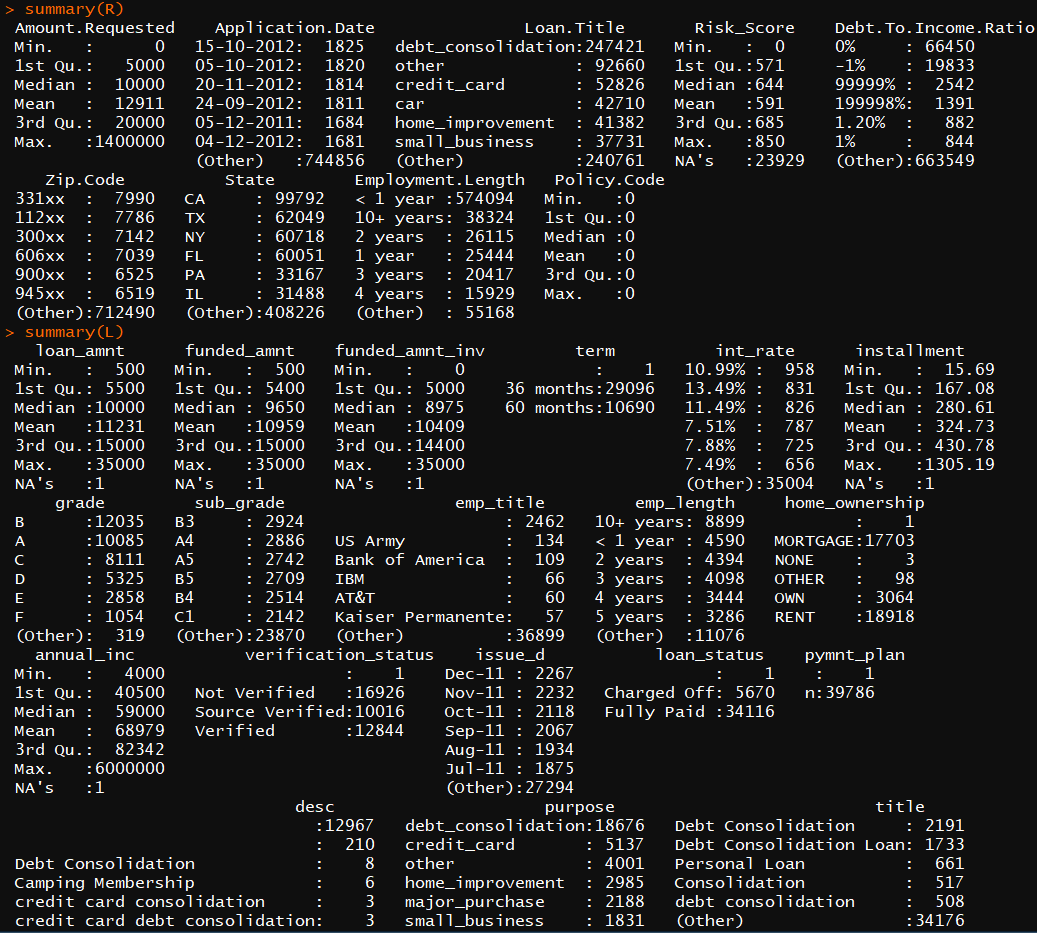
**Solution design and data collection**

**Data Collection:** As the project is based on the possibility of investing in a lending club, we need data related to past ventures of the lending club. This data is available on their website in form of a set that contains two csv files. ‘LoanData.csv’ contains all the details about the loan that was approved and the details of that loan over the period. ‘DeclinedLoanData.csv’ contains the records where the loan was declined because the conditions of the lending club were not met. We will attempt to first merge these two files and then use that single file for our case study.

Here is the summary of the datasets,



**Solution Design:** The aim consists of three parts, first is classification of a person to decide eligibility for loan, second is clustering that person according some decided characteristics, third is depending on the classification deciding what interest rate to offer to that person.

For the first part, we build a logistic regression model and Random forest model for classification. We then test out these two models based on the training and test data set to check for accuracy and then select the most accurate one.

For the second part, we segment data into clusters using k-means clustering algorithms or segment data into clusters manually using categorical or numerical features.

For the third part, we need to decide the best interest rate for a candidate. We try building various prediction models for each cluster. Using Linear regression, Neural Network models and KNN algorithms. Check for accuracy of the models based on MAE, RMS, MAPE for training and testing datasets. Then, we select the most accurate model for prediction.