LENDING CLUB CASE STUDY ASSIGNMENT

Name:
Vaibhav Kumbhar
Rutuja Uttarwar

Business Understanding

- Lending Club is the largest online loan marketplace, facilitating personal loans, business loans, and financing of medical procedures. Borrowers can easily access lower interest rate loans through a fast online interface.

Context

- lending loans to 'risky' applicants is the largest source of financial loss (called credit loss). The credit loss is the amount of money lost by the lender when the borrower refuses to pay or runs away with the money owed. In other words, borrowers who default cause the largest amount of loss to the lenders.

Problem Statement

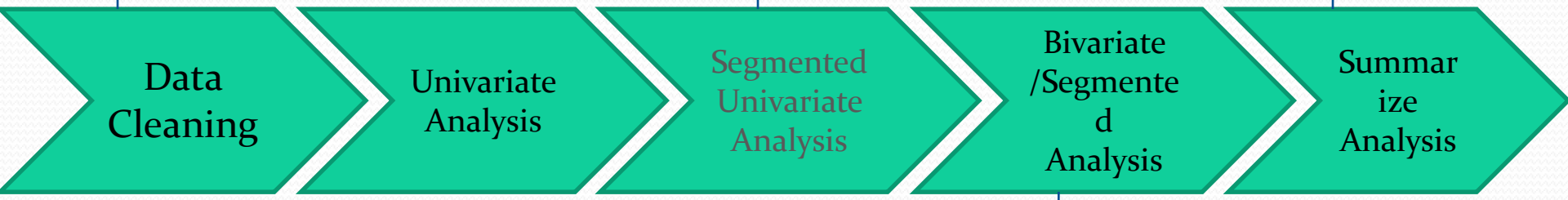
- Therefore, using **Data Science**, **Exploratory Data Analysis** and public data from **Lending Club**, we will be exploring and crunching out the driving factors that exists behind the **loan default**, i.e. the variables which are strong indicators of default.

Analytical Approach

- Drop all null columns and single value presented column
- Drop not needed columns
- Drop null rows or imputation
- Convert al fields into proper data types

•Check distributions and frequencies of various numerical and categorical variables >
• Create derived variables

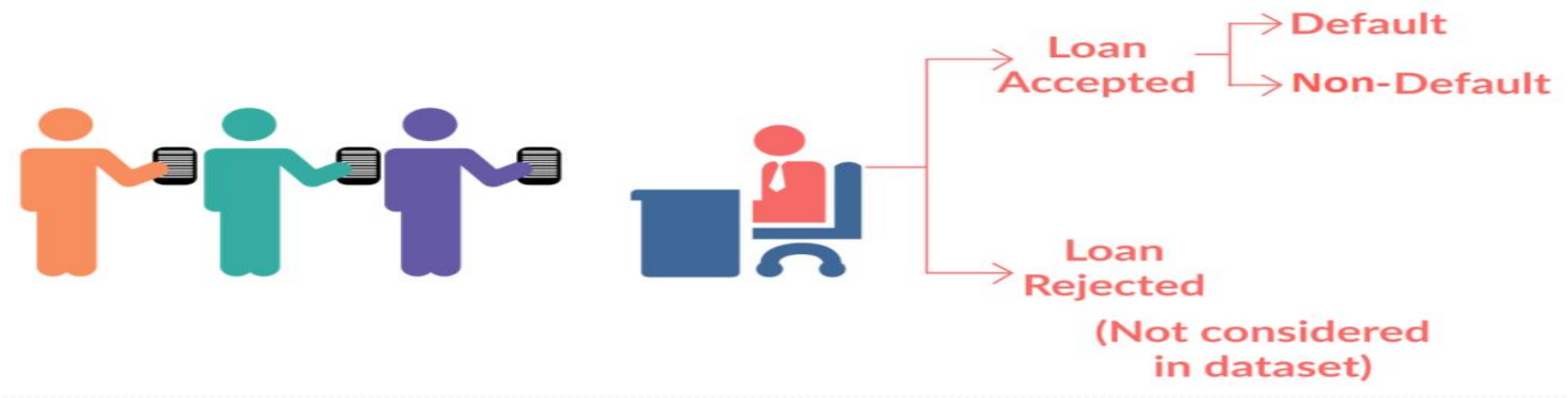
Conclusion and Suggestion to company



•>Check distributions and frequencies of various numerical and categorical variables
• Create derived variables

•Do correlation analysis Check how two variables affect each other or a third variable >
• Analyze joint distributions

About Lending club dataset



If the company approves the loan, there are 3 possible scenarios described below

- **Fully paid**: Applicant has fully paid the loan (the principal and the interest rate)
- **Current**: Applicant is in the process of paying the installments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as '**not defaulted/good**.'
- **Charged-off**: Applicant has not paid the installments in due time for a long period of time, i.e. he/she has **defaulted/bad** on the loan .

Derived Question For Analysis

- What set of loan data are we working with?
- What types of **features** do we have?
- Do we need to treat **missing values**?
- What is the distribution of Loan Status?
- What is the distribution of Loan Default with other features?
- What all plots we can draw for **inferring the relation** with Loan Default?
- Majorly, what are the **driving features** that describes the Loan Default?

Feature Distribution

- **Loan Characteristics** such as **loan amount, term, purpose** which shows the information about the loan that will help us in finding loan default.

- **Demographic Variables** such as **age, employment status, relationship status** which shows the information about the borrower profile which is not useful for us.

- **Behavioural Variables** such as **next payment date, EMI, delinquency** which shows the information which is updated after providing the loan which in our case is not useful as we need to decide whether we should approve the loan or not by default analysis

Feature Distribution

- **Dataset Overview** (Distribution of Loans)
- **Data Cleaning** (Missing Values, Standardize Data, Outlier Treatment)
- **Metrics Derivation** (Binning)
- **Univariate Analysis** (Categorical/Continuous Features)
- **Bivariate Analysis** (Box Plots, Scatter Plots, Violin Plots)
- **Multivariate Analysis** (Correlation Heatmap)

Analysis in case study

- The essence of the whole project is to analyze and understand how consumer attributes and loan attributes are influencing the tendency of defaulting.
- We performed data cleaning and preparation on the Loan dataset
- Histograms and Bar charts to check out the distribution of all the driver variables
- Box plots to detect the Outliers
- Performed the Multivariate analysis to understand how different variables interact with each other.

Data Cleaning

- With respective to the case statement, we analyzed all given and taken some decision on data cleaning. Shape of given data is "**(39717, 111)**"
- Dropped all columns having more than **50** % missing values.**(39717, 54)**
- We have few more customer behavioral columns. Here we are not doing behavioral analysis. So we dropped it too.**(39717, 54)**
- There were very less percent of mission values at row level i.e less than **3 %.**(like emp_length column has 1075 records of missing data)

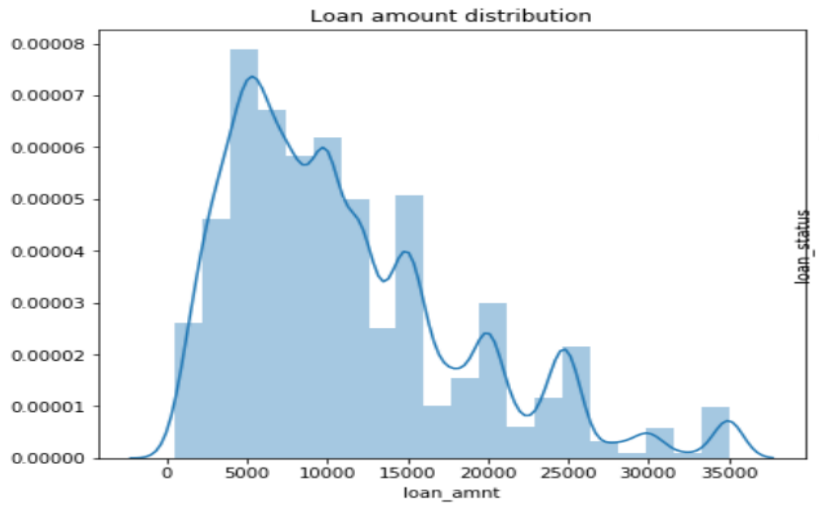
Data Cleaning cont.

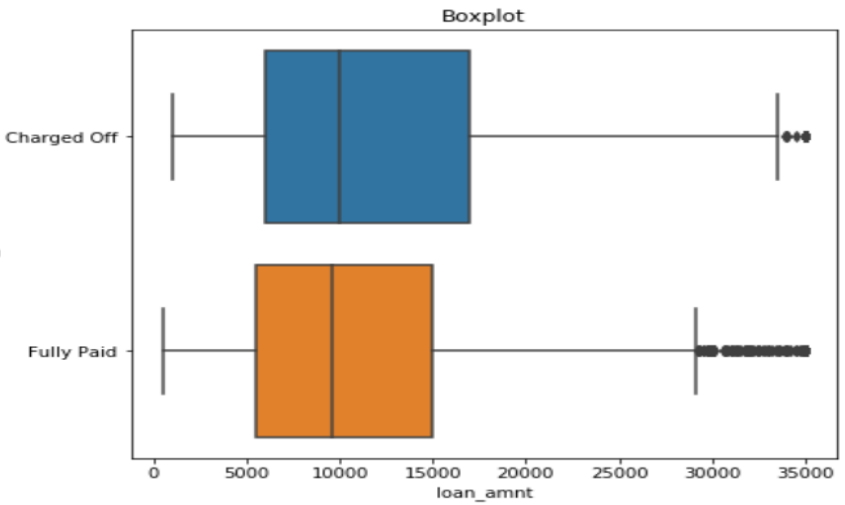
- We filter out loan status “Current” variable as we are analyzing loan borrowers data in this case study.
- Referred “**Fully Paid**” variable as “**Non-Defaulters/Good Variable**”
- Referred “**Charged Off**” as “**Defaulters/Bad Variable**”
- Created bins for quantitative and also created some new column for analyze like to check ratio between annual income and annual installment amount.

Univariate /Bivariate Analysis

- Loan Amount Distribution
- Interest Rate Percentage Distribution
- Installment Distribution
- Dti Distribution
- Compared Term, grade , purpose , emp_length_year, Issue_year, issue_month, verification w.r.t number of loan count, and created bins for quantitative measure and checked its distribution

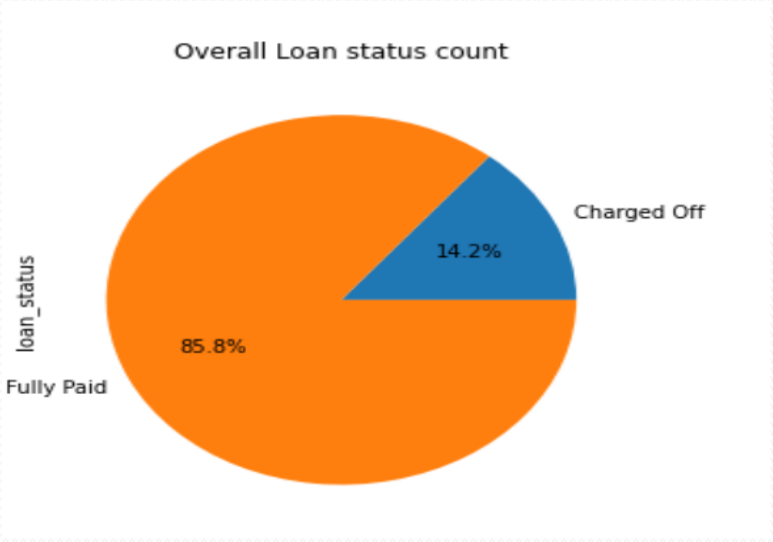
Loan Amount Distribution



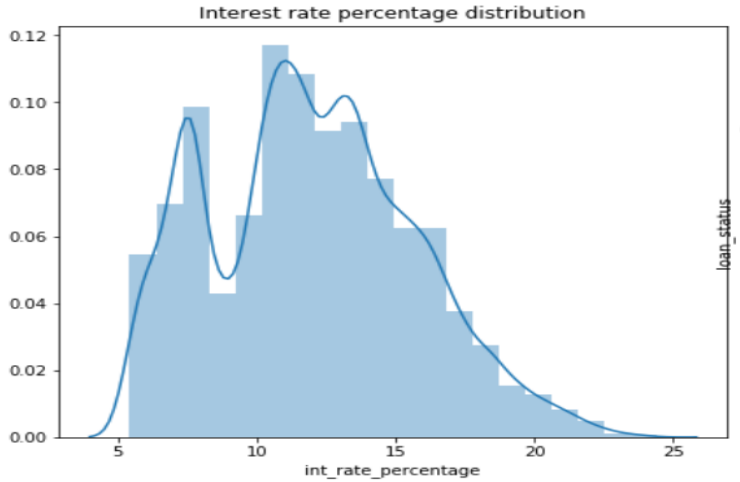


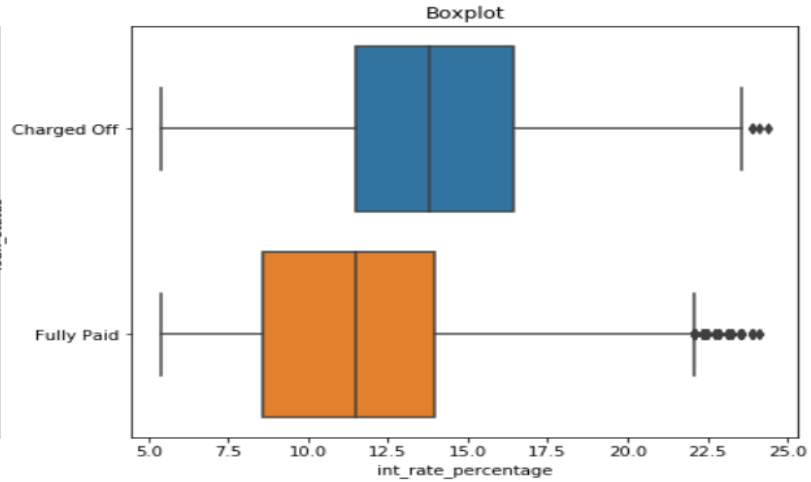
Observation

- Approximate **14% falls into default category**
- The loan amount varies from 0 to 35,000
- -Most of the loan amounts are in the range of **0-15K**
- Mean of default loan amount is approximate **12292.6**



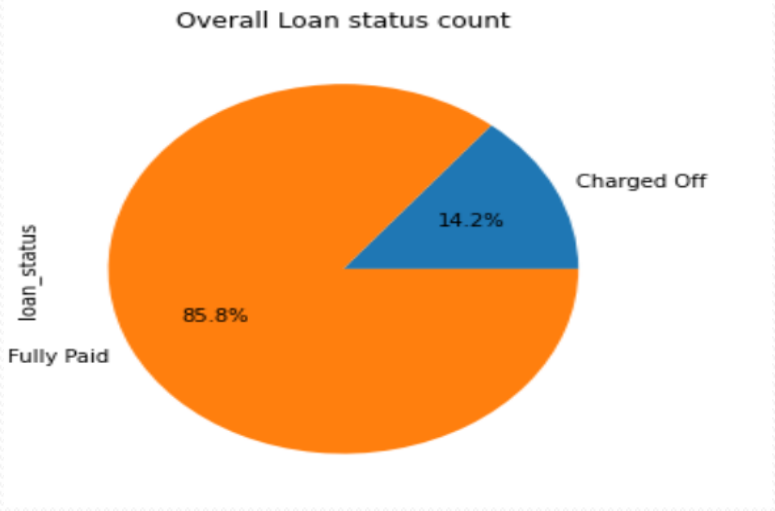
Interest rate percentage distribution



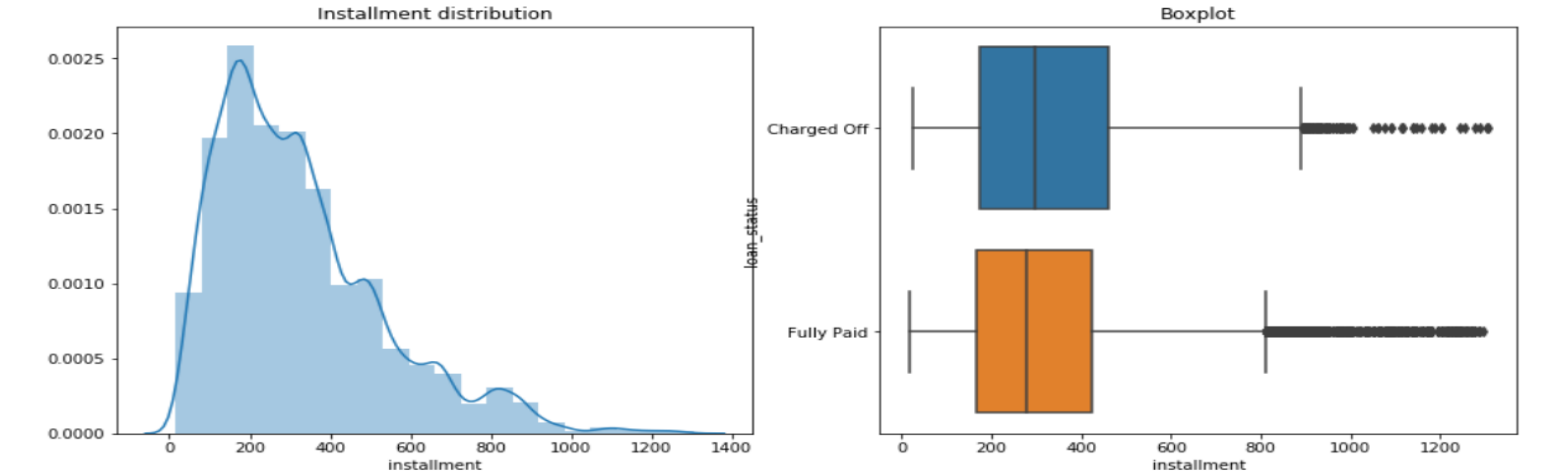


Observation

- Interest rate percentage varies from 5.42% to 24.40%
- Average interest of Charged off (13.9%) is **higher** than Fully paid (11.6%)



Installment Distribution

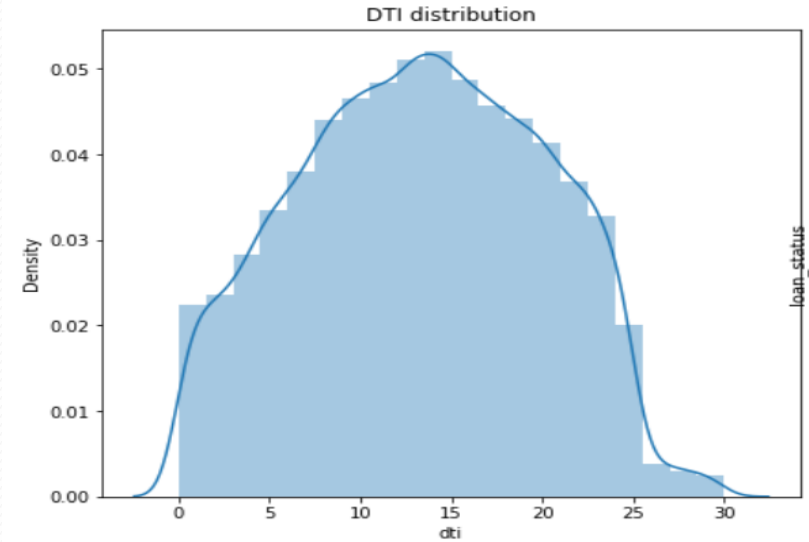


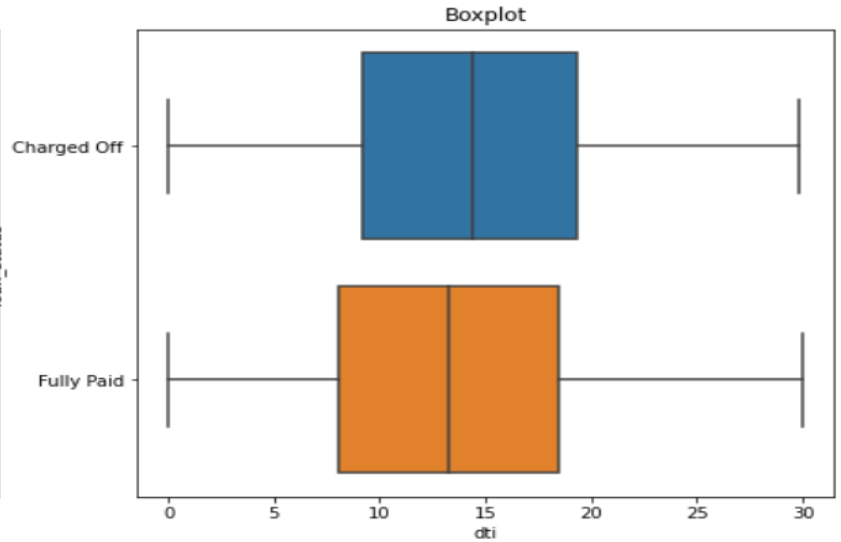
Observation

- **Charged Off**
  - Installment amount varies from 22 to 1305 for Charged Off
  - Average installment amount is 339.51
- **Fully Paid**
  - Installment amount varies from 16 to 1295 for Charged Off
  - Average installment amount is 322.61

Average Installment amount for defaulters is **higher** than good category people

dti distribution





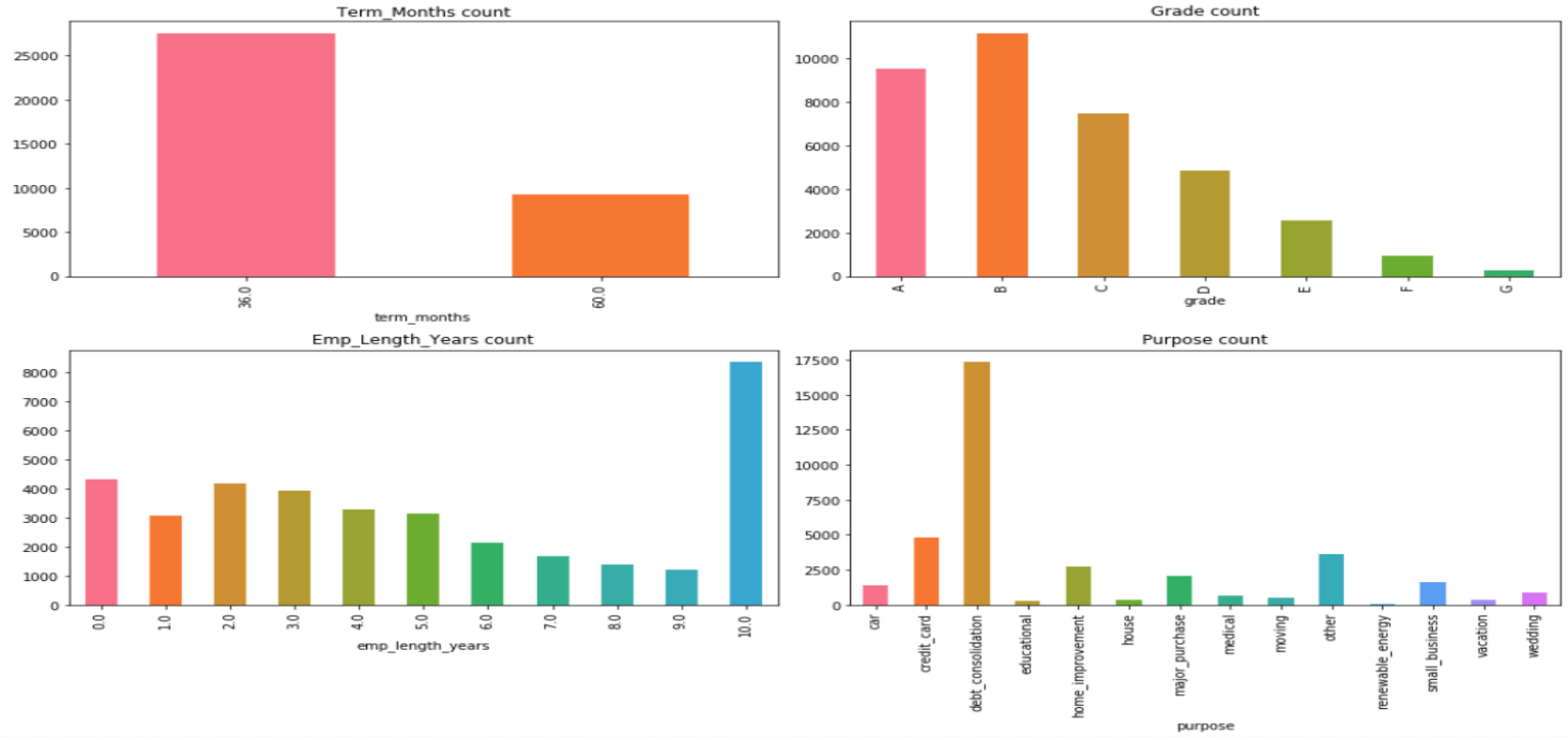
Observation

- **Charged Off**
  - dti amount varies from 0 to 29.85 for Charged Off
  - Average dti is 14.05
- **Fully Paid**
  - dti amount varies from 0 to 29.99 for Fully paid
  - Average dti is 13.21

Average dti amount is **higher for defaulters** than good category people

Analysis over categorical Variable

(defaulters and non defaulters)



Statement:

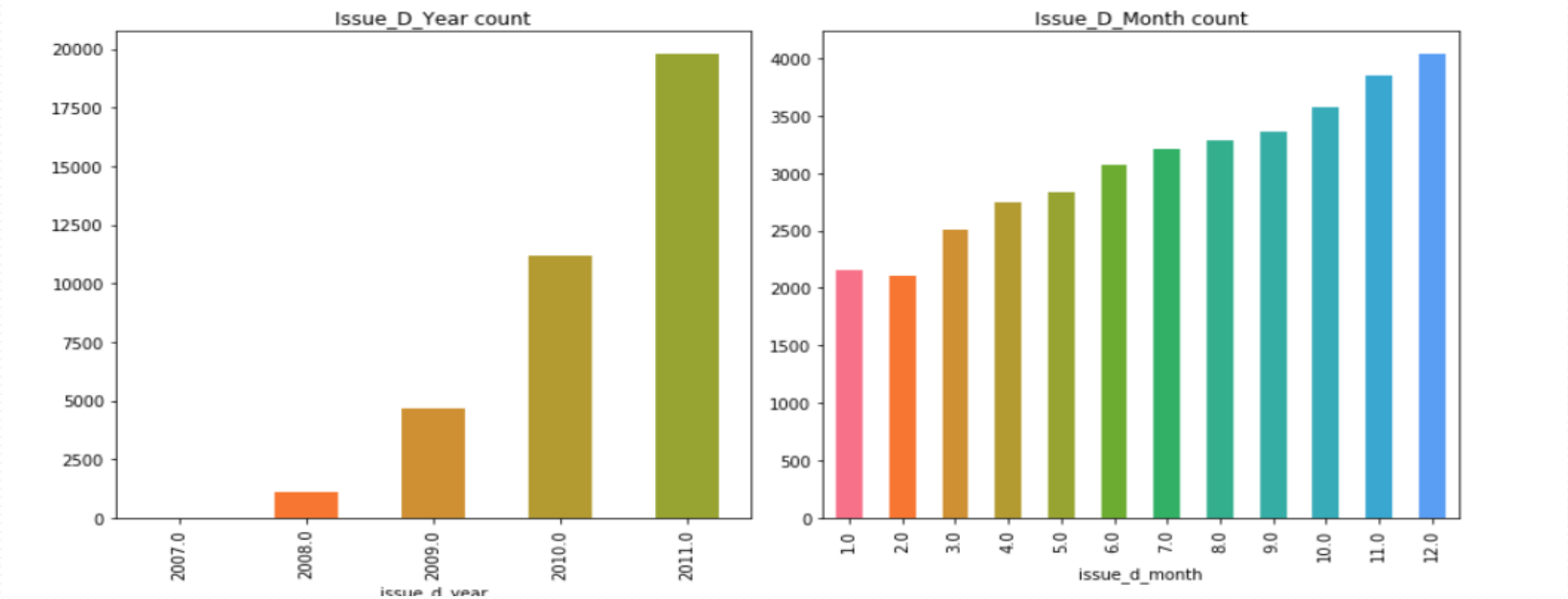
Here we did analysis on loan_amount over categorical fields like **Term, Grade,Purpose,Emp_length for default and non defaulters**

Observation:

- Most of the loans are taken for **36 months**
- Most of the loans have grade **A and B**, i.e most of the loans are high graded loans.
- Most of the employees applying for loan have experience more than **10**.
- Most of the loans are issued for the purpose of **dept consolidation**.

Analysis over categorical Variable

(defaulters and non defaulters)



Statement:

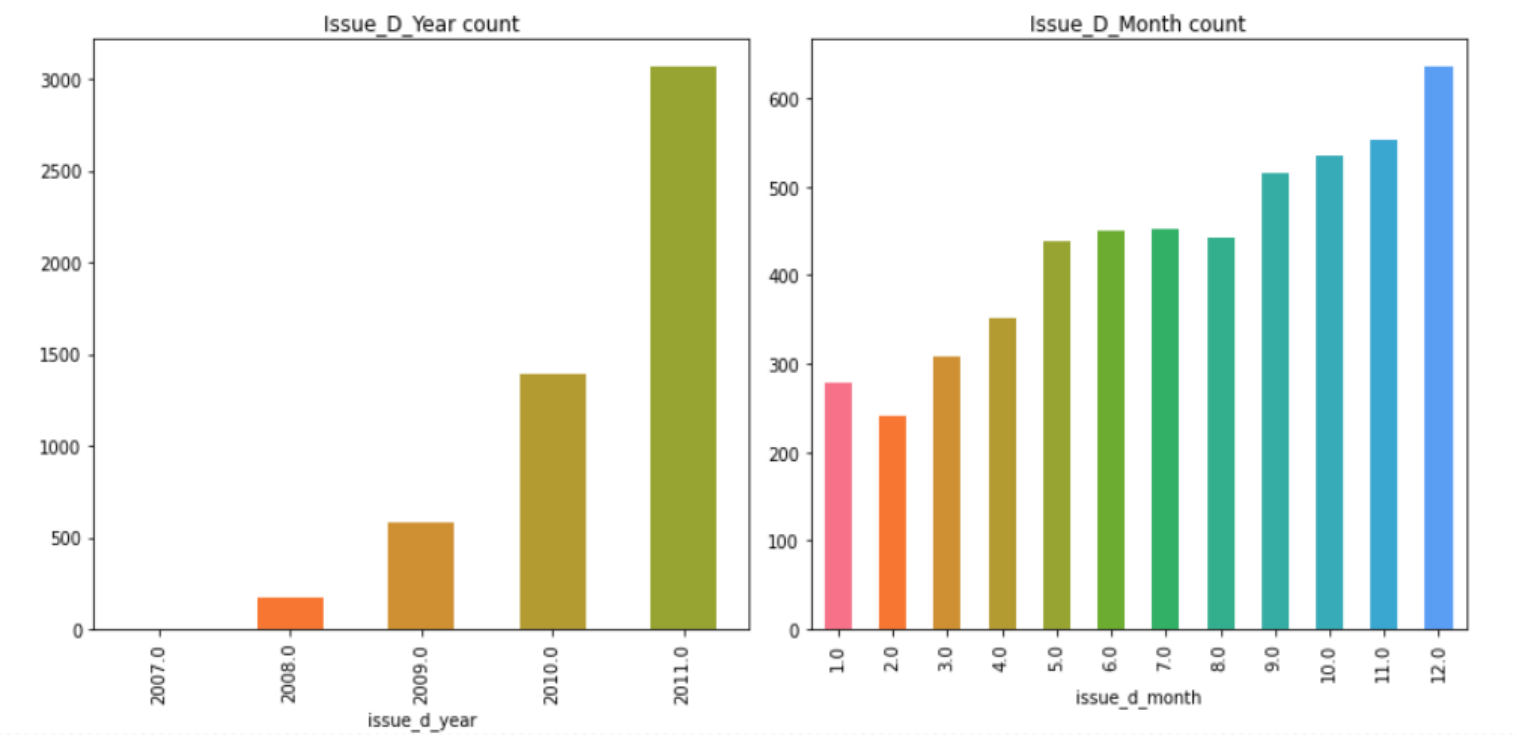
Here We did analysis on loan_amount over categorical fields like **Issue_date and issue_month** default and non defaulters

Observation:

- LC has been giving loans from year 2007 to 2011.
- Most of the loans are given in ```**2011**``` year
- -Most of the loans are issued in **December**

Analysis over categorical Variable

(only defaulters)



Statement:

Here We did analysis on loan_amount over categorical fields like **Issue_date and issue_month Using <u>defaulters value</u>**

Observation:

- Most of the loans are given in **2011** year
- -Most of the loans are issued in **December**
- **Graph is showing same pattern for these two fields**

Only for default values

Segmented Univariant Analysis
(only defulters)

loan_amnt_bins	0-5k	5-10k	10-15k	15-20k	20-25k	25-30k	30-35k
loan_status							
Charged Off	13.4952	12.637273	13.422999	17.151024	18.579235	20.520231	24.166667
Fully Paid	86.5048	87.362727	86.577001	82.848976	81.420765	79.479769	75.833333

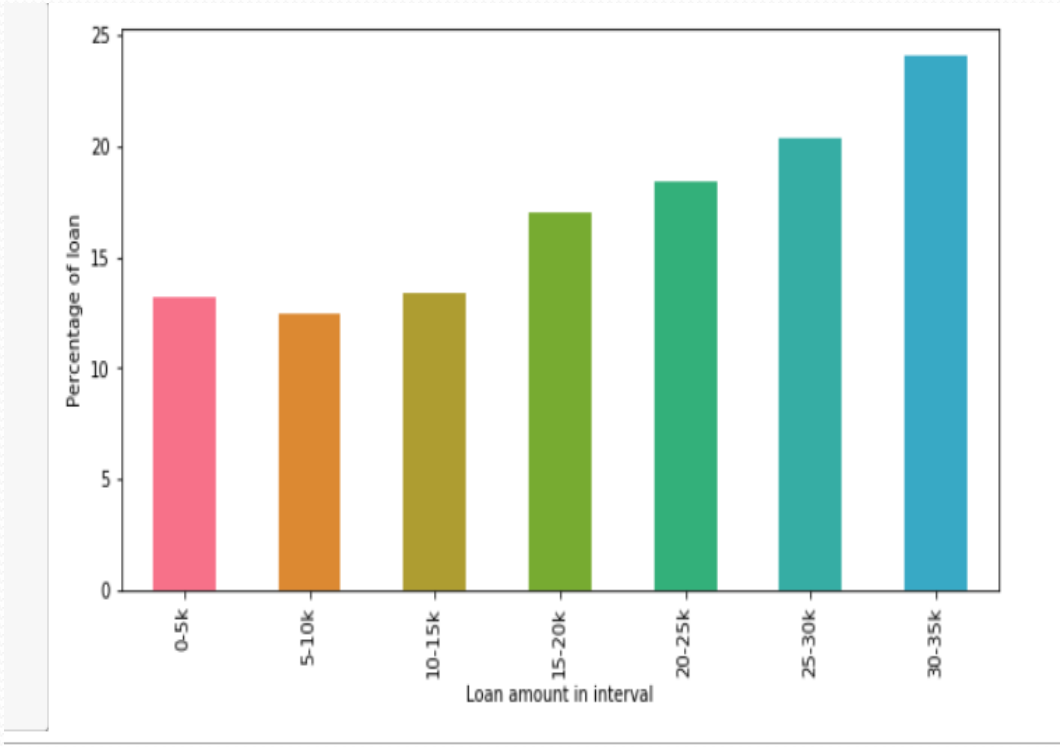
Loan Amount Distribution

•We created bins using cut method here.

•Checked distribution of number of loans of **defaulter** over this bin

Observation:

Most of the **defaulters loans falls under 30-35k** bin



DTI distribution

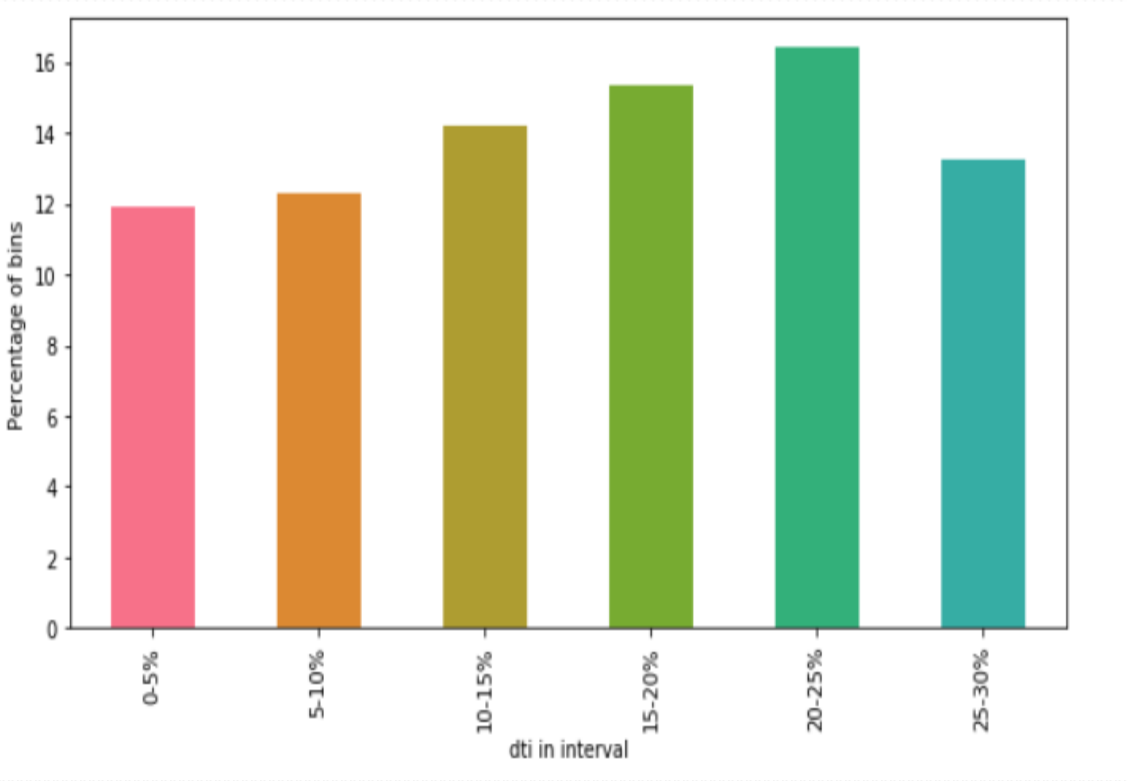
dti_bins	0-5%	5-10%	10-15%	15-20%	20-25%	25-30%
loan_status						
Charged Off	11.931444	12.280702	14.18809	15.332622	16.447681	13.255034
Fully Paid	88.068556	87.719298	85.81191	84.667378	83.552319	86.744966

•We created bins using cut method here.

•Checked distribution of percentage of bins **defaulter**

Observation:

Most of the **defaulters loan having 20-25% dti**



Installment distribution

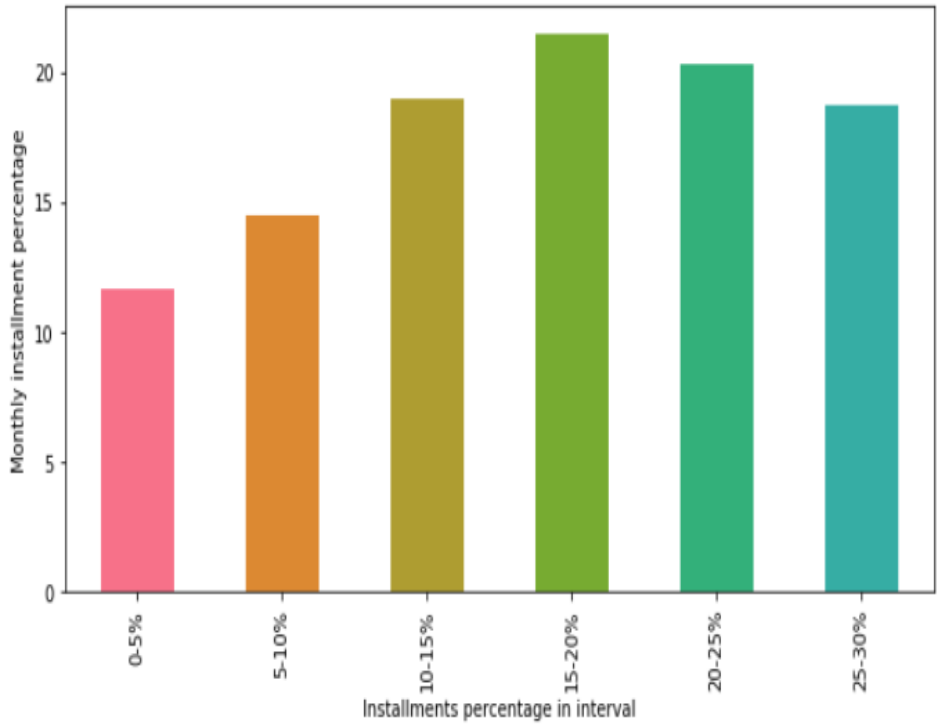
installment_monthly_perc_bins	0-5%	5-10%	10-15%	15-20%	20-25%	25-30%
loan_status						
Charged Off	11.636646	14.448539	18.935246	21.500843	20.27972	18.75
Fully Paid	88.363354	85.551461	81.064754	78.499157	79.72028	81.25

•We created bins using cut method here.

•Checked installment bin for **defaulter's**

Observation:

Most of the **defaulters loans having 15-20 % dti**



Annual Income Analysis

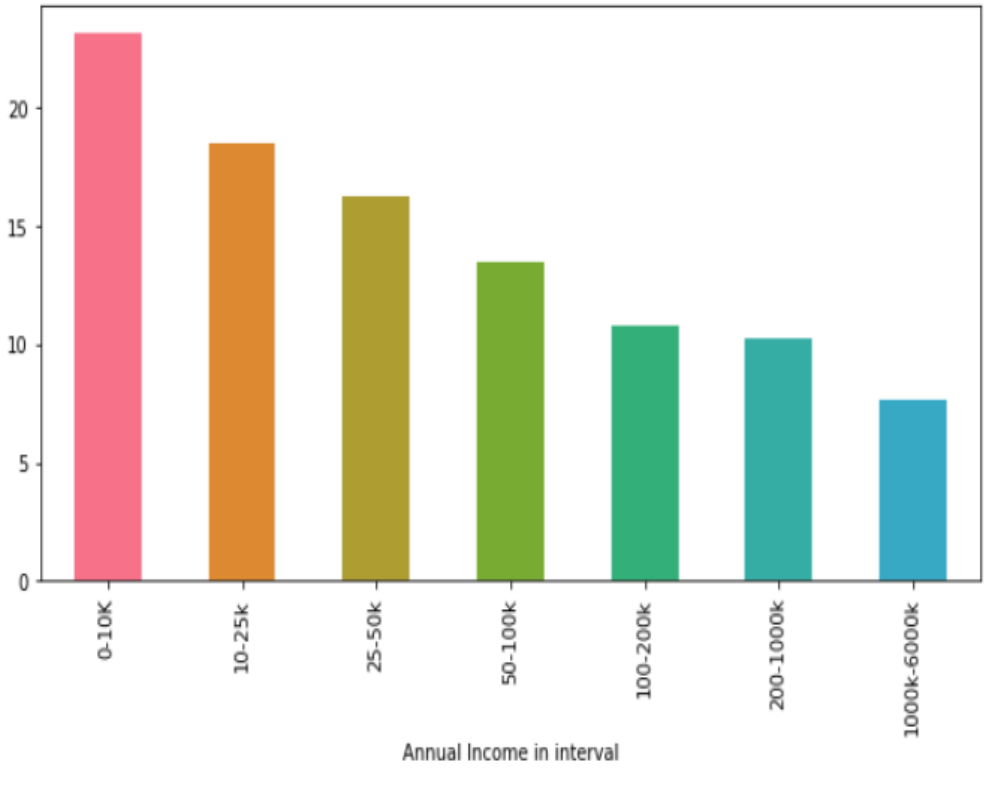
annual_inc_bins	0-10K	10-25k	25-50k	50-100k	100-200k	200-1000k	1000k-6000k
loan_status							
Charged Off	23.188406	18.497653	16.296662	13.515247	10.758304	10.280374	7.692308
Fully Paid	76.811594	81.502347	83.703338	86.484753	89.241696	89.719626	92.307692

•We created bins using cut method here.

•Checked installment bin for **defaulter's**

Observation:

Customers having annual income between 0-10k becomes bad customers in future and has more than 20 % chances



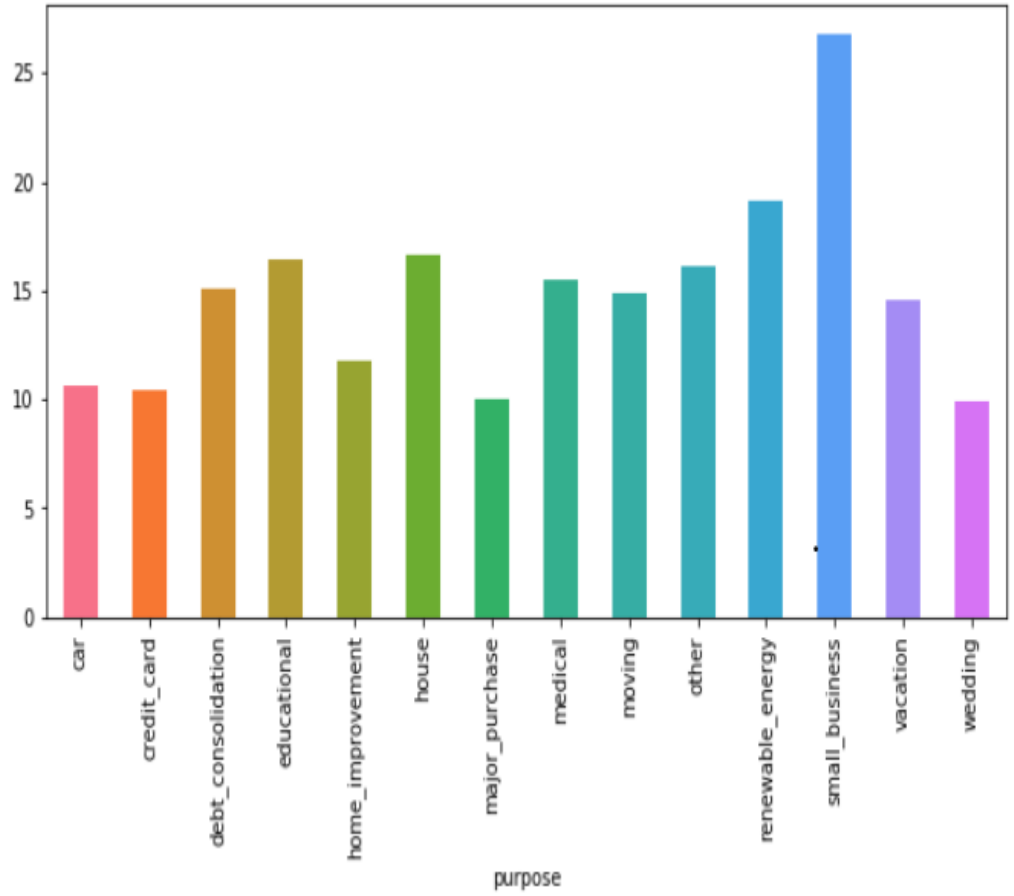
Defaulters percentage w.r.t Purpose

purpose loan_status	car	credit_card	debt_consolidation	educational	home_improvement	house	major_purchase	medical	moving	other
Charged Off	10.621943	10.456155	15.117349	16.38796	11.72817	16.618911	10.043668	15.455951	14.917127	16.13082
Fully Paid	89.378057	89.543845	84.882651	83.61204	88.27183	83.381089	89.956332	84.544049	85.082873	83.86918

Observation:

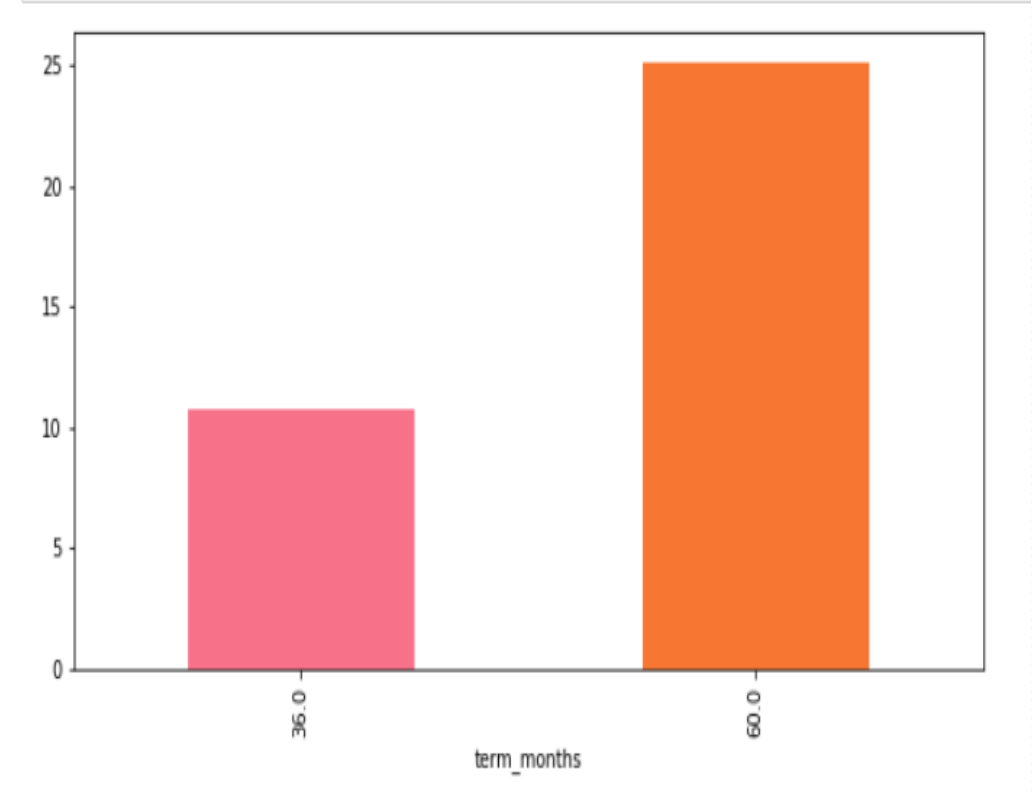
•People has taken loan for various purpose.

• Default customer with purpose **small business has high percentage (26.7 %) and renewable energy percentage (19.14%)**



Defaulters w.r.t terms month

term_months	36.0	60.0
loan_status		
Charged Off	10.710402	25.126931
Fully Paid	89.289598	74.873069



Lending company provide customer two windows to repay an amount.

Observation:.

- Default customer with terms **60 month (24.9 %)** have more percentage than 30 month (10.5 %)
- Term month with **60 are seems to be high risky**

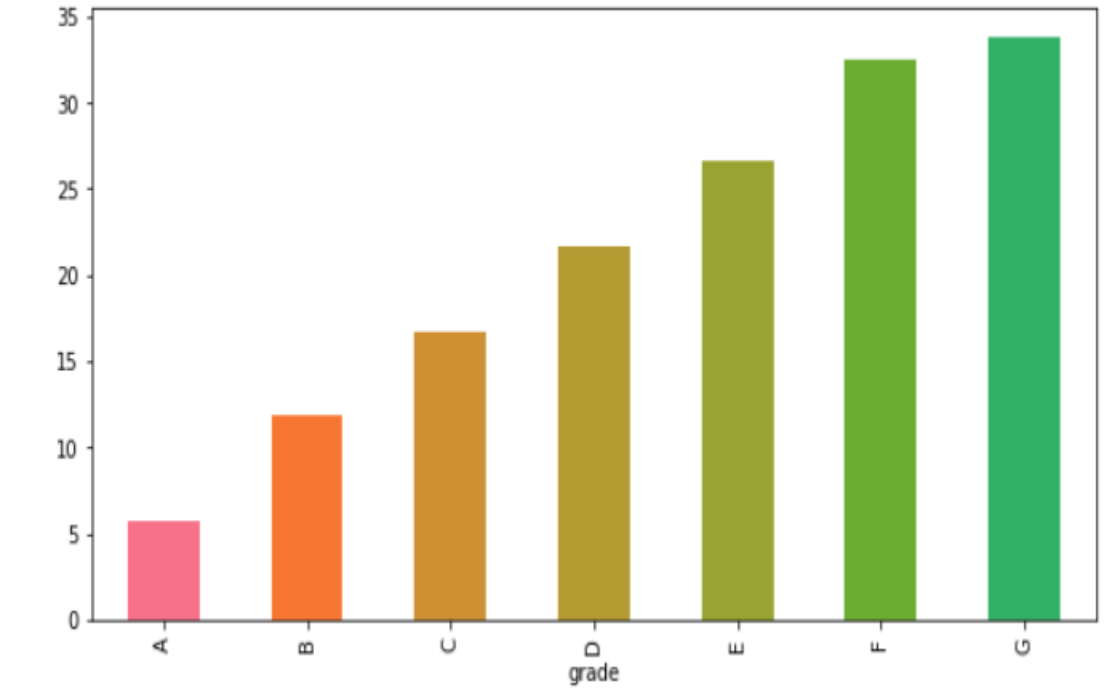
Defaulters w.r.t grade

grade	A	B	C	D	E	F	G
loan_status							
Charged Off	5.799538	11.892377	16.751269	21.610518	26.692456	32.521186	33.783784
Fully Paid	94.200462	88.107623	83.248731	78.389482	73.307544	67.478814	66.216216

Lending company give grade to its customer depending on various factor as credit score and previous loan amount.

Observation:.

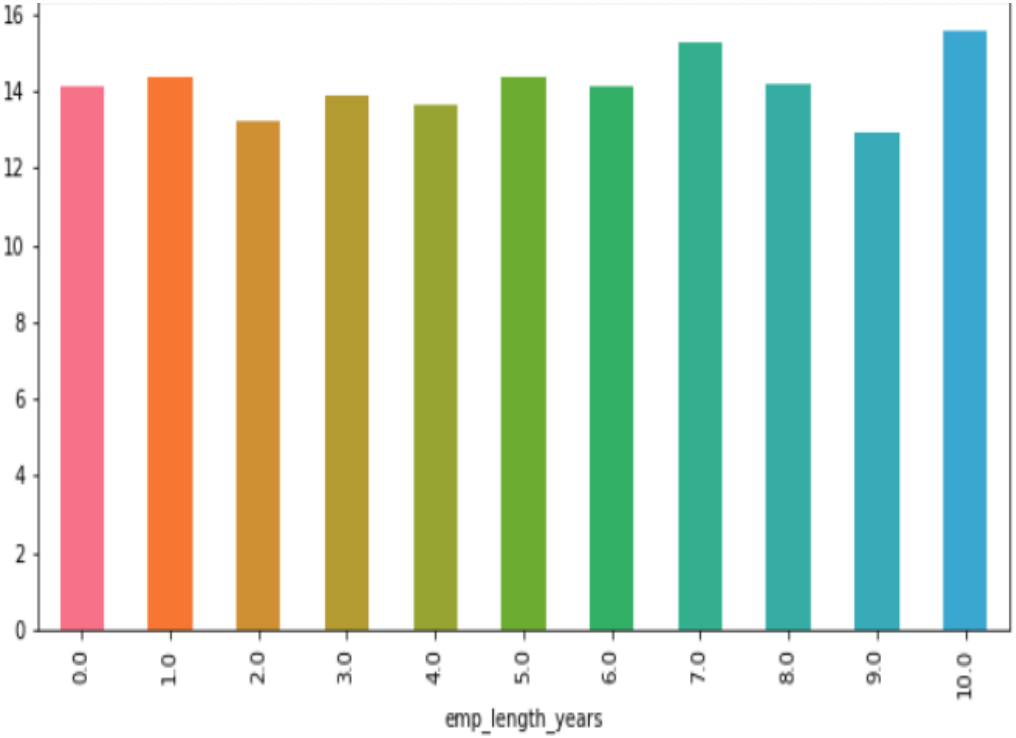
- E, F and G grade have high percentage of defaulter
- E, F and G grade seems to be high risky



Defaulters w.r.t emp_length

emp_length_years	0.0	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0
loan_status											
Charged Off	14.098134	14.364641	13.216068	13.895216	13.648772	14.362543	14.09176	15.275311	14.184397	12.903226	15.569363
Fully Paid	85.901866	85.635359	86.783932	86.104784	86.351228	85.637457	85.90824	84.724689	85.815603	87.096774	84.430637

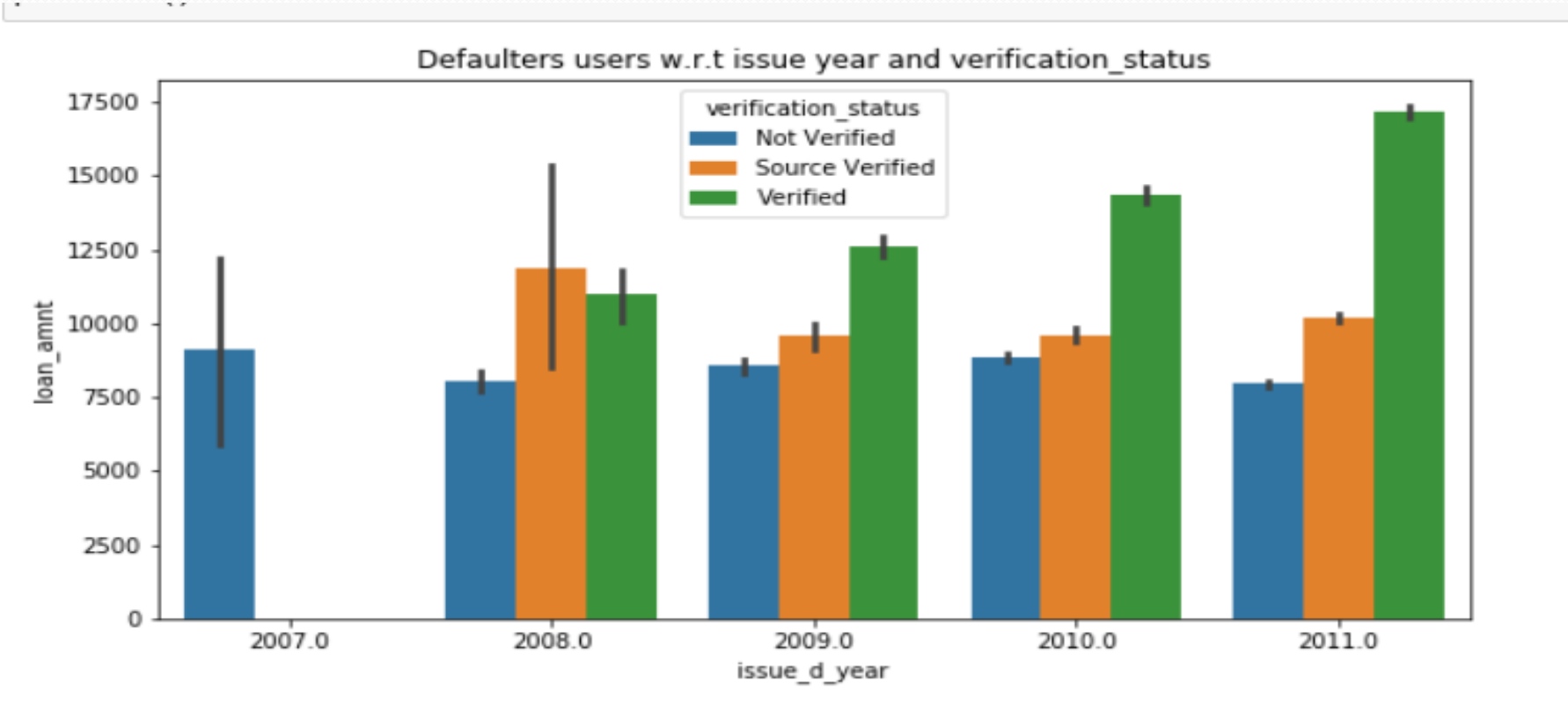
Emp_length means customer length with lending company



Observation:

There is no much difference in the defaulters with respect to years of experience

Default users w.r.t issue year and verification status



Lending company really need to improve its verification process.
As we can see in graph, There is a lot money stuck under not verified

Conclusion

- After studying and analyzing all the data , we conclude following categorical and quantitative fields which are impacting defaulters and increased risk factor to lending company.

1. **Term-** Duration of loan having 60 month is impacting more as compare to 36 month.
2. **Verification** : Lending company really need to improve its verification process. Non verified customers are more in number
3. **Grade:** E, F, G grade has high percentage of defaulter which seems to be risky.
4. **Purpose:** Customer having purpose as Small business , house , renewable energy seems to be more defaulters.
5. **Annual Income**: Annual income in range of 0 to 50K are seems to be more defaulter
6. There is no much difference in the defaulters with respect to **employee years of experience**
7. **Home ownership** of most of the loan borrowers is RENT and MORTAGE.
8. Most of the loans are from California(CA), following that is from New york(NY) and Florida(FL) **state**

We able to derived all theses risky factor after comparing it with following fields:

- total loan amount installment , annual income , dti, ratio of annual income to installment etc.

Suggestion

- Company needs to improve its verification process.
- They should analyze customer more on following factors:

purpose of loan, their income source , fixed annual income amount and their home_ownership factors before providing the loan.