An Attractive Template of a Reproducible Data Analysis Document for an Awesome Class Project

Mahbubul Majumder, PhD
Department of Mathematics
University of Nebraska at Omaha

September 4, 2014

Abstract

This first sentence of the abstract starts with a little background of the problem. The second sentence starts describing the problem with some motivation. If needed we add another sentence to increase the reader's interest and clarity. After the problem description is over, we need to describe what we have done, how we did it and what results we obtained. We include any challenges we face and solution of that challenges we made. We describe any experiment we did with their findings. We also clearly mention how our results or methods are better than any existing results or methods. Sometimes we like to mention how this work opens up some future works. Abstract should not be more than a paragraph preferably having less than 200 words.

1 Introduction

The first sentence of this section is so attractive that it made the reader concentrate on reading. The second sentence is so great that it made the reader forget the date. Rest of the sentences are so nice that at the end of the paragraph the reader can't just wait to see what is coming on the next paragraph. At this point reader does not mind if it gets a little technical.

The first sentence of any paragraph presents a clear message. The rest of the sentences just describe that idea and establish the facts so that the reader see the logical conclusion of the paragraph. The last sentence of the paragraph connects the following paragraphs or section.

Add some motivational pictures in this section whenever possible. This will provide the reader some relief from reading text after text. For example Figure 1 indeed make us happy that we have something else to concentrate. This motivational picture does not need to be generated from the data you are going to analyze. Notice that we added the **R** codes of generating the Figure 1.

```
plot(women)
```

Also please don't forget to explain in details about what this figure is telling. It is really a bad idea not to say anything about the figure when you added it. Its like a product you are selling to someone who don't want to buy it. So, you have to be very serious about selling it with convincing argument.

1.1 Preparing this document

This whole document is prepared using **R** [**R-base**] package `knitr` [**R-knitr**]. It is a dynamic document and reproducible any number of times for any data sets. To start our work conveniently we need to install **R**, `RStudio` and LaTeX [**lamport94**] . Once our installation is done we will configure `RStudio` to work with `knitr`. For this first install `knitr` using command `install.packages("knitr")` and include the `knitr` library by command `library(knitr)`. Once `knitr` is installed go to the `RStudio` menu `Tools > Global`

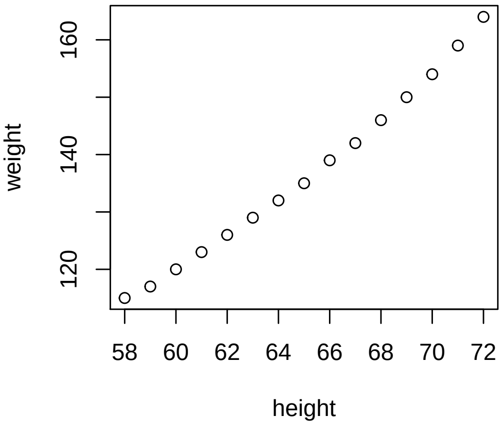

Figure 1: My awesome figure caption really describes what this figure is about and what we see in this figure. Also notice that the figure size is kept in such a way that it fits in the text nicely - not too big nor too small

`Options...Sweave` and change ‘Weave Rnw files using’ to indicate `Knitr`.

Now we are ready to create our first document using `knitr`. Go to `File > New File > R Sweave` and it will start with a new template for a document. If you save this minimal template it will be saved as a `.Rnw` file. Now we can just start filling the template with our texts. To create a human readable pdf file from `.Rnw` we just click on `Compile PDF` in `RStudio` toolbar.

PDF latex failure: If you encounter any problem such as `Running pdflatex on ...failed` it could be due to the bibliography. To solve that problem what you can do is: Go back to the folder where you saved your `.Rnw` file and find the `.tex` file that is created automatically. Now run the `.tex` file from LaTeX editor to create the pdf. Once you do this multiple times your bibliographies would be updated and you will be ready to work from `RStudio` as long as you don't change any object that has references in the file. There may be a better solution for this, but so far this worked for me.

The solution for this problem: just add `\usepackage[backend=bibtex]biblatex` in your preamble of the `.Rnw` file.

2 About the data

In this section we would like to describe the data we are going to use for the analysis. For example we intend to study the trees data that comes with default **R** installation. There are 31 data points and 3 variables in this data set. The variables are Girth, Height, Volume. The data may not be tidy and we may have to prepare the data before our analysis can be done. We will discuss how we prepared the data in the following section.

2.1 Preparing data

You invested lot of times preparing your data for exploration. Why not you describe what you did and how you did. You may add your R codes so that others know what exactly you did. For example let us view the summary of the data as below.

```
summary(trees)

##      Girth          Height       Volume
##  Min.   : 8.3   Min.   :63   Min.   :10.2
##  1st Qu.:11.1   1st Qu.:72   1st Qu.:19.4
##  Median :12.9   Median :76   Median :24.2
##  Mean   :13.2   Mean   :76   Mean   :30.2
##  3rd Qu.:15.2   3rd Qu.:80   3rd Qu.:37.3
##  Max.   :20.6   Max.   :87   Max.   :77.0
```

Try to avoid putting raw output like this in your final report. Instead make a clean table as shown in table 1. If you have to keep some raw output of your analysis please put them in a section called appendix at the end of the document. If you really believe that you have to put them here, you can do that and thats why we have this example here.

2.2 What is funny

This section may not be necessary. But if you notice something about the data that does not make any sense you can mention them in a section like that. Or if you think of anything interesting about the data, just discuss them here.

3 Methods

This section will include the methods you are planning to use for your analysis. You should include some theoretical justification here. For example, why you think the method is applicable, what are the assumptions about the methods, whether your data satisfies those assumption or not etc.

3.1 The model

These theories may require you to type mathematical equations and we need to refer them in the text like equation 1.

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

where $\epsilon \sim N(0, 1)$.

You should discuss the exploratory steps and the logical conclusion of adopting equation 1 for fitting to your data. Clearly mention the conditions and the assumptions of the model. Do not write any result of the model in this section. This section is only for theoretical discussion and any results of these models should be discussed in results section.

3.2 Data product

You may end up building a data product in your project. You may discuss about the plan here.

4 Results

In result section you can start with an overview of what you have found during the exploration of data.

4.1 Including tables

Include some summary tables of the data as as shown in table 1. Make sure you discuss about the table you have included and explain the facts it is revealing. You have to sell your table in a way that the reader will understand that this table was awesome and it reveals a fact the reader would otherwise not recognize.

Notice that we used the function `xtable()` form the **R** package `xtable` [**xtab**] to generate a pretty table. `knitr` does this using LaTeX codes generated by `xtable` and automatically put it in a nicer we and we don't have to worry about its position. Also notice how we write the caption of the table as well as refer the table 1 from the text.

```r
# Creating and printing summary data table
library(xtable)
summary_data <- apply(trees, 2, function(x) {
    return(c(Average = mean(x), Median = median(x), SD = sd(x), Range = range(x)))
})
print(xtable(summary_data, digits = 2, caption = paste("This table caption really",
    "describes what this table is about and what interesting facts it is revealing."),
    label = "summary-data"), caption.placement = getOption("xtable.caption.placement",
    "top"))
```

Table 1: This table caption really describes what this table is about and what interesting facts it is revealing.

	Girth	Height	Volume
Average	13.25	76.00	30.17
Median	12.90	76.00	24.20
SD	3.14	6.37	16.44
Range1	8.30	63.00	10.20
Range2	20.60	87.00	77.00

4.1.1 Book quality table

We can add tables that look like the tables in the book. For this we need to add package `booktabs` in the preamble of this .Rnw file. This will include a package called `booktabs` onto LaTeX. Once we add that we can now put option `booktabs = TRUE` in the **R** code as below.

```r
library(knitr)
x <- head(mtcars)
kable(x,format = 'latex', booktabs = TRUE)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

4.2 Including figures

Please don't forget to add nice data plots in your documents. Plots are nice to conveying message and much better than tables. Discuss what facts the figure is revealing and refer the figure from the text as figure 2.

```r
plot(trees)
```

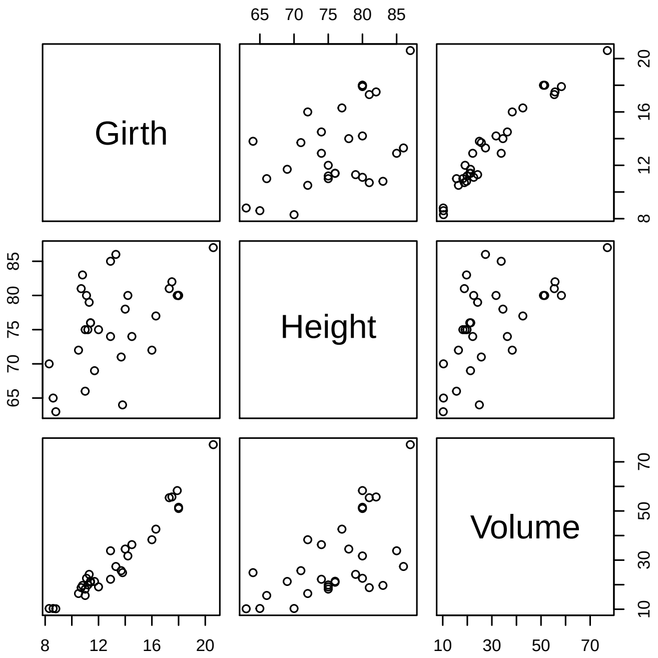

Figure 2: Awesome figure caption

4.3 How the data product works

If you build a data product you may discuss here how it works and what it provides. For data product being your main purpose, your main section may be different from just saying `Results`. You may think how you rename your sections to naturally fit in your work and the purpose.

5 Conclusion

The conclusion is an elaboration of your abstract. Here you will discuss what you have done and how. The gist of the results need to be mentioned here. It needs to be convincing and the reader will never regret forgetting the date. Please keep it in mind that there may be readers who only read your conclusion. So, make your conclusion complete so that no reader misses anything even if they don't want to read the whole document.

Each paragraph of the conclusion may discuss one result you have found or one concept you are proposing. Discuss your findings and why it is better and how it is compared to any existing methods may exist.

Please don't forget to cite the works of others if you used it in your analysis. The citation is important for two reasons. Fist of all it acknowledges the good works other people have done which encourages them keep continue doing their good work. Second, it protects you from plagiarism which is a very nasty task everyone should avoid.

There should be one paragraph about the future direction of the work you have done. You would like to make it so fascinating that the reader would wish to be involved in this work in future.

Finally this is just a template. Your exact document may have a very different outlook. It demonstrates how you can start to write a document. Our biggest problem is to figure out where to start from. And this documents provides a guide for that. I hope it turns out to be helpful for some of the readers. If you have any comments or concern about this document please let me know so that I can improve this document.

References

[1] David B. Dahl, *xtable: Export tables to LaTeX or HTML*, R package version 1.7-3, http://CRAN.R-project.org/package=xtable, 2014

[2] Leslie Lamport, *LaTeX: A Document Preparation System.* Addison Wesley, Massachusetts, 2nd Edition, 1994.

[3] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, http://www.R-project.org/ , 2014

[4] Yihui Xie *knitr: A general-purpose package for dynamic report generation in R*, http://yihui.name/knitr/, 2014