Group Facilitator Name

Goutam Debnath

Problem Statement:

You work for a consumer finance company which specializes 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:

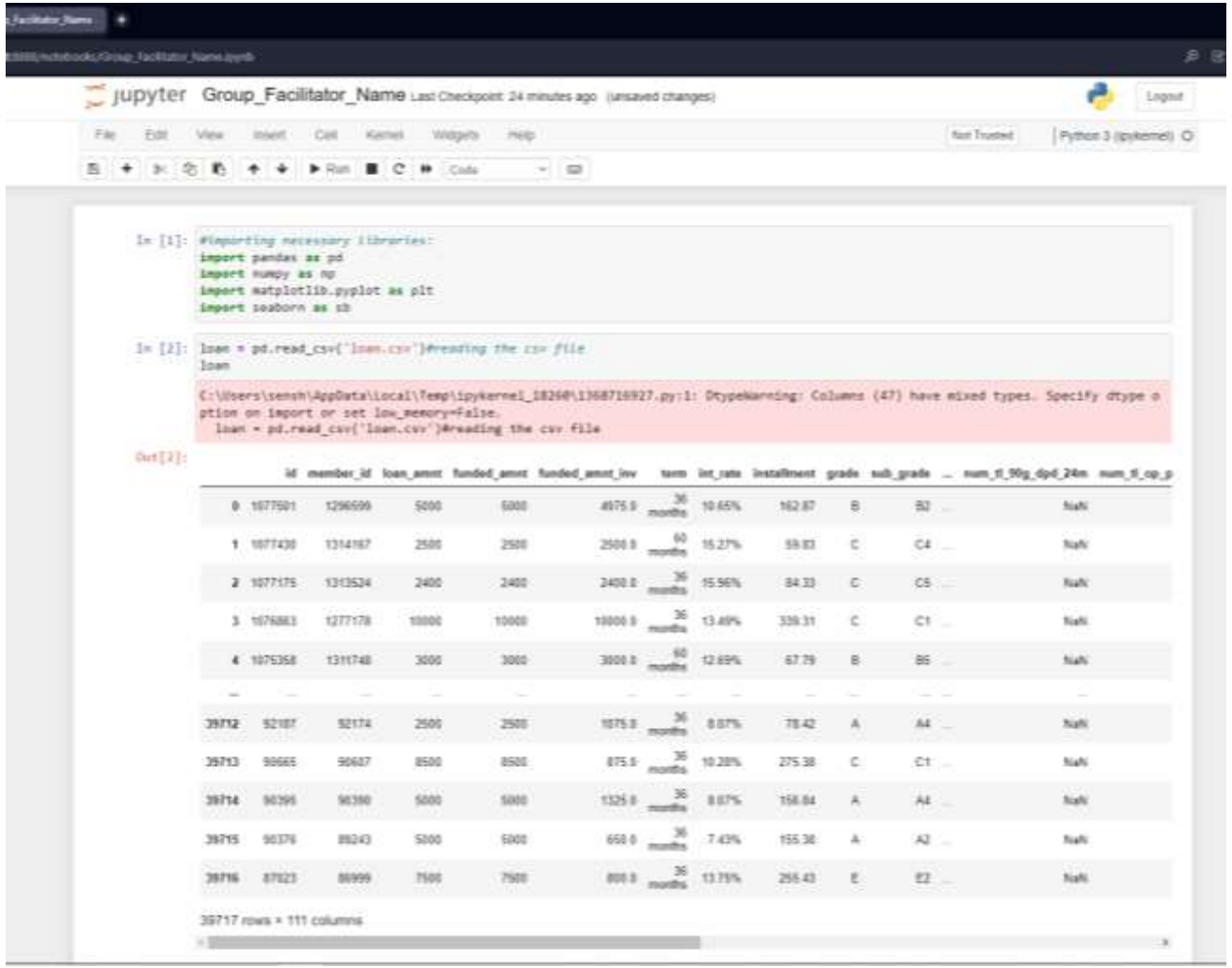
- If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company

- 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 data given below contains information about past loan applicants and whether they 'defaulted' or not. 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.

Loading the dataset:

Imported the necessary libraries and read the dataset from a .csv file.



Cleaning the dataset:

- Checked for duplicate rows
- Removed columns having only null values.
- Removed columns with majority null values
- Removed columns unnecessary for analysis

jupyter Group_Facilitator_Name Last Checkpoint: 33 minutes ago (autosaved)

Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

```
4        1075358
          ...
39712      92187
39713      90665
39714      90395
39715      90376
39716      87023
Name: id, Length: 39717, dtype: int64>
```

In [4]:
```
loan.columns[loan.isnull().sum()==39717] #identifying null columns
```

Out[4]:
```
Index(['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 [5]:
```
#removing the null columns.
loan.drop(columns=['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'], inplace=True)
```

In [6]:
```
loan.info()
```

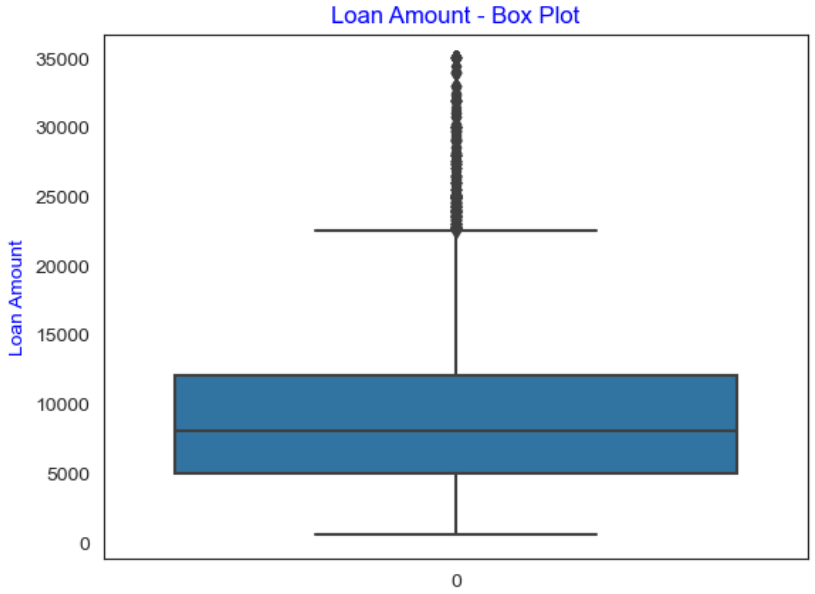
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Data columns (total 57 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   id                          39717 non-null  int64
 1   member_id                   39717 non-null  int64
 2   loan_amnt                   39717 non-null  int64
 3   funded_amnt                 39717 non-null  int64
 4   funded_amnt_inv             39717 non-null  float64
 5   term                        39717 non-null  object
 6   int_rate                    39717 non-null  object
 7   installment                 39717 non-null  float64
 8   grade                       39717 non-null  object
 9   sub_grade                   39717 non-null  object
 10  emp_title                   37258 non-null  object
 11  emp_length                  38642 non-null  object
 12  home_ownership              39717 non-null  object
```

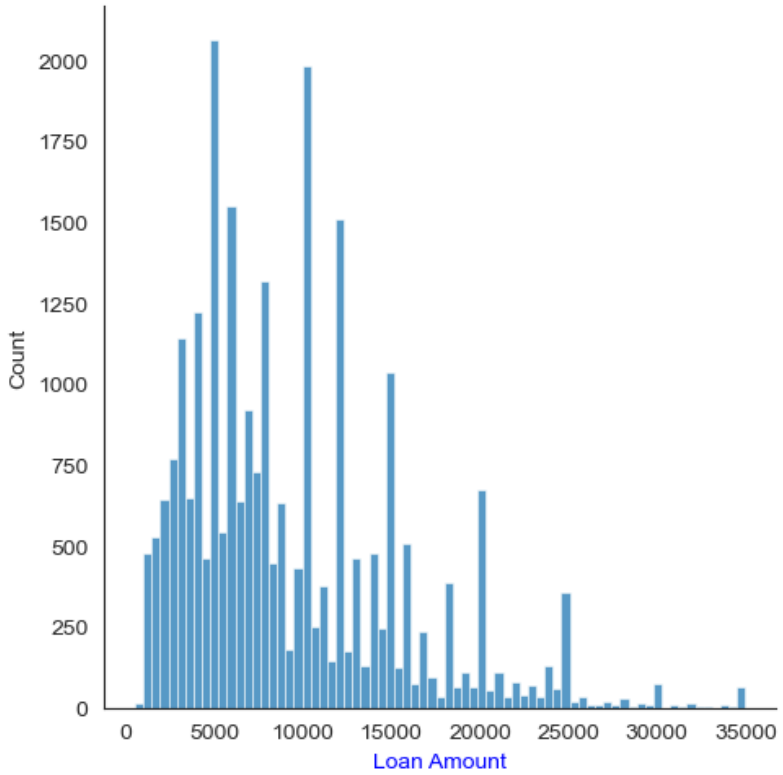
Analysis:

Univariate Analysis

Analyzing Loan Amount

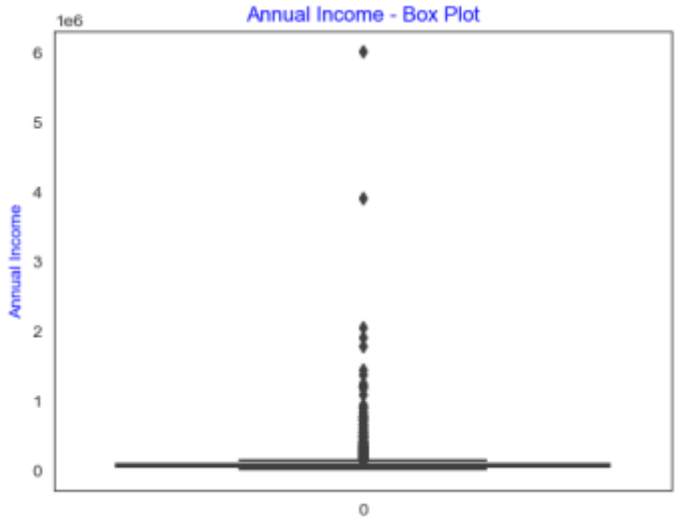
- Plotted the loan amount column in boxplot and distribution plot
- From the plot, it is evident that majority of the loans are in the range $4,500 to $15,000
- Majority of the clients have taken a loan of approx. $8,000

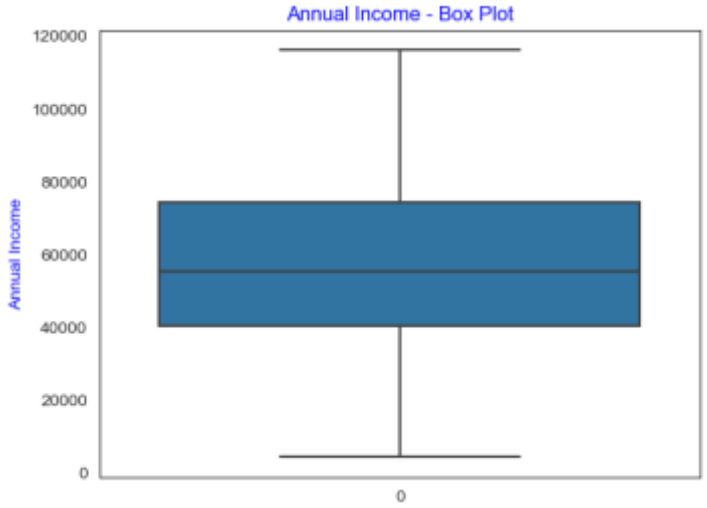


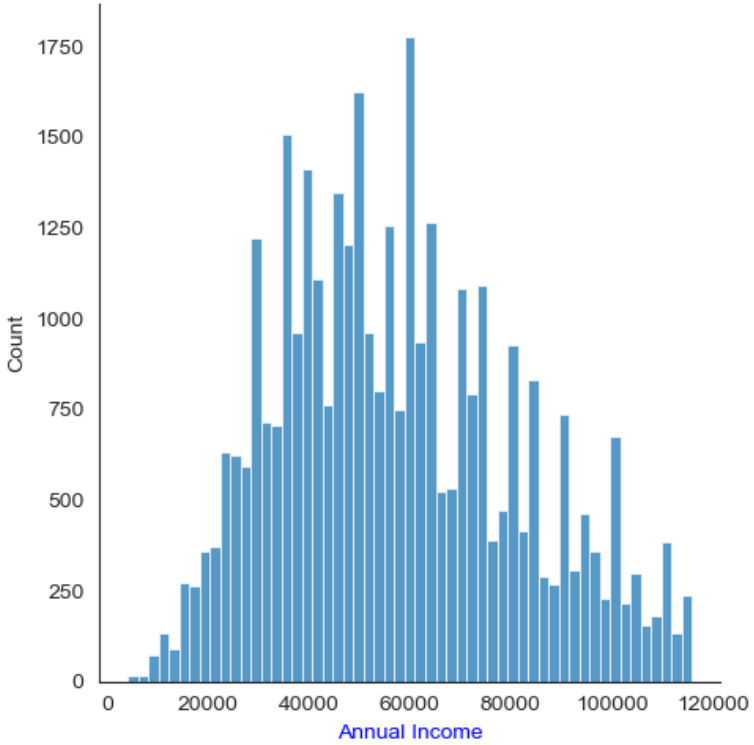


Analyzing Annual Income

- Plotted the annual income column in boxplot and distribution plot
- The annual income contained some outliers which needed to be removed for clearer analysis
- After removing the outlier, we notice a readable plot.
- From the plot, it is evident that most of the borrower's annual income falls in the range of $4,000 to $8,000
- Around half of the borrower's annual income is approx. $5,900

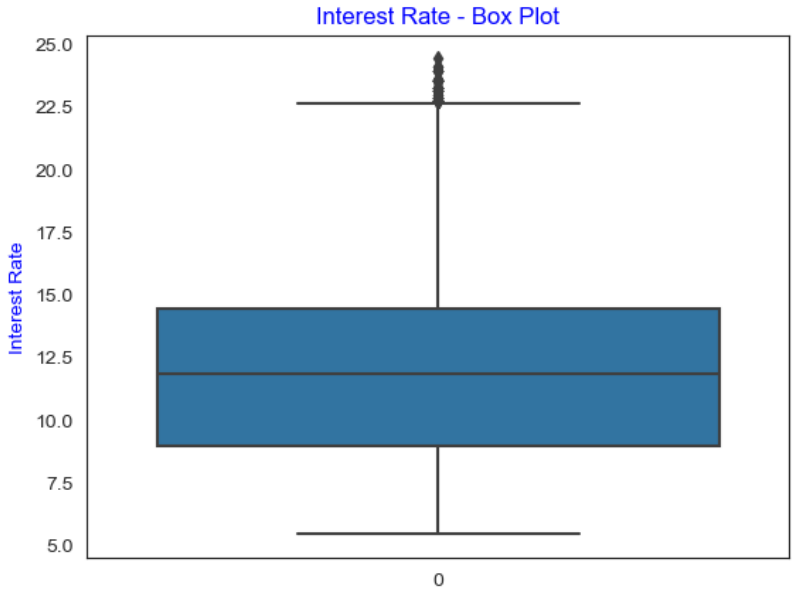


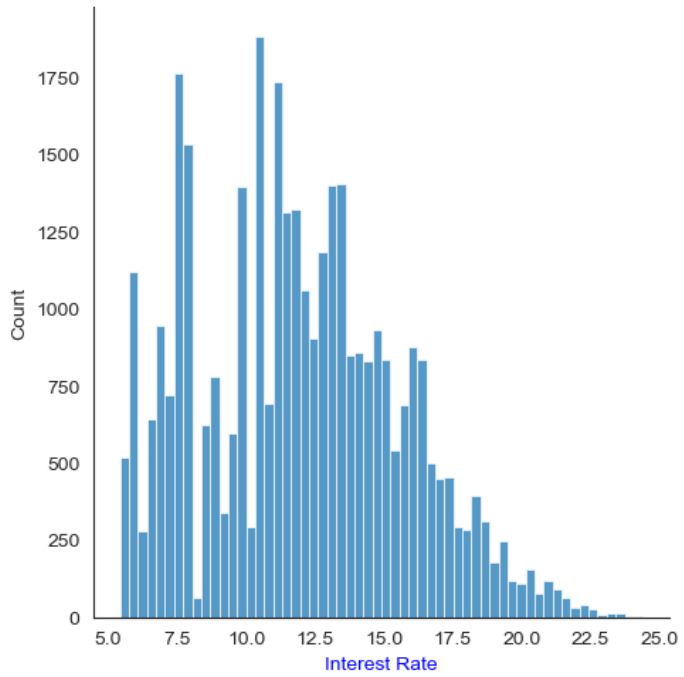




Analyzing Interest Rate

- Plotted the annual income column in boxplot and distribution plot
- From the plot, it is evident that most of the interest on the loans are between 10% to 15%
- Around half of the loan's interest is approx. 11.83%

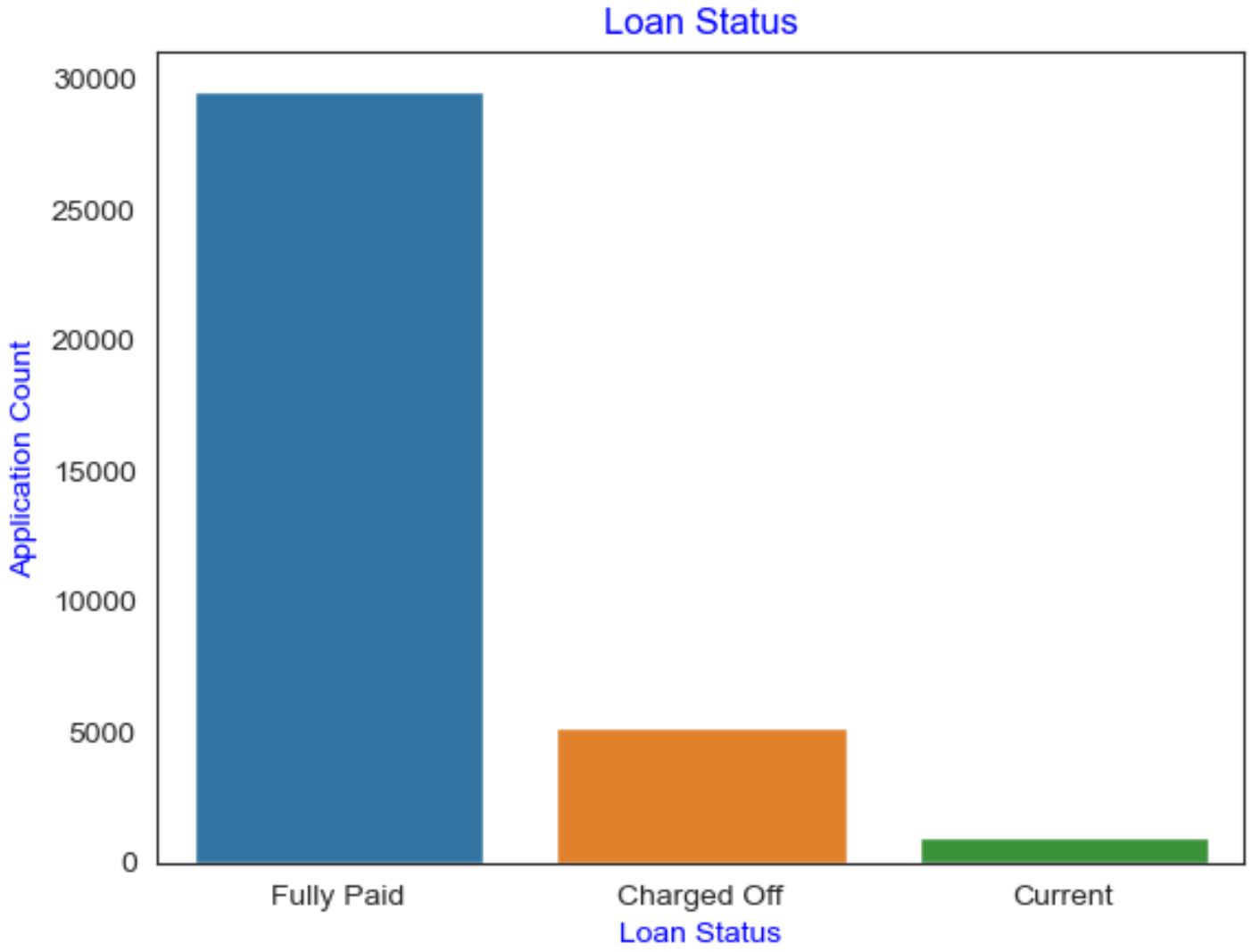




Analyzing Loan Status

From the plot of loan status we concur that:

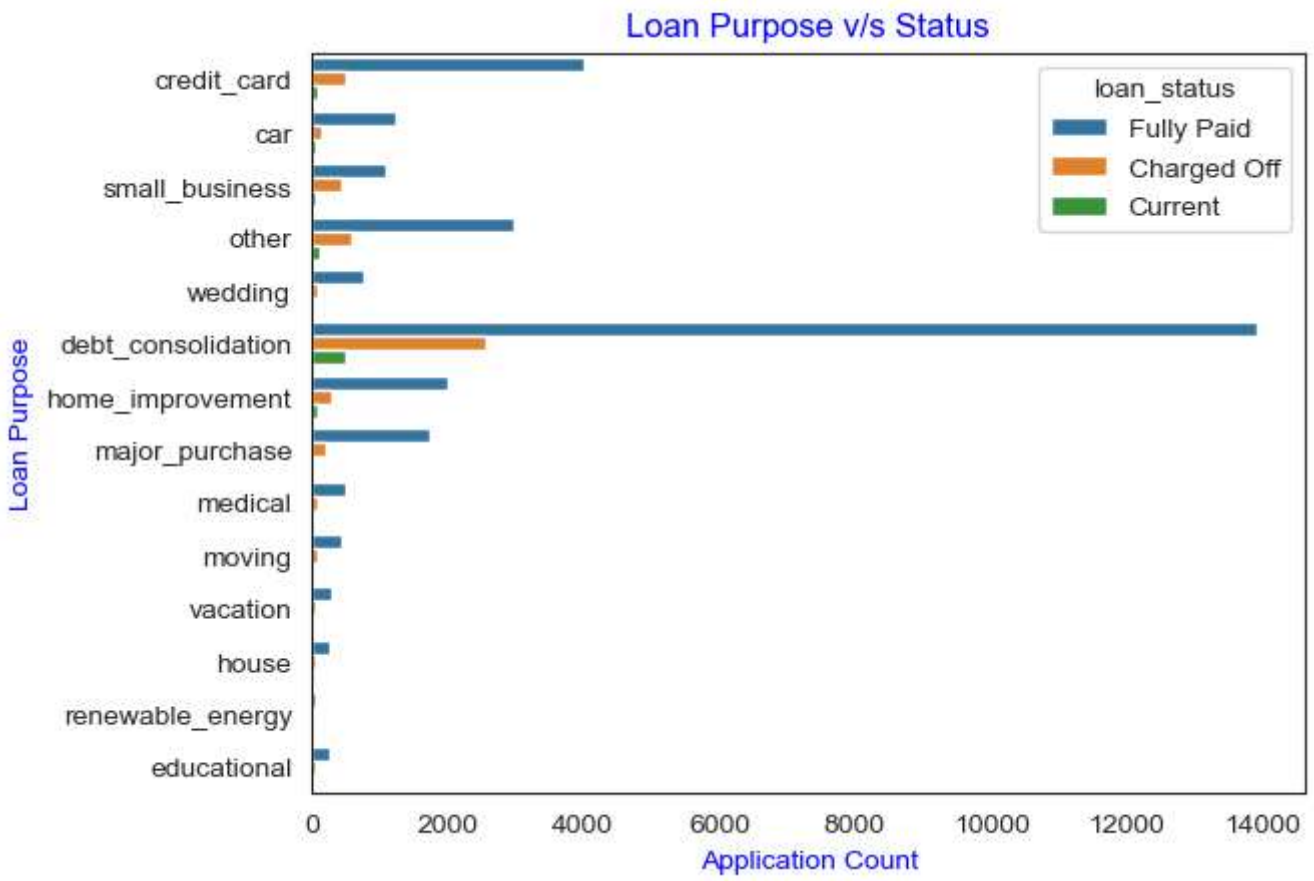
- 82.67% of loans have been fully paid
- 14.55% of loans have been charged off
- 2.78% of loans are currently active



Analyzing Loan Purpose

Plotting the loan purposes shows the following:

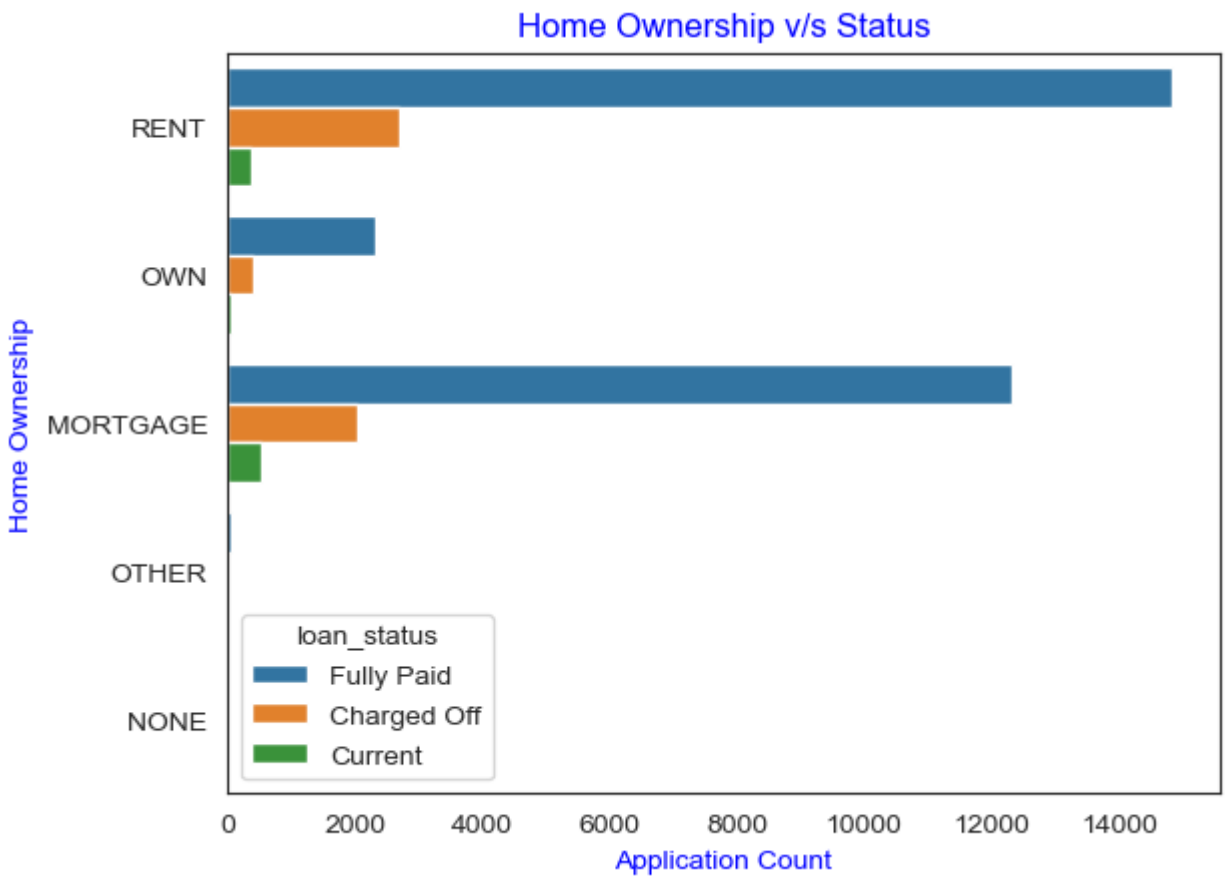
- Debt consolidation is the most common purpose for loans
- Debt consolidation has the most charged off loans



Analyzing Home Ownership

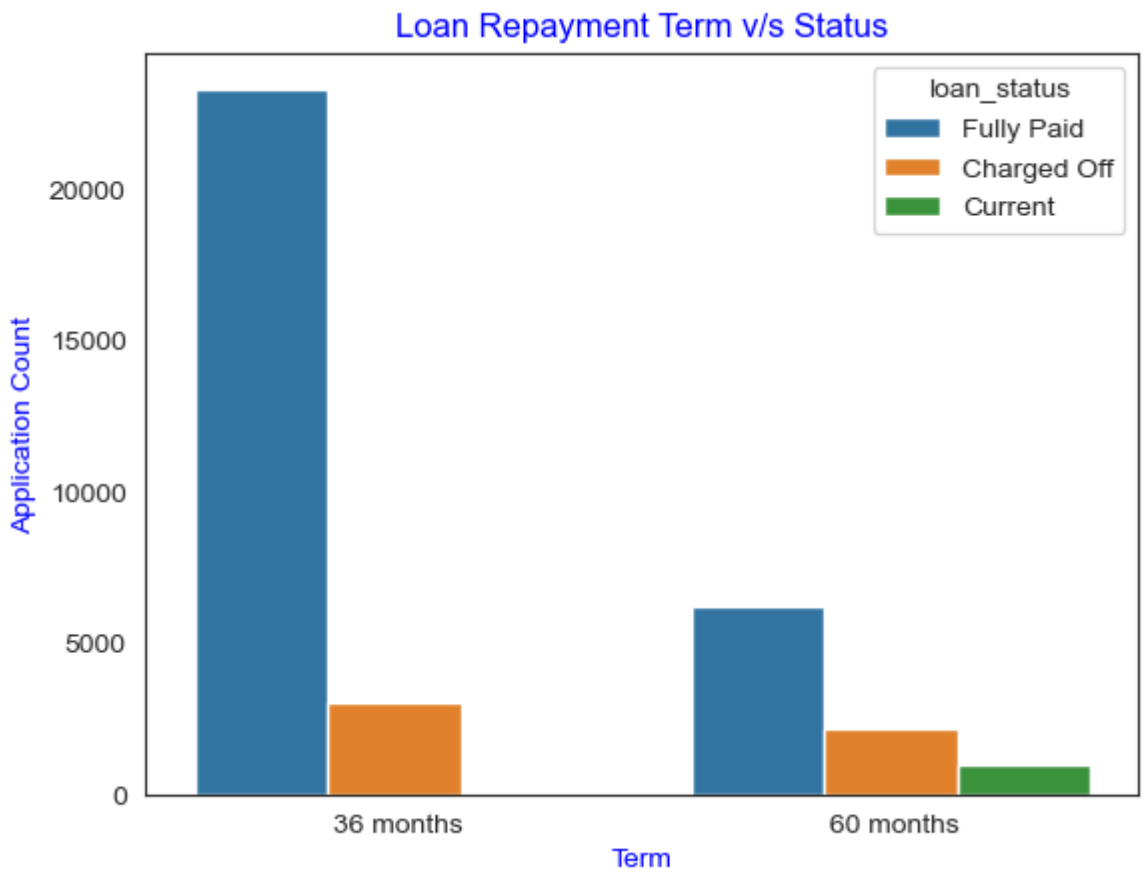
Plotting a graph for the home owners shows the following:

- Clients in rented houses have the most borrowings followed by mortgage
- The above two also have the most amount of charged off loans



Analyzing Loan Repayment Term

- Plotting the loan repayment term shows that the term of 36months is the most sought after by borrowers
- The 36months have the most amount of charged off loans



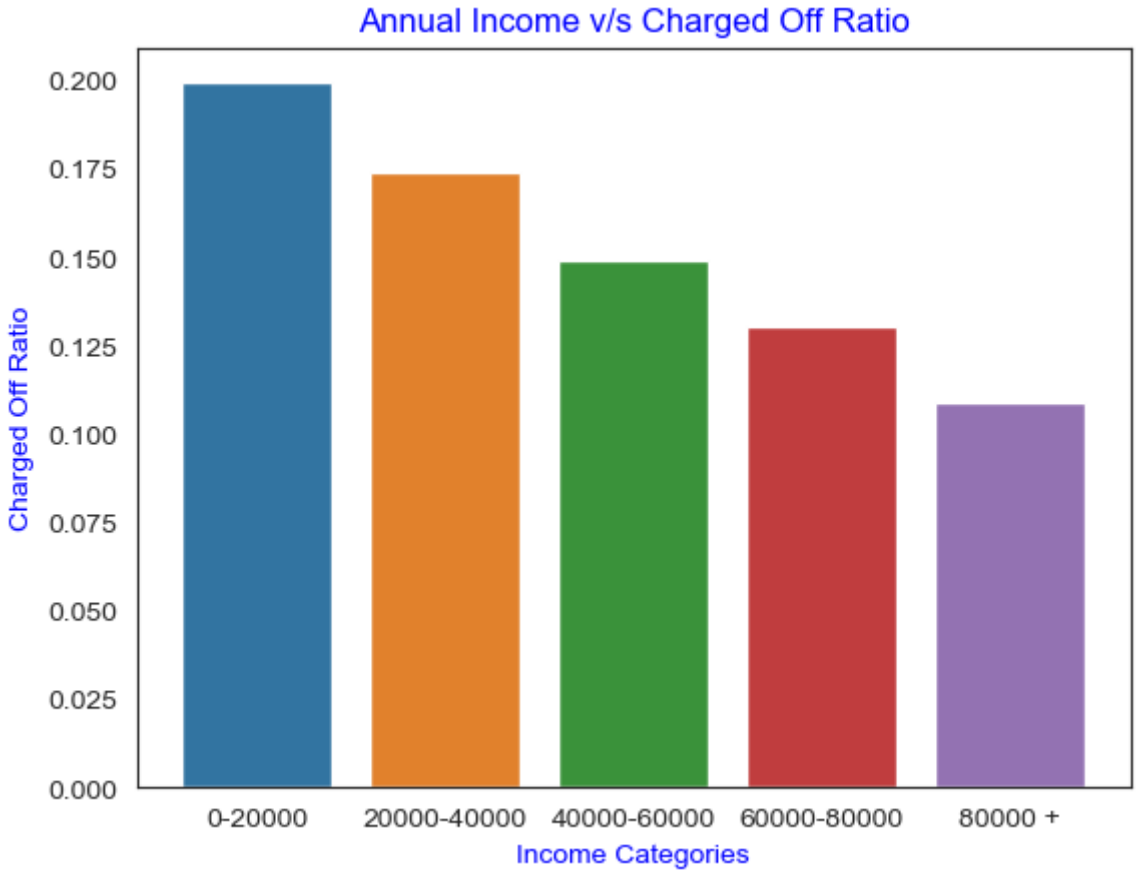
Analysis:

Bivariate Analysis

Annual Income v/s Charged Off Ratio

- Plotted categorized annual income against charged off ratio.
- From the plot we can concur that borrowers in the income range of $0-20000 have the highest chance of defaulting.
- The charged off ratio decreases as the annual income increases

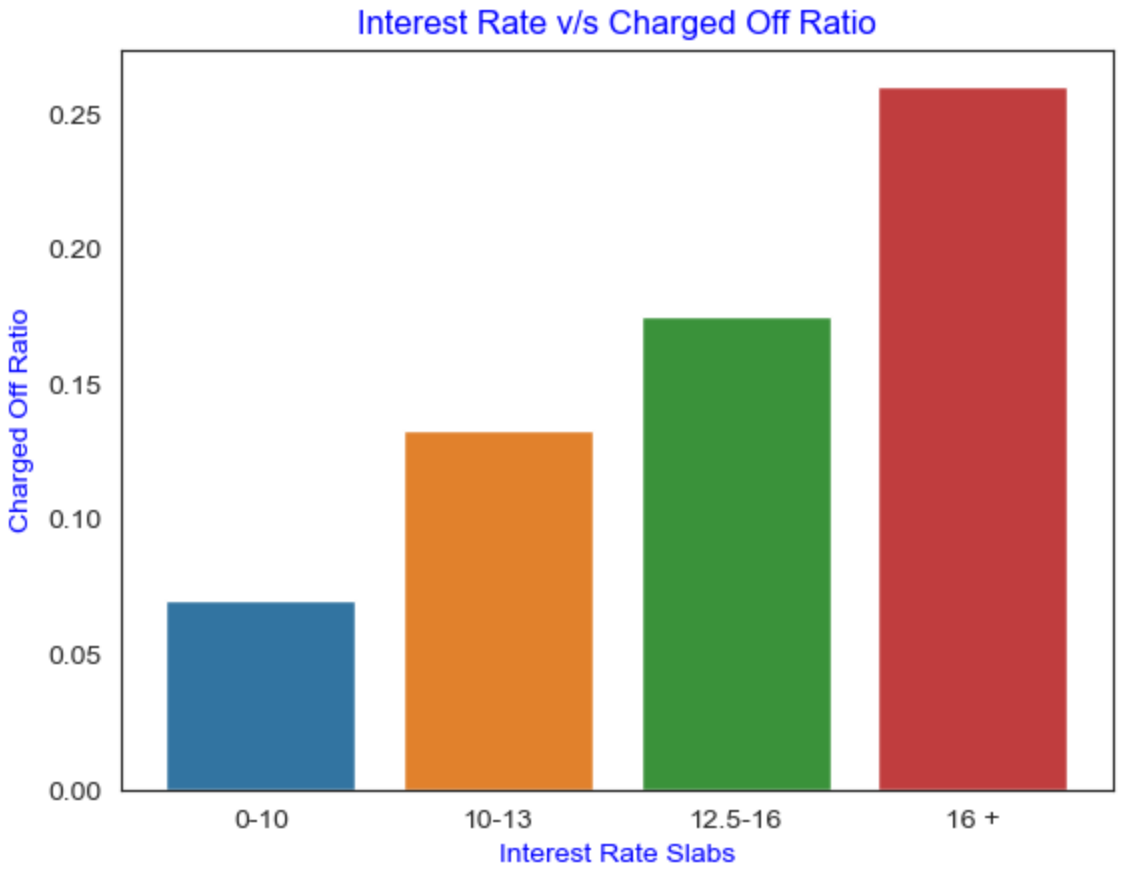
loan_status	annual_inc_cat	Charged Off	Current	Fully Paid	total	ratio
0	0-20000	237	9	943	1189	0.20
1	20000-40000	1514	170	7004	8688	0.17
2	40000-60000	1729	345	9534	11608	0.15
3	60000-80000	1024	240	6597	7861	0.13
4	80000 +	696	230	5468	6394	0.11



Interest Rate v/s Charged Off Ratio

- Plotted categorized interest rate against charged off ratio.
- From the plot we can concur interest rate of 16% or above has the highest charged off ratio.
- The charged off ratio decreases as the interest rate decreases.

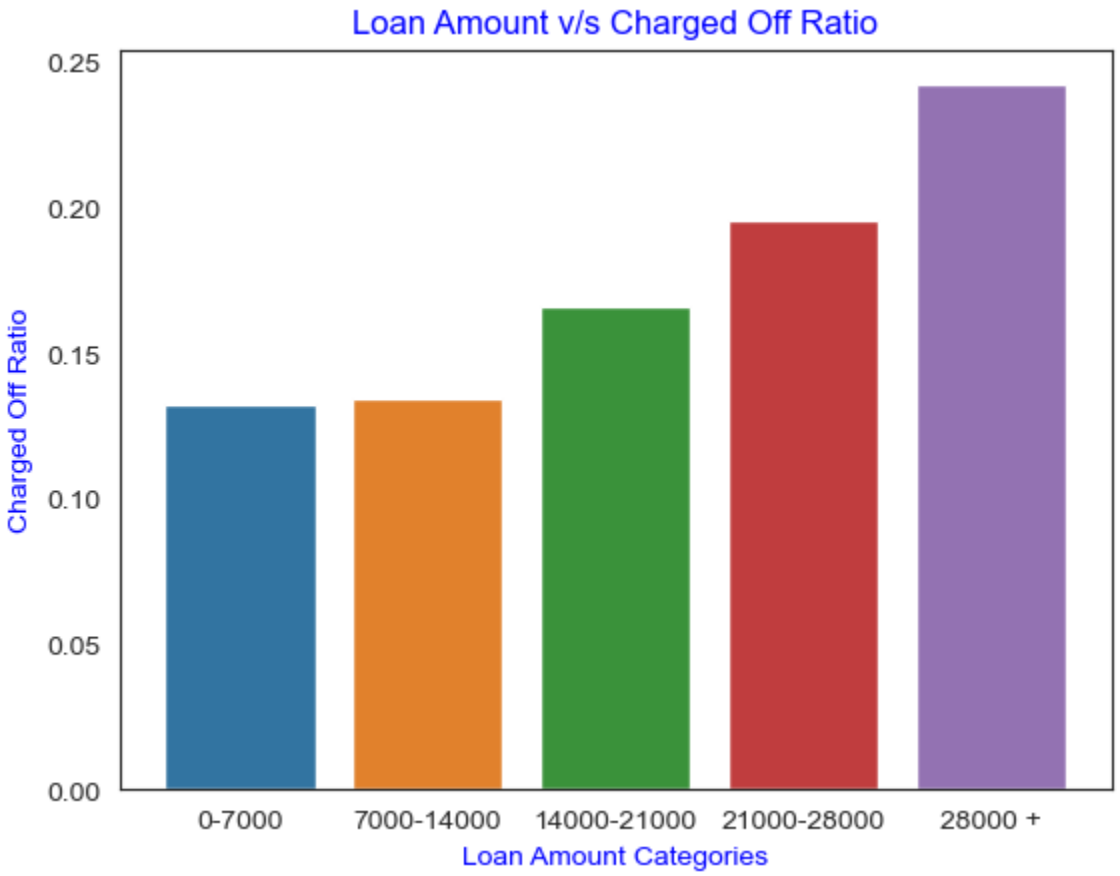
loan_status	int_rate_cat	Charged Off	Current	Fully Paid	total	ratio
0	0-10	792	74	10455	11321	0.07
1	10-13	1160	242	7334	8736	0.13
2	12.5-16	1870	294	8512	10676	0.18
3	16 +	1129	311	2892	4332	0.26



Loan Amount v/s Charged Off Ratio

- Plotted categorized loan amount against charged off ratio.
- From the plot we can concur that loan amount of $28,000 and above has the highest charge off ratio.
- The charged off ratio decreases as the loan amount decreases
- The charge off ratio below the amount of $14,000 is more or less same.

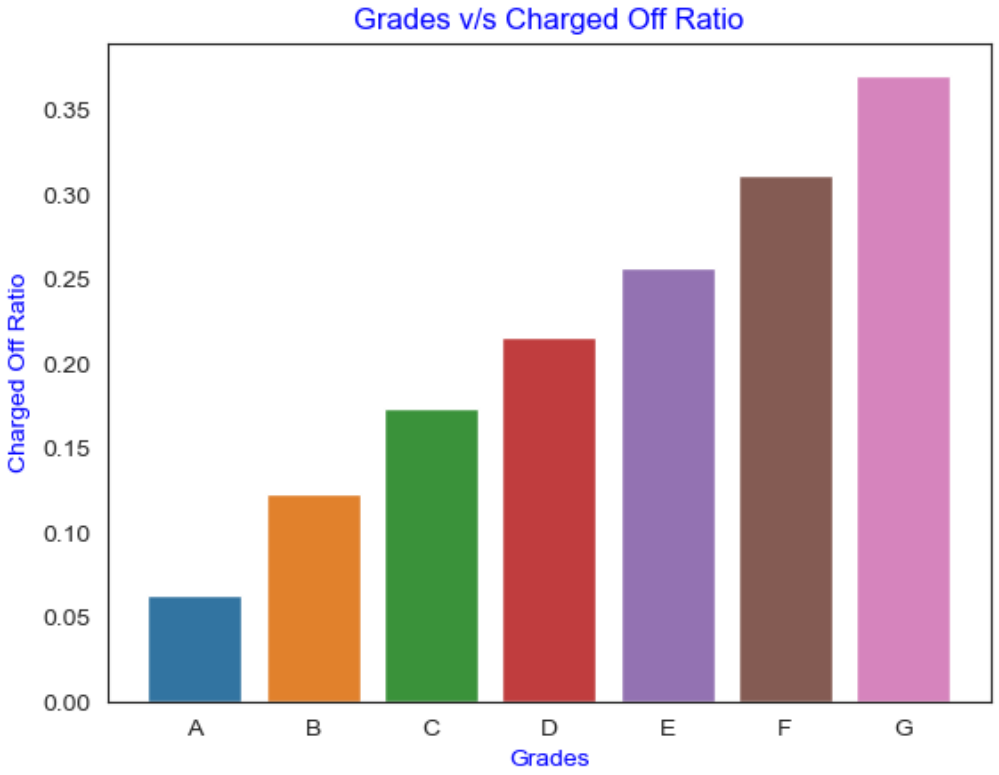
loan_status	loan_amnt_cat	Charged Off	Current	Fully Paid	total	ratio
0	0-7000	1837	160	11878	13875	0.13
1	7000-14000	1687	293	10541	12521	0.13
2	14000-21000	1061	308	5016	6385	0.17
3	21000-28000	427	144	1612	2183	0.20
4	28000 +	188	89	499	776	0.24



Grades v/s Charged Off Ratio

- Plotted grades against charged off ratio.
- From the plot we can concur that grades F and G have the highest charge off ratio
- The charged off ratio decreases as the grade moves towards A

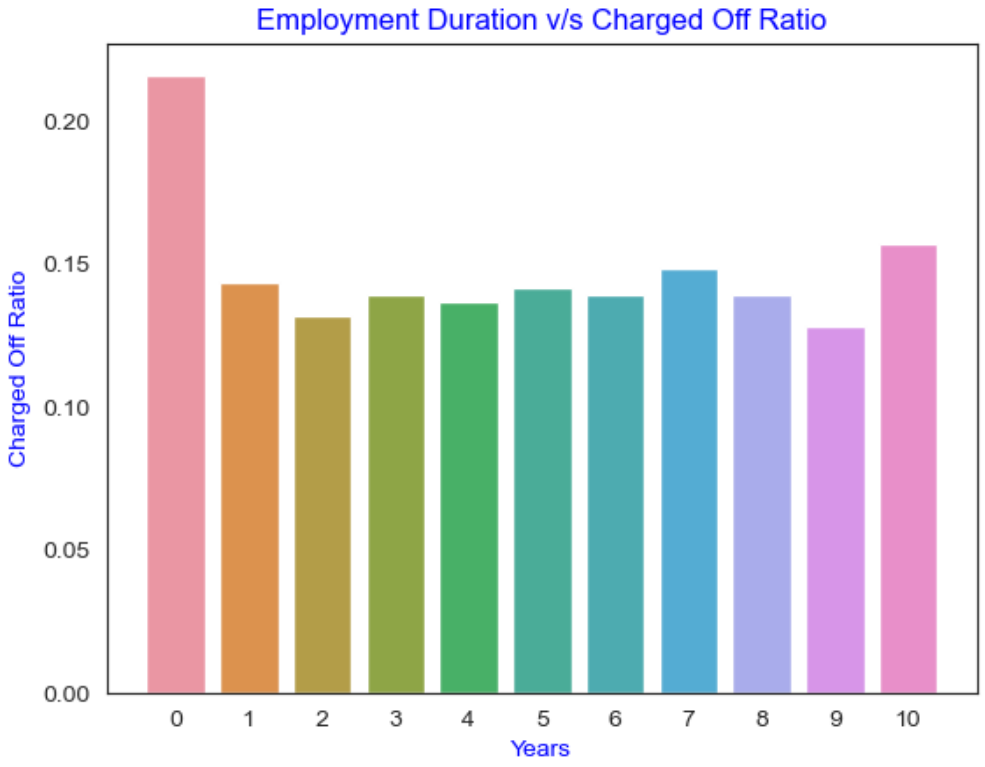
loan_status	grade	Charged Off	Current	Fully Paid	total	ratio
0	A	577	39	8588	9204	0.06
1	B	1337	310	9245	10892	0.12
2	C	1281	235	5848	7364	0.17
3	D	1027	197	3542	4766	0.22
4	E	620	149	1643	2412	0.26
5	F	269	52	541	862	0.31
6	G	89	12	139	240	0.37



Employment Duration v/s Charged Off Ratio

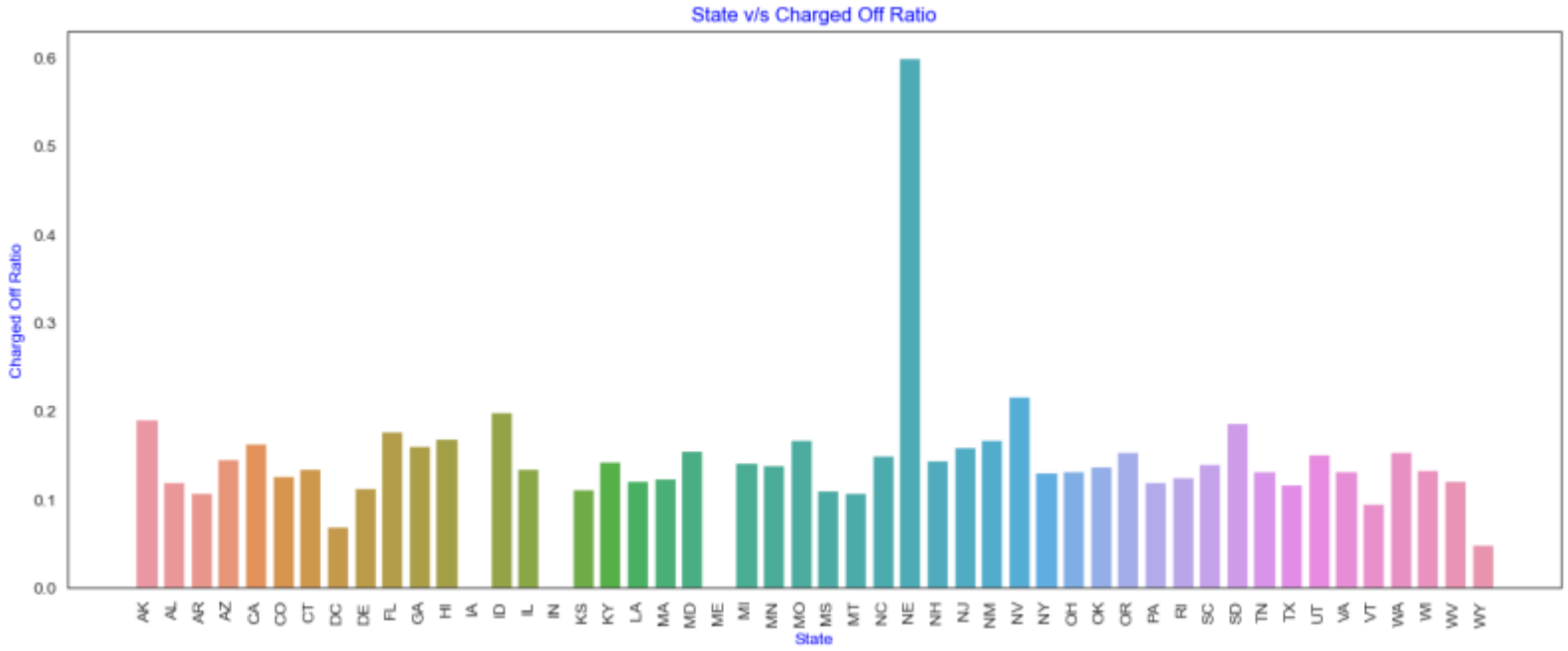
- Plotted employment duration against charged off ratio.
- From the plot we can concur that borrowers who have been employed for less than a year have a high chance of defaulting.
- The charged off ratio pretty much remains the same throughout else.

loan_status	emp_length	Charged Off	Current	Fully Paid	total	ratio
0	0	225	40	775	1040	0.22
1	1	1030	132	6019	7181	0.14
2	2	531	84	3409	4024	0.13
3	3	517	71	3115	3703	0.14
4	4	430	89	2626	3145	0.14
5	5	423	76	2486	2985	0.14
6	6	283	57	1689	2029	0.14
7	7	241	56	1326	1623	0.15
8	8	181	40	1080	1301	0.14
9	9	143	28	942	1113	0.13
10	10	1196	321	6079	7596	0.16



State v/s Charged Off Ratio

- Plotted states against charged off ratio.
- From the plot even though it seems that Nevada has the highest charge off ratio, the small number of applications from the state renders the inference inaccurate.



Debt-To-Income Ratio v/s Charged Off Ratio

- Plotted categorized debt-to-income ratio against charged off ratio.
- From the plot we the debt-to-income ratio of 25 and above seems to have the highest charge off ratio.
- The small variance in the charge off ratio across the categorized debt-to-income ratio might even be called pretty much same

loan_status	dti_cat	Charged Off	Current	Fully Paid	total	ratio
0	0-5	537	74	3627	4238	0.13
1	05-10	889	154	5854	6897	0.13
2	10-15	1289	226	7365	8880	0.15
3	15-20	1293	256	6861	8410	0.15
4	25+	1081	227	5179	6487	0.17