Georgia Tech Crime Data Wrangling

Phillip Spratling

This dataset consists of 9 separate csv files detailing crime on and near the Georgia Tech campus. Each file corresponds to a certain year, from 2010-2018. Each entry has multiple attributes, including the time of the incident, the code and description, the location, the disposition, and the patrol zone among others. The formatting is consistent throughout the years.

1. Merging the data

Each file was read and combined into a single dataframe containing records for all 9 years.

2. Extracting relevant columns, renaming columns, and reindexing

Only the columns relevant to the analysis were extracted. The ‘IncidentFromDate’ and ‘IncidentFromTime’ columns were combined into a single column, parsed as a datetime object, and was renamed as ‘time’. The other columns were renamed as well to keep naming consistent. The ‘time’ column was then set as the index of the dataframe.

3. Dropping bad records

Rows that were labeled as a non-crime were removed from the dataframe. Rows that didn’t contain any location data or any descriptive data were also removed.

4. Correcting the type of columns ‘street_number’, ‘lat’, and ‘code’

The ‘street_number’ column was changed from a float to an int, the ‘lat’ column was changed from a string to a float, and the ‘code’ column was changed from a string to an int. This helped merge duplicate values.

5. Cleaning ‘landmark’ column

Multiple entries in the ‘landmark’ column were the same value, but with different formatting. These values were changed to merge duplicate values.

6. Adding coordinate information

Many records did not contain any coordinate information. Coordinate information was added for many of these records using the coordinate information of their corresponding landmark.