# Exercise Data scientist - Data wrangling (3 days to complete)

Create an R script to perform the following task, do as much as you can. Think about documenting your code so other team members can follow your logic, think about your choice of packages, as well as the structure of the code so it is easily maintainable. If you have any questions email [louise.donnison@ed.ac.uk.](mailto:louise.donnison@ed.ac.uk)

**Materials**

The file to process with R code:

*Data\_Scientist\_Exercise\_File.xlsx* file containing the sheet to read: EXERCISE\_INPUT\_FILE

We have also provided an example of what the R code should output:

Data\_Scientist\_Exercise\_Output\_File*.xlsx* containing the sheet to output: EXERCISE\_OUTPUT\_FILE

**Exercise**

Write an R script to perform the following data wrangling actions:

1. Write code to read in the Excel file *Data\_Scientist\_Exercise\_File.xlsx* sheet *EXERCISE\_INPUT\_FILE* into a dataframe
2. Clean the text of the entries in the *DISEASE* column

* TRYPs should be Trypanosomosis
* PPR should be “Peste des petits ruminants”

1. Filter the dataframe, so that only rows that have values in the *YEAR\_PUBLICATION* column later than 2010 are kept - hint there is only one article like that.
2. Remove the columns *START\_DATE\_DATA* and *END\_DATE\_DATA* from the dataframe they are not needed.
3. Reorder the dataframe so that columns STATE and DISEASE are the last two columns of the dataframe.
4. Split the IDENTIFIER column and create a new column called AUTHOR which does not contain the date, but just the text. Still keep the IDENTIFIER column.
5. Write code to output the dataframe to an Excel File called Data\_Scientist\_Exercise\_Output\_File*.xlsx* with the sheet name EXERCISE\_OUTPUT\_FILE

Print out to the console the following summary numbers based on the cleaned dataframe:

1. The IDENTIFIER column value with the highest value in the PERCENTAGE column
2. The sum of the values in the column NUMBER\_TESTED

Consider introducing a folder structure to organise the files, code, and a readme.

**Deliverable**

Your R script that you have written to perform the exercise emailed within the timeframe given.

# Presentation Data scientist - Data pipeline (In interview)

Prepare and present a 5 minute presentation on the topic: Using examples, demonstrate your knowledge and experience of working with data pipelines, with a particular focus on ETL (extract, transform and load). Highlight any successes and challenges that you have encountered.