Lending Club Case Study

ANALYSIS ON THE LOAN DATA OF A FINANCIAL INSTITUTION

Analysis Overview

- Data from consumer finance company which specializes in lending various types of loans to urban customers.
- Given data 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.
- Identify driving factors behind loan default

Data Overview

- Data set contains 39,717 rows and 111 columns
- Few columns like ‘tot_hi_cred_lim’, ‘total_bal_ex_mort’ etc. have only ‘NA’ values
- Some columns like ‘member_id’, ‘Id’, ‘pymnt_plan’ etc. have almost similar values across the rows
- Loan status column has three values ‘Fully Paid’, ‘Current’ and ‘Charged Off’
- There are few columns which can be further used to derive new columns

	id	member_id	loan_amnt	funded_amnt	funded_amnt_inv	term	int_rate	inst
0	1077501	1296599	5000	5000	4975.0	36 months	10.65%	
1	1077430	1314167	2500	2500	2500.0	60 months	15.27%	
2	1077175	1313524	2400	2400	2400.0	36 months	15.96%	
3	1076863	1277178	10000	10000	10000.0	36 months	13.49%	
4	1075358	1311748	3000	3000	3000.0	60 months	12.69%	

5 rows × 111 columns

```
print("Shape of data is :",Ldata.shape)
print("*"*100)
Ldata.info()
```

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

Manipulations in Data to achieve desired results

- Deleting few columns as they are not needed for analysis

```
Ldata.drop(Ldata.columns[[1,17,18,19,22,35,50,51,52,53,54,55,56,57,58,59,60,61,62,63,64,65,66,67,68,69,70,71,
                          72,73,74,75,76,77,79,80,81,82,83,84,85,86,87,88,89,90,91,92,93,94,95,96,97,98,99,100,
                          101,102,103,104,107,108,109,110]],axis=1,inplace=True)
Ldata.head()
```

- Removing the records with loan status as ‘Current’ as these records will serve no purpose in default loans analysis

```
Ldata.drop(Ldata.index[Ldata['loan_status'] == 'Current'], inplace=True)
print("Shape of data is :",Ldata.shape)
```

```
Shape of data is : (38577, 47)
```

Manipulations in Data to achieve desired results

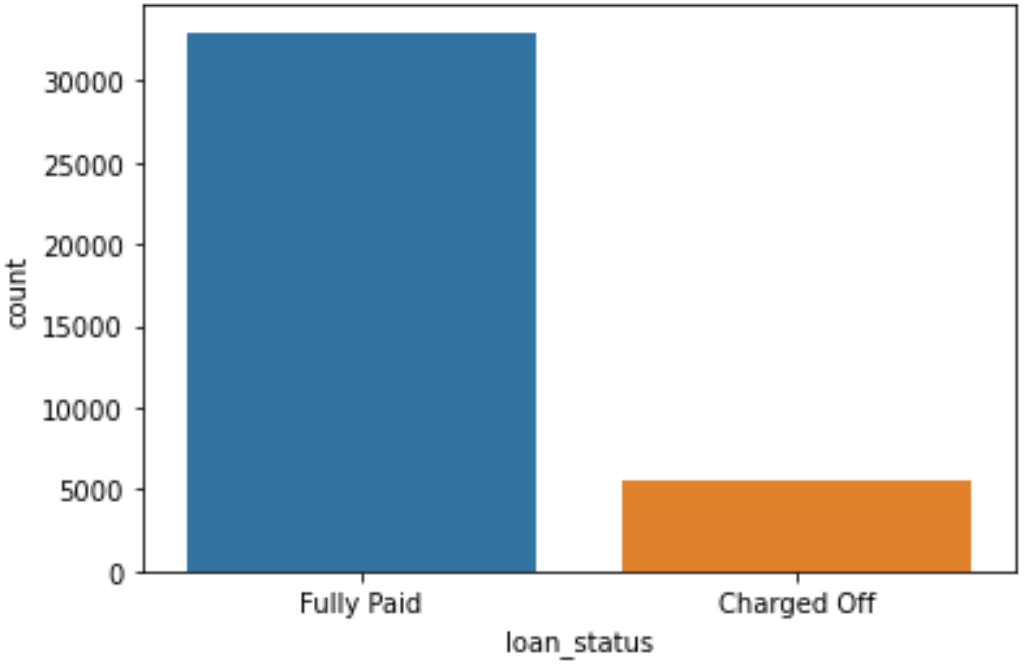
- Deriving a new column from ‘loan status’ to help us make some bar charts

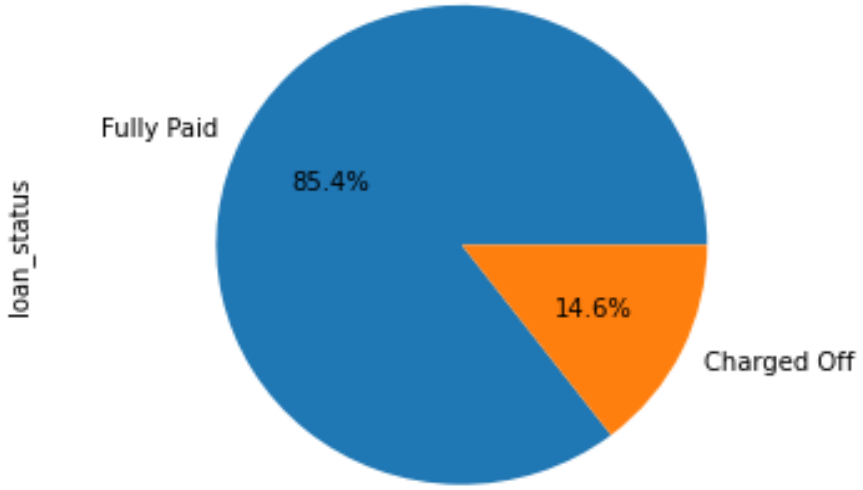
```
Ldata['loan_status_update'] = Ldata['loan_status'].apply(lambda x: 0 if x=='Fully Paid' else 1)
Ldata['loan_status_update'] = Ldata['loan_status_update'].apply(lambda x: pd.to_numeric(x))
```

- Creating a new field out of ‘Loan Amount’ to define some specific categories

```
## If we look at the data in loan amount column, we have minimum 500 and maximum 35000. Categorizing based on this range.
def loan_grouping(x):
    if x < 10000:
        return 'Less 10K'
    elif x >=10000 and x < 20000:
        return '10K-20K'
    else:
        return 'Greater 20K'

Ldata['loan_amnt_group'] = Ldata['loan_amnt'].apply(lambda x: loan_grouping(x))
```

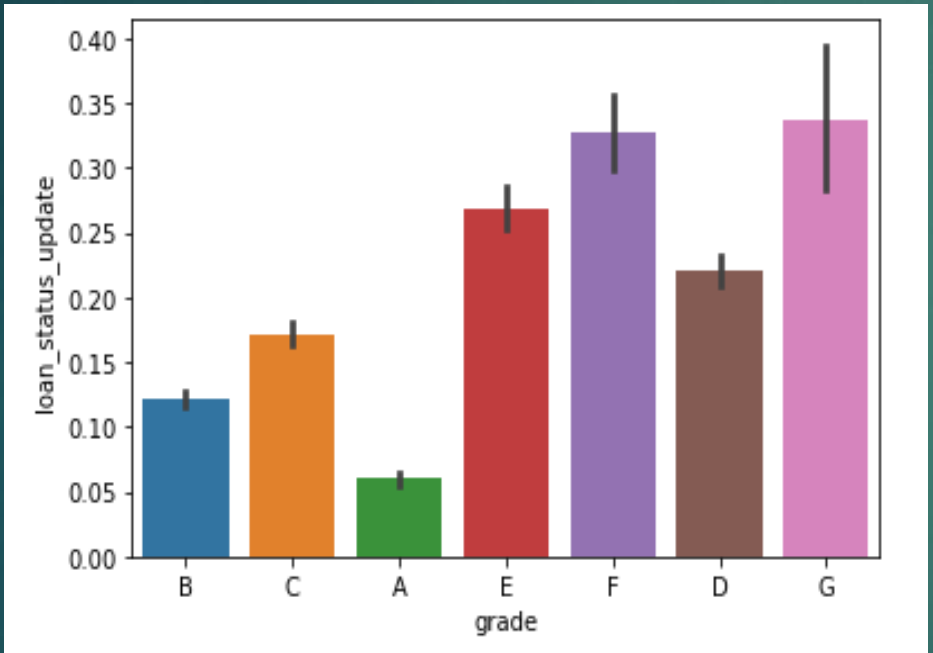




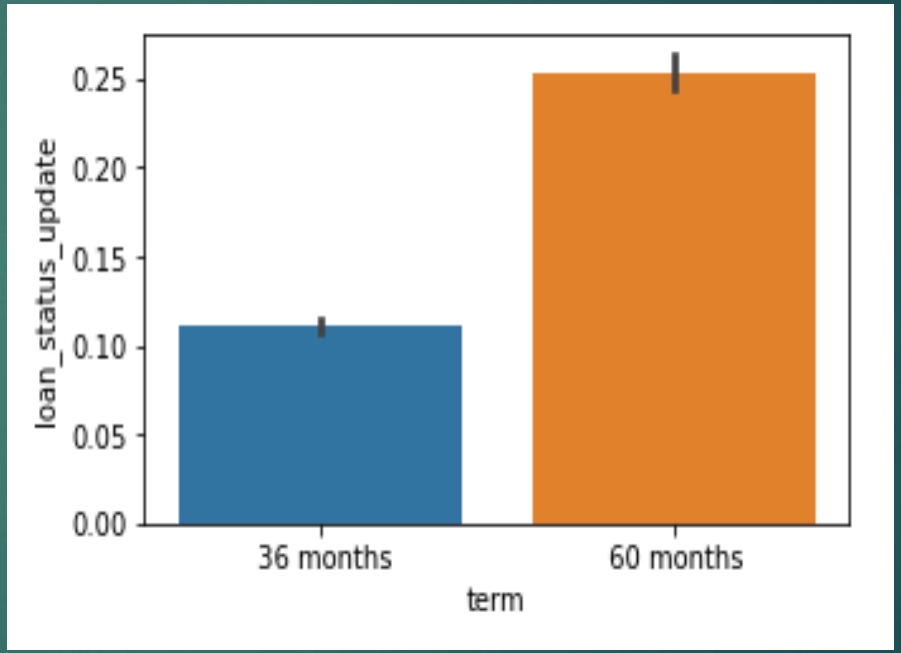
Overview of Default Cases

Bar Plots Against Loan Status

Grade Vs Default Rate



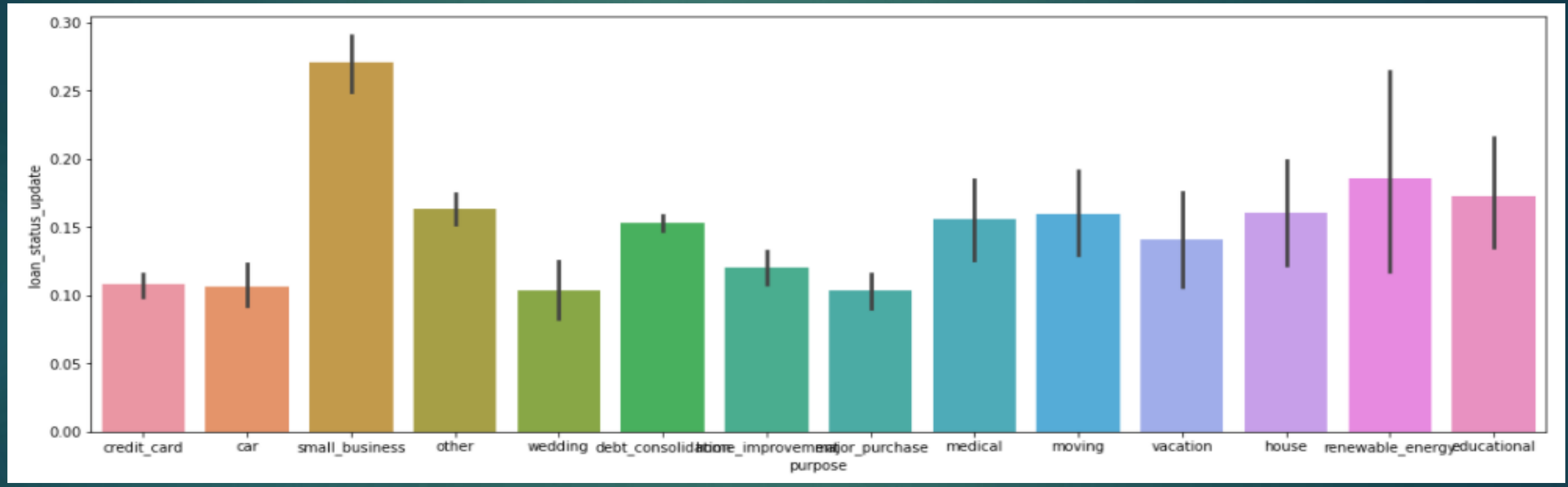
Tenure Vs Default Rate



- Grade ‘G’ is having high rate of loan default as compared to other grades and grade ‘A’ has the lowest default rate
- As the tenure of loan increases the default rate also increases

Bar Plots Against Loan Status

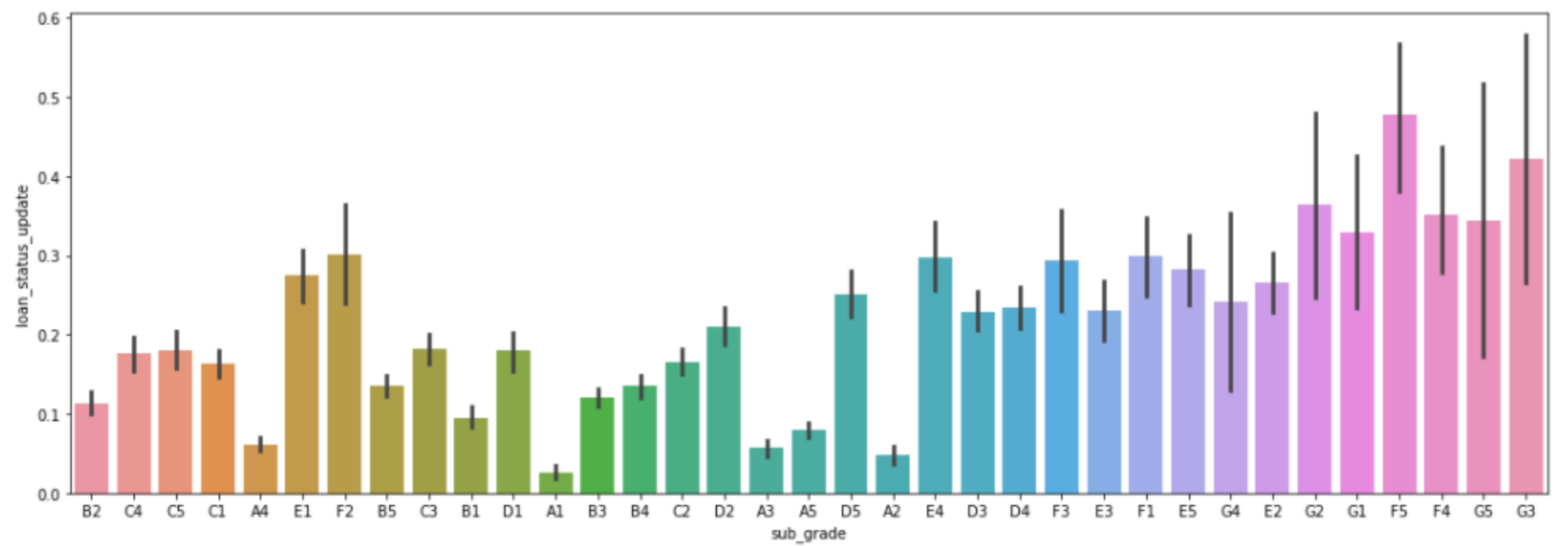
Purpose Vs Default Rate



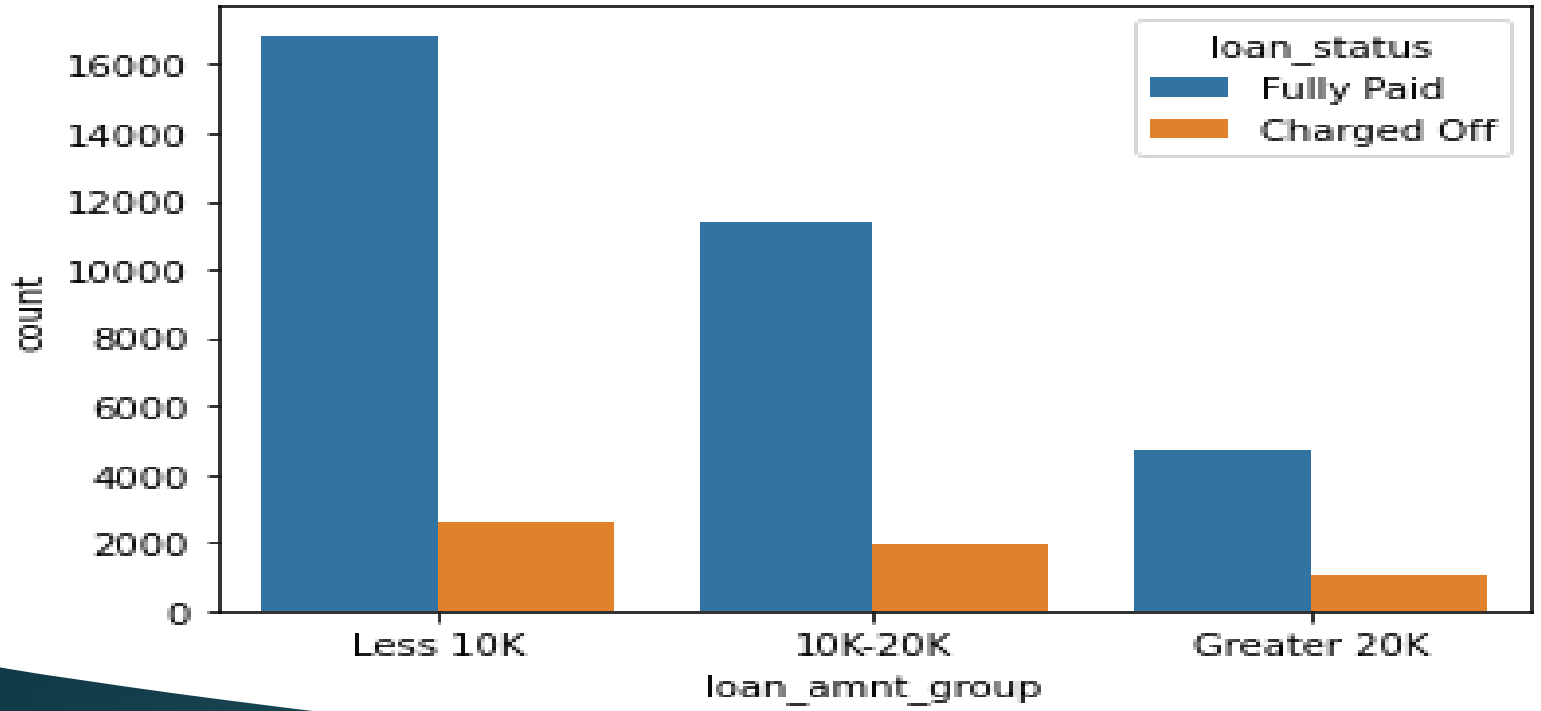
- ‘Small Business’ has the highest rate of loan default as compared to other categories
- ‘Wedding’, ‘credit Card’ and ‘Car’ have almost the same default rates and these rates are quite low as compared to other categories

Bar Plots Against Loan Status

Sub Grade Vs Default Rate



- Sub Grade ‘F5’ has the highest default rate
- Sub Grade ‘A1’ has the lowest default rate



Loan Amount Trend

The Percentage of Default increases as we move towards the higher category of loan amount. We have very high ratio of default in case of ‘Greater 20 K’ as compare other categories.

Conclusions

As per our analysis, we have the following driving factors behind loan default:

- In the entire data around 14.5% of the cases are 'Default' and from the bar chart it can be seen that almost 5K loans are default.
- The conclusion is that the provided data proves a significant correlation between the default rates and following variables:
  - Grade
  - Term
  - Purpose
  - Sub Grade
- If we derive one field from the loan amount, then we can clearly see that as the loan amount increases the chances of default also increases
- From the purpose of loan one can notice that 'Small Business' have the highest rate of default
- Similarly, we can identify the risk from Grade and Sub Grade as well. Grade G, F and E have high volume of defaults. Grade G has the highest default rate and grade A has lowest rate. Sub Grade 'F5' clearly has high default as compared to others and 'A1' falls to the lowest.

Thank You