Yelp and Crime Data

Kenneth Lin, Tom McCormick, Sid Naik

November 27, 2014

1 Introduction

For our CS 194 final project, we decided to investigate the potential relationship between the City of San Francisco public safety data set and the data set provided by the Yelp API. At first, we explored each data set individually to discover general patterns in the data set. After that was completed, we looked at the nearest 20 restaurants near each crime to see if we could discover any common patterns of restaurants close to crimes.

2 City of San Francisco Public Safety Data Set

	IncidntNum	Category	Descript	DayOfWeek	Date
	click to scroll output; double click to hide				
0	140001966	NON-CRIMINAL	DEATH REPORT, CAUSE UNKNOWN	Wednesday	01/01/2014
1	140003025	NON-CRIMINAL	DEATH REPORT, CAUSE UNKNOWN	Thursday	01/02/2014
2	140004487	NON-CRIMINAL	DEATH REPORT, CAUSE UNKNOWN	Thursday	01/02/2014
3	140000059	NON-CRIMINAL	AIDED CASE	Wednesday	01/01/2014
4	140000071	VANDALISM	MALICIOUS MISCHIEF, VANDALISM OF VEHICLES	Wednesday	01/01/2014

Time	PdDistrict	Resolution	Location	X	Y
16:21	TENDERLOIN	NONE	400.0 Block of ELLIS ST	-122.413794	37.784772
02:00	MISSION	NONE	500.0 Block of JERSEY ST	-122.438235	37.750203
14:30	BAYVIEW	NONE	100.0 Block of CORAL CT	-122.371925	37.727898
00:17	SOUTHERN	NONE	1500.0 Block of MISSION ST	-122.417566	37.773892
00:30	SOUTHERN	NONE	0.0 Block of MARKET ST	-122.393966	37.795028

Figure 1: Sample of San Francisco crime data set

The majority of the fields are self-explanatory. However, there are a few things to note:

1. Category and descript are both categories, but category is more general. There are only 36 different "Categories" while there are 499 different "Descript"s in the year of 2014.

2. Resolution, though none are shown in the sample above, denote whether any action was taken and what that action was.

3. X denotes longitude, while Y denote latitude.

A more detailed analysis is in the attached `analysis.ipynb`.

3 Yelp Data Set