Lending Club Case Study

EDA by Ramesh Velivela

ML61 EPGP in AI & ML

Agenda

- Introduction
- Problem Statement & Business Objective
- Primary Goals & Approach
- Data Handling/Prep for EDA
- Exploratory Data Analysis (EDA)
- Conclusion with Recommendation

Introduction

- **Contributor**: **Ramesh Velivela**
- **Topic**: **EDA**
- **Subject area**: **Lending Club Case Study**
- **Introduction**:
  - We work for a consumer finance company which specializes in lending various types of loans to urban customers. **Company has to make a decision for loan approval based on the applicant's profile**.
  - Company is **looking for detailed insights on various dimensions / driving factors causing the loan default** (strong indicators of default), which they can **utilize during portfolio planning and risk assessment** by their underwriters.

Problem Statement & Business Objective

- **Problem Statement:**
  - This company is the **largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures**. Borrowers can easily access lower interest rate loans through a fast online interface.
  - **Lending loans to 'risky' applicants is the largest source of financial loss** (called credit loss). **Credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed**.
  - Borrowers who **default** cause the largest amount of loss to the lenders. In this case, the **customers labelled** as '**charged-off**' are the **'defaulters**'.
  - Problem is to **avoid or mitigate risk** for lending company while accepting loan applications; **minimize "Credit Loss".**
- **Business Objective:**
  - **Generate** valuable **insights** into the complex dynamics **of loan default risk**, enable **Financial Institution to make informed decisions** and **minimize potential losses**.

Primary Goals & Approach

- **Primary Goals**:
  - The company wish to understand the **driving factors (or driver variables) behind loan default**, i.e. the variables which are **strong indicators of default**.
  - The company wish to utilize this knowledge for its portfolio and risk assessment.
- **Approach**:
  - Provided Source Data as **loan.csv** which contains the **complete loan data for all loans issued through the time period 2007 to 2011.**
  - "**Know your Data**" by going through the **Data Dictionary** spreadsheet to get awareness on the attributes and their values.
  - Use **Python** and various other libraries like **Numpy**, **Pandas**, **Matplotlib**, **Seaborn** to **load, cleanse, standardize the inbound data** and make it ready for EDA – Uni / Bi / Multi variate analysis. Generate various graphs on multiple variable of data set to present observations.

Data Handling/Prep for EDA

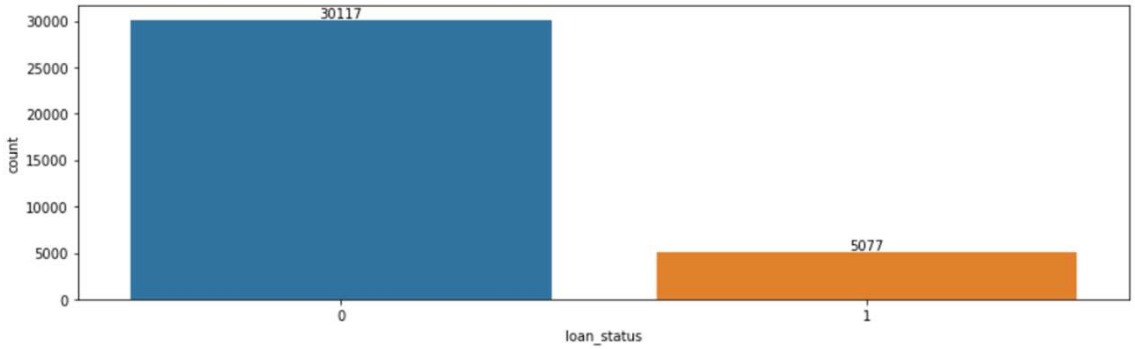
Data Loading followed by Data Preparation for Analysis

- **Dataset Overview:**
  - The source dataset is containing loan data for all loans **issued** through the 2007–2011
  - This includes the current **loan status** (**Current**, **Charged-off**, **Fully Paid**) and latest payment information.
  - **Additional attributes** include credit scores, number of finance inquiries, and collections among others.
  - Default shape of the dataset is about **39 thousand rows and 111 columns**.
- **As a first step**, **Load Dataset** using Python
- **Perform Data Cleaning & Data Handling**
  - Remove complete NULL columns, irrelevant columns to identify Risk Indicators
  - Remove rows pertaining to fully paid loans as no risk with such paid loans
  - Handle Missing Values (remove null rows),
  - Standardize Data (remove % from Rate of Interest and other column values, move "months" word from term variable), Find Outliers & Treat.
  - Encode Loan Status as Charged Off as 1 and Fully Paid as 0
- **Add** required **derived columns** like "month" and "year" values for DATE type variables
- **Add Bins on required columns with required ranges – i.e., Annual Income Range, Loan Amount Range, DTI Range, and Interest Rate Range.**
- **As part of EDA, perform Univariate Analysis** (Categorical/Continuous Features), **Bivariate Analysis** (Box Plots), **Multivariate Analysis** (Correlation Heatmaps)

Exploratory Data Analysis (EDA)

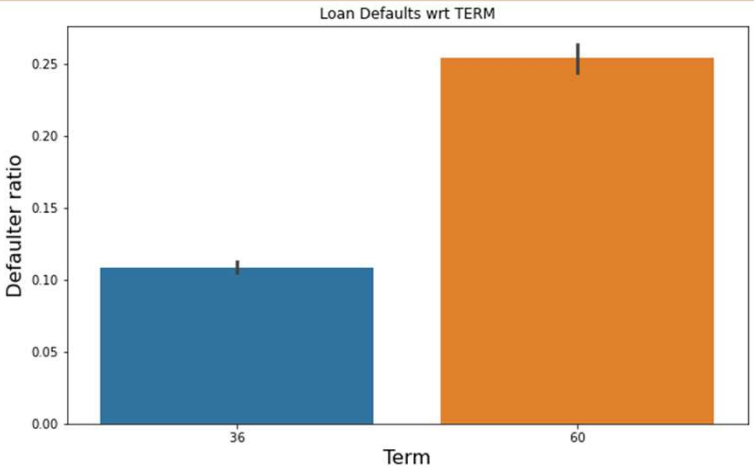
Univariate Analysis - 1

Distribution of Loans by Loan Status



Observation:
Defaulted loans a.k.a **Charged Off loans are less when compared to Fully paid loans**

Defaulters group by loan term



Observation:
When compared to 36 months tenure, **60 months term loans are more defaulted**.

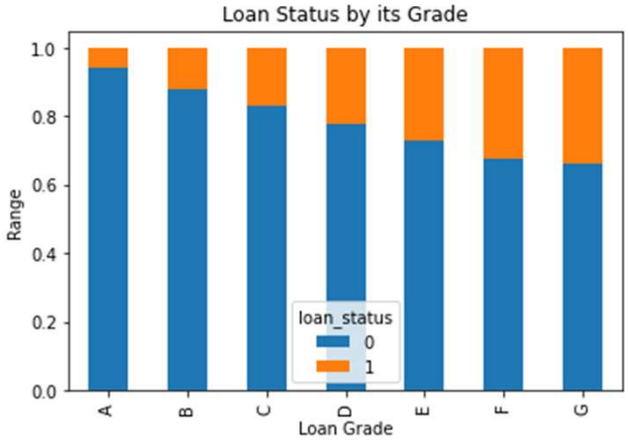
Exploratory Data Analysis (EDA)

Univariate Analysis - 2

Loan Status by Loan Grade

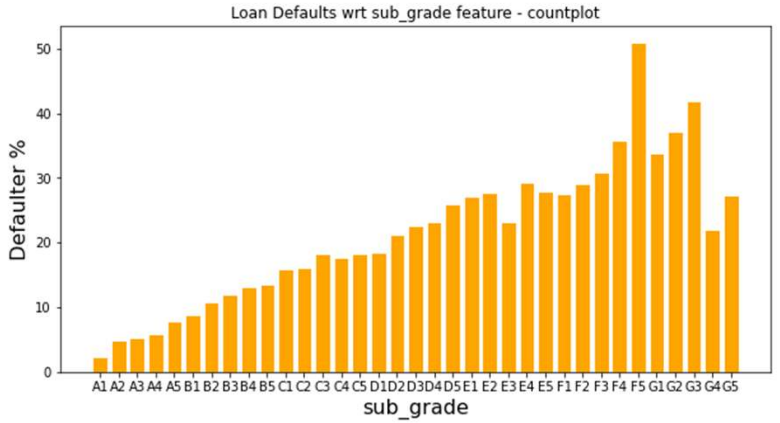
Observation:
Loans with **LOW grades** (towards grade A) **are resulting into highly probable fully paid loans.**

Loans with HIGH grades (towards grade **G**) **are highly potential as "Defaulted" loans.**



Defaulted loans (%) by "sub Grades"

Observation:
Loan Sub Grades dimension is a great indicative and **Defaulted loans are gradually increasing from A1 towards G5 sub grade.**

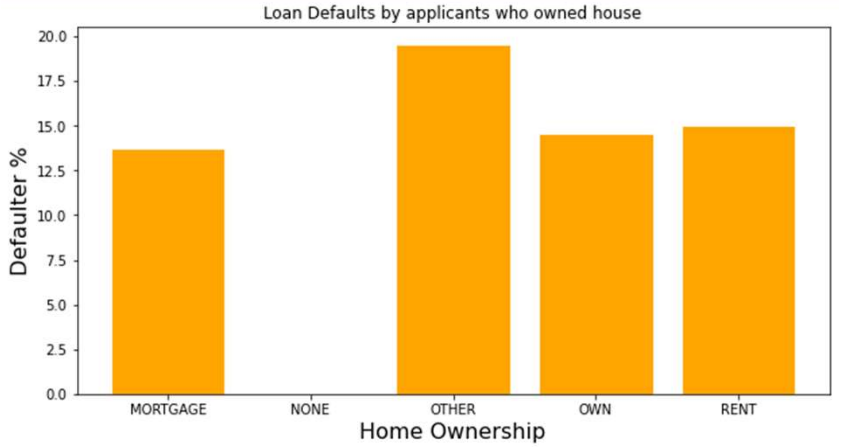


Exploratory Data Analysis (EDA)

Univariate Analysis - 3

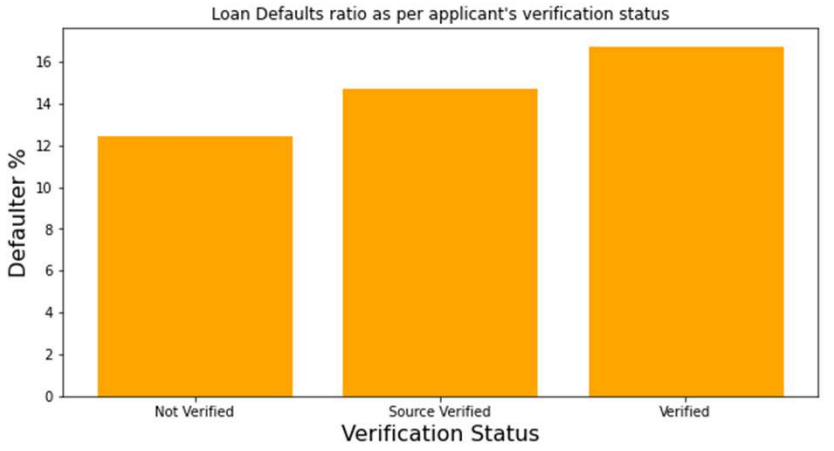
Defaulted loans (%) as per applicant's home ownership status

Observation:
Home ownership dimension is not helpful to find any risk trends as we observe defaulted loans in every category of home ownership



Defaulted loans (%) basing on applicant's Verification Status

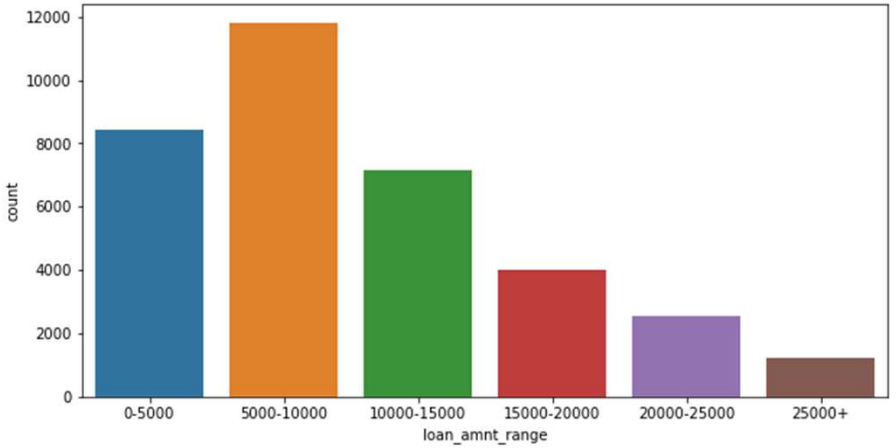
Observation: Similar to Home ownership dimension observation, Verification Status is also **not helpful to find any risk trends** as we observe defaulted loans in every category of this dimension.



Exploratory Data Analysis (EDA)

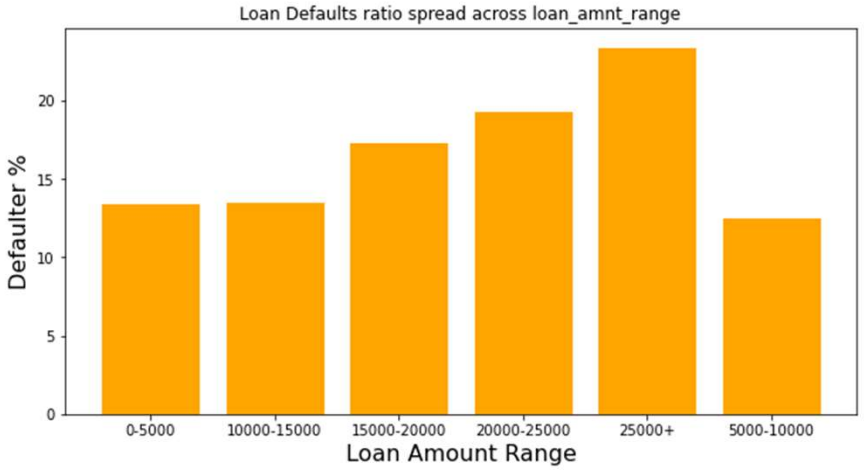
Univariate Analysis - 4

Loan amount range distribution and find any risk indicators



Observation:
Majority of the loans given in the range of 5k to 10k.
But, we observe defaulters are spread across every range, though 25k+ is dominated followed by 15k-20k and 20k to 25k range.

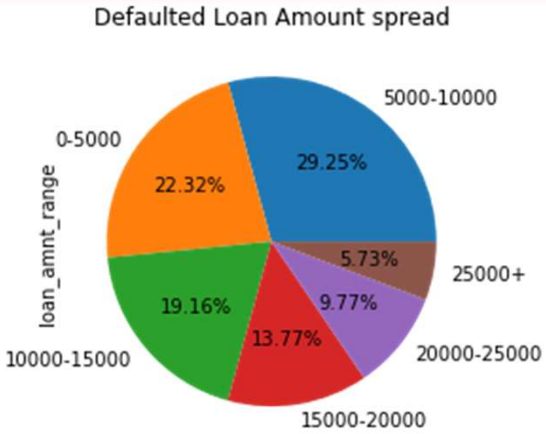
Hence, Loan Amount Range doesn't give any specific clear risk indicator



Exploratory Data Analysis (EDA)

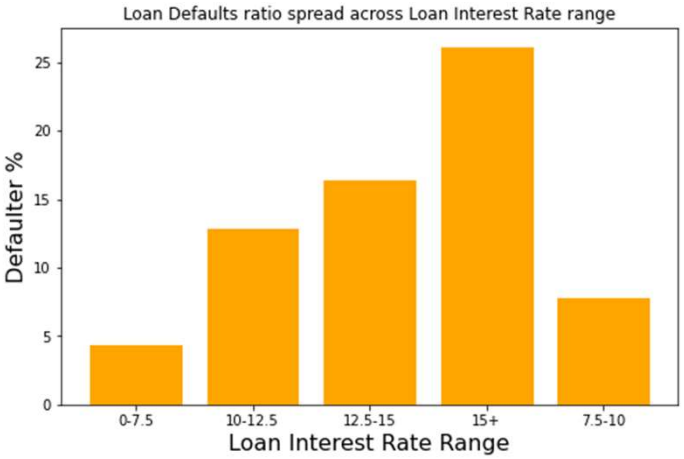
Univariate Analysis - 5

Defaulted Loans Amount Range



Observation:
Loan amount 5k-10k range is dominated followed by 0-5k and 10k-15k

Defaulted Loans spread across Loan Interest range



Observation:
Rate of Interest seems **directly proportional to get the loan defaulted as we observe highest rate of interest is having more number of defaulters**.

Lowest defaulters' loan interest rate is between 0 to 7.5, while highest is at 15+ rate of interest.

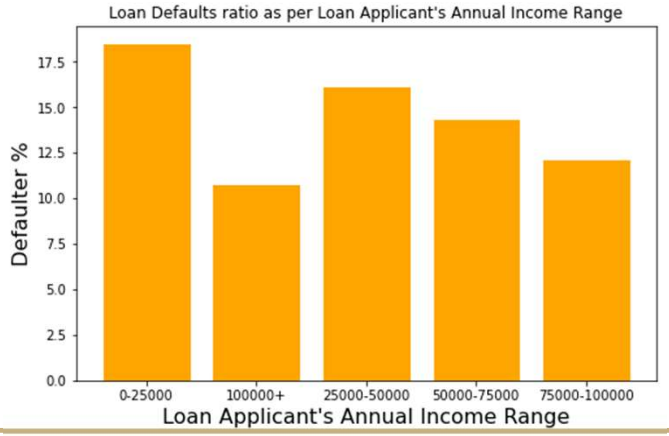
Largest chunk of defaulters interest rate is above 12.5.

Exploratory Data Analysis (EDA)

Univariate Analysis - 6

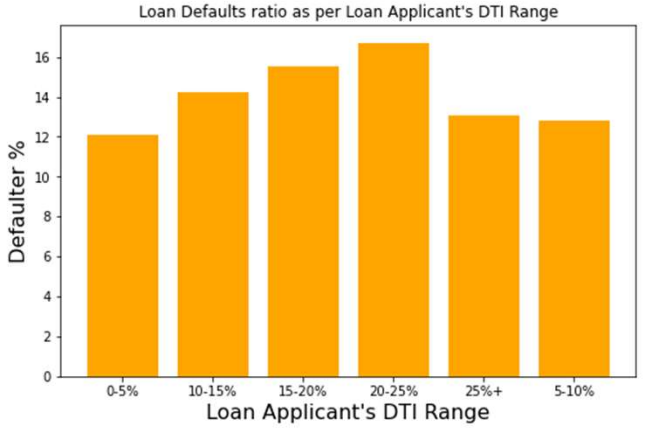
Defaulted Loans spread across various Annual Income Range

Observation: We have Defaulters spread across every range of applicant's annual income. So, **this annual income range dimension is not a dominating factor to realize defaulted loans**.
Not a potential risk indicator



Defaulted Loans spread across various DTI Range

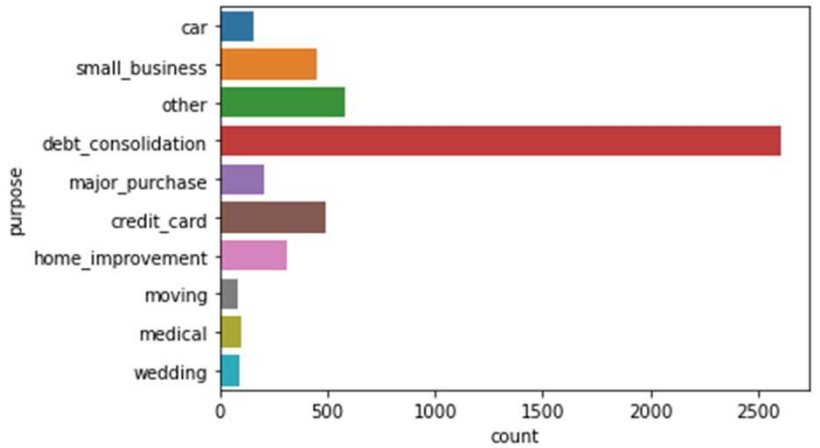
Observation: We have Defaulters spread across every range of applicant's DTI. So, **this DTI range dimension is not a dominating factor to realize defaulted loans**.
Not a potential risk indicator



Exploratory Data Analysis (EDA)

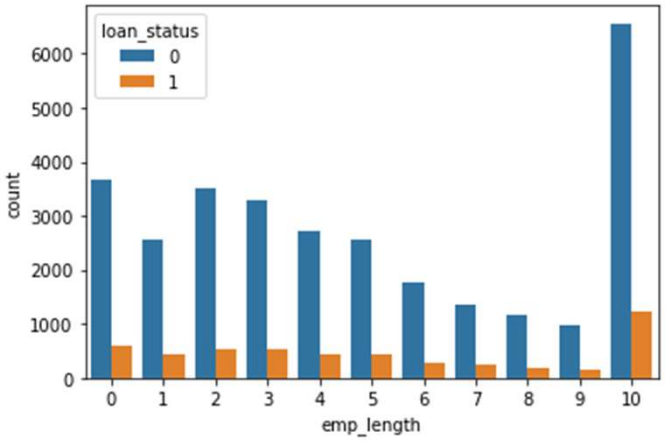
Univariate Analysis - 7

Defaulted Loans across various "loan purposes" being called out while applying the loan



Observation:
Very clear indication that **"DEBT CONSOLIDATION" purpose is dominated need for applying loans and the same is acting as major cause to have defaulted loans.**
So, this is an important risk indicator to secure from defaulted loans

Loan Status as per applicant's employment tenure



Observation:
Clearly evident that **10 and above experienced applicant's are becoming defaulters on their loans**.

So, Financial institutions need to observe closely before they approve the loan.

Exploratory Data Analysis (EDA)

Univariate Analysis - 8

Crosstab on Loan Grade by Loan Purpose

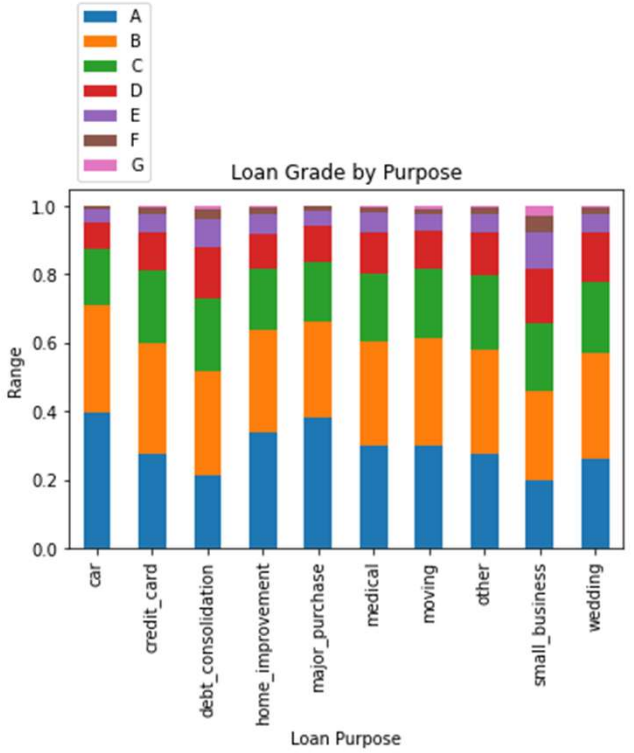
Observation:

This stacked bar chart depicts loan purpose and grade it belongs to.

- **Debt consolidation** need is the most dominant purpose to apply loans.
- Across any loan purpose, **majority is contributed from Grades A, B, and C**.
- **Loans for credit card payments is highest in the E and F loan grades**, while the proportion of loans for home improvement and **small business are highest in the G loan grade**.

These are very important risk factors to consider while underwriters are analyzing the loan application and taking the final decision.

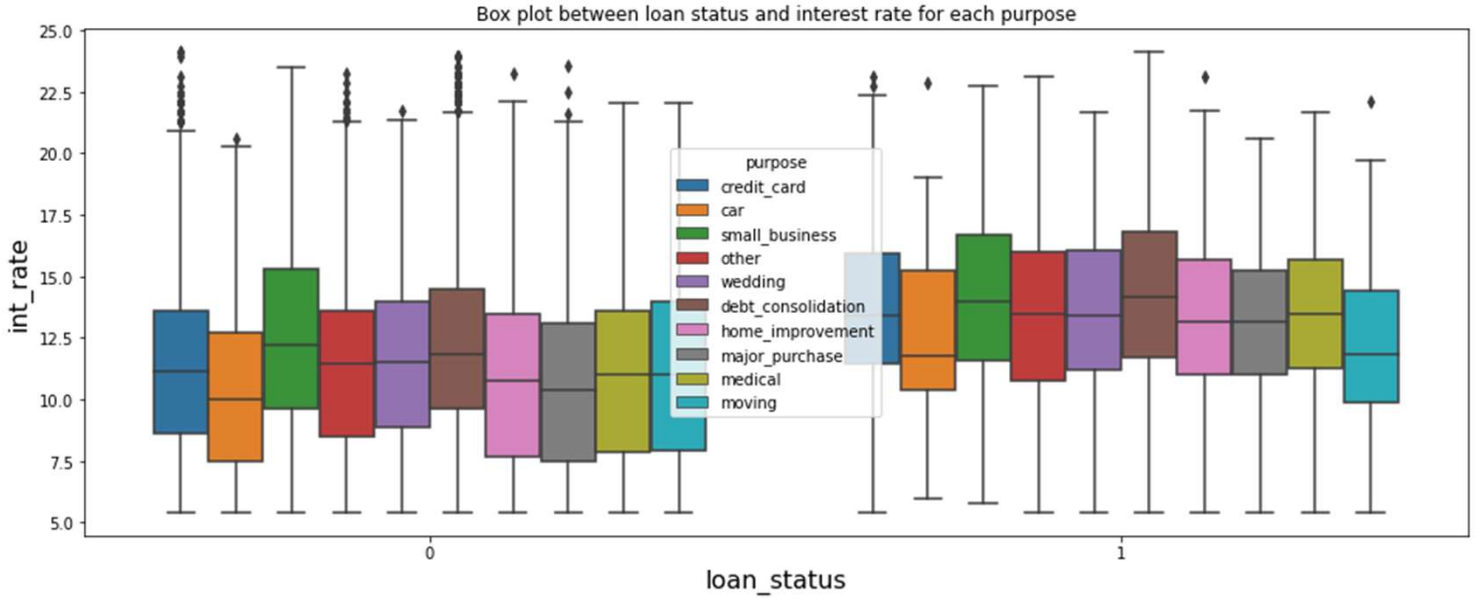
They may have to increase the interest rate to mitigate risks based on the purpose they apply for loan. But, this also might not be helpful to avoid Credit Losses.



Exploratory Data Analysis (EDA)

Bivariate Analysis - 1

Box plot between Loan Status and Interest Rate for each Loan Purpose



Observation:

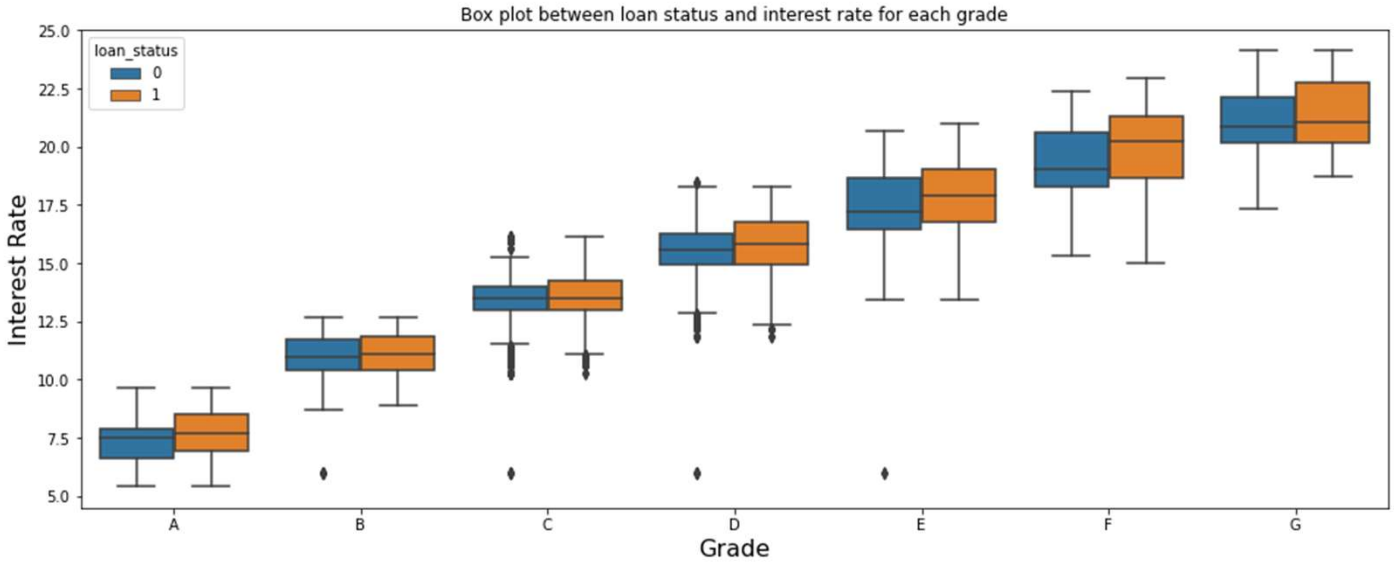
Most importantly, **If loan Interest rate increases, then corresponding loan may potentially get defaulted.** This scenario seems valid for any kind of loan purpose.

So, it is critical for companies not just simply increase the Rate of Interest (RoI) to mitigate the risk, which may deviate their hypothesis and lead towards facing high number of defaulted loans.

Exploratory Data Analysis (EDA)

Bivariate Analysis - 2

Box plot between Loan Status and Grade for each Interest Rate



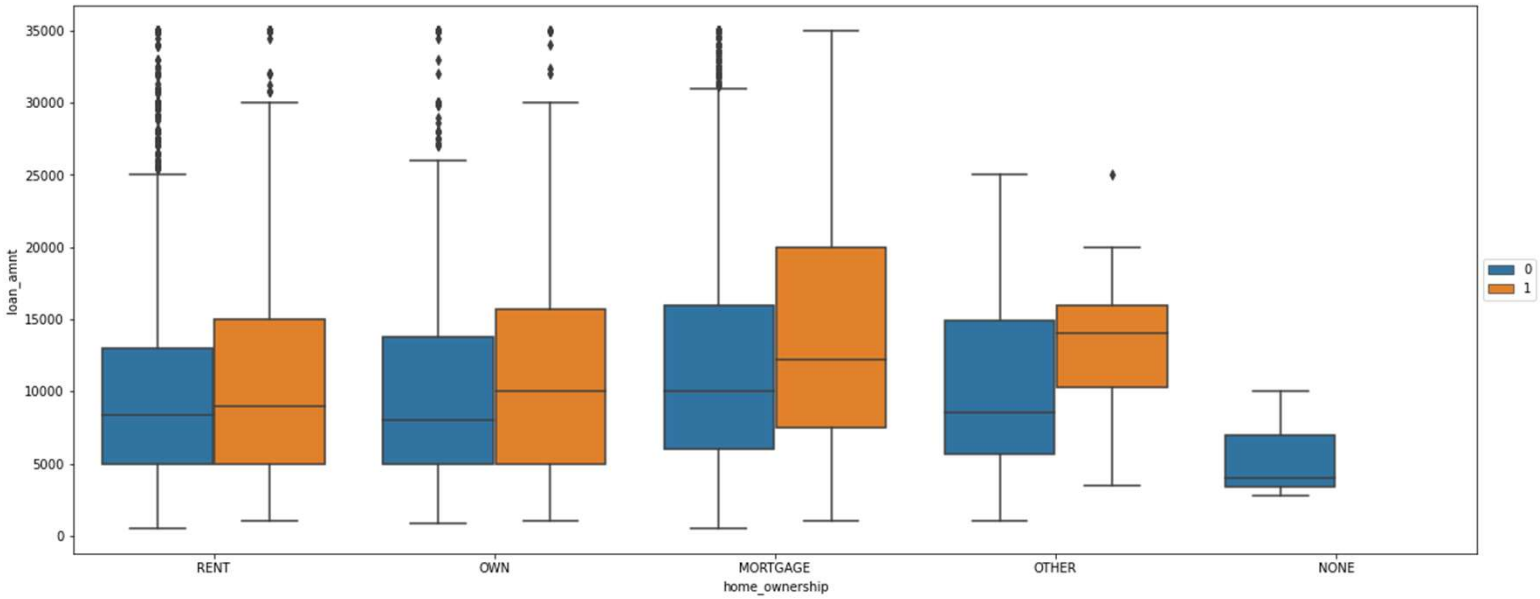
Observation:

- Grade and Interest rate are proportional which leads to face high number of defaulters.
- **Interest rate is increasing for every grade and also the defaulters for every grade are having their median near the non-defaulter 75% quantile of int_rate.**

Exploratory Data Analysis (EDA)

Bivariate Analysis - 3

Box plot between Home Ownership and Loan Amount for Loan Status



Observation:
Out of all types of Home Ownership, **loan applicants who has Mortgage are taking high amount as loan and also, they are potential defaulters too**.

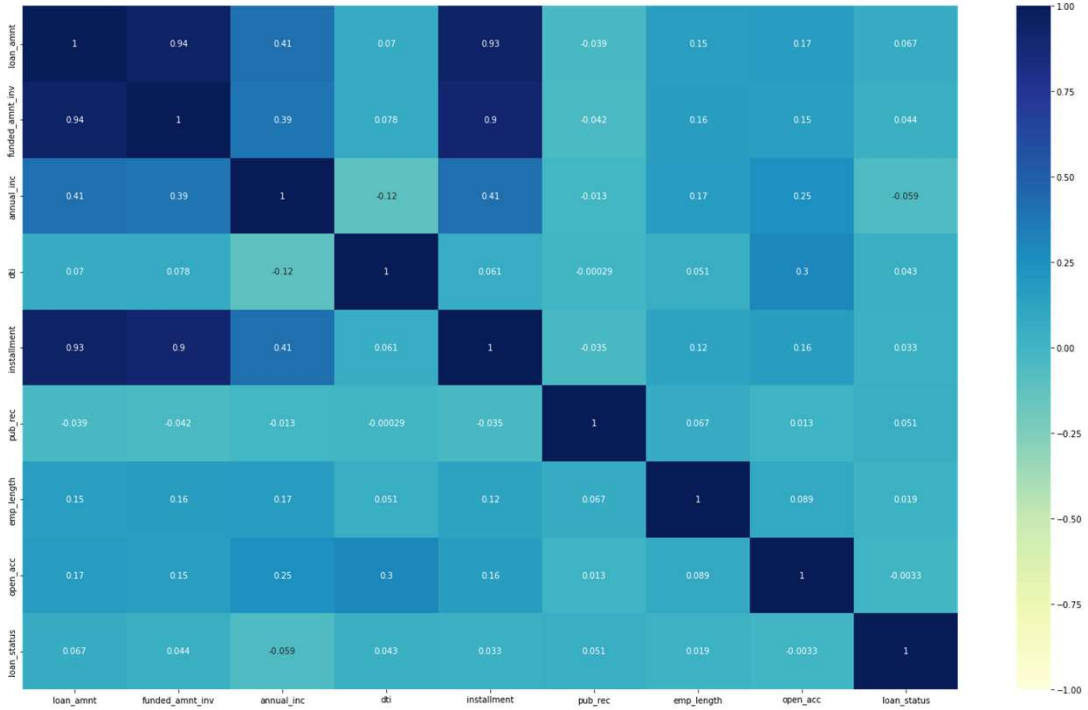
Exploratory Data Analysis (EDA)

Correlation Analysis (Multivariate)

Generate the **heat map** for columns: **'loan_amnt','funded_amnt_inv','annual_inc', 'dti', 'installment','pub_rec', 'emp_length','open_acc','int_rate_range', 'loan_status'**

Observation:

- **Highest positive correlation between loan_amnt, funded_amnt_inv and installment**
- **Negative correlation is between annual_inc and dti**



Conclusion with Recommendation

- As per the given source loan data, **following are few factors / risk indicators which conveys potential loan default scenarios**.
  - **High** Interest Rate (especially beyond ~13%)
  - **High** Annual Income (35+ k Annual income range)
  - **Long** Term Loans (60 months)
  - Loan purpose to **consolidate debts** and for **credit card bill payments**
  - Loan applications from **Highly** experienced employees who are taking high amounts (better to avoid to mitigate risk) and **High Annual Income**
- In summary, **Annual Income**, **Home Ownership**, **Purpose of Loan**, **Loan Amount**, **Interest Rate**, and **Loan Term** are **critical factors** to be considered by Company's Underwriters who perform Risk analysis and decide whether to "accept" or "reject" the loan application.
- Final recommendation as per this EDA is to consider the above set of attributes as **driving factors (or driver variables) behind loan default**, i.e. the variables which are **strong indicators of default**. Hence, this Lending company can refer this above during upcoming loan application's risk assessment.

Thanks