CityFibre Data Analysis Challenge

Background

CityFibre (cityfibre.com) are the UK's largest alternative provider of full fibre broadband networks. With a focus on UK cities, they have plans to roll out high speed broadband networks to 8 million premises across the UK. They currently have networks in 7 cities in Scotland (Aberdeen, Dundee, Edinburgh, Glasgow, Inverness, Paisley (Renfrewshire) and Stirling).

The Scottish Government has committed to having superfast broadband access to all by the end of 2021 (https://www.scotlandsuperfast.com/how-can-i-get-it/superfast-access-for-all/). Grants are in place for suppliers to help deliver this mission.

CityFibre would like to combine their own data with open data to help them make decisions about the prioritisation of their broadband rollout programme. They would like to identify the postcodes within their current network regions that are most in need of access to full fibre broadband with a view to providing this at a lower cost. This need may be due to a lack of existing broadband, being located in a more rural location, being an existing area of multiple deprivation with low income and employment, or a combination of all these factors together.

They are particularly keen to combine their own data with public/open data on:

- Superfast broadband schemes
- SIMD data on income, employment and broadband access
- Urban rural classification

They have approached you to help them with the data analysis behind this real business problem.

The Problem

The problem set by CityFibre is:
"Understand which residents potentially need access to broadband at a lower cost in our Scottish cities?"

Translating this into a data question, it is easier to answer the following question:
"Which are the top postcodes within CityFibre's existing network that they should focus on to maximise access to superfast broadband to those most in need? Where are these postcodes located?"

The Datasets

There are 4 datasets available to address this problem:

1. **cityfibre_scotland.csv** - A CityFibre provided dataset containing information about the postcodes and network nodes within their current Scottish networks

2. **scheme_references.csv** – A file containing information about the voucher schemes available for different postcodes, provided by the Scottish Broadband Voucher Scheme (SBVS). More information is available here: https://www.scotlandsuperfast.com/how-can-i-get-it/voucher-scheme/

3. **SG_SIMD_2020** – A folder containing ranks, indices and geometry for the Scottish Index of Multiple Deprivation (SIMD). More information is available here: https://www.gov.scot/collections/scottish-index-of-multiple-deprivation-2020/ This information is provided at datazone level, which is larger than a single postcode and covers the areas that are used for reporting Scottish Neighbourhood Statistics, or the census.

4. **PC_Cut_20_2** – A folder containing the latest information from the Scottish Postcode Directory. This contains postcode level information and geometry for every postcode in Scotland. More information is available here: https://www.nrscotland.gov.uk/statistics-and-data/geography/our-products/scottish-postcode-directory/2020-2

Approach

This is a broad question and the analysis can be approached in a number of different ways, leading to different output sets of priority postcodes. However, prior to any analysis the input datasets will all need to be prepared, through thorough cleaning, validation and manipulation.

After the analysis step the priority postcodes will need to be extracted so that they can be shared with CityFibre and their location or boundaries visualised on a mapping tool.

Step 1: Data access and understanding

Before commencing any analysis, it is important to understand the data that is being used.

- **Access**: download local copies of the input datasets
- **Documentation**: download any documentation and data dictionaries from supporting websites that can provide background into the datasets and the variables
- **Read in data**: read in all datasets
- **Tidy**: tidy up any variable names to make the datasets easy to work with
- **Visual inspection**: review each input dataset to see the type of data it contains, identify any potential issues and get a feel for the type of data it contains

- **Size, shape and format**: record the size and shape and data type of each input dataset and variable
- **Missing and unique values**: for each variable identify the level of missing data and the number of unique values
- **Relationships**: for related columns identify the level of consistency between them

Step 2: Data cleaning and manipulation

Undertake cleaning on each input dataset separately.
The supporting datasets are prepared first so that they are ready to be merged with the CityFibre dataset to create the final analysis dataset.

1. **Prepare the postcode dataset**

- **Subset**: subset the small and large user postcode files to a smaller set of useful variables
- **Combine**: combine the small and large postcode files together, creating an additional variable to identify the postcode type
- **Duplicates**: identify and fix duplicate entries. Retain only live postcodes

2. **Prepare the SIMD dataset**

- **Subset**: retain only the variables of interest. Referring to the supporting documents and technical notes will help to identify interesting variables that will be useful in the analysis
- **Reformat**: convert numeric variables from string format, removing any unnecessary characters e.g. “%”
- **Rename variable**: Rename variables to have a more descriptive name

3. **Prepare the CityFibre dataset**

- **Duplicates**: identify and remove exact duplicates
- **Duplicate postcodes**: review and fix remaining duplicates. Glasgow postcodes should begin “G” and Renfrewshire postcodes should begin “PA”.
- **Postcode format**: Ensure consistency of postcode formats to enable joining between datasets
- **Postcode validity**: Use the postcode dataset to ensure all remaining postcodes are live and valid
- **Scheme references**: merge in the scheme reference table
- **Postcode information**: merge in additional postcode information including the location, urban rural classification and datazone
- **SIMD information**: merge in a subset of SIMD information
- **Review**: ensure all postcodes are unique and no duplicates remain

- **Datazones**: write out an additional dataset with the datazones that are relevant to the CityFibre postcodes

Whilst undertaking the dataset cleaning and manipulation, ensure that at each stage the datasets are named clearly and not overwritten. This will allow errors to be spotted and comparisons to be made between datasets at different stages of preparation.

Step 3: Data analysis

Revisit the problem statement prior to commencing the analysis. The focus should be on identifying the postcodes with the biggest "need" for broadband access.

Need can be identified in different ways:

- Areas with low SIMD ranks, or low income, or low employment
- Areas with current low rates of access to superfast broadband
- Areas with voucher schemes in place, identified by the government as in need broadband access
- Those in rural locations for which it has historically been difficult to supply broadband to.

An approach for undertaking the analysis could include:

1. **Identify correlations**: calculate the correlation coefficients and plot the relationship between the different numeric measurements of need to understand if and how they are related.
   For categorical variables plot the distribution and relationship to the rate of broadband access.
   Identify the independent variables to take forward into the scoring.

2. **Create a prioritisation score**: Create a score that can be used to prioritise postcodes based on the independent need estimation variables.

3. **Determine a score cut-off**: Plot a score distribution and determine a suitable score cut-off that will identify the top 1000-2000 postcodes.

4. **Output prioritised postcodes dataset**: write out a dataset of the priority postcodes. For mapping, write out a csv containing as a minimum the postcode and its latitude and longitude.

Ensure any graphs created have titles and axes labels to allow the reader to understand what they are showing. The choice of graph type should take into consideration the type of data being plotted.

Step 4: Output mapping

Mapping can be done in either of two ways:

1. **Google Mymaps**

The simplest way is to use google's mymaps and upload a file containing location information to it. This will place a pin at the location of each priority postcode.

Go to https://www.google.com/mymaps

- Select "Create a new map"
- Give the map a suitable title
- Import the exported priority postcode .csv file
- Select latitude and longitude as the position fields
- Select postcode as the marker title
- Any additional data will also be imported and will be visible on the marker.

2. **Creating an interactive HTML map**

Interactive HTML maps can also be created using the open-source Leaflet library (https://leafletjs.com/). This uses OpenStreetMap for the underlying maps.

- The geometry of the postcode and datazone boundaries are available in the provided .shp files. These geometries have been provided in the form of Eastings and Northings and require converting to values of longitude and latitude.
- Create a popup label for each postcode to display any information of interest.
- Display the geometry information of the priority postcodes in a leaflet map using suitable colours and transparency.
- Write the map out as an HTML file for sharing.

Step 5: Interpretation and communication

The final step is review the outputs and prepare a presentation for the CityFibre team to explain any insights captured from the analysis and make your recommendations about how to address their problem statement.

This presentation should focus on the problem statement, rather than the analysis journey you undertook to arrive at your recommendations.

Supporting Information

Since this is a real-world problem the input dataset has a number of data quality issues. Addressing these data quality issues and then merging all the prepared input datasets together will take up the majority of the time.

The other challenge is the different levels of data being used. Identifying what each row in the dataset represents. For the initial CityFibre dataset it is postcode + node. For the SIMD data this is datazone, which is a group of postcodes. For the postcode dataset this is postcode.

It is recommended to undertake the analysis at postcode level, ensuring only relevant information is merged in at any stage.

Possible Extensions

Priority nodes: Once the priority postcodes have been identified, map these back to CityFibre's network nodes and recommend the top 10 nodes for them to focus on.

Alternative scoring: experiment with different scoring approaches, identifying how the priority postcodes will differ depending on which needs is being optimised.