Lending club is a leading online marketplace that connects lenders with borrowers for short-term personal loans up to $40,000. They have other offerings as well, but this case study will focus only on personal loans. The company has existed since 2007 and have provided millions of personal loans to individuals. Lending Club announced IPO (Initial Public Offering) in December 2014 since when their business model changed slightly with the company seeming to make risky lending. You will study various aspects of lending using R programing. You will submit your findings in a formal report format. The report must be at least 10 pages long with written description and explanation of your findings to the questions asked below.

**Details:** You are given six years of lending data (2012 – 2017) in csv format. The data files are relatively larger than what you have used during this course so far. The size of each file is different and depends upon the number of loans the company issued in a year. It can be noted that the file size are relatively larger 2015 onward, which is when the company went public and started lending more loans. Each file has 31 columns (variables) and the description of each column is provided in the DataDictionary.xls file.

In addition to that, you are also given the states characteristics in a file called states.csv. This file contains demographic information like population size, median income, unemployment rate etc.

Lastly, you are given a regions file called states\_regions.csv that contains larger regions and divisions that each state falls in. For example, New Hampshire is in the Northeast region and New England division.

**Merging and Cleaning**

Stack all six Lending Club files together on top of each other. Now join the states.csv file with the stacked file using state name as the primary key. Finally, merge the state\_regions file with the combined file so that you have one large file containing lending club and states geographic and demographic information.

**Analysis**

Use the above file to analyze and answer the following questions:

1. Find the distribution of number of loans by state, regions and divisions. Describe in your own words the geographic differences in the number of loans. Also, analyze your results by comparing number of loans per capita. Did you notice any missing states in the Lending Club data? If yes, then find out why.
2. Compare the average amount of loans granted by all states and divisions. Which states and divisions have the highest and lowest average loan amounts?
3. Compare the average interest rate charged and average loan amount by the loan Grade. Do you notice any patterns?
4. Run a frequency distribution of number of loans, average loan amount and average interest rate for each state by year (2012 through 2017). Describe the changing patterns in those numbers.
5. Is there a relationship with the population size of a state and the average loan amount given? Is there a relationship between Grade of loans and median income level in a state?

**Visualization**

1. Create a plot of interest rates and Grade or a loan and describe the pattern.
2. Create a map of US states and color code the map with the average amount of loans given.
3. Show visually the relationship between the annual income of the recipient and the loan amount obtained from Lending Club
4. Create a plot that shows the relationship between the length of employment and amount of loan obtained.
5. Create a “regional” map and show an interesting relationship of your liking.