Lending Club Case Study

Group : Sagar Thacker
Bhaskar Nag
Ansuman Gangapuram
Nitin Shukla

Lending Club Analysis Overview

- Identifying the Business Problem
- Data Description
- Data Preparation and Cleaning
- Data Classification and Analysis
  - Univariate Analysis - Unordered Categorical
  - Univariate Analysis – Ordered Categorical
  - Univariate Analysis – Quantitative
  - Segmented Analysis
  - Derived Analysis
  - Bivariate Analysis
- Conclusion

IDENTIFYING THE BUSINESS PROBLEM

- Lending Club is a Consumer Finance Company which specializes in lending various types of loans to urban costumers.

- Lending Club's Model for risk assessment categorizes borrowers by assigning them a grade and subgrade based on their credit history.

- Investors are presented with a list of borrowers, along with their assigned risk assessment grades, and they have the opportunity to choose which borrowers they will fund, and the percentage of funding they will cover.

- Our business problem is to provide Lending club with a more comprehensive assessment of these borrowers in order to make a smart business decision by identifying new borrowers who will default on their loans and which will not.

Data Description

- Lending club provides us with historical data. The Dataset contained information pertaining to borrower's past credit history and Lending club loan information. The total Dataset consisted of 39717 records, which was sufficient for our team to conduct Exploratory Analysis on the Data.

- Variables present in the Dataset provided ample information which we could use to identify relationships and gauge their effect on the success or failure of borrower fulfilling their loan commitments.

- We required only the variables that had a direct or indirect response to a borrower's potential to default. To achieve this we prepared the data by choosing select variables that would best fit this criteria.

Data Preparation and Cleaning

- We removed all columns that were completely empty eg: bc_util
- We removed all columns that had more than 60% of it's data missing.
- We also removed the description (desc) column as it did not contribute to our final goal.
- We removed columns irrelevant to our analysis like member_is, url etc.
- We removed all columns that contained just a single value in all of it's rows.
- We removed all missing values from the Dataset resulting in 8% data loss.
- We converted all columns having string values to lower case.
- We removed special characters like % to perform numerical operations on the columns.
- We Converted several columns to DateTime format.
- Some columns needed their type changes from string to float

DATA CLASSIFICATION AND ANALYSIS

Univariate Analysis : Unordered Categorical

We did Univariate analysis on the following Unordered Categorical variables:

Variable Name	Values	Most Frequent	Second
Term	36,60	36	60
Home Ownership	Rent,Own,Mortgage,Other	Rent	Mortgage
Verification Status	Verified, not verified, source	Not verified	verified
Loan status	Fully paid, charged off, current	Fully paid	Charged off
Loan Purpose	> 14	Debt consolidation	credit
Addr_state	=50	California	NY
Pub_rec	0,1,2,3,4	0	1

DATA CLASSIFICATION AND ANALYSIS

Univariate Analysis : Ordered Categorical

We did Univariate analysis on the following Ordered Categorical variables:

Variable Name	Values	Most Frequent	Second
Emp_length	1,2,3,4,5,6,7,8,9,10+	10+	2
Grade	A,B,C,D,E,F,G	B	A
Sub Grade	1-5 within Grades A-G	A4,B3	B5
Pub_rec_bankruptcies	0,1,2	0	1
Delinq_2_yrs	0 to 11	0	1
Inq_last_6mnths	0 to 8	0	1

DATA CLASSIFICATION AND ANALYSIS

Univariate Analysis : Quantitative

We did Univariate analysis on the following Quantitative variables:

Variable Name	Range	Median	Mean	75 Quartile
Loan Amount	500-35000	10,000	11307	15000
Funded_amount	500-35000	10,000	11032	15,000
Int_rate	5.42 – 24.40	11.86	12.07	14.65
Annual_inc	4000 – 600,000	60,000	69,305.22	83,000
DTI	0 – 29.99	13.54	13.44	18.69
Total_pymnt	33.73 – 58,563.68	10,101.59	12310.97	16700.72
Revol_bal	0 – 149588.00	9030	13465.66	17231
Revol_util	0 – 99.78	49.90	49.28	72.70

DATA CLASSIFICATION AND ANALYSIS

- **Univariate Analysis : Segmented (Loan Status = Charged Off)**

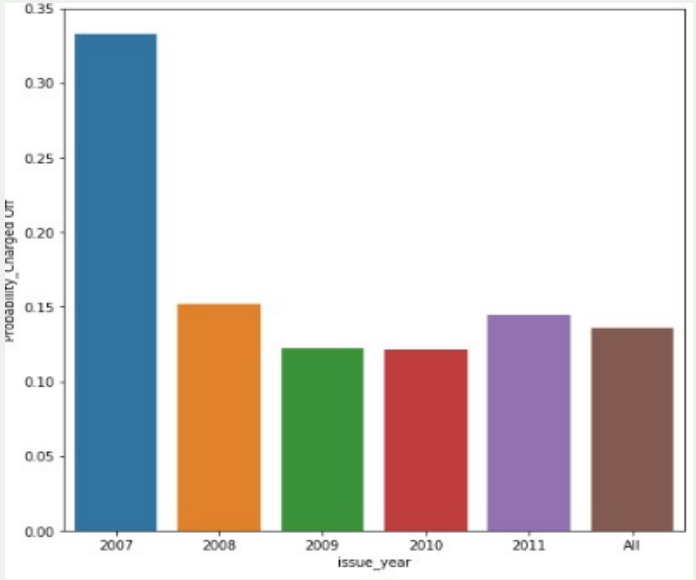
Variable	Inference (max along)	Variable	Inference (max along)	Variable	Inference (max along)
Term	60	**Home Ownership**	Rent	**Verification Status**	Not verified
Purpose	Small Business	**Addr_State**	Charged Off	**Pub_rec**	0
Emp_length	10+	**Grade**	B	**Sub_grade**	F5
Pub_rec_bankruptcies	0	**Delinq_2yrs**	8	**Inq_last6mnths**	6
Funded_amnt	Most at current	**Loan amount**	Max at current	**Int rate**	21.35 -24
Annual_inc	<2500	**Dti**	20-24	**Total_pymnt**	<=5000
Open_acc	38	**Revol_bal**	30-60k	**Revol_util**	10-20

DATA CLASSIFICATION AND ANALYSIS

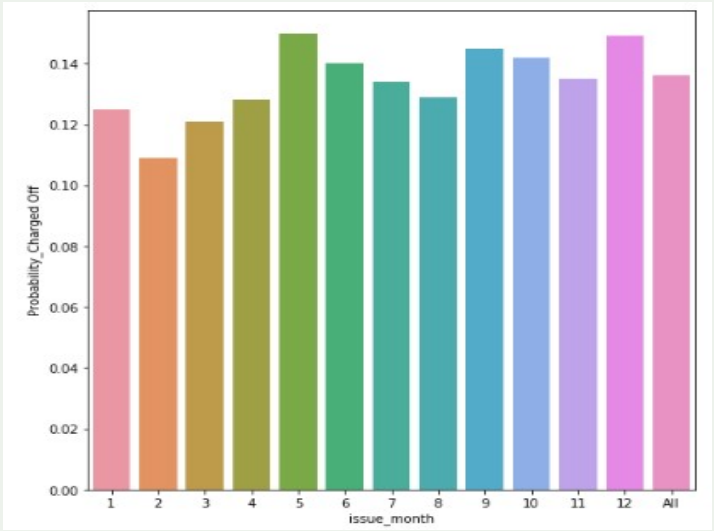
Univariate Analysis : Derived

We did Derived Metrices analysis on the following Derived variables:

Issue_year v/s Prob Chargeoff



issue_month v/s Prob Chargeoff



Inference : Probability Chargeoff is maximum for May 2017.

DATA CLASSIFICATION AND ANALYSIS

Bivariate Analysis :

- **Loan amount and purpose**

  Plots used :- Box Plot

  Inference :- Small business have more variation for charged off, followed by debt_consolidation

- **Loan amount and Interest Rate**

  Plots Used :- Line Plot

  Inference :- As loan_amnt and int_rate increases charged off rate also increases

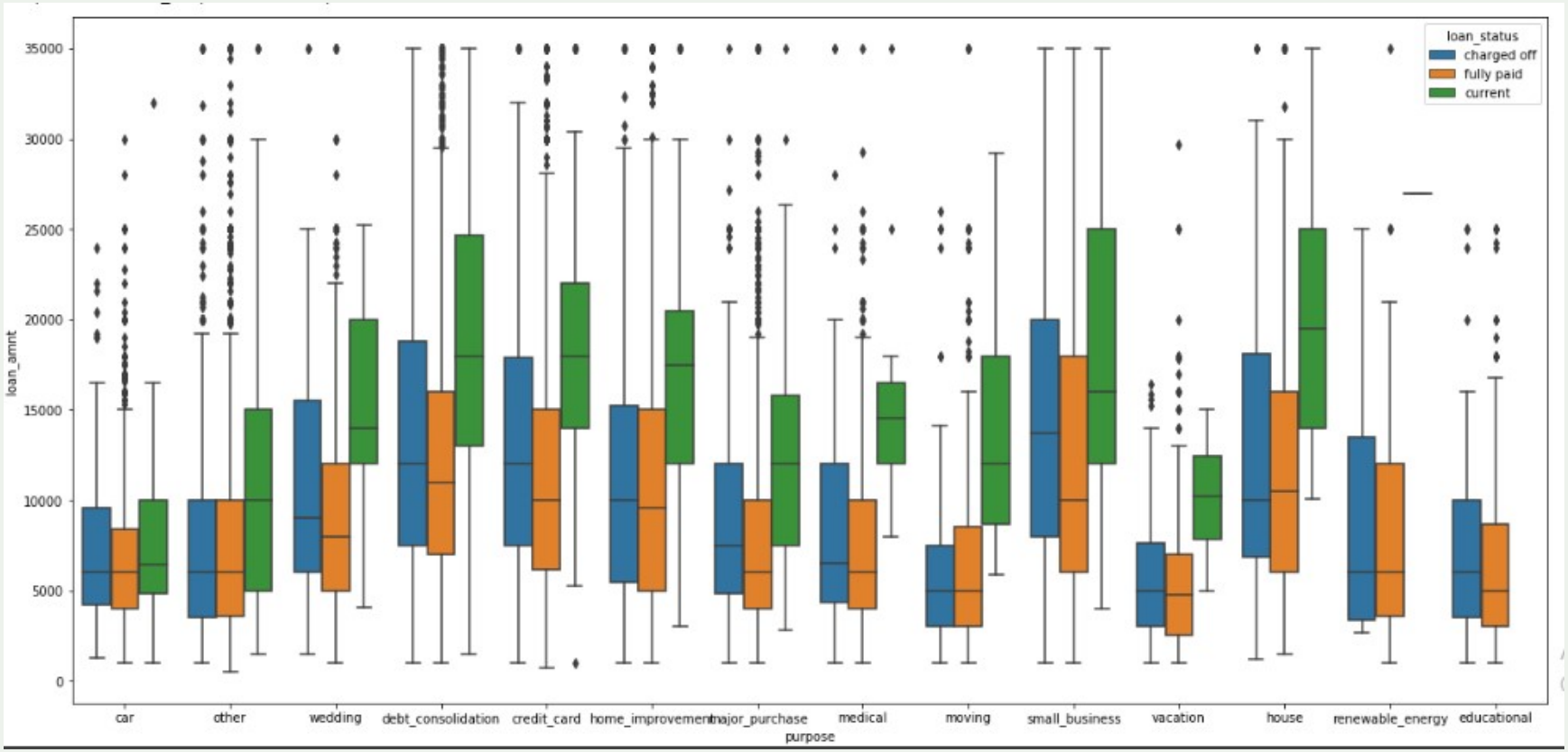
- **Int_rate and grade**

  Plots Used:- Boxplot

  Inference :- As grade increases int_rate also increases and along with the charged off also increases.

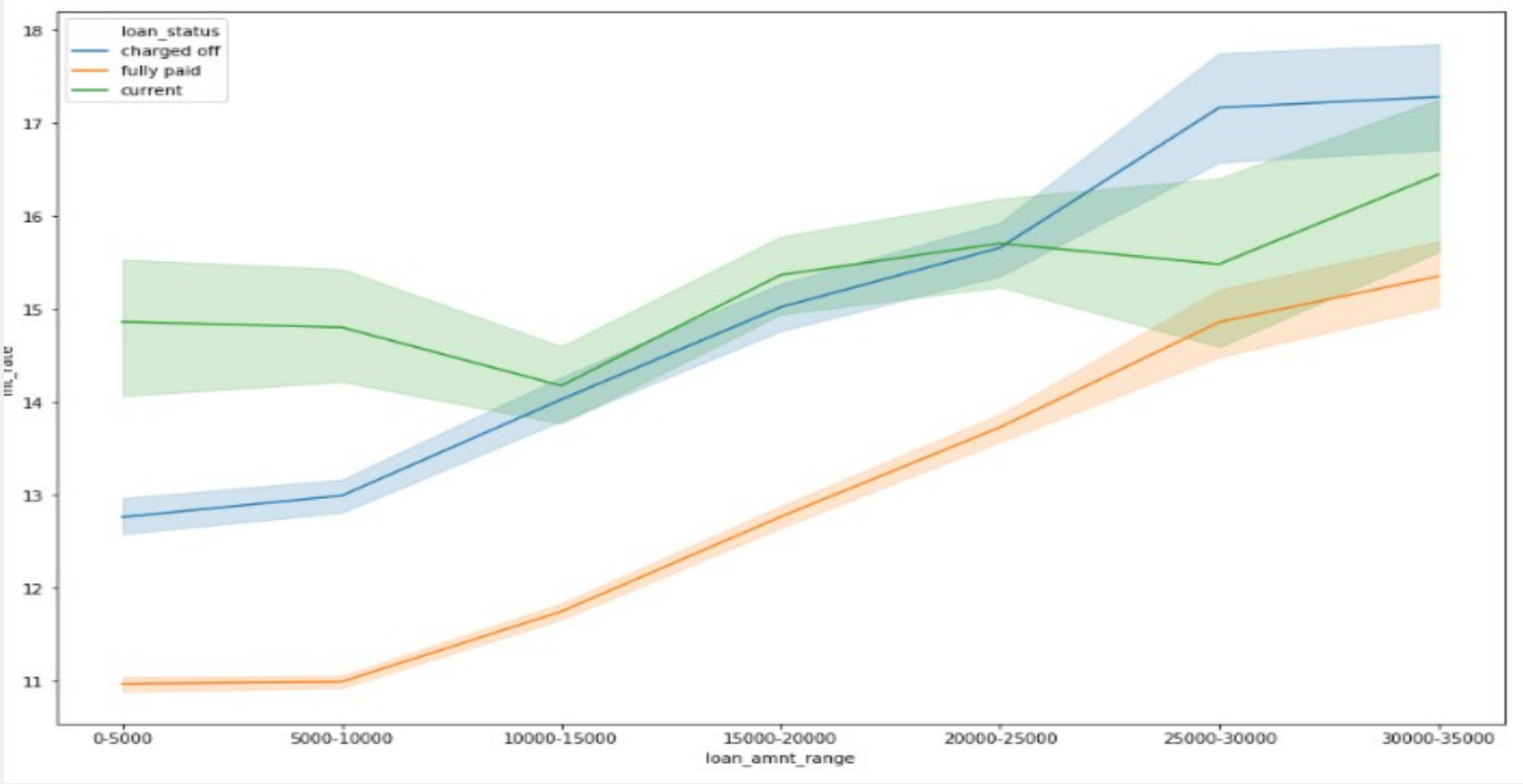
DATA CLASSIFICATION AND ANALYSIS

Loan_amnt vs purpose:
Small Business and Debt Consolidation contribute the most



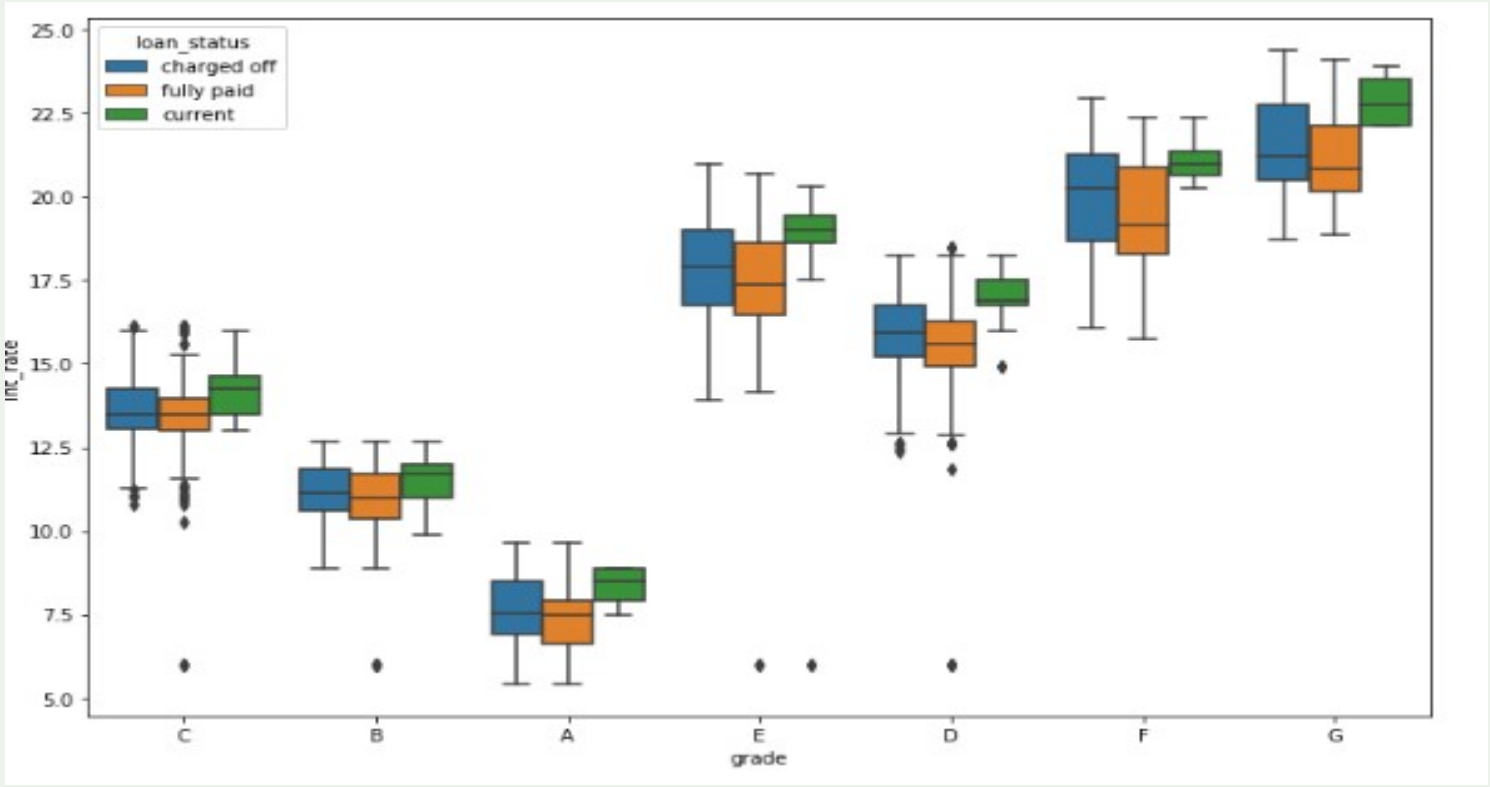
Loan_amnt vs int_rate

As the loan_amnt increases and int_rate increases, chances of getting default increases

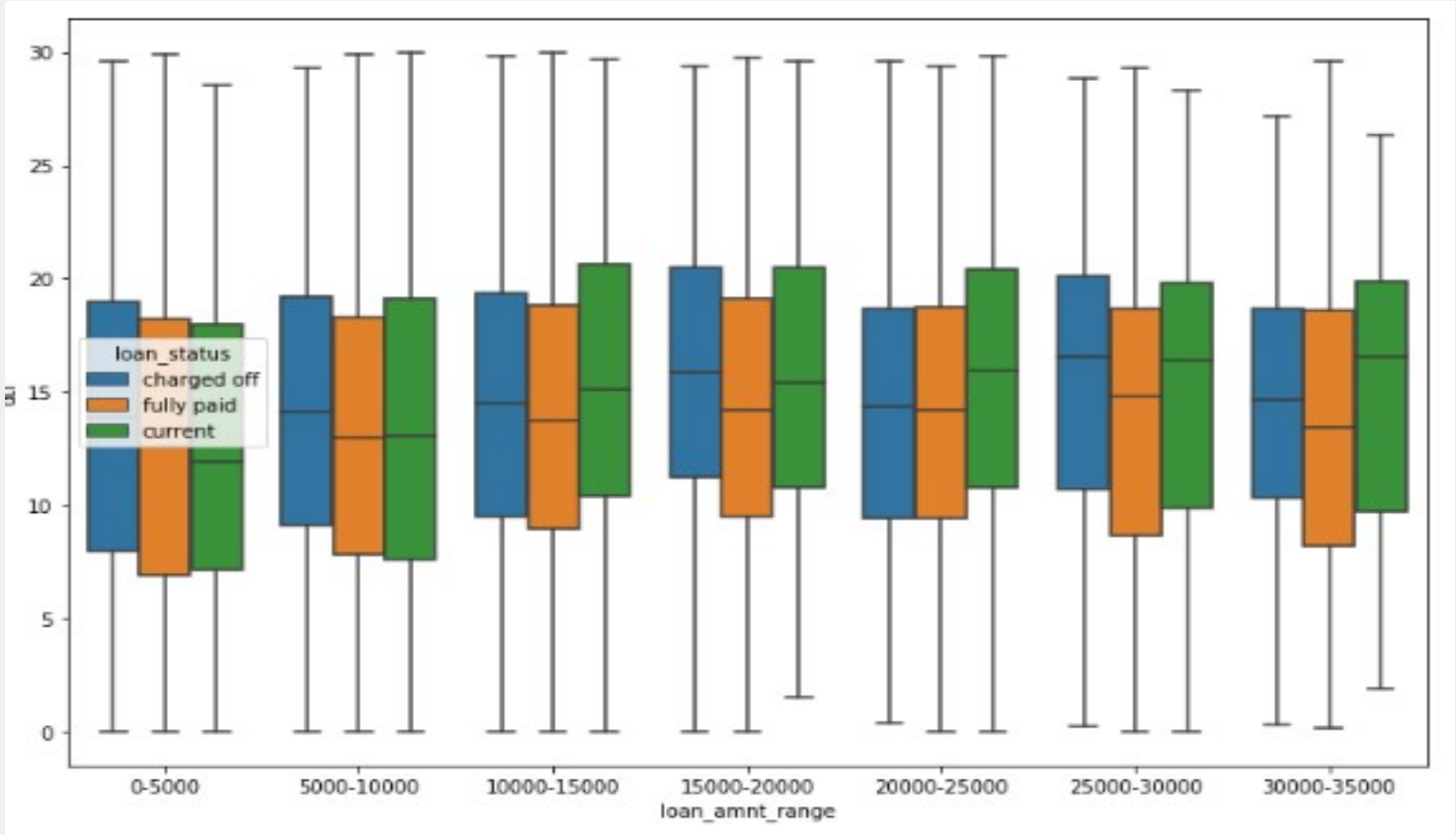


Grade vs int_rate

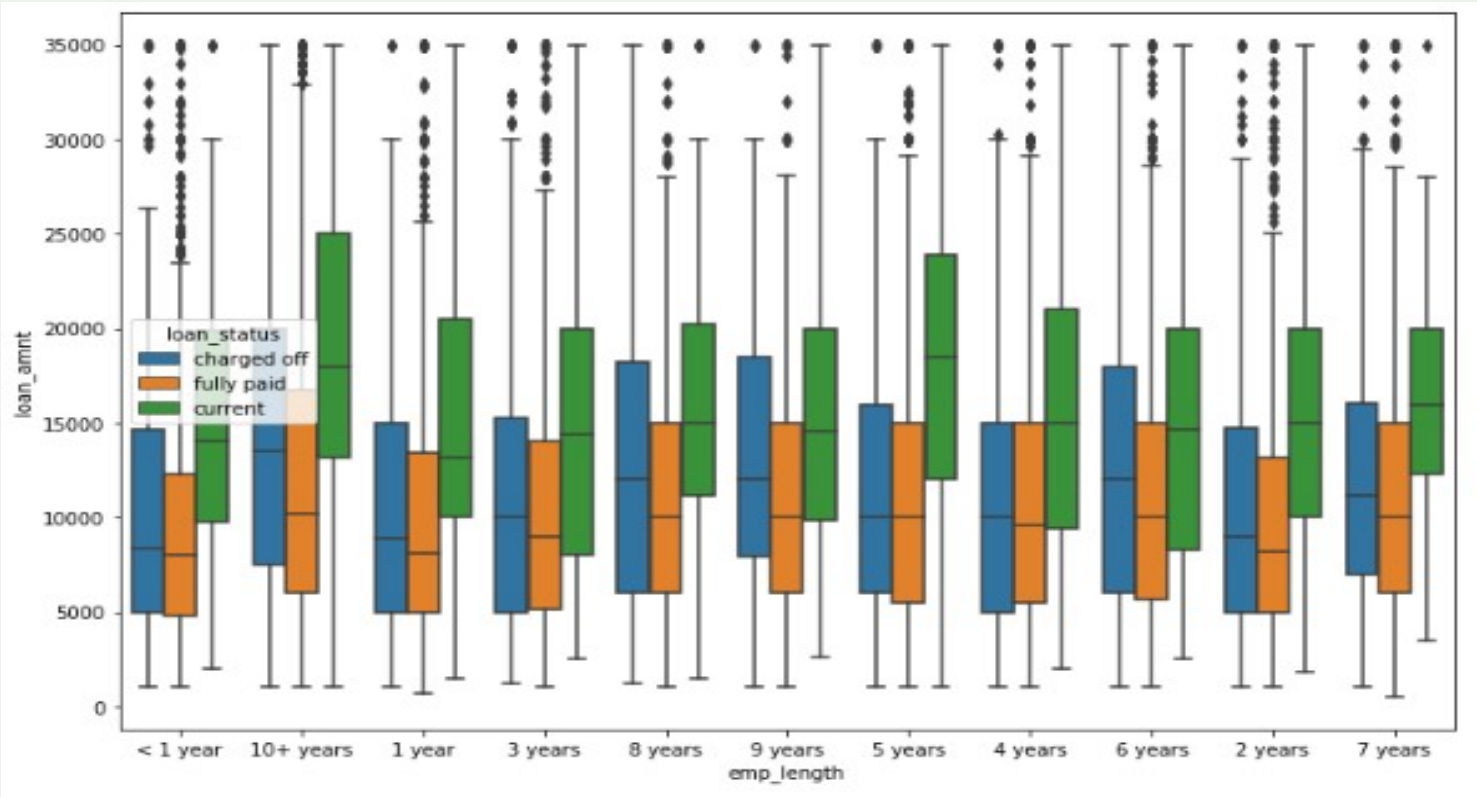
As the grade increases, with the increase in int_rate chances of getting charged off increases



Term vs loan_amnt
Applications have term 60 have higher chances of getting defaulted



Emp_length vs loan_amnt
10+ years, 5 years have a greater chance of getting defaulted



Conclusion

After performing Exploratory Data Analysis on the given Loan Dataset. We conclude that the below mentioned 5 features are most important for predicting the ability of a new borrower to default on his loan commitments.

- **Term**
- **Purpose**
- **Grade**
- **Employment length**
- **Interest rate**