# Reading in Data with R

Sean Conway

8/1/2021

# Introduction

- Today, you will learn how to read in datasets with R. By read in, we mean take a data file that exists outside of R and import it such that we can analyse with R code.
- When working with data, it often comes in different file formats. While there are more file formats than we can cover in a single slide show, today we will focus on three especially common file types: *.csv*, *.xls/.xlsx*, and *.RData*.

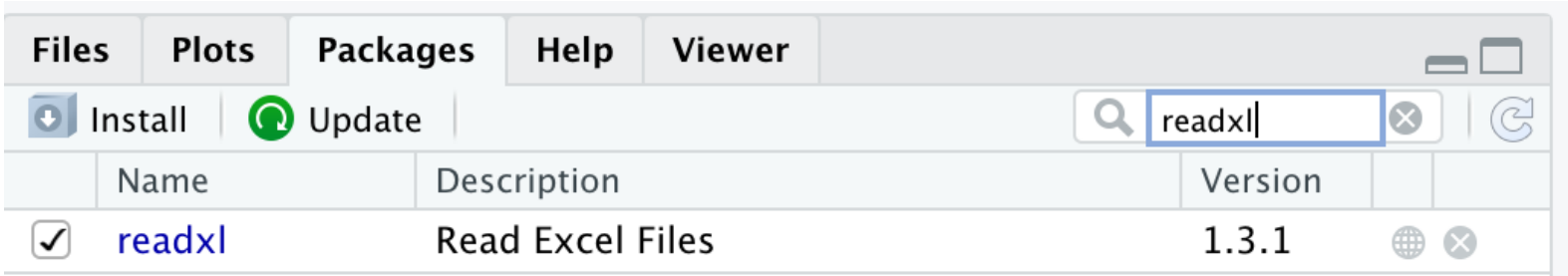
# Getting Started

- Before we get started, you will need to make sure you have two packages installed on your computer: `readr` and `readxl`.
- To install these packages, simply run the following code in your RStudio console:

```
install.packages(list("readr","readxl"))
```

# Getting Started

- If you're unsure whether or not you have these packages, you can navigate over to your packages tab (bottom right panel in RStudio) and search for them.



# Tabular Data

- Now that we're ready to go, let's talk a little about tabular data.
- Tabular data is simply data made up of rows and columns. Here is an example of a tabular dataset:

| name | age | major |
|---|---|---|
| Bob | 25 | Political Science |
| Jane | 20 | Psychology |

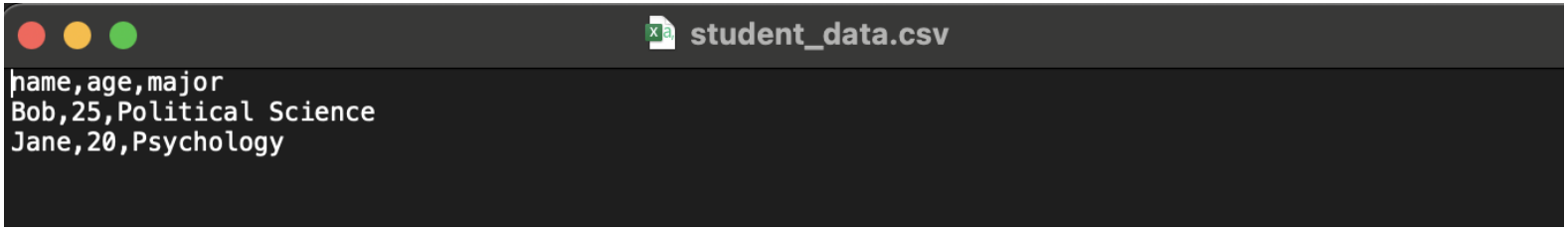
# Tabular Data

| name | age | major |
|---|---|---|
| Bob | 25 | Political Science |
| Jane | 20 | Psychology |

- Here we have a dataset with 2 rows and 3 columns (we don't include the first row in our count because this row only contains our column names). The column names denote **variables**, while the data in the rows contains individual values on these variables.
- This dataset contains data about two college students. Specifically, each student's *name*, *age*, and *major*.
- We call this dataset *tidy* because each row is a single observation (i.e., a single student), and each column is a single variable.

# Comma-Separated Value (.csv) Files

- A comma-separated value file (.csv) is a data file where a comma separates one value from another.
- Here's our student data, stored as a .csv file:




- Note that it looks *very* similar to the format we saw earlier, only now columns (rather than lines) separate the values.

# Reading in .csv files

- While `R` has a base `read.csv()` command, we will use the function `read_csv()` from the `readr` package.
- Reading in our file is fairly simple. All we need to do is specify the file location (or **path**) on our computer with `file`, and save the result to a new object in `R` using `<-`.

# Reading in .csv files

```
student_data <- read_csv(file="data/student_data.csv")
```

Now we have a nice `tibble` (a special type of `dataframe`), which we use to store data in R, called `student_data`, containing the dataset we need.

```
student_data
```

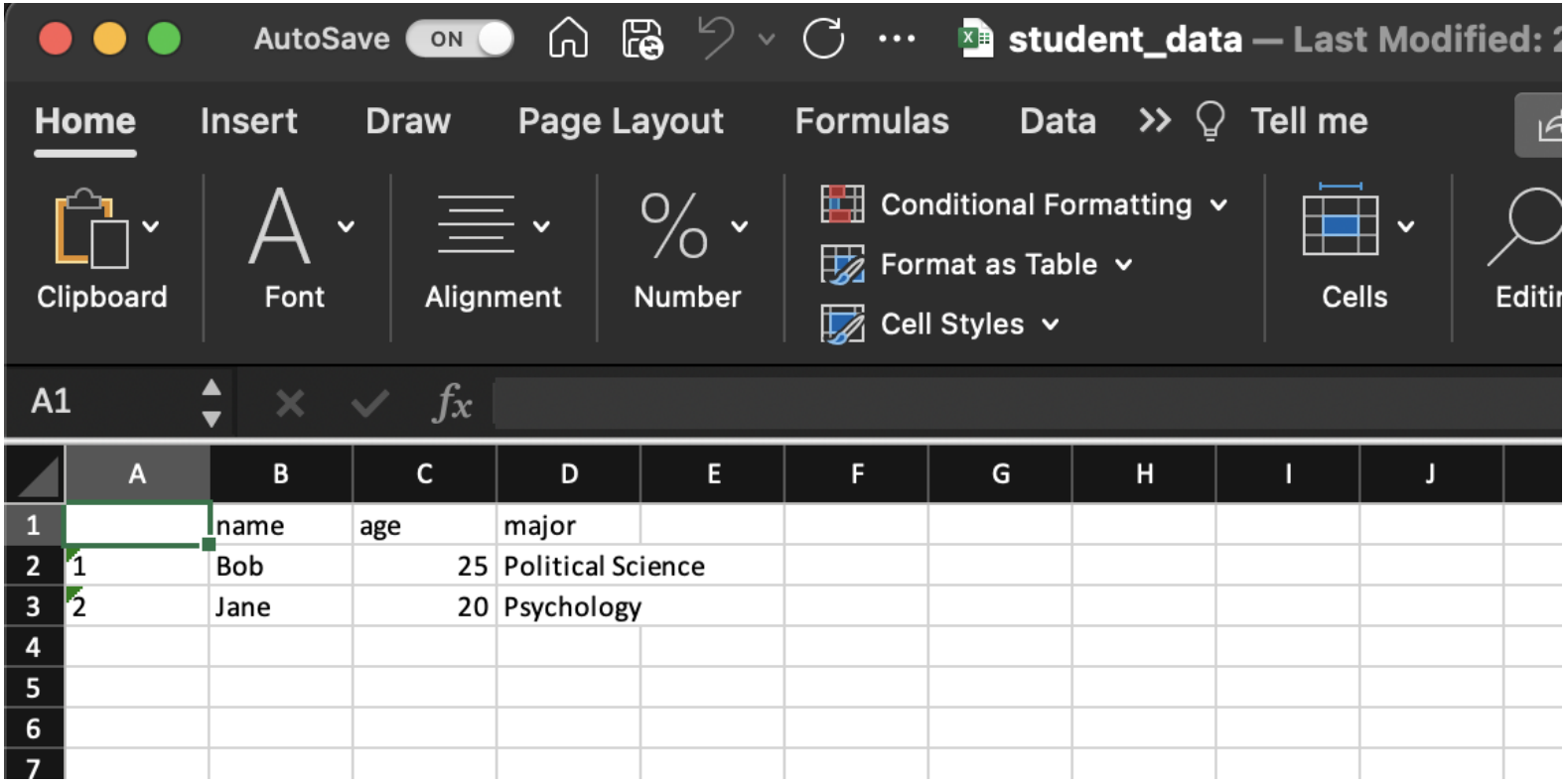
```
## # A tibble: 2 x 3
##   name    age major
##   <chr> <dbl> <chr>
## 1 Bob      25 Political Science
## 2 Jane     20 Psychology
```

# A bit more about `read_csv()`

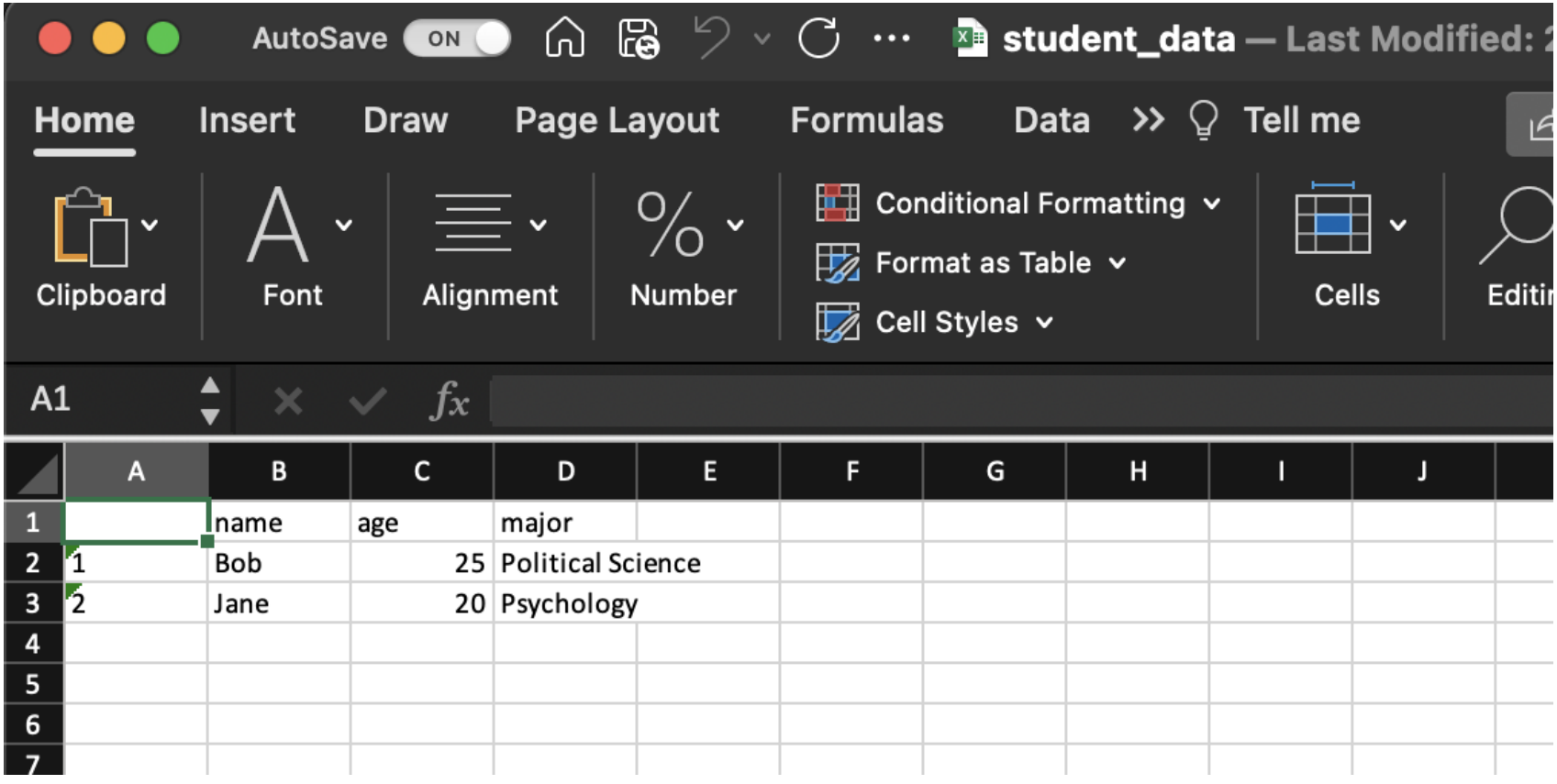
- `read_csv()` is smart. By default, it assumes that the first row of the file contains our column names, so it automatically names our columns based on that row.
- If by chance the data file does not have column names, you will need to specify `col_names=FALSE` when you run the code.
- Make sure to save your dataset to an `R` object using `<-`. You can name the data whatever you want, so long as you remember the name.

# Microsoft Excel Files

- If you have taken a computer class before, you probably have worked with Microsoft Excel. Excel is a program for storing and working with data. While we will be doing our data analyses in R, it's not uncommon to get a data file created with Excel. These files have the extension *.xls* or *.xlsx* .
- Here is our student data again, this time stored as a .xlsx file:



# Microsoft Excel Files




| | A | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | name | age | major | | | | | | |
| 2 | 1 | Bob | 25 | Political Science | | | | | | |
| 3 | 2 | Jane | 20 | Psychology | | | | | | |
| 4 | | | | | | | | | | |
| 5 | | | | | | | | | | |
| 6 | | | | | | | | | | |
| 7 | | | | | | | | | | |

- Notice that Excel decided to give this file row numbers - we will have to keep that in mind when we read the file into R, because we don't want R to think those numbers represent actual data.

# Microsoft Excel Files

- Sometimes you may need to open the .xlsx/.xls file before reading it in, in order to see what you're working with.
- You don't need Microsoft Excel installed on your computer to open one of these files (most computers come with free software that can open them), but if you would like it installed, UMass does offer Microsoft 365 at no cost to students - see [this link](https://www.umass.edu/it/software/microsoft-office-365-education) for more info.