Lending Club Case Study

Report by Divyansh and Amit

Problem statement

- A consumer finance company which specializes in lending various types of loans to urban customers. When the company receives a loan application, the company has to make a decision for loan approval based on the applicant’s profile.

- The data provided by company 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

- The company wants to understand the **driving factors** (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. The company can utilize this knowledge for its portfolio and risk assessment

Approach

❑ *We will use following 3 step approach to reach our Recommendations*

Data Understanding

- Understanding the data set, columns, their type and significance on output variable “Loan status”

Data cleaning & Manipulation

- Identifying the null values, missing values and not required columns and treating them to get refined data set

Data Analysis

- Understanding data and patterns through various plots to firm our understanding and recommendations

Data Understanding & Data clean up

```
In [53]: import numpy as np
         import pandas as pd
         import matplotlib.pyplot as plt
         import seaborn as sns
         import plotly.express as px
         import plotly.graph_objects as go
         loan = pd.read_csv("loan.csv", sep=",")
         loan.info()

<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
```

- Important to note that we started with 111 columns in data set but there were lots of columns which had missing values
- We removed any column that had more than 20% missing values

```
Index(['mths_since_last_record', 'next_pymnt_d', 'mths_since_last_major_derog',
       'annual_inc_joint', 'dti_joint', 'verification_status_joint',
       'tot_coll_amt', 'tot_cur_bal', 'open_acc_6m', 'open_il_6m',
       'open_il_12m', 'open_il_24m', 'mths_since_rcnt_il', 'total_bal_il',
       'il_util', 'open_rv_12m', 'open_rv_24m', 'max_bal_bc', 'all_util',
       'total_rev_hi_lim', 'inq_fi', 'total_cu_tl', 'inq_last_12m',
       'acc_open_past_24mths', 'avg_cur_bal', 'bc_open_to_buy', 'bc_util',
       'mo_sin_old_il_acct', 'mo_sin_old_rev_tl_op', 'mo_sin_rcnt_rev_tl_op',
       'mo_sin_rcnt_tl', 'mort_acc', 'mths_since_recent_bc',
       'mths_since_recent_bc_dlq', 'mths_since_recent_inq',
       'mths_since_recent_revol_delinq', 'num_accts_ever_120_pd',
       'num_actv_bc_tl', 'num_actv_rev_tl', 'num_bc_sats', 'num_bc_tl',
       'num_il_tl', 'num_op_rev_tl', 'num_rev_accts', 'num_rev_tl_bal_gt_0',
       'num_sats', 'num_tl_120dpd_2m', 'num_tl_30dpd', 'num_tl_90g_dpd_24m',
       'num_tl_op_past_12m', 'pct_tl_nvr_dlq', 'percent_bc_gt_75',
       'tot_hi_cred_lim', 'total_bal_ex_mort', 'total_bc_limit',
       'total_il_high_credit_limit'],
      dtype='object')
```

Data Understanding & Data clean up, Assumptions

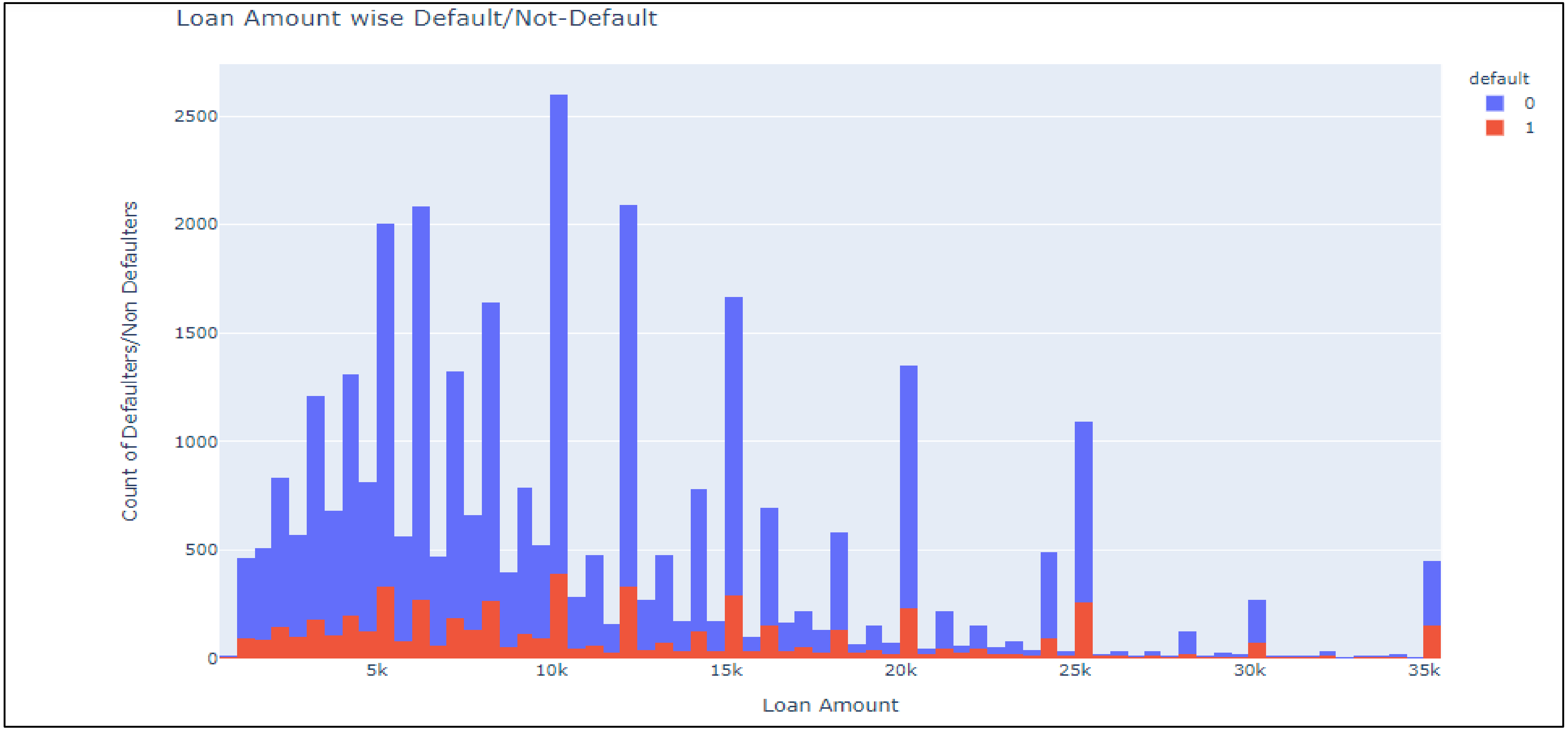
```
In [64]: round(loan.isnull().sum()/len(loan.index), 2)*100

Out[64]: id                            0.0
         member_id                     0.0
         loan_amnt                     0.0
         funded_amnt                   0.0
         funded_amnt_inv               0.0
         term                          0.0
         int_rate                      0.0
         installment                   0.0
         grade                         0.0
         sub_grade                     0.0
         emp_title                     6.0
         emp_length                    3.0
         home_ownership                0.0
         annual_inc                    0.0
         verification_status           0.0
         issue_d                       0.0
         loan_status                   0.0
         pymnt_plan                    0.0
         url                           0.0
         purpose                       0.0
         title                         0.0
         zip_code                      0.0
         addr_state                    0.0
         dti                           0.0
         delinq_2yrs                   0.0
         earliest_cr_line              0.0
         inq_last_6mths                0.0
         open_acc                      0.0
         pub_rec                       0.0
         revol_bal                     0.0
         revol_util                    0.0
         total_acc                     0.0
         initial_list_status           0.0
         out_prncp                     0.0
         out_prncp_inv                 0.0
         total_pymnt                   0.0
         total_pymnt_inv               0.0
         total_rec_prncp               0.0
         total_rec_int                 0.0
         total_rec_late_fee            0.0
         recoveries                    0.0
         collection_recovery_fee       0.0
         last_pymnt_d                  0.0
         last_pymnt_amnt               0.0
         last_credit_pull_d            0.0
         collections_12_mths_ex_med    0.0
         policy_code                   0.0
         application_type              0.0
         acc_now_delinq                0.0
         chargeoff_within_12_mths      0.0
         delinq_amnt                   0.0
         pub_rec_bankruptcies          2.0
         tax_liens                     0.0
         dtype: float64
```

Below three are the remaining ones with null values

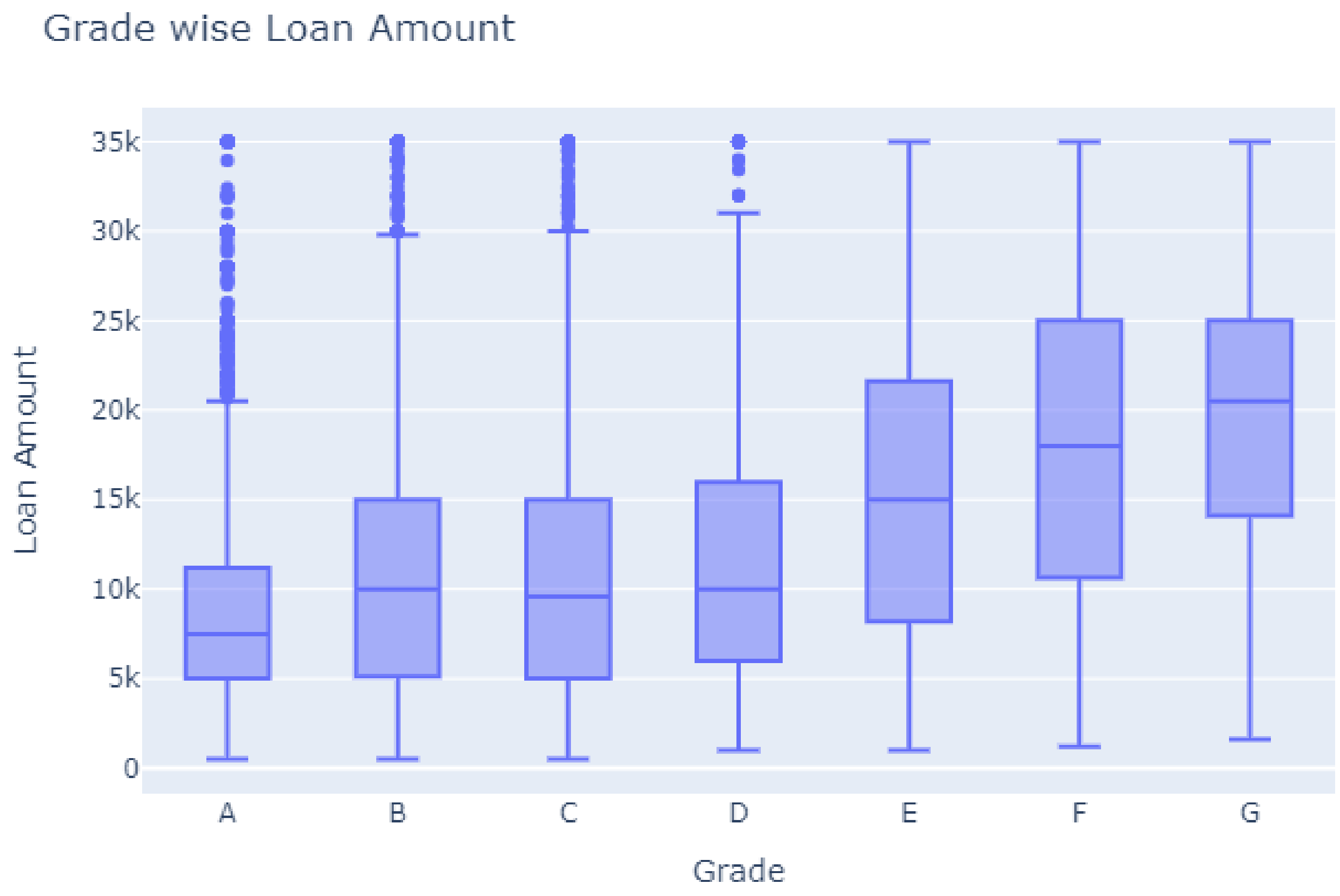
- ❑ Post missing value treatment there were no variable with more than 6% missing values.
- ❑ We assume that we must not treat these values or populate them with some sample statistic as this may change the insights which in terms of credit system can be very divergent.

Preliminary Analysis



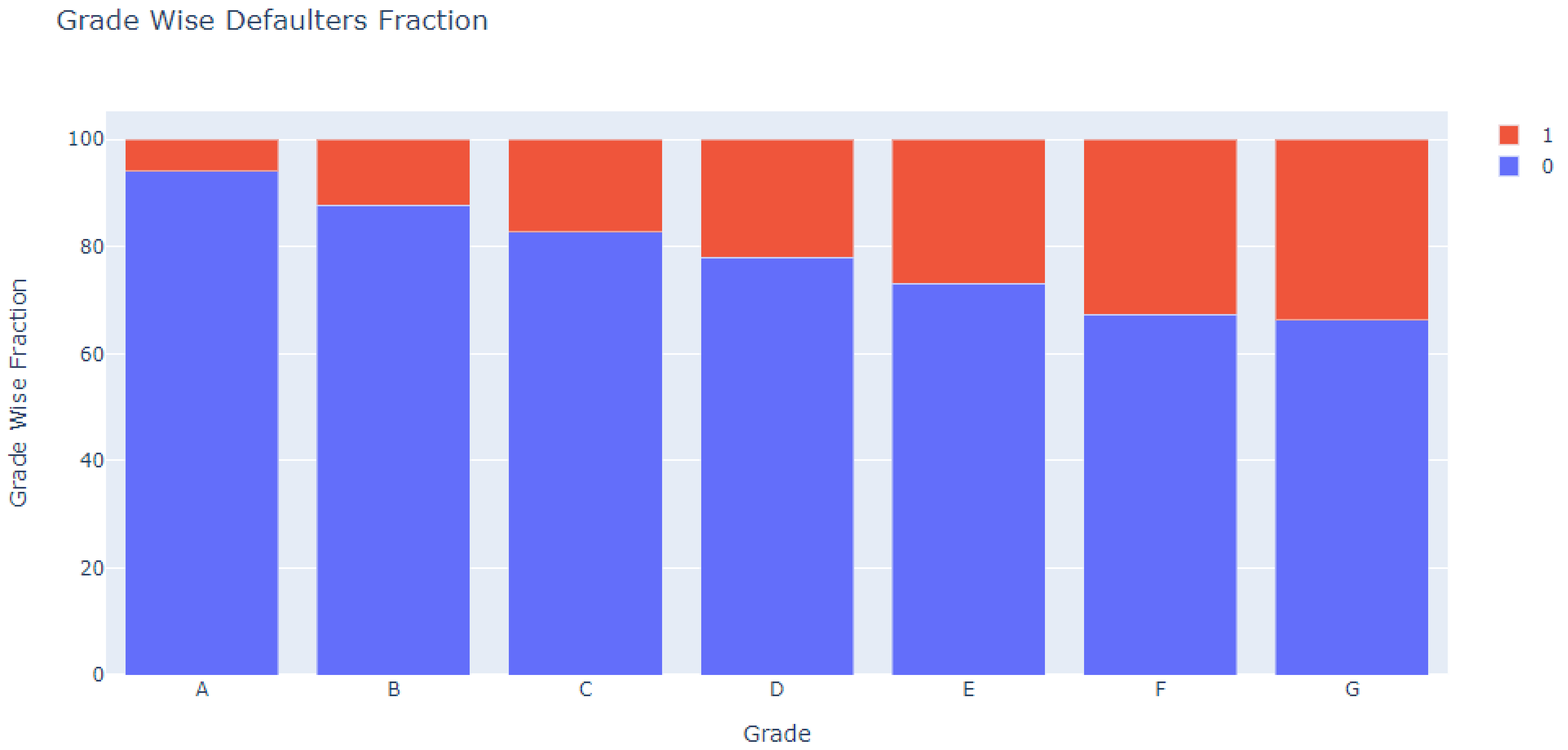
- During preliminary analysis we found out that customers usually have a interesting behavior of taking loans in multiples of 5000 (round off amounts).

Loan Amount vs Grade



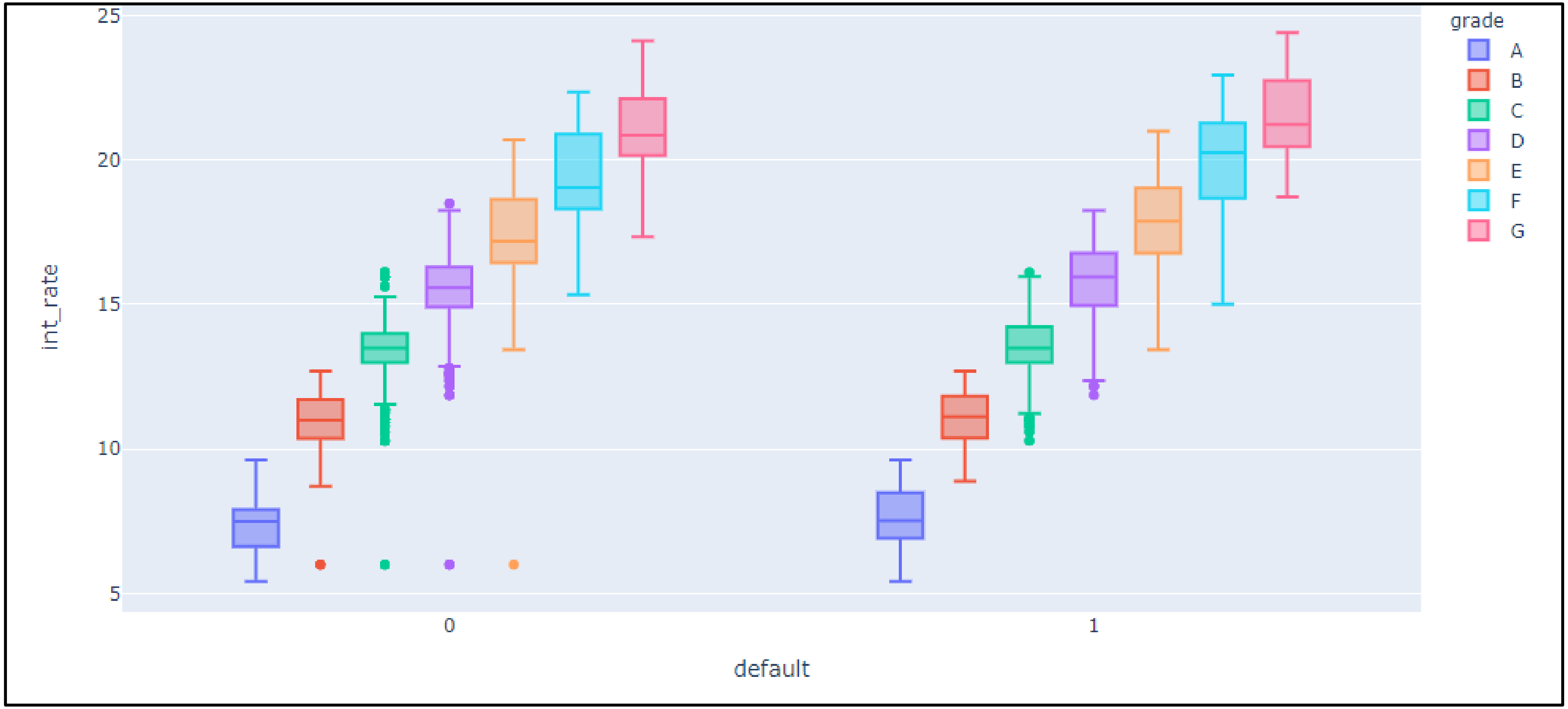
- ❑ Loan Amount Increases with the Grade assigned by LC from A to G.

Loan status wise defaulters fraction for each Grade



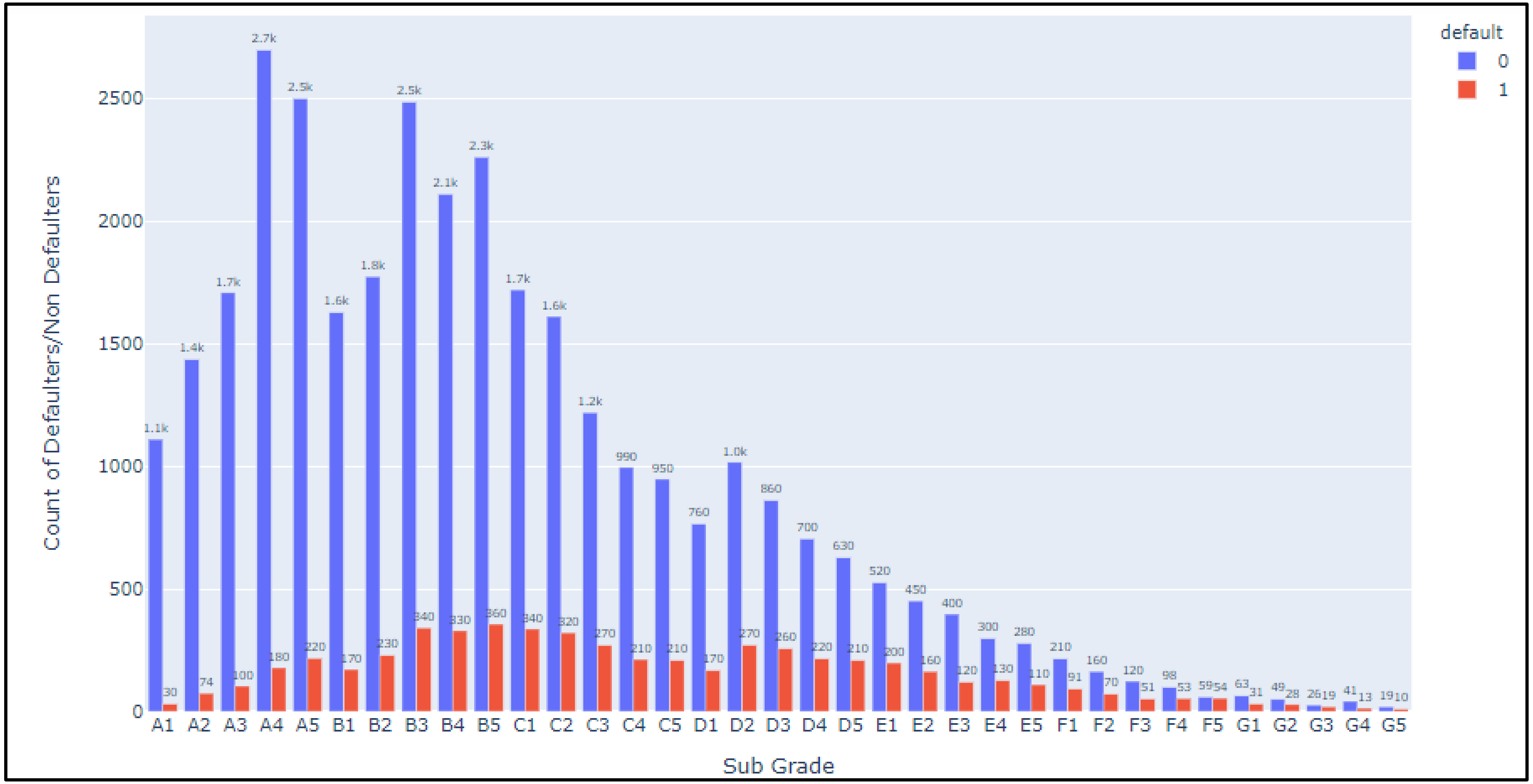
- ❑ Grade G has highest default rate followed by F. Almost 33% of the customers default in grade G.
- ❑ Grade A is safest with less than 10 % of the customers defaulting.

Loan status vs Interest rate



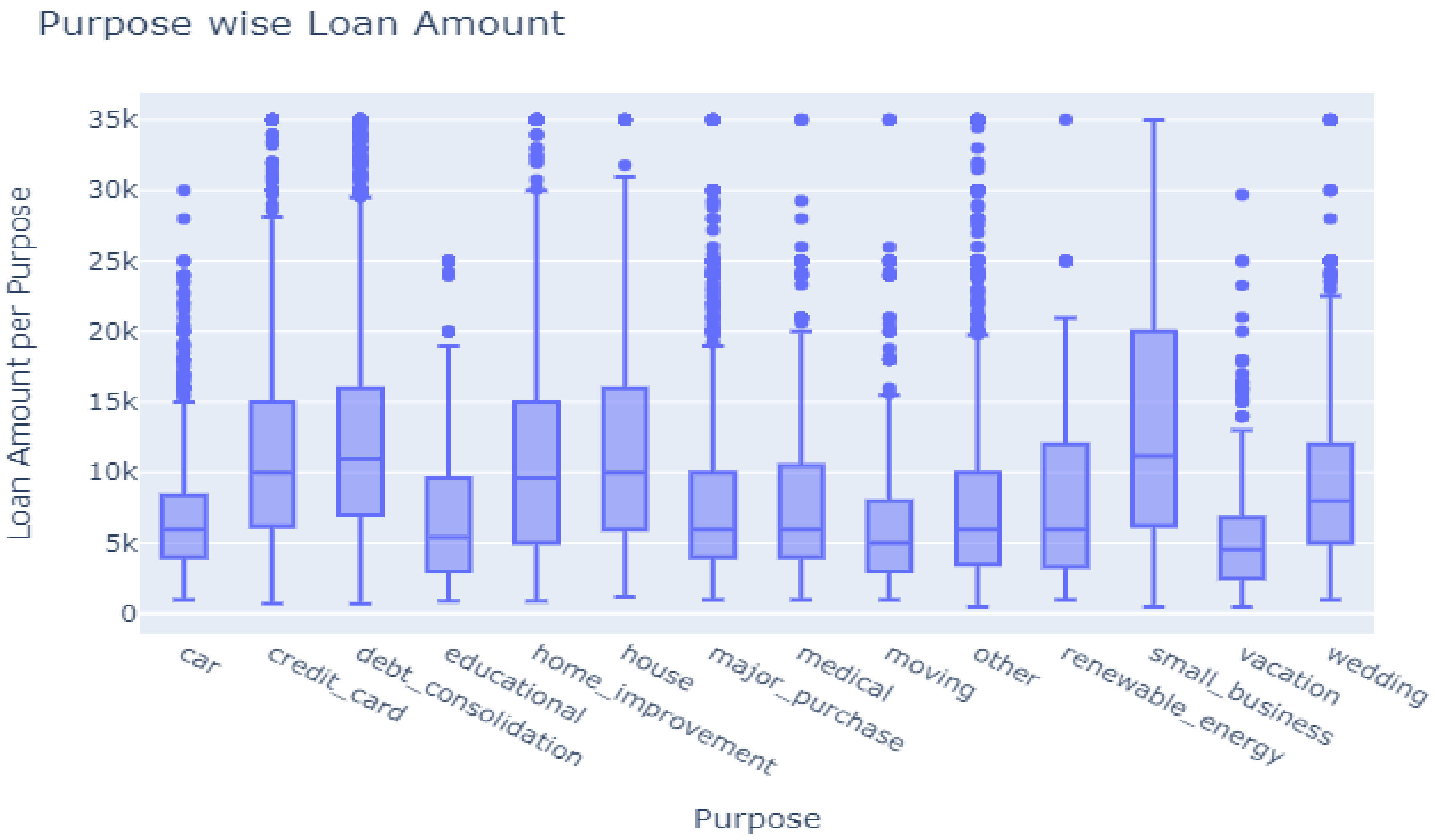
- ❑ We Identified that LC provided loan at higher interest rates to the customers who are allotted G grade vs the least interest rate for customers allotted A grade. Thus, people receiving loans in this grade default more.

Loan status vs Sub Grade



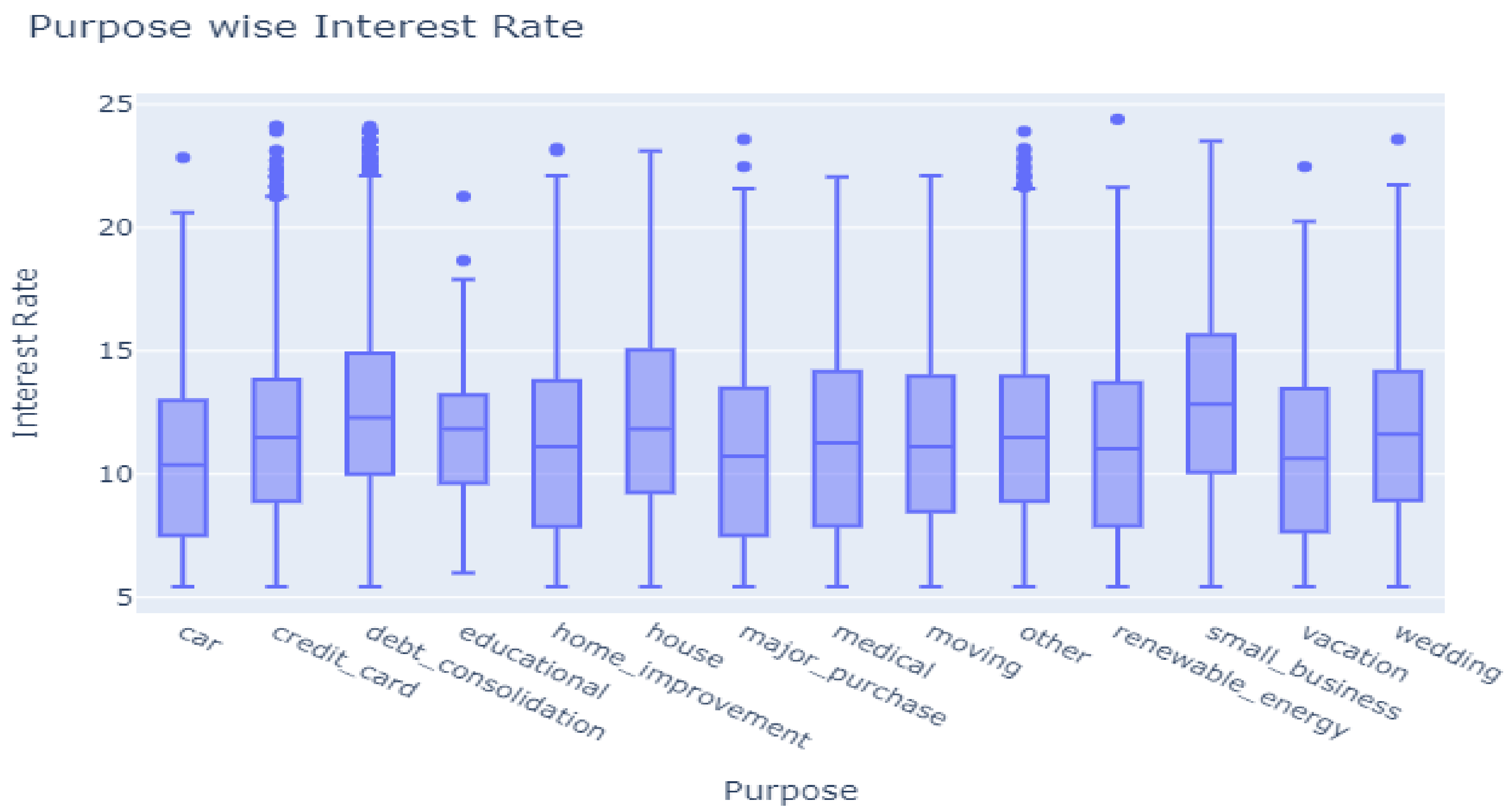
- ❑ Grade A is safest as sub categories also follow the same pattern
- ❑ Category G having the worst sub categories in terms of loan payment.

Loan Amount vs Purpose



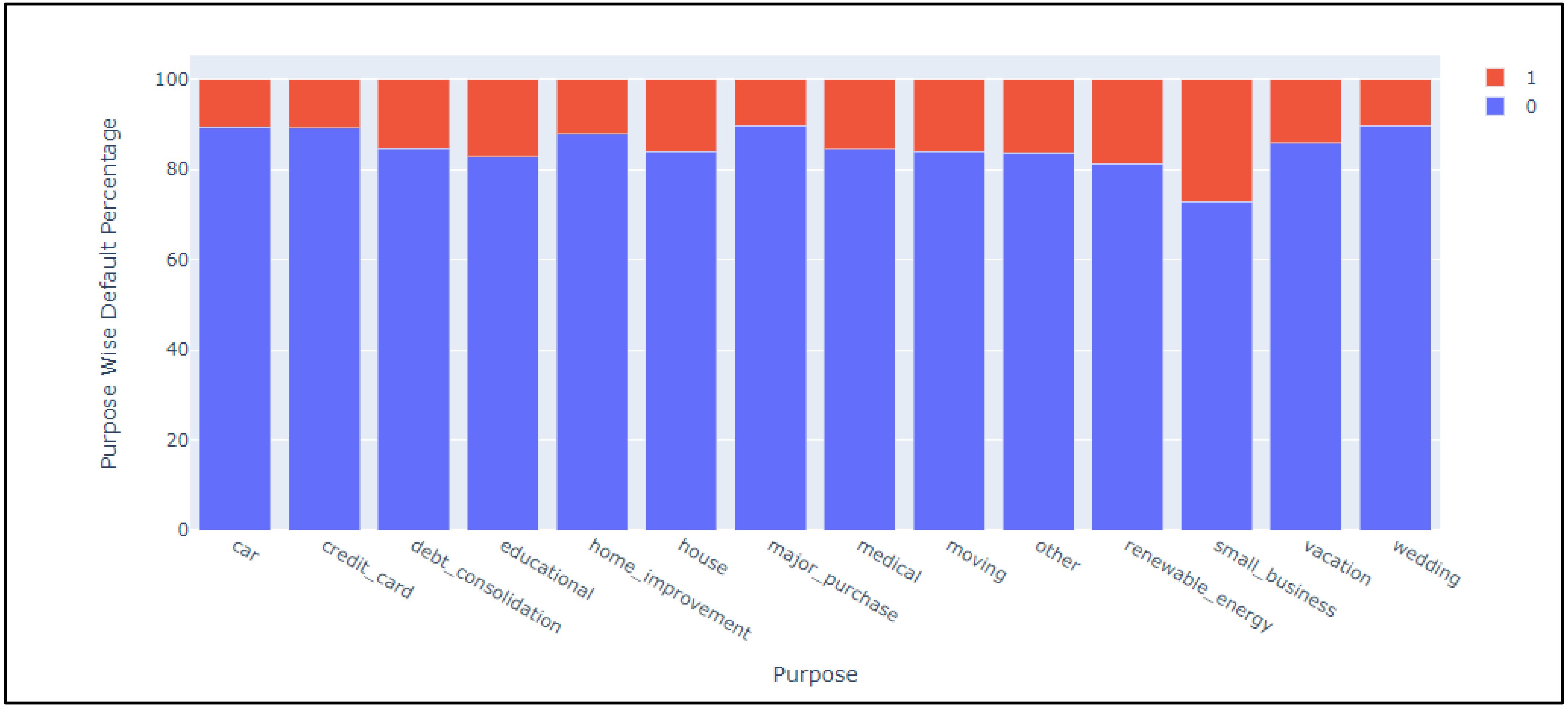
- ❑ Customers take maximum amounts of loan for Small Businesses, house, debt consolidation.
- ❑ Customers take maximum amounts of loan for vacations, moving and cars.

Interest Rate vs Purpose



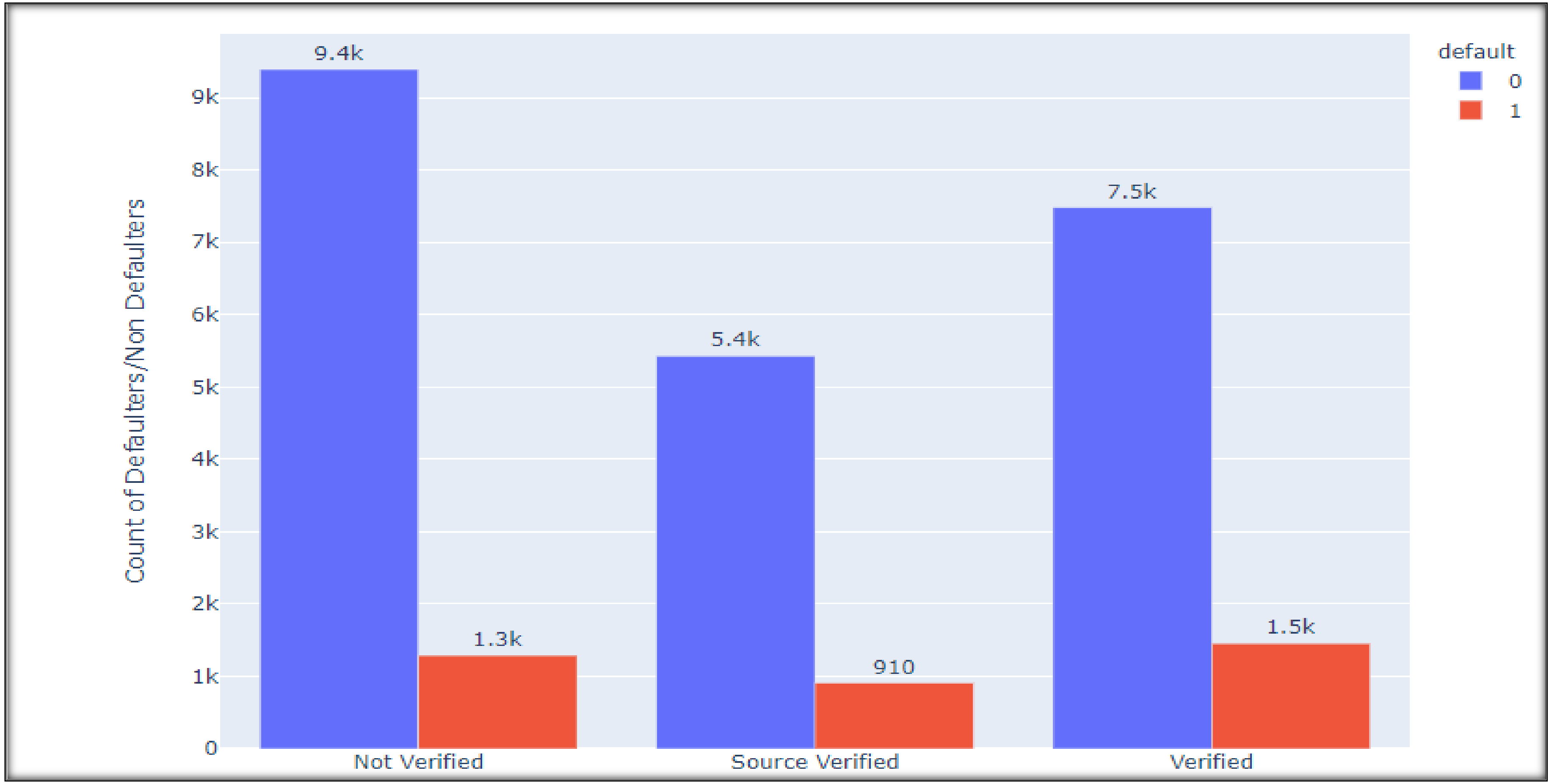
- ❑ Interest Rates are higher for Small Businesses, house, debt consolidation.
- ❑ Interest Rates are smaller for vacations, major purchase and cars.

Loan status vs Purpose



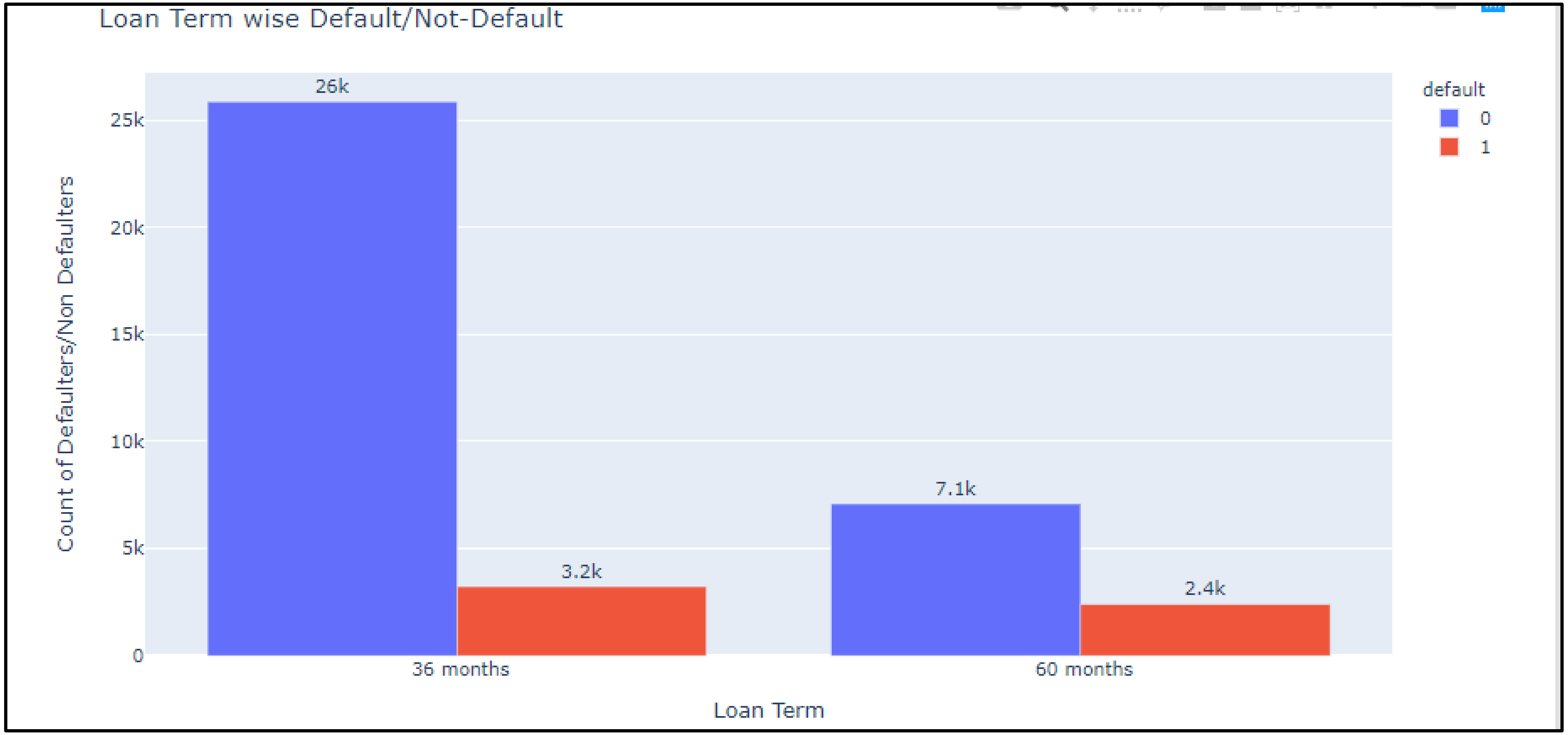
- ❑ Purpose wise default rate is maximum for small business, while wedding and car loans look safer option with the least default rate.

Loan status vs Verified Income Sources



- Interestingly, the customers with verified income sources are defaulting more as compared to those whose income sources are Not Verified. Thus, Income source verification looks less impacting in reducing default.

Loan status vs term of loan



- Customers who take loans for longer term (60 months) often default more. Thus, loans with less term period (36 months) are recommended to reduce chances of default.

Interest vs term of loan

- Customers who take loans for longer term (60 months) need to pay higher Interest Rates as compared to those taking loans for less time period of 36 months.

Recommendations

Based on our Analysis, we came up with following recommendations for Investors :

- Customers with long term loans tend to default more, recommendation is to encourage shorter term loans (36 month) as the interest rates are also higher in long term loans and customers fail to repay the loans.
- Higher LC Grade Loans have a impact on default rate as the interest rates also increase from grade A to grade G. Recommendation is to prefer Grade A, B, C over D, E, F, G. This is a Loan Attribute influencing tendency to default.
- Purpose of loan is also a strong indicator, recommendation is to scrutinize more for small business and renewable energy related loans. Wedding, cars, big purchases remains safe bet. Also one can provide small business loans at lower interest rates to reduce defaults. This is a consumer Attribute influencing tendency to default.
- Income source verification is not showing impact on tendency to default, recommendation is to not rely on income source verification for approving loans. But customers with lower incomes are tending to default more.
- Loan Default tendency driving variables are purpose, Grade, Interest rate, Annual Income, Term of Loan.