# Data Sets and Formats for the Assignment

Jeremy Ellman 17/12/20 V5

There are several datasets available for this assignment. These have been installed on AWS S3. Further data is available on data.gov.uk should you wish more detail.

The data sets are:

* all\_crimes18\_hdr.txt.gz (14GB Compressed, 43x10^6 records)
* LSOA\_pop\_v2.csv (2.4MB uncompressed)
* postcodes.gz (0.6GB Compressed)
* posttrans.csv (23.5 kB uncompressed)

All the files are csv format and may be compressed with gzip. Spark natively understands this compression format, so you may use the files just as CSV files.

### Location measurement

In these datasets location is specified in several ways:

1. ‘crime data’ uses ‘anonymized’ longitude and latitude AND LSOA code,
2. The LSOA dataset uses LSOA code.
3. Postcodes uses area centred longitude and latitude, postcode, LSOA code, Map Grid reference and many other measures

When considering which to use, you should bear in mind the level of detail needed:

1. The LSOA covers a mean of 1500 people. This means that LSOA will not have the same level of accuracy as the other measures. (but will be easier to handle)
2. Longitude and Latitude are accurate to 2-3m. you will need to convert these into postocdes.
3. A full postcode (e.g NE1 8ST, NE2 1XE) corresponds to approximately 6 households. You can generate a larger area by using the summing data over the first part (NE1, NE2)

### The Crimes Data

all\_crimes18\_hdr.txt.gz contains about 43million reported and logged crimes from 2010-2017. The data were downloaded from <https://data.police.uk/data/>. This site offers data by month, and by force. Consequently, they have been merged into one file for this assignment.

You can find out more about the data here (<https://data.police.uk/about/#columns>). Only ‘street’ files have been included. Outcomes are included.

The header row of the crimes data is:

**'Crime ID', 'Month', 'Reported by', 'Falls within', 'Longitude', 'Latitude', 'Location', 'LSOA code', 'LSOA name', 'Crime type', 'Last outcome category'**

Note that Longitude and Latitude are anonymized as described on the police web site here: <https://data.police.uk/about/#location-anonymisation>. Since the police use around 750,000 'anonymous' map points it is unlikely that these coincide with the longitudes and latitudes given in the postcode dataset. **For this reason, it is best to use LSOA (Lower Layer Super Output Area, UK Office for National Statistics )as a region indication**.

The file posttrans.csv will allow the translation of crimes’ longitude and latitude into actual postcodes.

The URL for the dataset is:

https://s3.amazonaws.com/kf7032-20.northumbria.ac.uk/all\_crimes18\_hdr.txt.gz

### Location Data.

The headers of the LSOA\_pop\_v2.csv. file are:

**"date","geography","geography code","Rural Urban","Variable: All usual residents; measures: Value","Variable: Males; measures: Value","Variable: Females; measures: Value","Variable: Lives in a household; measures: Value","Variable: Lives in a communal establishment; measures: Value","Variable: Schoolchild or full-time student aged 4 and over at their non term-time address; measures: Value","Variable: Area (Hectares); measures: Value","Variable: Density (number of persons per hectare); measures: Value"**

### Postcodes Data.

<https://www.freemaptools.com/download-uk-postcode-lat-lng.htm>

The headers of the postcodes.gz file are:

**'Postcode','InUse?','Latitude','Longitude','Easting','Northing','GridRef','County','District','Ward','DistrictCode','WardCode','Country','CountyCode','Constituency','Introduced','Terminated','Parish','NationalPark','Population','Households','BuiltUpArea','Builtupsubdivision','Lowerlayersuperoutputarea','Rural/urban','Region','Altitude'**

https://s3.amazonaws.com/kf7032-20.northumbria.ac.uk/postcodes.gz

### Posttrtans Data.

The headers of the posttrans.csv file are:

**Postcode,Lon,Lat**

https://s3.amazonaws.com/kf7032-20.northumbria.ac.uk/posttrans.csv