# STAT0023 Computing for Practical Statistics

## In-course assessment 2, take-home component (2022–23 session)

Table of Contents

## Rubric

- Your solutions should be your own work and are to be submitted electronically to the course Moodle page by **12 noon on MONDAY, 24TH APRIL 2023**.
- You can work either alone or in pairs for this assessment. It is up to you to form your own pairs. **You MUST register your choices on Moodle by 12 noon on WEDNESDAY, 22ND MARCH 2023, even if you choose to work alone**.
- If you choose to work in a pair, you will be jointly responsible for the work that is submitted and you will be awarded the same mark.
- Ensure that you electronically 'sign' the plagiarism declaration on the Moodle page when submitting your work. If you choose to work in a pair, both of you should check what has been submitted before signing this declaration: if any plagiarism or collusion is identified with anyone outside your pair, you will share responsibility for it.
- Late submission will incur a penalty unless there are extenuating circumstances (e.g. medical) supported by appropriate documentation and notified within one week of the deadline above. Penalties, and the procedure in case of extenuating circumstances, are set out in the latest editions of the Statistical Science Department student handbooks which are available from the departmental web pages.
- Failure to submit this in-course assessment will mean that your overall examination mark is recorded as "non-complete", i.e. you will not obtain a pass for the course.
- Submitted work that exceeds the specified word count will be penalized. The penalties are described in the detailed instructions below.
- Your solutions should be your own work. When uploading your scripts, you will be required to electronically sign a statement confirming this, and that you have read the Statistical Science department's guidelines on plagiarism and collusion (see below).
- Any plagiarism or collusion can lead to serious penalties for all students involved, and may also mean that your overall examination mark is recorded as non-complete. Guidelines as to what constitutes plagiarism may be found in the departmental student handbooks: the relevant extract is provided on the 'In-course assessment 2' tab on the STAT0023 Moodle page. The Turn-It-In plagiarism detection system may be used to scan your submission for evidence of plagiarism and collusion.
- You will receive feedback on your work via Moodle, and you will receive a provisional grade. **Grades are provisional until confirmed by the Statistics Examiners' Meeting in June 2023**.

---

## Background and overview

On 23rd June 2016, a referendum was held in the UK to decide whether or not to remain part of the European Union (EU). 72% of registered voters took part. Of those, 51.2% voted to leave the EU, and 48.1% voted to remain.

This result was unexpected, and there was extensive commentary on the reasons for it at the time. On 6th February 2017, the BBC News web site carried an article entitled "Local voting figures shed new light on EU referendum" (the article is at http://www.bbc.co.uk/news/uk-politics-38762034). The article is by Martin Rosenbaum, a Freedom of Information specialist at the BBC. He obtained data from 1070 *electoral wards*,[1] giving the numbers of 'Leave' and 'Remain' votes cast in each ward. The Appendix to these instructions provides details of how he obtained the data.

In his article, Martin Rosenbaum calculated some statistical associations between the proportion of 'Leave' votes in a ward, and some of its social, economic and demographic characteristics according to the most recent UK census which was conducted in 2011. He examined characteristics such as education, age and ethnicity taken individually. However, he did not investigate them jointly. This could be important, because there may be other variables that simultaneously influence (say) education level and the propensity to vote 'Leave', and which thereby create the illusion of a causal link between them.

The BBC web page provides the voting data that were used in Martin Rosenbaum's analysis, but not the census data. However, he very kindly shared his census data in response to a request from us at the time − and, for this in-course assessment, they have been supplemented with some additional information as well.

The data are provided to you in the `CSV` file `ReferendumResults.csv` in the 'In-course assessment 2' section of the STAT0023 `Moodle` page. For each of the 1070 electoral wards, this file provides values of around 45 variables that may be relevant in understanding why people voted as they did (the Appendix to these instructions gives a full list of variables, along with other metadata). In addition, for the first 803 wards in the data file, the numbers of 'Leave' votes are provided, as well as the total number of votes for 'Leave' and 'Remain' combined. For the final 267 wards however, the numbers of 'Leave' votes are not provided to you: they are given as `-1` in the data file.

Your task in this assessment is to use the data on the first 803 wards, to build a statistical model that will help you to:

- Understand the social, economic and demographic characteristics that are associated with the voting outcome for a ward; and
- Estimate the proportion of 'Leave' votes in each of the 267 wards for which you don't have this information.

[1] An electoral ward is the smallest administrative division for election purposes in the UK, typically with a population of around 5500. There are almost 9500 electoral wards in the UK.

# Detailed instructions

You may use either `R` or `SAS` for this assessment.

## Your tasks

1. Read the data into your chosen software package, and carry out any necessary recoding (e.g. to deal with the fact that `-1` represents a missing value).
2. Carry out an exploratory analysis that will help you to start building a sensible statistical model to explain and predict the proportion of 'Leave' votes in a ward. This analysis should aim to reduce the number of candidate variables to take into the subsequent modelling exercise, as well as to identify any important features of the data that may have some implications for the modelling. You will need to consider the context of the problem to guide your choice of exploratory analysis. See the 'Hints' below for some ideas.
3. Using your exploratory analysis as a starting point, develop a statistical model that enables you to *predict* the proportion of 'Leave' votes in a ward, based on (a subset of) the ward characteristics; and also to *understand* the variation in proportions of 'Leave' votes between different wards. To be convincing, you will need to consider a range of models and to use an appropriate suite of diagnostics to assess them. Ultimately however, you are required to recommend a single model that is suitable for interpretation, and to justify your recommendation. Your chosen model should be either a linear model, a generalized linear model or a generalized additive model.
4. Use your chosen model to predict the proportion of 'Leave' votes for each of the 267 wards with missing voting data, and also to estimate the standard deviation of your prediction errors.

## Submission requirements

Submission for this assessment is electronic, via the STAT0023 `Moodle` page. You are required to submit three files, as follows:

1. A report on your analysis, not exceeding 2500 words of text plus two pages of graphs and / or tables. The word count includes titles, footnotes, appendices, references etc. – in fact it includes everything except the two pages of graphs / tables and, if present, the separate page describing the contribution of each pair member (see below).

   Your report should be in three sections, as follows:

   I. Introduce the problem context and describe briefly what aspects you considered at the outset, how you used these to start your exploratory analysis, and what were the important points to emerge from this exploratory analysis.
   II. Describe briefly (without too many technical details) what models you considered in step (3) above, and why you chose the model that you did.
   III. State your final model clearly, summarise what your model tells you about the characteristics associated with the proportion of 'Leave' votes, and discuss any potential limitations of the model.

   Your report should not include any computer code. It should include some graphs and / or tables, but only those that support your main points. **Graphs and tables must**

**appear on separate pages**, or they will be not be marked and will contribute to your word count.

In addition to your data analysis, **if you are working as a pair then you must include an additional page at the end of their report where each pair member briefly describes their contribution to the project**. You will need to agree this in your pairs before submitting the report. If both pair members agree that they contributed equally then it is sufficient to write a single sentence to that effect, or alternatively you are very welcome to describe your own personal contribution to the project. Note that this page will not be marked and does not contribute to the word count; nor will different marks be allocated to different pair members based on this. The purpose is to encourage you all to be mindful about contributing to this piece of group-work.

**Your report should be submitted as a `PDF` file named as `########_rpt.pdf`, where `########` is your group ID, with any spaces replaced by underscores (IMPORTANT!!!)**. For example, if your group ID is 'ICA2 Group C', your report should be named `ICA2_Group_C_rpt.pdf`.

2. An R script or SAS program corresponding to your analysis and predictions. Your script / program should run *without user intervention* on any computer with R or SAS installed, providing the file `ReferendumResults.csv` is present in the current working directory / current folder. When run, it should produce any results that are mentioned in your report, together with the predictions and the associated standard deviations. **The script / program should be named `########.r` or `########.sas` as appropriate, where `########` is your group ID with underscores instead of spaces**. For example, if your group ID is 'ICA2 Group C' and you use `R`, your script should be named `ICA2_Group_C.r`.

   You may not create any additional input files that can be referenced by your script; nor should you write any code that requires access to the internet in order to run it. If you use R however, you may use the following additional libraries if you wish (together with other libraries that are loaded automatically by these): `mgcv`, `ggplot2`, `grDevices`, `RColorbrewer`, `lattice` and `MASS`. You may not use any other add-on libraries: for present purposes, an "add-on library" is one that requires a `library()` or `require()` command or equivalent (e.g. the `package::command` syntax) before it can be used, if your `R` system is installed using default settings.

3. A text file containing your predictions for the 267 wards with missing voting data. **This file should be named `########_pred.dat`, where `########` is your group ID with underscores instead of spaces**. The file should contain three columns, separated by spaces and with *no header*. The first column should be the ward identifier (corresponding to variable ID in file `ReferendumResults.csv`); the second should be the predicted proportion of 'Leave' votes for that ward, and the third should be the standard deviation of your prediction error.

**NOTE:** if you work in pairs, **both members of a pair must confirm their submission on Moodle before the submission deadline**.

## Marking criteria

There are 75 marks for this exercise. These are broken down as follows:

- **Report: 40 marks.** The marks here are for: displaying awareness of the context for the problem and using this to inform the statistical analysis; good judgement in the choice of exploratory analysis and in the model-building process; a clear and well-justified argument; clear conclusions that are supported by the analysis; and appropriate choice and presentation of graphs and / or tables. The mark breakdown is as follows:
  - **Awareness of context: 5 marks.**
  - **Exploratory analysis: 10 marks.** These marks are for (a) tackling the problem in a sensible way that is justified by the context (b) carrying out analyses that are designed to inform the subsequent modelling.
  - **Model-building: 10 marks.** The marks are for (a) starting in a sensible place that is justified from the exploratory analysis (b) appropriate use of model output and diagnostics to identify potential areas for improvement (c) awareness of different modelling options and their advantages and disadvantages (d) consideration of the social, economic and demographic context during the model-building process.
  - **Quality of argument: 5 marks.** The marks are for assembling a coherent 'narrative', for example by drawing together the results of the exploratory analysis so as to provide a clear starting point for model development, presenting the model-building exercise in a structured and systematic way and, at each stage, linking the development to what has gone before.
  - **Clarity and validity of conclusions: 5 marks.** These marks are for stating clearly what you have learned about the social, economic and demographic characteristics that are related to the voting outcome in a ward, and for ensuring that this is supported by your analysis and modelling.
  - **Graphs and / or tables: 5 marks.** Graphs and / or tables need to be relevant, clear and well presented (for example, with appropriate choices of symbols, line types, captions, axis labels and so forth). There is a one-slide guide to 'Using graphics effectively' in the Week 1 slides for the course. Note that **you will only receive credit for the graphs in your report if your submitted script / program generates and automatically saves all of these graphs when it is run**.

  **Word and page limits.** You will be penalised if your report exceeds EITHER the specified 2500-word limit or the number of pages of graphs and / or tables. Following UCL guidelines, the maximum penalty is 7 marks, and no penalty will be imposed that takes the final mark below 30/75 if it was originally higher. Subject to these conditions, penalties are as follows:

  - *More than two pages of graphs and / or tables:* zero marks for graphs and / or tables, in the marking scheme given above.
  - *Exceeding the word count by 10% or less:* mark reduced by 4.
  - *Exceeding the word count by more than 10%:* mark reduced by 7.

  In the event of disagreement between reported word counts on different software systems, the count used will be that from the examiner's system. The examiners will use

an R function called `PDFcount` to obtain the word count in your PDF report: this function is available from the Moodle page in file `PDFcount.r`.

- **Coding: 15 marks.** There are 3 marks here for reading the data and handling missing values correctly; 7 marks for effective use of your chosen software (e.g. programming efficiently and correctly); and 5 marks for clarity of your code − commenting, layout, choice of variable / object names and so forth.
- **Prediction quality: 20 marks.** The remaining 20 marks are for the quality of your predictions. Note, however, that **you will only receive credit for your predictions if your submitted prediction file is the same as that produced by your submitted script / program when it is run: if this is not the case, your predictions will earn zero marks**.

  For these marks, *you are competing against each other*. Your predictions will be assessed using the following score:

$$S = \sum_{i=1}^{267} \left[ \log \hat{\sigma}_i + \frac{(Y_i - \hat{p}_i)^2}{2\hat{\sigma}_i^2} \, , \right]$$

  where:

  - $Y_i$ is the actual proportion of 'Leave' votes (which the examiners know) for the $i$th prediction;
  - $\hat{p}_i = \widehat{\mathbb{E}}(Y_i)$ is your corresponding prediction;
  - $\hat{\sigma}_i$ is your quoted standard deviation for the prediction error.

  The score $S$ is an approximate version of a *proper scoring rule*, which is designed to reward predictions that are close to the actual observation and are also accompanied by an accurate assessment of uncertainty (this was discussed during the Week 10 lecture, along with the rationale for using this score for the assessment). Low values are better. The scores of all of the students in the class (and the lecturer) will be compared: students with the lowest scores will receive all 20 marks, whereas those with the highest scores will receive fewer marks. The precise allocation of marks will depend on the distribution of scores in the class.

  If you don't supply standard deviations for your prediction errors, the value of $\hat{\sigma}_i$ will be taken as $1/2$ for all of your predictions: this is the largest possible standard deviation for any random variable taking values between 0 and 1, and the value of $S$ will be correspondingly large so that you will receive few if any marks for your predictions.

## Hints on tackling the assessment

1. There is not a single 'right' answer to this assignment. There is a huge range of options available to you, and many of them will be sensible.
2. You are being assessed not only on your computing skills, but also on your ability to carry out an informed statistical analysis: material from other statistics courses (in particular STAT0006, for students who have taken it) will be relevant here. To earn high marks, you need to take a structured and critical approach to the analysis and to demonstrate appropriate judgement in your choice of material to present.

3. At first sight, the task will appear challenging to many of you. However, there is a lot that we already know: Martin Rosenbaum's article is an obvious starting point. You may also want to search for other commentaries on the UK referendum result, to gain some understanding of what kinds of relationships you might look for in the data.

4. When building your model, you have two main decisions to make. The first is: should it be a linear, generalized linear or generalized additive model? The second is: which covariates should you include? You might consider the following points:

   - **Linear, generalized linear or generalized additive?** This is best broken down into two further questions, as follows:

     - *Conditional on the covariates, can the response variable be assumed to follow a normal distribution with constant variance?* In this assignment, the response variable is a proportion and therefore cannot have exactly a normal distribution. However, there are thousands of votes in each ward: the Central Limit Theorem may apply, therefore, so that the response distribution has *approximately* a normal distribution – in which case you may judge that the approximation is adequate for your purposes.

       The 'constant variance' assumption may also be suspect: given that the response is a proportion, you might think that a binomial distribution would be appropriate, but the variance of a binomial proportion is $p(1-p)/n$ in an obvious notation. Since this depends on $p$, and $p$ varies between wards, the variance cannot be constant. Whether this is a problem depends on how much the 'Leave' probability $p$ varies: if it doesn't vary much, then you may wish to claim that the variance is approximately constant. If it varies a lot however, then you could probably improve your predictions (and hence your score!) by accounting for it. You might consider using your exploratory analysis to gain some preliminary insights into this point.

     - *Are the covariate effects best represented parametrically or nonparametrically?* Again, your exploratory analysis can be used to gain some preliminary insights into this. You may want to look at the material from week 6, for examples of situations where a nonparametric approach is needed.

   - **Which covariates?** The data file contains many potential covariates, some of which are more important than others. You have many choices here, and you will need to take a structured approach to the problem in order to avoid running into difficulties. The following are some potentially useful ideas:
     - *Look at other published commentaries on the referendum result*. What measures are considered useful? Can these be linked to covariates for which you have information? Obviously, if you do this then you will need to acknowledge your sources in your report.
     - *Define useful summary measures on contextual grounds, and work with these*. For example, 16 of the potential covariates in the data file are percentages of the population in different age categories (0 to 4, 5 to 7, … , all the way up to '90 plus'). You may decide just to work with 'young voters' (18 to 29 – 18 is the minimum voting age in the UK), 'working age' (30 to 64 say) and 'retirement age'

(65 and above). Or, indeed, to adopt your own categories − the results are unlikely to be sensitive to the *precise* definitions. Similar comments apply to the potential covariates representing ethnicity, household deprivation and so on.

- *Define new variables based on the correlations between the existing variables, and work with these*. If several continuous variables are highly correlated, then it is difficult to disentangle their effects and it may be preferable to work with a single 'index' that combines all of them. This is the basis of techniques such as Principal Components Analysis, that were discussed during the Week 10 lecture (along with how to implement them in `R` and `SAS`).

You should not start to build any models until you have formed a fairly clear strategy for how to proceed. Your decisions should be guided by your exploratory analysis, as well as your understanding of the context.

5. Don't forget to look for interactions! For example, one of the variables in the data set is `RegionName`, which is a factor (i.e. categorical covariate) indicating the UK region in which each ward is located. Possibly there is regional variation in the strength of dependence between other characteristics and the proportion of 'Leave' votes. Look at the analysis of the iris data from Workshop 2, for a similar kind of situation.

   Sometimes people get confused about the difference between interactions and collinearity. **Reminder:**

   - An *interaction* describes the way in which covariates must be considered in combination to characterise their relationship with the response variable.
   - By contrast, *collinearity* is just about correlations between the covariates: this has no reference to the response variable. Collinearity just makes it harder to identify which covariates are genuinely associated with the response (recall the "sheep energy" example from Week 9).

6. You probably won't find a perfect model in which all the assumptions are satisfied: models are just models. Moreover, you should not necessarily *expect* that your model will have much predictive power: maybe the covariates in the data set just don't provide very much useful information. You should focus on finding the best model that you can, therefore − and acknowledge any deficiencies in your discussion.

7. To obtain the standard deviations of your prediction errors, you need to do some calculations. Specifically:

   i. Suppose $\hat{p}_i = \widehat{\mathbb{E}}(Y_i)$ is your predicted probability of voting 'Leave' for the $i$th ward, and that $Y_i$ is the corresponding actual value.
   ii. Then your prediction error will be $Y_i - \hat{p}_i$.
   iii. $Y_i$ and $\hat{p}_i$ are independent, because $\hat{p}_i$ is computed using only information from the first 803 wards and $Y_i$ relates to one of the 'new' wards.
   iv. The *variance* of your prediction error is thus equal to $\text{Var}(Y_i) + \text{Var}(\hat{p}_i)$.
   v. You can calculate the standard error of $\hat{p}_i$ in both `R` and `SAS`, when making predictions for new observations − see Workshops 6 and 9. Squaring this standard error gives you $\text{Var}(\hat{p}_i)$.

vi. You can estimate $\text{Var}(Y_i)$ by plugging in the appropriate formula for your chosen distribution − for example, if you're using a binomial distribution then $\widehat{\text{Var}}(Y_i) = \hat{p}_i(1-\hat{p}_i)/n_i$, where $n_i$ is the number of votes for the $i$th ward.

vii. Hence you can estimate the standard deviation of your prediction error as $\hat{\sigma}_i = \sqrt{\widehat{\text{Var}}(Y_i) + \text{Var}(\hat{p}_i)}$. In fact, for the case of linear models this is exactly the calculation that is used in the construction of prediction intervals (see your STAT0006 notes or equivalent).

---

# Appendix: the `ReferendumResults.csv` data set

## Data sources

The data provided in `ReferendumResults.csv` are from several different sources, as follows:

### Martin Rosenbaum's BBC article (data source MR1)

The article at http://www.bbc.co.uk/news/uk-politics-38762034 provides an `Excel` spreadsheet, containing localised voting data supplied to the BBC by councils which counted the EU referendum. Results are provided for all individual wards where data were available at this level of detail: there are 1283 such wards, of a total of 9291 wards in the UK. Reasons for the figures not being available at the remaining wards are:

- Three councils did not respond to the BBC's request.
- Some councils refused to give the information to the BBC.
- For some councils, ballot boxes were mixed before counting so it was not possible to identify the precise numbers of votes in each ward.

**Important caveat**: in many wards, some postal votes were mixed in prior to counting. The BBC spreadsheet states "Figures which include postal votes cannot be treated as exact. However broad patterns can still be identified in the data."

### Martin Rosenbaum's data set from the 2011 UK census (data source MR2)

The variables in this data set are those that form the basis for the analysis reported in Martin Rosenbaum's article. He provided the following information when supplying the data to us:

*'All the 2011 census data was downloaded via selecting datasets at https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp?opt=3&theme=&subgrp=. (I calculated adult mean age from the raw counts of adults of each age in each ward). Please note that some areas have seen boundary changes to wards since 2011, so some wards with referendum voting data do not figure in this list.'*

This spreadsheet contains data for 8570 wards.

### The UK Register of Geographic Codes (data source RGC)

Geographical information for each ward was obtained from the UK 'Register of Geographic Codes', downloaded on 14th March 2017 and located by searching at http://geoportal.statistics.gov.uk/.

### UK age structure by ward, from the 2011 UK census (data source ASW)

Martin Rosenbaum's census data contains information on the mean adult age in each ward, but it is possible that more detailed information on age profiles would be useful. Percentages of population in different age bands were obtained for the 2011 census, from https://www.nomisweb.co.uk/query/select/getdatasetbytheme.asp?collapse=yes under *Census 2011*, *Key Statistics* and then *Age Structure*. This provides information on the same 8570 wards that are present in data source MR2.

## Data processing

The data sources have been combined in the following way to create `ReferendumResults.csv`:

1. The spreadsheets from sources MR1, MR2 and ASW were merged using the nine-digit ward identification code (identified as `WardCode` in MR1). The `Remain` variable, giving the number of 'Remain' votes in each ward, was replaced by an `NVotes` column giving the total number of 'Leave' and 'Remain' votes. There are 1070 wards remaining after this merge: this decrease from the original 1283 wards in MR1 is due to the exclusion of wards for which the boundaries changed between the 2011 census and the 2016 referendum.
2. Source RGC was used to identify the administrative area type and region name for each ward, again based on its nine-digit identification code. Some ward codes were found to be duplicated in source RGC, but in all cases the administrative area type and region name were identical for the duplicates.
3. The rows of the data table were randomly shuffled, so that the order of wards no longer corresponds to that in any of the data sources. This was done in order to prevent 'cheating' when making predictions.
4. A subset of 267 wards was identified for the 'prediction' part of the assessment. This was done in such a way that the distributions of all of the covariates in this subset is very similar to the distribution in the remaining 803 wards. Specifically:
   a. In each region, 25% of the wards were sampled at random as candidates for making predictions. This sample will be referred to as 'Group 2' below, with 'Group 1' comprising the remaining wards.
   b. For each of the numeric covariates in the data set, a Kolmogorov-Smirnov test was performed to test the null hypothesis that the underlying distributions in Groups 1 and 2 are the same.
   c. The prediction sample was accepted only if the $p$-values for *all* of the Kolmogorov-Smirnov tests were greater than 0.5 (this is not a typo). Otherwise, a new candidate sample was drawn in step (a) and the procedure was repeated.

   The Kolmogorov-Smirnov test is used here as a convenient way to measure whether two distributions are similar: the use of a high $p$-value threshold is chosen to ensure that the

resulting Groups 1 and 2 are very well balanced with respect to all of the covariate values. Note, however, that the voting numbers were *not* included in this balancing exercise: this is because the performance of predictions would be artificially enhanced if the voting numbers were included (for example, we would know that the distribution of Group 2 voting proportions is similar to that of Group 1 proportions). Note also that no attempt has been made to balance the groups in terms of *combinations* of the covariates.

5. The 'Group 2' rows were placed at the end of the data table, with their 'Leave' vote numbers set to `-1`; a new `ID` variable was created so that each ward has an ID number between 1 and 1070; the original nine-digit ward identification code was removed so that the wards cannot be identified easily using online information; a few other redundant covariates were removed (e.g. where one variable was the sum of two or more others); some variables were renamed for ease of interpretation; and the columns were re-ordered for convenience.
6. Some (but not all) of the covariates were multiplied by random numbers close to 1. This makes no difference to any models that you fit, because the corresponding regression coefficients will scale correspondingly; but it makes it even more difficult for you to identify the wards using online information.

## Description of variables

This section gives a brief description of each of the variables in `ReferendumResults.csv`, and an indication of which data source it came from. Descriptions are compiled on the basis of correspondence with Martin Rosenbaum and online documentation, including the UK Census User Guide (notably Part 4 of the 'Variables and Classifications' section, together with the Glossary) at https://www.ons.gov.uk/census/2011census/2011censusdata/2011censususerguide/.

| Variable name | Source | Description |
|---|---|---|
| `ID` | – | Ward ID number, from 1 to 1070 |
| `AreaType` | RGC | Type of administrative area in which the ward is situated. This takes one of four values: `E06`, `E07`, `E08` and `E09` representing 'unitary authorities', 'non-metropolitan districts', 'metropolitan districts' and 'London boroughs' respectively. This is included because `E08` and `E09` indicate a large urban area. |
| `RegionName` | RGC | Name of the region within which the ward is situated. The possible values are `North East`, `North West`, `Yorkshire and The Humber`, `East Midlands`, `West Midlands`, `East of England`, `London`, `South East` and `South West`. |
| `NVotes` | MR1 | Total number of votes ('Leave' plus 'Remain') |
| `Leave` | MR1 | Number of 'Leave' votes |

| Variable name | Source | Description |
|---|---|---|
| Postals | MR1 | Indicates whether postal votes were mixed in with the data prior to counting. P indicates yes, NP indicates no. |
| Residents | MR2 | Number of permanent residents |
| Households | MR2 | Number of households |
| MeanAge | ASW | Mean age of permanent residents |
| AdultMeanAge | MR2 | Mean age of adult permanent residents |
| Age_0to4 | ASW | % of permanent residents aged 0-4 |
| Age_5to7 | ASW | % of permanent residents aged 5-7 |
| Age_8to9 | ASW | % of permanent residents aged 8-9 |
| Age_10to14 | ASW | % of permanent residents aged 10-14 |
| Age_15 | ASW | % of permanent residents aged 15 |
| Age_16to17 | ASW | % of permanent residents aged 16-17 |
| Age_18to19 | ASW | % of permanent residents aged 18-19 |
| Age_20to24 | ASW | % of permanent residents aged 20-24 |
| Age_25to29 | ASW | % of permanent residents aged 25-29 |
| Age_30to44 | ASW | % of permanent residents aged 30-44 |
| Age_45to59 | ASW | % of permanent residents aged 45-59 |
| Age_60to64 | ASW | % of permanent residents aged 60-64 |
| Age_65to74 | ASW | % of permanent residents aged 65-74 |
| Age_75to84 | ASW | % of permanent residents aged 75-84 |
| Age_85to89 | ASW | % of permanent residents aged 85-89 |
| Age_90plus | ASW | % of permanent residents aged 90 and above |
| White | MR2 | % of permanent residents self-identifying as white |
| Black | MR2 | % of permanent residents self-identifying as black |
| Asian | MR2 | % of permanent residents self-identifying as Asian |
| Indian | MR2 | % of permanent residents self-identifying as Indian |
| Pakistani | MR2 | % of permanent residents self-identifying as Pakistani |
| Owned | MR2 | % of households owning their accomodation |
| OwnedOutright | MR2 | % of households owning their accomodation outright (i.e. with no mortgage) |

| Variable name | Source | Description |
|---|---|---|
| SocialRent | MR2 | % of households renting from social landlords, housing associations, charities and similar |
| PrivateRent | MR2 | % of households renting from private landlords |
| NoQuals | MR2 | % of permanent residents with no academic or professional qualifications |
| L1Quals | MR2 | % of permanent residents with only 'Level 1' qualifications (1–4 GCSEs or equivalent – a GCSE being a qualification that is usually gained at age 16) |
| L4Quals_plus | MR2 | % of permanent residents educated to the equivalent of degree level or above |
| Students | MR2 | % of permanent residents who are students |
| Unemp | MR2 | % of permanent residents who are unemployed |
| UnempRate_EA | MR2 | % of economically active residents who are unemployed |
| HigherOccup | MR2 | % of permanent residents in 'higher-level' occupations according to the 2011 census definition |
| RoutineOccupOrLTU | MR2 | % of permanent residents in 'routine' occupations or who are long-term unemployed |
| Density | MR2 | Population density (permanent residents per hectare) |
| Deprived | MR2 | % of households that are 'deprived' in at least one of four dimensions, according to the 2011 census definition |
| MultiDepriv | MR2 | % of households that are 'deprived' in at least two dimensions |
| C1C2DE | MR2 | % of households in 'social grades' (as defined by the UK Office for National Statistics) C1, C2, D and E, where the grades are defined as follows: |
| | | C1: Supervisory, clerical and junior managerial, administrative, professional occupations |
| | | C2: Skilled manual occupations |
| | | DE: Semi-skilled and unskilled manual occupations, unemployed and lowest grade occupations |
| C2DE | MR2 | % of households in social grades C2, D and E |
| DE | MR2 | % of households in social grades D and E |