Lending Case Club Study

Goals of data analysis:

- The main objective is to be able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

- Perform an analysis to understand the driving factors (or driver variables) behind loan default, i.e. The variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment.

Submitted By

Arjun Singh

Paryanshu Sourav

Method or techniques used to perform the analysis

1. Data Sourcing
2. Data Cleaning
3. Univariate Analysis
4. Bivariate analysis
5. Derived Matrices

Data Sourcing

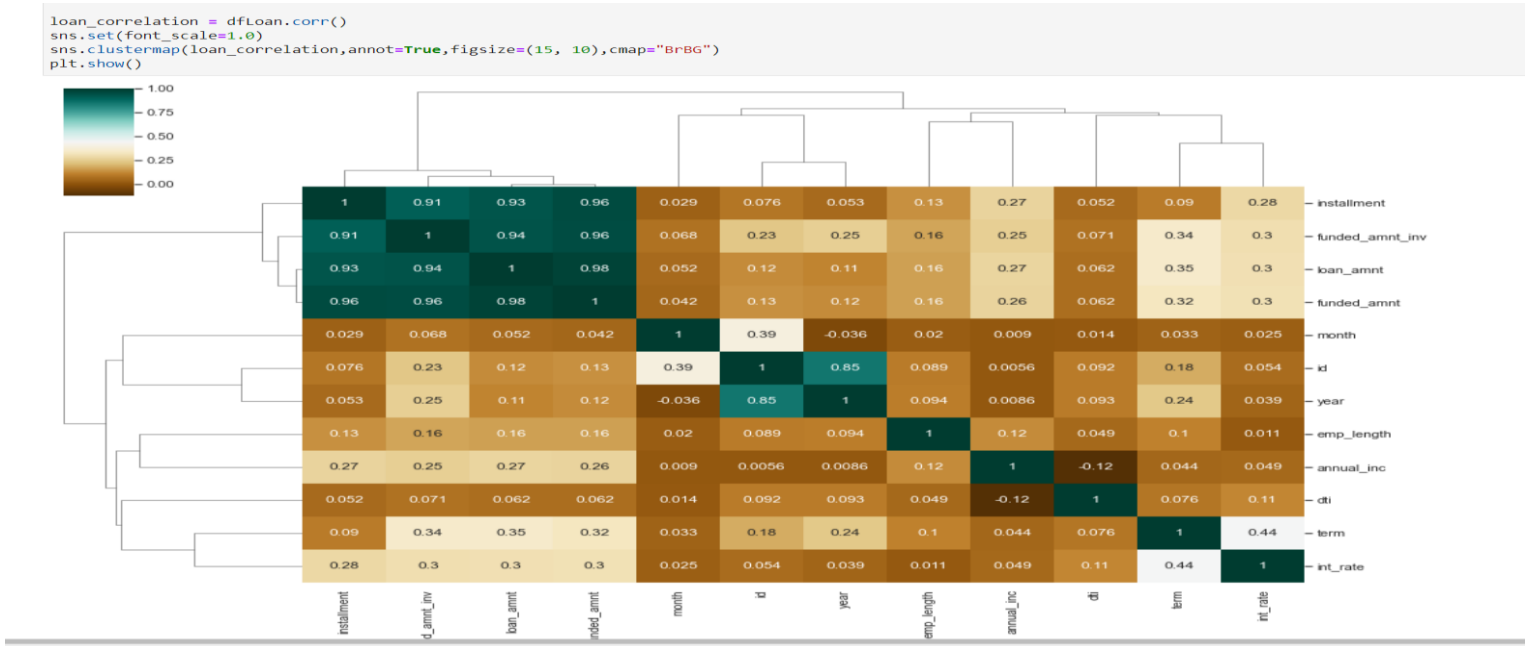
1. Following information extracted from data frame using various methods
2. Convert the excel data into the data frame using python library as pandas
3. Extract the information of data frame using info and description methods
4. Identify the columns with null values and null values percentage
5. Identify the columns with zero values and zero values percentage
6. Identify the customer behavioural columns
7. Identified the target columns and other important columns to include in the analysis
8. Calculated the number of rows present with all NA value
9. Calculated the shape of data frame
10. Identified the columns with same value throughout the column.
11. Updating the null values with most occurrence values in emp_title columns and zero in emp_length

Data Cleaning

1. Removed all the columns with all values as null
2. Removed all the columns with all values as Zero
3. Removed all the columns with the higher percentage of null values(50).
4. Removed all the columns with the higher percentage of zero values(80).
5. Removed all the columns with the same value throughout the column
6. Removed all the rows with loan status as current, it can be default and fully paid so deleting the relates record from analysis process.
7. Removed all the customer behavioural variables columns.
8. Removed columns with unnecessary values(URL,desc ..)
9. Replaced the NAN value with None in emp_title column

Correlation Matrix - Quantitative Variables

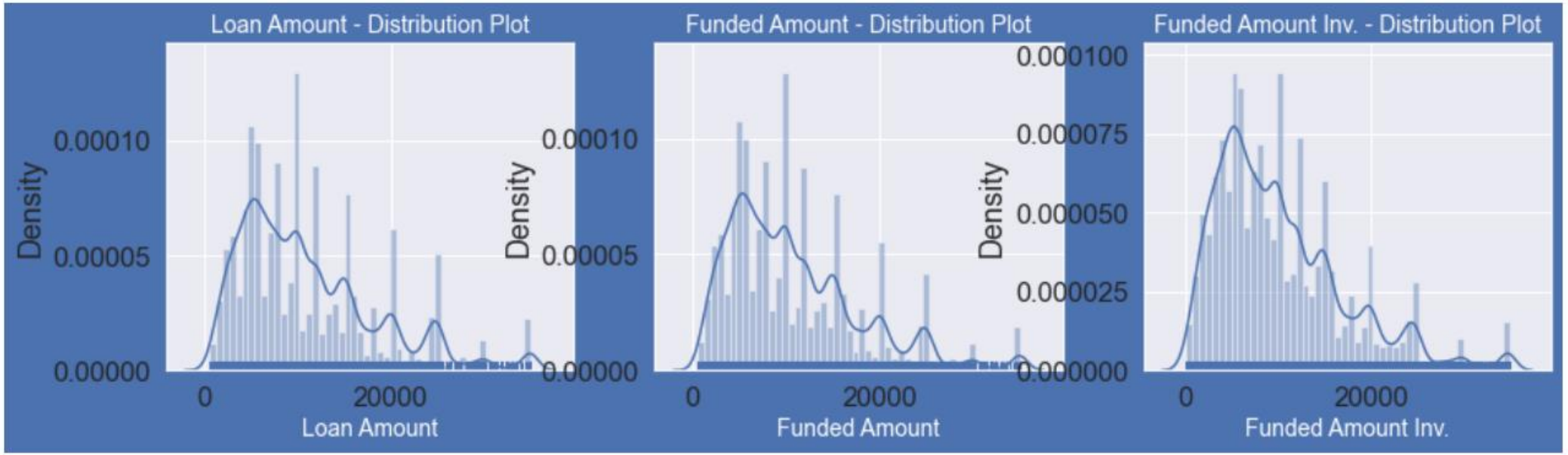
```
loan_correlation = dfLoan.corr()
sns.set(font_scale=1.0)
sns.clustermap(loan_correlation,annot=True,figsize=(15, 10),cmap="BrBG")
plt.show()
```



Observation:

Loan amount, investor amount, funding amount are strongly correlated.

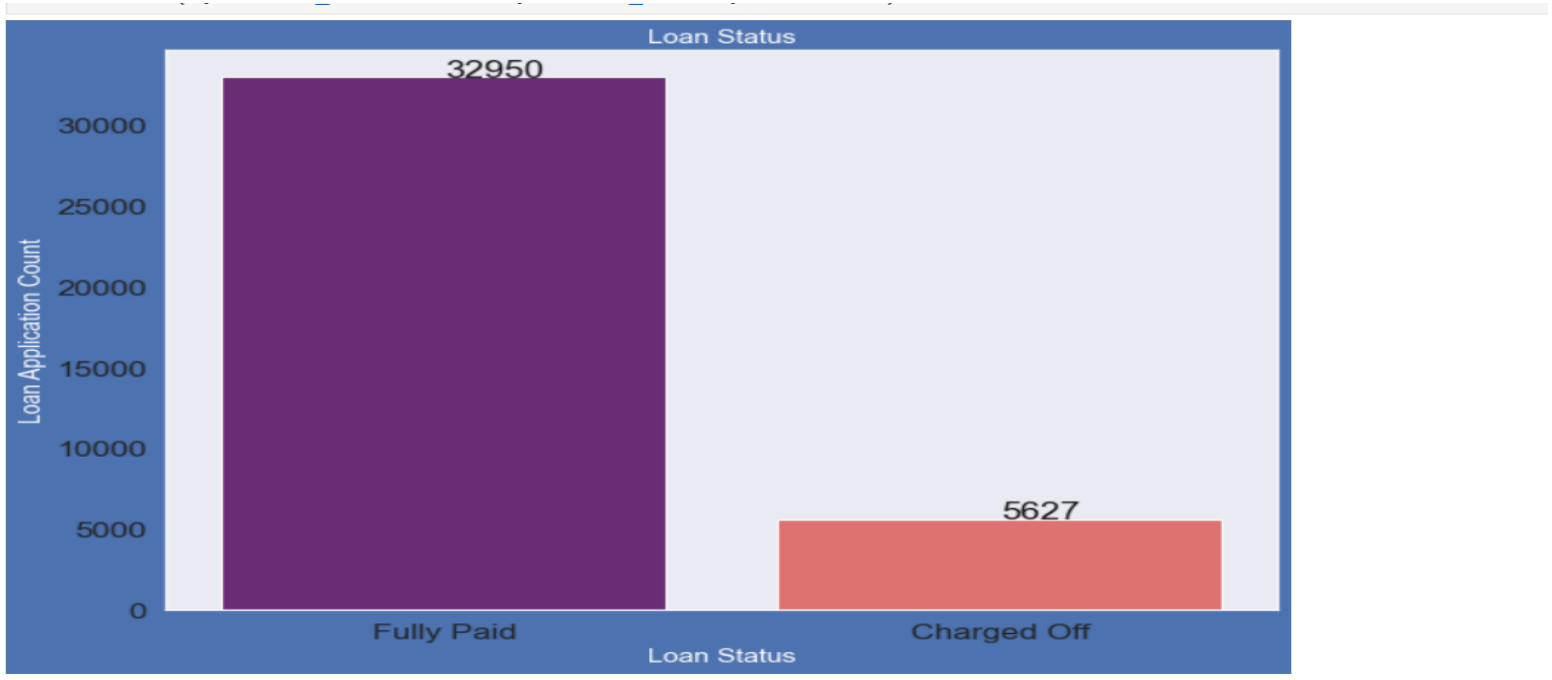
Univariate Analysis



Observation:

Distribution of amounts for all (Loan amount, investor amount, funding amount) these are strongly correlated.

Univariate Analysis - Loan Status

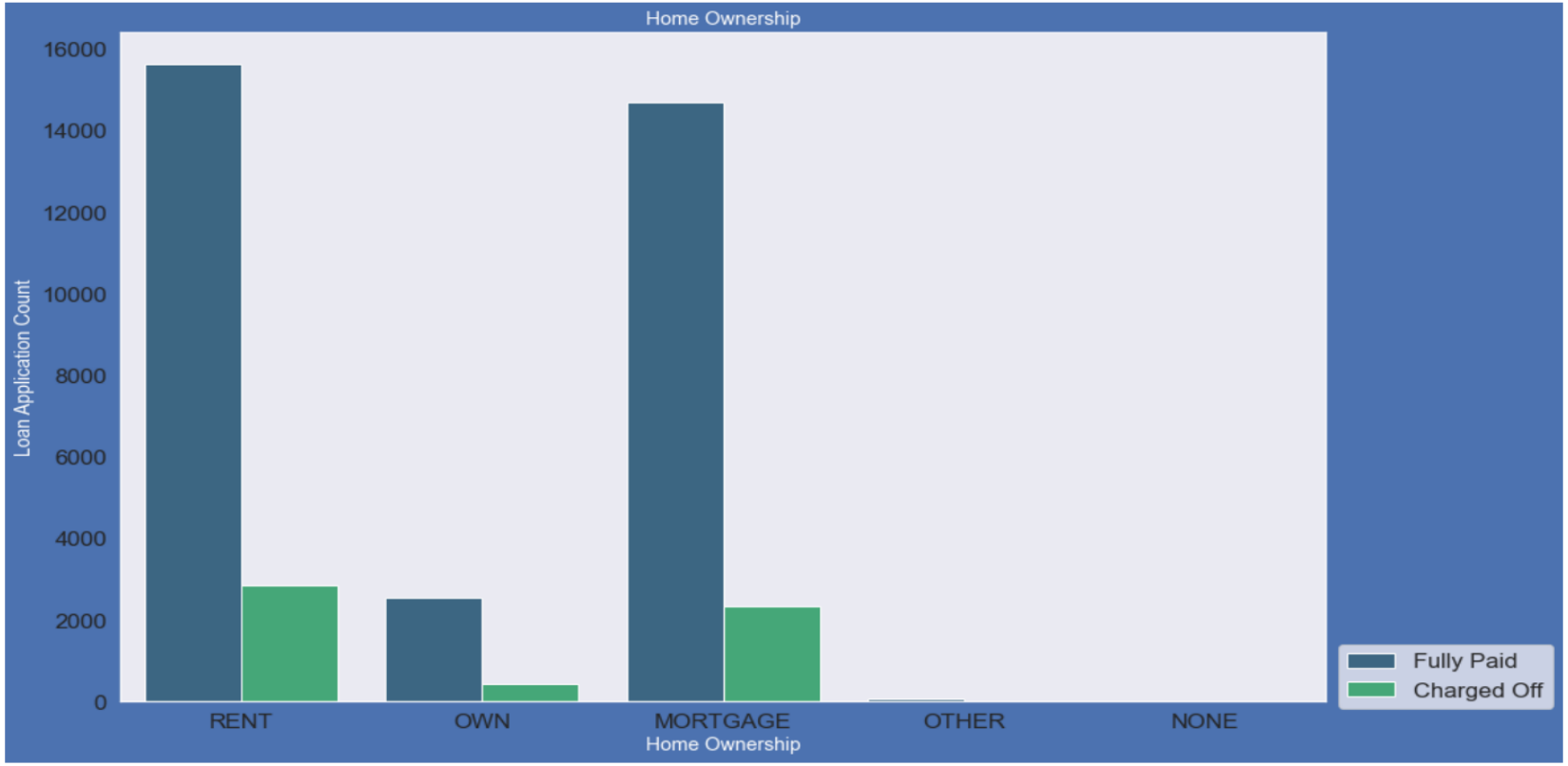


Observation:

Above plot illustrations that close to 14.5% loans were charged off .

Univariate Analysis - Home Ownership

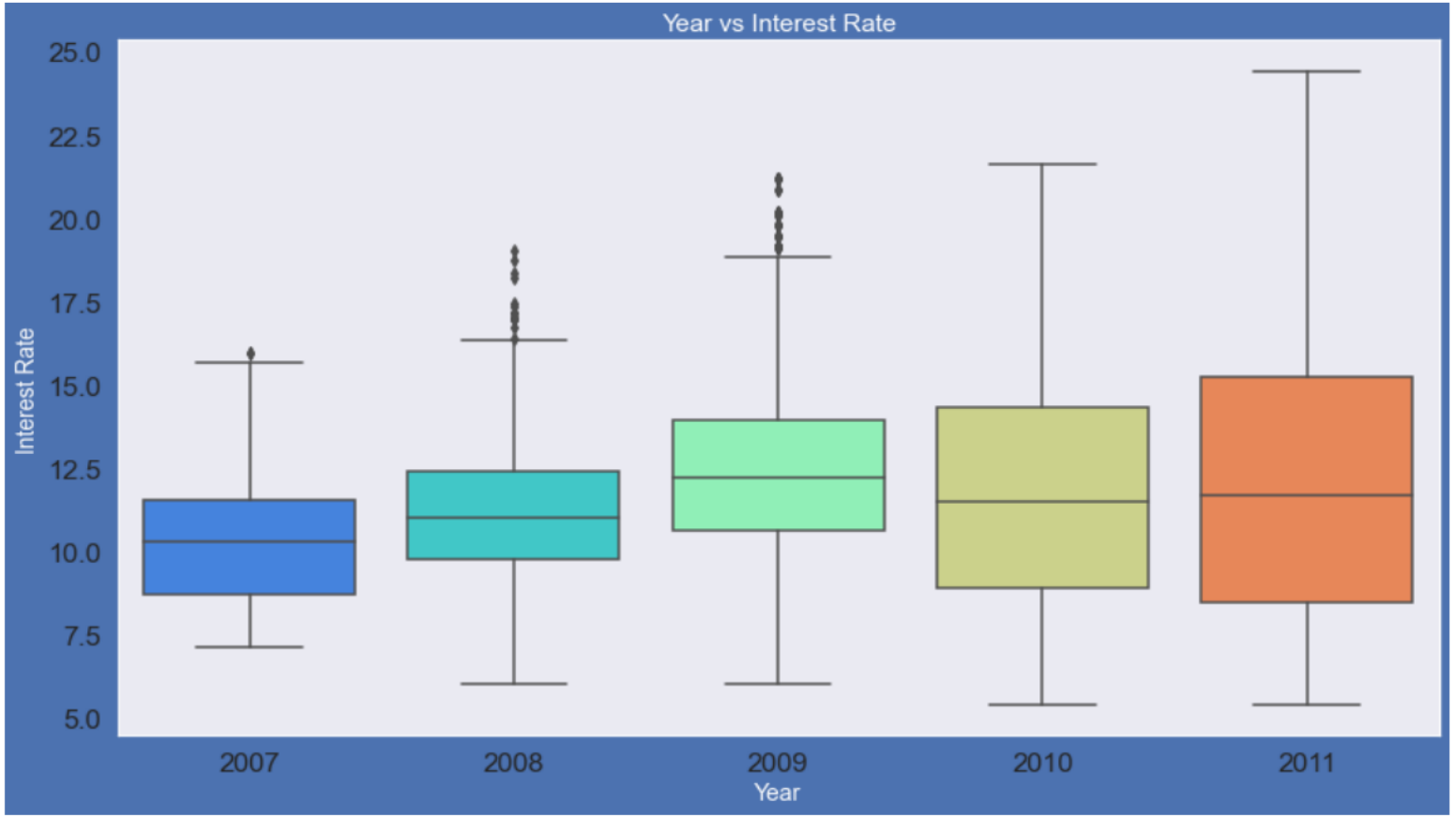
```
plt.show()
```



Observations:

Charged off is high as mentioned categories(Rent, Mortgage).

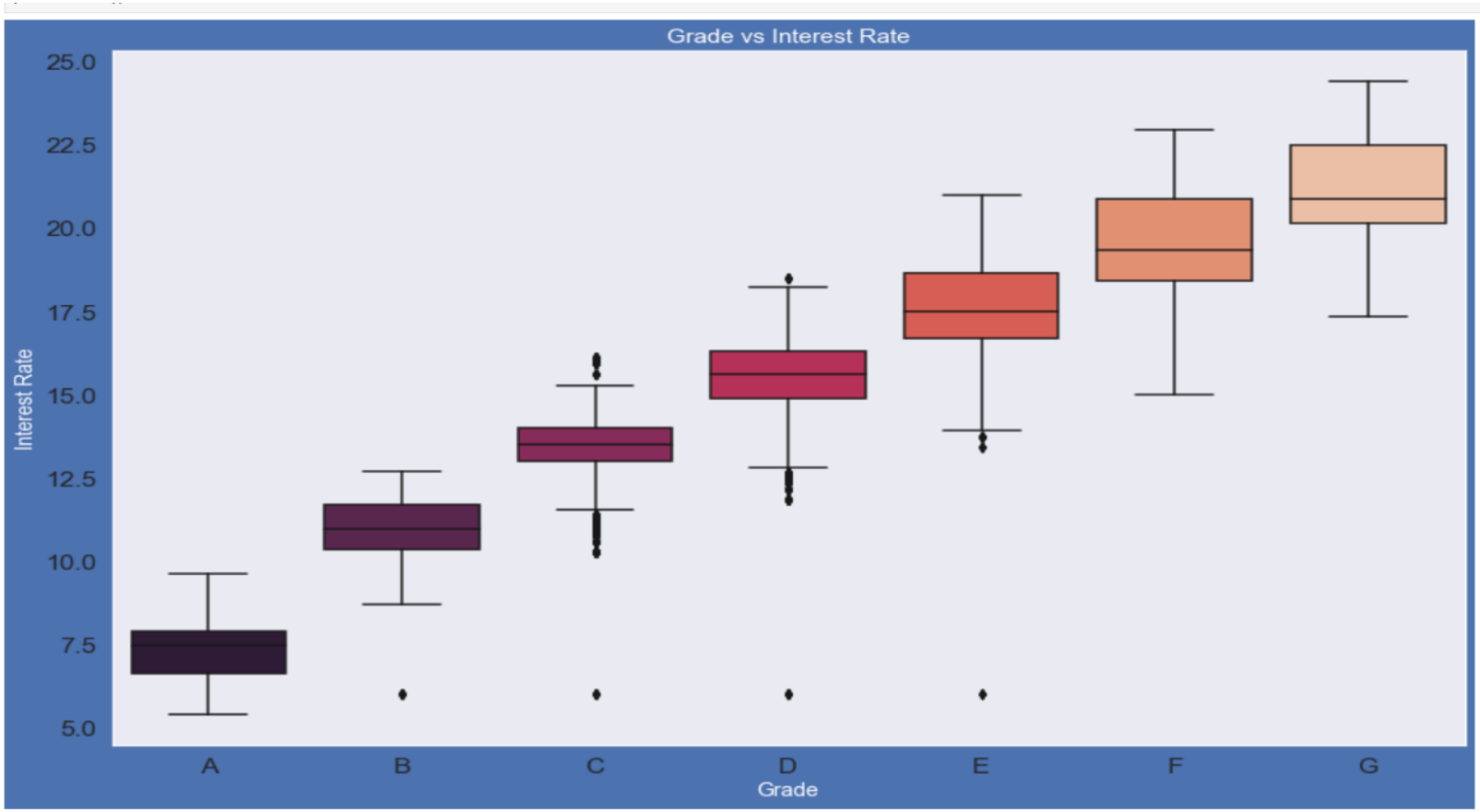
Bivariate Analysis - year vs Interest Rate



Observation:

Interest rate is increasing slowly with increase in year

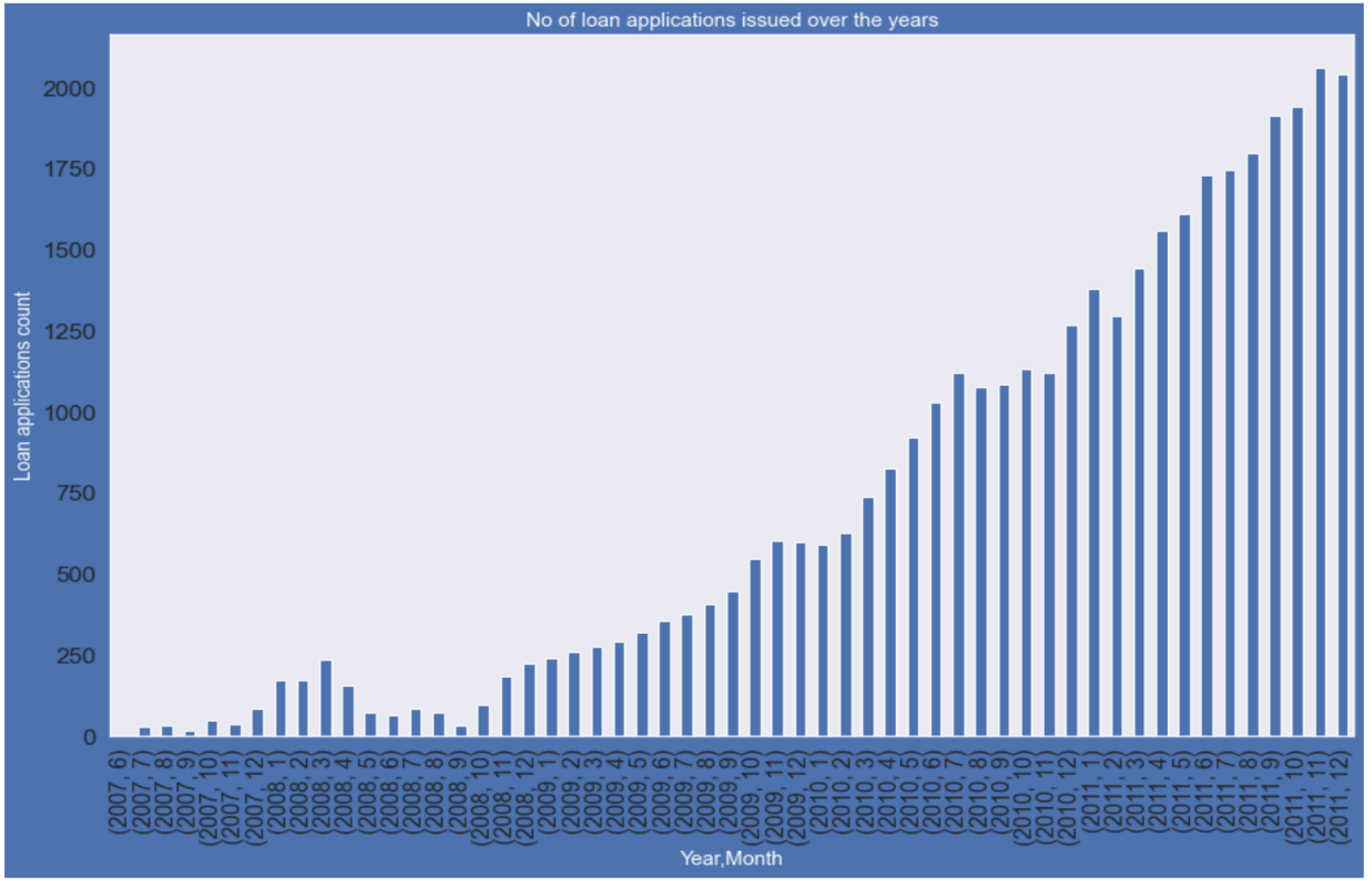
Bivariate Analysis - Grade vs Interest Rate



Observation:

The higher the borrower's credit grade, the lower the interest rate offered to that borrowed loan

Derived Column - Ordered Categorical Variables



Observation:

As the loan applications increased charged off applications Increased

Recommendation

1. To identify patterns which indicate if a person is likely to defaulter.
2. Those who have the own home will be preferred customer to provide the loan