Vizualizacija v R

Namen

- Eksploratorna analiza (identificiranje hipotez)
  - relativna primerjava
  - identificiranje vzročnosti, mehanizma vpliva, razlage
  - opazovanje več kot dveh spremenljivk
- Bolj jasna in prepričljiva predstavitev podatkov

Sistemi za risanje

- base
  - osnoven (star) sistem
  - risanje na platno
- lattice
  - konstruiranje funkcije, ki izvede izris
- ggplot2
  - moderen pristop na osnovi določenega teoretičnega okvira
- ggvis
  - moderna nadgradnja ggvis (v izgradnji), ki uporablja spletne tehnologije

Sistem `ggplot2`

- Avtor Hadley Wickham
- Moderen sistem izgrajen na praktični "teoriji" o grafiki
- Vizualizacija je preslikava iz podatkov v 2D (ali 3D) prostor z izbranim koordinatnim sistemom, in sicer v:
  - estetske atribute (barva, oblika, velikost) in
  - geometrijske objekte (točke, črte, stolpiče)
- Pametno premišljene privzete nastavitve

Komponente `ggplot2`

- vhodni podatki so vedno v tabelah (`data.frame`)
- `aes` - estestke preslikave v barvo, obliko in velikost
- `geoms` - geometrijski objekti (točke, črte, liki)
- `facets` - izrisi pogojno na vrednosti faktorjev
- `stats` - statistične transformacije (delitev v koše, kvantili, glajenje)
- `scales` - lestvice
- koordinatni sistem

Primeri

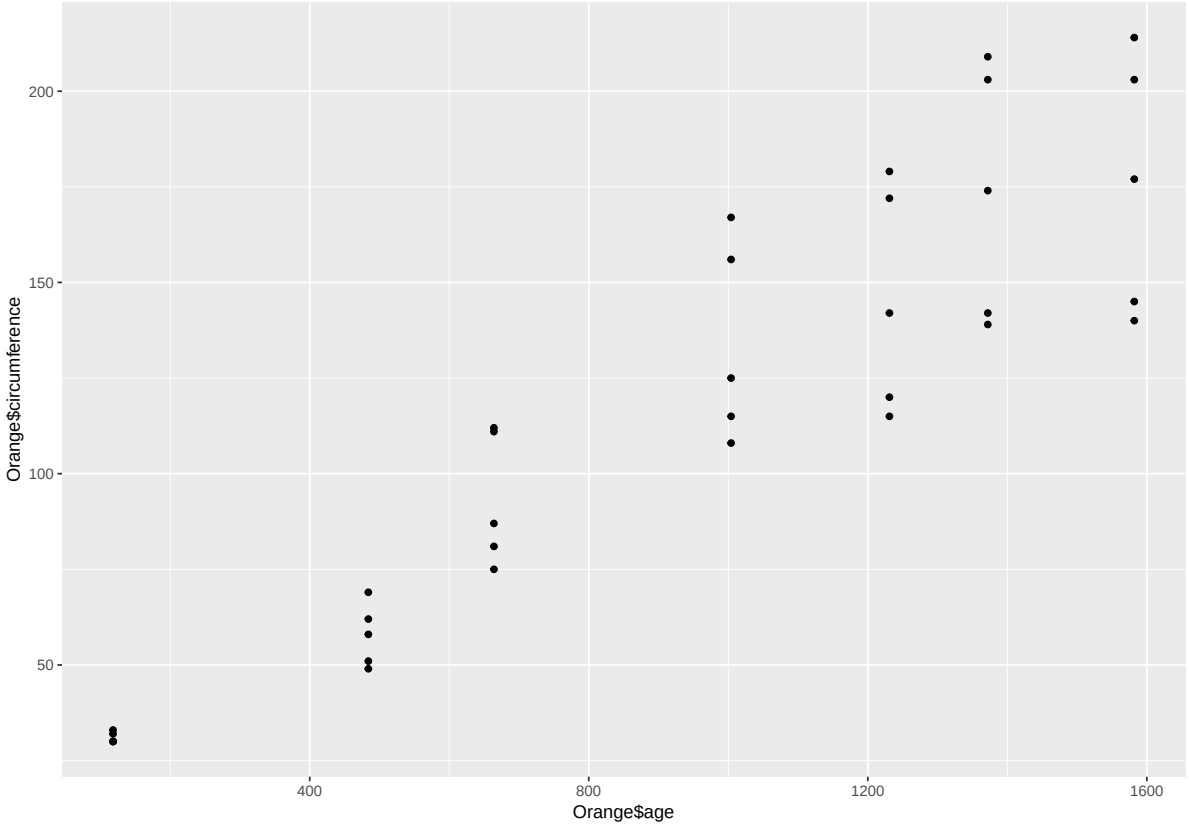
```
require(ggplot2)
require(dplyr)
```

```
head(Orange)
```

```
##   Tree  age circumference
## 1    1  118            30
## 2    1  484            58
## 3    1  664            87
## 4    1 1004           115
## 5    1 1231           120
## 6    1 1372           142
```

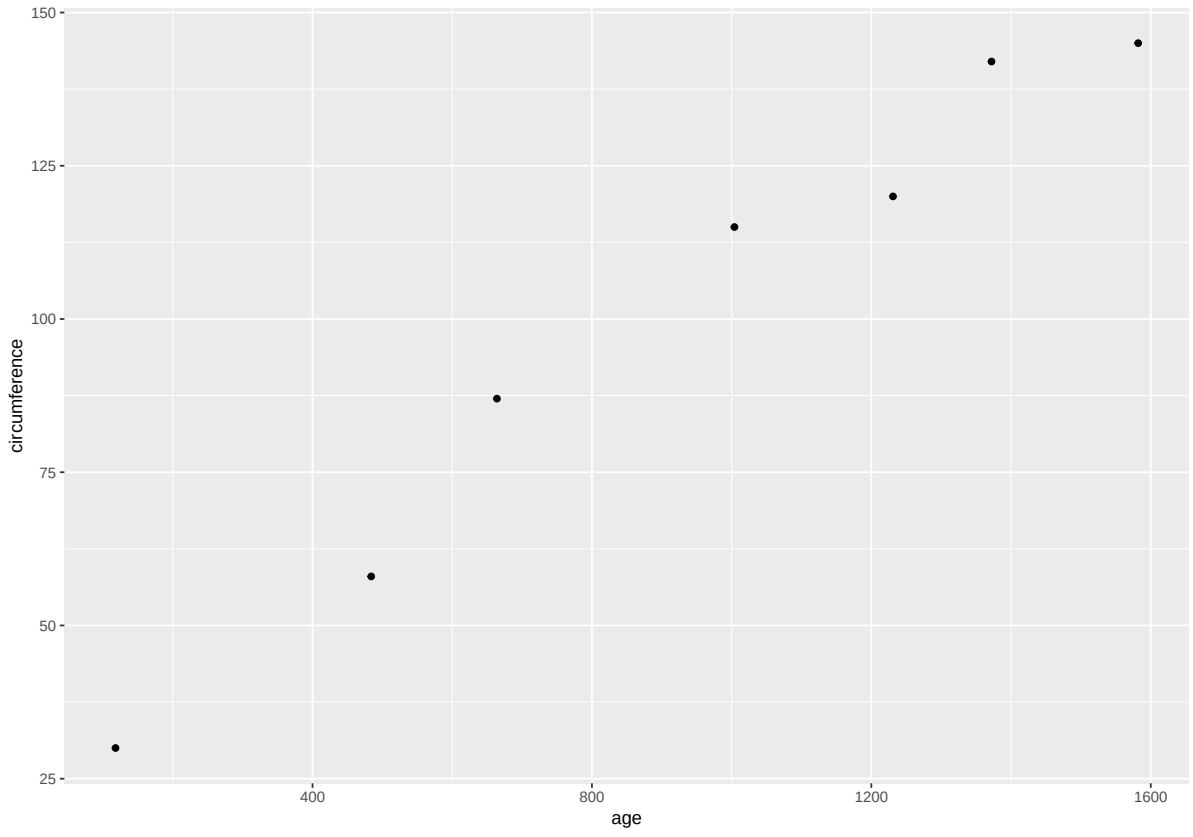
Primeri

```
ggplot(data=Orange, aes(x=Orange$age,
  y=Orange$circumference)) + geom_point()
```



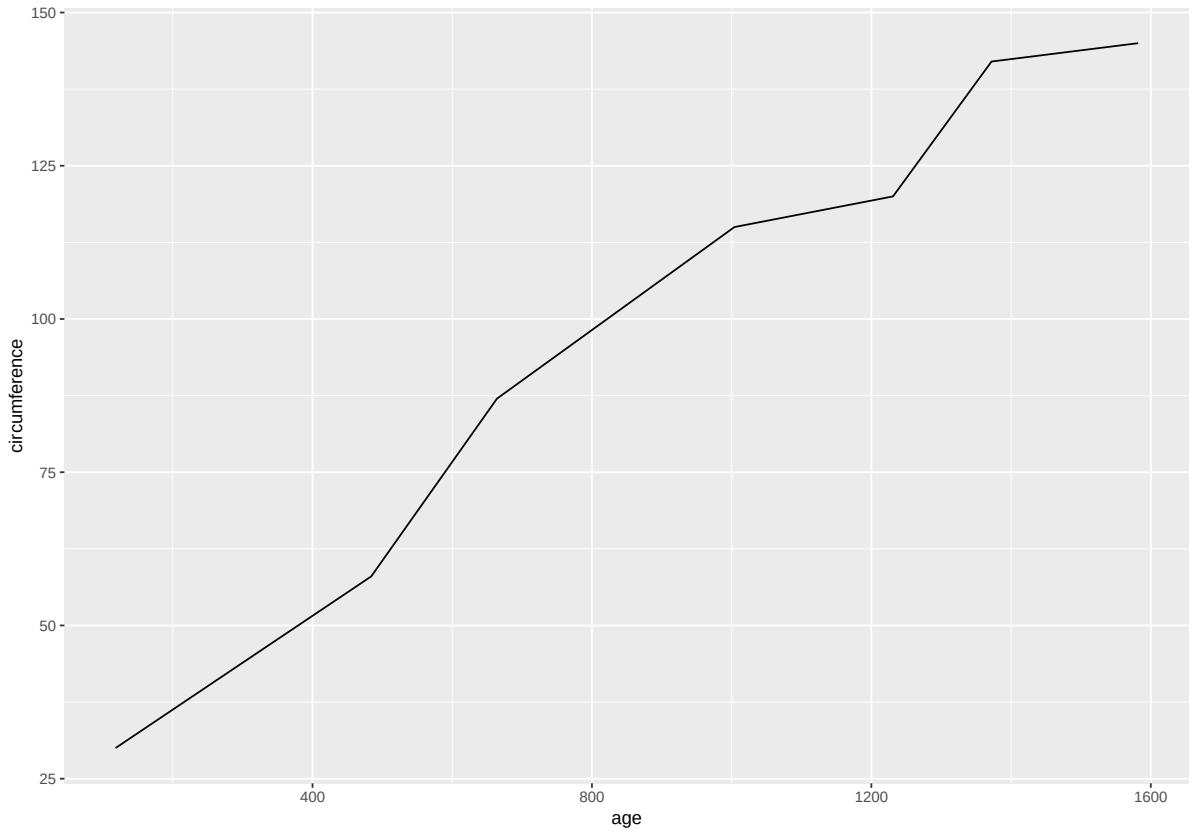
Primeri

```
ggplot(data=Orange %>% filter(Tree==1),
aes(x=age, y=circumference)) + geom_point()
```



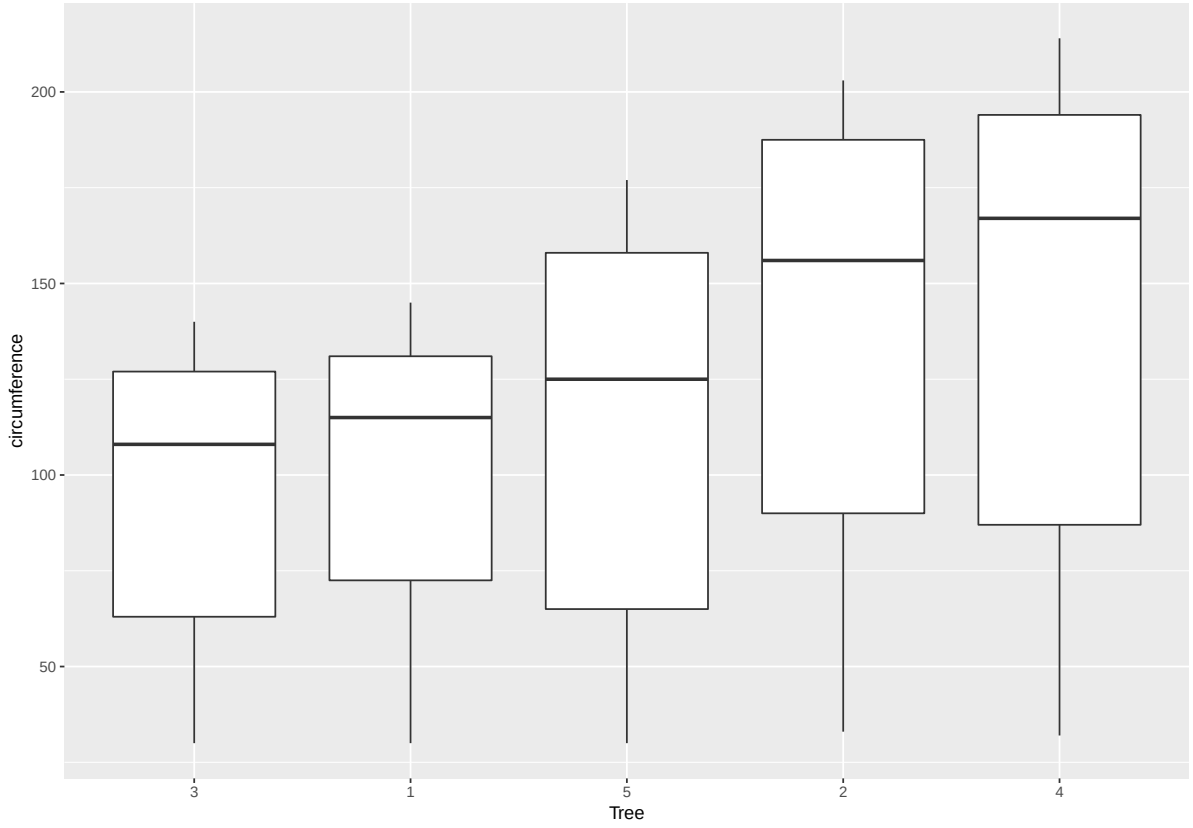
Primeri

```
ggplot(data=Orange %>% filter(Tree==1),
  aes(x=age, y=circumference)) + geom_line()
```



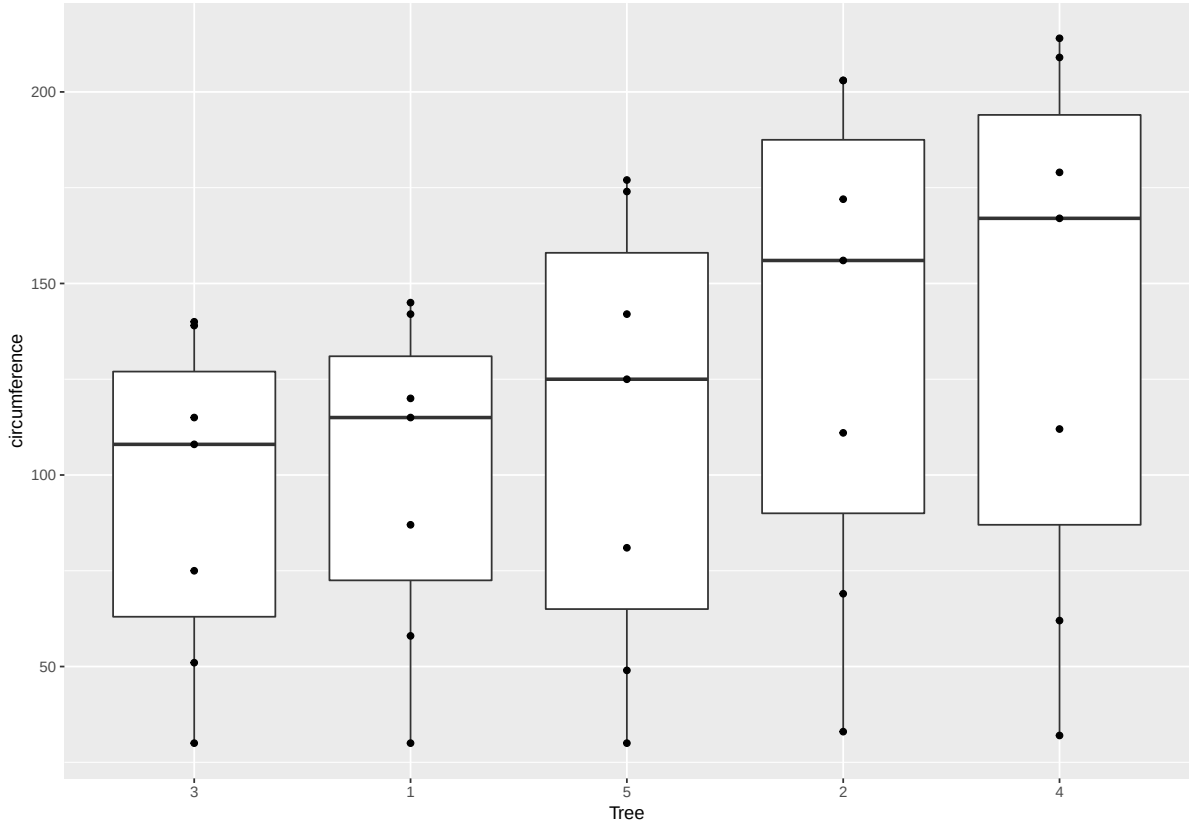
Primeri

```
ggplot(data=Orange, aes(x=Tree, y=circumference)) +
    geom_boxplot()
```



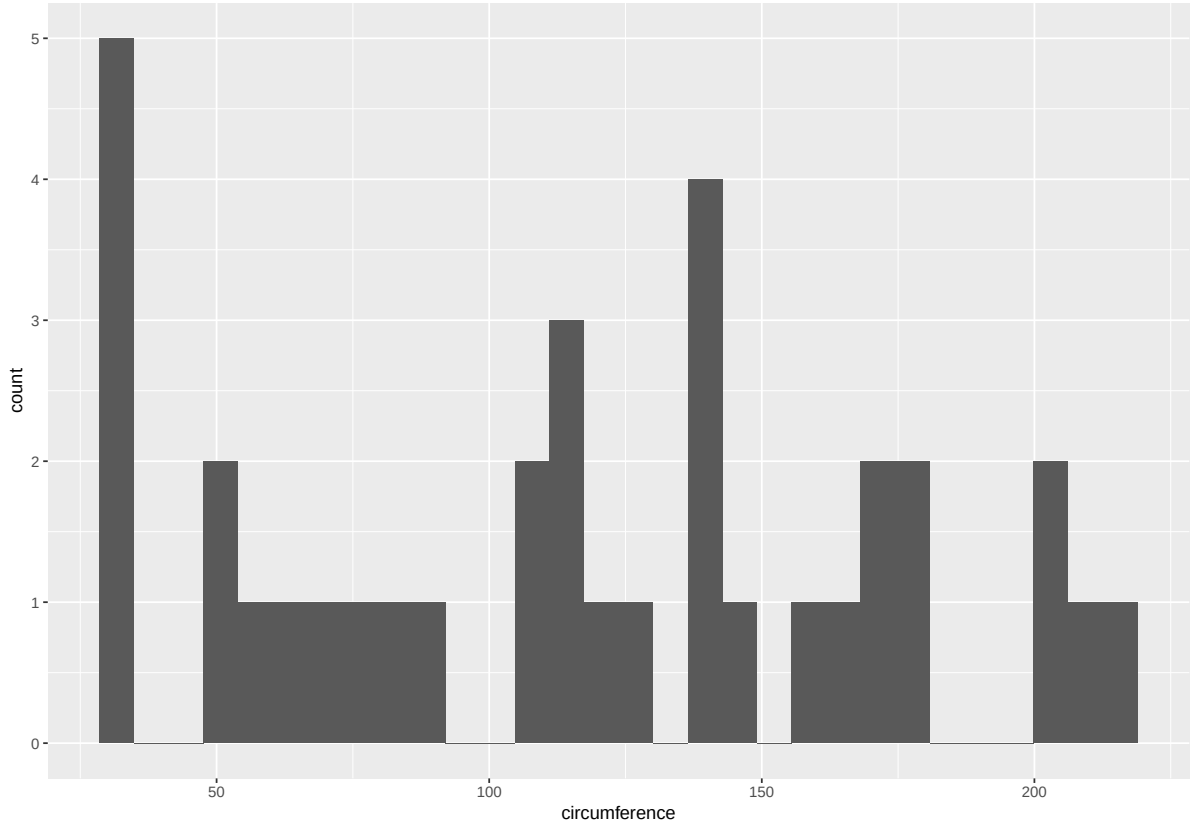
Primeri

```
ggplot(data=Orange, aes(x=Tree, y=circumference)) +
    geom_boxplot() + geom_point()
```



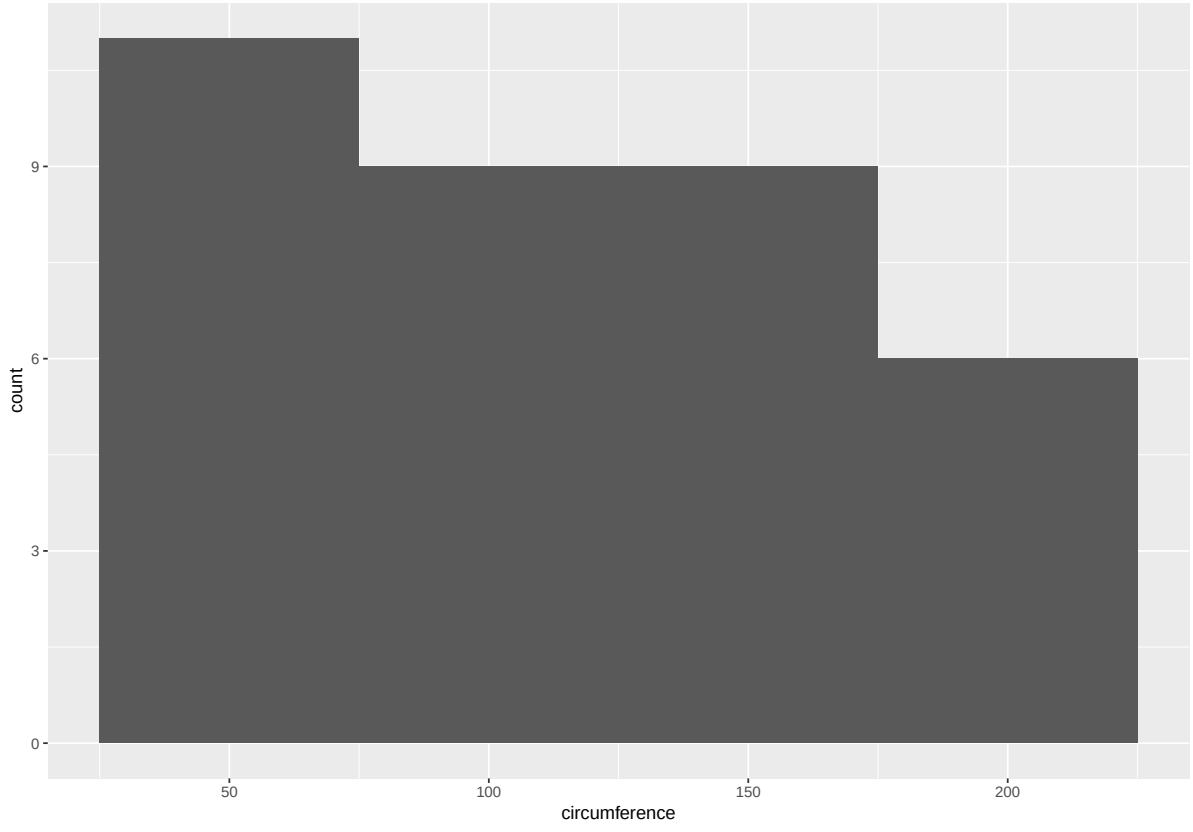
Primeri

```
ggplot(data=Orange, aes(x=circumference)) +
  geom_histogram()
```



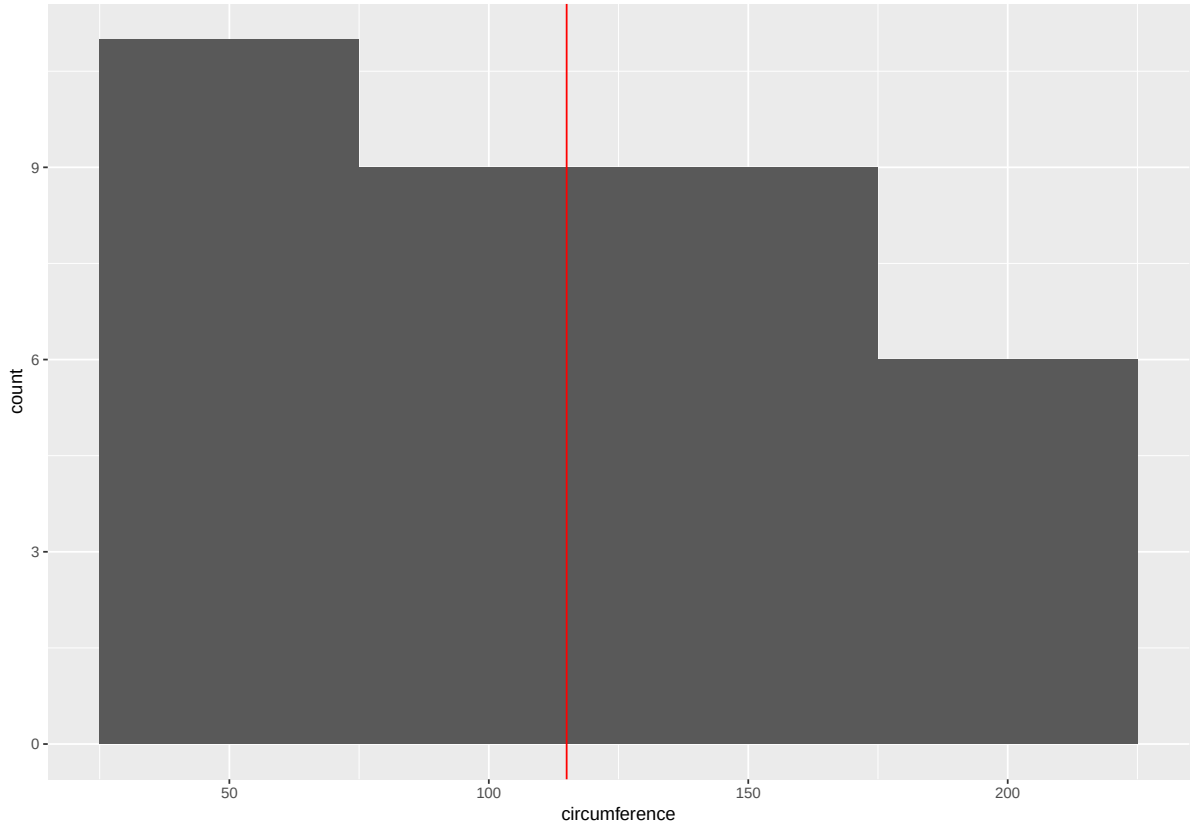
Primeri

```
ggplot(data=Orange, aes(x=circumference)) +
  geom_histogram(binwidth=50)
```



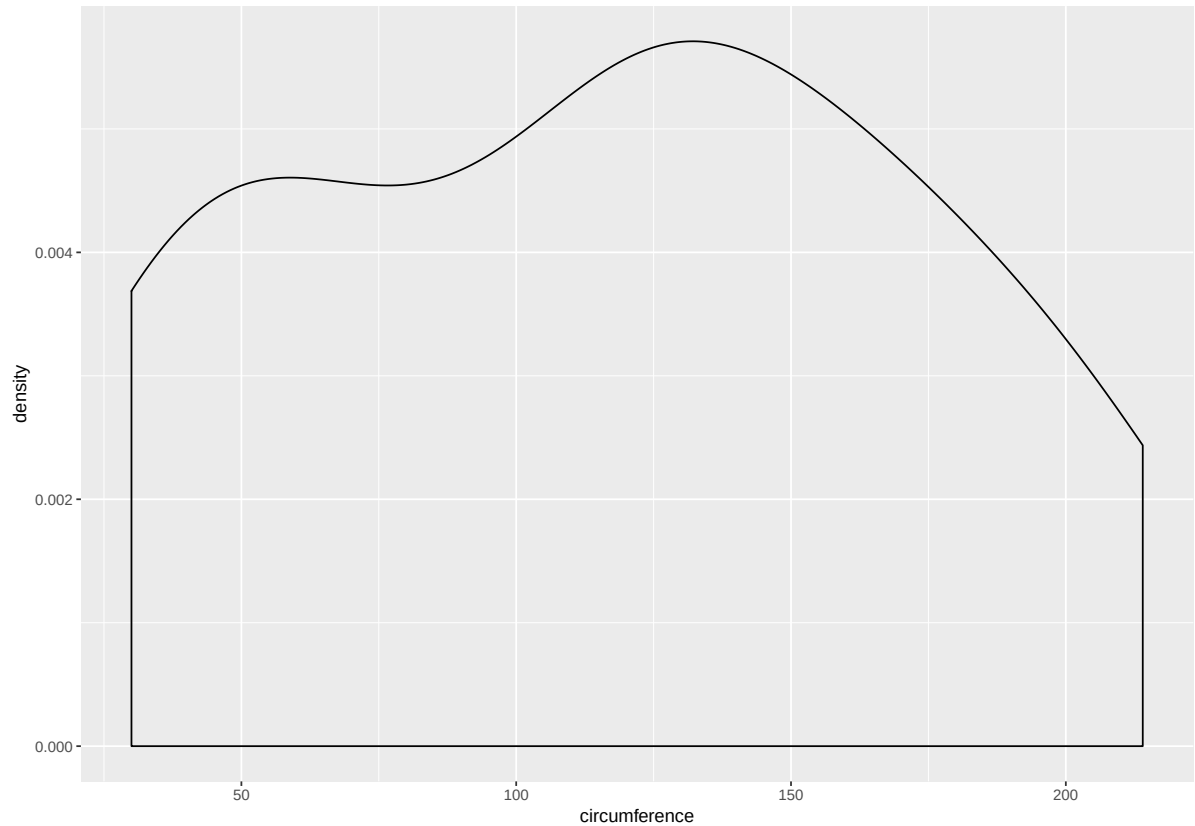
Primeri

```
ggplot(data=Orange, aes(x=circumference)) +
  geom_histogram(binwidth=50) +
  geom_vline(xintercept=median(Orange$circumference),
             col="red")
```



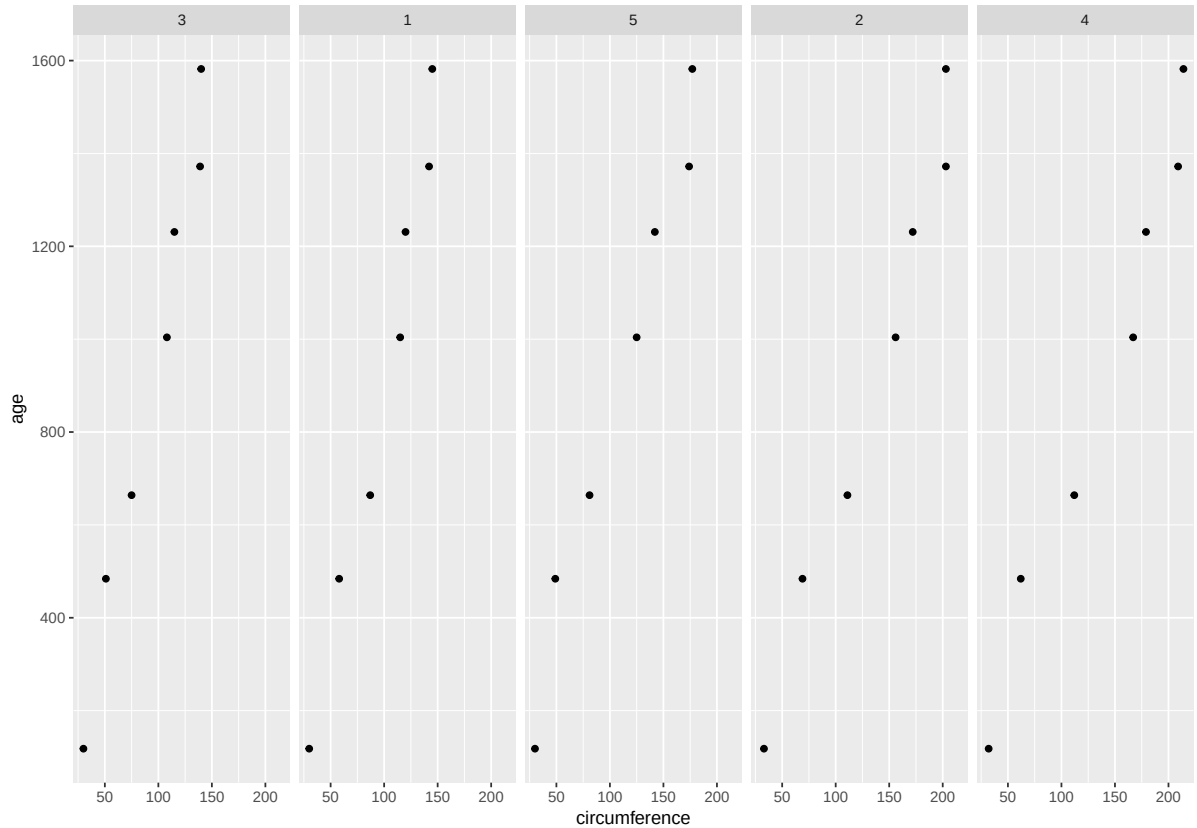
Primeri

```
ggplot(data=Orange, aes(x=circumference)) +
  geom_density()
```



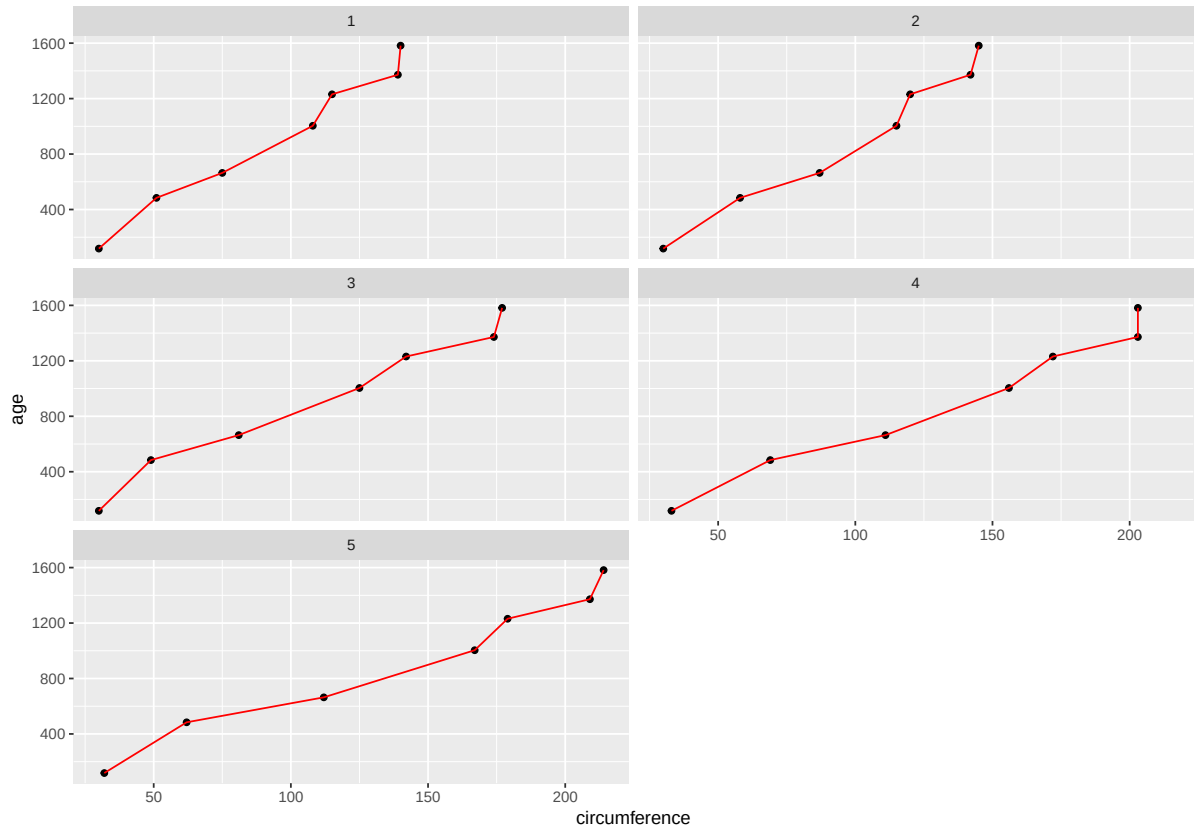
Primeri

```
ggplot(data=Orange, aes(x=circumference, y=age)) +
  geom_point() + facet_grid(~Tree)
```



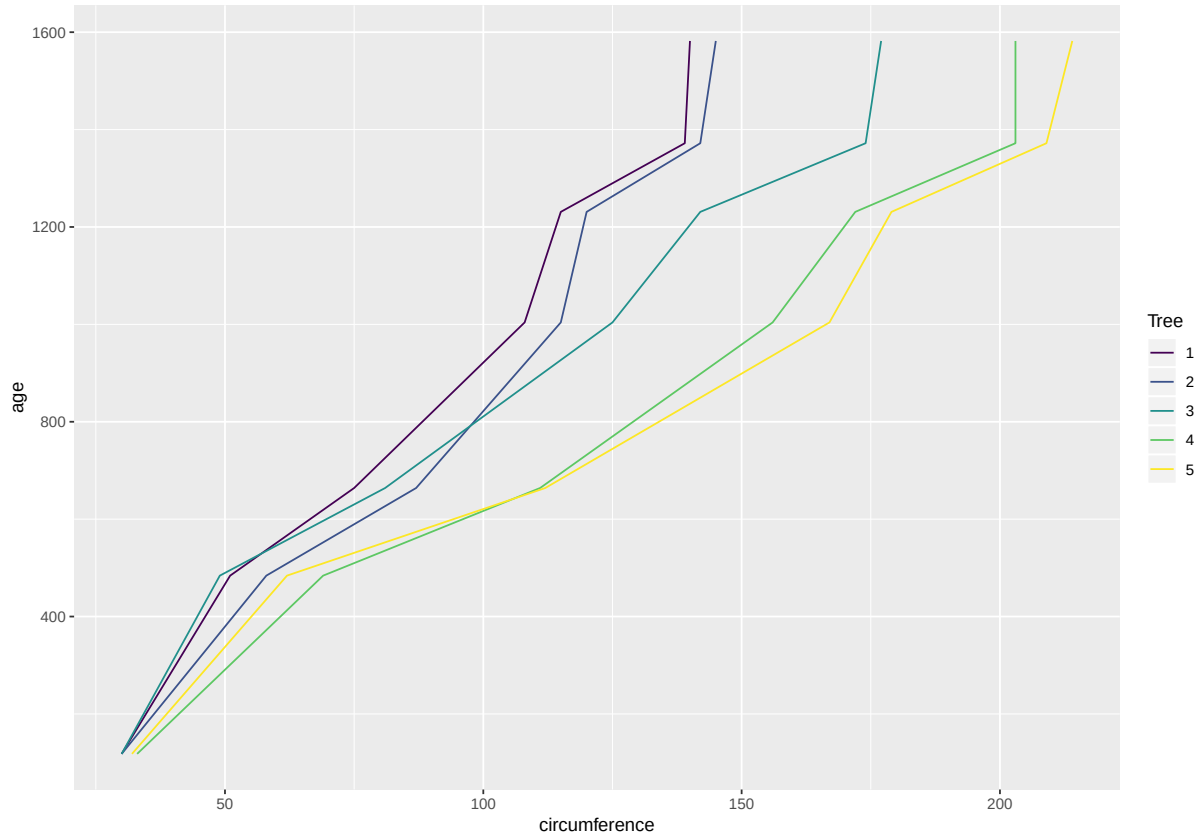
Primeri

```
levels(Orange$Tree) <- sort(levels(Orange$Tree))
ggplot(data=Orange, aes(x=circumference, y=age)) +
  geom_point() +
  geom_line(col="red") +
  facet_wrap(~Tree, ncol=2)
```



Primeri

```
ggplot(data=Orange, aes(x=circumference, y=age, col=Tree))
  geom_line()
```



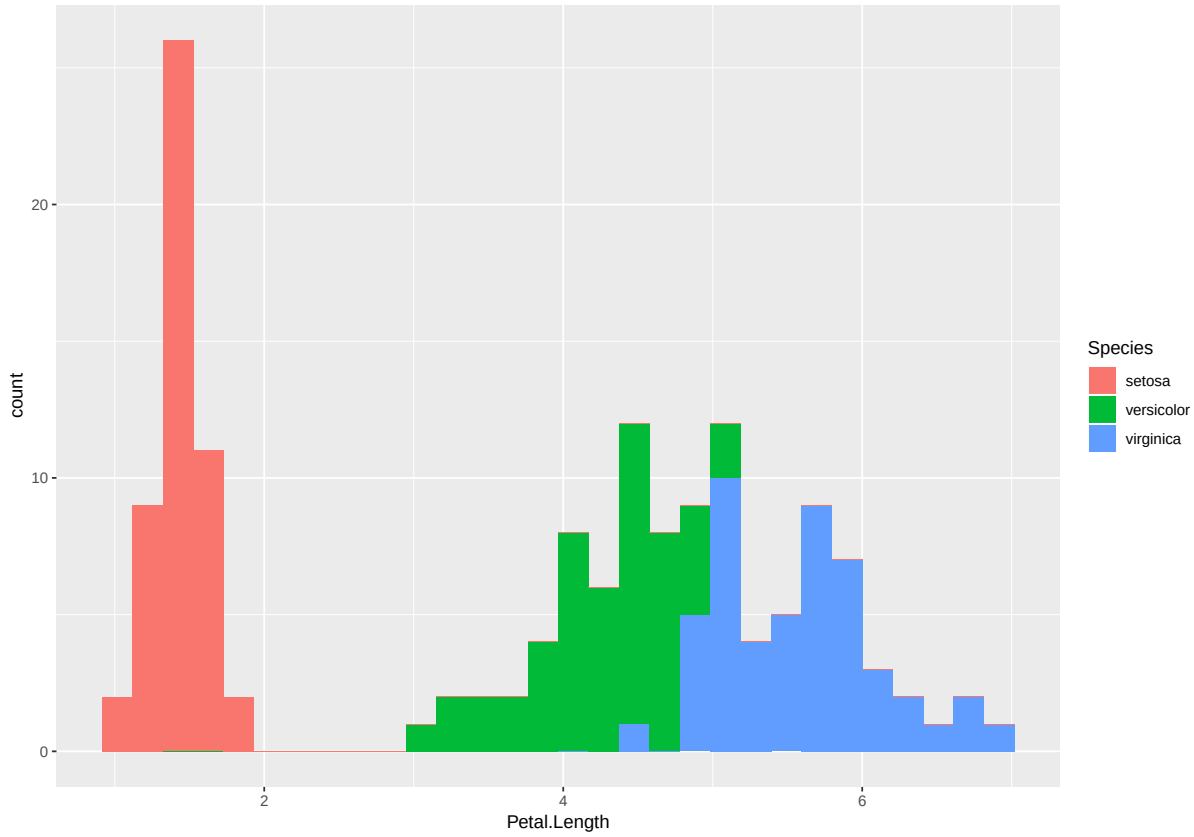
Primeri

```
head(iris)
```

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

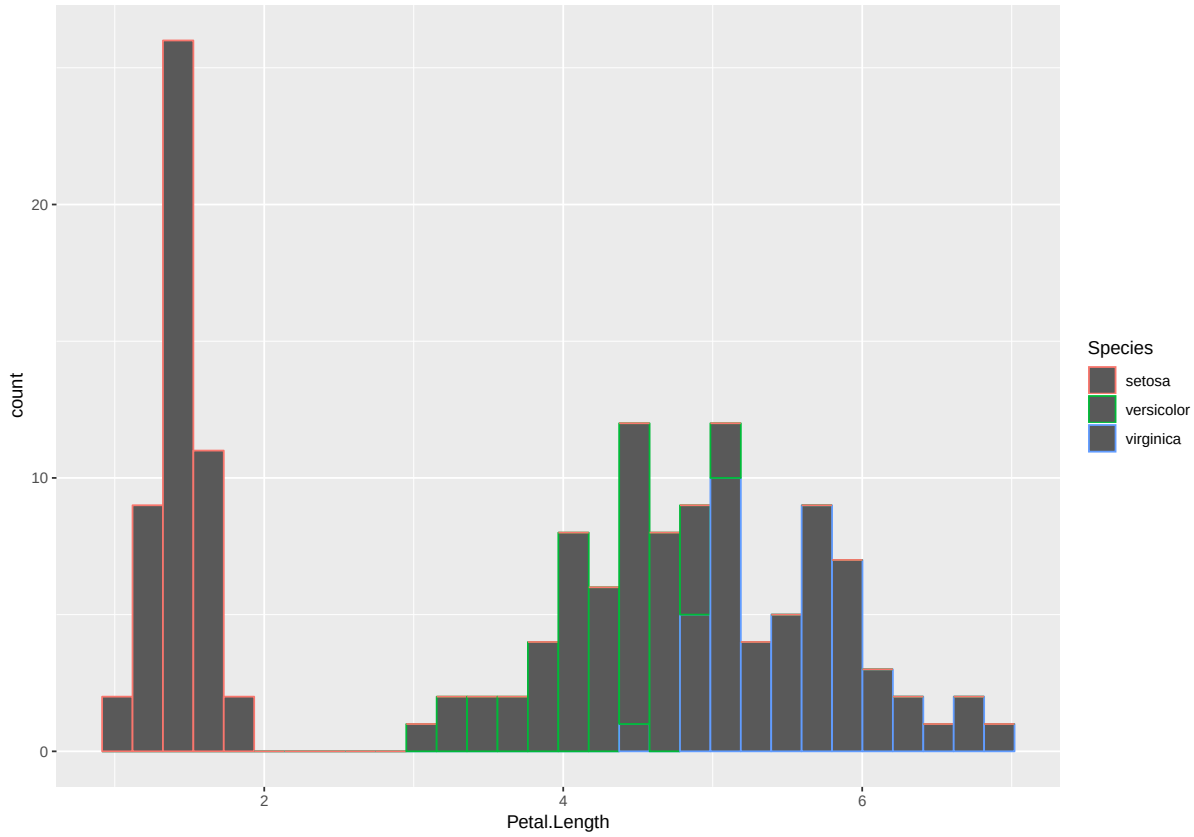
Primeri

```
ggplot(data=iris, aes(x=Petal.Length, fill=Species)) +
  geom_histogram()
```



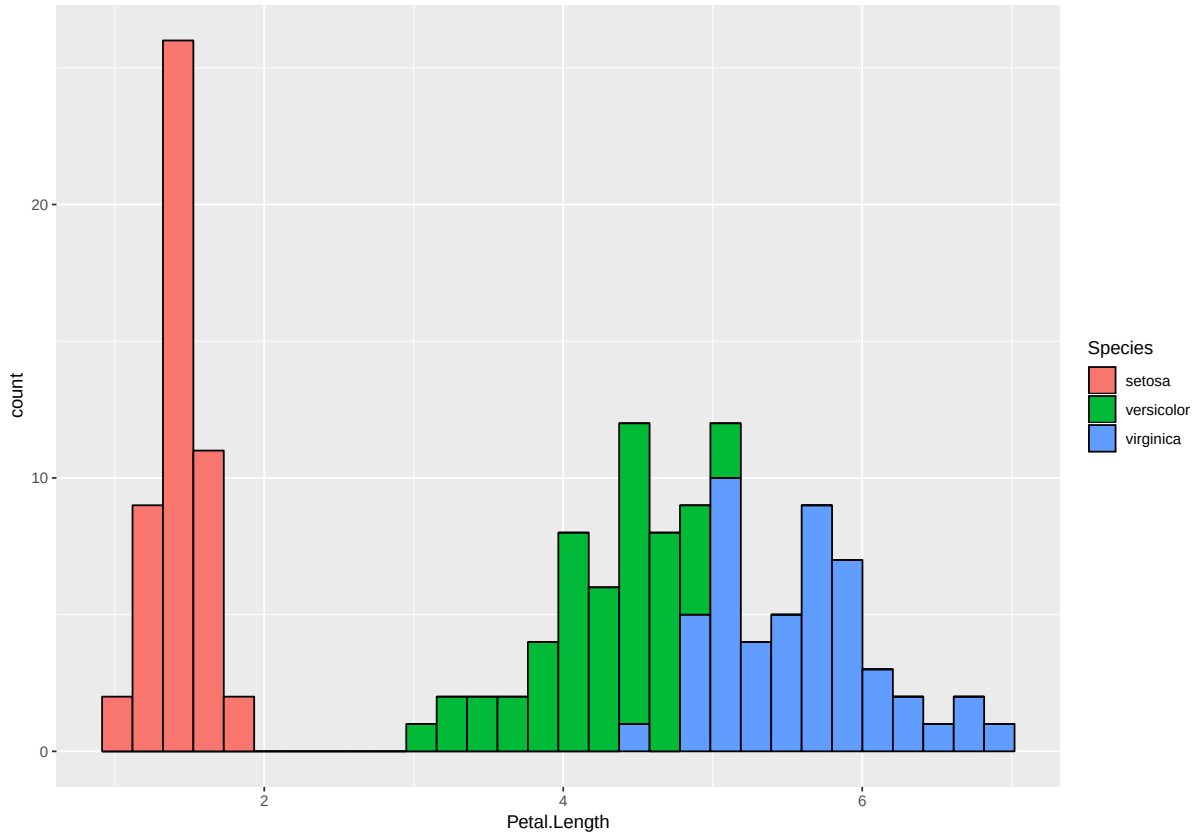
Primeri

```
ggplot(data=iris, aes(x=Petal.Length, color=Species)) +
  geom_histogram()
```



Primeri

```
ggplot(data=iris, aes(x=Petal.Length, fill=Species)) +
  geom_histogram(color="black")
```



Primeri

```
require(ggplot2movies)
head(movies)
```

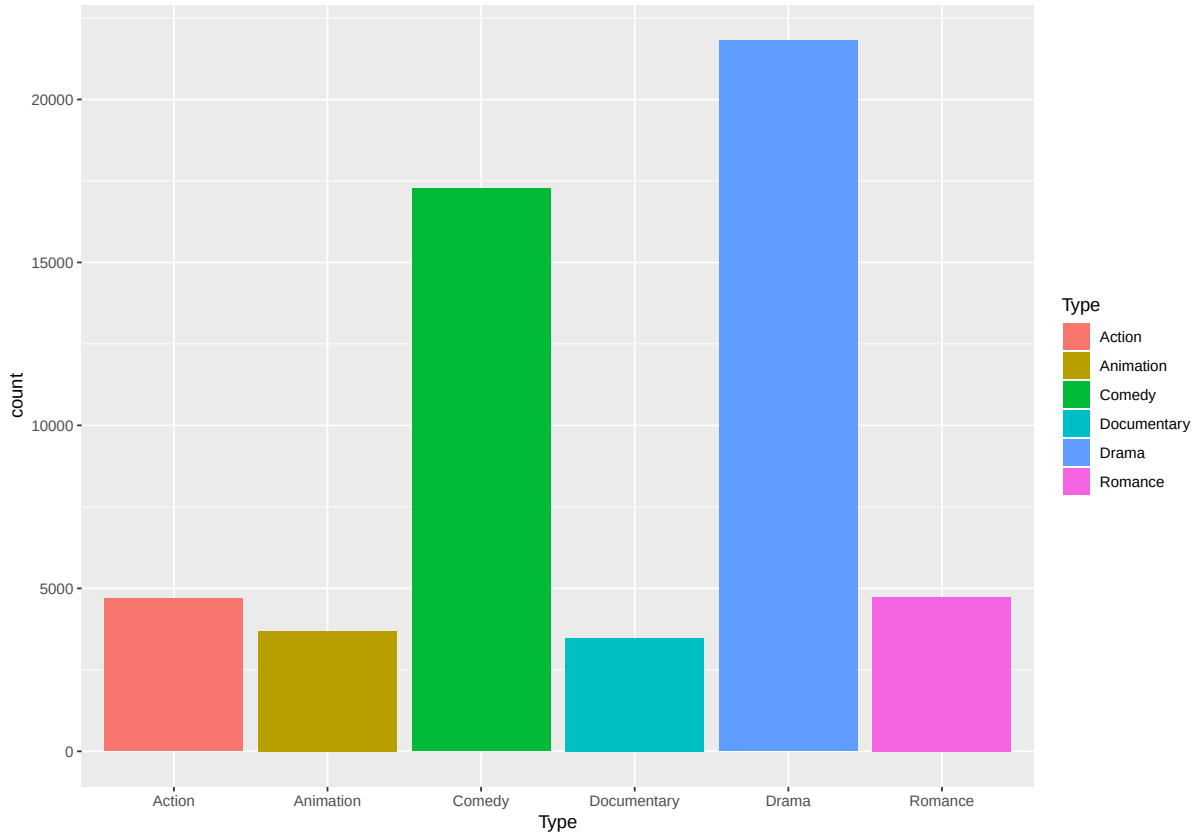
```
## # A tibble: 6 x 24
##   title  year length budget rating votes    r1    r2
##   <chr> <int>  <int>  <int>  <dbl> <int> <dbl> <dbl> <db
## 1 $      1971    121     NA    6.4   348   4.5   4.5   4
## 2 $100~  1939     71     NA    6      20   0    14.5   4
## 3 $21 ~  1941      7     NA    8.2     5   0     0     0
## 4 $40,~  1996     70     NA    8.2     6  14.5   0     0
## 5 $50,~  1975     71     NA    3.4    17  24.5   4.5   0
## 6 $pent  2000     91     NA    4.3    45   4.5   4.5   4
## # ... with 13 more variables: r6 <dbl>, r7 <dbl>, r8 <db
## #   r10 <dbl>, mpaa <chr>, Action <int>, Animation <int>
## #   Drama <int>, Documentary <int>, Romance <int>, Short
```

Primeri

```
tipi = names(movies)[18:23]
seznam = list()
for (i in 1:length(tipi)) {
    tip = tipi[[i]]
    seznam[[i]] <- movies %>%
        filter_(paste(tip, "==", 1)) %>%
        select(Budget=budget, Short, Year=year) %>%
        mutate(Type=tip)
}
myMovies <- do.call(rbind, seznam)
```

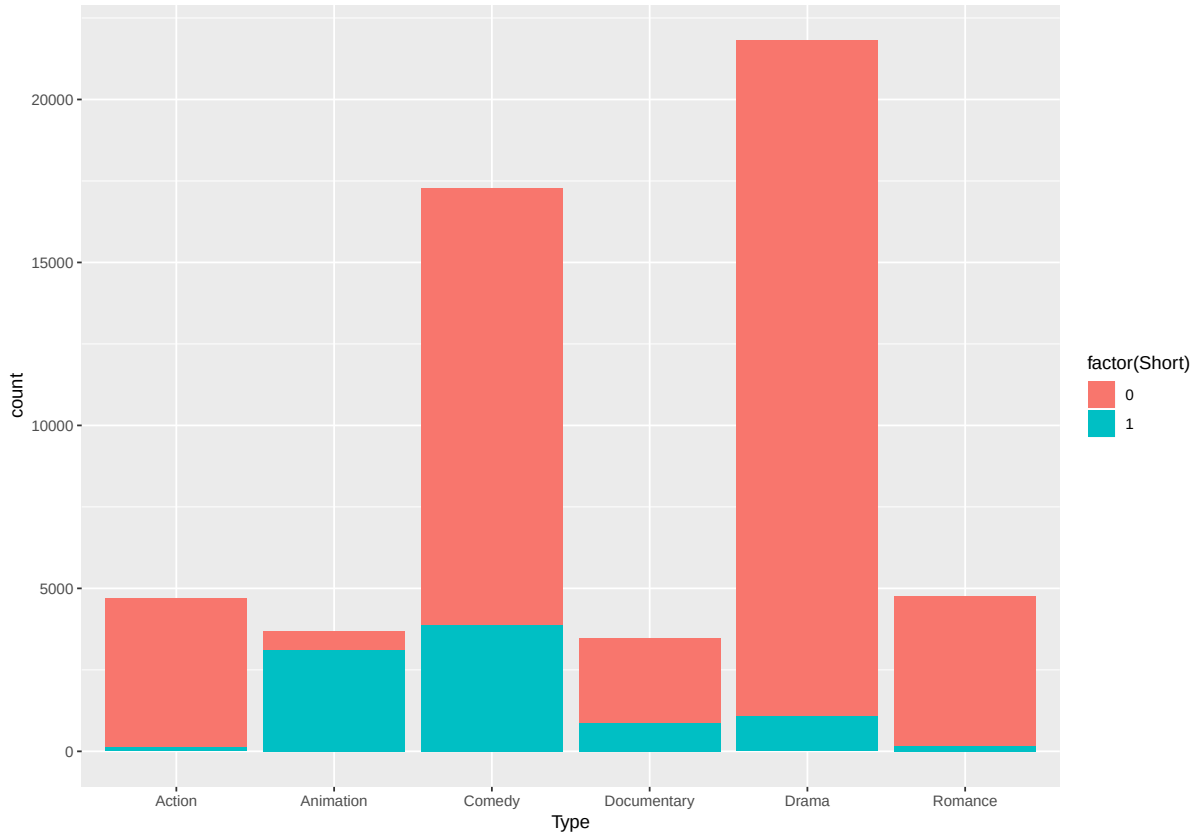
Primeri

```
ggplot(data=myMovies, aes(x=Type, fill=Type)) +
  geom_bar()
```



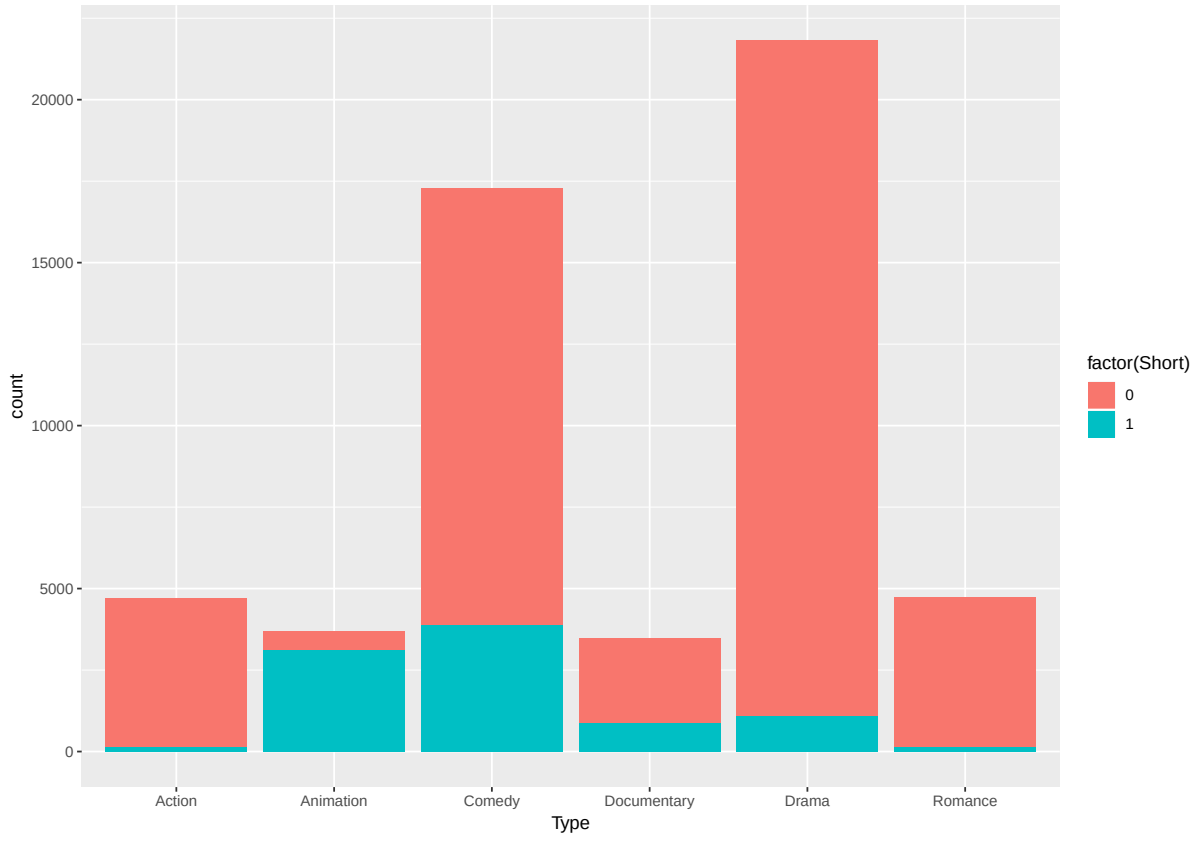
Primeri

```
ggplot(data=myMovies, aes(x=Type, fill=factor(Short))) +
  geom_bar()
```



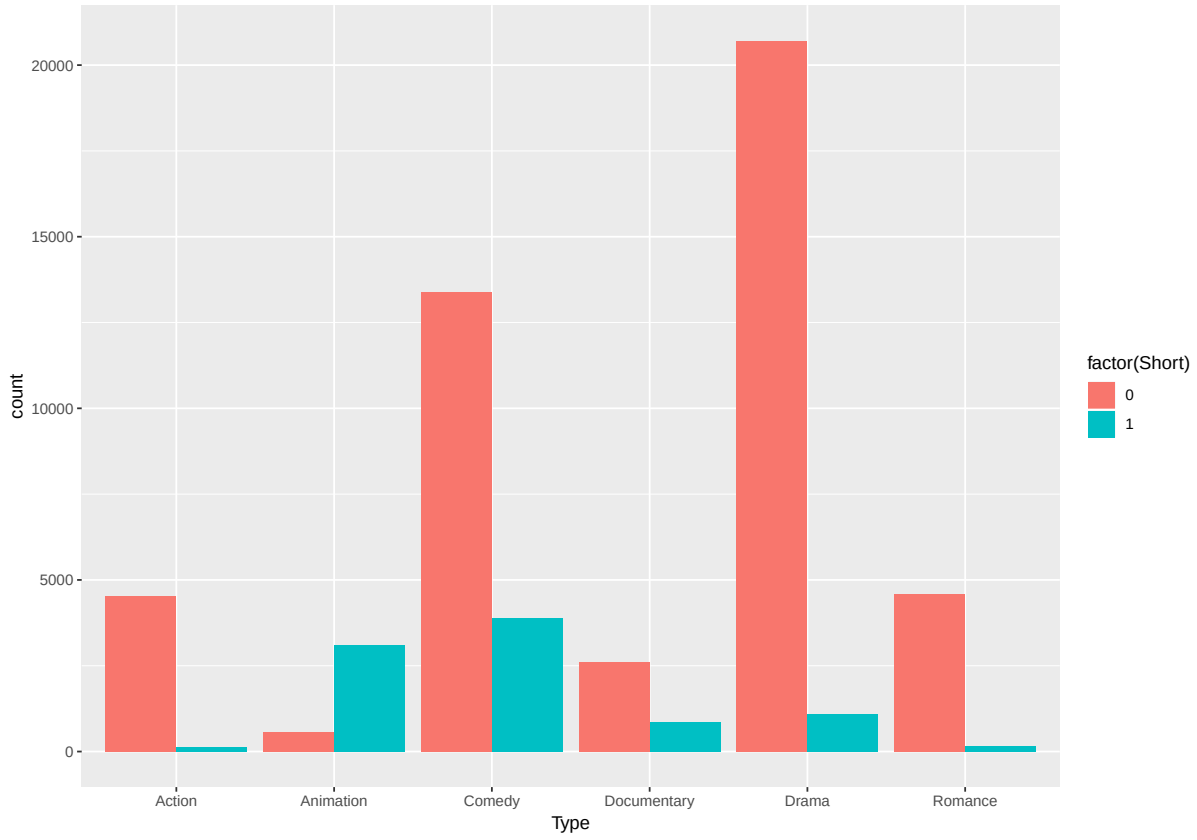
Primeri

```
ggplot(data=myMovies, aes(x=Type, fill=factor(Short))) +
  geom_bar(position="stack")
```

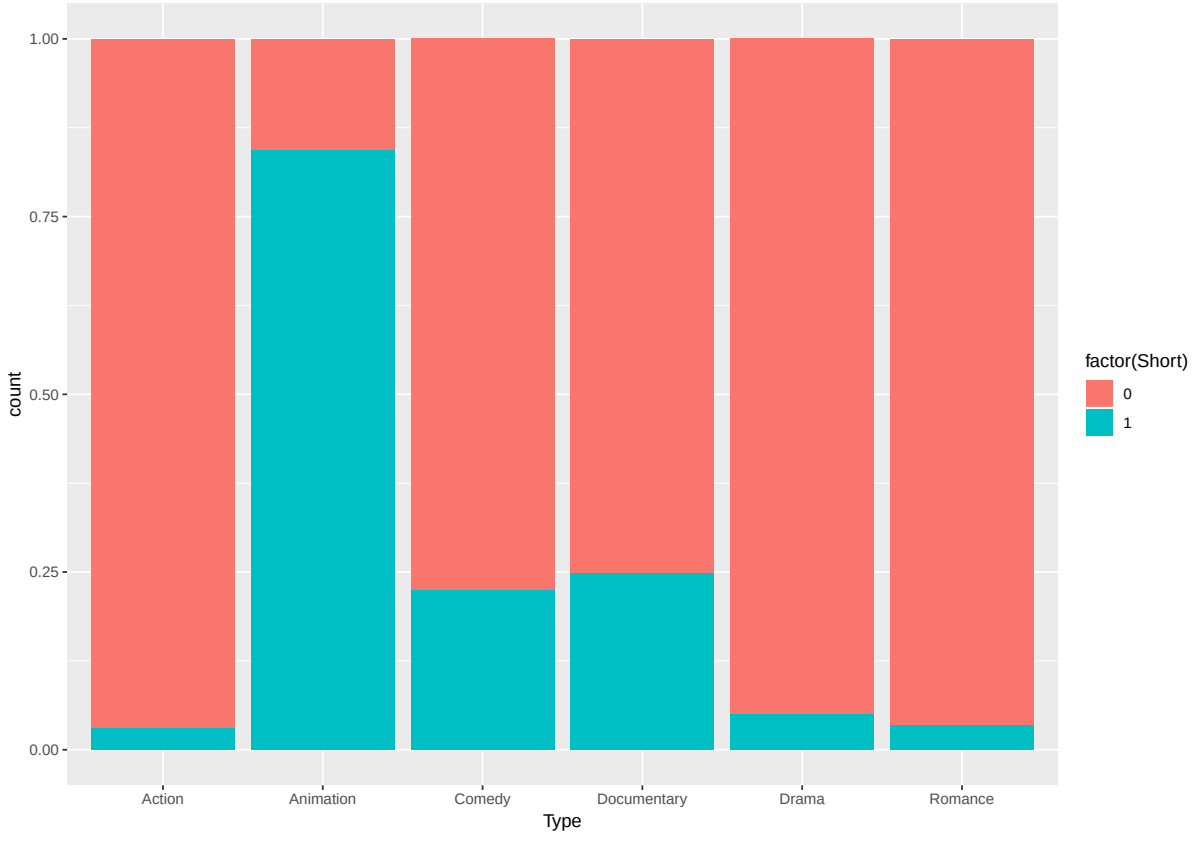


Primeri

```
ggplot(data=myMovies, aes(x=Type, fill=factor(Short))) +
  geom_bar(position="dodge")
```

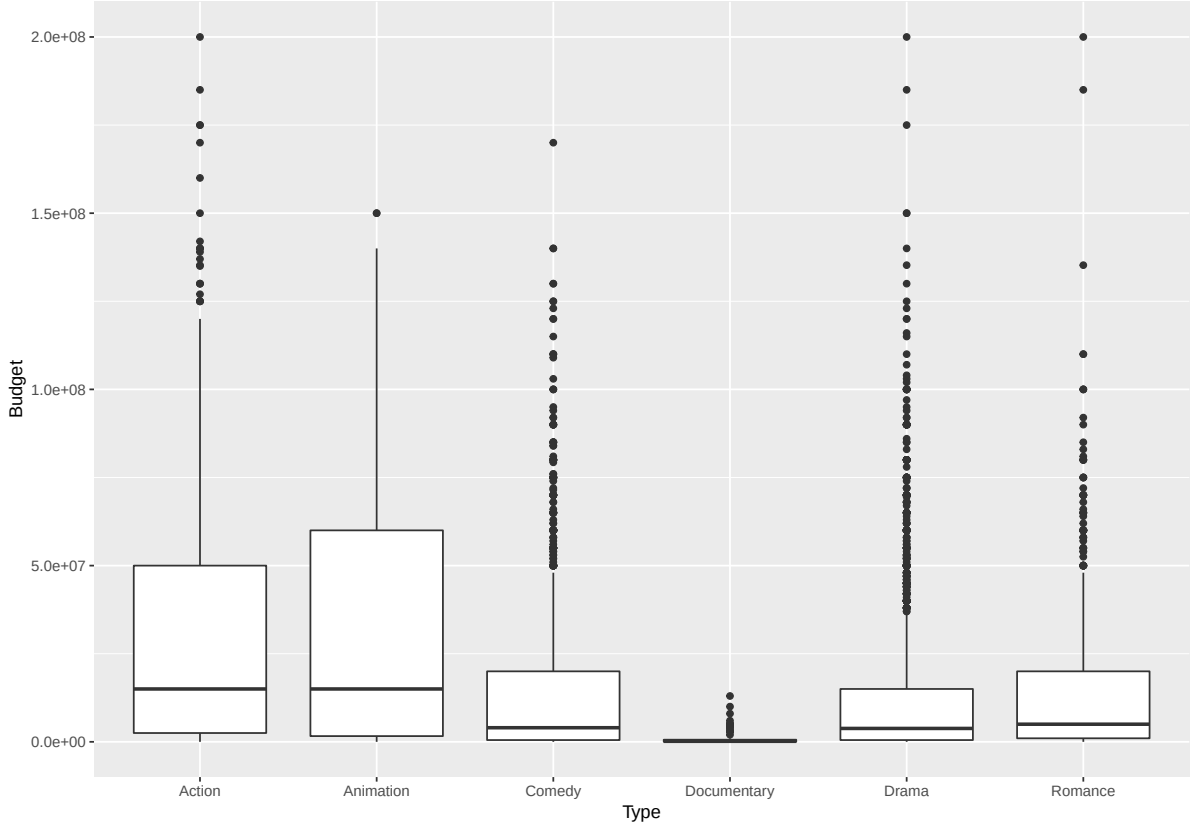


Primeri


1.00
0.75
0.50
0.25
0.00
count
Action
Animation
Comedy
Documentary
Drama
Romance
Type
factor(Short)
0
1

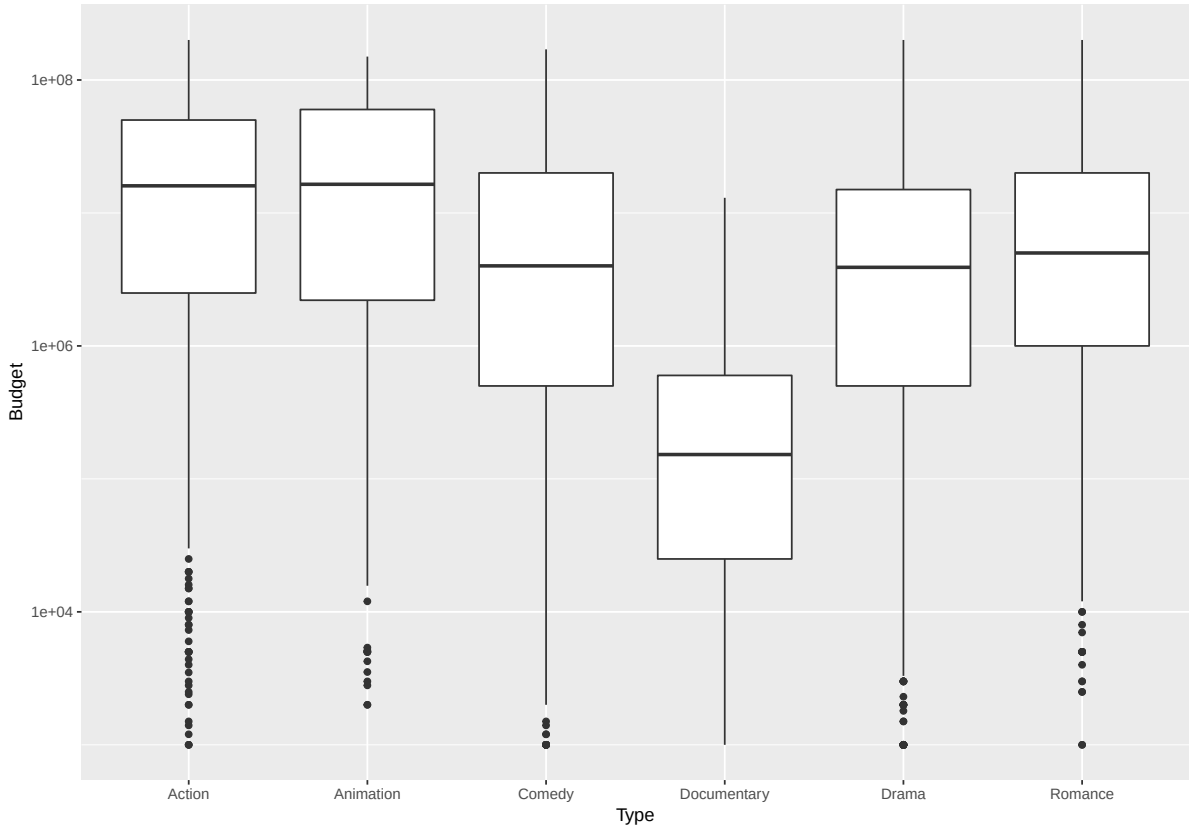
Primeri

```
ggplot(data=myMovies, aes(x=Type, y=Budget)) +
  geom_boxplot()
```



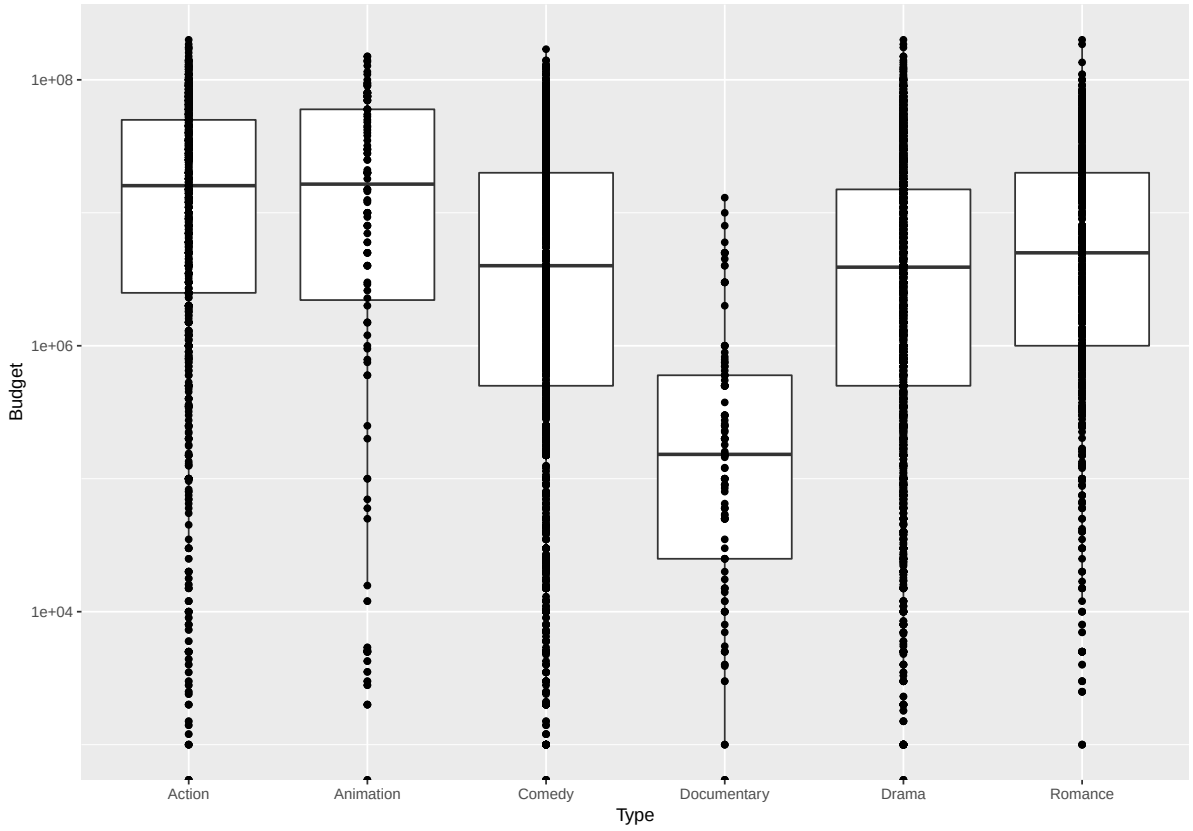
Primeri

```
ggplot(data=myMovies, aes(x=Type, y=Budget)) +
  geom_boxplot() + scale_y_log10()
```



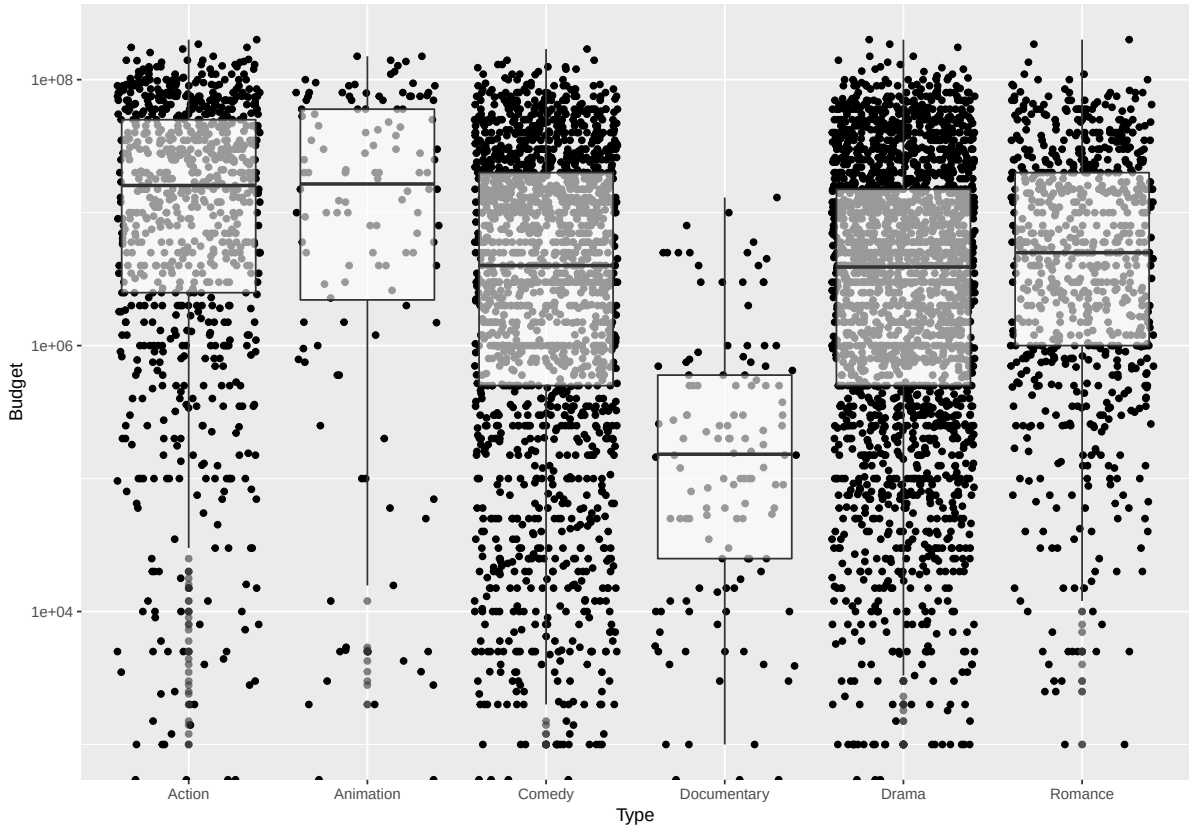
Primeri

```
ggplot(data=myMovies, aes(x=Type, y=Budget)) +
  geom_boxplot() +
  scale_y_log10() + geom_point()
```



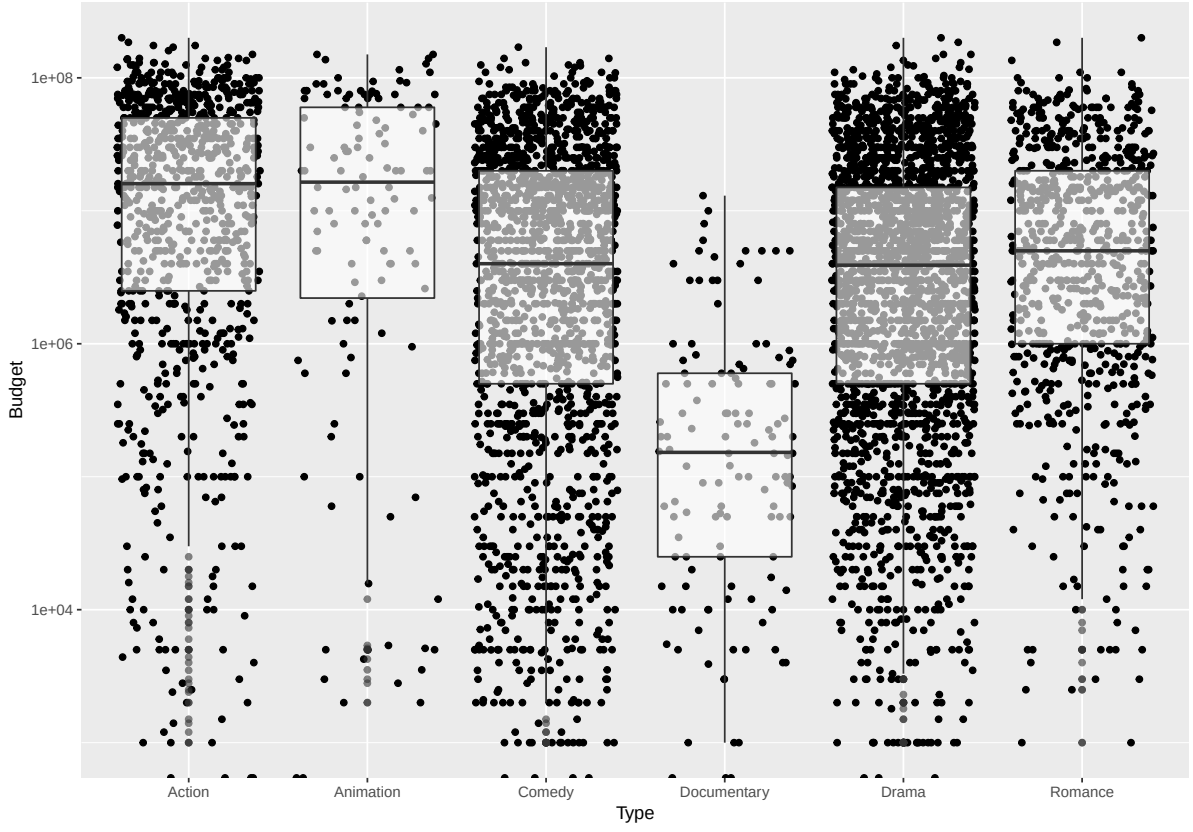
Primeri

```
ggplot(data=myMovies, aes(x=Type, y=Budget)) +
  geom_jitter() +
  geom_boxplot(alpha=I(0.6)) +
  scale_y_log10()
```



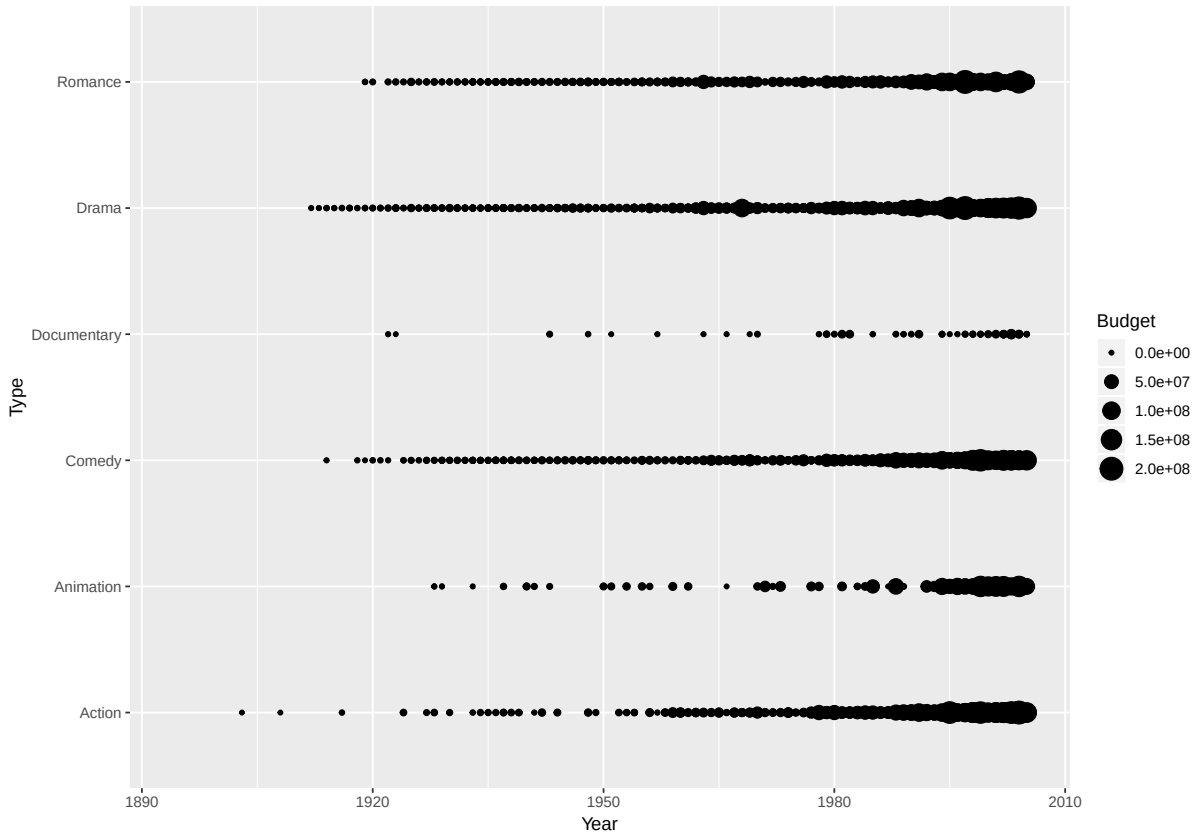
Primeri

```
ggplot(data=myMovies, aes(x=Type, y=Budget)) +
  geom_jitter() + geom_boxplot(alpha=I(0.6)) +
  scale_y_log10()
```



Balončki

```
ggplot(data=myMovies, aes(x=Year, y=Type, size=Budget)) +
  geom_point()
```



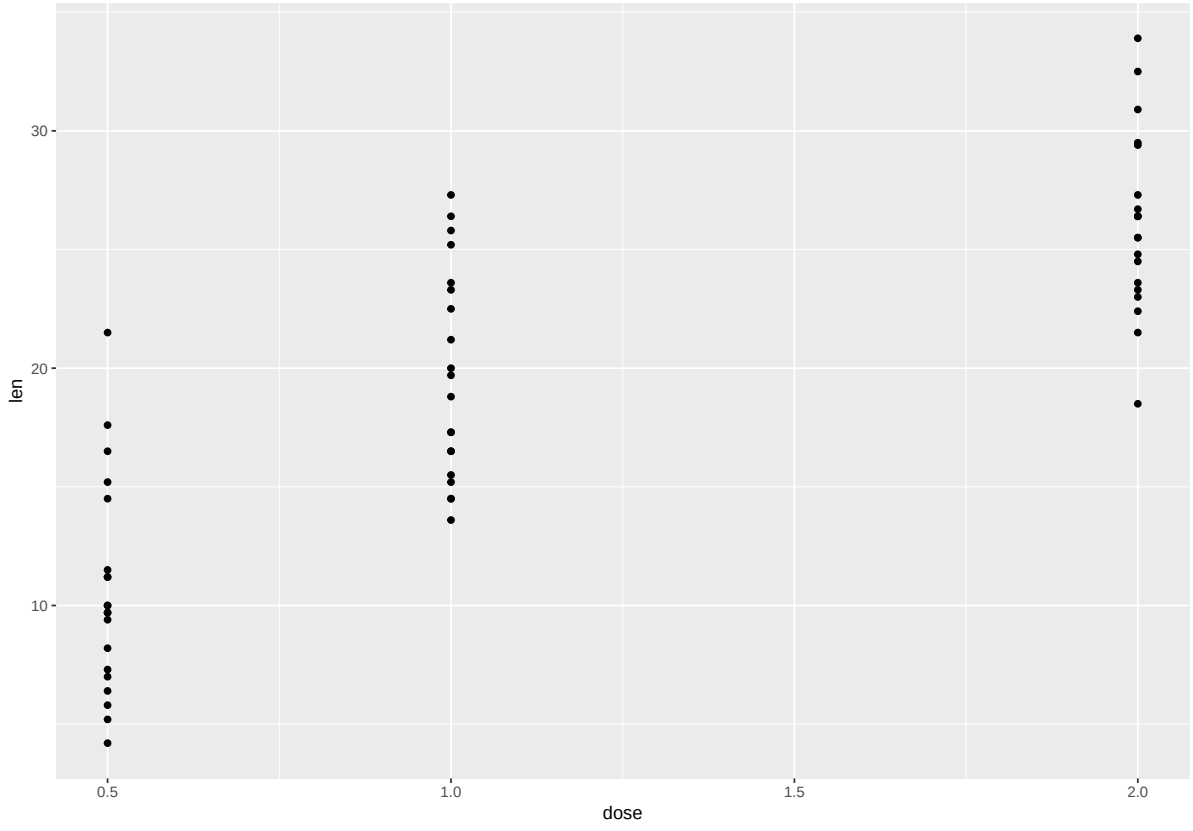
Primeri

```
head(ToothGrowth)
```

```
##    len supp dose
## 1  4.2   VC  0.5
## 2 11.5   VC  0.5
## 3  7.3   VC  0.5
## 4  5.8   VC  0.5
## 5  6.4   VC  0.5
## 6 10.0   VC  0.5
```

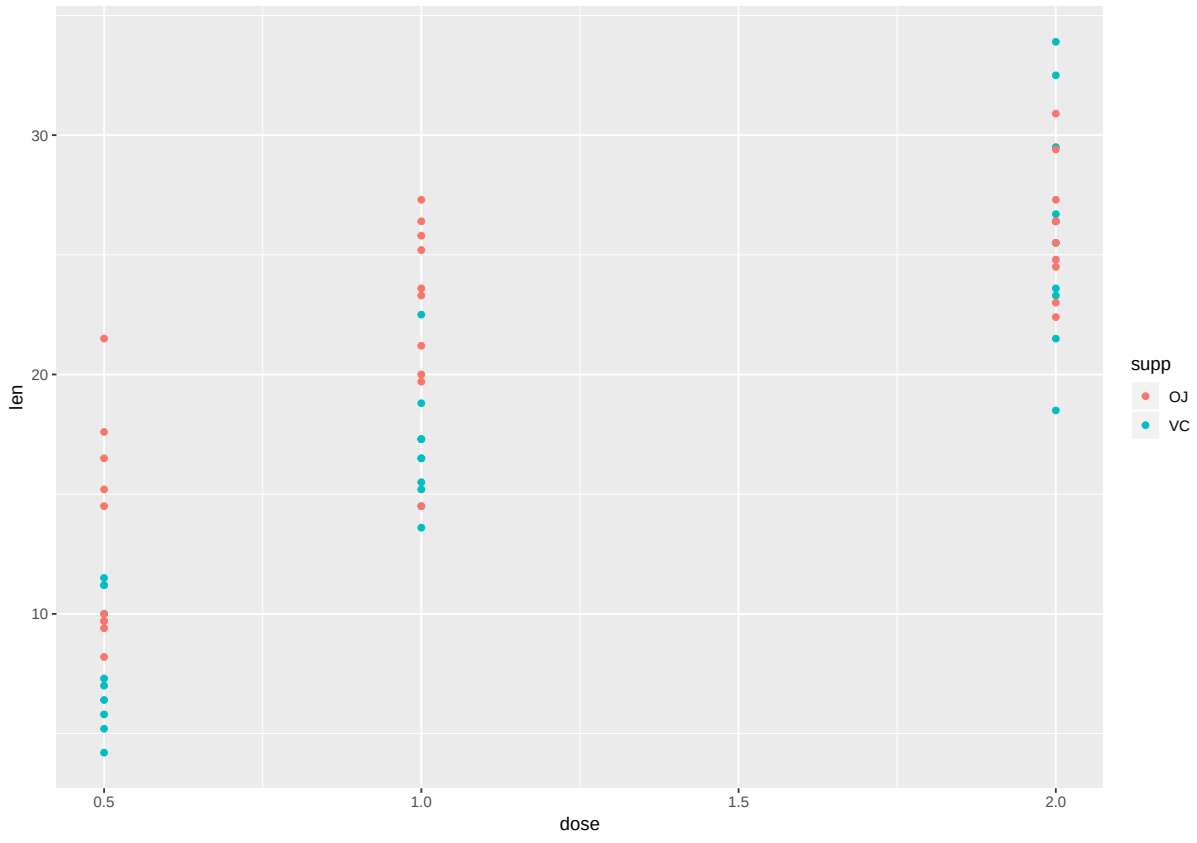
Primeri

```
ggplot(data=ToothGrowth, aes(x=dose, y=len)) +
  geom_point()
```



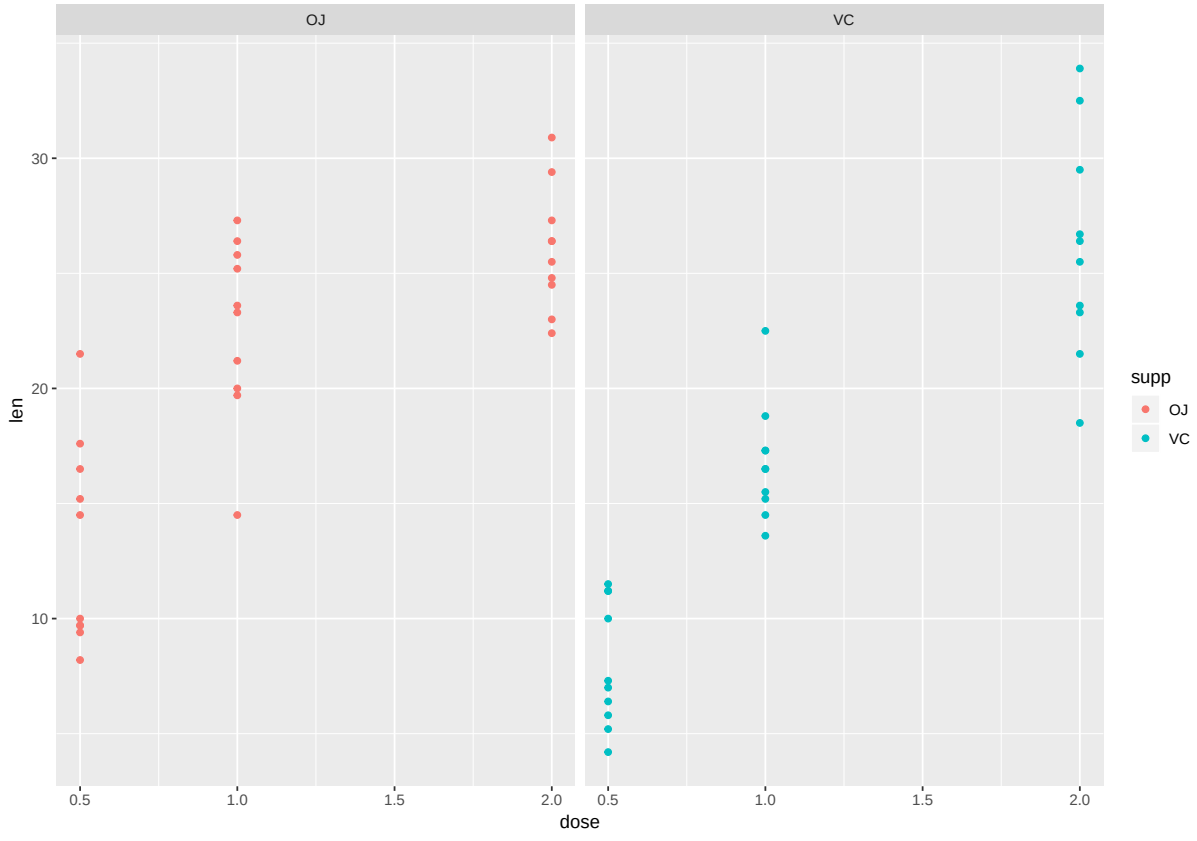
Primeri

```
ggplot(data=ToothGrowth, aes(x=dose, y=len, col=supp)) +
  geom_point()
```



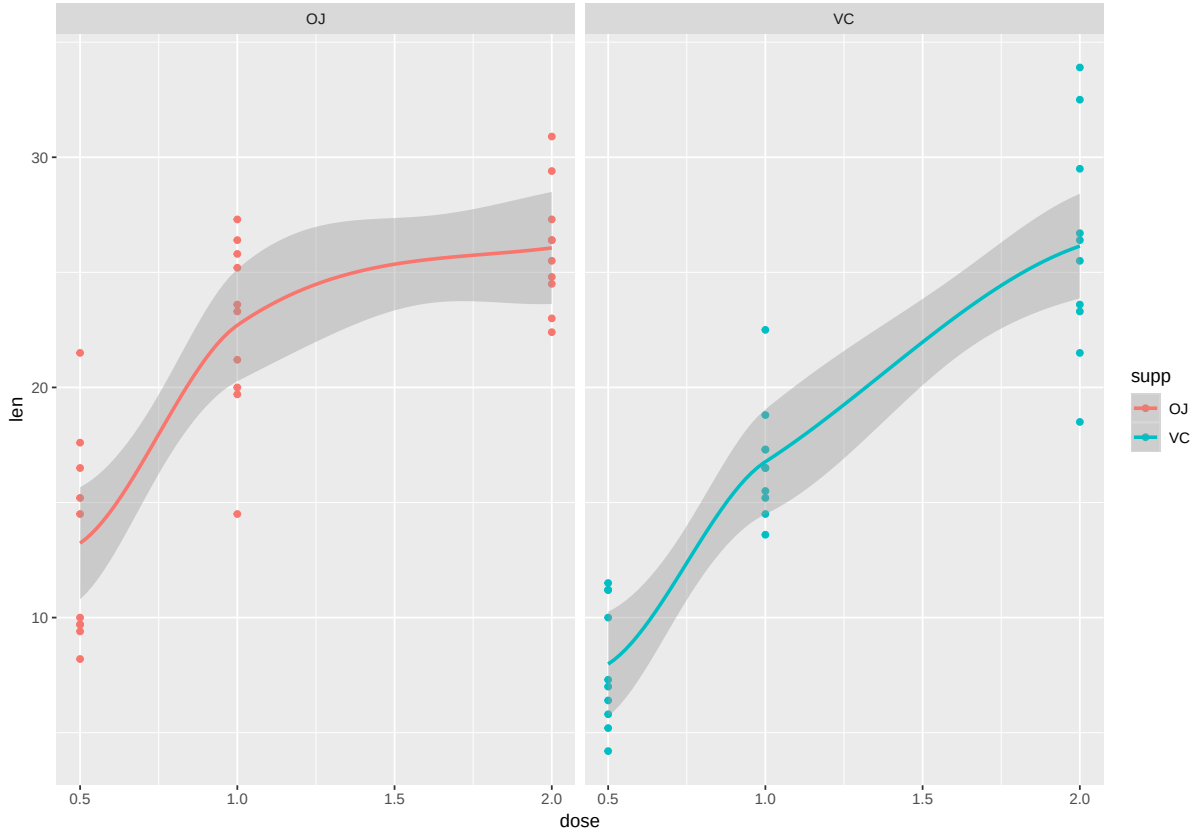
Primeri

```
ggplot(data=ToothGrowth, aes(x=dose, y=len, col=supp)) +
  geom_point() +
  facet_grid(. ~ supp)
```



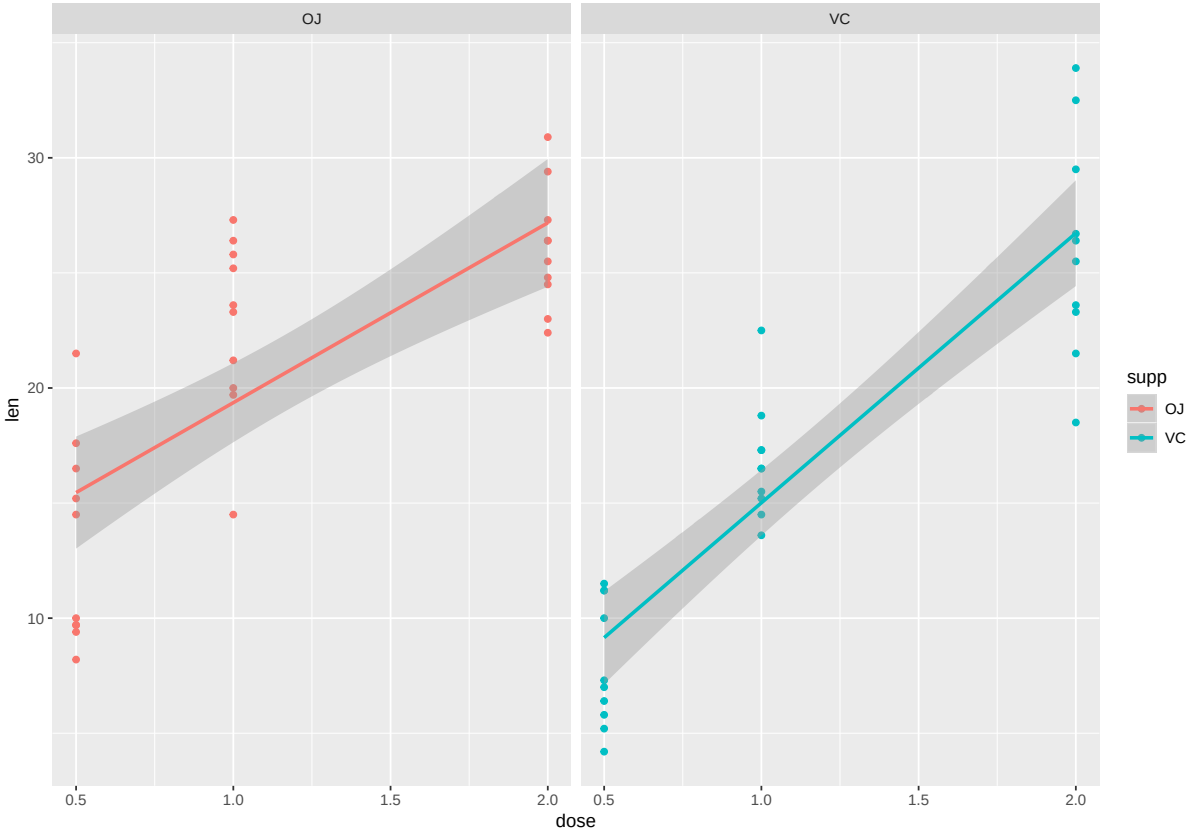
Primeri

```
ggplot(data=ToothGrowth, aes(x=dose, y=len, col=supp)) +
  geom_point() +
  facet_grid(. ~ supp) + stat_smooth()
```



Primeri

```
ggplot(data=ToothGrowth, aes(x=dose, y=len, col=supp)) +
  geom_point() +
  facet_grid(.~supp) +
  stat_smooth(method="lm")
```



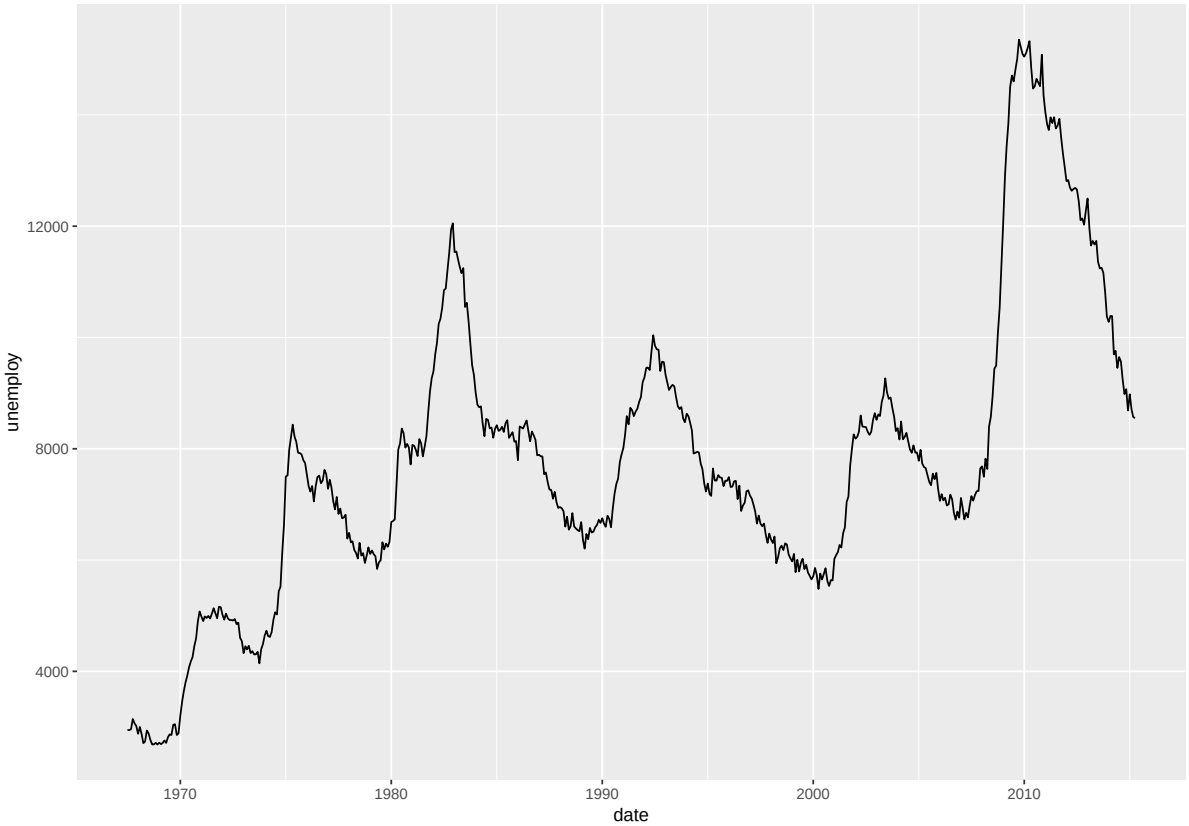
Primeri

```
head(economics)
```

```
## # A tibble: 6 x 6
##   date         pce    pop psavert uempmed unemploy
##   <date>     <dbl>  <int>   <dbl>   <dbl>    <int>
## 1 1967-07-01  507. 198712    12.5     4.5     2944
## 2 1967-08-01  510. 198911    12.5     4.7     2945
## 3 1967-09-01  516. 199113    11.7     4.6     2958
## 4 1967-10-01  513. 199311    12.5     4.9     3143
## 5 1967-11-01  518. 199498    12.5     4.7     3066
## 6 1967-12-01  526. 199657    12.1     4.8     3018
```

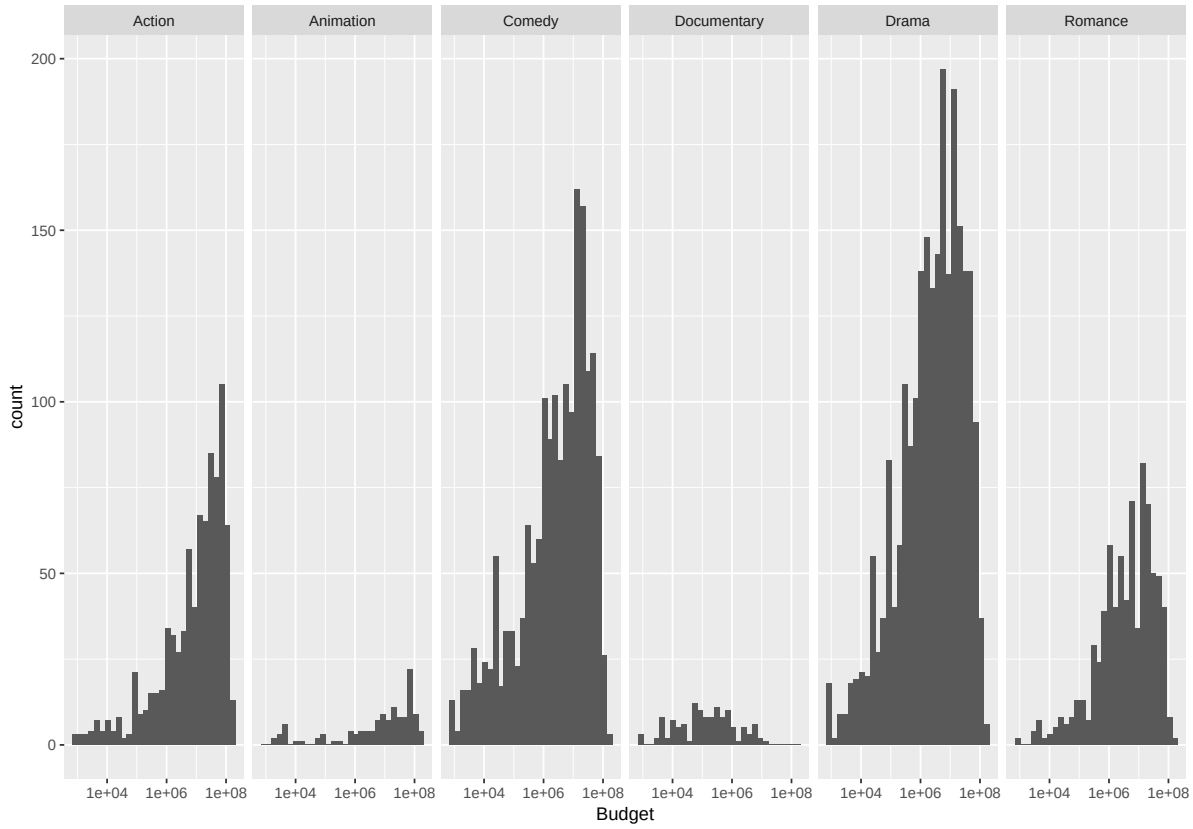
Primeri

```
ggplot(data=economics, aes(x=date, y=unemploy)) + geom_line
```



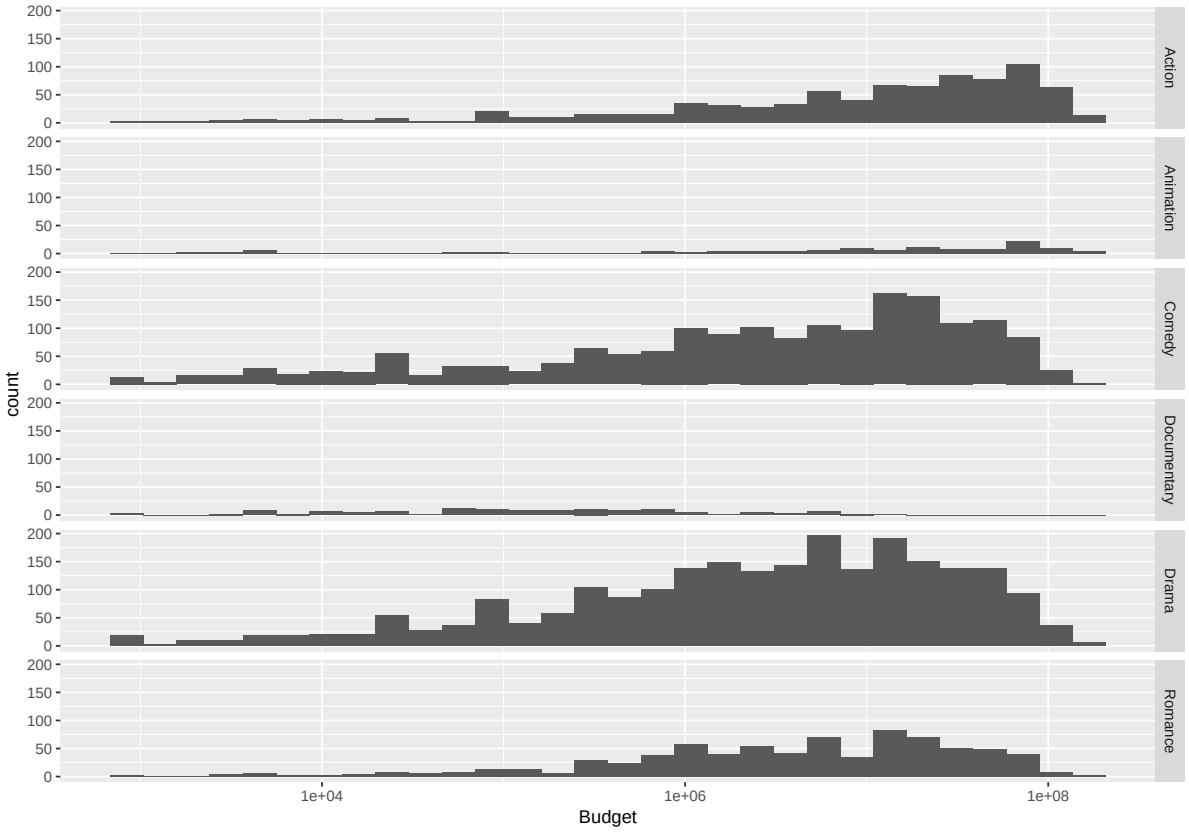
Primeri

```
myMovies$RoundYear <- signif(myMovies$Year, digits = 3)
ggplot(data=myMovies,aes(Budget)) +
  geom_histogram(binwith=1) +
  facet_grid(. ~ Type) + scale_x_log10()
```



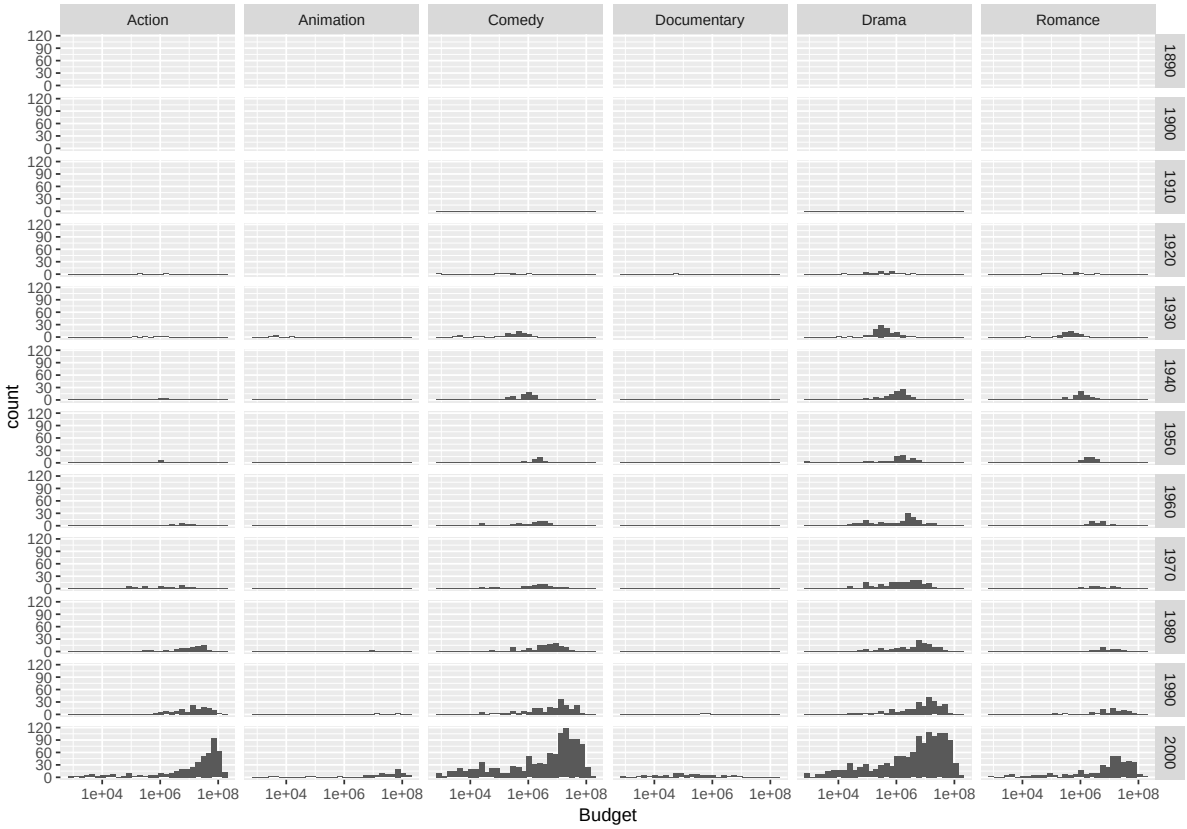
Primeri

```
ggplot(data=myMovies,aes(Budget)) +
  geom_histogram(binwith=1) +
  facet_grid(Type~.) + scale_x_log10()
```



Primeri

```
ggplot(data=myMovies,aes(Budget)) +
  geom_histogram(binwith=1) +
  facet_grid(RoundYear~Type) +
  scale_x_log10()
```



Primeri

```
ggplot(data=subset(myMovies, RoundYear>1980), aes(Budget))
  geom_histogram(binwith=1) +
  facet_grid(. ~ Type+RoundYear) +
  scale_x_log10()
```