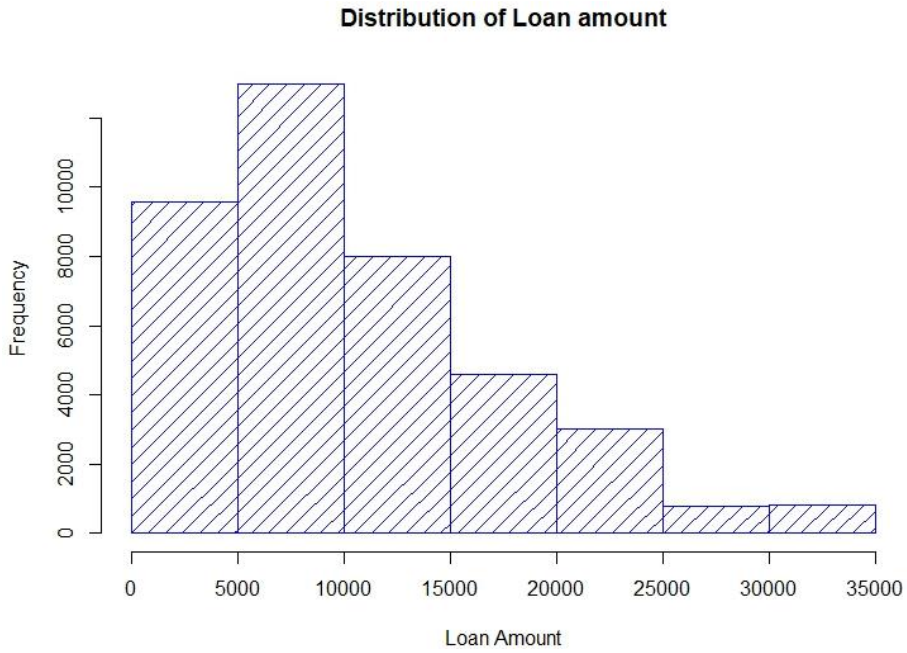
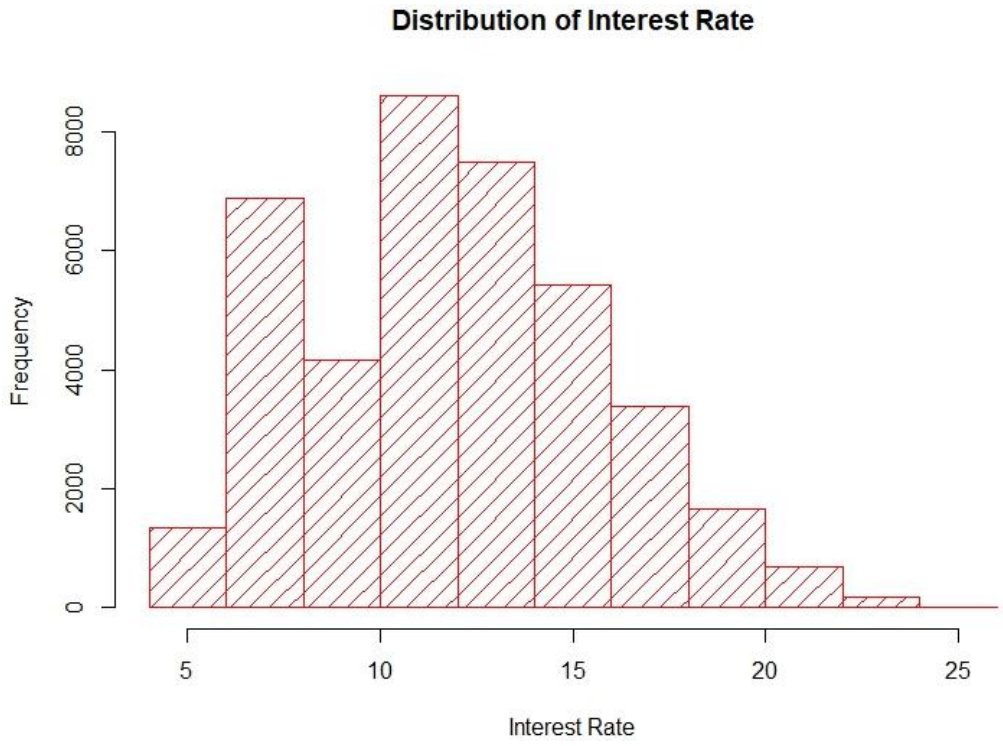
Data Visualization, Data Mining Techniques, and Implementation

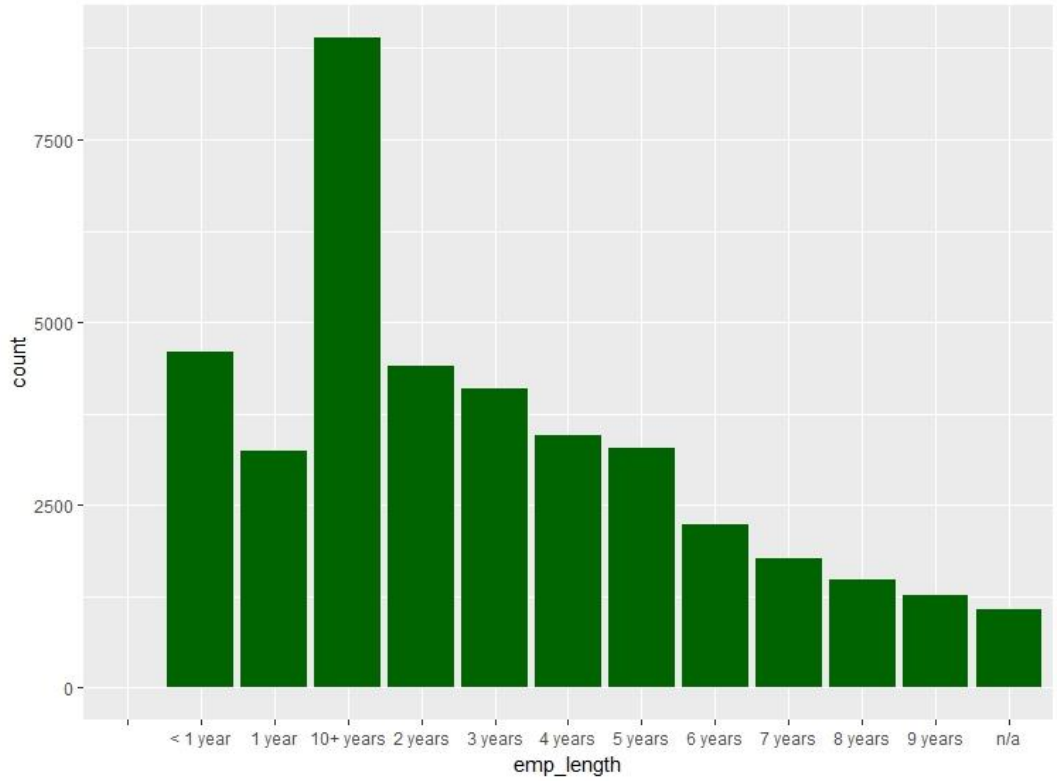
The data set includes two types of data, the loan details of the applications that were accepted for lending club and the details of the applications that were rejected. By performing Exploratory data analysis on these datasets, we have learned the following, the distribution of the loan amount is mostly based in the $5000-$10000 region.



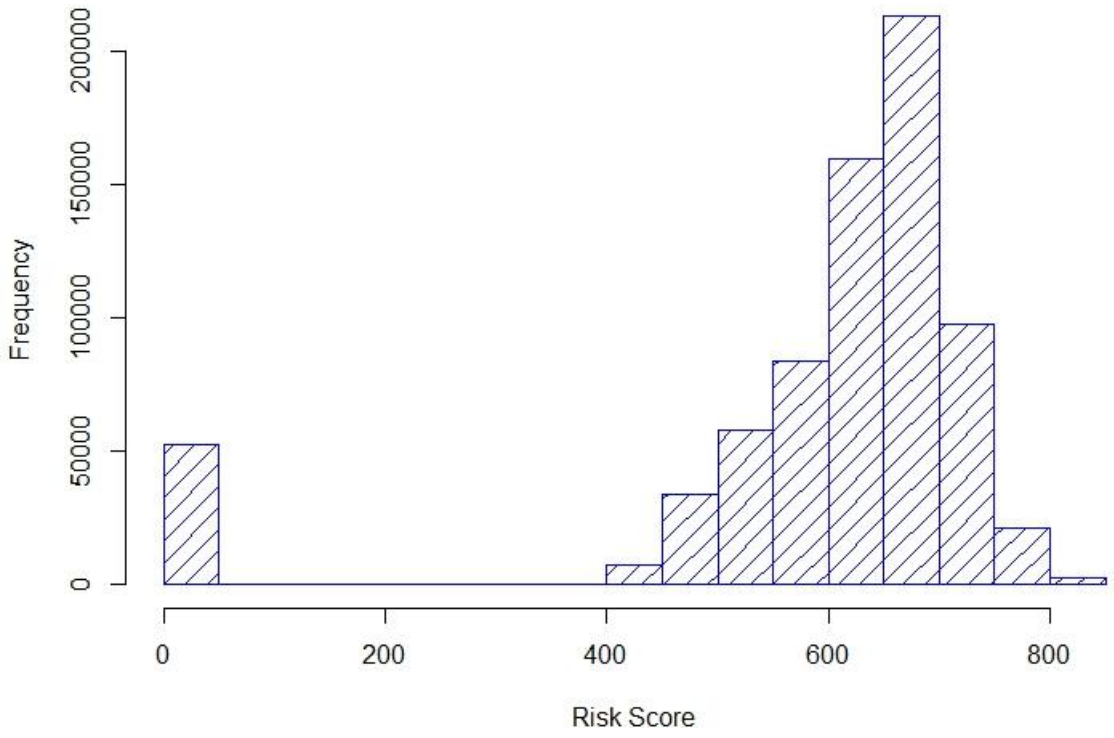
The interest rate that was provided after the loan was given lies mostly between 10%-15%



The employment length of the approved candidates is ranked as follows,



And the Risk score of the rejected candidates was mostly high which can be confirmed using visualization and thus these candidates were rejected.



This dataset will need a lot of cleaning as the numerical values are stored with a string for units as factors. These variables must be processed to help in the upcoming models that we are going to build for classification and prediction.

Firstly, we need to decide whether to give a loan or not, and for that we need the employment details and the amount requested along with a few other elements. Then our next task is to cluster them based on some attributes and then according to the clusters we need to predict the interest that should be given to these clients.

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.