



Биоәртүрлілікті зерттеудегі цифрлық технологиялар
Digital technologies in biodiversity research
Цифровые технологии в исследовании биоразнообразия



ЛЕКЦИЯ 9

Зачем биологу
программирование?
Может ли R заменить
другие инструменты.



Слайды CC BY:

Артём Созонтов, ИЭРИИЖ УрО РАН
Наталья Иванова,
Максим Шашков

План лекции

Инструменты для статистического анализа данных

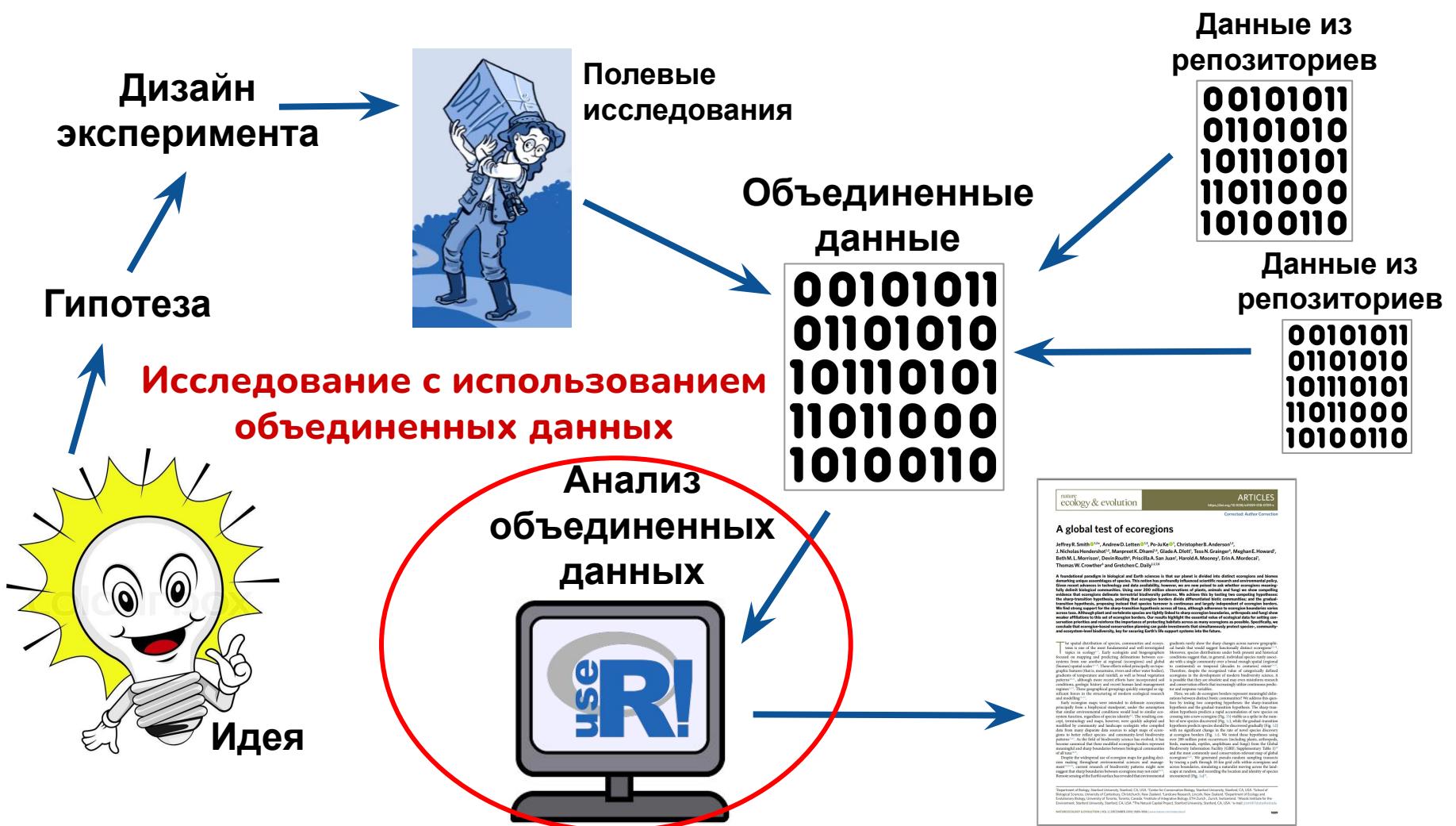
Язык статистического программирования R:

- Возможности
- Область применения
- Основы использования



A screenshot of a Mac OS X window showing a terminal session. The window has a dark grey header bar with three colored dots (red, yellow, green) in the top-left corner. The main area of the window is a terminal window with a black background. It contains the following text:

```
1 hello <- "Hello, World!"  
2 print(hello)  
3
```



Гипотеза (биологическая)

Планирование (полевого) эксперимента

Сбор материала

Проверка статистических гипотез

Вывод о биологической гипотезе

**В какой программе
считать?**

File		Edit	View	Insert	Format	Styles	Sheet	Data	Tools	Window	Help
		File	Open	Save	Print	Format	Cell	Range	Tools	Window	Help
		Font	Font	Font	Font	Font	Font	Font	Font	Font	Font
B40											
1											
2	1	45	заповедник								
3	2	65	заповедник								
4	3	74	заповедник								
5	4	23	заповедник								
6	5	56	заповедник								
7	6	12	заповедник								
8	7	43	заповедник								
9	8	56	заповедник								
10	9	23	заповедник								
11	10	54	заповедник								
12	11	65	заповедник								
13	12	23	заповедник								
14	13	54	заповедник								
15	14	87	заповедник								
16	15	58	заповедник								
17	16	24	заповедник								
18	17	74	заповедник								
19	18	85	заповедник								
20	19	31	заповедник								
21	20	69	парк								
22	21	32	парк								
23	22	52	парк								
24	23	80	парк								
25	24	42	парк								
26	25	46	парк								
27	26	23	парк								
28	27	86	парк								
29	28	32	парк								
30	29	46	парк								
31	30	85	парк								
32	31	35	парк								
33	32	64	парк								
34	33	23	парк								
35	34	45	парк								
36	35	75	парк								
37	36	23	парк								
38	37	72	парк								
39	38	42	парк								

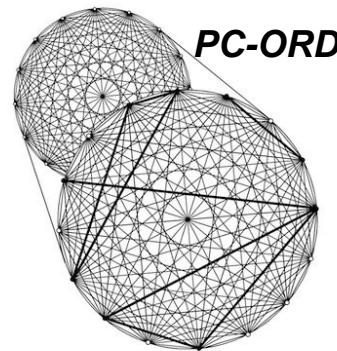
Есть ли различия в индивидуальной массе жуков *Carabus granulatus* в заповеднике и в городском парке?

- Каким методом сравнить?
- В какой программе посчитать?



Sultanov-rinat CC-BY
<https://www.gbif.org/occurrence/2980798058>

А что, если мы хотим проверить более сложные гипотезы?



Специализированные
статистические программы

Большинство из них платные

Каждую нужно осваивать
отдельно

Возможно, в данной
программе не будет нужного
вам метода



— язык программирования для статистической обработки данных и работы с графикой, а также свободная программная среда вычислений с открытым исходным кодом.

в мире 2 млн
пользователей R
(сентябрь 2022)

Developed R by Ross Ihaka and Robert Gentleman

1991



1976

Statistical programming Language S developed at Bell Labs

R got GNU General Public License and become free software.

1995



1993

Announcement of R to the public

R Version 1.0.0 released

2000



1997

R core group team is formed who controls the source code of R



2019

R 3.5.2 is the currently stable active version.





R Console

```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 The R Foundation for Statistical Computing
Platform: x86_64-w64-mingw32/x64 (64-bit)

R -- это свободное ПО, и оно поставляется безо всяких гарантий.
Вы вольны распространять его при соблюдении некоторых условий.
Ведите 'license()' для получения более подробной информации.

R -- это проект, в котором сотрудничает множество разработчиков.
Ведите 'contributors()' для получения дополнительной информации и
'citation()' для ознакомления с правилами упоминания R и его пакетов
в публикациях.

Ведите 'demo()' для запуска демонстрационных программ, 'help()' -- для
получения справки, 'help.start()' -- для доступа к справке через браузер.
Ведите 'q()', чтобы выйти из R.

[Загружено ранее сохраненное рабочее пространство]

> |
```



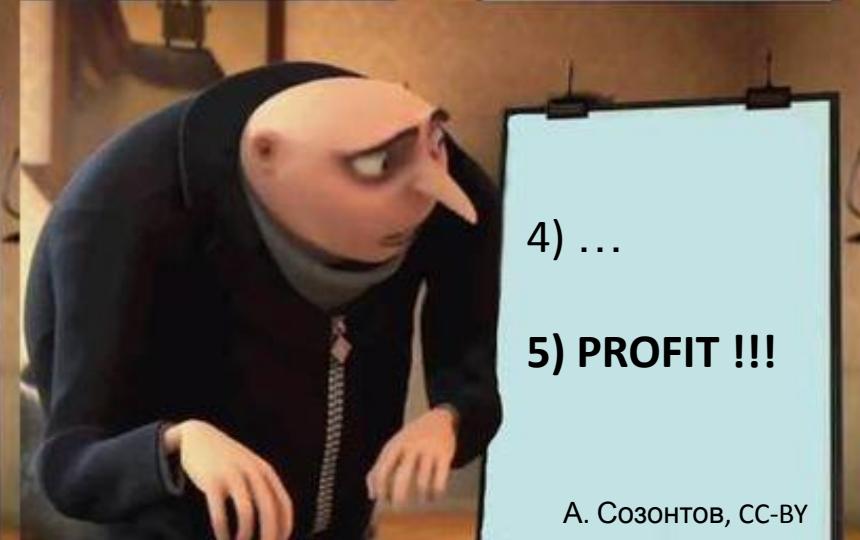
1) Что такое
этот ваш R
и почему я
должен
тратить на
это своё
время?



2) Чего R
умеет такого,
что я не могу
сделать в
Excel / Past /
Statistica и
другом ПО?



3) Считать
это ладно, но
сумеет ли R в
два клика
вывести мне
на экран
график,
карту,
диплом или



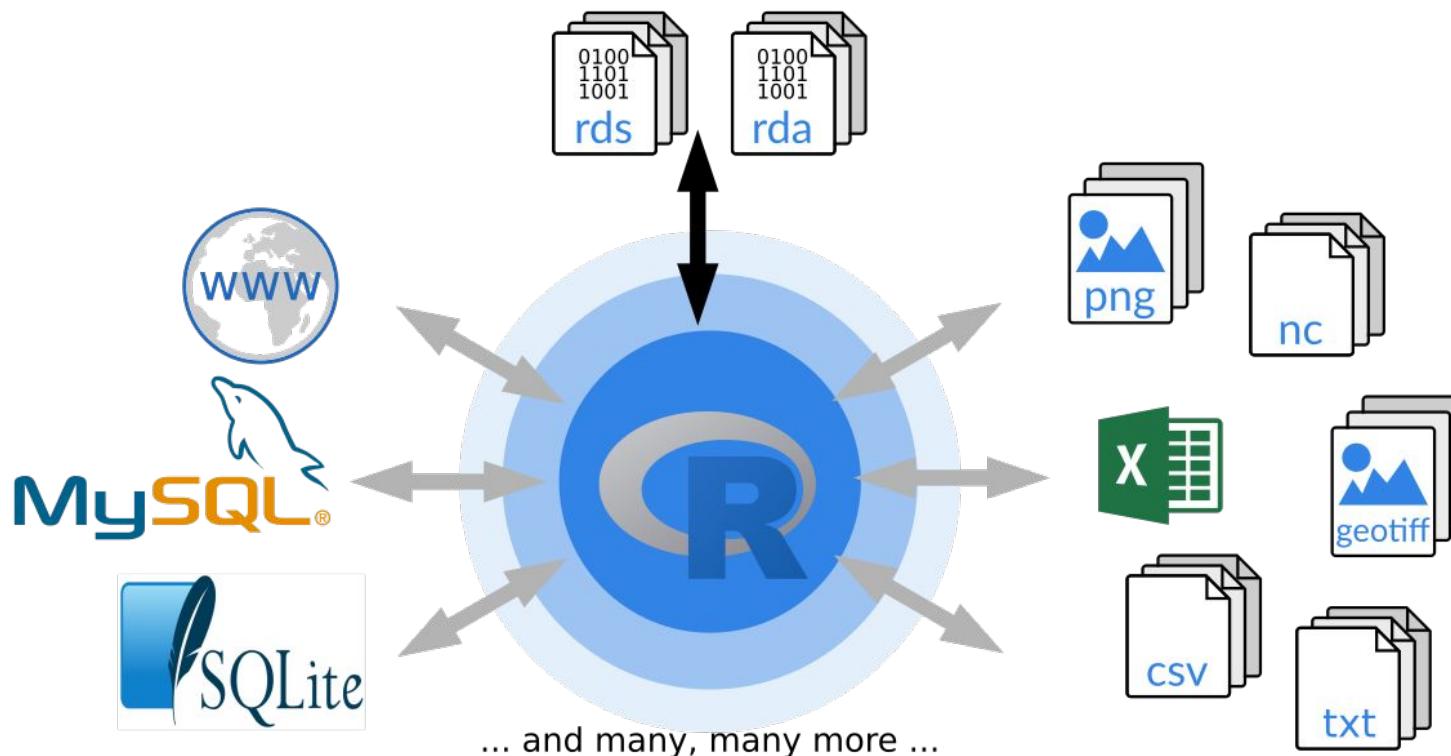
4) ...

5) PROFIT !!!

Что умеет R: работает на **всех основных ОС**

- Windows
- MacOS
- Linux

Что умеет R: работает с разными типами файлов и источниками данных



Что умеет R: статистика

- “Базовые” статистические методы (описательная статистика, одномерные и многомерные)
- 99.9% всех остальных существующих методов

R - любимый язык профессиональных статистиков. Новые методы очень быстро становятся доступными в R.

Use R!

Daniel Borcard
François Gillet
Pierre Legendre

Numerical Ecology with R

Second Edition

Milena Lajicevic
Nicholas Povak
Keith M. Reynolds

Springer

M. Henry H. Stevens

A Primer of Ecology with R

Springer

Introduction to R for Terrestrial Ecology

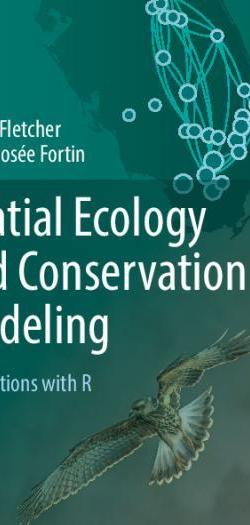
Basics of Numerical Analysis, Mapping,
Statistical Tests and Advanced Application of R

Springer

Robert Fletcher
Marie-Josée Fortin

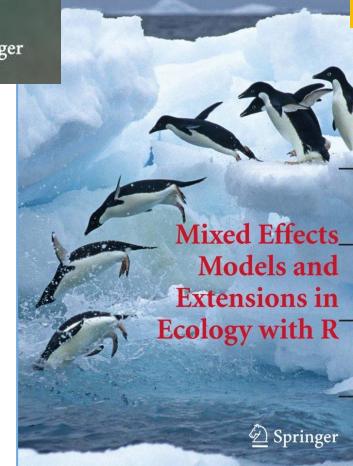
Spatial Ecology and Conservation Modeling

Applications with R



Springer

Alain F. Zuur • Elena N. Ieno
Neil J. Walker • Anatoly A. Saveliev
Graham M. Smith



Springer

Use R!

Nathan G. Swenson

Functional and Phylogenetic Ecology in R

Spatial Data
Analysis in Ecology
and Agriculture
Using R

SECOND EDITION

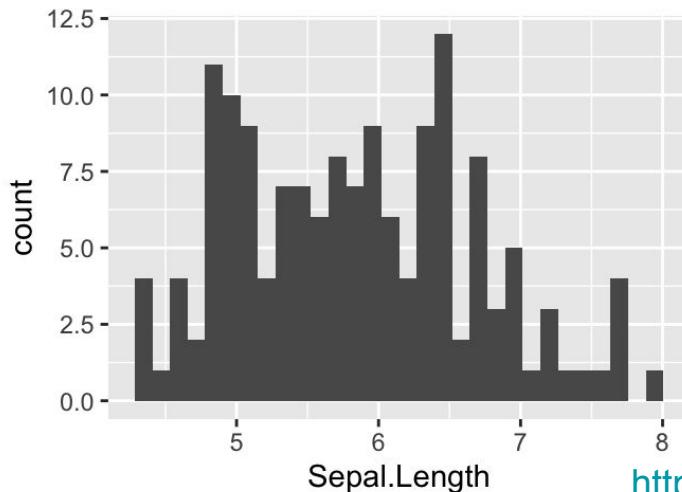
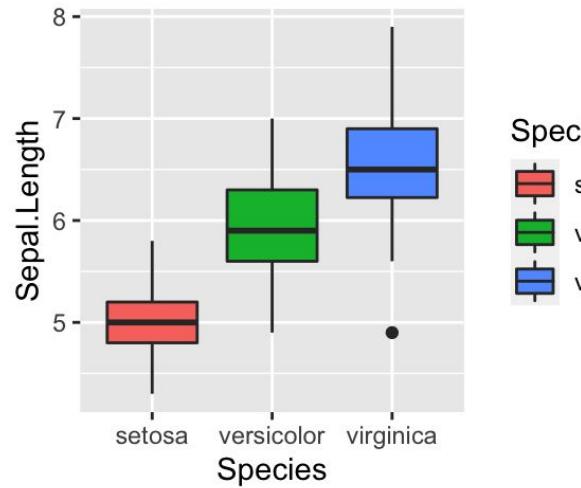
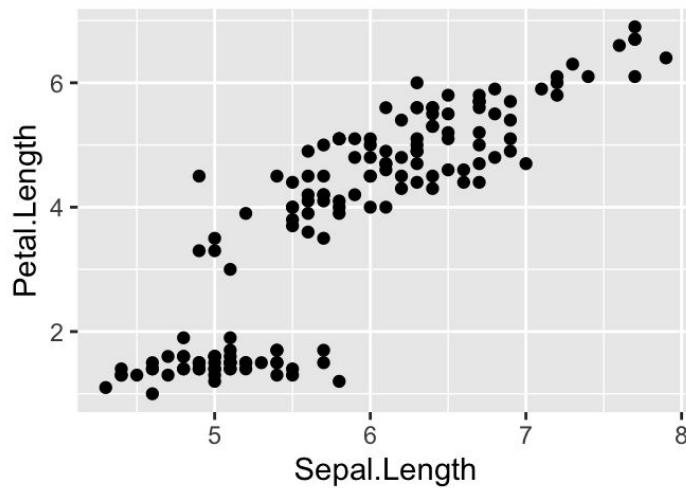


Richard E. Plant

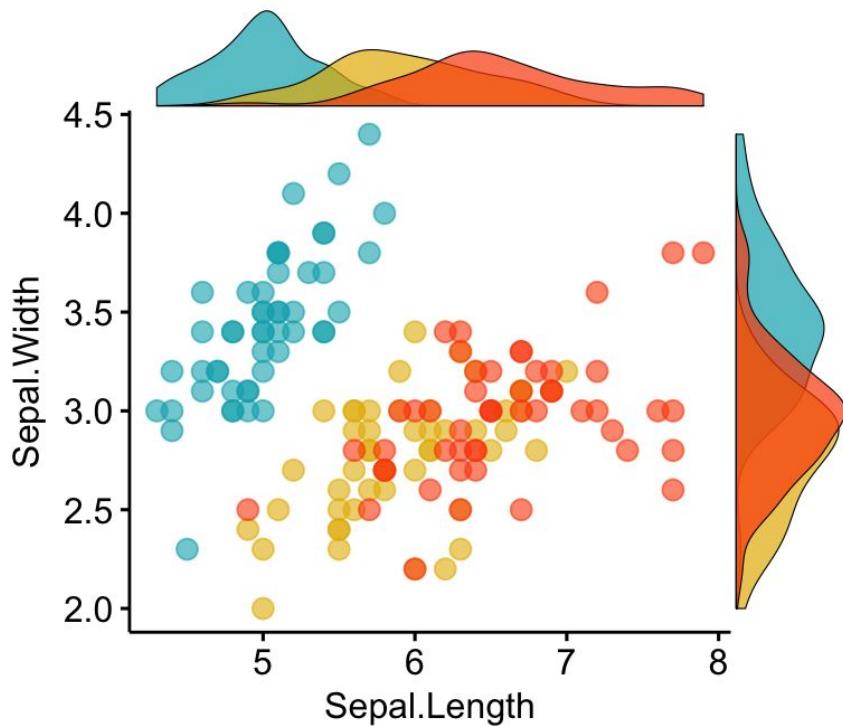
CRC Press
Taylor & Francis Group

Что умеет R: графика

- Все базовые графики (включая те, которые нельзя построить в Excel)
- Более сложные графики
- Визуализация специфических данных (например, 3D облака точек)
- Возможность настроить график (изменить цвет, линии, точки, размер шрифтов, заголовки и подписи, добавить легенду и т.д.)



Species ● setosa ● versicolor ● virginica

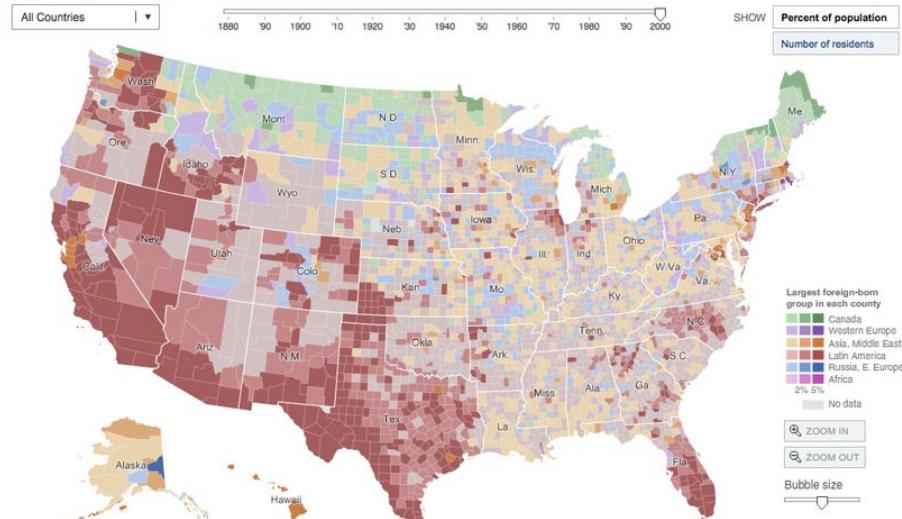


```
1 iris
2 library(ggpubr)
3
4
5 # Grouped Scatter plot with marginal density plots
6 ggscatterhist(
7   iris, x = "Sepal.Length", y = "Sepal.Width",
8   color = "Species", size = 3, alpha = 0.6,
9   palette = c("#00AFBB", "#E7B800", "#FC4E07"),
10  margin.params = list(fill = "Species", color = "black", size = 0.2)
11 )
```

Что умеет R: пространственные данные

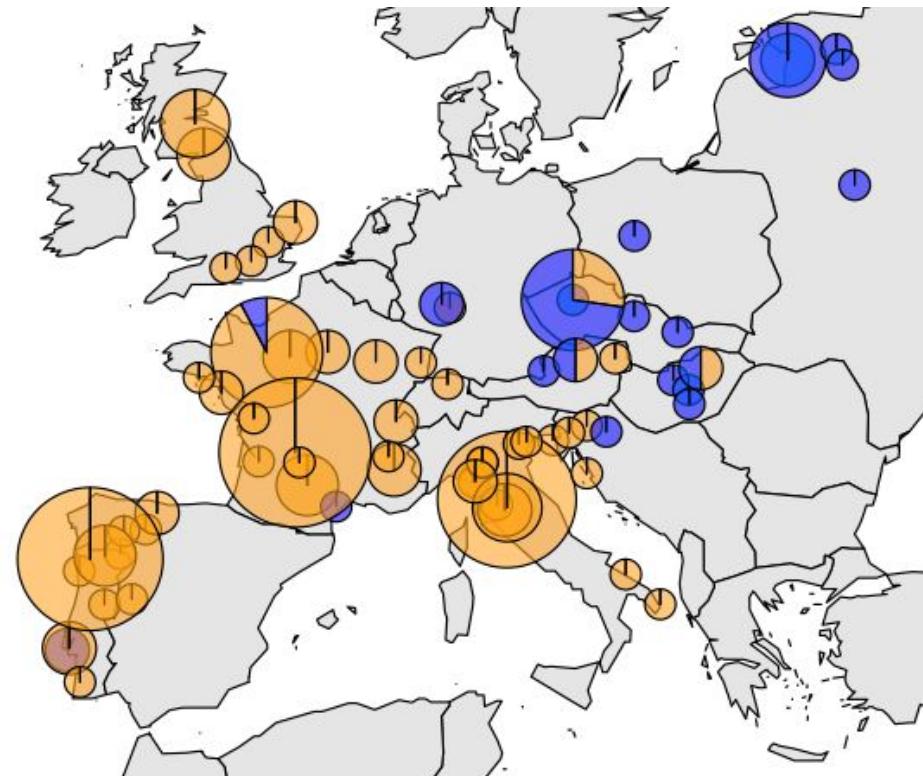
Immigration Explorer

Select a foreign-born group to see how they settled across the United States.



Matthew Bloch and Robert Gebeloff/The New York Times

Sources: Social Explorer, www.socialexplorer.com; Minnesota Population Center; U.S. Census Bureau



Что умеет R: web-страницы

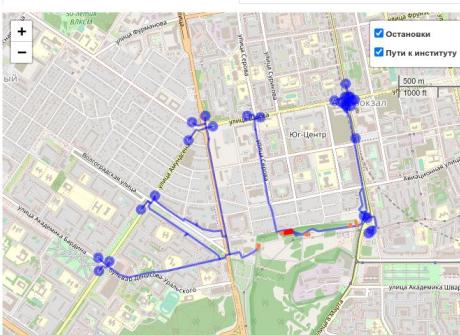
RPubs by RStudio

Sign in Register

Маршруты и ключевые точки для участников полевой школы

На карте отображаются корпуса ИЭРиЖ, остановки общественного транспорта и пути от них до института, продуктовые магазины и супермаркеты, кафе, столовые и другие места общепита, а также маршруты к ним от института.

Общественный транспорт и дорога к институту Места общественного питания и дорога к ним



500 м
1000 м

Остановки
Пути к институту

Leaflet | © OpenStreetMap contributors, CC-BY-SA

Вход и выход через главный корпус ИЭРиЖ и главный вход ботсада свободные. Проход с западной стороны через генетико-биологический корпус возможен, но только в сопровождении сотрудников. Проход через главное здание ботсада считаем предпочтительным, т.к. с западной стороны ведутся строительные и ремонтно-дорожные работы.

РЕГИСТРАЦИЯ

Главный корпус - четырехэтажное здание, выделенное красным цветом на карте.

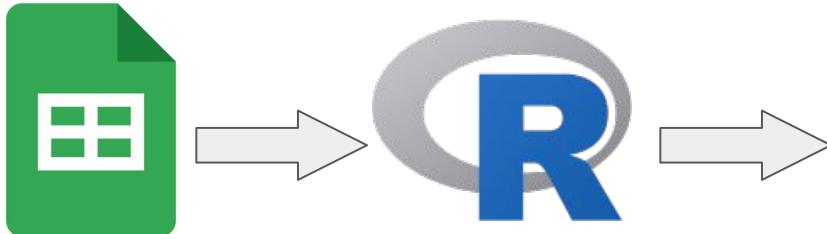
Comments (-) Share Hide Toolbar

Питание и дорога by Соzонтов Артем Last updated about 1 year ago

https://rpubs.com/Sozontov/soilzool21_food

<https://www.dsquintana.blog/free-website-in-r-easy/>

Что умеет R: скомпоновать текст по шаблону



Biotopic preferences: I – All the habitats (20 %), IV – Edges of mixed forests (15 %), III – Bogs (13 %), II – Anthropogenic habitats (12 %), II – Meadowlands (6 %).

Associated species. Quantitative: *Tetragnatha pinicola* (0.28), *Dictyna arundinacea* (0.25), *Tibellus oblongus* (0.22), *Xysticus ulmi* (0.22), *Tmarus piger* (0.22). **Qualitative:** *Dictyna arundinacea* (0.39), *Tetragnatha pinicola* (0.34), *Tibellus oblongus* (0.34), *Phylloneura impressa* (0.31), *Evarcha arcuata* (0.31).

Stratum: herb.

Range: Palaearctic (West-Central-Palaearctic Subboreal).

Neoscona adianta (Walckenaer, 1802)

Material (Fig. 29): Alnashy Distr.: Golushurma (47): 1 juv., sloping steppe meadow, 17.VI.2015, leg. A.N. Sozontov; Karakulino Distr.: Byrgynda

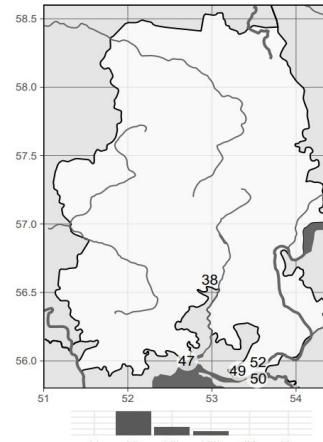


Fig. 29. Distribution map and phenology of *N. adianta*

Arthropoda Selecta. Supplement No. 5

A.N. Sozontov, S.L. Esyunin

Spiders of the Udmurt Republic:
fauna, ecology, phenology
and distribution

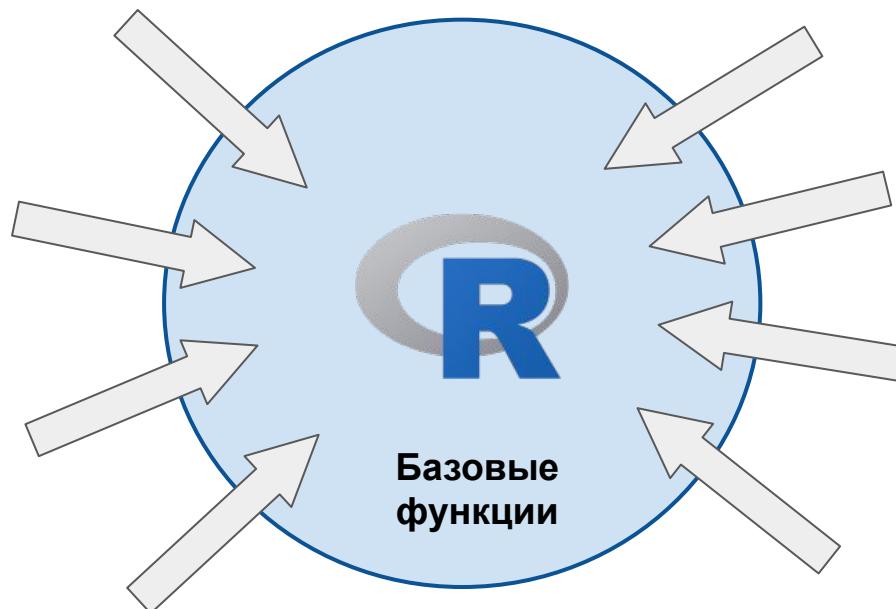
KMK Scientific Press
Moscow ♦ 2022

Spiders of the Udmurt Republic: fauna, ecology, phenology and distribution» M.: KMK, 2022. 285p.
https://ipae.uran.ru/sites/default/files/publications/users/Sozontov_Esyunin_2022.pdf

И многое другое

Как это устроено: пакеты

Пакет - набор функций для решения определенной задачи. Пакет включает код, описание функций (документацию) и часто учебные данные.



Где хранятся пакеты: репозитории

- **CRAN** (the Comprehensive R Archive Network)
<https://cran.r-project.org/> - основной репозиторий
- **Bioconductor** <https://www.bioconductor.org/> - для пакетов по биоинформатике
- **GitHub** <https://github.com/trending/r>



разработано
18 839
пакетов для



<https://cran.r-project.org/web/packages/>

Top Companies Using R

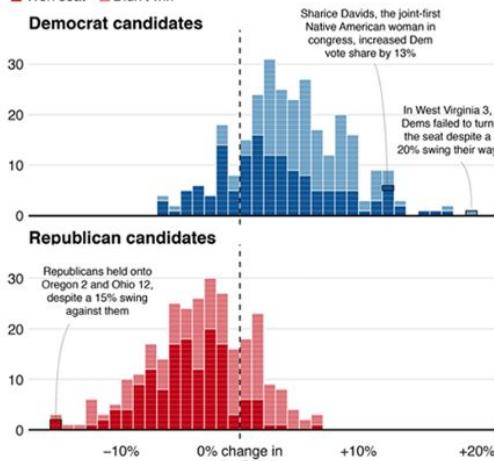


How the BBC Visual and Data Journalism team works with graphics in R

Blue wave

■ Won seat ■ Didn't win

Democrat candidates

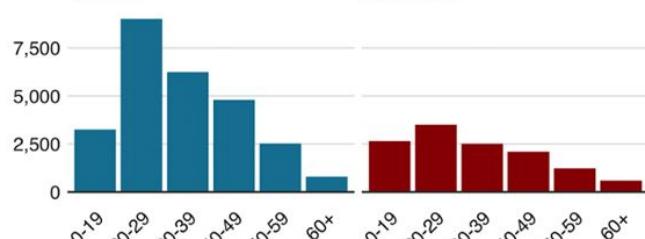


Republican candidates

Homophobic hate crimes are mainly committed by young people on young people

Number in each age group 2014 - 2017

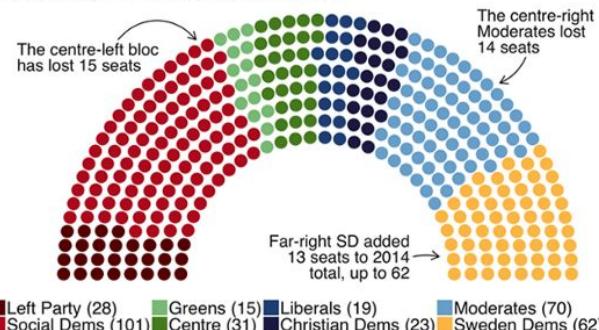
Victims



Suspects

Source: BBC Freedom of Information requests to UK police forces

Results of the 2018 election



Source: Reuters

BBC

Democrats take the House

Dem 232

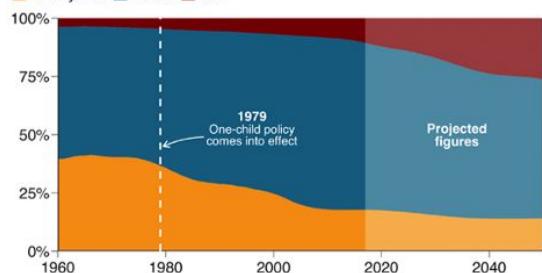
218 to win

Rep 198

Breakdown of China's population by age group

Proportion of total population (1960-2050)

■ 0-14 years ■ 15-64 ■ 65+



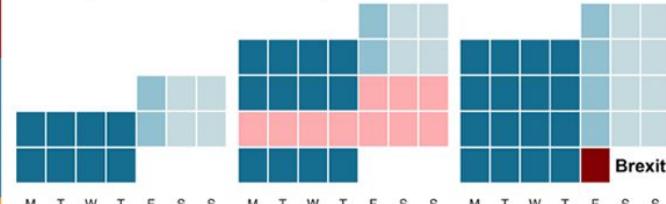
The Commons has 36 normal working days until Brexit

■ Monday to Thursday ■ Friday ■ Weekend ■ Recess

January

February

March



Note: The House of Commons sometimes sits on Fridays to debate individual MPs' bills

Source: The World Bank

BBC Source: Parliament

BBC Source: AP. Grey districts are undeclared

BBC

Минус Решение

Нет русскоязычного интерфейса и справки	<ul style="list-style-type: none">Набор команд ограничен, есть руководства на русском языке
Интерфейс R не дружественный	<ul style="list-style-type: none">Существуют оболочки для R, предоставляющие оконный интерфейсR можно интегрировать в Statistica (>10), SPSS и MS Excel
Чтобы работать в R, нужно уметь программировать	<ul style="list-style-type: none">Да, но научиться основам достаточно просто; уже базовые навыки позволяют автоматизировать однотипные операции и тем самым ускорять решение многих задач (вся мощь R кроется именно в возможности писать скрипты)
Чувствителен к синтаксису	<ul style="list-style-type: none">Смириться и быть внимательнее
Не существует технической поддержки R	<ul style="list-style-type: none">Существует обширная справочная и учебная литература по работе в R (в т.ч. на русском), большое сообщество пользователей, многие из которых готовы помочь.
Интерпретируемый и потому медленный	<ul style="list-style-type: none">Рядовой пользователь не заметит, что скорость вычислений низкая в сравнении с копмилируемыми или JIT-компилируемыми языками программированияИмеются простые и эффективные средства распараллеливания вычисленийПредоставляет базу для дальнейшего освоения более сложных и производительных языков (Python, Julia, Go, Ruby)
Сложно ориентироваться	<ul style="list-style-type: none">Использовать хорошие практики, узнать о которых можно здесь

Плюс Применение

Свободная лицензия	<ul style="list-style-type: none">Свобода и бесплатность использования
Автоматизация	<ul style="list-style-type: none">Экономия времени за счет автоматизации рутинных операций
Концепция «Все в 1»	<ul style="list-style-type: none">Не нужно учить много программ, достаточно одной для решения всех задач
Предобработка	<ul style="list-style-type: none">Простота и скорость чистки, доработки и преобразования первичных данных
Статистика	<ul style="list-style-type: none">Самый полный набор всевозможных статистических процедур: поправки на множественное сравнение, GLM с любыми распределениями, всевозможные варианты бутстрепа, весь арсенал методов многомерной статистики и т.д.
Графика	
Распараллеливание	<ul style="list-style-type: none">Ресурсоемкие вычисления можно делать эффективнее
ГИС-технологии	<ul style="list-style-type: none">Скорости обработки карт и других пространственных данных
Отчеты RMarkdown	<ul style="list-style-type: none">Не надо переделывать всю диссертацию, если есть изменения в данных
Веб-приложения Shiny	<ul style="list-style-type: none">Зачем запускать код многократно, если можно один раз и пойти пить чай?
SDM и rgbif	<ul style="list-style-type: none">Прямой доступ к данным с GBIF и использование их для моделирования ареалов

И все это только вершина айсберга...

Протоколирование

• Остаются задокументированы ВСЕ этапы работы с данными. Это сочетается с концепциями FAIR-data и OpenScience, улучшает воспроизводимость, позволяет коллегам учиться или заходить

01 Free and Open-source Tool

02 Large Community of Users

03 Latest cutting edge technology

04 Independent Platform

05 Gateway to lucrative career

06 Has a Robust visualization library

07 Go to language for Stats. & Data Science

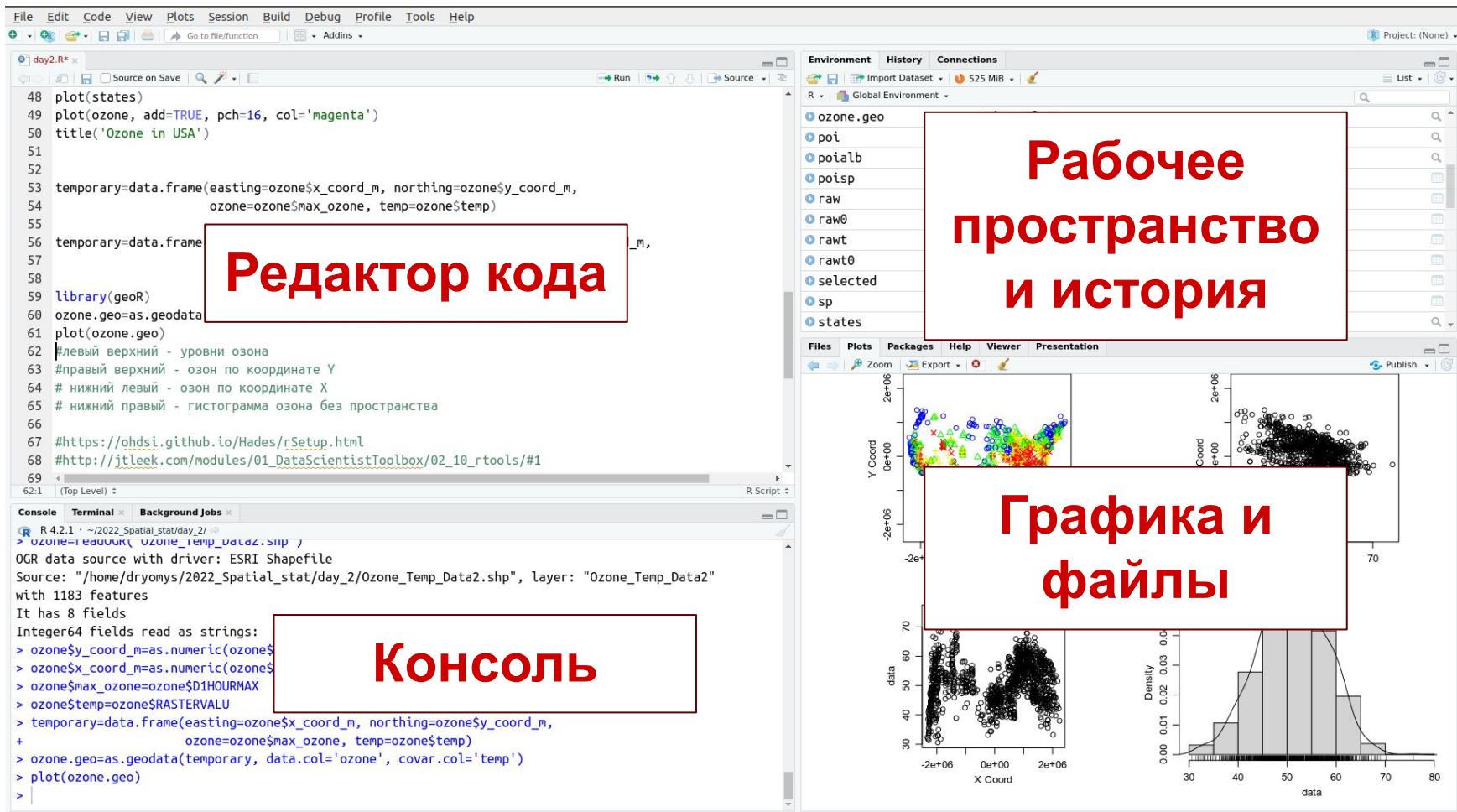
08 Used in almost every industry

Why Learn



Начало работы в R





Функции

Буквенные символы с круглыми скобками сразу после названия функции.

Мы подаем на вход (внутрь скобочек) какие-то данные, внутри этих функций происходят какие-то вычисления, которые выдают в ответ какие-то другие данные (или же функция записывает файл, рисует график и т.д.).

Данные на входе называются **аргументом** функции.

`sqrt()` # функция для вычисления квадратного корня

Чтобы узнать, какие аргументы есть у функции, нужно вызвать справку

`?sqrt()` # вызвать справку по функции `sqrt()`

Комментарии

Всё, что написано после значка **#** не будет читаться R как команда.

- Что делает функция
- Что записано в переменную
- Последовательность действий
- Любые другие заметки

Переменные

Каждый объект должен иметь своё имя

a<-c(1, 2, 3, 4, 5) или a=c(1, 2, 3, 4, 5)

Обязательно делайте названия переменных осмысленными!

Старайтесь делать при этом их понятными и короткими, это сохранит вам очень много времени, когда вы (или кто-то еще) будете пытаться разобраться в написанном ранее коде.

Типы объектов в R

Векторы

Матрицы

Списки

Таблицы данных

Факторы

Векторы

Вектор - это набор значений одного типа

`c()` # функция для создания вектора

`c(21, 15, 9, 45, 17)` - числовой вектор

`c('red', 'green', 'blue')` - текстовый вектор

Одна из самых раздражающих причин ошибок в коде — это использование с из кириллицы вместо с из латиницы. Видите разницу? А R видит!

Факторы

Используется для хранения категориальных данных. Они имеют градации или уровни (levels).

`factor()` #функция для создания фактора

red green blue orange magenta

Levels: blue green magenta orange red

Таблицы данных (data frames)

Таблицы данных – это основной класс объектов R, используемых для хранения данных.

Обычно такие таблицы подготавливаются при помощи внешних приложений и затем загружаются в среду R. Также таблицу можно собрать из нескольких векторов средствами самой системы R.

`data.frame()` # функция для создания таблиц

Рабочая директория

Это папка, в которой хранятся ваши рабочие файлы.

Программе R нужно будет указать, откуда брать файлы.

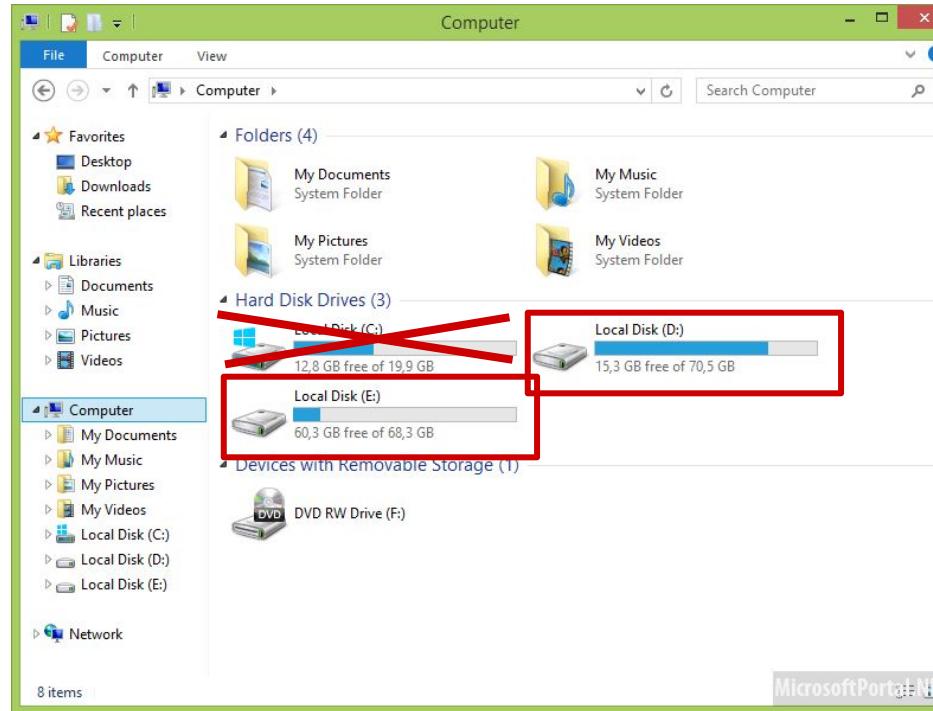
```
getwd() # узнать рабочую директорию
```

```
setwd() # задать рабочую директорию
```

Или воспользоваться меню R-studio (подробнее на практическом занятии)

Рабочая директория

Не заводите рабочую директорию на рабочем столе!



Материалы для самостоятельного изучения

The R Graph Gallery <https://r-graph-gallery.com/>

Quick-R <https://www.statmethods.net/> (<https://www.statmethods.net/stats/index.html> основная статистика)

STHDA R Basics: Quick and Easy <http://www.sthda.com/english/wiki/r-basics-quick-and-easy> и другие материалы на этом сайте

<https://r-coder.com/> код + статистика

<https://agricolamz.github.io/2020-2021-ds4dh/index.html> Наука о данных в R для программы Цифровых гуманитарных исследований (Г.А. Мороз, И.С. Поздняков)

[An Introduction to Statistical Programming Methods with R](#)

[Modern Statistics with R. From wrangling and exploring data to inference and predictive modelling](#)

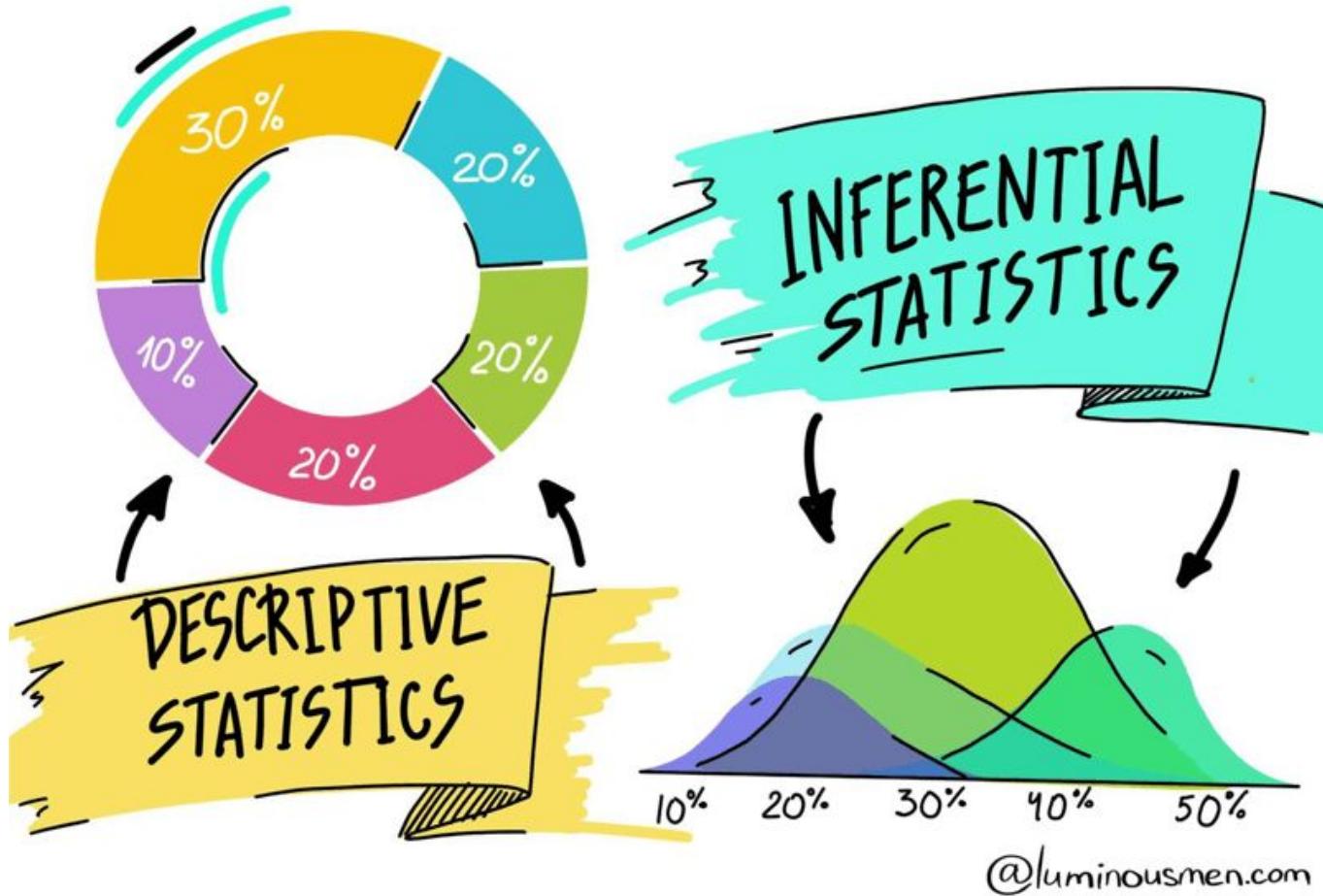
[Introduction to Data Exploration and Analysis with R](#)

Рекомендуемая литература на русском

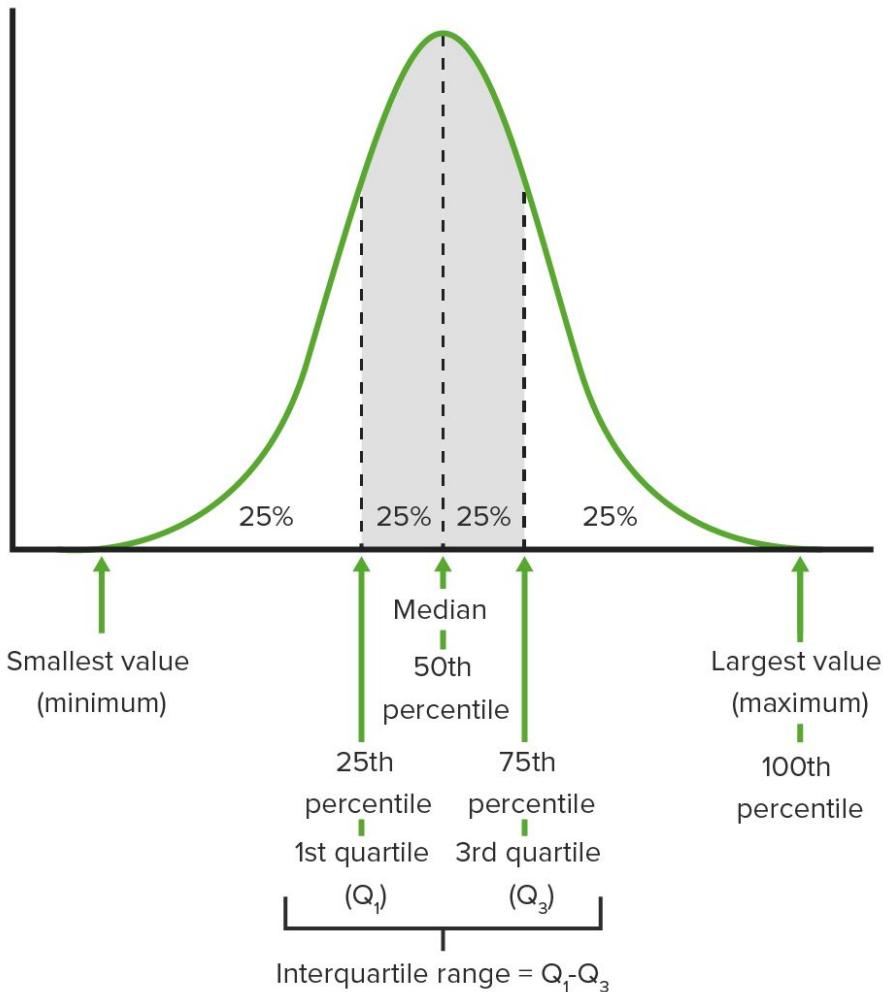
- **Кабаков Р.И. (2014)** R в действии. Анализ и визуализация данных в программе R. М.: ДМК Пресс. 588 с.
- **Шипунов А.Б. и др. (2014)** Наглядная статистика. Используем R! [Электронная книга]
- **Мастицкий С.Э., Шитиков В.К. (2014)** Статистический анализ и визуализация данных с помощью R [Электронная книга]
- **Шитиков В.К., Мастицкий С.Э. (2017)** Классификация, регрессия и другие алгоритмы Data Mining с использованием R [Электронная книга]
- **Эрве М. (2016)** Путеводитель по применению статистических методов с использованием R. Планирование исследований и анализ результатов в биологии с помощью программного обеспечения R [Электронная книга]
- **Зарядов И.С. (2010)** Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Изд-во РУДН. 207 с.
- **Зарядов И.С. (2010)** Статистический пакет R: теория вероятностей и

Разведочный анализ данных в R





@luminousmen.com



EXPLORING DATA

DESCRIPTIVE STATISTICS

CONTINUOUS DATA MoV

STANDARD DEVIATION

RANGE (MIN, MAX)

INTERQUARTILE RANGE

CONTINUOUS DATA MoL

MEAN

MEDIAN

MODE

CATEGORICAL DATA

FREQUENCY

PERCENTAGE (ROWS, COLUMNS, OR TOTAL)

VISUALIZATION

CONTINUOUS DATA

HISTOGRAM

BOX & WHISKERS PLOT

DOT PLOT

(can be used against categorical data)

SCATTER PLOT

(two continuous variables)

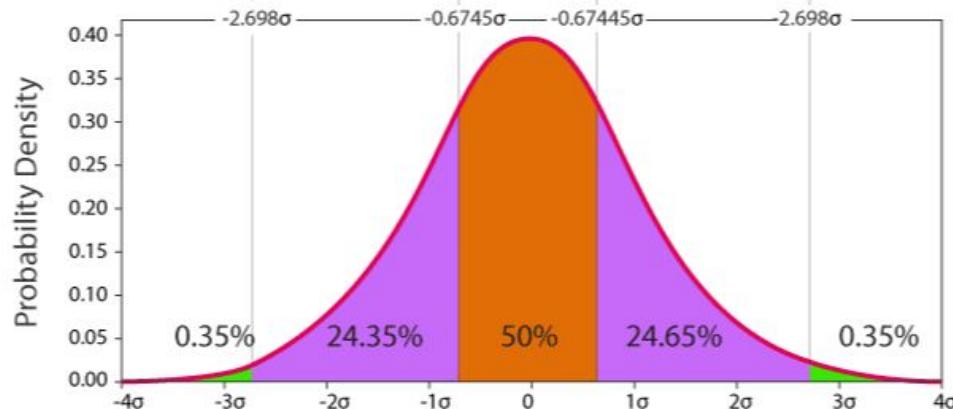
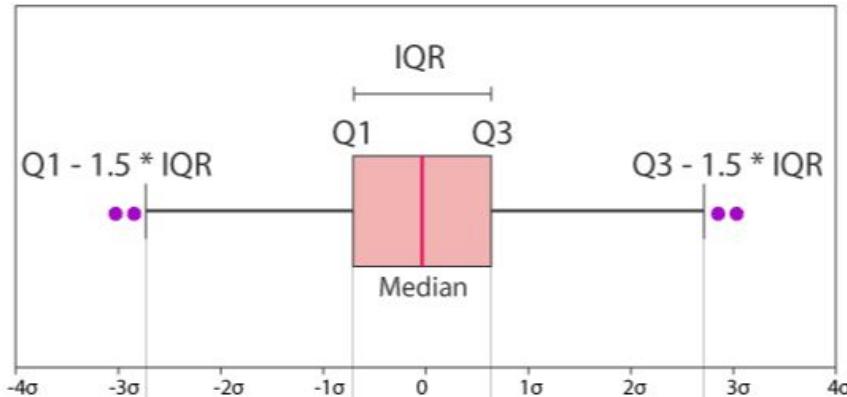
CATEGORICAL DATA

BAR CHART

CLUSTERED BAR CHART

(2 categorical variables)

BAR CHART WITH ERRORS



Boxplot on a normal distribution

Independent variable		Dependent variable		
		Nominal	Ordinal (continuous non-normal)	Continuous
Nominal, dichotomous	Independent samples	Z test proportions comparison Chi-squared test Fisher's exact test	Mann-Whitney's U test	Student's t test (independent samples)
	Paired samples	McNemar's test Z test and binomial method	Wilcoxon rank sum test	Student's t test (paired samples)
Nominal, polytomous		Chi-squared test Binomial method	Kruskal-Wallis' test Friedman's test (paired samples)	Analysis of variance (ANOVA)
Continuous		Student's t test	Spearman's correlation coefficient	Pearson's correlation coefficient Lineal regression

From: Ochoa Sangrador C, ed. Diseño y análisis en investigación. International Marketing & Communication, Madrid; 2019.

