Lending club case study

EDA for defaulter on credits loans

Vivek Kumar
Samruddhi Salunkhe

Intro

- The purpose of this exercise is to analyse the data given by bank and check what factor leads to customer becoming defaulter, so that they can mitigate this

Approach

1. We look at the raw data for some basic understanding.
2. Removed all column with nothing but null values and then removed the columns where null was there as a part of data cleaning
3. Created some new columns for the columns that have range and perform binning on them
4. Then we analysing data (columns) in univariate fashion so that we will get to know which columns are of use for use.
5. Plots some graphs so that we will know which column have more relation with a defaulter and fully paid loan
6. After that we did Bivariate analysis of two columns with respect to defaulter ratio and plot some graphs accordingly.
7. In the end we plot a multivariate graph which shows correlation of diff columns that we thought are useful for our analysis.

Data Cleaning

- While looking at the data we found that almost 50% of columns have NA value.
  - There can be a case whether NA actually means something but while looking at the columns data we found nothing else was there.
  - We found few rows also have some NA data that has to be removed.

mo_sin	mo_sin	mo_sin	mo_sin	mort_a	mths_si	mths_si	mths_si	mths_si
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA
NA	NA	NA	NA	NA	NA	NA	NA	NA

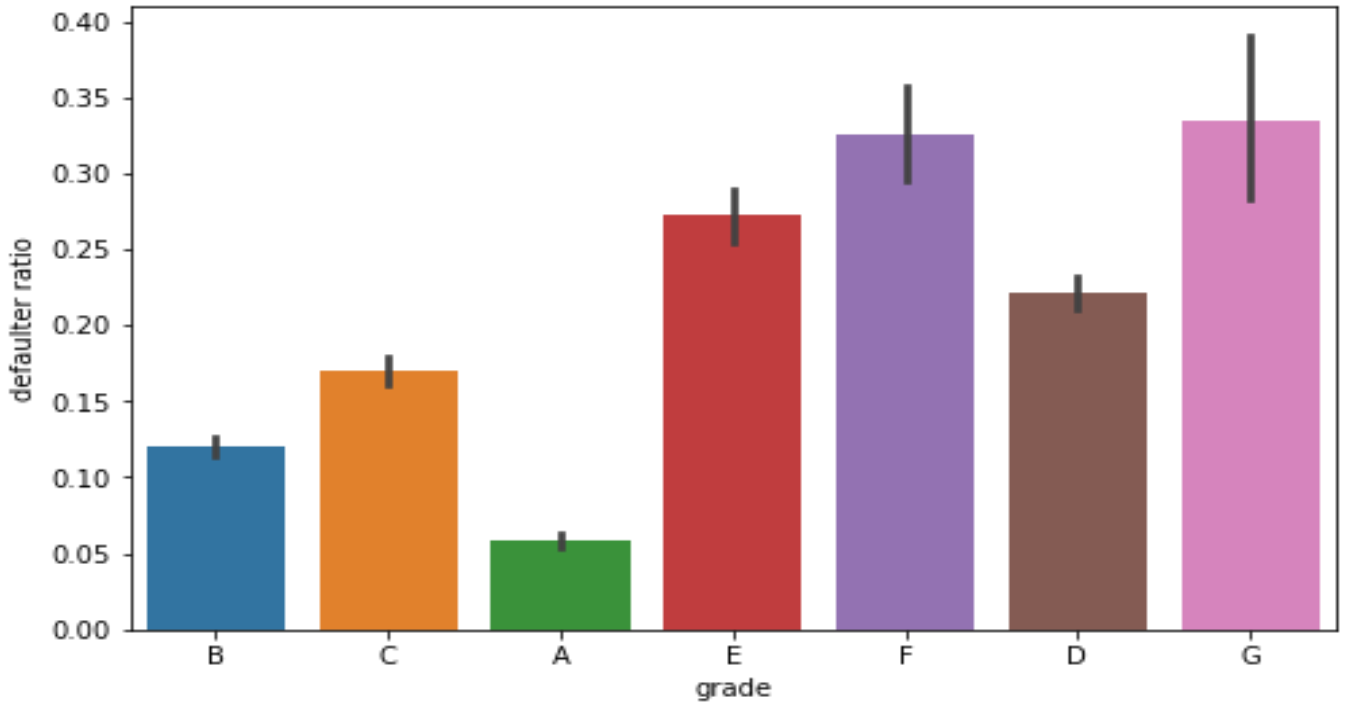
New columns for range data and Binning them

- We found out that few columns have some range of data so we tried to create new columns and put there range in those new columns like for interest rate and annual income salary

oan_amnt_range	int_rate_range	annual_inc_range	dti_range
0-5000	10-12.5	0-25000	25%+
0-5000	15+	25000-50000	0-5%
0-5000	15+	0-25000	5-10%
5000-10000	12.5-15	25000-50000	15-20%
0-5000	7.5-10	25000-50000	10-15%

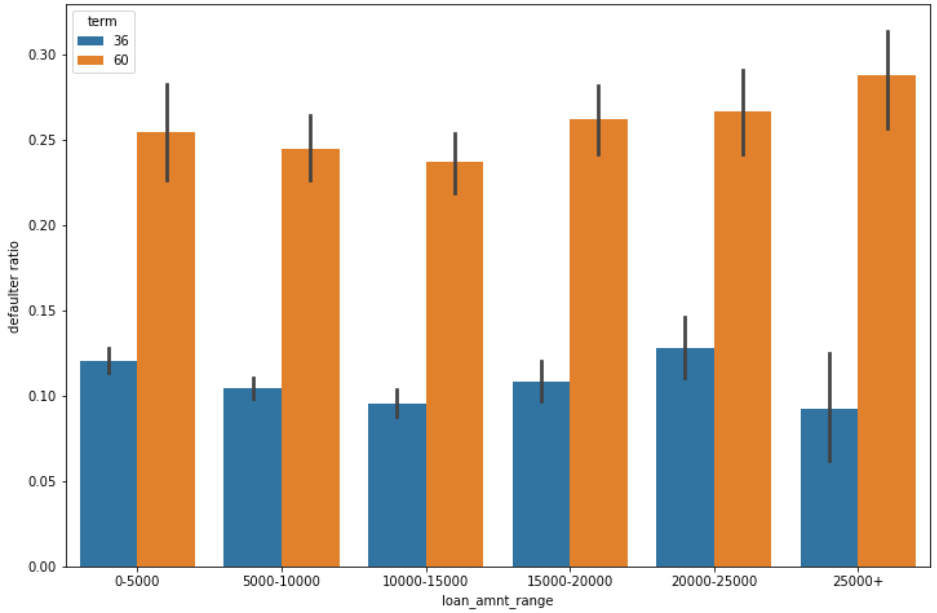
Univariate Analysis

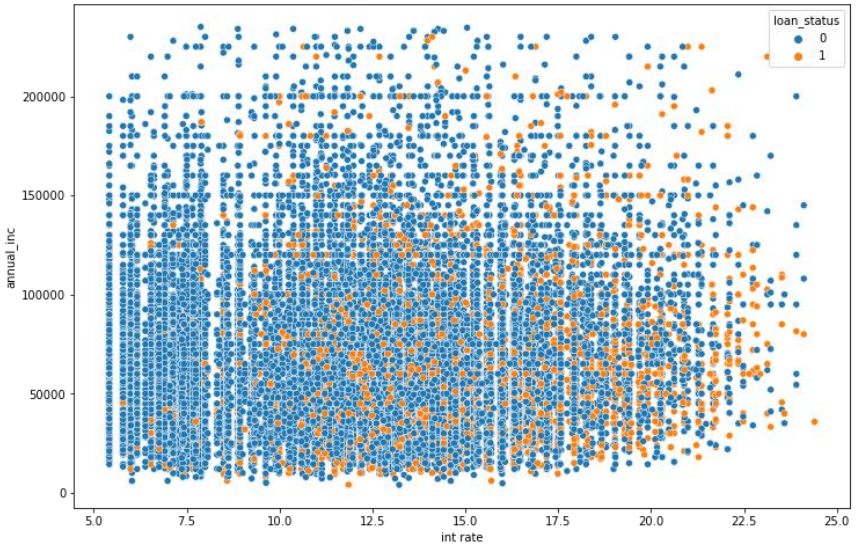
- After data cleaning and assembling data we compare our columns with defaulter ratio. Mean we check what is the relation between each column that can affect the customer payment of loan. We used bar plot for pictorial presentation. After this few more columns were removed from analysis as they were not impacting the case



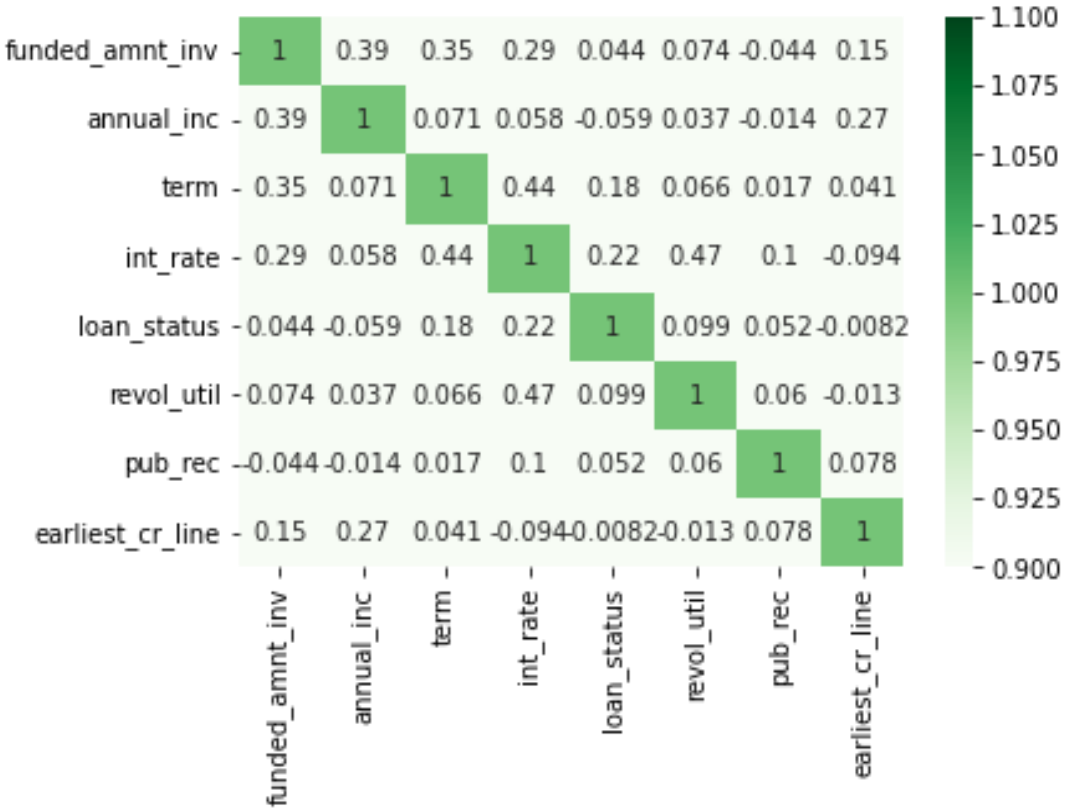
Bivariate Analysis

- Now we are left with comparatively less columns to go through bivariate analysis. So we combine two columns and compare them with defaulter ratio .for that we used bar plot and scatter plot as box plot was not able to plot all info like annual income with interest rate means for continuous data.





Multivariate Correlation Graph



In this graph we can see how these columns affect each other.

For example public record and funded amount invested are in –ve means so don’t pose much issue . whereas funded amount and annual income are dependant on each other.

Thanks