Lending Club Case Study - EDA

- SRINATH TUMMALA

Introduction

Lending Club is one of 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.

The dataset contains complete loan data for all loans issued through the 2007-2011, including the current loan status (Current, Charged-off, Fully Paid, etc.) and latest payment information.

For companies like Lending Club correctly predicting whether or not a loan will be a default is very important. In this project, using the historical data from 2007 to 2011, we will use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default. We'll analyze the data and derive some insights about customers.

Data Understanding

- The Data Dictionary File contains description of each column in the dataset.

- The dataset consists of **111 columns** which explain various parameters of the loan provided to the customer. There are **39717 rows** which have unique customer IDs corresponding to each customer.

- The columns represent various attributes of the loan and the customer. Like, Annual income of the customer, Debt to income ratio of the customer, interest rate on the loan provided to the customer, etc.

Data Cleaning

- Out of the 111 columns in the dataset, many columns have completely null values. So, we have dropped columns which have completely null values.

- In the remaining columns, we still observe there are columns with many null values. So, We have dropped the columns with null values.

- Our goal is to find what kind of customer defaults a loan. So among the columns, we have dropped columns which do not contribute in analyzing the loan status.

- As we are checking for potential default customers, the data which corresponds to “current” loan status is not required. So, we have dropped rows with “current” as loan status.

Missing Values

- Finally we are left with 21 columns. Among these, employee length and revolving line utilization rate have missing values.

- In the emp_length column, we have categorical data and we have 10+ years value as the mode. As the missing value percentage is very less (2.8%) , we have imputed the emp_length column with the mode.

- In revol_util, the values are widely spread and the revol_util might be different for each customer, we did not impute the column. we have dropped the rows with empty revol_util value.

Data Standardization

- Columns like revol_util, int_rate have “%” symbol in the values. Since we require only numerical values for analysis, we have removed “%” symbol from these two columns.

- We have converted emp_length into numbers by assigning 0 to “<1 year” values and 10 to “10+ years” values and removed “years” from remaining values.

- The column “term” also has “months” along with numerical value. So, we have removed “months” from the term column to make it numerical value.

- We have issue date in “MON-YY” format. We have derived issue month and issue year from the issue date. These two new columns are type-driven derived variables.

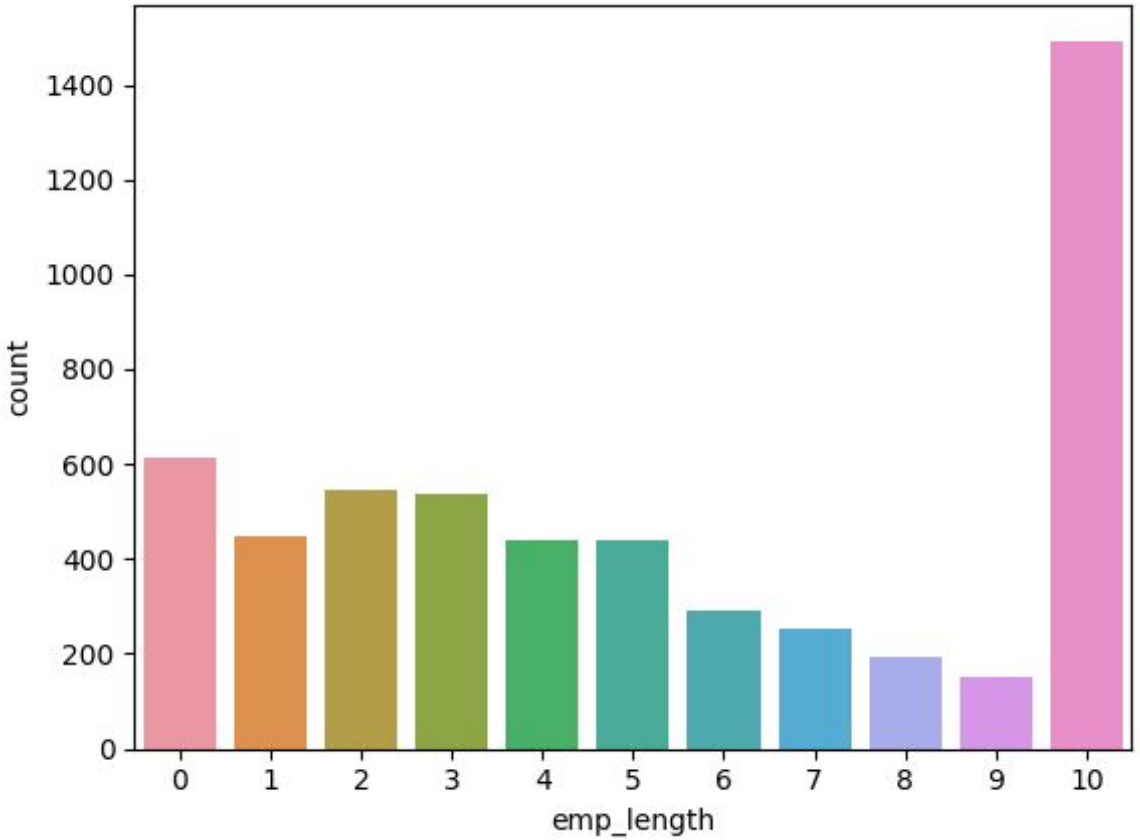
Outliers Check

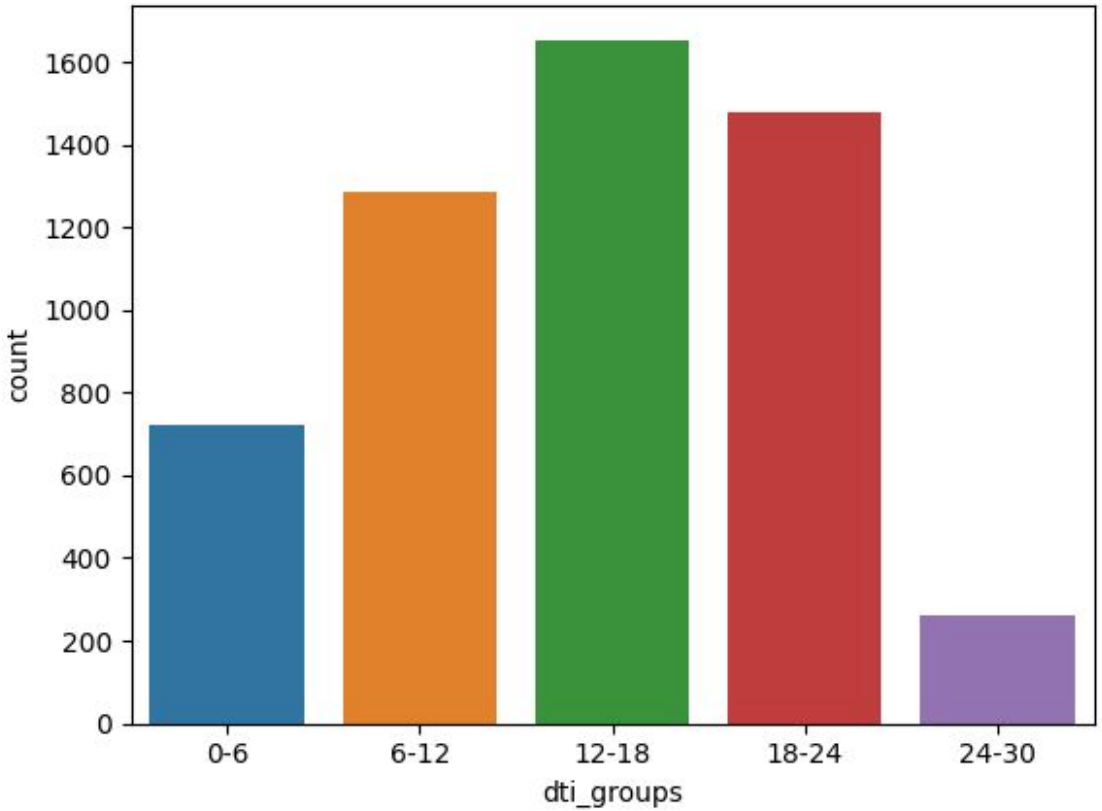
- We have checked the numerical columns “dti”, “annual_inc”, “loan_amnt”,”funded_amnt_inv” for outliers.
- Among these, the annual_inc column has outliers which are disconnected from the rest of the values.
- After checking the quantile information, we have found out that the values after 95th quantile are spread far away from the distribution. So, with trehold as 95%, we have removed the outliers.
- Finally, we are left with 36606 rows and 23 columns.

Univariate and Segmented Univariate Analysis

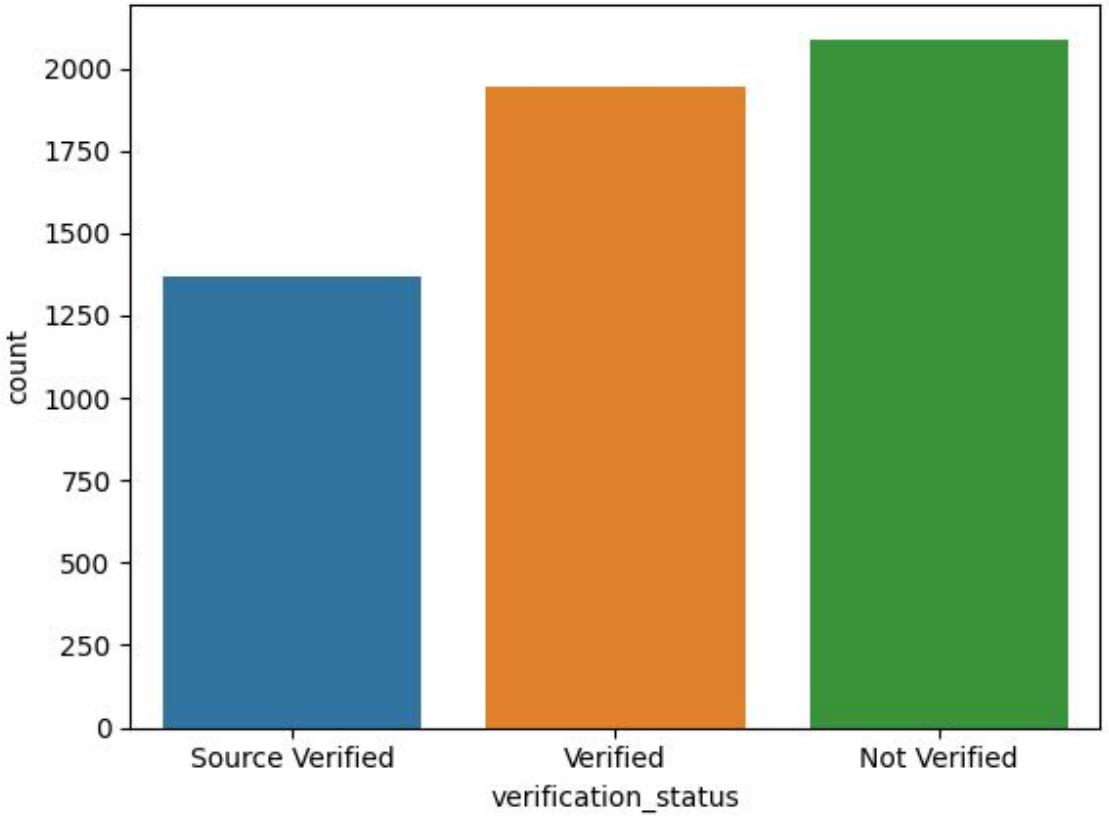
Top Driver variables or factors behind loan default

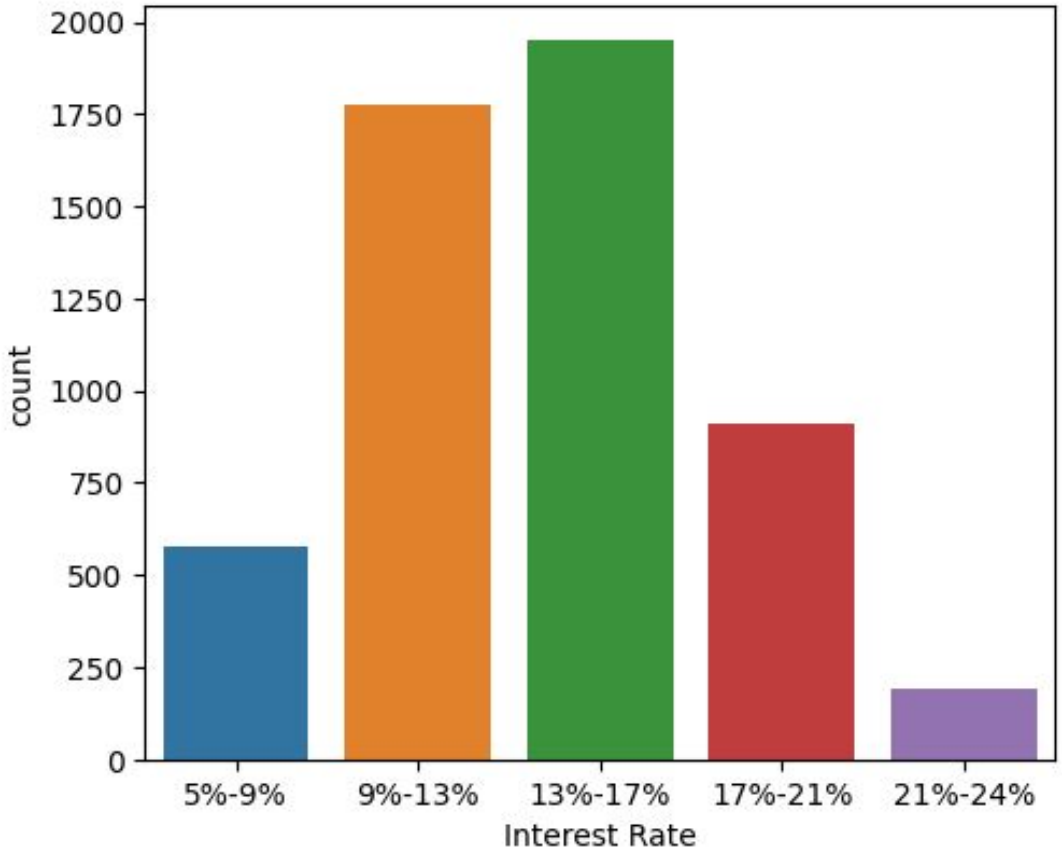
- If an applicant has **10 or more years** of work experience, the applicant is most likely to default the loan.
- If the **Debt to Income ratio** of an applicant in the range of **12%-18%**, the applicant is most likely to default the loan.
- If the applicant's source is **not verified**, the applicant is most likely to default the loan.
- If the **interest rate** on the loan is in the range of **13%-17%**, the applicant is most likely to default the loan.
- If the loan amount is in the range of **5k-10k**, the applicant is most likely to default the loan.
- If the LC assigned loan **grade** is **B**, then the applicant is most likely to default the loan.
- If the applicant applies the loan to **consolidate their debts**, the applicant is most likely to default the loan.



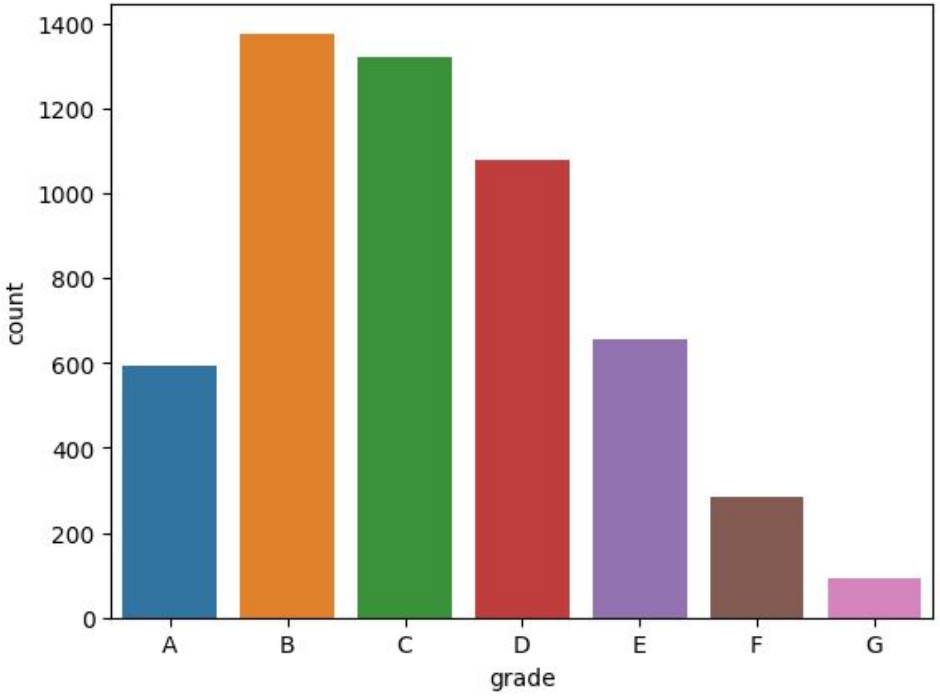


These plots represent Number of charged off loans with respect to employee work history and Debt to income ratio of applicant.

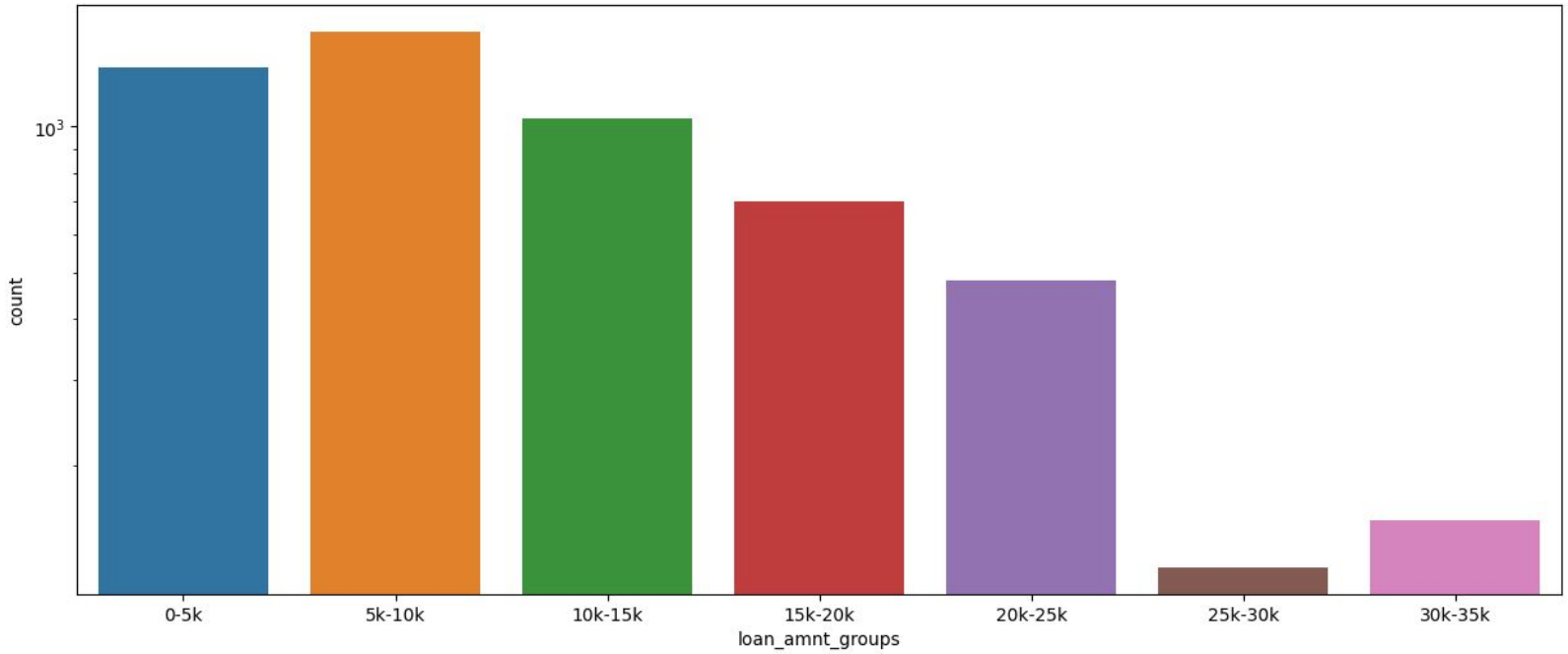


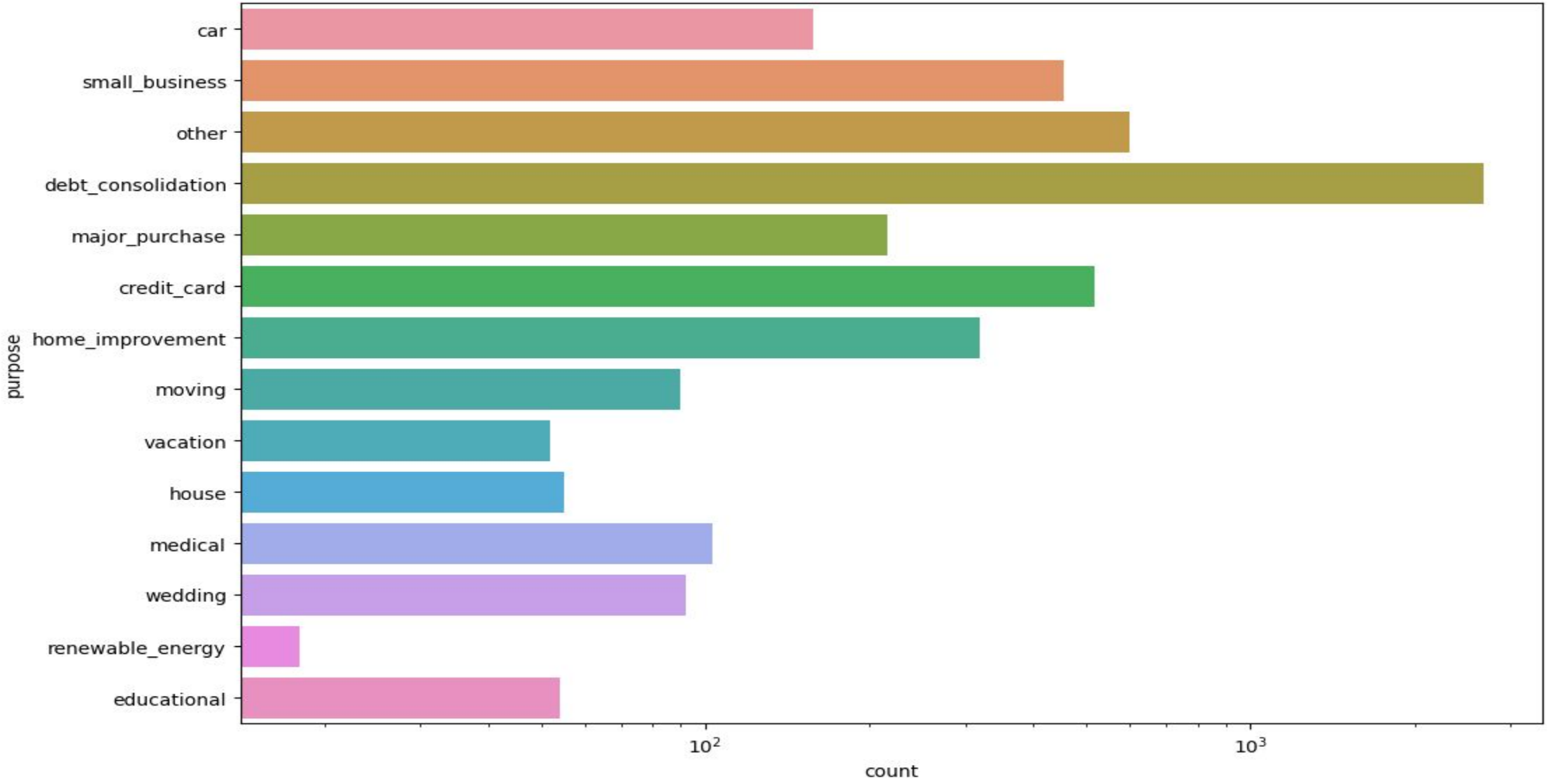


These plots represent number of charged off loans with respect to source verification status of the applicant and Interest rate on the loan.



These plots illustrate count of number of charged off loans across LC grade of the loan and the loan amount.





This plot illustrates the number of charged off loans across the purpose of loan mentioned by the customer

Bivariate Analysis

Important Combinations of Driver Variables

- Through correlation analysis among numerical variables, we found out that (loan amount and funded amount by investors) , (loan amount and installments) , (fundended amount by investors and installments) are **highly correlated** with each other.

- If the interest rate is in the range **14-16%** and if the purpose of the loan is to buy a **home**, then it is possible that the applicant might default the loan.

- If the loan amount is in the range of **30k-35k** and if the interest rate is in the range of **16%-18%**, there is a possibility that the applicant might default the loan.

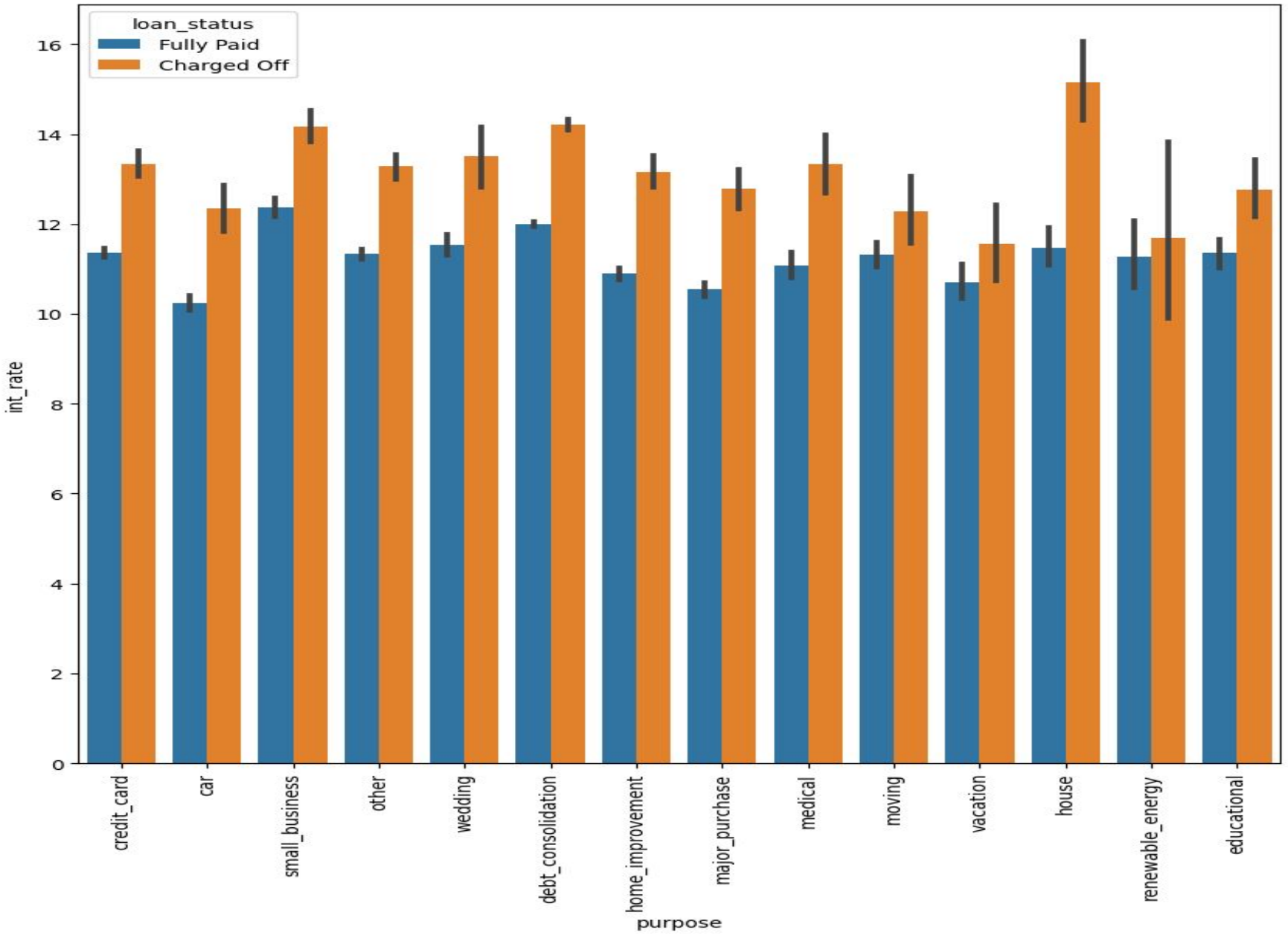
- An applicant might default the loan if the purpose of the loan is for a **small business** and the loan amount is in the range of **12k-14k**.

- There is a possibility that an applicant will default the loan if the LC assigned loan **grade** is "**F**" and the loan amount is in the range **17k-20k**.

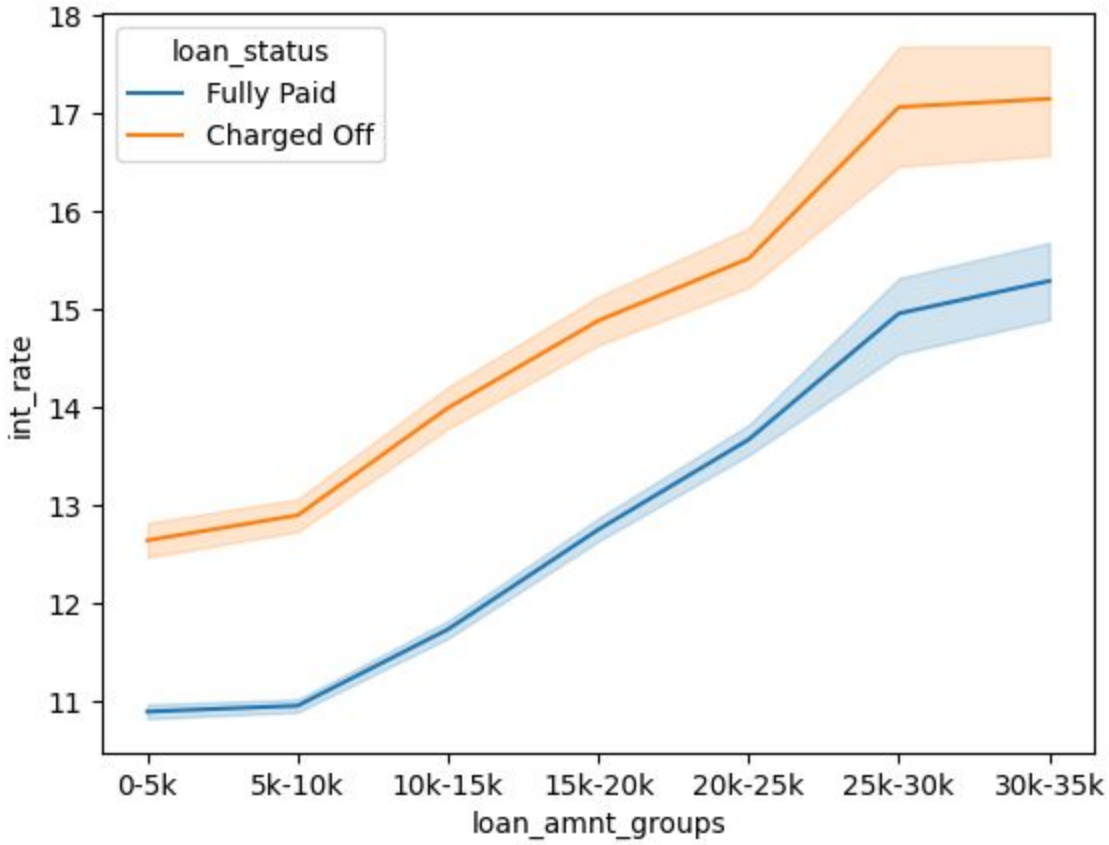
- If the applicant has **10** or more years of **work** experience and the **loan amount** is in the range **12k-14k**, the applicant is most likely to default the loan.

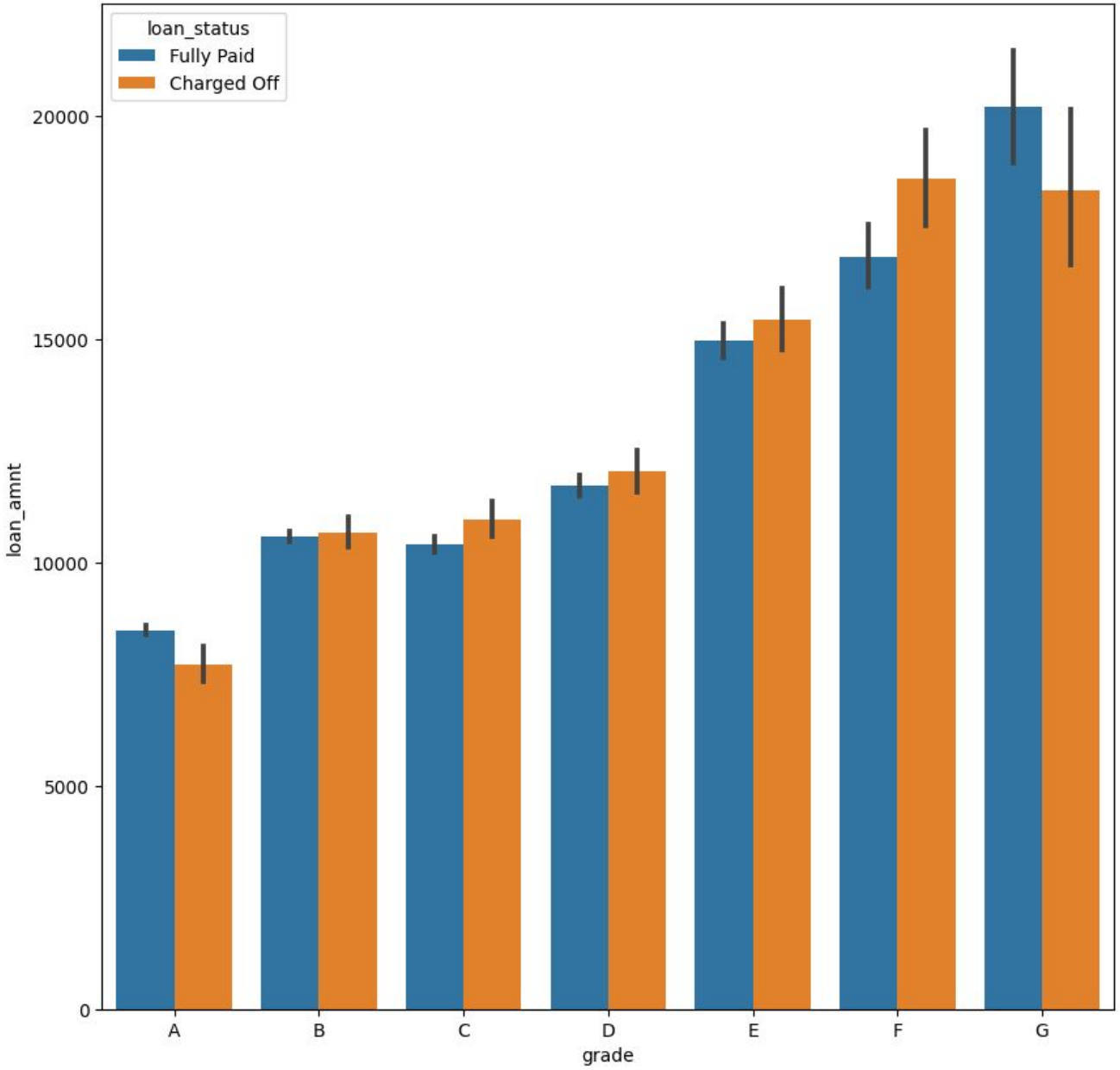
- If the applicant's source is **verified** and the loan amount is above **16k**, the applicant is most likely to default the loan.

- If the Debt to Income ratio of the applicant is in the range of **18-24** and the interest rate is **above 14%**, the applicant is most likely to default the loan.

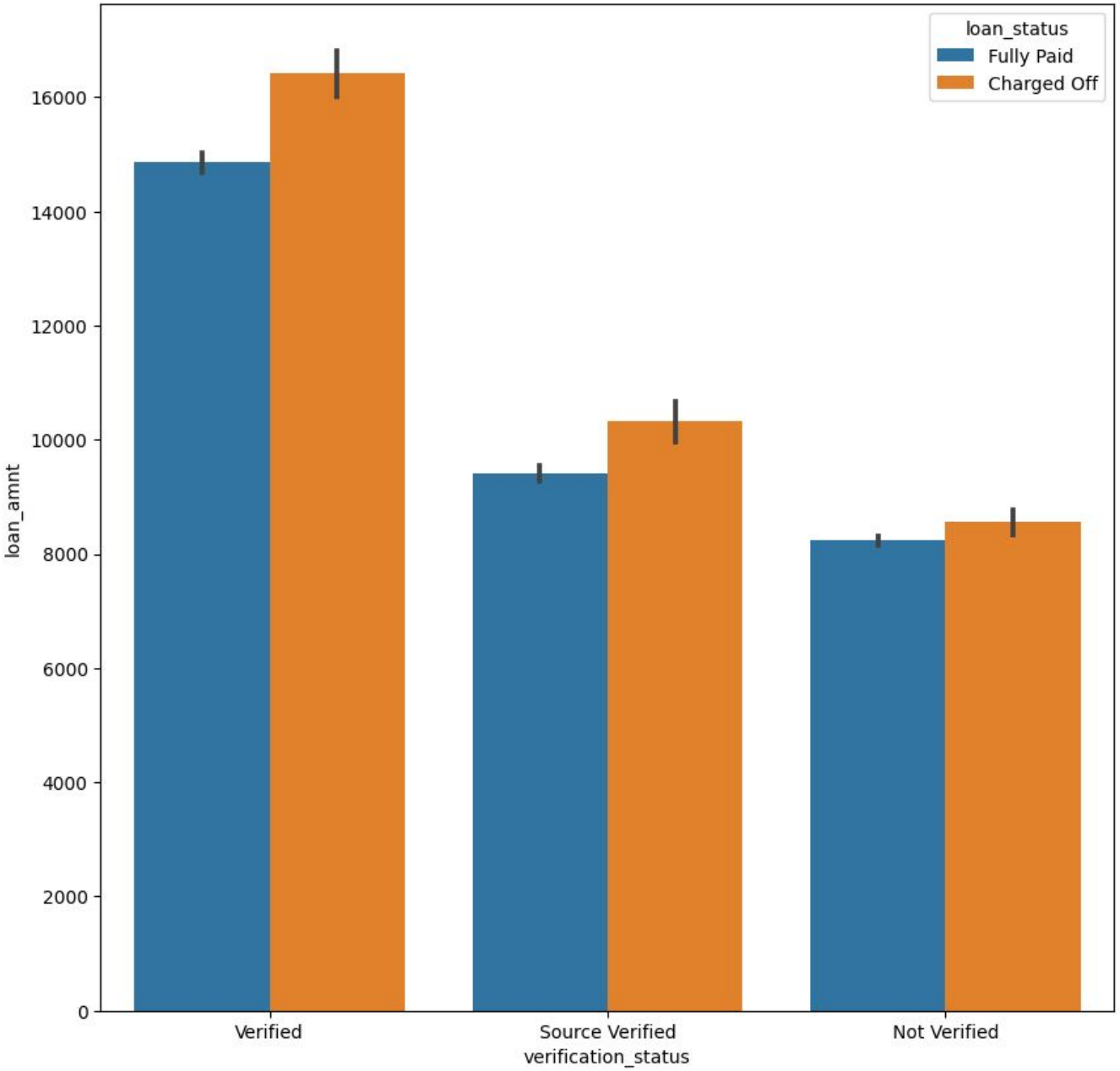


This plot shows the effect of purpose of loan and interest rate on the loan status

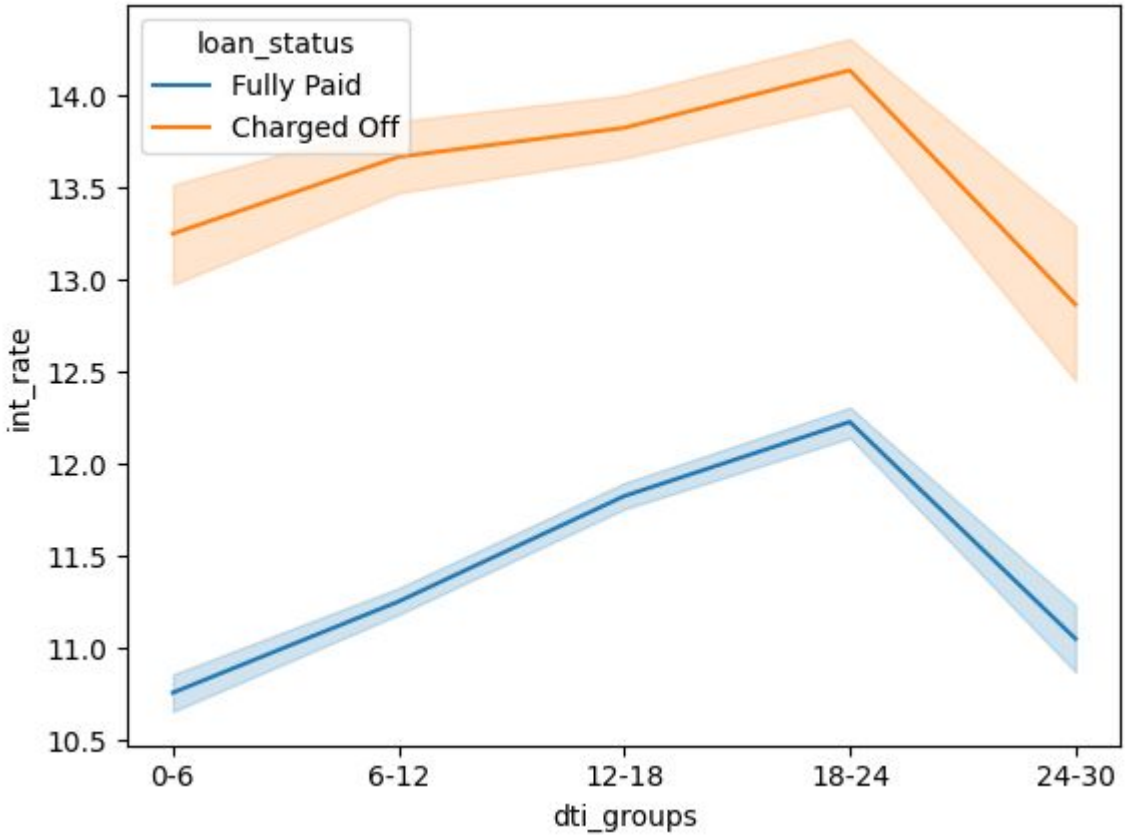


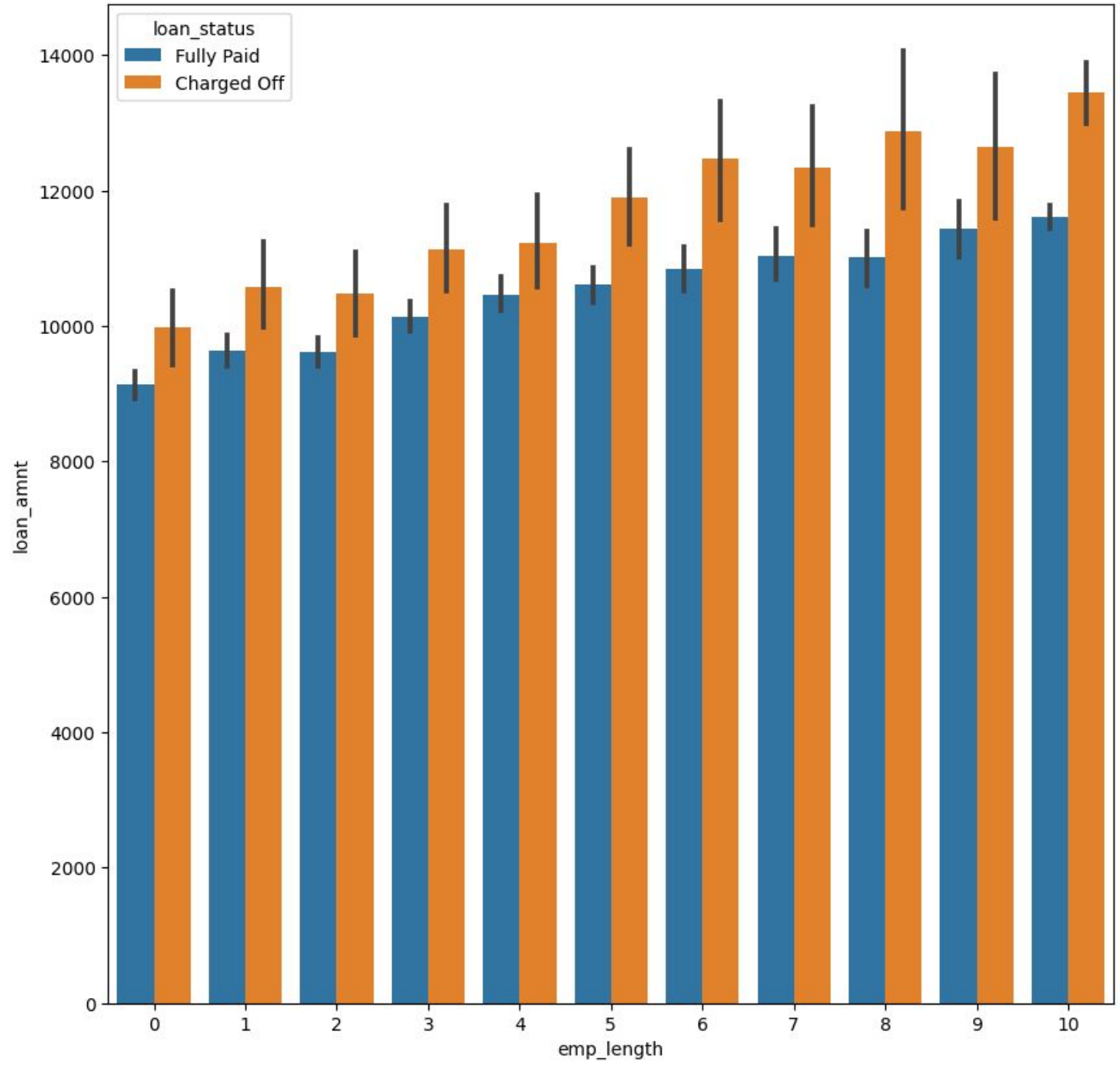


These plots show how charged off loan status is affected by (interest rate, loan amount) and (grade, loan amount).



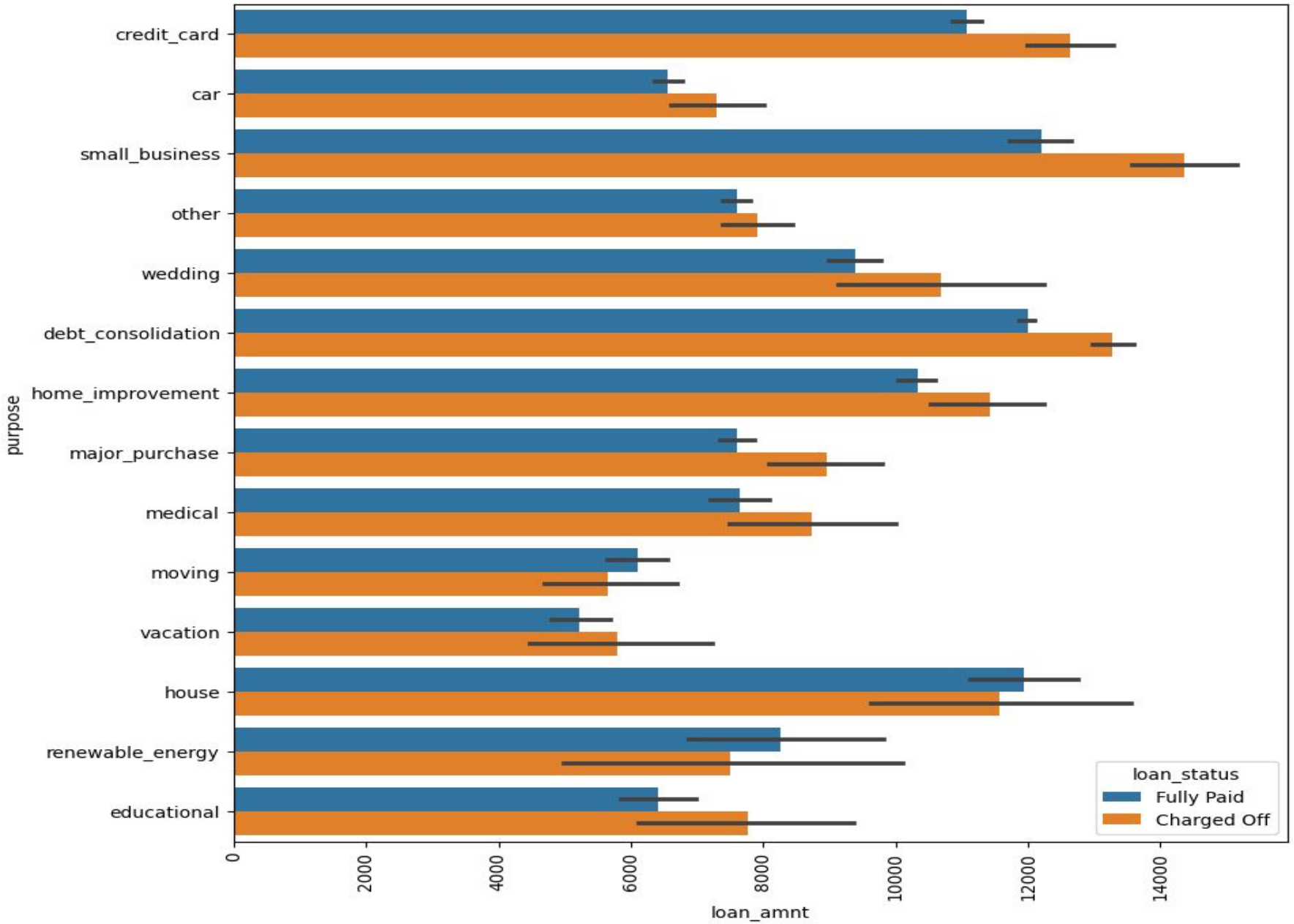
These plots illustrate how charged off loan status is affected by (interest rate, debt to income ratio) and (verification status, loan amount).





This plot show how charged off loan status is affected by loan amount and employment length of applicant.

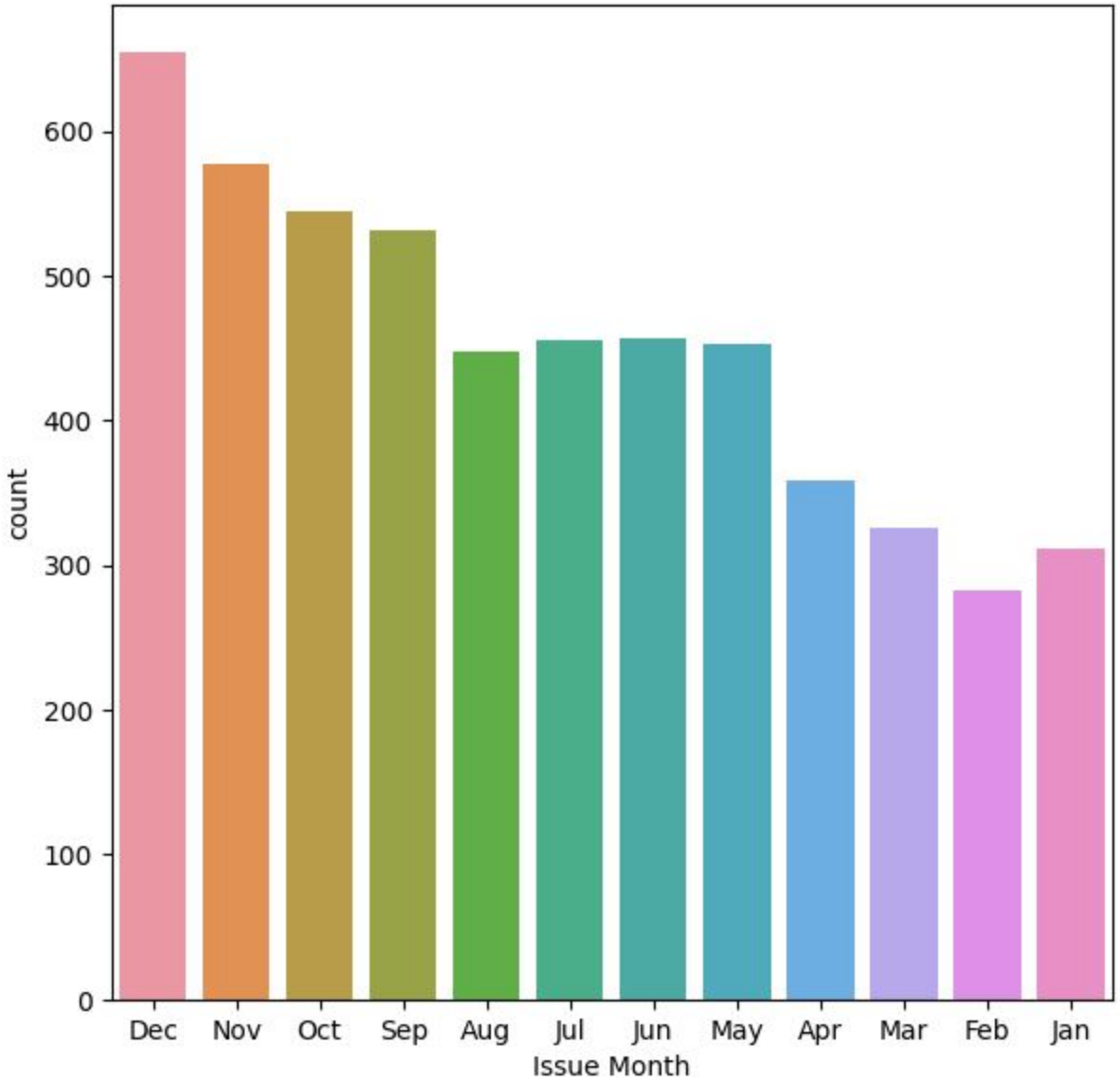
This plot illustrates how charged off loan status is affected by loan amount and the purpose of loan application.

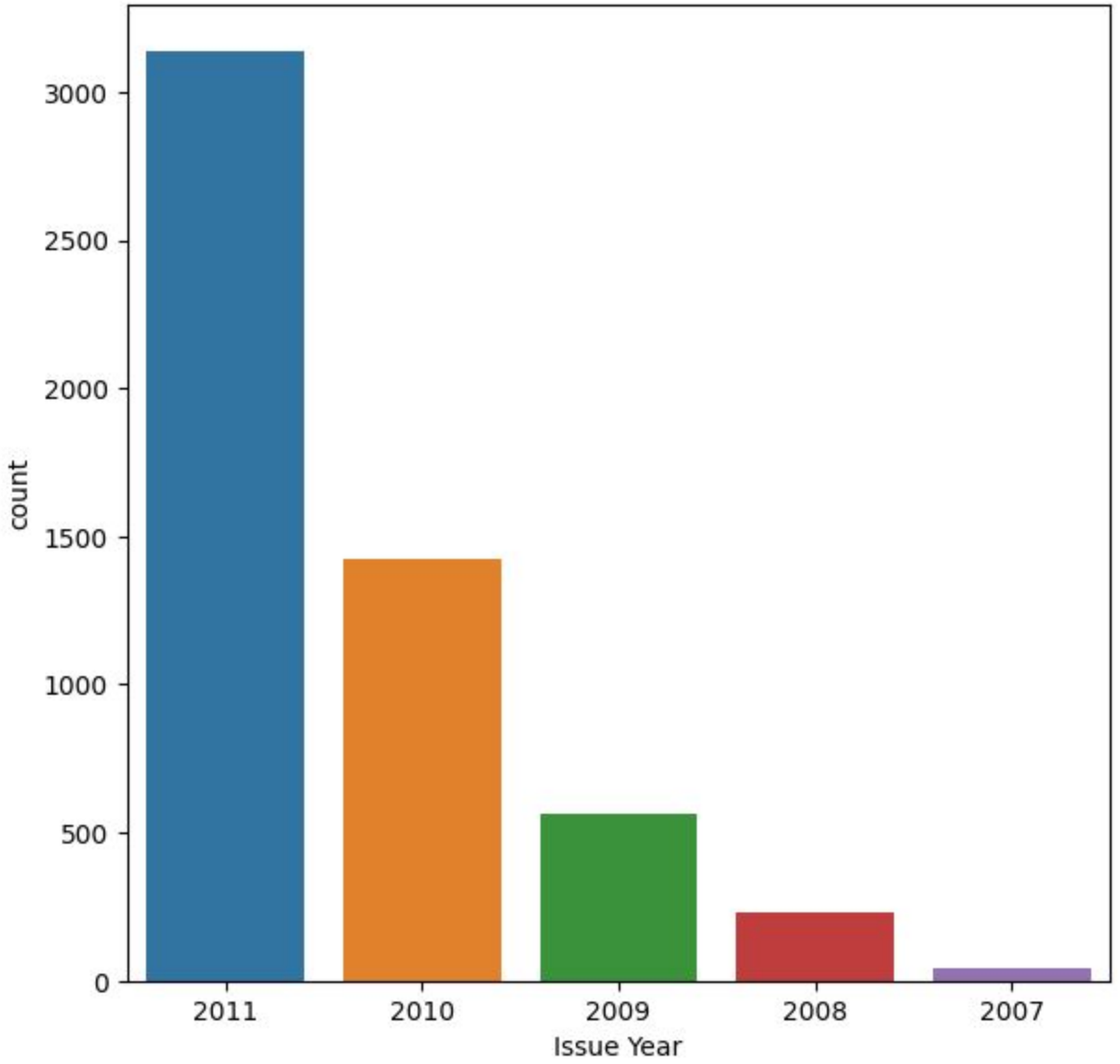


Observations

- It is observed that loans issued in last quarter of the calendar year (Oct, Nov, Dec) are being defaulted in high number when compared to other months. This might be because of the financial year in the USA. As financial year starts from OCT in the USA, people might be taking loans to document them in tax filing, but they are defaulting once the loan is sanctioned.

- As we have 2007-2011 data, we can observe that 2011 has highest number of defaults. This might be because of the financial crisis in the USA during 2011.





These plots illustrate month and year wise count of number of charged off loans.

Thank you!

Univariate and Segmented Univariate Analysis

Data Understanding

```
#Check size of dataset
data.shape
```

```
(39717, 111)
```

```
#Information of dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 39717 entries, 0 to 39716
Columns: 111 entries, id to total_il_high_credit_limi
dtypes: float64(74), int64(13), object(24)
memory usage: 33.6+ MB
```

```
#Data types of columns
data.dtypes
```

```
id                            int64
member_id                     int64
loan_amnt                     int64
funded_amnt                   int64
funded_amnt_inv             float64
                             ...
tax_liens                   float64
tot_hi_cred_lim             float64
total_bal_ex_mort           float64
total_bc_limit              float64
total_il_high_credit_limit  float64
Length: 111, dtype: object
```

In this section, we delved into fundamental aspects of our dataset. The shape, information, and data types collectively equip us with a foundational understanding of the data. This knowledge forms the basis for making informed decisions throughout our analysis and ensures that our insights are accurate and relevant.

Data Cleaning

```
[12]: #Check for total count of null values
      data.isnull().sum()
```

```
[12]: id                            0
      member_id                     0
      loan_amnt                     0
      funded_amnt                   0
      funded_amnt_inv               0
                                   ...
      tax_liens                    39
      tot_hi_cred_lim           39717
      total_bal_ex_mort         39717
      total_bc_limit            39717
      total_il_high_credit_limit    39717
      Length: 111, dtype: int64
```

There are lot of columns with all null values. Let's remove them

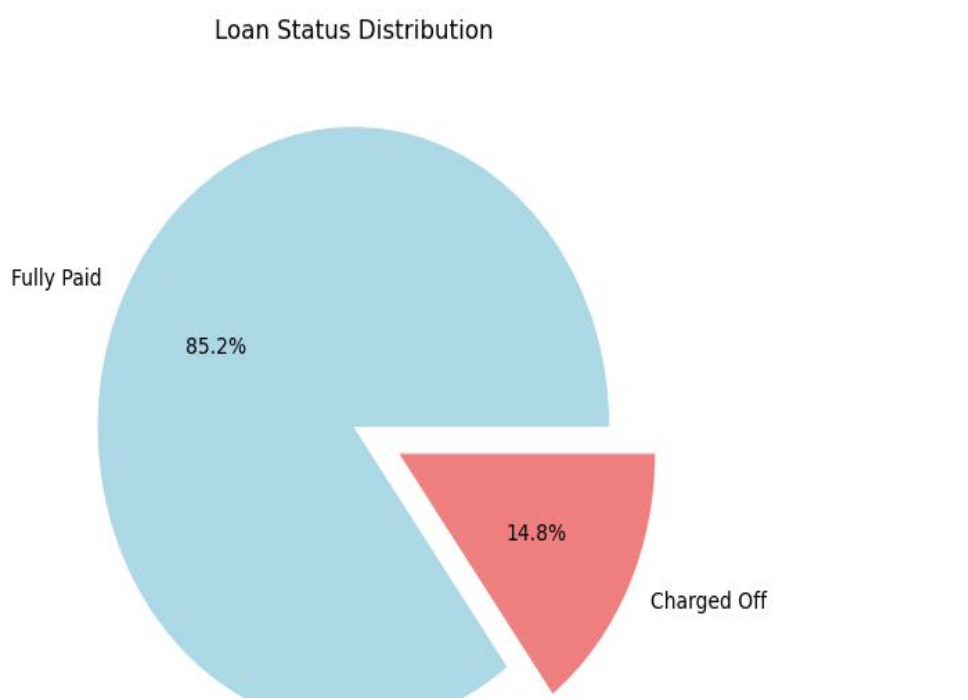
```
[13]: #Remove columns which have all null values
      data.dropna(axis = 1, how = 'all', inplace = True)
```

We have Identify and evaluate columns that contribute minimally to the analysis objectives or contain redundant information. By addressing missing values, we have enhanced the dataset's reliability and ensure that insights drawn are more representative and accurate.

Data Analysis

We have conducted a thorough examination to identify outliers within the quantitative variables of the dataset. Additionally, we have standardized the data to ensure uniformity and comparability. Our analysis further encompassed both univariate and bivariate approaches, allowing us to ascertain the default rate and identify individuals more inclined to avail the loan. In our upcoming presentation, we will delve into illustrative examples showcasing these analytical processes

Univariate Analysis

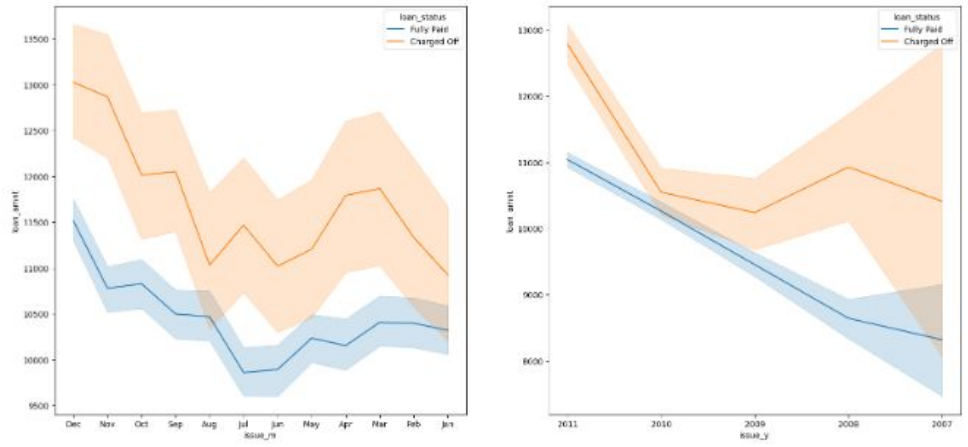


We have looked at each variable separately to better understand the data, focusing on its distribution, central tendency, and spread. This approach has allowed us to uncover patterns, trends, and potential outliers within individual variables, contributing to a deeper comprehension of the dataset and we have gathered some valuable insights from this analysis.

Bivariate Analysis

```
In [76]: #loan amount across months and years
         plt.figure(figsize=(20,20))
         plt.subplot(221)
         sns.lineplot(data =data,y='loan_amnt', x='issue_m', hue ='loan_status')
         plt.subplot(222)
         sns.lineplot(data =data,y='loan_amnt', x='issue_y', hue ='loan_status')

Out[76]: <Axes: xlabel='issue_y', ylabel='loan_amnt'>
```



From the above subplots we can observe that, higher loan amounts are being charged off in the year end. Also, 2011 has more defaulters with high loan amount when compared to other years

Bivariate analysis involves studying the relationship between two variables to understand how they interact or influence each other. By comparing pairs of variables, we have aimed to uncover meaningful connections and patterns that might not be apparent when looking at each variable in isolation. we have also examined how changes in one variable corresponded to changes in another.

Conclusion

The probability of an applicant to default a loan is high based on following factors.

The applicant has the house ownership category as "RENT."

The loan application category is of "B" Grade.

The applicant uses the loan to clear other debts (debt_consolidation).

The loan tenure is 36 months.

The applicant verification status is "not verified.

Conclusion

The probability of an applicant to default a loan is high based on following factors.

Applicants who have taken a loan amount in the range of 30k-35k are charged an interest rate of 15-17.5%.

Applicants who have taken a loan for a small business and the loan amount is greater than 14k.

The applicant's loan category is grade F, and the loan amount is 15k-20k.

The applicant's employment length is ten or more years, and the loan amount is 12k-14k.

The applicant is verified, and the loan amount is above 16k

Thankyou