

# First Meeting: crash into R

Eduard Bukin

February 21, 2018



## Plan of the first meeting of the R-Users at IAMO

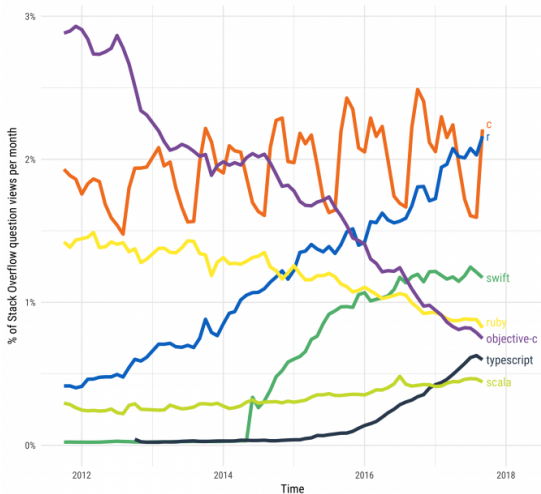
- 1 R and what is so cool about R!
- 2 Main data types and structures: `objects`, `vectors`, `c()`, `list()`, `matrix()`, `data.frame`, `$`, `[ ]`, `[[ ]]`
- 3 Packages: `install.packages()`, `tidyverse`
- 4 **Get data in R**: `base::read.csv()` - why we should never use it; `readr::read_csv()`; `readxl::read_excel()`
- 5 Glance at data in R: `str()`; `glimpse()`; `tibble::tbl_df()`
- 6 ***tidy data***
- 7 Basic grammar of data manipulation `dplyr`: `select()`, `filter()`, `mutate()`, `summaries()`, `group_by()`
- 8 How to learn? Where to get help to learn? Cheat sheets

# R and what is so cool about R!

# Growth of R

## Stack Overflow Traffic to Programming Languages

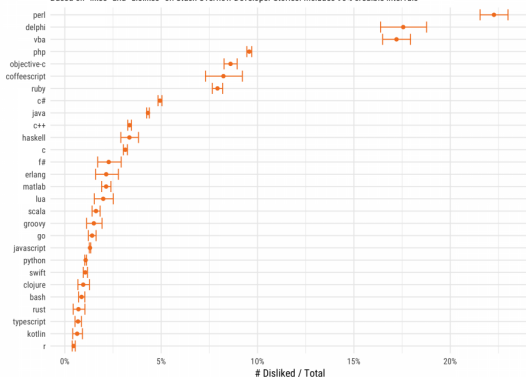
Based on visits to Stack Overflow questions from World Bank high-income countries.  
The more-visited languages of Python, JavaScript, Java, C#, and PHP were omitted.



## Least disliked language

### How disliked is each programming language?

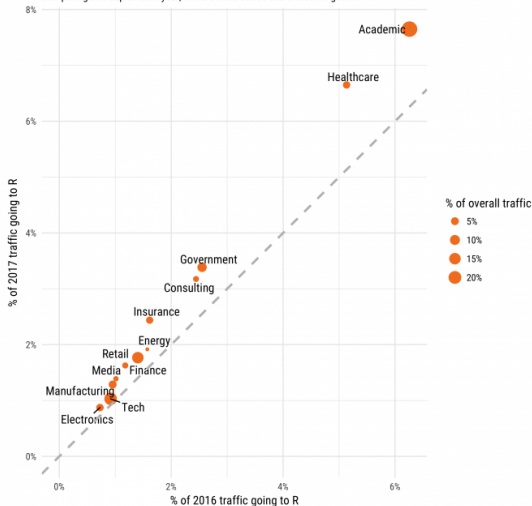
Based on "likes" and "dislikes" on Stack Overflow Developer Stories. Includes 95% credible intervals



# R-UseRs

## Traffic by industry to R

Comparing Jan-Sep of each year, in the United States and United Kingdom.



## Some other cool things

- Access and handle data with tidyverse and other tools
- Easy to analyse data with specifically developed tools CRAN Task View
- Build dynamic and interactive data visualization with Shiny, plotly, and many other widgets
- Built dashboards
- Do reproducible research and dynamic documentations with rmarkdown
- Teach and learn in R with swirl



## Some other cool things

- Do research
- Meet incredible R community:
- With Rstudio, R-studio developers' blogs and community
- With The Comprehensive R Archive Network, with their **12 000+** packages in the repository and Task Views
- Open source and publicly developed software using github
- Multiple communities: R Open science, stakexchange, R-pubs and many other things!

# Moving to Tutorial in R

Let us practice

# Resources

Useful resources for learning R

## Where are how to learn?

### PRACTICE

- refer to the R Studio cheetsheets;

### Read:

- R for Data Science;
- Official CRAN documentation (sometimes boring to read);
- Search questions and answers on [stackexchange](#);
- [r pubs](#);
- CRAN Task View;
- R Studio Community;

## Where are how to learn?

Learn:

- TRY SWIRL;
- E-courses DataCamp, coursera - R programming, edx - Statistics in R, edx - R programming.;
- do exercise;
- see R by example;

## Some other resources:

- Impatient R
- R Wikibooks
- Examples of data analysis in R
- GOOGLE

Thank you

NOT COVERED



## Tidy data

- What is tidy?
- How to make it tidy?
- Get data in R: `readr::read_csv()`; `readxl::read_excel()`

## Data sets examples 1 (1)

##	country	year	cases	population
## 1	Afghanistan	1999	745	19987071
## 2	Afghanistan	2000	2666	20595360
## 3	Brazil	1999	37737	172006362
## 4	Brazil	2000	80488	174504898
## 5	China	1999	212258	1272915272
## 6	China	2000	213766	1280428583

## Data sets examples 1 (2)

##	country	year	rate
## 1	Afghanistan	1999	745/19987071
## 2	Afghanistan	2000	2666/20595360
## 3	Brazil	1999	37737/172006362
## 4	Brazil	2000	80488/174504898
## 5	China	1999	212258/1272915272
## 6	China	2000	213766/1280428583

## Data sets examples 1 - structure

country	year	population
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 17206362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

table3

country	year	population
Afghanistan	1999	745 / 19987071
Afghanistan	2000	2666 / 20595360
Brazil	1999	37737 / 17206362
Brazil	2000	80488 / 174504898
China	1999	212258 / 1272915272
China	2000	213766 / 1280428583

variables

country	year	population
Afghanistan	99	745 / 19987071
Afghanistan	00	2666 / 20595360
Brazil	99	37737 / 17206362
Brazil	00	80488 / 174504898
China	99	212258 / 1272915272
China	00	213766 / 1280428583

values

## Data sets examples 2

##	country	year	key	value
## 1	Afghanistan	1999	cases	745
## 2	Afghanistan	1999	population	19987071
## 3	Afghanistan	2000	cases	2666
## 4	Afghanistan	2000	population	20595360
## 5	Brazil	1999	cases	37737
## 6	Brazil	1999	population	172006362
## 7	Brazil	2000	cases	80488
## 8	Brazil	2000	population	174504898
## 9	China	1999	cases	212258
## 10	China	1999	population	1272915272
## 11	China	2000	cases	213766
## 12	China	2000	population	1280428583

## Data sets examples 2 - structure

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

variables

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

observations

### Data sets examples 3

```
##          country    1999    2000
## 1 Afghanistan      745     2666
## 2          Brazil 37737   80488
## 3           China 212258 213766
```

```
##          country      1999      2000
## 1 Afghanistan 19987071 20595360
## 2          Brazil 172006362 174504898
## 3           China 1272915272 1280428583
```

## Data sets examples 3 - structure

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

country	1999	2000
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

table5

country		
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

country		
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

variables

country		
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

country		
Afghanistan	19987071	20595360
Brazil	172006362	174504898
China	1272915272	1280428583

observations



## Tidy data

Your data will be easier to work with in R if it follows three rules:

- Each variable in the data set is placed in its own column
- Each observation is placed in its own row
- Each value is placed in its own cell

Data that satisfies these rules is known as tidy data.

Borrowed from Data science with R: Tidying

## Tidy data example - structure

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

table1

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

variables

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

observations

## Tidy data example

DSR::table1

##		country	year	cases	population
## 1		Afghanistan	1999	745	19987071
## 2		Afghanistan	2000	2666	20595360
## 3		Brazil	1999	37737	172006362
## 4		Brazil	2000	80488	174504898
## 5		China	1999	212258	1272915272
## 6		China	2000	213766	1280428583

## How to make tidy data?

Use R package `tidyr`.

Functions:

- `spread()`
- `gather()`

spread()

country	year	key	value
Afghanistan	1999	cases	745
Afghanistan	1999	population	19987071
Afghanistan	2000	cases	2666
Afghanistan	2000	population	20595360
Brazil	1999	cases	37737
Brazil	1999	population	172006362
Brazil	2000	cases	80488
Brazil	2000	population	174504898
China	1999	cases	212258
China	1999	population	1272915272
China	2000	cases	213766
China	2000	population	1280428583

table2

country	year	cases	population
Afghanistan	1999	745	19987071
Afghanistan	2000	2666	20595360
Brazil	1999	37737	172006362
Brazil	2000	80488	174504898
China	1999	212258	1272915272
China	2000	213766	1280428583

gather()

The diagram illustrates the use of the `gather()` function in R. It shows a transformation from a wide table to a long table. The wide table on the right has columns for country, 1999, and 2000. The long table on the left has columns for country, year, and cases. Arrows indicate that the '1999' and '2000' columns from the wide table are being gathered into the 'year' column of the long table, while the 'country' column remains unchanged.

country	year	cases
Afghanistan	1999	745
Afghanistan	2000	2666
Brazil	1999	37737
Brazil	2000	80488
China	1999	212258
China	2000	213766

country	1999	2000
Afghanistan	745	2666
Brazil	37737	80488
China	212258	213766

table4

## Non-tidy data

Be aware, that sometimes, data cannot be tidy and in fact is it easier to work with such data.

For more information, see Non-tidy data.

Where to go next?

Data manipulation with `dplyr` - next meeting