# LENDING CLUB CASE STUDY

# ASSIGNMENT SUBMISSION

Name: **Rahul Kumar Jha**

**Ragesh Hajela**

# Problem Statement

## Business Challenge

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.

**Consumer finance company** which specialises in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

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

## Business Context

When a person applies for a loan, there are **two types of decisions** that could be taken by the company:

**Loan accepted:** The company approves the loan after careful considerations

**Loan rejected**: The company had rejected the loan (because the candidate does not meet their requirements etc.)

3 possible scenarios for accepted loan are as:

**Fully paid**: Applicant has fully paid the loan

**Current**: Applicant is in the process of paying the instalments.These candidates are not labelled as 'defaulted'.

**Charged-off**: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has **defaulted** on the loan

## Business Objective

Like most other lending companies, lending loans to 'risky' applicants is the largest source of financial loss,borrowers who **default** cause the largest amount of loss to the lenders.

If anyone is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Identification of risky loan applicants is the aim of this case study so such loans can be reduced thereby cutting down the amount of credit loss.

Business wants to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

# Data Analysis Challenges

## Data Understanding

Familiarizing and understanding domain and how lending business work in general

Deducing correct meaning of the terminologies and variables present in loan dataset

Identifying the Customer & Loan attributes correctly from dataset

Correctly understanding the significance of the different loan variables

Identifying data quality issues looking through the dataset

## Data Preparation

The dataset have more than 100 variables, we will have to limit the features based on domain knowledge and data cleaning techniques

Preparing metadata dictionary for keeping important info of variables easily accessible at once

Reducing & removing the noise in data , treating missing values, transforming & reshaping data to fit our needs for analysis

Performing quality checks on data to make it appropriate and useful for EDA

## Data Analysis

Deriving insights from variables following various EDA techniques such as univariate, bivariate, segmented univariate analysis e.t.c.

Performing Facts and data supported analysis to find the important factors and driving variables for loan default

Deriving business goal attached conclusions with clear and precise informations based on important factors

Communicating the finding to the stakeholders in simple & easy to understand yet precise informations

# Driver Variable 1: Loan Term (The number, in months, of payments on the loan)

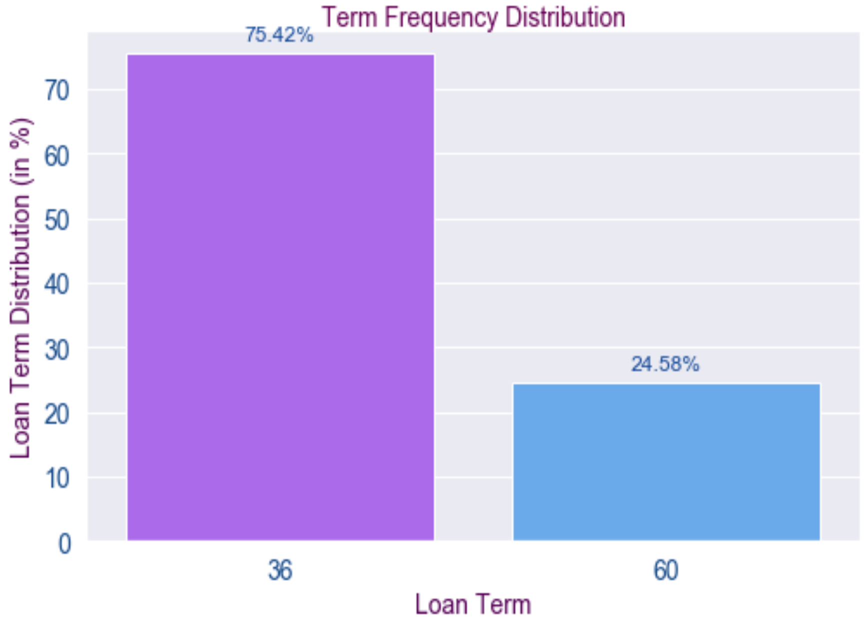
UpGrad

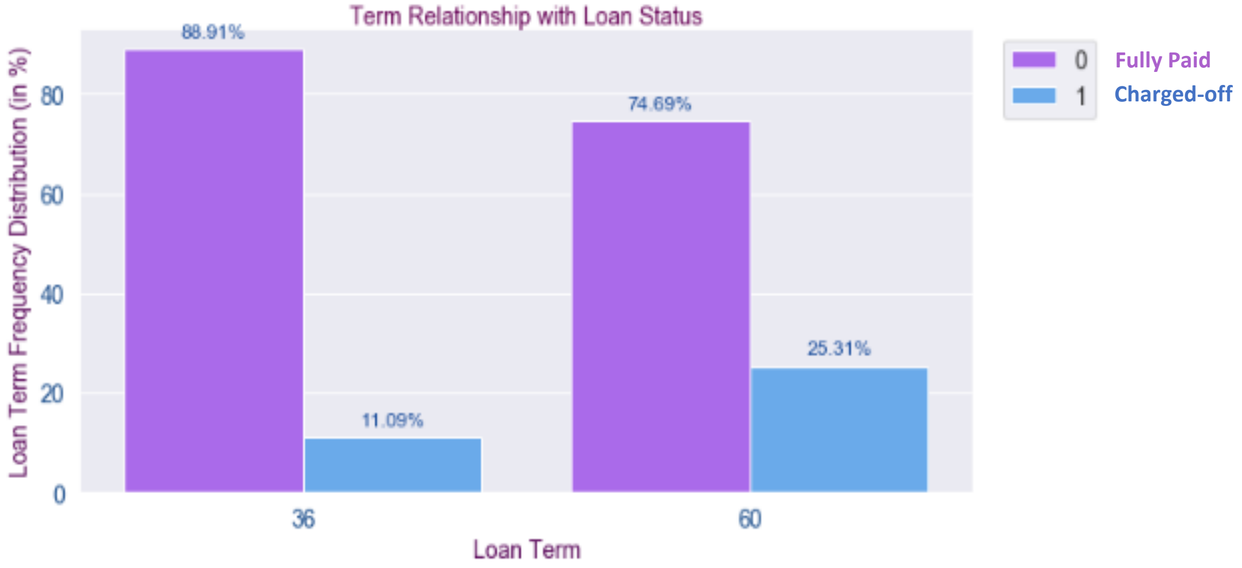
## Term

**Term** is a **loan attribute,** the options are **36 Months** and **60 Months**.

**75%** of the loan term fall under 36 months term category & **25%** in 60 months term category

This indicates applicants prefer shorter term for loan more




## Term Relationship with Loan Status

**89%** of the 36 months term loan have been **fully paid** and **11%** have the likelihood for **loan default**

**75%** of the 60 months term loan have been **fully paid** and **25%** have the likelihood for **loan default**

**60 months** term **charged-off rate** is more than **twice** of the **36 months** term **charged-off rate**

**Longer the term greater the chances of applicants not repaying the loan**

# Driver Variable 2: Interest Rate (Interest rate on the loan)

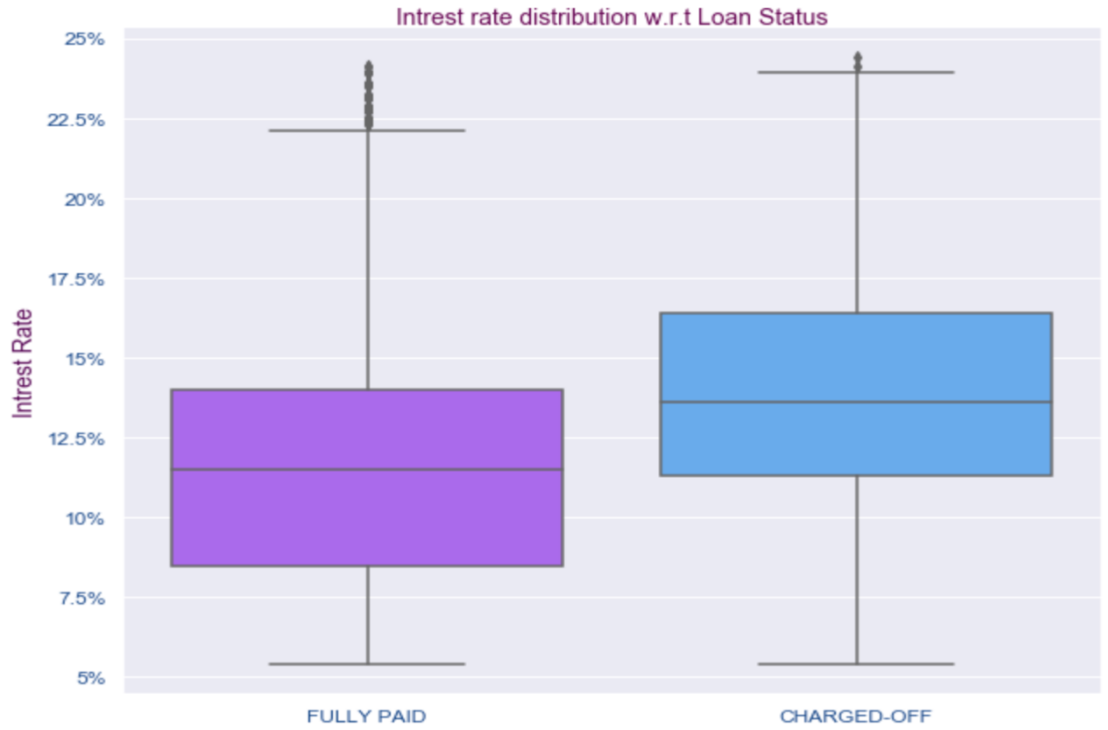
## Interest Rate Relationship with Loan Status

**Interest rate** ranges from as low as **5%** to as high as **25%**

**Average** interest rate for **fully paid** loan is **12%** & for **charged-off** loan the **average** interest rate is **14**%

As the interest rate increases so is the likelihood of loan default

**Higher interest rate poses greater risk for business and the applicants with high interest rate are more likely to not repay their loan**





## Interest Rate & Term Variations with Loan Status

Interest rate & Term are lower for fully paid loan, 36 months term loan have lower interest rate than 60 months loan term

Interest rate & Term are higher for charged-off loan, 60 months term loan have higher interest rate than 36 months loan term

Applicants with 36 months loan term and have considerably lower interest rates are more likely to repay their loan

**Applicants with 60 months loan term and considerably higher interest rates are more likely to default ot not repaying the loan.**

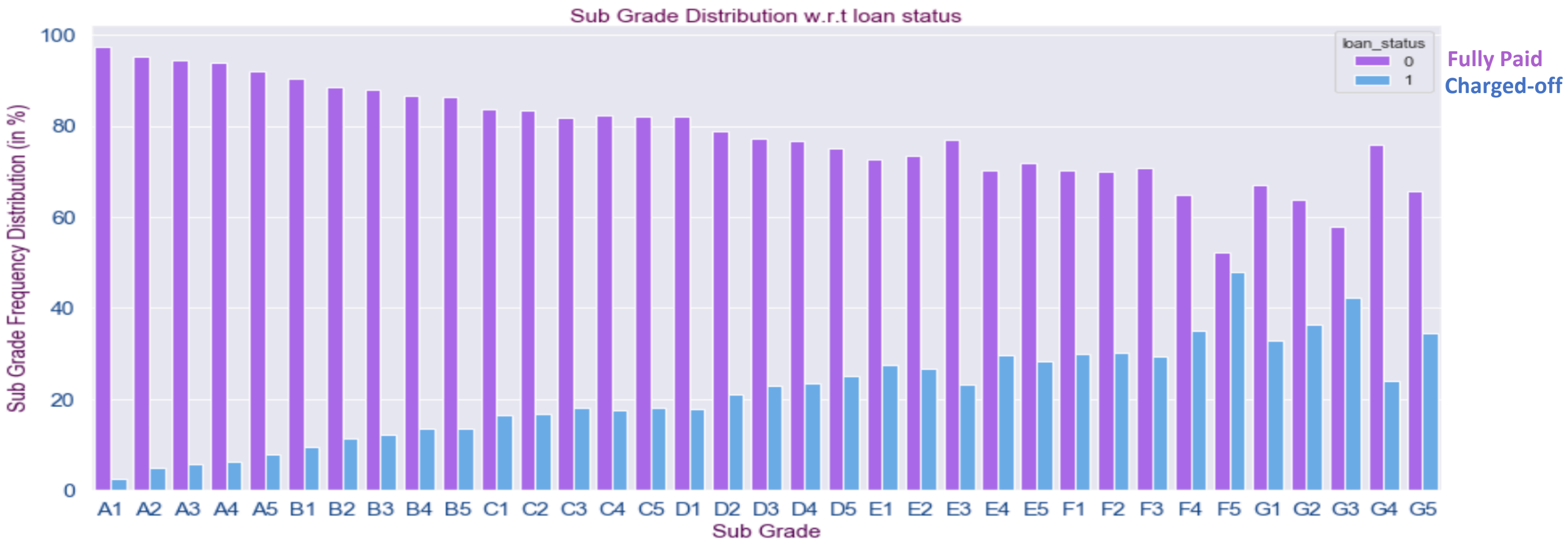
# Driver Variable 3: Sub grade (LC assigned loan sub grade)

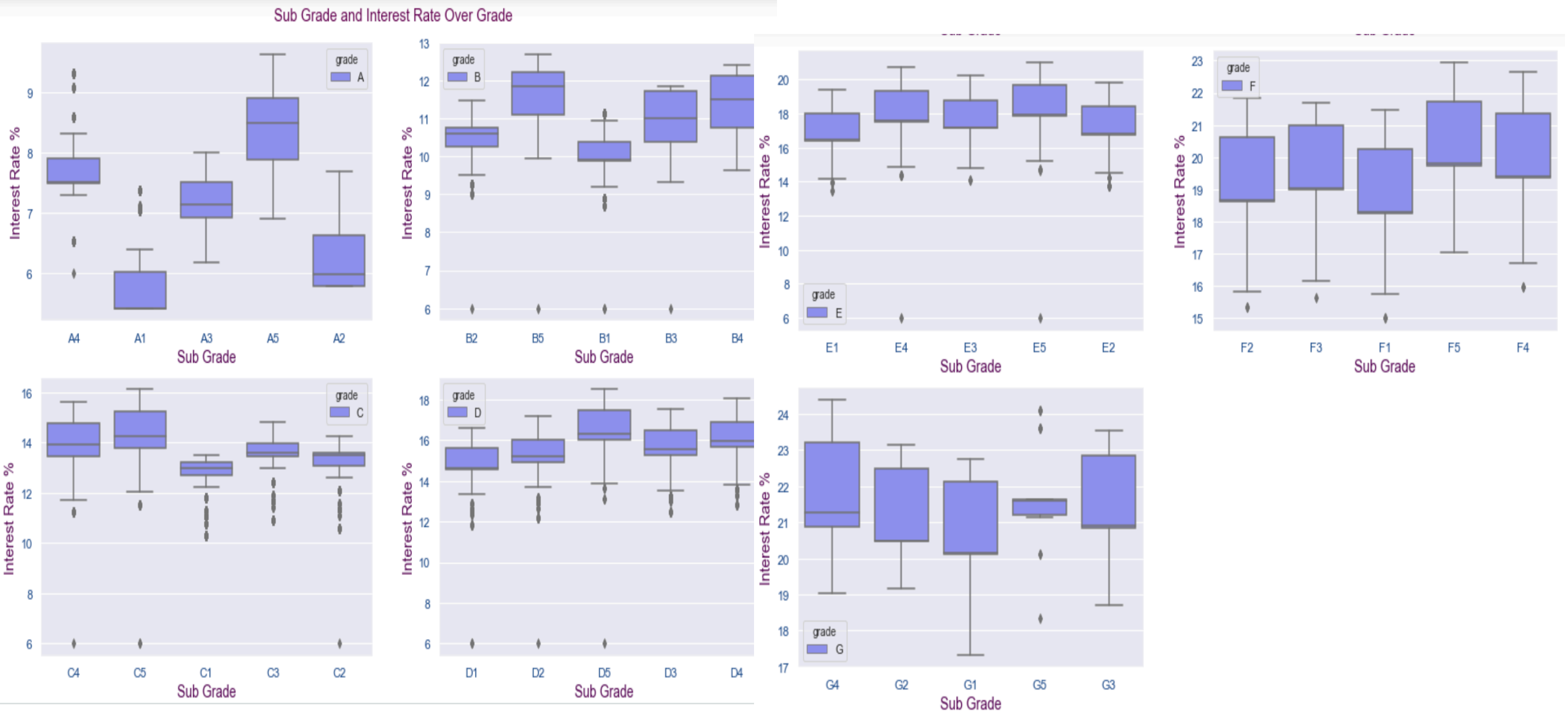
## Sub-grade relationship with loan status

There are **35 unique** sub grades, each associated with a grade.

Fully paid loans tend to have better sub grade, so for lender, better sub grades are good indicator of good loans.

There exists almost linear relationship between sub grades and loan status. **As the sub grades worsen so is the likelihood that loan repayment probability declines too.**




## Sub-grade relationship with grade and Interest rates

Interest rates tends to **increase** for subgrades in moving direction of grades from A to G.

For loan applicants having poor sub-grades, the interest rate will be higher than applicants having better grades/sub-grades.

**As subgrade worsens, chances of loan default rises significantly.**

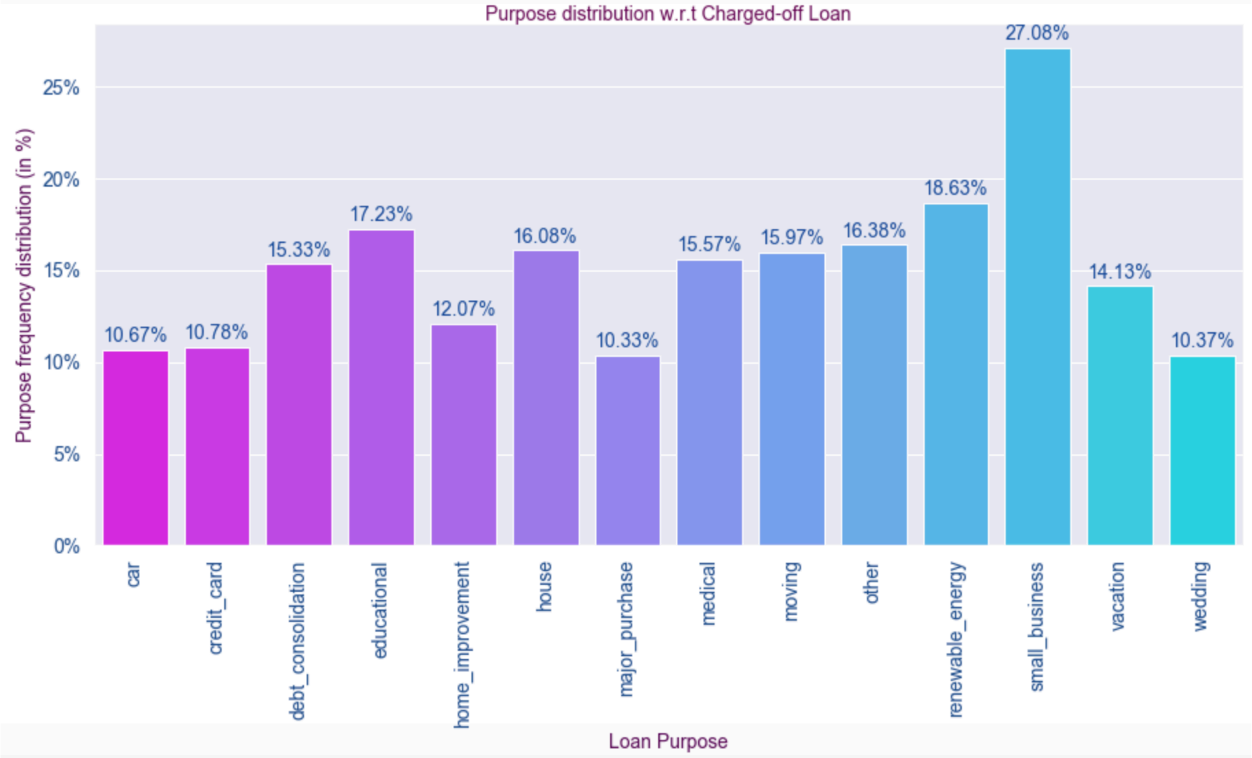
# Driver Variable 4: Purpose (A category provided by the borrower for the loan request)

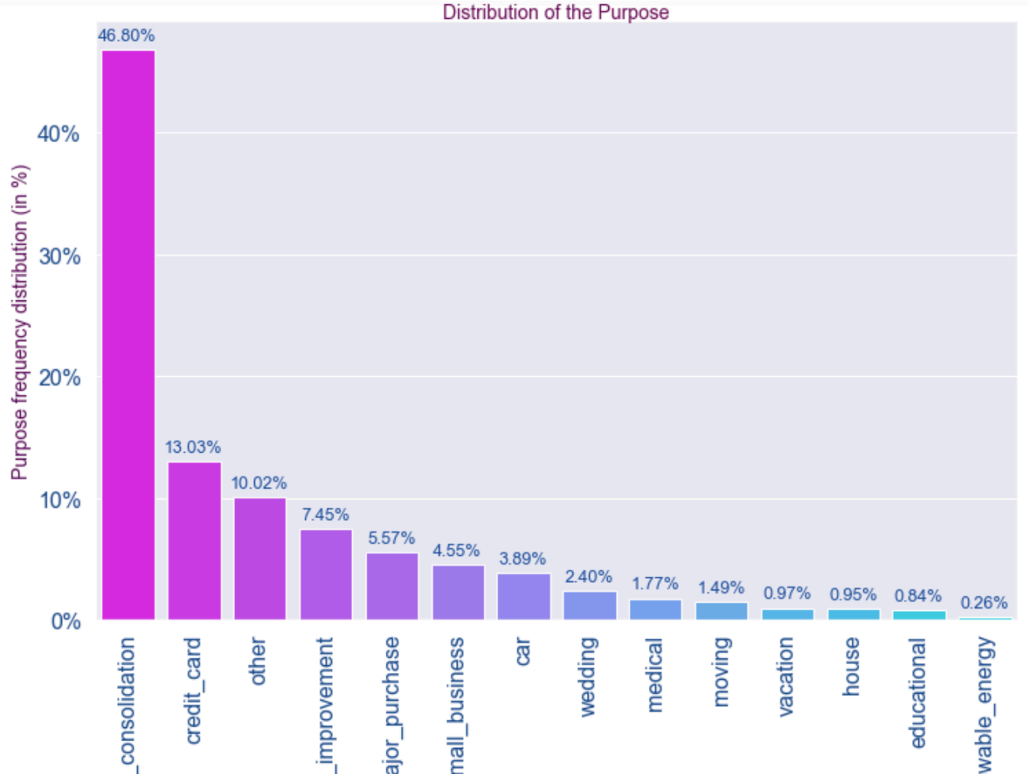
## Purpose

It is a consumer attribute.

**Debt consolidation** have the **highest frequency (approx 47%)** as purpose cited by loan applicants.

Purpose and title were observed to be having mostly similar information as they both have same most frequent values.




## Purpose relationship with loan status

**Small business** loan purpose has the highest charged-off frequency.

Applicants with Loan for **small business purposes are more likely not to repay the loan.**

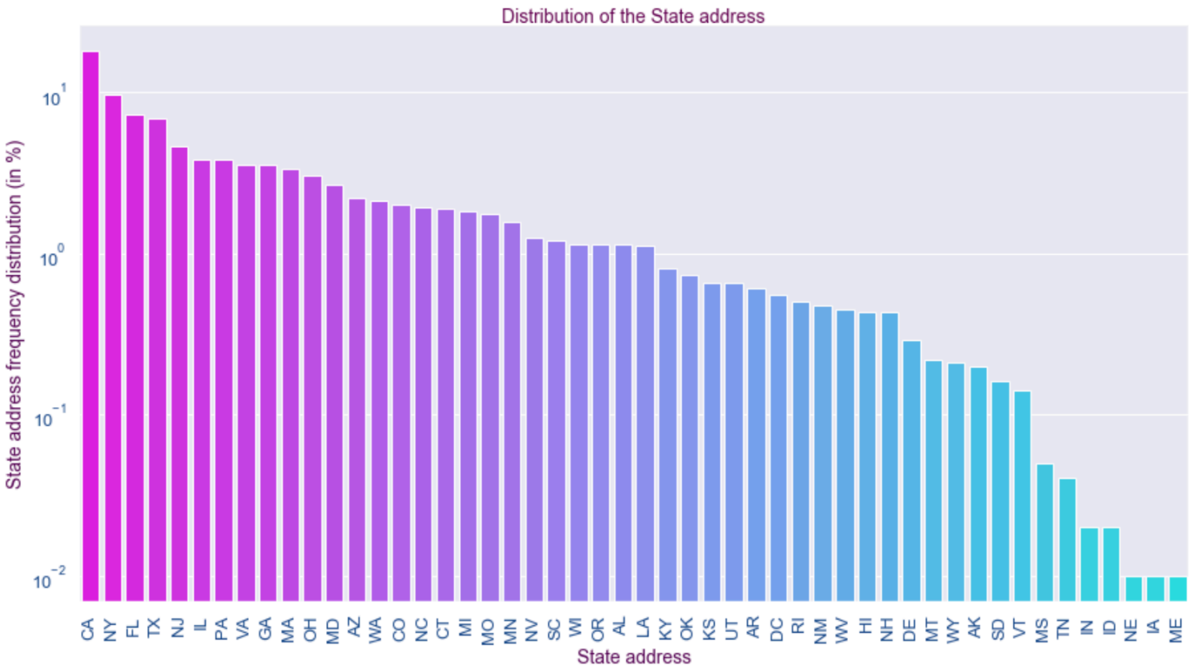
# Driver Variable 5: Address State (The state provided by the borrower in the loan application)

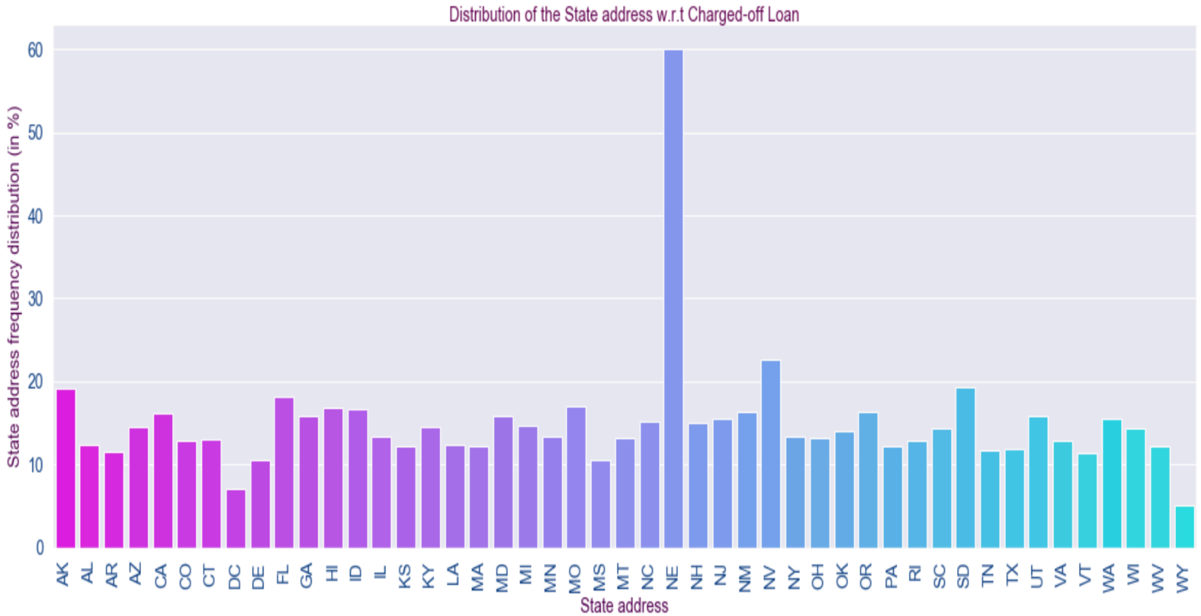
UpGrad

## Address State

It is a consumer attribute.

Highest number of loan applicants are from **California (CA)** state.




## Address state relationship with loan status

**Nebraska with state code NE & Nevada with state code NV**, both have very high charged-off rate.

E.g. Out of each 100 applicants from NE on an average **60 applicants default** on their loan.

For business, loan application from NE & NV states pose **very high risk for loan default.**

# Driver Variable 6: Loan Amount (Listed amount of the loan applied for by the borrower)

UpGrad

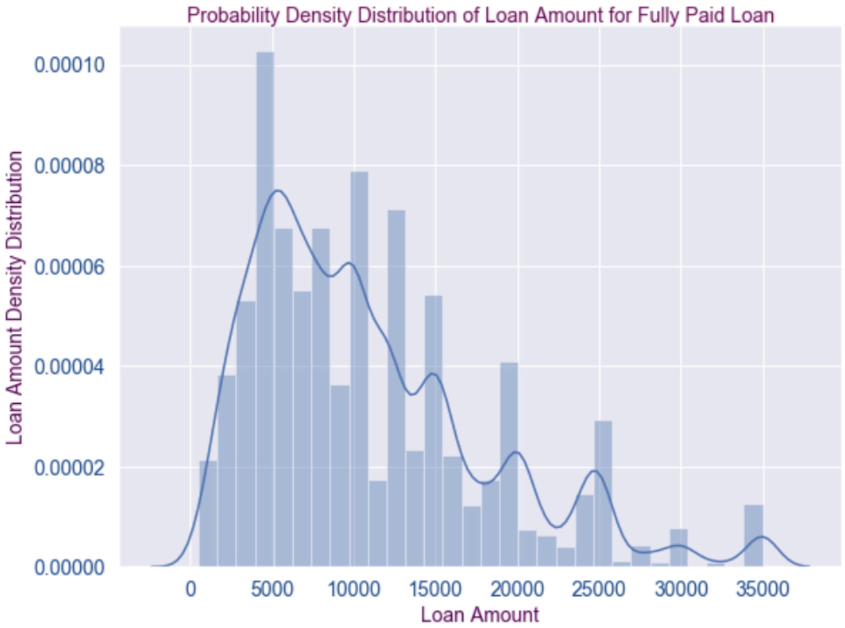
## Loan Amount relationship with loan status

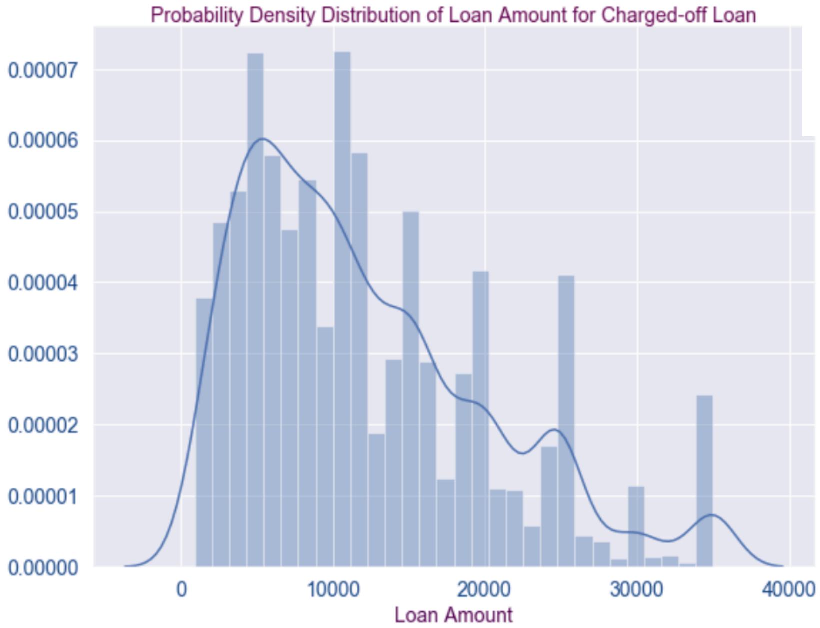
Observed **high correlation and linearity** between loan_amnt, funded_amnt, funded_amnt_inv & installment. Loan amount has the highest positive correlation with loan status.

**Average loan amount for charged-off loan is greater than the fully paid loan**

Charged-off loan average amount is greater, also there are more data points above 50% (between q2 - q3) in charged-off loan amount, along with outlier presence.

**As the loan amount increases, so is the likelihood of loan default.**


## Loan Amount observations

Density curve of loan amount for charged-off loan is **wider at center and spike in loan default** as the loan amount increases.

Observe **huge spread** in the density distribution and **right skewness** in the data for loan amount.

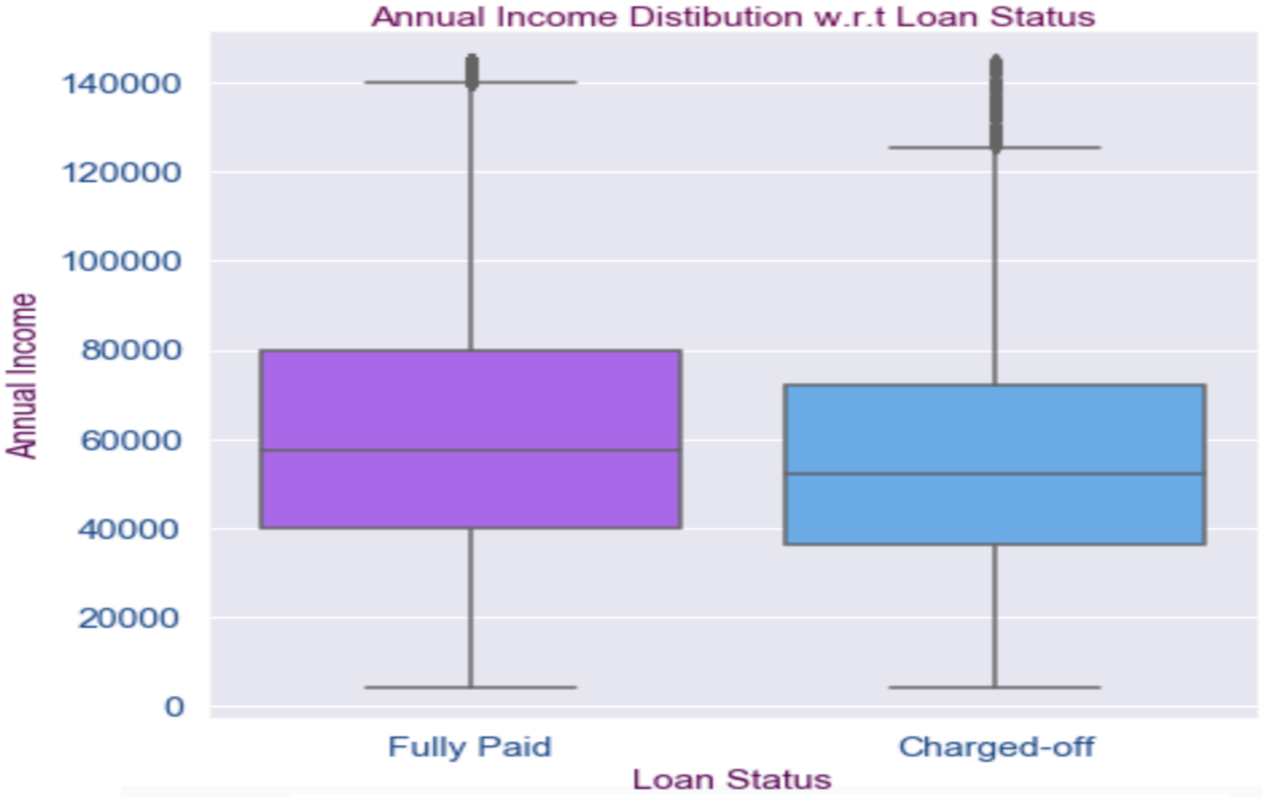
# Driver Variable 7: Annual Income (Self reported income provided by the borrower)

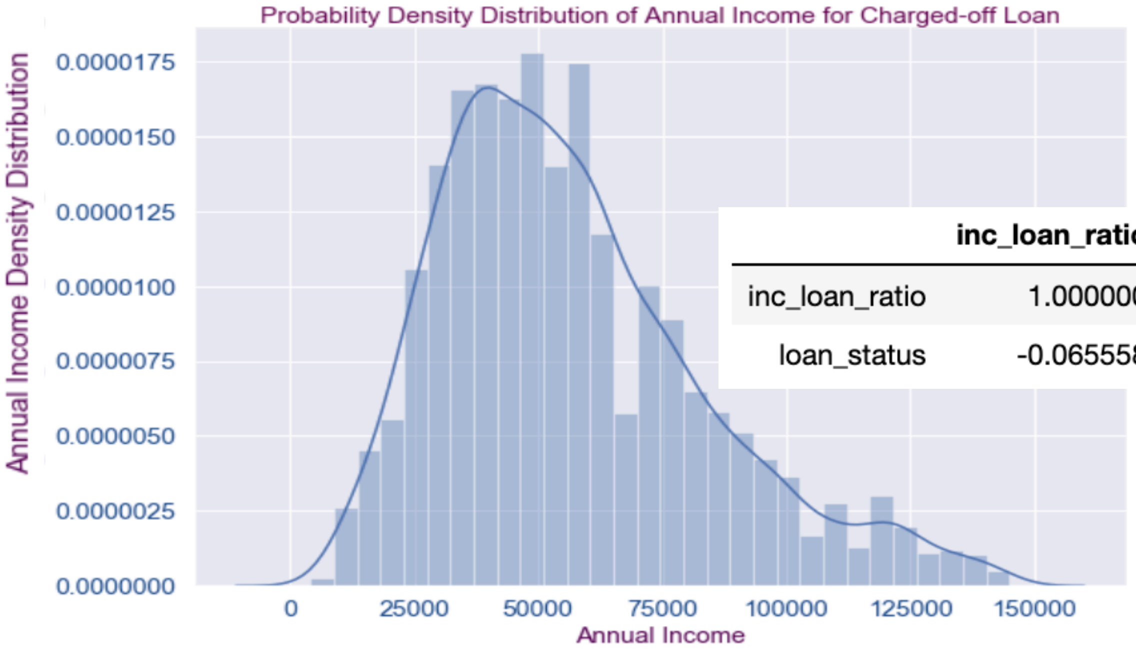
UpGrad

## Annual Income relationship with loan status

Average annual income for charged-off loan is **lesser** than the fully paid loan

**Lesser annual income have greater chances of defaulting or not repaying the loan.**




| | inc_loan_ratio | loan_status |
|---|---|---|
| inc_loan_ratio | 1.000000 | -0.065558 |
| loan_status | -0.065558 | 1.000000 |

## Annual Income relationship with loan amount

**Ratio of annual income & loan amount is lower for charged-off loan** than fully paid loan, so we have found that annual income have greater impact with combination of loan amount on the loan default.

Applicants with **lesser annual income to loan amount ratio** are more likely to not repay the loan.

# Driver Variable 8: DTI

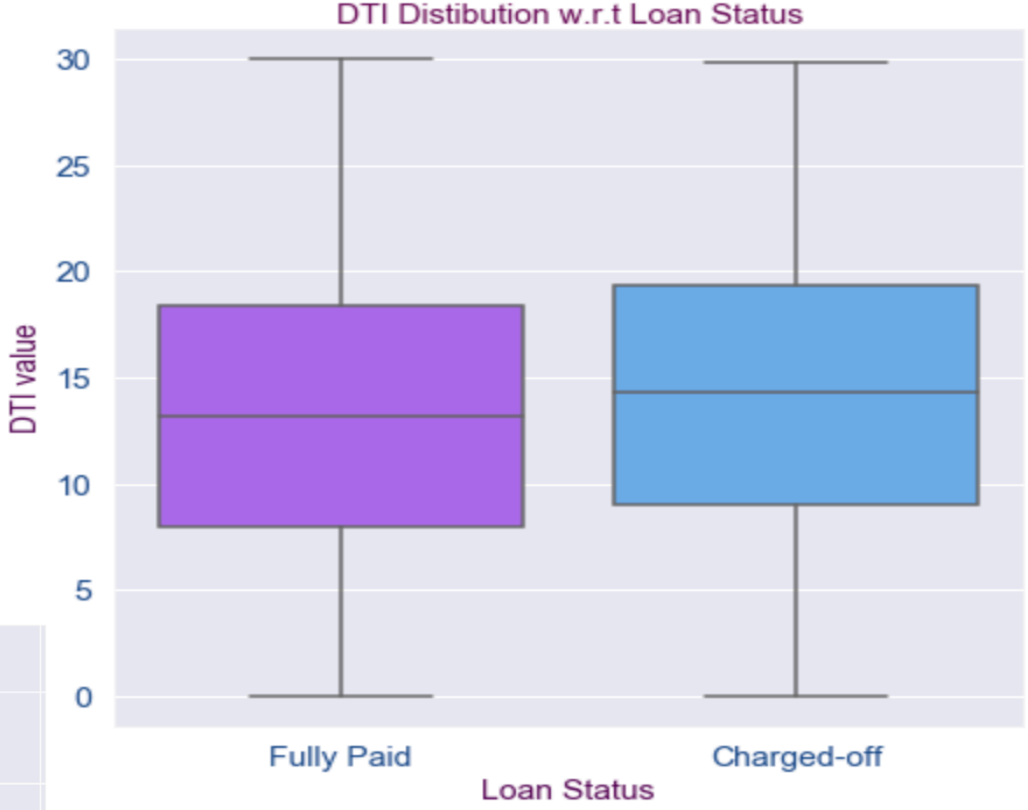
(Ratio of borrower's total monthly debt payments on the total debt obligation, excluding mortgage and the requested LC loan, divided by the borrower's self reported monthly income)

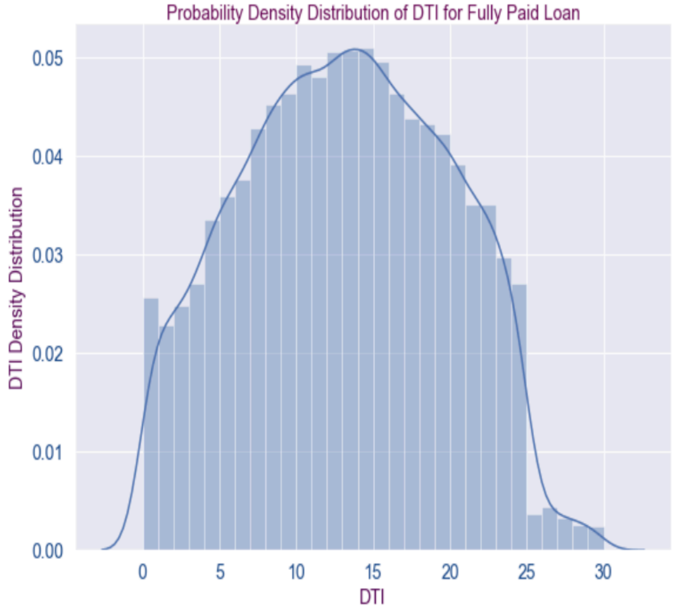
## DTI relationship with loan status

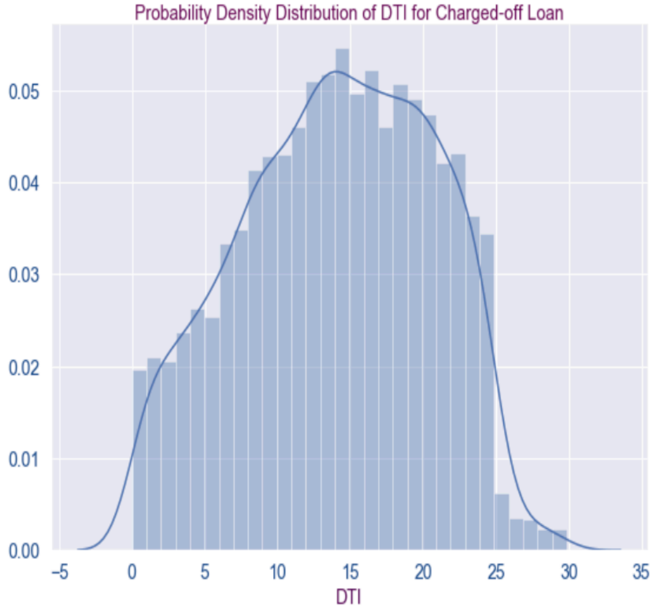
Average DTI is **more** in cases of charged_off loan compared to the fully paid loan.

**As DTI increases, so is the likelihood of defaulting on loan.**

DTI is a strong driver variable for the charged-off loan.




## DTI observation

DTI (**Debt to income ratio**) average for charged-off loan is **higher** than for the fully paid loan.

DTI can be used in combination with other driver variables to flag or identify the risky loan applicants.

# Conclusions

1. These variables are the strongest indicators of whether a borrower may charge-off a loan in future.

2. They are good fit to be used together for the identification of the risky applicants.

3. Business can use these variables metrics in combinations to take calculated decision of accepting or rejecting the loan, decreasing the loan amount or lending at higher interest rate.