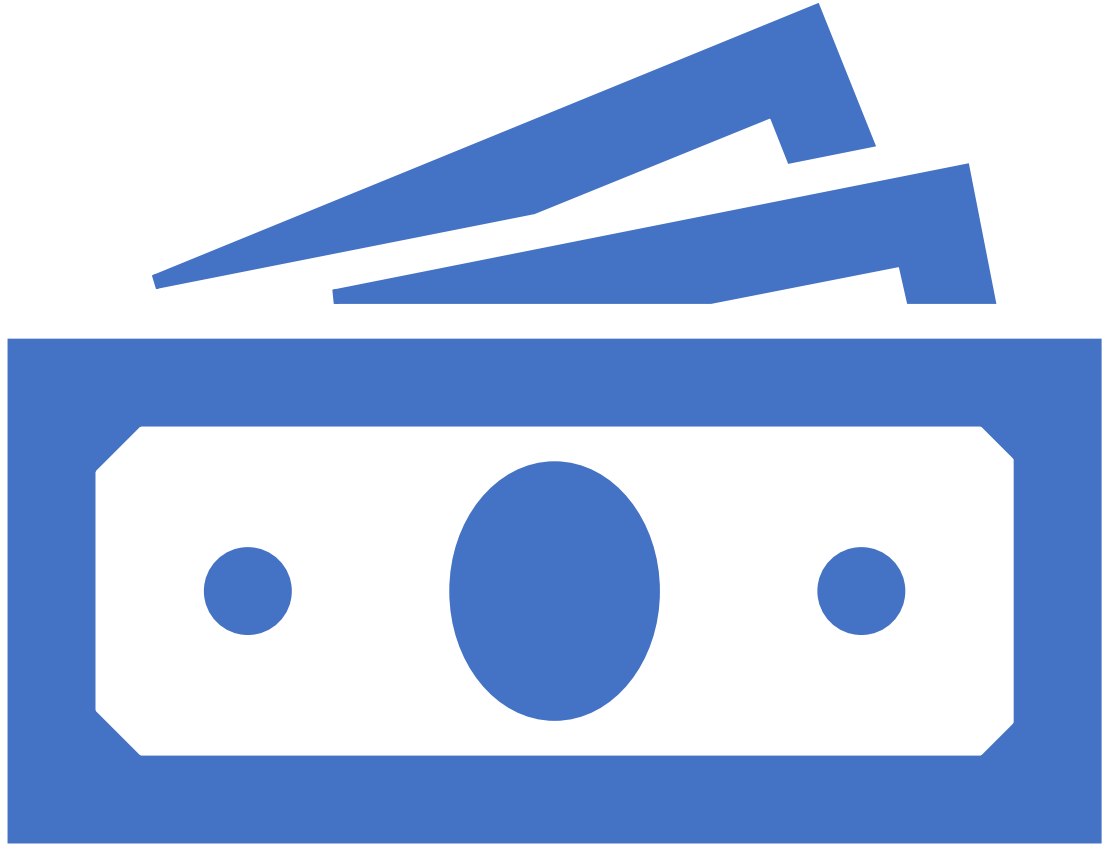
Lending Club Case Study

EDA Module Case Study

Group Members:
1. Vishvmitra Belsare
2. Akash Prasad



Introduction:

- Lending Club is a marketplace for personal loans that matches borrowers who are seeking a loan with investors looking to lend money and make a return.
- When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.
- Two types of risks are associated with the bank's decision:
  1. If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company
  2. 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

Objective

- In this project, we aim to understand the driving factors (or driver variables) behind Loan ‘Default’, i.e. the variables which are strong indicators of Loan ‘Default’.
- The loan data is given about past loan applicants and whether they ‘defaulted’ or not. The aim is to use **Exploratory Data Analysis (EDA)** to identify patterns which indicate if a person is likely to ‘Default’.
- The company can utilize this knowledge for its portfolio and risk assessment and help make better decisions while granting loans to the customers.

Problem Solving Methodology:

We perform the Exploratory Data Analysis (EDA) on the past loan data which involves following steps:

- Data Understanding
- Data Preparation (Data cleaning/formatting/filling missing values)
- Univariate Analysis
- Segmented Univariate Analysis
- Bi-Variate Analysis
- Results

Data Cleaning

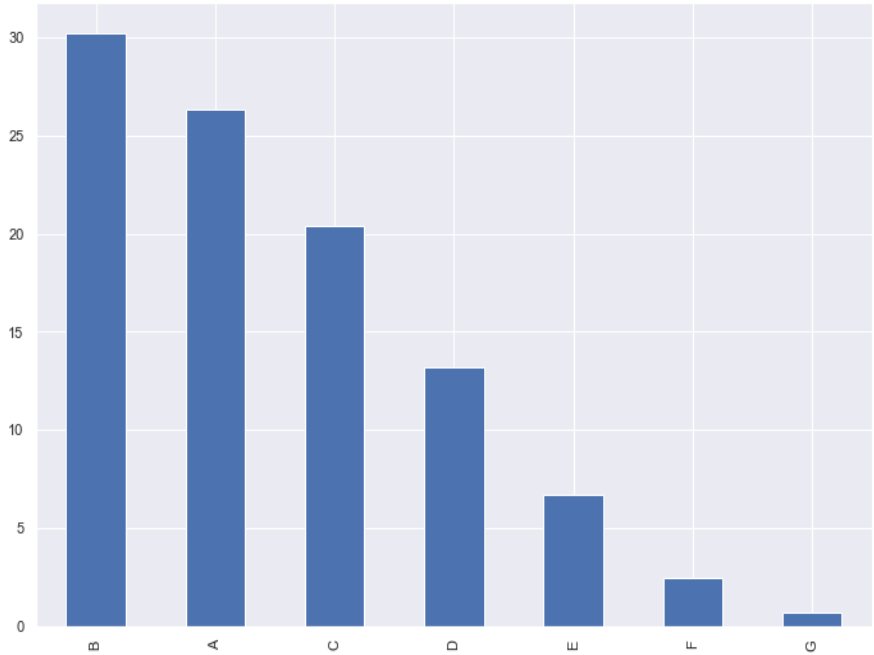
1. Remove all the columns having 70% or more data missing (Assumption: Columns not useful for our analysis as maximum data in that is missing)
2. Drop all columns having ‘single unique value’ as we can’t derive any insights from it
3. We remove the rows with loan_status as “Current” because as we can’t decide based on it
4. Emp_length has 1033 missing values. Since, it’s a categorical variable we replace missing values with **Mode** (10+ years)
5. Customer Behaviour variables that are created after a loan is taken doesn't make sense for our analysis.
6. Also, fields like 'id', 'member_id' & 'url' are not needed as we are doing overall analysis.
7. 'zip_code' values are masked so not useful anymore.
8. 'desc' has too many unique values which can't be analysed
9. Columns like 'loan_amnt', 'funded_amnt', 'funded_amnt_inv' have high correlation value and are similar fields. So, keeping just ‘loan_amnt’.

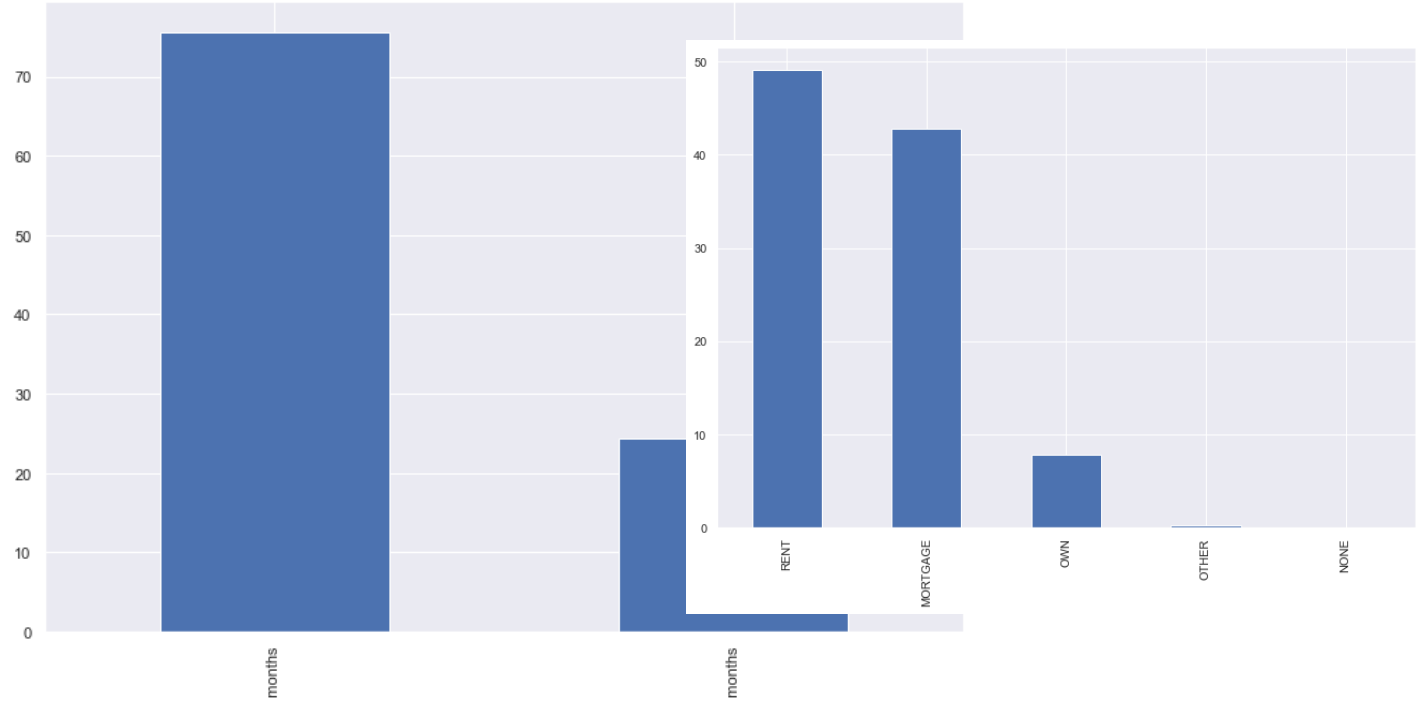
Data Formatting

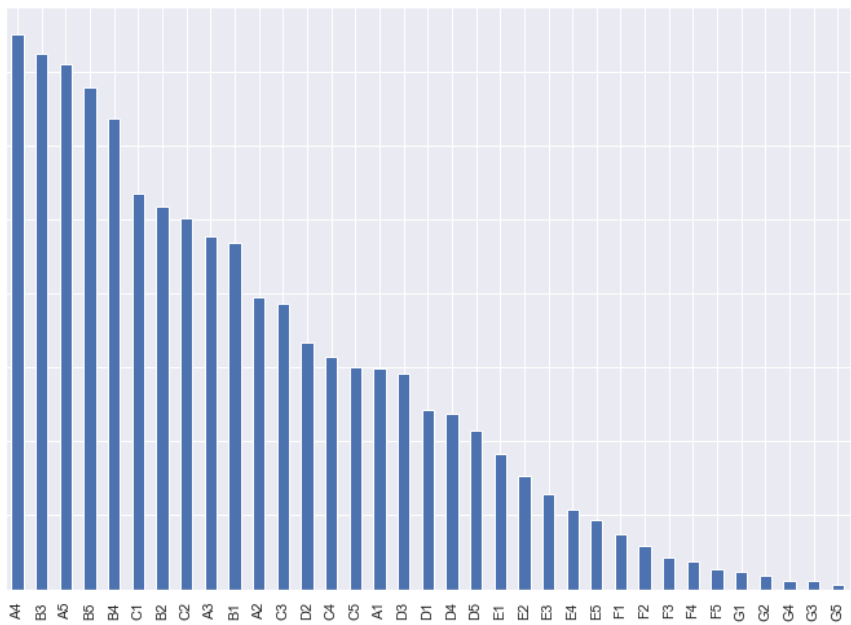
- We remove the outliers for “annual_inc” field -
  - annual_income_upper_limit = Q3_annual_income(75%ile) + 1.5* IQR_annual_income(75%ile – 25%ile)
  - We remove all rows with ‘annual_income’ above ‘annual_income_upper_limit’
- We extract the month and year from ‘issue_d’ column and create new columns for it. Also, we append “20” at the start of the year as data is between 2007-11
- For “int_rate”, we remove “%” at the last and convert it to float type
- We convert “loan_amnt” to int
- We convert “annual_inc” to float
- We convert “installment” to float
- We convert “dti” to float

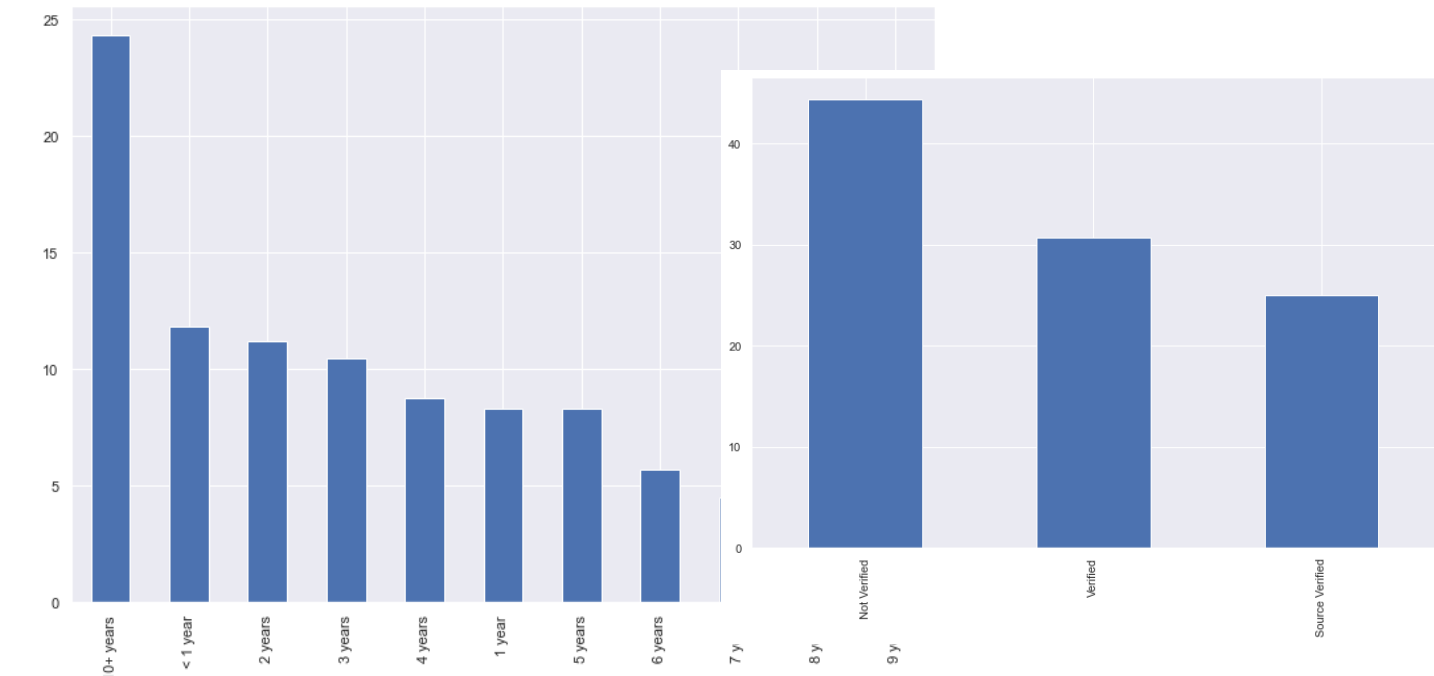
Data Visualization

- Following columns are kept are used for analysis
- Categorical Columns:
  - 'term', 'grade’,
  - 'sub_grade’,
  - 'emp_length’,
  - 'home_ownership',
  - 'verification_status’,
  - 'purpose', 'addr_state',
  - 'issue_d_month’
  - 'issue_d_year', 'loan_status’
- Continuous columns:
  - ‘annual_inc’,
  - 'loan_amnt’,
  - 'int_rate’,
  - 'dti’,
  - 'installment'


B
A
C
D
E
F
G


months
months
RENT
MORTGAGE
OWN
OTHER
NONE


A4
B3
A5
B5
B4
C1
B2
C2
A3
B1
A2
C3
D2
C4
C5
A1
D3
D1
D4
D5
E1
E2
E3
E4
E5
F1
F2
F3
F4
F5
G1
G2
G4
G3
G5


10+ years
< 1 year
2 years
3 years
4 years
1 year
5 years
6 years
7 y
8 y
9 y
Not Verified
Verified
Source Verified

Observation about categorical variables:

- Loans are given for 2 terms (36 months and 60 months). Above 70% of the total loans are given for 36 months (3 years)
- Maximum loans are given with Grade B followed by A and C, then D,E,F and very few with G.
- Maximum loan are given with sub-grade A4 and then B3
- Maximum number of loans were given to perons with employment length '10+ years'
- Home ownership is highest for - 'RENT'
- Maximum loans given are 'NOT VERIFIED'
- Maximum loans were given for the purpose - 'debt_consolidation'
- Maximum loans are given for the state - 'CA' (California)
- Maximum loans are given for the month of Dec
- Maximum loans are given for the year 2011
- Maximum loans given out were 'Fully Paid' (around 85%) and only around 15% were "Charged-Off".

Univariate Analysis
Categorical Variables - Observations

We have most loan with status "Charged off" in following categories:

- when term = 36 months i.e number of payments on the loan is '36 months'
- When Grade = B i.e when LC assigned loan grade is 'B'
- When Sub Grage = B5 i.e when LC assinged loan grade is 'B5'
- When Employee length = '10+ years' i.e when employment length/experience of the person is 10+ years
- When the home ownership provided by the person is 'RENT'
- When verification status is 'Not verified'
- When the purpose for the loan is 'debt consolidation'
- When address state is 'CA’
- When loan was issue in 'Dec' month
- Number of loans are increasing every year, max loans were given in '2011’

But this doesn't necessarily mean these conditions lead to default as its just frequency for each bucket and not normalized

Segmented Univariate Analysis
Continuous variables - Observations

We have most loan with status "Charged off" in following categories –

- Applicants who receive interest at the rate of 13.5-18.5
- Applicants who have an income of range 30800-56080
- Loan amount is between 0 - 7000
- Dti is between 13-19.5
- When monthly installments are between 145-274

Bi-Variate Analysis:

For Continuous variables, from heat map we can know:

1. Interest rate has positive correlation with loan amount

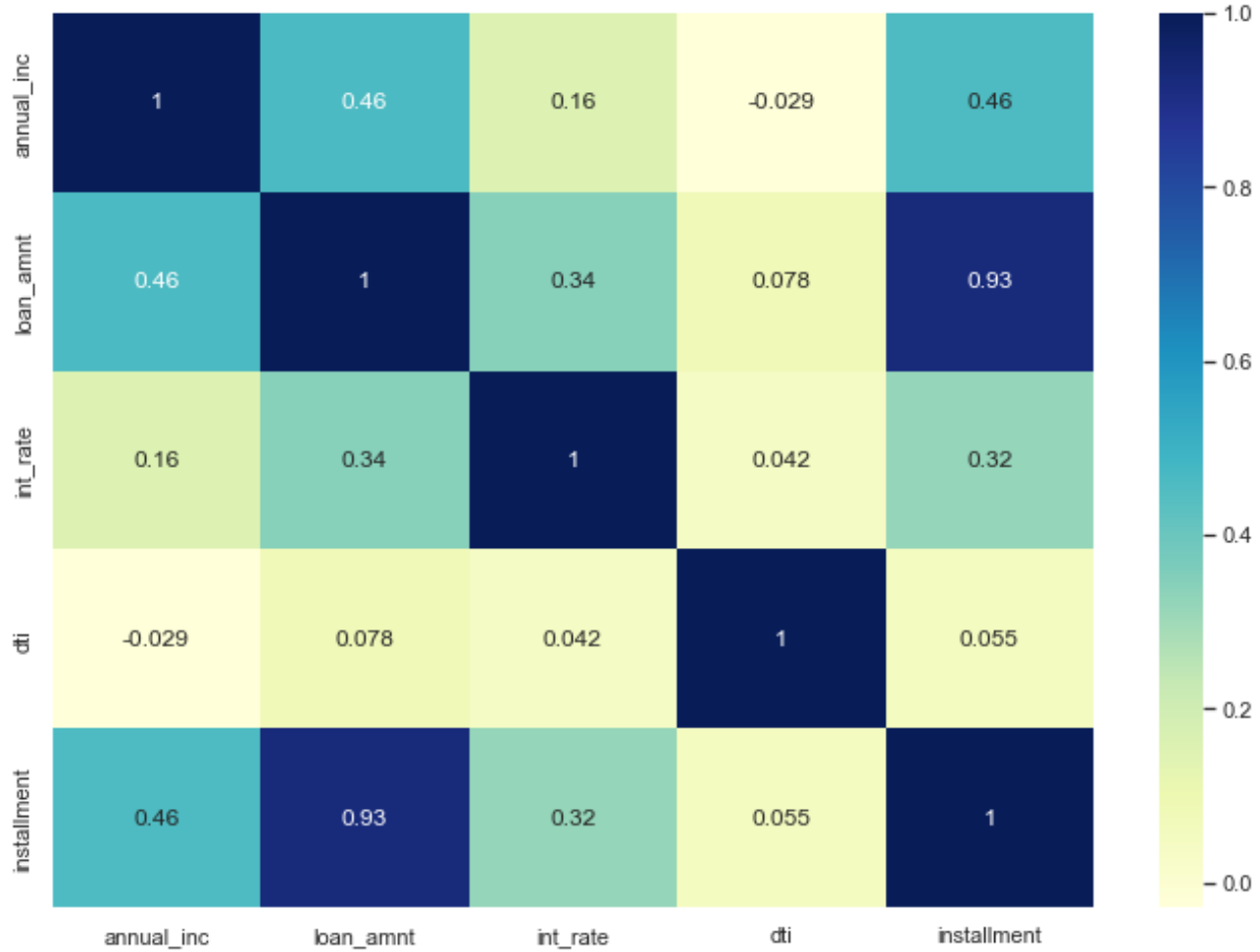
2. Loan amount is highly correlated with installment

3. DTI has negative correlation with annual income

4. DTI has small positive correlation with loan amount and installment

5. Annual income has positive correlation with installment, loan amount

6. Annual income has negative correlation with DTI



BiVariate Analysis (Categorical variables):

Observations:

1. Loan_status vs Loan term: We have 2 terms (36 months, 60 months). We see that the percentage of loans that are "Charged off" is much higher for longer term.
2. Loan_status vs Grade: Percentage of loans getting "Charged Off" increases as the Grade decreases from A to G.
3. Loan_status vs sub-grade: Similar trend is seen for sub-grades as well, with some exceptions, that percentage of loans getting "Charged off" increases as the sub-grade decreases (A1 to G5)
4. Loan_status vs Employment_length: Employee length doesn't have a clear trend wrt. loan status
5. Loan_status vs Home_status: Home status doesn't have a clear trend wrt loan_status
6. Loan_status vs addr_state: Percentage of loans "Charged off" is highest in state - 'NE' (Nebraska)
7. Loan_status vs issue_month: Percentage of loans "Charged off" is highest for month December. Then September and May.
8. Loan_status vs issue_year: No clear trend on percentage of loans "Charged off" over year
9. Loan_status vs purpose: Percentage of loans "Charged off" is highest for purpose 'Small Business'.

Multi-Variate Analysis

- Observation:
- customer who is applying for loan with amount 28000 to 35000 and installment above 800 is having higher chance to default
- customer who is applying for Grade "G" & interest above 20% is having higher chance to default
- customer who is applying for "home loan" & interest between 14% to 16% is having higher chance to default
- MORTGAGE loan having loan amount between 12000 to 14000 has higher chance of defaulting