

Advanced Data Analytics Coursework Specification

Spring 2023

1. Overview

The goal of this coursework is to give you experience of the whole lifecycle of carrying out a full visual analytics project.

Your goals are:

- To follow a sound visual analytics process
- To develop a visualisation that displays important features of a dataset
- To write a clear report on your findings.

The outputs from this work should be

1. a Tableau dashboard and associated worksheets (as a packaged workbook: see https://help.tableau.com/current/pro/desktop/en-us/save_savework_packagedworkbooks.htm);
2. a written report with sections as defined below.

The submission deadline is **13:00 on Wednesday 24th May** through Blackboard: create a single zip file containing all the files in your submission. This coursework is worth 60% of the marks for the unit.

2. Task Details

The task you are asked to carry out for the coursework is to design, construct, and evaluate an exploratory analysis of a complex dataset using both information visualisation *and* data projection. This dataset should be based on census data for England and Wales. You should design the visualisation to address some socio-economic issue that is important to you.

You must submit at least two data projections using different algorithms. My expectation is that you will do this work in Python (following the methods you have practiced in the labs) and for each projection, create a matrix with two columns representing the two variables the data is projected onto. If you save this matrix in a file (e.g. CSV format) it can then be imported easily into Tableau and used in your visualisations. I would like to review the Python code used to generate the projections, so please include it in your submission. The purpose of data projection is to show the data structure: clusters, outliers, relationships between different labels.

You may use data taken from the 2011 census in England and Wales which is indexed by the Excel file 2011CensusIndexofTablesandTopics_v11_4_2.xlsx. The tab labelled 'All Tables' provides a list of tables and links to the underlying data. (I have found that the Excel file links are valid, the NESS links don't work as the server can't be found, and the links to NOMIS take you to a website where additional data can be downloaded.) You may find Tableau's Data Interpreter useful, and you may also need to edit some files to create usable datasets.

There are more than 1600 tables in total: clearly this is far too many to create an interesting report. You should focus on a limited number of tables (probably around three or four) that allow you to explore a particular aspect of socio-economic life in England and Wales: for example, health and links to nationality or occupation.

A new census was carried out in 2021 (during the pandemic). Some of the results have been released by the Office for National Statistics, but so far these have only been in certain topics. A link to the topics that have been released can be found here <https://census.gov.uk/census-2021-results/phase-one-topic-summaries>. You should find that you can click through on a topic to a map display <https://www.ons.gov.uk/census/maps> and from here select a topic such as 'Housing'. Selecting a variable changes the map and also

provides a link to download the data for that variable. Perhaps simpler is to visit the bulk downloads page https://www.nomisweb.co.uk/sources/census_2021_bulk

You may use either the 2011 data, 2021 data, or both (e.g. if you wanted to show changes over time).

Your report should contain the following sections:

- Abstract. A brief description of the key points in the report.
- Introduction. The background of the problem.
- Data Preparation and Abstraction. Describe the data manipulation necessary to create a dataset for analysis and the principal data types and semantics that you have analysed.
- Task Definition. A description of the tasks using Munzner's task taxonomy for which you have created the visualisations. N.B. many reports last year did not use the taxonomy and lost marks accordingly.
- Visualisation Justification. Define the visualization techniques you use and provide a justification for your choices. You should refer to the principles of info vis, relevant aspects of human perception and cognition, and the scientific literature where appropriate. You should also explain why you have chosen the data projection methods that you have used. This justification and explanation is a very important assessment criterion, so do not skimp on this and make sure that it is grounded in the theoretical concepts we have covered during the course.
- Conclusion. I expect you to address two aspects.
 - What you have learned about the socio-economic problem that was the basis of the visualization.
 - What you have learned about information visualisation from doing the coursework.

I am expecting the report to be about six to ten pages in length. This is an expectation, not a strict limit, so there will be no penalty for exceeding it. But if you find yourself writing much more than this, you are almost certainly providing too much detail. In particular, note that I will see the visualisation you generate, so there should be little or no need for screenshots.

I use the term 'dashboard' in the Tableau sense of a set of visualisations on a single screen. It is permissible to submit more than one Tableau dashboard or workbook if that supports the task better. Do not feel you have to squeeze everything onto a single dashboard. You may remember the system for visualising American census data that had every possible graph interacting in lots of ways. It was just too crowded and complex to be useful.

Geocoding issues

It can be hard to plot the census data in Tableau because it does not contain outcode information. This [blog](#) contains some geocoding packages and a video on how to use them that support geographic information at many different levels of granularity. It should be helpful for you.

You may have some problems with using geocoding packages, in which case this link to Tableau help should be useful.

<https://kb.tableau.com/articles/issue/error-the-custom-geocoding-folder-has-errors-when-creating-map>

I have also provided a short guidance note written by Joshua Ramini on the Blackboard site.

3. Assessment

The assessment criteria are:

- Problem understanding: how well you have explained the goals of the tasks, taking account of end-user requirements. (10 marks)
- Data preparation and task analysis: care taken over extracting and manipulating the data; insights gained through the task analysis. (15 marks)

- Data visualisation: appropriateness of visualization and modelling approaches; systematic use of statistical and visualisation methods; justification of visualization approach used. (50 marks)
- Conclusions: what the user should learn from your analysis and what you have learned about larger-scale data visualisation. (15 marks)
- Presentation: fluency and coherence of the written text; quality of images and graphics used. (10 marks)

You may also find it helpful to read the general feedback that I gave on a similar coursework last year.

- Ensure that questions you set out to ask are answered by the visualisation and in the report.
- Having the option of switching between absolute values and proportions is often a useful feature. This is particularly helpful when comparing areas with different populations.
- When using dimensionality reduction it is important to communicate to the user which variables were used in the original data space as otherwise it is hard to interpret the plots.
- Tooltips should identify the corresponding point (e.g. a location) particularly for projected data.
- The introduction should contain some discussion of the type of user the visualization is intended for.
- The report should note data anomalies (e.g. missing values) in report, in particular, quantifying the number of missing values etc.
- The abstract should describe the main findings of the work.
- Data cleaning matters.
- The use of section and page numbers helps the reader to navigate the report.
- References to secondary literature are valuable tools to provide context.