**HACKATHON: EXPLORING THE RELATIONSHIP BETWEEN CROWDS, SAFETY AND CRIME**

**University of Manchester, Methods @ Manchester, 9 &11 December 2020.**

**DATA DESCRIPTION**

**Summary**

We have provided four primary sources of data: two for measuring crimes and two for measuring crowding. These are supplemented by some additional datasets which might help you along the way. Feel free to download these files from the GitHub or Dropbox page and make use of them throughout the hackathon! Some of the data have been cleaned either to make them more user-friendly or to reduce the file size. We have provided the R scripts used for this process, along with links to the original data sources, should anyone be interested in looking into this.

|  |  |  |  |
| --- | --- | --- | --- |
| **File name** | **R script** | **Description** | **Original source** |
| la\_crime\_june.csv | la\_crime.r | Police-recorded crime incidents during June 2020 in Los Angeles | [Los Angeles open data archives](https://data.lacity.org/A-Safe-City/Crime-Data-from-2020-to-Present/2nrs-mtv8/data). |
| us\_crime\_2018.csv | crime\_2018.r | Police-recorded crime incidents in 16 major US cities during 2018. | R package *crimedata* (Ashby, 2018). |
| retrosheet\_gl\_2018.csv | retrosheet.r | Major League Baseball game logs for 2018. | [Retrosheet website](https://www.retrosheet.org/gamelogs/index.html). |
| stadium\_coords.csv | None. | Stadium coordinates for Major League Baseball stadiums in US crime data cities. | Open Street Map API. |
| california\_cb\_shp.zip | None. | Shapefile of census block groups in California. | [Government open data portal.](https://catalog.data.gov/dataset/tiger-line-shapefile-2016-state-california-current-block-group-state-based) |
| la\_cb.csv | None. | List of census block groups nested within Los Angeles. | [Government open data portal.](https://catalog.data.gov/dataset/monthly-dockless-vehicles-trip-start-by-census-block-group) |
| cali\_nhood\_safegraph.csv | safegraph.r | Footfall counts in California during June 2020 at census block group level. | [Safegraph website](https://docs.safegraph.com/docs/neighborhood-patterns-2020?mkt_tok=eyJpIjoiTm1KaU5qVmlPRGRqT1RZeCIsInQiOiIySVJhYmYxbFF3TlpJYlArb1wvN0RcL2F3ZHNZc001SDA5WjZcL25JaVRPb0lEcDdIUVhjN1UrbGo3RU1iakYwWFpXdkVBdDRQZ1FqUzFHbkRlcGU2cjhkTXQ2c1FYdGpSYUdZQ05jVmQ2MDh2NUNmank1VlFwcDd2NFg4d0ZWUXhwSiJ9). |

**Data details**

***LA crime data -*** la\_crime\_june.csv

The original data was downloaded manually from the [Los Angeles open data archives](https://data.lacity.org/Public-Safety/Crime-Data-from-2020-to-Present/2nrs-mtv8/data) which contains data for 2020. The R script cleans the variable names and subsets for June 2020 to match the data from Safegraph (see below). The data contains individual-level records (i.e. each row is one crime) for incidents of crime in the city of Los Angeles. We recommend you explore the website for details on the variables, but below is a summary of some variables that you might find particularly useful.

|  |  |
| --- | --- |
| **Variable name** | **Brief description** |
| dr\_no | Individual identification |
| date\_rptd | Date crime was reported |
| date\_occ | Date crime occurred |
| time\_occ | Time crime occurred |
| crm\_cd\_desc | Crime type |
| location | Street address location |
| lat | Latitude coordinate |
| lon | Longitude coordinate |

***United States crime data -*** us\_crime\_2018.csv

Police-recorded crime data for multiple major cities is available through an R package *crimedata* (Ashby. 2018). We have specifically chosen the year 2018 to match the time period of the Retrosheet Game Logs (see below). The R script simply pulls this data using the package and saves the subset, so feel free to edit this code to get more (or less) data. If you are new to R or have not used the *crimedata* package before, we recommend the online [vignette](https://cran.r-project.org/web/packages/crimedata/vignettes/introduction.html) (and feel free to ask us for help!). There is also a [pre-print article](https://osf.io/preprints/socarxiv/9y7qz/) which provides specifics on the package and open crime data more generally. Variable names are intuitive but a full description can be found in the [package documentation](https://cran.r-project.org/web/packages/crimedata/crimedata.pdf) on pages 5-6.

***Retrosheet -*** retrosheet\_gl\_2018.csv

Crowding is difficult to measure, but certain events (e.g. concerts, sport games) provide insight into how many people were in a specific place at a specific time. The [Retrosheet website](https://www.retrosheet.org/gamelogs/index.html) contains information on historical Major League Baseball games in the United States, including where the game took place, on what date, and how many people attended. The dataset we provide is a subset and cleaned version of a Game Log for the 2018 season. Please take a look at the R script if you are interested in what we did. Variable names and descriptions are summarised below. Note that coordinate locations of stadiums are also contained in a separate dataset (stadium\_coords.csv).

|  |  |
| --- | --- |
| **Variable** | **Description** |
| date | Date of game |
| away\_team | Away team |
| home\_team | Home team |
| day\_night | Day or night game |
| completed | Details if game completed at later date |
| pitch | Pitch ID |
| attendance | Attendance |
| game\_time | Game time in minutes |

***Stadium coordinates*** - stadium\_coords.csv

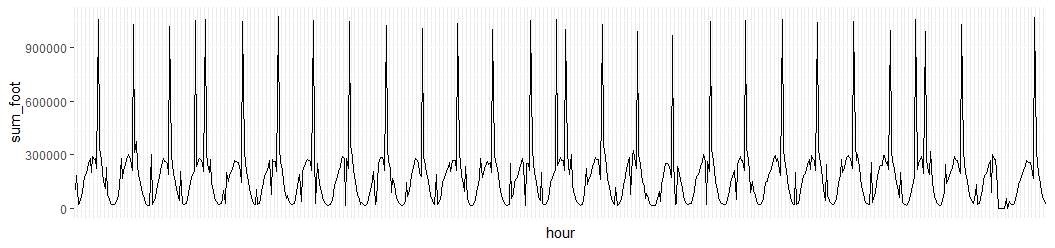
A dataset containing the coordinate locations of Major League Baseball stadiums from cities in the 2018 crime data. You might want to use these locations to capture crime occurring near the stadium (e.g. all crimes within a 1-mile buffer around the stadium) on game days in the Retrosheet game logs from 2018.

***Safegraph -*** cali\_nhood\_safegraph.csv

Safegraph provide spatially sensitive data on Points of Interest (POI), building footprints and footfall counts, amongst other things. Some of this data is available for free. We have provided their [publicly available data](https://docs.safegraph.com/docs/neighborhood-patterns-2020?mkt_tok=eyJpIjoiTm1KaU5qVmlPRGRqT1RZeCIsInQiOiIySVJhYmYxbFF3TlpJYlArb1wvN0RcL2F3ZHNZc001SDA5WjZcL25JaVRPb0lEcDdIUVhjN1UrbGo3RU1iakYwWFpXdkVBdDRQZ1FqUzFHbkRlcGU2cjhkTXQ2c1FYdGpSYUdZQ05jVmQ2MDh2NUNmank1VlFwcDd2NFg4d0ZWUXhwSiJ9) on footfall counts (measured by pings from anonymous mobile phones) aggregated to census block units in June, 2020. It is a subset of the raw data (which is very large) to only include California. You can match this information with the crime data for Los Angeles by using la\_cb.csv which contains a list of the census block groups in the city.

There is a detailed account of this data on the [Safegraph website](https://docs.safegraph.com/docs/neighborhood-patterns-2020?mkt_tok=eyJpIjoiTm1KaU5qVmlPRGRqT1RZeCIsInQiOiIySVJhYmYxbFF3TlpJYlArb1wvN0RcL2F3ZHNZc001SDA5WjZcL25JaVRPb0lEcDdIUVhjN1UrbGo3RU1iakYwWFpXdkVBdDRQZ1FqUzFHbkRlcGU2cjhkTXQ2c1FYdGpSYUdZQ05jVmQ2MDh2NUNmank1VlFwcDd2NFg4d0ZWUXhwSiJ9) and in an [open Google doc](https://docs.google.com/document/d/1FWYfnjfd4yY8Ip2MPe7Q815Gcu5oGCnRh-UZv8BlwDU/edit). However, it is worth highlighting a few things about this dataset. Firstly, it has a slightly odd structure. Each row is a census block group. Important variables on footfall counts in each of these blocks often consist of lists of numbers which correspond to different time periods. For example, the variable ‘stops\_by\_each\_hour’ contains a list of 720 numbers (footfall counts) per cell: one for each hour of the month in chronological order (30 days in June multiplied by 24 hours in a day). It might take some creative data handling to extract these counts into a useful form!

Another thing is known issues around how Safegraph counts mobile phone pings. This may result in some abnormal patterns that you might want to remove, or replace with an estimate. For example, the figure below visualises hour-by-hour summed footfall counts for the whole of California. We can clearly see the daily rhythms of the state, but there are anomalies (often at the same time each day).



***California census block boundaries* -** california\_cb\_shp.zip

This data contains the shapefiles for census blocks in California. It might be a useful way to look at the spatial patterning of footfall counts in the Safegraph data. You can subset these boundaries to include Los Angles in isolation using la\_cb.csv which contains a list of census blocks in the city.

**References**

Ashby, M P J. (2018, August 12). Studying crime and place with the Crime Open Database. DOI: http://doi.org/10.31235/osf.io/9y7qz