# Lending Club Case Study:

upGrad

**Group:** Individual
Name: Vishal Jain

# Business Requirement

upGrad

- **Problems**: I have lending loan data of consumer finance company which are providing various type of loan to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant's profile. Two types of risks are associated with the bank's decision.
  - If the applicant is likely to repay the loan, then not approving the loan results in a loss of business to the company.
  - If the applicant is not likely to repay the loan, i.e. he/she is likely to default, then approving the loan may lead to a financial loss for the company.
- The data given contains the information about past loan applicants and whether they 'defaulted' or not. 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, etc.
- I will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.
- When a person applies for a loan, there are two types of decisions that could be taken by the company:

1. Loan accepted: If the company approves the loan, there are 3 possible scenarios described below:
   1. Fully paid: Applicant has fully paid the loan (the principal and the interest rate).
   2. Current: Applicant is in the process of paying the instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
   3. Charged-off: Applicant has not paid the instalments in due time for a long period of time, i.e. he/she has defaulted on the loan
2. Loan rejected: The company had rejected the loan (because the candidate does not meet their requirements etc.). Since the loan was rejected, there is no transactional history of those applicants with the company and so this data is not available with the company (and thus in this dataset)

# Process to Achieve Goal

upGrad

- We have lending club loan history data in csv format where we can do the analysis on approved loan on the behalf of many features.

- EDA Process :
    - ❑ Data Cleaning
    - ❑ Data sanity Testing
    - ❑ Binning
    - ❑ Deciding Target Frames
    - ❑ Univariate Analysis
    - ❑ Multivariate Analysis
    - ❑ Conclusion

# Data Cleaning

- Firstly we need to remove attached irrlevant-variables-removed.txt file features from loan data frames

irrlevant-variables-removed.txt

- I imputed **emp_length** feature from median that needs to be imputed since its low number to drop these values.
- I have handled data type conversion for features: **emp_length**, **term**, **int_rate** etc.
- I have dropped some records due to missing values into **emp_title** etc.

# Sanity Check

- I have filtered data and do some sanity testing like funded amount should not be equal or less than zero etc.
- eg: **loandf.funded_amnt_inv > 0**

# Deciding Target Frames

- I am extracting 2 target frames using below loan status:
  - Fully Paid
  - Charged off

## Binning

- I have created binning for interest rate, funded amount, annual amount to identify the relation with deciding factor.

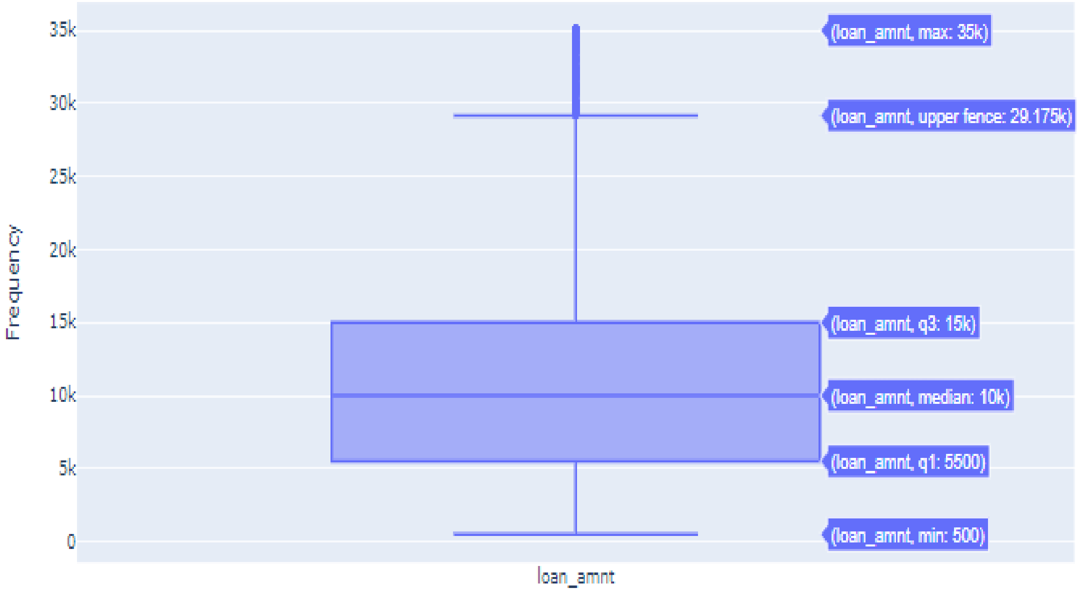
## Univariate Analysis

- I have done univariate analysis on the behalf of some of the important fields like Annual Income, Funded Income Inv, Verification Status, Grade, Interest rate etc.
- I can achieve univariate analysis with the help of boxplot, pie chart, bar plot etc.
- I am using univariate analysis to identify the frequency order, outliers etc.

# Loan Amount

Please have a look below stats:

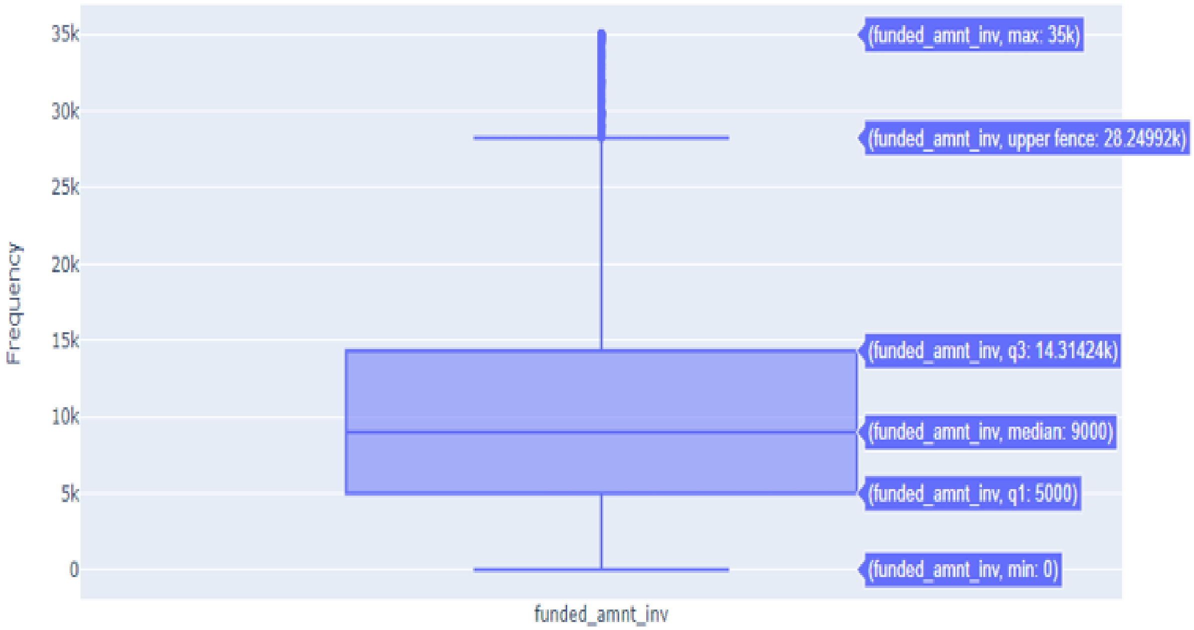
- Count 37057.000000
- Mean 11230.901044
- Std 7383.178753
- Min 500.000000
- 25% 5500.000000
- 50% 10000.000000
- 75% 15000.000000
- Max 35000.000000
- Name: loan_amnt, dtype: float64



# Funded Amount Inv

Please have a look below stats:

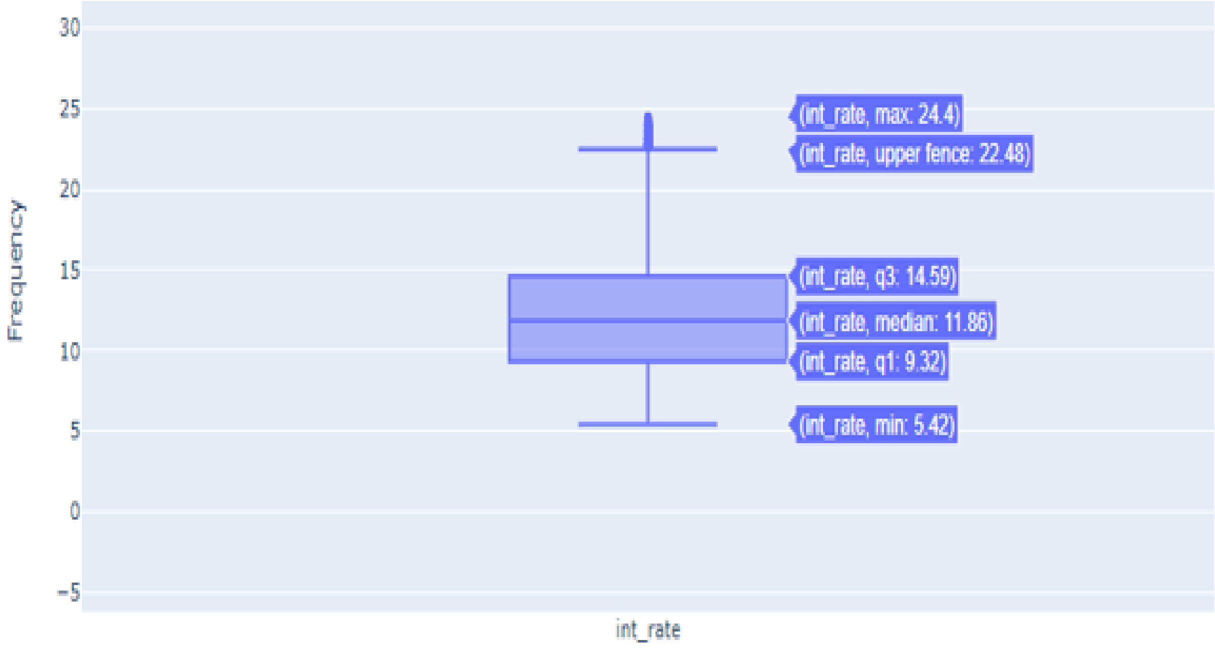
- Count 36844.000000
- Mean 10357.971056
- Std 7017.898541
- Min 0.000000
- 25% 5000.000000
- 50% 9000.000000
- 75% 14313.362500
- Max 35000.000000
- Name: funded_amnt_inv
  dtype: float64



# Interest rate

Please have a look stats:

- Count 36844.000000
- Mean 12.039300
- Std 3.709402
- Min 5.420000
- 25% 9.320000
- 50% 11.860000
- 75% 14.590000
- Max 24.400000
- Name: int_rate, dtype: float64

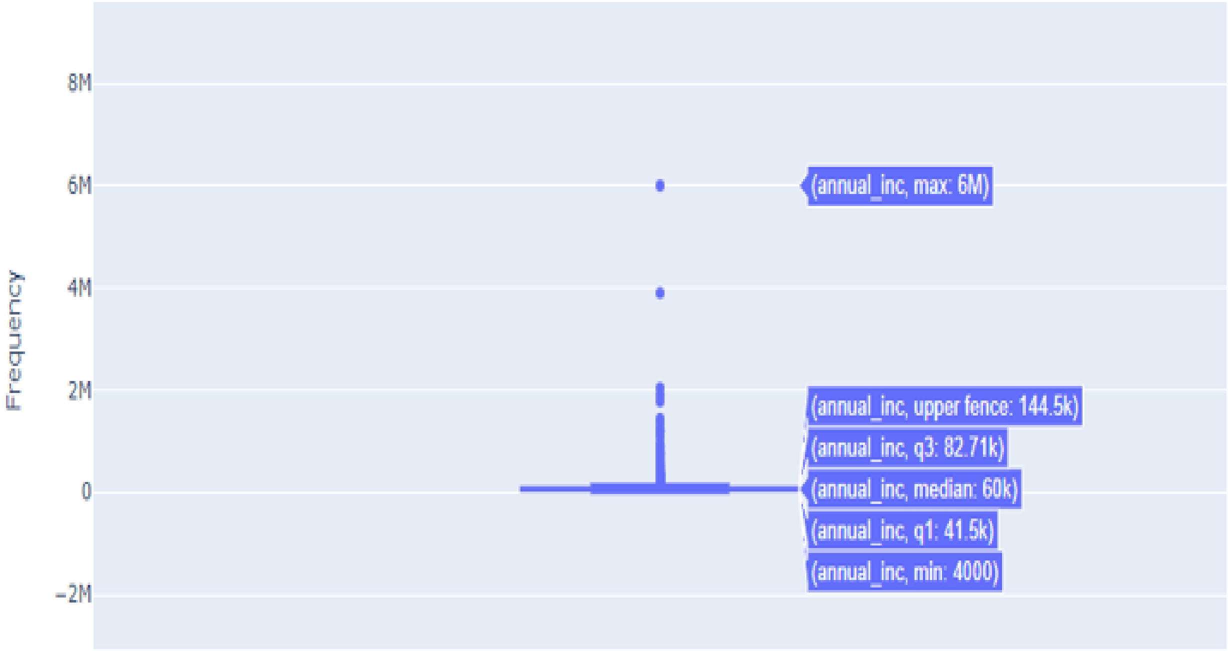


upGrad

# Annual Income

Please have a look stats:

- Count 36844.000000
- Mean 65592.048841
- Std 34142.373826
- Min 4000.000000
- 25% 41000.000000
- 50% 59000.000000
- 75% 81000.000000
- Max 224000.000000
- Name: annual_inc, dtype: float64

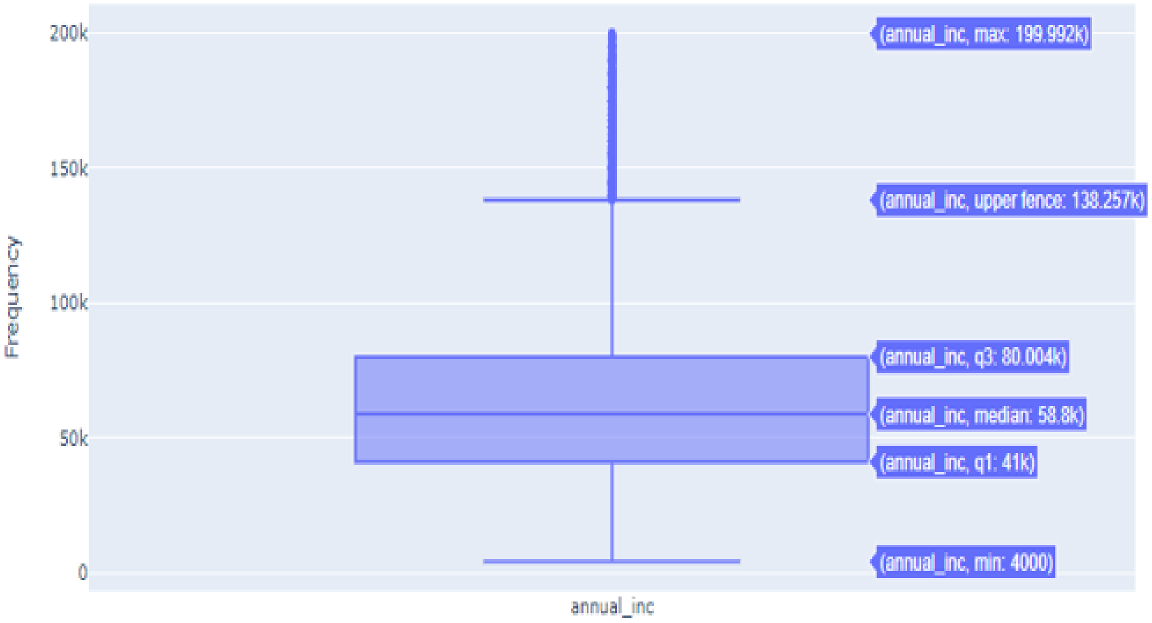


# Annual Income

**Annual income after removing the outliers.**

**Please have a look stats:**

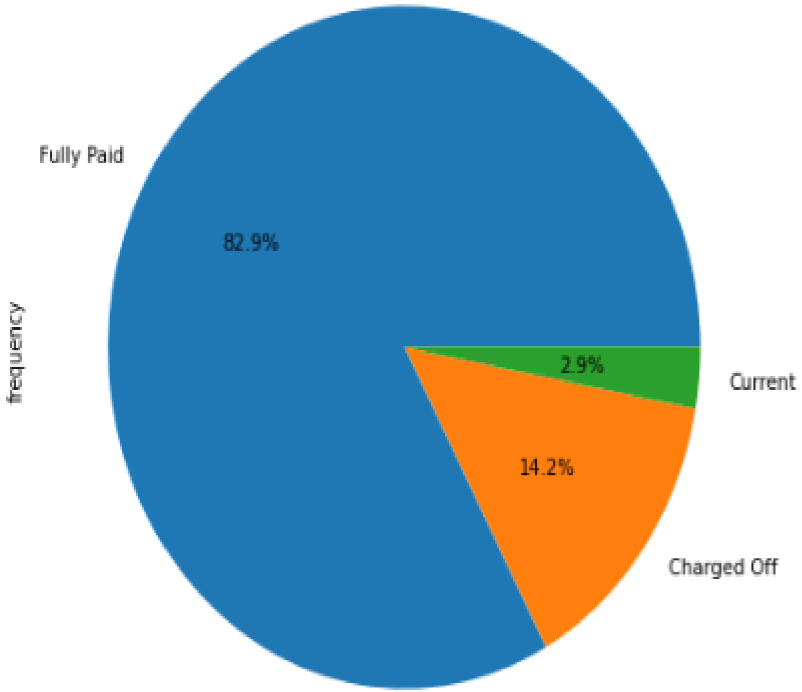
- **Count** **36634.000000**
- **Mean** **64792.179819**
- **Std** **32555.369624**
- **Min** **4000.000000**
- **25%** **41000.000000**
- **50%** **58800.000000**
- **75%** **80004.000000**
- **Max** **199992.000000**
- **Name:** **annual_inc,**
  **dtype:** **float64**



# Loan Status Frequency

**We can see here stats of Loan status:**

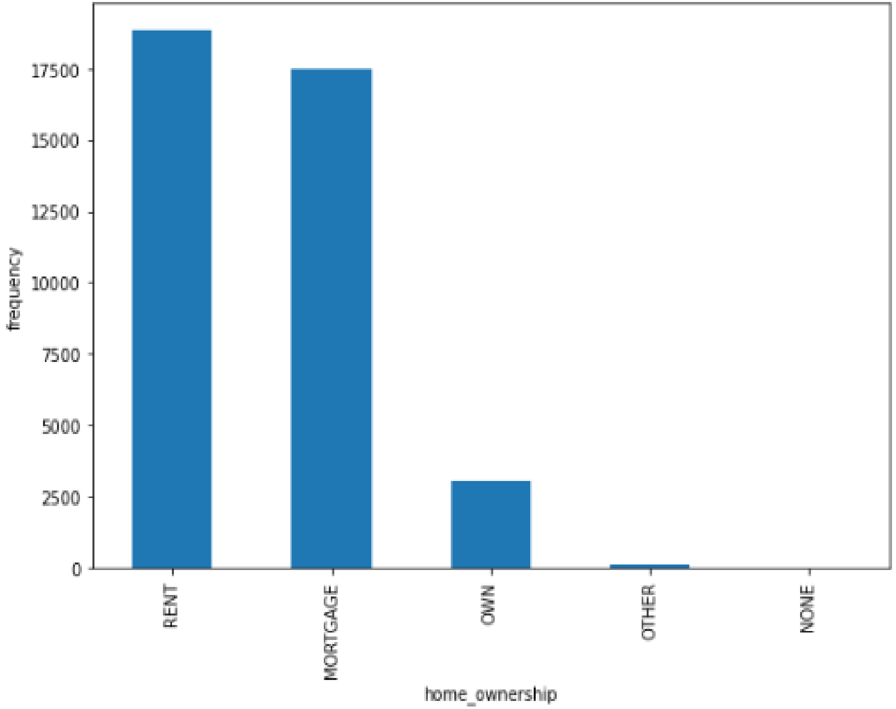
- **Fully Paid: 82.9%**
- **Current: 2.9%**
- **Charged Off: 14.%**



# Home Ownership Frequency

**We can see here stats of Loan status:**

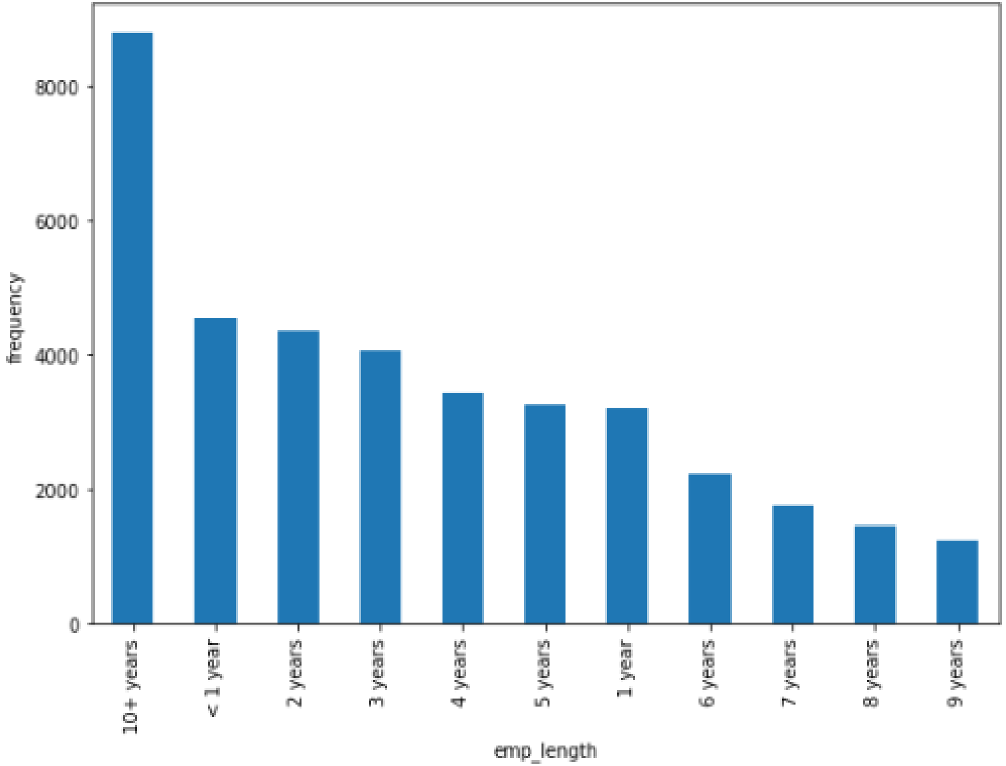
- **Count:** **39494**
- **Unique:** **5**
- **Top:** **RENT**
- **Freq:** **18848**
- **Name:** **home_ownership, dtype: object**



# Emp Length Frequency

We can see here stats of Loan status:

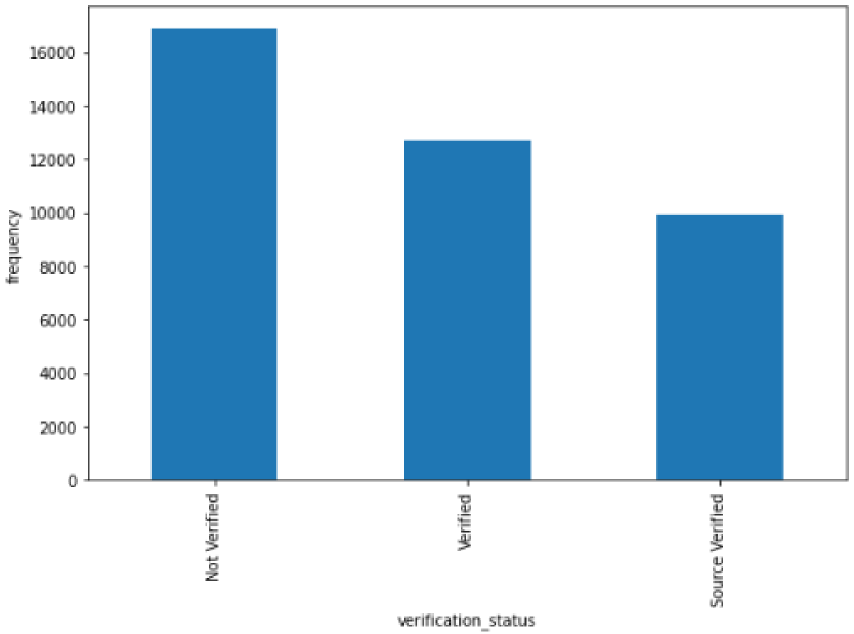
- count: 38422
- Unique: 11
- Top: 10+ years
- Freq: 8796
- Name: emp_length, dtype: object



# Verification Status Frequency

We can see here stats of Loan status:

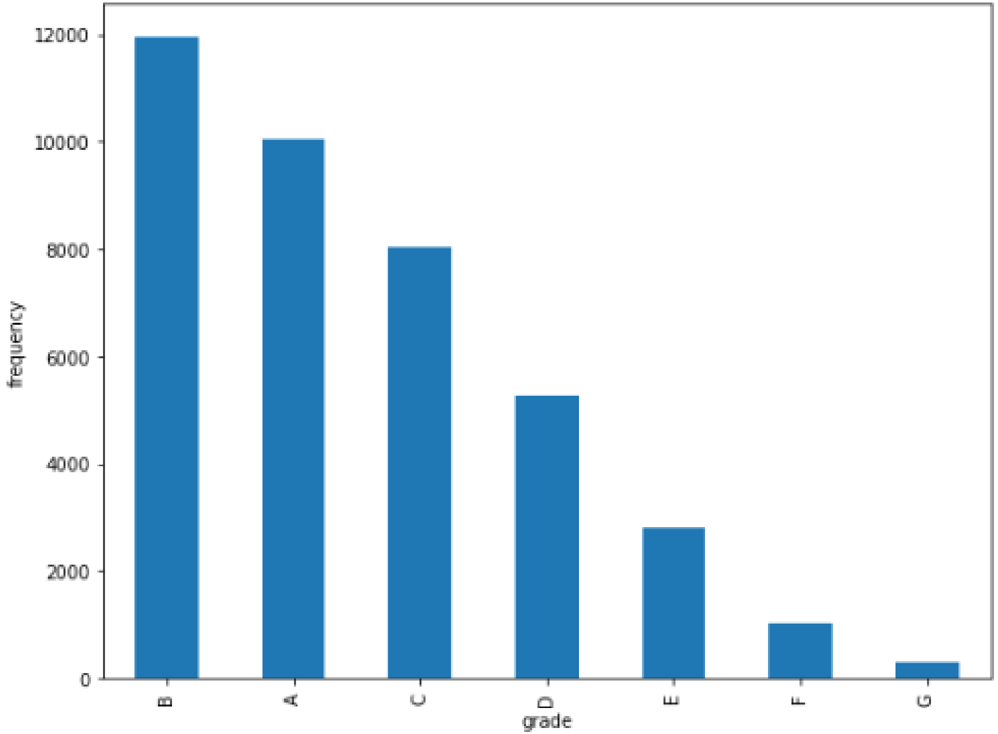
- Count: 39494
- Unique: 3
- Top: Not Verified
- Freq: 16860
- Name: verification_status, dtype: object



# Grade Frequency

We can see here stats of Loan status:

- Count 39494
- Unique 7
- Top B
- Freq 11967
- Name: grade, dtype: object



# Multivariate Analysis

- We are performing multivariate analysis to identify the correlation status among the features.

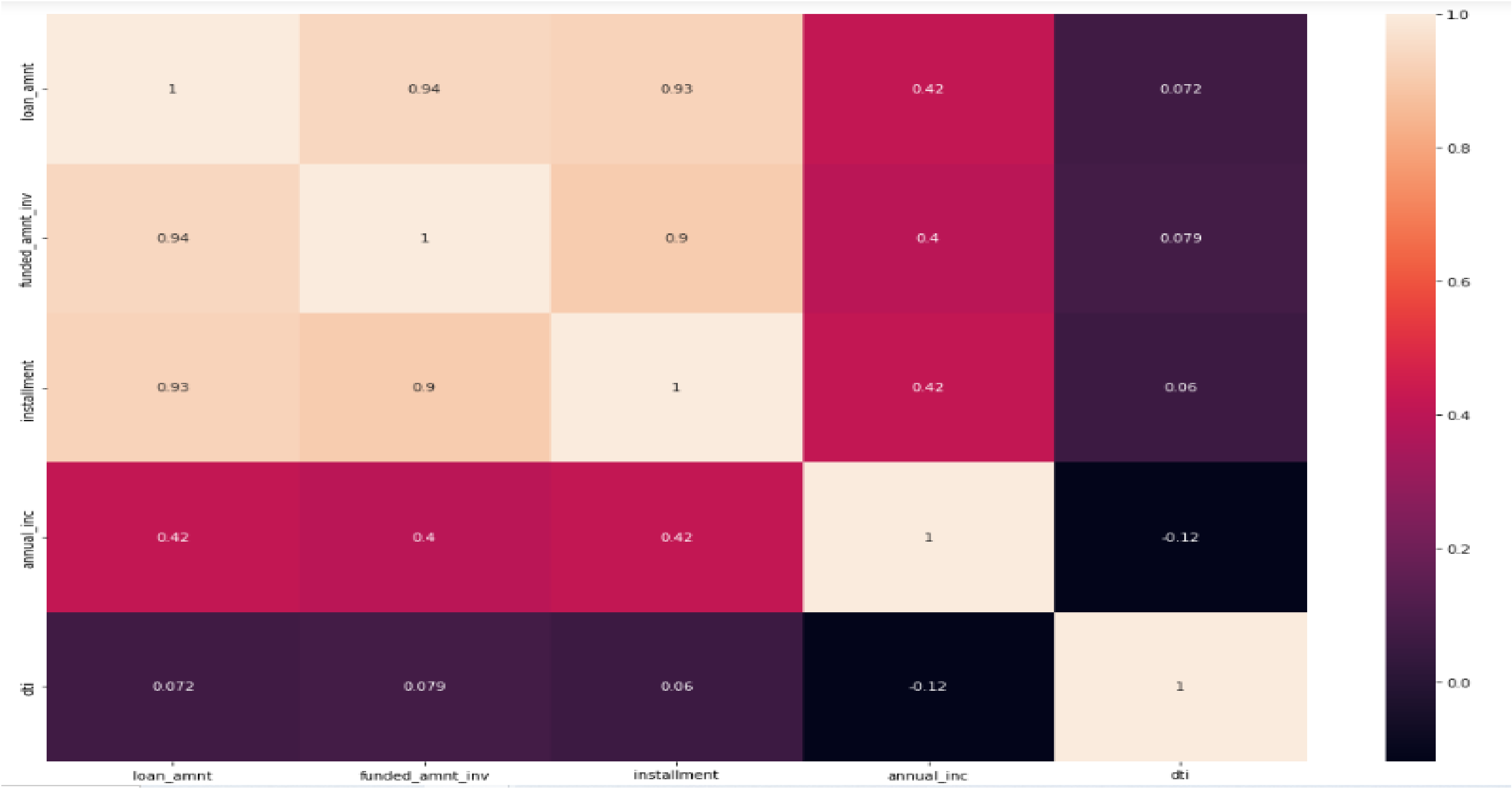
# Create the Heatmap identify the correlation

**Observation:**

- We are creating the heatmap to identify the correlation.
- We can see Loan Amount is highly correlated with Funded Amount Inv

# Lending Club: EDA Case Study

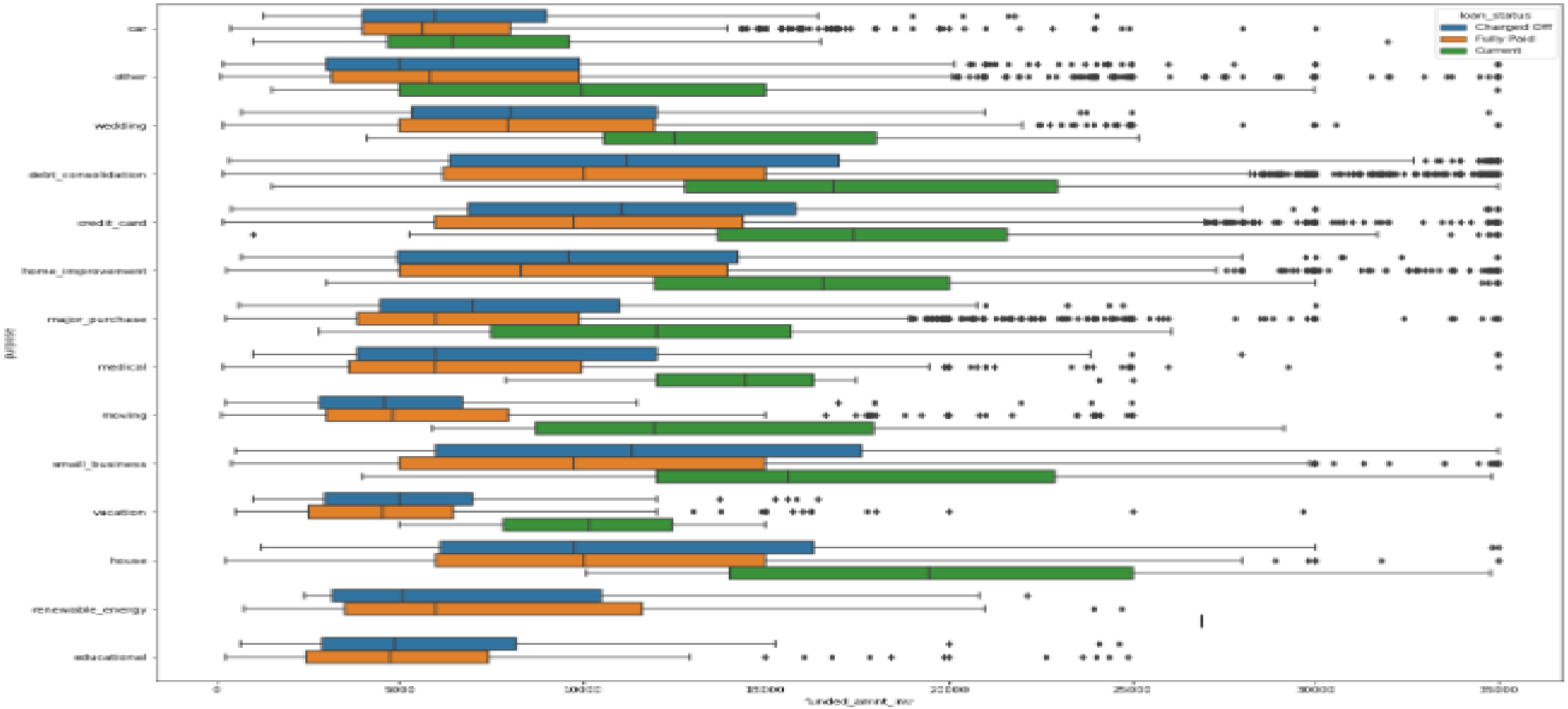
upGrad



# Identify Purpose Of Loan VS Loan Amount per loan status: Boxplot

**Observation:**

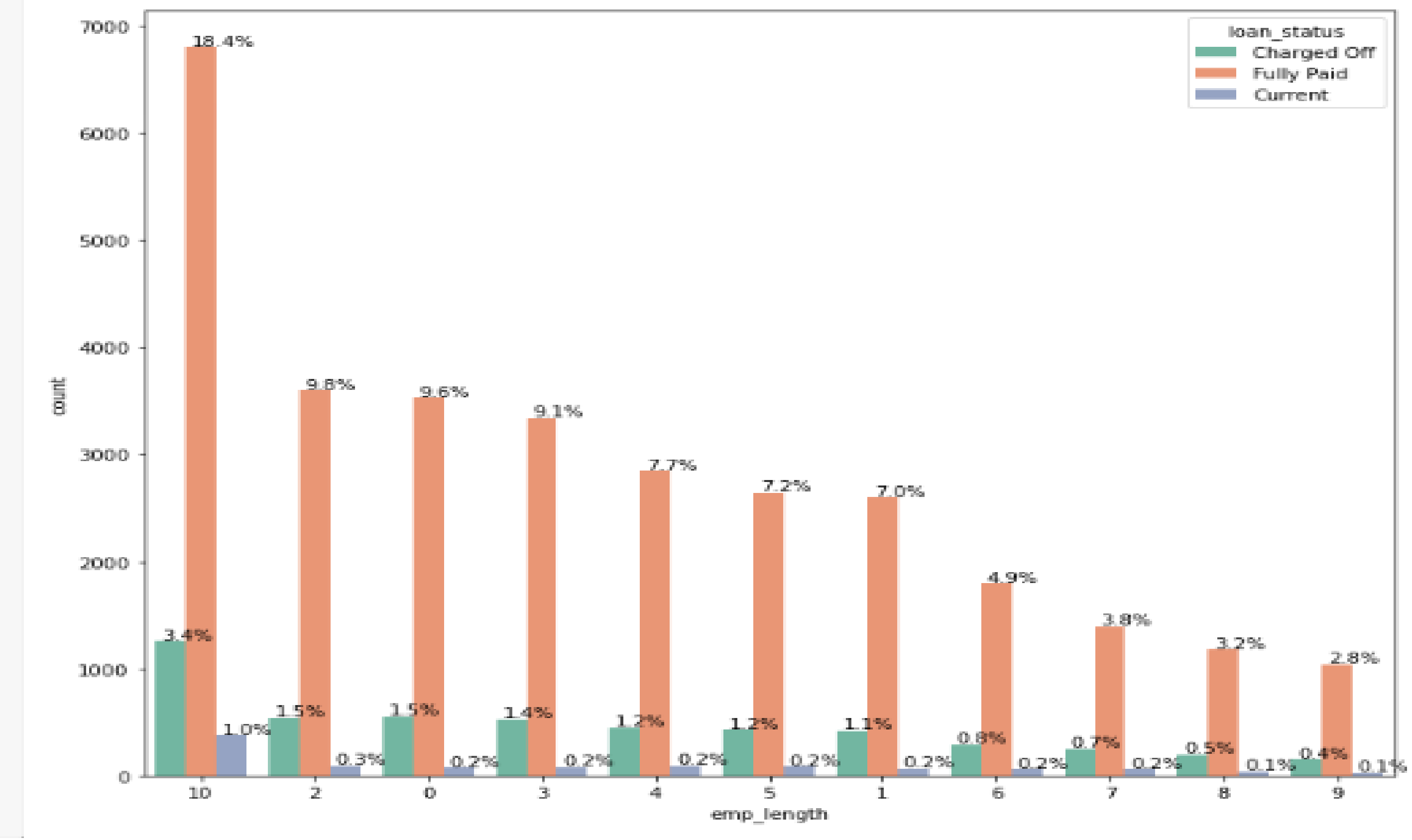
- Small Business: We can not see any clear indication of outliers for Small Business. So Bank should do more analysis while giving the loan.
- We can see that Others and Major Purchase categories have lots of outliers which can be contribute more losses.


loan_status
Charged Off
Fully Paid
Current
car
other
wedding
debt_consolidation
credit_card
home_improvement
major_purchase
medical
moving
small_business
vacation
house
renewable_energy
educational
purpose
0
5000
10000
15000
20000
25000
30000
35000
funded_amnt_inv

# Employee Length VS Loan Status: Bar plot

**Observation:**

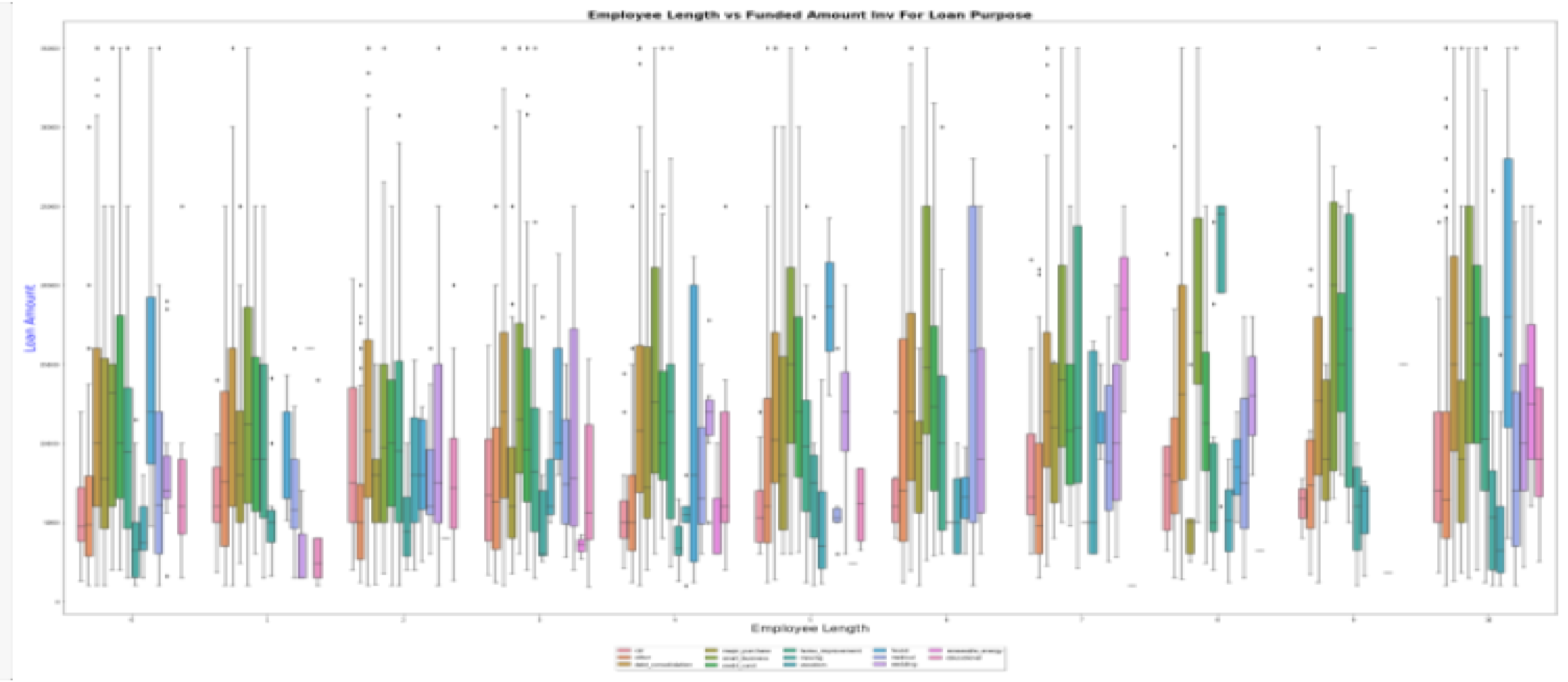
- Loan taken in maximum number whose employee length is >= 10 Years. 18 % Loan has been taken.
- Loan taken in minimum number whose employee length is 1 year as compared to others[0 to 5 years] and has low defaulting rate.


loan_status
Charged Off
Fully Paid
Current
count
emp_length
7000
6000
5000
4000
3000
2000
1000
0
10
2
0
3
4
5
1
6
7
8
9
18.4%
9.8%
9.6%
9.1%
7.7%
7.2%
7.0%
4.9%
3.8%
3.2%
2.8%
3.4%
1.5%
1.5%
1.4%
1.2%
1.2%
1.1%
0.8%
0.7%
0.5%
0.4%
1.0%
0.3%
0.2%
0.2%
0.2%
0.2%
0.2%
0.2%
0.2%
0.1%
0.1%

# Employee Length V/S Funded Amount For Loan Purpose: Boxplot

**Observation:**

- Loan taken in maximum number whose employee length is >= 10 Years. 18 % Loan has been taken.
- Loan taken in minimum number whose employee length is 1 year as compared to others[0 to 5 years] and has low defaulting rate.

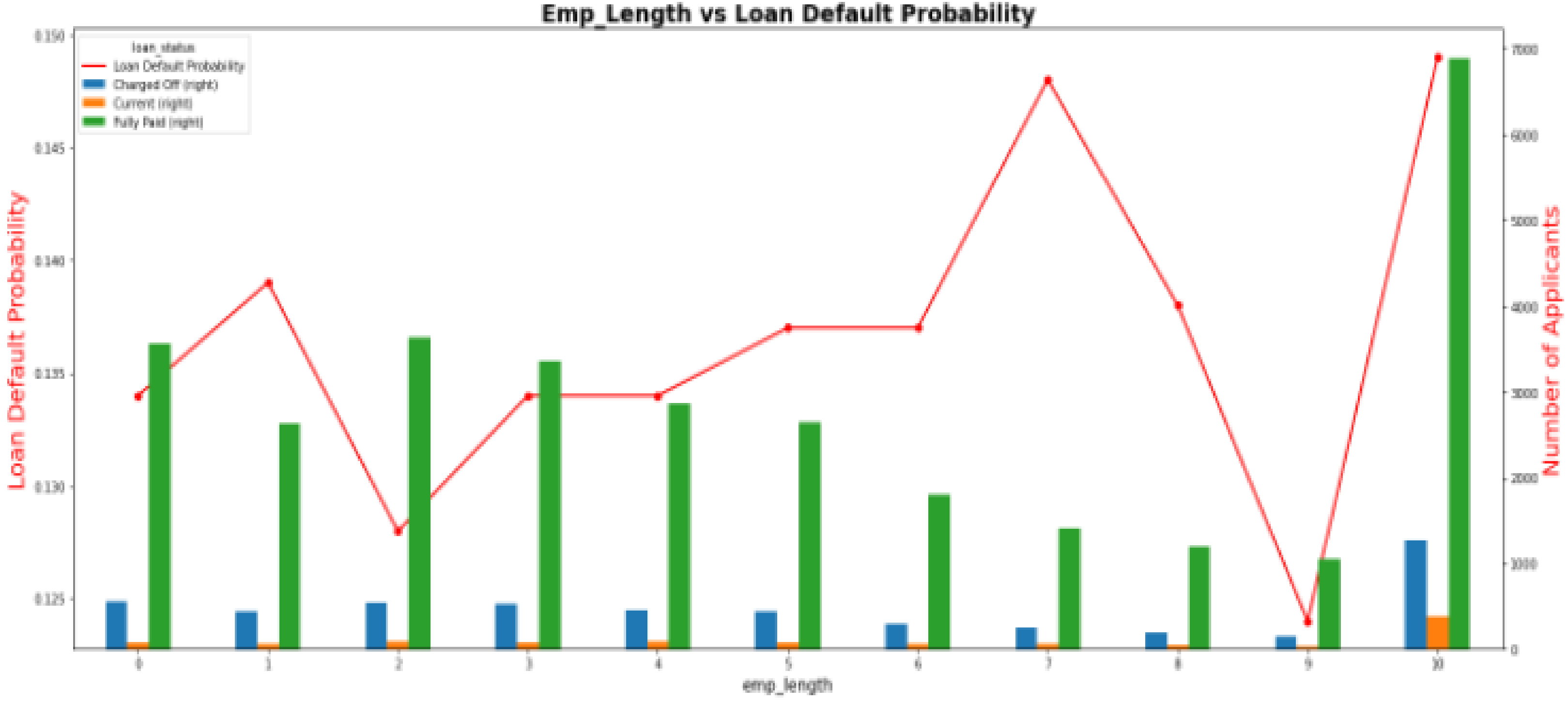

Employee Length vs Funded Amount Inv For Loan Purpose
Loan Amount
Employee Length

# Driving Variables Contributing to Loan Default

**Observation:**

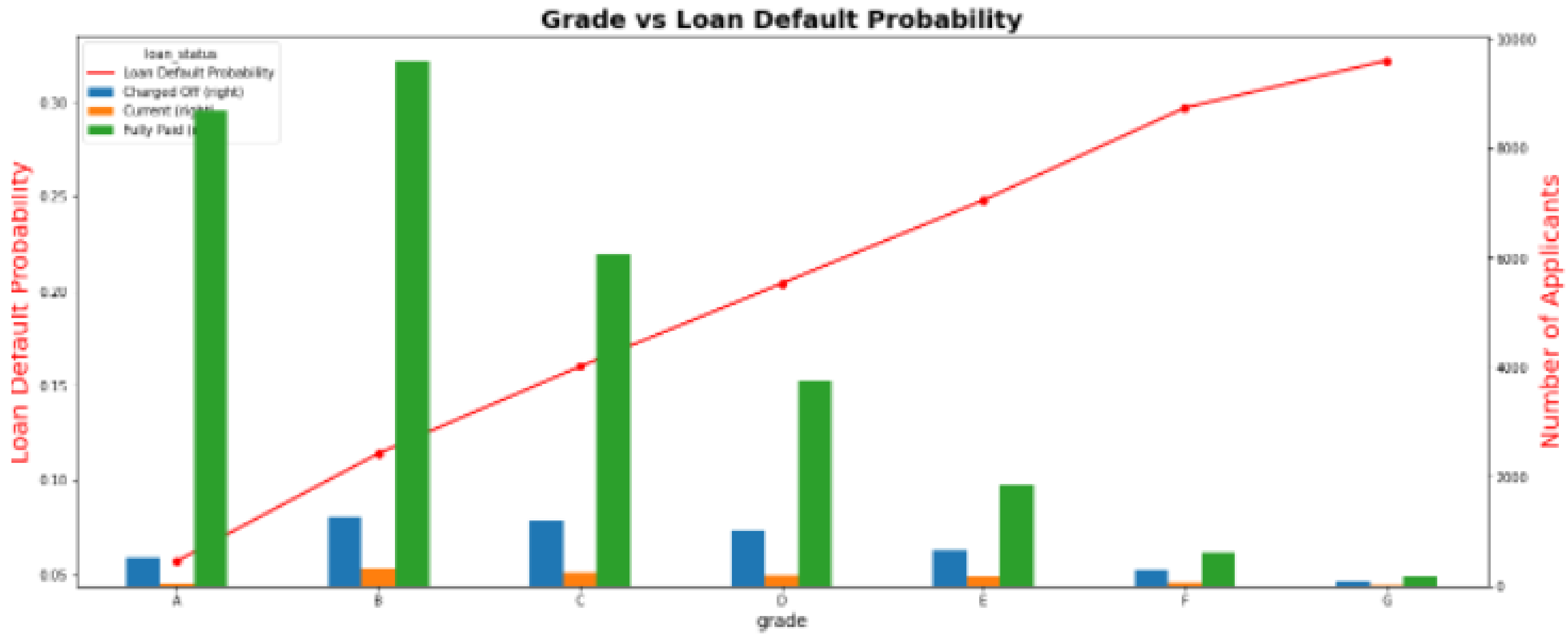
- Identifying probability of risky applicants through variables that are responsible for triggering defaulters
- Below Variables might trigger "Charged-Off":
    - 1. employment length -----> Categorical Variable
    - 2. grades ------> Categorical Variable
    - 3. purpose ------> Categorical Variable
    - 4. loan_amnt -----> Categorical Variable ( After Conversion )
    - 5. int_rate -----> Categorical Variable ( After Conversion )
    - 6. annual_inc -----> Categorical Variable ( After Conversion )

# Employee length vs Default Probability

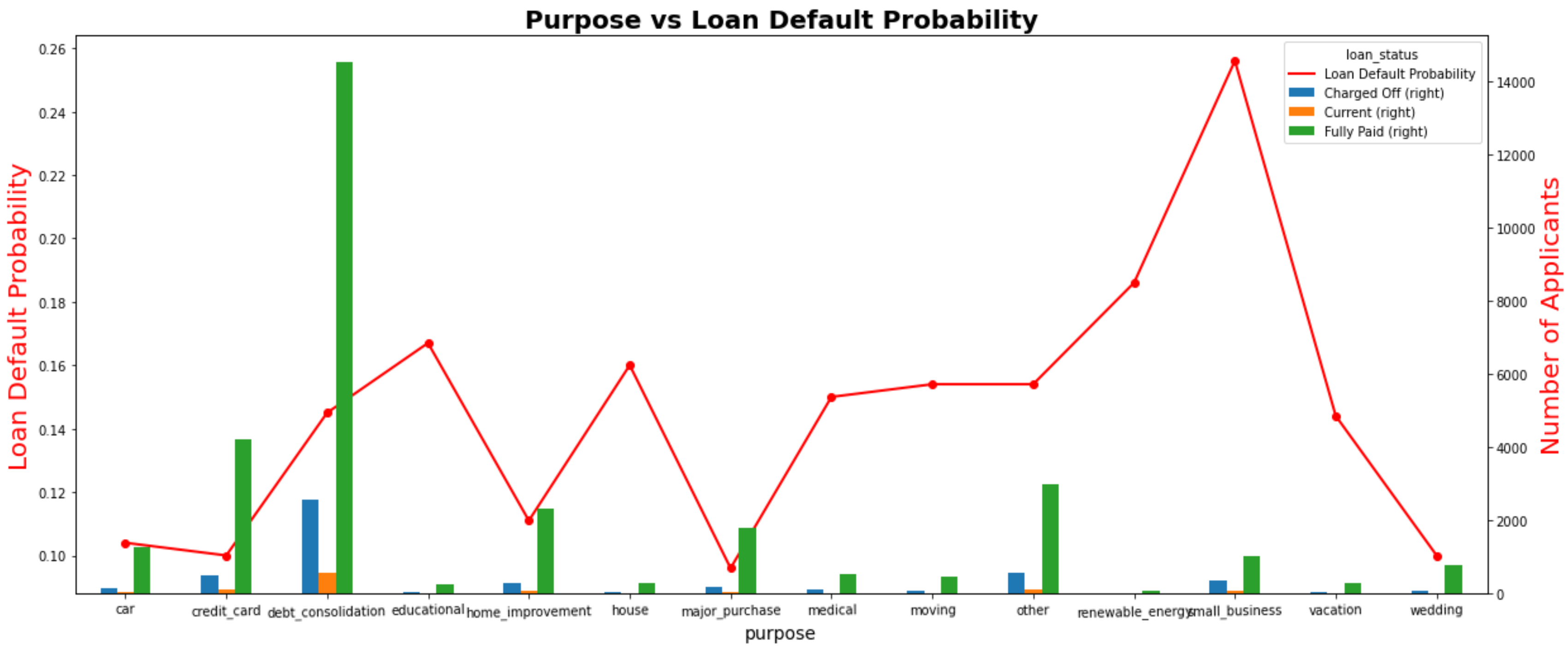


upGrad

# Grade vs Default Probability



# Purpose vs Default Probability



# Driving Variables Contributing to Loan Default

We have observed many things from above 3 slides:-

1. Employee length vs Default Probability Observation:
    - Less Defaulter rate when your employee length is 9 years
    - Higher Defaulter rate for employee length >= 10 years

2. Grades with default chances Observation :-
    - From grade A to G , Loan Probability Defaulter increasing

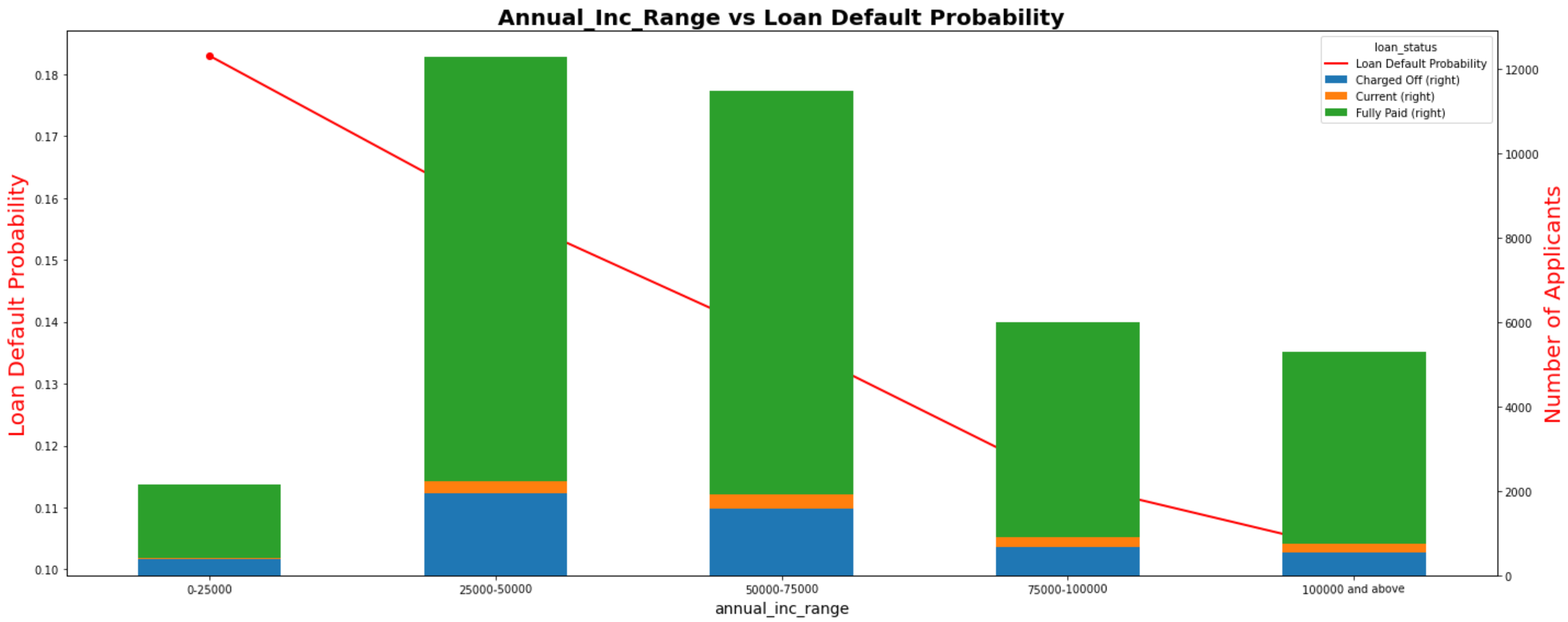
3. Purpose With Default Chances Observation :-
    - We can see that , most of the loan default probability is seen for small_business, so bank should be extra careful while approving the loan for such businesses
    - Minimum defaulter rate showing for Major Purchase

# BINNING AS WE FOUND OUTLIERS EARLIER IN ANNUAL INCOME

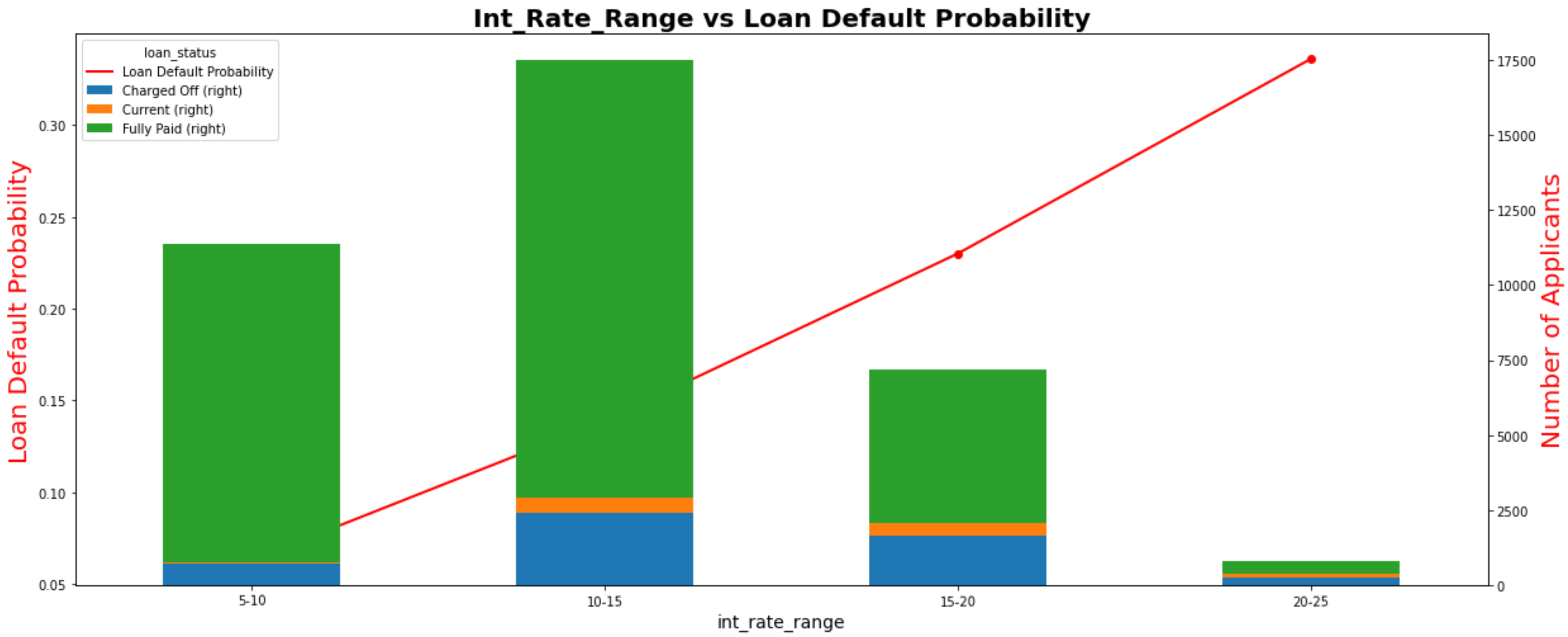
- **Annual Income range vs Loanda default Probability**
  - We can see, as annual income is increasing , probabality of being defaulter is also increasing , reaching up to 19%
- **Interest Rate vs Loan Defaulter Probability**
  - We can see that , as interest rate is increasing , chance of being defaulter is also increasing , when the interest rate touches more than 15% , risk of default rate is increasing
- **funded_amnt_range vs Loan Defaulter Probability**
  - We can see that default rate is increasing , when the loan amount/funded amount is increasing at the alarming rate

upGrad

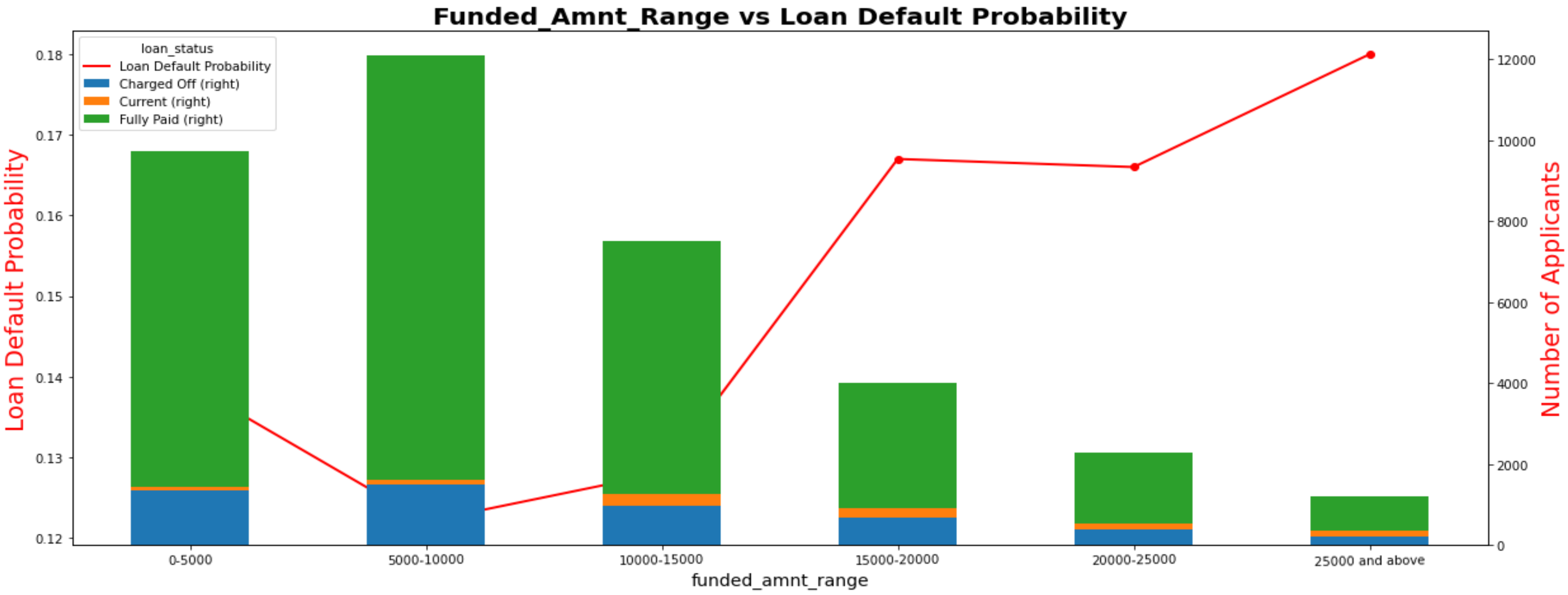
# Annual Income Range chance to be defaulter



# Interest Range and chance to be defaulter



# Loan Amount (Approved Amount) and chance of being defaulter



# Case Study Conclusion

1. Defaulter rate is higher when employee length are 1 year, 7 years, and >=10 years.
2. Minimum defaulter rate exist for 9 years employee length.
3. Maximum defaulter rate for employee length >=10 years.
4. Grade is proportional to Loan Defaulter Probability.
5. We can see that , most of the loan default probability is seen for small_business,so bank should be extra careful while approving the loan for such businesses
6. Minimum defaulter rate showing for Major Purchase purpose.
7. We can see, as annual income is @[proportional to] probability of being defaulter. It is reaching up to 19%
8. We can see that as interest rate is increasing chance of being defaulter is also increased. when the interest rate touches more than 15% , risk of default rate is increasing
9. We can see that default rate is increasing , when the loan amount/funded amount is increasing at the alarming rate

# Thank You!