Adithyan Subramanian
Capstone Project 1 Milestone report
Mentors: Mukesh Mithrakumar

Health disparities in Chicago based on Socio-economic and Business Factors

Problem Statement

Chicago is an extremely diverse city with people of all walks of life living, working, and playing within the city. Despite Chicago's diversity, it is still considered one of the most segregated cities in America today[1]. This segregation has caused many health disparities within Chicago[2]. In this project, we aim to further understand which neighborhoods are the healthiest. We will focus on cancer rates, lung cancer rates and infant mortality rates. Finally, we will explore how the number and types of businesses in a neighborhood affects the healthiness of the citizens in that neighborhood. The end goal is to predict lung cancer and infant mortality rates for neighborhoods in Chicago.

Data Set

All data comes from the Chicago Data Portal, Chicago's hub for all public data sets. The first dataset is the Selected Public Health Indicators by Chicago Community Area (https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-public-health-in/iqnk-2tcu). This dataset documents a variety of different public health metrics including teen birth rate, infant mortality, crowded housing, and unemployment. The second dataset is the Business Licenses dataset (https://data.cityofchicago.org/Community-Economic-Development/Business-Licenses/r5kz-chrr). This dataset has all business licenses awarded within the city starting from 2002. It also documents the location of the businesses, as well as the nature of the license. The final dataset is the Selected underlying causes of death in Chicago, 2006 – 2010 data set (https://data.cityofchicago.org/Health-Human-Services/Public-Health-Statistics-Selected-underlying-cause/j6cj-r444), which documents all deaths under the age of 75 in the city based on the neighborhood the death occurred. All data is available as CSV files.

Data Wrangling

The cleaning for this project was largely challenging. My project to analyze the effects of business locations on the health outcomes of Chicago neighborhoods utilized 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 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 "Gonorrhea 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.

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.

Exploratory Data Analysis

In order to answer the question of if businesses affected the health outcomes, we must first analyze which neighborhoods have significant health disparities. To begin, we first looked at the disparities among the neighborhoods with high cancer rates and low cancer rates.

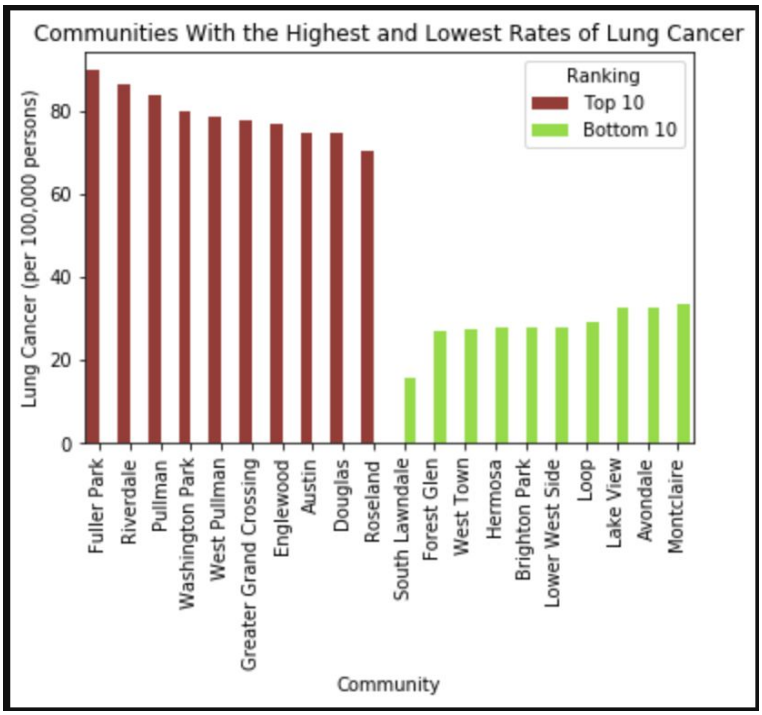

Fig. 1

Figure 1 shows the lung cancer rates of the 10 neighborhoods with the highest lung cancer rates, and the 10 neighborhoods with the lowest lung cancer rates. It is important to notice the vast disparity between the two groups. the neighborhoods with the highest lung cancer rates have rates that are almost twice the rate of the neighborhoods with the lowest lung cancer rates. In fact, Fuller Park, which is the neighborhood with the highest lung cancer rate, has over 4 times the lung cancer rate of South Lawndale, which is the neighborhood with the lowest lung cancer rate in Chicago. This is a serious disparity within the city and should be made a priority.

It is also important to visualize the neighborhoods location in relation to the health metric in question in order to identify any geographic trends that may be apparent. To do this we created interactive choropleth maps using the Bokeh library.

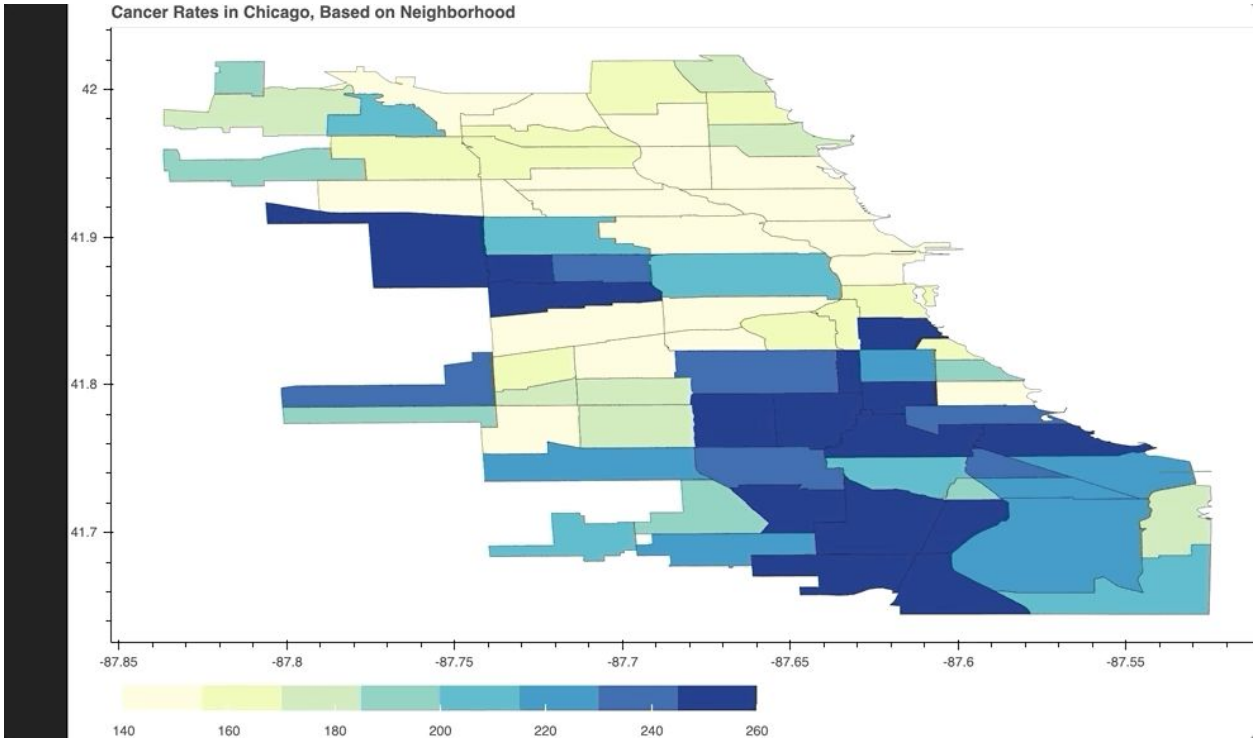

Fig. 2

Fig. 2 is a GIF showing the choropleth created for cancer rates within Chicago. An end user can mouse over separate neighborhoods to visualize the cancer rates within the neighborhoods. Holistically, it is remarkable that the south and far west sides of the city seem to have the highest cancer rates, while the northside has a relatively low rate of cancer.

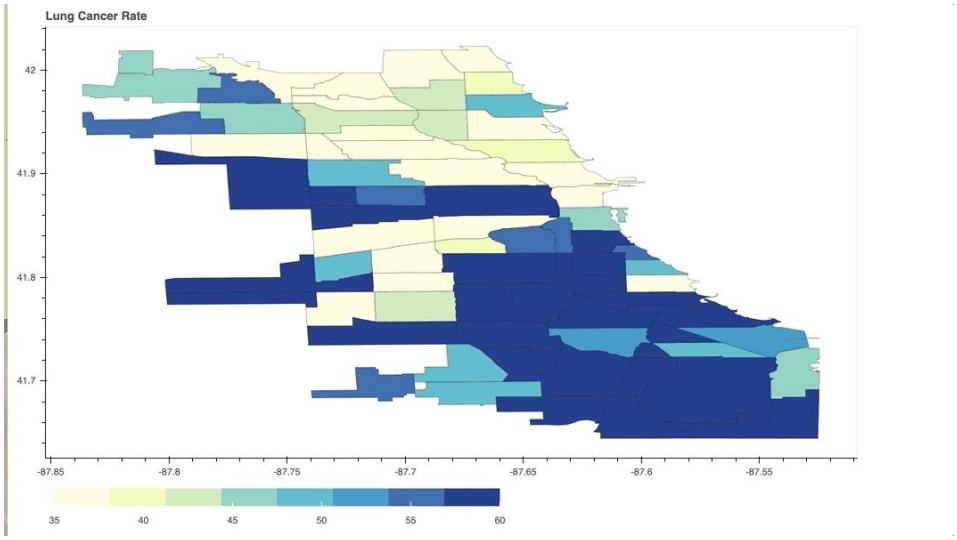

Fig. 3

Fig. 3 shows a GIF of the choropleth created visualizing lung cancer rates within the City of Chicago. Much like the general cancer rates map, this shows that the south and west sides have higher rates of lung cancer compared to the north side. However more neighborhoods seem to have relatively higher comparative lung cancer rates, compared to the general cancer rates. This shows a higher disparity level throughout the city.

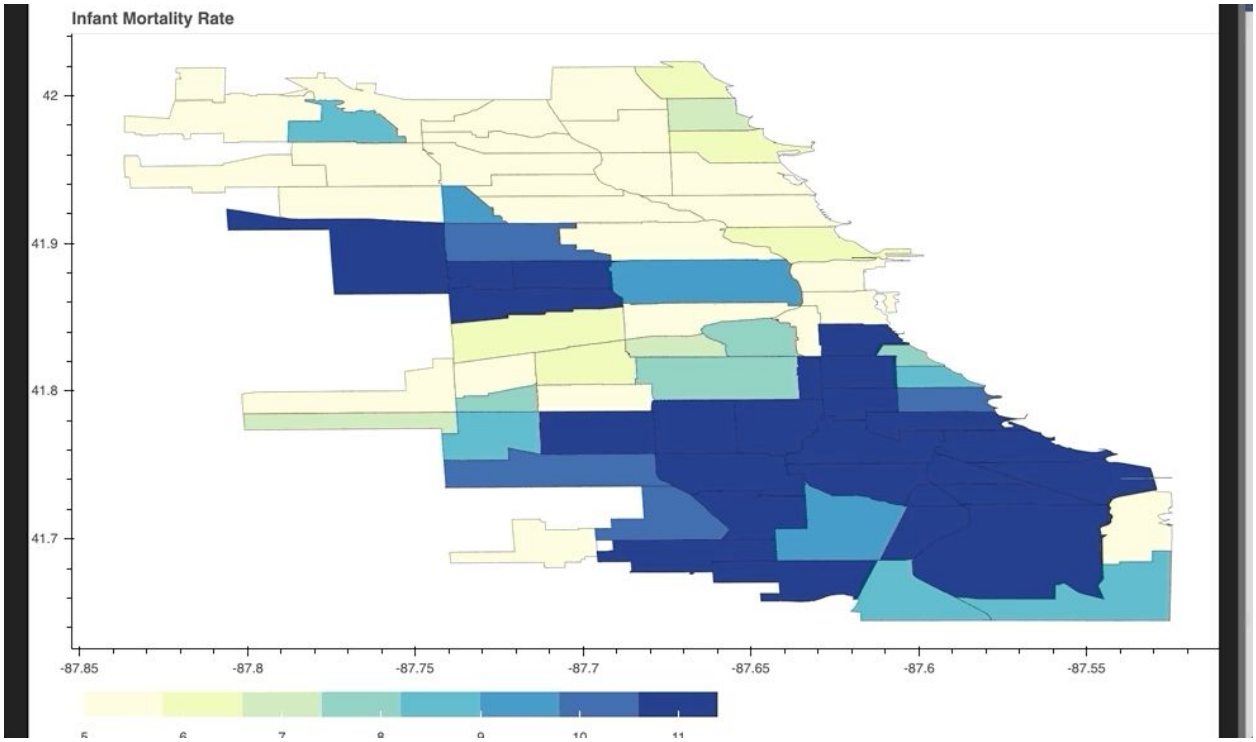

Fig. 4

Fig 4. Is a GIF of the choropleth visualizing infant mortality rates throughout the City of Chicago. Once again, the south and far west sides have elevated levels of infant mortality compared to the north side.

In all three of the choropleths, it is evident that the South and West sides of the city suffer from significant healthcare disparities. It is important to note that a majority of these underprivileged neighborhoods are minority majority neighborhoods. The role of race in healthcare disparities is well documented. However, since the datasets used make no note of race, any further analysis on the factor of race is out of the scope of this project.

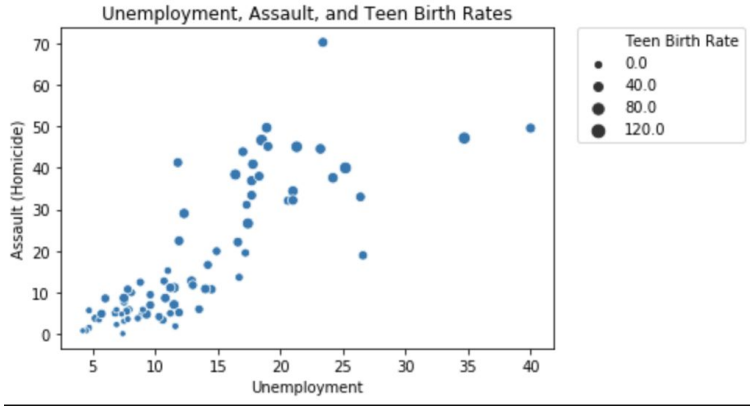

Fig. 5

We also thought it was vital we saw how the health metrics compared to one another. In fig. 2, we can see the relationship between Unemployment rates, assault rates, and teen birth rates. Teen birth rates are visualized by the size of the points on the graph. As seen, there is a noticeable positive correlation between all three of these metrics. It seems that neighborhoods with high unemployment rates also suffer from increased violence and a higher teen birth rate. This while not a revolutionary discovery, does indicate how important socio-economic factors play in the health outcomes of the Citizens of Chicago.

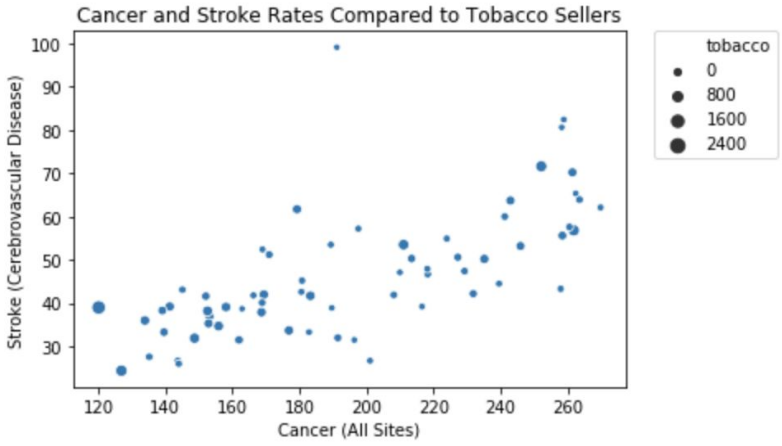

Fig. 5

We are finally ready to visualize how businesses affect health outcomes in the City of Chicago. In Fig. 5 , we see that there seems to be a minor positive correlation between stroke and cancer rates. However, it is interesting to note that two neighborhoods with a relatively high number of tobacco sellers have relatively low stroke and cancer rates. This is rather unexpected, but I suspect they are neighborhoods where people may work, but not live in, inflating the number of tobacco sellers, and deflating the stroke and cancer rates.

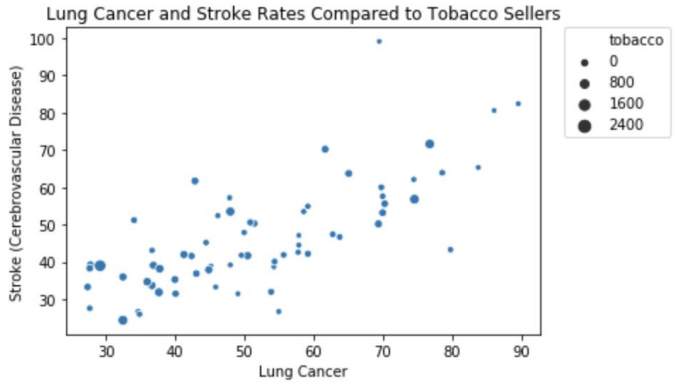

Fig. 6

In Fig. 6, we see the relationship between tobacco sellers in a neighborhood, and lung cancer and stroke rates. is a stronger positive correlation between stroke and lung cancer rates compared to stroke and all cancer rates. Noticeably, there does not seem to be much of a correlation with the number of tobacco sellers within a neighborhood.

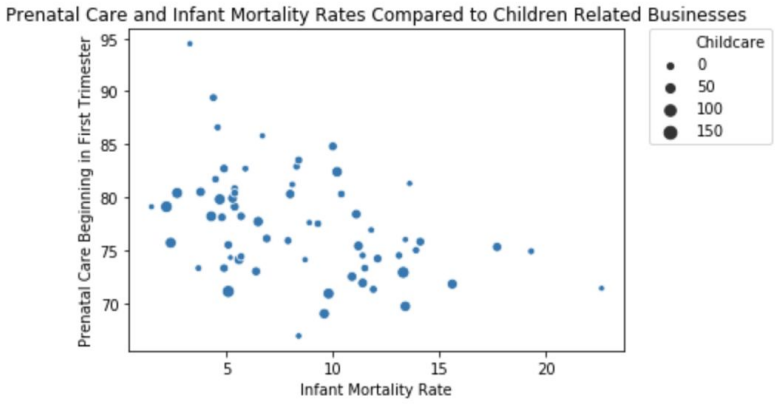

Fig. 7

Finally, we visualized the relationship between infant mortality rate, prenatal care beginning in the first trimester and the number of children related businesses in a neighborhood. which includes tutoring operations, daycare centers etc. There seems to be a negative correlation between prenatal care starting in the first semester and infant mortality rate. This logically makes sense. However it seems to be that the lower the rate of prenatal care beginning in the first trimester, the more child related businesses are within the neighborhood. This is remarkable and may merit further inspection.

Inferential Statistics

The overarching goal of this project is to recommend actions to the City of Chicago to promote healthiness and health equity within all neighborhoods. In addition, we will attempt to model and predict the lung cancer and infant mortality rates for neighborhoods in Chicago. In order to do this we must first perform some inferential statistical analysis. To do so I chose to perform a frequentist inference study.
This study aimed to answer two questions.

1) *Is there a significant difference in lung cancer rates in neighborhoods with high number of tobacco sellers compared to neighborhoods with a low number of tobacco sellers?*
2) *Is there a significant difference in infant mortality rates in neighborhoods with a high number of child related businesses compared to neighborhoods with a low number of child related businesses?*

The null hypothesis for the first question is that there is no difference in lung cancer rates in neighborhoods with a high number of tobacco sellers compared to neighborhoods with a low number of tobacco sellers. To run my frequentist test, I split the dataset so that any neighborhood with a lower than average number of tobacco sellers was considered “low tobacco neighborhood.” any neighborhood with an above average number of tobacco sellers was considered a “high tobacco neighborhood.” I then performed a 2-sided T-test. With the following results.

Ttest_indResult(statistic=-2.6192258278255816, pvalue=0.011372935200506825)

Given that the statistic is less than the p-value, and the p-value is less than 0.05, we reject the null hypothesis, meaning that there is a significant difference in lung cancer rates in neighborhoods with a high number of tobacco sellers compared to neighborhoods with low number of tobacco sellers. The null hypothesis for the second question is that there is no difference in infant mortality rates in neighborhoods with a high number of child related businesses compared to neighborhoods with a low number of child related businesses. Similarly to the first analysis, I split my dataset based on if a neighborhood had an above average number of child related businesses, or if a neighborhoods had a below average number of child related businesses. I again did a 2-sided T-test, with the following results.

Ttest_indResult(statistic=-0.06516825791640102, pvalue=0.9482522439304938)

Given that the p-value is greater than 0.05, we accept our null hypothesis. There is no significant difference in infant mortality rate based on the number of child related businesses.

Given the results of these two frequentist analysis, we can safely say that lung cancer rates are affected by the number of tobacco sellers within a neighborhood, and that infant mortality rates are not affected by the number of child-related businesses. Further frequentist testing can be used to analyze other businesses affect on various health metrics, but is beyond the scope of this project.

Conclusion

The goals of this project was to analyze overarching health trends in the City of Chicago, and to understand the affects certain businesses have on health metrics, specifically lung cancer rates, and infant mortality rates. From our exploratory data analysis, we can conclude that neighborhoods with increased amounts of tobacco sellers tend to have a higher rate of lung cancer. The number of child related businesses do not have an effect on infant mortality rates within a neighborhood. Overarchingly, it is evident that the south and west side of the city suffer the most from health care disparities. The next step of the project is creating an algorithm that will be able to predict the lung cancer and infant mortality rates within a neighborhood. This can potentially be used to measure the impact a new business would have within a neighborhood.