DSA1101

Introduction to Data Science

Week 4 Tutorial 2

TODAY'S AGENDA

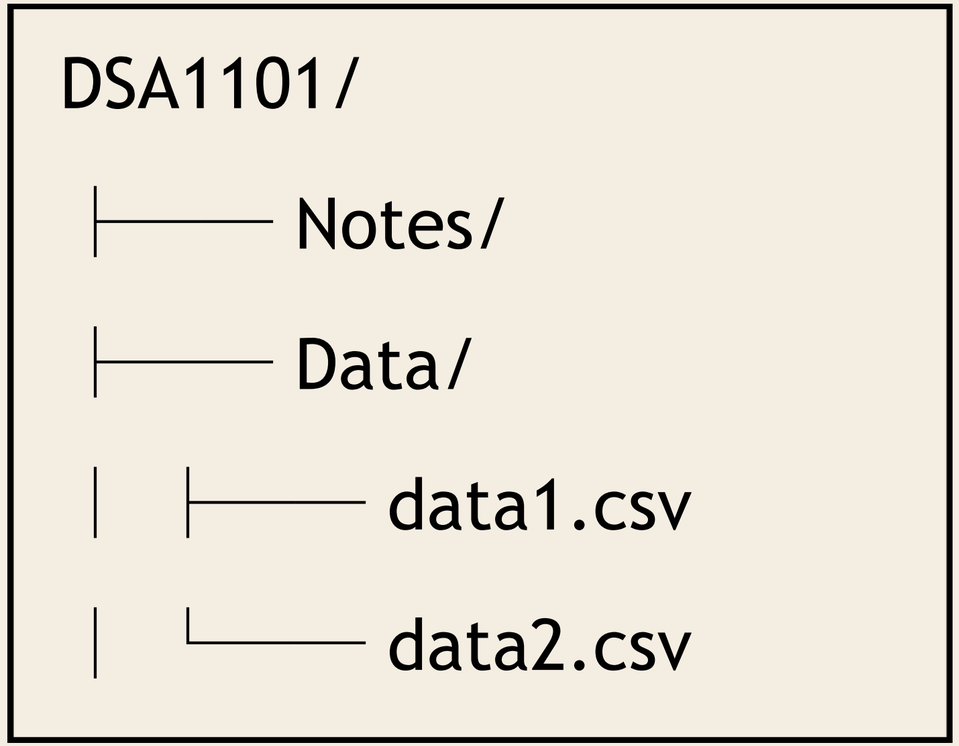
1. Quick Revision
2. On-site Questions Attempt
3. On-site Questions Discussion
4. Off-site Questions Discussion

QUICK REVISION

Importing Data in R

Key to remember!

1. Ensure your working directory is correct → check using getwd()
2. Import files by their **relative** path to the working directory



Eg. setwd(“.../DSA1101”)

read.csv(“Data/data1.csv”)

METHOD 1

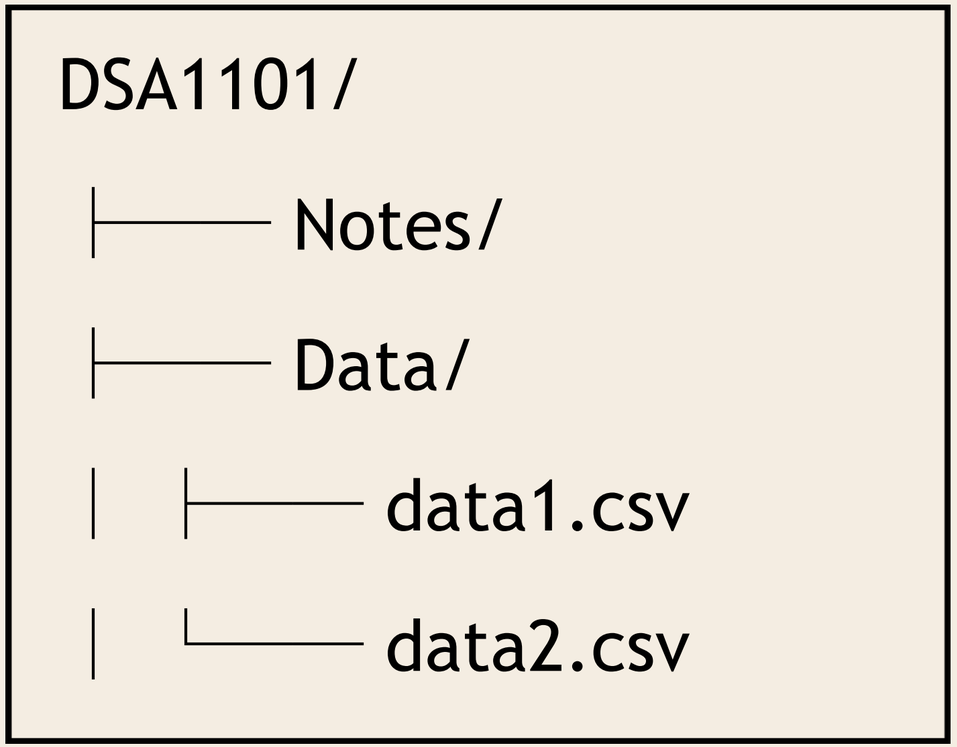
SETTING WORKING DIRECTORY

For Windows:

```
setwd("C:/Users/.../DSA1101/Data")
```

For Mac:

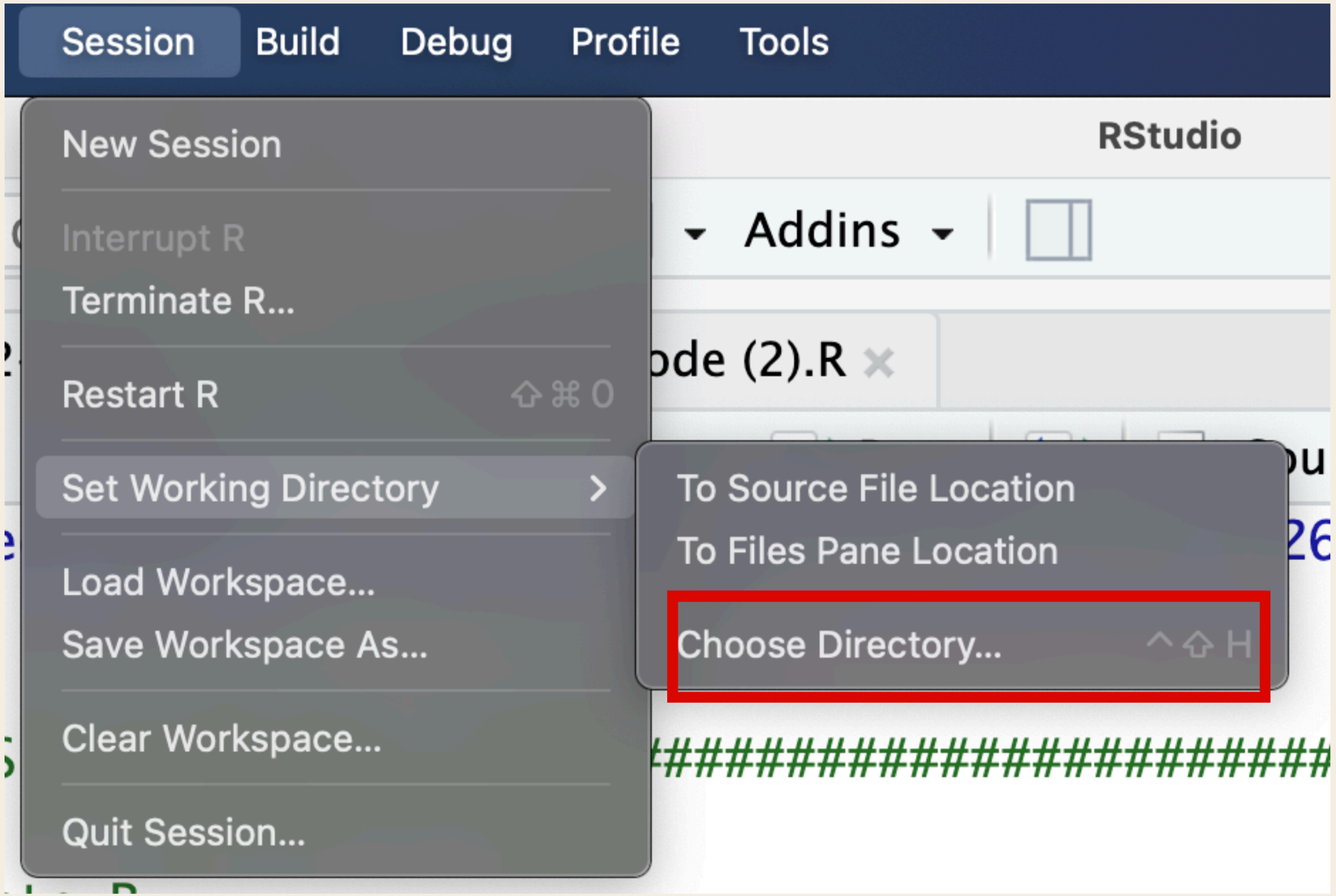
```
setwd("~/DSA1101/Data")
```




```
read.csv("data1.csv")
```

METHOD 2

SETTING WORKING DIRECTORY



read.csv() function

```
df = read.csv(...)
```

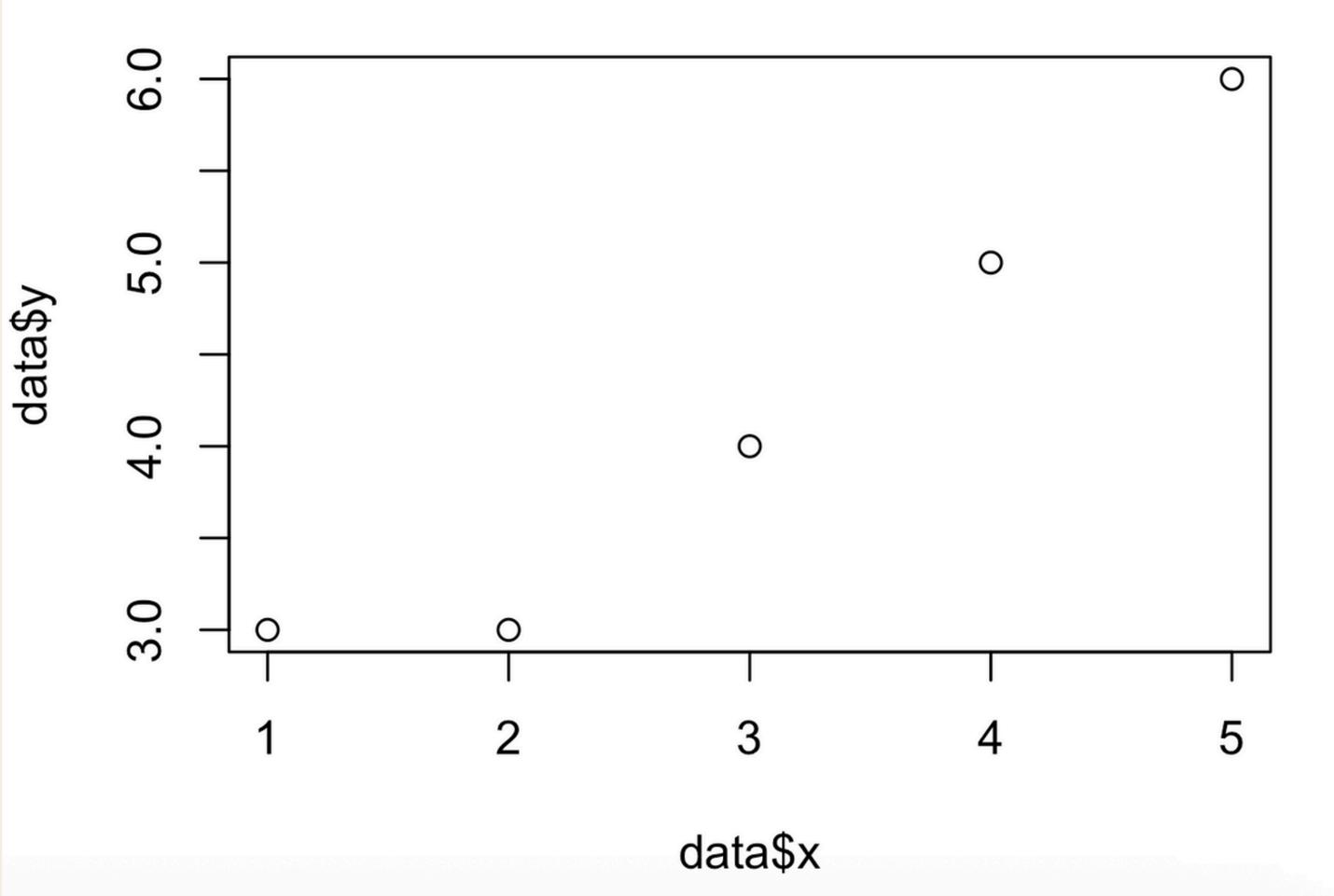
Imports file into a dataframe in R

BY DEFAULT, READ.CSV ASSUMES:

1. Values are separated by commas (sep = ‘,’)
2. Data contains headers (header = TRUE)

plot() function

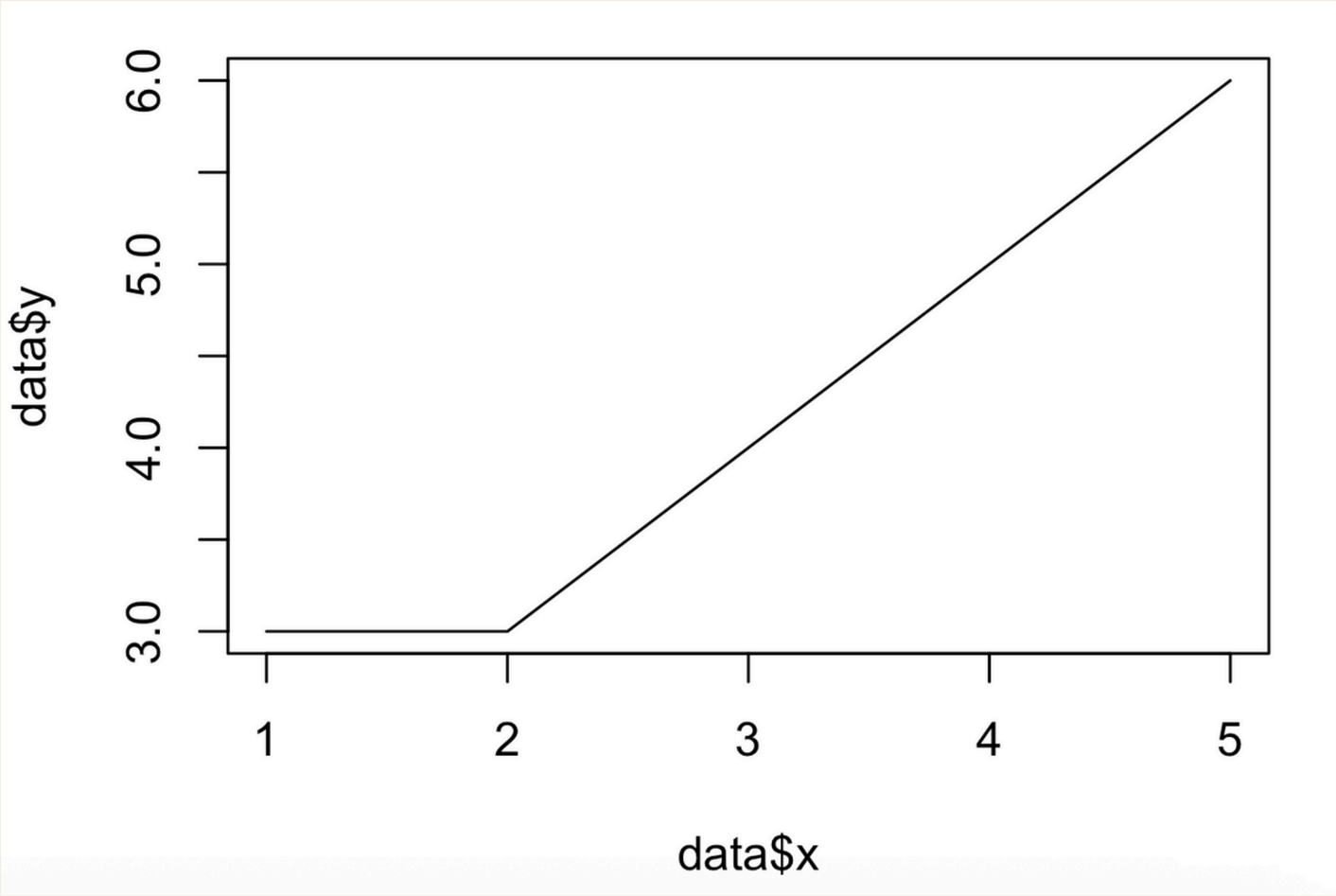
```
data = data.frame(
  x = c(1,2,3,4,5),
  y = c(3,3,4,5,6)
)

# Method 1: plot(x, y)
plot(data$x, data$y)

# Method 2: plot(y ~ x)
plot(data$y ~ data$x)
```



plot scatterplot

plot() function

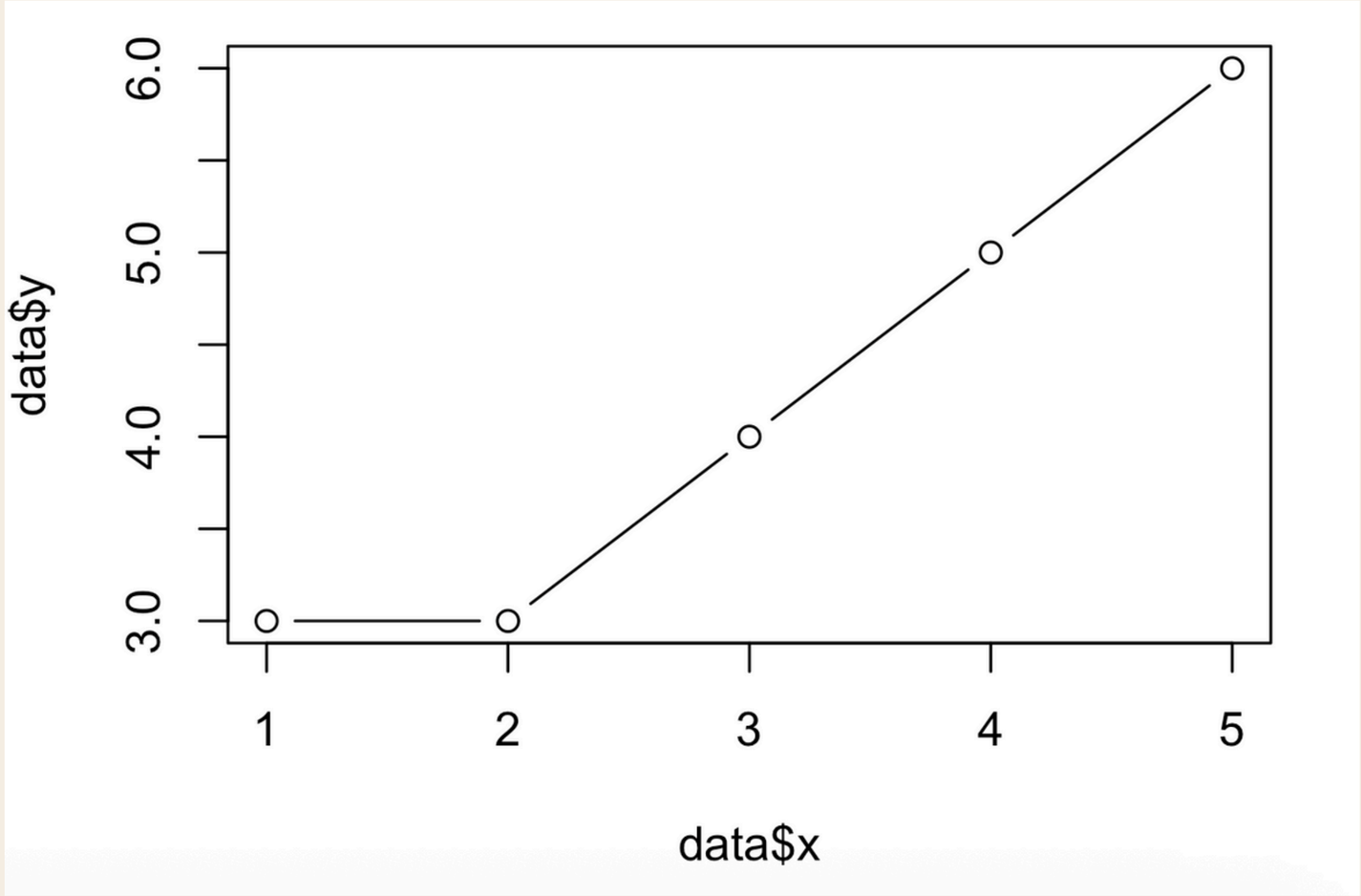
```
data = data.frame(
  x = c(1,2,3,4,5),
  y = c(3,3,4,5,6)
)

plot(data$x, data$y,
     type = "l"
)
```



plot line graph

plot() function

```
data = data.frame(
   x = c(1,2,3,4,5),
   y = c(3,3,4,5,6)
)

plot(data$x, data$y,
      type = "b"
)
```

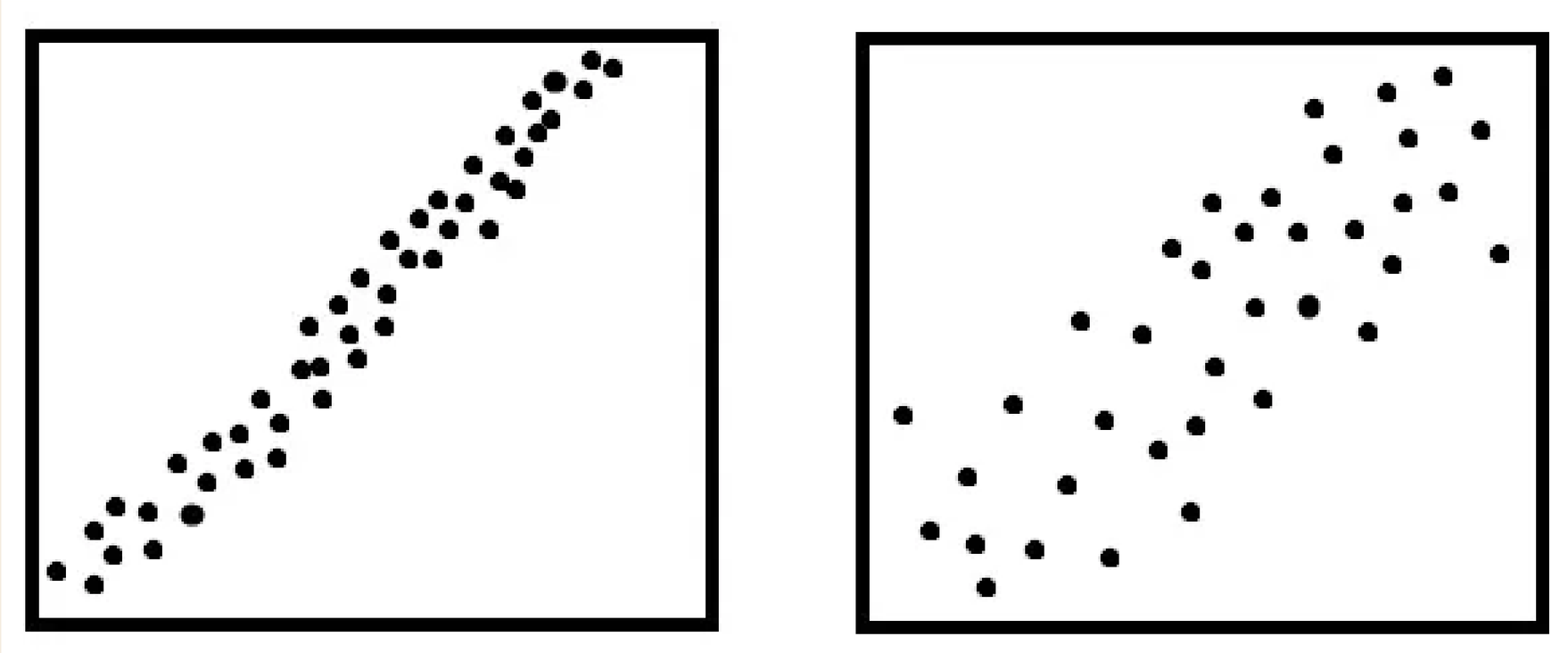


Interpreting Scatterplot

1. Is there any relationship? Is it strong?
2. If there is, is it positive or negative?
3. Relationship is linear or non-linear?
4. Special observations (Outliers)?
5. Is the variability of the response stable when x changes?

Interpreting Scatterplot

1. Is there any relationship? Is it strong?

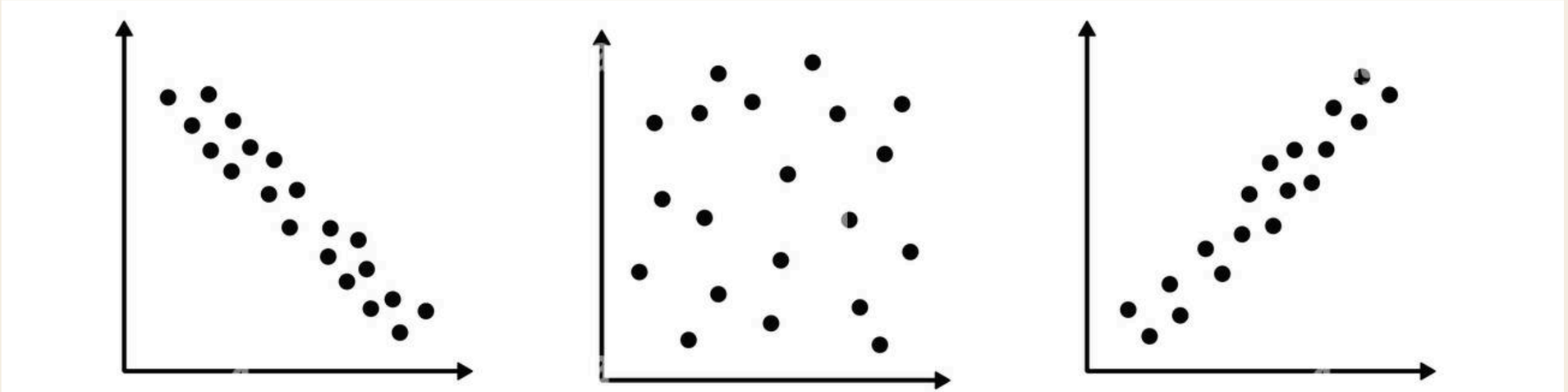


strong

weak

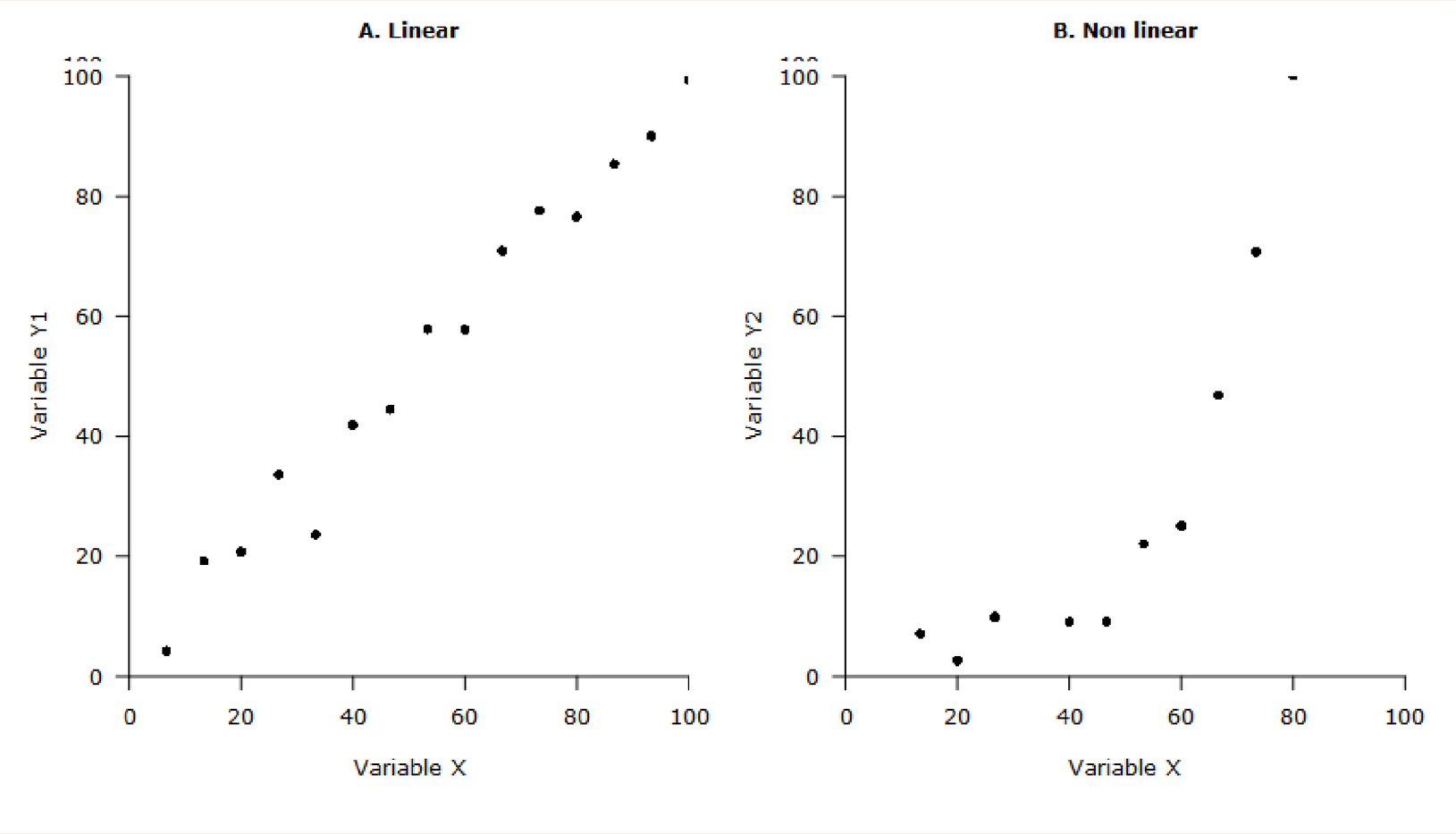
Interpreting Scatterplot

2. If there is, is it positive or negative?



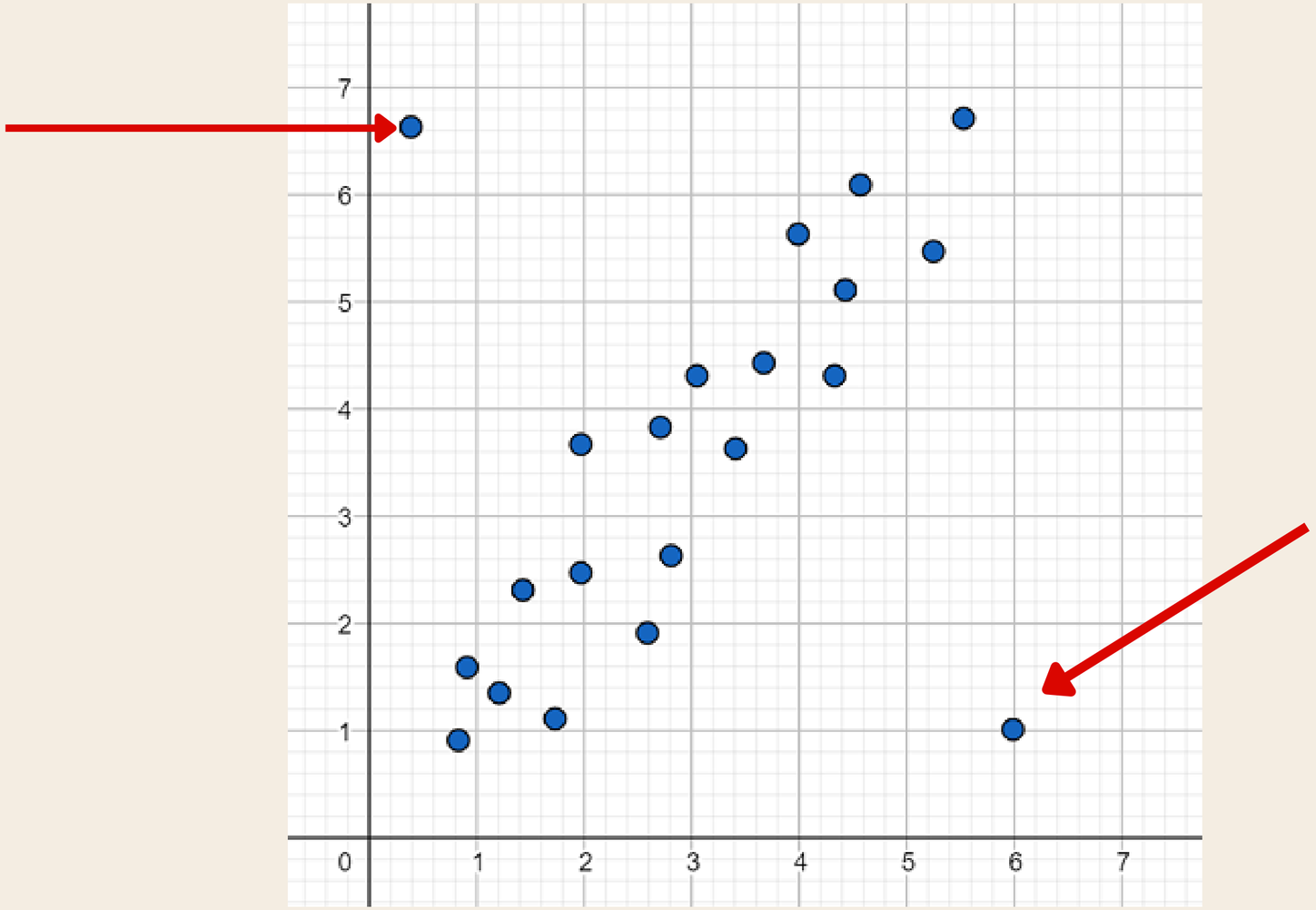
Interpreting Scatterplot

3. Relationship is linear or non-linear?



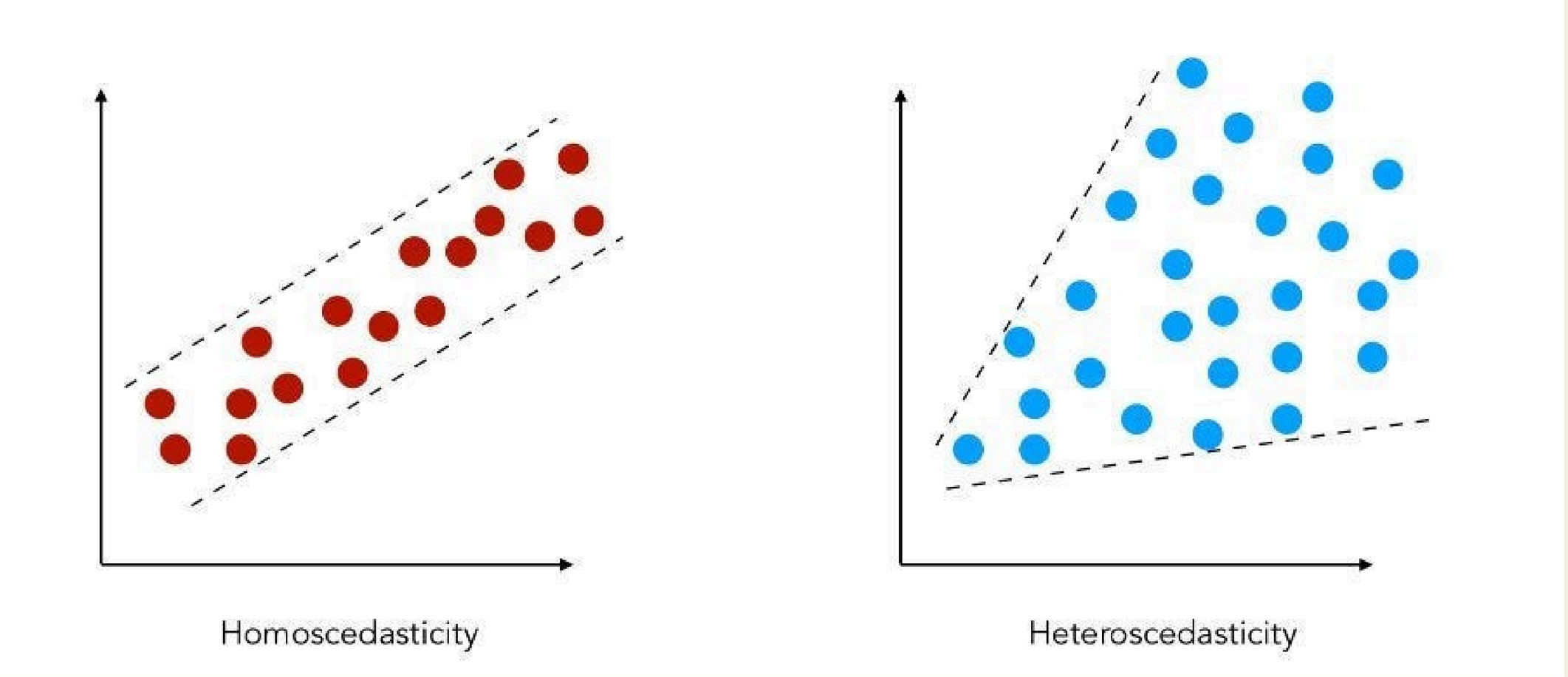
Interpreting Scatterplot

4. Special observations (Outliers)?



Interpreting Scatterplot

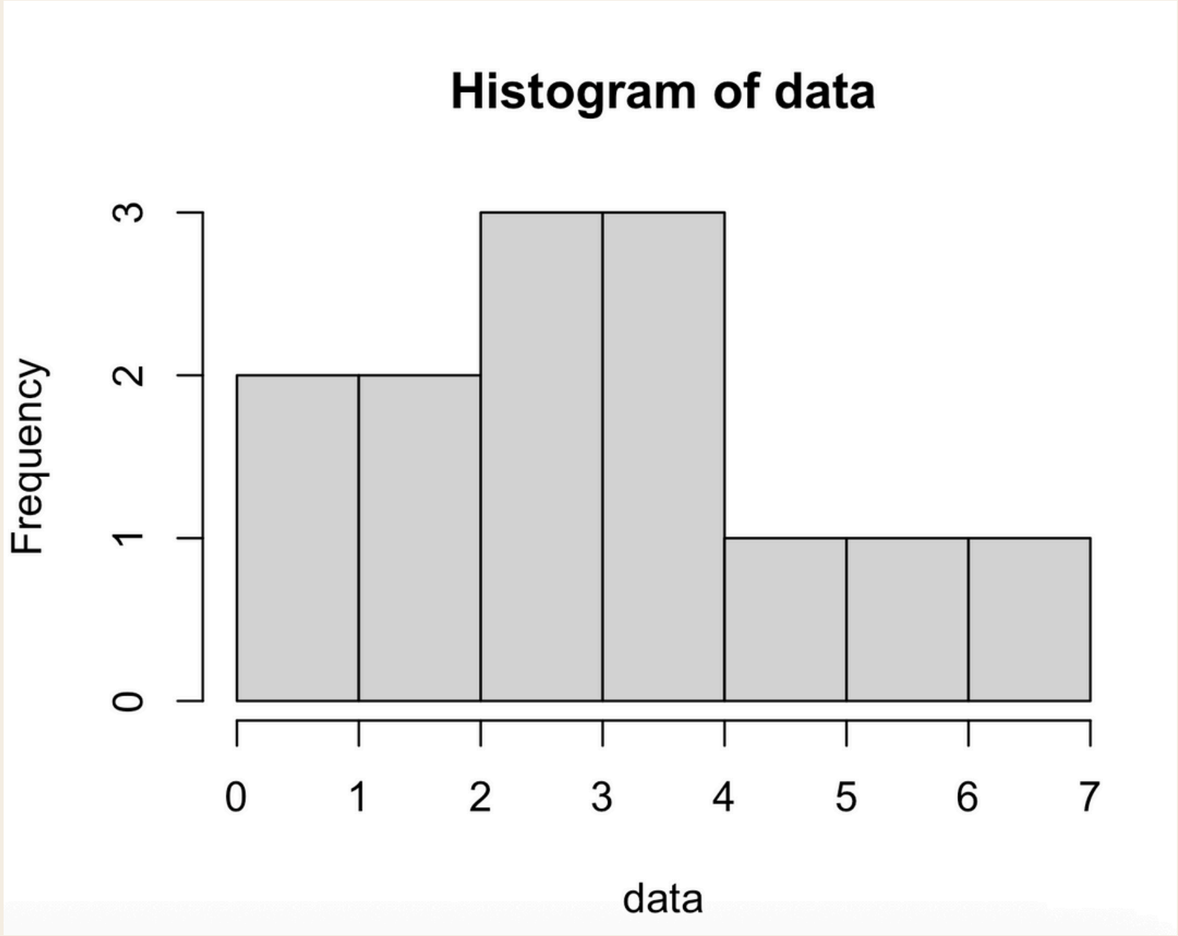
5. Is the variability of the response stable when x changes?



hist() function

```
data = c(0,1,2,2,3,3,3,4,4,4,5,6,7)

hist(data)
```

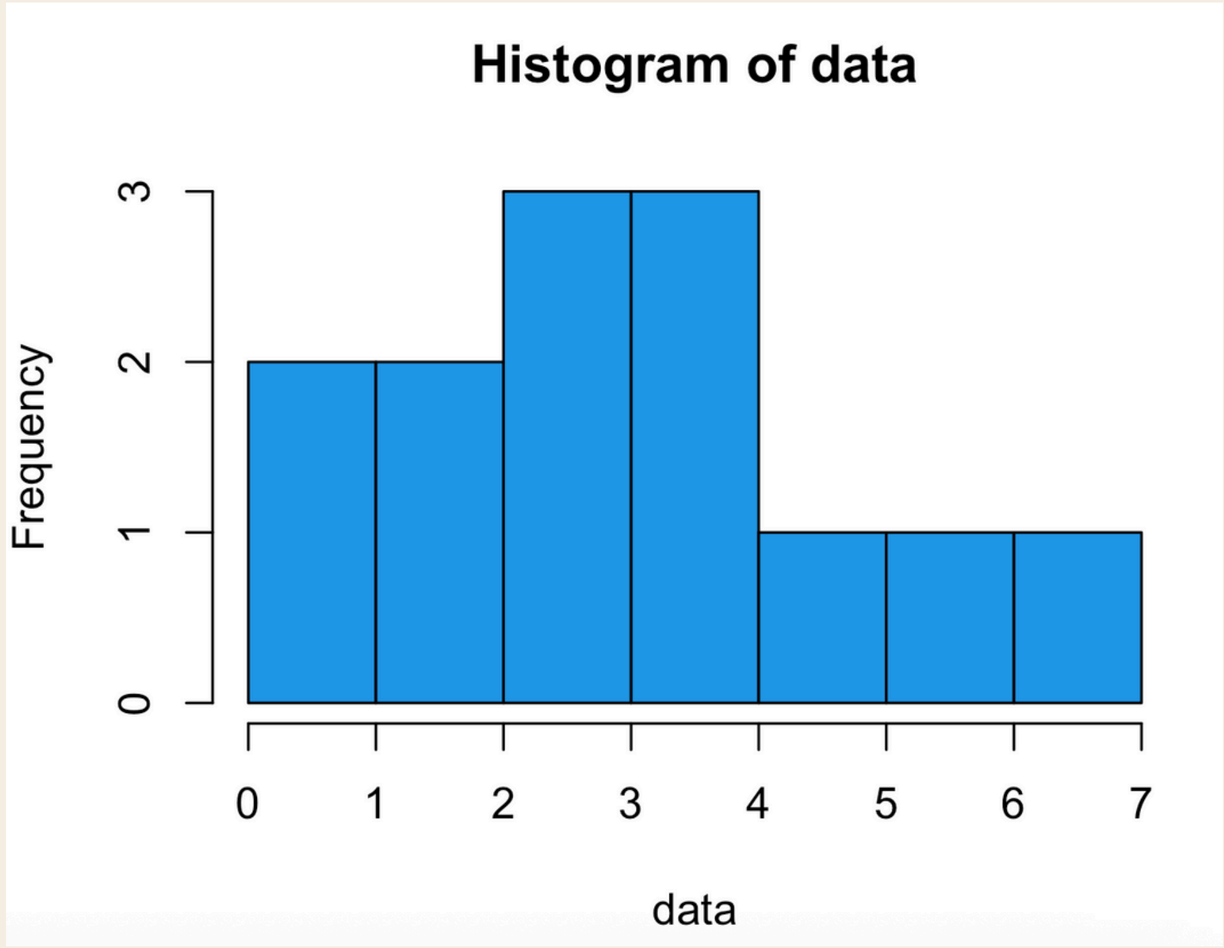
hist() creates a histogram



hist() function

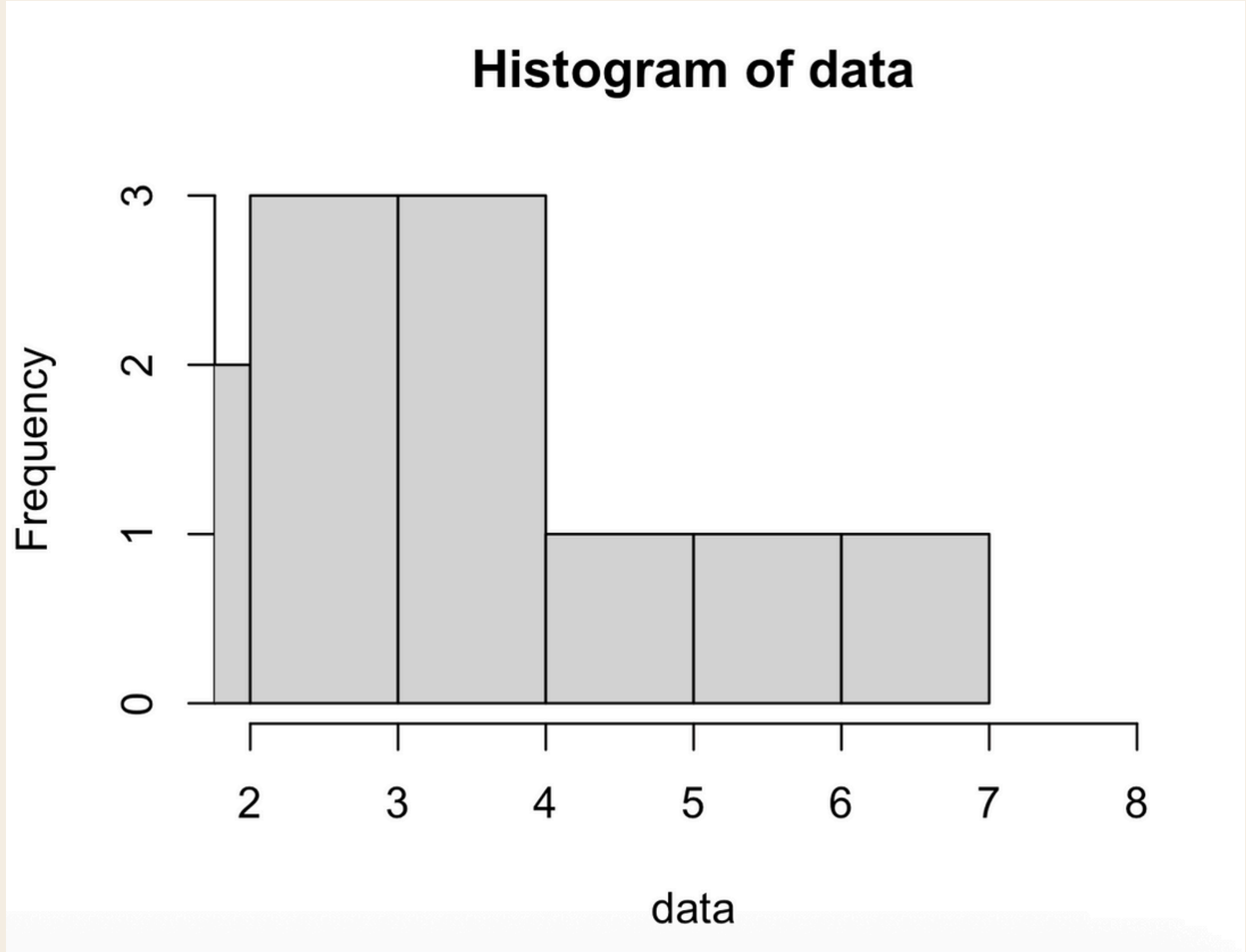
```
data = c(0,1,2,2,3,3,3,4,4,4,5,6,7)

hist(data, col = 20)
```

col: to change colour



hist() function

```
data = c(0,1,2,2,3,3,3,4,4,4,5,6,7)

hist(data, xlim = c(2,8))
```

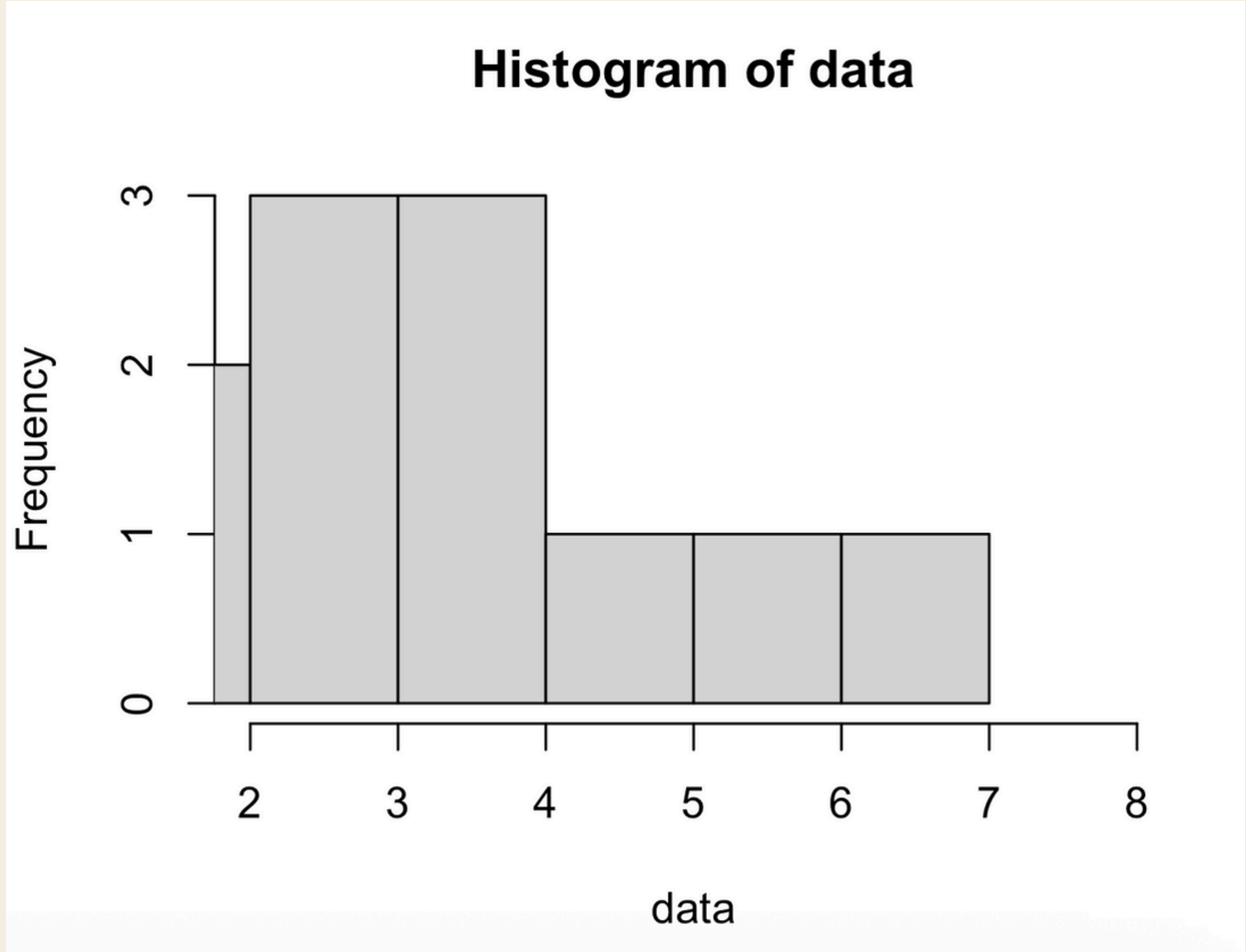


xlim: set x-axis limits

hist() function

```
data = c(0,1,2,2,3,3,3,4,4,4,5,6,7)

hist(data, ylim = c(0,4))
```

ylim: set y-axis limits

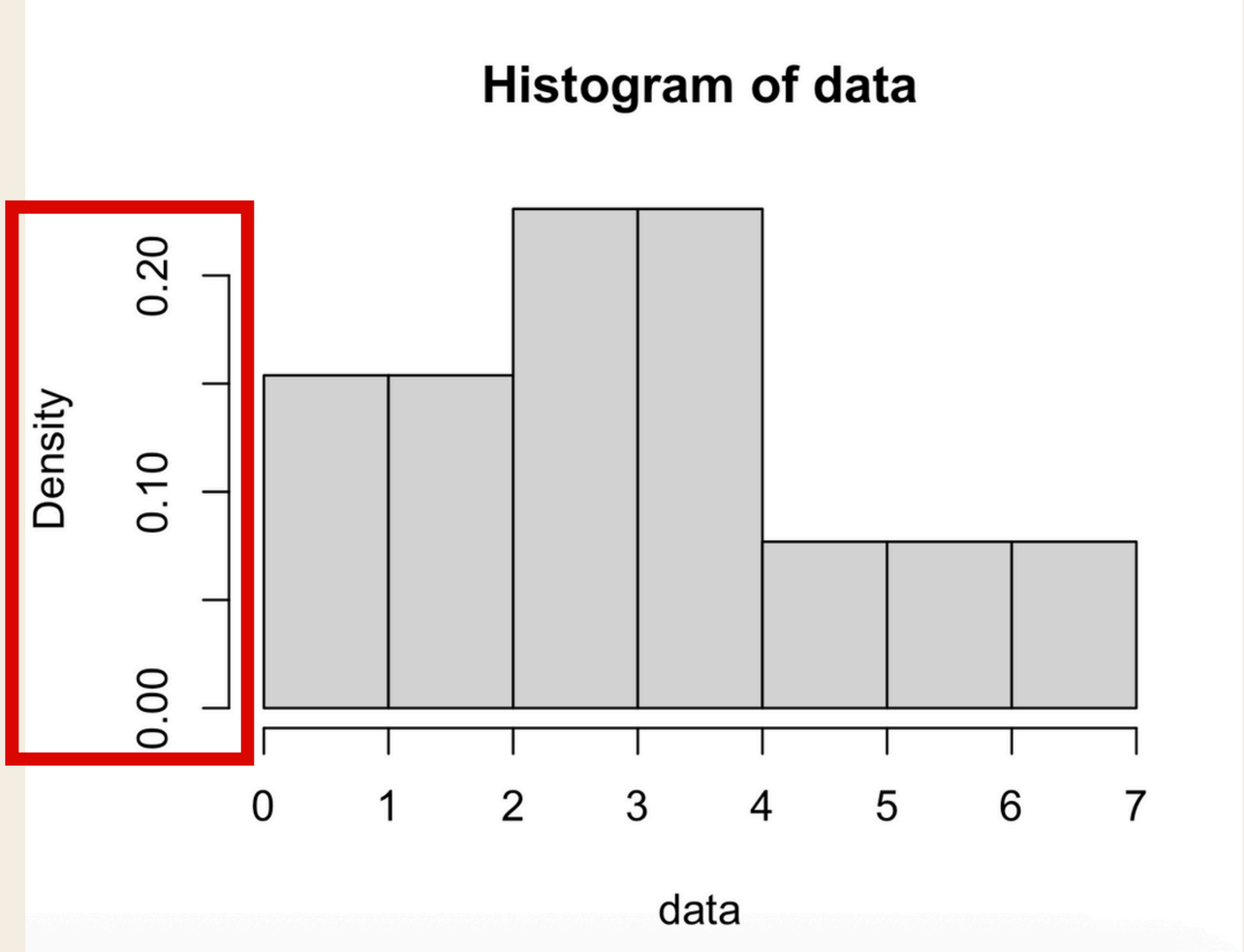


hist() function

```
data = c(0,1,2,2,3,3,3,4,4,4,5,6,7)

hist(data, freq = FALSE)
```

default: freq = TRUE

freq = FALSE plots densities instead of freq (counts)

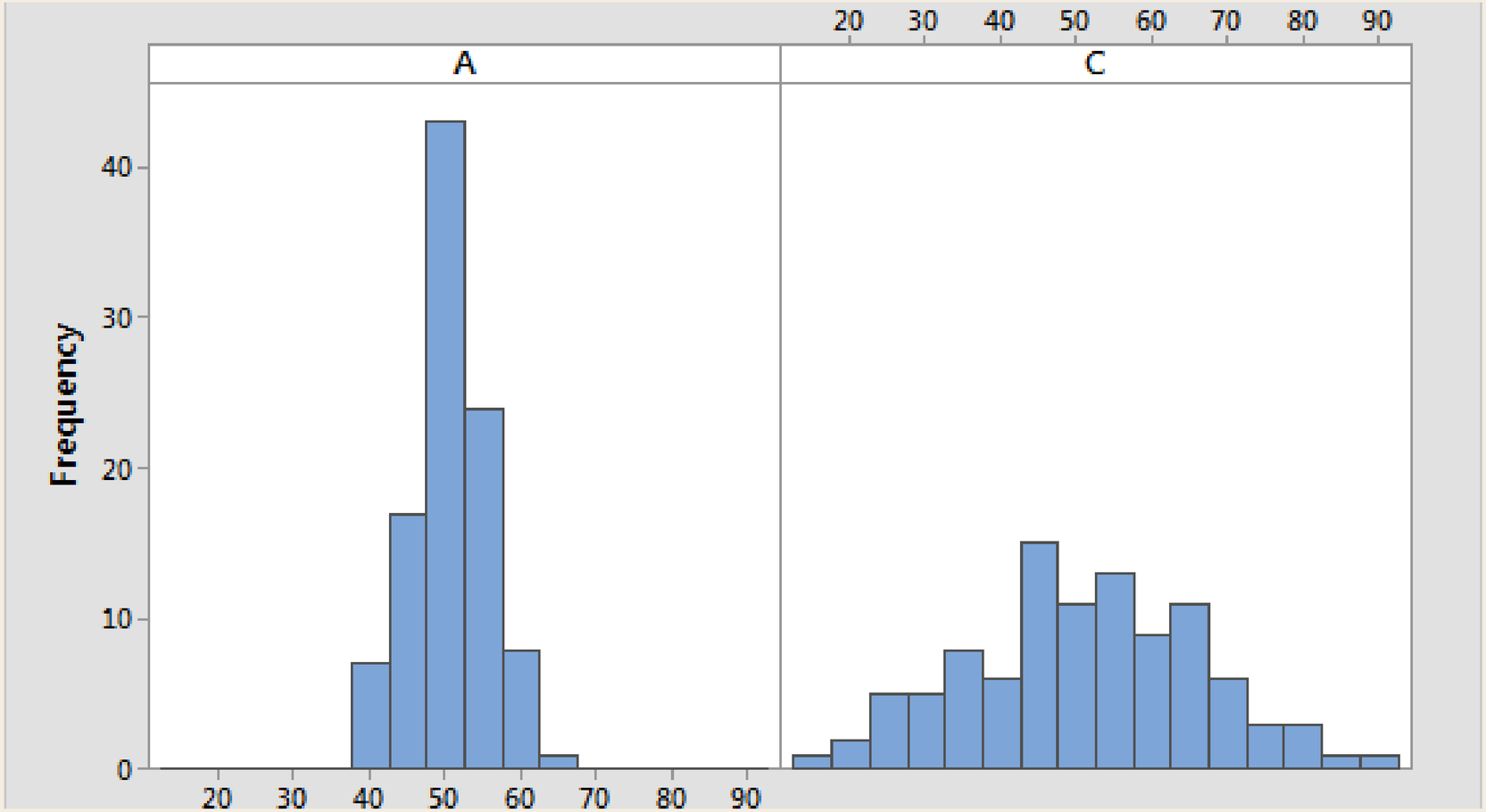


Interpreting Histogram

1. Range
2. Unimodal/ Bimodal/ Multimodal
3. Symmetric/ Skewed
4. Data has Gaps/ is Clustered
5. Suspected outliers

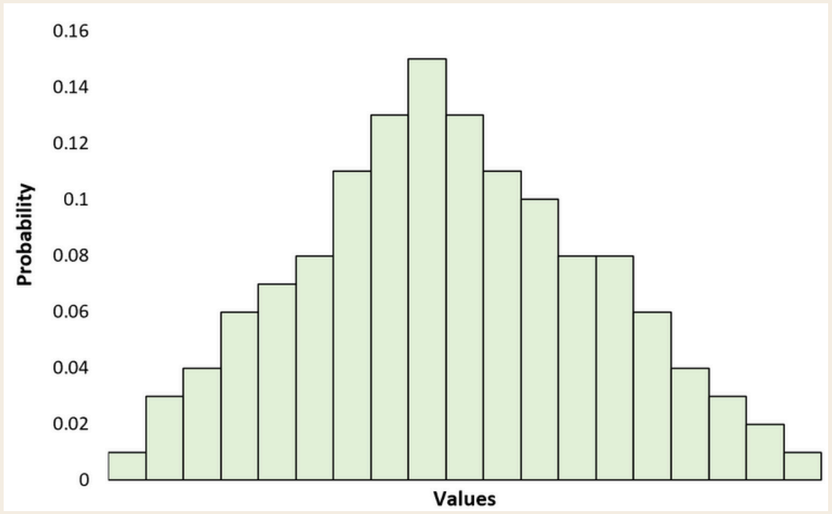
Interpreting Histogram

1. Range

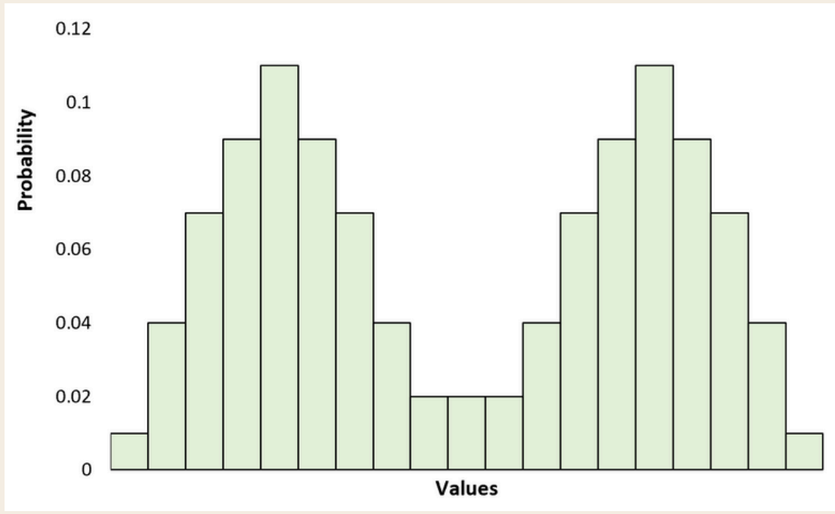


Interpreting Histogram

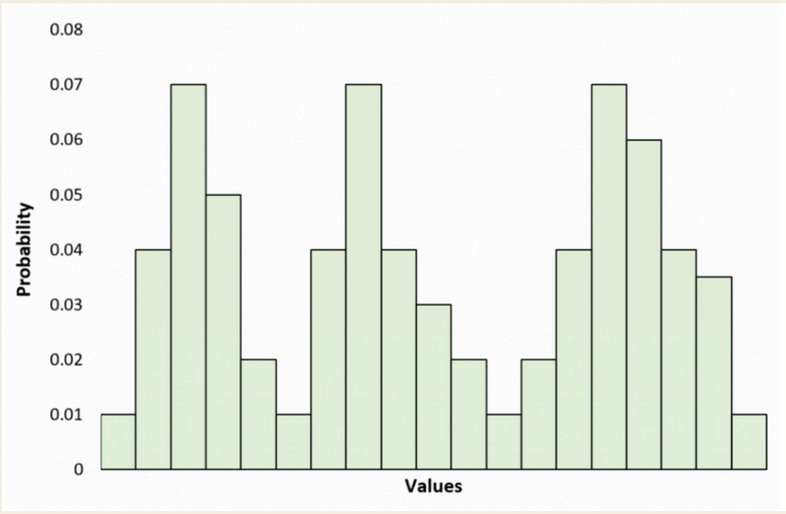
2. Unimodal/ Bimodal/ Multimodal



unimodal



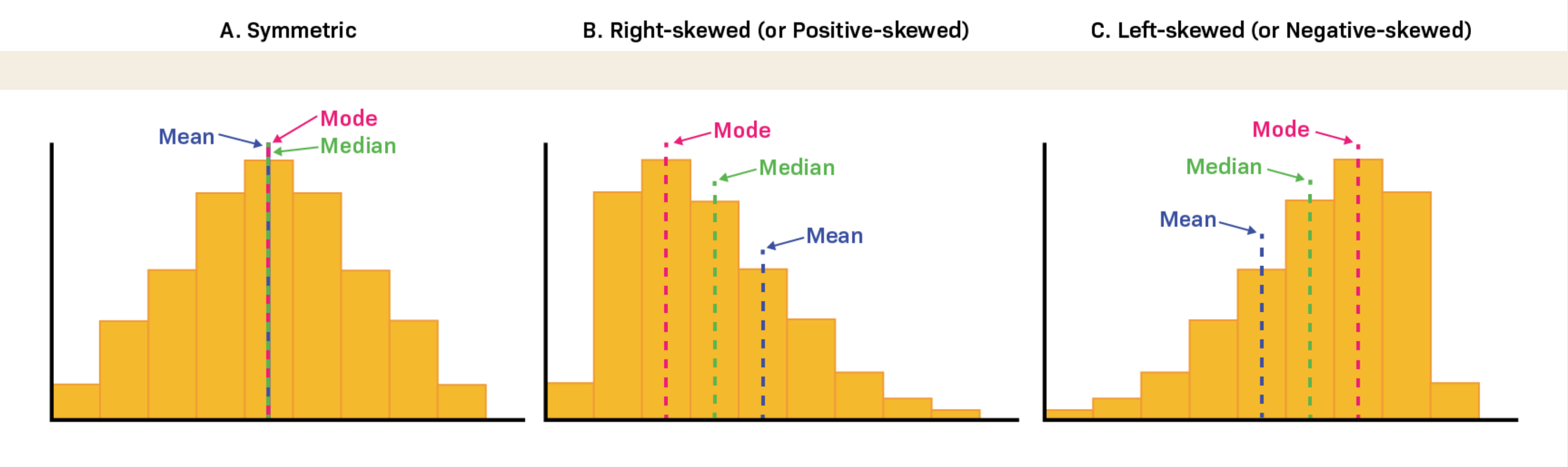
bimodal



multimodal

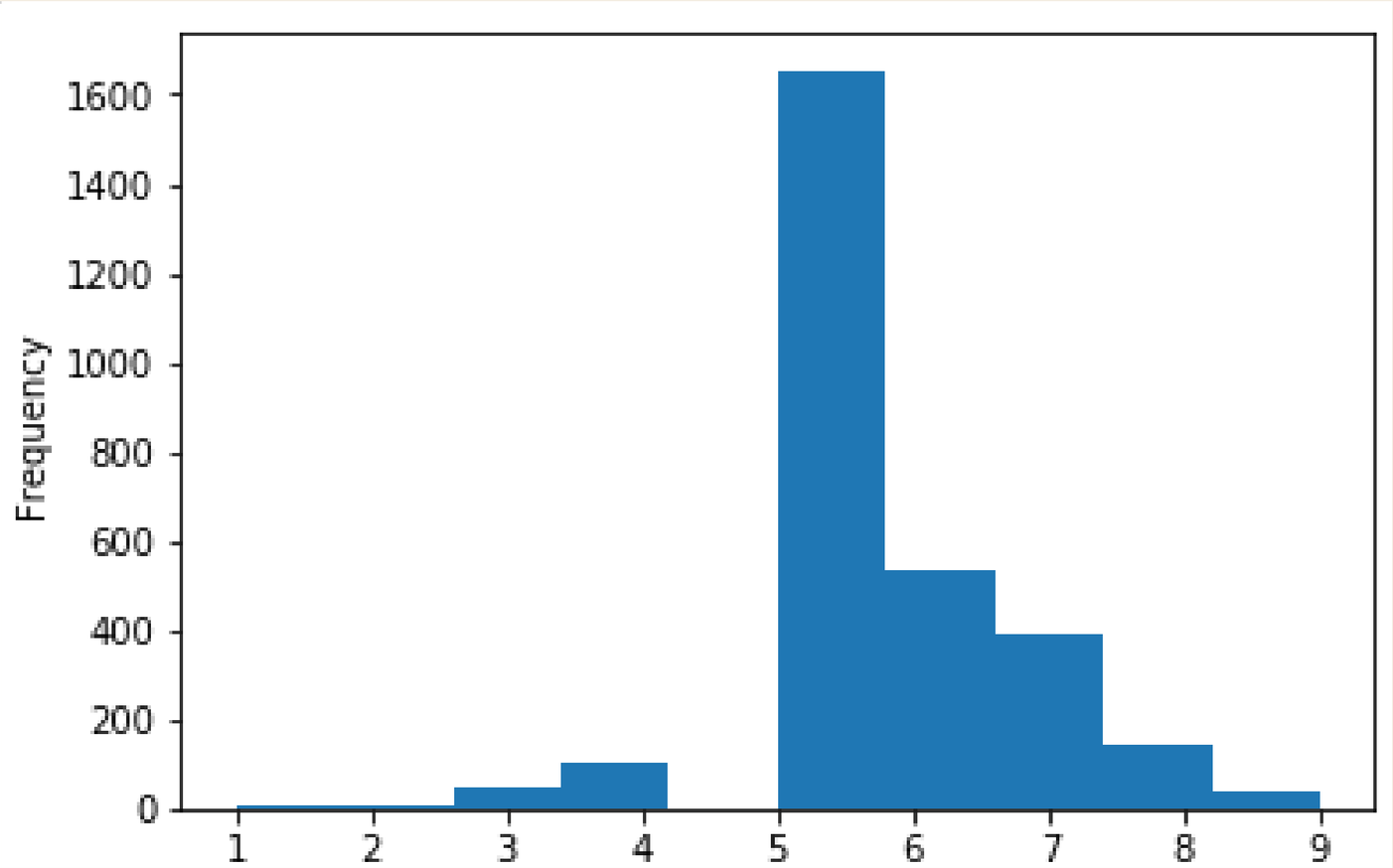
Interpreting Histogram

3. Symmetric/ Skewed



Interpreting Histogram

4. Data has Gaps/ is Clustered

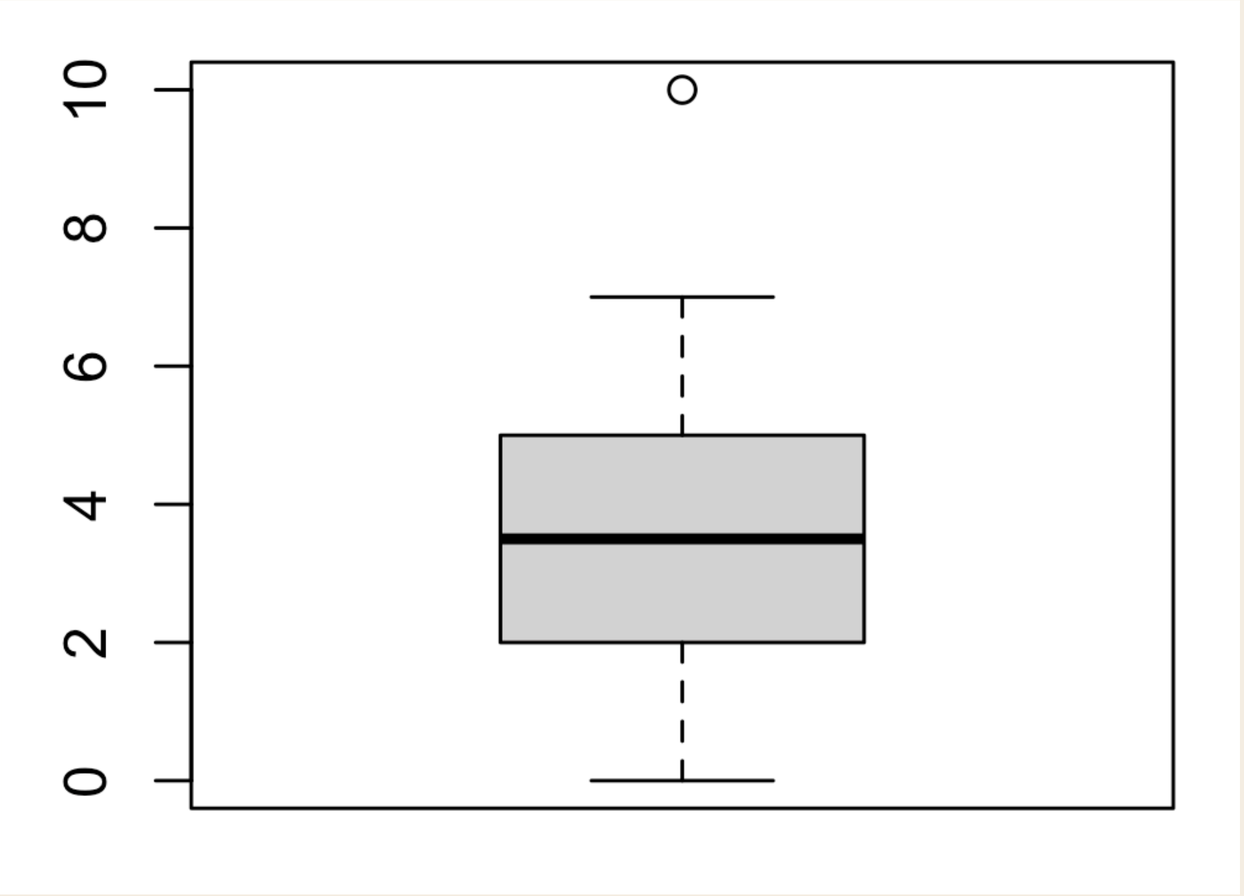


Interpreting Histogram

5. Suspected outliers

boxplot() function

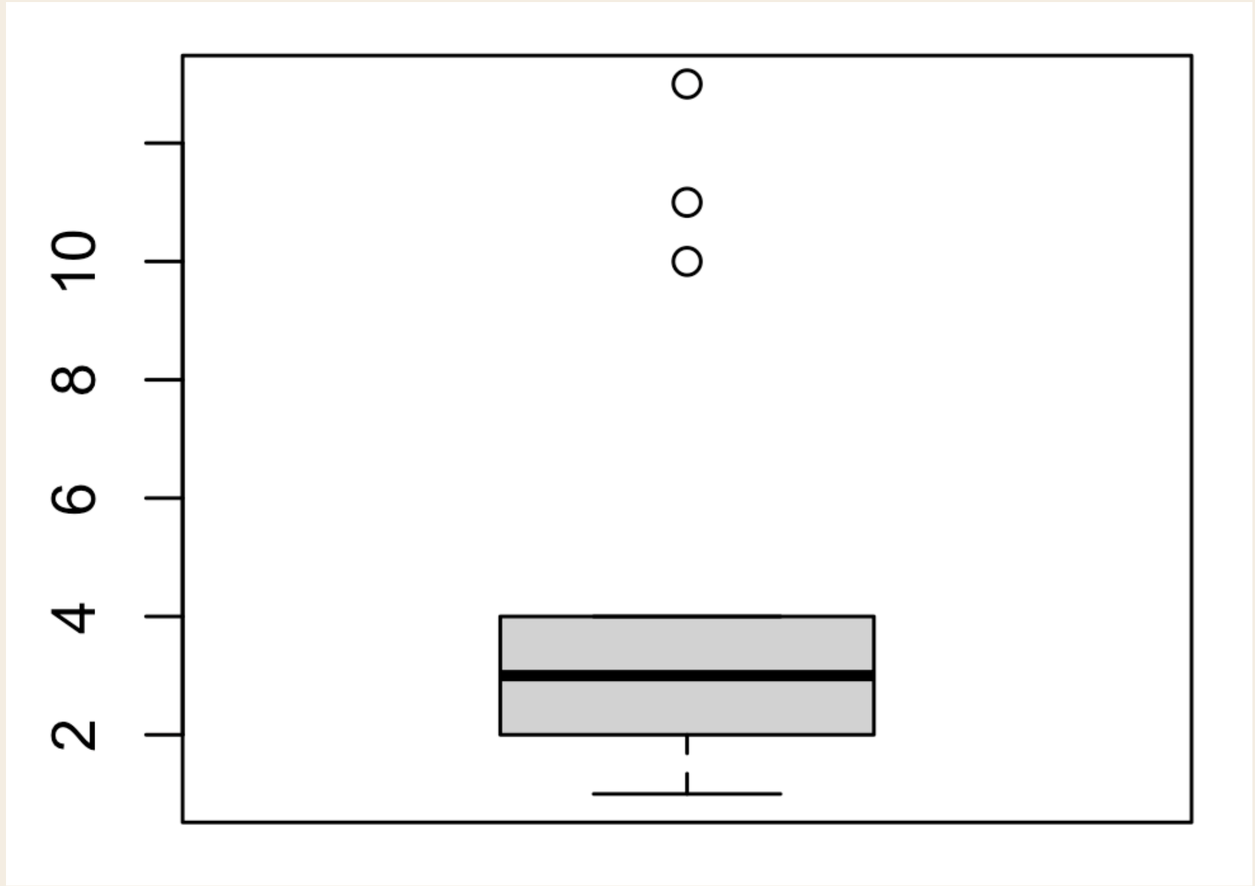
```
data = c(0,1,2,2,3,3,3,4,4,4,5,6,7,10)

boxplot(data)
```



boxplot() function

```
data = c(1,1,2,2,2,2,3,3,3,3,4,10,11,13)

bp = boxplot(data)

bp$out
```

```
> bp$out

[1] 10 11 13
```



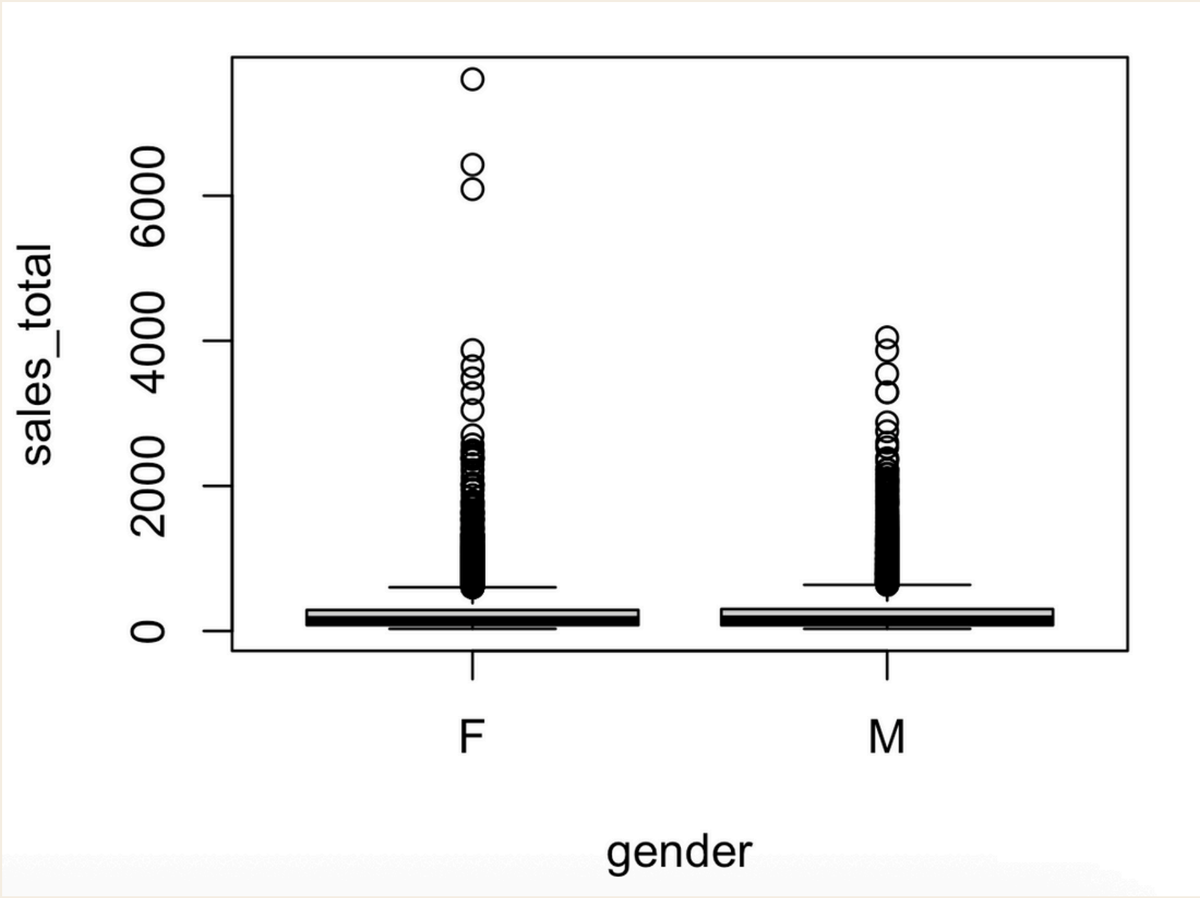
prints outlier’s value

boxplot() function

	cust_id	sales_total	num_of_orders	gender
1	100001	800.64	3	F
2	100002	217.53	3	F
3	100003	74.58	2	M
4	100004	498.60	3	M
5	100005	723.11	4	F
6	100006	69.43	2	F

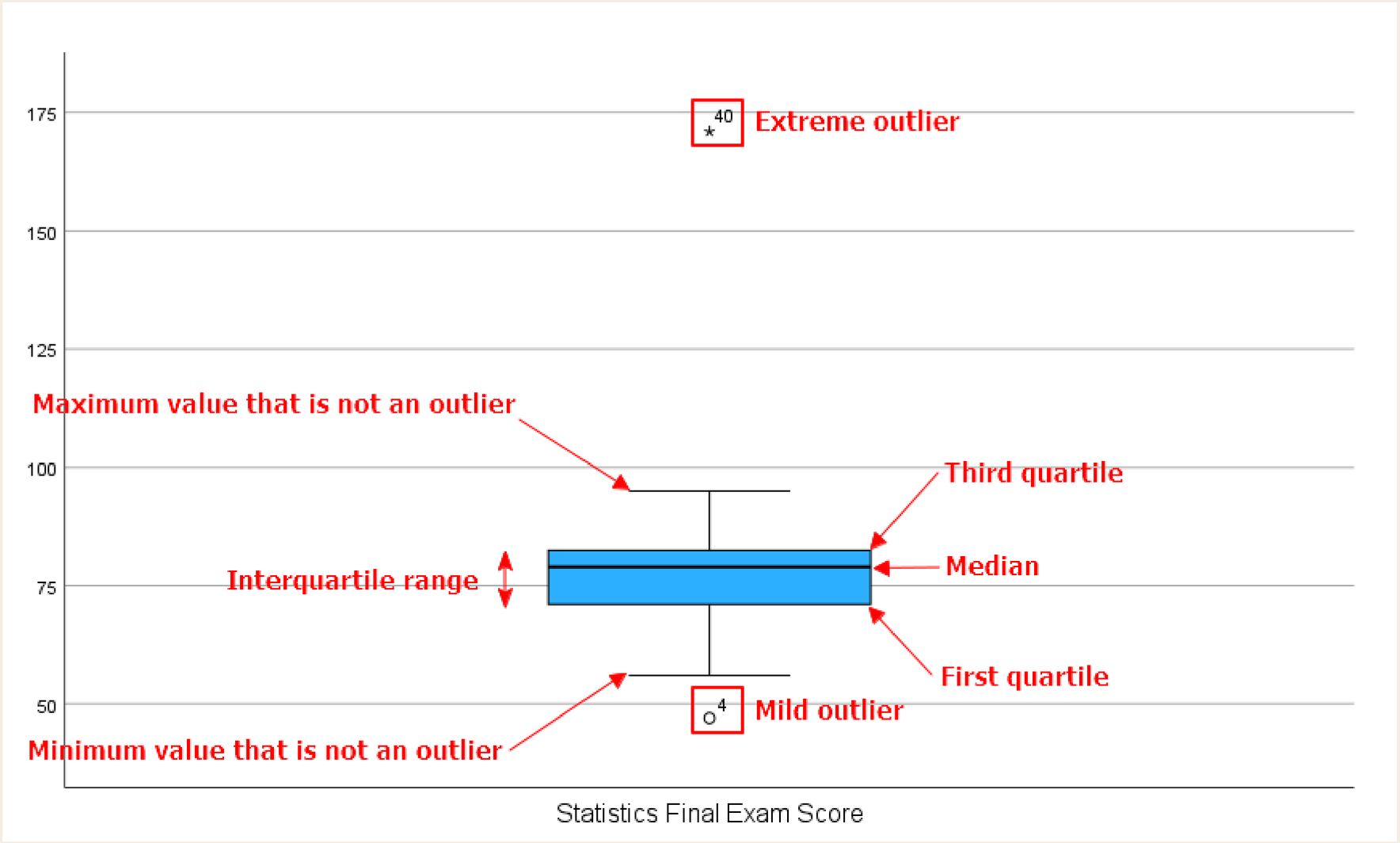
boxplot(sales_total ~ gender)

boxplot(response ~ independent)



side-by-side boxplot

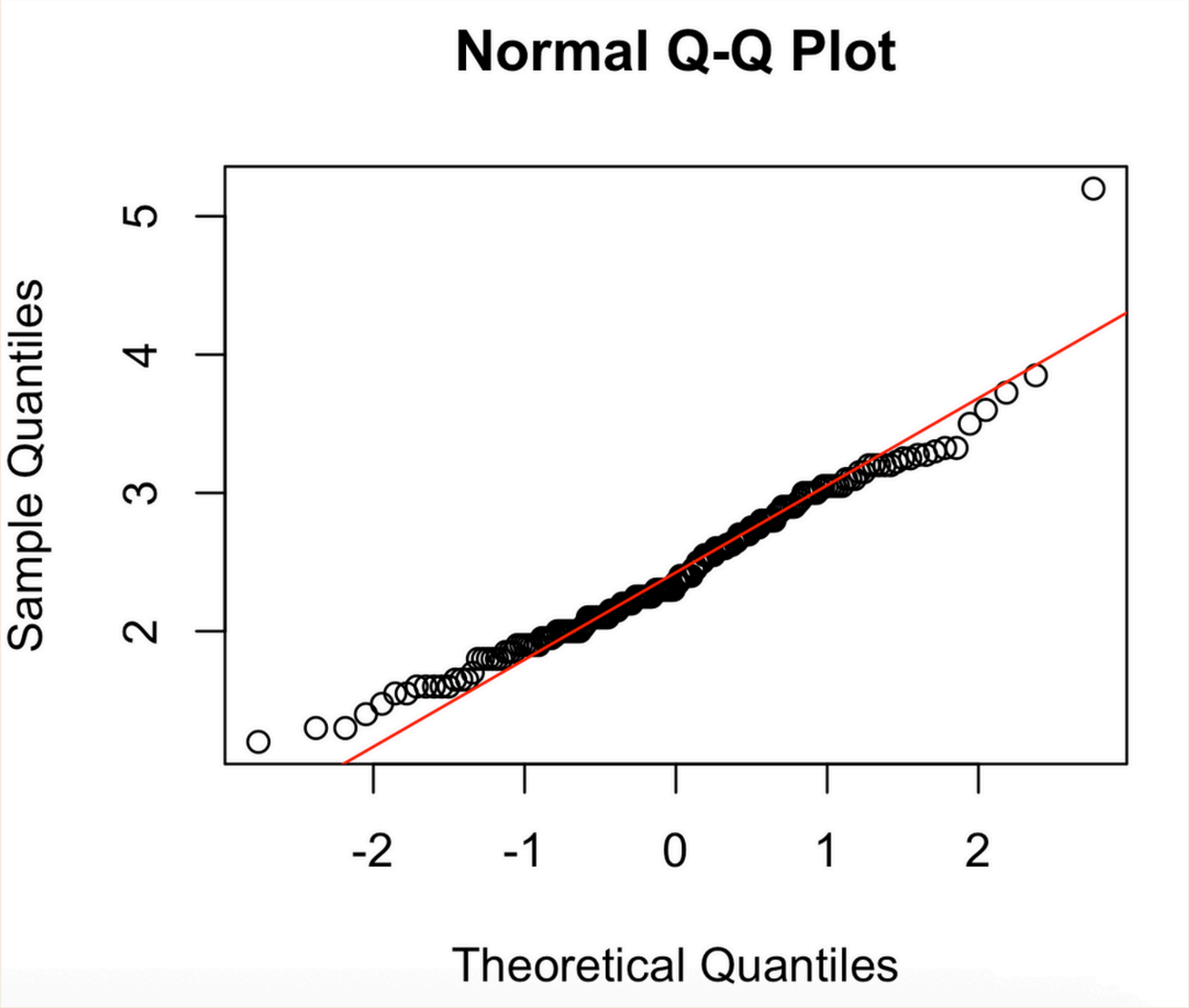
Interpreting Histogram



qqnorm() & qqline() function

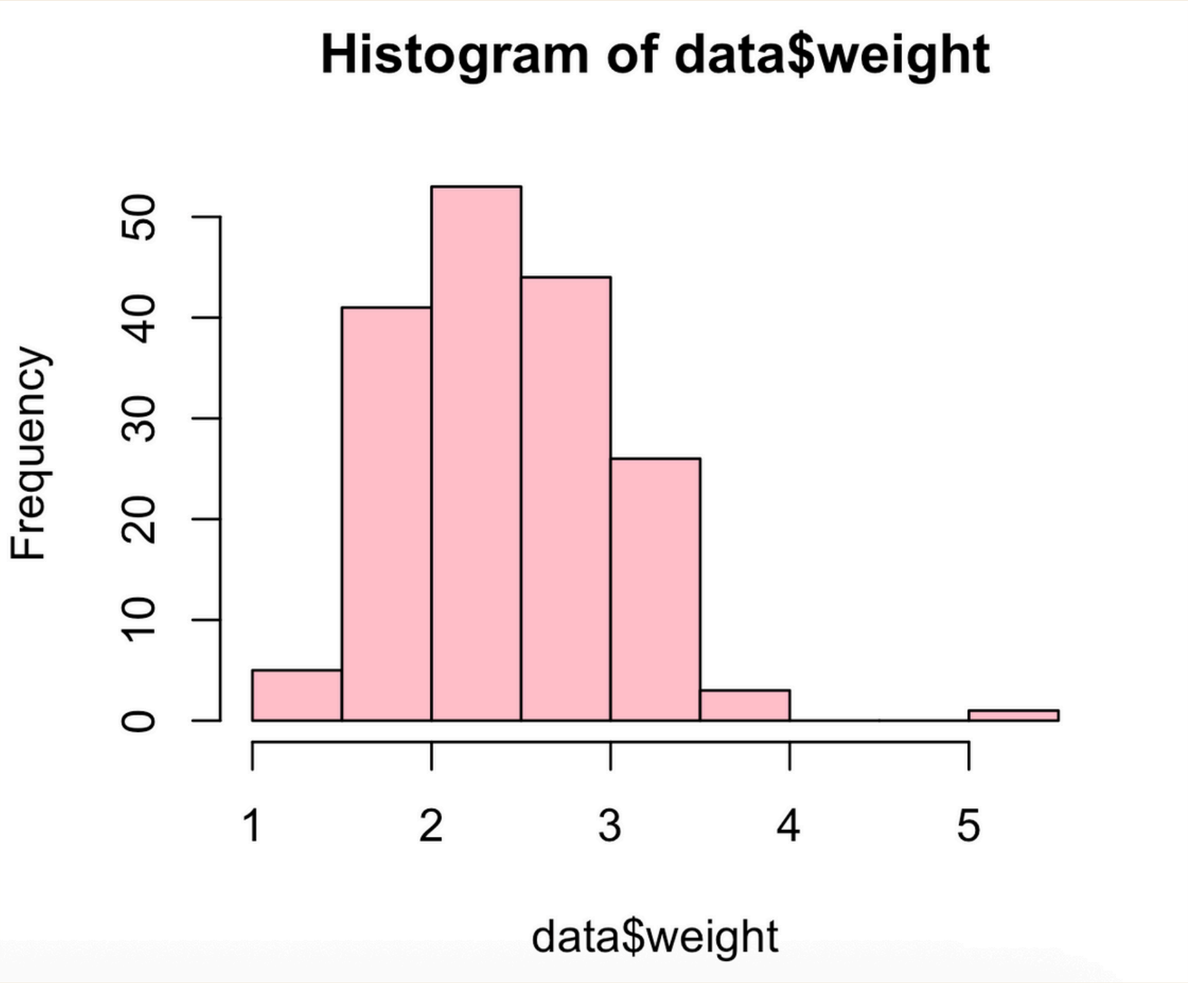
check if your data is normally distributed

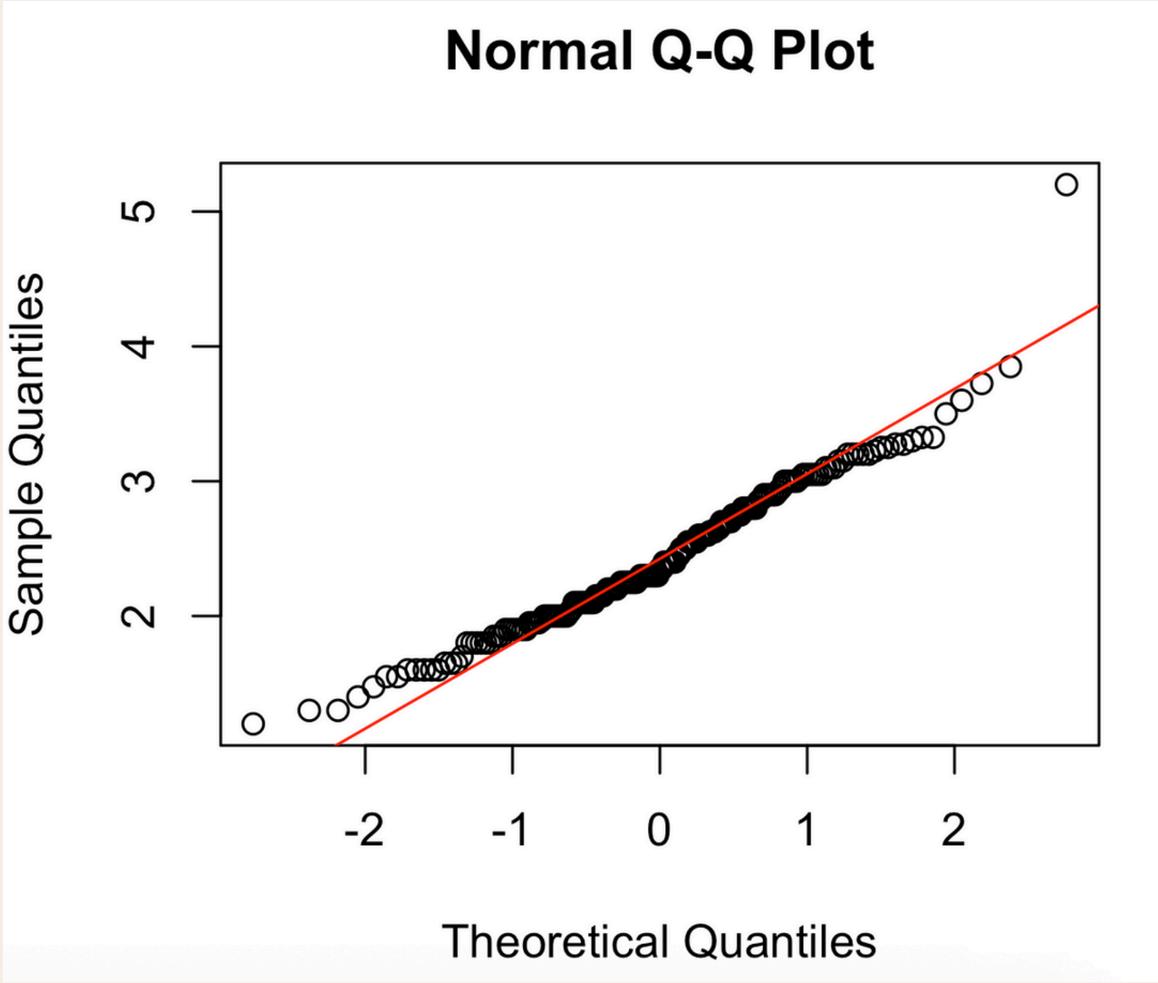
```
qqnorm(data$weight)

qqline(data$weight, col = "red")
```



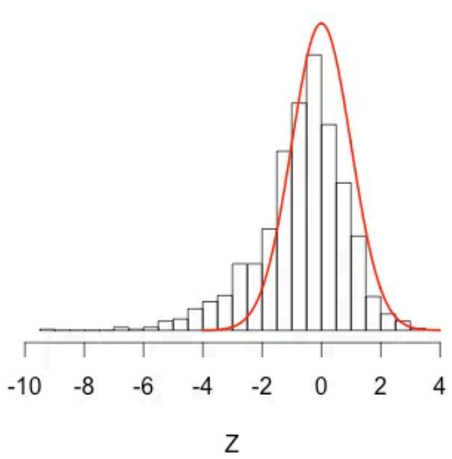
Interpreting Q-Q Plot

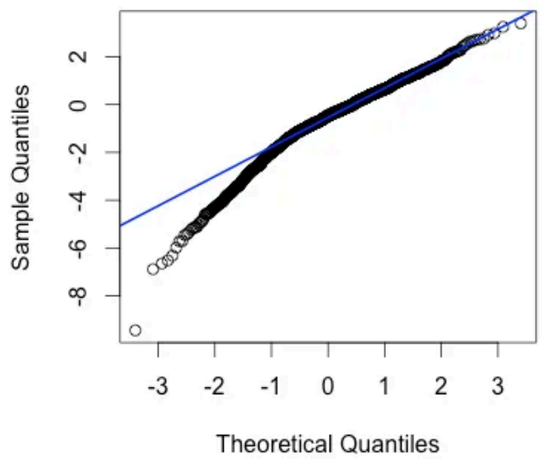
check if your data is normally distributed

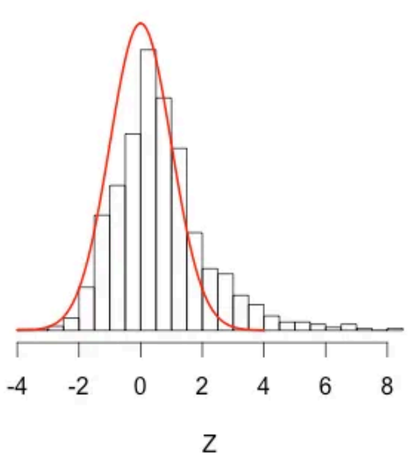


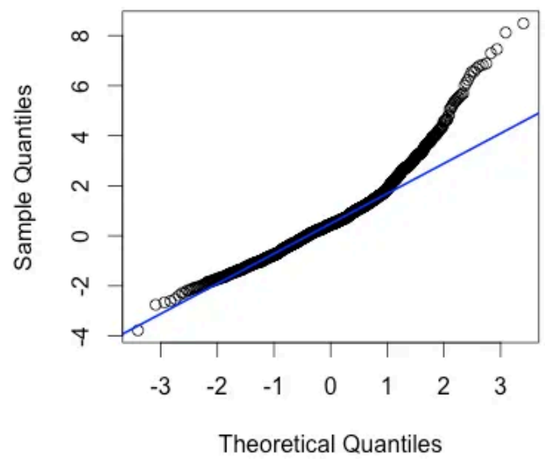


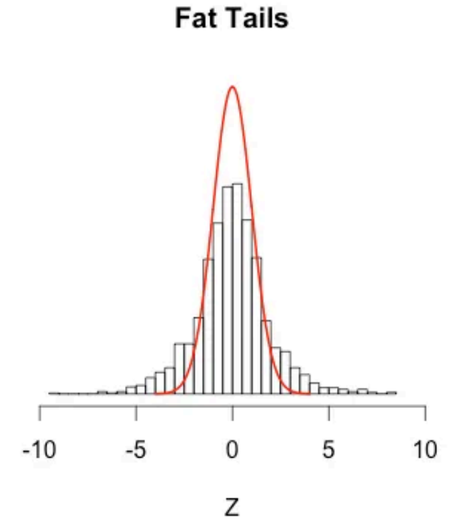
Interpreting Q-Q Plot

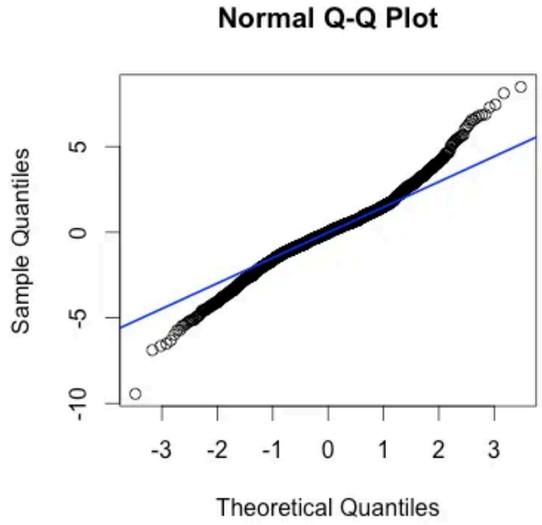


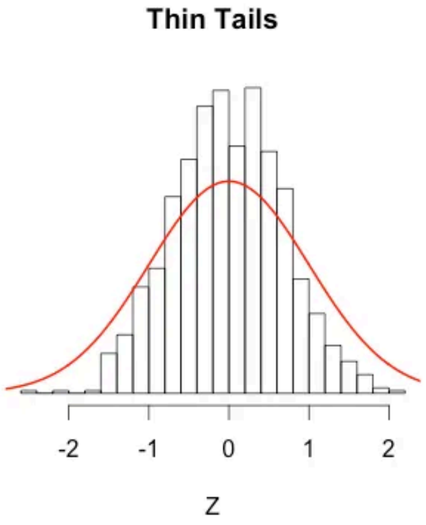


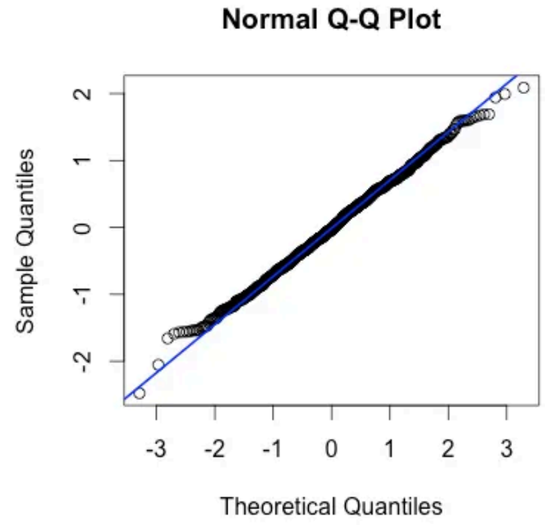












attach() function

```
data = data.frame(
   x = c(1,2,3,4,5),
   y = c(T,T,F,F,F)
)

hist(x)
```

❌ Error: object 'x' not found

```
data = data.frame(
   x = c(1,2,3,4,5),
   y = c(T,T,F,F,F)
)

hist(data$x)
```

attach() function

```
data = data.frame(
   x = c(1,2,3,4,5),
   y = c(T,T,F,F,F)
)

attach(data)

hist(x)
```

attach() makes the variables in a data frame directly accessible by their names **without needing to use the $ operator**

attach() function

```
data = data.frame(
x = c(1,2,3,4,5)
)

x = 'some random value'

attach(data)

print(x)
```

```
[1] "some random value"
```

You must be very careful when using attach()!

attach() function

```
data = data.frame(
x = c(1,2,3,4,5)
)

attach(data)

attach(data)
```

The following object is masked from data (pos = 3):

x

attach() function

```
data = data.frame(
  x = c(1,2,3,4,5)
)

attach(data)

# ...code that uses data....

detach(data)
```

Always remember to **detach() when you're done!

attach() function

```
search()
```

```
> search()
 [1] ".GlobalEnv"        "data"                "data"                "df"
 [5] "tools:rstudio"     "package:stats"       "package:graphics"    "package:grDevices"
 [9] "package:utils"     "package:datasets"    "package:methods"     "Autoloads"
[13] "package:base"
```

Get a list of all currently attached packages and R objects in the R search path

ON-SITE QUESTIONS

ON-SITE QUESTION 1,2

Consider a dataset about HDB resale flats in Singapore give in `hdbresale_reg.csv` which helps to investigate the factors that affect the resale price of the flats.

1. Import the dataset into R. *Do ask for help if you cannot do it on your own by the end of Week 4.*
2. How many flats in the sample given?

Discuss with your group! Time limit: 5 mins

ON-SITE QUESTION 1 ANSWER

Consider a dataset about HDB resale flats in Singapore give in `hdbresale_reg.csv` which helps to investigate the factors that affect the resale price of the flats.

1. Import the dataset into R. *Do ask for help if you cannot do it on your own by the end of Week 4.*

Straightforward answer:

```
setwd(...)

hdb <- read.csv("Data/hdbresale_reg.csv")
```

But how are we sure that our header and sep matches the default?

ON-SITE QUESTION 1 ANSWER

Here's what I like to do before calling read.csv():

Step 1: Look up the default values for read.csv()

```
?read.csv
```

```
read.csv(file, header = TRUE, sep = ",", quote = "\"",
         dec = ".", fill = TRUE, comment.char = "", ...)
```

ON-SITE QUESTION 1 ANSWER

Here's what I like to do before calling read.csv():

Step 2: Examine the data structure

(mainly to check separator)

```
readLines("Data/hdbresale_reg.csv", n=3)
```

```
> readLines("Data/hdbresale_reg.csv", n=3)
[1] "\"\",\"month\",\"town\",\"flat_type\",\"block\",\"street_name\",\"storey_rang
e\",\"floor_area_sqm\",\"flat_model\",\"lease_commence_date\",\"resale_price\""
[2] "\"580\",\"2012-03\",\"CENTRAL AREA\",\"3 ROOM\",\"640\",\"ROWELL RD\",\"01 TO
05\",74,\"Model A\",1984,380000"
[3] "\"581\",\"2012-03\",\"CENTRAL AREA\",\"3 ROOM\",\"640\",\"ROWELL RD\",\"06 TO
10\",74,\"Model A\",1984,388000"
```

sep = ","

quote = "\""

ON-SITE QUESTION 1 ANSWER

```r
# From here, we could examine the default values of read.csv()
# To know if we need to define input parameters such as header, sep etc
?read.csv

# To know what separator a dataset is, you can do:
readLines("Data/hdbresale_reg.csv", n=3)

# Import CSV file into data frame
hdb <- read.csv("Data/hdbresale_reg.csv")
```

MY FINAL ANSWER

ON-SITE QUESTION 1 ANSWER

```r
# Inspect the data (what I usually inspect first):
names(hdb) # Columns of the data frame

head(hdb, n=3) # First 3 rows

str(hdb) # Data types, first few values of each column

dim(hdb) # Number of rows and columns of data
```

After importing data, inspect it 🕵️

ON-SITE QUESTION 2 ANSWER

2. How many flats in the sample given?

Official answer:

```
dim(hdb) # 6055  11
```

Ans: 6055

This solution is correct!

But we are assuming that one row = one flat

ON-SITE QUESTION 2 ANSWER

2. How many flats in the sample given?

Verify assumption (Good practice):

```
length(unique(hdb$X)) # 6055

length(unique(hdb$X)) == dim(hdb)[1] # TRUE
```

After examination of data, X seems to be the unique identifier for data.

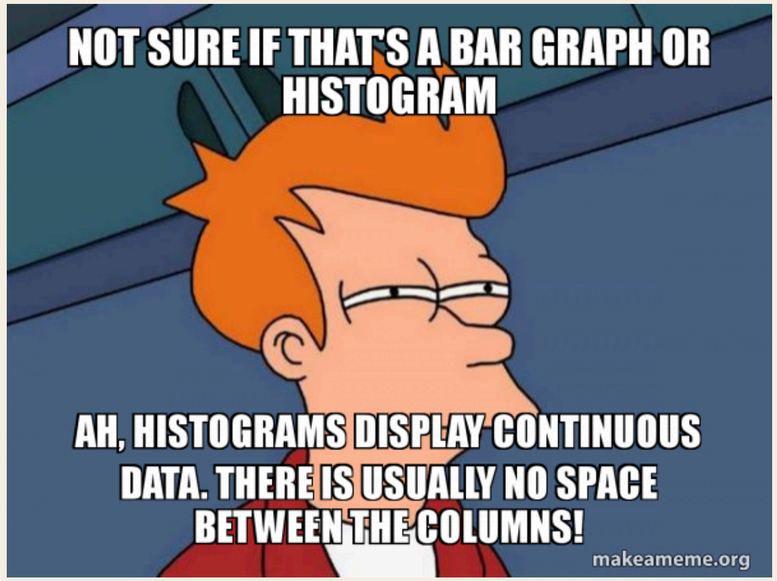
ON-SITE QUESTION 3

3. Exploring variable `resale_price`:

   (a) Create a histogram. Give your comments. Is the sample of resale price normally distributed?

   (b) Create a box plot.

If you have extra time, try to customize title, axis titles and colour of your plots!



Discuss with your group! Time limit: 8 mins

ON-SITE QUESTION 3A ANSWER

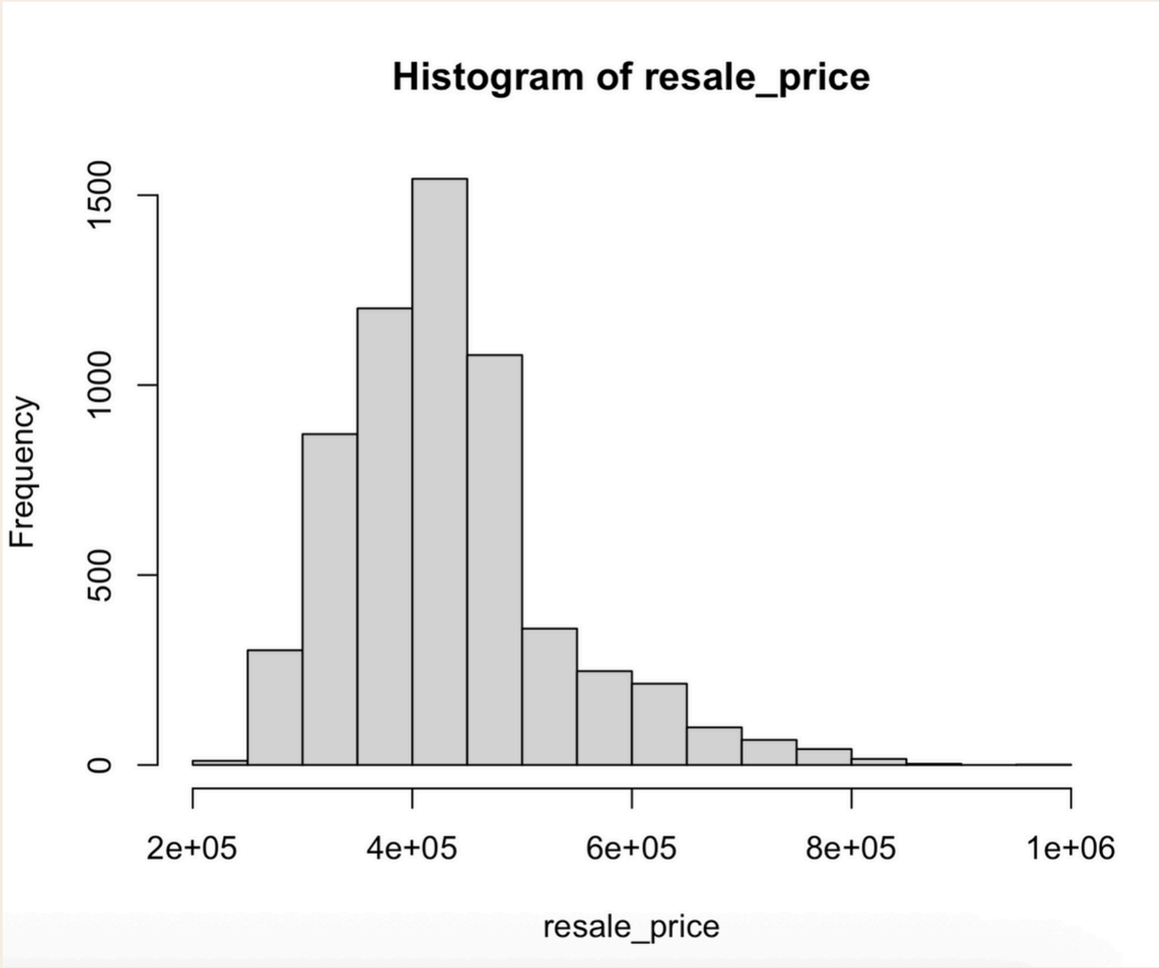
(a) Create a histogram. Give your comments. Is the sample of resale price normally distributed?

Creating a histogram:

```
hist(resale_price)
```

Comments:

- Range: 200k to 1M
- Not normally distributed → Right-skewed (Long right tail)
- Unimodal (One peak) at 400k-450k
- possible upper-tail outliers

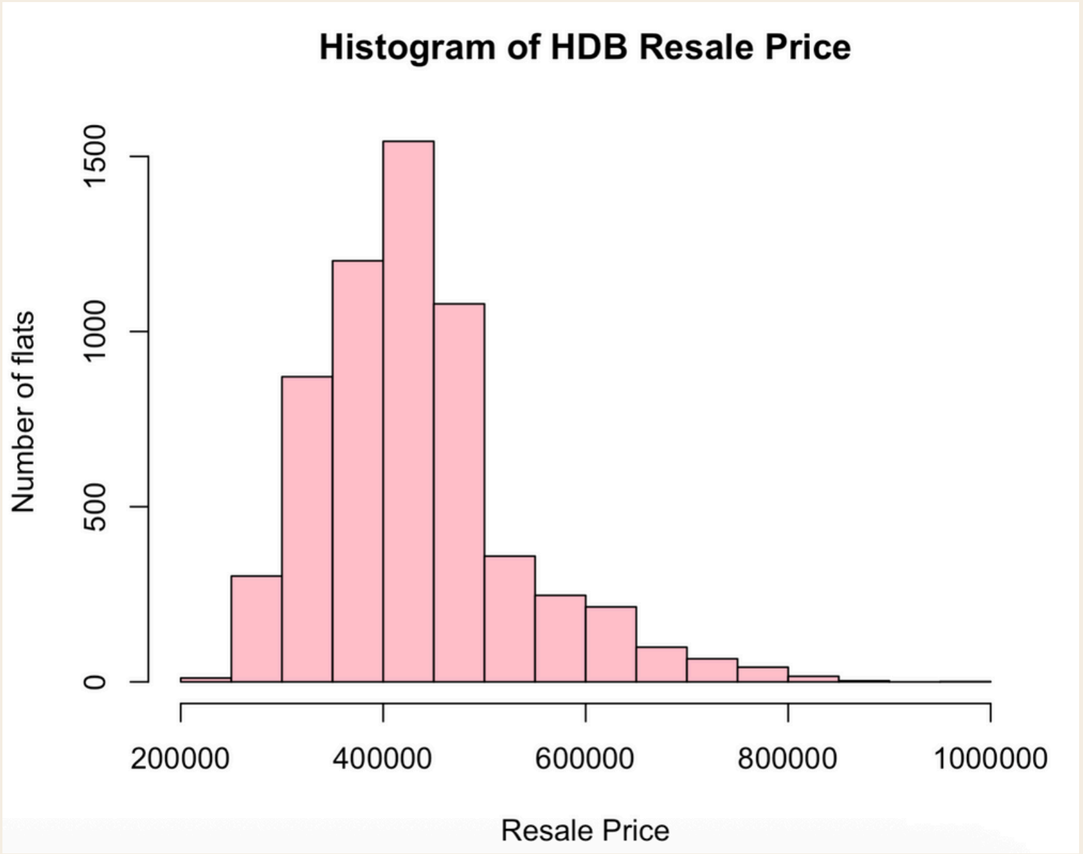


ON-SITE QUESTION 3A ANSWER

(a) Create a histogram. Give your comments. Is the sample of resale price normally distributed?

(EXTRA) Make plot prettier:

```
 # extra: turn off scientific notation
options(scipen = 999)

hist(resale_price,

     col = "pink", # colour

     main = "Histogram of HDB Resale Price", # main title

     xlab = "Resale Price", # x-axis title

     ylab = "Number of flats") # y-axis title

# to change back to default
options(scipen = 0)
```



ON-SITE QUESTION 3B ANSWER

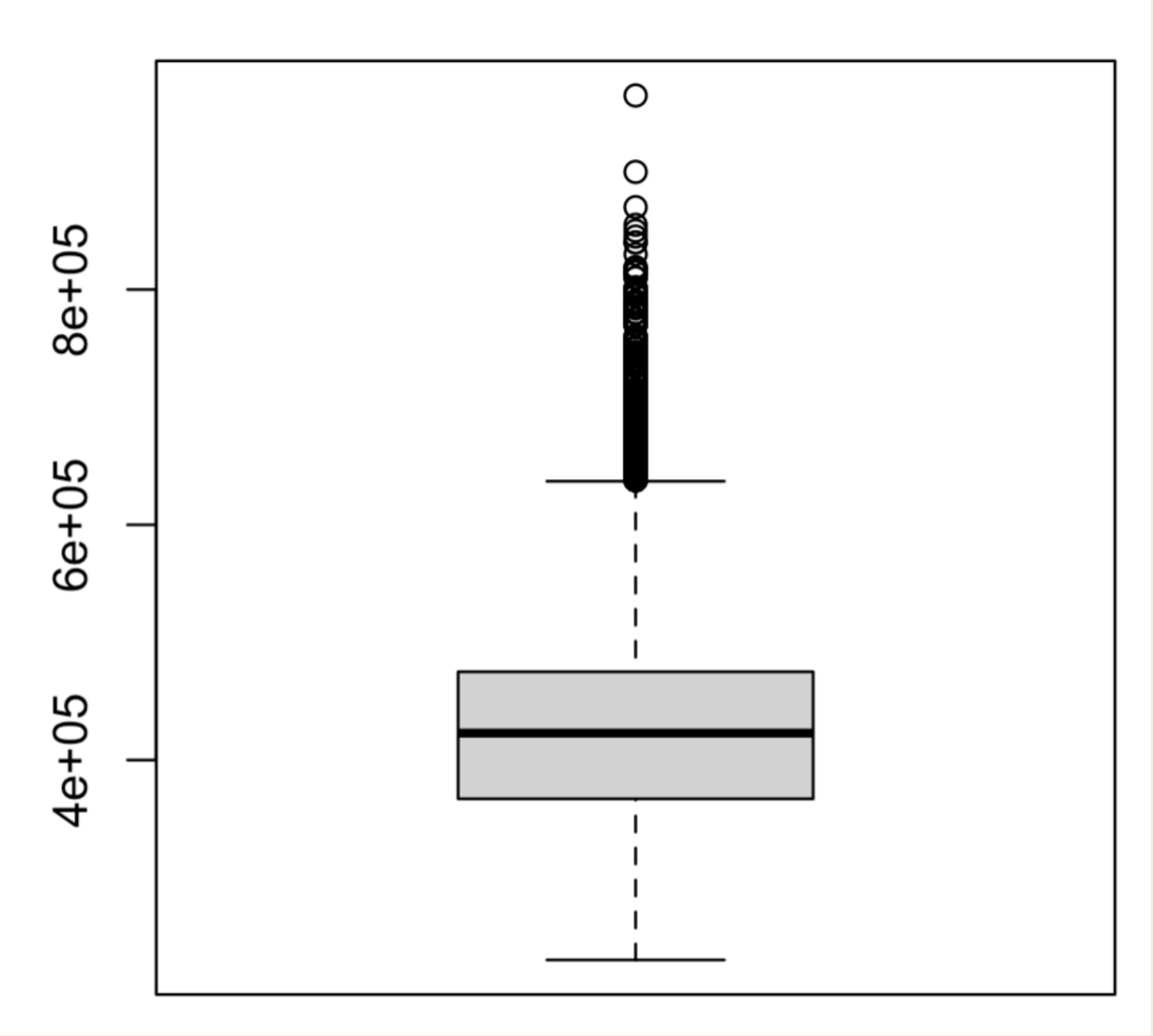
(b) Create a box plot.

Creating a boxplot:

```
boxplot(resale_price)
```

Comments:

- Many upper-tail outliers → 650k and above
- Lower Quartile slightly below 400k
- Median slightly above 400k
- Upper Quartile slightly below 500k



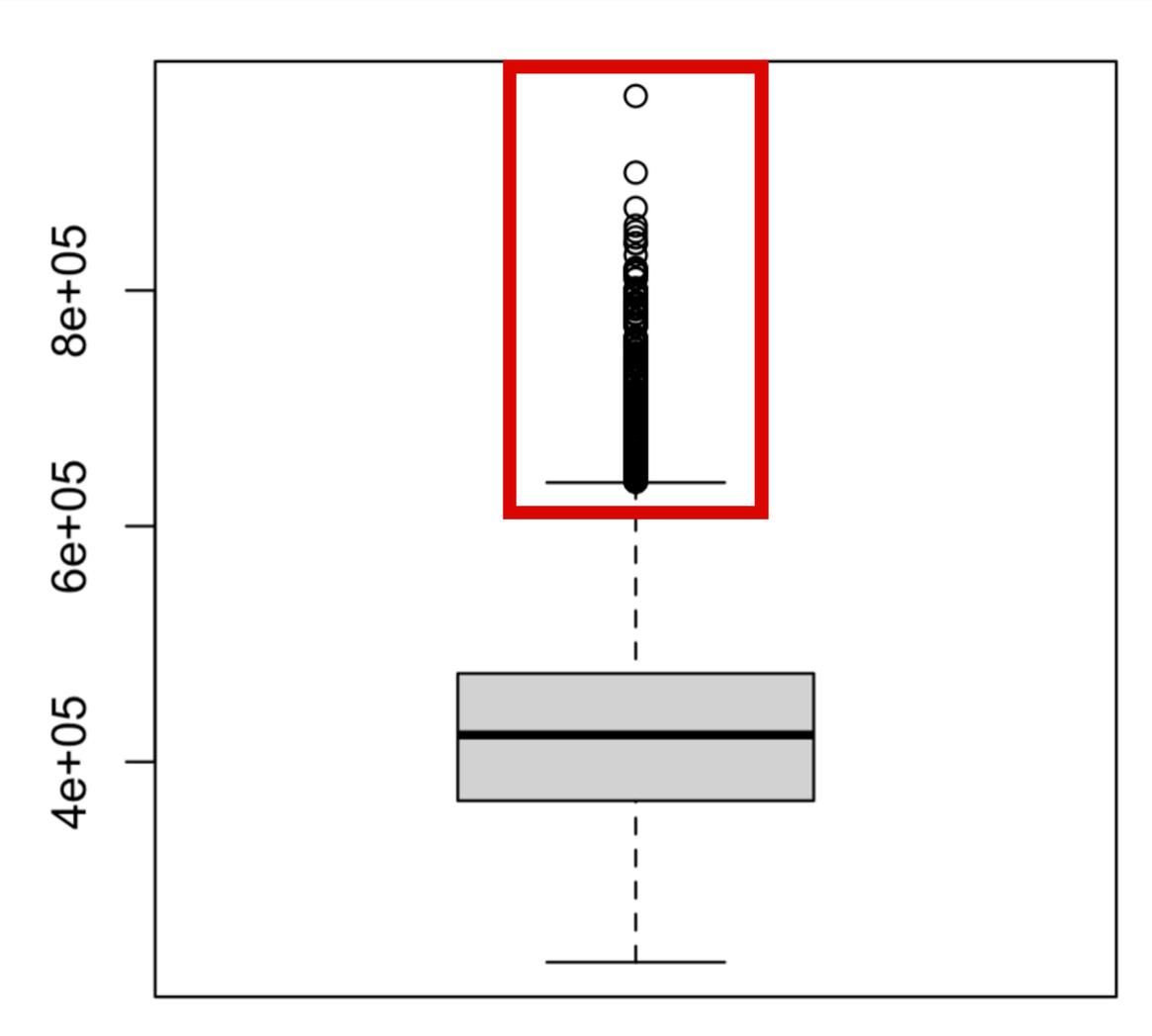
ON-SITE QUESTION 3B ANSWER

(b) Create a box plot.

Investigating Outliers

```
outliers = boxplot(resale_price)$out

length(outliers)
```

284 outliers (A lot!)



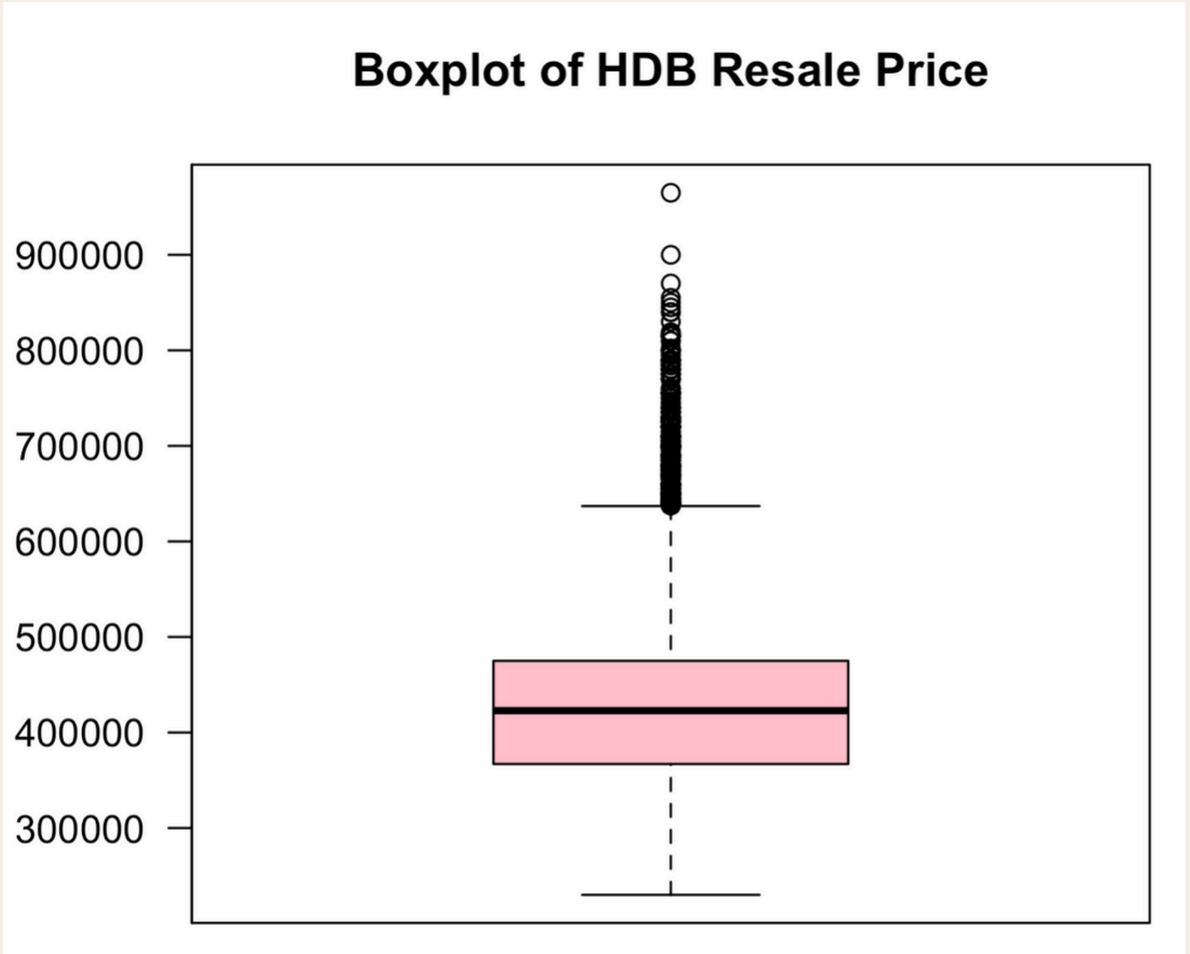
ON-SITE QUESTION 3B ANSWER

(b) Create a box plot.

(EXTRA) Make plot prettier:

```
boxplot(resale_price,
        col = "pink",
        main = "Boxplot of HDB Resale Price",
        yaxt = "n" # suppresses default y-axis
)

# Create customizable axis
axis(side = 2,
     at = seq(200000, 1000000,
     by = 100000), las = 1)
```

- **side = 2**: left axis
- **at**: points at which tick-marks are drawn
- **las = 1**: make labels horizontol



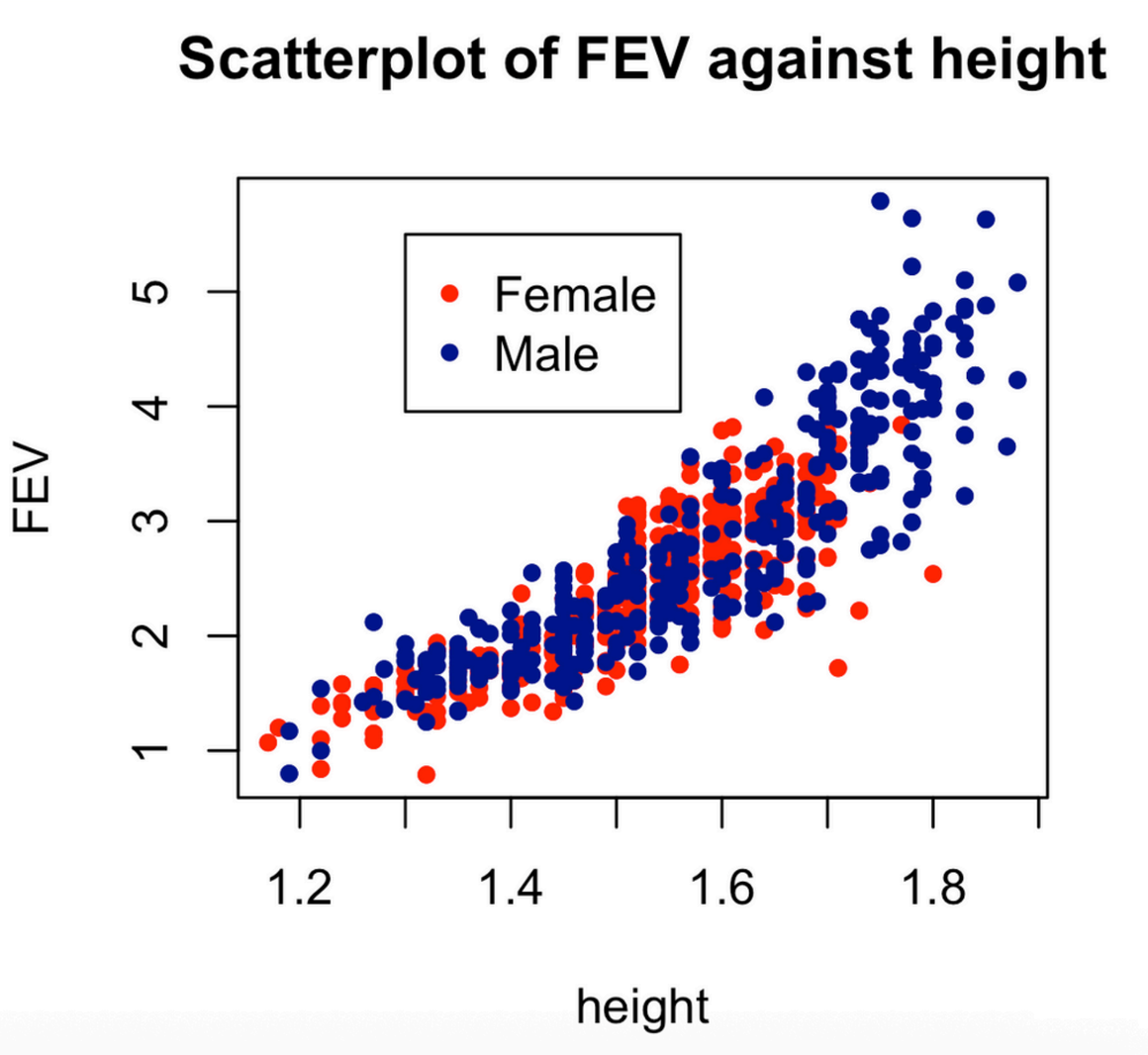
OFF-SITE QUESTIONS

This week, I'm trying something where I go through answers in R-studio rather than slides.

Would love for your feedback which you prefer after the class :D

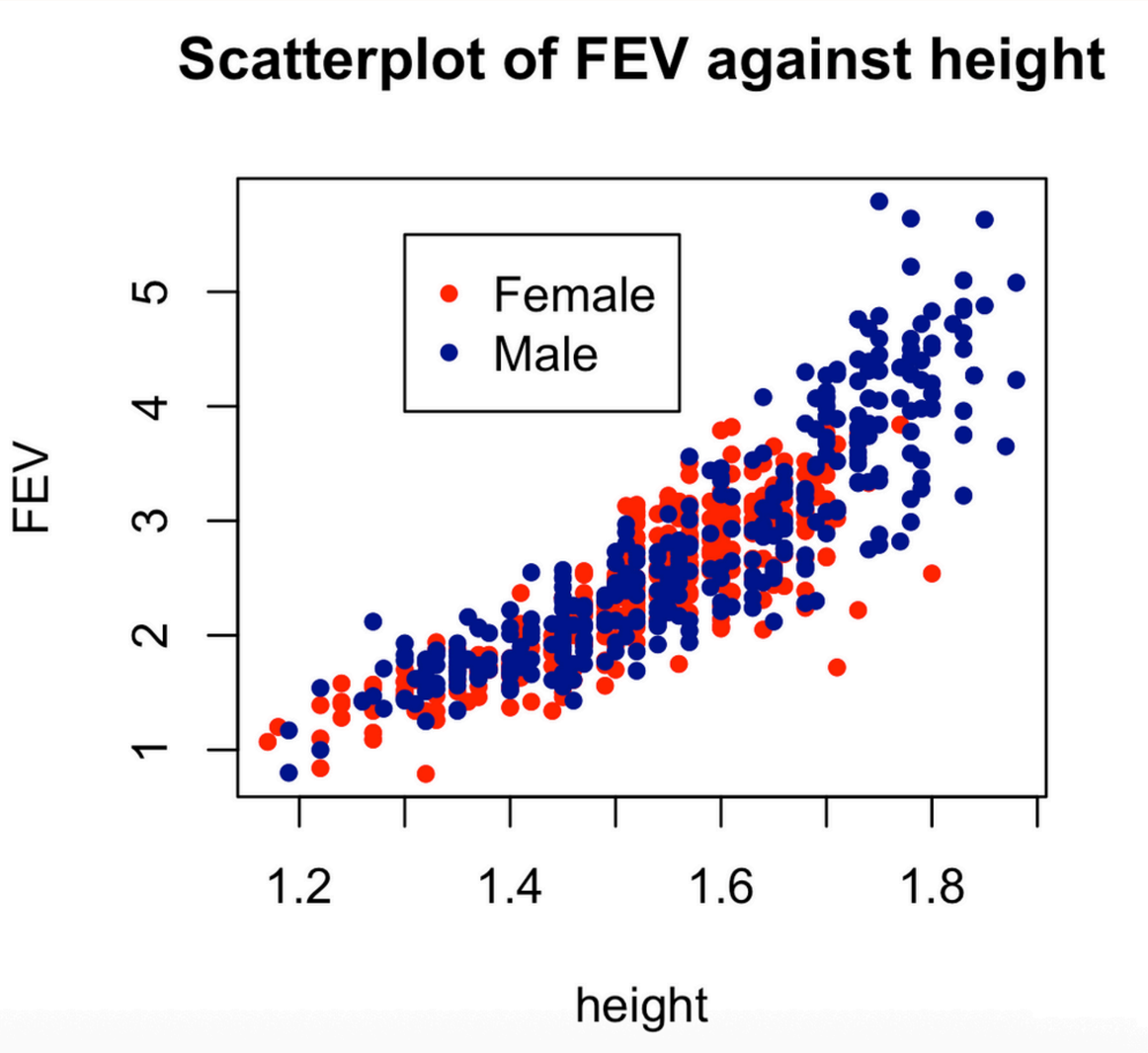
EXTRA QUESTIONS

EXTRA QUESTION



Based on this plot that we created. Can we conclude higher height causes higher FEV?

EXTRA QUESTION



No. Correlation != Causation

While there's a strong association, we cannot conclude that height directly causes higher FEV.

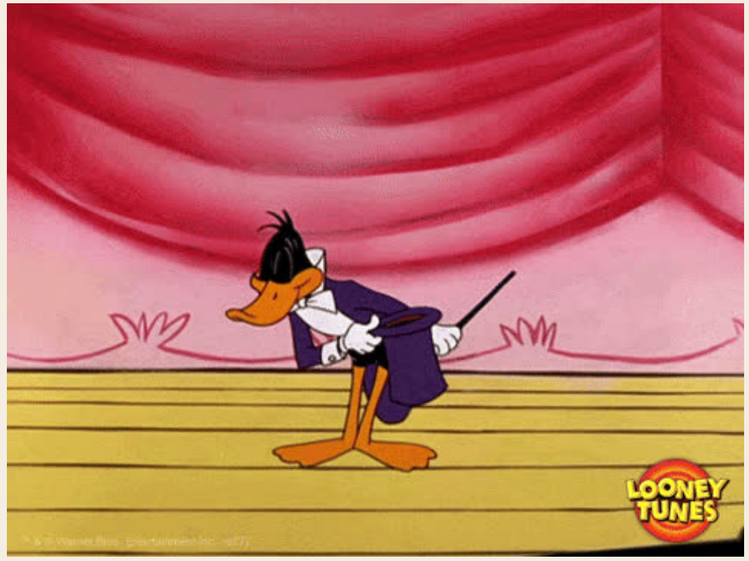
If you want to learn more, can look into **confounding variables**

FINAL QUESTION

WHICH DO YOU PREFER, TEACHING MAINLY USING CODE OR SLIDES?

PollEv.com
/kaironglee

THANK YOU!



Weekly Slides & Code will be shared on Github (⭐ star it for bookmark):

https://github.com/kr7001/DSA1101-teaching-materials