Data Analysis and Visualization: R Workflow

Dr. Olga Scrivner, Research Scientist
January 16, 2018
Indiana University
Upcoming Events:

Monday, January 22, 2018 | 4:00 PM

Visual/Data Literacy Related to Maps
Theresa Quill
Indiana University Libraries

Monday, January 29, 2018 | 4:00 PM

Visualizing Science Using VOSviewer
Ludo Waltman
Centre for Science and Technology Studies (CWTS) at Leiden University

http://cns.iu.edu/netscitalks.html
Goals

1. Understand RStudio
2. Understand the difference between R scripts and R projects
3. Learn how to plan and manage R project
4. Learn how to deploy and share R project

TBA:

- R crash course
- Shiny basics and Shiny advanced
Overview

1. Project set-up and planning
2. Reporting and documenting
3. Visualizing
4. Sharing

(Wickham and Grolemund, 2017)
Project-oriented Workflow
Organize each data into a project:

File → New Project
Create Project

New Project

Create Project

New Directory
Start a project in a brand new working directory

Existing Directory
Associate a project with an existing working directory

Version Control
Checkout a project from a version control repository

Cancel
Project Type

New Project

Project Type

Empty Project
Create a new project in an empty directory

R Package
Create a new R package

Shiny Web Application
Create a new Shiny web application

Back

Cancel

7
Create New Project

Directory name:
Workshop–R–Workflow

Create project as subdirectory of:
~/Documents/CNS/Visual Insights/Rworkflow

- [ ] Create a git repository
- [ ] Use packrat with this project

Open in new session

Create Project

Cancel
New Project

R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"
Copyright (C) 2016 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin13.4.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY. You are welcome to redistribute it under certain conditions. Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors. Type 'contributors()' for more information and 'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or 'help.start()' for an HTML browser interface to help. Type 'q()' to quit R.

>
Building a Directory Structure

1. Install library **ProjectTemplate**

   ![Install Packages]

   - Install from: Repository (CRAN)
   - Packages (separate multiple with space or comma): ProjectTemplate
   - Install to Library: /Users/ogascriever/Library/R/3.3/library (Default)
   - ✓ Install dependencies

   ![Install]

2. Open a new R script: File → New File → R Script

   ```
   library('ProjectTemplate')
   create.project('Project Name', minimal = TRUE, 
                  merge.strategy = 'allow.non.conflict')
   ```

3. TIP: Change your working directory - one level up from your project
   Session → Set Working Directory → Choose Directory
Directory Structure

Extensions:

- R script - .R
- Readme file - .md (Markdown)
- Project - .Rproj
Packrat - stores your package dependencies inside the project.

Advantages:

1. **Isolation**: Installing a new or updated package for one project will not break your other projects.
2. **Portability**: Easily transport your projects from one computer to another, even across different platforms.
3. **Reproducibility**: Packrat records the exact package versions you depend on.

http://rstudio.github.io/packrat/
Project Planning
“smart preparation minimizes work” (Berkun, 2005)

https://csgillespie.github.io/efficientR/workflow.html
SMART Criteria

1. **Specific**: is the objective clearly defined and self-contained?
2. **Measurable**: is there a clear indication of its completion?
3. **Attainable**: can the target be achieved?
4. **Realistic**: have sufficient resources been allocated to the task?
5. **Time-bound**: is there an associated completion date or milestone?

https://csgillespie.github.io/efficientR/workflow.html
library('DiagrammeR')
# Define the Gantt chart and plot the result
mermaid("gantt
    Section Initiation
    Planning :a1, 2018-01-16, 10d
    Data processing :after a1 , 30d")

- **Section** refers to the project’s section (useful for large projects, with milestones)
- **Line** refers to a task
- **Example:** **Planning** begins on Jan 16 2018 and lasts for 10 days
Project Management

Large projects: regular meetings, division of labour, tracking progress, issues and priorities (Gillespie and Lovelace, 2017, Chapter 4)

1. The interactive code sharing site GitHub
2. ZenHub, a browser plugin that is “the first and only project management suite that works natively within GitHub”
3. Web-based and easy-to-use tools such as Trello
4. Dedicated desktop project management software such as ProjectLibre and GanttProject
5. Fully featured, enterprise scale open source project management systems such as OpenProject and redmine
Documenting and Reporting
R Script

https://example.com
1. Copyright statement comment
2. Author comment (Use #)
3. File description comment, including purpose of program, inputs, and outputs
4. `source()` and `library()` statements
   - `source("file name")` - read R code from a file
   - `library(name)` - package name
5. Function definitions
6. Executed statements, if applicable (e.g., plot)
### R Style Guide: Naming

#### File Names

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>predict_ad_revenue.R</td>
<td>foo.R</td>
</tr>
</tbody>
</table>

#### Variable Names

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>variable.name (preferred)</td>
<td>variable_Name</td>
</tr>
<tr>
<td>variableName (accepted)</td>
<td></td>
</tr>
</tbody>
</table>

#### Function Names (use action verbs)

<table>
<thead>
<tr>
<th>Good</th>
<th>Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>CalculateAvgClicks</td>
<td>calculate_avg_clicks</td>
</tr>
<tr>
<td>CalculateAvgClicks</td>
<td>calculateAvgClicks</td>
</tr>
</tbody>
</table>
Spaces around all binary operators (=, +, -, <-)

Use <-, not =, for assignment

Space after a comma

Incorrect: total == sum(x[1,])
Spaces around all binary operators (=, +, -, <-)

Use <-, not =, for assignment

Space after a comma

Incorrect: total == sum(x[1,])

Correct: total <- sum(x[1,])

Learn more: http://adv-r.had.co.nz/Style.html
“R Markdown files are the ultimate R reporting tool” (Grolemund, 2014)

R Markdown is a file format for making dynamic documents with R.

Markdown - an easy-to-write plain text format.

R Markdown files can be converted into HTML, PDF, and Word documents.
R markdown sample

Olga Scrivner
January 16, 2018

R Markdown

Click the **Knit** to generate the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```{r cars}
summary(cars)
```

<table>
<thead>
<tr>
<th>speed</th>
<th>dist</th>
</tr>
</thead>
<tbody>
<tr>
<td>Min.</td>
<td>Min.</td>
</tr>
<tr>
<td>4.0</td>
<td>2.00</td>
</tr>
<tr>
<td>1st Qu.</td>
<td>1st Qu.</td>
</tr>
<tr>
<td>12.0</td>
<td>26.00</td>
</tr>
<tr>
<td>Median</td>
<td>Median</td>
</tr>
<tr>
<td>15.0</td>
<td>36.00</td>
</tr>
<tr>
<td>Mean</td>
<td>Mean</td>
</tr>
<tr>
<td>15.4</td>
<td>42.98</td>
</tr>
<tr>
<td>3rd Qu.</td>
<td>3rd Qu.</td>
</tr>
<tr>
<td>19.0</td>
<td>56.00</td>
</tr>
<tr>
<td>Max.</td>
<td>Max.</td>
</tr>
<tr>
<td>25.0</td>
<td>120.00</td>
</tr>
</tbody>
</table>
Did You Know?

Markdown is used:

- Github
- StackOverflow
- Reddit
# R Markdown

<table>
<thead>
<tr>
<th><strong>Text using Markdown syntax</strong></th>
<th><strong>Corresponding HTML produced by a Markdown processor</strong></th>
<th><strong>Text viewed in a browser</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Heading</strong></td>
<td>&lt;h1&gt;Heading&lt;/h1&gt;</td>
<td><strong>Heading</strong></td>
</tr>
<tr>
<td><strong>## Sub-heading</strong></td>
<td>&lt;h2&gt;Sub-heading&lt;/h2&gt;</td>
<td><strong>Sub-heading</strong></td>
</tr>
<tr>
<td>Paragraphs are separated by a blank line.</td>
<td>&lt;p&gt;Paragraphs are separated by a blank line.&lt;/p&gt;</td>
<td>Paragraphs are separated by a blank line.</td>
</tr>
<tr>
<td>Two spaces at the end of a line leave a line break.</td>
<td>&lt;p&gt;Two spaces at the end of a line&lt;br /&gt; leave a line break.&lt;/p&gt;</td>
<td>Two spaces at the end of a line leave a line break.</td>
</tr>
<tr>
<td>Text attributes <em>italic</em>, <strong>bold</strong>, <code>monospace</code>.</td>
<td>&lt;p&gt;Text attributes &lt;em&gt;italic&lt;/em&gt;, &lt;strong&gt;bold&lt;/strong&gt;, &lt;code&gt;monospace&lt;/code&gt;.&lt;/p&gt;</td>
<td>Text attributes <em>italic</em>, <strong>bold</strong>, <code>monospace</code>.</td>
</tr>
<tr>
<td>Horizontal rule:</td>
<td>&lt;p&gt;Horizontal rule:&lt;/p&gt;</td>
<td>Horizontal rule:</td>
</tr>
<tr>
<td>---</td>
<td>&lt;hr /&gt;</td>
<td></td>
</tr>
<tr>
<td><strong>Bullet list:</strong></td>
<td>&lt;p&gt;Bullet list:&lt;/p&gt;</td>
<td></td>
</tr>
<tr>
<td>* apples</td>
<td>&lt;ul&gt;</td>
<td></td>
</tr>
<tr>
<td>* oranges</td>
<td>&lt;li&gt;apples&lt;/li&gt;</td>
<td></td>
</tr>
<tr>
<td>* pears</td>
<td>&lt;li&gt;oranges&lt;/li&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;li&gt;pears&lt;/li&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;/ul&gt;</td>
<td></td>
</tr>
<tr>
<td><strong>Numbered list:</strong></td>
<td>&lt;p&gt;Numbered list:&lt;/p&gt;</td>
<td></td>
</tr>
<tr>
<td>1. wash</td>
<td>&lt;ol&gt;</td>
<td></td>
</tr>
<tr>
<td>2. rinse</td>
<td>&lt;li&gt;wash&lt;/li&gt;</td>
<td></td>
</tr>
<tr>
<td>3. repeat</td>
<td>&lt;li&gt;rinse&lt;/li&gt;</td>
<td></td>
</tr>
<tr>
<td></td>
<td>&lt;/ol&gt;</td>
<td></td>
</tr>
</tbody>
</table>

[https://en.wikipedia.org/wiki/Markdown](https://en.wikipedia.org/wiki/Markdown)
Rendering

Run R code chunks separately

Render the entire document as HTML, PDF, or Word

https://en.wikipedia.org/wiki/Markdown
Here’s some code
```
\{r
  dim(iris)
\```

http://shiny.rstudio.com/articles/rmarkdown.html
Here’s some code
```
\{r echo=FALSE\}
dim(iris)
```

```
## [1] 150  5
```
Here’s a plot
```r
# {r echo=FALSE}
hist(iris[[2]])
```

http://shiny.rstudio.com/articles/rmarkdown.html
R Chunks - Inline

Two plus two equals `r 2 + 2`.

Two plus two equals 4.

http://shiny.rstudio.com/articles/rmarkdown.html
Websites

R Markdown makes it easy to build webpages straight from .Rmd files.

- **R Markdown**
  The R Markdown website is itself built with R Markdown. [Example Code](#).

- **flexDashboard**
  flexdashboard extends R Markdown to make administrative dashboards. Its website is also built from R Markdown. [Example Code](#).

- **bookdown**
  Bookdown extends R Markdown to make books. Its website is built with R Markdown and CSS styling. [Example Code](#).

- **profvis**
  profvis provides profiling tools for R code, as well as a website made with R Markdown. [Example Code](#).

http://rmarkdown.rstudio.com/gallery.html
Gallery: Get Inspiration

Interactive Documents
Combine R Markdown with htmlwidgets or the shiny package to make interactive documents.

**HTML Widgets**
Add interactive graphics with htmlwidgets, such as the leaflet map widget.

**HTML Widgets**
Embed htmlwidgets such as dygraphs and datatables directly into your reports.

**Shiny**
Add interactive analysis with shiny, which lets your user rerun the actual analysis within your report.

**Shiny**
Shiny components and htmlwidgets will work in any HTML based output, such as a file, slide show or dashboard.

http://rmarkdown.rstudio.com/gallery.html
Gallery: Get Inspiration

Presentations

R Markdown supports several presentation (slide show) formats.

Beamer slideshow
Create pdf slides with Beamer. Example Code.

Slidy slideshow
Create HTML-based slides with Slidy.

ioslides slideshow
Create HTML-slides with ioslides. Example Code

reveal.js slideshow
Create HTML-based slides with reveal.js. Example Code

http://rmarkdown.rstudio.com/gallery.html
Gallery: Get Inspiration

Dashboards
Combine R Markdown with the flexdashboard package to quickly assemble R components into administrative dashboards. Each example below contains a link to the source code within the dashboard.

Dashboard with gauges and value boxes
Use flexdashboard to create dashboards with gauges and value boxes.

Dashboard with htmlwidgets
Add interactive graphics to a dashboard with htmlwidgets.

Dashboard with Shiny
Add interactive analysis to a dashboard with Shiny.

Dashboard with storyboard
Organize dashboards around a storyboard.

http://rmarkdown.rstudio.com/gallery.html
Data Analysis and Visualization Workflow
1. **Import** data into R: `read_csv()`, `read_line()`, `read_delim()`
2. **Tidy** data - variables per column, observation per row
3. **Transform** with `dplyr`
4. **Visualize** with `ggplot` and `plotly`

(Wickham and Grolemund, 2017)
Data Transformation - dplyr

- Pick observations by their values - `filter()`
- Reorder the rows - `arrange()`
- Pick variables by their names - `select()`
- Create new variables with functions of existing variables - `mutate()`
- Collapse many values down to a single summary - `summarise()`

(Wickham and Grolemund, 2017, Chapter 5)
Practice: http://r4ds.had.co.nz/transform.html
Recommended Reading -
http://vita.had.co.nz/papers/layered-grammar.pdf

Visualization Template

```r
ggplot(data = <DATA>) +
  <GEOM_FUNCTION>(mapping = aes(<MAPPINGS>))
```
ggplot(data = mpg) +
  geom_point(mapping = aes(x = displ, y = hwy))
Exploratory Data Visualization

- `geom_point()`
- `geom_bar()`
- `geom_boxplot()`

http://r4ds.had.co.nz/exploratory-data-analysis.html
install.packages("plotly")
library(plotly)

https://plot.ly/ggplot2/
Publishing
Publishing

```r
---
title: "R markdown sample"
author: "Olga Scrivner"
date: "January 16, 2018"
output: html_document
---
```

**R Markdown**

Olga Scrivner  
January 16, 2018

R Markdown

Click the Knit to generate the output of any embedded R code chunks within the document. You can embed an R code chunk like this:
### Publishing

#### Publish To

<table>
<thead>
<tr>
<th>Publish To</th>
<th>RPubsv</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RPubs</strong></td>
<td>RPubs is a free service from RStudio for sharing documents on the web.</td>
</tr>
<tr>
<td><strong>RStudio Connect</strong></td>
<td>RStudio Connect is a server product from RStudio for secure sharing of applications, reports, and plots.</td>
</tr>
</tbody>
</table>

#### Publish to RPubsv

RPubs is a free service from RStudio for sharing documents on the web. Click Publish to get started.

**IMPORTANT:** All documents published to RPubs are publicly visible. You should only publish documents you wish to share publicly.

---

[Publish] [Cancel]
Easy web publishing from R

Write R Markdown documents in RStudio.
Share them here on RPubs. (It’s free, and couldn’t be simpler!)

Get Started

Recently Published
General Information
E583 | Z637 | Information Visualization MOOC 2018

This graduate level course provides an overview of the state of the art in information visualization. The course teaches visualization theory and the process of producing effective and actionable visualizations that take the needs of users into account. Students apply the visualization knowledge and skills that they gain in the course by working in teams on real-world client projects.

Data Visualization Literacy

[http://scimaps.org/atlas2](http://scimaps.org/atlas2)
THE END
@katycns
@obscrivn
#IVMOOC