Introduction to Data Science
Coursework 2
Deadline: Monday August 9th, 12 noon
Submission: EBART

This assessment is worth 40% of the overall mark. This is an individual assessment. You are reminded of the University's regulations on plagiarism.

Brief

In this coursework you are asked to explore a large dataset of job adverts collected from the Reed.co.uk API. Most of the tasks below should be possible to accomplish using the skills you have learned in this course. The assignment tests your ability to apply your skills and knowledge of data science to an unseen dataset. You may also need to do some self-guided learning to work out how to complete some of the tasks – this is an important part of data science in the real world!

There are several data analysis and critical reflection tasks to complete. Your submission should be in the form of a single pdf containing code, results, figures and text where appropriate. Do not simply copy paste your code into a document, include only important code snippets, the majority of the document should be text and images. Please separate and label the different tasks clearly to aid assessment.

Dataset

The dataset you will use consists of job adverts collected from the Reed.co.uk API during the period 19/12/2019 to 11/01/2021.

Each file contains a job ad. Ads are stored as JSON objects.

Files can be downloaded at the link below (you will need to login via your University account):
https://universityofexeteruk-my.sharepoint.com/:f:/g/personal/r_arthur_exeter_ac_uk/EvKwLGPrEfhAvPxIvq5_868BXBE8h3AqSdwR-GstSUTTWA

If you cannot access the data please contact me: r.arthur@exeter.ac.uk

Hints for data processing:

- Look at the *json* module in Python for processing the ads.
- One of the challenges with this coursework is handling large data files. You may wish to think about your strategy. The JSON files contain a lot of redundant information. Maybe make smaller JSON or CSV files by writing out just the fields you are interested in from the raw JSON files? Think about what you will need to do with the data to complete the coursework. A good strategy will require less storage and make the files quicker to process.
- Don't use the entire dataset to debug your code.

Tasks

Part 1. Basic Stats (40 marks)

1. Count the total number of adverts. [5 marks]
2. Count the number of ads for full-time and part-time jobs. [5 marks]
3. Plot a time-series of the number of ads by day. Comment on what you see. [10 marks]
4. Plot the distribution of advertised salaries. Discuss how you cleaned the data and discuss the distribution you obtain. [10 marks]
5. Using a box and whisker diagram (https://en.wikipedia.org/wiki/Box_plot), display the average number of ads posted on each day of the week. What pattern do you observe? [10 marks]

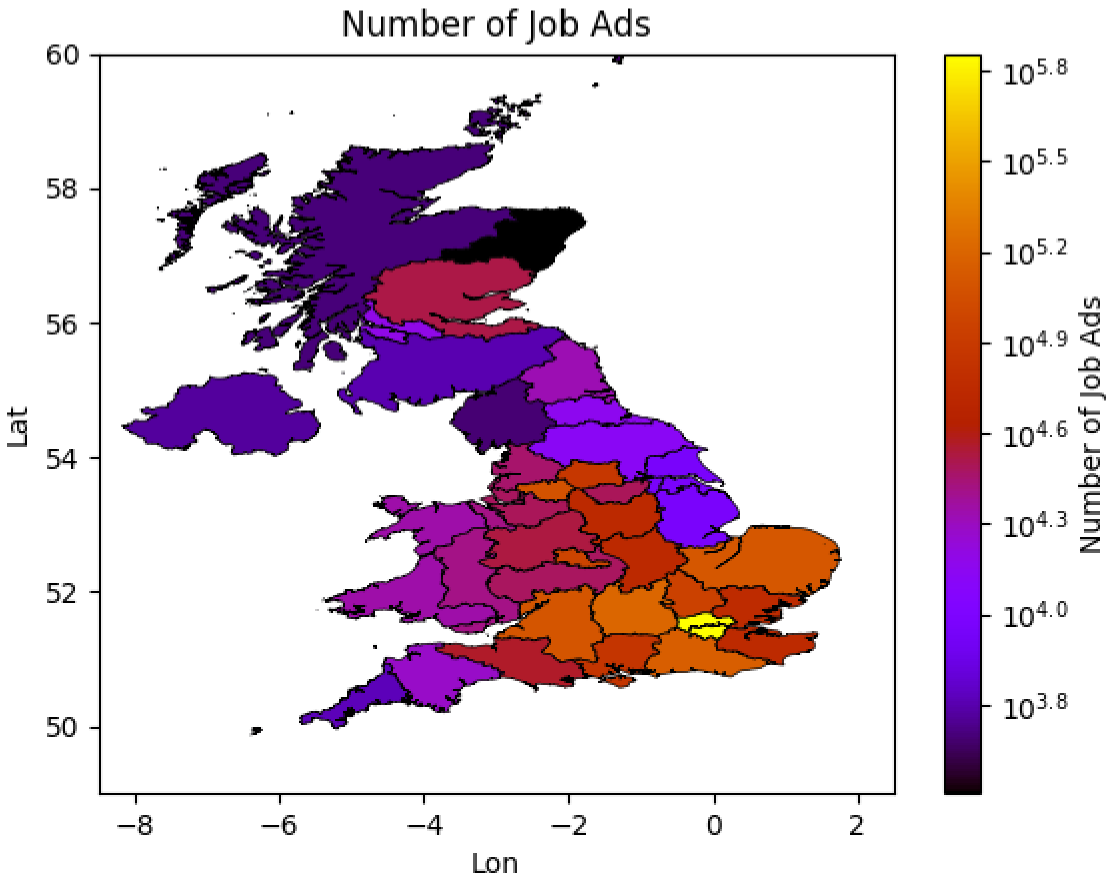
Part 2. Topics (30 marks)

1. Find the top 5 users by total number of job ads posted and discuss your findings. [5 marks]
2. List the 5 most common bigrams found in the job titles. [5 marks]
3. Create a time series of job adverts posted for a particular type of job e.g. nurse, software developer, driving instructor etc. and comment on what you observe. [10 marks]
4. Make a word cloud that summarises the job descriptions for the job type you identified in part 3 and comment on any interesting features. [10 marks]

Part 3. Mapping (20 marks)

1. Draw a map of the UK showing the location of the job ads, similar to the figure below. UK county borders can be obtained from https://data.gov.uk/dataset/11302ddc-65bc-4a8f-96a9-af5c456e442c/counties-

and-unitary-authorities-december-2016-full-clipped-boundaries-in-england-and-wales. The locations are given in the “locationName” field in the JSON. Based on the text you will need to figure out which county to assign each advert NOTE: y**ou do not have to include every ad**, you should be able to approximate the spatial distribution of job ads using a subsample of the data. [15 marks]



2. Explain any patterns you observe. [5 marks]

HINT: geopandas might be a useful library: https://geopandas.org/

Part 4. Reflection (10 marks)

1. Critically reflect on the analysis of job ads to understand changes in the labour market.
    - What are the strengths and weaknesses of the approach?
    - What aspects of the job market can **not** be learned from this data alone?
    - Are there any ethical or legal implications to this kind of analysis?

Write no more than 500 words. [10 marks]

Marking Guide

This is a general marking guide to how your document will be assessed. In general, for each question some of the marks will be awarded for producing the requested plot or data and some will be awarded for the strength of the analysis or the insightfulness of the comment, the exact split will vary with each question.

Criterion	**What is expected for a good mark?**
Writing	Your writing should be clear, well-structured and concise.
Structure of the document	The structure should be clear, easy to navigate and with useful headings.
Presentation	Your document should conform to a clear and consistent visual style, be well-spaced, with appropriate font sizes and consistent and complete labelling and captions.
Code	We are interested in seeing relevant short code snippets that add meaning and context to your document. The code you include should allow the reader to understand the method or process you used, but it is not necessary to include all of it. Be selective. Your code should be well-structured and readable but you will not be required to adhere to specific conventions e.g. PEP
Graphs and maps	Your graphic outputs should be well labelled and captioned, readable, meaningful and relevant.
Analysis	This is the most important area in which you will be assessed. Your analysis of the data should be thorough and we are looking to be impressed by your background research, verification of conclusions and exploration of the available techniques.
Explanation	Your methods and approaches should be clearly described and justified, and your

	comments and conclusions should be robust, valid and verified against additional sources where possible.

Submission

Your work must be submitted by 12pm (noon) on the hand-in date shown at the top of this descriptor. Please allow time for the submission process.

You should submit one PDF document containing all of your answers. Instructions for submission will be made available through the ELE page closer to the submission date.