Analsyis_of_bankruptcies_by_ZIP

Jacob Steimer

2023-07-27

What causes high bankruptcy rates in some U.S. ZIP codes?

Introduction

During the height of the COVID-19 pandemic, the number of personal bankruptcy filings dropped significantly across the country. However, some parts of the country continue to see far more bankruptcies than others, in a way that doesn't seem clearly correlated with any one factor, according to data from the Federal Judicial Center. Also, people's choices of type of bankruptcy — Chapter 7 or Chapter 13 — vary widely.

The goal of this project is to investigate the factors contributing to varying bankruptcy rates in different U.S. ZIP codes, with a particular focus on understanding why some areas have higher bankruptcy rates than others. I explore the relationship between bankruptcy rates in U.S. ZIP codes and various demographic variables pulled from the U.S. Census Bureau. Using machine learning techniques, I explore the importance of the different variables and build models that can predict bankruptcy rates given Census data.

Among other purposes, this research was designed to inform a news article in MLK50: Justice Through Journalism that was published in September 2023.

Data Overview

I obtained bankruptcy data from the Federal Judicial Center website and demographic data from the U.S. Census Bureau. Because the FJC and Census website links are subject to change, I uploaded the relevant data from the websites to GitHub and provided code for data retrieval.

The bankruptcy data includes one row for every bankruptcy filed during 2022. Columns included the type of bankruptcy and the ZIP code connected with each debtor.

The Census Data included the following data for each ZIP code:

- The Labor Force Participation Rate (the percentage of adults who are either working or looking for work)
- The Unemployment Rate (The percentage of the Labor Force that isn't working)
- The Homeownership Rate (the percentage of households that own their residence)
- The percentage of residents with bachelor's degrees

- The percentage of residents who are Black
- The population and median income in each ZIP

I chose most of these factors because they're important economic characteristics. I chose the percentage of Black residents because prior research has shown African Americans are more likely to file Chapter 13 than Chapter 7.

Methods/Analysis

Data Preparation

To prepare the data for analysis, I determined the count of Chapter 13 bankruptcies, count of total bankruptcies and the percentage of Chapter 13 bankruptcies in each ZIP using "summarize" and "mutate".

Additionally, I transformed the median income variable for better interpretability: I divided the raw median income by 10,000 because a difference in $10,000 between ZIPs is more understandable than a $1 difference.

I then split the data into a holdout set, a training set and a testing set. Since the data only included about 18,000 observations, I used a 10% partition each time, to ensure both the holdout and testing sets were large enough.

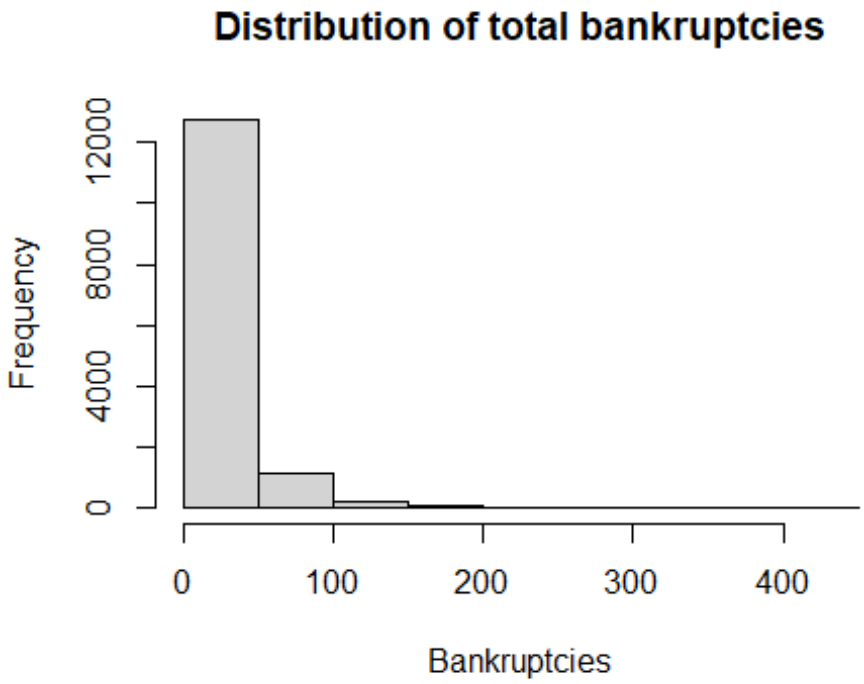
Data Exploration

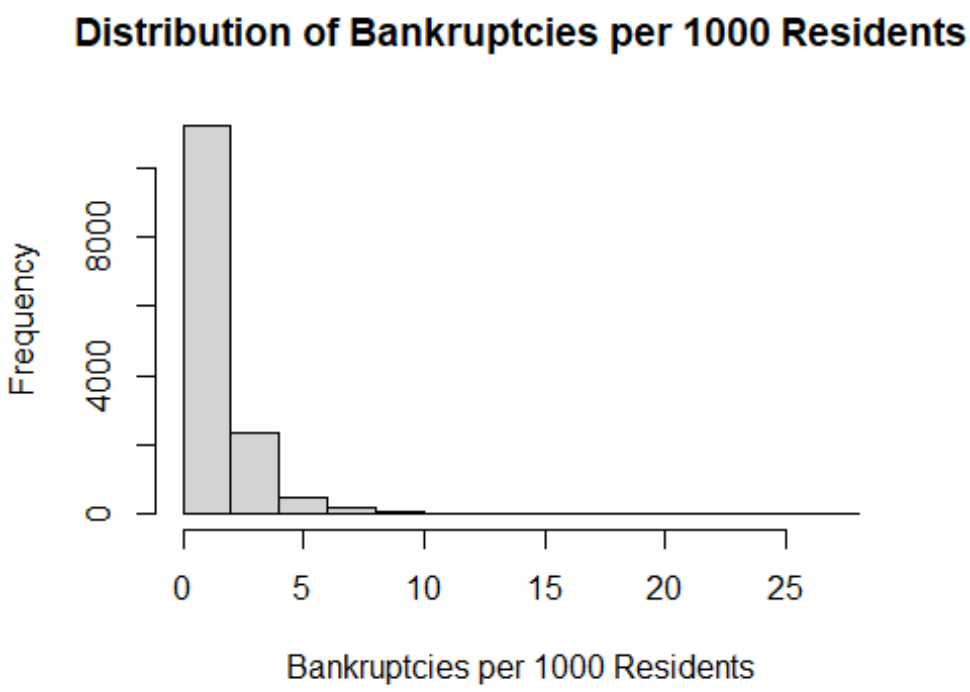
To start my data exploration, I produced a basic statistical summary of my dataset, using the summary() function. From the summary, the presence of outliers became clear. Also, it confirmed the wide variation in bankruptcy rates I had been told about.

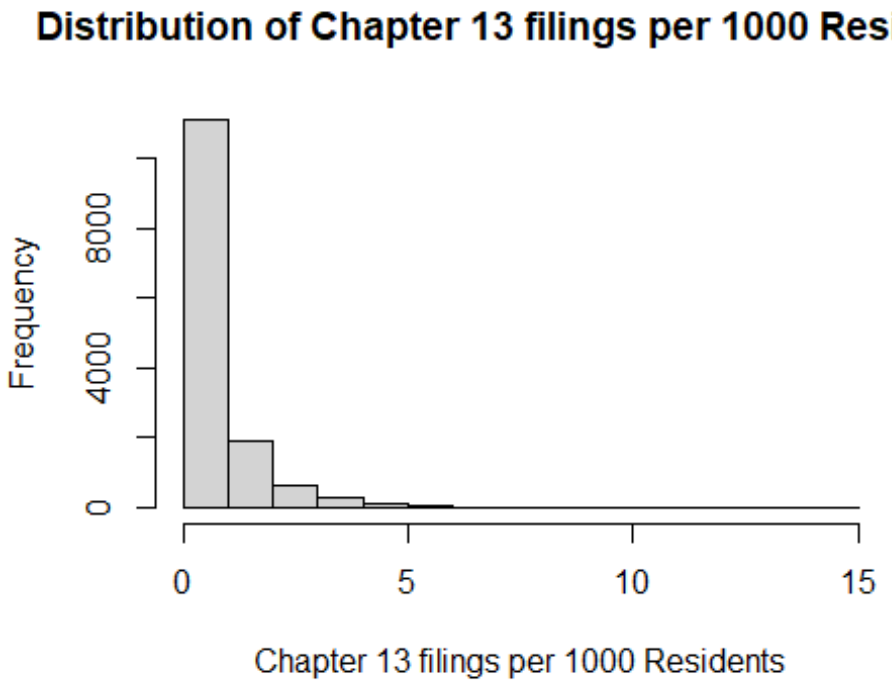
```
##     GEOID              Bankruptcies       Thirteens        ThirteenPerc      
##  Length:14232       Min.   :  1.00   Min.   :  1.000   Min.   :0.02174  
##  Class :character   1st Qu.:  4.00   1st Qu.:  2.000   1st Qu.:0.26667  
##  Mode  :character   Median : 11.00   Median :  4.000   Median :0.44444  
##                     Mean   : 21.14   Mean   :  8.886   Mean   :0.48424  
##                     3rd Qu.: 27.00   3rd Qu.: 10.000   3rd Qu.:0.66667  
##                     Max.   :446.00   Max.   :372.000   Max.   :1.00000  
##                                                                         
##  LabForceParticipation     UnRate          HomeOwnRate       BachPercent     
##  Min.   :0.0000        Min.   :0.00000   Min.   :0.0000   Min.   :0.0000  
##  1st Qu.:0.5546        1st Qu.:0.03315   1st Qu.:0.6279   1st Qu.:0.1615  
##  Median :0.6212        Median :0.04870   Median :0.7401   Median :0.2384  
##  Mean   :0.6066        Mean   :0.05568   Mean   :0.7081   Mean   :0.2852  
##  3rd Qu.:0.6718        3rd Qu.:0.06899   3rd Qu.:0.8255   3rd Qu.:0.3737  
##  Max.   :1.0000        Max.   :0.80000   Max.   :1.0000   Max.   :1.0000  
##                        NA's   :4         NA's   :11                       
##    MedIncome       BlackPercent         Population       BrupPerThou      
##  Min.   : 1.179   Min.   :0.000000   Min.   :   252   Min.   : 0.03568  
##  1st Qu.: 5.015   1st Qu.:0.007803   1st Qu.:  3750   1st Qu.: 0.69213  
##  Median : 6.323   Median :0.035532   Median : 11448   Median : 1.10998  
```

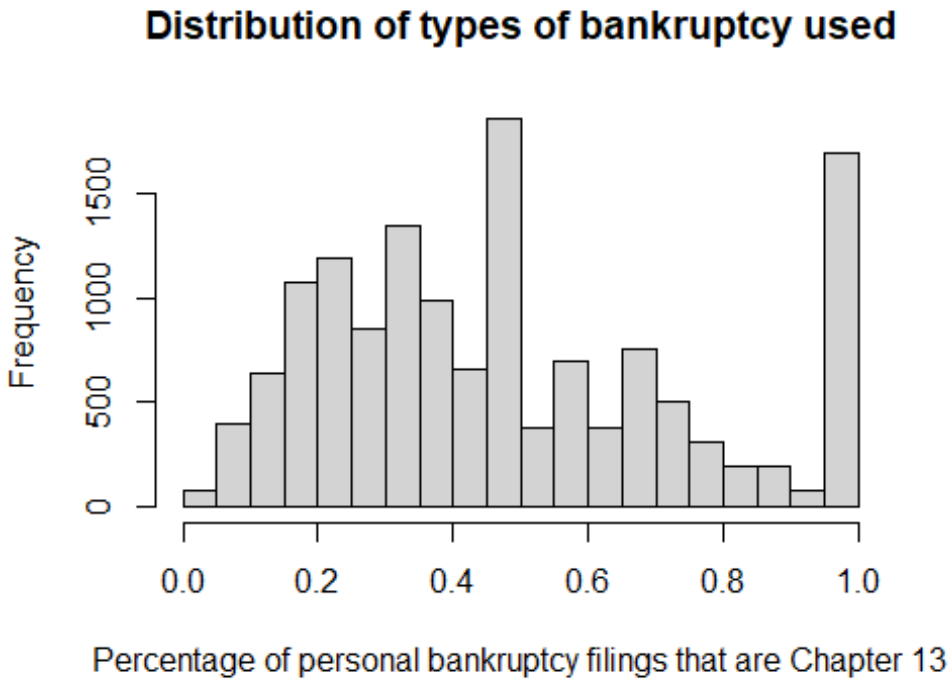
```
##  Mean   : 6.960   Mean   :0.115151   Mean   : 17355   Mean   : 1.47458
##  3rd Qu.: 8.264   3rd Qu.:0.136524   3rd Qu.: 26679   3rd Qu.: 1.80075
##  Max.   :25.000   Max.   :1.000000   Max.   :130352   Max.   :26.59575
##  NA's   :109
##  ThirteenPerThou
##  Min.   : 0.01074
##  1st Qu.: 0.22516
##  Median : 0.44322
##  Mean   : 0.76928
##  3rd Qu.: 0.89077
##  Max.   :14.33692
##
```

Next, I created histograms for my bankruptcy variables, to better visualize the range of outcomes I would be predicting and the variation of bankruptcy usage across the country.


Distribution of total bankruptcies
Frequency
0
4000
8000
12000
0
100
200
300
400
Bankruptcies

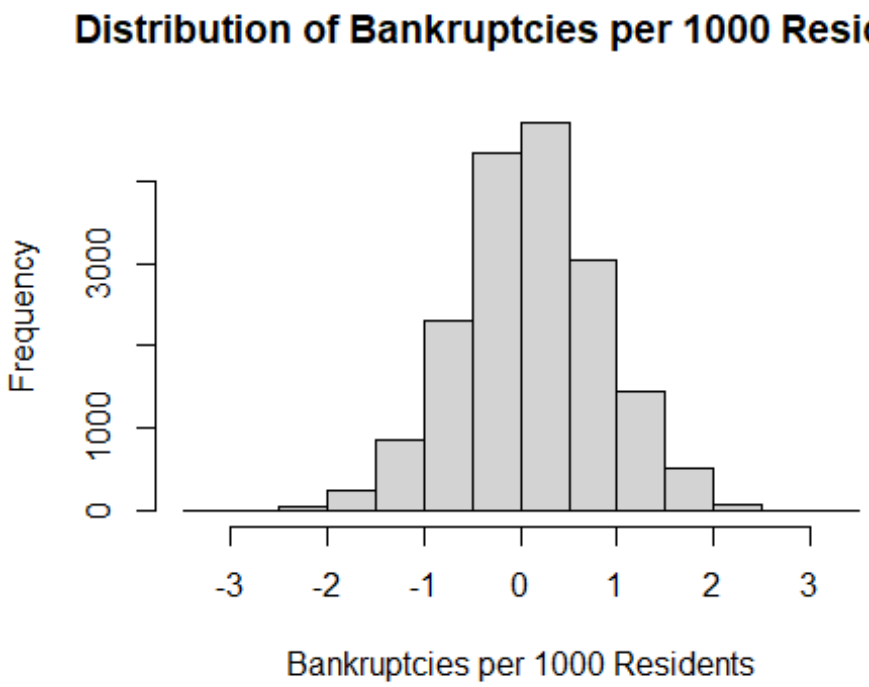

Distribution of Bankruptcies per 1000 Residents
Frequency
0
4000
8000
0
5
10
15
20
25
Bankruptcies per 1000 Residents

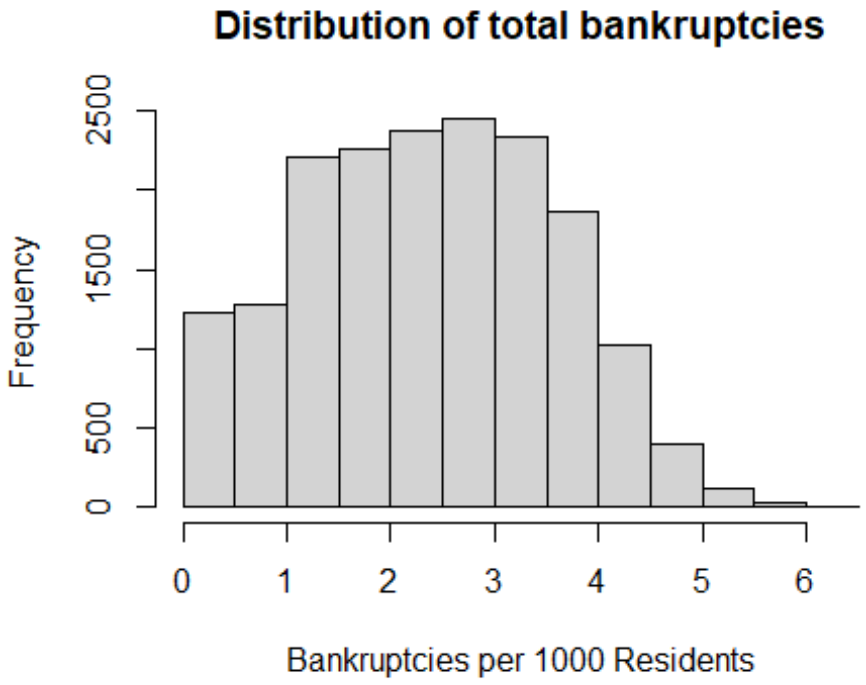


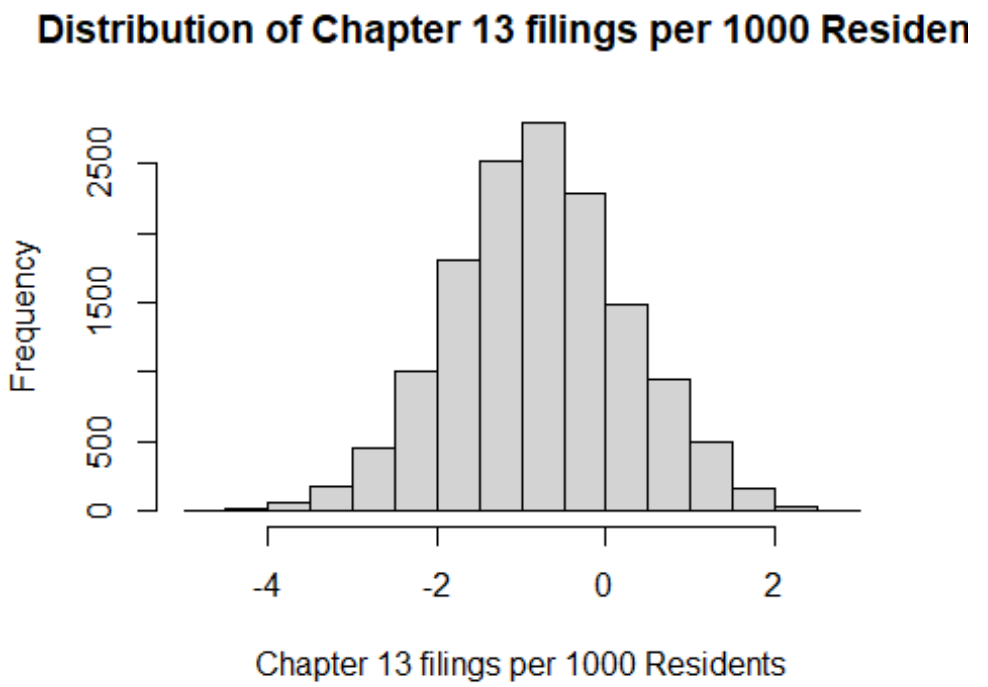


The fourth histogram, showing the percentage of filings that are Chapter 13, was quite useful. I didn't expect so many ZIPs to have exclusively Chapter 13 bankruptcies. And it also showed that fewer than 40% of bankruptcies are Chapter 13 in most ZIPs.

The other histograms didn't show much other than the presence of outliers. So, I decided to use logarithms to explore the data further.


Distribution of Bankruptcies per 1000 Residents
Frequency
0
1000
3000
-3
-2
-1
0
1
2
3
Bankruptcies per 1000 Residents

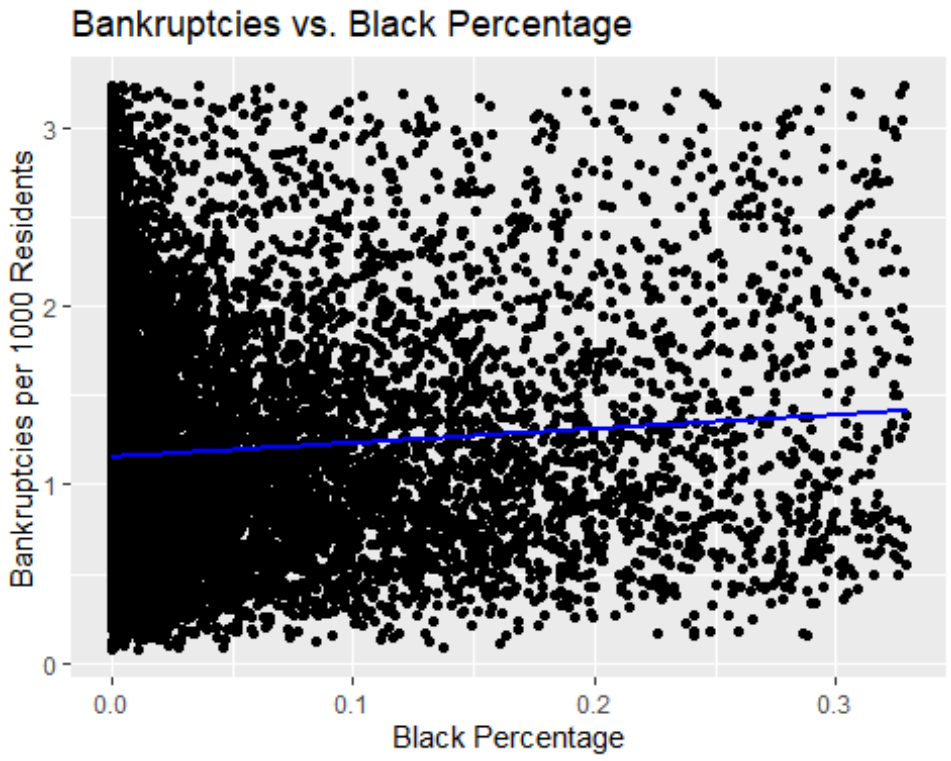

Distribution of total bankruptcies
Frequency
0
500
1500
2500
0
1
2
3
4
5
6
Bankruptcies per 1000 Residents

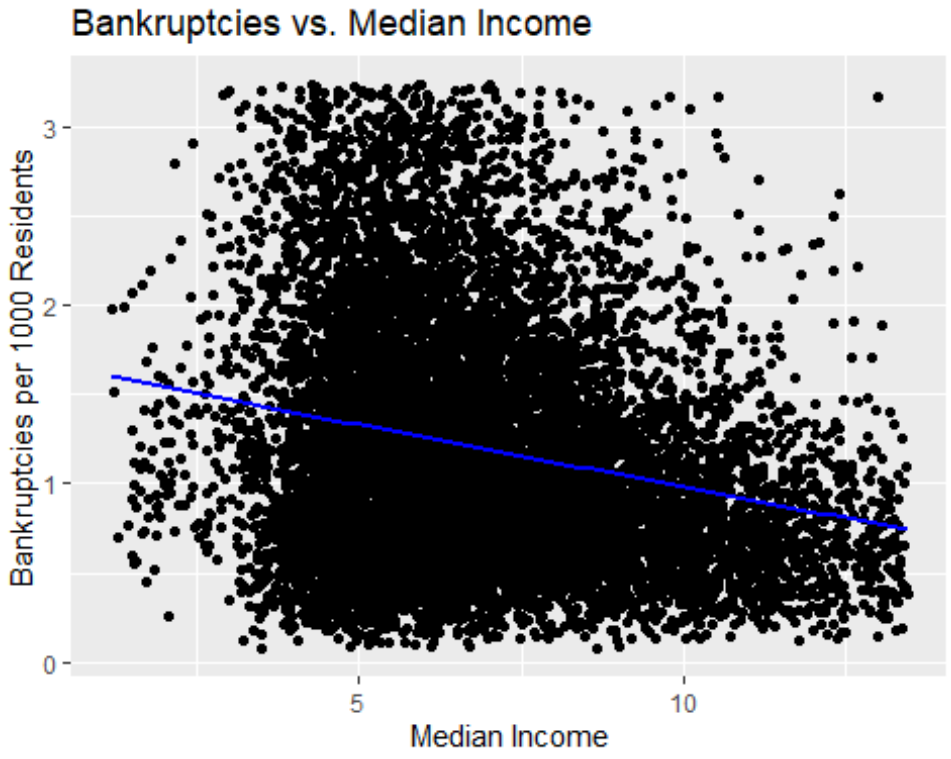


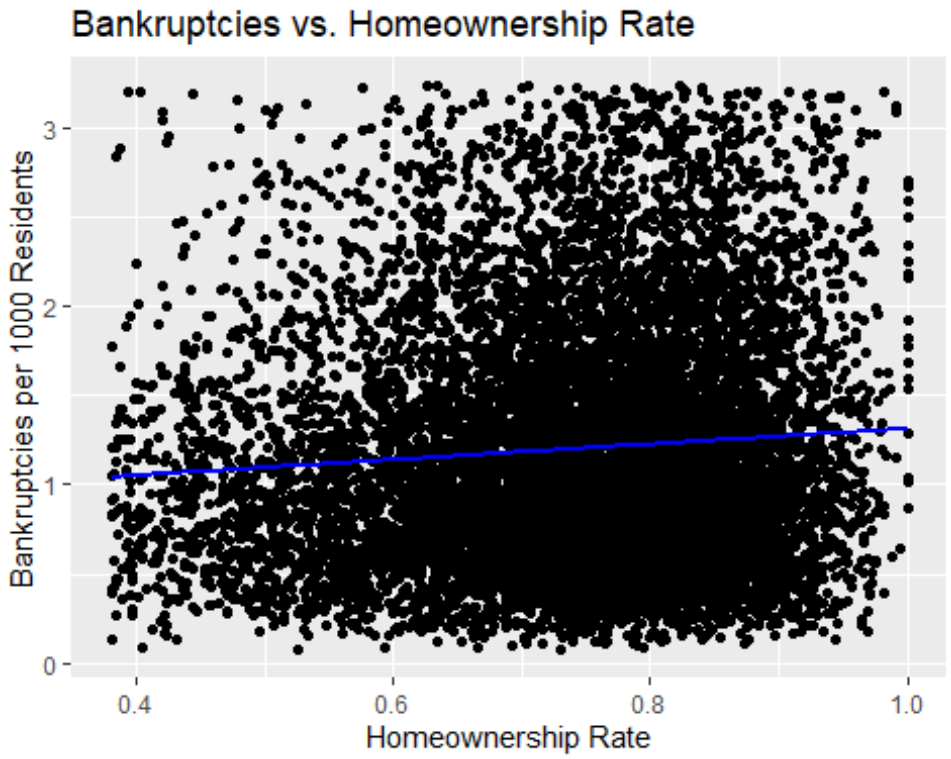
These histograms showed that while the distribution of total bankruptcies is pretty wide and uneven, both of the per-thousand measures roughly follow a normal distribution.

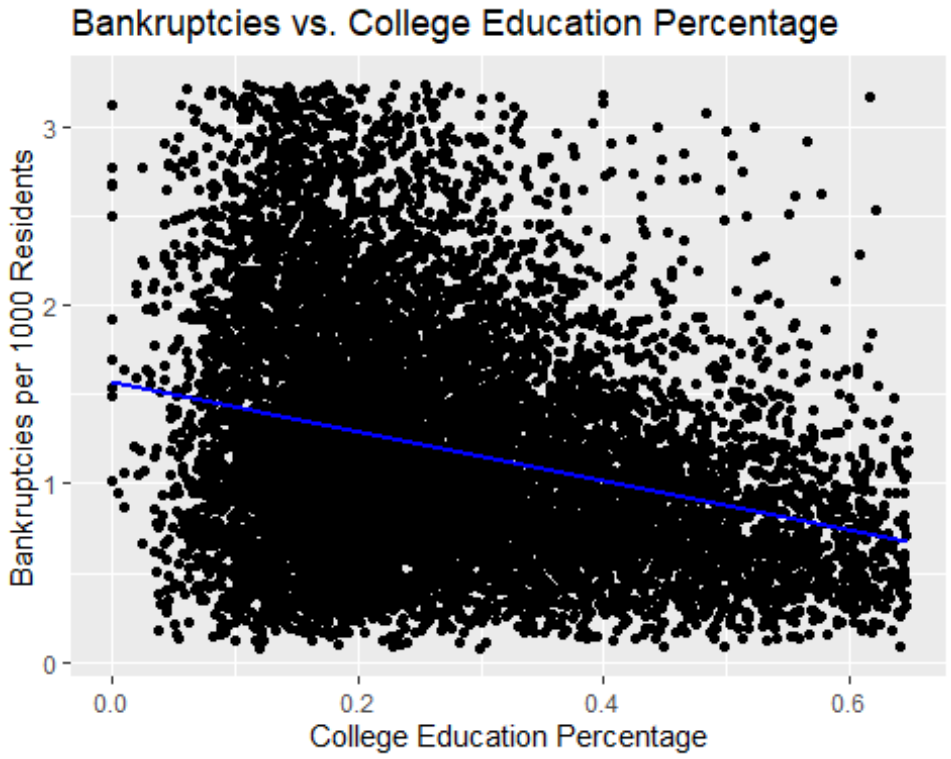
To explore relationships between the Census variables and bankruptcies, I decided to create a few scatter plots. However, these didn't show much besides the presence of outliers. So, I decided to remove outliers and re-plot.

With so many data points, the results are still difficult to interpret but did give me a rough idea of the effects of some of the key Census variables.


Bankruptcies vs. Black Percentage
Bankruptcies per 1000 Residents
0
1
2
3
0.0
0.1
0.2
0.3
Black Percentage


Bankruptcies vs. Median Income
Bankruptcies per 1000 Residents
0
1
2
3
5
10
Median Income





Modeling with linear regression

To determine which demographic characteristics correlate most with elevated bankruptcy rates, I started with linear models.

My first two linear models (fitlm and fitlm2) analyzed how demographic characteristics affected total bankruptcies per 1,000 residents in a ZIP. My first linear model (fitlm) included all six demographic variables I studied (listed above) and found race, college education and the homeownership rate to be the strongest predictors, with race having by far the largest effect. While I expected race to have a significant impact, I was surprised by this result.

Because there is a significant correlation between many of these demographic characteristics, I decided to run a second total-bankruptcies-per-person model (fitlm2) that included fewer variables — race, college education and homeownership. It showed college education to be almost as impactful as race — a result that likely didn't show up in the first model because of college education's strong correlation with unemployment and median income.

My second pair of linear models (fitlm3 and fitlm4) analyzed how the demographic characteristics affected Chapter 13 bankruptcies per 1,000 residents. Fitlm3 included all six demographic characteristics and found race, homeownership rate, labor force participation rate and college education to be the most impactful. Race's importance here was less surprising, given existing research on African Americans' increased likelihood of filing for Chapter 13 bankruptcy.

In fitlm4, I included each of the four variables that fitlm3 found to be most important and left out the unemployment rate and median income. Because of the labor force participation rate's strong correlation with those two variables, it appeared far more important in fitlm4 than in any prior model. In fact, surprisingly to me, its estimated impact surpassed the race variable. This showed me that the economic demographics — in the aggregate — are still quite important for predicting bankruptcy rates even if the individual variables don't appear as important.

My third pair of linear models (fitlm5 and fitlm6) analyzed the demographic variables' effects on the percentage of bankruptcies that are Chapter 13. Fitlm5 showed race and homeownership rate to be the two most important variables, with homeownership rate the most important. This gives credence to the common expression that Chapter 13 is "the homeowner chapter." It's also interesting that the race variable (percent of Black residents) and homeownership both positively increase the percent of bankruptcies that are Chapter 13 while being negatively correlated with each other. Additionally, I was surprised to see that increases in income have a small negative effect on the percentage of bankruptcies that are Chapter 13, given the fact that Chapter 13 was designed for the middle class. When stripping away some of the less powerful economic variables (fitlm6), labor force participation rate once again showed a larger effect.

Modeling with Random Forrest

Next, I applied the Random Forrest algorithm to predict the same outcomes as in the linear regression models. The Random Forrest's variable importance measurements did not highlight the same variables that jumped out in our linear models. Also, while the linear models consistently saw large effects caused by similar variables, the Random Forrest importance rankings were significantly different from each other, depending on whether

total bankruptcies (fitrf) , Chapter 13 bankruptcies (fitrf13) or percentage of Chapter 13 bankruptcies (fitrf13perc) were being predicted.

For the number of total bankruptcies and Chapter 13 bankruptcies, it listed the bachelor's degree and race variables as most important and saw little importance for the homeownership rate. But for the percentage of bankruptcies that are Chapter 13, it listed homeownership rate and race as most important and median income and labor force participation rate as least important.

These results are difficult to interpret. In aggregate, they seem to say that labor force participation rate is a bit more important than the linear models seemed to indicate and that race is a bit less important. They also reaffirm how poor of a predictor median income is. Going into this analysis, I had expected it to be one of the most important variables.

Results

Linear models performance

Our linear models exhibited relatively low Root Mean Square Errors (RMSE) across various predictions. RMSE is a commonly used metric for assessing the accuracy of predictive models. It measures the average difference between the predicted values and the actual observed values. A lower RMSE indicates a better fit of the model to the data.

The models had their lowest RMSEs (around .25) when predicting the percentage of bankruptcies in a given ZIP that were Chapter 13 (fitlm5 and fitlm6). It struggled the most in predicting the total number of bankruptcies in a ZIP (fitlm and fitlm2). Fitlm had an RMSE of 1.189, and fitlm2 had an RMSE of 1.246. When predicting the number of Chapter 13 bankruptcies, the RMSEs were .859 (fitlm3) and .882 (fitlm4)

These RMSE values show that bankruptcy prevalence and, especially, type can be accurately predicted based on readily available demographic data. This proves the systemic nature of bankruptcy in America.

Random Forrest performance

(see lines 145-158 in code)

All three of our Random Forrest models performed slightly better — in terms of RMSE — than their linear model counterparts. The Random Forrest analyzing total bankruptcy filings (fitrf) had an RMSE of 1.149, the one analyzing Chapter 13 filings (fitrf13) had an RMSE of .839, and the one analyzing the percentage of filings that were Chapter 13 (fitrf13perc) had an RMSE of .233. However, since the differences were small, we shouldn't necessarily consider the Random Forrest results more important than the linear model results.

Again, these low RMSEs show the relative ease with which bankrutpcy rates and types can be predicted given demographic data.

Final holdout performance

The goal we stated originally was finding out why some parts of the country have more bankruptcies than others. And the random forest models performed better than the linear models. Therefore, we considered "fitrf" our final model and applied it to the final holdout set.

It turns out the model performed even better on the holdout set than the test set, with an RMSE of 1.05.

Aside #1: KNN model and performance

(see lines 159-165 in code)

I also attempted to use K-nearest neighbors (KNN) for prediction and analysis, but the results showed higher root mean square errors (RMSE) compared to linear regression and Random Forrest. Also, I realized that for the purposes of this analysis, KNN is not the best type of model, since it makes it hard to interpret the importance of certain variables. Consequently, I decided to focus on linear regression and Random Forrest.

Aside #2: Using the bankruptcy data itself to predict Chapter type

As I've continued to research the causes of someone filing a Chapter 13 instead of Chapter 7, I decided to analyze this data in one more way. The bankruptcy data itself contains far more information than Chapter and ZIP. So, I wanted to analyze which variables within that dataset can help us accurately predict Chapter choice.

To do this, I pared the data down into include 9 variables that seemed like they could affect Chapter choice: Total Assets (TOTASSTS), Real Property (REALPROP), Personal Property (PERSPROP), Total Liabilities (TOTLBLTS), Secured Claims (SECURED), Unsecured Priority Claims (UNSECPR), Unsecured Nonpriority Claims UNSECNPR, Total Debt (TOTDBT), and Current Monthly Income (CNTMNTHI). I then made Chapter choice a binary variable.

I then chopped the dataset up into four parts: A testing set, a training set and a small training set. I created the small training set because the data has over 388,000 observations.

Using the small training set, I ran a linear regression to find the most important variables. These were, in order, secured debt, real property, and current monthly income.

Then, using just these variables, I ran a linear regression on the larger training set. With fewer variables, current monthly income rose to the top as the most predictive variable. The RMSE turned out quite small (.484), which gives me confidence that these three factors play large roles in the Chapter someone chooses.

Conclusions

This project primarily accomplished three tasks:

1. It demonstrated the ability of multiple machine learning techniques to predict bankruptcy rates and Chapter types using Census data. These techniques achieved low RMSEs.This proved how closely bankruptcy is intertwined with economic conditions.

2. It proved that the percentage of Black residents in a given ZIP plays a major role in the number and type of bankruptcies there. In fact, it seems to play a slightly larger role than any of the economic or educational characteristics we evaluated. This further proves the previously documented role of racial bias in the bankruptcy system. This finding informed a news article I wrote, which can be found here: https://mlk50.com/2023/09/05/shelby-county-ranks-no-1-in-bankruptcy-heres-why/. In that article, I wrote, "The percentage of Black residents in a given U.S. ZIP code is a strong predictor of the number of Chapter 13 bankruptcy filers there, even when accounting for other demographic variables. When MLK50 analyzed the effects of race, college education, homeownership, unemployment, labor force participation and median income on Chapter 13 rates across the country, the percentage of Black residents and the labor force participation rate had the largest effects."

3. It also showed that the homeownership rate, college education, and labor force participation rate are better predictors of bankruptcy than median income or the unemployment rate.

These results have potential implications for policymakers looking to address systemic racism and researchers interested in understanding why some parts of the country have much higher bankruptcy rates than others.