1. Data given in .xlsx format.
2. Save the two sheets of the excel spreadsheet as 2 separate csv files
3. Set the working directory to that which the data is currently in.
4. Ask R to read sheet 1 and change it to a csv
5. Do the same with sheet 2
6. With the raw data, I’ve made the top row the only header row
7. As a separate df, I’ve made the sample names the row names and deleted all columns before the lipid data
8. I’ve asked the script to tell me if there are any duplicate columns in binary format (i.e. TRUE)
9. I’ve read in the lipids tested in a wide format
10. I’ve made the format long, so every lipid is assigned to a family i.e. the previous column name
11. I’ve reordered all of the family names alphabetically in order to make sure every family is grouped. Because of the previous wide format, this leaves a lot of family names with empty cells for lipid names as they were lower down.
12. I’ve deleted all rows with lipid family names that don’t have a corresponding lipid in the adjoining cell.
13. I changed the syntax of the (): symbols to . symbols in order to match the column names of the raw data
14. I made new objects of the column names for the raw data and the lipids in the families object. These were then compared against each other to see which ones didn’t match. (this solved the number issues I had). I got it to tell me how many were mismatched and then list the names of the columns in the raw data that wouldn’t match to the lipids tested list.
15. I got the programme to split up all of the lipid families into separate data frames and export them all as new CSV files

First aim:

* Get the programme to organise the data in a logical way
* Get the programme to split the lipid families into different data frames in order to be able to compare the results between and within groups