Adithyan Subramanian
Capstone Project 1 Final report
Mentors: Mukesh Mithrakumar and Danny Wells

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. 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.
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, we 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 we had to use the longitude and latitude to locate each business in each neighborhood. We were able to find a json file with a list of polygons for each neighborhood. We read this file into the Jupyter notebook, changed the list from a list of lists to a list of tuples, and wrote a function to take a latitude longitude set to determine which neighborhood that point was in. we 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 lung cancer rates.

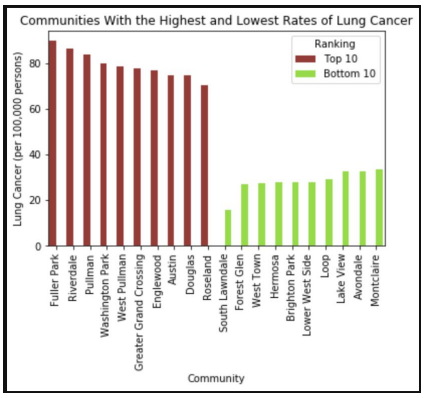

Fig. 1

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.

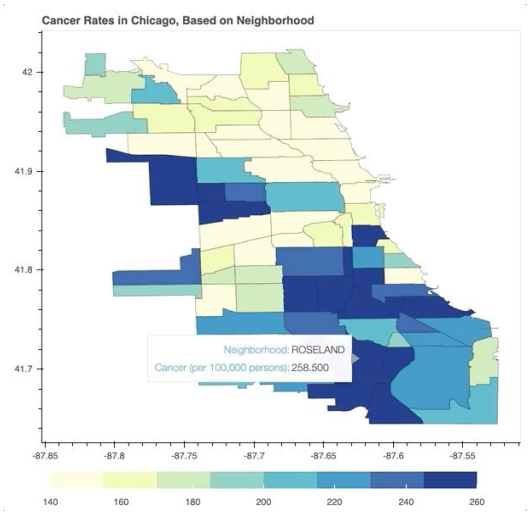

Fig. 2

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.

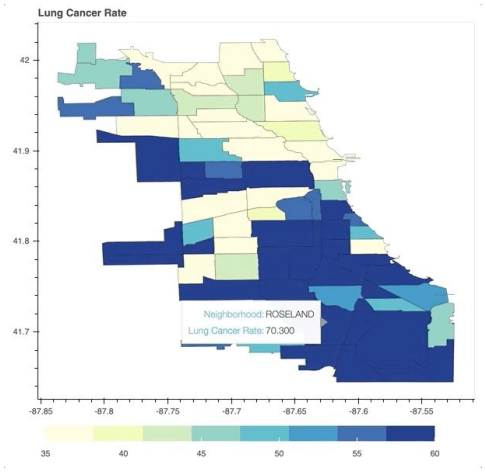

Fig. 3

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.

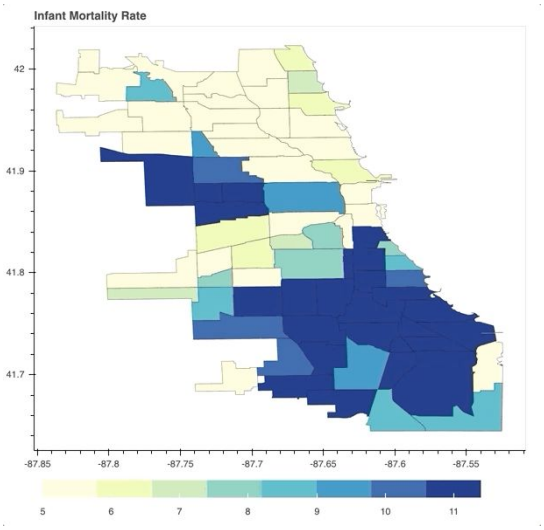

Fig. 4

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.

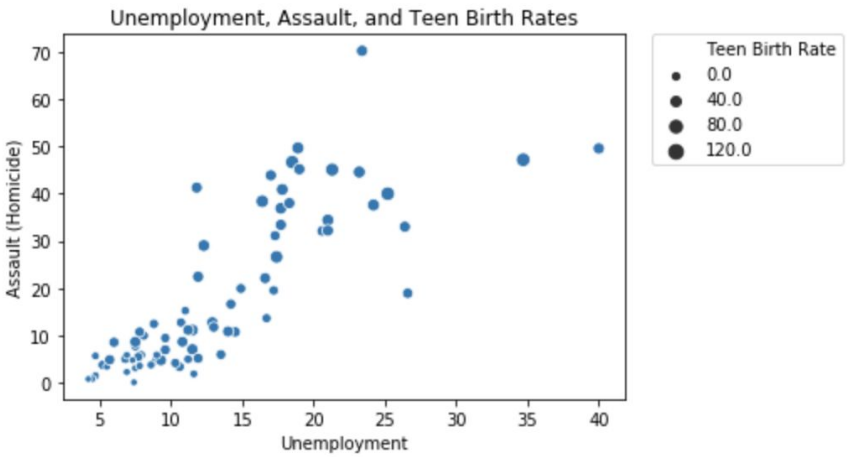

Fig. 5

We also thought it was vital we saw how the health metrics compared to one another. In fig. 5, 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.

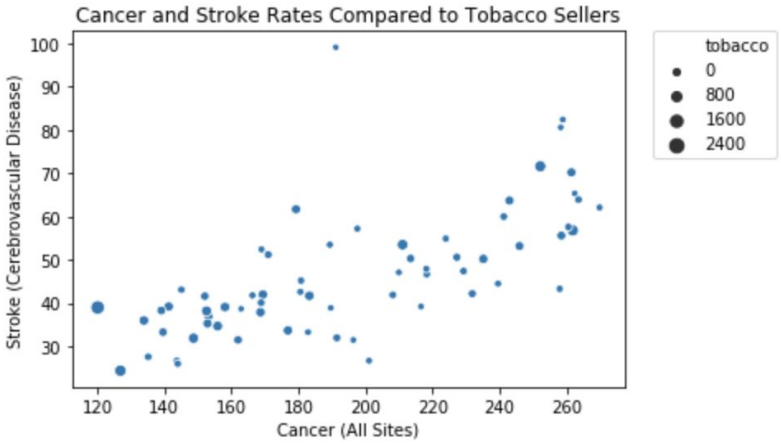

Fig. 5

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 we 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.

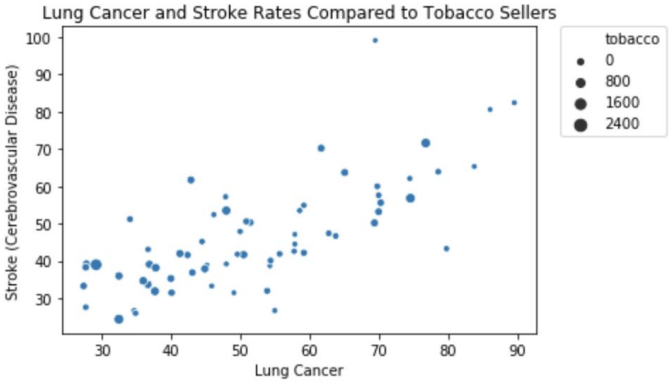

Fig. 6

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.

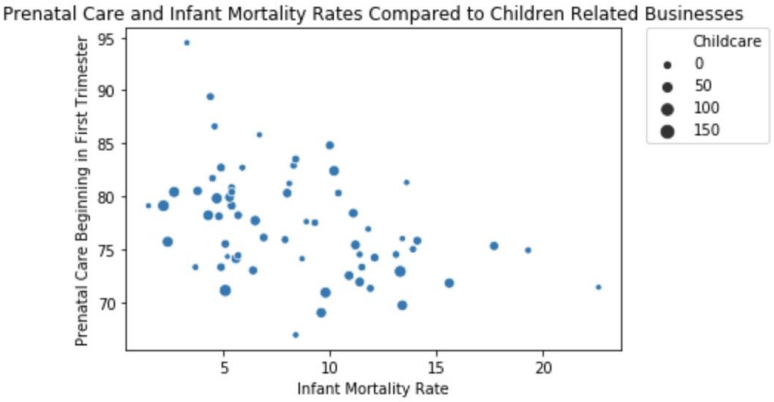

Fig. 7

Finally, in Fig 7, 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 we 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 the frequentist test, we 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.” we 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, we 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. We 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.

Machine Learning

The final goal of this project was to attempt to model and predict the lung cancer and infant mortality rates for neighborhoods in Chicago. In order to do this, we used several methods.
To start, we initially performed a linear regression to predict lung cancer and infant mortality rates. Since there were 30 possible features to choose from, from the final dataset, we had to choose features which would give us the best model. In order to create our model for lung cancer, we chose to focus on the following features: number of tobacco sellers, percentage of residents with no high school diploma, per capita income, number of liquor sellers, percentage of occupied housing, the percentage of households living under the poverty line, and percentage of adults who are dependent to predict lung cancer rates. Initially we fit an ordinary least squares model using these features. We were able to create a model with an R-squared value of 0.632. We also took a look at the coefficients of the features and came away with the following equation to predict lung cancer

$$LungCancer = 38.9335 + 0.0001x_1 - 0.0532x_2 - 0.0003x_3 - 0.0004x_4 - 2.4849x_5 + 0.6146x_6 + 0.6155x_7$$

Where x_1 is the number of tobacco sellers in the neighborhood, x_2 is the percentage of residents with no high school diploma, x_3 is the per capita income, x_4 is the number of liquor sellers in the neighborhood, x_5 is the percentage of occupied housing, x_6 is the percentage of households below the poverty level, and x_7 is the percentage of adults who are dependent.

From this equation, we can see that the strongest factors in predicting lung cancer are percentage of occupied housing, percentage of households below the poverty level, and percentage of adults who are dependent. With that said, We would suggest that public health workers working to lower lung cancer rates focus on increasing the percentage of occupied housing, and decrease poverty and dependency rates.

In order to create a more accurate model, we chose to implement scikit-learn's decision tree regressor. Initially we used the same features with no optimization. This model had a mean-squared error of 160.9, and an accuracy of 0.138 with the testing set. We then used a gridsearchCV function to optimize the maximum depth of the tree, the number of features used in the model, as well as the minimum number of samples needed before the tree split. Using the gridsearchCV function, we found that the best model would use 7 features, have a maximum depth of 2, and the minimum number of samples before branching would be 2. This model gave an improved MSE of 149.85, and an accuracy of 0.198 on the testing set.
As the accuracy ratings for the decision tree was low, we attempted using other tools to create a better model. We chose to perform a random forest model, as well a lasso and ridge regression using the same features. We were concerned that our decision tree and random forest models were overfitting to the training data set, so we decided to utilize the lasso and ridge regression to examine if there was any improvement to the model. The result of these three models are shown in the table below.

Model Type	MSE	Test Set Accuracy
Random Forest	89.9	0.519
Lasso Regression	154.5	0.172
Ridge Regression	154.4	0.173

It is important to note that we did optimize the random forest model, using a GridSearchCV, like we did for the decision tree. The optimal parameters were 20 estimators, and 5 minimum number of samples before branching. Since none of these techniques were able to exceed the accuracy of the ordinary least squares model, we would suggest using the ordinary least squares method to predict lung cancer rates and to find insight.

We followed a similar approach to predict infant mortality rates. We used the following features to create our model:
Number of “child-related businesses”, percentage of residents with no high school diploma, per capita income, percentage of occupied housing, percentage of households living under the poverty line, percentage of babies born with low birth weight, the general fertility rate, the number of childhood lead poisoning per 100 children, number of females per 100 with gonorrhea, and the number of liquor sellers within a community. We created an ordinary least squares model using these features. This model had an R-squared value of 0.634. Using the coefficients we came with the following equation

$$InfantMortality = 6.9660 - 0.0120x_1 - 0.0042x_2 - 3.572e-05x_3 - 0.1713x_4 + 0.0588x_5 + 0.0945x_6 - 0.0105x_7 + 1.5477x_8 + 0.0022x_9 - 0.0006x_{10}$$

Where x_1 is the number of child related businesses in the neighborhood, x_2 is the percentage of residents with no high school diploma, x_3 is the per capita income, x_4 is the percent of occupied housing, x_5 is the percentage of households below the poverty level, x_6 is the percentage of babies born with low birth weight, x_7 is the general fertility rate, x_8 is the number of childhood lead poisoning per 100 children, x_9 is the number of females per 100 with gonorrhea, and x_{10} is the number of liquor sellers in a neighborhood. From this equation we can see that the strongest factors in predicting infant mortality is the number of childhood lead poisoning per 100, the percentage of occupied housing, and the percentage of babies born with low birth weight. It is recommended that public health workers working to lower infant mortality rates focus on lowering the number of childhood lead poisoning per 100 as well the percentage of babies born with low weight, and increase the percentage of occupied housing. It is also interesting to note that per capita income has a marginal effect on infant mortality. Further investigation will be needed to understand this relationship.

Using the same process as what was used to create our lung cancer model, we created a decision tree model. The best model gave us a low MSE of 13.029, and an accuracy of 0.221 on the test set. As in the case of predicting lung cancer rates, since the decision tree model was unable to predict

infant mortality as accurately as the ordinary least squares, we would suggest that the ordinary least squares model be used to predict infant mortality rates and find insight.

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.

From a machine learning perspective, it is important to note that although the main purpose was to create models that can predict infant mortality and lung cancer rates, another goal was to use these models to formulate a better understanding of how these features affected the rates in question. We can recommend that public health workers working on reducing lung cancer rates prioritize solutions where occupied housing increases, and poverty and dependency rates decreases, while public health workers working on reducing infant mortality rates prioritize solutions where the number of childhood lead poisoning per 100 as well the percentage of babies born with low weight decreases, and the percentage of occupied housing increases.

There are several reasons as to why all the models had low accuracy. Forthmost is the fact that our dataset is small. Given we only had data for the 77 communities of Chicago, the dataset may have been too small and too limited for any accurate machine learning to take place. It is also important to note that while the dataset gives a good picture of health outcomes to all Chicago neighborhoods, it is not an extensive dataset that takes into account all socioeconomic and health factors that may play a part within the proliferation within these diseases. In short, more data is the only way to create more accurate models.

There are a variety of ways that this work can be expanded upon. To start, we only explored two particular types of businesses and how they affected two specific health outcomes. Mainly tobacco sellers affect on lung cancer, and child related businesses affect on infant mortality. The type of businesses and the effect they have on various health outcomes is virtually inexhaustible. One can very simply look at the various types of businesses and attempt to analyze correlations to various health outcomes. It may also be interesting to look at distances between certain types of businesses and if that has any correlation with negative health outcomes.

From a machine learning perspective, A next step to this project would be to incorporate neighborhood health data from other major US cities and see if the increase in data is able to create a more accurate model. One can also use the current data to create models for other healthcare metrics, such as diabetes, or prostate cancer in males.

Works Cited:

[1]https://www.chicagotribune.com/real-estate/ct-re-0603-housing-segregation-20180525-story.html

[2]https://news.wttw.com/2017/03/23/survey-reveals-alarming-health-disparities-chicago-neighborhoods