Exploratory Data Analysis on LENDING CLUB's Loan data

Business Objective

- the objective of this EDA study is to understand the driving factors (or driver variables)behind loan default, i.e., the variables which are strong indicators of default.

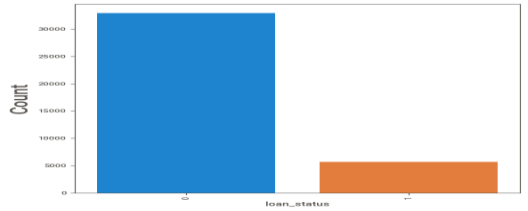
Section: Exploratory data Analysis

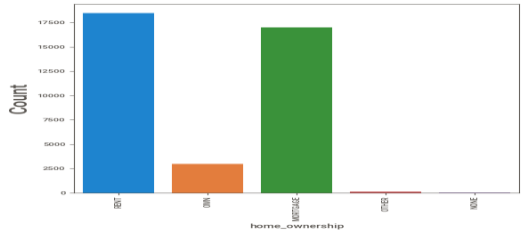
- Introduction
- Section 1: UNDERSTANDING THE DATA
- Section 2: DATA CLEANING AND INITIAL ANALYSIS
- Section 3:Plotting Correlations between continous variables
- Section 4:Plotting Association among variables
- Section 5:Analyzing Outliers(Visualization)
- Section 6:Imputing Missing Value
- Section 7: Findings

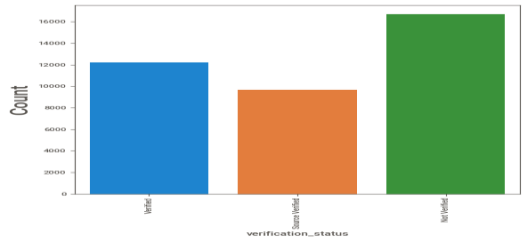
Distribution of listed variables w.r.t their count

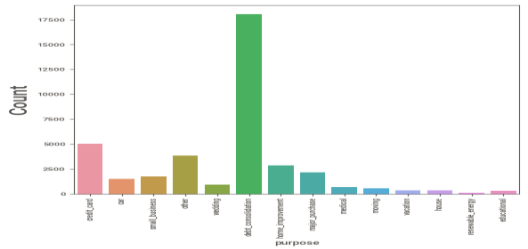
FINDINGS

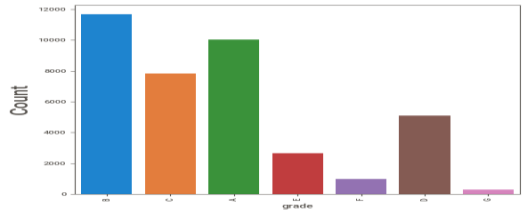
- Loan status plot clearly highlights 6.5 to 1 of Fully paid to defaulters' ratio.
- Company has been making one loss to every 6.5 non defaulting account
- home ownership plot shows that majority of the borrowers are Rented/mortgage
- Verification status plot shows that company has been accepting non verified customers compared to verified and source verified.
- Purpose plot shows that majority of the loans are borrowed for debt consolidation.
- Grade A and B loan are distributed proportionately more than other grades.
- Sub grades' A4, A5,B3,B5,C2 are distributed proportionately higher than other sub grades.
- Demand of 36 months loan has been almost 3 times the demand of 60 months loan.
- Company has been funding more loan to 10+ years of employment length than other years. Company has relatively lesser customer with employment length of 8

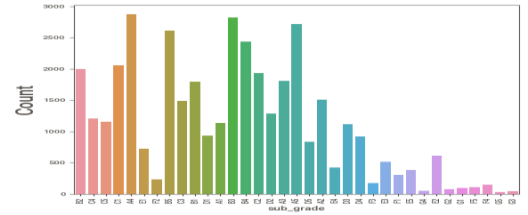


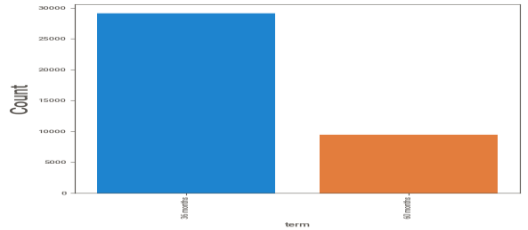


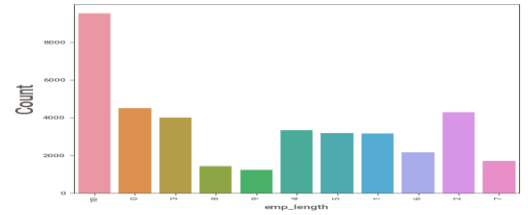












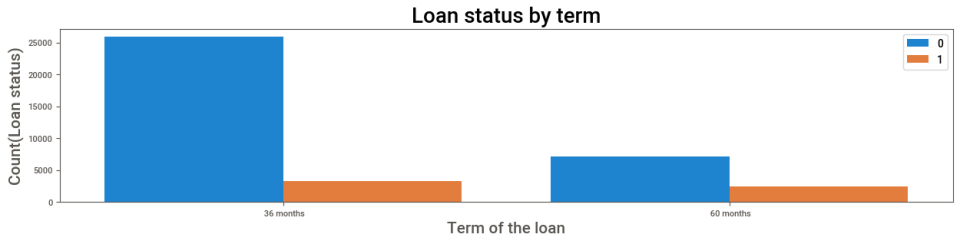
Proportion of fully paid to defaulters are much better in 36 months term than 60 months.

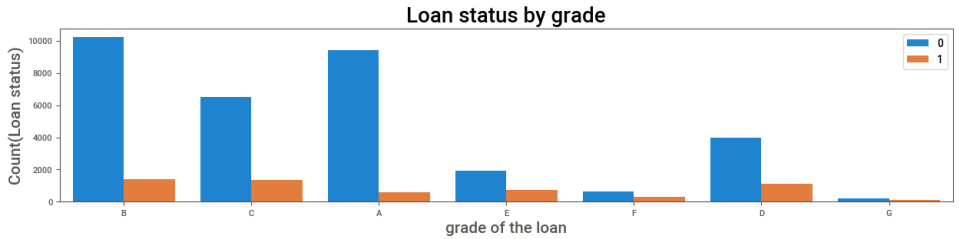
Proportion of fully paid to defaulters are much better in grade A,B and C than grade E,F,G.

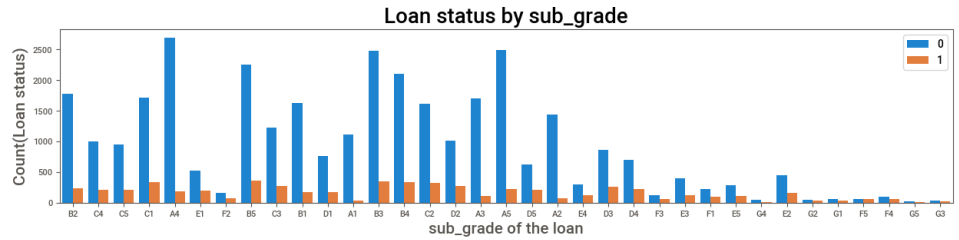
Grade G is the worst performer.

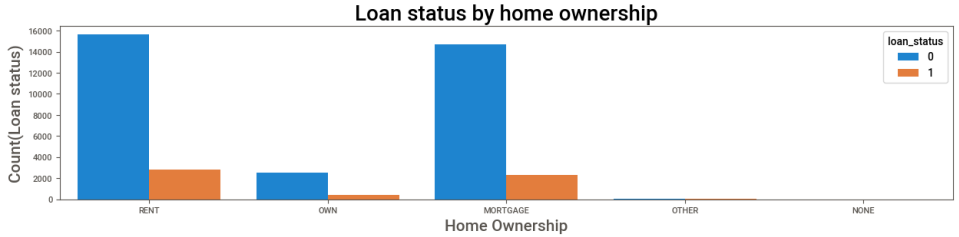
Subgrades' A4,A5,b3,B4,B5 are best performers whereas G5,G3,F5 are worst performers.

Default risk across the home ownership level is low.





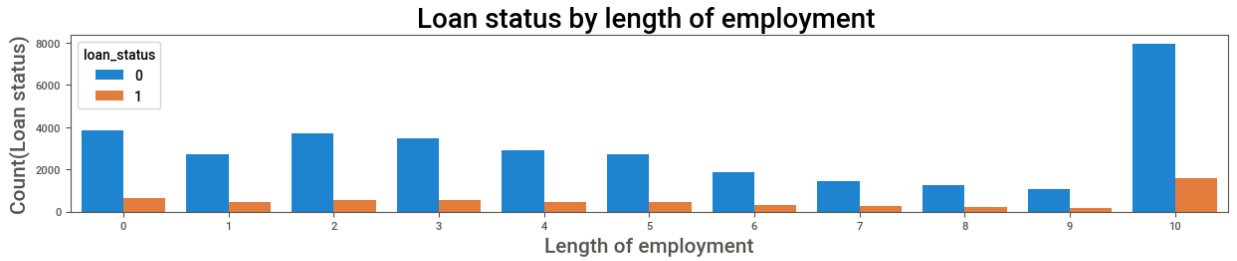


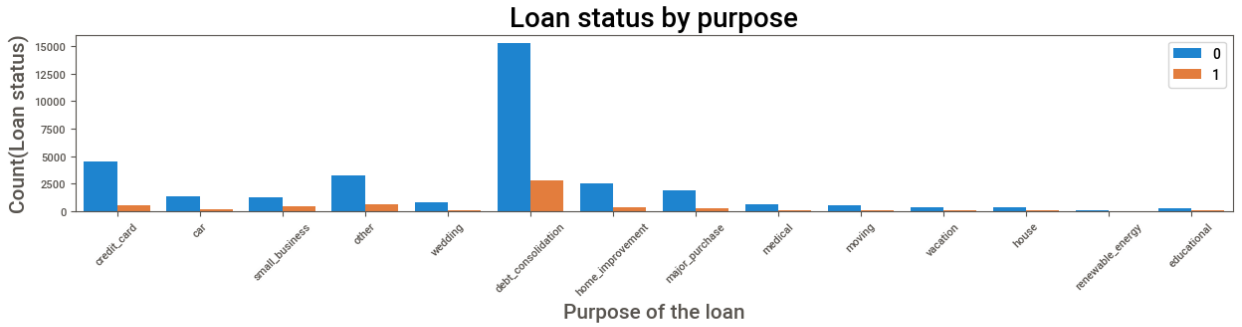


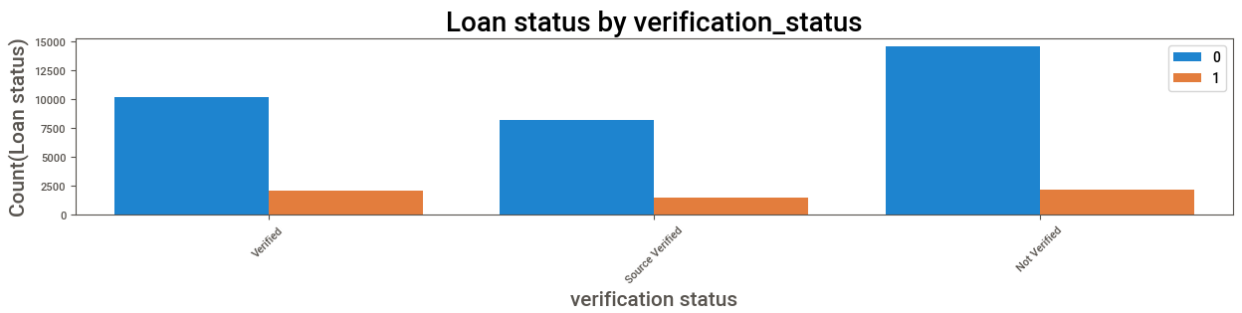
Proportion of fully paid to defaulters are considerably similar at all level of employment length.

Proportion of fully paid to defaulters are much better in verification status of non verified better than verified and source verified.

Default risk is considerably negligible where the purpose have been Moving, Vacation, and educational.

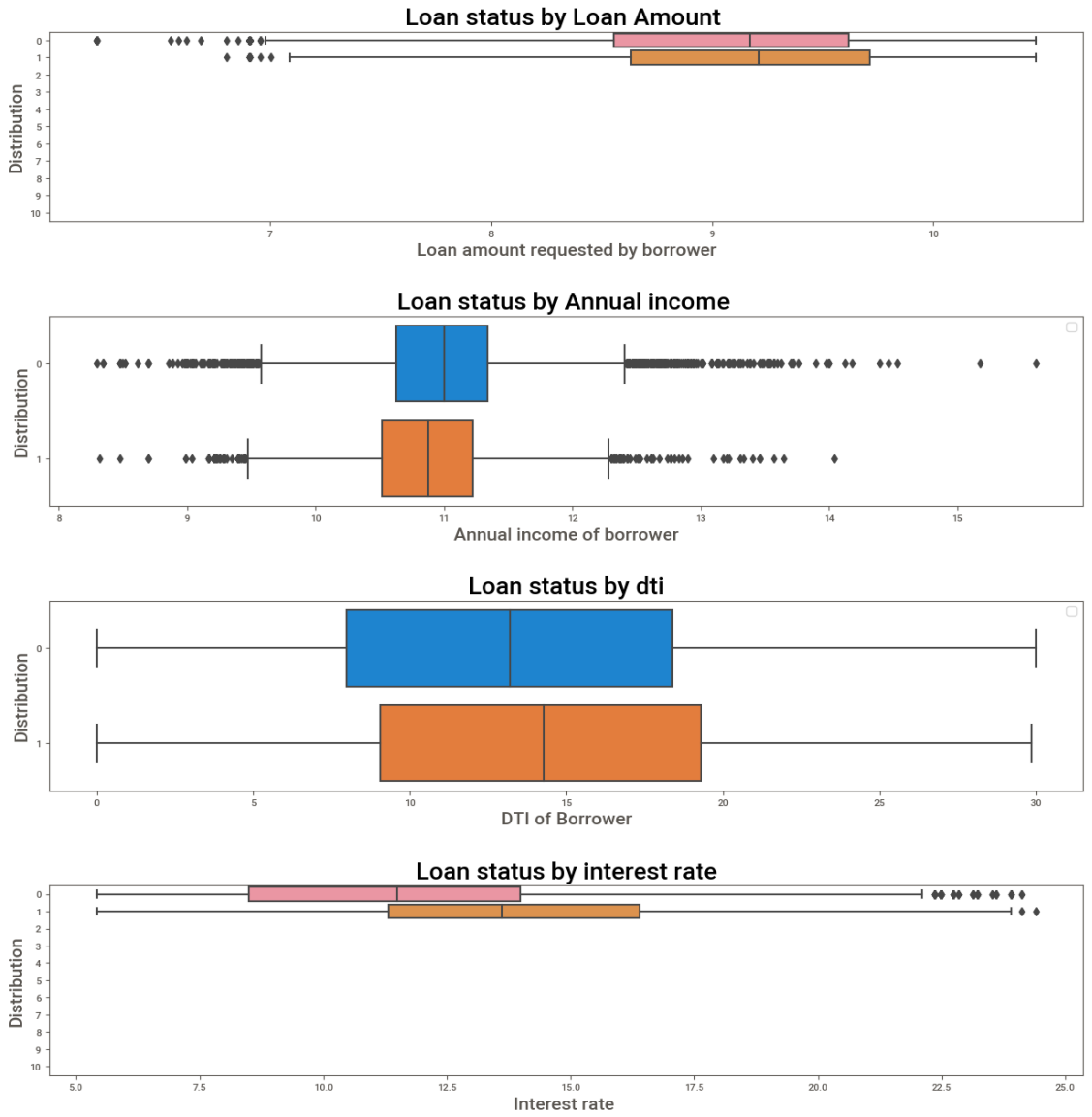






Loan Status wrt. Continuous variables

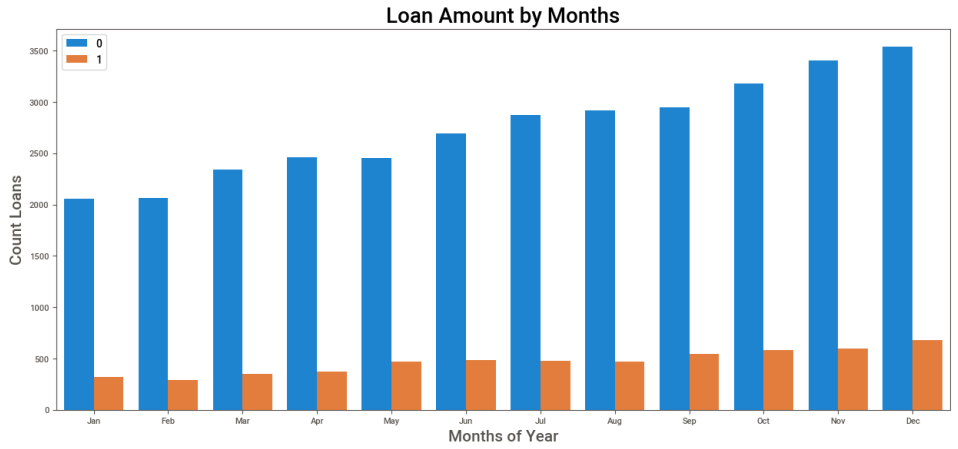
- Median loan amount is higher for defaulters.
- Median DTI is higher for Defaulters.
- Median Interest rate is higher for defaulters.
- Median Income for non defaulters are high.

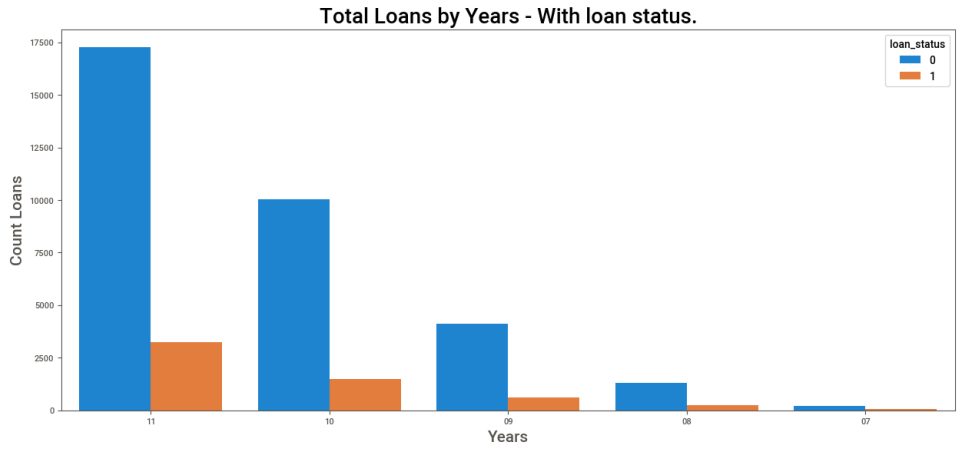


Loan Amount by months/years wrt. Loan status

Finding

- defaulter and fully paid customers have grown from 2007 to 2011.
- defaulter and fully paid customers have grown from January to December.





Loan status wrt. Employment Title

- Employee title such as Lockheed martin, USAF ,Wells Fargo, Walgreenshave perform better
- Walmart is the worst performer.

loan_status	0	1
emp_title		
AT&T	79.000000	21.000000
Bank of America	81.000000	19.000000
Department of Defense	89.000000	11.000000
IBM	86.000000	14.000000
Kaiser Permanente	86.000000	14.000000
Lockheed Martin	95.000000	5.000000
Self Employed	88.000000	12.000000
State of California	89.000000	11.000000
U.S. Army	80.000000	20.000000
UPS	77.000000	23.000000
US ARMY	87.000000	13.000000
US Air Force	88.000000	12.000000
US Army	86.000000	14.000000
USAF	96.000000	4.000000
USPS	82.000000	18.000000
United States Air Force	88.000000	12.000000
Verizon Wireless	90.000000	10.000000
Walgreens	92.000000	8.000000
Walmart	69.000000	31.000000
Wells Fargo	92.000000	8.000000

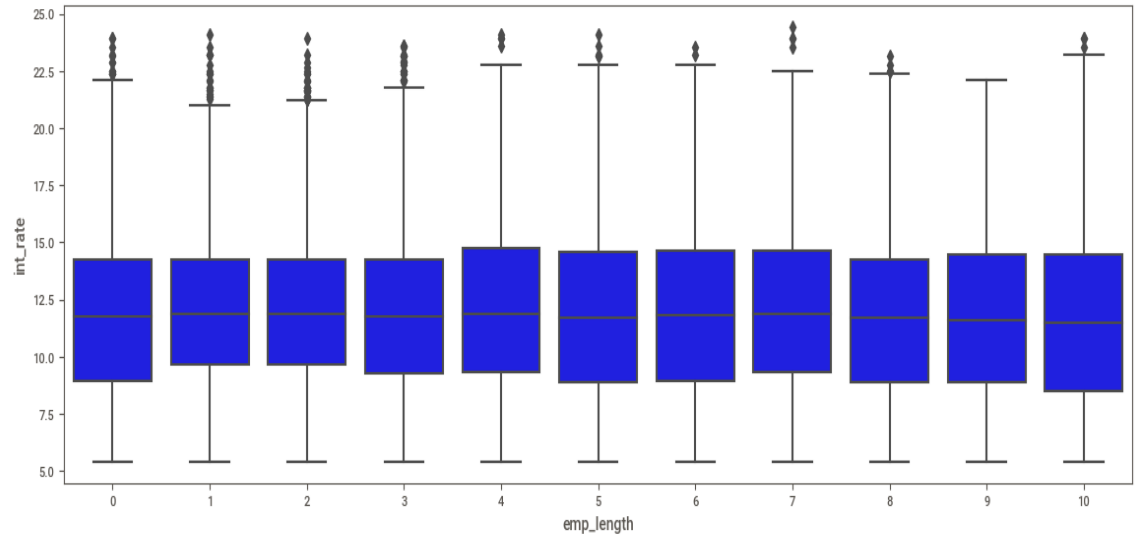
State Wrt. To Loan status

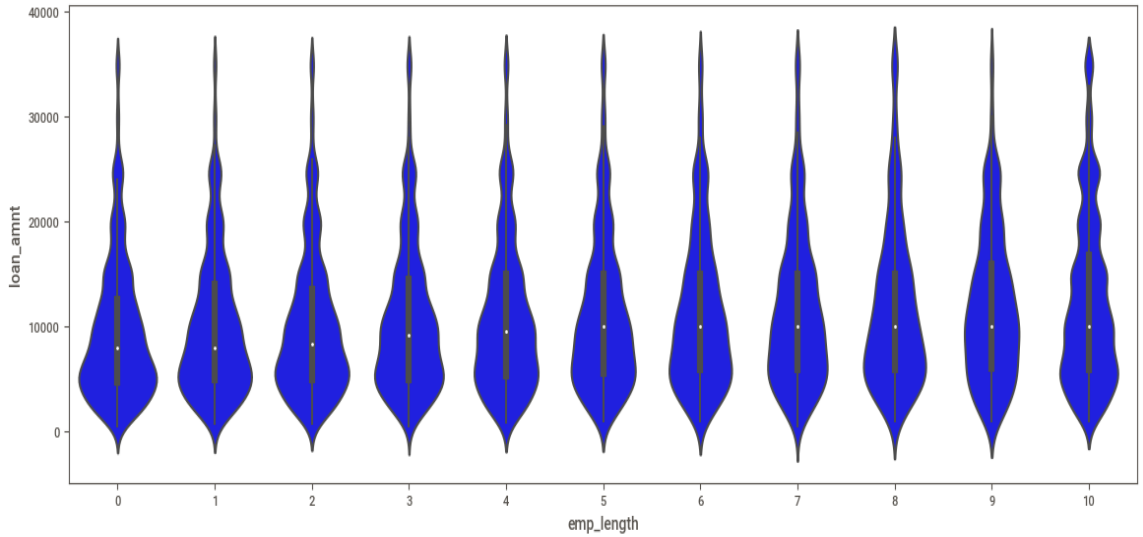
- State such as California, new York , Texas and Florida have state where company are experiencing high number of fully paid as well as defaulters.

loan_status	0	1
addr_state		
AK	0.160000	0.040000
AL	0.990000	0.140000
AR	0.540000	0.070000
AZ	1.880000	0.320000
CA	15.100000	2.920000
CO	1.730000	0.250000
CT	1.640000	0.240000
DC	0.510000	0.040000
DE	0.260000	0.030000
FL	5.900000	1.310000
GA	2.970000	0.560000
HI	0.360000	0.070000
IA	0.010000	0.000000
ID	0.010000	0.000000
IL	3.320000	0.510000
IN	0.020000	0.000000
KS	0.580000	0.080000
KY	0.690000	0.120000
LA	0.970000	0.140000
MA	2.950000	0.410000
MD	2.230000	0.420000
ME	0.010000	0.000000
MI	1.560000	0.270000
MN	1.360000	0.210000
MO	1.440000	0.300000
MS	0.040000	0.010000
MT	0.190000	0.030000
NC	1.650000	0.300000
NE	0.010000	0.010000
NH	0.370000	0.060000
NJ	3.920000	0.720000
NM	0.400000	0.080000
NV	0.960000	0.280000
NY	8.300000	1.280000
OH	2.650000	0.400000
OK	0.640000	0.100000
OR	0.940000	0.180000
PA	3.340000	0.470000
RI	0.440000	0.060000
SC	1.020000	0.170000
SD	0.130000	0.030000
TN	0.040000	0.010000
TX	6.070000	0.820000
UT	0.550000	0.100000
VA	3.090000	0.460000
VT	0.120000	0.020000
WA	1.790000	0.330000
WI	0.980000	0.160000
WV	0.390000	0.050000
WY	0.200000	0.010000

Relationship: Interest rate, Employment level, loan amount

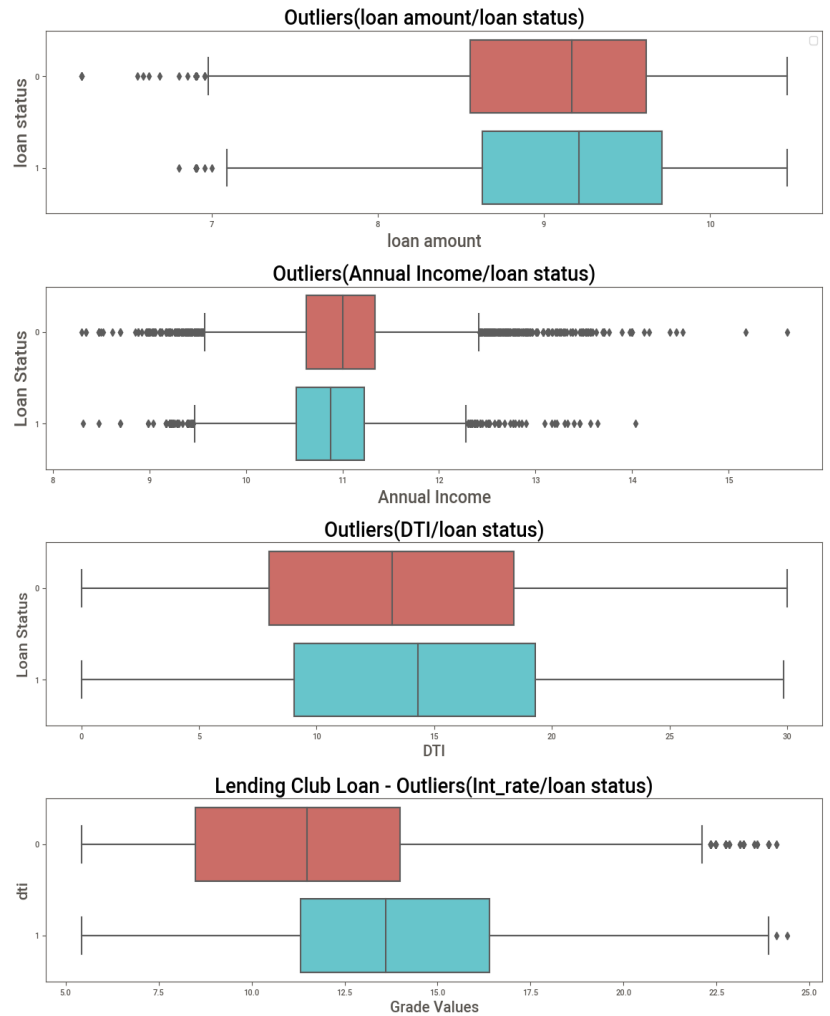
- There is not much difference in interest rate among different level of work experience.
- loan amount has been higher for employment length of 10 years.





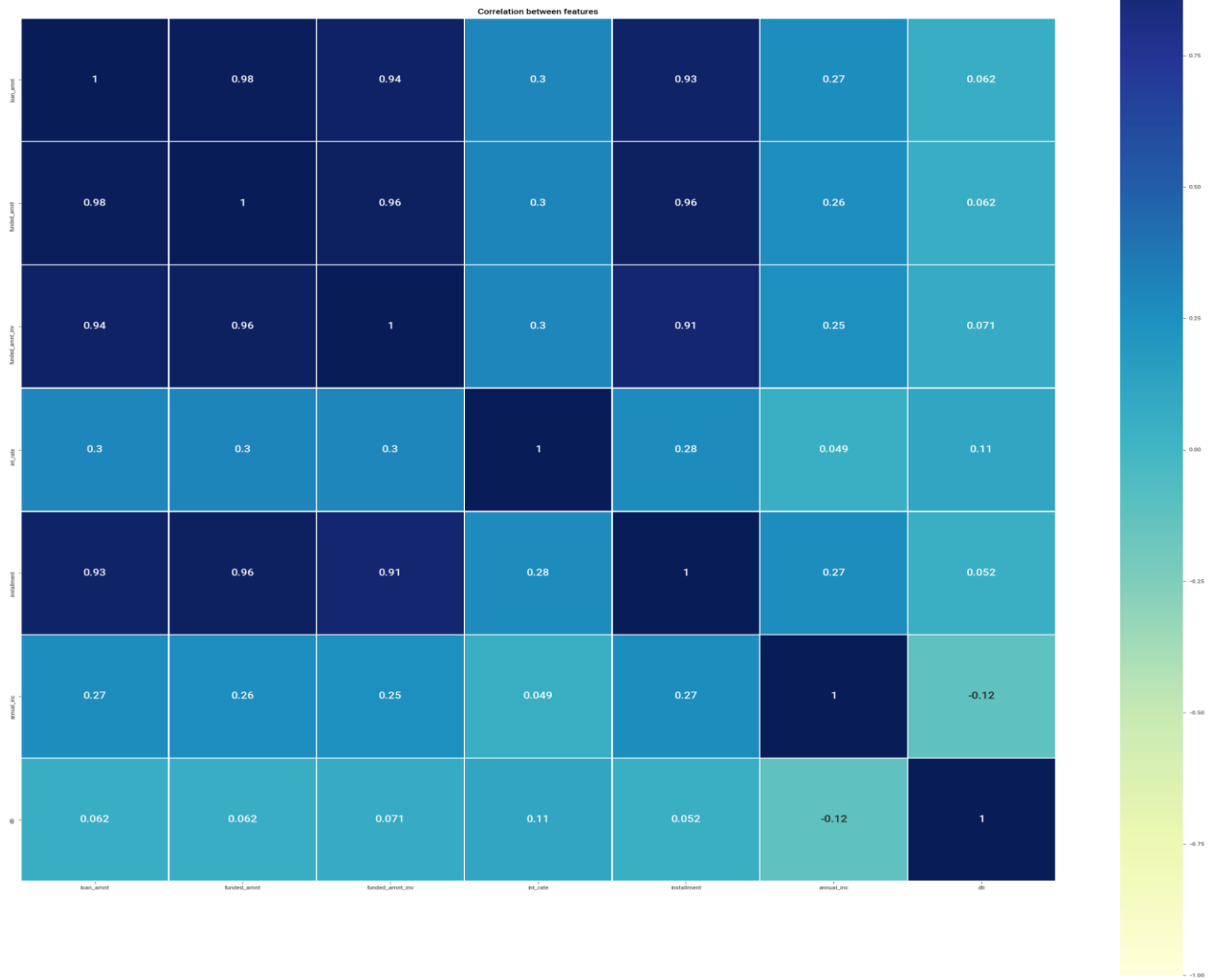
- no outliers in DTI
- High outliers in Annual income and loan amount.

Outliers in terms of loan amount,Annual Income,interest rate and dti .wrt. Loan Status

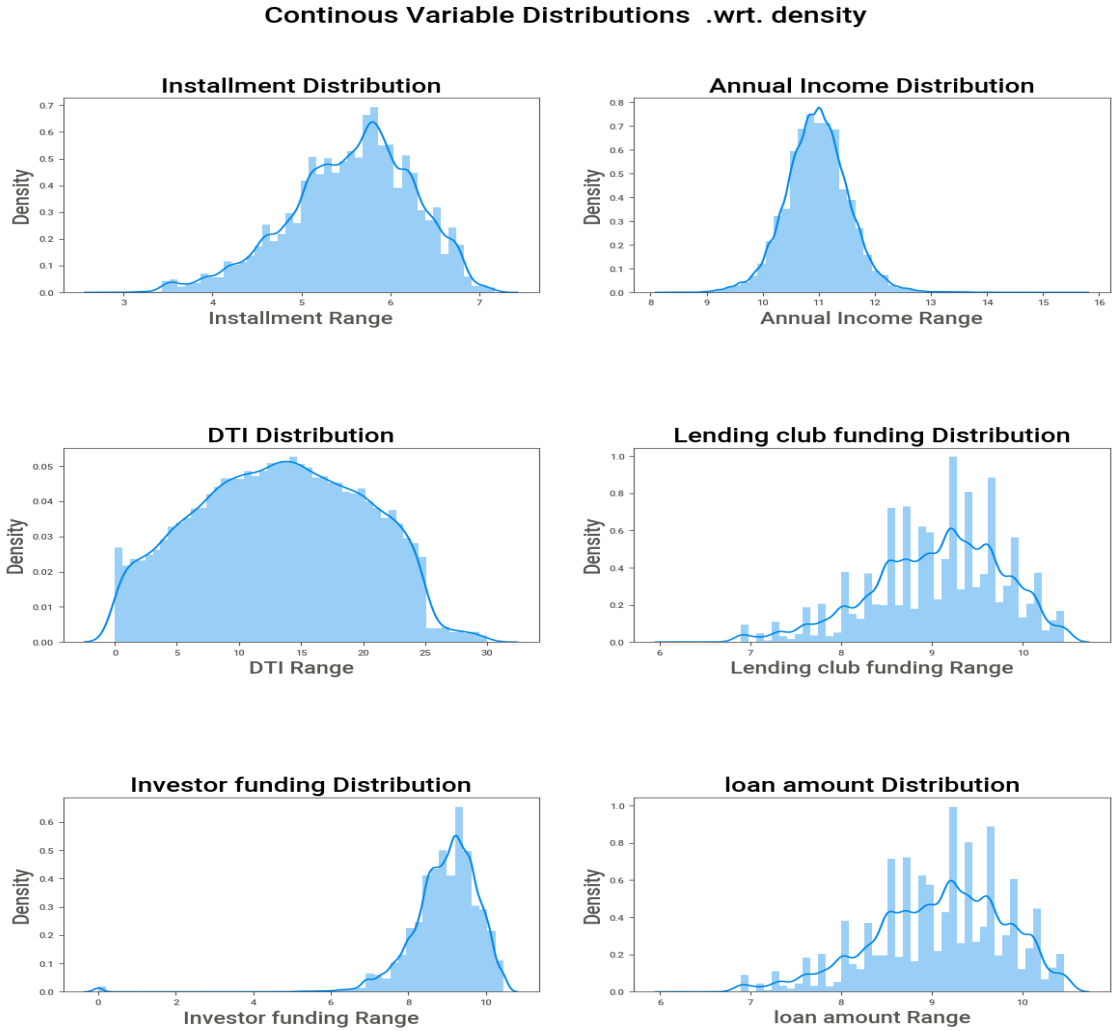


Correlation Matrix

- Strong Correlation between loan amount, amount funded by investor, amount funded by LC and installment.
- They form a segment with very high correlation.



Comparing distribution before and after log1 Transformation



Plot findings

- DTI has normal distribution
- Installment, Annual income, amount funded by LC, Loan amount, amount funded by investor and interest rate are positively skewed.
- Skewed variables have now normal distribution after log1 transformation as we managed the outliers efficiently without dropping them.

Findings

- Loan status plot clearly highlights 6.5 to 1 of Fully paid to defaulters' ratio.
- Company has been making one loss to every 6.5 non defaulting account
- home ownership plot shows that majority of the borrowers are Rented/mortgage
- Verification status plot shows that company has been accepting non verified customers compared to verified and source verified.
- Purpose plot shows that majority of the loans are borrowed for debt consolidation.
- Grade A and B loan are distributed proportionately more than other grades.
- Sub grades' A4, A5,B3,B5,C2 are distributed proportionately higher than other sub grades.
- Demand of 36 months loan has been almost 3 times the demand of 60 months loan.
- Company has been funding more loan to 10+ years of employment length than other years. Company has relatively lesser customer with employment length of 8

Findings

- Proportion of fully paid to defaulters are much better in 36 months term than 60 months.
- Proportion of fully paid to defaulters are much better in grade A,B and C than grade E,F,G.
- Grade G is the worst performer.
- Subgrades' A4,A5,b3,B4,B5 are best performers whereas G5,G3,F5 are worst performers.
- Default risk across all the home ownership level is low.

Findings

- Proportion of fully paid to defaulters are considerably similar at all level of employment length.
- Proportion of fully paid to defaulters are much better in verification status of non verified better than verified and source verified.
- Default risk is considerably negligible where the purpose have been Moving, Vacation, and educational.
- defaulter and fully paid customers have grown from 2007 to 2011.
- defaulter and fully paid customers have grown from January to December.

Findings

- defaulter and fully paid customers have grown from 2007 to 2011.
- defaulter and fully paid customers have grown from January to December.
- State such as California, new York , Texas and Florida have state where company are experiencing high number of fully paid as well as defaulters.
- There is not much difference in interest rate among different level of work experience.
- loan amount has been higher for employment length of 10 years.

Findings

- no outliers in DTI
- High outliers in Annual income and loan amount.
- DTI has normal distribution
- Installment, Annual income, amount funded by lending club, Loan amount, amount funded by investor and interest rate are positively skewed.
- Skewed variables have now normal distribution after log1 transformation as we managed the outliers efficiently without dropping them.

Findings

- Strong Correlation between loan amount, amount funded by investor, amount funded by LC and installment.
- They form a segment with very high correlation.
- Interest rate, grade and sub grade shows strong association among each other, and they form a segment with strong association.
- Median loan amount is higher for defaulters.
- Median DTI is higher for Defaulters.
- Median Interest rate is higher for defaulters.
- Median Income for non defaulters are high.

Recommendation

- Company should fund more loans with 36-month term as there are lesser defaulters compared to 60-month loan term.
- Company should fund more loans with Grade A,B and C as there are lesser defaulters.
- Company should halt it’s funding for Grade G loans as it performs worst among all grades.
- Company should focus on Subgrades’ A4,A5,b3,B4,B5 as these are best performers whereas put stringent criteria for G5,G3,F5 as they are worst performers.
- Company should increase the interest rate for borrowers with less than 5 years of employment level as there is not much difference in interested rate across the employment level.

- Company should lend more to borrowers with purpose: Moving, Vacation, and educational as default has been negligible in this segment.
- Company should fund more loan for borrowers with higher income with employment level of at least 5 years.
- Company should revisit their funding strategy for verified and source verified borrowers .Surprisingly non verified borrowers have performed better.
- Six variables : DTI,Interest rate, Term, Grade, Sub Grade and Annual Income are most significant driving factors. They don’t share segments with high correlation/association.

Thank you