Язык R и анализ данных

Занятие 4

Елена Ставровская

29 сентября 2017

Dplyr

- Позволяет легко извлекать и преобразовывать данные
- Имеет простой синтаксис
- Позволяет выстраивать конвейер и не плодить лишние промежуточные переменные

Устанавливаем пакет

```
>install.packages(“dplyr”)
```

Основные функции dplyr

- select() и rename() доставать столбцы по именам.
- filter() доставать строки по условию.
- mutate() и transmute() добавлять новые столбцы, которые являются результатом функции от имеющихся столбцов.
- summarise() получить одно значение по набору значений.
- sample_n() и sample_frac() вытащить случайные подвыборки данных.
- arrange() упорядочить.
- inner_join(), left_join(), right_join() склеить два набора данных по столбцу (столбцам)

Анализируем данные о качестве воздуха. Select

>head(airquality)

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8.0	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6

>select(airquality, Wind, Temp)

Wind Temp 1 7.4 67 2 8.0 72 3 12.6 74 4 11.5 62 5 14.3 56 6 14.9 66

Filter

- Выберем все строки, где температура >70 (не пугайтесь, это в фаренгейтах)

```
>f_airquality<-filter(airquality, Temp > 70)
>head(f_airquality)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	36	118	8.0	72	5	2
2	12	149	12.6	74	5	3
3	7	NA	6.9	74	5	11
4	11	320	16.6	73	5	22
5	45	252	14.9	81	5	29
6	115	223	5.7	79	5	30

Filter

- Можно комбинировать. Выведем строки, где температура >80, а месяц позже чем май

```
> f_airquality<- filter(airquality, Temp > 80 & Month > 5)
>head(f_airquality)
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	NA	186	9.2	84	6	4
2	NA	220	8.6	85	6	5
3	29	127	9.7	82	6	7
4	NA	273	6.9	87	6	8
5	71	291	13.8	90	6	9
6	39	323	11.5	87	6	10

Задание 1

- Выведите данные airquality за август, когда температура была больше 75, но меньше 90

Mutate

- Позволяет добавить новый столбец к данным.

Например, добавим столбец с температурой в градусах Цельсия

```
>m_airquality<-mutate(airquality, TempInC = (Temp - 32) * 5 / 9)
> head(m_airquality)
```

	Ozone	Solar.R	Wind	Temp	Month	Day	TempInC
1	41	190	7.4	67	5	1	19.44444
2	36	118	8.0	72	5	2	22.22222
3	12	149	12.6	74	5	3	23.33333
4	18	313	11.5	62	5	4	16.66667
5	NA	NA	14.3	56	5	5	13.33333
6	28	NA	14.9	66	5	6	18.88889

transmute

- Делает тоже самое, что mutate, только возвращает один новый столбец

```
>c_temp<-transmute(airquality, TempInC = (Temp - 32) * 5 / 9)
>head(c_temp)
   TempInC
1 19.44444
2 22.22222
3 23.33333
4 16.66667
5 13.33333
6 18.88889
```

Задание 2

- Добавьте столбец Wind_kmh – скорость в км/ч (в столбце Wind данные в милях/ч). Считайте, что миля=1.6 км

Summarise

- Позволяет применять функцию к данным (очень мощная штука!)

Вычислим среднюю температуру

```
>summarise(airquality, mean(Temp, na.rm = TRUE))
mean(Temp, na.rm = TRUE)
1 77.88235
```

Group By

- Используется для группировки данных по одной и более переменных.

Выведем среднюю температуру по месяцам

```
>summarise(group_by(airquality, Month), mean(Temp,
    na.rm = TRUE))
```

```
#   A tibble: 5 x 2
    Month `mean(Temp, na.rm = TRUE)`
    <int>   <dbl>
1   5        65.54839
2   6        79.10000
3   7        83.90323
4   8        83.96774
5   9        76.90000
```

Задание 3

- Выведите среднюю скорость ветра по месяцам

Sample

- Позволяет выбирать случайные строки из датафрейма

>sample_n(airquality, size = 5)

	Ozone	Solar.R	Wind	Temp	Month	Day
151	14	191	14.3	75	9	28
73	10	264	14.3	73	7	12
105	28	273	11.5	82	8	13
64	32	236	9.2	81	7	3
117	168	238	3.4	81	8	25

>sample_frac(airquality, size = 0.05)

	Ozone	Solar.R	Wind	Temp	Month	Day
129	32	92	15.5	84	9	6
28	23	13	12.0	67	5	28
80	79	187	5.1	87	7	19
106	65	157	9.7	80	8	14
137	9	24	10.9	71	9	14
39	NA	273	6.9	87	6	8
117	168	238	3.4	81	8	25
90	50	275	7.4	86	7	29

Count

- Считает строки, группируя их по заданным столбцам.

Посчитаем количество строк для каждого месяца

```
> count(airquality, Month)
# A tibble: 5 x 2
    Month n
    <int> <int>
1   5     31
2   6     30
3   7     31
4   8     31
5   9     30
```

Arrange

- Позволяет упорядочить датафрейм по столбцам.

Упорядочим наш датафрейм по месяцам (в порядке убывания), затем по дням

```
>new_airquality<-arrange(airquality, desc(Month), Day)
> head(new_airquality)
```

```
  Ozone Solar.R Wind Temp Month Day
1    96     167  6.9   91     9   1
2    78     197  5.1   92     9   2
3    73     183  2.8   93     9   3
4    91     189  4.6   93     9   4
5    47      95  7.4   87     9   5
6    32      92 15.5   84     9   6
```

Конвейер

- Можно передавать данные из одной команды в другую.

```
> airquality %>% filter(Month != 5) %>%
  group_by(Month) %>% summarise(mean(Temp,
  na.rm = TRUE))
```

- Тоже что и

```
> filteredData <- filter(airquality, Month != 5)
> groupedData <- group_by(filteredData, Month)
> summarise(groupedData, mean(Temp, na.rm =
  TRUE))
```

Задание 4

- Для встроенного датасета flights посчитать среднюю задержку рейсов за первые 7 дней каждого месяца

join

Студент	Факультет	Курс	Пол
Иванов			
Петров			

Студент	КР1	КР2	КР3
Иванов			
Сидоров			

Студент	Факультет	Курс	Пол	КР1	КР2	КР3
Иванов						

join

- Пример: 2 таблицы
  - id -> оценки,
  - id -> метаданные студента
- Объединить таблицы в одну по идентификатору студента

```
> head(grades)
    id write math science socst
1   70    52   41      47    57
2  121    59   53      63    61
3   86    33   54      58    31
4  141    44   47      53    56
5  172    52   57      53    61
6  113    52   51      63    61
```

```
> head(metadata)
   id female race schtyp prog
1   1      1    1      1    3
2   2      1    1      1    3
3   3      0    1      1    2
4   4      1    1      1    2
5   5      0    1      1    2
6   6      1    1      1    2
```

Inner_join

```
>stud<-inner_join(metadata, grades, by=c("id"))
```

```
> head(stud_data)
```

	X.x	id	female	race	schtyp	prog	X.y	write	math	science	socst
1	1	1	1	1	1	3	99	44	40	39	41
2	2	2	1	1	1	3	139	41	33	42	41
3	3	3	0	1	1	2	84	65	48	63	56
4	4	4	1	1	1	2	112	50	41	39	51
5	5	5	0	1	1	2	76	40	43	45	31
6	6	6	1	1	1	2	149	41	46	40	41

left_join оставляет все строки из первого датафрейма

right_join оставляет все строки из второго датафрейма

Для строк без соответствия добавляет **NA**

left_join

```
>grades1<-grades[1:100,]
> stud1<-left_join(metadata, grades1, by=c("id"))
>head(stud1)
X.x id female race schtyp prog X.y write math science socst
1 1 1 1 1 1 3 99 44 40 39 41
2 2 2 1 1 1 3 NA NA NA NA NA
3 3 3 0 1 1 2 84 65 48 63 56
4 4 4 1 1 1 2 NA NA NA NA NA
5 5 5 0 1 1 2 76 40 43 45 31
6 6 6 1 1 1 2 NA NA NA NA NA
```

Задание 5

- Загрузите данные об оценках студентов grades.csv и данные с информацией о студентах metadata.csv
- Посчитайте среднюю оценку за письмо (write) для мальчиков и девочек
- Отличаются ли оценки по письму у мальчиков и девочек? (вспоминаем статистику)

R markdown

Как писать код и одновременно делать отчет

Зачем это надо?

- Чтобы не забыть, что мы делаем
- Красиво оформить результаты
- Протокол исследования: результаты для разных наборов параметров (попробовали так, попробовали сяк…)

Преимущества:

- Одновременно пишем код и делаем отчет
- Все ходы записаны – понятно, что делали
- Ничего не путается – картинки результатов строго соответствуют коду
- Можно форматировать текст, вставлять ссылки, внешние картинки, формулы и многое другое!

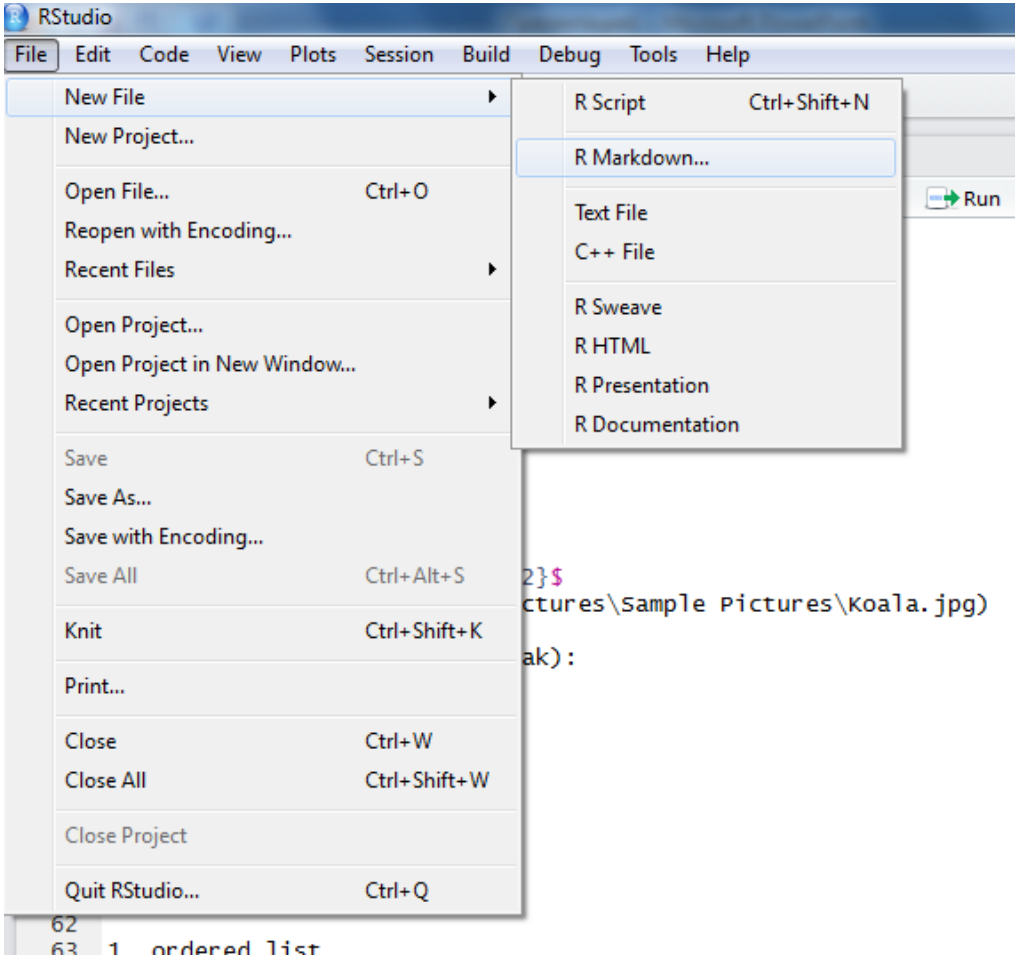
Ставим необходимый пакет

```
install.packages("rmarkdown")
```

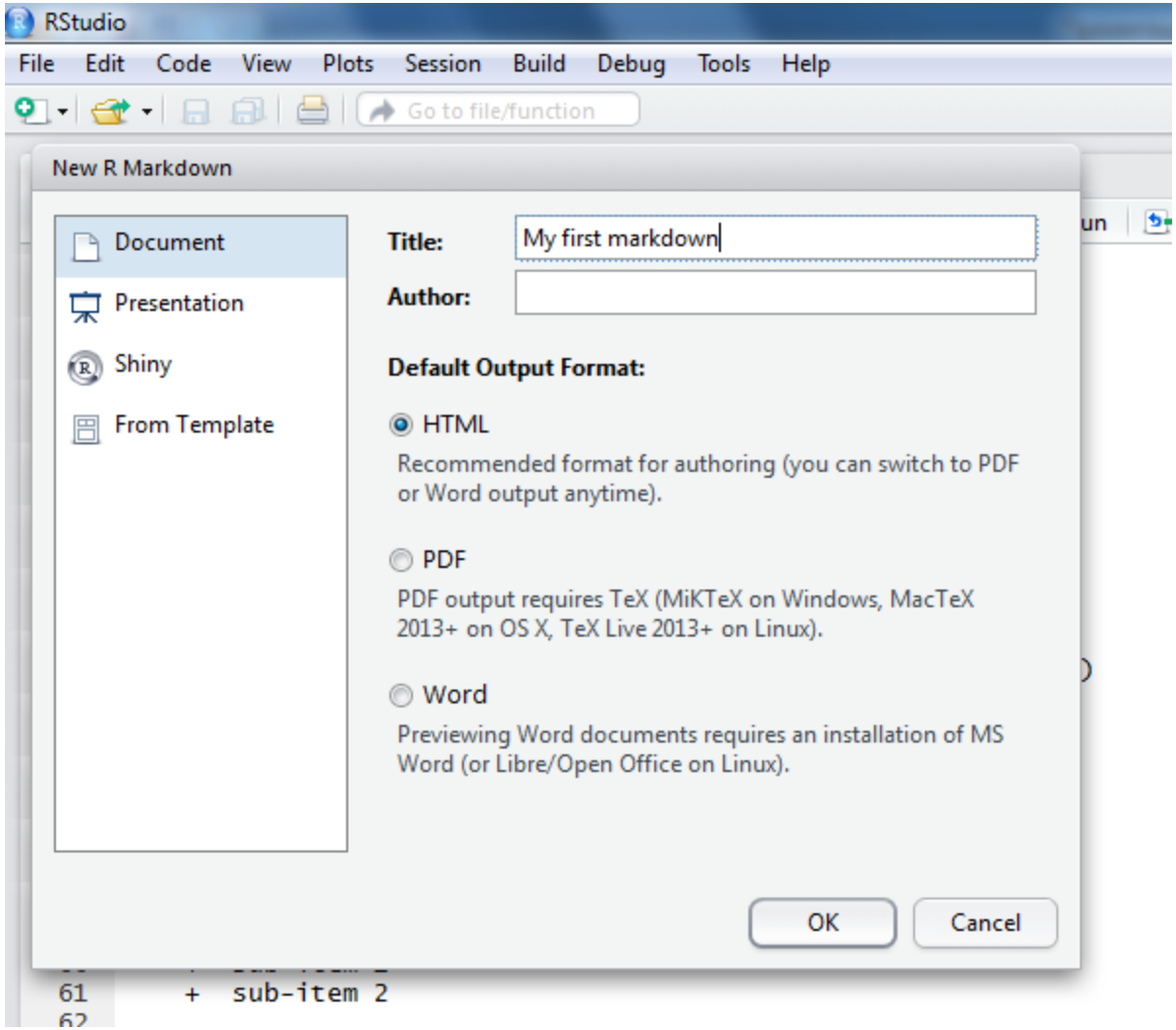
Порядок работы

- Создаем документ doc_name.Rmd
- Пишем протокол нашего исследования
- Наполняем R кодом (собственно, наше исследование)
- Форматируем текст (чтобы все было красиво)
- Генерим с помощью Rstudio наш отчет (в виде html, pdf, документа word или презентации)

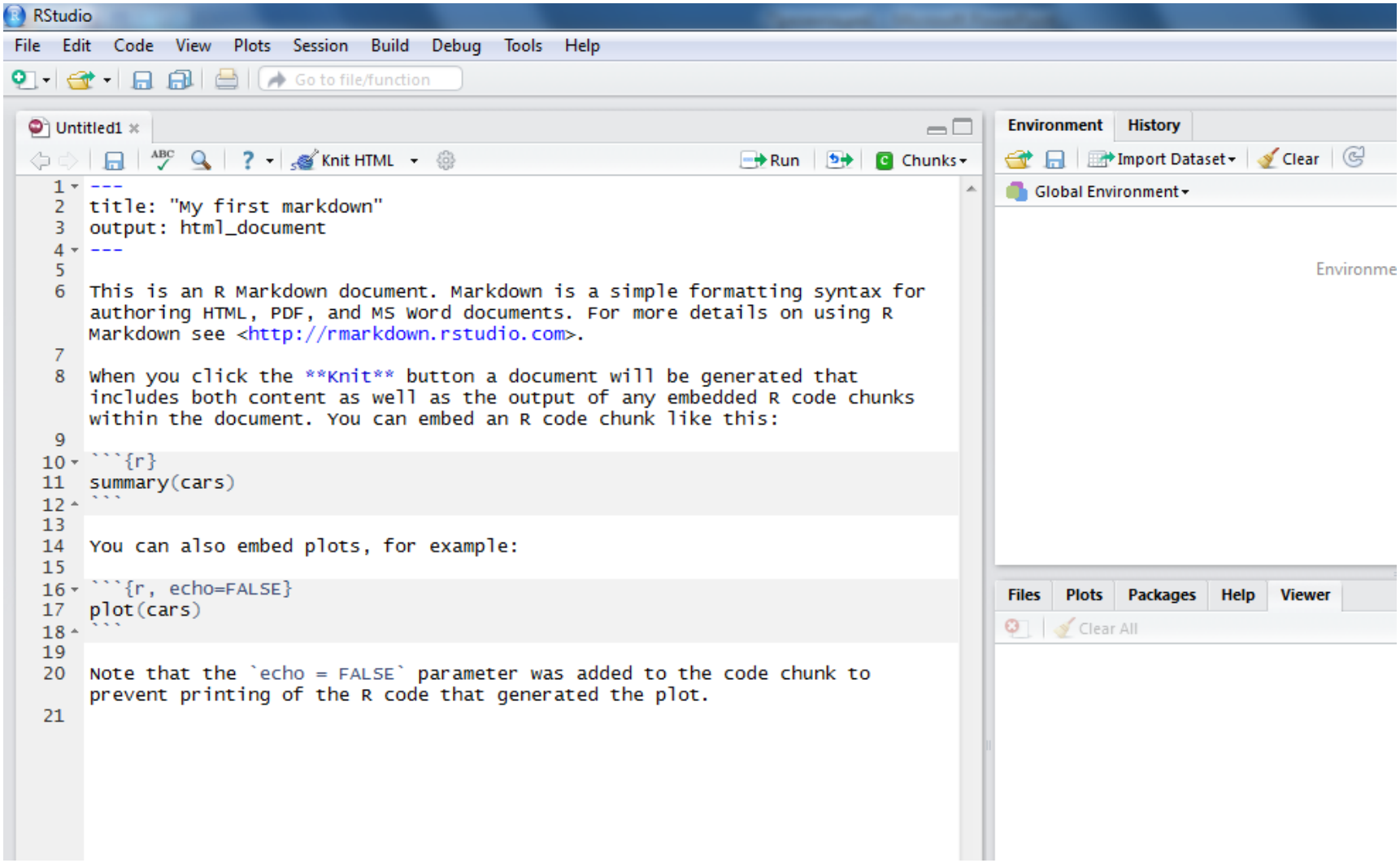
Как создать отчет в RStudio



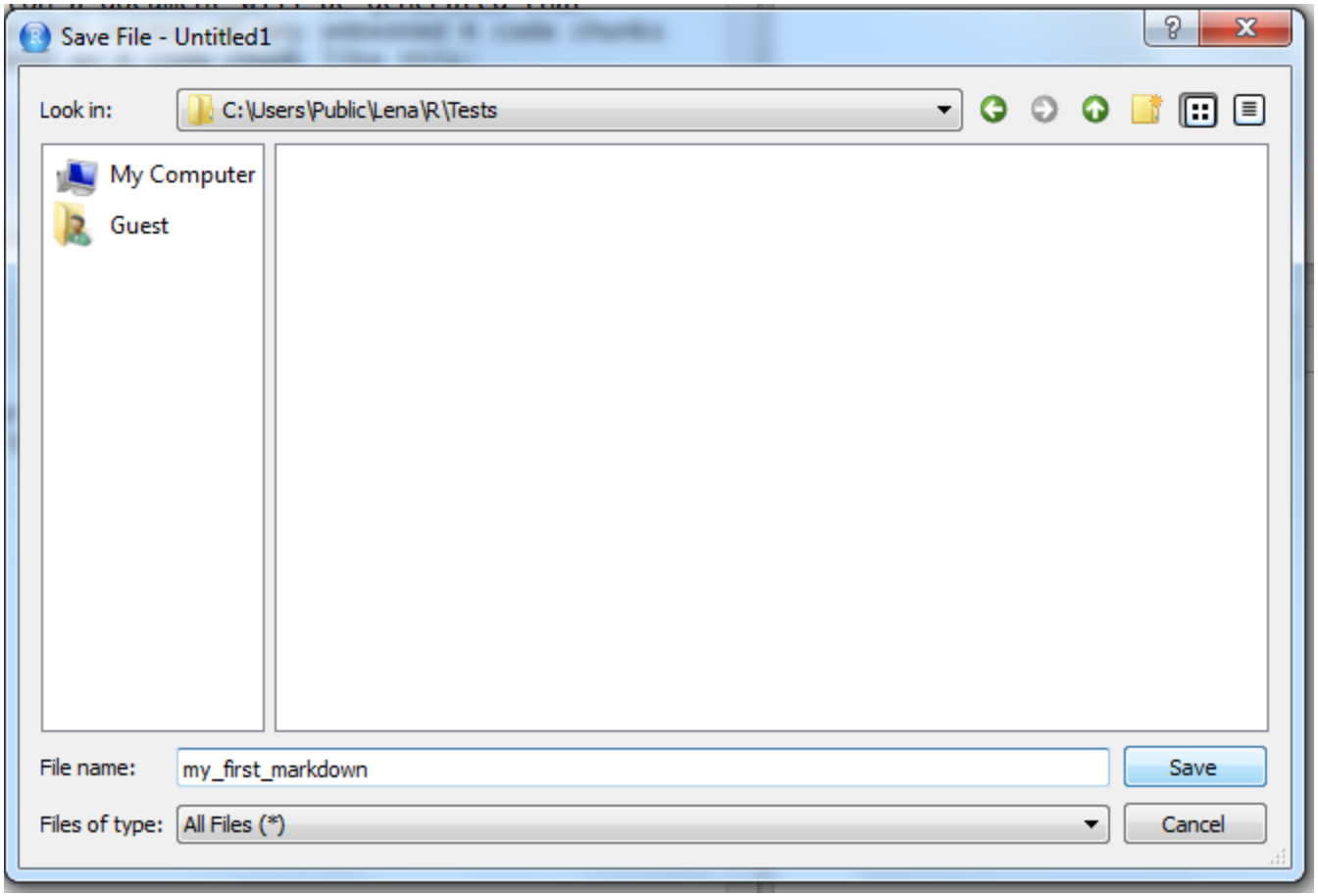
Как создать отчет в RStudio



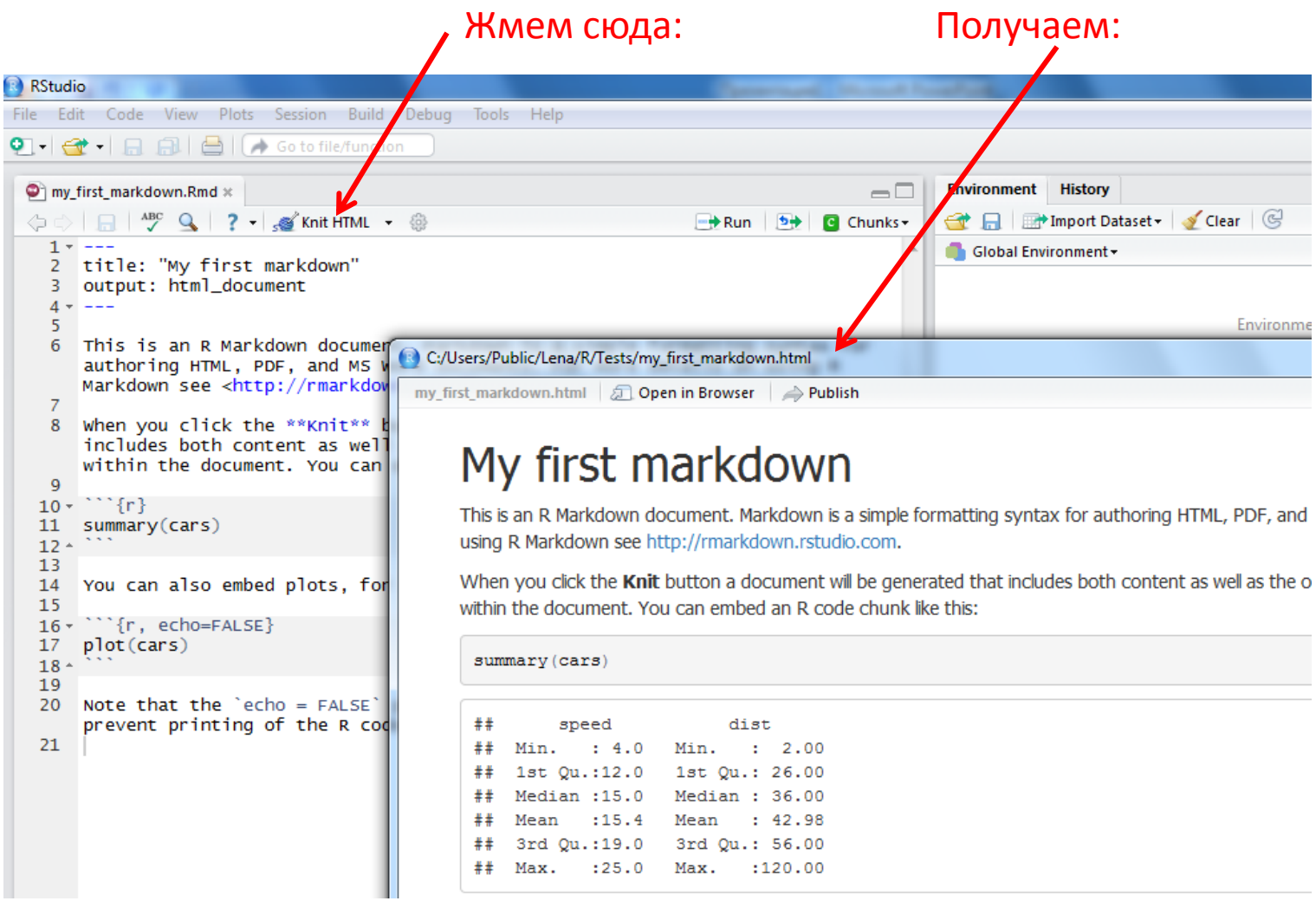
Как создать отчет в RStudio



Как создать отчет в RStudio



Запускаем базовый пример



Добавление блока R кода

- В отчете отображается код и результат

```
```{r}
dim(iris)
```
```

```
dim(iris)

## [1] 150   5
```

- Только код

```
```{r eval=FALSE}
dim(iris)
```
```

```
dim(iris)
```

- Только результат

```
```{r echo=FALSE}
dim(iris)
```
```

```
## [1] 150   5
```

Форматирование текста

- Вставка кусочка кода в текст

We have 2+2=`r 2+2`

We have 2+2=4

- Италик

italics and _italics_

italics and *italics*

- Жирный текст

bold and __bold__

bold and **bold**

- Верхний индекс

superscript^2^

superscript^2

- Зачеркнутый текст

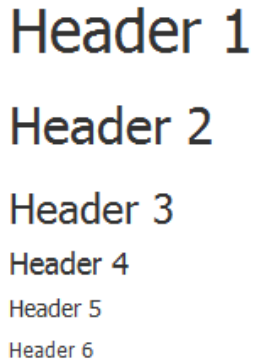
~~strikethrough~~

~~strikethrough~~

Форматирование текста

- Заголовки разного уровня

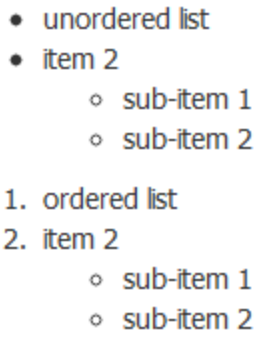
```
# Header 1
## Header 2
### Header 3
#### Header 4
##### Header 5
###### Header 6
```



- Списки

```
* unordered list
* item 2
   + sub-item 1
   + sub-item 2

1. ordered list
2. item 2
   + sub-item 1
   + sub-item 2
```



Форматирование текста

- Формула

$A = \pi*r^{2}$

$$A = \pi * r^2$$

- Внешняя картинка

- Ссылка

[link](www.rstudio.com)

<http://rmarkdown.rstudio.com>

link

http://rmarkdown.rstudio.com

Форматирование текста

- Разрыв страницы (новый слайд)

***

- Цитата (выделенный блок текста)

> block quote

block quote

Форматирование текста

- Таблица

Table Header | Second Header

---------- | ----------

Table Cell | Cell 2

Cell 3 | Cell 4

Table Header	Second Header
Table Cell	Cell 2
Cell 3	Cell 4

Задание 6

- Оформить анализ задания 5 в виде отчета в pdf формате. Отчет должен включать графический анализ выборок