Honors Assignment Report: PPP Loan Data

Report by Jennifer Case
7/3/2022

Automated fraud detection has numerous applications that promises to save money for both businesses and governments. At the start of the COVID-19 pandemic, the US government offered Paycheck Protection Program (PPP) loans through the Small Business Administration (SBA) to help pay business costs and to increase job retention due to the pandemic-related shutdown. Due to the rapid release of funding through the PPP, the program has had issues with fraud[1,2].

In this report, a subset of the total PPP loan data released by the SBA is examined to determine if there may be recognizable patterns that may assist in fraud detection. Since the data is not labeled whether or not it is fraud, this report takes a broad look over patterns in the data, which may be relevant for fraud detection algorithms in the future.

Data Description

The PPP loan data[3] is broken out into two categories: loans above $150k and loans equal to or below $150k. Since fraud patterns are likely to differ based on loan size (e.g., does an individual apply for one large fraudulent loan or a series of small fraudulent loans?), only the subset of loans above $150k is considered in this analysis.

There are 986,532 records for loans above $150k with the following 53 attributes:

	Attribute Name	Attribute Descriptions	Data Type
1	LoanNumber	Loan Number (unique identifier)	int64
2	DateApproved	Loan Funded Date	object
3	SBAOfficeCode	SBA Origination Office Code	int64
4	ProcessingMethod	Loan Delivery Method (PPP for first draw; PPS for second draw)	object
5	BorrowerName	Borrower Name	object
6	BorrowerAddress	Borrower Street Address	object
7	BorrowerCity	Borrower City	object
8	BorrowerState	Borrower State	object
9	BorrowerZip	Borrower Zip Code	object
10	LoanStatusDate	Loan Status Date - Loan Status Date is blank when the loan is disbursed but not Paid In Full or Charged Off	object

[1] "Man Convicted for $27 Million PPP Fraud Scheme." *The United States Department of Justice*, 29 March 2022, https://www.justice.gov/opa/pr/man-convicted-27-million-ppp-fraud-scheme.
[2] "Woman Pleads Guilty for $43.8 Million COVID-19 Relief Fraud Scheme." *The United States Department of Justice*, 6 April 2022, https://www.justice.gov/opa/pr/woman-pleads-guilty-438-million-covid-19-relief-fraud-scheme.
[3] "PPP FOIA: Data and Resources." *U.S. Small Business Administration*, 4 April 2022, https://data.sba.gov/dataset/ppp-foia.

11	LoanStatus	Loan Status Description - Loan Status is replaced by 'Exemption 4' when the loan is disbursed but not Paid in Full or Charged Off	object
12	Term	Loan Maturity in Months	int64
13	SBAGuarantyPercentage	SBA Guaranty Percentage	int64
14	InitialApprovalAmount	Loan Approval Amount (at origination)	float64
15	CurrentApprovalAmount	Loan Approval Amount (current)	float64
16	UndisbursedAmount	Undisbursed Amount	float64
17	FranchiseName	Franchise Name	object
18	ServicingLenderLocationID	Lender Location ID (unique identifier)	int64
19	ServicingLenderName	Servicing Lender Name	object
20	ServicingLenderAddress	Servicing Lender Street Address	object
21	ServicingLenderCity	Servicing Lender City	object
22	ServicingLenderState	Servicing Lender State	object
23	ServicingLenderZip	Servicing Lender Zip Code	object
24	RuralUrbanIndicator	Rural or Urban Indicator (R/U)	object
25	HubzoneIndicator	Historically Underutilized Business zone (Hubzone) Indicator (Y/N)	object
26	LMIIndicator	Low- and Moderate-Income (LMI) Indicator (Y/N)	object
27	BusinessAgeDescription	Business Age Description	object
28	ProjectCity	Project City	object
29	ProjectCountyName	Project County Name	object
30	ProjectState	Project State	object
31	ProjectZip	Project Zip Code	object
32	CD	Project Congressional District	object
33	JobsReported	Number of Employees	float64
34	NAICSCode	North American Industry Classification System (NAICS) 6 digit code	float64
35	Race	Borrower Race Description	object
36	Ethnicity	Borrower Ethnicity Description	object
37	UTILITIES_PROCEED	Note: Proceed data is lender reported at origination. On the PPP application the proceeds fields were check boxes.	float64
38	PAYROLL_PROCEED		float64
39	MORTGAGE_INTEREST_PROCEED		float64
40	RENT_PROCEED		float64
41	REFINANCE_EIDL_PROCEED		float64
42	HEALTH_CARE_PROCEED		float64
43	DEBT_INTEREST_PROCEED		float64
44	BusinessType	Business Type Description	object
45	OriginatingLenderLocationID	Originating Lender ID (unique identifier)	int64
46	OriginatingLender	Originating Lender Name	object
47	OriginatingLenderCity	Originating Lender City	object
48	OriginatingLenderState	Originating Lender State	object
49	Gender	Gender Indicator	object
50	Veteran	Veteran Indicator	object

51	NonProfit	'Yes' if Business Type = Non-Profit Organization or Non-Profit Childcare Center or 501(c) Non Profit	object
52	ForgivenessAmount	Forgiveness Amount	float64
53	ForgivenessDate	Forgiveness Paid Date	object

In addition to the PPP loan data, this analysis will also consider NAICS data[4] that specifies the number of businesses that fall under various industries to enable proper scaling of the number of businesses that received PPP loans over $150k per industry given the total number of businesses in that industry. To join the data, the NAICSCode attribute that contains a 6 digit NAICS code, which indicates a specific industry, in the PPP loan data will be converted to a 2 digit NAICS code, which indicates a general industry, and connected to the NAICS data on the Code attribute. The goal of identifying industries is to enable the stratification of the data by industry as well as identify if certain industries are dramatically over- or under-represented in the data. That data set will include:

	Attribute Name	Attribute Descriptions	Data Type
1	Code	NAICS 2 digit code	int64
2	IndustryTitle	Industry Title	object
3	NumBusinesses	Number of Business Establishments	object

Data Exploration Plan

- Perform data cleaning.
    - Evaluate for duplicate records.
    - Validate data value ranges.
    - Check for manual data entry errors.
    - Outliers will be maintained since they could assist in fraud detection.
- Perform feature engineering using cleaned data.
    - Select variables based on available data:
        - LoanNumber, DateApproved, ProcessingMethod, BorrowerState, LoanStatus, Term, SBAGuarantyPercentage, InitialApprovalAmount, CurrentApprovalAmount, UndisbursedAmount, RuralUrbanIndicator, HubzoneIndicator, LMIIndicator, BusinessAgeDescription, ProjectState, JobsReported, NAICSCode, Race, Ethnicity, BusinessType, Gender, Veteran, NonProfit, and ForgivenessAmount.
    - Generate new features:
        - Industry: use the NAICS code to identify which of 20 main industries the project falls under.
        - StateMatch: a binary data column that identifies if the project state matches the borrower state.
        - OwedAmount: the difference between the current approval amount and the forgiveness amount.
    - Scale column data as needed:

[4] "NAICS & SIC Identification Tools." *NAICS Association*, https://www.naics.com/search/#naics.

        - When presenting data using the NAICS code, it is scaled per 1,000 for that industry type.
        - Ensure data follows a normal distribution if needed for statistical analysis.
- Perform exploratory data analysis.
    - Generate statistics to understand the mean, standard deviation, median, maximum, and minimum values for the following features: initial approval amount, current loan approval amount, undisbursed amount, forgiveness amount, and owed amount.
    - Generate percentages of loans based on race, ethnicity, gender, veteran status, non-profit, rural/urban indicator, HUBzone indicator, LMI indicator, business age, business type, and state match.
    - Generate visualizations of the data grouping by Industry.
        - Generate box plots of current loan amounts and owed amounts.
        - Generate scatter plots based on number of employees and current loan amounts.
    - Generate visualizations of the data grouping by Gender.
        - Generate box plots of current loan amounts and owed amounts.
        - Generate scatter plots based on number of employees and current loan amounts.
    - Generate pair plots of features to determine trends.

Data Cleaning and Feature Engineering

- Evaluated for duplicates via ensuring that the loan numbers were unique. No duplicate entries were found.
- The following columns are being trimmed from the data set:

	Attribute Name	Reason for Removal
3	SBAOfficeCode	Not being used in analysis
5	BorrowerName	Not being used in analysis
6	BorrowerAddress	Not being used in analysis
7	BorrowerCity	Not being used in analysis
9	BorrowerZip	Not being used in analysis
10	LoanStatusDate	Not being used in analysis
13	SBAGuarantyPercentage	SBA guaranty percentage is 100% for all loans.
16	UndisbursedAmount	Only 17 samples have undisbursed amounts, so the sample is not large enough to learn anything meaningful.
17	FranchiseName	Not being used in analysis
18	ServicingLenderLocationID	Not being used in analysis
19	ServicingLenderName	Not being used in analysis
20	ServicingLenderAddress	Not being used in analysis
21	ServicingLenderCity	Not being used in analysis
22	ServicingLenderState	Not being used in analysis
23	ServicingLenderZip	Not being used in analysis
28	ProjectCity	Not being used in analysis
29	ProjectCountyName	Not being used in analysis
31	ProjectZip	Not being used in analysis

32	CD	Not being used in analysis
37	UTILITIES_PROCEED	The lender provided data is inconsistent
38	PAYROLL_PROCEED	The lender provided data is inconsistent
39	MORTGAGE_INTEREST_PROCEED	The lender provided data is inconsistent
40	RENT_PROCEED	The lender provided data is inconsistent
41	REFINANCE_EIDL_PROCEED	The lender provided data is inconsistent
42	HEALTH_CARE_PROCEED	The lender provided data is inconsistent
43	DEBT_INTEREST_PROCEED	The lender provided data is inconsistent
45	OriginatingLenderLocationID	Not being used in analysis
46	OriginatingLender	Not being used in analysis
47	OriginatingLenderCity	Not being used in analysis
48	OriginatingLenderState	Not being used in analysis
53	ForgivenessDate	Not being used in analysis

- The null values for the remaining columns were treated as follows:

	Attribute Name	**Treatment of Null Values**
1	LoanNumber	N/A
2	DateApproved	N/A
4	ProcessingMethod	N/A
8	BorrowerState	Rows removed
11	LoanStatus	N/A
12	Term	N/A
14	InitialApprovalAmount	N/A
15	CurrentApprovalAmount	N/A
24	RuralUrbanIndicator	N/A
25	HubzoneIndicator	N/A
26	LMIIndicator	N/A
27	BusinessAgeDescription	Null values moved to the pre-existing ‘Unanswered’ category
30	ProjectState	Rows removed
33	JobsReported	Rows removed
34	NAICSCode	Null values changed to zero to be translated to ‘Unanswered’ in the Industry column
35	Race	N/A
36	Ethnicity	N/A
44	BusinessType	Null values moved to an ‘Unanswered’ category
49	Gender	N/A
50	Veteran	N/A
51	NonProfit	Null values are equivalent to ‘No’ since ‘Yes’ was the only possible answer
52	ForgivenessAmount	Null values treated as zero

After removing rows, we have 968,517 samples remaining.

- The StateMatch column was created by comparing the BorrowerState to the ProjectState (Y/N).
- The OwedAmount column was created by subtracting the ForgivenessAmount from the CurrentApprovalAmount.

- The Industry column was created by converting the 6-digit NAICS code in the PPP loan database to a 2-digit code and cross referencing the IndustryTitle from the NAICS database. Null answers or 6-digit codes from the PPP loan database that do not have a corresponding matching in the NAICS database (e.g., 999990) were placed in an 'Unanswered' category.

Key Findings and Insights

1. When looking at the dates that PPP loans got approved, certain days stand out as having a high volume of loan applications approved. If agencies feel pressured to get through a higher volume of applications on specific days, it may result in shorter review times per loan which may result in more fraudulent loans getting accepted.

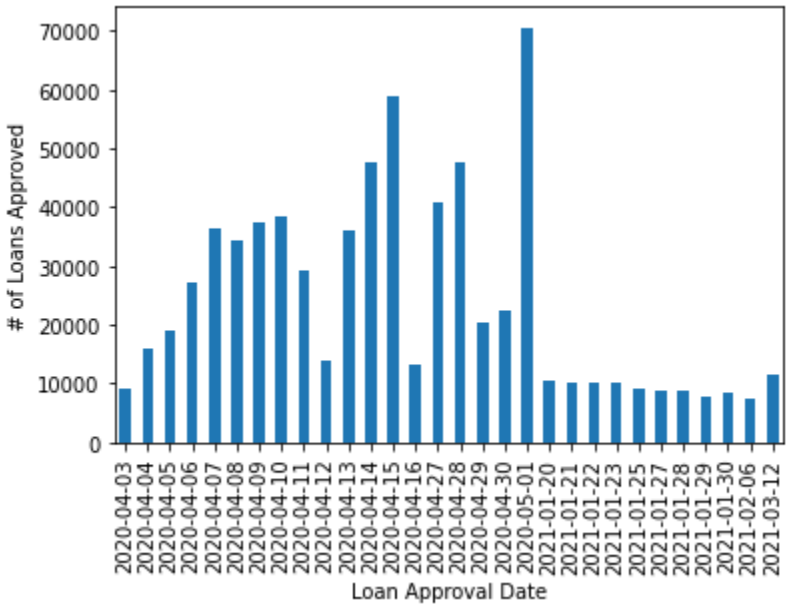

Figure 1. Graph showing the number of loans approved for the top 30 loan approval days.

2. Female owned businesses appear to receive smaller loan amounts per number of employees compared to male owned businesses or when applicants do not specify a gender. There are several reasons why this may be the case, such as women preferring to own businesses in less lucrative industries (which would result in a lower loan request) or may be tied to the same social tendencies that result in women asking for lower wages or raises.[5]

[5] Ro, Christine. "How the salary 'ask gap' perpetuates unequal pay." *BBC*, 18 June 2021, https://www.bbc.com/worklife/article/20210615-how-the-salary-ask-gap-perpetuates-unequal-pay.

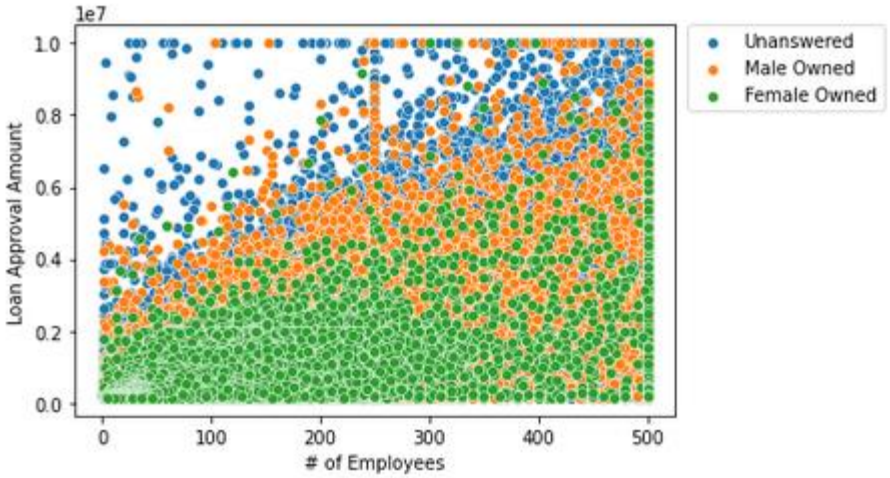

Figure 2. Graph of the number of employees vs the loan approval amount stratified by gender. The unanswered data was plotted first, then male owned businesses, and lastly female owned businesses to better visualize the trend.

3. When the data is stratified by industry, it becomes easier to identify outliers that may be indicative of fraud where the applicant requested more money than may have been reasonable given the size of the company. While many industries had outliers, two specific industries are highlighted in the graphics below: (1) Agriculture, Forestry, Fishing and Hunting and (2) Retail Trade. Using machine learning to flag these outliers for review can help reduce the burden on employees of identifying suspicious data.

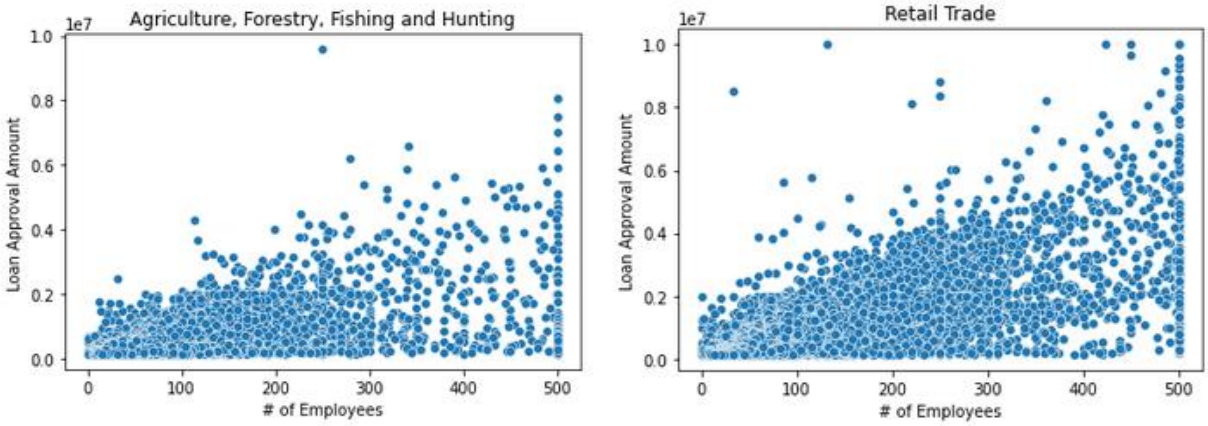

Figure 3. Graphs showing the number of employees vs the loan approval amount for the Agricultural, Forestry, Fishing and Hunting Industry and the Retail Trade Industry.

4. When the industry-stratified data is additionally split by gender, it reveals:
    a. Women appear to own smaller businesses (i.e., less employees) for many industries.
    b. More male-owned businesses received larger loans than female-owned businesses for many industries.

Figure 4. Graph showing the number of employees vs the loan approval amount for Retail Trade Industry for each gender. The unanswered data was plotted first, then male owned businesses, and lastly female owned businesses to better visualize the trend.

5. Although 80% of the PPP loan applications over $150k did not enter a race, of the remaining 192,213 samples, on 3.8% went to Black or African American applicants while 81.5% went to White applicants, 11.1% went to Asian American applicants, and 3.2% went to American Indian or Alaska Native applicants. According to the U.S. Census Bureau[6], the percentages of the population based on these races are: 76.3% White, 13.4% Black or African American, 5.9% Asian American, and 1.3% American Indian/Alaska Native. Based on this overview, we can identify that the percentage of PPP loans over $150k that went to Black or African American applicants was significantly lower (around 10%) than their representation in the population, while both Asian American and American Indian/Alaska Native applicants were overrepresented by almost double. While there are numerous possible reasons for these under- and overrepresentations of various races, more data is required (e.g., the data on denied loan applications or data on business owners across the county) before any conclusions can be drawn.

6. Considering the # of loans per 1,000 taken by each industry shows which types of industries requested more PPP loans proportional to the number of businesses in that industry, which can indicate which industries suffered the most losses during the pandemic shutdown. It can be seen that Mining, Manufacturing, and Accommodation and Food Services proportionally requested more PPP loans, which is likely due to the inability of jobs in these industries to be accomplished remotely affecting their ability to make money to pay their workers during the shutdown.

[6] "Quick Facts." *United States Census Bureau*, 2021, https://www.census.gov/quickfacts/fact/table/US/PST045221.

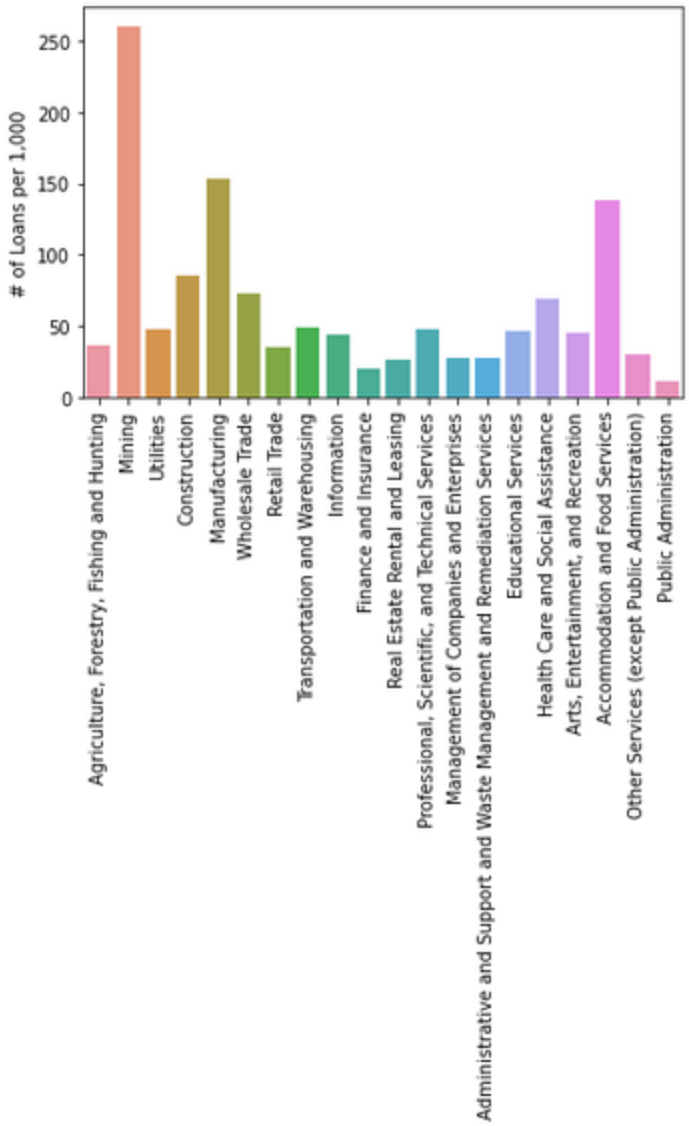

Figure 5. Distribution of the number of loans per 1,000 businesses by industry.

Proposed Hypotheses

1. Loan approval days that have more than 30,000 loans approved have a lower loan forgiveness rating than other loan approval days.
2. There is a difference between the CurrentApprovalAmount between female owned and male owned businesses for PPP loans over $150k.
3. There is a difference in the size of female owned and male owned businesses per industry for PPP loans over $150k.

Significance Test and Results

Hypothesis: Loan approval days that have more than 30,000 loans approved have a lower loan forgiveness rating than other loan approval days.

For this hypothesis, loan forgiveness rate is calculated as:

$$LoanForgivenessRate = \frac{\#\ of\ loans\ forgiven}{\#\ of\ loans}$$

where the # of loans forgiven is calculated as a loan that has a ForgivenessAmount > 0. This will group loans that were partially forgiven with loans that were fully forgiven.

To test this hypothesis, the data was (1) plotted visually and (2) subjected to a t-test with $\alpha = 0.05$.

When the distributions of forgiveness rates are plotted based on high loan approval days (i.e., days that approved more than 30,000 loans) and low loan approval days, it can be seen that distributions vary.

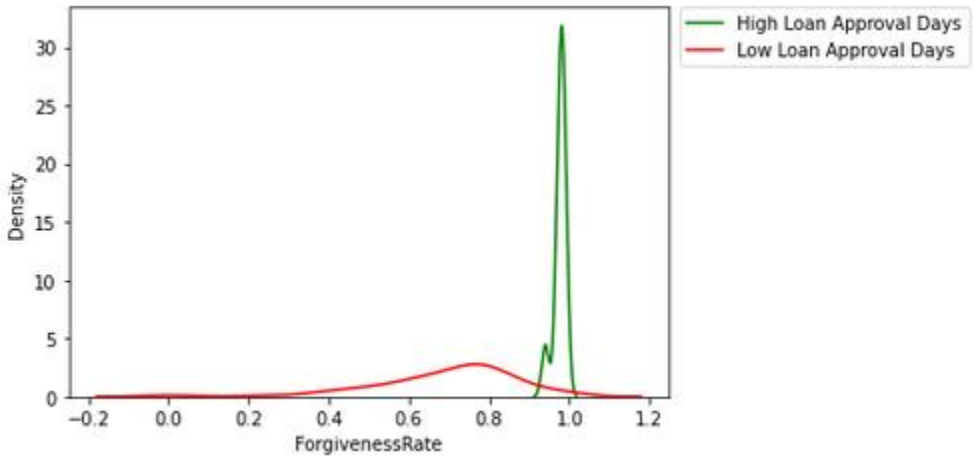

Figure 6. Density plot of the forgiveness rate based on high or low loan approval days.

Due to the non-uniformity of the data, the box cox transformation was used for each set of data before conducting the t-test.

The results of the t-test were t-value = -8 and p-value = 5e-14, which indicates that there is a difference in the loan forgiveness rate based on how many loans were approved that day. However, the initial hypothesis is incorrect. Loan approval days that have more than 30,000 loans approved have a *higher* loan forgiveness rating than other loan approval days.

Next Steps

Within this data set, there is further analyses that can be done comparing the number of employees vs loan approval amounts across industries for Race, Ethnicity, Veteran status, Non-Profit status, Rural/Urban indicator, HUBZone Indicator, LMI Indicator, Business type, and Business age.

Summary

This data set contains a large amount of information about accepted PPP loans. Since some of the data was entered manually, it is difficult to use city data. For example, looking at only Illinois cities, there are entries for “SAINT CHARLES”, “Saint Charles”, “ST CHARLES”, “St Charles”, “ST. CHARLES”, “St. Charles”, and “St.Charles” that all indicate the same city. This can additionally be seen in misspellings such as “Chicacgo” and “Chicagoi” for Chicago. Additionally, no analysis can be done based on how the money

was spent since this was lender reported and the data was largely missing and the existing data may have inaccuracies.

This data set only provides information on PPP loan forgiveness for loans above $150k. To draw more conclusions from this data set, it would be useful to collect additional data on:

- Rejected PPP loan applications for loans above $150k to see statistics on the loan applications based on Gender, Race, Ethnicity, Veteran status, etc. to be able to draw conclusions on the accepted loan applications.
- Statistics on businesses based on Gender, Race, Ethnicity, Veteran status, etc. to be able to draw conclusions on the diversity of the accepted loan applications.
- Labeled fraud data that can be used to train an algorithm for fraud detection.