Descriptive Analysis of Bank Loan Data

Cheng Peng

Contents

1 Introduction

In this brief report, we will summarize the working data set collected from banks that received loan applications from small businesses with a partial warranty from the Small Business Association (SBA). The data is in a public domain and can be found at https://pengdsci.github.io/datasets/#sba-loan. We use the free open-source R (http://www.r-project.com) to perform descriptive analysis.

```
loan01 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational01.csv", header = TRUE)[,
loan02 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational02.csv", header = TRUE)[,
loan03 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational03.csv", header = TRUE)[,
loan04 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational04.csv", header = TRUE)[,
loan05 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational05.csv", header = TRUE)[,
loan06 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational06.csv", header = TRUE)[,
loan07 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational07.csv", header = TRUE)[,
loan08 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational08.csv", header = TRUE)[,
loan09 = read.csv("https://pengdsci.github.io/datasets/SBAloan/w06-SBAnational09.csv", header = TRUE)[,
loan = rbind(loan01, loan02, loan03, loan04, loan05, loan06, loan07, loan08, loan09)
# dim(bankLoan)
#names(bankLoan)
```

You can use the information in the description that is given in the document on the web page .....

```
var.names = names(loan)
my.var = var.names[c(1,6,21,26)]
my.new.data = loan[1:15, my.var]
dim(my.new.data)
```

```
[1] 15  4
```

Based on the information given in the data set description, we can create a basic RMD table.

Variable Name	Type	Description
LoanNr_ChkDgt	Text	Identifier – Primary Key
Name	Text	Borrower Name
City	Text	Borrower City
State	Text	Borrower State
Zip	Text	Borrower Zip Code
Bank	Text	Bank Name
BankState	Text	Bank State
NAICS	Text	North American Industry Classification System code
ApprovalDate	Date	Date SBA Commitment Issued
ApprovalFY	Time	Fiscal Year of Commitment
Term	Numeric	Loan term in months
NoEmp	Numeric	Number of Business Employees
NewExist	Text	1 = Existing Business, 2 = New Business
CreateJob	Numeric	Number of jobs created
RetainedJob	Numeric	Number of jobs retained
FranchiseCode	Text	Franchise Code 00000 or 00001 = No Franchise
UrbanRural	Text	1= Urban, 2= Rural, 0 = Undefined
RevLineCr	Text	Revolving Line of Credit : Y = Yes
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date	The date when a loan is declared to be in default
DisbursementDate	Date	Disbursement Date
DisbursementGross	Numeric	Amount Disbursed
BalanceGross	Numeric	Gross amount outstanding
MIS_Status	Text	Loan Status
ChgOffPrinGr	Numeric	Charged-off Amount
GrAppv	Numeric	Gross Amount of Loan Approved by Bank
SBA_Appv	Numeric	SBA's Guaranteed Amount of Approved Loan
New	Text	=1 if NewExist=2 (New Business), =0 if NewExist=1 (Existing Business)
Portion	Numeric	Proportion of Gross Amount Guaranteed by SBA
RealEstate	Text	=1 if loan is backed by real estate, =0 otherwise
Recession	Text	=1 if loan is active during Great Recession, =0 otherwise
Selected	Text	=1 if the data are selected as training data to build model for assignment, =0 if the data are selected as testing data to validate model
Default	Text	=1 if MIS_Status=CHGOFF, =0 if MIS_Status=PIF (pay in full)

One nice feature of this method is that you have parameters you can control to make a better good looking graphic. For more information on embedding images in RMD, please visit https://yihui.org/knitr/options/.

In the next sections, we will perform frequency analysis for character variables and descriptive numeric measures for numerical variables.

2 Variable Inspection

I use an informal term *variable inspection* in this note to have a glance at the distribution of the individual variables of interest to be used in the analysis. We only want to use this data for sampling purposes for now. A categorical variable is needed to do stratified sampling.

2.1 Frequency Analysis for Category Variables

There are AAA categorical variables in the data set, variables XXX, YYY, and ZZZ are related to the geographic regions, the type of industry, and the size of the business. These variables can be used to define independent sub-populations.

Some of the categories may have only a small number of loans, we should combine small sub-populations in a **meaningful** way so that we can use the "balanced" variable for stratification.

For example, we can look at the variable NAICS. Since R is case sensitive, before you write code to select a variable for analysis, you may want to see the exact form of the variable name in your data set to avoid unnecessary frustration.

```
name.vec=names(loan)                                  # this is a character vector
name.mtx = matrix(name.vec, ncol=3)                   # define a matrix and then make a table
colnames(name.mtx)=c("variable name", "variable name", "variable name" ) # assign column names: This is
kable(name.mtx)                                       # make a nice looking table: kable() isa fun function in
```

variable name	variable name	variable name
LoanNr_ChkDgt	ApprovalFY	LowDoc
Name	Term	ChgOffDate
City	NoEmp	DisbursementDate
State	NewExist	DisbursementGross
Zip	CreateJob	BalanceGross
Bank	RetainedJob	MIS_Status
BankState	FranchiseCode	ChgOffPrinGr
NAICS	UrbanRural	GrAppv
ApprovalDate	RevLineCr	SBA_Appv

Next, I use the North America Industry Classification System (NAICS) as an example to demonstrate how to obtain a frequency table from R and what information you need to know in designing a stratified sampling. The following code chunk produces a frequency distribution (table).

```
naics =as.character(loan$NAICS)  # make a character vector
N=length(naics)                  # find the size of the data.
f.table = -sort(-table(naics))   # sort the vector in descending order
n = length(f.table)              # find the number of distinct industries
n.0 = sum(f.table < 900)         # industry with less than 0.1% of the population size
# A note of length of R variable: the latest version of R has upped
# the maximum length of variable names from 256 characters to a whopping 10,000.
# We should try our best to give meaningful names to R variables.
kable(cbind(Population.size = N, Number.of.Industries=n, Sub.Pop.less.1000 = n.0))
```

Population.size	Number.of.Industries	Sub.Pop.less.1000
899164	1312	1140

In the above frequency distribution table, we observe the following

- About 87% of NAICS classified industries have less than 900 applications. If the sample size is not large enough, 87% of the industries in the population will be unlikely represented in an SRS sample.
- It makes sense to combine these small sub-populations with a larger similar population.
- It also makes sense to define a study population by including a set of sub-populations with a certain size.

2.2 Summary of Numerical Variables

In general, descriptive analyses of the numerical variable mainly focus on the distribution of the variable. For example, for a multi-modal or an extremely skewed variable, we may want to discretize it and to make a discrete variable for analysis.

We can also discretize a numerical variable based on the distribution to make a stratification variable.

3 Initial Data Management

We will do some basic data management to create a data set that will be used as a study population. We can add a variable to the data set and used a sampling list. Some R functions are convenient to use for setting data based on observations (not variable!).

3.1 Add New Variable to a Dataframe

Here is an example. I would add an ID variable to the existing data frame so that you can use it as the sampling frame in the SRS sampling scheme.

```
pop.N = length(loan$LoanNr_ChkDgt)          # population size
loan$obs.ID = 1:pop.N                       # adding an observation ID to the dataframe
# Next, we choose first 10 row of a subset of variables to make a table for inspection.
kable(as.matrix(loan[1:10, c("obs.ID", "LoanNr_ChkDgt", "Name", "City", "State", "Bank")]))
```

obs.ID	LoanNr_ChkDgt	Name	City	State	Bank
1	1000014003	ABC HOBBYCRAFT	EVANSVILLE	IN	FIFTH THIRD BANK
2	1000024006	LANDMARK BAR & GRILLE (THE)	NEW PARIS	IN	1ST SOURCE BANK
3	1000034009	WHITLOCK DDS, TODD M.	BLOOMINGTON	IN	GRANT COUNTY STATE BANK
4	1000044001	BIG BUCKS PAWN & JEWELRY, LLC	BROKEN ARROW	OK	1ST NATL BK & TR CO OF BROKEN
5	1000054004	ANASTASIA CONFECTIONS, INC.	ORLANDO	FL	FLORIDA BUS. DEVEL CORP
6	1000084002	B&T SCREW MACHINE COMPANY, INC	PLAINVILLE	CT	TD BANK, NATIONAL ASSOCIATION
7	1000093009	MIDDLE ATLANTIC SPORTS CO INC	UNION	NJ	WELLS FARGO BANK NATL ASSOC
8	1000094005	WEAVER PRODUCTS	SUMMERFIELD	FL	REGIONS BANK
9	1000104006	TURTLE BEACH INN	PORT SAINT JOE	FL	CENTENNIAL BANK
10	1000124001	INTEXT BUILDING SYS LLC	GLASTONBURY	CT	WEBSTER BANK NATL ASSOC

3.2 Accessing Observations Meeting Certain Conditions

Setting a data set based on variables is straightforward. you need the following command to define a subset with only the listed 6 variables.

```
loan[ , c("obs.ID", "LoanNr_ChkDgt", "Name", "City", "State", "Bank")]
```

If we define a subset with conditions in one or more variables, we can use a powerful function *which*(). The following example defines a subset

```r
subset.id = which(loan$Bank=="FIFTH THIRD BANK" | loan$Bank=="CENTENNIAL BANK") # "|" means "or".
centennial.5th3rd = loan[subset.id, ]
sizes = dim(centennial.5th3rd)
mtx=cbind(obs.n = sizes[1], var.n = sizes[2])
kable(mtx)
```

obs.n	var.n
5628	28

3.3 Random Sampling

As you have already seen that SRS is the foundation of probabilistic sampling plans. There are two types of SRS: SRS with and without replacement. The built-in R function *sample*() can do both with and without replacement plans. It can also do weighted sampling.

```
sample(x, size, replace = FALSE, prob = NULL)
# x = vector
# size = sample size
# replace = FALSE, without replacement
# replace = TRUE, with replacement
# prob =  a vector of probabilities (weights)
          of corresponding observations
          to do weighted sampling.
```

4 Some Useful Tips

4.1 Embed External Images Into RMD

For example, you can make an image of the variable list and embed it into this document. To do this, create an image and save it in a folder where you save this R markdown document. You then can use the following R code chunk to embed the image into a file in the format of html, word, or pdf when you knit this RMD document.

Variable Name	*Data Type*	*Description of variable*
LoanNr_ChkDgt	Text	Identifier – Primary Key
Name	Text	Borrower Name
City	Text	Borrower City
State	Text	Borrower State
Zip	Text	Borrower Zip Code
Bank	Text	Bank Name
BankState	Text	Bank State
NAICS	Text	North American Industry Classification System code
ApprovalDate	Date/Time	Date SBA Commitment Issued
ApprovalFY	Text	Fiscal Year of Commitment
Term	Number	Loan term in months
NoEmp	Number	Number of Business Employees
NewExist	Text	1 = Existing Business, 2 = New Business
CreateJob	Number	Number of jobs created
RetainedJob	Number	Number of jobs retained
FranchiseCode	Text	Franchise Code 00000 or 00001 = No Franchise
UrbanRural	Text	1= Urban, 2= Rural, 0 = Undefined
RevLineCr	Text	Revolving Line of Credit : Y = Yes
LowDoc	Text	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	Date/Time	The date when a loan is declared to be in default
DisbursementDate	Date/Time	Disbursement Date
DisbursementGross	Currency	Amount Disbursed
BalanceGross	Currency	Gross amount outstanding
MIS_Status	Text	Loan Status
ChgOffPrinGr	Currency	Charged-off Amount
GrAppv	Currency	Gross Amount of Loan Approved by Bank
SBA_Appv	Currency	SBA's Guaranteed Amount of Approved Loan
New	Number	=1 if NewExist=2 (New Business), =0 if NewExist=1 (Existing Business)
Portion	Number	Proportion of Gross Amount Guaranteed by SBA
RealEstate	Number	=1 if loan is backed by real estate, =0 otherwise
Recession	Number	=1 if loan is active during Great Recession, =0 otherwise
Selected	Number	=1 if the data are selected as training data to build model for assignment, =0 if the data are selected as testing data to validate model
Default	**Number**	**=1 if MIS_Status=CHGOFF, =0 if MIS_Status=P I F**

Figure 1: List of all variables and the description of each variable