# Big Data Analytics: London Crime Data Analysis

Gianmarco Ricciarelli[1]

[1]University of Pisa,
gianmarcoricciarelli@gmail.com

# Overview

# The analysis' purpose

To discover the patterns among the criminal activities in the London metropolitan area in a distinct window of time.

# 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 eguals 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.
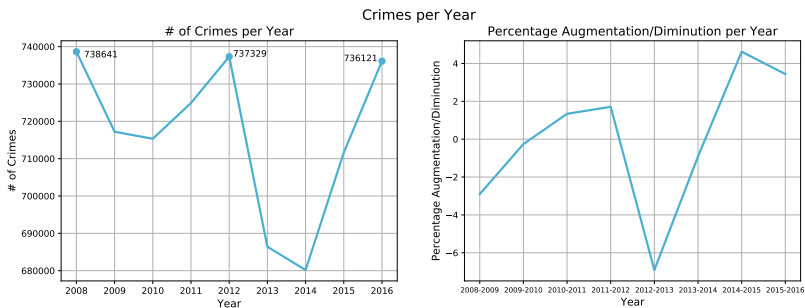
# Crimes per Year



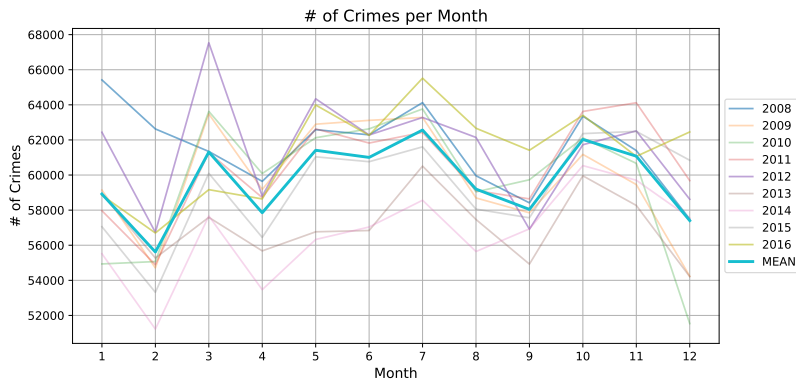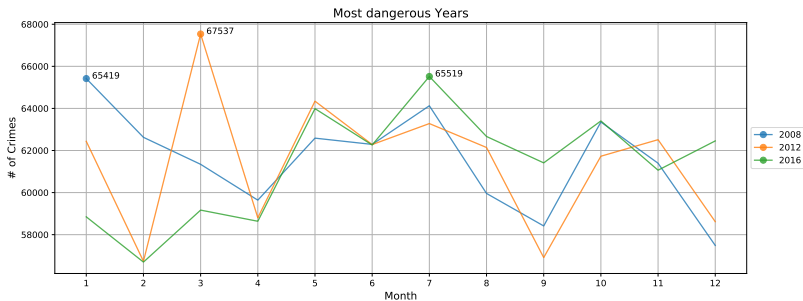Figure: Crime's progress over the years

# Crimes per Month



Figure: Crime's progress over the months

# 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;
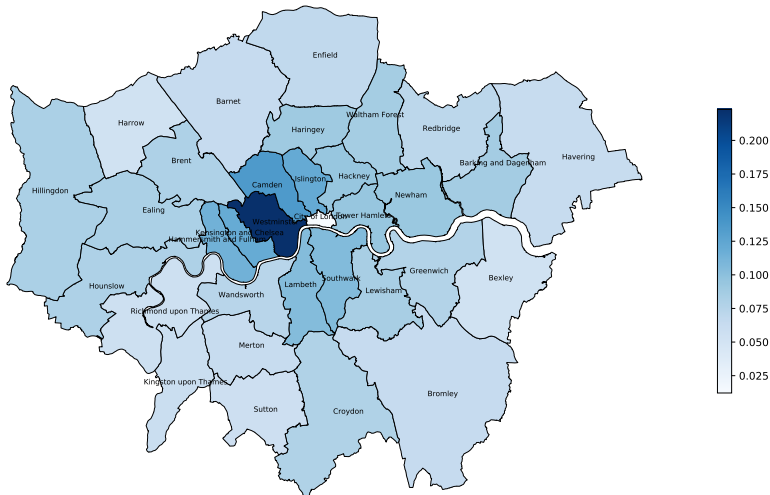
# Crimes per Borough
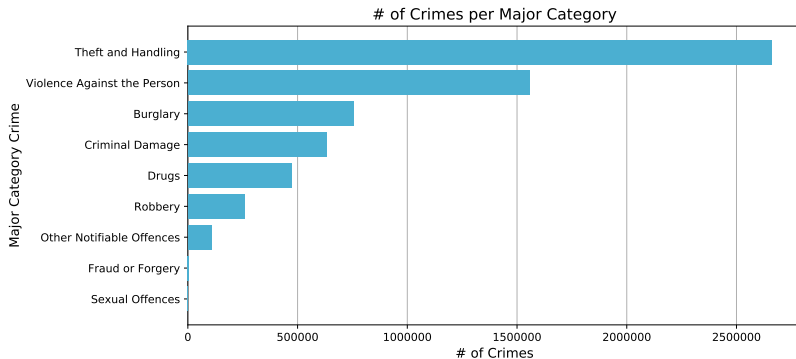


# of Crimes per Borough
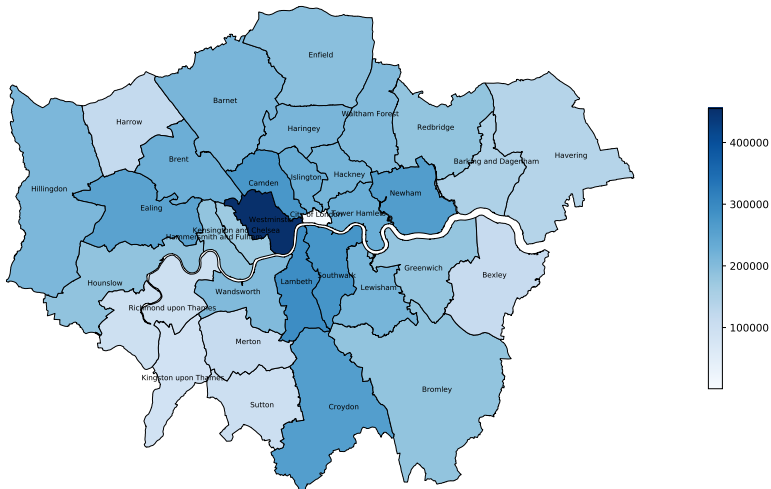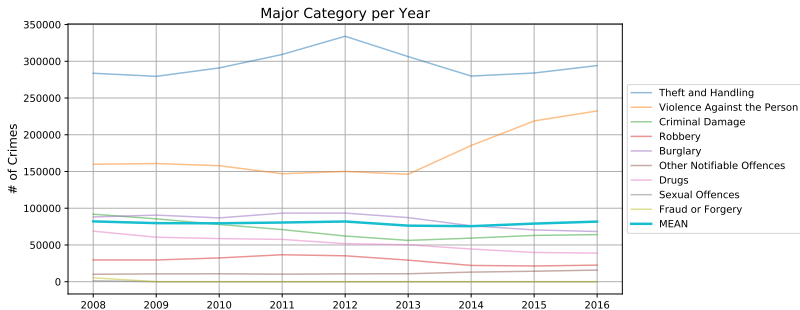
# Crimes per Borough over Population Density



# 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 | D | I |
| **borough** | D | D | D | D | D | D | D |
| **major_category** | D | D | D | D | D | D | D |
| **minor_category** | D | D | D | D | D | D | D |
| **value** | D | D | D | D | D | D | D |
| **year** | D | D | D | D | D | D | D |
| **month** | I | D | D | D | D | D | D |

# Introduction

I have decided to enrich the informations provided with the data understaning 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
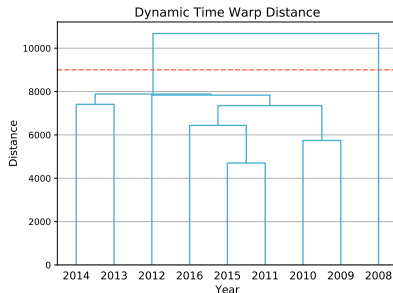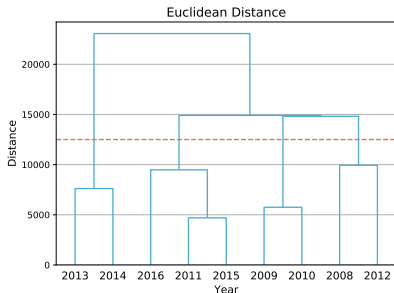
# By-Year Series: KMeans Algorithm and DBSCAN Clustering

| 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-Borough Series: Hierarchical Agglomerative Clustering



By-Borough Series: Hierarchical Agglomerative Clustering

# By-Major Category Series: Hierarchical Agglomerative Clustering



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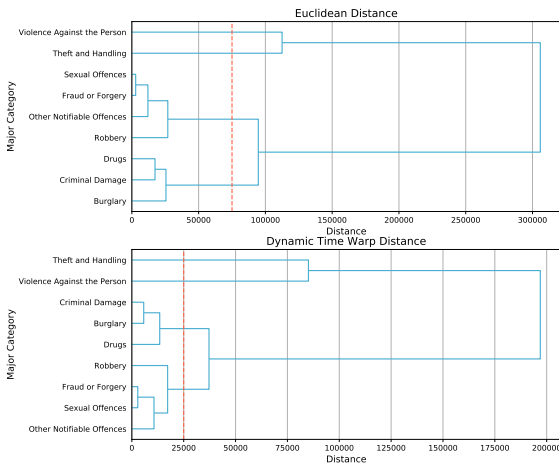
# By-Major Category Series: KMeans Algorithm and DBSCAN Clustering

| | Cluster |
|---|---|
| **Major Category Crime** | |
| 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

| | Cluster |
|---|---|
| **Major Category** | |
| 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