Lending Club Case Study

Divyesh Sharma
Rahul Pandey

Problem Statement

- A Lending Company dataset with details corresponding to Default and Non-Default customers. The target is to figure out the parameters around which a customer can default.

Agenda

- Analysing the Data
- Cleaning
- Categorizing\Visualizing the Data
- Filtering & Imputing
- Univariate Analysis
- Bivariate Analysis
- Multivariate Analysis
- Conclusion

Analysing & Visualising the Data

We have around 111 columns to analyse data.

On straightforward analysis, there are around 55 columns which has more than 75% of null data.

Hence, we can remove those columns.

Cleaning

Cleaning ¶

```
# Get the percentage of null values in each column
null_percentage = df.isnull().mean() * 100

# Extract column names with 100% null values
columns_with_nulls = null_percentage[null_percentage >= 75].index.tolist()

len(columns_with_nulls)
```

56

```
# Drop columns with >=75% null values
df = df.drop(columns=columns_with_nulls)
```

```
df.shape
```

(39717, 55)

```
#Getting columns which has only single unique values.
df_combined = combinedInfo(df)
single_valued_columns = df_combined[df_combined["Unique Values"] == 1]
df[single_valued_columns.index].head(2)
```

	acc_now_delinq	application_type	chargeoff_within_12_mths	collections_12_mths_ex_med	delinq_amnt	initial_list_status	policy_code	pymnt_plan	tax_liens
0	0	INDIVIDUAL	0.0	0.0	0	f	1	n	0.0
1	0	INDIVIDUAL	0.0	0.0	0	f	1	n	0.0

- Dropped around 56 columns which were having >=75% of null values.
- Also around 9 columns have only one unique value.
- After Removing above not required columns only 46 columns remain.

Categorizing\Visualizing the Data

Behavior Variables
Delinquency Year -2 (delinq_2yrs)
Debt to income (Dti)
Earliest Credit Date (earliest_cr_line)
Revolving balance (revol_bal)
Loan Purpose (purpose)
Term (term)
Annual Income (annual_inc)
Employeement Length (annual_inc)
Home Ownership (home_ownership)
Number of Credit Lines(total_acc)
Open Credit Lines (open_acc)
Derogatory Record (pub_rec)
Record of Bankruptcies (pub_rec_bankruptcies)
Revolving Credit Balance(revol_bal)
Revolving Utilization Rate (revol_util)
Loan Title (title)

Loan Characteristics
Loan Amount (loan_amnt)
Funded Amount (funded_amnt)
Funded Amount Investment (funded_amnt_inv)
Interest Rate (int_rate)
Loan Status (loan_status)
Loan Grade (grade)
Loan Issue Date (issue_d)
Loan Sub Grade (sub_grade)
Income Verification (verification_status)

Customers Location
Employment Title (title)
Employment Length (emp_length)
Zip Code (zip_code)
Description (desc)

Post Loan Sanction
Total Payment Recieved (total_pymnt)
Investor Payment Received (total_pymnt_inv)
Interest Received (total_rec_int)
Late Fees Received (total_rec_late_fee)
Principal Received (total_rec_prncp)
Recovery Fee (collection_recovery_fee)

Contd

Columns to drop

Inline similar columns

- State & zipcode, are similar, hence we can drop addr_state column.

Post Loan Sanction columns

- last_pymnt_amnt, out_prncp, out_prncp_inv, total_pymnt_inv recoveries, total_pymnt, total_pymnt_inv, total_rec_int, total_rec_late_fee, total_rec_prncp : Useful post loan approves

Row Unique columns

- member_id, id, url, last_credit_pull_d, emp_title : fields not useful in analysis as they all have unique values

Filtering & Imputing

Filtering

Filtering loan_status with Current, as we are only interested in either Fully paid or Charged off

```
# removing current loan amount
df = df[df.loan_status != "Current"]
df.shape
```

```
(38577, 35)
```

Filtering data on Current, as we only need analysis around Fully Paid or Charged Off Loan

Imputing columns

Below columns are having null values, hence we can evaluate whether to drop the null values or impute them.

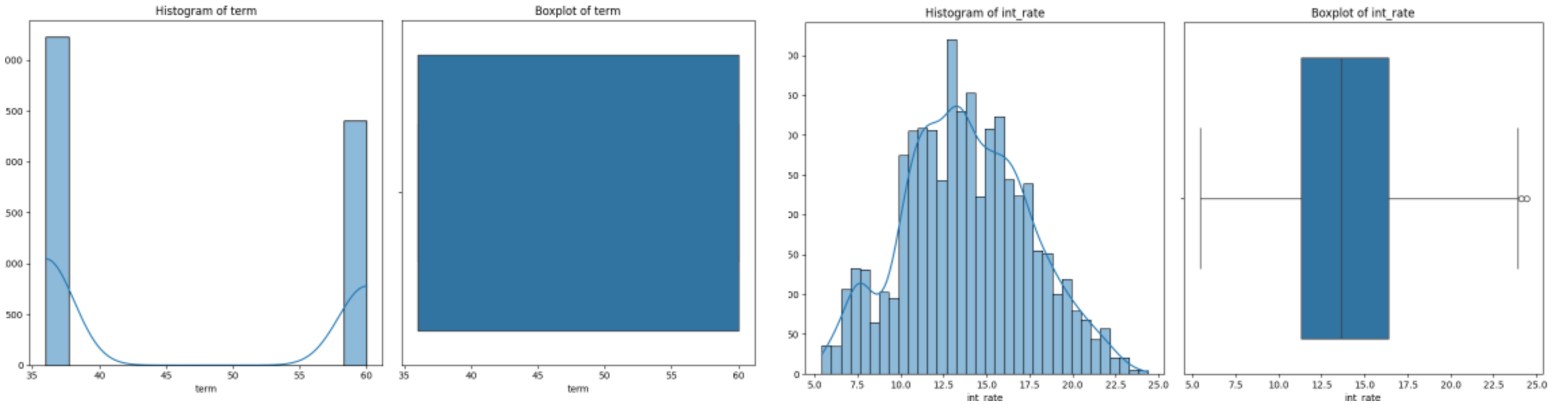
- emp_length
- title
- revol_util
- pub_rec_bankruptcies

Below columns are object type, converting it to correct type.

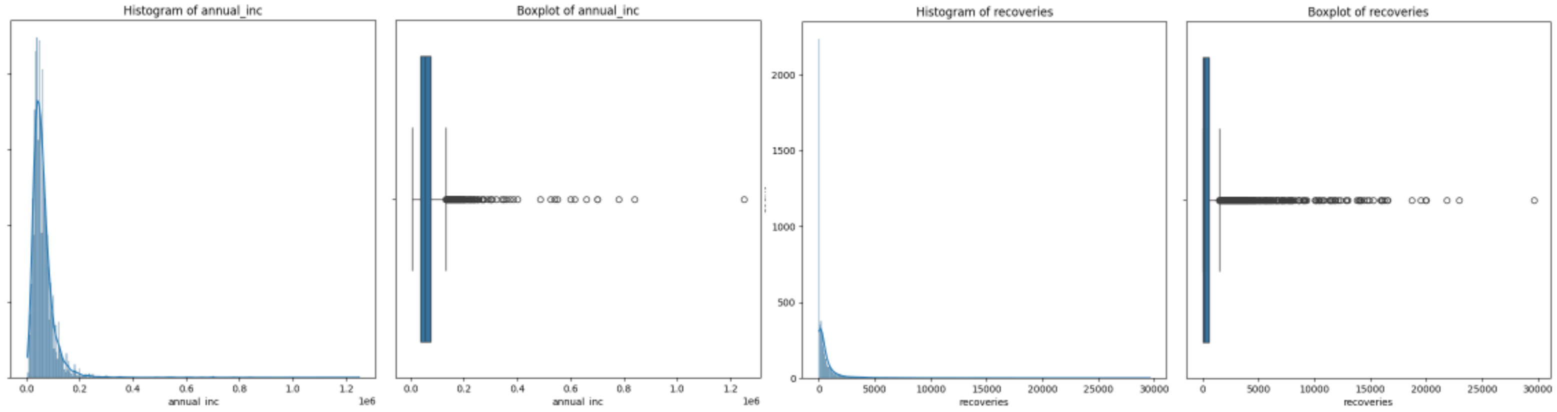
- term : to integer
- int_rate : to float
- loan_status : to category

Univariate Analysis - Numerical

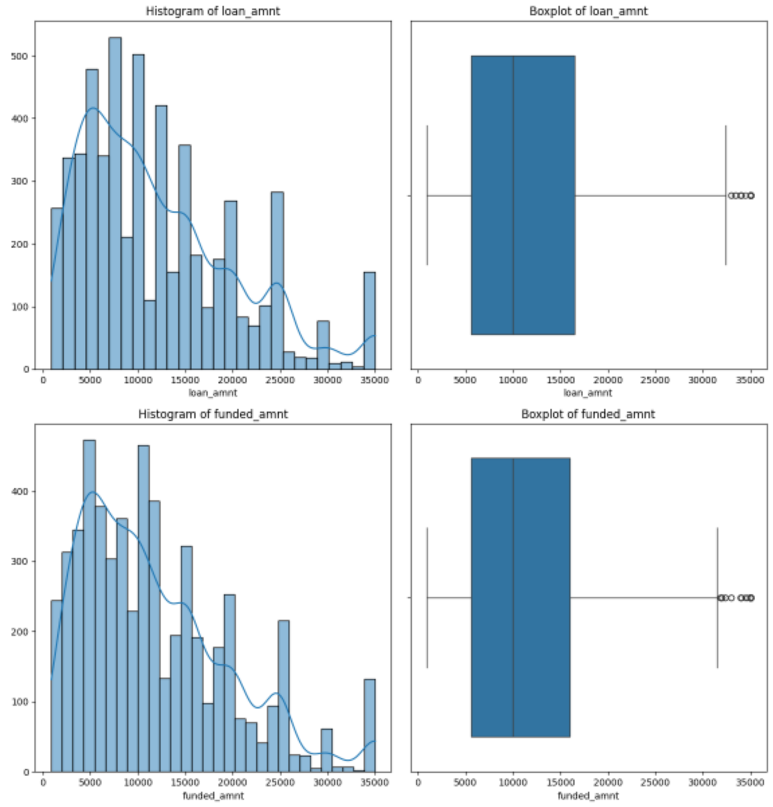
Impact creating parameters



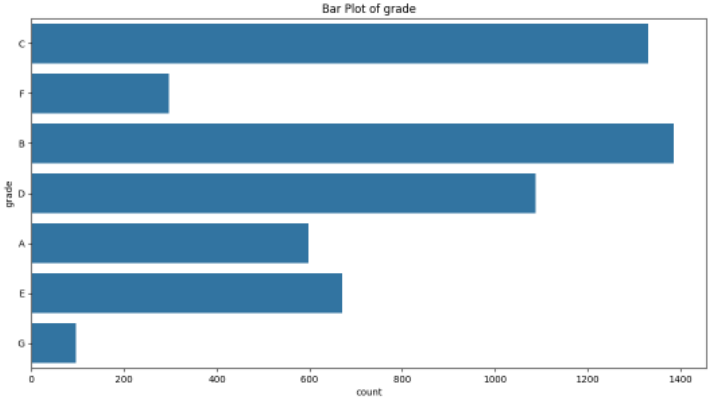
Outliers

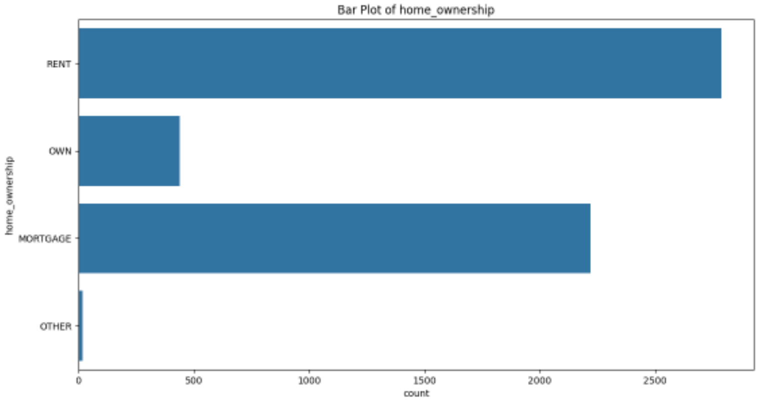


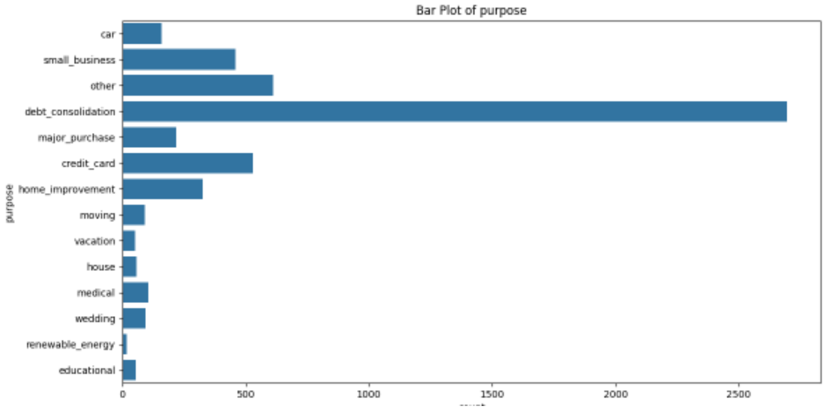
Correlated data or Similarly Skewed

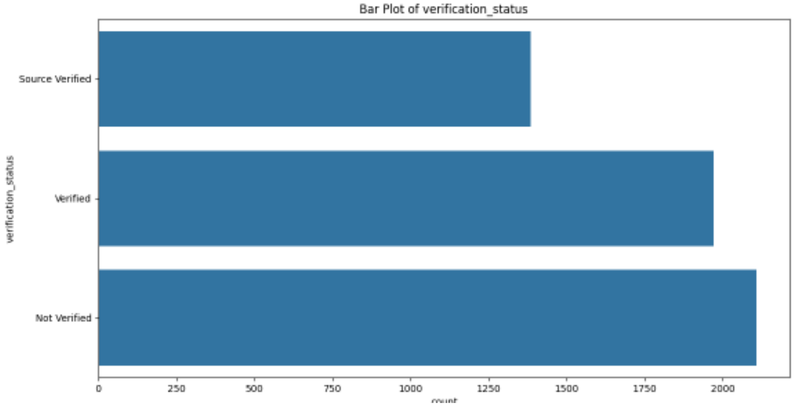


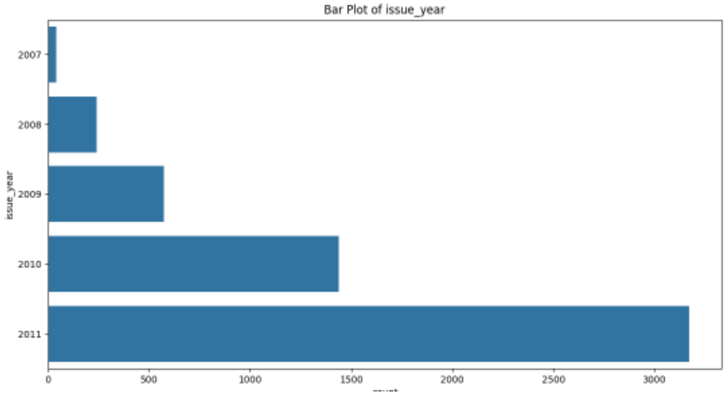
Univariate Analysis - Categorical

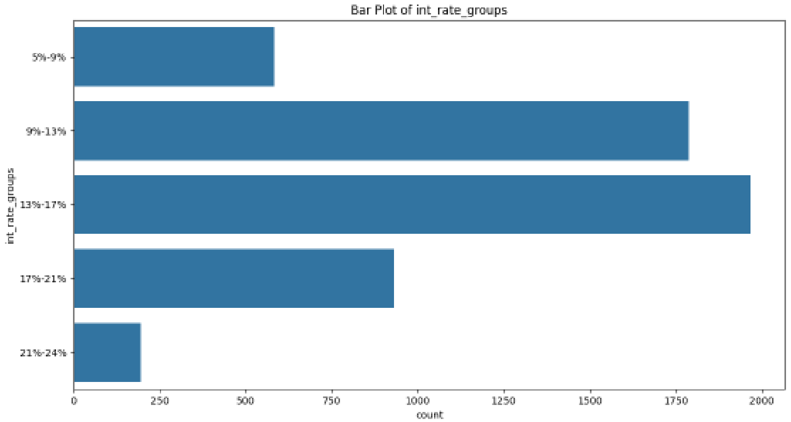






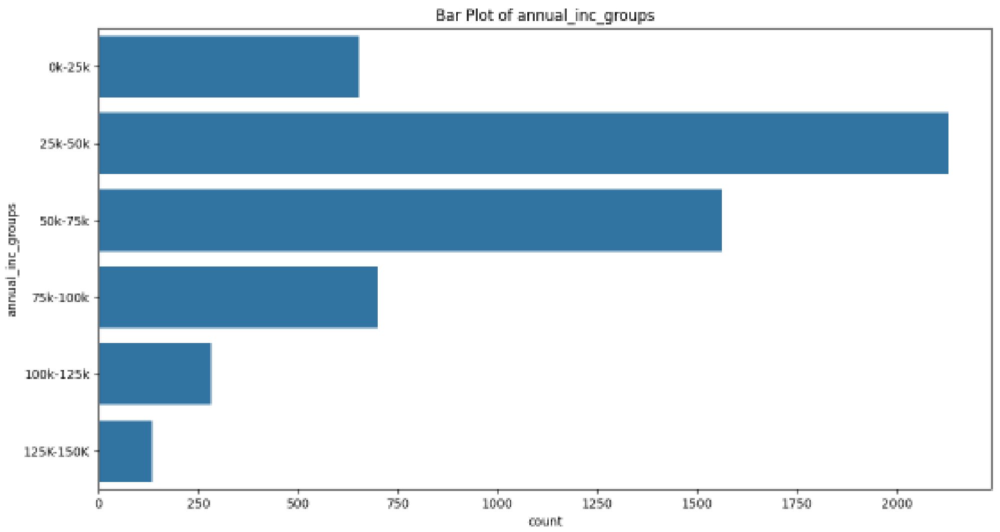


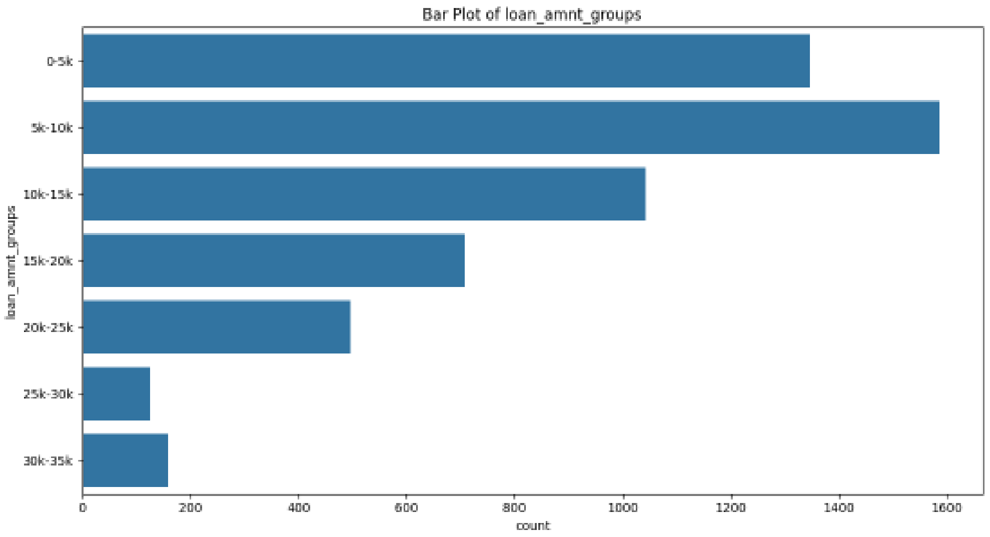




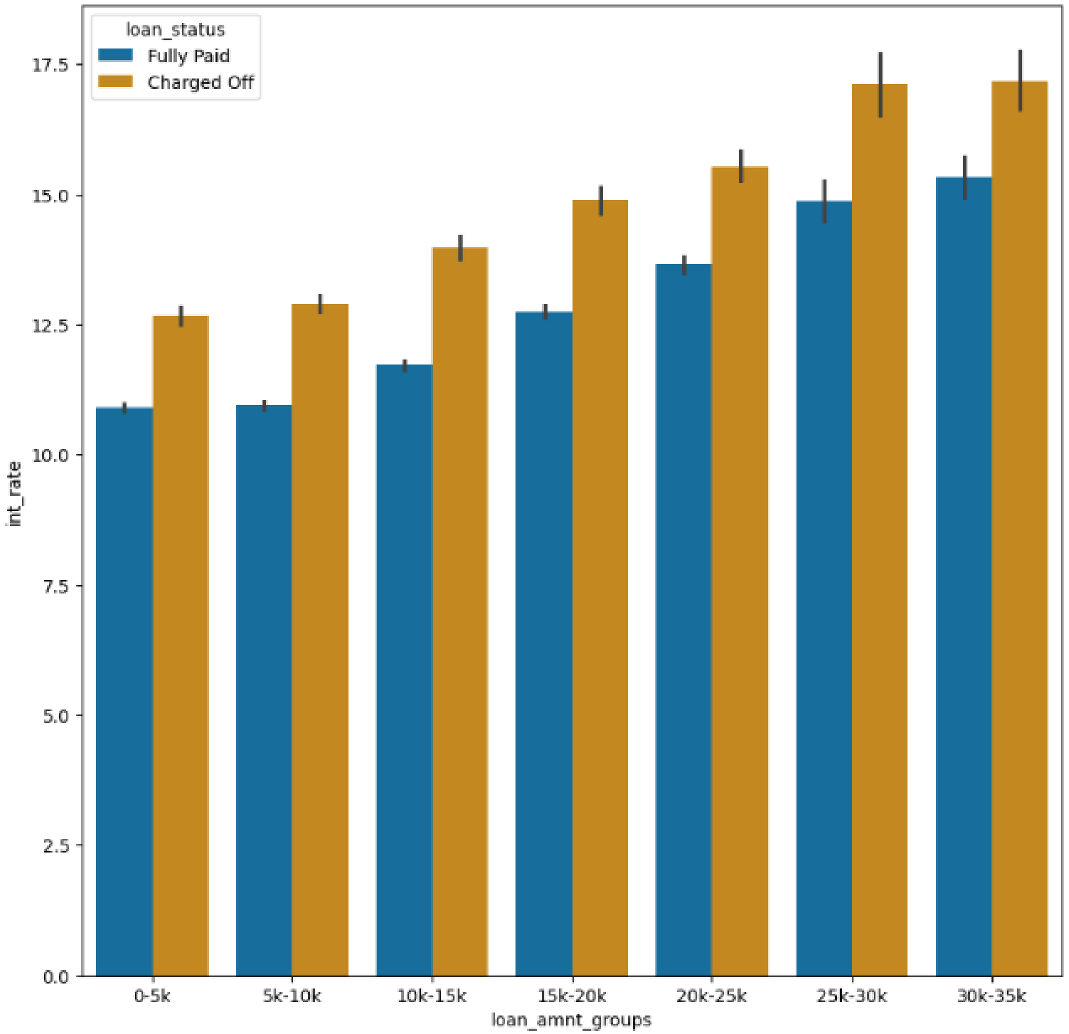
contd

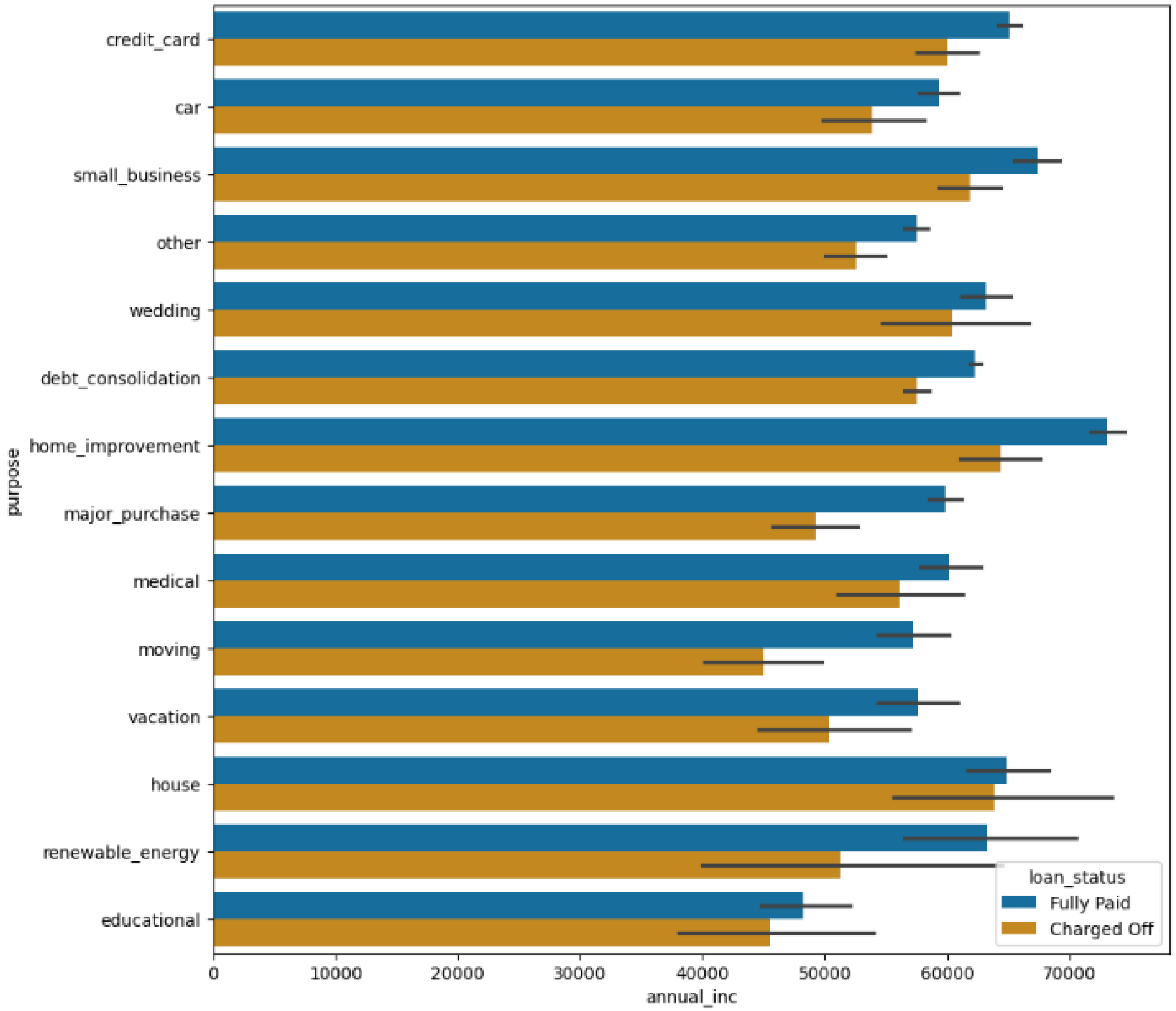
- **grade** : Major contributers are B, C, D
- **home_ownership** : Rent and Mortgage is having more have more defaulters. Rent is the leading.
- **verification_status** : verified and not verified has more contribution in defaulters
- **purpose** : debt_consolidation has the highest defaulting rate.
- **earliest_credit_line** : Need more analysis, as the graph shows some trend on the credit line.
- **last_credit_date** : shows that for one date the defaulter count is more.
- **issue_d**: shows the trend that from jan to dec 2011 more defaulters were there and especially the graph for issue_year shows more in 2011.
- **int_rate** : 9-13% and 13-17% are major contributors
- **loan amount** : 510K loan is a major contributer
- **annual income**: 2530K is a major contributer



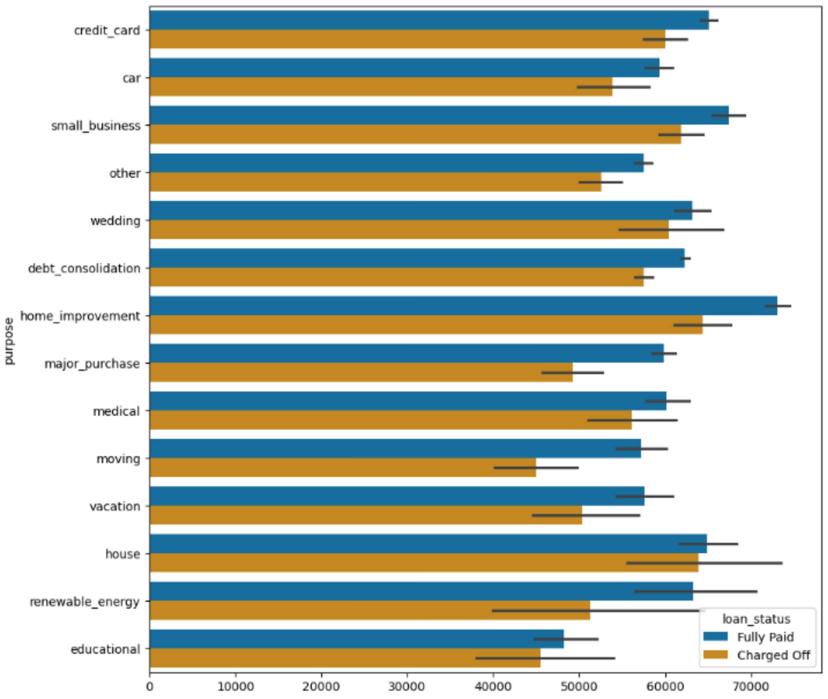


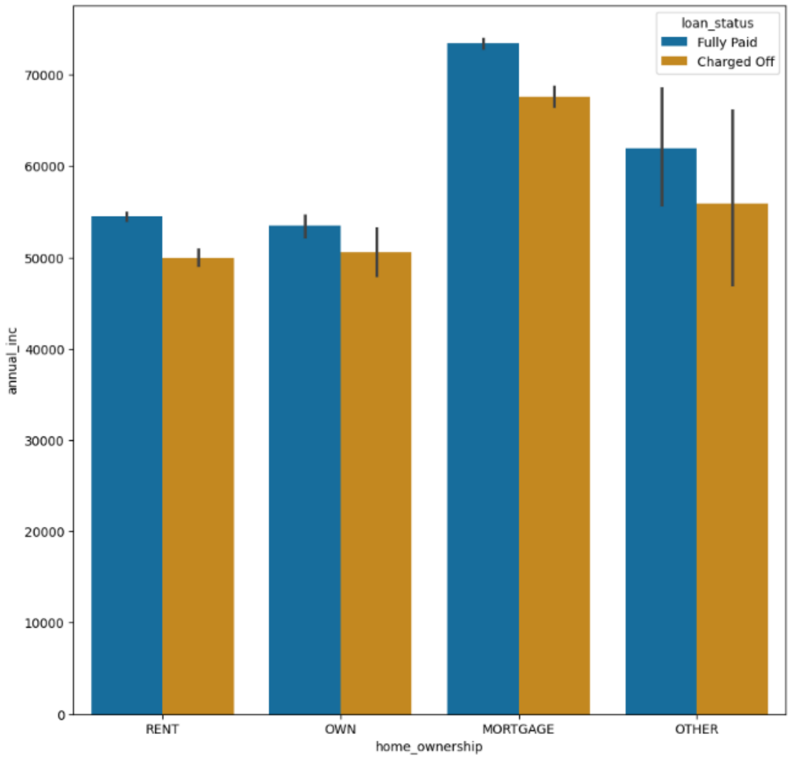
Bivariate Analysis





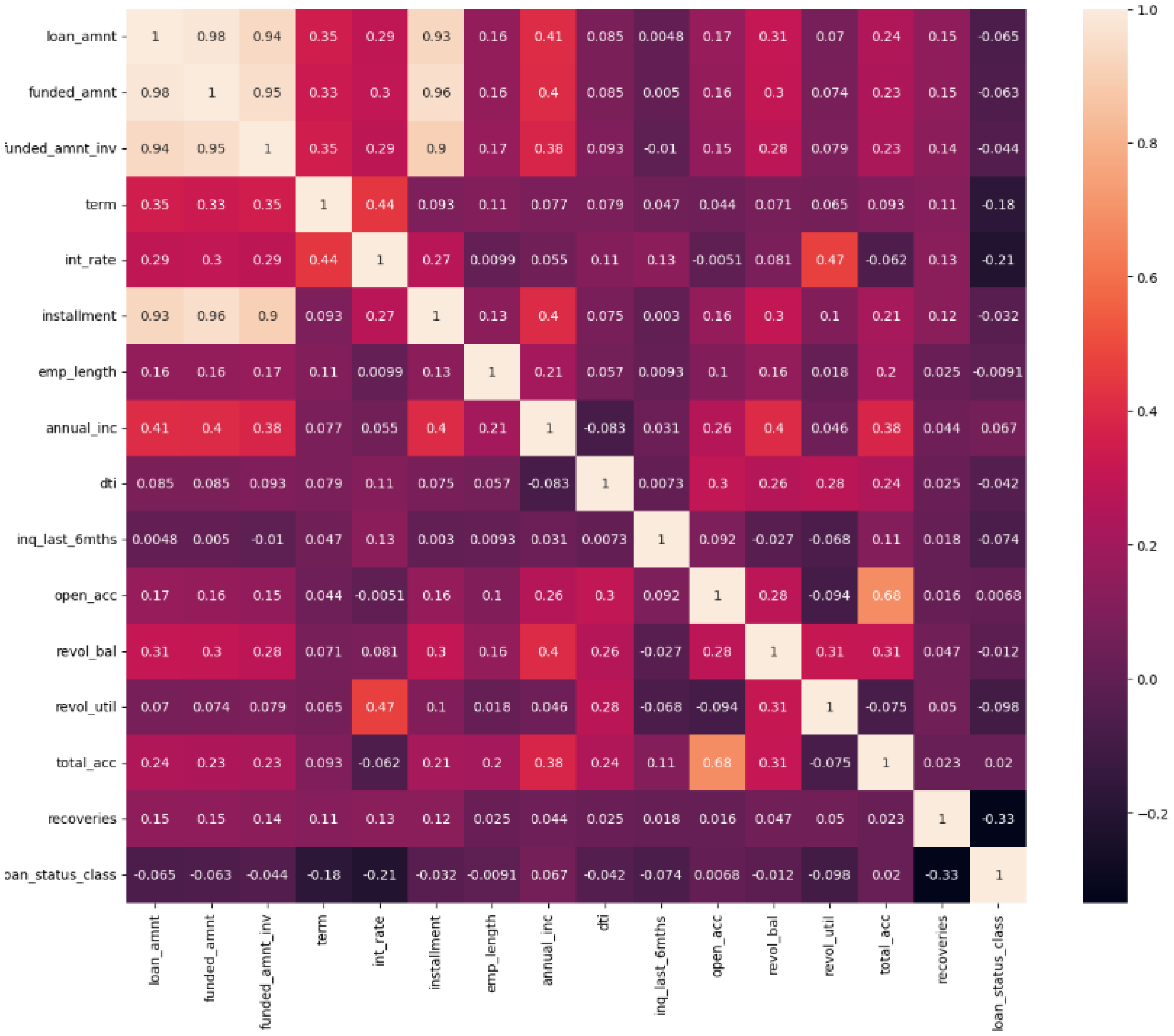
Contd





- annual income vs loan_amnt: betwen 3035K with interest rate between 15 to 17.5 (higher side)
- annual income vs int rate: interest rate between 2124% and income > 80K.
- loan amount vs int rate : loan amount between 3035K and interest rate between 16 to 17.5%
- annual income vs purpose: with home_improvement and annual income between 60 to 80K
- annual income vs home ownership : ownership type Mortgage and income between 70-80K

Multivariate Analysis



[479]:

	loan_status	grade	int_rate_groups	annual_inc_groups	term	purpose	home_ownership	verification_status	loan_default_count	chargeoff_percentage
12465	Charged Off	B	9%-13%	25k-50k	36	debt_consolidation	RENT	Not Verified	68	1.244510
24561	Charged Off	C	13%-17%	25k-50k	36	debt_consolidation	RENT	Not Verified	52	0.951684
369	Charged Off	A	5%-9%	25k-50k	36	debt_consolidation	RENT	Not Verified	35	0.640556
12466	Charged Off	B	9%-13%	25k-50k	36	debt_consolidation	RENT	Source Verified	35	0.640556
12456	Charged Off	B	9%-13%	25k-50k	36	debt_consolidation	MORTGAGE	Not Verified	30	0.549048
46907	Charged Off	E	17%-21%	25k-50k	60	debt_consolidation	RENT	Verified	30	0.549048
34641	Charged Off	D	13%-17%	25k-50k	36	debt_consolidation	RENT	Not Verified	29	0.530747
47243	Charged Off	E	17%-21%	50k-75k	60	debt_consolidation	RENT	Verified	28	0.512445
47234	Charged Off	E	17%-21%	50k-75k	60	debt_consolidation	MORTGAGE	Verified	24	0.439239
12962	Charged Off	B	9%-13%	50k-75k	60	debt_consolidation	MORTGAGE	Verified	24	0.439239

Conclusion

- The chance of a consumer defaulting is more when below conditions are there.
  - The term is 36
  - if the grade is B or C
  - if the purpose is debt_consolidation
  - home_ownership is RENT or Mortgage
  - verification_status is Not Verified
  - loan_amount between 5k-10k
  - int_rate between 9-17%