Exploratory Data Analysis : Loan Default Prevention

OBJECTIVE:
To identify the driving factors for loan defaults from the data of past loan applications

APPROACH OF ANALYSIS to identify the driving factors for loan defaults and thereby minimize the risks in future:

On the past data of the lending club, Exploratory Data Analysis had been carried, mainly (1) Univariate Data Analysis, (2) Segmented Univariate Analysis and (3) Bivariate Analysis.

Since the target is to identify factors linked with loan default, the "**loan_status"** column is chosen as target column and **"Charged Off"** as target variable

To find the number of Charged Off borrowers in the Loan Status, Frequency Distribution Plots, Cross Tables, Grouped Bar Charts, Scatter Plots, Correlation Analysis are used

Data Cleaning

Columns with missing data more than 25 % are dropped

After removing the columns, rows with missing data are dropped, since this represent few percentage of the total dataset

Numeric data with percentage characters are converted into float by stripping the ‘%’ character

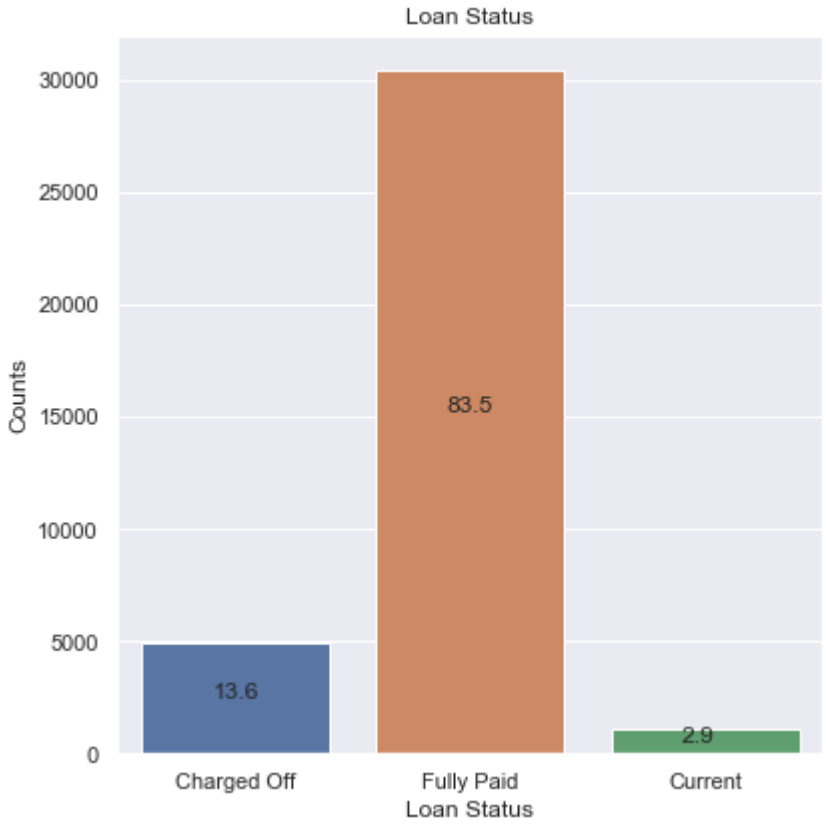
Columns with constant value are not considered in the analysis

Columns like url,…. are removed based on the business understanding

In the Original dataset, there are 111 columns and 39717 rows

The cleaned dataset contains:
- 53 Columns and 36431 Rows
- There are 33 numerical variables and 20 object datatypes

Loan Status: Fully Paid vs. Default

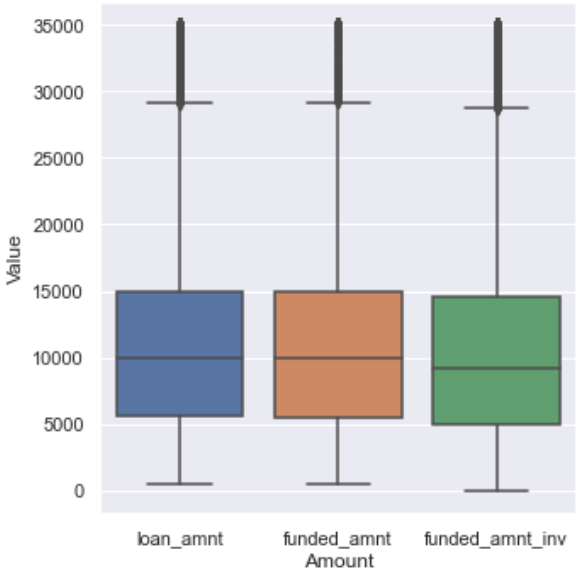


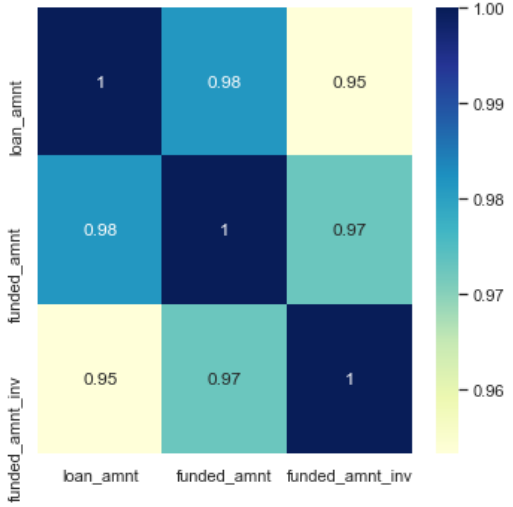
The Frequency Distribution plot on the column “Loan Status” clearly shows that the data is highly imbalanced

83.5 % loans are fully paid

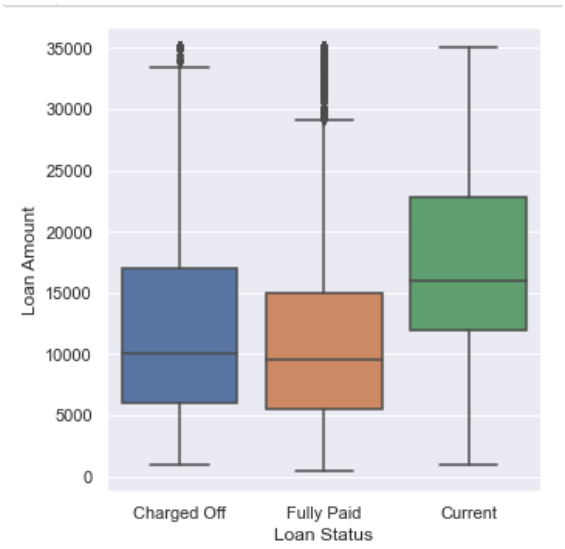
Only 13.6 % loans are Charged Off

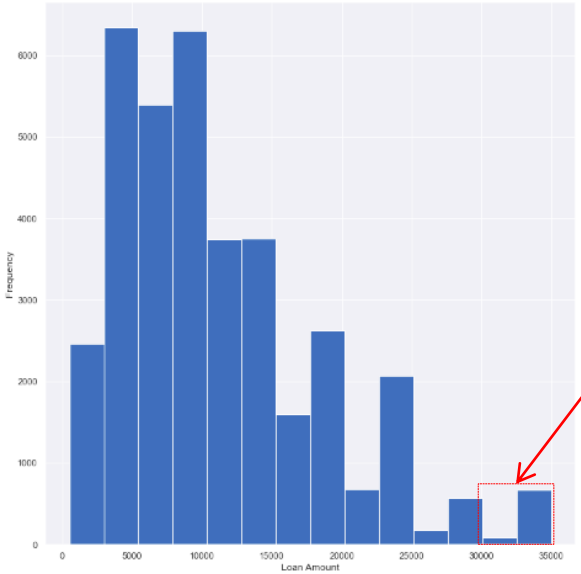
Loan Amount Value on Loan Default





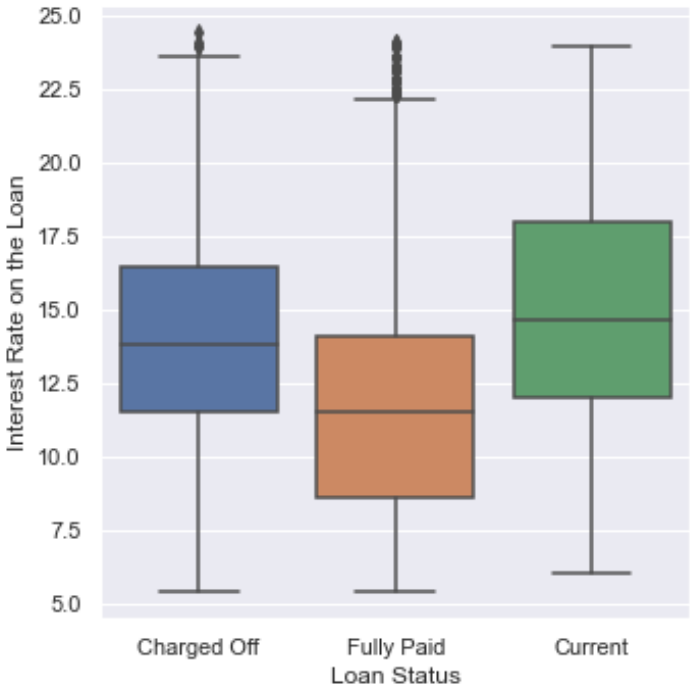
- From the Grouped Box plot, the distribution of values of 3 variables namely, Loan Amount, Funded Amount and the Funded Amount by the Investors are almost similar.
- Also these 3 variables are highly correlated as seen in the Pearson Correlation Matrix.
- So, for the further analysis only the column "Loan Amount" is considered.

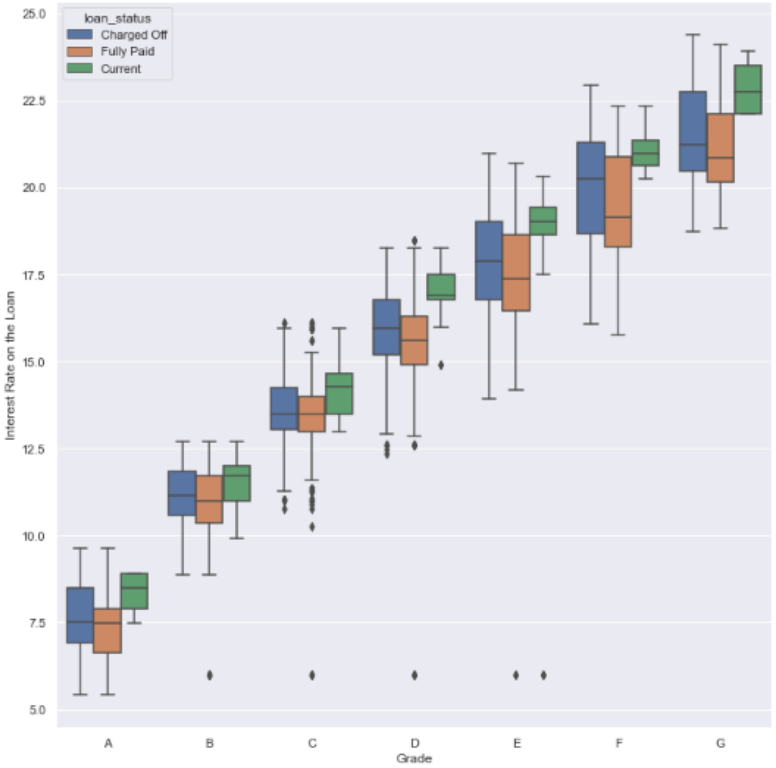




Although the median of the Loan Amount for Fully Paid and Charged Off are comparable, **when loan amount exceeds 30000 more likely it could be default**

Loan Grading & Interest Rate vs. Loan Default





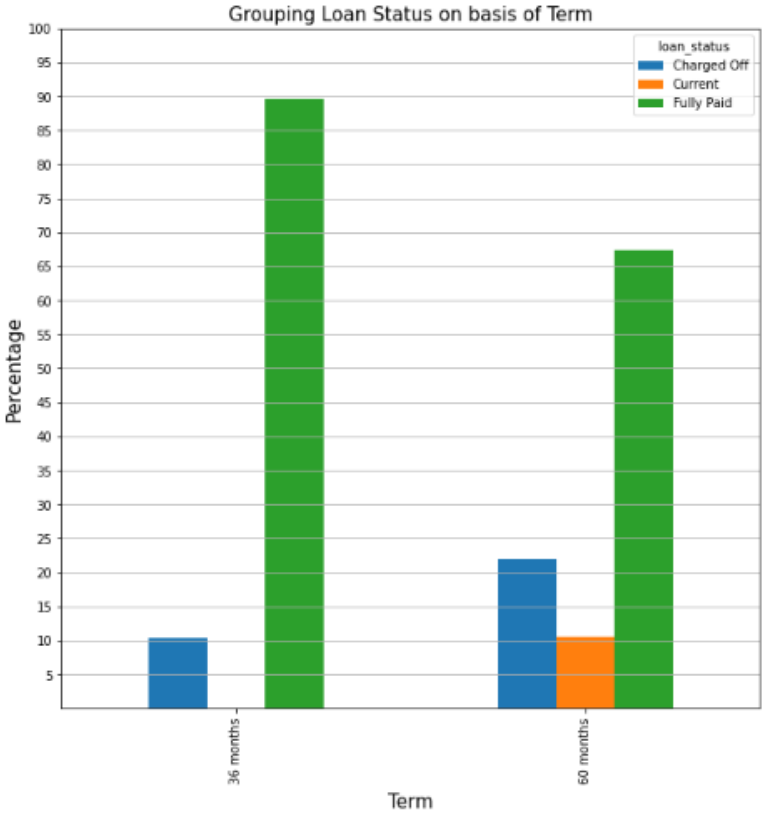
Cross Table Comparing Loan Status (in Number of Percentage) against Loan Grades

Grade / Loan Status	A	B	C	D	E	F	G
Charged Off	5.59%	11.35%	15.73%	20.11%	24.45%	**29.51%**	**31.21%**
Current	0.38%	2.87%	3.36%	4.35%	6.38%	6.91%	5.37%
Fully Paid	94.03%	85.77%	80.90%	75.54%	69.17%	63.57%	63.42%

At interest rates higher than 20-22.5% chances for loan default. Loan grades F and G are associated with risky applicants, where interest rates are also high and as per the loan grading more percentage of loan defaults are at grades F and G.

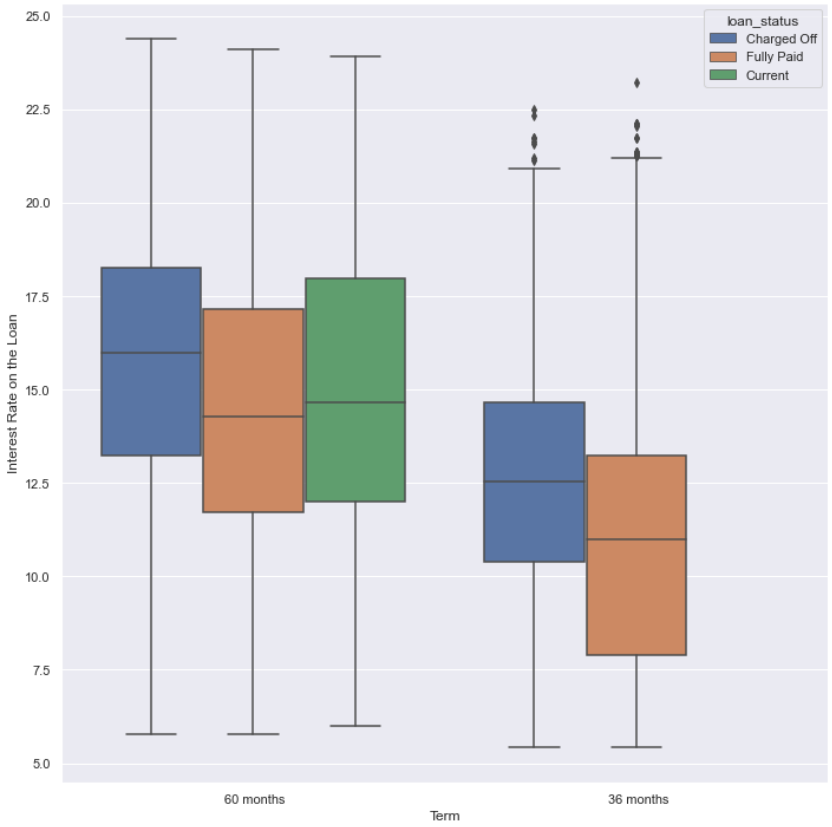
Loan Term Period vs. Loan Default

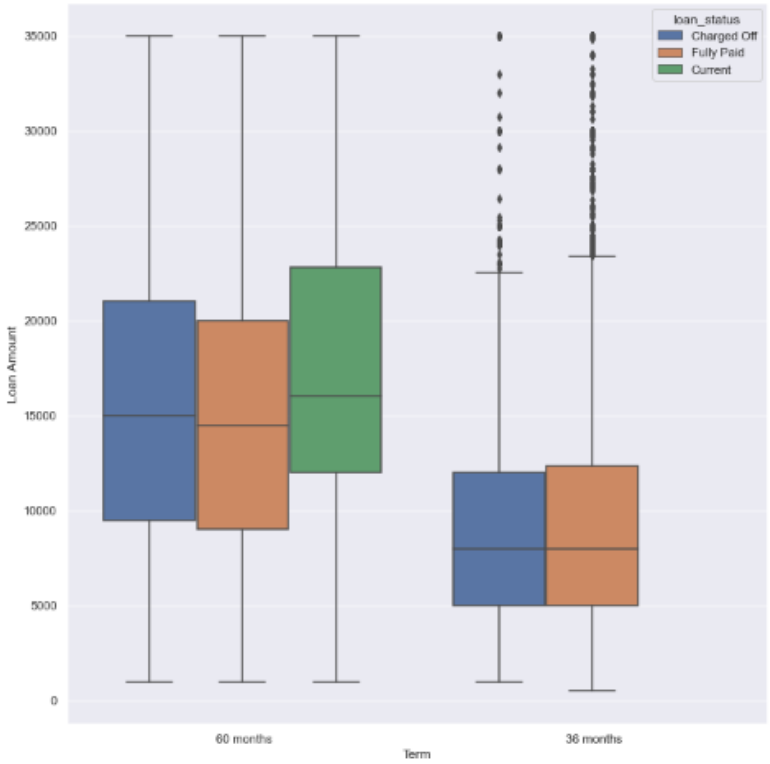
Term / Loan Status	36 months		60 months	
	Count	Percentage	Count	Percentage
Charged Off	2729	10.35%	**2213**	**22.01%**
Current	0	0.00%	1066	10.60%
Fully Paid	23646	89.65%	6777	67.39%



Loans with 60 months term are more likely to default

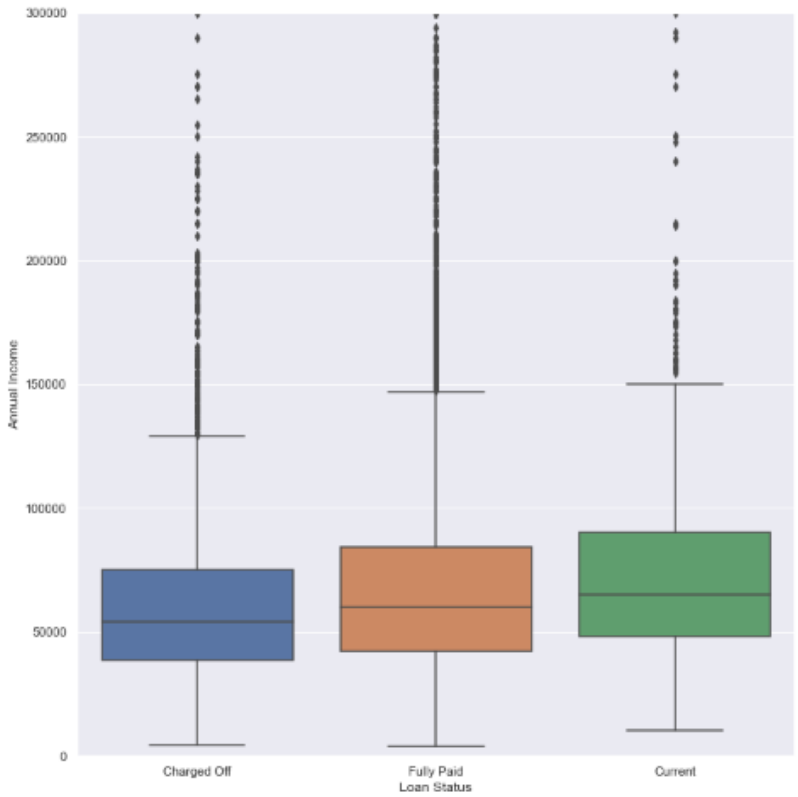
Loan Term Period vs. Loan Default

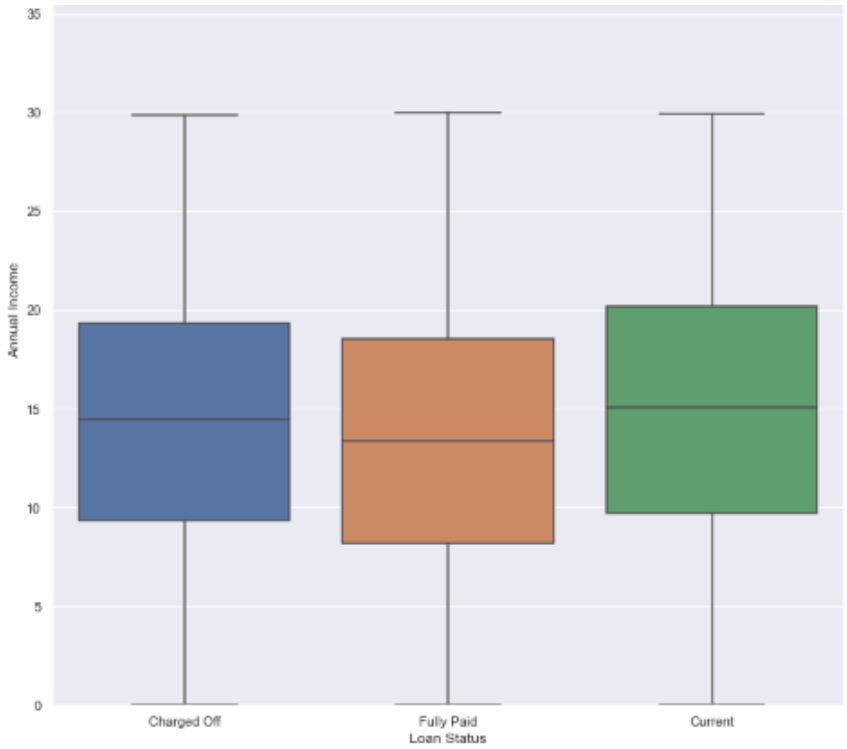




- Loans with 60 months term are more likely to default
- Loans with 60 months term are with high interest rates
- Loan amount of the 60 month term period are higher than that of 36 months

Annual Income of the Borrower on Loan Status

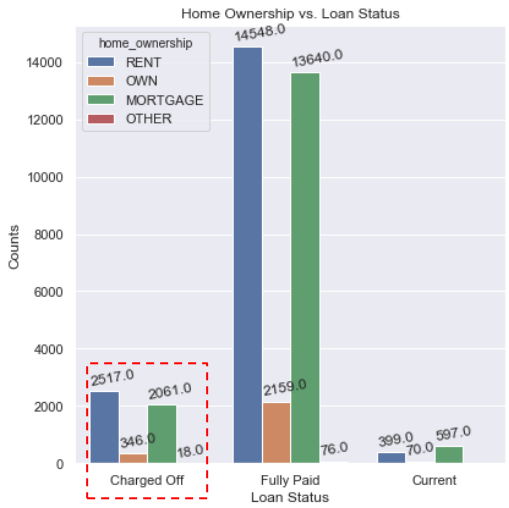




From the Grouped Barplot, it reveals no significant trend observed in Loan Status categories for Annual Income and Debt to Income Ratio

Home Ownership of the Borrower on Loan Status

HOME OWNERSHIP / LOAN STATUS	MORTGAGE		OTHER		OWN		RENT	
	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
Charged Off	**2061**	**12.65%**	18	19.15%	346	13.44%	**2517**	**14.41%**
Current	597	3.66%	0	0.00%	70	2.72%	399	2.28%
Fully Paid	13640	83.69%	76	80.85%	2159	83.84%	14548	83.30%



Among the defaulters, those in Rental house are higher than those in own house.

This suggest that borrowers with own home are less likely to default

Important Observations from the Analysis

1. Loan Status in this dataset clearly shows that the data is highly imbalanced. 83.5 % loans are fully paid and only 13.6 % loans are Charged Off.
2. When loan amount exceeds 30000 more likely it could be default
3. At interest rates higher than 20 - 22.5 % chances for loan default.
4. Loans with 60 months term are more likely to default.
5. Loans with 60 months term are with high interest rates
6. Borrowers with own home are less likely to default

Other Analysis form this Dataset

(Other variables are also analyzed in this dataset, however they doesn’t seem to provide valuable insights for understanding loan default from Exploratory Data Analysis)

Few Results from Other Analysis form this Dataset

Debt Consolidation as purpose given by borrower have higher number among the other purposes in Charged Off