**Raspberry Pi Foundation: Deprivation data analysis**

Jonathan Dickins

Approach and tools

* R/RStudio – used for reading in and cleaning data, combining datasets, visualisations, and writing the command line tool.
* Shapefiles are public datasets pertaining to the location and shape of geographic areas. They are used to generate maps and perform other geospatial analyses. I used shapefiles in combination with the postcode dataset to generate maps.
* I used ChatGPT to write a regex to reformat postcodes in an older version of the code – in the end I used a more manual approach to ensure edge cases were being handled correctly.
* See readme.txt for data sources.

**Merging IMD data – merging\_indices.R**

Takes IMD data from across the four nations and creates a single dataset with IMD decile (postcodes.csv)

* Assumptions:
  + Use existing IMD calculations and merge datasets together, rather than create all-new IMD calculations across nations. IMD is calculated differently across the four nations so direct comparisons in IMD score are not valid.
  + Dataset is to be used to identify deprived small areas across all four nations, and not as a tool to directly compare small areas between nations.
* Gathering public data across countries is challenging because of different availability and accessibility of information. My approach was to identify only essential columns across all datasets, align these columns and combine the datasets.
* Essential columns were:
  + Small area codes: A unique identifier to join with postcode data.
  + Small area name: Not useful for joining as these can be inconsistent between sources, but as a visual check for accuracy and to display if needed in the command line tool.
  + IMD Rank: The IMD score rankings for each nation.
  + I wrote a simple function to rename these three columns for consistency.
* I wrote a function to take IMD rank and return a user-specified number of quantiles. I did this to calculate deciles, quintiles, or quartiles flexibly as needed and because different quantiles were not available across all datasets.
* The postcode lookup data supplied in the task briefing did not include Scottish data zone codes. I found a more recent postcode lookup dataset that included them (see readme.txt for source).
* I combined the four consistent datasets and joined to the postcode lookup data on small area code.
* Limitations:
  + There are some specific problems compared IMD data across nations:
    - Different criteria and/or weighting of criteria
    - IMD analyses conducted at different points in time
    - Small area designations differ between nations (data zones in Scotland have smaller population on average than LSOAs in England, for instance).

**Visualising IMD data – mapping.R**

* I chose to visualise the data using shapefile maps to show the distribution of IMD deciles across the UK.
* My approach was to combine the data created in the last step with shapefiles to shade individual areas according to their IMD decile.
* Limitations:
  + It can be difficult to gain a true picture of IMD distribution due to the number of small areas and the resolution of the plots.
  + Given more time, it would be useful to identify whether any characteristics of small areas are associated with higher Foundation engagement, and identify potential clusters of areas to engage with more.

**Postcode to IMD lookup tool – cl\_tool.R**

* My command line tool runs using R. It takes a user-supplied postcode and returns the small area name and latest IMD decile.
* The tool does two things:
  + Reformats user-supplied postcode to align with the ONS Postcode Lookup format, in which postcodes are always seven characters and uppercase.
  + Pulls data from the dataset created in the first task (postcodes.csv) according to the user-supplied postcode.
* Limitations:
  + Relies on the user having both R (and required packages) installed on their machine and a copy of the postcodes.csv dataset available.