Dyplr&Data.table

Simon Ress

17 4 2020

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
library(dplyr)
library(data.table)
library(gapminder)
```

#Import data

dplyr

```
data <- read.csv2("data.csv")
```

data.table

```
data.dt <- as.data.table(
                        read.csv2("data.csv")
                        )
```

Getting the number of distinct/unique values in a variable

dplyr

```
data %>%
  summarize(n = n_distinct(Country))
```

```
##    n
## 1 32
```

data.table

```
data.dt[,.(n = uniqueN(Country))]
```

```
##     n
## 1: 32
```

... grouped by another variable

dplyr

```
data %>%
  group_by(fed) %>%
  summarize(n = n_distinct(Country)) %>%
  head(4)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 2
##     fed     n
##   <int> <int>
## 1     0    15
## 2     1     2
## 3     2     5
## 4    NA    32
```

data.table

```
data.dt[,
        .(n = uniqueN(Country)),
        by = "fed"][1:4]
```

```
##    fed  n
## 1:   0 15
## 2:  NA 32
## 3:   1  2
## 4:   2  5
```

... filter by values of two other variables

dplyr

```
gapminder %>%
  filter(lifeExp >= 30, lifeExp <= 60) %>%
  group_by(continent) %>%
  summarize(n = n_distinct(country)) %>%
  head(4)
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 4 x 2
##   continent     n
##   <fct>     <int>
## 1 Africa       52
## 2 Americas     19
## 3 Asia         29
## 4 Europe        7
```

data.table

```
data.dt <- as.data.table(gapminder)
data.dt[lifeExp >= 30 & lifeExp <= 60,
        .(n = uniqueN(country)),
        by = "continent"][1:4]
```

```
##    continent  n
## 1:      Asia 29
## 2:    Europe  7
## 3:    Africa 52
## 4:  Americas 19
```

... filter by values of two other variables

base-R

```
aggregate(country ~ continent,
          data = subset(gapmin
          function(x) count=n_
```

```
##   continent country
## 1    Africa      52
## 2  Americas      25
## 3      Asia      33
## 4    Europe      30
## 5   Oceania       2
```

dplyr

```
gapminder %>%
  filter(lifeExp >= 30, lifeEx
  group_by(continent) %>%
  summarize(n = n_distinct(cou
```

```
## `summarise()` ungrouping ou
```

```
## # A tibble: 4 x 2
##   continent     n
##   <fct>     <int>
## 1 Africa       52
## 2 Americas     19
## 3 Asia         29
## 4 Europe        7
```

data.table

```
data.dt <- as.data.table(gapminder)
data.dt[lifeExp >= 30 & lifeExp <= 60,
        .(n = uniqueN(country)),
        by = "continent"]
```

```
## put (override with `.groups` argument)
## 1:      Asia 29
## 2:    Europe  7
## 3:    Africa 52
## 4:  Americas 19
```