Materials & Methods:

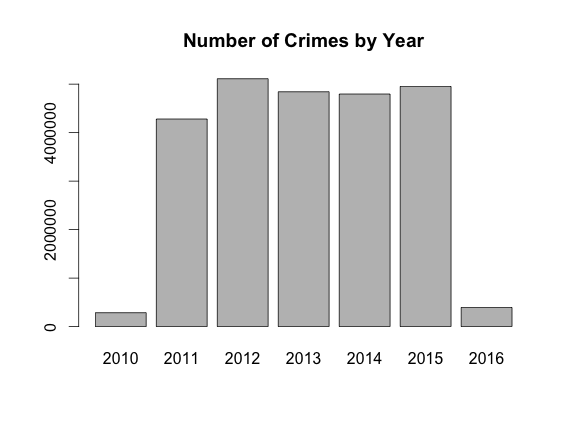
We collected data from a variety of sources for our project. First of all, we had crime data from the UK Home Office. This dataset contains every incident of criminal activity reported in England, Wales, and Northern Ireland from 2010-2016[[1]](#footnote-1), organized by month and responding police department. In all, our dataset contained 24,668,162 incidents, organized by which LSOA they occurred in. The breakdown by year is presented in Figure 1, and the breakdown by crime type is presented in Figure 2. 

Figure 1 –There is only one month of data available for 2010 and 2016. There was a big jump in crime from 2011-2012; aside from that, the number of crimes was relatively constant.

In addition, we collected demographic and economic statistics from Nomis[[2]](#footnote-2), a service provided by the Office for National Statistics. This data consisted of the monthly unemployment rate (among those aged 16-64) and population for each of 381 LADs from 2010-2016. We calculated the yearly unemployment rate by averaging the monthly rates, and ended up with 1905 year-district observations, of which 120 had to be thrown out for lacking data. Figure 3 shows the national yearly unemployment rate for 2011-2015. Our third data source was a crosswalk file[[3]](#footnote-3) which we used to aggregate the crime data from the LSOA level to the LAD level. This was also published by the Office for National Statistics. Finally, we gathered LAD-level shapefiles[[4]](#footnote-4) for England from the Census Support division of the UK Data Service. This dataset consisted of 326 records, one for each LAD in England.

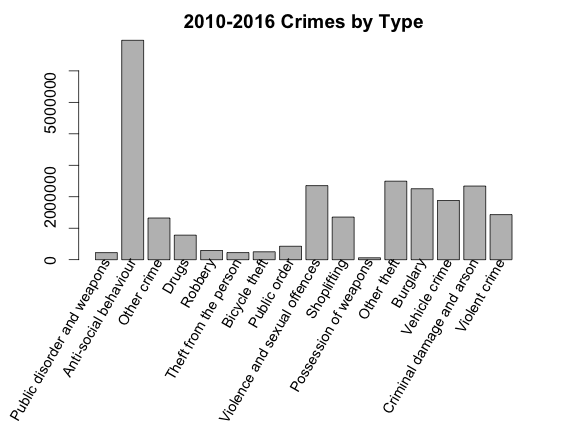


Figure 2 – Anti-social behaviour, which includes crimes like littering, vandalism, and street drinking, is by far the most common crime type

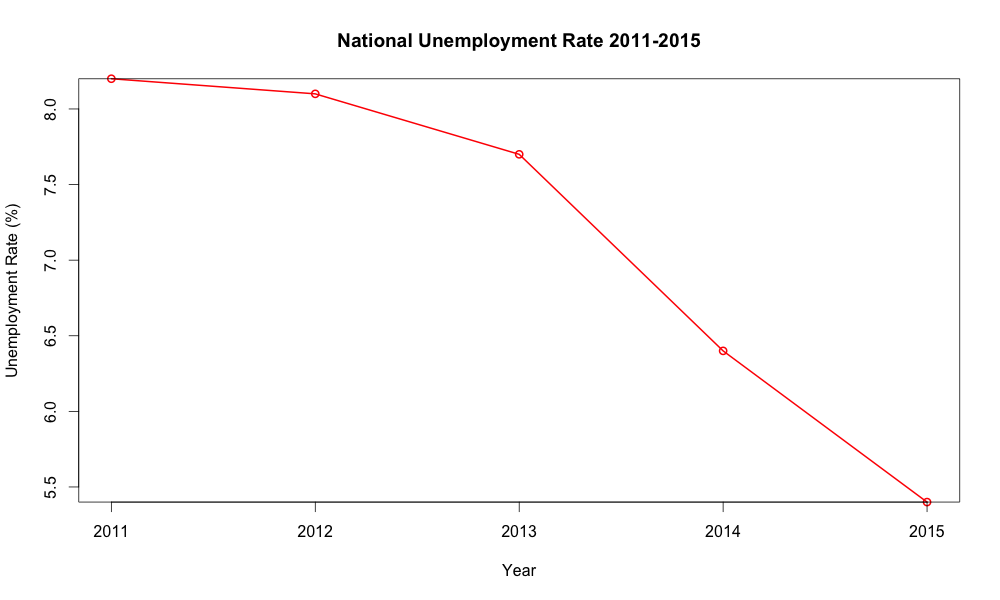


Figure 3 –Unemployment rate by year. There was a significant downward trend in unemployment over the years we looked at

The coordinate system used by the shapefiles we found was quite obscure – OSGB 1936 / British National Grid – we used Python to convert them into a standard system, so that we would be able to create visualizations over a world map background. We decided to use WGS 84, “the standard U.S. Department of Defense definition of a global reference system for geospatial information and the reference system for the Global Positioning System.”[[5]](#footnote-5)

We used PySpark and SparkSQL to aggregate the crime data up to the LAD level. First, to avoid the small files problem, we collapsed all 6000 tables in the crime dataset into a single 6.6gb table. After that we cleaned the data as best as we could. In particular, we removed duplicate crime IDs by month, favoring records with more information, expecting that they were perhaps put in multiple times by accident. We also removed all records that were missing their location information or appeared as complete duplicates.

At this point in our project we discovered something interesting about how Spark interacts with Amazon’s Elastic MapReduce software; when we ran our cleaning program in Spark 1.6.0 on EMR 4.3, the cores in our clusters were not being utilized effectively (Figure 4). However, when we ran it in the same version of Spark on EMR 4.4, the work was parceled out much more efficiently (Figure 5), and as a result the program’s run-time was cut nearly in half.

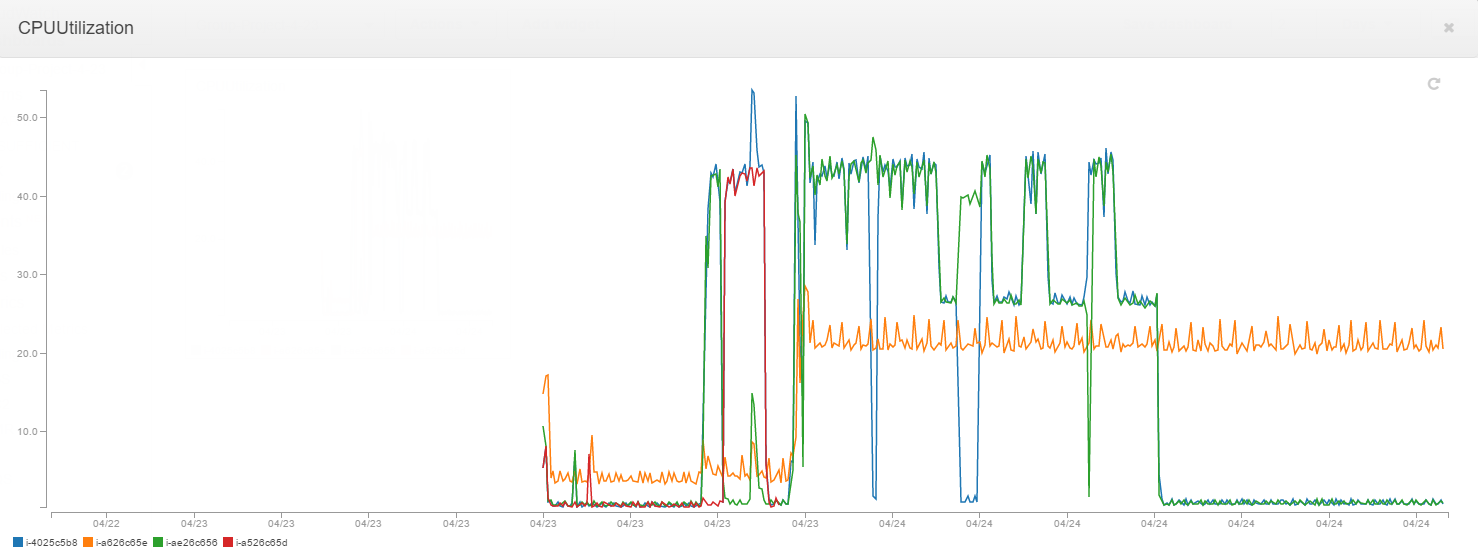


Figure 4 –EMR 4.3, Spark 1.6.0. Most of the time, most notes are not even 50% utilized. The master node is not pictured here.

The unemployment data came in the standard rate format; we needed to convert the crime data to a similar format for easy visual comparison. We did this in R – by dividing the number of crimes in a year in a LAD by the LAD’s total population that year, and multiplying the result by 100, we were able to create a crime rate statistic measured in units of crimes per 100 people.

We also made liberal use of Amazon’s Simple Storage Service (S3) to store data and to host our maps online. \*Tim’s part about making maps in Bokeh goes here – I don’t have anymore to say about S3 so see if you can add on to the end of this paragraph to transition into Bokeh\*

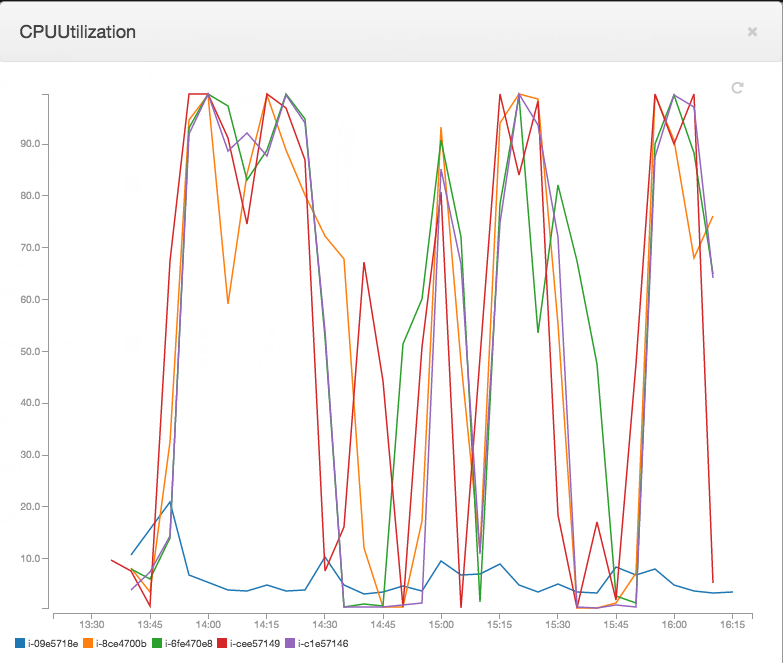


Figure 5 - EMR 4.4, Spark 1.6.0. Most of the time, most nodes are over 90% CPU utilization. Master node shown in blue

1. "Data Downloads." Home | Data.police.uk. UK Home Office. Web. 10 Apr. 2016. <https://data.police.uk/data/>. [↑](#footnote-ref-1)
2. "Nomis Official Labour Market Statistics." Home. Office for National Statistics. Web. 10 Apr. 2016. <https://www.nomisweb.co.uk/>. [↑](#footnote-ref-2)
3. "Lookups between 2011 Census Output Areas and Other Geographies." Office for National Statistics. Web. 10 Apr. 2016. <http://webarchive.nationalarchives.gov.uk/20160105160709/http://www.ons.gov.uk/ons/guide-method/geography/products/census/lookup/2011/index.html>. [↑](#footnote-ref-3)
4. "Census Support Easy Download: English Boundary datasets." UK Data Service. Web. 10 Apr. 2016. <https://census.edina.ac.uk/easy\_download.html>. [↑](#footnote-ref-4)
5. “World Geodetic System 1984.” National Geospatial-Intelligence Agency. Web. 10 Apr. 2016. <http://www.unoosa.org/pdf/icg/2012/template/WGS\_84.pdf>.

   [↑](#footnote-ref-5)