Public Health Report
Julian Mack

Your task is to inform the health authorities on the relations between health indicators and socioeconomic and demographic indicators. You will need to describe an analysis strategy, model a particular relation and describe how you would inform the authorities on possible future outcomes.

Part 1

The indicators are grouped into the five categories listed in Table 1. Broadly speaking, the indicators in category D and E are dependent on those in A, B and C. In this report, I will attempt to explore and predict the relationships between the independent variables “ABC” and the dependent variables “DE”.

	Category	**Indicators Include**
A	Our communities	Deprivation, Proportion of children in poverty, Statutory homelessness, 4 GCSE achieved (5A*-C inc. Eng & Maths), Violent crime, Long term unemployment.
B	Children's and young people's health	Breast-feeding initiation, Obese Children (Year 6), Alcohol-specific hospital stays (under 18), Smoking in pregnancy, Teenage pregnancy (under 18).
C	Adults' health and lifestyle	Healthy eating adults, Increasing and higher risk drinking, Adults smoking, Physically active adults, Obese adults.
D	Disease and poor health	Incidence of malignant melanoma, Hospital stays for self-harm, Hospital stays for alcohol related harm, Drug misuse, People diagnosed with diabetes, New cases of tuberculosis, Acute sexually transmitted infections, Hip fracture in 65s and over.
E	Life expectancy and causes of death	Excess winter deaths, Life expectancy – male & female, Life expectancy – female, Infant deaths, Smoking related deaths, Early deaths: heart disease and stroke, Early deaths: cancer, Road injuries and deaths.

Table 1: Indicators segmented by group[1]

Time series data is only available for three indicators: the mortality rates as a result of a) cancer, b) heart disease, c) all causes. As these are all dependent variables, there is no opportunity to investigate how they are affected by the independent variables as time progresses. The rest of the data was collected from a variety of time periods in the years 2006-2013, with the majority in the range 2009 - 2012. This complicates matters as it does not make sense to suggest one measurement is “dependent” on another that took place later in time. However, I will make the assumption that the indicators are static in time and hence their observed values at the point of interest are the same as when they were measured. Clearly, this will not be exactly true in practice but making it will allow us to make progress without massively limiting the scale of the analysis. In the visualizations in Part 3, I only focus explicitly on relationships where the independent variable was measured before or at the same time as the dependent variable.

For this reason, as well as many others, the ABC indicators are not truly independent from the DE indicators. Therefore, all findings should be interpreted in terms of correlation rather than causation.

[1] These labels were obtained according to the official segmentations given here: https://bit.ly/2SSEQHu

My analysis method is as follows:

1. Explore data including with visualizations
2. Create preliminary candidate models, and
3. Use the information gathered in steps 1. and 2. to inform the next steps. Given the open-ended nature of this task, I intend to focus on the relationships that have high predictive power.

As this information will be conveyed to the health authorities, it is very important that the final predictive model is "explainable". As such, I plan to use a Linear Regression or Random Forest rather than a deep learning method[2].

Part 2

In order to prepare the data for analysis the following steps were performed:

1. A PostgreSQL database was created in PGAdmin and hosted locally.
2. The .csv files were parsed and inserted into the database.
3. A database view was constructed to provide fast data access. This would have been useful if the amount of data was considerably greater but, in retrospect, was unnecessary here.
4. Each indicator was manually labeled with a category label A, B, C, D or E, allowing the data to be split into training data X and target data Y.
5. The data was mean-centered and scaled to a unit standard deviation (per feature).
6. The empty fields were filled with the feature's mean. I did this rather than fill with zero because the indicators with missing values are not counts but ratios (making 0 an atypical value). An alternative would have been to exclude the areas with missing indicators, but I decided against this as the amount of data was already small.

These code to achieve these steps is in preprocessing.py and the pipeline is in preprocessing_and_analysis.ipynb.

Part 3

I visualized the data to give guidance on which dependencies were interesting. During this process, I found that there were a small number (< 10) of very extreme outliers. After removing those with Z-score > 4, I found a marked improvement in the performance of my preliminary model. In a regression setting, it is valid to remove outliers as the calculated parameters will be unreasonably affected by these points[3].

Although there is a great deal of noise in the data, there appear to be strong correlations between demographic factors and public health. After investigating other models, I found that multivariate target first-order Linear regression produced the best results. Using, 17 of the provided independent indicators I was able to predict 11 dependent indicators simultaneously with an overall R^2 value of

[2] Neural Networks would also be very difficult to train with such a small amount of data.

[3] If, on the other hand, this had been a classification task, there would no need to remove these samples as they would likely be a long way from the decision boundary.

0.71 when the entire dataset was used and of 0.51 under 10-fold cross-validation (i.e. when tested on unseen data).

It is not-possible to visualize these results in 17x11 dimensions, so I have extracted the predictions, P-values and R-values for a number of interesting relationships in Figure 1. Note that the p-values are 0 to 4.d.p in all cases, suggesting that there is < 0.005% chance that there is zero correlation between these variables. The full list of indicators with good predictive power are given in the Appendix.

The four independent variables with the highest predictive importance were deprivation, whether a mother breast feeds, incidence of long term unemployment and teenage pregnancy under 18 years old. If I was presenting these findings to the authorities, I would create graphs to demonstrate the effect of these indicators on public health outcomes to emphasize the importance of making change in these areas.

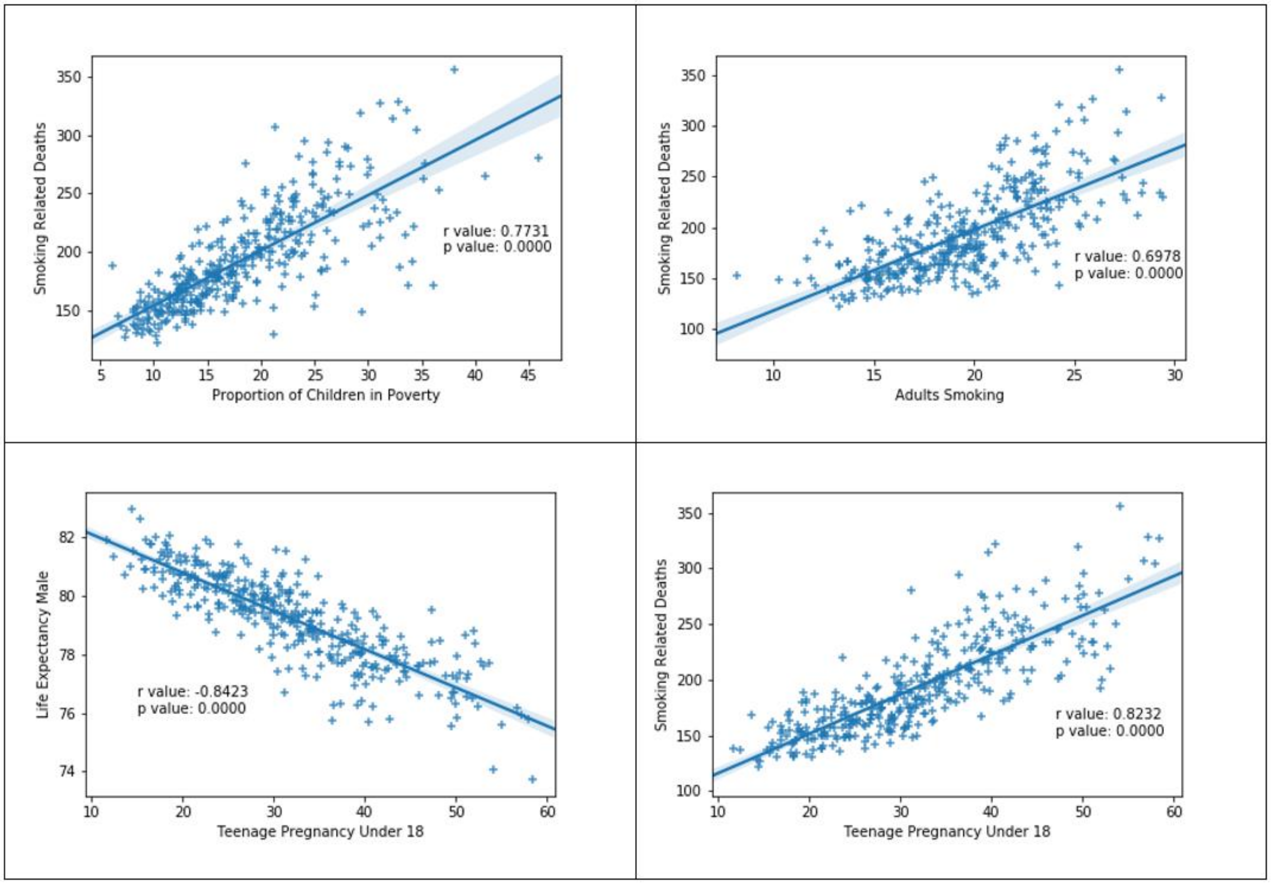

Figure 1: Linear Regression fit for a subset of indicators. R values are the correlation coefficient, and p-values are the probability that there is no correlation between the variables in question. Regression uncertainty values are shaded in blue.

Appendix

Independent Variables in final model	Dependent Variables in final model
Physically active adults, Proportion of children in poverty, adults smoking, average deprived quintile, deprivation, healthy eating adults, increasing and higher risk drinking, least deprived quintile, less deprived quintile, long term unemployment, more deprived quintile, most deprived quintile, obese adults, starting breastfeeding, statutory homelessness, teenagepregnancyunder18, violent crime.	People diagnosed with diabetes, smoking related deaths, acute sexually transmitted infections, drug misuse, early deaths due to cancer, early deaths due to heart disease and strokes, hospital stays for alcohol related harm, hospital stays for self-harm, life expectancy female, life expectancy male, hospital visits that were an emergency for demographic: white.