# Guidelines to use Dataset.R for data pre-processing

A document contains a brief guidance on how to pre-process raw data before uploading to the **PredPLSR App.** This document is prepared assuming the user record spectrum data in the following format. The first sheet in the Excel Workbook contains a Legend indicating **Melamine Concentrations** and corresponding **Sample IDs.** It is **mandatory** to store data in the exact format in an Excel Workbook (First sheet – Legend, remaining sheets – spectrum data for each concentration) as shown in Figure 01.

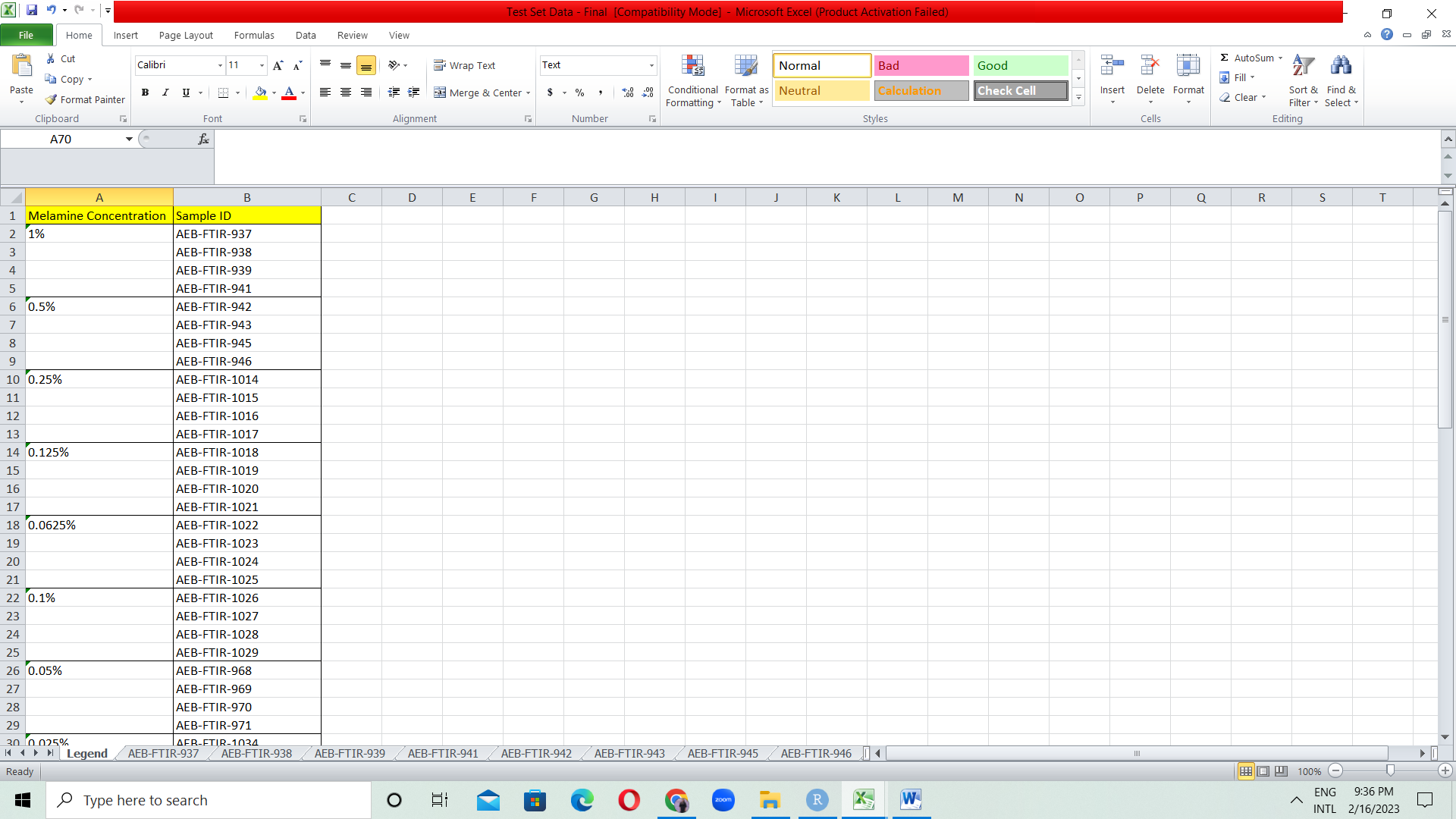


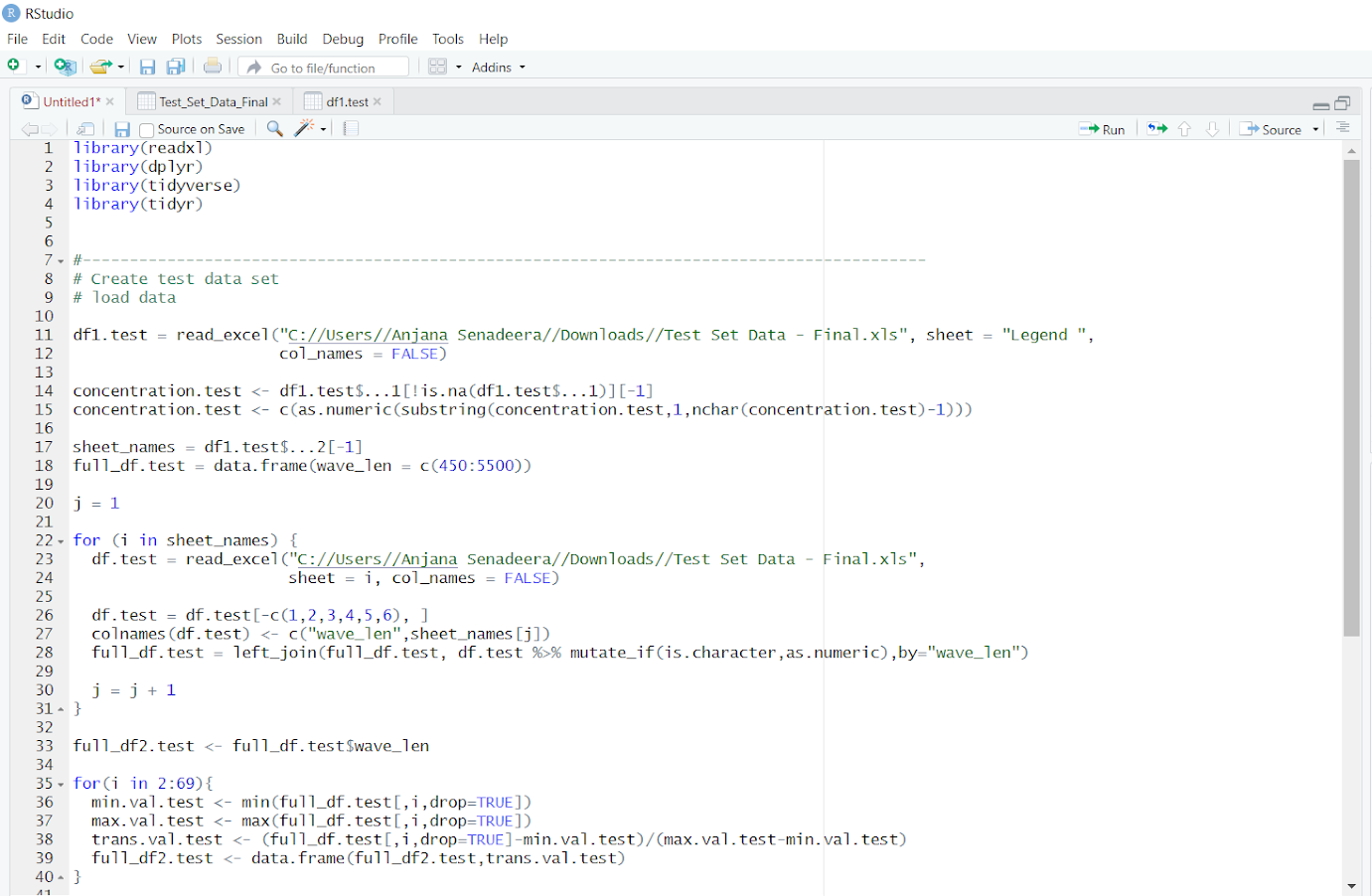
Figure 01: Data format provided by the client

This dataset will be used as the main input for **Dataset.R** which processes the dataset into a tidy format compatible with PredPLSR App. You can upload any processed spectrum dataset into this app and perform PLS regression.

Following is a demonstration on how to customize values inside Dataset.R to gain the final tidy data set. (This is performed to get the final testing set).

**Testing set**

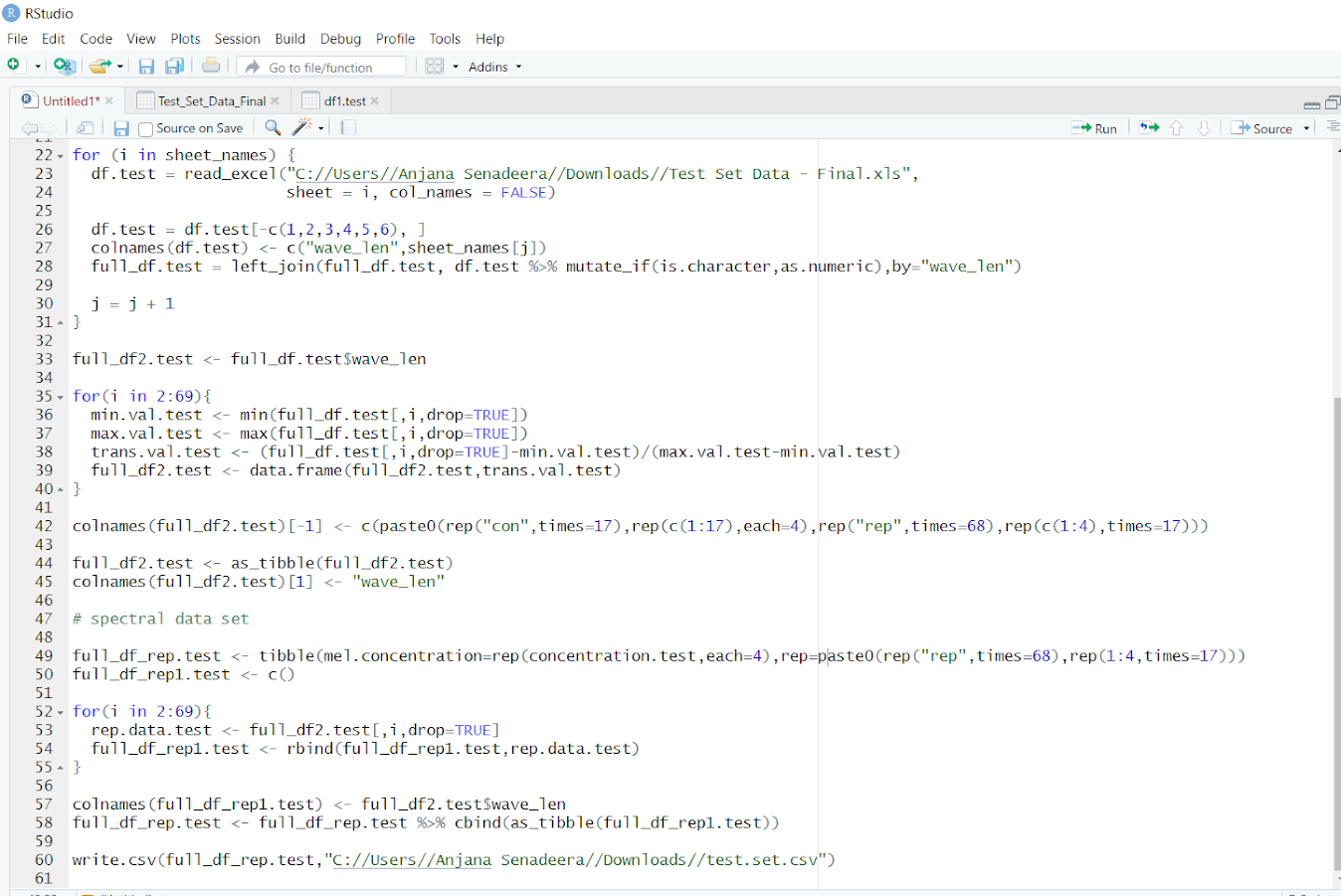
The Testing Set contained spectrums of 18 melamine concentrations with each having 4 replicates. However one concentration had only 3 replicates therefore we had to remove that concentration. Note that the program may indicate an error if at least one replicate is missing for a particular concentration. You can either add the missing observation or you have to remove the entire concentration from the Legend in the Excel workbook. We proceeded data pre-processing stage for 17 concentrations each with 4 replicates, therefore with 68 observations in total.



**Sheet = “Legend ” : keep a single space before the last quotation mark if you get an error message as ‘sheet’ = “Legend” not found**

**Insert the File Path where the Excel Workbook is stored**

Figure 02: Customizing Dataset.R – Screenshot 01



**2: 69 : 2 is a fixed value. 69 = total number of observations + 1**

**(in this demonstration we considered 68 observations:**

**17 concentrations \* 4 replicates)**

**Insert the File Path where you wish to store the final output (.csv file)**

**values in this line should also be changed accordingly**

**c(1: number of concentrations)**

**times = number of concentrations**

**times = number of obs.**

**times = number of concentrations**

Figure 02: Customizing Dataset.R – Screenshot 02

Customizing Dataset.R to obtain a training set can also be done in a similar manner. But note that you have to pre-decide which observations to be included in the training set and prepare a separate Excel Workbook accordingly.