# LENDING CLUB CASE STUDY

# SUBMISSION

Group facilitator: Budhaditya Saha

Group member: Manohar Simons

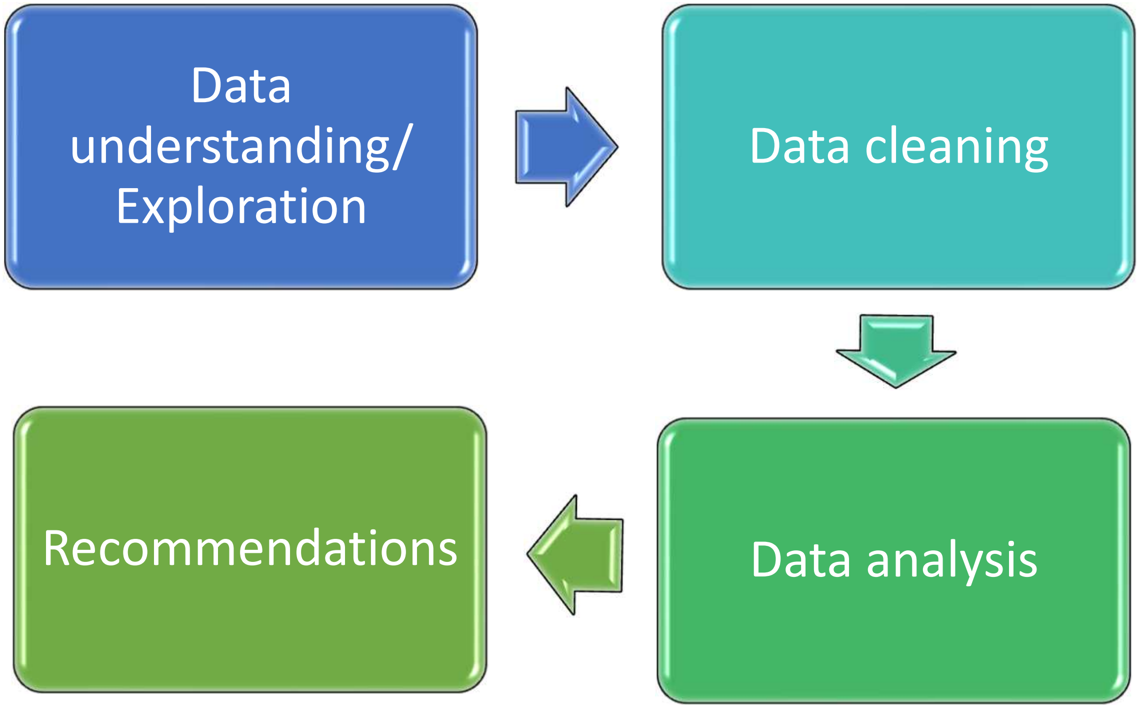
# Problem Statement

- When the Lending Club receives a loan application, it has to make a decision for loan approval based on the applicant's profile. Two **types of risks** are associated with the LC's decision:
- If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
- If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

# Analysis approach

- The given data contains the information about past loan applicants and whether they ‘defaulted’ or not.
- The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.
- We will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

# Problem solving methodology



Data understanding/ Exploration

Data cleaning

Data analysis

Recommendations

# Data exploration

- ❑There are 39717 rows and 111 columns in the given dataset loan.csv.
- ❑Identified the data types of the columns and observed that interest rate and employment length are of type “object”. This would limit the numerical univariate and bivariate analysis and hence these columns are converted to numerical values.
- ❑The target column is the loan status. We retain the fully paid and charged off loans and drop loans that have a current status.
- ❑Convert fully paid and charged off loans to 0's an 1’s respectively.

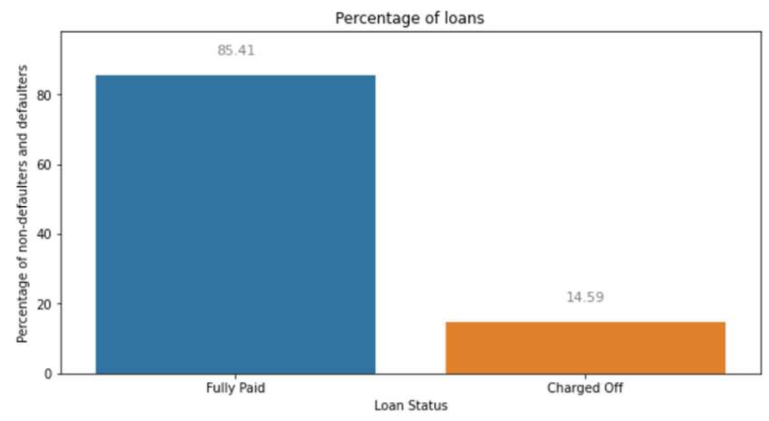
# Data cleaning

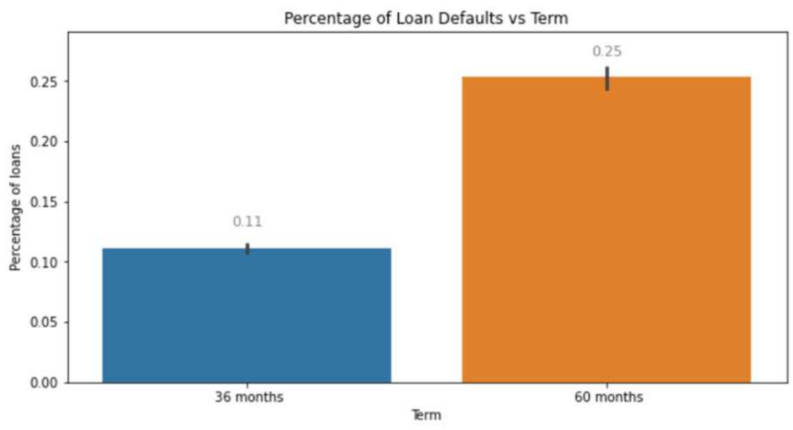
- ❑There are 54 columns that do not have any data and has been dropped from the dataset.
- ❑All columns with “0” value has been dropped.
- ❑All redundant columns has been dropped.
- ❑Identified columns which are not relevant for the analysis and has been dropped.
- ❑The customer behavioral variables are not available at the time of loan application and thus they cannot be used as predictors for credit approval. There are 20 such columns which has been dropped from the data set.
- ❑The employment length column has 2.67% n/a values. Also all but 18 rows that have employment length as n/a also have the employer title as blank. Which probably implies that these applicants are not employed and we can treat these values as 0's.

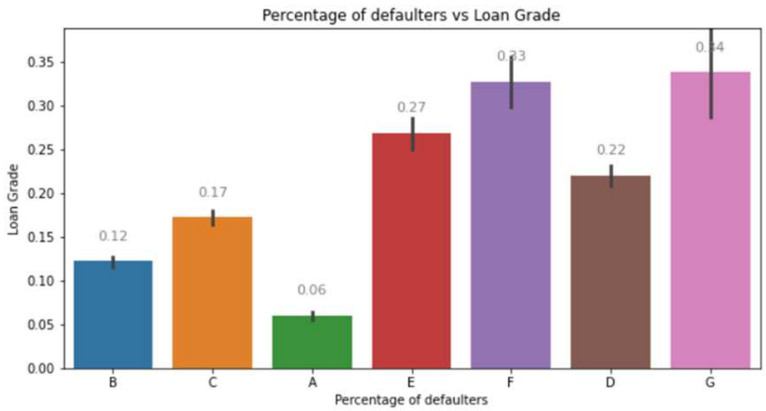
# Probable predictors of default

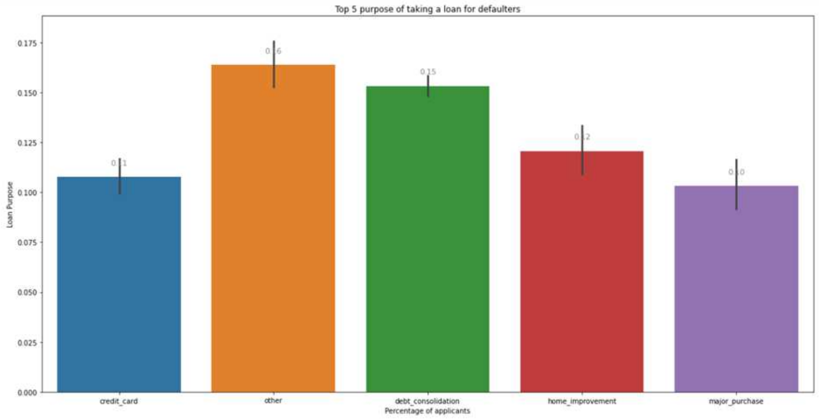
- Loan amount
- Term : 36 months & 60 months
- Interest rate
- Grade & subgrade
- Annual income
- Purpose of the loan
- Verification status
- Loan date
- Debt to Income (DTI)
- Employment length
- Home ownership

# Plots – Categorical univariate

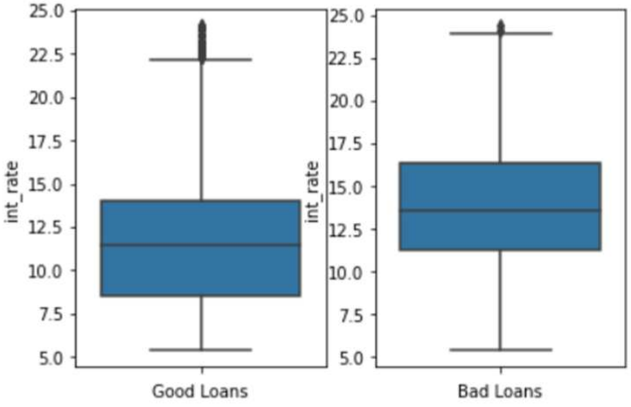


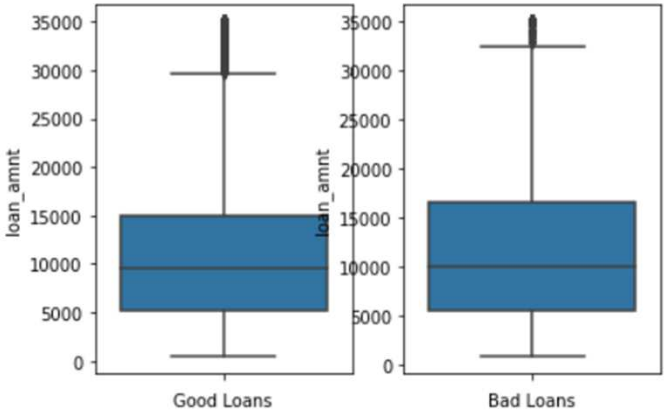


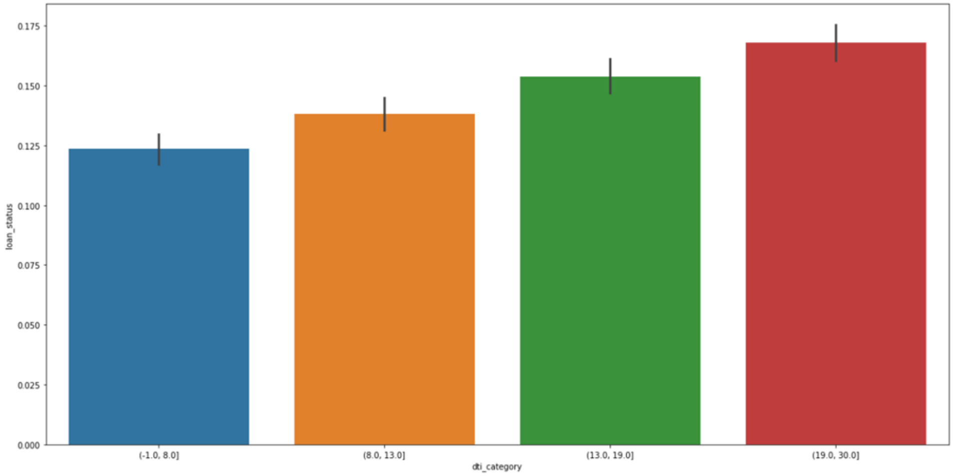


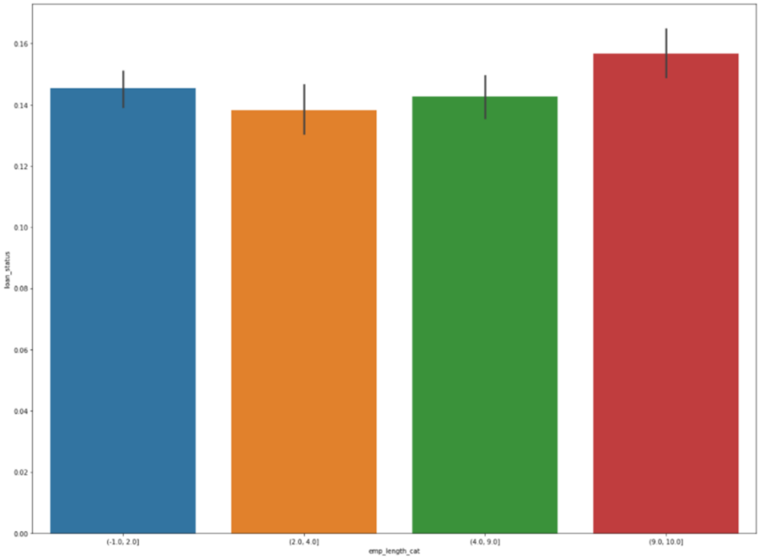


# Plots – Numerical univariate







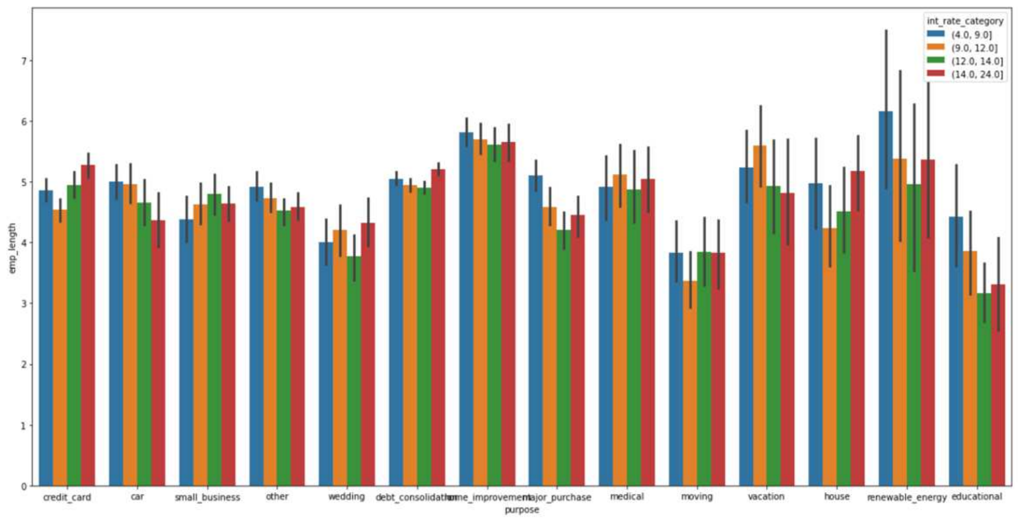


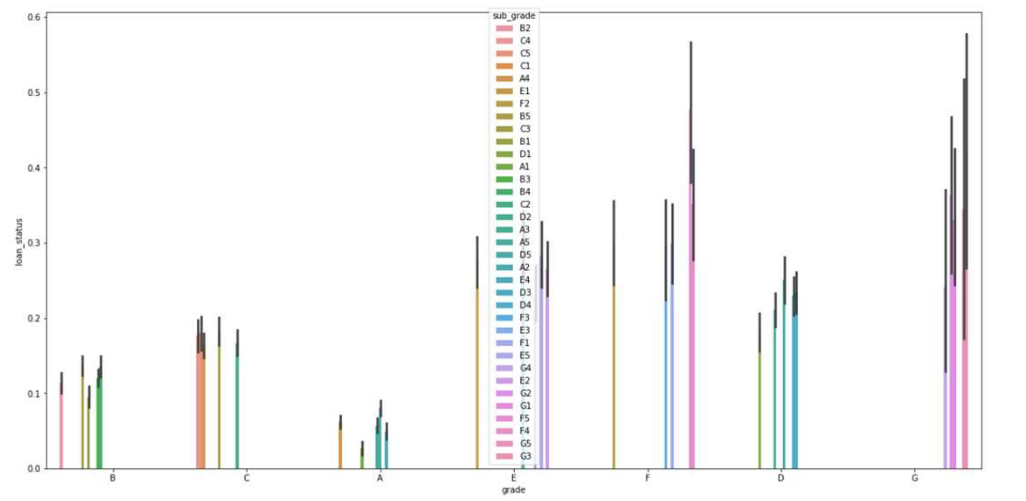
# Univariate Analysis

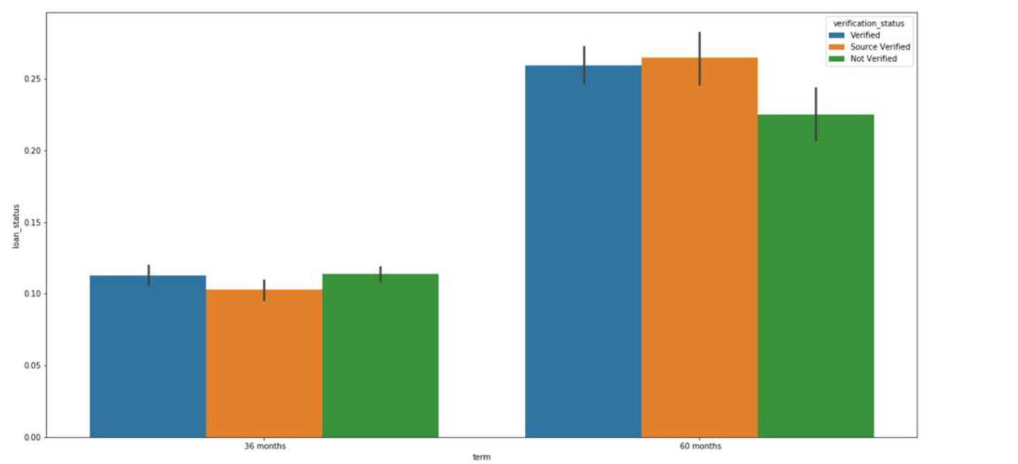
Analysis of the categorical variables indicate the following:

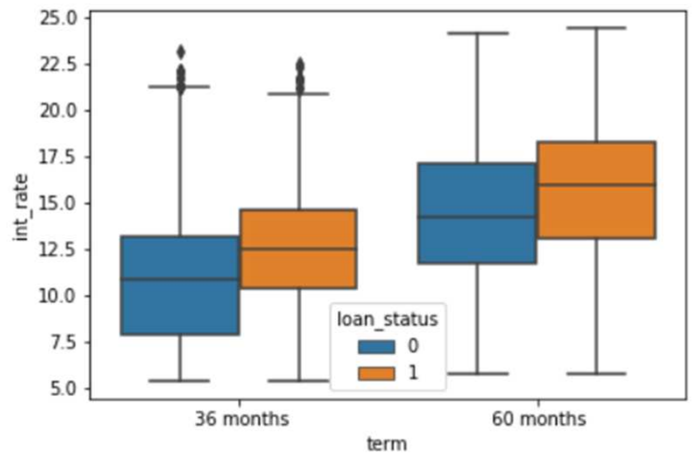
1. Fully paid loan constitute 85% of the total dataset and 15% loan has defaulted.
2. People taken loan for longer term i.e. 60 months has defaulted more than those of shorter term i.e. 36 months by 56%. The information content is 0.56.
3. The variance in the defaulted rate across the different grades is 82% where Grade A has defaulted for 6% (lowest) and Grade G has defaulted for 34% (highest). Hence Grade can be considered as one of the most important predictors as the information content is high at 0.82.
4. The categorization of the purpose of the loan indicates that the top 5 reasons to apply for loan are: debt consolidation; credit card; other; home improvement; major purchase. However the most defaulted is not falling within the “top 5 loan purpose”. Loans taken for the purpose of a small business have the highest rate of default (27%) and the lowest being wedding (10%). Hence the IC is 0.63.
5. Loans that are defaulted have a higher rate of interest than loans that are re-payed. The higher the interest rate the greater is the probability for default. Int rate for the range 14-24% has defaulted for 22% while for interest rate of range 4-9% has defaulted for 6%. IC is 0.72.
6. The principal loan amount does not seem to be a predictor of a default as the IC is 0.22 which is low.
7. People have multiple loans tend to default more with the minimum defaulters being 12.5% and the maximum defaulters being 17%. Hence the IC is 0.26.
8. People with a lower annual income tend to default the most. 18% of people in lowest income bracket becomes the defaulter and 12.5% of the people in the highest income bracket becomes the defaulter. The IC is 0.3.
9. People who are in the employment bracket 9-10+ years have a strong tendency to default. Around 15% of the people from the highest bracket will default. However the IC is low (0.12) and the employment bracket is not a strong predictor for the default.

# Plots – Bivariate

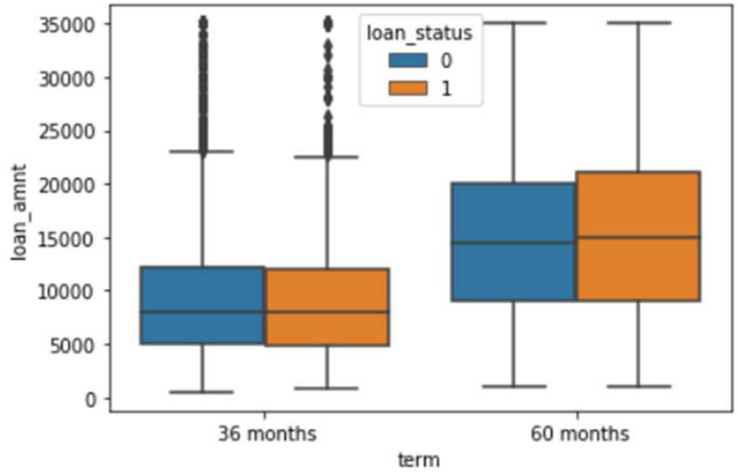


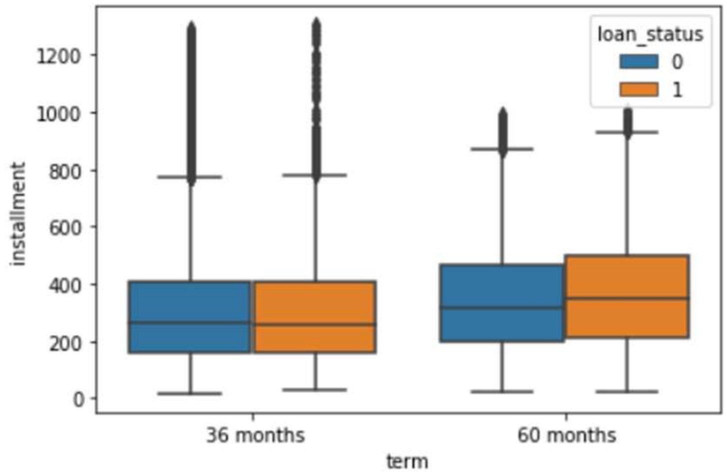


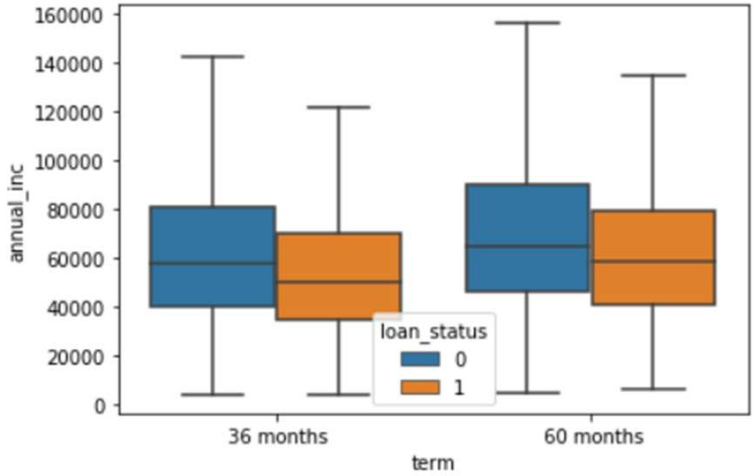


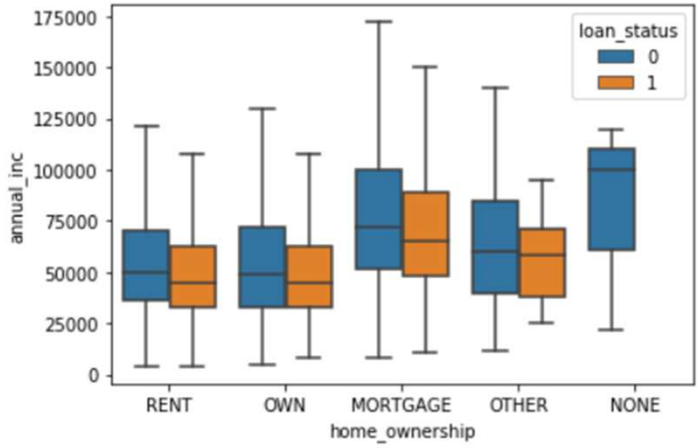


# Plots – Bivariate









# Bivariate analysis

Analysis of the categorical variables indicate the following:

- Verification status in not good predictor for defaulters. For loan term of 36 months the % of people defaulted for verified and non-verified is the same which is 11% respectively. For loan term of 60 months the % of people defaulted the most is from "source verified" category around 26%.
- Loans that are defaulted have a higher rate of interest than loans that are re-payed.
- People paying higher installment on longer term loans have a larger tendency to default.
- People who take shorter term loans tend to have a higher annual incomes than those who take longer term loans across all home ownership categories.
- The purpose of the loans are not dependent on the interest rates, people tend to apply for all the various types of loans irrespective of the different interest rates. However it has been observed that more the no. of years of employment the top three purpose for taking loans linked to the interest rates - renewable energy, improvements and vacation.
- The purpose of the loans are not dependent on the grades, people tend to apply for all the various types of loans for all the different grades. However it has been observed that more the no. of years of employment the top three purpose for taking loans linked to the different grades - renewable energy, house and improvements.

# Key Predictors

| S.No | PREDICTOR | TYPE | IMPACT (std. deviation) |
|---|---|---|---|
| 1 | Sub Grade | categorical | 0.11 |
| 2 | Term | categorical | 0.1 |
| 3 | Grade | categorical | 0.1 |
| 4 | Address | categorical | 0.08 |
| 5 | Home ownership | categorical | 0.07 |
| 6 | Purpose of Loan | categorical | 0.04 |
| 7 | Employment years | categorical | 0.03 |
| 8 | Verification status | categorical | 0.02 |
| 9 | Interest rate | numerical | 0.08 |
| 10 | Loan amount | numerical | 0.03 |
| 11 | Funded amount | numerical | 0.03 |
| 12 | Annual Income | numerical | 0.03 |
| 13 | Investor funded amount | numerical | 0.02 |
| 14 | No. of installments | numerical | 0.01 |

# Most important predictors

Considering the cut-off threshold for the Impact as 0.08 we can conclude that the most important predictors are the ones listed below:

| S.No | PREDICTOR | TYPE | IMPACT (std. deviation) |
|---|---|---|---|
| 1 | Sub Grade | categorical | 0.11 |
| 2 | Term | categorical | 0.1 |
| 3 | Grade | categorical | 0.1 |
| 4 | Address | categorical | 0.08 |
| 9 | Interest rate | numerical | 0.08 |

# Recommendations

Following are the key recommendations to the LC on a business model to review and approve or reject a loan application:

1. The key predictors are identified by doing the various analysis from the historical data shared by the Lending Club and those predictors are to be considered while reviewing a loan application from an applicant. A threshold will be defined for the list of the key predictors based on which the most important predictor will be identified. Any values from the loan application which falls below the threshold for all the important predictors has to be rejected to reduce the number of bad loans.
2. A mathematical function can be defined to assign a numeric value to all the key predictors in the loan application form based on the information provided.
3. The threshold to approve a loan can be earmarked as 0.08 based on the analysis performed on the given data set.
4. Using this threshold we can conclude that the most important predictors to be catered for are
   - i. **Grade**
   - ii. **Sub-grade**
   - iii. **Term**
   - iv. **Address**
   - v. **Interest rate**
5. The numerical values obtained in step 2 will be mapped against the threshold and for any application in which the important predictors are having less value than the threshold can be safely rejected to reduce the number of bad loans.