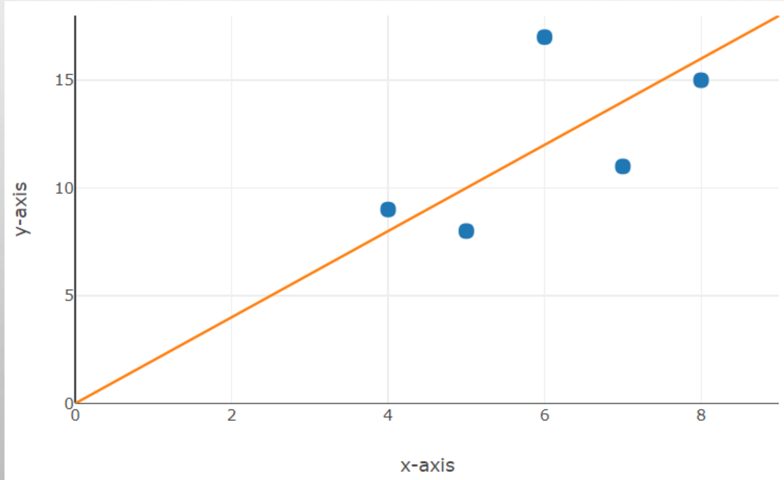
CHAPTER 5 - LINEAR REGRESSION AND PLOTTINGS

LINEAR REGRESSION

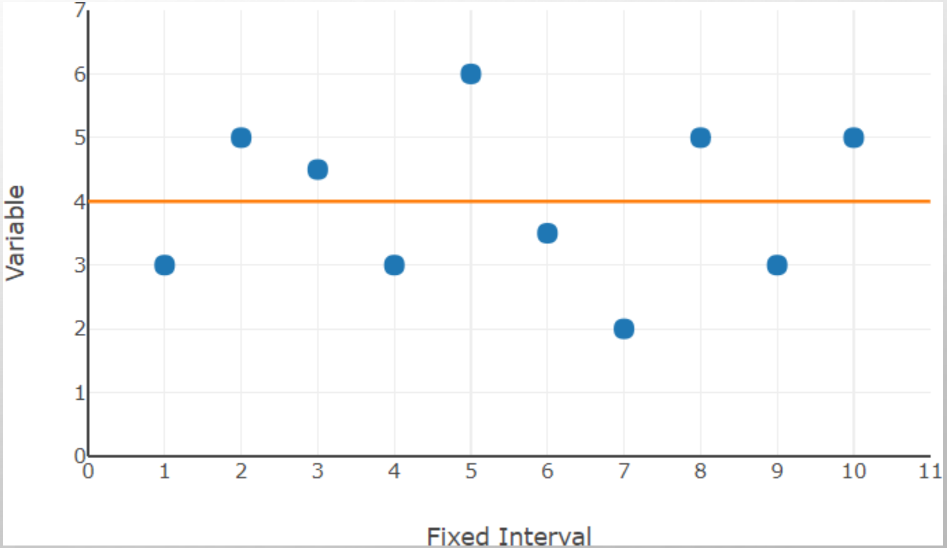
- The goal of regression is to develop an equation or formula that best describes the relationship between variables.



$y = 2x$

LINEAR REGRESSION

- How do we find a best-fit line?
- Consider a dataset with only one variable
- The best-fit line is just the mean value of the data points

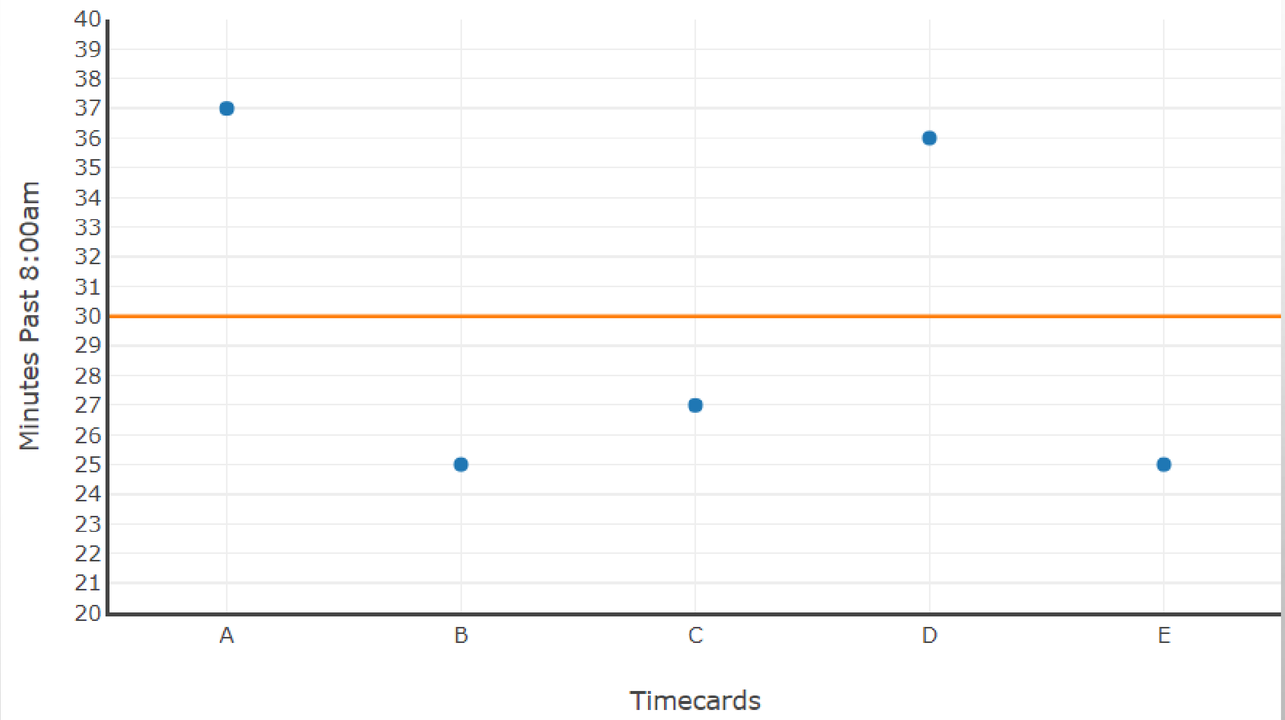


UNDERSTANDING BEST FIT

- A plant manager wants to know when employees arrive at work
- The shift starts a 8:30am
- She takes five random timecards and plots the minutes of arrival on a chart

UNDERSTANDING BEST FIT

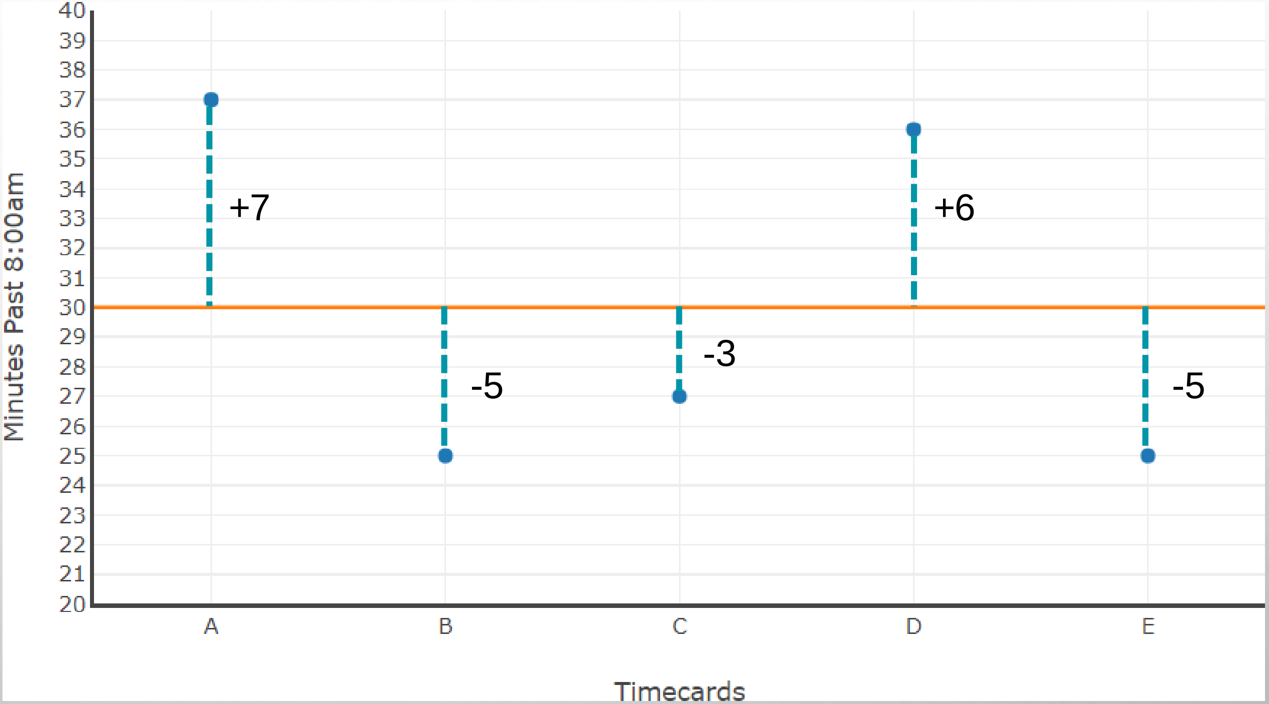
Timecard	Minutes past 8:00am
A	37
B	25
C	27
D	36
E	25
Total:	**150**
Mean	**30**



UNDERSTANDING BEST FIT

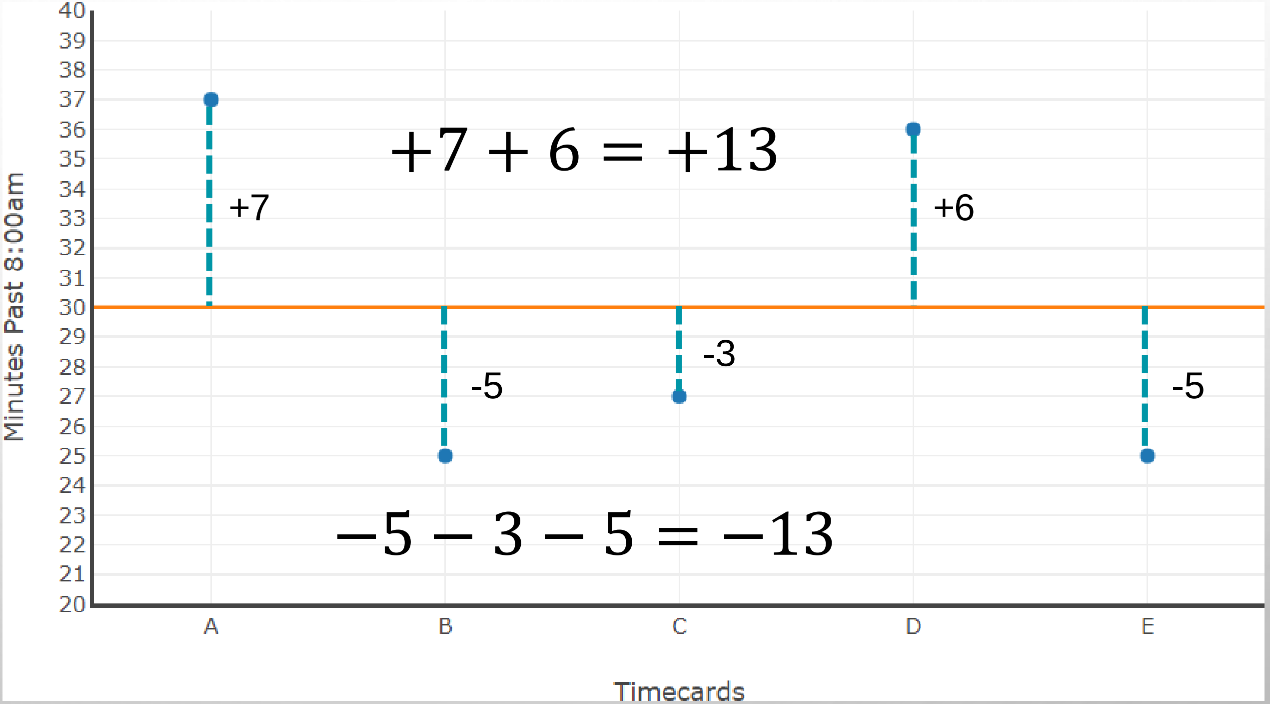
What makes $y = 30$ a best-fit line?

Consider the error



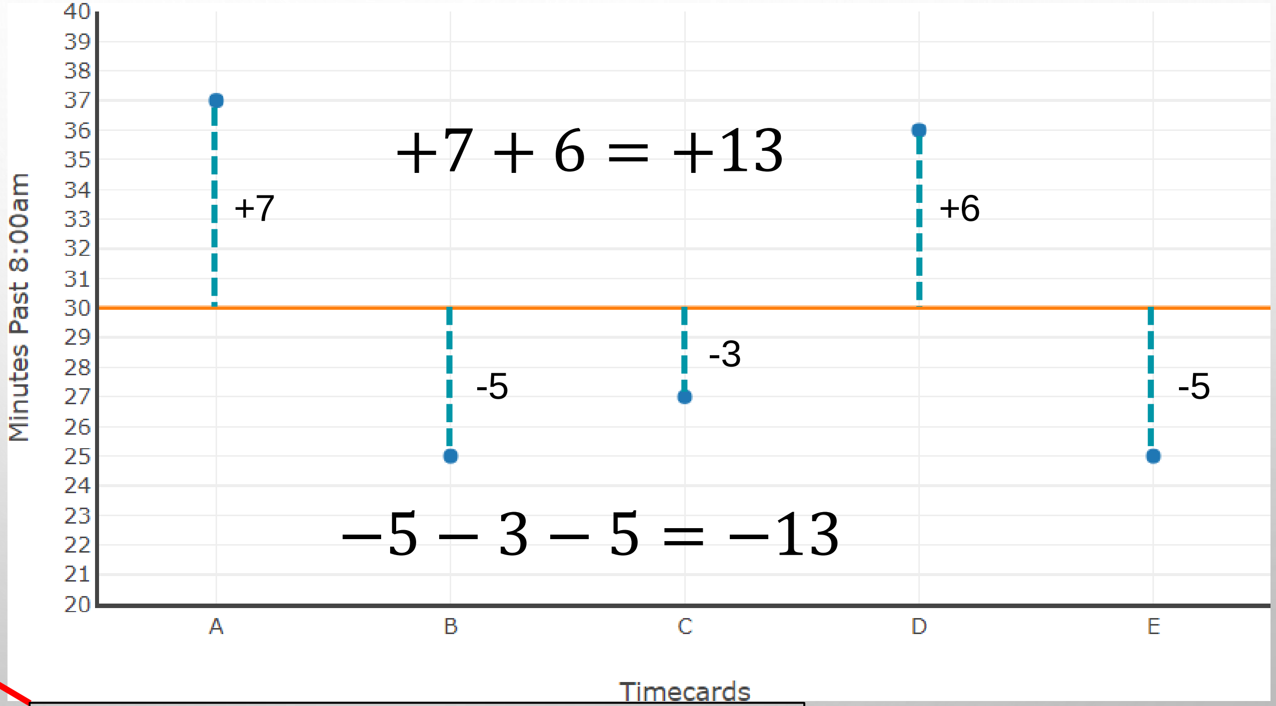
UNDERSTANDING BEST FIT

See that the sum of the distances above the line balances the sum of those below the line



UNDERSTANDING BEST FIT

Error (E)	Square Error (SE)
+7	49
-5	25
-3	9
+6	36
-5	25
Sum of Squares Error (SSE)	**144**



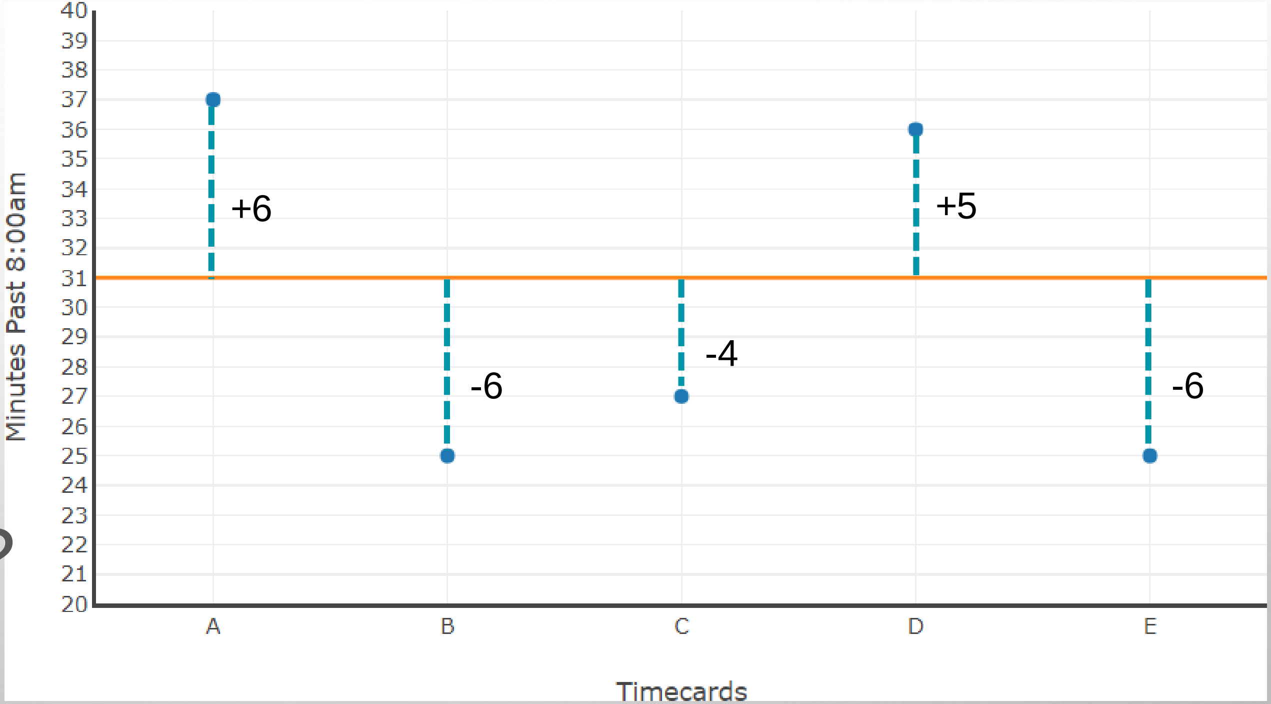
we want to MINIMIZE the SSE

UNDERSTANDING BEST FIT

What if we move the line?

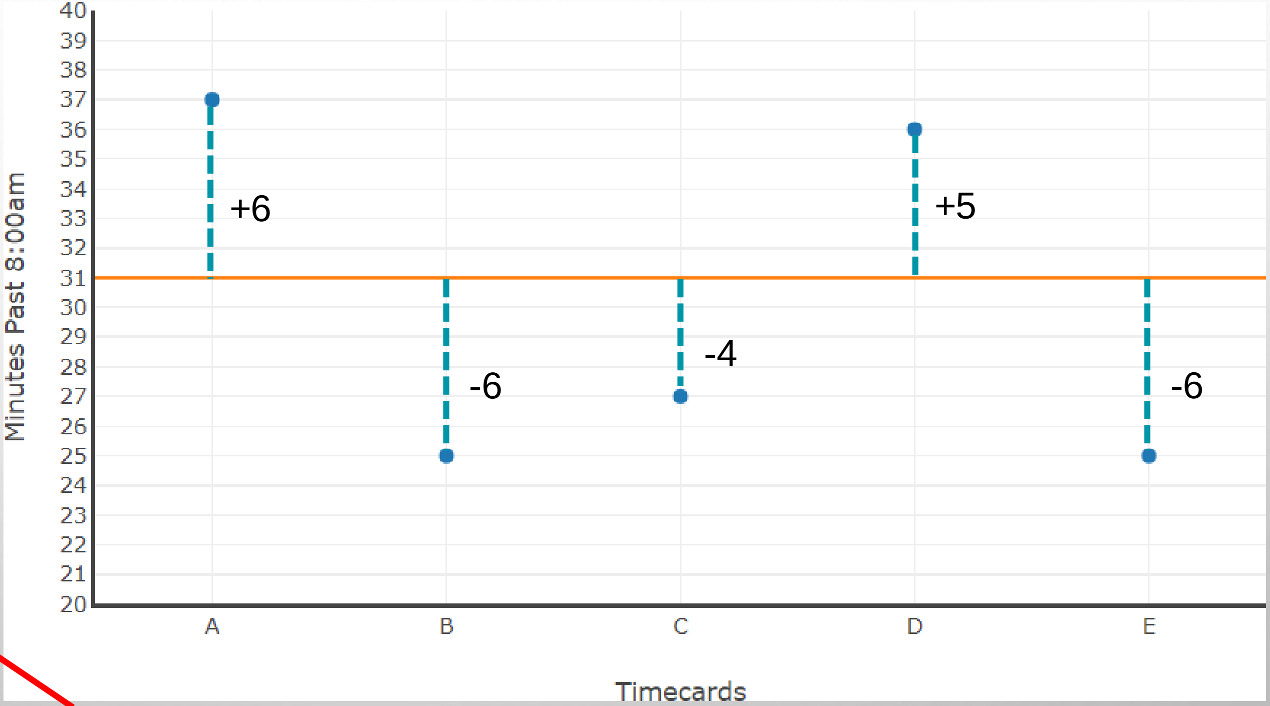
Let's set it to $y = 31$ instead

How does it affect the SSE?



UNDERSTANDING BEST FIT

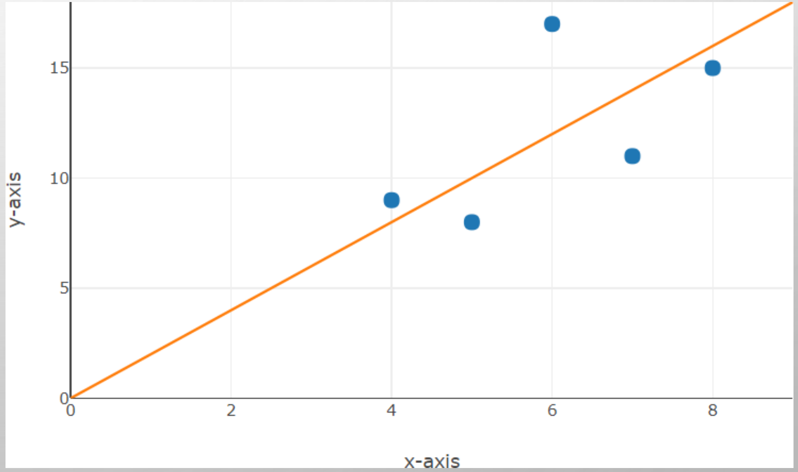
Error (E)		Square Error (SE)	
+7	+6	49	36
-5	-6	25	36
-3	-4	9	16
+6	+5	36	25
-5	-6	25	36
Sum of Squares Error (SSE)		144	149



moving the line INCREASED the SSE

LINEAR REGRESSION

- That's it! The goal of regression is to find the line that best describes our data.
- Fortunately, we don't have to rely on trial-and-error.
- We have algebra!

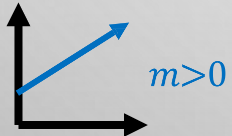


LINEAR REGRESSION

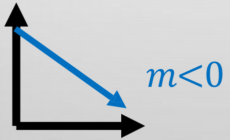
- Recall that the equation of a line follows the form $y = mx + b$ where

  m is the slope of the line, and

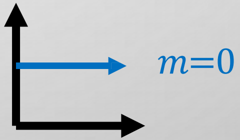
  b is where the line crosses the y-axis when x=0 (b is the y-intercept)



positive slope



negative slope



zero slope

LINEAR REGRESSION

- In a linear regression, where we try to formulate the relationship between variables, $y = mx + b$ becomes

$$\hat{y} = b_0 + b_1 x$$

- Our goal is to predict the value of a dependent variable (y) based on that of an independent variable (x).

$$\hat{y} = b_0 + b_1 x$$

LINEAR REGRESSION

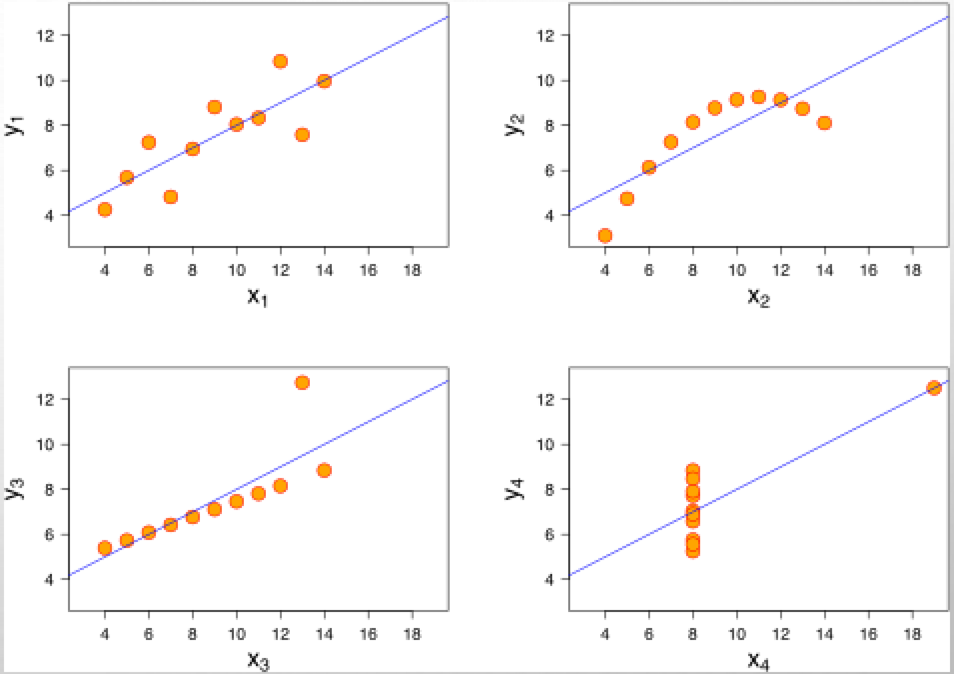
- How to derive b_1 and b_0:

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

LIMITATIONS OF LINEAR REGRESSION

Anscombe's Quartet illustrates the pitfalls of relying on pure calculation.
Each graph results in the same calculated regression line.



REGRESSION EXERCISE #1

- A manager wants to find the relationship between the number of hours that a plant is operational in a week and weekly production.

REGRESSION EXERCISE #1

- Here the independent variable x is hours of operation, and the dependent variable y is production volume.

REGRESSION EXERCISE #1

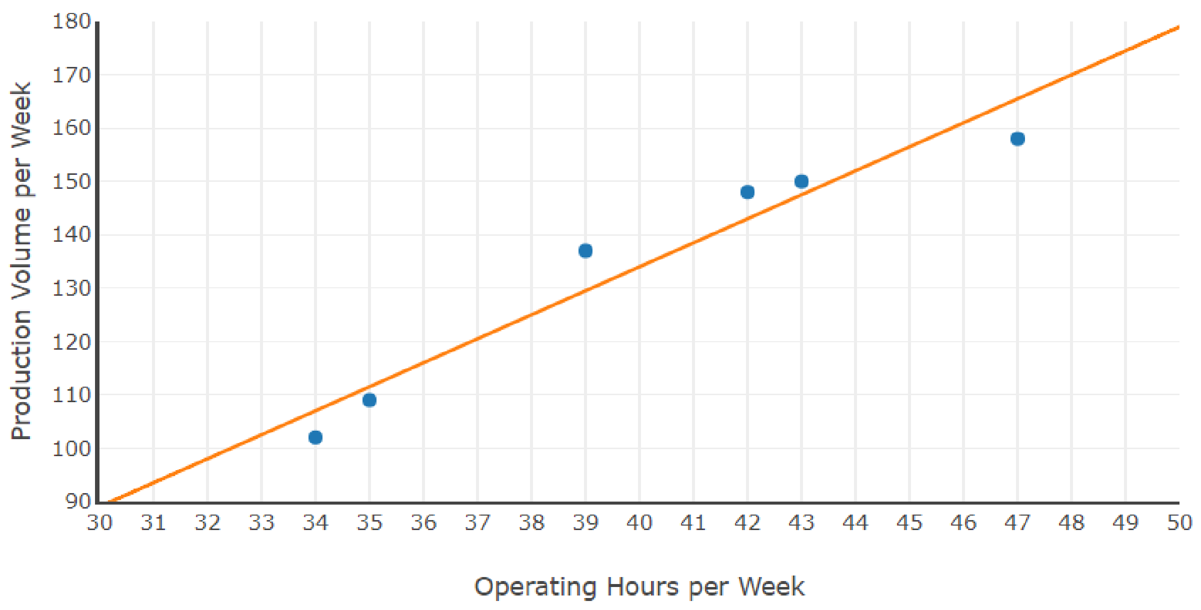
- The manager develops the following table:

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158

REGRESSION EXERCISE #1

- First, plot the data Is there a linear pattern?

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158



REGRESSION EXERCISE #1

$$\hat{y} = b_0 + b_1 x$$

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

- Run calculations:

	Production Hours (x)	Production Volume (y)	$(x-\bar{x})$	$(y-\bar{y})$	$(x-\bar{x})(y-\bar{y})$	$(x-\bar{x})^2$
	34	102	-6	-32	192	36
	35	109	-5	-25	125	25
	39	137	-1	3	-3	1
	42	148	2	14	28	4
	43	150	3	16	48	9
	47	158	7	24	168	49
$\bar{x}, \bar{y}$	**40**	**134**		Sum:	**558**	**124**
					$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

REGRESSION EXERCISE #1

$$\hat{y} = b_0 + b_1 x$$
$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$b_0 = \bar{y} - b_1\bar{x}$$

- Run calculations:

	Production Hours (x)	Production Volume (y)
	34	102
	35	109
	39	137
	42	148
	43	150
	47	158
$\bar{x}, \bar{y}$	**40**	**134**

$$b_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{558}{124} = \mathbf{4.5}$$

$$b_0 = \bar{y} - b_1\bar{x} = 134 - (4.5 \times 40) = \mathbf{-46}$$

Sum:	**558**	**124**
	$\Sigma(x-\bar{x})(y-\bar{y})$	$\Sigma(x-\bar{x})^2$

REGRESSION EXERCISE #1

Based on the formula, if the manager wants to produce 125 units per week, the plant should run for:

Production Hours (x)	Production Volume (y)
34	102
35	109
39	137
42	148
43	150
47	158

$$\hat{y} = b_0 + b_1 x$$

$$125 = -46 + 4.5x$$

$$x = \frac{171}{4.5} = \mathbf{38}\ hours\ per\ week$$

MULTIPLE REGRESSION

LINEAR VS MULTIPLE REGRESSION

- In linear regression we have one independent variable that may relate to a dependent variable with the formula

$$\hat{y} = b_0 + b_1 x$$

LINEAR VS MULTIPLE REGRESSION

- Multiple regression lets us compare several independent variables to one dependent variable at the same time.
- Each independent variable is assigned a subscript: x_1, x_2, x_3 etc.

LINEAR VS MULTIPLE REGRESSION

- The general formula is expanded:

linear regression

$$\hat{y} = b_0 + b_1 x$$

multiple regression

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots$$

- b_1 is the coefficient on x_1
- b_1 reflects the change in $\hat{y}$ for a given change in x_1, all else remaining constant

LINEAR VS MULTIPLE REGRESSION

- The formulas for coefficients also expand:

multiple regression

$$b_1 = \frac{\sum(x_2-\overline{x_2})^2 \sum(x_1-\overline{x_1})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_2-\overline{x_2})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - (\sum(x_1-\overline{x_1})(x_2-\overline{x_2}))^2}$$

$$b_2 = \frac{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_1-\overline{x_1})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - (\sum(x_1-\overline{x_1})(x_2-\overline{x_2}))^2}$$

$$b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2}$$

MULTIPLE REGRESSION

- For example, a used car lot may want to know what variables affect net profits
- They would create a list of predictors that might correlate with profit:

age
brand
price
color
style

MULTIPLE REGRESSION

- They would want to measure the correlation of each variable to net profit
- However, some predictors might correlate with each other:

price age color brand style

MULTIPLE REGRESSION

- The age of a car would have a direct impact on its sales price
- You can’t adjust one without affecting the other
- This is called multicollinearity

age

price

brand

color

style

REGRESSION EXERCISE #2

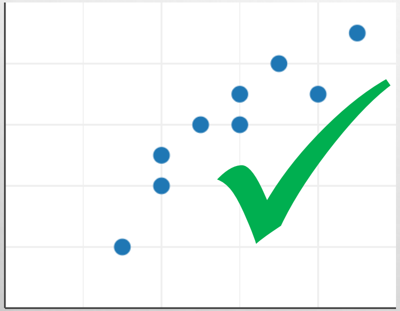
- A pharmacy delivers medications to the surrounding community.
- Drivers can make several stops per delivery.
- The owner would like to predict the length of time a delivery will take based on one or two related variables.

REGRESSION EXERCISE #2

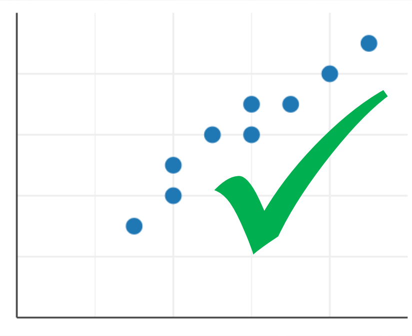
- First, consider what variables may have an effect on delivery time:
  - number of stops
  - driving distance
  - outside temperature
  - gasoline prices

REGRESSION EXERCISE #2

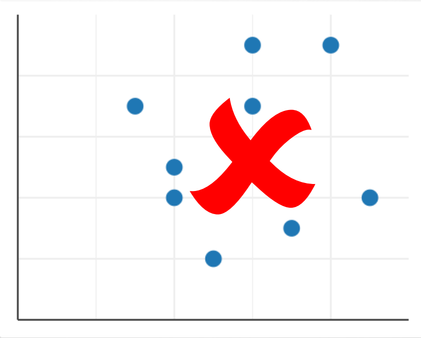
- Next, plot each variable against delivery time to see if there may be a relationship



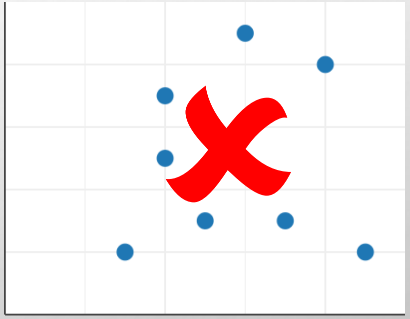
Time vs Distance



Time vs Stops



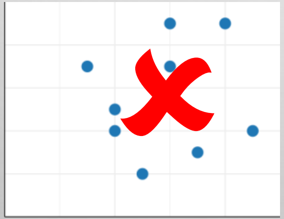
Time vs Temperature



Time vs Gas Price

REGRESSION EXERCISE #2

- Once we've chosen our variables x_1 and x_2, we'll usually test for multicollinearity
- We want to know if our two independent variables are closely related to each other
- If they are, it makes sense to discard one!



Stops vs Distance

A delivery might go to one customer that lives far away, or to a group of stops close by

REGRESSION EXERCISE #2

$y = Delivery\ Time\ (minutes)$
$x_1 = Number\ of\ Stops$
$x_2 = Distance\ (miles)$

y	x_1	x_2	$(y-\bar{y})$	$(x_1-\bar{x}_1)$	$(x_1-\bar{x}_1)^2$	$(x_2-\bar{x}_2)$	$(x_2-\bar{x}_2)^2$
29	**1**	**8**	-1	-1	1	2	4
31	**3**	**4**	1	1	1	-2	4
36	**2**	**9**	6	0	0	3	9
35	**3**	**6**	5	1	1	0	0
19	**1**	**3**	-11	-1	1	-3	9

$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$	$\Sigma(x_1-\bar{x}_1)^2$	$\Sigma(x_2-\bar{x}_2)^2$
30	**2**	**6**	**4**	**26**

$(x_1-\bar{x}_1)(y-\bar{y})$	$(x_2-\bar{x}_2)(y-\bar{y})$	$(x_1-\bar{x}_1)(x_2-\bar{x}_2)$
1	-2	-2
1	-2	-2
0	18	0
5	0	0
11	33	3
$\Sigma(x_1-\bar{x}_1)(y-\bar{y})$	$\Sigma(x_2-\bar{x}_2)(y-\bar{y})$	$\Sigma(x_1-\bar{x}_1)(x_2-\bar{x}_2)$
18	**47**	**-1**

REGRESSION EXERCISE #2

y = *Delivery Time (minutes)*
x_1 = *Number of Stops*
x_2 = *Distance (miles)*

$$b_1 = \frac{\sum(x_2-\overline{x_2})^2 \sum(x_1-\overline{x_1})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_2-\overline{x_2})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - \left(\sum(x_1-\overline{x_1})(x_2-\overline{x_2})\right)^2}$$

$$b_2 = \frac{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})(y-\bar{y}) - \sum(x_1-\overline{x_1})(x_2-\overline{x_2}) \sum(x_1-\overline{x_1})(y-\bar{y})}{\sum(x_1-\overline{x_1})^2 \sum(x_2-\overline{x_2})^2 - \left(\sum(x_1-\overline{x_1})(x_2-\overline{x_2})\right)^2}$$

$\bar{y}$	$\overline{x_1}$	$\overline{x_2}$
30	2	6

$\Sigma(x_1-\overline{x_1})^2$
4

$\Sigma(x_2-\overline{x_2})^2$
26

$\Sigma(x_1-\overline{x_1})(y-\bar{y})$	$\Sigma(x_2-\overline{x_2})(y-\bar{y})$	$\Sigma(x_1-\overline{x_1})(x_2-\overline{x_2})$
18	47	-1

REGRESSION EXERCISE #2

y = $Delivery\ Time\ (minutes)$
$x_1 = Number\ of\ Stops$
$x_2 = Distance\ (miles)$

$$b_1 = \frac{(26)(18) - (-1)(47)}{(4)(26) - ((-1))^2} = \frac{515}{103} = \mathbf{5}$$

$$b_2 = \frac{(4)\ (47) - (-1)(18)}{(4)(26) - ((-1))^2} = \frac{206}{103} = \mathbf{2}$$

$\bar{y}$	$\bar{x}_1$	$\bar{x}_2$
30	2	6

$\Sigma(x_1 - \bar{x}_1)^2$
4

$\Sigma(x_2 - \bar{x}_2)^2$
26

$\Sigma(x_1 - \bar{x}_1)(y - \bar{y})$	$\Sigma(x_2 - \bar{x}_2)(y - \bar{y})$	$\Sigma(x_1 - \bar{x}_1)(x_2 - \bar{x}_2)$
18	47	-1

REGRESSION EXERCISE #2

$y = Delivery\ Time\ (minutes)$
$x_1 = Number\ of\ Stops$
$x_2 = Distance\ (miles)$

$$\hat{y} = 8 + 5x_1 + 2x_2$$

$$b_1 = \frac{(26)(18) - (-1)(47)}{(4)(26) - ((-1))^2} = \frac{515}{103} = \mathbf{5}$$

$$b_2 = \frac{(4)\ (47) - (-1)(18)}{(4)(26) - ((-1))^2} = \frac{206}{103} = \mathbf{2}$$

$$b_0 = \bar{y} - b_1\overline{x_1} - b_2\overline{x_2}$$

$$= 30 - (5)(2) - (2)(6)$$

$$= 30 - 10 - 12 = \mathbf{8}$$

$\bar{y}$	$\overline{x_1}$	$\overline{x_2}$
30	2	6

$\Sigma(x_1 - \overline{x_1})^2$
4

$\Sigma(x_2 - \overline{x_2})^2$
26

$\Sigma(x_1 - \overline{x_1})(y - \bar{y})$	$\Sigma(x_2 - \overline{x_2})(y - \bar{y})$	$\Sigma(x_1 - \overline{x_1})(x_2 - \overline{x_2})$
18	47	-1

REGRESSION EXERCISE #2

$y = Delivery\ Time\ (minutes)$
$x_1 = Number\ of\ Stops$
$x_2 = Distance\ (miles)$

$$\hat{y} = 8 + 5x_1 + 2x_2$$

y	x_1	x_2
29	1	8
31	3	4
36	2	9
35	3	6
19	1	3

D ON OUR ANALYSIS, PHARMACY DELIVERIES HAVE A FIXED TIME OF 8 MINUTES, PLUS 5

S FOR EACH STOP,

2 MINUTES FOR EACH MILE TRAVELED

IMPLEMENTATION IN R

USAGE OF LM()

Syntax: lm(formula, data)

Example:

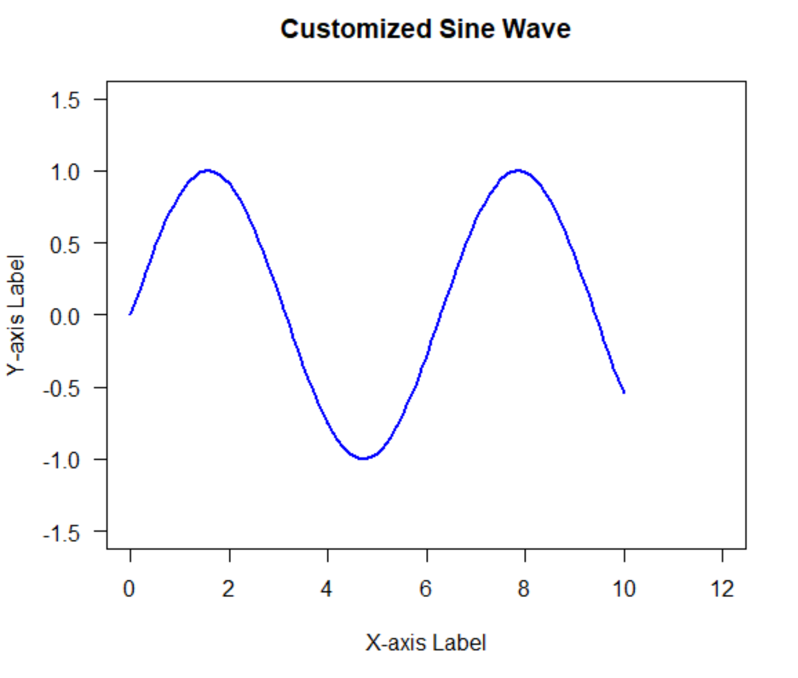
lm(y~x, data=dataset)

```
R 4.4.1 · ~/
> data(mtcars)
> # Example dataset
> model <- lm(mpg ~ wt + hp, data = mtcars)
> summary(model)

Call:
lm(formula = mpg ~ wt + hp, data = mtcars)

Residuals:
   Min     1Q Median     3Q 
-3.941 -1.600 -0.182  1.050 
   Max 
 5.854 

Coefficients:
            Estimate
(Intercept) 37.22727
wt          -3.87783
hp          -0.03177
            Std. Error
(Intercept)    1.59879
wt             0.63273
hp             0.00903
```

ADVANCED PLOTTING

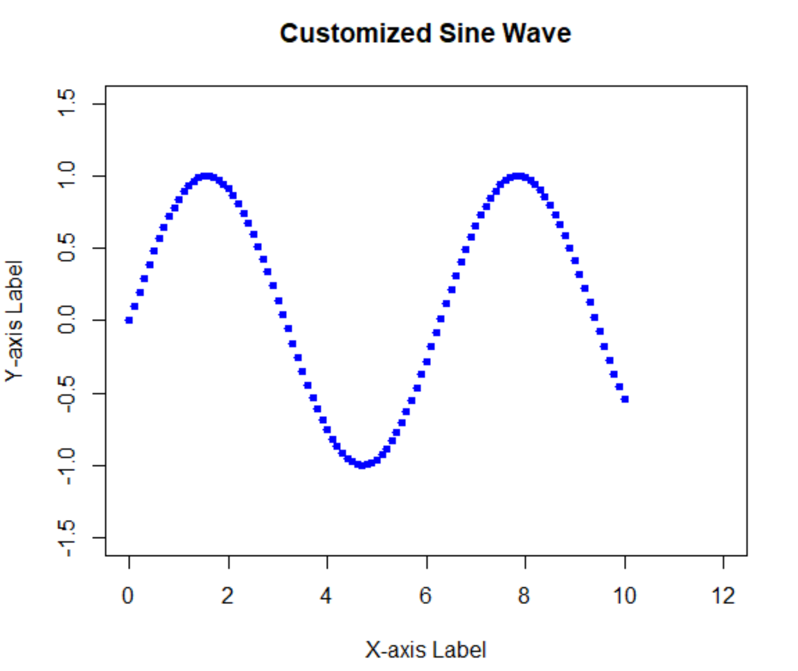
SIMPLE PLOT

```r
# Generate sample data
x <- seq(0, 10, by = 0.1)
y <- sin(x)

# Basic plot with customization
plot(
  x, y,
  type = "l",                # Line plot
  col = "blue",              # Line color
  lwd = 2,                   # Line width
  main = "Customized Sine Wave",  # Title
  xlab = "X-axis Label",     # X-axis label
  ylab = "Y-axis Label",     # Y-axis label
  xlim = c(0, 12),           # X-axis limits
  ylim = c(-1.5, 1.5),       # Y-axis limits
  las = 1                    # Rotate axis labels
)
```



USING PCH

```r
# Generate sample data
x <- seq(0, 10, by = 0.1)
y <- sin(x)

# Basic plot with customization
plot(
  x, y,
  type = "p",                # Line plot
  col = "blue",              # Line color
  lwd = 2,                   # Line width
  main = "Customized Sine Wave",  # Title
  xlab = "X-axis Label",     # X-axis label
  ylab = "Y-axis Label",     # Y-axis label
  xlim = c(0, 12),           # X-axis limits
  ylim = c(-1.5, 1.5),       # Y-axis limits
  pch = 20
)


```

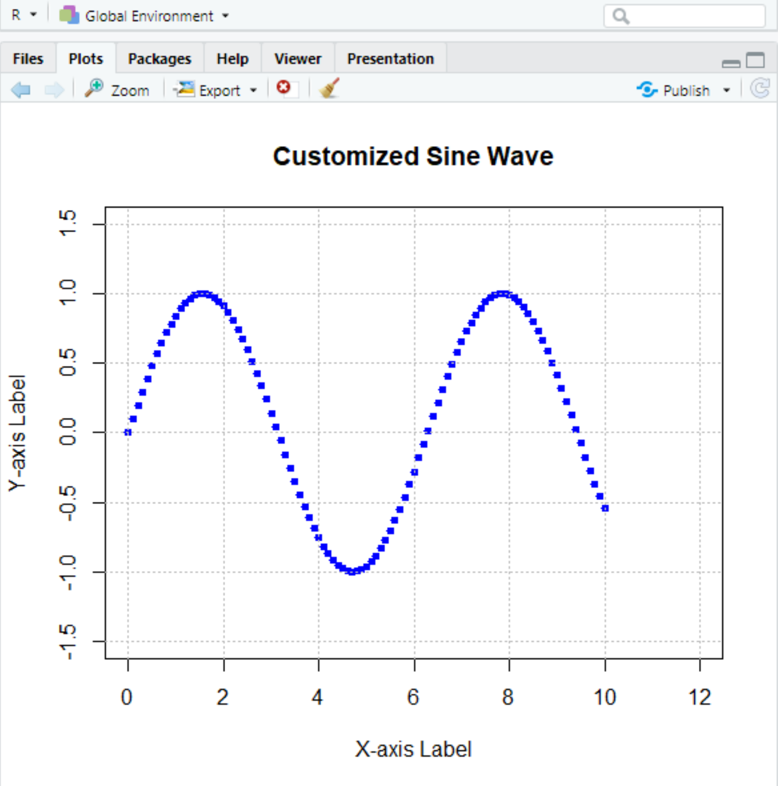


ADDING GRID LINES

```r

# Adding grid lines
grid(col = "gray", lty = "dotted", lwd = 0.5)
```

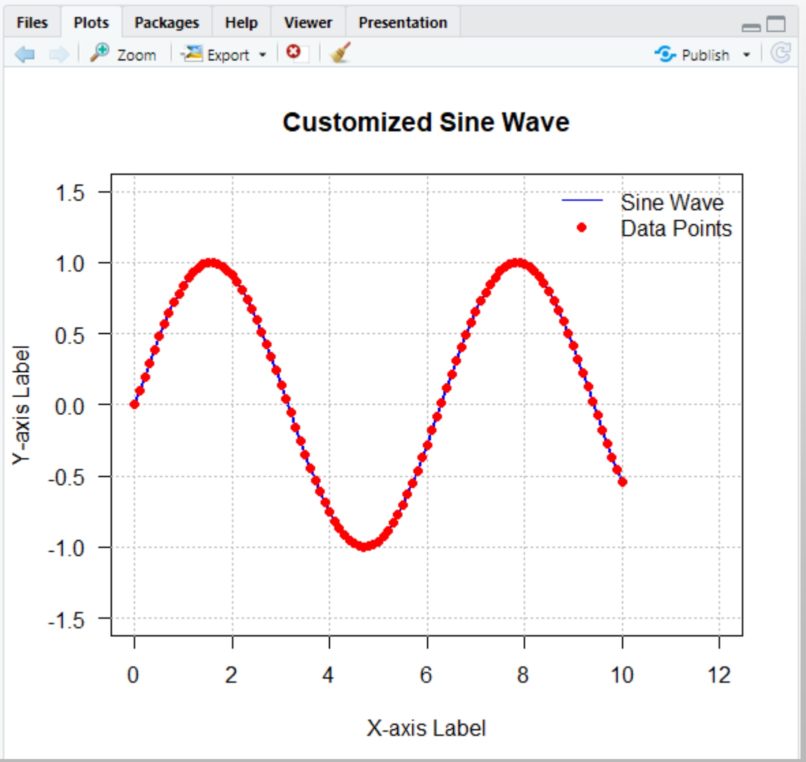
3:46 (Top Level) R Script

Console



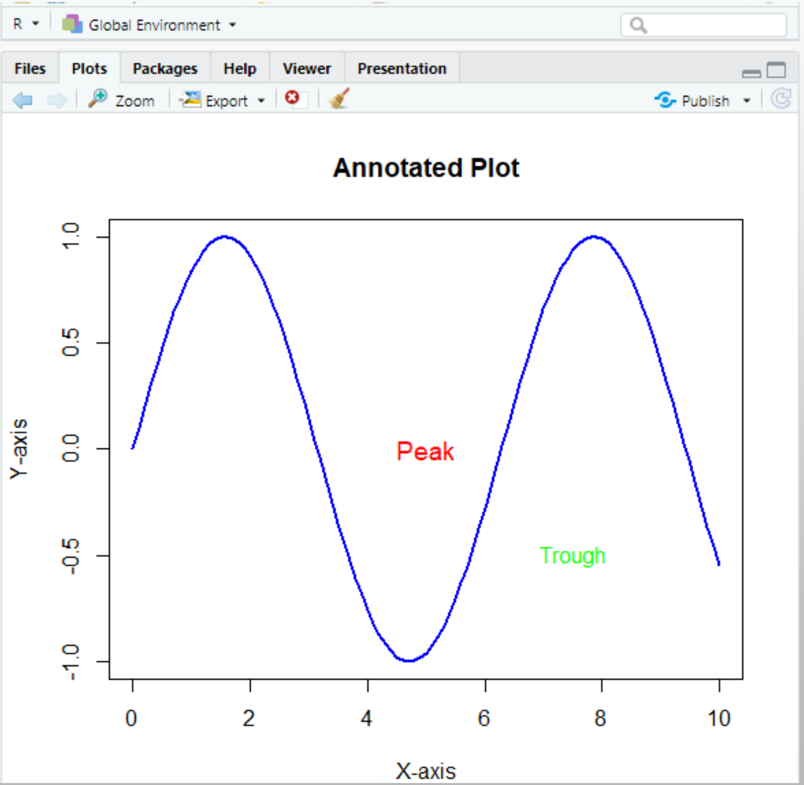
ADDING A LEGEND

```r
# Adding a legend
legend(
  "topright",                      # Position
  legend = c("Sine Wave", "Data Points"),  # Labels
  col = c("blue", "red"),          # Colors
  lty = c(1, NA),                  # Line type (solid for sin
  pch = c(NA, 16),                 # Point type
)

```



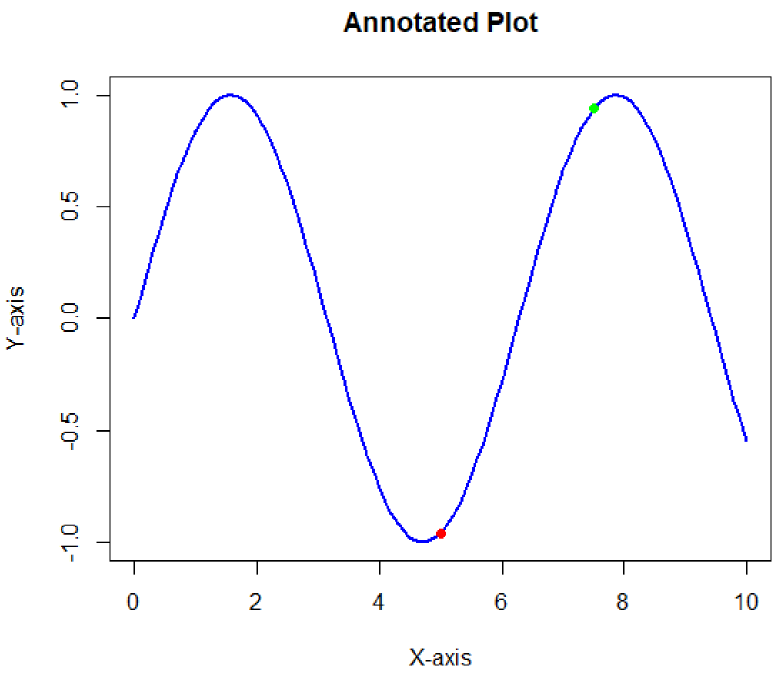
ADDING A TEXT

```r

# Add text annotations

text(5, 0, "Peak", col = "red", cex = 1.2)
# Add text at (7.5, -0.5)
text(7.5, -0.5, "Trough", col = "green", cex = 1)
```



LABEL SPECIFIC POINTS

```r
# Label specific points

# Highlight point at x = 5
points(5, sin(5), col = "red", pch = 16)
# Highlight point at x = 7.5
points(7.5, sin(7.5), col = "green", pch = 16)
```



COLORS()

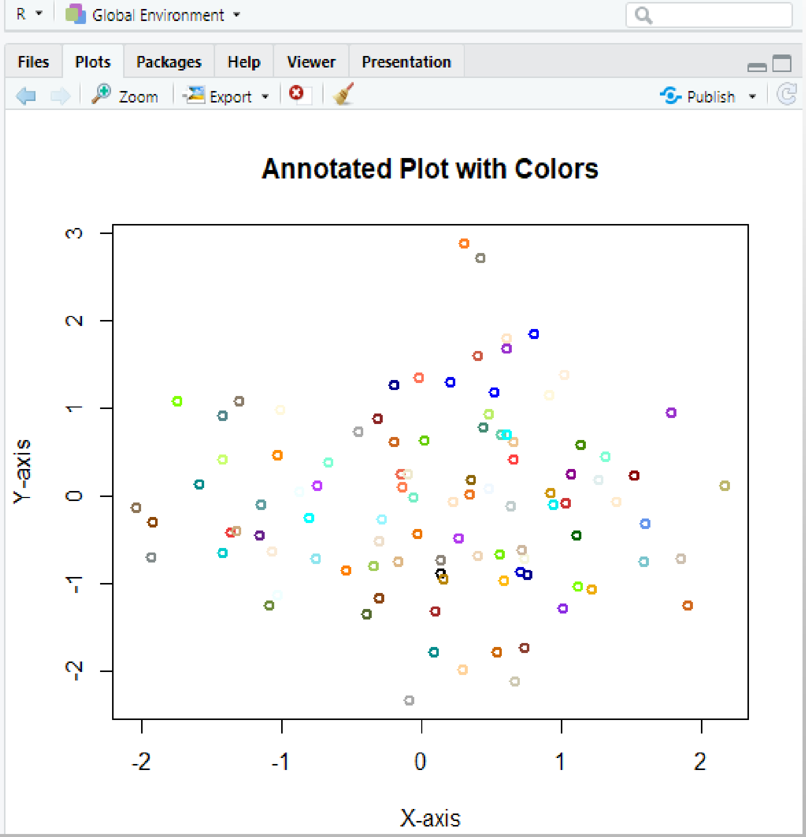
```
> colors(distinct = TRUE)
  [1] "white"
  [2] "aliceblue"
  [3] "antiquewhite"
  [4] "antiquewhite1"
  [5] "antiquewhite2"
  [6] "antiquewhite3"
  [7] "antiquewhite4"
  [8] "aquamarine"
  [9] "aquamarine2"
 [10] "aquamarine3"
 [11] "aquamarine4"
 [12] "azure"
 [13] "azure2"
 [14] "azure3"
 [15] "azure4"
 [16] "beige"
 [17] "bisque"
 [18] "bisque2"
 [19] "bisque3"
 [20] "bisque4"
 [21] "black"
 [22] "blanchedalmond"
```

THE COLORS() FUNCTION GENERATES ALL BUILT IN COLORS IN R

THIS CAN BE USED LIKE:

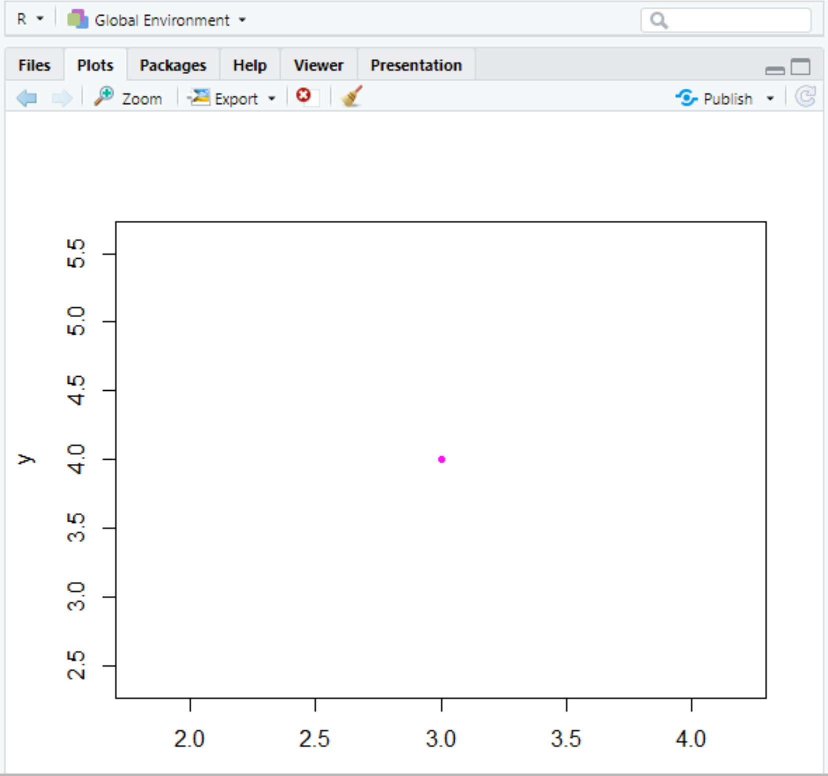
```r
# Generate sample data
x <- rnorm(100)
y <- rnorm(100)

# Basic plot with customized colors
plot(
  x, y, type = "p", col = colors(), lwd = 2,  # Hex cod
  main = "Annotated Plot with Colors", xlab = "X-axis",
)
```

7:35 (Top Level) R Script



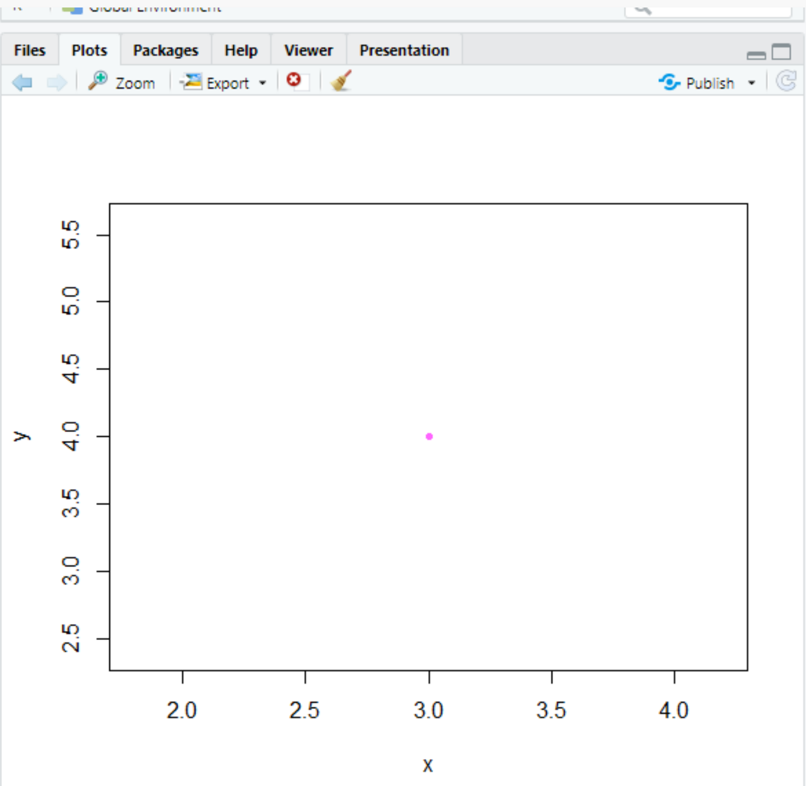
YOU CAN ALSO USE HEX CODE TO COLOR

```r
# Generate sample data
x <- 3
y <- 4

# Basic plot with customized colors
plot(
  x, y, type = "p", col = "#FF11F3", pch=20
)
```



YOU CAN ALSO USE RGB()

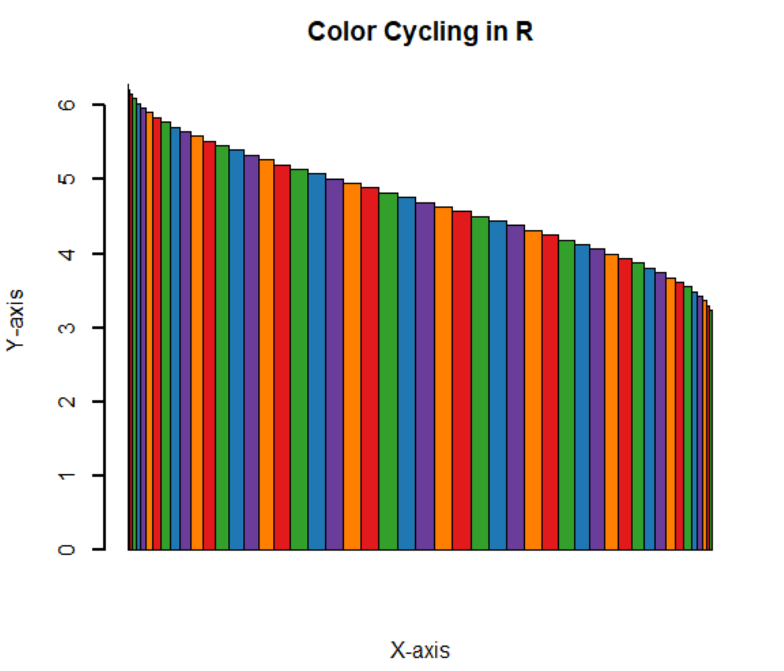
```
> rgb(1,0.4,1)
[1] "#FF66FF"
> plot(x,y,col=rgb(1,0.4,1),pch=20)
> |
```



COLOR CYCLING IN R

```
# Create a custom palette
custom_colors <- c("#1f78b4",
                   "#33a02c",
                   "#e31a1c",
                   "#ff7f00",
                   "#6a3d9a")  # Hex colors
# Plot with color cycling
barplot(
  x, y, type = "l", lty = 1, lwd = 2,
  col = custom_colors,  # Use custom color palette
  xlab = "X-axis", ylab = "Y-axis", main = "Color Cycl
)
```

12:2 (Top Level) R Script

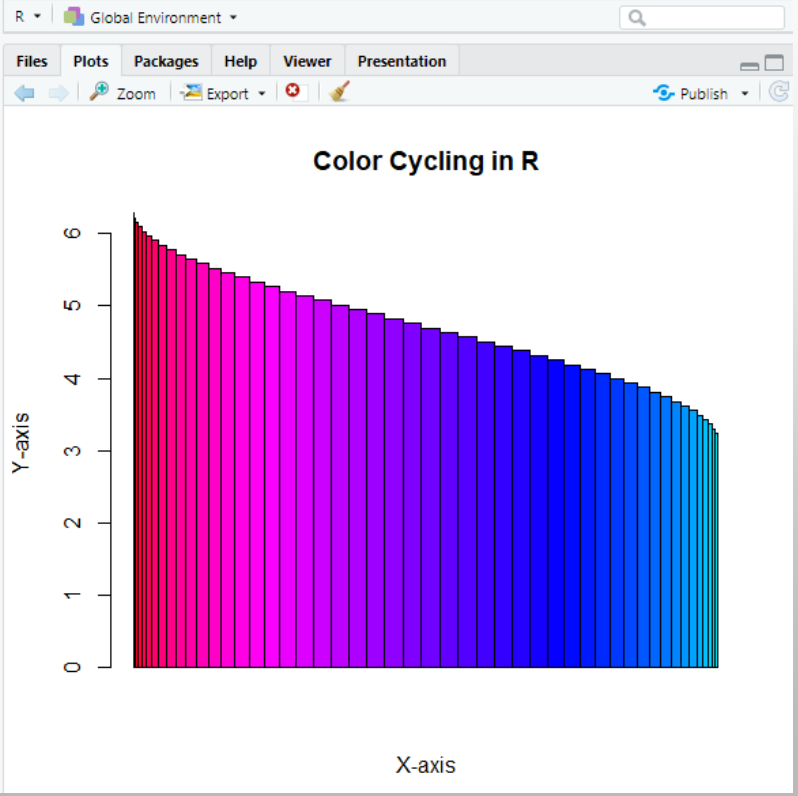


USING COLOR PALETTES – RAINBOW(N)

```r

# Plot with color palette
barplot(
  x, y, col = rainbow(100),
  xlab = "X-axis", ylab = "Y-axis",
  main = "Color Cycling in R"
)
```

7:2 (Top Level) R Script

Console

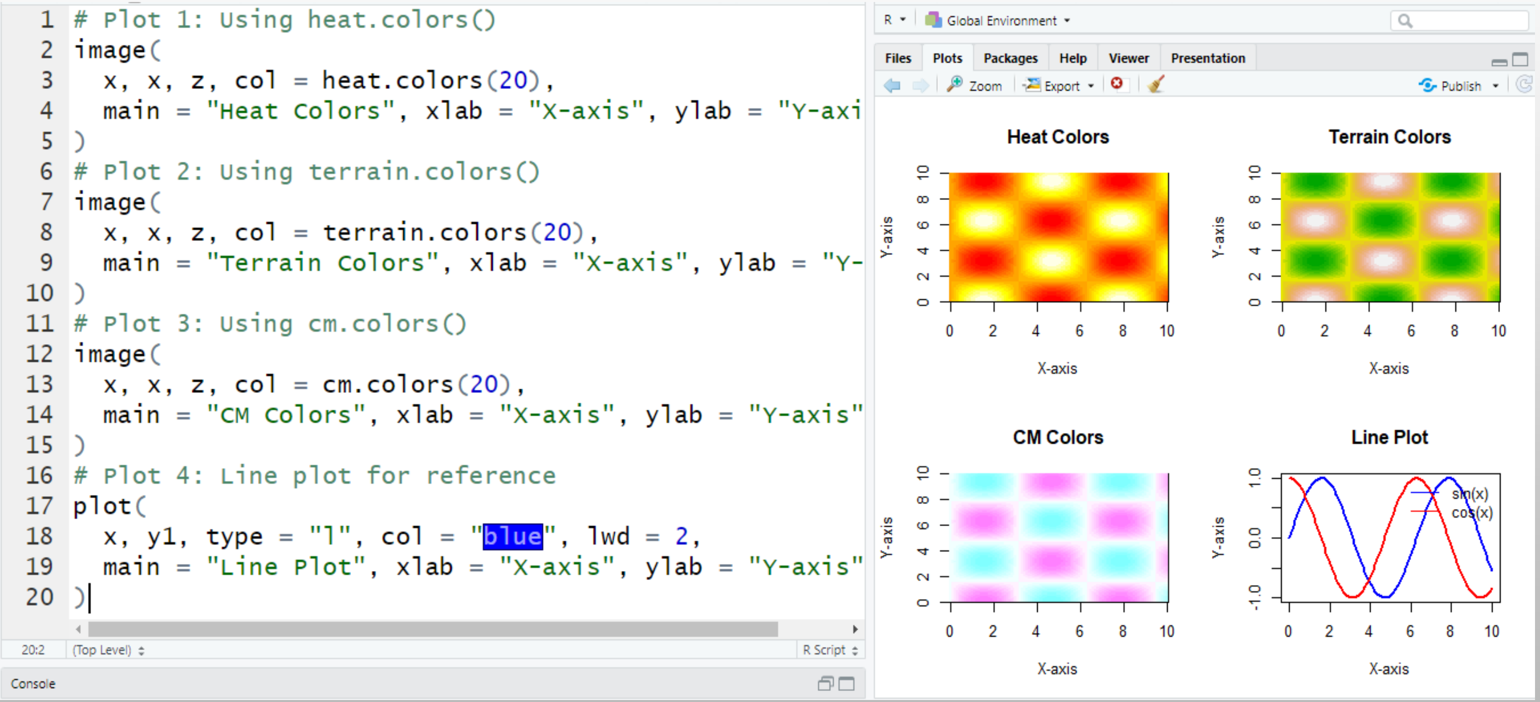


OTHER COLOR PALETTES

heat.colors()

terrain.colors()

cm.colors()



3D SCATTER PLOT

```r
# Install the scatterplot3d package
install.packages("scatterplot3d")

# Load the package
library(scatterplot3d)

# Generate random data
x <- rnorm(50)
y <- rnorm(50)
z <- rnorm(50)

# Create a 3D scatterplot
scatterplot3d(
  x, y, z, pch = 16, color = "blue",
  main = "3D Scatterplot", xlab = "X-axis",
  ylab = "Y-axis", zlab = "Z-axis"
)

```