INFSCI - 2725 (Data Analytics)

Assignment 1

Create a clear problem statement that this data is going to address. (You don't need to solve the problem, but the data needs to appear to have enough information to address the problem.)

1. Take your dataset (at least 10 different variables and 500 observations). Select 10 variables (text and numeric) to analyze.

2. Write a text description of the dataset (1 paragraph), a description of why you selected the variables you selected (what question is that collection of variables going to answer) (1 or 2 paragraphs).

3. Describe the range of values for each variable (can be a table).

4. Describe how the variables relate to each other, correlations for numeric variables, crosstabs for textual variables, conditional probability distributions for text by numeric variables (e.g., for females describe variable x, for males describe variable y, does y differ by gender?)

5. Can have plots if they show interesting information.

(Not more than 5 pages in total.)

Submitted by:

Sl. No.	Name	Email Id	Student ID
1	Piu Mallick	pim16@pitt.edu	4374215
2	Shruti Gupta	shg104@pitt.edu	4374956
3	Debdas Ghosh	deg107@pitt.edu	4366821

Background/Introduction to our problem

Since we will be focusing and analyzing on the dataset of the **LendingClub**, we need to know a little bit about the organization.

"**LendingClub** is the world's largest online marketplace for connecting the borrowers and the investors. Through personal loans, auto-financing loans, business loans and many other kinds of loans, **Lending Club** offers the borrowing and investing solutions for the customers."

The organization enables borrowers to create unsecured personal loans between *$1,000* and *$40,000*. The standard loan period is three years. Investors can search and browse the loan listings on **LendingClub** website and select loans that they want to invest in based on the information supplied about the borrower, amount of loan, loan grade, and loan purpose.
Definition Source: Lending Club Article, Wikipedia

Problem Statement: Predicting the probability that a loan will **charge off*,** or the loan is going to turn as a 'bad loan'.

* *"A **charge-off** is a debt, for example on a credit card or for personal purpose, that is deemed unlikely to be collected by the creditor because the borrower becomes substantially delinquent after a period of time. However, a charge-off does not mean a write-off of the debt entirely."*

Link to Data Set that is going to be used for analysis purpose: Data is here.

Description of the data: The data file contains complete loan data for all loans issued by the LendingClub, including the current loan status (Current, Late, Fully Paid, etc.) and latest payment information.
The file is a matrix of about 1646801 thousand observations and 150 variables.

Some of the mentionable variables in the dataset are:

`loan_amnt, purpose, annual_inc, emp_title, emp_length, term, int_rate, addr_state, verification_status, dti, bc_util, term`

However, for the assignment, we would be considering 10 variables for analysis purpose.

The **variables** that we would be considering for analysis are:

a. **`loan_amnt`** (*Loan Amount*): The loan amount asked by the borrower is key factor in determining whether the loan would be charged-off by **LendingClub**. The higher the loan amount, thee greater is the risk in paying off the loan.

b. **`purpose`** (*Purpose of the loan*): The purpose of the loan amount plays an important factor too. The various reasons/purpose for applying loan might be – credit card payoff, re-financing, home buying/improving, medical expenses, etc. If the loan is taken to pay off other loans or pay off credit card bills, chances are there that it might result in a bad loan.

c. **`annual_inc`** (*Annual Income*): The annual income of the borrower is one of the prime factors in determining the loan amount issued to the borrower. The higher the annual income, chances are greater that loan would be paid off. If the annual income is on the lower side, probabilities are more that the loan might get charged off.

d. **`emp_title`** (*Employee Title*): The loan amount dispersed by Lending Club is also determined by the employee title or designation in the organization.

e. **`emp_length`** (*Employment Length*): This variable determines the tenure of the borrower employed in a company. If a person is employed for a lesser time, it means he/she might take some time to get stabilized and it might get riskier. Also, if a person is nearing his retirement age, lending loan to him/her can be riskier as he/she might not have enough time to repay the loan. Employment length in the range of greater or equal to 3 years can be considered as a good indicator for lending loans.

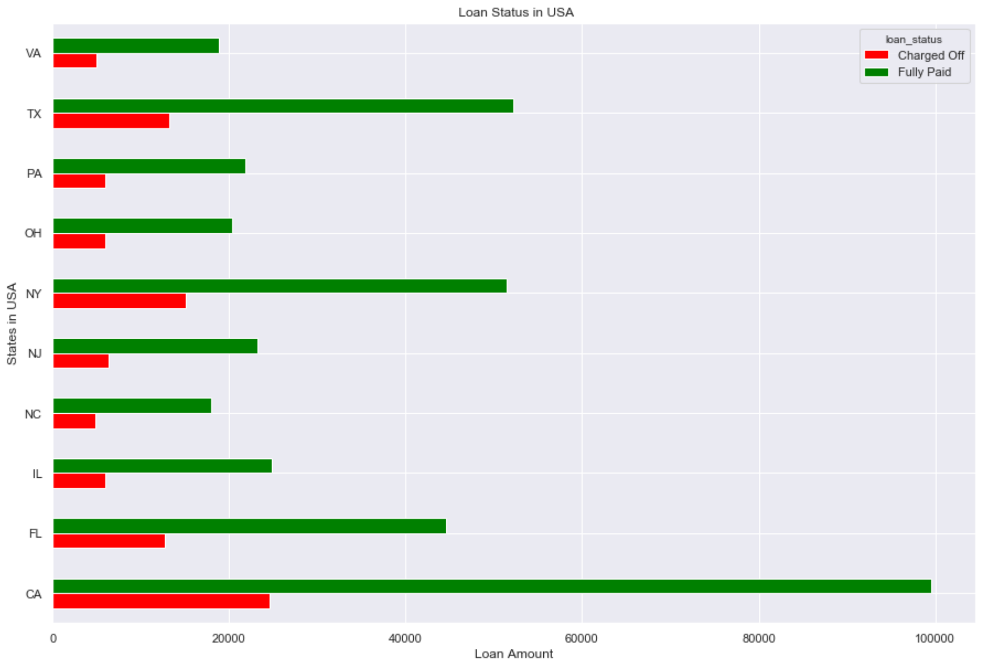
f. **`term`** (*Term of the loan, in months*): in how much time the loaner will repay the loan. The lesser the better. If the rate of interest is high and term is also long, then there are high chances that it will result in bad loan. Below is the data based on the dataset, lending club offers only two options in terms, 36 months and 60 months.

term	**loan_status**	**% of population**
36 months	Fully Paid	0.83429
	Charged Off	0.16571
60 months	Fully Paid	0.666207
	Charged Off	0.333793

g. **`int_rate`** (*Interest Rate against the loan*): Interest rate is calculated by adding the base interest rate and the sub grade. Higher grade means higher interest rate, which implies that it tends to bad loans. Hence, the higher value of the variable contributes to the factor that the loans would be charged off.

h. **`addr_state`** (*State from which loan is applied*): The dataset contains values from various states of USA. There are some states like California, New York, Texas, etc. which are expensive for living. So, the place/address plays a deciding factor for the loan to turn bad.

A bar chart (prepared based on the data set) proves that the same thing – the percentage of loans going bad is high where the cost of living is high.



i. **verification_status** (*Verification status of the loan*): If a person's information is not verified, he/she is most likely to default the loan, as there is no guarantee of the authenticity of his/her information.

j. **dti** (*Debt Payment to Income*): DTI is a comparison of a borrower's monthly debt payments with monthly income. The calculation is simple: total monthly debt divided by total monthly income equals DTI. The lower the DTI, the better.

k. **bc_util:** It is the ratio of total current balance to high credit/credit limit for all bankcard accounts. This variable determines that if the credit limit is too high and the current balance is low, then it is possible that the person will spend more than his/her limits, which implies that he/she has credit card debts.

(Statistical) Range of the values:

The statistical range of the selected values are shown below:

Numeric Values

Variable Name	Count	Mean	Std	Min	25%	50%	75%	Max
loan_amnt	814986	14315.45821	8499.799241	500	8000	12000	20000	40000
annual_inc	814986	75230.39	65243.73	0	45000	65000	90000	9550000
dti	814950	17.867719	8.856477	-1	11.64	17.36	23.63	999
int_rate	814986	13.490993	4.618486	5.32	9.99	13.11	16.29	30.99
bc_util	759321	61.575664	27.87117	0	40.8	65.4	86	339.6

Text Values

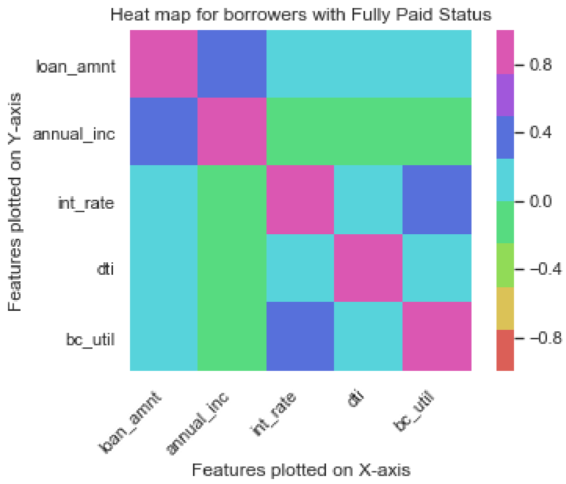
Variable	Count	Unique	Top	Frequency
purpose	814986	14	debt_consolidation	481652
emp_title	766415	280473	Teacher	11351
emp_length	772733	11	10+ years	165
addr_state	814986	51	CA	124204
verification_status	814986	3	Source Verified	293897
term	814986	2	36 months	618460

How the variables relate to each other?

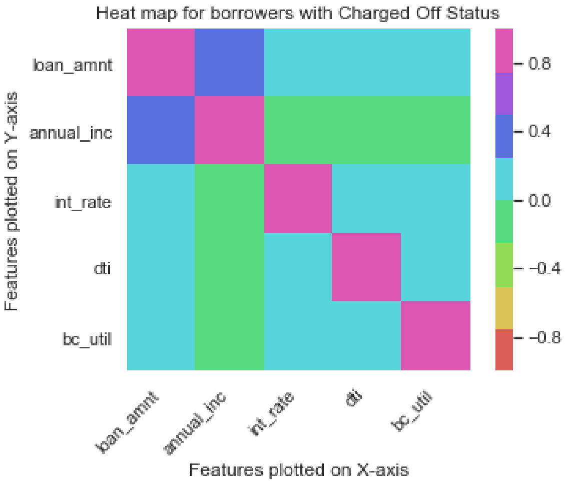
The dataset used for the analysis has been split out into two categories of the loan status - "**Fully Cleared**" and "**Charged Off**".

The correlation of two different subsets of data are shown below:

Plot for "Fully Paid" Status



Plot for "Charged Off" Status



We could see that the differentiating factor between the two figures is the correlation between the variables `int_rate` and `bc_util`.

There is no correlation between `int_rate` and `bc_util` in the Charged Off case, which means people who have less `bc_util` have less chances to get a good interest rate (less interest rate), hence they might fall under the category of people whose loan may turn bad.