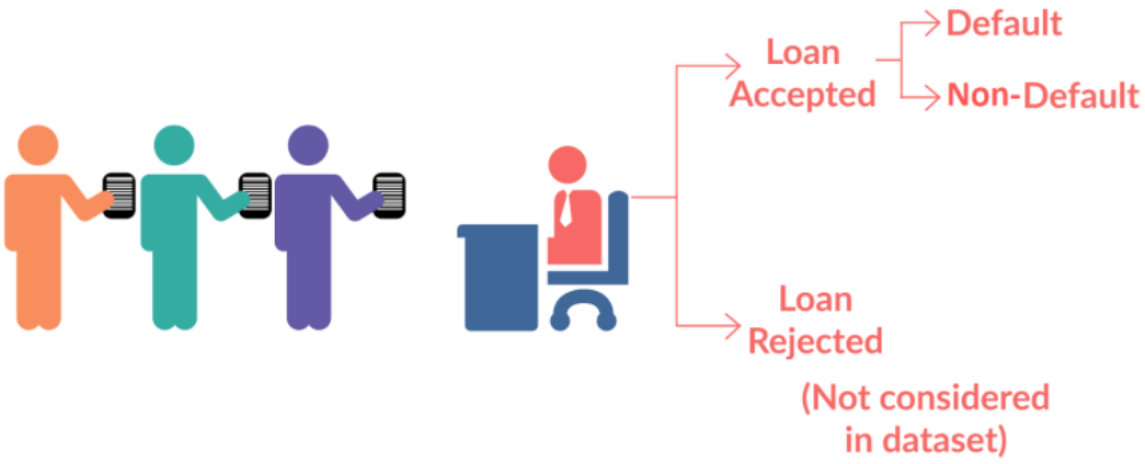
Lending Club Case Study

Business Objective: To analyse the Customer loan data and understand how **consumer attributes** and **loan attributes** influence the tendency of default which will help the company to reassess its portfolio and risk assessment.



Steps Involved:

- Loading the loan Dataset onto python platform (jupyter notebook).
- Sanity Check of the Data. Checking for any errors in particular rows or columns , Summary statistics , Missing values in data.
- Treating Missing values .Checking for outliers and treating them.
- Deriving new columns from the existing columns by process of binning.
- Performing EDA on the Data to extract key insights to satisfy our business objective. This process is divided into three parts univariate , segmented univariate and bivariate analysis.

Data:

- The loan dataset is divided into 111 columns and 39717 rows. The columns are divided into 74 float ,24 object, 13 integer type variables.

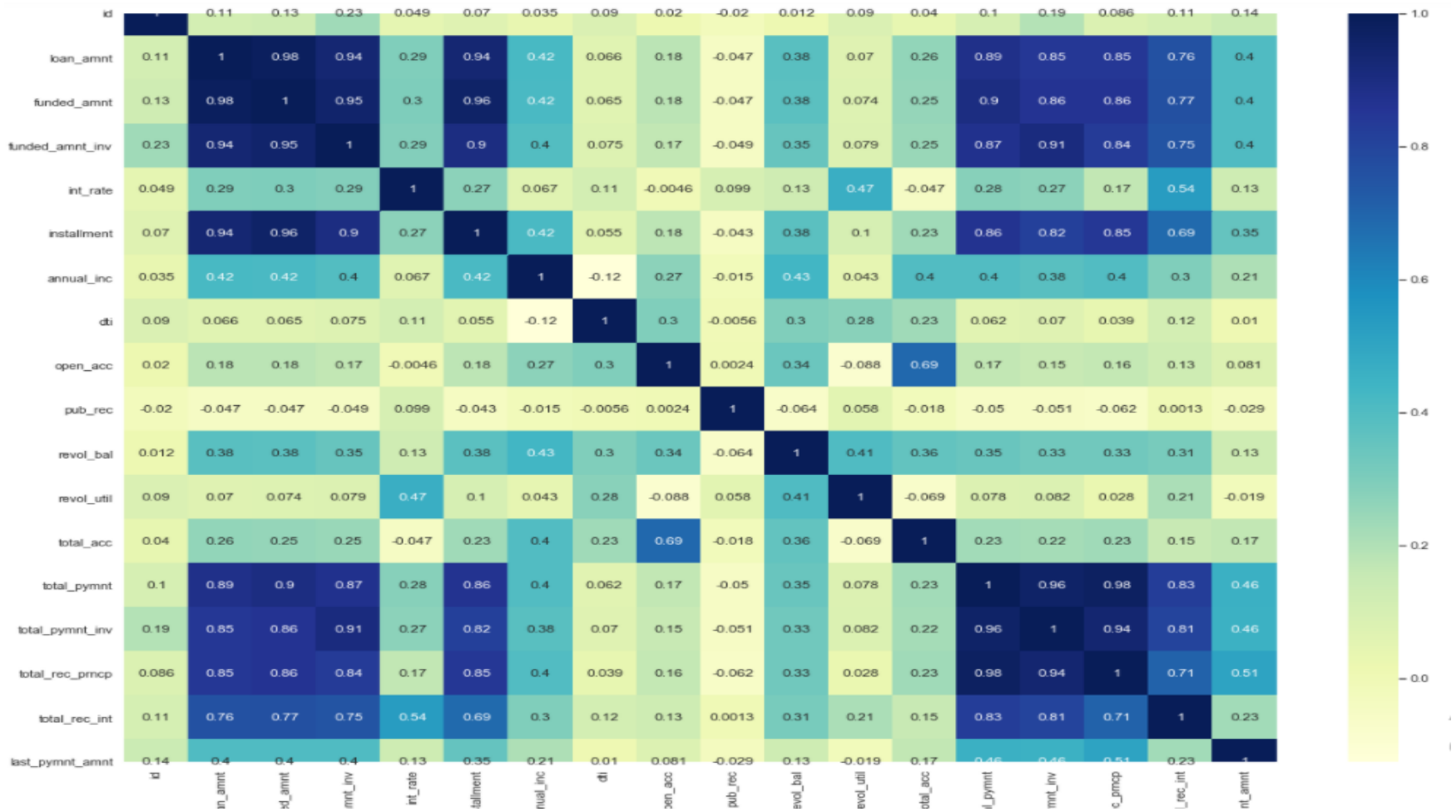
	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	installment	grade	sub_grade	emp_title	emp_length
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	162.87	B	B2	NaN	10+ years
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	59.83	C	C4	Ryder	< 1 year
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	84.33	C	C5	NaN	10+ years
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	339.31	C	C1	AIR RESOURCES BOARD	10+ years
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	67.79	B	B5	University Medical Group	1 year

Data Cleaning and Missing value imputation:

- Large no. of columns consist of missing values . Most of the columns have 100% missing values .
- Missing value imputation has been carried out such that columns having missing values more than 30% have been dropped.
- Row wise dropping has been done for columns having less than 10% missing values.
- imputing missing values in columns pub_rec_bankruptcies and last_pymnt_d with mode.

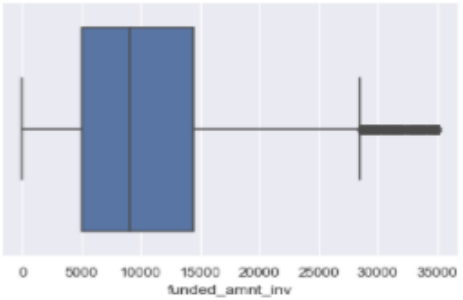
last_credit_pull_d	2
collections_12_mths_ex_med	56
mths_since_last_major_derog	39717
policy_code	0
application_type	0
annual_inc_joint	39717
dti_joint	39717
verification_status_joint	39717
acc_now_delinq	0
tot_coll_amt	39717
tot_cur_bal	39717
open_acc_6m	39717
open_il_6m	39717
open_il_12m	39717
open_il_24m	39717
mths_since_rcnt_il	39717
total_bal_il	39717
il_util	39717
open_rv_12m	39717
open_rv_24m	39717
max_bal_bc	39717
all_util	39717
total_rev_hi_lim	39717
inq_fi	39717
total_cu_tl	39717
inq_last_12m	39717
acc_open_past_24mths	39717
avg_cur_bal	39717
bc_open_to_buy	39717
bc_util	39717
chargeoff_within_12_mths	56
delinq_amnt	0
mo_sin_old_il_acct	39717
mo_sin_old_rev_tl_op	39717
mo_sin_rcnt_rev_tl_op	39717

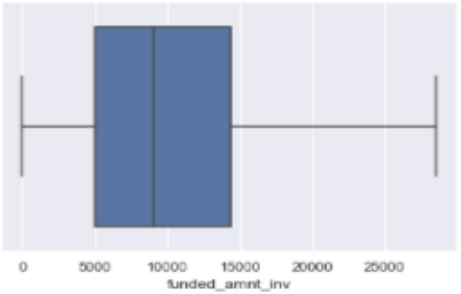
- Multicollinearity is checked by plotting heat map and such columns are dropped of.
- Columns which have a single category/value throughout are also dropped .

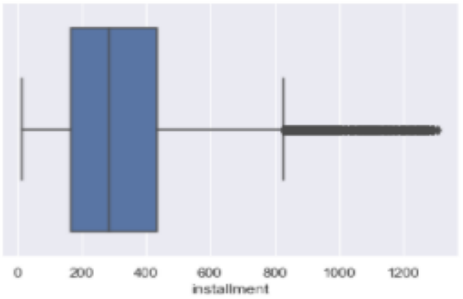


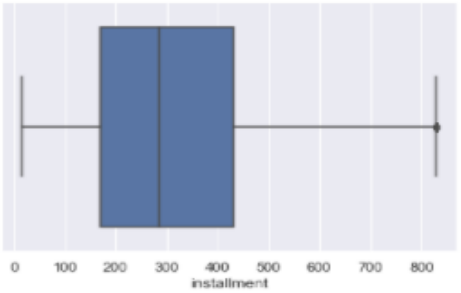
Outlier check and Treatment:

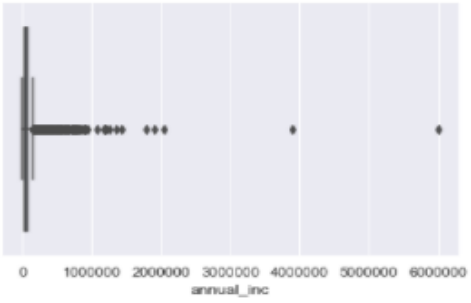
- Outlier Check is done on numeric continuous variables by plotting Box plots.
- Outliers have been Treated by find out Upper capping limit for each variable and capping them with this value.

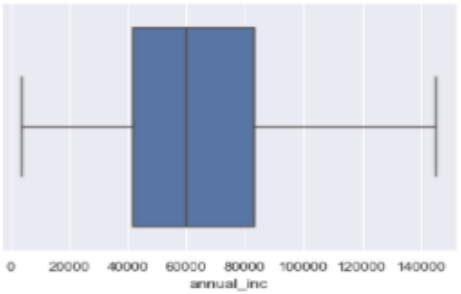












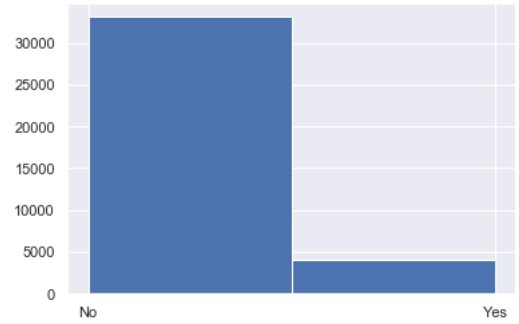
Before Capping

After Capping

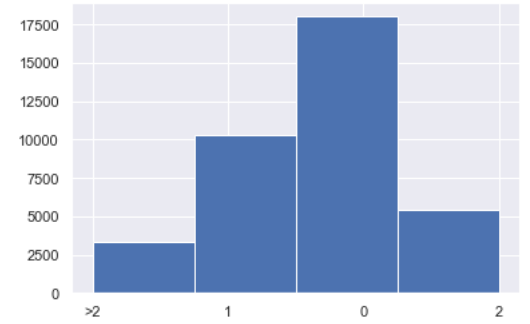
Binning of Columns(Derived Metrics):

- Binning has been done as follows:
- Binning of Delinquency in 2 years column to category no, yes.
- Binning of inq_last_6mths column (Value counts before and after) binned to category 0,1,2,>2.
- As outstanding pr and outstanding pr_inv are highly +vely correlated we form a column which is addition of both = total_out_principal.
- Binning of total_out_principal column (Value counts before and after) binned to category 0,>0.

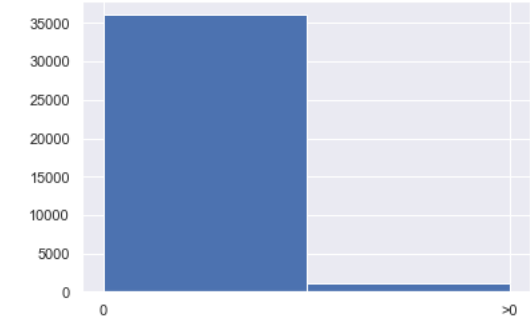
```
Name: delinq_2yrs, dtype: int64
```




```
Name: inq_last_6mths, dtype: int64
```

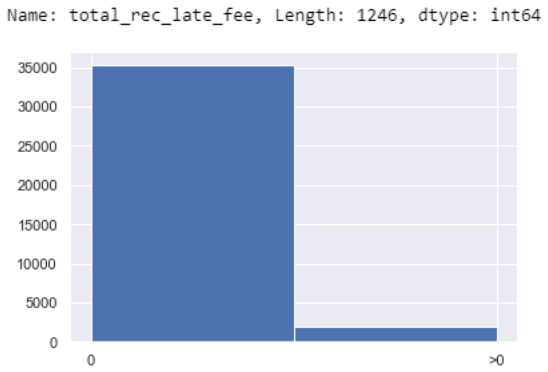


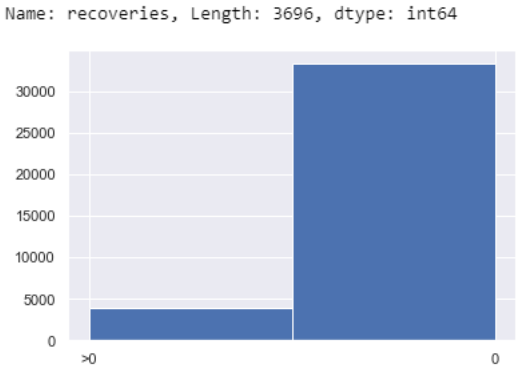

```
Name: total_out_principal, Length: 1066, dtype: int64
```

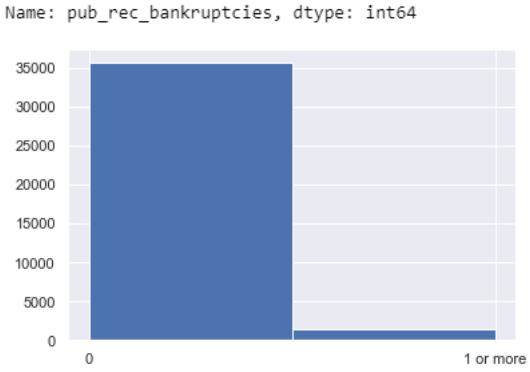


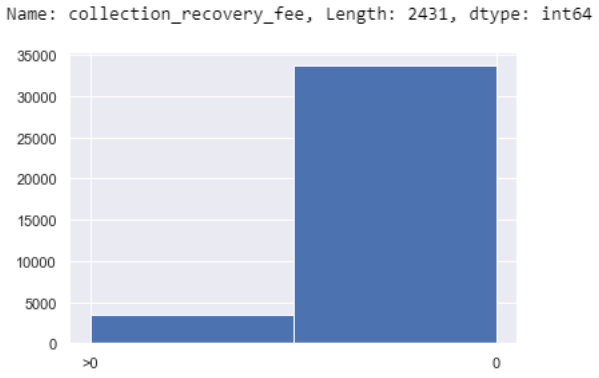
- Binning of total_rec_late_fee column binned to category 0,>0.
- Binning of recoveries column binned to category 0,>0.
- Binning of collection_recovery_fee column binned to category 0,>0.
- Binning of pub_rec_bankruptcies column (Value counts before and after) binned to category 0,>0.

- Apart from that columns: annual income , loan amount , DTI and interest rate are also binned to help perform Bivariate analysis.



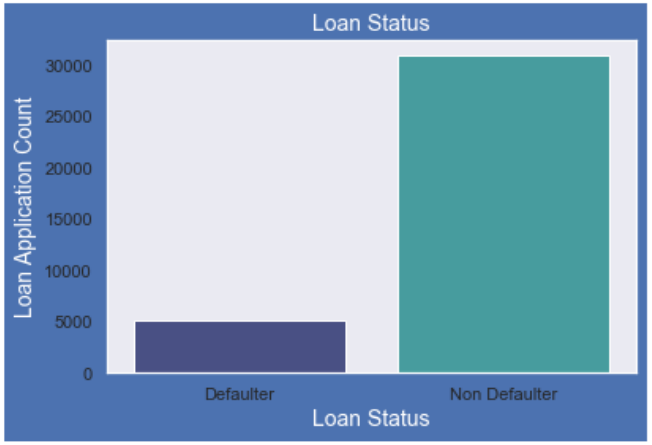


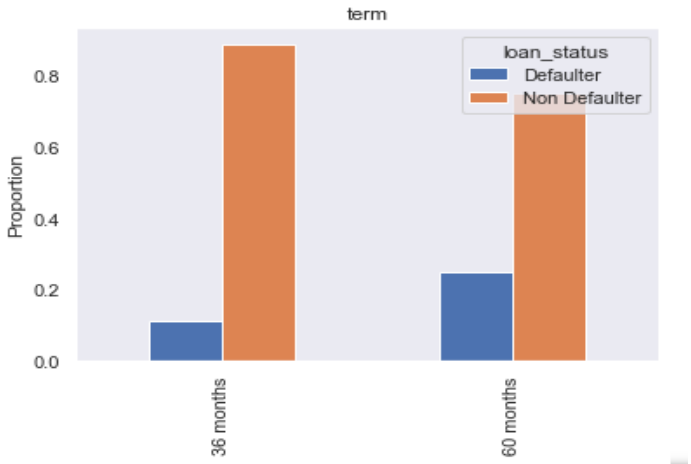


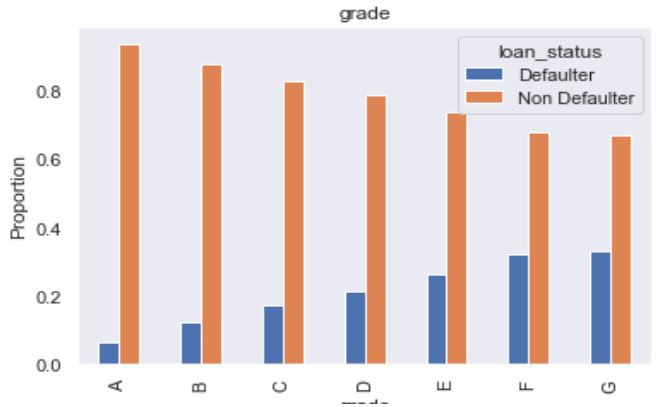


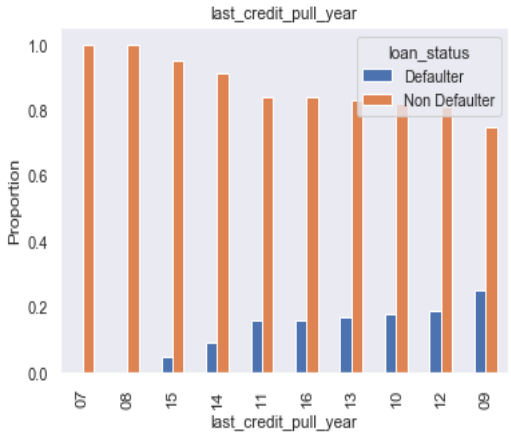
Univariate analysis:

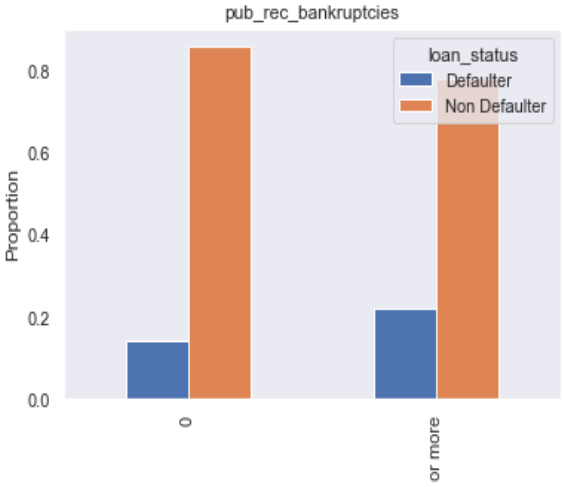
- Loan status plot shows that close to 14% loans were Defaulters out of total loan issued.
- Customer who have taken loan term of 60 months have defaulted the most.
- Customers belonging to the E,F,G grade consisted of most defaulters.

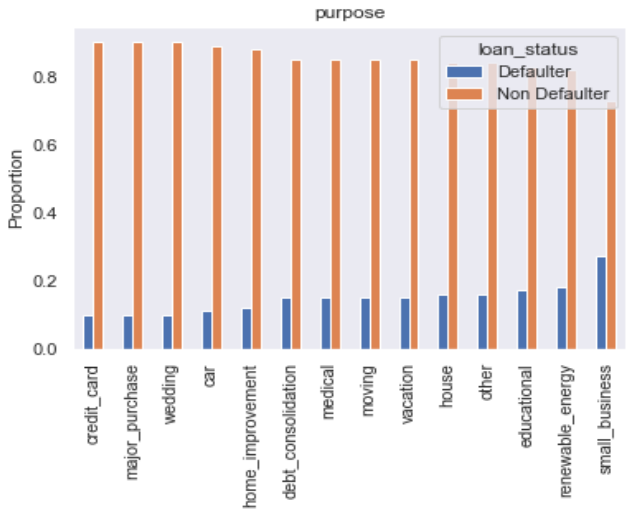


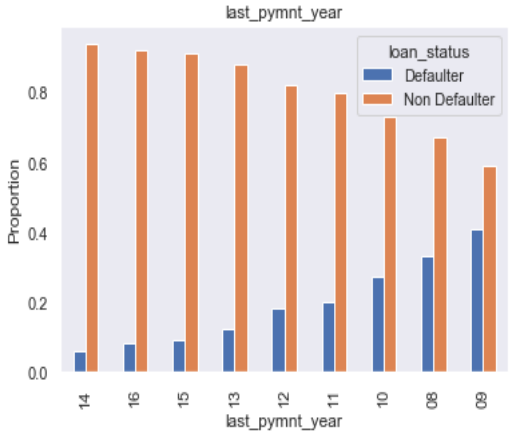


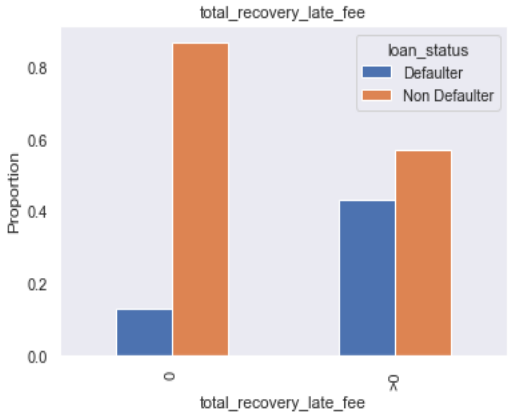


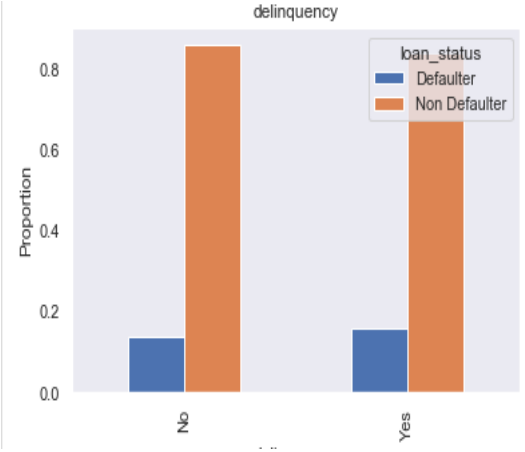




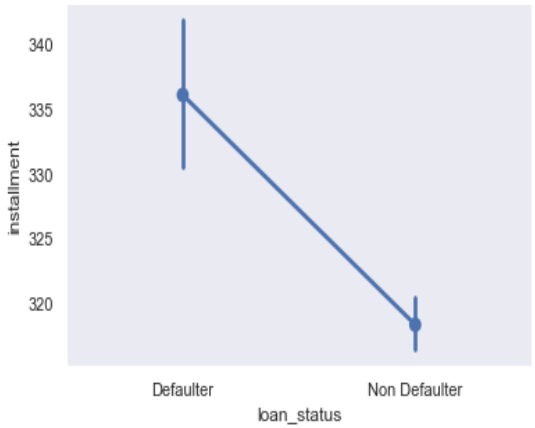


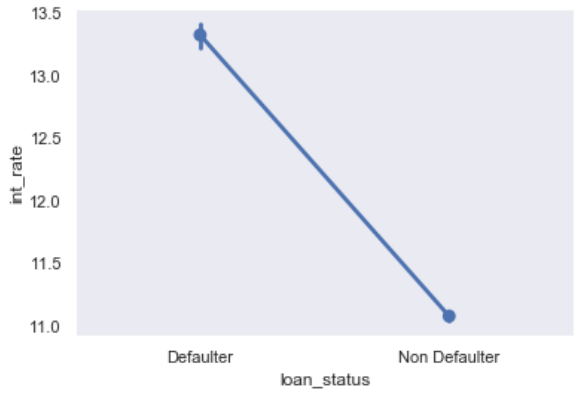


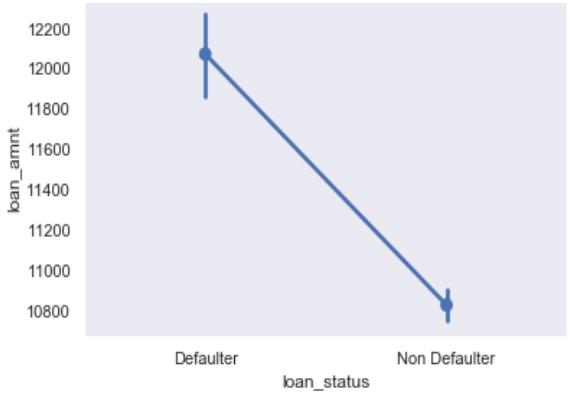


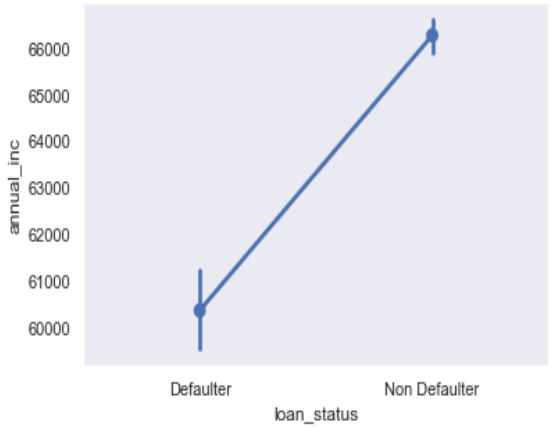


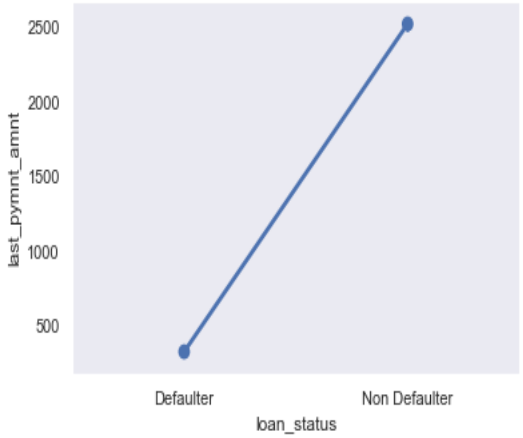
- Customer taking loan for small businesses are more likely to default.
- Customer who have done there last credit pull and last payment in year 2009 have defaulted the most.
- Customers having bankrupt records are mostly defaulters.
- Customer which have paid late recovery fees and have been delinquent (>30 days) are most likely to default.
- * Here cross tabs depict the proportion of defaulters and non defaulters in each category. These are not the counts.

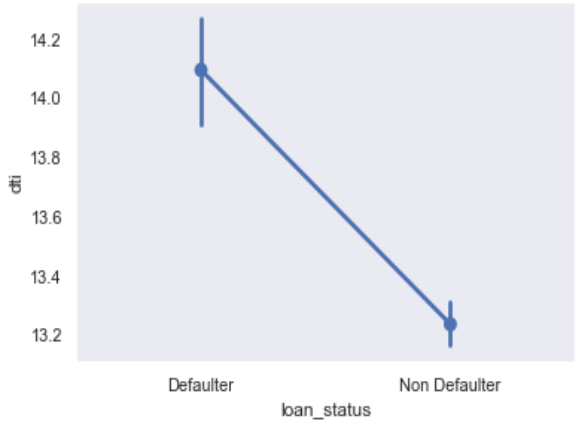








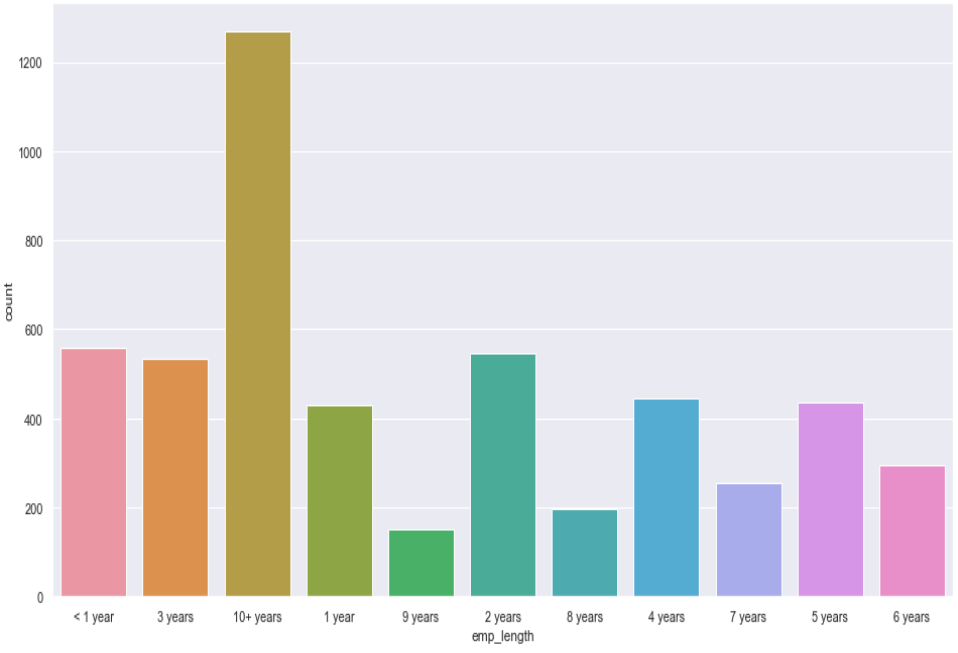


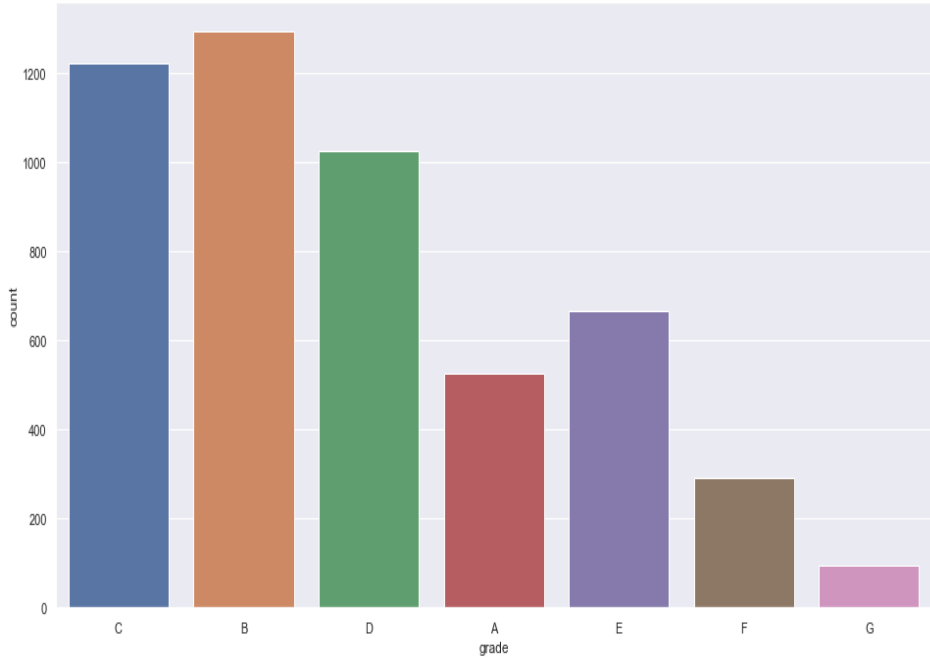


- From the above point plots we observe: Customers who have taken a higher loan amount are defaulting more.
- Customers having low annual income are the one's defaulting.
- Customer having higher interest rates on their loan amount are the one's defaulting. Customers who have to pay higher instalment are defaulting more.
- Customers having high DTI ratio which is debt payment to income ratio are the ones defaulting.
- * the points depicted in the plots are mean values of variables and the vertical lines are the std dev from mean.

Segmented Univariate analysis:

- Here we have segmented the data on Loan Status so now we have 2 data frames 1st of defaulters and 2nd of non defaulters.
- We will study each variable in defaulter category.
- Here we can see the counts of defaulters in each category in a column variable.
- Customers who are employed for 10+ years have high no of defaulters.
- In the second graph we see that customers belonging to C,B,D grade have high no of defaulters although the proportion of defaulter to non defaulter is high in the G grade as seen earlier.





Defaulters:

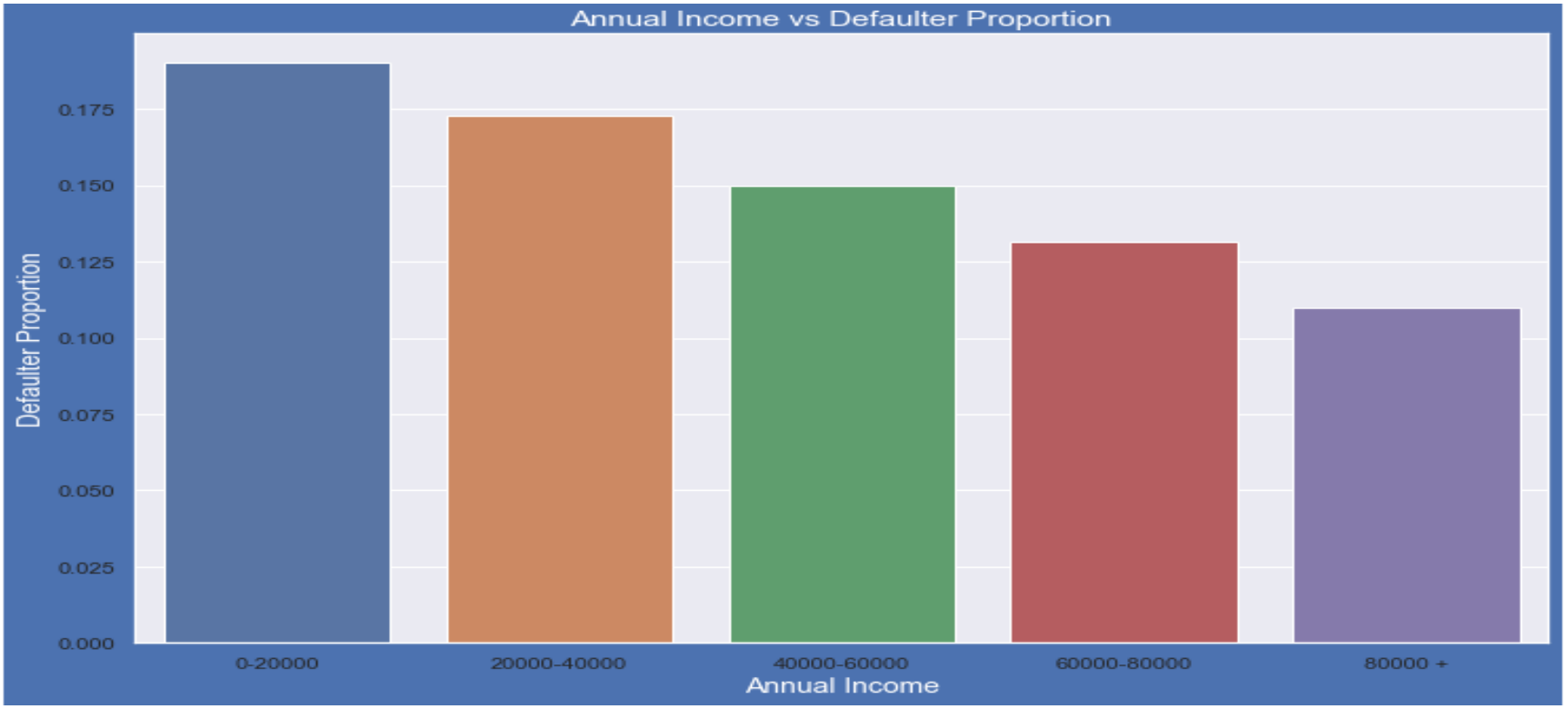
	id	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	open_acc	pub_rec
count	5.115000e+03	5115.000000	5115.000000	5115.000000	5115.000000	5115.000000	5115.000000	5115.000000	5115.000000	5115.000000
mean	6.981147e+05	12071.911046	11754.139785	10828.598363	13.319844	336.143760	60371.714909	14.095406	9.228739	0.082307
std	2.175044e+05	7595.594790	7343.967032	7226.801610	3.676622	206.096011	30663.044927	6.541546	4.301103	0.281883
min	6.141900e+04	900.000000	900.000000	0.000000	5.000000	22.790000	4080.000000	0.000000	2.000000	0.000000
25%	5.308660e+05	6000.000000	6000.000000	5000.000000	11.000000	171.635000	38398.500000	9.240000	6.000000	0.000000
50%	7.009680e+05	10000.000000	10000.000000	9600.000000	13.000000	299.040000	54000.000000	14.400000	9.000000	0.000000
75%	8.560645e+05	17000.000000	16000.000000	15000.000000	16.000000	461.830000	75000.000000	19.300000	12.000000	0.000000
max	1.077430e+06	29100.000000	29250.000000	28500.000000	24.000000	828.000000	145100.000000	29.850000	21.000000	2.000000

Non Defaulters:

	id	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	open_acc	pub_rec
count	3.096600e+04	30966.000000	30966.000000	30966.000000	30966.000000	30966.000000	30966.000000	30966.000000	30966.000000	30966.000000
mean	6.720510e+05	10825.494898	10595.201996	10076.358091	11.069851	318.360026	66280.027101	13.236912	9.291772	0.049151
std	2.077065e+05	6878.142402	6682.712013	6612.251948	3.626828	196.150032	32441.755690	6.654275	4.214416	0.226116
min	5.574200e+04	500.000000	500.000000	0.000000	5.000000	16.080000	4000.000000	0.000000	2.000000	0.000000
25%	5.087822e+05	5500.000000	5375.000000	5000.000000	8.000000	167.080000	42000.000000	8.110000	6.000000	0.000000
50%	6.497385e+05	9600.000000	9500.000000	8775.000000	11.000000	278.230000	60000.000000	13.300000	9.000000	0.000000
75%	8.225218e+05	15000.000000	14600.000000	13975.000000	13.000000	421.875000	84000.000000	18.480000	12.000000	0.000000
max	1.076863e+06	29100.000000	29250.000000	28500.000000	24.000000	828.000000	145100.000000	29.990000	21.000000	4.000000

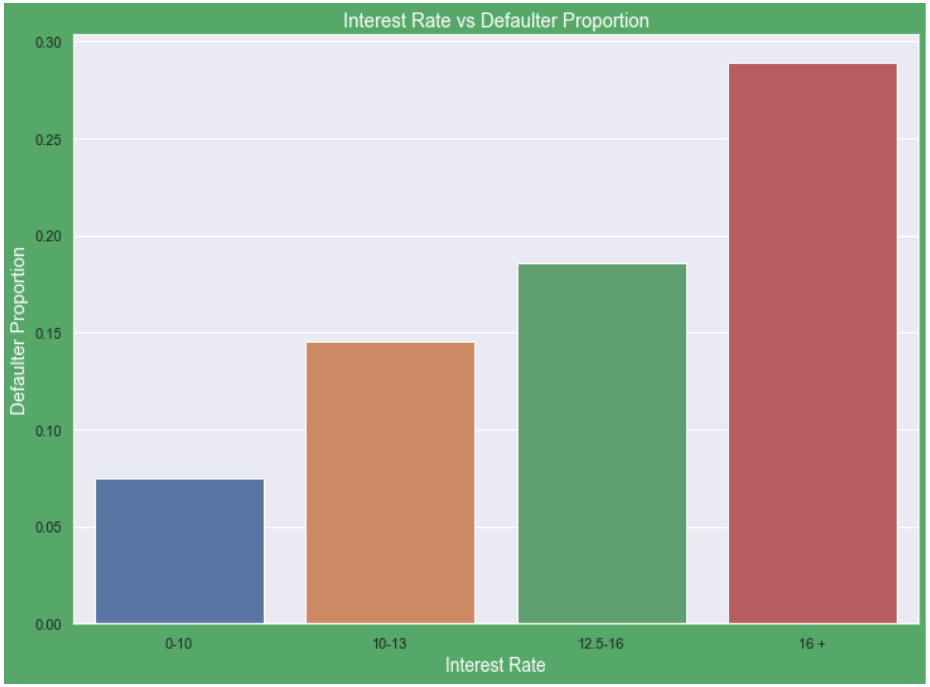
- From the summary stats of the defaulters and non defaulters we can clearly see the differences in mean loan amounts, annual incomes , interest rates, DTI ratios and derogatory records(pub_rec).
- *the mean values depicted would have more difference in scale but as the capping(outlier treatment) of each variable is done and both have same maxima’s so the results are a bit altered.

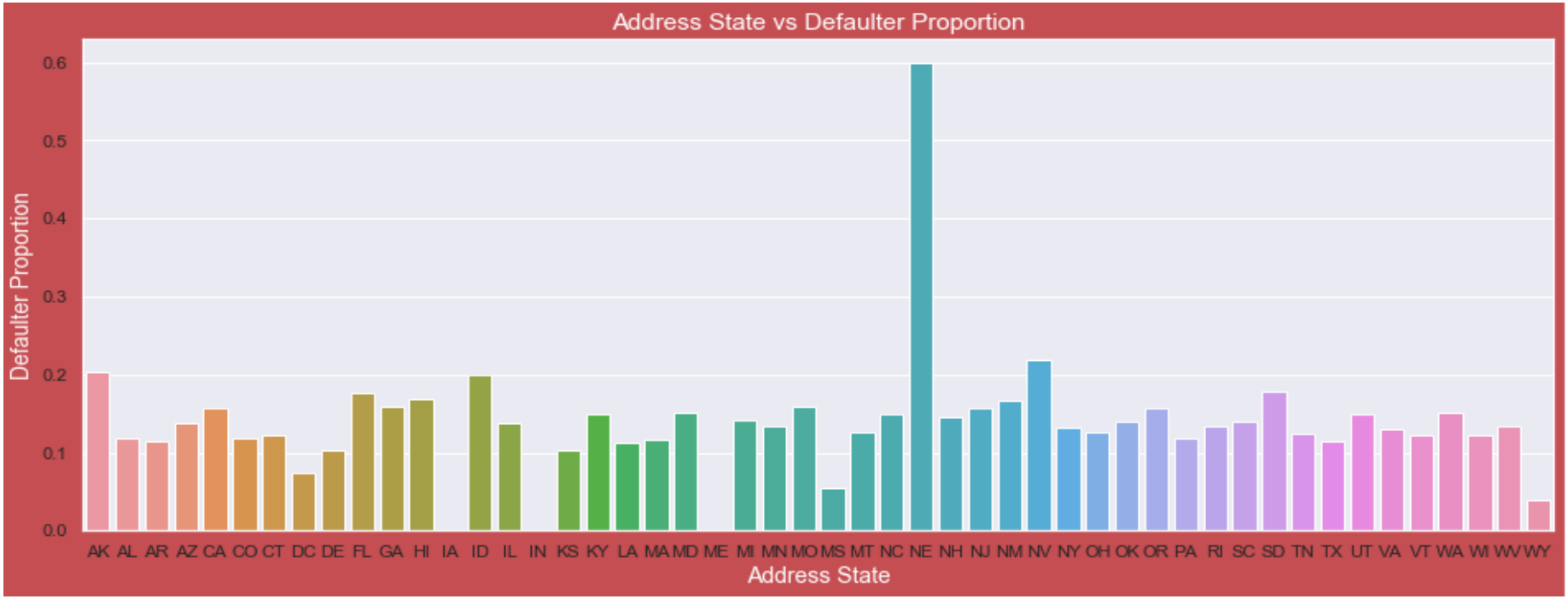
Bivariate Analysis:



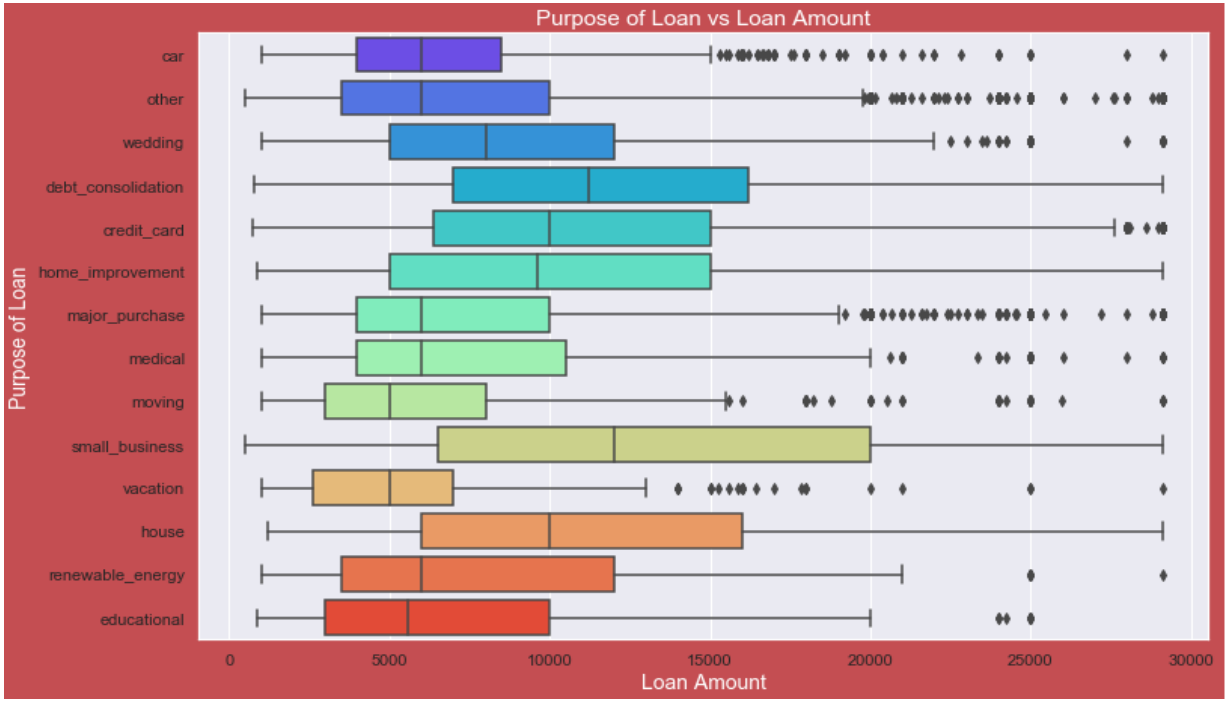
- Income range 80000+ has less chances of Default.
- Income range 0-20000 has high chances of Default.
- Notice that with increase in annual income Defaulter proportion got decreased.

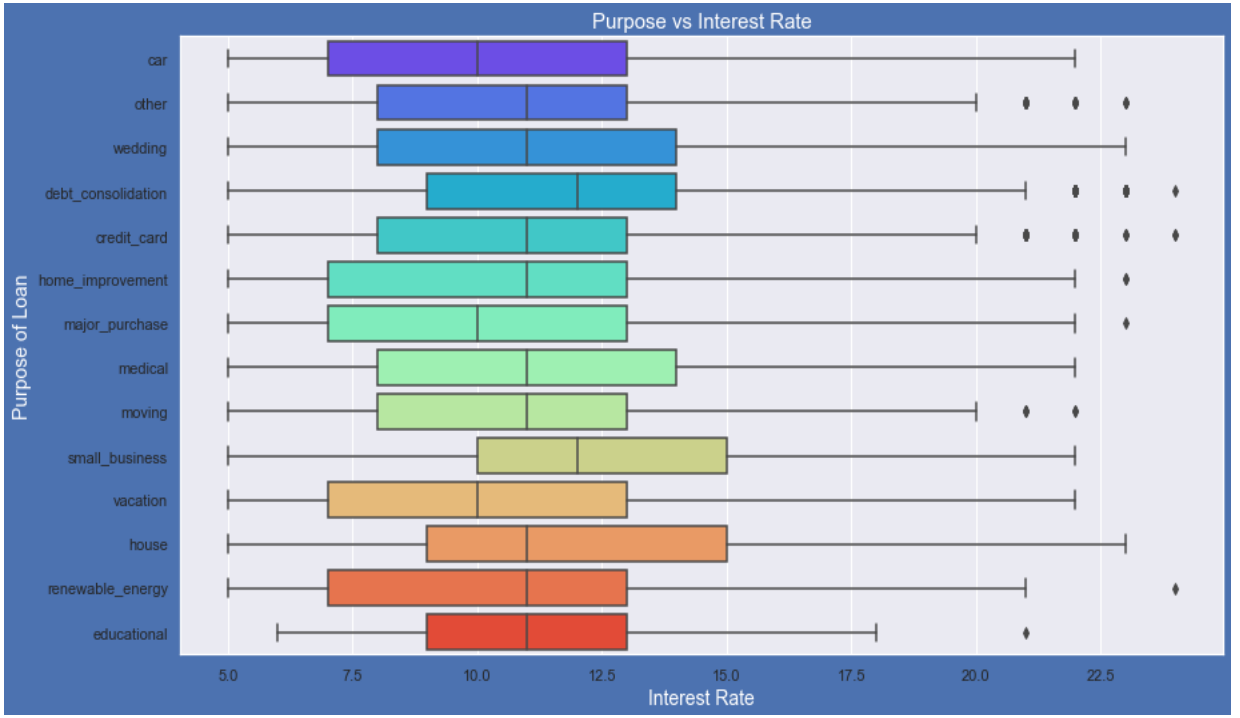
- interest rate less than 10% has very less chances of Default. Interest rates are starting from minimum 5 %.
- interest rate more than 16% has good chances of Default as compared to other category interest rates.
- Defaulter proportion is increasing with higher interest rates.
- states NE has very high chances of Defaulting but number of applications are too low to make any decisions.
- NV,CA and FL states shows good number of Defaulters in good number of applications.



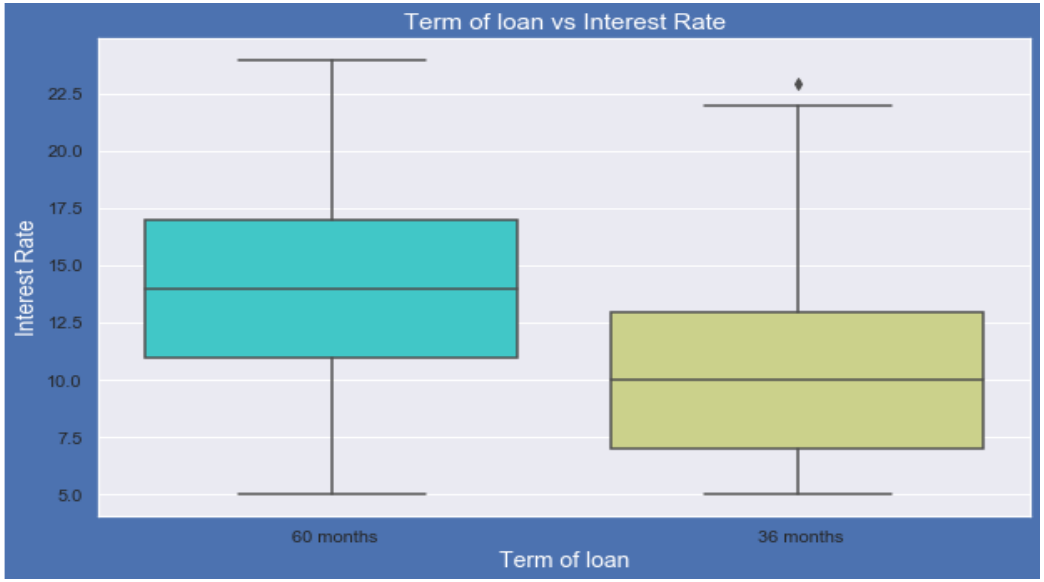


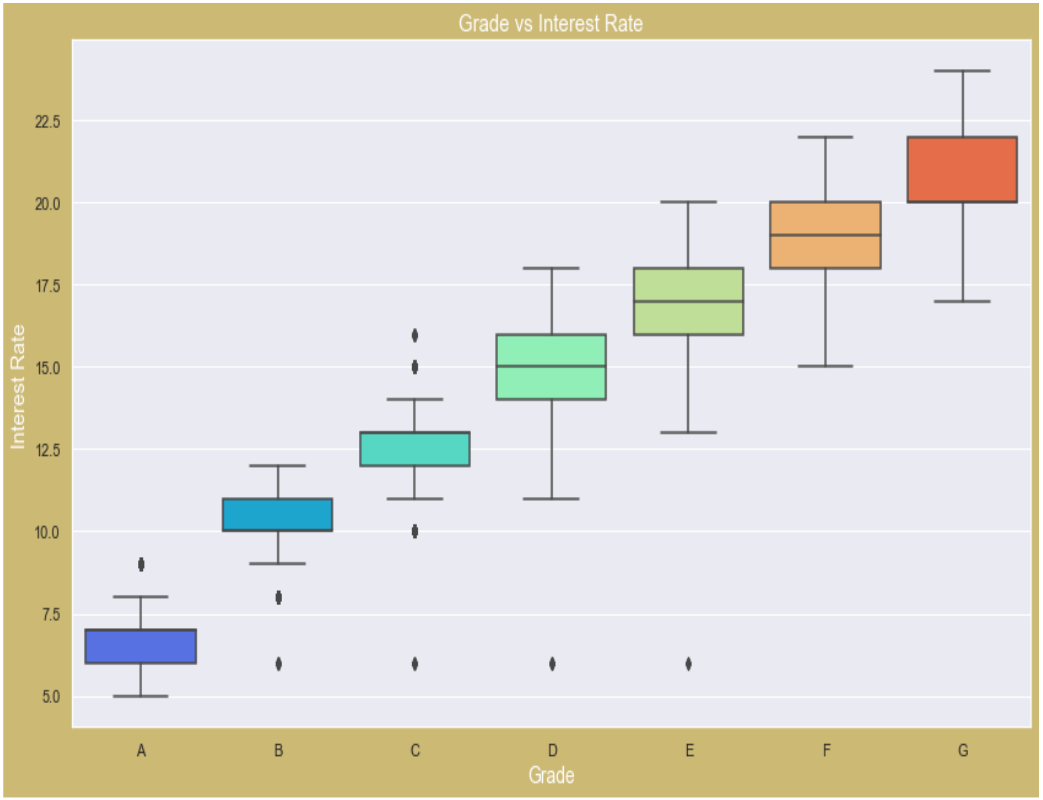
- Median,95th percentile,75th percentile of loan amount is highest for loan taken for small business purpose among all purposes.
- Debt consolidation is second and Credit card comes 3rd.
- It is clear that average interest rate is highest for small business purpose.
- Loans taken for small business purposes had to repay the loan with more interest rate as compared to other.
- Debt consolidation is 2nd where borrowers had to pay more interest rate.



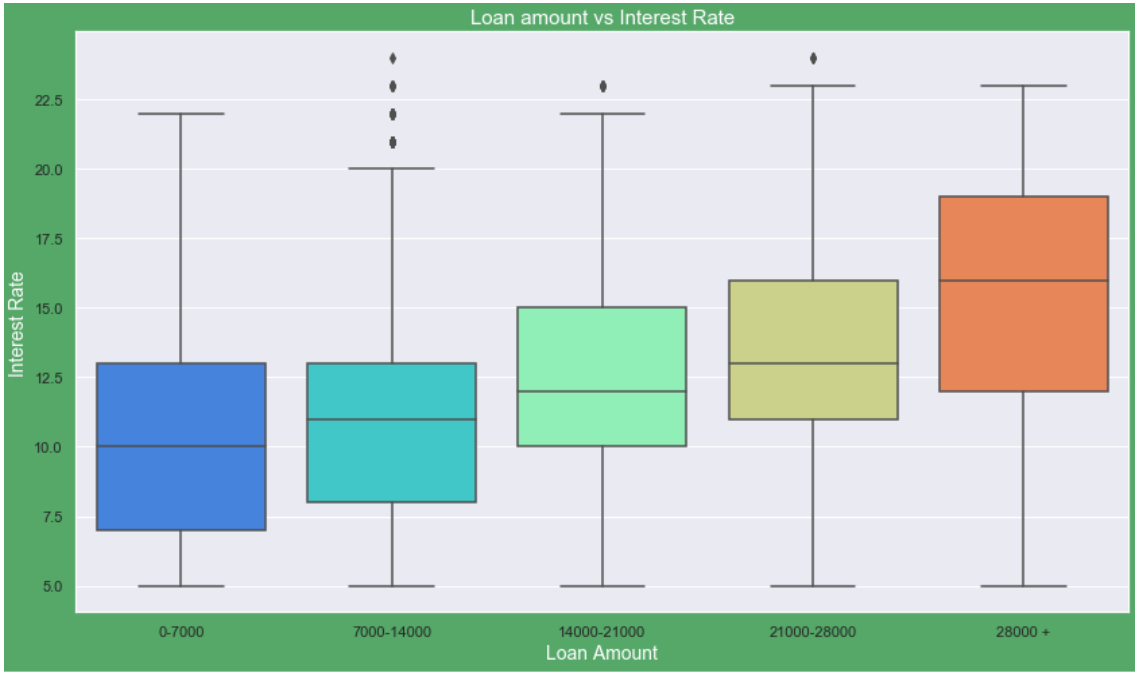


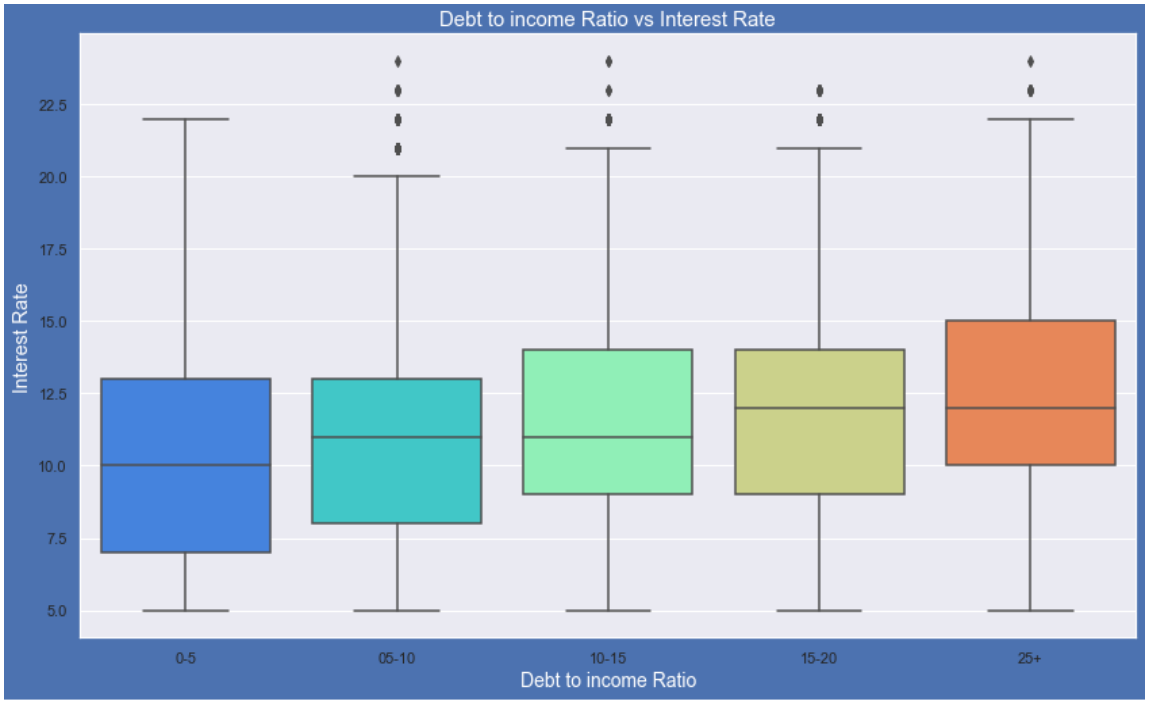
- It is clear that average interest rate is higher for 60 months loan term.
- Most of the loans issued for longer term had higher interest rates for repayment.
- A-grade is a top grade for a lender to assign to a borrower.
- The higher the borrower's credit grade, the lower the interest rate offered to that borrower on a loan.(Grade hierarchy is A to G A being the highest Grade).
- It is clear that interest rate is increasing with grades moving from A to G.

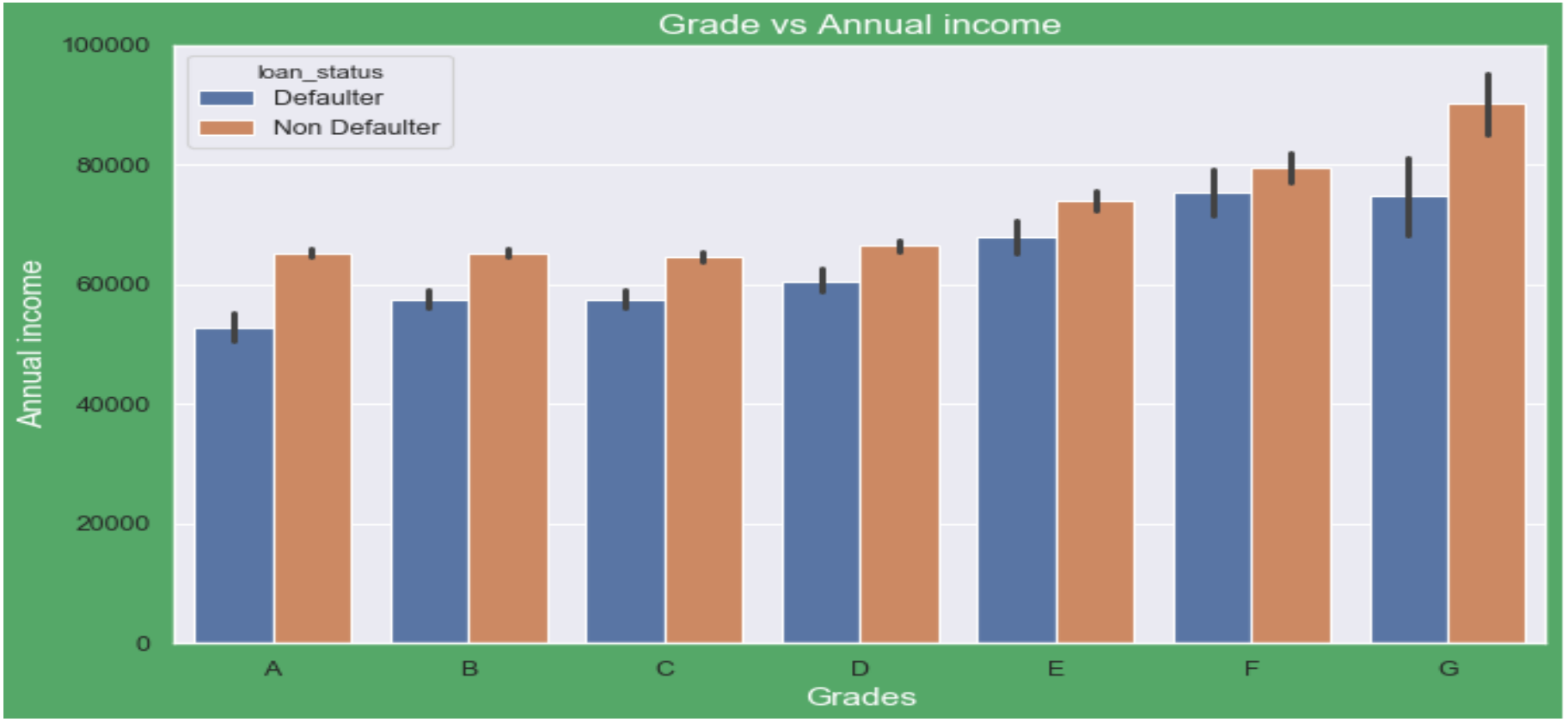




- It is clear that interest rate is increasing with loan amount increase.
- Probably when loan amount is more it is taken for longer loan term, we saw earlier that longer the loan term more the interest rate.
- If your DTI is low enough you may get a lower interest rate.
- Plot shows no significant variation but there is slight increase in interest rate with increase in DTI.







From this we can conclude that the ones getting 'Defaulter' have lower annual incomes than the ones who 'Non Defaulted' for each and every grade (i.e. at same interest range)

Conclusion:

- Customers having low annual income should be thought twice before giving them loan.
- Long term loans with high interest rates is the bracket where most of the defaulters lie , so company should think twice before issuing a long term loan at higher interest rate. Loans with higher interest are almost never paid off according to data.
- Company should give loans with larger amount to only trust worthy customers as most of the defaulters have taken a large loan amount and not paid them off. One thing to notice is larger the loan amount more is the interest rate hence more probability to default.
- Customers with low DTI (debt to loan ratio) are less likely to default so loan should be issued to such customers. Again we see that higher the DTI ratio higher the interest rate and hence more probability to default.
- Customers under grade A,B and C are the best borrowers and loan should be given to customers under these grades only .As we go down to G grade the probability to default increases. Customers in G grade have the highest default rate.
- We have also observed that customers who have taken loan for small businesses , house rent and debt consolidation had a higher amount of loan and at a high interest too and these were the customers who defaulted the most. So company should think before giving loan for such purposes.
- States NE has very high chances of Defaulting but number of applications are too low to make any decisions . NV,CA and FL states shows good number of Defaulters in good number of applications.
- Customers having derogatory public records and bankruptcies record should not be given loans.
- Customers which are delinquent for more than 2 months and who always tend to pay late fee should not be issued loans in future as there defaulting rate is becomes high.