Lending club case study

Contents

Table of Contents

Study for Loan Data

❖ Data Sourcing

- **Import the useful libraries**
  - Numpy
  - Pandas
  - Seaborn
  - Matplotlib.pyplot

- **Read the Data set**

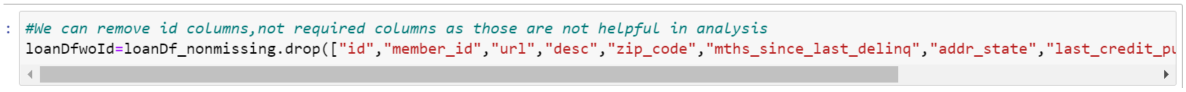
- **Apply Filter to keep only required records for the analysis**

  loanDf=loan[(loan.loan_status.str.strip() =='Fully Paid') | (loan.loan_status.str.strip() =='Charged Off')

❖ Data Cleaning

- **Dropping unnecessary columns**

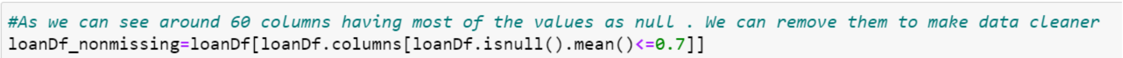
  Drop columns that are not required in analysis



```
#We can remove id columns,not required columns as those are not helpful in analysis
loanDfwoId=loanDf_nonmissing.drop(["id","member_id","url","desc","zip_code","mths_since_last_delinq","addr_state","last_credit_pu
```

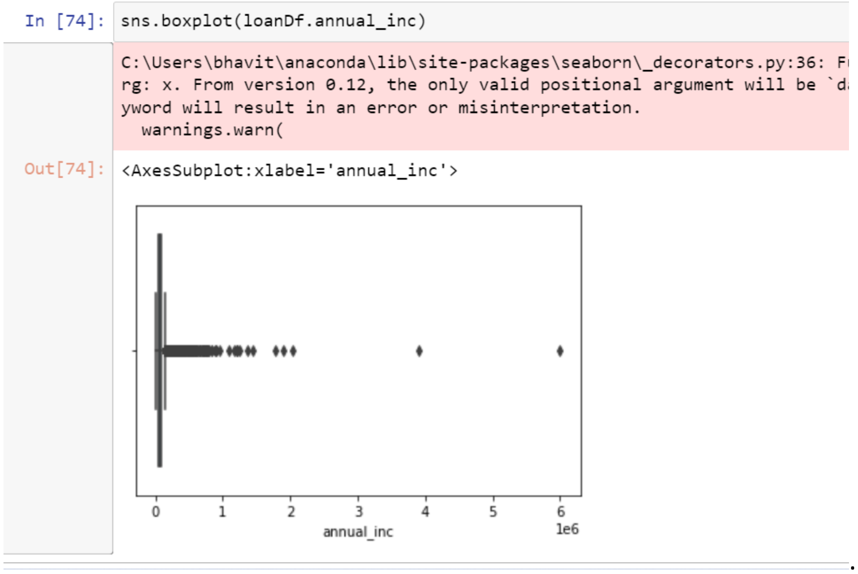
- **Impute/Remove missing values**

  1. calculate the percentage of the missing values
  2. drop the columns having more than 30% missing

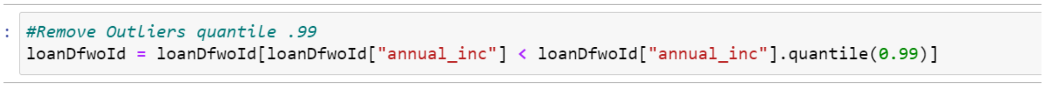


```
#As we can see around 60 columns having most of the values as null . We can remove them to make data cleaner
loanDf_nonmissing=loanDf[loanDf.columns[loanDf.isnull().mean()<=0.7]]
```

  3. Imputing missing values for emp_length. We replaced missing with 0.
  4. Imputing pub_rec_bankruptcies with ‘unknown’ category.
  5. Remove outliers on annual income as we can see from boxplot:

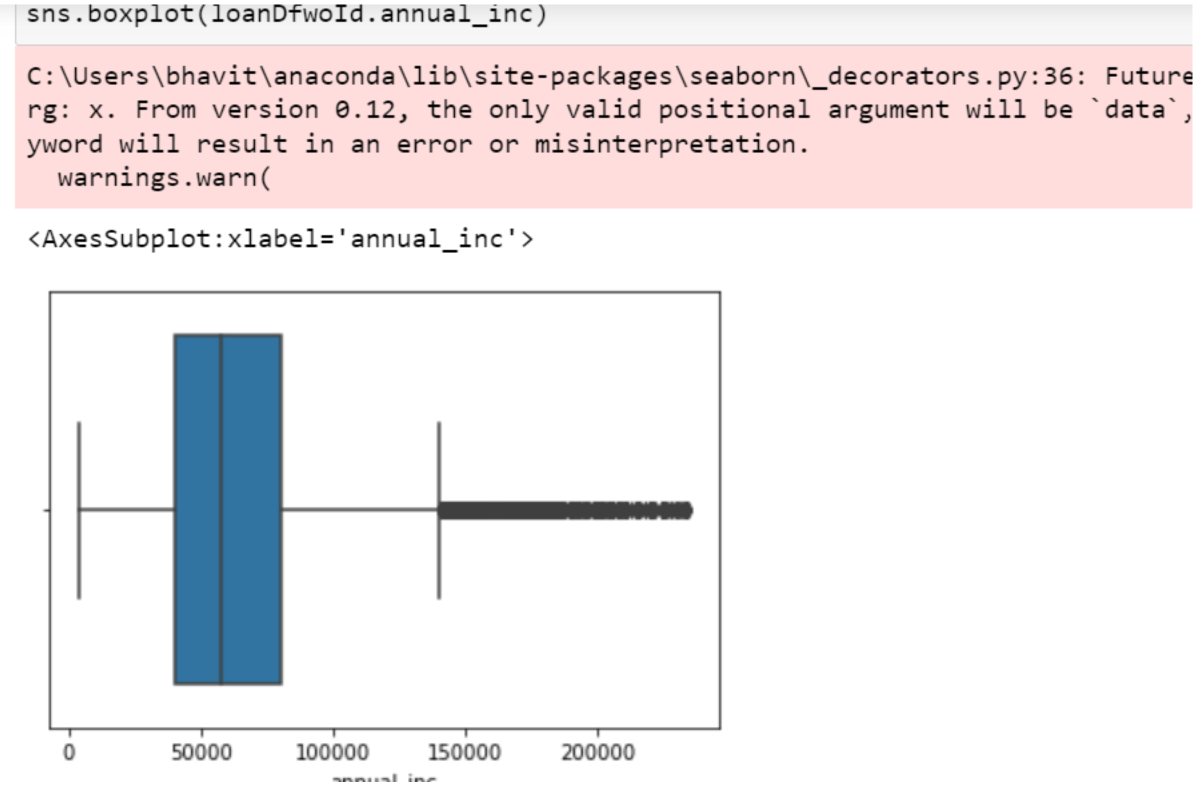


```
In [74]: sns.boxplot(loanDf.annual_inc)

C:\Users\bhavit\anaconda\lib\site-packages\seaborn\_decorators.py:36: Fu
rg: x. From version 0.12, the only valid positional argument will be `da
yword will result in an error or misinterpretation.
  warnings.warn(

Out[74]: <AxesSubplot:xlabel='annual_inc'>
```



```
#Remove Outliers quantile .99
loanDfwoId = loanDfwoId[loanDfwoId["annual_inc"] < loanDfwoId["annual_inc"].quantile(0.99)]
```

Post removal of outliers:

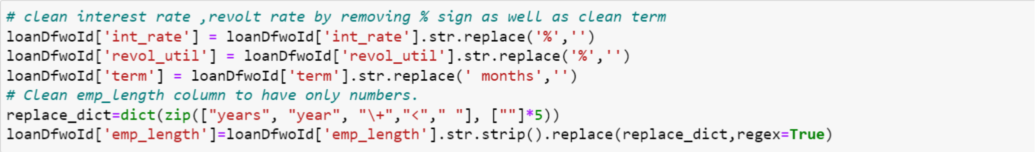


```
sns.boxplot(loanDfwoId.annual_inc)
```

```
C:\Users\bhavit\anaconda\lib\site-packages\seaborn\_decorators.py:36: Future
rg: x. From version 0.12, the only valid positional argument will be `data`,
yword will result in an error or misinterpretation.
  warnings.warn(
```

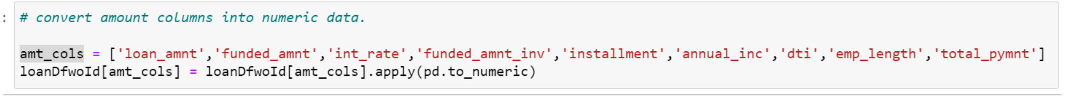
```
<AxesSubplot:xlabel='annual_inc'>
```

6. clean interest rate ,revolt rate by removing % sign as well as clean term.



```
# clean interest rate ,revolt rate by removing % sign as well as clean term
loanDfwoId['int_rate'] = loanDfwoId['int_rate'].str.replace('%','')
loanDfwoId['revol_util'] = loanDfwoId['revol_util'].str.replace('%','')
loanDfwoId['term'] = loanDfwoId['term'].str.replace(' months','')
# Clean emp_length column to have only numbers.
replace_dict=dict(zip(["years", "year", "\+","<"," "], [""]*5))
loanDfwoId['emp_length']=loanDfwoId['emp_length'].str.strip().replace(replace_dict,regex=True)
```

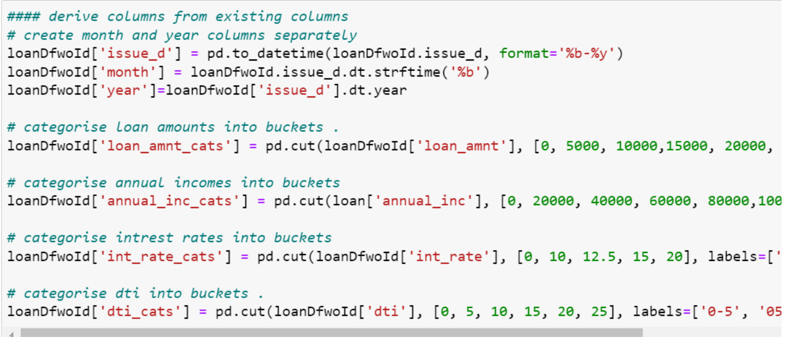
7. convert amount columns into numeric data. So that we can see the correlation between these columns.



```
# convert amount columns into numeric data.

amt_cols = ['loan_amnt','funded_amnt','int_rate','funded_amnt_inv','installment','annual_inc','dti','emp_length','total_pymnt']
loanDfwoId[amt_cols] = loanDfwoId[amt_cols].apply(pd.to_numeric)
```

- **Derive new columns from existing ones**

Derived new columns from the existing one like month and year from issue date. Also created bucketed column on loan amount, interest rates, debt to income ratio(dti) as well as on annual income.

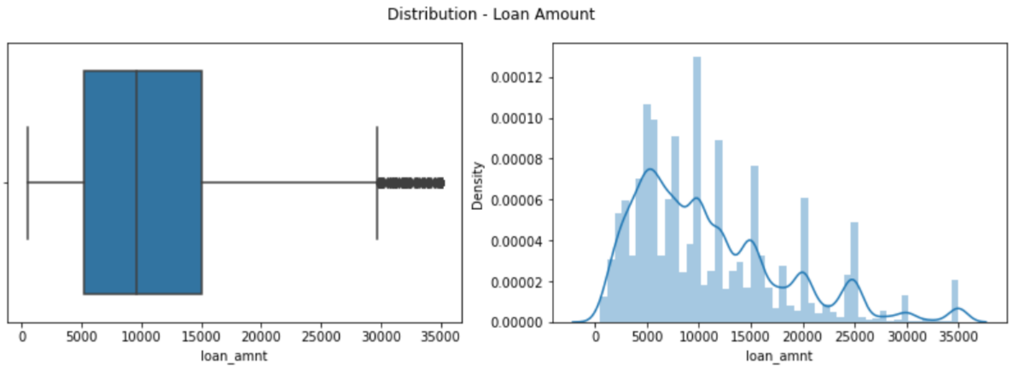


```
#### derive columns from existing columns
# create month and year columns separately
loanDfwoId['issue_d'] = pd.to_datetime(loanDfwoId.issue_d, format='%b-%y')
loanDfwoId['month'] = loanDfwoId.issue_d.dt.strftime('%b')
loanDfwoId['year']=loanDfwoId['issue_d'].dt.year

# categorise loan amounts into buckets .
loanDfwoId['loan_amnt_cats'] = pd.cut(loanDfwoId['loan_amnt'], [0, 5000, 10000,15000, 20000,

# categorise annual incomes into buckets
loanDfwoId['annual_inc_cats'] = pd.cut(loan['annual_inc'], [0, 20000, 40000, 60000, 80000,100

# categorise intrest rates into buckets
loanDfwoId['int_rate_cats'] = pd.cut(loanDfwoId['int_rate'], [0, 10, 12.5, 15, 20], labels=['

# categorise dti into buckets .
loanDfwoId['dti_cats'] = pd.cut(loanDfwoId['dti'], [0, 5, 10, 15, 20, 25], labels=['0-5', '05
```

❖ Analysis

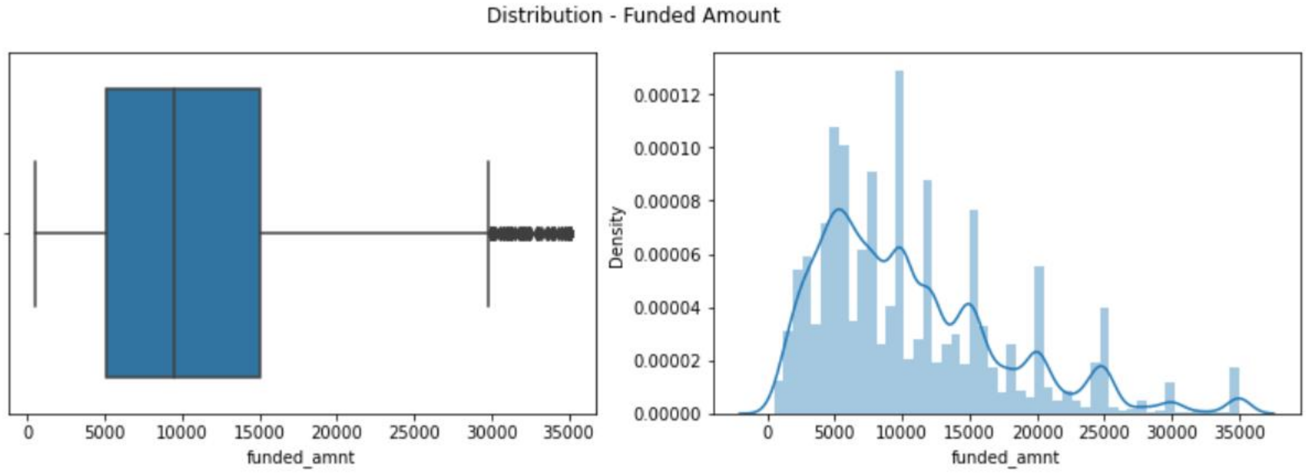
Univariate Analysis

- **Distribution of loan amount**



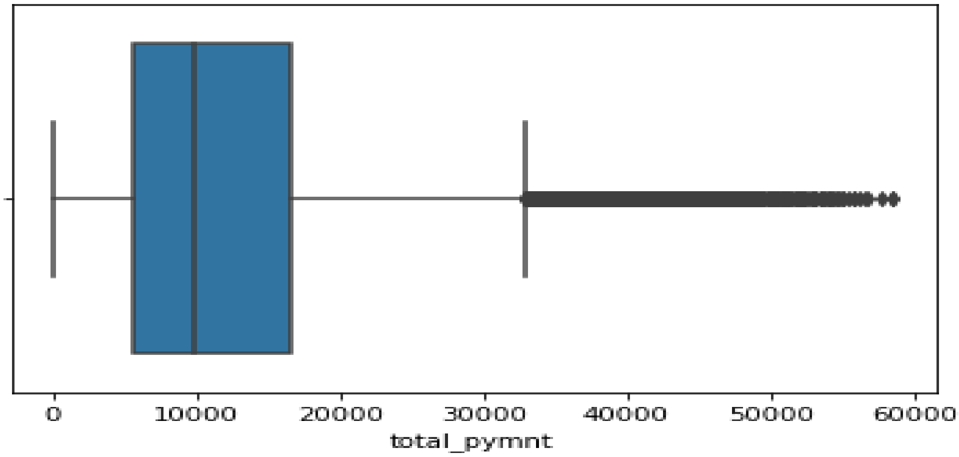
From the above plots, we can see that loan amount is varying between 5000 and 15000 for almost 50 % people and median is around 9600. 90 percent of loans are below 21,000.

- **Distribution of funded amount:**



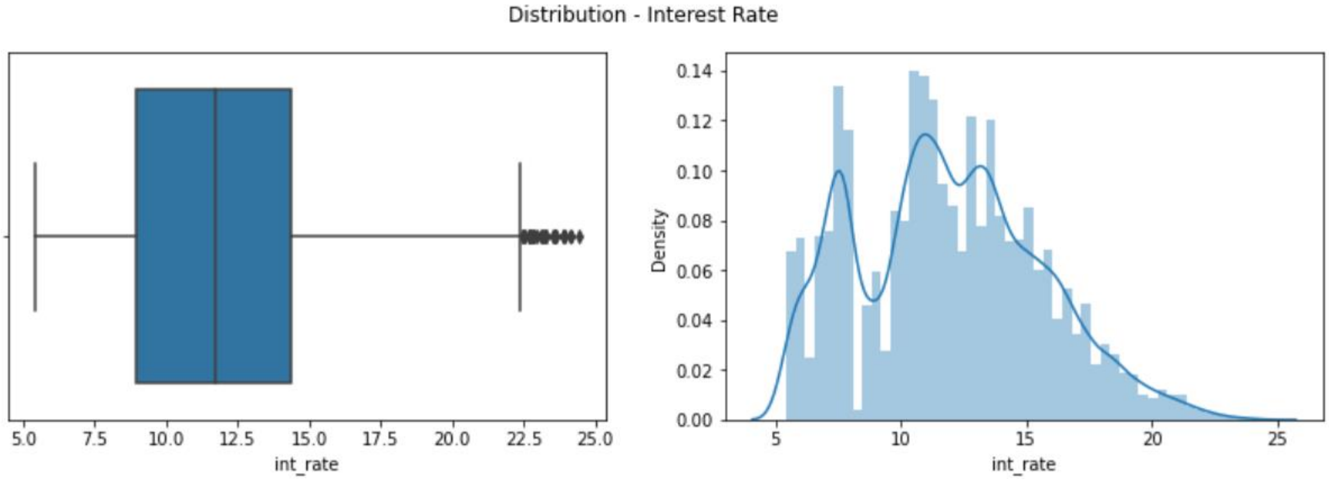
Observations:
Funded amount data has same distribution as of loan Amount, so we can say that approved loan is almost same as Applied loan amount.

- **Distribution of total payment:**



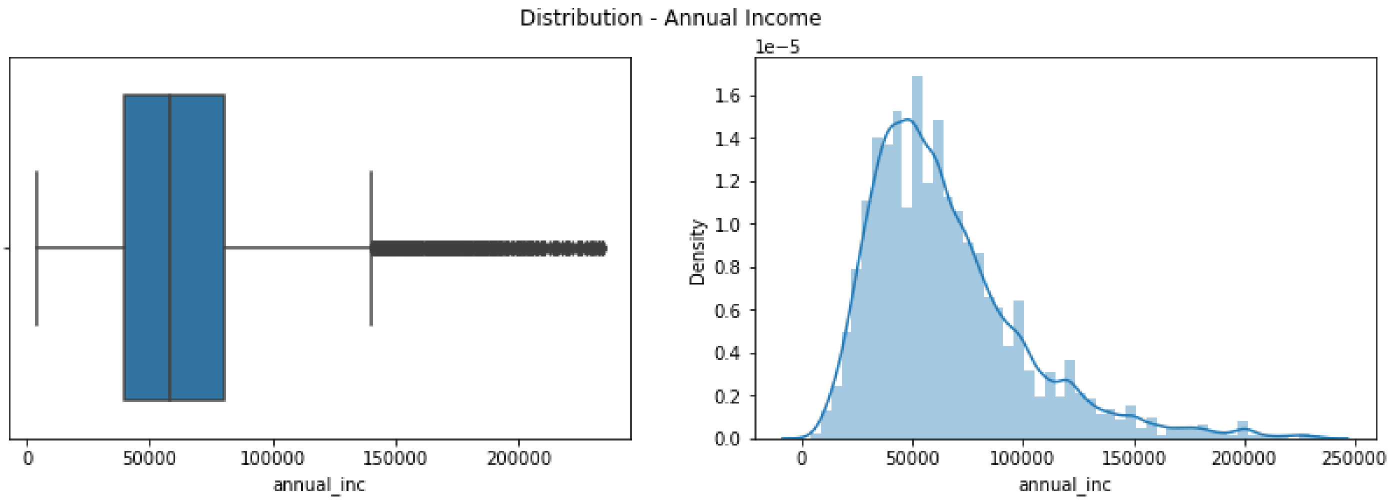
Payment amount data shows variation between 6000 and 16000 for 50% people, so we can say that loan having around 10% return on loan_amount (as its having variation between 5000 and 15000) .

- **Distribution of interest rates:**



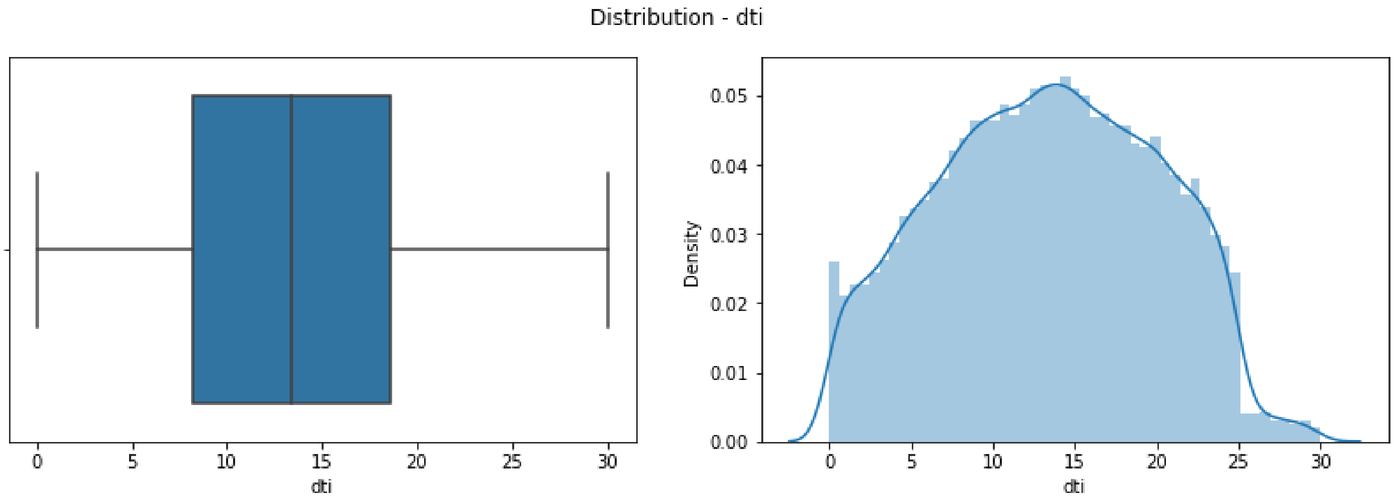
Observations:
From the above 2 plots of interest rates we can conclude that most of the interest rates lies are in the range of **9% to 14.5%**. There are some exceptions/outliars i.e., **22.5+** %

- **Distribution of Annual Income:**



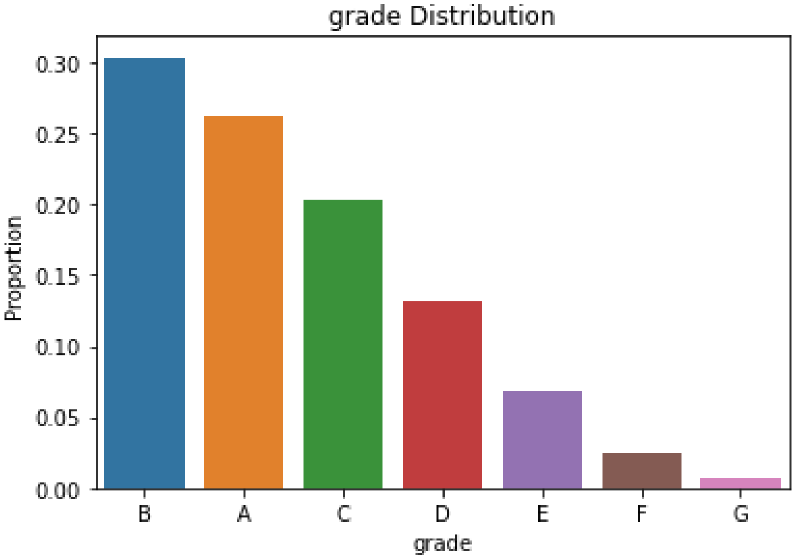
Observations:
Business observation: Most of the people having annual income between 40k and 80K who got loan approved.

- **Distribution of Debt-to-Income ratio:**



Observations:
As you can see dti is between 2 and 30. So, it’s a healthy sign that loan is disbursed to good saving people so that loan defaulters can be less.

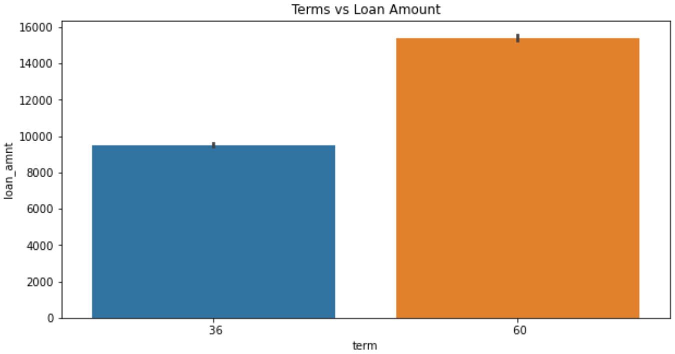
- **Analysis of grade:**

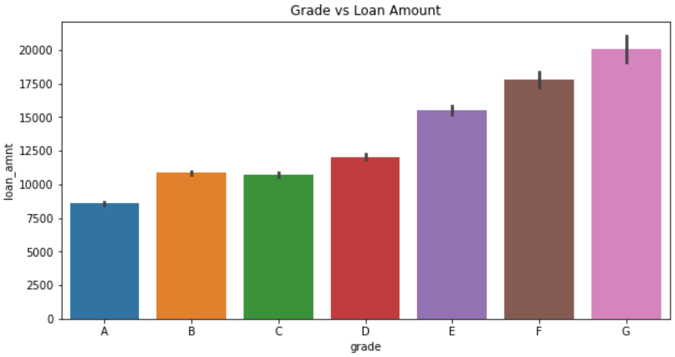


Observations: As you can see most of the borrower having B and A grades.

Segmented Univariate Analysis

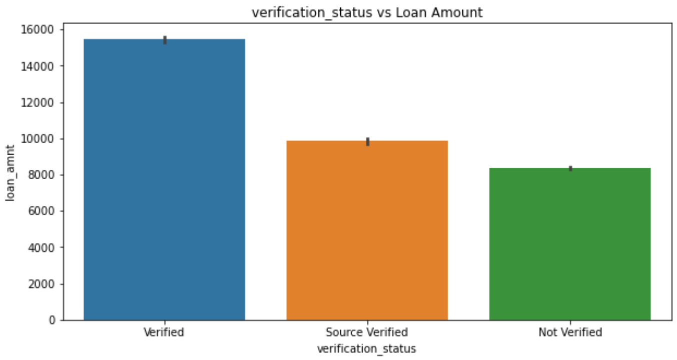
- **Analysis of loan amount with respect to Term and Grade :**

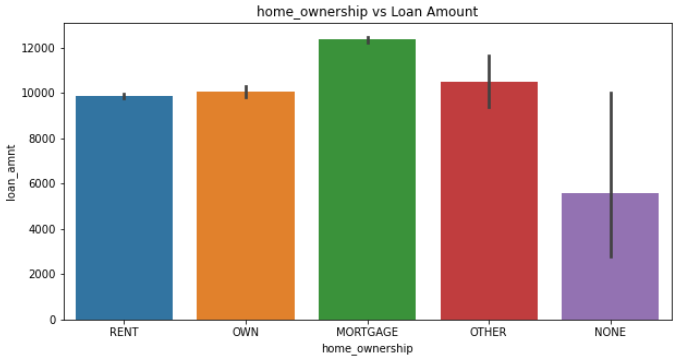




Observations: AS you can see Higher amount loans have high term. Grade 'G' and 'F' have taken max loan amount. As Grades are decreasing the loan amount is increasing.

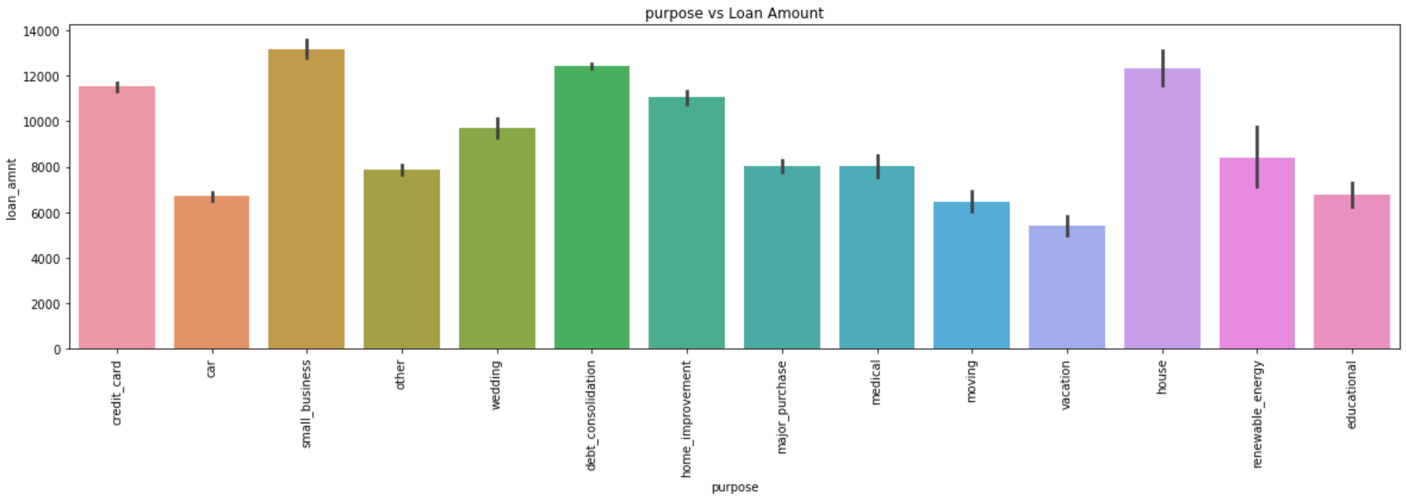
- **Analysis of loan amount with respect to verification status and home ownership:**





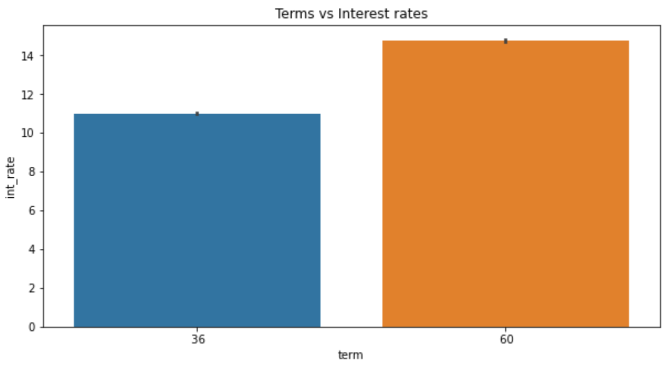
Observations: Verified applicants have higher loans. Mortgage category having higher loan amount.

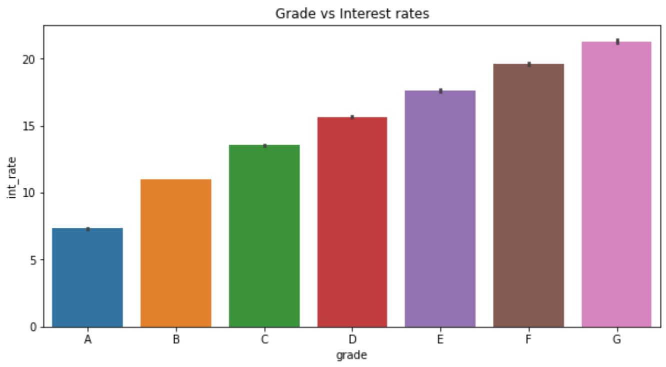
- **Analysis of loan amount with respect to purpose:**



Observations: small business and house purchase have higher loan amount.

- **Analysis of interest rates with respect to Terms and Grade:**

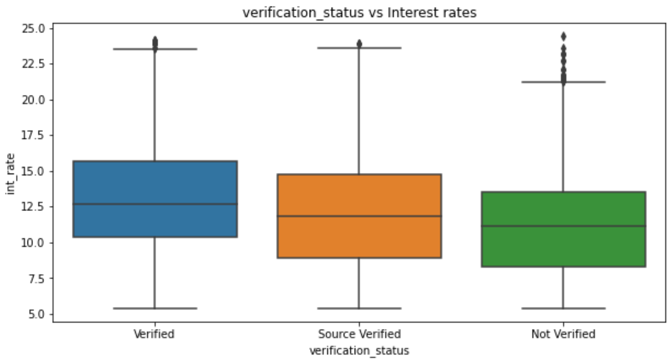


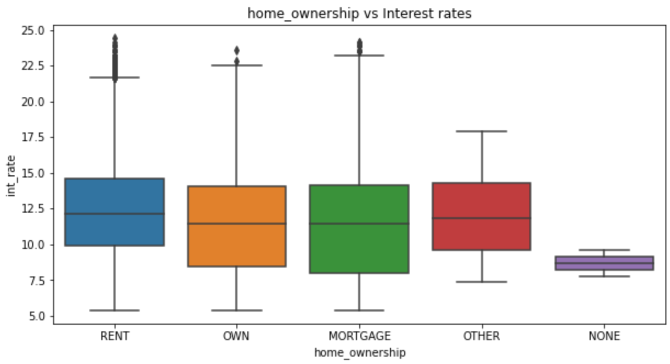


Observations:

- Higher term having high interest rates.
- Lower grades having higher interest rates.

- **Analysis of interest rates with respect to home_ownership:**

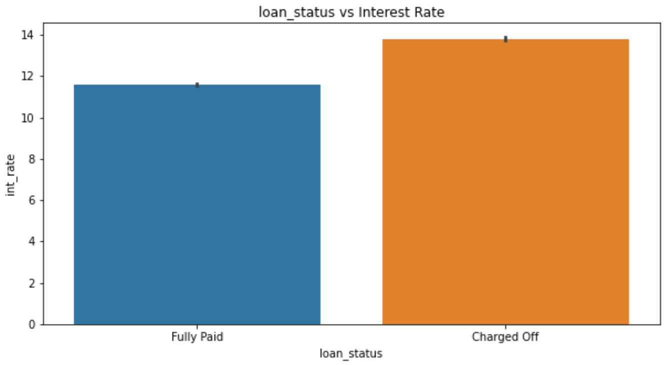


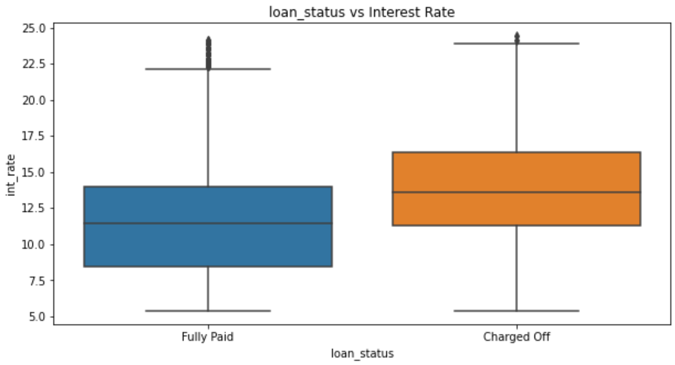


Observations:

- There is a small variation in interest rates for different categories of verification status.
- For Mortgage category of home ownership we do have lower interest rates.

- **Analysis of interest rates with respect to home_ownership :**

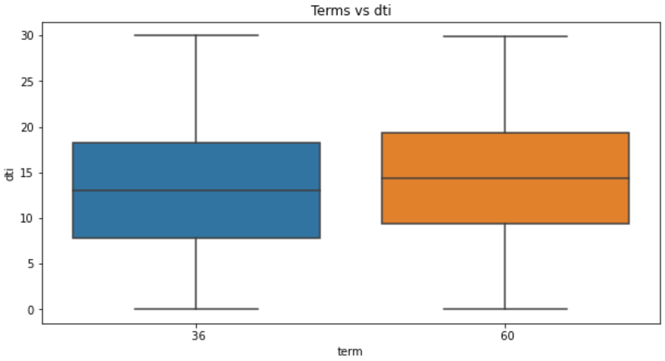


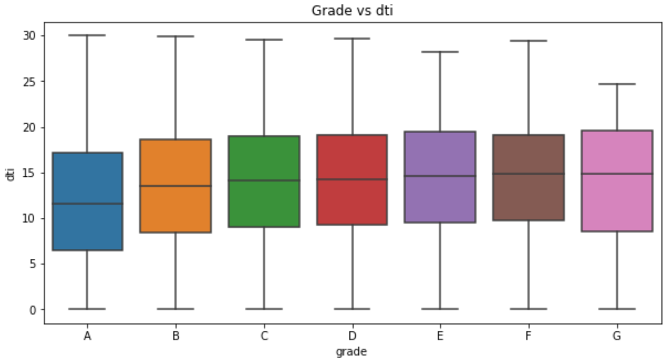


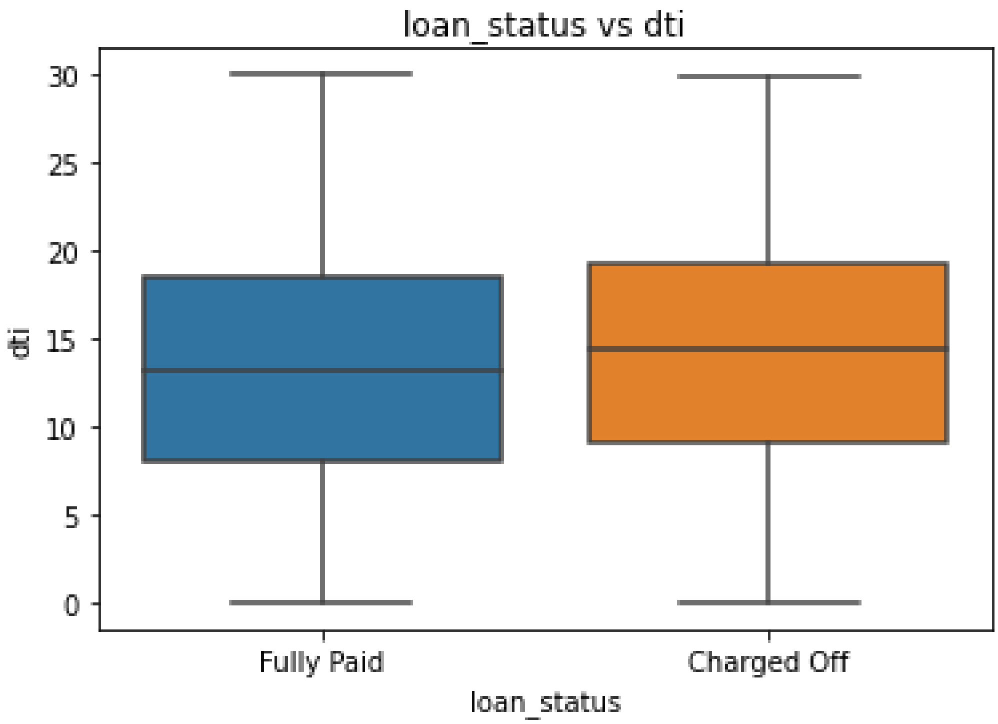
Observations:

- Most defualters had high interest rates.

- **Analysis of DTI w.r.t to loan_status , terms and grades :**





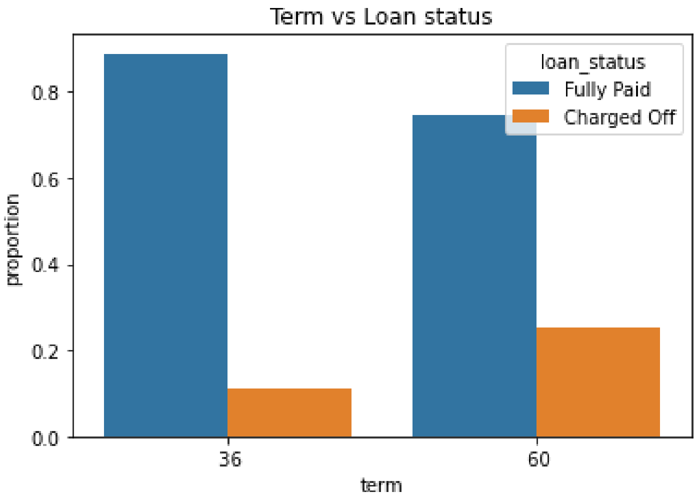


Observations:

- debt to income ratio is lower for lower loan duration or we say with lower dti applicants mostly take lower duration loans.
- As dti increasing grading is decreasing (from A to G) so higher dti means lower grading from above data snapshot
- Higher dti is converting applicant to defulter category.i.e. Borrowers with high DTI has more probability to default

Bivariate Analysis

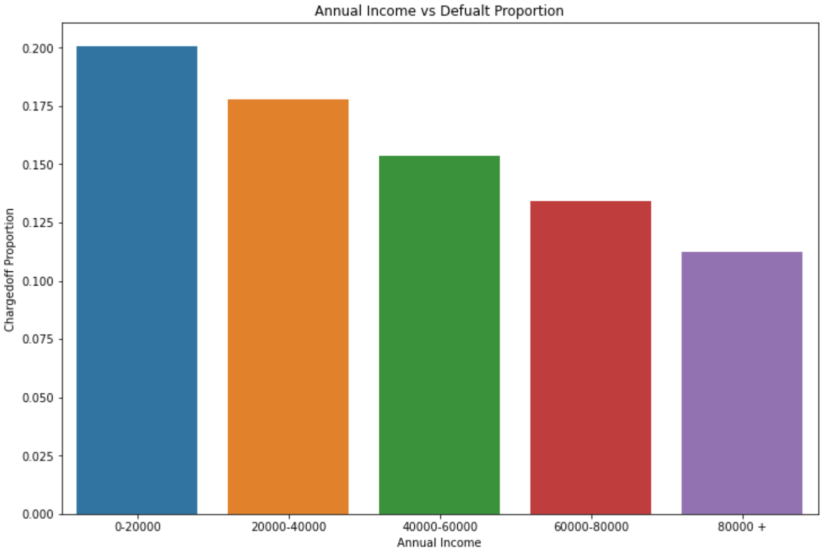
- **Term vs Loan Status**



Observations:

- % defualter borrowers in 60 months term are more then 36 months.
- Fully Paid rate is higher in 36 months tenure.

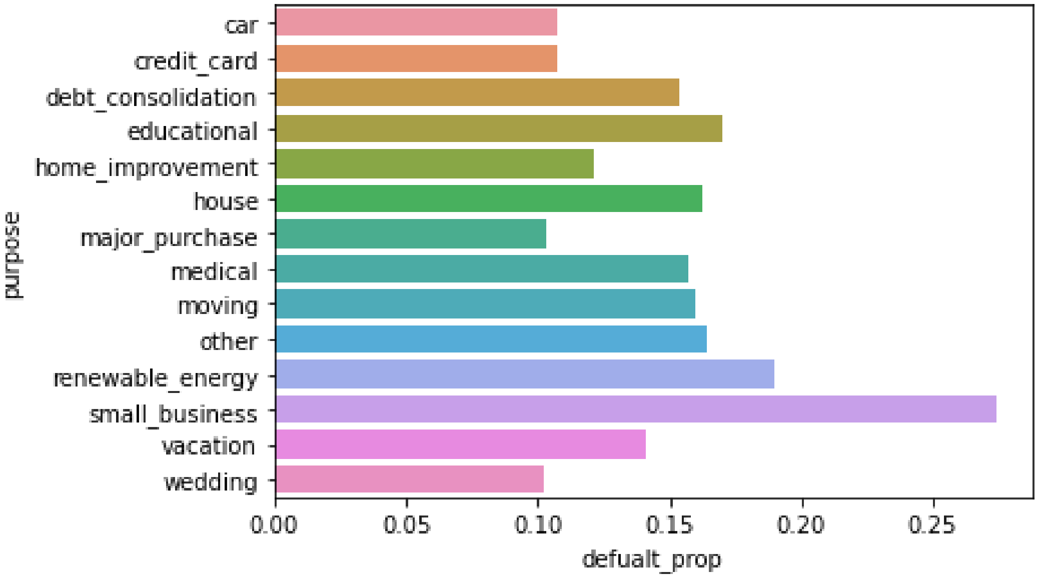
- **Bivariate analysis between annual income category and default %**



Observation

- As you can see higher annual income people have lesser defaults

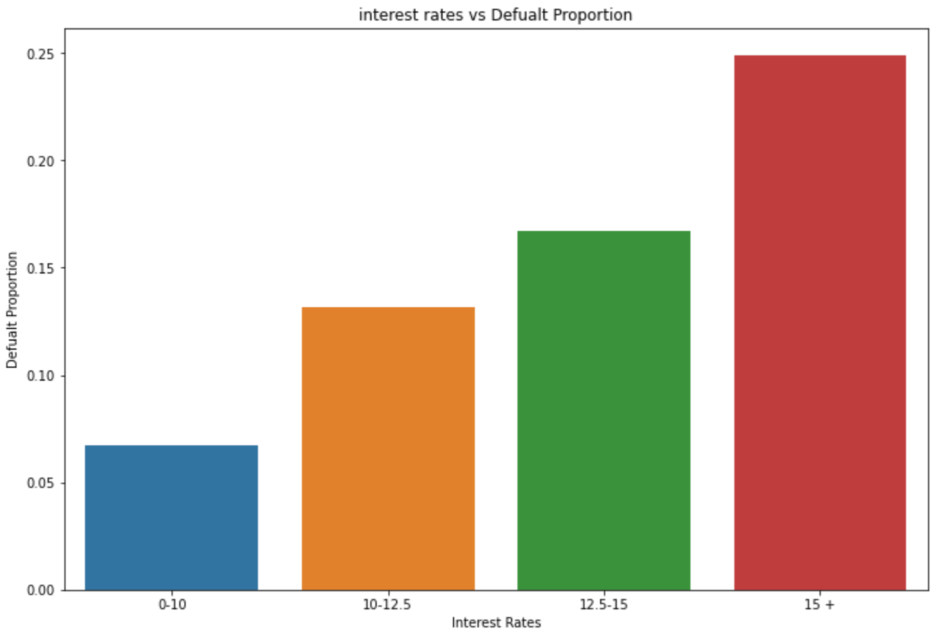
- **Bivariate analysis between purpose category and default proportion**



Observation

- small business has higher chances to defult the loan and next is renewable energy

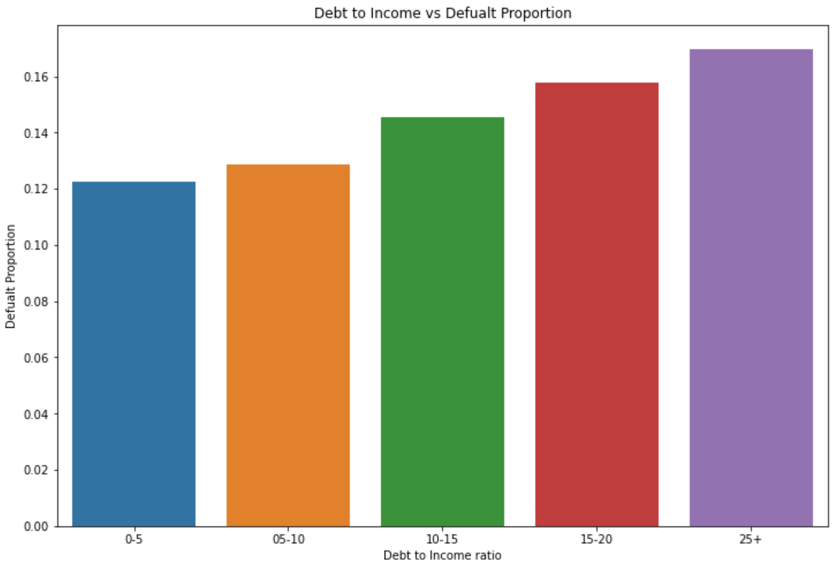
- **Bivariate analysis between interest rate category and default proportion**



Observation

- Higher Interest rates has higher chances to defult the loan

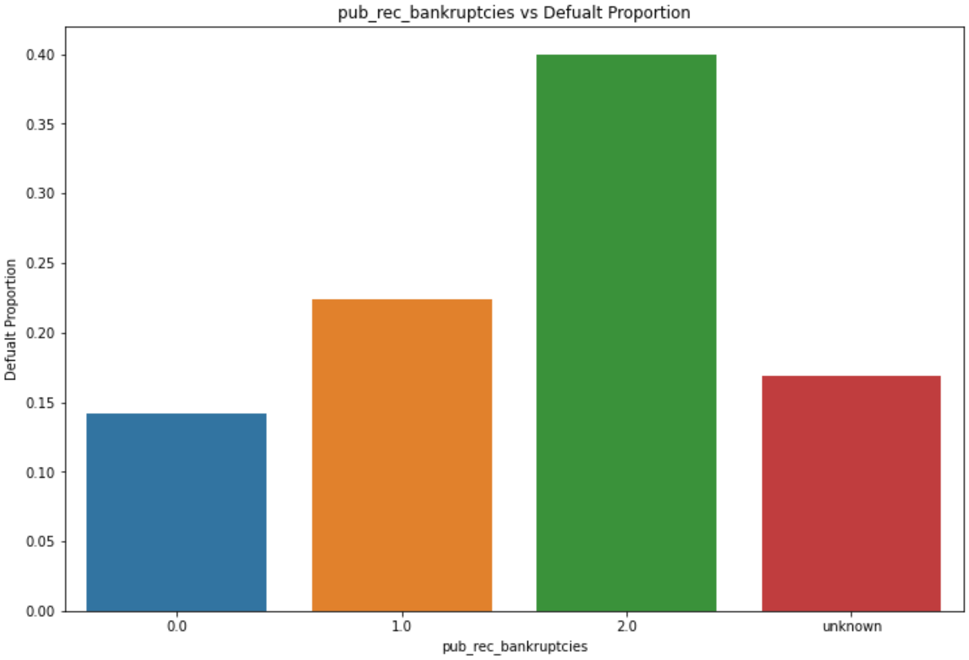
- **Bivariate analysis between dti category and default proportion**



Observation

- Higher DTI (Debt to Income) has higher chances to default

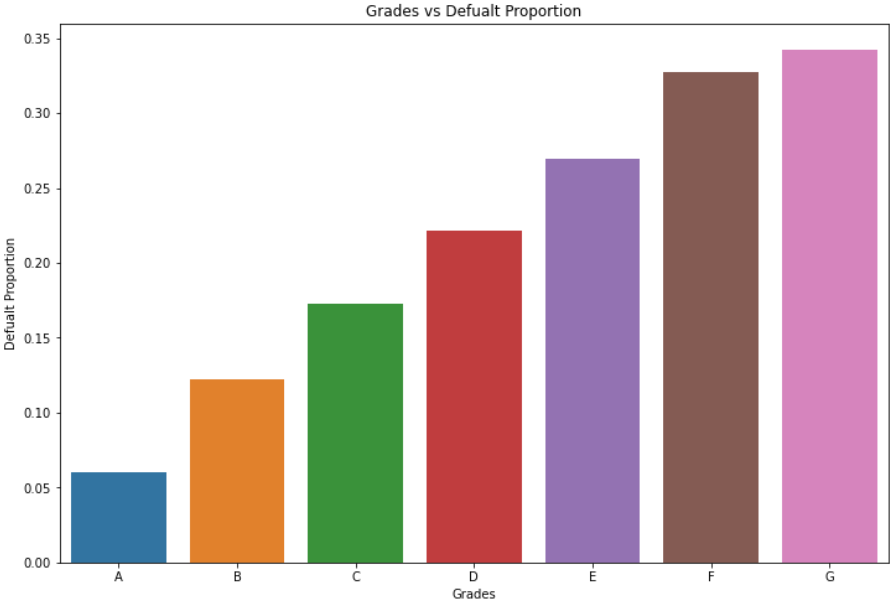
- **Bivariate analysis between public record bank corrupt category and default proportion**



Observation

- Higher the number of public record bankruptcies, higher the chances of default.

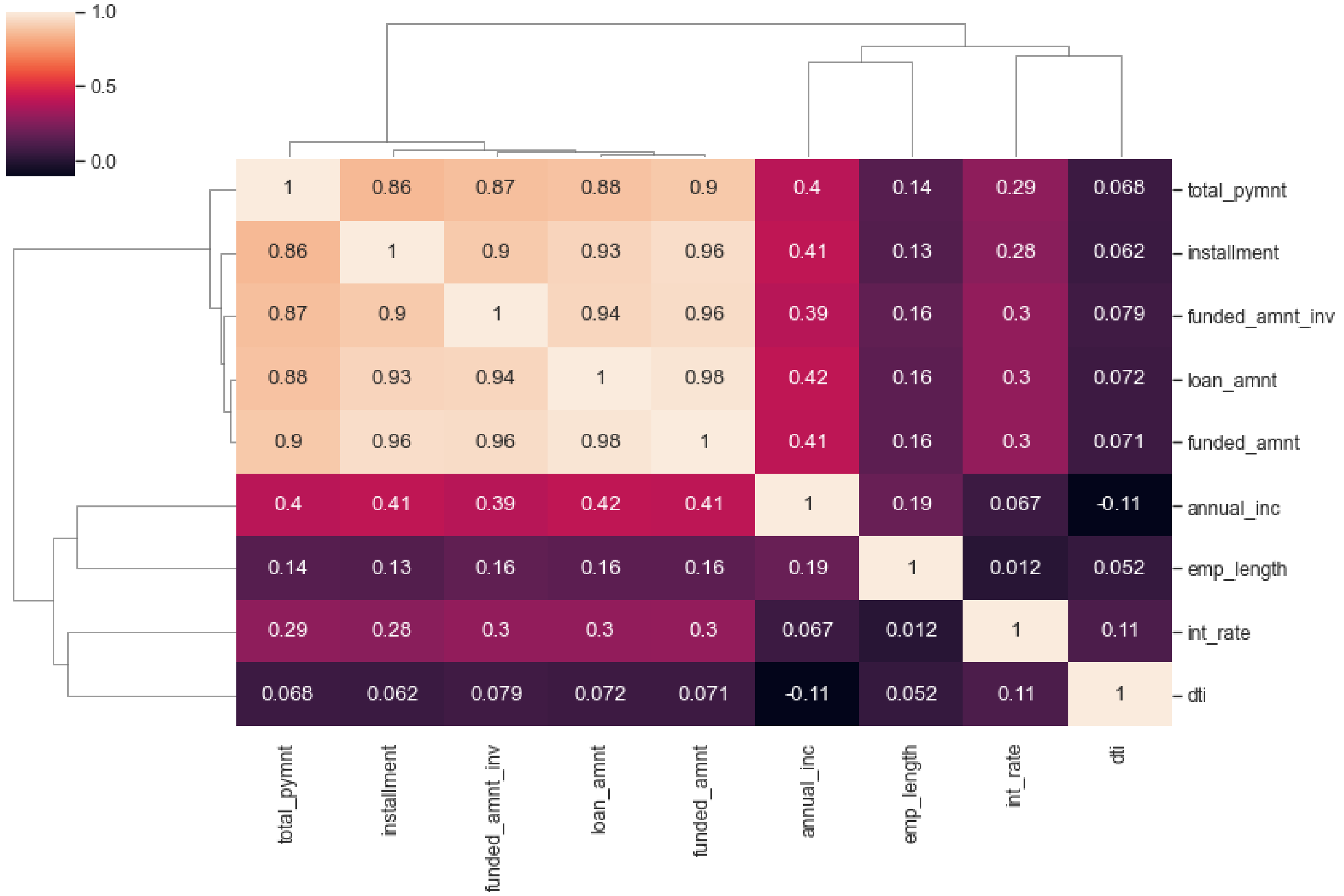
- **Bivariate analysis between grade category and default proportion**



Observation

- Lower the grade, higher the chances of default

Overall correlation between important numeric feature columns



	total_pymnt	installment	funded_amnt_inv	loan_amnt	funded_amnt	annual_inc	emp_length	int_rate	dti
total_pymnt	1	0.86	0.87	0.88	0.9	0.4	0.14	0.29	0.068
installment	0.86	1	0.9	0.93	0.96	0.41	0.13	0.28	0.062
funded_amnt_inv	0.87	0.9	1	0.94	0.96	0.39	0.16	0.3	0.079
loan_amnt	0.88	0.93	0.94	1	0.98	0.42	0.16	0.3	0.072
funded_amnt	0.9	0.96	0.96	0.98	1	0.41	0.16	0.3	0.071
annual_inc	0.4	0.41	0.39	0.42	0.41	1	0.19	0.067	-0.11
emp_length	0.14	0.13	0.16	0.16	0.16	0.19	1	0.012	0.052
int_rate	0.29	0.28	0.3	0.3	0.3	0.067	0.012	1	0.11
dti	0.068	0.062	0.079	0.072	0.071	-0.11	0.052	0.11	1

Observation

- Loan amount, investor amount, funding amount are strongly correlated.
- emp_length has less correlation with loan amount which is surprsing either it should be less or high
- DTI has negative correlation with annual income.
- annual income and loan amount are positive correlated i.e. Higher annual income has higher loan amount approved.
- emp_lenth and annual income are not strongly correlated which is surprising