**Steps in cleaning Capstone data**

I saved the full list of UCAS data files into a ‘Data’ folder in my Capstone Project folder, and also saved a reference Excel document that showed the makeup of each file. All filenames were in a similar format, and for each file the first 5 lines were descriptive, with the data content starting on line 6

I first created lists of all the files, and then all the filenames. I shortened the list to only include csv file names.

I experimented with using for loops and lapply to import and name all files using their content and UCAS-given filenames, and to list out the content of files so that I could choose the most useful ones in R, but could not make this work.

Instead, I reviewed the Excel reference document and chose and imported the three files that combined in any way the crucial variables that I needed to answer my capstone project question (detailed subject, student UK domicile region, student age, higher education institution). I then renamed the files to show their content.

I checked the data frames for NA values, and found they had additional empty columns on the end of them. I then removed blank columns from files and renamed the files, bringing the description to the front.

I checked the data frames for NA values again and found there were now none.

I viewed a summary of each of the three files to see which columns were included, and then checked the summary for each column to see whether there were any values that didn't make sense or were outliers. I found some columns had ‘Unknown’ as one of their values, specifically for student domicile region.

Where ‘Unknown’ was a value in each file, I checked how many values this applied to per file. I considered removing the ‘Unknown’ domicile examples, but was wary of doing so as the data isn’t at individual student level.

I had had trouble searching for ‘Unknown’ values (because of double and single quote marks), so viewed the data frame again to try and understand why. I noticed that character values in the data frames (including those for domicile region such as ‘Unknown’) were surrounded by single quotes, so used gsub to remove those.

I then considered the three data frames I had available in light of the capstone project question, and realised the crucial starting point was understanding which subjects students were applying for, from which domicile regions, so focused on this data frame.

I realised I needed first to understand the picture in 2016 (could then look at trends later if relevant), so created a 2016-only version of the subject by domicile data frame.

I decided against reshaping the dataframes, as the data is not available down to individual observation level, and was already quite ‘long’ in shape.