Adithyan Subramanian
Data Wrangling report
Mentors: Danny Wells and Mukesh Mithrakumar

The cleaning for this project was largely challenging. My project to analyze the effects of business locations on the health outcomes of Chicago neighborhoods utliized data that was incomplete or unclean. In this report I will discuss the various methods I used to make bring the data to a usable state for this project.

The project will utilize three data sets. All of these datasets are easily available on the City of Chicago's data portal.

The easiest dataset to clean was the Public Health Statistics-Selected public health indicators by Chicago community area data set (Health). The health dataset consisted of public health indicators such as unemployment, assault rate, and the rates of various diseases such as gonorrhea or tuberculosis for each neighborhood. The "Gonnorhea in Females" column had 12 nan values, whereas the "Childhood Blood level screening" and "Childhood lead poisoning" columns both had 1 nan value each. In order to rectify this, I chose to substitute the nan values with the mean of the column. Although clearly not an accurate measure of the indicator, this will place a reasonable estimate of the indicator within these neighborhoods.

The Public Health Statistics-Selected underlying causes of death in Chicago 2006-2010 data set (Deaths) documented the prevalence of several causes of death by neighborhood. The Death dataset was rather easy to clean as well. There were 18 nan values in the Cumulative Deaths Rank, Crude Rate Rank, Adjusted Rate Rank, YPLL Rate RANK columns. Upon further investigation, it was apparent that the nan values were part of rows documenting the cumulative Chicago city data for the pertinent factors. As the entire city cannot be ranked among the communities, there could be no rank given. It was decided to drop those rows with a simple pd.dropna function.

Finally the Business_Licenses (Businesses) data set was by far the hardest set to clean and prepare. The dataset consisted of all the businesses licenses awarded by the City over the last 30 (?) years. There were two major issues. The first was that although this data set was large, there

were many nan values, of which none were easily replaceable by numerical values. For this reason we had to drop a large size of the dataset, and the final businesses dataset only has some 13,000 odd rows. Secondly, while each business had a latitude and longitude, it did not specify which neighborhood each business was located. To solve this issue I had to use the longitude and latitude to locate each business in each neighborhood. I was able to find a json file with a list of polygons for each neighborhood. I read this file into the Jupyter notebook, changed the list from a list of lists to a list of tuples, and wrote a to take a latitude longitude set to determine which neighborhood that point was in. I then used another for loop to loop through the dataset assigning neighborhoods to the businesses.

This was a long process, but I hope it will help bring insight into the various health factors within the various neighborhoods of Chicago.