Lending Club Case Study

Exploratory Data Analysis

- Vivek Paradkar

Problem Statement

This is a largest consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile.

Lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who **default** cause the largest amount of loss to the lenders. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

The company wants 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.

As a Data Analyst, perform EDA to understand which consumer and loan attributes are the most important driving factors behind loan default.

Data Source

- loan.csv

Loan data set is available in a csv file format. It contains the complete loan data for all the loans issued through the time period 2007 to 2011. This does not contain the loans rejected by the company, since there is no transactional history of those applicants with the company

- Data_Dictionary.xlsx

This file contains the meaning of the all the variables present in the Loan data set.

Data Analysis

Following steps were performed as part of this analysis

- Drop all columns which have more than 60% null values.
- Drop all columns which are unique or have single value for all the rows
- Drop columns that are related to customer behavior or are redundant.
- Impute values for columns with less percentage of null values
- Convert columns to proper data type

- Perform Univariate Analysis on numerical and categorical variables

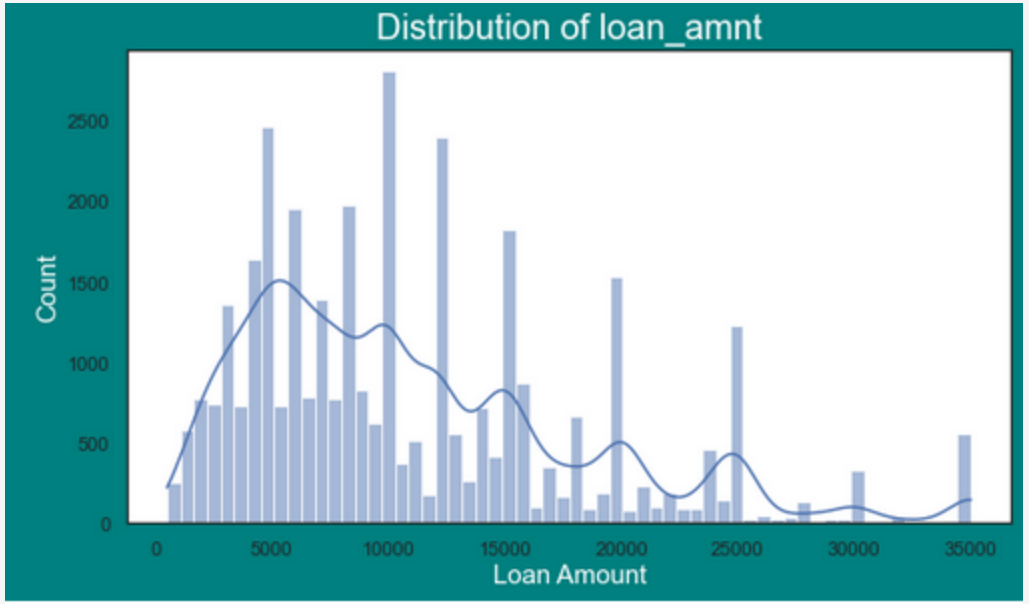
- Based on the analysis done, document and present the recommendations.

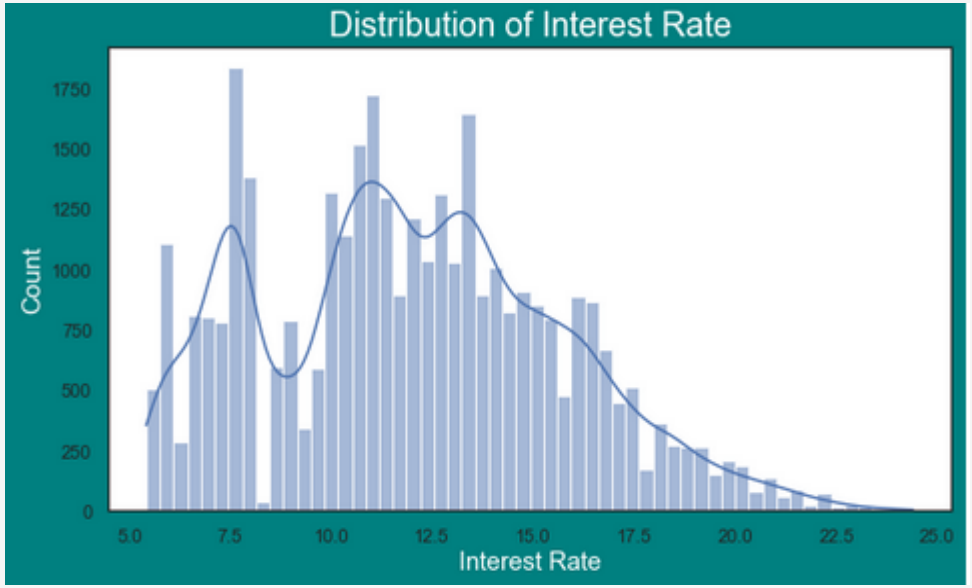
- Create Derived Metrics for Loan Issue Year and month
- Create bins for numeric columns to perform analysis such as loan amount, annual income and interest rate.

- Perform bivariate analysis to see how one variable affects another
- Find correlation between variables and identify variable pairs that are positively or negatively correlated.

Data Analysis

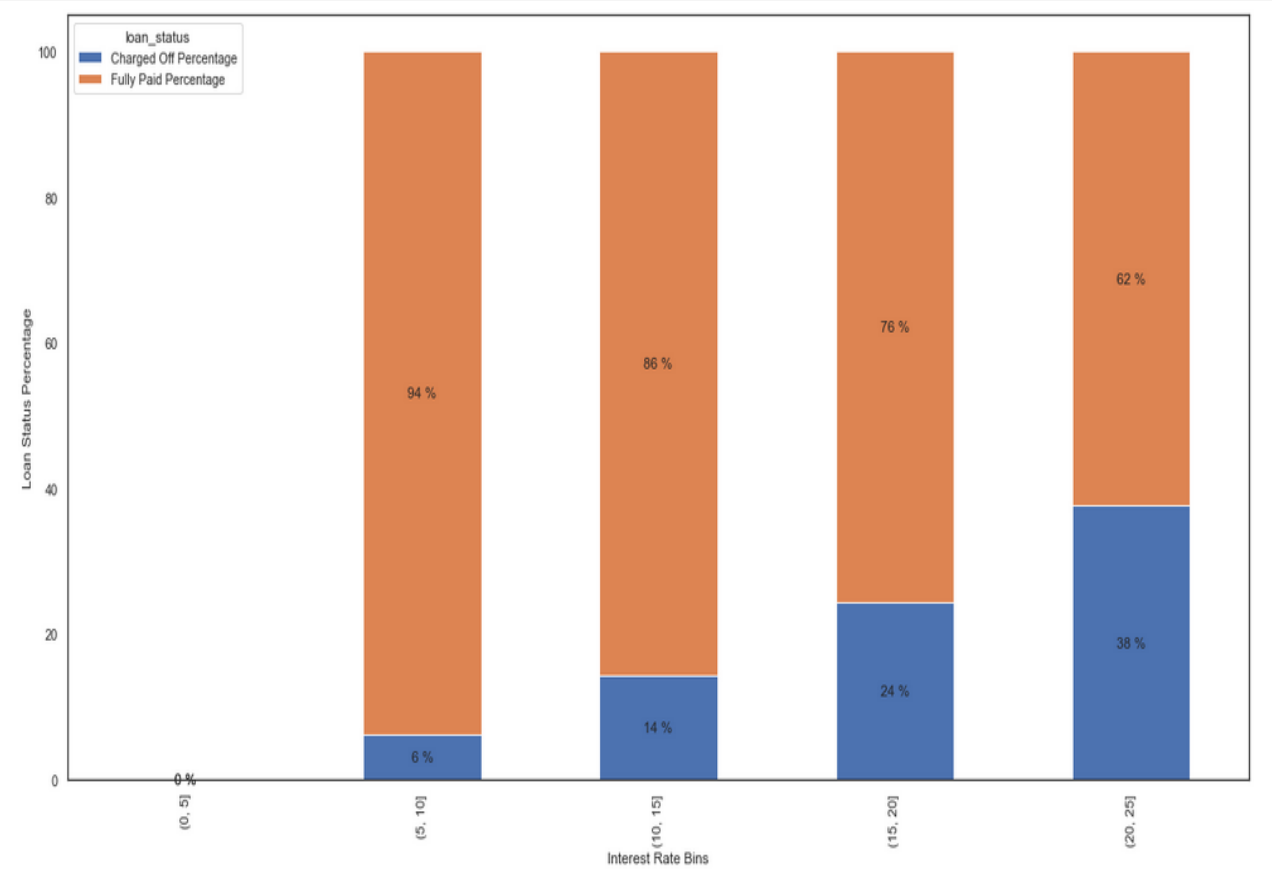
Checked the distribution of loan amount as well as distribution of interest rate. Both the distributions have tail towards the end(high value).





Data Analysis

Interest Rate and Loan Status

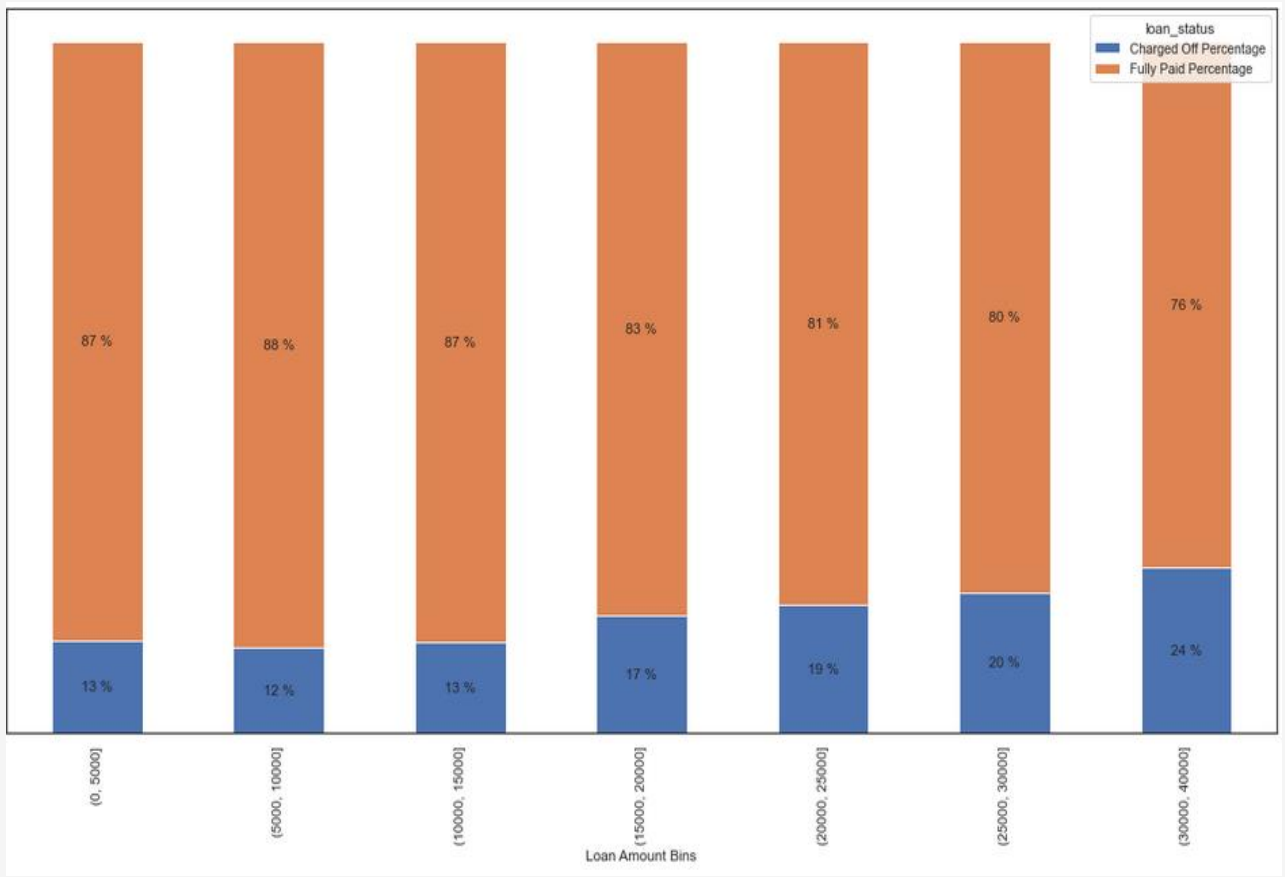


It can be noticed from the graph that percentage of loans defaulted increases with the increase in Interest Rate.

38% Loans are defaulted with Interest rate greater than 20%.

Data Analysis

Loan Amount and Loan Status

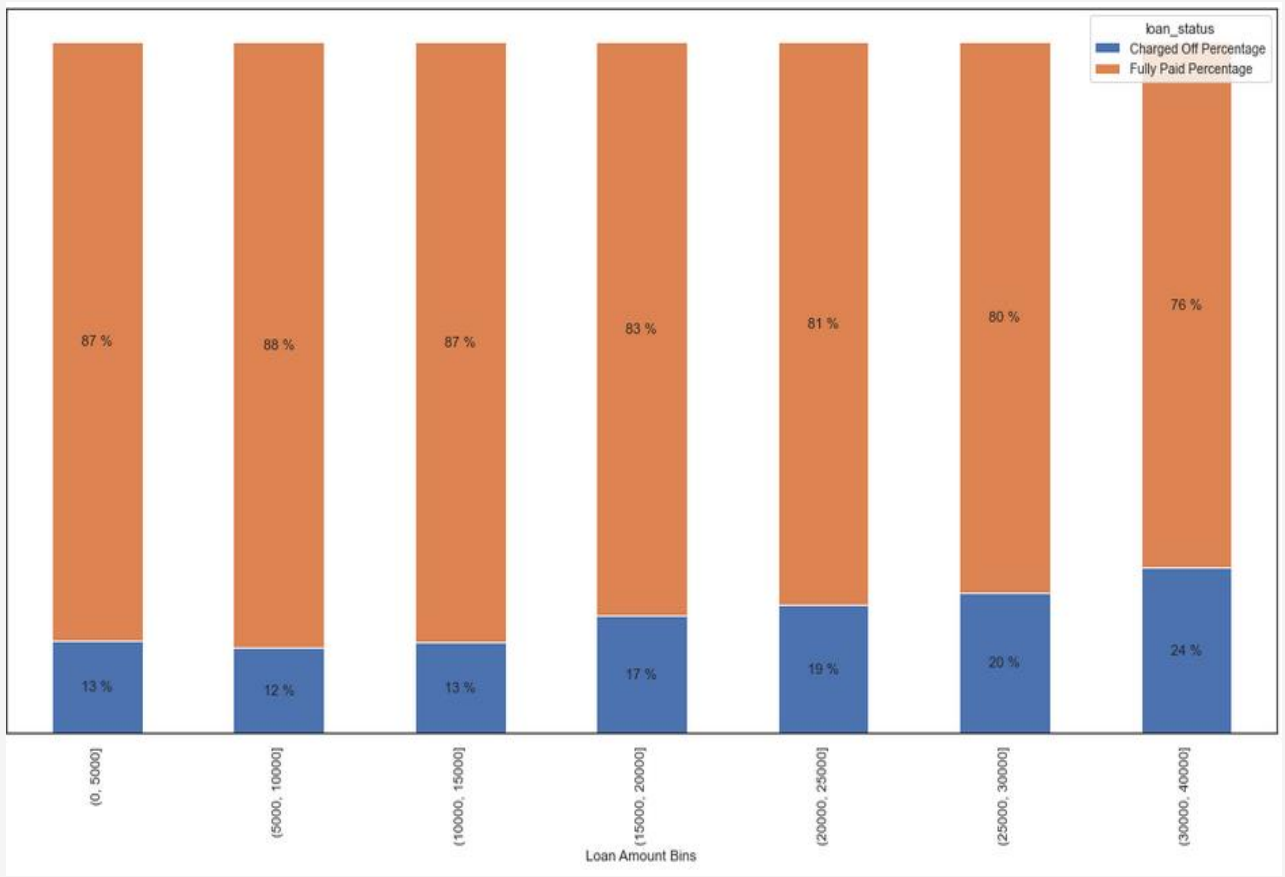


It can be noticed from the graph that percentage of loans defaulted increases with the increase in Loan Amount.

24% Loans are defaulted with loan amount greater than 30K.

Data Analysis

Loan Term and Loan Status

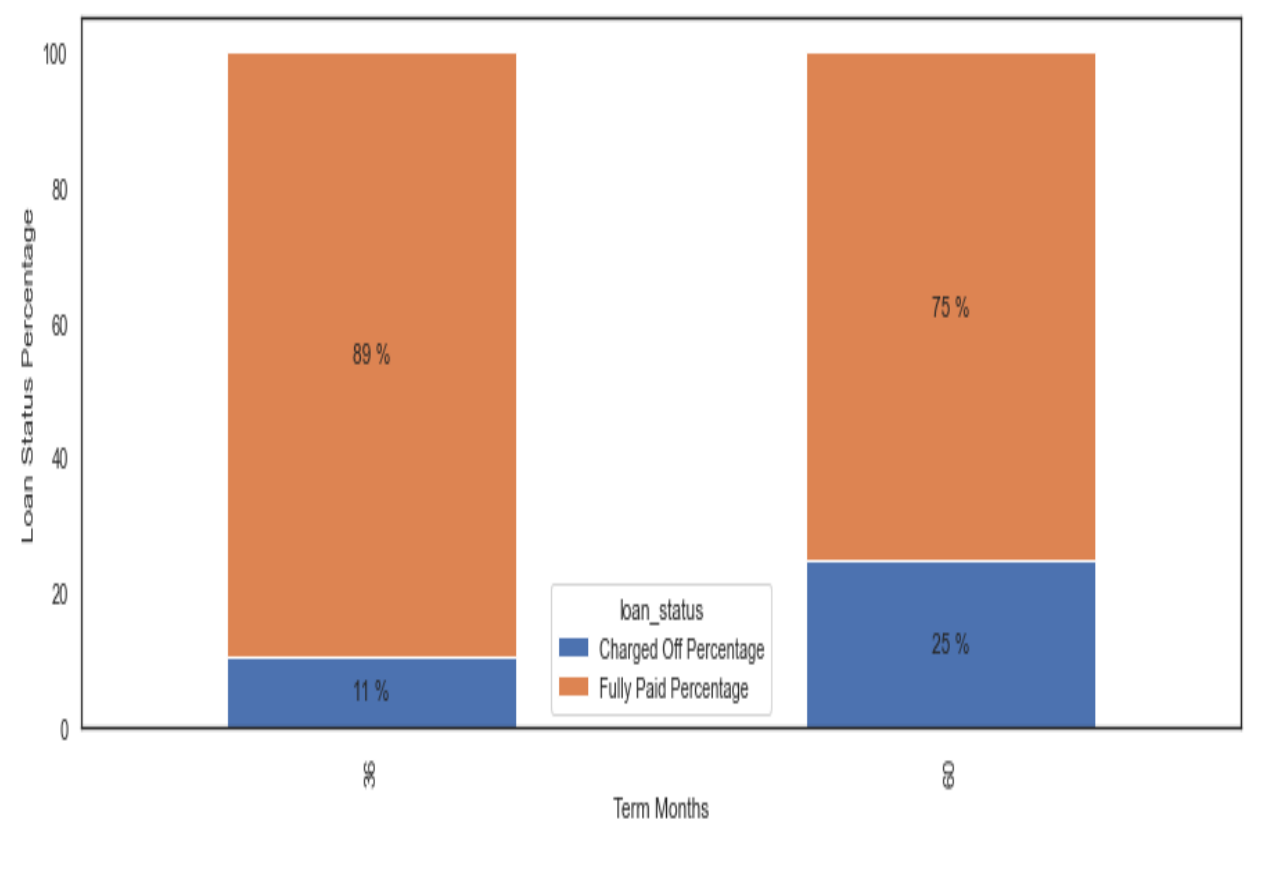


It can be noticed from the graph that percentage of loans defaulted increases with the increase in Loan Amount.

24% Loans are defaulted with loan amount greater than 30K.

Data Analysis

Loan Grade and Loan Status

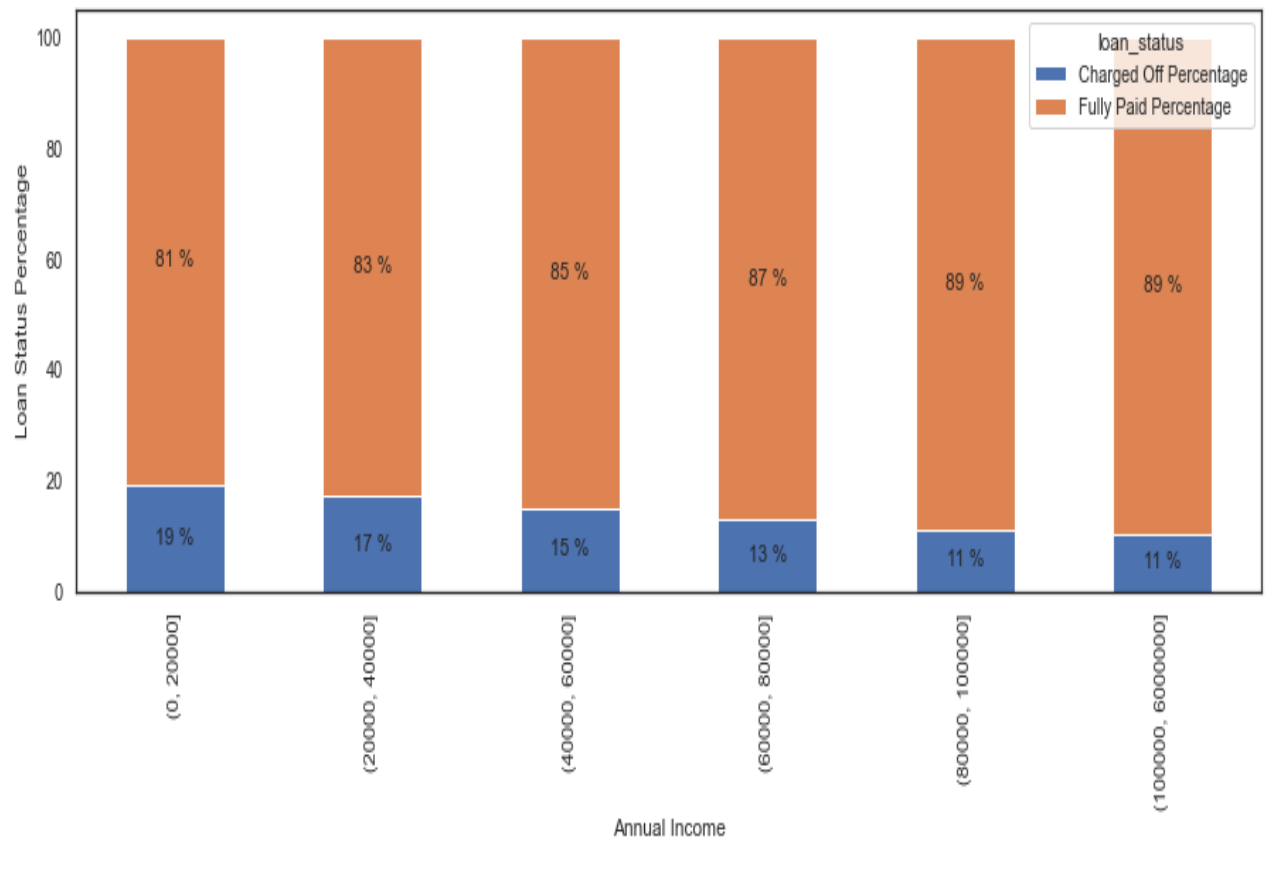


It can be noticed from the graph that percentage of loans defaulted is more for loans with term as 60 months.

Loan applications with 60 months term have 25% defaults.

Data Analysis

Annual Income and Loan Status



It can be noticed from the graph that percentage of loans defaulted increases as the Annual income decreases

19% of Loans are defaulted with Annual income less than 20K.

Recommendation

1. Avoid approving loans with Interest Rate higher than 15%. As the interest rate increases the number of loans charged off increases

2. Avoid approving loans with Loan Term 60 Months. As the Loan term increases the number of loans charged off also increases

3. Avoid approving loans with Loan Amount more than 30000. As the loan amount increases there is an increase in loan defaults

4. Avoid approving loans with Loan Grade E,F and G, these have higher loan defaults

5. Avoid approving loans with Annual Income less than 20K, as the annual income decreases the percentage of loan defaults also increases.