Steven Cloud, Yasmin Givens, Dileka Gunawardana, Sebe DiPrampero & Matthew Zamarripa
November 30, 2021

STAT 405 - Final Project Report

I. About Datasets

In conducting our final project analysis, we decided to focus on business data, particularly the nuances and relationships that exist between the characteristics of a business/ its owner and the successes/ hardships that must be dealt with. We will begin with a brief introduction to the two datasets that guided our analysis followed by the major components of our methodology and takeaways of our analysis.

A. Dataset I: US Census Survey of Business Owners (SBO)

The primary dataset used for our group project was called the "2007 Survey of Business Owners (SBO) Public Use Microdata Sample (PUMS)" and can be accessed by the following link. It contains 199 columns and 2,165,680 rows. Each row corresponds to a registered non-farm business located in the United States. Each column describes various demographic characteristics regarding the owners of each business such as their race, ethnicity, veteran status, age, and education level attained in addition to the business itself such as the employment size. There are also columns focusing on the business owners' relationship with their business (i.e the number of hours spent working on the business and whether or not the business is their primary income source). For the purposes of consistency, we will focus our analysis on the first-listed owner (corresponding to whoever owns the largest share of the business). For the sake of space, we have included a limited number of columns of the dataset below:

Because the survey allowed businesses to report up to four owners, there are demographic characteristics listed for each of these theoretical four owners. Because some businesses don't contain four owners, we decided to only use data from the first owner (also referred to as "primary" owner). This allowed consistency and ease of comparisons throughout the analysis.

B. Dataset II: Small Business Administration Disaster Loan Data

The secondary dataset used was from a government source - the Small Business Administration (SBA). Essentially, this is a record of all loans administered to small businesses in the United States and is accompanied by parameters related to the disaster that occured, the size of the loan, the location/ characteristics of the business. In order to ensure compatibility between the two datasets, we only included data based from 2007.

County	State	Total Verified Loss	Verified Loss Real Estate	Verified Loss Content	Total Approved Loan Amount	Approved Amount Real Estate	Approved Amount Content	Approved Amount EIDL
LAKE	IN	0	0	0	20,000.00	0	0	20,000.00
LAKE	IN	230,505.00	221,705.00	8,800.00	90,600.00	82,500.00	5,400.00	2,700.00
VANDERBURGH	IN	24,242.00	22,742.00	1,500.00	24,300.00	22,800.00	1,500.00	0
VANDERBURGH	IN	99,021.00	78,736.00	20,285.00	0	0	0	0
VANDERBURGH	IN	10,757.00	10,757.00	0	10,000.00	10,000.00	0	0
LAKE	IN	46,169.00	40,869.00	5,300.00	45,500.00	29,900.00	5,300.00	600
LAKE	IN	27,293.00	27,293.00	0	12,000.00	12,000.00	0	0
LAKE	IN	26,100.00	26,100.00	0	0	0	0	0
LAKE	IN	26,829.00	23,279.00	3,550.00	0	0	0	0
LAKE	IN	478,776.00	361,554.00	117,222.00	283,400.00	176,900.00	84,300.00	22,200.00
FLOYD	IN	31,411.00	21,546.00	9,865.00	25,200.00	21,600.00	0	3,600.00
RICHMOND	VA	14,975.00	14,975.00	0	16,700.00	15,000.00	0	1,700.00

II. Data Cleaning & Preparation

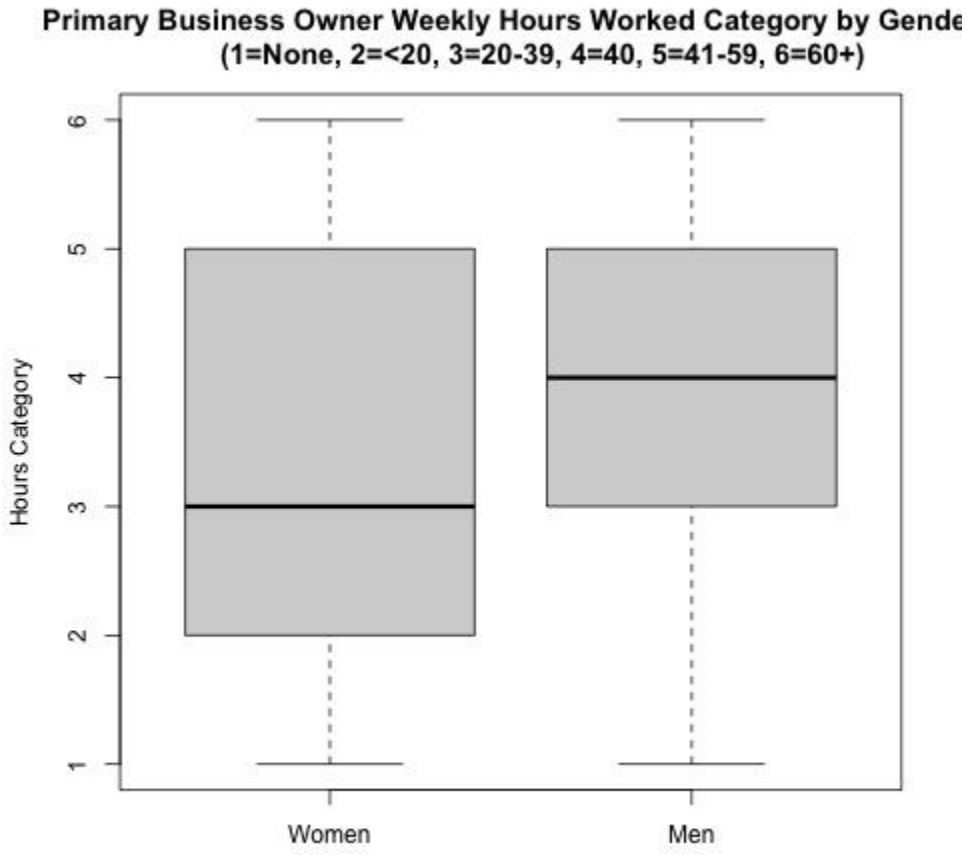
A. SQL

B. StringR

III. Plots

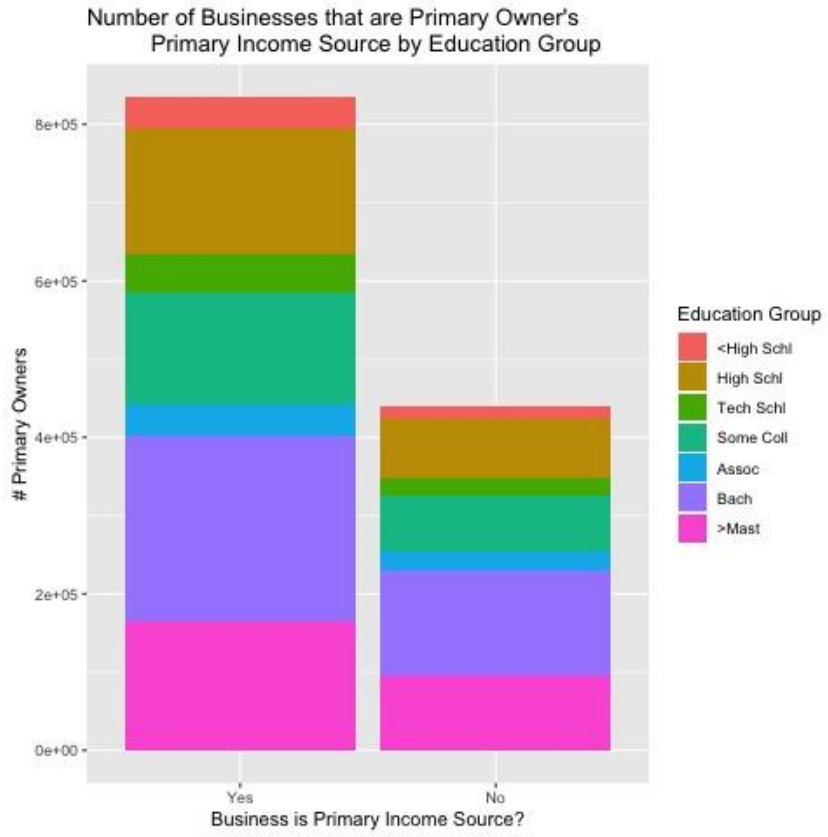
A. Business Loan Analysis

B. Analysis of Relationships Between Gender & Businesses

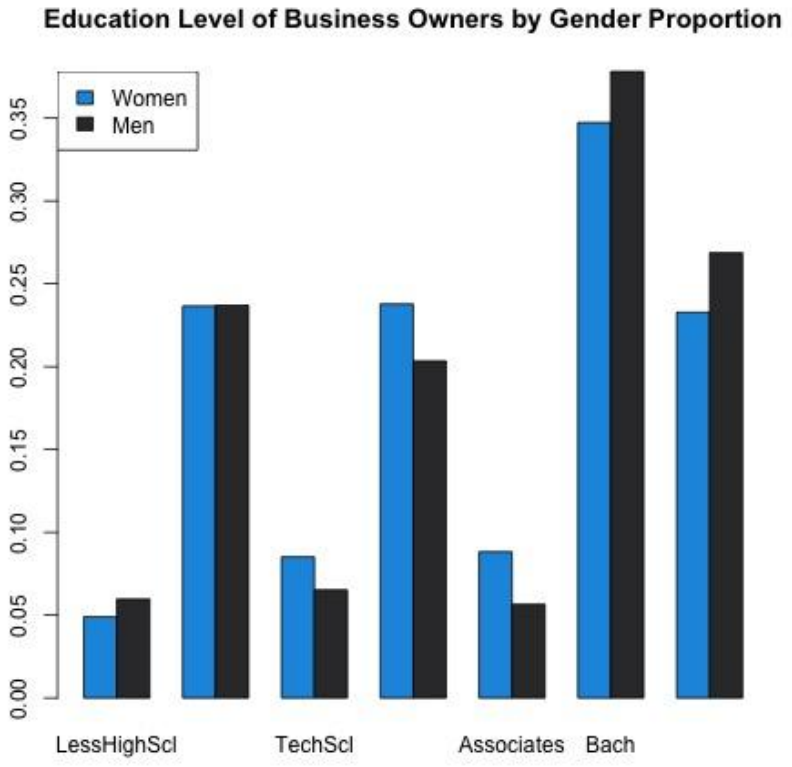


The boxplots show the distribution of the number of hours per week the first owner works, these distributions are then separated based on the first owner's gender. Based on the boxplots, the male business owners tend to work more hours per week on their business, 40 hours per week, while the women business owners spend an average of 20 to 39 hours per week.

C. Analysis of Relationships Between Education Level & Businesses



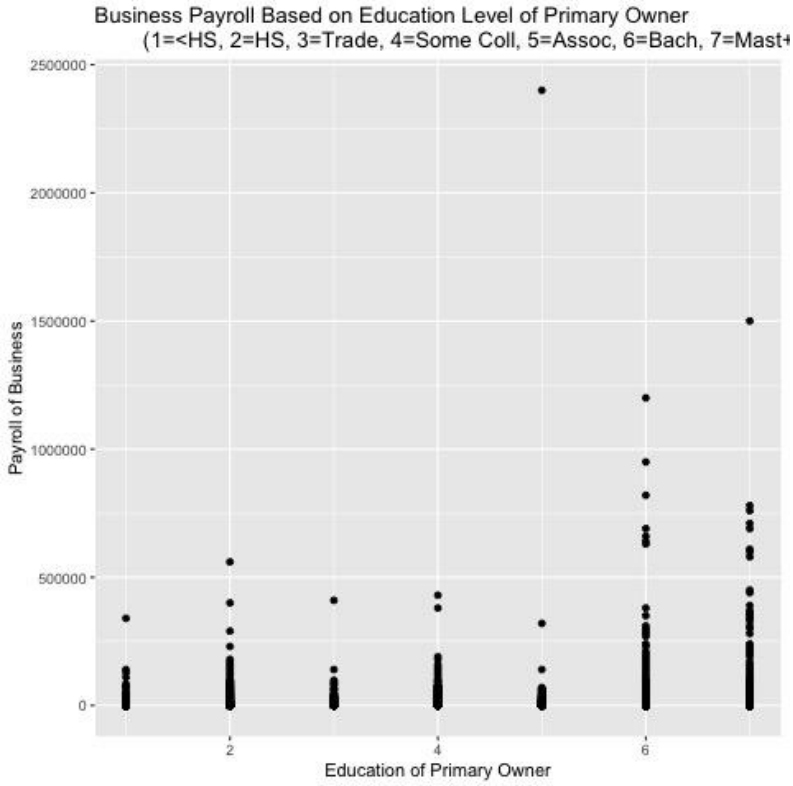
This plot suggests that whether or not a business is someone's primary income source doesn't particularly depend on their education level as the two groups appear to have the same general distribution in terms of education group.



This graph shows the educational level of business owners by gender from the PUMS 2007 dataset. The proportions shown in this graph reflect the relationship between business owners in a gender from a specific educational level to all business owners of that gender. So for example, approximately 5% of female business owners have less than high school education while approximately 6% of men have the same education level. From this graph, we can see that at lower levels of education, male and female business owners are at similar levels. At the higher education level, female business owners have higher educational rates in tech schools and associates degrees. However, male business owners have higher rates of completion for higher-level education degrees specifically for bachelor and greater than master degrees. And to note, this graph does not account for the ~200,000 female and ~430,000 male business owners who did not fill out this portion of the survey.

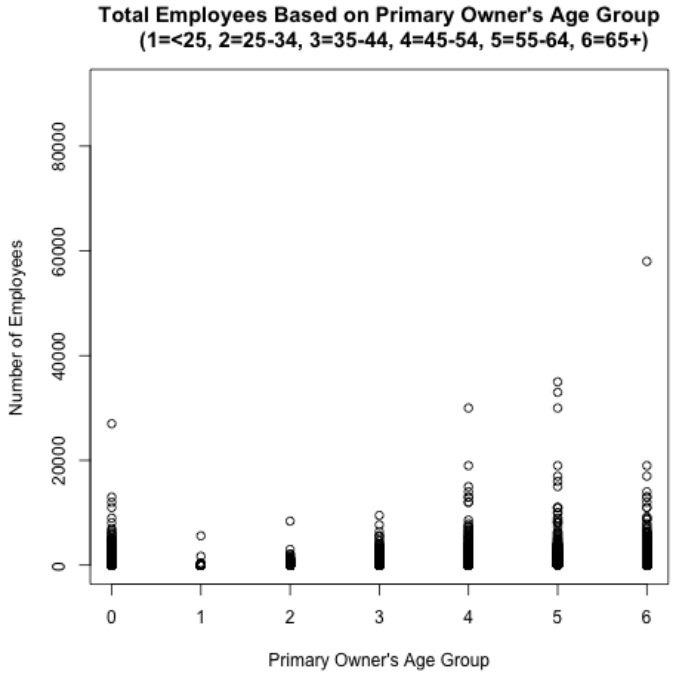
D. Analysis of Relationships Between Sector & Businesses

E. Analysis of Relationships Between Owner's Education Level & Business Payroll



This plot presents a visualization of the education of owner 1, with ratings from less than high school to masters+, on the x-axis against the payroll of the business on the y-axis

F. Analysis of Relationships Between Owner's Age & Business Number of Employees



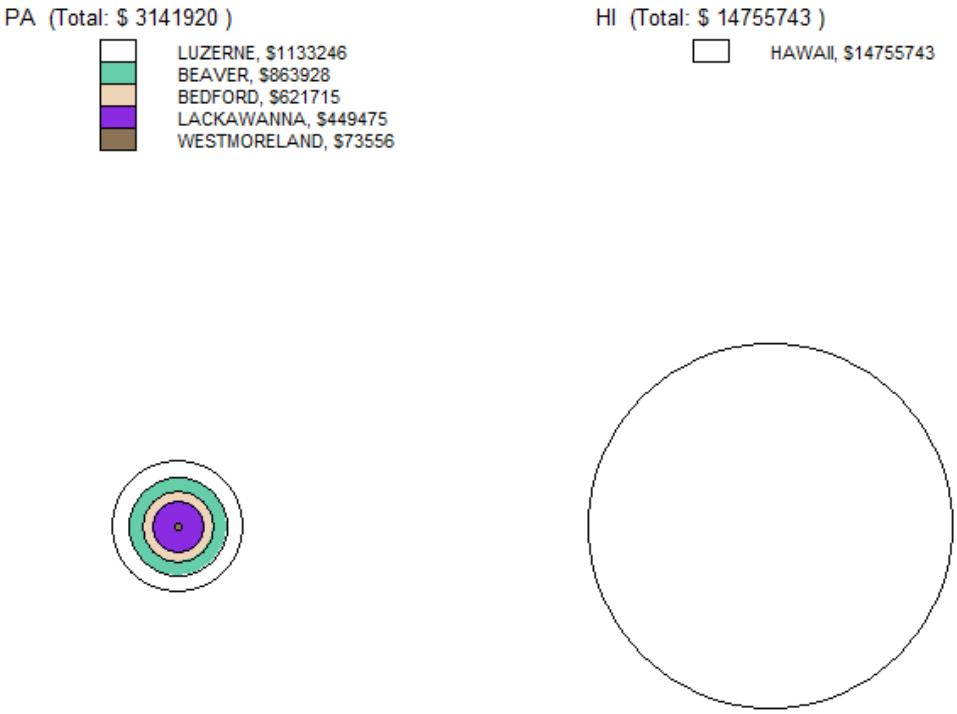
The results of this plot makes sense, since a younger owner results in less employees generally. With older owners, one can find more businesses that have a larger number of employees due to the fact that they had more time to develop their business.

IV. R Shiny/ Grid Killer Plot

Our killer plot takes in two states as inputs and creates a visualization using our secondary dataset, Small Business Administration Disaster Loan Data. The graph essentially takes every county from the inputted state that had small businesses impacted by a natural disaster and it displays the distribution of costs for each county in the form of a target. Each ring of the target represents one county, which is color coded and described in the legend above each state's target. Additionally, in the legend next to the name of the county, one can find the actual amount of money that small businesses collectively had to spend for that respective county.

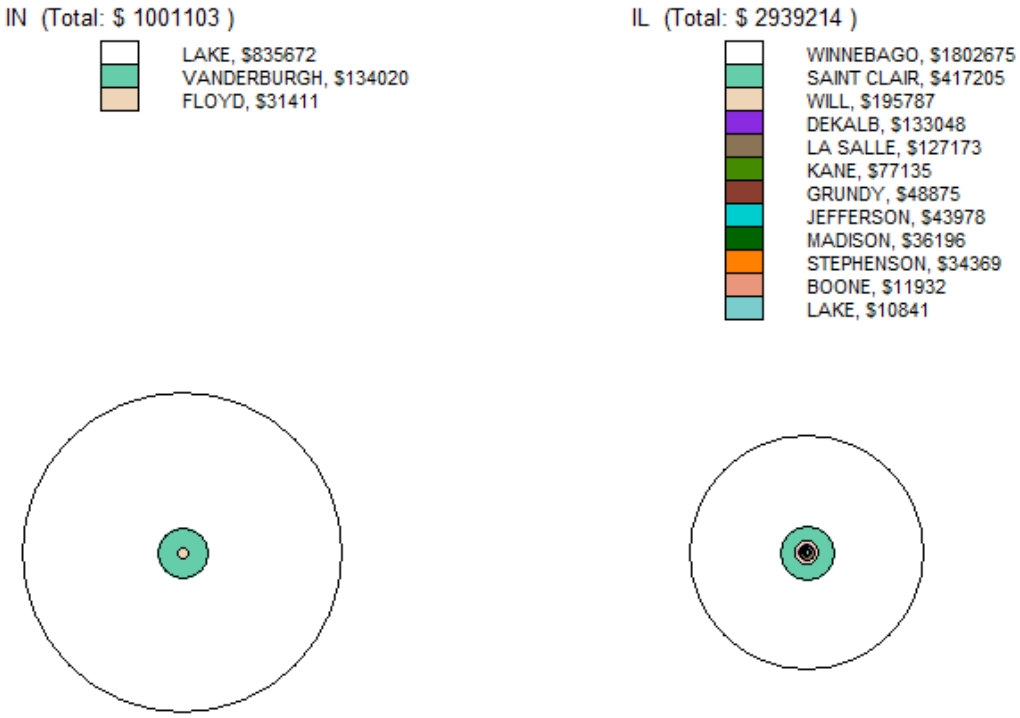
There are many interesting takeaways from our killer plot, especially when juxtaposing two states next to each other. Our killer plot allows a viewer to be able to see every county that had small businesses hit by natural disasters, each county's total expenses from the natural disasters, the total amount of money that small businesses had to pay for natural disasters in the whole state, and how these costs were distributed throughout the respective counties. Based on these observations, when comparing two states side by side, one can see the difference in number of counties hit in each state, the difference in total expenses for each state, and the difference in distribution of expenses across the counties. Using shiny, one can quickly and efficiently switch between any state in the U.S. but for the purposes of this final report, we have added a few examples of our killer graph below, paired with some brief analysis.

A. Pennsylvania vs Hawaii



Some clear initial observations that one can see from the figure above include the fact that five counties had small businesses impacted by natural disasters in Pennsylvania, and only one county in Hawaii had small businesses impacted by natural disasters. These numbers are logically consistent, as Hawaii only has four counties and Pennsylvania was not affected by a large number of natural disasters during the year we studied. After more careful inspection, one can see that, based on the size of each ring in Pennsylvania's graph, the total expenses for the natural disasters was pretty evenly distributed across the counties. It is obviously useless to look at the distribution of cost for Hawaii, since only one county was affected, however the large size of the entire target for Hawaii compared to the smaller target for Pennsylvania indicates that Hawaii's one impacted county had to pay a disproportionately large amount compared to the counties of Pennsylvania.

B. Illinois vs. Indiana



The aim of this second example was to take two bordering states and see if the data collected on the impact of natural disasters on small businesses was similar due to their geographic proximity to each other. At first glance, the two targets look very alike - each has one large outer ring with some inner rings that seem to be mostly evenly distributed in terms of cost per county. Based on the Illinois graph, one can see that more counties were affected by natural disasters, but upon further research the majority of these counties lie in the section of Illinois that is furthest from Indiana. When looking at the largest ring of each graph, one can discover that Winnebago county and Lake county, the two white labeled counties, both lie in the northern part of their state, and the fact that they are relatively close to each other could explain their comparatively heavy burden of cost in regards to the rest of their state's counties.

V. Conclusions

Based on our research and exploration of our two datasets, we have found three major conlcusions:

- In our data, there is little difference between male and female business owners in terms of hours worked, education level, and the business being a primary income source
- There seemed to be little correlation between natural disasters striking a state and that state's reported payroll
- The distribution of loans by the government seemed to favor states with more counties and larger populations