On June 23rd, 2016, The UK had a national referendum to decide whether the country should leave the EU ('Brexit'). The result, a win for the Leave campaign, surprised many political commentators, who had expected that people would vote to Remain. Immediately people began to look for patterns that could explain the Leave vote: cities had generally voted to Remain, while small towns had voted to Leave. England and Wales voted to Leave, while Northern Ireland and especially Scotland voted to Remain.

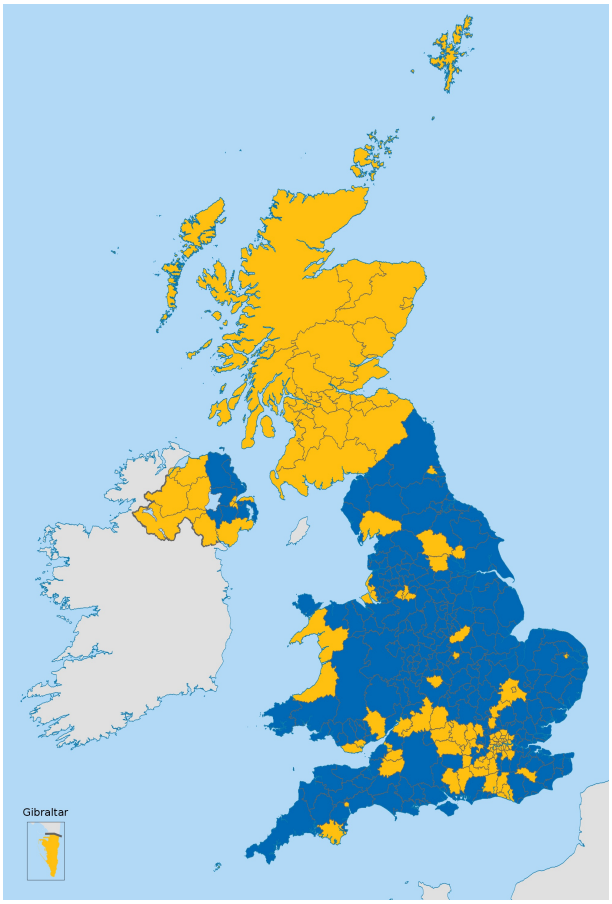

Figure 1: EU referendum vote by electoral ward. Yellow indicates Remain, blue indicates Leave

In the next few days, the Guardian newspaper presented some apparent demographic trends in the vote, based

on the ages, incomes, education and class of different electoral wards (https://www.theguardian.com/politics/ng-interactive/2016/jun/23/eu-referendum-live-results-and-analysis). The Guardian's analysis stopped at showing these results graphically, and commenting on the apparent patterns. In this practical, you will do statistical analysis of the data.

The data from the Guardian's plots can be downloaded from MINERVA (brexit.csv). There are 6 attributes in the data. The 5 possible input variables are:

- abc1: proportion of individuals who are in the ABC1 social classes (middle to upper class)
- medianIncome: the median income of all residents
- medianAge: median age of residents
- withHigherEd: proportion of residents with any university-level education
- notBornUK: the proportion of residents who were born outside the UK

These are normalised so that the lowest value is zero and the highest value is one.

The output variable is called voteBrexit, and gives a TRUE/FALSE answer to the question 'did this electoral ward vote for Brexit?' (i.e. did more than 50% of people vote to Leave?).

Tasks

1. Consider a logistic regression model for the data. Use the model with all inputs to find which input variables are relevant for explaining the output by interpreting the model. Identify the direction and magnitude of each input effect from the fitted coefficients. Which inputs would you say have strong effects? Order the inputs in terms of decreasing effect. Justify your reasoning. Compare your findings with the plots shown on the Guardian website. Do your findings agree with these plots? Comment on your findings.
2. Discuss factors that may affect interpretability of the regression coefficients of the fitted model. Based on your discussion, explain whether you can reliably determine which inputs are relevant for modelling the output and order the input variables based on their relevance in decreasing order. Justify your reasoning.
3. Based on your discussion for Task 2, present and carry out an alternative approach to carry out the analysis for Task 1. Discuss benefits and disadvantages of your approach.