Наука о данных в R для программы Цифровых гуманитарных исследований

Г. А. Мороз

Оглавление

1 Вступление **5**

2 Введение в R **7**

2.1 Наука о данных 7

2.2 Установка R и RStudio 8

2.3 Полезные ссылки 8

2.4 Rstudio 9

3 Трансформация данных **11**

4 Вмзуализация данных **13**

5 Условия и работа со списками **15**

6 Представление данных: rmarkdown, github, shiny **17**

7 Работа со строками **19**

8 Работа с текстами: tidytext, udpipe **21**

9 Сбор данных из интернета: rvest **23**

10 Нестандартные данные: время, OCR, карты **25**

Глава 1

Вступление

Материалы для курса Наука о данных для магистерской программы Цифровых гуманитарные исследования НИУ ВШЭ.

Глава 2

Введение в R

2.1 Наука о данных

Наука о данных — это новая область знаний, которая активно развивается в последнее время. Она находиться на пересечении компьютерных наук, статистики и математики и трудно сказать, действительно ли это наука. При этом это движение развивается в самых разных научных направлениях, иногда даже оформляясь в отдельную отрасль:

- биоинформатика
- цифровые гуманитарные исследования
- датажурналистика
- ...

Все больше книг “Data Scince for ...”:

- psychologists (Hansjörg, 2019)
- immunologists (Thomas and Pallett, 2019)
- buisness (Provost and Fawcett, 2013)
- public policy (Brooks and Cooper, 2013)
- fraud detection (Baesens et al., 2015)
- ...

Среди умений датасаентистов можно перечислить следующие:

- сбор и обработка данных
- трансформация данных
- визуализация данных
- моделирование данных
- представление полученных результатов

Все эти темы в той или иной мере будут представлены на нашем курсе.

2.2 Установка R и RStudio

В данной книге используется исключительно R (R Core Team, 2019), так что для занятий понадобятся:

- R
  - на Windows
  - на Mac
  - на Linux, также можно добавить зеркало и установить из командной строки:

```
sudo apt-get install r-cran-base
```

- RStudio — IDE для R (можно скачать здесь)
- и некоторые пакеты на R

Часто можно увидеть или услышать, что R — язык программирования для "статистической обработки данных". Изначально это, конечно, было правдой, но уже давно R — это полноценный язык программирования, который при помощи своих пакетом позволяет решать огромный спектр задач. В данной книге используются следующая версия R:

```
sessionInfo()$R.version$version.string
```

```
## [1] "R version 3.6.1 (2019-07-05)"
```

Некоторые люди не любят устанавливать лишние программы себе на компьютер, несколько вариантов есть и для них:

- RStudio cloud — полная функциональность RStudio, пока бесплатная, но скоро это исправят;
- RStudio on rollApp — облачная среда, позволяющая разворачивать программы.

2.3 Полезные ссылки

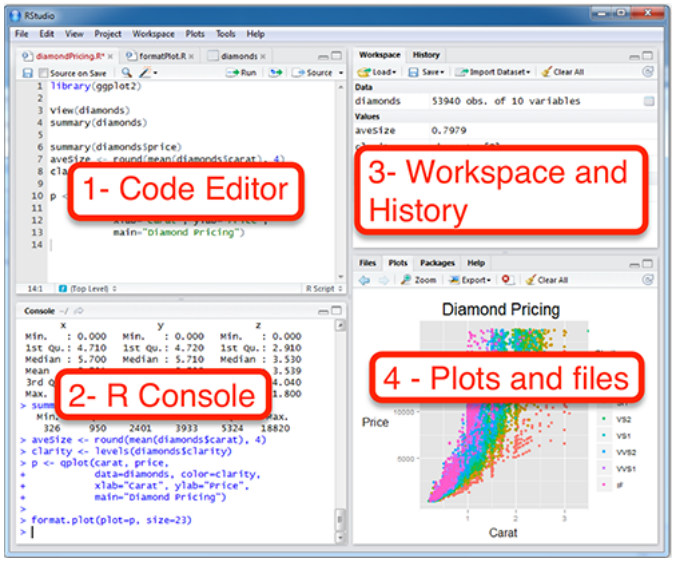
В интернете легко найти документацию и туториалы по самым разным вопросам в R, так что главный залог успеха — грамматно пользоваться поисковиком, и лучше на английском языке.

- книга (Wickham and Grolemund, 2016) является достаточно сильной альтернативой всему курсу
- stackoverflow — сервис, где достаточно быстро отвечают на любые вопросы (не обязательно по R)
- RStudio community — быстро отвечают на вопросы, связанные с R
- русский stackoverflow
- R-bloggers — сайт, где собираются новинки, связанные с R
- чат, где можно спрашивать про R на русском (но почитайте правила чата, перед тем как спрашивать)

- чат по визуализации данных, чат датажурналистов
- канал про визуализацию, дата-блог “Новой газеты”, ...

2.4 Rstudio

Когда вы откроете RStudio первый раз, вы увидите три панели: консоль, окружение и историю, а также панель для всего остального. Если ткнуть в консоли на значок уменьшения, то можно открыть дополнительную панель, где можно писать скрипт.



Существуют разные типы пользователей: одни любят работать в консоли, другие предпочитают скрипты. Консоль позволяет иметь интерактивный режим команда-ответ, а скрипт является по сути текстовым документом, фрагменты которого можно для отладки запускать в консоли.

Глава 3

Трансформация данных

Глава 4

Вмзуализация данных

Глава 5

Условия и работа со списками

Глава 6

Представление данных: rmarkdown, github, shiny

Глава 7

Работа со строками

Глава 8

Работа с текстами: tidytext, udpipe

Глава 9

Сбор данных из интернета: rvest

Глава 10

Нестандартные данные: время, OCR, карты

Литература

Baesens, B., Van Vlasselaer, V., and Verbeke, W. (2015). *Fraud analytics using descriptive, predictive, and social network techniques: a guide to data science for fraud detection*. John Wiley & Sons.

Brooks, H. and Cooper, C. L. (2013). *Science for public policy*. Elsevier.

Hansjörg, N. (2019). *Data Science for Psychologists*. self published.

Provost, F. and Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media, Inc.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Thomas, N. and Pallett, L. (2019). *Data Science for Immunologists*. CreateSpace Independent Publishing Platform.

Wickham, H. and Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data*. O'Reilly Media, Inc.