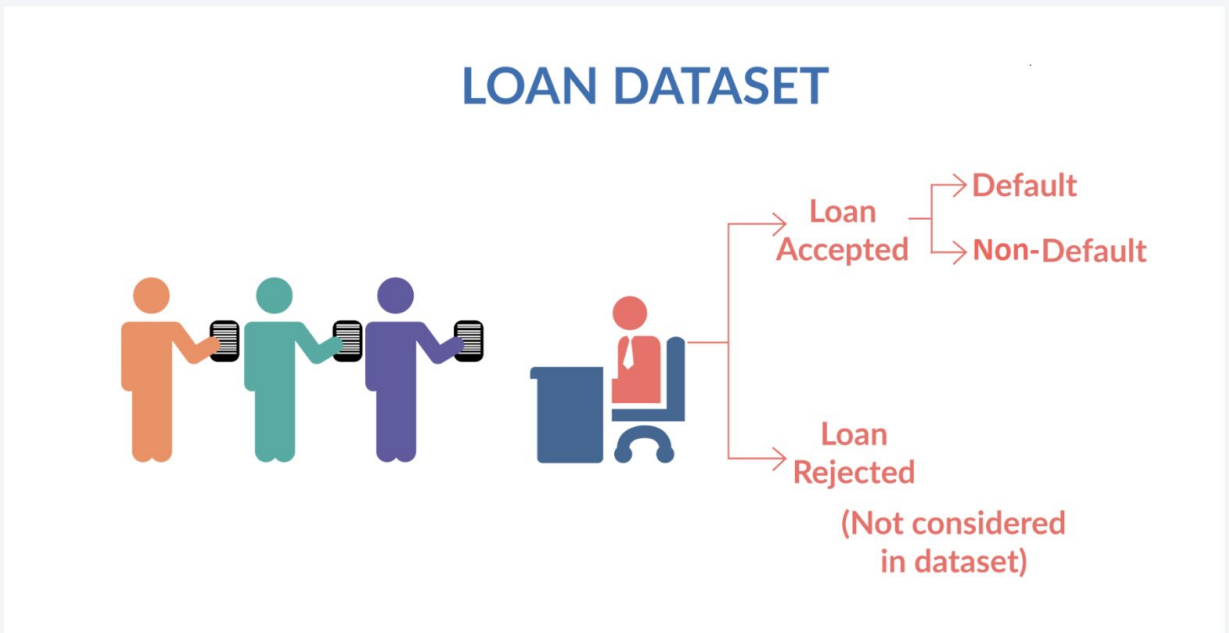


Lending Club Case Study - Loan Dataset

Group Members:

1. Pravin Tawade
2. Naveen

Agenda

- Overview - Business Understanding
- Problem Statement - Business Objectives
- Data in depth
- Approach
- Exploratory Data Analysis on Loan Dataset
- Conclusion

Overview - Business Understanding

This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures.

- ➔ Like most other lending companies, 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. In other words, borrowers who default cause the largest amount of loss to the lenders.
- ➔ In this case, the customers labelled as 'charged-off' are the 'defaulters'. If one 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.

Problem Statement - Business Objectives

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

Want 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 in depth

- We are going to analyze loan datasets,
- The datasets contains details information related of the loans such as loan amount, funded amount, interest rate, repay terms in months etc.
- The dataset comprises of **38717 observations of 111 columns**. Below is a table showing names of the few columns.

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	..
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	..
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	..
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	..
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	..
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	..

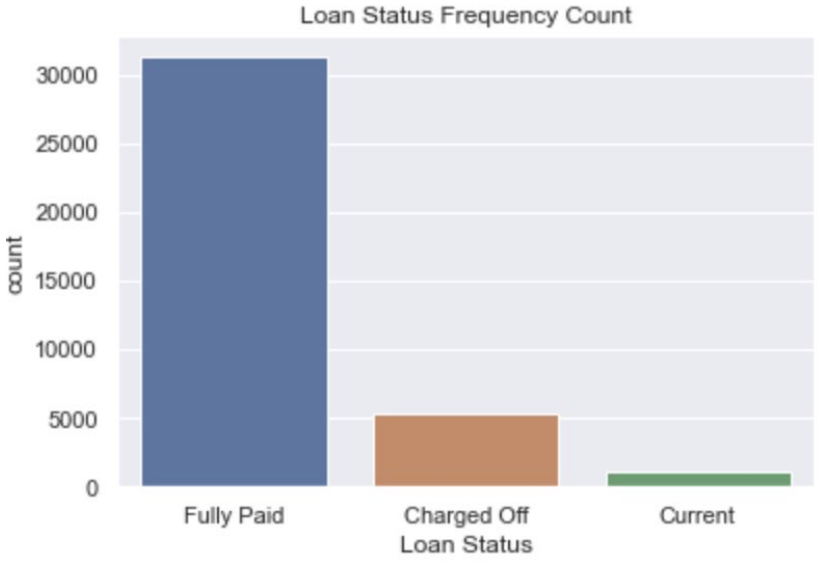
Approach

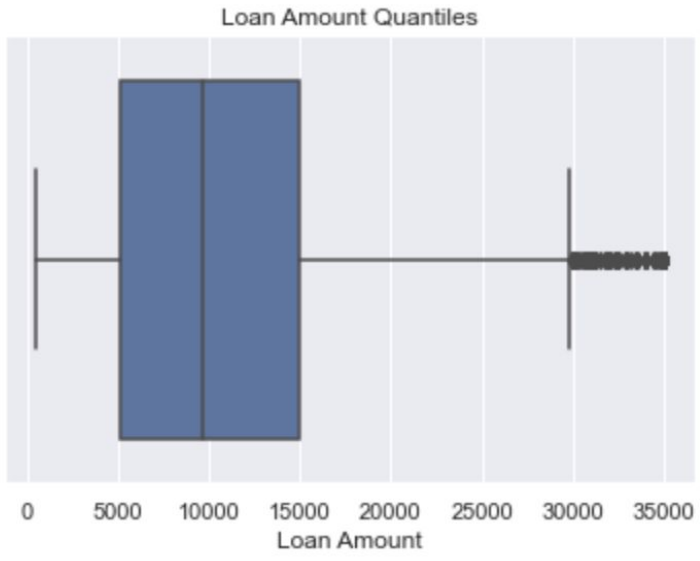
1. Understanding the Dataset
    - To gain insights from data we must look into each aspect of it very carefully. We will start with observing few rows and columns of data both from the starting and from the end.
2. Preprocessing
    - We will deal with erroneous and missing values of columns.
    - Correlation between different columns
    - See how preprocessing have transformed our dataset.
    - Derive new data from existing data to get more insite
3. Exploratory Data Analysis on Loan Dataset
    - Try to find out answers of some set of questions

Lending Club Case Study - Loan Dataset

What is loan status?

- The **loan amount varies** from **0 to 30,000** having **mean of 10,000**. Also there are **few outliers above 30,000**.
- Most of the loans are **Fully Paid**.
- About **14%** of **loan** are having status as **defaulters**.

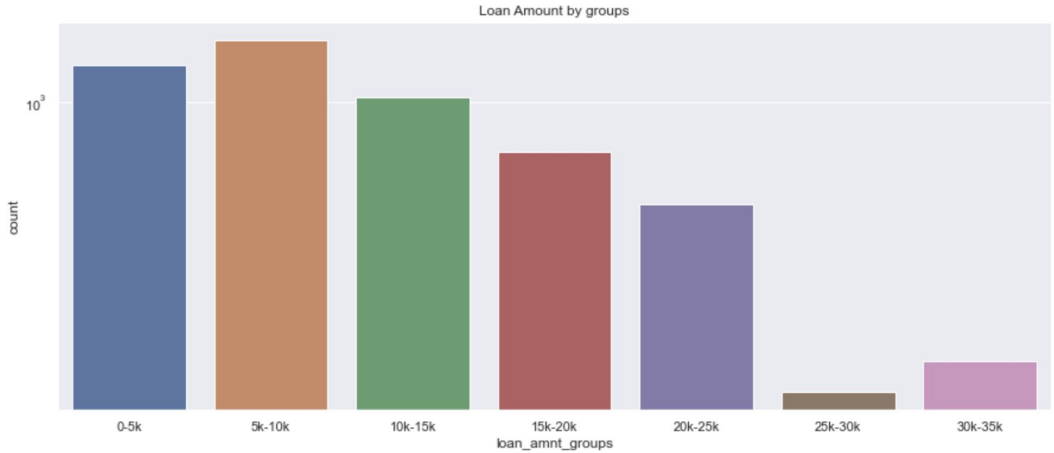


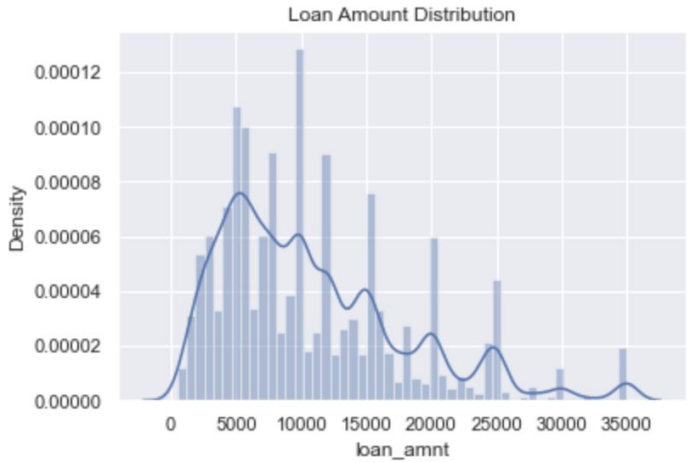


Lending Club Case Study - Loan Dataset

What is the min & max loan amount?

- The **loan amount varies** from **0 to 30,000** having **mean of 10,000**.
- **Loan amount** is also **left skewed**.
- Most of the **loan amount** given is **5K-10K lakhs**

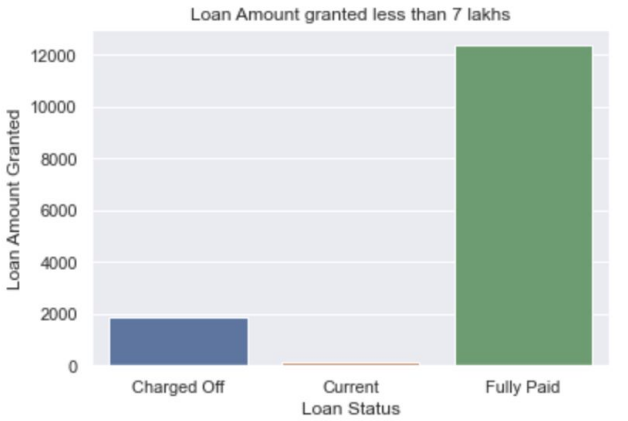


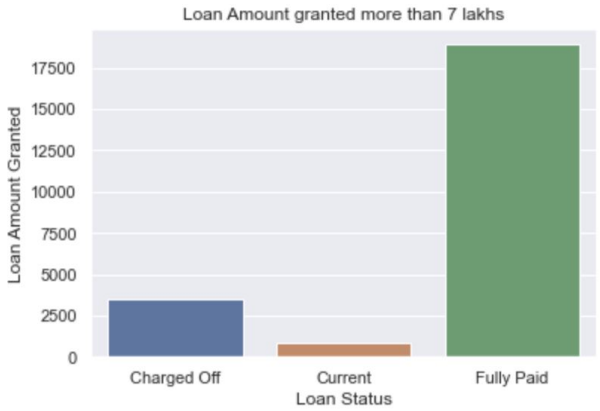


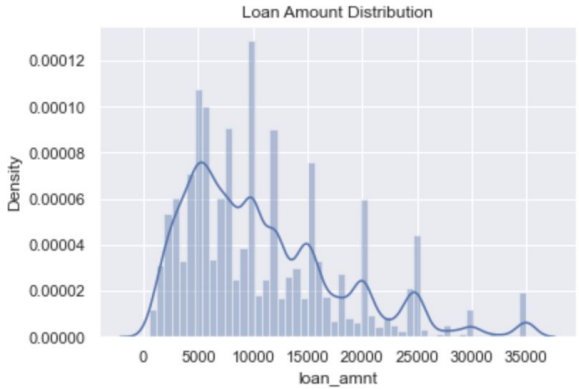
Lending Club Case Study - Loan Dataset

Comparison of the loan status for certain loan amount

- Most of the **loan amount** given is **below 7 lakhs**.
- Probability of people with loan amount **greater than 7 lakhs** tends to **default** is **more** than the people with **less than 7 lakh**

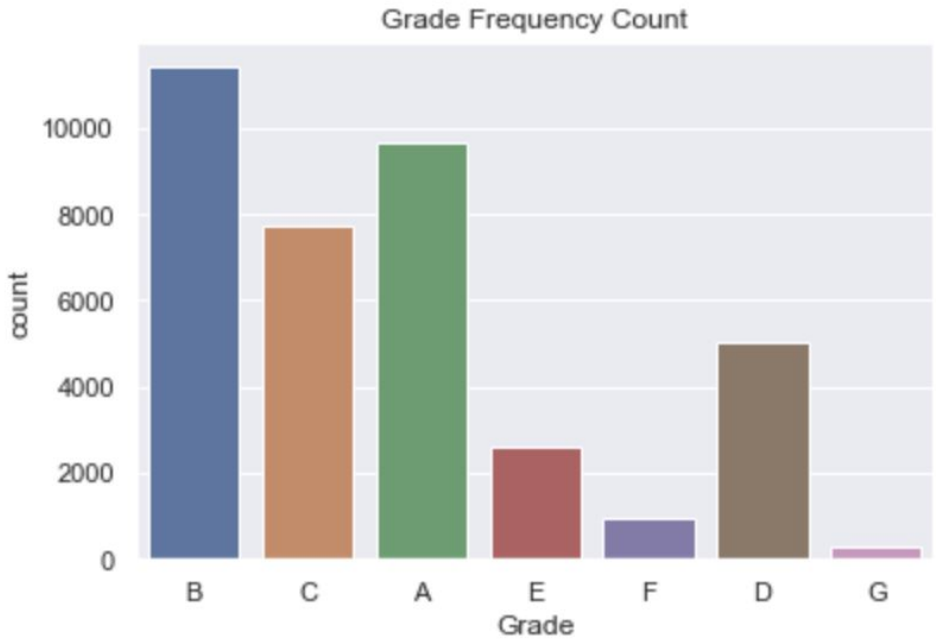






Lending Club Case Study - Loan Dataset… continue

Top Loan Grades that have produced the most number of successful loan application?

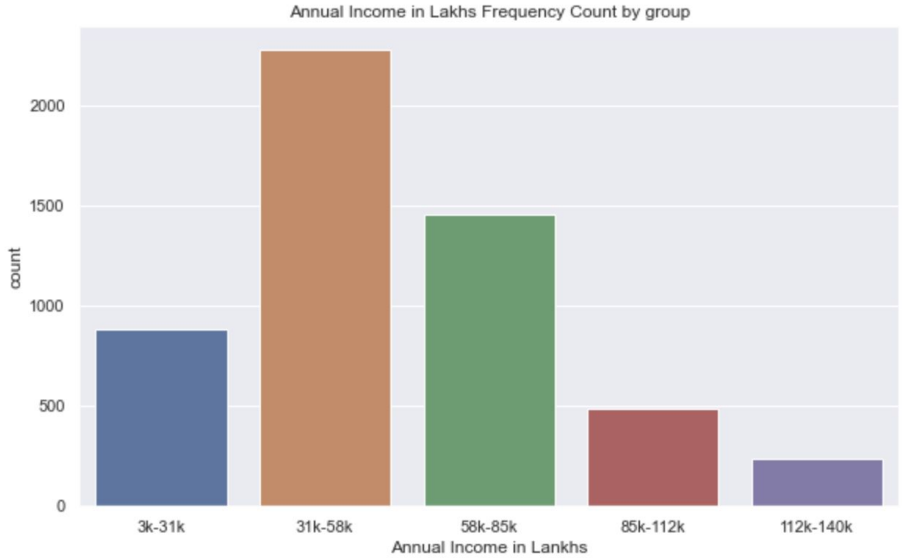


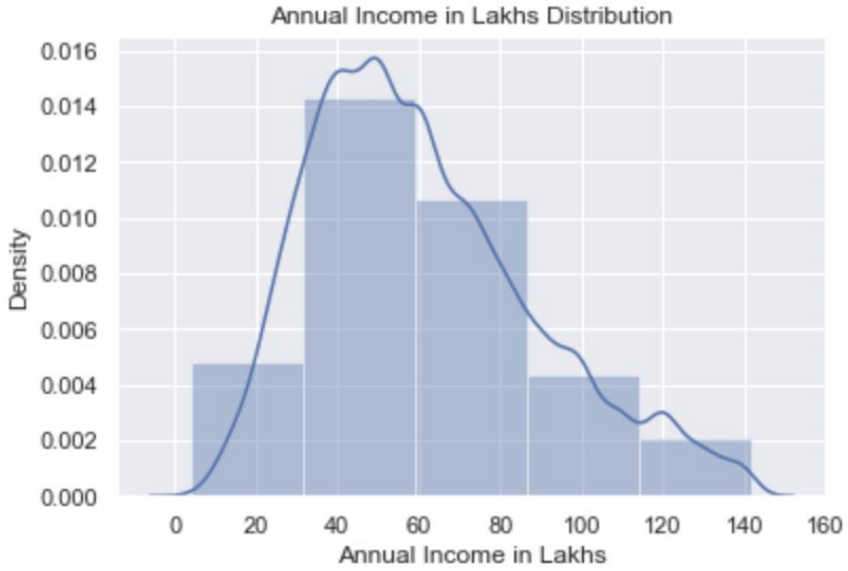
- Most of the loans have **grade** of **A** and B.
- Therefore stating most of the **loans are high graded loans**.

Lending Club Case Study - Loan Dataset… continue

What is the higher income range for loan applicant?

- Most of the application having **highest income** range **31K-58K** followed by 58K-85K

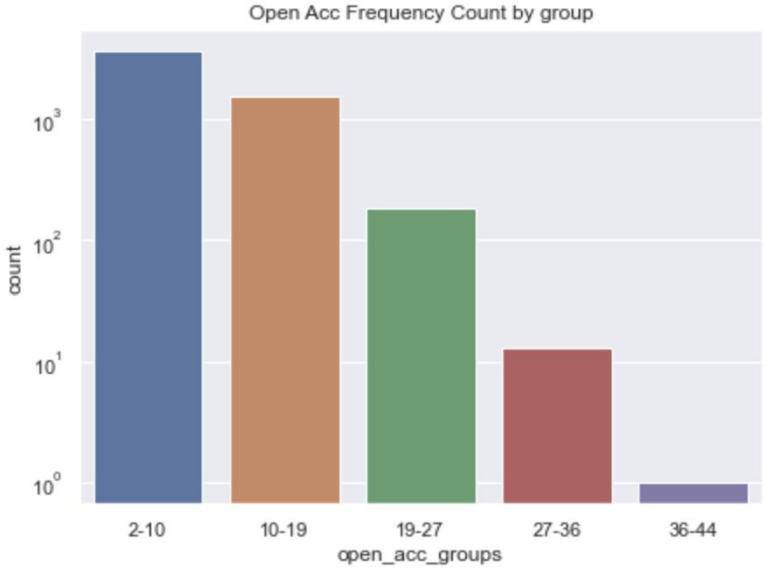


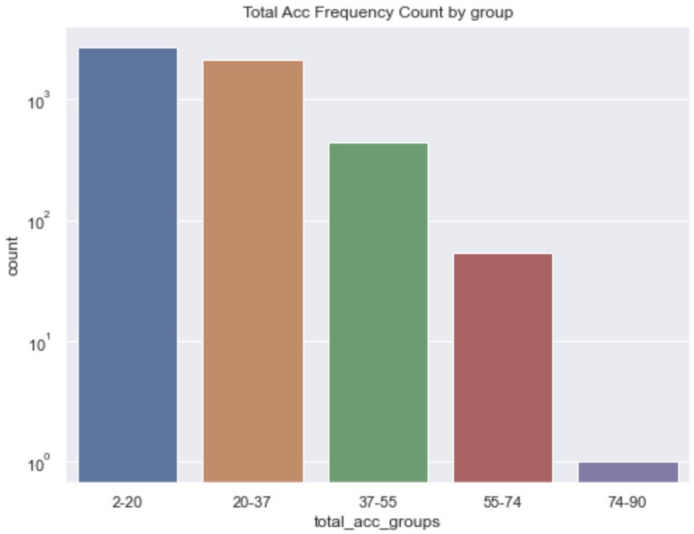


Lending Club Case Study - Loan Dataset… continue

What is the higher open acc & total acc range for loan applicant?

- Most of the application having **highest open acc 2-10** followed by 10-19
- Also application having **highest total acc 2-20** followed by 20-37

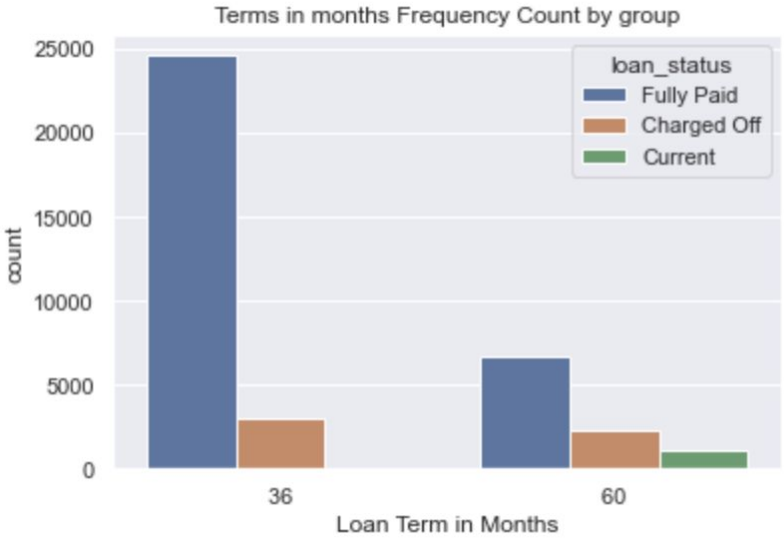


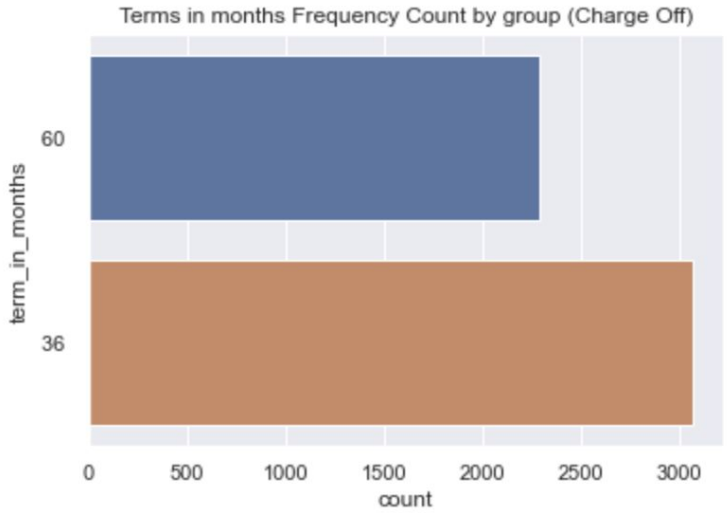


Lending Club Case Study - Loan Dataset… continue

Which tenure in months have have high chances to be defaulters?

- **Tenure of 36 months** have have **high chances** to be **defaulters**

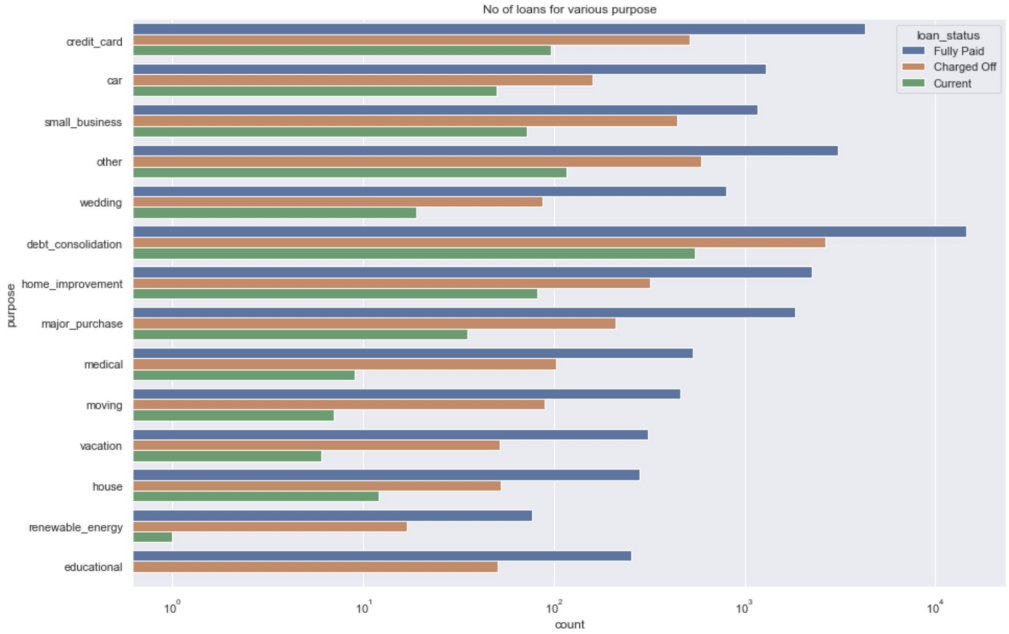


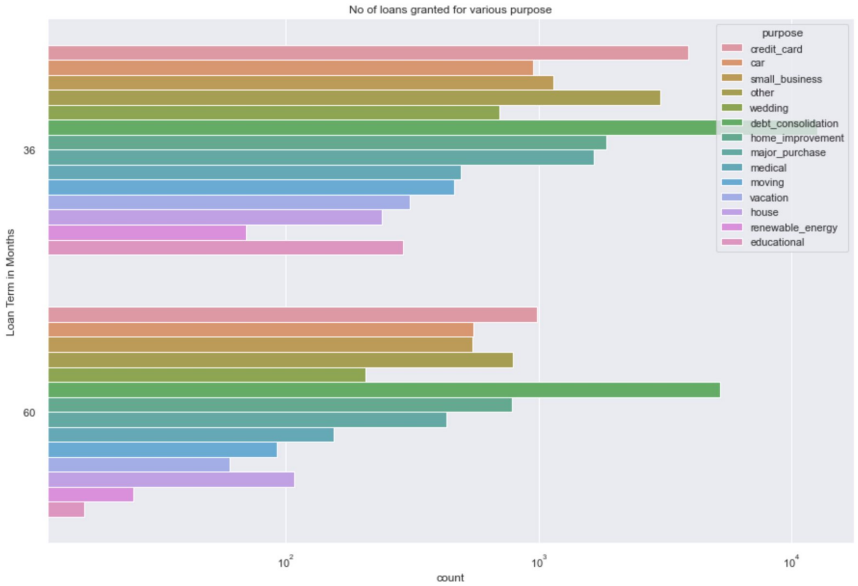


Lending Club Case Study - Loan Dataset… continue

What kind of loan purpose having high probability for defaulter's?

- Majority of loan has been given for the **debt consolidation** purpose having high probability for **defaulter's**

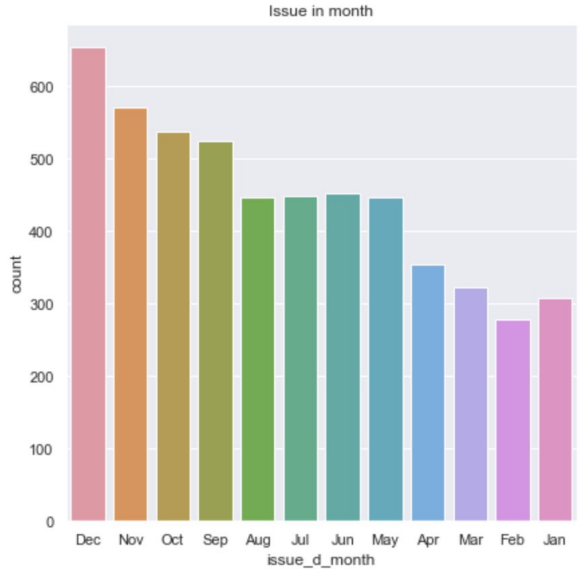


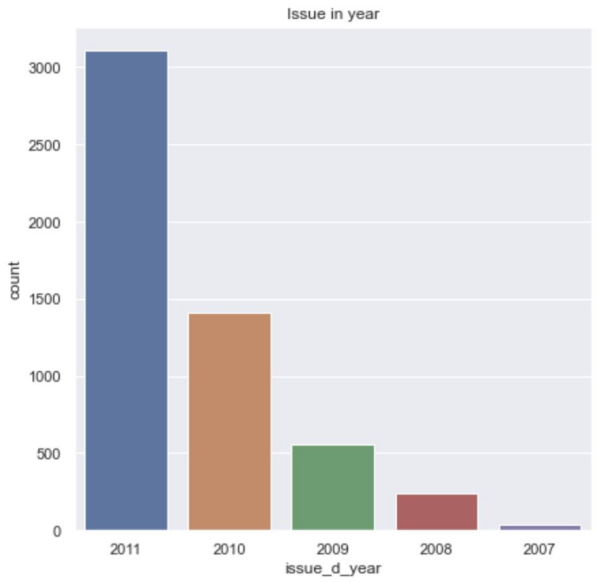


Lending Club Case Study - Loan Dataset… continue

What kind of pattern do we have for loan distribution over the year and month?

- Maximum **number of defaults occured** when the loan was **sanctioned/issued in Dec**.
- Also Loan Applicants have been increasing **exponentially** from year 2007 to 2011

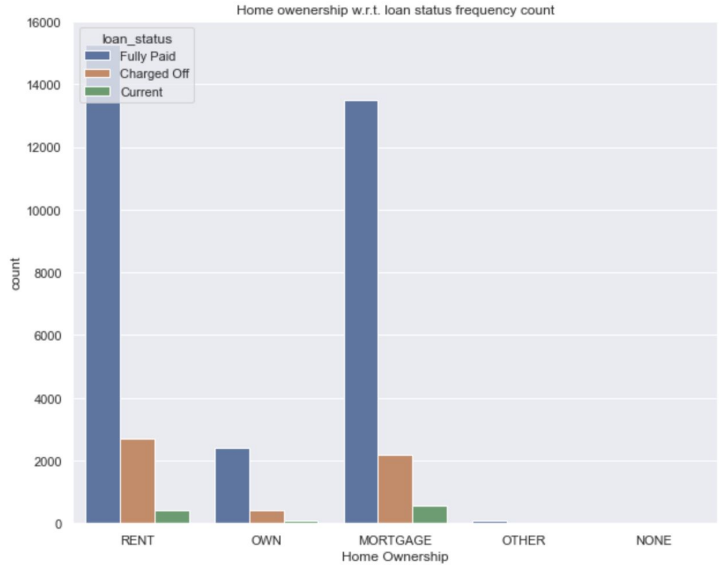


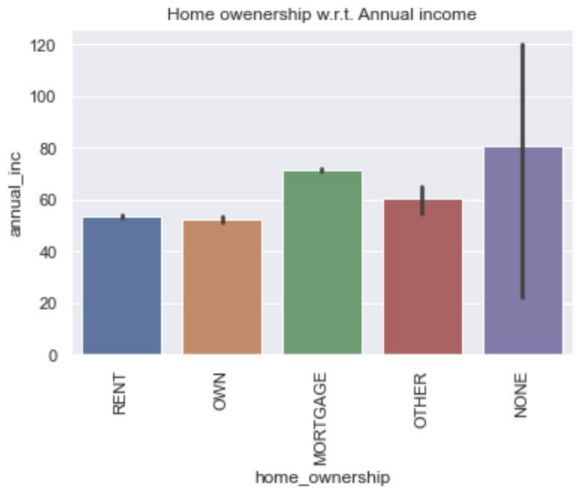


Lending Club Case Study - Loan Dataset… continue

Does a home ownership play major roles in loan defaulter?

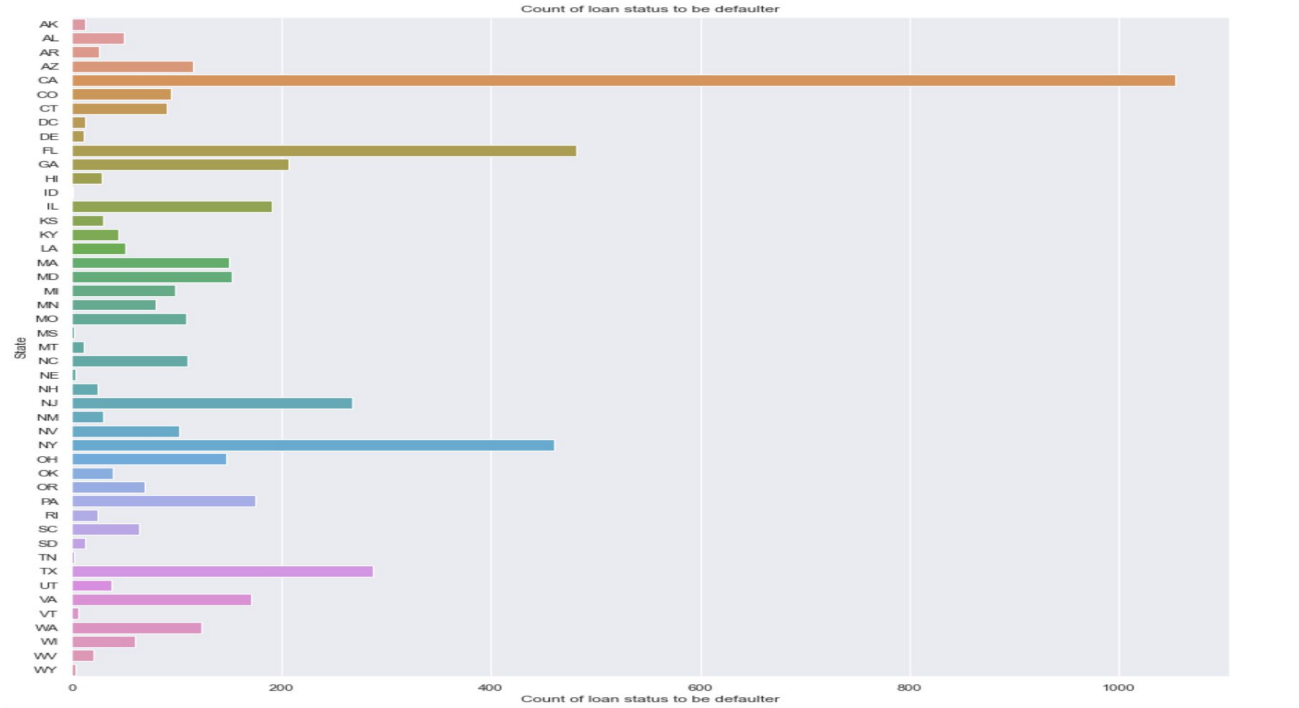
- There is high probability if applicants taking loan whose **home ownership is 'MORTGAGE** and have income of 31k -58k





Lending Club Case Study - Loan Dataset… continue

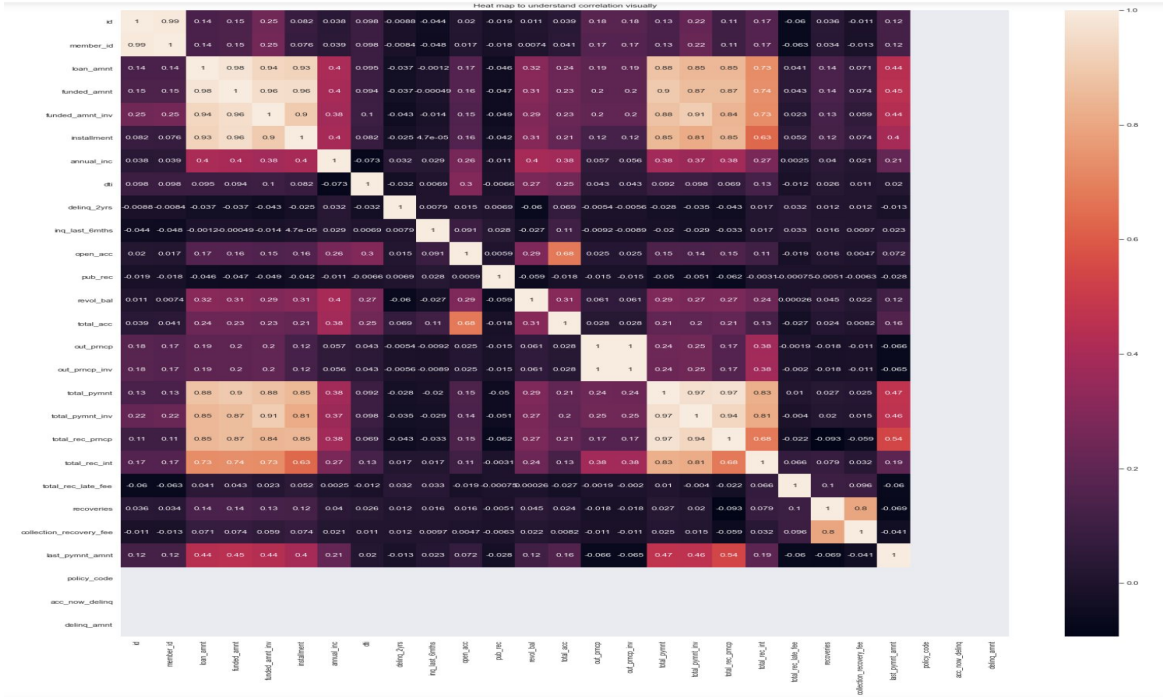
Should we also consider a residential state in the loan approval process?



- It's seem that applicants from **state CA** have high tendency to default the loan

Lending Club Case Study - Loan Dataset… continue

Which factors play a key role in deciding loan approval will be.



While pre-processing, we can clearly say all other features didn't make much difference in results, except below mentioned variables which has relationships & will help in lto decide loan defaulters:

- **loan amount**
- **interest rate**
- **installment,**
- **repayment term,**
- **home ownership,**
- **loan purpose,**
- **residential state**
- **income group**

Conclusion

Based on our analysis there is high probability if applicants taking loan:

- for '**debt consolidation**' and have **income** of **31k -58k**
- whose **home ownership** is '**MORTGAGE** and have **income** of **31k -58k**
- when **grade** is **A** and **loan amount** is between **31k -58k**
- when **dti** is between **12-18** and **loan amount** is between **31k -58k**
- maximum number of defaults occured when the loan was sanctioned/issued in Dec

We can see **loan amount** & **interest rate, installment, repayment term, home ownership, loan purpose, residential state** and **income** group plays an important role's to decide loan defaulters.

Thank you