GRAMENER CASE STUDY

PROBLEM STATEMENT

- There are 2 types of risks associated with a company's decision to give/not to give the loan:

  - If the applicant is **likely to repay the loan**, then not approving the loan results in a **loss of business** to the company
  - If the applicant is **not likely to repay the loan,** i.e. he/she is likely to default, then approving the loan may lead to a **financial loss** for the company

- In this analysis, we use EDA to understand how **consumer attributes** and **loan attributes** influence the tendency of default.

ANALYSIS APPROACH

- We employ various EDA techniques like Univariate, Segmented Univariate and Bivariate analysis to find the most important driver variables and combinations of those that indicate the tendency of a loan to default.

- Univariate Analysis to understand the properties of each variable.

- Segmented Univariate Analysis to understand the impact of the loan_status variable on the others.

- Bivariate Analysis is to identify patterns that lead to loan default by combining the driver variables.

TARGET DATA

- Data pertaining to only the loan statuses ‘Fully Paid’ and ‘Charged Off’ are considered for the analysis.
- The following types of variables are not considered in the analysis:
  - Variables with all null values.
  - Variables with only single value.
  - Variables with all different values.
  - Variables with significant null values.
  - Variables that are related to date/month/time.
  - Variables that are related to payment receiving.
  - Variables related to funding and installment.
  - Variables Zip Code, Sub Grade, Employer Title, Title of Loan, Total Credit Lines Open.

DATA UNDERSTANDING/STANDARDIZATION/CLEANUP

- A general understanding of data has been attempted like statistical info, % of missing values in rows & columns, data types, unique values and their frequencies etc.

- Wherever necessary, variables are manipulated to bring them to a standard format.

- Missing value imputation has been done by taking the IQR as the missing value.

- Presence of duplicate records have been checked and removed if found any.

CHOICE OF PLOTS

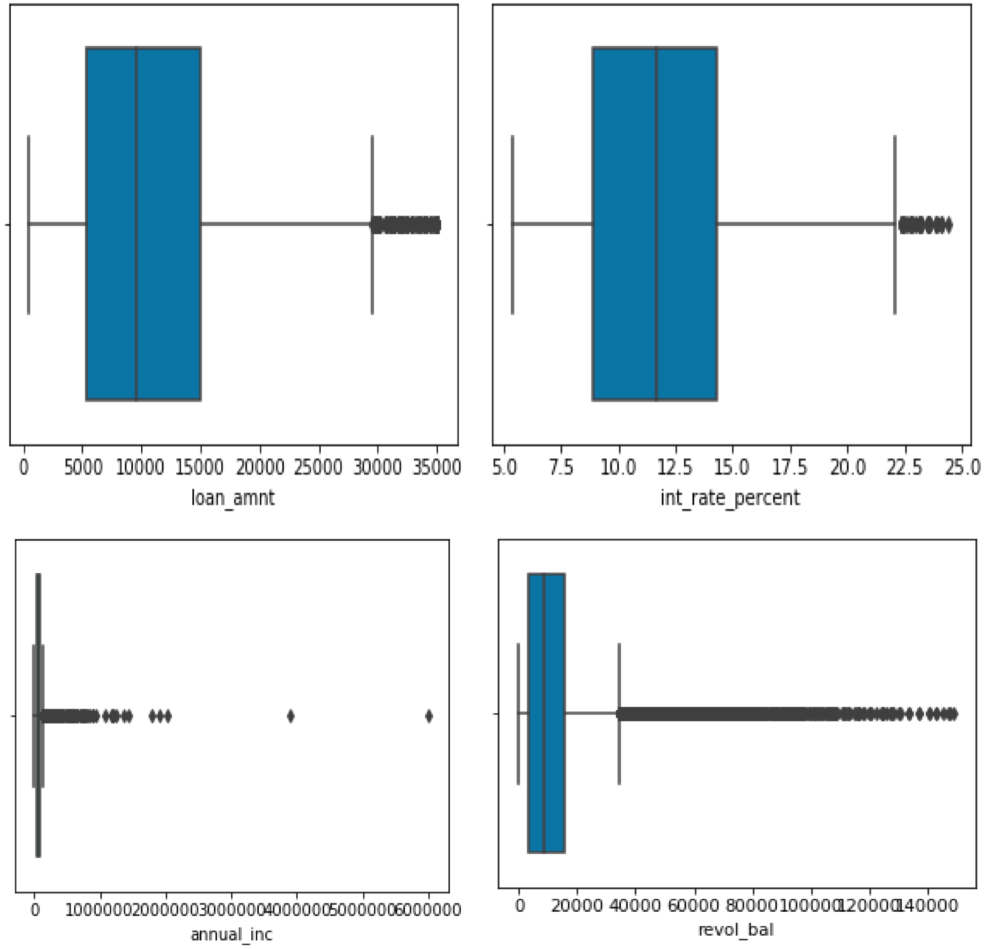
- Univariate Analysis
  - Quantitative Variable => Box plot to see for outliers.
  - Unordered Categorical Variable => Count plot to see the frequency distribution.
  - Numerical Categorical Variable => Count plot & Distribution plot to see the frequency/density distribution.

- Segmented Univariate Analysis
  - 1 Categorical + 1 Quantitative Variable => Bar plots to see the mean and median values across the loan statuses.
  - 1 Categorical + 1 Numeric Categorical Variable => Count plots to see the frequency distribution across the loan statuses.
  - 2 Categorical => Bar plots with count/percent distributions to see the effect of loan status on the others.

- Bivariate Analysis
  - 3 Categorical Variables => Bar plots with count/percent distributions to see the effect one variable on the other.

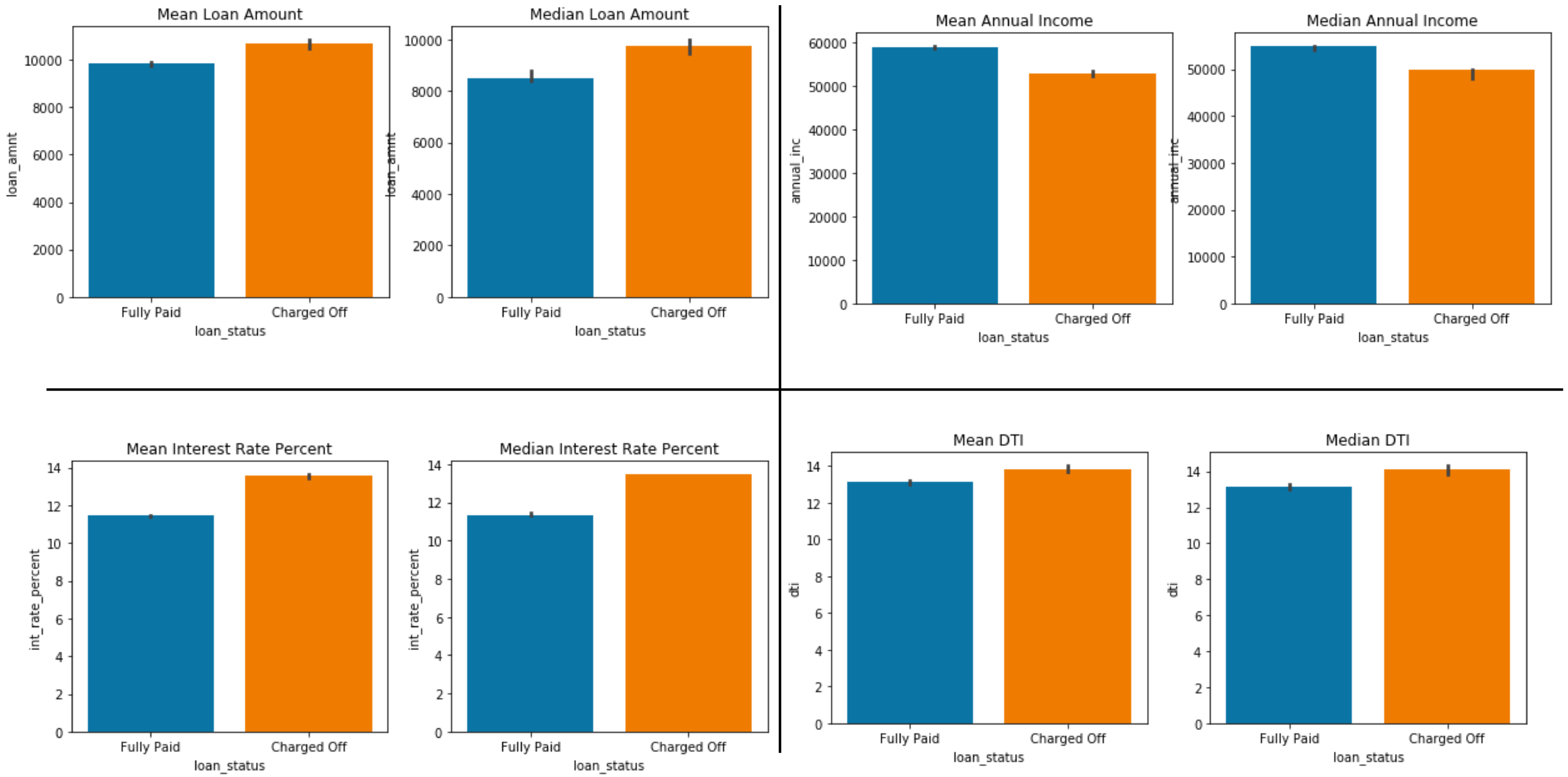
UNIVARITE/SEGMENTED UNIVARIATE ANALYSIS

OUTLIER TREATMENT



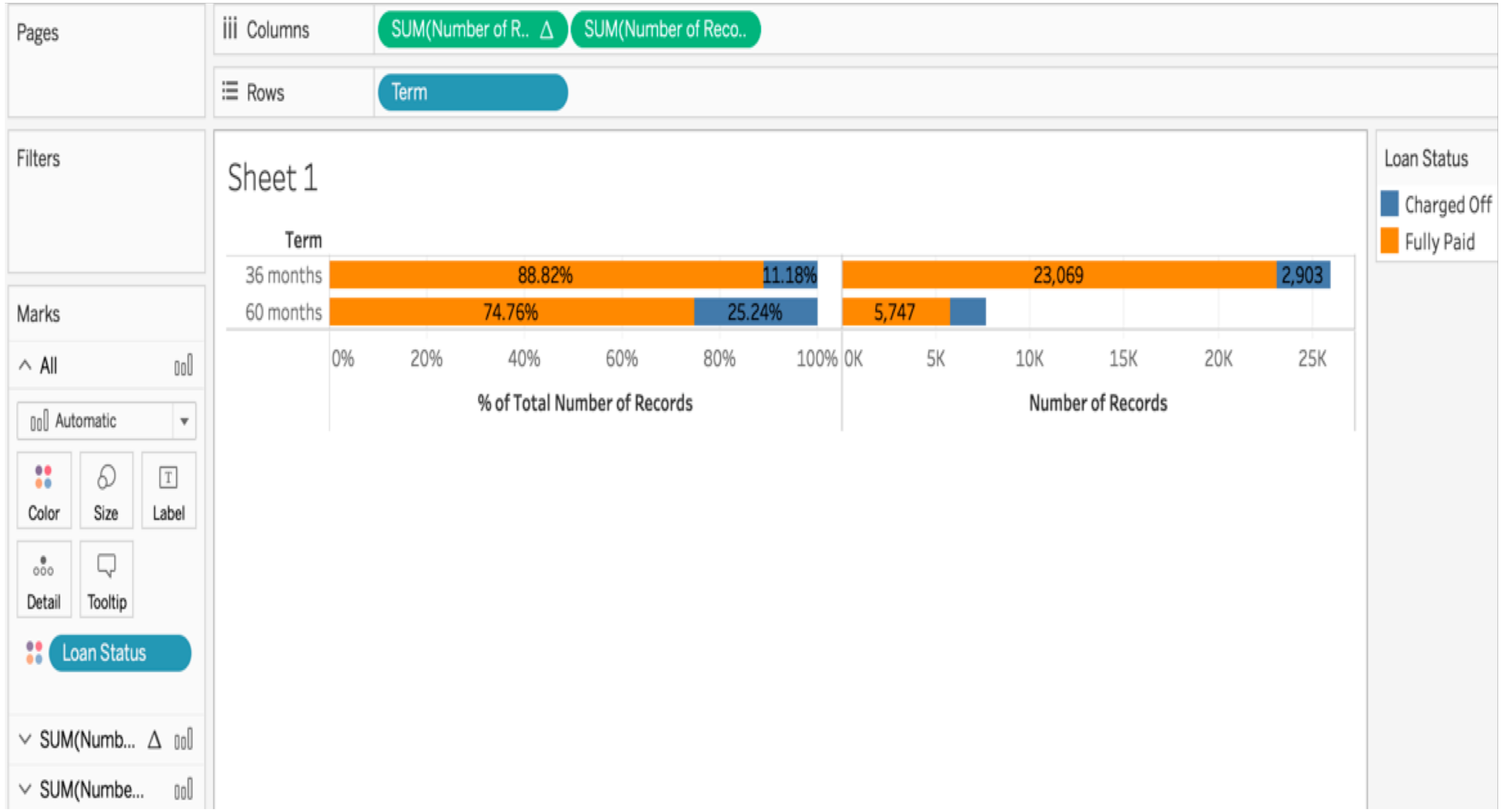
- During univariate analysis, using box plots, outliers are identified in loan_amount, int_rate_percent, annual_inc and revolv_bal columns.
- 1.5 x IQR method has been employed to deal with the outliers:
  - Data below Q1-1.5xIQR and data above Q3+1.5xIQR has been discarded.

LOAN AMOUNT, INTEREST RATE, ANNUAL INCOME & DTI



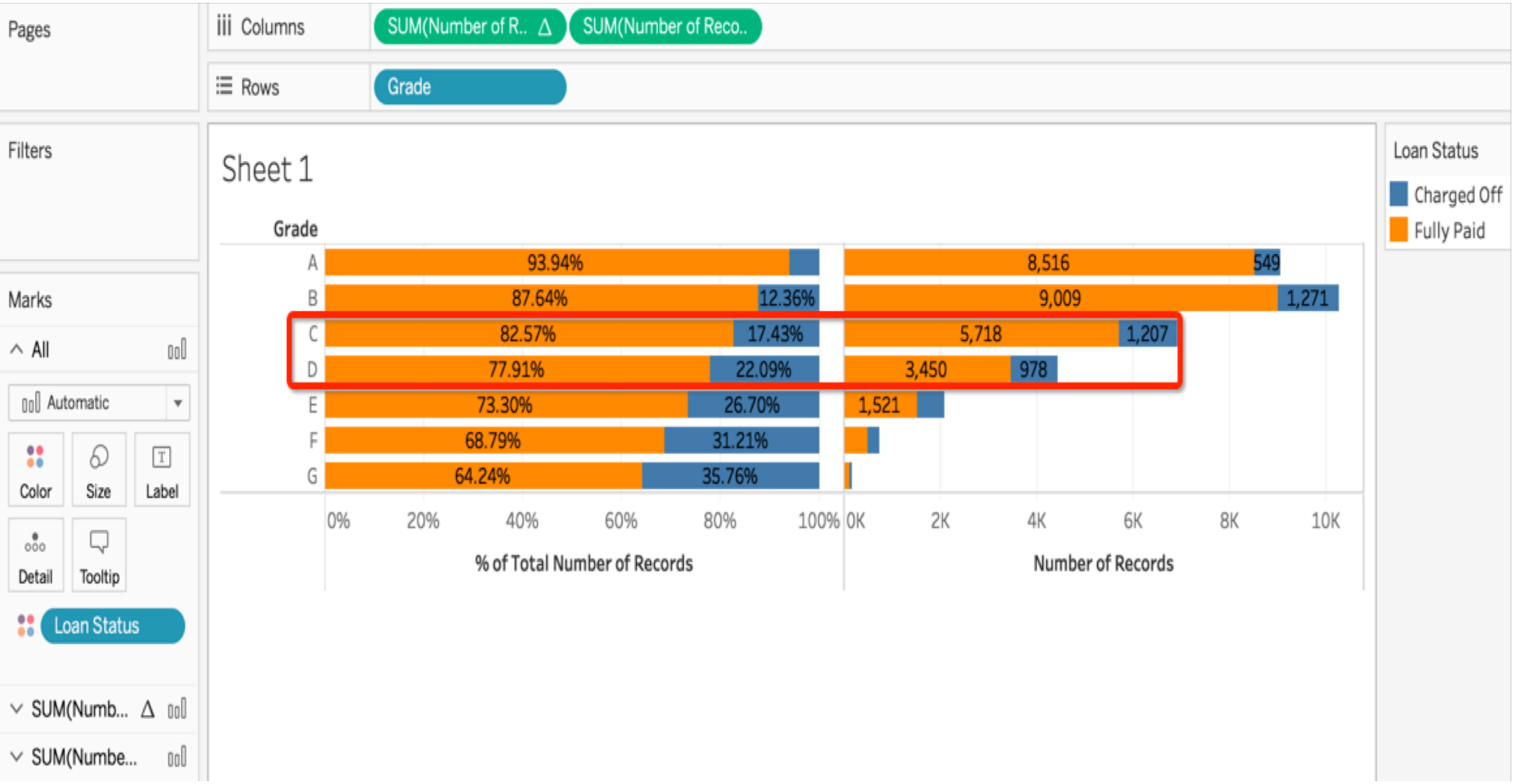
- The mean and median values of these variables for each loan status don't differ much.
- We can't conclude the significance of these using this observation.

LOAN TERM



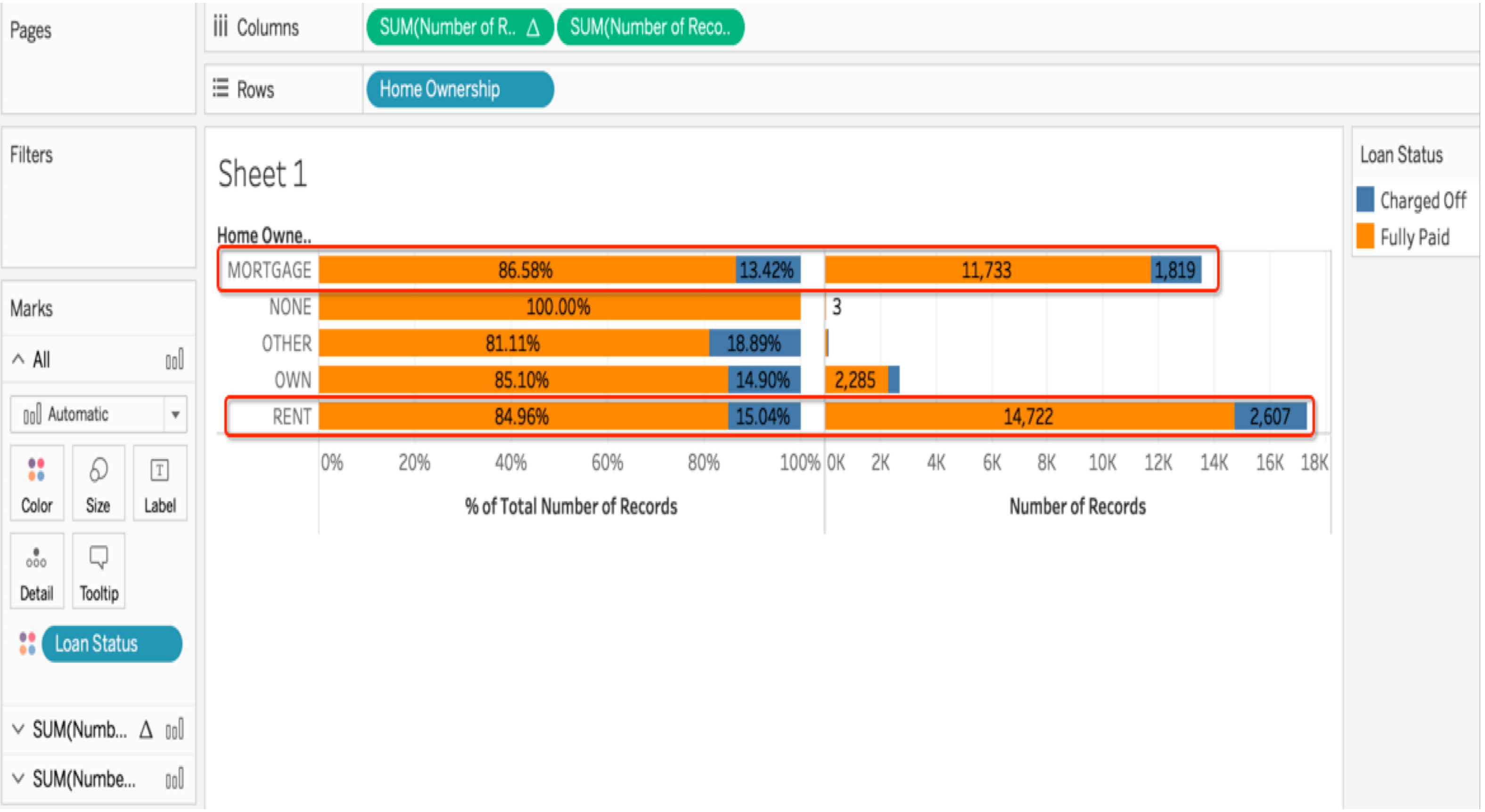
60 months term has the right combination of proportion & count and hence has higher chance of default.

BORROWER GRADE



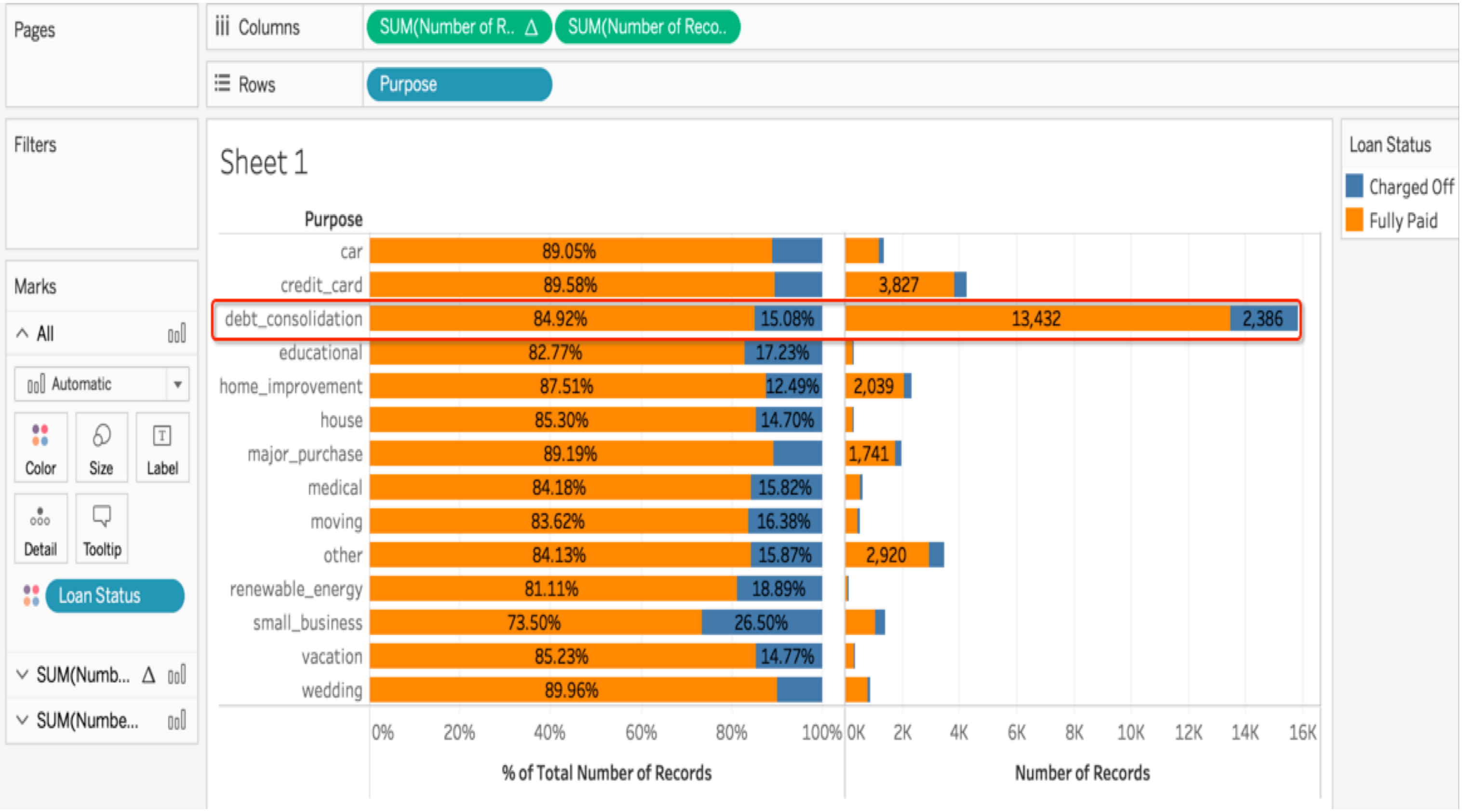
Grades C & D has the right combination of proportion & count and hence has higher chance of default.

BORROWER HOME OWNERSHIP



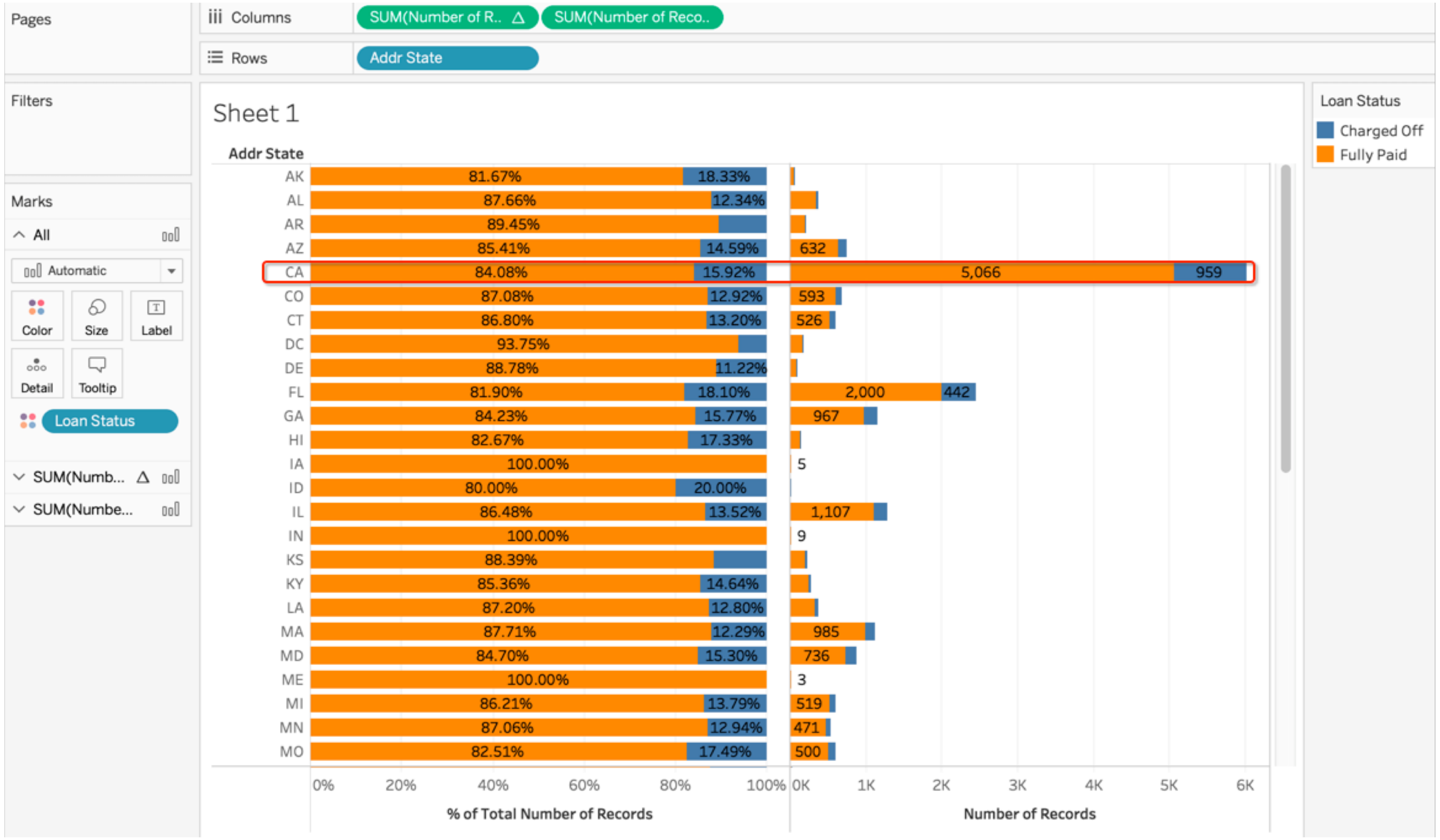
Borrowers whose are currently living in a rented accommodation or their house is on mortgage are more likely to default.

PURPOSE OF LOAN



Loans taken for the purpose of debt_consolidation have higher chance of default.

BORROWER ORIGIN STATE



Loans by borrowers from the state 'CA' have higher chance of default.

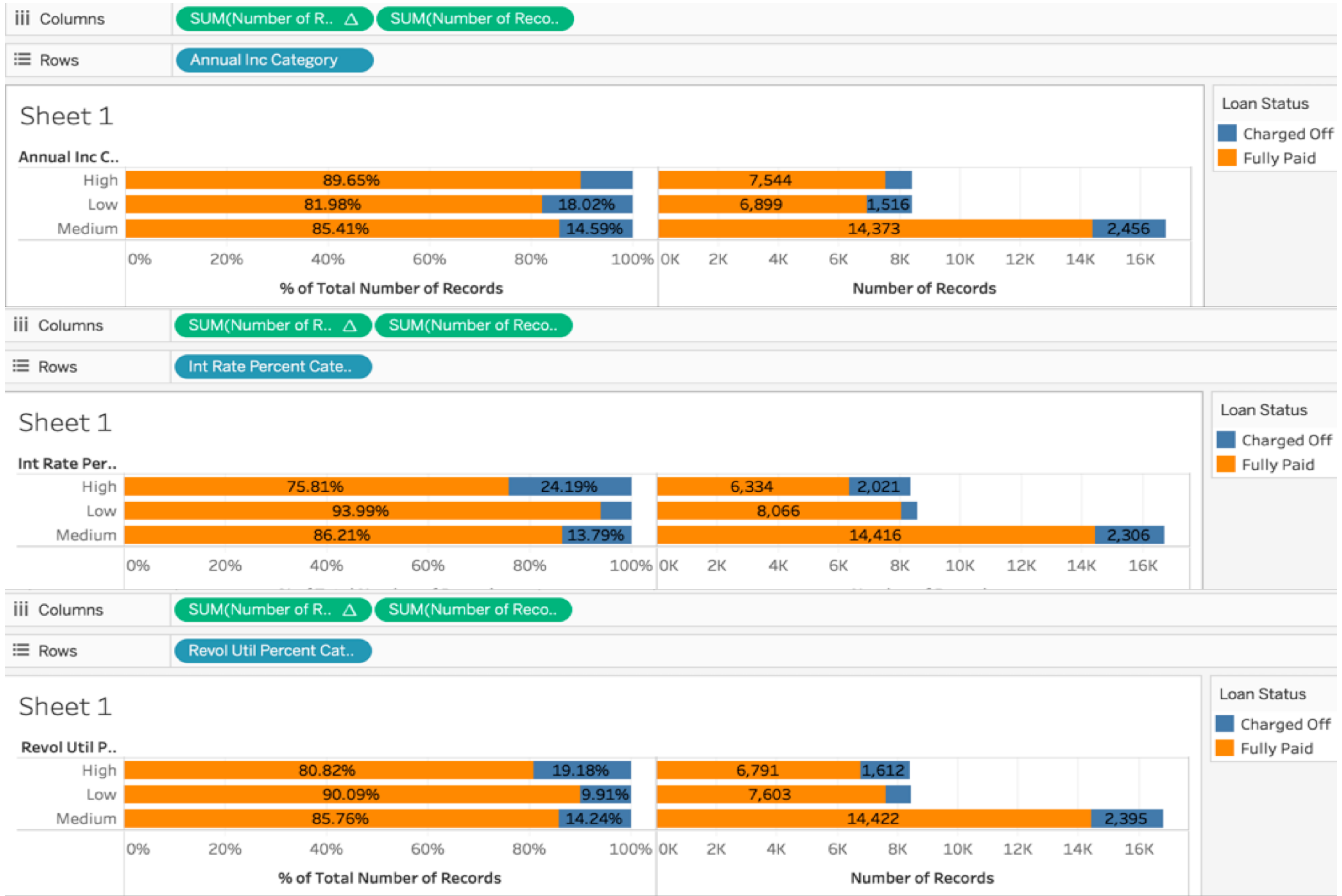
DERIVED METRICS

- The variables annual_inc, int_rate_percent and revol_util_percent all have outliers even after treating them and also are widely spread. It makes sense to categorize them into low, medium and high and then try to analyse their effect on loan status.

- 3 bins are created for each variable based on the quantiles 0, 0.25, 0.75 and 1.

- The obtained derived metrics are type driven ordinal variables which follow an order but mathematical difference between categories is not meaningful.

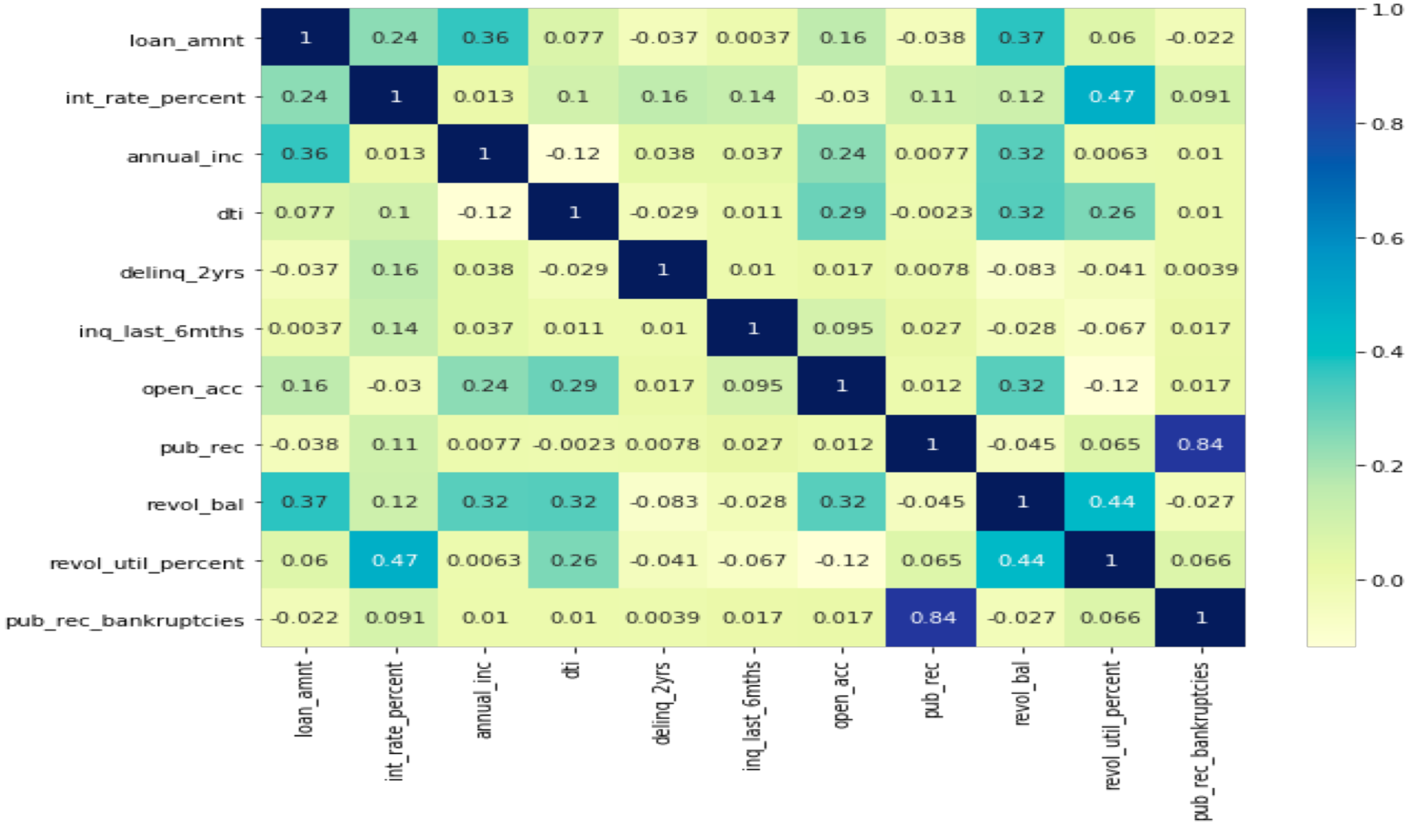
DERIVED METRICS ANALYSIS



None of the derived metrics show significant impact on the loan default.

BIVARIATE ANALYSIS

CORRELATION



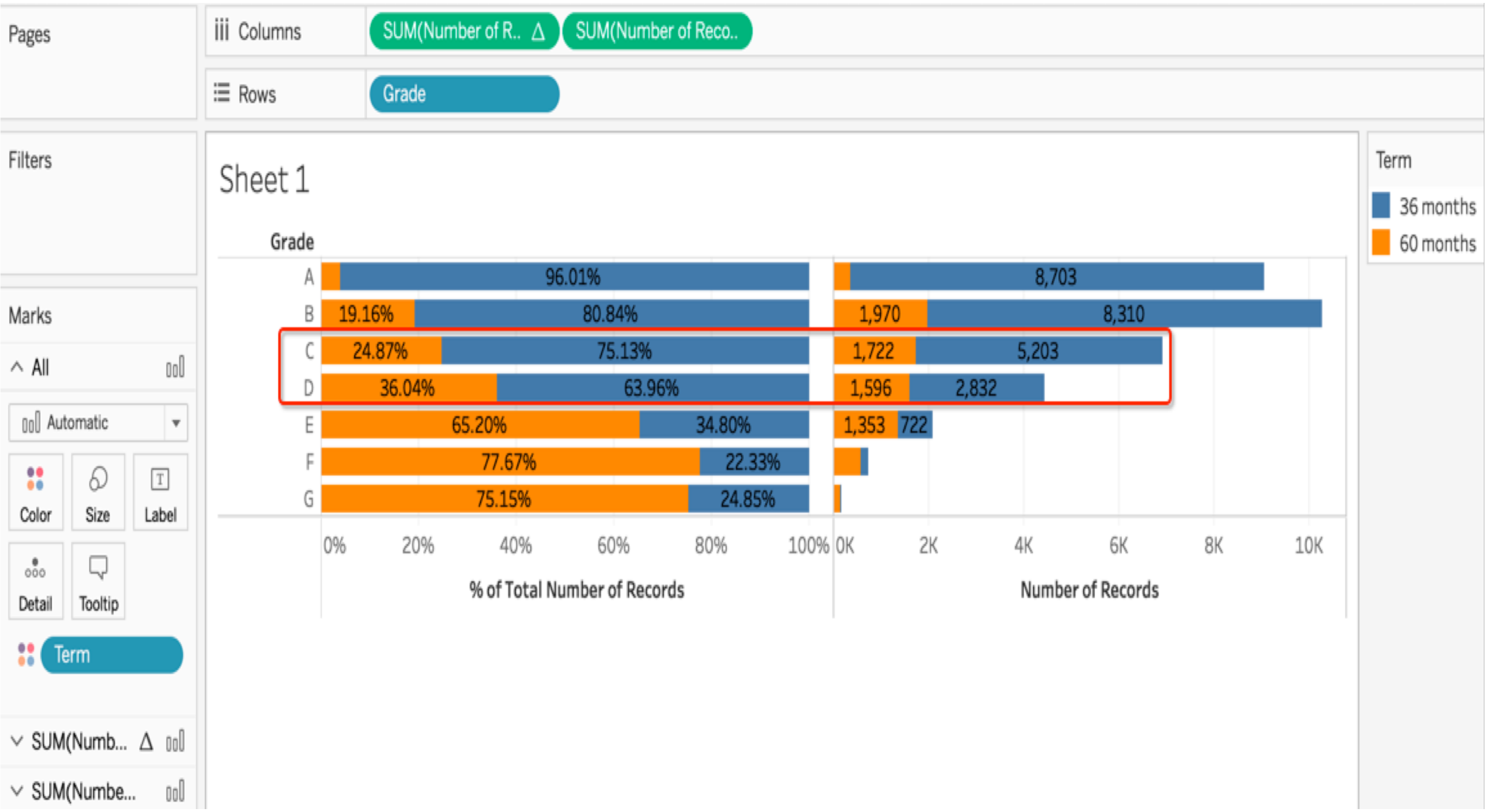
There is no significant correlation between the variables that we are interested in that helps the analysis.

TERM OF THE LOAN BY BORROWER GRADE 1

Loans with term of 60 months borrowed by individuals with grades C or D are more likely to default:

term	grade	% of deault
60 Months	C	24.8
	D	28.57

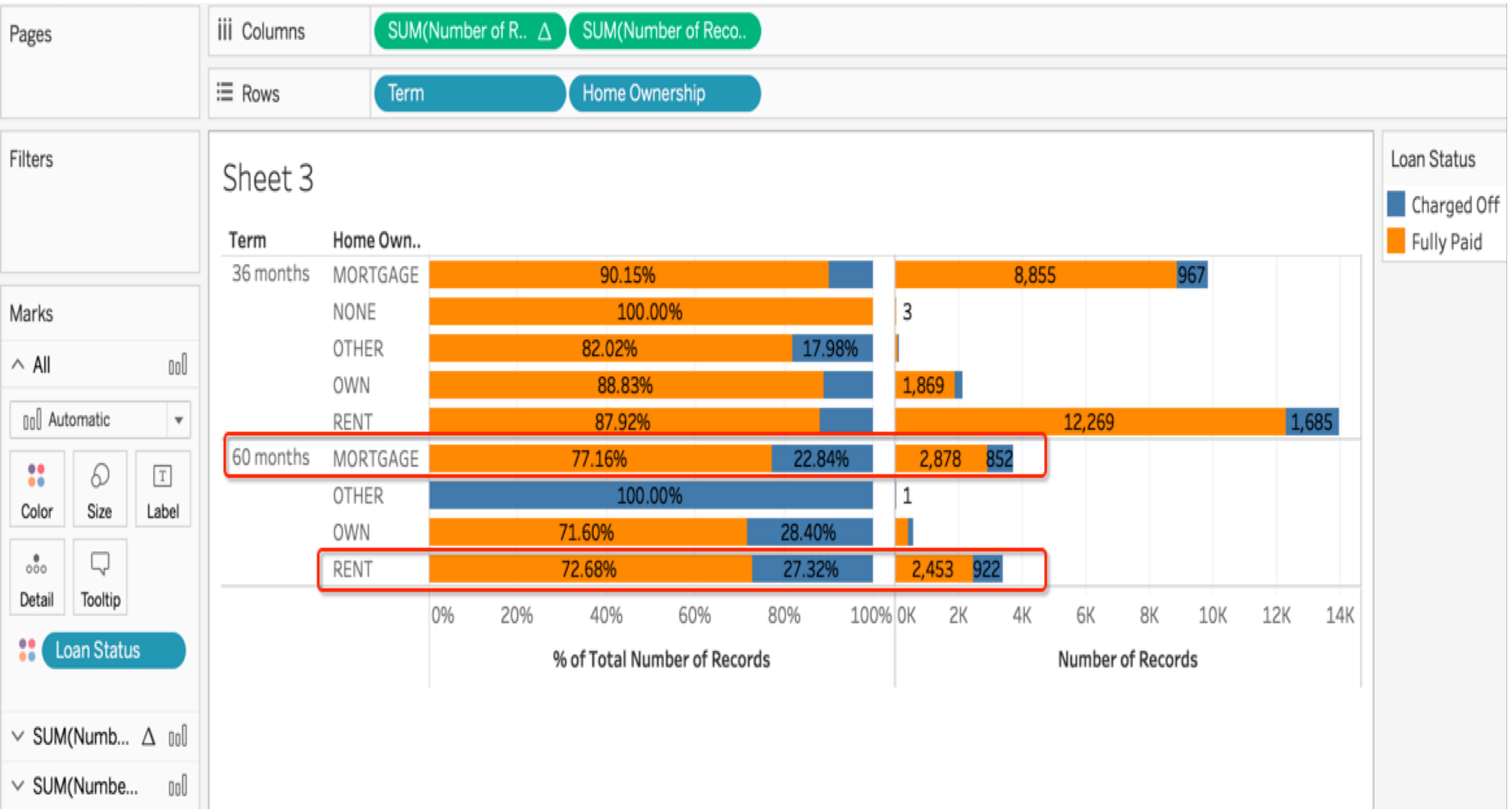
TERM OF THE LOAN BY BORROWER GRADE 2



Borrowers of grades C & D have substantiate percentages of 60 months loans.

There is an increase in taking 60 months loans as we move from grades A to G. Meaning, lower grade people taking loans of higher tenure.

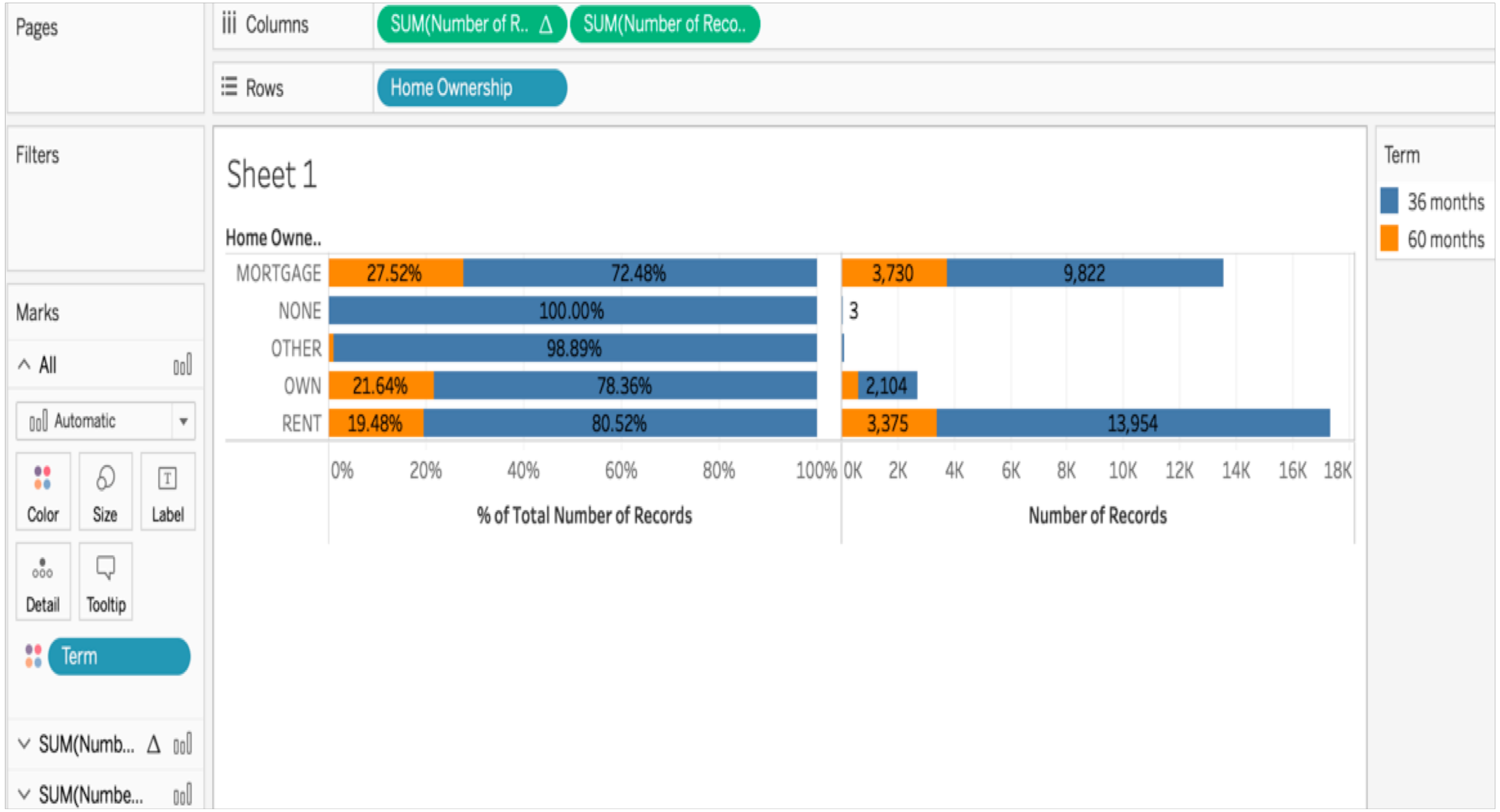
LOAN TERM BY BORROWER HOUSE OWNERSHIP 1



Loans with term of 60 months borrowed by individuals who are either living in a rented accommodation or whose house is in mortgage are more likely to default:

term	home_ownership	% of deault
60 Months	RENT	27.32
	MORTGAGE	22.48

LOAN TERM BY BORROWER HOUSE OWNERSHIP 2



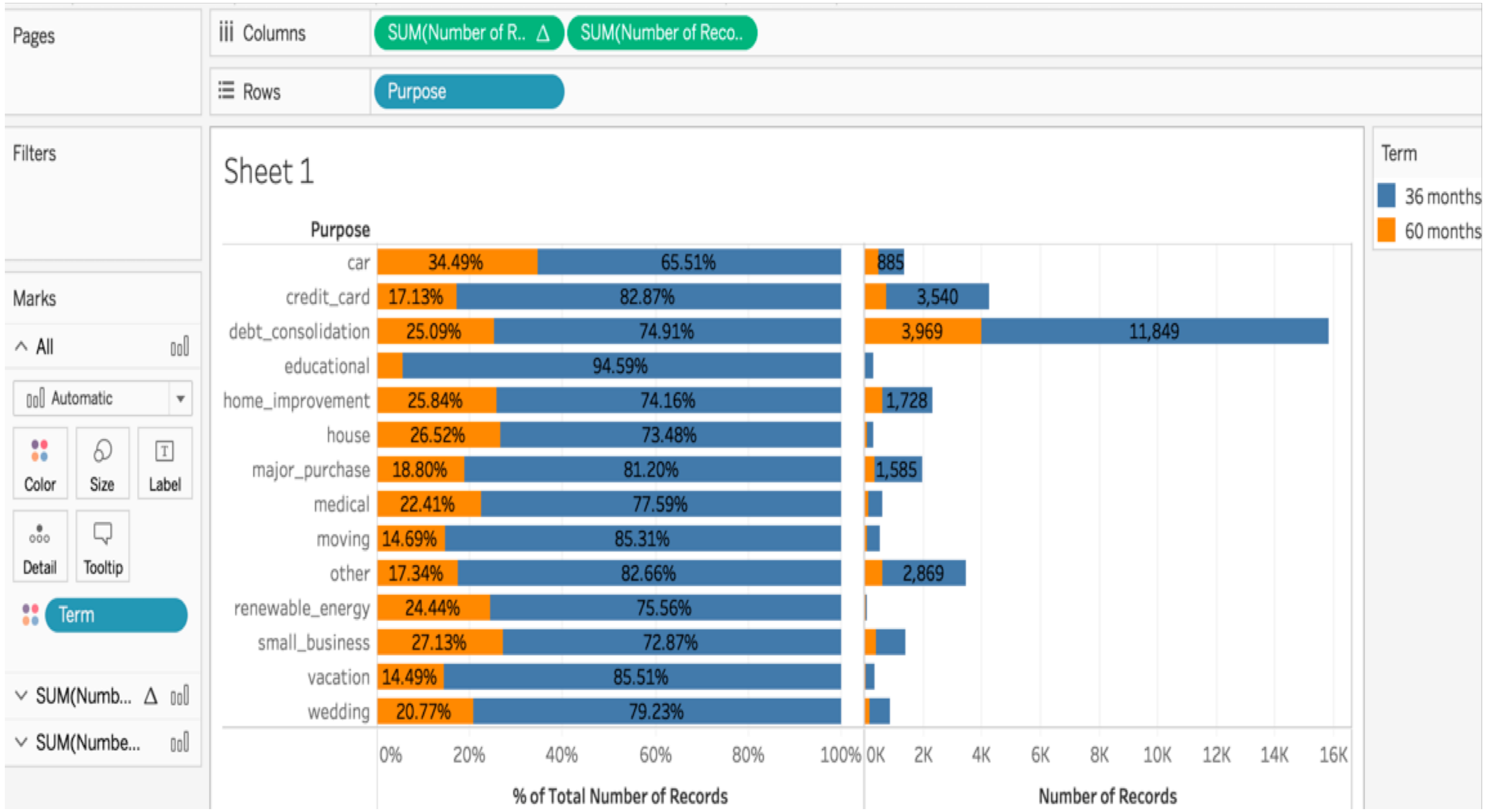
Borrowers who fall under the category of Rent and Mortgage home ownership have substantiate percentages of 60 months loans.

LOAN TERM BY PURPOSE OF LOAN 1

Loans with term of 60 months and are taken for the purpose of debt consolidation are more likely to default:

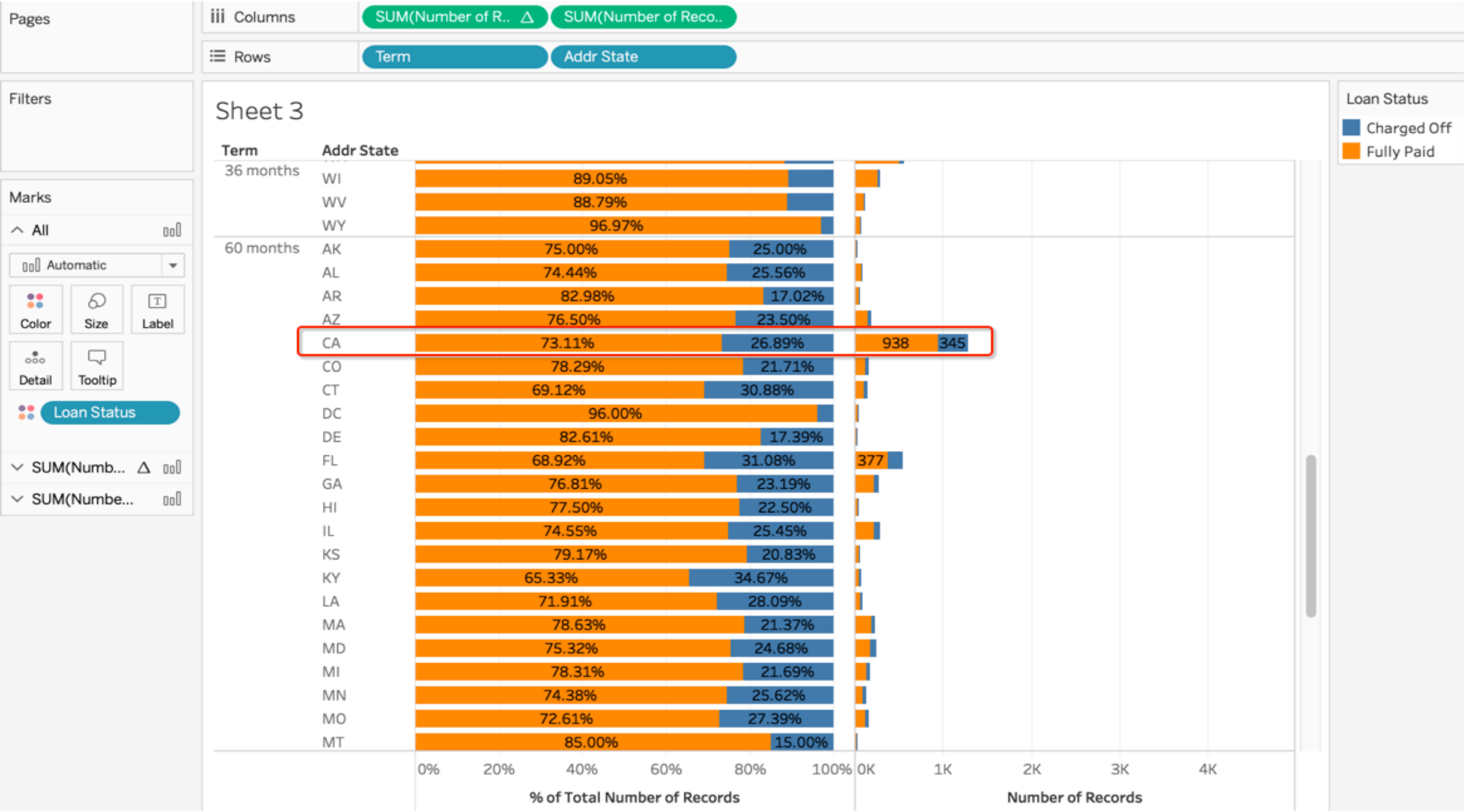
term	purpose	% of deault
60 Months	Debt Consolidation	26.28

LOAN TERM BY PURPOSE OF LOAN 2



Borrowers whose purpose of taking loan is debt consolidation have substantiate percentages of 60 month loans.

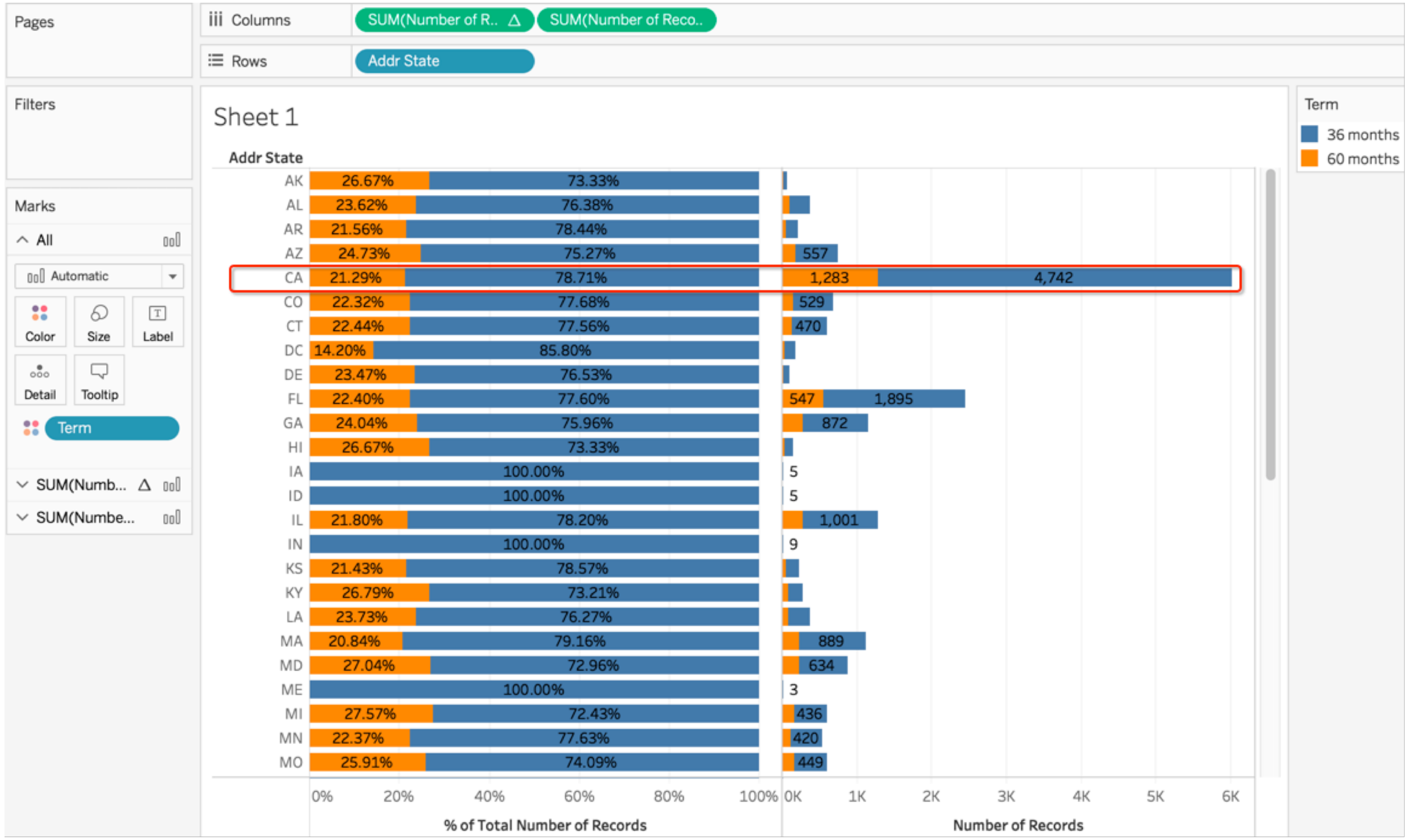
LOAN TERM BY BORROWER ORIGIN STATE 1



Loans with term of 60 months and are taken by borrowers from state CA are more likely to default:

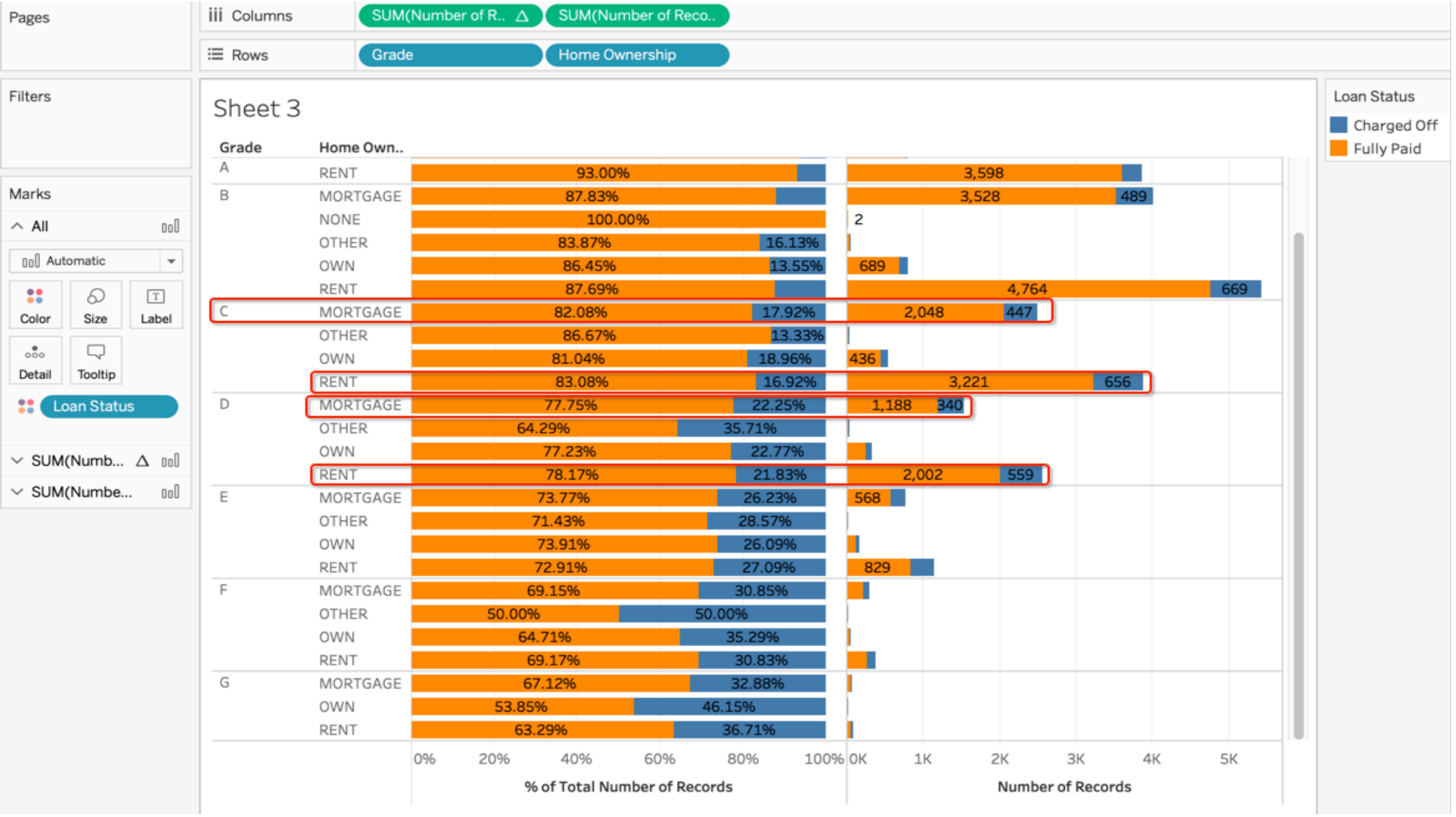
term	state	% of deault
60 Months	CA	26.89

LOAN TERM BY BORROWER ORIGIN STATE 2



Borrowers whose from CA state have substantiate percentages of 60 month loans.

BORROWER GRADE BY BORROWER HOME OWNERSHIP



Borrowers with grade D and who are living in a rented accommodation or whose house is in mortgage are more likely to default:

grade	home_ownership	% of deault
D	RENT	21.83
	MORTGAGE	22.25

BORROWER GRADE BY PURPOSE OF LOAN

Borrowers with grade D and have taken loans for the purpose of debt consolidation are more likely to default:

grade	purpose	% of deault
D	Debt Consolidation	22.2

BORROWER GRADE BY BORROWER ORIGIN STATE

Borrowers with grade D and have originated from state CA are more likely to default:

grade	state	% of deault
D	CA	24.82

OTHER WEAKER CORRELATIONS

home_ownership	purpose	% of deault
MORTGAGE	Debt Consolidation	15.33
RENT	Debt Consolidation	14.89

home_ownership	state	% of deault
MORTGAGE	CA	14.52
RENT	CA	16.25

purpose	state	% of deault
Debt Consolidation	CA	16.05

Borrower's home ownership by purpose & state, Loan purpose by borrower state also revealed some interesting chances but are relatively weaker than the earlier ones.

SUMMARY OF OBSERVATIONS

- Term of the loan combined with other driver variables is observed to be a strong indicator of default as explained in the bivariate analysis.
- Grade, House Ownership combined with other driver variables also act as important variables but only next to term of the loan.
- House Ownership and State combined with other driver variables are weaker indicators of loan default.

RECOMMENDATIONS

- Lending company should be wary of loans whose tenure is 60 months, whose borrowers fall in grades C & D, whose borrower's home ownership is rent or mortgage and whose origin state is California.

- Any combination of the above variable values is also a key indicator of loan default.

- Lending company should either decline the loan, decrease the loan amount or increase the interest rate to curb the losses due to loan applications.