**ЛР 8 Пакет dplyr: продолжение**

Здесь мы продолжим знакомиться (как всегда на примерах) с популярным пакетом dplyr, который позволяет проводить различные манипуляции с таблицами (датафреймами) в R. В этом занятии рассмотрим функции **mutate, transmute, summarise**

**Порядок работы,**

1. Работу можно продолжать в том же проекте, назвав файл LR 8
2. Откроем литературу [1] и посмотрим видео 5 (18 мин.) и видео 6 (.
3. Наберем примеры ( в одном файле), приведенныев литературе [1], заменив их данные на свои (имя файла, путь к файлу, содержимое файлов.
4. С помощью Help ознакомимся с новыми функциями, которое здесь используются: **mutate, transmute, summarise**
5. В отчете привести код **своей** программы, скриншоты выполненных заданий и вывод.

**Литература**

1. Курс "Язык R для пользователей Excel". Урок 5 и 6 / <https://github.com/selesnow/r4excel_users>, 2020.
2. Введение в R / <https://rstudio-pubs-static.s3.amazonaws.com/33285_9ffa5341065d4bff9607fd6ccba75d98.html#->

2014

**Примечания**

1 Код примера

Урок 5 [1]

|  |  |
| --- | --- |
|  | # подключение пакетов |
|  | library(vroom) |
|  | library(dplyr) |
|  |  |
|  | # загрузка данных |
|  | ga\_data <- vroom("https://raw.githubusercontent.com/selesnow/publications/master/code\_example/from\_excel\_to\_r/lesson\_3/ga\_nowember.csv") |
|  |  |
|  | # mutate |
|  | # добавл¤ем новый столбец |
|  | ga\_data <- mutate(ga\_data, |
|  |  bounce\_rate = bounces / sessions) |
|  |  |
|  | ga\_data <- mutate(ga\_data, |
|  |  bounce\_rate = bounces / sessions, |
|  |  br\_group = if\_else(bounce\_rate > 0.6, "high\_br", "normal\_br")) |
|  |  |
|  | # примен¤ем преобразование к уже существующим столбцам |
|  | ga\_data %>% mutate\_if(is.character, toupper) |
|  |  |
|  | # преобразуем значени¤ существующих столбцов примен¤¤ регул¤рные выражени¤ |
|  | ga\_data %>% mutate\_at(vars(matches("s$")), sqrt ) |
|  |  |
|  | # transemute |
|  | # убираем все столбцы кроме преобразованных |
|  | transmute(ga\_data, |
|  |  bounce\_rate = bounces / sessions, |
|  |  date = format(date, "%d %B %Y"), |
|  |  source) |
|  |  |
|  | # преобразовываем столбцы по их признаку и удал¤ем все остальные |
|  | transmute\_if(ga\_data,  |
|  |  is.character, toupper) |

Урок 6 [1]

|  |
| --- |
| # загрузка данных |
|  | ga\_data <- vroom("https://raw.githubusercontent.com/selesnow/publications/master/code\_example/from\_excel\_to\_r/lesson\_3/ga\_nowember.csv") |
|  |  |
|  | # сгруппировать данные по дням |
|  | ga\_data %>% |
|  |  group\_by(date) %>% |
|  |  summarise(sessions = sum(sessions), |
|  |  bounces = sum(bounces)) |
|  |  |
|  | # другой вариант |
|  | ga\_data %>% |
|  |  group\_by(date) %>% |
|  |  summarise\_at(c("sessions", "bounces"),  |
|  |  sum) |
|  |  |
|  | # группировка по двум полям |
|  | ga\_data %>% |
|  |  group\_by(date, medium) %>% |
|  |  summarise\_at(c("sessions", "bounces"),  |
|  |  sum) |
|  |  |
|  | ga\_data %>% |
|  |  group\_by(date, medium) %>% |
|  |  summarise(ses = sum(sessions), |
|  |  bounc = sum(bounces)) |
|  |  |
|  | # сумма по всем числовым столбцам |
|  | ga\_data %>% |
|  |  group\_by(medium) %>% |
|  |  summarise\_if(is.numeric, mean) |
|  |  |
|  | # среднесуточное количество сеансов по каналам |
|  | ndays <- unique( ga\_data$date ) %>% length |
|  |  |
|  | ga\_data %>% |
|  |  group\_by(medium) %>% |
|  |  summarise(daily\_sessions = sum(sessions) / ndays) |
|  |  |
|  | # применяем сразу несколько агрегирующий функций |
|  | # сумма по всем числовым столбцам |
|  | result <- |
|  | ga\_data %>% |
|  |  group\_by(medium) %>% |
|  |  summarise\_if(is.numeric,  |
|  |  list( avg = mean, |
|  |  med = median, |
|  |  sum = sum, |
|  |  count = length, |
|  |  min = min, |
|  |  max = max)) |