PROJECT 1 SUMMARY

John Pospisil

Our team selected to study relationships between Chicago parks, crimes and potholes. In the brief time we had to discuss what we wanted to study. There could be a relationship between number of parks and crimes. One hypothesis was that number of parks and number of crimes by ward would be negatively correlated. Park availability has been used as a surrogate for higher socioeconomic status and quality of life in a neighborhood (in this case we used ward) and it’s likely areas with higher socioeconomic status and quality of life have less crime. Alternately, parks are great places for people to hide and thus could be a conduit for greater crimes levels. The potholes idea was a fun idea one team member had.

Data was retrieved via the Chicago Data Portal (CDP). Park data was only for 2014, while we retrieved a greater number of years for committed crimes. I volunteered to take the parks and crimes hypothesis. Thus, I only used 2014 crimes as it matched the parks data. It is reasonable to try using all crime years related to parks. However, it was thought that the number of parks would have changed over time, so I only used 2014 data.

I filtered crimes by overall top 5 crime types rather than using relatively obscure crimes types. The, I eliminated park types if it could not be determined what a park definition was. For example, what is a “spray feature?” Is that a wall painted with murals or literally some water device that sprays?

While it was clear some wards had significantly higher levels of crime, it was hard to find a statistical relationship. The data being counts, it would most likely follow a Poisson distribution. I never got a Poisson regression to work in Python (but could certainly do it in R, SAS or SPSS). Chi-square is a great way to look at counts, but there were so many categorical variables they would have to be collapsed, and honestly there wasn’t enough time to go that route.

There were lessons learned from the process. First, the entire team wasn’t on the same page, which led to individuals just doing whatever we wanted with our data. Next, only two team members has significant experience with statistics. This wasn’t optimal. Finally, the project needed to be scoped down, as looking at everything and see what happens is a difficult project to manage in 10 days.

So, was the project successful. We did collect data, it was analyzed at a very low level and a presentation was made. However, given the short time, overly broad hypotheses and lack of team coordination I really don’t consider it successful.