GRAMENER CASE STUDY SUBMISSION

GROUP 100:

1. RICHA GAUTAM
2. RISHI VERMA
3. SAGAR MALVIYA
4. SAKEER VALAPPIL

Business Understanding

- The company wants to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default.

Problem statement:

- Company is facing decrease in profitability due to high loan defaults.

What we AIM to achieve:

- Increase profitability of the company, by identifying characteristics of ‘risky’ loan seeking applicants, to be extra cautious while lending loan to such applicants.

Data Exploration

- Total 39718 records with 111 columns
- 54 columns contain all NA values

- Assumptions:

- Analysis has to be done for the characteristics that can be determined before the start of loan
- Columns containing a single value/ all ‘na’ values are not meant for analysis
- Columns containing mostly ‘na’ values (more than 60%) are not meant for analysis

Data Cleaning and manipulation

Observed data cleaning parameters:

- Columns with all NA values.
- Columns with single values.
- Outliers in data.
- Irrelevant columns (columns that come into picture after the loan has been sanctioned, user provided columns like desc, url etc.)

Other issues:

- In-consistency in letter case
- Replacing missing values
- Numeric variables like ‘employment length’ saved as ‘< 1 years’ and ‘10+ years’
- Date treatment
- Created a variable "default_status" with values 'default' and 'non default’
- Created a numeric field 'default_statusN' from 'default_status' in order to do correlation analysis

Data Analysis

- After cleaning the data, the final data frame consists of :

  - 35639 records and 28 columns

- Exploratory Data Analysis techniques used:
  - Univariate
  - Segmented Univariate
  - Bivariate
  - Derived Metric

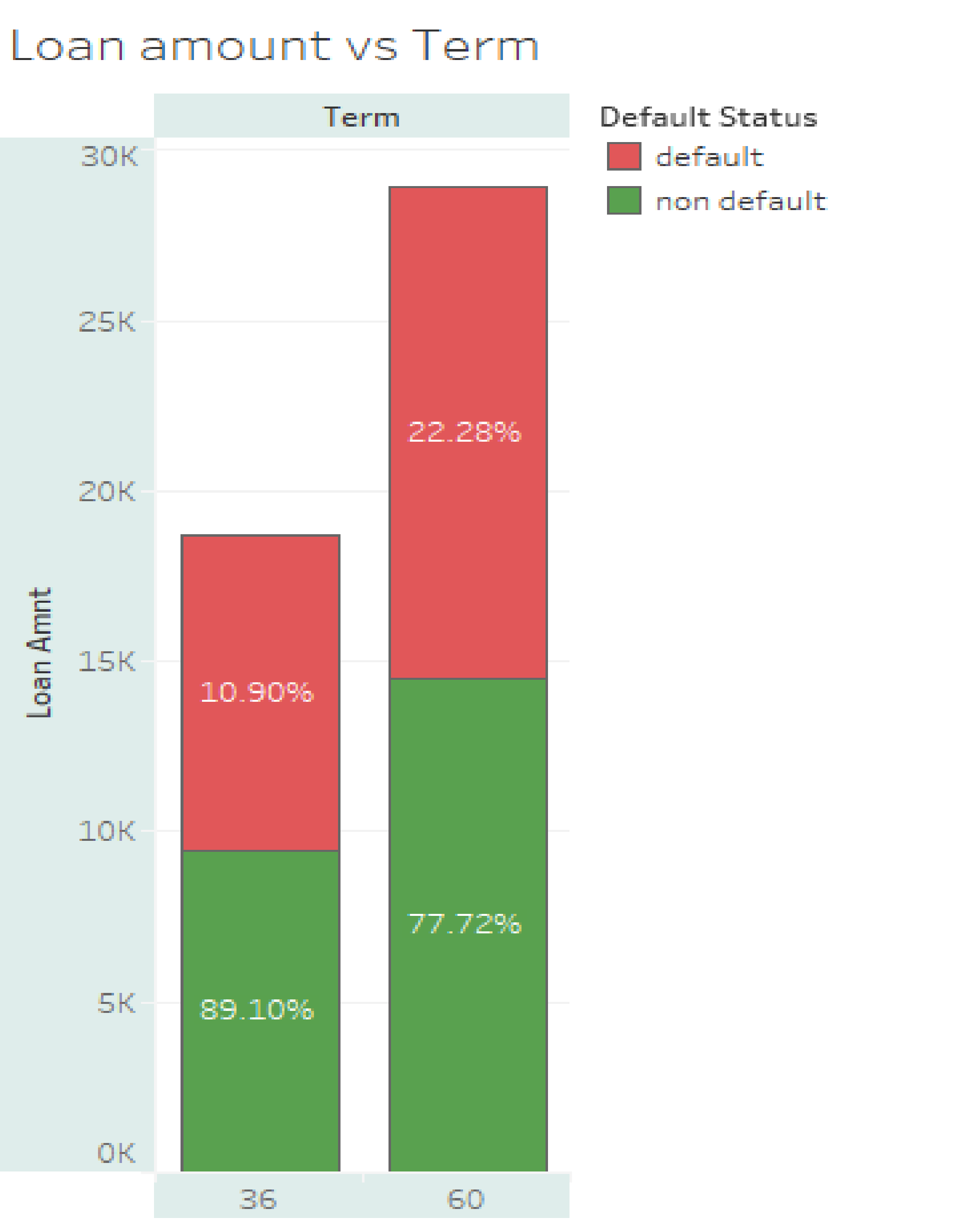
  we came across several variables that impact the default status of an applicant.

#term

The number of payments on the loan. Values are in months and can be either 36 or 60.

- **It is observed that 22% of loans with term 60 months have been defaulted whereas only 11% defaulted in case of term 36 months**

Term	Chances of Default
36 months	10.90%
60 months	22.28%

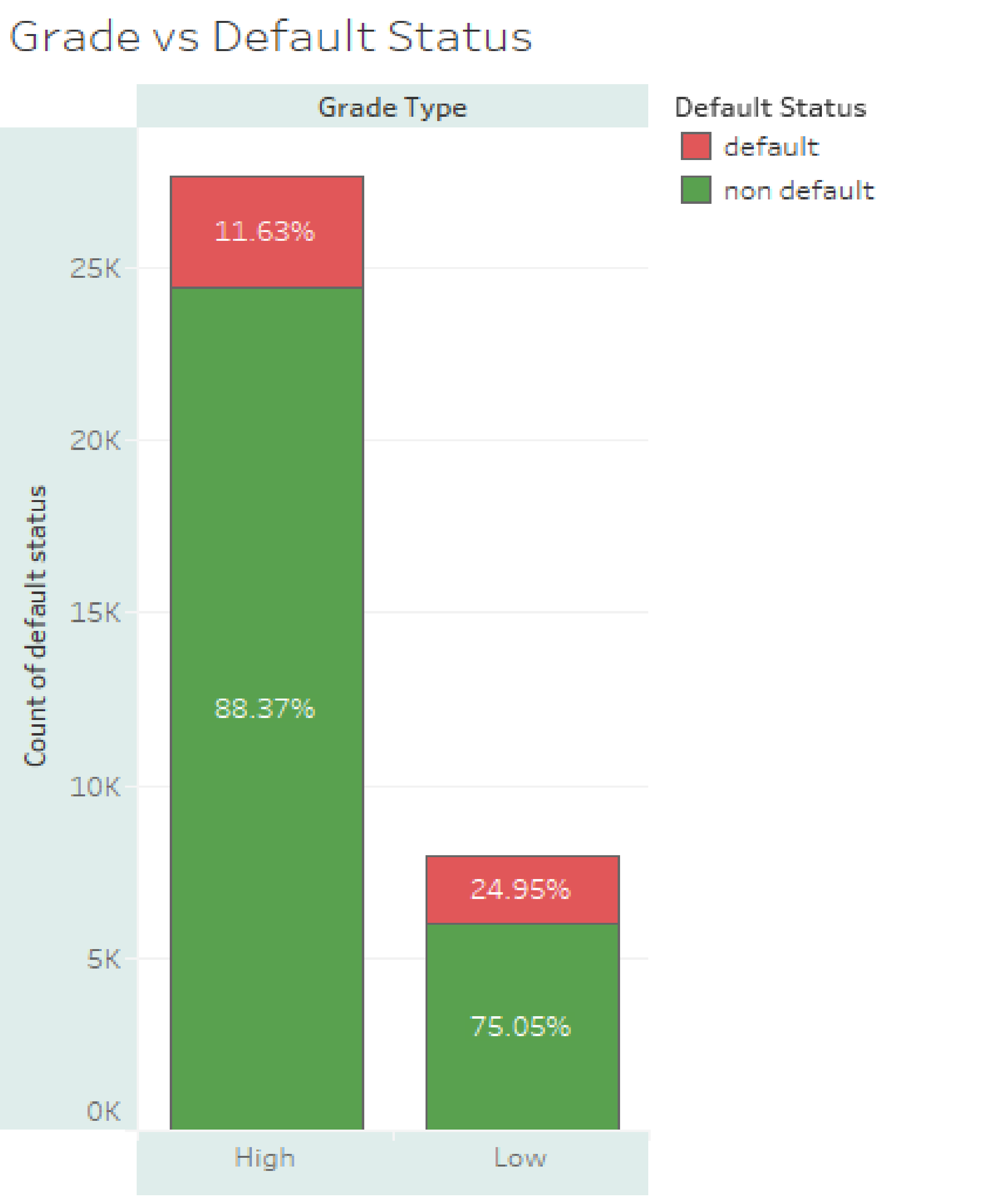


Average of Loan Amnt for each Term. Color shows details about Default Status. The marks are labeled by % of Total Loan Amnt.

#grade

LC assigned loan grade

- **Grades have been binned into two categories: A,B and C in High and D,E,F,G in Low.**
- **It is observed that Higher grades categories have defaulting rates of 13% less default rate than lower categories.**



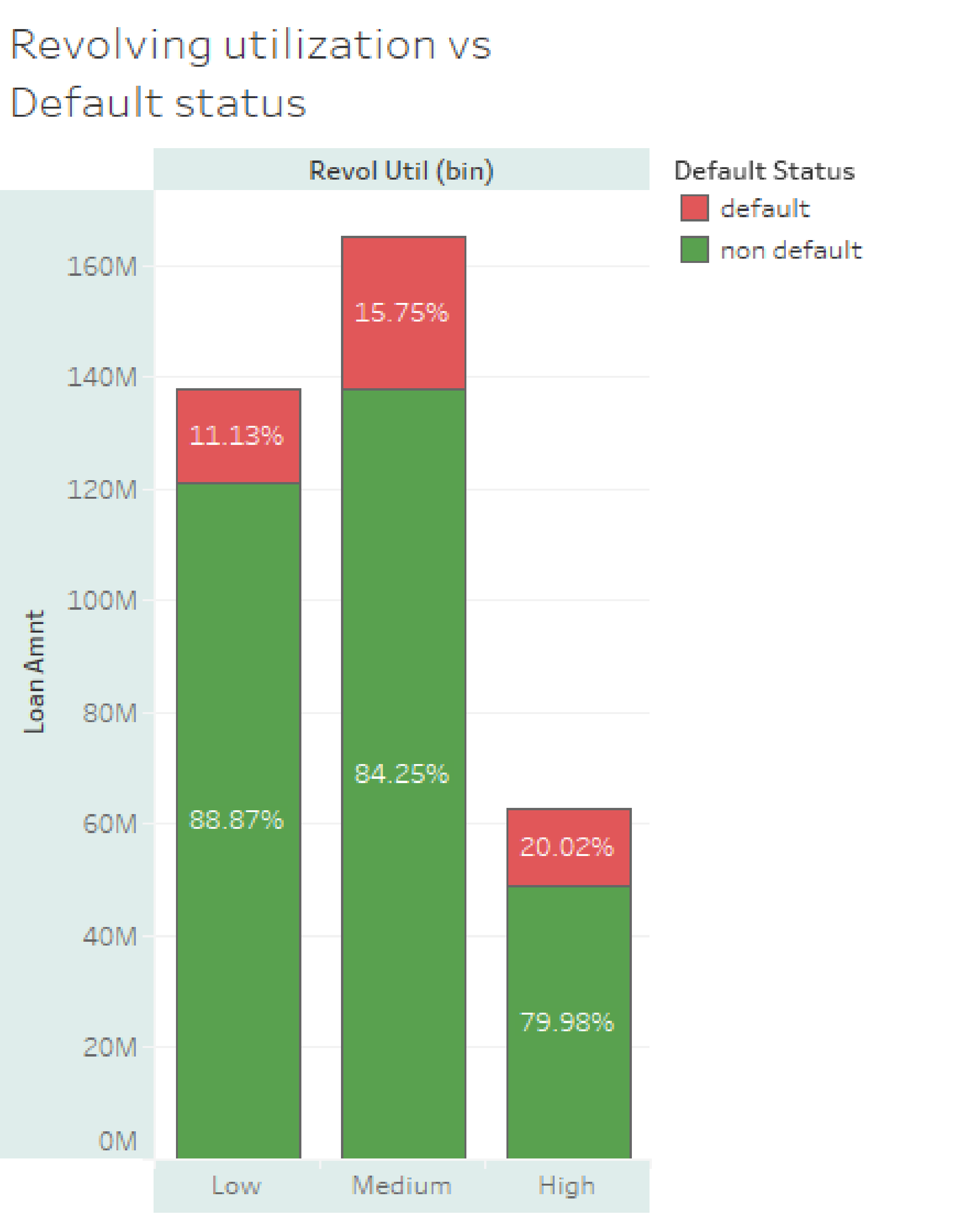
Count of default statusN for each Grade Type. Color shows details about Default Status. The marks are labeled by % of Total Count of default statusN.

#revol_util

Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.

- **If the revolving utilization is more that 25% customers are more likely to default**

Credit Utilization	Chances of Default
Low(0-40)	11%
Medium(40-80)	16%
High(80-100)	20%

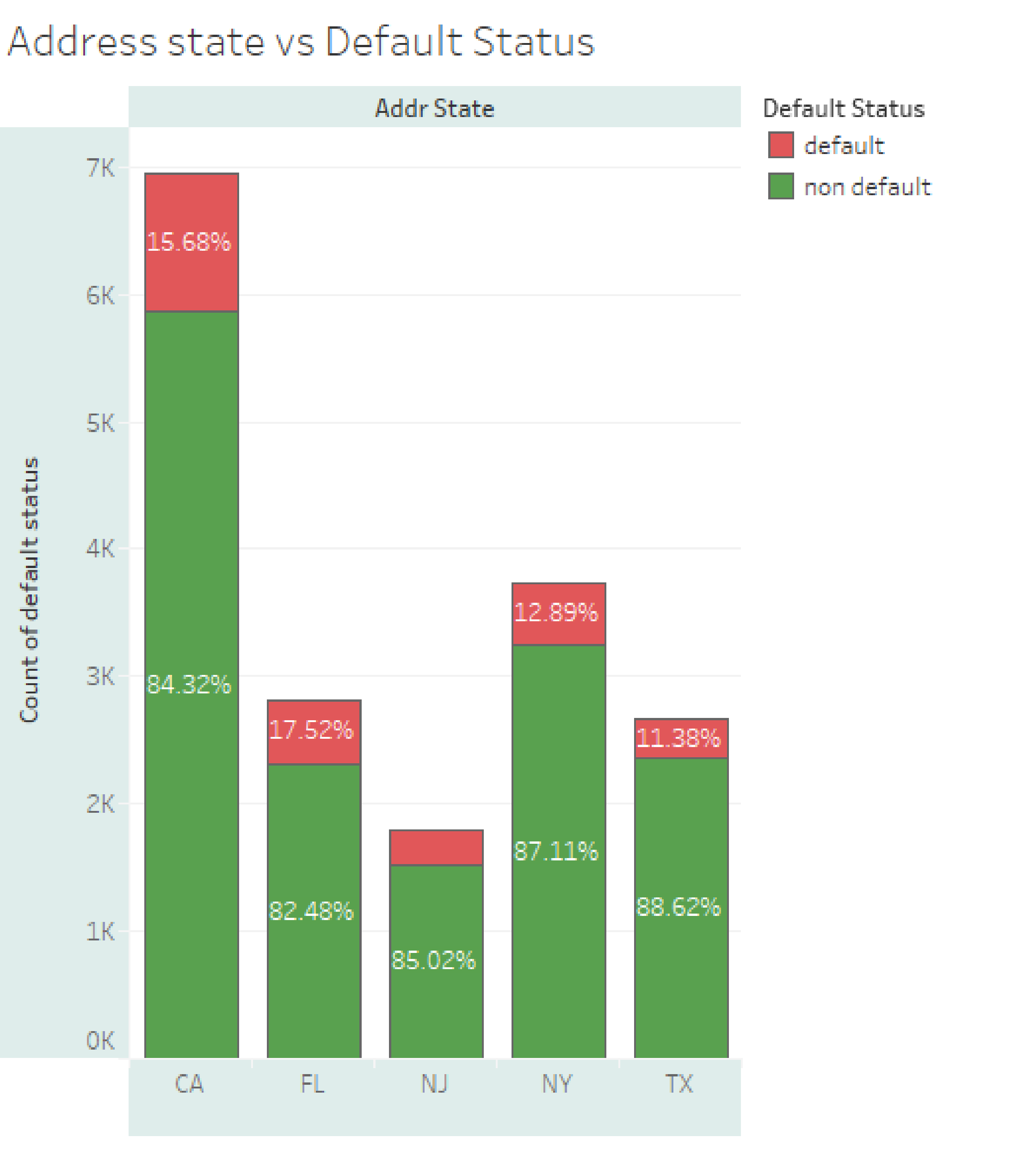


Sum of Loan Amnt for each Revol Util (bin). Color shows details about Default Status. The marks are labeled by % of Total Count of default statusN.

#addr_state

The state provided by the borrower in the loan application

- **It is observed that states like CA, FL, NJ have higher number of defaulters as compared to other states.**
- **Minimum 1500 Loans.**
- **It can be a case that these states does not have very strict rules for loan defaulters**



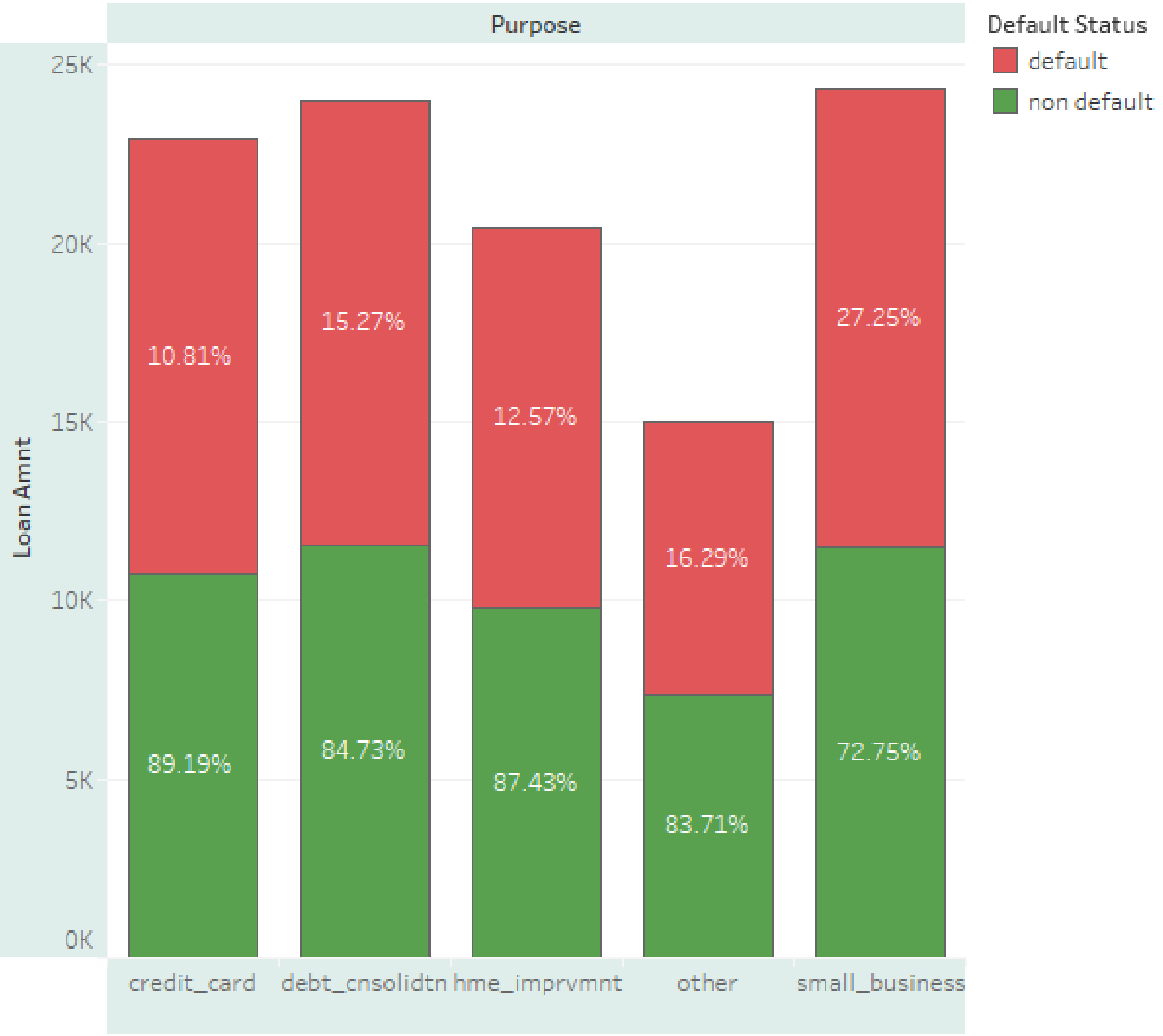
Count of default statusN for each Addr State. Color shows details about Default Status. The marks are labeled by % of Total Count of default statusN. The view is filtered on Addr State, which keeps CA, FL, NJ, NY and TX.

#purpose

A category provided by the borrower for the loan request

- **It is observed that out of the 14 categories, people taking loans for 'Small Business' have defaulted for maximum of 27% of the time.**

Purpose vs Default Status



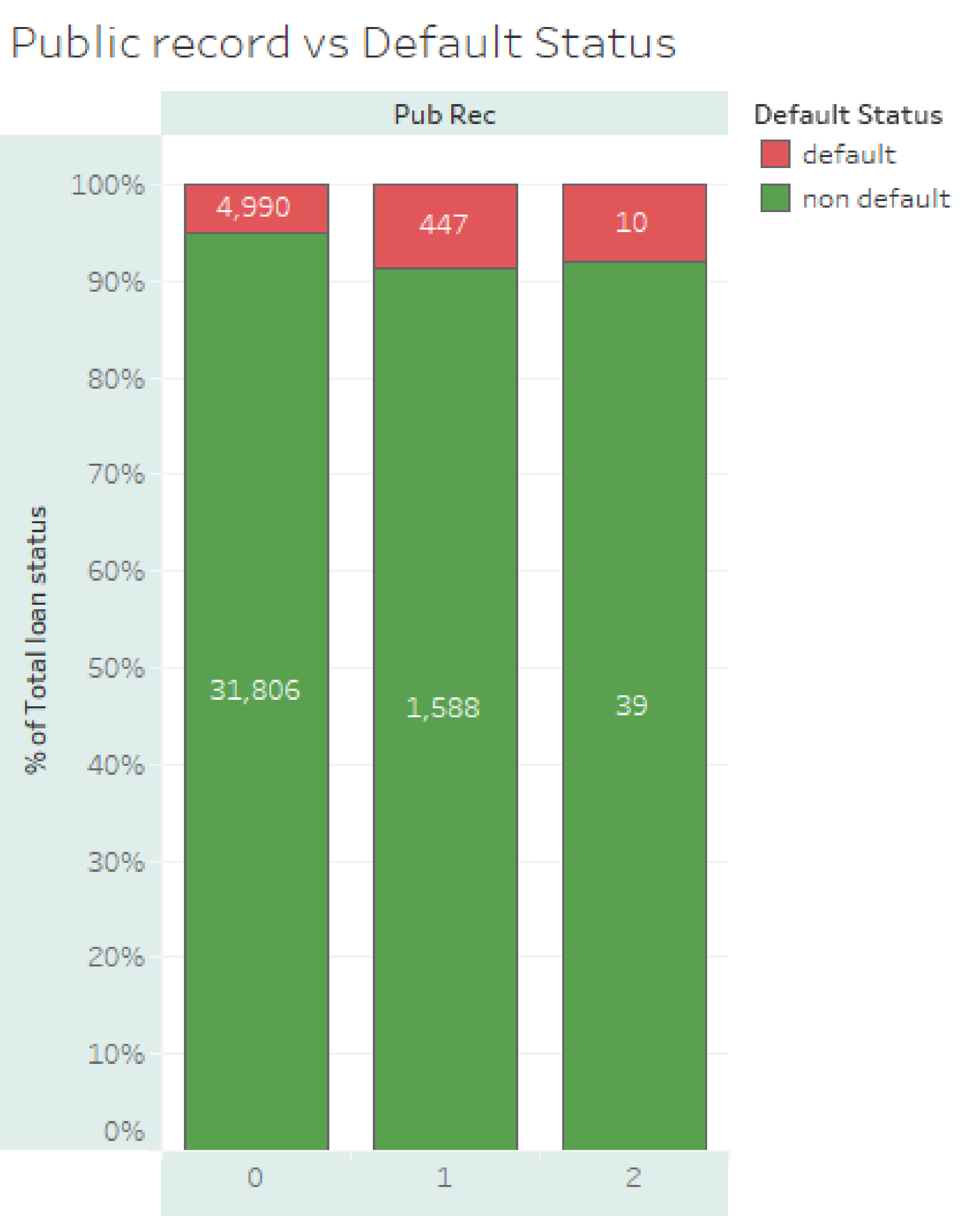
Average of Loan Amnt for each Purpose. Color shows details about Default Status. The marks are labeled by % of Total Count of default statusN. The view is filtered on Purpose, which keeps credit_card, debt_cnsolidtn, hme_imprvmnt, other and small_business.

#pub_rec

Number of derogatory public records

- **It is observed that if pub_rec is greater than 0, default rate is higher**

Pub_rec	Count of Default
0	4990
1	447
2	10

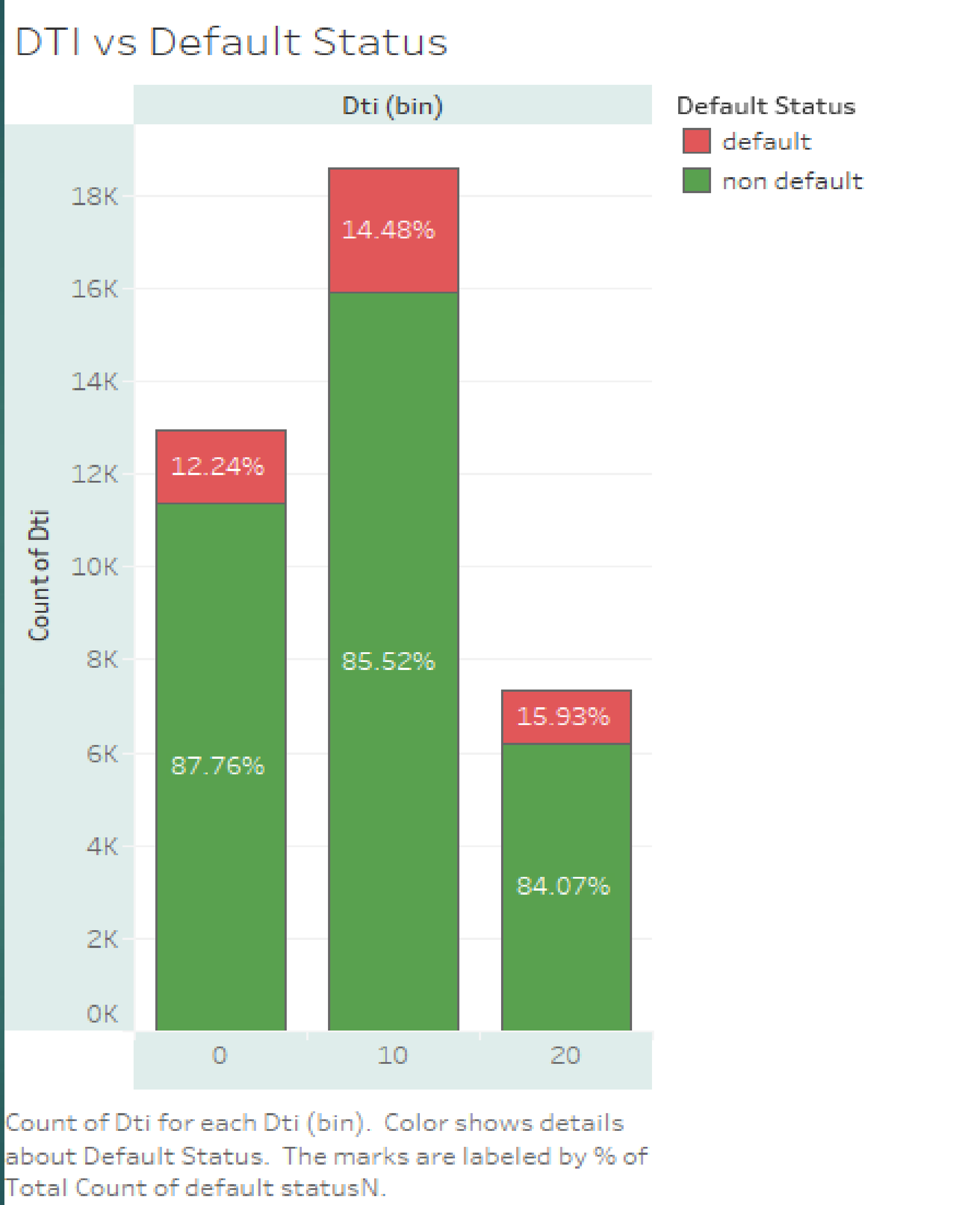


% of Total loan statusN for each Pub Rec. Color shows details about Default Status. The marks are labeled by count of loan statusN. The view is filtered on Pub Rec, which keeps 0, 1 and 2.

#dti

A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, divided by the borrower's self-reported monthly income

- It is observed that 12.4% people defaulted where dti is less than or equal to 12, whereas the default %age shoots up to 15.93% when dti is greater than 12.
- This clearly shows that higher dti means higher chances of defaulting



Conclusions

Below is the priority order list of the variables that drive the number of defaulters for LC:

1. term
2. grade
3. revol_util
4. addr_state
5. purpose
6. pub_rec
7. dti

It is advised that LC must use these variable to filter out the applications that can be future loan defaulters

Thankyou