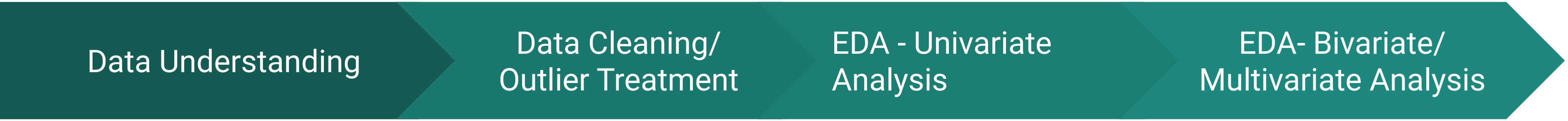
Lending Club Case Study

Vaibhav Holker || Victor Mohanty

Aug’23 batch

Process Flow



Data Understanding

Data Cleaning/ Outlier Treatment

EDA - Univariate Analysis

EDA- Bivariate/ Multivariate Analysis

Based on the Dataset and dictionary provided, get an understanding of the dataset, intuition based relationship between categorical and continuous variable

- Null Value Treatment
- Datatype Treatment
- Outlier Treatment
- Derived Metrics (Buckets to create categorical variables)

(Value Counts, Histograms, Outliers)

1. Understand the variation of output metric (Loan Status) across different categorical & continuous variables
2. Understand distributions of different variables

(Pie charts, Bar charts)

Based on trends from Univariate analysis, check trends in Output metric across multiple variables (2+ at a time) and identify trends as combination of 2 or more variables (Scatter plots, Stacked Bar charts, multi column charts)

Data Understanding and basic EDA

01 **Shape of the Dataset**

- There are 39717 rows with 111 columns with their definitions
- The Output Column i.e. Loan_Status has 3 values (Current, Fully Paid and Charged Off)

02 Level of Data & Unique Identifier

- Dataset has ID column that is unique ID assigned to a loan listing
- Each row can be identified basis this ID

03 **Key Categorical Variables**

- Loan Duration, Grade, SubGrade, Employment length, Home Ownership, Verification Status, Purpose, User Address State are the key major categorical variables that we must study in order to understand the variation of defaulters across these identifiers

04 **Key Numerical Variables**

- Loan Amount, Interest Rate, Annual Income, DTI, Open Accounts are some of the key continuous variables
- Some of these should be bucketed to get better understanding of outcome variable

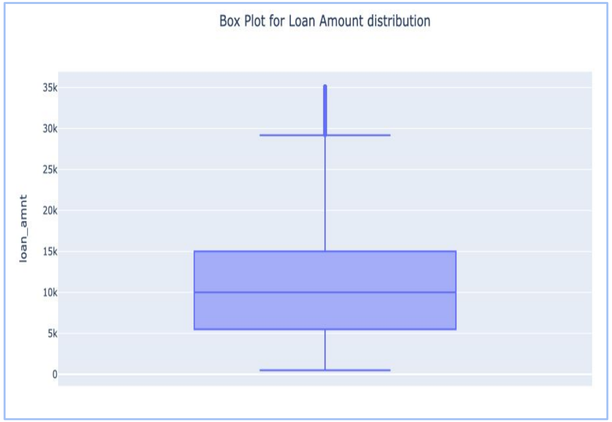
Data Checks & Treatment

Checks Done		Pass/Fail	Final State	Treatments Done
1	Duplicates in Unique Identifier i.e. ID column	✓	✓	• NA
2	Empty rows and Columns & Columns with more than 25% null values	✗	✓	• There were 54 columns with all null values, and 4 columns with >25% null values • These columns were dropped
3	Columns with Single values	✗	✓	• Removed these columns as these would not add any values to the analysis
4	Null Value Checks and Data format	✗	✓	• Nulls were replaced in Emp_length • String removal from Emp Length and Term • Data type changed to Integer
5	Data Type check for date columns (to be used for extraction)	✗	✓	• Changed to to_datetiem format • Year and Month extracted for further analysis for Issue Date

** Left with 39717 rows and 50 columns post derived columns and data treatment*

Outlier Treatment - *Box Plots for Major continuous Variables*

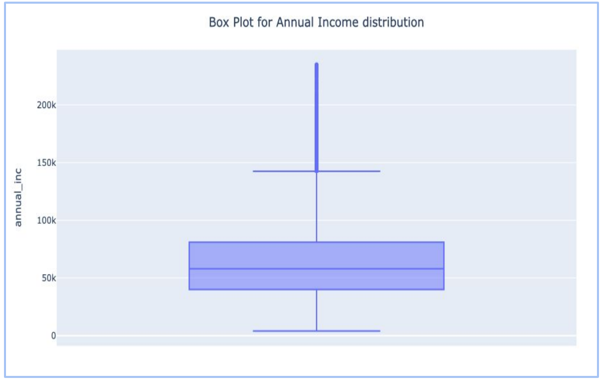
Loan Amount



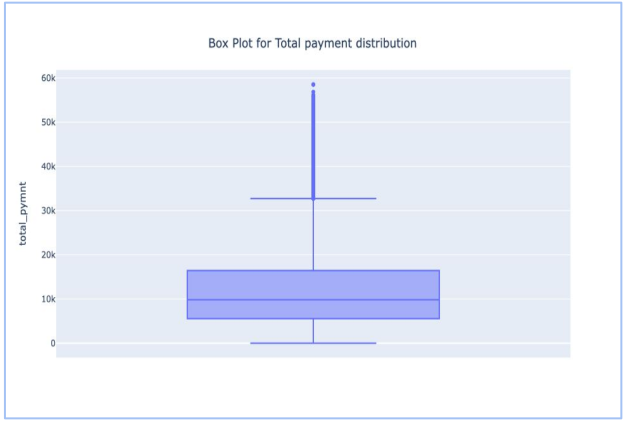
Loan Amount lies within the range of 0-35k $, with 25th and 75th percentiles being at 5.5k and 15k $.
The median is at 10k $

Annual Income

- There were outliers observed in Annual Income going as high as 6M. **Outlier Treatment** was done to remove top 1 percentile of accounts
- Post Treatment, Annual Income lies within the range of 0-235k $, with 25th and 75th percentiles being at 40k and 81k $. The median is at 58k $

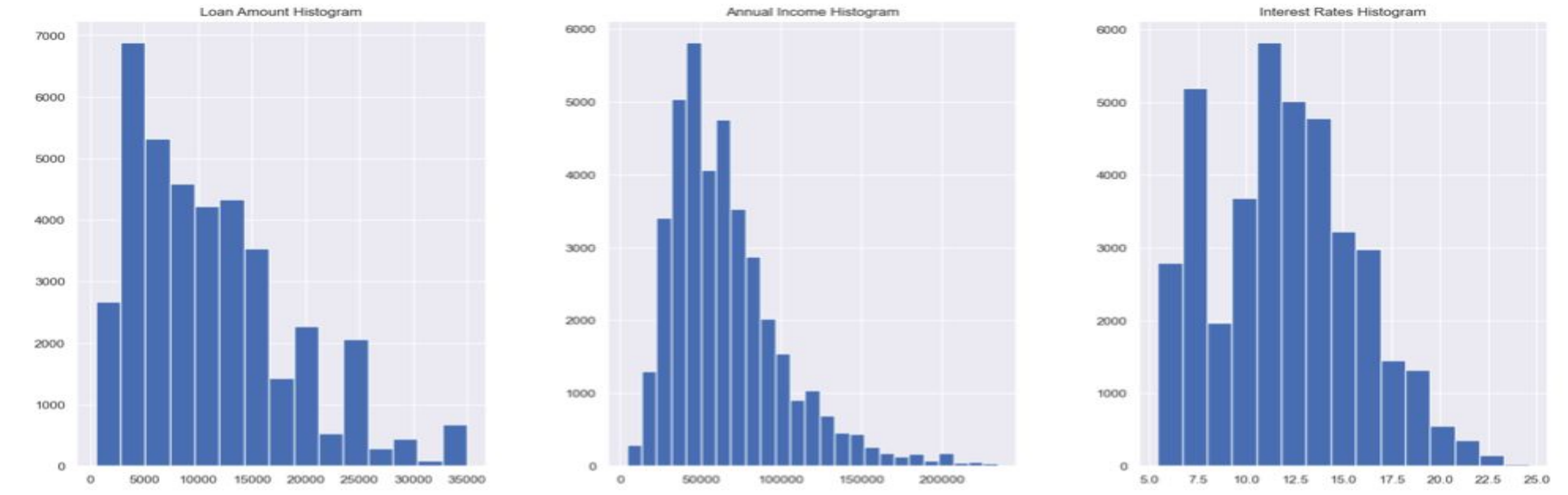


Annual Income



Total Payment lies within the range of 0-58k $. 25th and 75th percentile being at 5.5k and 16k respectively.
The Median lies at ~10k $

EDA - Histograms 1 *(Loan Amount, Income & Interest Rates)*

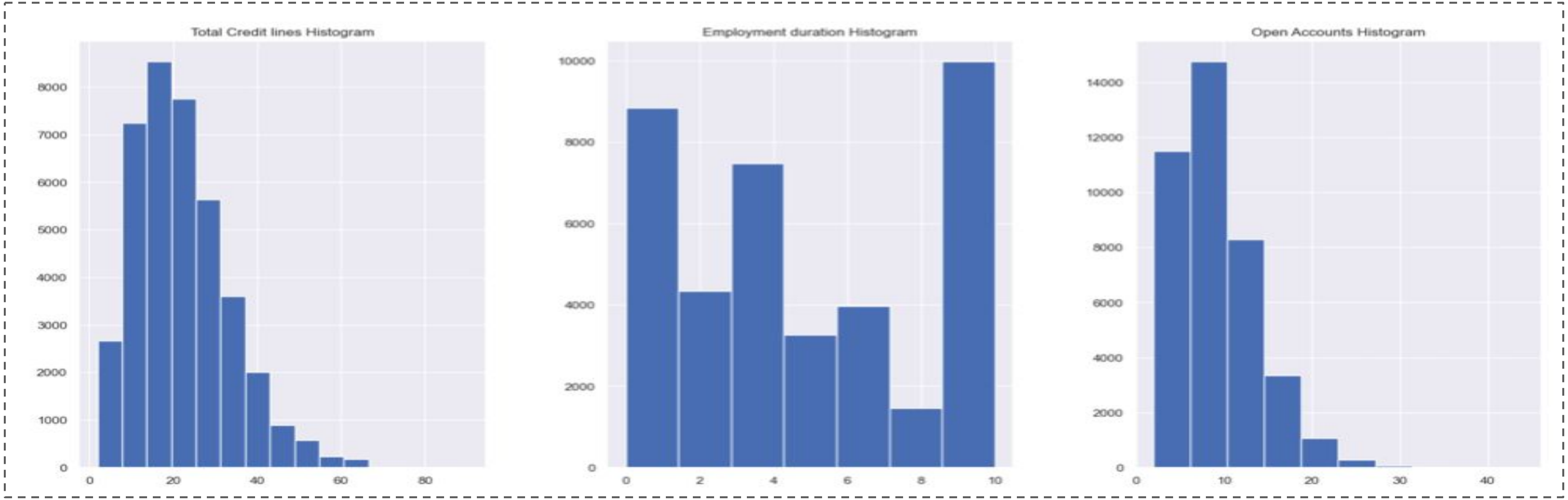


Loan Amount: Basis histogram in 1st chart, # accounts with loan amount peaks before 5k and concentrated between 5-15k

Annual Income: Basis the histogram in the middle chart, Annual Income peaks at 50k

Interest rates: Basis the histogram in the rightmost chart, Interest Rates seem to peak at 11-12% followed by 7%

EDA - Histograms 2 *(Credit Lines, Emp Length & # Open Accounts)*

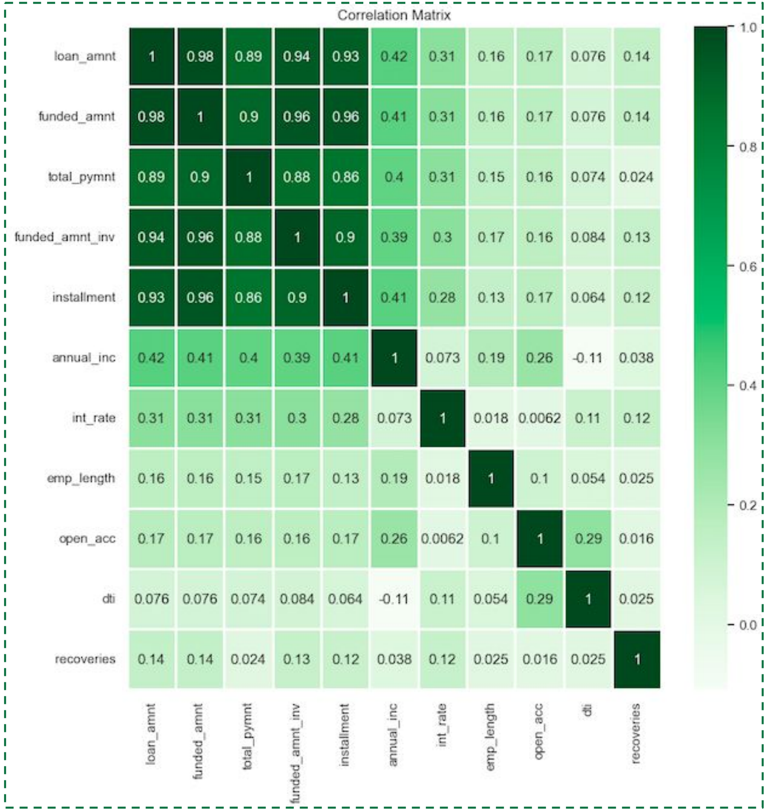


Credit Lines: Basis histogram in 1st chart, # accounts credit lines peaks within 15-20 and tapers down beyond 20

Employment Duration: Basis the histogram in the middle chart, highest number of users lie in > 10 years exp range followed by 0-1 years of work exp.

Open Accounts: Basis the histogram in the right-most chart, majority population seems to have 5-10 open accounts

EDA - Correlation Matrix for major continuous variables



	loan_amnt	funded_amnt	total_pymnt	funded_amnt_inv	installment	annual_inc	int_rate	emp_length	open_acc	dti	recoveries
loan_amnt	1	0.98	0.89	0.94	0.93	0.42	0.31	0.16	0.17	0.076	0.14
funded_amnt	0.98	1	0.9	0.96	0.96	0.41	0.31	0.16	0.17	0.076	0.14
total_pymnt	0.89	0.9	1	0.88	0.86	0.4	0.31	0.15	0.16	0.074	0.024
funded_amnt_inv	0.94	0.96	0.88	1	0.9	0.39	0.3	0.17	0.16	0.084	0.13
installment	0.93	0.96	0.86	0.9	1	0.41	0.28	0.13	0.17	0.064	0.12
annual_inc	0.42	0.41	0.4	0.39	0.41	1	0.073	0.19	0.26	-0.11	0.038
int_rate	0.31	0.31	0.31	0.3	0.28	0.073	1	0.018	0.0062	0.11	0.12
emp_length	0.16	0.16	0.15	0.17	0.13	0.19	0.018	1	0.1	0.054	0.025
open_acc	0.17	0.17	0.16	0.16	0.17	0.26	0.0062	0.1	1	0.29	0.016
dti	0.076	0.076	0.074	0.084	0.064	-0.11	0.11	0.054	0.29	1	0.025
recoveries	0.14	0.14	0.024	0.13	0.12	0.038	0.12	0.025	0.016	0.025	1

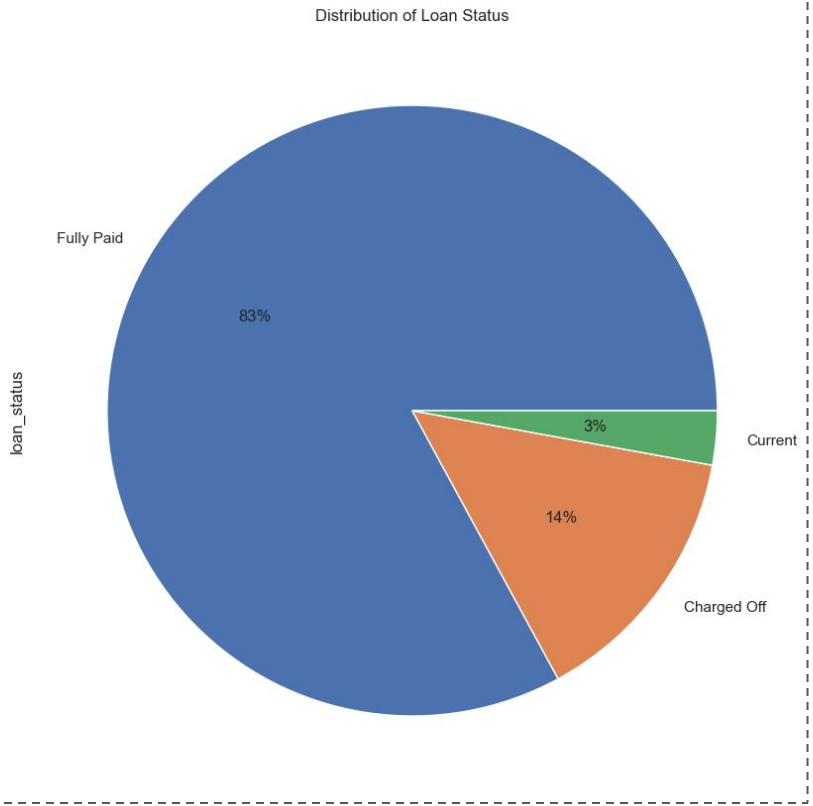
Below are the Correlations observed –

- Loan Amount vs Funded Amount: 0.98
- Loan Amount vs Total Payment: 0.89
- Loan Amount vs Investor Funded Amount: 0.94
- Loan Amount vs Installment: 0.93

- All of these metrics have high correlation within them and hence going forward we will only be using Loan Amount as a continuous metric (or bucketed as categorical metric) as that would gather most of the variation across all the set of above metric

- There is no specific -ve correlation observed in any of the other metrics

EDA - Distribution of Loan Status (using Pie chart)

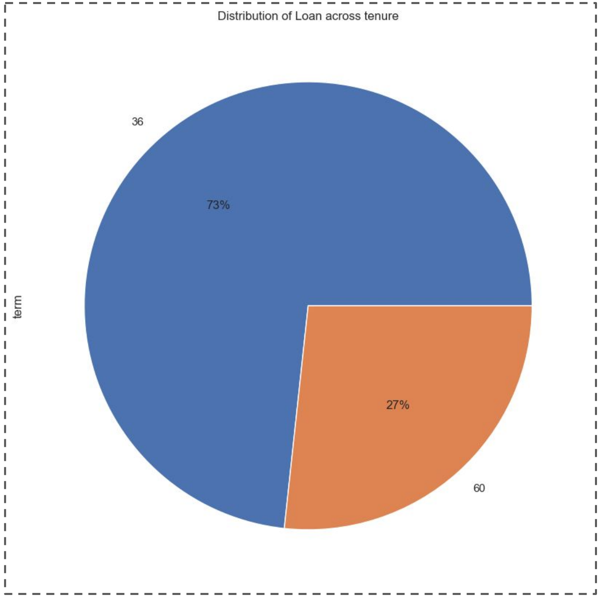


Below are the distributions observed –

- About 83% of the Loans are Fully Paid off
- 14% of the Loans have been charged Off which would be the focus area of the Analysis
- About 3% of the accounts mentioned in the data have current ongoing loans. *The learnings from this analysis can be used to get an idea of how many of these loan are potential risks*

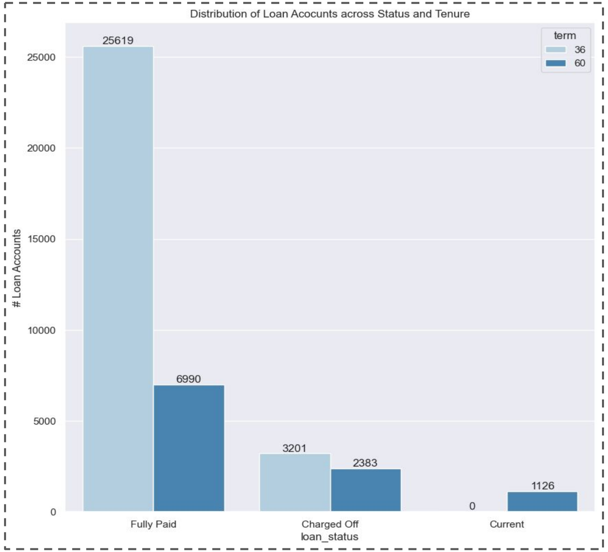
For further study, we can ignore current loans i.e. 3% of the dataset as the outcome is not yet clear for the objective function i.e. Default or Charged Off loans

EDA - Distribution of Tenure & Loan Status (using Pie/Bar chart)



Below are the distributions observed –

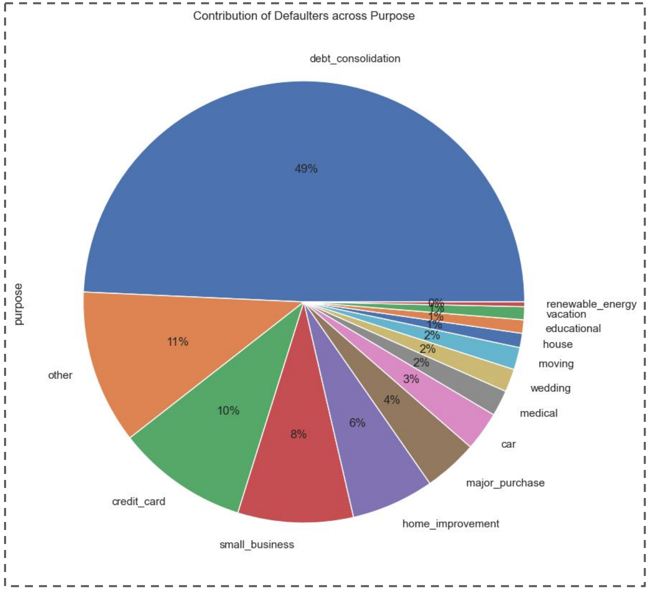
- About 73% of the Loans are 36 months tenured whereas, 27% loans are 60 months tenured



Below are the distributions observed –

- About 74% of the Charged Off Loans are from 60 month tenure loans
- Excluding current ongoing loans, 25% of the 60m term loans are Charged off as opposed to 11% for 36m term

EDA - Distribution of Loan Purpose & Status (using Pie/Bar chart)

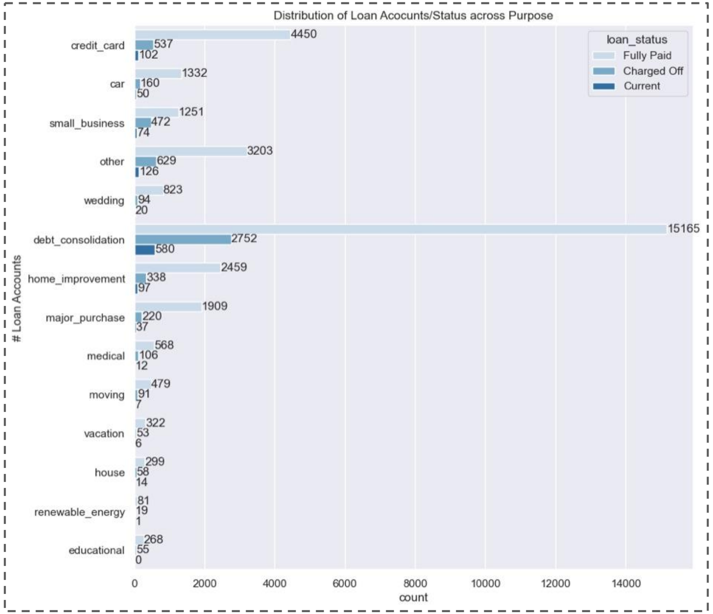


- About half (49%) of the Loans are taken with a purpose of Debt Consolidation
- Excluding others, Credit Card and Small Business loans are 2nd and 3rd highest contributors at 10% & 8% respectively

Bivariate analysis for Purpose against Loan Status

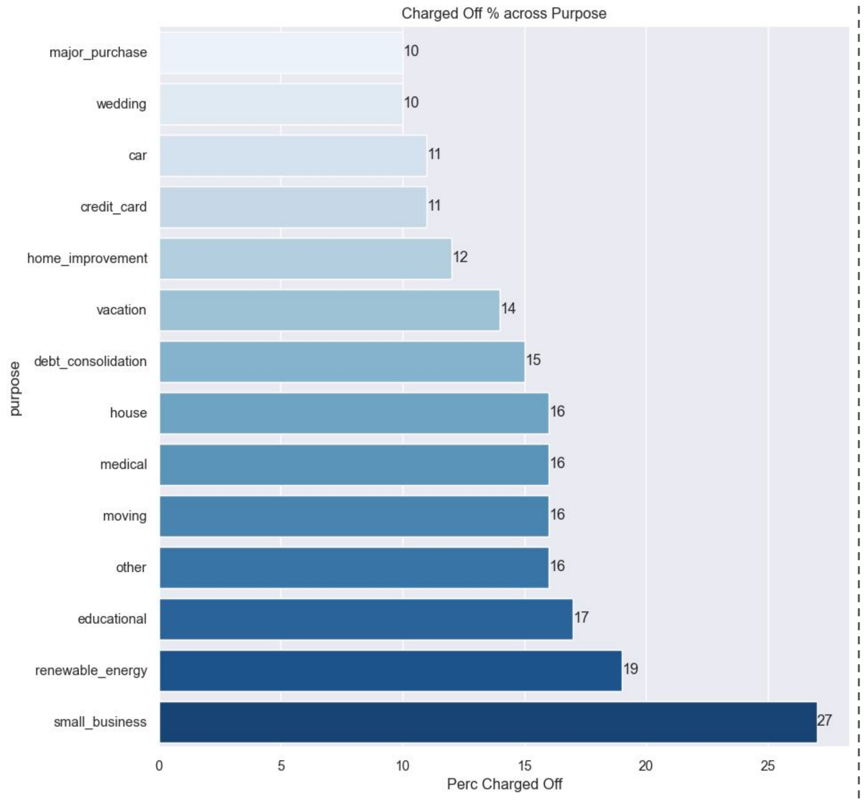
Below are the distributions observed –

- Similarly almost 50% of the Defaulters lie in loans taken under the purpose of debt consolidation



EDA - Loan Purpose wise % Charged off Loans

** Charged off loan is defined as Total charged off loans upon (Charged Off + Fully Paid Loans)*

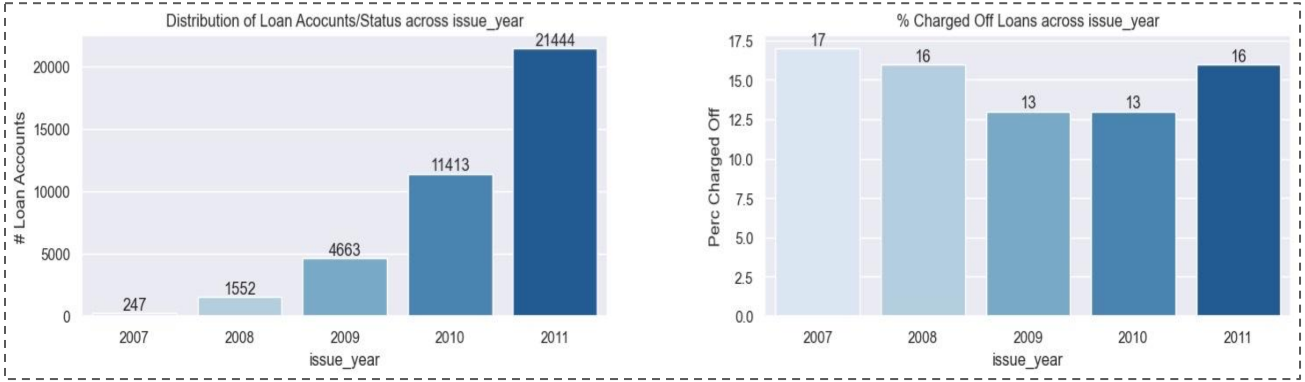


Below are the observations –

- Small Business Loans, that had the 3rd highest contribution to loans has the highest Default % at 27%
- This is followed by Renewable Energy and Educational purpose loans with Defaulter percentage at 19% & 17% respectively

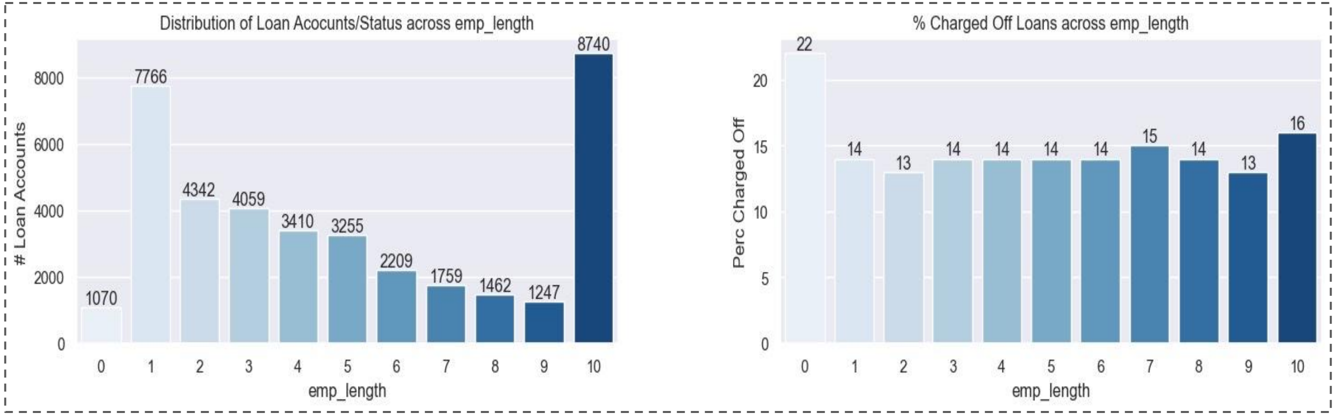
Small Business loans and Renewable Energy loans are highly risky purposes leading to highest defaulters. We should check if interest rates are higher here, if not there is scope for either reducing risk or increasing interest rates

EDA - Study across Issue Year & Employment Length - Contribution to loan and & % Defaulters

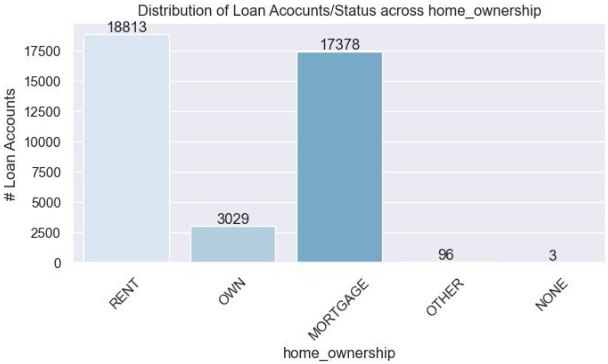


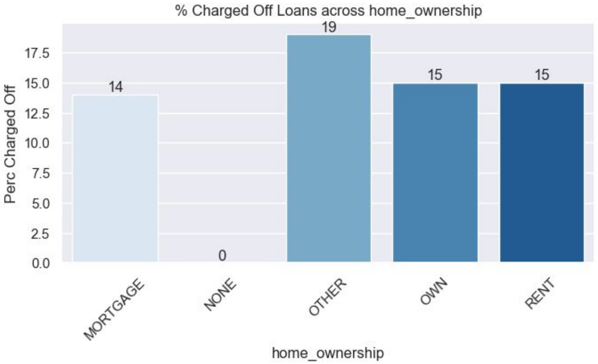
- *There has been a gradual increase in Loans across the years 2007 to 2011*
- *There is no specific trend in % defaulters across years*

- *Major portion of the loan account holder sare either > 10 years experienced or with <= 1 yr experience*
- *< 1 yr experience acc holders have highest Defaulter %. There is no trend beyond that*



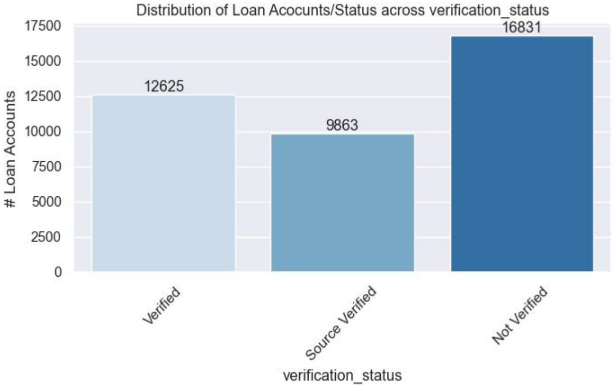
EDA - Study across House Ownership & Verification Status - Contribution to loan and & % Defaulters

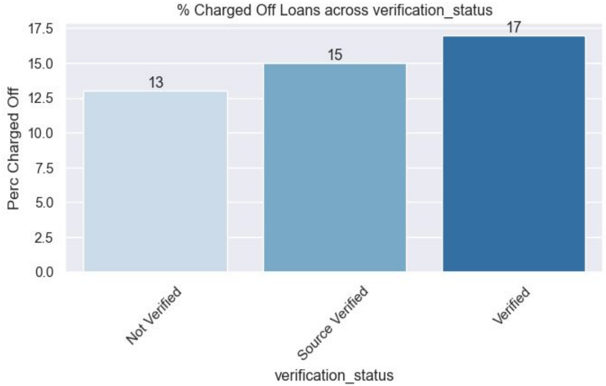




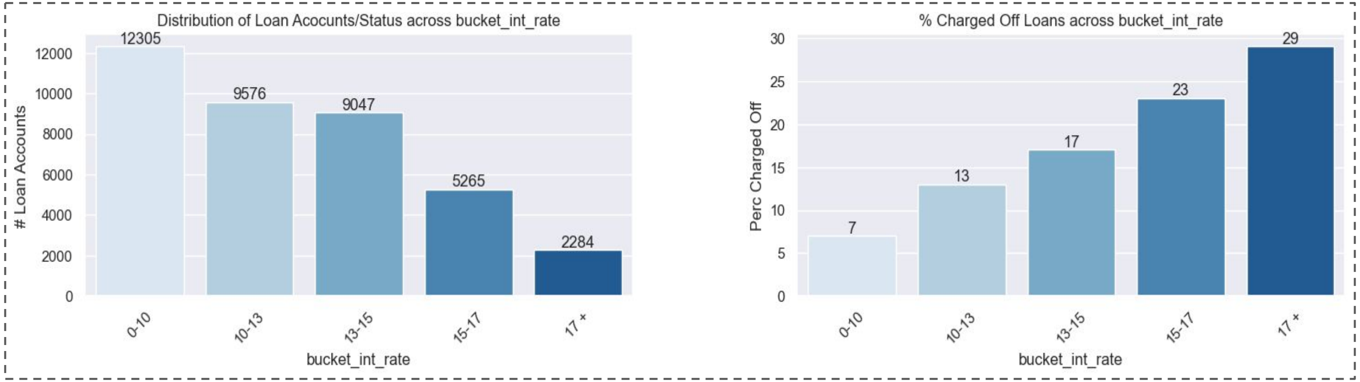
- *Mortgage and Rent users contribute to the majority loan users*
- *Even with only 96 loan accounts from others section, Defaulter % is highest at 19%*

- *About 43% of loans are contributed by Non verified users*
- *Surprisingly, non-verified users have the lowest defaulter rates at 13% compared to verified users at 17%*



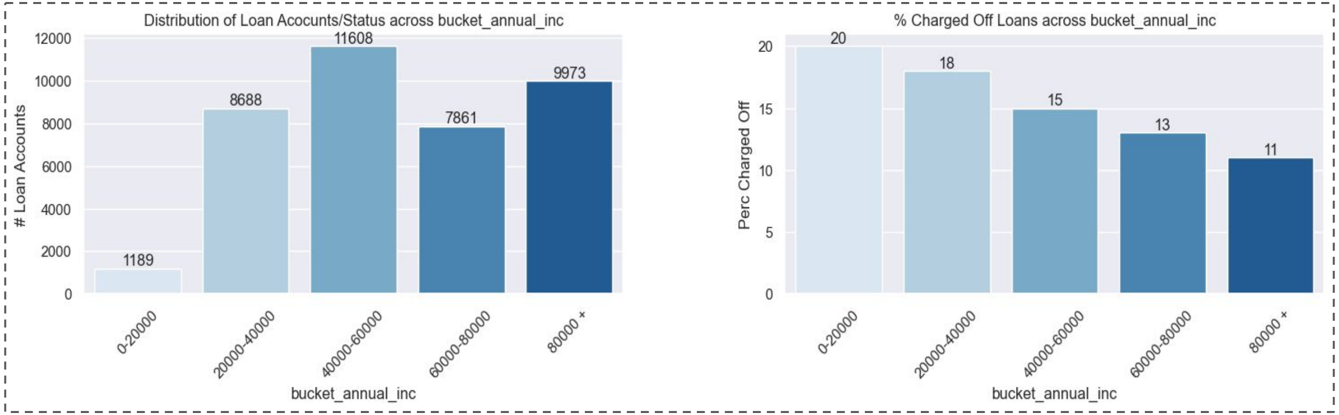


EDA - Study across Interest Rates & Verification Status - Contribution to loan and & % Defaulters

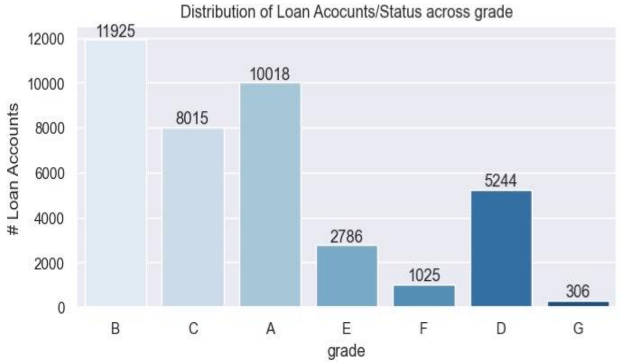


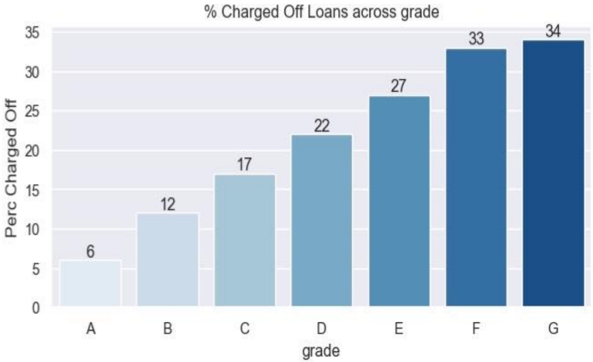
- *There is a steep increase in Defaulter % as the interest rate increases with ~29% for > 17% interest rate bucket*

- *There is no trends in # of loans based on Income buckets*
- *However, there is a stark declining trend in Defaulter % as income buckets move higher (highest at 20% in the < 20k bucket)*



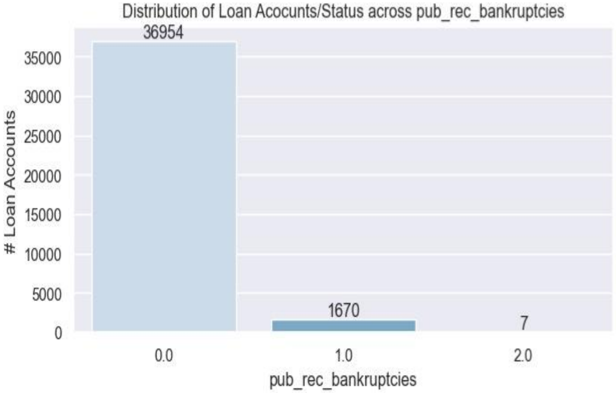
EDA - Study across Grades & Bankruptcy records - Contribution to loan and & % Defaulters

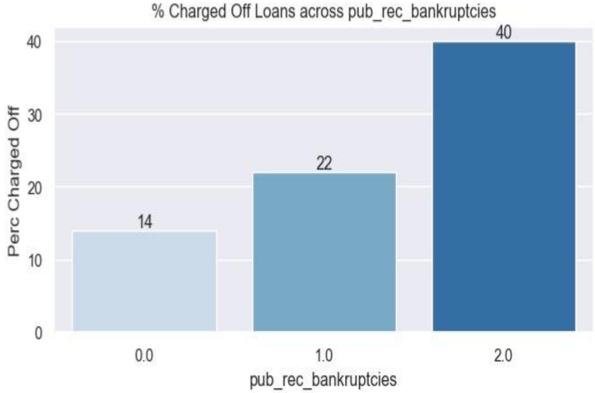

Distribution of Loan Acocunts/Status across grade


% Charged Off Loans across grade

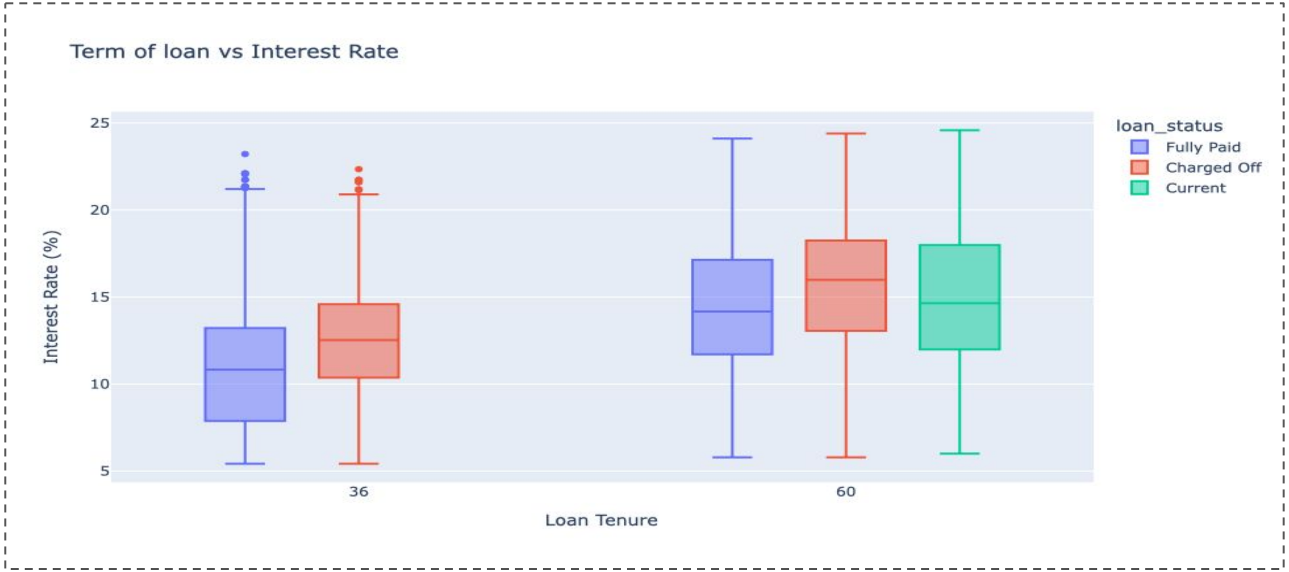
- *There is a steep increase in Defaulter % at higher Grades (E, F, G) at 27%, 33% and 34% respectively*

- *While the # of loans for users with >= 1 bankruptcies is minimal, the default rates are as high as 40% for users with 2 records*


Distribution of Loan Acocunts/Status across pub_rec_bankruptcies

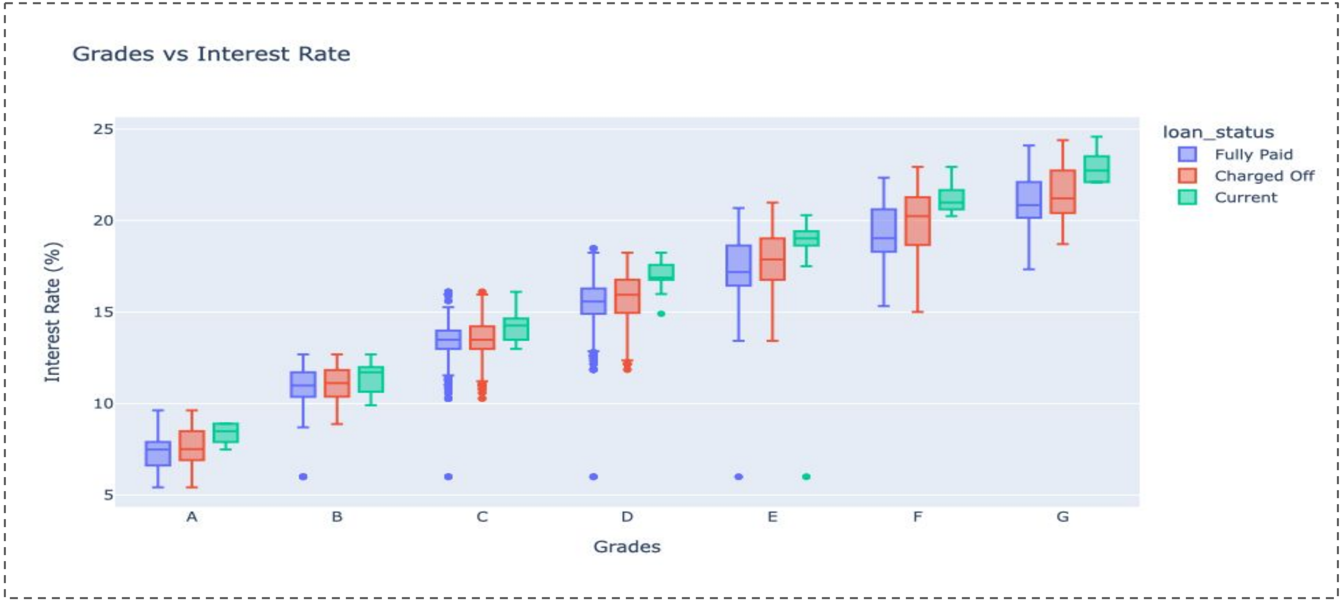

% Charged Off Loans across pub_rec_bankruptcies

EDA - Multivariate Analysis - Term, IR vs Loan Status



- Interest Rates are significantly higher for 60 months loan tenure compared to 36 month tenure (median at 14.79% and 11% respectively).
- Within 60 months tenure as well the Defaulter's median IR is at 16% compared to 14.7% for Fully Paid ones

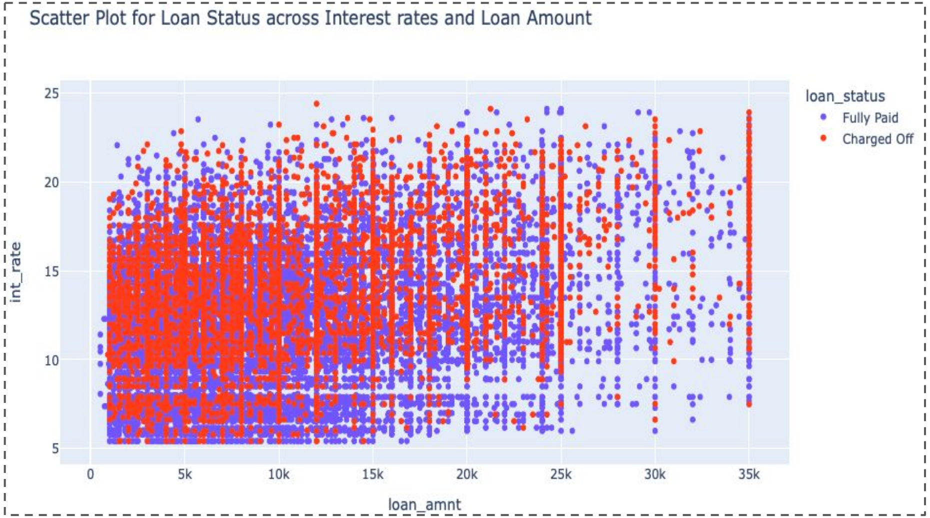
EDA - Multivariate Analysis - Grades, IR vs Loan Status

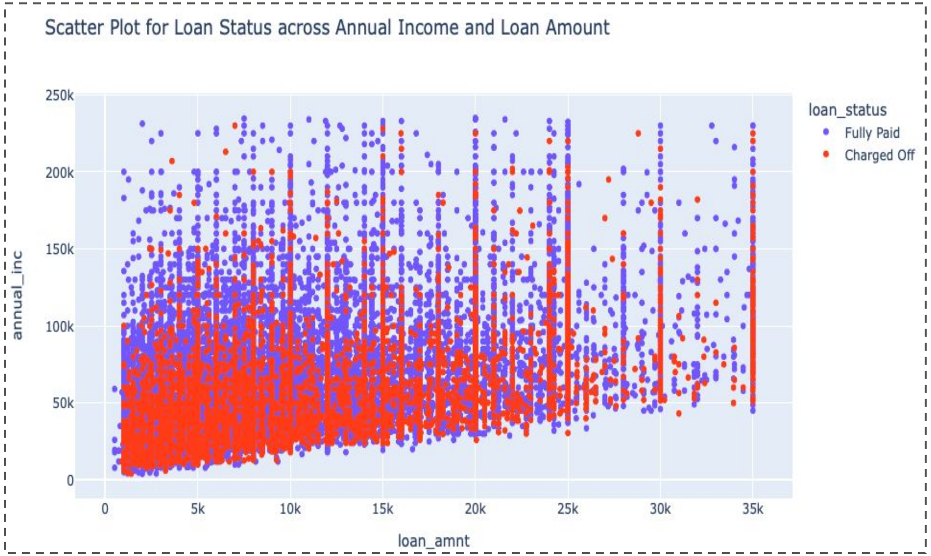


- There is a clear Trend of Increase in IR as one moves from Grade A to Grade G Loans.
- Interestingly, it is observed that the loans with Status as current have the highest Interest Rates within Each bucket

EDA - Multivariate Analysis using Scatter Plots (Interest rates vs Loan Amount & Income vs Loan Amount

There is high consolidation of Charged Off loans at lower Loan Amount and higher Interest rates


Scatter Plot for Loan Status across Interest rates and Loan Amount


Scatter Plot for Loan Status across Annual Income and Loan Amount

There is an increase in bottom limit of loan amount with increasing annual income

There is high consolidation of Charged Off loans at low annual income and lower loan amount