Lesson Outline

> Data wrangling process:
>
> - Gather
> - Assess
> - Clean (this lesson)

Cleaning your data is the third step in data wrangling. It is where you fix the quality and tidiness issues that you identified in the assess step. In this lesson, you'll clean all of the issues you identified in Lesson 3 using Python and pandas.

This lesson will be structured as follows:

- You'll get **re**motivated *(if you aren't already)* to clean the dataset for lessons 3 and 4: Phase II clinical trial data that compares the efficacy and safety of a new oral insulin to treat diabetes to injectable insulin
- You'll learn about the data cleaning process: defining, coding, and testing
- You'll address the missing data first (and learn why it is usually important to address these completeness issues first)
- You'll tackle the tidiness issues next (and learn why this is usually the next logical step)
- And finally, you'll clean up the quality issues

This lesson will consist primarily of Jupyter Notebooks, of which there will be two types: one quiz notebook that you'll work with throughout the whole lesson (i.e. your work will carry over from page to page) and three solution notebooks. I'll pop in and out to introduce the larger conceptual bits.

You will leverage the most common cleaning functions and methods in the pandas library to clean the nineteen quality issues and four tidiness issues identified in Lesson 3. Given your pandas experience and that this isn't a course on pandas, these functions and methods won't be covered in detail. Regardless, with this experience and your research

Lesson Outline

NEXT