William Spagnola

Prof. Shroff

APSTA-2047: Messy Data & Machine Learning

HW #1 (Part 1): Dataset Datasheet

September 27, 2018

**Data:** 2010-2016 NYC School Safety Report

**Motivation for Dataset Creation**

**Why was this dataset created?**

This dataset was created for the New York City Department of Education (DOE) to keep track of incidents of crime that occur in New York City schools. As well as informing the DOE about the distribution and trend of crimes in schools across the city, the dataset could be used to provide information to parents about the relative safety of the schools in which their children attend.

**What other tasks could the dataset be used for?**

This dataset could be used by both education and criminology researchers to investigate correlations between higher rates of crime in schools and other school-level and neighborhood level variables. The data could also be used to inform ongoing political debates such as the presence of school safety officers and metal detectors in schools.

**Who funded the creation of the dataset?**

The City of New York funded the creation of this dataset through the DOE and the New York City Police Department (NYPD).

**Dataset Composition**

**What does the data consist of?**

The dataset is compiled at the building level. In New York City, many schools are co-located within the same building. The dataset includes a list of schools within each building. The dataset records the number of crimes by different categories. The crimes are first categorized under three categories: “major crimes” such as burglary and felony assault, “other crimes” such as drug possession and misdemeanor assault, and “non-criminal incidents” that were not crimes but were disruptive to the school environment. The dataset also includes the number of property and violent crimes at each building. The dataset also categorized buildings by their population and presents the city-wide average for each type of crime for the buildings in the given population range.

**Data Collection Process:**

**Who was involved in the data collection process?**

The NYPD collected data of crimes that occurred in school and then reported it to the DOE. According to the “Data Dictionary”, this process has been ongoing since 1998.

**How was the data collected?**

The data were presumably collected based on NYPD reports within schools. However, there is no information on how the NYPD or DOE aggregated the data by building.

**Over what time-frame was the data collected?**

The data were collected annually. The data are organized by school year.

**Is there information missing from the dataset?**

Some schools do not have data in certain fields. Missing data is marked as “N/A”.

**Dataset Pre-processing**

It is not clear whether or not there was any pre-processing of the data other than aggregating the data by building and indicating where data were missing.

**Dataset Distribution**

**How is the data distributed?**

The data is available on the NYC OpenData website at <https://data.cityofnewyork.us/Education/2010-2016-School-Safety-Report/qybk-bjjc>. It can be accessed by downloading the entire CSV or through an API.

**When was the dataset first distributed?**

According to the NYC OpenData, the dataset was first made public on February 21, 2013.

**What license (if any ) is it distributed under?**

The data does not have any license or copyright restrictions most likely because it is public data compiled by the city government.

**Are there any fees or access/export restrictions?**

There are no fees for downloading the data.

**Dataset Maintenance**

**Who is hosting the dataset?**

The dataset is hosted on the NYC OpenData website.

**Will the dataset be updated?**

According to the NYC OpenData website, the data are updated annually. The last data update was from September 16, 2017 and the last metadata update was from September 10, 2018. There is no description on the website concerning what was updated.

**Legal and Ethical Considerations**

**Were people informed about the data collection?**

There is no indication that the students involved in the incidents were informed about the data collection process.

**Could this dataset expose people to harm or legal action?**

Since the data was aggregated at the building-level, it is unlikely that the data will expose any harm to individuals or any additional legal consequences.