Predavanje 08 – Odgovori na vprašanja

Iskanje in zamenjava podnizov v nizih znakov

Podnize v nizih iščemo in menjamo s funkcijama `grep()` in `gsub()`. Obe funkciji delujeta na vektorju nizov, zato ju lahko uporabimo tudi npr. na imenih stolpcev ali na nekem stolpcu data.frame-a. Obe funkciji podpirata ujemanje s pomočjo regularnih izrazov, kar nam omogoča napredna iskanja, kot je npr. določeno število alfanumeričnih znakov ali pa številke.

```
grep("ana", c("Anakonda", "Banana", "Ananas", "Kolo"))
```

```
## [1] 2 3
```

```
gsub("Predpona", "", c("Predpona1", "Predpona2"))
```

```
## [1] "1" "2"
```

```
gsub("Predpona", "Prefix", c("Predpona1", "Predpona2"))
```

```
## [1] "Prefix1" "Prefix2"
```

ggplot2 – boxplot

Diagram s *škatlami in brčicami* (angl. boxplot) je del nabora geom-ov ggplot2.

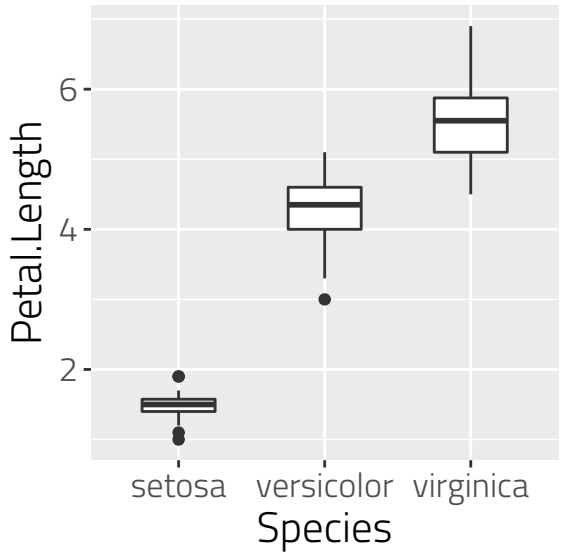
```
data("iris")
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Primer bomo izkoristili, da pokažemo, kako spremeniti pisavo v ggplot2:

```
library(ggplot2)
library(extrafont)
font_import("c:/Windows/Fonts/", pattern = "TitilliumWeb-Light", prompt = F)
loadfonts(device = "win")
```

```
ggplot(iris, aes(x = Species, y = Petal.Length)) +
  geom_boxplot(width = 0.5) +
  theme(text=element_text(size=16,  family="Titillium Web Light"))
```



ggplot2 – errorbar

Na statističnih grafih, ki vsebujejo opisne statistike, kot je npr. povprečje, pogosto prikažemo še negotovost v obliki standardnih odklonov ali standardnih napak. S knjižnico ggplot2 to storimo z uporabo geom-a `errorbar`. Pred tem moramo ustrezno pripraviti podatke tako, da dodamo še stolpec s spodnjo in zgornjo mejo napake. Če je napaka simetrična, potrebujemo le en stolpec.

```
data("mtcars")
head(mtcars)
```

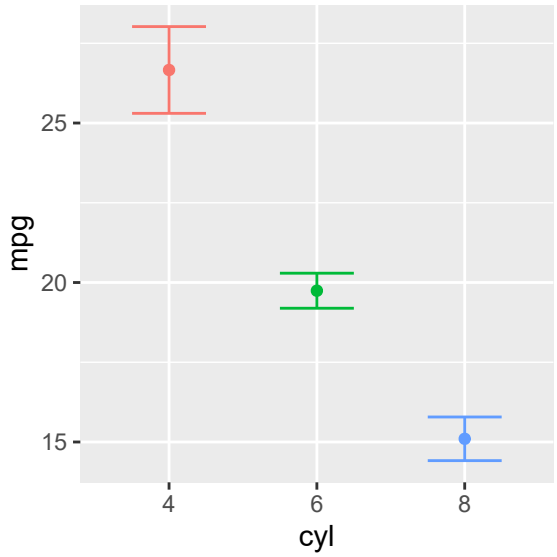
```
##                    mpg cyl disp  hp drat    wt  qsec vs am gear carb
## Mazda RX4         21.0   6  160 110 3.90 2.620 16.46  0  1    4    4
## Mazda RX4 Wag     21.0   6  160 110 3.90 2.875 17.02  0  1    4    4
## Datsun 710        22.8   4  108  93 3.85 2.320 18.61  1  1    4    1
## Hornet 4 Drive    21.4   6  258 110 3.08 3.215 19.44  1  0    3    1
## Hornet Sportabout 18.7   8  360 175 3.15 3.440 17.02  0  0    3    2
## Valiant           18.1   6  225 105 2.76 3.460 20.22  1  0    3    1
```

```
mus <- aggregate(mpg ~ cyl, mtcars, FUN = mean)
sds <- aggregate(mpg ~ cyl, mtcars, FUN = function(x) {sd(x) / sqrt(length(x))})
df  <- cbind(mus, SE = sds$mpg)
df$cyl <- as.character(df$cyl)

head(df)
```

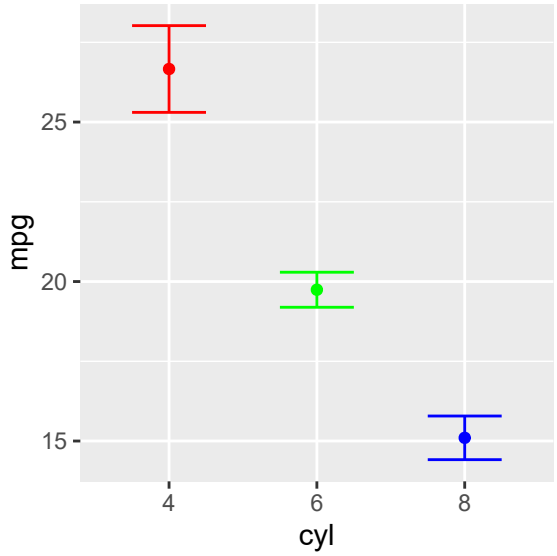
```
##   cyl      mpg        SE
## 1   4 26.66364 1.3597642
## 2   6 19.74286 0.5493967
## 3   8 15.10000 0.6842016
```

```
library(ggplot2)
ggplot(df, aes(x = cyl, y = mpg, colour = cyl)) +
  geom_point() +
  geom_errorbar(aes(ymin = mpg - SE, ymax = mpg + SE), width = 0.5) +
  theme(legend.position = "none")
```



Priložnost lahko izkoristimo se za uporabo lastne palete barv:

```
ggplot(df, aes(x = cyl, y = mpg, colour = cyl)) +
  geom_point() +
  geom_errorbar(aes(ymin = mpg - SE, ymax = mpg + SE), width = 0.5) +
  theme(legend.position = "none") +
  scale_colour_manual(values = c("#FF0000", "#00FF00", "#0000FF"))
```



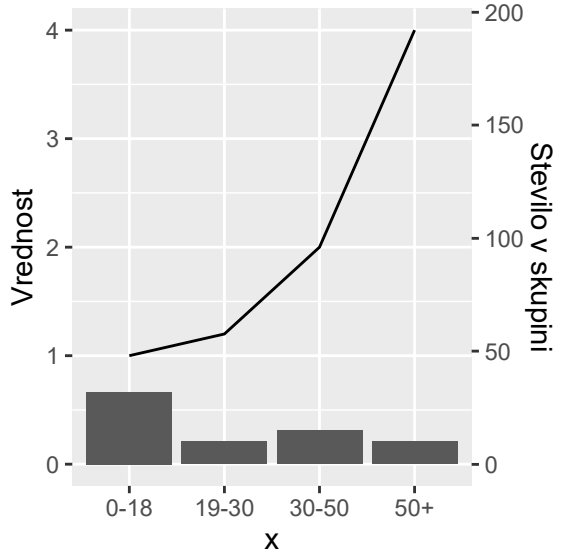
Kako v ggplot2 narediti graf z dvema y-osema?

Predlagamo, da ne uporabljate grafov z dvema y-osema: (https://blog.datawrapper.de/dualaxis/).

V kolikor zadeve ne morete rešiti drugače, pa se lahko poslužite sledečega trika:

```
x  <- c("0-18", "19-30", "30-50", "50+")
y1 <- c(1, 1.2, 2, 4)
y2 <- c(32, 10, 15, 10)
df <- data.frame(x = x, y1 = y1, y2 = y2)

ggplot(df) +
  geom_line(aes(x = x, y = y1, group = 1)) +
  geom_bar(aes(x = x, y = y2 / 8 / 6), stat = "identity") +
  scale_y_continuous(
    name = "Vrednost",
    sec.axis = sec_axis(trans=~. * 8 * 6, name="Stevilo v skupini")
  )
```



Kar smo naredili je, da smo drugo spremenljivko `y2` transformirali, da je bila na enakem razponu, kot prva spremenljivka `y` (deljenje z 8). Potem smo jo še nekoliko zmanjšali, da smo dobili škatle pod črto, za lepši izgled (deljenje s 6, ta korak bi lahko preskočili). Potem moramo samo še dodati drugo os, kjer definiramo obratno transformacijo z `sec_axis(trans=~. * 8 * 6)`. Torej, če želite imeti dve osi, je najprej potrebno drugo spremenljivko ustrezno transformirati in nato dodati obratno transformacijo v argument `sec.axis` funkcije `scale_y_continuous`.

Formatiranje nizov

Kadar želimo bolj strukturiran izpis nizov, si pomagamo s funkcijo `sprintf()` in izpisom s funkcijo `cat()`.

```
for (i in unique(mtcars$cyl)) {
  tmp <- mtcars[mtcars$cyl == i, ]
  mu  <- mean(tmp$mpg)
```

```
  med <- median(tmp$mpg)
  # C-jevsko formatiranje
  cat(sprintf("cyl = %d | mean mpg = %3.2f | median mpg = %3.2f\n", i, mu, med))
}
```

```
## cyl = 6 | mean mpg = 19.74 | median mpg = 19.70
## cyl = 4 | mean mpg = 26.66 | median mpg = 26.00
## cyl = 8 | mean mpg = 15.10 | median mpg = 15.20
```

Statistični testi

Večina klasičnih statističnih testov in modelov je vgrajenih že v osnovni R. Poglejmo si uporabo dveh izmed najbolj popularnih, t-testa in linearne regresije.

```
# modelirajmo porabo goriva, pri cemer kot neodvisne spremenljivke uporabimo:
# stevilo cilindrov, konjsko moc in tezo
lr <- lm(mpg ~ cyl + hp + wt, data = mtcars)
summary(lr)
```

```
##
## Call:
## lm(formula = mpg ~ cyl + hp + wt, data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9290 -1.5598 -0.5311  1.1850  5.8986
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.75179    1.78686  21.687  < 2e-16 ***
## cyl         -0.94162    0.55092  -1.709 0.098480 .
## hp          -0.01804    0.01188  -1.519 0.140015
## wt          -3.16697    0.74058  -4.276 0.000199 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.512 on 28 degrees of freedom
## Multiple R-squared:  0.8431, Adjusted R-squared:  0.8263
## F-statistic: 50.17 on 3 and 28 DF,  p-value: 2.184e-11
```

```
# t-test uporabimo za statisticno primerjavo pricakovane sirine listov
# dveh vrst perunike

x_vir <- iris$Sepal.Width[iris$Species == "virginica"]
x_ver <- iris$Sepal.Width[iris$Species == "versicolor"]

t.test(x_vir, x_ver)
```

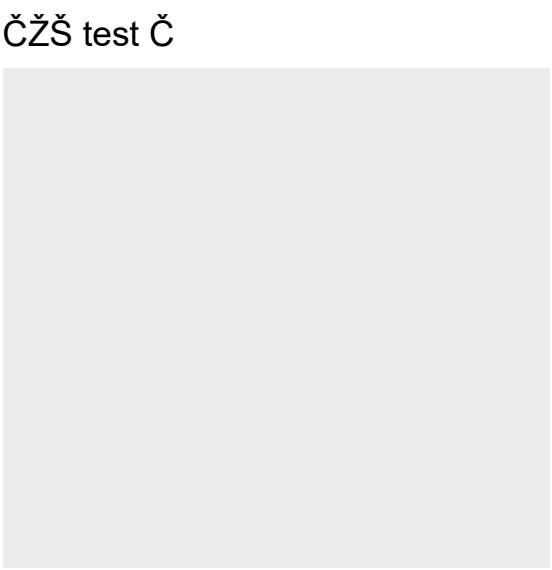
```
##
##  Welch Two Sample t-test
```

```
##
## data:  x_vir and x_ver
## t = 3.2058, df = 97.927, p-value = 0.001819
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.07771636 0.33028364
## sample estimates:
## mean of x mean of y
##     2.974     2.770
```

Č

```
Sys.setlocale("LC_ALL", "Slovenian")
```

```
options(encoding = "UTF-8")
g1 <- ggplot() + ggtitle("ČŽŠ test \u010C")
plot(g1)
```




```
ggsave("test.jpg", g1) # device = cairo_pdf for pdf
```

```
## Saving 3 x 3 in image
```