Lending club case study

- Bharath veer KS
- Sivaskandan P

Problem statement

- Analysing the given loan dataset to identify the patterns to determine if
  - The loan applicant will be most likely to pay the loan
  - The loan applicant most likely to default the loan

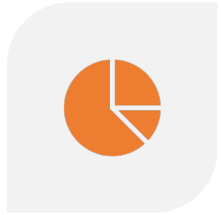
Structured apporach

DATA ANALYSIS

DATA CLEANING

UNIVARIATE ANALYSIS

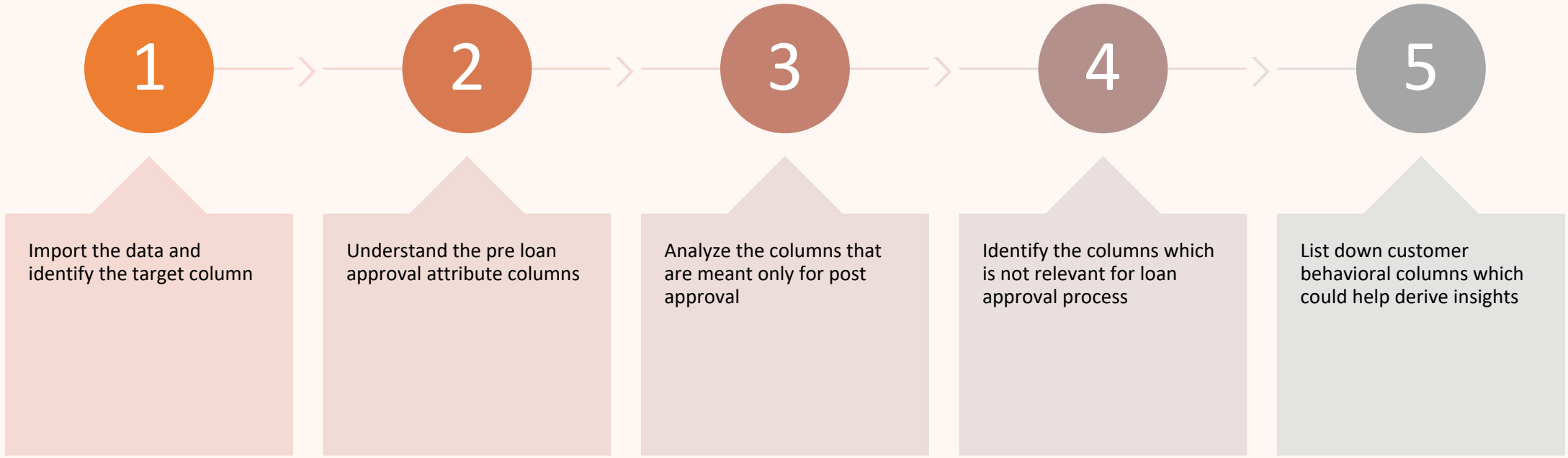
BI VARIATE ANALYSIS



MULTIVARIATE ANALYSIS

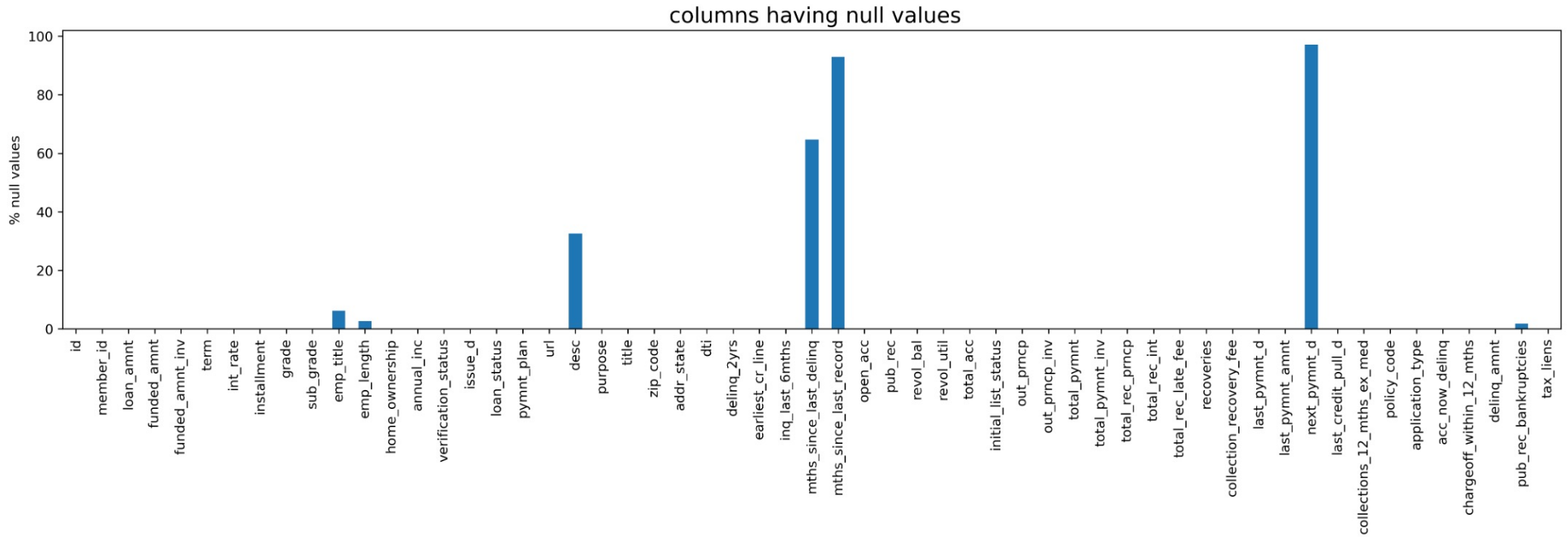
RECOMMENDATIONS

Data analysis



Data clean up

- We removed all columns that has only null values
- Analyze columns with redundant values and eliminate
- We standardized number columns by removing unwanted characters and convert them to numeric
- Derive new columns from existing columns to plot various graphs



Insights on data cleanup/correction

- There are columns with redundant values which has to be eliminated
- There are numerical columns with special characters which will be cleaned and converted to numeric type
- Remove the Current ongoing loan application as it is not relevant to our problem statement
- Drop columns where we observe 50% NA values
- Drop post loan approval columns

Univariate analysis

Broadly done on three types of columns

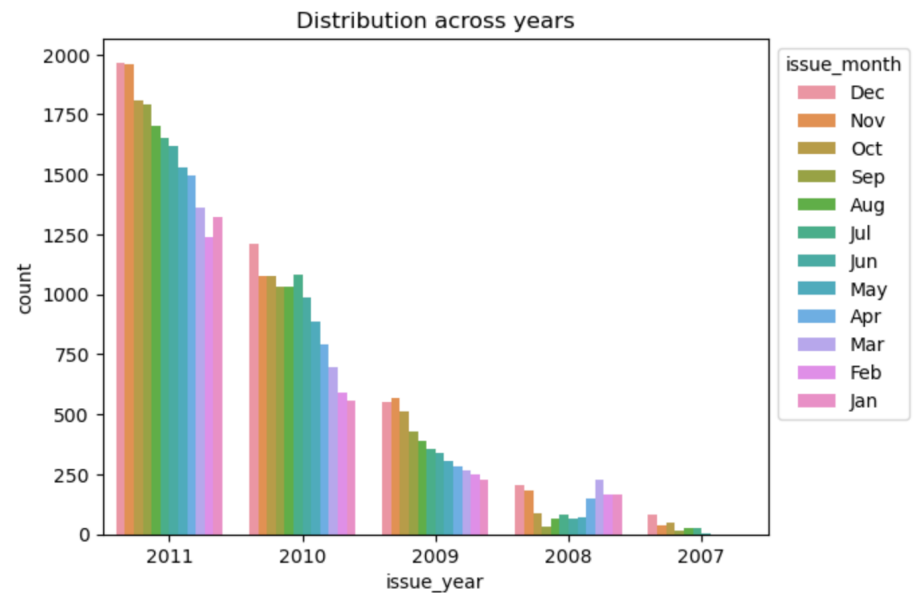
- Segmented categorical
- Ordered categorical
- Quantitative

What strategies used

- Single frequency tables
- Derive new columns wherever possible
- Curate columns that could correlate target variable

```
#plot countplot with issue_year hueed by issue_month to find the distribution of each year and month
ax = sns.countplot(x='issue_year', hue='issue_month', data=loan)
sns.move_legend(ax, "upper left", bbox_to_anchor=(1, 1))
plt.title("Distribution across years")
```

```
Text(0.5, 1.0, 'Distribution across years')
```



Summary on Univariate analysis

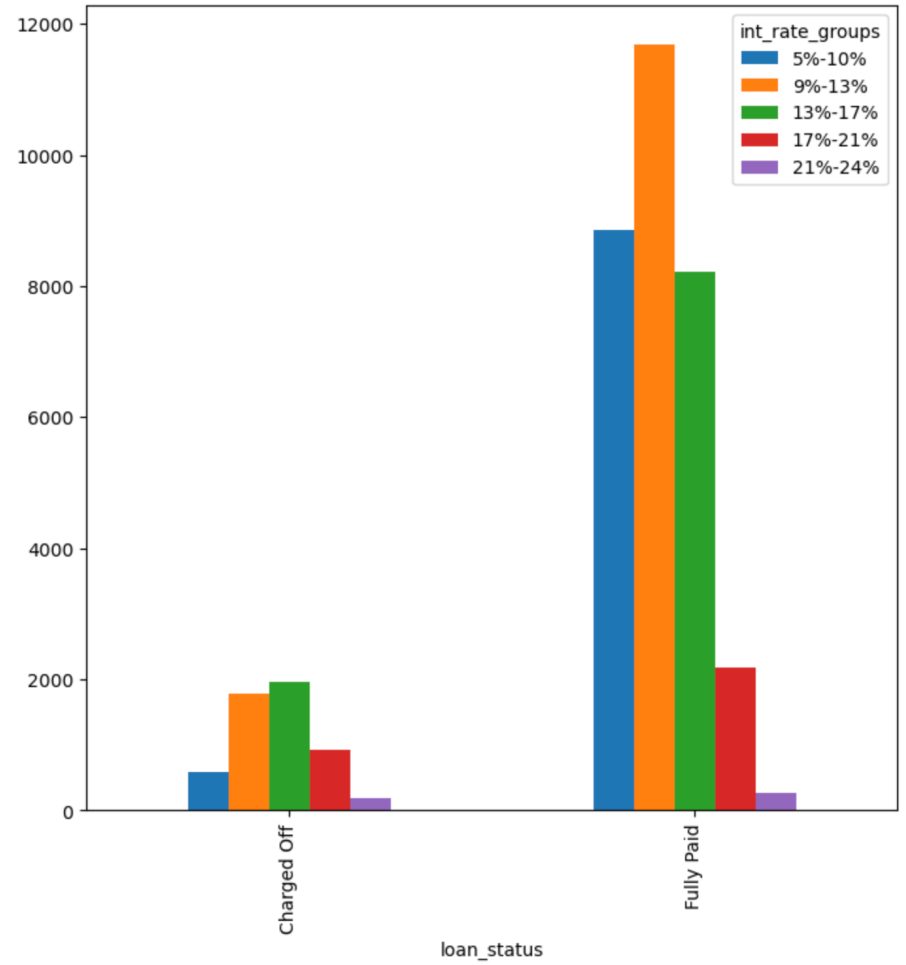
- Year end has more applicants for loan than year start
- Rent and Mortgage house holders are seeking for loan more than own house people
- People above 10+years seeks loan more than lower experienced
- Removed outliers in Annual income field to get proper distribution of data.
- Annual income of 50K is the median of people who needs loan
- In int_rate charging off is less in case of interest rate 21-24% but the total people taken 21-24% is less
- If we see the total proportion in interest rate 9-13% interest rate has increased chance of being fully paid.
- The interest rate 13-17% high percentage of getting charged off.
- In annual_inc groups people with 112k - 140k has less percentage of getting charged off.
- The annual_inc group with 31k-58k has higher percentage of getting charged off.

Bi variate analysis

Plotting ideas used to understand the defaulters are mentioned below

- Heatmap to find correlation matrix
- Scatter plot to understand correlation between loan_amnt
- Cross tabs to correlate categorical vs categorical
- Count plots to visualize the grades again target variable

```
# crosstab function is used to create derived table by plotting against loan_status and int_rate_groups
int_crosstab = pd.crosstab(index=loan["loan_status"],
                           columns=loan["int_rate_groups"])
int_crosstab.plot(kind="bar",
                  figsize=(8,8),
                  stacked=False )
plt.show()
```



Summary of Bivariate analysis

Below are the probable defaulters' reasons

- Higher the interest rate as higher (13-17%) interest range has high defaulters
- Debt consolidation stands out in the defaulter's loan where it could be borrowers took a loan to close existing loan but failed again
- We can clearly see during the year 2011 end there are more defaulters which is directly correlating to recession in U.S hence market situation-based loan approvals would solve most of these.
- Employees with 10+ years has more probability of being a defaulter
- Employees with annual income range 31k-58k has more probability of being a defaulter
- The people who are marked as grade G has higher chance of getting charged off
- The Sub Grade F5 has the highest percentage of Charged off

Good candidates for providing loan which is inversely proportional to defaulters

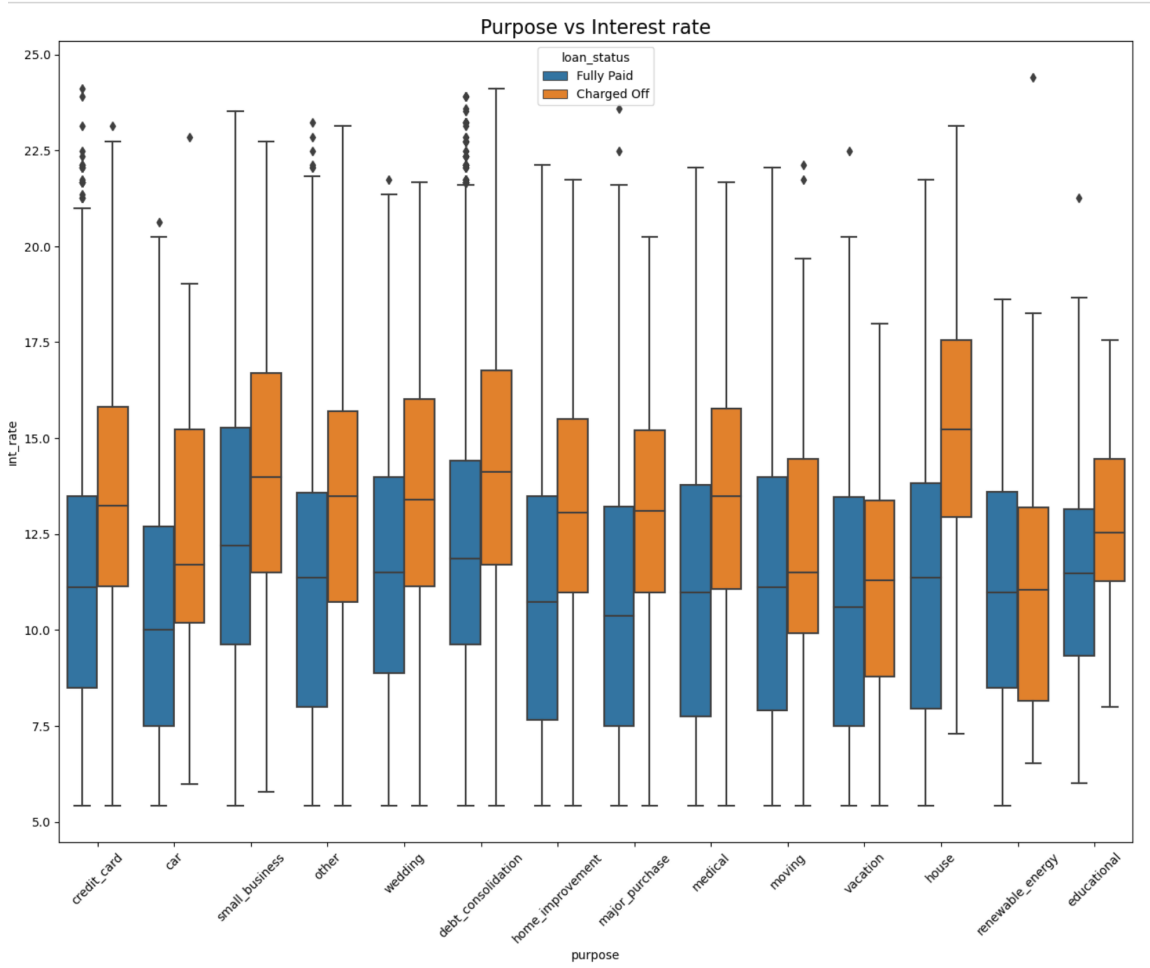
- Obviously, lesser interest rate helps borrowers repay the loan without fail
- People who owns a house have good records of closing the committed loan amount
- The grade B has higher chance of getting fully paid
- The interest rate 9-13% has a higher chance of getting paid.
- The Sub Grade A1 has the highest percentage of fully paid
- The income group 112k - 140k has less charged off count. Could be due to their high income.

Multi variate analysis

We did multi variate analysis on top of our bi variate attributes to understand the impact of a third variable on the target

For example, when we validated purpose column debt_consolidation appears to be on higher level on defaulters, however when interest rate is imposed as third column we could see the percentage small_business and house has higher defaulters in real

By doing above step we were able to derive some depth insights on the influence of one variable over other



Summary of Multi variate

Unlikely to repay

- The People in 31k - 58 k income group with a own house has higher chance of paying the loan
- Loan amnt higher than 12000 dollars has high chance for small business of getting Charged off.
- We could see here that irrespective of emp_length when the loan_amnt increases probablity of Charged off is slightly increasing

Likely to repay

- The People in 31k - 58 k income group with house in mortgage or living in a rented house has higher chance of defaulting the loan
- Loan amnt lesser than 12000 dollars has high chance for small business of getting fully paid.
- Interest rate group in 21 - 24% having annual income 70k+ has the good candidates for offering the loan

Recommendations

Probability of defaulters

- Loan amnt higher than 12000 dollars for small buisiness has slightly high chance of geting Charged off.
- The employees with 10+ years of experience and a annual income of 31k - 58k are more likely to get charged off
- People who are buying loan for their house between interest rate 12.5 - 17.5 are more likely to get charged off
- Car loan with interest rates above 15 are more likely to get charged off
- Small_buisness loans with higher interest (above 11%) are more likely to be charged off
- The People in 31k - 58 k income group with house in mortgage or living in a rented house has higher chance of defaulting the loan
- The people who are buying loan for debt consolidation with home ownership has rent or mortgage has higher chance of getting charged off

Probability of eligible candidates

- The employees with 10+ years of experience and a annual income of 51k - 85k are more likely to fully pay the loan
- People who are buying loan for their house less than 13% interest are more likely to fully pay the loan
- Car loan within interest rates 7.5 to 10 are more likely to get Fully Paid
- Small_buisness loans with less interest (11%) are more likely to be fully paid
- Loan amount lesser than 12000 dollars for small buisiness has high chance of geting fully paid.
- The People in 31k - 58 k income group with a own house has higher chance of paying the loan
- Intrest rate group in 21 - 24% having annual income 70k+ has the good candidated for offering the loan