

The background is a dark blue gradient with a faint, abstract line graph. The graph features several data points connected by lines, with some points highlighted in white and others in blue. A specific data point is labeled with the value '289.33'.

EDA : Lending club case study

- Manasi Pawar

Problem Statement



Consumer finance company wants to understand the **driving factors (or driver variables)** behind loan default



The company can utilize this knowledge for its portfolio and risk assessment



Analysis will help company to make a decision for loan approval based on the applicant's profile



Analysis approach

Data understanding and sourcing

Check data quality issues and fix missing values

Perform univariate data analysis – categorical and numerical variables

Perform bivariate data analysis - categorical and numerical variables

Identify correlation between continues variables

Data Understanding



- There are 111 columns having various data types like object, int, float and 305711 rows.
- There are many columns with 0 values or NULL values.
- There are columns with special characters like interest rate, employment length, standardizing is required

Data quality check

- There are 58 columns having missing value more than 30%. Hence we exclude these columns from analysis

```
## list of columns where missing values are above 30%
nullcol_gt30 = (round((data.isnull().sum()*100/len(data)).sort_values(ascending = False),2))[round((data.isnull().sum()*100/len(data)).sort_values(ascending = False),2)>30]
```

✓ 0.9s

Python

```
print("Num of columns having missing values more than 30% :",len(nullcol_gt30))
```

✓ 0.3s

Python

Num of columns having missing values more than 30% : 58

Data standardizing

- Convert emp_length column to int by removing special characters and alphabets

```
# Remove alphabets and extra characters and convert to numeric
data1['emp_length'].head(2)
```

```
[55] ✓ 0.5s
```

```
... 0    10+ years
     1    < 1 year
     Name: emp_length, dtype: object
```

```
# create new column 'emp_length_int' to get integer part from 'emp_length' column,
# possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
data1['emp_length_int'] = data1['emp_length'].str.replace(" years", '')
data1['emp_length_int'] = data1['emp_length_int'].str.replace(".*1 year", '0')
data1['emp_length_int'] = data1['emp_length_int'].str.replace("10\\+", '10')
```

```
[21] ✓ 0.8s
```

```
data1[['emp_length_int']]
```

```
[23] ✓ 0.5s
```

```
... emp_length_int
     0             10
     1              0
     2             10
```

Data standardizing

- Clean up int_rate column to int by removing '%' character and convert to numeric

```
# Cleanup -remove % sign and convert to int
data1['int_rate'] = data1['int_rate'].str.replace("%", '')
[29] ✓ 0.4s

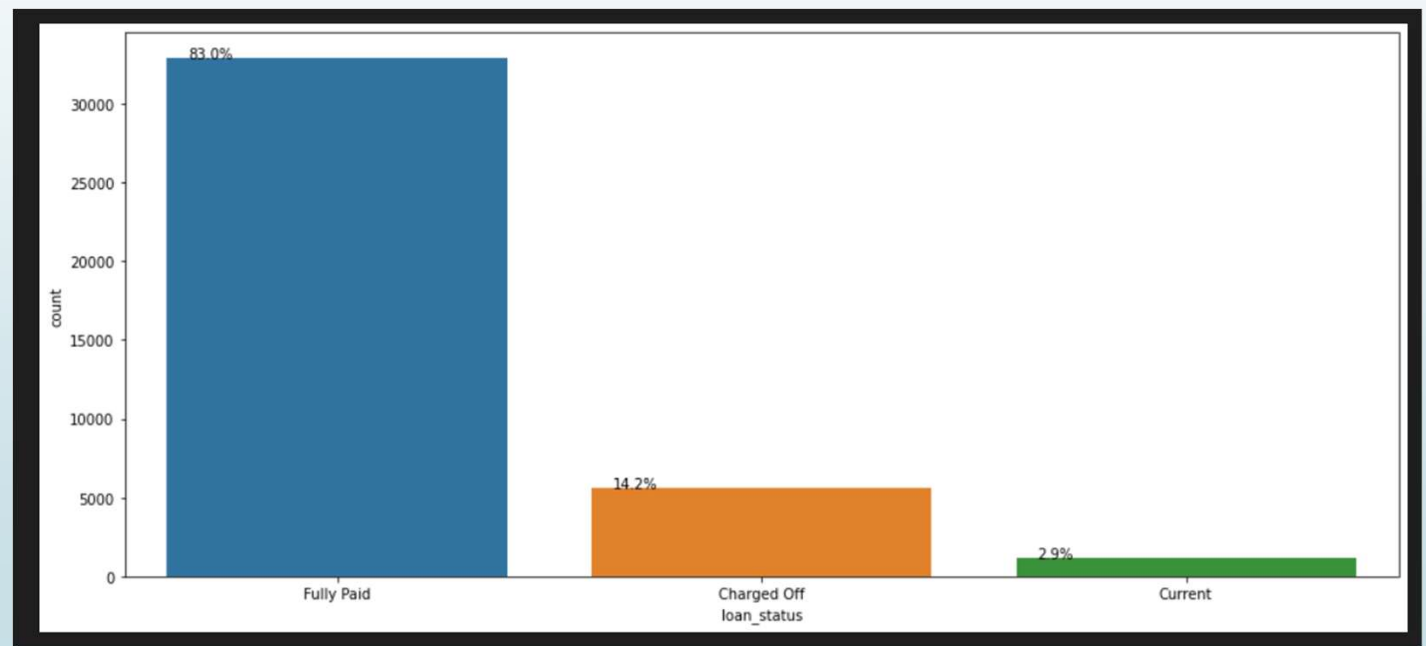
data1['int_rate'] = pd.to_numeric(data1['int_rate'])
[30] ✓ 0.6s

data1['int_rate'].head()
[31] ✓ 0.4s
... 0    10.65
    1    15.27
    2    15.96
    3    13.49
    4    12.69
    Name: int_rate, dtype: float64
```

Univariate analysis

Categorical column 'loan_status':

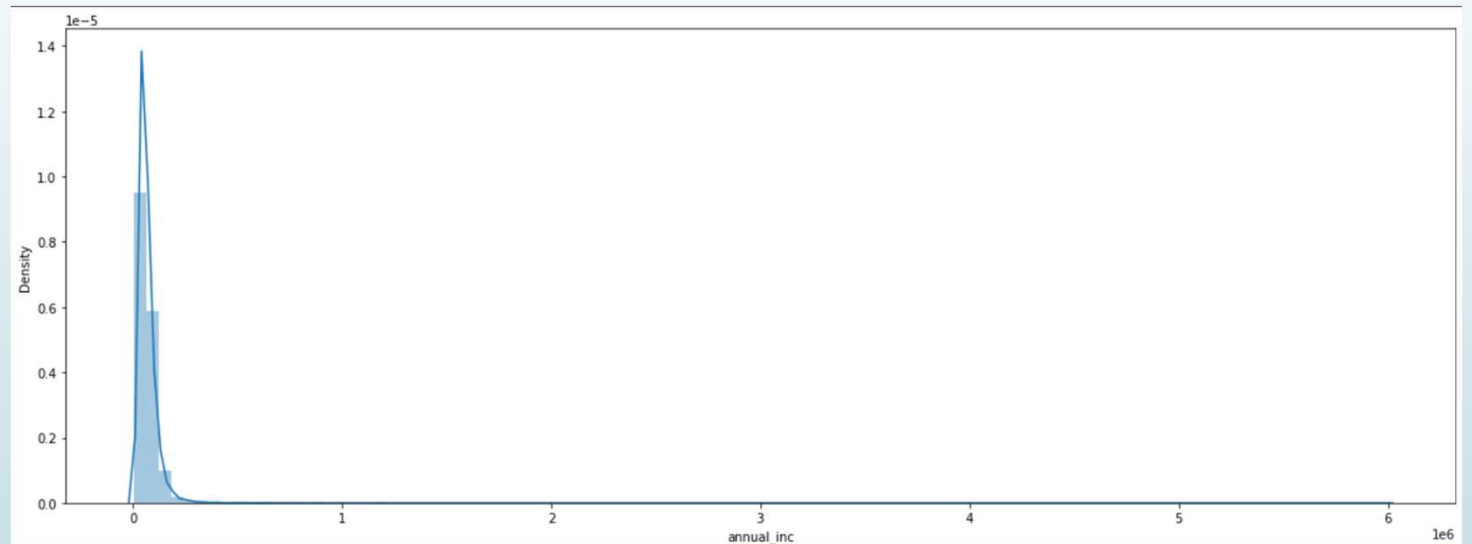
Out of 39717 applicants, 14.2% that is 5609 applicants defaulted.



Univariate analysis

Annual Income:

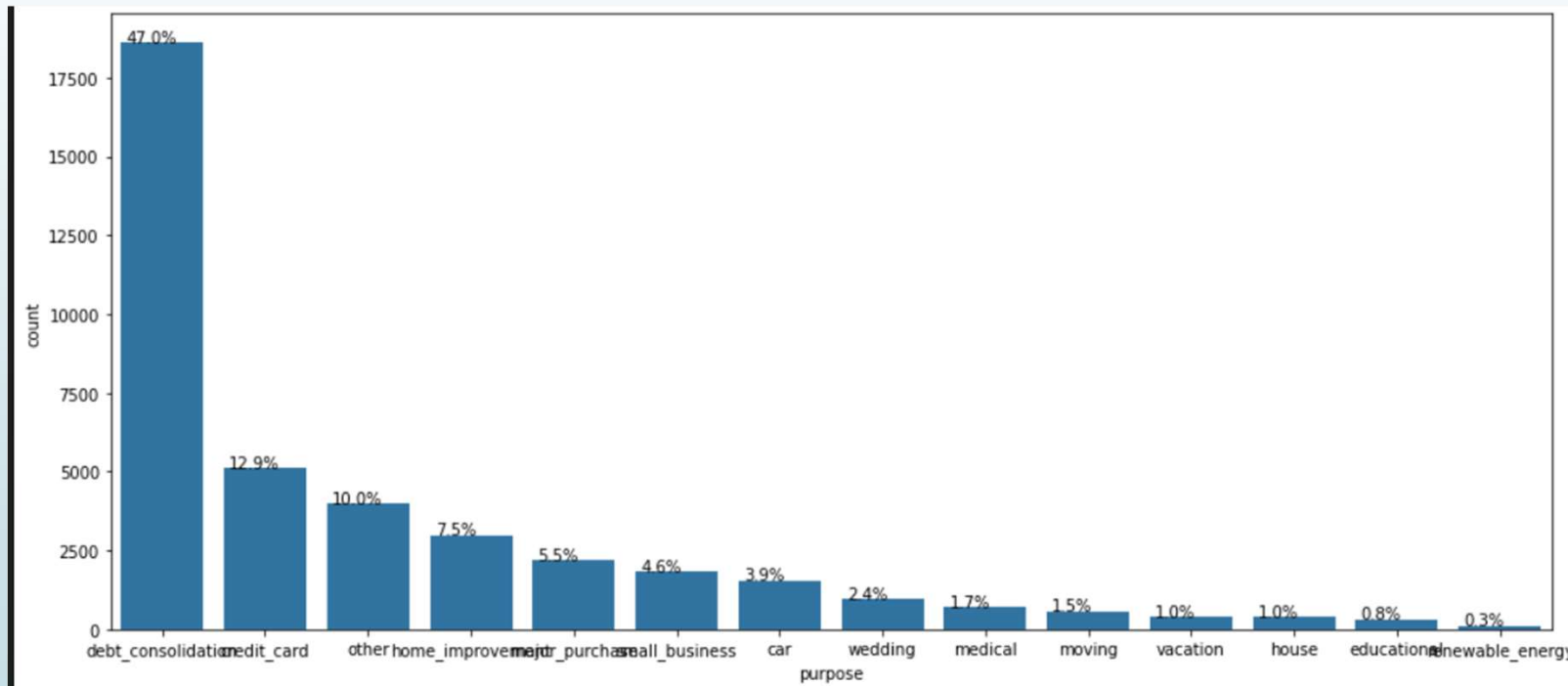
Most of applicants have income less than 50000



Univariate analysis

Loan Purpose:

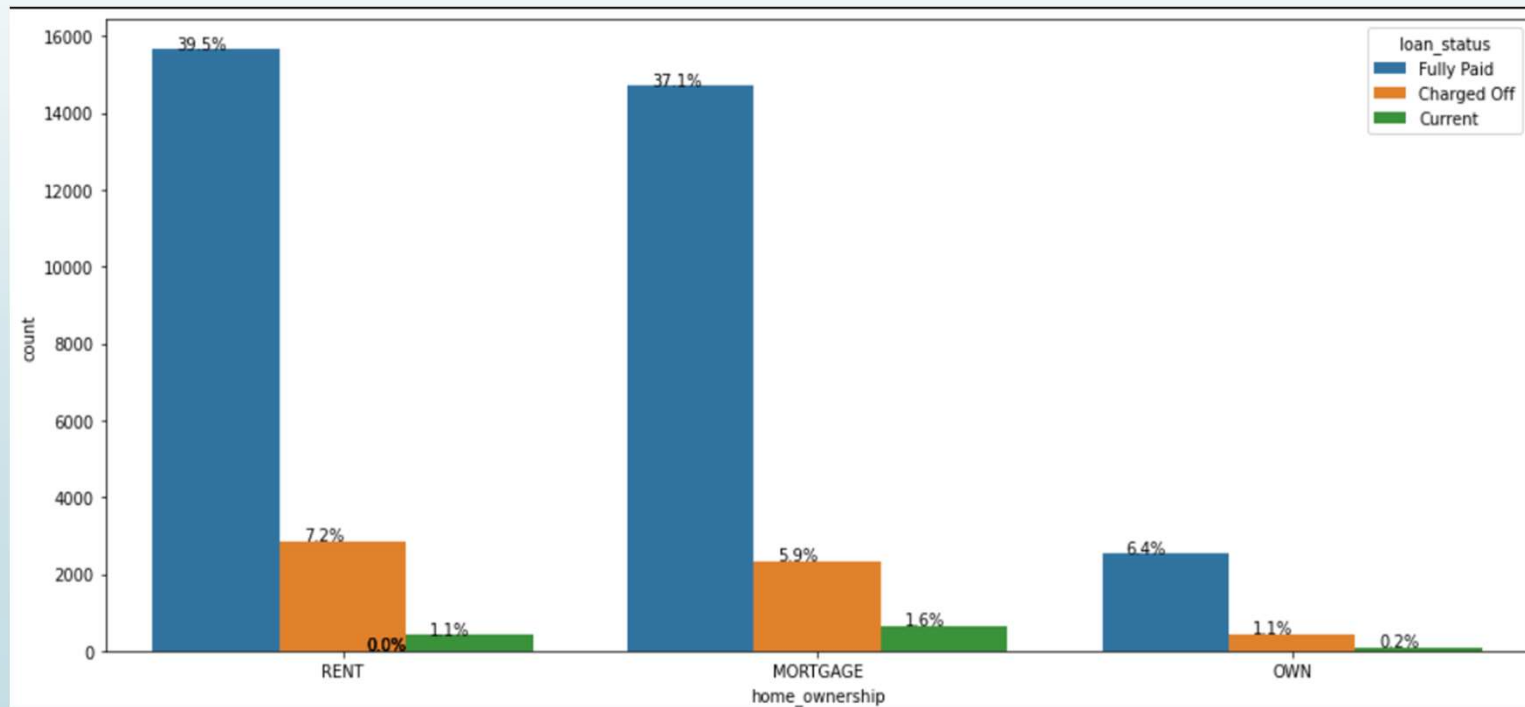
Majority of loan - 47% falls under debt consolidation category



Bivariate analysis

Home Ownership vs loan status of applicants:

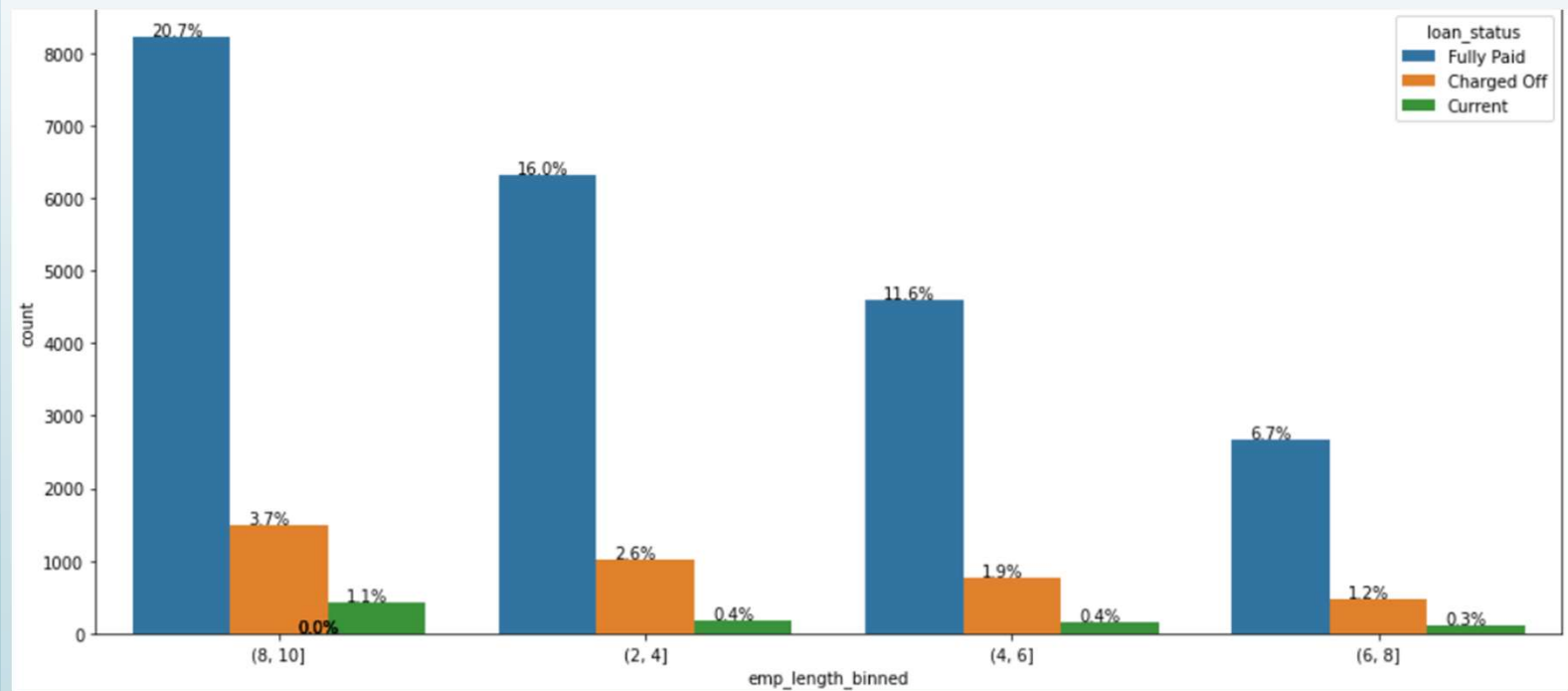
Approx 48% of applicants live in rented home and 45% of applicants have mortgage on their house. There are very less loan applications from house owners.



Bivariate analysis

Employment length vs Loan status of applicants:

There is not much difference in defaulters with respect employment length

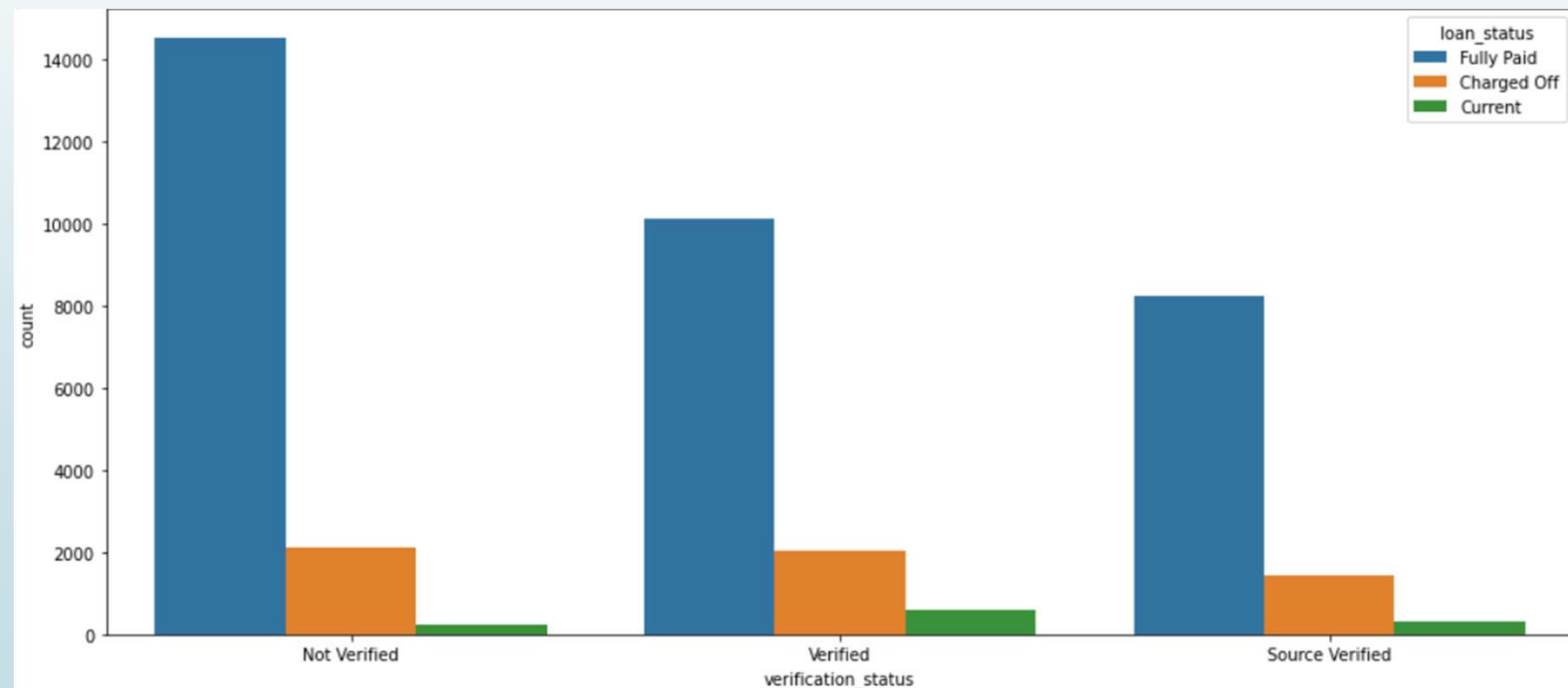


Univariate analysis

Application verification vs loan status:

From not verified, around 86% applicants fully repaid loan, around 12.64% charged off

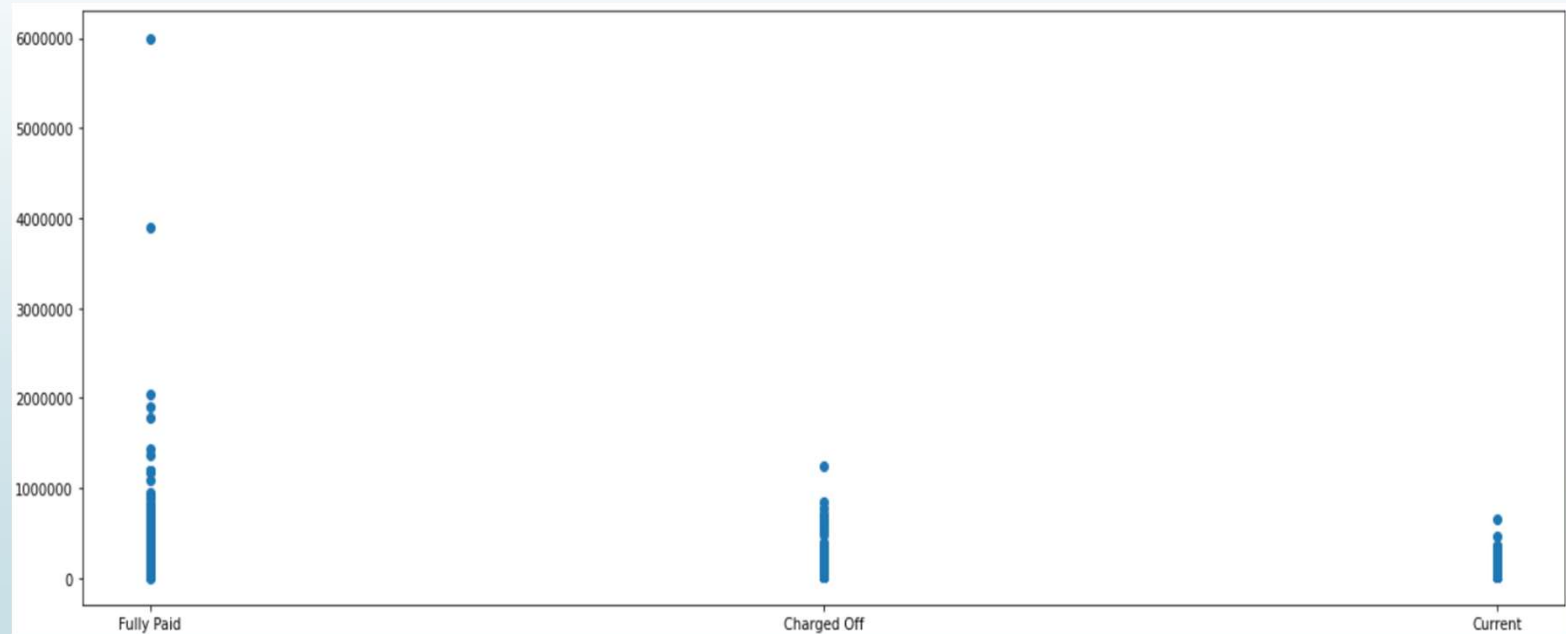
Charged off percentage is higher 16% in Source verified category, compared to other two.



Bivariate Analysis

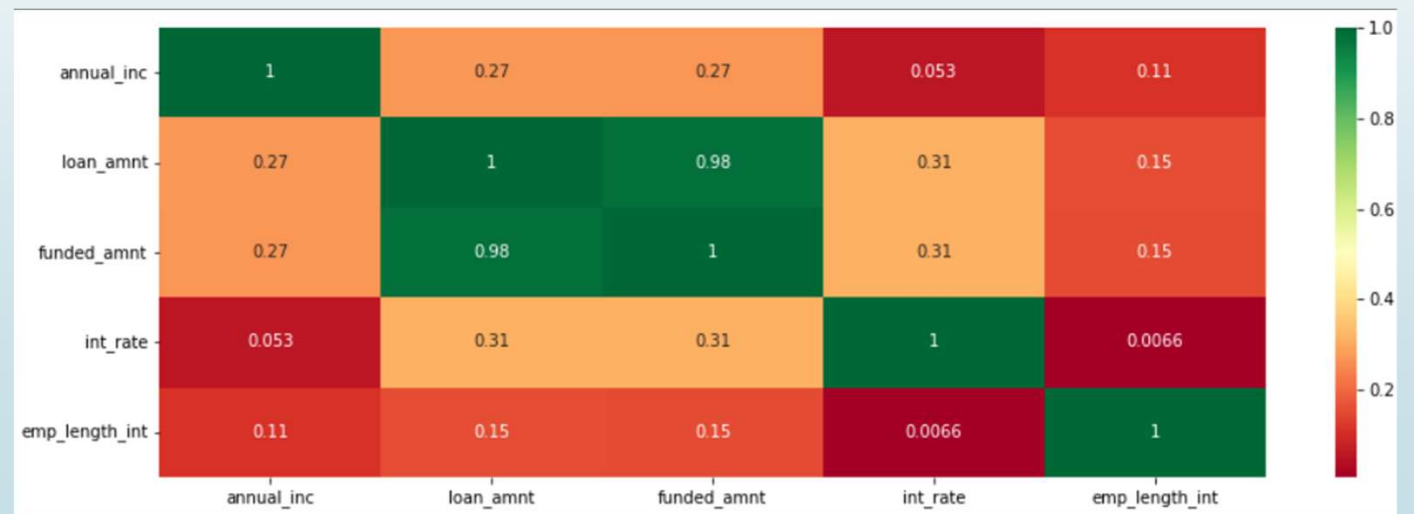
Annual Income:

All the defaulters have annual income below 150000



Correlation

- There is strong correlation between funded amount and loan amount
- Loan amount and funded amount shows positive correlation with Annual income
- Loan amount and funded amount shows positive correlation with interest rate
- Correlation is less between employment length and annual income, loan amount and funded amount.





Conclusion

Decisive Factor whether an applicant will be Repayer:

1. Annual_income: Annual income below 150000 have more defaults.
2. Verification status: Source verification has less defaults than not verified and verified.
3. Housing status with Rent is having more defaults.



Thank you