Lending Club – Case Study

Lending Club – an online agency which hosts a marketplace to mediate between investors and borrowers wants to analyse their data to minimise the risk of losing money while lending to customers.

We were offered a small subset of data between 2007 and 2011 (5 years) to identify driving factors for a profitable business model.

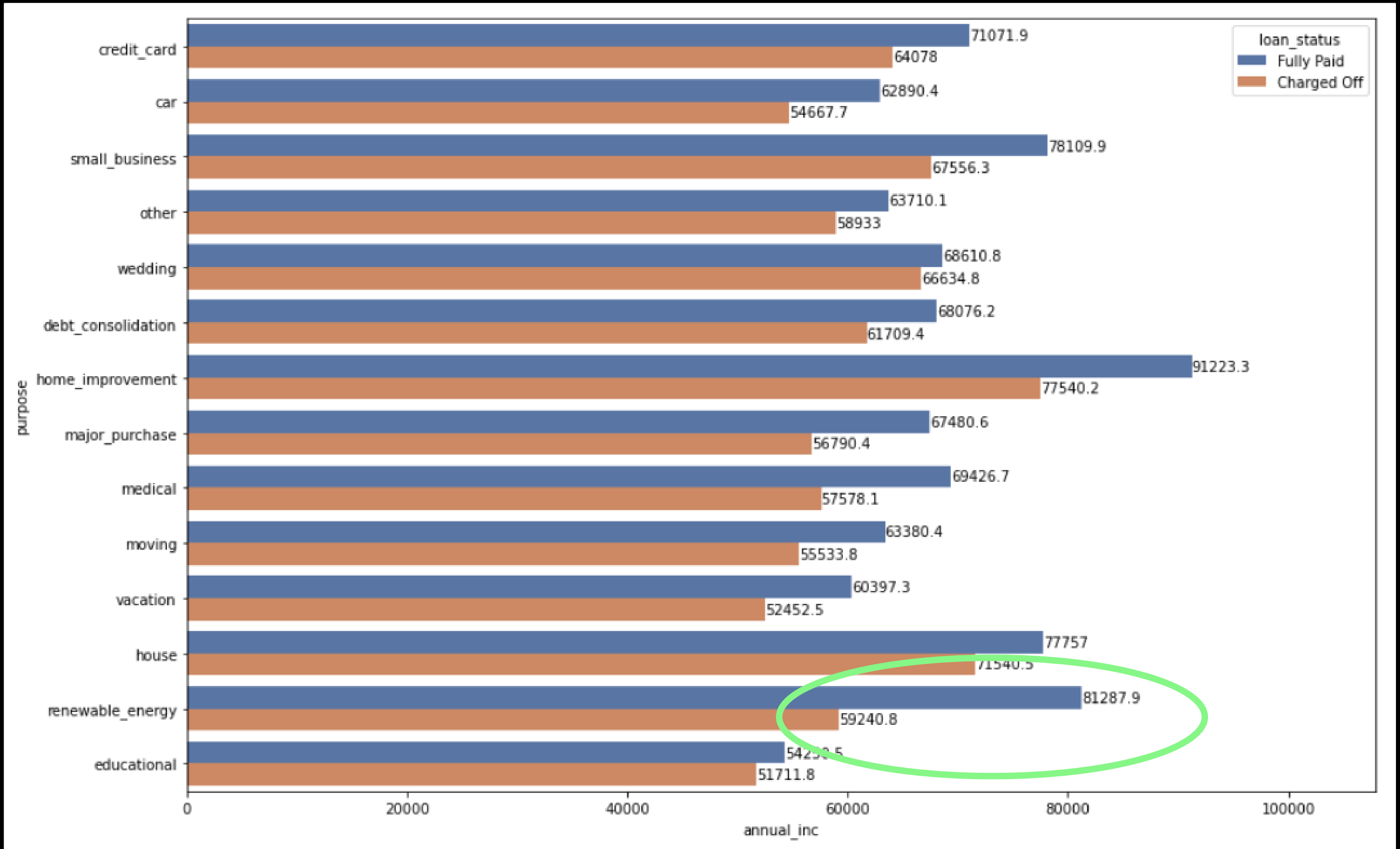
James Jeyabalan / Akik Ranade

Business Findings

- 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.
- Observations
- The above analysis with respect to the charged off loans. There is a more probability of defaulting when :
- Applicants taking loan for 'home improvement' and have income of 60k -70k
- Applicants whose home ownership is 'MORTGAGE and have income of 60-70k
- Applicants who receive interest at the rate of 21-24% and have an income of 70k-80k
- 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
- When grade is F and loan amount is between 15k-20k
- When employment length is 10yrs and loan amount is 12k-14k
- When the loan is verified and loan amount is above 16k
- For grade G and interest rate above 20%

Recommendation :

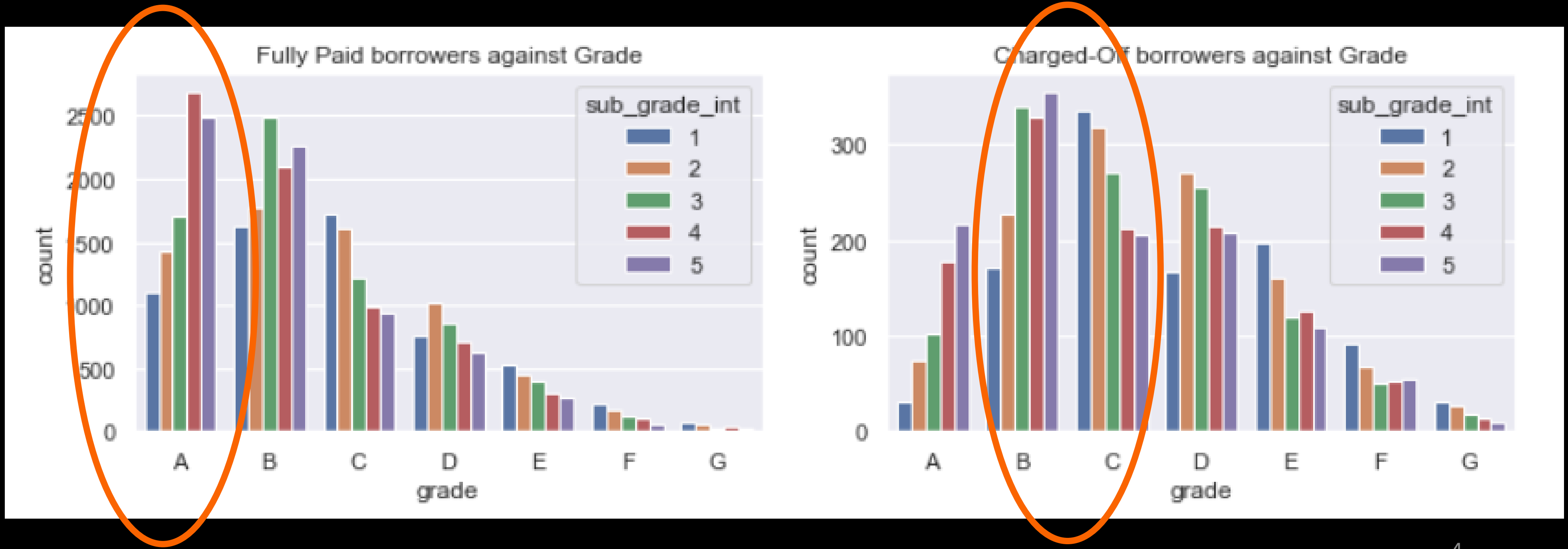
When borrower's annual income is below USD 60K and the purpose is for renewable energy, there is a high chance of default



Recommendation:

Grade 'A5' is only as good as 'B1' – Default percentage raises in Subgrade5 (Grade A)

- *"Fully Paid" Category has borrowers in grade A & B predominantly*
- *"Charged Off" Category borrowers shift towards grade B & C*



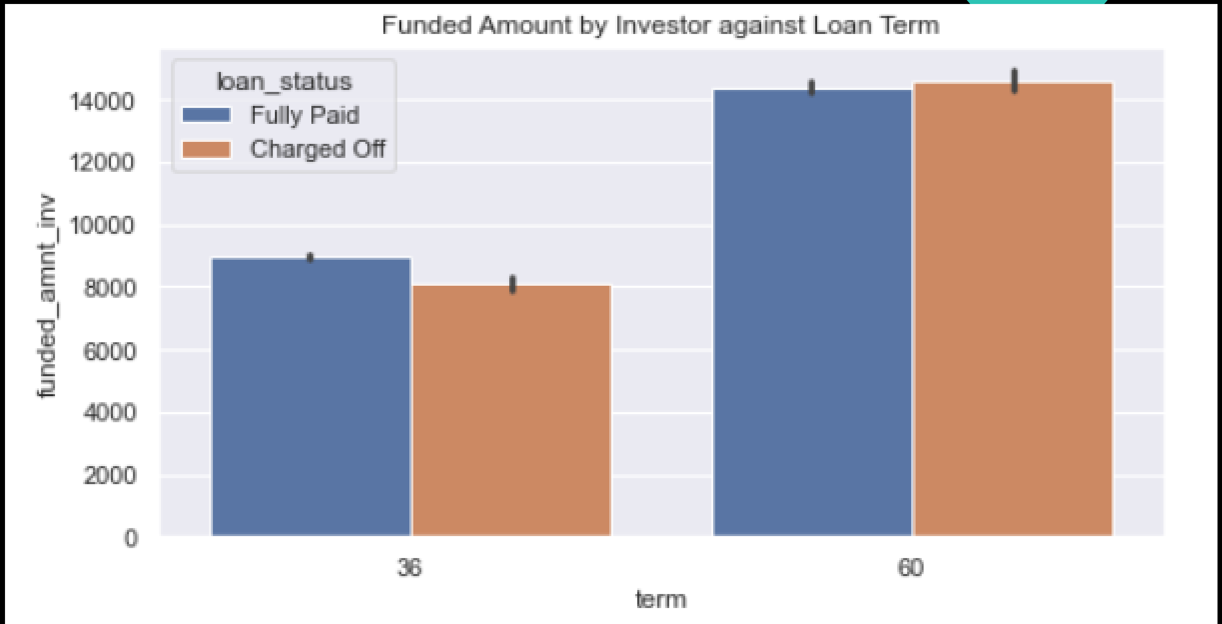
Recommendation:

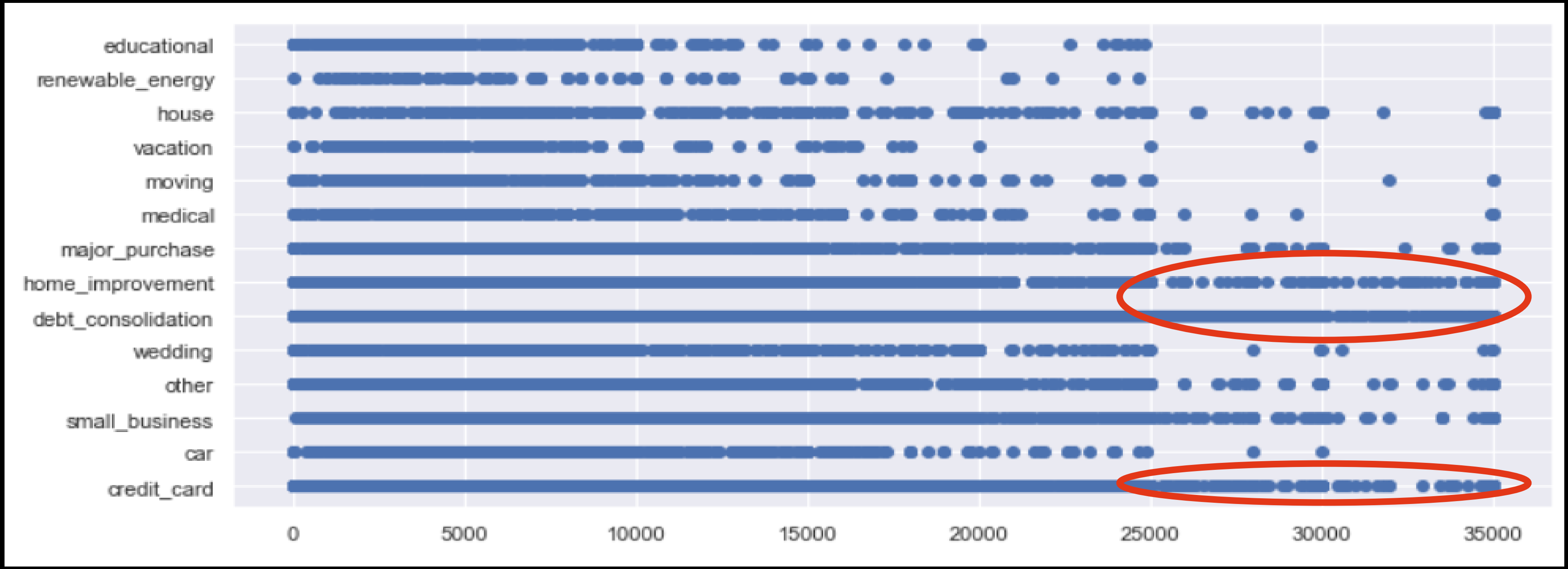
Wedding and Medical loans have higher default rate
Vacation, Car and Moving loans have No defaulters

* *Considering only the loan amounts above 25K USD*

Observation – 1

- *Longer the term, higher the loan amount taken*
- *A large portion of **high** loan amounts are for Debt Consolidation, Credit Card & Home Improvement*
- ***3/4th of the loans** taken are for the above 3 categories*





Observation - 2

- *There are couple of borrower's Annual Income values that are very high but it doesn't skew the mean or median. Hence those values were **not considered as outliers***

```
loan.annual_inc.quantile([0.70, 0.80, 0.90, 0.99, 1])
```

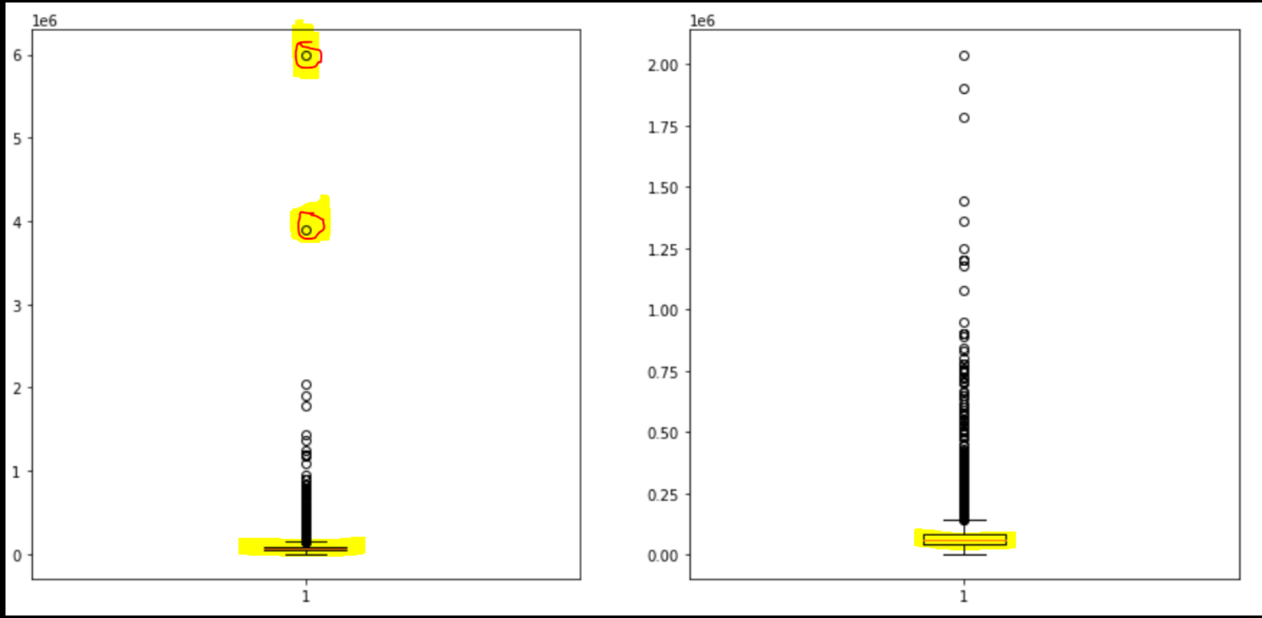
```
0.70      75000.0
0.80      90000.0
0.90     115000.0
0.99     234000.0
1.00    6000000.0
Name: annual_inc, dtype: float64
```

```
loan.annual_inc.mean()
```

```
68809.22861110396
```

```
loan[loan['annual_inc'] < 3850000].annual_inc.mean()
```

```
68555.82480726806
```



Observation – 3

- *There is no correlation between Borrower's Employment period & Grade* *(Financial Rating)*

```
corr = loan.emp_length.corr(loan.grade_int)
print(corr)
corr = loan.emp_length.corr(loan.sub_grade_int)
print(corr)
```

```
-0.009659424947207684
-0.016887063002591185
```

Observation – 4

- *Considering a subset where the investor hasn't funded the loan, there are borrowers marked as "delinquent in the last 2 years"*
- *These borrowers where probably funded by Lending Club*

    12% had defaulted (18 out of 148) in this "Lending Club Funded" category

Assumption: *Since the MemberID is unique across the dataset each row is considered as a new borrower.*

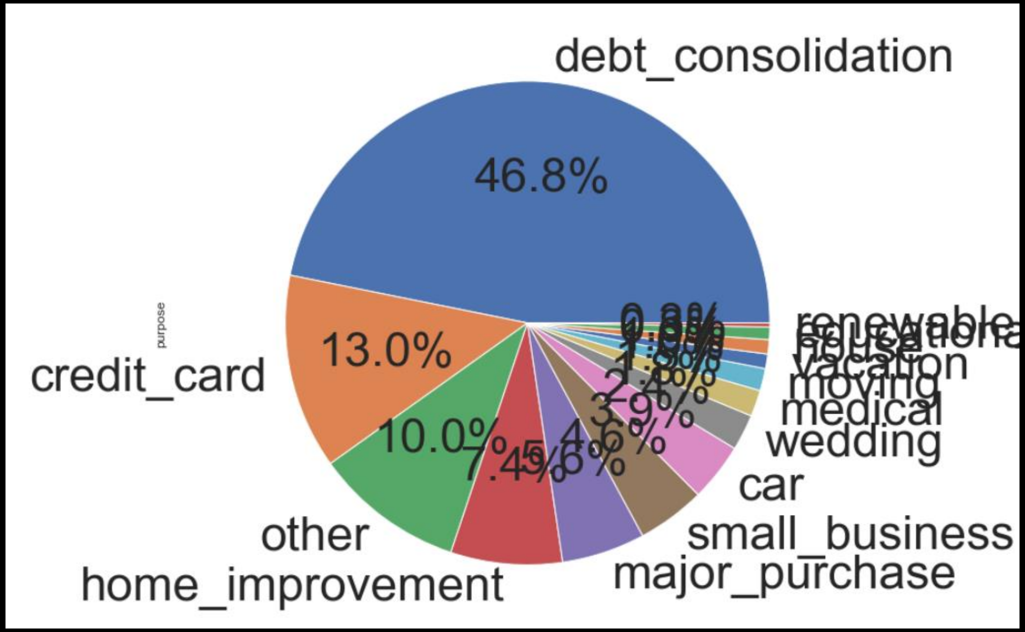
```
loan[loan['funded_amnt_inv_bin']=='Not Funded'].filter(['delinq_2yrs']).value_counts()
```

```
delinq_2yrs
0              130
1               12
2                4
3                2
dtype: int64
```

Observation – 5

- *Distribution of loans by "**Purpose**"*

  *"**Debt Consolidation**" is the single largest loan purpose across "Fully Paid" and "Charged-Off" borrowers*



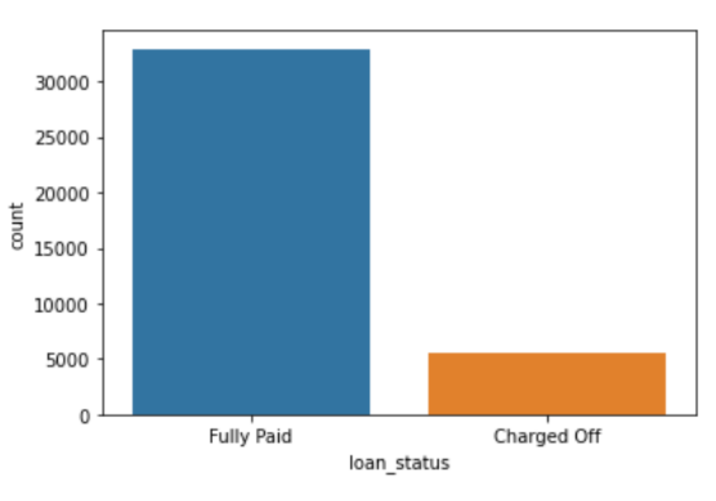
- *Distribution by "**Loan Status**"*

  (after removing 'Current' accounts)

  17% Defaulted loans

```
loan.loan_status.value_counts()

Fully Paid      32916
Charged Off      5611
```



Technical Prep Summary

*Raw data size (Rows: **39717**, Columns: **111**)*

1. *After removing **"NA" columns** (Columns: **57**)*
   - *Effective memory utilization technique as the data set size reduced **50%** (34 Mb to 17 Mb)*
2. *After removing **descriptive columns** (Columns: **50**)*
3. *After removing **uni-value columns** (Columns: **41**)*
4. *After removing >80% **missing columns** (Columns: **38**)*
5. *Eliminating **Current accounts** as they don't authoritatively say whether the borrower will be a defaulter or not. (Rows: **38577**, 38)*
   - *# of Current accounts is 1140, of which only 121 rows show delinquency, and the last delinquency is an average 3 Years*
   - *Hence considering this subset of data has insignificant for any concrete decision and removing the "Current" loan accounts.*
6. *Remove 50 rows with 'NA' value for revol_util column since we don't want to impute this 'Ratio' field (Rows: **38527**, 38)*
7. *Remove %, + **symbols**, String objects which could potentially be Numeric data after removing the unwanted text suffixing the number value*
8. *Add 2 **derived** attributes/**categories** like Year/Month*
9. *Add **derived metrics** like 'Annual_Installment Amount' to 'Annual Income' ratio (i2i) for analysis*
10. *Add fields to dataset by casting AlphaNumeric Codes in 'grade', 'subgrade' columns as "Int" for correlation analysis*
11. *Impute Employment Length of borrower with "mode" value for ~1000 rows which have NaN*
12. *Manual Reassign Home Ownership from one bin to another to reduce the # of categories*

Thank you