**Introduction**

Around the world police are receiving greater scrutiny in all aspects of duty. In the interests of increased transparency of police operations, the United Kingdom has compiled and released detailed data covering street-level crime, outcomes, and police response. However, the database is vast, and in its raw form, the data is opaque. For our final project, we wanted to start the process of bringing order to that data in an effort to find out what it would say about crime in the United Kingdom. In particular, we will couple the data from the UK Home Office (the crime data) with data from the Office for National Statistics (Socioeconomic Status (SES) data) in an effort to gain a better idea of the overall health of society.

This project seeks to answer the question of whether or not there is a relationship between the health of society, which will be represented by the unemployment rate of residents between 16 and 64 year of age, and the crime rate within society. It is our hypothesis that higher levels of unemployment directly lead to higher levels of crime. In order to test the hypothesis, we will attempt to predict crime rates with unemployment through linear regression analysis using data available at various organizational levels of society. Those organizational levels of society are hierarchical, and we will be using the following three: Lower Layer Super Output Areas (LSOAs), Middle Layer Super Output Areas (MSOAs), and Local Authority Districts (LADs). Information at the LSOA level can be aggregated up to the MSOA level or the LAD level. The MSOA level is between the LSOA and LAD levels.

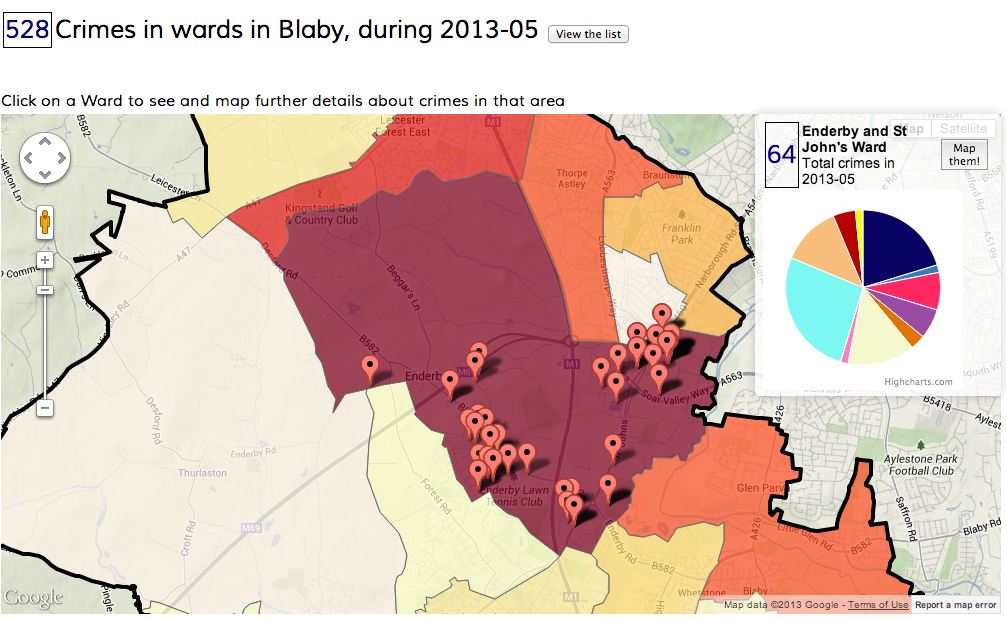
This project also strives to make the knowledge within the data more accessible to the average user. In particular, we will provide access to the data through development of descriptive statistics and an application of those statistics to maps of the UK. Visualizations of the data through maps make it much easier to see trends across the country and within output areas simultaneously and across time.

Information from the above analyses and statistics contribute to the health of society in a number of ways. First, it can educate the citizenry, helping citizens to make safe choices about where to work, live, and travel. Second, a better understanding of crime will yield further hypotheses about how to reduce it, paving the way for policies and programs in government. Finally, such groundwork can lay the foundation for future research focused around identifying what police action is contributing to crime prevention and which may be abuse.

Through this paper, each section will provide the reader with another piece of the puzzle behind the hypothesis. First, there will be a short literature review. Second, an explanation of the materials and methods we employed to explore the data and hypothesis. Third, we will present the results of testing the hypothesis directly through linear regression analysis. Finally, there will be a conclusion.

**Prior Work and Influences**

The United Kingdom government has been at the forefront of the open data movement, and among the available data, there have been a variety of applications created to visualize crime data. One of the first attempts we found was from a blog that described an app (Figure 1) that provided street-level crime data for each ward within a local authority district. It provided a map of the district and clicking on a ward provided a pie chart with information about the number and types of crimes in that area for a selected month. The blog post mentioned that there were ideas on combining crime data with related public sources, but unfortunately, the app and associated website are now defunct.



*Figure 1 – A screenshot of the now-defunct interactive application on crime data in England*

A very similar website[[1]](#footnote-1) by the City of London Police maps the crime data and is filterable by crime type. Individual crimes can be viewed at the street level, and they offer basic statistics on the data over the most recent one-year period. However, the search is only provided by individual months and is limited to the City of London (Figure 2).

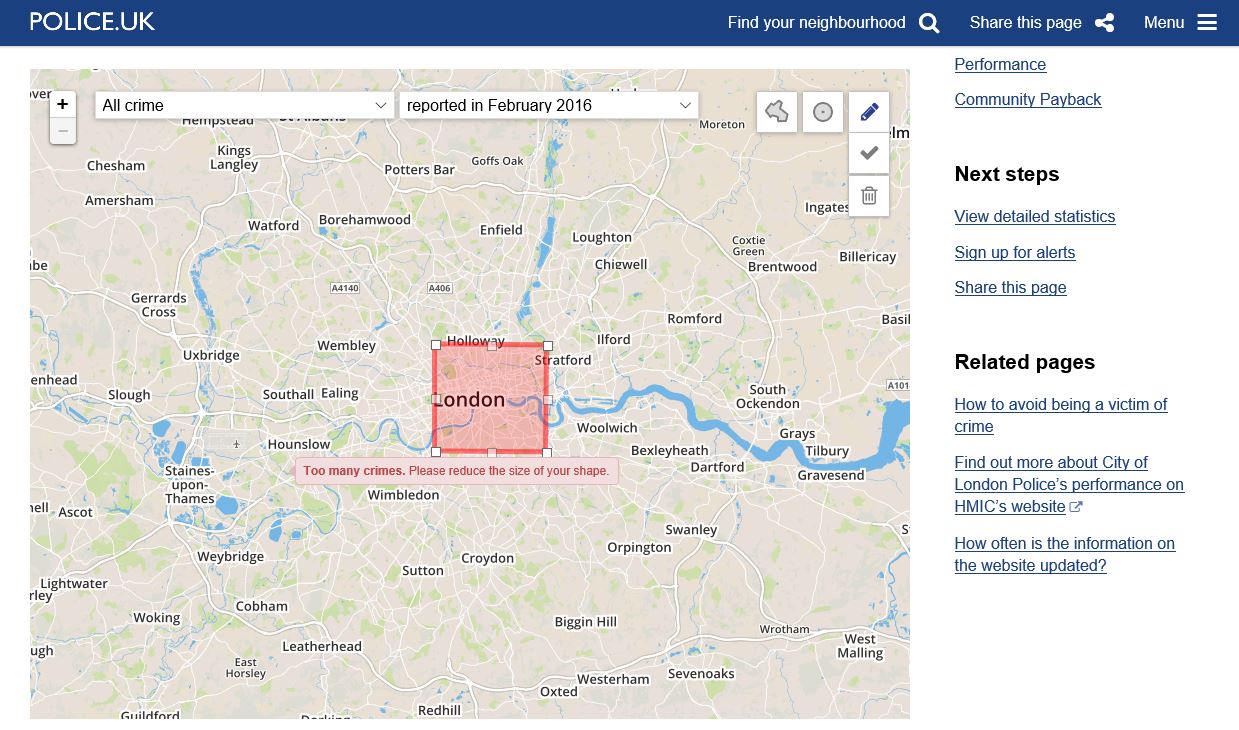


Figure 2 – A screenshot of the City of London Police website. The image is of the crime map section

Finally, one of the more impressive visualizations that we came across was not on crime data, but on information from the UK census[[2]](#footnote-2) (Figure 3). There are several different levels of filtering, an impressive map that is mapped out down to the Lower-Layer Super Output Area (LSOA), and very robust data.

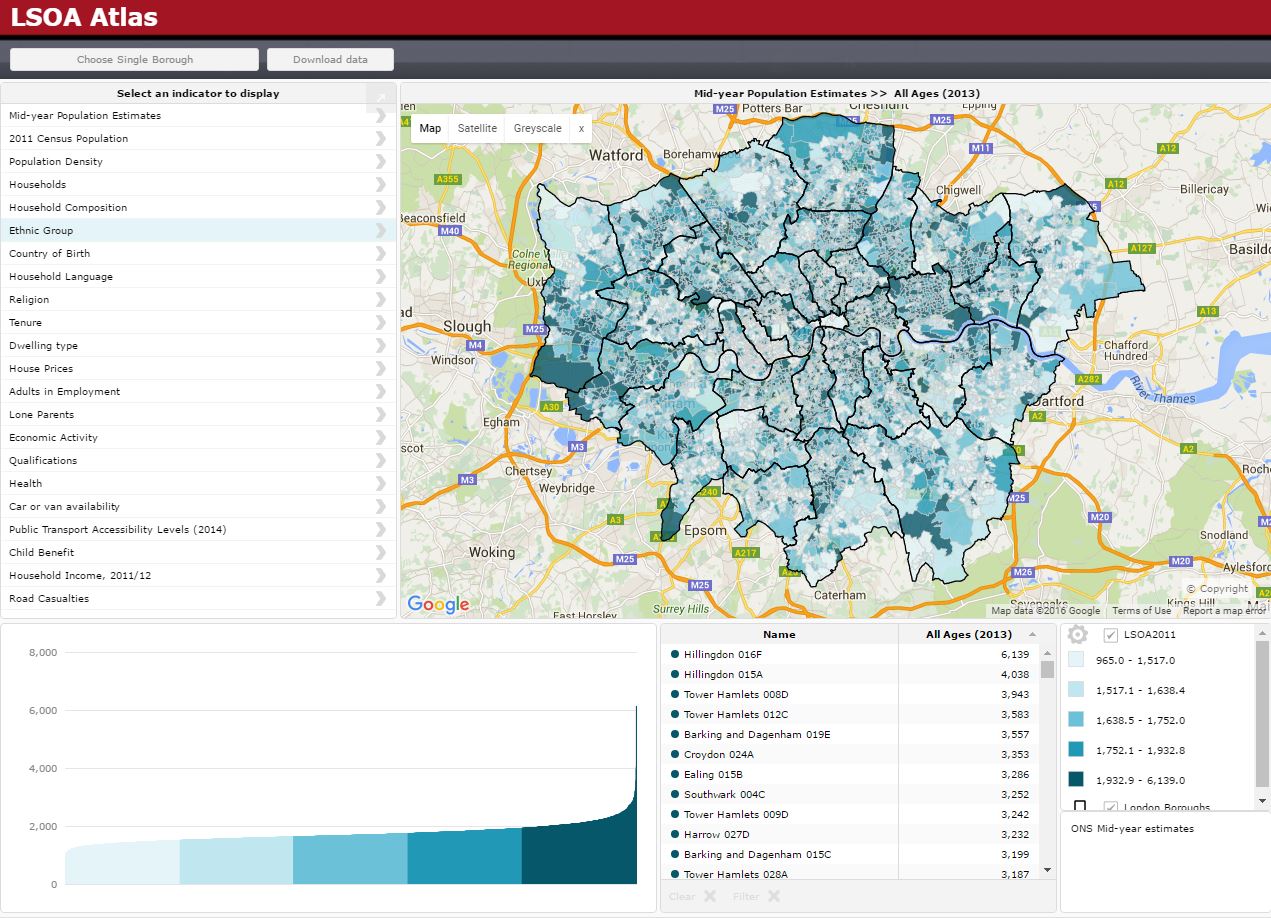


Figure 3 – A screenshot of the LSOA Atlas website which provides census information

These previous works, among others, provided stunning interactive visualizations, but they largely just relayed existing information. Our goal was to uncover new information by uniting these rich data sources.

Some prior research that influenced our work[[3]](#footnote-3) incorporated the complex feedback between social and economic change in the domain of criminal justice, but also weaved in the political ideas of Thatcherism and its implications to their analyses. Our interpretations are strictly from the data at hand, without significant prior knowledge of British political history. Another source of inspiration was “The Economy, Crime and Time: An Analysis of Property Crime in England & Wales 1961-2006 (Jennings, Farrall, Bevan[[4]](#footnote-4))” which provided a solid example of inferences and relationships between variables of crime, economy, and time in a time series regression model. The approached used in that study specifically focused on property crime.

**Materials & Methods:**

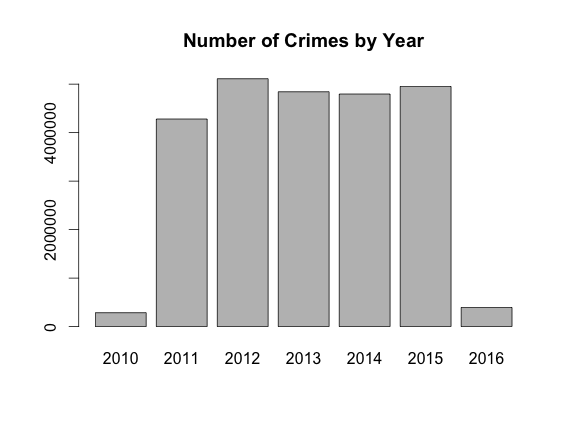
We collected data from a variety of sources for our project. First of all, we had crime data from the UK Home Office. This dataset contains every incident of criminal activity reported in England, Wales, and Northern Ireland from 2010-2016[[5]](#footnote-5), organized by month and responding police department. In all, our dataset contained 24,668,162 incidents, organized by which LSOA they occurred in. The breakdown by year is presented in Figure 4, and the breakdown by crime type is presented in Figure 5. 

Figure 4 –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[[6]](#footnote-6), 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 6 shows the national yearly unemployment rate for 2011-2015. Our third data source was a crosswalk file[[7]](#footnote-7) 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[[8]](#footnote-8) for England from the Census Support division of the UK Data Service. This dataset consisted of 326 records, one for each LAD in England.

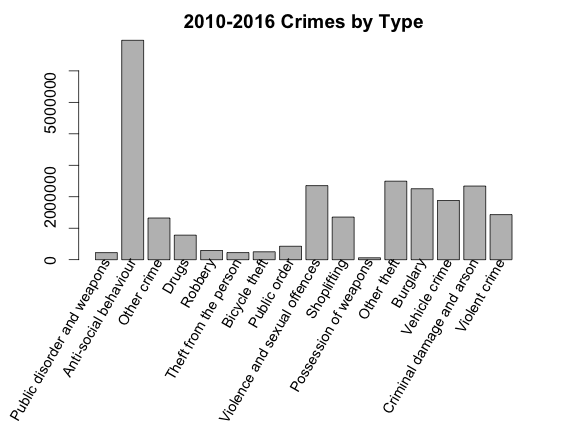


Figure 5 – Anti-social behaviour, which includes crimes like littering, vandalism, and street drinking, is by far the most common crime type

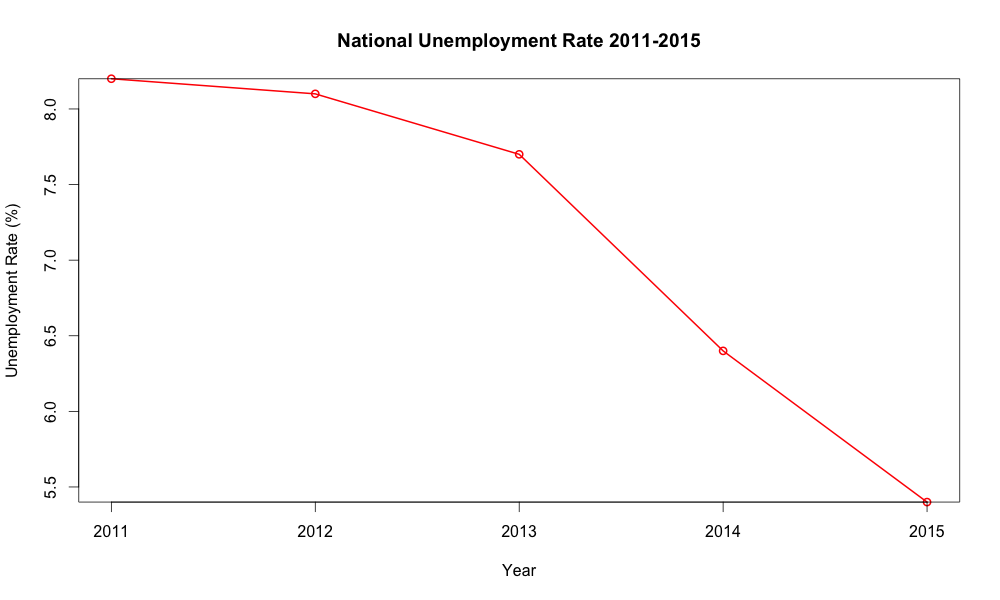


Figure 6 –Unemployment rate by year. There was a significant downward trend in unemployment over the years we looked at

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.”[[9]](#footnote-9)

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. Finally, we merged on information from crosswalk file that would allowed us to aggregate from the LSOA level to the LAD level.

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 7). However, when we ran it in the same version of Spark on EMR 4.4, the work was parceled out much more efficiently (Figure 8), and as a result the program’s run-time was cut nearly in half.

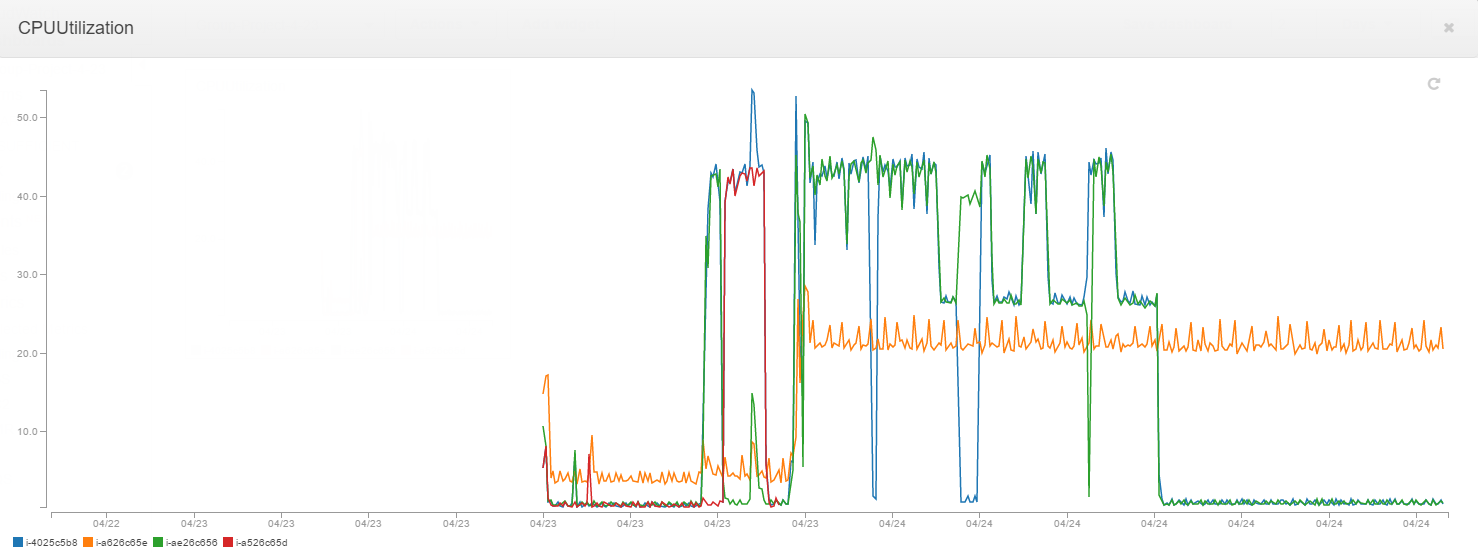


Figure 7 – EMR 4.3, Spark 1.6.0. Most of the time, most notes are not even 50% utilized. The master node is not pictured here.

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\*

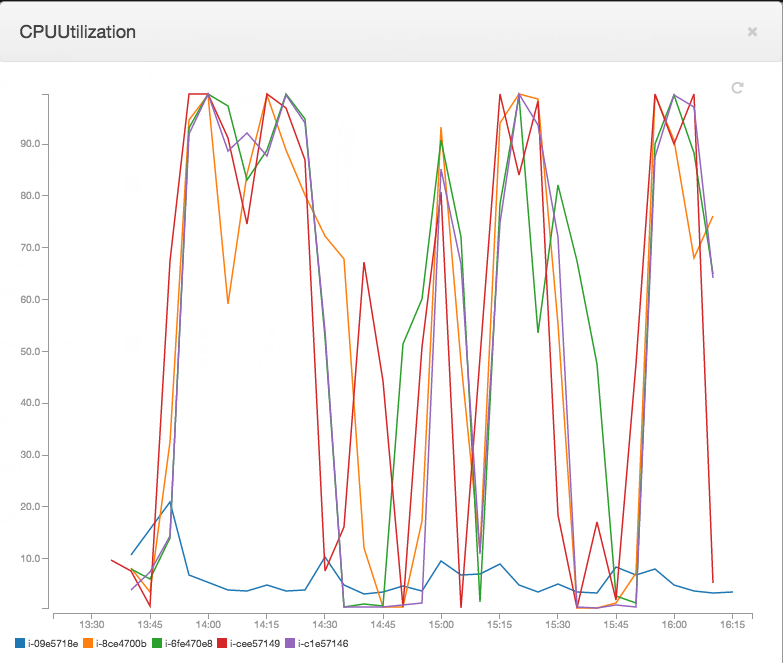


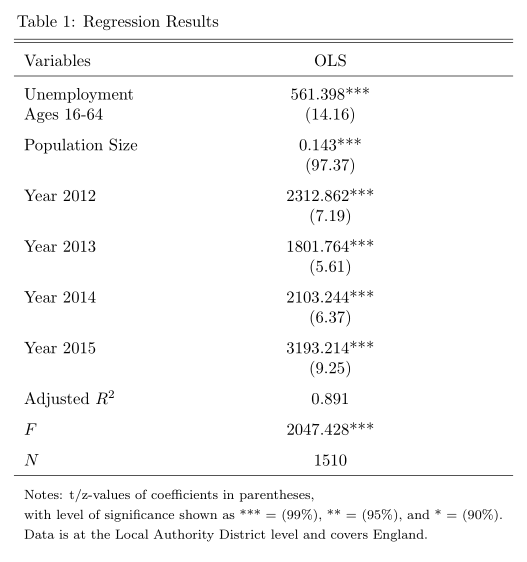
Figure 8 - EMR 4.4, Spark 1.6.0. Most of the time, most nodes are over 90% CPU utilization. Master node shown in blue

**Findings:**

*Statistical Analysis*

In order to directly test the hypothesis that higher levels of unemployment directly lead to higher levels of crime, we used ordinary least squares linear regression analysis using base R. We performed this analysis at the LAD level, and all variables were at the LAD level when this analysis was performed. Total number of crimes in an LAD was the dependent variable. The independent variables were the unemployment rate of people between the ages of 16 and 64, the population size of the LAD, and a dummy variable for each year in the study except 2011 (omitted to avoid collinearity).

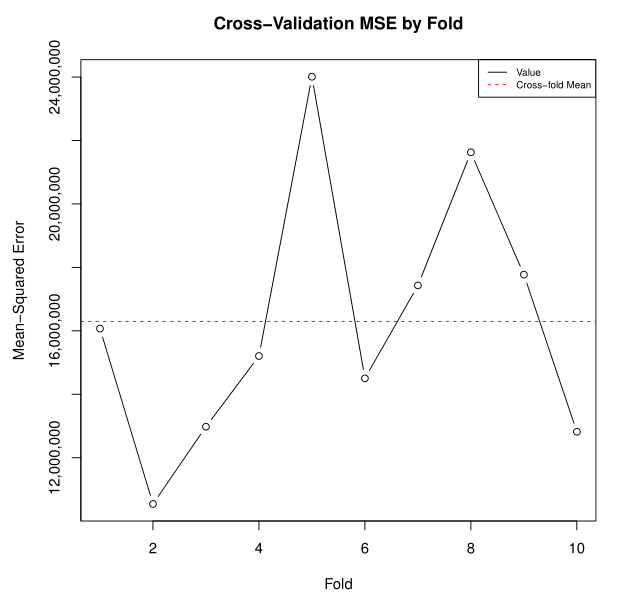
Of course while the unemployment variable was the focus of the analysis, we added other predictors to act as controls. Population size controlled for the large variation in population size (and therefore some of the variation in crime counts) in the LADs. Furthermore, we are not naïve enough to believe that the only predictor of total crime is going to be unemployment, though we do believe it makes methodological sense for it to be one of the strongest. As such, including a Year dummy variable for each year helps us to approximately control for various economic factors that were not readily available, such as yearly mean or median household income, or GDP. I say that it would approximately control for those things because in their absence, the dummy variable for each year is going to be given “credit” by the regression for variation that is actually caused by these missing macro-economic indicators changing between years. Recall that any coefficient of these year dummy variables should be interpreted as the additional change in the crime count for the given year compared to the effect of 2011.



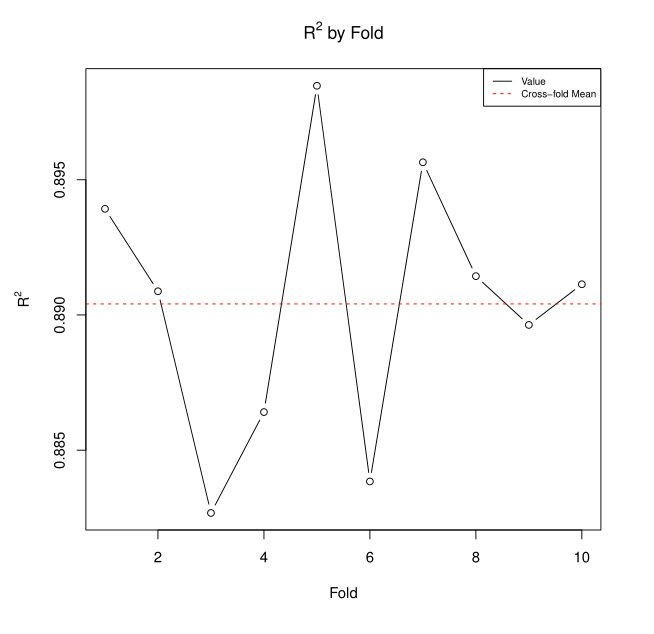
Consider Table 1. As you can see, all of the predictors were statistically significant at a 99% confidence level, and the adjusted R2 value was 0.891, indicating that these 6 predictors can, together, explain approximately 89.1% of the variation in total crimes across local authority districts. It is clear that the model has explanatory power as indicated by the F-statistic, which is far greater than 1. There were 1510 records included in the regression. That is one record per year per local authority district between 2011 and 2015. There were about 150 records that had to be removed prior to running the regression due to missing data.

In order to perform validation on our analysis, we used 10-fold cross-validation. For each fold, we tracked the mean-squared error (MSE), adjusted R2 value, and the value of the coefficients. Through the validation and analysis of our results we found that the model was fairly stable, but that there may be ways that the model could be improved in the future.

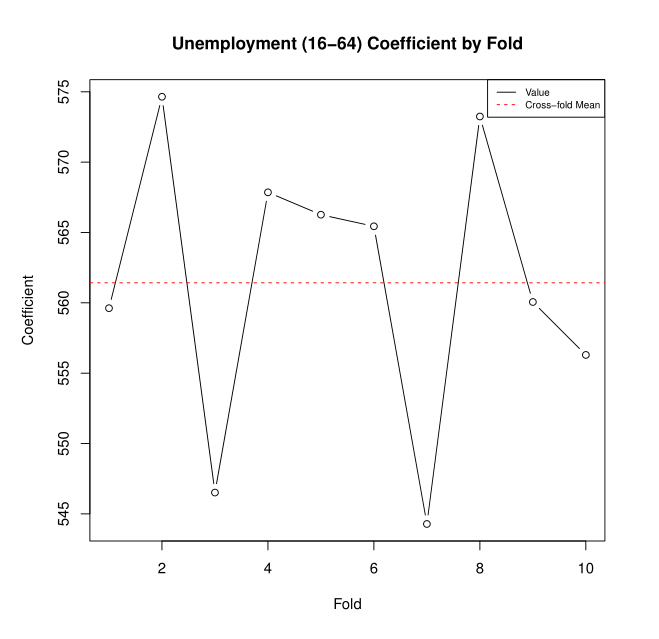
In the graph below, you can see our mean-squared error, which is pretty high, averaging just above 16 million. This indicates that, if we were to use this model to perform prediction, the prediction of crimes for any given LAD would be under or over-predicted by about 4,000 crimes. For larger LADs, that isn’t so erroneous. However, the mean value of total crime in our data is about 14,280 crimes, and the median is just under 10,000 crimes. As a result, we would recommend against using this model for prediction.



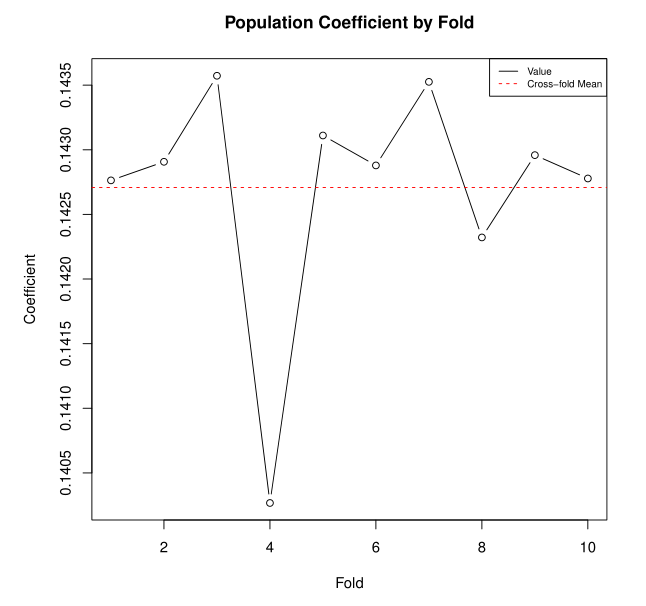
Given the graph below of the adjusted R2, you might think that such a high error rate of the model is unexpected. The adjusted R2 graph indicates that on average (across folds) we explain approximately 89% of the variation in the data. However, there is a way that this could be occurring, and that is due to omitted variable bias, which will be a theme as we continue to look at the cross-validation results and analyze this model. It is what we tried to control for with the year dummy variables.



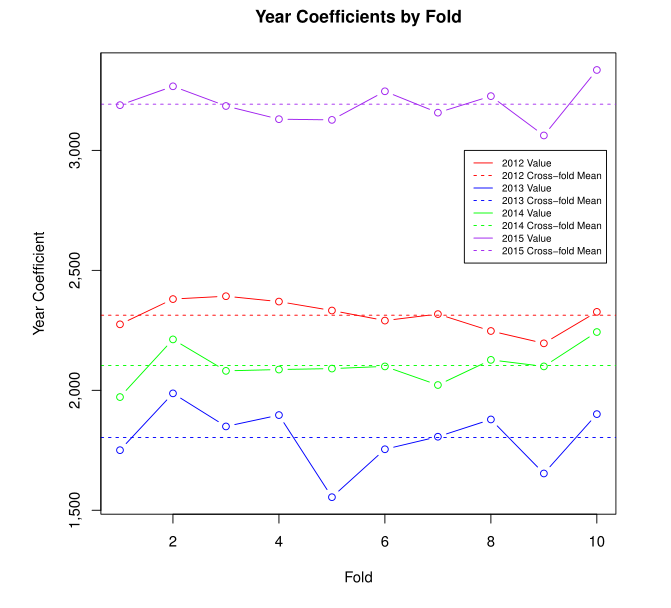
The next three graphs show the stability of the coefficient estimates across folds. Note that although the graphs look relatively variable, the scale on the y-axis is very small, so in reality we are looking at very little variation.



Note that our average coefficient for Unemployment was approximately 562, indicating that for each additional 1% of economically active 16 to 64 year olds that are unemployed, there will be, on average, 562 more crimes.



The coefficient for Population indicates that, ceteris paribus, we would expect another crime for every 7 people in the local authority district. Recall that “crimes” are loosely defined here, as “anti-social behavior” is the most common crime in our data.



Finally, we can take a look at the variance of the coefficients across years and within year across folds. Again, these values are very stable. Recall that these have to be considered in relation to 2011. So if 2011 is considered the baseline, then we could interpret the 2012 year coefficient on the first fold to mean that, on average, in 2012 there are roughly 2,300 more crimes than there were in 2011.

Unfortunately, with such high magnitudes and a non-trivial variance, since we have no methodological reason to believe that one year should be so different from the others, we can be sure that the analysis suffers from omitted variable bias. That is to say, the year variable is clearly acting as a proxy for other variables that should have been included in the regression, were they available. Given the large magnitude of these coefficients and the high variance among them, we would need to be even more careful about predicting values outside of the time range of our data than we are with predictions without our time range.

*Maps*

One of our main goals in working with this dataset was making it easy to visualize our large and disparate dataset. To this end, we made interactive maps (hosted on our S3 bucket website) showing the crime and unemployment rate (Figure 9) by LAD from 2011-2015. The interactive component does not show up in the map below, but the helpfulness of maps of this sort in clearly and quickly communicating a large amount of data is obvious – one glance at the map and you can immediately pick out features like the cluster of high-unemployment LADs around Sheffield. We made the corresponding map to Figure 9 for crime rate, as well as a series of tabbed maps (meaning they all show up on the same web page, you just click a tab to pick the year) for 2011-2015. The tabbed maps unfortunately do not look as nice due to difficulties plotting them over Google Maps, but they are more functional.

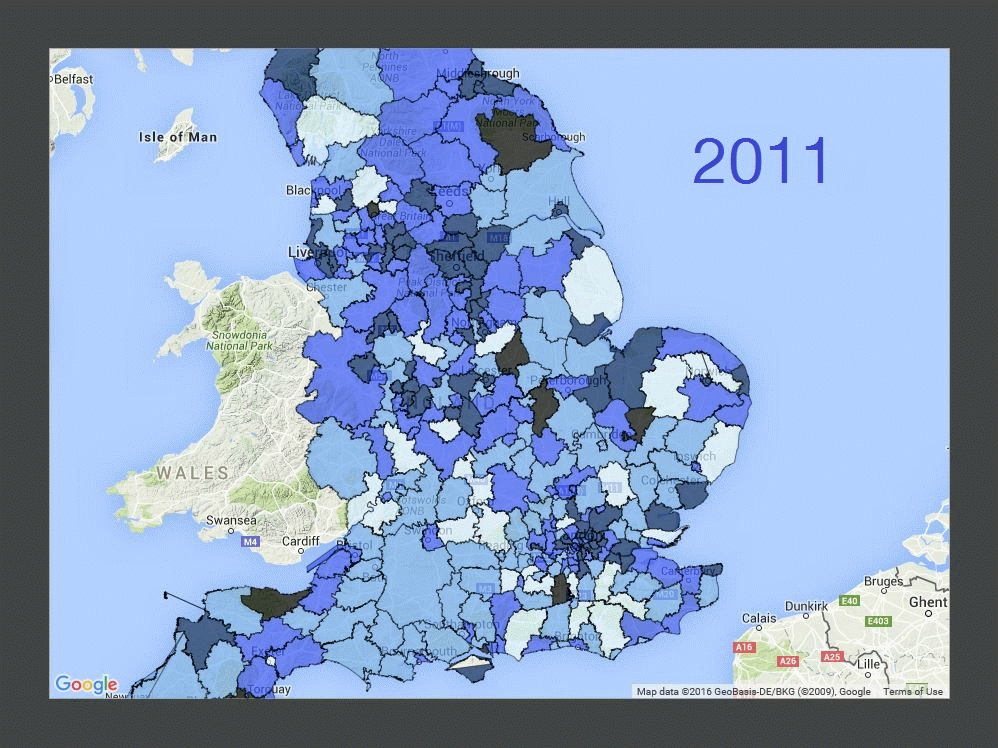
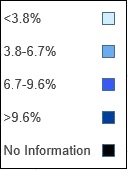


Figure 9 – Unemployment Rate by Local Authority District, 2011

**Conclusion:**

Direct evaluation of the hypothesis through ordinary least squares regression showed that the prevalent unemployment rate between people of 16 and 64 years of age in any given LAD has a strong, significant effect on the total number of crimes in that LAD. However, to speak to the magnitude of the effect or in greater detail about crime numbers in an LAD, further research and data collection would need to be conducted to, among other things, address the clear omitted variable bias.

While the statistical analysis was revealing, a major focus of our project was developing interesting and informative interactive visualizations. In this regard we feel that, for the most part, we achieved what we set out to do. Using Bokeh, we were able to create interactive maps to show changing crime (Figure 9) and unemployment rates over the time frame of our dataset. To some degree, we were able to put the visualizations over a Google Maps background (we were not able to replicate all of the functionality on this background). Considering that we started the project with zero experience producing this sort of visualization, we are proud of what we were able to accomplish in the limited time we had. However, we wish we could have had more time to learn how to add additional features to our maps.

**Opportunities for Future Research:**

While our analysis did bear fruit, it left something to be desired. We had quite a bit of interesting data that, due to time constraints, we ended up not being able to use. For example, we looked at the breakdown by crime type in the exploratory stage, but we were not able to utilize this information in our analysis. This would be an obvious first step in developing a better model of criminal activity in England.

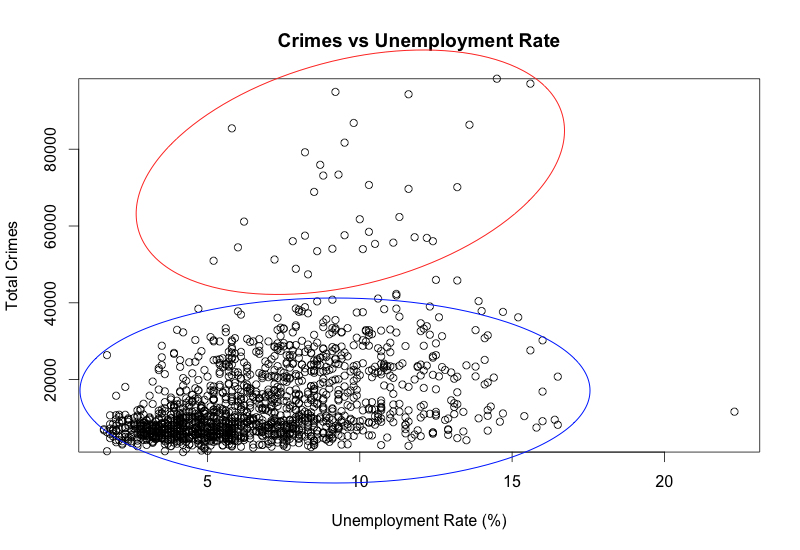
Another area where our project could have been improved was in creating the visualizations. This was the main focus of our project but, unfortunately, we spent too much time trying to make GMapPlot work well (it needs a lot of work before it should be used in any professional capacity), and by the time we switched back to normal figure plots in Bokeh, we didn’t have time left to figure out how to add desirable features like multiple overlays (one for crime, one for unemployment), date selection options, or crime type selection options. In the exploratory analysis phase, we discovered that our dataset contains a distinct cluster of high-crime districts (Figure 10). We did not have time to look into this phenomenon in any great detail. Looking at the clusters individually might have pointed us in the direction of the root causes of high crime levels. 

Figure 10 – In this plot of Total Crime vs. Unemployment Rate, there are distinct clusters of high-crime and low-crime LADs.

Finally, our analysis was limited to the UK. The US is in the midst of a push to open police data to the public, and as a result datasets similar in scope and detail to what we worked with are now available for other areas. Other countries also have public repositories of police data. Conducting the same type of research on countries besides the UK would provide insight into the differences in crime across countries, and might reveal which economic and social factors are associated with higher crime rates.

1. "Crime Map." For Community Policing, City of London Police. Web. 12 May 2016. <https://www.police.uk/city-of-london/cp/crime/>. [↑](#footnote-ref-1)
2. "LSOA Atlas - 2011 Boundaries." L*SOA Atlas - 2011 Boundaries*. Web. 12 May 2016. <http://londondatastore-upload.s3.amazonaws.com/instant-atlas/lsoa-atlas/atlas.html>. [↑](#footnote-ref-2)
3. Farrall, Stephen, and Will Jennings. "Policy Feedback and the Criminal Justice Agenda: An Analysis of the Economy, Crime Rates, Politics and Public Opinion in Post-War Britain." *Contemporary British History* 26.4 (2012): 467-88. Web. [↑](#footnote-ref-3)
4. Jennings, Will, Stephen Farrall, and Shaun Bevan. "The Economy, Crime and Time: An Analysis of Recorded Property Crime in England & Wales 1961–2006." *International Journal of Law, Crime and Justice* 40.3 (2012): 192-210. Web. [↑](#footnote-ref-4)
5. "Data Downloads." Home | Data.police.uk. UK Home Office. Web. 10 Apr. 2016. <https://data.police.uk/data/>. [↑](#footnote-ref-5)
6. "Nomis Official Labour Market Statistics." Home. Office for National Statistics. Web. 10 Apr. 2016. <https://www.nomisweb.co.uk/>. [↑](#footnote-ref-6)
7. "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-7)
8. "Census Support Easy Download: English Boundary datasets." UK Data Service. Web. 10 Apr. 2016. <https://census.edina.ac.uk/easy\_download.html>. [↑](#footnote-ref-8)
9. “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-9)