Brittany Scott

Midterm Project

[bmscott.95@gmail.com](mailto:bmscott.95@gmail.com)

1. **Economic Question**
   1. *Question*: Does a community’s economic hardship have an effect on the number of school expulsions in that community area?
      1. Do expulsions in an area have any relationship to the number of people that are 25 without a high school diploma?
   2. *Importance*: This question is important because education is a key factor in determining the success of a child. If a child is growing up in an area with a high hardship index, then it may have a direct effect on the child’s school environment, and in turn performance. If a child cannot successfully complete a k-12 education, they will only perpetuate the current cycle of high hardship.
   3. *Hypothesis*: The community areas with the highest hardship index will also have the highest expulsion rates
2. **Raw Data Set Description**
3. The Census Data provided by the City of Chicago is a selection of six socioeconomic indicators. The data was collected using census tract-level estimates acquired form the Census Bureau 2008-2012 American Community Survey 5-year estimates. The data covers the years of 2008-2012 and is limited to the Chicago community area. (<https://catalog.data.gov/dataset/census-data-selected-socioeconomic-indicators-in-chicago-2008-2012-36e55)>
4. The school data was collected and provided by Chicago Public Schools from the 2015-2016 school year. (<http://cps.edu/SchoolData/Pages/SchoolData.aspx>)
   1. The Census data set is a time series data set
      1. Unit of Observation: the people in each community area
   2. The Misconduct data is a cross sectional data set
      1. Unit of Observation: Students in a certain school
5. The variables that I used for the purpose of merging were community area name, and school id. There are a number of observations in my chosen data sets. I chose to use the number of expulsions per 100 students, the hardship index, and the number of people 25 without a diploma to test my hypothesis.
6. The key economic variables include the dependent variables, hardship index and community area depending on which hypothesis I was testing.
7. As I was cleaning up the data, I had to perform a number of tasks such as the following:
   1. Renaming variables
   2. Calculating aggregate data
   3. Merging datasets
   4. Changing the vector type
   5. Identifying and Correcting Misspellings in Data sets
8. **\*\*Refer to PowerPoint for graphs and summary statistics\*\***
9. **A Discussion of Next Steps**
   1. How do you intend to identify causal relationships?
      1. I will identify relationship be aggregating the data and analyzing based on community areas
      2. Utilizing bar graphs, and scatter plots to identify relationships
   2. Any additional data work that you think you will need to do
      1. Moving forward, I plan on looking into school data from 25+ years ago to see if there is any relationship between the number of people age 25 and unemployed. I will also investigate school attendance rates against the hardship index, and identify the top performing school to determine whether they have high expulsion rates or not.