We will be exploring Lending Club’s loan origination data from 2007-2015. Please download the dataset in CSV format (loan.csv) and associated dictionary (LCDataDictionary.xlsx) fromhttps://[www.kaggle.com/wendykan/lending-club-loan-data](http://www.kaggle.com/wendykan/lending-club-loan-data).

**Part 1:** Data Exploration and Evaluation Create an exploratory data analysis project. Load the data and perform any necessary cleaning and aggregations to explore and better understand the dataset. Based on your exploration, please describe your high level findings in a few sentences. Please include two data visualizations and two summary statistics to support these findings.

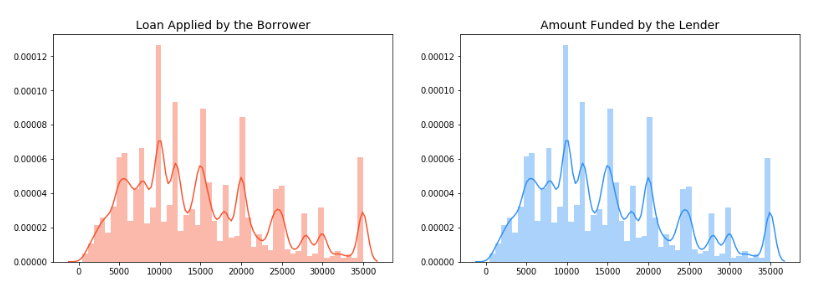
**Solution =>** Jupyter NoteBook is in files section of this project, file is named as "loan\_club.ipynb".

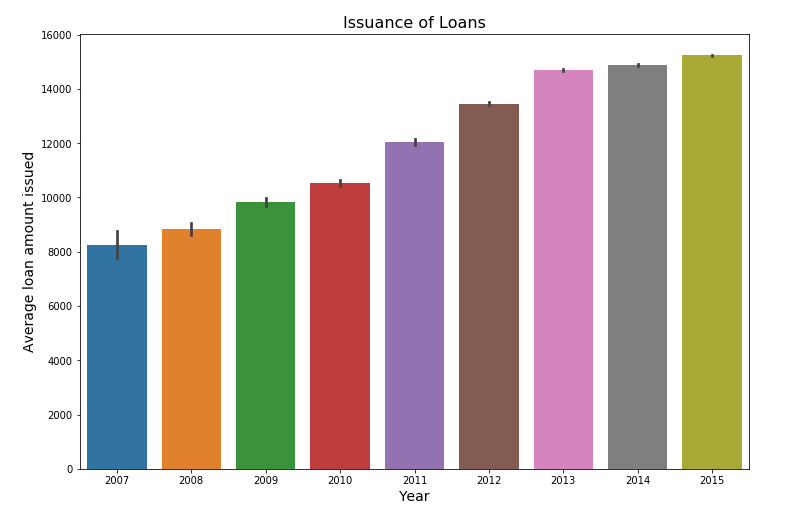
**Similar Distributions:**

We will start by exploring the distribution of the loan amounts and see when did the loan amount issued increased significantly.

**Summary:**

* Most of the **loans issued** were in the range of 10,000 to 20,000 USD.
* The **year of 2015** was the year were most loans were issued.
* Loans were issued in an **incremental manner**. (Possible due to a recovery in the U.S economy)
* The loans **applied** by potential borrowers, the amount **issued** to the borrowers are similarly distributed, **meaning** that it is most likely that qualified borrowers are going to get the loan they had applied for.





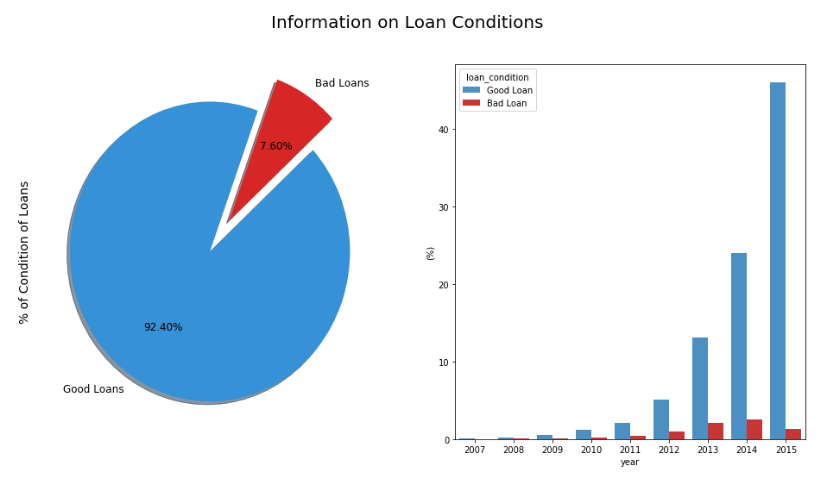
# Good Loans vs Bad Loans:

## Types of Loans:

In this section, we will see what is the amount of bad loans Lending Club has declared so far, of course we have to understand that there are still loans that are at a risk of defaulting in the future.

#### Summary:

* Currently, **bad loans** consist 7.60% of total loans but remember that we still have **current loans** which have the risk of becoming bad loans. (So this percentage is subjected to possible changes.)



**Note :-** To view this you may upload in Kaggle and check the visualizationa and code. You may aslo use Jupyter NoteBook in your local system but in that case you will need all the Libraries required in the code and Anaconda installed in the local system.

**Part 2:** Engineering Please build a prototype of a production data pipeline that will feed an analysis system (data warehouse) based on this dataset. This system will allow data scientists and data analysts to interactively query and explore the data, and will also be used for machine learning model building and testing. You may drop fields that you consider are not important for your analysis based on your evaluation in Part1. • Create a data model / schema in a database engine of your choice • Develop code that will persist the dataset into this storage system • Include any data validation routines that you think may be necessary Prioritize simplicity in your data model and processing code. Explain your thought process and document any alternate data models you considered along the way.

**Solution =>** A high Level design of the Engineering part:-

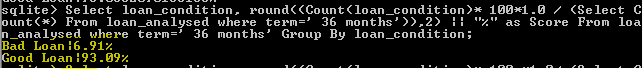
1. Here we can store the source data in SQLLite.
2. After that we can write a python code which can pick up the Table and crunch and Analyze it as in Part 1 using DataFrames. Here we add a new column named "Loan\_Type" which is either Good Loan or Bad Load. Which will be further used in Part 3
3. This Python code after crunching and Analysing the table will store data in new table in SQLite which could be "loan\_club\_analysed"
4. To make this Part an automated part we can use Windows task Scheduler or CRON in Unix Systems to schedule this job(Python Code).
5. The ETL code is written in python in file named "loan\_ETL.py".

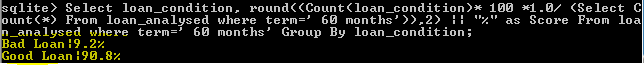
**Part 3:** Business Analysis After getting the data in the warehouse, your business analysts are interested in getting answers to the following, please write SQL queries and share the resultant data.

**Note :-** Since in Part 2 Actual system and Table was not built hence the resultant data is not present. Only SQLs are provided

1. Assuming the loans with status that are ‘Current’, ‘Issued’ and ‘Fully Paid’ as “Good Loans”, what is the percentage of good loans across each the 36- and 60-month terms

**Solution =>**

Select loan\_condition, round((Count(loan\_condition)\* 100\*1.0 / (Select Count(\*) From loan\_analysed where term=' 36 months')),2) || "%" as Score From loan\_analysed where term=' 36 months' Group By loan\_condition;

Select loan\_condition, round((Count(loan\_condition)\* 100 \*1.0/ (Select Count(\*) From loan\_analysed where term=' 60 months')),2) || "%" as Score From loan\_analysed where term=' 60 months' Group By loan\_condition;

1. What are the title(s) of employee(s) who took the most loans and least number of loans.

**Solution =>**

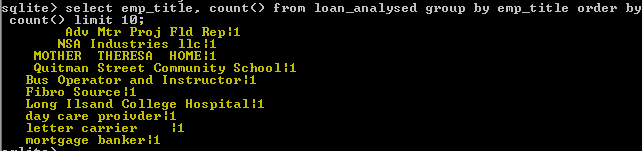
select emp\_title, count() from loan\_analysed where emp\_title !="" group by emp\_title order by count() desc limit 1;

So Teacher is the most common title of the employee who took the loan.



select emp\_title, count() from loan\_analysed group by emp\_title order by count() limit 10;

For this we have alot of answers as below :-

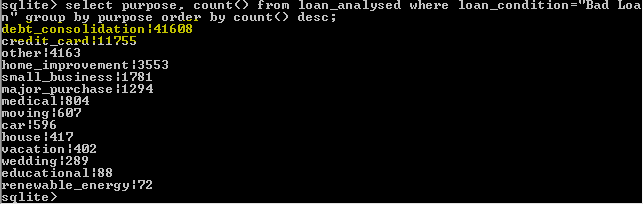


1. What is the most common purpose of the loans that are considered “Bad Loans” (please use definition mentioned for “Good Loans” in #1 above).

**Solution =>**

select purpose, count() from loan\_analysed where loan\_condition="Bad Loan" group by purpose order by count() desc limit 1;

So Debt\_Consolidation is the most common purpose of loans.



Artifacts Submit all code and documentation via GitHub. Please include all code files, outputs, and visualizations in the repository. Include a separate documentation file with your project detailing the design, approach and implementation.