Intro to R

Sean Raleigh

Welcome to R! This document will walk you through some of the important features of R.

How these assignments work

In the first assignment on getting to know R Markdown, you were provided with a PDF and asked to type things into a new R Markdown file. From here on out, I will be giving you the R Markdown file and asking you to modify it in various places.

Sometimes you will be asked to add your own R code. That will appear in this document as a code chunk with a request for you to add your own code, like so:

```
## Add code here to [do some task]...
```

When you see that in a code chunk, you need to type some R code to complete a task.

Sometimes you will be asked to type up your thoughts. Instructions to do that will be labeled as follows. If you are currently reading this in the knit output, please look back at the R Markdown file to see the following text:

(When you knit the document, you can't see the text from the line above. That's because the crazy notation surrounding that text is an HTML comment, and therefore doesn't appear in the output.) In these areas of the assignment, please use full sentences and proper spelling, grammar, punctuation, etc. This is not R code, but rather a free response section where you talk about your analysis and conclusions.

Getting started

Make sure you're in a project

In the introduction to R Markdown, you were taught how to start a new project in RStudio. If you're looking at this document, you should have already created a project and uploaded this R Markdown file to that project folder.

Save your file!

The first thing we **always** do is save our file. You'll probably want to save this under a new name. Go to the "File" menu and then "Save As". Once you've saved the file with the new name, from them on it's easier to just hit Ctrl-S (or cmd-S on a Mac) to keep saving it periodically.

Remember that file names should not have any spaces in them. (In fact, you should avoid other kinds of special character as well, like periods, commas, number signs, etc. Stick to letters and numerals and you should be just fine.) If you want a multiword file name, I recommend using underscores like this: `this_filename_has_spaces_in_it`.

Load Packages

Packages are sets of commands and functions that people all over the world write. These packages extend the capabilities of R and add useful tools. The following code chunk will load the `mosaic` package that you

installed previously and an additional package `MASS` that comes prepackaged with R. We'll need the `MASS` packages because it comes with an interesting data set on risk factors associated with low infant birth weight.

```
library(mosaic)
library(MASS)
```

Exploring data

Let's look at the birth weight data. You'll need to type library(MASS) in the Console to load the MASS package into your current R session. (It's not enough that the command appears in this R Markdown document a few lines up. That only loads the package for purposes of knitting the R Markdown document.) Then type ?birthwt in the Console to get the help file for the birth weight data. Take a moment to familiarize yourself with the variables. You should also type View(birthwt) in the Console to get a "spreadsheet" view of the data. (Don't forget that the `View` command has an uppercase V.)

Heads and tails

We can now talk about ways to summarize the data. The `head` command shows the first six rows of the dataset.

```
head(birthwt)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
```

Remember that you will need to knit the document to see the result of doing this. (As I mentioned in the previous assignment, I would knit to HTML while you are working on the document, and only knit to PDF when you are completely finished and ready to turn in the final product.)

Verify that these really are the first six rows by looking at the spreadsheet version that resulted from the `View` command.

If you want to see more/fewer rows, you can change this:

```
head(birthwt, 10)
```

```
##    low age lwt race smoke ptl ht ui ftv  bwt
## 85   0  19 182    2     0   0  0  1   0 2523
## 86   0  33 155    3     0   0  0  0   3 2551
## 87   0  20 105    1     1   0  0  0   1 2557
## 88   0  21 108    1     1   0  0  1   2 2594
## 89   0  18 107    1     1   0  0  1   0 2600
## 91   0  21 124    3     0   0  0  0   0 2622
## 92   0  22 118    1     0   0  0  0   1 2637
## 93   0  17 103    3     0   0  0  0   1 2637
## 94   0  29 123    1     1   0  0  0   1 2663
## 95   0  26 113    1     1   0  0  0   0 2665
```

Now you try:

```
## Add code here to show only the first three cases of the dataset.
```

Experiment with the `tail` command.

```
## Try figuring out how to use the tail command. Verify that it works as expected by comparing your outp
```

Summary

We need to be able to summarize variables. The `summary` command is one way:

```
summary(birthwt)
```

```
##       low              age             lwt             race      
##  Min.   :0.0000   Min.   :14.00   Min.   : 80.0   Min.   :1.000  
##  1st Qu.:0.0000   1st Qu.:19.00   1st Qu.:110.0   1st Qu.:1.000  
##  Median :0.0000   Median :23.00   Median :121.0   Median :1.000  
##  Mean   :0.3122   Mean   :23.24   Mean   :129.8   Mean   :1.847  
##  3rd Qu.:1.0000   3rd Qu.:26.00   3rd Qu.:140.0   3rd Qu.:3.000  
##  Max.   :1.0000   Max.   :45.00   Max.   :250.0   Max.   :3.000  
##      smoke            ptl               ht                ui        
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000   Min.   :0.0000  
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.0000  
##  Median :0.0000   Median :0.0000   Median :0.00000   Median :0.0000  
##  Mean   :0.3915   Mean   :0.1958   Mean   :0.06349   Mean   :0.1481  
##  3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.00000   3rd Qu.:0.0000  
##  Max.   :1.0000   Max.   :3.0000   Max.   :1.00000   Max.   :1.0000  
##       ftv              bwt        
##  Min.   :0.0000   Min.   : 709  
##  1st Qu.:0.0000   1st Qu.:2414  
##  Median :0.0000   Median :2977  
##  Mean   :0.7937   Mean   :2945  
##  3rd Qu.:1.0000   3rd Qu.:3487  
##  Max.   :6.0000   Max.   :4990
```

Take a look at the output of the previous code chunk in the knit document. We may not have talked about all this in class yet, so you may not recognize the "Median" or the "1st Quartile" or "3rd Quartile". Nevertheless, you can see why this would come in handy.

Question: There are only some of the variables for which this summary makes sense. For example, something is weird about taking the mean of the `race` variable. What's wrong? List all the variables for which this summary is inappropriate.