Data Wrangling of Lending Club Data
Intermediate Data Science – SpringBoard
David Ermer

Data Set From Kaggle:
https://www.kaggle.com/wendykan/lending-club-loan-data/downloads/lending-club-loan-data.zip)

The data set is provided in various file formats. Actual loan data is provided in both CSV and SQLite formats. A dictionary that provides detailed information about data in the various data columns is provided in CSV and XLSX format.

The first step was to compare the data as stored in each format. The dictionary file has 79 columns. The loan data in CSV format has 74 columns and the data in SQLite format has 75 columns. The differences in the loan data are 1) the SQLite data has an index. The difference between the dictionary file and the loan data file columns is that the loan data does not include four columns of FICO score data. There are some differences in the naming of the columns in the two loan data files that has been noted. Because working with CSV format is less complicated I will work with the data that is stored in this format.

The size of both loan data files is ~500MB. Two functions were written to pull columns out of the loan data files by column name to automate this process. Because of the size of the data files I plan to pull individual columns from the CSV file and build needed data frames from them.

My initial interest is in the difference between 'good' and 'bad' loans. There are 10 different categories for the status of loans and 35 sub-grade categories. The sub-grades are assigned by Lending Club.

Checks for missing data were preformed on all columns look at so far. None were found.

This data is very clean and very little wrangling has been necessary. Because of the large number of data columns on going wrangling (i.e. creation of data frames from individual columns) will be necessary.

Some initial calculations from the data were made.

Total Amount of Loans	$13,093,511,950.00
Average Loan Amount	$14,755.26
Total Number of Loans	887379
% of Loans In Default or Charged Off	~5.24%