Section 3. data

Data Collection

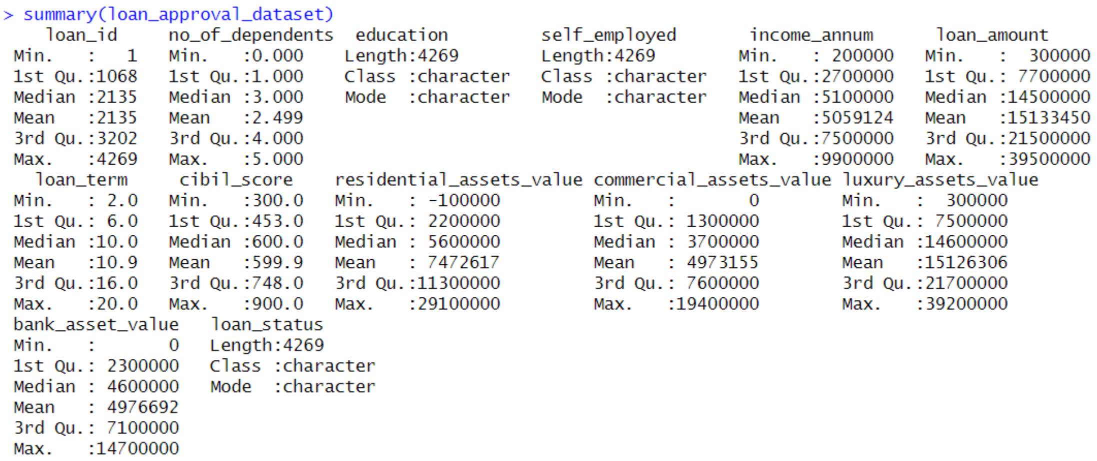
Data source: https://www.kaggle.com/datasets/architsharma01/loan-approval-prediction-dataset

Data Description

We collected these data from a website call Kaggle, where you can find many shared data by others. We also discussed that should we use real-world data from our group member's company but most of our members' companies' data are strictly confidential. Therefore we chose to find shared data online.

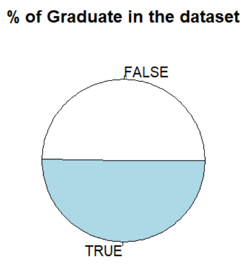
The data set we selected is called Loan Approval Prediction Dataset. It is a set of financial records and related information used to determine the possibility of individual for obtaining loans from a financial institution like bank. It is well structured so we saved a lot of time for data cleaning. It includes various numerical variables like annual income, loan amount, and credit scores. It also includes categorical variables like employment status and education background of the applicant.

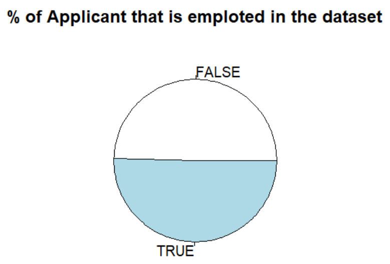
Data summary:

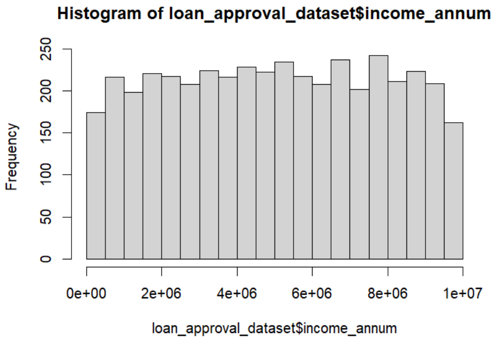


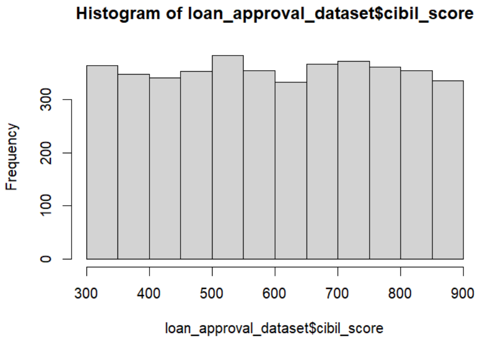
```
> summary(loan_approval_dataset)
    loan_id     no_of_dependents  education          self_employed       income_annum       loan_amount     
 Min.   :   1   Min.   :0.000    Length:4269        Length:4269        Min.   : 200000   Min.   :  300000  
 1st Qu.:1068   1st Qu.:1.000    Class :character   Class :character   1st Qu.:2700000   1st Qu.: 7700000  
 Median :2135   Median :3.000    Mode  :character   Mode  :character   Median :5100000   Median :14500000  
 Mean   :2135   Mean   :2.499                                          Mean   :5059124   Mean   :15133450  
 3rd Qu.:3202   3rd Qu.:4.000                                          3rd Qu.:7500000   3rd Qu.:21500000  
 Max.   :4269   Max.   :5.000                                          Max.   :9900000   Max.   :39500000  
   loan_term     cibil_score    residential_assets_value commercial_assets_value luxury_assets_value
 Min.   : 2.0   Min.   :300.0   Min.   : -100000         Min.   :       0        Min.   :  300000   
 1st Qu.: 6.0   1st Qu.:453.0   1st Qu.: 2200000         1st Qu.: 1300000        1st Qu.: 7500000   
 Median :10.0   Median :600.0   Median : 5600000         Median : 3700000        Median :14600000   
 Mean   :10.9   Mean   :599.9   Mean   : 7472617         Mean   : 4973155        Mean   :15126306   
 3rd Qu.:16.0   3rd Qu.:748.0   3rd Qu.:11300000         3rd Qu.: 7600000        3rd Qu.:21700000   
 Max.   :20.0   Max.   :900.0   Max.   :29100000         Max.   :19400000        Max.   :39200000   
 bank_asset_value   loan_status       
 Min.   :       0   Length:4269       
 1st Qu.: 2300000   Class :character  
 Median : 4600000   Mode  :character  
 Mean   : 4976692                     
 3rd Qu.: 7100000                     
 Max.   :14700000                     
```

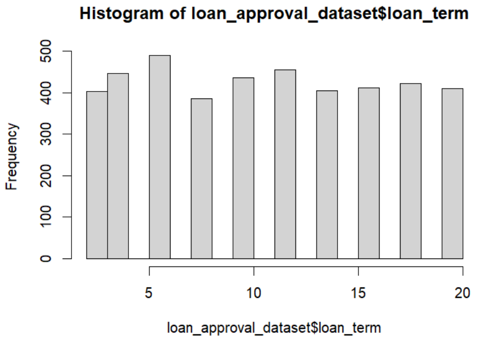
Data visualization

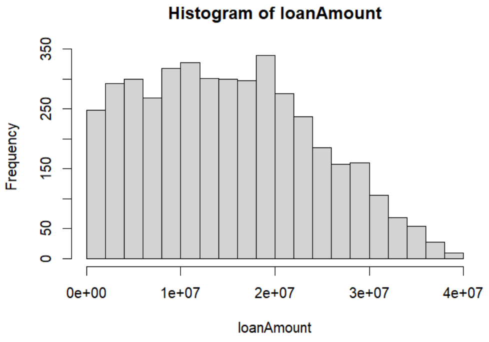





Histogram of loan_approval_dataset$income_annum
Frequency
0
50
100
150
200
250
0e+00
2e+06
4e+06
6e+06
8e+06
1e+07
loan_approval_dataset$income_annum


Histogram of loan_approval_dataset$cibil_score
Frequency
0
100
200
300
300
400
500
600
700
800
900
loan_approval_dataset$cibil_score


Histogram of loan_approval_dataset$loan_term
Frequency
0
100
200
300
400
500
5
10
15
20
loan_approval_dataset$loan_term


Histogram of loanAmount
Frequency
0
50
150
250
350
0e+00
1e+07
2e+07
3e+07
4e+07
loanAmount