LendingClub CASE STUDY

SUBMISSION

Name: Rajib Biswas

Date: 21/04/2021

Abstract

- Lending Club is an American peer-to-peer online based lending company which provides mostly unsecured loans to applicants for reasons like investing in small business, medical expenses etc..
- As with most financial organisations, issuing of loans comes with significant credit loss risks. I am analysing past loan data to identify "Features" which can be used in pre-empting future defaults and credit –risk applicants.

Business Understanding

- With every loan application there are 2 types of risks associated-
  - 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.
- If one is able to identify these risky loan applicants, then such loans can be reduced thereby cutting down the amount of credit loss. Identification of such applicants using EDA is the aim of this case study.

Objective

- Aim is to understand the driving factors (or driver variables) behind loan default, i.e. the variables which are strong indicators of default. Lending Club can utilise this knowledge for its portfolio and risk assessment.

Data Understanding

- Lending Club has provided complete loan data for all loans issued through the time period 2007 to 2011. This dataset contained information related to borrower's credit information along with loan information.
- Dataset has 39717 records with 111 columns . Column description has been provided in Data Dictionary excel file.
- Loans are classified into 3 main categories –
  - **Fully Paid** – Loans that are fully paid and closed
  - **Charged Off** – Bad loans
  - **Current** – Loans with on-going repayment plans

Problem Solving Methodology

START

Understand-
- Business Objective
- Goal of Data Analysis

Input Data –
- Loan Data
- Data dictionary

Data Analysis and Preparation

- Identify Target Column(loan_status)
- Identify consumer and loan attributes

- Ensure Data Quality
- Identify and treat missing data
- Filter relevant data and standardize for analysis purpose

A

Data Analysis using Pandas, Numpy, Matplotlib, Seaborn

A

Identify the correct features

Univariate Analysis –
- Categorical
- Quantitative

Derived Metrics
- Type Driven
- Business Driven
- Data Driven

Segmented Univariate Analysis

Bivariate Analysis

Plot Graphs

Identify the features impacting the decision

END

Data Analysis & Cleaning

- *Key features are grouped as Loan and Consumer attributes as below -*

Loan Attributes	Customer Attributes
Interest rate	Annual Income
Term	Employment Length
Funded Amount	Debt to Income ratio
Grade	Address
Purpose	Number of Public record bankruptcies

Statistics

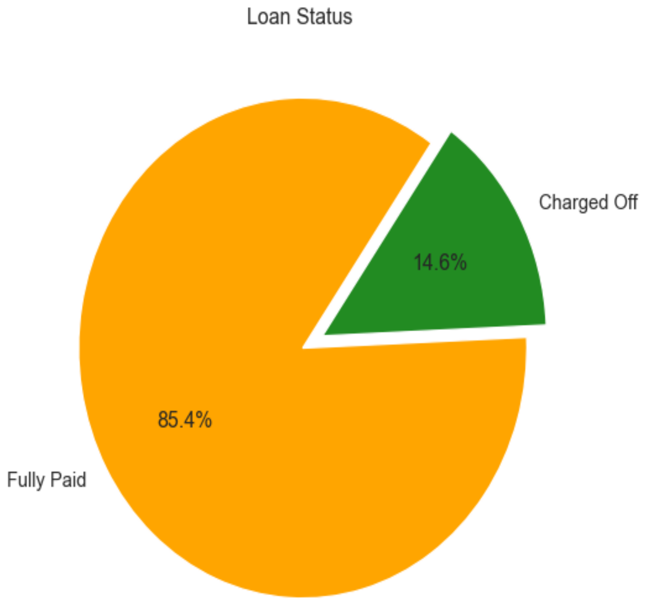
57 — 57 Columns had no value or completely blank. All were dropped

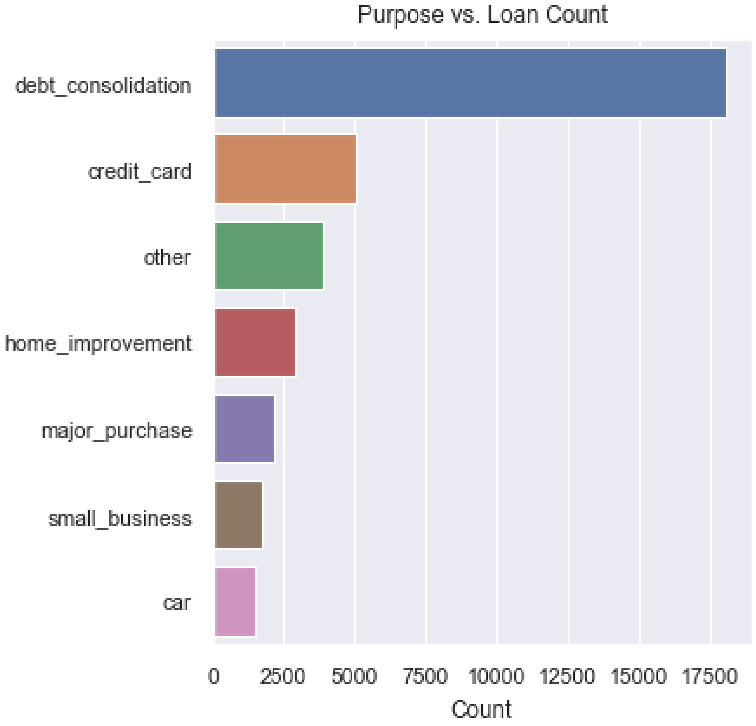
33 — 33 Columns were related to customer behaviour or demographic data. All were dropped

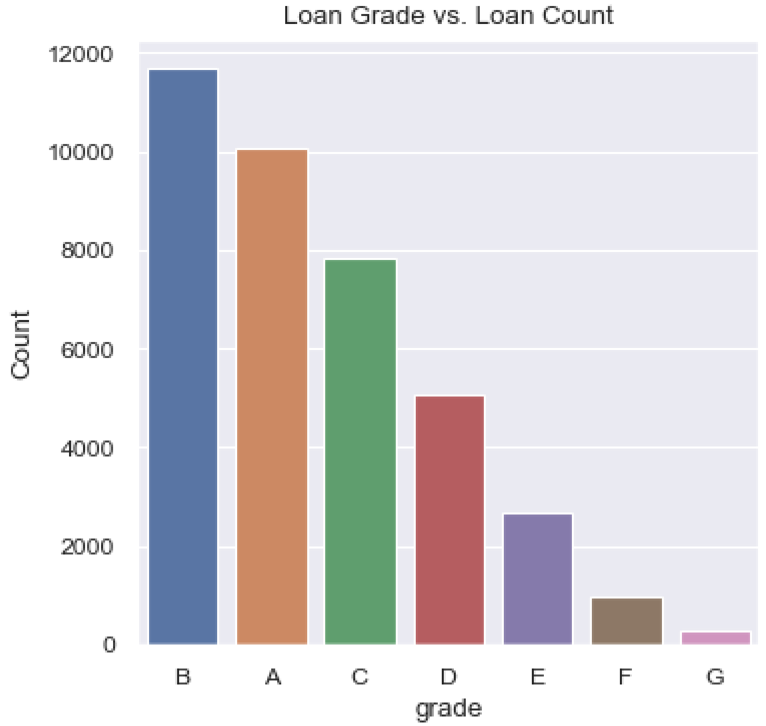
4 — 4 numerical columns binned into ranges for easy plotting and analysis

1 — 1 business driven derived metric – ***"dti_updated"*** created for adding valuable insight in the analysis

Univariate Analysis – Categorical Data







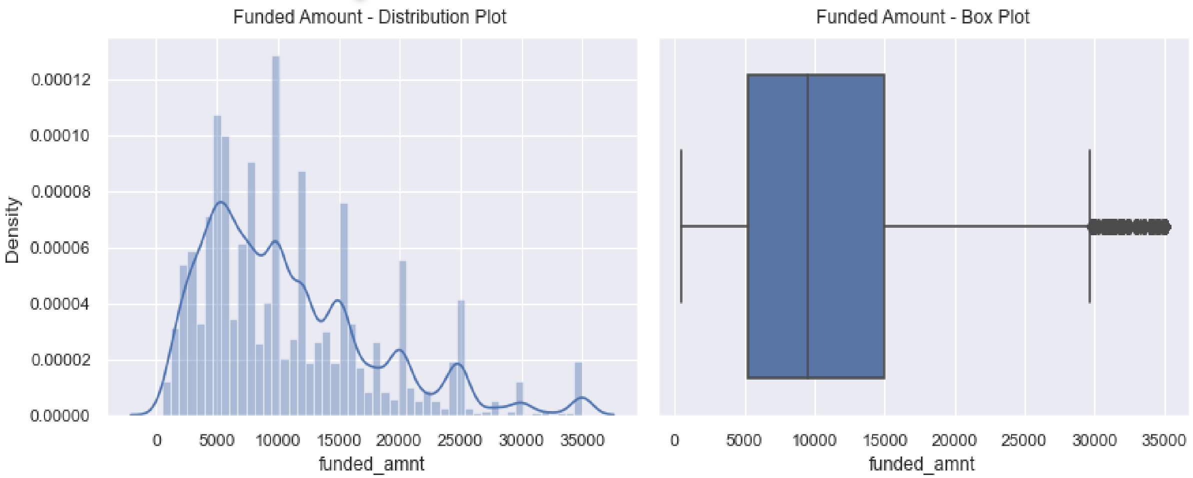
- ❑ **14.6%** of total loans has resulted in default.
- ❑ Remaining loans were fully paid or in current status.

- ❑ Majority of existing loans were taken to pay off ***existing loan obligations*** followed by ***credit card payments***.

- ❑ Majority of loans approved by Lending Club seems to be of ***Grade A to D*** which are low risk and low interest rate.

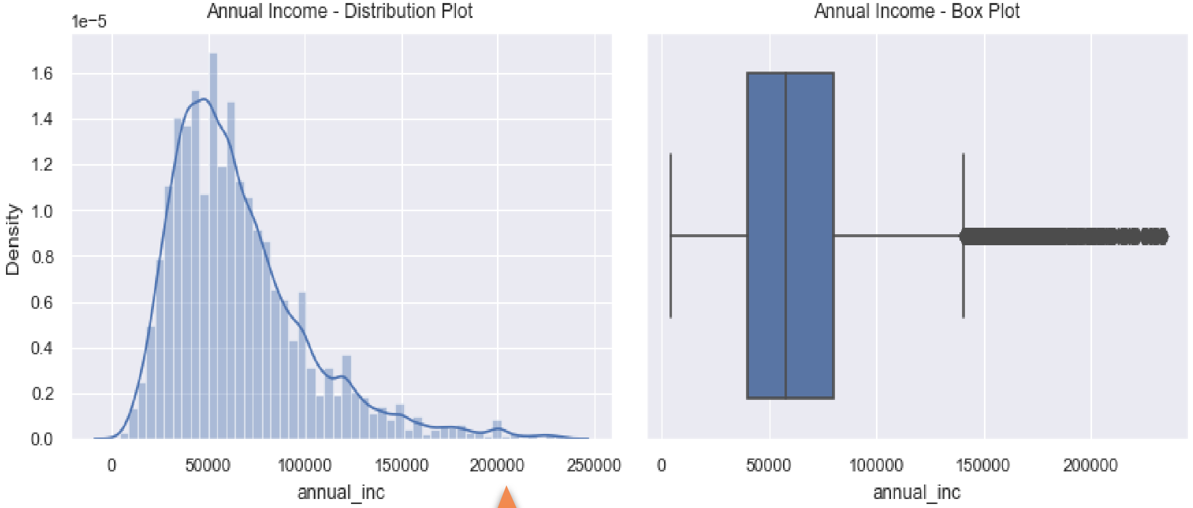
Univariate Analysis –Quantitative Data

Funded Amount



- Lending Club approves loan between ***500 USD*** and ***35,000 USD***.
- As evident from the plot above, majority of applicants received funded loans between ***5000 to 20000 USD***.

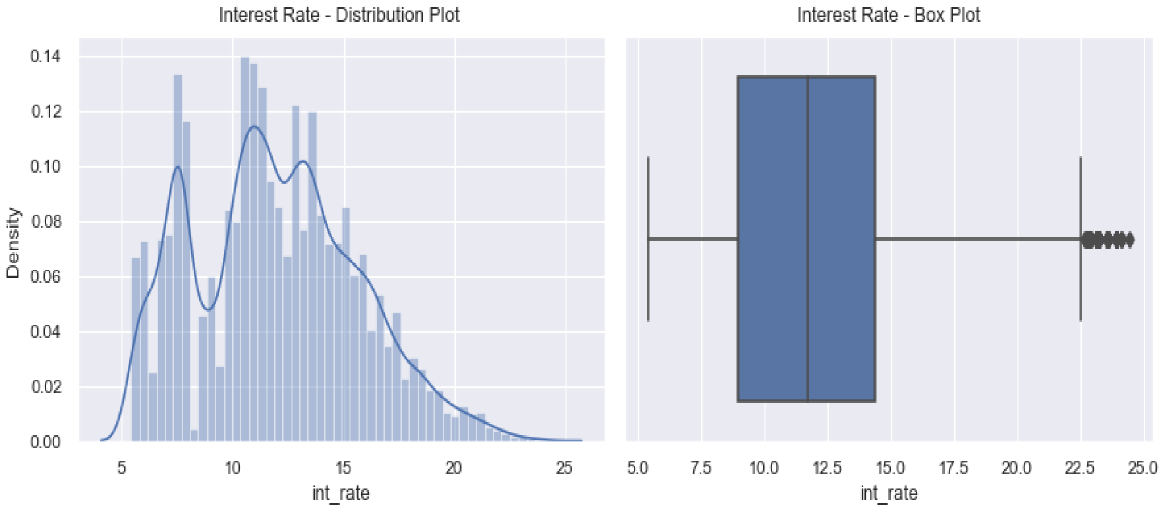
- Annual Income for majority of applicants are between ***20000 to 150000 USD***.
- Maximum count of applicants are at ***60000 and 50000 USD***.



Annual Income

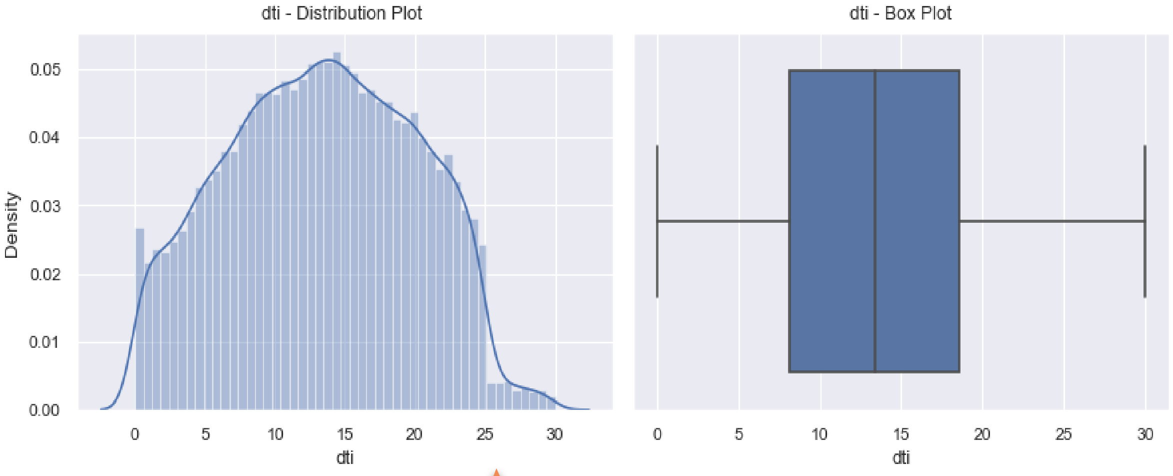
Univariate Analysis –Quantitative Data

Interest Rate



- From the distribution plot we can infer interest rates for loans varies between ***5% to 24%.***
- However, majority of the loans were sanctioned with interest rate hovering around ***5% and 15%***.

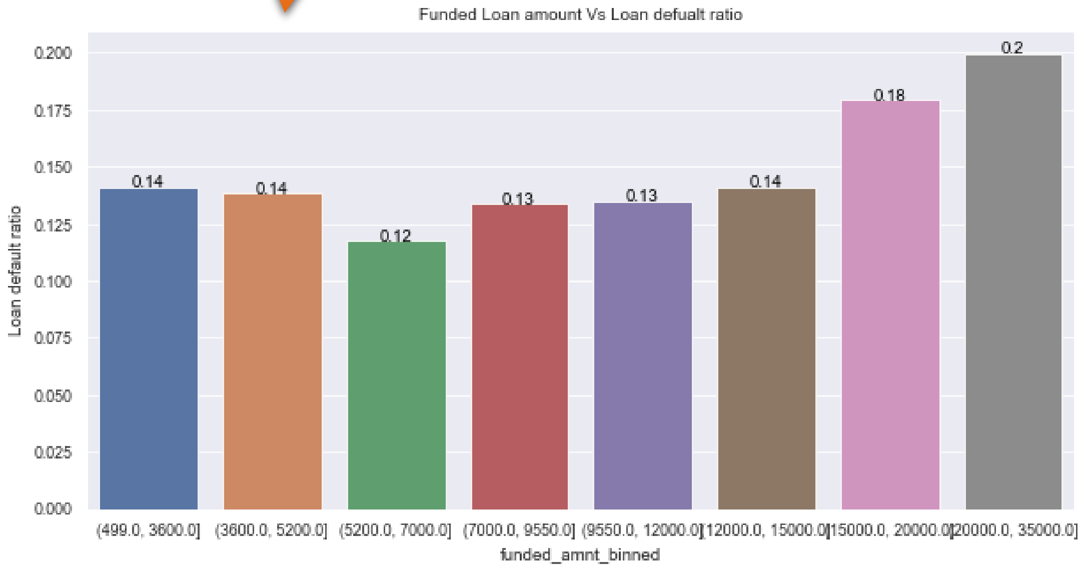
- Majority of applicants with approved loan has dti between ***0 to 25***.
- Between ***25 and 30*** the loan approval drops significantly and beyond ***30*** there are no applicants with approved loans.



Debt to Income

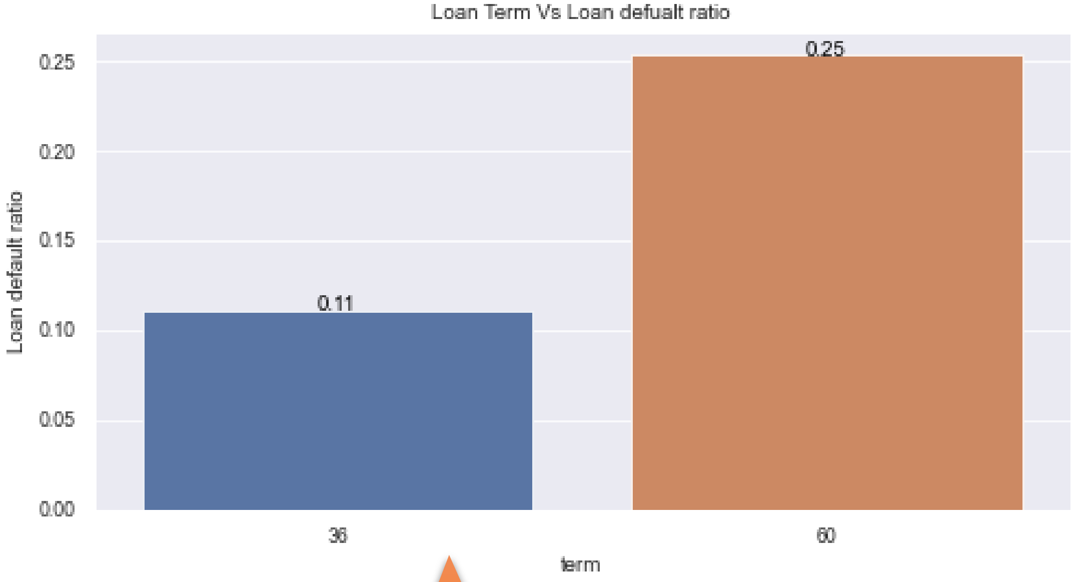
Segmented Univariate Analysis

Funded Loan Amount



- With increase in loan amount the default rate also rises.
- Between ***20k-35 k USD, the default ratio zooms past 20% mark.***

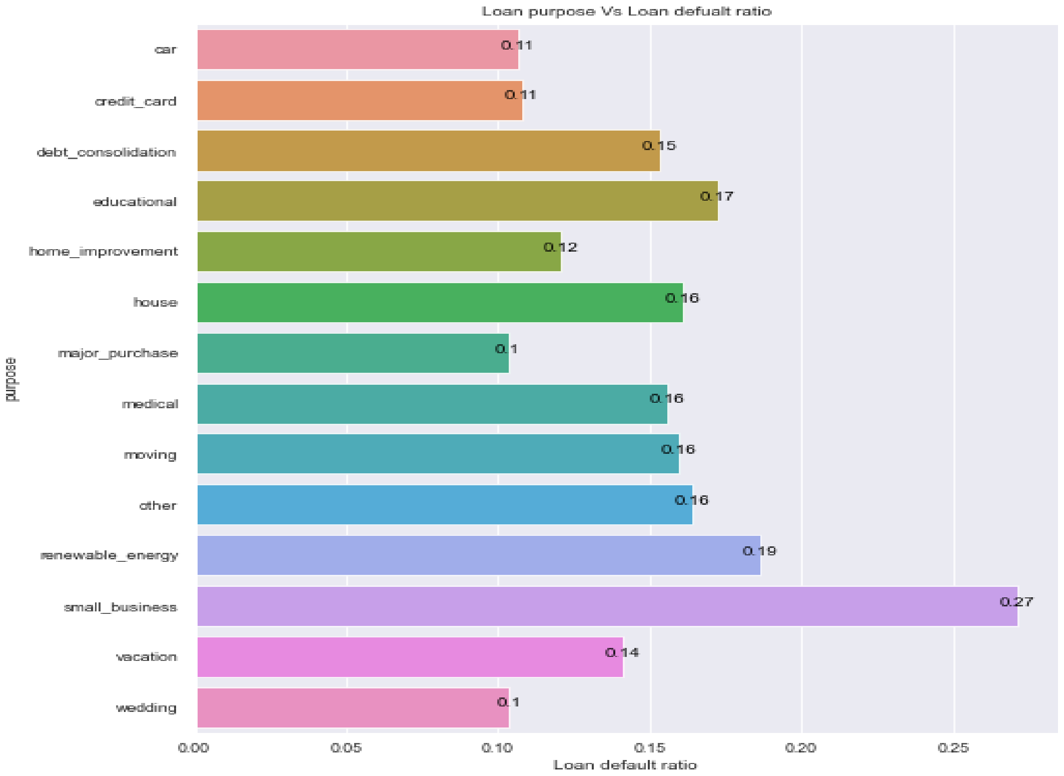
- Defaulters rate is increasing wrt term, ***hence the chances of loan getting defaulted is less for 36m than 60m.***



Term

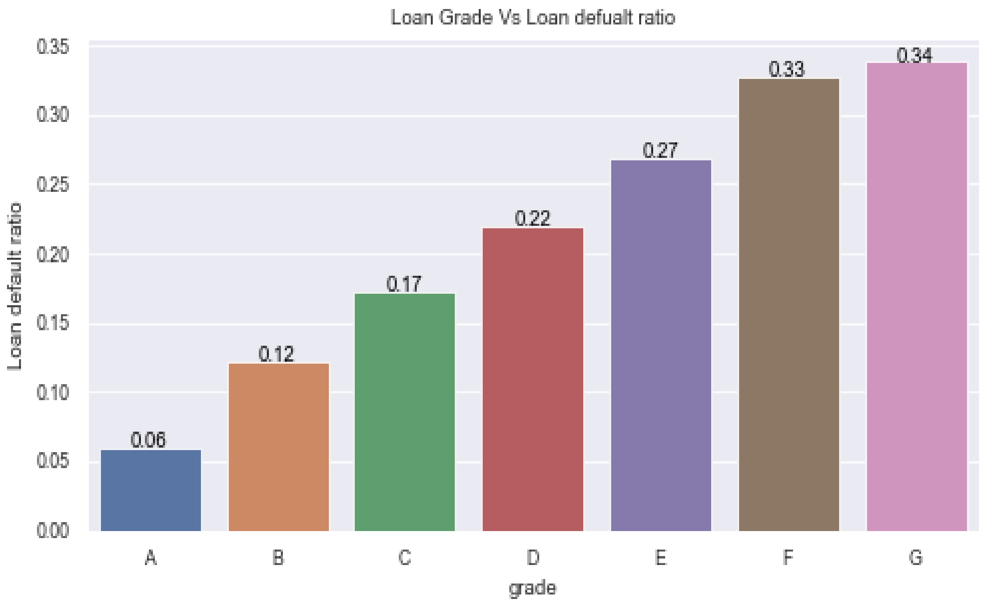
Segmented Univariate Analysis

Funded Loan Amount



- Applicants who have taken loans for small business is more likely to default.

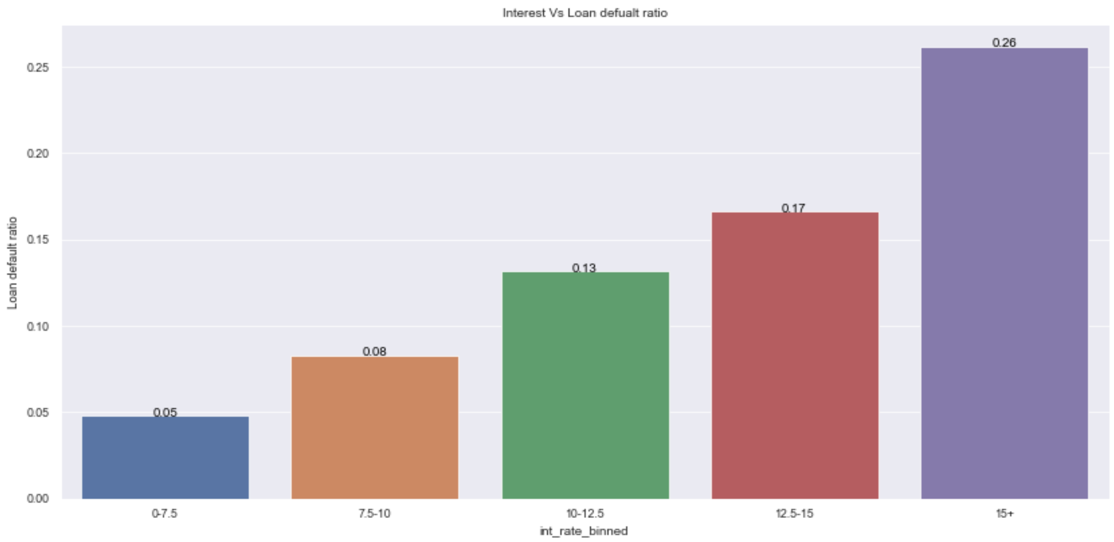
- Defaulters rate is increasing wrt grade, hence the chances of loan getting defaulted ***increases with the grade from A moving towards G.***



Grade

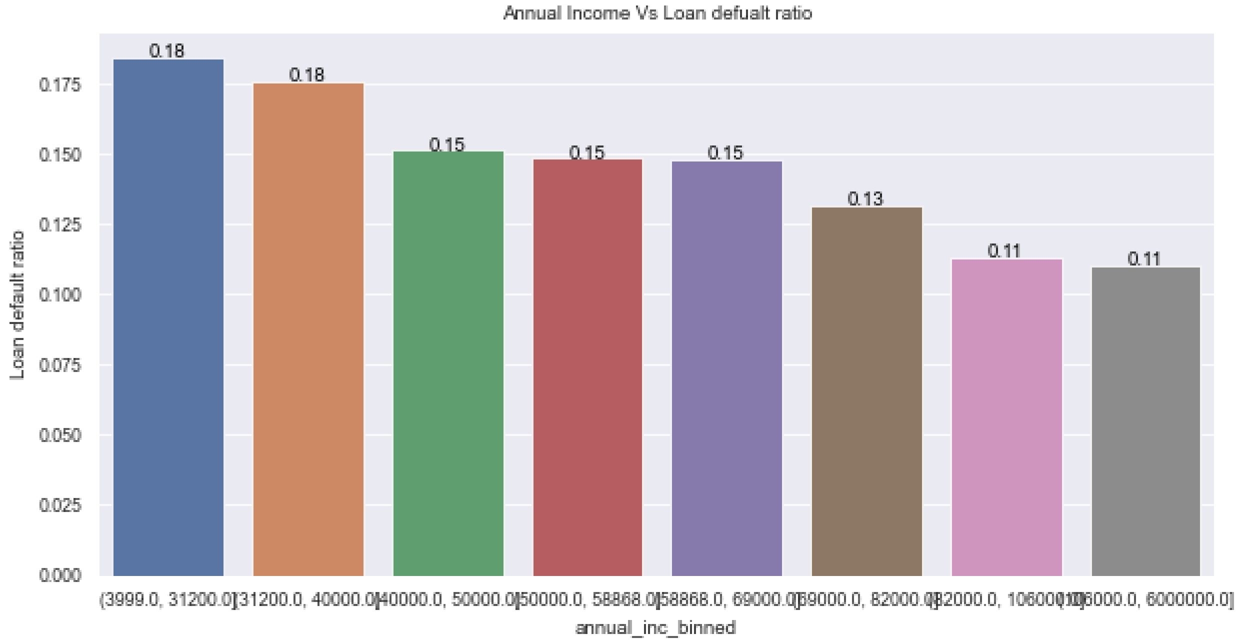
Segmented Univariate Analysis

Interest Rate



- ❑ Default rate increases with the increase in the interest rate.

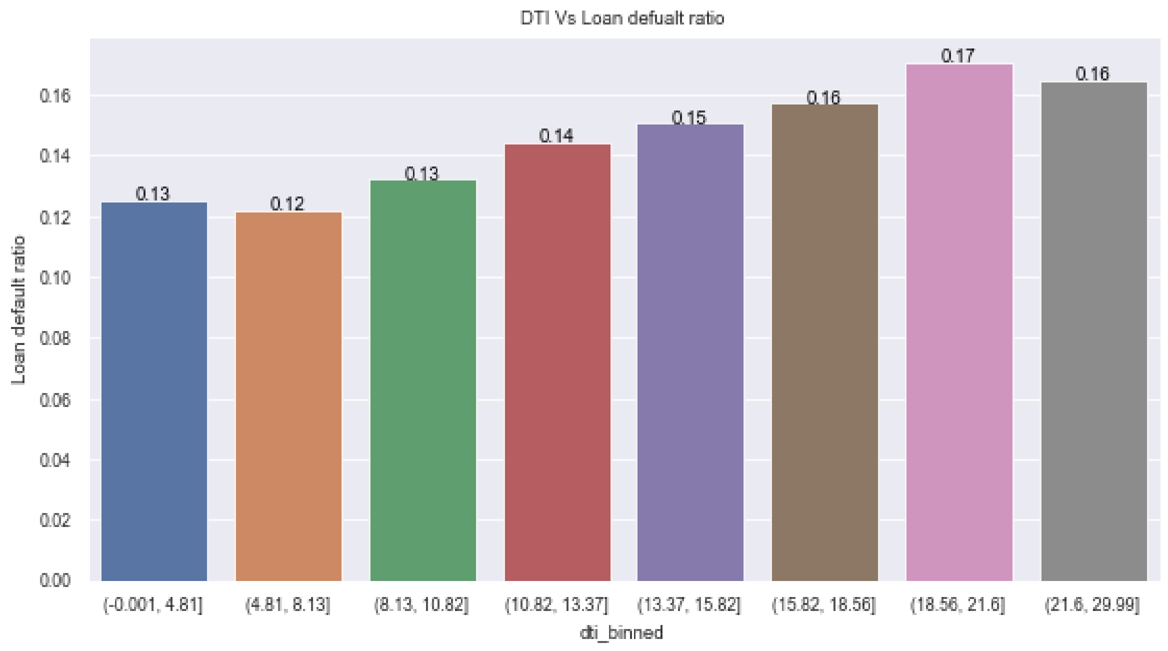
- ❑ Applicants with lower incomes are more likely to default.
- ❑ ***Between 4k-40K the total default rate is around 36%.***



Annual Income

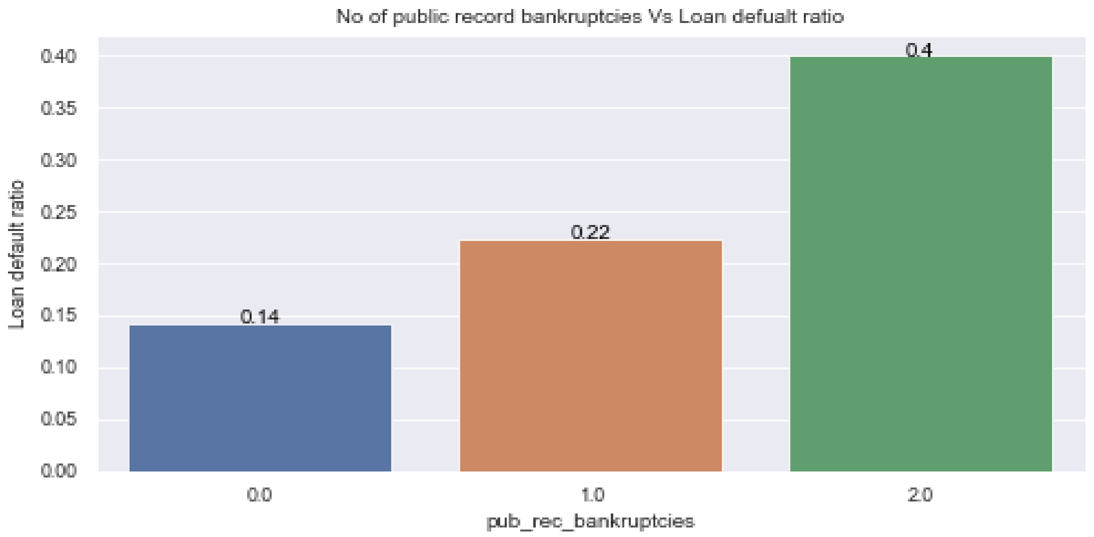
Segmented Univariate Analysis

Debt to Income



- ❑ Default rate increases with the increase in the debt to income ratio.

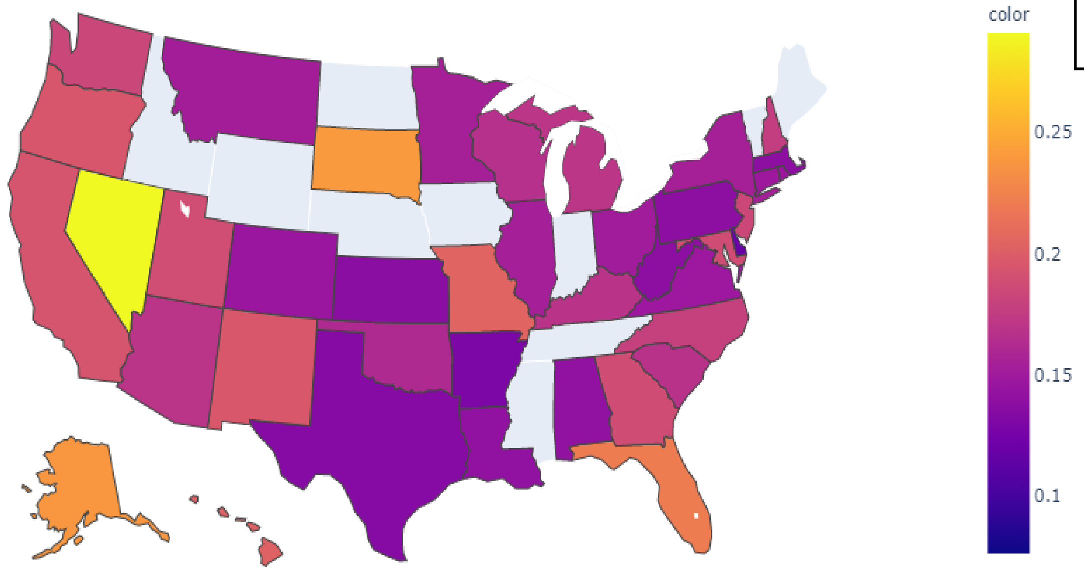
- ❑ Applicants with more than 1 past bankruptcy record is more likely to default.



Bankruptcy Records

Segmented Univariate Analysis

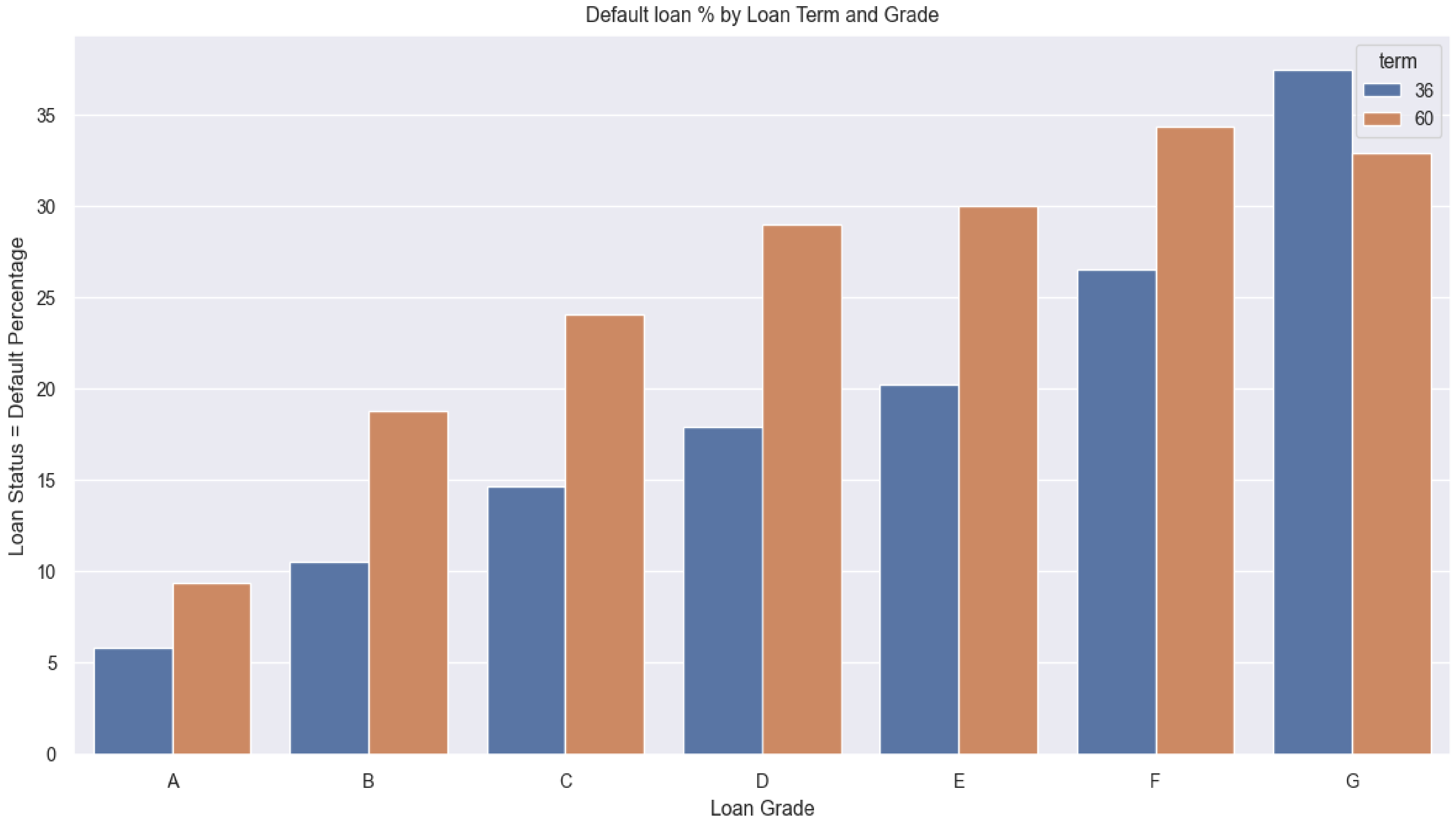
States



- Applicants from Nevada and South Dakota has higher default ratio.
- ***Most populous states in US like California, Florida,Texas has more default applicants. But this is also because there were possibly more number of applicants in first place.***

Bivariate Analysis

Loan Term to Loan Grade



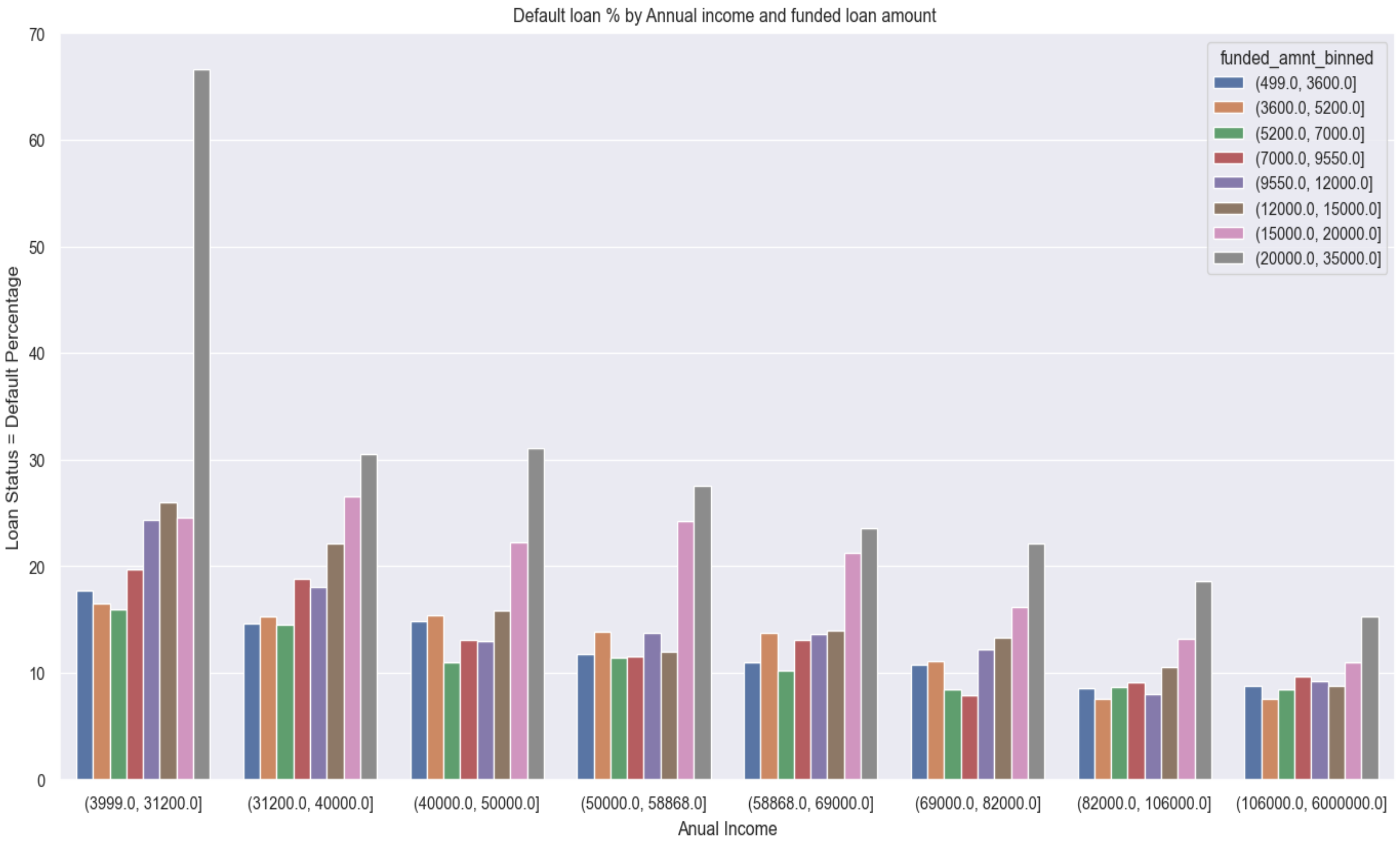
- ❑ Loans with Grade A and B irrespective of loan terms and grade C and D with term of 36 months are likely to repay.

- ❑ Other lower grades loans with any loan terms are relatively risky.

Bivariate Analysis

Annual Income to Funded Loan Amount

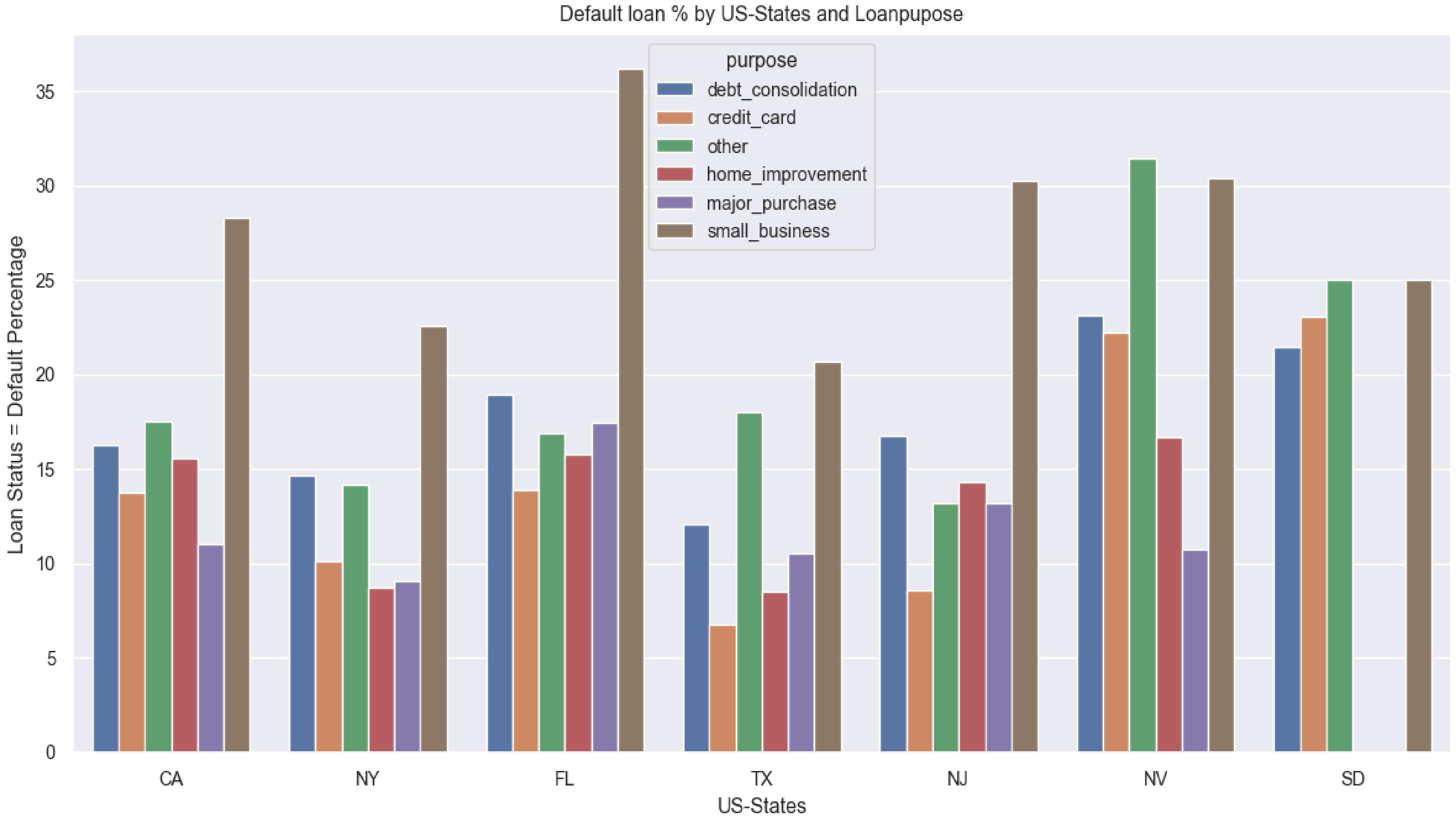
- Funded loan amount of greater than ***20000 USD*** are most likely to default across all income ranges.
- Applicants with less than ***70000 USD/year*** income are more likely to default on loan amount ***>=15000 USD.***



Bivariate Analysis

States to Loan Purpose

- Across all the top states applicants who have ***loaned the money for small business are most likely to default.***
- Nevada is an exception where ***"Others"*** category is actually higher - but we don't know what this refers to.



Conclusion – Recommendation for Approving future loans

1. ***'Lending Club'** should **Stop** – approving loans where amount/income i.e. **'dti'** is higher than **30%.***

2. ***Reduce** – number of approvals **where purpose is small business** ;these are more likely to default with **ratio > 25%.***
 *Loan purpose of car, wedding, major purchase and credit are **having 90% of probability of repayment**.*

3. ***Stop** – approving loans to people with **prior bad record**. Or at least stop approving high-value loans.*

4. *Funded loan amount above **15,000 USD for customer with < 70,000 USD annual income is risky and should be avoided**. Customers with more than **10,000 USD annual income are having more than 85% of probability of repayment.***

5. *Loans with **Grade A and B irrespective of loan terms and grade C and D with term of 36 months are likely to repay**. Other lower grades loans **(Grade E,F, G) with any loan terms are relatively risky.***