

REPRODUCIBLE RESEARCH

R MARKDOWN + KNITR

PETRA ISENBERG

VISUAL ANALYTICS

MARKDOWN

"Markdown is a text-to-HTML conversion tool for web writers. Markdown allows you to write using an easy-to-read, easy-to-write plain text format, then convert it to structurally valid XHTML (or HTML)."

John Gruber, creator of Markdown

EXAMPLE

This text will appear italicized!



This text will appear italicized!

EXAMPLE

`**This text will appear bold!**`



This text will appear bold!

EXAMPLE

This is a secondary heading

This is a tertiary heading

This is a secondary heading

This is a tertiary heading

RESOURCES

- <http://daringfireball.net/projects/markdown>

R MARKDOWN

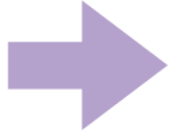
- R markdown files can be used to generate reproducible reports
 - Text and R code are integrated
 - Very easy to create in Rstudio
- no need to generate Readme files

R MARKDOWN

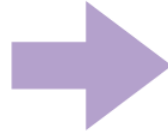
- R markdown is the integration of R code with markdown
- Allows one to create documents containing "live" R code
- R code is evaluated as part of the processing of the markdown
- Results from R code are inserted into markdown document
- A core tool in literate statistical programming

WORKFLOW

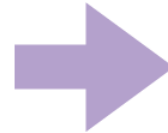
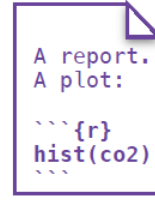
i. Open - Open a file that uses the .Rmd extension.



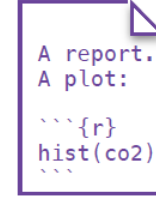
ii. Write - Write content with the easy to use R Markdown syntax



iii. Embed - Embed R code that creates output to include in the report



iv. Render - Replace R code with its output and transform the report into a slideshow, pdf, html or ms Word file.



=



=



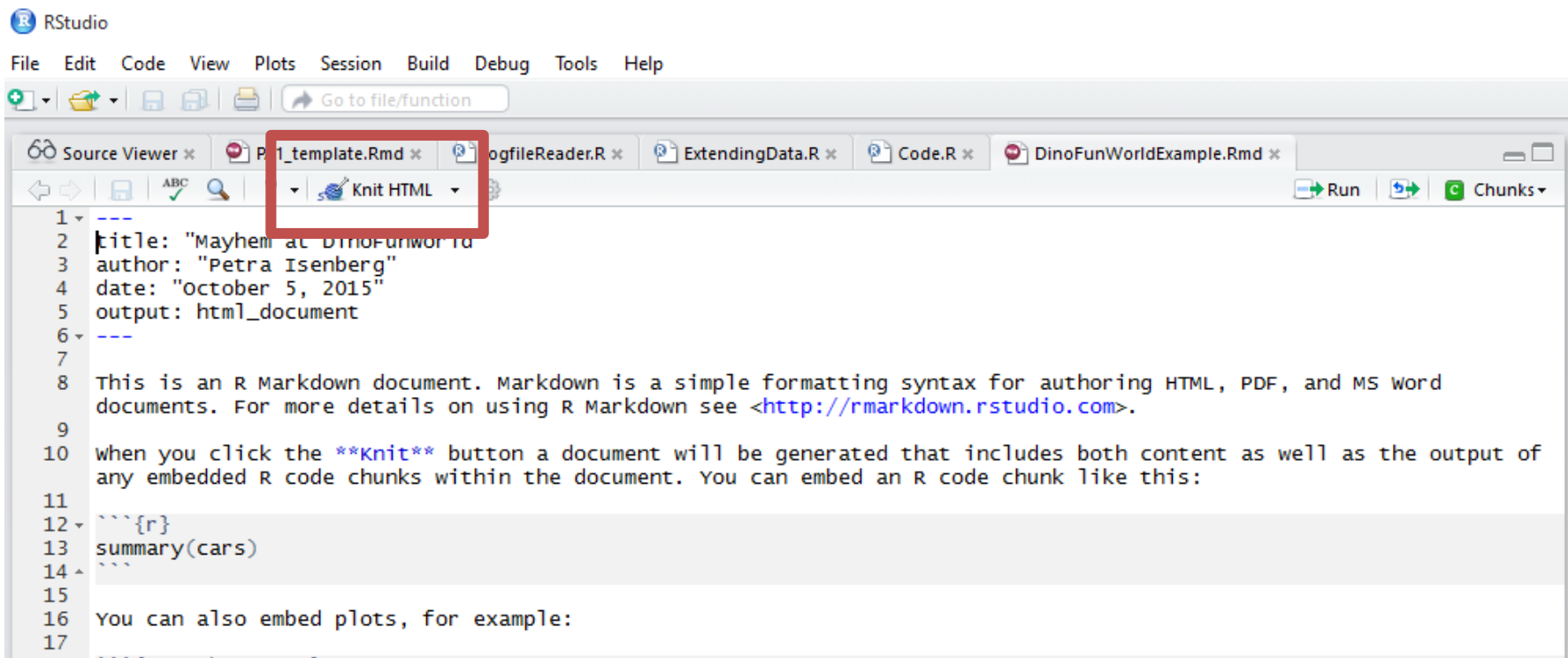
OPEN FILE

- In Rstudio
 - File -> New File -> R Markdown...
 - Give it a title, leave defaults, click OK

```
1 ---
2 title: "Mayhem at DinoFunWorld"
3 author: "Petra Isenberg"
4 date: "October 5, 2015"
5 output: html_document
6 ---
7
8 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS word
9 documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.
10
11 when you click the **knit** button a document will be generated that includes both content as well as the output of
12 any embedded R code chunks within the document. You can embed an R code chunk like this:
13
14 ```{r}
15 summary(cars)
16 ```
17
18 You can also embed plots, for example:
19
20 ```{r, echo=FALSE}
21 plot(cars)
22 ```
23
24 Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated
25 the plot.
```

NEXT

- Click on the Knit HTML Button



The screenshot shows the RStudio interface. The top menu bar includes File, Edit, Code, View, Plots, Session, Build, Debug, Tools, and Help. Below the menu bar is a toolbar with icons for file operations and a search bar labeled "Go to file/function". The main window displays several open files: Source Viewer, P1_template.Rmd, logfileReader.R, ExtendingData.R, Code.R, and DinoFunWorldExample.Rmd. The "P1_template.Rmd" file is selected, and its content is visible in the editor. A red box highlights the "Knit HTML" button in the toolbar. The code in the editor includes a YAML header, a paragraph of text, and an R code chunk.

```
1 ---  
2 title: "Mayhem at DinoFunworld"  
3 author: "Petra Isenberg"  
4 date: "October 5, 2015"  
5 output: html_document  
6 ---  
7  
8 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word  
documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.  
9  
10 when you click the **knit** button a document will be generated that includes both content as well as the output of  
any embedded R code chunks within the document. You can embed an R code chunk like this:  
11  
12 ```{r}  
13 summary(cars)  
14 ```  
15  
16 You can also embed plots, for example:  
17
```

```
1 ---
2 title: "Mayhem at DinoFunworld"
3 author: "Petra Isenberg"
4 date: "October 5, 2015"
5 output: html_document
6 ---
7
8 |
```

Erase the example code

```
1 ---
2 title: "Mayhem at DinoFunworld"
3 author: "Petra Isenberg"
4 date: "October 5, 2015"
5 output: html_document
6 ---
7
8 |
```

Lets do something useful for your assignment...

EXPLORATORY DATA ANALYSIS

```
#Exploratory Data Analysis with RMarkdown
```

```
##Loading the Data
```

```
` `` {r}
```

```
data <- read.csv("Paper-Author.csv")
```

```
` ``
```


Exploratory Analysis Tutorial 4

Petra Isenberg

September 28, 2016

Exploratory Data Analysis with RMarkdown

Loading the Data

```
data <- read.csv("Paper-Author.csv")
```

```
##Inspecting the Data with R
```

```
###The first few lines of the dataset
```

```
` `{r}  
head(data)  
` ` `
```

```
###A summary of the dataset
```

```
` `{r}  
str(data)  
` ` `
```

Loading the Data

```
data <- read.csv("Paper-Author.csv")
```

Inspecting the Data with R

The first few lines of the dataset

```
head(data)
```

```
##           Paper.DOI Deduped.author.names
## 1 10.1109/TVCG.2015.2467324 Rubio-Sanchez, M.
## 2 10.1109/TVCG.2015.2467324           Raya, L.
## 3 10.1109/TVCG.2015.2467324           Diaz, F.
## 4 10.1109/TVCG.2015.2467324 Sanchez, A.
## 5 10.1109/TVCG.2015.2467471           Setlur, V.
## 6 10.1109/TVCG.2015.2467471           Stone, M.C
```

ARE THE FIELDS CORRECT?

WHAT ABOUT THE VALUES?

WHAT ABOUT THE DATA TYPES?

A summary of the dataset

```
str(data)
```

```
## 'data.frame':   9667 obs. of  2 variables:
## $ Paper.DOI      : Factor w/ 2752 levels "10.0000/000000001",...: 1166 1166 1166 1166 1182 1182 1220 122
0 1229 1229 ...
## $ Deduped.author.names: Factor w/ 4890 levels "", "Abbasloo, A.",...: 3528 3379 882 3598 3756 4018 1937 1785
3860 115 ...
```

SOME FIRST STEPS TO START WITH

1. Plot the raw data
2. **Plot simple statistics**
3. Look at plots together

```
##Some Simple Statistics
```

```
` `` {r}
```

```
summary(data)
```

```
` ``
```

Some Simple Statistics

```
summary(data)
```

```
##          Paper.DOI          Deduped.author.names
## 10.1109/VAST.2011.6102498 : 17 Groller, E.      : 58
## 10.1109/VISUAL.2005.1532845: 17 Kaufman, A.    : 57
## 10.1109/VISUAL.2002.1183812: 15 Kwan-Liu Ma   : 51
## 10.1109/TVCG.2009.164      : 14 Ertl, T.      : 45
## 10.1109/TVCG.2012.278     : 14 Keim, D.A.    : 44
## 10.1109/TVCG.2014.2346911 : 14 van Wijk, J.J.: 38
## (Other)                   :9576 (Other)       :9374
```

```
##Some Simple Statistics
```

```
```{r}
```

```
summary(data)
```

```
countTable <- as.data.frame(table(data$Deduped.author.names))
```

```
colnames(countTable) <- c("Author", "Freq")
```

```
median <- c(median(countTable$Freq))
```

```
mean <- c(mean(countTable$Freq))
```

```
stdev <- c(sd(countTable$Freq))
```

```
measures <- c("mean", "stdev", "median")
```

```
values <- c(mean, stdev, median)
```

```
descriptiveStats <-data.frame(measures, values)
```

```
descriptiveStats
```

```
```
```

```
###Plotting some simple statistics
```

The average and standard deviation of average paper counts per author

```
` `` {r}
```

```
library(ggplot2)
```

```
ggplot(descriptiveStats, aes(x=measures, y=values))  
+ geom_bar(stat="identity")
```

```
`` `
```


Next, let's look at the distribution

```
```{r}  
qplot(countTable$Freq,
binwidth=1,geom="histogram")

boxplot(countTable$Freq,data=countTable)
```
```

Now try to find out how many authors are on a paper on average

RESOURCE

- <https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf>