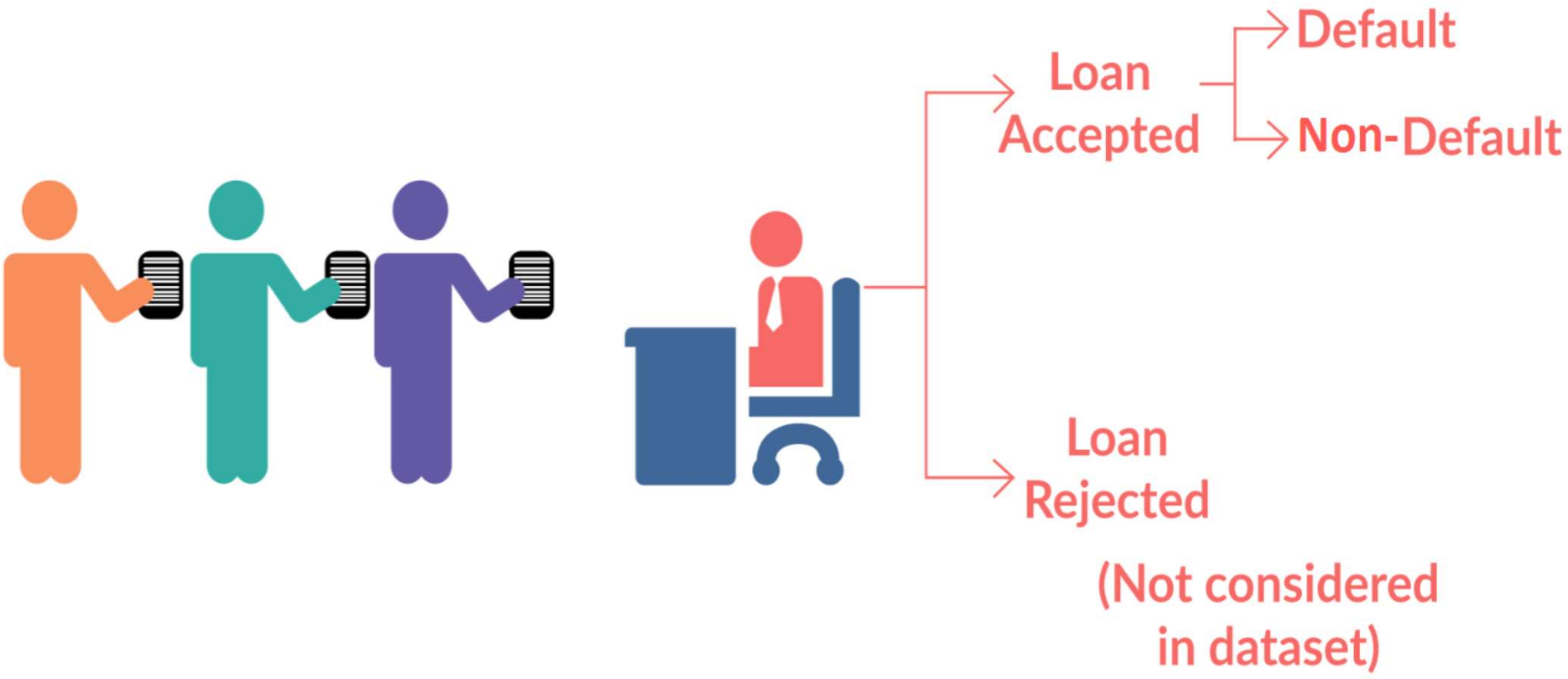
Lending Club Case Study

By

Anjan Kumar Pal

LOAN DATASET



Loan Accepted

Default

Non-Default

Loan Rejected

(Not considered in dataset)

Loan Approval Prerequisite

- 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:
    1. **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
    2. **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'.
    3. **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)

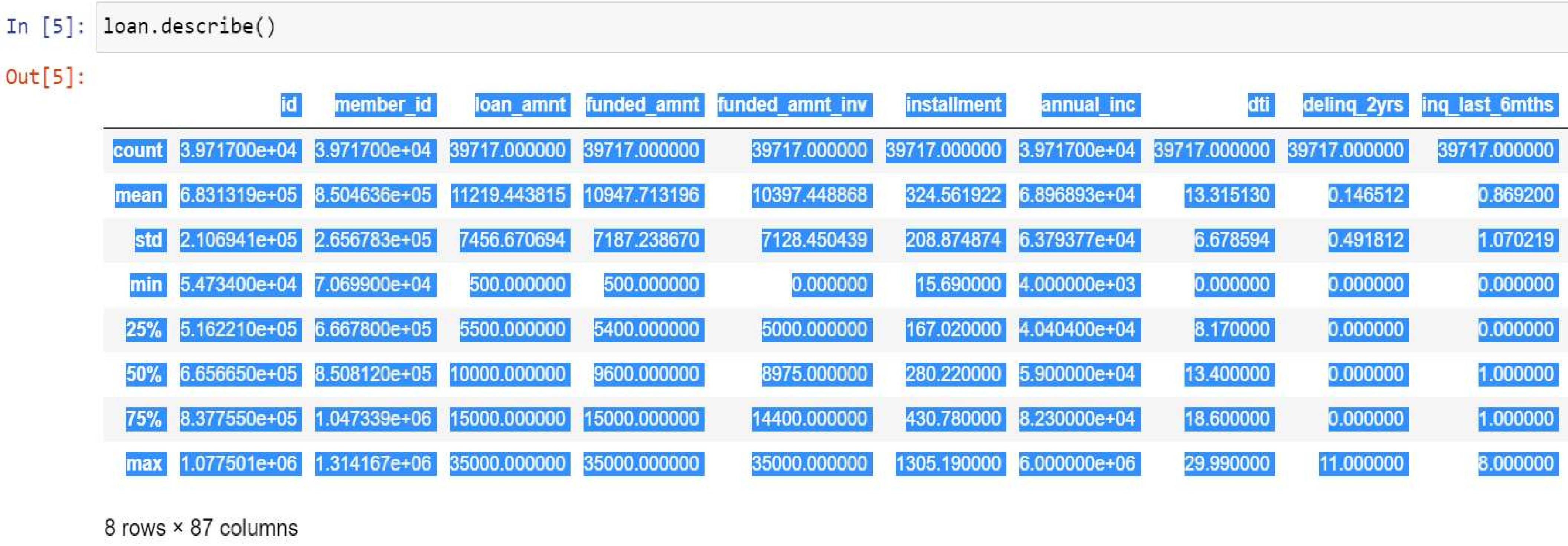
Business Objectives:

- This company is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.
-
- Like most other lending companies, 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'.
-
- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicant using EDA is the aim of this case study.
-
- In other words, 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.
-

  To develop your understanding of the domain, you are advised to independently research a little about risk analytics (understanding the types of variables and their significance should be enough).

Data Dictionary Understanding:

- Let us understand the dataset
- loan.shape gives us the number of records & columns(39717,111)
- Loan.describe() will give us a description as follows



```
In [5]: loan.describe()
```

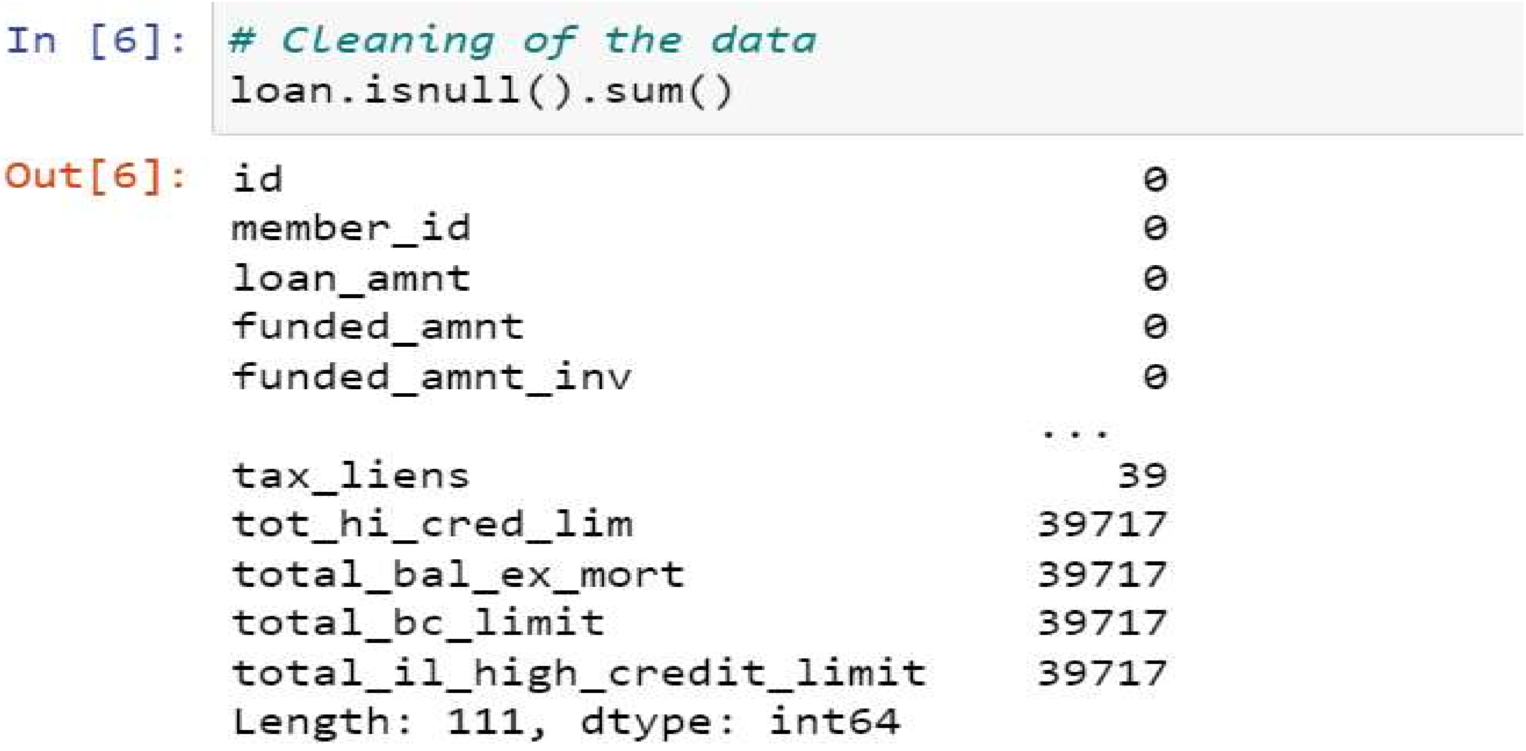
Out[5]:

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths
count	3.971700e+04	3.971700e+04	39717.000000	39717.000000	39717.000000	39717.000000	3.971700e+04	39717.000000	39717.000000	39717.000000
mean	6.831319e+05	8.504636e+05	11219.443815	10947.713196	10397.448868	324.561922	6.896893e+04	13.315130	0.146512	0.869200
std	2.106941e+05	2.656783e+05	7456.670694	7187.238670	7128.450439	208.874874	6.379377e+04	6.678594	0.491812	1.070219
min	5.473400e+04	7.069900e+04	500.000000	500.000000	0.000000	15.690000	4.000000e+03	0.000000	0.000000	0.000000
25%	5.162210e+05	6.667800e+05	5500.000000	5400.000000	5000.000000	167.020000	4.040400e+04	8.170000	0.000000	0.000000
50%	6.656650e+05	8.508120e+05	10000.000000	9600.000000	8975.000000	280.220000	5.900000e+04	13.400000	0.000000	1.000000
75%	8.377550e+05	1.047339e+06	15000.000000	15000.000000	14400.000000	430.780000	8.230000e+04	18.600000	0.000000	1.000000
max	1.077501e+06	1.314167e+06	35000.000000	35000.000000	35000.000000	1305.190000	6.000000e+06	29.990000	11.000000	8.000000

8 rows × 87 columns

Understanding Missing Data & Cleaning:

- Let us understand the missing columns
- loan.isnull().sum() gives us an information as follows:

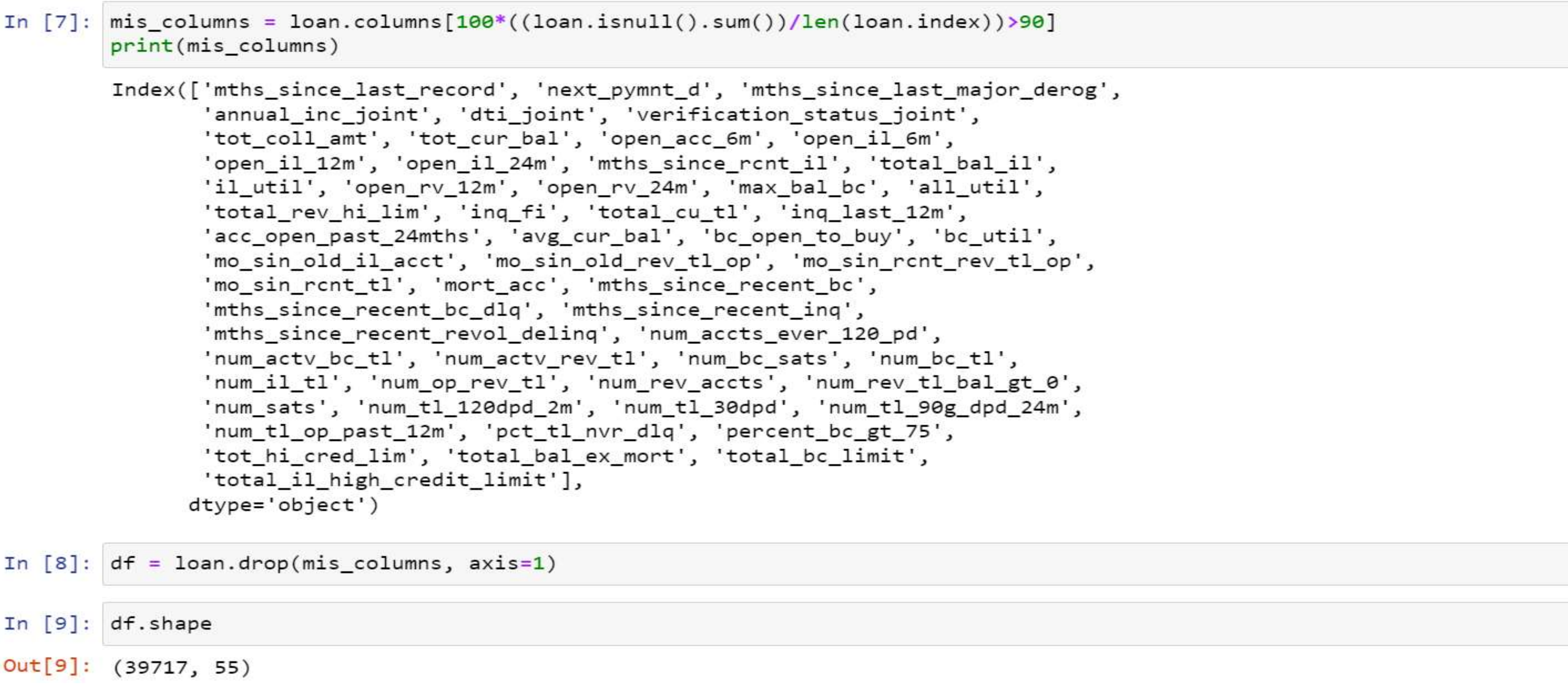


```
In [6]:  # Cleaning of the data
          loan.isnull().sum()
```

```
Out[6]:  id                                0
         member_id                         0
         loan_amnt                         0
         funded_amnt                       0
         funded_amnt_inv                   0
                                        ...
         tax_liens                        39
         tot_hi_cred_lim               39717
         total_bal_ex_mort             39717
         total_bc_limit                39717
         total_il_high_credit_limit    39717
         Length: 111, dtype: int64
```

Understanding Missing Data & Cleaning…Contd:

- Let us filter out the missing data columns which have more than 90% of missing data
- Drop those missing data columns from the loan dataset. This gives rise to a dataset of 55 columns



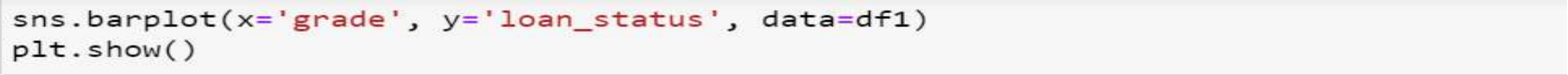
```
In [7]:  mis_columns = loan.columns[100*((loan.isnull().sum())/len(loan.index))>90]
         print(mis_columns)

         Index(['mths_since_last_record', 'next_pymnt_d', 'mths_since_last_major_derog',
                'annual_inc_joint', 'dti_joint', 'verification_status_joint',
                'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m',
                'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il',
                'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util',
                'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m',
                'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util',
                'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op',
                'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc',
                'mths_since_recent_bc_dlq', 'mths_since_recent_inq',
                'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd',
                'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl',
                'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0',
                'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m',
                'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75',
                'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit',
                'total_il_high_credit_limit'],
               dtype='object')

In [8]:  df = loan.drop(mis_columns, axis=1)

In [9]:  df.shape

Out[9]:  (39717, 55)
```

Understanding Missing Data & Cleaning…Contd:

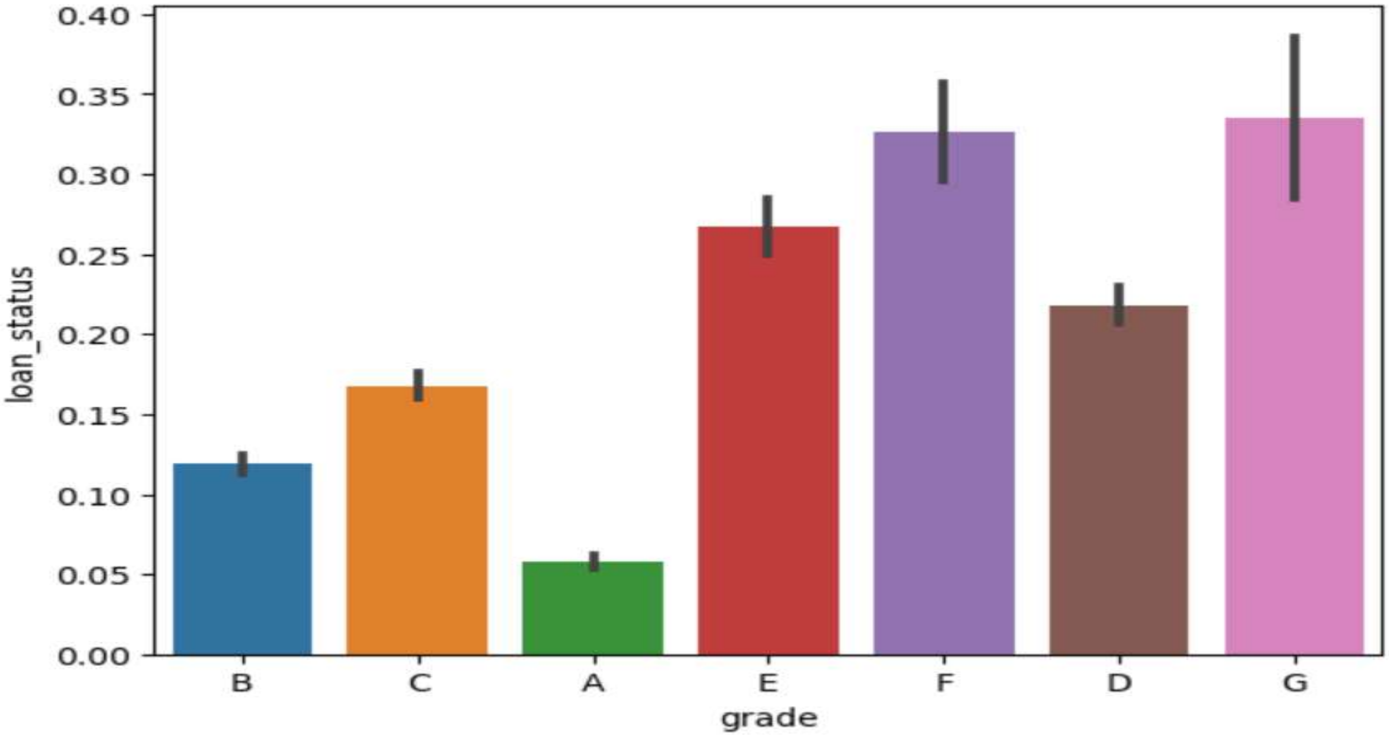
- Let us drop ‘desc’ column
- Drop the rows having missing data values more than 5
- Convert columns which are supposed to be numeric, however they are defined as objects in the dataset
- These are ‘int_rate’ and ‘emp_length’; These are done as below
- df['int_rate'] = df['int_rate'].apply(lambda x:pd.to_numeric(x.split('%')[0]))
- df['emp_length']=df['emp_length'].apply(lambda x: re.findall('\d+', str(x))[0])
- df['emp_length'] = df['emp_length'].apply(lambda x:pd.to_numeric(x))

Univariate & Bivariate Analysis of Dataset:

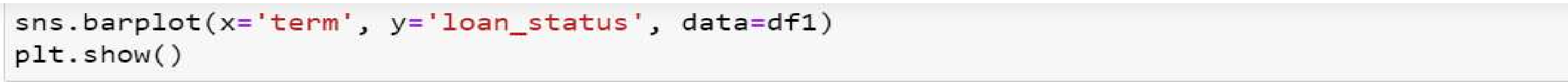
- Now having cleaned the dataset we want to have some univariate and bivariate analysis on the loan dataset
- We will have some distribution analysis with respect to different parameters and show the bar graph with those analysis



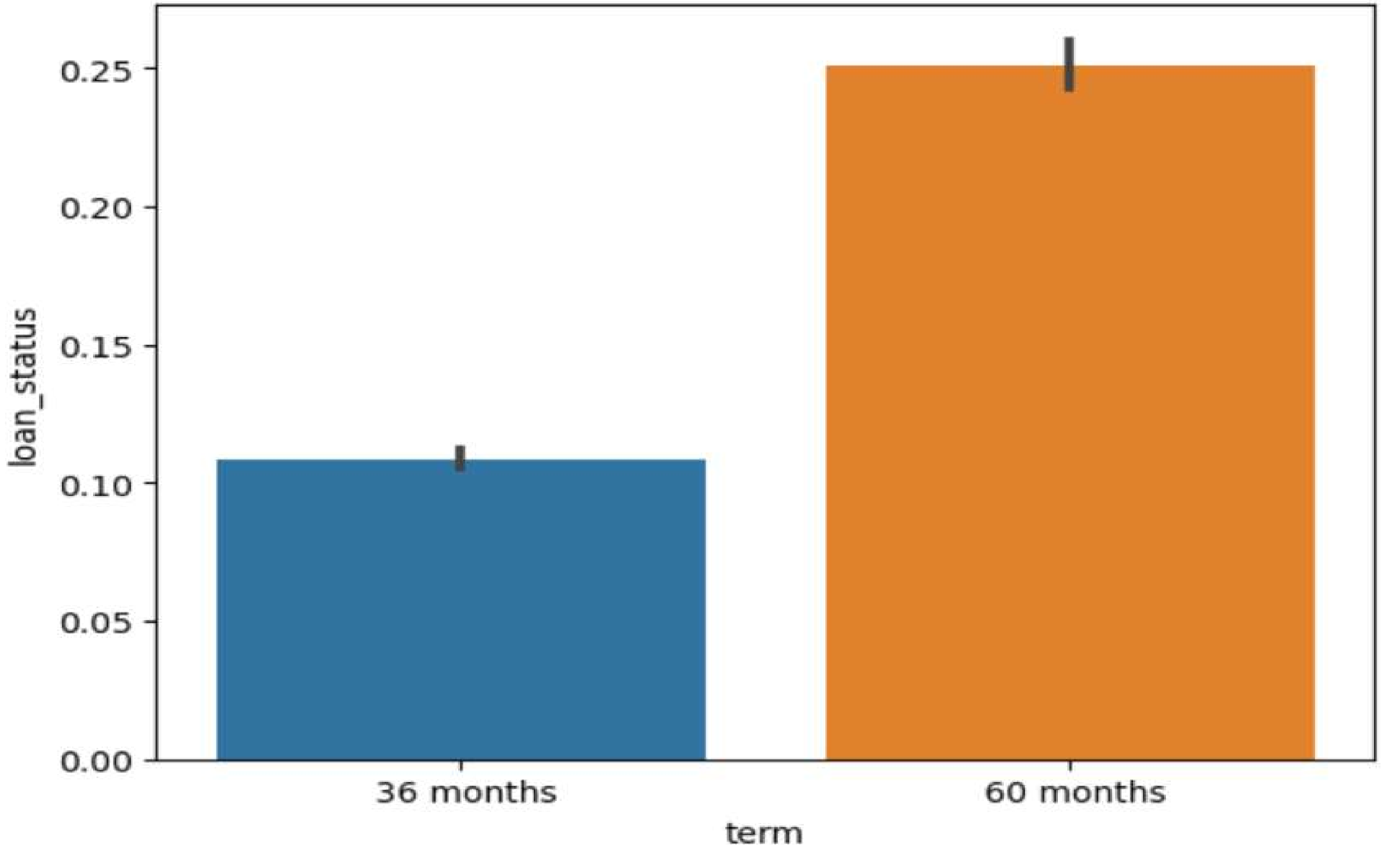
```
sns.barplot(x='grade', y='loan_status', data=df1)
plt.show()
```



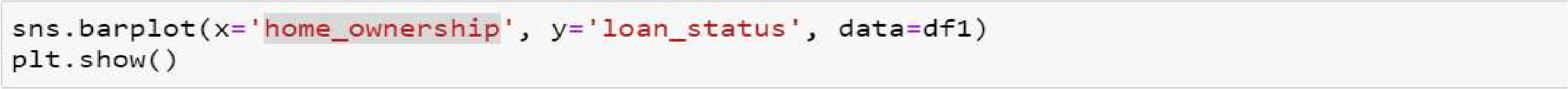
Univariate & Bivariate Analysis of Dataset:



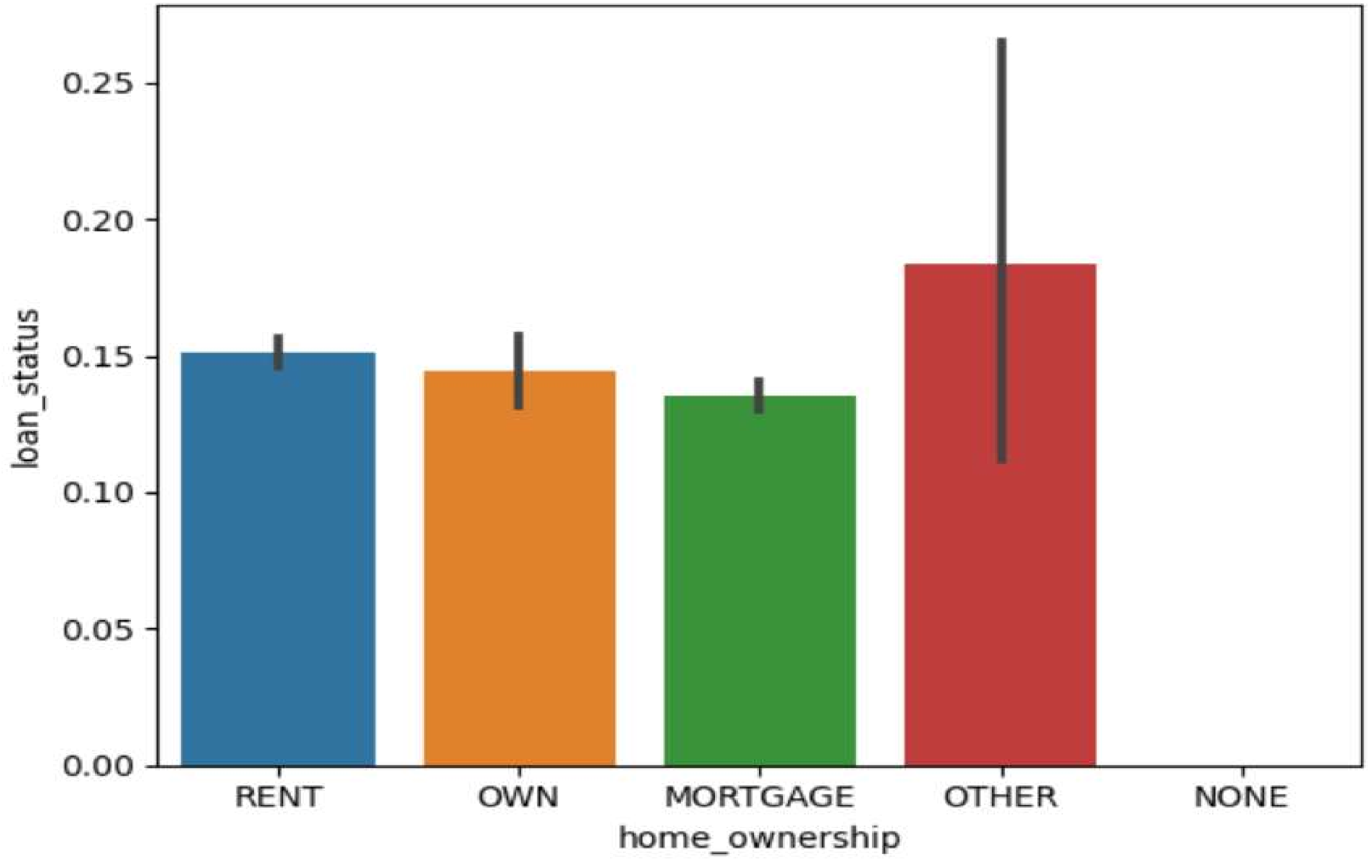
```
sns.barplot(x='term', y='loan_status', data=df1)
plt.show()
```



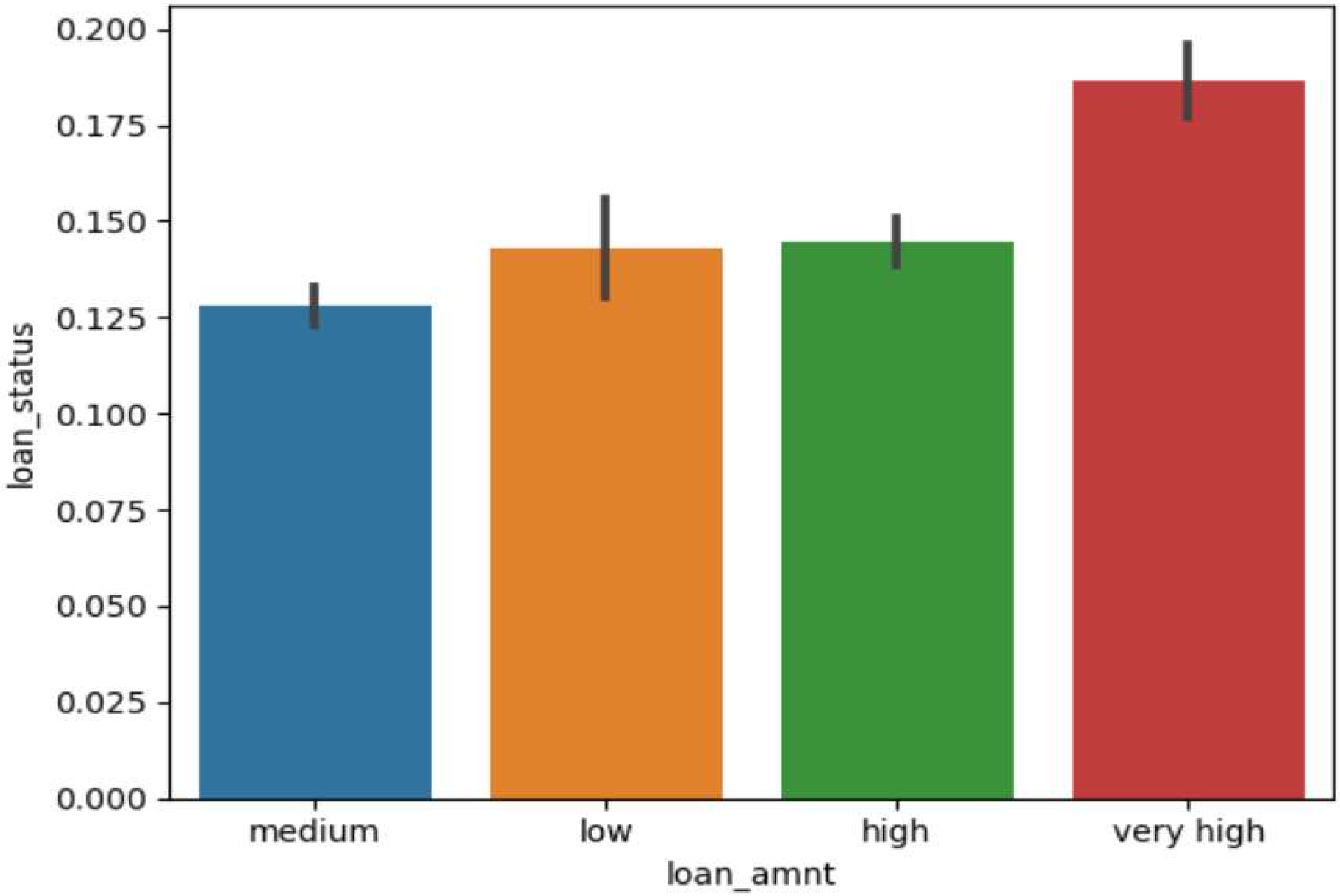
Univariate & Bivariate Analysis of Dataset:



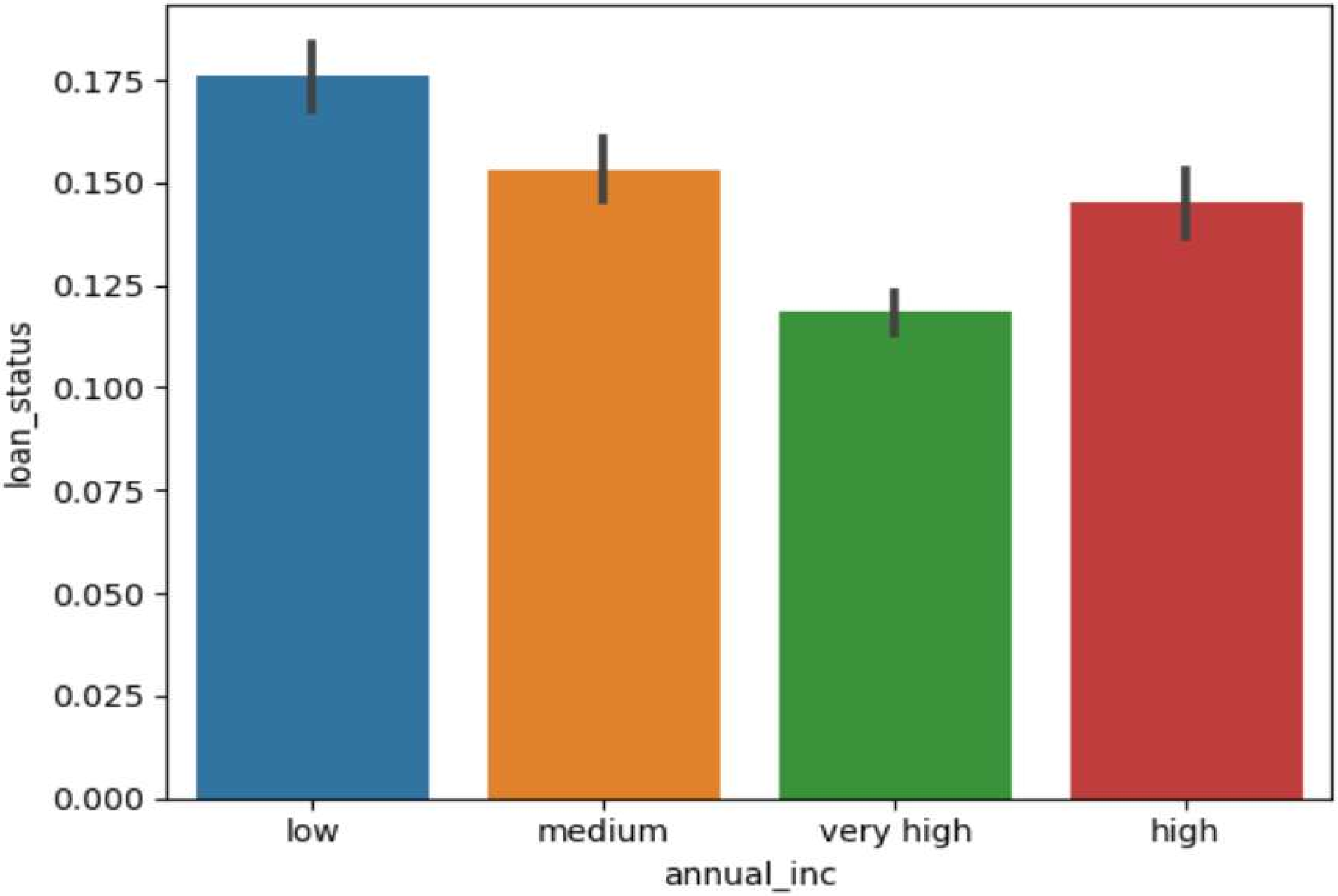
```
sns.barplot(x='home_ownership', y='loan_status', data=df1)
plt.show()
```



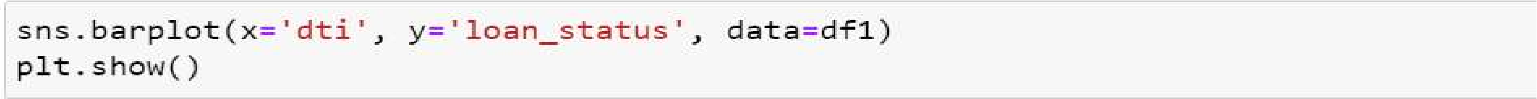
Univariate & Bivariate Analysis of Dataset:



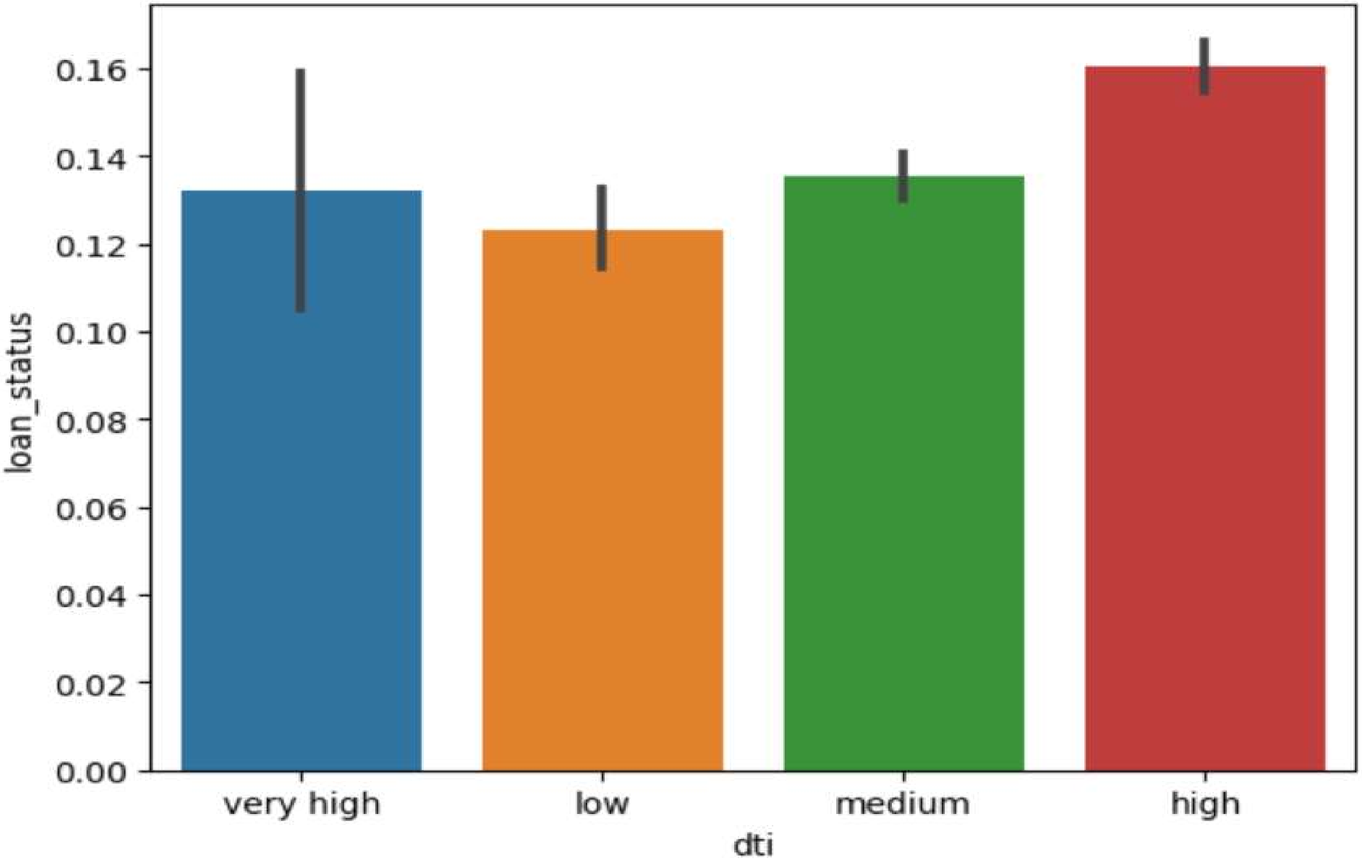
Univariate & Bivariate Analysis of Dataset:



Univariate & Bivariate Analysis of Dataset:



```
sns.barplot(x='dti', y='loan_status', data=df1)
plt.show()
```



Conclusions and recommendations:

- Lending club may reduce high rate for 60 months tenure, as those are susceptible to default.
- It should check more data based on borrower grades (G to A) as it is a good metric for finding out defaulters.
- Lending club should stop giving loans to borrowers with mortgage home ownership when loan amount requested is more than 12000 as they are likely to default approved loans after taking higher amount of loans.
- Lending club should DTI records before giving loan to borrowers.