2020

# Capstone Project I Milestone Report

## 2018 LENDING CLUB'S LOAN DATA

NGOC PHAN

## Project Background

LendingClub (LC) is an online credit marketplace that enables a borrower to apply for a loan and an investor to select a loan to invest. When a borrower applies for a loan at LC, the company will screen the applicant. If the loan application gets approved, LC will provide the borrower the interest rate for the loan. Once the borrower accepts the loan, the loan is made available for the investors to select. The investor may choose to invest in a whole loan or a fractional part of a loan.

## Problem Statement

This project studies the 2018 LC loan data to assist prospective investors in making investment decision by:

- Exploring the characteristics of fully paid and default loans.
- Estimating the return on investment and loss of investment.
- Developing machine learning algorithms that predicts the likeliness of a default loan.

## Dataset

The loan datasets are collected on LendingClub website at https://www.lendingclub.com/info/download-data.action that include 2018 loan data by quarter. Each dataset is in csv format and stored in a zip file. There are 495,242 rows and 144 columns in the dataset. All columns contain cleaned data, and no duplicates have been found. The following rows are ignored when reading the csv files:

- The first row of each csv file that contains general note.
- The last two rows of each csv file that contain the total amount funded in policy code 1 and 2.

Columns id, member_id, url, and desc contain no values and are excluded from the data.

# Data Preparation

## Missing Values

Remove columns that have more than 25% of missing values. There are 39 columns. The missing-value percentage for those columns ranges from 55% to 99.73%. Since the missing-value percentage for those columns is high, it is necessary to remove the columns. Refer to appendix A for a list of columns that have high percentage of missing values.

For columns that have less than or equal to 25% of missing values, fill in the median value for numerical variables and leave the values as null for non-numerical variables. There are 5 non-numerical columns and 12 numerical columns. The missing-value percentage for those columns ranges from 0.01% to 18.79%. Appendix B shows a list of columns that have low percentage of missing values.

## Outliers

Compute z-scores to obtain the records and variables that contain outliers (z-score < -3 or z-score > 3). There are 195,492 rows and 71 columns that contain outliers. Since the number of outliers in the dataset is high (195,492 / 495,242 = 39%), it is necessary to keep the outliers because they may contain significant information, and there are also some models that work well with outliers.

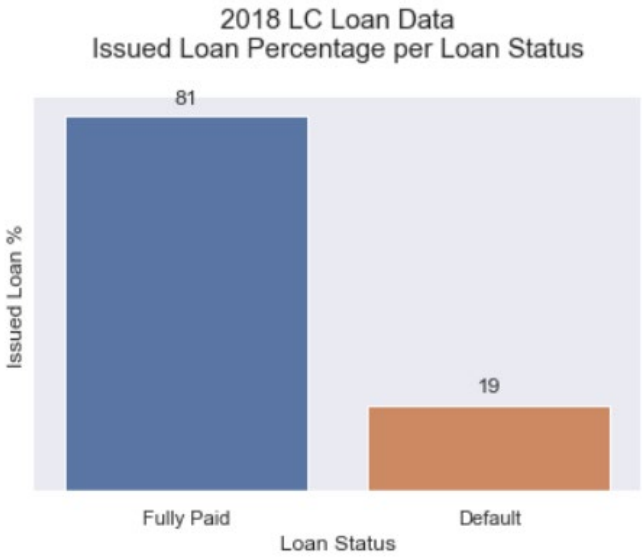
## New Columns

- Return on Investment (ROI) = total payment amount – loan amount
- Months on Loan = last payment date – loan issued date

# Exploratory Data Analysis

## Number of Issued Loans in 2018

In 2018, LendingClub issued 93,853 loans in which 81% of the loans were paid-off and 19% of the loans were defaulted.

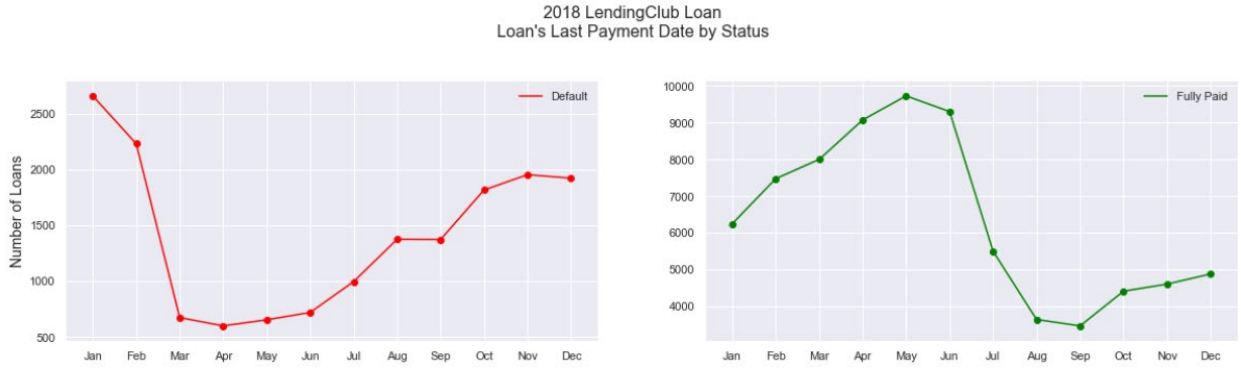



The trend for the number of issued loans per month is downward. LC issued more loans in the first five months of 2018. The number of issued loans decreased at a higher rate after May.

## Loan's Last Payment Date

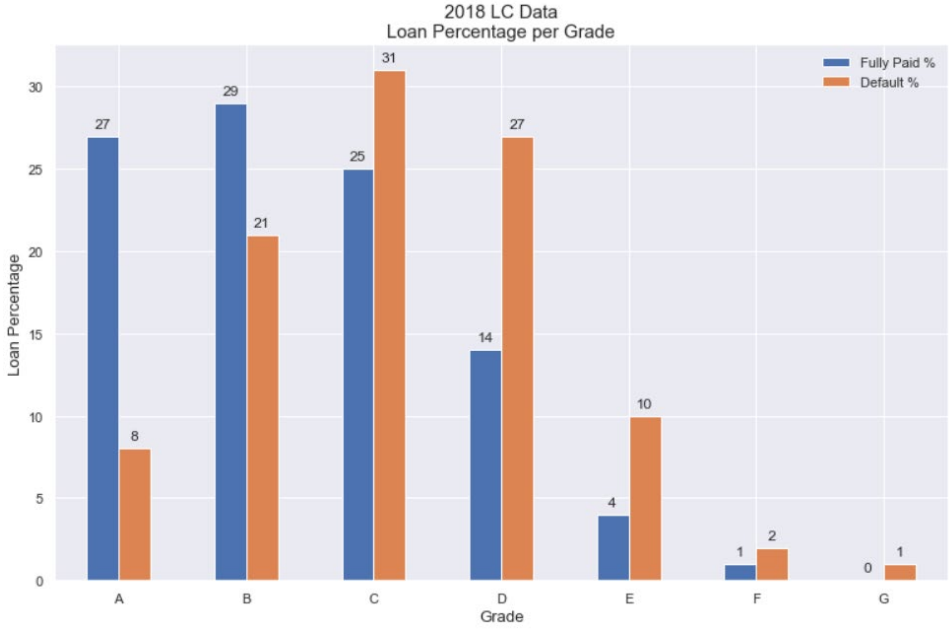
Many borrowers made their last payment on January and February before defaulting on their loans. The number of last payment date is low between March and June for default loans, and then increased at a higher rate after June.

In contrast to default loans, many fully paid loans were paid off during the first 6 months of 2018. The number of paid off loans was sharply decreased from June to August and then slow increased until December.

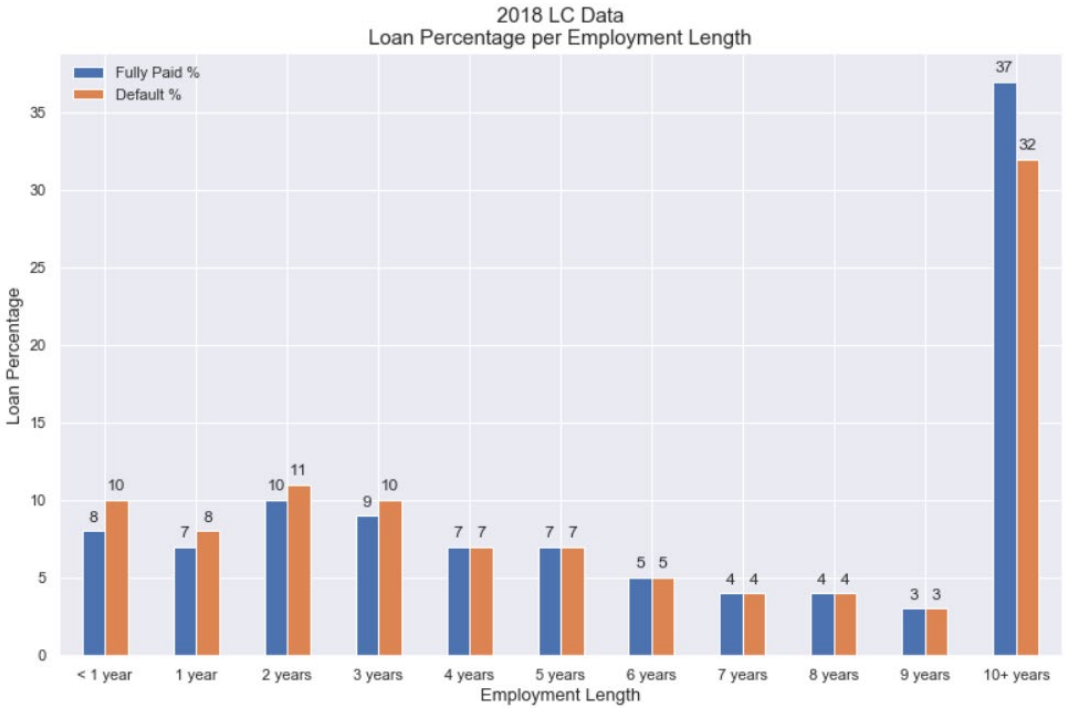



## Data Distribution

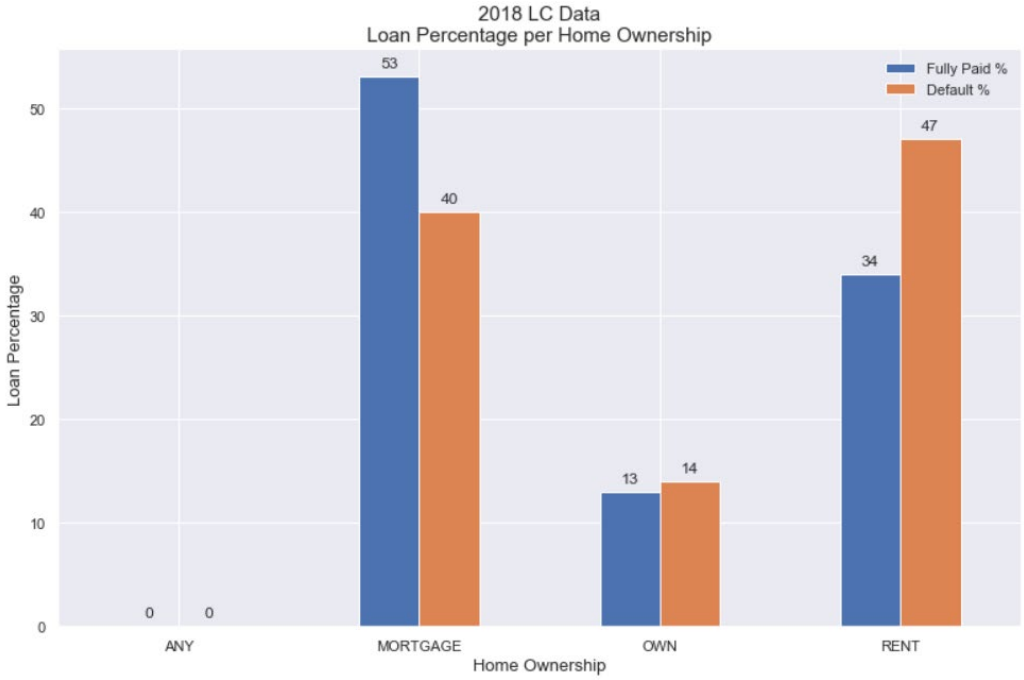
Most fully paid loans have a grade of A, B, or C while most default loans have a grade of B, C, D.



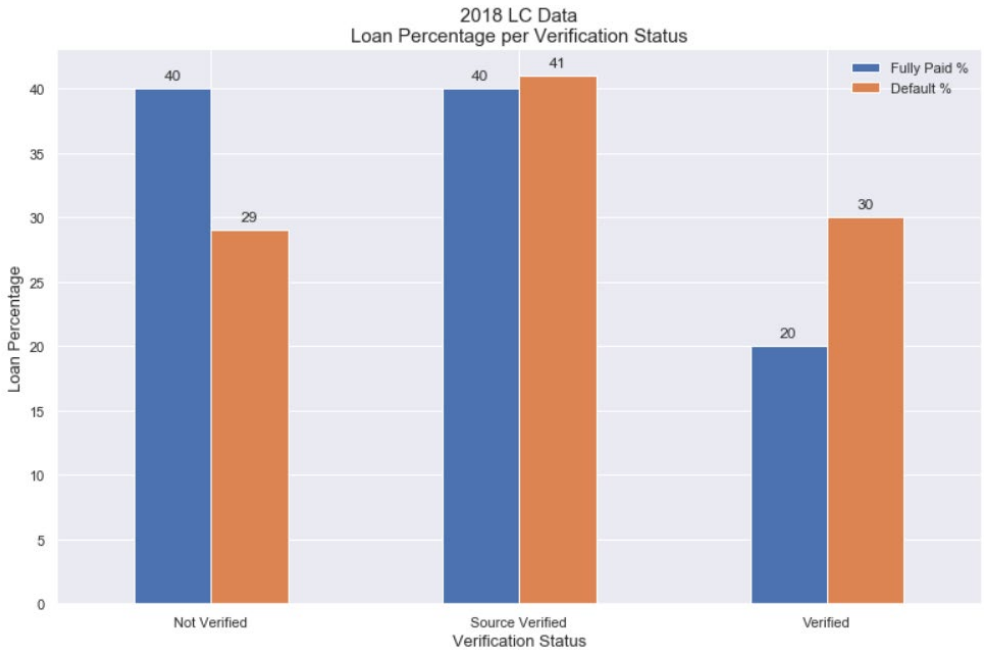
The distribution of employment length for both fully paid and default loans are similar. Both have high percentage of borrowers who have been employed 10 or more years.



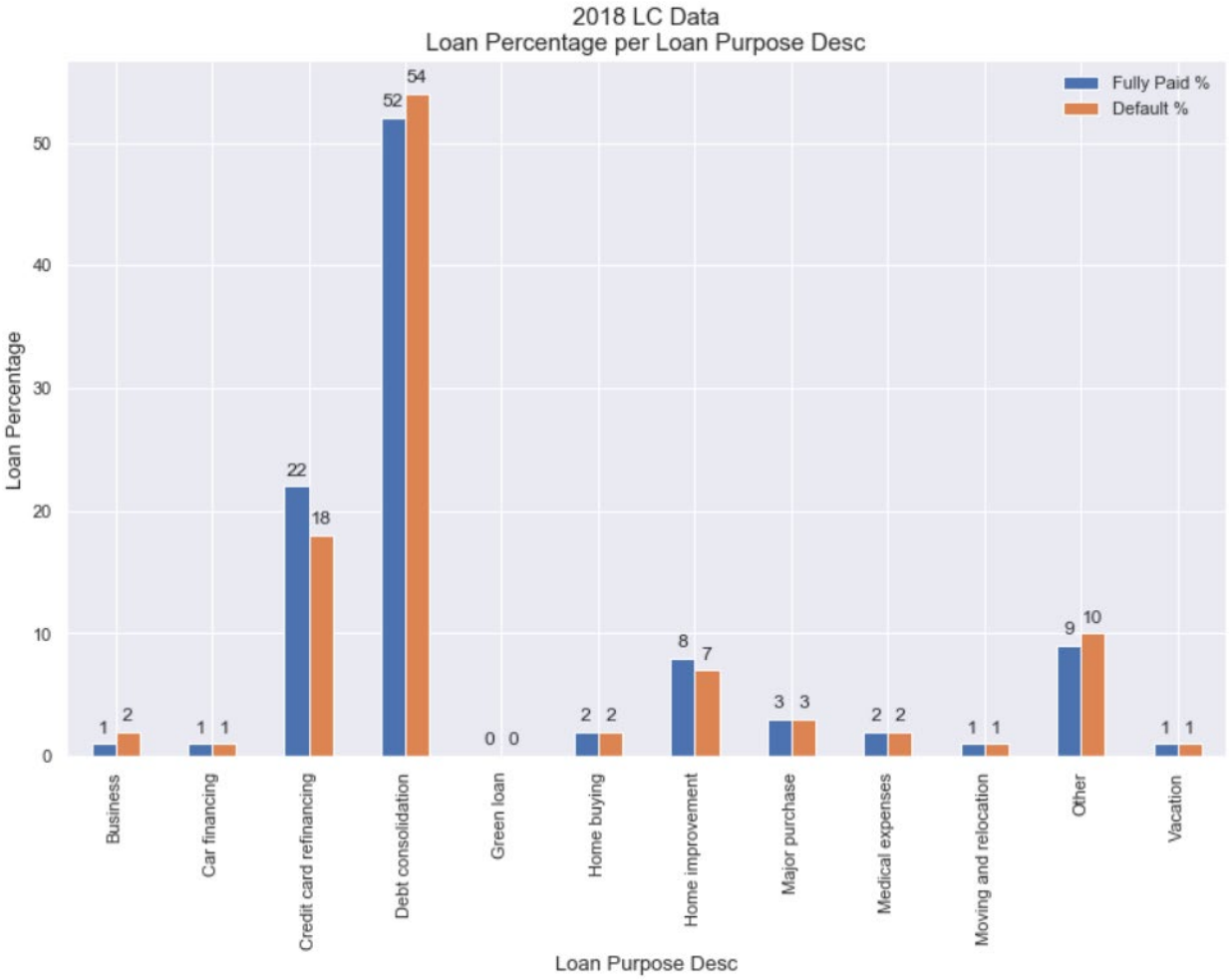

Most fully paid loans have borrowers who are on mortgage while most default loans have borrowers who are on house rental.



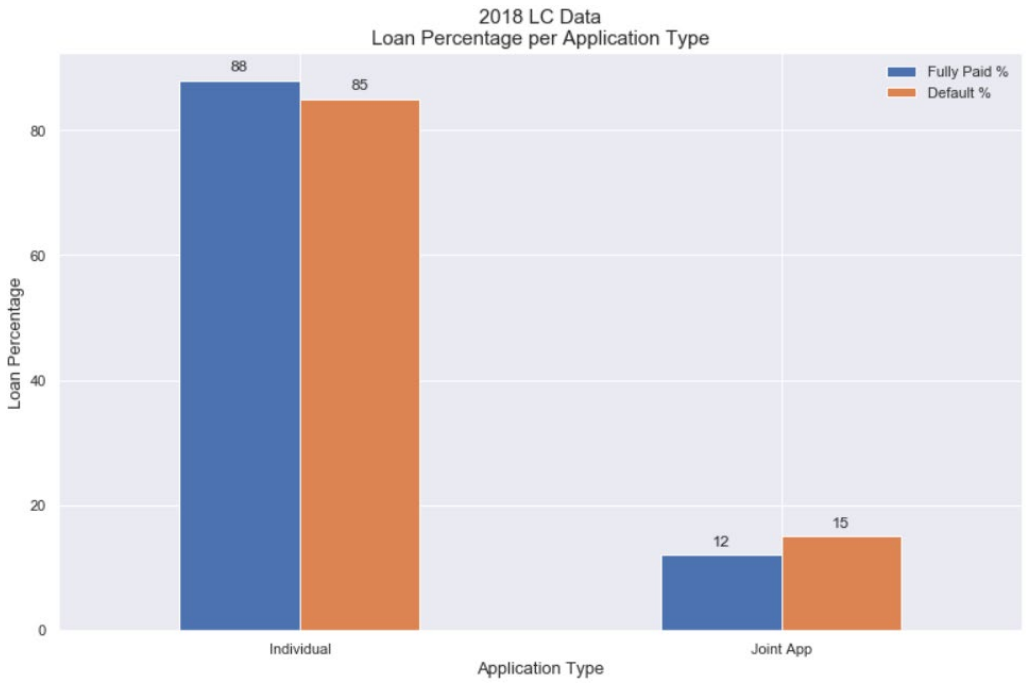
Many borrowers for both fully paid and default loans had their information verified by LC (source verified) before having their loan issued. The number of borrowers who did not have their information verified (not verified) is higher for fully paid loans than that of default loans. In contrast, the number of borrowers who had their information verified by other sources (verified) is higher for default loans than that of fully paid loans.




The distribution of loan purpose for both fully paid and default loans are similar. Many loans were issued to borrowers for debt consolidation and credit card refinancing.

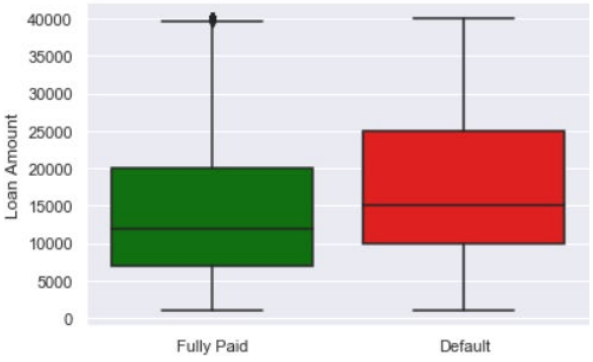


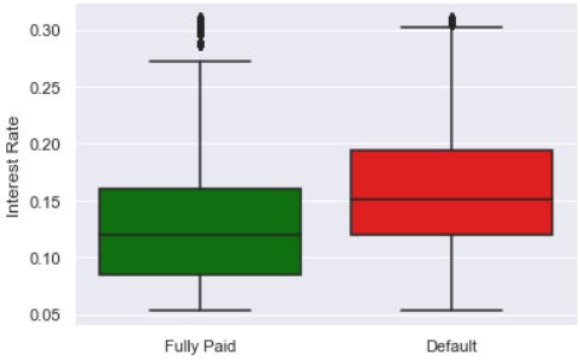

Likewise, the distribution of loan's application type are similar for both default and fully paid loans. There are more individual applications than joint applications.

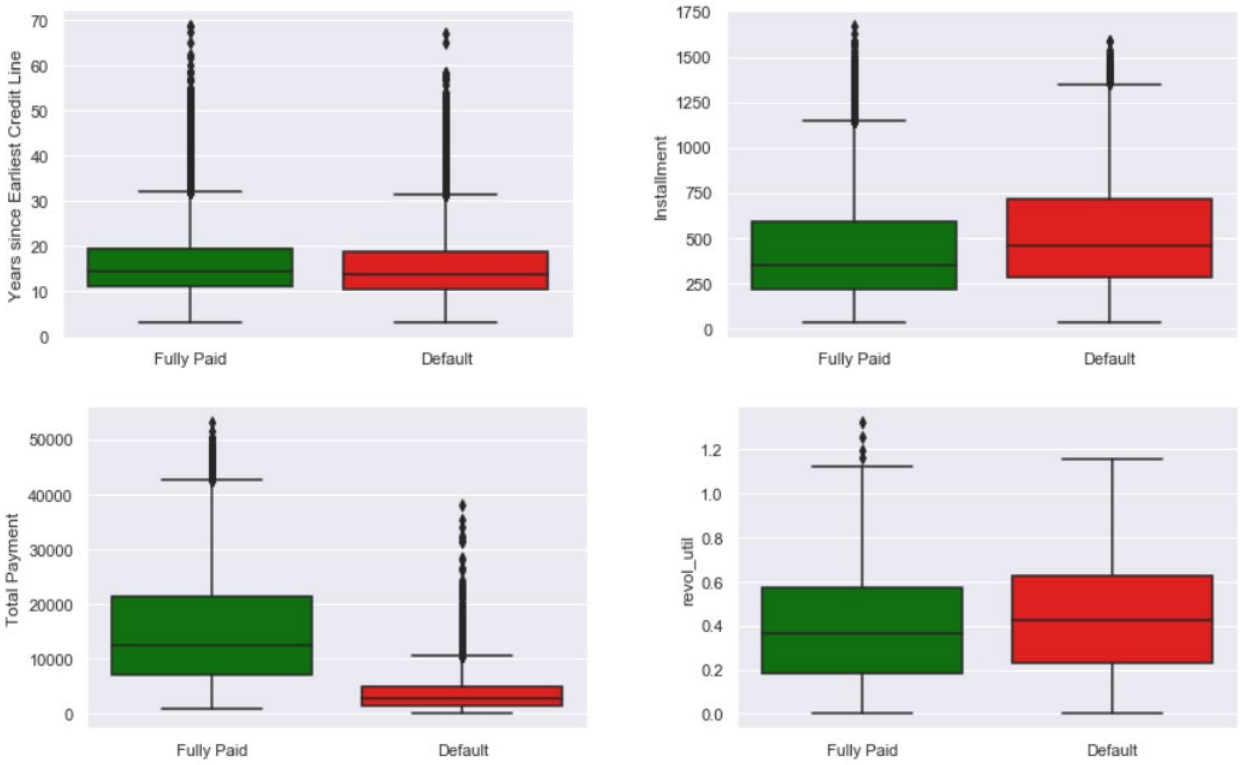



## Data Location

Most fully paid loans have lower loan amount, interest rate, and installment compared to those of default loans. The total payment for most fully paid loans are also higher than that of default loans. Years since earliest credit line and the amount of credit the borrower is using relative to all available revolving credit (revol_util) are almost the same for both fully paid and default loans.



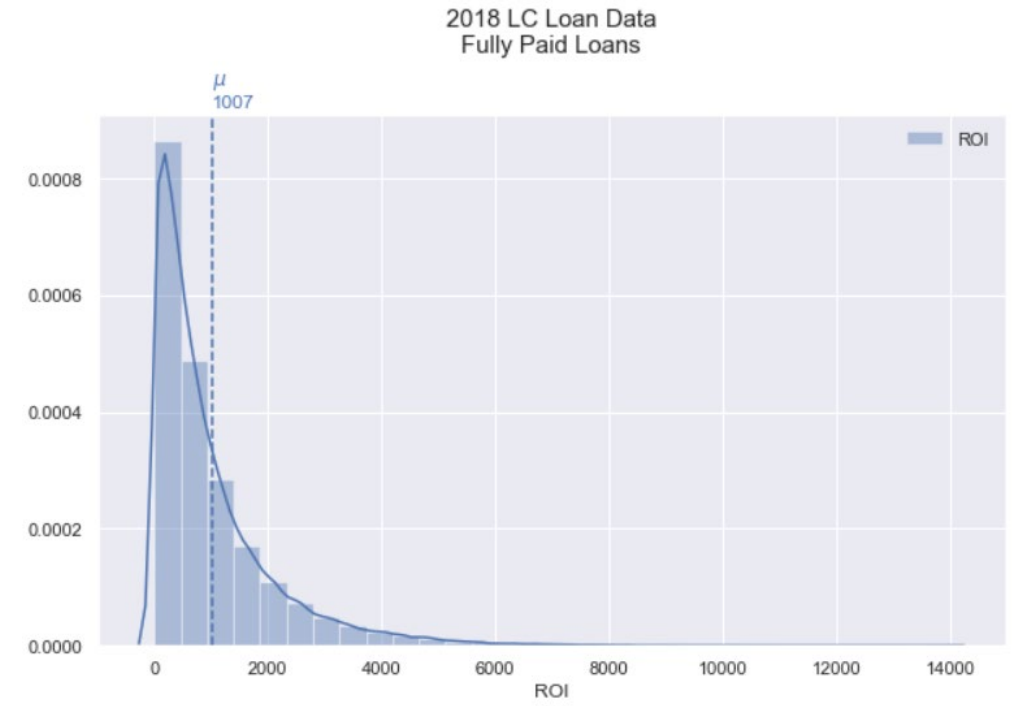



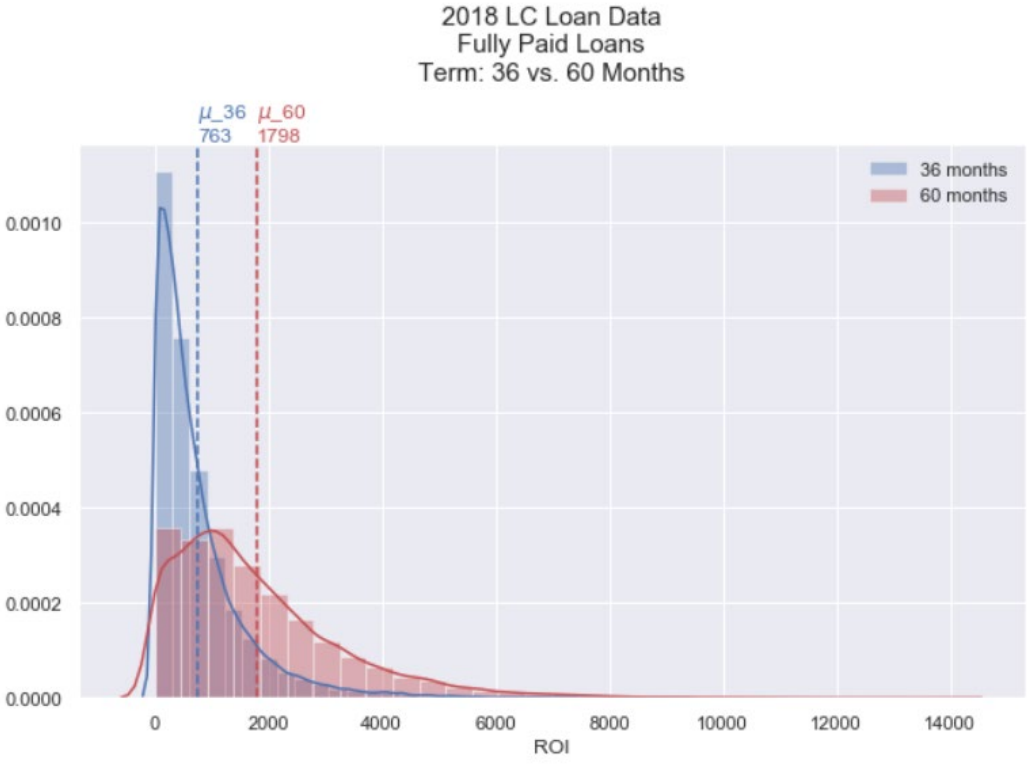



## Return on Investment (ROI)

For fully paid loans, we are 95% confident that the average ROI is:

- between 990 and 1015 dollars, regardless of loan term.
- between 756 and 770 dollars for 36 months loan.
- between 1776 and 1820 dollars for 60 months loan.



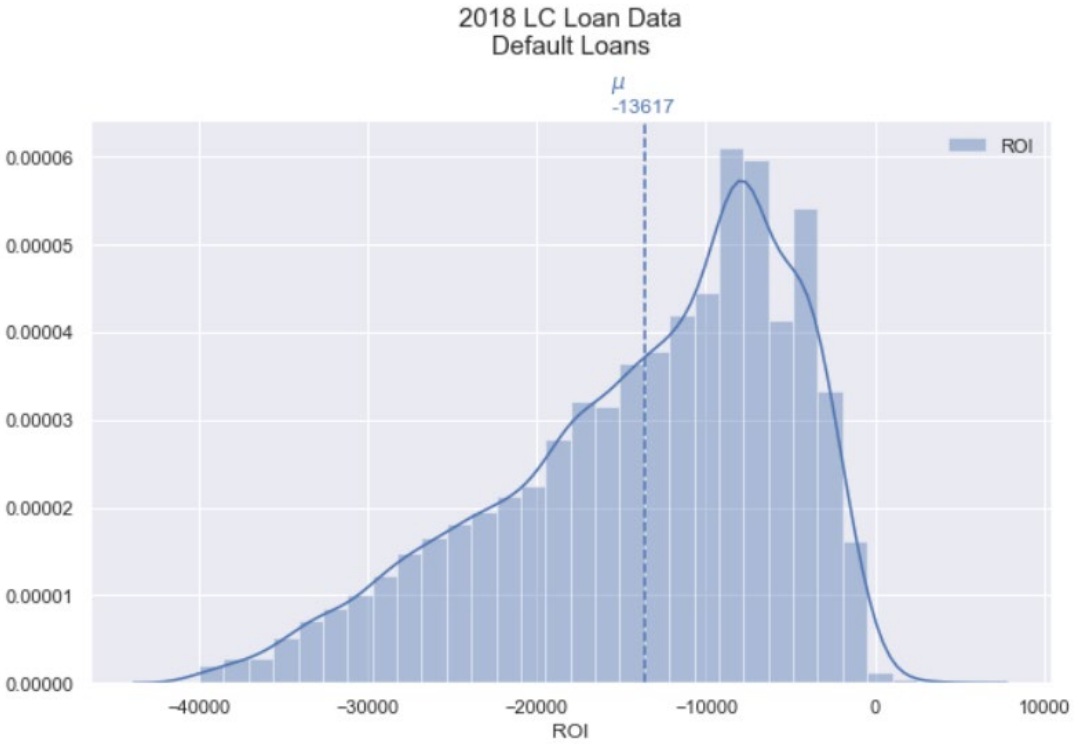



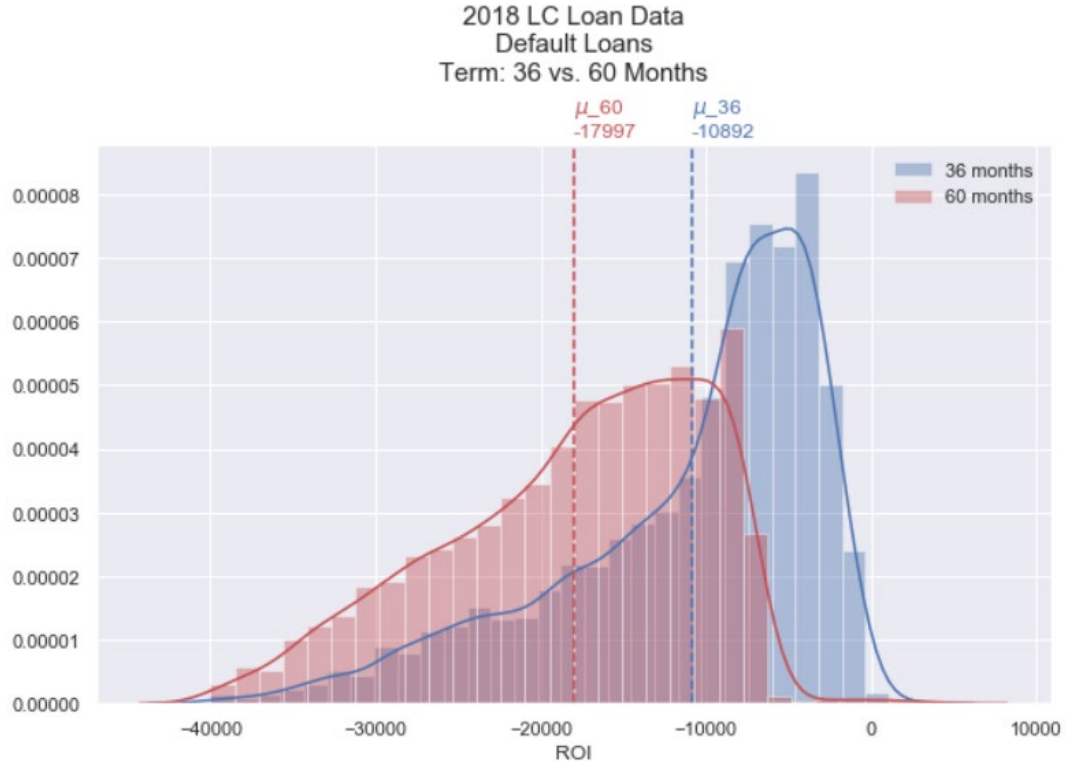

## Loss of Investment

For default loans, we are 95% confident that the average loss is:

- between 13,490 and 13,744 dollars, regardless of loan term.

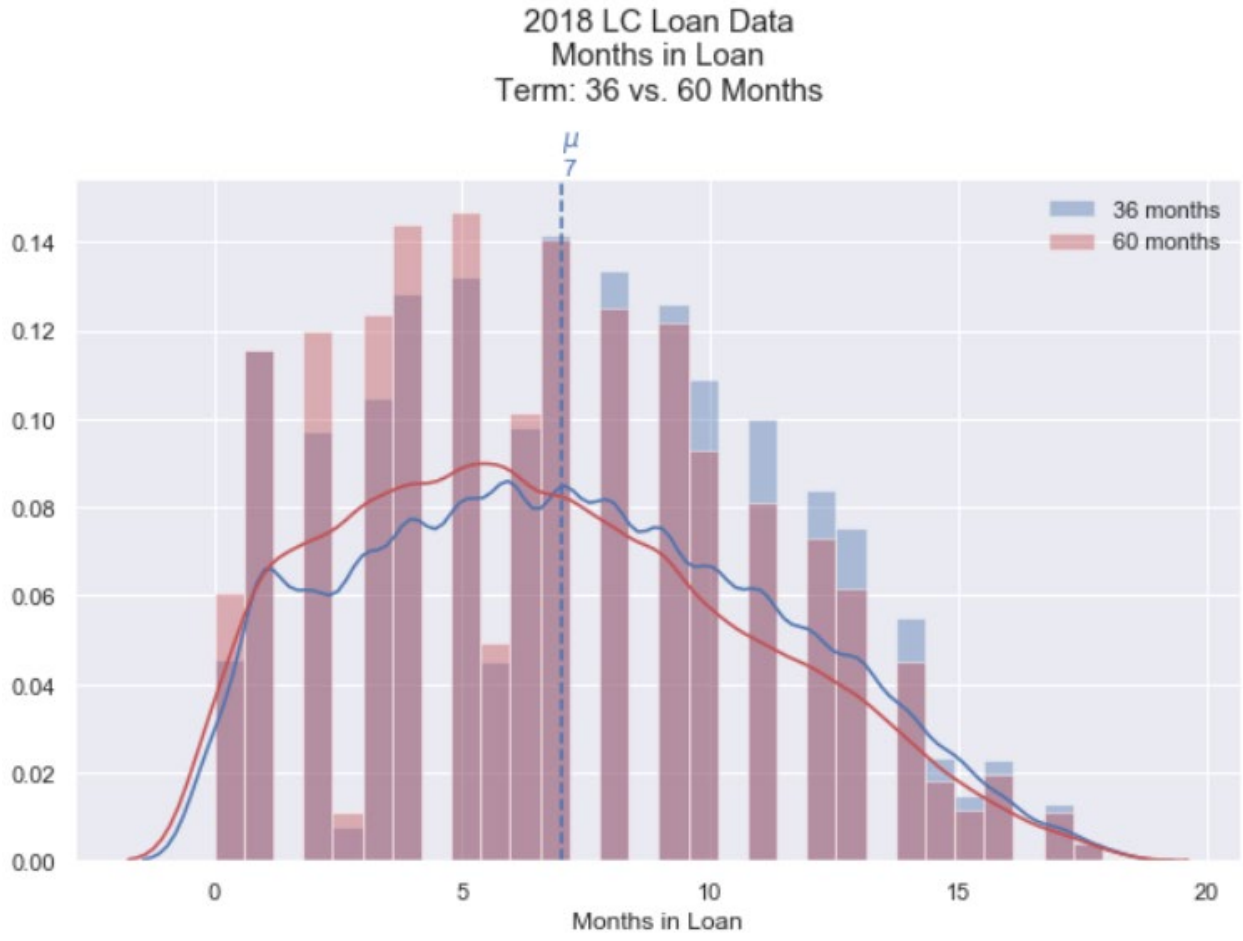
- between 10,744 and 11,040 dollars for 36 months loan.
- between 17,810 and 18,185 dollars for 60 months loan.



## Months in Loan

For both fully paid and default loans, most borrowers stay on the loan for an average of 7 months.




## Investment Portfolio: Estimating Loss of Investment

The table below shows the expected number of defaults and loss based on the number of invested loans. The data generated is based on the following assumptions:

- The lender invests $250 in each loan.
- The lender loses all money if the loan is default.
- There is no other hidden fees or costs.

Variables

- ***no_loans***: number of loans to be invested in

- ***expected_no_defaults***: the expected number of defaults is the mean of the binomial distribution with the probability of default of 17,589 / 93,853.
- ***tot_inv***: total amount invested = amount invested in each loan * no_loans
- ***expected_loss***: expected loss = amount invested in each loan * expected_no_defaults

```
Amount to be invested in each loan: 250
```

| | no_loans | expected_no_defaults | tot_inv | expected_loss |
|---|---|---|---|---|
| **0** | 50 | 9.0 | 12500.0 | 2250.0 |
| **1** | 100 | 19.0 | 25000.0 | 4750.0 |
| **2** | 150 | 28.0 | 37500.0 | 7000.0 |
| **3** | 200 | 37.0 | 50000.0 | 9250.0 |
| **4** | 250 | 47.0 | 62500.0 | 11750.0 |
| **5** | 300 | 56.0 | 75000.0 | 14000.0 |
| **6** | 350 | 66.0 | 87500.0 | 16500.0 |
| **7** | 400 | 75.0 | 100000.0 | 18750.0 |
| **8** | 450 | 84.0 | 112500.0 | 21000.0 |
| **9** | 500 | 94.0 | 125000.0 | 23500.0 |
| **10** | 550 | 103.0 | 137500.0 | 25750.0 |
| **11** | 600 | 112.0 | 150000.0 | 28000.0 |
| **12** | 650 | 122.0 | 162500.0 | 30500.0 |
| **13** | 700 | 131.0 | 175000.0 | 32750.0 |
| **14** | 750 | 141.0 | 187500.0 | 35250.0 |
| **15** | 800 | 150.0 | 200000.0 | 37500.0 |
| **16** | 850 | 159.0 | 212500.0 | 39750.0 |
| **17** | 900 | 169.0 | 225000.0 | 42250.0 |
| **18** | 950 | 178.0 | 237500.0 | 44500.0 |
| **19** | 1000 | 187.0 | 250000.0 | 46750.0 |