



Казанский
Государственный
Медицинский
Университет

Основные компоненты статистической среды R.

Лакман Ирина Александровна

Казань, 2025



План лекции



Что такое R, RStudio, RTools?



Как создать скрипты и проекты?



Представление данных в R



Стандарт аккуратных данных в R



Формирование отчетов в RMarkdown



Особенности работы в Rmarkdown

Что такое R, RStudio, RTools?

R – язык программирования для статистической обработки данных и работы с графикой

RStudio – кроссплатформенная оболочка языка R

RTools – требуется для установки пакетов в Windows

Установка

Необходимо скачать:

- **R**
 - Для win cran.r-project.org/bin/windows/base/
 - Для macOS cran.r-project.org/bin/macosx/
- **RStudio** rstudio.com/products/rstudio/download/#download
- **RTools** (только для win) cran.r-project.org/bin/windows/Rtools/



Порядок установки: **R → RStudio → RTools**

Для пользователей Windows: **важно**, чтобы путь установки **НЕ содержал кириллицу**.

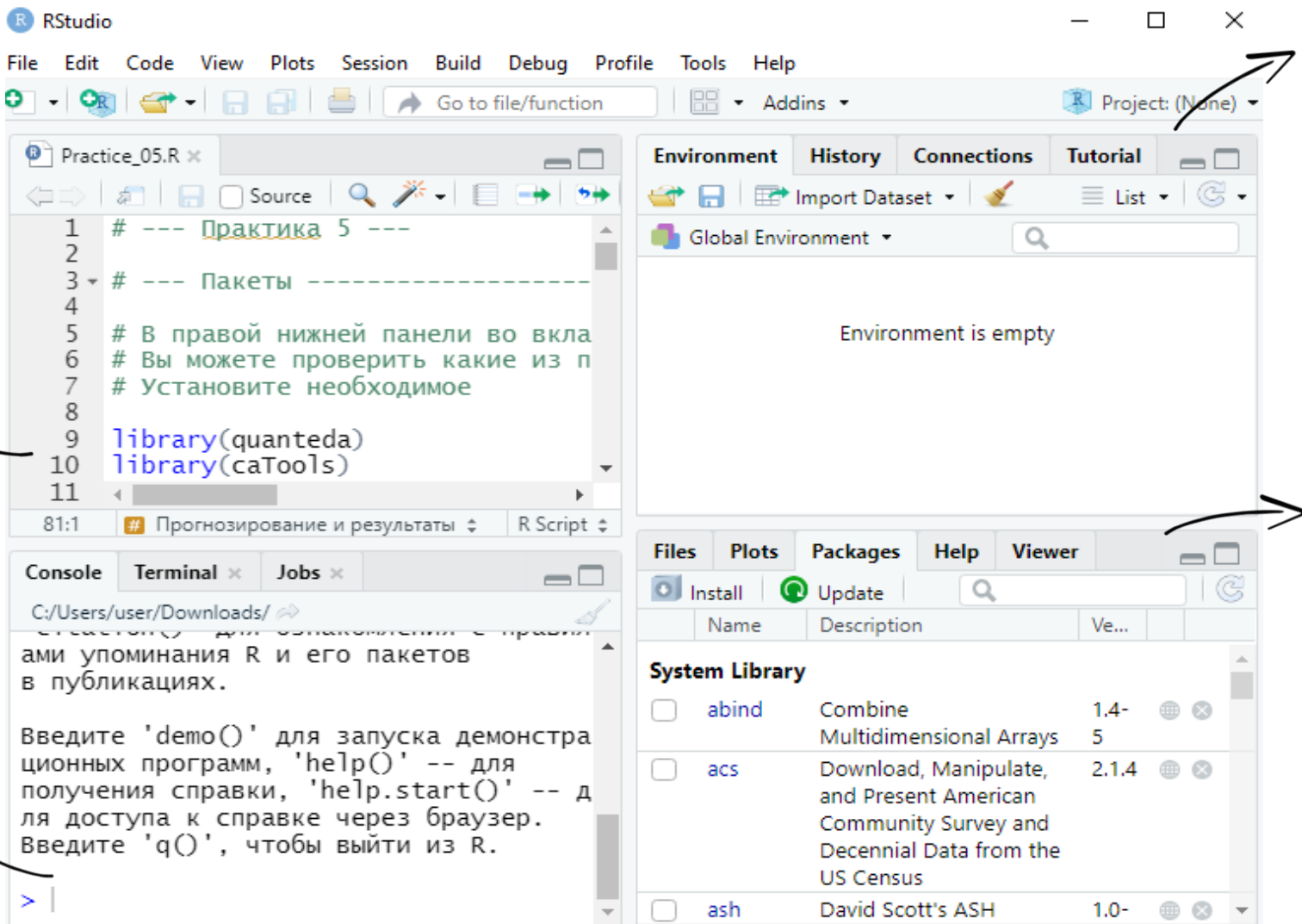
Не устанавливайте программы в стандартную папку **«Program Files»**, поскольку в дальнейшем возникнут проблемы при установке дополнительных библиотек.

Лучше всего установить R в «C:/R», RStudio в «C:/RStudio», а Rtools в «C:/Rtools»

Вид и значение консолей в интерфейсе RStudio

Окно со скриптами, где можно писать основной код

Консоль, где отображаются результаты последних выполненных команд



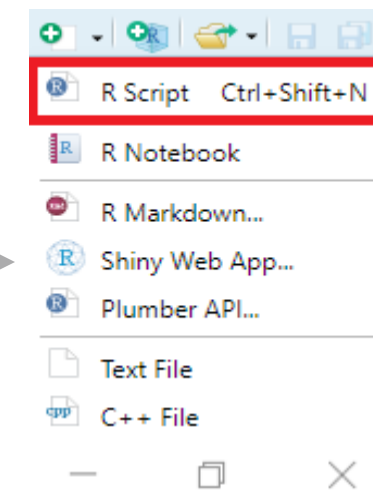
Environment - показаны все подключённые данные в этой сессии
History - история выполненных в консоли команд
Connections - список подключений к базам данных

Plots – отображение построенных графиков
Packages - вкладка с библиотеками (пакетами). Из этой вкладки можно установить новые пакеты или обновить существующие
Help - вкладка со справкой. Следует искать через поисковую строку, как пользоваться нужными функциями или пакетами

Как создать скрипты и проекты?

Создание R-скриптов

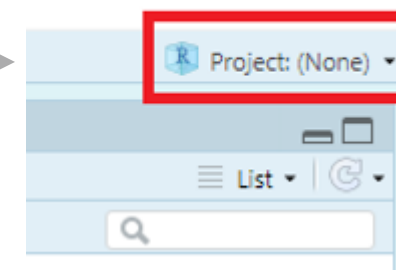
Чтобы создать скрипт, следует нажать на зеленую кнопку с плюсом в левом верхнем углу и затем на **R Script**. Альтернативно можно использовать сочетание клавиш **Ctrl + Shift + N**



Создание проектов

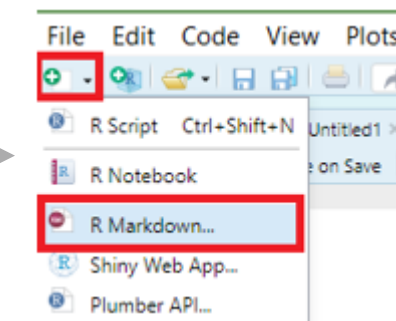
Проекты устанавливают рабочую папку, где можно хранить свои скрипты и данные. Это упрощает работу, поскольку не приходится каждый раз прописывать весь путь до данных.

Для того, чтобы создать проект следует нажать на кнопку **Project: (None)** в правом верхнем углу



Создание R Markdown скриптов

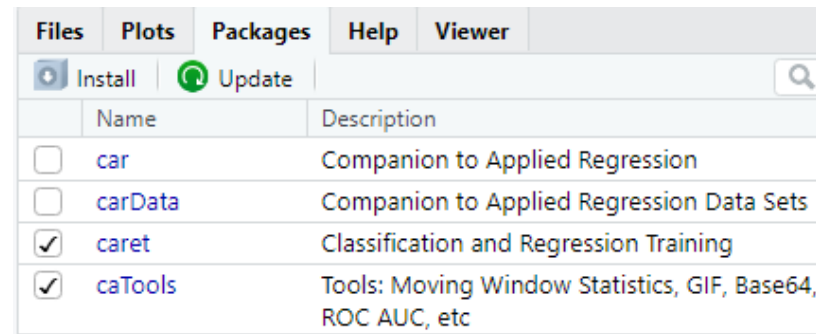
В RStudio можно создавать отчёты с помощью языка разметки **Markdown**. Чтобы создать RMarkdown скрипт следует выполнить следующие действия: **Files → New Files → R Markdown**. Другой способ создания R Markdown скрипта: нажать на зеленую кнопку с плюсом и затем выбрать R Markdown.



Что такое библиотеки в R, и чем они отличаются от пакетов?

Многие полезные функции R поставляются в пакетах, бесплатных библиотеках кода, написанных активным сообществом пользователей R (учеными, исследователями). В Rstudio пакеты можно установить разными способами.

1. Написать функцию в консоли **install.packages(*название пакета*)**
2. Во вкладке **Packages** выбрать кнопку **Install**, в появившемся диалоговом окне указать название библиотеки и нажать кнопку **Install**



Пакеты устанавливаются один раз, в дальнейшем их необходимо только подгружать. Для подключения установленных пакетов используется команда **library()**

Для проверки какие пакеты установлены и какие из них подключены следует зайти во вкладку Packages. Установленные есть в списке, а подключённые отмечены галочкой.

Представление данных в R

R позволяет загружать данных из практически всех источников данных. Основные форматы данных: .csv, .xlsx

У формата CSV (comma-separated values - значения, разделенные запятыми) существует свой стандарт, но многие csv-файлы собраны по своим собственным правилам

Перед тем как загрузить данные в R необходимо открыть файл в текстовом редакторе и проверьте как собран csv-файл.

На что обратить внимание:

1. Наличие заголовка в файле с названиями переменных (header)
2. Разделитель целой и дробной частей: точка или запятая (в России чаще всего запятая)
3. Разделитель между переменными (столбцами): пробел, табуляция, запятая, точка с запятой (в России чаще всего точка с запятой)
4. Кодировка данных: UTF-8, CP1251, KOI8-R (старайтесь использовать UTF-8)

Стандарт csv-файлов: с заголовком, разделитель целой и дробной частей: точка, разделитель между столбцами: запятая, кодировка UTF-8

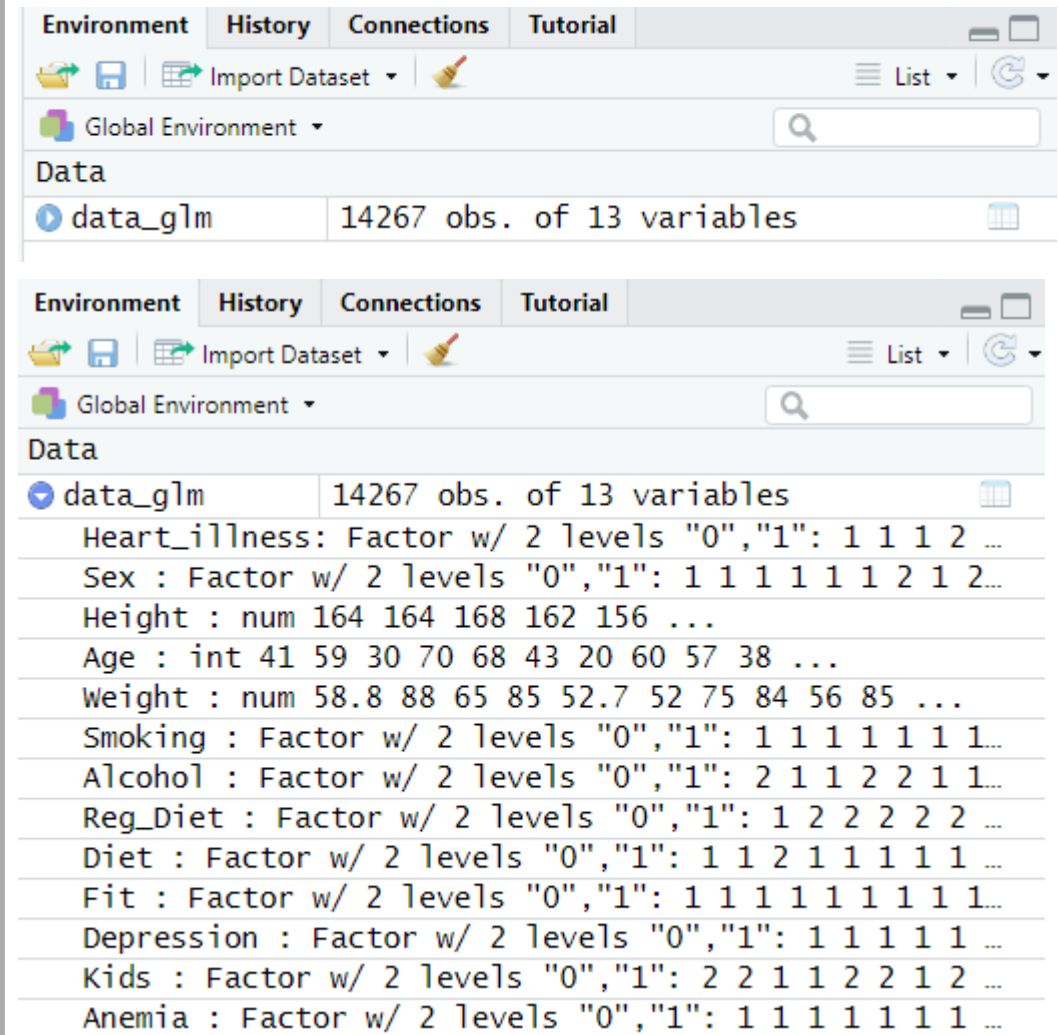
Просмотр данных

Наборы данных можно просмотреть во вкладке Environment в правой верхней консоли:

Для просмотра информации о типах данных следует раскрыть данные:

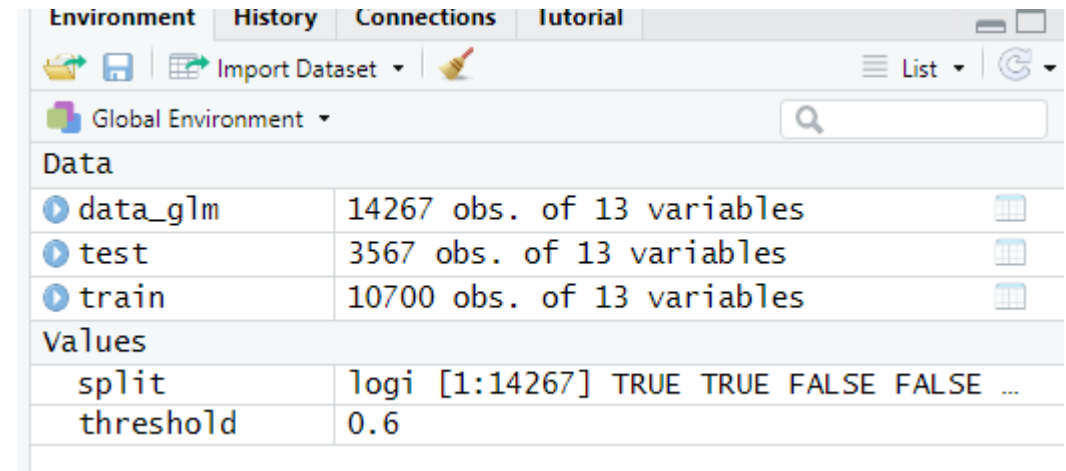
Для просмотра самой информации можно:

- 1) 2 раза щёлкнуть по имени набора данных
- 2) Набрать команду View(данные)



Отражение типов данных в R

- Ordered – упорядоченные
- Int – целочисленные
- Num – числовые
- Factor – номинальные (факторные)
- Chr – текстовые
- Logi – логическая (“True”, “False”)



The screenshot shows the RStudio Environment pane. At the top, there are tabs for 'Environment', 'History', 'Connections', and 'Tutorial'. Below the tabs, there is a toolbar with icons for file operations and a search bar. The main area displays the 'Global Environment' with a list of objects. The objects are categorized into 'Data' and 'Values'.

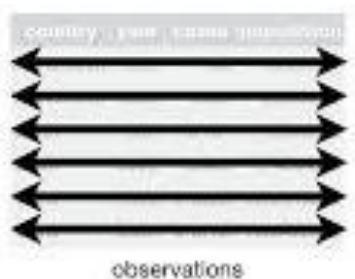
Data	
data_glm	14267 obs. of 13 variables
test	3567 obs. of 13 variables
train	10700 obs. of 13 variables
Values	
split	logi [1:14267] TRUE TRUE FALSE FALSE ...
threshold	0.6

Стандарт аккуратных данных в R



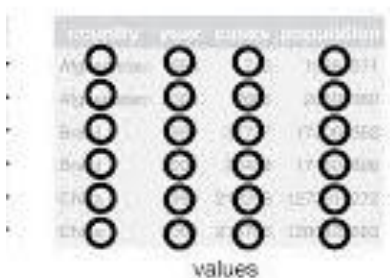
country	year	cases	population
Australia	2015	15	230071
Australia	2016	15	230071
Brunei	2015	2137	436800
Brunei	2016	2137	436800
China	2015	2137	13713272
China	2016	2137	13713272

variables



country	year	cases	population
Australia	2015	15	230071
Australia	2016	15	230071
Brunei	2015	2137	436800
Brunei	2016	2137	436800
China	2015	2137	13713272
China	2016	2137	13713272

observations



country	year	cases	population
Australia	2015	15	230071
Australia	2016	15	230071
Brunei	2015	2137	436800
Brunei	2016	2137	436800
China	2015	2137	13713272
China	2016	2137	13713272

values

Преимущество R - обработка данных

tidyverse – флагманская библиотека языка R, разработанная для комплексного анализа данных и Data Science.

Входящие в экосистему tidyverse пакеты позволяют производить комплексную обработку данных – начиная от загрузки данных из различных источников и заканчивая построением визуализаций.

Tidy Data – основная концепция пакетов семейства tidyverse.

Есть определенные законы и определения по поводу того, как должны выглядеть аккуратные данные.

- Каждая переменная находится в столбце.
- Каждое наблюдение — это строка.
- Каждое значение является ячейкой.





RMarkdown



Требование к отчетам по анализ данных:

- Поддержка визуализации
- Создание шаблонов отчета
- воспроизводимость,
- Возможность динамических изменений
- Возможность открытия в виде ссылки

R Markdown объединяет анализ данных, визуализацию данных и другой код R с описательным текстом для создания документа, который можно экспортировать во многие форматы, включая Word, PDF и HTML.

Текст в RMarkdown

R Markdown позволяет легко и просто компилировать отчёты в формате: **.docx**, **.pdf** и **.html**

Создастся новый скрипт (файл формата **.Rmd**) с уже готовым кодом.

Небольшая демонстрация возможностей RMarkdown. Здесь можно видеть, что RMarkdown позволяет сочетать обычный текст с кодом R. Эти куски кода называются **chunk**. Наверху прописана преамбула (выделена тремя тире сверху и снизу) документа: название документа, автор, дата и выходной формат

```
1 |  
2 |---  
3 |title: "Untitled"  
4 |author: ""  
5 |output: html_document  
6 |---  
7 |```{r setup, include=FALSE}  
8 |knitr::opts_chunk$set(echo = TRUE)  
9 |```  
10 |  
11 |## R Markdown  
12 |  
13 |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  
14 |see <http://rmarkdown.rstudio.com>.  
15 |  
16 |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  
17 |document. You can embed an R code chunk like this:  
18 |  
19 |```{r cars}  
20 |summary(cars)  
21 |```  
22 |  
23 |## Including Plots  
24 |  
25 |You can also embed plots, for example:  
26 |  
27 |```{r pressure, echo=FALSE}  
28 |plot(pressure)  
29 |```  
30 |  
31 |Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.
```



R Markdown: компиляция

Чтобы сформировать отчет, следует нажать кнопку **Knit** в RMarkdown-скрипте.

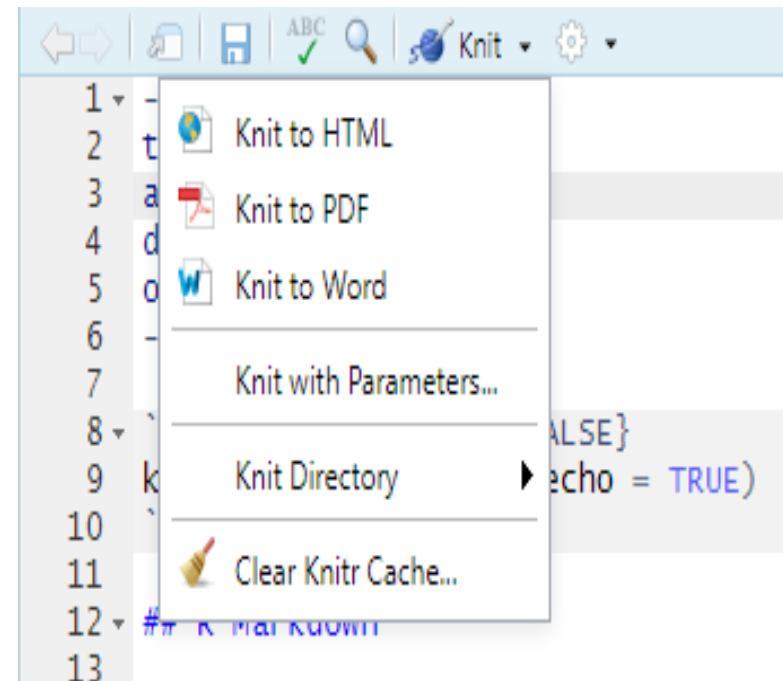
Для выбора определённого формата отчёта (**HTML**, **PDF (требуется LaTeX)**, **Word**), следует нажать на «стрелку вниз» рядом с надписью Knit.

Здесь можно выбрать 'Knit to HTML', 'Knit to PDF' или 'Knit to Word'. В любой момент можно сменить тип выходного документа.

После компиляции откроется отчет в выбранном формате. При этом сам файл отчета будет находиться в одной папке со скриптом RMarkdown.

При этом:

- обычный текст из скрипта RMarkdown «перебирается» в отчет в практически неизменном виде
- куски кода R выполняются при компиляции, а результат отображается в отчете (таблица с данными и рисунок)





Метаданные YAML

```
---  
title: "Анализ больших бизнес-данных"  
author: "Лакман И.А., lackmania@mail.ru"  
output:  
  html_document:  
    df_print: paged  
---
```

Раздел YAML — это самое начало документа R Markdown.

Название YAML происходит от аббревиатуры:

YAML ain't markup language.

Три тире обозначают начало и конец, а текст внутри него содержит метаданные о документе R Markdown.

YAML предоставляет:

- заголовок
- автора
- дату
- формат вывода

Все элементы YAML даны в **key: value** синтаксисе, где каждый ключ — это метка для фрагмента метаданных (например, заголовка), за которым следует значение.



Фрагменты кода

```
205
206 ```{r echo=TRUE, message=FALSE, warning=FALSE}
207
208 summary(cars)
209
210 ```
211
```

Блок установочного
кода

Документы R Markdown имеют структуру, отличную от файлов скриптов R.

Файлы скриптов R обрабатывают весь контент как код, за исключением комментариев, отмеченных #.

В R Markdown все после YAML обрабатывается как текст, если нет создания фрагменты кода .

фрагменты кода начинаются с трех обратных кавычек (`` ` ``), за которыми следует строчная буква `r`, указывающая, что код создан в R

три обратных кавычки обозначают конец фрагмента кода:

Фрагменты кода, называются **Chunk** и выделены в документе markdown светло-серым фоном.

R Markdown обрабатывает все в фрагменте кода как код R.

Блок кода в верхней части каждого документа R Markdown, называется **установочным**.



Настройки установок кода

Список и описание основных настроек установочного блока кода:

- **`echo`** - отображение (TRUE)/скрытие (FALSE) кода.

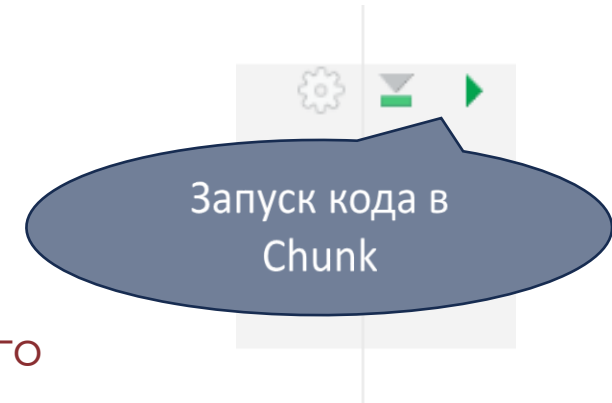
Можно включить, когда требуется показать и код, и результаты выполнения.

- **`eval`** - вычисление (TRUE) кода.

Для случаев, когда требуется показать код, но нет необходимости его считать

- **`warning`** - отображение (TRUE)/скрытие (FALSE) предупреждений от пакетов и команд
- **`message`** - отображение (TRUE)/скрытие (FALSE) сообщений от пакетов и команд.
- **`fig.height`**, **`fig.width`** и др. - различные настройки отображения графиков.

По умолчанию **`echo`**, **`eval`**, **`warning`** и **`message`** - включены, то есть у них стоит значение **TRUE**.





Текст в формате Markdown



Markdown — это способ стилизации текста.

Поддерживает:

Списки (нумерованные и нет)

Полужирный шрифт и курсив

Подчеркивание и зачеркивание текста

Различного формата заголовки

Отступы и выступления

-Позволяет вставлять в отчет:

Цитаты, Блоки кода, Ссылки, Разделительные линии

Позволяет вставлять рисунки и формулы

Для формул используется язык LaTeX

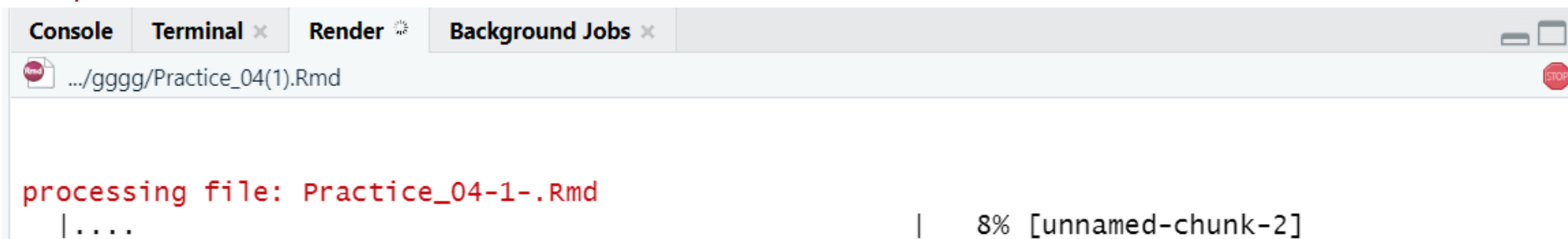
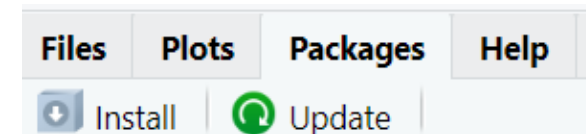
Есть аналог **Markdown - bookdown**, используемый для верстки статей

Есть аналог для ведения блогов - **Rblogdown**



Особенности работы в Rmarkdown

1. Файл со скриптом .rmd должен размещаться в одном проекте с данными, если в нем есть обращение к этим данным.
2. Не рекомендуется использовать команду `install, package(" ")` в скрипте .rmd, так как не будет скомпилирован отчет. Лучше установить пакет заранее используя вкладку Packages, кнопка Install
3. Для того чтобы вставился в отчет рисунок определенного размера, надо дополнительно делать настройки этого размера с помощью команд ``fig.height``, ``fig.width``
4. В начале скрипта также требуется (желательно в отдельном chunk) сделать обращение к предустановленным библиотекам (команда `library()`)
5. Для компиляции отчета в формат .pdf требуется заранее установить LaTeX
6. При компиляции отчета во вкладке Render отображается процесс его формирования в %





**Спасибо за
внимание!**