



# LENDING CLUB CASE STUDY

# Project Brief

Solving this assignment will give you an idea about how real business problems are solved using EDA. In this case study, apart from applying the techniques you have learnt in EDA, you will also develop a basic understanding of risk analytics in banking and financial services and understand how data is used to minimise the risk of losing money while lending to customers.



# Business Understanding

You work for a consumer finance company which specialises in lending various types of loans 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 below 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.

In this case study, you will use EDA to understand how consumer attributes and loan attributes influence the tendency of default.

# LOAN DATASET

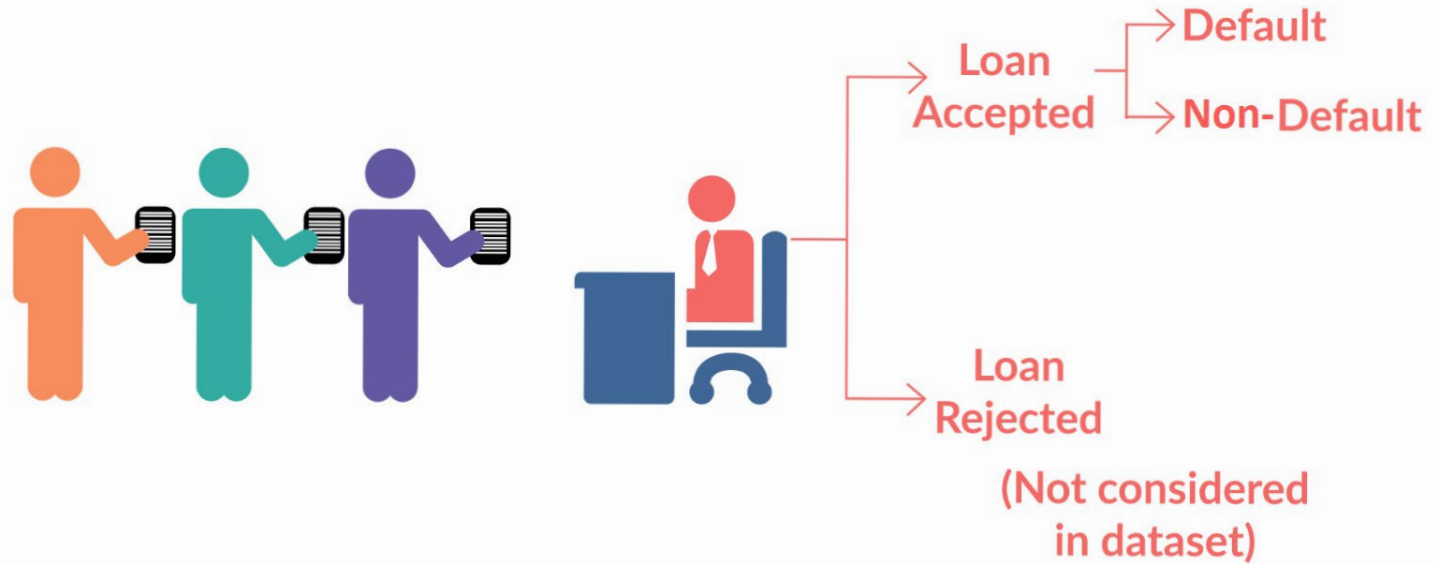


Figure 1. Loan Data Set

## Figure 1. Loan Data Set

When a person applies for a loan, there are two types of decisions that could be taken by the company:

**Loan accepted** : 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 instalments, i.e. the tenure of the loan is not yet completed. These candidates are not labelled as 'defaulted'.
- 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

**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)

# Business Objectives

This company 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.

Like most other lending companies, 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. In this case, the customers labelled as 'charged-off' are the 'defaulters'.

If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

In other words, the company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilise this knowledge for its portfolio and risk assessment.

---

# Data understanding

Understanding data in term of business understanding We categorize our data in three types of variables below and consider below variables as our basis of analysis.

## Customer's Demographic Information:

- emp\_title,
- emp\_length,
- home\_ownership,
- annual\_inc,
- verification\_status,
- addr\_state,
- zip\_code,
- title,
- purpose,
- desc,
- url,

## Loan Characteristics Information:

- loan\_amnt,
- funded\_amnt,
- funded\_amnt\_inv,
- int\_rate,
- loan\_status,
- grade,
- sub\_grade,
- dti,
- loan\_issue\_d,
- term,
- installment,



## Credit information (Customer Behaviour variables):

- 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,
- next\_pymnt\_d,
- last\_credit\_pull\_d,
- application\_type,

All other variables are not associated in identifying the default as they come in picture when the loan is approved. But we are focused if we want to approve loan in starting or not.

Removing title, desc and url and zip\_code columns from customer demographic as these will not be associated in identifying defaults.

**Business Objective:** To find the applicants which have strong probability of defaulting and also to identify the applicants which can repay their loan. To meet business objective we will only consider the Customer Demographic and Loan Attributes and we will ignore the Customer behaviour attributes as these will not be known during the time of application.

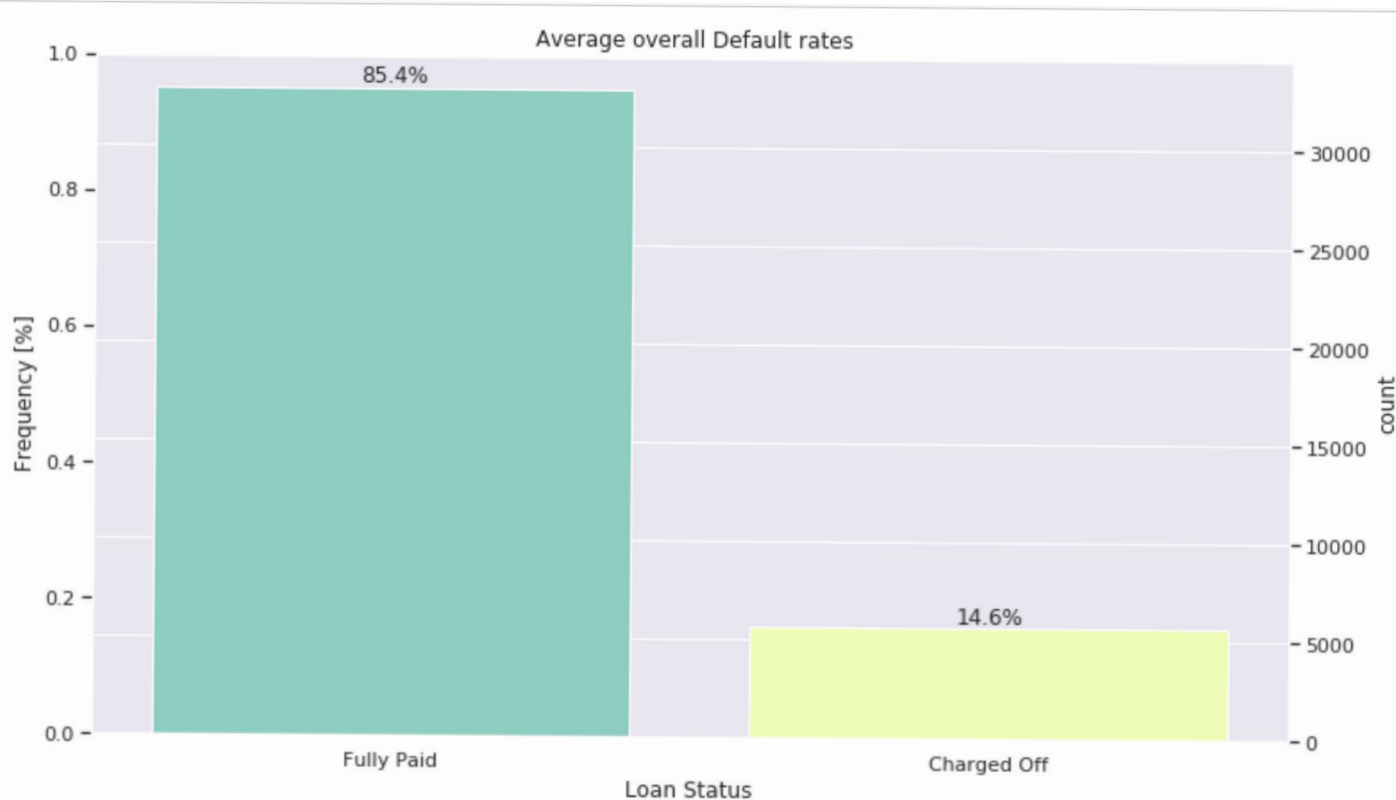
## Data Quality Issues:

- Blank data in some columns identified and removed on below steps.
- Term column of loan had string values along with integer value, dealt with it below for further analysis.
- Emp\_length (work experience) field has string values along with integer value for experience, created a new column named work\_experience in months and year for further analysis.
- Date format of some fields need to be converted to months , year etc for analysis
- loan\_status column contained some random dates along with it.



# Univariate Analysis

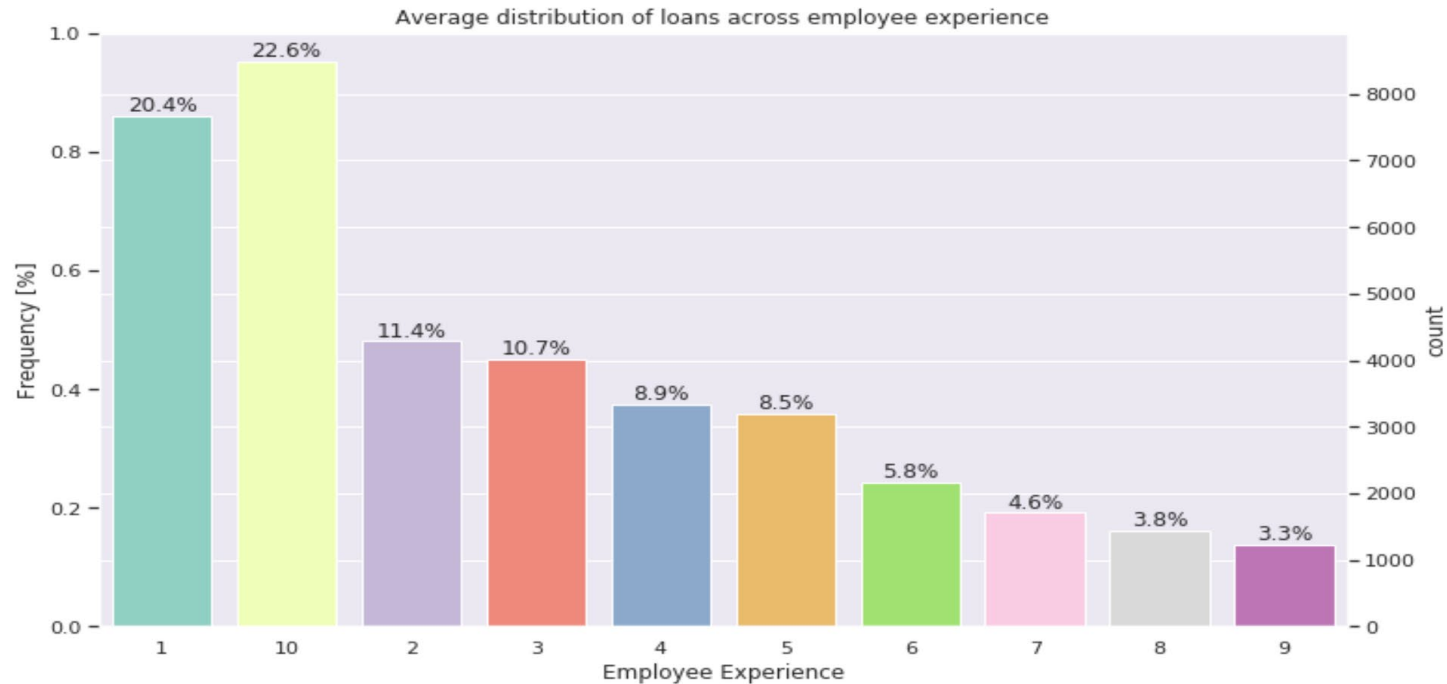
# 1) Loan Status Column Analysis



## **Conclusion:**

- **From above plots we can see that average default rate across all categories is 14.4% and non-default rate is 85.4%**

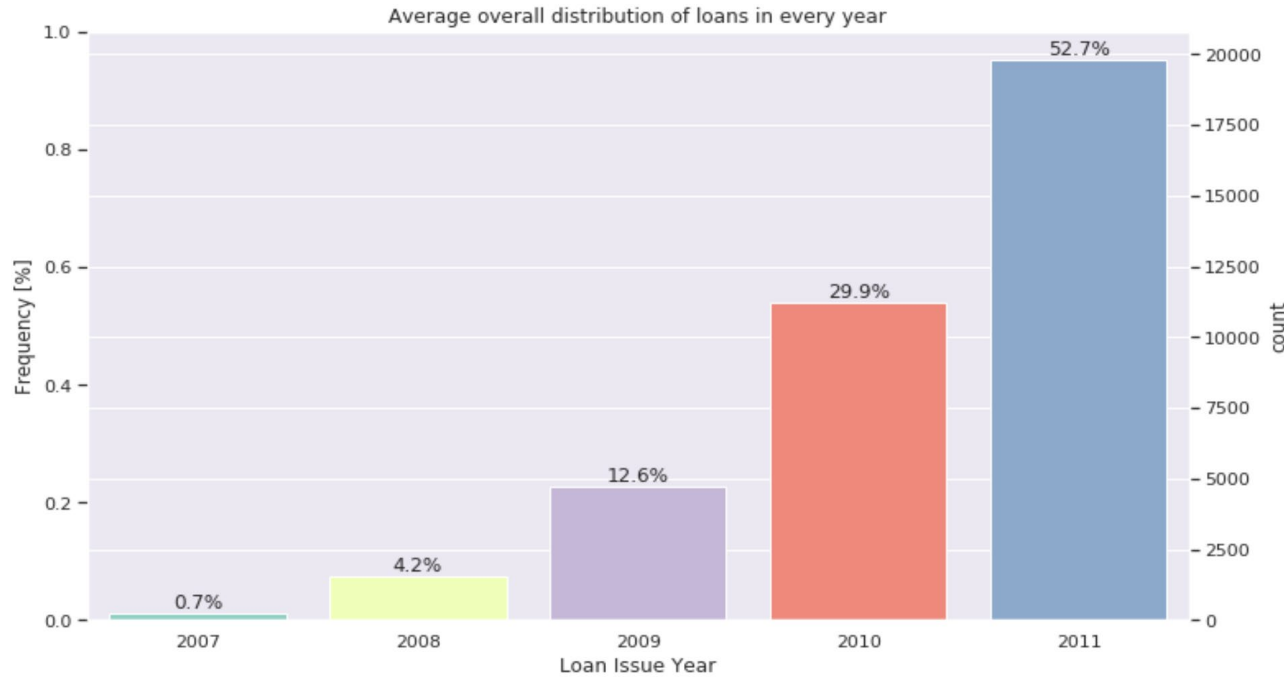
## 2) Employee Work Experience Analysis



## **Conclusion:**

- **People with experience 1 and 10 years are purchasing more loans.**

### 3) Loan Year Analysis

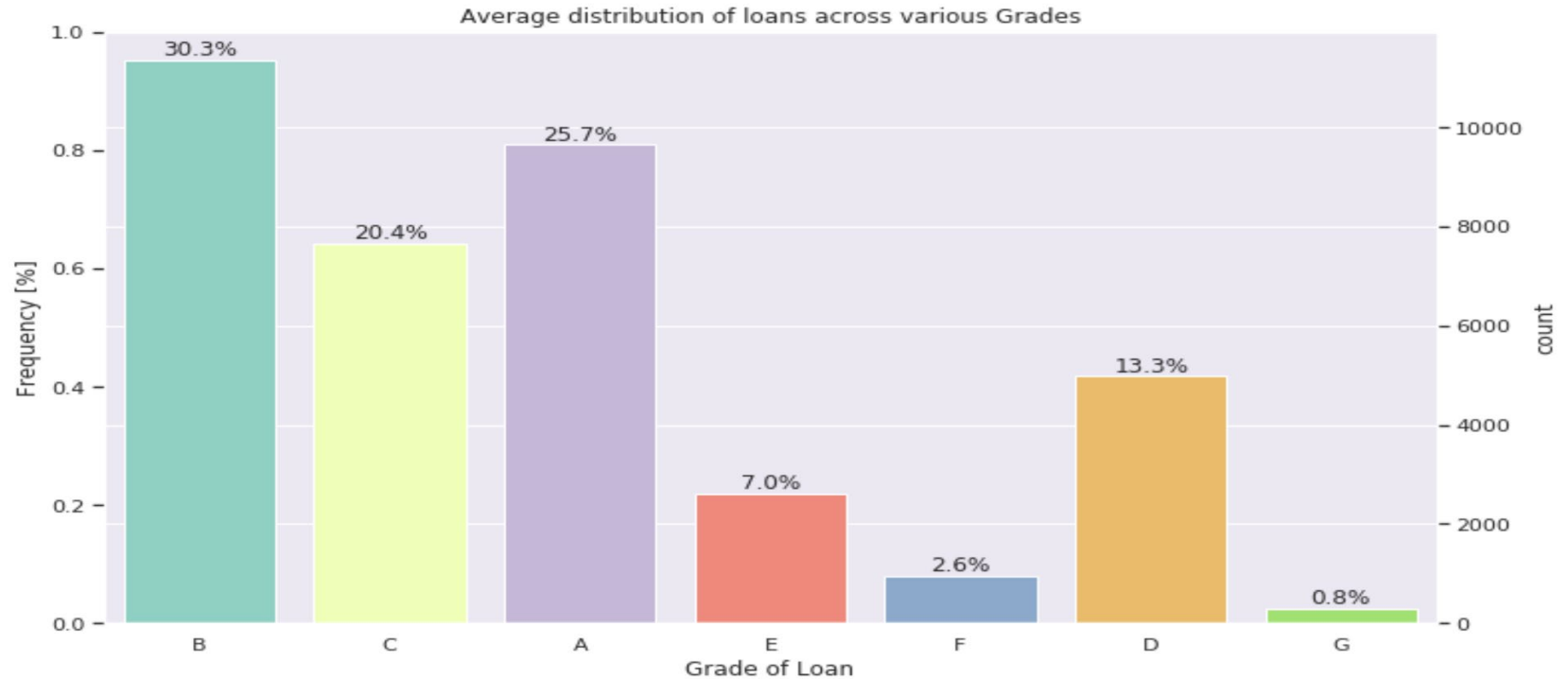




## **Conclusion:**

- **Evaluation from year 2007 to 2011 the loan purchase are increasing every year.**

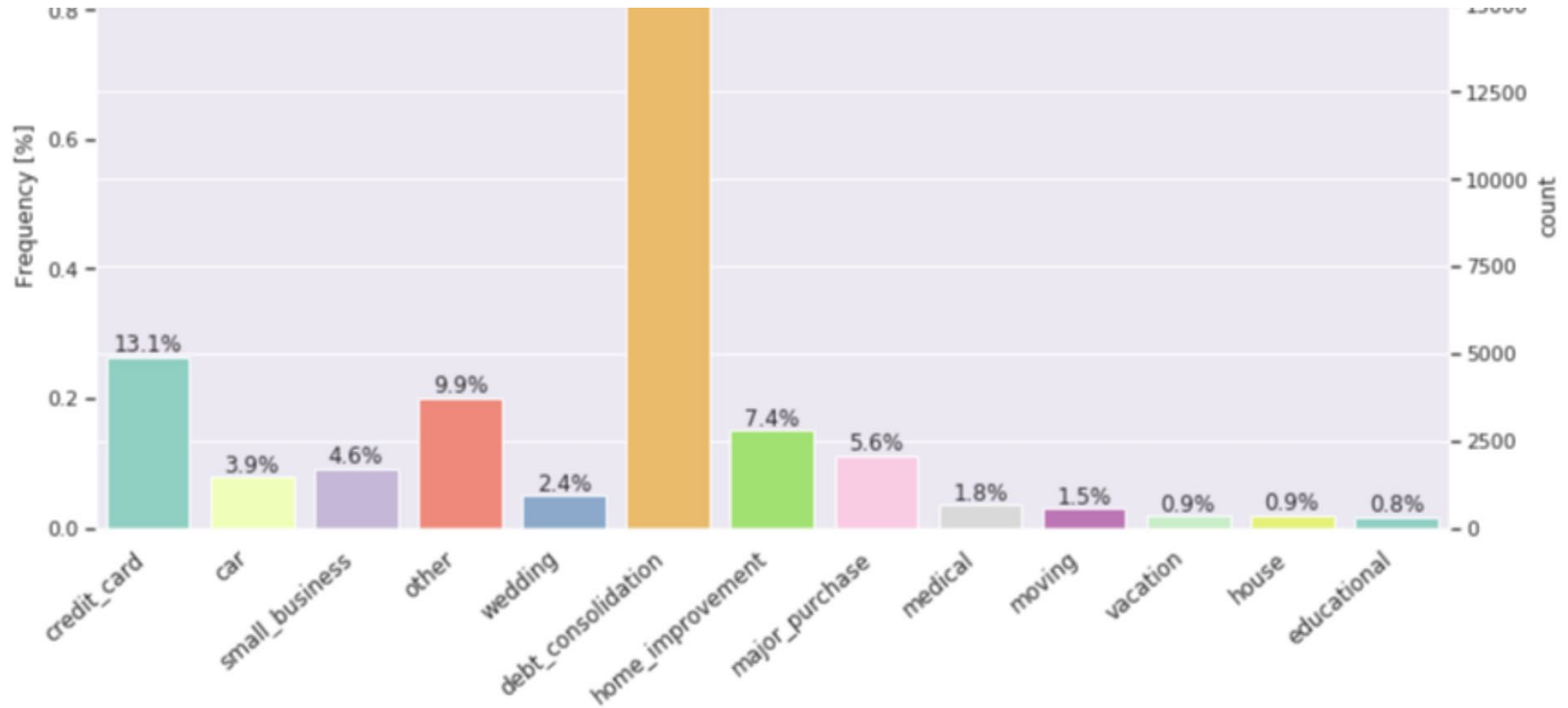
## 4) Grade Column Analysis



## **Conclusion:**

- **These are the percentage of loans across various grades.**
- **B type grade take the lead(30.3%), followed by the A type grades(26%).**

## 5) Purpose Column Analysis



## **Conclusion:**

- The following are the top 13 categories where maximum loan applications have recieved and hence high is there defaulting probability. ¶
- The highest probality to default is debt\_consolidation.

# Insights from above univariate plots

**There is a more probability of defaulting when :**

- From above plots we can see that average default rate across all categories is 14.4% and non-default rate is 85.4%.
- People with experience 1 and 10 years are purchasing more loans.
- B type grade take the lead(30.3%), followed by the A type grades(26%).
- The highest probability to default is debt\_consolidation.

# Univariate segmented Analysis on new derived variables

The background of the slide features a dark blue world map. Overlaid on the map are several glowing green and blue line charts. These charts represent data trends across different geographical regions. Numerous data points are marked along the lines, each accompanied by a numerical value. The overall aesthetic is high-tech and data-driven, with a network of glowing lines connecting various points across the globe.

**For better analysis we will limit our analysis on top6 loan categories.**

- debt\_consolidation
- credit\_card
- other
- home\_improvement
- major\_purchase
- small\_business

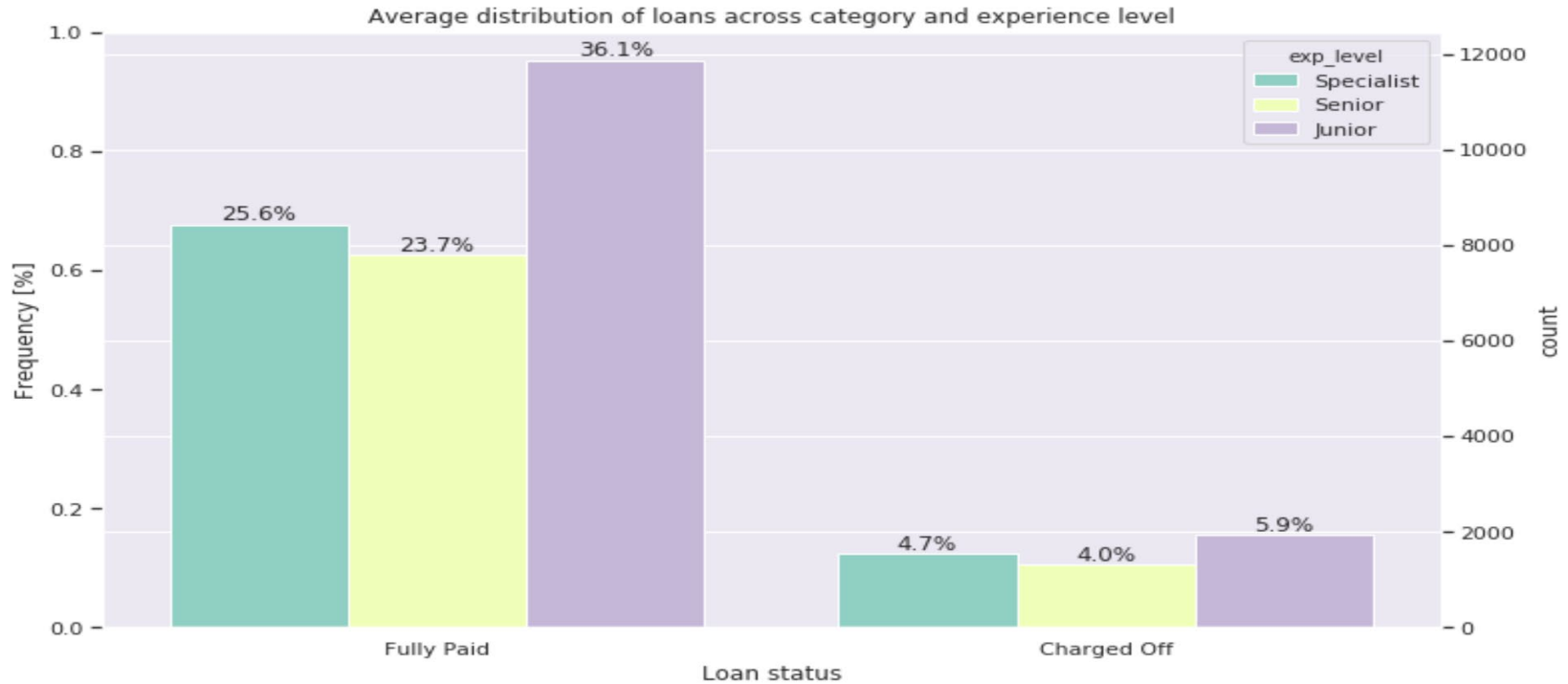


# Now we will filter the data for these categories and also we will derive new variables for our analysis

We will derive below variables for our analysis

- **dti range variable** - (Higher the dti ratio, lessen the chances of loan getting accepted)
  - $dti < 10$  (low),
  - $dti > 10$  and  $dt < 20$  (medium),
  - else 'high'
- **Loan-to-annual-income range variable** - (Higher the ratio , more chances of defaulting)
  - $l\_t\_ai < 0.1$  (low),
  - $l\_t\_ai > 0.1$  and  $< 0.2$  (medium),
  - $l\_t\_ai > 0.2$  (high)
- **Experience category** -
  - $exp < 3$  years (junior),
  - $exp$  between 3 - 7 years (senior),
  - $exp > 7$  years (specialist)

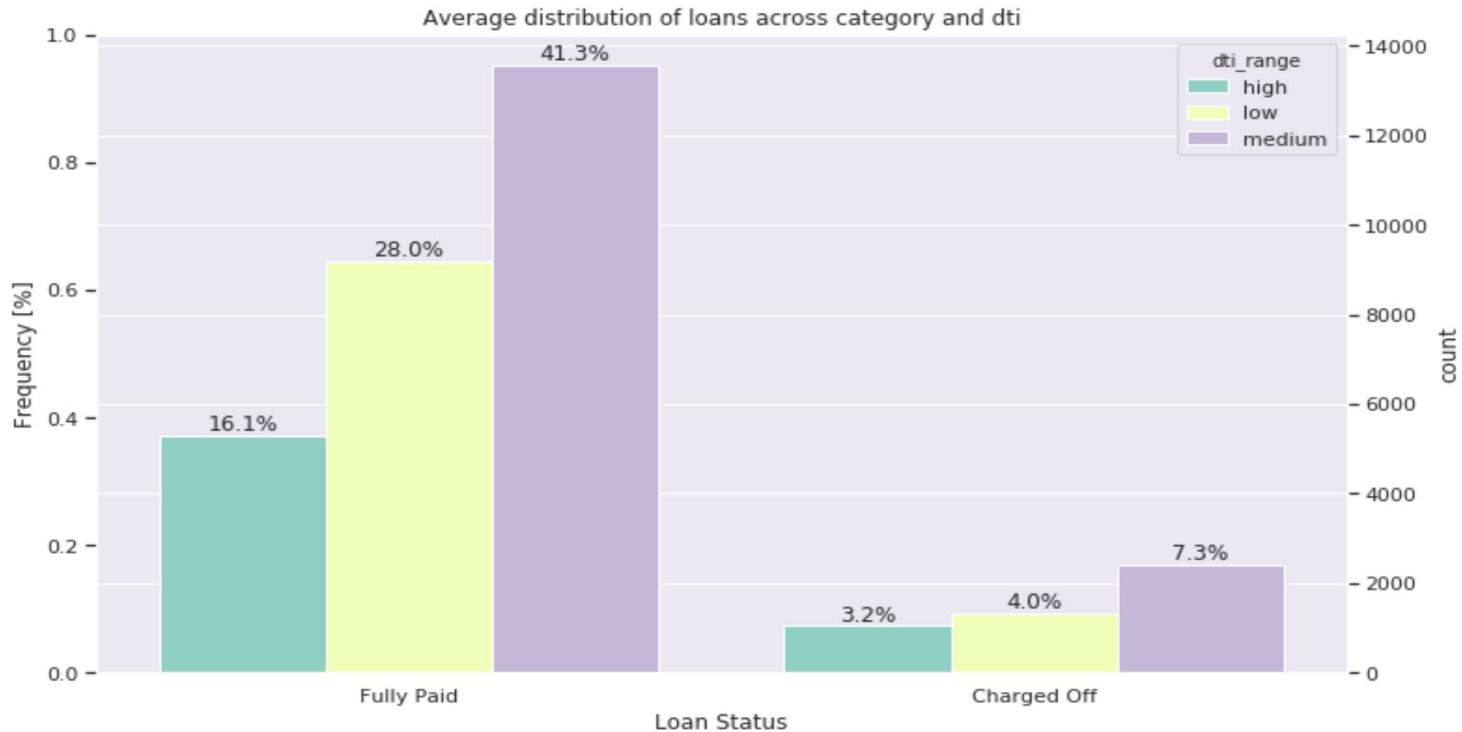
# 1.Experience level based on loan category



## **Conclusion :**

- **From the above graphs we can see that people with less experience have high chance of default.**

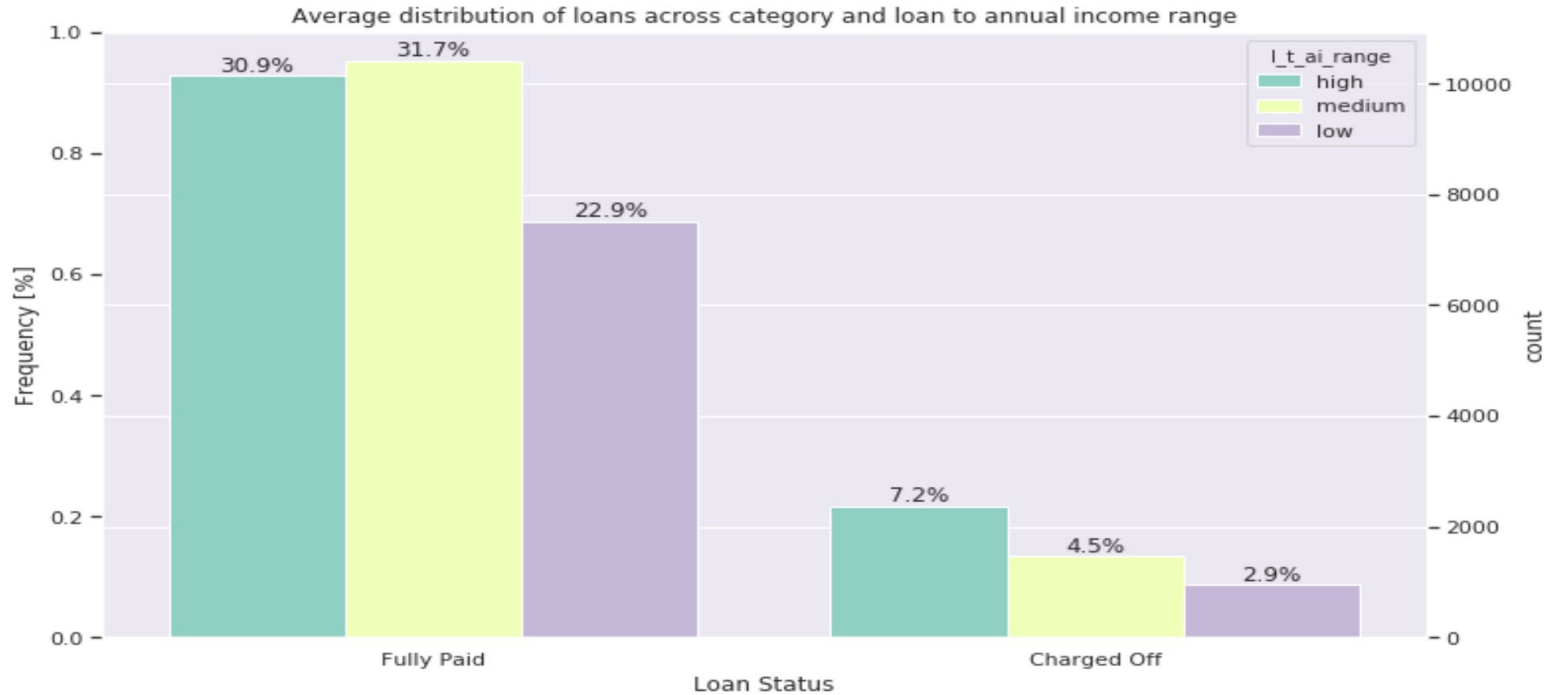
## 2. Debt-income-range based on loan category



## **Conclusion:**

- **People lying in medium dti range have high chances of default**

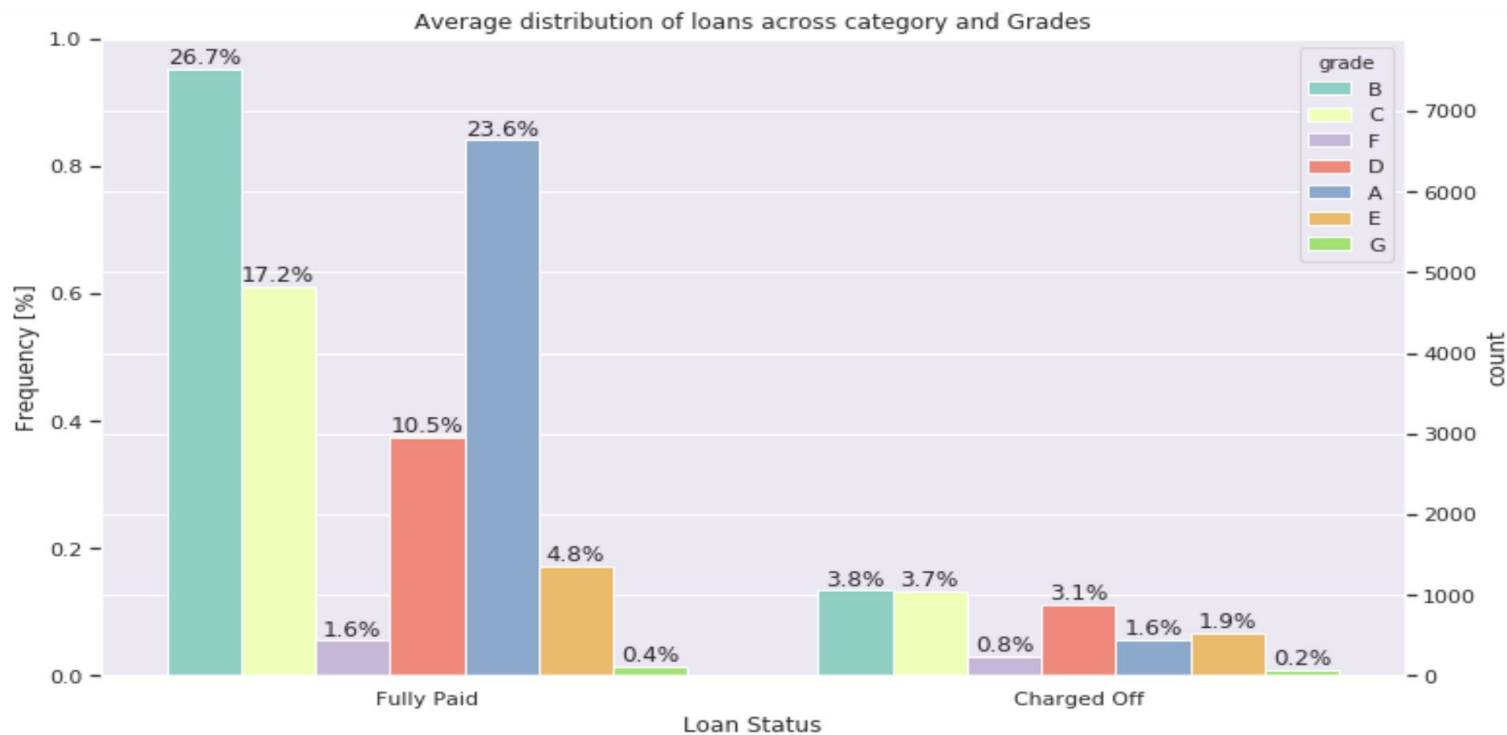
### 3.Loan-to-annual-icome-range based on loan category



## **Conclusion :**

- **People who have high loan to annual income ratio are at high risk of defaulting.**

## 4. Grade based on loan category





## **Conclusion :**

- **Grades B,C and D are at high probability of defaulting.**

# Insights from above graphs from Univariate segmented Analysis on derived variables

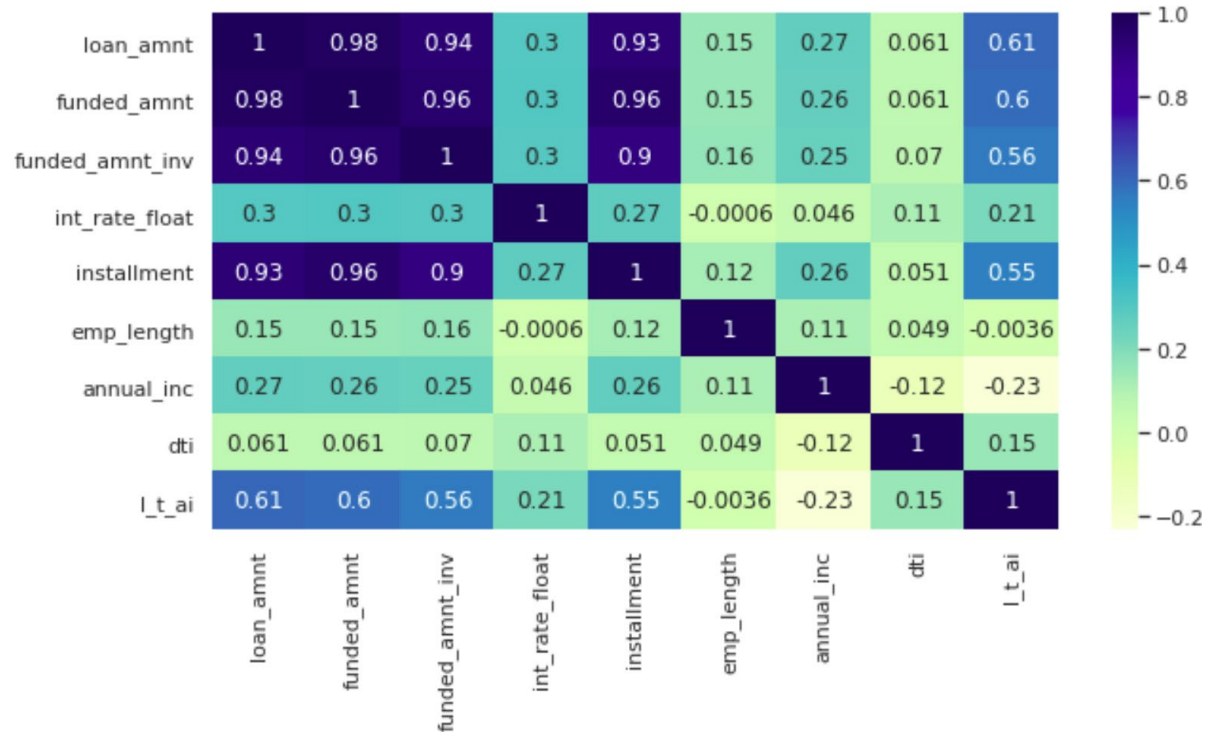
**There is a more probability of defaulting when :**

- From the above graphs we can see that people with less experience have high chance of default.
- People lying in medium dti range have high chances of default.
- People who have high loan to annual income ratio are at high risk of defaulting.
- Grades B,Cand D are at high probablity of defaulting.

# Bivariate Analysis




# Correlation Plot of numerical variables



## **Conclusions:**

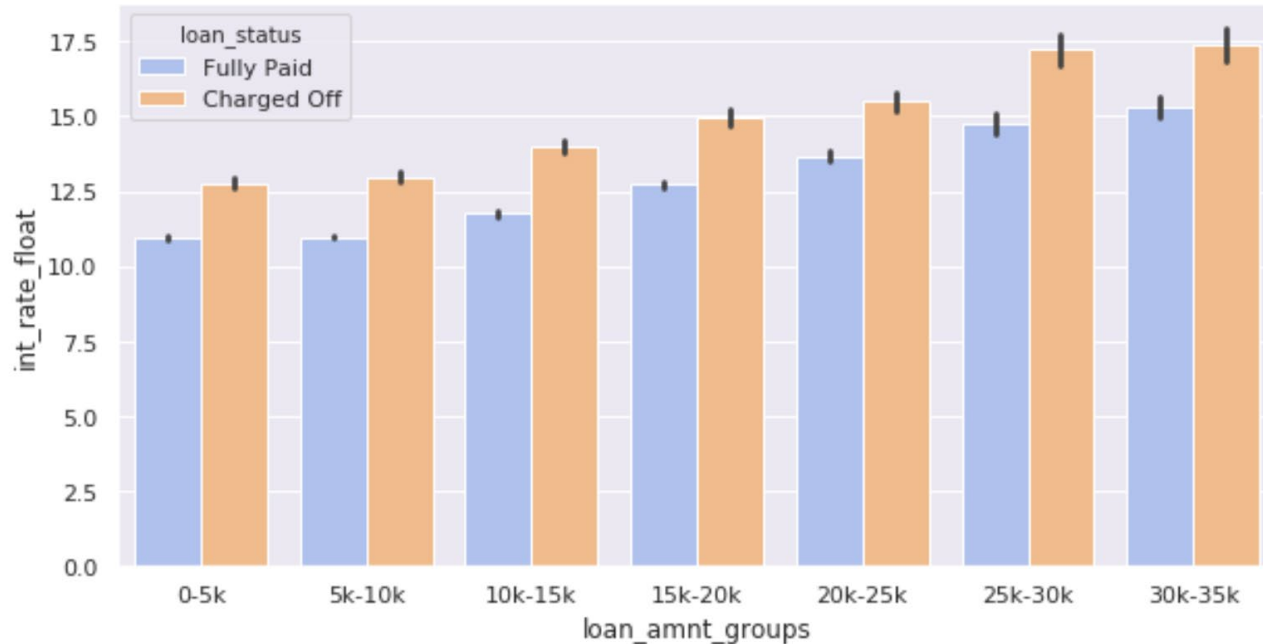
- **loan\_amnt and installment has a positive correlation**
- **loan\_amnt and int\_rate\_float has a moderate correlation**

The background is a complex, abstract network diagram. It features numerous blue and red nodes of varying sizes, connected by thin lines of the same colors. Some nodes are surrounded by concentric circles, suggesting a central or hub-like position. The overall composition is dense and interconnected, with lines crisscrossing the frame. The text is overlaid on a semi-transparent blue rectangular background.

Analysing loan amount with other columns for  
more insights

# 1.Loan Amount vs Interest Rate

loan\_amnt\_groups are created from loan amnt by grouping them based on the list ['0-5k','5k-10k','10k-15k','15k-20k','20k-25k','25k-30k','30k-35k']



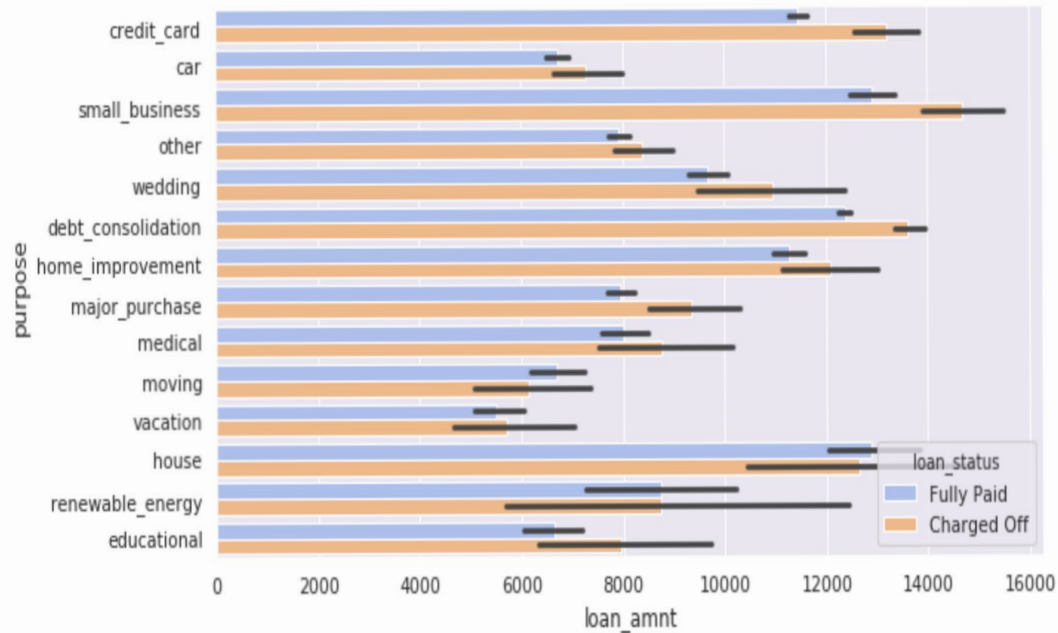
## **Conclusion:**

- **Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %, they have the highest probability to default.**



## 2.Loan vs Loan purpose

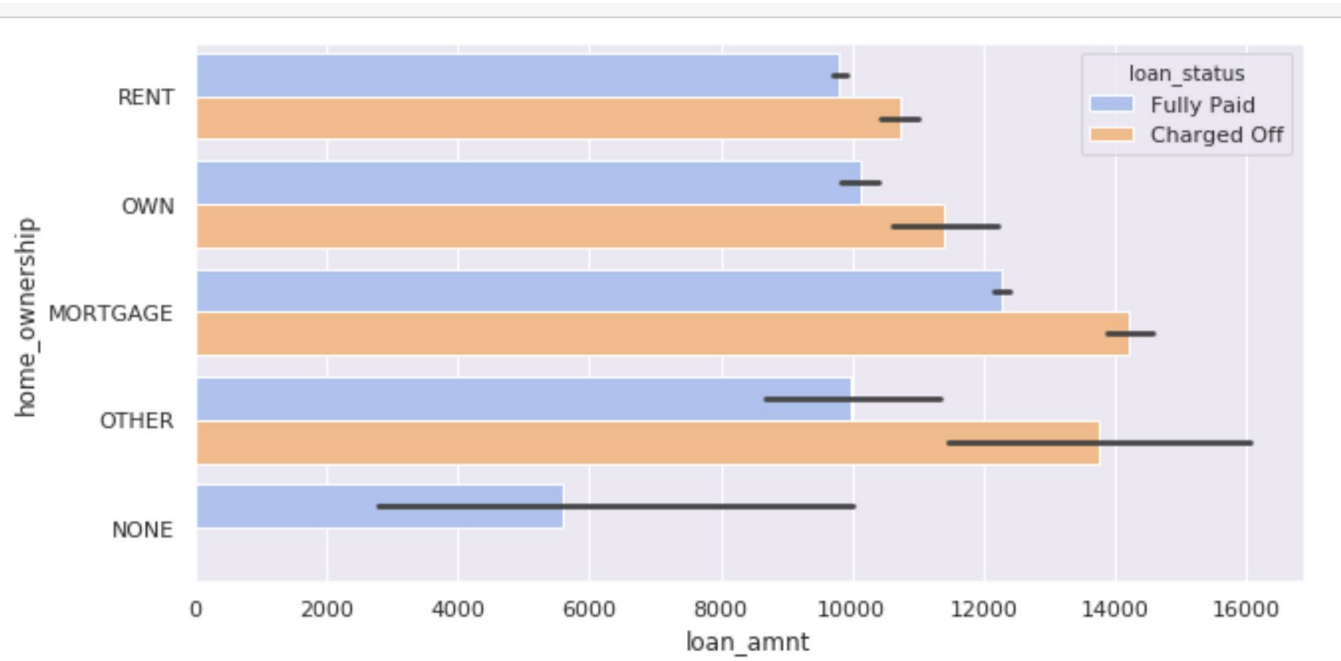
```
plt.show()
```



## **Conclusion:**

- **Applicants who have taken a loan for small business and the loan amount is greater than 14k have the highest chances to default.**

### 3.Loan vs House Ownership

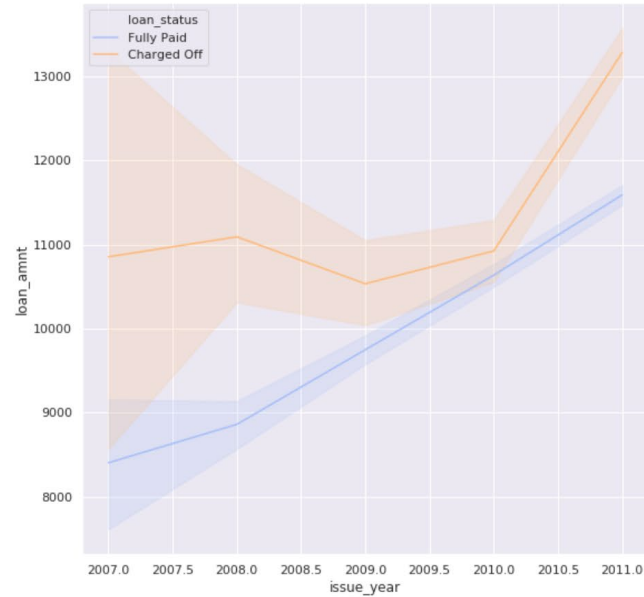
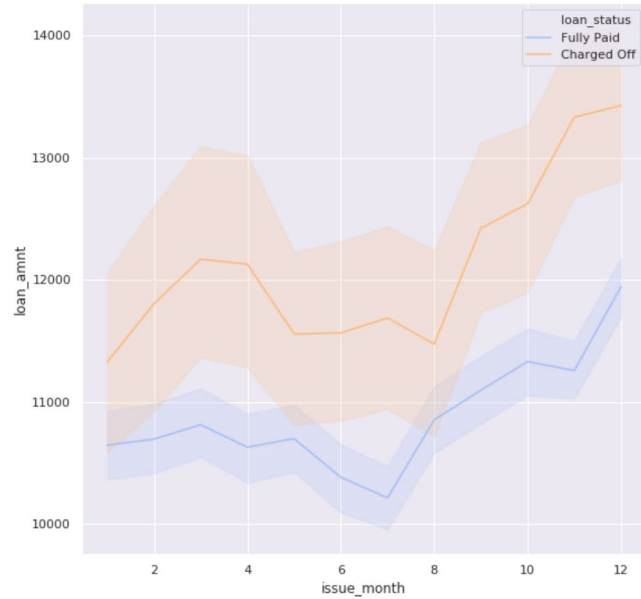


## **Conclusion:**

- Applicants whose home ownership is 'MORTGAGE' and have loan of 14-16k has the highest probability to default.

## 4.Loan amount vs month issued and year issued

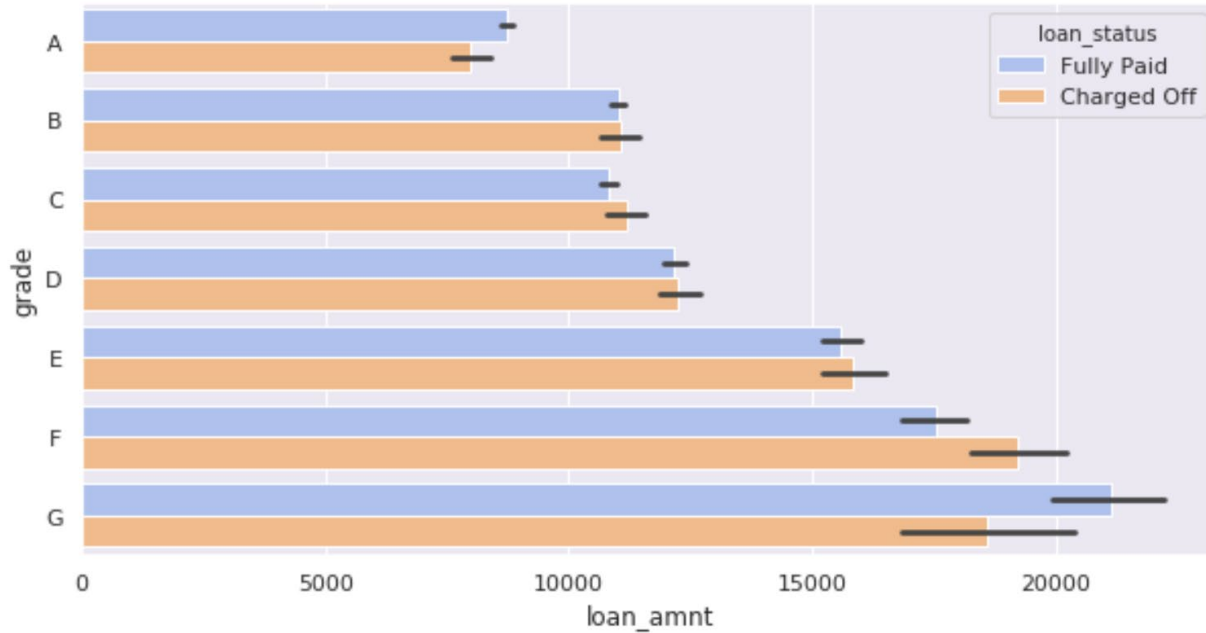
```
plt.show()
```



## **Conclusion:**

- **Maximum number of defaults occurred when the loan was sanctioned/issued in month December.**
- **Loan issued in the year 2011 also defaulted more as compared to other years.**

## 5.Loan amount vs Grade

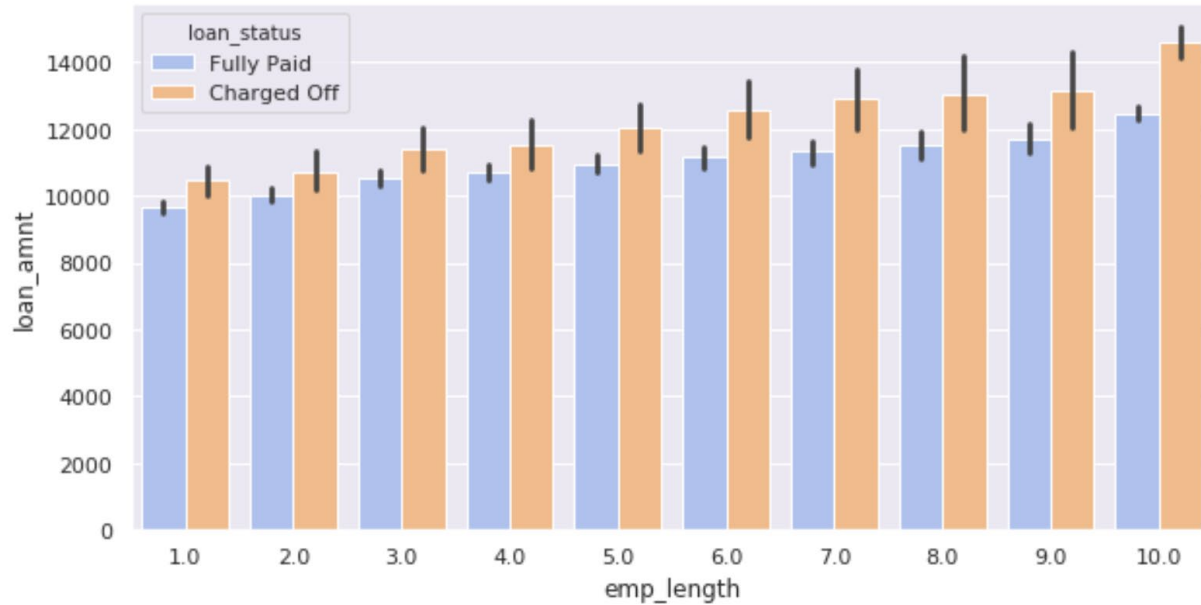


## **Conclusion:**

- **When grade is F and loan amount is between 15k-20k, there is a greater chance to default.**



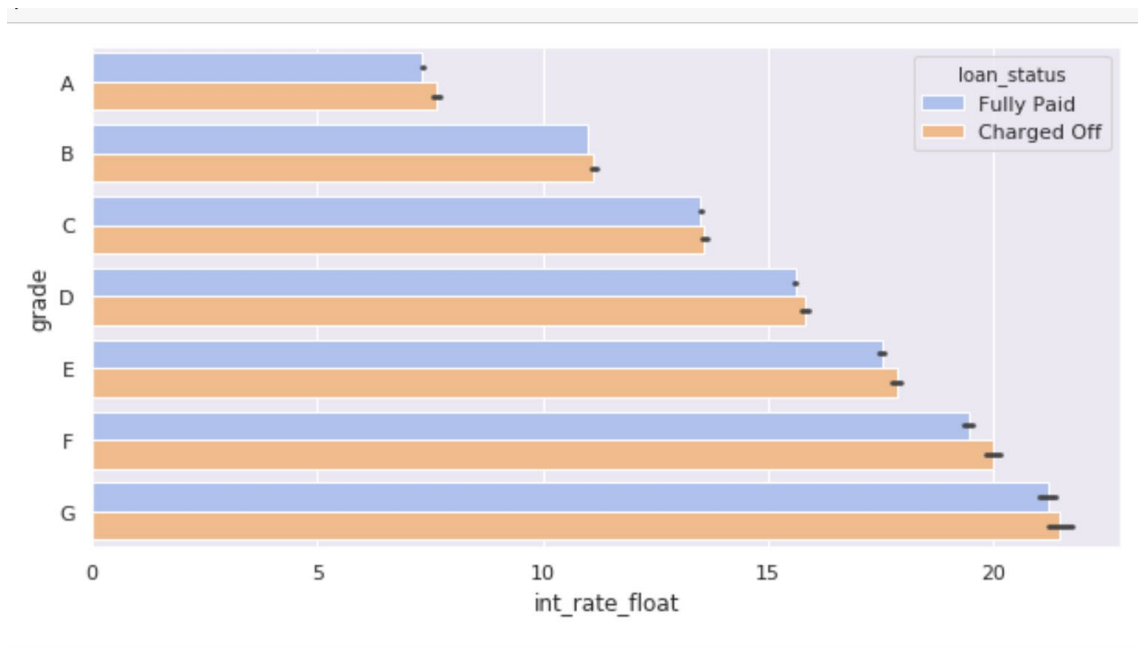
## 6.Loan amount vs employee length



## **Conclusion:**

- **Employees with longer working history got the loan approved for a higher amount.**
- **When employment length is 10yrs and loan amount is 12k-14k, there are higher chance to default.**

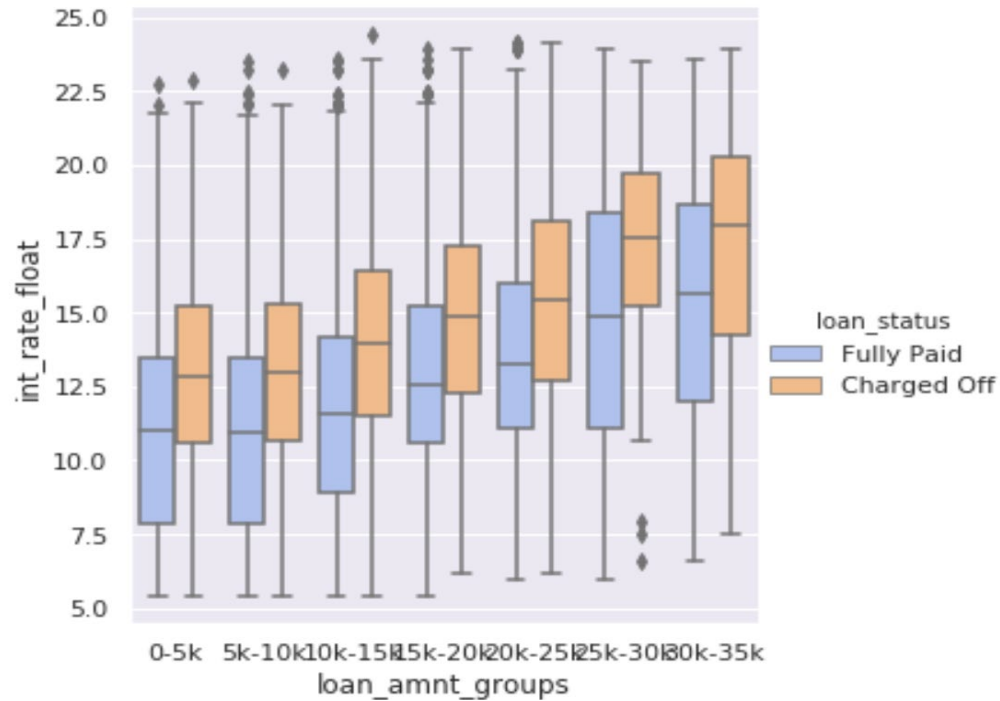
## 7. Grade vs interest rate



## **Conclusion:**

- **For grade G and interest rate above 20% there are high chances to default.**

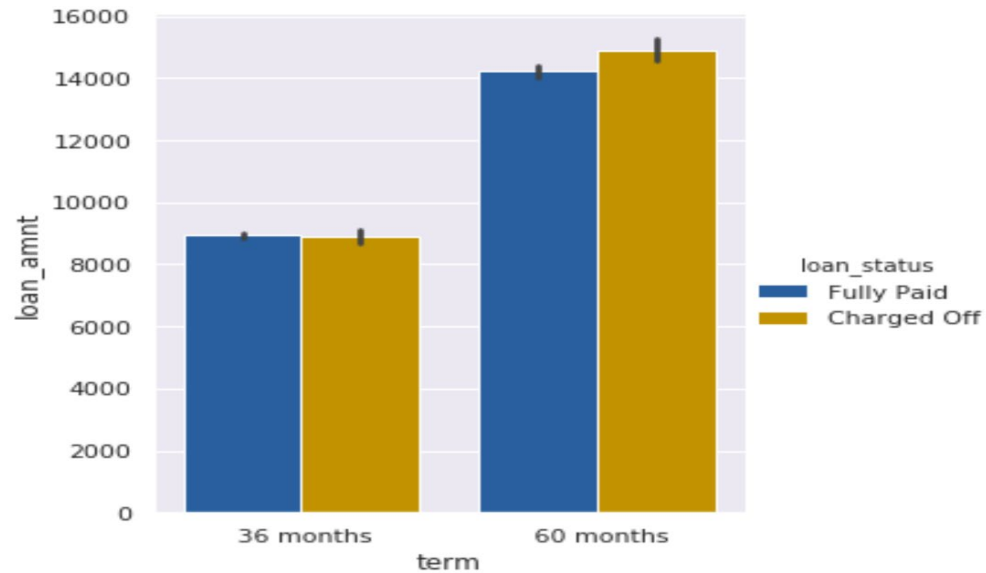
## 8. Interest rate and loan amount



## **Conclusion:**

- **The interest rate for charged off loans is pretty high than that of fully paid loans in all the loan\_amount groups.**
- **Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %**

## 8.Term and loan amount



## **Conclusion:**

- **Applicants who applied and defaulted have no significant difference in loan\_amounts**  
**Which means that applicants applying for long term has applied for more loan. ¶**

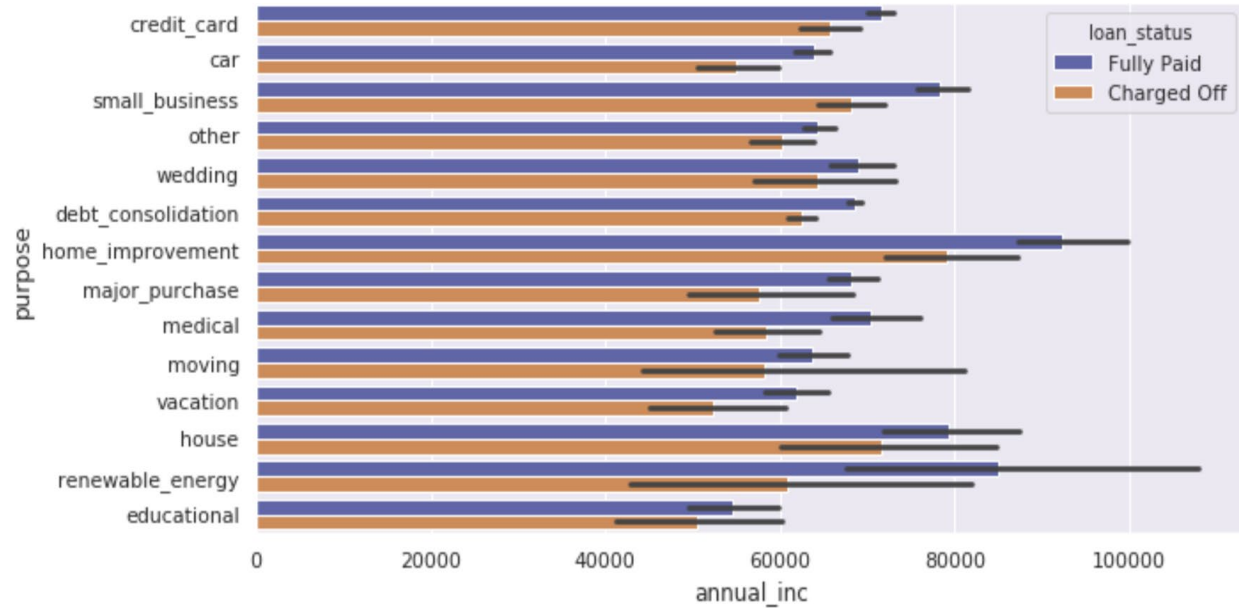




**Analysing annual income with other  
columns for more insights**

# 1. Annual income vs loan purpose

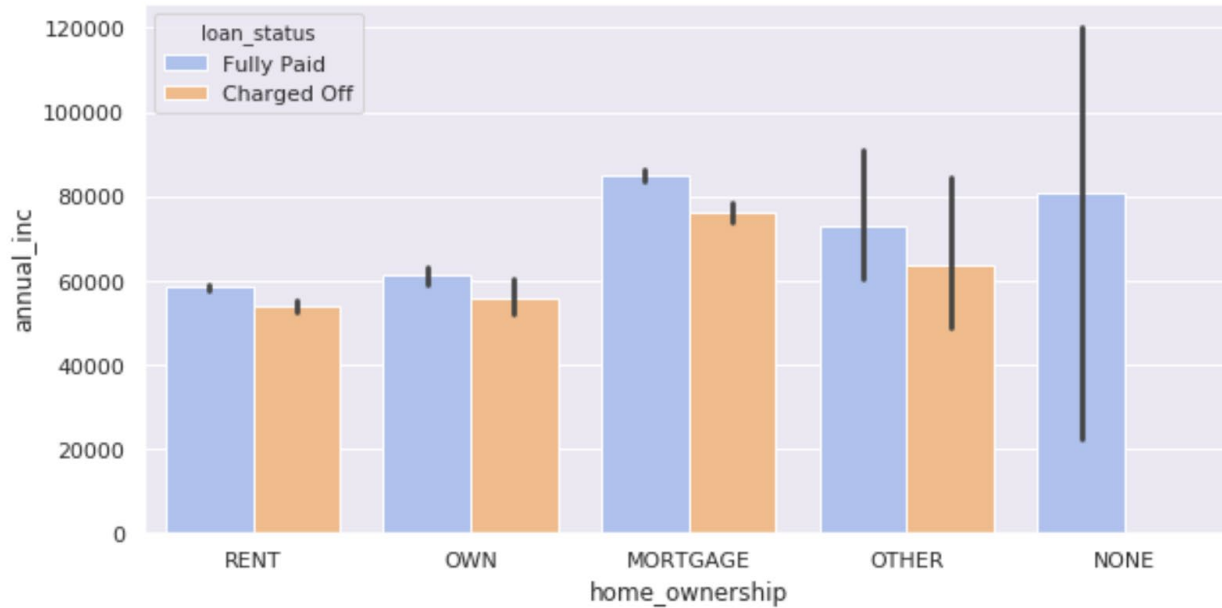
purpose (y)



## **Conclusion:**

- **Applicants taking loan for 'home improvement' and have income of 60k -70k have greater chance to default.**
- **Applicants with higher salary mostly applied loans for "home\_improvment", "house", "renewable\_energy" and "small\_businesses".**

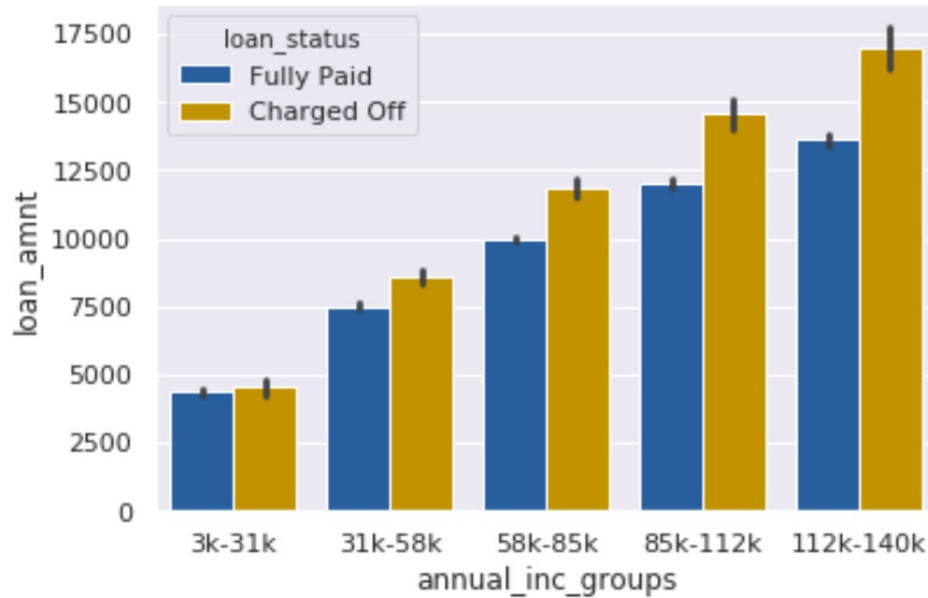
## 2. Annual income vs home ownership



## **Conclusion:**

- **Applicants whose home ownership is 'MORTGAGE and have income of 60-80k has greater chance to default.**

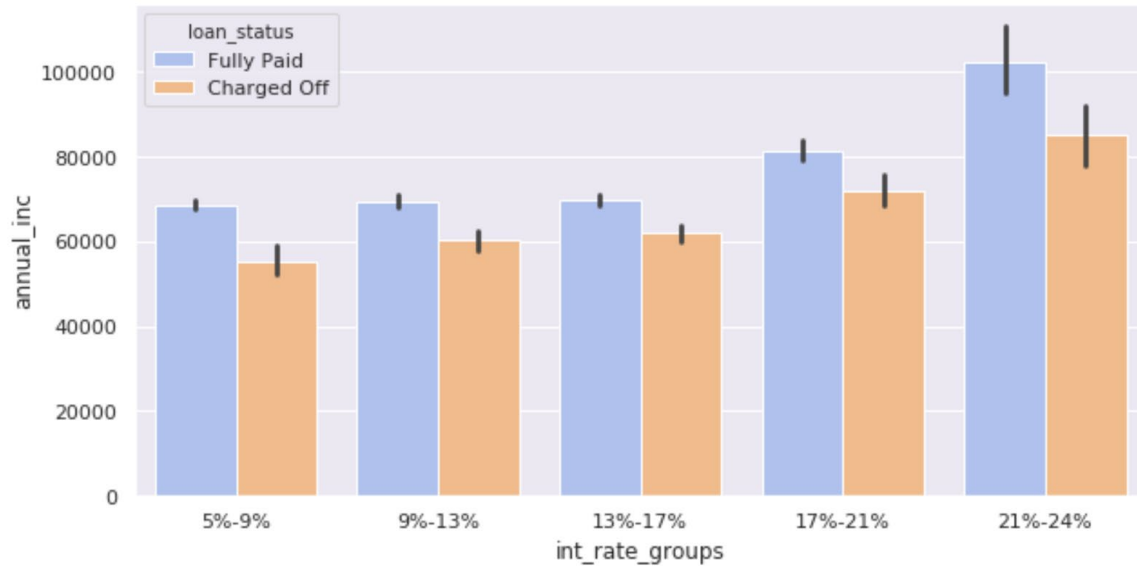
### 3. Annual Income vs Loan amount



## **Conclusion:**

- **Across all the income groups, the loan\_amount is higher for people who defaulted.**

### 3. Annual income vs int\_rate





## **Conclusion:**

- **Applicants who receive interest at the rate of 21-24% and have an income of 60k-70k, have high chance to default. ¶**

# Observations From Bivariate Analysis

**The above analysis with respect to the charged off loans. There is a more probability of defaulting when :**

- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %.
- Applicants who have taken a loan for small business and the loan amount is greater than 14k.
- Applicants whose home ownership is 'MORTGAGE' and have loan of 14-16k.
- Maximum number of defaults occurred when the loan was sanctioned/issued in month December.
- Loan issued in the year 2011 also defaulted more as compared to other years.
- When grade is F and loan amount is between 15k-20k.
- When employment length is 10yrs and loan amount is 12k-14k.
- For grade G and interest rate above 20%.
- This can be a pretty strong driving factor for loan defaulting.
- The interest rate for charged off loans is pretty high than that of fully paid loans in all the loan\_amount groups.
- Applicants who have taken a loan in the range 30k - 35k and are charged interest rate of 15-17.5 %.
- Applicants taking loan for 'home improvement' and have income of 60k -70k.
- Applicants whose home ownership is 'MORTGAGE' and have income of 60-80k.
- Across all the income groups, the loan\_amount is higher for people who defaulted.
- Applicants who receive interest at the rate of 21-24% and have an income of 60k-70k.



*It requires a very unusual mind to undertake the analysis of the obvious.*