Lending Club Case Study

Group_Facilitator_(Soham Kulkarni)

Problem statement

Lending club is a portal where lenders and borrowers meet for financial needs and benefits. In order to make this platform full proof from frauds or avoid approving loans higher than borrower's capacity, portal needs to maintain pre loan checks of applicants.

Problem statement: With available data set, indicators to be established in terms of borrower's ability/tendency of repayment. These indicators shall be useful for future loan approval process.

Approach:

- Given data would be studied first for information available within.
- Data which is not relevant for the study, or incomplete data will be then removed as part of data cleaning.
- Sanity checks, inconsistent data points also be taken care.
- As a part of analysis, overall percentages of loan amounts, charged off/fully paid loan details, interest rates distribution, localities of which client details involved etc will be studied.
- Next step will be to find driving parameters towards loan repayment ability. The generic details mentioned above such as borrowers personal detail like age, income, purpose of loan, employer, debt to income ratio, also terms of loan like interest rate, loan amount will be plotted against whether loan is repaid or no.

Understanding Data

Given data basically contains the information of customers who have borrowed the money in the form of as shown in below

Borrowing details: borrowed amount, borrowing date, purpose , interest rate. Loan paid or charged off etc

Borrower's details: member id, employment background, income details, debt to income ratio, home ownership, demographic details etc.

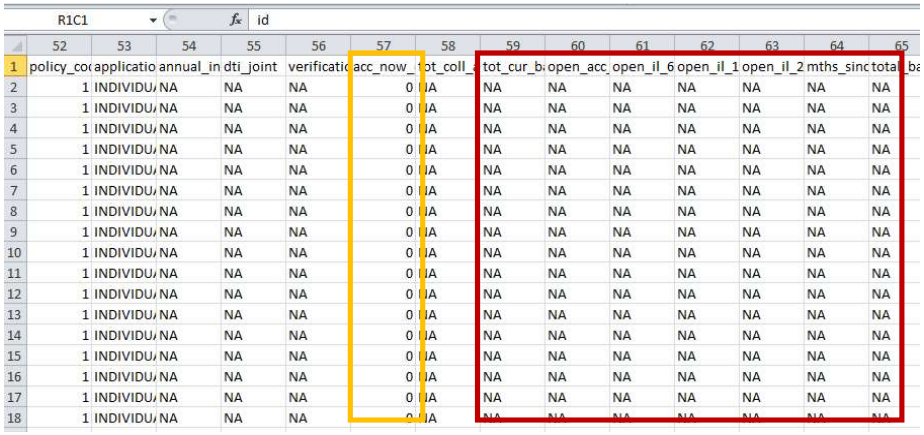
Some terminologies and ratios such as dti, funded_amnt_inv, funded_amnt etc are understood with help of reference data sheet.

Loan status with "Current" (ongoing loans), data is not considered for this analysis, as it will have 50-50 probability of defaulting or fully paid

Other details such as behavioral aspects such as earliest_cr_line, revol_bal, revol_util will not be considered. As these aspect come to existence once loan is approved.

Data Cleaning

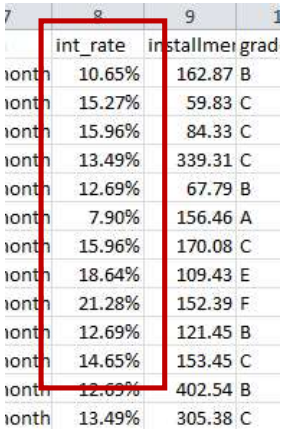
NA, 0 value treatment



	52	53	54	55	56	57	58	59	60	61	62	63	64	65
1	policy_co	applicatio	annual_in	dti_joint	verificatio	acc_now_	ot_coll_	tot_cur_b	open_acc_	open_il_6	open_il_1	open_il_2	mths_sinc	total_ba
2	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
3	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
4	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
5	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
6	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
7	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
8	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
9	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
10	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
11	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
12	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
13	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
14	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
15	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
16	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
17	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA
18	1	INDIVIDU	NA	NA	NA	0	NA	NA	NA	NA	NA	NA	NA	NA

Columns with more than 75% "NA" values removed

Columns with more than 90% "0" values removed



	8	9	1
	int_rate	installmer	grad
onth	10.65%	162.87	B
onth	15.27%	59.83	C
onth	15.96%	84.33	C
onth	13.49%	339.31	C
onth	12.69%	67.79	B
onth	7.90%	156.46	A
onth	15.96%	170.08	C
onth	18.64%	109.43	E
onth	21.28%	152.39	F
onth	12.69%	121.45	B
onth	14.65%	153.45	C
onth	12.69%	402.54	B
onth	13.49%	305.38	C

Values in percentages such as "int_rate", string separated and "%" sign removed, then converted back to float values again

```
df2['intrate']=df2['int_rate'].str.replace('%','').astype(float)
```

Data Cleaning

22	23
tle	zip_code
ɔmputer	860xx
ike	309xx
eal estat	606xx
ersonel	917xx
ersonal	972xx
ly wedd	852xx
ɔan	280xx
ar Down	900xx
xpand B	958xx
uilding	774xx
igh intre	853xx
ɔnsolida	913xx
eedom	245xx
ticard fu	606xx
ther Loa	951xx

Derived Metrics

49	50
zipcode	zpcod_bin
860	(850, 860]
309	(300, 310]
606	(600, 610]
917	(910, 920]
972	(970, 980]
852	(850, 860]
280	(270, 280]
900	(890, 900]
958	(950, 960]
774	(770, 780]
853	(850, 860]
913	(910, 920]

```
df2['zipcode']=df2['zip_code'].str.replace('xx','').astype(int)
bins_zpcod=np.arange(0,1000,10)
df2['zpcod_bin']=pd.cut(df2['zipcode'],bins_zpcod)
```

From columns like "zip_code", letters "xx" removed, then converted to int values

Later zip codes were binned in range of 10s, and used as Derived Metrics

27
earliest_c i
Jan-85
Apr-99
1-Nov
Feb-96
Jan-96
4-Nov
5-Jul
7-Jan

27
earliest_c
Jan-85
Apr-99
Nov-22
Feb-96
Jan-96
Nov-22
Jul-22
Jan-22

Date formats were unified, for ex. Conversion of dd_yyyy to mm_yyyy.
This further helped in converting data to DateTimeIndex

Data Cleaning

emp_title
Baptist Health
Baptist Health of South Florida
Baptist Health
Baptist Health Systems of South Florida
Baptist Health Systems
Baptist Health South Florida
Baptist Health South Florida
Baptist Health South Florida
Baptist Health South Florida
Baptist Health Systems

emp_title_original
Baptist Health
Baptist Health of South Florida
Baptist Health
Baptist Health of South Florida
Baptist Health of South Florida
Baptist Health
Baptist Health of South Florida
Baptist Hospital East
Baptist Health Medical Center NLR
Baptist Health of South Florida
Baptist Health of South Florida
Baptist Health Medical Center NLR
Baptist Health

Ex. 1

Python SequenceMatcher is used to unify these names, with sequence ratios above 0.7-0.8

```
import time
a=0
df2['emp_title'].fillna('NAA', inplace=True)
t0 = time.time()
for i in df2['emp_title']:
    print(i)
    if (i =='NAA'):
        break
        b=a+1
        for j in df2['emp_title'][a+1:]:
            print('1')
            seqMatch = SequenceMatcher(None,i,j)
            if ((seqMatch.ratio()>0.7) & (0.75<(len(i)/len(j))<1.3)):
                df2['emp_title'][b]=str(i)
            elif (seqMatch.ratio()>0.8):
                df2['emp_title'][b]=str(i)
        b=b+1
    a=a+1
t1 = time.time()
numpy_time = t1 - t0
print (t1-t0)
```

Same names spelled differently, or with no space or with additional character had to be treated and corrected. For eg "BanofAmerica" & "Bank Of America" are same. Or "Baptist Health South Florida" & "Baptist Health **Of** South Florida" are same

Finally names are converted to lower case to avoid any further mismatch

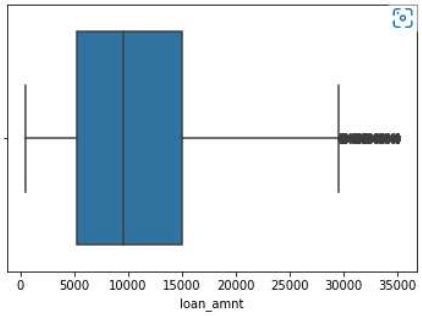
Ex. 2

emp_title
Bank of America
Bank of America
Bank of America
Bank of America Merrill Lynch
Bank of America Merrill Lynch
Bank Of America
Bank of America
BankofAmerica
Bank of America
Bank of America
Bank Of America
Bank of America
Bank Of America
Bank of America
Bank of America
Bank of America
Bank of America Merrill Services
bank of america
Bank of America
Bank of America
Bank of America / Merrill Lynch

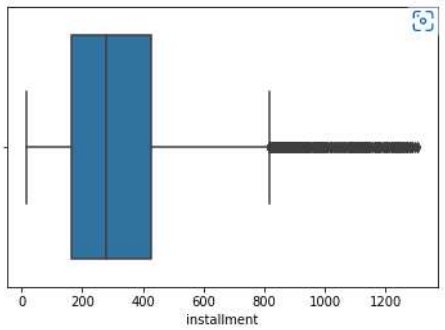
11	12
emp_title_original	emp_title
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America Merrill Lynch	bank of america merrill lynch
Bank of America Merrill Lynch	bank of america merrill lynch
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America Merrill Lynch	bank of america merrill lynch
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America	bank of america
Bank of America Merchant Services	bank of america merchant servi
Bank of America	bank of america

Distribution check outliers removal

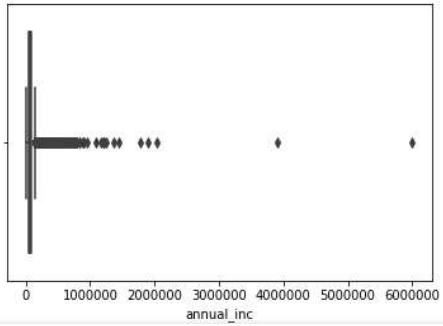
loan_amnt



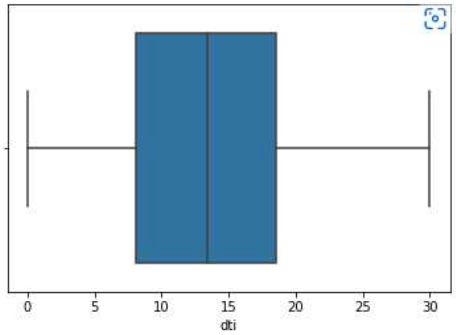
installment



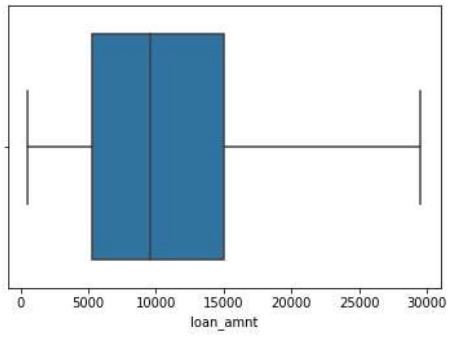
annual_inc



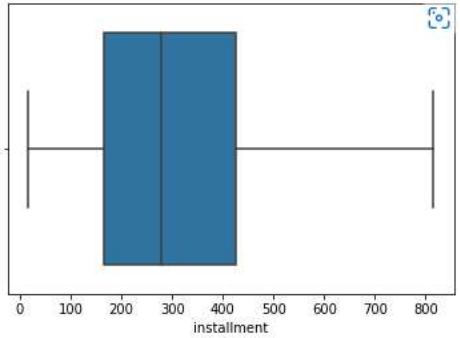
dti



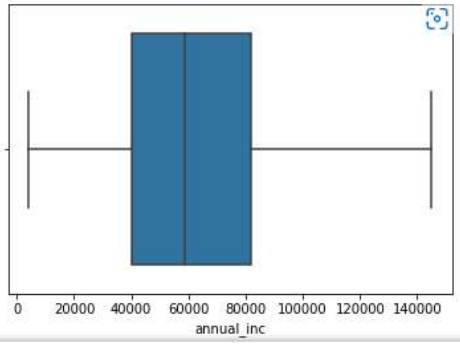
loan_amnt



installment

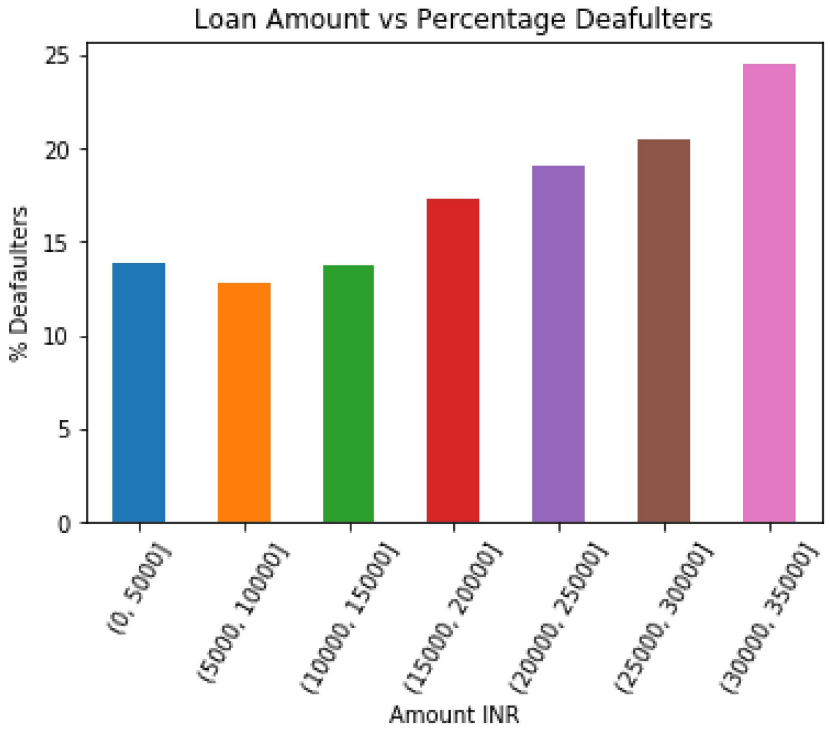


annual_inc



Loan amount vs % Defaulters

- To check co-relation in requested loan amount on defaulter rate, loan amount bins prepared in range of INR 5000
- It can be seen from adjacent graph, as loan amount increasing, rate of defaulters is also increasing.
- The rate is highest or last bin 30-35k
- The analysis is performed based on number of charged off loans and not sum of amount
- Recommendation:
- Special focus can be given on higher loan amount customer if not already been given.
- Additional checks shall be assigned as loan amount increases.



Employee title co-relation

- Borrowers' firm in pivot table including number of borrowers to particular firm, their loan status, with ratio showing defaulters and finally the ratio where overall amount is charged off to fully paid amount.
- The ratio **more than 1** means the loan amount charged off is higher than loan amount fully paid. For ex in table WalMart is having highest loss2revenue ratio
- "US Army" is having **highest Charged Off** count(defaulters count), but it has **high fully paid count as well**
- Recommendation :The top firms in list with high loss2revenue ratio as well high perc_defaulters or Charged Off count **shall be on radar for debt repayments**.
- And **caution** must be taken **while lending money** to these firms

Number of borrowers

Out of total charged off or fully paid amount what percentage particular firm's employee account for

Ratio shows charged off amount ratio

loan_status	Charged Off	Fully Paid	perc_defaulters	perc_amnt_chargedoff	perc_amnt_fullypaid	loss2revenue_ratio
emp_title						
wal-mart	30.0	85.0	26.1	0.74	0.43	1.72
postal service	13.0	49.0	21.0	0.46	0.37	1.24
va medical center	10.0	42.0	19.2	0.34	0.29	1.17
ups	19.0	75.0	20.2	0.73	0.63	1.16
self employed	19.0	80.0	19.2	0.50	0.52	0.96
at&t	17.0	68.0	20.0	0.50	0.55	0.91
united states postal service	27.0	113.0	19.3	0.75	0.82	0.91
central school district 301	10.0	45.0	18.2	0.27	0.30	0.90
cox communications	24.0	104.0	18.8	0.79	0.90	0.88
harlem consolidated school district	14.0	80.0	14.9	0.51	0.58	0.88
steris corporation	16.0	92.0	14.8	0.58	0.71	0.82
bank of america	26.0	138.0	15.9	0.72	1.01	0.71
verizon wireless	15.0	67.0	18.3	0.36	0.53	0.68
department of justice	13.0	108.0	10.7	0.51	0.82	0.62
us army	34.0	201.0	14.5	1.08	1.82	0.59
kaiser permanente	10.0	66.0	13.2	0.30	0.58	0.52
us navy	10.0	73.0	12.0	0.29	0.63	0.46
us air force	10.0	71.0	12.3	0.20	0.45	0.44
wells fargo bank	10.0	94.0	9.6	0.31	0.72	0.43

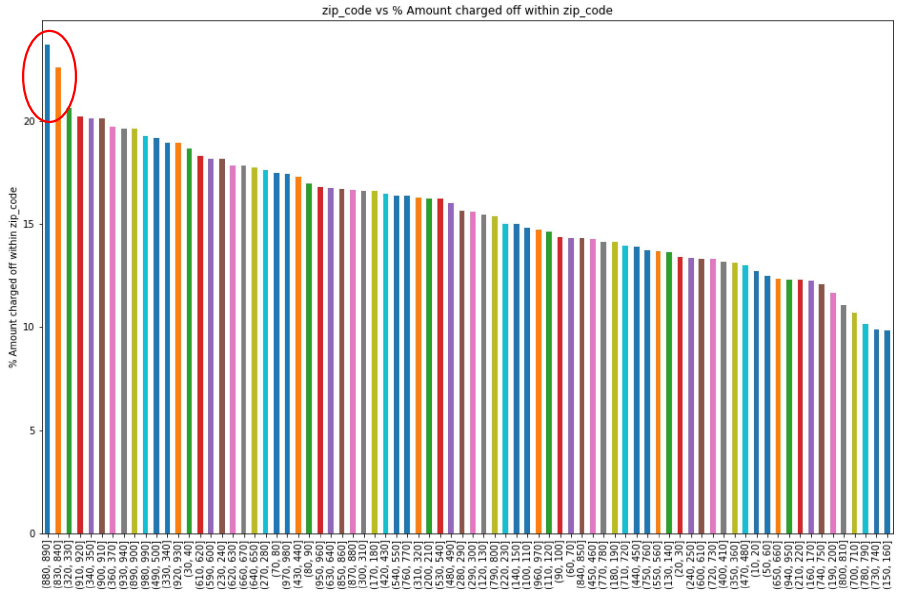
Empl_Title needs highest caution

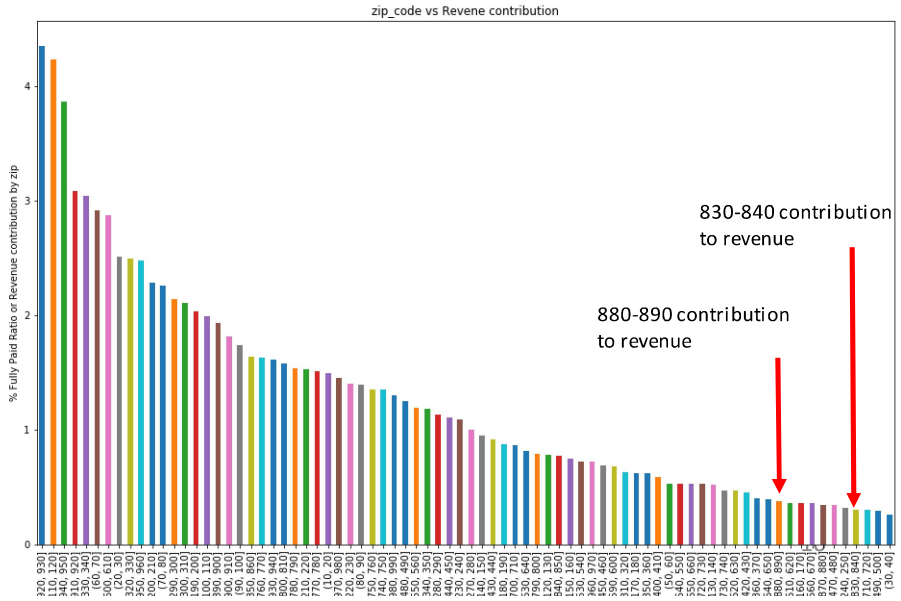
Empl_Title needs moderate caution

Empl_Title needs low caution

Area/Locality connection with defaulters

- The barplot shows percentage amount charged off within particular zip_code
- $\text{Perc\_amnt\_chargedoff} = \frac{amount\ charged\ off\ within\ zip\ code}{total\ loan\ amount\ assigned\ to\ zip\ code(chared\ off+fully\ paid)}$
- 1st Barplot shows zip_codes with higher charged off rates, these zip codes can be checked in 2nd barplot, their contribution to overall revenue.
- For ex, (880-890) & (830-840) have higher rate of charged off amount. Also, from 2nd graph can be seen, these localities contribute to revenue by small percentage.
- Recommendation: This shows, necessary action can be taken easily with these zip_codes, without hampering business. So more stringent rules can be applied for these localities.
- Additional research can be done by getting different datas, relating to these zip_codes. Such as poverty rate, fraud, crime etc....



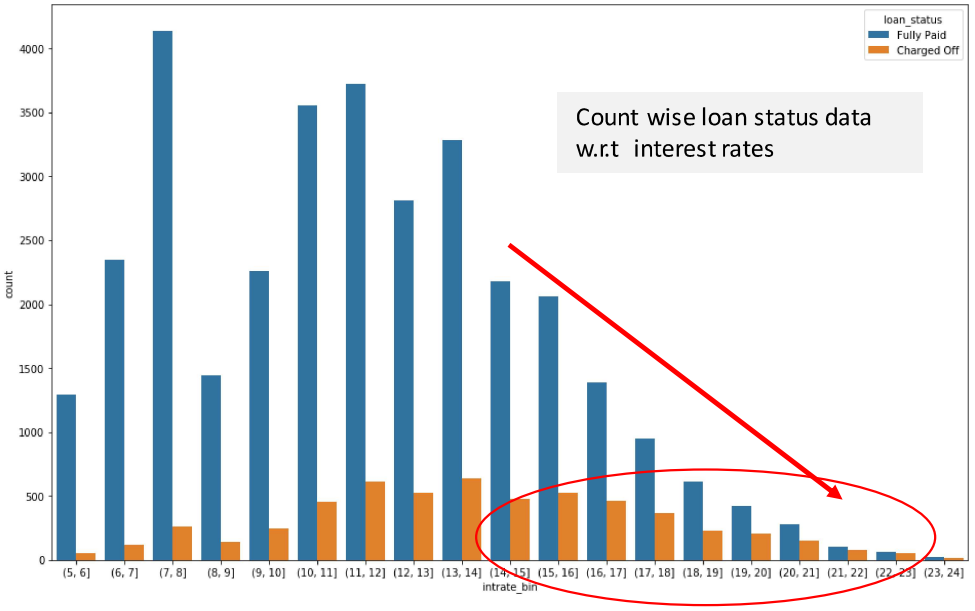


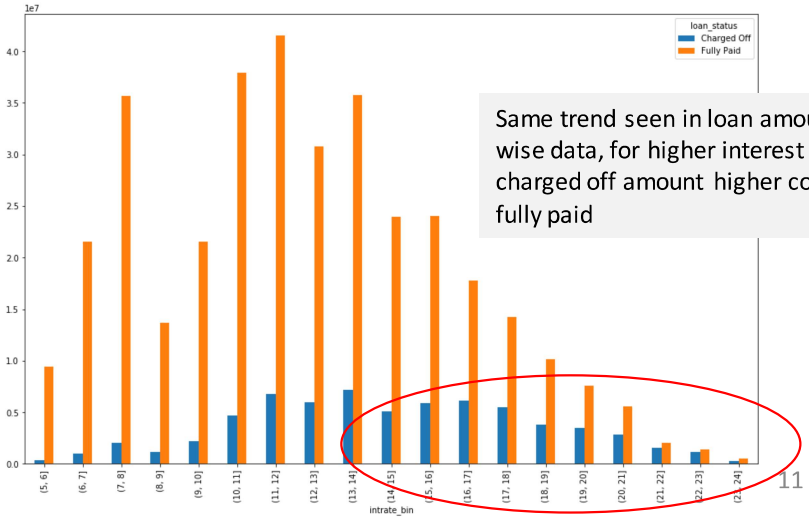
Interest rate effect

- 1st plot shows count-wise charged off fully paid loans against Interest rates
- As seen for lower interest rates **upto 14 %**, charged offs/defaulters **increase steadily** with overall count
- For higher interest rates **beyond 15%**, **higher defaulters** can be seen compared to fully paid.

- Recommendation:

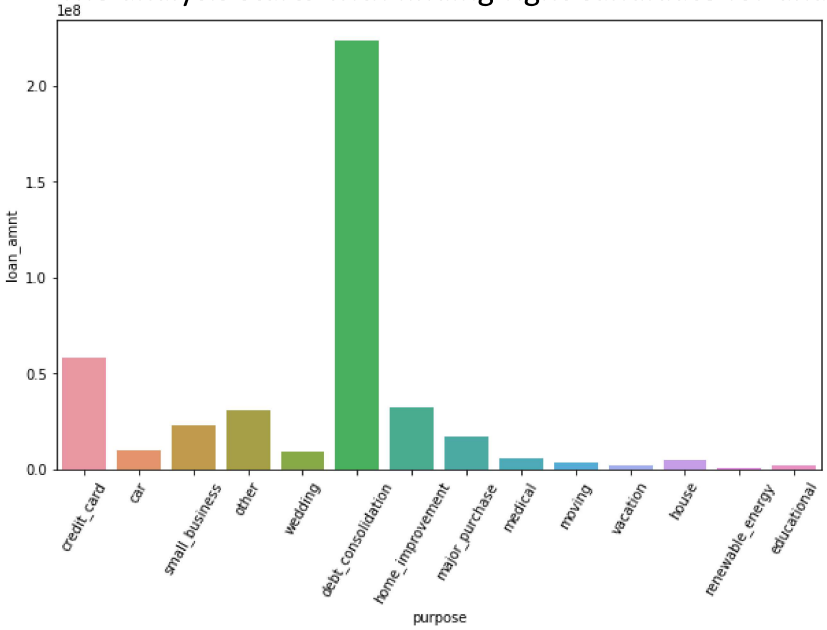
For higher interest rates, installments would be higher for same tenure. So additional checks are recommended such as annual income & debt to income ratio or previous loans



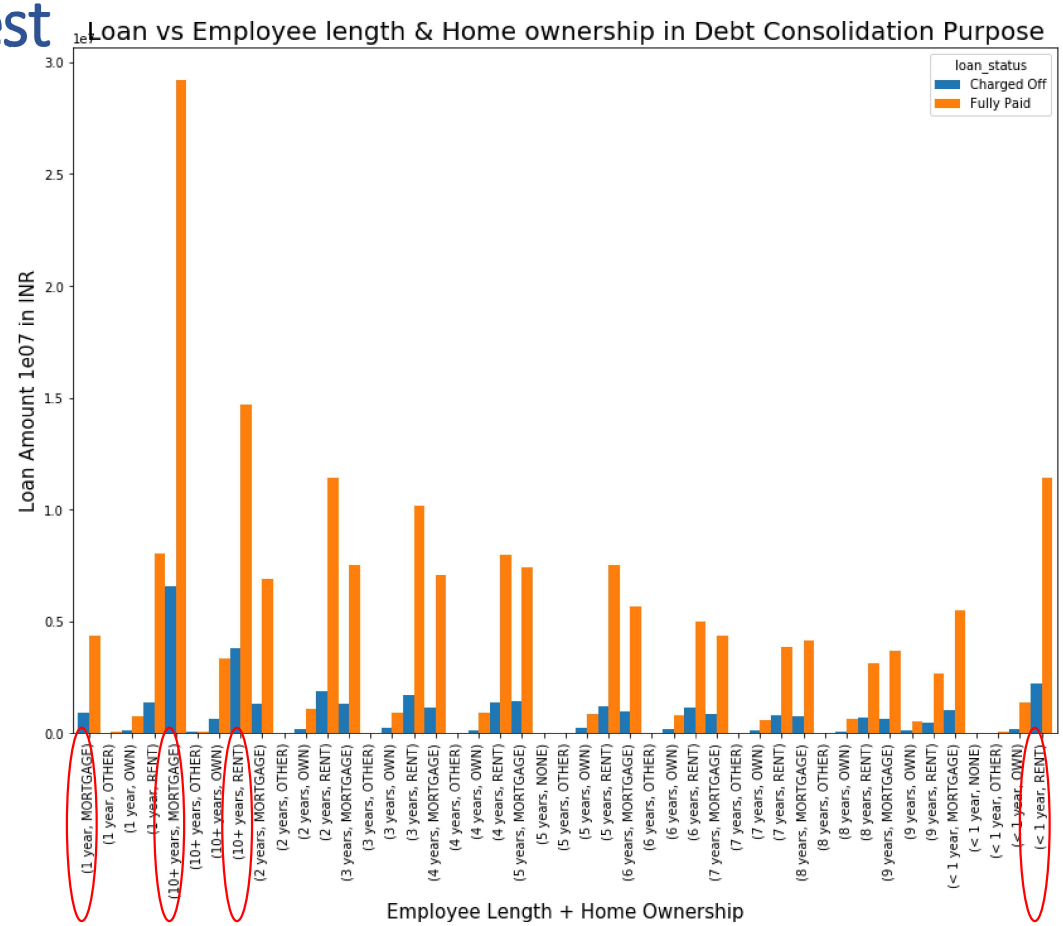


Common purpose for loan request

The analysis starts with finding right candidate for analysis



Highest loan requisition with purpose "Debt consolidation" is explored further with employee details



Applicants with home mortgage or home rent are seen with charged off loans. On next slide we will focus on <1year & 10+year experienced applicants

Applicant income vs installment

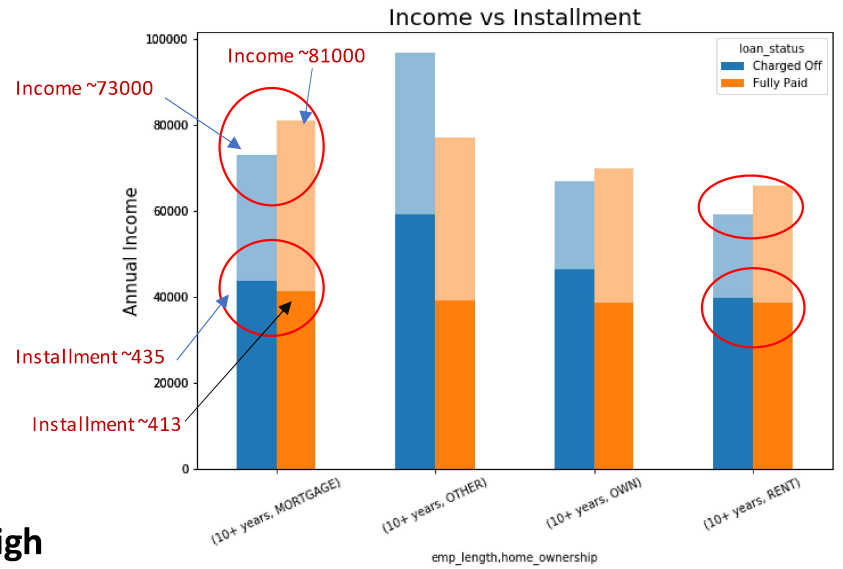
Continuing from last slide, income vs installment for both <1year & 10+year are shown besides.

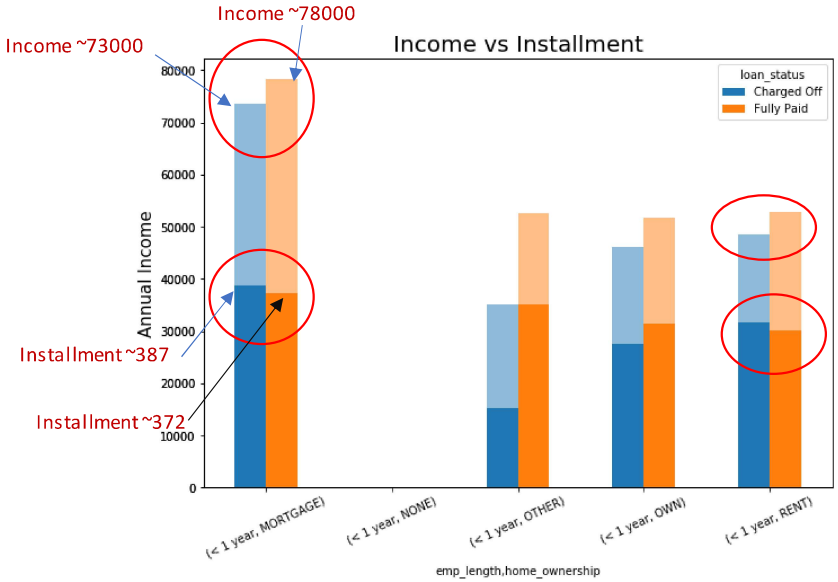
Installment (opaque) is imposed on Income(transparent)
Installment is multiplied by factor 100

These 2 section in applicants have more charged offs comparatively.

In both plots applicants from "Charged Off" section have high installment with low income compared to "Fully Paid"

Recommendation:
Installments can be adjusted for comparatively low income applicants





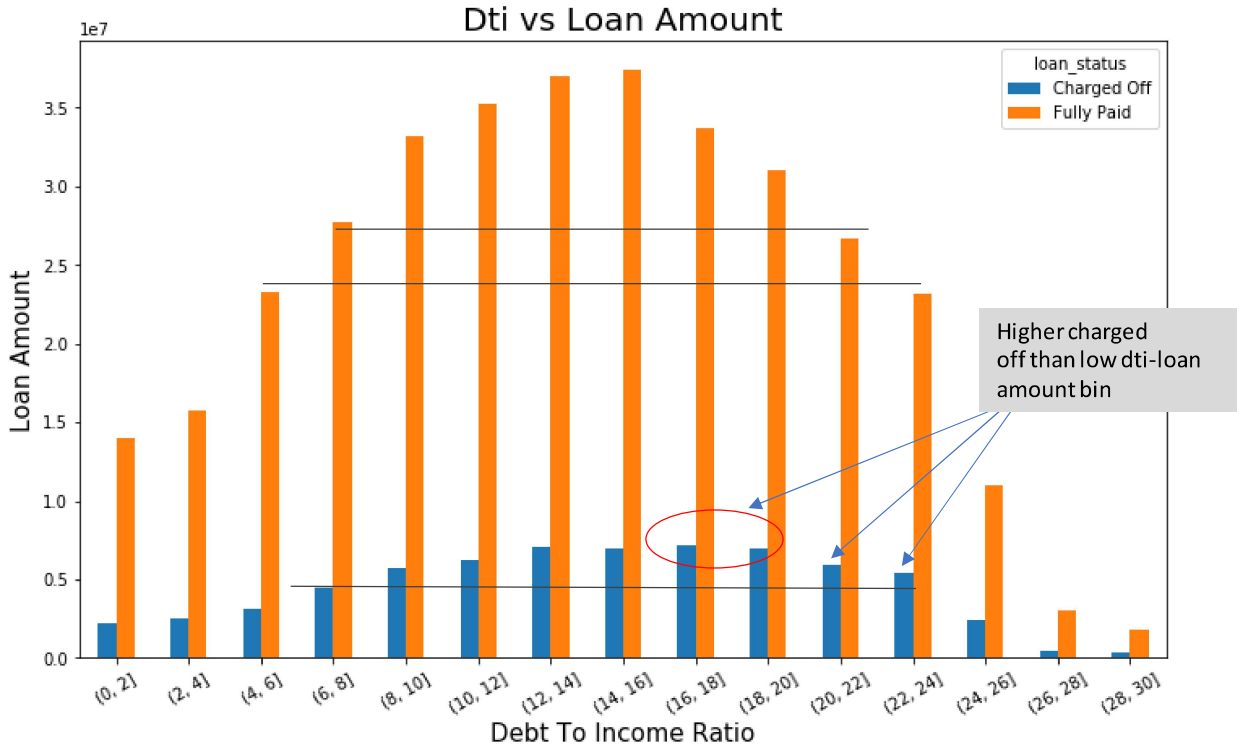
Debt to Income ratio consideration

As DTI increases, until value 16, charged off amount increases steadily with fully paid amount.

After 16, "charged of amount" stake becomes higher comparatively than in lower dti ratio.

Recommendation:
For applicants with dti ratio more than 16, additional checks or more loan prerequisite rules can be added

Standards "dti" limits to be set, after loan funding dti shall be in standard limits.
Loan scheme to be decided by dti.



Summary/Conclusion

The loan repayment ability has certain connection to applicant's professional, personal, demographic background

To some extent co-relation can be found on applicants loan paying probability with applicant background and loan attributes. Some co-relations such as

1. Probability of applicant likely to repay loan in case annual income above 82000/- with installment as 410/- is higher than with less income & higher installment 410/-
2. Applicants with debt to income ratio less than 16 perform better in term repaying loan than with dti higher than 16.
3. The loan terms I.e. interest rate also influenced the repayment ability. Interest rates higher than 15% have less probability of repayment. Maybe higher interest rates linked to higher installments and causing 3rd point.
4. Some localities are seen to have less repayment ratio. E.g. (880-890) & (830-840) zip code have more defaulters than fully paid applicants.
5. Some firms applicants, like "wal-mart", "postal service" have higher loss making customers than usual.

These are some correlations, which can be used to put additional checks such dti before loan & after loan to keep dti within limit, or making specific loan schemes to certain neighborhoods(zip-codes) or consider guarantors within same firm to reduce defaulters