Statistical Computing with R
Masters in Data Science 503 (S2)
First Batch, SMS, TU, 2021

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Public Health Informatics

Patan Academy of Health Sciences, Lalitpur, Nepal

Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

Faculty, FAIMER Fellowship in Health Professions Education, India/USA.

Unit 1: R Software for Basic Programming [7 hrs.]

- R software, Statistics, Big Data and Data Science.
- Downloading and installing R software in Windows, Linux and Unix systems.
- Variables, Data types, Vectors, Lists and Matrix in R. Factors, Data Frames and Dealing with missing values in R.
- Logical statements, Loops, Functions and Pipes in R. Coding and naming conventions in R.
- Reproducible Analysis: Markdown Language, YAML Language; R Markdown/knitr document in R IDE (RStudio).
- Profiling and optimizing codes/scripts in R.

Review Preview

- R installation
- R Studio installation
- R console
- R objects
- R functions
- R plots
- Summary statistics
- Frequencies
- Multiple Response Frequencies

R Consoles:

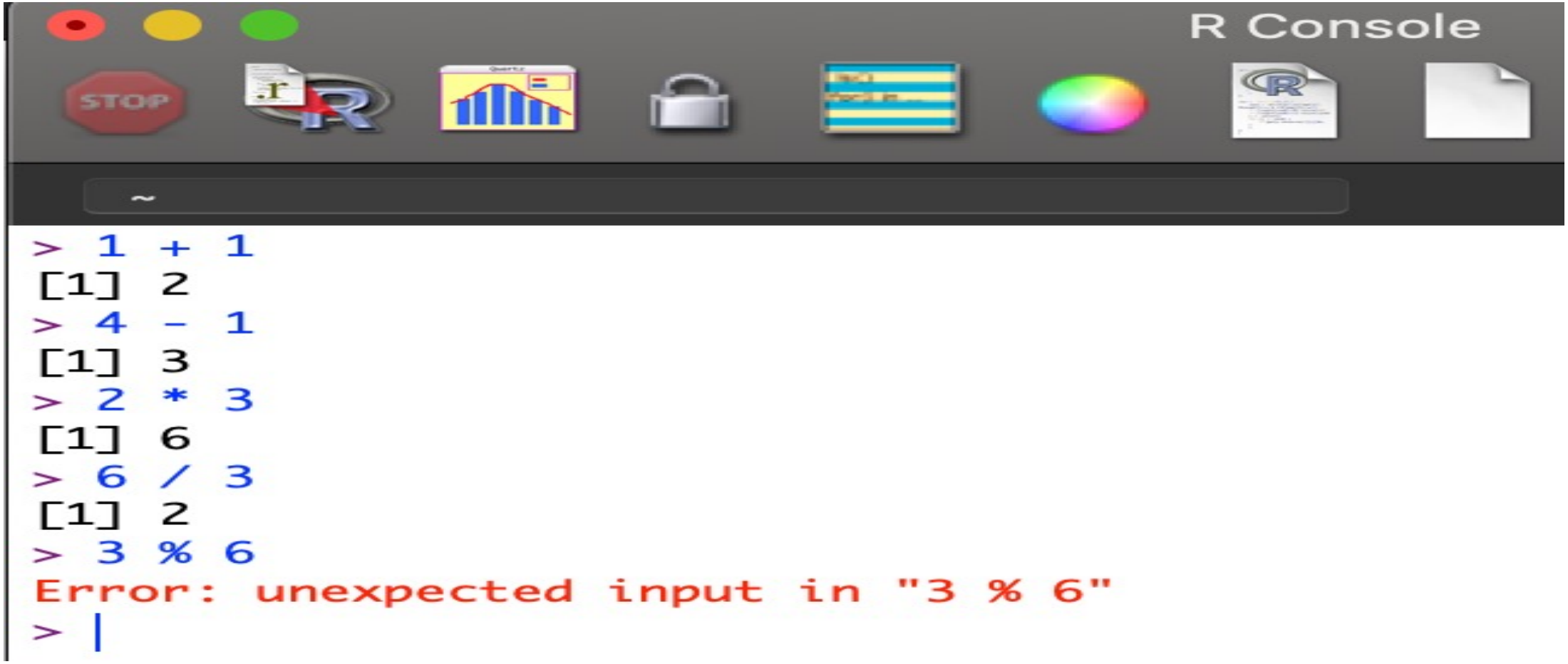
R Console

```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[R.app GUI 1.72 (7847) x86_64-apple-darwin17.0]

[Workspace restored from /Users/mac/.RData]
[History restored from /Users/mac/.Rapp.history]

>
```

mac — R — 80×24

```
Last login: Sat Aug 14 19:58:02 on console
macs-MBP:~ mac$ R

R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Previously saved workspace restored]

>
```

Basic Mathematical Operations in R console: https://rstudio-education.github.io/hopr/basics.html



RStudio File Edit Code View Plots Session Build Debug Profile Tools Window Help 50% Mon 8:54 AM Shital Bhandary

RStudio

Go to file/function Addins Project: (None)

Console Terminal Jobs

~/

```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from ~/.RData]

>
```

R Console

Environment History Connections Tutorial

Import Dataset List

Global Environment

Environment is empty

Objects, Data etc.

Files Plots Packages Help Viewer

Zoom Export

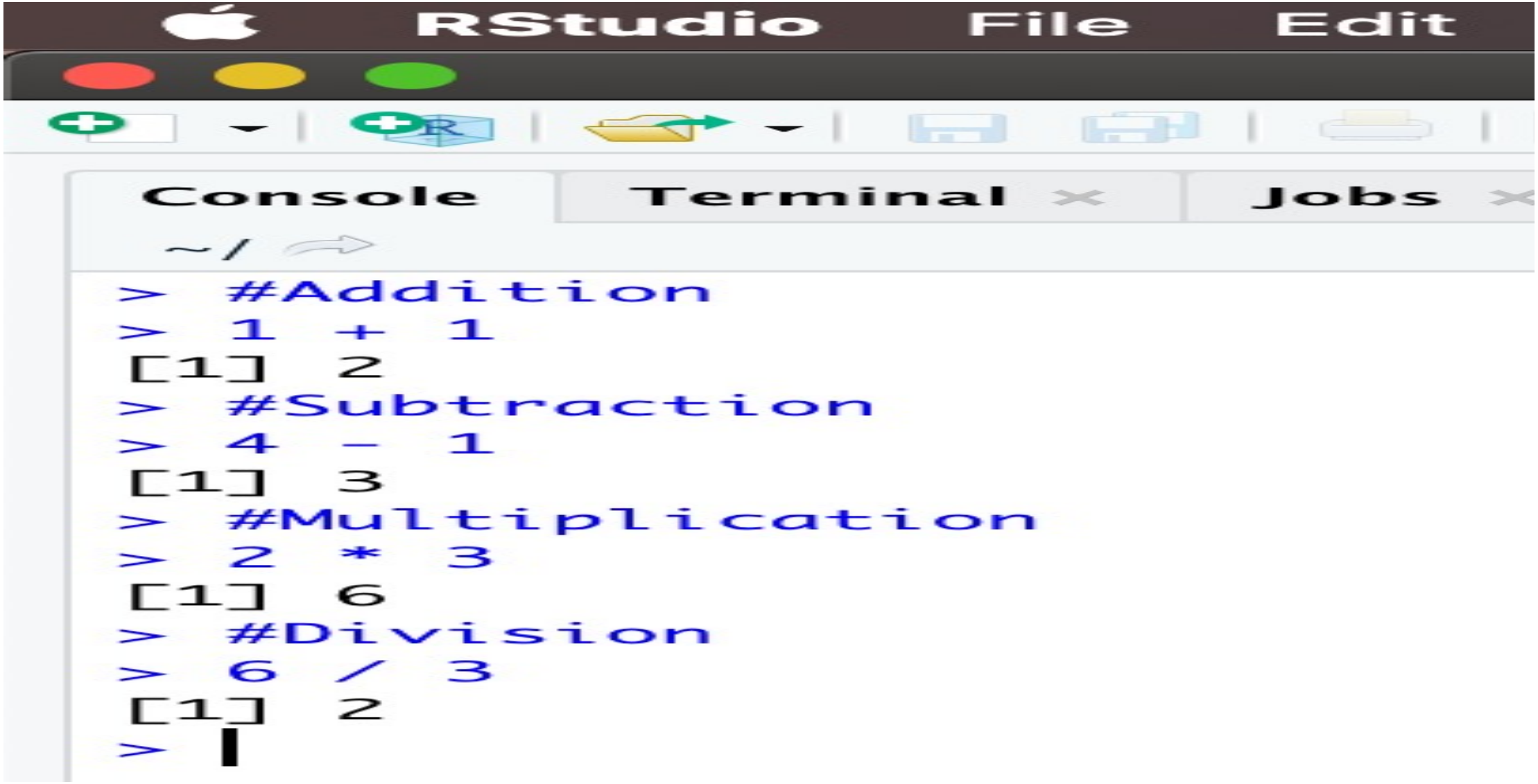
Outputs

Files | Plots | Packages | Help | Viewer

Install | Update

	Name	Description	Version
System Library			
☐	abind	Combine Multidimensional Arrays	1.4-5
☐	ade4	Analysis of Ecological Data: Exploratory and Euclidean Methods in Environmental Sciences	1.7-16
☐	aplpack	Another Plot Package: 'Bagplots', 'Iconplots', 'Summaryplots', Slider Functions and Others	1.3.3
☐	arm	Data Analysis Using Regression and Multilevel/Hierarchical Models	1.11-2
☐	askpass	Safe Password Entry for R, Git, and SSH	1.1
☐	assertthat	Easy Pre and Post Assertions	0.2.1
☐	aweek	Convert Dates to Arbitrary Week Definitions	1.0.2
☐	backports	Reimplementations of Functions Introduced Since R-3.0.0	1.2.1
☑	base	The R Base Package	4.0.2
☐	base64enc	Tools for base64 encoding	0.1-3
☐	BH	Boost C++ Header Files	1.72.0-3
☐	bit	Classes and Methods for Fast Memory-Efficient Boolean Selections	4.0.4
☐	bit64	A S3 Class for Vectors of 64bit Integers	4.0.5

Mathematical operator in R Studio:



R objects:

- Arrays: x and y defined in session 1, can be of any dimension
- Matrices: cbind of x and y (try it on your own and get class)
- Lists: Array with Strings, Integers, Numbers, Matrices, Boolean etc.)
- Data frame (data.frame to work with up to 1-2 gb data)
- Data table (data.table to work with more than 2 gb data)

RStudio File

Console Terminal

~/

```
> #Column vector
> x <- c(1:30)
> y <- x^3
> plot(x,y)
>
```

Environment History Connections Tutorial

Import Dataset List

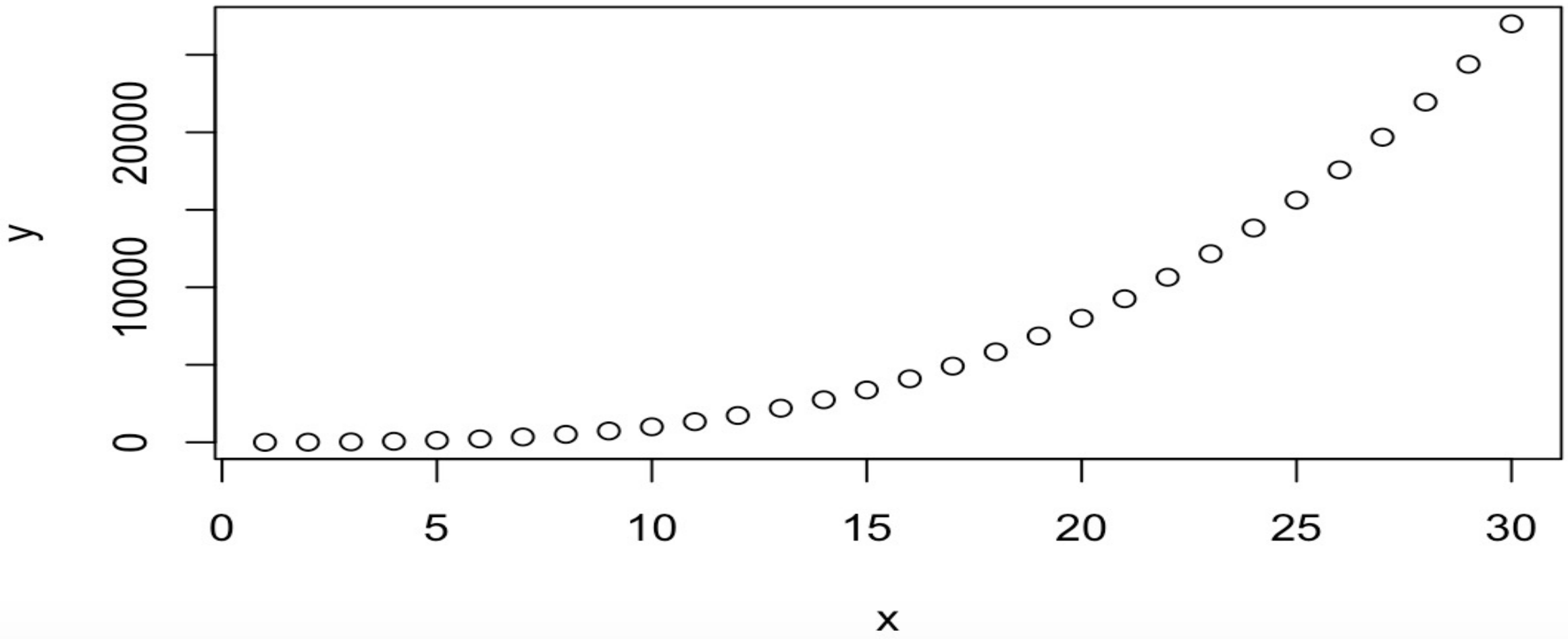
Global Environment

Values

x	int [1:30] 1 2 3 4 5 6 7 8 9 10 ...
y	num [1:30] 1 8 27 64 125 216 343 512 729 1000 ...

Files Plots Packages Help Viewer

Zoom Export Publish



RStudio File Edit Code View

Console Terminal Jobs

~/

```
> #Data Frame
> df <- data.frame(x<-c(1:30), y<-x^3)
> plot(df$x, df$y)
```

Environment History Connections Tutorial

Import Dataset List

Global Environment

Data

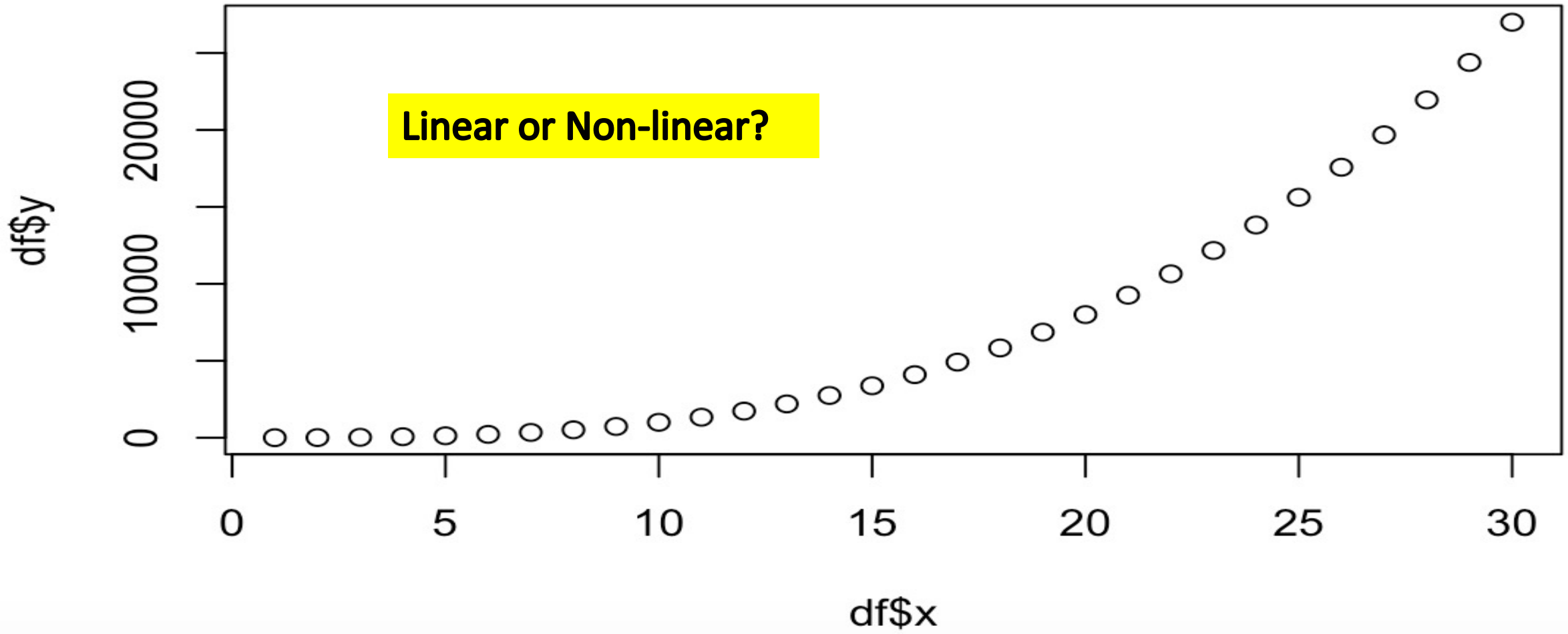
df	30 obs. of 2 variables

Values

x	int [1:30] 1 2 3 4 5 6 7 8 9 10 ...
y	num [1:30] 1 8 27 64 125 216 343 512 729 1000 ...

Files Plots Packages Help Viewer

Zoom Export Publish



Can we “transform” to make it “linear”?

- Yes, we can!
- **We can log transform the y and x variables and check it again**
- Let us define z as log(y) in r as follows:

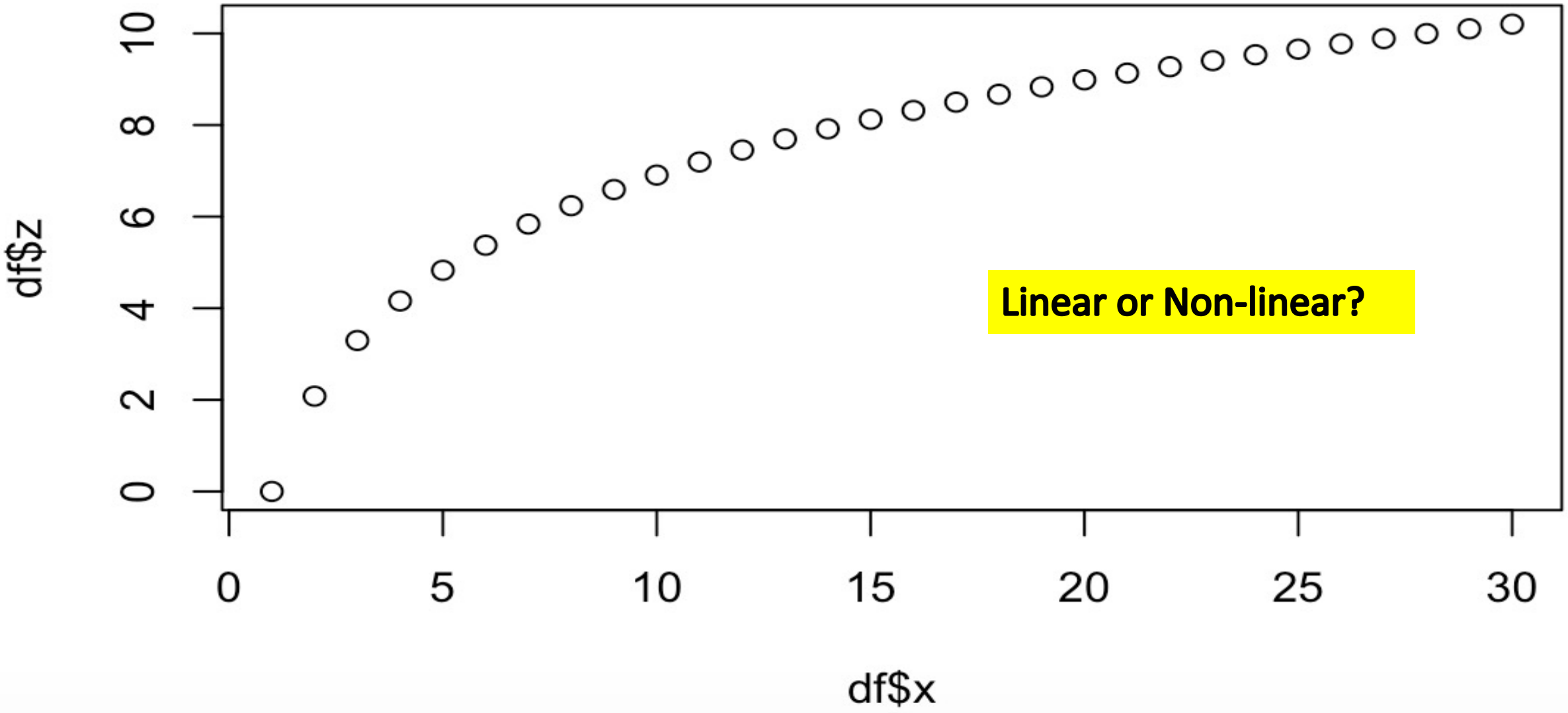
- df$z <- log(df$y)

- Let us plot the scatterplot again as:

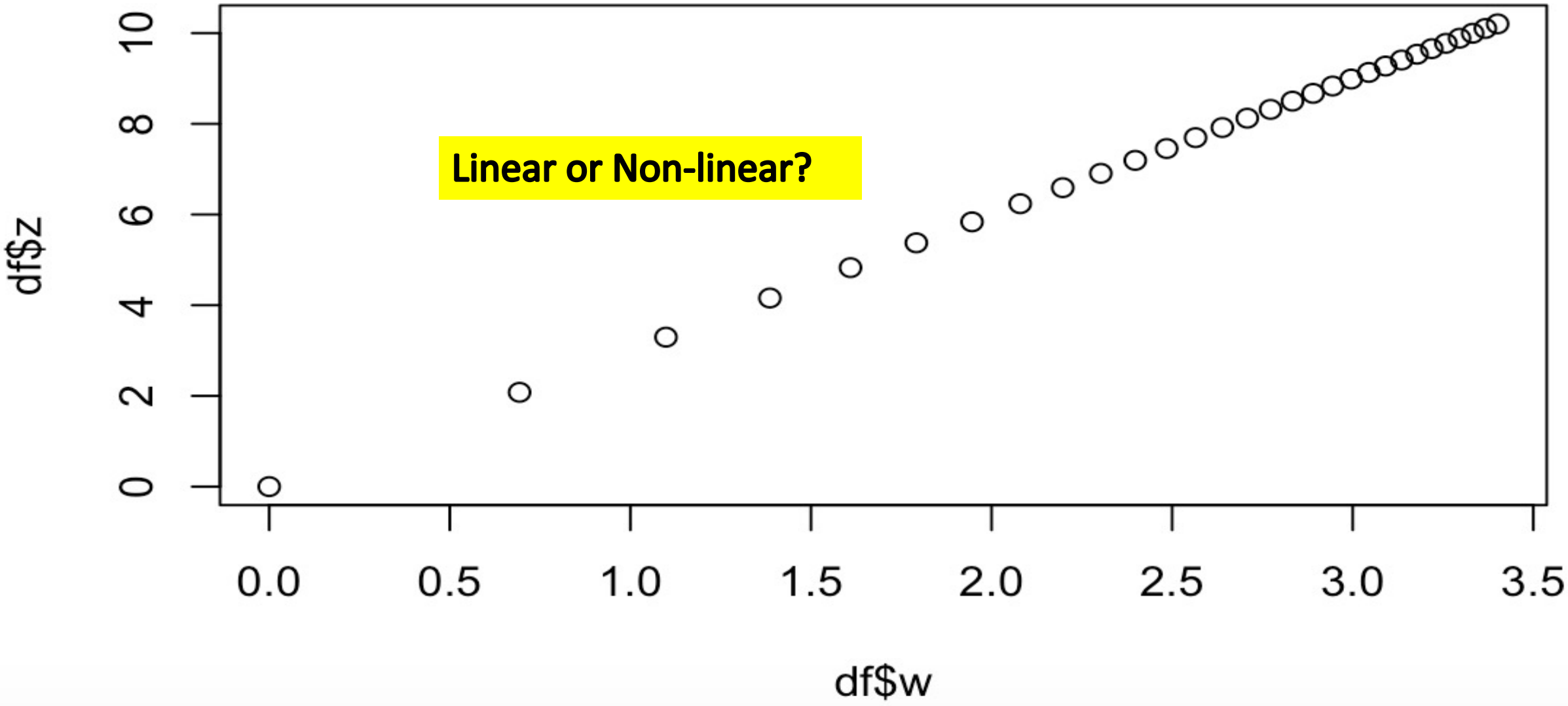
- plot(dfx, dfz)

- How does the graph look now?

Scatterplot of x and log(y)



Exercise1: Scatterplot of log(x) and log(y)



Console Terminal Jobs

~/

```
> #Data Frame
> df <- data.frame(x<-c(1:30), y<-x^3)
> plot(df$x, df$y)
> View(df)
> print(df)
   x....c.1.30. y....x.3
1             1       1
2             2       8
3             3      27
4             4      64
5             5     125
```

```
> colnames(df) <- c('x', 'y')
> View(df)
>
```

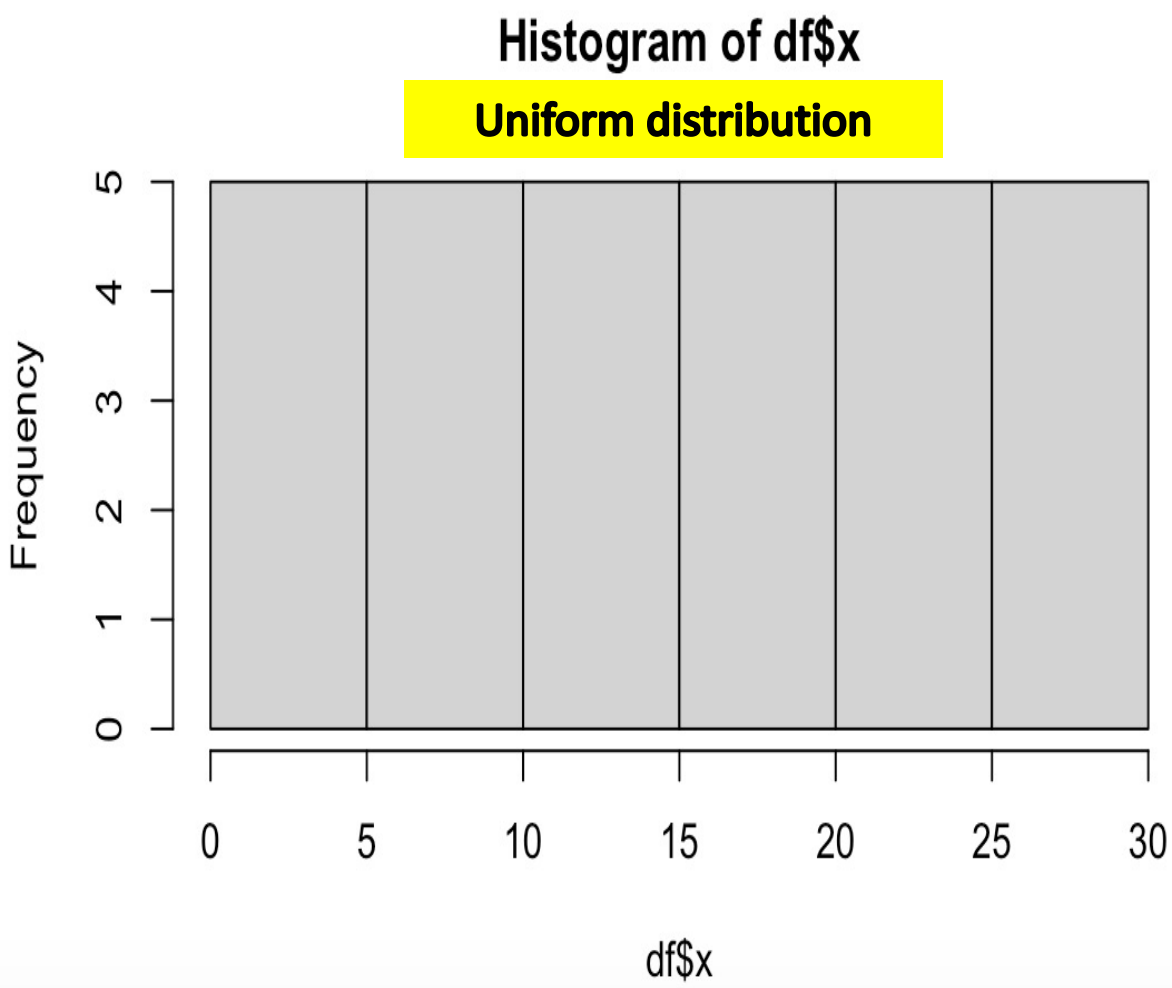
df

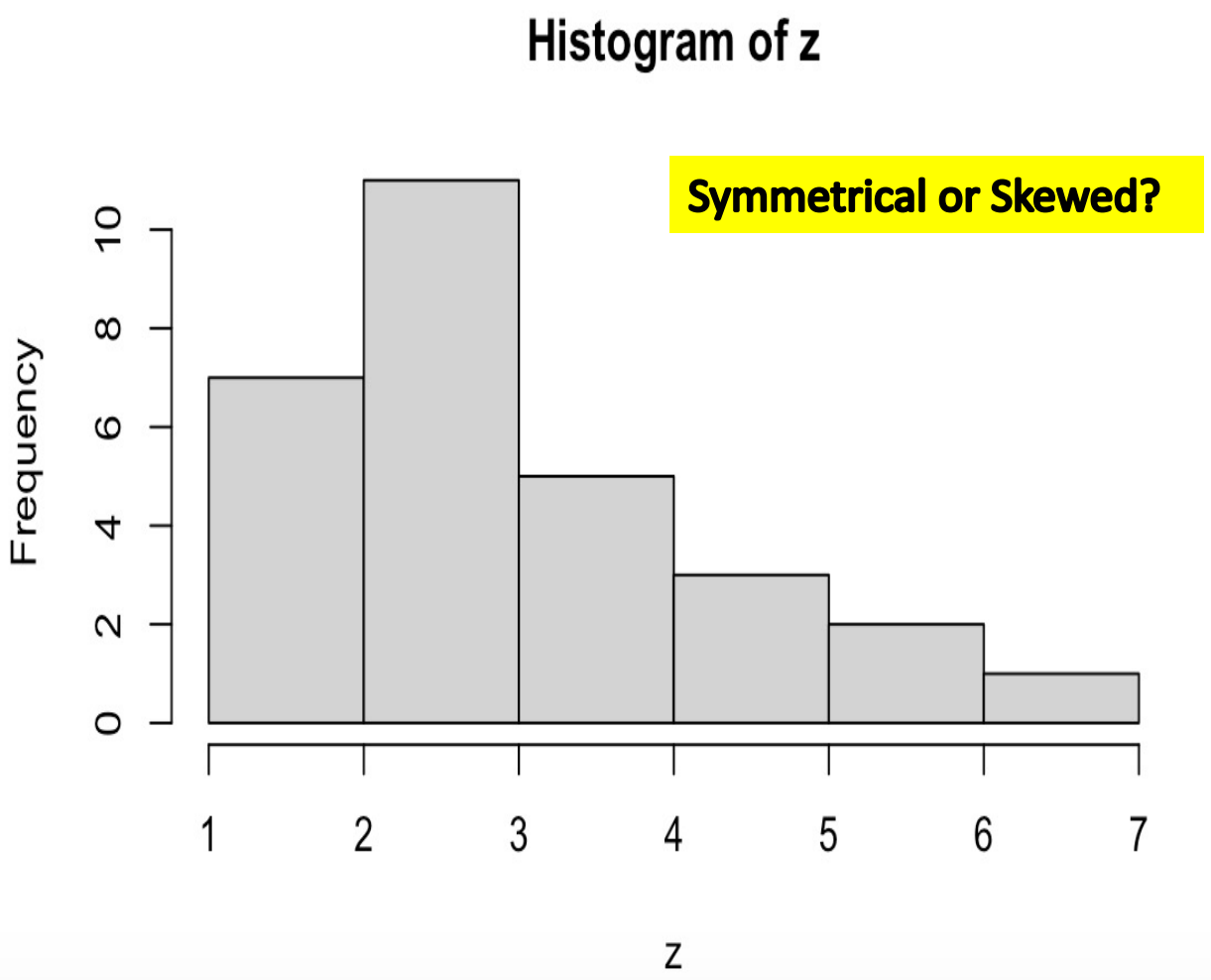
Filter

	x	y
1	1	1
2	2	8
3	3	27
4	4	64
5	5	125
6	6	216
7	7	343
8	8	512
9	9	729
10	10	1000
11	11	1331
12	12	1728
13	13	2197

Showing 1 to 14 of 30 entries, 2 total columns

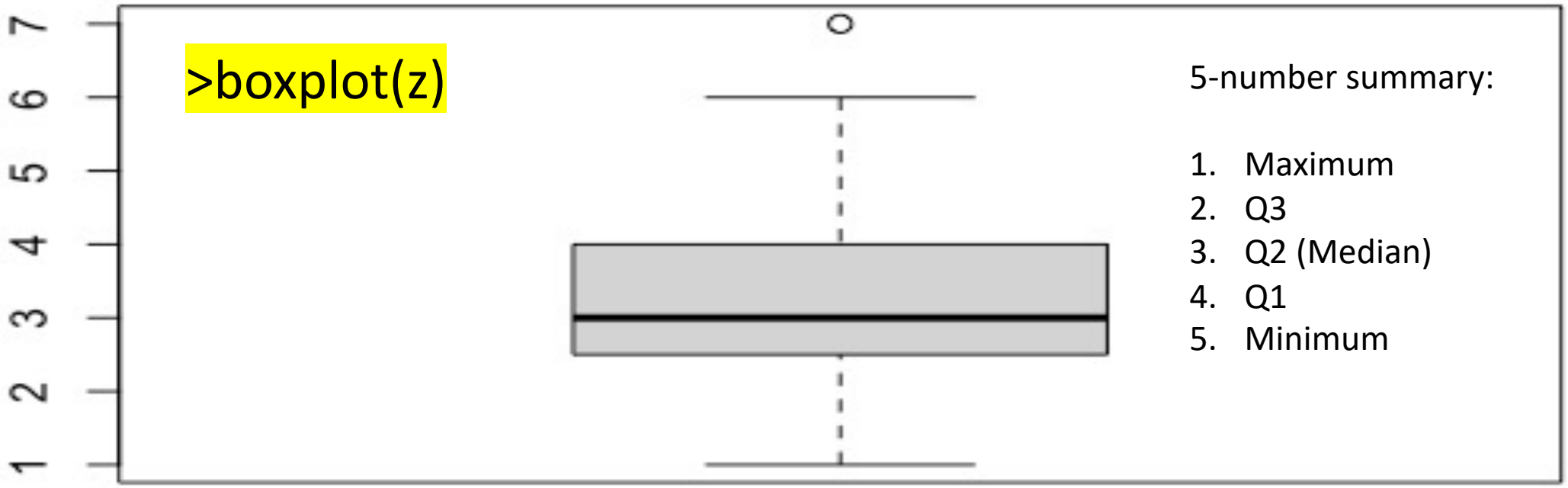
```
> hist(df$x)
> z <- c(1,1,2,2,2,2,2,3,3,3,3,3,3,3,3,3,3,3,4,4,4,4,4,5,5,5,6,6,7)
> hist(z)
>
```





Summary statistics of z variable:

- > summary(z)
- Min. 1st Qu. Median Mean 3rd Qu. Max.
- 1.000 2.750 3.000 3.407 4.000 7.000



5-number summary:

1. Maximum
2. Q3
3. Q2 (Median)
4. Q1
5. Minimum

What is “O” shown in the box and whisker plot? Why is it important in the statistical analysis?

Factors and attributes in R:

- Factor is used to create and store categorical variable in R like Gender (Male/Female), Blood group (A, B, AB, O) and Blood Rh factor (Positive/Negative) etc.

- > gender <- factor(c(“male”, “female”, “female”, “male”)
- > typeof(gender) #datatype
- > attributes(gender) #Levels and class
- > unclass(gender) #Check how it is stored in R

```r
gender <- factor(c("male", "female", "female", "male"))

typeof(gender)
## "integer"

attributes(gender)
## $levels
## [1] "female" "male"
##
## $class
## [1] "factor"
```

You can see exactly how R is storing your factor with `unclass`:

```r
unclass(gender)
## [1] 2 1 1 2
## attr(,"levels")
## [1] "female" "male"
```

https://rstudio-education.github.io/hopr/r-objects.html#attributes

Functions in R: Built-in functions

- round()
  - round(3.1415)
  - 3

round()
round(3.1415, digits = 2)
3.14

- factorial()
  - factorial(3)
  - 6
  - 3! = 3 x 2 x 1

die <- 1:6
sample(x = die, size = 1)
sample(x = die, size = 2)
sample(x = die, size = 2, replace = TRUE)

- mean()
  - mean(1:6)
  - =(1+2+3+4+5+6)/6 = 3.5

mean(die)
mean(round(die))

User-defined function:

- my_function <- function() {}
- Where,
- my_function = name of the function e.g. roll (roll the die)

- function() = telling R that it is a user-defined function

- { = We need to start our code after this braces
- } = We need to close our codes before this braces

User-defined function: roll()

```
roll <- function() {
        die <- 1:6
        dice <- sample(die, size = 2, replace = TRUE)
        sum(dice)
}
```

First roll:	roll()
Second roll:	roll()
Third roll:	roll()

User-defined function: roll2()

```
roll2 <- function(dice = 1:6) {
        dice <- sample(dice, size = 2, replace = TRUE)
        sum(dice)
}
```

First roll:	roll2()
Second roll:	roll2()
Third roll:	roll2()

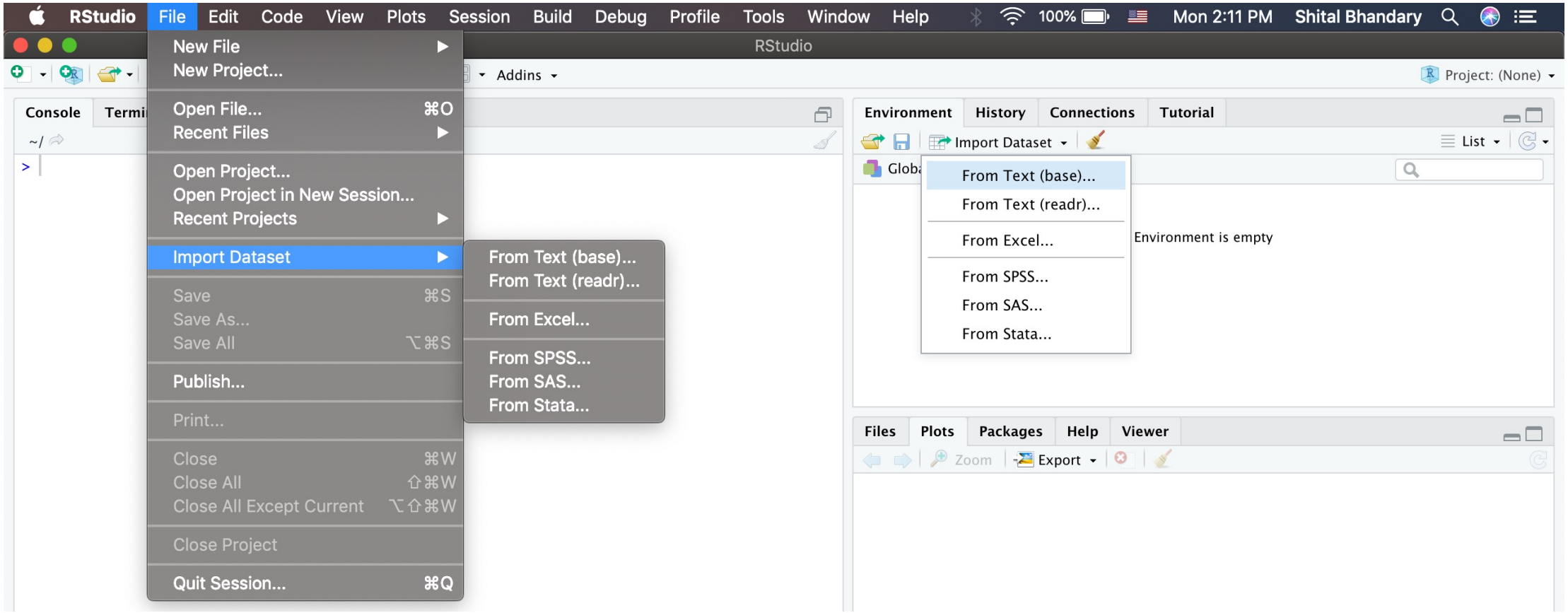
User-defined function: roll3(dice = ?:?)

```
roll3 <- function(dice) {
        dice <- sample(dice, size = 2, replace = TRUE)
        sum(dice)
}
```

First roll:	roll3(dice = 1:6)
Second roll:	roll3(dice = 1:12)
Third roll:	roll3(dice = 1:24)

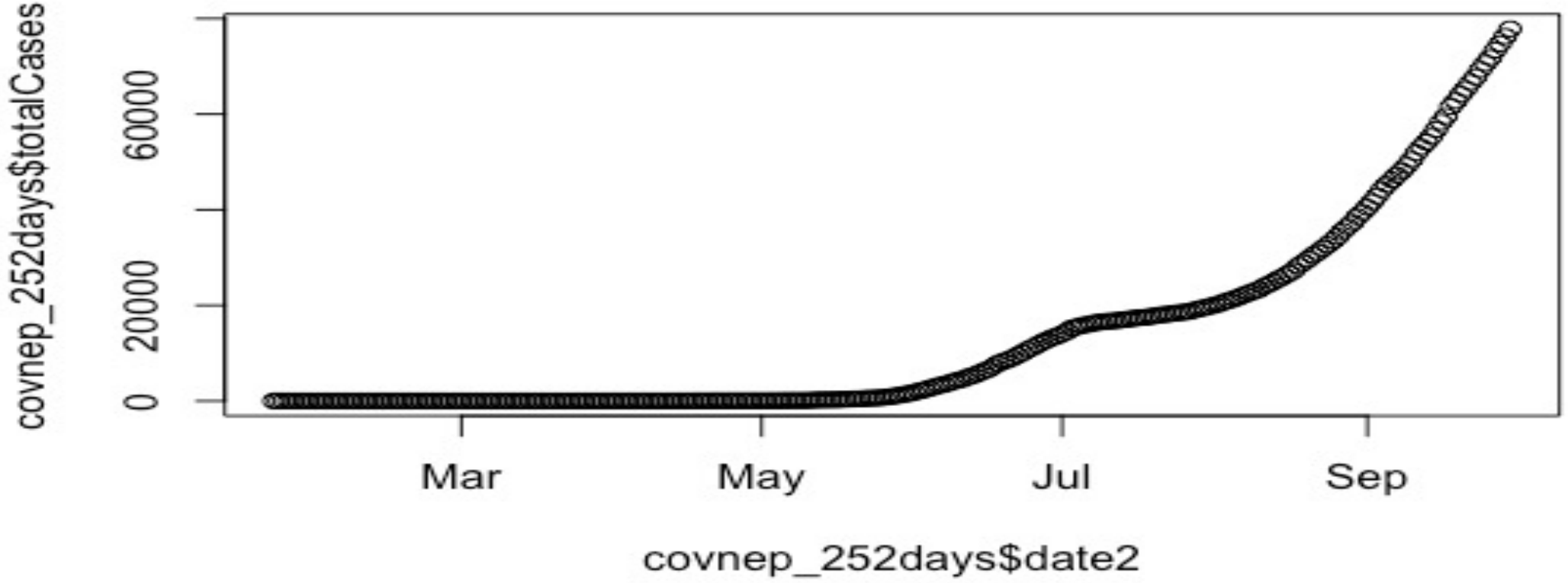
Questions/queries?

Import “covnep_252days.csv” data in R Studio: I recommend the “readr” package



Then get this chart in R Studio:

Cumulative COVID-19 cases in Nepal: First 252 days since onset at 23/01/2021



Then get summary of “totalCases” variable:

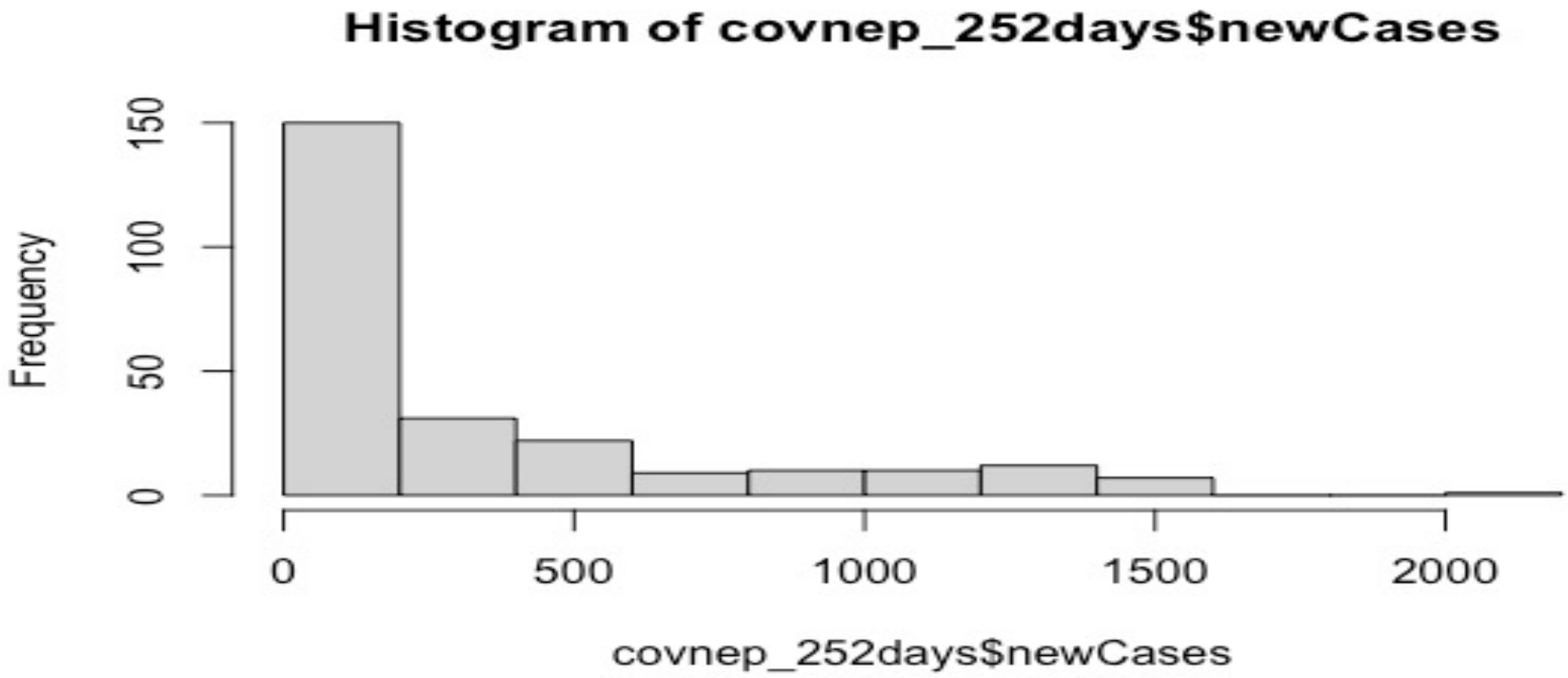
- > summary(covnep_252days$totalCases)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0	2	963	13376	19340	77816

- What is the problem with this result?

- The minimum value can’t be 0 as first case was detected on 23 Jan 2020 (first case in the data) so the minimum cumulative case must be 1. What to do now?

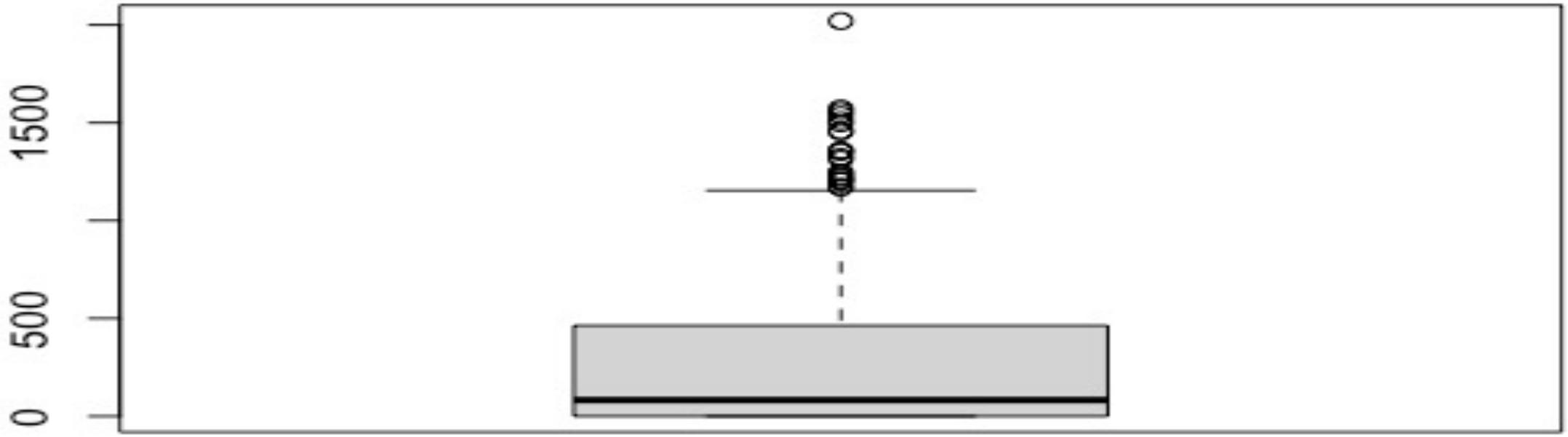
Then get this chart in R Studio:



Then get summary of “newCases” variable and interpret the result carefully:

- > summary(covnep_252days$newCases)

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0	0.0	82.5	308.8	463.2	2020.0



Import “SAQ8.sav” data in R Studio and get frequencies of q01, q03, q06 & q08 variables:

Statistics makes me cry					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly agree	270	10.5	10.5	10.5
	Agree	1338	52.0	52.0	62.5
	Neither	735	28.6	28.6	91.1
	Disagree	187	7.3	7.3	98.4
	Strongly disagree	41	1.6	1.6	100.0
	Total	2571	100.0	100.0	

Standard deviations excite me					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly agree	497	19.3	19.3	19.3
	Agree	672	26.1	26.1	45.5
	Neither	878	34.2	34.2	79.6
	Disagree	448	17.4	17.4	97.0
	Strongly disagree	76	3.0	3.0	100.0
	Total	2571	100.0	100.0	

I have little experience of computers					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly agree	702	27.3	27.3	27.3
	Agree	1127	43.8	43.8	71.1
	Neither	344	13.4	13.4	84.5
	Disagree	252	9.8	9.8	94.3
	Strongly disagree	146	5.7	5.7	100.0
	Total	2571	100.0	100.0	

I have never been good at mathematics					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Strongly agree	383	14.9	14.9	14.9
	Agree	1487	57.8	57.8	72.7
	Neither	482	18.7	18.7	91.5
	Disagree	147	5.7	5.7	97.2
	Strongly disagree	72	2.8	2.8	100.0
	Total	2571	100.0	100.0	

Hint: Base R does not have any function for it!

You can write your own function?

- 'count' function included in the 'plyr' package is very helpful

- install.packages("plyr") #Install the 'plyr' package
- library(plyr) #Load the 'plyr' package

- count(SAQ8, 'q01') #It will give you the frequencies

- How to get percentage? #Individual assignment

Import “MR_drugs.xls” file in R Studio and get the following table: MR variables are binary!

$Income Frequencies				
		Responses		Percent of Cases
		N	Percent	
Income - Multiple Response[a]	inco1	226	12.8%	23.5%
	inco2	607	34.5%	63.0%
	inco3	293	16.6%	30.4%
	inco4	50	2.8%	5.2%
	inco5	82	4.7%	8.5%
	inco6	151	8.6%	15.7%
	inco7	352	20.0%	36.6%
Total		1761	100.0%	182.9%

a. Dichotomy group tabulated at value 1.

Question/queries?

Thank you!

@shitalbhandary