Small Business Loan Exploratory Data Analysis

Heather Qiu, Aditya John, Isha Singh, Jenny Shen

2022-10-21

Data Overview

The original dataset is from U.S. Small Business Administration (SBA), a governmental agency founded in 1953 to provide financial assistance to small enterprises in the U.S. credit market. The file contains 899,164 loan records from small businesses across the country and 27 variables. A detailed description of all variables can be found in Table 1. For the project, our primary research interests can be summarized into the following questions:

1. How does the requested loan duration (***Term***), revolving line of credit (***RevLineCr***), and employee count (***NoEmp***) of a small business impact the amount of the approved SBA loan (***GrAppv***)?
2. What are the different factors that impact the defaulting of a loan?

In addition, there are complexities in the original dataset to be aware of:

- Missing Values. Given the size of the dataset, it is not uncommon to find missing values on some variables. Table 2 shows the count of missing values for all variables in the original file. For the scope of the current project, we will exclude loan records with missing values except for the variable, ***ChgOffDate***, which indicates a date when a loan is declared to be in default. Given the rigorous review procedure that banks have set in place, only a small portion of loans are expected to default. Therefore, loans in good standing or have been paid off won't indicate a date when the loan is declared to be in default (***ChgOffDate***).

- Inconsistent Data Format. The following categorical variables contain values other than yes and no. Without further clarification from the original dataset, we will exclude loan records beyond the binary choices.

    - Revolving Line of Credit (***RevLineCr***)
    - Business Status: New or Existing Businesses (***NewExist***)
    - Business Locale: Rural or Urban (***UrbanRural***)
    - LowDoc Loan Program (***LowDoc***)
- Incorrect Data Type. The following variables that contain monetary values are represented as strings rather than numeric numbers in the original file. These variables will be converted to the numeric data type for our analysis.

    - SBA's Guaranteed Amount of Approved Loan (***SBA_Appv***)
    - Gross Amount of Loan Approved by Bank (***GrAppv***)
    - Charged-off Amount (***ChgOffPrinGr***)
    - Gross Amount Outstanding (***BalanceGross***)
    - Amount Disbursed (***DisbursementGross***)
- Data Distribution. For the following variables, the distribution is skewed to either the left or the right. In addition, there are significant outliers that require attention. Their impact can be minimized by either performing data transformations or merely excluding these data points from the file before modeling. Categorical variables have factor imbalances (i.e., more occurrences of one factor compared to the rest).

    - Gross Amount of Loan Approved by Bank (***GrAppv***)
    - Number of Business Employees (***NoEmp***)

- Urban/Rural Columns (***UrbanRurac_fac***)

Primary Relationship of Interest

For research question number one, we first looked at the distribution of the response variable ***GrAppv_num***. According to the histogram and the density plot (Figure1), the response variable is not normally distributed. After transforming the variable by applying the natural log, the distribution of the ***GrAppv_num*** looks normal now.

We then explored the relationship between ***GrAppv_num*** and each predictor. We applied scatter plots to continuous/numeric predictors while using boxplots to explore the categorical predictors (Figure2 and Figure3 in the Appendix). The graphs indicate a non-linear relationship between the response variable and predictors ***Term*** and ***NoEmp***. There are also differences in mean for the gross amount of loan approved by bank vs. revolving line of credit, which shows the relation between ***GrAppv_num*** and ***RevLineCr_fac*** is worth considering.

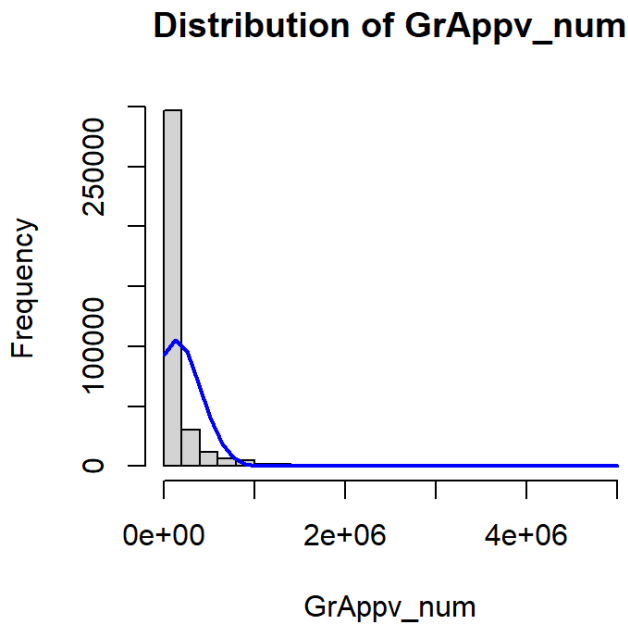
For the second research question, Tables 4 through 7 and Figure 4 in the Appendix present the descriptive statistics. We have summarized the key findings below:

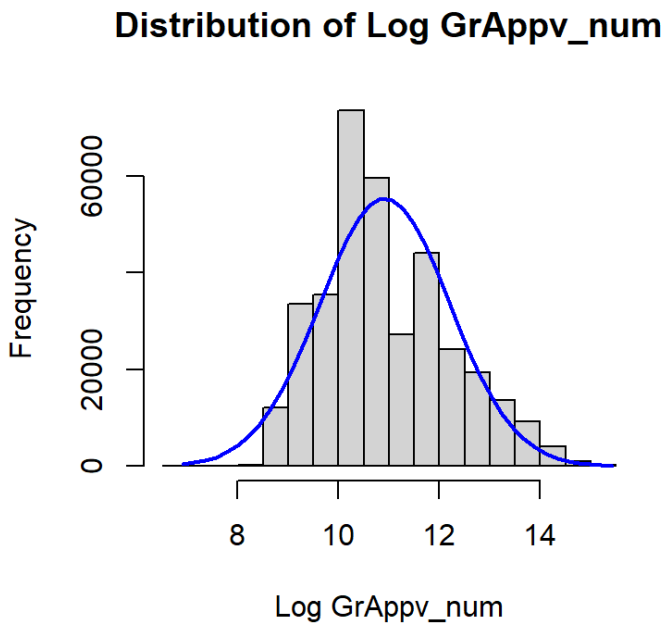
- Majority of the establishments applying for small business loans are existing businesses that operate in urban settings.
- The default rate varies between states. California, New York, and Florida have the highest percentage of small businesses defaulting on their loans.
- 77% of small business loans in default occurred between 2005 and 2008. However, this is likely because about 76% of loans were distributed around the same time.

Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Term	357597	82.828	59.081	0	50	84	527
RevLineCr_fac	357597						
… N	167293	46.8%					
… Y	190304	53.2%					
NoEmp	357597	8.344	31.757	0	2	8	8000
GrAppv_num	357597	139403.767	269915.621	1000	25000	120000	5e+06

Figure 1.Distribution of Gross Amount of Loan Approved by Bank





Other Characteristics

For the first research question:

- From the summary statistics, it is observed that ***NoEmp*** is right-skewed.
- Most loan terms in months are shorter than 100 months.
- The ratio of whether the loan is approved depends almost equally (46.8% v.s. 53.2%) on whether the business has a revolving line of credit.

For the second research question:

- From Table 4, we can observe that small businesses that create fewer jobs are generally at a higher risk of default.
- On the other hand, companies are less likely to default when the number of jobs retained is high.
- Almost all of the small businesses in this analysis are not involved in the LowDoc Loan Program.

Potential Challenges

The following are the challenges identified when performing EDA on the dataset.

1. The first challenge is the exclusion of loan record values beyond the binary (Y/N) choices that were showing up within binary variables. The reason for exclusion is that we lack information about what the different values represent. By retaining only Y/N values, we face the challenge of information loss.
2. The second challenge is that a handful of variables are in line with the outcome variable, balance gross and charge-off amount. These variables have to be excluded from the analysis as they are all bi-variate outcomes.
3. The dataset was too large, causing runtime inefficiency in R.
4. Due to the uncleaned data, we had to remove empty values and data that did not make sense in context. If we could have talked to the original data collector, our analysis would likely be more comprehensive.

Appendix

Code Book

Variable	Description
LoanNr_ChkDgt	Identifier Primary Key
Name	Borrower Name
City	Borrower City
State	Borrower State
Zip	Borrower Zip Code
Bank	Bank Name
BankState	Bank State
NAICS	North American Industry Classification System Code
ApprovalDate	Date SBA Commitment Issued
ApprovalFY	Fiscal Year of Commitment
Term	Loan Term in Months
NoEmp	Number of Business Employees
NewExist	1 = Existing Business, 2 = New Business, 0 = Undefined
CreateJob	Number of Jobs Created
RetainedJob	Number of Jobs Retained
FranchiseCode	Franchise Code, (00000 or 00001) = No Franchise
UrbanRural	1 = Urban, 2 = Rural, 0 = Undefined
RevLineCr	Revolving Line of Credit: Y = Yes, N = No
LowDoc	LowDoc Loan Program: Y = Yes, N = No
ChgOffDate	The date when a loan is declared to be in default
DisbursementDate	Disbursement Date
DisbursementGross	Amount Disbursed

Variable	Description
BalanceGross	Gross Amount Outstanding
MIS_Status	Loan Status Charged off = CHGOFF, Paid in Full =PIF
ChgOffPrinGr	Charged-off Amount
GrAppv	Gross Amount of Loan Approved by Bank
SBA_Appv	SBA's Guaranteed Amount of Approved Loan

Missing Value Overview

Variable	Count of Missing Values
LoanNr_ChkDgt	0
Name	8
City	30
State	14
Zip	0
Bank	1559
BankState	1566
NAICS	0
ApprovalDate	0
ApprovalFY	0
Term	0
NoEmp	0
NewExist	136
CreateJob	0
RetainedJob	0
FranchiseCode	0
UrbanRural	0

Variable	Count of Missing Values
RevLineCr	4528
LowDoc	2582
ChgOffDate	736465
DisbursementDate	2368
DisbursementGross	0
BalanceGross	0
MIS_Status	1997
ChgOffPrinGr	0
GrAppv	0
SBA_Appv	0

Figure 2.Scatter Plot of Gross Amount of Loan Approved by Bank and Loan term in months/Number of business employees

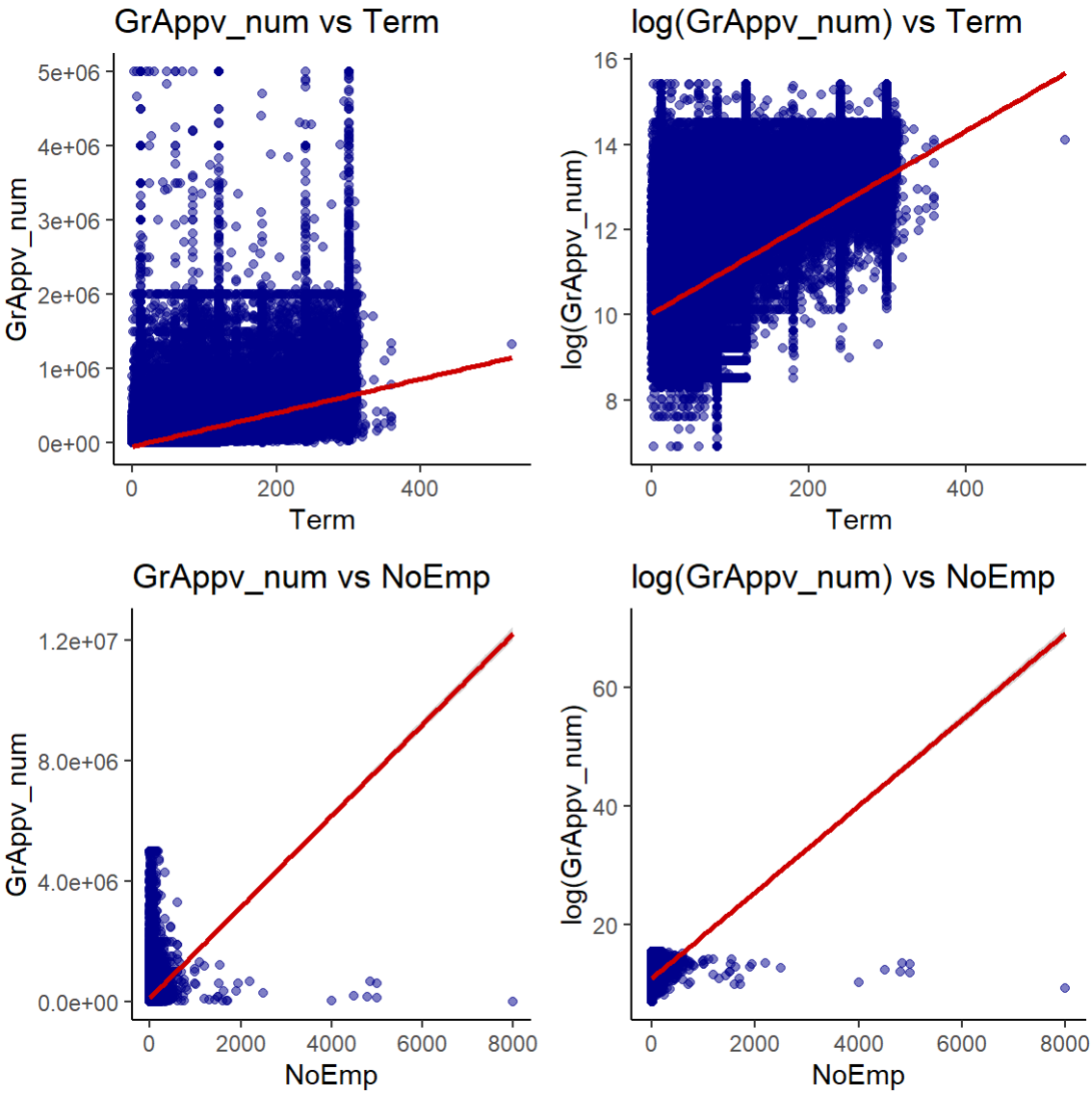


Figure 3.Comparison of Revolving Line of Credit in the Gross Amount of Loan Approved by Bank

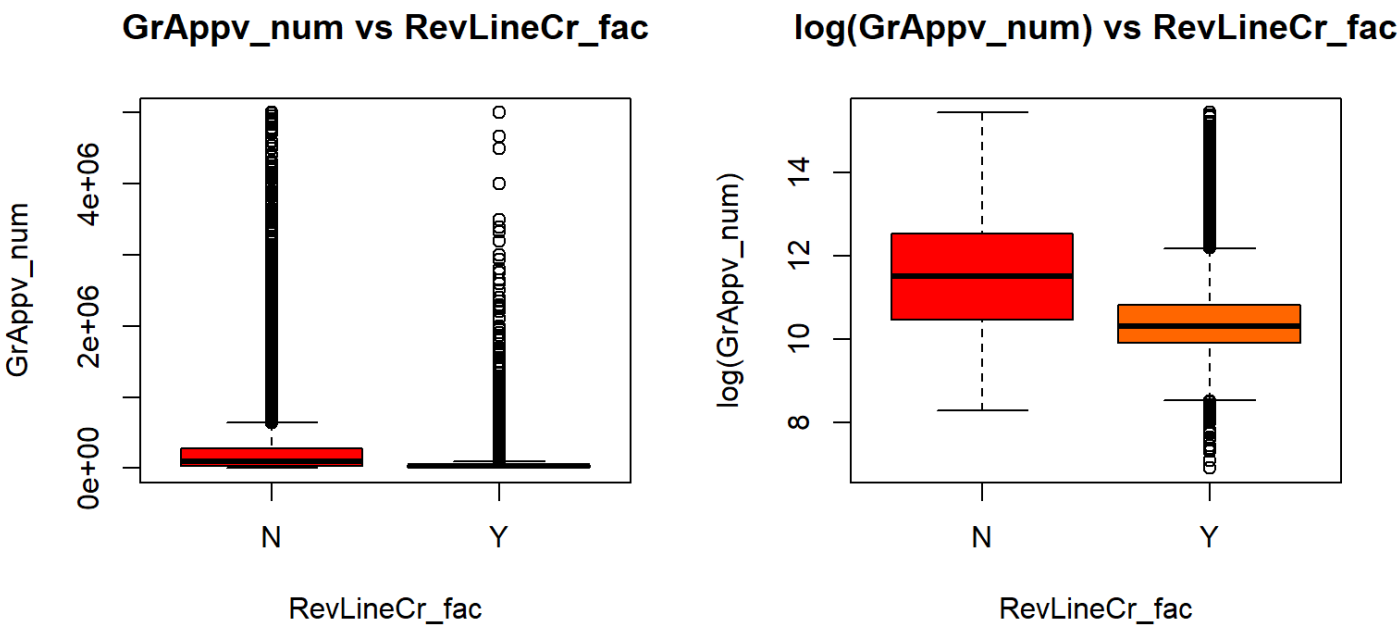


Table 4.Descriptive of Small Business Loans by Loan Status (I)

	CHGOFF (N=94522)	P I F (N=263075)	Overall (N=357597)
Term			
Mean (SD)	48.4 (34.3)	95.2 (61.2)	82.8 (59.1)
Median [Min, Max]	46.0 [0, 339]	84.0 [0, 527]	84.0 [0, 527]
NoEmp			
Mean (SD)	5.81 (35.8)	9.25 (30.1)	8.34 (31.8)
Median [Min, Max]	3.00 [0, 8000]	4.00 [0, 5000]	3.00 [0, 8000]
NewExist_fac			
Existing Business	69692 (73.7%)	191552 (72.8%)	261244 (73.1%)
New Business	24830 (26.3%)	71523 (27.2%)	96353 (26.9%)
CreateJob			
Mean (SD)	1.73 (13.3)	2.14 (14.7)	2.03 (14.3)
Median [Min, Max]	0 [0, 1620]	0 [0, 5090]	0 [0, 5090]
RetainedJob			
Mean (SD)	4.80 (16.8)	5.99 (21.2)	5.68 (20.2)
Median [Min, Max]	3.00 [0, 4440]	2.00 [0, 7250]	2.00 [0, 7250]
UrbanRural_fac			
Urban	81242 (86.0%)	211441 (80.4%)	292683 (81.8%)
Rural	13280 (14.0%)	51634 (19.6%)	64914 (18.2%)
RevLineCr			
N	44385 (47.0%)	122908 (46.7%)	167293 (46.8%)
Y	50137 (53.0%)	140167 (53.3%)	190304 (53.2%)
LowDoc			
N	94294 (99.8%)	261742 (99.5%)	356036 (99.6%)
Y	228 (0.2%)	1333 (0.5%)	1561 (0.4%)
SBA_Appv_num			
Mean (SD)	52200 (123000)	120000 (245000)	102000 (222000)
Median [Min, Max]	20000 [500, 3410000]	25000 [500, 4500000]	25000 [500, 4500000]
GrAppv_num			
Mean (SD)	80400 (165000)	161000 (296000)	139000 (270000)
Median [Min, Max]	35000 [1000, 3500000]	50000 [1000, 5000000]	50000 [1000, 5000000]
ChgOffPrinGr_num			
Mean (SD)	58000 (113000)	162 (3030)	15500 (63300)
Median [Min, Max]	28100 [0, 2220000]	0 [0, 634000]	0 [0, 2220000]
BalanceGross_num			
Mean (SD)	0 (0)	7.66 (2530)	5.63 (2170)
Median [Min, Max]	0 [0, 0]	0 [0, 996000]	0 [0, 996000]

	CHGOFF (N=94522)	P I F (N=263075)	Overall (N=357597)
DisbursementGross_num			
Mean (SD)	102000 (171000)	184000 (315000)	163000 (286000)
Median [Min, Max]	50000 [4000, 4360000]	72900 [4000, 11400000]	63900 [4000, 11400000]

Table 5.Descriptive of Small Business Loans by Loan Status (II)

	CHGOFF (N=94522)	P I F (N=263075)	Overall (N=357597)
State			
AK	73 (0.1%)	425 (0.2%)	498 (0.1%)
AL	723 (0.8%)	1619 (0.6%)	2342 (0.7%)
AR	454 (0.5%)	1385 (0.5%)	1839 (0.5%)
AZ	2471 (2.6%)	4375 (1.7%)	6846 (1.9%)
CA	15935 (16.9%)	32327 (12.3%)	48262 (13.5%)
CO	2315 (2.4%)	5466 (2.1%)	7781 (2.2%)
CT	1044 (1.1%)	4683 (1.8%)	5727 (1.6%)
DC	156 (0.2%)	495 (0.2%)	651 (0.2%)
DE	243 (0.3%)	661 (0.3%)	904 (0.3%)
FL	7465 (7.9%)	12696 (4.8%)	20161 (5.6%)
GA	3039 (3.2%)	4947 (1.9%)	7986 (2.2%)
HI	251 (0.3%)	923 (0.4%)	1174 (0.3%)
IA	474 (0.5%)	2341 (0.9%)	2815 (0.8%)
ID	855 (0.9%)	3061 (1.2%)	3916 (1.1%)
IL	4346 (4.6%)	9193 (3.5%)	13539 (3.8%)
IN	1504 (1.6%)	5033 (1.9%)	6537 (1.8%)
KS	528 (0.6%)	2331 (0.9%)	2859 (0.8%)
KY	728 (0.8%)	2351 (0.9%)	3079 (0.9%)
LA	713 (0.8%)	1957 (0.7%)	2670 (0.7%)
MA	2260 (2.4%)	10392 (4.0%)	12652 (3.5%)
MD	1402 (1.5%)	4084 (1.6%)	5486 (1.5%)
ME	316 (0.3%)	2001 (0.8%)	2317 (0.6%)
MI	3287 (3.5%)	7148 (2.7%)	10435 (2.9%)
MN	1651 (1.7%)	7205 (2.7%)	8856 (2.5%)
MO	1491 (1.6%)	4910 (1.9%)	6401 (1.8%)
MS	629 (0.7%)	2487 (0.9%)	3116 (0.9%)
MT	224 (0.2%)	2081 (0.8%)	2305 (0.6%)
NC	1397 (1.5%)	4051 (1.5%)	5448 (1.5%)
ND	124 (0.1%)	1384 (0.5%)	1508 (0.4%)
NE	270 (0.3%)	1460 (0.6%)	1730 (0.5%)
NH	902 (1.0%)	4771 (1.8%)	5673 (1.6%)
NJ	3178 (3.4%)	7322 (2.8%)	10500 (2.9%)
NM	282 (0.3%)	1461 (0.6%)	1743 (0.5%)
NV	1226 (1.3%)	1995 (0.8%)	3221 (0.9%)
NY	8018 (8.5%)	20665 (7.9%)	28683 (8.0%)

	CHGOFF (N=94522)	P I F (N=263075)	Overall (N=357597)
OH	3573 (3.8%)	11986 (4.6%)	15559 (4.4%)
OK	711 (0.8%)	2588 (1.0%)	3299 (0.9%)
OR	1143 (1.2%)	3756 (1.4%)	4899 (1.4%)
PA	3078 (3.3%)	12898 (4.9%)	15976 (4.5%)
RI	715 (0.8%)	3619 (1.4%)	4334 (1.2%)
SC	582 (0.6%)	1528 (0.6%)	2110 (0.6%)
SD	123 (0.1%)	867 (0.3%)	990 (0.3%)
TN	945 (1.0%)	2396 (0.9%)	3341 (0.9%)
TX	5848 (6.2%)	16266 (6.2%)	22114 (6.2%)
UT	2602 (2.8%)	7511 (2.9%)	10113 (2.8%)
VA	1361 (1.4%)	3990 (1.5%)	5351 (1.5%)
VT	172 (0.2%)	1365 (0.5%)	1537 (0.4%)
WA	2057 (2.2%)	6776 (2.6%)	8833 (2.5%)
WI	1408 (1.5%)	6489 (2.5%)	7897 (2.2%)
WV	170 (0.2%)	750 (0.3%)	920 (0.3%)
WY	60 (0.1%)	604 (0.2%)	664 (0.2%)

Figure 4.Default Rate of Loan by States in Continental US

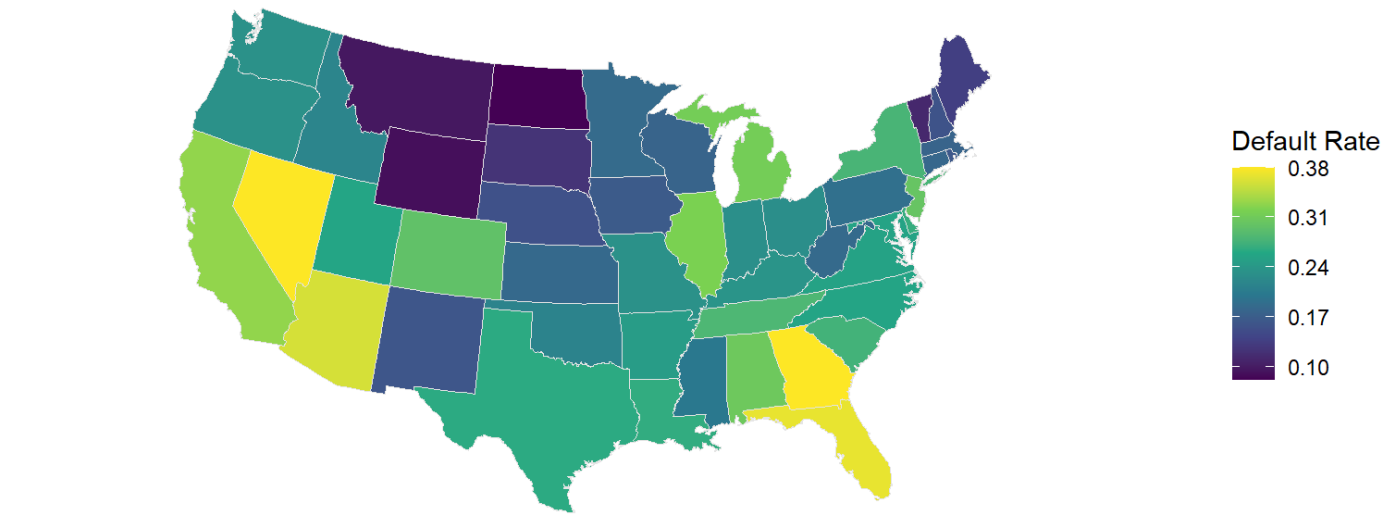


Table 6.Descriptive of Small Business Loans by Loan Status (III)

	CHGOFF (N=94522)	P I F (N=263075)	Overall (N=357597)
ApprovalFY			
1994	2 (0.0%)	60 (0.0%)	62 (0.0%)
1995	11 (0.0%)	268 (0.1%)	279 (0.1%)
1996	8 (0.0%)	79 (0.0%)	87 (0.0%)
1997	21 (0.0%)	336 (0.1%)	357 (0.1%)
1998	32 (0.0%)	160 (0.1%)	192 (0.1%)
1999	507 (0.5%)	5208 (2.0%)	5715 (1.6%)
2000	1495 (1.6%)	13099 (5.0%)	14594 (4.1%)
2001	1718 (1.8%)	13658 (5.2%)	15376 (4.3%)
2002	2204 (2.3%)	16581 (6.3%)	18785 (5.3%)
2003	3670 (3.9%)	24272 (9.2%)	27942 (7.8%)
2004	5347 (5.7%)	27620 (10.5%)	32967 (9.2%)
2005	12723 (13.5%)	34614 (13.2%)	47337 (13.2%)
2006	20220 (21.4%)	35425 (13.5%)	55645 (15.6%)
2007	26293 (27.8%)	30691 (11.7%)	56984 (15.9%)
2008	13726 (14.5%)	15760 (6.0%)	29486 (8.2%)
2009	3269 (3.5%)	13103 (5.0%)	16372 (4.6%)
2010	2039 (2.2%)	13691 (5.2%)	15730 (4.4%)
2011	866 (0.9%)	10740 (4.1%)	11606 (3.2%)
2012	301 (0.3%)	5169 (2.0%)	5470 (1.5%)
2013	65 (0.1%)	2282 (0.9%)	2347 (0.7%)
2014	5 (0.0%)	259 (0.1%)	264 (0.1%)

Table 7.Descriptive of Small Business Loans by Loan Status (IV)

	CHGOFF (N=94522)	P I F (N=263075)	Overall (N=357597)
factor(Disbursement_Year)			
1994	2 (0.0%)	58 (0.0%)	60 (0.0%)
1995	11 (0.0%)	248 (0.1%)	259 (0.1%)
1996	8 (0.0%)	80 (0.0%)	88 (0.0%)
1997	20 (0.0%)	321 (0.1%)	341 (0.1%)
1998	29 (0.0%)	188 (0.1%)	217 (0.1%)
1999	553 (0.6%)	4348 (1.7%)	4901 (1.4%)
2000	1533 (1.6%)	12270 (4.7%)	13803 (3.9%)
2001	1684 (1.8%)	13074 (5.0%)	14758 (4.1%)
2002	2374 (2.5%)	16764 (6.4%)	19138 (5.4%)
2003	3871 (4.1%)	23659 (9.0%)	27530 (7.7%)
2004	6231 (6.6%)	29036 (11.0%)	35267 (9.9%)
2005	13423 (14.2%)	34558 (13.1%)	47981 (13.4%)
2006	22050 (23.3%)	36973 (14.1%)	59023 (16.5%)
2007	25964 (27.5%)	28767 (10.9%)	54731 (15.3%)
2008	10483 (11.1%)	16443 (6.3%)	26926 (7.5%)
2009	3457 (3.7%)	15048 (5.7%)	18505 (5.2%)
2010	1759 (1.9%)	13874 (5.3%)	15633 (4.4%)
2011	730 (0.8%)	9954 (3.8%)	10684 (3.0%)
2012	267 (0.3%)	5025 (1.9%)	5292 (1.5%)
2013	68 (0.1%)	2161 (0.8%)	2229 (0.6%)
2014	5 (0.0%)	224 (0.1%)	229 (0.1%)
2020	0 (0%)	1 (0.0%)	1 (0.0%)
2028	0 (0%)	1 (0.0%)	1 (0.0%)