Project I Proposal

1 MOTIVATION AND SUMMARY

This project will focus on Detroit City's neighborhoods' characteristics and crime incidents. Detroit has experienced continuous demographic and economic decline during recent decades. The population of Detroit has dropped from 1.85 million in 1950 to 670,000 in 2015. Due to its considerable economic growth during the peak of the auto industry, the population of Detroit peaked in the 1950s. Thus, some neighborhoods in this city got lots of demolished houses and empty lots, while others may be relatively more well-maintained. After World War II, deindustrialization and decentralizing trends of the auto industry have made Detroit a representative shrinking city. While population keeps losing and many houses were abandoned and demolished in Detroit, worsened economic condition has brought crimes in this city. According to Federal Bureau of Investigation Uniform Crime Reports, Detroit ranked 2nd in the list of most violent cities in the U.S. As a student in the School of Environment and Sustainability, I am always interested in looking at data relevant to urban informatics. Thus, I want to take a look at the relationship between the city's shrinkage and crime incidents.

2 DATASETS

Two datasets will be involved in this project. The first one is Motor City Mapping dataset. The Motor City Mapping is a project gathering and digitizing Detroit's parcel information. The API link for this dataset is http://portal.datadrivendetroit.org/datasets/80f30d7f6683441cacef62574a22d8a9_0.geojson. The dataset should include information on:

- whether there is a built-up structure in this parcel.
- what neighborhood (in this project, analyses are focusing on the level of census tract) the parcel is in,
- and, the use of the parcels (e.g. residential, commercial, industrial, etc.). These two properties are critical for us to identify the demolition rate for a neighborhood.

The second dataset would be all crime incidents. This dataset is based on information provided by Detroit Police Department (DPD). The API doc for this dataset can be found at https://dev.socrata.com/foundry/data.detroitmi.gov/fxch-8vn6. The dataset should provide information about:

- the category of each crime incident,
- the date, and the hour of incidents,
- and the census tract where the incident happened, etc.

3 MANIPULATION AND JOIN

The basic idea joining these two datasets is that both these datasets should share the same column which indicates the census tract. Before this, necessary data tidying should be conducted. For instance, some rows do not contain the information we are interested in. Based on the parcel dataset, we could calculate the demolition rate simply by (# all parcels)/(# demolished parcels) within one census tract. For the crime dataset, we may calculate frequency of crime incidents by census tract, frequencies of certain categories of crimes, and how these frequencies changed over time.

4 MAP-REDUCE TASKS

First, we want to check if demolitions are correlated to crime incidents. Intuitively, we may think that crimes are more likely to happen in city's sketchy areas, but crime can be influenced by many other factors as well. To test this, for each incident we can generate maps like (neighborhood code, 1) to calculate the number of crime incidents within each neighborhood. The output may be sorted by the calculated demolition rate. The task can be done using *spark*.

Second, I am also interested in how crime rates changed among "good" and "bad" neighborhoods over time. For identifying "good" and "bad" neighborhoods, we may extract a certain (say, 10% or 25%?) percentage of neighborhoods with highest and lowest demolition rates, which could be implemented in *sparksql*. Since the crime dataset should contain

information on the date and hour of crime incidents, we can see how frequencies of crimes changed in different types of neighborhoods at the level of month or year. Then we may compare if the patterns of change will be different for those neighborhoods. The map key could be the combination of neighborhood and time, which could be reduced in spark.

Third, I am also assuming that major categories of crimes may vary in different neighborhoods. For example, it is possible that sketchy areas may see a higher proportion of violent crime incidents. Thus, using the attribute "category" of the crime dataset can help with this analysis. We can compare the counts, or the proportions of violent crime incidents in the "good" and "bad" neighborhoods. Similarly, the map key could be a combination of crime category and neighborhood, and could be further reduced with *spark*.

Other possible analysis

Also, I am curious about whether parcel use could possibly influence the crime incidents. Are crimes more likely to occur in residential areas, or areas with more commercial/industrial parcels?

5 VISUALIZATION

I may use R to present my results. For example, scatterplots can be used for presenting the relationship between demolition and crime frequency. With R I can also present the changing patterns in crime incidents. Also, a geographical map for Detroit may be used to show how the crimes are distributed and differentiated among neighborhoods.