Lending Club Case Study

DEBABRATO SENGUPTA

Lending Club Case Study

Operates a "peer-to-peer" lending website for personal loans. The company assesses applicants' risk and lets investors lend directly to individuals or spread their money across a number of loans.

They want to assess the risk by analysing a dataset with previous loan related data

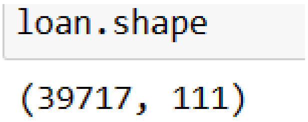
The purpose is to find out some variables which might help them to predict the occurrence of default in order to lower their risks and increase profits

Cleaning the Dataset

Data Cleaning Summary

Initial number of rows : 39717

Initial number of columns : 111



```
loan.shape
```

```
(39717, 111)
```

No Summary rows at the top or bottom of the table found

Number of columns with all “N/A” values = 54

Number of columns found to be not required for analysis = 35

Removed all rows with loan status as “Current”, as current loans will not help us to determine if the loan will end up being default or not

Removed outliers

Cleaned numeric data by removing %, <, >, string values

After Cleaning

Number of rows: 36970

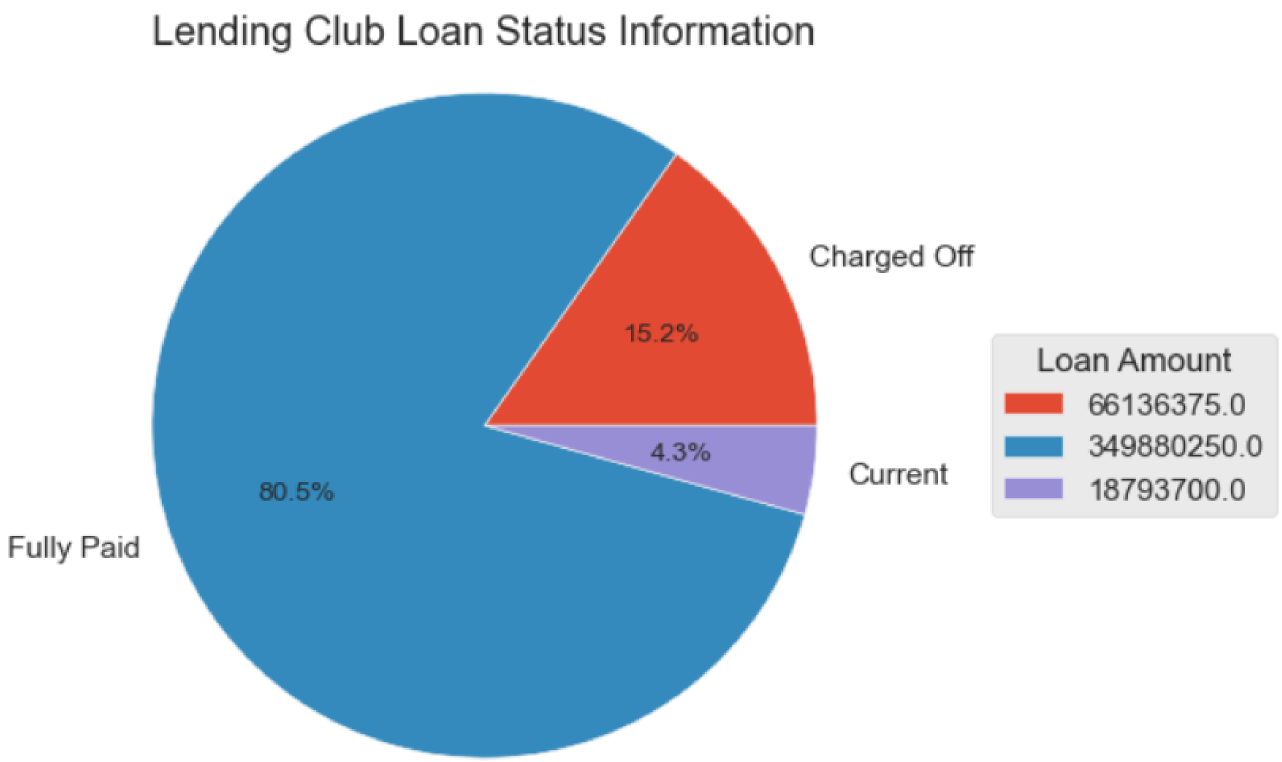
Number of Columns: 23

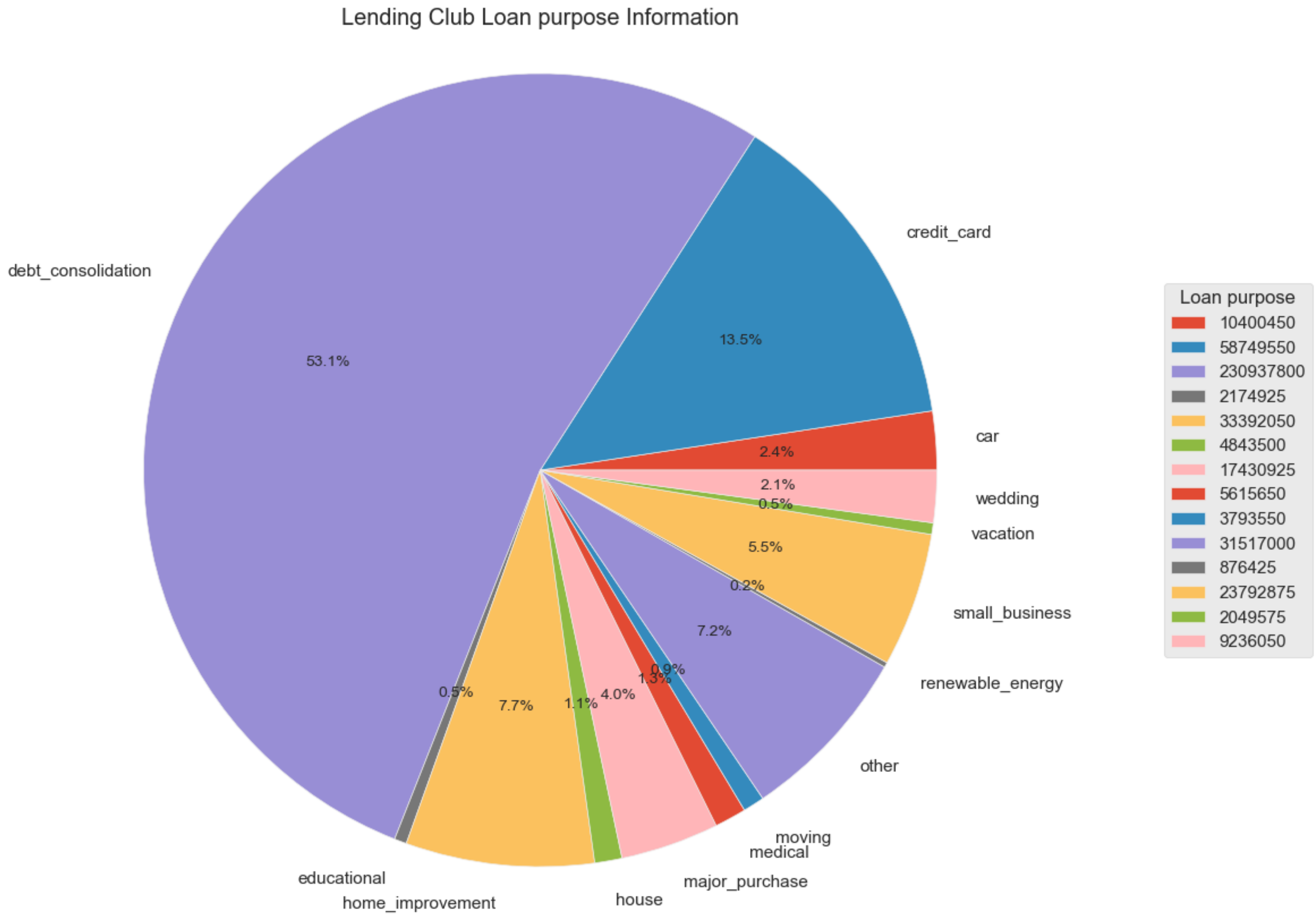
Loan Status Info

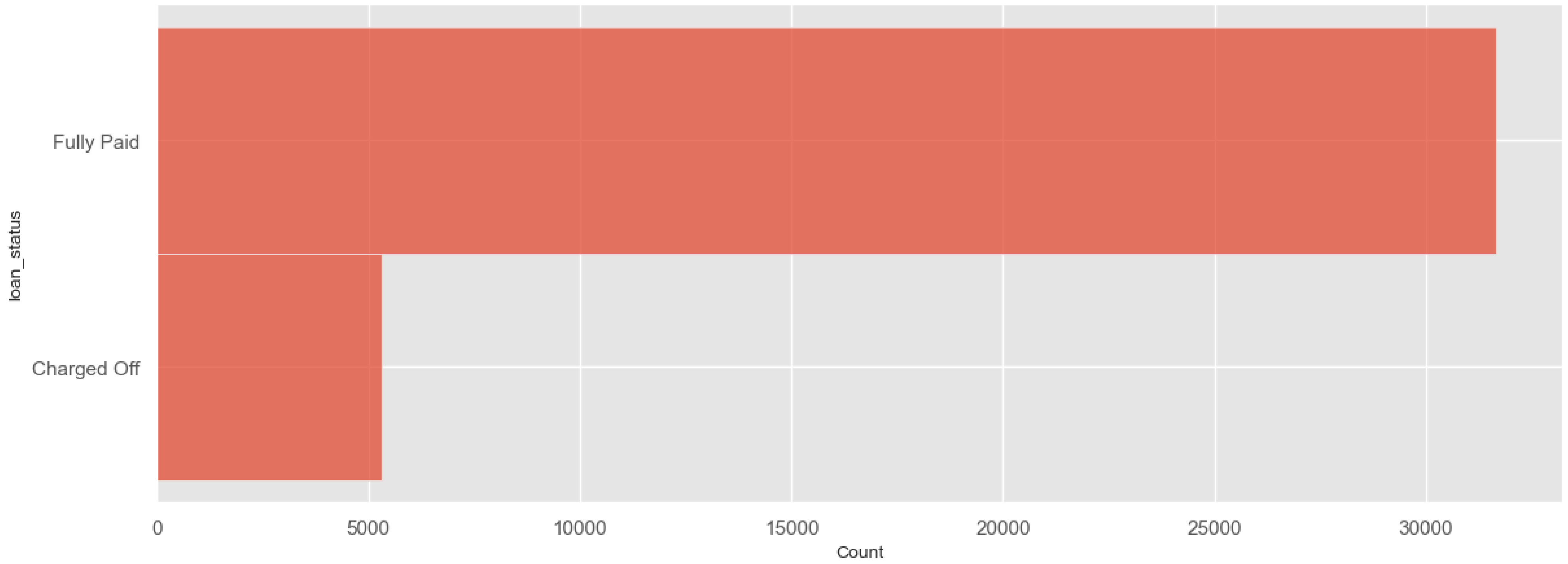
“Current” Loans would not help in determining if the loan would default.

The Pie chart shows the loan status.

We can see that 4.3 % of the loans are current




Lending Club Loan purpose Information
debt_consolidation
53.1%
credit_card
13.5%
car
2.4%
wedding
2.1%
vacation
0.5%
small_business
5.5%
renewable_energy
0.2%
other
7.2%
moving
0.9%
medical
1.3%
major_purchase
4.0%
house
1.1%
home_improvement
7.7%
educational
0.5%
Loan purpose
10400450
58749550
230937800
2174925
33392050
4843500
17430925
5615650
3793550
31517000
876425
23792875
2049575
9236050



Loan Status post cleanup

Univariate Analysis

Purpose and approach

The purpose of the Univariate analysis is to identify the variables which have an impact on the loan default.

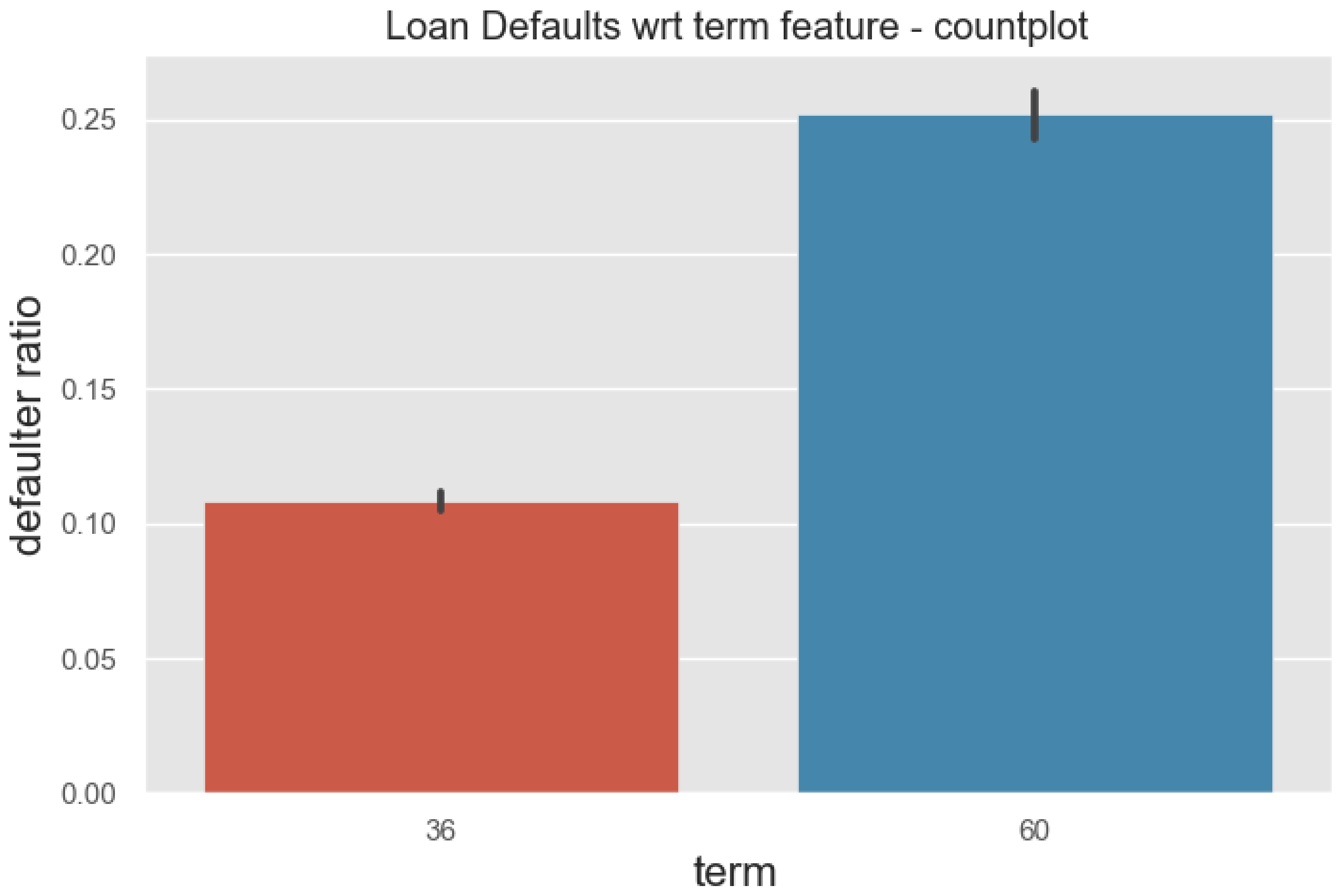
This will help Lending Club to make an informed decision when approving the loan

A function has been written which taken the feature as input. The feature can be any of the columns

The function plots a countplot with the data and can be used to infer the impact

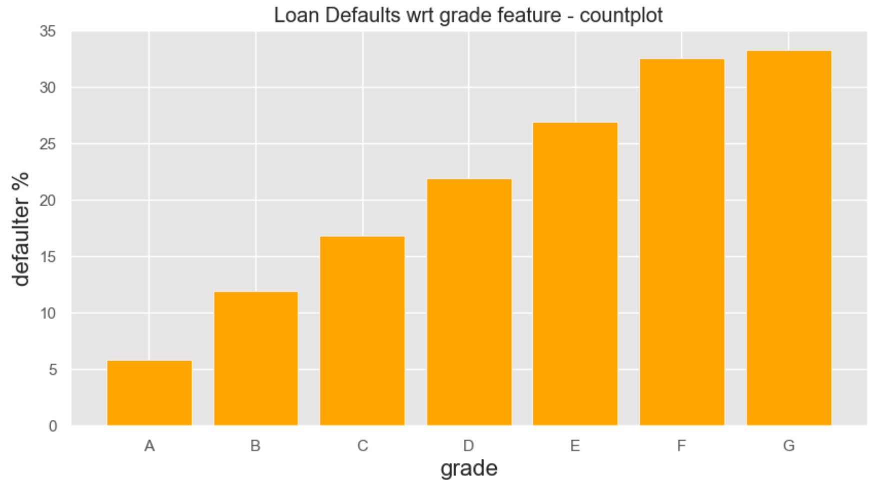
Default wrt term

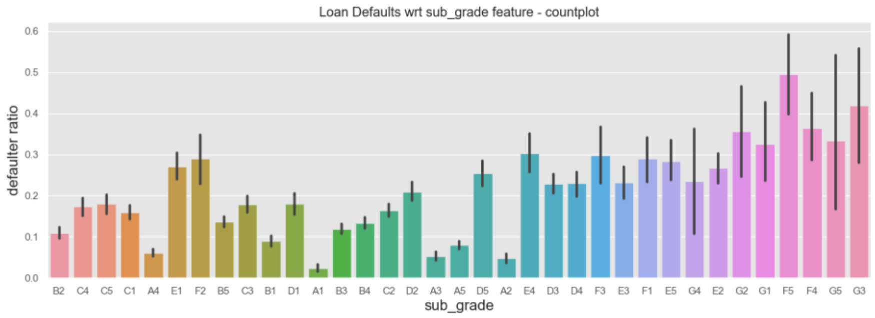
Defaulter rate is higher for loan term of 60 months



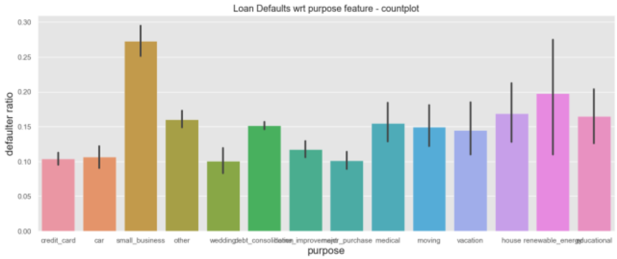
Default wrt Grade and Subgrade

It can be inferred from the plots that both grade and sub_grade have a relation with default rate

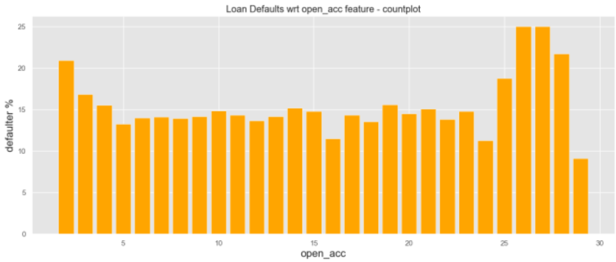




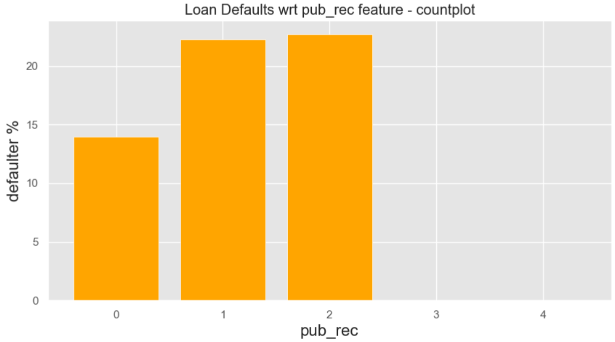
Default wrt purpose, open_acc, pub_rec, home_ownership



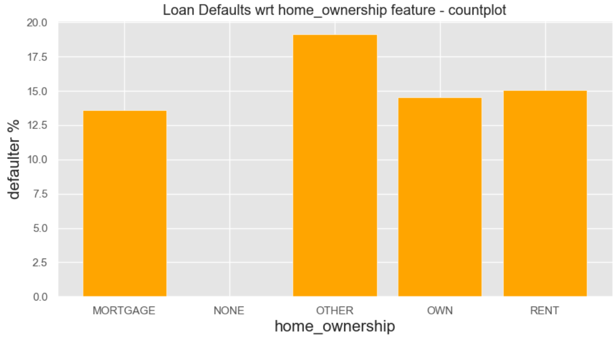
Loan Defaults wrt purpose feature - countplot



Loan Defaults wrt open_acc feature - countplot



Loan Defaults wrt pub_rec feature - countplot



Loan Defaults wrt home_ownership feature - countplot

We can infer the following from the plots:

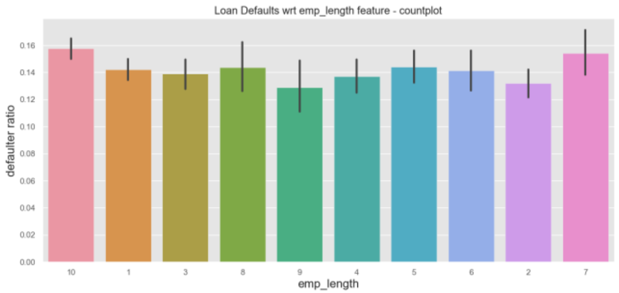
Purpose – Big spike noticed for “small_business”. Thus purpose is relevant and useful.

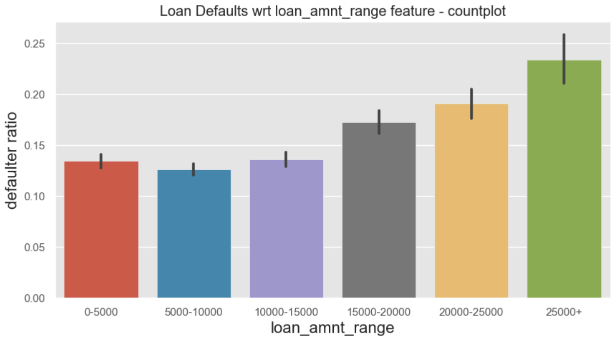
Open_acc – Almost constant. Ts variable is not useful.

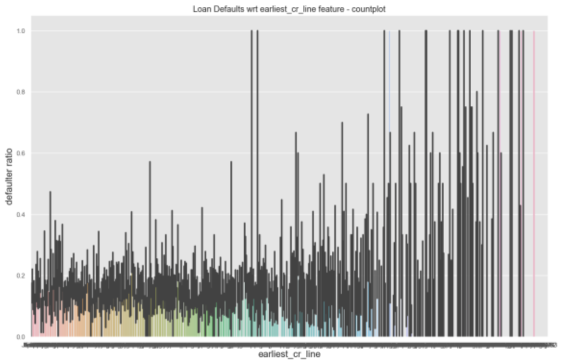
Pub_rec – rate of defaulters rises by > 5% for pub_rec > 0. We will consider this as useful

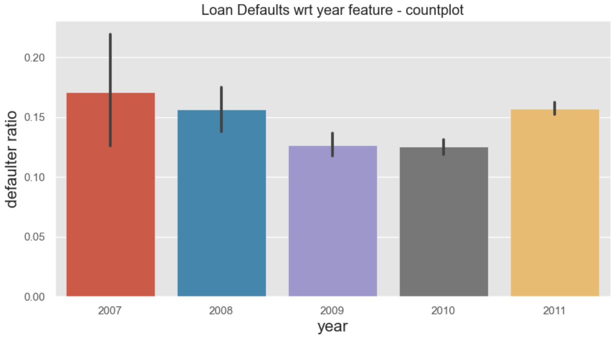
Home_ownership – Only some increase is seen for “Other”, but as we don’t have data on what “Other” comprises of, we will not consider this variable.

Default wrt emp_length, loan_amt_range, year, earliest_cr_line









The following can be inferred from the charts:

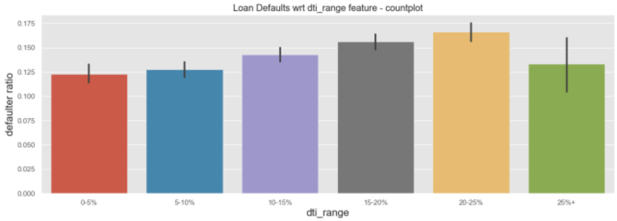
Emp_length – defaulter ratio is almost constant, thus this variable is not useful.

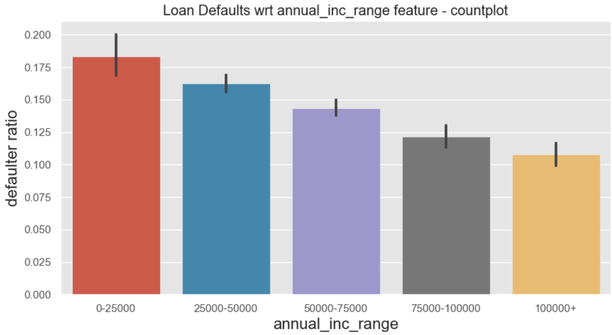
Loan_amnt_range – Ratio is increasing, thus this variable is useful.

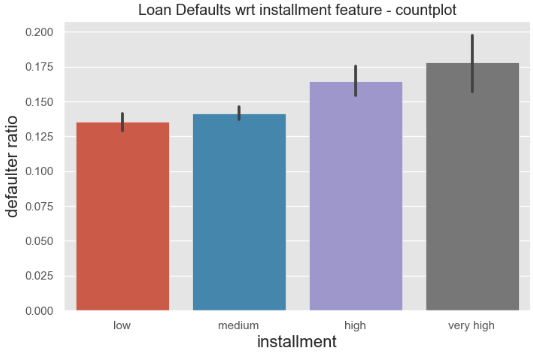
Earliest_cr_line – Ratio is almost constant, thus the variable is not useful.

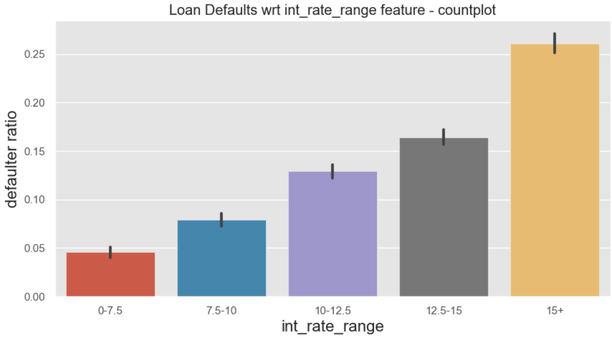
Year – Ratio is almost constant over the years, thus this variable also cannot be useful.

Default wrt dti_range, annual_inc_range, installment, int_rate_range









Following can be inferred:

Dti_range – Defaulter rate increases with increase of dti. Thus this variable is also useful.

Annual_inc_range – Defaulter rate reduces with increase in annual income. Thus this variable is useful.

Installment – Defaulter ratio is increasing with increase in installment, thus this variable is useful.

Int_rate_change – Defaulter ratio is increasing with increase in int_rate. This this variable is useful.

Univariate Results Sumary

The following are the important features we have deduced using univariate analysis:-

term, grade, sub_grade, purpose, pub_rec, funded_amnt, int_rate, annual_inc, dti, installment

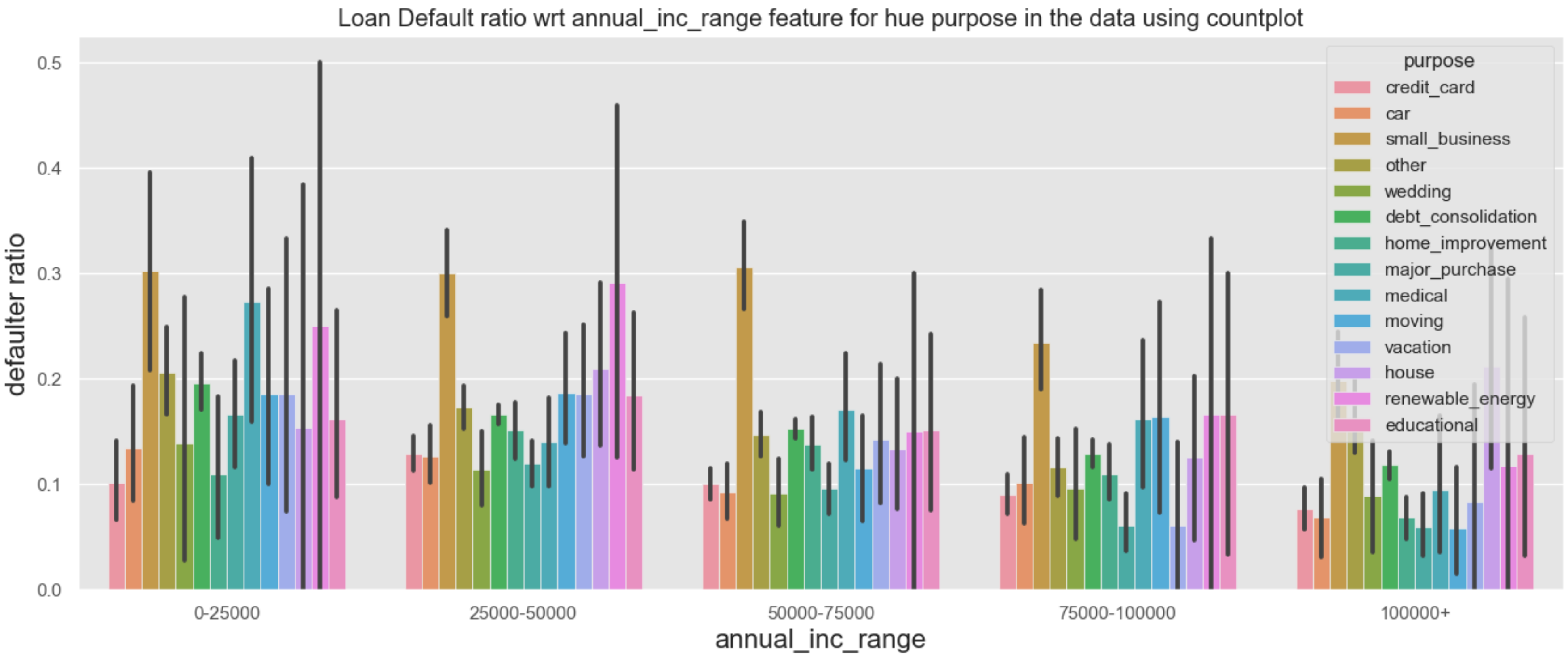
Bivarate Analysis

Purpose and Approach

The purpose of the Bivariate analysis is to identify the relation between 2 variables and identify if they have an impact on the loan default.

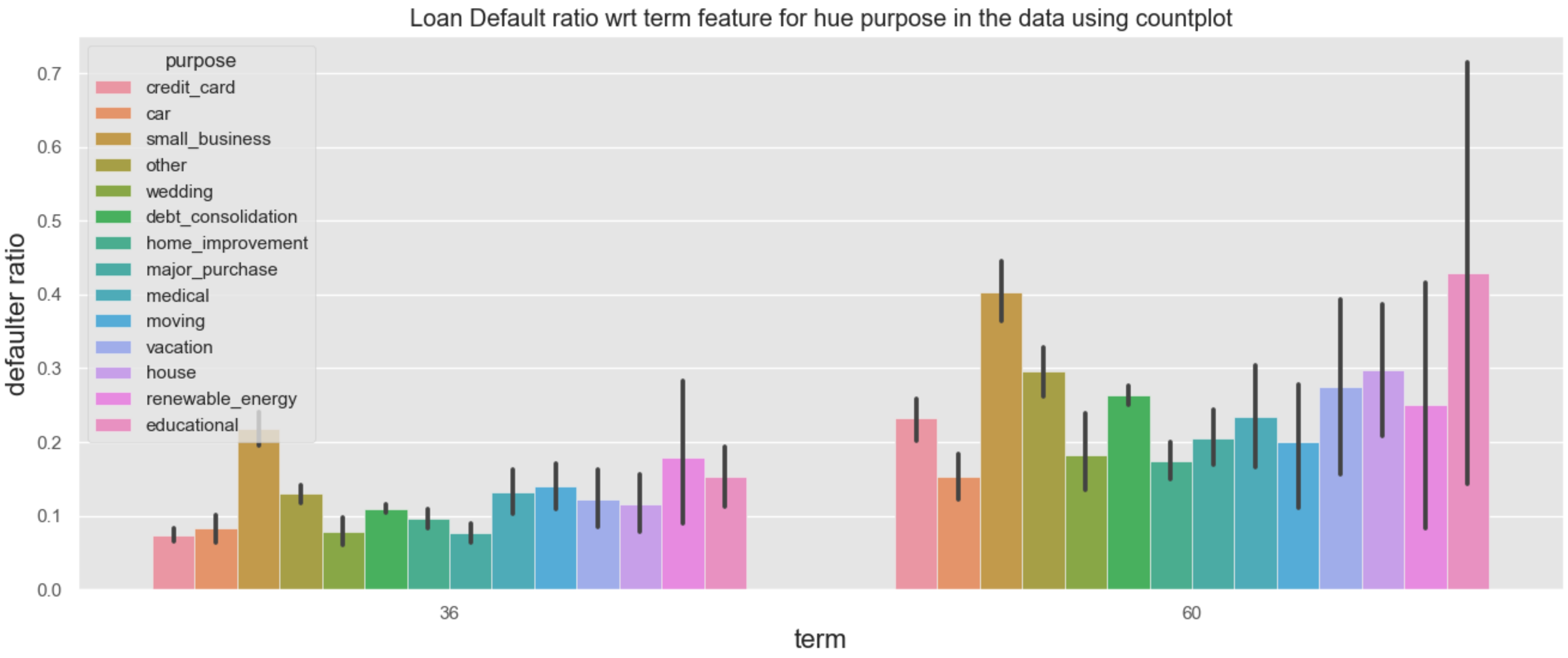
This will help Lending Club to make an informed decision when approving the loan

A function has been written which takes 2 features as input. The features can be any of the columns



Annual_inc_range and purpose

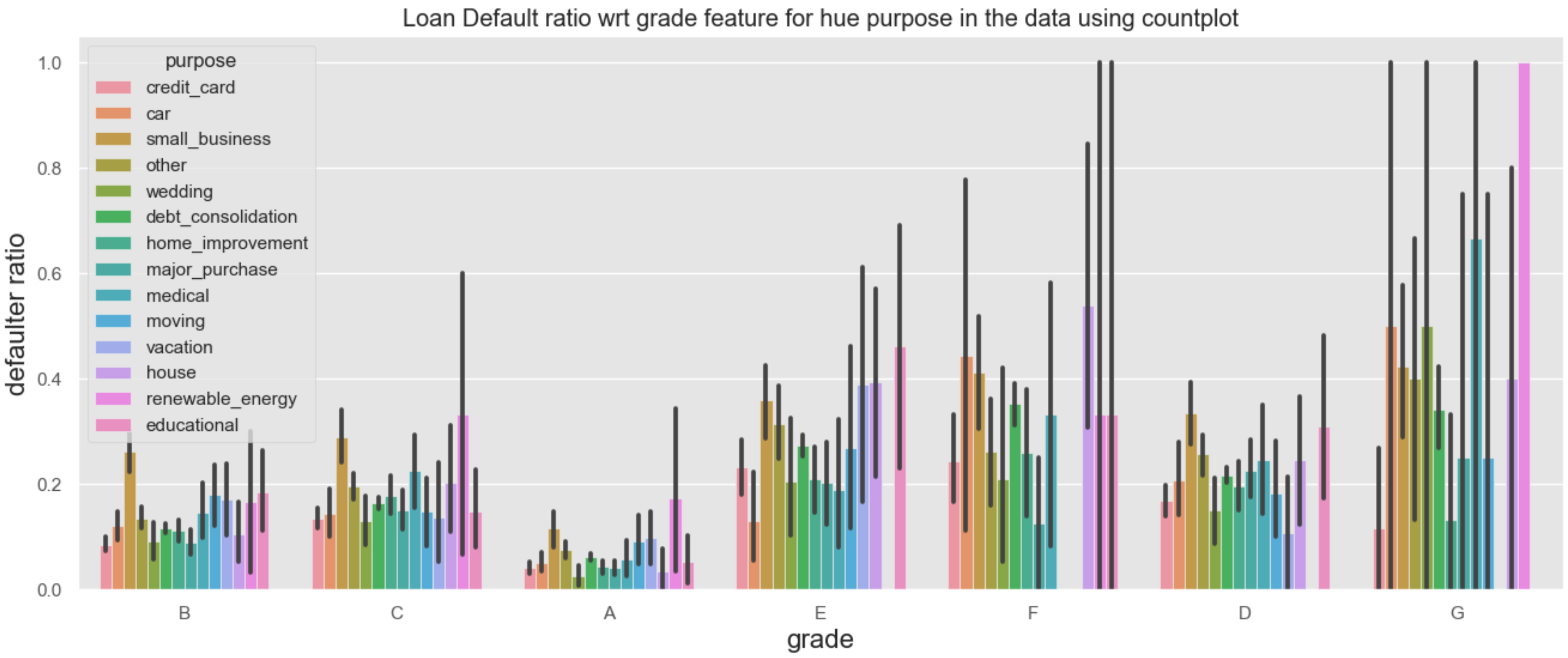
From the above, we can infer that there is no correlation between the annual_income_range and purpose wrt defaulter ratio



Term and purpose

As we can see straight lines on the plot, default ratio increases for every purpose wrt term

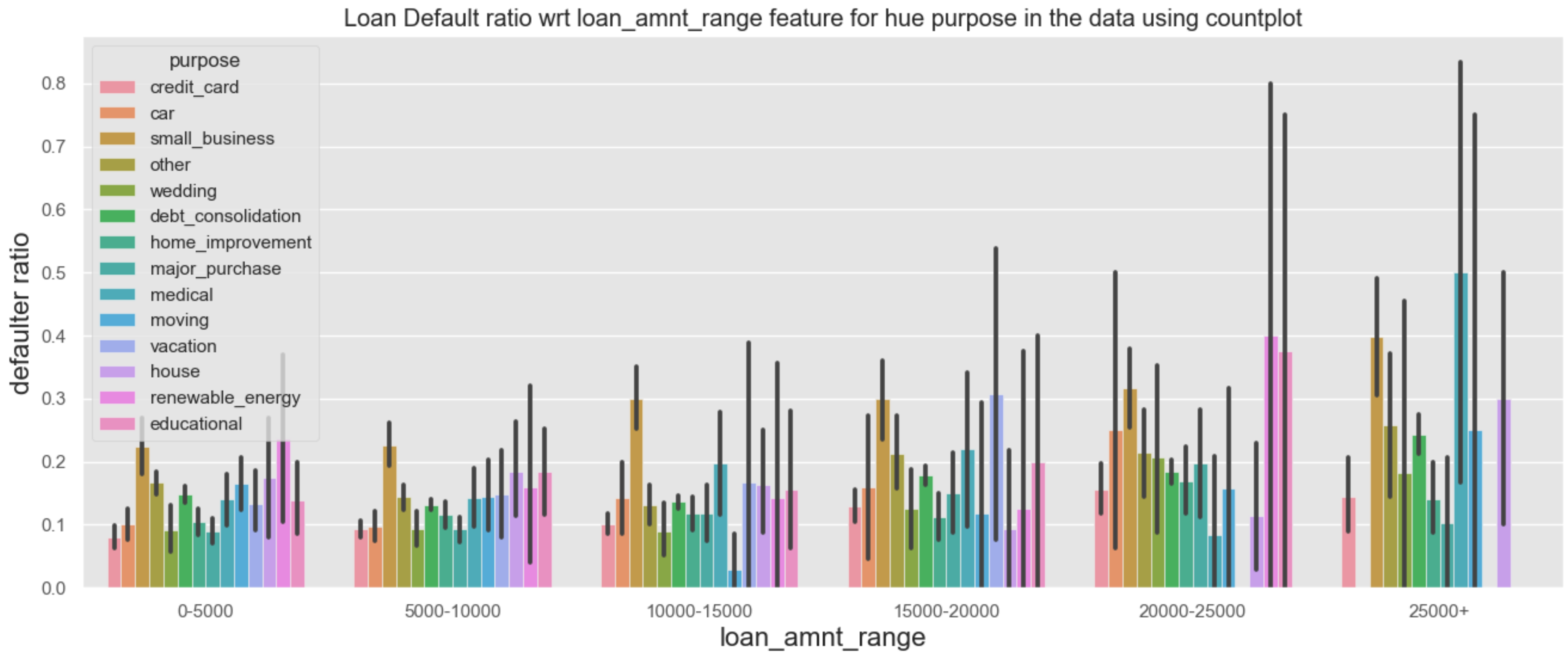
There is a relation



Grade and prpose

As we can see straight lines on the plot, default ratio increases for every purpose wrt grade

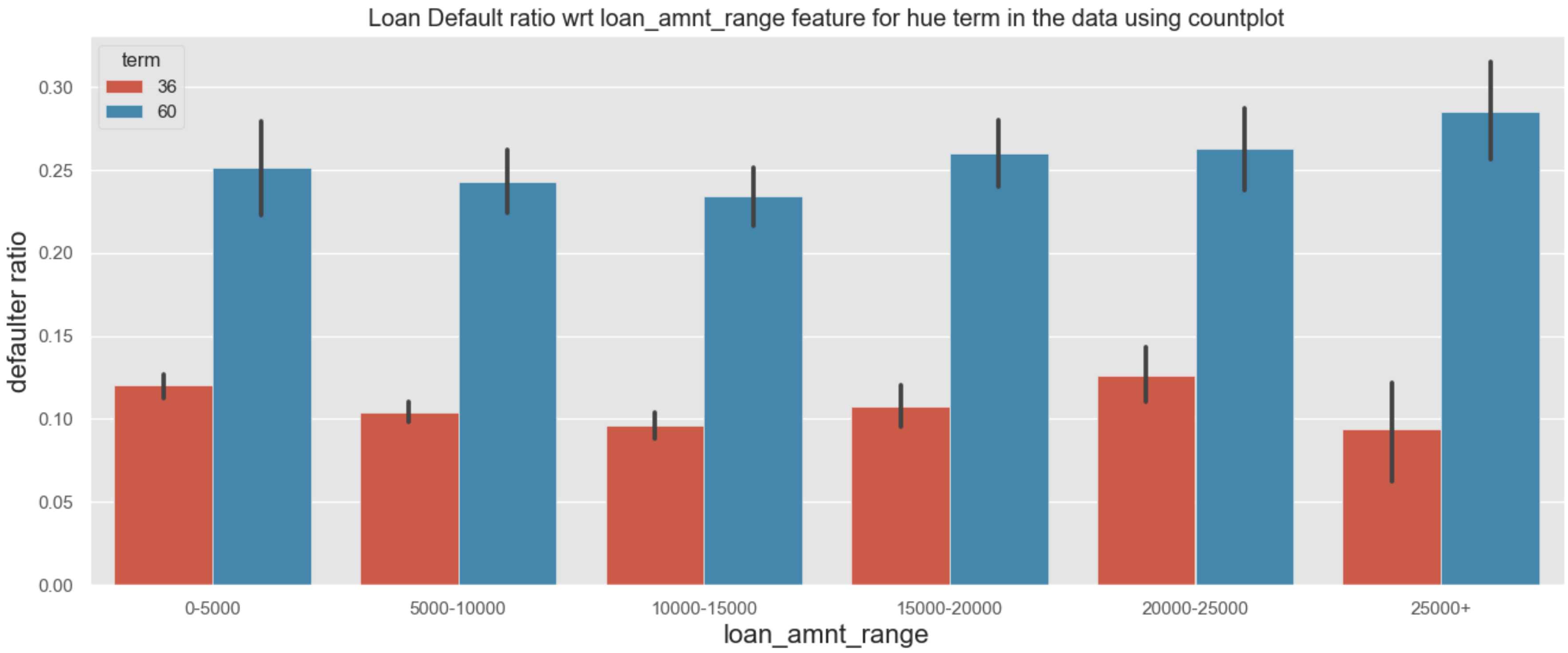
There is a relation



Loan_amnt_range and purpose

As we can see straight lines on the plot, default ratio increases for every purpose wrt loan_amnt_range

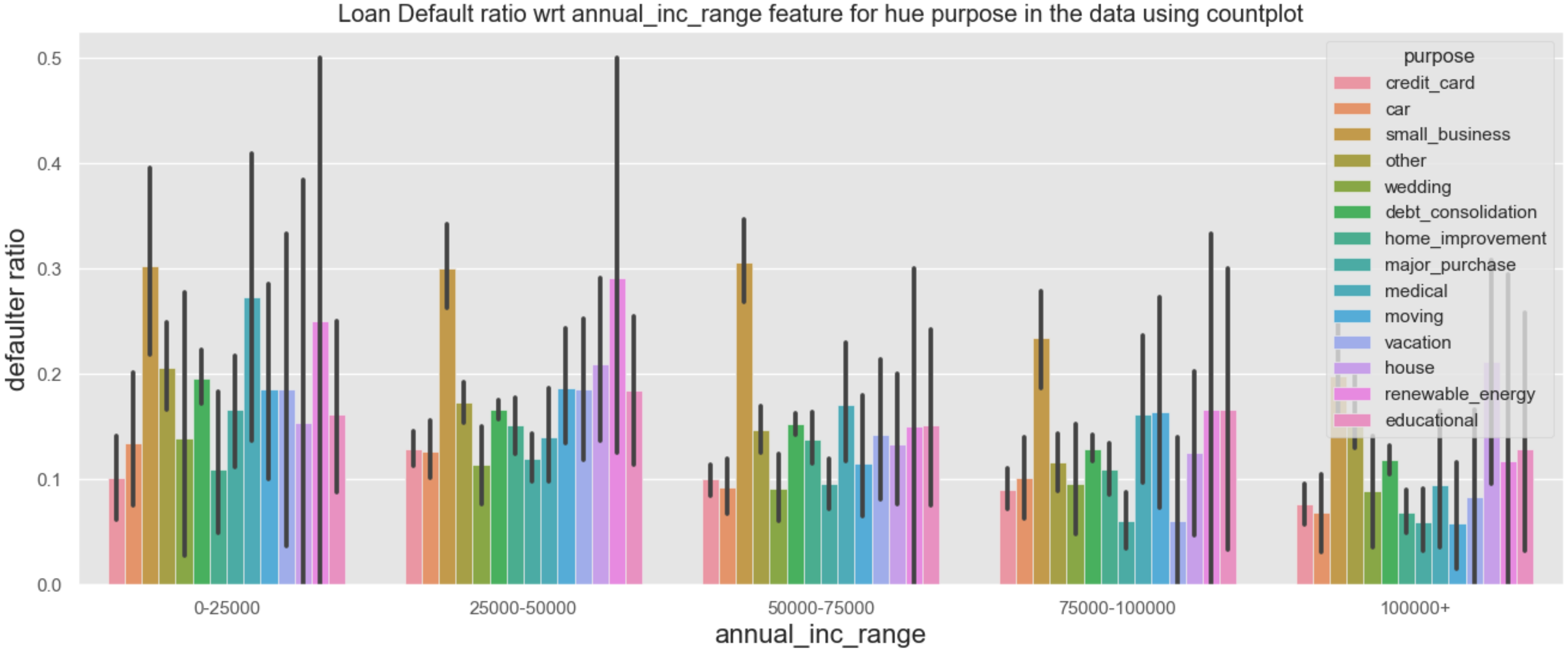
There is a relation



Loan_amnt_range and Term

As we can see straight lines on the plot, default ratio increases for every term wrt loan_amnt_range

There is a relation



Annual_inc_range and purpose

As we can see straight lines on the plot, default ratio increases for every purpose wrt annual_inc_range

There is a relation

Final Findings

After analysing all the related features available in the dataset, we have come to an end, deducing the main driving features for the Lending Club Loan Default analysis:

The best driving features for the Loan default analysis are: term, grade, purpose, pub_rec, int_rate, installment, annual_inc, funded_amnt