Модели и технологии оперативного анализа данных

Лекция 7
Трансформация данных в R

Гедранович Ольга Брониславовна, старший преподаватель кафедры ИТ, МИУ volha.b.k@gmail.com

#### Вопросы лекции

- Фильтрация (filter)
- Сортировка (arrange)
- Выборка (select)
- Модификация (mutate)
- Групповые операции
- Принципы Tidy Data и пакет tidyr



#### Трансформация данных

Ключевые функции пакета dplyr для трансформации данных:

- Pick observations by their values filter().
- Reorder the rows arrange().
- Pick variables by their names select().
- Create new variables with functions of existing variables mutate().
- Collapse many values down to a single summary summarise().
- These can all be used in conjunction with group\_by()
   which changes the scope of each function from operating
   on the entire dataset to operating on it group-by-group.



#### Трансформация данных

All verbs work similarly:

- The first argument is a data frame.
- The subsequent arguments describe what to do with the data frame, using the variable names (without quotes).
- The result is a new data frame.

Together these properties make it easy to chain together multiple simple steps to achieve a complex result.



#### Фильтрация (filter) filter() allows you to subset observations based on their values 2013 1 2013 1 529 4 540 2 545 -1 600 -6 558 -4 600 -5 600 -3 1 1 533 1 1 542 1 1 544 1 1 554 1 1 555 1 1 557 1 1 557 2013 2013 812 913 709 2013 854 2013 600 600 838 # ... with 832 more rows, and 10 more variables: carrier <chr>>, flight <int>>, tailnum <chr>>, # origin <chr>>, dest <chr>>, air\_time <dbl>>, distance <dbl>>, hour <dbl>>, minute <dbl>>, time\_hour <dttm>

## Сортировка (arrange)

arrange() changes rows' order.

It takes a data frame and a set of column names (or more complicated expressions) to order by.

If you provide more than one column name, each additional column will be used to break ties in the values of preceding columns.



## Сортировка (arrange)

arrange(flights, year, month, day)

# A tibble: 336,776 × 19

	year	month	day	dep_time	sched_dep_time	dep_delay	arr_time	sched_arr_time	arr_delay
	<int></int>	<int></int>	<int $>$	<int></int>	<int></int>	<dbl></dbl>	<int></int>	<int></int>	<dbl></dbl>
1	2013	1	1	517	515	2	830	819	11
2	2013	1	1	533	529	4	850	830	20
3	2013	1	1	542	540	2	923	850	33
4	2013	1	1	544	545	-1	1004	1022	-18
5	2013	1	1	554	600	-6	812	837	-25
6	2013	1	1	554	558	-4	740	728	12
7	2013	1	1	555	600	-5	913	854	19
8	2013	1	1	557	600	-3	709	723	-14
9	2013	1	1	557	600	-3	838	846	-8
10	2013	1	1	558	600	-2	753	745	8

- # ... with 336,766 more rows, and 10 more variables: carrier <chr>, flight <int>,
- # minute <dbl>, time\_hour <dttm>



```
Выборка (select)
select() allows to zoom in on a useful subset basing on the names of the variables.
select(flights, year, month, day)
# A tibble: 336,776 \times 3
   year month day
  <int> <int> <int>
  2013
  2013 1
  2013 1
  2013 1
  2013 1
  2013 1
  2013 1 1
  2013 1 1
  2013
         1
10 2013
# ... with 336,766 more rows
```

```
There are a number of helper functions you can use within select():

• starts_with("abc"):matches names that begin with "abc"

• ends_with("xyz"):matches names that end with "xyz"

• contains("ijk"):matches names that contain "ijk"

• matches("(.)\\1"):selects variables that match a regular expression. This one matches any variables that contain repeated characters

• num_range("x", 1:3):matches x1, x2 and x3.

rename(flights, tail_num = tailnum)

select(flights, time_hour, air_time, everything())
```

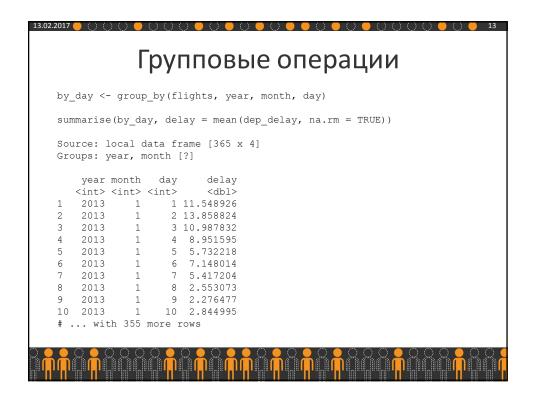
# 

## Групповые операции

summarise() collapses a data frame to a
single row.

summarise() is not terribly useful unless we
pair it with group\_by(). This changes the unit
of analysis from the complete dataset to
individual groups.





## Групповые операции

If you need to remove grouping, and return to operations on ungrouped data, use ungroup ().

```
daily %>%
  ungroup() %>%  # no longer grouped by date
  summarise(flights = n()) # all flights
```



#### Принципы Tidy Data и пакет tidyr

**Hadley Wickham** is a statistician from New Zealand who is currently Chief Scientist at RStudio.

He is best known for his development of packages for R, that implement logics of data visualisation and data transformation, including tidyr, ggplot2, plyr, dplyr, and reshape2.

Wickham's data analysis packages for R are collectively known as the 'tidyverse'.





#### Принципы Tidy Data и пакет tidyr

A dataset is a collection of **values**, usually either numbers (if quantitative) or strings (if qualitative).

Values are organized in two ways. Every value belongs to a **variable** and an **observation**.

A variable contains all values that measure the same underlying attribute across units.

An observation contains all values measured on the same unit across attributes.



#### Принципы Tidy Data и пакет tidyr

A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

#### In tidy data:

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

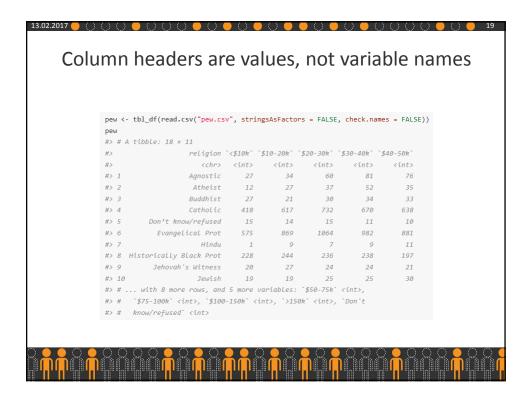


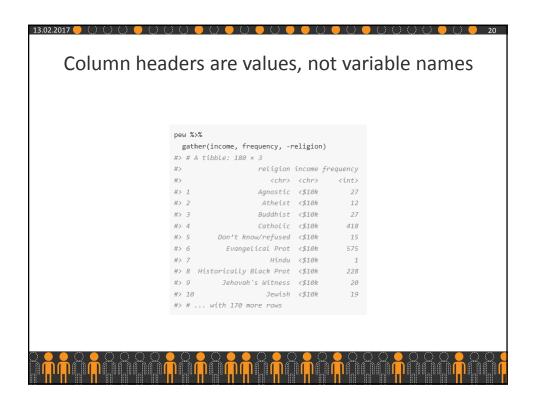
#### Принципы Tidy Data и пакет tidyr

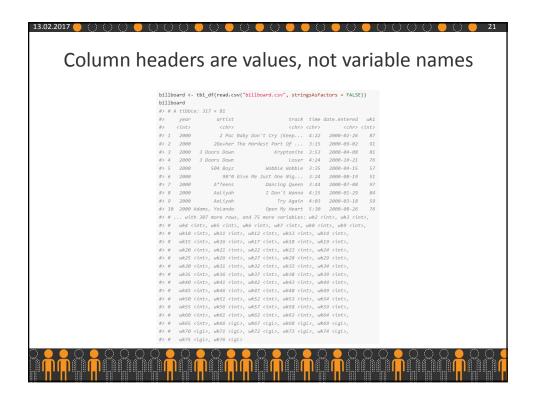
The five **most common problems** with messy datasets, along with their remedies:

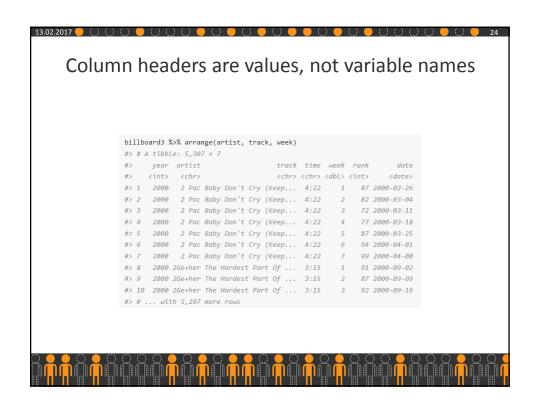
- Column headers are values, not variable names.
- Multiple variables are stored in one column.
- Variables are stored in both rows and columns.
- Multiple types of observational units are stored in the same table.
- A single observational unit is stored in multiple tables.

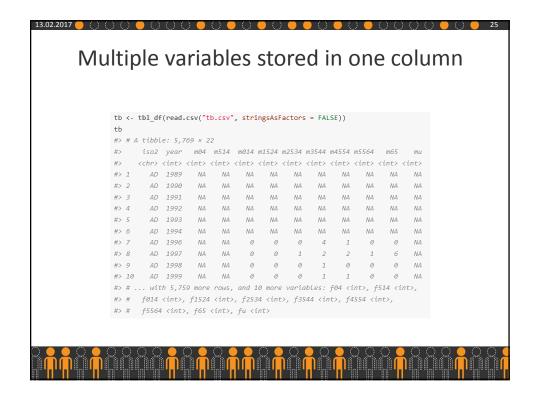


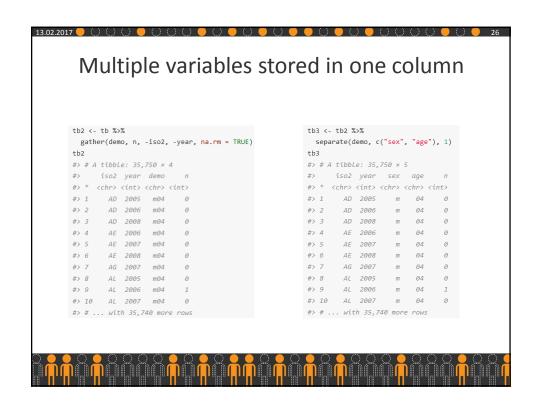


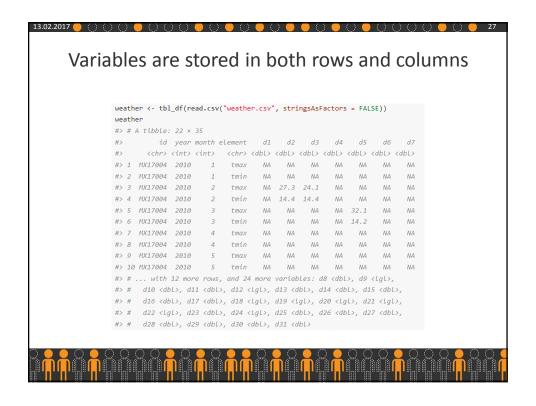


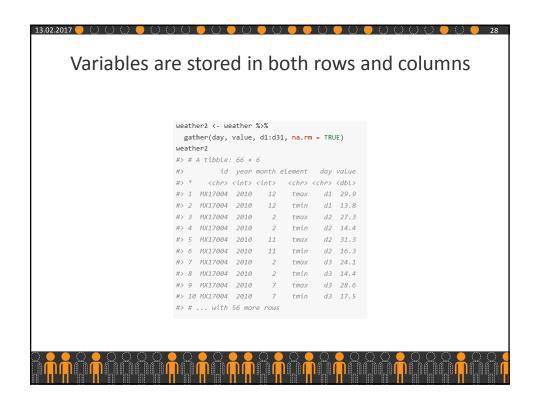


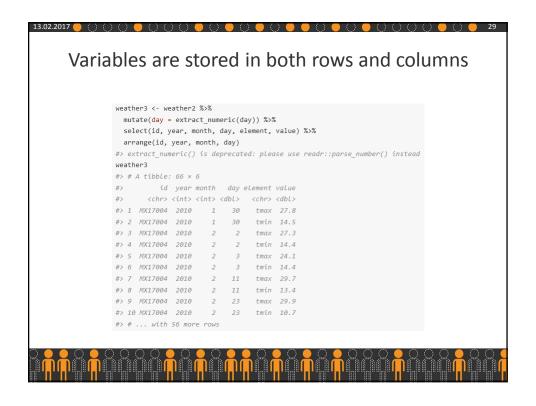


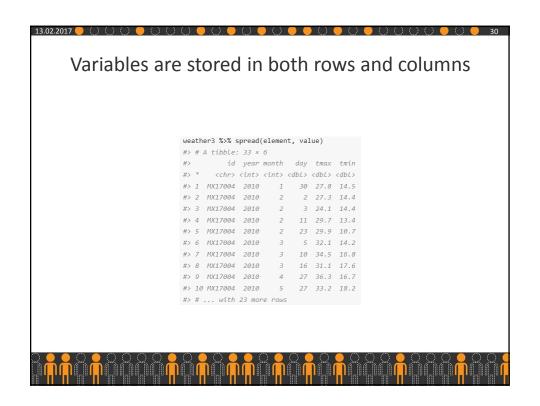


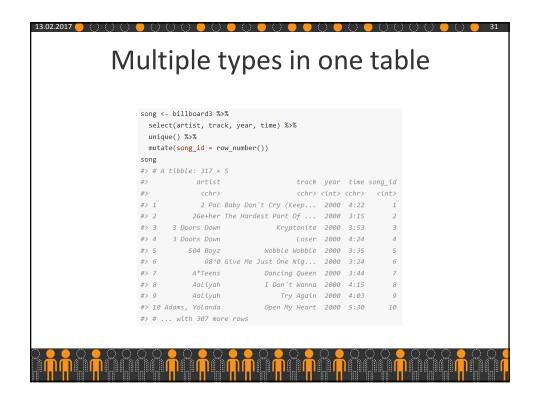












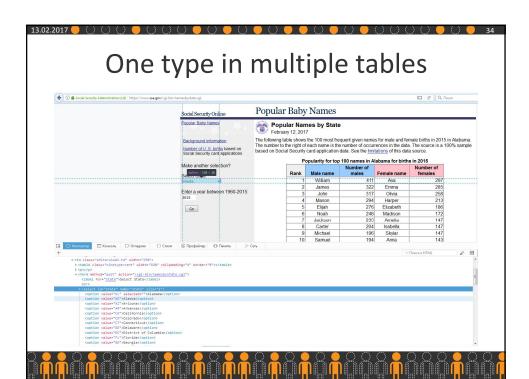
```
### Company of Company
```

## One type in multiple tables

It's also common to find data values about a single type of observational unit spread out over multiple tables or files. These tables and files are often split up by another variable, so that each represents a single year, person, or location.

As long as the format for individual records is consistent, this is an easy problem to fix:

- Read the files into a list of tables.
- For each table, add a new column that records the original file name (the file name is often the value of an important variable).
- Combine all tables into a single table.





## Основная литература

- Grolemund, G. R for Data Science [Electronic resource] / Garrett Grolemund, Hadley Wickham. – 2016. – Mode of access: http://r4ds.had.co.nz/index.html. – Date of access: 01.09.2016.
- Wickham, H. Tidy data [Electronic resource] / Hadley Wickham // The Journal of Statistical Software. – Vol. 59. – 2014. – Mode of access: http://vita.had.co.nz/papers/tidy-data.html. – Date of access: 01.09.2016.
- Wickham, H. Tidy data [Electronic resource]. 2017. Mode of access: ftp://cran.rproject.org/pub/R/web/packages/tidyr/vignettes/tidy-data.html. – Date of access: 01.02.2017.
- Ogurtsov, A. Tidy data (Перевод) [Electronic resource]. 2015. Mode of access: http://biostat-r.blogspot.com.by/2016/01/tidy-data.html. – Date of access: 01.02.2017.

