## EPA1351 – Assignment 1: Data cleaning – Group 8

Erin Bartholomew

Juan David Godoy Landinez

Ellen Mok

Cyntha Nijmeijer – 4157974

Bastiaan Reijm

#### Step 1: HTML to CSV

The first step to take is converting the original html data files of the roads and the bridges to two csv files; one for the roads and one for the bridges. The reason the original data is used instead of the ‘overview’ excel files, is that potential calculation errors in the excel files are avoided and the semantic accuracy is increased. Semantic accuracy indicates the conformity of a data value to its real-world counterpart, calculation errors result in a value that does not correspond to the real-life value any more.

To convert the html files to csv files BeautifulSoup is used to parse the original file and classes for roads and bridges are defined to include all attributes of each road and bridge. For each individual road and bridge a separate csv file is created as a check and two complete files with all roads and bridges respectively are created.

#### Step 2: Match Road & bridge coordinates

#### Step 3: