R Notebook

Code

Hide

```
library(datasets)
?datasets
```

Hide

```
library(help = "datasets")
```

Hide

```
head(datasets::ToothGrowth)
```

	len	supp	dose
	<dbl>	<fctr>	<dbl>
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

6 rows

Hide

```
?ToothGrowth
```

Hide

```
head(datasets::UKgas)
```

```
[1] 160.1 129.7  84.8 120.1 160.1 124.9
```

Hide

```
?UKgas
```

Hide

```
head(datasets::PlantGrowth)
```

	weight <dbl>	group <fctr>
1	4.17	ctrl
2	5.58	ctrl
3	5.18	ctrl
4	6.11	ctrl
5	4.50	ctrl
6	4.61	ctrl

6 rows

Hide

```
?PlantGrowth
```

Hide

```
head(datasets::quakes)
```

	lat <dbl>	long <dbl>	depth <int>	mag <dbl>	stations <int>
1	-20.42	181.62	562	4.8	41
2	-20.62	181.03	650	4.2	15
3	-26.00	184.10	42	5.4	43
4	-17.97	181.66	626	4.1	19
5	-20.42	181.96	649	4.0	11
6	-19.68	184.31	195	4.0	12

6 rows

Hide

```
?quakes
```

Hide

```
head(datasets::trees)
```

	Girth <dbl>	Height <dbl>	Volume <dbl>
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7

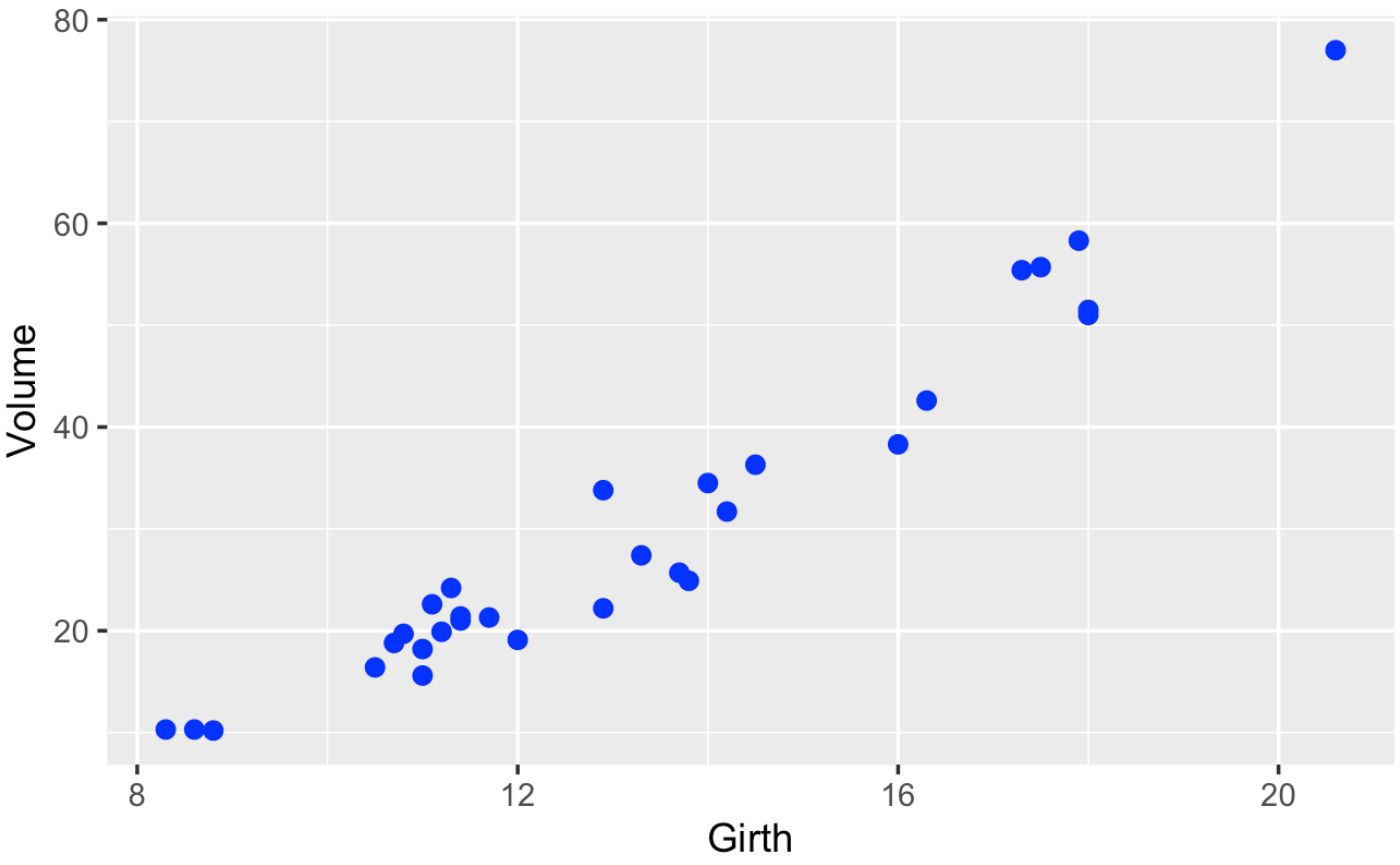
6 rows

Hide

```
?trees
```

Hide

```
library(ggplot2)
plot1 <- ggplot(trees, aes(Girth,Volume)) + geom_point(colour = "blue", size = 2)
plot1
```



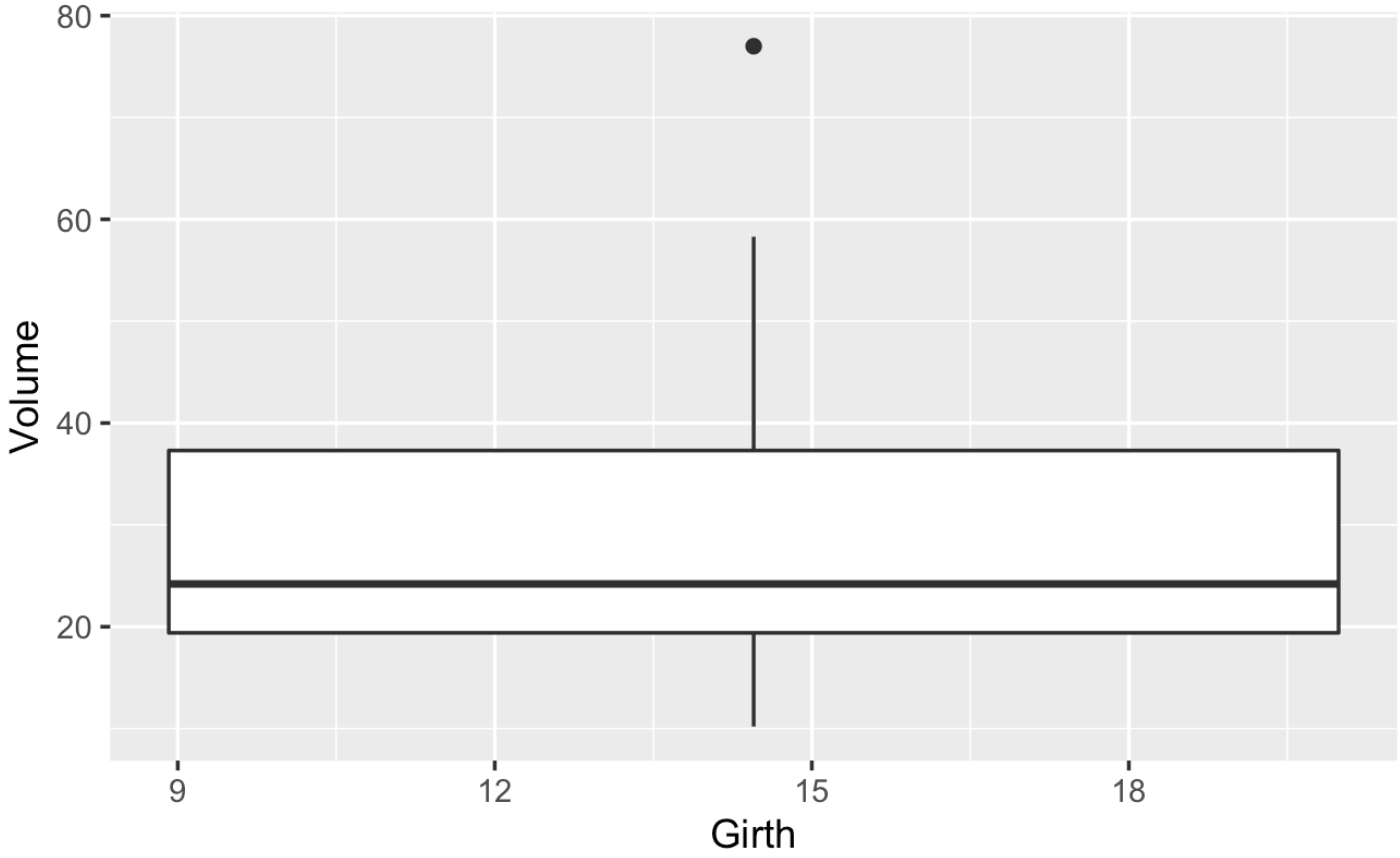
Hide

```
cor.test(trees$Girth, trees$Volume)
```

```
    Pearson's product-moment correlation

data:  trees$Girth and trees$Volume
t = 20.478, df = 29, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9322519 0.9841887
sample estimates:
      cor
0.9671194
```

Hide

```
plot2_bp <- ggplot(trees,aes(Girth,Volume,group = 1)) + geom_boxplot()
plot2_bp
```



Hide

```
head(datasets::cars)
```

	speed	dist
	<dbl>	<dbl>

1	4	2
2	4	10
3	7	4
4	7	22
5	8	16
6	9	10

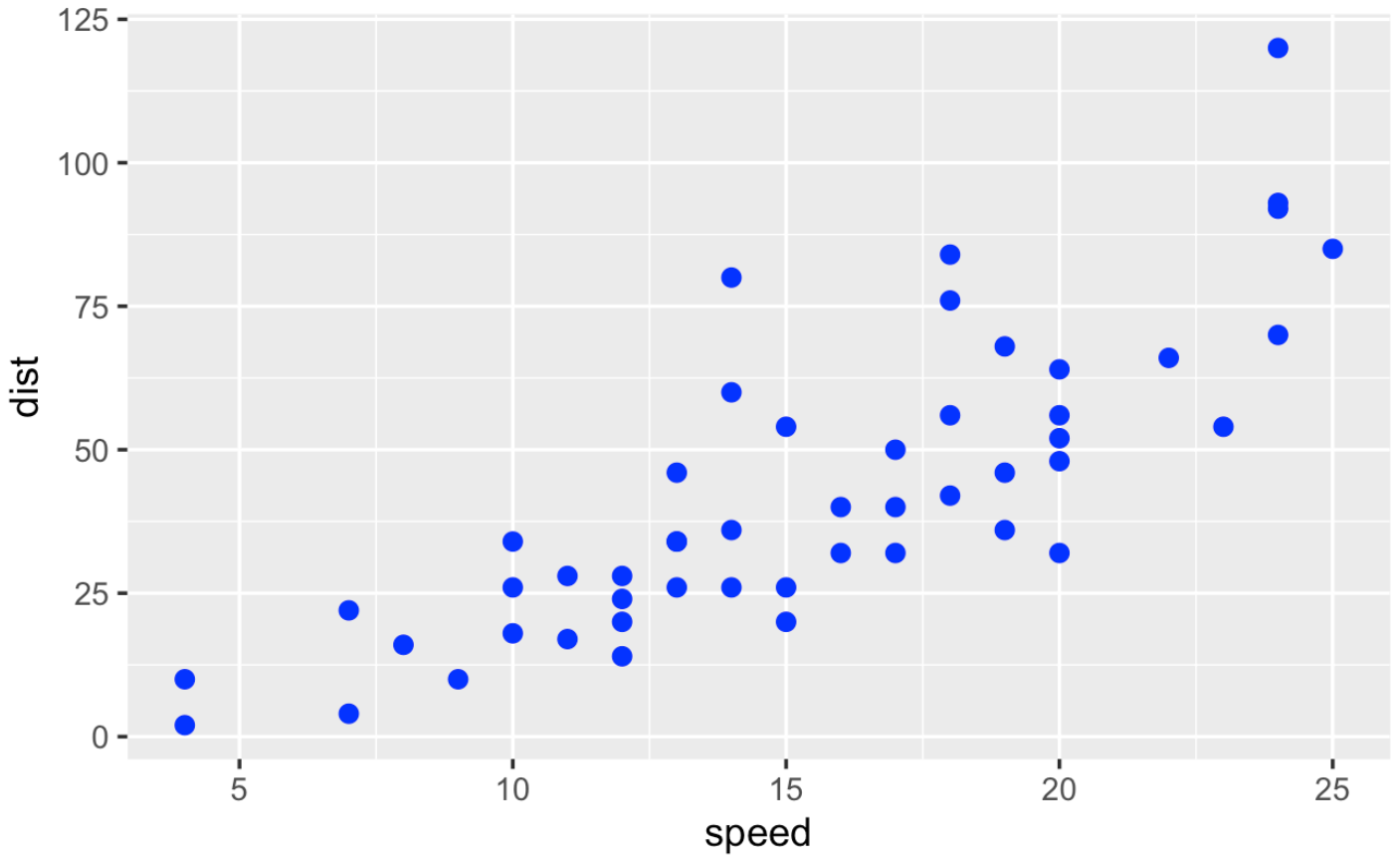
6 rows

Hide

```
?cars
```

Hide

```
plot2 <- ggplot(cars, aes(speed,dist)) + geom_point(colour = "blue", size = 2)
plot2
```



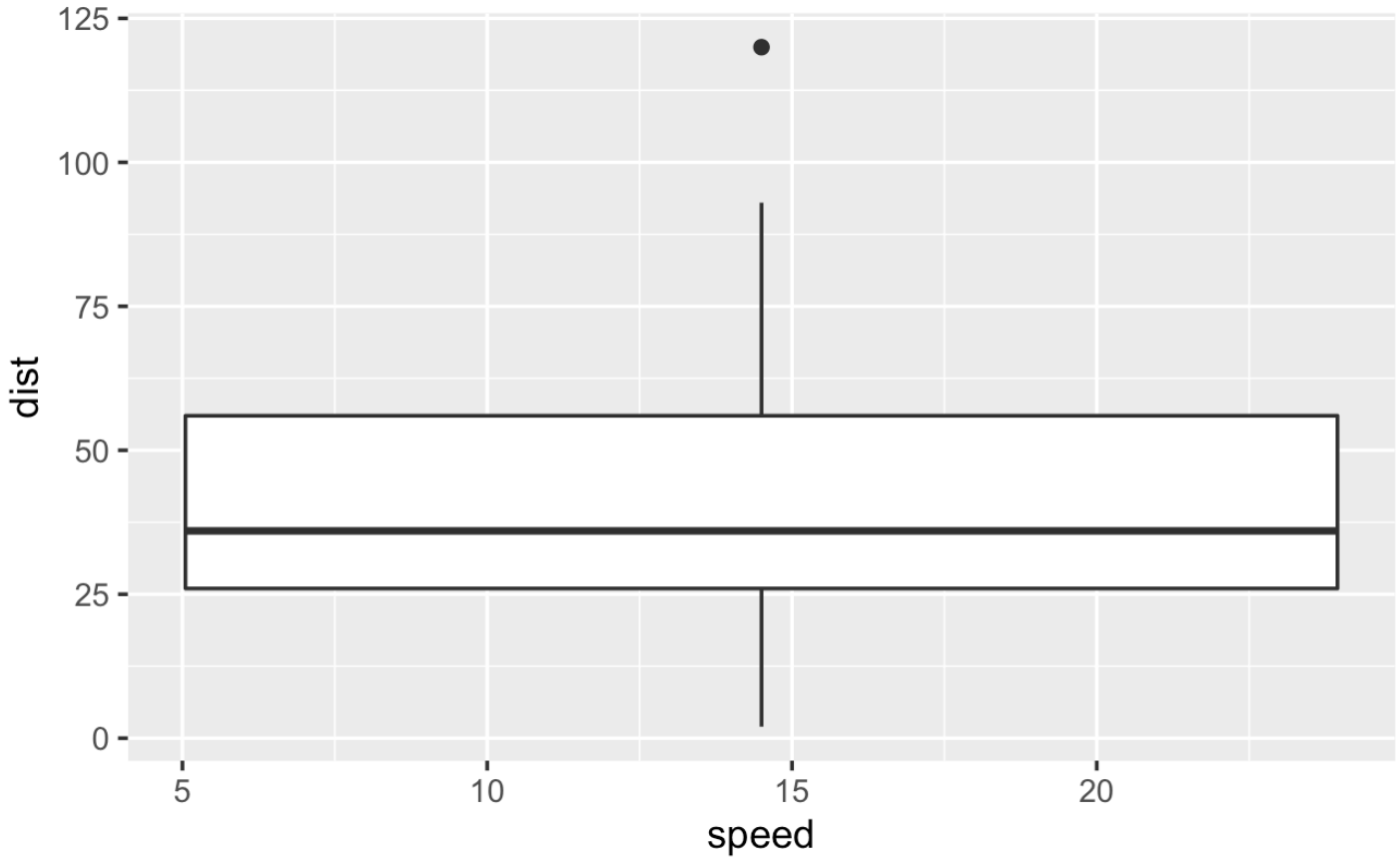
Hide

```
cor.test(cars$speed, cars$dist)
```

```
    Pearson's product-moment correlation

data:  cars$speed and cars$dist
t = 9.464, df = 48, p-value = 1.49e-12
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6816422 0.8862036
sample estimates:
      cor
0.8068949
```

Hide

```
plot2_bp <- ggplot(cars,aes(speed,dist,group = 1)) + geom_boxplot()
plot2_bp
```



Hide

```
head(datasets::state.x77)
```

```
           Population Income Illiteracy Life Exp Murder HS Grad
Alabama          3615   3624        2.1    69.05   15.1    41.3
Alaska            365   6315        1.5    69.31   11.3    66.7
Arizona          2212   4530        1.8    70.55    7.8    58.1
Arkansas         2110   3378        1.9    70.66   10.1    39.9
California      21198   5114        1.1    71.71   10.3    62.6
Colorado         2541   4884        0.7    72.06    6.8    63.9
           Frost   Area
Alabama       20  50708
Alaska       152 566432
Arizona       15 113417
Arkansas      65  51945
California    20 156361
Colorado     166 103766
```

Hide

```
?state.x77
as.data.frame(state.x77)
```

	Population <dbl>	Inco... <dbl>	Illiteracy <dbl>	Life Exp <dbl>	Mur... <dbl>	HS Grad <dbl>	Frost <dbl>	Area <dbl>
Alabama	3615	3624	2.1	69.05	15.1	41.3	20	50708
Alaska	365	6315	1.5	69.31	11.3	66.7	152	566432
Arizona	2212	4530	1.8	70.55	7.8	58.1	15	113417
Arkansas	2110	3378	1.9	70.66	10.1	39.9	65	51945
California	21198	5114	1.1	71.71	10.3	62.6	20	156361
Colorado	2541	4884	0.7	72.06	6.8	63.9	166	103766
Connecticut	3100	5348	1.1	72.48	3.1	56.0	139	4862
Delaware	579	4809	0.9	70.06	6.2	54.6	103	1982
Florida	8277	4815	1.3	70.66	10.7	52.6	11	54090
Georgia	4931	4091	2.0	68.54	13.9	40.6	60	58073

1-10 of 50 rows Previous **1** 2 3 4 5 Next

Hide

```
# Setting the randomizer's seed
set.seed(123)

#Split your data into training and test sets
trees_sample <- sample(c(TRUE, FALSE), nrow(trees), replace = T, prob = c(0.6,0.4)
)
trees_train <- trees[trees_sample, ]
trees_test <- trees[!trees_sample, ]
```

Hide

```
trees_model <- lm(Volume ~ Girth, data = trees_train)
```

Hide

```
cars_sample <- sample(c(TRUE, FALSE), nrow(cars), replace = T, prob = c(0.6,0.4))
cars_train <- cars[cars_sample, ]
cars_test <- cars[!cars_sample, ]
```

Hide

```
cars_model <- lm(dist ~ speed, data = cars_train)
```

Hide

```
# Find the RSE of each model
sigma(trees_model)
```

```
[1] 3.917072
```

Hide

```
sigma(cars_model)
```

```
[1] 16.4502
```

Hide

```
# Find the R Squared value of each model
summary(trees_model)$r.squared
```

```
[1] 0.9450944
```

Hide

```
summary(cars_model)$r.squared
```

```
[1] 0.6233043
```

Hide

```
# Print the summary of each model
summary(trees_model)
```

```

Call:
lm(formula = Volume ~ Girth, data = trees_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.7789 -2.6169 -0.3262  3.1373  5.3413 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -34.3241     4.1916  -8.189 1.04e-06 ***
Girth         4.8761     0.3141  15.524 3.23e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.917 on 14 degrees of freedom
Multiple R-squared:  0.9451,    Adjusted R-squared:  0.9412 
F-statistic:   241 on 1 and 14 DF,  p-value: 3.23e-10
```

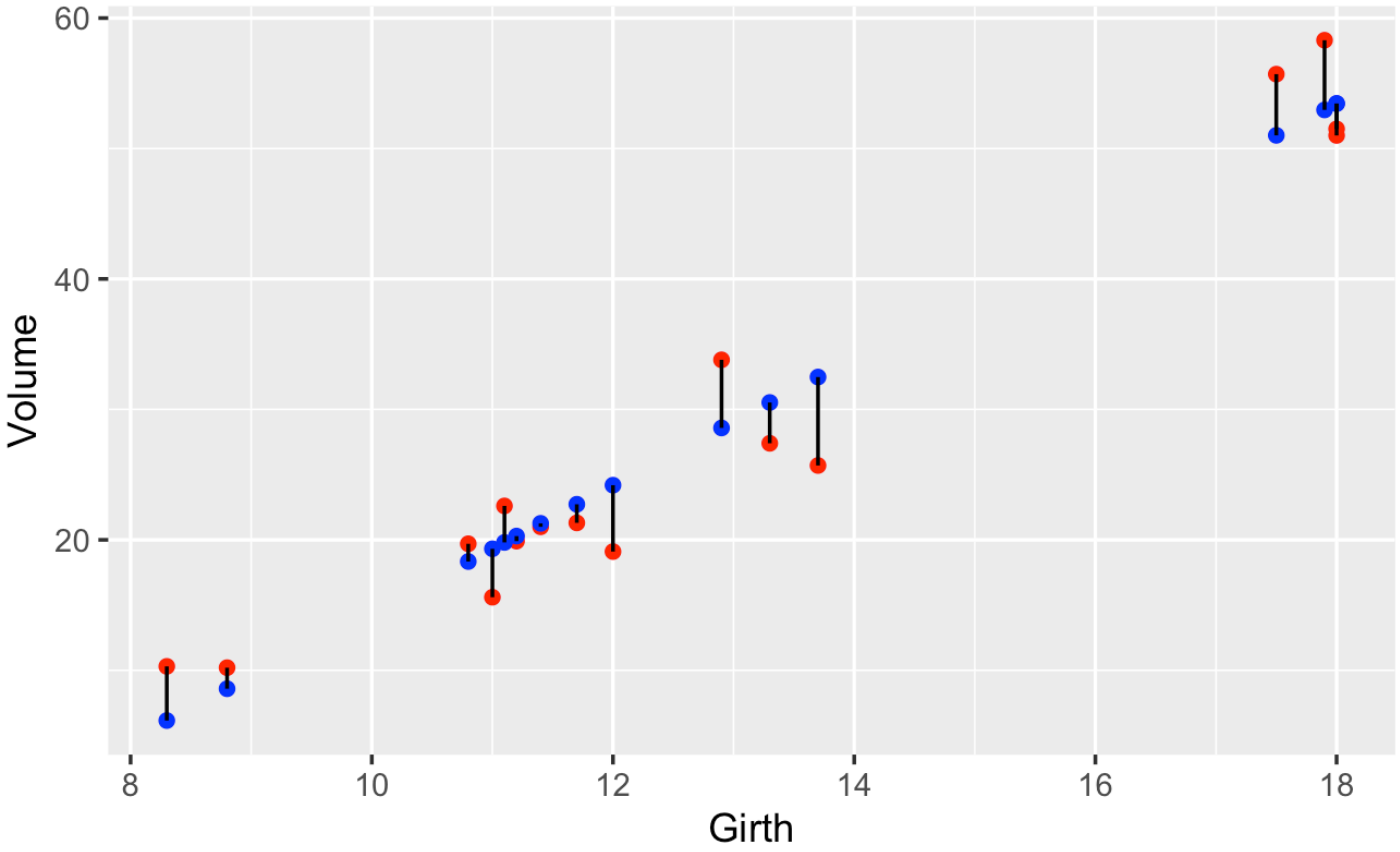
Hide

```
summary(cars_model)
```

```

Call:
lm(formula = dist ~ speed, data = cars_train)

Residuals:
    Min      1Q  Median      3Q     Max 
-30.122  -8.886  -0.886   3.732  42.805 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -20.9673     9.4277  -2.224   0.0336 *  
speed         4.1544     0.5801   7.162 4.74e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.45 on 31 degrees of freedom
Multiple R-squared:  0.6233,    Adjusted R-squared:  0.6112 
F-statistic: 51.29 on 1 and 31 DF,  p-value: 4.738e-08
```

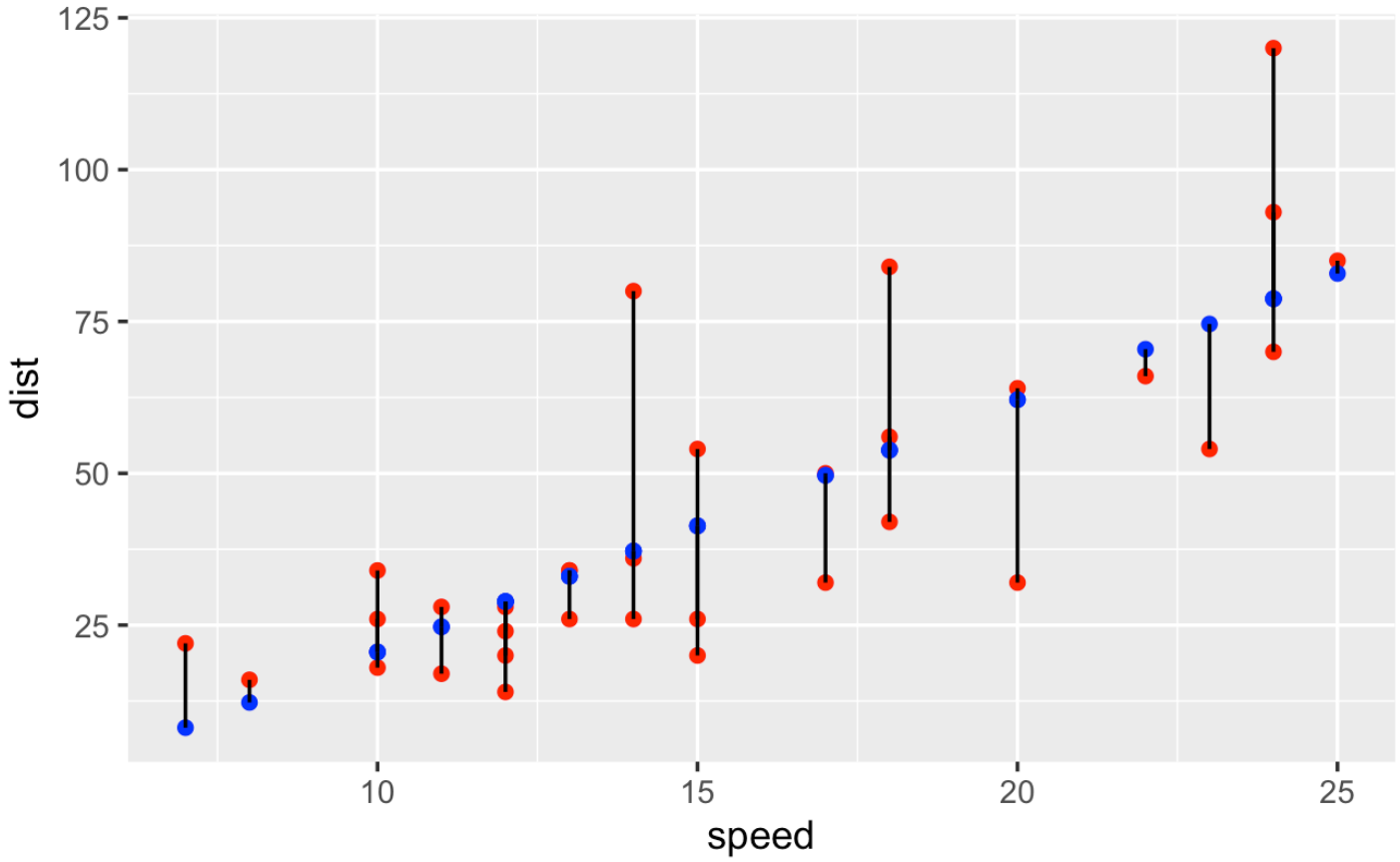
Hide

```
#save predicted and residual values to df
trees_train$estimate <- predict(trees_model)
trees_train$residuals <- residuals(trees_model)
cars_train$estimate <- predict(cars_model)
cars_train$residuals <- residuals(cars_model)
#create visualization
ggplot(trees_train, aes(Girth, Volume)) +
geom_point(colour = "red") +
geom_point(aes(y = estimate), colour = "blue") +
geom_segment(aes(xend = Girth, yend = estimate), colour = "black")
```



Hide

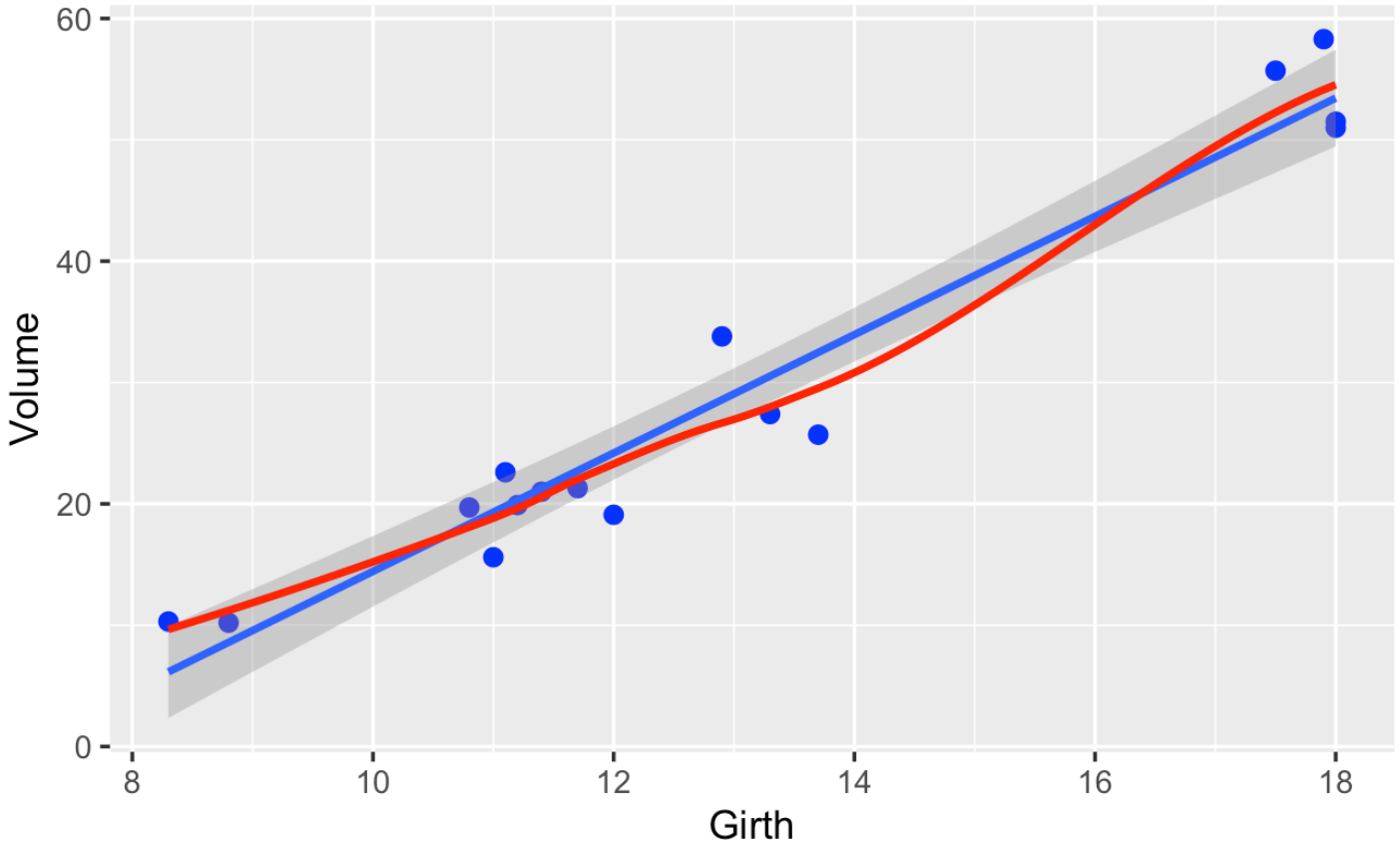
```
ggplot(cars_train, aes(speed, dist)) +
geom_point(colour = "red") +
geom_point(aes(y = estimate), colour = "blue") +
geom_segment(aes(xend = speed, yend = estimate), colour = "black")
```



Hide

```
# Add a LOESS smoother to your scatterplot
plot1 <- ggplot(trees_train, aes(Girth,Volume)) + geom_point(colour = "blue", size
= 2) + geom_smooth(method = "lm") + geom_smooth(se = FALSE, color = "red")
plot1
```

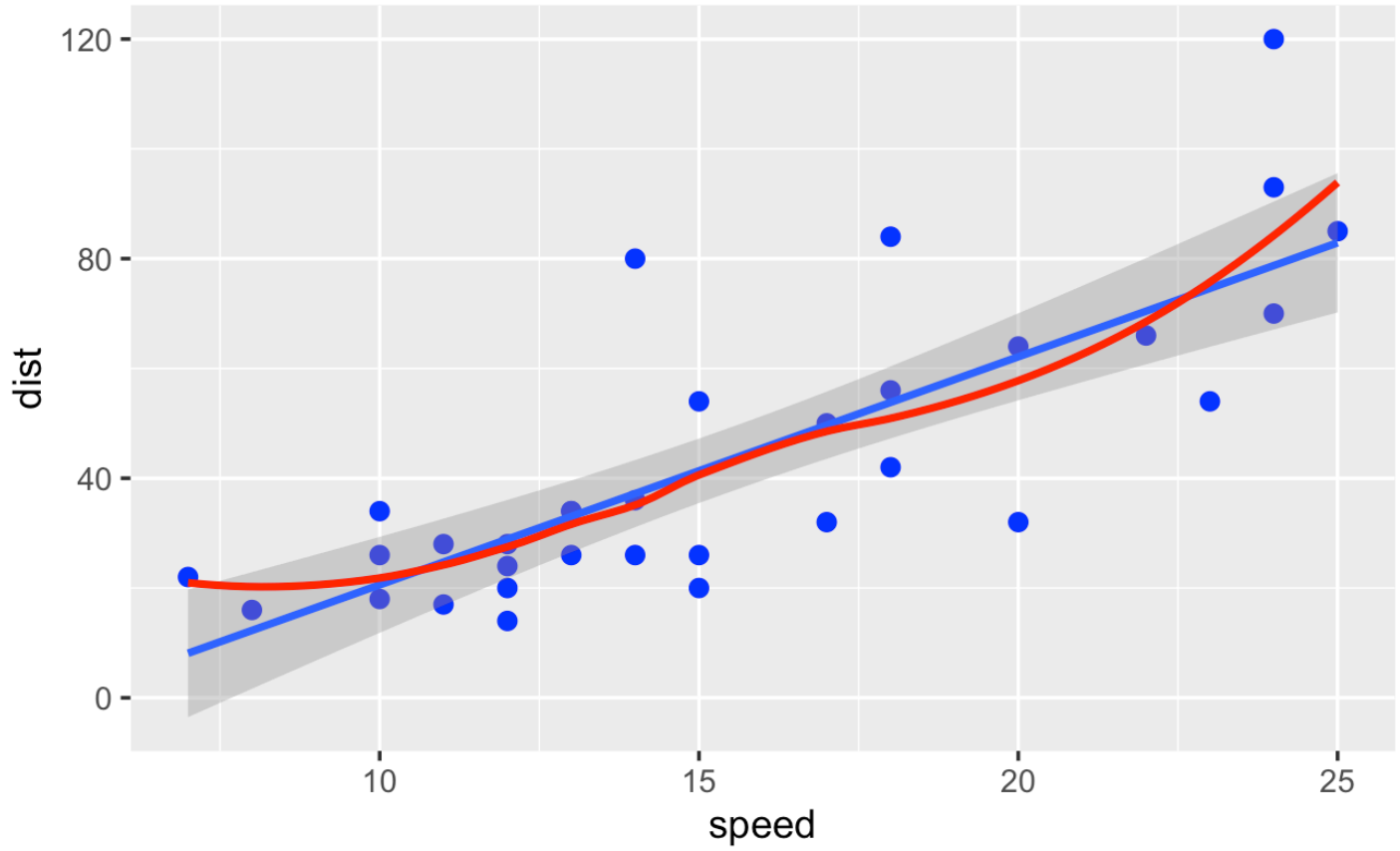
```
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Hide

```
plot2 <- ggplot(cars_train, aes(speed,dist)) + geom_point(colour = "blue", size =
2) + geom_smooth(method = "lm") + geom_smooth(se = FALSE, color = "red")
plot2
```

```
`geom_smooth()` using formula 'y ~ x'
`geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Hide

```
# Create a multiple linear regression model
trees_model_multivar <- lm(Volume ~ Girth + Height, data = trees_train)
summary(trees_model_multivar)
```

```

Call:
lm(formula = Volume ~ Girth + Height, data = trees_train)

Residuals:
   Min     1Q Median     3Q    Max 
-5.462 -1.990 -0.561  3.367  5.753 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -47.5291    11.2544  -4.223 0.000996 ***
Girth         4.6319     0.3637  12.737 1.02e-08 ***
Height        0.2146     0.1703   1.260 0.229751    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.837 on 13 degrees of freedom
Multiple R-squared:  0.9511,    Adjusted R-squared:  0.9435 
F-statistic: 126.3 on 2 and 13 DF,  p-value: 3.035e-09
```

Hide

Hide

```
#create visualization
ggplot(trees_train, aes(Girth + Height, Volume)) +
geom_point(colour = "red") +
geom_point(aes(y = estimate), colour = "blue") +
geom_segment(aes(xend = Girth + Height, yend = estimate), colour = "black")
```