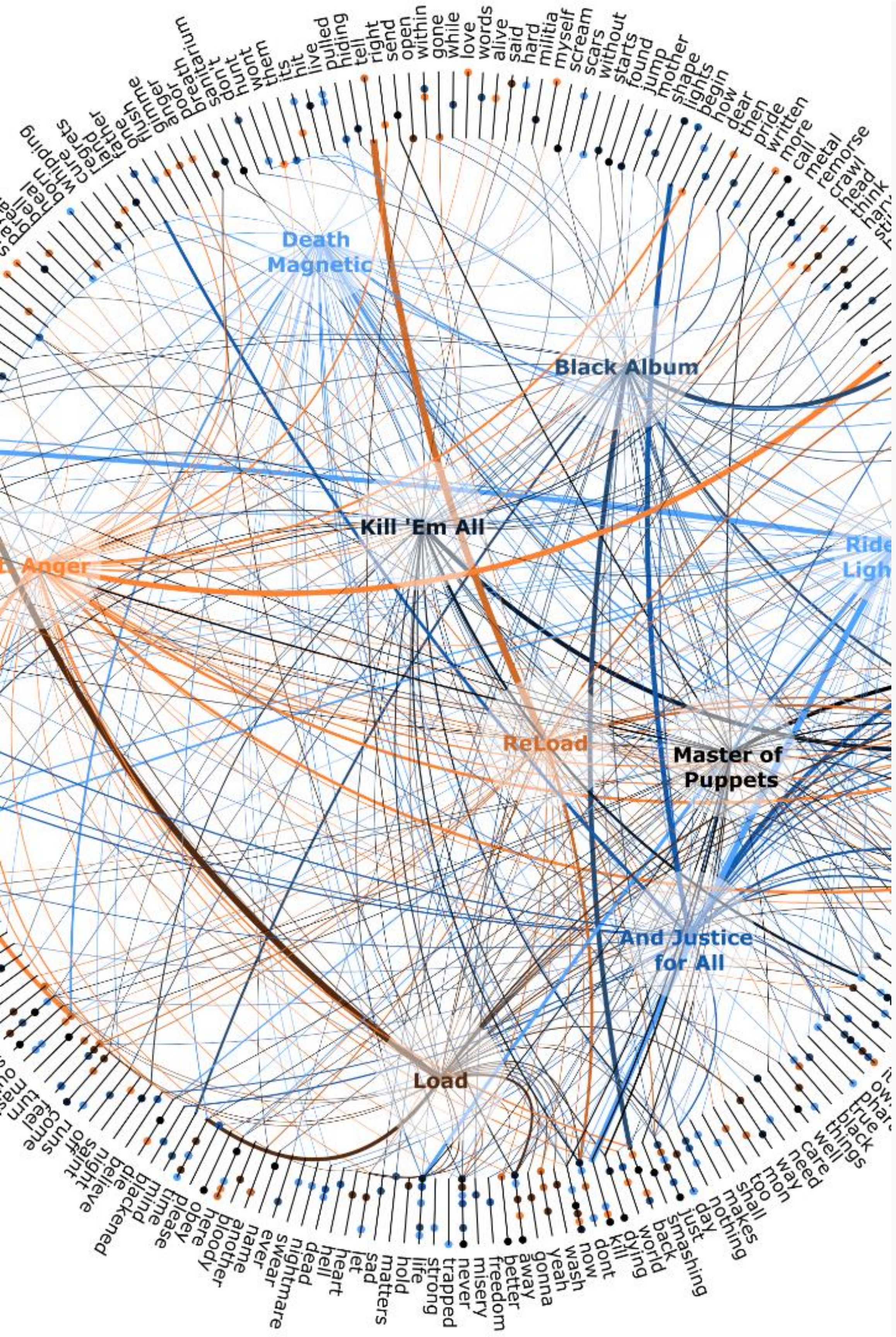
Lending Club Case Study

Lending Club 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.

It provides loans to valued customers. In this presentation, we will walk you through our customer data, understand the data, data cleaning and manipulation, data analysis done with python code and presentation and recommendations.

by Prem Subudhi and Shilpa Pandey



Data Understanding:

In this presentation, we will explore the key steps in data understanding. From identifying data quality issues to analyzing and interpreting variables.

Data Cleaning and Manipulation

1

Identifying and Addressing Data Quality Issues

All the missing values, outliers, and redundancies, are correctly identified and cleaned. The necessary steps are taken to address these issues.

Out of 111 columns 71 column have been removed.

2

Converting Data to Suitable Format

The data is formatted using appropriate methods, which ensured that the data is ready for analysis and manipulation.

Example : For numerical analysis the relevant columns modified to floats.

3

Manipulation of Strings and Dates

String and date manipulation is performed accurately and effectively whenever necessary. This allows for deeper analysis and understanding of the data. Example : The data attempted to numeric for few categorical data like loan status "Fully Paid Charged Off Current" to "1,2,3“ and emp_length is formatted to number of years by removing the > + years from the value string. The % character is removed in column int_rate 10.65% to 10.65.

Data Analysis: Uncovering the Insights

1

Coherent Problem Solving

The analysis focuses on a relevant problem that aligns with the business needs. The structure of the analysis is clear and easy to follow, ensuring that the insights are easily understood.

2

Univariate and Segmented Univariate Analysis

Univariate analysis is conducted accurately, taking into account segments and making realistic assumptions. At least five important driver variables, which strongly indicate default, are identified through the analysis.

3

Creation and Utilization of Metrics

Business-driven, type-driven, and data-driven metrics are created for important variables and effectively used for analysis. The derived metrics are well-explained and reasonable in their approach.

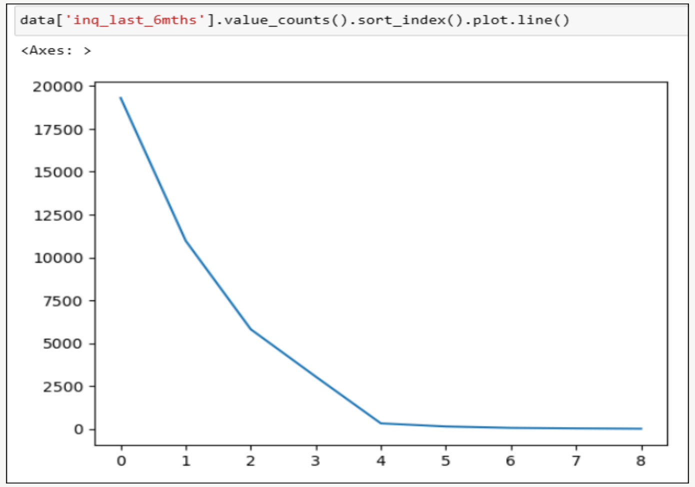
4

Bivariate Analysis and Identifying Combinations

Bivariate analysis is performed correctly, identifying the crucial combinations of driver variables. The chosen combinations make both business and analytical sense, leading to valuable insights.

```
0 100.0
```

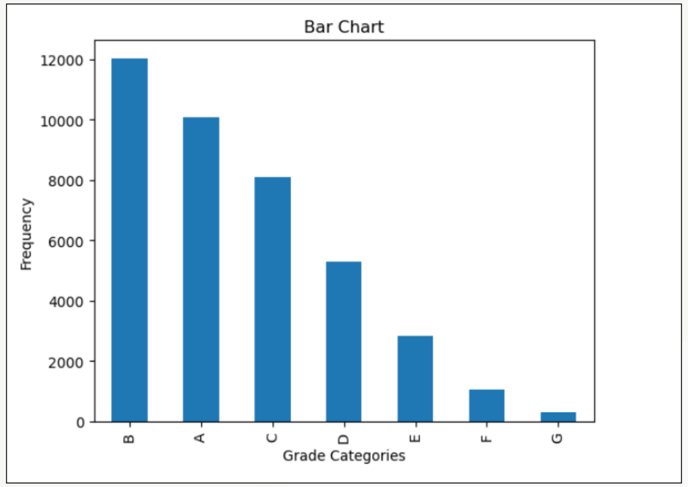
Visualizing Insights: Univariate Variable Analysis



```
data['inq_last_6mths'].value_counts().sort_index().plot.line()
```

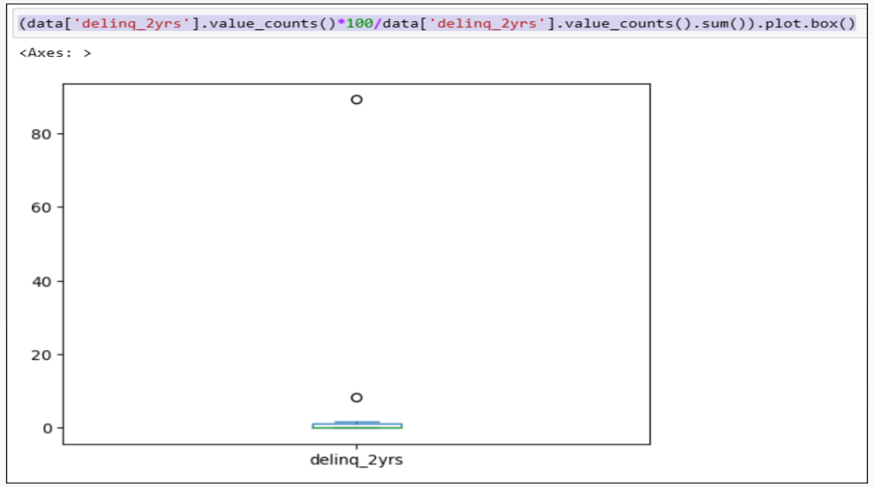
```
<Axes: >
```

The enquiries made during last six month shows around 20000 made by customer which demonstrates a huge credit requirement by customers which declined over 6months, It provide a business decision to capture this market opportunity. 0 : 100 for query (data['delinq_amnt'].value_counts()/data['delinq_amnt'].value_coun



The grades of loan offered over the time period.

The quality of loan reduces form A to G. the above plot shows mix of all the customer have and it lets us understand the consumer type and how this grading is made w.r.t consumer profile.



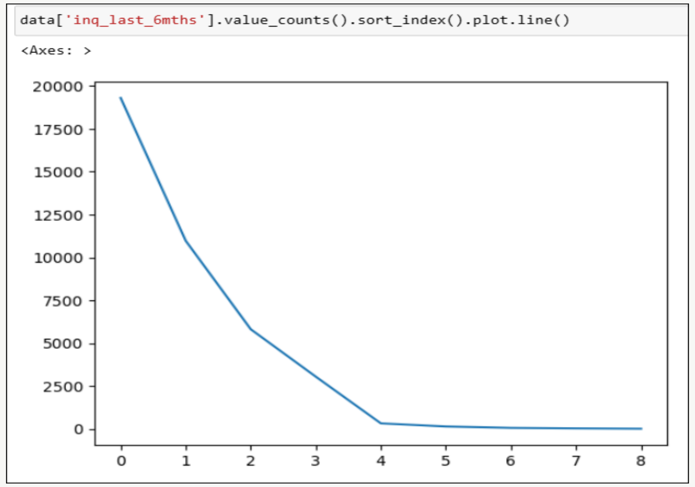
```
(data['delinq_2yrs'].value_counts()*100/data['delinq_2yrs'].value_counts().sum()).plot.box()
```

```
<Axes: >
```

Deli nq reported in 2 years

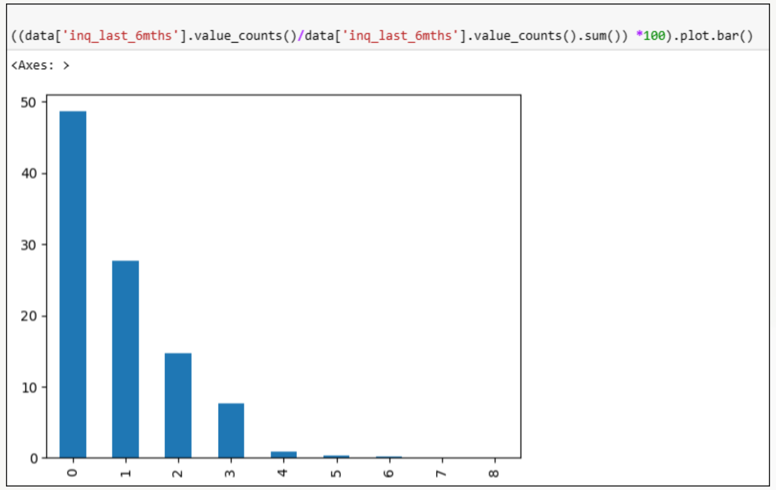
The box plot shows data points where around 11 time the payments were not made, and good number of defaults seen with customers to understand the risk of the loan can be given in this customer segment.

Visualizing Insights: Univariate Variable Analysis



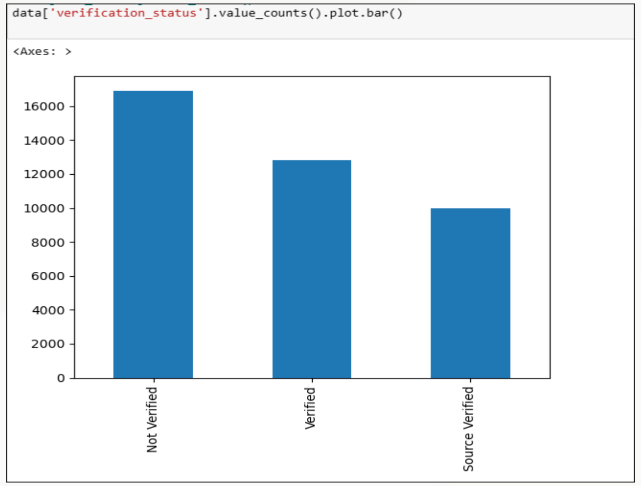
```
data['inq_last_6mths'].value_counts().sort_index().plot.line()
```

```
<Axes: >
```



```
((data['inq_last_6mths'].value_counts()/data['inq_last_6mths'].value_counts().sum()) *100).plot.bar()
```

```
<Axes: >
```



```
data['verification_status'].value_counts().plot.bar()
```

```
<Axes: >
```

(data['acc_now_delinq'].value_counts()/data['acc_now_delinq'].value_counts().sum()) *100

(data['delinq_amnt'].value_counts()/data['delinq_amnt'].value_counts().sum()) *100

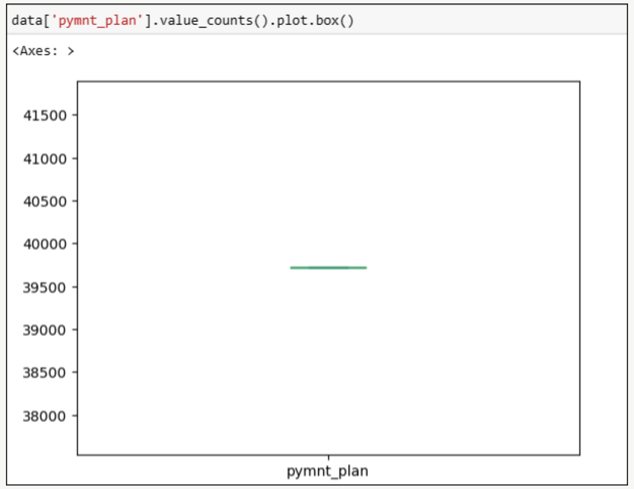
Returns 0 : 100

The enquiries does provide the input to the risk assessment. With number of enquiry increases the chances of disbursing the loan will be subject to enquiry.

Verification Status seems in little risk as all are not verified yet.

This column poses some kind risk as all the customers information is not verified yet. Around 16+k records status does not have data yet to provide loan.

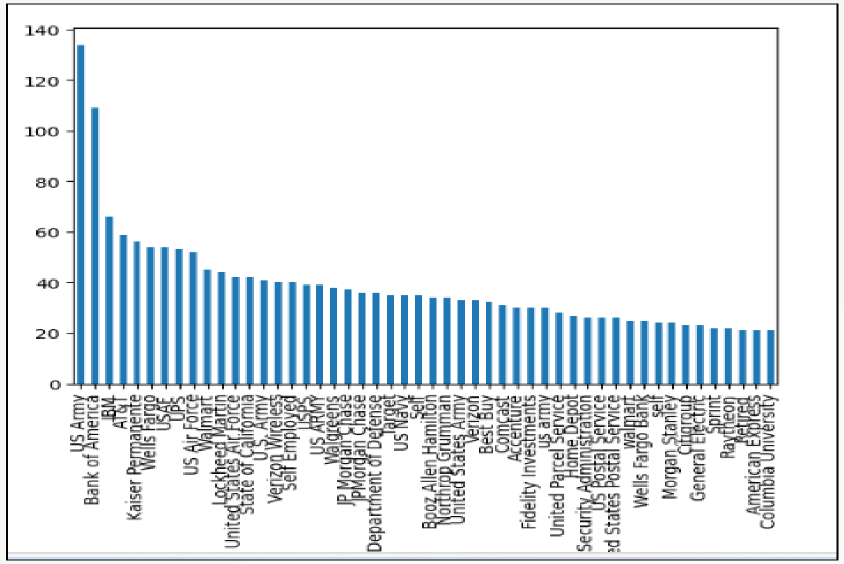
Visualizing Insights: Univariate Variable Analysis



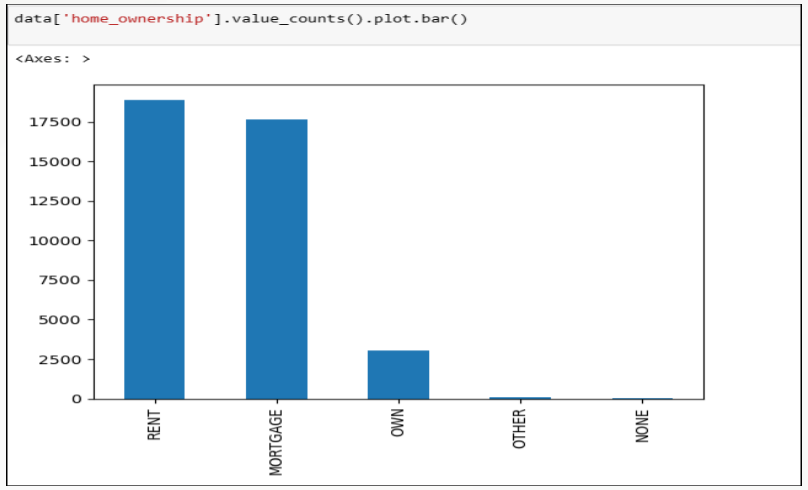
```
data['pymnt_plan'].value_counts().plot.box()
```

```
<Axes: >
```

All the records does not have a plan i.e all are expected to pay they emi's in regular as per installments chosen. This is a positive factor as it does not have any complexity to deal with.



The employee title graph looks good and if all the information verified true then the risk level are low.

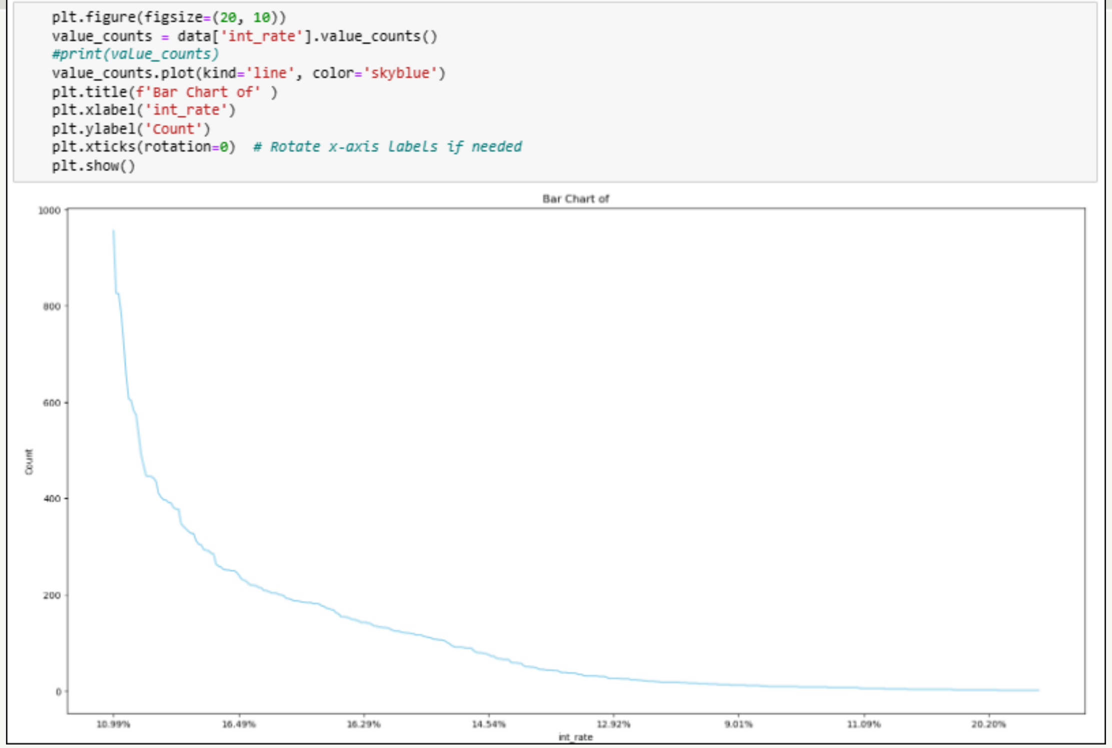


```
data['home_ownership'].value_counts().plot.bar()
```

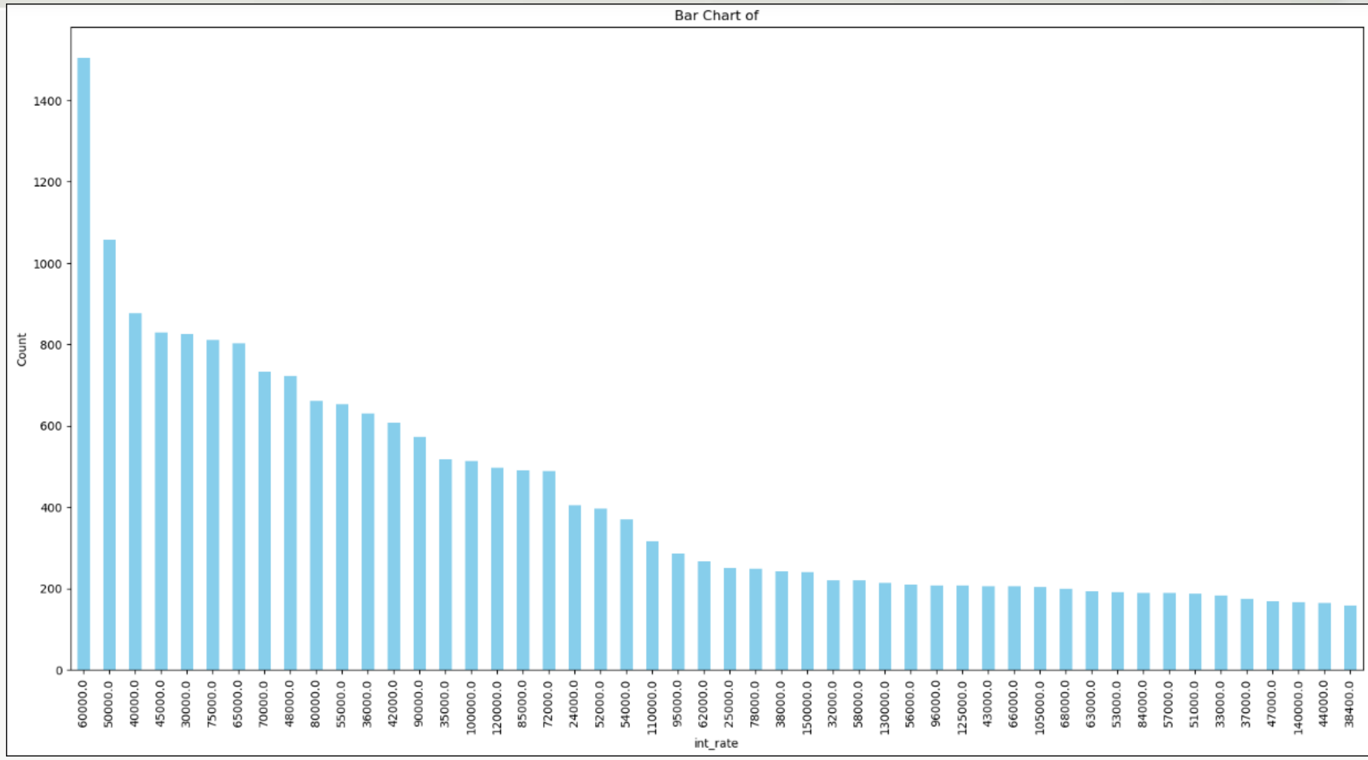
```
<Axes: >
```

The home ownership shows most customers have Rental and Mortgage.

The ownership has very low number it's a one of risk factor to consider before providing the approval to the loan. We may derive another information, like there may Many first time loan property who stays in Rental house and going for a own home, then the risk factor would be low which can be can be verified.

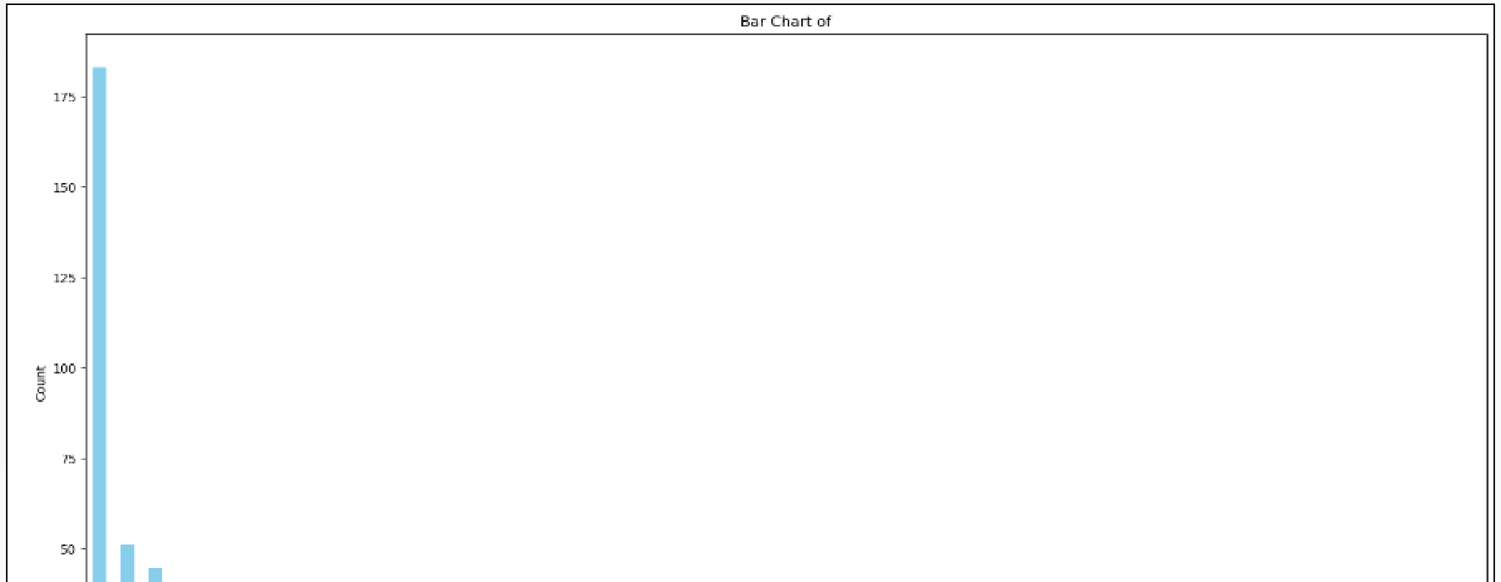


```
plt.figure(figsize=(20, 10))
value_counts = data['int_rate'].value_counts()
#print(value_counts)
value_counts.plot(kind='line', color='skyblue')
plt.title(f'Bar Chart of' )
plt.xlabel('int_rate')
plt.ylabel('Count')
plt.xticks(rotation=0)  # Rotate x-axis labels if needed
plt.show()
```

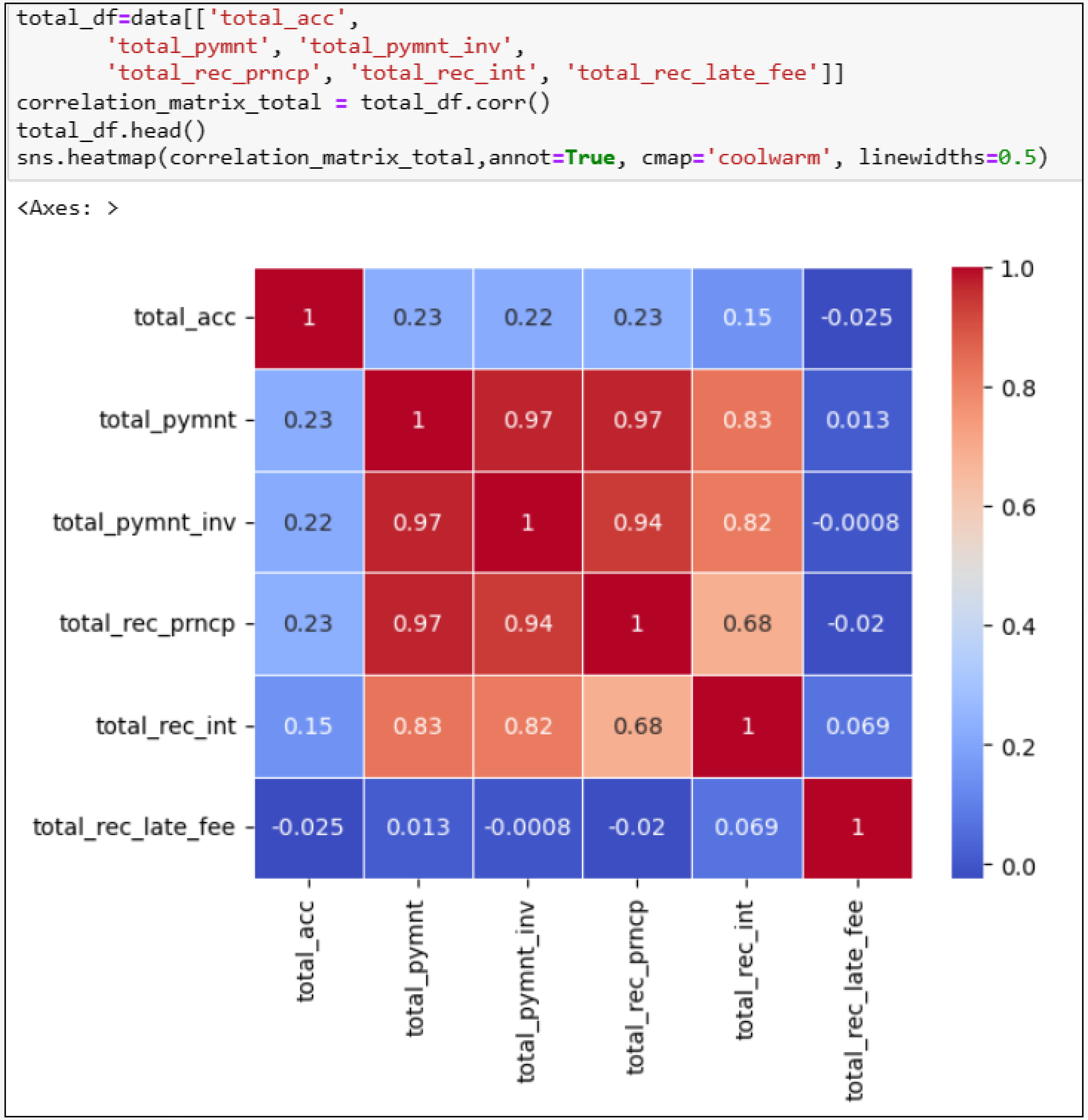


The int_rate column has most customers
Fall in the range interest for 10.99 to 0,
the risk factor for lending the loan interest
rate of 10.99 may posses risk in repayment
 for lower income group

The highest income group who has applied for loan is 60k it give
a great insight the market segment who is taking loan in this
amount. For less amount the risk is distributed to may customers.
It may be viable option for lender with lower risk.



For DTI column the plot show a good number of amount who has
not taken loan any before and it has good market segment where
The risk is the lowest to invest.



```
total_df=data[['total_acc',
       'total_pymnt', 'total_pymnt_inv',
       'total_rec_prncp', 'total_rec_int', 'total_rec_late_fee']]
correlation_matrix_total = total_df.corr()
total_df.head()
sns.heatmap(correlation_matrix_total,annot=True, cmap='coolwarm', linewidths=0.5)
```

```
<Axes: >
```

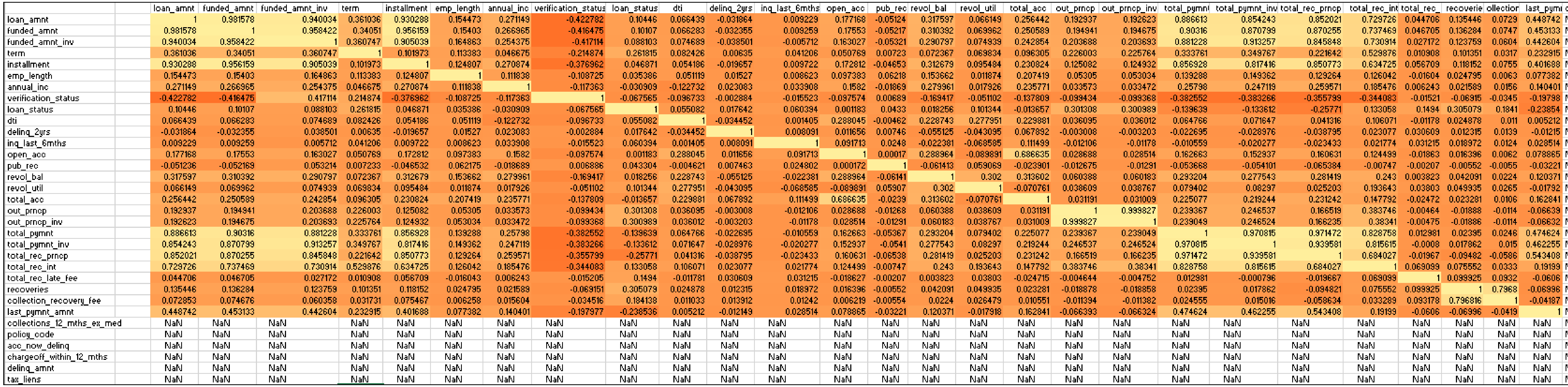
	total_acc	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	total_rec_late_fee
total_acc	1	0.23	0.22	0.23	0.15	-0.025
total_pymnt	0.23	1	0.97	0.97	0.83	0.013
total_pymnt_inv	0.22	0.97	1	0.94	0.82	-0.0008
total_rec_prncp	0.23	0.97	0.94	1	0.68	-0.02
total_rec_int	0.15	0.83	0.82	0.68	1	0.069
total_rec_late_fee	-0.025	0.013	-0.0008	-0.02	0.069	1

This heatmap shows the total values of account, payments, principal , interest and late fee the correlation matrix. The total payments done by investors and principal got high correlation with the total payments made.

And the correlation between late fee and received principal is negative which is obvious.

The data seems as expected and a good input for the risk analysis.

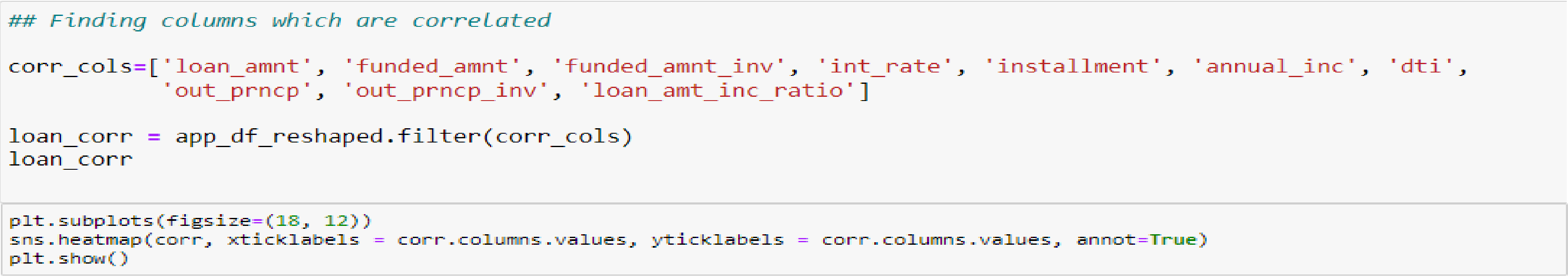
Visualizing Insights: Correlation Metrics



	loan_amnt	funded_amnt	funded_amnt_inv	term	installment	emp_length	annual_inc	verification_status	loan_status	dti	delinq_2yrs	inq_last_6mths	open_acc	pub_rec	revol_bal	revol_util	total_acc	out_prncp	out_prncp_inv	total_pymnt	total_pymnt_inv	total_rec_prncp	total_rec_int	total_rec_	recoverie	ollection	last_pymi	c
loan_amnt	1	0.981578	0.940034	0.361036	0.930288	0.154473	0.271149	-0.422782	0.10446	0.066439	-0.031864	0.009229	0.177168	-0.05124	0.317597	0.066149	0.256442	0.192937	0.192623	0.886613	0.854243	0.852021	0.729726	0.044706	0.135446	0.0729	0.448742	N
funded_amnt	0.981578	1	0.958422	0.34051	0.956159	0.15403	0.266965	-0.416475	0.10107	0.066283	-0.032355	0.009259	0.17553	-0.05217	0.310392	0.069962	0.250589	0.194941	0.194675	0.90316	0.870799	0.870255	0.737469	0.046705	0.136284	0.0747	0.453133	N
funded_amnt_inv	0.940034	0.958422	1	0.360747	0.905039	0.164863	0.254375	-0.417114	0.088103	0.074689	-0.038501	-0.005712	0.163027	-0.05321	0.290797	0.074939	0.242854	0.203688	0.203693	0.881228	0.913257	0.845848	0.730914	0.027172	0.123759	0.0604	0.442604	N
term	0.361036	0.34051	0.360747	1	0.101973	0.113383	0.046675	-0.214874	0.261815	0.082426	0.00635	0.041206	0.050769	0.00723	0.072367	0.069834	0.096305	0.226003	0.225764	0.333761	0.349767	0.221642	0.529876	0.010908	0.101351	0.0317	0.232915	N
installment	0.930288	0.956159	0.905039	0.101973	1	0.124807	0.270874	-0.376962	0.046871	0.054186	-0.019657	0.009722	0.172812	-0.04653	0.312679	0.095484	0.230824	0.125082	0.124932	0.856928	0.817416	0.850773	0.634725	0.056709	0.118152	0.0755	0.401688	N
emp_length	0.154473	0.15403	0.164863	0.113383	0.124807	1	0.111838	-0.108725	0.035386	0.051119	0.01527	0.008623	0.097383	0.06218	0.153662	0.011874	0.207419	0.05305	0.053034	0.139288	0.149362	0.129264	0.126042	-0.01604	0.024795	0.0063	0.077382	N
annual_inc	0.271149	0.266965	0.254375	0.046675	0.270874	0.111838	1	-0.117363	-0.030909	-0.122732	0.023083	0.033908	0.1582	-0.01869	0.279961	0.017926	0.235771	0.033573	0.033472	0.25798	0.247119	0.259571	0.185476	0.006243	0.021589	0.0156	0.140401	N
verification_status	-0.422782	-0.416475	0.417114	0.214874	-0.376962	-0.108725	-0.117363	1	-0.067565	-0.096733	-0.002884	-0.015523	-0.097574	0.00689	-0.169417	-0.051102	-0.137809	-0.099434	-0.099368	-0.382552	-0.383266	-0.355799	-0.344083	-0.01521	-0.06915	-0.0345	-0.19798	N
loan_status	0.10446	0.10107	0.088103	0.261815	0.046871	0.035386	-0.030909	-0.067565	1	0.055082	0.017642	0.060394	0.001183	0.0433	0.018256	0.101344	-0.013657	0.301308	0.300989	-0.139639	-0.133612	-0.25771	0.133058	0.1494	0.305079	0.1841	-0.23854	N
dti	0.066439	0.066283	0.074689	0.082426	0.054186	0.051119	-0.122732	-0.096733	0.055082	1	-0.034452	0.001405	0.288045	-0.00462	0.228743	0.277951	0.229881	0.036095	0.036012	0.064766	0.071647	0.041316	0.106071	-0.01178	0.024878	0.011	0.005212	N
delinq_2yrs	-0.031864	-0.032355	0.038501	0.00635	-0.019657	0.01527	0.023083	-0.002884	0.017642	-0.034452	1	0.008091	0.011656	0.00746	-0.055125	-0.043095	0.067892	-0.003008	-0.003203	-0.022695	-0.028976	-0.038795	0.023077	0.030609	0.012315	0.0139	-0.01215	N
inq_last_6mths	0.009229	0.009259	0.005712	0.041206	0.009722	0.008623	0.033908	-0.015523	0.060394	0.001405	0.008091	1	0.091713	0.0248	-0.022381	-0.068585	0.111499	-0.012106	-0.01178	-0.010559	-0.020277	-0.023433	0.021774	0.031215	0.018972	0.0124	0.028514	N
open_acc	0.177168	0.17553	0.163027	0.050769	0.172812	0.097383	0.1582	-0.097574	0.001183	0.288045	0.011656	0.091713	1	0.00017	0.288964	-0.089891	0.686635	0.028688	0.028514	0.162663	0.152937	0.160631	0.124499	-0.01863	0.016396	0.0062	0.078865	N
pub_rec	-0.051236	-0.052169	0.053214	0.007233	-0.046532	0.062175	-0.018689	0.006886	0.043304	-0.004621	0.007463	0.024802	0.000172	1	-0.061413	0.059069	-0.023901	-0.012675	-0.01291	-0.053668	-0.054101	-0.065384	-0.00747	-0.00207	-0.00552	-0.0055	-0.03221	N
revol_bal	0.317597	0.310392	0.290797	0.072367	0.312679	0.153662	0.279961	-0.169417	0.018256	0.228743	-0.055125	-0.022381	0.288964	-0.06141	1	0.302	0.313602	0.060388	0.060183	0.293204	0.277543	0.281419	0.243	0.003823	0.042091	0.0224	0.120371	N
revol_util	0.066149	0.069962	0.074939	0.069834	0.095484	0.011874	0.017926	-0.051102	0.101344	0.277951	-0.043095	-0.068585	-0.089891	0.05907	0.302	1	-0.070761	0.038609	0.038767	0.079402	0.08297	0.025203	0.193643	0.03803	0.049935	0.0265	-0.01792	N
total_acc	0.256442	0.250589	0.242854	0.096305	0.230824	0.207419	0.235771	-0.137809	-0.013657	0.229881	0.067892	0.111499	0.686635	-0.0239	0.313602	-0.070761	1	0.031191	0.031009	0.225077	0.219244	0.231242	0.147792	-0.02472	0.023281	0.0106	0.162841	N
out_prncp	0.192937	0.194941	0.203688	0.226003	0.125082	0.05305	0.033573	-0.099434	0.301308	0.036095	-0.003008	-0.012106	0.028688	-0.01268	0.060388	0.038609	0.031191	1	0.999827	0.239367	0.246537	0.166519	0.383746	-0.00464	-0.01888	-0.0114	-0.06639	N
out_prncp_inv	0.192623	0.194675	0.203693	0.225764	0.124932	0.053034	0.033472	-0.099368	0.300989	0.036012	-0.003203	-0.01178	0.028514	-0.01291	0.060183	0.038767	0.031009	0.999827	1	0.239049	0.246524	0.166235	0.38341	-0.00475	-0.01886	-0.0114	-0.06632	N
total_pymnt	0.886613	0.90316	0.881228	0.333761	0.856928	0.139288	0.25798	-0.382552	-0.139639	0.064766	-0.022695	-0.010559	0.162663	-0.05367	0.293204	0.079402	0.225077	0.239367	0.239049	1	0.970815	0.971472	0.828758	0.012981	0.02395	0.0246	0.474624	N
total_pymnt_inv	0.854243	0.870799	0.913257	0.349767	0.817416	0.149362	0.247119	-0.383266	-0.133612	0.071647	-0.028976	-0.020277	0.152937	-0.0541	0.277543	0.08297	0.219244	0.246537	0.246524	0.970815	1	0.939581	0.815615	-0.0008	0.017862	0.015	0.462255	N
total_rec_prncp	0.852021	0.870255	0.845848	0.221642	0.850773	0.129264	0.259571	-0.355799	-0.25771	0.041316	-0.038795	-0.023433	0.160631	-0.06538	0.281419	0.025203	0.231242	0.166519	0.166235	0.971472	0.939581	1	0.684027	-0.01967	-0.09482	-0.0586	0.543408	N
total_rec_int	0.729726	0.737469	0.730914	0.529876	0.634725	0.126042	0.185476	-0.344083	0.133058	0.106071	0.023077	0.021774	0.124499	-0.00747	0.243	0.193643	0.147792	0.383746	0.38341	0.828758	0.815615	0.684027	1	0.069099	0.075552	0.0333	0.19199	N
total_rec_late_fee	0.044706	0.046705	0.027172	0.010908	0.056709	-0.016043	0.006243	-0.015205	0.1494	-0.011781	0.030609	0.031215	-0.018627	-0.00207	0.003823	0.03803	-0.024715	-0.004644	-0.004752	0.012981	-0.000796	-0.019667	0.069099	1	0.099925	0.0932	-0.0606	N
recoveries	0.135446	0.136284	0.123759	0.101351	0.118152	0.024795	0.021589	-0.069151	0.305079	0.024878	0.012315	0.018972	0.016396	-0.00552	0.042091	0.049935	0.023281	-0.018878	-0.018858	0.02395	0.017862	-0.094821	0.075552	0.099925	1	0.7968	-0.06996	N
collection_recovery_fee	0.072853	0.074676	0.060358	0.031731	0.075467	0.006258	0.015604	-0.034516	0.184138	0.011033	0.013912	0.01242	0.006219	-0.00554	0.0224	0.026479	0.010551	-0.011394	-0.011382	0.024555	0.015016	-0.058634	0.033289	0.093178	0.796816	1	-0.04187	N
last_pymnt_amnt	0.448742	0.453133	0.442604	0.232915	0.401688	0.077382	0.140401	-0.197977	-0.238536	0.005212	-0.012149	0.028514	0.078865	-0.03221	0.120371	-0.017918	0.162841	-0.066393	-0.066324	0.474624	0.462255	0.543408	0.19199	-0.0606	-0.06996	-0.0419	1	N
collections_12_mths_ex_med	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
policy_code	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
acc_now_delinq	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
chargeoff_within_12_mths	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
delinq_amnt	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N
tax_liens	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	N

On the cleaned data we created a correlation metrics demonstrates the correlation among the columns, it will help to understand how the columns are affected by the other variables.

Visualizing Insights: Bivariate Analysis

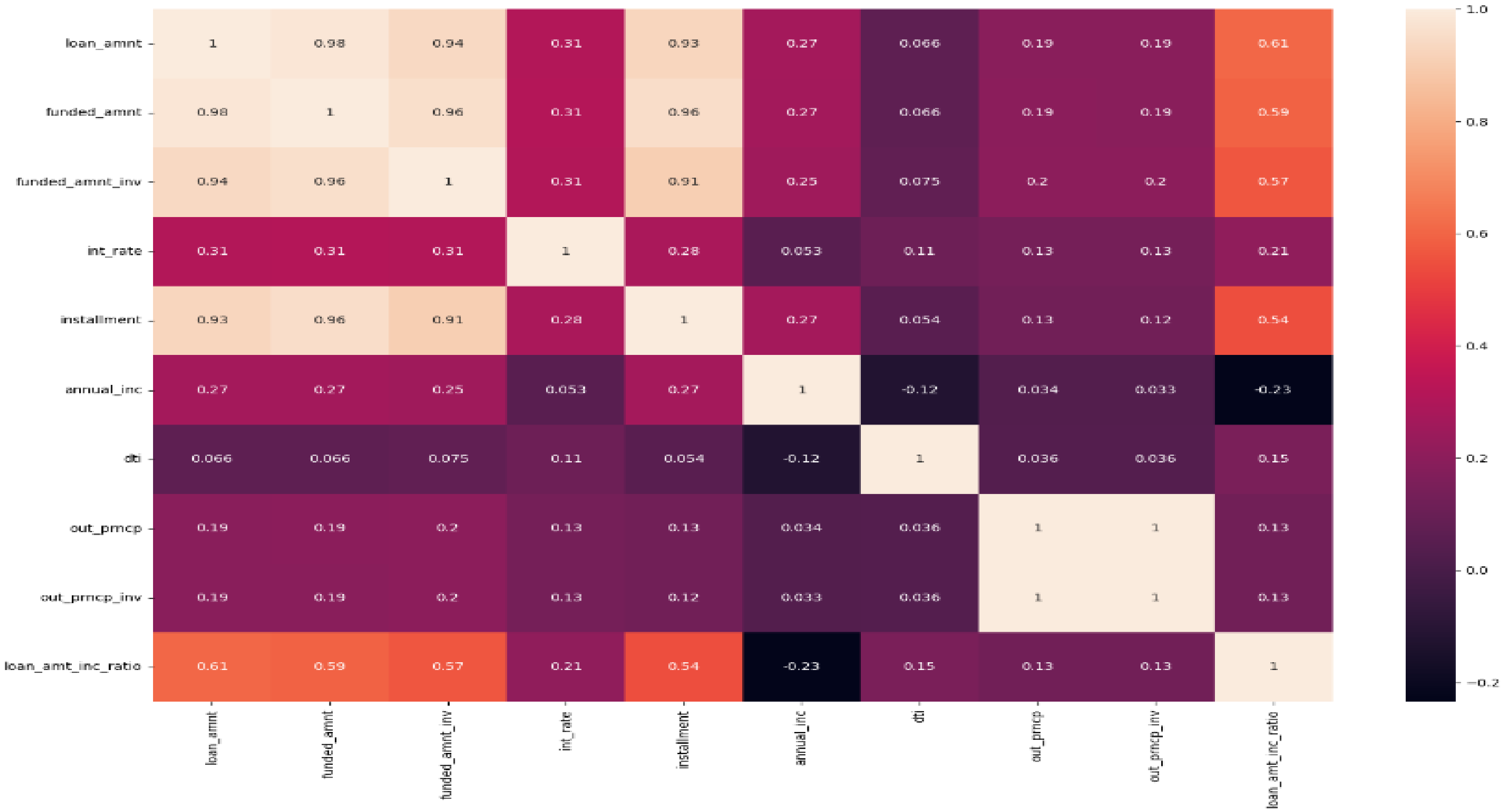


```
## Finding columns which are correlated

corr_cols=['loan_amnt', 'funded_amnt', 'funded_amnt_inv', 'int_rate', 'installment', 'annual_inc', 'dti',
           'out_prncp', 'out_prncp_inv', 'loan_amt_inc_ratio']

loan_corr = app_df_reshaped.filter(corr_cols)
loan_corr
```

```
plt.subplots(figsize=(18, 12))
sns.heatmap(corr, xticklabels = corr.columns.values, yticklabels = corr.columns.values, annot=True)
plt.show()
```

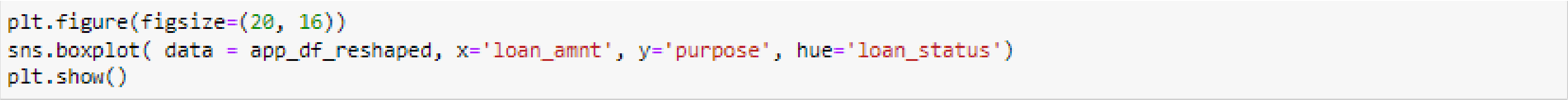


	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	out_prncp	out_prncp_inv	loan_amt_inc_ratio
loan_amnt	1	0.98	0.94	0.31	0.93	0.27	0.066	0.19	0.19	0.61
funded_amnt	0.98	1	0.96	0.31	0.96	0.27	0.066	0.19	0.19	0.59
funded_amnt_inv	0.94	0.96	1	0.31	0.91	0.25	0.075	0.2	0.2	0.57
int_rate	0.31	0.31	0.31	1	0.28	0.053	0.11	0.13	0.13	0.21
installment	0.93	0.96	0.91	0.28	1	0.27	0.054	0.13	0.12	0.54
annual_inc	0.27	0.27	0.25	0.053	0.27	1	-0.12	0.034	0.033	-0.23
dti	0.066	0.066	0.075	0.11	0.054	-0.12	1	0.036	0.036	0.15
out_prncp	0.19	0.19	0.2	0.13	0.13	0.034	0.036	1	1	0.13
out_prncp_inv	0.19	0.19	0.2	0.13	0.12	0.033	0.036	1	1	0.13
loan_amt_inc_ratio	0.61	0.59	0.57	0.21	0.54	-0.23	0.15	0.13	0.13	1

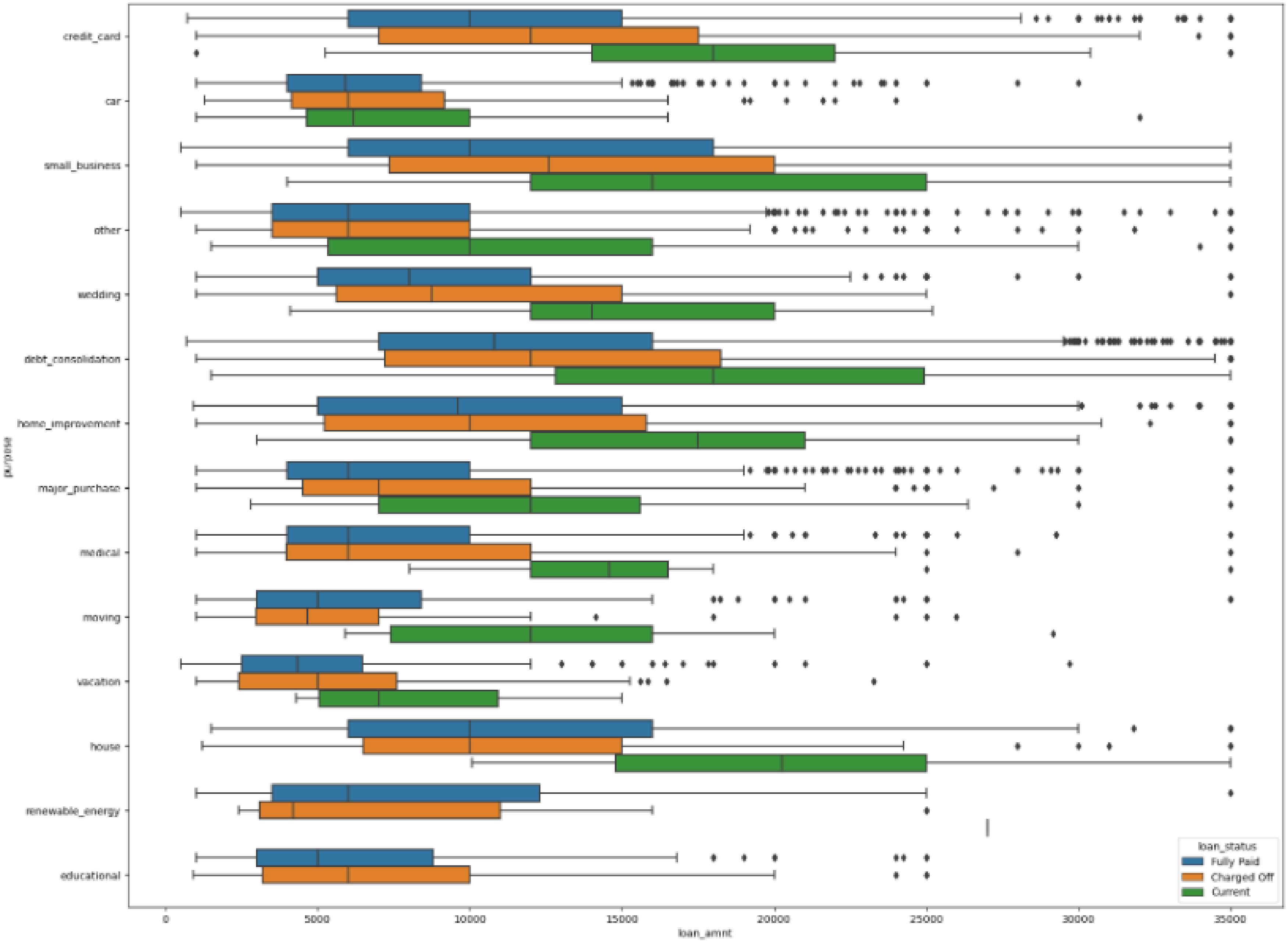
Generating another heatmap to find correlation between loan amount, funded amount, interest rate, installment, annual income, out principal, dti, loan amount income ration.

Inference from the heatmap : *Loan Amount, Funded Amount and Annual Income is highly correlated.*

Visualizing Insights: Bivariate Analysis



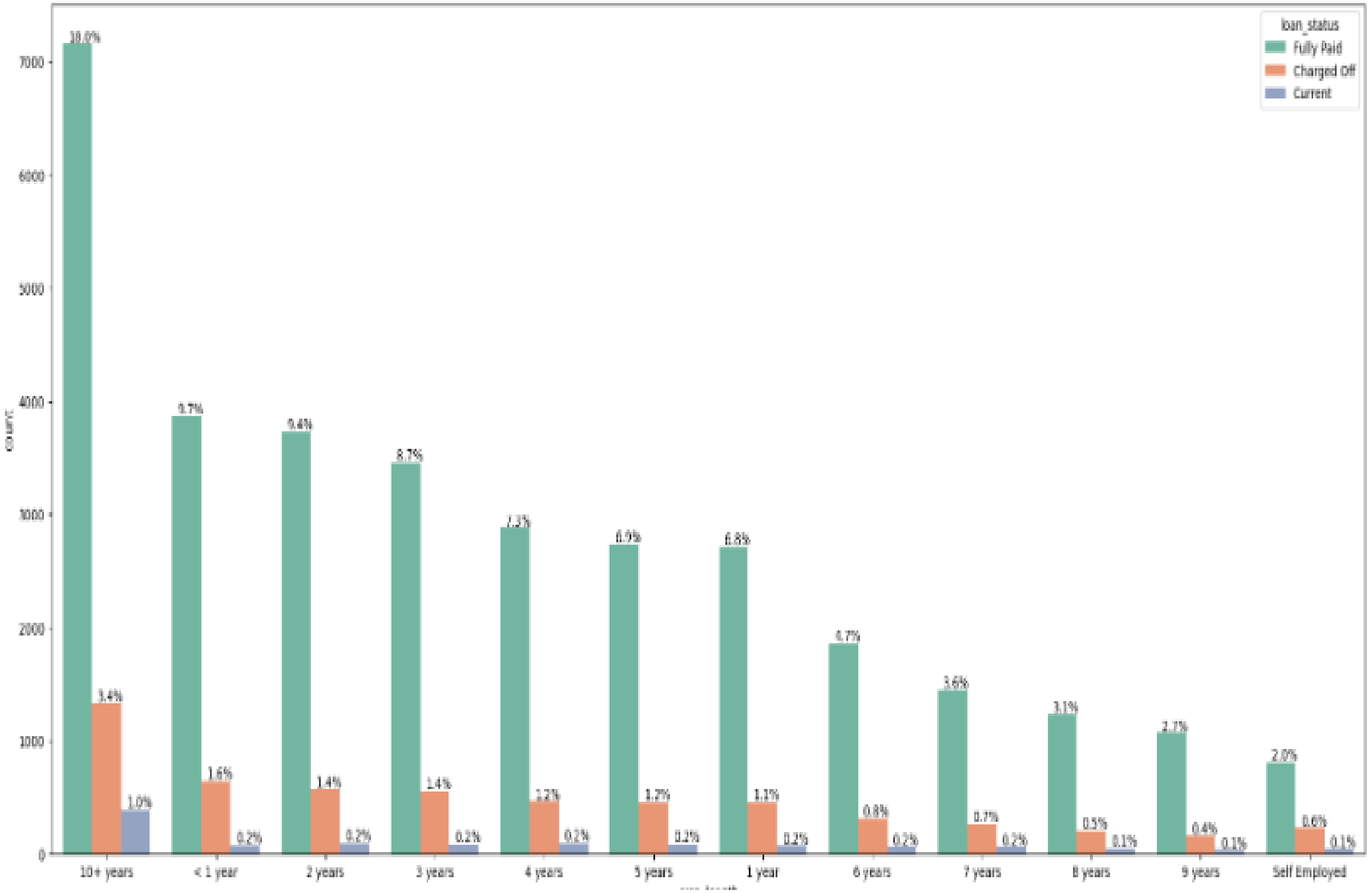
```
plt.figure(figsize=(20, 16))
sns.boxplot( data = app_df_reshaped, x='loan_amnt', y='purpose', hue='loan_status')
plt.show()
```



Analyze the purpose of loan to find the motive between loan and the amount with respect to each loan status.

Inference : *Purpose having 'other' and 'major purchase' type are having lot of outliers in charge off which will result in more loss to the bank. 'Small business' purpose type also needs an attention. So overall, 'other', 'major purchase' and 'small business' types needs to be considered while lending loan to the customers.*

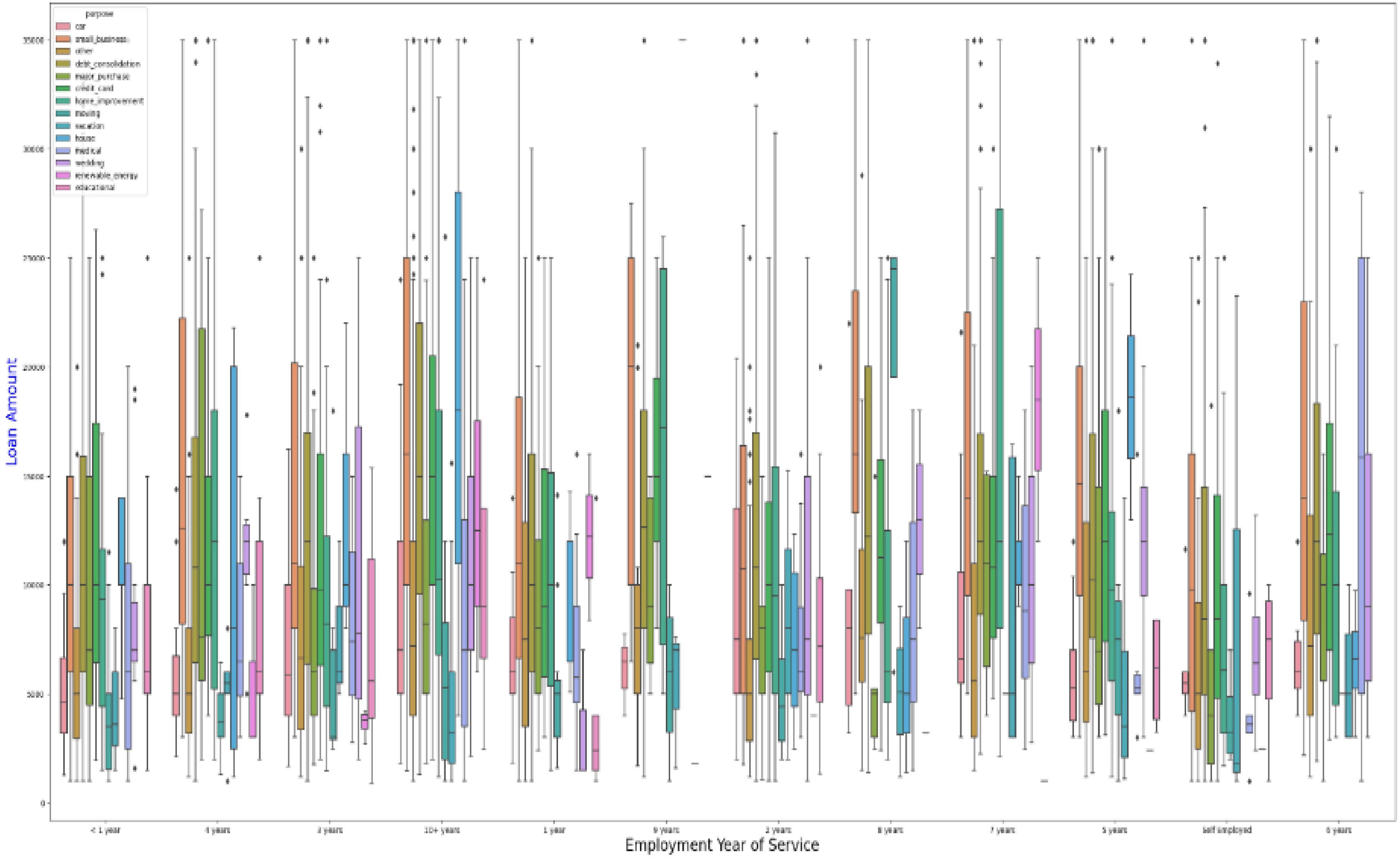
Visualizing Insights: Bivariate Analysis



Analyze loan status and employee’s year of service/experience.

Inference : *Maximum loan is taken by 10+ years employees, approximately 18%.*

Visualizing Insights: Bivariate Analysis

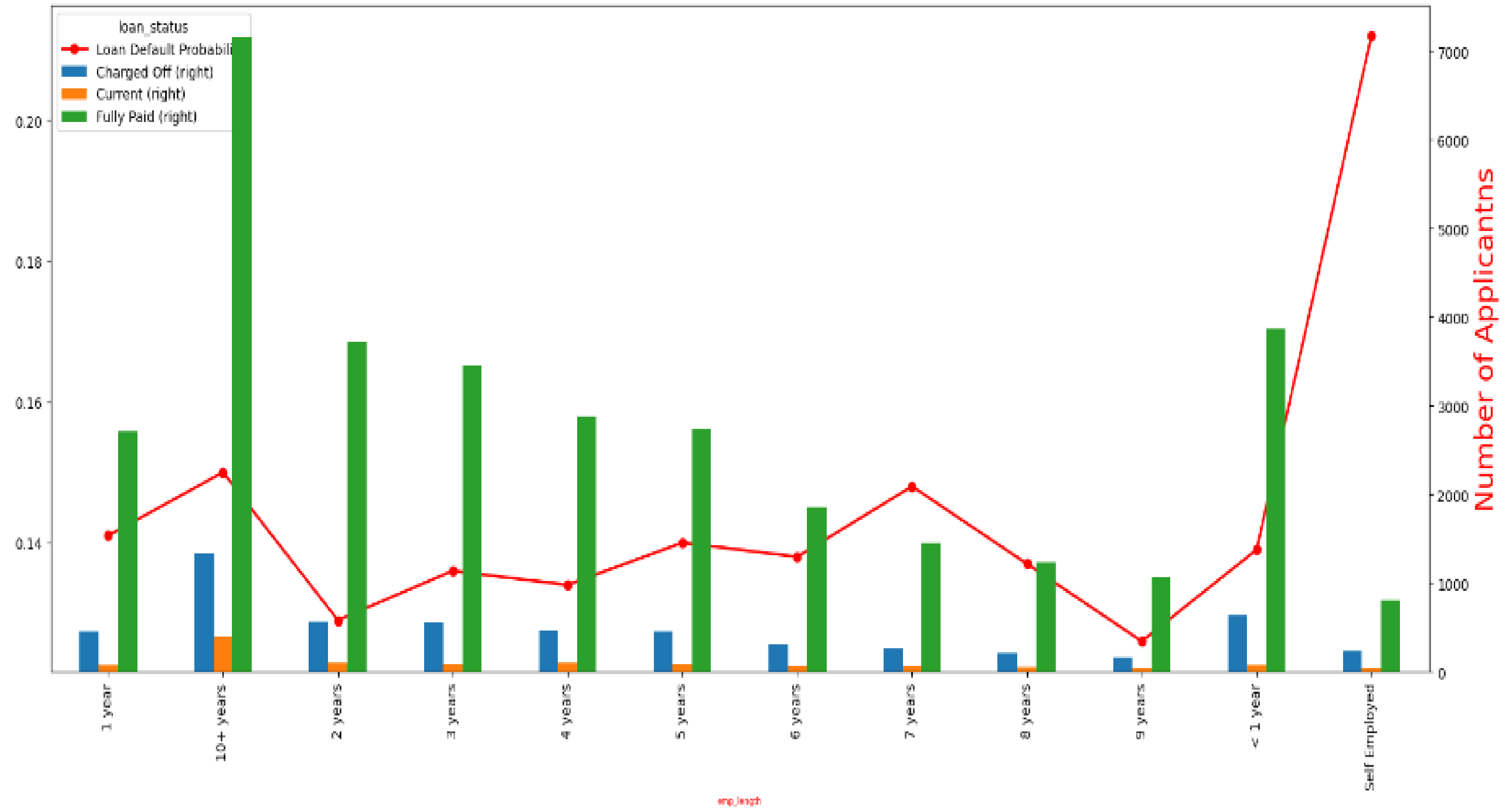


Deriving relationship between employee's years of experience and purpose of loan

<u>Inference</u> : *Employees with 10+ years of experience have taken loan in 'small business' , 'other, 'car', credit card' categories.*

Visualizing Insights: Bivariate Analysis

Deriving defaulter probability for employment year of service



Finding out factors for loan defaulters, in other words, we are trying to find derived variables. Following factors, we assume will be good for analysis :

1. Loan status
2. Charged Off
3. Employment Year of Service
4. Employment Grades
5. Annual Income

<u>Inference</u> : *Applicants who are Self Employed and having less than 1 year of experience are more likely to be Defaulter.*

Visualizing Insights: Bivariate Analysis

Analyze defaulter probability for grades and sub-grades

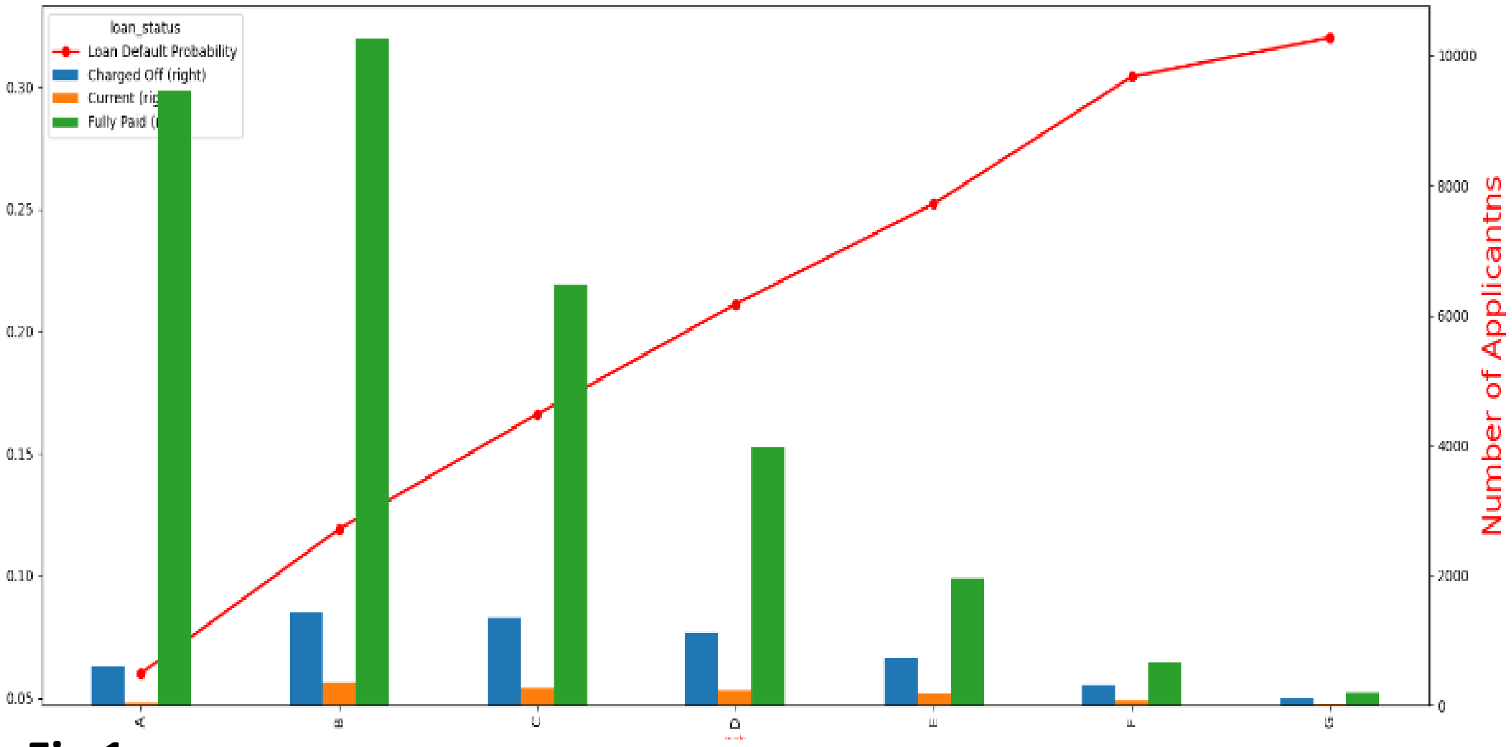
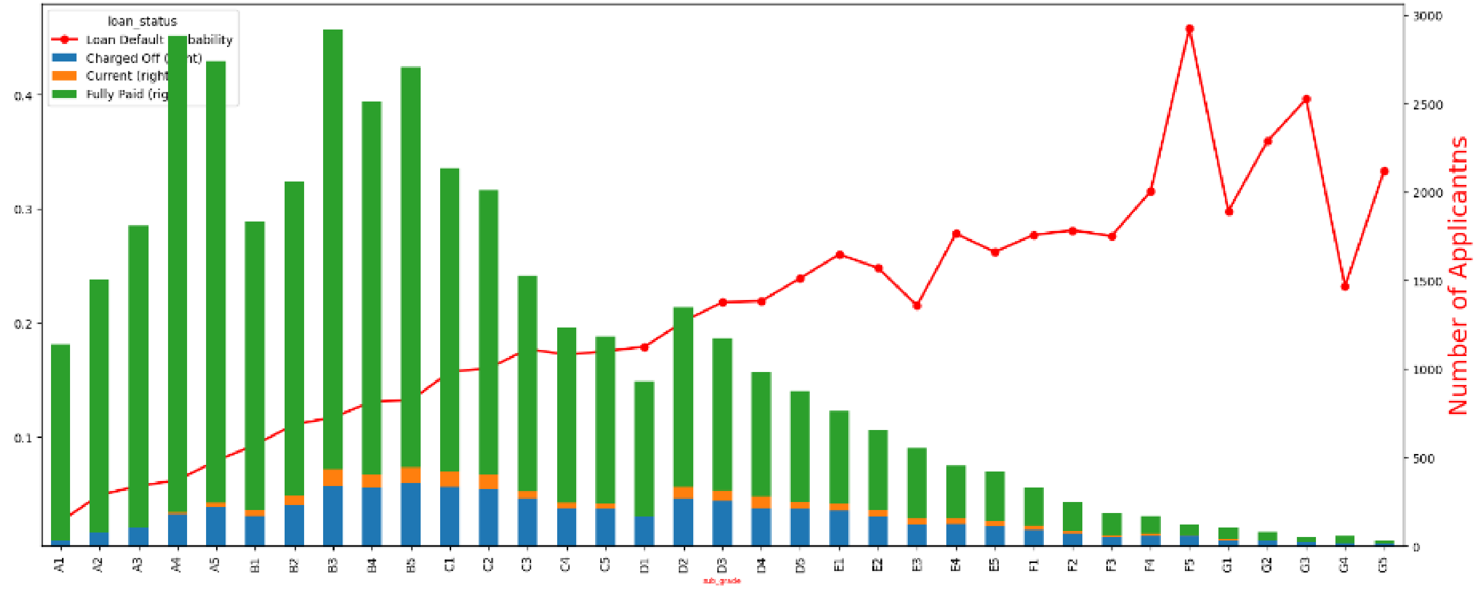


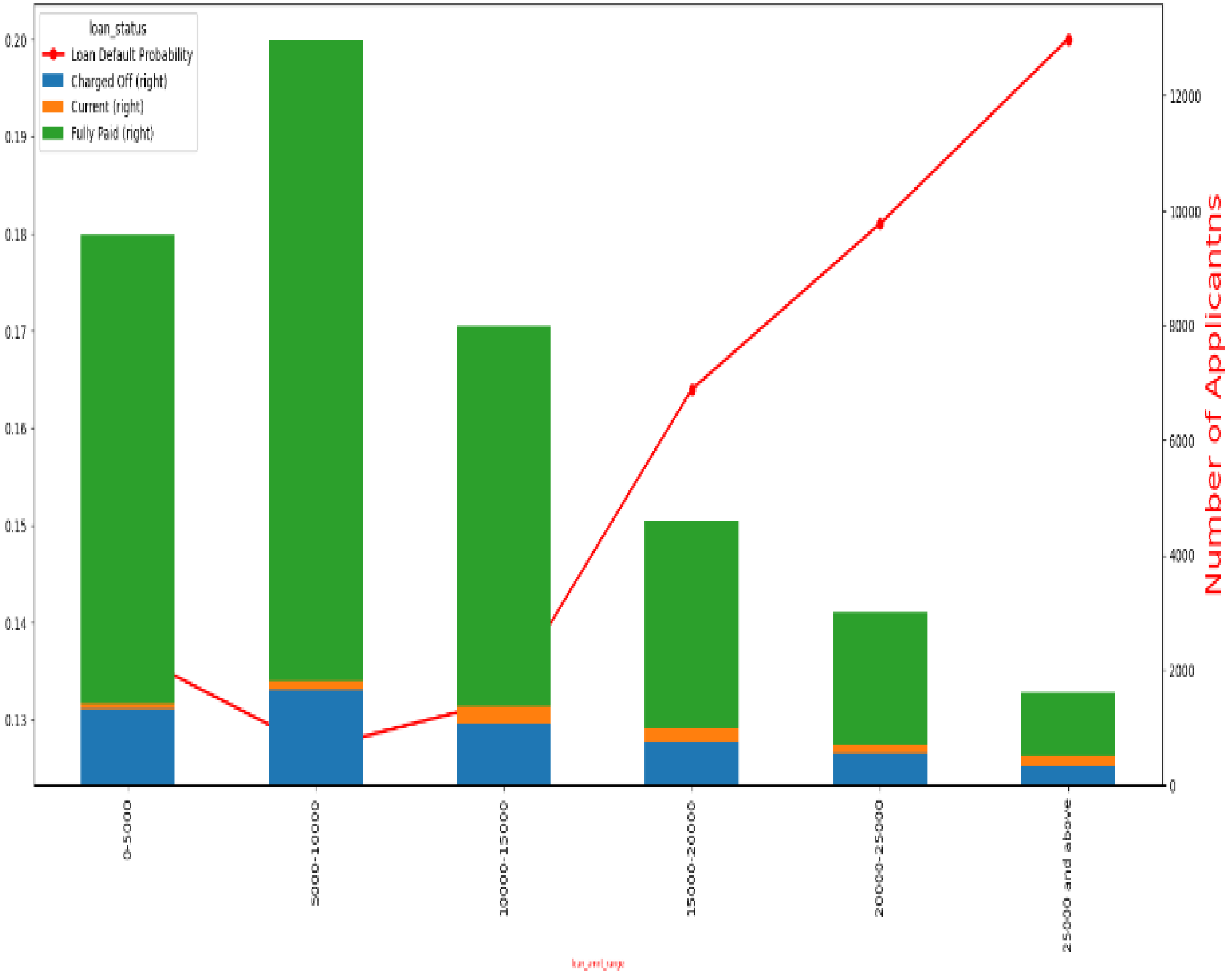
Fig 1: **Grade vs Number of Applications**

Fig 2: **Sub Grade vs Number of Applications**



Inference : *From Grade A to G defaulter probability is increasing*

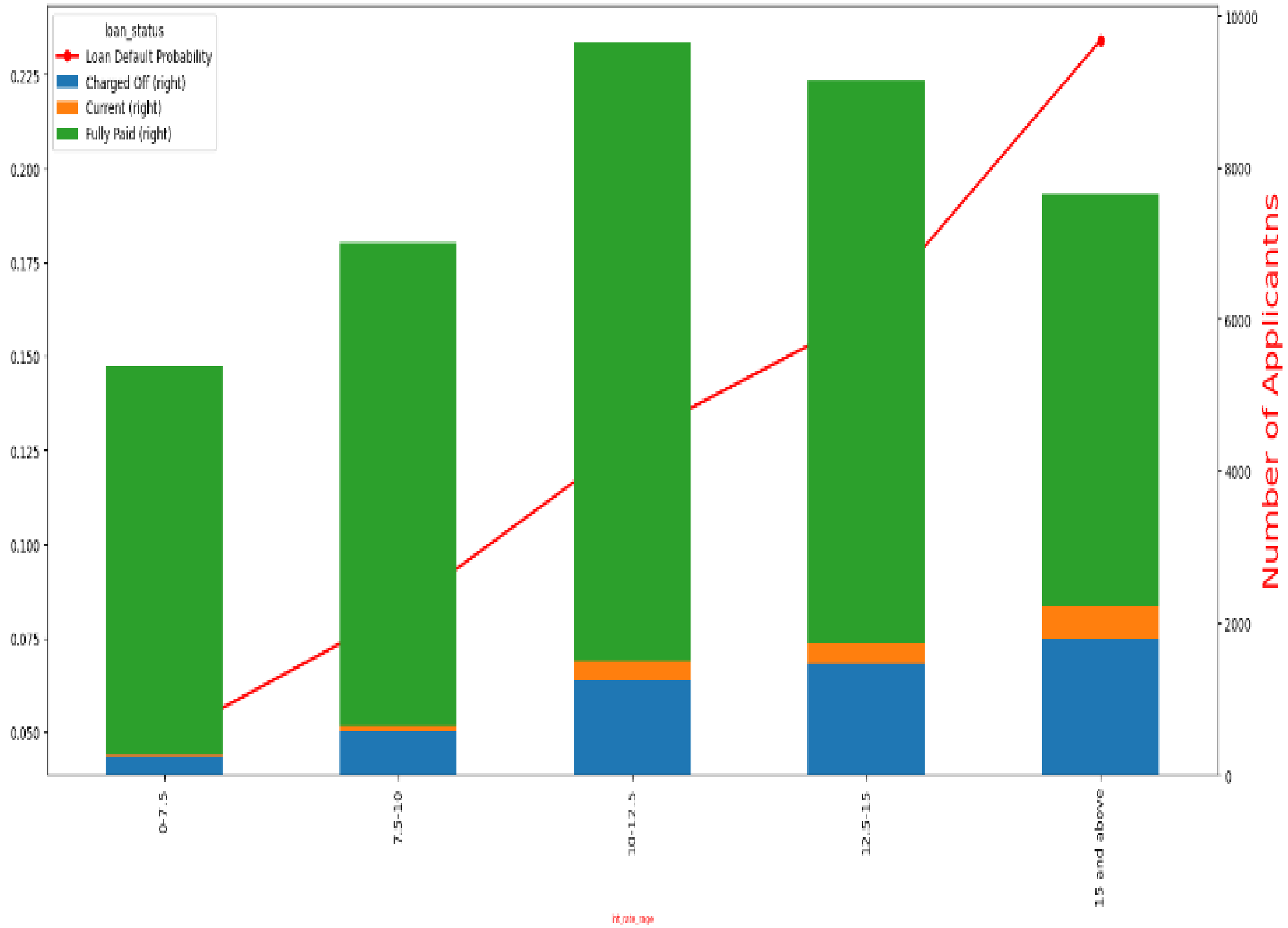
Visualizing Insights: Bivariate Analysis



Analyze defaulter probability based on Loan Amount

Inference : *From the graph, we can say that when loan amount is increasing defaulter rate is increasing so bank should pay attention for higher loan amount*

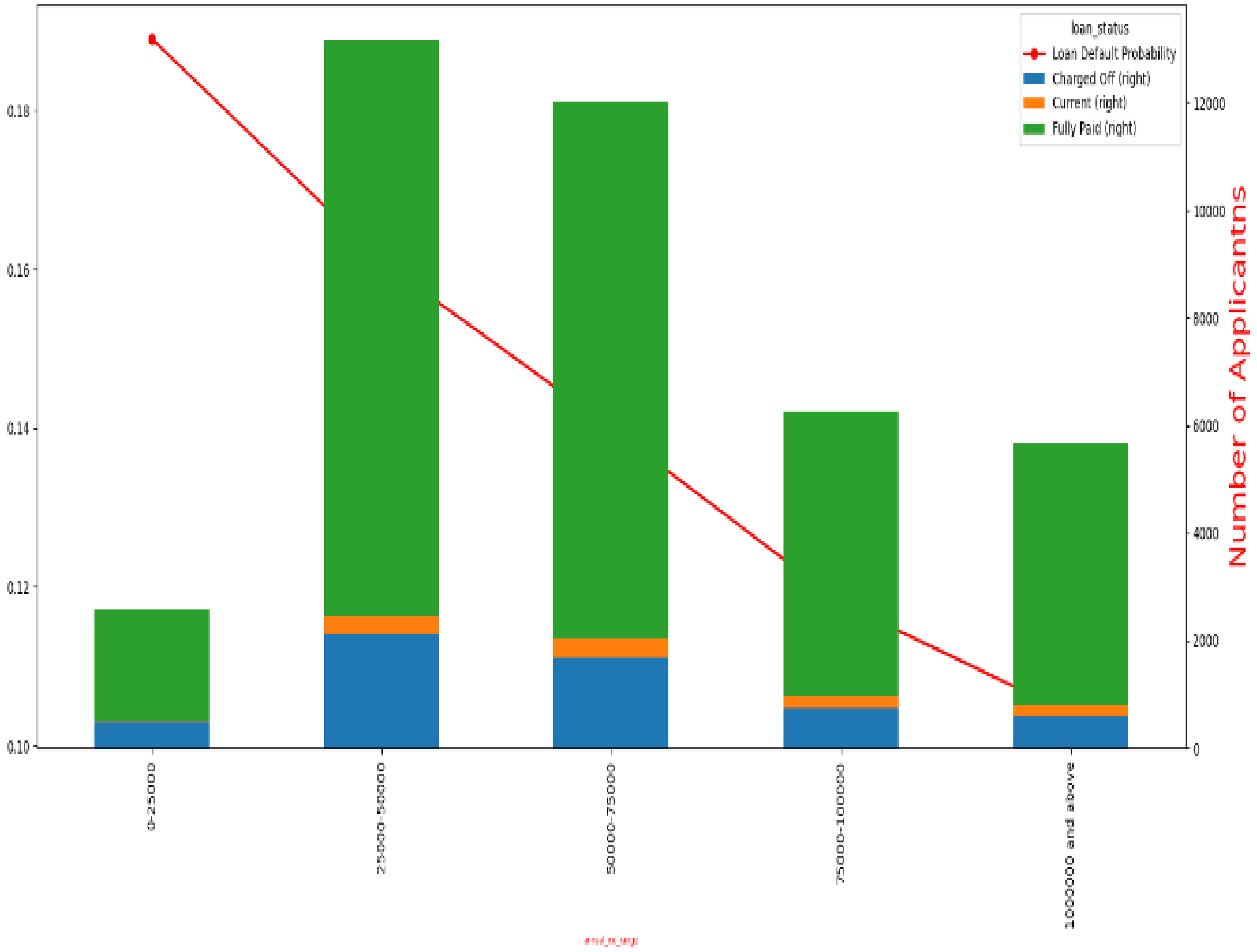
Visualizing Insights: Bivariate Analysis



Analyze defaulter probability based on Interest Rate

<u>**Inference**</u> **:** *As the interest rate is increasing the defaulter rate, so banker should be cautious on this front.*

Visualizing Insights: Bivariate Analysis



Analyze defaulter probability based on Annual Income

Inference : *As annual income is decreasing loan defaulter is increasing with highest of 19% approx. (annual income range - 0-25000)*

Conclusion: Unlocking the Power of Data

Through this comprehensive guide, we have explored the essential steps in data understanding. By correctly identifying data quality issues, conducting thorough analysis, and presenting actionable recommendations, this analysis empower organizations to harness the power of data to drive success and growth.