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

Adithyan Subramanian
Mentors: Mukesh Mithrakumar and Danny Wells

Background and Goals

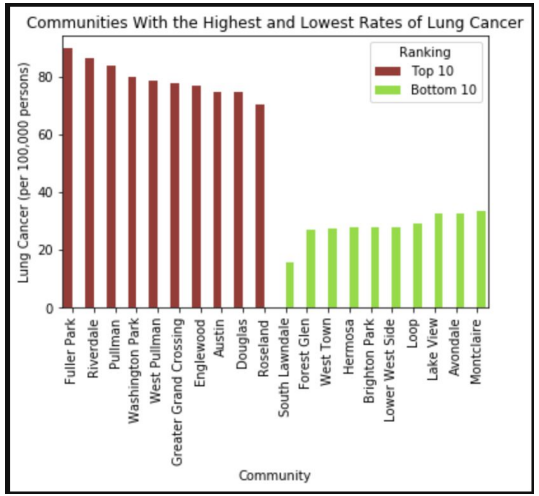
- Chicago - Extremely diverse and segregated
- Segregation = health disparities
- Which neighborhoods are more healthy?
- How do businesses affect neighborhood health?
- Predict lung cancer and infant mortality rates

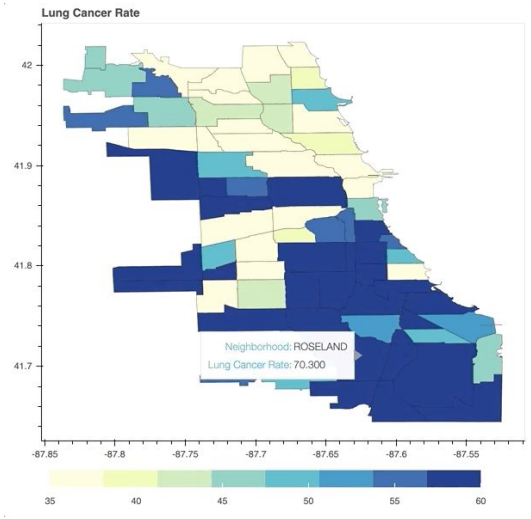
Data

- 2 Datasets
  - Health dataset
  - Business dataset
- Both from the Chicago Data Portal

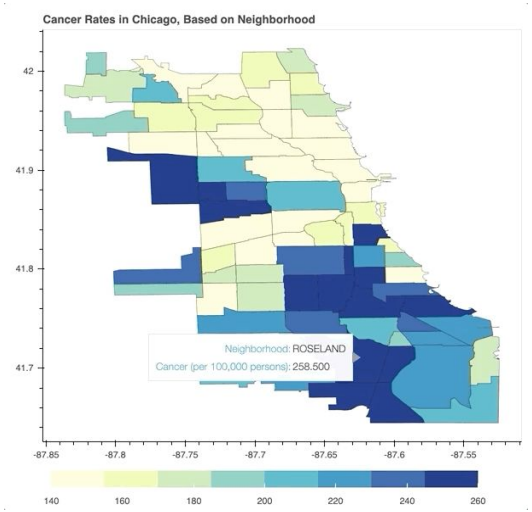
Exploratory Data Analysis

Lung Cancer Disparities: A Case study

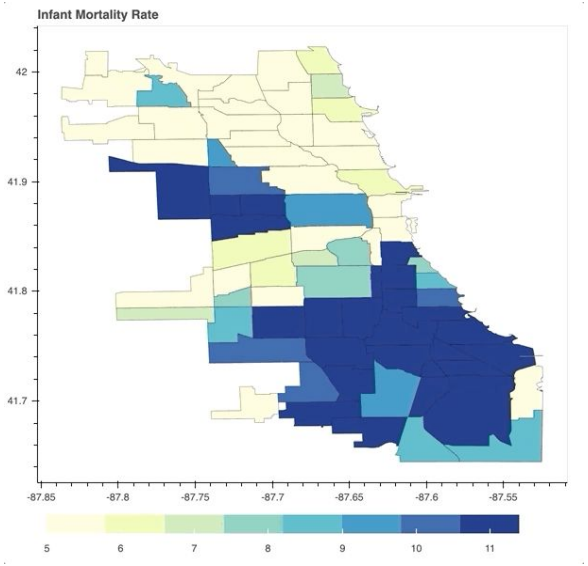




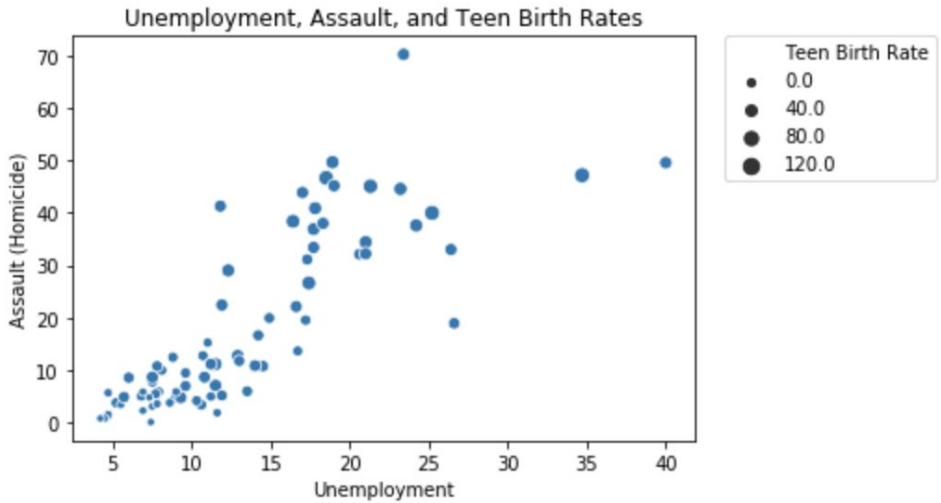
General Cancer Rates



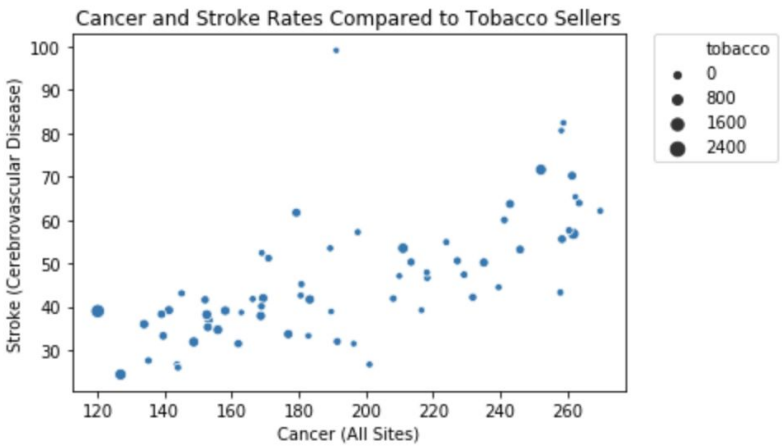
Infant Mortality Rates

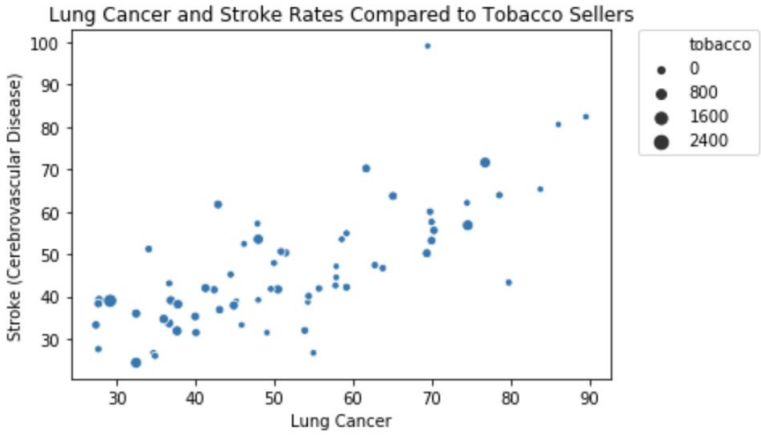


Comparison of Socioeconomic Factors

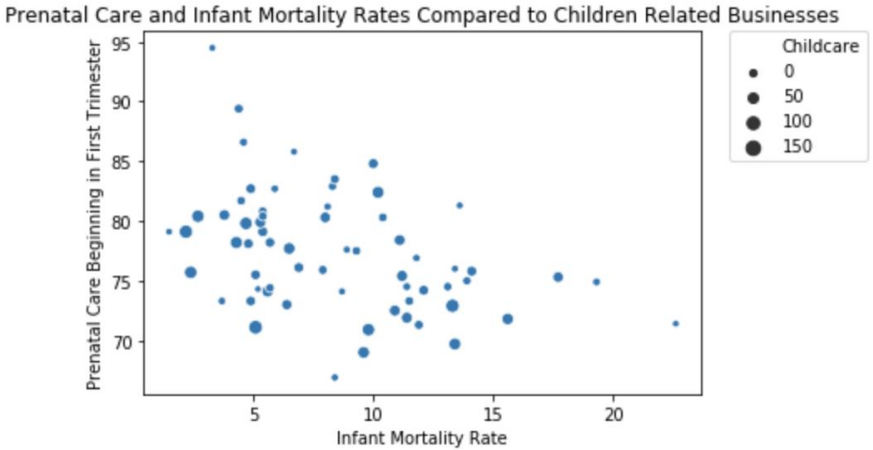


The Effect of Tobacco





Infant Mortality and Child-Care Businesses



EDA Conclusions

- Significant difference in lung cancer rates in neighborhoods with a high number of tobacco sellers compared to neighborhoods with low number of tobacco sellers.

- No significant difference in infant mortality rate based on the number of child related businesses.

Machine Learning

Infant Mortality

- Used following metrics as features
  - 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,
  - General fertility rate
  - Number of childhood lead poisoning per 100 children
  - Number of females per 100 with gonorrhea
  - Number of liquor sellers within a community.

Infant Mortality Results

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

- Linear Regression
  - R^2: 0.63
- Highest factors: Childhood lead poisoning, percentage of occupied housing, percentage of babies born with low birth weight
- Also created and optimized decision tree, but with accuracy lower than linear regression model

Lung Cancer

- Used Following Metrics as features
  - Number of tobacco sellers
  - Percentage of residents with no high school diploma
  - Per capita income
  - Number of liquor sellers
  - Percentage of occupied housing
  - Percentage of households living under the poverty line
  - Percentage of adults who are dependent

Lung Cancer Linear Regression Model

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

- R^2: 0.62
- Highest factors: percentage of occupied housing, percentage of households below the poverty level, and percentage of adults who are dependent

Other Algorithms

Method	MSE	Test Set Accuracy
Non-optimized Decision tree	160.9	0.138
Optimized Decision Tree	149.9	0.198
Non-optimized Random Forest	119.2	0.36
Optimized Random Forest	89.9	0.52
Lasso Regression	154.5	0.17
Ridge Regression	154.4	0.17

What Went Wrong

- Small data set
  - Only 77 neighborhoods
- Not extensive
  - Does not incorporate all business and health care data

Conclusion

Final Thoughts and Recommendations

- Must prioritize South and West sides of the city
- Less tobacco sellers = less lung cancer
- Prioritize solutions where occupied housing rises, and the poverty and dependency rates decreases to lower lung cancer rates
- Prioritize solutions where occupied housing rises, and where childhood lead poisoning and low birth weight rates decreases

Next Steps

- Other businesses?
- Other outcomes?
- How location of businesses affect health outcomes?
- More data?
    - From other cities?
- Models on other metrics