**EXTRACT**

Data for different wine ratings was found on Kaggle and Dataworld websites. The orginal files came in various formats including CSV, JSON and XLXS.

The Kaggle data was from winemag.com (Wine Magazine’s web site). The CSV file was too large to load into GitHub, so we converted it to an XLXS file and it was just small enough to avoid needing to parse it into separate files.

The Dataworld data was originally in both CSV and XLXS format, but some of the symbols were not UTF-8 characters. This was addressed by reading the XLXS file version.

A column was added to both extracted pandas indicating the source of the data WM or DW. This was to allow duplicates to be removed and to average the rating and price of any wine in both WM and DW and change the source to indicate it was both.

The Winemag file did not have vintage as a column. The title of a wine is the winery, the vintage and the designation. The vintage was extracted from the column of string variables under title. The results were stored as integers in a vintage column.

The Dataworld file had the Vintage column as a mixture of datetime variables and strings. A subroutine converted them into integers for the vintage year and replaced the Vintage column.