Lending club case study

Avinash S R

26/05/2023

Batch: EPGP ML52 April 2023

Problem statement

- It is required to find the driving factors (loan attributes and customer attributes) that can lead to loan defaults

Analysis approach

- Removal of columns in which all or majority of the values are missing or zero
- Univariate and bivariate study to ascertain their impact on the target column- loan_status

Dataset description and data cleaning

- The dataset has 39717 rows and 111 columns
- Target column is “loan_status” which is further categorized by “Fully paid”, “Charged off” and “Current”
  - Fully paid: indicates the customer has repaid the loan
  - Charged off: this indicates that customer has defaulted on loan payment
  - Current: customers who are currently paying the installments
- Several of the columns have all their values as “NA”, these were removed from the dataset resulting in 39717 rows and 57 columns

Data cleaning

- It is observed that many columns have “0” for all the values or majority of the values are zero, these were removed as well, resulting in 37 columns
- The columns “term”, “int_rate”, “emp_length” and “revol_util” were modified so as to convert their datatype to int32 or float
- Month and year were extracted from the column “issue_d”(the month which the loan was funded)

Univariate analysis-checking for outliers

- Outliers are observed for some the variables(total_pymnt and total_pyment_inv) shown in boxplot

- Additionally the median values of all these columns are very close to each other

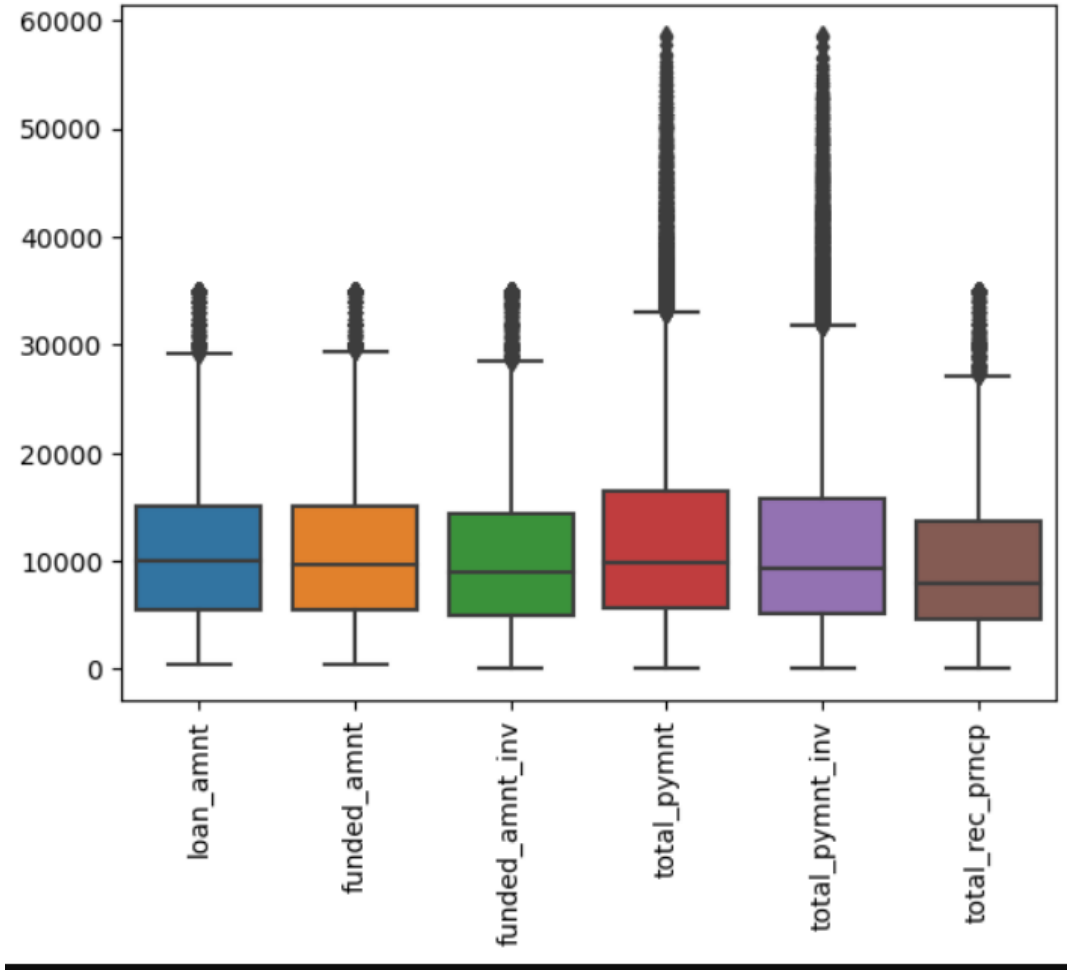

Fig: Boxplot showing outliers

Univariate analysis-checking for outliers

- Figure shows the boxplot of variable “annual_inc”, this represents the annual income of the customers

- Outliers are observed in this column and it is possible that few customers may have very high annual incomes

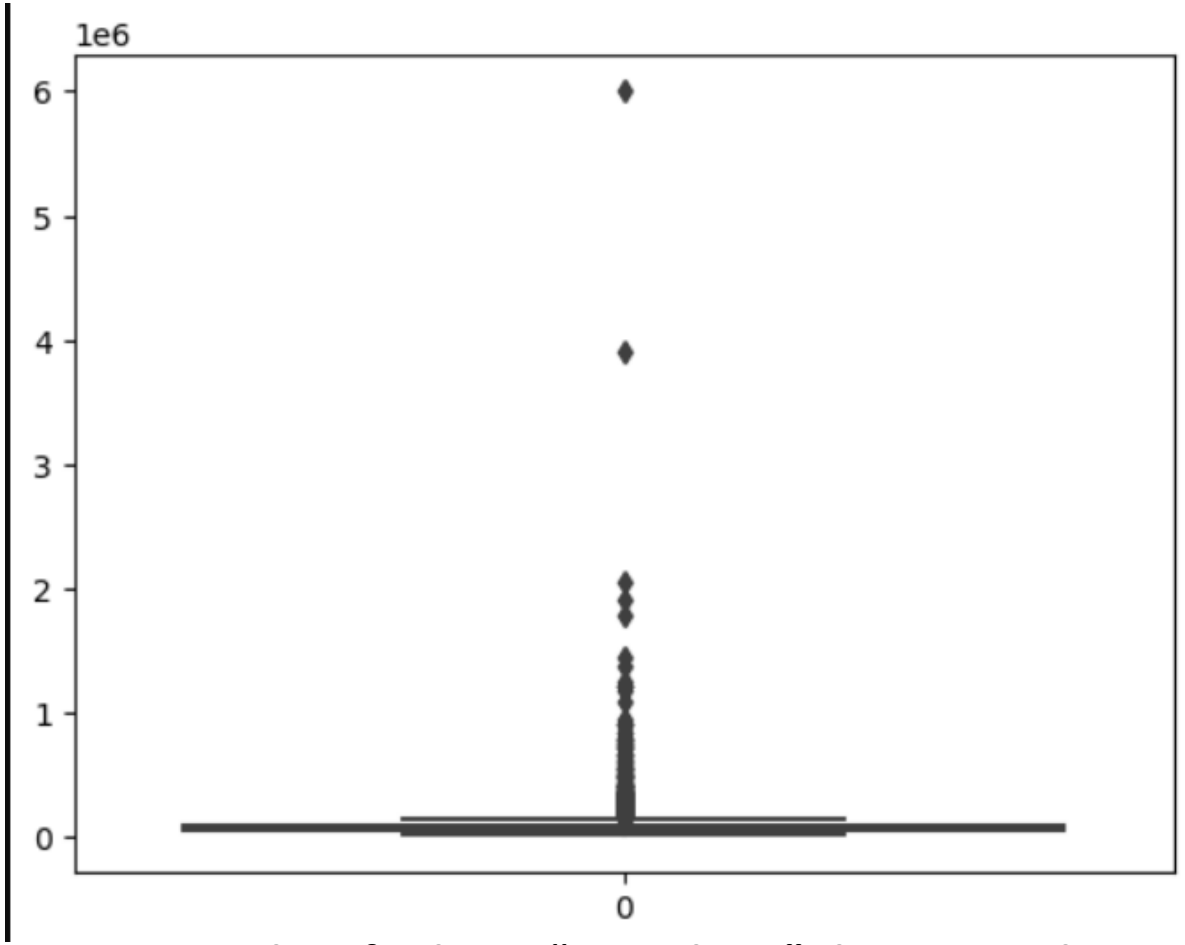

Fig: Boxplot of column “annual_inc” showing outliers

Univariate analysis-checking for outliers

- Figure shows the boxplot of variables ‘int_rate” and “dti”
- Outliers are not observed in this plot

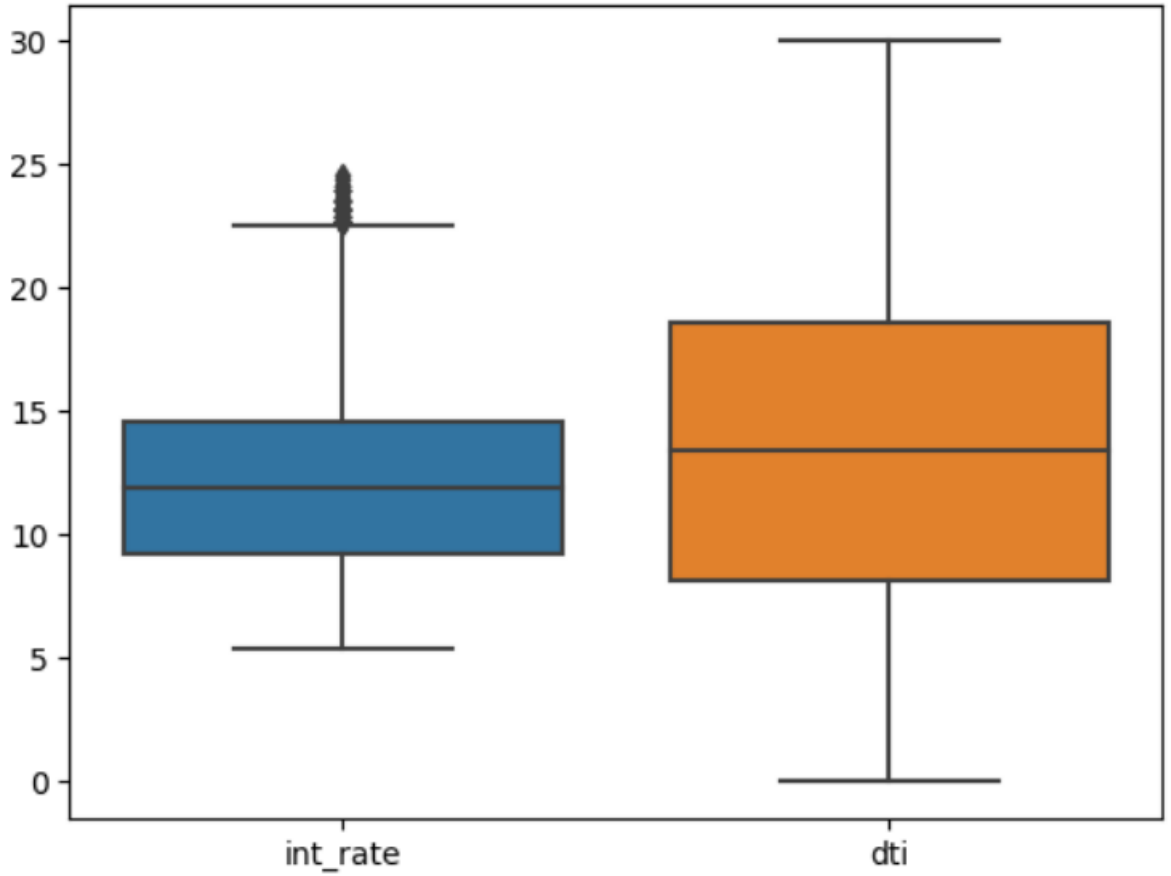

Fig: Boxplot of columns interest rate and dti ratio

Bivariate analysis

- Figure shows the barplot of variables ‘int_rate” and “loan_status”

- It can be observed that loans with higher interest rates(>12%), have higher possibilities of getting charged off.

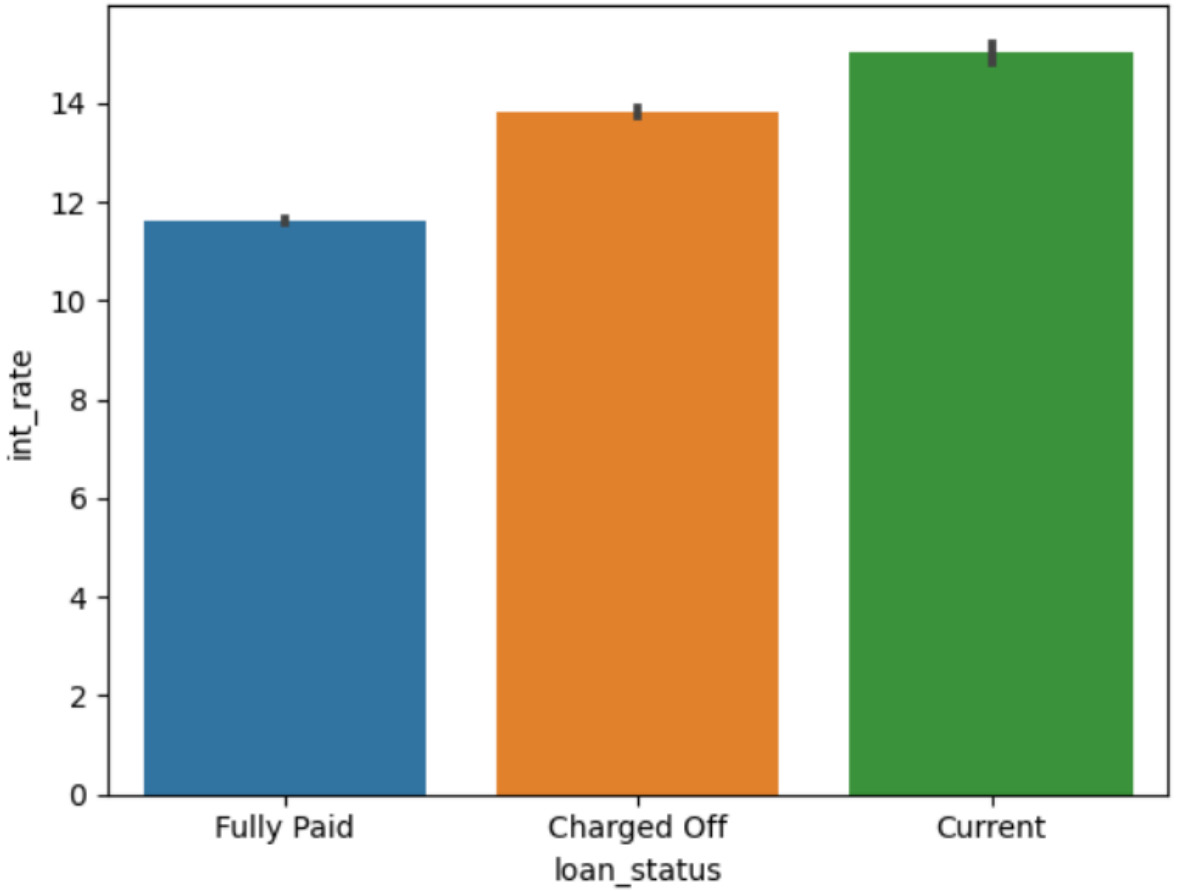

Fig: Barplot of columns interest rate and loan status

Bivariate analysis

- Figure shows the boxplot of variables ‘dti” and “loan_status”
- The dti ratio does not seem to have any significant impact on the loan status

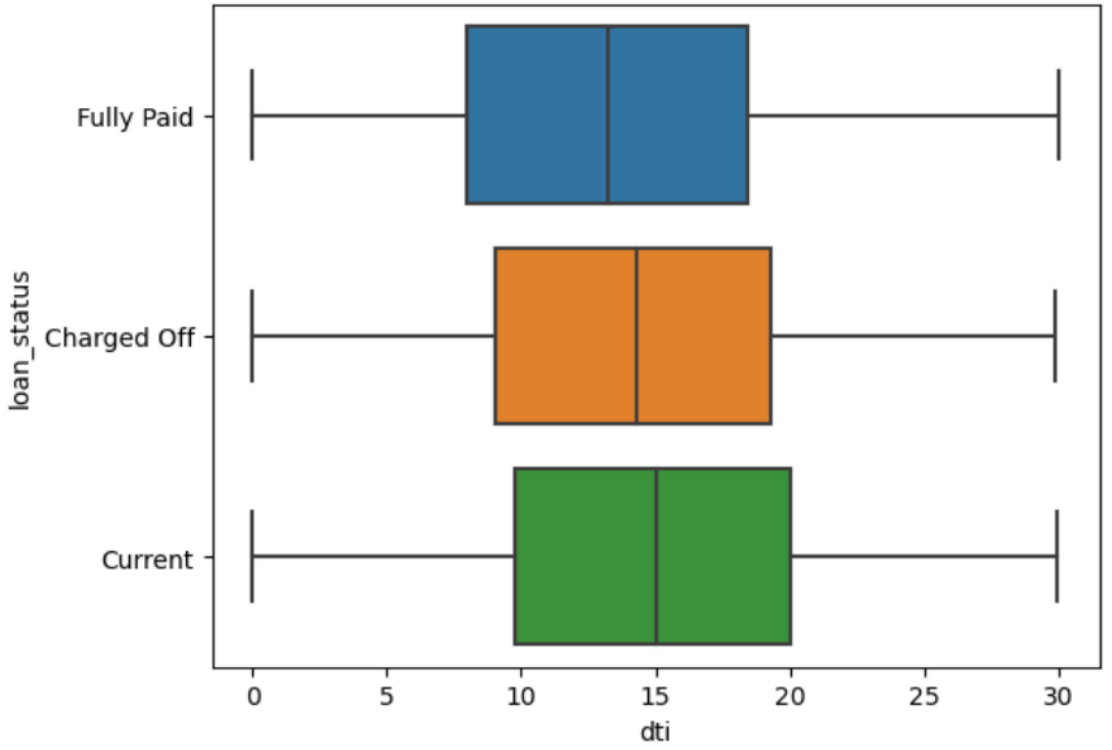

Fig: Boxplot of columns dti ratio and loan status

Bivariate analysis

- Figure shows the barplot of variables ‘annual_inc” and “loan_status”

- It can be observed that customers with annual income greater than 60000 have lower chances of defaulting on their loans

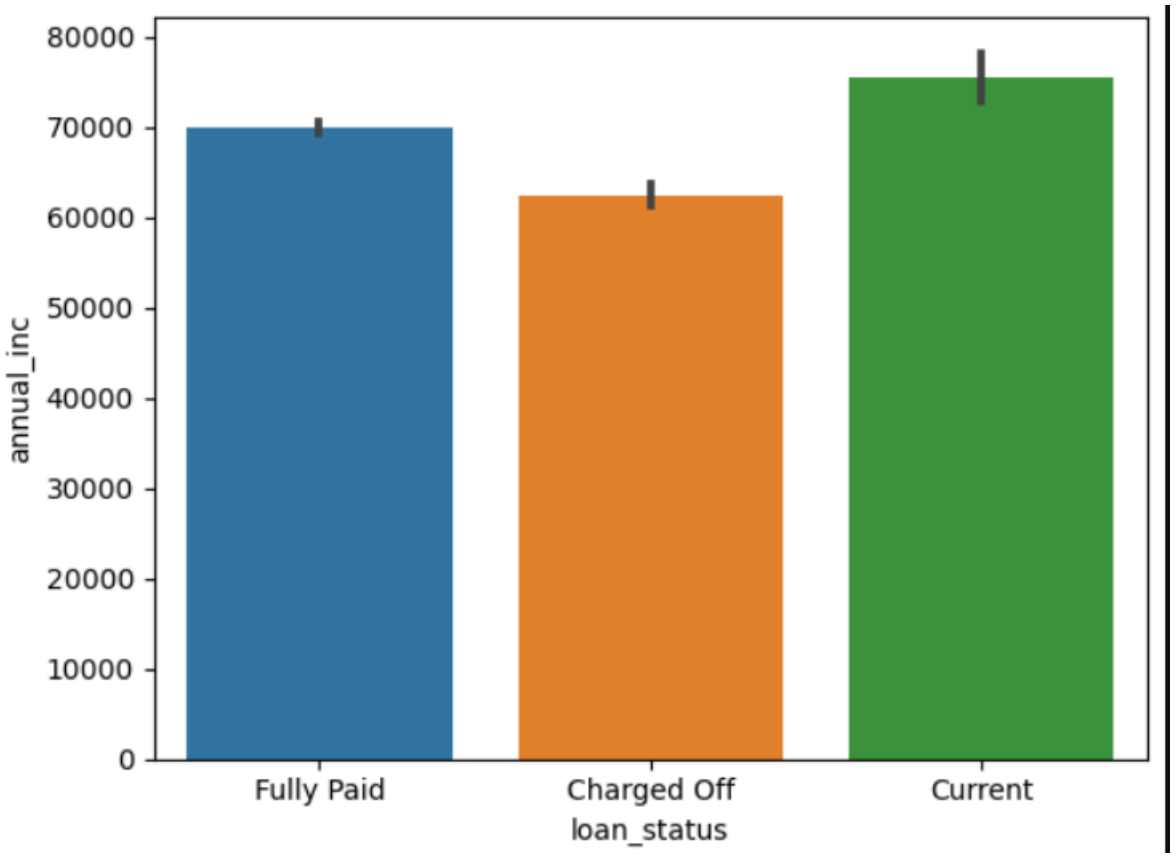

Fig: Barplot of columns annual income and loan status

Bivariate analysis

- Figure shows the barplot of variables ‘purpose” and “loan_status”
- The column “purpose” provides information on the purpose of taking a loan by respective customer
- “Debt_consolidation” being more in number, show higher number of loan defaults

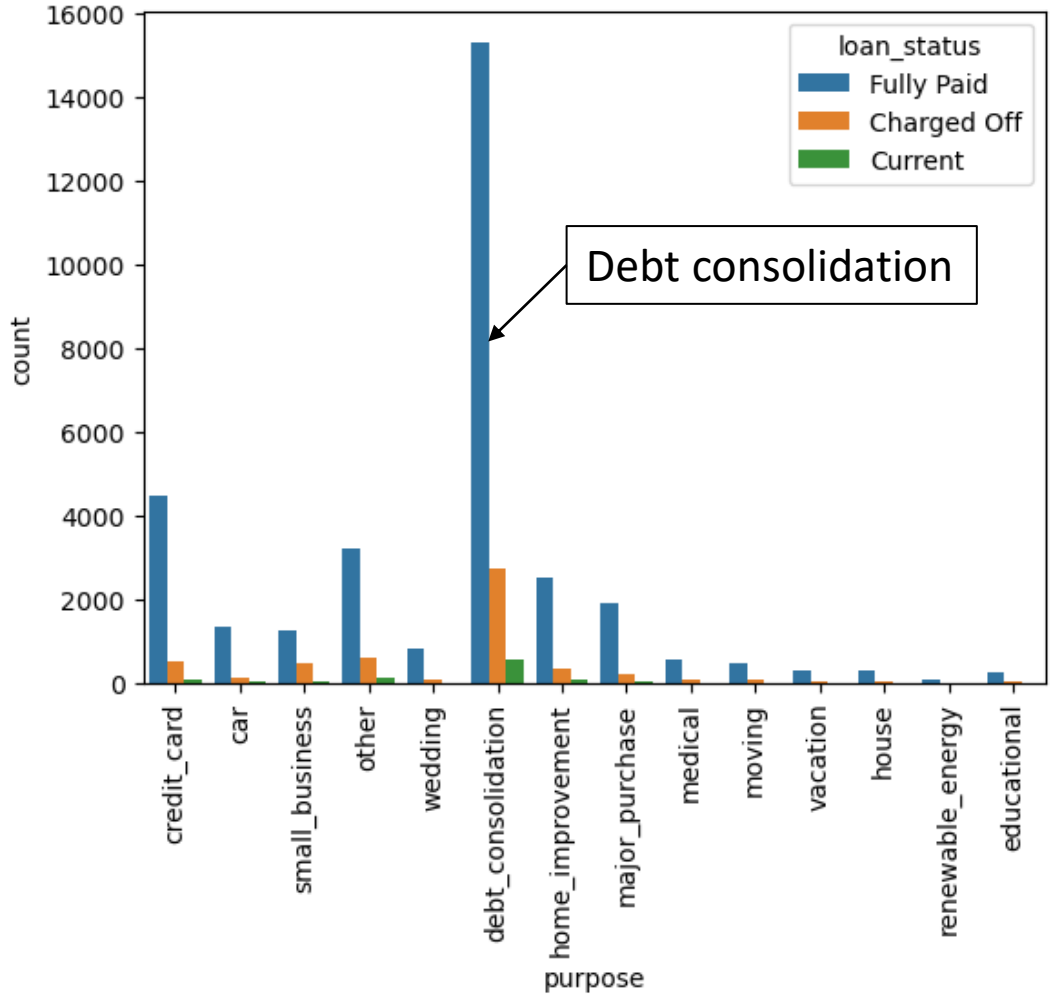

Fig: Barplot of columns “purpose” and loan status

Multivariate analysis

- Figure shows the barplot of variables ‘int_rate” ,“loan_status” and “home_ownership”

- It can be observed higher interest rates(>12%) irrespective of "home_ownership" seems to be driving loan defaults

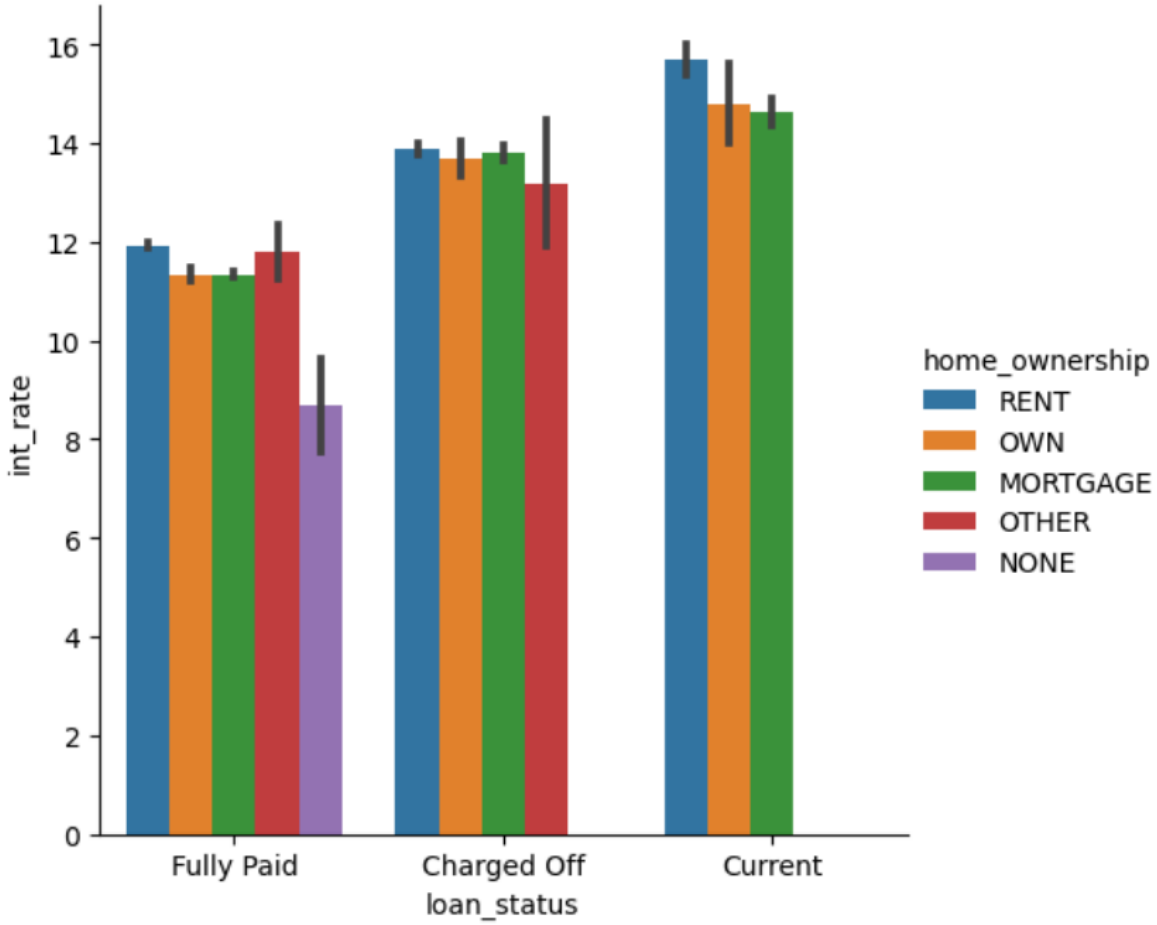

Fig: Barplot of columns int_rate, home_ownership and loan status

Multivariate analysis

- Figure shows the barplot of variables ‘int_rate” ,“loan_status” and loan grade

- The grade of the loan does not have any impact on loan defaults

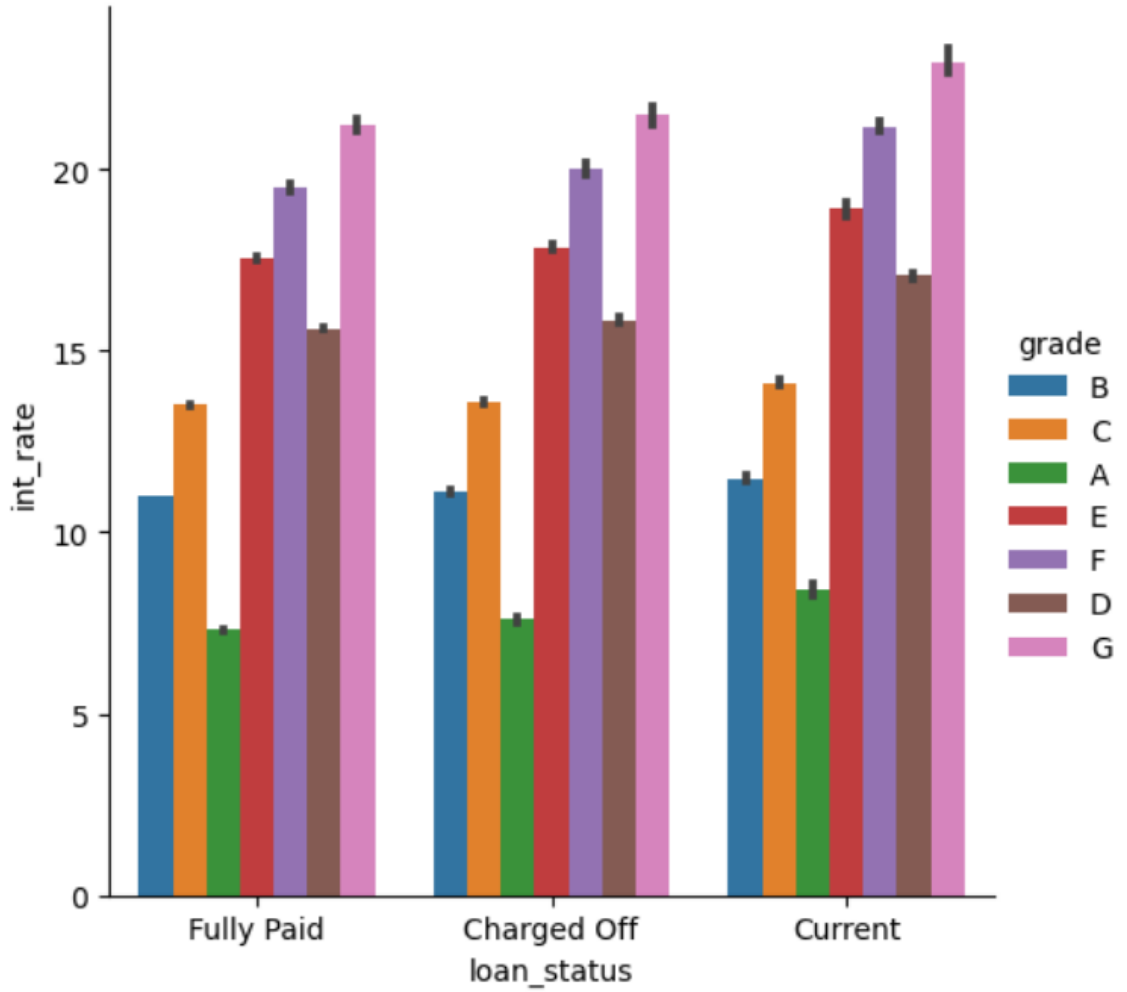

Fig: Fig: Barplot of columns int_rate, loan grade and loan status

Multivariate analysis

- Figure shows the barplot of variables ‘int_rate” ,“purpose” and loan status
- It is observed that the category of housing loan tends have higher chances of being charged off compared to other categories as the interest rate increases beyond 12-14%
- This assessment requires further investigation

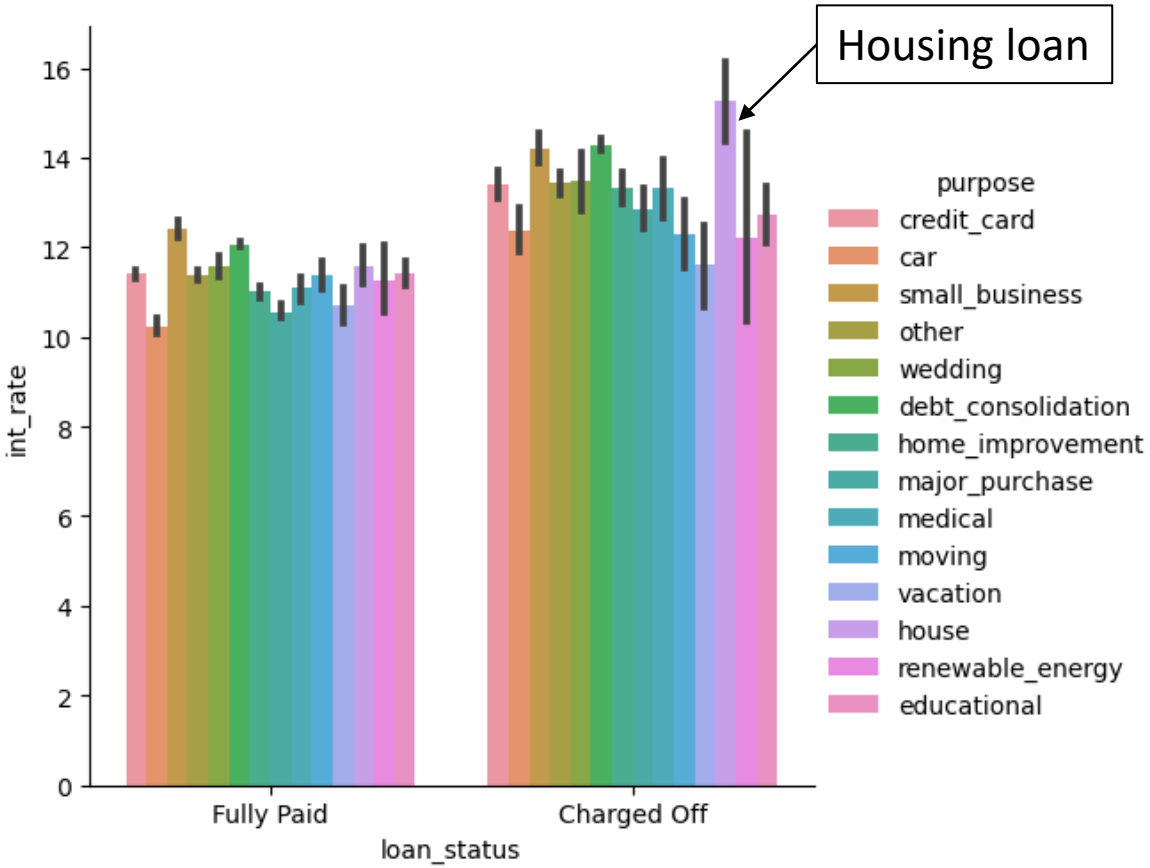

Fig: Fig: Barplot of columns int_rate, purpose and loan status

Multivariate analysis

- Figure shows the pivot table of variables ‘int_rate” ,“purpose” and loan status

- From the pivot table, loans such as- "housing", small_business“ and "debt_consolidation" are at a higher risk of being charged off compared to other categories, if their interest rates are more than 14%

loan_status	Charged Off	Fully Paid
purpose		
car	12.372813	10.229178
credit_card	13.405000	11.408515
debt_consolidation	14.275598	12.056410
educational	12.724107	11.427398
home_improvement	13.304006	11.011341
house	15.257966	11.570812
major_purchase	12.860495	10.559061
medical	13.322642	11.083791
moving	12.307717	11.384277
other	13.434202	11.379220
renewable_energy	12.221053	11.257590
small_business	14.203747	12.410188
vacation	11.610189	10.699534
wedding	13.488125	11.582867

Fig: Pivot table of columns int_rate, purpose and loan status

Multivariate analysis

- Here an additional column “percentage of defaults” has been added for each loan purpose

- It is observed that the category “small_business” has got the highest proportion of being charged off and second to it are “renewable energy” and “educational”.

Sl.no	Loan purpose	Count	Percentage of defaults
1	small_business	1828	25.98
2	renewable_energy	103	18.44
3	educational	325	17.23
4	other	3993	15.85
5	moving	583	15.78
6	house	381	15.48
7	medical	693	15.29
8	debt_consolidation	18641	14.84
9	vacation	381	13.91
10	home_improvement	2976	11.65
11	credit_card	5130	10.56
12	car	1549	10.32
13	major_purchase	2187	10.15
14	wedding	947	10.13

Fig: Table of loan purpose and percentage of defaults

Conclusion

- Loans with interest rates greater than 12% have higher chances of being charged off
- Customers with annual income greater than 60000, have fewer chances of defaulting on their loans
- “Housing", small_business“ and "debt_consolidation“ loans are at a higher risk of being charged off compared to other categories, if their interest rates are more than 14%
- It is observed that the category “small_business” has got the highest proportion of being charged off