Stacey Stevens and Sam Duffield

8 December 2017

Introduction to Data Science: Illinois Health Outcomes of Top 18 Most Populated Cities

St. Louis College of Pharmacy

We looked at various databases available online and decided to use a data set called 500 cities available from <https://www.healthdata.gov/dataset/500-cities-local-data-better-health-0>. After using RStudio to browse through the full data set, we decided that our working hypothesis would concentrate on identifying relationships between health outcome measurements listed in the data set and various geographic regions. We used the code contained in the Filter Data R file to open and filter the original online data set to concentrate on health outcomes measurements for cities in Illinois and other variables in the data set that could be useful for developing our analysis. We chose variables that would be useful in determining our formal hypothesis and methods for supporting or disproving it.

As we were preparing to visualize our data, we decided that it would be necessary to create variables to determine incidence rates as percentages that would be representative of the entire city for the year 2014 in order to avoid counting patients twice. The code used to create this analysis is contained in the Variable Analysis R file. Our graphs were each created with the R scripts contained in the Visualize Your Data files. All of the scripts need to be set to a working directory containing the downloaded data, as our original data set was too large to upload to GitHub. Our graphs represent the various incidences of each negative health outcome as a percentage of the population of each town treated and recorded by the original data set. We synthesized the results from this analysis into the table in our results section by using sorting functionalities for data sorting in R and recording the results in an Excel file.

We only included the cities with the top ten average rankings for high incidences of each health outcome measured for the sake of space, but we included all eighteen cities in our analysis. The graphs and table in our poster represent an initial investigation into this subject meant to use the techniques learned in class to select, sort, adjust, and graph a much less manageable database found online. All of the results from our R scripts were utilized in creating and presenting the results of this analysis. Although the data suggested our initial hypothesis was incorrect, it conclusively suggested that there were discrepancies between regions for the health outcomes that we selected. Our analysis was insufficient to suggest causal mechanisms for our data, but we did successfully use our original data set to point to trends that may be worth further study. Particularly, the differences in the health outcome data between the two groups of nonmetropolitan cities as well as the high incidences found in certain metropolitan areas caught our attention as worthy of further study.