The data for this project was taken from

<https://data.gov.uk/dataset/d0be1ed2-9907-4ec4-b552-c048f6aec16a/gb-driving-licence-data>

The source file (Driving-Licence-data-March-2020.xlsx) is made up of multiple sheets, each containing slightly different sets of data. In this project I will use the data from sheets DRL0102 (Licence types by postcode district) and DRL0132 (Licences with penalty points by postcode district). To make this data suitable for use in a database, some preliminary cleaning is required.

First, I removed unnecessary data (summary rows, descriptions) and formatting from DRL0102 and DRL0132 to make it csv compatible, producing ‘Licence\_types\_by\_district.csv’ and ‘Licences\_with\_points\_by\_district.csv’.

Here I found a mismatch between the two tables. DRL0132 shows data for 2808 postcode districts, but DRL0102 shows 3194. Before I proceed, I want to remove the districts from DRL0102 that do not appear in DRL0132.

After removing districts that did not have a corresponding record, I created ‘Cleaned\_licence\_types\_by\_district.csv’. Both datasets now contain 2808 records.

Here I found that there the number of rows in the table was fewer than expected. For any amount of penalty points with zero corresponding licences, that entire column was omitted from the source data. To correct this, I added the missing columns and then also added extra columns to allow for a maximum of 100 points.

With the two cleaned tables ready, I had to consider how they would be arranged once in the database. I could store them as separate tables, linked by postcode district as a foreign key, or I could join them together into a single table. As this project is limited in scope, I decided to join them together for a simpler database structure. If I planned to make this project more complex in future then I might need to reconsider whether separate tables might be better.

My finalised dataset is ‘Joined\_licences\_by\_district.csv’.