

Getting and Cleaning Data

Coursera Data Specialization Track

4 May 2015 – 8 June 2015

Week 1 Lecture Notes

0.1 Obtaining Data

0.2 Raw and Processed Data

0.3

0.4 Reading Excel Files

Excel files are still the most widely used format for sharing data. But we need to be able to extract this data.

We load the `xlsx` package. (Not the only library for this functionality.)

The `write.xlsx` function will write to an Excel file. `read.xlsx2` is much faster than `read.xlsx` but may be unstable when reading subsets of rows. The `XLConnect` package has more options for writing and manipulating Excel files. In general though, it's better to store data in a database or in a comma separated (*.csv) or tab separated (*.tab / *.txt) as it's just easier to distribute these formats.

Week 2 Lecture Notes

Week 3 Lecture Notes

Week 4 Lecture Notes