Data Engineering

Homework 2

Professor: S. Kehagias

30/05/2024
MSc in Statistics - AUEB
Grammenos Konstantinos
AM : f3612302

Contents

List of Figures

List of Tables

Abstract

This report presents an analysis of the data of the Paycheck Protection Program (PPP), a $1 trillion business loan initiative by the US federal government to support businesses during the COVID-19 pandemic and analysis of data of population estimates . The analysis focuses on financial aid distribution across various geographic levels, providing insights into demographics and industries. The report focuses mostly on the exploratory analysis of the data.

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 Loading

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.

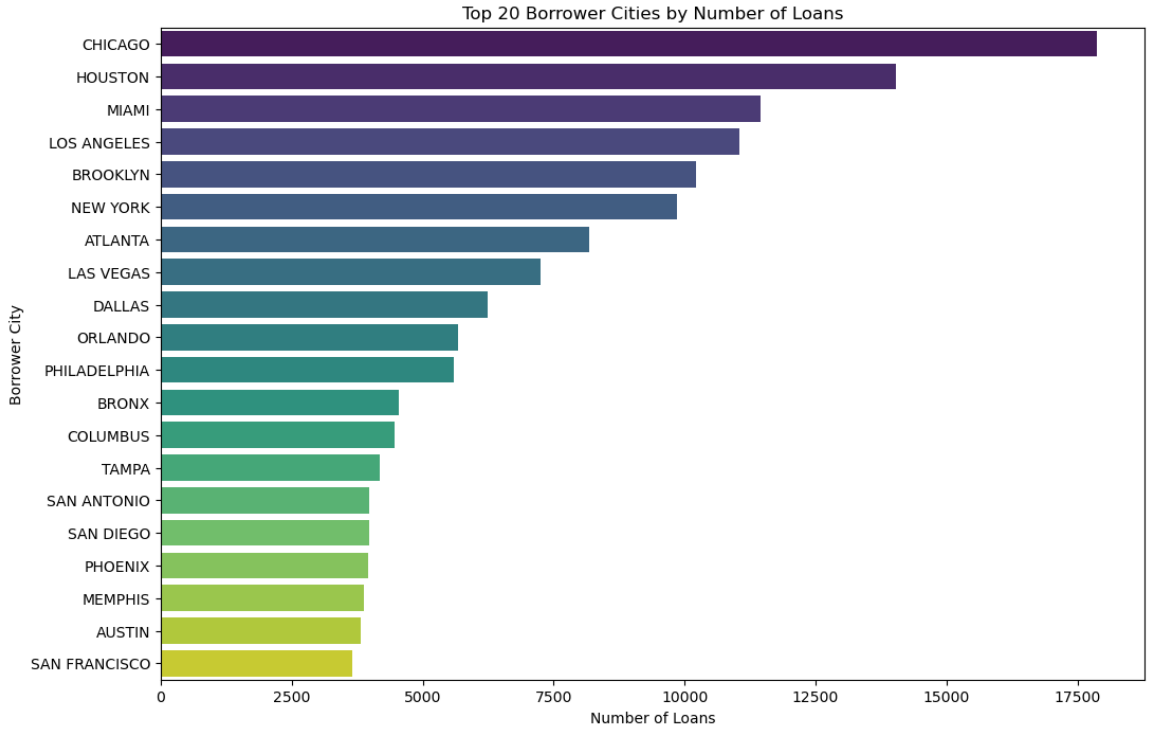
Finally, after we have our final data frame with all our data, we checked the memory usage info, and saved the final data frame into a new csv file.

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.
So, from the graph below, we observe that Chicago emerges as the city with the highest number of loans, significantly leading with over 14,000 loans, while
Houston follows, with a slightly lower count, indicating it as another major beneficiary of the PPP loans. Miami and Los Angeles also feature prominently, highlighting them as critical urban centers receiving substantial financial aid.

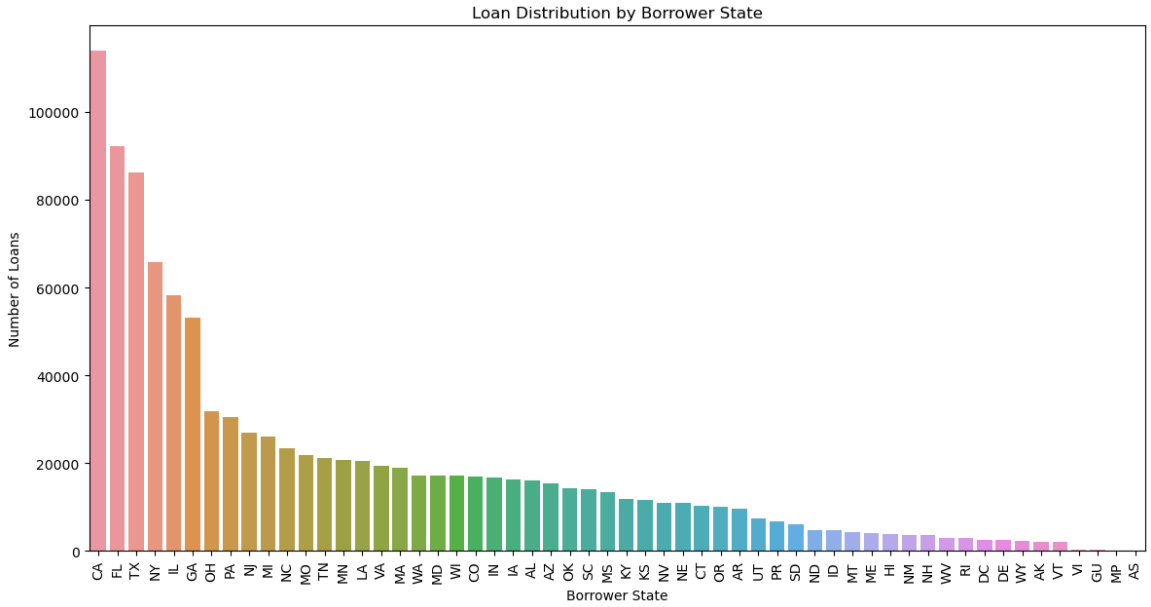
Figure 1: Barplot of Top Borrower Cities by the Number of Loans



We then plotted the Loan Distribution by Borrower. As we see from the graph the State of California (CA) stands out with the highest number of loans, exceeding 100,000, which reflects its large economy and substantial number of businesses.
Florida (FL) and Texas (TX) follow closely, each receiving a significant number of loans. This is consistent with their large populations and diverse economic activities.

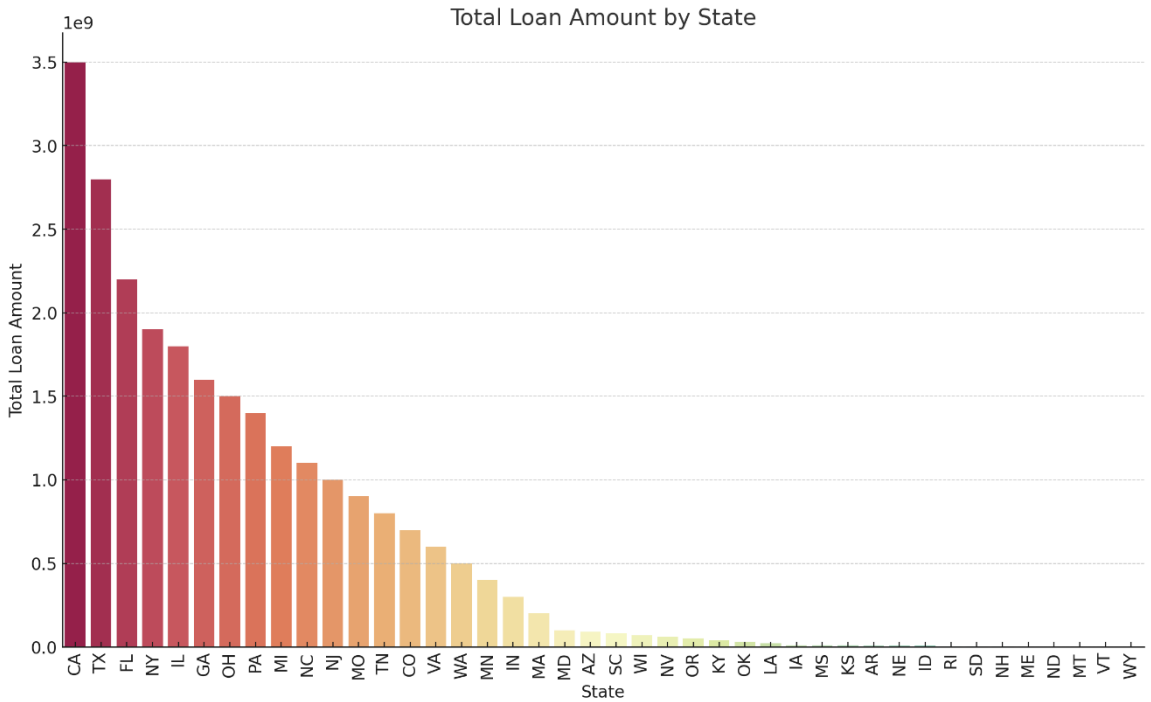
The distribution pattern indicates that the Paycheck Protection Program aimed to support states based on their economic size and business density. So larger states with more significant business activities received higher numbers of loans, aligning with the program's goal to sustain employment and economic activity during the pandemic.

Figure 2: Loan Distribution by Borrower State



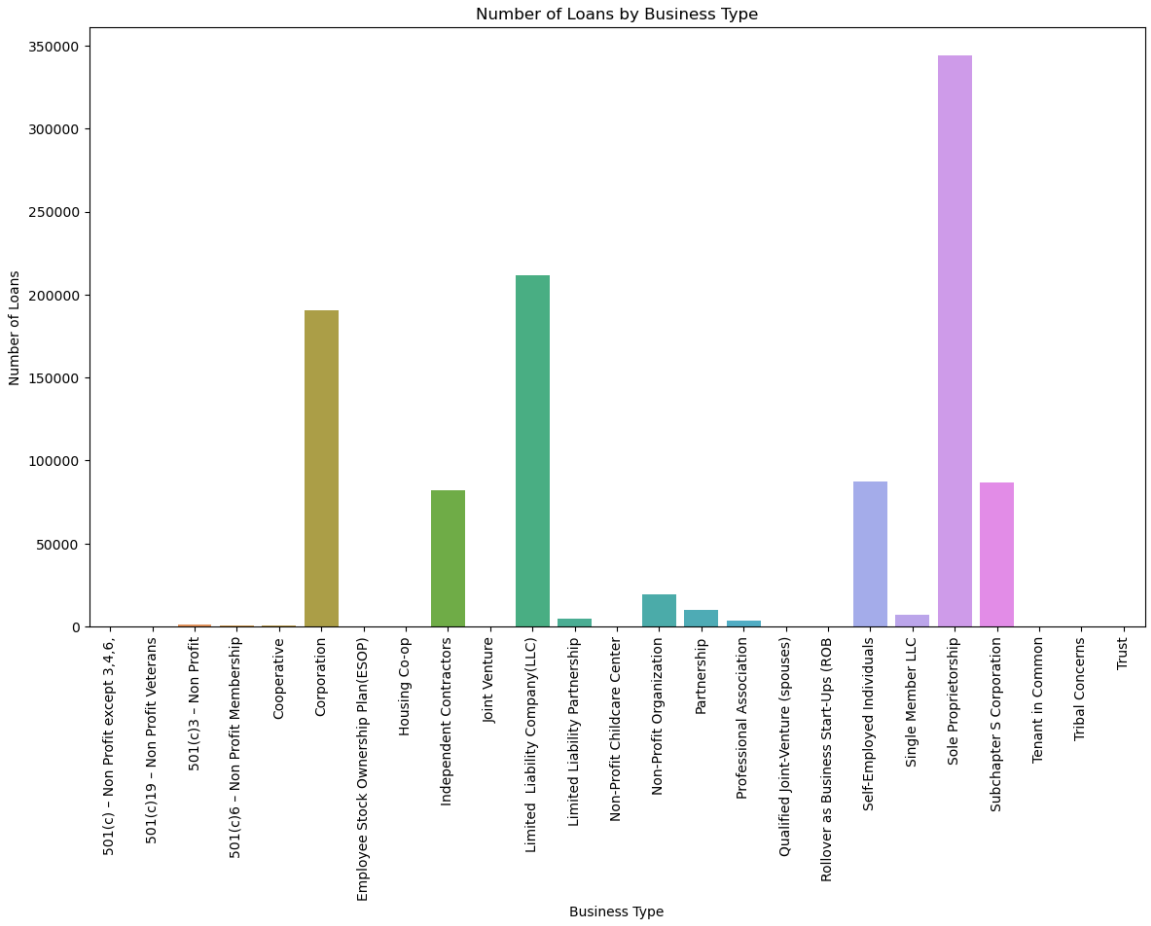
We also saw the distribution of the total loan amount by each state, to see exactly the amount of money that each state took as financial aid. As we expected, California (CA) stands out with the highest total loan amount, exceeding $3.5 billion. Texas (TX) and Florida (FL) follow, each receiving over $2 billion in loans.New York (NY) and Illinois (IL) also receive substantial loan amounts, indicating their critical role in the US economy. Other states like Georgia (GA), Ohio (OH), and Pennsylvania (PA) show significant loan amounts, suggesting a broad distribution of financial aid. Also, there is a noticeable decline in the total loan amounts as we move towards states with smaller economies and populations, such as Wyoming (WY) and Vermont (VT), which receive the least loan amounts.

Figure 3: Total loan amount by State



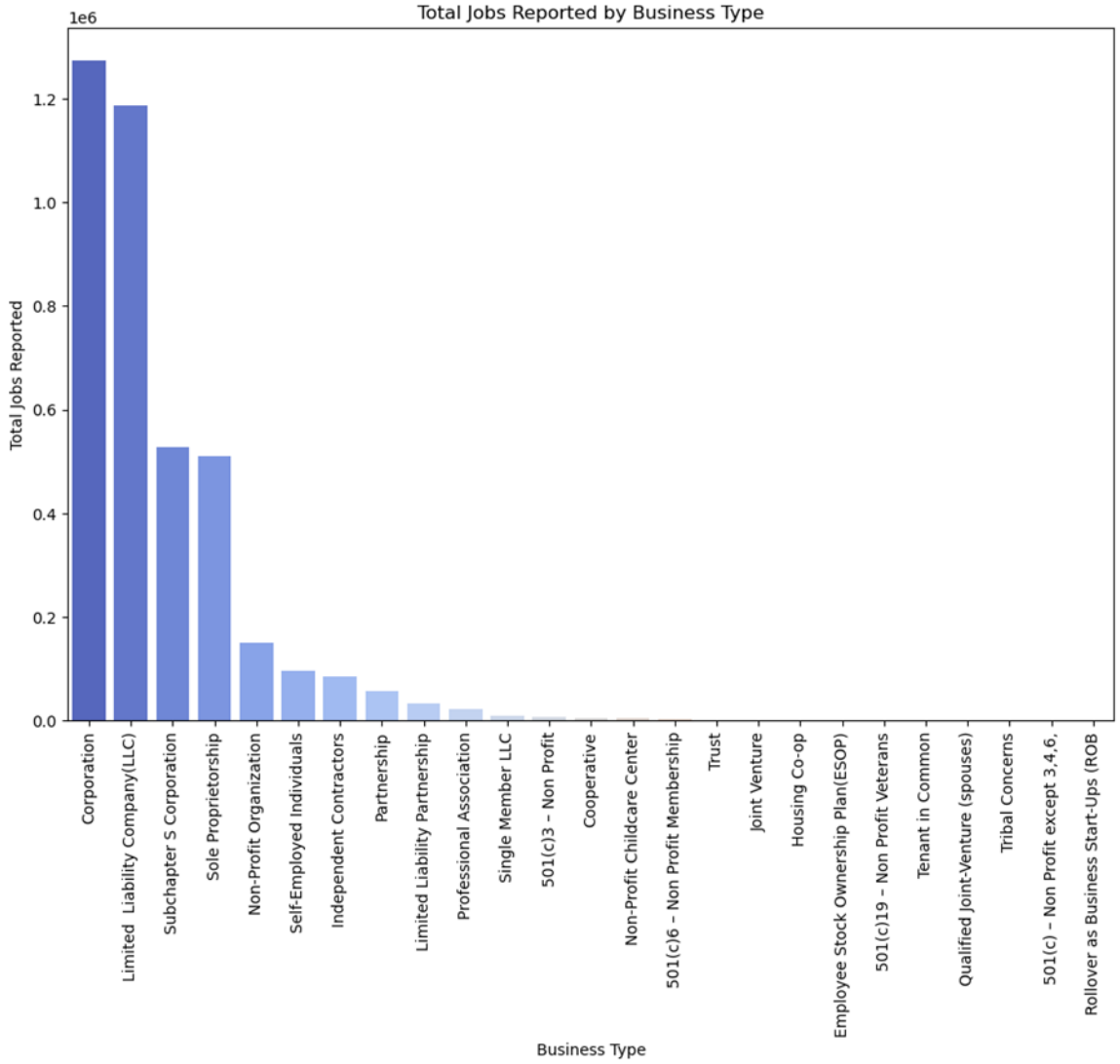
Next, we see a bar plot of the number of Loans by Business Type. We can see the distribution of PPP loans across various business types.
So in the graph below we see that Sole Proprietorships received the highest number of loans, exceeding 300,000, while Corporations and Limited Liability Companies (LLC) 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



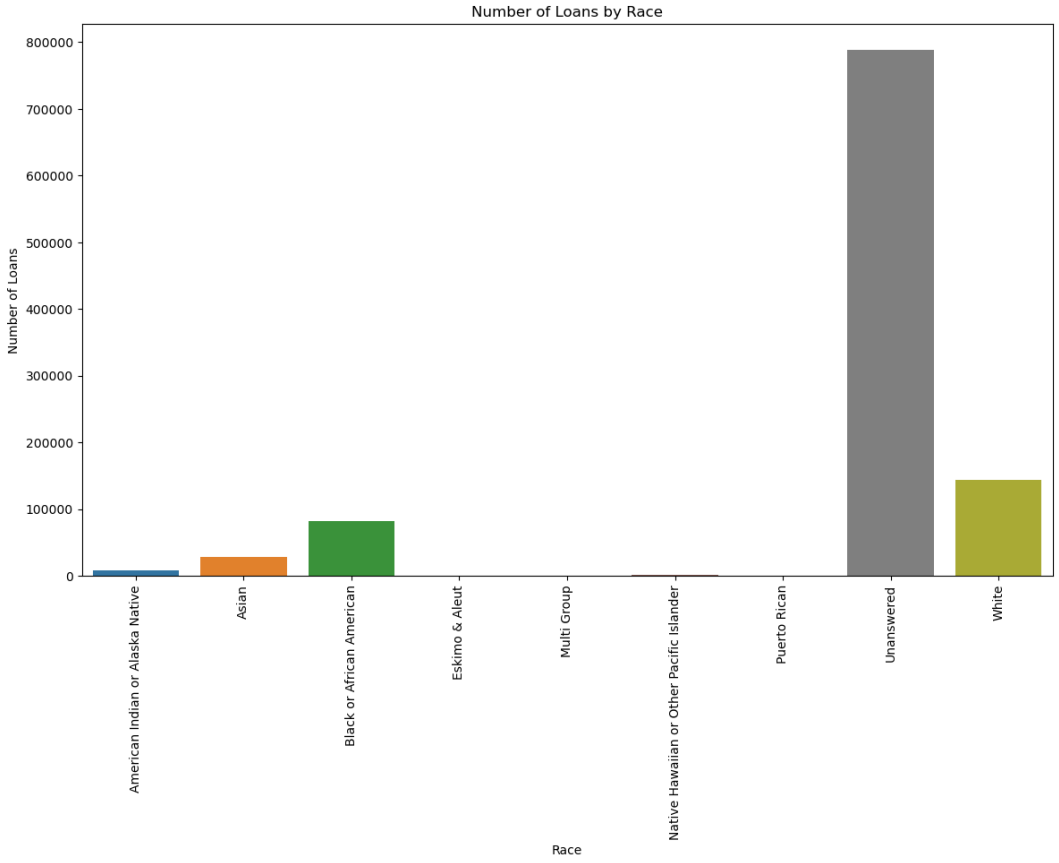
Next, we did a bar plot to illustrate the total number of jobs reported by different business types under the Paycheck Protection Program (PPP). The majority of reported jobs are from Corporations, Limited Liability Companies (LLCs), and Subchapter S Corporations. These business types account for the 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



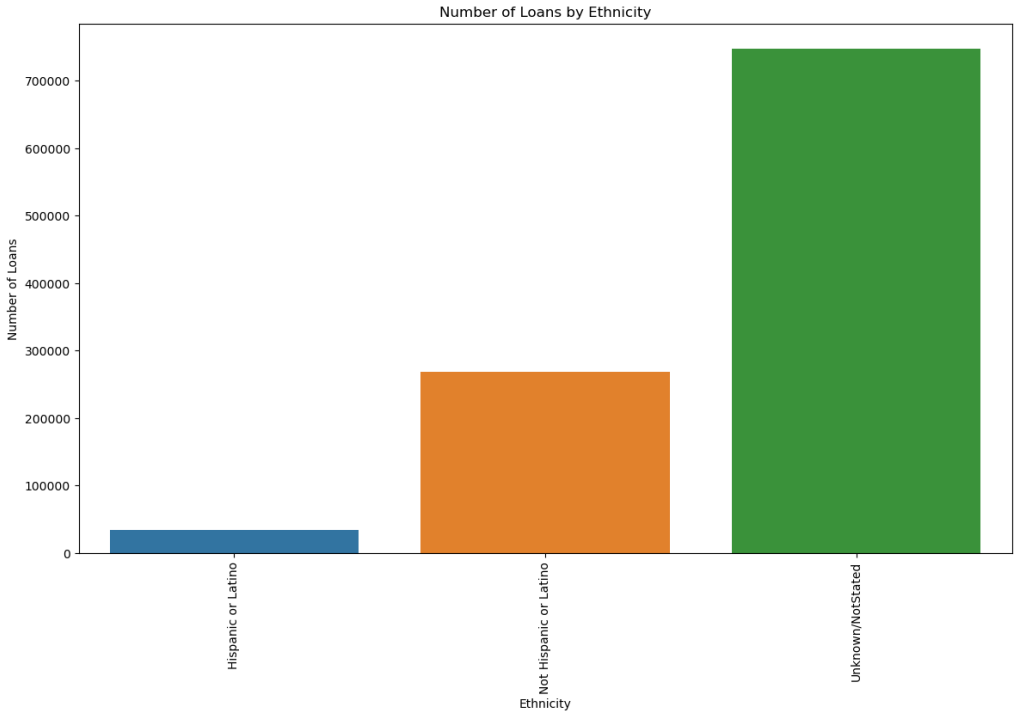
In addition, we illustrated the distribution of PPP loans across different racial groups. The majority of the loans fall under the "Unanswered" category, indicating a significant amount of missing data in terms of racial identification.
Among those who specified their race, White borrowers received the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

Figure 6: Number of Loans by Race



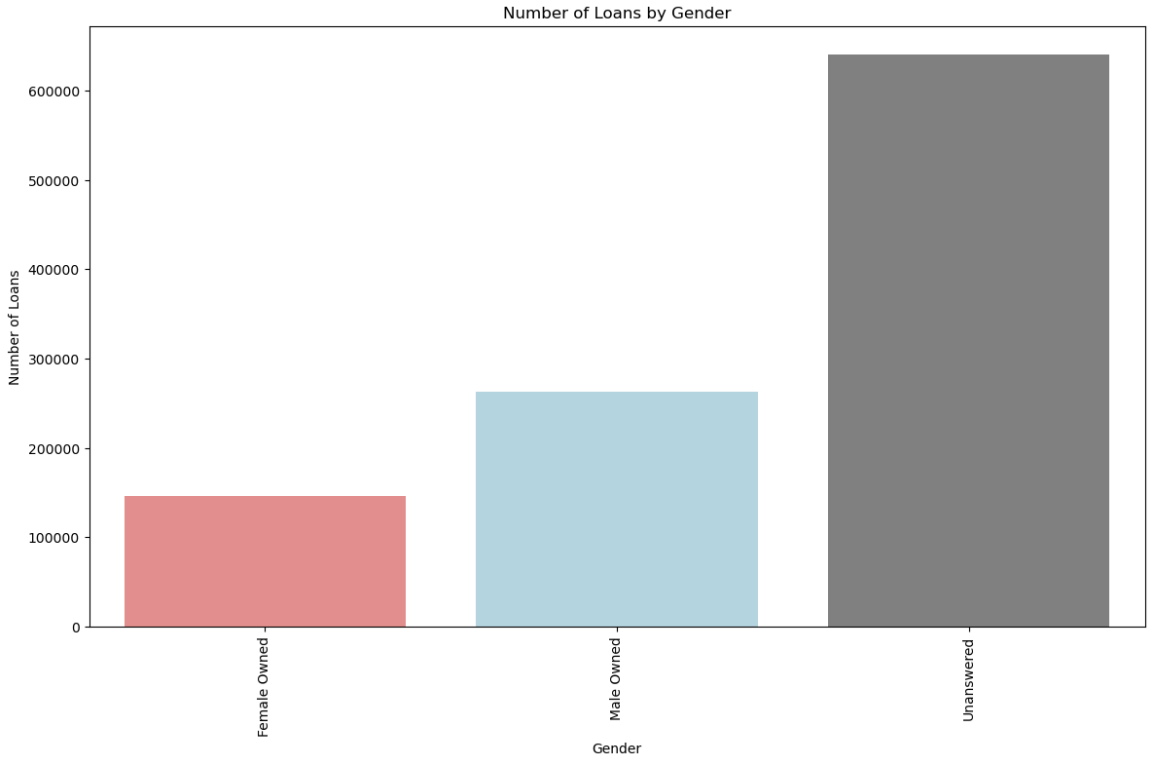
After that, we plotted the distribution of PPP loans across different ethnicities.
The majority of the loans fall under the "Unknown" category, indicating that we do not have many information about the ethnicity of the loan borrowers.
Among those who know their ethnicity, Not Hispanic or Latino borrowers 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 number 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 is a problem when we will try to merge those. We have to do again 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 convert the data type of the columns to display numbers with fewer decimal points and in the right format, 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

Borrower State	Total Loan Amount	Population Estimate 2023	Total Loan Amount Per100k Residents
North Dakota	115,378,856	1,567,852	7,359,040.01
South Dakota	131,262,912	1,838,636	7,139,146.19
Nebraska	231,835,296	3,956,758	5,859,223.54
District of Columbia	79,343,640	1,579,444	5,842,924.30
Illinois	1,412,761,728	25,099,378	5,628,672.26
Wyoming	64,247,328	1,168,114	5,500,090.57
Georgia	1,194,599,040	22,058,454	5,415,606.37
Iowa	343,131,936	6,414,008	5,349,727.28
Louisiana	478,641,152	9,147,498	5,232,481.63
Vermont	65,222,280	1,294,928	5,036,749.53
Alaska	72,090,384	1,466,812	4,914,766.45
Montana	110,943,240	2,265,624	4,896,807.24
Florida	2,194,535,680	45,221,452	4,852,864.26
Mississippi	279,033,376	5,879,380	4,745,966.00
New York	1,854,199,936	39,142,432	4,737,058.59
Minnesota	524,843,104	11,475,830	4,573,465.31
New Hampshire	125,534,632	2,804,108	4,476,811.59
Kansas	261,100,336	5,581,092	4,439,657.40
Connecticut	320,844,256	7,234,352	4,435,010.30
Massachusetts	617,847,360	14,002,798	4,412,313.60

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.