Health Data Science

Data Engineering Homework 2

Kalatzi Marilena

AM: $f3612303$

Professor: Kechagias S.

Friday 31st May, 2024

Introduction

The Paycheck Protection Program (PPP) was a $1 trillion business loan initiative launched by the US federal government in 2020 to support businesses and sole proprietors in maintaining their workforce during the economic downturn caused by the COVID-19 pandemic. This report provides a comprehensive analysis of the financial aid distributed through the PPP, focusing on state level. Additionally, we offer insights regarding the demographics that benefited from this program.

In order to analyze the financial aid distributed through the PPP, we merged different datasets covering public loans. We then assessed the distribution of these loans with exploratory data analysis.

Exploratory Data Analysis

Borrower State

Firstly, we counted the loans provided to each state using the information of the borrower state. We plotted the counts with the following results:

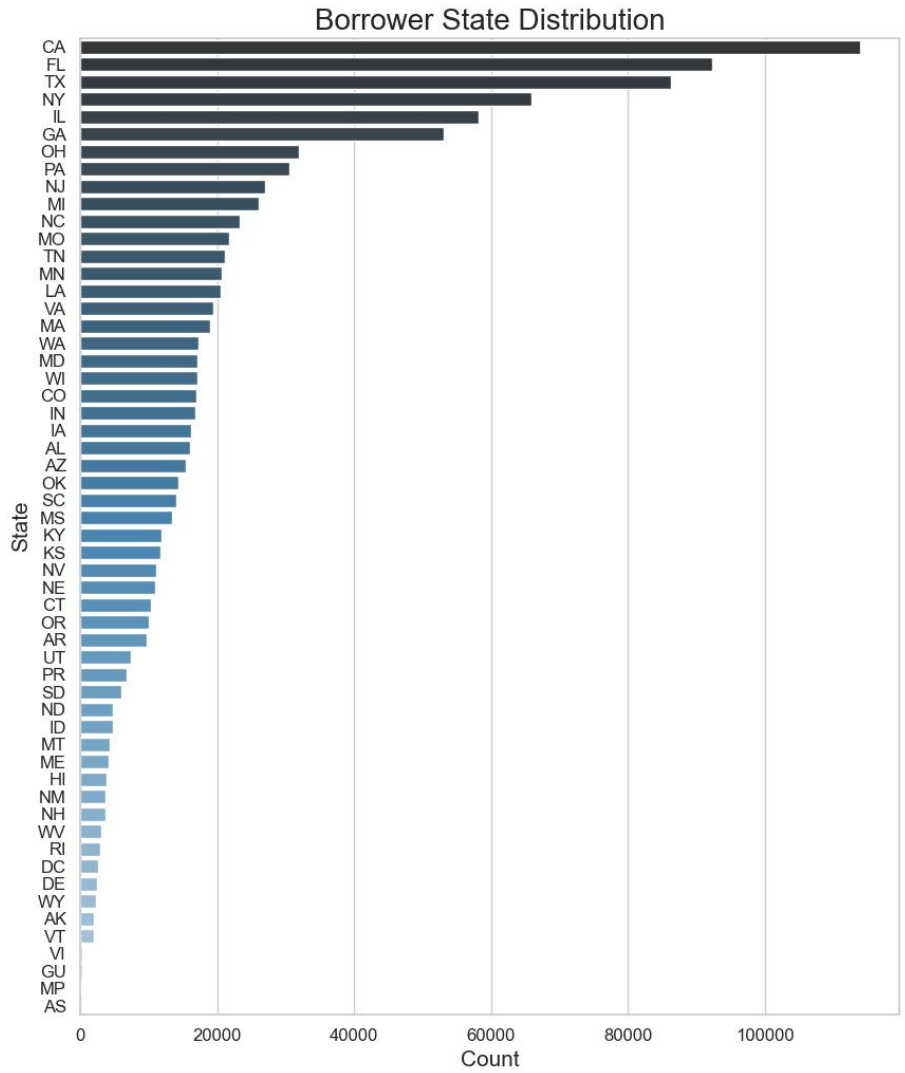

Figure 1: Borrower state distribution

The highest number of loans were distributed in California (CA), followed by Florida (FL), Texas (TX), and New York (NY). This distribution reflects the larger populations and economic activities in these states, while states with smaller populations and economies received fewer loans. These information helps understand regional disparities in loan distribution and can inform future financial aid programs to ensure more support across different regions.

We proceed with exploring the demographics in more depth.

Ethnicity

The loan distribution regarding the ethnicity is shown below:

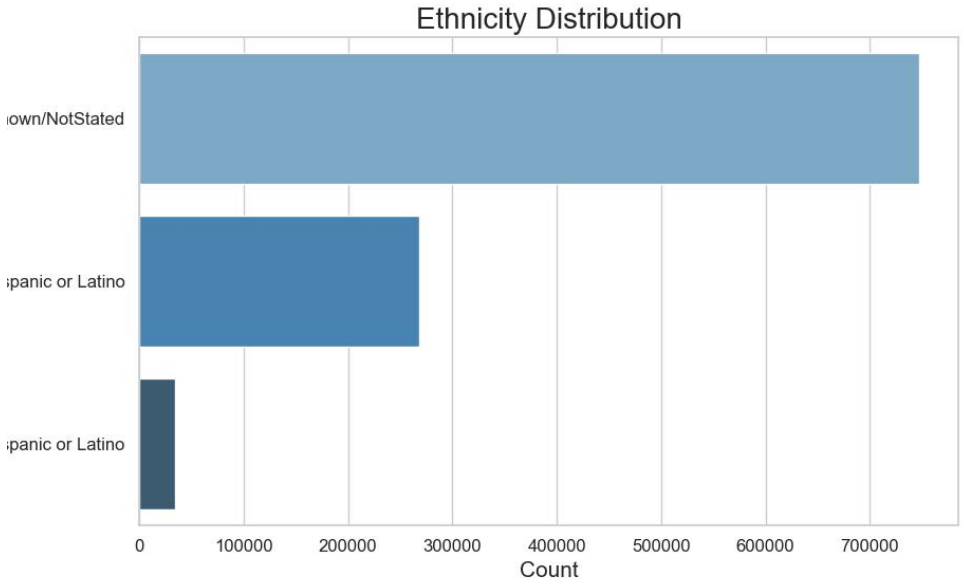

Figure 2: Ethnicity distribution

The Ethnicity Distribution plot reveals that the majority of the loans were provided to individuals that did not report their ethnicity, categorized as "Unknown/NotStated". Among those who did report, there is a significant portion identifying as "Not His-

panic or Latino", followed by a smaller group identifying as "Hispanic or Latino".

Race

The loan distribution by race is shown below:

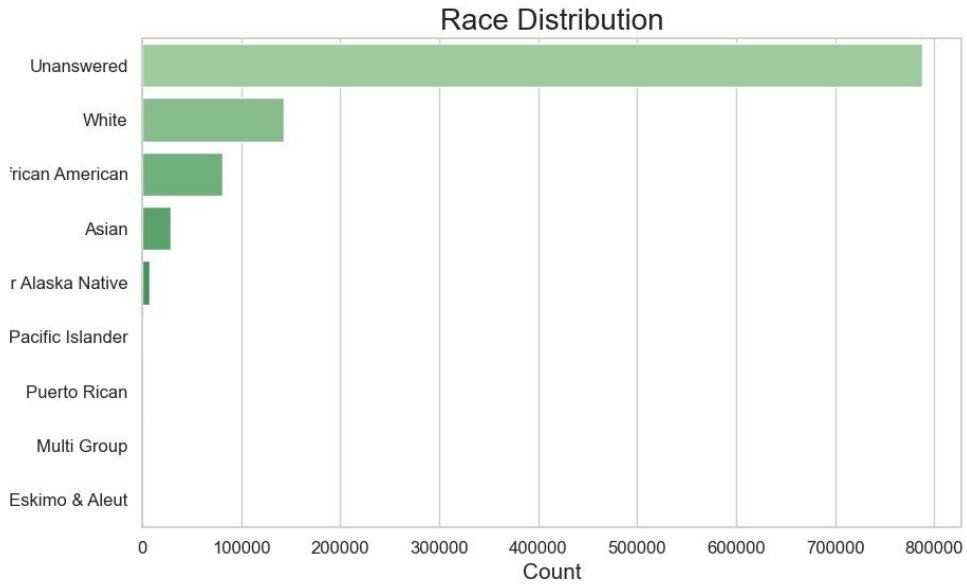

Figure 3: Race distribution

The Race Distribution plot shows that the vast majority of the loans were given to individuals that did not disclose their race, categorized as "Unanswered". Among those who reported their race, "White" and "Black or African American" were the most common, followed by "Asian". The large "Unanswered" category suggests a significant gap in the demographic data, which could be crucial for assessing the equity of the loan distribution.

Gender

Finally, the gender loan distribution is presented:

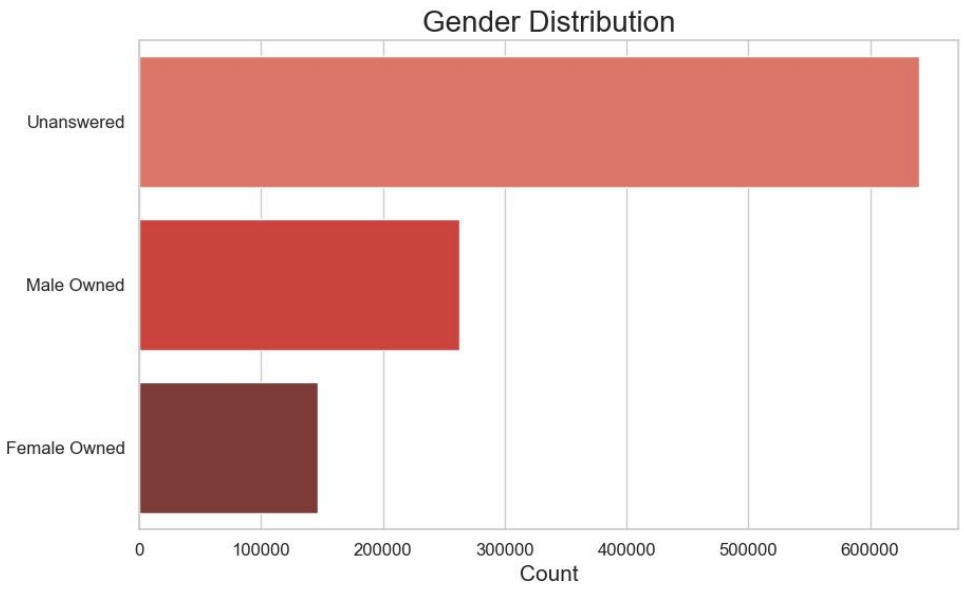

Figure 4: Gender distribution

The Gender Distribution plot highlights that a substantial number of loan recipients did not report their gender, labeled as "Unanswered". Among those who did, "Male Owned" businesses received more loans compared to "Female Owned" businesses. This distribution might point to gender disparities in business ownership or loan approval rates that warrant further investigation.

Conclusion

Overall, these plots provide a clear visual representation of the PPP loan distribution across various demographics and geographic locations. The significant amount of "Unanswered" data in ethnicity, race, and gender categories highlights a crucial gap in the dataset. Addressing these gaps in future data collection could improve the analysis of the program's effectiveness and equity. Additionally, the state-wise

distribution underscores the importance of considering regional economic contexts in designing financial support initiatives. Finally, future analysis should also account for the population distribution within each category. For instance, normalizing the loan counts by the respective population sizes of each ethnicity, race, and gender would allow us to determine if certain groups are over- or under-represented in the loan distribution relative to their population proportion. Similarly, considering the state-wise population when analyzing the Borrower State Distribution would provide a more accurate picture of regional disparities. By integrating population context into the analysis, we can gain deeper insights into the equity and effectiveness of the PPP loan program, ensuring that future financial support initiatives are better targeted and more inclusive.