# LENDING CLUB CASE STUDY

# PROBLEM STATEMENT

The consumer finance company faces two types of risks when deciding 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.

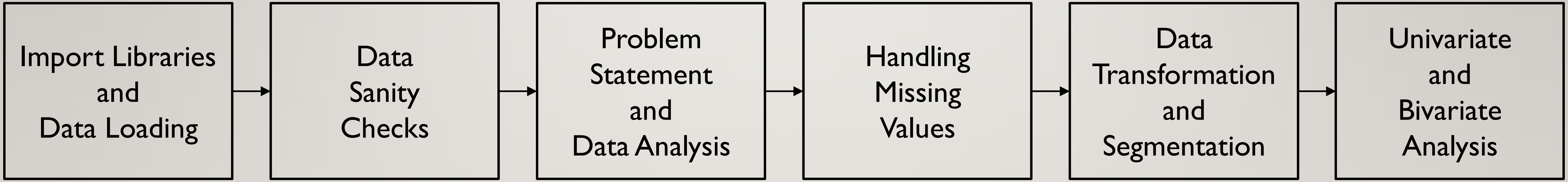
**Identify the driving factors behind loan default** thereby empowering the company in making a wise decision on loan approval to minimize the financial loss to the company.

# ASSUMPTIONS

- Decision of loan approval is taken when we have two kinds of information:
  1. The data filled in by the consumer in the loan application.
  2. Background verification and consumer specific attributes identified by the company.

  Hence, the variables which are related to information post the loan approval, will not be relevant candidates for the decision of loan approval.
- Variables containing random information such as identifiers, names, descriptive fields, etc. are not relevant to analysis.
- Constant (single) valued variables will not provide any insights since they have the same value across the dataset.
- Masked variables will not be useful for analysis due to incomplete information available for the respective variable.

# APPROACH



Import Libraries and Data Loading → Data Sanity Checks → Problem Statement and Data Analysis → Handling Missing Values → Data Transformation and Segmentation → Univariate and Bivariate Analysis

# APPROACH

1. Data for only "Fully Paid" and "Charged Off" consumers is considered for analysis since information from "Current" consumer data does not give insight about the loan default.
2. Variables not useful for analysis are dropped from the dataset as highlighted under Assumptions.
3. To handle missing values:
   a. Variables with only NA are dropped since no analysis can be done for them.
   b. Variables having acceptable range (<40%) of missing values are imputed.
      - Median is used for imputation in the case of numerical variables. (Since median is a better representative than mean.)
      - Mode is used for imputation in the case of categorical variables.
   c. Variables having high percentage (>=40%) of missing values are dropped (since imputation for more than half of the dataset for the respective variable will be not useful for analysis).

# APPROACH

4. Transformation on datatypes is performed for the variables, int_rate, annual_inc and pub_rec_bankruptcies.

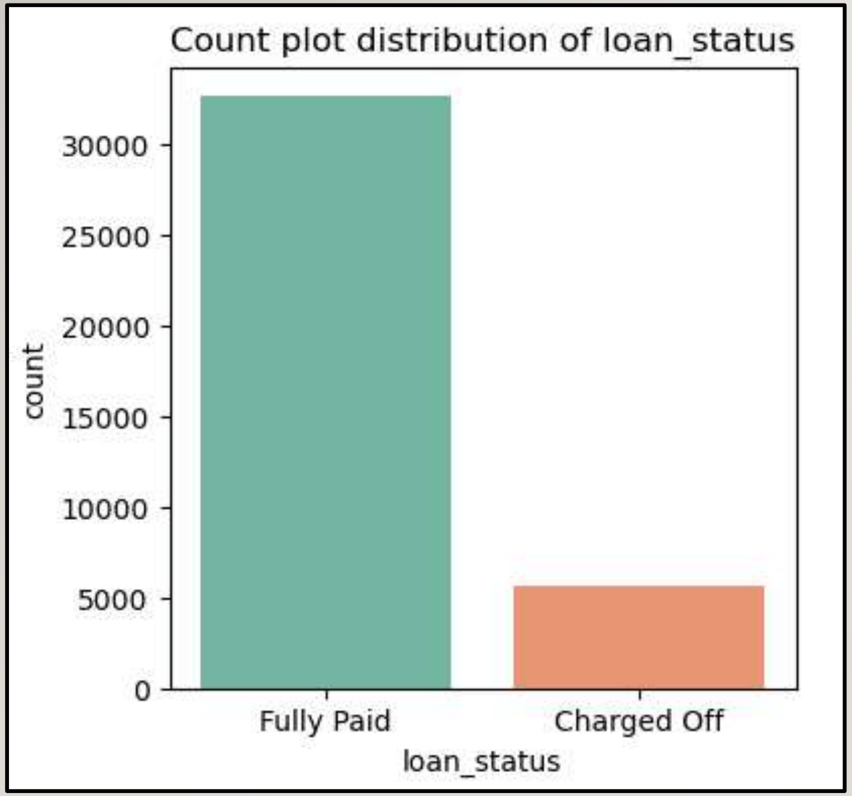
5. Outliers are handled for the variable annual_inc.

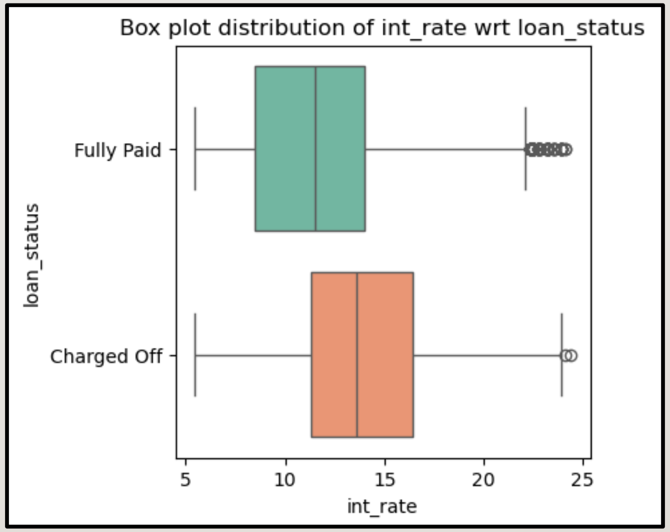
6. Derived metrics are created:
   - Month and year are extracted for each of the variables issue_d and earliest_cr_line, to separately analyze the impact.
   - We have two variables grade and subgrade. Since the grade value is part of the subgrade variable, we have extracted the actual subgrade value to perform analysis on the impact of grade and subgrade separately.

7. Segmentation into numerical and categorical variables is done for easier (similar) analysis on the target variable.

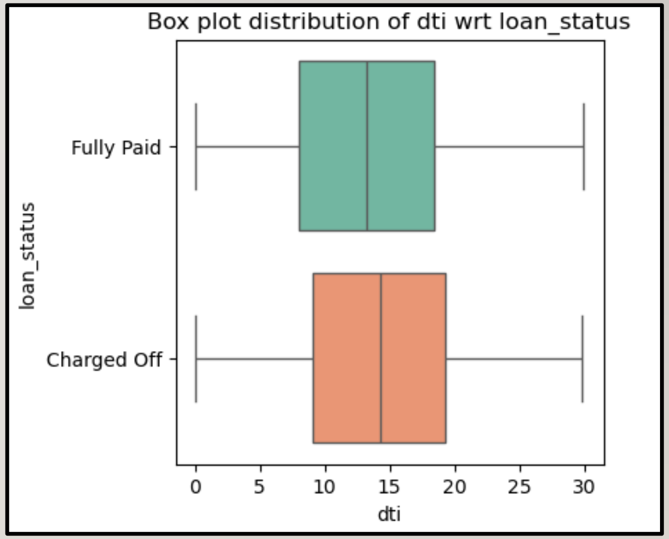
All trends, distributions, plots for every feature variable is analyzed against both the "Fully Paid" and the "Charged Off" loan statuses. This helps to identify whether the respective feature variable has an impact and is correlated to the possibility of loan default or has just a general behavior irrespective of the loan status.

# OBSERVATIONS



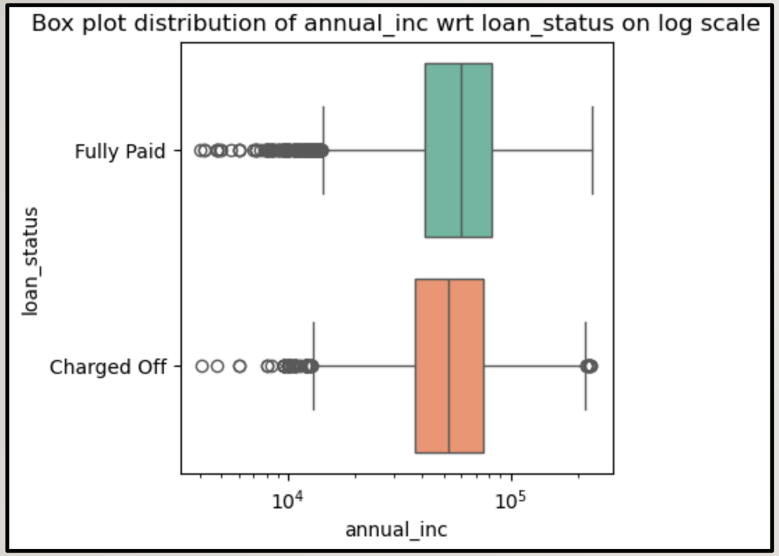


Higher value of **int_rate** can lead to higher chances of loan default.

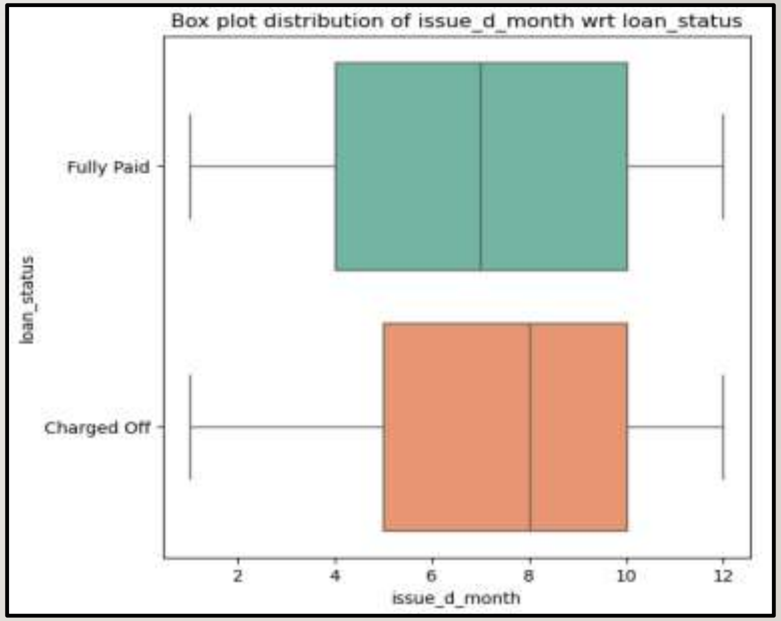


Higher value of **dti** can lead to higher chances of loan default.

# OBSERVATIONS

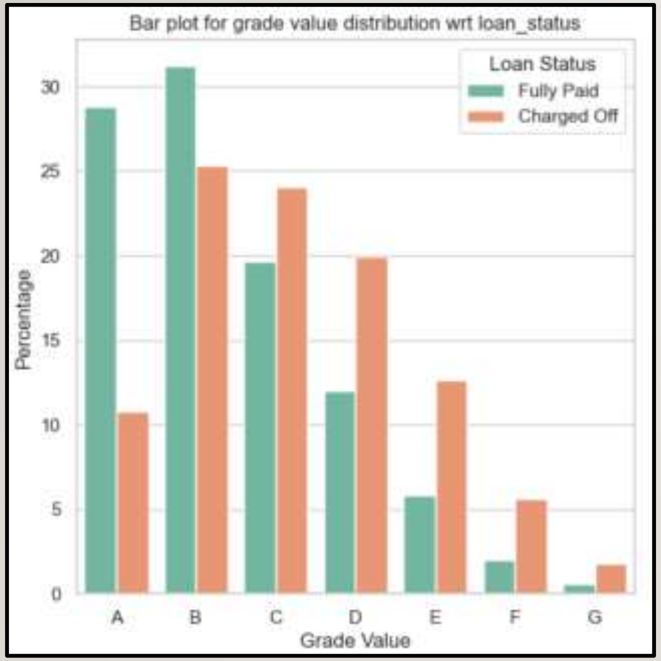


Lower value of **annual_inc** can lead to higher chances of loan default.

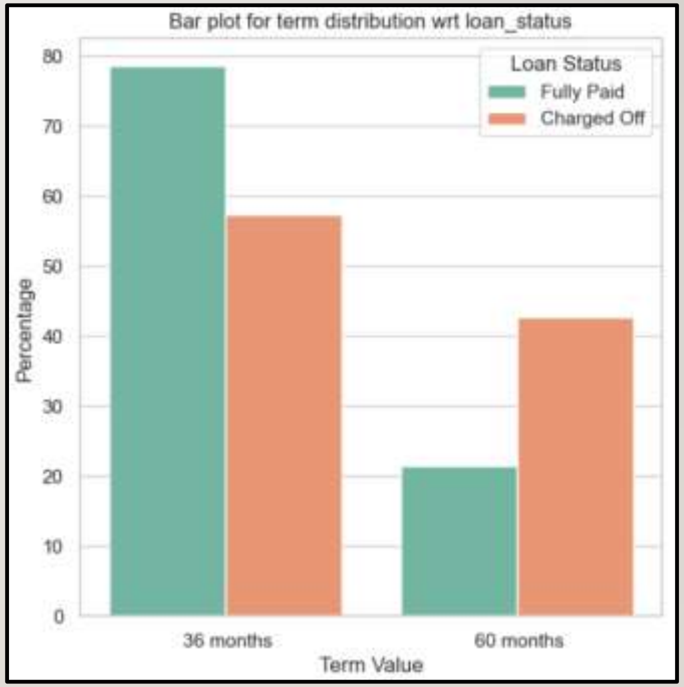


Higher value of **month in issue_d** can lead to higher chances of loan default.

# OBSERVATIONS

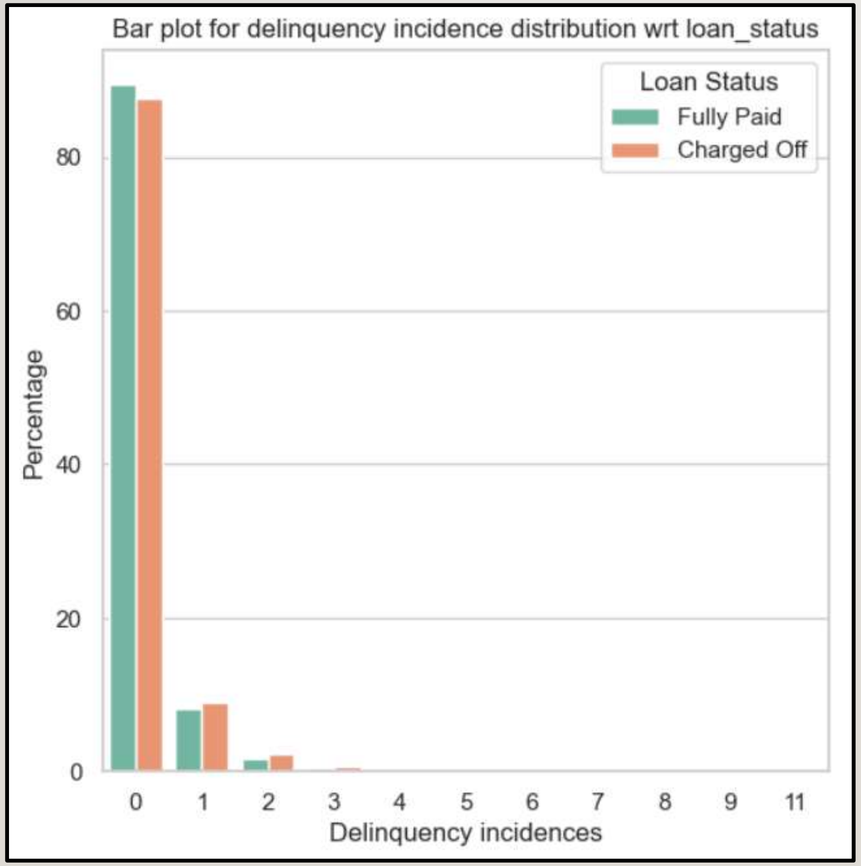


Lower value of **grade** (B, C, D etc.) can lead to higher chances of loan default.



Higher value of **term** duration can lead to higher chances of loan default.

# OBSERVATIONS



Higher value of **delinq_2yrs** i.e. number of delinquency incidents in the past 2 years, can lead to higher chances of loan default.

# CONCLUSION

Major driving factors behind loan default are mentioned below:

- Positively correlated variables:
  - **int_rate**: Higher interest rate implies higher probability of loan default.
  - **dti**: Higher dti ratio implies higher probability of loan default.
  - **issue_d**: Higher month part of issue_d (loan applicants which have issue date in the later months of the year especially December) implies higher probability of loan default.
  - **term**: Higher term duration (60 months) implies higher probability of loan default.
  - **delinq_2yrs**: Higher number of delinquency incidences in the past 2 years implies higher probability of loan default.
- Negatively correlated variables:
  - **annual_inc**: Lower annual income implies higher probability of loan default.
  - **grade**: Lower grade value (B, C, D etc.) implies higher probability of loan default.

# CONCLUSION

Other driving factors slightly impacting loan default are mentioned below:

- **loan_amnt**: Higher loan amount slightly increases the probability of loan default.
- **inq_last_6mths**: Higher number of inquiries made in the last 6 months slightly increases the probability of loan default.
- **open_acc**: Lower number of open accounts slightly increases the probability of loan default.
- **home_ownership**: Home ownership value "RENT" slightly increases the probability of loan default.
- **purpose**: Purpose of availing loan being "debt_consolidation" slightly increases the probability of loan default.
- **address_state**: Loan applicant's address state being "CA", "FL" slightly increases the probability of loan default.