Lending Club Case Study

For,

Indian Institute of Information Technology, Bangalore, Upgrad Education

Prepared By:

Shashank Mishra
Naveen Cheruku

Objective:

The Objective of this Case study is to apply concepts of Exploratory Data analysis to arrive at conclusions which will help Consumer Finance companies to make better decisions while lending to Borrowers.

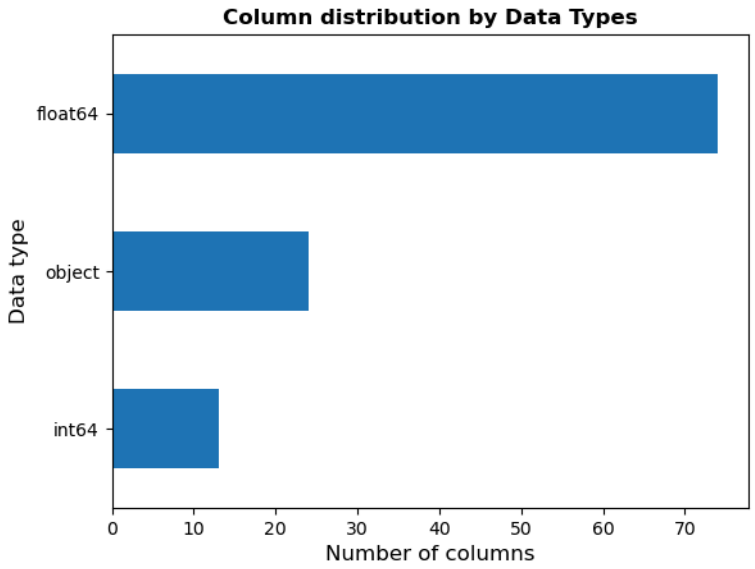
Data Source:

The data has been provided in the form of a csv file. It has an associated Data dictionary to explain the meaning of various data columns present in the cs file.

Approach:

The approach we will follow is the standard approach to analyze a data set which is raw in nature. Hence we will apply the following methodologies to clean the data, understand the nature of data and arrive at a list of data columns which are required from a final analysis perspective. A list of steps that we are going to follow are as follows. Data sourcing has already been mentioned :

1. Data cleaning
2. Univariate Analysis
3. Segmented variate Analysis
4. BiVariate Analysis
5. Derived Metrics, if any
6. Collate and conclude on findings to arrive at suggestions



Data Cleaning and Manipulation

- Identifying columns with most missing values and removing them
- Identifying columns which can be removed based on having missing values and no impact on analysis
- Identifying and removing columns which are mostly unique like 'id'
- Check if there are rows having too many missing or null values which might be candidate for removal
- Finally, remove the columns based on the requirement for the final analysis as some columns may not be required to be part of final analysis

Identification and Removal of Irrelevant Columns

- Columns with no unique values tax_liens, delinq_amnt, chargeoff_within_12_mths, acc_now_delinq, application_type, policy_code, collections_12_mths_ex_med, initial_list_status, pymnt_plan
- Columns with All unique values id, url and member_id are all unique. We can use one of them as identifier and drop other two
- Columns which are not useful emp_title and title columns seem to be just designations. Hence they can be removed

Missing values in rows

- Check what is the maximum number of missing values in a row. If it is a small number say less than 5. Ignore it
- The number of null values seem to be in irrelevant columns which we will drop in our final evaluation of important columns. Hence, leaving them be for now

Handling 'null' values

- Evaluate whether we can impute 'null' values. If not, remove them

Other Issues

- 'Term' column has padded spaces in it which needs to be removed

Data shortlisting required from Business Perspective

The basis for revisiting the leftover column is to check whether they are relevant customer attributes or loan attributes which are useful to make a decision whether to provide the loan or not. So, any column which helps us to make a decision 'before' loan sanction, is 'relevant' otherwise it is irrelevant.

- In order to make a decision on whether to Lend money or not, the Loans which are currently being serviced are not counted. Only the loans which are Fully paid or Charged Off are required. Hence, we will remove the Loans with status Current
- As per provided Data Dictionary following columns represent some data which is gathered after the Loan is sanctioned and hence we will drop these columns 'delinq_2yrs', 'earliest_cr_line', '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_late_fee', 'recoveries', 'collection_recovery_fee', 'last_pymnt_d', 'last_pymnt_amnt','last_credit_pull_d'
- 'Zip_code' is Masked column, hence it will be dropped
- 'Funded_amnt_inv' is Internal data of the Company, not required for making decision
- In order to make a decision on whether to Lend money or not, the Loans which are currently being serviced are not counted. Only the loans which are Fully paid or Charged Off are required. Hence, we will remove the Loans with status Current

Data Conversion

- term column needs to be converted to 'int' after removing string literal 'months'
- grade column can be converted to 'category' type
- sub_grade column can be converted to 'category' type
- emp_length column needs to keep only number part and drop any string or alphanumeric
- home_ownership column can be converted to 'category' type
- verification_status column can be converted to 'category' type
- purpose column can be converted to 'category' type
- addr_state column can be converted to 'category' type
- issue_d can be converted to 'datetime'
- int_rate columns has '%' sign in it. This needs to be removed. Also, data type needs to be converted to float

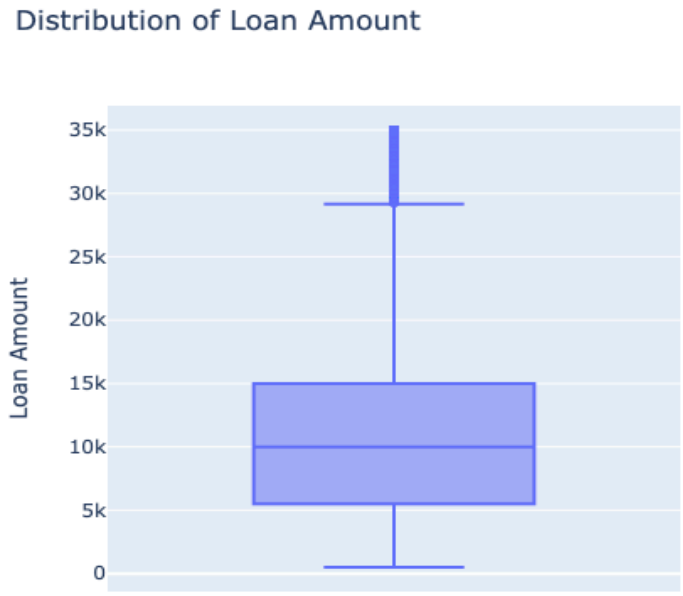
Outlier analysis and Handling

Outlier check was performed on following continuous variables and handled for Annual income to make data uniformly distributed

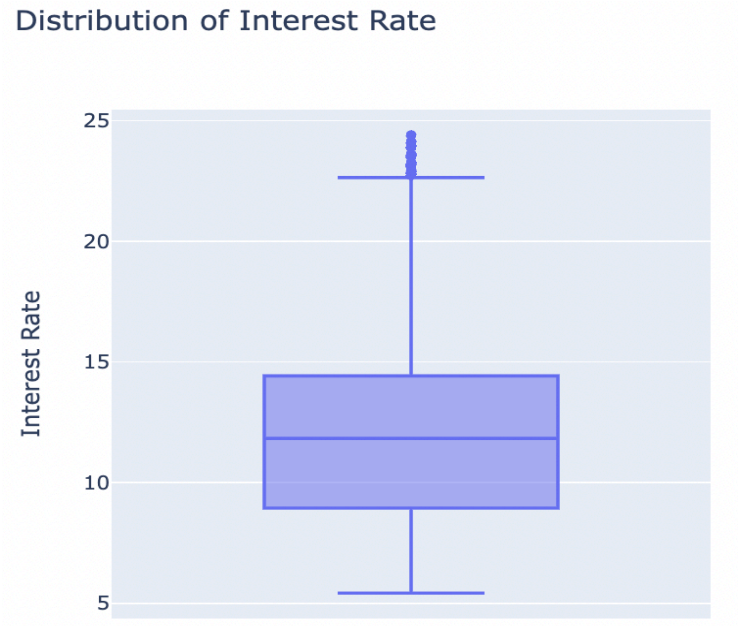
1. Loan amount
2. Interest rate
3. Annual Income
4. DTI - Debt to Income ratio

Univariate Analysis and Outlier Handling

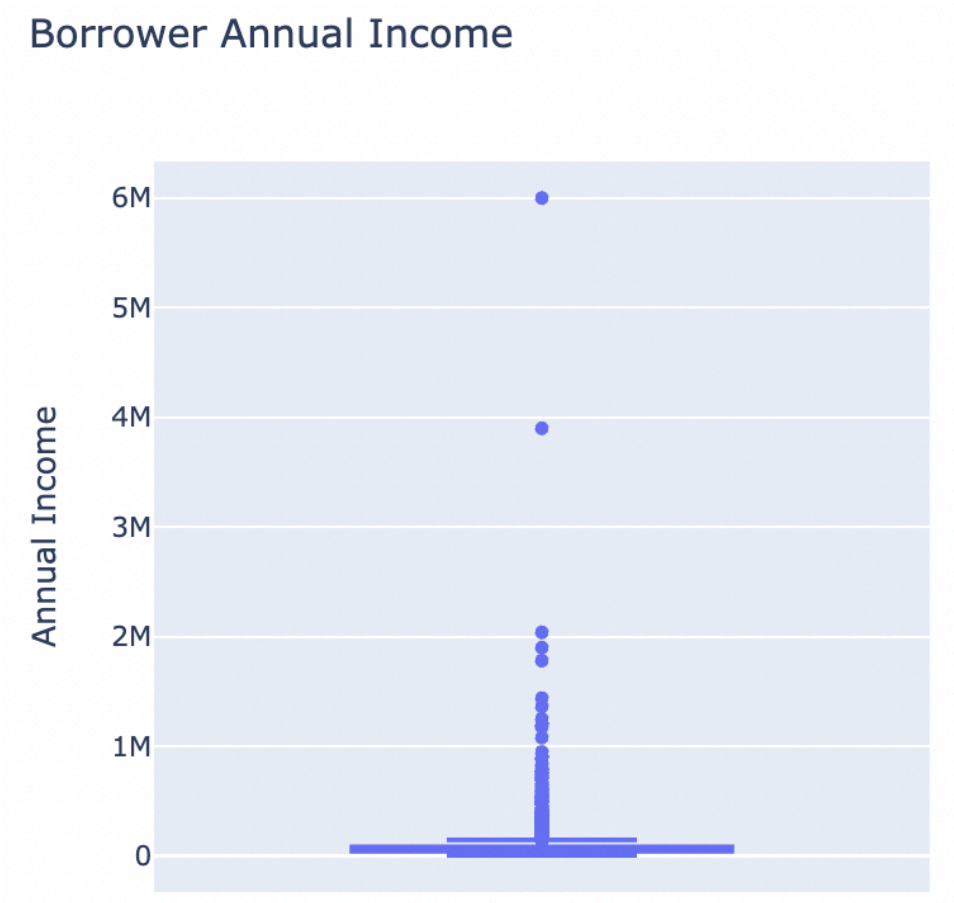
1. **DTI - Debt to Income ratio**

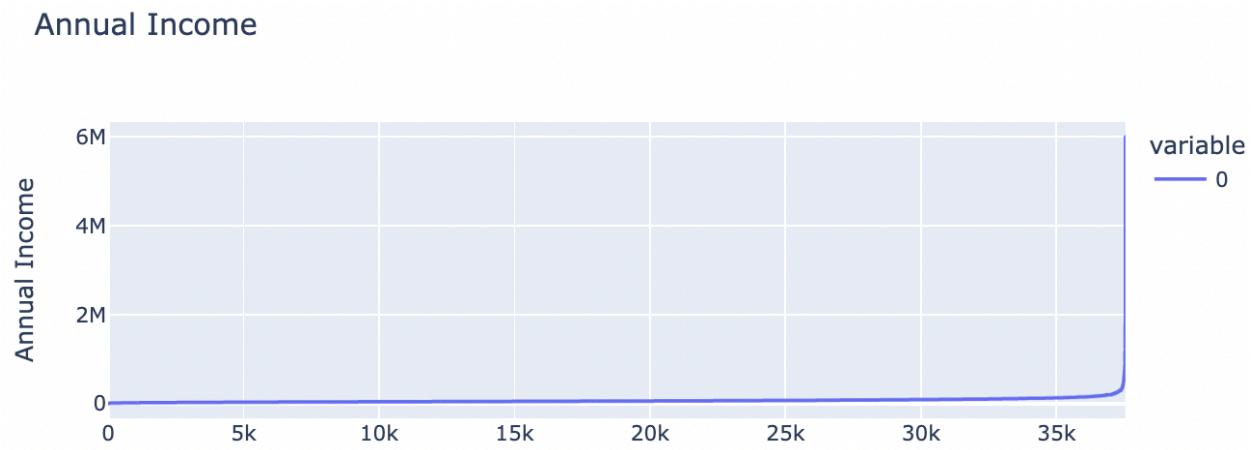


2. **Interest Rate**

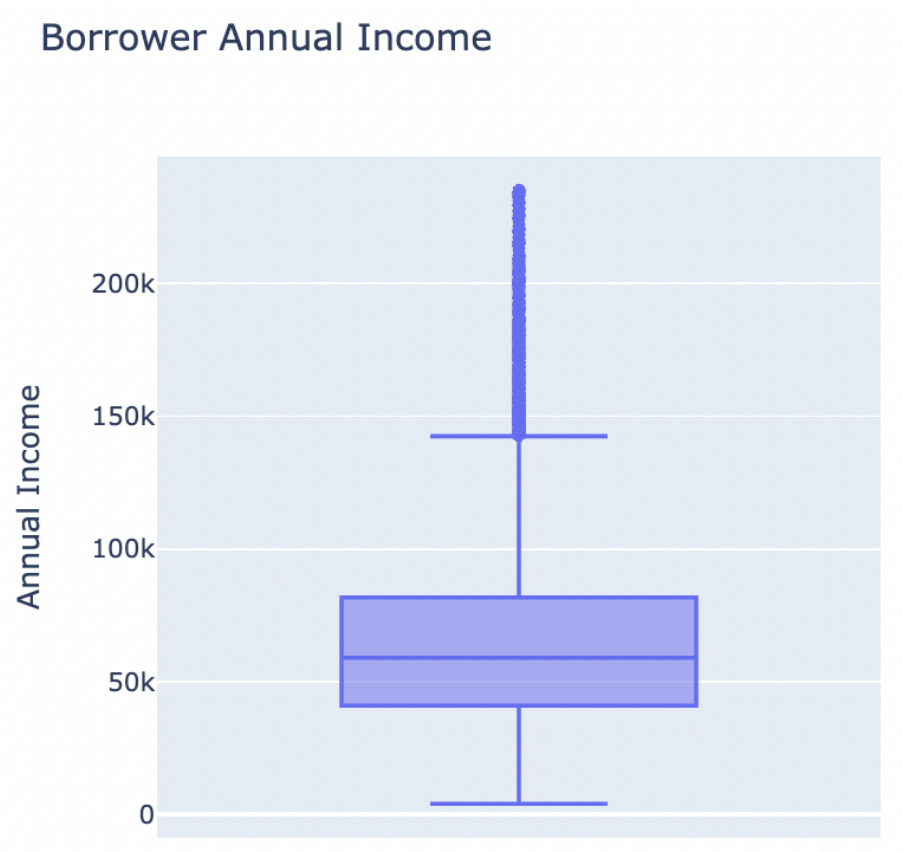


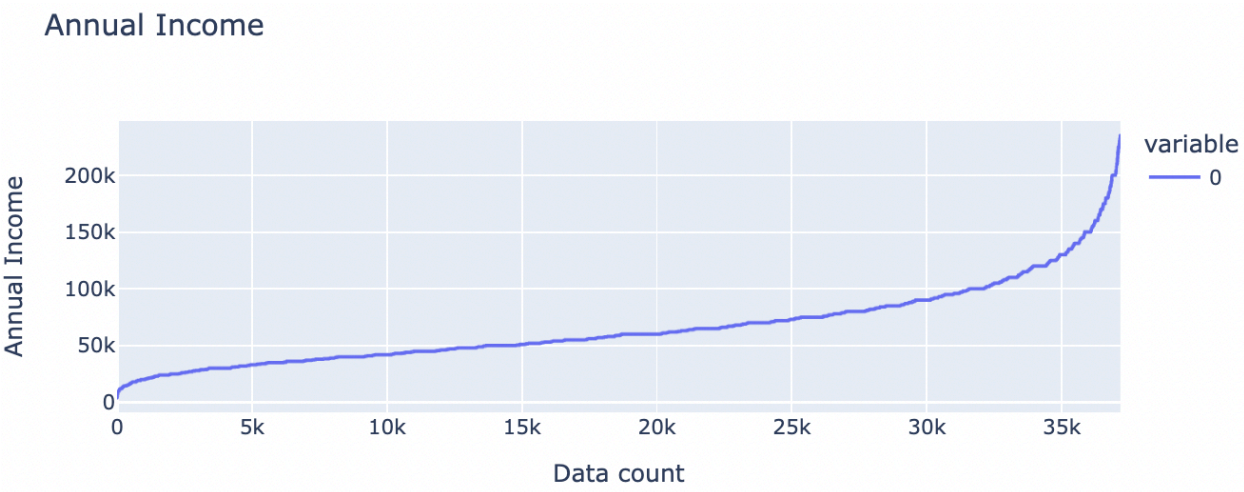
3. Annual Income with Outliers in raw data



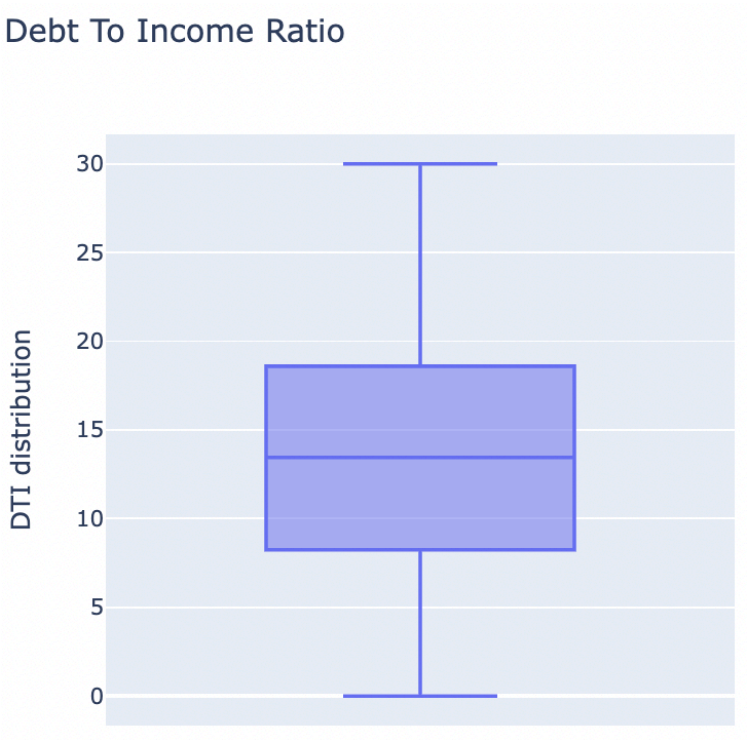


4. Annual Income with 99th percentile outlier removed





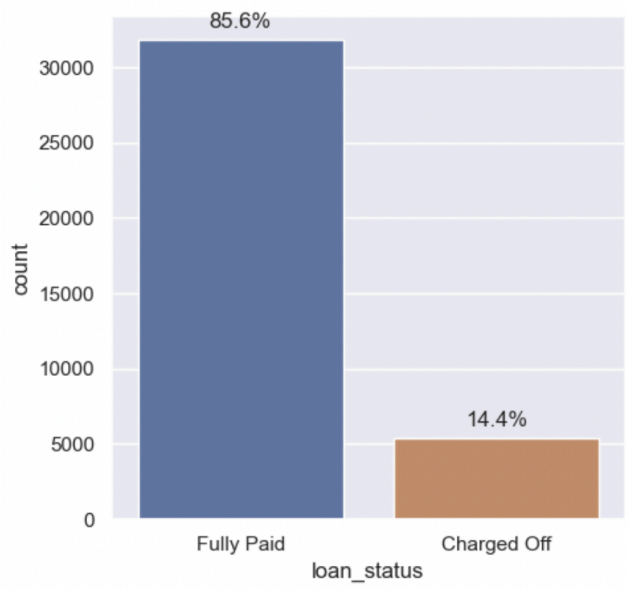
5. DTI - Debt to Income Ratio



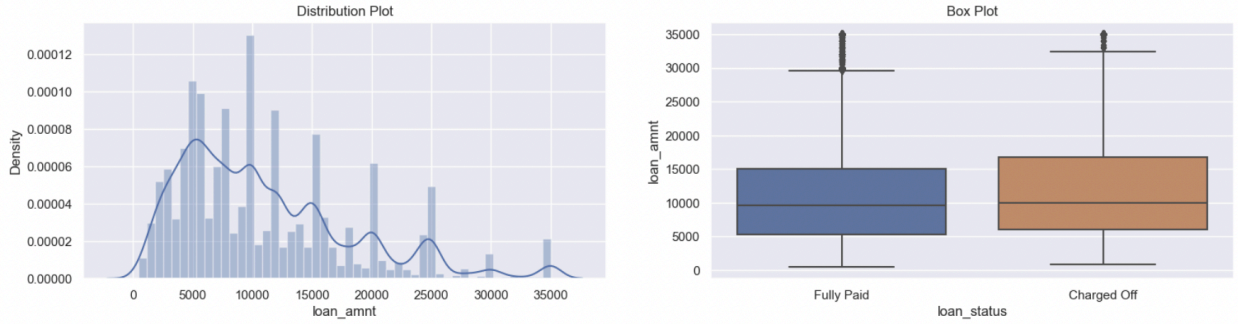
6. Loan_status

The target column for our analysis.

Further we will try to find out how loan_status compares with other columns (Customer and Loan attributes)

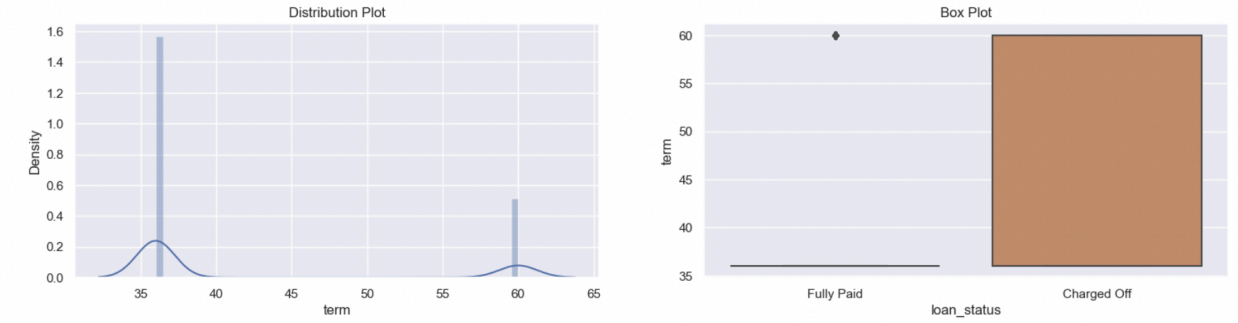


7. loan amount



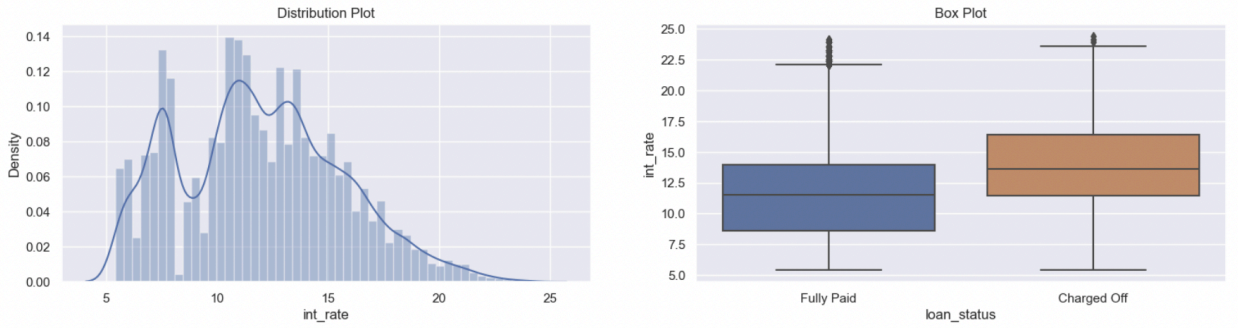
Inference: The loan amount varies from 500 to 35000 with a mean of 9800.

8. term



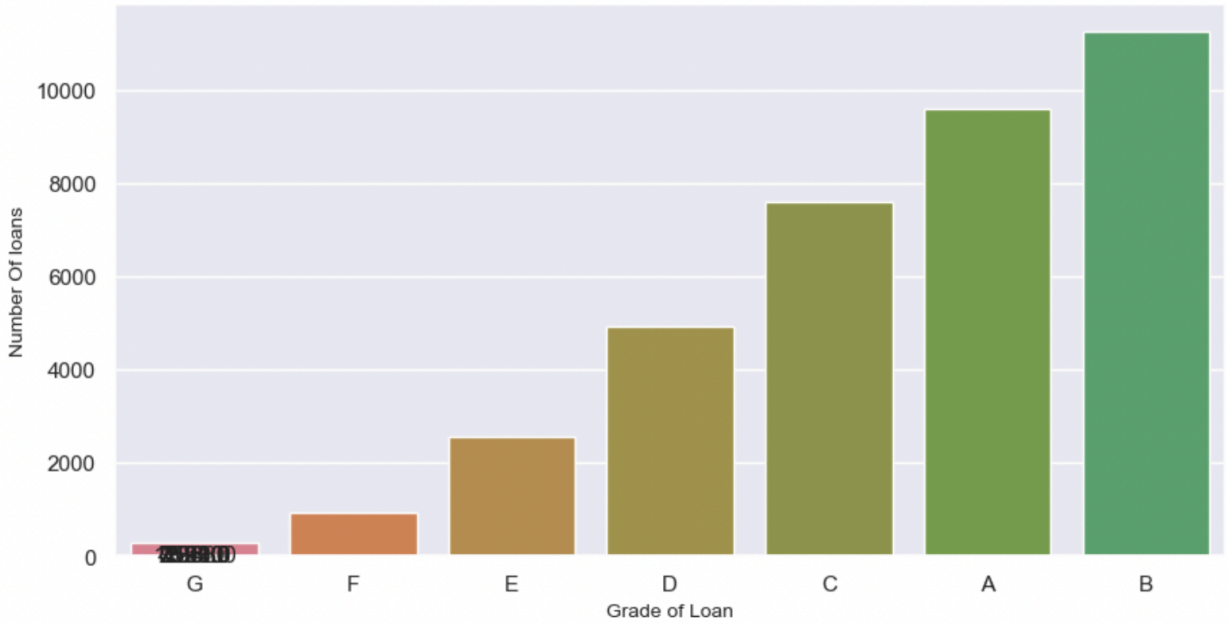
Inference: 75% of the loan taken has term of 36 months compared to 60 months.

9. Interest rate



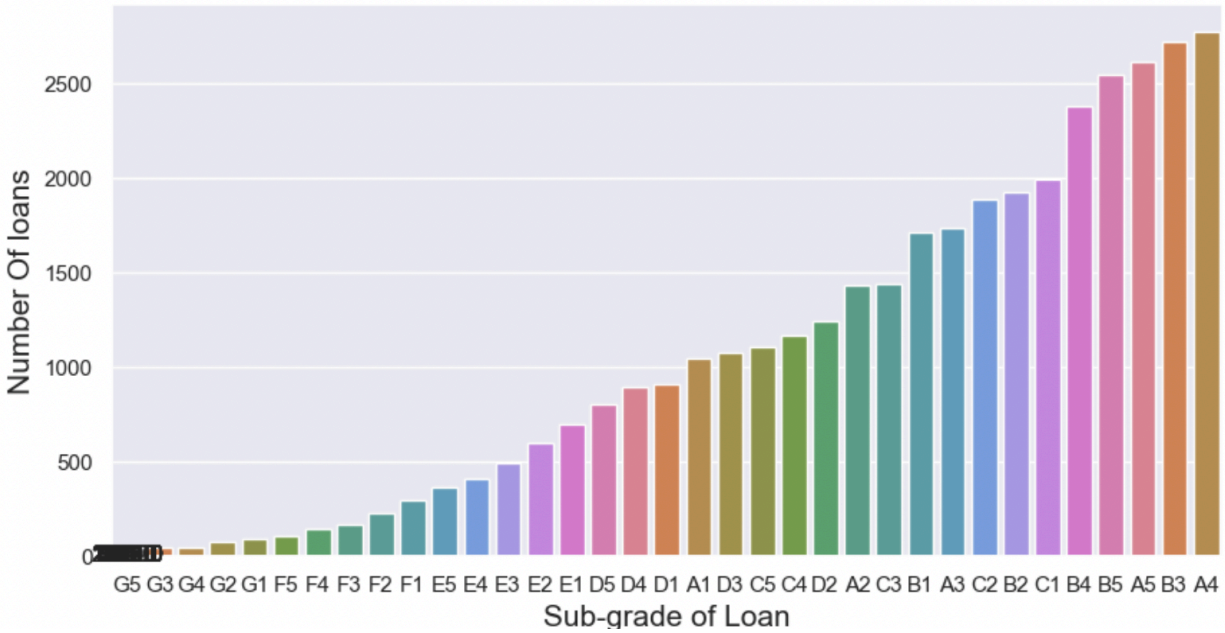
Inference: Most Loans are issued at an interest rate of around 5-10 and 10-15 with a drop near

10. Grade



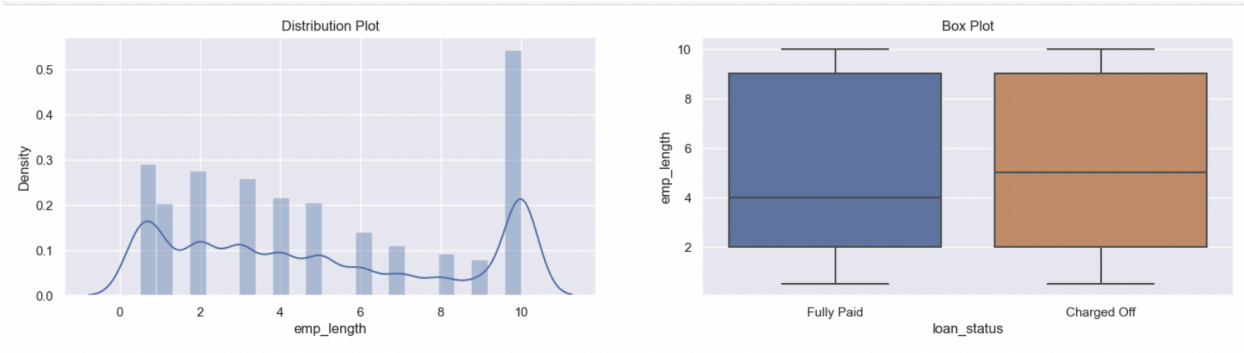
Inference: High Grade Loans are high in numbers with Grade B being highest in number.

11. Sub-grade



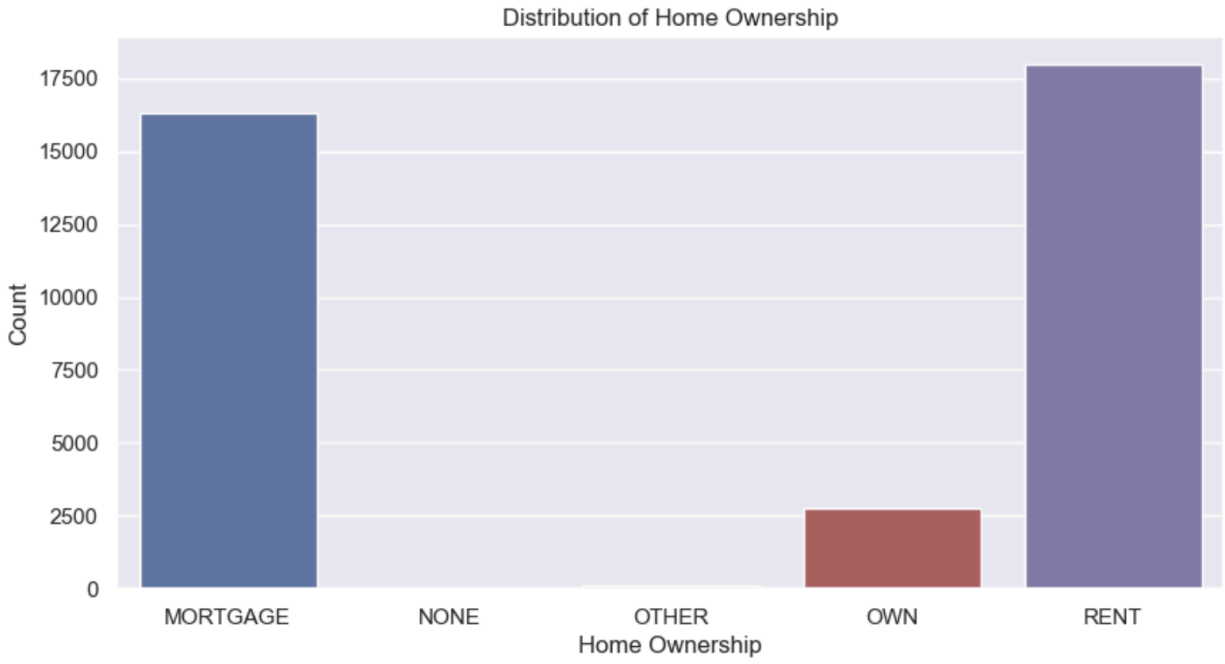
Inference: Higher subgrades in each grade just confirm that higher grade loans are higher in number

12. Employment Length



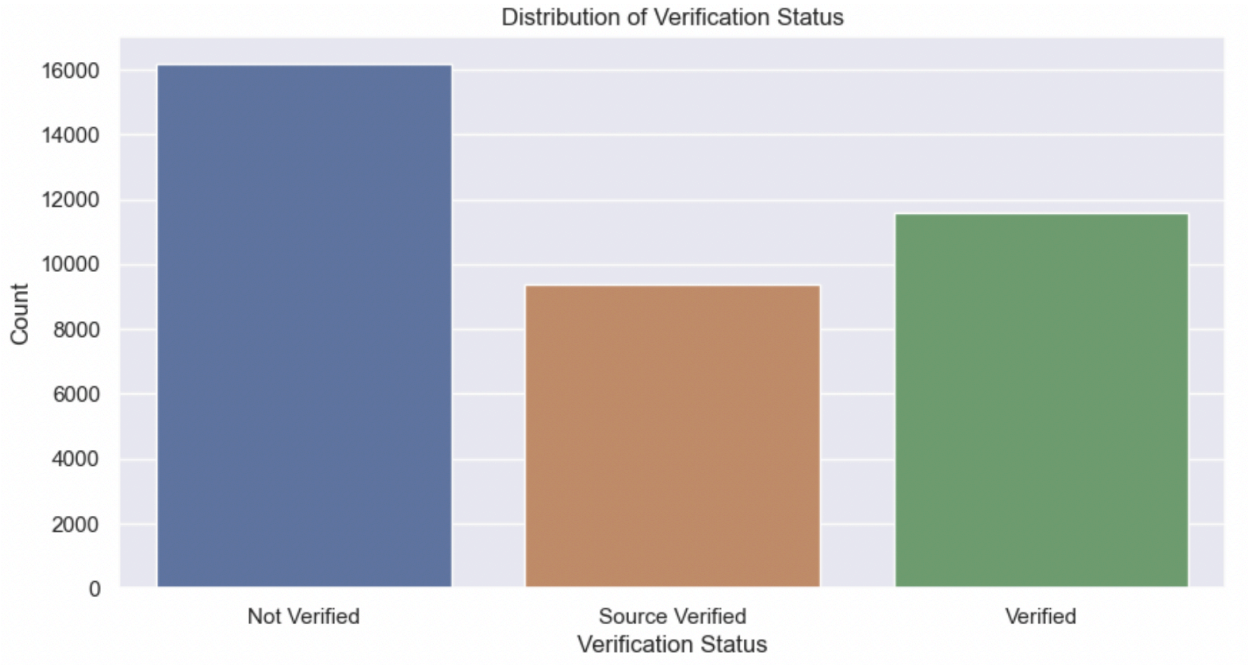
Inference: Largest number of borrowers are the ones having highest employment length with 10+ years being the most common.

13. Home ownership



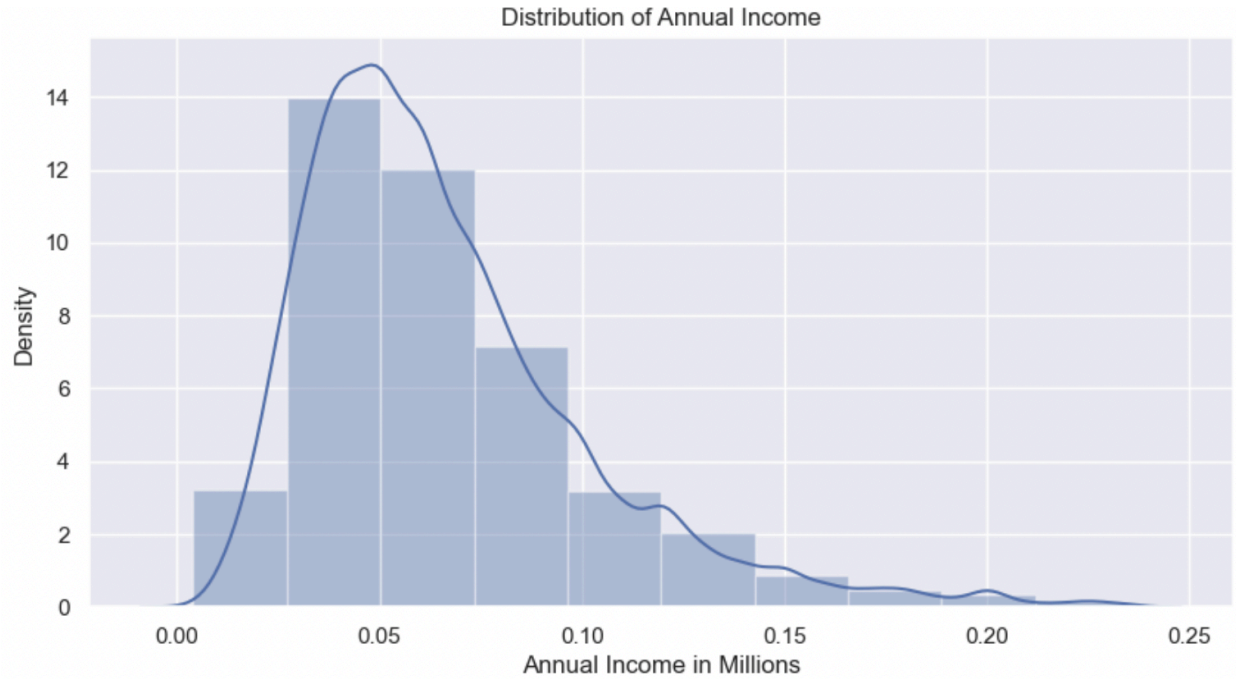
Inference: Borrowers living on rented accomodation are most common followed by those living in mortgaged accomodation. Owners are the least borrowers.

14. Verification status



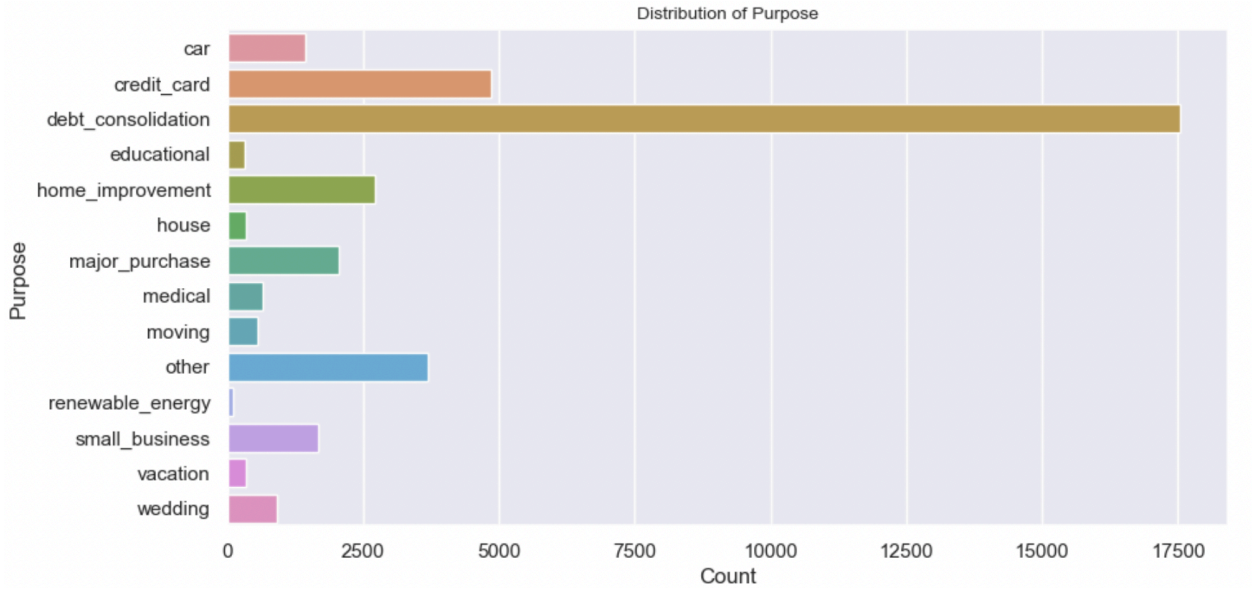
Inference: Most borrowers have been verified by company or have the source verified

15. Annual Income



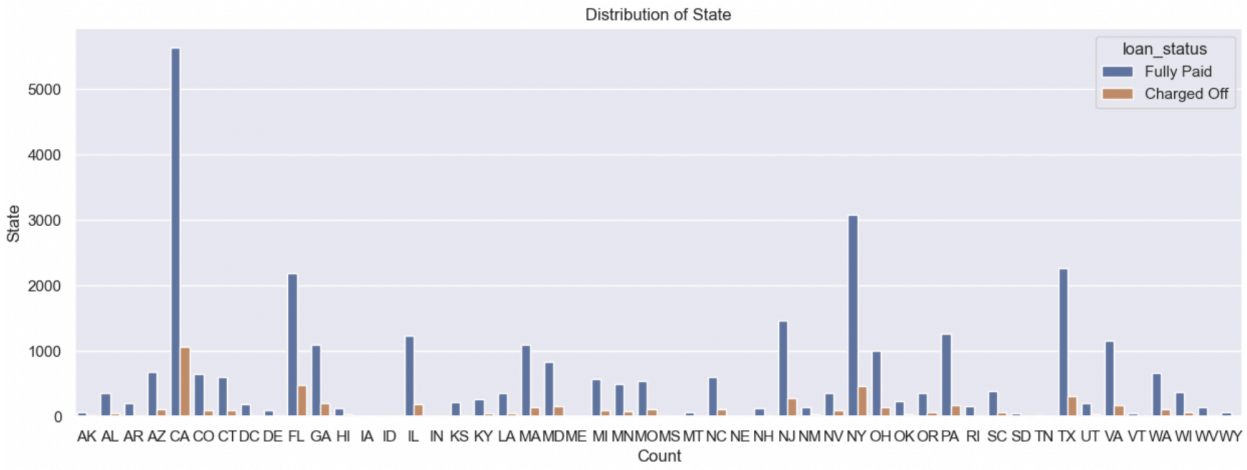
Inference: Majority of borrowers have very low annual income.

16. Purpose



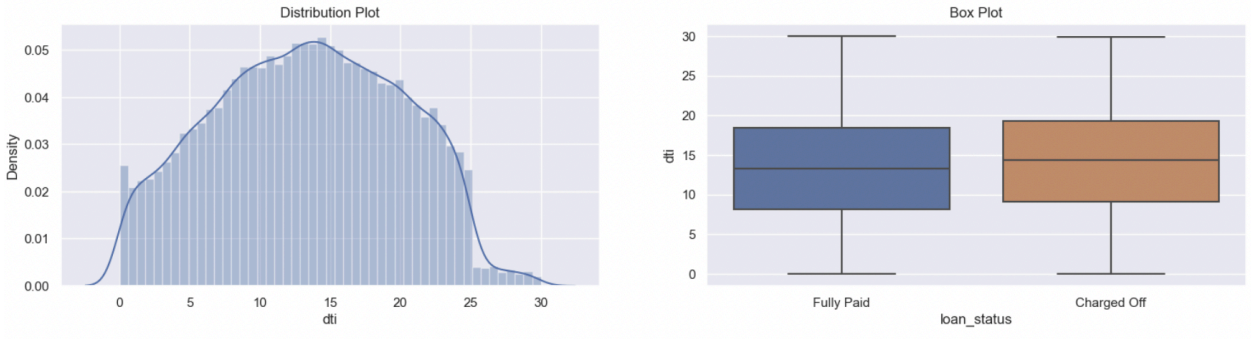
Inference: Borrowers are taking more loan to consolidate their existing debts. Debt_consolidation is the most specified reason for taking loan

17. Address State



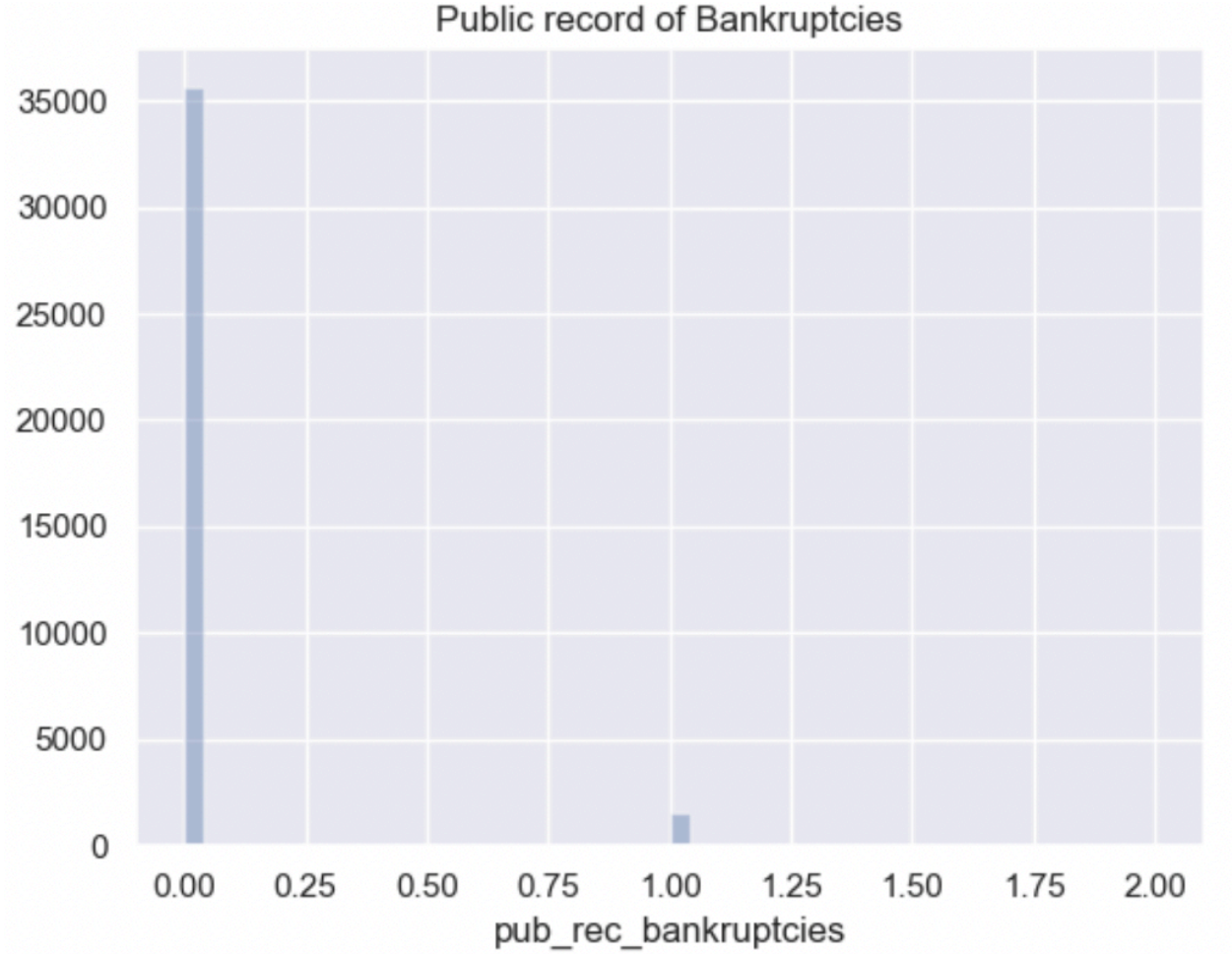
Inference: Most of the borrowers are from states california, new york, texas, florida.

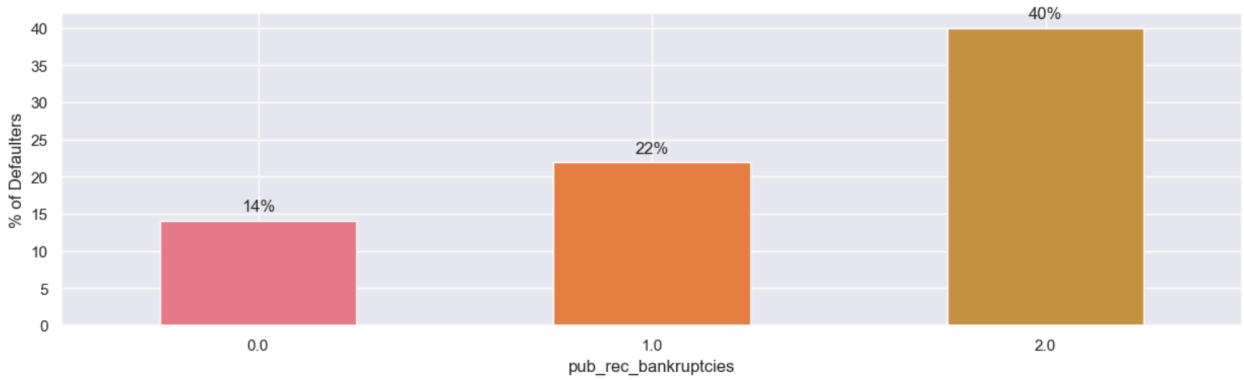
18. DTI - Debt to Income ratio



Inference: Debt-to-Income ratio of the majority of the borrowers is in 10-15 zone which means that the debt is high as compared to Income for these borrowers

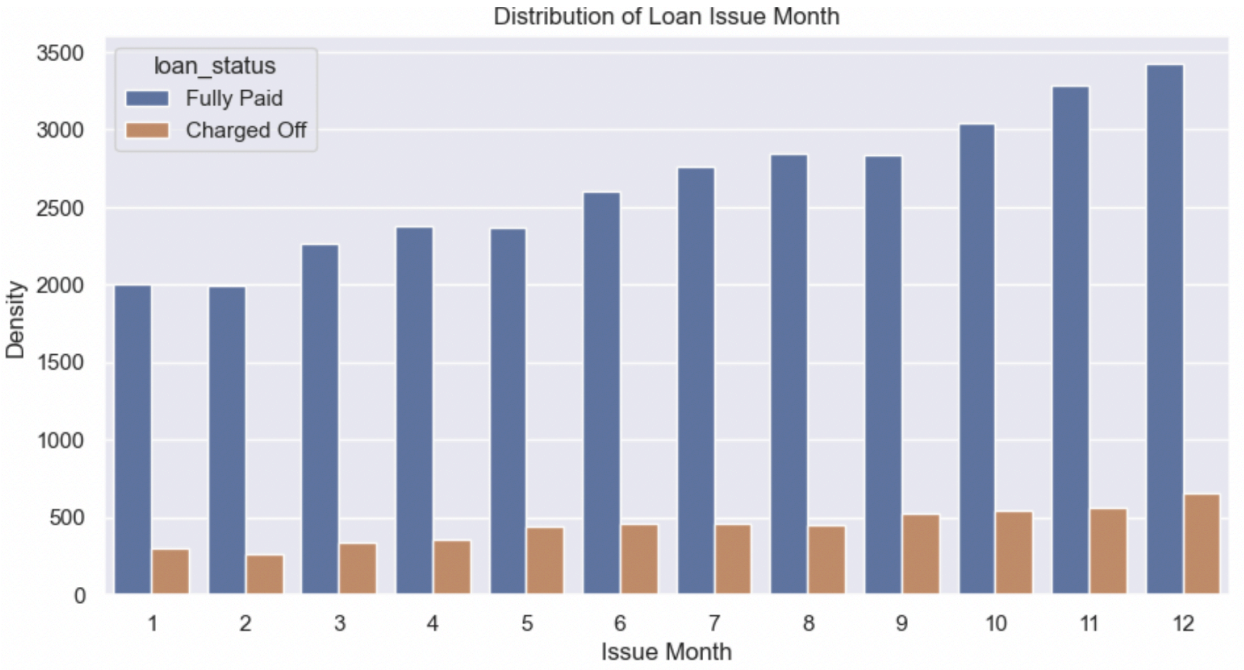
19. Public record of Bankruptcy





Inference: Out of 37173 records 31813 have no public bankruptcy records (85.5%) Rest 5360 (14.5%) have atleast 1 public bankruptcy records Following is the statistics for the same. 15% defaulters have no Public bankruptcy records, 22% defaulters have 1 record, 40% defaulters have 2 public bankruptcy record. It is almost clear that each derogatory record almost doubles the chances of Charge Off.

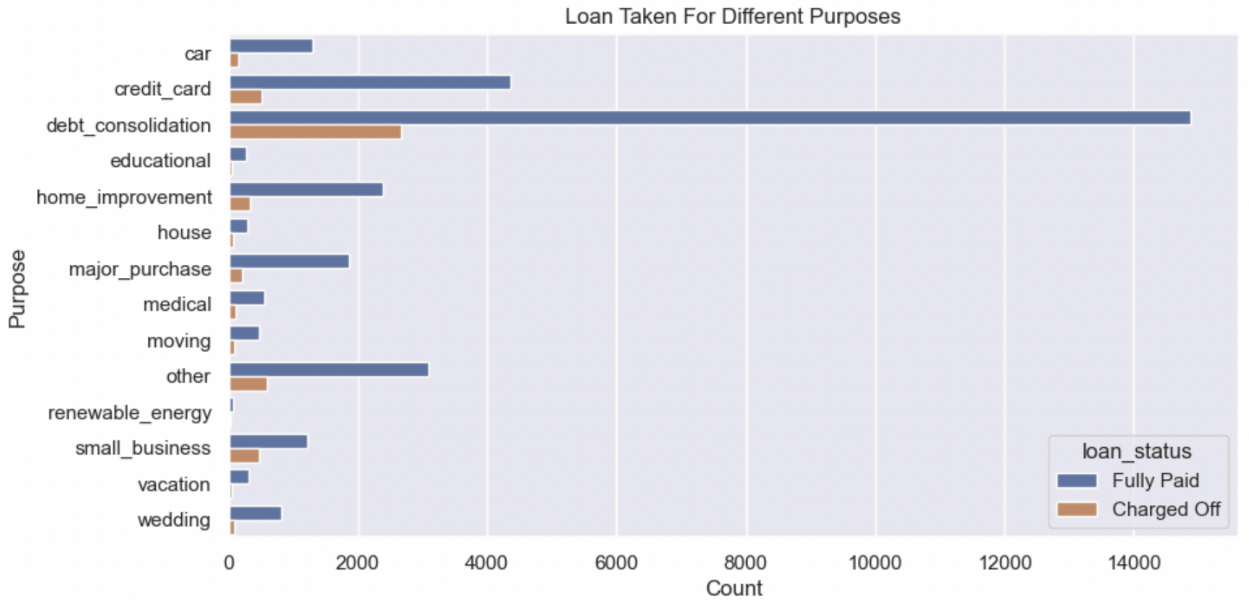
20. Derived Metric - Loan Issue Month



Inference: Last quarter of the year shows the highest disbursal of loans.

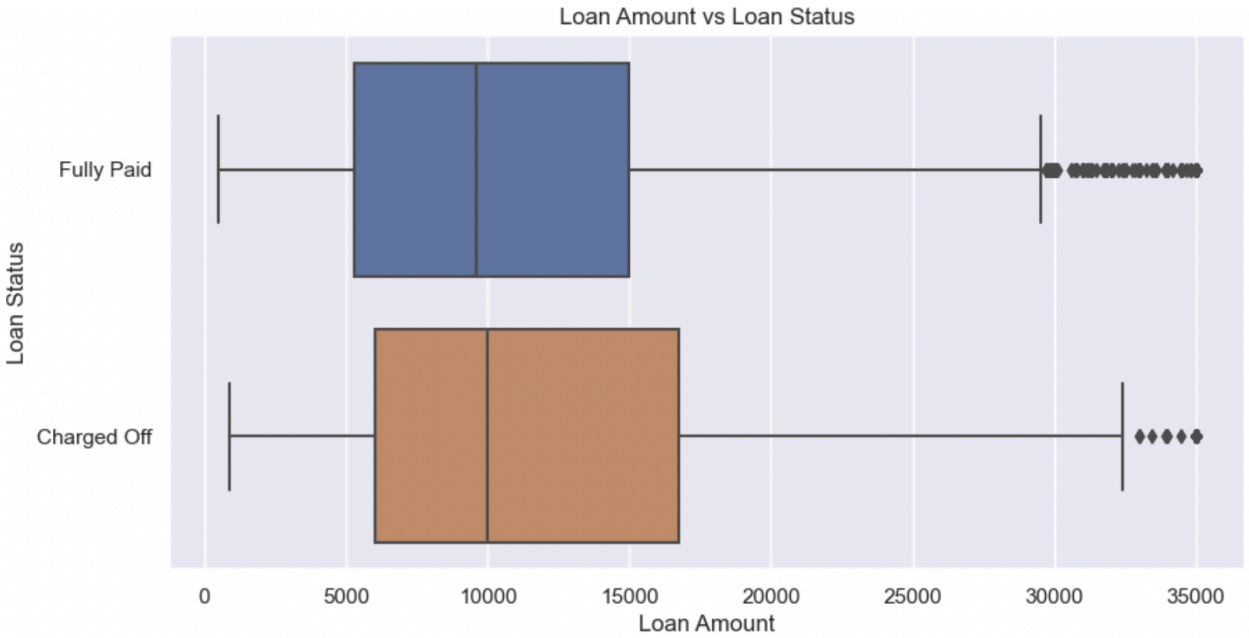
Segmented Univariate Analysis

1. Purpose



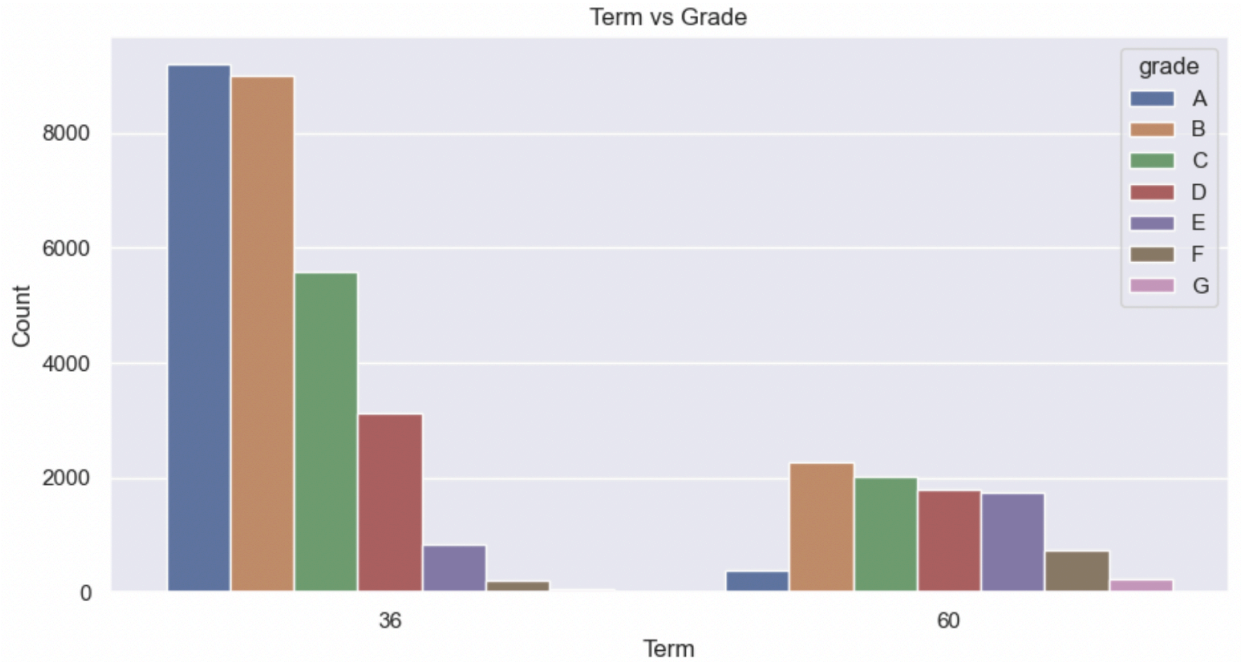
Inference: Debt Consolidation has highest number of both fully paid and defaulted loans. Credit card is the second in the list.

2. Loan Amount vs Loan Status



Inference: While the 25th and mean percentile for Fully paid and Charged off case are same. However, a loan ticket of 10K to 17K shows higher Charged off incidence lying in the 50th to 75th percentile.

3. Loan Term vs Grade

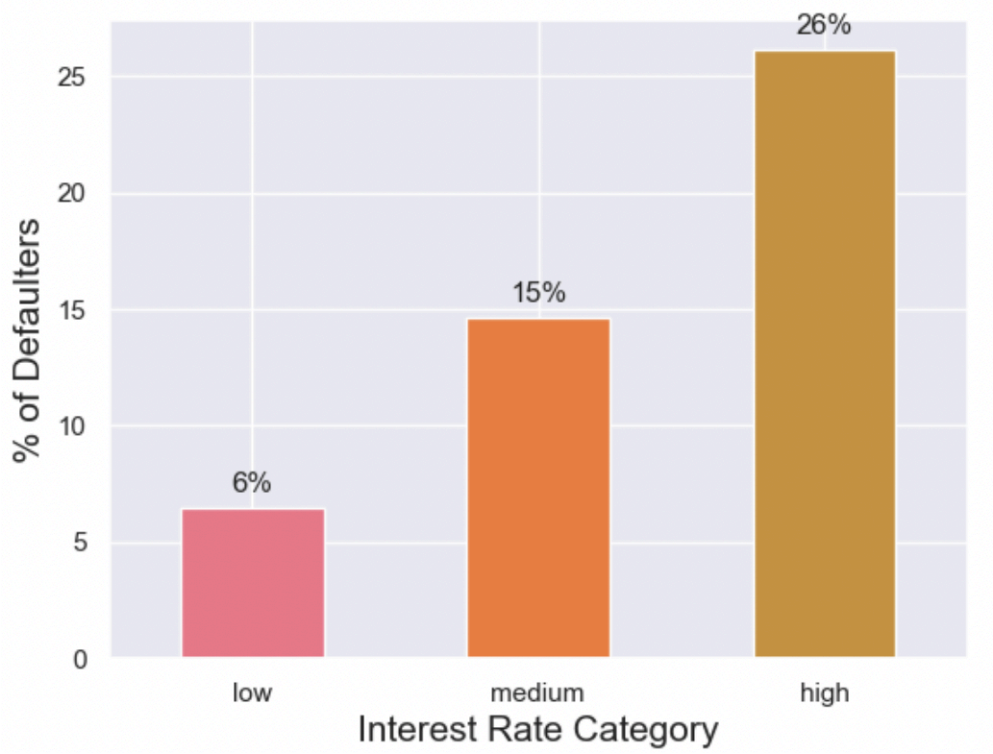


Inference: Higher grade loans of type A,B,C,D are generally taken for shorter term of 36 months as compared to Grade B,C,D,E loans taken for 60 months term

Derived Metrics

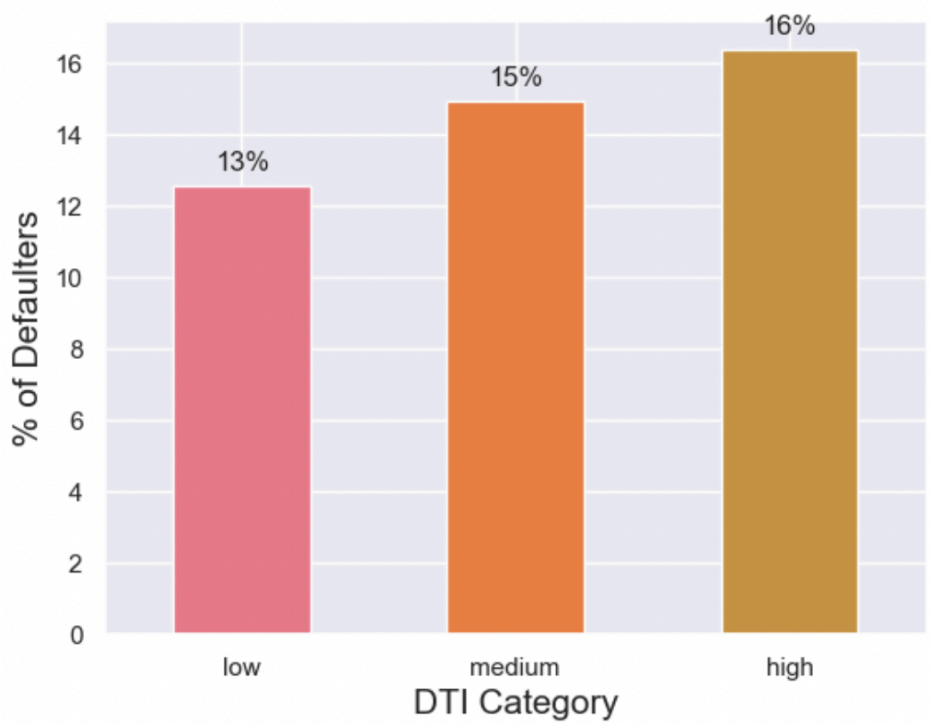
We will create bins or indicators for continous variables and group into discrete categories to group data into categores as low, small, medium, high, very high.

Type Driven Metrics : int_rate_bin



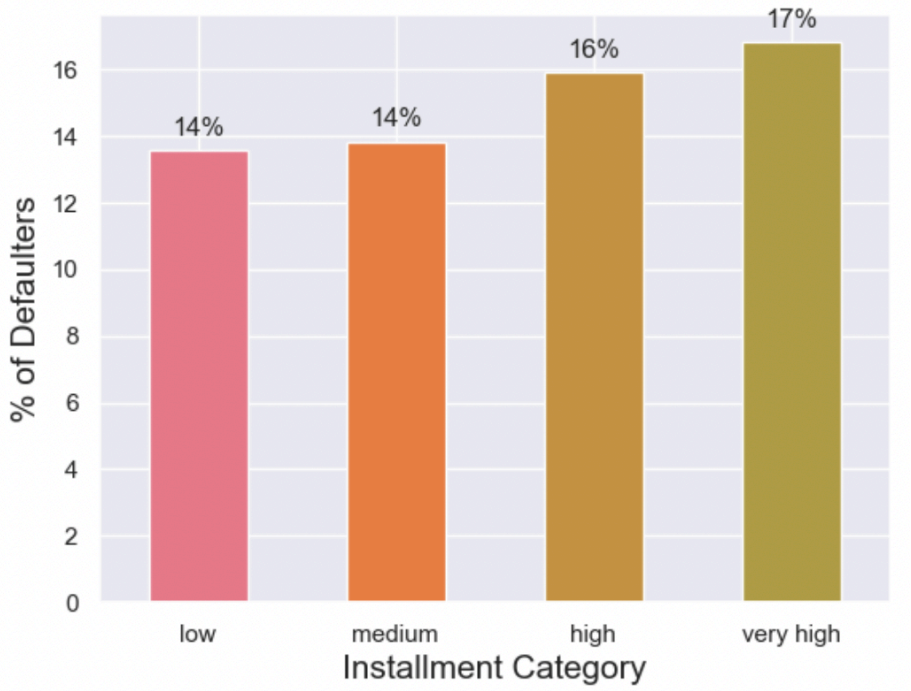
Inference: Borrowers facing Higher interest rate tend to default more as compared to one with lower interest rate.

Type Driven Metrics : dti_bin



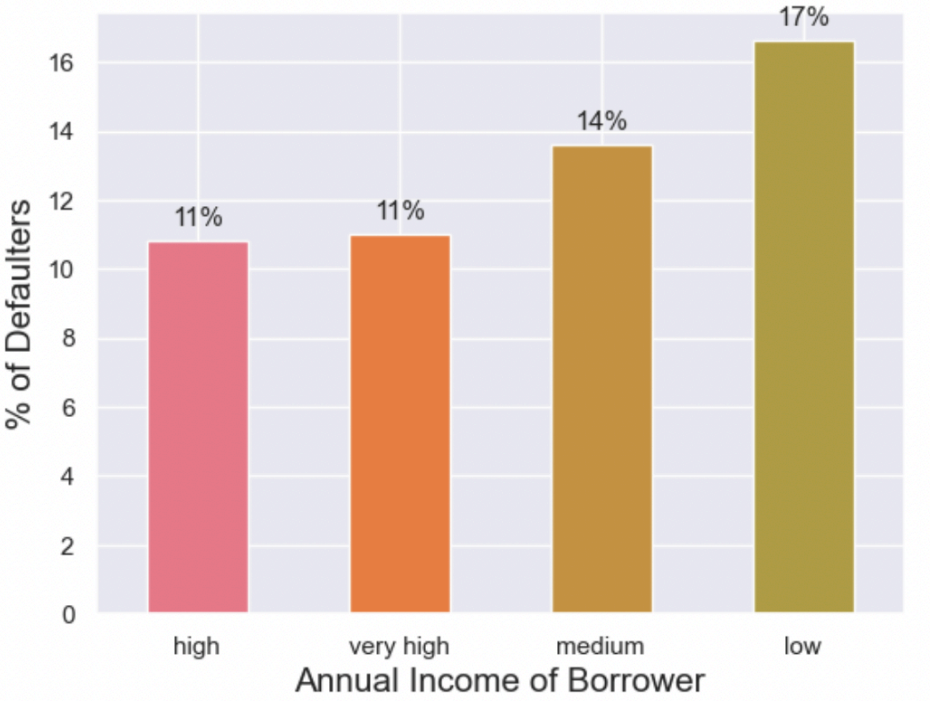
Inference: Borrowers with high DTI(>20) are likely to default more than the borrowers with less DTI(<10) or (<20)

Type Driven Metrics : installment



Inference: There is not much difference in the number of defaulter paying high installments. However, technically, high installments are related to high defaults.

Type Driven Metrics : annual_inc



Inference: Borrowers with low Annual income (in the range < 50K) are likely to default more.

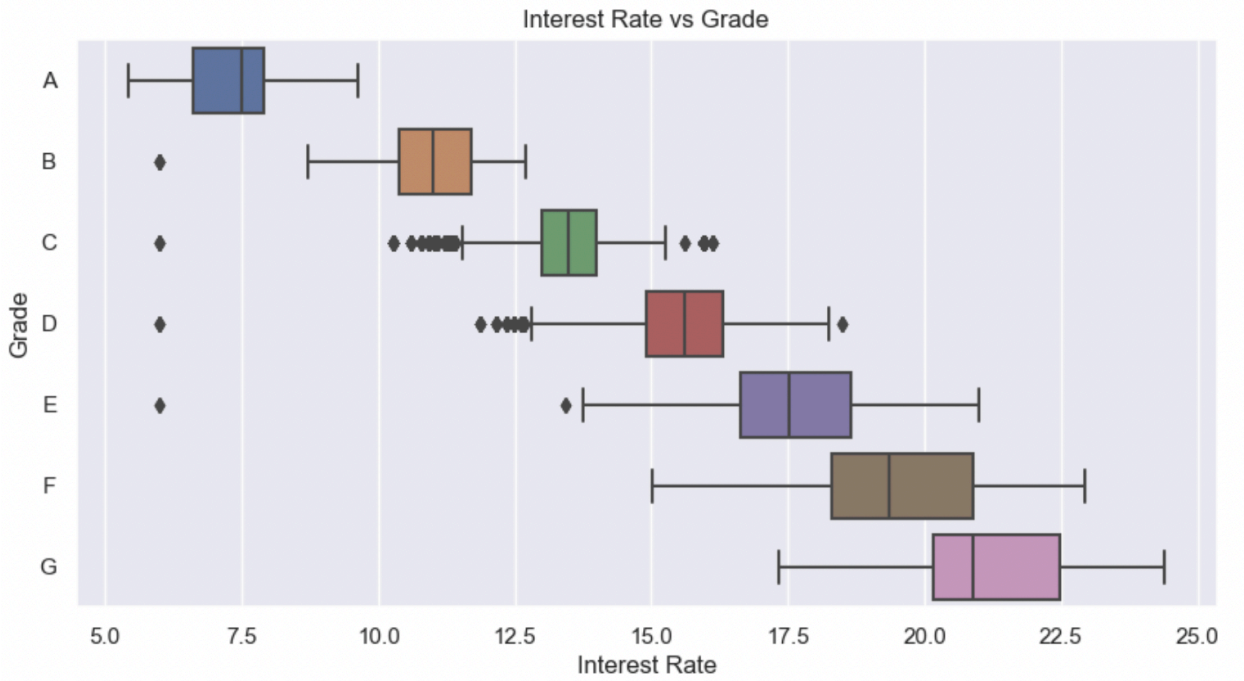
Data points of Interest

As per our univariate analysis, following variables can be considered for bivariate analysis

- **Loan Amount (loan_amnt)**
- **Purpose (purpose)**
- **Home Ownership (home_ownership)**
- **Issue Date (issue_d)**
- **Sub-Grade (sub_grade)**
- **Term (term)**
- **Annual Income (annual_inc)**
- **DTI (dti)**
- **Public Record Bankruptcies (pub_rec_bankruptcies)**
- **Employment Length (emp_length)**

Bivariate Analysis

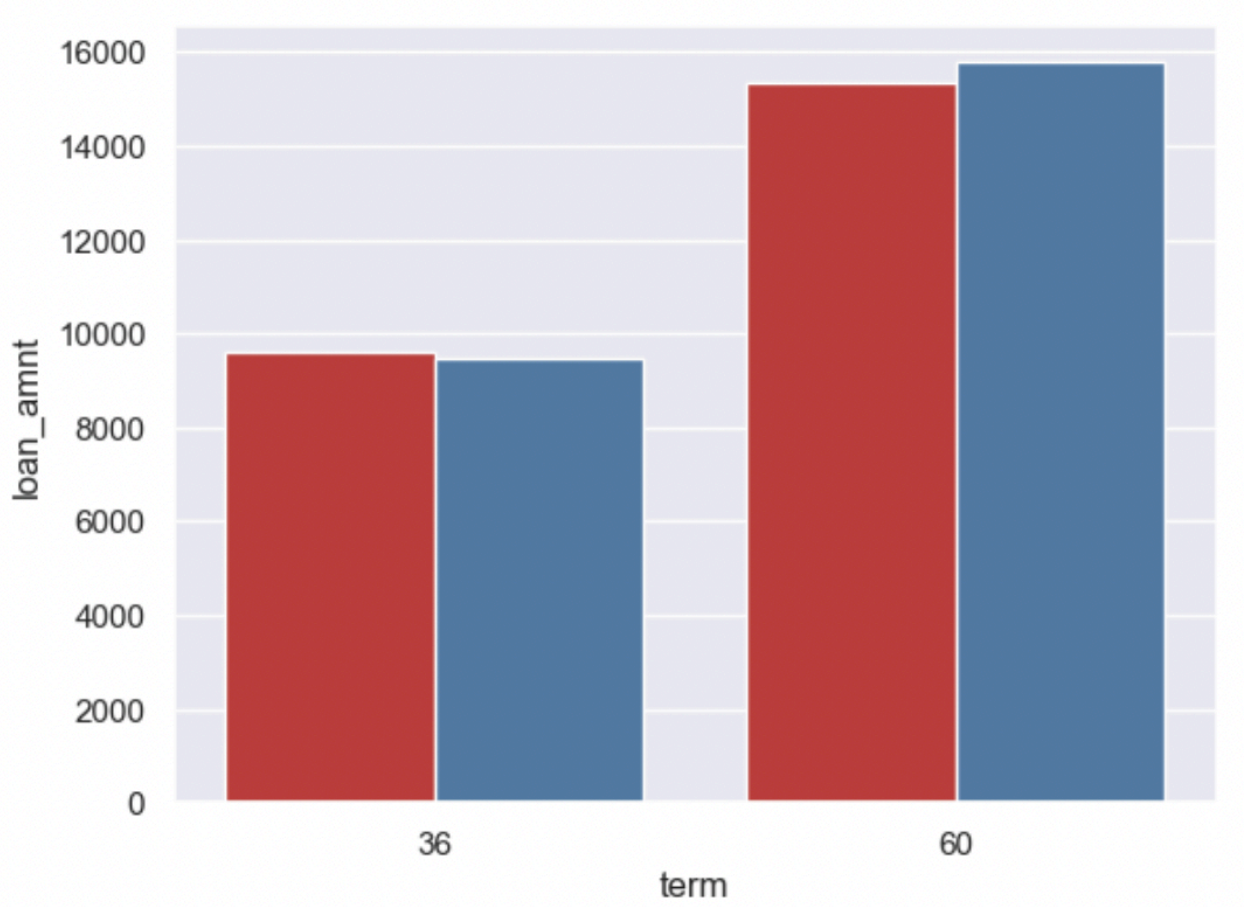
Interest Rate vs Grade



Inference: Lower Grade = Higher Risk. Thus Lower grade with high interst rate increases the risk.

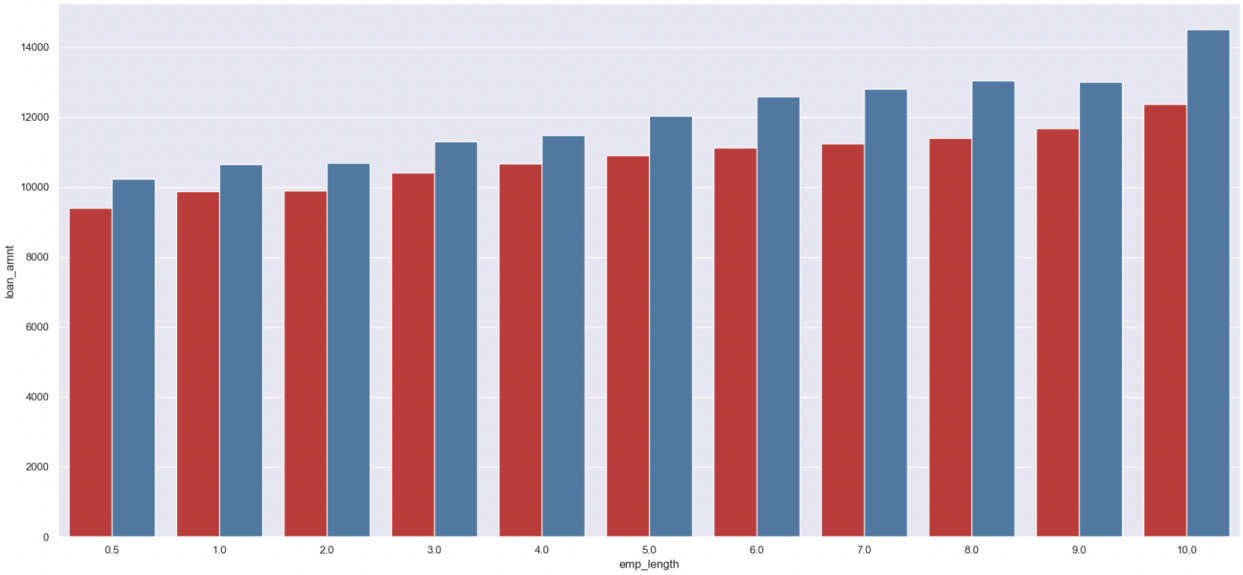
Multi-Variate Analysis

1. **Loan Amount vs Term vs loan_status**



Inference: Higher loan amount with longer terms has more defaulters as well fully paid loans.

2. **Loan Amount vs Employment Length vs loan_status**



Inference:
1. Borrowers with more than 10 years of employment length take large loans.
2. Defaulters who take Large loan amounts have more than 10 years of employment length.

Correlation HeatMap

	loan_amnt	funded_amnt	term	int_rate	installment	grade	emp_length	annual_inc	loan_status	dti	pub_rec_bankruptc
loan_amnt	1.000000	0.981574	0.347447	0.296455	0.931098	0.280973	0.149245	0.412178	0.064965	0.070918	-0.028
funded_amnt	0.981574	1.000000	0.325587	0.300729	0.957259	0.282535	0.148890	0.407337	0.061781	0.070580	-0.029
term	0.347447	0.325587	1.000000	0.439440	0.088707	0.426705	0.105716	0.072589	0.176019	0.079515	0.019
int_rate	0.296455	0.300729	0.439440	1.000000	0.272861	0.947899	-0.000144	0.063253	0.214626	0.114593	0.085
installment	0.931098	0.957259	0.088707	0.272861	1.000000	0.258188	0.121445	0.408688	0.031707	0.060419	-0.026
grade	0.280973	0.282535	0.426705	0.947899	0.258188	1.000000	-0.000507	0.066128	0.204583	0.100309	0.078
emp_length	0.149245	0.148890	0.105716	-0.000144	0.121445	-0.000507	1.000000	0.172843	0.016942	0.052581	0.063
annual_inc	0.412178	0.407337	0.072589	0.063253	0.408688	0.066128	0.172843	1.000000	-0.059898	-0.112099	-0.010
loan_status	0.064965	0.061781	0.176019	0.214626	0.031707	0.204583	0.016942	-0.059898	1.000000	0.041832	0.044
dti	0.070918	0.070580	0.079515	0.114593	0.060419	0.100309	0.052581	-0.112099	0.041832	1.000000	0.007
pub_rec_bankruptcies	-0.028905	-0.029832	0.019322	0.085265	-0.026723	0.078934	0.063764	-0.010598	0.044683	0.007245	1.000

Correlation heat map of Correlated variables

Key Points:

High correlation is observed between

1. funded_amnt and loan_amnt
2. Loan_amnt and installment
3. Grade and interest rate

Analysis Results collated

Univariate Analysis

- Majority of borrowers don't possess property and are on mortgage or rent.
- About 50% of the borrowers are verified by the company or have source verified.
- Annual Income shows left skewed normal distribution thus we can say that the majority of borrowers have very low annual income compared to rest.
- The number of charged off loan is 7 times less than the number of fully paid loan.
- The majority of loan has a term of 36 months compared to 60 months.
- The interest rate is more crowded around 5-10 and 10-15 with a drop near 10.
- A large amount of loans are with grade 'A' and 'B' compared to the rest showing most loans are high grade loans.
- Majority of borrowers have working experience greater than 10 years.
- A large percentage of loans are taken for debt consolidation followed by credit card.
- Majority of the borrowers are from the large urban cities like california, new york, texas, florida etc.
- Majority of the borrowers have very large debt compared to the income registered, concentrated in the 10-15 DTI ratio.
- Majority of the borrowers have no record of Public Recorded Bankruptcy.
- Majority of the loans are given in the last quarter of the year.
- The number of loans approved increases with the time at exponential rate, thus we can say that the loan approval rate is increasing with the time.

Segmented Univariate Analysis

- Borrowers with 10+ years of experience are likely to default and have a higher chance of fully paying the loan.
- The 60 month term has a higher chance of defaulting than the 36 month term whereas the 36 month term has a higher chance of fully paid loan.
- The loans in the 36 month term majorly consist of grade A and B loans whereas the loans in 60 month term mostly consist of grade B, C and D loans.
- The Loan Status varies with DTI ratio, we can see that the loans in DTI ratio 10-15 have higher number of defaulted loans but higher dti has higher chance of defaulting.
- The Defaulted loans are lower for the borrowers which own their property compared to on mortgage or rent.
- borrowers with less than 50000 annual income are more likely to default and higher annual income are less likely to default.
- The mean and 25% are the same for both but we see a larger 75% in the defaulted loan which indicates a large amount of loan has a higher chance of defaulting.
- Debt Consolidation is the most popular loan purpose and has the highest number of fully paid loans and defaulted loans.
- Fully paid loans are increasing exponentially with the time compared to defaulted loans.
- The default loan amount increases with interest rate and shows decline after 17.5 % interest rate.

Bivariate Analysis

- The Grade represents the risk factor. Lower grade high risk of default.
- The borrowers with no record of Public Recorded Bankruptcy can be a safe choice for lending.

Recommendations

- Data points that can be used to predict whether there will be a default and avoiding Credit Loss:
    - DTI
    - Grades
    - Verification Status
    - Annual income
    - Pub_rec_bankruptcies
- With respect to borrowers following considerations can be considered as risk for 'defaults' :
    - Should not be from large urban cities like california, new york, texas, florida etc.
    - not having annual income in the range 50000-100000.
    - not having Public Recorded Bankruptcy.
    - have least grades like E,F,G which indicates high risk.
    - Have very high Debt to Income value.
    - Have working experience of 10+ years.