# An exploration of London Crime

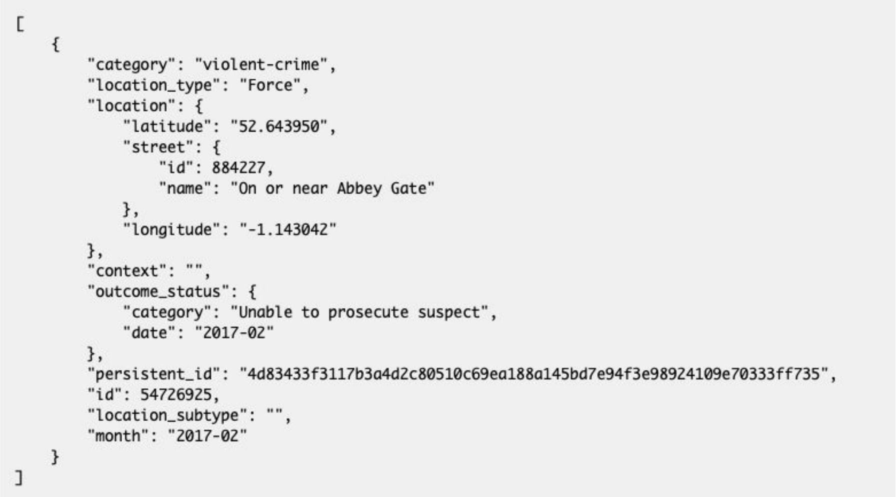
By Mahima and Tom

# Introduction to the Data

Fortunately for us, Crime data in the UK is extremely open to the general public. There is a government run central database, which is updated constantly, providing APIs for general queries about crime. The one we used, at https://data.police.uk/docs/, returned all crimes that occured at a particular location, in a specified month and year.

We also accessed data through archive CSV files put out by the Metropolitan police in London, giving us all the crimes in the occured in any particular borough. They were massive files, of course, that required some cleaning.

We also had a look into some excel files put together by other data scientists working on similar projects to try help put some of the data into context



```
[
    {
        "category": "violent-crime",
        "location_type": "Force",
        "location": {
            "latitude": "52.643950",
            "street": {
                "id": 884227,
                "name": "On or near Abbey Gate"
            },
            "longitude": "-1.143042"
        },
        "context": "",
        "outcome_status": {
            "category": "Unable to prosecute suspect",
            "date": "2017-02"
        },
        "persistent_id": "4d83433f3117b3a4d2c80510c69ea188a145bd7e94f3e98924109e70333ff735",
        "id": 54726925,
        "location_subtype": "",
        "month": "2017-02"
    }
]
```

# Our Hypotheses

With such massive data, it was hard to know where to start!

1. Bicycle crime is less common in 2016 than it was in 2015 in London

2. Crime got proportionately less violent in Camden in October to November 2011

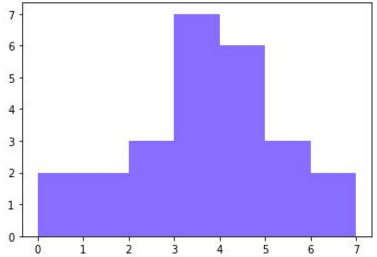
3. Unemployment improved between 2009 and 2010 in London

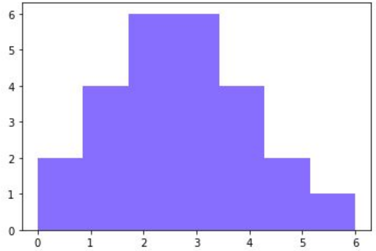
4. Robberies in Lambeth and Lewisham remain unchanged over time

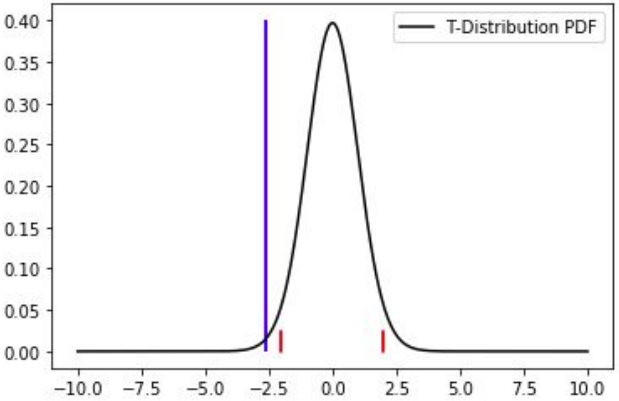
5. Anti-social behaviour in Westminster and Croydon increased over time

(H0: p1=p2 for all, where p1 and p2 are the means for the two samples)

# Hypothesis 1 - Bicycle Crime







1 - We ran the API over a loop, to get a random selection of crimes occurring at different locations in London, for both 2011 and 2013

2 - We ran a Welch's T Test to see if the change in prevalence in Bicycle theft was relevant:

We first checked for normality of our data, which was confirmed

We calculated the t statistic and compared it to the t distribution

We calculated the critical value at the alpha=0.05 level for this specific t distribution, and found that our t statistic was lower than the t critical value, which means we can reject the null hypothesis!

We have evidence to suggest Bicycle theft is improving in London!

# Hypothesis 2 - Unemployment

1 - We cleaned up and extracted data on unemployment in London, and how it changed over the last few years

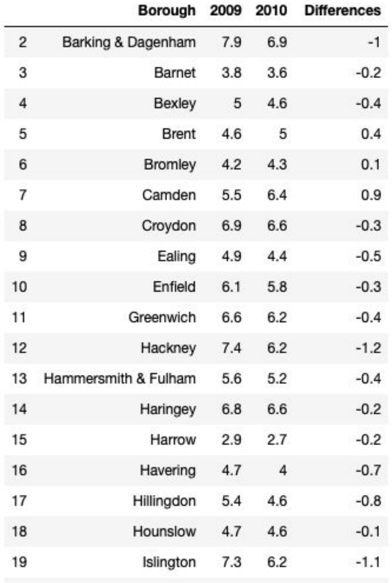
2 - We ran a Two Sample Paired T Test to see if there was a genuine change in unemployment from 2009 to 2010:

We calculated the t statistic and p value for the differences between the years, using the relevant stats python library

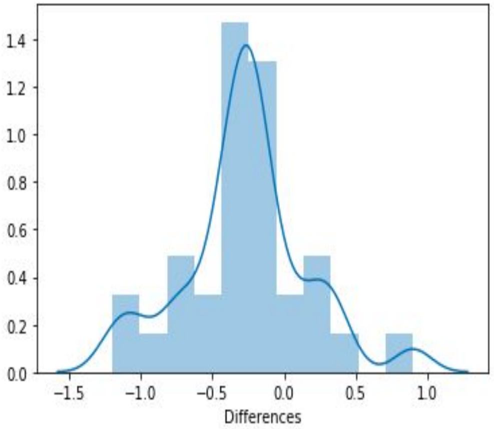
Such a small p value is plenty of evidence to reject the null hypothesis at the 0.05 level

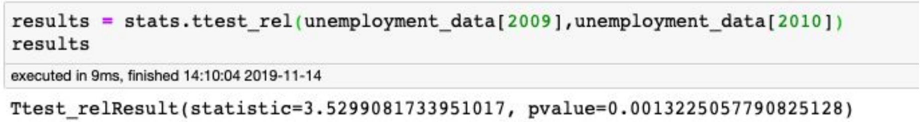
Cohen’s d came out as -0.2810698560212756, suggesting a small/medium effect size, which was successfully picked up on by this sample size

We have evidence to suggest Unemployment genuinely improved in London!



| | Borough | 2009 | 2010 | Differences |
|---|---|---|---|---|
| 2 | Barking & Dagenham | 7.9 | 6.9 | -1 |
| 3 | Barnet | 3.8 | 3.6 | -0.2 |
| 4 | Bexley | 5 | 4.6 | -0.4 |
| 5 | Brent | 4.6 | 5 | 0.4 |
| 6 | Bromley | 4.2 | 4.3 | 0.1 |
| 7 | Camden | 5.5 | 6.4 | 0.9 |
| 8 | Croydon | 6.9 | 6.6 | -0.3 |
| 9 | Ealing | 4.9 | 4.4 | -0.5 |
| 10 | Enfield | 6.1 | 5.8 | -0.3 |
| 11 | Greenwich | 6.6 | 6.2 | -0.4 |
| 12 | Hackney | 7.4 | 6.2 | -1.2 |
| 13 | Hammersmith & Fulham | 5.6 | 5.2 | -0.4 |
| 14 | Haringey | 6.8 | 6.6 | -0.2 |
| 15 | Harrow | 2.9 | 2.7 | -0.2 |
| 16 | Havering | 4.7 | 4 | -0.7 |
| 17 | Hillingdon | 5.4 | 4.6 | -0.8 |
| 18 | Hounslow | 4.7 | 4.6 | -0.1 |
| 19 | Islington | 7.3 | 6.2 | -1.1 |



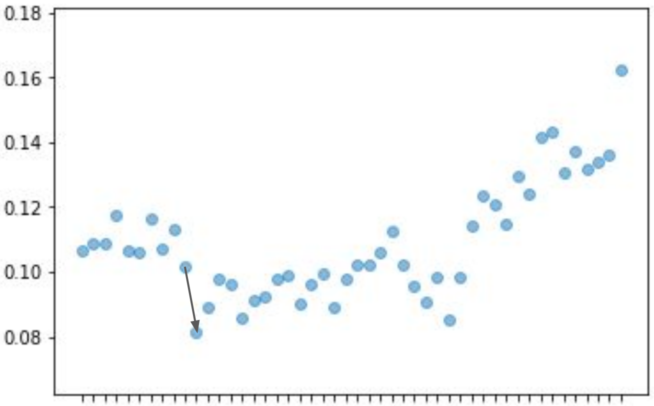


```
results = stats.ttest_rel(unemployment_data[2009],unemployment_data[2010])
results
```

executed in 9ms, finished 14:10:04 2019-11-14

```
Ttest_relResult(statistic=3.5299081733951017, pvalue=0.0013225057790825128)
```

# Hypothesis 3 - Violent Crime



1 - We ran a function through our massive CSVs on all crime data in London, which returned the proportion of crime that is violent in a particular borough

We had a closer look, as we noticed a substantial drop from October 2012, to November 2012, in Camden

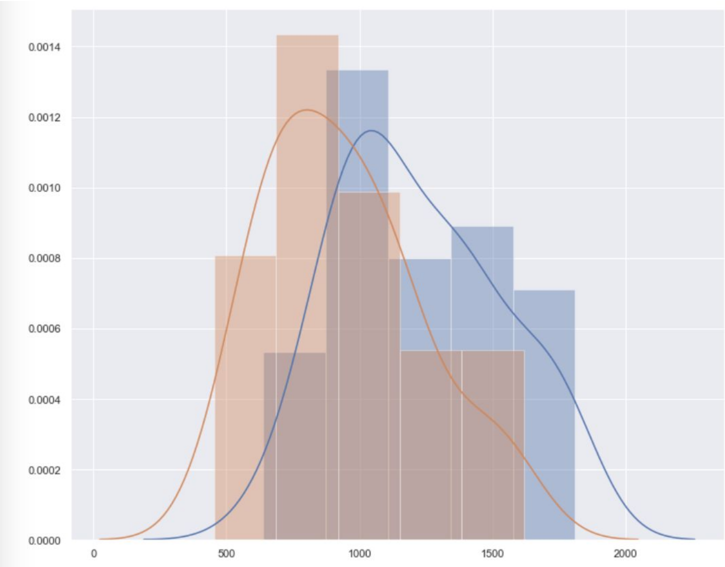
We did a one sample p test, assuming that the October month was the correct mean, to see the likelihood that November was a fluke.

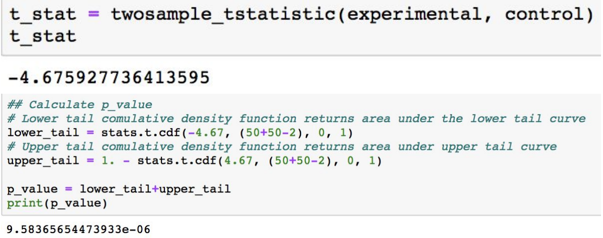
The z value came out at z=3.17, which was well over the z=1.96 we needed to reject the null hypothesis.

We have evidence to suggest Violent Crime genuinely improved in London!

# Hypothesis 4 - Robberies

1. The UK police defines robbery as mugging, theft and snatch theft involving violence or threats
2. To evaluate the incidence of robbery we looked at two boroughs - Lambeth and Lewisham - and checked if there was any change in the incidence over our time period Jan 2011- Dec 2014
3. As the first step, we cleaned the dataset and created a function to loop through all the CSV files
4. Running a two-sample t-test, we calculated the t-stat and p-value, keeping alpha = 0.05
5. As the p-value is much smaller than our alpha, we reject the null hypothesis, and conclude that there is a high  difference in the robbery levels between the two boroughs





```
t_stat = twosample_tstatistic(experimental, control)
t_stat
```

```
-4.675927736413595
```

```
## Calculate p_value
# Lower tail comulative density function returns area under the lower tail curve
lower_tail = stats.t.cdf(-4.67, (50+50-2), 0, 1)
# Upper tail comulative density function returns area under upper tail curve
upper_tail = 1. - stats.t.cdf(4.67, (50+50-2), 0, 1)

p_value = lower_tail+upper_tail
print(p_value)
```

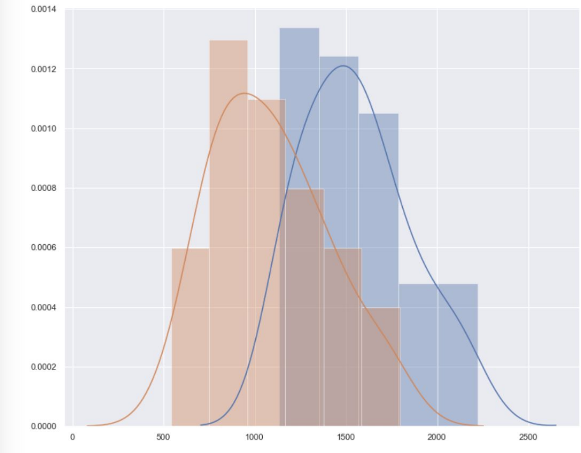
```
9.58365654473933e-06
```

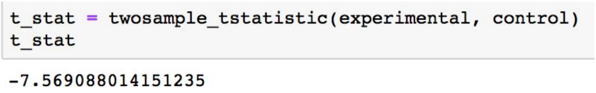
*Blue - Lambeth(control)
Orange - Lewisham (experiment)

# Hypothesis 4 - Anti-social behaviour

1. The UK police defines robbery as drunken or threatening behaviour, vandalism, graffiti and playing loud music at night
2. To evaluate the incidence of robbery we looked at two boroughs - Westminster and Croydon - and checked if there was any change in the incidence over our time period Jan 2011- Dec 2014
3. As the first step, we cleaned the dataset and created a function to loop through all the CSV files
4. Running a two-sample t-test, we calculated the t-stat and p-value, keeping alpha = 0.05
5. As the p-value is smaller than our alpha, we reject the null hypothesis, and conclude that there is a significant difference in the robbery levels between the two boroughs

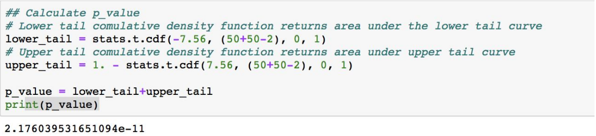
*Blue - Westminster (control)
Orange - Croydon (experiment)





```
t_stat = twosample_tstatistic(experimental, control)
t_stat
```

```
-7.569088014151235
```



```
## Calculate p_value
# Lower tail comulative density function returns area under the lower tail curve
lower_tail = stats.t.cdf(-7.56, (50+50-2), 0, 1)
# Upper tail comulative density function returns area under upper tail curve
upper_tail = 1. - stats.t.cdf(7.56, (50+50-2), 0, 1)

p_value = lower_tail+upper_tail
print(p_value)
```

```
2.176039531651094e-11
```