# LENDING CLUB CASE STUDY

# SUBMISSION

Group Members:

1. Ravi Kiran Kanneganti
2. Chaitra C R
3. Indranil Ganguly
4. Nikhil Gupta

# Objective of Data Analysis

- LendingClub is a platform that connects borrowers and investors.

- Borrowers are required to give their personal, professional and purpose of loan requirement.

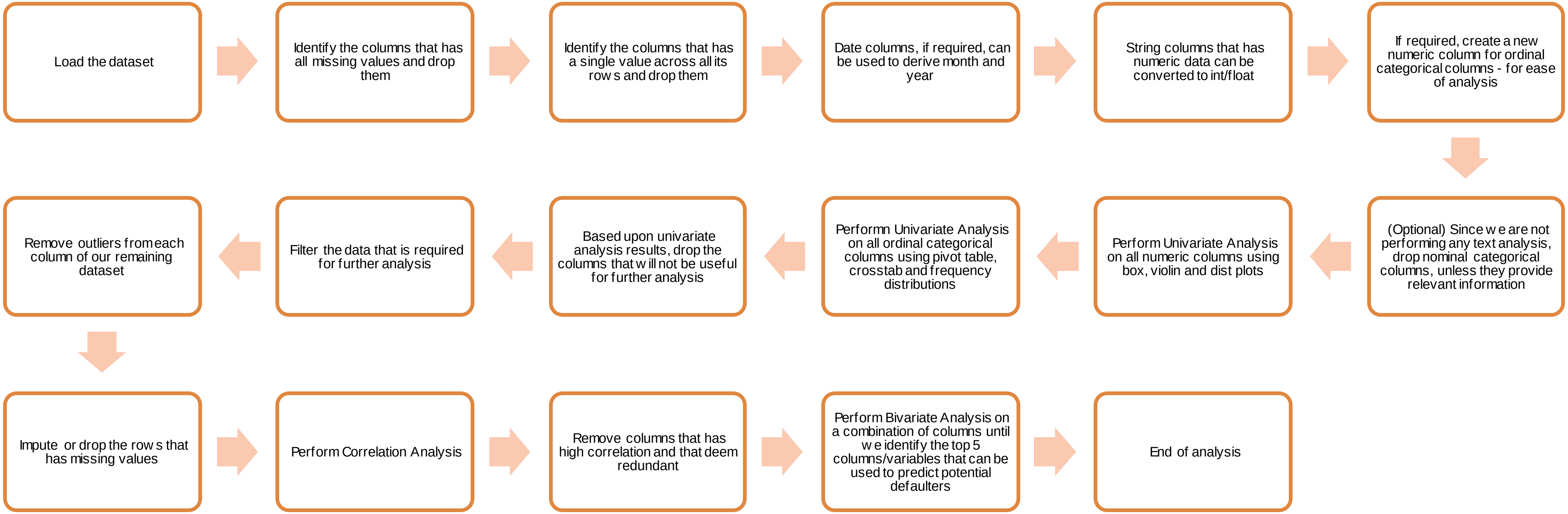
- LendingClub then pulls the credit history of the loan applicant, audits the profile and either approves or rejects the application.

- Investors can browse through various approved loan applications and invest in them - for which, they will be shared the profits.

- Like any credit business, LendingClub faces loss majorly because of the defaulted loans.

- Our objective is to perform Exploratory Data Analysis on the give data and identify the top 5 drivers that can be used to predict potential defaulters.

# Data Analysis Strategy



Load the dataset

Identify the columns that has all missing values and drop them

Identify the columns that has a single value across all its row s and drop them

Date columns, if required, can be used to derive month and year

String columns that has numeric data can be converted to int/float

If required, create a new numeric column for ordinal categorical columns - for ease of analysis

(Optional) Since w e are not performing any text analysis, drop nominal categorical columns, unless they provide relevant information

Perform Univariate Analysis on all numeric columns using box, violin and dist plots

Performn Univariate Analysis on all ordinal categorical columns using pivot table, crosstab and frequency distributions

Based upon univariate analysis results, drop the columns that w ill not be useful for further analysis

Filter the data that is required for further analysis

Remove outliers from each column of our remaining dataset

Impute or drop the row s that has missing values

Perform Correlation Analysis

Remove columns that has high correlation and that deem redundant

Perform Bivariate Analysis on a combination of columns until w e identify the top 5 columns/variables that can be used to predict potential defaulters

End of analysis

# Understanding Data

- A dataset that contains loan application details, credit profile of loan application along with actual status of the loan repayment is provided.
- There are a total of 111 columns.
- Out of the 111 columns, there are 54 columns that has no data.
- No. of unique records in loans dataframe = 39717
- There is no column that gives the information of FICO score for each loan applicant directly.
- Dataset has nominal categorical, ordinal categorical and numeric variables.
- Dataset has input variables - which are filled by the loan applicant and output variables - which are possibly filled by LendingClub from various sources or after analysis.
- member_id and id are unique across the dataset and this means that no two rows are associated with each other.
- dti is ratio of debt to income. The highest value is ~30% which means 70% of annual income of the member is intact. Usually, if a member has dti < 40%, that member can be eligible for getting a loan. Hence, we can drop this column

# Cleaning Data

- Identify the list of columns that can be dropped based upon the following criteria:
  - A column that has a single value across all its rows will not be useful for our analysis.
  - A column that has a single value along with missing values will not be useful for our analysis.
  - Nominal categorical variables that has random values are not useful for our analysis.
  - Output Columns that are not useful for analysis.
  - Columns that have more than 30% of missing values.
  - Columns that are redundant.
  - Other columns that are not useful for analysis because of the business objective.

- From the dataset, it is evident that there ~83% of data is related to Fully Paid, ~2.8% of data is related to Current and ~14.2% of data is related to Charged Off loan status.

- Since the purpose of analysis is to identify the drivers that can predict potential loan defaulters, filter all the give data to have only the Charged Off loan status for further analysis.

# Data Manipulation

- Since each variable of given dataset is populated either by user input, by a financial function or from a public source and since the highest no. of missing values in a row is < 4% of the overall filtered data, we can remove them instead of imputing them with mean or median or mode as it may skew the data.

- Follow the steps below to manipulate existing data:
  - Extract month and year from date columns.
  - Convert percentage string columns to float columns
  - Create numerical columns for ordinal categorical columns

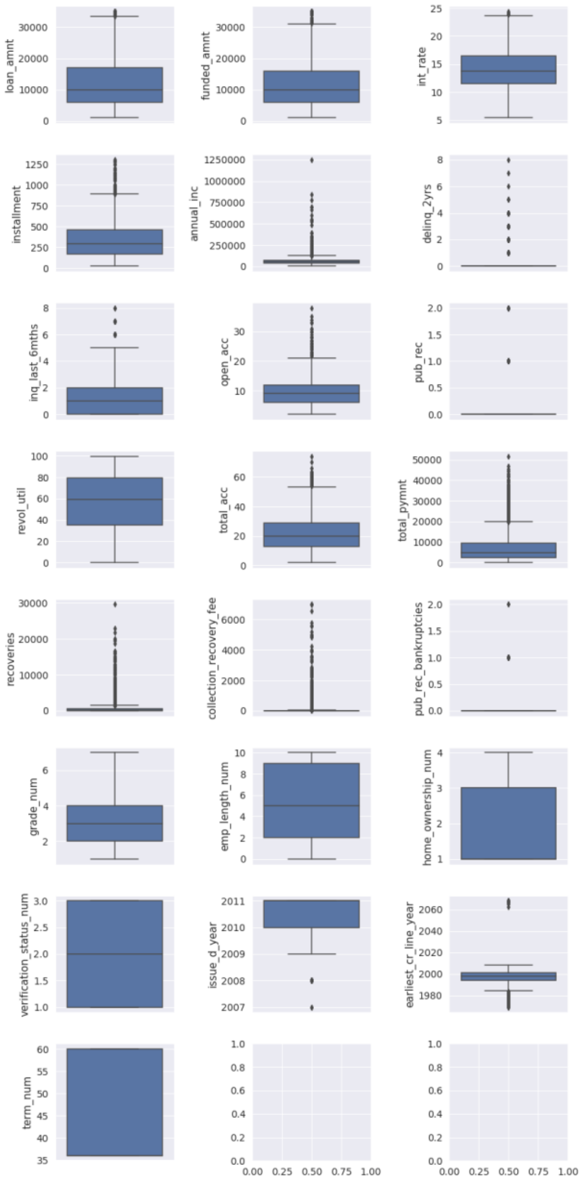
- Extracting required information using above steps will ensure that we followed data-driven, business-driven and type-driven approaches that resulted in new columns.

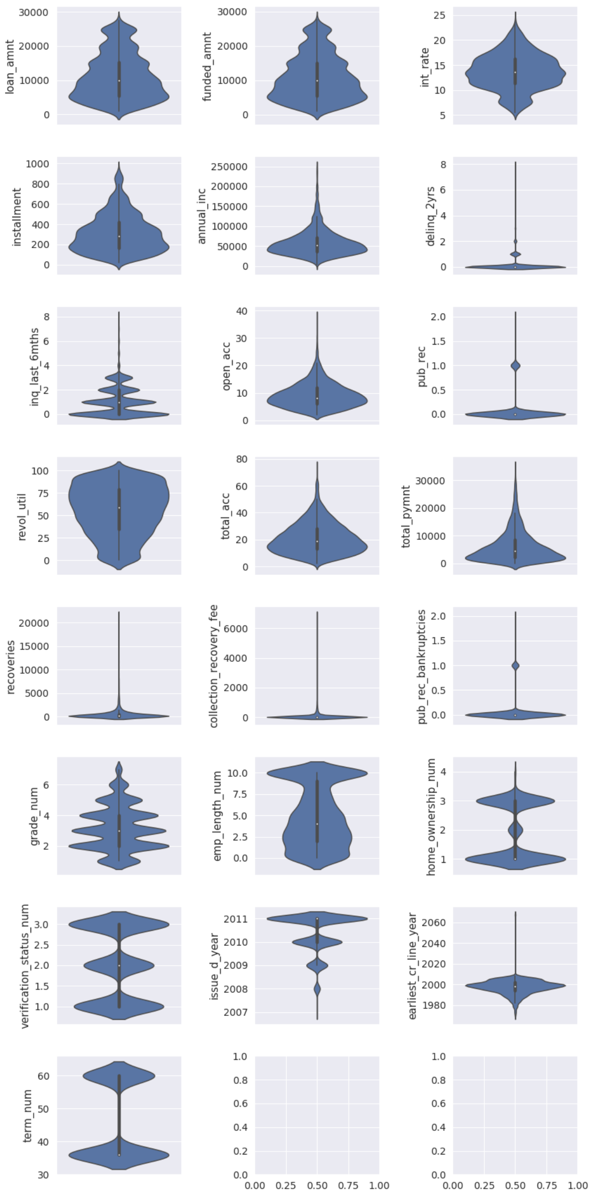
- Drop the redundant old columns from which all the required data is extracted.

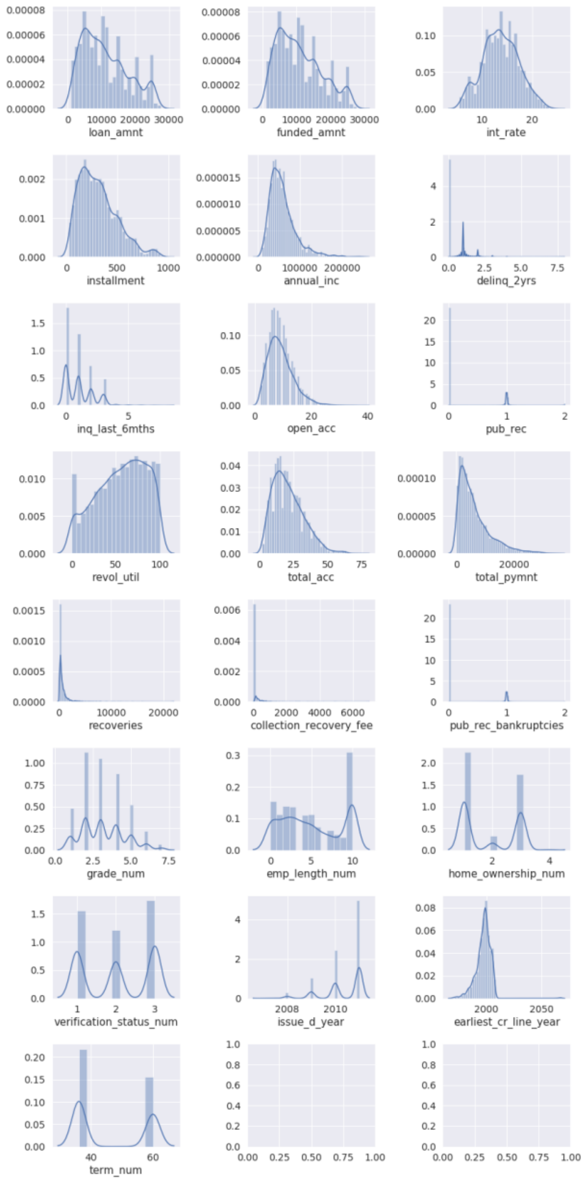
# Univariate Analysis – Numerical

- Plot Box, Violin and Distribution plots for all numerical variables.
- Observe the range of values that are present for each column, their mean, median, standard deviation, 25th and 75th quantiles, minimum and maximum.
- If there are any outliers, identify the important columns that may skew the data because of outliers.
- Remove the outliers on necessary columns by selecting only the quantiles that span across 2 or 3 sigmas away from mean.
- Outliers in less important columns can be dealt whenever required.

# Univariate Analysis – Numerical







# Univariate Observations– Numerical

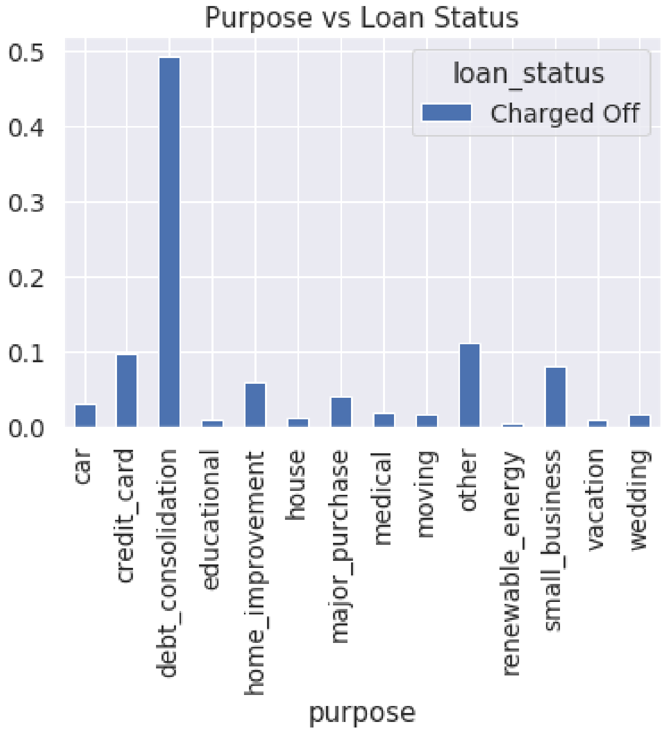
For all the loan applicants who have defaulted, below are the observations:

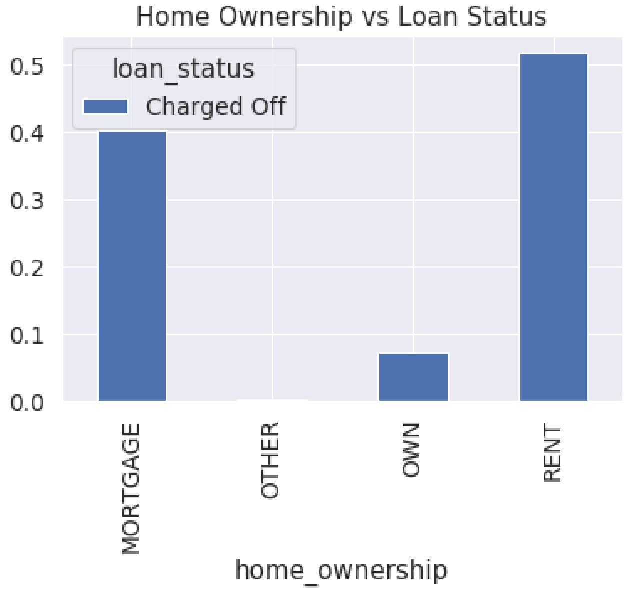
- Most of the loan amounts are in the range of $5500to $15250
- Most of the annual incomes are in the range of $37000to $72000
- Most of the interest rates are in the range of 11.28% to 16.29%
- Most of the instalment amounts are in the range of $167 to $417
- Most of the revolving credit utilization is in the range of 34% to 99%
- Most of the loan grades are in the range B to D
- Most of the employee length are in the range of 2 years to 9 years

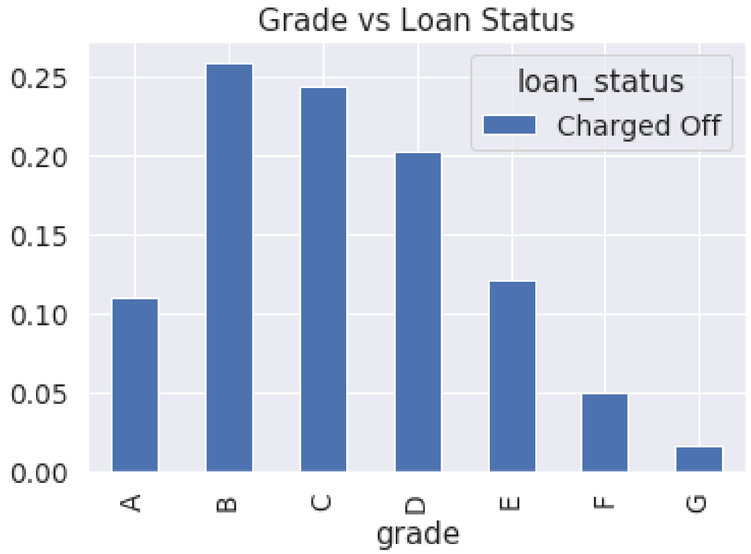
# Univariate Analysis – Categorical

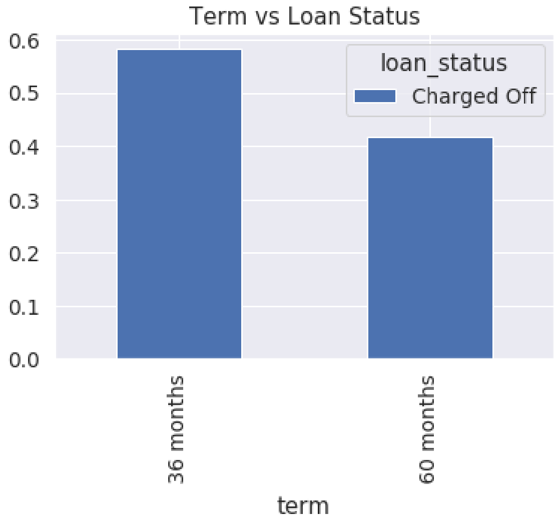
- Create pivot tables and plot frequency distribution plots for all categorical variables.
- Observe the frequency of values against "Charged Off" loan status for each variable.
- If there are any outliers, identify the important columns that may skew the data because of outliers.
- Remove the outliers on necessary columns by selecting only the quantiles that span across 2 or 3 sigmas away from mean.
- Outliers in less important columns can be dealt whenever required.

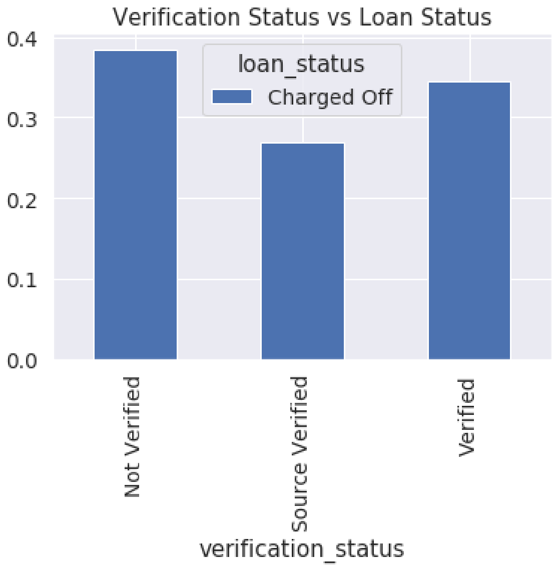
# Univariate Observations– Categorical


Purpose vs Loan Status


Home Ownership vs Loan Status


Grade vs Loan Status


Term vs Loan Status

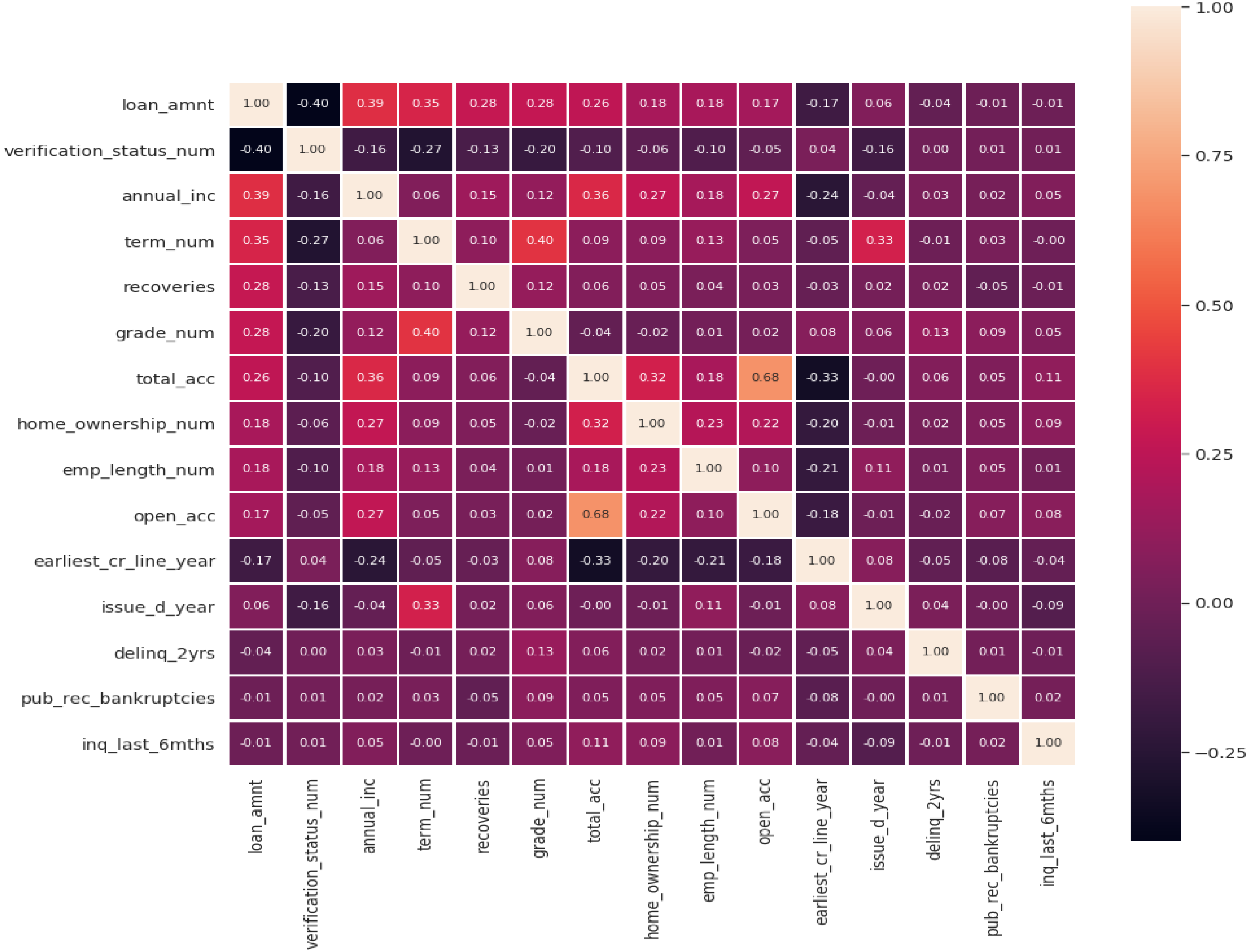

Verification Status vs Loan Status

# Univariate Observations– Categorical

For all the loan applicants who have defaulted, below are the observations:

- Debt Consolidation purpose has highest default rate
- Both **Rent** and **Mortgage** home ownerships have higher default rate
- Interestingly, **36 months (lower)** term has higher default rates
- Interestingly, **B, C and D grade** loans have higher default rates
- Employee length **10+ years followed by < 1 year** have higher default rates
- **"Not Verified" followed by "Verified"** loans have higher default rates
- **California followed by Florida and New York** states have higher default rates
- Loans **without any prior inquiries** have higher default rates
- Applicant having **6 to 10 open credit lines** have higher default rates
- Interestingly, applicants **who don't have any public record bankruptcies** have higher default rates
- Interestingly, default rates are higher **as year goes by - that is higher in December than in January**
- Applicants whose earliest credit line is **between 1998 to 2001** have higher default rates

# Bivariate Analysis Correlation



| | loan_amnt | verification_status_num | annual_inc | term_num | recoveries | grade_num | total_acc | home_ownership_num | emp_length_num | open_acc | earliest_cr_line_year | issue_d_year | delinq_2yrs | pub_rec_bankruptcies | inq_last_6mths |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| loan_amnt | 1.00 | -0.40 | 0.39 | 0.35 | 0.28 | 0.28 | 0.26 | 0.18 | 0.18 | 0.17 | -0.17 | 0.06 | -0.04 | -0.01 | -0.01 |
| verification_status_num | -0.40 | 1.00 | -0.16 | -0.27 | -0.13 | -0.20 | -0.10 | -0.06 | -0.10 | -0.05 | 0.04 | -0.16 | 0.00 | 0.01 | 0.01 |
| annual_inc | 0.39 | -0.16 | 1.00 | 0.06 | 0.15 | 0.12 | 0.36 | 0.27 | 0.18 | 0.27 | -0.24 | -0.04 | 0.03 | 0.02 | 0.05 |
| term_num | 0.35 | -0.27 | 0.06 | 1.00 | 0.10 | 0.40 | 0.09 | 0.09 | 0.13 | 0.05 | -0.05 | 0.33 | -0.01 | 0.03 | -0.00 |
| recoveries | 0.28 | -0.13 | 0.15 | 0.10 | 1.00 | 0.12 | 0.06 | 0.05 | 0.04 | 0.03 | -0.03 | 0.02 | 0.02 | -0.05 | -0.01 |
| grade_num | 0.28 | -0.20 | 0.12 | 0.40 | 0.12 | 1.00 | -0.04 | -0.02 | 0.01 | 0.02 | 0.08 | 0.06 | 0.13 | 0.09 | 0.05 |
| total_acc | 0.26 | -0.10 | 0.36 | 0.09 | 0.06 | -0.04 | 1.00 | 0.32 | 0.18 | 0.68 | -0.33 | -0.00 | 0.06 | 0.05 | 0.11 |
| home_ownership_num | 0.18 | -0.06 | 0.27 | 0.09 | 0.05 | -0.02 | 0.32 | 1.00 | 0.23 | 0.22 | -0.20 | -0.01 | 0.02 | 0.05 | 0.09 |
| emp_length_num | 0.18 | -0.10 | 0.18 | 0.13 | 0.04 | 0.01 | 0.18 | 0.23 | 1.00 | 0.10 | -0.21 | 0.11 | 0.01 | 0.05 | 0.01 |
| open_acc | 0.17 | -0.05 | 0.27 | 0.05 | 0.03 | 0.02 | 0.68 | 0.22 | 0.10 | 1.00 | -0.18 | -0.01 | -0.02 | 0.07 | 0.08 |
| earliest_cr_line_year | -0.17 | 0.04 | -0.24 | -0.05 | -0.03 | 0.08 | -0.33 | -0.20 | -0.21 | -0.18 | 1.00 | 0.08 | -0.05 | -0.08 | -0.04 |
| issue_d_year | 0.06 | -0.16 | -0.04 | 0.33 | 0.02 | 0.06 | -0.00 | -0.01 | 0.11 | -0.01 | 0.08 | 1.00 | 0.04 | -0.00 | -0.09 |
| delinq_2yrs | -0.04 | 0.00 | 0.03 | -0.01 | 0.02 | 0.13 | 0.06 | 0.02 | 0.01 | -0.02 | -0.05 | 0.04 | 1.00 | 0.01 | -0.01 |
| pub_rec_bankruptcies | -0.01 | 0.01 | 0.02 | 0.03 | -0.05 | 0.09 | 0.05 | 0.05 | 0.05 | 0.07 | -0.08 | -0.00 | 0.01 | 1.00 | 0.02 |
| inq_last_6mths | -0.01 | 0.01 | 0.05 | -0.00 | -0.01 | 0.05 | 0.11 | 0.09 | 0.01 | 0.08 | -0.04 | -0.09 | -0.01 | 0.02 | 1.00 |

# Bivariate Analysis

- Plot correlations for all numerical columns
- Use only one of the highly correlated columns in a group - as using all may be redundant

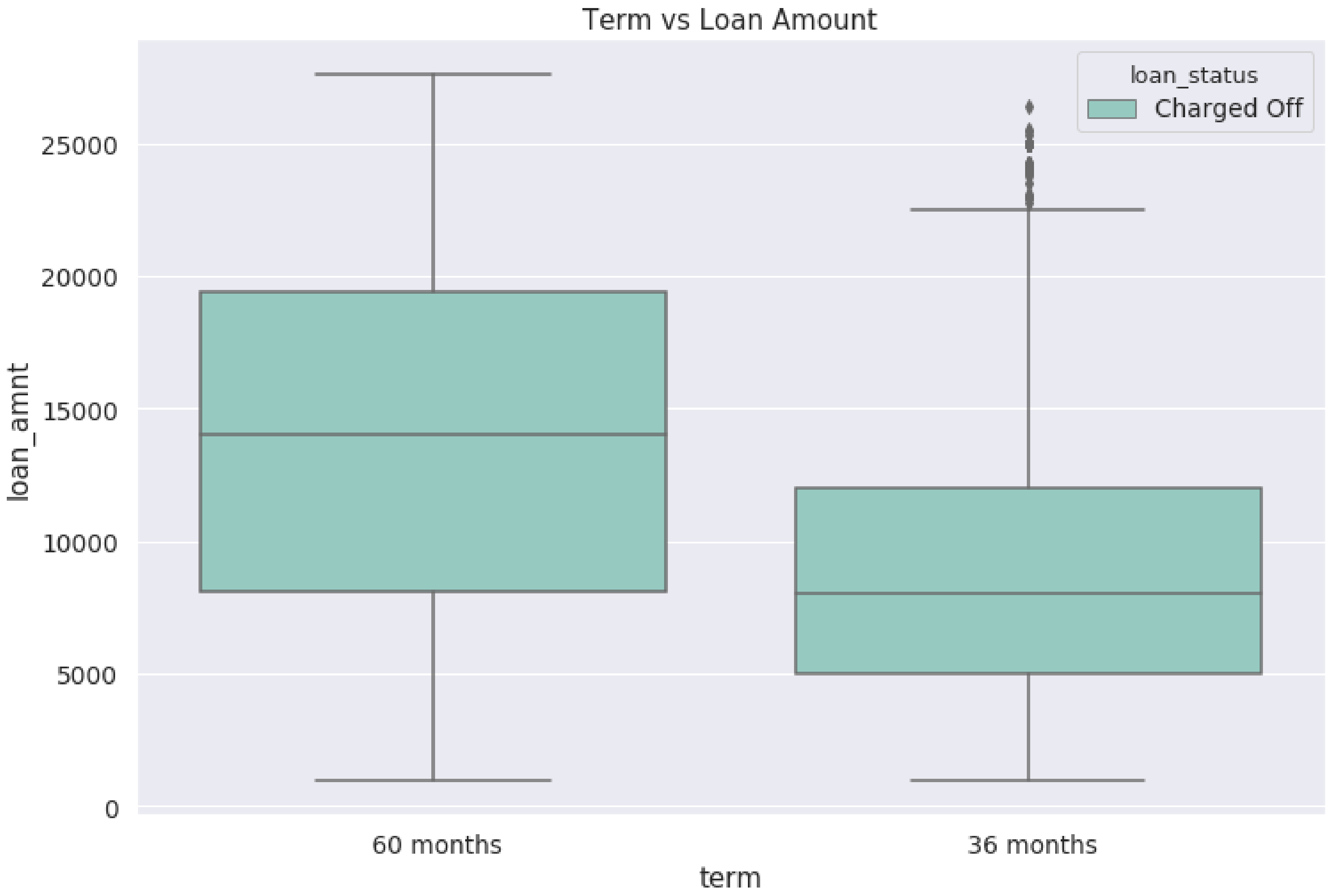
After plotting the correlations for all numerical columns, below are the observations:

- loan_amnt, funded_amnt, installment, total_pymnt are highly correlated and only one of them can be used.
- int_rate and grade are highly correlated and only one of them can be used.
- recoveries and collection_recovery_fee are highly correlated and hence, only one of them can be used.
- pub_rec and pub_rec_bankruptcies are highly correlated and hence, only one of them can be used.
- revol_util and int_rate are slightly correlated which means that higher revol_util will have higher interest rate too.
- emp_length and int_rate are not at all correlated - which means that int_rate will be higher for people with lesser emp experience.
- open_acc and verification_status has no clear correlation with any of the columns.

After removing redundant columns, we will be left with 25 columns for further analysis.

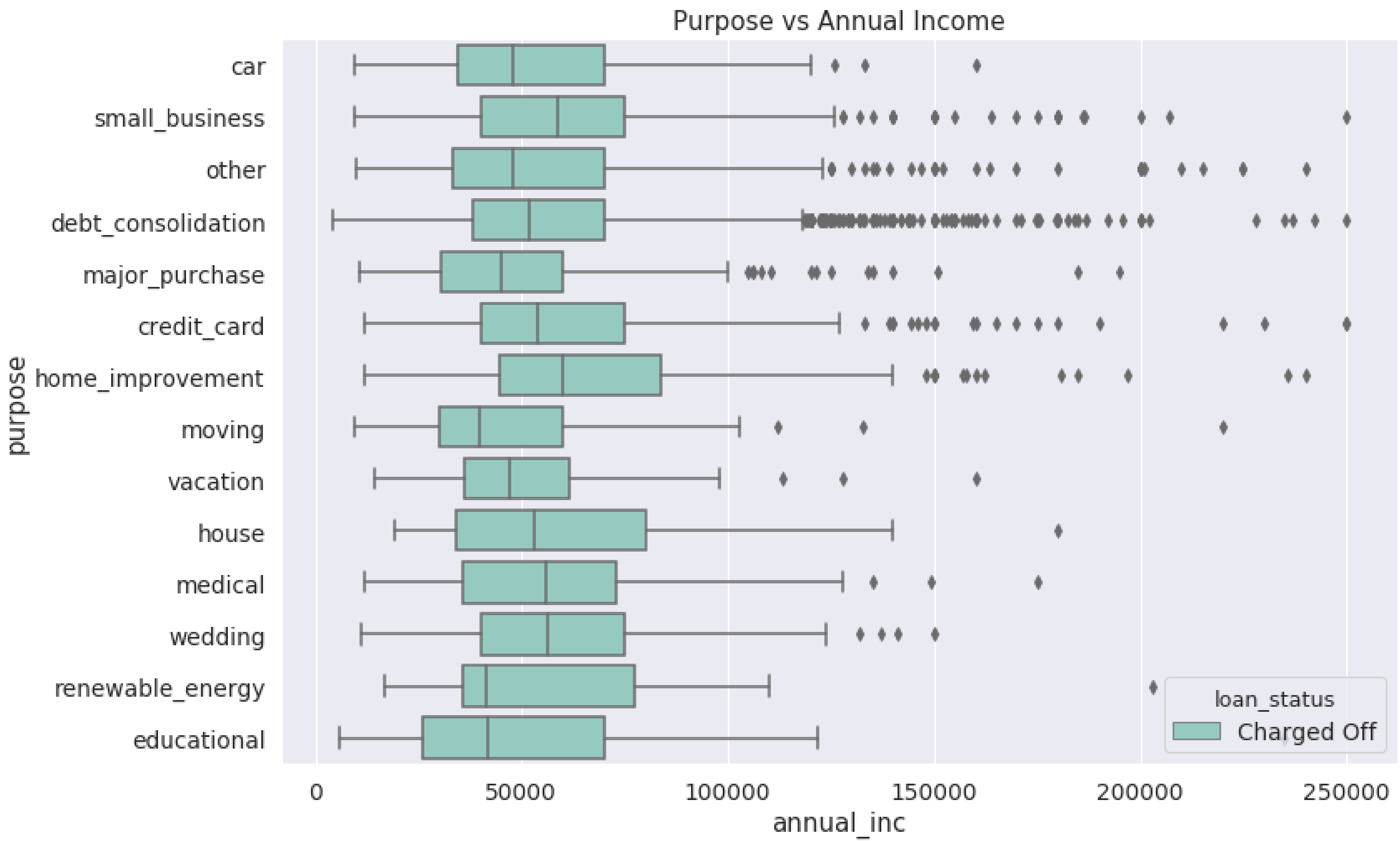
# Bivariate Analysis on Loan Status

General trend of Term vs most of the numerical variables is like below which means that higher the term, higher the risk of a loan turning into default.



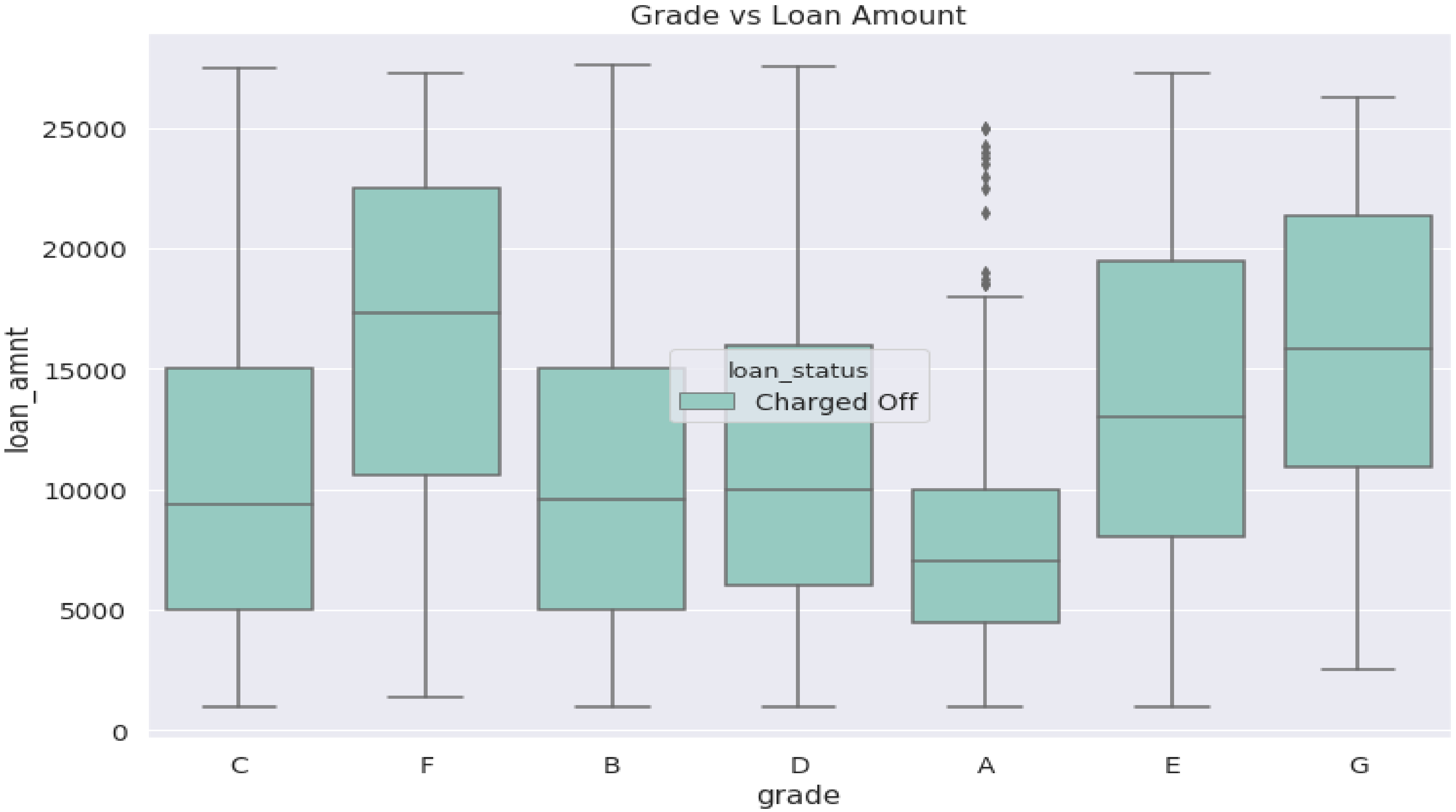
# Bivariate Analysis on Loan Status

General trend of Purpose vs most of the numerical variables is like below which means that the impact of purpose of loan is not very significant.



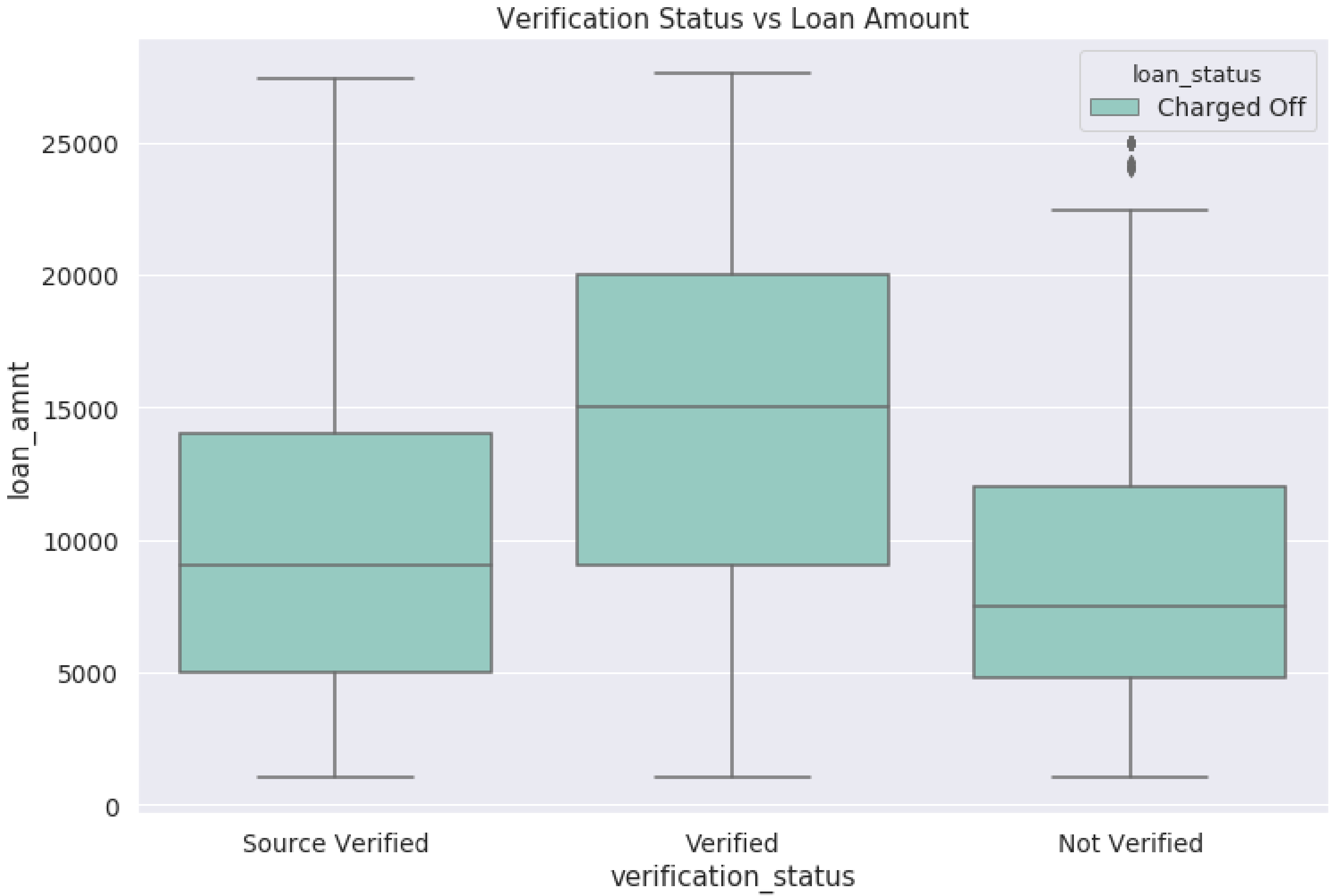
# Bivariate Analysis on Loan Status

General trend of Grade vs most of the numerical variables is like below which means that loan grades E, F, G have higher chances of turning a loan into default



# Bivariate Analysis on Loan Status

General trend of Verification Status vs most of the numerical variables is like below which means that most of the defaulted loan applicants have "Verified" and "Source Verified" status



# Bivariate Analysis on Loan Status

Other key observations are mentioned below:

- Defaulted loan applicants annual income range is 30000 - 80000 dollars for most of the loans
- Defaulted loan amount range is 4000 to 22500 dollars for most of the loans
- "Rent" home ownership category has higher default rates
- Employee Length range of 4 - 5 has higher default rates
- Grades E, F, G loan grades with respect to loan amount and B, C, D loan grades with respect to loan purpose
- Different purposes have different risks of being default. Small Business, Debt Consolidation and Home Improvement have top default rates

# Conclusions

The top driving factors that can identify potential loan defaulters are:

- **Grade** - E, F, G loan grades with respect to loan amount and B, C, D loan grades with respect to loan purpose
- **Term** - Higher the term, higher the risk of a loan becoming default except with respect to verification status
- **Annual Income** - All defaulted loan applicants have annual income in the range of 30000 - 80000 dollars
- **Loan Amount** - All defaulted loan applicants have sanctioned loan amount in the range of 5000 - 22500 dollars
- **Home Ownership** - Most of the defaulted loan applicants are living on "Rent"
- **Verification Status** - Most of the defaulted loan applicants have "Verified" and "Source Verified" status

# Recommendations

- Since Verification Status is one of the driving factors, ***ensure*** that the loan applicant's ***income and the source of income are thoroughly verified***

- Since loan applicants living on "Rent" are already paying significant amount of their income towards their rent, try ***adding "Rent" as a debt while calculating the debt-to-income ratio***

- Since lower loan amounts are at high risk of being defaulted, be a little ***stringent on the "purpose" of loan***. Avoid giving loans for Debt Consolidation and Small Business. For all purposes, ***give loans if loan applicant's annual income is higher than 80000 dollars***

- Try to have ***multiple and flexible loan payment term*** options - in the range of 24 months to 60 months

- Even if a loan has good grades assigned to it, ***bad purpose of the loan might push a loan to default***.