ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

Data Engineering Homework 2

By Antonis Xylas
AM:F3612314
MSc in Statistics
Professor: Stefanos Kechagias

Introduction

The COVID-19 pandemic has had a profound impact on the global economy, with businesses across the United States facing unprecedented challenges. In response, the US federal government established the Paycheck Protection Program (PPP) in 2020 as part of the Coronavirus Aid, Relief, and Economic Security (CARES) Act. The PPP is a $1 trillion business loan initiative designed to provide financial support to businesses and sole proprietors, ensuring they can continue paying their employees during the economic downturn caused by the pandemic.

The primary objective of the PPP is to prevent mass layoffs and unemployment by offering forgivable loans to cover payroll costs, rent, utilities, and mortgage interest. By maintaining the workforce, the program aims to stabilize the economy and expedite recovery once public health measures are lifted.

This report provides a comprehensive analysis of the financial aid distributed through the PPP, focusing on various geographical levels, including national, state, and local perspectives. Additionally, the report delves into demographic insights and industry-specific impacts, highlighting how different sectors and populations have benefited from the program.

Data

For PPP data

In this report, we detail the comprehensive process undertaken to collect, clean, and sample data from various CSV files related to the Paycheck Protection Program (PPP). The download process involved verifying the success of each file retrieval by checking the HTTP response status, ensuring that all datasets were saved in the designated CovidRecovery/RawData directory without errors. We then proceeded to check each dataset for the presence of NaN values, an essential step to guarantee data integrity. Datasets containing NaNs were cleaned by removing incomplete rows, significantly improving data quality. This integrated, sampled dataset offers a manageable yet representative subset, enabling efficient analysis while maintaining data integrity.

Following the initial steps of data collection, loading, cleaning, and sampling, we proceeded to further preprocess the datasets by focusing on specific columns relevant to our analysis. We defined a set of columns to retain, ensuring that we concentrate on the most pertinent data points for the Paycheck Protection Program (PPP) analysis. The columns selected for retention included DateApproved, BorrowerState, InitialApprovalAmount, CurrentApprovalAmount, ServicingLenderState, JobsReported, BusinessType, Race, Ethnicity, Gender, and Veteran.

For all information on job postings from the Opportunity Insights

The same procedure will be followed for data relating to jobs. To streamline the data for analysis, specific columns were retained: year, month, day end of week, statefips, posts, and state. These columns were selected for their relevance to tracking job postings over time and across different states, providing valuable insights into economic recovery trends.

Population estimates of US Counties from US Census Bureau

To enhance our analysis, we incorporated population estimates data sourced from the U.S. Census Bureau. We focused on retaining only the relevant columns: STATE and POPESTIMATE2020, which provide the state identifiers and population estimates for 2020, respectively. we have chosen the year 2020 to see how the pandemic period has affected the labour sector and the economy.

Exploratory Data Analysis

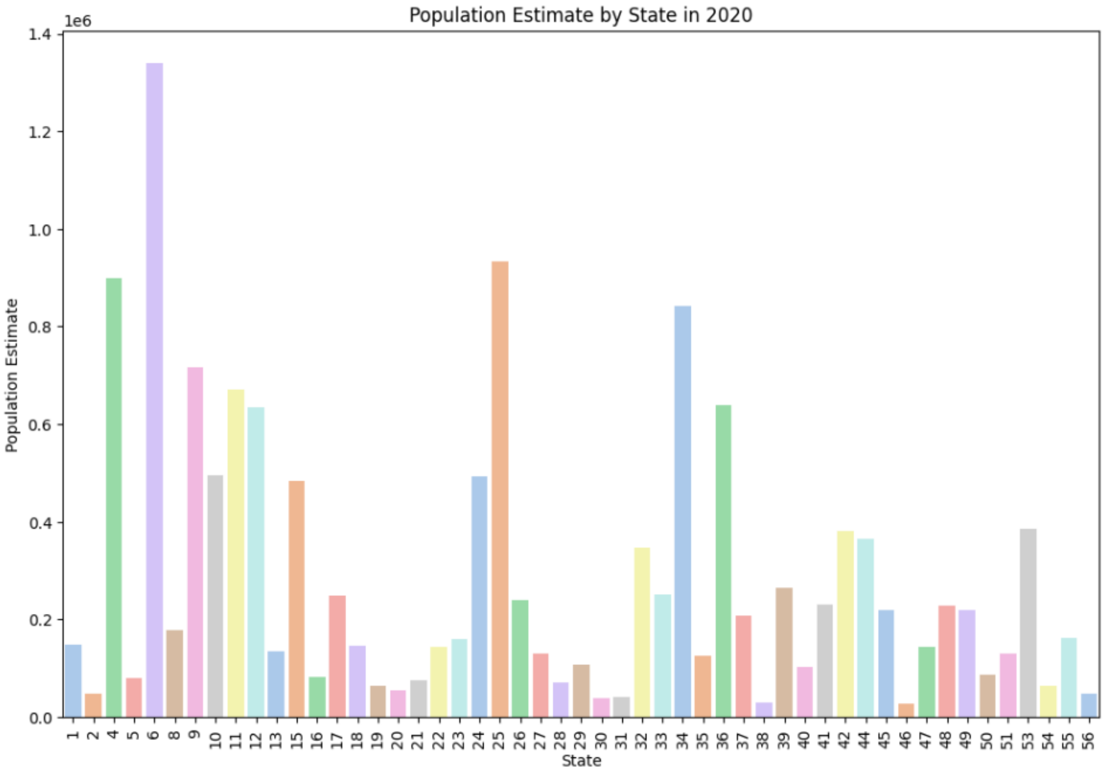

Figure 1: Population estimate barplot

The bar plot illustrates the population estimates for various states in the year 2020. Each bar represents a state, and the height of the bar corresponds to the population estimate for that state. There is a noticeable variation in population estimates across different states, with some states having populations under 200,000, while others exceed 1 million. Many states fall into a mid-range population estimate, demonstrating a diverse distribution of population across the country.

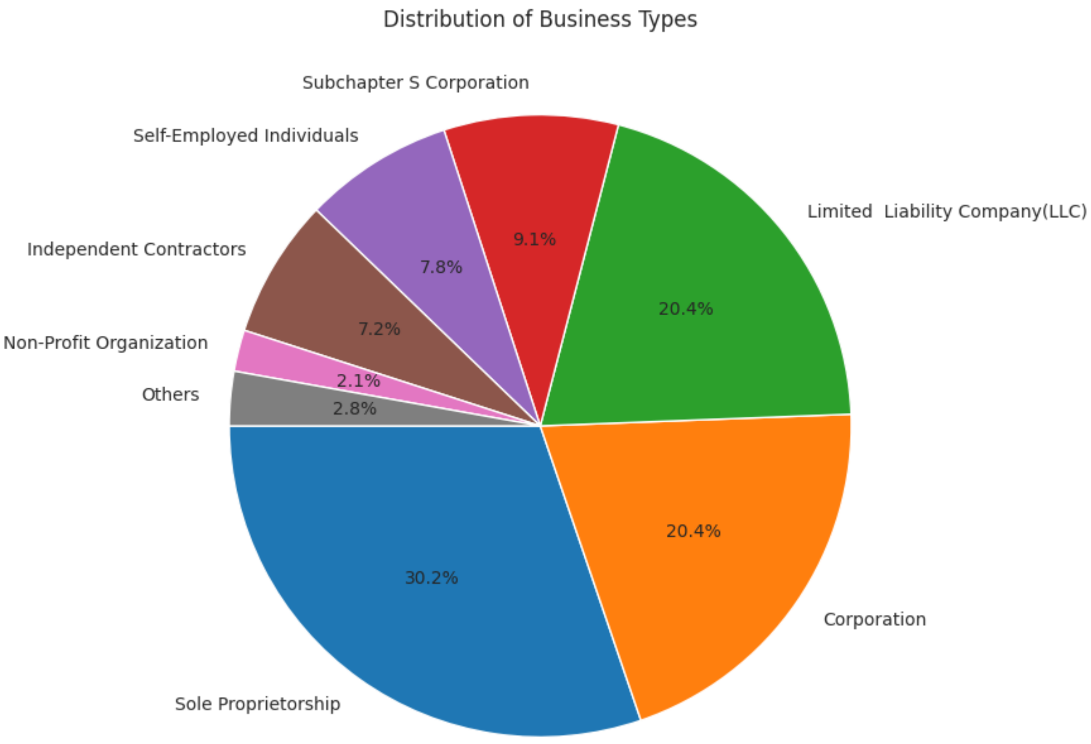

Figure 2: Distribution of business types

-Sole Proprietorship: The largest segment of the chart, constituting 30.2% of the total, indicates that Sole Proprietorship is the most common business type.
-Corporation and LLC: Both Corporations and Limited Liability Companies (LLCs) are equally prevalent, each making up 20.4% of the total. This suggests that these two business structures are also very popular among business owners.
-Subchapter S Corporation: Comprising 9.1%, this segment represents a significant portion of businesses choosing this tax-advantaged structure.
-Self-Employed Individuals and Independent Contractors: Representing 7.8% and 7.2% respectively, these segments highlight the presence of a considerable number of self-employed professionals and contractors.
-Non-Profit Organizations: At 2.1%, this segment indicates a smaller but important portion of businesses operating on a non-profit basis.
-Others: The smallest segment, "Others," at 2.8%, encompasses various other business structures not categorized separately in the chart.

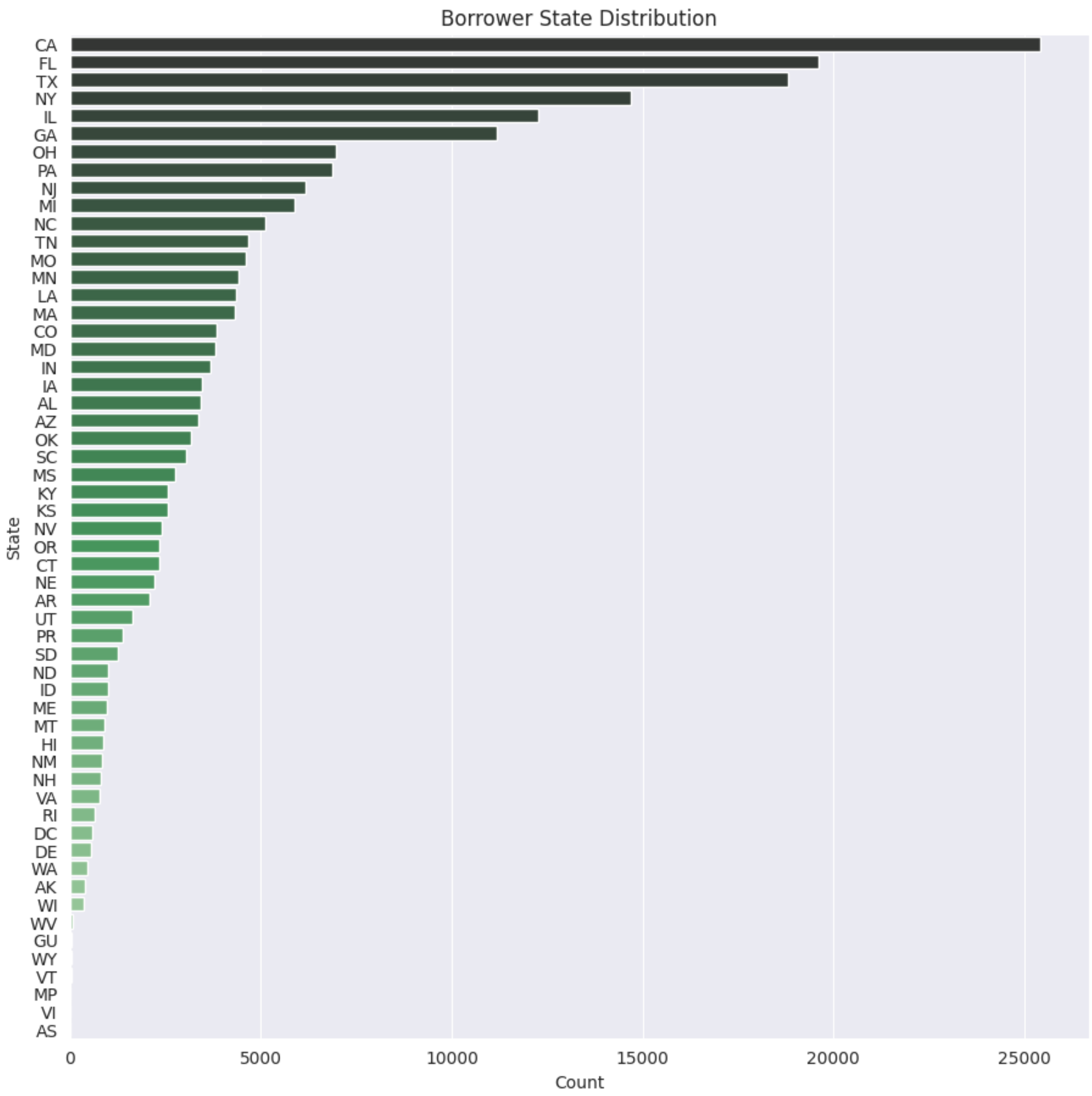

Figure 3: Borrower State Distribution

With over 20,000 borrowers, California (CA) stands out as the state with the highest number of borrowers. Texas (TX) and Florida (FL) come in second and third place, respectively. Both states have significant borrower counts, reflecting their diverse demographics and strong housing markets. States like Illinois (IL), Ohio (OH), and Georgia (GA) also contribute to the borrower distribution. These states play a crucial role in the overall lending landscape. While the larger states dominate, it’s interesting to see smaller states like Rhode Island (RI), Delaware (DE), and Wyoming (WY) represented.

Merge columns

By mapping state FIPS codes to state abbreviations and merging the PPP loan data with population estimates, we achieved a comprehensive dataset that combines economic support information with demographic context. This enriched dataset provides a deeper understanding of the distribution and impact of PPP loans across different states, enhancing the overall analysis and insights derived from the data.