# Investigating COVID-19 Paycheck Protection Program Loans

**Lauren Lunsford**

## Motivations

With the onset of Covid-19 and increased social safety guidelines, the economy and many businesses suffered. Paycheck Protection Program (PPP) loans were extended to business to provide relief during Covid-19. I am interested in understanding the demographic breakdowns of the PPP loans to examine if there is a relationship between race, population size, income and the size of the approved PPP loan.

## Data

The initial project proposal included a zillow dataset for the purposes of gathering median home prices as a proxy for economic status of each county. After a more thorough look, income statistics were included in the census data, and the Zillow dataset was no longer necessary.

**PPP Application Data**
Location: PPP Loan Data (Paycheck Protection Program)
Format: Returned in a CSV
Variables: Rows are approved PPP loan. The columns used are the zip code, a range of the approved loan size.
There is demographic information, however, most of the information is missing (~86% of values are missing for race & ethnicity). Therefore, instead of using the incomplete R&E data per loan, I aggregate loans by county and use census data to find county demographic data.
Time: The data is from 2017

**Zip Code to County Data**
Location: US Zipcode to County State
Format: Returned in a CSV
Variables: Rows are instances and columns are zip codes and counties. This will be used to convert PPP zip codes into counties.

**State Abbreviations**
Location: List of State Abbreviations
Format: Returned in a CSV
Variables: Rows are states and the used columns are the full state name, and the two letter abbreviation.
I needed to ensure that counties with the same name in different states were separate, so I needed an ID by state which required creating a column with a common reference type to state.

**Census Data**
Location: Census_Data
Format: Returned in a CSV

Variables: Rows are counties. Columns are total population, state, number of men, number of women, number of White residents, number of Black residents, number of Native residents, number of Asian residents, number of Pacific residents, income, income per capita.
The purpose of this dataset is to gather demographic information. Most PPP applications did not have relevant demographic data (~86% for race & ethnicity). Therefore, an additional metric may provide insight. The census dataset includes demographic information by county such race and gender.

## Data Manipulation

To clean my data I used Google CoLab and Pandas. I first imported the PPP data and selected the loan range and zip code columns to create a smaller dataframe called **ppp_small**. The value of the approved PPP loan is in a range instead of an exact number. To change the data from categorical to numerical, I created a new dataframe called **ranges_values** that included each of the loan ranges and the corresponding average of the boundaries. The negative externalities of this maneuver are discussed in challenges. I joined **ranges_values** and **ppp_small** into a dataframe called **joined**, and cleaned up the column names so that the PPP data had numerical values. There was no missing data at this stage.

Then I imported the zip code to county data and selected the relevant columns (zip code, county name, and two letter state code) into a dataframe called **ziptocounty**. I then merged this with **joined** (the dataframe that included the PPP data) on the zip code to get all county and state names for each instance of a PPP loan. Since multiple states have counties with the same names, I created a unique ID by combining the county name with the two letter state ID to create a column called ID. All this information is held in the **values_county** variable. I checked to make sure this has no null values.

The census data has only full state names, not the two letter state codes. To create the unique identifier that I want the census data and the PPP data to merge on, I need to translate the full state names to the state codes. I do so by importing state abbreviations data and selecting the state and code into the **state** variable. I then import the census data. It is organized by county so I read it in as **bycounty**. I join the census data (**bycounty**) and the state abbreviation data (**state**) by state name into the variable **state_abb**, where I concatenate the county name and state code to create a unique identifier similar to those formed in **values_county** (the PPP data)**.** I also select for the relevant columns in this step.

Finally, I merge the PPP data, **values_county**, and the census data, **bycounty** on the unique county ID such that each instance of a PPP loan gets the relevant county's demographic information.  I then export to a csv file that is uploaded to the cavium cluster for analysis.

## Computations

Please note: I have structured my SparkSQL such that each query is a letter. I will reference the queries relevant to each computation through a bolded letter in parentheses.

1. Which counties have the most approved PPP loans?

To compute this, I used SparkSQL. I loaded data.csv into a dataframe (**data)**. From that I selected ID, and counted the instances of a loan across ID (county + state) to find the number of rows (number of approved loans) per county. I ordered by the number of approved loans in descending order in a new table **(B)**. The top ten results are below.

| County, State | Number of Approved PPP Loans |
|---|---|
| Los Angeles County, CA | 24803 |
| Cook County, IL | 14341 |
| New York County NY | 13082 |
| Harris County TX | 11493 |
| Orange County CA | 10265 |
| Maricopa County AZ | 8593 |
| Dallas County TX | 8205 |
| San Diego County CA | 7773 |
| Miami-Dade County FL | 7233 |
| King County WA | 6653 |

The top county is Los Angeles which is one of the most populated counties in the country. The other top counties include highly populated cities which indicates that perhaps the more populous a county, the more loans it received.

2. Is there a relationship between the number of loans and the population of each county?

After seeing the above data, I wanted to check my assumption that there was a strong, positive correlation between population size and the number of approved loans. To check this, I performed a correlation analysis. I imported **Correlation** from **pyspark.ml.stat**. I created a new dataframe (**J)** where I grouped the data by county again and selected the counted instances of ID (county + state) and selected the average of the total population since the total population is the same across each PPP loan instance. I then used **J.corr()** and the two selected columns to calculate a correlation coefficient **(K)**.

The correlation coefficient was r= 0.951, which confirmed my assumption that there was a strong, positive correlation between population and number of approved loans. Since there are more people, it makes sense there would need to be a number of businesses proportionate to the population to service the people and provide jobs and that those businesses would need loans.

3. Which counties have approximately the highest PPP loan values per capita?

To identify this, I selected the sum of the PPP values per county and divided it by the average total population (which yields simply the total population per county), and the county ID (**N)**.

| County, State | Approved PPP Loan Size Per Capita |
|---|---|
| San Juan County CO | 100131.8 |
| McPherson County NE | 57638.8 |
| Sioux County NE | 39251.5 |
| Dunn County ND | 33915.8 |
| Roberts County TX | 28740.1 |
| Irion County TX\| | 28721.7 |
| Kent County TX | 26250.0 |
| Lane County KS | 25822.5 |
| Billings County ND | 23296.0 |
| Hodgeman County KS | 21466.6 |

I was curious as to what made San Juan County, Colorado so unique. I googled it and found it has the smallest population in a county (711). For fun, I very quickly calculated the correlation coefficient between the PPP Size per capita and the population (**P**). If small counties had an inherent advantage in the metric, then there would be a (likely weak) negative correlation. The R value was -0.024, indicating practically no relationship. This means that San Juan County was a likely outlier in the data.

4. Is there a relationship between the racial diversity of a county and the approved PPP loan values per capita? (D,E,F)

To identify this, I created a new column to use as a metric of racial diversity that summed the percents of non-white residents **(A)**. I then selected the county ID, total PPP value per capita, and the non-white percent calculated in **A**, and ordered by the total PPP value per capita in **(D).** From there, I selected only the PPP per capita, and non-white percent per county into **(E)** and calculated a correlation coefficient into **(F)**.

The R value was -0.063 which indicates no relationship. I was initially curious to see if there may have been racial disparity in PPP loan size. While the county data doesn't describe the business demographic, it is a small proxy that can be used as an initial indicator. Having no relationship means that there is no racial bias seen within this metric. However, of course, a more detailed analysis with more complete data may reveal a more complex perspective.

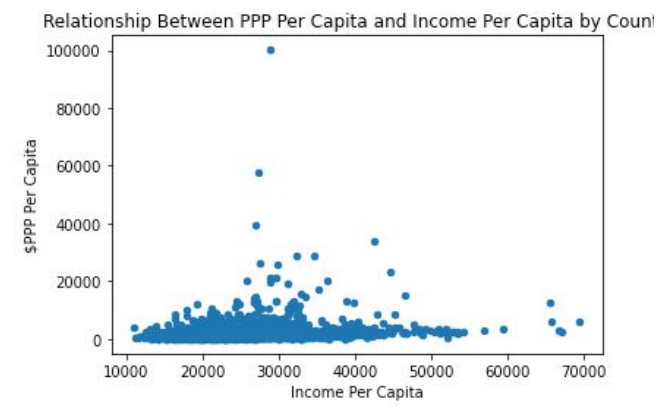
5. Is there a relationship between the PPP loan values per capita and the Income per Capita? (M)

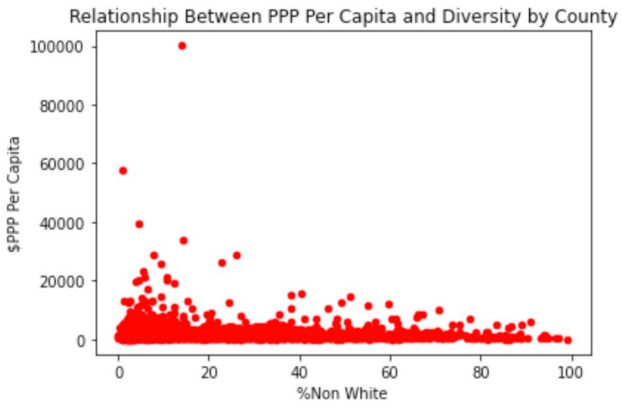
I was also curious to know if counties with higher incomes per capita led to higher PPP loans. I thought perhaps if the income of the area was higher, businesses would need more

money to pay the wages of those workers. To find this out I selected the average PPP loan per capita and the income per capita grouped by county ID (**L)**. Then I performed a correlation analysis (**M)**. The r value was r=0.165 which indicates no real relationship. This could be because not every business employs individuals from their immediate area as large corporations may be headquartered in other counties.

## Visualizations

I depicted scatterplots calculation 4 and 5. Neither of the two charts have any relationship between their X and Y variables.




## Challenges

**Ranges PPP Data:**

The PPP data is bucketed into size ranges. This made it incredibly difficult to perform the types of statistical analyses I wanted to. Proper statistical techniques would use the data as categorical data and perform an ordered logit analysis. However, I don't know how to do that on python, so, for the purposes of informal data exploration, I used the average value of the range to make continuous calculations possible. This, of course, is by no means the actual average of the PPP loans, however, without the data, that was the best option. By performing this, the calculations assume that there is no variability within each range. This generally makes any results found less exact. Results from this should be taken as indications of trends, not fact.