Suicide Rates in London Boroughs for Targeted Preventative Measures

1604827
Foundations of Data Analytics CS910: Project

December 19, 2019

Abstract

This report aims to understand some of the factors that affect the rates of suicide in London and link it to the population in each borough, to hopefully allow for more targeted suicide prevention measures. Classification and clustering tools have been used to establish relationships between suicide rates and other attributes. Some clear attributes such as sex are found to be key factors affect the rate of suicide and other attributes such as the type of employment also correlate to an increase in the suicide rate, and as hoped area does correlate to the suicide rate as well. While further research would be needed over a large time period to have a higher confidence level in the results, this report shows a clear relation between the boroughs and suicide rates.

Introduction

The ecological associations between suicide rates and different factors that affect these rates has been studied in London for at least 60 years. One of the earliest examples being [Sainsbury(1955)], which showed significant correlations between the rates of suicide and social isolation, as well as other factors that also correlated. The aim of this report is to see if the area in which someone lives in London has an affect on the suicide rates and explore some possible reasons as to why that may be. The hope is that this data analysis could be used to help target suicide prevention measures in areas that have higher rates.

Data

Collection

All the data has been collected from the the London DataStore [DataStore(2019)], which contains many open-source datasets relating to London. Four of these datasets have been compiled together to build the data that will be used for the analysis in this report. The datasets that have been used are:

- Housing Benefit Claimants, Borough
- Qualifications by Economic Activity Status, Borough
- Personal Well-being (Happiness) by Borough
- Suicide Mortality Rates, Borough

All of the datasets have some redundancies, the data used in this report will only have data for the years 2011, 2012, 2013, 2014 this is because these are the only years where all the datasets overlap. As some of the datasets contain information for male, female and the population as a whole there will be three sets of data for each year, in total there are about 500 data points. Data about different parts of the country has also been kept so a comparison between London and the rest of the UK can be made, in order to gain a better understanding of where London stands in regards to the rest of the country.

Housing Benefit Claimants, Borough

This dataset contains the housing benefit claimants by case load and per 100 people aged over 18. The data is categorised by borough and also contains some national data. Only data on the rates per 100 people has been kept due to it being independent of the population of the borough. This data isn't gender specific so has been used in all three sets of data that are used in the analysis, hence it's the same in all of the data sets of that year. This attribute will be most reliable when looking at the population as a whole not specifically by sex.

Qualifications by Economic Activity Status, Borough

This dataset is again categorised by borough and also contains some national information. The data shows the percent of people in a borough who are economically active with NVQ levels 1-4+ and also includes trade apprenticeships, no qualifications and other qualifications. NVQ level 1 is equivalent to GCSE grade D-C, level 2 GCSE A^*-C, level 3 A levels, level 4 any degree (foundation, Bachelors), these are just a few equivalents for example. This data is used to infer the type of employment a borough may have, it is particularly useful for the data on unemployment. The data is split into the three categories, those being male, female and population as a whole.

Personal Well-being (Happiness) by Borough

This dataset contains estimates of personal well-being from the Annual Population Survey Well-being dataset. The data contains estimates of life satisfaction, worthwhile, happiness and anxiety. The data was collected by asking people four questions: Overall, how satisfied are you with your life nowadays? Overall, to what extent do you feel the things you do in your life are worthwhile? Overall, how happy did you feel yesterday? Overall, how anxious did you feel yesterday? The participants where asked to respond with a numeric value between 0 and 10, where 0 is "not at all" and 10 is "completely". The results are given by borough and the mean score summary will be used in the data analysis.

Suicide Mortality Rates, Borough

This dataset contains a table of directly age-standardised rates of suicides per 100,000 population, for age 15+. The suicide rate also includes deaths where the cause couldn't be determined between a self inflicted act or an accident. From this dataset the gender specific data as well as the non-gender specific data will be used in the analysis. This dataset will also be used for the information about population of each of the boroughs it contains.

Hypothesis

This report will attempt to test three hypotheses, that will hopefully give an understanding of some of the key factors that affect suicide and the relation to the borough. The hypotheses being tested are:

1. What are the major factors that affect suicide rates.
2. Is there a relationship between London boroughs and the rate of suicide.
3. Are there any factors that can be be attributed to high rates of suicide in the boroughs.

Preprocessing

Programs used

- **Excel** was used to load and some initial preprocessing as it supports a large number of documentation types and can produce csv files, which can be read by my other analytical tools including ones used in this report.
- **Python** is a program language that can allow for some more in-depth analysis and preprocessing. In this report the pandas data frames package was used, predominantly for the analysis of the data and to build a class variable for further analysis in weka.
- **Weka** provides a graphical and command line interface for data analysis. In this report it has mainly been used for classification and clustering. While also some of its graphical properties especially when looking at the trees it can produce which I've used to infer what attributes are important factors in suicide rates.

Data manipulation

In order to analyse the data and gain insight into the topic the data needed to be preprocessed. The data obtained from the London DataStore is in excels file format and for the ease of data analysis they need to be convert it to a csv file type this can be done in excel. The data was first combined so the relevant data from the datasets was formatted into three files which contains information for just males, females and the population. These can later be combined and used together for various different analysis'. With the csv files pythons panda package was able to be use to adapt the data by making a new classification column based on the data for use in wekas

classifiers. This was done by calculating the mean and standard deviation (std) of the suicide rates per 100000 people and classify everything less than the $mean - 0.45sdt$ as low and everything above $mean + 0.45std$ as high and everything in between as average. 0.45 was chosen as it best split the data into categories with nearly the same number of data points. This was done on the data for each sex and population datasets independently. This method was also applied on all the male and female combined data, this was done to analyse the difference in suicide rate between sexes.

Data cleaning

Excel was also used to remove all data about the city of London borough, this was done as it had many missing values from several of the datasets (over 50% was missing). The datasets also contained some missing values to deal with this they where all set to zero. This was justified as the missing values where only seen in the data from "Qualifications by Economic Activity Status, Borough", where missing values where near zero in all cases. This was shown when summing the percentages (after assuming zero for the missing data) of all employment columns from the qualifications dataset, the sum was very close to 100% which is to be expected, justifying the zeroing of data.

Data Analysis

Testing the first hypothesis, that being if there are any major factors that affect suicide rates, python was used to calculate the correlation between the different all the factors in the datasets as shown in figure 1 below.

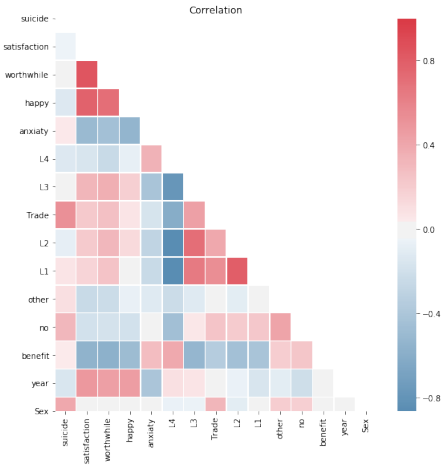


Figure 1: Heat-map of correlations between various numeric categories in the whole dataset

Where "L" followed by a number represents the NVQ level of employment. As can be seen it shows correlation even between factors that this report has no interest in exploring. By looking at the first column only the correlation between suicide and the other factors can be seen. It can be seen there is positive correlation for having no job, sex, working in a trade and also weaker correlation with not having a job with the suicide rate. This method only works with numeric data, as area isn't numeric a correlation can't be found. However as sex is binary it can be converted into a number and allow a correlation to be made.

Python was used to calculate the correlations between suicide rates and each of the attributes. To do this the pandas package has been used, calling the corr() function, to calculate the Pearson correlation. This method calculates a number between -1 and 1. The correlation is a statistical method for measuring the association between two variables, thus can be used to see what factors are closely linked so a further investigation can be made into them. The values of the correlation are given below in the table. As can be seen sex, having a trade employment (i.e. builder) and being unemployed all have a large correlations with suicide rates.

Attribute	correlation
suicide	1.000000
satisfaction	-0.042449
worthwhile	-0.017139
happy	-0.123533
anxiety	0.04842692
L4	-0.132191
L3	0.016425
Trade	0.522237
L2	-0.080756
L1	0.070792
other	0.106682
no	0.309303
benefit	0.041766
year	-0.154374
Sex	0.4010566

Classification and clustering

To further test which factors affect suicide rates an attempt to classify the data will be used, the dataset that will be used contains all of the available data points for male, female and person (the average population of the borough). Using weka's classifiers and clustering algorithms with a 66% training split and trying to classify the data into

the three categories low, average, high rates of suicide generated by python in the data preprocessing.

The following results are obtained: using a Naive Bayes classifier gives an accuracy of 62.3%, logistic regression gives an accuracy of 68.3% and with a J48 tree an accuracy of 69.5%, visualising the tree it is clear that sex has a large influence on the rates of suicide.

To investigate the affect of sex on suicide further the dataset containing only the male and female data will be used. Using the J48 classifier again an accuracy of 75% is now achieved and a 69.6% accuracy with NaiveBayes and when using a nearest neighbour classifier (IBk) with 5 neighbours the accuracy is 70.5%. These are all higher than the results when using the data that contains the data on a person in an area, this again goes to show the importance of sex. If sex has such an important role in the rates of suicide using clustering for classification should also work well, testing this in weka using a K means clustering to predict three clusters (as there are three possible target groups) an accuracy of 72% was achieved.

Testing this to the extreme, weka has the ability to remove columns of data, removing all data except that of sex and attempt to predict the class based solely on the sex attribute. Produce results for the classifiers of 73.2% for all the classifiers that have been previously used on this dataset, this is because it classifies all males as high and females as low. This can be visualised using the J48 tree and also when looking at the confusion matrix for the classifiers.

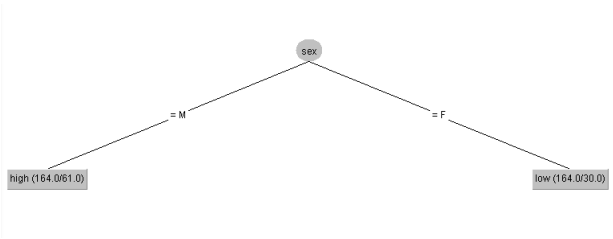


Figure 2: The J48 classifier tree made when using only sex as a predictor

This part of my analysis strongly suggests that your sex has an important part in the rate of suicide, this result is similar to results from many other studies. [Hawton(2018)]

Relation between area and suicide rate

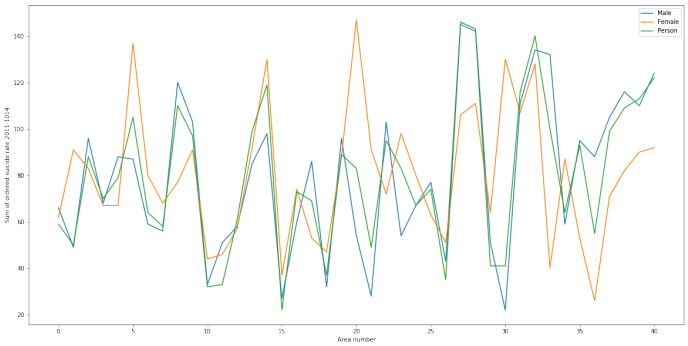
Now moving on to test the second hypothesis, looking at area and its affect on suicide rates. As the area data isn't numeric and can't easily be converted to a numeric value the correlation with suicide can't be made and classification in weka also wouldn't be very affective due to the low number of examples of each location in the data. A different approach will be used. Looking at the data for each of the three datasets (male, female, person), the suicide rates for each year have been ranked in order. Then each area has been given a number between 0 and 40 based on the rate of suicide for that year, the results for the four years of data was then summed. This gives a value of how high the suicide rates of a borough are compared with other the the other boroughs. Areas with high suicide rates over the last 4 years will have a high value this method is better than looking at the data for each year as it reduces the fluctuations caused by the relatively low number of suicides each year. The results are shown below in the heat-maps of London, where darker colours correspond to high suicide rates and lighter to lower rates.

Figure 3: Heat-map of London suicides: Male

Figure 4: Heat-map of London suicides: Female

Figure 5: Heat-map of London suicides: Person

As the heat-maps don't show the exact size of each group a graph to show the exact numbers which where used in creating the heat-maps has been included below with a key:



Key Area name: sum, 'Barking and Dagenham LB': 0, 'Barnet LB': 1, 'Bexley LB': 2, 'Brent LB': 3, 'Bromley LB': 4, 'Camden LB': 5, 'Croydon LB': 6, 'Ealing LB': 7, 'East Midlands': 8, 'East of England': 9, 'Enfield LB': 10, 'Greenwich LB': 11, 'Hackney LB': 12, 'Hammersmith and Fulham LB': 13, 'Haringey LB': 14, 'Harrow LB': 15, 'Havering LB': 16, 'Hillingdon LB': 17, 'Hounslow LB': 18, 'Islington LB': 19, 'Kensington and Chelsea LB': 20, 'Kingston upon Thames LB': 21, 'Lambeth LB': 22, 'Lewisham LB': 23, 'London': 24, 'Merton LB': 25, 'Newham LB': 26, 'North East': 27, 'North West': 28, 'Redbridge LB': 29, 'Richmond upon Thames LB': 30, 'South East ': 31, 'South West': 32, 'Southwark LB': 33, 'Sutton LB': 34, 'Tower Hamlets LB': 35, 'Waltham Forest LB': 36, 'Wandsworth LB': 37, 'West Midlands': 38, 'Westminster, City of LB': 39, 'Yorkshire and The Humber': 40

As can be seen from the maps there are clearly areas that have higher rates than others and they seem to be geographically linked. This can most readily be seen on the persons heat-map where all the centrally located boroughs have high rates gradually decreasing away from the centre. Between the three there are many areas that share common characteristics such as harrow having one of the lowest suicide rates. The data for the three datasets (male, female, person) also is closely linked sharing some key features, such as areas near the centre having higher rates than areas at the outer edges especially, on the west side of London. They also show that areas that have high suicide rates for males also have high suicide rates for females.

From the graph it can be seen that the areas suicide rates measured in this way do correlate well with each other with areas having peaks and troughs in similar places, and that the amplitude is roughly equal.

The data also includes information about the whole of England, preforming a brief analysis of this data and using a similar method outlined above in the creation of the London heat-maps. The data shows that the north has a higher suicide rate than than southern parts of England and that London as a whole has a lower suicide rate than other areas of England. [Kontopantelis and Buchan(2018)]

Further Analysis

Finally investigating the third hypothesis, that there maybe factors that can explain the reasons for the high suicide rates in some areas. Firstly a factor other than sex will need to be identified to have a correlation with suicide. This is because there will be minimal difference in the number of each sex in an area. The data will be analysed to see if there are any social, economic or well-being factors that affect the the rate of suicide.

To do this weka has been used to remove all data about factors already know affect suicide those being sex and area. Now using a wekas J48 classifier to get a general understanding, an accuracy of 58% is achieved. Looking at the tree it becomes apparent that employment with various amounts of education plays a key role in the suicide rate, this is supported by the previous work that showed strong correlations between these factors and suicide rate. It can also be seen as employment factors make up almost all of attributes at the higher branches of the tree.

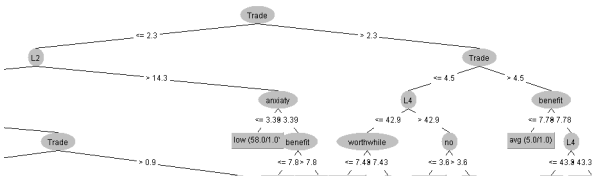


Figure 6: Sample of the J48 classifier tree made when using all attributes other than sex and area

The data contains information on the employment that each area has, using this data weka can be used for its multi-linear regression function. After removing all data except suicide rates and employment data. Weka can be used to preform multi-linear regression, this produces the equation:
$suicide = 0.1713 * L4 + 1.8007 * Trade + 0.1725 * other + 0.7055 * no - 8.4771$
It can be seen that working in a trade has a strong relation with suicide as does having no employment, it also shows that working with a level 4 NVQ or "other" qualifications have a slight increase in suicide rates.

A possible reason for trades having such a high correlation with suicide is that there was very little data on trade apprenticeships and what data was available was male dominated and as discussed earlier being male increases the suicide rate so making it a good predictor. This is supported from my previous analysis done in python, where it was seen that it has the highest correlation with the suicide rate per 100000.

The results from weka show that using just information about the employment can classify with an accuracy of 65.2% for a nearest neighbours algorithm with k=5 and 66% with k=1. Using logistic regression gives an accuracy of 71.4% meaning that the employment someone has does play a important role in the rate of suicide. This rate isn't quite as high as using sex but is still very high. A similar result has been shown in other studies. [Andriy Yuryev(2010)]

Surprisingly the data on well-being doesn't seem to have an affect on the rates of suicide, and is not a good predictor for the rates of suicide only achieving a 43.8% accuracy in prediction with a J48 classifier and 39.3% with a nearest neighbour algorithm with k=5, both of these classifiers have previously preformed well on these datasets. The models are as good as just assigning all of the data to a single class as can be seen from the confusion matrix. Clustering algorithms also don't produce accurate predictions either. This may be due to the relatively small number of suicides per 100000 not affecting the areas well-being as a whole, but further research would be required to prove this hypothesis.

Now applying this knowledge that trade and not being employed are strongly correlated to suicide rates. Investigate if there's a link between areas with high levels of trade employment or unemployment and suicide rates will be done by attempting to match them up with the areas that have high suicide rates. Using the same method as before to make heat-maps of London a comparison of the male suicide rates and male unemployment will be made and can be seen in the heat-maps below. The reason trade wasn't used even though it has a higher correlation is due to the amount of missing data in the female category, its dependency on sex and relatively small percentages of males in that type of employment.

Figure 7: Heat-map of London unemployment: Male

Figure 8: Heat-map of London suicides: Male

Unfortunately the heat-maps don't match up well, with very few areas showing the correlation that was expected. This could be due to the relatively low rates of unemployment in all of the boroughs. Only being able to include one of the jobs attributes may also be a reason behind the disappointing result, again further research would be required on more data to prove this hypothesis.

Conclusion

It is clear that there are several factors involved in the suicide rates one of which being sex and for London at least the area someone lives and what employment they have. In testing the first hypothesis it was found that a major factor in suicide rates is sex, this finding agrees with many other studies on suicide. [Hawton(2018)]. In testing the second hypothesis it was found that the borough of London someone lives in does indeed affect the rate of suicide.

When investigating the third hypothesis it was found that the employment someone has is another important factor that affects the rate of suicide, however it was not shown that linked to the suicide rates in the individual boroughs. However from other studies the same results about employment affecting suicide rates was also shown. [Andriy Yuryev(2010)]

Limitations

A lot of the data that that was collected from the various sources wasn't used as it wasn't compatible with other sources meaning a large amount of time and data was lost. This means there wasn't a huge amount of data to be analysed (only 500 data points), this may potential explain the lack of correlation between figure 7 and figure 8. Also due to the compatibility issue data comes from several years ago, even though some of the datasets had more modern data available. This means that the result obtained in this report may not be as relevant as they once where, so for future studies finding more recent data may prove to be more useful.

The relative rates of suicides is likely to be subject to high levels of variability due to the relatively small number of suicides that occur, using combined data over a longer period of time will produce results with a higher confidence level. However that wasn't possible for this data analysis.

Also not having data separated by sex for all of the variables (i.e. well-being) may have been a contributing factor as to why it showed very little correlation. Having data on all the factors that where investigated separated by sex may improve the strength of my results.

Extensions

Possible extensions of this report would be to include other attributes, that may have an impact on the rates of suicide. As there are many more possible reasons that the boroughs with high suicide rates have them which weren't discussed and no data was gathered on. For instance in another study it was shown that an index of social fragmentation (i.e. private renting, single-person household, unmarried persons and other social factors) [Congdon(1996)] [Rezaeian et al.(2007)Rezaeian, Dunn, Leger, and Appleby] had a stronger correlation between suicides than the indices of socio-economic status. Which could potentially be a factor if an area had a particularly high level of single-person house holds for instance.

Another extension may be to look at the country as a whole in more detail than this study did and try to find the same results that where obtained from the London area but for the entirety of the country.

Applications

As the title of the report says the main aim is to provide insight into the areas of London that have higher suicide rates to hopefully better implement suicide prevention help. An attempt was also made to try and explain why these areas have higher rates to further understand the cause of the higher rates. However the data analysis preformed unfortunately didn't show any relations between factors that where known to increase suicide rates and the boroughs. However knowing which areas have higher suicide rates is still an important way to target the preventative measure to the areas that need them most.

References

[Andriy Yuryev(2010)] Airi Vrnik Andriy Yuryev. Employment status influences suicide mortality in europe, Nov 2010.

[Congdon(1996)] Peter Congdon. Suicide and parasuicide in london: A small-area study - peter congdon, 1996, 1996.

[DataStore(2019)] DataStore. Datastore london, 2019. URL `https://data.london.gov.uk/dataset`.

[Hawton(2018)] Keith Hawton. Sex and suicide: The british journal of psychiatry, Jan 2018.

[Kontopantelis and Buchan(2018)] Evangelos Kontopantelis and Iain Buchan. Disparities in mortality among 2544-year-olds in england: a longitudinal, population-based study, Oct 2018.

[Rezaeian et al.(2007)Rezaeian, Dunn, Leger, and Appleby] Mohsen Rezaeian, Graham Dunn, Selwyn St Leger, and Louis Appleby. Do hot spots of deprivation predict the rates of suicide within london boroughs?, Apr 2007.

[Sainsbury(1955)] P. Sainsbury. Suicide in london, 1955.