



Data Engineering

Homework 2

Professor: S. Kehagias

30/05/2024

MSc in Statistics - AUEB

Grammenos Konstantinos

AM : f3612302

Contents

Contents.....	3
List of Figures	4
List of Tables.....	4
Introduction	5
Data	5
Exploratory Data Analysis.....	5
Analysis of Total Loan Amount per 100k residents.....	9
Conclusions	10

List of Figures

Figure 1: Bar plot of Top 20 Borrower Cities by the Number of Loans 6

Figure 2: Count Of Loans Distribution by Borrower State 6

Figure 3: Total loan amount by State 7

Figure 4: Number of Loans by Business Type..... 7

Figure 5: Total Number of Jobs by Business Type 8

Figure 6: Number of Loans by Race 8

Figure 7: Number of Loans by Ethnicity..... 8

Figure 8: Number Of Loans by Gender..... 9

List of Tables

Table 1:Table of State Name and Total Loan Amount per 100k residents sorted by the Loan Amount Per 100k residents 9

Introduction

The PPP was established to help businesses and sole proprietors continue paying their employees during the COVID-19 pandemic. This analysis aims to provide a comprehensive view of the financial aid distributed across the United States, focusing on state-level data. The report highlights the key steps taken to clean and analyze the data, addressing challenges such as large dataset sizes and derive useful insights.

Data

In the initial stage of the project, a sample of 1,000 rows from one of the PPP datasets was loaded to understand the structure and data types of each column. This was essential for planning the data cleaning and processing steps.

This preliminary examination revealed that the 'NonProfit' and SBAGuarantyPercentage column predominantly contains NaN values. Such insights guided the subsequent data cleaning steps. Also, we checked the Data Type of each column and we observed that most of the columns have not the correct Data Type.

After this step we created a data type mapping dictionary in order to define the correct data types for each column when loading the dataset. We did this so to optimize memory usage and ensure that all the columns have the datatype that they should have.

Then, we selected the columns that we thought that we need for the aim of our analysis, in order to optimize memory usage. It would not be right to load the dataset with all the columns as we may do not need them all.

Next we loaded a random sample of each file and merge all the random samples in a single data frame. We did this through these steps:

- First we set the paths for the input and output folders.
- Then we defined a function to load the data of the file to a dataframe, random sample this dataframe, and save a csv file with this random sample.
- We iterate through our csv files, processes each file and save the sampled data.
- Next, we collect sampled data from all files and concatenates them into a single Data Frame.

Finally, after we have our final data frame with all our data, we checked the memory usage info, and saved the final data frame into a new csv file.

We also Loaded the data that we found from the Latest Population estimates of US Counties from US Census Bureau and the data containing all the information on job postings from the Opportunity Insights

Exploratory Data Analysis

First of all we wanted to see the number of loans distributed to the top 20 cities under the Paycheck Protection Program (PPP). We had an issue of duplicate entries due to inconsistent data entry (e.g., "HOUSTON" and "NEW YORK" appearing in different cases). In order to address this issue, we converted all city names to uppercase to ensure consistency and grouped the data by city names and summed the loan counts to get accurate totals.

Figure 1: Bar plot of Top 20 Borrower Cities by the Number of Loans

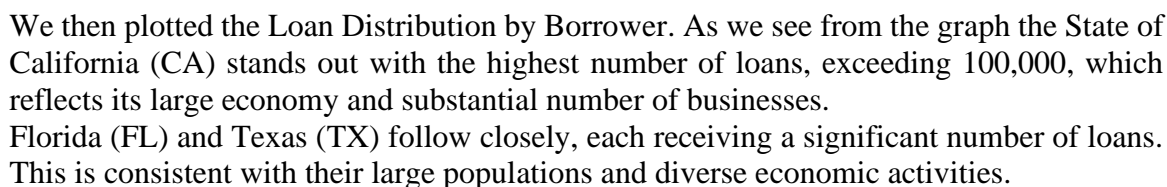
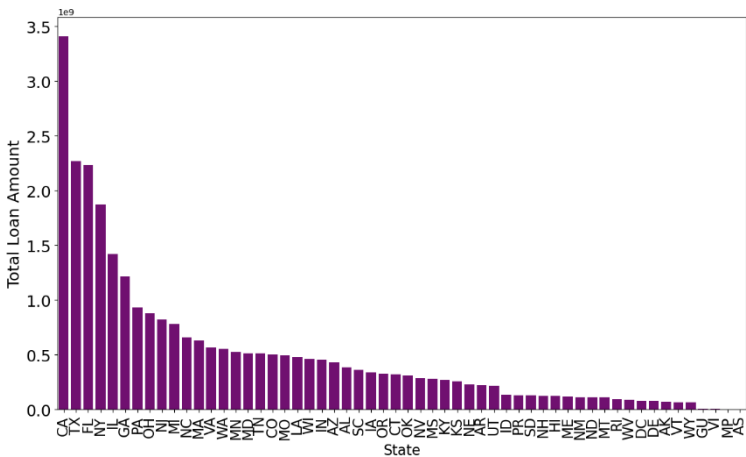


Figure 2: Count Of Loans Distribution by Borrower State



populations, such as Wyoming (WY) and Vermont (VT), which receive the least loan amounts.

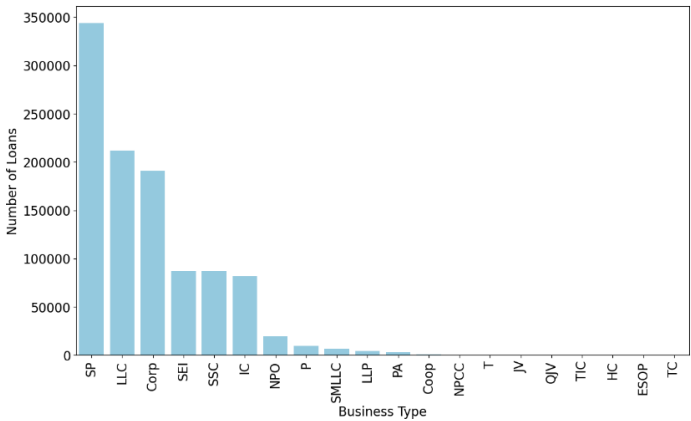
Figure 3: Total loan amount by State



Next, we see a bar plot of the number of Loans by Business Type. We can see the distribution of PPP loans across various business types.

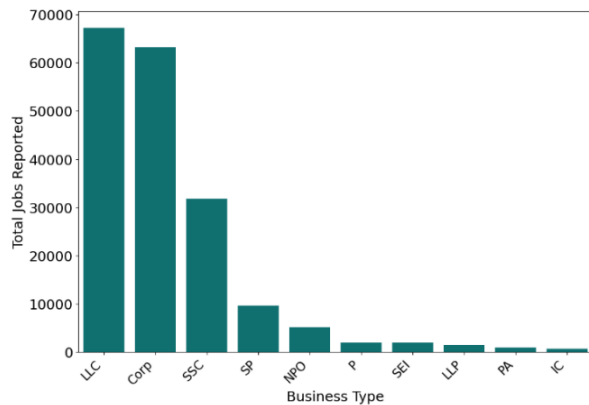
So in the graph below we see that Sole Proprietorships received the highest number of loans, exceeding 300,000, while Corporations and Limited Liability Companies (LLC) also received also significant numbers of loans. Other business types, such as Subchapter S Corporations and Independent Contractors, received moderate numbers of loans. Less common business types, like Non-Profit Organizations and Trusts, received fewer loans. This distribution highlights the emphasis on supporting smaller, independent businesses and sole proprietors during the COVID-19 pandemic.

Figure 4: Number of Loans by Business Type



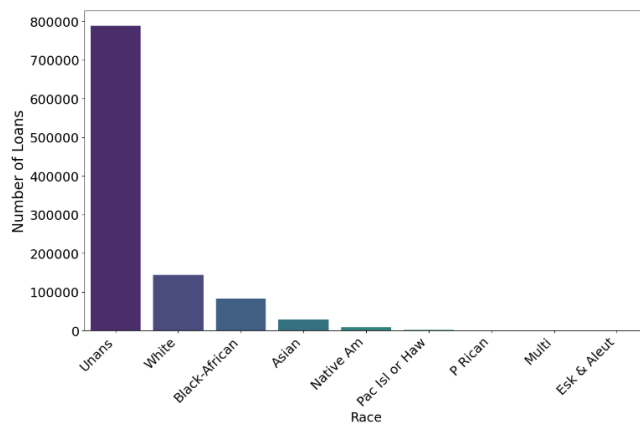
Next, we did a bar plot to illustrate the total number of jobs reported by different business types under the Paycheck Protection Program (PPP). The majority of reported jobs are from Corporations, Limited Liability Companies (LLCs), and Subchapter S Corporations. These business types account for the highest number of jobs, reflecting their significant role in the economy and the impact of PPP loans on sustaining employment during COVID-19. This visualization helps to understand the distribution of job support across various business types.

Figure 5: Total Number of Jobs by Business Type



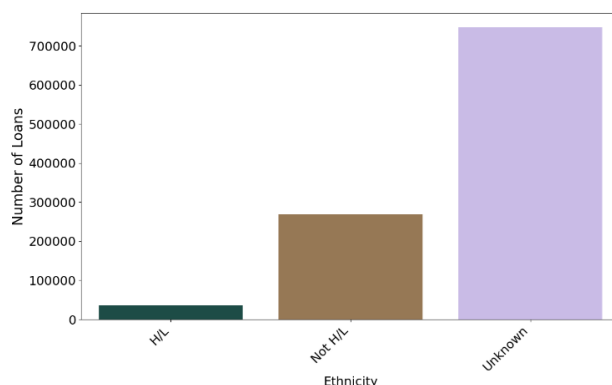
In addition, we illustrated the distribution of PPP loans across different racial groups. The majority of the loans fall under the "Unanswered" category, indicating a significant amount of missing data in terms of racial identification. Among those who specified their race, White borrowers received the highest number of loans, followed by Black or African American.

Figure 6: Number of Loans by Race



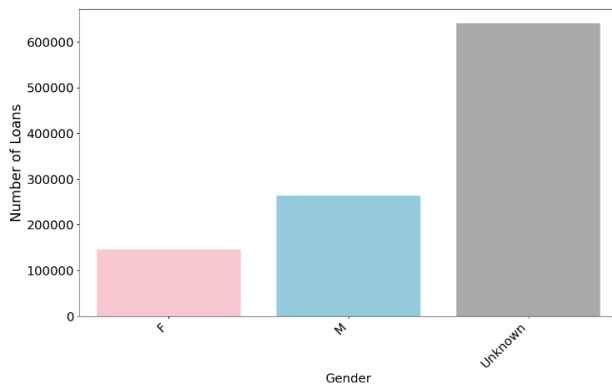
After that, we plotted the distribution of PPP loans across different ethnicities. The majority of the loans fall under the "Unknown" category, indicating that we do not have many information about the ethnicity of the loan borrowers. Among those who know their ethnicity, Not Hispanic or Latino borrowers received the highest number of loans.

Figure 7: Number of Loans by Ethnicity



Finally, we plotted the distribution of PPP loans by gender. The majority of the loans fall again under the "Unanswered" category, indicating that we also do not have many information about the gender of the loan borrowers. Among those who know their gender, Male borrowers received the highest # of loans.

Figure 8: Number Of Loans by Gender



Analysis of Total Loan Amount per 100k residents

First, in order to do this analysis we need to merge the data frame of the PPP data and the data frame of the population estimates. Throughout the route of my analysis we observe that while in the PPP data the state name is denoted by only 2 letters, in the population estimate data the state is defined by the whole name of the state and that was a problem when we tried to merge those two. We had to do a mapping in order to map the coded state name of the one dataset (PPP data) with the state name of the other dataset. In order to solve this issue we constructed a dictionary to map state decoded states of the one data frame, to full state names of the other data frame. Then we aggregated loan amounts by state and we summarized population estimates by state and we calculated the Total Loan Amount per 100k residents. Finally, we merged the PPP data with population estimation data and we sorted the dataset based on the Total Loan Amount per 100k residents. So, in the Table below we see the State Name, the Total Loan Amount, the Population Estimate and the Total Loan Amount per 100k residents.

Table 1: Table of State Name and Total Loan Amount per 100k residents sorted by the Loan Amount Per 100k residents

Borrower State	Total Loan Amount	Population Estimate 2020	Loan Amount Per100k Residents
North Dakota	115,378,856	1,530,618	7,538,056.92
South Dakota	131,262,912	1,785,434	7,351,877.02
Nebraska	231,835,296	3,875,104	5,982,685.78
Illinois	1,412,761,728	25,175,060	5,611,751.19
Georgia	1,194,599,040	21,420,034	5,577,017.48
Columbia	79,343,640	1,425,632	5,565,506.39
Wyoming	64,247,328	1,164,656	5,516,420.99
Iowa	343,131,936	6,327,122	5,423,191.40
Vermont	65,222,280	1,246,694	5,231,618.99
Louisiana	478,641,152	9,290,636	5,151,866.37
Montana	110,943,240	2,161,154	5,133,518.48
Florida	2,194,535,680	43,466,624	5,048,783.36
Alaska	72,090,384	1,462,316	4,929,877.26
New York	1,854,199,936	38,673,552	4,794,490.91
Mississippi	279,033,376	5,933,572	4,702,620.55
Minnesota	524,843,104	11,314,684	4,638,601.52
New Hampshire	125,534,632	2,732,550	4,594,047.03
New Jersey	815,126,080	17,764,742	4,588,448.74
Connecticut	320,844,256	7,114,012	4,510,032.54
Massachusetts	617,847,360	13,787,148	4,481,328.26

Conclusions

The analysis of the Paycheck Protection Program (PPP) data provided several insights into the distribution of financial aid during the COVID-19 pandemic. Overall, all these plots provide a clear visual representation of the PPP loan distribution across various demographics and geographic locations.

Firstly, we found out that the states with the highest number of loans were California, Texas, and Florida, reflecting their large economies and substantial business activities. These states also received the highest total loan amounts, with California leading at over \$3.5 billion. Also, major urban centers like Chicago, Houston, Miami, and Los Angeles emerged as top recipients, highlighting the concentration of financial aid in densely populated and economically active regions.

Also, as regards the business we found that Sole Proprietorships received the highest number of loans, indicating a significant emphasis on supporting small, independent businesses during the pandemic. Corporations and Limited Liability Companies (LLCs) also received substantial numbers of loans, underscoring their critical role in the economy. Finally, the analysis revealed significant gaps in the demographic data, with a large proportion of loans falling under the "Unanswered" category for race, ethnicity, and gender. This indicates a need for better data collection and reporting practices.

Among those who specified their demographics, White borrowers received the highest number of loans, followed by Black or African American and Asian borrowers. For gender, Male-owned businesses received more loans than Female-owned businesses.

All these insights and the focus of our analysis is to enhance the effectiveness of financial aid programs and ensure that support is distributed equitably and efficiently across different regions, business types, and demographic groups.