Module 2 – lesson 01: Document Components

Script

In this lesson, you are going to learn a lot more about R markdown. I encourage you to read through the supplementary materials provided and spend some time reviewing the R markdown website by RStudio at <http://rmarkdown.rstudio.com/index.html>

[DEMO here of the website and quick overview of what is available – especially point out the steps and information provided in “Getting Started” – PLUS the Gallery, Formats and Articles]

However, let’s also do a quick Overview of the components contained in an R Markdown Document:

At the top of an R markdown document you will typically have the “document metadata.” This metadata is contained in a YAML header that has:

* + information about the document;
  + various parameters; and
  + formatting options; and
  + other options that can be customized

After the YAML header, comes the BODY of Document. The rest of the document will consist of:

* + Plain Text – this plain text will contain the content of your document along with
    - Rmarkdown syntax for this course which is based on markdown
  + Most likely you’ll also have some programming code included – these are called code chunks
    - For this course you will be learning the R language, but in RStudio using the R markdown package there are options for including other computer languages like Python, Rcpp (R’s version of C plus plus), the SQL database language and STAN (used for Bayesian statistical analyses).
  + You may also have other embedded “objects” such as:
    - Figures; images; photos; pictures
    - Tables
    - Videos; animations
    - Equations
    - References;
    - Or footnotes

So, let’s start at the beginning with the YAML document header which contains your document’s “metadata”. What is Document “Metadata”?

Every document (and file for that matter) you create has metadata. Metadata is often referred to as “data about data” – in other words, metadata is information about your document or file – not “data” within your document.

[DEMO on computer]

Let’s see an example. On a Windows machine, if you open the file explorer and right click on a file (like a Word DOC file) and view “properties” and then click on “details” you get a lot of information about the file or document including:

* The original author or user who created the file or document
* Last date and time the file was saved and by who (which user)
* The revision number
* The program used to view/edit the file
* The name of the company or manager for that file or software license (if available)
* Depending on the file type sometimes there is a description of the content in the file
* The file size (amount of hard drive space used to store the file)
* And much more
* Other file types (like computer programs or scripts) may have different information – for example, images and pictures will usually have information on the image size (e.g. width and height in pixels)

When we create an `R markdown` document we will have direct control over some of this metadata through information contained in the header of the document. This header information is provided and structured using YAML. So, what is YAML? It is pronounced to rhyme with camel.

YAML stands for “yet another markup language” or “YAML ain’t markup language” depending on who you ask. You can learn a lot about YAML simply by searching the Internet using your favorite search engine. You can also visit the official YAML website at yaml.org but this website is aimed at programmers and is not very user friendly. However, the yaml.org website does offer some insight into how many programming languages and platforms YAML supports besides R markdown.

Technically, it is possible to create an R markdown file without a YAML header, but for this course we will always have some information contained and defined within the YAML header. We will be using the YAML header to define the parameters (or options) used by the Rmarkdown R package to “render” the final document. We’ll talk more about the render function in future lessons, but for now, just understand that the information contained in the YAML header is used as instructions and input parameters for creating the final document.

[DEMO on computer]

Let’s create a Github repository for this next exercise. Log into your Github account and create a new repository. Name your repository “Module2\_rmd1” for your first R markdown document exercise in Module 2. Type in a description and add a README file and click create repository.

Once you have your Github repository created – click on “Clone or Download” and copy the URL to the clipboard.

Open RStudio and Click on File/New Project. Choose Version Control using Git. Paste in the URL you just copied and MAKE SURE you are creating the repository in the folder you created for this course “C:\RepResearchCourse” – that way you’ll have all of the content for this course organized in one central location on your local drive.

Now that we’ve got a new RStudio project created, let’s go ahead and create a new R markdown document. Click on File/New File/R Markdown. We will be creating a “document” – type in a title for your document like “Module 2 - R Markdown Document 1”. Type in your name as author – it may already be entered. And keep the default output format as HTML. Click OK.

Everything you just typed in – your title, author name and the output format – you will see again when we review the YAML header information next.

When the document opens, you’ll notice that the TAB is called “untitled” and there is an icon next to it that looks like a document with a red circle over it. It is hard to see on my screen but the letters “RMD” are inside the red circle. Let’s go ahead and save this file and name it module2\_rmd1. Click Save. The file format defaults to the RMD format “module2\_rmd1.Rmd” and the file name changes in the TAB at the top.

Now, let’s look at the document. At the very top there are 3 dashes - - - these indicate the BEGINning of the YAML header content. You should see information for the:

* Title
* Author
* Date
* Output

After the output there are 3 more dashes - - - indicating the END of the YAML header.

The words “title” “author” “date” and “output” are all YAML “key words” or parameters or options used by the render function in the R markdown package which compiles and creates the final document.

After each key word there is a colon : followed by your input for each parameter. For example, your title input is contained within beginning and ending double quotation marks “Module 2 - R Markdown Document 1”. Similarly on the next line you have the key word author followed by a colon : and then in between beginning and ending double quotation marks you have your name. Likewise for the date. In a future lesson, I will show you how to use r code to automatically change the date to the current date and time if you wish.

Currently, the last line in the YAML header has the key word output : html\_document. This was defined when we first created the New File/R Markdown document.

We will be learning a lot more about this YAML key word for output since this is where we define the parameters for customizing the various output formats we want. For now, watch what happens if we select different KNIT options. Let’s first KNIT to HTML. Your document should open in the VIEWER window to the right.

If your document does not appear in the VIEWER window, check your KNIT options – click the gear icon next to KNIT and see which options are selected. The preview in window opens a new separate window. Preview in Viewer Pan open in the VIEWER window at the bottom right.

Now, let’s KNIT to WORD to create a DOCX formatted file. As soon as you select this option, watch the RMD file, a second output option is entered for word\_document: default. This additional option is added to your YAML header automatically. This is one example of how changing the options in your YAML header directly affects the output produced from your R markdown document.

You’ll notice that the DOCX file also opens in a new window to preview the WORD Document.

OPTIONALLY if you have LaTeX installed on your computer, you can also try KNIT to PDF and again this will open in a new window to preview the PDF document. And a 3rd line of text is added to the YAML header for pdf\_document: default.

For now this completes the overview of the YAML header. But we will be adding and removing more YAML keywords as we work through this course.

Go ahead and keep this R markdown file open for the next lesson.