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,

```
> summary(R)
 Amount.Requested    Application.Date              Loan.Title       Risk_Score    Debt.To.Income.Ratio
 Min.   :      0   15-10-2012:  1825   debt_consolidation:247421   Min.   :  0     0%     : 66450
 1st Qu.:   5000   05-10-2012:  1820   other             : 92660   1st Qu.:571     -1%    : 19833
 Median :  10000   20-11-2012:  1814   credit_card       : 52826   Median :644     99999% :  2542
 Mean   :  12911   24-09-2012:  1811   car               : 42710   Mean   :591     199998%:  1391
 3rd Qu.:  20000   05-12-2011:  1684   home_improvement  : 41382   3rd Qu.:685     1.20%  :   882
 Max.   :1400000   04-12-2012:  1681   small_business    : 37731   Max.   :850     1%     :   844
                   (Other)   :744856   (Other)           :240761   NA's   :23929   (Other):663549
    Zip.Code          State        Employment.Length   Policy.Code
 331xx  :  7990   CA     : 99792   < 1 year :574094   Min.   :0
 112xx  :  7786   TX     : 62049   10+ years: 38324   1st Qu.:0
 300xx  :  7142   NY     : 60718   2 years  : 26115   Median :0
 606xx  :  7039   FL     : 60051   1 year   : 25444   Mean   :0
 900xx  :  6525   PA     : 33167   3 years  : 20417   3rd Qu.:0
 945xx  :  6519   IL     : 31488   4 years  : 15929   Max.   :0
 (Other):712490   (Other):408226   (Other)  : 55168
> summary(L)
   loan_amnt      funded_amnt    funded_amnt_inv         term          int_rate      installment
 Min.   :  500   Min.   :  500   Min.   :    0            :    1   10.99% :  958   Min.   :  15.69
 1st Qu.: 5500   1st Qu.: 5400   1st Qu.: 5000    36 months:29096   13.49% :  831   1st Qu.: 167.08
 Median :10000   Median : 9650   Median : 8975    60 months:10690   11.49% :  826   Median : 280.61
 Mean   :11231   Mean   :10959   Mean   :10409                      7.51%  :  787   Mean   : 324.73
 3rd Qu.:15000   3rd Qu.:15000   3rd Qu.:14400                      7.88%  :  725   3rd Qu.: 430.78
 Max.   :35000   Max.   :35000   Max.   :35000                      7.49%  :  656   Max.   :1305.19
 NA's   :1       NA's   :1       NA's   :1                          (Other):35004   NA's   :1
    grade         sub_grade                emp_title         emp_length     home_ownership
 B      :12035   B3     : 2924                   : 2462   10+ years: 8899           :    1
 A      :10085   A4     : 2886   US Army         :  134   < 1 year : 4590   MORTGAGE:17703
 C      : 8111   A5     : 2742   Bank of America :  109   2 years  : 4394   NONE     :    3
 D      : 5325   B5     : 2709   IBM             :   66   3 years  : 4098   OTHER    :   98
 E      : 2858   B4     : 2514   AT&T            :   60   4 years  : 3444   OWN      : 3064
 F      : 1054   C1     : 2142   Kaiser Permanente:   57   5 years  : 3286   RENT     :18918
 (Other):  319   (Other):23870   (Other)         :36899   (Other)  :11076
  annual_inc           verification_status    issue_d          loan_status    pymnt_plan
 Min.   :   4000                    :    1   Dec-11 : 2267              :    1    :    1
 1st Qu.:  40500   Not Verified     :16926   Nov-11 : 2232   Charged Off: 5670   n:39786
 Median :  59000   Source Verified:10016    Oct-11 : 2118   Fully Paid :34116
 Mean   :  68979   Verified        :12844   Sep-11 : 2067
 3rd Qu.:  82342                            Aug-11 : 1934
 Max.   :6000000                            Jul-11 : 1875
 NA's   :1                                  (Other):27294
                        desc                     purpose                          title
                           :12967   debt_consolidation:18676   Debt Consolidation     : 2191
                           :  210   credit_card       : 5137   Debt Consolidation Loan: 1733
 Debt Consolidation        :    8   other             : 4001   Personal Loan          :  661
 Camping Membership        :    6   home_improvement  : 2985   Consolidation          :  517
 credit card consolidation :    3   major_purchase    : 2188   debt consolidation     :  508
 credit card debt consolidation:    3   small_business    : 1831   (Other)                :34176
```

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.