

# Big Data Analytics: London Crime Data Analysis

Gianmarco Ricciarelli<sup>1</sup>

<sup>1</sup>University of Pisa,  
gianmarcoricciarelli@gmail.com

# Overview

- 1 Introduction
- 2 Data Understanding
- 3 Cluster Analysis
- 4 Forecasting

# The analysis' purpose

To **discover** the clusters among the criminal activities in the London metropolitan area in a distinct window of time and to **forecast** a possible development for future crimes.

# The Dataset(1)

**London Crime Data, 2008-2016:** this dataset, hosted by **Kaggle**, is composed by 13 millions rows describing the London metropolitan area's criminal activities by *Borough*, *Category*, *Month* and *Year* in a window of time that ranges from January 2008 to December 2016.

# The Dataset(2)

The dataset is composed by 7 variables:

- **lsoa\_code**: code for Lower Super Output Area in Greater London;
- **borough**: common name for London borough;
- **major\_category**: high level categorization of crime;
- **minor\_category**: low level categorization of crime within major category;
- **year**: year of reported counts, 2008 – 2016;
- **month**: month of reported counts, 1 – 12;
- **value**: monthly reported count of categorical crime in given borough;

# The Dataset(3)

The variables *lsoa\_code*, *borough*, *major\_category*, *minor\_category*, *year* and *month* are **categorical** variables, while *value* is a **discrete numerical** variable.

# Numeric Variables' Analysis(1)

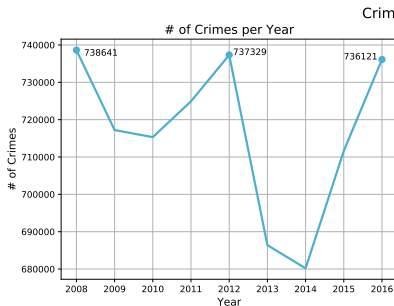
**value** is the only numeric variables in the dataset, it represents the monthly reported count of categorical crime in given borough and has 247 unique values. Its minimum value is 0 and its maximum value is 309, the mode is 0, which appears in the 74.56% of the dataset's samples.

## Numeric Variables' Analysis(2)

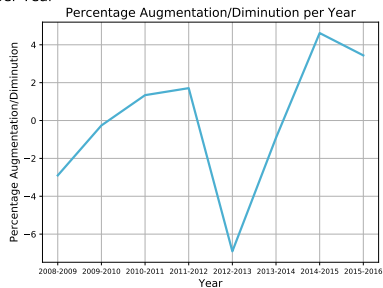
Since 10,071,505, that is, the 74.56% of the dataset's samples have the variable value equals to 0, we can conclude that, on a superficial level, the window of time from 2008 to 2016 wasn't too dense of criminal activities.



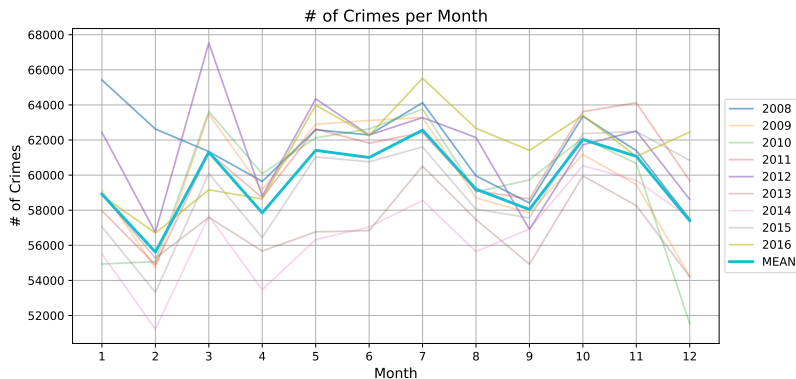
# Crimes per Year



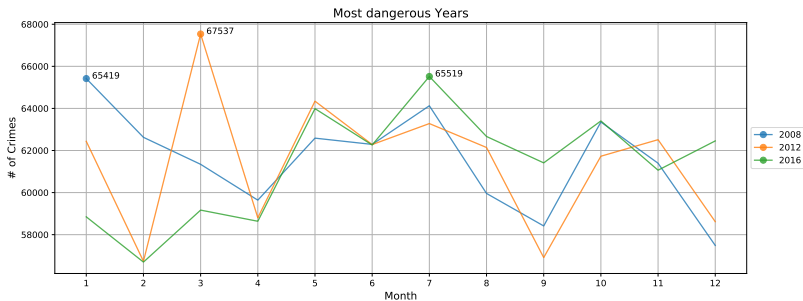
Crimes per Year



# Crimes per Month



# Most Dangerous Years

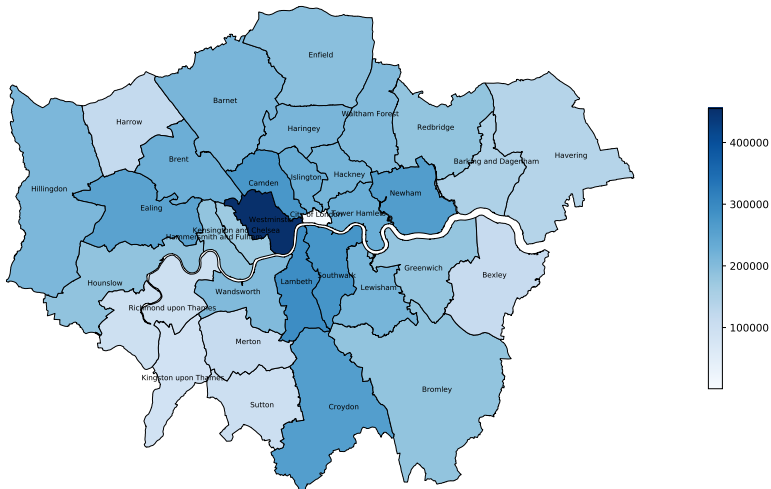


# Categorical Variables' Analysis

- **borough** has 33 unique values, of which Lambeth is the most frequent, appearing in the 4.47% of the cropped dataset's records;
- **major\_category** has 9 unique values, of which Theft and Handling is the most frequent, appearing in the 33.25% of the cropped dataset's records;
- **year** has 9 unique values, of which 2016 is the most frequent, appearing in the 11.45% of the cropped dataset's records;
- **month** has 12 unique values, of which 7 is the most frequent, appearing in the 8.66% of the cropped dataset's records;

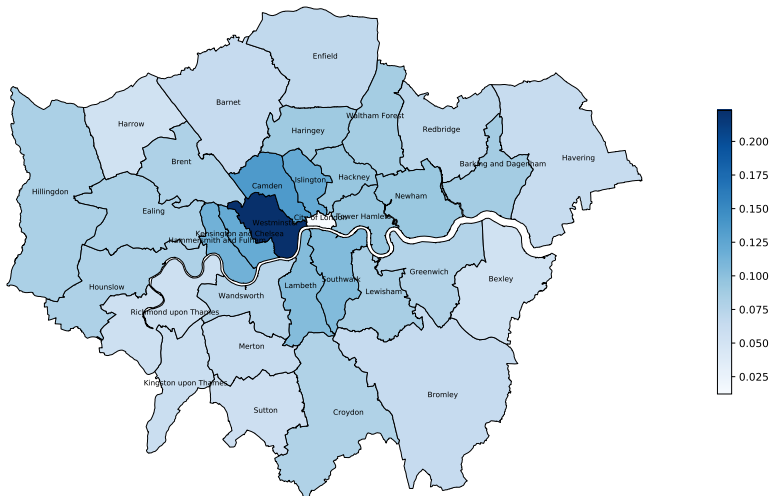
# Crimes per Borough

# of Crimes per Borough

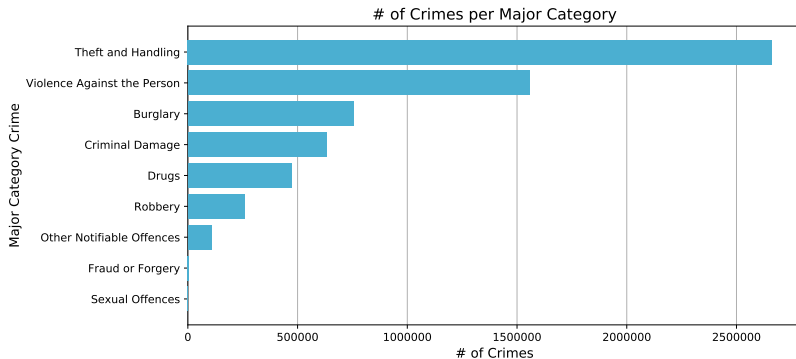


# Crimes per Borough over Population Density

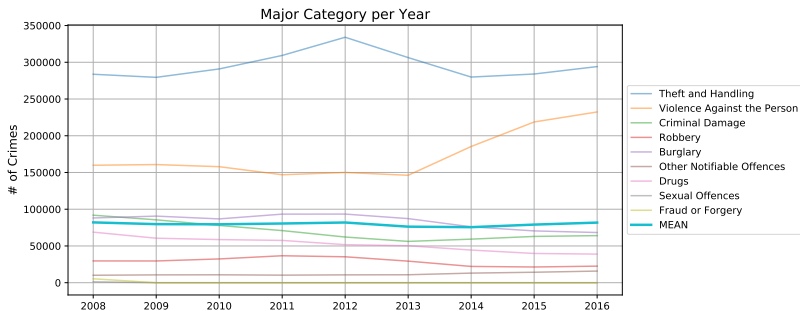
# of Crimes per Borough over Population



# Crimes per Major Category



# Major Category Crimes per Year





# Correlation Analysis

↕ lsoa_code ↕	↕ borough ↕	↕ major_category ↕	↕ minor_category ↕	↕ value ↕	↕ year ↕	↕ month ↕
lsoa_code	D	D	D	D	D	I
borough	D	D	D	D	D	D
major_category	D	D	D	D	D	D
minor_category	D	D	D	D	D	D
value	D	D	D	D	D	D
year	D	D	D	D	D	D
month	I	D	D	D	D	D

# Introduction

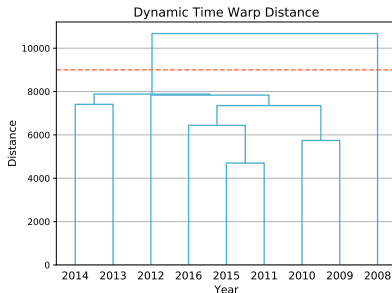
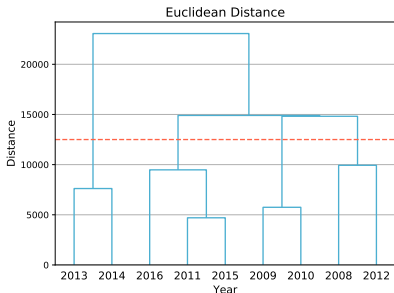
I have decided to enrich the informations provided with the data understanding by searching for possible **cluster-like structures** in the **time series** extracted from the main dataset, that is, hoping to discover the similarities (or dissimilarities) among the series describing the criminal activities from 2008 to 2016.

# Choice of the Algorithms

I have used three popular clustering algorithms, that is, the **KMeans algorithm**, the **Hierarchical Agglomerative algorithm** and the **DBSCAN**. The three algorithms were adapted depending on the different series they were applied on.

# By-Year Series: Hierarchical Agglomerative Clustering

By-Year Series: Hierarchical Agglomerative Clustering



# By-Year Series: KMeans Algorithm and DBSCAN Clustering

◆ Cluster ◆	
Year ◆	◆
2009	0
2010	0
2011	1
2015	1
2016	1
2008	2
2012	2
2013	3
2014	3

(a) KMeans

◆ Cluster ◆	
Year ◆	◆
2008	-1
2012	-1
2009	0
2010	0
2011	1
2015	1
2016	1
2013	2
2014	2

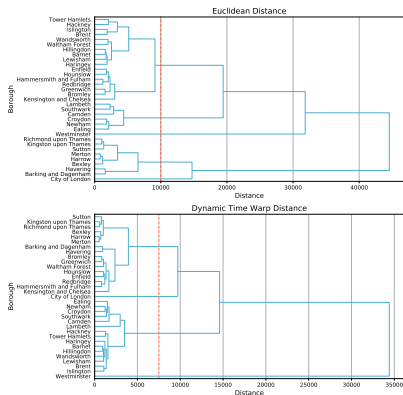
(b) DBSCAN

# By-Year Series: Conclusions

- the series representing years 2013 and 2014 are the least dense of criminal activities, hence are clustered together;
- the series representing years 2008 and 2012 are the most dense of criminal activities, hence are clustered together;
- the remaining series are splitted into two distinct clusters;

# By-Borough Series: Hierarchical Agglomerative Clustering

By-Borough Series: Hierarchical Agglomerative Clustering

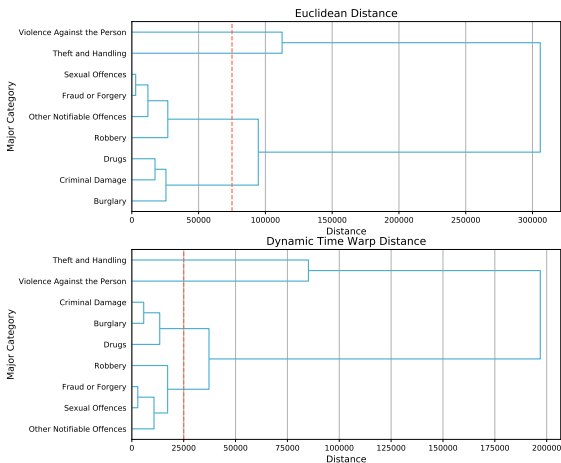


# By-Borough Series: Conclusions



# By-Major Category Series: Hierarchical Agglomerative Clustering

By-Major Category Series: Hierarchical Agglomerative Clustering



# By-Major Category Series: KMeans Algorithm and DBSCAN Clustering

Major Category Crime	Cluster
Robbery	0
Fraud or Forgery	0
Sexual Offences	0
Other Notifiable Offences	0
Theft and Handling	1
Violence Against the Person	2
Criminal Damage	3
Burglary	3
Drugs	3

(a) KMeans

Major Category	Cluster
Theft and Handling	-1
Violence Against the Person	-1
Criminal Damage	0
Burglary	0
Drugs	0
Robbery	1
Fraud or Forgery	1
Sexual Offences	1
Other Notifiable Offences	1

(b) DBSCAN

# By-Major Category Series: Conclusions

- Fraud or Forgery, Sexual Offences, Other Notifiable Offences and Robbery are the least popular types of crimes, hence their series are clustered together;
- Theft and Handling and Violence Against the Person are the most popular types of crimes, hence they form distinct clusters for themselves;
- the other categories are clustered together;

# The Models - Autoregressive Model of order $p$

Represent by the formula

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t$$

where  $\varphi_1, \dots, \varphi_p$  are the parameters of the model,  $c$  is a constant, and  $\varepsilon_t$  is white noise.

# The Models - Autoregressive-Moving-Average

## Model of orders $p$ and $q$

Represented by the formula

$$X_t = c + \sum_{i=1}^p \varphi_i X_{t-i} + \varepsilon_t + \sum_{i=1}^q \theta_i \varepsilon_{t-i}$$

where  $p$  and  $q$  are, respectively, the autoregressive terms and the moving-average terms,  $\varphi_1, \dots, \varphi_p$  are the parameters of the autoregressive model,  $\theta_1, \dots, \theta_q$  are the parameters of the moving-average model and  $\varepsilon_t, \varepsilon_{t-1}, \dots$  are white noise error terms.

# The Models - Autoregressive-Integrated-Average Model of orders $p$ , $q$ and $d$

Represented by the formula

$$\left(1 - \sum_{i=1}^p \varphi_i L^i\right) (1 - L)^d X_t = \left(1 + \sum_{i=1}^q \theta_i L^i\right) \varepsilon_t$$

where  $L$  is the lag operator,  $\varphi_1, \dots, \varphi_p$  are the parameters of the autoregressive model,  $\theta_1, \dots, \theta_q$  are the parameters of the moving-average model and  $\varepsilon_t$  is white noise.

# General Comparison