# Project Proposal

due October 11, 2021 by 11:59 PM

Isabelle Xiong and Maya Ghanem

10/10/2021

## Load Data

```
Carolina_Small_Business_Dataset <- read_excel("~/Carolina Small Business: Stats 198 Project/data/Caroli
```

```
north_carolina_map <- map_data('county', 'north carolina')
```

## Introduction and Data, including Research Questions

### Motivation for research:

Recently, we've seen more and more posts of small businesses on social media platforms. Walking on the streets of Durham, I've noticed that the area around ninth street is filled with businesses run by BiPOC and marginalized folks. According to data from the NC Secretary of State's Office, the number of small businesses in North Carolina boomed during 2021, increasing by 50% compared to 2020. We wanted to understand more about the socioeconomic and locational distributions of small businesses in North Carolina, and the support they are receiving to stay open during COVID, where there's even more competition due to the increasing amount of new small businesses. For this project, we will be studying a dataset from the Carolina Small Business Development Fund (see references).

### Research Questions:

1) Is more money lent to minority and women owned businesses? Is the difference in lending money significant based on whether a business is minority and/or women owned?
2) Is there a difference in social, community trust, and financial stability scores based on whether the business is minority and/or women owned?
3) What do employment outcomes look like for minority owned and women owned businesses versus businesses that are not minority or women owned?

## Glimpse

### Carolina Small Business Dataset:

```
glimpse(Carolina_Small_Business_Dataset)
```

```
## Rows: 175
## Columns: 15
## $ `Program Name`              <chr> "Durham Small Business Recovery Loan", "Me~
## $ `Application No.`           <chr> "03492", "00290", "01130", "01087", "03638~
```

```
## $ Amount                       <dbl> 35000, 30746, 25000, 35000, 25011, 7668, 5~
## $ `NAICS 6-Digit Code`         <chr> "312120", "423450", "541840", "541910", "5~
## $ `Social Capital Score`       <dbl> 2.0, 1.4, 1.4, 1.6, -0.4, 2.0, -0.2, 0.4, ~
## $ `Community Trust Score`      <dbl> 2.9, 2.6, 3.1, 3.0, 3.0, 3.7, 2.5, 2.0, 2.~
## $ `Financial Stability Score` <dbl> 0.5, 0.5, 0.5, -1.0, 1.0, 0.5, 1.3, -0.5, ~
## $ `Minority-Owned Firm`        <chr> "True", "True", "False", "False", "False",~
## $ `Women-Owned Firm`           <chr> "True", "False", "False", "True", "False",~
## $ `Veteran-Owned Firm`         <chr> "False", "False", "False", "False", "False~
## $ County                       <chr> "Durham", "Mecklenburg", "Mecklenburg", "M~
## $ `Shipping Zip/Postal Code`   <chr> "27713", "28211", "28203", "28202", "28217~
## $ `Created FTE`                <dbl> 4.0, 0.3, 9.1, 0.0, 1.1, 0.0, 2.6, 1.1, 1.~
## $ `Retained FTE`               <dbl> 0.0, 1.7, 17.1, 1.1, 1.4, 1.0, 1.0, 2.4, 2~
## $ `Current FTE`                <dbl> 8.0, 1.7, 17.1, 10.3, 1.4, 1.0, 1.0, 2.4, ~
```

### Map of North Carolina Counties

```
glimpse(north_carolina_map)
```

```
## Rows: 3,670
## Columns: 6
## $ long      <dbl> -79.53800, -79.54372, -79.54372, -79.53800, -79.52081, -79.2~
## $ lat       <dbl> 35.84424, 35.89008, 35.89008, 35.98175, 36.23385, 36.23385, ~
## $ group     <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, ~
## $ order     <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 1~
## $ region    <chr> "north carolina", "north carolina", "north carolina", "north~
## $ subregion <chr> "alamance", "alamance", "alamance", "alamance", "alamance", ~
```

## Data Analysis Plan

For the three research questions, the explanatory and response variables are as follows:

1) Explanatory Variable: Whether business is minority and/or women owned. Response Variable: Money lent to the business

2) Explanatory Variable: Whether business is minority and/or women owned. Response variable: Social, Community Trust, and Financial Stability Scores

3) Explanatory Variable: Whether business is minority and/or women owned. Response Variable: Employment Outcomes

The data analysis methodology is as follows:

### Step 1: Contextualize and Situate the Data

The dataset will be contextualized with a description of the metadata and the report that published the dataset, including an explanation of the variables. Additionally, visualizations will be made for the following:

a) To preview the overall difference between minority and/or women owned and not minority and/or women owned businesses, a bar plot will be made comparing the total amount of money lent to minority owned business, women owned business, both minority and women owned businesses, and not minority or women owned businesses. Similar geom_bar plots will be made based on the social, community trust, financial stability scores and based on employment outcome response variables.

b) To preview the spread of data and overall difference between minority and/or women owned and not minority and/or women owned businesses, a box plot will be made comparing the total amount of money lent to minority owned business, women owned business, both minority and women owned

businesses, and not minority or women owned businesses. Similar geom_bar plots will be made based on the social, community trust, financial stability scores and based on employment outcome response variables.

c) A histogram with the distribution of the amount of money lent, with a fill variable for women, minority, women+minority, and not women or minority businesses. Similar histograms will be made for the distribution of scores and employment outcomes.

d) A density map of North Carolina counties based on the amount of lending projects in each county.

e) A table, grouped by county, with the fraction of women owned/total businesses in country, minority owned/total businesses in county, women+minority/total businesses in county, and not women or minority/ total businesses in county.

### Step 2: Conduct Statistical Tests

The primary form of testing for each research question will be a paired two-sided hypothesis t test. For each test, the null hypothesis is that there is no difference in the response variable (money lent, scores, or employment outcomes) based on the explanatory variable (women, minority, women+minority, or neither women/minority business), and the alternative hypothesis is that there is a difference in the response variable (money lent, scores, or employment outcomes) based on the explanatory variable (women, minority, women+minority, or neither women/minority business). For the t-tests, this is how we would like to compare the explanatory variables:

a) women vs. not women businesses
b) minority vs not minority business
c) women+minority vs. everything else
d) neither women or minority vs. everything else

### Step 3: Display Results of Statistical Tests

The results of the paired two-sided hypothesis t-tests will be displayed by creating table p-values, confidence intervals, and t-values for each t-test. We cannot create a graph for the distribution of the difference between the explanatory variables compared because the t-test is not comparing two points in time.

### Step 4: Draw Conclusions

If the p value for a given test is above 0.05 and the null hypothesis mean (difference = 0) is within the confidence interval, the null hypothesis cannot be rejected.

If the p value for a given test is below 0.05 and the null hypothesis mean (difference = 0) is not within the confidence interval, the null hypothesis can be rejected. More testing must be done before any conclusion can be made about the alternative hypothesis, but some inferences and predictions could be made based on the context of the data.

## References

https://www.carolinasmallbusiness.org/post/small-business-covid-19-lending-programs-fostering-social-capital-and-financial-stability