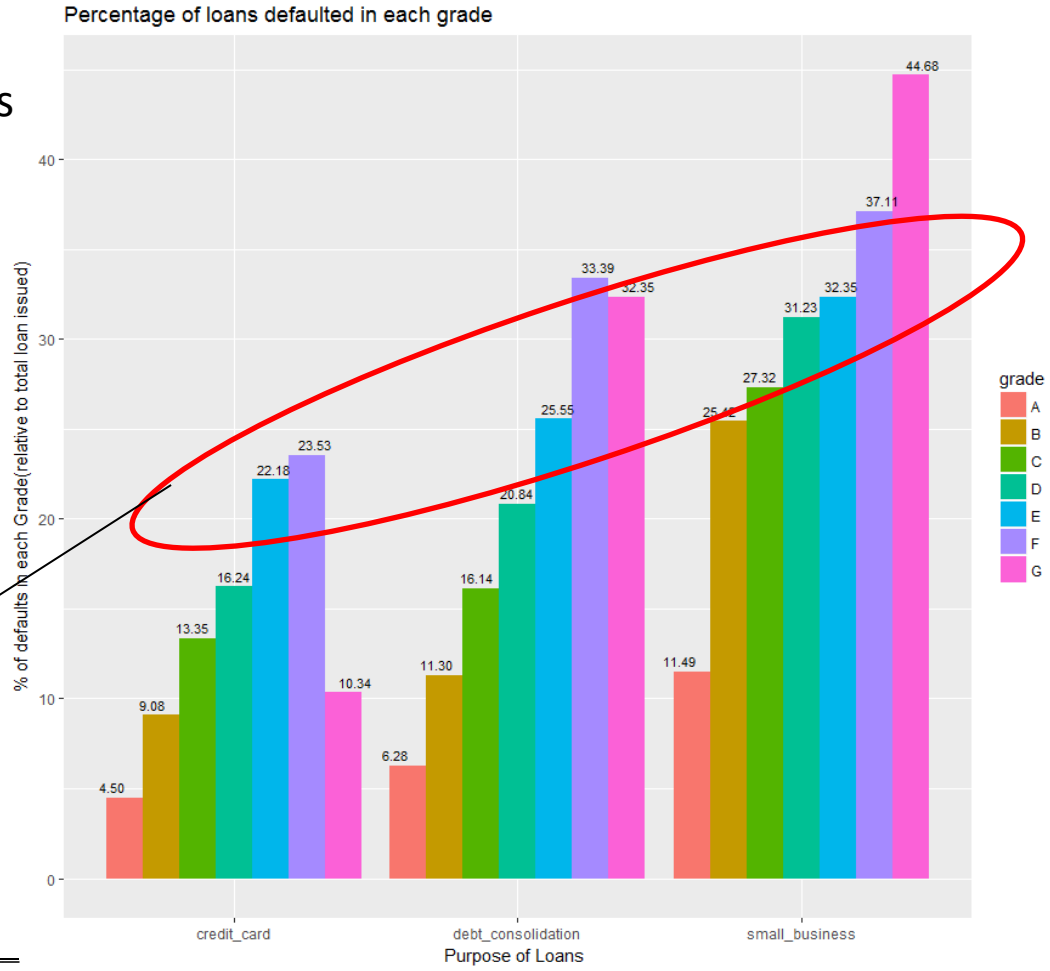


Fig 3.2

Bivariate Analysis

Analysis 4 (Based on Employment_length):

4.1 Does Employment length have impact on Credit Loss Risk ?

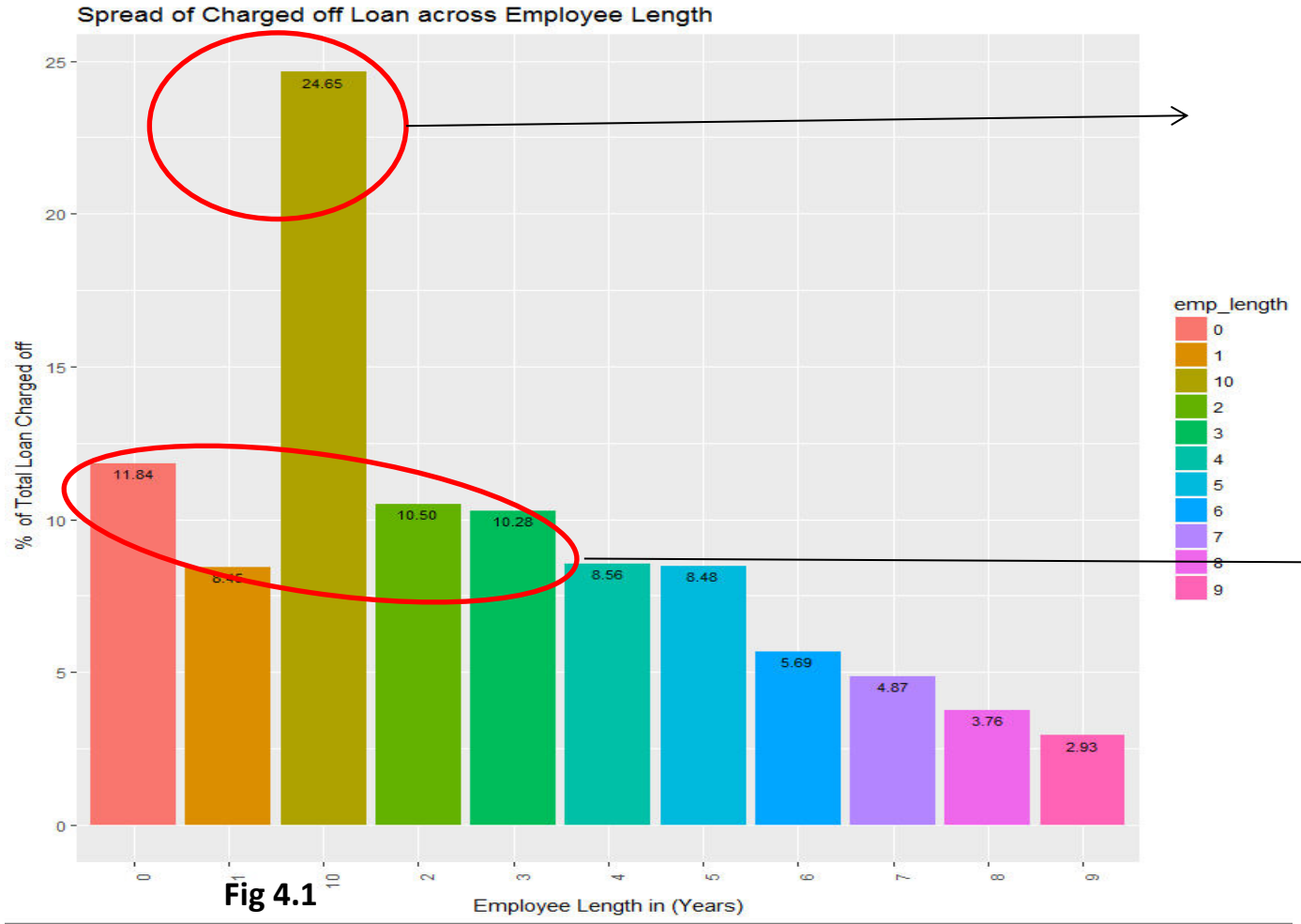


Fig 4.1

➢ Employment length >10 years account for maximum number of charged off loans

➢ Employment length 0,2 ,3 account for next largest defaulters

Bivariate Analysis

Analysis 5 (Based on Income):

5.1 Is there any co-relation of Income with other categorical variables such as Loan_status, Grade ?

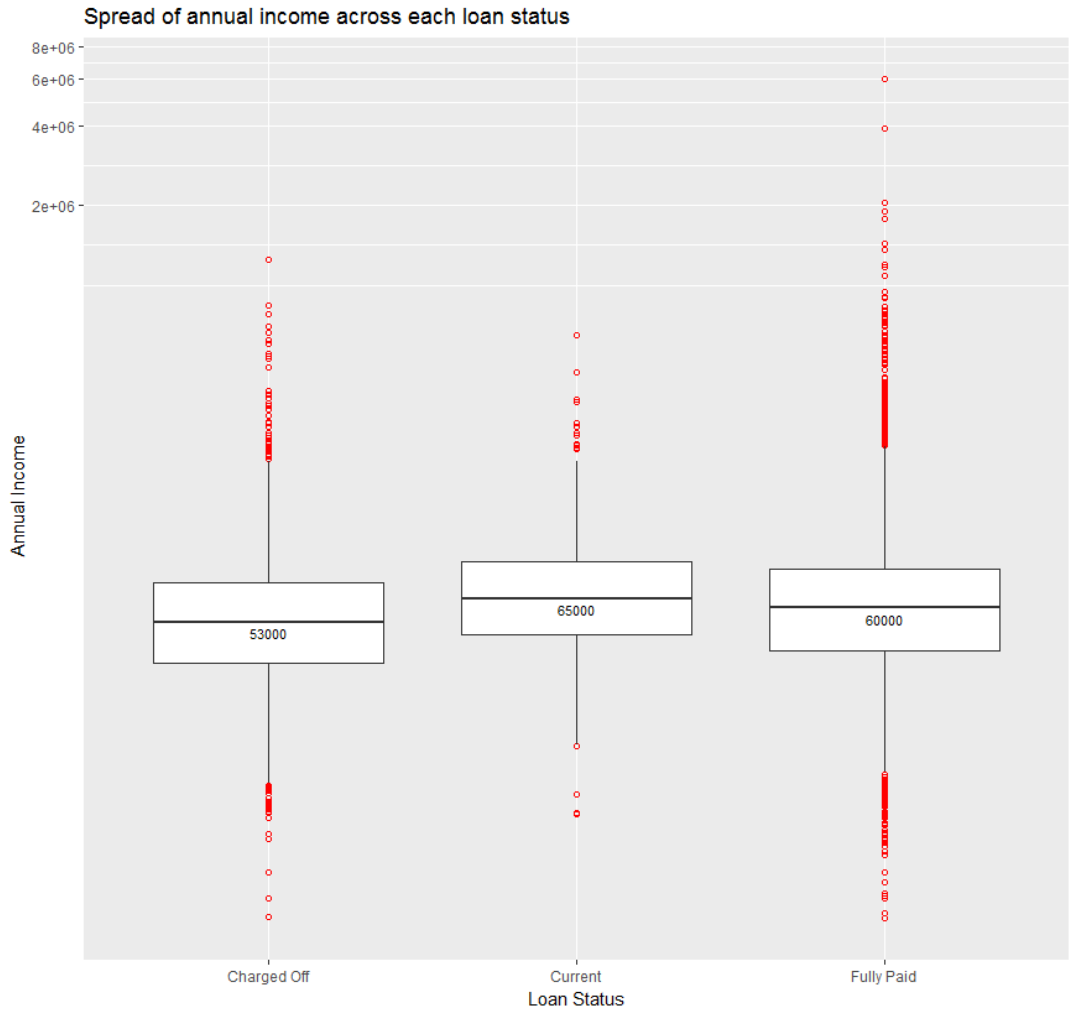


Fig 5.1 (a)

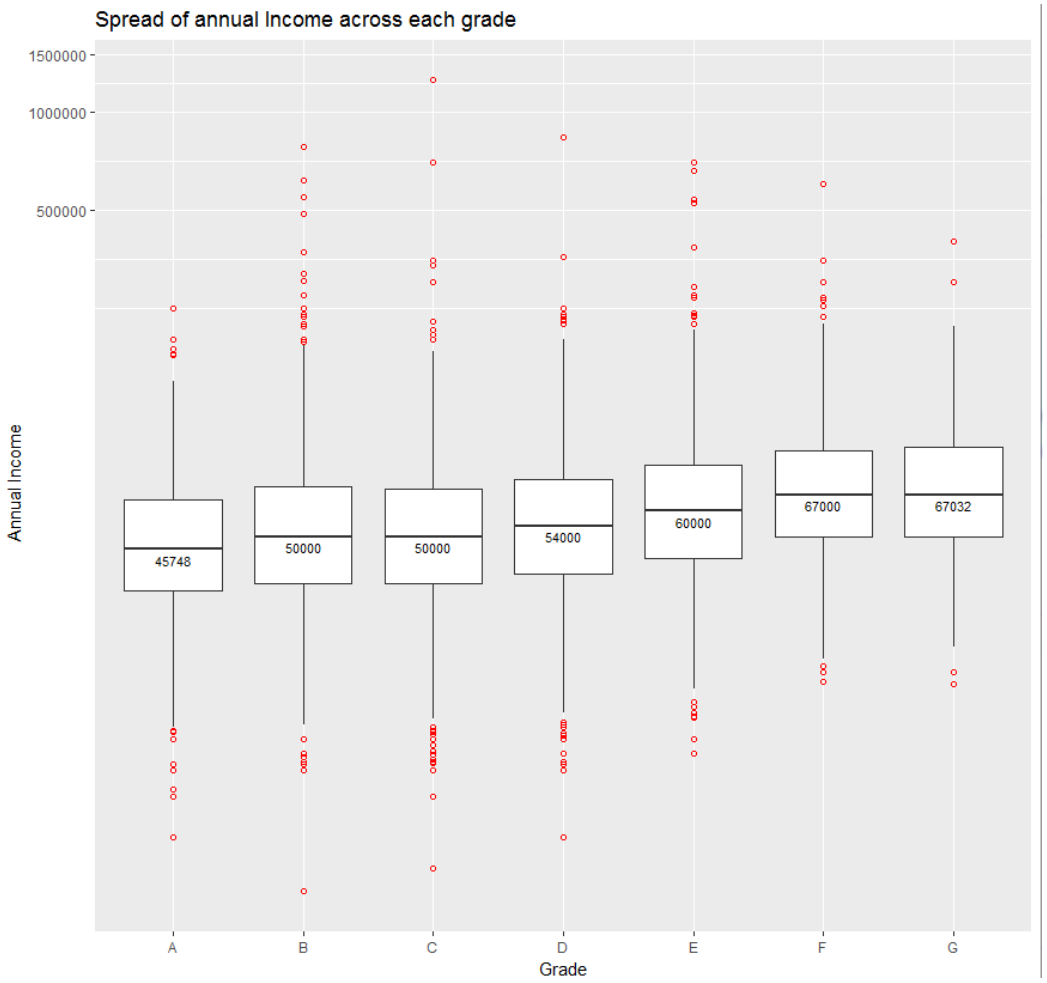


Fig 5.1 (b)

Bivariate Analysis

Analysis 5 (Based on Income):

5.2 Can we infer anything from the Interest-rate v/s Grade patterns ?

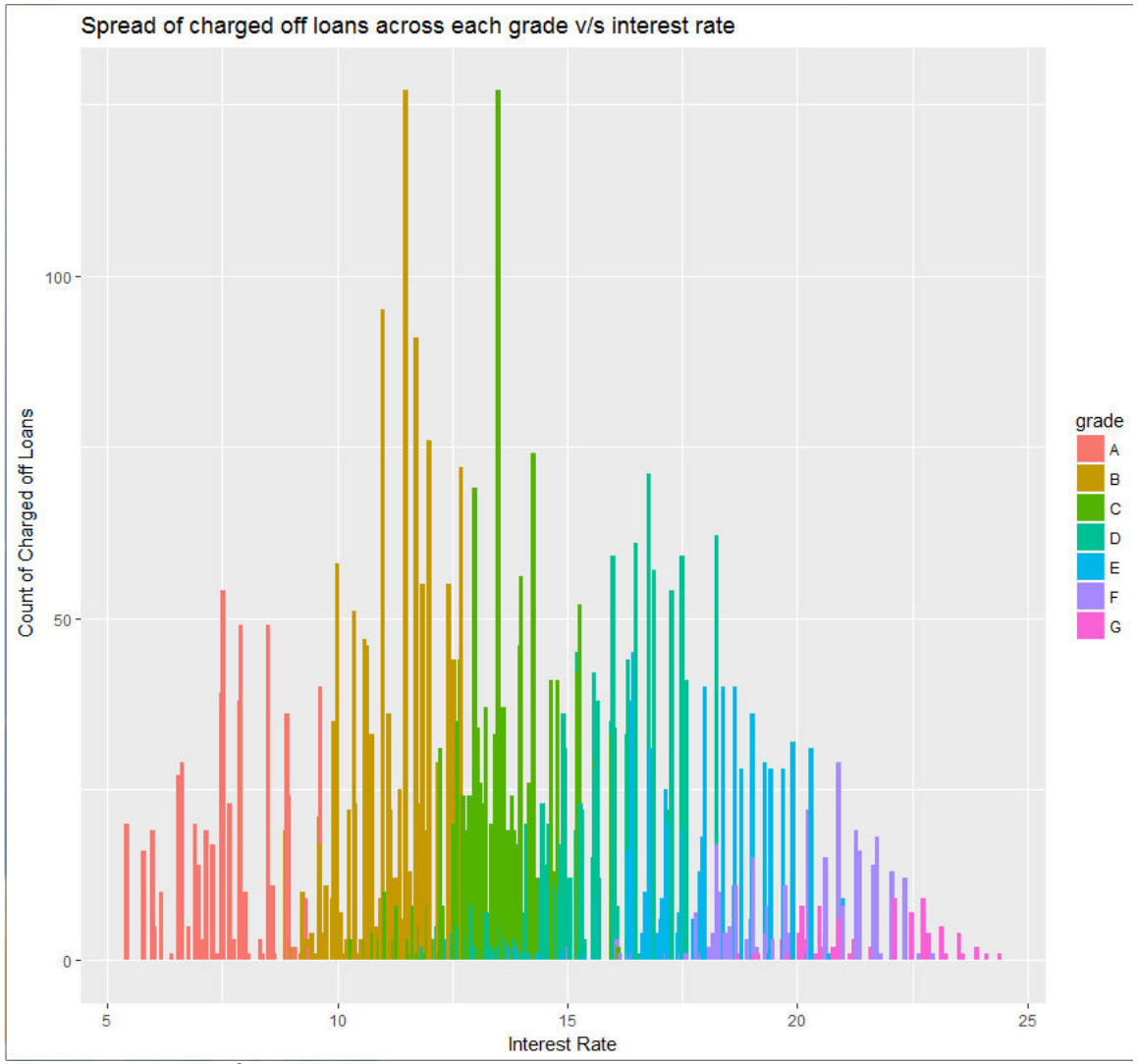


Fig 5.2

Summary :

- The average income of full paid group is higher
- Fully paid group has lot of outliers above the third quartile range
- Interest Rate is directly proportional to Salary.
- A with lower average salary have defaulted for lower interest rates
- B,C have defaulted at slightly high interest rates,with their mean salary simialar,the overlap on interest rates are also high
- E,F,G who have average salary higher have defaulted only at highter interest rates.

Important factors that contribute to default

- Grade
- Loan_purpose
- Annual_income
- Interest_rate
- DTI
- Employee length
- State

Note: *term, house ownership , % of monthly income as installment etc did not bear any major influence on loan charge off pattern.*

Conclusions

- California has highest number of loan requests and high percentage of loan default
- Top 3 contributors of charged off loans are debt_consolidation,credit_card and small business over the period of analysis.
- While small_business contributes higher % of loan defaulted
- deb_consolidation contributes maximum loss ie ,to about 50% of the total credit loss
- For the top 3 contributors ,customers belonging to grades B, C, D have defaulted high
- Grades E,F,G have higher risk of default , considering the % of total loan v/s defaulted under each grade
- Around one-fourth of the loans defaulted are from employees who have over 10+ years of experience
- Either low experience below years or more than 10 years have high risk of default
- The income of charged off loans is relatively less as compared to fully paid ,and majorly higher and above the third quartile region
- Interest rate and income have direct relation .
- Grade wise average income increases and higher grade employees have risk to default with higher interest rates while they dont default at lower interest rates.
- DTI of above 10 to around 25 have higher risk of default.