Senior Data Scientist - Chimnie

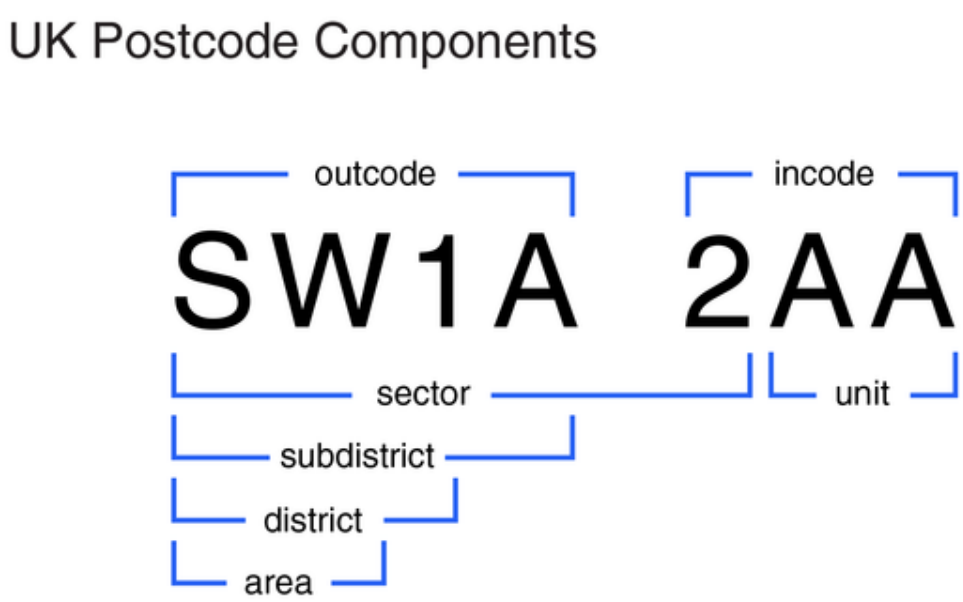
The exercises are aimed at your familiarity with big datasets, your adaptability in working with property insight, and your comfort with applying statistical techniques to manipulate data. You will need to download and work with approx 10GB of data files, the sources for which are provided below.

Please submit your response as a zip file, containing your solutions to the exercises as well as any code used. Use whichever format you think is most suitable, e.g. notebook / markdown.

Data sources

Dataset	Link	Notes
Land Registry	https://www.gov.uk/government/statistical-data-sets/price-paid-data-downloads	Use the complete dataset, not the current month
Postcodes (NSPL)	https://geoportal.statistics.gov.uk/datasets/f7464f3658ba439ba577651b32014cfe/about	The NSPL contains codes for various administrative geographies. The corresponding names can be found under "Names and Codes" in the web portal
British National Grid shapefiles	https://github.com/charlesroper/OSGB_Grids?tab=readme-ov-file	Use 10km resolution

Note that UK Postcodes are comprised by the following components:



Exercises

Using the sources above:

1. Find two land registry records that are likely to be errors
    - In each case, provide a potential explanation
2. Complete the following table by calculating the number of sales and average sale price for all London Boroughs in 2023:

London Borough	Count of sales in 2023	Average sale price in 2023
Barking and Dagenham	?	?
Barnet	?	?
...		

3. Count the number of new build Flats sold in each UK region since the start of 2020
4. Plot the number of sales per week since the start of 2020 as a line chart
    - Discuss the chart, and provide potential explanations for any patterns or anomalies
5. Plot a histogram of sale prices and discuss which distribution best represents the data
    - Feel free to transform the data before plotting, but explain your reasoning if you choose to do so

For the following exercises, use the complete price paid dataset and British National grid tiles at 10km resolution. If you are unfamiliar with geospatial data, try using GeoPandas

6. Using the BNG tiles and an appropriate scale, plot a map showing the number of sales per 10km grid square
7. Plot a map showing the average sale price per 10km grid square
8. Comment on your findings
9. Given everything you have learned from the exercises above, discuss the following model.
    - In your discussion, provide an approximate R^2 value that you would expect from the model
    - How would you improve the model?
    - What range of R^2 would you be happy with?

$$Price \sim \beta * Year + \sum_{i=1}^{n_\alpha} \alpha_i + \sum_{i=1}^{n_\gamma} \gamma_i + \epsilon$$

Where:

- β is the coefficient for the year of sale
- α_i and γ_i are dummy variables corresponding to postcode area and property type respectively

9. Energy Performance Certificates (EPCs) are published for each property transaction recorded by the Land Registry. The data schema for these is available here

I. Which fields from this dataset would be useful in determining sale price?

II. How would you approach the task of joining the EPC database with the Land Registry?
    - N.B. We don't expect you to actually join the databases - but please provide as much detail as possible about how you would approach the task.