UpGrad

International Institute of Information Technology Bangalore

GRAMENER CASE STUDY

Submitted By-

Shradha Katakdhond

Avanni Gudimetla

Karan Hindocha

Vaishali Ramachandran

ABSTRACT

- The main objective of the case study is to understand how consumer and loan attributes influence the tendency of default by understanding the driving variables behind a loan default.
- The dataset contains the history of all the applicants who have defaulted in the past.
- The aim is to identify patterns which indicate if a person is likely to default, which may be used for taking actions such as denying the loan, reducing the amount of loan, lending to risky applicants at a higher interest rate and so on using Exploratory Data Analysis.
- Lending loans to risky applicants is the largest source of financial loss called credit loss. The customers labelled as 'charged-off' are the 'defaulters'. Identification of such loan applicants is the main objective of the study.
- We use EDA to understand how the various parameters such as consumer attributes and loan attributes play a vital role in determining if an applicant would default or not.
- Once we fully comprehend the driving factors behind loan default, these driver variables and the insights we gather from them can be utilized by the company for its portfolio and risk assessment.

ASSUMPTIONS

- The analysis is not carried out for current accounts because it is not possible to determine whether those applications will default or not. The categories - charged-off and fully paid only are considered.
- The variables such as recoveries, total_pymnt, total_pymnt_inv, total_rec_prncp and so on normally get captured only after a loan is accepted and are not available during a new loan application. Hence, they are removed from the dataset.
- Since bankruptcy filings, tax liens and judgments are the three kinds of public records that appears on a credit report, this information was captured in the column pub_rec which contains derogatory public records. The number of values in pub_rec_bankruptcies is greater then pub_rec, hence, the column pub_rec_bankruptcies is dropped.
- The column emp_title contains discrepancies and several unique values that will not yield useful insights about the pattern for loan defaulting. (e.g. The same employer name is mentioned in various formats).
- It is assumed that the deciding parameters for loan defaulters are loan amount, Grade, Purpose of Loan, State, Home Ownership and Verification Status.

PROBLEM SOLVING METHODOLOGY

Data Sourcing and Data Cleaning

- Read the loan.csv data file and understand the different columns and terminologies involved in the study
- Drop unwanted columns and those with significantly high number of missing values.
- Create unique identifier columns and also remove outliers (eg: annual_inc). Also, standardize precision.
- Fix columns with incorrect datatypes and delete invalid values

Univariate and Segmented Univariate Analysis

- Perform univariate analysis of ordered and unordered categorical variables.
- Perform univariate analysis of quantitative variables and create derived columns as and when required.
- Perform Segmented Univariate analysis for the “Defaulted” and “Non-defaulted” segments.

Bivariate Analysis and interpreting the results

- Perform bivariate analysis on the continuous and categorical variables.
- Document the observations of EDA i.e., we make a list of the driver variables.
- Plot the results in Python / Tableau to draw better insights to the results

DATA CLEANING

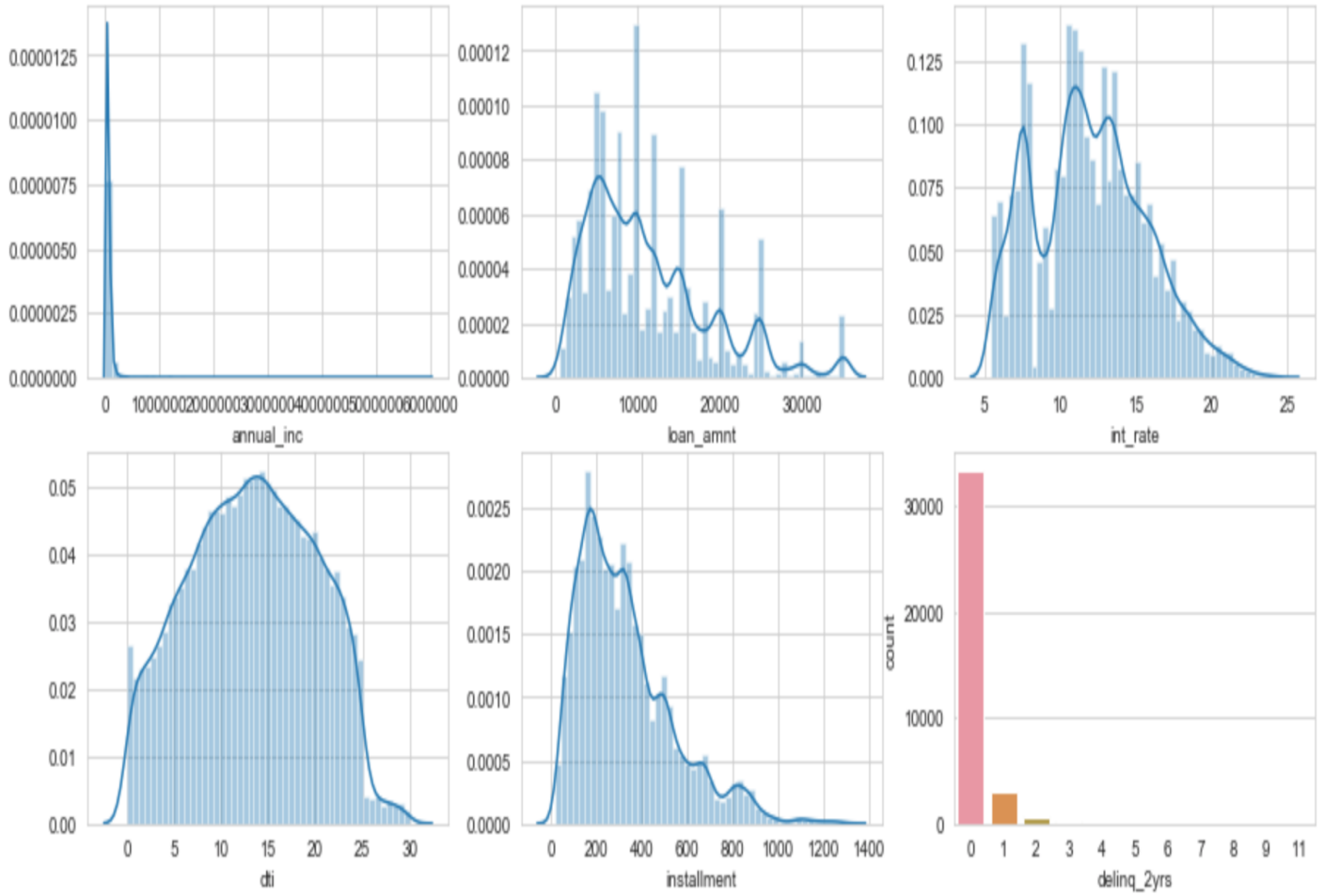
A crucial step to analytics is the data cleansing activity. The objective is to remove unwanted data from the dataset to make it more concise and easy to analyse.

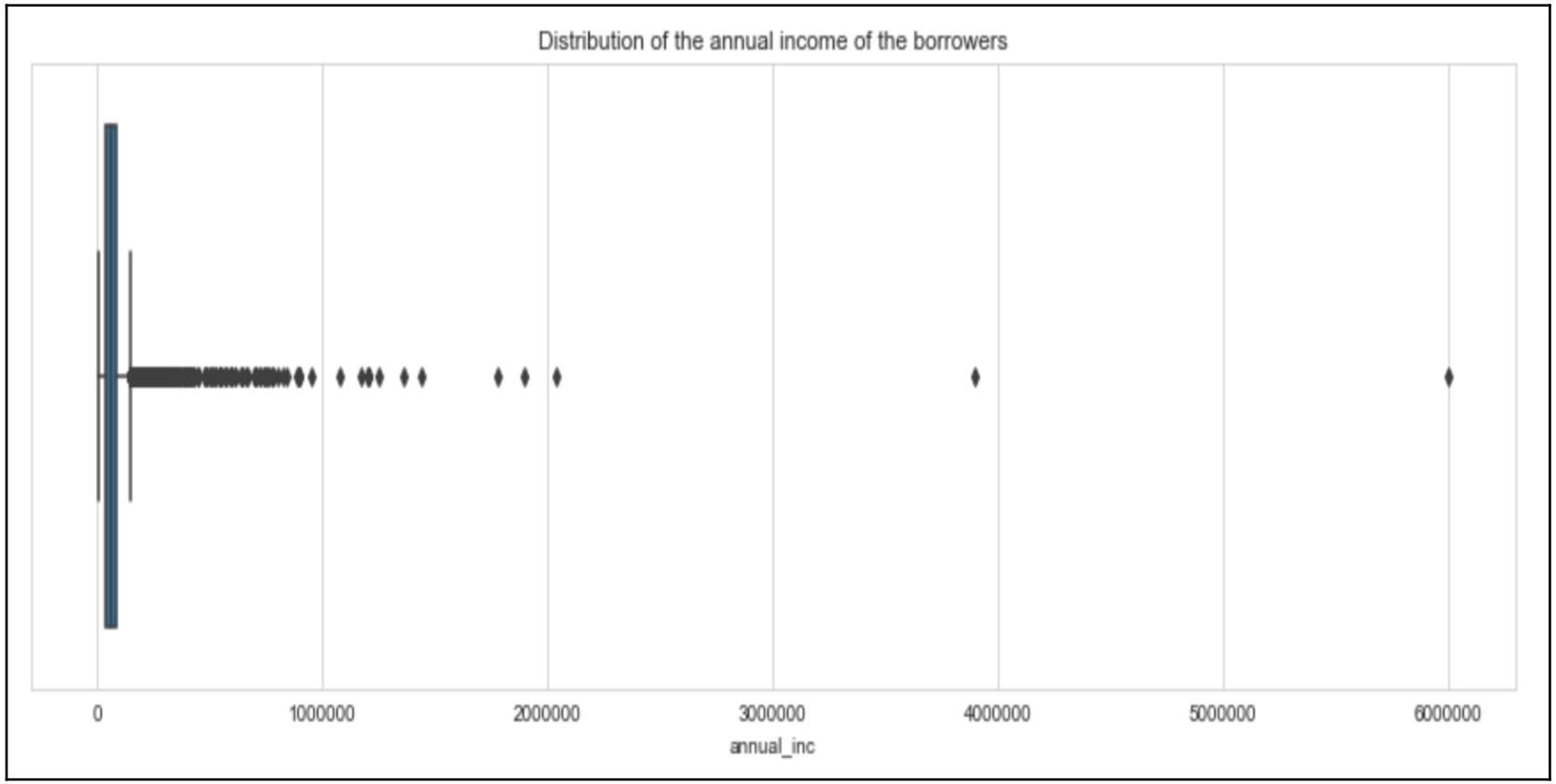
The following steps are taken to clean the dataset:

- Drop the columns which have all the values as “NA” or “0”.
- Drop the columns which have only one unique value and those that are least significant.
- Drop the rows which have all null values.
- Drop “*emp_title*” column since it has 28776 unique values and is not very significant.
- Convert dates to python datetime object.
- Remove *loan_status* rows with values as “Current” since we have no way of finding out if those applicants have defaulted or not.
- Standardize the data formats such as removing the ‘%’ symbol in the ‘*interest_rate*’ and ‘*revol_util*’ columns and also removing the string ‘months’ from the ‘*term*’ column.
- Analysing and removing outliers in columns such as ‘*annual_income*’.
- Conversion of fields to a more readable format as and when necessary.

UNIVARIATE ANALYSIS

- The figure shows the distribution of the following six variables - annual_inc, loan_amnt, int_rate, dti, installment and deling_2yrs.
- The distribution of annual income variable shows a large spike for 0-10,00,000 income range followed by a relatively flat line throughout the plot.
- The reason is the presence of large number of outliers.
- The outliers need to be removed as such values may return false statistical results.
- The other variables display a normal distribution.





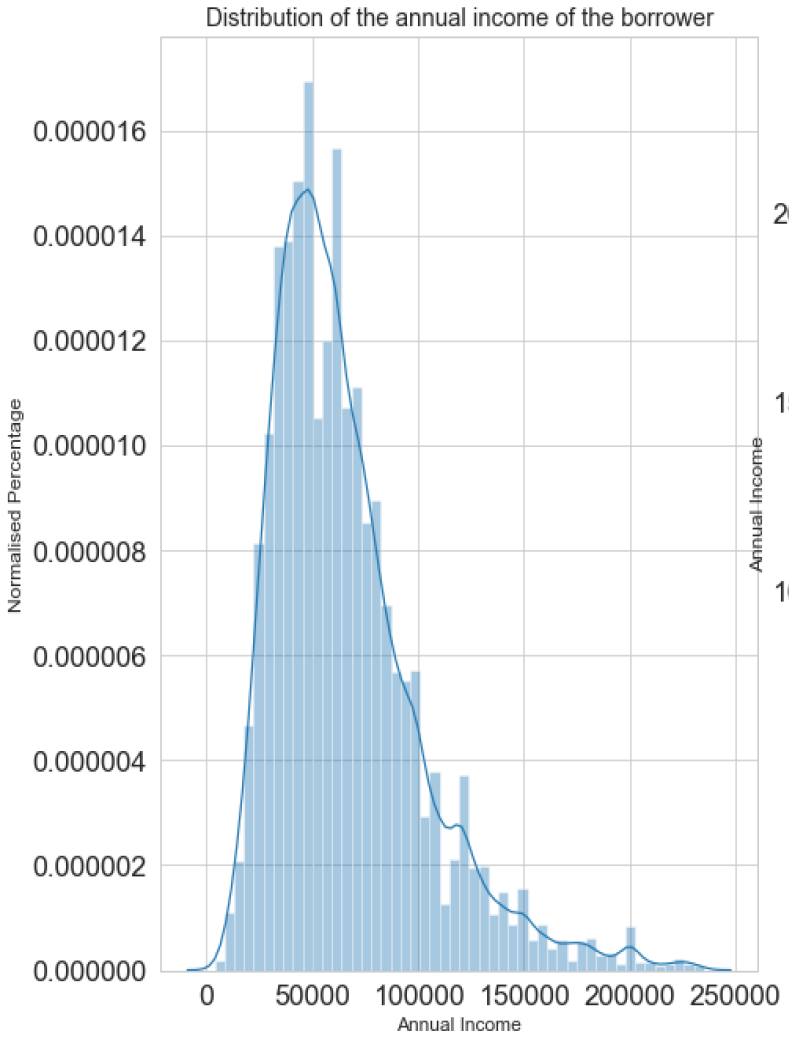
- The box plot presents the distribution of the annual income.
- There are numerous outliers that exceed the maximum and most of the values lie within the Interquartile Range (IQR).
- The annual income variable needs further investigation to correctly interpret the data.

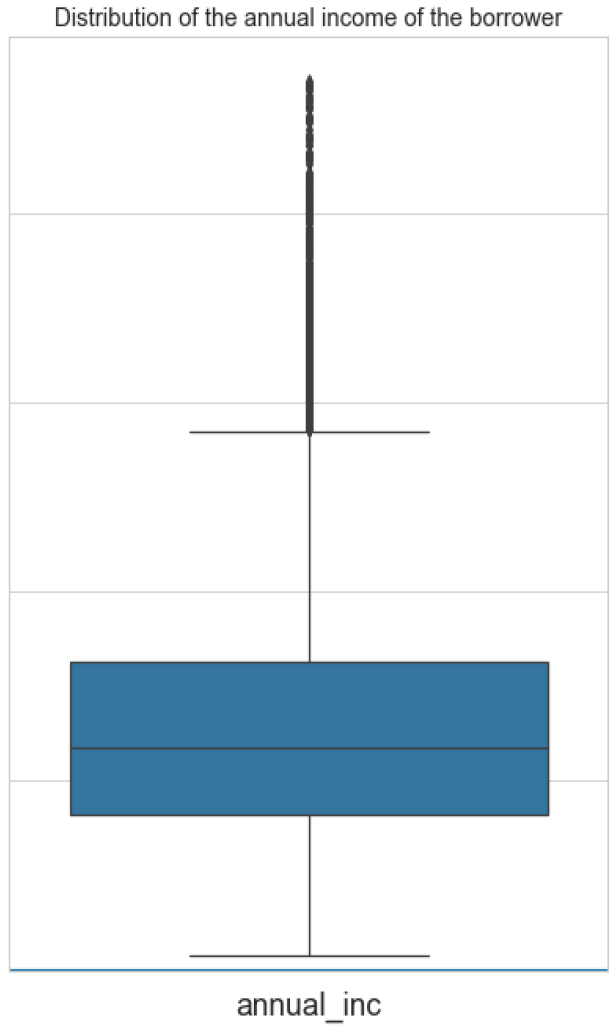
```
df['annual_inc']=df['annual_inc']/1000
df['annual_inc'].describe(percentiles = [0.0,0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,0.9,0.95,0.97,0.99,1])
```

```
count    37544.000000
mean        69.407080
std         64.676984
min          4.000000
0%           4.000000
10%         30.000000
20%         38.000000
30%         45.000000
40%         51.000000
50%         60.000000
60%         66.000000
70%         76.000000
80%         90.000000
90%        116.672400
95%        142.000000
97%        166.213000
99%        235.000000
100%      6000.000000
max       6000.000000
Name: annual_inc, dtype: float64
```

- The values of the annual income column are separated into buckets of thousands described by the percentile range.
- The figure shows that the mean of annual income is 69.407080. Additionally, the 50th percentile annual income is around 60.000000. The minimum income is 4.000000 while the maximum is 6000.000000.
- Also, 99% of the borrowers have income within 235000USD. The remaining 1% are clearly outliers. Hence, the analysis is carried forward with the details of all the borrowers with annual income within 235000USD by dropping the last 1% of the borrowers.

- The figure exhibits the distribution of annual income in the form of histogram and box plots.
- Removing the outliers, annual income shows a normal distribution.
- The box plot shows that the median of annual income is around 60000 USD, the 25th and 75th percentile values are around 45000 and 80000 USD respectively. The minimum is around 0 and the maximum is around 150000 USD.



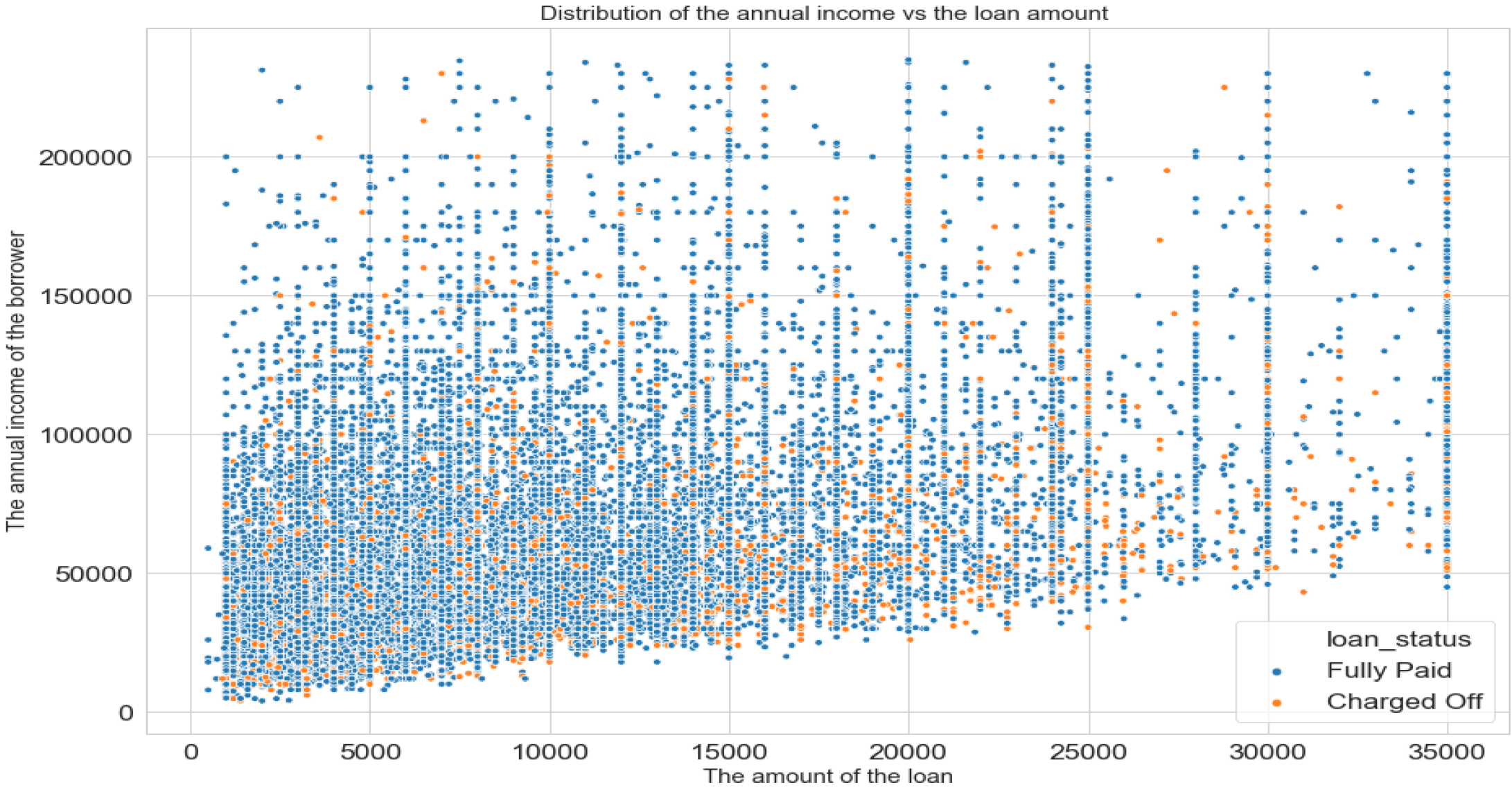


STATISTICAL DETAILS

loan_amnt	funded_amnt	term	int_rate	installment	emp_length	annual_inc	dti	delinq_2yrs	revol_util	pub_rec_bankruptcies	year
12179.181445	11824.015307	46.344596	13.869591	337.877174	5.195072	60699.96003	14.023760	0.169685	55.734577	0.064492	2010.350009
10839.236984	10593.173101	41.162475	11.625589	319.385178	5.027980	66779.36112	13.231143	0.142103	47.695202	0.038353	2010.289959

From the above result, the following observations can be made :

- Credit delinquencies reveal interesting insights. We see that delinquent consumers often borrow hefty loans compared to those who pay their EMIs on time and do not default. Also, on an average, the annual income of non-defaulters is usually higher than delinquent customers.
- The Interest Rate for the defaulters is on an average higher than the rates for the customers who do not default; while there is no significant difference in the installment amount.
- 'Debt to Income' ratio for charged-off(defaulters) is more than the ratio for fully-paid customers.
- Revolving line utilization rate is significantly higher for the charged-off applicants than the fully paid applicants.



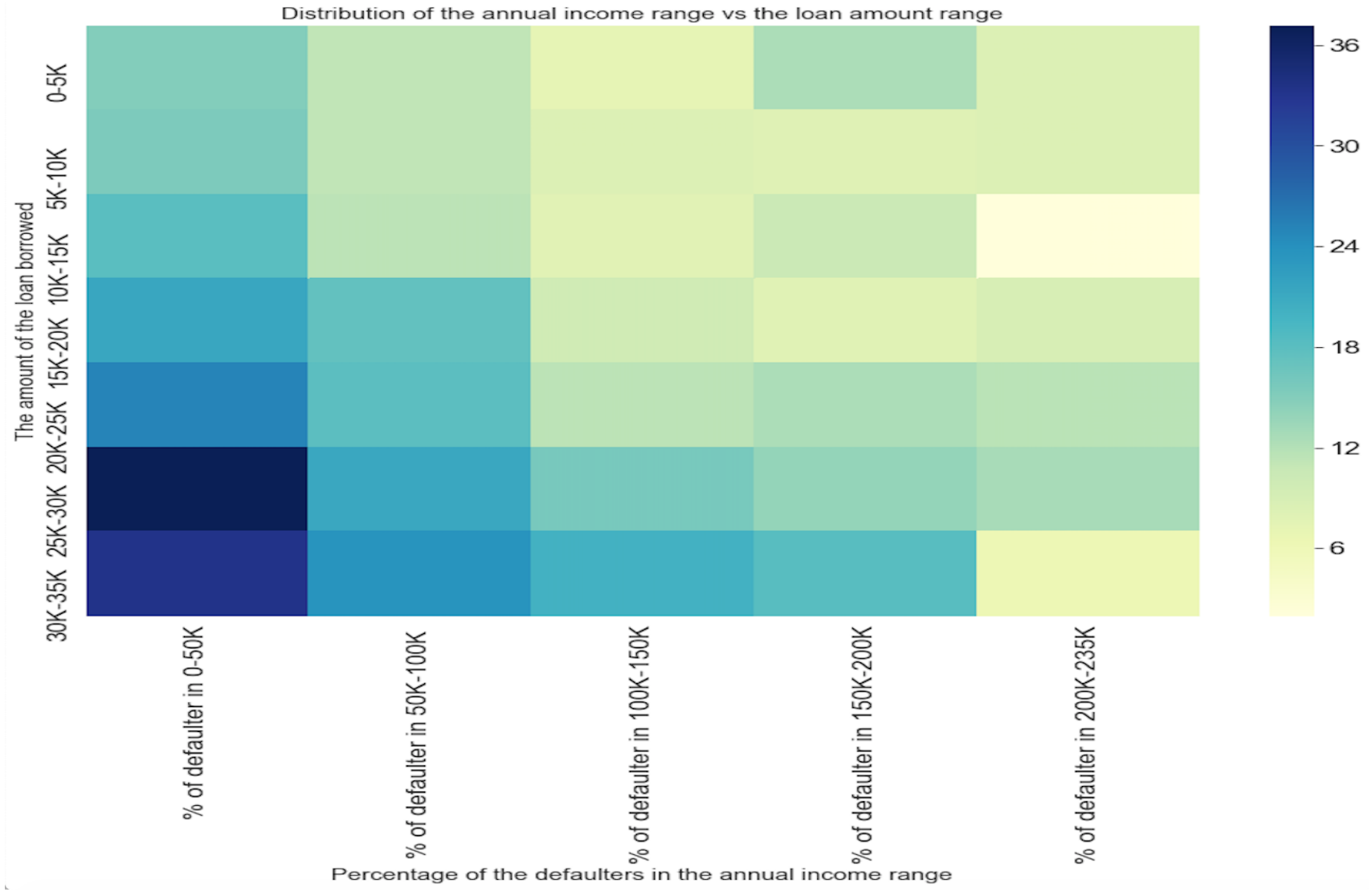
- The above scatter plot tells us that the defaulters generally lie in the area where annual income is low but the borrowed loan amount is significantly high.

Loan_amount_range	% of defaulter in 0-50K	% of defaulter in 50K-100K	% of defaulter in 100K-150K	% of defaulter in 150K-200K	% of defaulter in 200K-235K
0-5K	15.118223	11.191336	7.250755	12.500000	8.333333
5K-10K	15.481411	11.193182	8.370536	8.074534	8.333333
10K-15K	18.073136	11.467577	7.863974	10.447761	1.960784
15K-20K	21.463897	17.476460	9.863429	7.913669	8.928571
20K-25K	25.081433	17.931422	11.501597	12.432432	11.666667
25K-30K	37.142857	21.315193	15.909091	14.000000	12.765957
30K-35K	33.333333	23.706897	20.143885	18.181818	6.250000

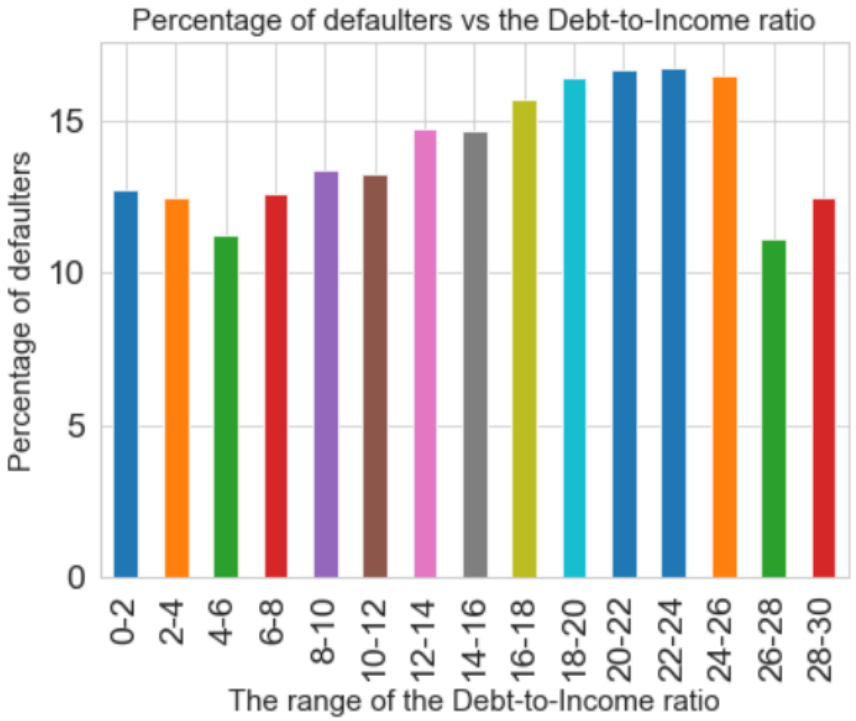
- The above image illustrates the fact that in the categories 0-50k ,50-100k and 100-150k, the number of defaulters increase as the loan amount increases. However, in the categories above 150k , we observe a number of peaks and troughs.

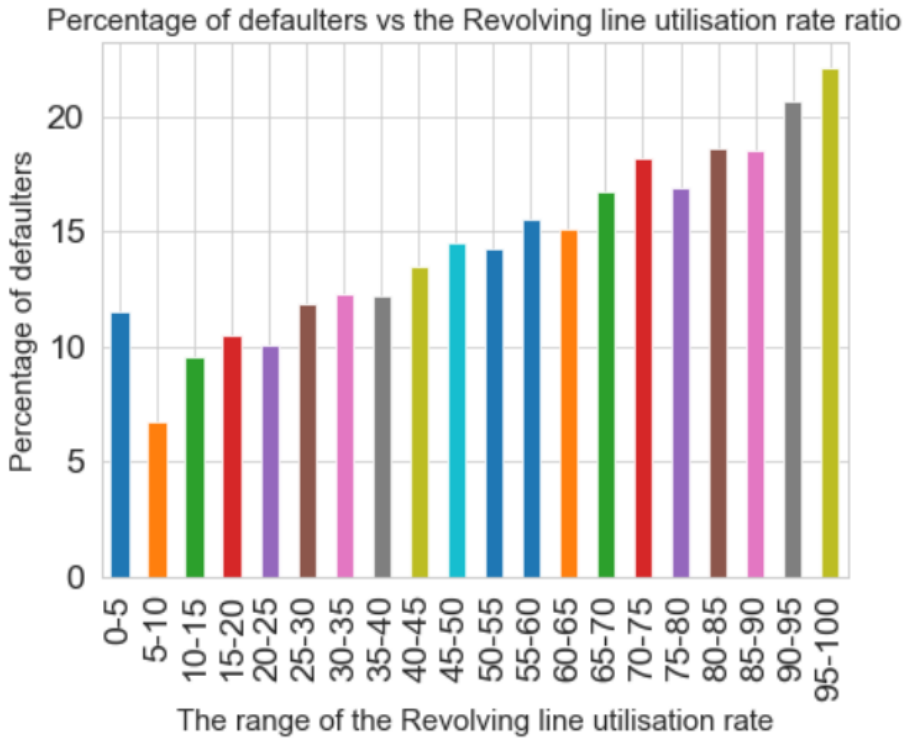
- This implies that the percentage of defaulters is higher in the lower income range and for significantly higher loan amounts.

- The same can be illustrated in a heat map as shown in the following slide.



- The heat map shows that money is lent to borrowers who are not in a position to repay the loans. The lending clubs should restrict the amount of money loaned to such applicants in order to avoid financial loss .

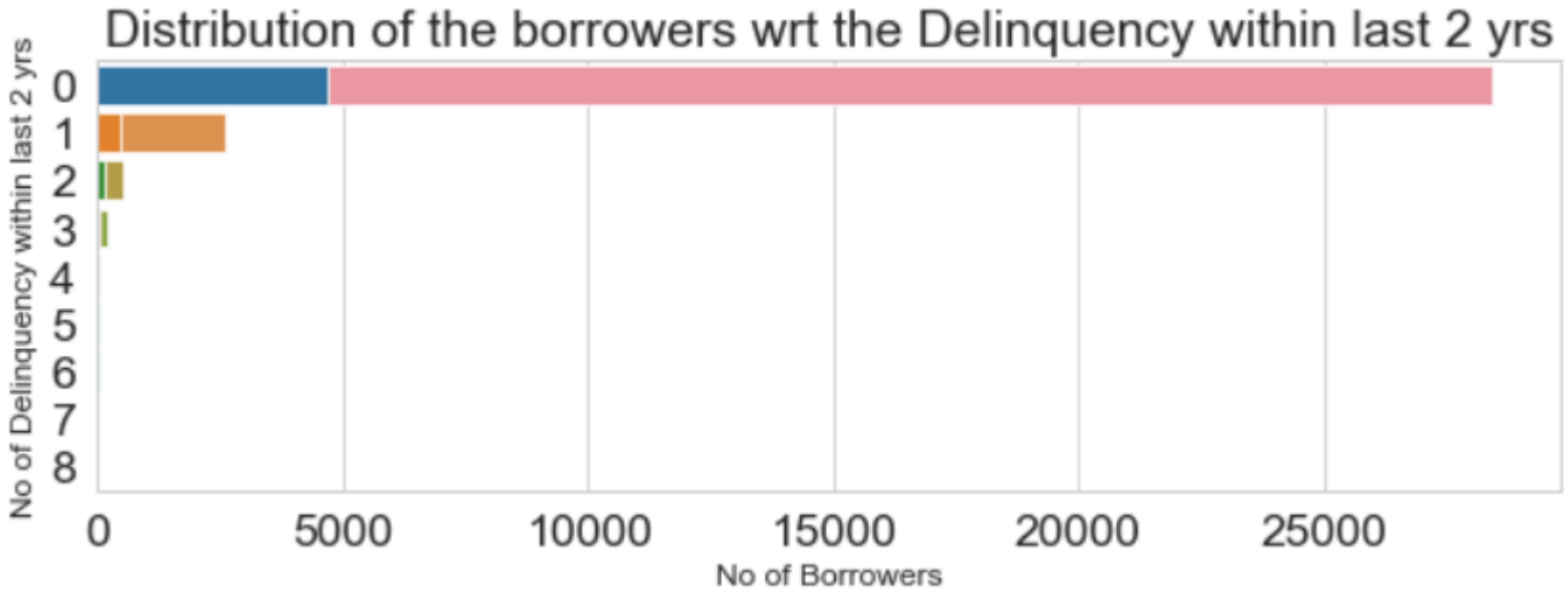




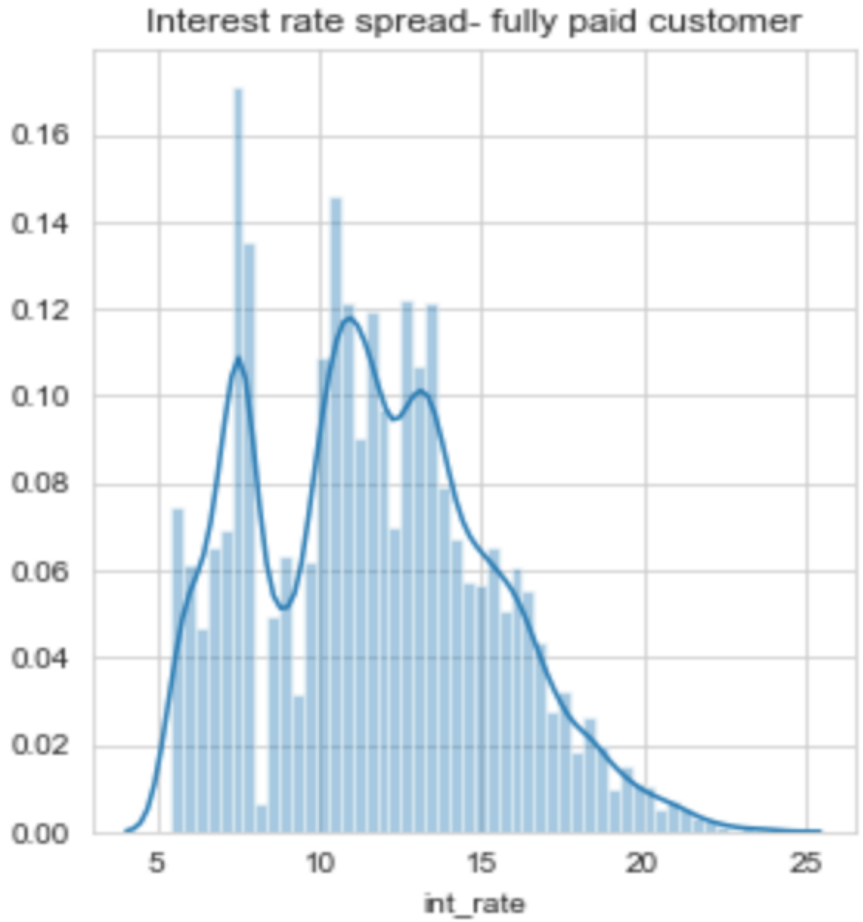
Debt-to-income ratio : A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.

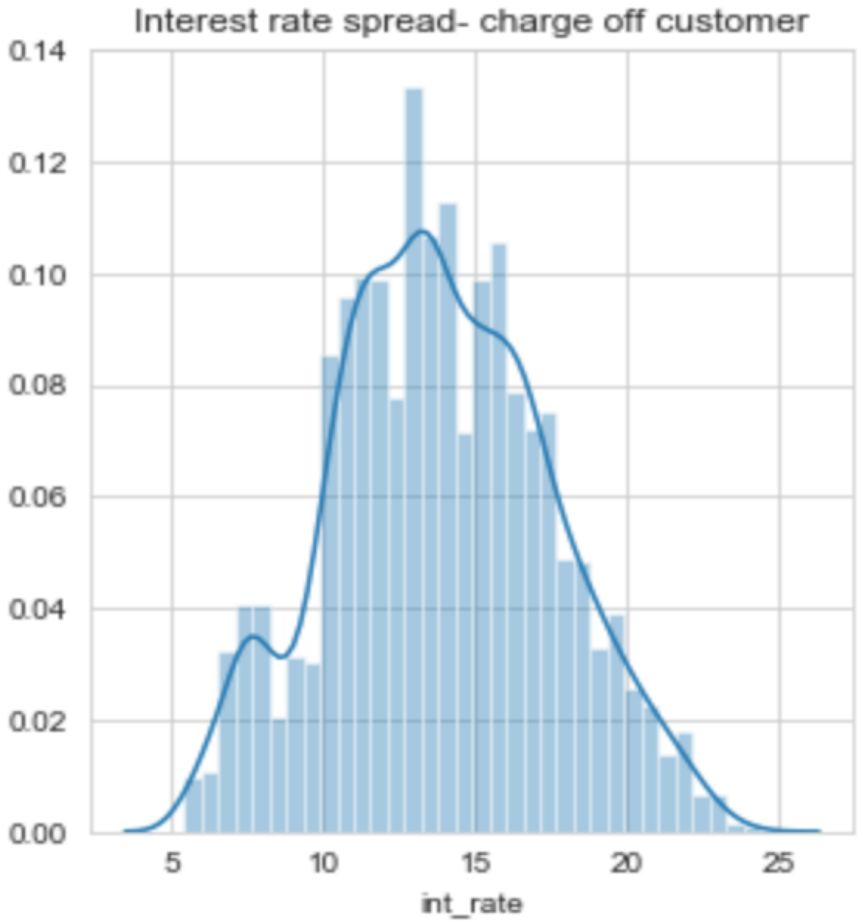
Revolving line utilization ratio : Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

- From the above graph , we draw the conclusion that as the debt-to-income ratio and the revolving line utilization ratio increases, the percentage of defaulters also increases on an average.



- From the graph, we see that there are more borrowers with delinquency within 2yrs equal to 0. Borrowers with delinquency more than 0 have higher number of defaulters compared to that with 0. Hence, borrowers with delinquency greater than 0 within the last two years are risky applicants.





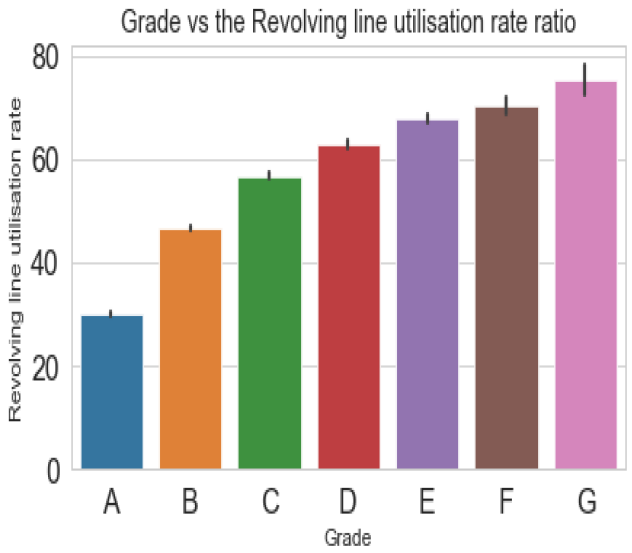
- The figure exhibits the spread of interest rate for fully paid customers and for charged-off customers.
- The rate of interest of defaulters is higher compared to fully paid customers. The are two possible outcomes:

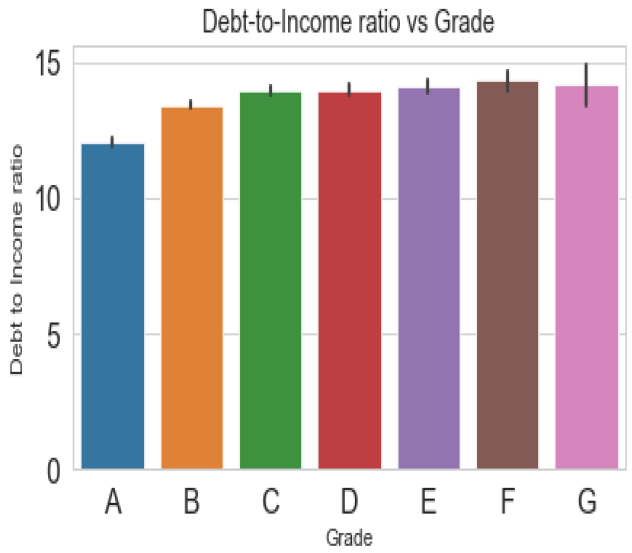
1) Borrowers with high rate of interest tend to default.
2) Risky applicants are identified based on their grade and given high rate of interest for loans in order to expect returns soon since the borrower is likely to default.

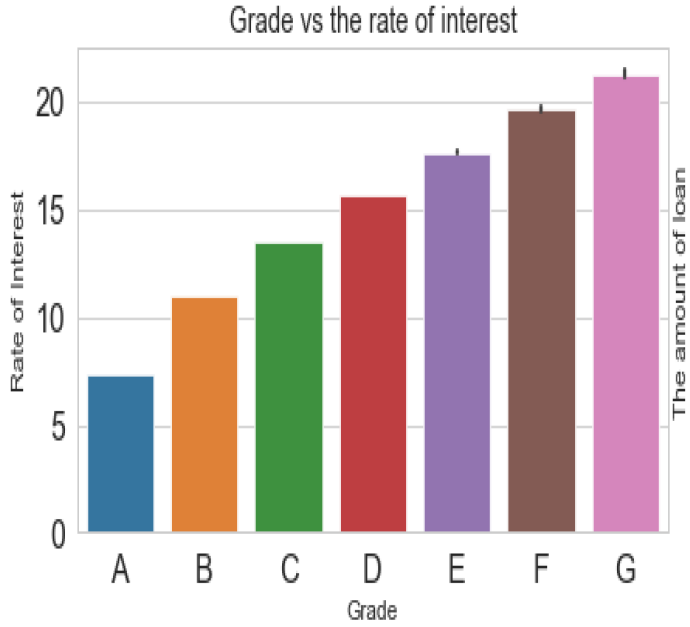
grade	Percentage of defaulters
A	5.81
B	11.94
C	16.90
D	21.92
E	26.83
F	32.66
G	34.03

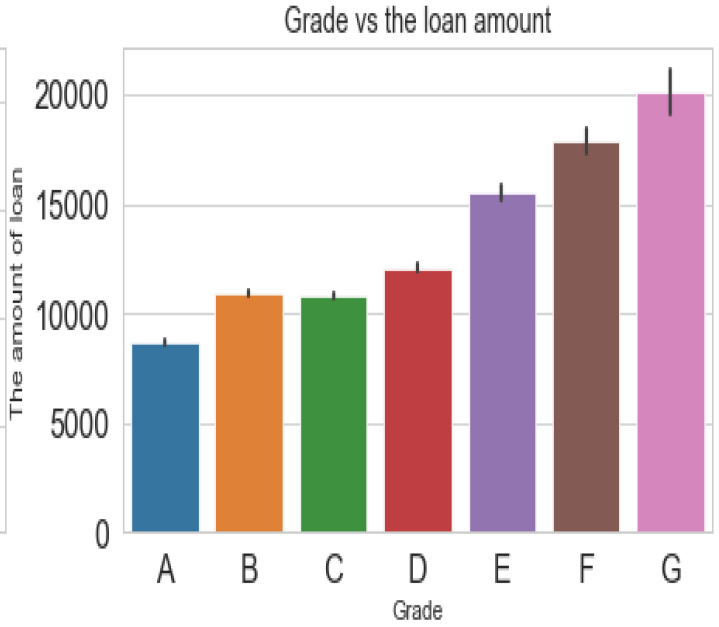
- The table shows the percentage of defaulters for all the seven LC assigned loan grades.
- The percentage of loan defaulters is inversely proportional to the category of loan grade. In other words, as the loan amount increases, the number of charged-off customers increases.

- The bar plots exhibit the relationship between revol_util vs grade, dti vs grade, int_rate vs grade and loan_amnt vs grade in the order of A-G loan grades.
- The observations are that the revolving line utilization rate (revol_uti) and dti increases as the grade of the borrower decreases.
- Hence, lower grade borrowers are likely to default and in order to avoid major losses, the rate of interest for such applicants are set higher.
- Additionally, as the grade reduces, the amount of loan borrowed increases which is a critical matter. This is due to the reason that highly risky applicants should not borrow large amounts of loan as their dti is more.
- Hence, allowance of large amount of loans to highly risky customers should be avoided.

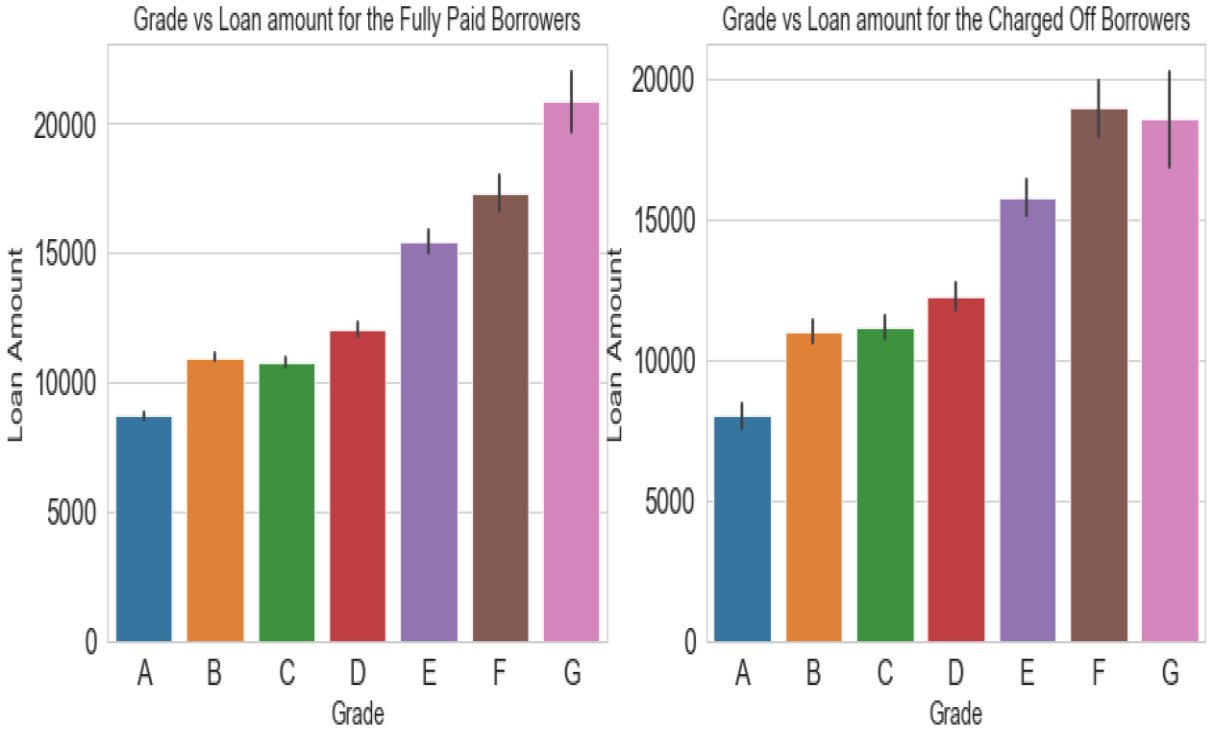


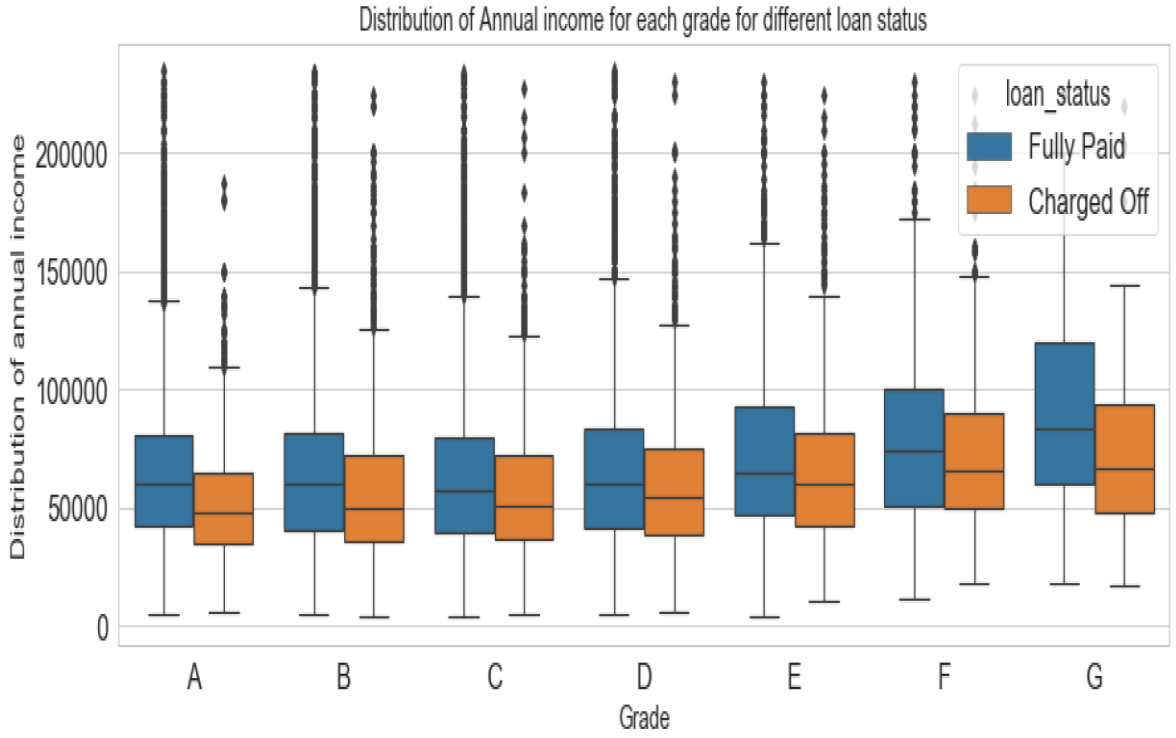




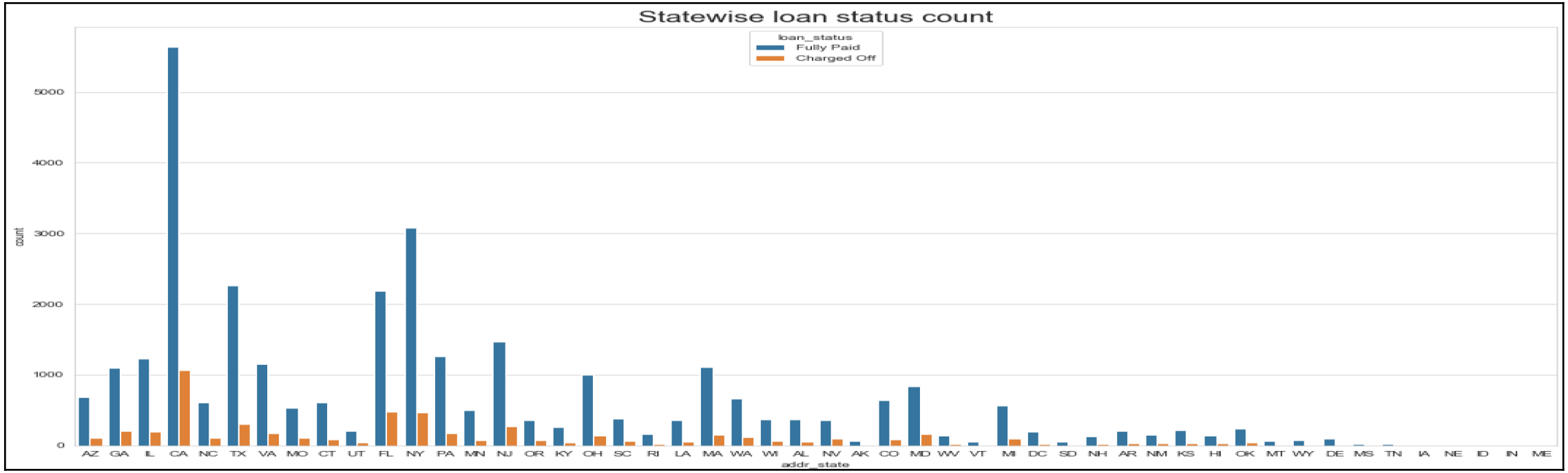


- Figure shows the bar plots of loan amount vs grade.
- In the case of fully paid customers, as the grade increases, the amount of loan borrowed also increases.
- Whereas, in the case of charged-off customers, the number of defaulters are highest in the largest loan grade categories and subsequently lower for lower loan grades.
- As the grade decreases, the amount of loan increases. Hence, more number of defaulters. Lending Club company need to take care of this by applying more rules for ensuring the repayment of the loan.





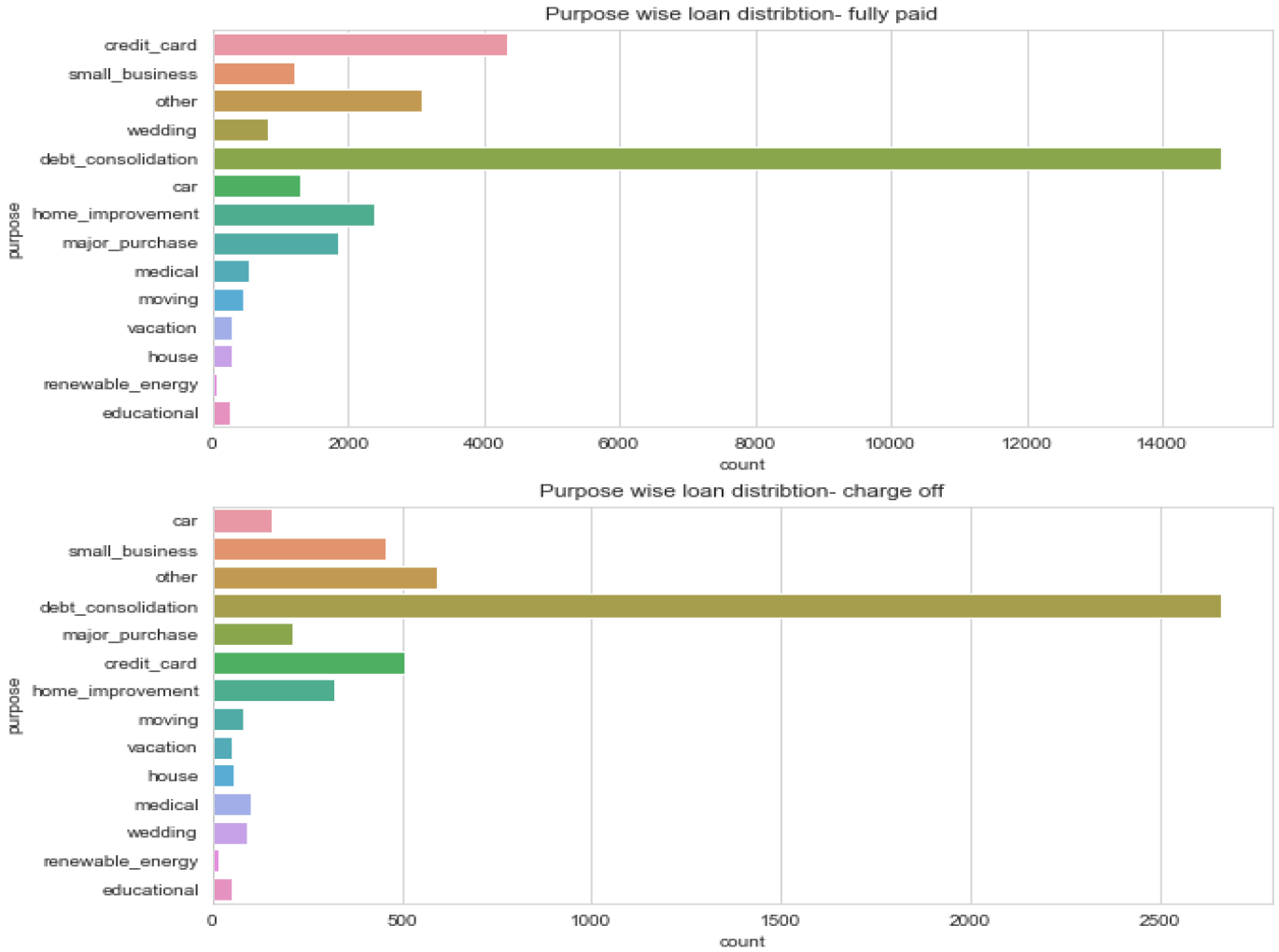
CATEGORICAL ANALYSIS

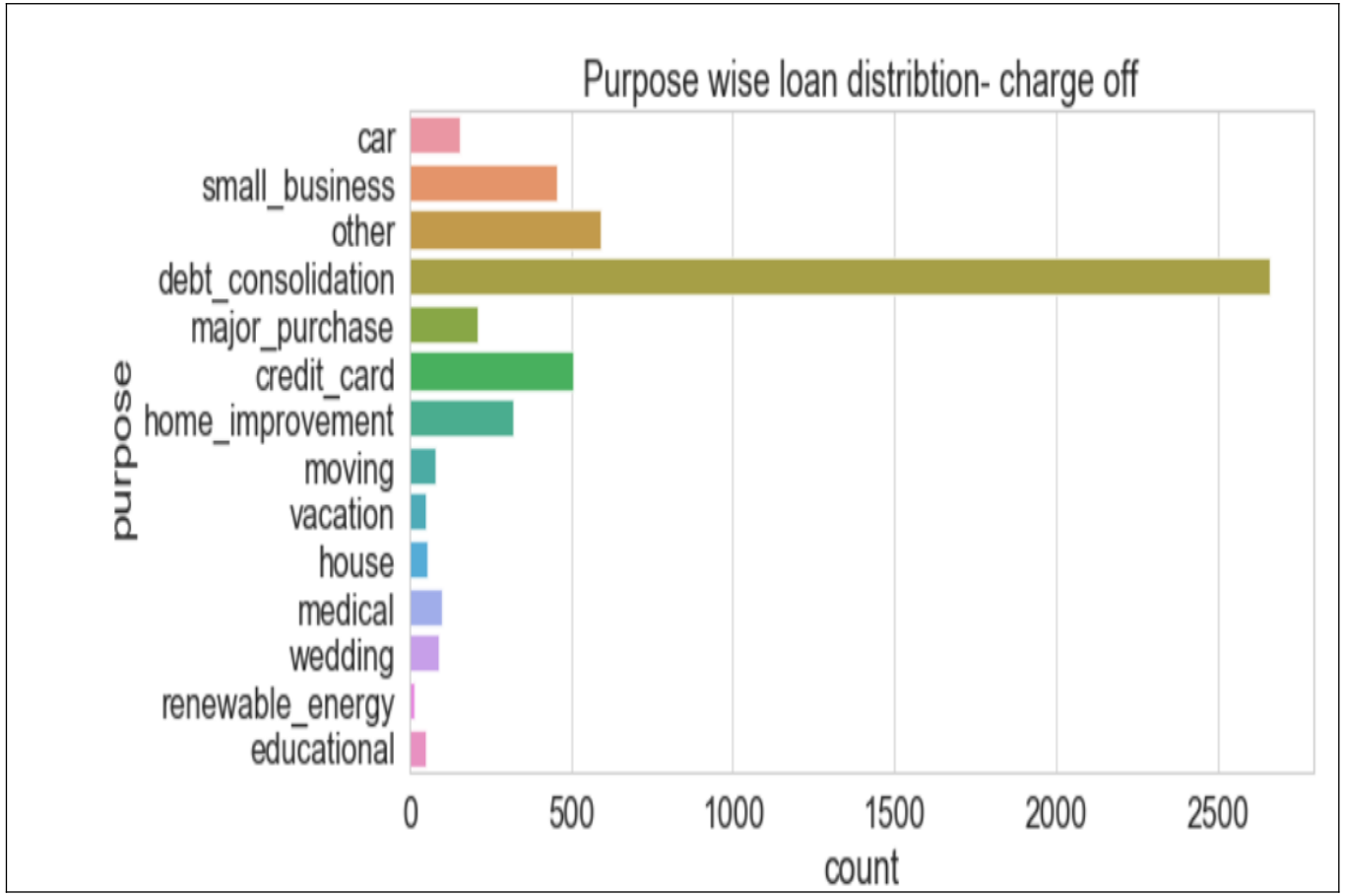


- Categorical analysis is conducted to observe the different variables such as loan amount as well as the interest rates for a particular segment.
- The bar plot shows the status of loans for both fully paid and charged-off customers across all the states in the US.
- The amount of loans fully paid by the customers are highest in the state of California. Likewise, the highest amount of defaulters are also present in the same state.
- The states of New York, Texas and Florida have the next highest number of customers that have completely repaid their loans. The number of charged-off customers of Florida and New York are the same.

PURPOSE WISE LOAN DISTRIBUTION

- The figure represents the bar plots demonstrating the purpose wise distribution of loans for fully paid and charged-off customers.
- Debit consolidation is the main reason for both sets of customers to apply for loans with 14000 and 2600 customers respectively.
- Renewable energy is the category where the least amount of customers applied for loans.
- The categories where generally people applied for loans and repaid in full are credit card, small business, other, wedding, car, house improvement, major purchase whereas in major number of cases, the company suffered a credit loss for loans applied in car, small business, other, major purchase, credit card and house improvement.



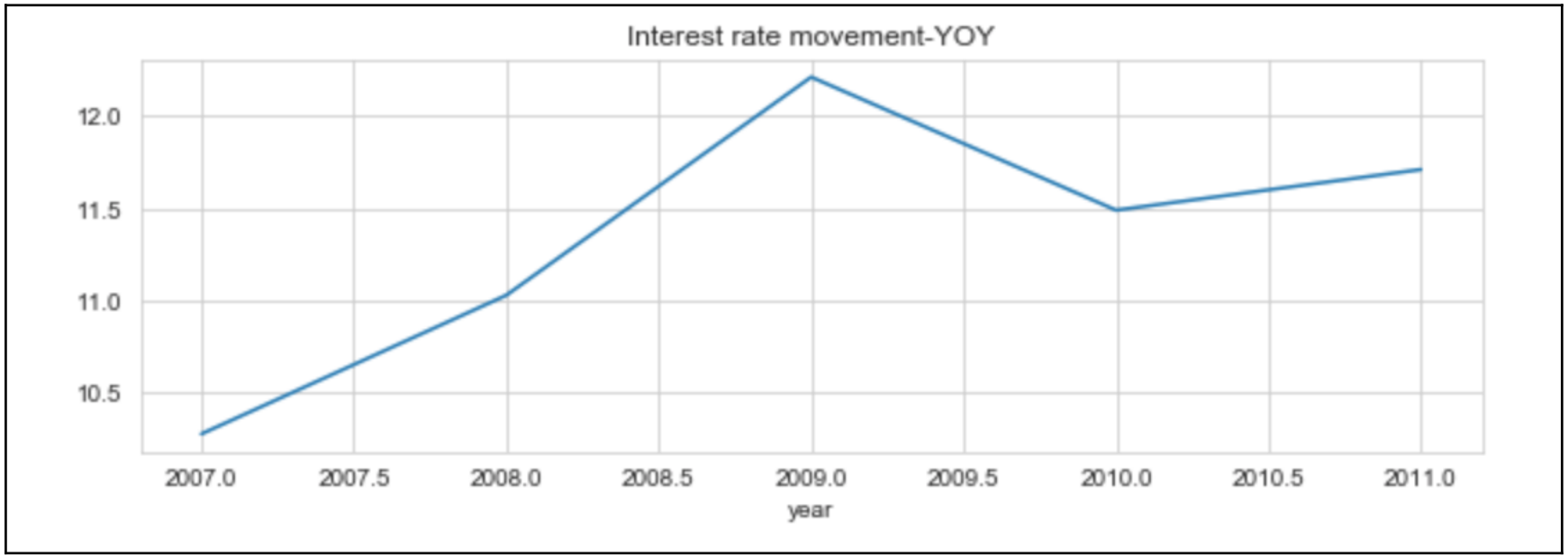


debt_consolidation	47.0
credit_card	13.0
other	10.0
home_improvement	7.0
major_purchase	6.0
small_business	5.0
car	4.0
wedding	2.0
medical	2.0
moving	1.0
vacation	1.0
house	1.0
educational	1.0
renewable_energy	0.0

Name: purpose, dtype: float64 2

- The graph above and the tabular results shows us that *debt_consolidation,credit_card,others and home_improvements* constitute 77 % of the total defaulters and out of this we observe that debt_consolidation has the highest number of defaulters.
- One recommendation to the lending clubs to reduce the risk of financial loss would be to ensure that such applications mortgage some property which gives the lending clubs some assurance that in the event of a delinquency, they can auction the mortgaged property to recover their losses.

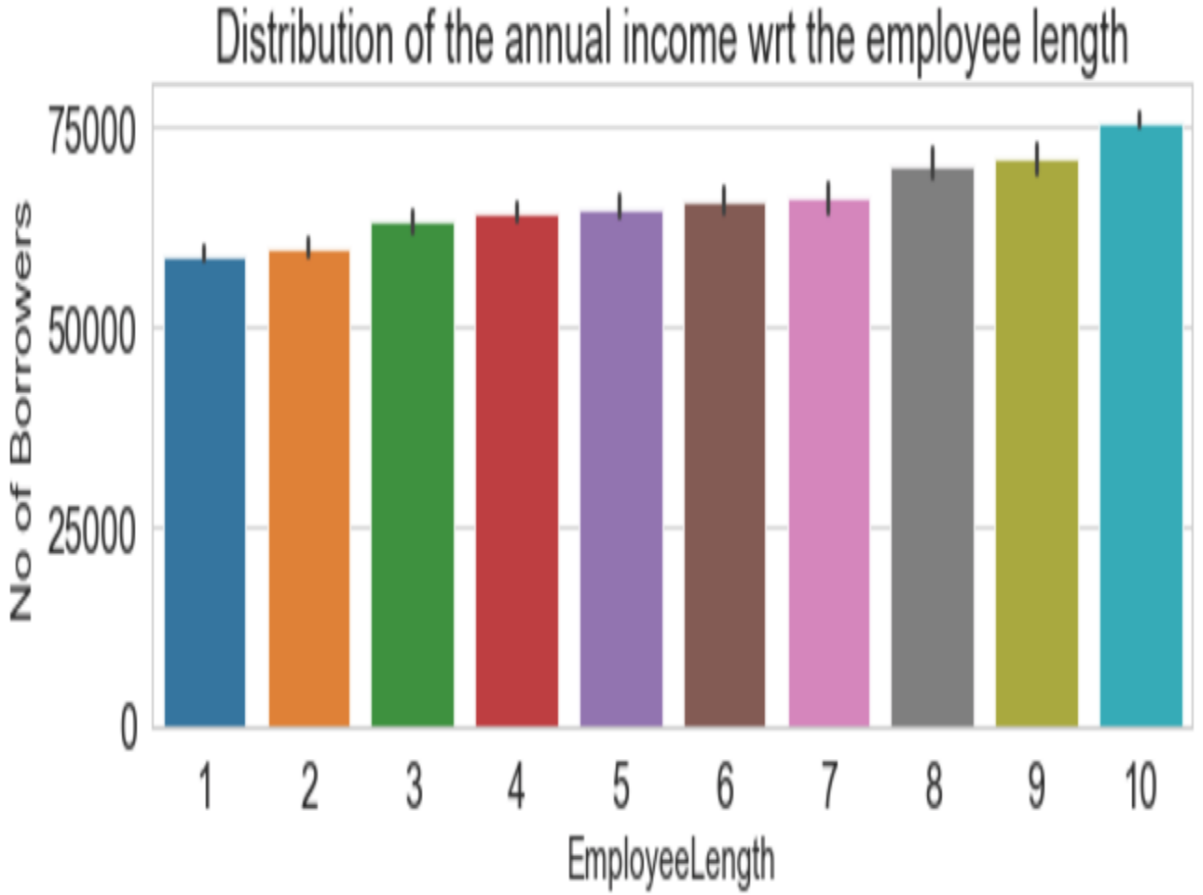
INTEREST RATE DISTRIBUTION



- The graph shown above displays the variation in the rate of interest year over year.
- The interest rate was around 10.0% in the year 2007 and observed a steady growth in 2008 as well as in the year 2009 where it peaked at around 12.5%.
- A downward trend in the rate of interest was observed in the year following 2009.

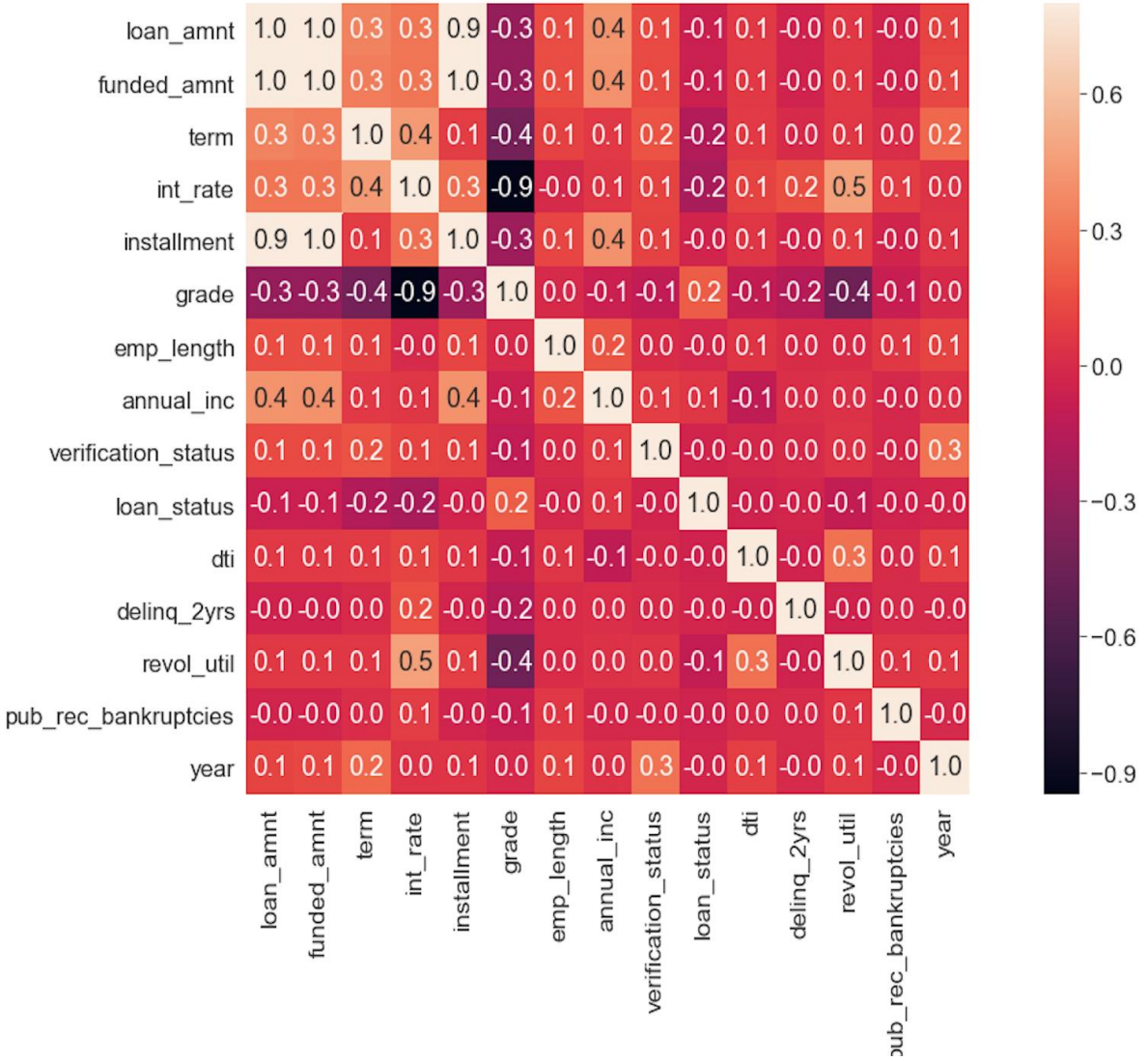
ANNUAL INCOME VS EMPLOYMENT LENGTH

- The figure represents the relationship between employment length in years and the annual income in an interval of 10's of thousands.
- The possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
- The bar plot shows that the employees with a tenure of 10 or more years have the highest annual income of around 75000 possibly due to promotion and better incentives.
- On the other hand, the employees with an experience of 1 year and higher observe a slow and steady rise in the annual income drawn.



CORRELATION

- The figure shows the heatmap or matrix of the correlation of all the variables of the loan dataset.
- The term "**Correlation**" refers to the mutual association between different variables or factors.
- The intensity of the shade represents the correlation index between two variables.
- Lighter the shade, better the correlation. As the shade gets darker, the value of correlation index between those two variables decreases.
- The correlation index between loan amount and funded amount is 1.0. This indicates that these two variables vary equally when either factor increases or decreases.
- A negative correlation value indicates a mutually inverse relationship between two quantities meaning when one variable increases, the other decreases.
- For an instance loan amount vs grade has a correlation index of -0.3.



SUMMARY

- On Average there is difference in the annual income of defaulters and fully paid. Annual income of defaulters is less and fully paid is high. But the amount of loan given on average is nearly same for different grades of borrowers. Since the borrowers are likely to default we have set more interest rate compared to borrowers which is less likely or with better grade.
- Delinquency should be equal to 0 within last two years, Dti should be less , Revolving line utilization rate should be less. People belonging to different geographical area have different financial conditions and standard of living. Accordingly the loan amount and interest rate should be set.
- Number of borrowers are belonging to lower grades , more scrutiny or more rules in order to ensure repayment of loan such as mortgage in return along with proofs of owning the asset .
- Also, we see the purpose of borrowing the loan are debt consolation , credit bill payment etc the borrowers with such purposes are highly likely of defaulting. Hence, we need to add more scrutiny or set limit to the loan amount based on their debts.

THANK YOU