Sexual Offences in Greater London: How could UK Police Resources be better placed

Laura Foulquier

Home Office

Key Findings

- The areas within Greater London where sexual offences are most prevalent are tourist zones, like Westminster borough and City of London. The mean number of assaults per head in one of this area is approaching 0.47 crimes per head by example.
- In all but one of the areas studied with the highest prevalence of sexual offences, gender distribution shows an average ratio of roughly 60% of males. This suggests that fewer females per males could lead to more sexual assaults.
- Surprisingly, there is no correlation between the age distribution and the amount of sexual offences.
- The occurrence of sexual offences does not depend on the season: in 2014 and 2015, the attacks are equally spread within the year.

Agenda

- Data choices, assumptions and profiling
- Structure Query Language (SQL) operations
- Analysis, Results and Recommendations (Tableau)
- References

Data

Data choices, assumptions and profiling

Home Office

Data Choices

- Police and Population Data
  - Police data is available from December 2010 until September 2016 *(ref 1)*
  - LSOA (Lower Super Output Area) population density from the Office of National Statistics is available from mid-2011 until mid 2015 *(ref 2)*
- Therefore this study focuses on a data range from **January 2011 until December 2015,** with the assumption that the population recorded on June of each year is valid for the entire year
- The police data used is a representative sample from crime files.
- The assessment focusses on the following police forces in the London area:
  - City of London Police
  - Metropolitan Police
  - Surrey Police

How to Join the Datasets

- From the police dataset, the following parameters are used to determine event location:
  - Longitude, Latitude,
  - Location,
  - LSOA code and name
- From the population files, the location is derived from:
  - LSOA code
  - Name
- The location from the two datasets can therefore be cross-referenced **using LSOA codes**

NB: The population files only cover England and Wales, not Scotland nor Northern Ireland

Data Profiling

- Using Data Profile Viewer was used to profile our data: identify null values, get maximum and minimum length, profile value distribution and analyse pattern. Below are some findings for both datasets
  - Police Data
    - The null values are given by empty cells (example in Latitude and Longitude below)

Length	Count	Percentage
0	51158	0.8875 %
8	972685	16.8743 %
9	4740445	82.2382 %

    - All crimes are defined by a crime type
  - Population Data
    - This dataset is very clean, no null values were identified
    - Due to the dataset layout, it is very difficult to get statistics out of it

SQL

Data Loading and Cleaning, Table formatting, Normalization,

Home Office

Loading and Cleaning Data in SQL

Data Loading

- Considering the amount of files to import into SQL (60 folders, containing each 3 datasets for police data only), an ETL (Extract Transform Load) process was built using Visual Studio 2013
- A new table called dbo.RawCrime was created where all the police data was inserted, using an iterative process (for each loop). The data access mode was set as “fast load” (data loaded in 3 minutes instead of 45 !)
- The same iterative process was used with the population data. The data was loaded into SQL under the table name dbo.RawPopulation

Data Cleaning

- The first step was to remove the rows with empty cells or update the empty cells with null values
- Then the tables were stripped by keeping only the LSOA of interest (Greater London only). By reducing the table size, queries were executed faster

Formatting Table dbo.RawPopulation

- Our table dbo.RawPopulation had a lot of information, but its layout was not designed for efficient queries, as illustrated below:

[Age between 12 and 79, hidden for display purpose]

Results | Messages

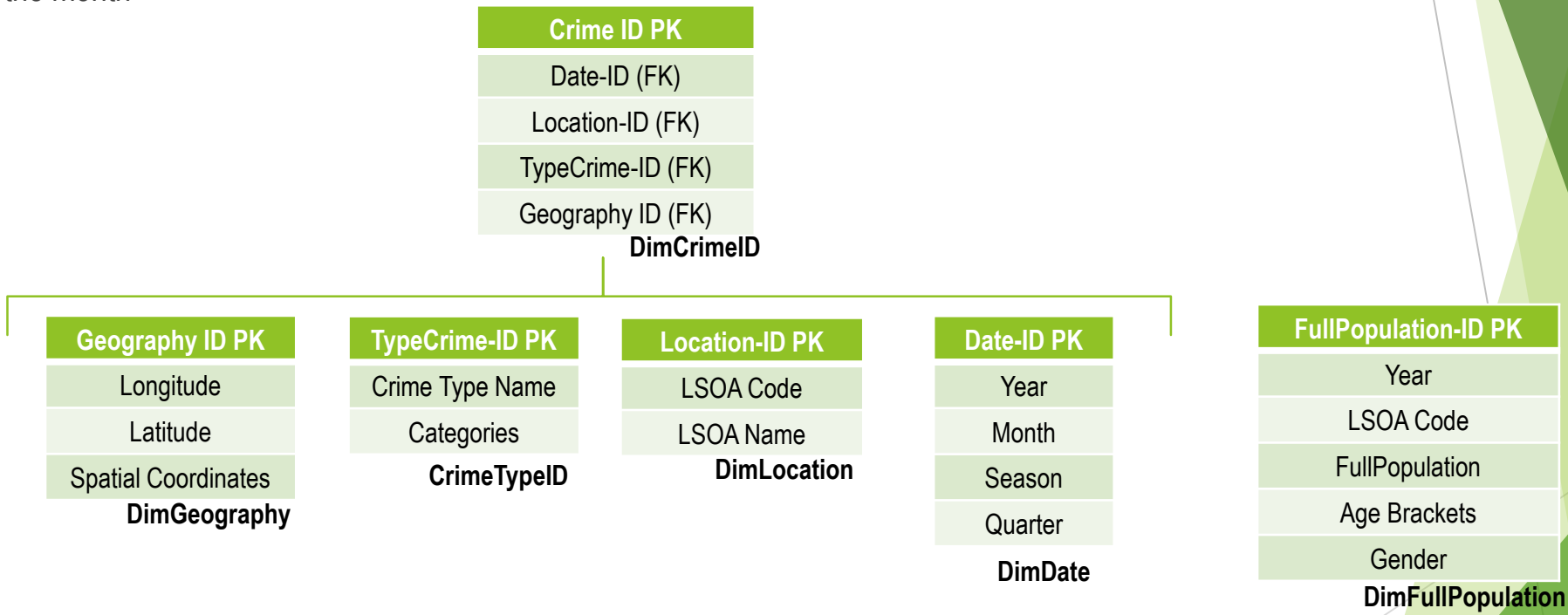
	Area Codes	Area Names	All Ages	0	1	2	3	4	5	6	7	8	9	10	11	80	81	82	83	84	85	86	87	88	89	90+	Gender	Year
1	E01004763	Westminster 013B	935	3	5	4	7	5	7	5	3	1	4	1	1	0	3	2	1	1	0	2	0	0	2	0	F	2014
2	E01033595	Westminster 013E	493	3	8	0	5	4	0	0	3	4	0	6	4	3	3	2	1	1	0	0	0	0	0	2	F	2014
3	E01004734	Westminster 018A	699	6	5	4	5	5	4	7	2	5	6	1	0	4	8	4	7	3	3	2	0	2	0	5	F	2014
4	E01004735	Westminster 018B	517	5	0	9	8	4	3	1	4	5	5	6	3	1	5	2	1	1	2	1	3	0	5	6	F	2014
5	E01004736	Westminster 018C	689	8	12	4	7	0	0	2	2	1	2	3	1	6	2	4	0	2	3	3	4	2	0	6	F	2014
6	E01001043	Croydon 027B	1513	26	36	34	42	26	19	24	17	14	16	19	12	2	3	0	0	3	0	1	0	1	0	4	M	2015
7	E01001043	Croydon 027B	1392	39	28	38	32	25	16	24	11	12	19	12	8	1	3	0	2	1	1	1	1	2	0	4	F	2014
8	E01033708	Hackney 027G	715	4	5	5	3	1	1	2	5	2	4	5	4	1	0	0	0	0	0	0	0	0	0	1	F	2015
9	E01004763	Westminster 013B	1448	4	6	2	0	4	3	5	2	1	1	1	2	1	1	2	1	2	0	0	1	0	0	2	M	2013
10	E01033595	Westminster 013E	725	4	5	10	1	1	0	1	0	0	1	1	2	2	3	2	2	2	3	2	2	1	1	2	M	2013
11	E01004734	Westminster 018A	1064	4	4	7	5	3	6	3	5	2	2	3	4	8	4	3	5	3	1	0	0	0	0	1	M	2013
12	E01004735	Westminster 018B	865	9	2	6	4	3	16	5	2	4	6	1	2	7	2	5	1	1	3	0	1	1	1	7	M	2013
13	E01004736	Westminster 018C	1150	3	2	5	4	1	4	1	7	3	1	2	2	4	2	2	6	7	2	1	2	1	1	2	M	2013
14	E01004763	Westminster 013B	1520	3	5	2	4	1	3	3	6	0	1	2	0	0	1	0	1	0	1	0	0	1	0	2	M	2014

- In order to get a more comprehensive table, Area Codes, Area Names, All Ages were kept as they are. The ages however were bracketed in range of 4 years (0-4 , 5-9, 10-14, etc). The new age columns were then pivoted, and the final table **dim.FullPopulation** looks as follow:

	Area Codes	Area Names	All Ages	Gender	Year	Population	Age Bracket
1	E01004763	Westminster 013B	1635	M	2015	24	0-4
2	E01004763	Westminster 013B	1635	M	2015	17	5-9
3	E01004763	Westminster 013B	1635	M	2015	10	10-14
4	E01004763	Westminster 013B	1635	M	2015	28	15-19
5	E01004763	Westminster 013B	1635	M	2015	189	20-24
6	E01004763	Westminster 013B	1635	M	2015	286	25-29
7	E01004763	Westminster 013B	1635	M	2015	293	30-34
8	E01004763	Westminster 013B	1635	M	2015	188	35-39
9	E01004763	Westminster 013B	1635	M	2015	165	40-44
10	E01004763	Westminster 013B	1635	M	2015	148	45-49

Normalization

- Once the data was imported (population / year/ages / gender and crime/year), normalised tables were populated to reduce the size of our main crime table, following the normalization rules for relational database
- All but one primary key was generated using “identity” in SQL. The primary key in DimDate is a concatenation of year followed by the month



- In terms of memory, the main file is now **0.200 GB** (DimCrimeID), instead of **1.24** (RawCrime) previously. Hence it will be **quicker** to query
- The DimFullPopulation table was used with the DimCrimID table using “join” in SQL (joining on LSOA Code)

Examples of SQL Queries

- The following queries were used in SQL, each showing a specific feature:
  - How a view was created with a selection of LSOA
  - How a view was created with joining different normalised tables
  - How the table DimDate was created using “case … then”

```
/* Create view with age brackets, their percentage, per gender, per year, per restricted area names*/

create view vw_PercentagePerAgePerGender
as
select distinct
[Area Names]
,[Year]
,[Age Bracket]
,Gender
,PercentagePerAge

from DimFullPopulation

where 1 = 1
and [Area Names] in ('Westminster 018A','City of London 001F','Westminster 013E','Westminster 013B',
'Westminster 018C','Hackney 027G','Westminster 018B','Camden 021A','Croydon 027B')
and [Year] in ('2013','2014','2015')
```

```
/* Create View for the crime occurrence by year and categories*/

create view  vw_CrimeCategoriesOccurrenceperYear
as
select
ct.Categories
,d.[year]
,count(*) [Crime Occurence]
from DimCrimeID c

join DimLocation l
on l.[Location ID]=c.LocationID

join DimDate d
on d.[Date ID]=c.DateID

join CrimeTypeID ct
on ct.CrimeTypeID=c.TypeCrimeID

where 1 = 1
group by ct.Categories,d.[year]
```

```
/* Create Normalized Table DimDate */

create table DimDate
(
[Date ID] varchar(10) not null primary key,
[Year] varchar(10),
[Month] varchar(10),
[Season] varchar(10),
[Quarter] varchar(10)
)
GO

insert into DimDate
([Date ID], [Year], [Month], [Season], [Quarter])
Select
[Month]
,left([Month], 4)
,right([Month], 2)
,case when right([Month], 2) in ('12', '01', '02') then 'Winter'
when right([Month], 2) in ('03', '04', '05') then 'Spring'
when right([Month], 2) in ('06', '07', '08') then 'Summer'
when right([Month], 2) in ('09', '10', '11') then 'Autumn'
else right([Month], 2)
end
,case when right([Month], 2) in ('01', '02', '03') then 'Q1'
when right([Month], 2) in ('04', '05', '06') then 'Q2'
when right([Month], 2) in ('07', '08', '09') then 'Q3'
when right([Month], 2) in ('10', '11', '12') then 'Q4'
else right([Month], 2)
end
from dbo.RawCrime
group by [Month]
```

SQL Project

NB : Please note that the SQL project is accessible on Git

Analysis – See Tableau workbook / pdf file

Sexual Assaults in Greater London for 2013, 2014 and 2015

Home Office

Results on Tableau Workbook

Note: If you open the Tableau Workbook, you can choose to display the different results by year.

Links / References

1. Crime Dataset

https://data.police.uk/

2. Population Dataset

https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/datasets/lowersuperoutputareapopulationdensity

3. Sex ratio

https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/200527/Gender_birth_ratio_in_the_UK

4. Mean and Median Salary

https://data.london.gov.uk/dataset/household-income-estimates-small-areas

5. Average salary London

http://metro.co.uk/2015/11/18/how-does-your-salary-compare-with-the-rest-of-the-country-5511194/

6. RAINN

https://www.rainn.org/index.php