**Grade: 100%**

Risk Assessment for Loan Investment

Group 7

Sheetal Chowdhary

Maaz Kamal

Rohit R Patil

# Objective

LendingClub is a peer-to-peer loan lending platform. It enables borrowers to get unsecured loans between $1000 and $4000. The investors can then search and lend money to borrowers based on the information about the borrower, the loan amount, and the purpose of the loan.

For evaluating the creditworthiness of their borrowers, Lending Club relies on many factors related to borrowers such as credit history, employment, income, ratings etc. Lending club then assigns rating/sub-rating to their borrowers based on their credit-history. This rating information is then made available to investors who fund the loan requests. Investors use this information to analyze loan request and adjudicate the approved funded amount. In addition to the grade information, Lending Club provides historical loan performance data to investors for more comprehensive analysis. Borrowers with higher credit score get lower interest rate, whereas borrowers with low credit score get higher interest rates. From the investors perspective lending to borrowers with high interest rate seems more profitable as it will give a higher Return on Investment.

But at the same time there is a risk of the loan not being returned. As per the recent studies, 3-4% of the total loans defaults every year. There is a huge risk for the investors who is funding the loans. Investors require more comprehensive assessment of these borrowers than what is presented by Lending Club to make a smart business decision. Data mining techniques and Machine Learning modelling/analysis could help predicting the loan default likelihood which may allow investors to avoid loan defaults thus limiting the risk of their investments. The goal of our project is to assist such investors in predicting which high interest loans are more likely to be returned and help them in finding worthy borrowers to lend their money.

# Data Set Description

We are using [Lending Club dataset provided by Nathan George](https://www.kaggle.com/wordsforthewise/lending-club)  from Kaggle. The dataset contains 151 columns and 2260701 rows which are a mix of categorical and numerical type. This dataset contains complete loan data for all loans issued through the 2007-2018, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information. The file contains complete loan data for all loans issued through the previous completed calendar quarter. Additional features include credit scores, number of finance inquiries, address including zip codes, and state, and collections, among others.

Some of the variables that are relevant for our model building would be the borrower’s income, employment length, FICO score, debt-to-income ratio, the loan amount, loan purpose, loan grade, interest rate and installment. All these variables could be a good potential predictor for our model. There might be few other predictors which might not be obvious to us now. Therefore, this is not the final list, and we plan to explore for more predictors. Since the data is quite extensive, doing this project will help us learn how to work efficiently with Big Data.

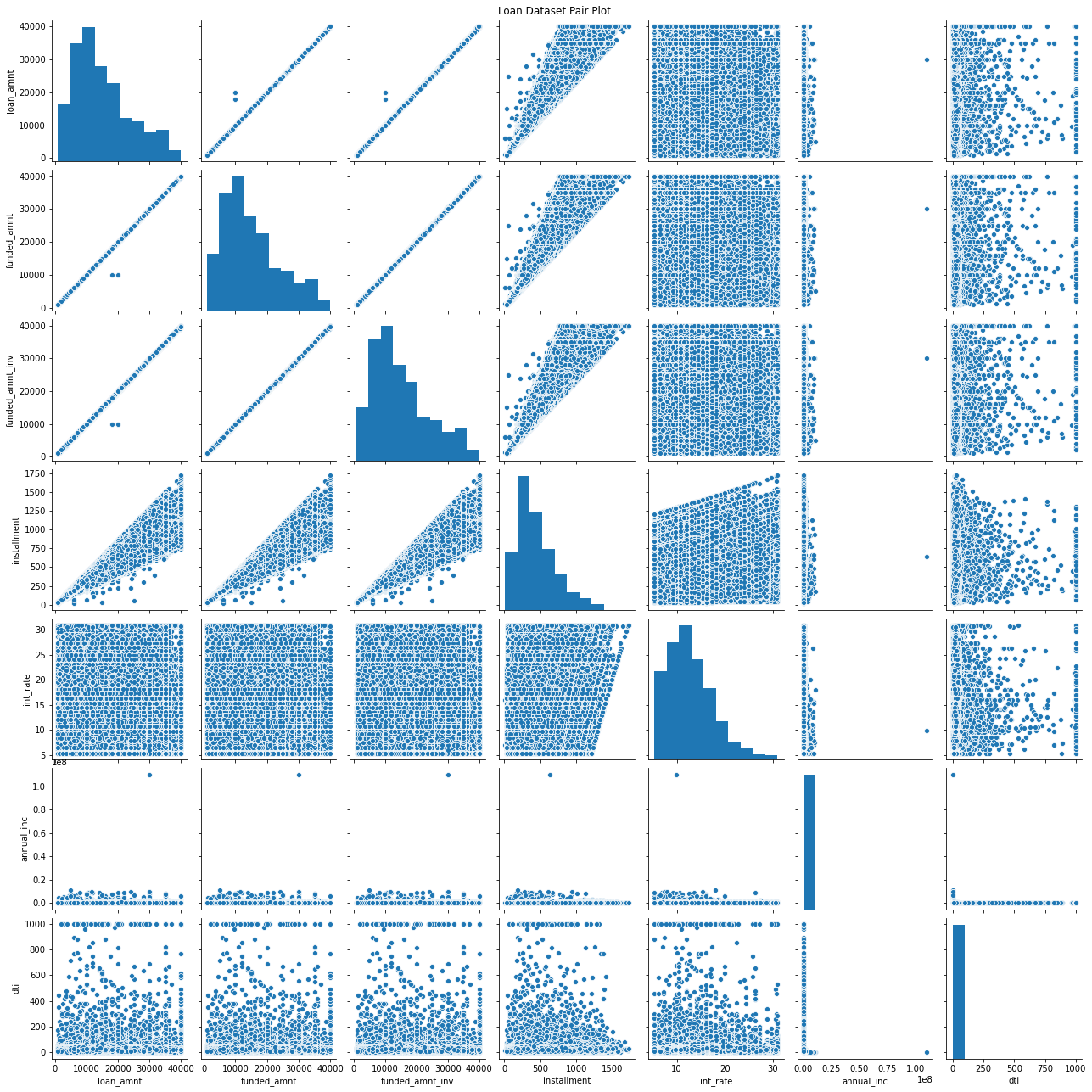
The data can be found on the following link: https://www.kaggle.com/wordsforthewise/lending-club

# Preliminary Data Exploration

The original dataset is about 1.6 GB size and we faced issues with uploading it on Databricks. Therefore, we are using only half of the dataset, with 1048575 rows and 151 columns for the purpose of this project. Please note that if we manage to load the original 1.6 GB data on Colab environment we will go ahead with it else will use half of the dataset that we have used for EDA.

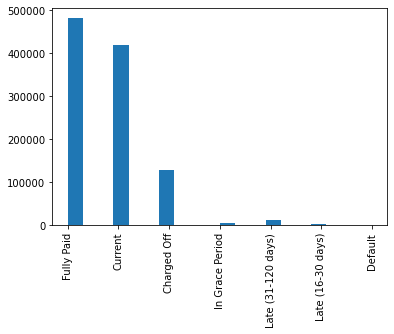
After initial null value analysis, we found that there are 43 columns with more than 50% null values. We dropped these columns as it is not feasible to impute missing values for those columns. After dropping these values, we are left with 108 columns. We also dropped some other irrelevant columns like ‘url’ which are of no use in our modelling and analysis.

**Loan Dataset Pair Plot**



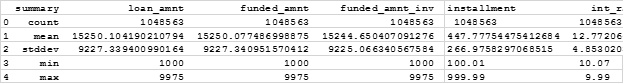
Looking at the pair-plots, it is clear that other than a few predicting variables, there is no real correlation which might be helpful when using linear regression.

**Loan Status Histogram**



The “loan\_status” appears to be unbalanced. It might be a good idea to lump some of the less occurring labels. We will be using ‘Fully Paid’ and ‘Charged Off’ loan records to perform classification and build a model to find probability of a loan being paid off or charged off.

**Summary statistics of some important predicting variables**



# Predictions

We plan to predict the following from our models -

1. Probability that a particular borrower returns the loan - borrower will default the loan or payoff
2. Fraction of the expected loan return that a prospective borrower will pay back
3. Total Return on Investment for the investor
4. Time Period in which investor can expect return of loan amount

# Inference

We plan to gain insights about the following points -

1. Which factors affect the burrower to charge-off or default the most?
2. Is the risk of investing in high interest rate loans good in the long run? Is the expected return on high risk investments positive?
3. What factors affect the total return on investment for the investor and to what extent?

# Non-spark packages

We do not plan to use any other packages outside of spark, pandas, NumPy, matplotlib and seaborn.