Crimes in the UK
An awesome Data Mining Project

1. Phase 1: Crimes Data only

In the first phase of our project, we focused on the crimes data. …

1.1. Original Data Sets

1.1.1. data.police.uk

Our main data set is provided by data.police.uk. It is structured in 70 folders. Each folder represents the crime data for a single month of the respective year, broken down by territorial police force. The source files are stored as .CSV and have 13.2 GB. There are three types of data: Street-level crimes (called "*streets*"), street-level outcomes (called "*outcomes*") and *stop and search* information.

The Street-level crimes data ranges from december 2010 till September 2016, the street-level outcomes data ranges from January 2012 till September 2016, and the stop and search data ranges from December 2014 till September 2016.

For privacy reasons, the data is being anonymized. For this, the location points of the data have been mapped to more than 750.000 map points all around UK. The generated data points represent some Points Of Interest (POI). If the nearest point is more than 20km away, the coordinates of the crimes are being removed. Also other fields like id or date have been anonymized.

Cleanings

In outcomes, there are 2547177 "crime_id"s that exist more than one time. In streets, there are 145124 such cases. The tuples with the same "crime_id" are very similar, so that we assume that they concern the same crime and that they accrued because of updates. Therefore, we removed the older tuples. In total, we removed 656261 tuples in streets and 3405907 tuples in outcomes.

We also removed all tuples before September 2011 because in September 2011, a lot of new types of crimes were introduced. We deleted 5069641 tuples in streets.

We dropped the column "reported_by" in outcomes, because it has always the same value like "falls_within". Moreover, we dropped "context", because the values are not very meaningful and is null in 99.64% of the cases.

Streets and outcomes have beside of the coordinates-columns also a column named "location", that contains a small description of a place. In some cases, the same place have different location values, because some values are different spelled (e.g. "On or near Moretons Lane" and "On or near Moreton'S Lane"). We corrected the differences.

In the data sets from September 2013, "lsoa_code" is sometimes missing although a corresponding lsoa does exist and is given in other tuples with the same coordinates. We corrected it.

Data Attributes and Statistics

After cleanings, we have the following attributes and number of tuples.

Streets

The streets relation has 29289398 tuples. The table below describes each attribute in the "streets" files of the dataset.

Attribute name	Missing values	Number of distinct values	Example value	Comment
crime_id	11758313 (40.1%)	17531085	42a1d4194f0a7d0f7f6e0d4e709c7f2d68a321e2b28f9a8fc5203c329bbdc5a4	
year_month	0 (0%)	61	2015-09	
reported_by	0 (0%)	45	Hertfordshire Constabulary	Except 82506 cases, it is the same like falls_within.
falls_within	0 (0%)	45	Hertfordshire Constabulary	
longitude	355661 (1.2%)	749319	-0.57560600	
latitude	355661 (1.2%)	720917	51.76402600	
location	355661 (1.2%)	275005	On or near BECKETTS	
lsoa_code	699241 (2.4%)	34750	E01023370	LSOA are only given for crimes in England and Wales. Scotland and Ireland have similar geographical areas[1], but they are not given by this data set.

---

[1]   http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/beginner-s-guide/census/super-output-areas--soas-/index.html

lsoa_name	699241 (2.4%)	34750	Dacorum 012B		
crime_type	0 (0%)	16	Anti-social behaviour		
last_outcome _category	1282379 (43.8%)	25	Unable to prosecute suspect		
police_force	0 (0%)	45	hertfordshire		A variant spelling of falls_within, that is used for joining with stop and search.

Outcomes

The outcomes relation has 14218059 tuples. The table below describes each attribute in the "outcomes" files of the dataset

Attribute name	Missing values	Number of distinct values	Example value	Comment
crime_id	0 (0%)	14218059	42a1d4194f0a7d0f7f6e0d4e709c7f2d68a321e2b28f9a8fc5203c329bbdc5a4	
year_month	0 (0%)	57	2015-09	
police_force	0 (0%)	43	hertfordshire	See above
falls_within	0 (0%)	43	Hertfordshire Constabulary	
longitude	257468 (1.8%)	648721	-0.57560600	
latitude	257468 (1.8%)	630170	51.76402600	
location	257468 (1.8%)	253550	On or near BECKETTS	
lsoa_code	257924 (1.8%)	34749	E01023370	See above
lsoa_name	257924 (1.8%)	34749	Dacorum 012B	

outcome_type	0 (0%)	23	Offender given a caution	

Stop and Search

The stop and search relation has 584841 tuples. The table below describes each attribute in the stop and search files of the dataset

Attribute name	Missing values	Number of distinct values	Example value	Comment
type	0 (0%)	3	Person search	
year_month_day	0 (0%)	278644	2015-09-01	
part_of_a_policing _operation	410668 (70.2%)	2	False	
policing_operation	569290 (97.3%)	1316	OPERATION BLACKJACK	
latitude	274057 (46.8%)	97563	53.73099100	
longitude	274057 (46.8%)	98148	-1.75077100	
gender	57273 (9.7%)	3	Male	
age_range	33856 (5.7%)	5	18-24	
self_defined _ethnicity	8979 (1.5%)	17	White - White British (W1)	
officer_defined _ethnicity	48310 (8.2%)	5	White	
legislation	47621 (8.1%)	14	Police and Criminal Evidence Act 1984 (section 1)	
object_of_search	113984 (19.4%)	14	Article for use in theft	
outcome	17957 (3%)	14	Nothing found - no further action	
outcome_linked_to _object_of_search	445630 (76.1%)	2	True	
removal_of_more _than _just_outer_clothing	426419 (72.9%)	2	False	
police_force	0 (0%)	44	west yorkshire	

1.1.2. Points of Interest

As an enrichment to the crimes data, we used a data set[2] that provides around 191,000 POI all over the UK, broken down by the different categories listed below.

- Wifi hotspot
- Fuel: Gas station
- Shopping consumer electronics: electrical retailer selling home electronics and household appliances
- Transport: airports and train stations
- Accommodation: Hotels
- Entertainment: Casinos, Cinemas
- Banks
- Food and drink: restaurants, dining pups
- Government agency: Tax office, passport office
- Public service buildings: Police station, fire station
- Landmark: Mountains, hills, waterfalls
- Attraction: Aquarium, museums
- Book shop
- Special Interest: Madame Tussaud, Church bell
- Community: churches, Universities
- Courier: DHL, Express
- Healthcare: Hospital, practice
- Sports center: sports clubs
- Pharmacy
- Shopping supermarkets: Lidl, Aldi
- Association
- Holiday parks
- ATM

1.1.3. Geographical Data

In addition to the coordinates, we also had the Lower Layer Super Output Areas (LSOA) code of the area where a crime happened. LSOAs have an average population size of around 1,500 citizens. This low granularity was i) too small to be visualized on a map in an informative manner ii) not enough to discover location-based patterns. Therefore, we integrated more geographical data[3]. This allowed us to come up with the following geographical structure and therefore derive new features about the crime locations:

Area Type	Description
Latitude/Longitude	The anonymised coordinates. (~800.000 values)
Lower Layer Super Output Area (LSOA)	Small areas defined by the Office for National Statistics with about 1,500 citizens. (~35,000 values)
Middle Layer Super Output Areas (MSOA)	Like LSOA but with about 7,200 citizens. (~7,200 values)
Postcode Area	There are 121 postcode areas in the UK. We have data

[2]   http://www.pocketgpsworld.com/modules.php?name=POIs
[3]   https://data.gov.uk/dataset/national-statistics-postcode-lookup-uk/resource/3206f3b9-854a-46ec-8fd2-a6823e836b65

	about crimes in 105 areas.
Region	England can be divided into 9 regions. Furthermore, we also consider Wales, Scotland and Northern Ireland as regions.

1.2. Combining Data Sets

1.2.1. Crime-centric

Using the three data sets, we created two relations. The first one is crime-centric and called *crimes*.

In order to detect the nearest stop and search incidents and POIs within the range of 500m from crimes, we used PostGIS, a spatial database extender for PostgreSQL. Because of performance reasons, we used as distance metric the L-infinity norm, so we could build indices with the coordinates of Stop-and-Search and Points-of-Interest points. Afterwards, we were able to perform a fast search of surrounding points for every crime entry - which still took many hours on our local machines.

The *crimes* relation has the following attributes:

- id
- crime_id
- year_month
- reported_by
- falls_within
- latitude
- longitude
- location
- lsoa_code
- lsoa_name
- msoa_code
- msoa_name
- local_authority_code
- local_authority_name
- postcode_area
- region_code
- region_name
- **crime_type**
- last_outcome_category
- suspect_identified
- police_force
- outcome_type
- suspect_identified_outcome
- number_500m_ss
- number_500m_ss_person_searches
- number_500m_ss_vehicle_searches
- number_500m_ss_asian_searches

- number_500m_ss_black_searches
- number_500m_ss_white_searches
- number_500m_ss_outcome_suspect_arrested
- number_500m_ss_outcome_article_found
- number_500m_ss_outcome_offender_given_drugs_p_warning
- number_500m_ss_outcome_offender_cautioned
- number_500m_ss_outcome_suspect_summonsed_to_court
- number_500m_ss_outcome_nothing_found
- number_500m_ss_outcome_offender_given_penalty_notice
- number_500m_ss_outcome_local_resolution
- number_500m_ss_age_under_10
- number_500m_ss_age_10_17
- number_500m_ss_age_18_24
- number_500m_ss_age_25_34
- number_500m_ss_age_over_34
- number_500m_ss_object_controlled_drugs
- number_500m_ss_object_criminal_demage_articles
- number_500m_ss_object_stolen_goods
- number_500m_ss_object_weapons
- number_500m_ss_object_firearms
- number_500m_ss_object_evidence
- number_500m_poi
- number_500m_poi_community
- number_500m_poi_food_and_drink
- number_500m_poi_fuel
- number_500m_poi_atm
- number_500m_poi_transport
- number_500m_poi_motorcycle_dealership
- number_500m_poi_wifi_hotspot
- number_500m_poi_shopping_supermarkets
- number_500m_poi_special_interest
- number_500m_poi_shopping_others
- number_500m_poi_banks
- number_500m_poi_sports_center
- number_500m_poi_car_dealership
- number_500m_poi_commercial_site
- number_500m_poi_association
- number_500m_poi_car_parts_supplier
- number_500m_poi_builder_and_diy_supplies
- number_500m_poi_accommodation
- number_500m_poi_attraction
- number_500m_poi_pharmacy
- number_500m_poi_healthcare
- number_500m_poi_landmark
- number_500m_poi_public_service_buildings
- number_500m_poi_shopping_consumer_electrics
- number_500m_poi_entertainment
- number_500m_poi_car_servicing

- number_500m_poi_building_society
- number_500m_poi_bookshop
- number_500m_poi_truck_dealership
- number_500m_poi_holiday_park
- number_500m_poi_government_agency
- number_500m_poi_vehicle_hire
- number_500m_poi_car_auction
- number_500m_poi_caravan_dealership
- number_500m_poi_courier
- number_500m_poi_estate_agent

1.2.2. Location-centric

Our second approach is location-centric and called locations. The relation *locations* has latitude/longitude entries as rows. Therefore one row represents crimes data about one location. The table could be used to analyze crime trends and anomalies in different type of locations.

The locations table has the following attributes:

- latitude
- longitude
- location
- lsoa_code
- lsoa_name
- msoa_code
- msoa_name
- local_authority_code
- local_authority_name
- postcode_area
- region_code
- region_name
- number_500m_ss
- number_500m_ss_person_searches
- number_500m_ss_vehicle_searches
- number_500m_ss_asian_searches
- number_500m_ss_black_searches
- number_500m_ss_white_searches
- number_500m_ss_outcome_suspect_arrested
- number_500m_ss_outcome_article_found
- number_500m_ss_outcome_offender_given_drugs_p_warning
- number_500m_ss_outcome_offender_cautioned
- number_500m_ss_outcome_suspect_summonsed_to_court
- number_500m_ss_outcome_nothing_found
- number_500m_ss_outcome_offender_given_penalty_notice
- number_500m_ss_outcome_local_resolution
- number_500m_ss_age_under_10
- number_500m_ss_age_10_17
- number_500m_ss_age_18_24
- number_500m_ss_age_25_34

- number_500m_ss_age_over_34
- number_500m_ss_object_controlled_drugs
- number_500m_ss_object_criminal_demage_articles
- number_500m_ss_object_stolen_goods
- number_500m_ss_object_weapons
- number_500m_ss_object_firearms
- number_500m_ss_object_evidence
- number_500m_poi
- number_500m_poi_community
- number_500m_poi_food_and_drink
- number_500m_poi_fuel
- number_500m_poi_atm
- number_500m_poi_transport
- number_500m_poi_motorcycle_dealership
- number_500m_poi_wifi_hotspot
- number_500m_poi_shopping_supermarkets
- number_500m_poi_special_interest
- number_500m_poi_shopping_others
- number_500m_poi_banks
- number_500m_poi_sports_center
- number_500m_poi_car_dealership
- number_500m_poi_commercial_site
- number_500m_poi_association
- number_500m_poi_car_parts_supplier
- number_500m_poi_builder_and_diy_supplies
- number_500m_poi_accommodation
- number_500m_poi_attraction
- number_500m_poi_pharmacy
- number_500m_poi_healthcare
- number_500m_poi_landmark
- number_500m_poi_public_service_buildings
- number_500m_poi_shopping_consumer_electrics
- number_500m_poi_entertainment
- number_500m_poi_car_servicing
- number_500m_poi_building_society
- number_500m_poi_bookshop
- number_500m_poi_truck_dealership
- number_500m_poi_holiday_park
- number_500m_poi_government_agency
- number_500m_poi_vehicle_hire
- number_500m_poi_car_auction
- number_500m_poi_caravan_dealership
- number_500m_poi_courier
- number_500m_poi_estate_agent
- **number_crimes**
- **number_public_order**
- **number_anti_social**
- **number_bulgary**

- **number_other_crime**
- **number_violence_sexual_offences**
- **number_bicycle_theft**
- **number_drugs**
- **number_criminal_demage**
- **number_other_theft**
- **number_shoplifting**
- **number_vehicle_crime**
- **number_violent_crime**
- **number_public_disorder_weapon**
- **number_robbery**
- **number_theft_from_person**
- **number_possessions_weapons**

1.3. Data Availability

data.police.uk is a quite new platform, which composes data of several police forces all over the UK. These police forces entered the project at different times and use various systems so that the data from the different geographical areas are not available in the same volume.

1.3.1. Countries

Country	Number location points	Number crimes	1st crime entry	Number stop and searches	1st stop and search entry
England	732558	27121189	2011-09	297031	2014-03-31
Wales	44829	1470012	2011-09	11295	2014-09-30
Scotland	326	11370	2011-09	3	2016-02-08
Northern Ireland	41315	331166	2011-09	90	2014-12-08

From Scotland, we only have data from British Transport Police, that is responsible for the railways in Great Britain.

1.3.2. Police Forces

Police force	Number location points	Number crimes	Number stop and searches	1st stop and search entry
Avon and Somerset Constabulary	23406	823502	6125	2014-12-01
Bedfordshire Police	7491	313498	3203	2015-03-31
British Transport Police	2781	236302	5048	2015-04-01
Cambridgeshire Constabulary	11827	375010	2645	2015-07-31
Cheshire Constabulary	16767	460259	7288	2015-03-31

City of London Police	602	33953	2018	2015-03-02
Cleveland Police	8566	437957	4716	2014-12-01
Cumbria Constabulary	8030	232653	3657	2015-07-31
Derbyshire Constabulary	16502	510179	3918	2015-04-01
Devon & Cornwall Police	25332	668174	8103	2015-03-01
Dorset Police	10686	340816	9630	2014-12-01
Durham Constabulary	8761	319035	5427	2015-03-31
Dyfed-Powys Police	7957	195835	6218	2015-01-01
Essex Police	22331	803825	4518	2015-04-01
Gloucestershire Constabulary	9253	280632	2544	2015-04-01
Greater Manchester Police	40968	1671091	9102	2015-01-01
Gwent Police	8932	305207	1933	2014-09-30
Hampshire Constabulary	26882	887366	33886	2014-03-31
Hertfordshire Constabulary	14146	451034	11817	2015-03-01
Humberside Police	12804	443441	2868	2014-12-01
Kent Police	22655	810707	11512	2015-03-31
Lancashire Constabulary	24010	898381	7535	2015-03-31
Leicestershire Police	14239	418823	3268	2014-12-01
Lincolnshire Police	12424	304192	5954	2015-01-01
Merseyside Police	19064	777752	14676	2015-05-31
Metropolitan Police Service	93131	5001436	224533	2015-03-31
Norfolk Constabulary	15245	352927	8064	2015-01-01
North Wales Police	11516	317687	2951	2014-12-01
North Yorkshire Police	13704	354853	7943	2014-12-01
Northamptonshire Police	10862	396922	2598	2015-02-01
Northumbria Police	18599	803885	9726	2015-03-01
Nottinghamshire Police	16145	558848	4615	2014-12-01
Police Service of Northern Ireland	41307	331427	0	-
South Wales Police	17762	663060	9919	2015-03-31
South Yorkshire Police	18997	937120	5338	2015-03-01

| Staffordshire Police	16786	494211	6414	2015-01-01
Suffolk Constabulary	12109	320018	4333	2015-01-01
Surrey Police	16705	462186	10924	2014-12-01
Sussex Police	21708	734380	11829	2015-03-01
Thames Valley Police	30549	844882	16627	2015-03-31
Warwickshire Police	8197	245611	5184	2014-12-02
West Mercia Police	18856	579486	13652	2014-12-01
West Midlands Police	28219	1275494	31246	2014-12-01
West Yorkshire Police	35982	1330066	27540	2014-12-01
Wiltshire Police	10264	285275	3796	2015-01-01

We have data from all police forces since December 2010 except from Police Service of Northern Ireland. The first crime entries from them are from September 2011.
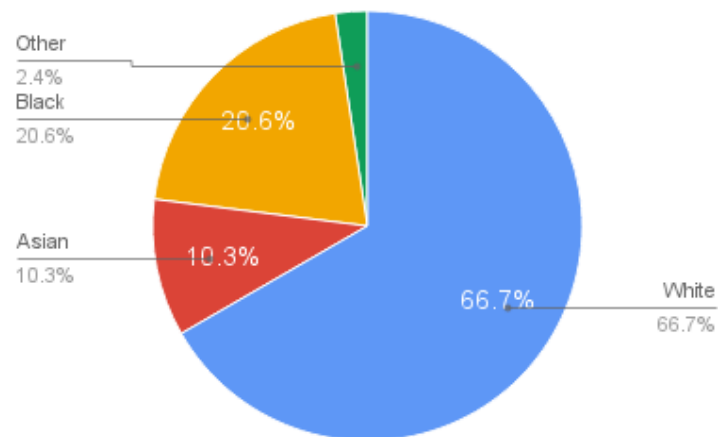
1.4. Stop and Searches
1.4.1. Simple Statistics

Age Range

The pie chart below describes the age range distribution in the stop and search table. Through this chart, it can be observed that there are far more people in the 18-24 age range that are being stopped and searched than in any other age range category. Except the 18-24 age range, the number of stop and searches by age range is similarly distributed.
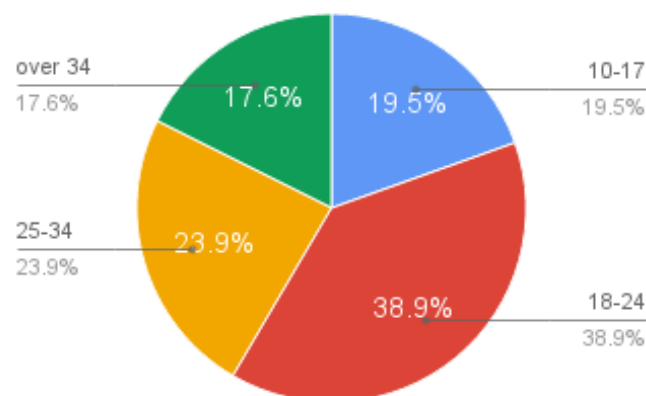

Ethnicity

The pie chart below describes the ethnicity distribution in the stop and search table. Through this chart, it can be observed that the most stopped and searched persons are white. However, these numbers do not express anything if not compared to the ethnicity distribution in the area where the stop and search has been made.


Gender

The pie chart below describes the gender distribution in the stop and search table. It is surprising to see that 9 out of 10 stopped and searched persons are males.
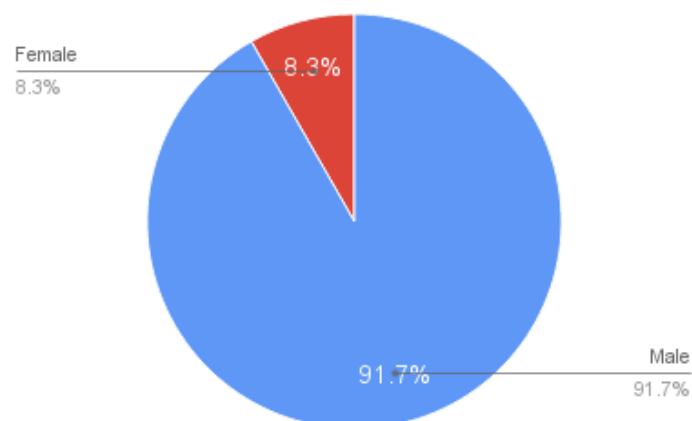
Over the Time

The time series graph below describes the amount of stop and searches that have been done until September 2016. It can be observed that the first year of this project is not very useful as data is missing. The second chart shows the days distribution on crimes. The crimes are similarly distributed among all week days.

Total number of stop and searches
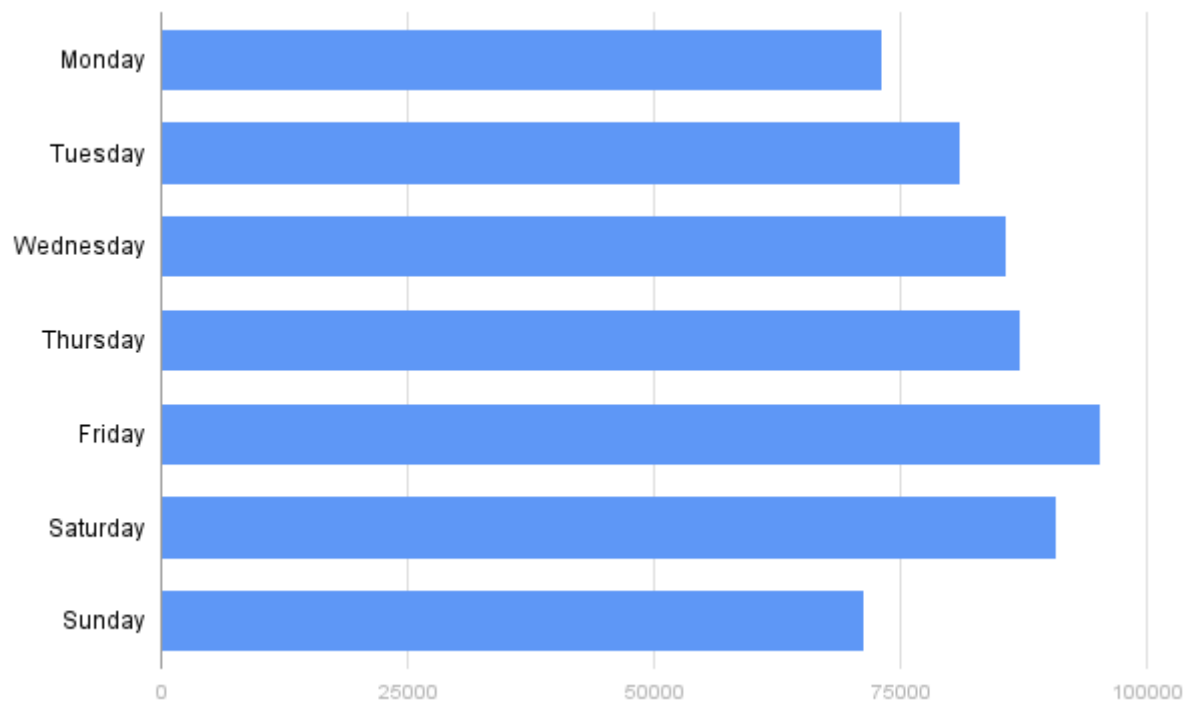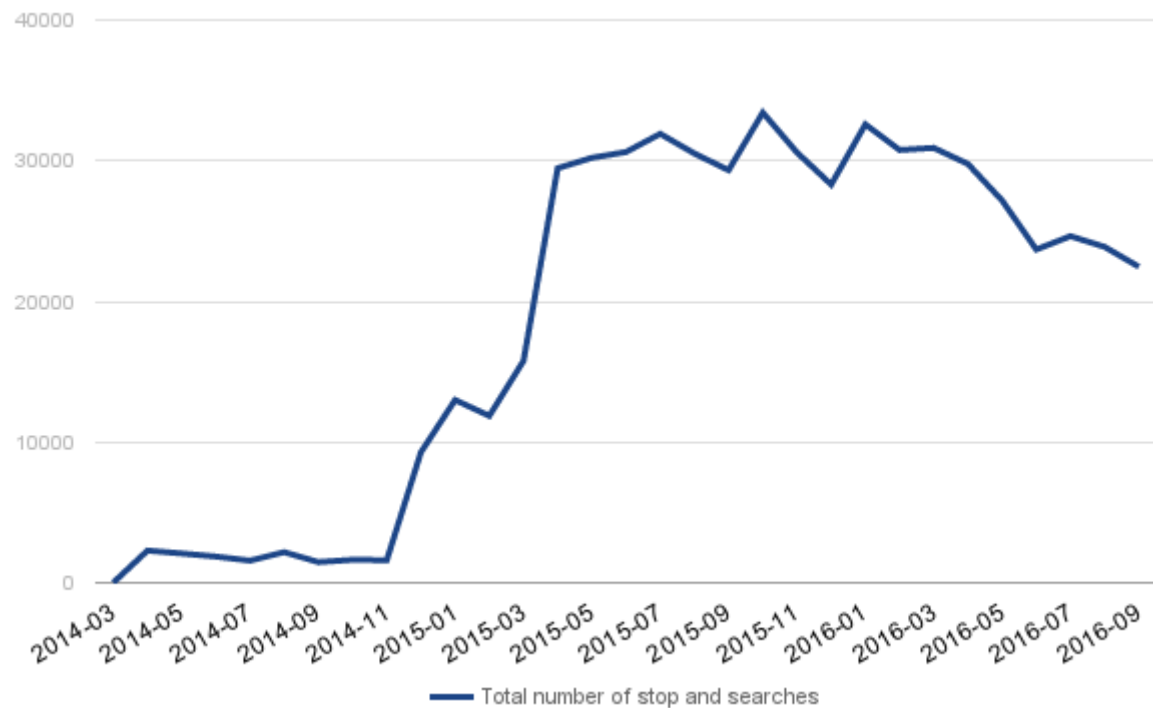
1.5. Crimes

1.5.1. Simple Statistics

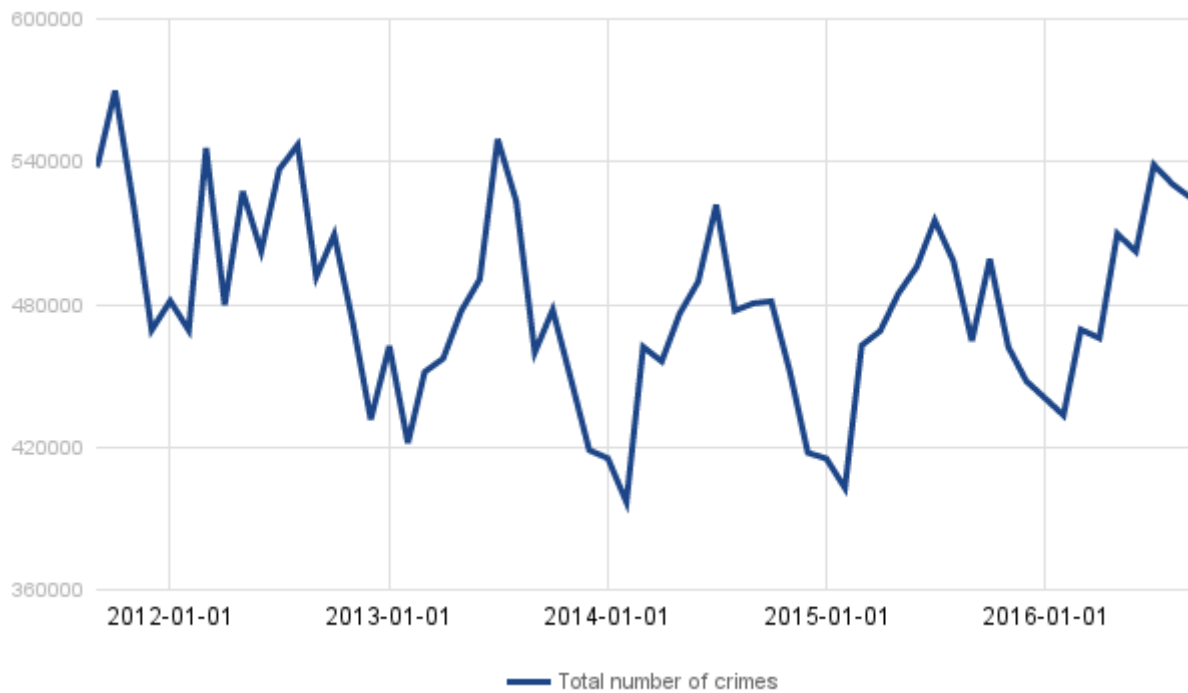The time series below depicts the total number of crimes over time. In this graph, it can be observed that during winter there are typically less crimes than in the rest of the year. Conversely, the highest number of crimes is always in the summer

1.5.2. Crime Types

The crimes are divided in 16 categories[4]. This section analyzes the crime types and their evolution. The table below contains for each crime type the total number of entries, the date of the first entry, and the date of the last entry.

Crime type	Number crimes	1st entry	Last (given) entry
Anti-social behaviour	10723897	2011-09-01	2016-09-01
Bicycle theft	315680	2013-05-01	2016-09-01
Burglary	2184796	2011-09-01	2016-09-01
Criminal damage and arson	2674455	2011-09-01	2016-09-01
Drugs	881852	2011-09-01	2016-09-01
Other crime	462642	2011-09-01	2016-09-01

---

[4]   https://www.police.uk/about-this-site/faqs/#what-do-the-crime-categories-mean

Other theft	2886095	2011-09-01	2016-09-01
Possession of weapons	80052	2013-05-01	2016-09-01
Public disorder and weapons	233368	2011-09-01	2013-04-01
Public order	603667	2013-05-01	2016-09-01
Robbery	281452	2011-09-01	2016-09-01
Shoplifting	1631288	2011-09-01	2016-09-01
Theft from the person	291188	2013-05-01	2016-09-01
Vehicle crime	1886783	2011-09-01	2016-09-01
Violence and sexual offences	3104835	2013-05-01	2016-09-01
Violent crime	1047348	2011-09-01	2013-04-01
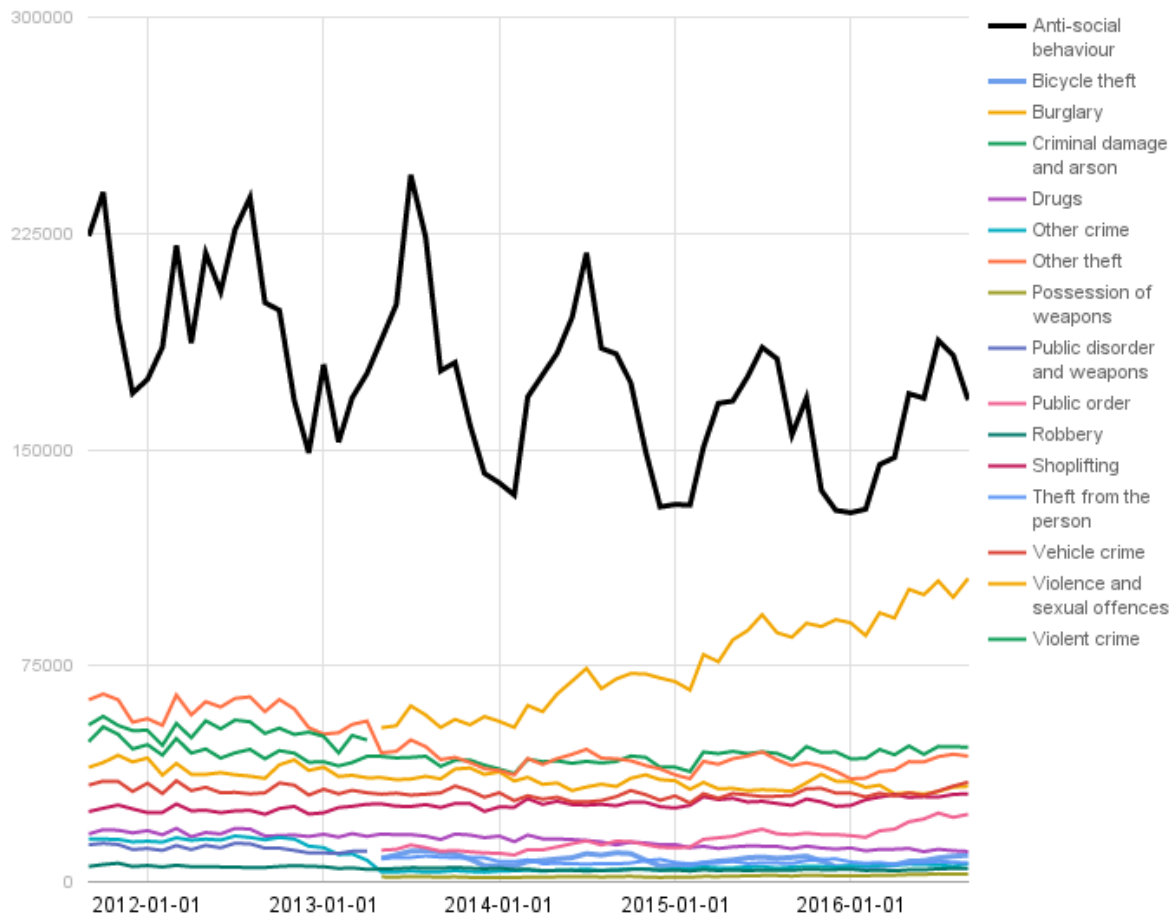
5

The time series above show the evolution of the number of crimes for each crime type. As it can be seen, the crime type "Public disorder and weapons" was splitted into "Possession of weapons" and "Public order" in April 2013. We want to have data for the whole period of

---

5   Used script: no. 2

time. Therefore we had to merge some of the columns. We decided to merge the splitted columns again to "Public disorder and weapons". Furthermore, the crime type "Violent crime" was removed by the police.uk since May 2013. Instead, "Violence and sexual offences" was introduced. Again, we decided to merge them and name them as "Violence and sexual offences".

After doing the cleanings described above, we have the following 13 crime types:

Crime type	Number crimes	1st entry	Last (given) entry
Anti-social behaviour	10723897	2011-09-01	2016-09-01
Bicycle theft	315680	2013-05-01	2016-09-01
Burglary	2184796	2011-09-01	2016-09-01
Criminal damage and arson	2674455	2011-09-01	2016-09-01
Drugs	881852	2011-09-01	2016-09-01
Other crime	462642	2011-09-01	2016-09-01
Other theft	2886095	2011-09-01	2016-09-01
Public disorder and weapons	917087	2011-09-01	2016-09-01
Robbery	281452	2011-09-01	2016-09-01
Shoplifting	1631288	2011-09-01	2016-09-01
Theft from the person	291188	2013-05-01	2016-09-01
Vehicle crime	1886783	2011-09-01	2016-09-01
Violence and sexual offences	4152183	2011-09-01	2016-09-01

1.5.3. Outcomes

Anti-social behavior does not appear in the outcomes table because of some difficulties that are described on the police.uk website[6]. There are 24 outcome types[7].

Outcome type	Number crimes
Action to be taken by another organisation	10283
Court case unable to proceed	98287
Defendant found not guilty	221406

---

[6]   https://www.police.uk/about-this-site/faqs/#why-does-police.uk-not-show-outcomes-for-every-crime

[7]   https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/217421/police-uk-statistics.pdf

Defendant sent to Crown Court	1597
Formal action is not in the public interest	93106
Further investigation is not in the public interest	2328
Investigation complete; no suspect identified	8815109
Local resolution	453389
Offender deprived of property	6480
Offender fined	162119
Offender given a caution	585943
Offender given a drugs possession warning	208388
Offender given absolute discharge	5283
Offender given community sentence	276617
Offender given conditional discharge	130177
Offender given penalty notice	155306
Offender given suspended prison sentence	115366
Offender ordered to pay compensation	18202
Offender otherwise dealt with	38270
Offender sent to prison	305281
Suspect charged	1004765
Suspect charged as part of another case	91758
Unable to prosecute suspect	1414560

The next chart depicts the amount of outcome types per crime type.

The chart legend (right side) lists:
- Unable to prosecute suspect
- Suspect charged as part of another case
- Suspect charged
- Offender sent to prison
- Offender otherwise dealt with
- Offender ordered to pay compensation
- Offender given suspended prison sent...
- Offender given penalty notice
- Offender given conditional discharge
- Offender given community sentence
- Offender given absolute discharge
- Offender given a drugs possession warning
- Offender given a caution
- Offender fined
- Offender deprived of property
- Local resolution
- Investigation complete
- Further investigation is not in the public interest
- Formal action is not in the public interest
- Defendant sent to Crown Court
- Defendant found not guilty
- Court case unable to proceed
- Action to be taken by another organisation

1.6. Geographical Distributions

This section analyzes the distribution of the number of crimes in the UK at different granularities.

1.6.1. Total Number of Crimes in the Regions

The first approach was to plot the number of crimes by region. The darker the color of an area, the higher the number of crimes in that area. We can definitely see some areas with a higher number of crimes here. However the data should also be visualized at a lower granularity, as in section 1.6.2.
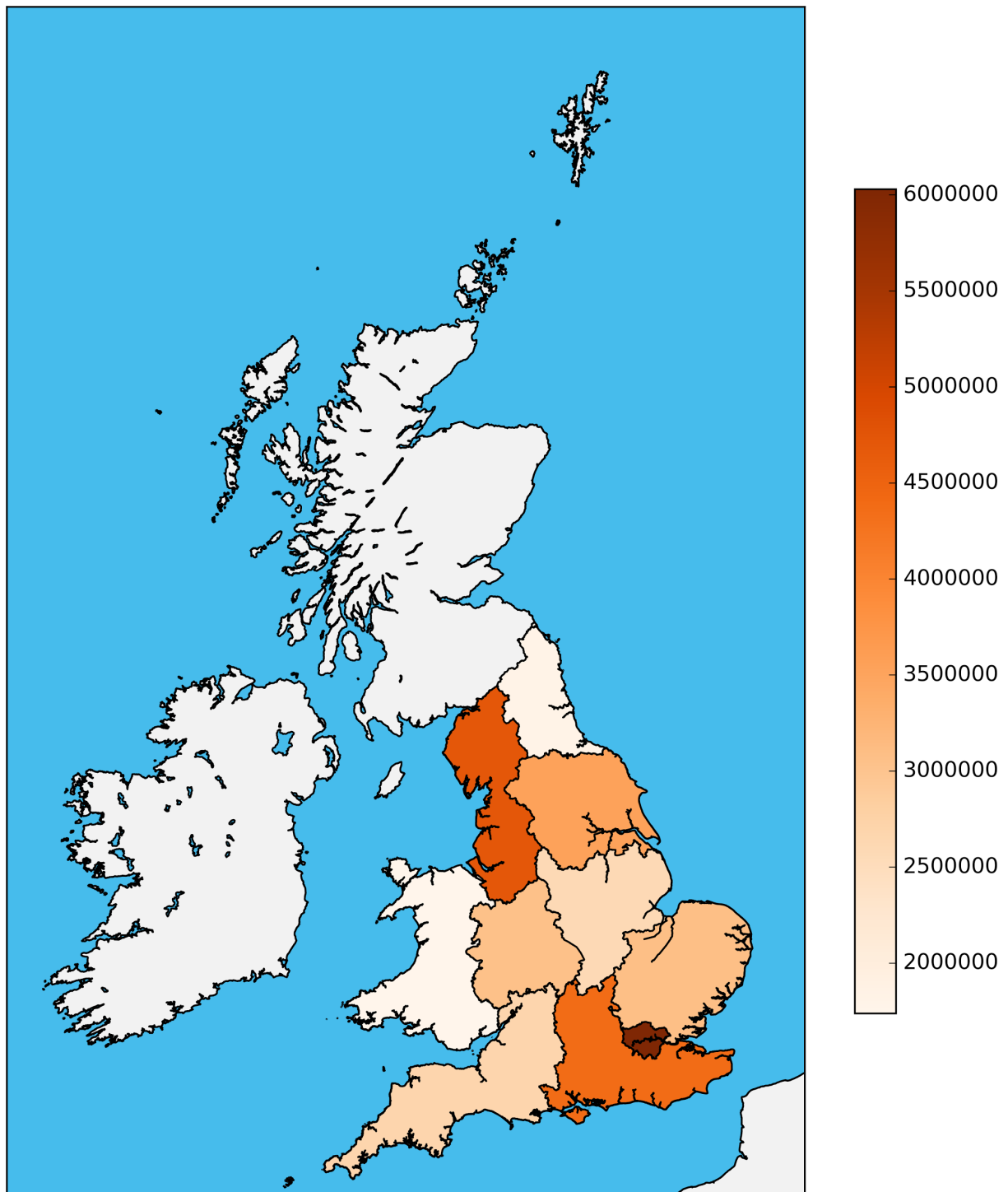
1.6.2. Total Number of Crimes in the Postcode Areas

The map below depicts the number of crimes by postal area. The darker the color of an area, the higher the number of crimes in that area. Some darker areas can be observed

which could be further analyzed to find the roots of the high number of crimes.