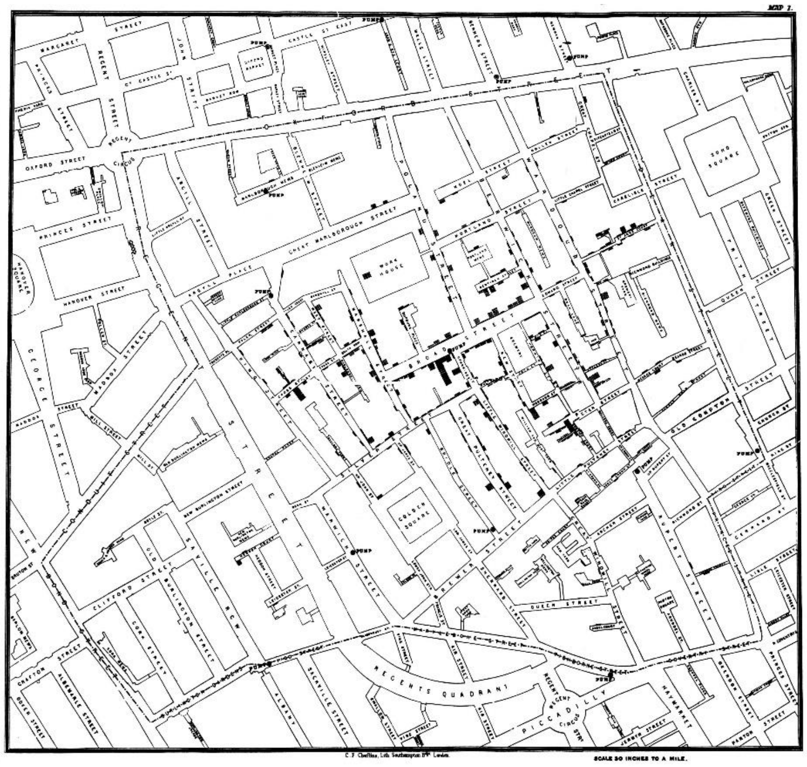
Data Preparation and Analysis

Martin Rosenfeld
A20350531

Homework 2

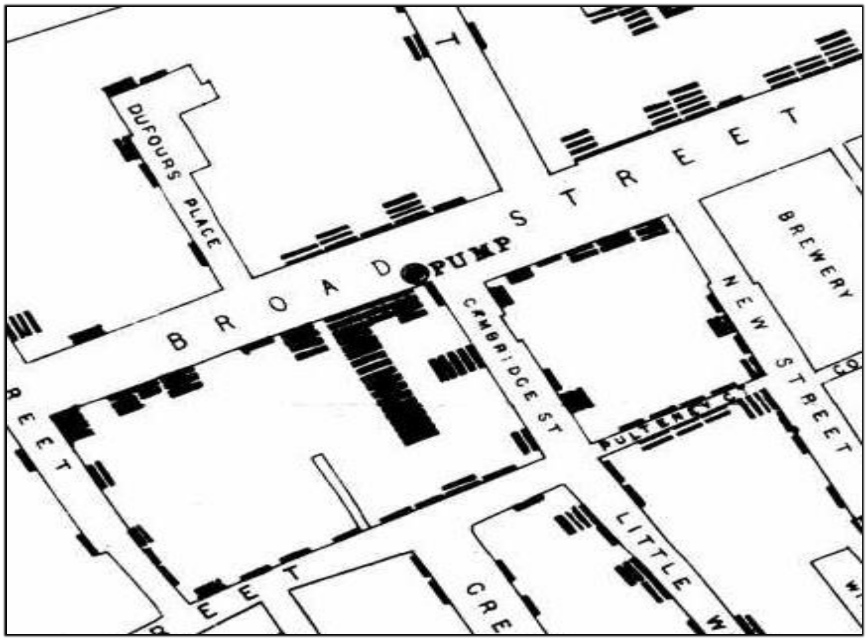
1. The London's 1854 cholera outbreak[1] was a severe outbreak of cholera that was centered near Broad Street in the Soho district of London. Until that moment, the prevailing theory to explain diseases was the miasma theory. The miasma theory stated that diseases were spread by "miasma" in the air[2], i.e. they were spread by pollution or a noxious form of "bad air".
   John Snow, whose house was closed to the Soho District, was a skeptic of the miasma theory. His hypothesis was that contaminated water, not air, spread cholera. He examined water samples from various wells under a microscope, and confirmed the presence of an unknown bacterium in the Broad Street samples. To support his hypothesis, and despite strong skepticism from the local authorities, he published the following map of the epidemic.



   Zooming the area of interest gives us the following map:

[1] https://en.wikipedia.org/wiki/1854_Broad_Street_cholera_outbreak

[2] https://www.udel.edu/johnmack/frec682/cholera/



The complete map shows the 578 cholera deaths mapped by home address, marked as black bars stacked perpendicular to the streets. In this way, Snow showed how cases of cholera were centered on the pumps.

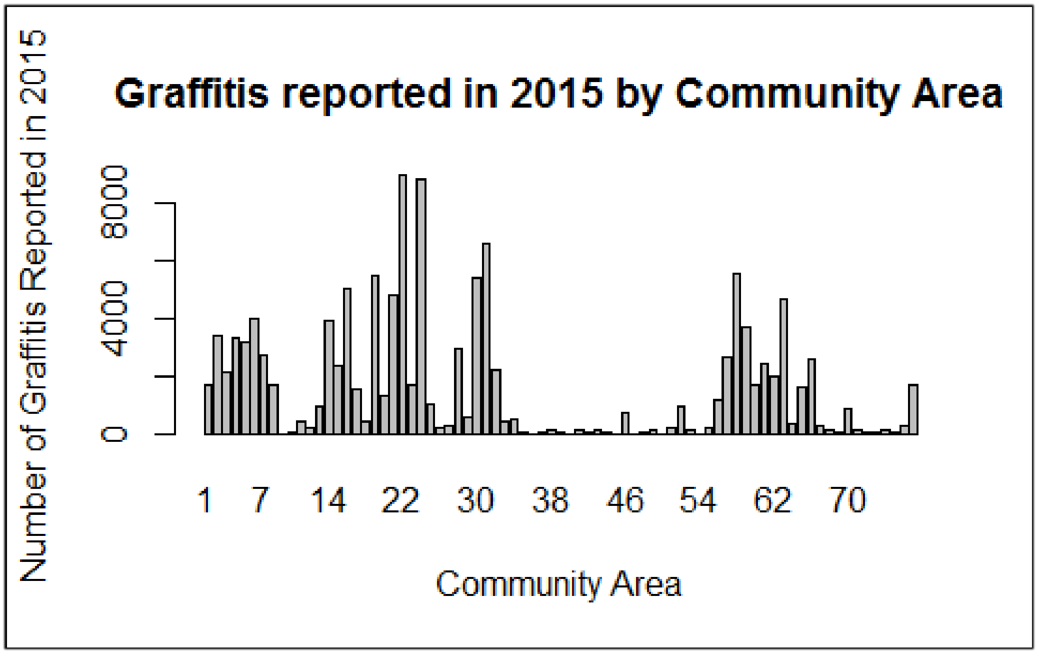
The important lesson from this history is that John Snow could convey his idea by creating a visualization of the data he collected. The created map helped him show in a simple and fast way his theory. Obviously, when transforming his raw data to a picture, some data were lost. However, the visualization was simple and abstract enough to convey his idea.

One other thing to mention is that the visualization alone is not enough to demonstrate his theory. Some questions can arise, e.g. regarding the way Snow collected the data, the reliability of the data, the interpretation of the map, etc. This means that a picture does not replace the whole data science life cycle. It is a powerful mean to convey an idea in a much simpler and more effective way.

As a data analysis problem, this history shows the whole data science life cycle. It started with data collection, which probably took more time than the rest of the phases. It included surveys and getting official data of the deaths. It followed by stating a hypothesis and doing a statistical analysis to illustrate the connection between the quality of the source of water and cholera cases. And a last phase of communicating and transmitting the results to the stakeholders. After that, actions were taken and new data was collected which was used to test Snow's hypothesis.

2.

   a. The following figure shows a barplot with the Community Area number in the x axis and the number of graffitis reported during 2015 in the y axis. I decided to show the community area number instead of the name, because showing 77 names in the x axis would make the graphic difficult to visualize.



In can be seen that most of the community areas have less than 4000 graffitis reported, and also the are many with small bars with very few graffitis reported. Besides, 2 pikes are identified which are community areas with more than 8000 graffitis reported. They are West Town and Logan Square.

b. The Chicago Community Areas date from the late 20's[3], when the University of Chicago in collaboration with Chicago's Department of Public Health produced a first map with 75 community areas. In the time, the existing "ward system"[4] used by the United States Bureau of the Census was unsuited because ward boundaries changed with each census cycle. The community areas were chosen such that they had distinctive histories and consistent rates of various social ills, regardless of who lived there.

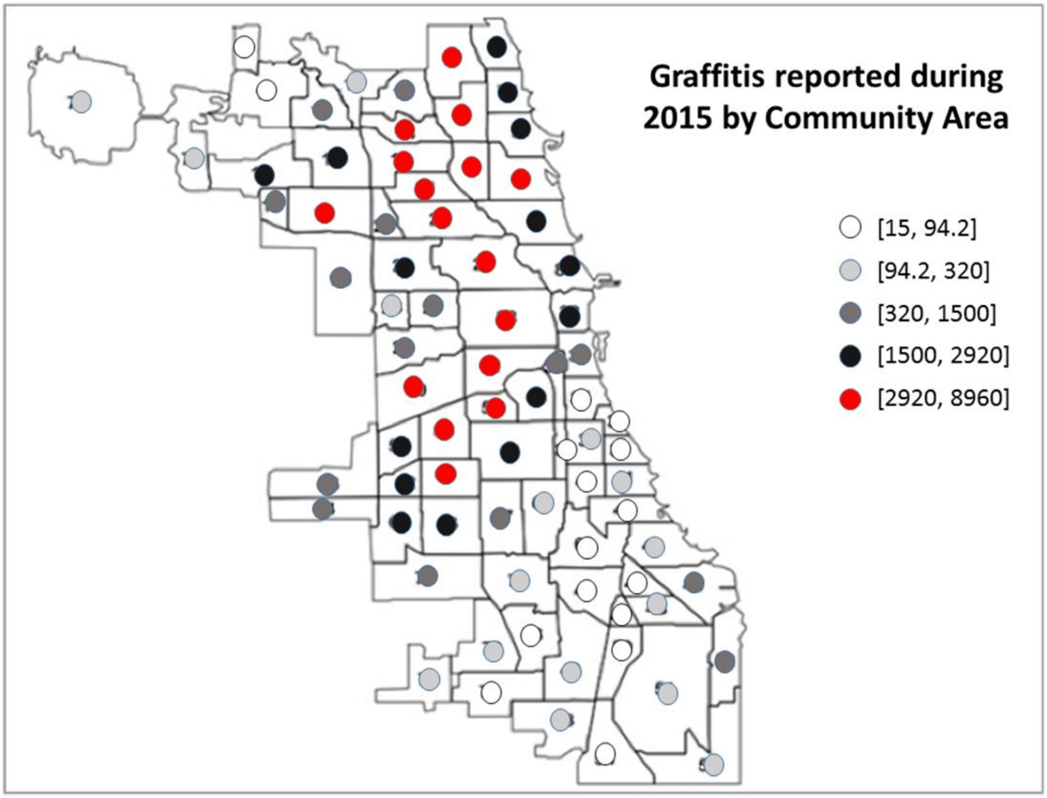
There have been only minor changes in the community area map so far. Specifically, two new areas were added (to see the official complete list refer to [5]). Maintaining the boundaries of the neighborhoods relatively stable, although not reflecting changes, allows comparisons of these areas over time.

The following figure is a map of the community areas. To reflect the graffitis reported by community area, I bucketized the data in 5 quantiles, and chose a different color per level. The community areas with less graffitis reported during 2015 contain a white circle. The increasing range of greys indicates more number of graffitis. Finally, the red color determines the community areas with most graffitis.

[3] https://en.wikipedia.org/wiki/Community_areas_in_Chicago

[4] http://www.encyclopedia.chicagohistory.org/pages/1316.html

[5] http://www.cityofchicago.org/content/dam/city/depts/doit/general/GIS/Chicago_Maps/Citywide_Maps/Community_Areas_W_Numbers.pdf



By doing this, I was able to visualize the effect of the geography in the number of graffitis reported.

Clearly, the community areas with most graffitis reported during 2015 are in the north and central sides of the city. While moving to the south side, there is a decrease which can be noted by the colors of the circles.

The above statement is a description of the visualization. The interpretation of this is a more complex task. It should be done by specialists in the sociology and historical aspects of Chicago population and its neighborhoods. One possible explanation is that the north and central community areas, including downtown, have a more artistic background than the south ones. Another possible explanation, taking into account that the data set comes from the graffitis “reported” to the Chicago government, is that people in those community areas are more concerned about graffitis and the visual aspects of the neighborhoods where they live and, thus, report them more often. This means that the number of graffitis reported is not necessarily equal to the number of graffitis which actually exists.