

# **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.

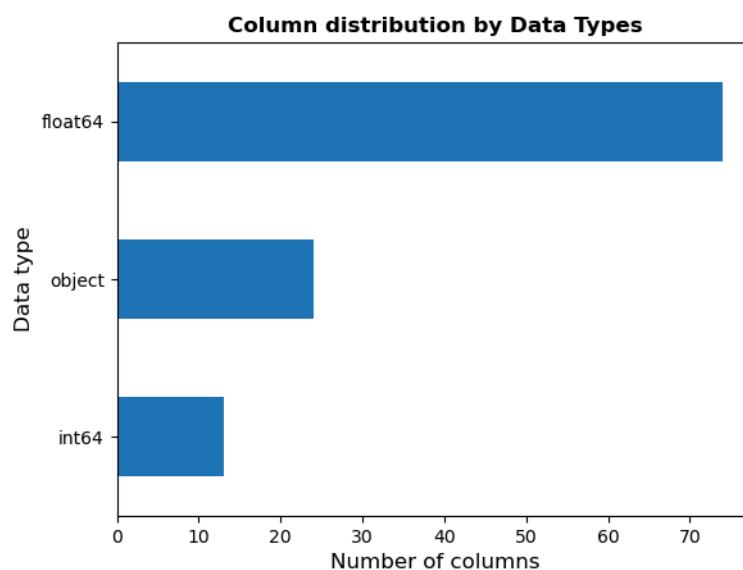
## **Data Source:**

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

## 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

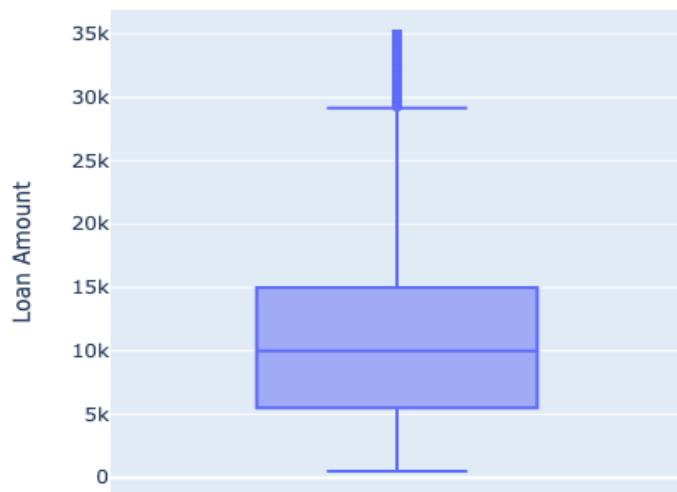
**Outlier check was performed on following continuous variables and handled for Annual income to make data uniformly distributed**

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

# Univariate Analysis and Outlier Handling

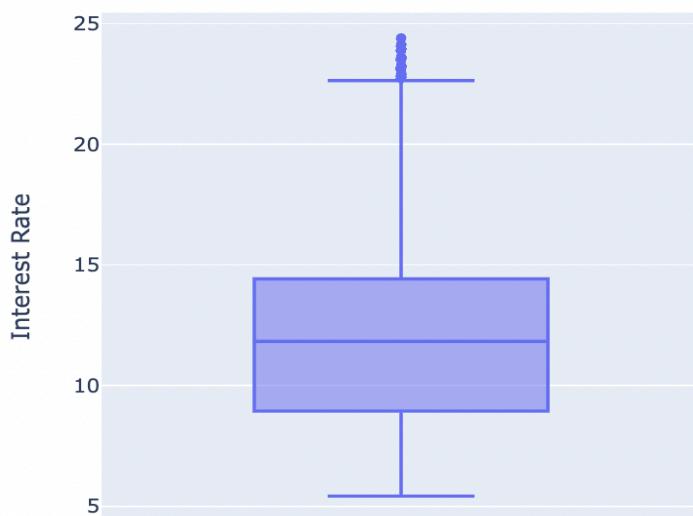
## 1. DTI - Debt to Income ratio

Distribution of Loan Amount



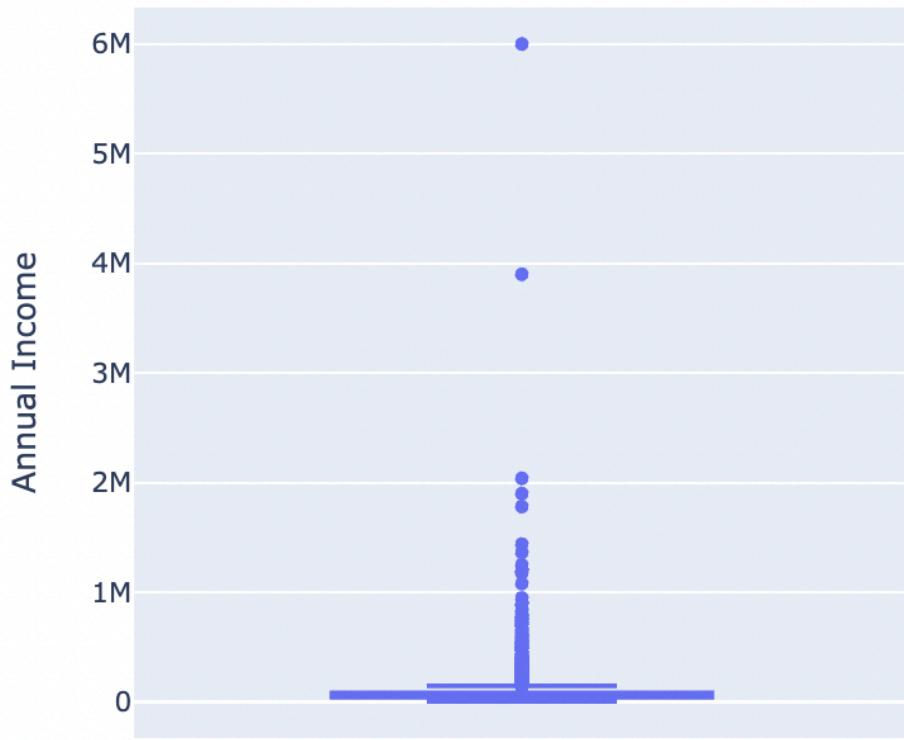
## 2. Interest Rate

Distribution of Interest Rate

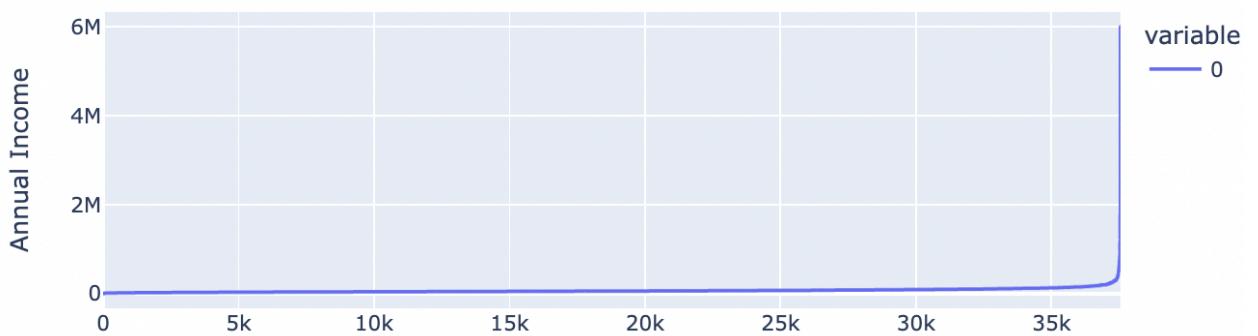


### 3. Annual Income with Outliers in raw data

Borrower Annual Income

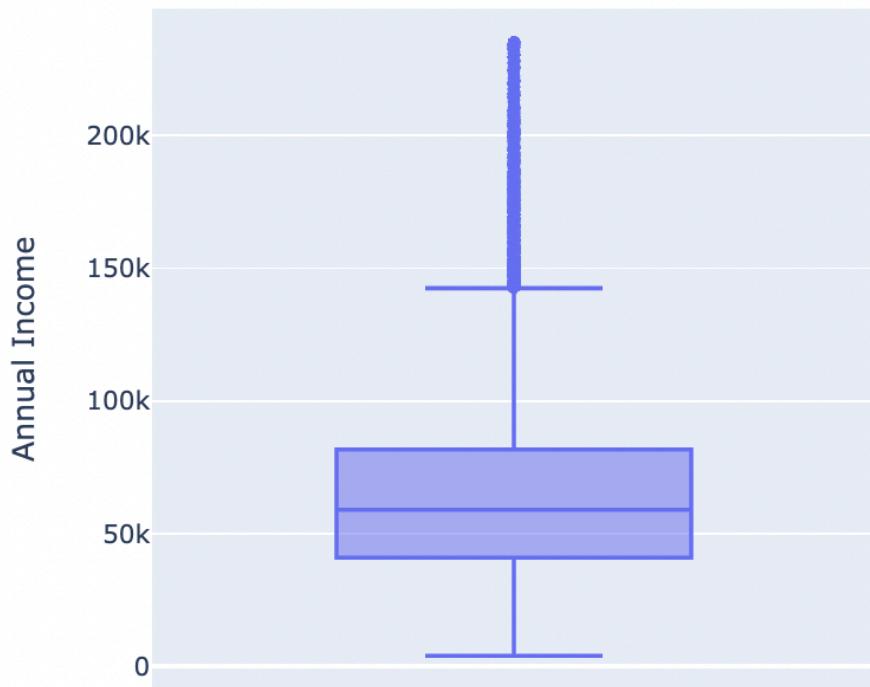


Annual Income

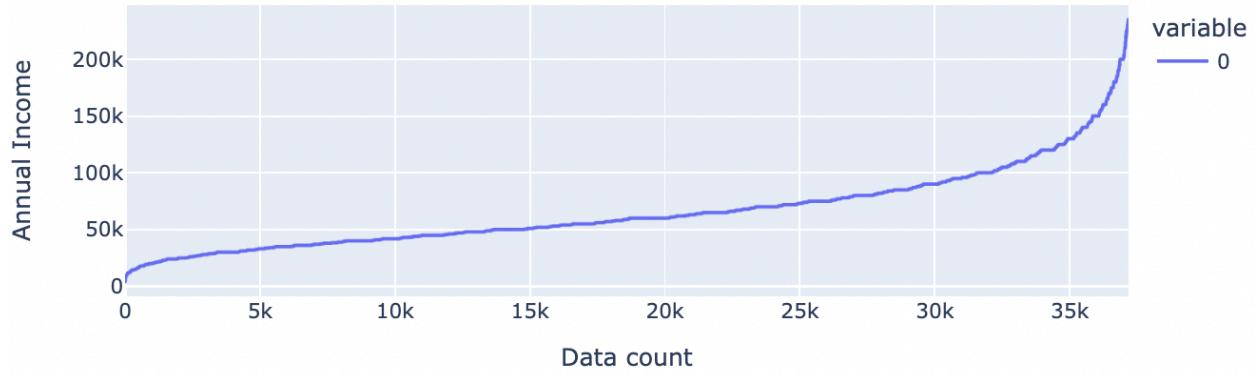


#### 4. Annual Income with 99th percentile outlier removed

Borrower Annual Income

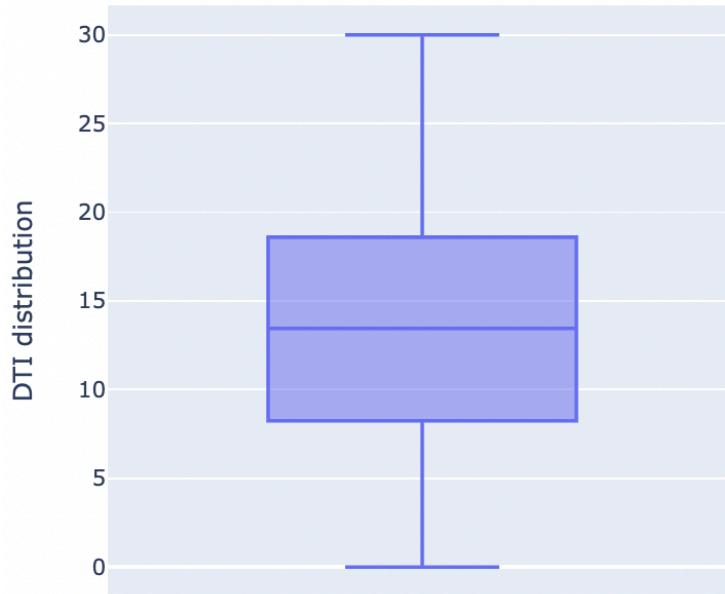


Annual Income



## 5. DTI - Debt to Income Ratio

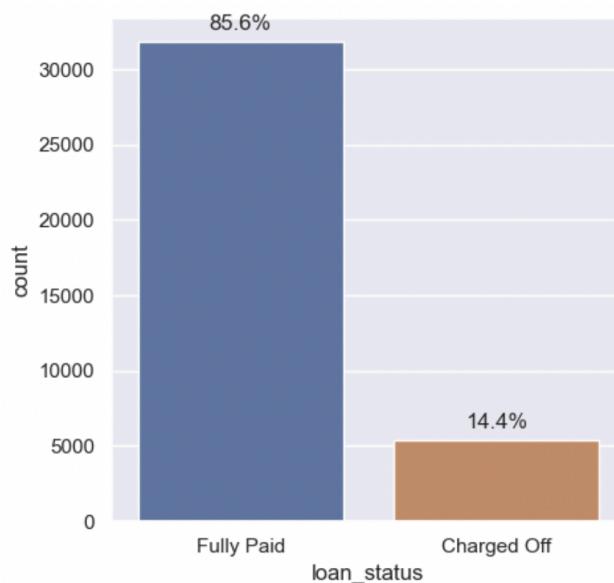
Debt To Income Ratio



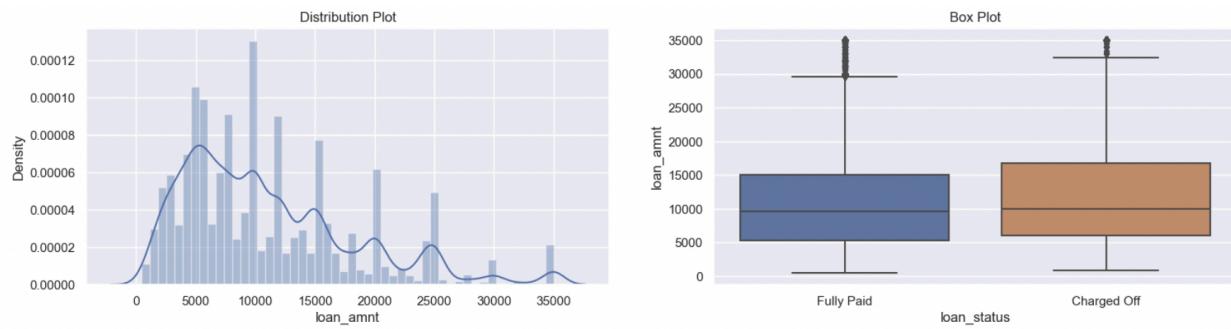
## 6. Loan\_status

The target column for our analysis.

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

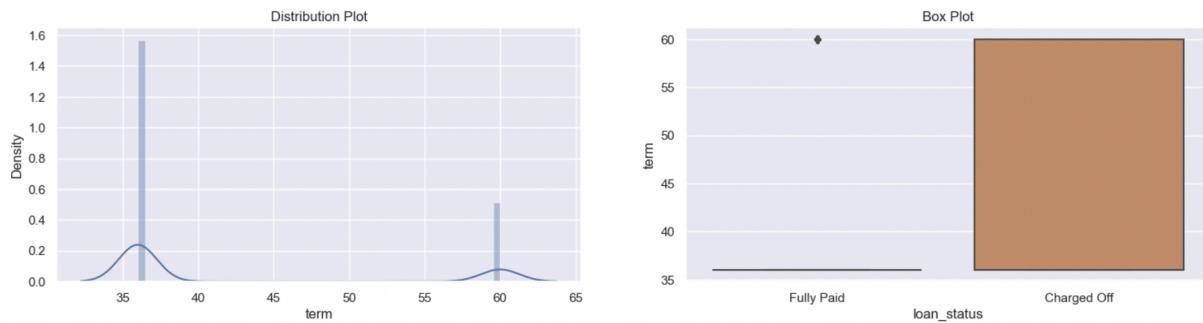


## 7. loan amount



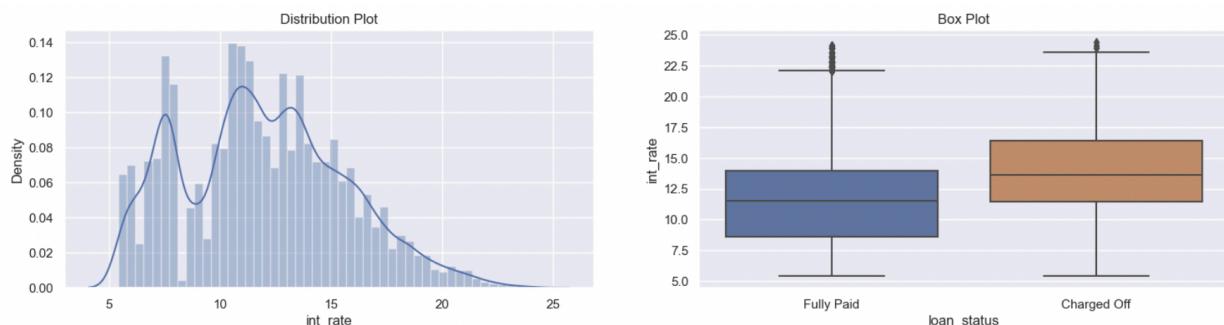
**Inference:** The loan amount varies from 500 to 35000 with a mean of 9800.

## 8. term



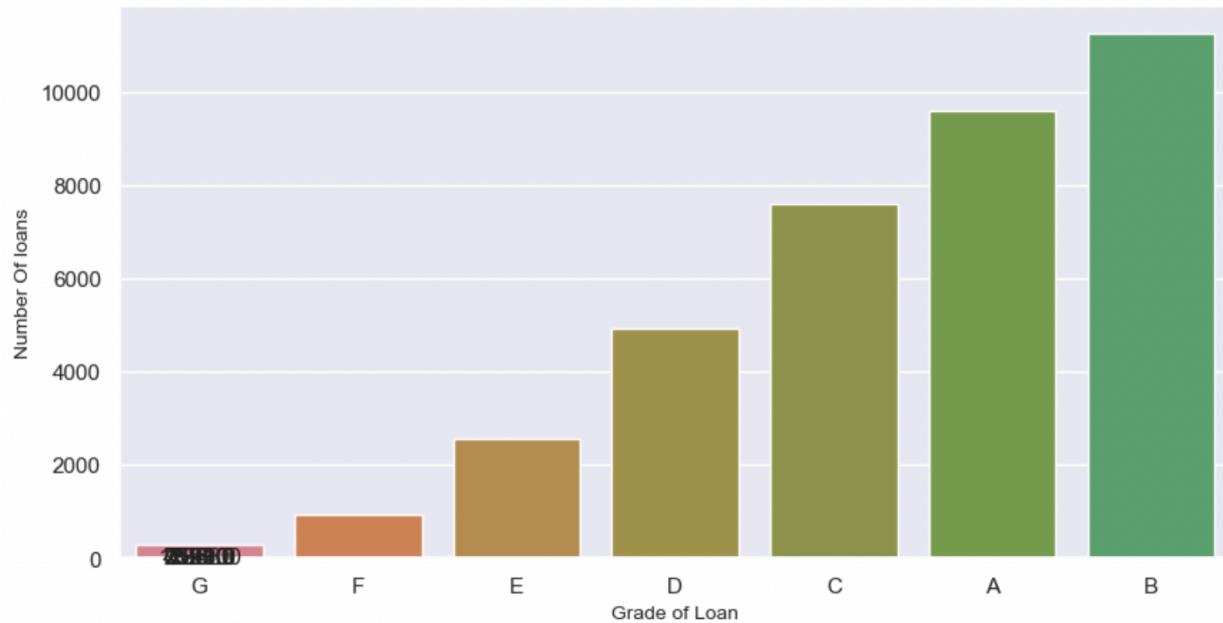
**Inference:** 75% of the loan taken has term of 36 months compared to 60 months.

## 9. Interest rate



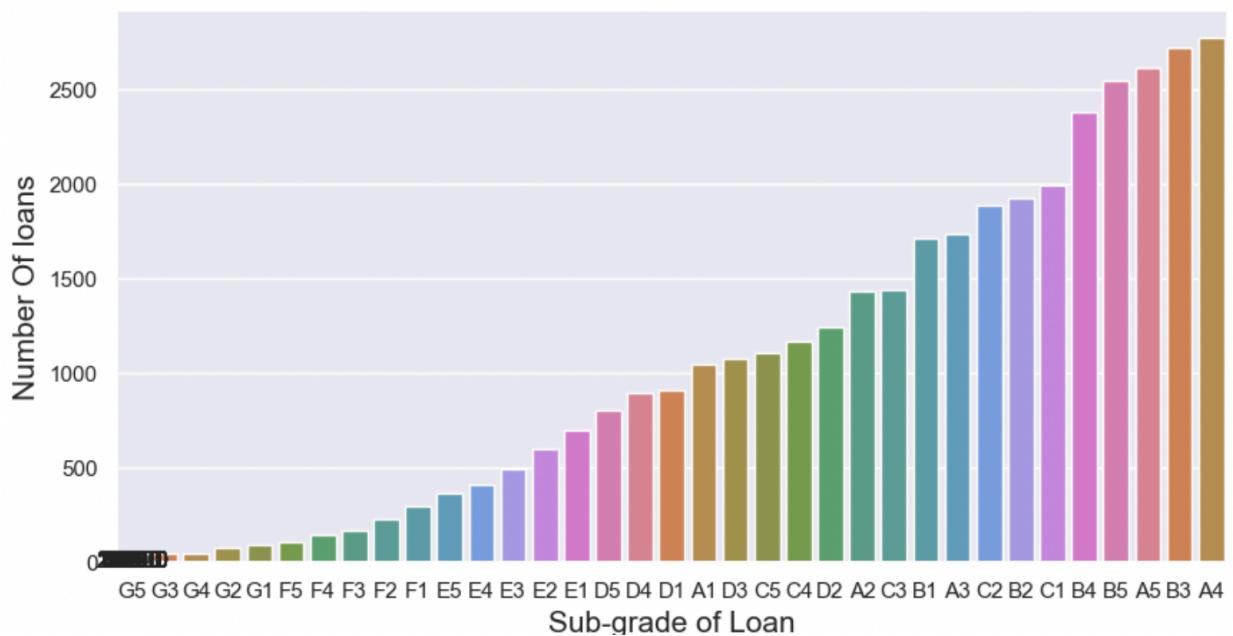
**Inference:** Most Loans are issued at an interest rate of around 5-10 and 10-15 with a drop near

## 10. Grade



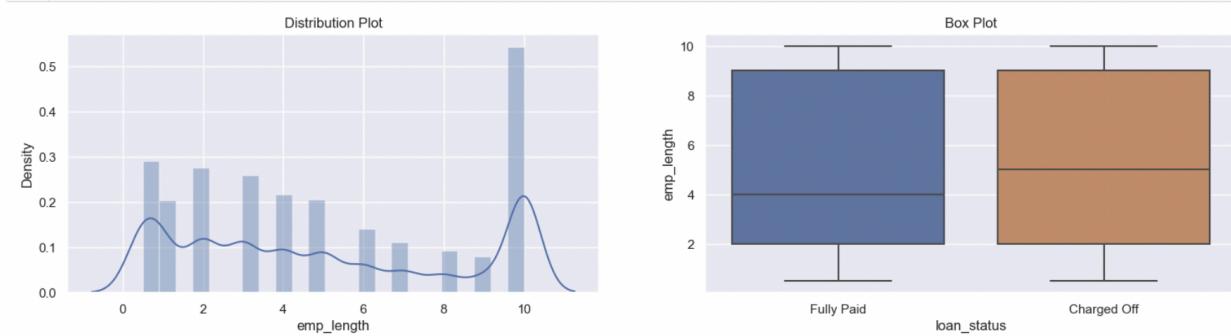
**Inference:** High Grade Loans are high in numbers with Grade B being highest in number.

## 11. Sub-grade



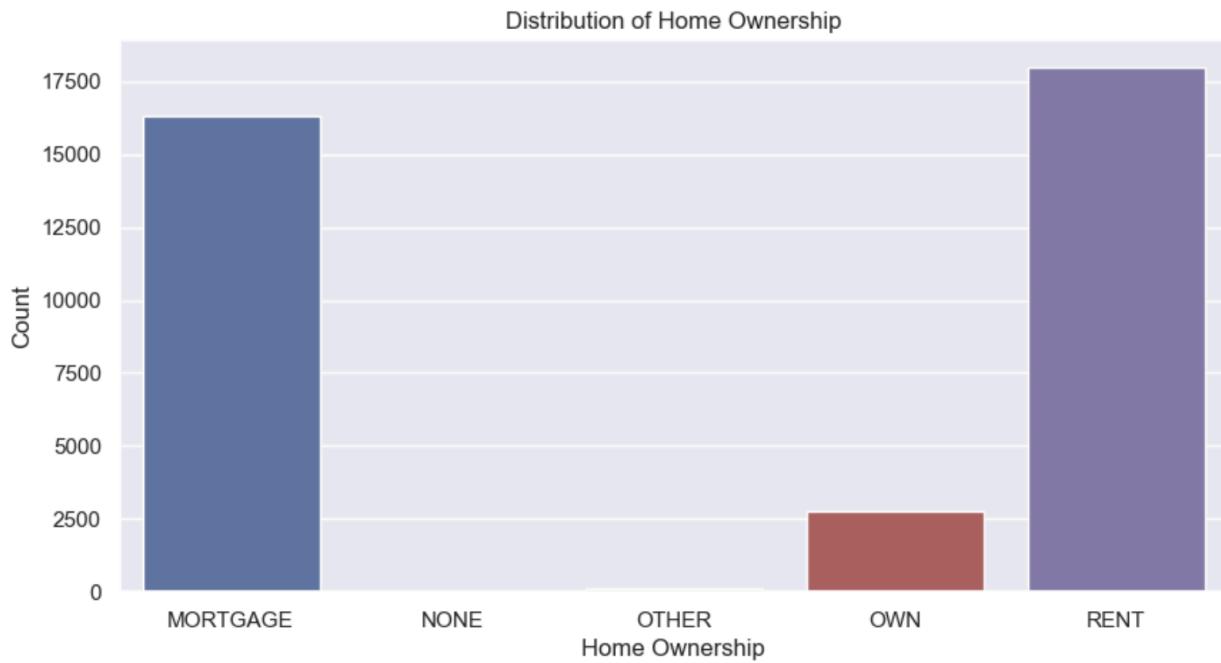
**Inference:** Higher subgrades in each grade just confirm that higher grade loans are higher in number

## 12. Employment Length



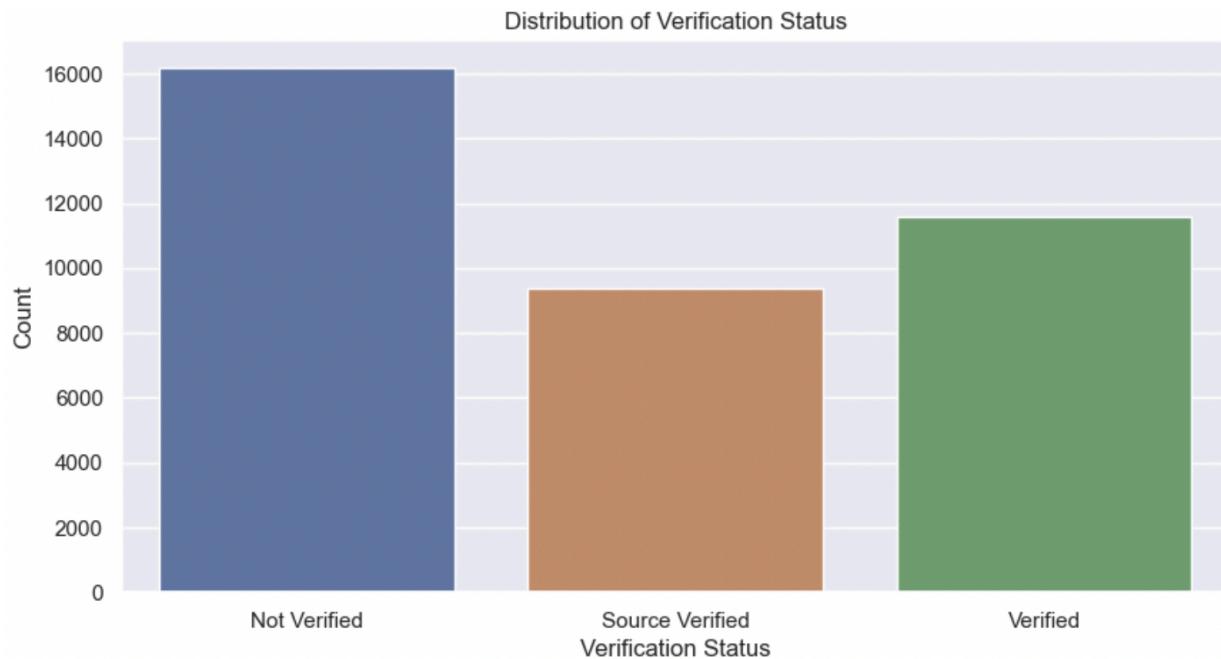
**Inference:** Most borrowers are the ones having highest employment length with 10+ years being the most common.

## 13. Home ownership



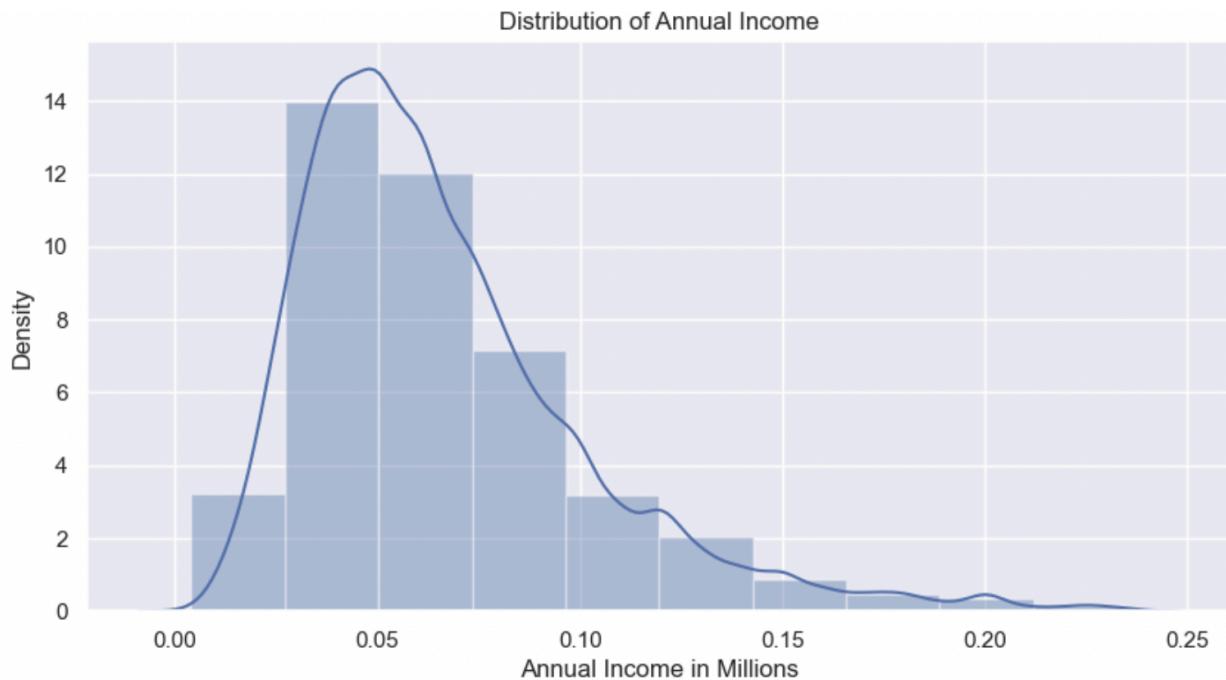
**Inference:** Borrowers living on rented accomodation are most common followed by those living in mortgaged accomodation. Owners are the least borrowers.

## 14. Verification status



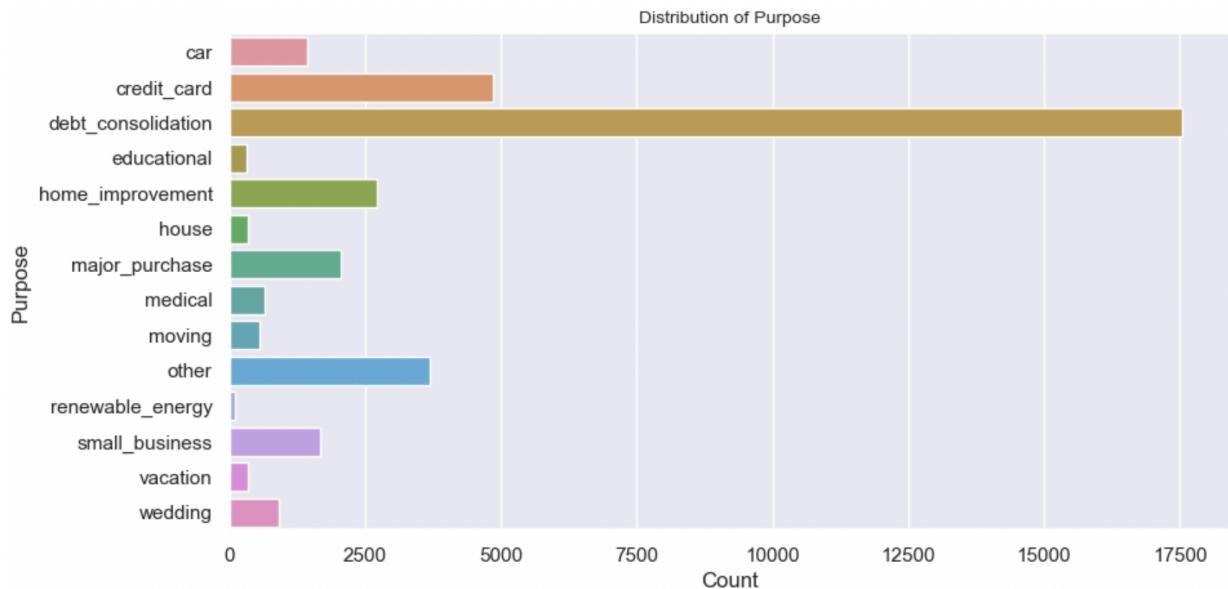
**Inference: Most borrowers have been verified by company or have the source verified**

## 15. Annual Income



**Inference: Majority of borrowers have very low annual income.**

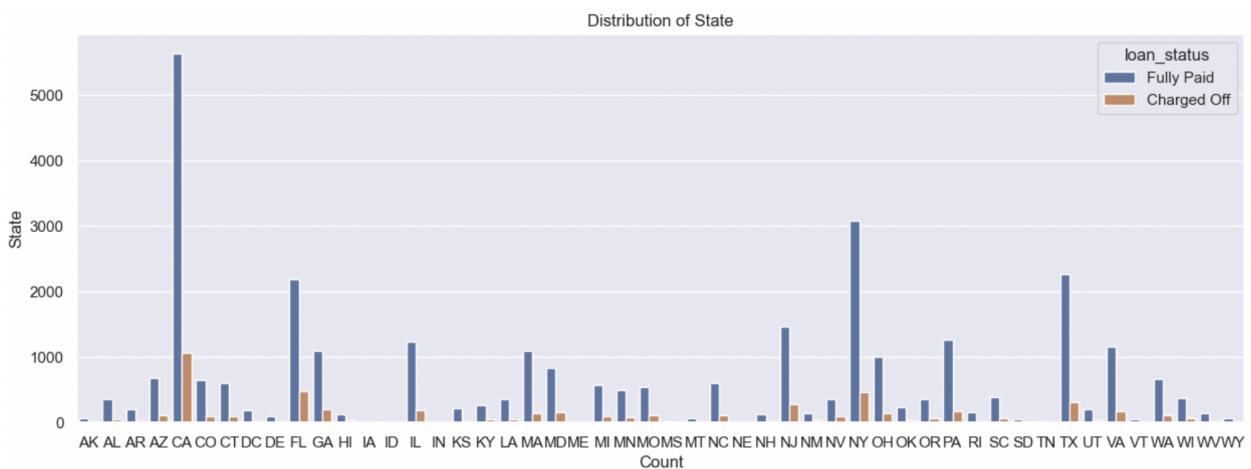
## 16. Purpose



**Inference: Borrowers are taking more loan to consolidate their existing debts.**

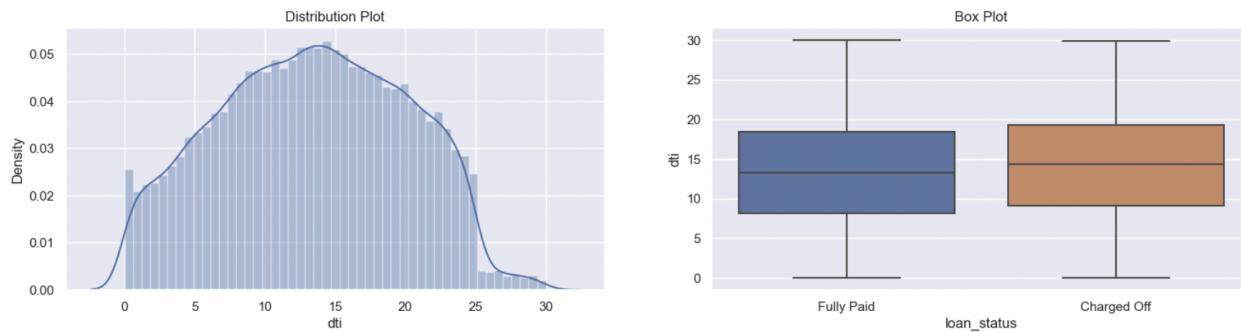
**Debt\_consolidation is the most specified reason for taking loan**

## 17. Address State



**Inference: Most of the borrowers are from states California, new york, Texas, florida.**

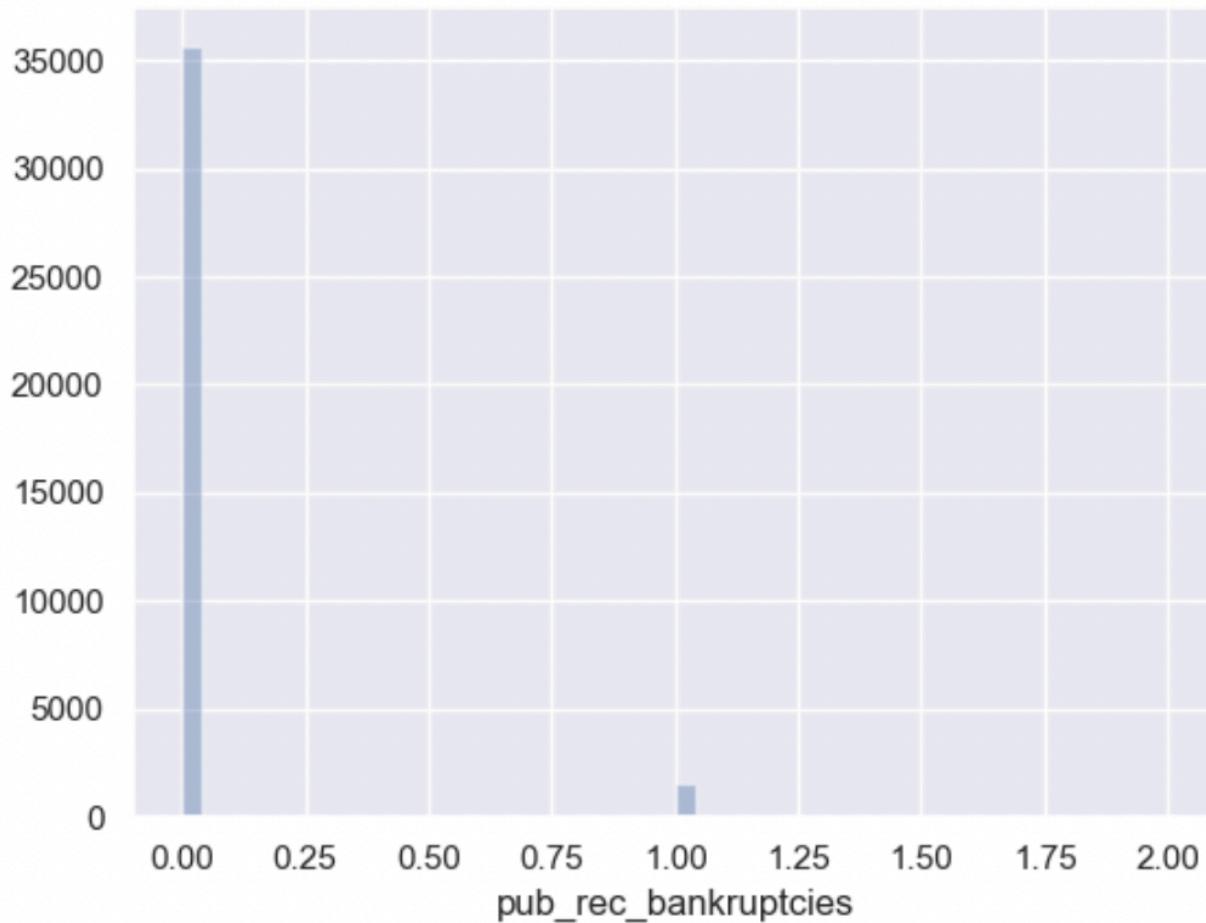
## 18. DTI - Debt to Income ratio

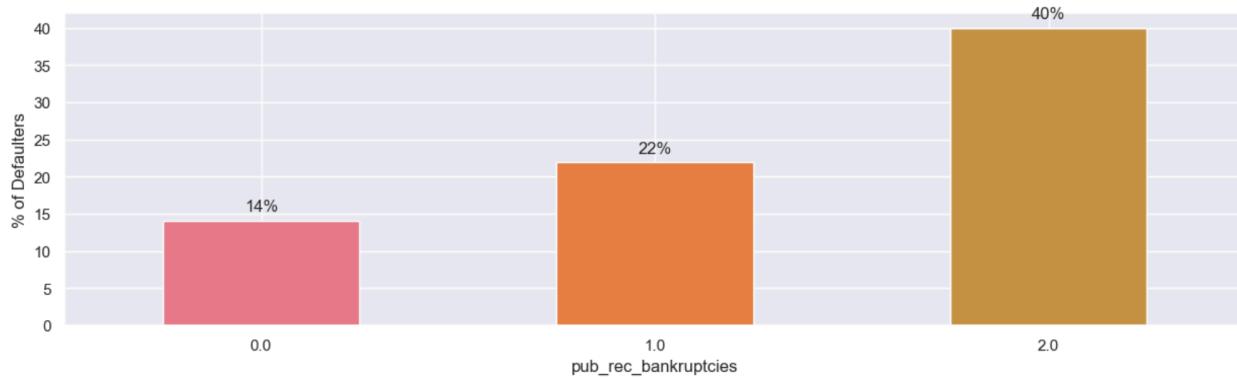


**Inference:** Debt-to-Income ratio of the majority of the borrowers is in 10-15 zone which means that the debt is high as compared to Income for these borrowers

## 19. Public record of Bankruptcy

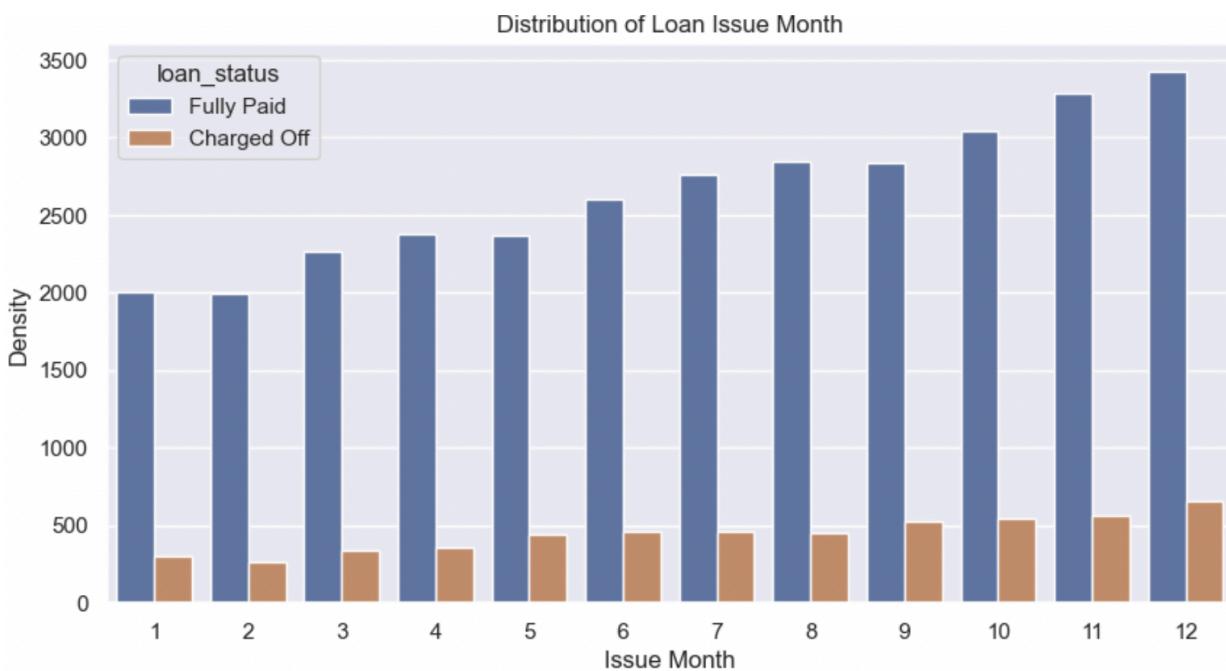
Public record of Bankruptcies





**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.

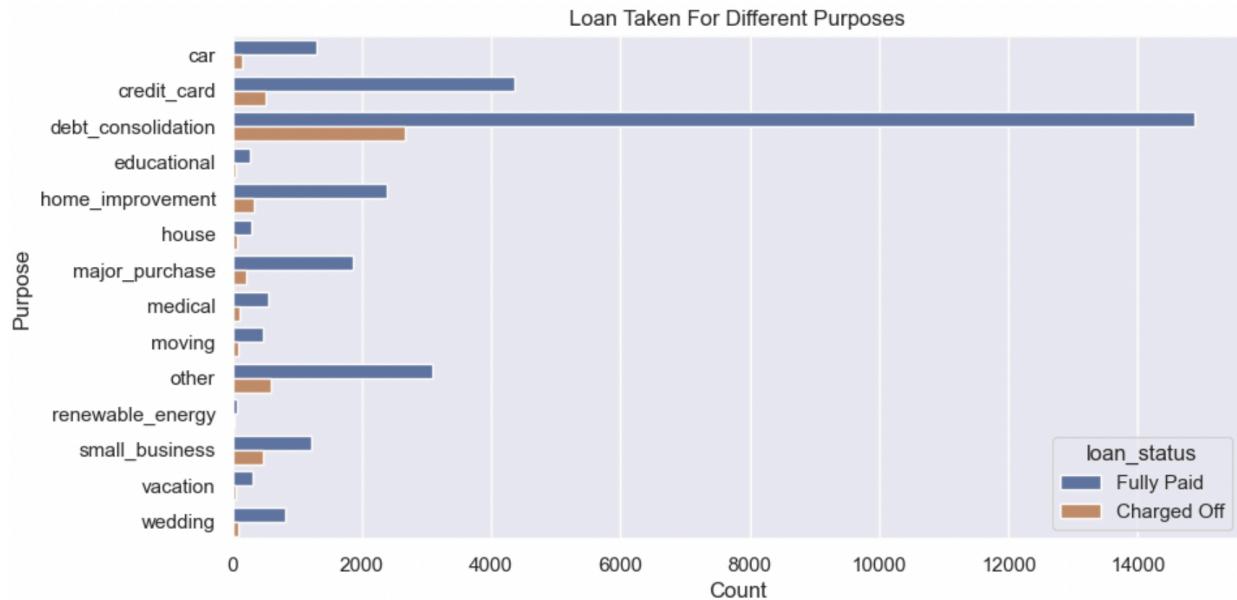
## 20. Derived Metric - Loan Issue Month



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

# Segmented Univariate Analysis

## 1. Purpose



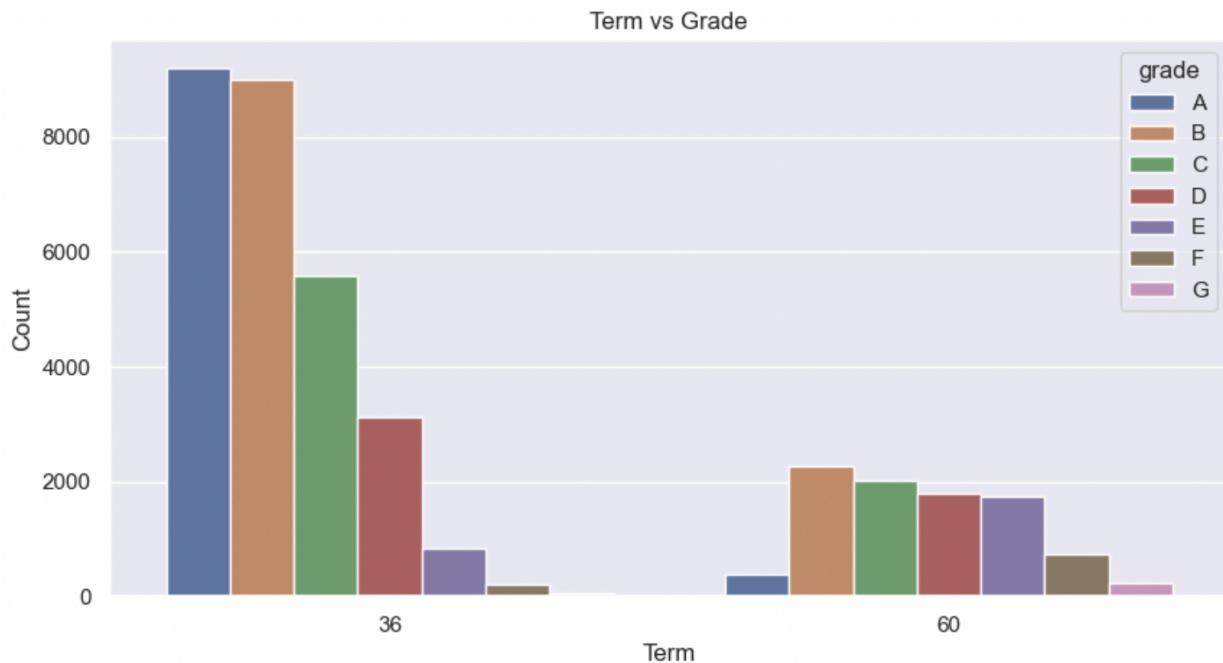
**Inference:** Debt Consolidation has highest number of both fully paid and defaulted loans.  
Credit card is the second in the list.

## 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

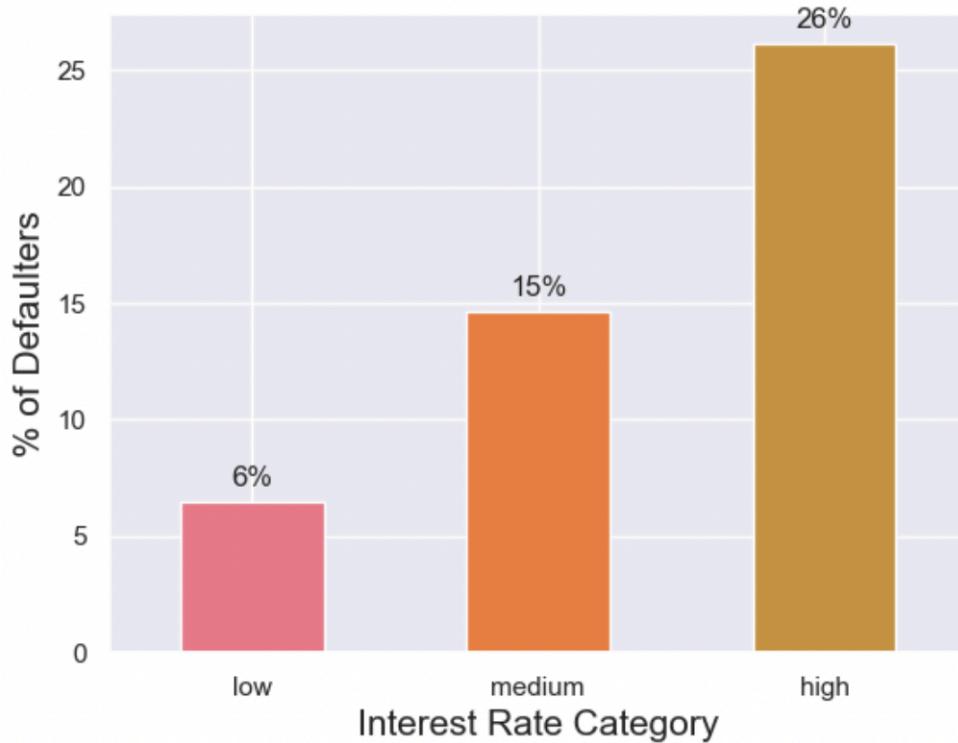


**Inference:** Higher grade loans of type A,B,C,D are generally taken for shorter term of 36 months as compared to Grade B,C,D,E loans taken for 60 months term

## Derived Metrics

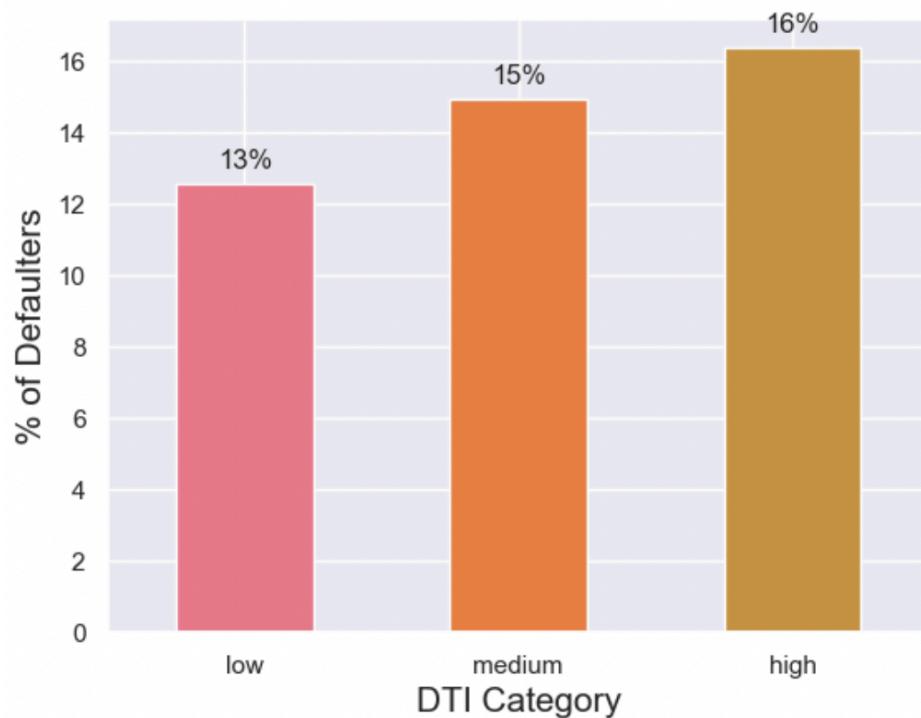
We will create bins or indicators for continuous variables and group into discrete categories to group data into categories as low, small, medium, high, very high.

Type Driven Metrics : int\_rate\_bin



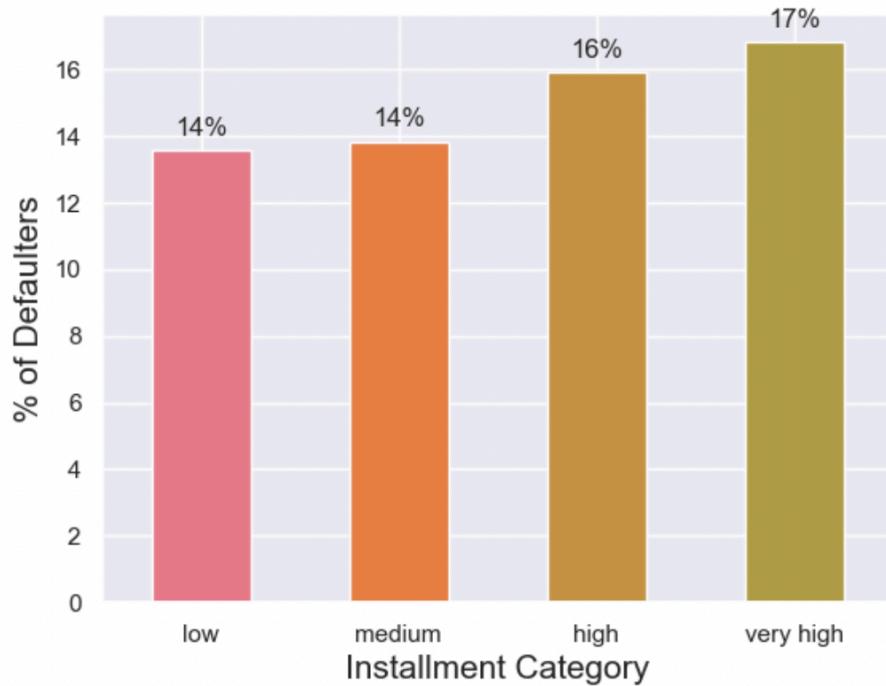
**Inference:** Borrowers facing Higher interest rate tend to default more as compared to one with lower interest rate.

### Type Driven Metrics : dti\_bin



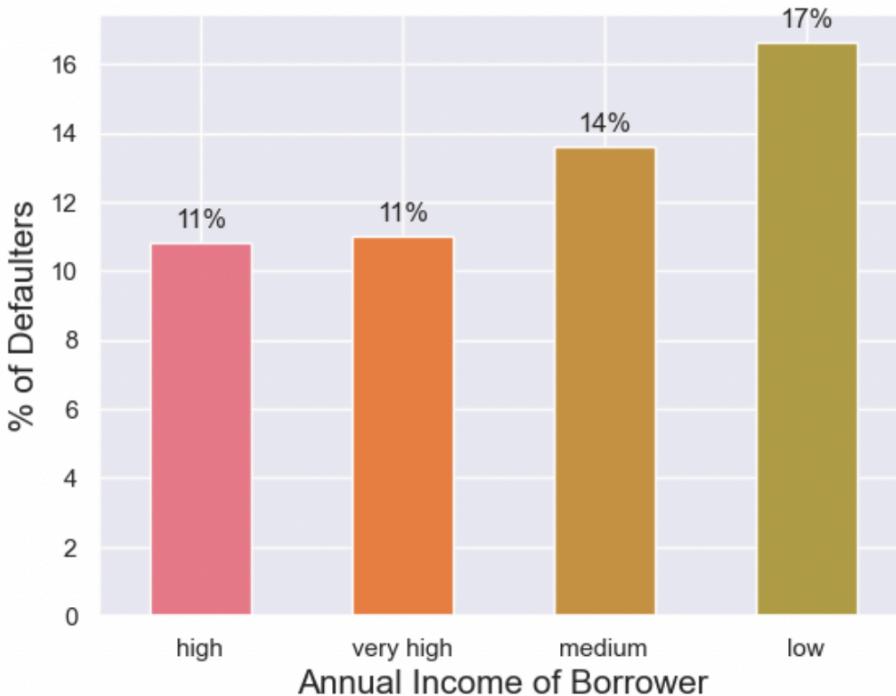
**Inference:** Borrowers with high DTI(>20) are likely to default more than the borrowers with less DTI(<10) or (<20)

#### 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.

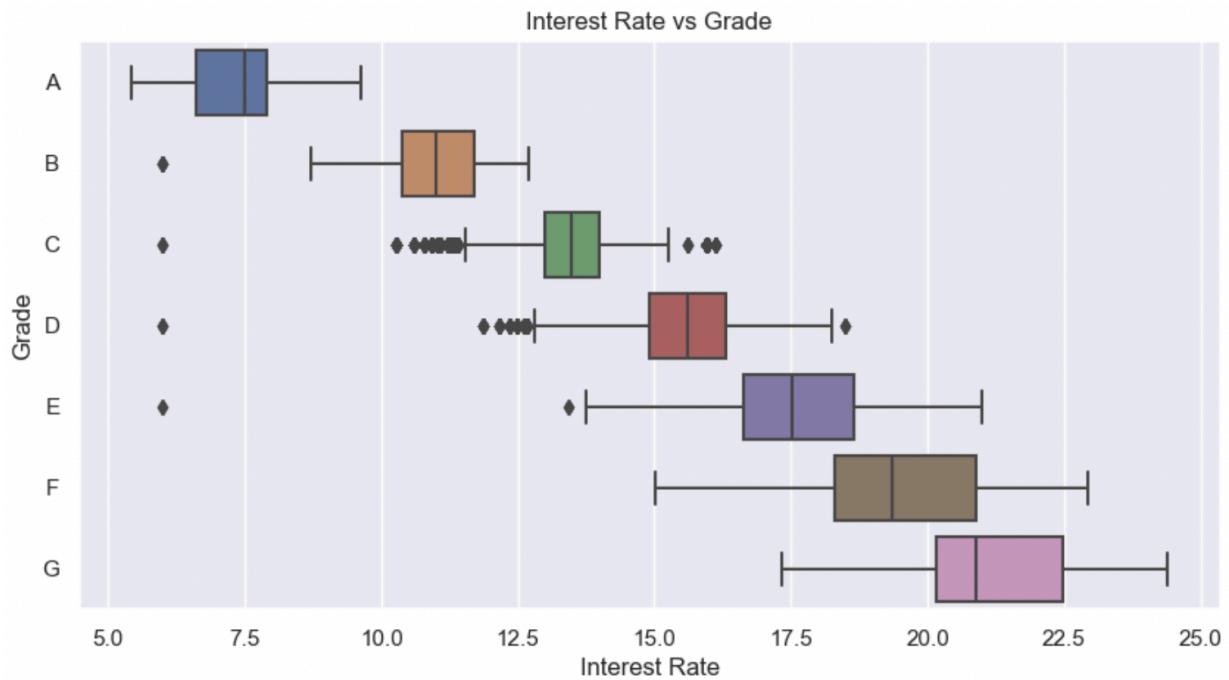
## Data points of Interest

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

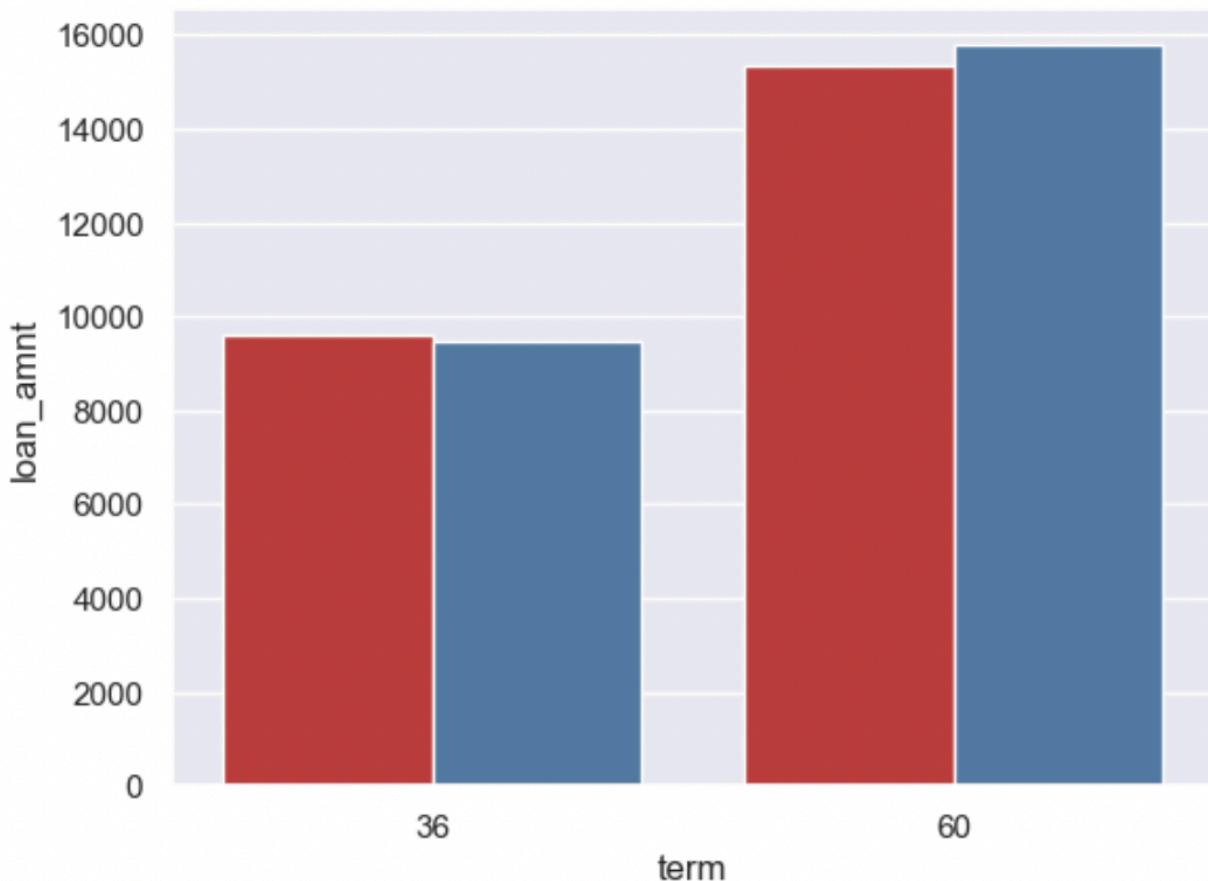
## Interest Rate vs Grade



**Inference:** Lower Grade = Higher Risk. Thus Lower grade with high interest 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.



# Analysis Results

## Univariate Analysis

- The loan amount varies from 500 to 35000 with a mean of 9800.
- 75% of the loan taken has a term of 36 months compared to 60 months.
- Most Loans are issued at an interest rate of around 5-10 and 10-15 with a drop near
- High Grade Loans are high in numbers with Grade B being highest in number.
- Higher subgrades in each grade just confirm that higher grade loans are higher in number
- Most borrowers are the ones having the highest employment length with 10+ years being the most common.
- Borrowers living on rented accomodation are most common followed by those living in mortgaged accommodation. Owners are the least borrowers.
- Most borrowers have been verified by company or have the source verified
- Majority of borrowers have very low annual income.
- Borrowers are taking more loans to consolidate their existing debts. Debt\_consolidation is the most specific reason for taking loan
- Most of the borrowers are from states california, new york, texas, florida.
- 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
- Out of 37173 records 31813 have no public bankruptcy records (85.5%) Rest 5360 (14.5%) have at least 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 records. It is almost clear that each derogatory record almost doubles the chances of Charge Off.
- Last quarter of the year shows the highest disbursal of loans.
- Debt Consolidation has the highest number of both fully paid and defaulted loans. Credit card is the second in the list.
- While the 25th and mean percentile for Fully paid and Charged off cases are the same. However, a loan ticket of 10K to 17K shows higher Charged off incidence lying in the 50th to 75th percentile.
- 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

## Segmented Univariate Analysis

- Borrowers facing Higher interest rate tend to default more as compared to one with a lower interest rate.
- Borrowers with high DTI(>20) are likely to default more than the borrowers with less DTI(<10) or (<20)
- There is not much difference in the number of defaulters paying high installments. However, technically, high installments are related to high defaults.
- Borrowers with low Annual income (in the range < 50K) are likely to default more.
- Lower Grade = Higher Risk. Thus Lower grade with high interest rate increases the risk.
- Higher loan amounts with longer terms have more charged off as well as fully paid loans.
- 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.
- Borrowers with more than 10 years of employment length take large loans.
- Most Defaulters who take Large loan amounts have more than 10 years of employment.

## 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.

## Correlation

High correlation is observed between

- funded\_amnt and loan\_amnt
- Loan\_amnt and installment
- Grade and interest rate

## Suggestions and 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.