Carballo, Catolico, Meija, Principe
Big Data Proposal - CS 129.1

Big Data Question:

Given the dataset of loaners' information, we will prove/disprove using Analysis of Variance if the means of the funded_income of the different types of loan_status are significantly different or not.

Data Source Description:

The dataset is retrieved from the lending and investing website called Lending Club. The dataset that we got came from the year 2015 (only chosen arbitrarily) and removed impertinent columns using Microsoft Excel that had 60-70% null values in it so that the size of the data is lessened and it will be faster to apply MapReduce to the dataset because only the significant columns are included.

Link to data source:

https://www.lendingclub.com/info/download-data.action

Updated link for the cleaned dataset:

https://drive.google.com/open?id=1MFWAO0L6E0Oybh1gMPudQquhdj_Jmyui