R Workshop Day 4

Importing messy data

- This analysis uses data available only to Hanover College employees, apologies to anyone else looking at this.
- Create a new project and RMarkdown document.
- First, download the xlsx file we shared and upload it into your new project.
- Add chunks that load hanoverbase and readxl.
- In a chunk we would execute:

```
path <- "course_based_instruction_stats_R.xlsx"
sheets <- excel_sheets(path)
```

  Notice that we want to ignore the "Trend Report" sheet for now. Use str_subset[1] for this.
  At the end of this, we have a variable called sheetNames_to_process.
- We first practice loading one sheet:

```
sheetName <- "18-19"
sheet <- readExcel(path, sheetName)
```

- Looking at the sheet, we see that it has loaded a lot more rows than we need. We need to find an unambiguous way to specify which rows we want. One way is to ask for 2 things:
  - The "level" should exist (not NA)
  - The "assigned librarian" should be one of .....

  Here is code that does that:

```
librarians <- c("Kelly", "Reiley", "Heather", "Jen")
sheet %>% drop_na(Level) %>%
    filter(`Assigned Librarian` %in% librarians)
```

- We load all sheets at once by use of the map command[2], and turning the above step into a function.
  - When we try this, it fails because one of the sheets has the columns named "Course Level" rather than "Level". We then add a step to fix that, using dplyr's rename command:

```
sheet %>% rename(Level="Course Level") %>% ...
```

[1] ../morsels/stringrStrSubset.html
[2] ../morsels/purrMap.html