LENDING CLUB CASE STUDY

EXPLORATORY DATA ANALYSIS

Group Facilitator:
Vishal Singh

Team Member:
Prathippa

PROBLEM STATEMENT

- The data given below contains information about past loan applicants and whether they 'defaulted' or not. The aim is to identify patterns that indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of the loan, lending (to risky applicants) at a higher interest rate, etc.

- Use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

- The company needs this to understand the **driving factors (or driver variables)** behind loan default, i.e. the variables that are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

ANALYSIS APPROACH

- Perform Data Pre-processing
  - Read CSV Data
  - Check and remove any header/footer or total/subtotals in the data
  - As we're focusing only on default factors, we can remove data where loan_status is 'Current' as it has incomplete info
  - Drop columns where we have missing values in all the rows to reduce clutter
  - Check the number of unique values in all the fields.
  - Drop columns that have the same value in all the rows as they do not give any additional information.
  - Review remaining columns to see if they make business sense to keep them for analysis else drop them
  - Check if the data type of the remaining columns is as per requirement. If not, then update the data type
  - Now check for missing values and perform treatment as required
  - Perform Outlier Treatment for Float & Integer Columns
  - Check to see if it makes sense to convert any numeric column into Categorical columns by binning/grouping
  - Classify the columns as categorical, numerical or Others(fields which are verbose or don't need to be analyzed)
- Data Analysis
  - Perform Univariate, Segmented Univariate, Bi-Variate & Multivariate Analysis
  - State observation & Insights

DATA REDUCTION

- Total Records in CSV File: 39, 717
- Remove records where loan_status is 'Current'. (Less: 1,140) i.e. 2.87% records
- Records Analyzed: (39,717 – 1,140) = 38,577

	Freq	Pct
Fully Paid	32950	82.96
Charged Off	5627	14.17
Current	1140	2.87

- Total Columns in CSV File: 111
- Drop columns with all values as missing. (Less 55)
- Drop columns that have the same value in all the rows: (Less 11)
- Columns that can be dropped as they don't make business sense to analyze: (Less 23)
- Columns Analyzed: (111 – 55 – 11 – 23) = 22.

UNIVARIATE ANALYSIS

- **Term:** Shorter term loans are more popular than longer term loans
- **Verification_Status:** Applications with verification_status have been given highest number of loans
- **Home_Ownership:** Highest number of loans approved are for applicants living on Rent
- **Grade:** Interestingly, Grade B loans are higher than Grade A (why?) while rest of the loans reduce with grade.
- **emp_length:** Maximum loans have been disbursed to applicants with emp_length > 10 years of <=1 year.
- **purpose:** 'Debt Consolidation' is the largest reason for loan
- **issue_d_month:** Number of loans issued increase with increase in month number except Feb which has least number of loans probably because it has least number of days
- **issue_d_year:** Number of loans seem to be increasing on an annual basis. Most likely due to company expansion

SEGMENTED UNIVARIATE (1/4)

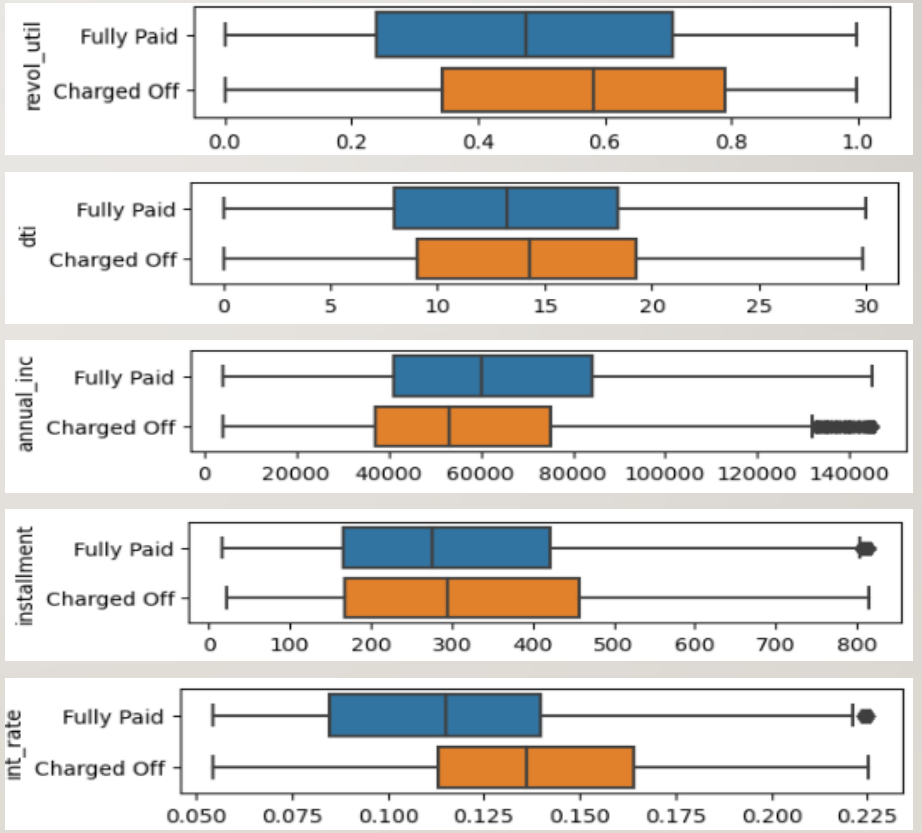
Charged_off loans tend to have higher revol_util

DTI of Charged_off loans seems to be higher

Annual_inc of Charged_off loans seems to be lower

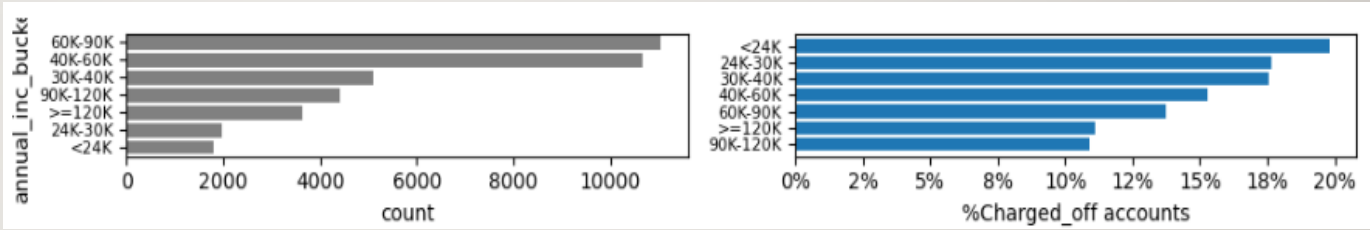
Installment amount tends to be higher for Charged_off loans

Interest rate for Charged_off loans is significantly higher

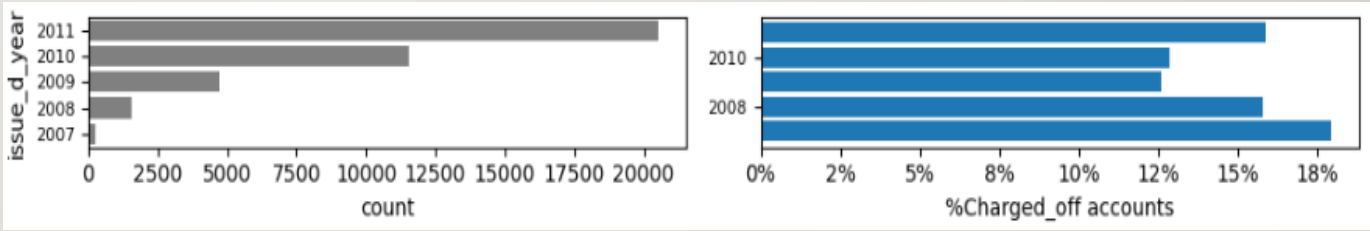


SEGMENTED UNIVARIATE (2/4)

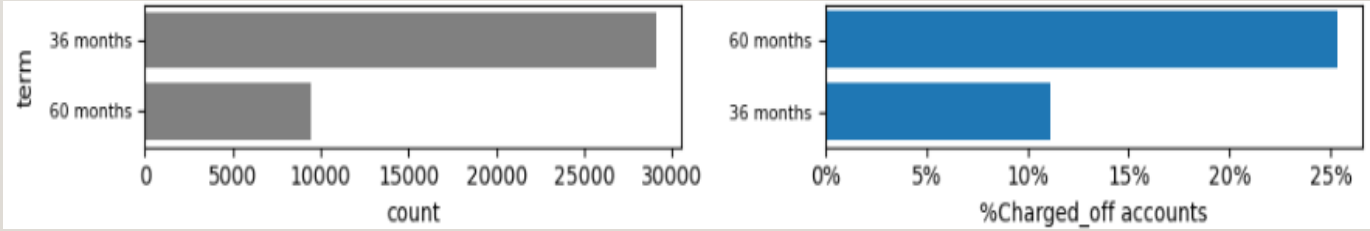
income_buckets: Medium Income buckets have highest number of loans. Higher income buckets have lowest charged_off rates compared to lower income buckets



issue_d_year: While loans issued in 2007 were the least, their charge_off rate is highest.

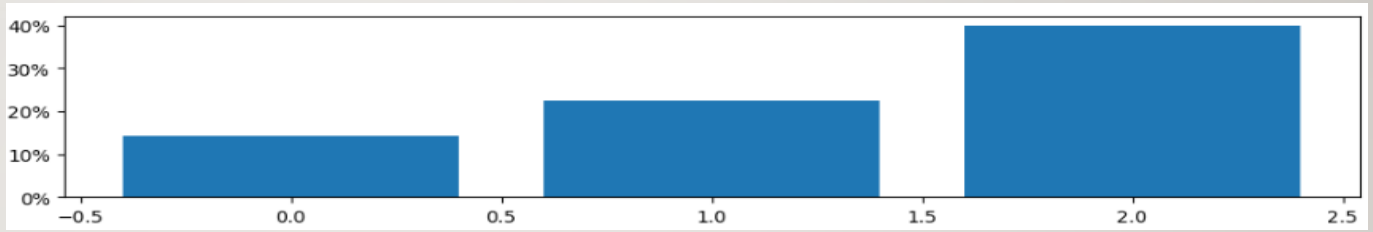


Term: While loans disbursed for 36 months term is higher, loans for 60 months term have significantly higher percentage of charged_off Loans

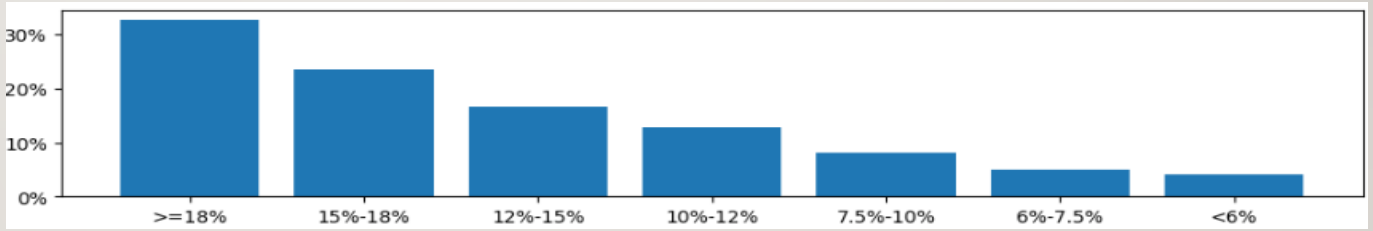


SEGMENTED UNIVARIATE (3/4)

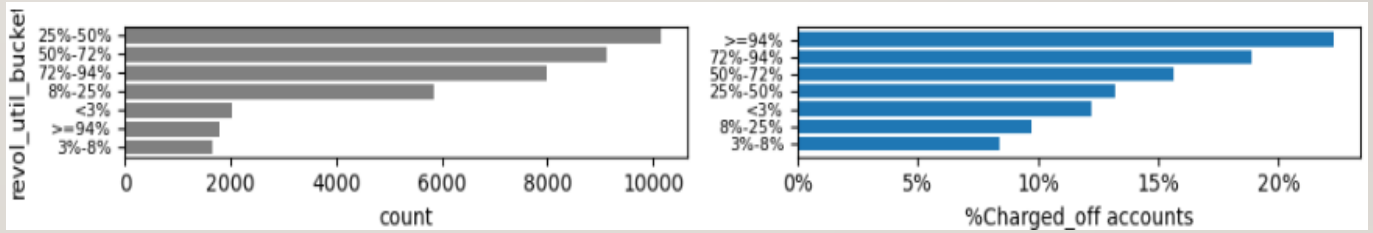
pub_rec_bankruptcies: Higher the number of pub_rec_bankruptcies, higher the Charge_off rates



int_rates: Higher the interest rates, higher the Charge_off rates

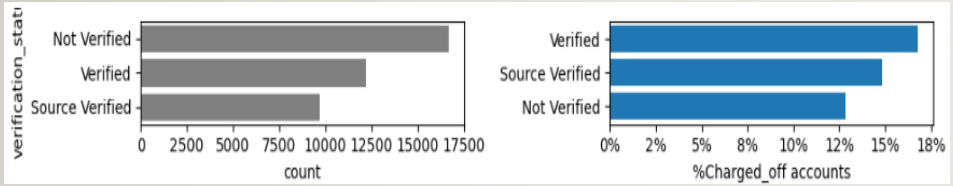


revol_util_buckets: Higher than average charged_off rates are seen in accounts with > 50% revol_util value

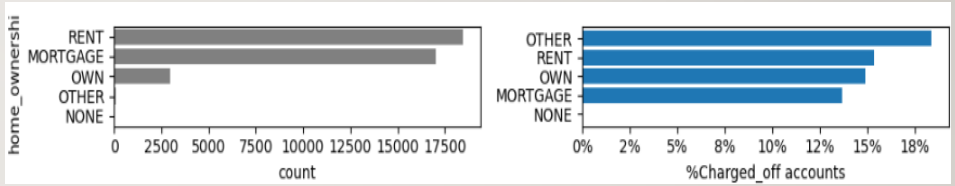


SEGMENTED UNIVARIATE (4/4)

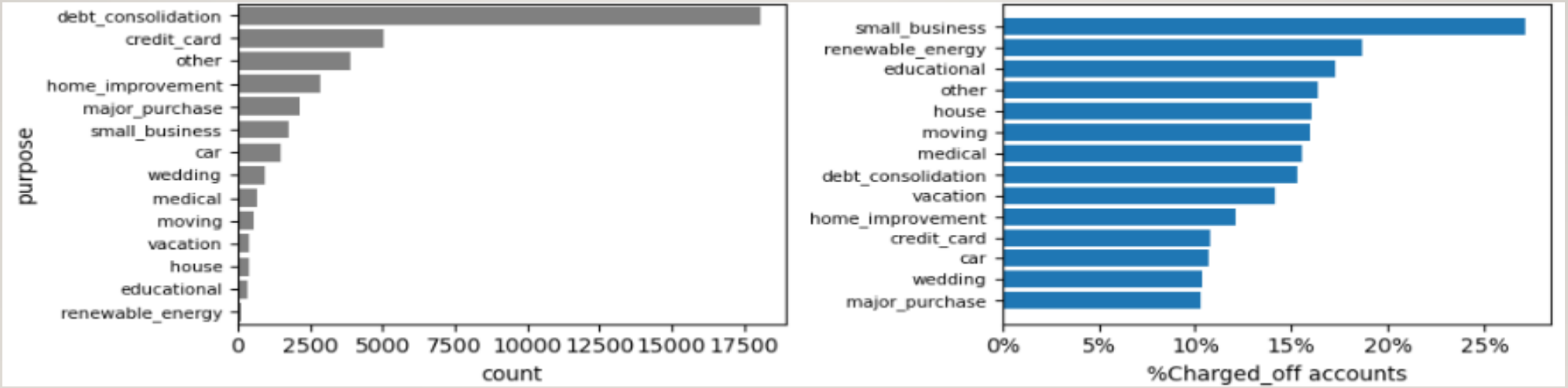
Verification Status: Accounts with a Verification status as Verified have the highest %Charged_off rate



Home_Ownership: Highest loans issued are for those who live on rent but charge off is highest for 'Others'



Purpose: Though Debt Consolidation has maximum loans, Small Business tends to have the highest Charge_off_rates



BI-VARIATE ANALYSIS (1/5)

term	36 months	60 months
issue_d_year		
2007	251	0
2008	1562	0
2009	4716	0
2010	8466	3066
2011	14101	6415

term	36 months	60 months
issue_d_year		
2007	17.93	NaN
2008	15.81	NaN
2009	12.60	NaN
2010	9.95	20.97
2011	10.63	27.39

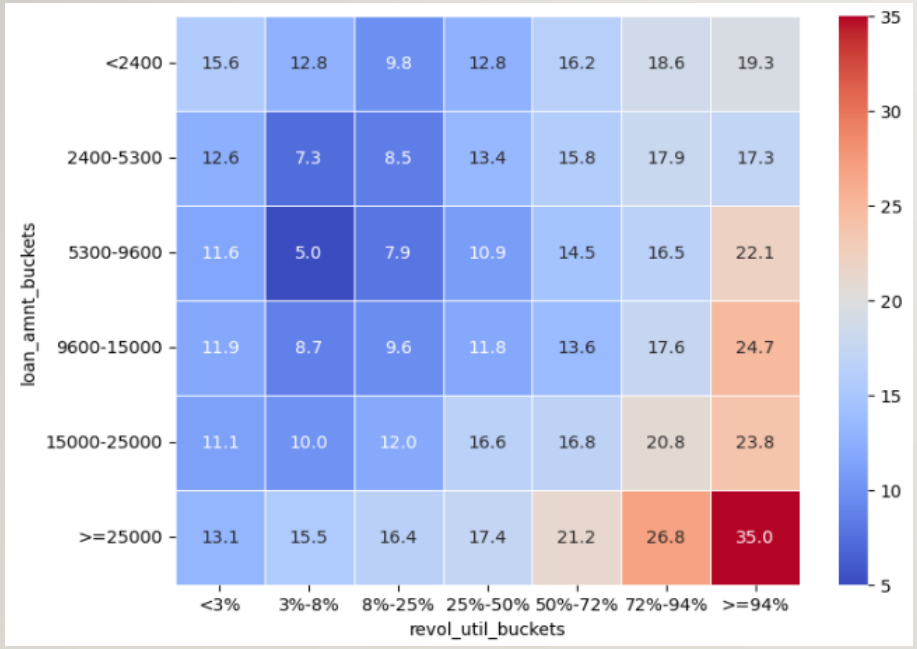
- Loans with term as 60-month started only in 2010
- Charge off rate for tenure 60-month is significantly higher than 36-month loans
- Charge-off rate for 36-months tenure has been constantly dropping except for 2011

term	36 months	60 months
loan_amnt_buckets		
<2400	93.58%	6.42%
2400-5300	88.12%	11.88%
5300-9600	84.58%	15.42%
9600-15000	74.98%	25.02%
15000-25000	60.03%	39.97%
>=25000	40.84%	59.16%

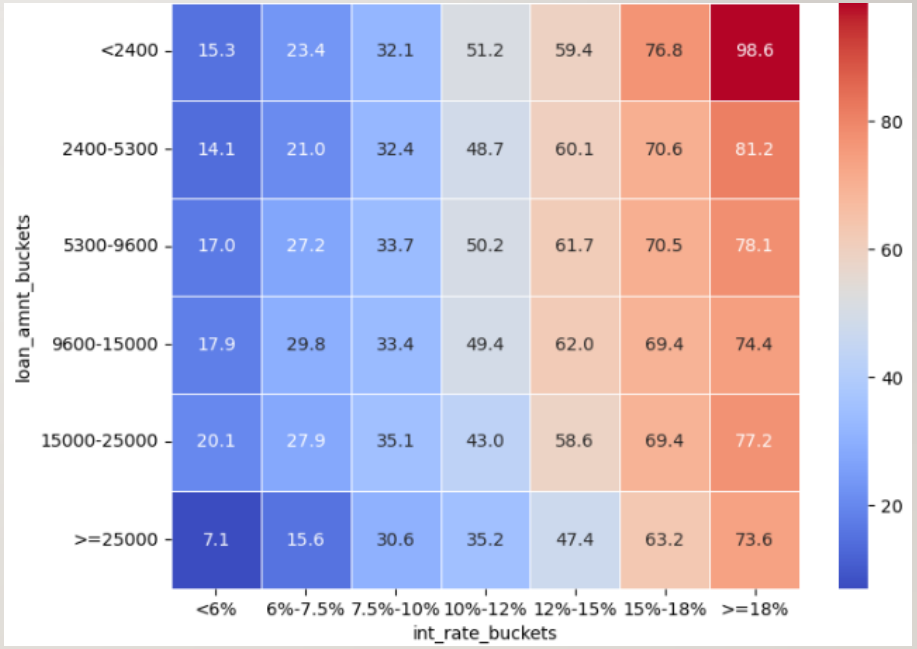
term	36 months	60 months
loan_amnt_buckets		
<2400	13.82	29.75
2400-5300	12.09	25.65
5300-9600	10.45	24.11
9600-15000	10.03	24.01
15000-25000	10.84	25.98
>=25000	12.68	26.44

- As the loan_amount increases, the proportion of loans for the term of 60-month increases.
- While the Charge_Off percent for 60-month tenure > 30-month tenure, the trend with respect to loan_amount buckets seems to be the same

BI-VARIATE ANALYSIS (2/5)

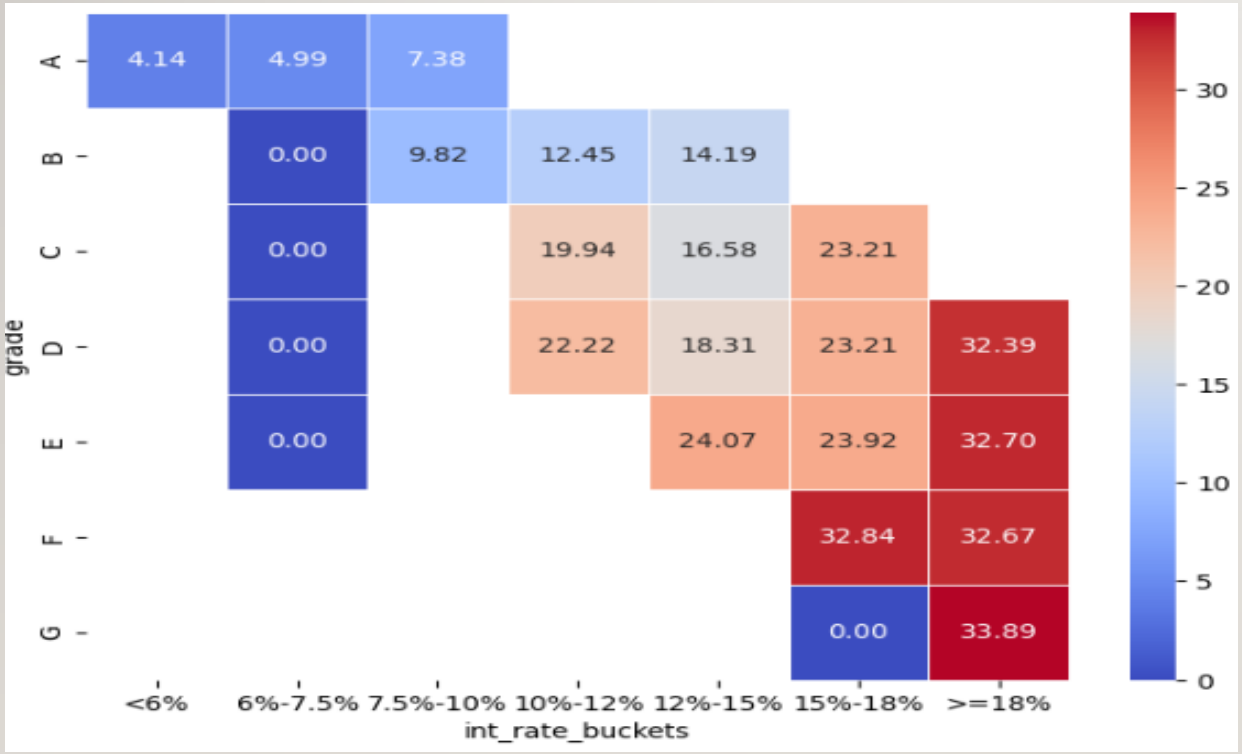


As the loan_amnt increases and revol_util increase, proportion of charge_off increases



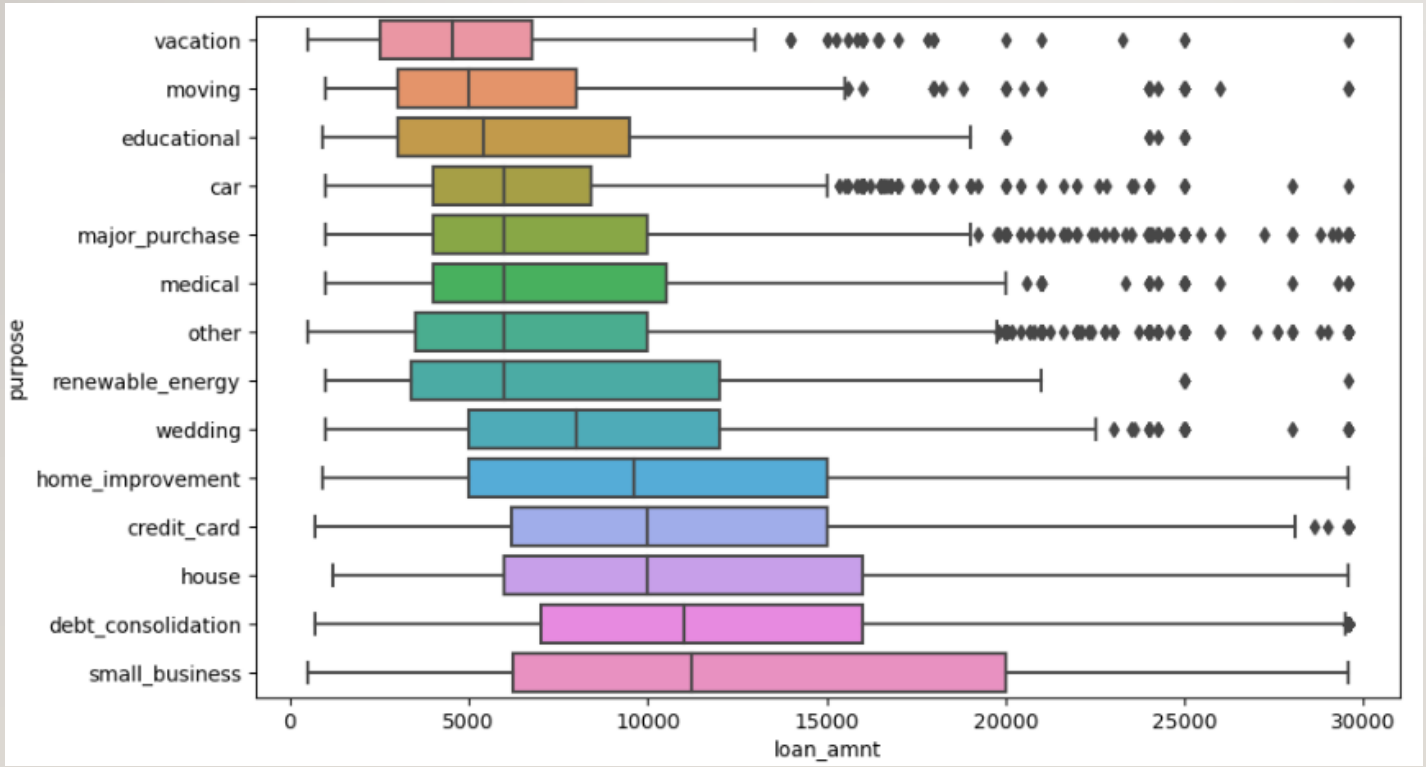
Higher chage off rate is seen as interest rate increases and loan amount decreases

BI-VARIATE ANALYSIS (3/5)



- **Grade of the loan corresponds to the interest rates.**
- **Higher charge off rates are observed as the grades declines and int_rates increase**

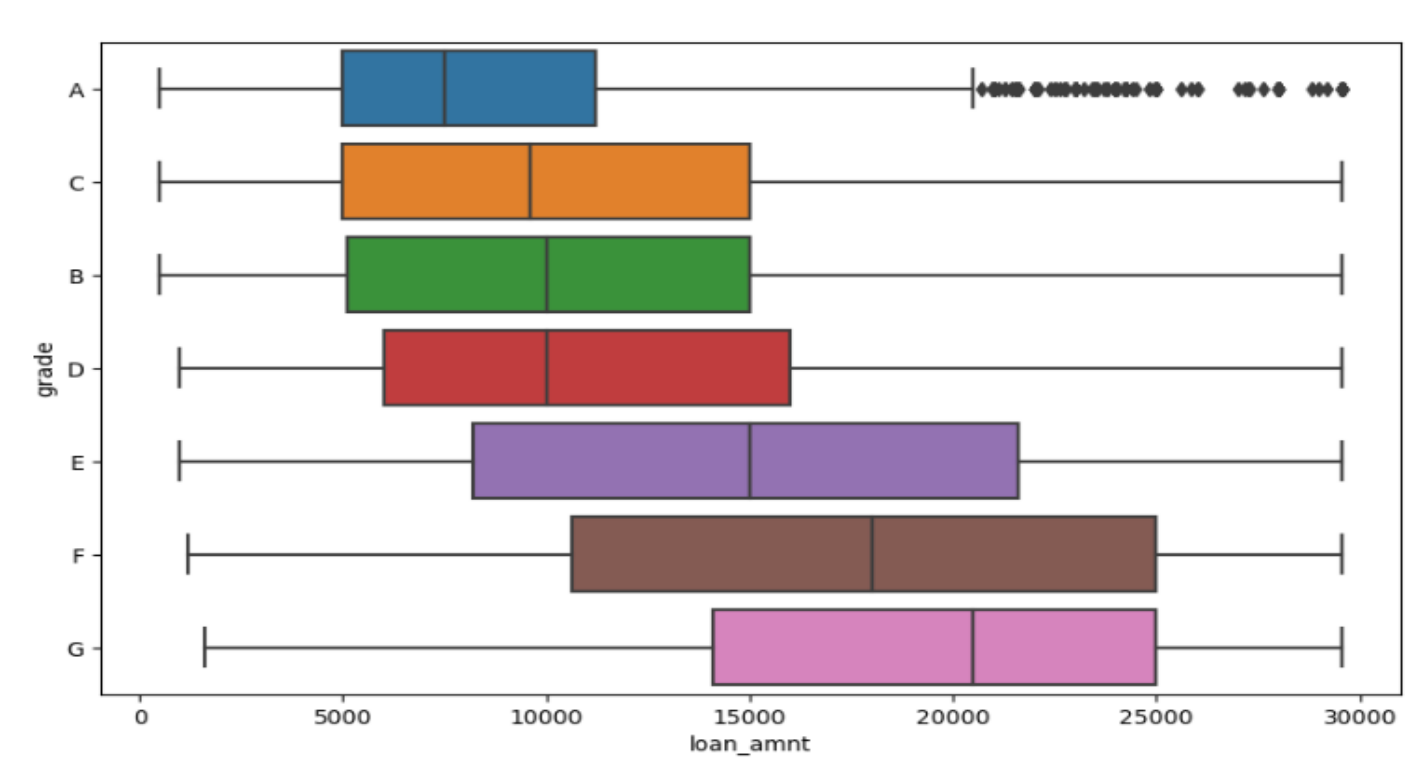
BI-VARIATE ANALYSIS (4/5)



purpose:

Small_Business and Debt_Consolidation seem to have the highest median loan amount disbursed

BI-VARIATE ANALYSIS (5/5)



grade:

Both interest rate and median loan_amnt tend to increase as grade deteriorates from A to G.

SUMMARY

- Factors critical to look at:
  - Loan-Attributes: [Loan_Amnt, Term, Int_Rate, grade, purpose]
  - Customer_Attributes:[pub_rec_bankrupticies, revol_util, dti, home_ownership, annual_income]
- Giving loans for Small Business, Renewable Energy & Education especially for 60-month term is riskier. Might want to look for lower loan amount and lower term
- Giving loans where revol_util is high is risky, especially for larger loans.
- Home_Ownership: The highest loans issued are for those who live on rent but charge off is highest for 'Others'. Need to be able to classify 'Others' better
- Higher charge-off rates are observed as the grade declines and int_rates increase
- Higher charge-off rates are seen among loans where interest rate increases and loan amount decreases. One might want to avoid such loans. People accepting high-interest small-ticket loans are riskier.