Data Mining - Spring 2024
Exercise 1

Tidy data

1. In general, I recommend to work through the chapters on "Tidy Data" in Hadley Wickham's book "R for Data Science" (2nd ed.), `https://r4ds.hadley.nz/data-tidy` to familiarize yourself with the concept of tidy data and the corresponding R functions.

2. Download the excel file "UntidyData_Badly-Structured-Sales-Dat.csv" from Teams and rearrange this data into the correct four columns. There has been a mix of rows and columns everywhere. Also, watch out for Grand Totals and Sub Totals, you do not need those in tidy data.

3. Download the excel file "UntidyInvoices-with-Merged-Categories-and-Merged-Amounts.csv" from Teams and tidy it. Because a single transaction (identified with an order id) has multiple items purchased, who ever captured this data decided to create a single row for each order, thereby lumping the different items purchased and the amounts together into 2 fields respectively. The better thing to do is to let each item purchased be on a single row with the amount. It is better to repeat the Order IDs on different rows than lumping up amounts in a single cell. As we will be analyzing items bought and amounts a lot, we need them separated into rows.