The Deluge Team Clyde

1. Data Pre-Processing

By merging the data with attribute joins, it was possible to group it up into two separate groups (See Appendix 1). Each group had at least one member with coordinate data: the two postcode datasets and the stations dataset, meaning both groups could be joined together spatially by finding the nearest station of each type (rain, river and tide) to each postcode coordinate. This allowed us to create two complete datasets (one labelled and one unlabelled) on which to test the feature importance and train our model on, with a record for each of the location points. Several other pre-processing steps were taken on each dataset, these are detailed in the diagram in Appendix 2.

2. Modelling Approaches

Model 1 – Median property price prediction

Default model: Random Forest, **Options include**: K-Neighbours, combined 'classification-regression' K-Neighbours

Cross-validation was used for selecting and tuning the median property price regression model, with feature permutations. The final default model is a Random Forest regression model, chosen based on RMSE performance on both our held-out test set and the evaluation set.

Both normal and improved K-Neighbours regression models are also included as options for the user.

The improved K-Neighbours model utilises a combined 'classification-regression' regression, in which the data is first classified as being of either 'low' or 'high' median price, and a separate regressor tuned to predict the price for each class. This approach showed much improved performance for predicting price for 'low' median price areas, and so should be selected by the user based on the data appropriately.

Model 2 – Flood risk prediction

Default model: K-Neighbours, **Options include**: Gradient Boost

We again used cross-validation for selecting and tuning the flood risk classification model, with feature permutations. The final default model is a K-Neighbours classification model, chosen based on performance on the evaluation set.

Interestingly, a Gradient Boost regression model was actually the best performing model on our held-out test set, using an augmented feature set that included the mean, range, and extreme rainfall levels for each location. Unintuitively however, this model performed worse on the evaluation set (as well as a K-Neighbours model also using the augmented feature set).

To generate the 'augmented' dataset, an additional K-Neighbours classification model is used to assign each location to a nearest 'measuring station', and its corresponding measured rainfall levels.

Both these 'augmented' models are still included as options for the user.

3. Data visualisation demonstration

Rainfall, river, tidal, and risk data can be visualized both scatter plots and heatmaps using Folium maps. These visualizations can be accessed through a webpage link and are shown in Appendix 3

4. User Interface

Appropriate scripts, detailed in the package repository READ.me, enable the user to run both the risk prediction model and visualisation tools from the command line.

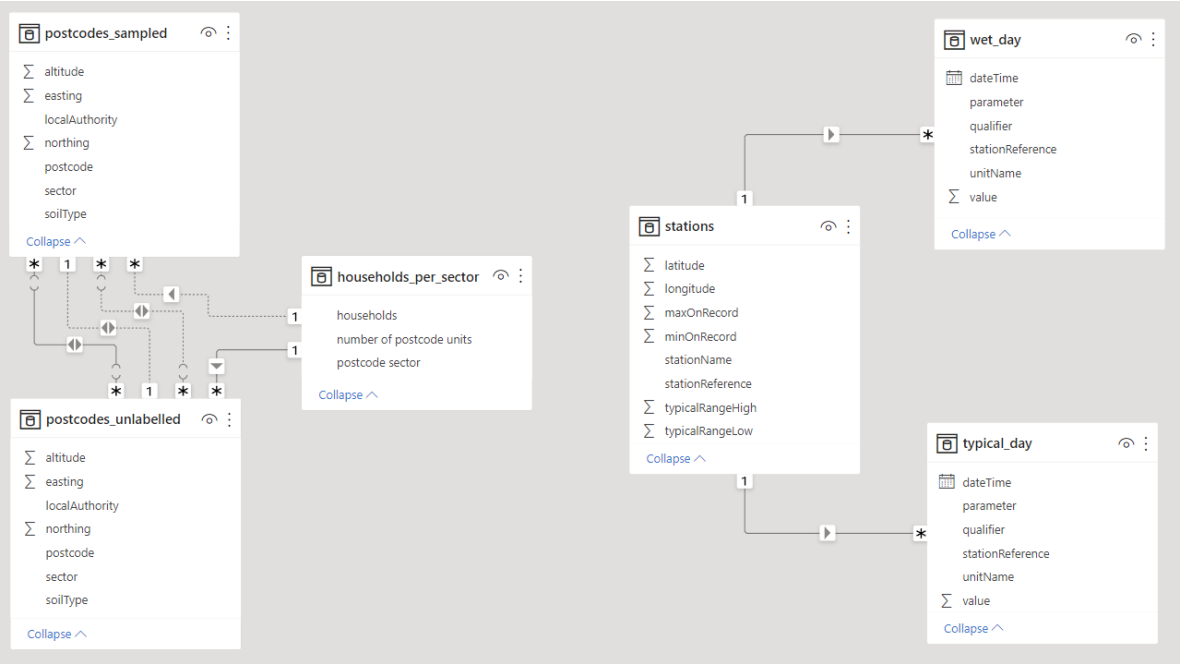
Instructions on how to do so can also be found in the READ.me.

5. Testing

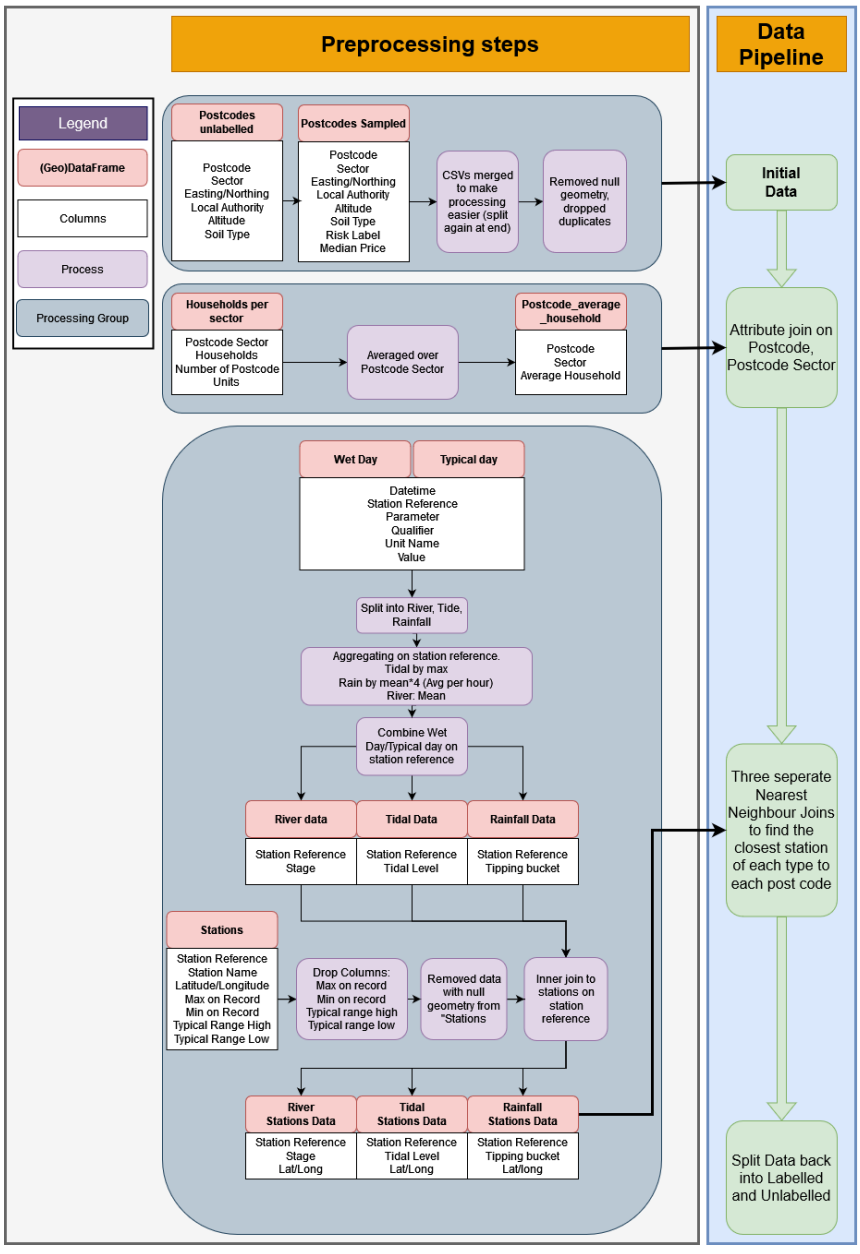
Appropriate unit tests are included for core functionality, using the pytest framework.

Instructions on running these tests can also be found in the package repository READ.me.

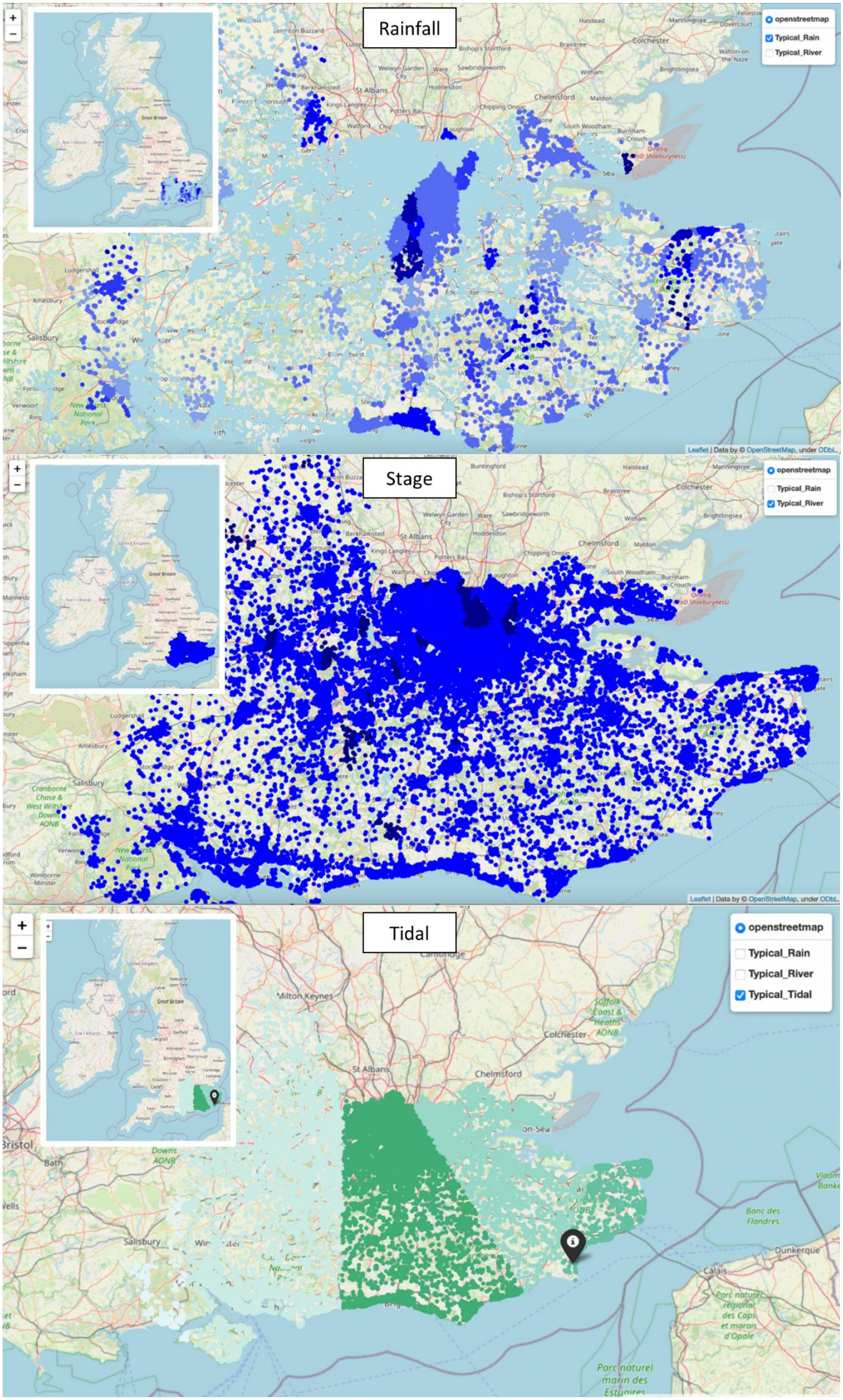
Appendix 1: Data grouped by shared attributes



Appendix 2: Data pre-processing diagram



Appendix 3: Data Visualisation outputs



Appendix 4: Overall Process Diagram