EDA CASE STUDY

PROBLEM STATEMENT

You work for a consumer finance company which specialises 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. Two types of risks are associated with the bank's decision:

- **If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company**
- **If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company**

The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending (to risky applicants) at a higher interest rate, etc.

DATA DESCRIPTION

When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. **Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:**
    - **Fully paid: Applicant has fully paid the loan (the principal and the interest rate)**
    - **Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.**
    - **Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan**
2. **Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)**

Objective

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 utilise this knowledge for its portfolio and risk assessment.

Data preprocessing

1. data cleaning
2. dealing with missing values
3. dropping unnecessary null values
4. modification of some columns

UNIVARIATE ANALYSIS

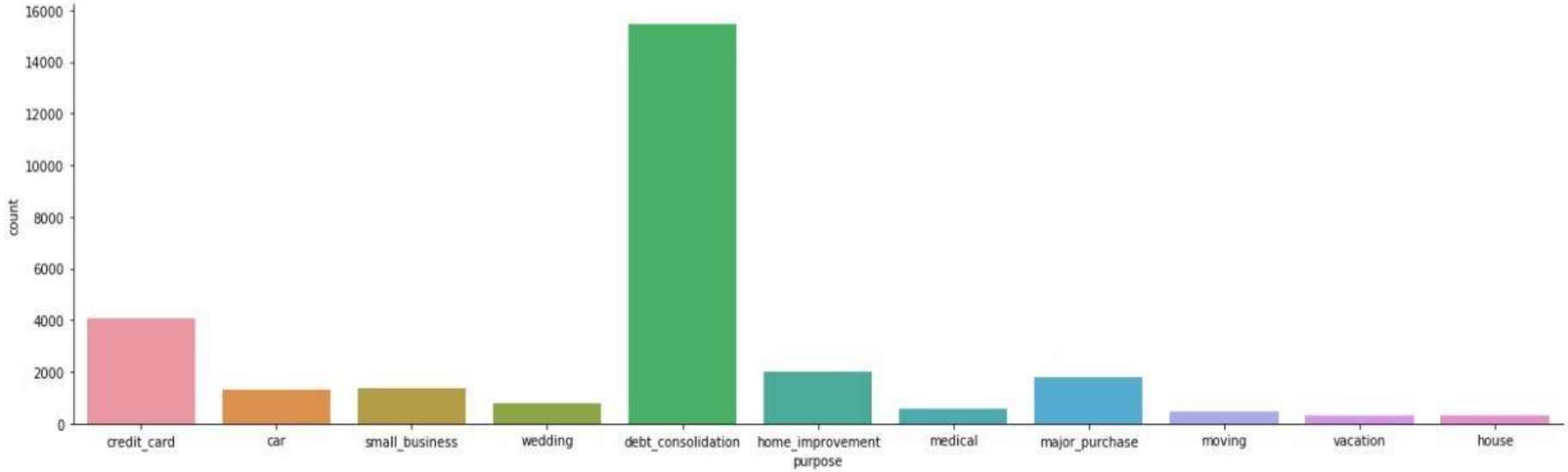
ON THE FOLLOWING COLUMNNS WE DEEP DIVE OUR ANALYSIS:

need_columns=['loan_amnt','funded_amnt','int_rate','home_ownership','annual_inc
','loan_status','purpose']

LOAN PURPOSE: 16000 people take debt for debt consolidation.

```
1 sns.catplot(x="purpose", kind="count", data=data,height=5, aspect=3.5)
2
3
```

```
<seaborn.axisgrid.FacetGrid at 0x21dce6b7388>
```



COUNT LOAN STATUS

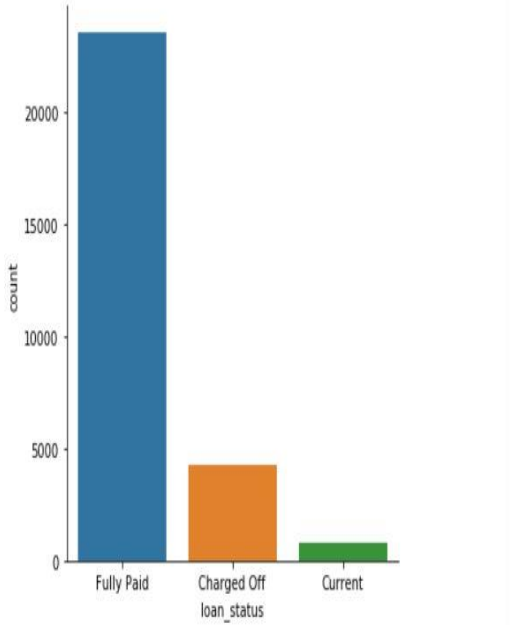
Insights: From above graph 2% customer chrged_off from the distribution.

```
1  ##count_loan status
2  data.loan_status.value_counts()*100/len(data)
3
```

```
Fully Paid     82.874233
Charged Off    14.339682
Current         2.786085
Name: loan_status, dtype: float64
```

```
1  sns.catplot(x="loan_status", kind="count", data=data)
```

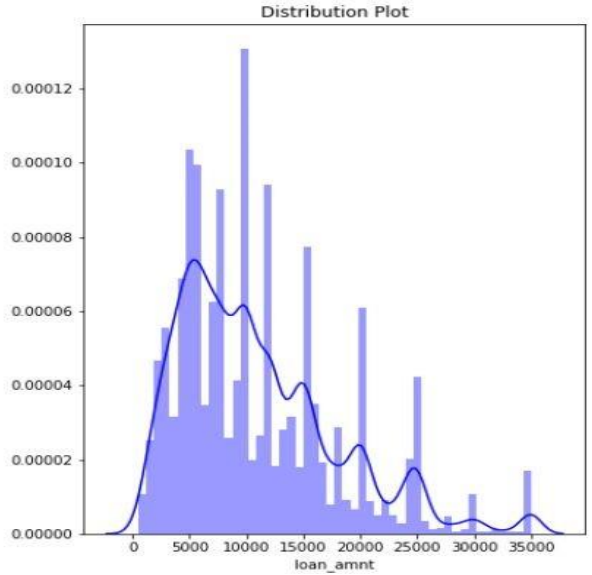
```
<seaborn.axisgrid.FacetGrid at 0x21dcc479c48>
```

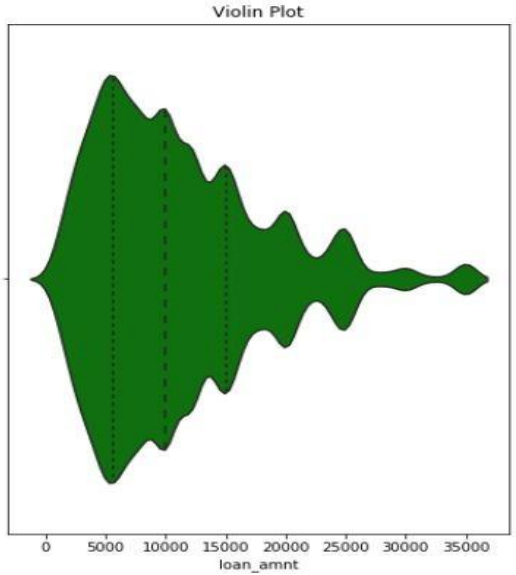


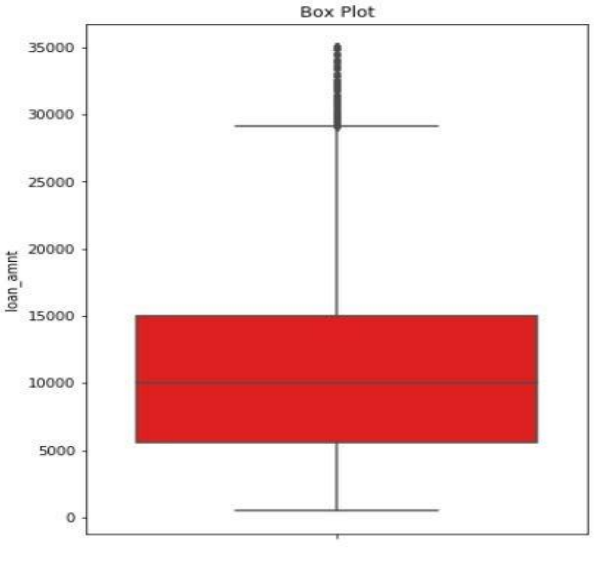
Loan Amount Analysis: **75% of the loan amount lie between 5000-15000**

```
## plotting loan amount
fig,ax=plt.subplots(nrows=1,ncols=3,figsize=(20,8))
ax[0].set_title("Distribution Plot")
sns.distplot(data['loan_amnt'],ax=ax[0],color='blue')
ax[1].set_title("Violin Plot")
sns.violinplot(data=data,x='loan_amnt',ax=ax[1],inner='quartile',color='green')
ax[2].set_title("Box Plot")
sns.boxplot(data=data,x='loan_amnt',ax=ax[2],orient='v',color='red')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x21dc58b1c08>
```







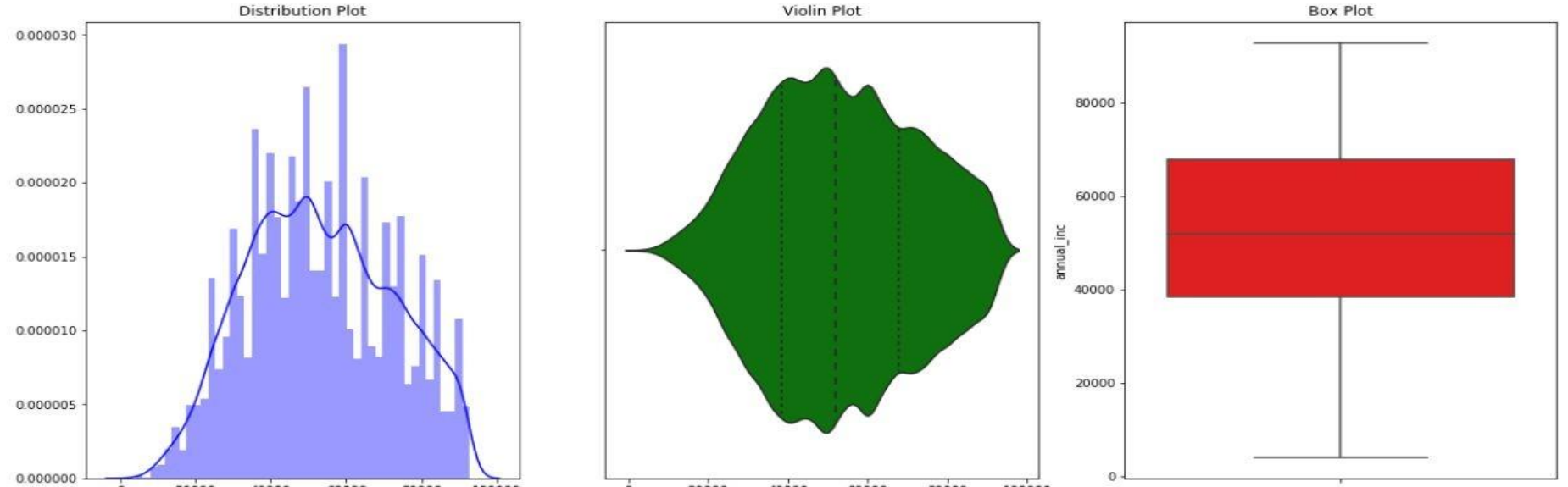
Annual Income Analysis:

75% of the data have anual income between 40000 - 78000 anually.

Bivariate analysis

```
fig,ax=plt.subplots(nrows=1,ncols=3,figsize=(20,8))
ax[0].set_title("Distribution Plot")
sns.distplot(data['annual_inc'],ax=ax[0],color='blue')
ax[1].set_title("Violin Plot")
sns.violinplot(data=data,x='annual_inc',ax=ax[1],inner='quartile',color='green')
ax[2].set_title("Box Plot")
sns.boxplot(data=data,x='annual_inc',ax=ax[2],orient='v',color='red')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x21dc56d1708>
```



Bivariate Analysis

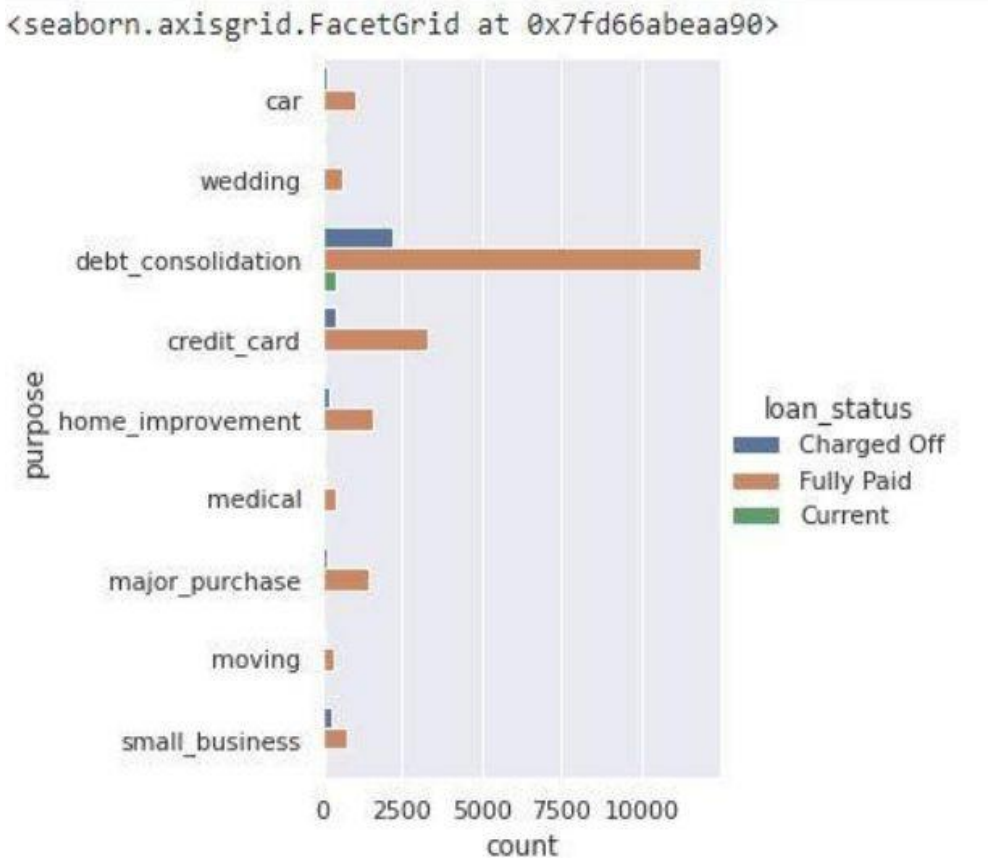
1. purpose vs loan status
2. home_ownership vs loan_status
3. loan_status vs term
4. homeownership vs loan status
5. loan_status vs term
6. loan_status vs verification status

Analysis between purpose and loan status .

Purpose can impact loan status and

help to company which loan are likely

To paid of

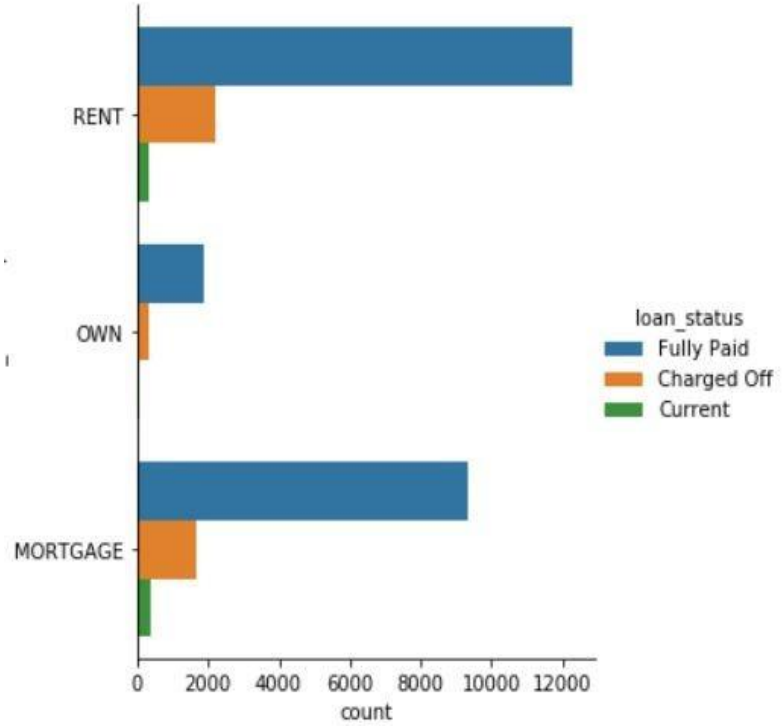


home_ownership vs loan_status

33% rent customers fully paid there debt

This will help company to take home ownership as a factor for giving debt to a customers or not .

```
sns.catplot(y='home_ownership', hue="loan_status", kind
eaborn.axisgrid.FacetGrid at 0x21dc5edef08>
```



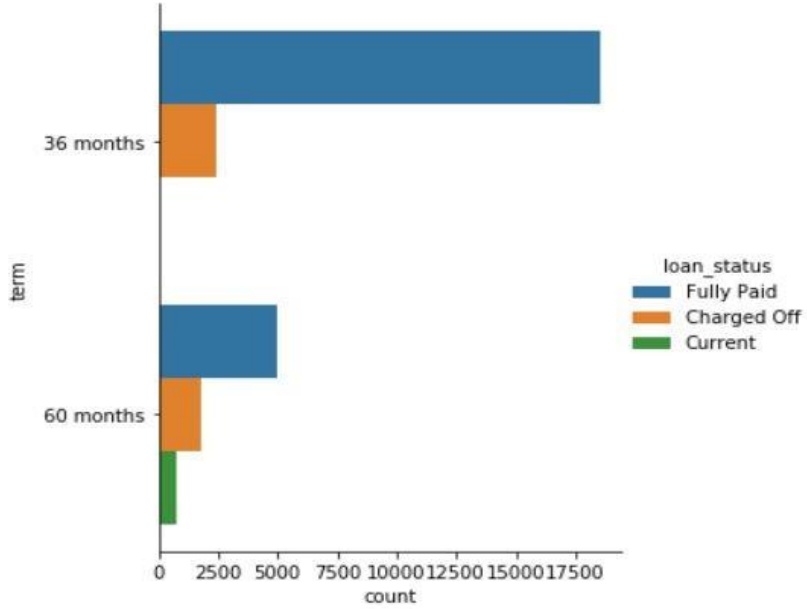
loan_status vs term

Duration of paying loan may be a very good factor for Company to give a loan.

As we observed short term period debts are 68% fully paid.

```
1 sns.catplot(y='term', hue="loan_status", kind="count",data=data)
```

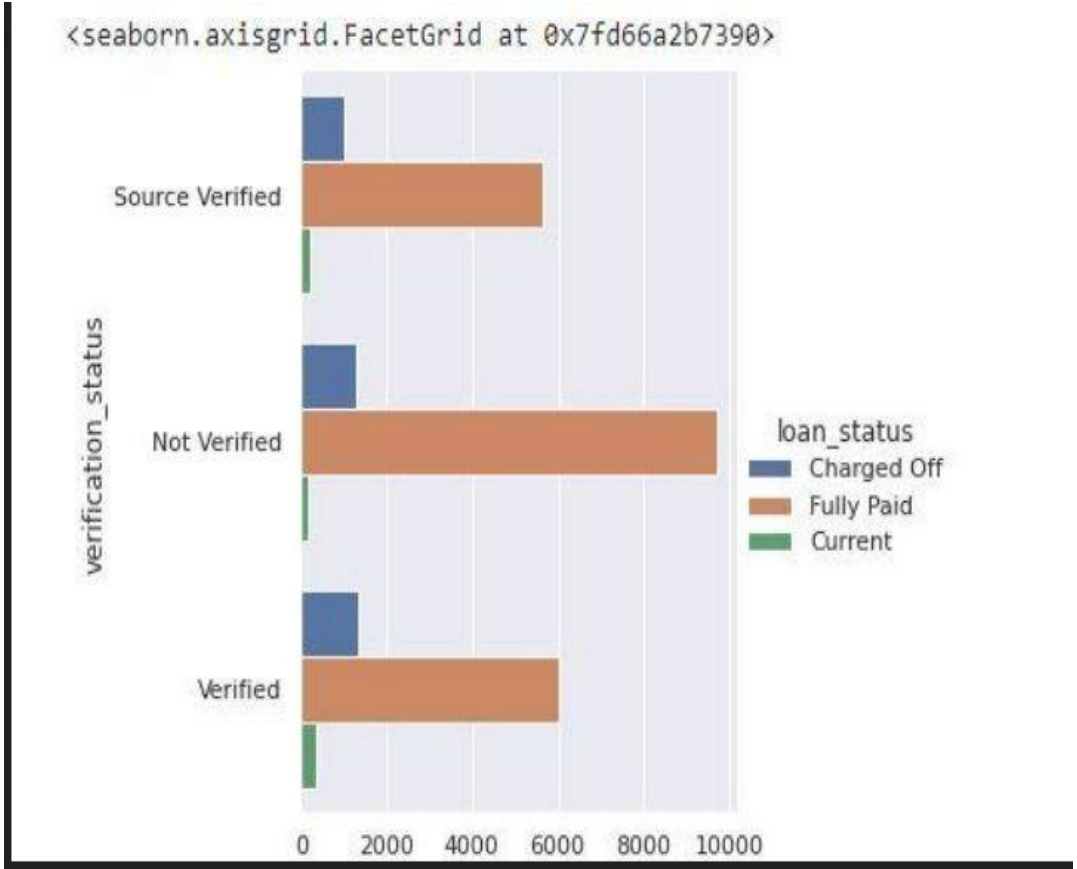
```
<seaborn.axisgrid.FacetGrid at 0x21dc5cb7fc8>
```



68% customers fully paid the loan in the term period of 36months

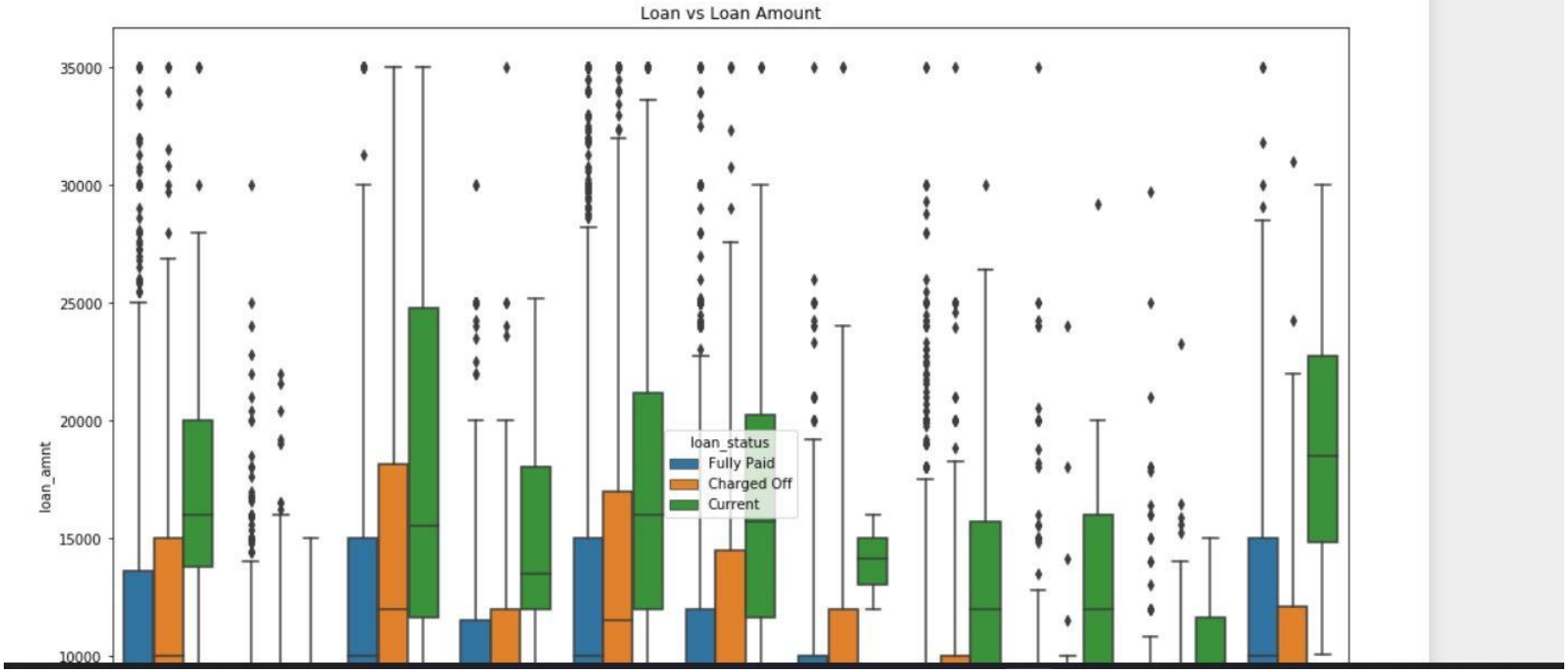
loan_status vs verification_status

Verified sources have the most defaulters



Data vs loan amount

```
:
1  plt.figure(figsize=(16,12))
2  sns.boxplot(data =data, x='purpose', y='loan_amnt', hue ='loan_status')
3  plt.title('Loan vs Loan Amount')
4  plt.show()
```



Correlation

Purpose and verification

status seems to have strong correlation

Homeownership doesn’t seem any correlation

Recommendations

1.)Company can focus on small loan like debit consolidation (credit card),small loans and purchases

2.)Process of income should be more accurate as it important parameter

3.)Target based/last minute loan disbursal should be avoided

Finish