Lending Club Case Study

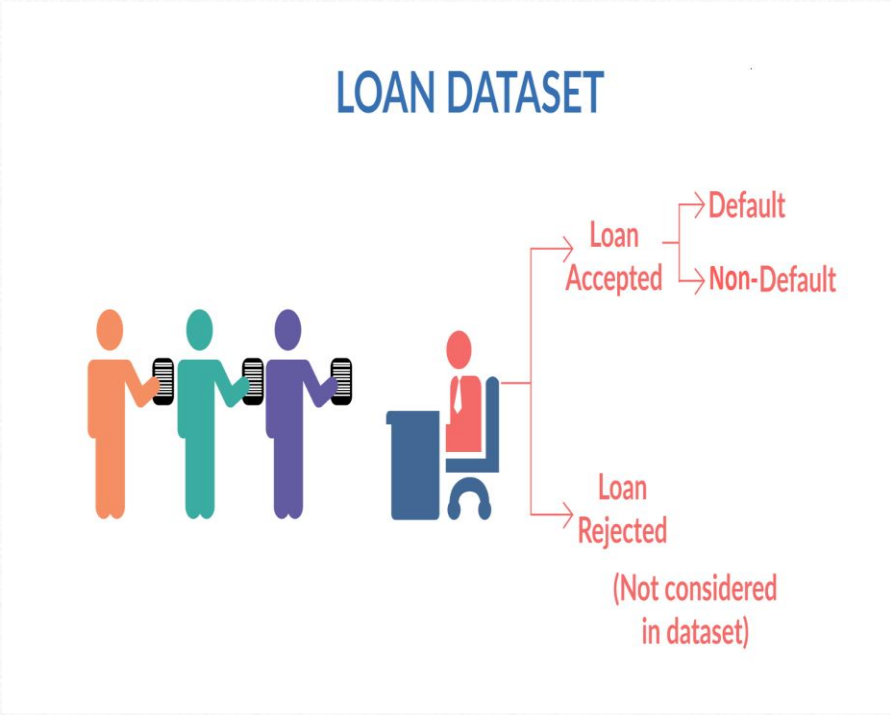
Submitted by:
Nithyashree Ravi

Agenda

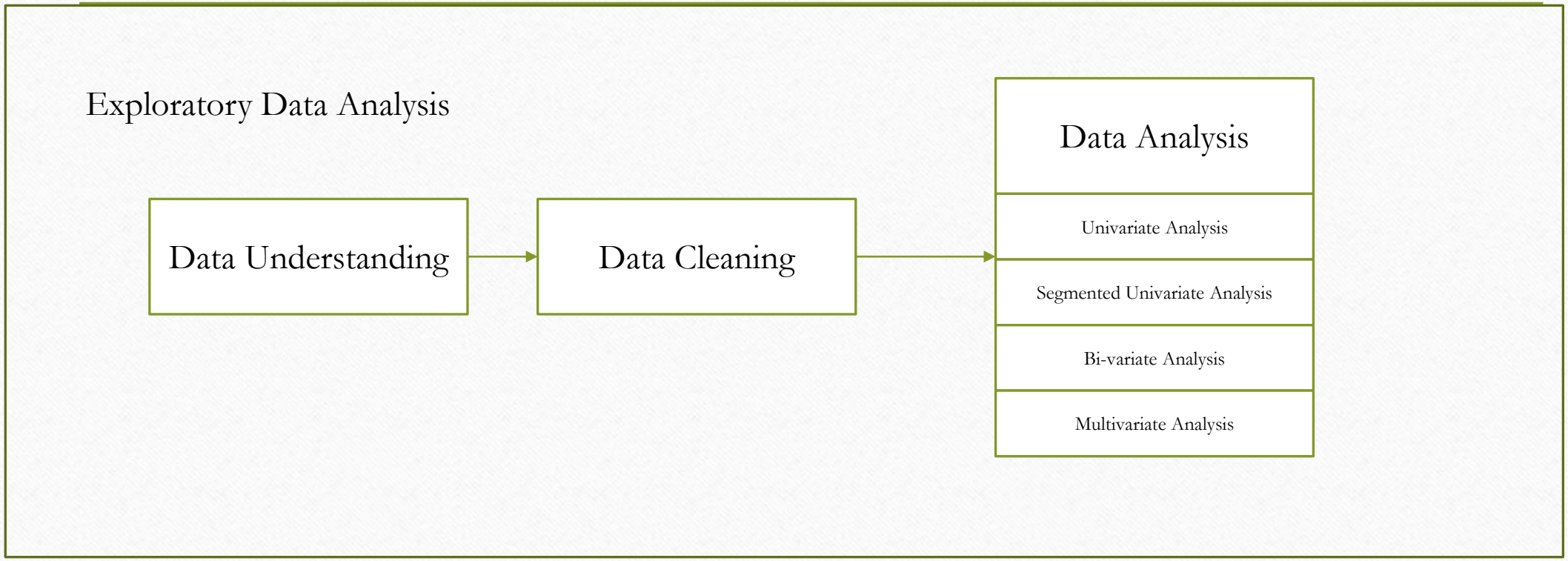
- Introduction
- Architecture
- Data Understanding Before Cleaning
- Data Cleaning
- Data Analysis
    1. Univariate Analysis
    2. Segmented Univariate Analysis
    3. Bivariate Analysis
    4. Multivariate Analysis
- Conclusion

Introduction

- When a person applies for a loan, it can either be accepted or rejected. If the loan is accepted, there can be defaulters (charged off) or the ones who have fully paid the loan.
- Our task is to predict the loan defaulters.
- The reason why we are doing this is because if one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss for the company.
- Identification of such applicants using EDA is the aim of this case study.
- To get the driver variables we will perform data analysis split by steps.



Architecture



Data Understanding Before Cleaning

- Data can have many null values, noise and outlier variables, for this to not impact while we train the machine learning model, we have to remove these and provide a cleansed data to the model.
- We can get a clear picture of the data by finding it's mean, datatype, median, mode, standard deviation, minimum value and maximum value.
- What 25% of data contains and what 50% of data contains. With this we can get a clear picture of what the data is.
- After having a rough estimate of the values in the data, we will check for unwanted columns.
- Unwanted columns are those which are empty (stored as NaN (Not a Number) in python) in excel csv.
- We must make use of median more then mean as mean before cleansing the data would not have a proper value.

	loan_amnt	funded_amnt	funded_amnt_inv
count	39717.000000	39717.000000	39717.000000
mean	11219.443815	10947.713196	10397.448868
std	7456.670694	7187.238670	7128.450439
min	500.000000	500.000000	0.000000
25%	5500.000000	5400.000000	5000.000000
50%	10000.000000	9600.000000	8975.000000
75%	15000.000000	15000.000000	14400.000000
max	35000.000000	35000.000000	35000.000000

Data Cleaning

- Now we will do the data cleaning process.

This is divided into two types:

-> Column Cleaning

-> Row Cleaning

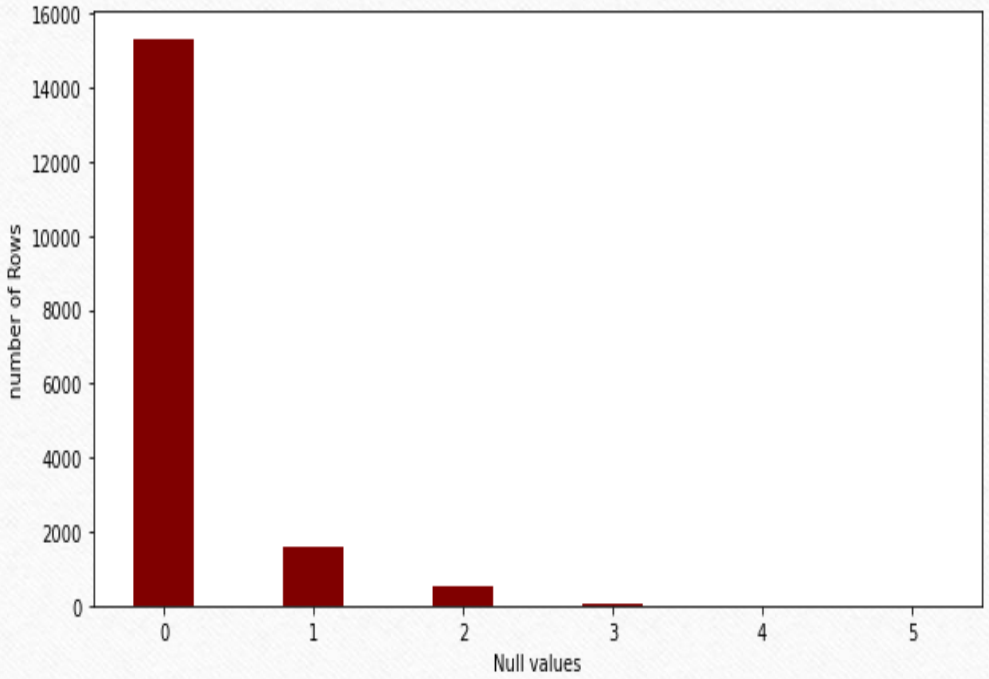
1. Column Cleaning :

- We have to first check if all values are null values in the column.
- Then remove the null columns.
- There are still columns which are all not null values but having higher number of null values.
- If there are more number of nulls in a specific column, then remove it.

Data Cleaning

2. Row Cleaning

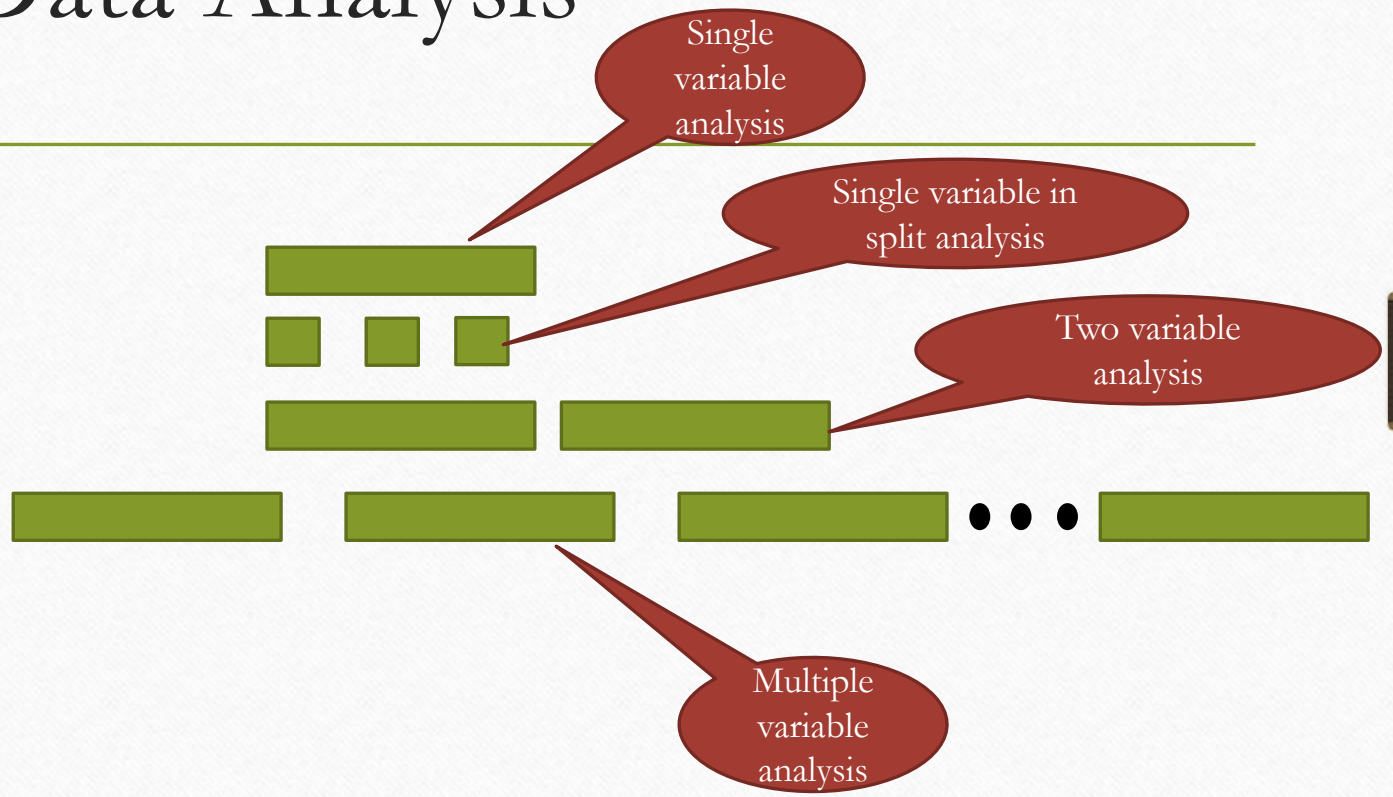
- We will check the number of null values in a particular row.
- Since we have removed significant number of columns having higher null values, it is assumed that we will have less number of null rows.
- For this dataset, there are rows containing both 1 and 2 null values.
- But if we remove the whole row we will have less data.
- So we again redo the column analysis and remove columns with a different threshold value.
- This way all the null values are removed effectively.



Data Analysis

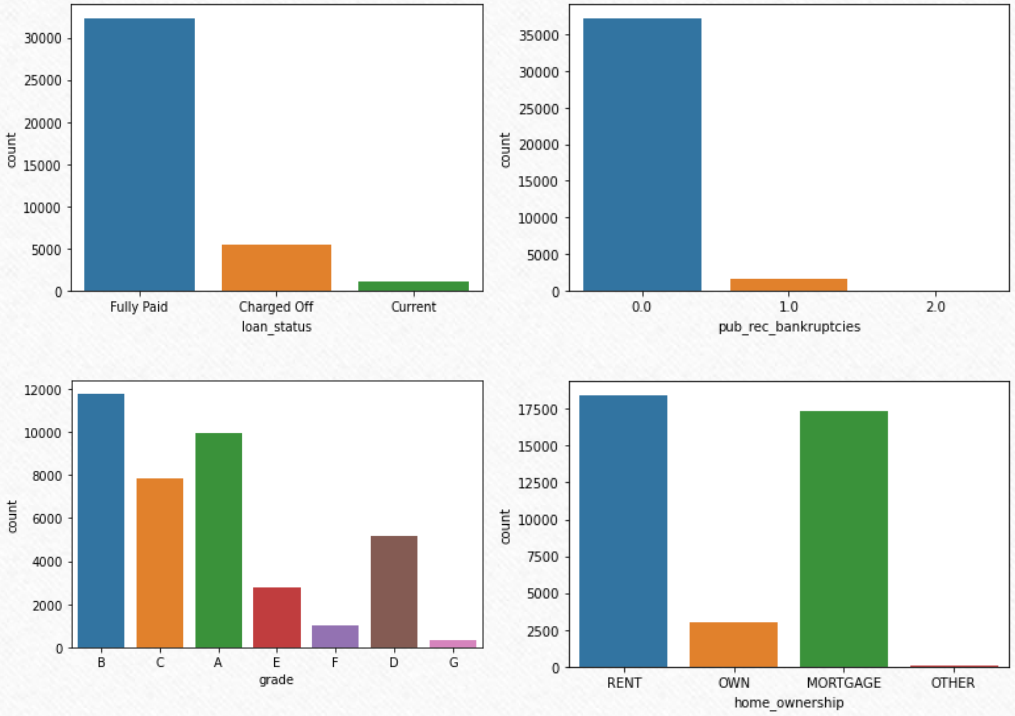
Data Analysis comprises of 4 types:

- Univariate Analysis
- Segmented Univariate Analysis
- Bi-variate Analysis
- Multivariate Analysis



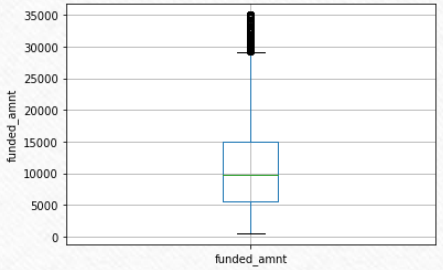
Univariate Analysis

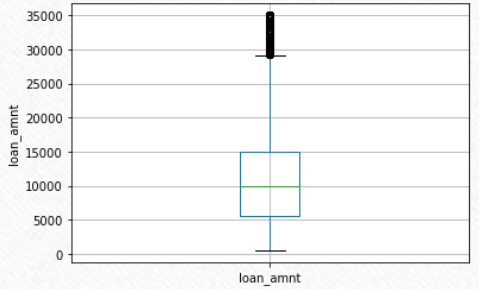
- The predicting variable has less values of Charged off which is the deciding factor of the dataset.
- There is a class imbalance in the data that is, we will actually be needing more data which are of status Charged off.
- We will separate categorical variables and numerical variables.
- After separating we will perform the count of column analysis for categorical variables.

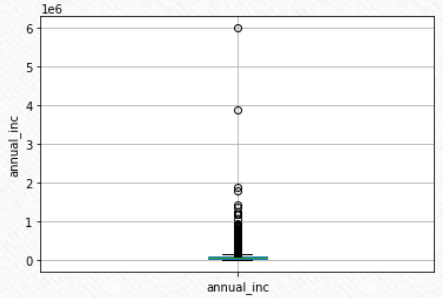


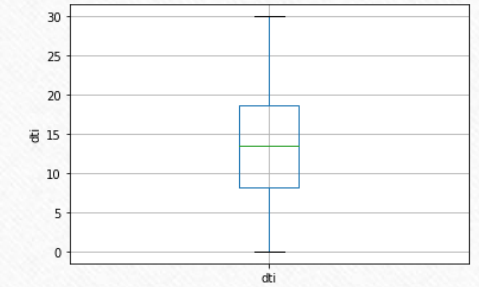
Univariate Analysis

- We will now be performing univariate analysis for continuous variables.
- We can notice that there are many outliers in the dataset.
- But if we remove them then there will be very less number of data for the model.
- So we will consider columns with less outliers.
- We can find that loan amount and funded amount plays an important factor in the dataset.
- Box plots are used to find the quartile ranges and check for outliers.
- Annual income has many outliers, but it plays a major factor in loan approval, if we remove it, then we will have less data.



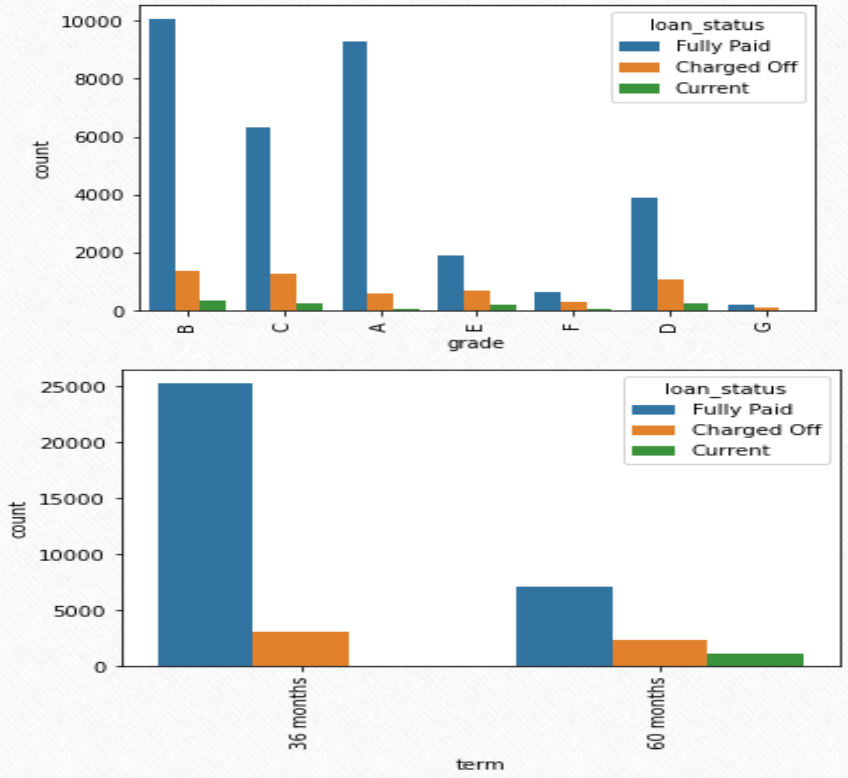






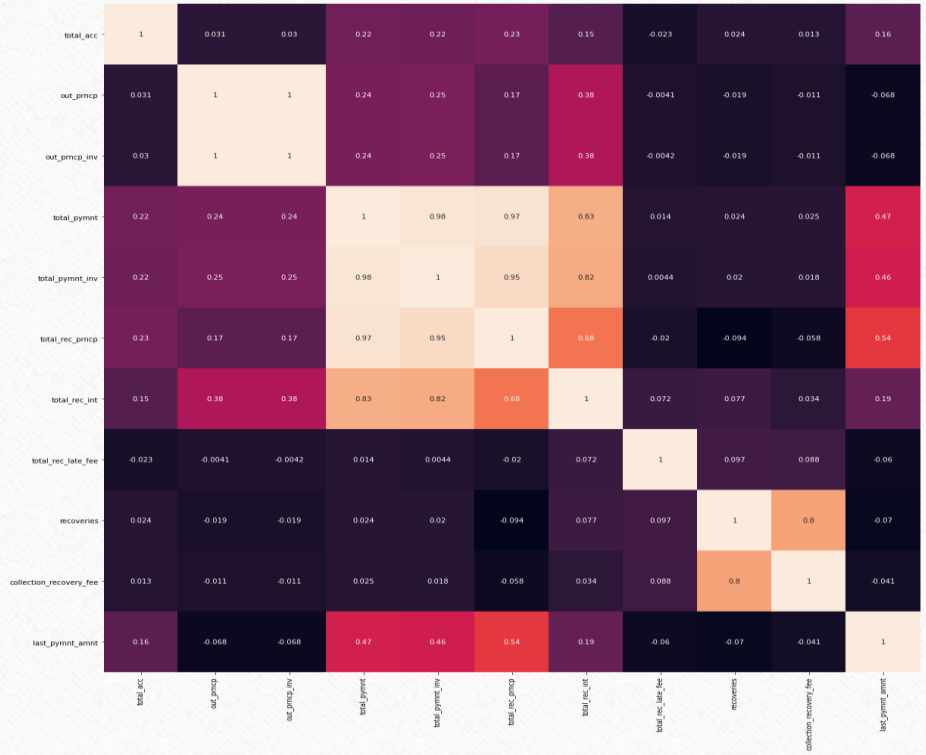
Segmented Univariate Analysis

- We will now find the segmented univariate analysis with respect to the loan status.
- We are checking the count of variables separately to understand the data even better.
- We can notice a clear class imbalance in segmented data analysis.
- Charged off is more for lesser number of months this is because there is a huge chance of the loan not being paid in lesser months.
- For term, there is a higher possibility of C and D getting charged off. Data of G and E are very less so we cannot arrive at any conclusion.
- A and B have higher values of loan being paid because of the higher value of Fully Paid status.



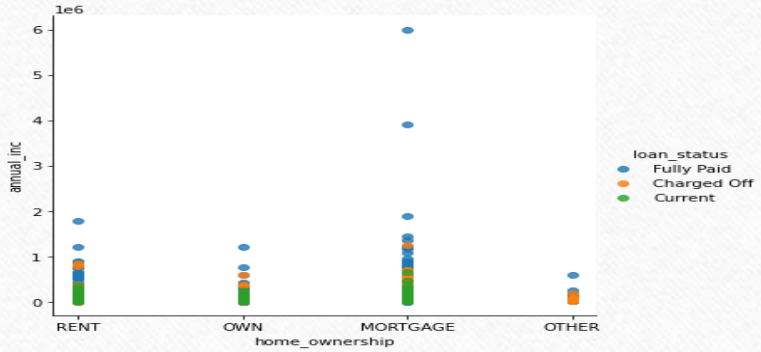
Bivariate Analysis

- Bi variate analysis involves comparison of 2 variables.
- The best way is the find the correlation between the variables.
- Correlation should involve numerical values accordingly.
- We can conclude negative, positive and neutral correlation between the two variables based on its values.
- We are seeing a positive correlation between Funded amount and total correlation.
- Most of newer column creations will happen in this process.



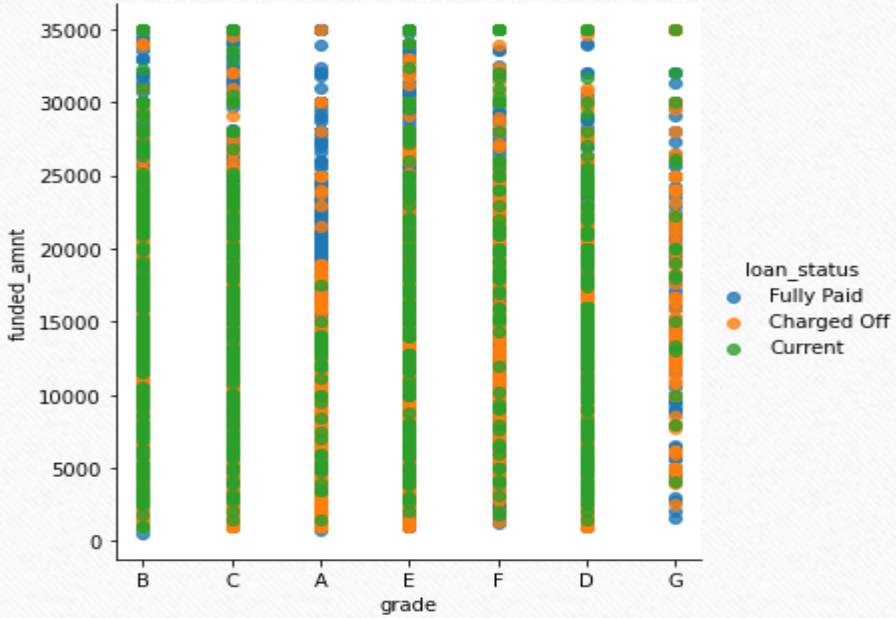
Multivariate Analysis

- Multivariate Analysis involves analysing using 3 or more variables simultaneously. We can analyse using 3 variables maximum as we will come to a better conclusion only when we compare 3 variables.
- We can conclude that home ownership as own and higher annual income implies a greater chance of the loan being paid.
- Home ownership as mortgage and less annual income has a greater chance of the loan status as default (charged off).
- Based on lesser term, the loan status of being charged off is more comparatively.

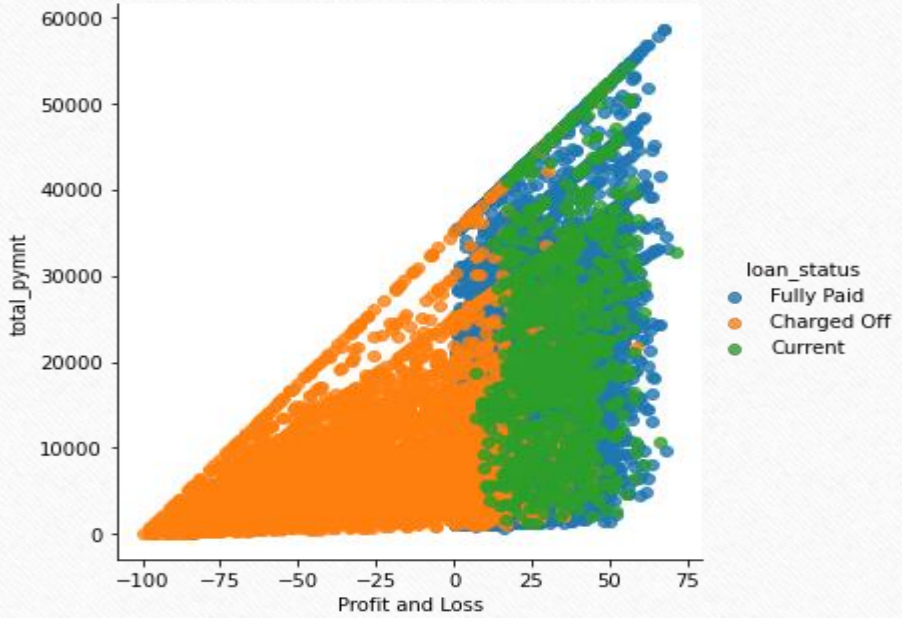


Multivariate Analysis

A and B have more number of data which is fully paid. D and C are likely to be charged off.



Profit is high when total payment amount is high and the status is likely to be fully paid.



Conclusion

- The most impacting variables for default values of loan are Profit and Loss, total payment, funded amount, home ownership, term, grade, annual income.
- This concludes that a person with Grade B or A, home status as own, State as Texas is likely to fully pay the loan. Higher annual income of a person concludes that the person will be paying the loan fully. Whereas, there is a higher chance of a person who is in places like CA, MD, MI as address state, Grade as C or D and home ownership status as Mortgage to be charged off (Loan defaulters).
- Impacting Variables to be charged off are Profit and Loss, total payment, funded amount, home ownership, term, grade, annual income.
- Based on the change in these values, the variables will behave accordingly.