Lending Club Case Study

Vineet Kumar Sharma

Dev Kashyap

Problem Statement / Goal

- Given the data of existing customers who have either paid off or defaulted the loan, business wants to answer the question whenever a new application for loan is requested – “Will this applicant default in future ?”

- If the question can be answered with a high accuracy using EDA the business grows by lending money to the genuine customers while saving potential losses by predicting and rejecting a defaulter applicant

Assumptions

- Grade is assumed to follow the order A>B>C i.e customer with grade A is less likely to default then grade B and so on.
- Subgrade is assumed to follow the order A1 > A2 > A3 and likewise

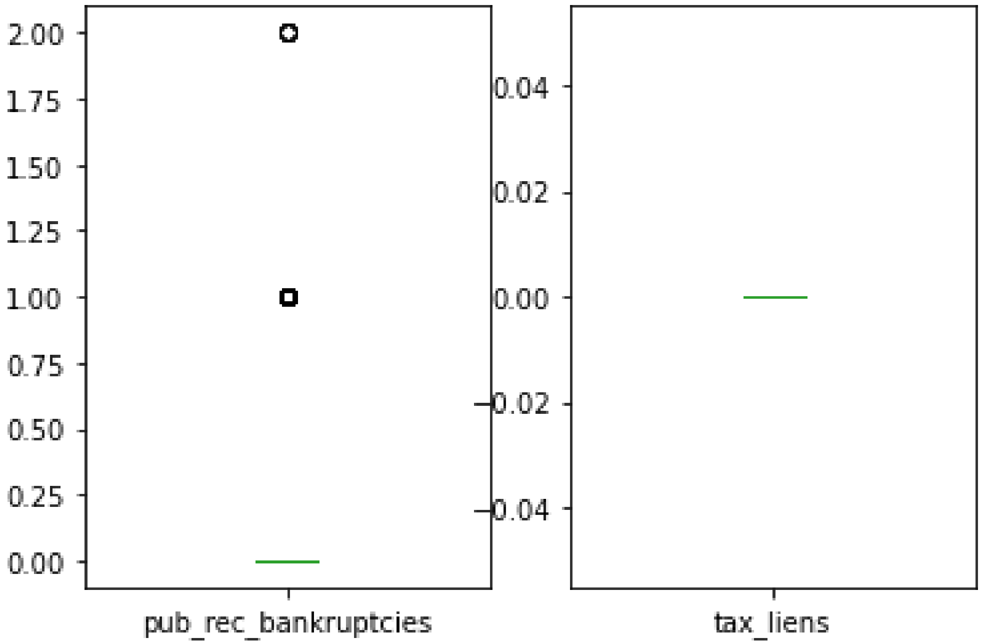
Approach -

- Understand the business domain and go through data dictionary.
- Based on understanding of domain think of the attributes that would affect the probability of an applicant to default.
- Drop all other columns from the data set. Keep only the attributes identified in last step.
- Clean the data so that aggregations and string operations can be performed by software. Remove or substitute the null values
- If a column has too many outliers then handle the outliers.

Approach - EDA

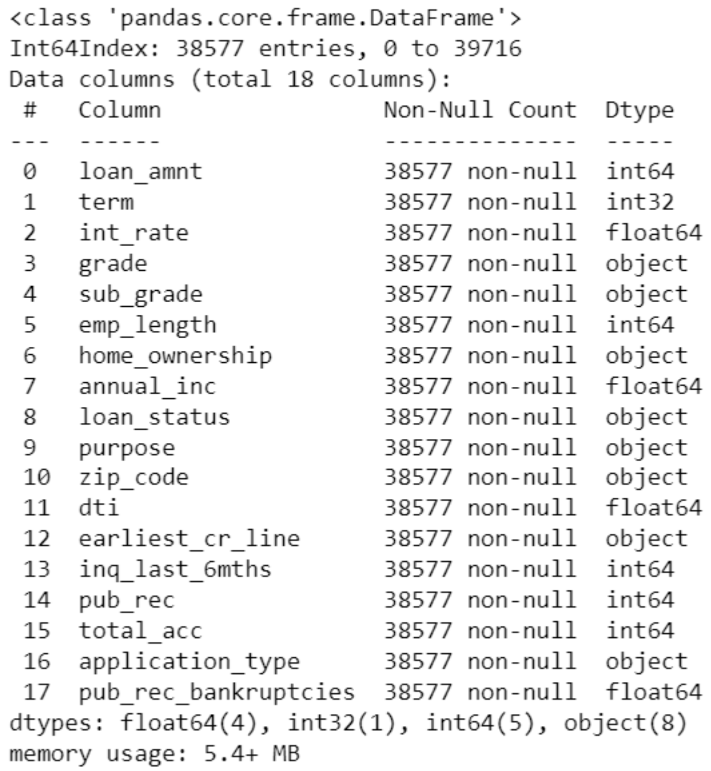
- First see basic pattern in data provided. It could tell how the data for a column is distributed. Is it good for identifying some trends ?
- Find corelation if any between any pair of columns
- Most important is segmented univariate analysis where we will keep target variable as loan_status and try to identify trend.
- Example mean of loan amount for fully paid could be 5000$ and for charged off mean of loan amount could be 10000$. Means a higher loan amount increases chances of default

Imputing data by using IQR



These two columns were chosen after studying domain. Turns out the captured data does not have much variation. We dropped tax_liens column and substituted nulls with 0 for pub_rec_bankruptcies

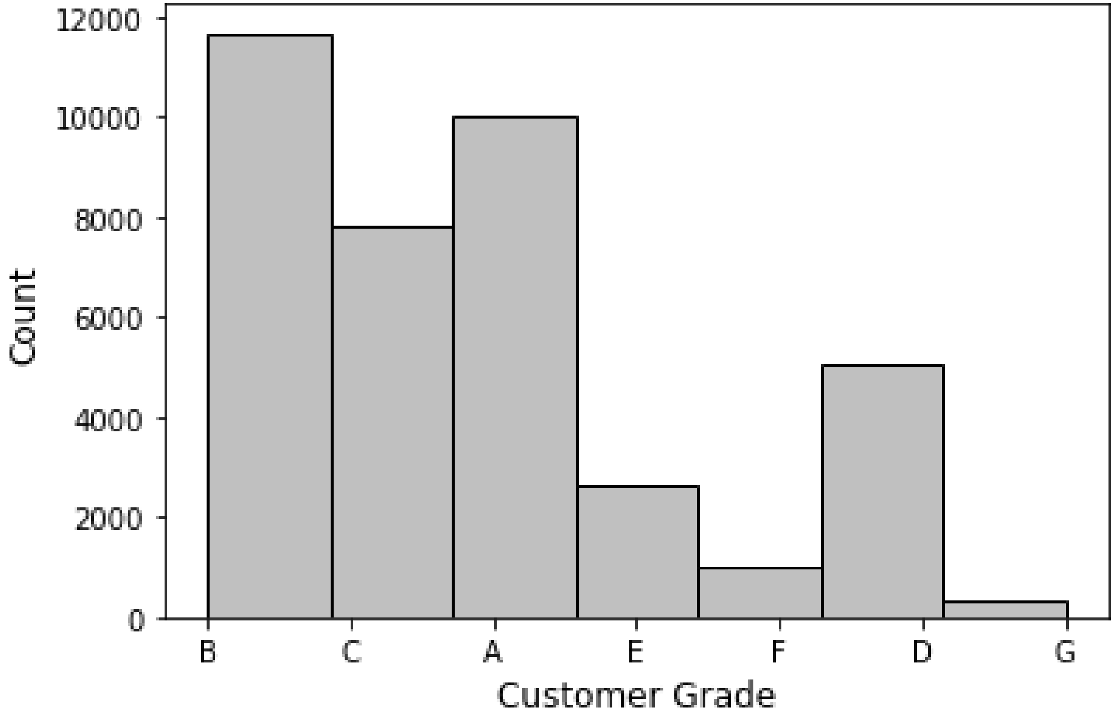
Final data set after clean up

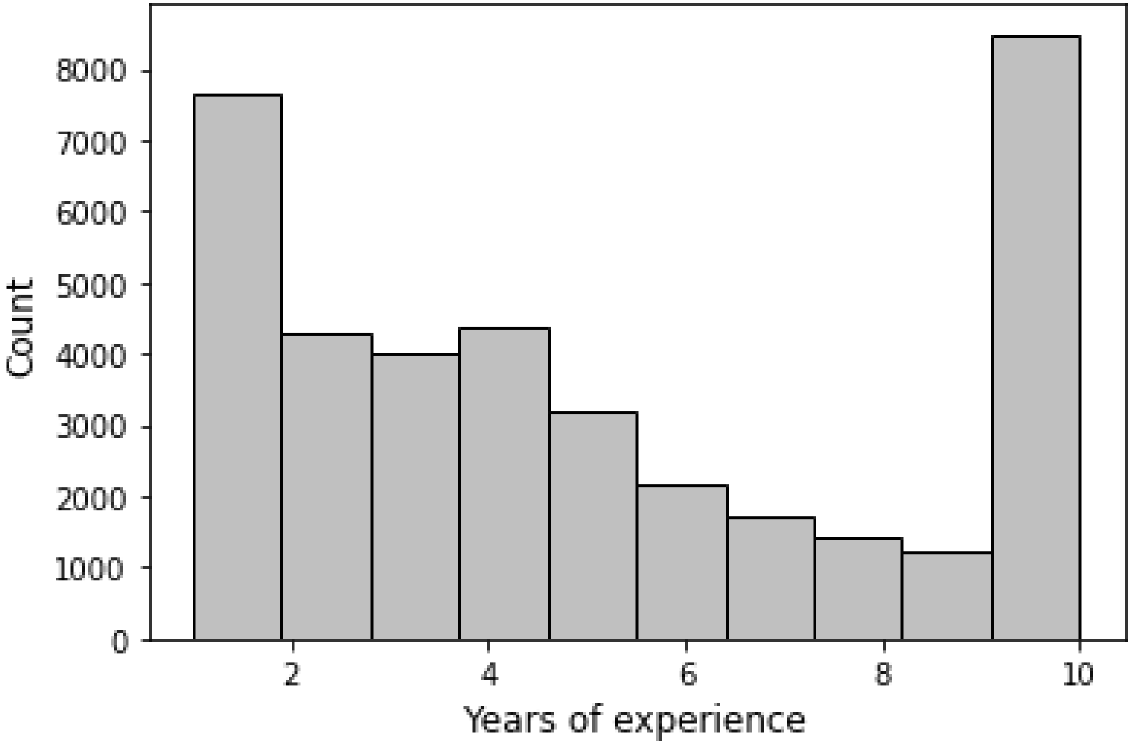


```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 38577 entries, 0 to 39716
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype  
---  ------                --------------  -----  
 0   loan_amnt             38577 non-null  int64  
 1   term                  38577 non-null  int32  
 2   int_rate              38577 non-null  float64
 3   grade                 38577 non-null  object 
 4   sub_grade             38577 non-null  object 
 5   emp_length            38577 non-null  int64  
 6   home_ownership        38577 non-null  object 
 7   annual_inc            38577 non-null  float64
 8   loan_status           38577 non-null  object 
 9   purpose               38577 non-null  object 
 10  zip_code              38577 non-null  object 
 11  dti                   38577 non-null  float64
 12  earliest_cr_line      38577 non-null  object 
 13  inq_last_6mths        38577 non-null  int64  
 14  pub_rec               38577 non-null  int64  
 15  total_acc             38577 non-null  int64  
 16  application_type      38577 non-null  object 
 17  pub_rec_bankruptcies  38577 non-null  float64
dtypes: float64(4), int32(1), int64(5), object(8)
memory usage: 5.4+ MB
```

EDI Analysis - Graphs

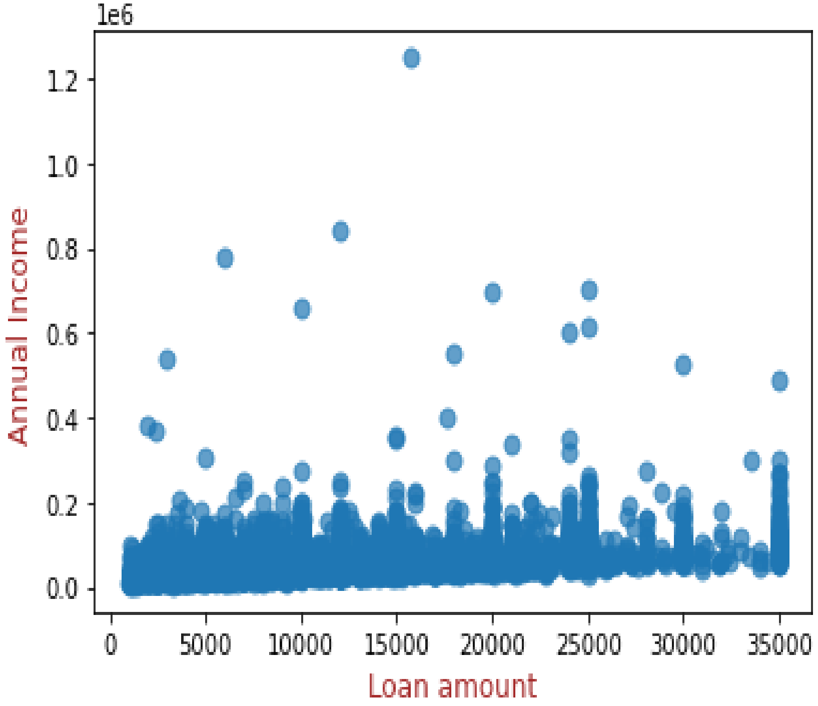
- Distribution of data by categories



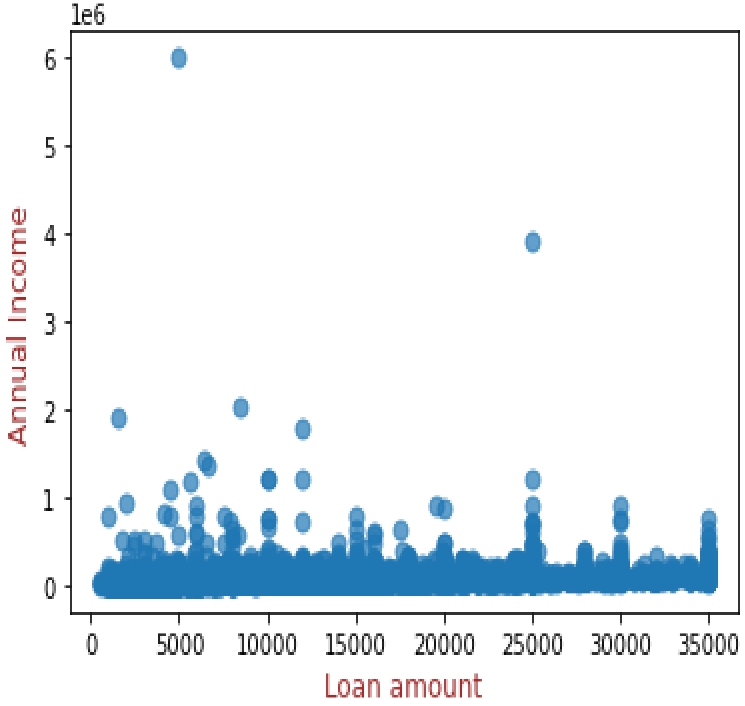


EDA – Graphs Contd

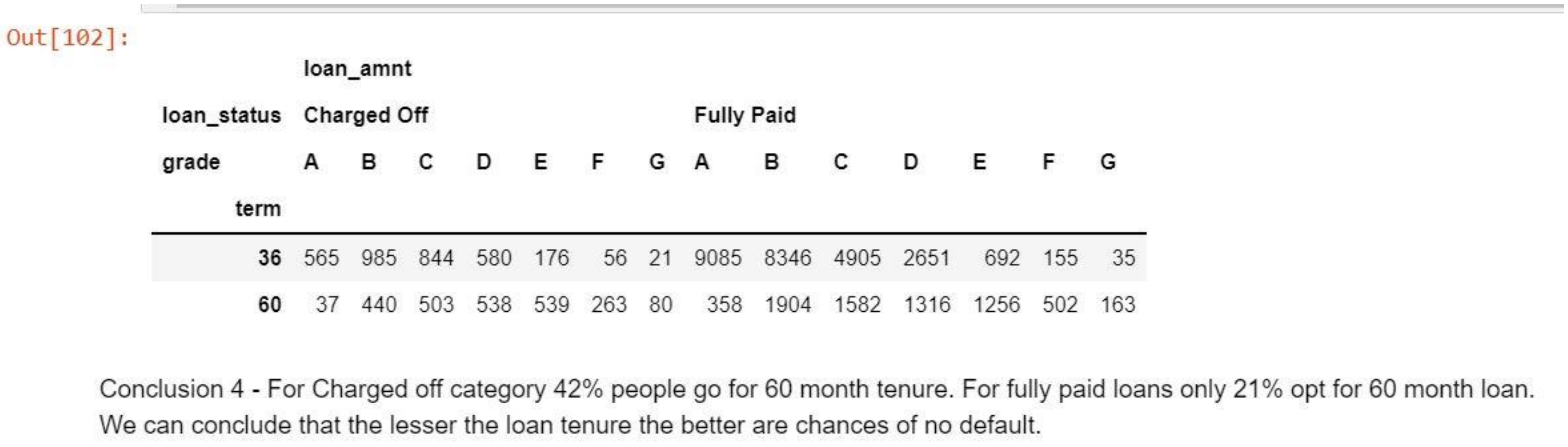
Annual Income vs Loan amount for defaulted loans



Annual Income vs Loan amount for Fully Paid loans



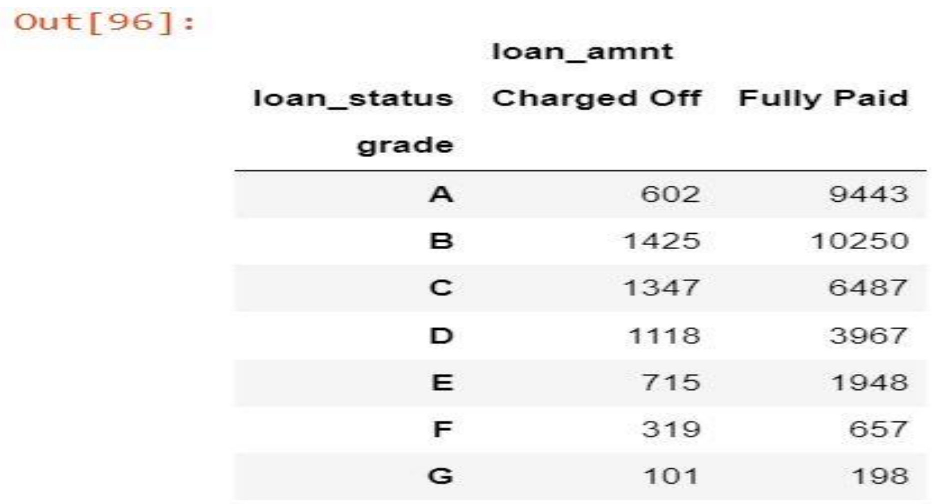
Pivot tables



Out[102]:

	loan_amnt													
loan_status	**Charged Off**							**Fully Paid**						
grade	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**A**	**B**	**C**	**D**	**E**	**F**	**G**
term														
36	565	985	844	580	176	56	21	9085	8346	4905	2651	692	155	35
60	37	440	503	538	539	263	80	358	1904	1582	1316	1256	502	163

Conclusion 4 - For Charged off category 42% people go for 60 month tenure. For fully paid loans only 21% opt for 60 month loan. We can conclude that the lesser the loan tenure the better are chances of no default.



Out[96]:

	loan_amnt	
loan_status	**Charged Off**	**Fully Paid**
grade		
A	602	9443
B	1425	10250
C	1347	6487
D	1118	3967
E	715	1948
F	319	657
G	101	198

Conclusions/Recommendations

- annual income is a very strong attribute to consider (especially for higher loan amounts)
- The riskiest of loan purpose is small_business followed by debt_consolidation and credit_card
- Grading system used by Loan Club is working almost accurately in real world.
- Lesser the loan tenure the better are chances of no default.
- DTI seems to have almost no effect on the loan paying capacity which is counterintuitive

Thank You