Reading Excel Data

Another really common type of data is Excel data. This is data that has a file extension of `.xls` or `.xlsx`. We often want to pull data from different *sheets* within these files.

`readxl` Package

The `readxl` package is part of the tidyverse (not loaded by default) that has functionality for reading in this type of data!

However, these types of files cannot be pulled from a URL. Instead, we'll need to download the files and provide a path to them.

- Download the dry beans data set available at: https://www4.stat.ncsu.edu/~online/datasets/Dry_Bean_Dataset.xlsx
- Store it in your R project folder, a datasets folder within there, or the folder with your `.qmd` file in it.
- Let's read it into R!

If the file exists in your `.qmd` file's directory, we can read it in via:

```
        library(readxl)
```

```
Warning: package 'readxl' was built under R version 4.1.3
```

```
        dry_bean_data <- read_excel("Dry_Bean_Dataset.xlsx")
        dry_bean_data
```

```
# A tibble: 13,611 x 17
    Area Perimeter MajorAxisLength MinorAxisLength AspectRatio Eccentricity
   <dbl>     <dbl>           <dbl>           <dbl>       <dbl>        <dbl>
 1 28395      610.            208.            174.        1.20        0.550
 2 28734      638.            201.            183.        1.10        0.412
 3 29380      624.            213.            176.        1.21        0.563
 4 30008      646.            211.            183.        1.15        0.499
 5 30140      620.            202.            190.        1.06        0.334
 6 30279      635.            213.            182.        1.17        0.520
 7 30477      670.            211.            184.        1.15        0.489
 8 30519      630.            213.            183.        1.17        0.514
 9 30685      636.            214.            183.        1.17        0.514
10 30834      632.            217.            181.        1.20        0.554
# i 13,601 more rows
# i 11 more variables: ConvexArea <dbl>, EquivDiameter <dbl>, Extent <dbl>,
#   Solidity <dbl>, Roundness <dbl>, Compactness <dbl>, ShapeFactor1 <dbl>,
#   ShapeFactor2 <dbl>, ShapeFactor3 <dbl>, ShapeFactor4 <dbl>, Class <chr>
```

Great! Easy enough. If the file was in one folder up from your `.qmd` file, you could read it in via

```
dry_bean_data <- read_excel("../Dry_Bean_Dataset.xlsx")
```

If the file had been in a folder called **datasets** located one folder up from your `.qmd` file, you could read it in via

```
dry_bean_data <- read_excel("../datasets/Dry_Bean_Dataset.xlsx")
```

Note: If you switch to have your chunk output in your console, the working directory used during the interactive modifying and submitting of code from your `.qmd` file will use your usual working directory for your R session. This can be annoying! When you render it will use the `.qmd` file's location as the working directory.

Reading From a Particular Sheet

We might want to programmatically look at the sheets available in the excel document. This can be done with the `excel_sheets()` function.

```
excel_sheets("Dry_Bean_Dataset.xlsx")
```

```
[1] "Dry_Beans_Dataset" "Citation_Request"
```

We can pull in data from a specific sheet with the name or via integers (or `NULL` for 1st)

```
citation_dry_bean_data <- read_excel("Dry_Bean_Dataset.xlsx",
                                     sheet = excel_sheets("Dry_Bean_Dataset
citation_dry_bean_data
```

```
# A tibble: 0 x 1
# i 1 variable:
#   Citation Request :
KOKLU, M. and OZKAN, I.A., (2020), "Multiclass Classification of Dry Beans
Using Computer Vision and Machine Learning Techniques." Computers and
Electronics in Agriculture, 174, 105507. DOI: https://doi.org/10.1016/
j.compag.2020.105507 <lgl>
```

Notice that didn't read in correctly! There is only one entry there (the 1st cell, 1st column) and it is currently being treated as the column name. Similar to the `read_csv()` function we can use `col_names = FALSE` here (thanks coherent ecosystem!!).

```
citation_dry_bean_data <- read_excel("Dry_Bean_Dataset.xlsx",
                                     sheet = excel_sheets("Dry_Bean_Dataset
                                     col_names = FALSE)
```

```
New names:
* `` -> `...1`
```

```
citation_dry_bean_data
```

```
# A tibble: 1 x 1
  ...1
  <chr>
1 "Citation Request :\r\nKOKLU, M. and OZKAN, I.A., (2020), “Multiclass
Classif~
```

We can see there are some special characters in there (like line break). If we use `cat()` it will print that out nicely.

```
        cat(dplyr::pull(citation_dry_bean_data, 1))
```

```
Citation Request :
KOKLU, M. and OZKAN, I.A., (2020), “Multiclass Classification of Dry Beans
Using Computer Vision and Machine Learning Techniques.” Computers and
Electronics in Agriculture, 174, 105507. DOI: https://doi.org/10.1016/
j.compag.2020.105507
```

Reading Only Specific Cells

Occasionally, we might want to read only some cells on a particular sheet. This can be done by specifying the `range` argument!

- Cells must be in a contiguous range

```
        dry_bean_range <- read_excel("Dry_Bean_Dataset.xlsx",
                          range = cell_cols("A:B")
                          )
        dry_bean_range
```

```
# A tibble: 13,611 x 2
    Area Perimeter
   <dbl>     <dbl>
 1 28395      610.
 2 28734      638.
 3 29380      624.
 4 30008      646.
 5 30140      620.
 6 30279      635.
 7 30477      670.
 8 30519      630.
 9 30685      636.
10 30834      632.
# i 13,601 more rows
```

Recap!

The `read_xl` package provides nice functionality for reading in excel type data.

- As it is part of the `tidyverse` it reads the data into a `tibble`
- Functionality to read in from different sheets or to read in particular ranges of data