EDA Assignment

- By Jithan A N
- 03/Jan/2022

Problem Statement

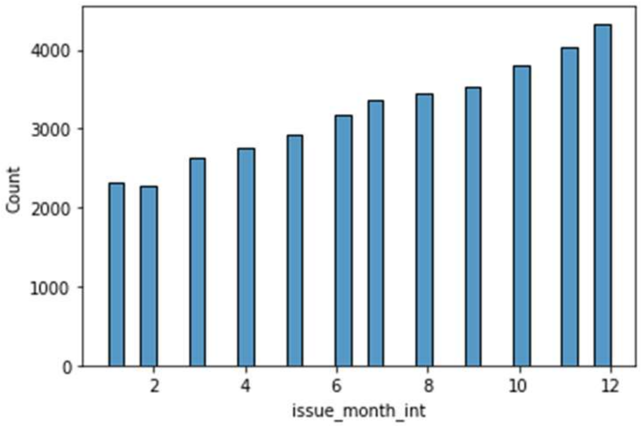
- Analysis of loan dataset history with regards to accepted loans
- The data contains loans fully paid and loans defaulted or charged off
- There are various other member information presented for data analysis

Approach

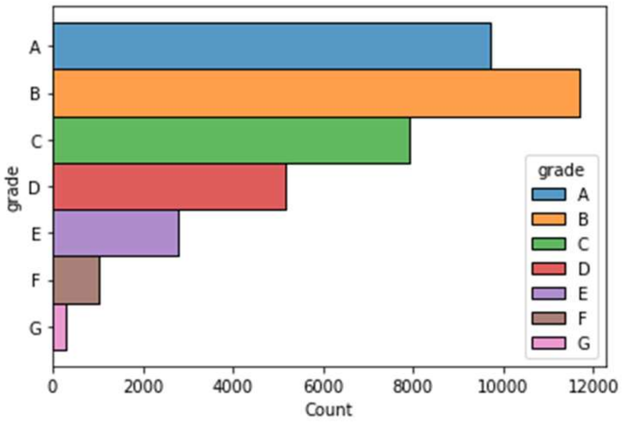
- Load the data
- Clean the data
- Impute or remove the missing values
- Univariate Analysis
- Segmented Univariate Analysis
- Bivariate Analysis

Univariate Analysis

- Plot of Count versus issue month in integer shows most of the loans being taken closer to the end of the year

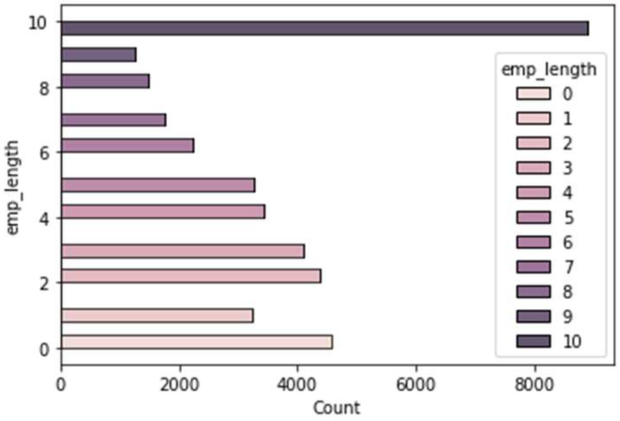


- Plot of grade versus count shows most of the loans being taken by grade B members followed by grade A members

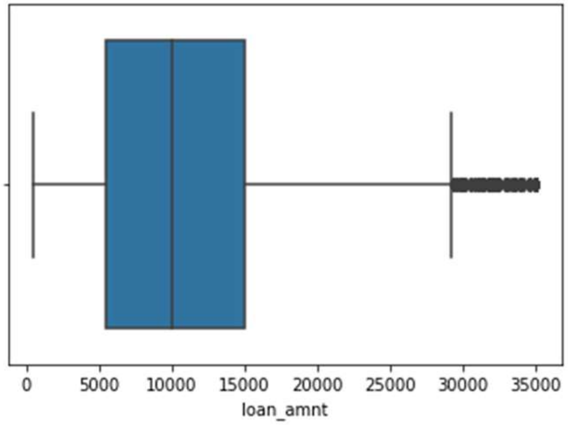


Univariate Analysis

- Plot of Count versus employee length shows most of the loans being taken by members with experience of 10+ years followed by new employees with below 1 year experience



- Box plot of loan amount shows that loans between 5000 to 15000 are very common and above 30000 are outliers



Segmented Univariate Analysis

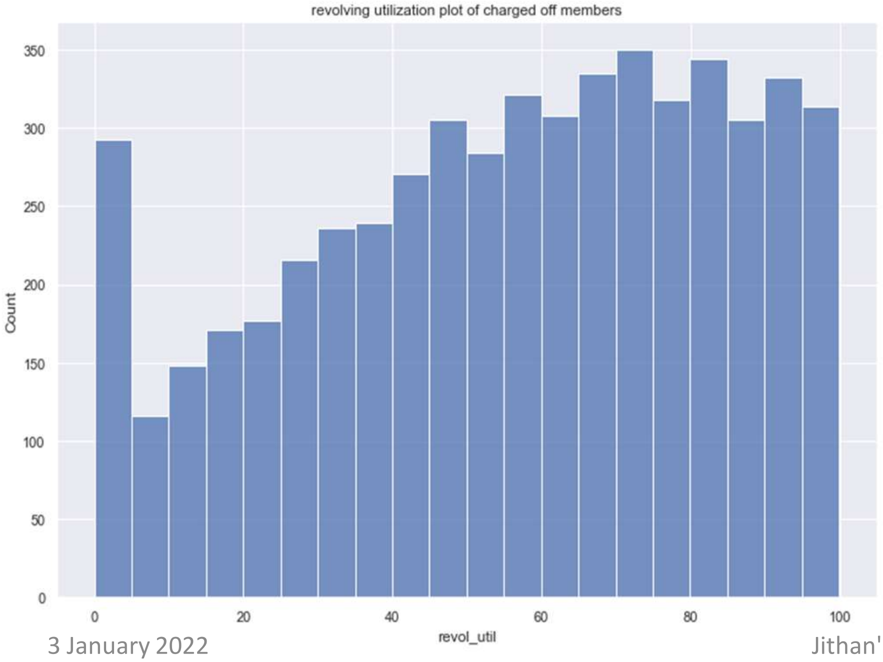
- We need to target that we need to
  - Give loans to members who can pay back
  - Don’t give loans to members which will become bad loans or charged off

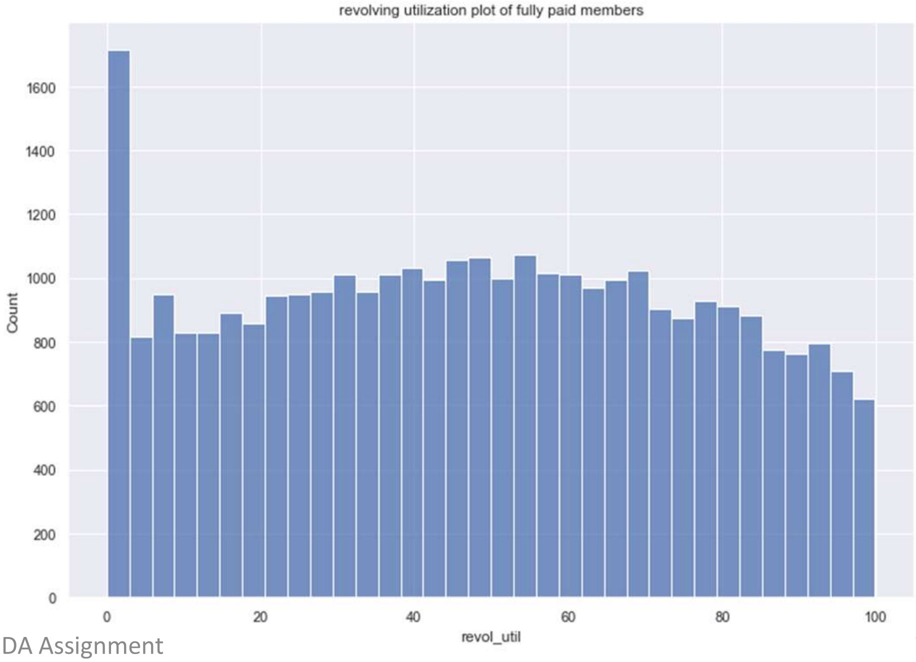
- For this, we will target our study on loan status

- We will segment data or create 2 dataframes using charged_off and fully_paid as the distinction criteria between them and study the different variables post this segmentation

Segmented Univariate Analysis

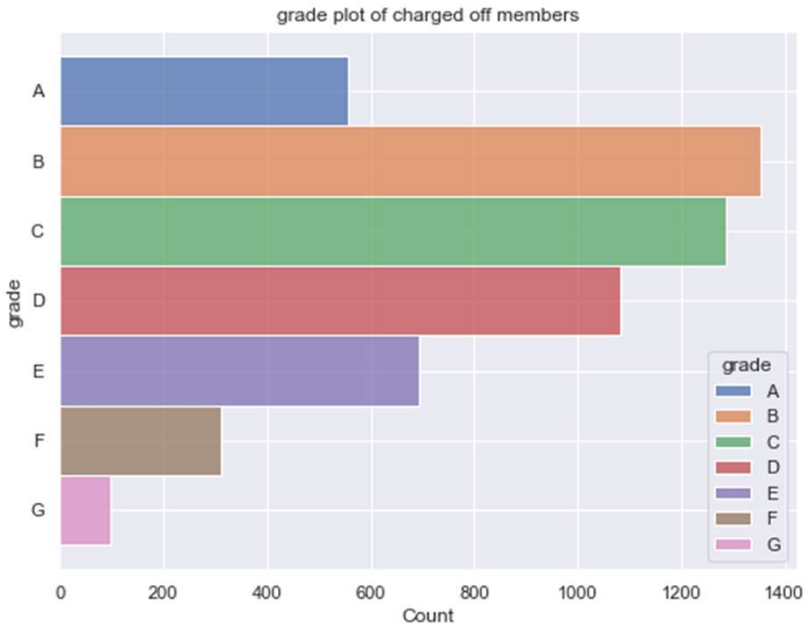
- Plotting revolving utilization after segmentation on loan status shows that higher revolving credit utilization is a big reason for charged off loans

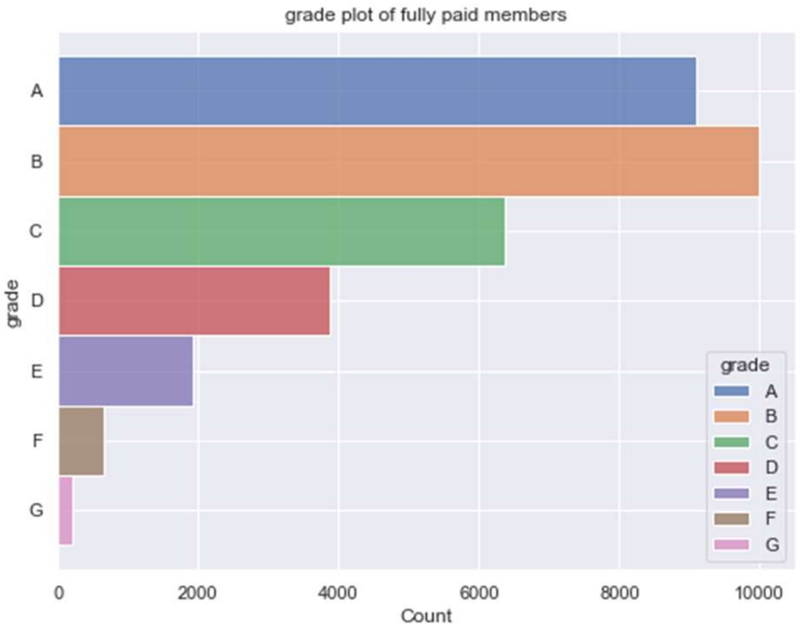




Segmented Univariate Analysis

- Plotting grade after segmentation on loan status shows that
  - Grade A members are better paying members
  - Grade C and D show comparatively risky behavior

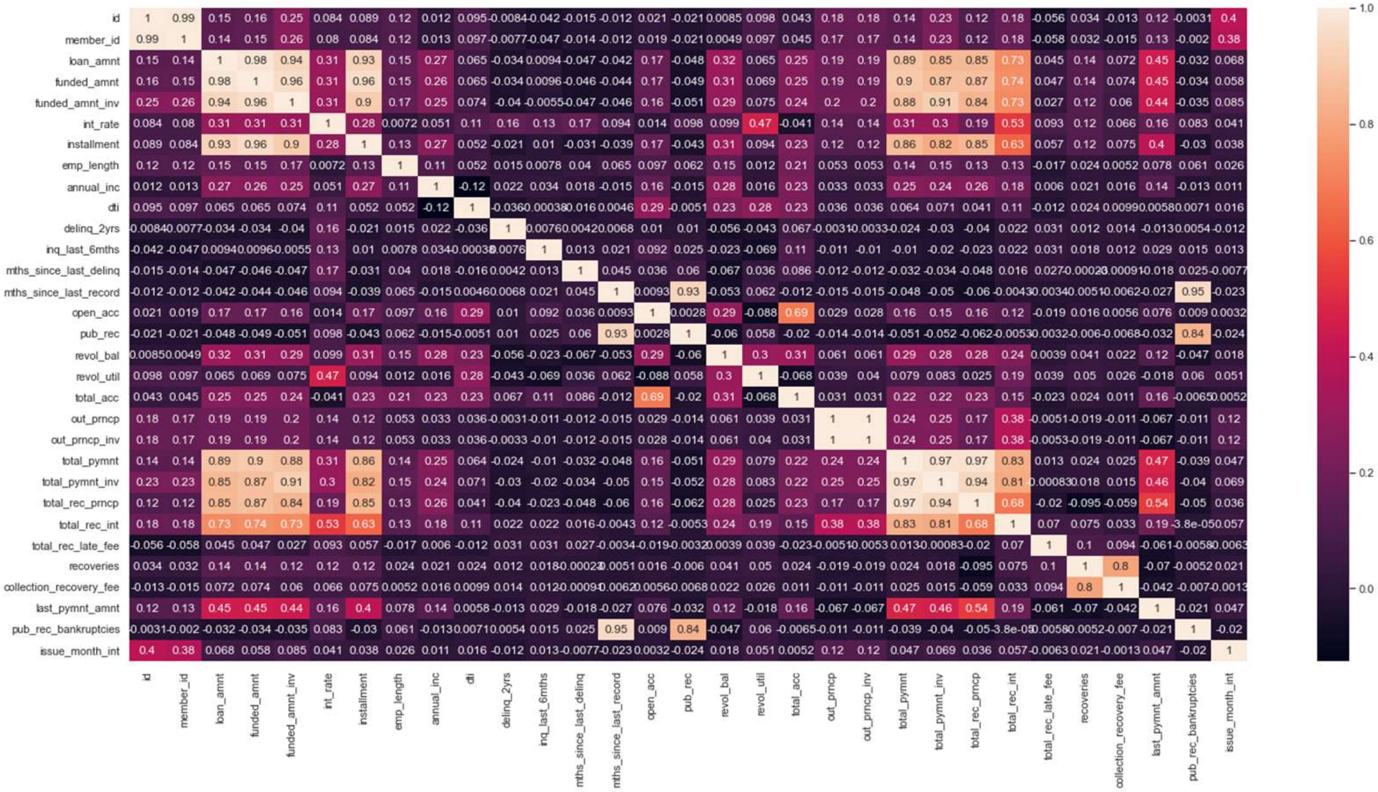

grade plot of charged off members


grade plot of fully paid members

Bivariate Analysis

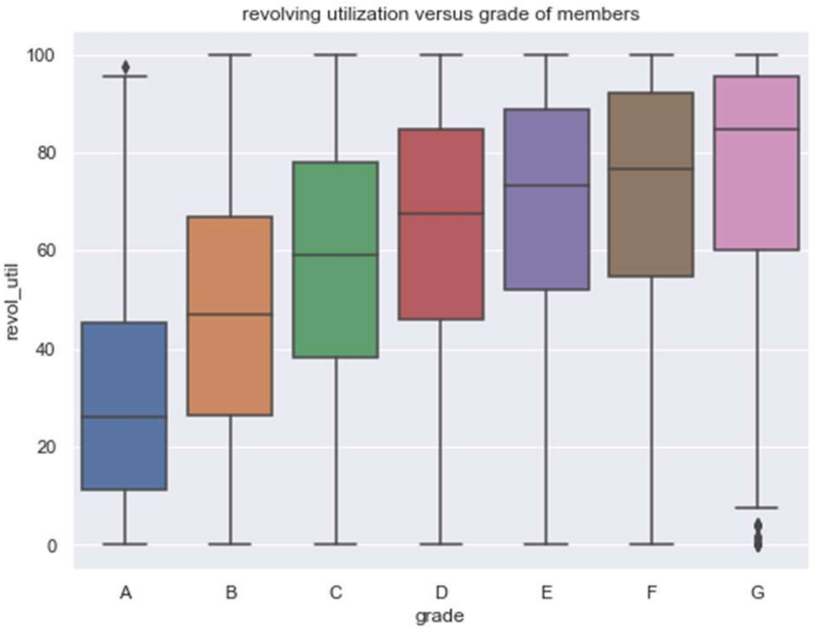
- For Bivariate analysis, we can run a correlation analysis plot using seaborn as depicted on right.

- There is not much that can be inferenced other than known facts like loan amount, funded amount, installment, total payment etc are correlated which is on expected lines



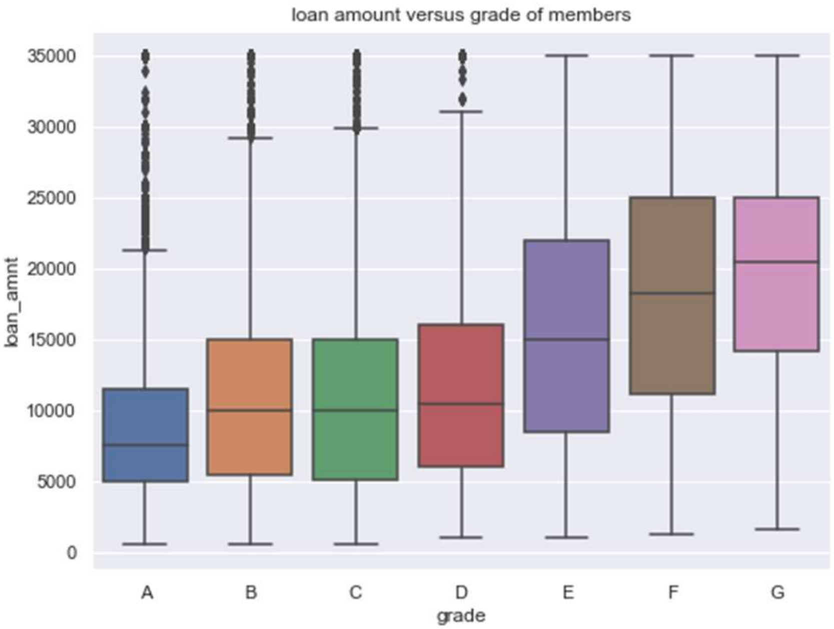
Bivariate Analysis

- Plotting grade against revolving utilization shows that revolving credit is found highest as the grades are increased
- In other words, grade A uses lowest revolving credit whereas G uses highest as per plot below



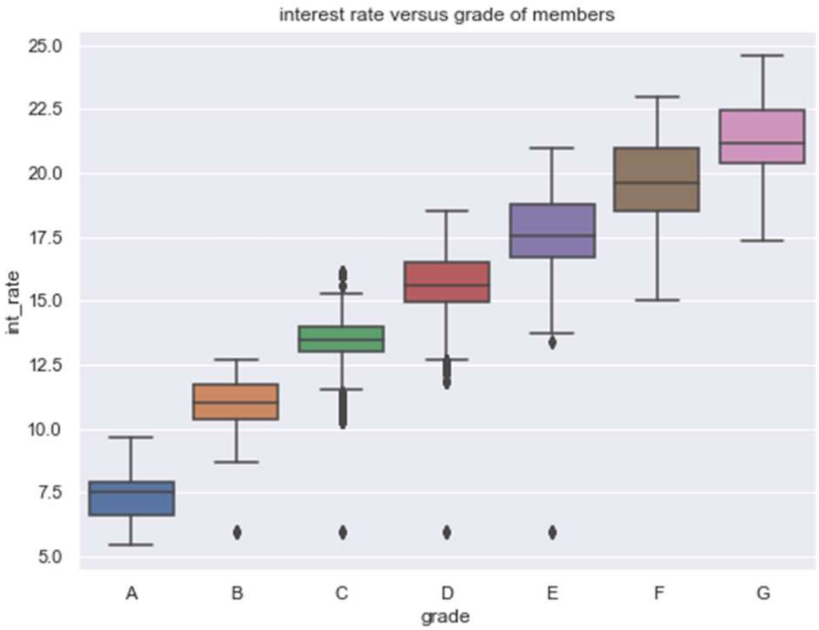
Bivariate Analysis

- Plotting grade against loan amount shows that loan amount is found highest as the grades are increased
- This follows and validates the previous observation where G grade members utilitize highest revolving credit and hence loan amounts are also higher along the grade chain.



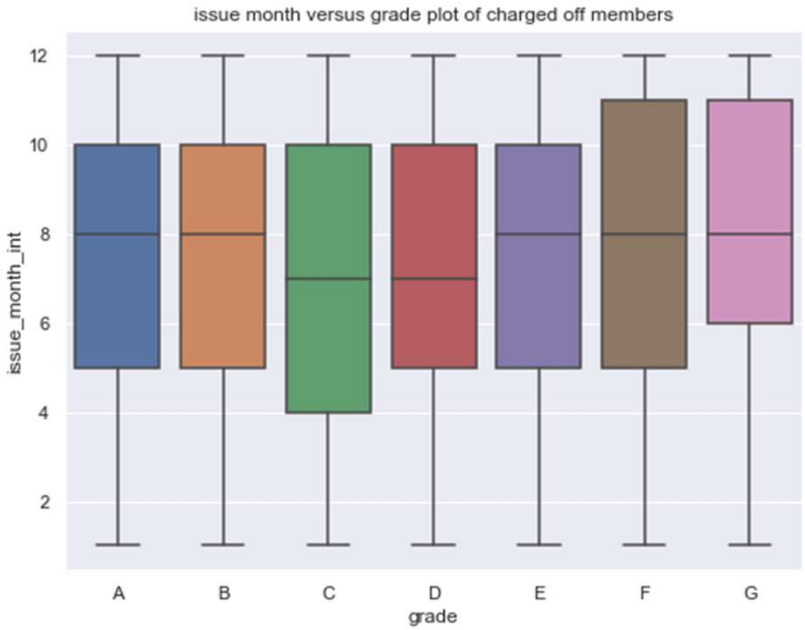
Bivariate Analysis

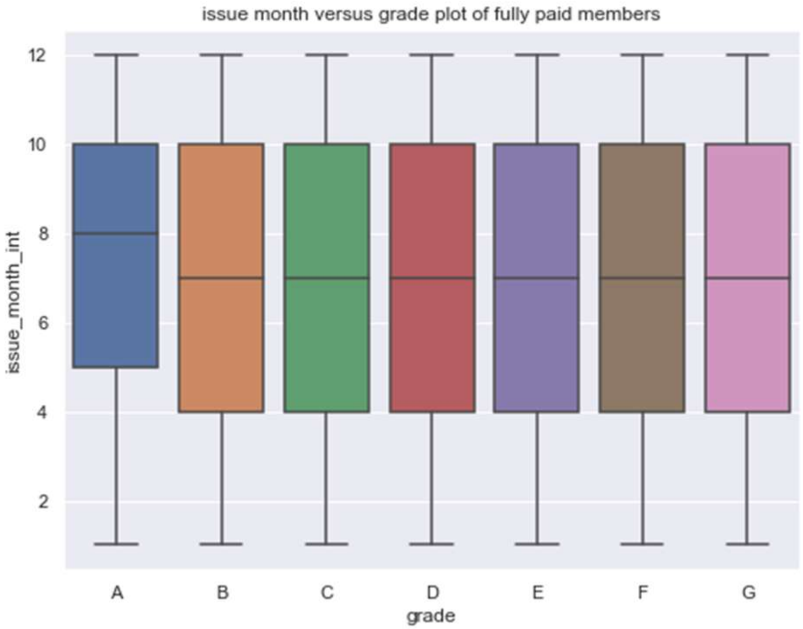
- Following the same analysis, we find interest rate is highest along the grade chain which complements the fact that higher revolving credit utilization is seen along the grade chain



Bivariate Analysis

- Plotting grade against issue month against both charged off members and fully paid members is shown below
- A unique observation is members of F & G grade taking loans in 11th month or November are at higher risk of default


issue month versus grade plot of charged off members
issue_month_int
A
B
C
D
E
F
G
grade


issue month versus grade plot of fully paid members
issue_month_int
A
B
C
D
E
F
G
grade