


Lending Club Case Study

Exploratory Data Analysis of Lending Club: Insights into Peer-to-Peer Lending



Problem Statement: Risk of losing money while lending to customers in banking and financial services



Overview: Based on historical data company wants to understand the driving factors(or driver variables) behind loan default, i.e, the variables that are strong indicators of default. Company can then utilize this knowledge for its portfolio and risk assessment/management.

Analysis Approach:

1. **Data understanding and exploration** pertaining to the financial transaction.
2. **Data Cleaning** - To filter/remove columns that are not very pertinent to problem solving.
3. **Data Analysis** - Univariate and Bivariate analysis of columns to determine its relevance for addressing the problem.
4. **Recommendation** - Based on findings, propose certain key insights which may reduce the probability of funding a defaulter.

Data Understanding and Cleaning



Dataset Overview/Understanding

- Dataset has records of 39717 individuals and 111 columns of attributes pertaining to them.
- Columns can be segregated into 2 categories:
 - **Consumer Attributes:** Employee title, Annual Income, Title, Employment Length, etc
 - **Loan Attributes:** Term, Interest Rate, Loan Amount Given, etc
- Key target column is **Loan Status** which has 3 categories - **Fully Paid, Charged Off & Current.**
- Based on analysis, some of the important columns are:
 - Loan Amount, Term, Grade, Annual Income, DTI, Interest Rate, etc

Data Understanding and Cleaning

Dataset Cleaning:

Below are some of the steps executed to clean the dataset

1. Drop columns which have only **Null** Data. Dataset had 54 columns which were null.
2. Removing **Duplicates**. Although we checked the dataset for duplicates, there weren't any.
3. Dropping **customer behaviour** variables since these columns do not aid in our analysis of identifying a likely defaulter.
4. Columns which had single values and unfavourable to our final analysis were also removed. Some of these columns were: "pymnt_plan", "tax_liens", "policy_code", etc
5. Removed the rows pertaining to category "**Current**" from the column "**Loan Status**". "Current" refers to people who loan is still ongoing making it unsure as to which group they belong to.
6. We identified the column "**emp_length**" as an important variable for our analysis. It had some missing values, we filled it with the mode - 10+ years.
7. Annual Income has some outliers which may skew our analysis. Hence removing values beyond 90 percentile

Univariate Analysis

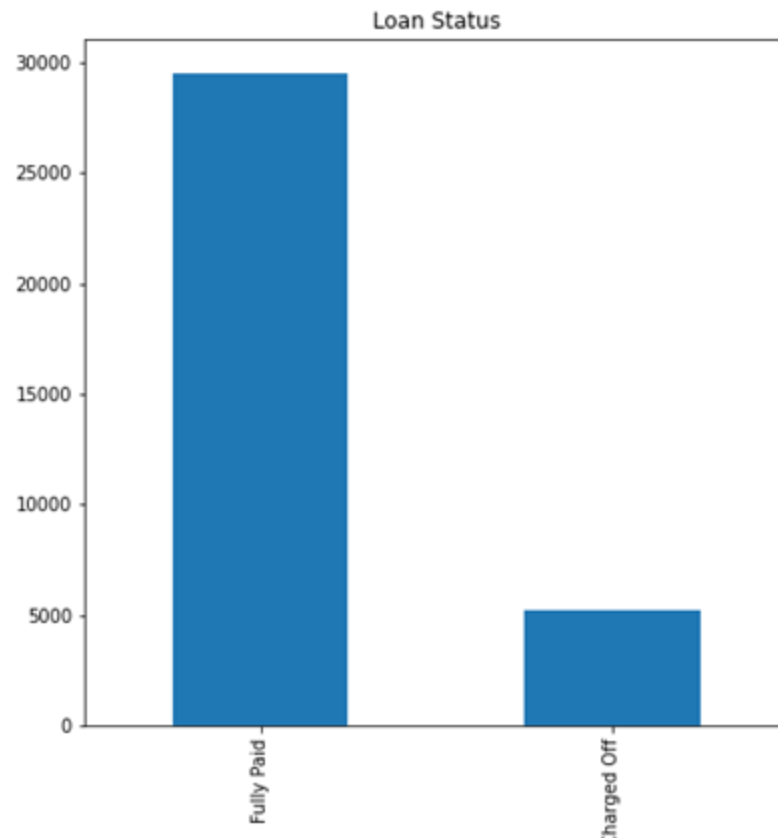
Loan Status:

There are 2 categories: Fully Paid and Charged Off.

14.5% of people default on their loan.

Full Paid: **29525**

Charged Off: **5198**



Univariate Analysis

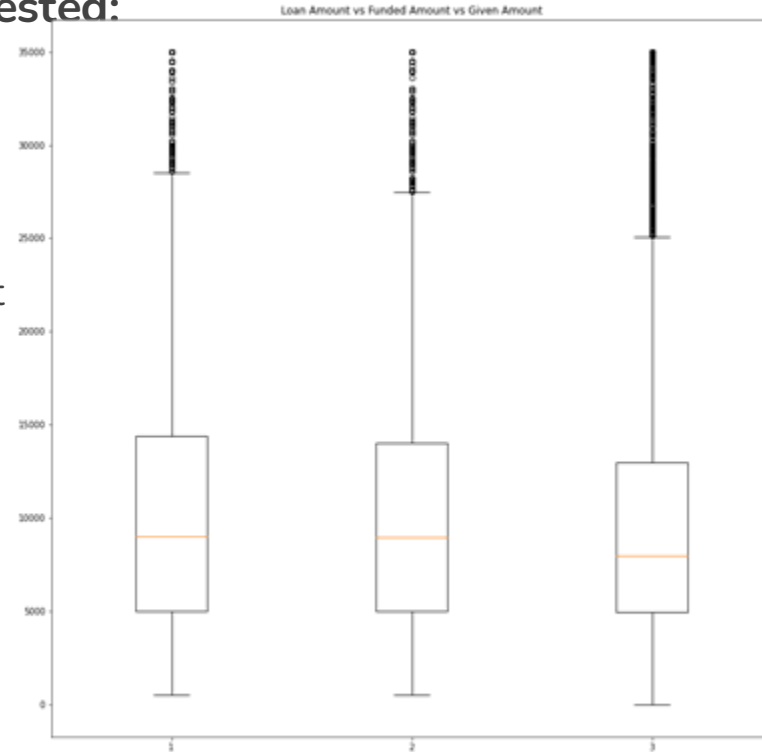
Loan Amount, Funded Amount, Funded Amount Invested:

Loan Amount: Loan amount applied by the customer

Funded Amount: Amount approved by the lending club

Funded Amount Invested: Final amount given as loan to customer

From the figure we can infer that the median Funded Amount Invested is lower than the Loan Amount showing that the given loan is less than what is requested.



Univariate Analysis

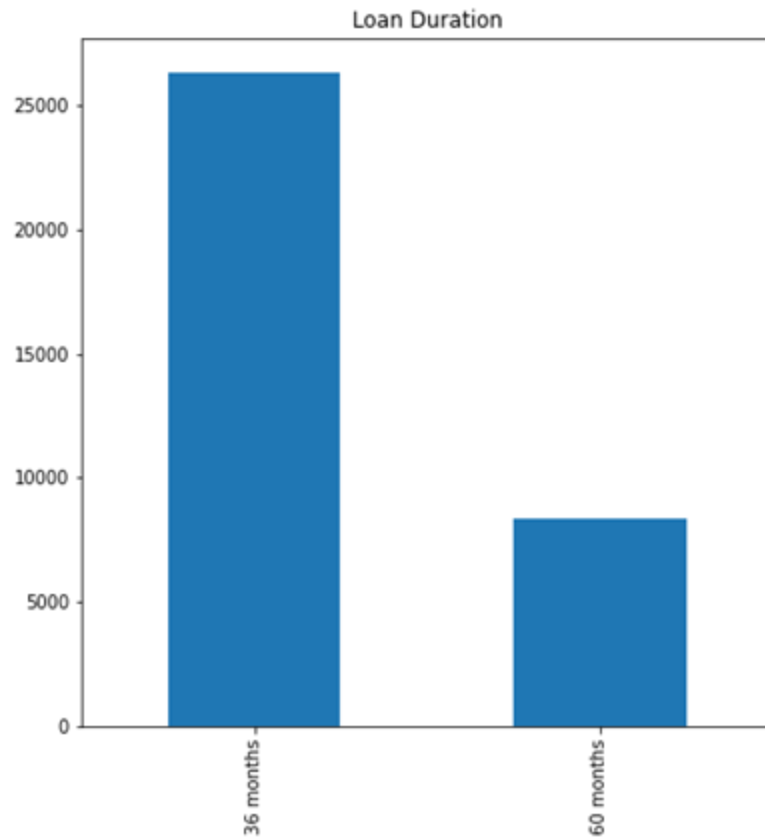
Term:

There are 2 categories: **36 months** and **60 months**

75% of all applicants have a term of 36 months

36 Months: **26341**

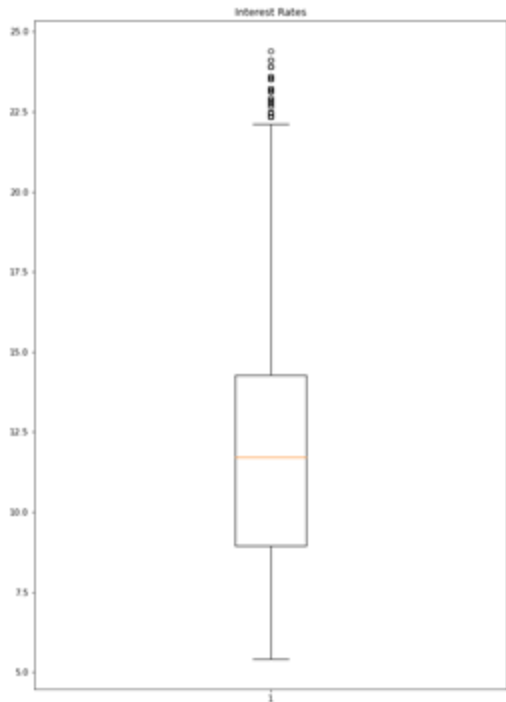
60 Months: **8382**



Univariate Analysis

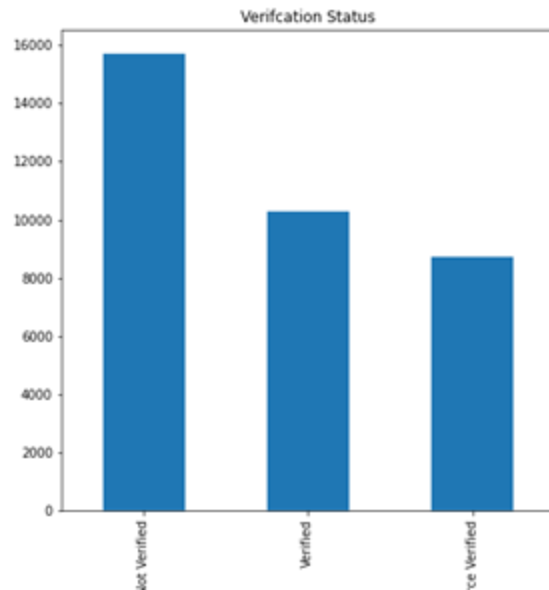
Interest Rate:

mean **11.86**
min 5.42
25% 8.94
50% 11.71
75% 14.27
max 24.4



Verification Status

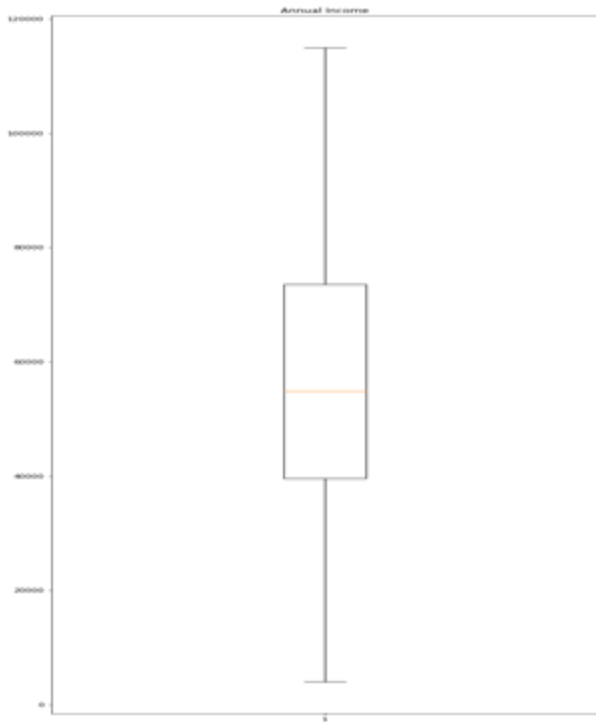
Not Verified **15723**
Verified **10298**
Source Verified **8702**



Univariate Analysis

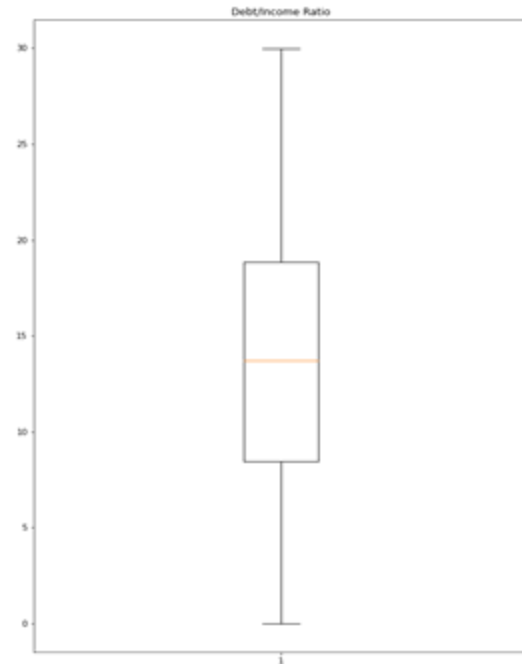
Annual Income:

mean 57209
min 4000
25% 39600
50% 54912
75% 73600
max 115000



Debt to Income Ratio

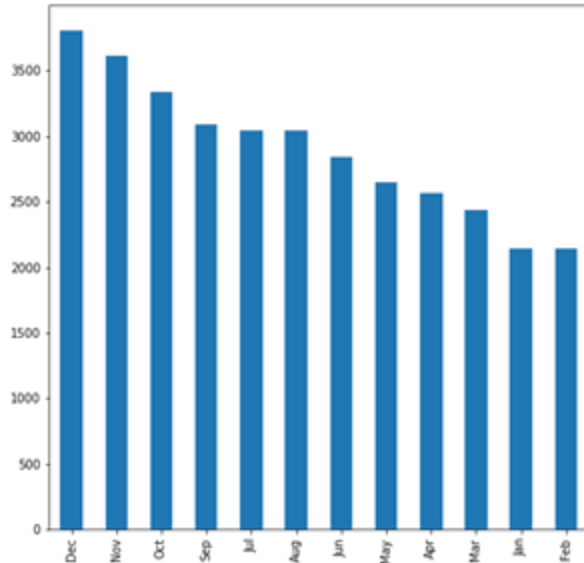
mean 13.54
min 0.0
25% 8.45
50% 13.71
75% 18.84
max 29.99



Univariate Analysis

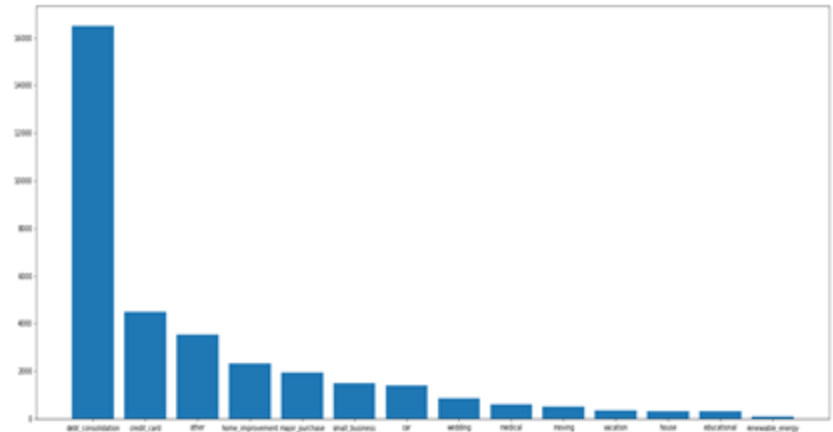
Issue Month:

From the graph it is visible that most of Loans are given at the last months of the Year. December has the highest number of Loans disbursed, followed by November.



Purpose:

Highest number of loans were given out for “**debt consolidation**” at **47.5%**. Followed by “credit card” at 12.9%.



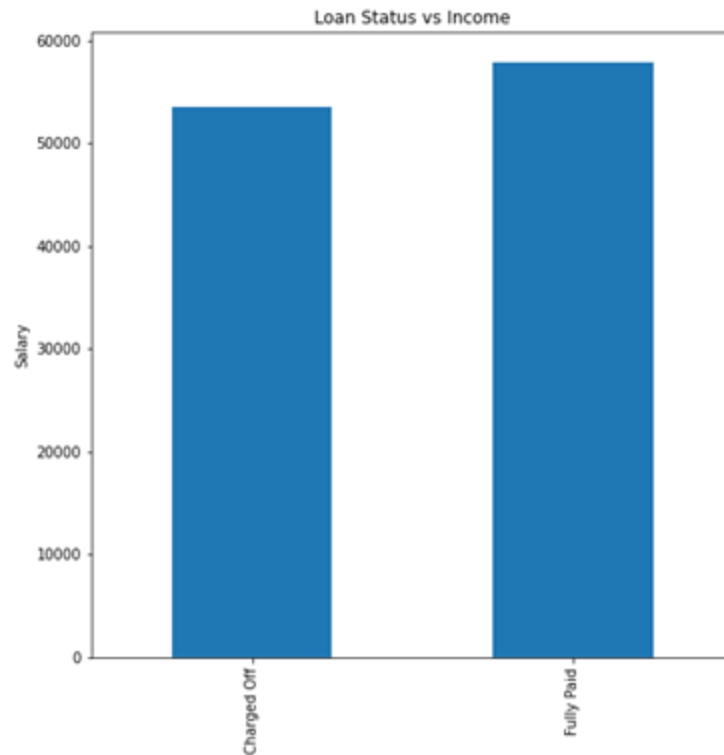
Bivariate Analysis

Loan Status vs Annual Income:

Average Annual Income - 57209

Mean Annual Income of people who default - **53491**.

Mean Annual income of people fully paid - **57864**



Bivariate Analysis

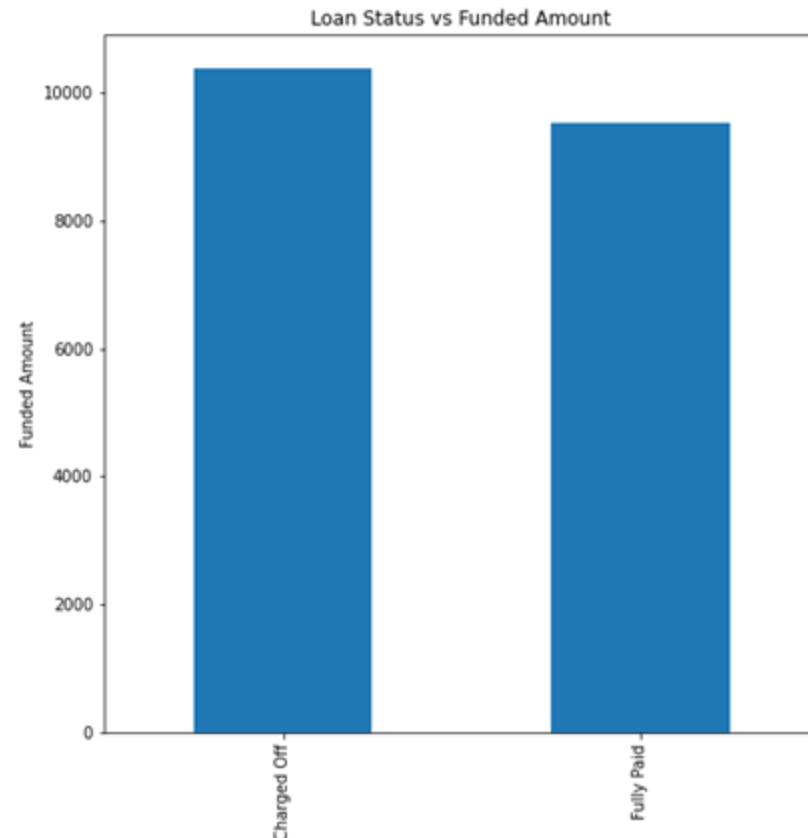
Loan Status vs Funded Amount Invested:

Mean Funded Amount - 9641

Mean funded amount of people who default - **10380**

Mean Annual income of people fully paid - **9511**

People who default take up loans larger than the mean amount funded.

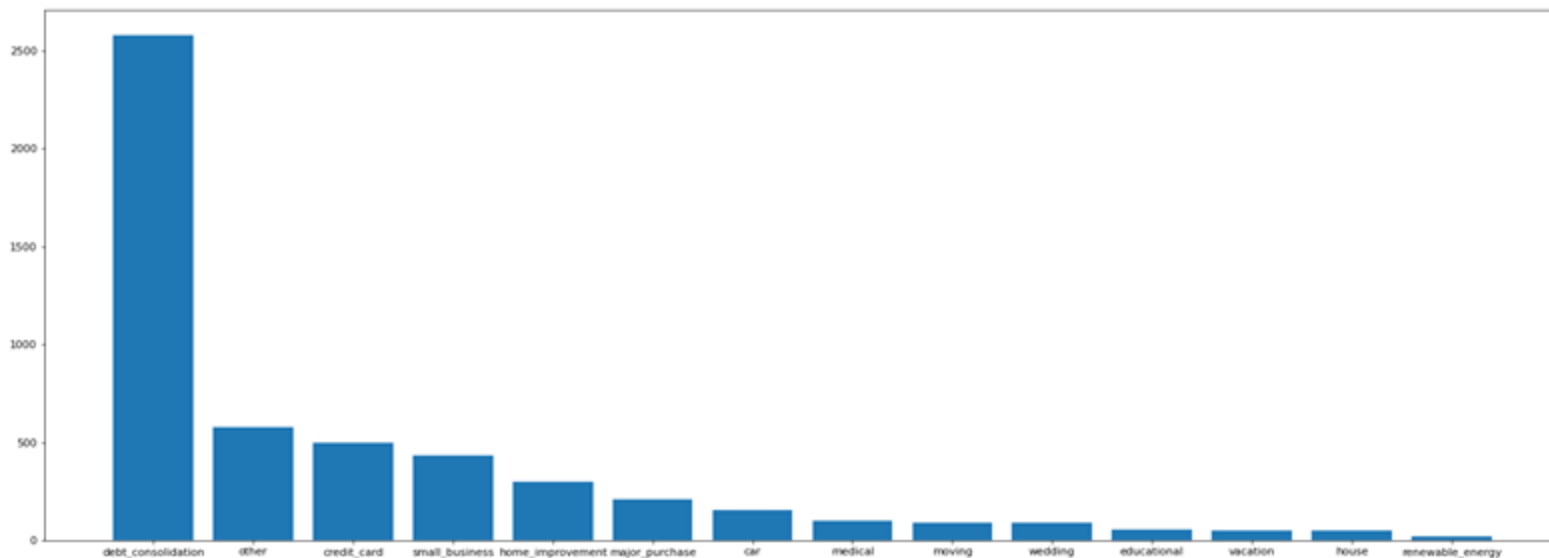


Bivariate Analysis

Loan Status vs Purpose:

~**48%** of all loans defaulted were taken for “**debt consolidation**”

Ratio of default to fully paid for debt consolidation is also highest at **18.52%**



Bivariate Analysis

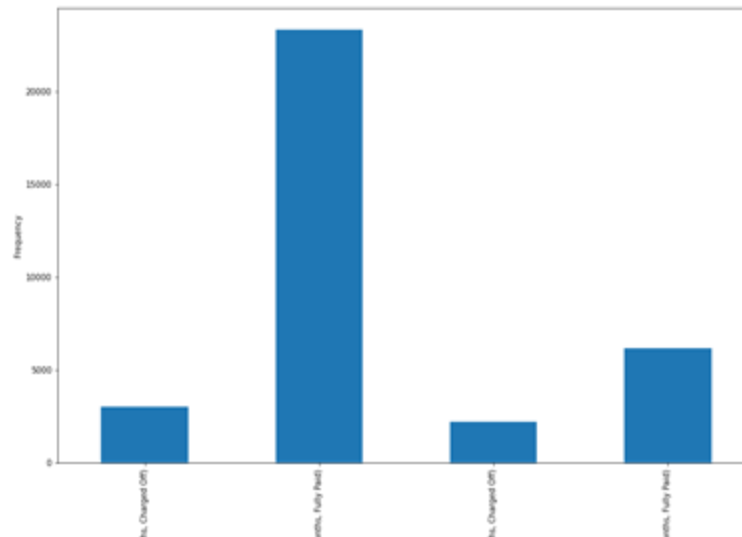
Loan Status vs Term:

25.31% of people who took 60 months term plan defaulted on their loans

11.4% of people who took 36 month term plan defaulted on their loans

People who take 60 month loan tend to default twice as much as those who take 36 month loan

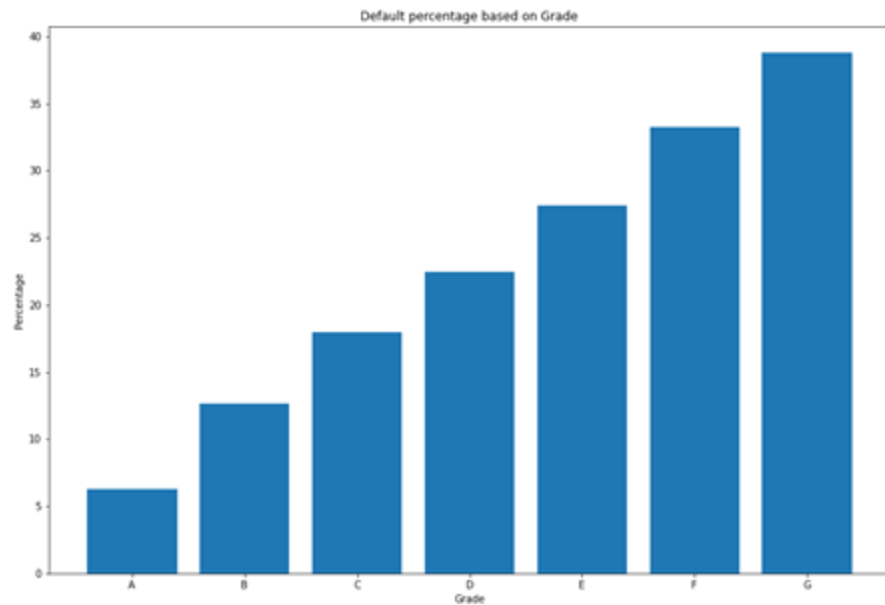
```
term      loan_status
36 months Charged Off    3003
          Fully Paid    23338
60 months Charged Off    2195
          Fully Paid    6187
Name: loan_status, dtype: int64
```



Bivariate Analysis

Loan Status vs Grade:

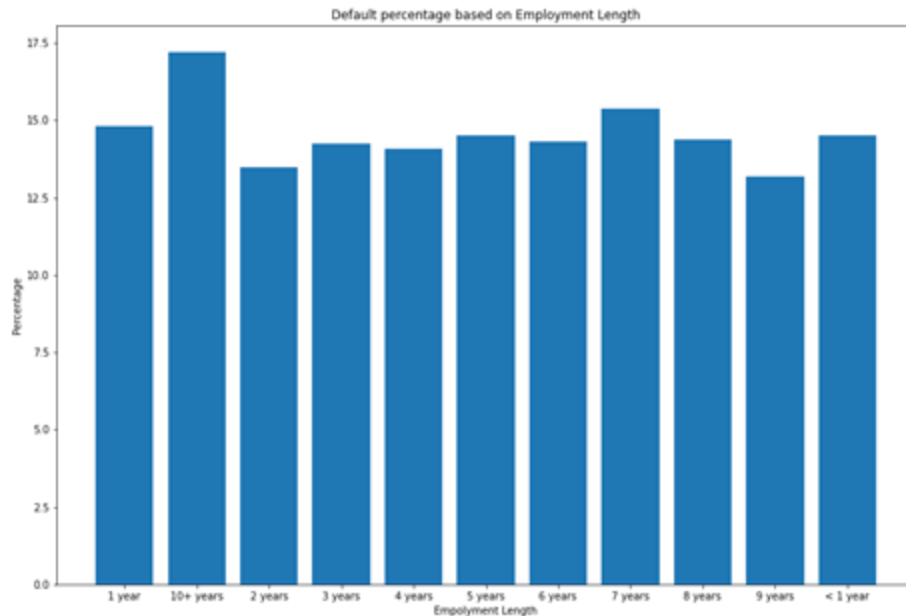
The default percentage increases as the Grades increase. The default rate is highest at for grade **G** at ~**39%**



Bivariate Analysis

Loan Status vs Employment Length:

People with employment length greater than 10 years have the highest default percentage of ~16%



Observation/Recommendation

Point to remember:

We have not added all the columns analysis in this PPT. For detailed analysis on all columns, please refer to the python notebook.

Observation:

Likelihood of Defaulting is when:

1. Annual salary is below the mean value
2. Loan is given for 60 months duration. 25% people default.
3. When the interest rate of the loan is above the mean value.
4. When the employment length is greater than 10+ years.
5. Default rate increases as the Grade of the customer increases.