STAT 19000 Project 8

Topics: Markdown, RMarkdown, computational reproducibility

Motivation: The ability to quickly reproduce an analysis is important. It is often necessary that other individuals will need to be able to understand and reproduce an analysis. This concept is so important there are classes solely on reproducible research! In fact, there are papers that investigate and highlight the lack of reproducibility in various fields. If you are interested in reading about this topic, a good place to start is the paper titled "Why Most Published Research Findings Are False", by John Ioannidis (2005).

Context: Making your work reproducible is extremely important. We will focus on the computational part of reproducibility. We will learn RMarkdown to document your analyses so others can easily understand and reproduce the computations that led to your conclusions. Pay close attention as future project templates will be RMarkdown templates.

Scope: Understand Markdown, RMarkdown, and how to use it to make your data analysis reproducible.

You can find useful examples that walk you through relevant material here or on scholar:

`/class/datamine/data/spring2020/stat19000project08examples.R`.

It is highly recommended to read through these to help solve problems.

Important note: It is highly recommended that you use https://rstudio.scholar.rcac.purdue.edu/. Use another system at your own risk. The version of RStudio on https://desktop.scholar.rcac.purdue.edu/ is 99.9.9, and is known to have some strange issues when running code chunks.

After each problem, we've provided you with a list of keywords. These keywords could be package names, functions, or important terms that will point you in the right direction when trying to solve the problem, and give you accurate terminology that you can further look into online. You are not required to utilize all the given keywords. You will receive full points as long as your code gives the correct result.

Don't forget the very useful documentation shortcut `?`. To use, simply type `?` in the console, followed by the name of the function you are interested in.

You can also look for package documentation by using `help(package=PACKAGENAME)`.

Sometimes it can be helpful to see the source code of a defined function. To do so, type the function's name without the `()`.

Question 1: understanding RMarkdown

Important note: You can either submit questions 1a - 1c in a new `question01.Rmd` file, or roll your solutions to 1a - 1c into the modified question 2 `stat19000project08question02.Rmd` file. If you choose to roll the solution into the `stat19000project08question02.Rmd` file, just add the solutions above the first header, "(Header 1) Worming through book data".

1a. *(1 pt)* Describe what Markdown is in 1-2 brief sentences. Describe what RMarkdown is in 1-2 brief sentences.

Item(s) to submit:
- 1-2 brief sentences inside a Markdown Cell describing what Markdown is.
- 1-2 brief sentences inside the same Markdown Cell describing what RMarkdown is.

1b. *(1 pt)* RMarkdown documents can be converted to other file types. Name two of these formats.

Item(s) to submit:

- A list of 2 formats in a Markdown Cell.

1c. *(1pt)* Assume you are writing a data analysis report for DataMine using RMarkdown, and you want to make the word "DataMine" bold in your text, and "data analysis" italicized. In addition, you want to provide a hyperlink in the word "article" to an article you read online. Type the commands you would use to do this. An example of markdown, would be if we wanted to strikethrough "DataMine", we would do `~~DataMine~~`.

Hint: *Your solution should be in a Markdown Cell. Much like the name implies, you can test out your markdown by running the cell like it contains code!*

Item(s) to submit:

- A single markdown cell with the word "DataMine" bolded, the words "data analysis" italicized, and the word "article" as a hyperlink to any article online.

Question 2:

(5 pts)

Open RStudio. We will use RStudio throughout the remaining of this project.

Your classmates created an RMarkdown document to continue exploring some of the data from STAT 19000 Project 2, and make their new analysis reproducible. However, they are running into some errors using RMarkdown, and aren't sure how to fix it. You volunteered to help them as long as they send you a list with the bugs they want fixed. Their RMarkdown can be found here or on scholar:

`/class/datamine/data/spring2020/stat19000project08question02.Rmd`.

Fix the document based on the list below. Do not modify any R code, just the RMarkdown parts.

1. *(0.5 pts)* Rename the author to be you. That is, replace "Your_name_here" with your name in the author information.
2. *(1 pt)* Your classmates did not understand how to make headers. Fix all "Headers" to be actual headers with the correct level instead of just a text in parenthesis. For example, for "(Header 1) This is a header", remove "(Header 1)" and make "This is a header" a Header 1.
3. *(0.5 pts)* They wanted to list the variable names they are going to use during the analysis in lines 21 to 28. Fix their code in lines 21 to 28 to make it a markdown *ordered* list.
4. They want the chunk of code named `testing-hypothesis` in the RMardown document found in lines 47 to 52 to:
    a. *(1 pt)* run but want neither the code nor the results from `length(title_length)` to be shown.
    b. *(1 pt)* have the plot (figure) be shown in the document.

Hint: *It is necessary to specify the code chunk options* ***after*** *the code chunk label.*

5. *(1 pt)* Your classmates tried to add the following chunk of code but the knitting process for RMarkdown is not working for some reason. Insert the chunk of code below at the end of the document (lines 62 to 75), and fix the issue so the document is rendered properly.

~~~
```{r testing-hypothesis}
# removing NAs
bookDF <- subset(subBookDF, subBookDF$language_code!='')
is_english <- grepl('en', bookDF$language_code)

# mean values
tapply(bookDF$average_rating, is_english, mean)
~~~

~~~
# histogram for en
hist(bookDF$average_rating[is_english], main = 'Average Rating - English')

# histogram for not en
hist(bookDF$average_rating[!is_english], main = 'Average Rating - Non-english')
```
~~~

Hint: *Great resources are the RMarkdown cheat sheet and the RMarkdown reference guide.*

Hint: *Taking a look at the examples we provided are always a good place to start.*

Hint: *For items 4 and 5 in the list of improvements, you may want to look at the chunk's options and name. We discussed these items in our example.*

Hint: *When you do item 5, read the error, it will tell you why there is an error.*

Item(s) to submit:

- The fixed version of `stat19000project08question02.Rmd`, in it's entirety as an `.Rmd`.
- The fixed version of `stat19000project08question02.pdf`, in it's entirety as a `.pdf`.

Question 3:

(2 pts)

Like problem (2), you should use RStudio for this question.

We started an RMarkdown document for you to document an analysis on a hotel-booking dataset. Our RMarkdown can be found here or on scholar:

`/class/datamine/data/spring2020/stat19000project08question03.Rmd`.

Copy, paste, and make additions to `stat19000project08question03.Rmd`, and use it to document your findings for this question. Make sure to modify your name in author's information (replace 'Your_name_here' with your name). The dataset within the document is pretty cool, and has information on 32 variables, including:

- type of hotel ("City Hotel", "Resort Hotel")
- type of deposit ("No Deposit", "Refundable", "Non Refund")
- how many booking changes occured
- customer type ("Transient", "Contract", "Transient-Party", "Group")
- how many adults/children/babies
- number of days between the booking date and the arrival date (`lead_time`)
- month of arrival
- number of special requests
- reserved room type
- assigned room type
- etc.

Use what you have learned so far (plots, apply functions, correlation, etc.) to find one factoid you think is interesting from this data. Make sure to include your code, and your factoid in the document. Your factoid and accompanying code can be as simple or complex as you want. Just have fun and play around with RMarkdown and the data.

You may or may not want to show certain figures, or code snippets. Maybe you don't want certain chunks of code to be evaluated, but you do want to show them. Mix and match the chunk options, how you see fit!

If you would like to "prettify" the RMarkdown document by including different headers, formatting, or other stylistic changes, please feel free to do so.

Note that we have included a small analysis to illustrate a factoid we thought was interesting, and to get you started.

Item(s) to submit:

- The modified version of `stat19000project08question03.Rmd`, in it's entirety as an `.Rmd`, including a cool factoid, and the code used to discover said factoid.
- The modified version of `stat19000project08question03.pdf`, in it's entirety as an `.pdf`. Note that the pdf may not contain all of the information depending on settings you've chosen – this is ok! That is why we are asking for the `.Rmd` as well.

Project Submission:

Submit your solutions for the project at this URL: https://classroom.github.com/a/0RdcQyjn using the instructions found in the GitHub Classroom instructions folder on Blackboard.

Important note: For this project you will definitely submit the following files: `stat19000project08question03.Rmd`, `stat19000project08question03.pdf`, `stat19000project08question02.Rmd`, `stat19000project08question02.pdf`. If you choose to submit question 1 in a separate file, you will also submit a `question01.Rmd` file. Otherwise, your solutions to questions 1a - 1c will be added to the top of your modified `stat19000project08question02.Rmd` above the first header, "(Header 1) Worming through book data".