# Statistical Computing with R Masters in Data Science 503 (S2) Second Batch, SMS, TU, 2023

Shital Bhandary

Associate Professor

Statistics/Bio-statistics, Demography and Medical Education

Patan Academy of Health Sciences, Lalitpur, Nepal

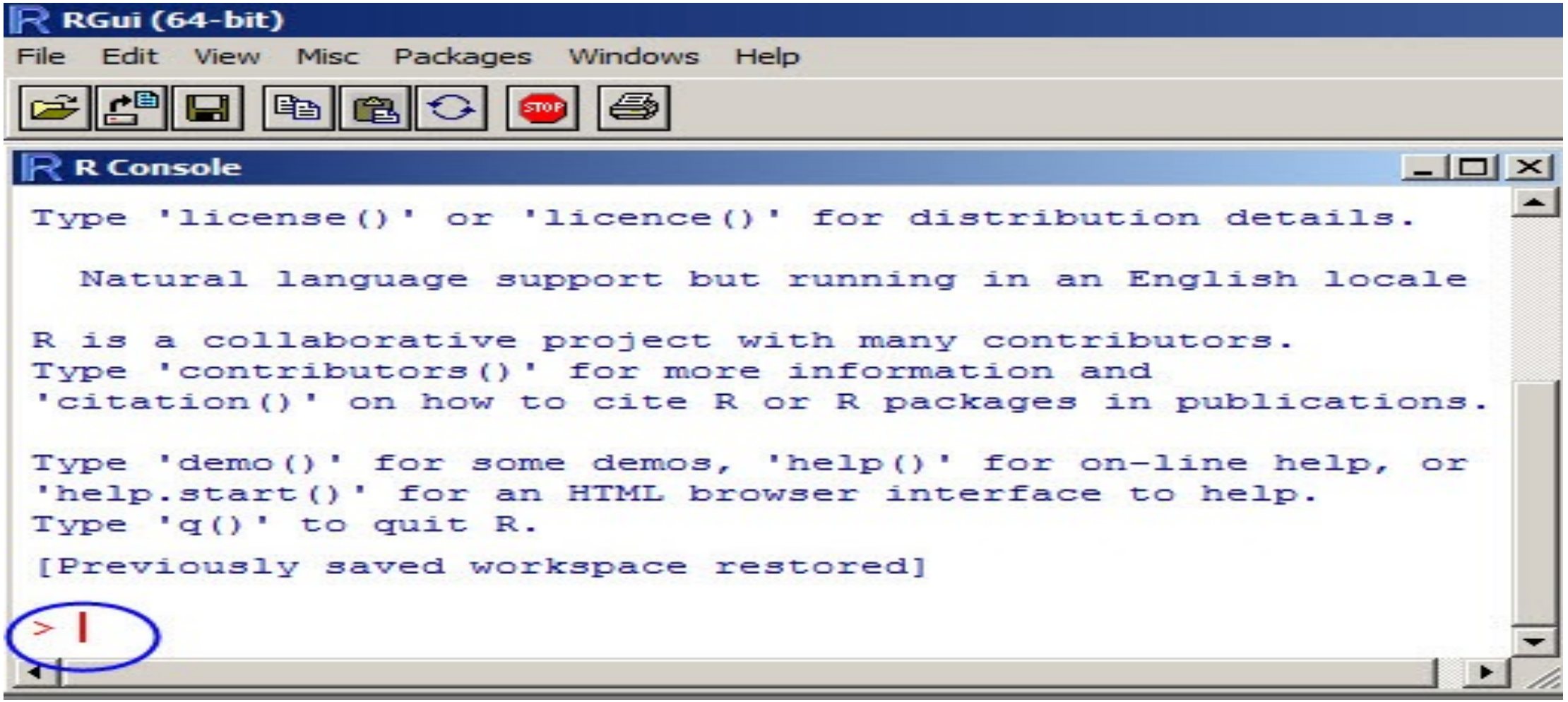
Faculty, Data Analysis and Decision Modeling, MBA, Pokhara University, Nepal

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

# Review Preview

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

# R console in Windows OS



# R Console in Mac OS:

R Console

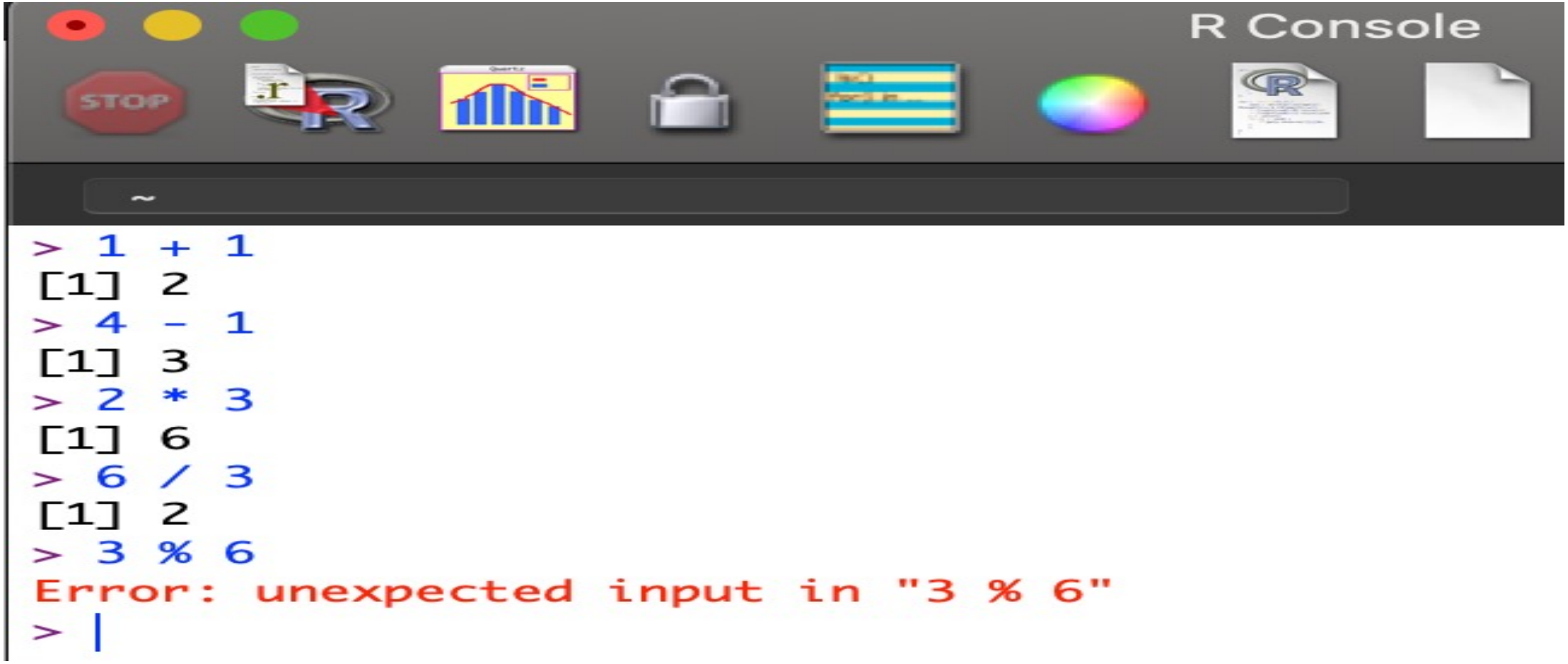
```
R version 4.0.2 (2020-06-22) -- "Taking Off Again"
Copyright (C) 2020 The R Foundation for Statistical Computing
Platform: x86_64-apple-darwin17.0 (64-bit)

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]

>
```

This is after installing the X-11 in MacOS

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 55% Wed 6:55 AM Shital Bhandary

RStudio

Go to file/function | Addins | Project: (None)

Untitled1

Source on Save | Run | Source

1

R Script

1:1 (Top Level) R Script

Console Terminal Jobs

R 4.1.1 · ~/

```
R version 4.1.1 (2021-08-10) -- "Kick Things"
Copyright (C) 2021 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 redistri                    ditions.
Type 'license()' or 'licenc                    ails.

  Natural language support                     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.
```

R Console

Environment History Connections Tutorial

Import Dataset | 90 MiB | List

R | Global Environment

Environment is empty

Objects, Data etc.

Files Plots Packages Help Viewer

Install | Update

| | Name | Description | Version |
|---|---|---|---|
| **System Library** | | | |
| | abind | Combine Multidimensional Arrays | 1.4-5 |
| | ada | The R Package Ada for Stochastic Boosting | 2.0-5 |
| | additivityTests | Additivity Tests in the Two Way Anova with Single Sub-class Numbers | 1.1-4 |
| | amap | Another Multidimensional Analysis Package | 0.8-18 |
| | aplpack | Another Plot Package: 'Bagplots', 'Iconplots', 'Summary ers | 1.3.5 |
| | arm | Data Analy Multilevel/ | 1.12-2 |
| | arules | Mining As emsets | 1.7-2 |
| | arulesViz | Visualizing Association Rules and Frequent Itemsets | 1.5-1 |
| | askpass | Safe Password Entry for R, Git, and SSH | 1.1 |
| | assertthat | Easy Pre and Post Assertions | 0.2.1 |
| | babynames | US Baby Names 1880-2017 | 1.0.1 |
| | backports | Reimplementations of Functions Introduced Since R-3.0.0 | 1.2.1 |

aplpack

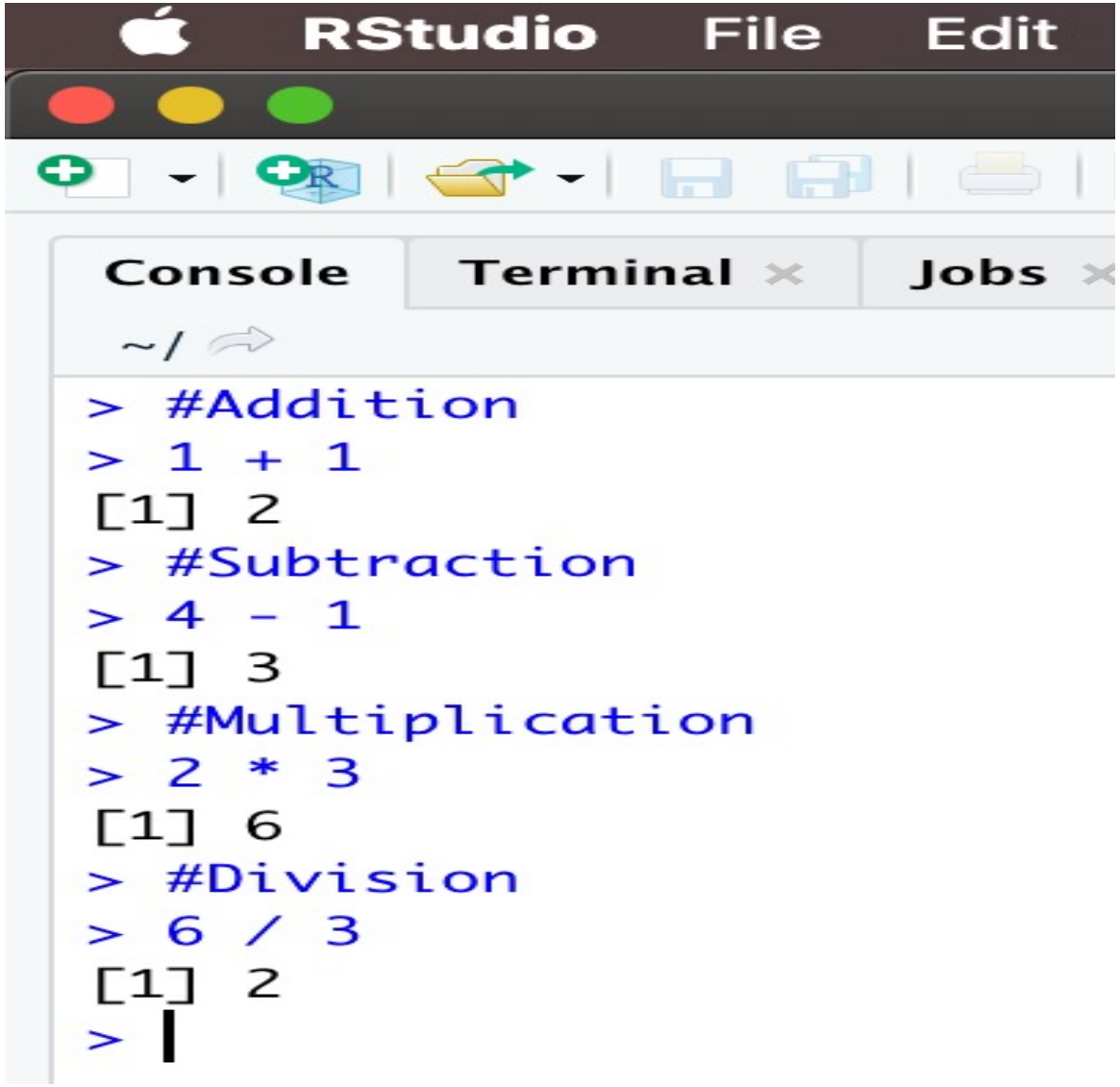
Outputs

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:




The comments are added using #

It is a good practice to use comments before any code

This will help us to understand our codes better when we re-visit them after a long time

# 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

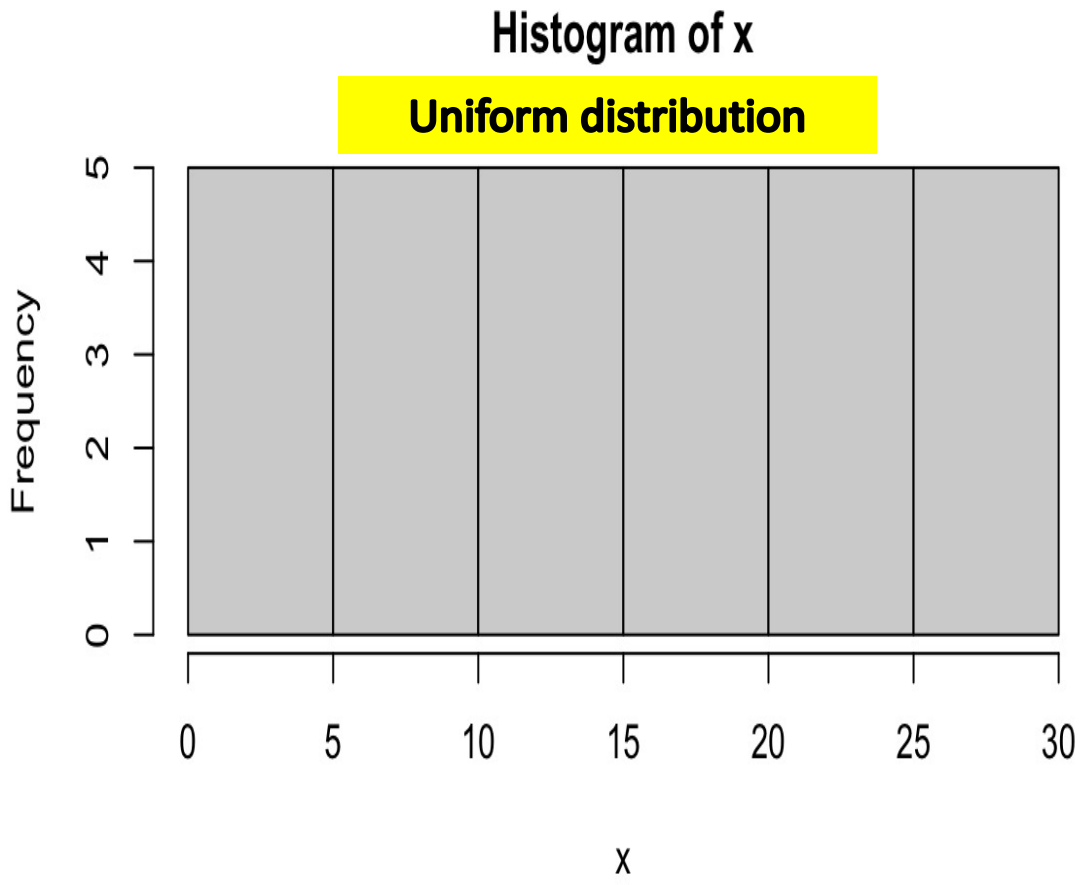
| | |
|---|---|
| x | int [1:30] 1 2 3 4 5 6 7 8 9 10 ... |
| y | num [1:30] 1 8 27 64 125 216 343 512 729 1000 ... |

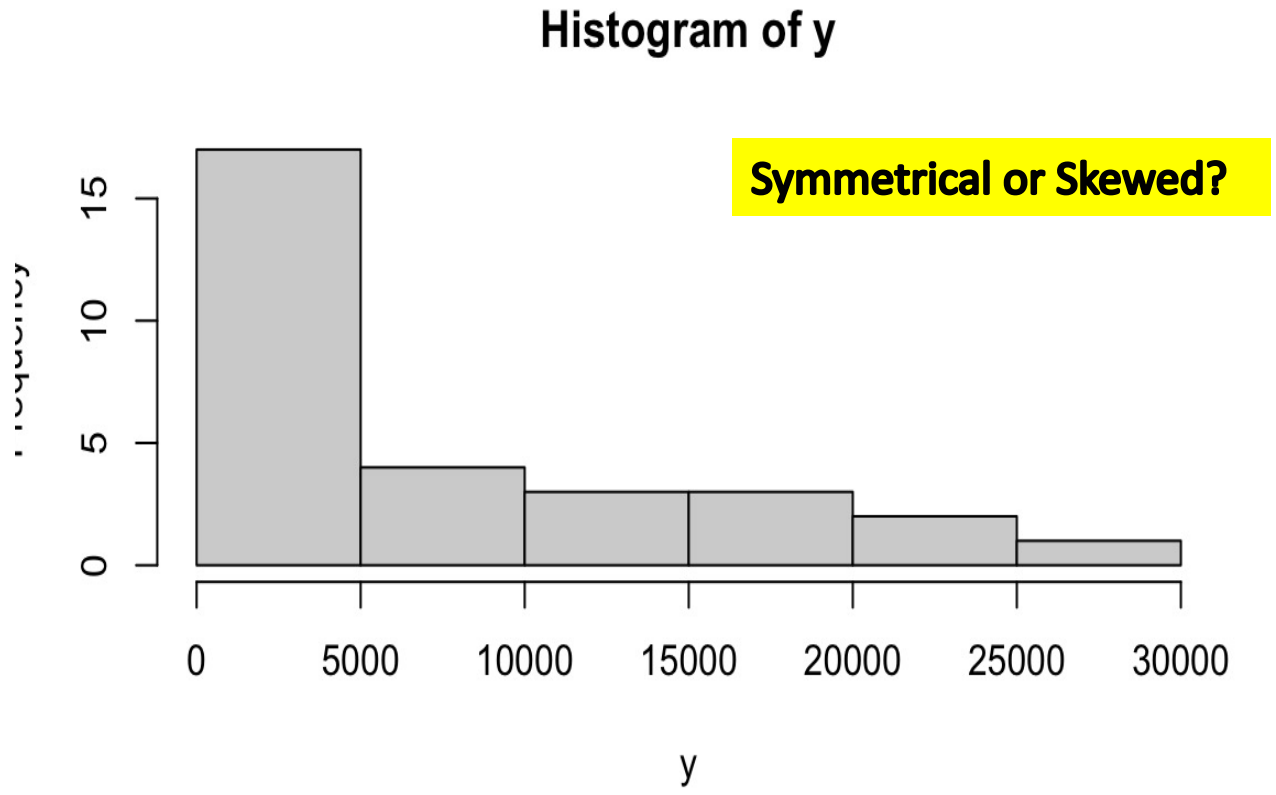
# Get the summary of x and y variables:

- summary(x)
- summary(y)
- Think and decide:
  - Which measure of central tendency must be used? Why?
  - Which measure of dispersion must be used? Why?
  - How to define the outliers using the central tendency and dispersion values?

# hist(x) and hist(y)

## How the class interval was formed here?

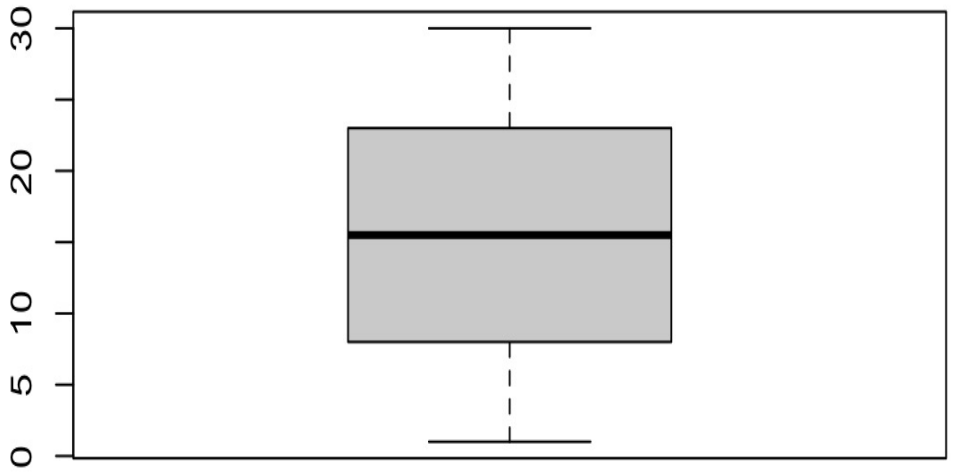


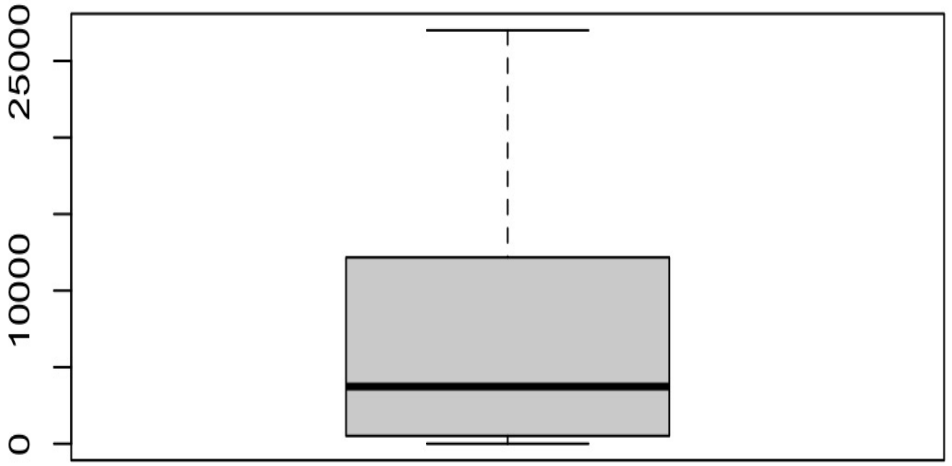
# Quick Think!

- What is a five number summary?
- How is it different from the Tukey's five number summary
- How to represent five number summary using visualization?
- How to represent Tukey's five number summary using graph?

# Boxplot of x and y variables:
# How to interpret these plots?




boxplot(x)




boxplot(y)

# Bi-variate analysis

- When we use two variables to find relationship or association then it is known as bi-variate analysis

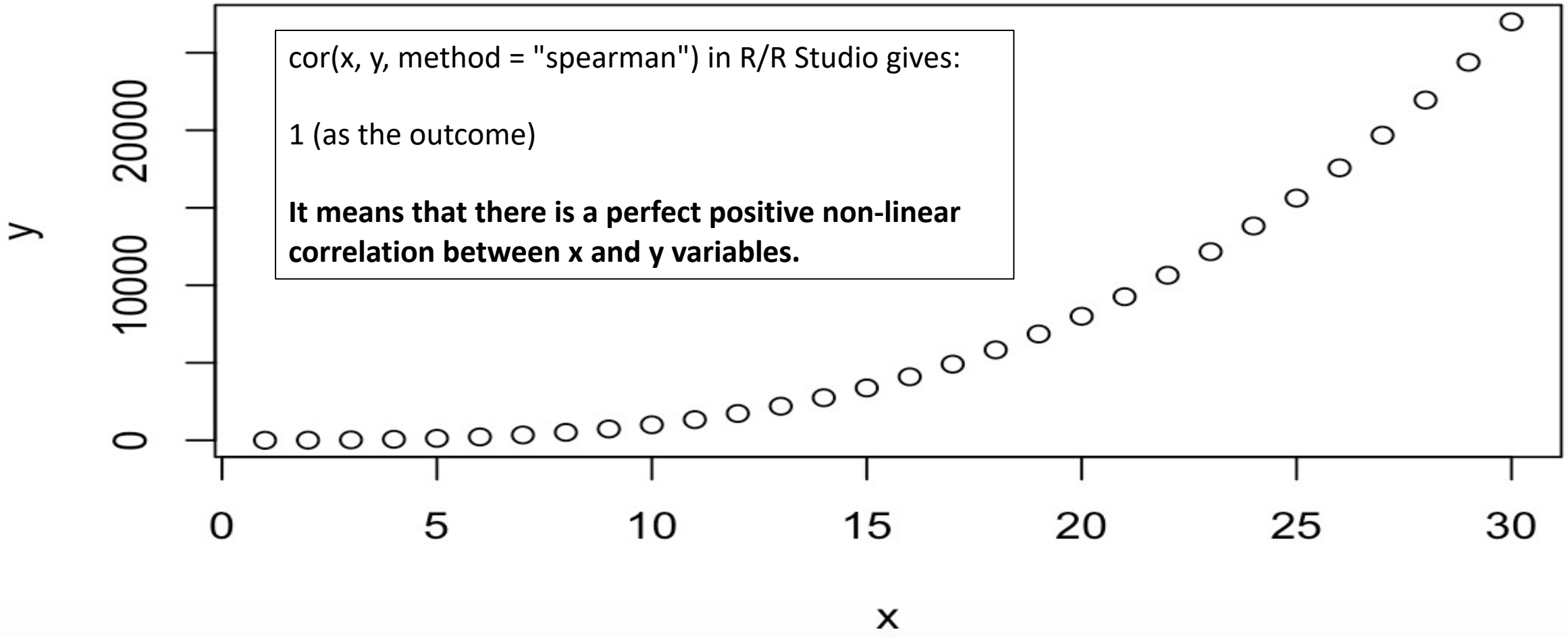
- The x and y variables created earlier are numerical variables and they need to be assessed with “Scatterplot” before correlation coefficient

- We can find relationship between these variables using:
  - Pearson correlation coefficient – if they show linear relationship
  - Spearman’s correlation coefficient – if they show non-linear relationship

Files Plots Packages Help Viewer

Zoom Export Publish

This scatterplot shows a non-linear relationship and we must used Spearman's rank correlation coefficient




cor(x, y, method = "spearman") in R/R Studio gives:

1 (as the outcome)

**It means that there is a perfect positive non-linear correlation between x and y variables.**

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 ... |

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
```

We need to change the variable name as “x” and “y”

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

df

Filter

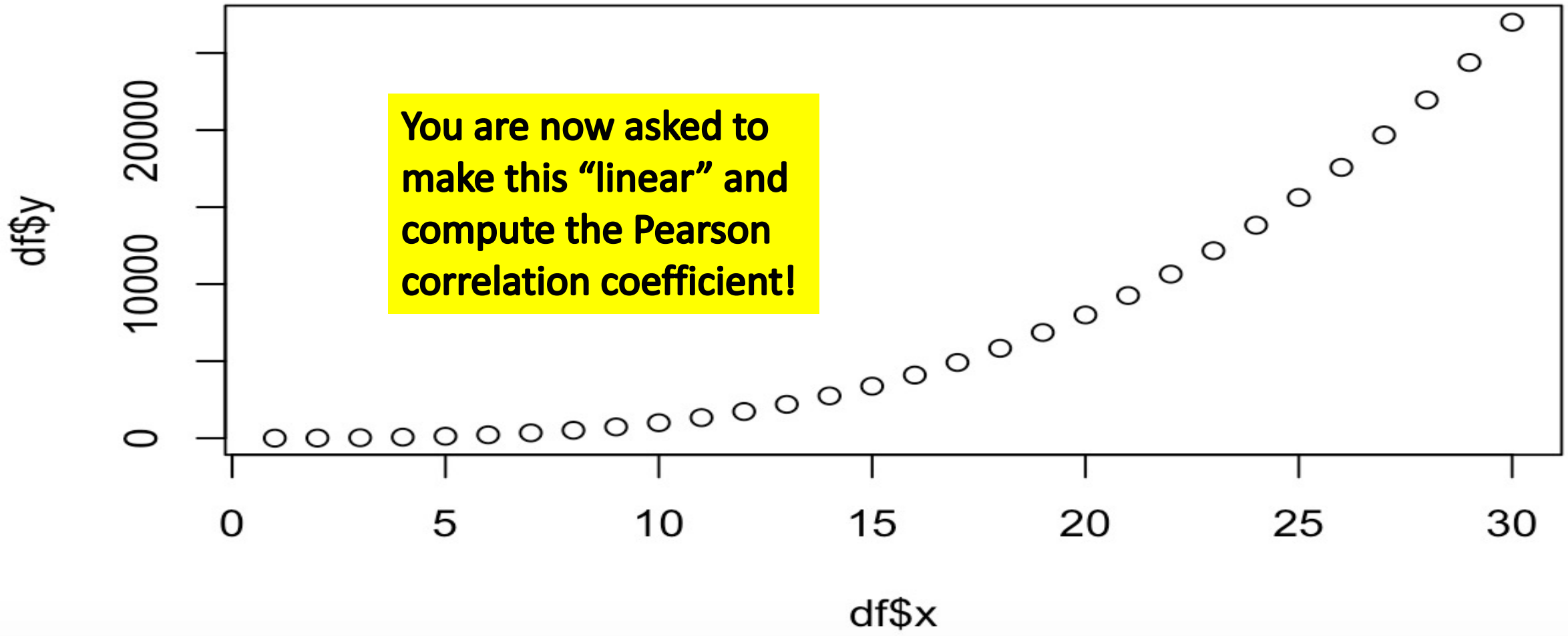
| | x | y |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 8 |
| 3 | 3 | 27 |
| 4 | 4 | 64 |
| 5 | 5 | 125 |
| 6 | 6 | 216 |
| 7 | 7 | 343 |
| 8 | 8 | 512 |
| 9 | 9 | 729 |
| 10 | 10 | 1000 |
| 11 | 11 | 1331 |
| 12 | 12 | 1728 |
| 13 | 13 | 2197 |

Showing 1 to 14 of 30 entries, 2 total columns

Files Plots Packages Help Viewer

Zoom Export Publish

cor(df$x,df$y)
Pearson= 0.92011
This is a biased estimate as the relationship is not linear!




**You are now asked to make this "linear" and compute the Pearson correlation coefficient!**

# 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:

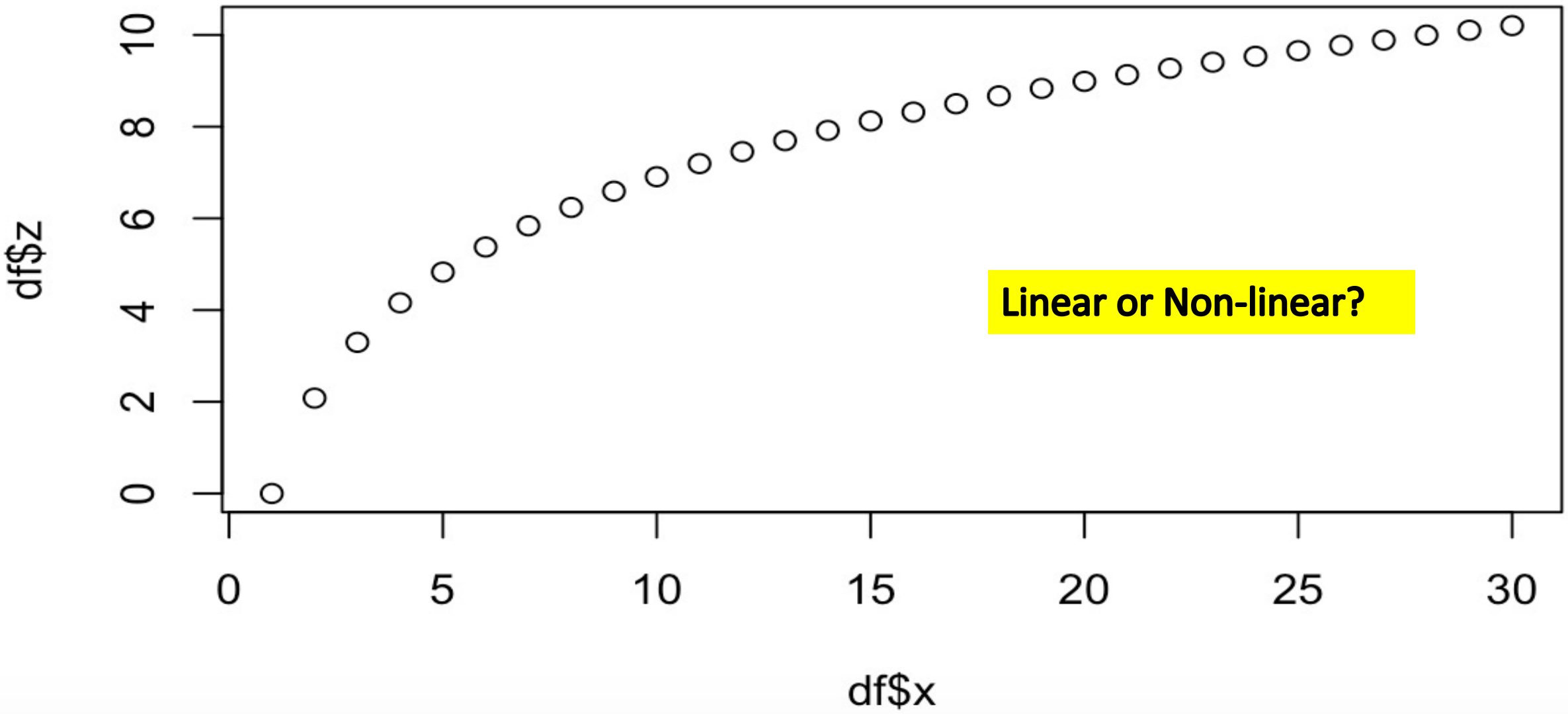
- df$z <- log(df$y)

- Let us plot the scatterplot again as:

- plot(df$x, df$z)

- How does the graph look now?

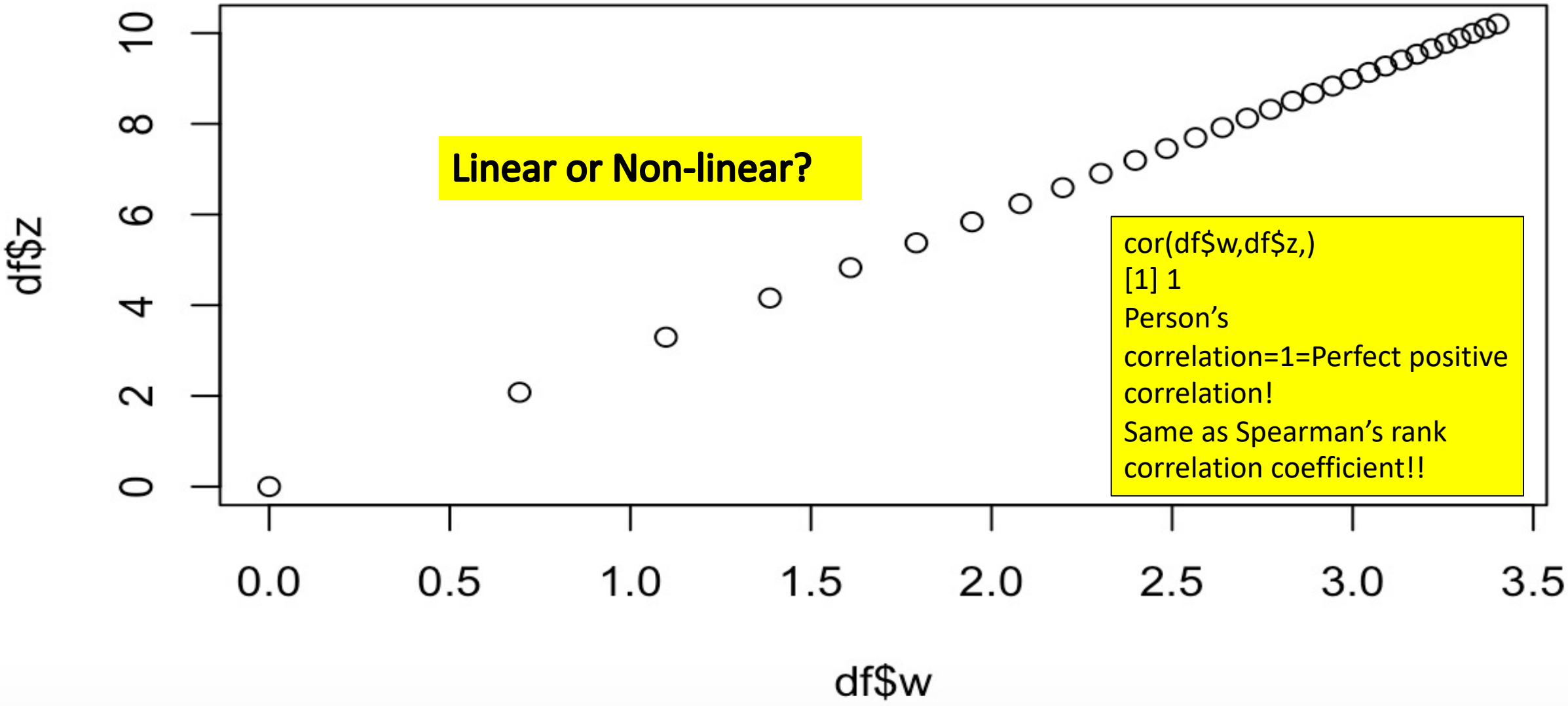
# Scatterplot of x and log(y)



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

- Let us define w as log(x) in r as follows:
- df$w <- log(df$x)
- Let us plot the scatterplot again as:
- plot(df$w, df$z)
- How does the graph look now?

# Scatterplot of log(x) and log(y)
# This is called log-log transformation!




**Linear or Non-linear?**

cor(df$w,df$z,)
[1] 1
Person's correlation=1=Perfect positive correlation!
Same as Spearman's rank correlation coefficient!!

# Questions/queries?

```
> 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)
```

**Work/Assignment 1:**

Show the histogram of z variable and interpret it carefully.

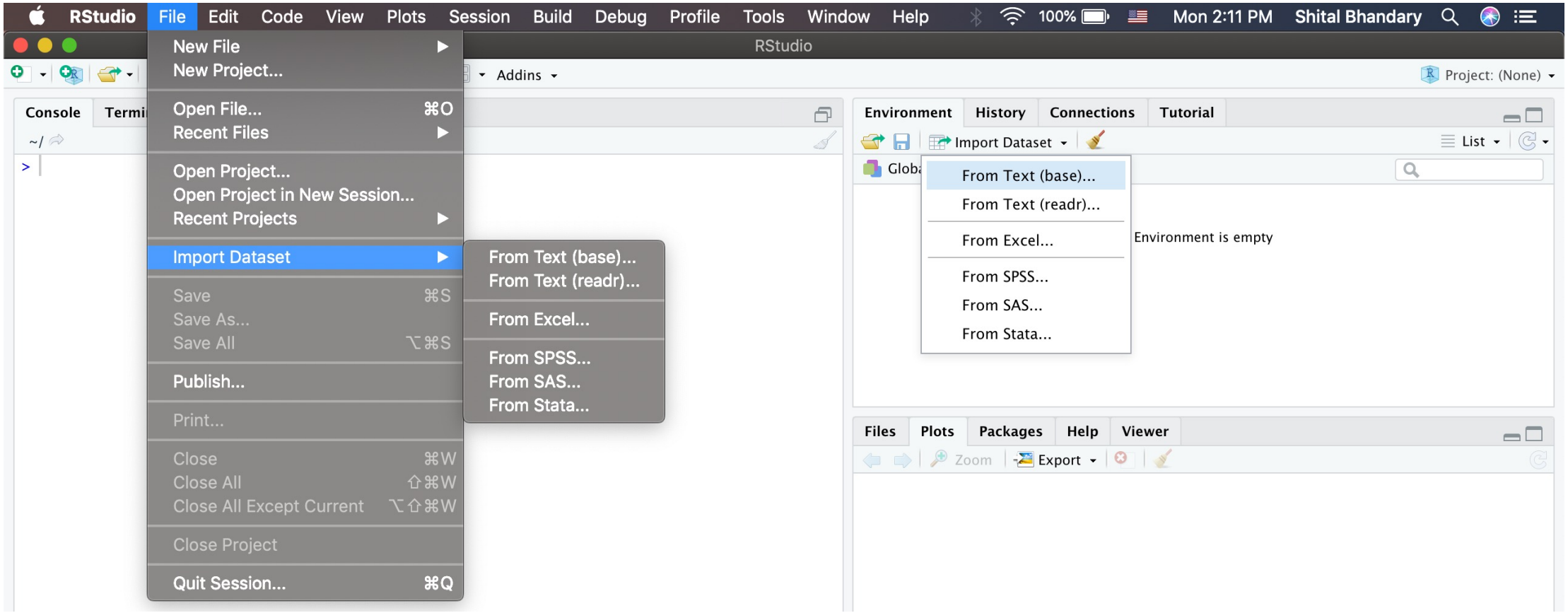
Get summary of this variable and decide which measure of central tendency and measure of dispersion must be used for this variable?

Get the five number summary of this variable and interpret them carefully.

Create boxplot of this variable and interpret it carefully.

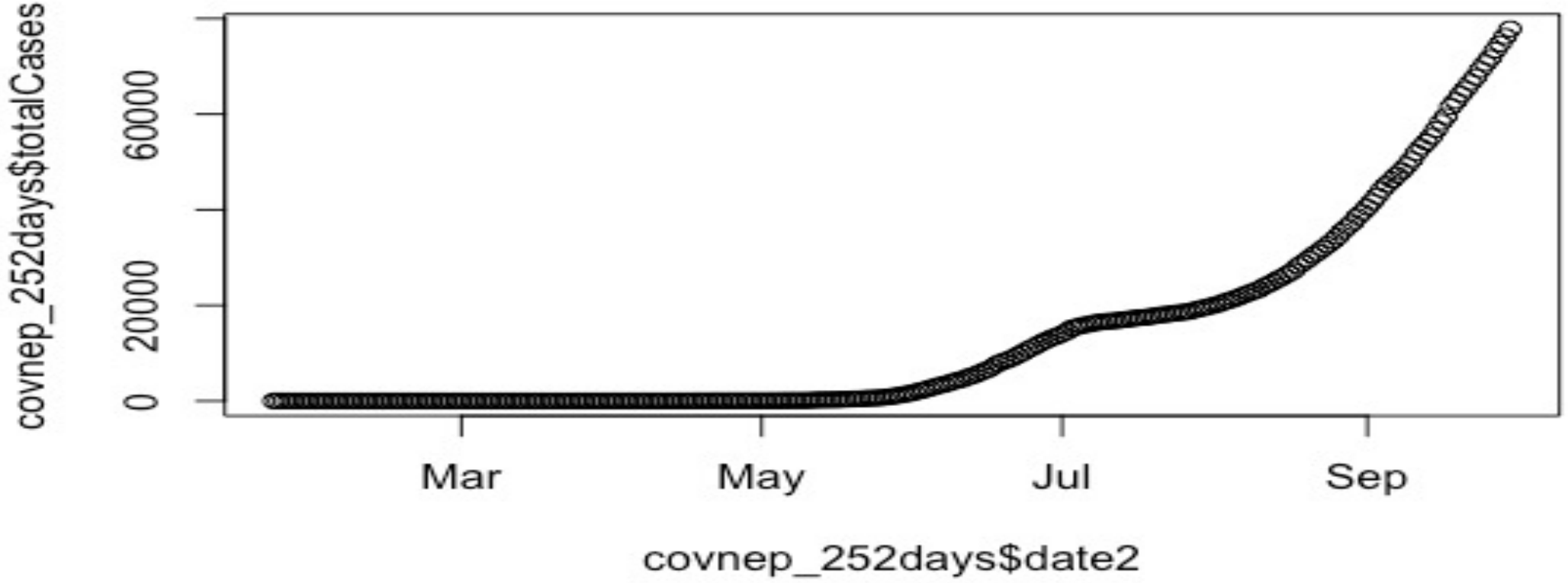
Do you get outlier for this variable in the boxplot? Why?

# Work2: 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



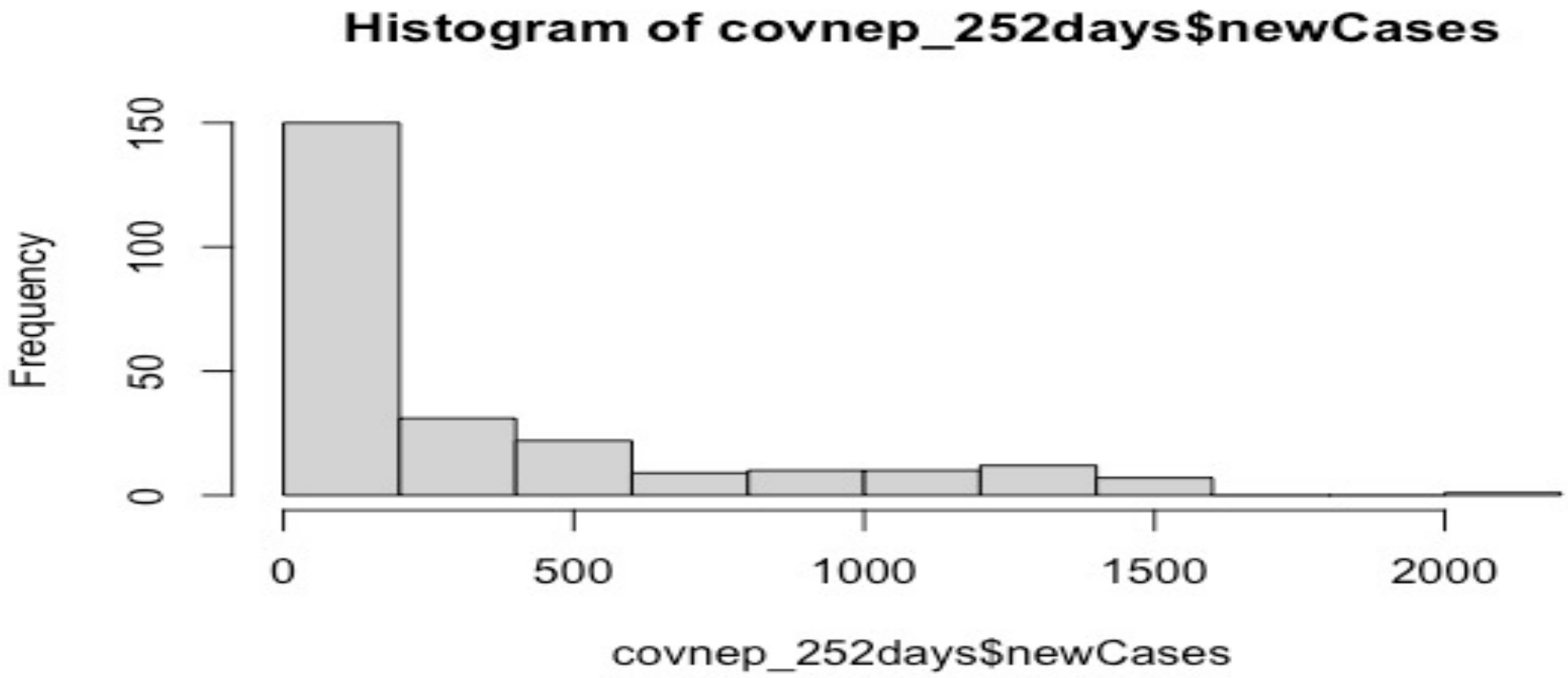
# Then get summary of “totalCases” variable:

- > 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?**
- Fix the problem and get the summary again.
- Interpret the revised summary carefully.

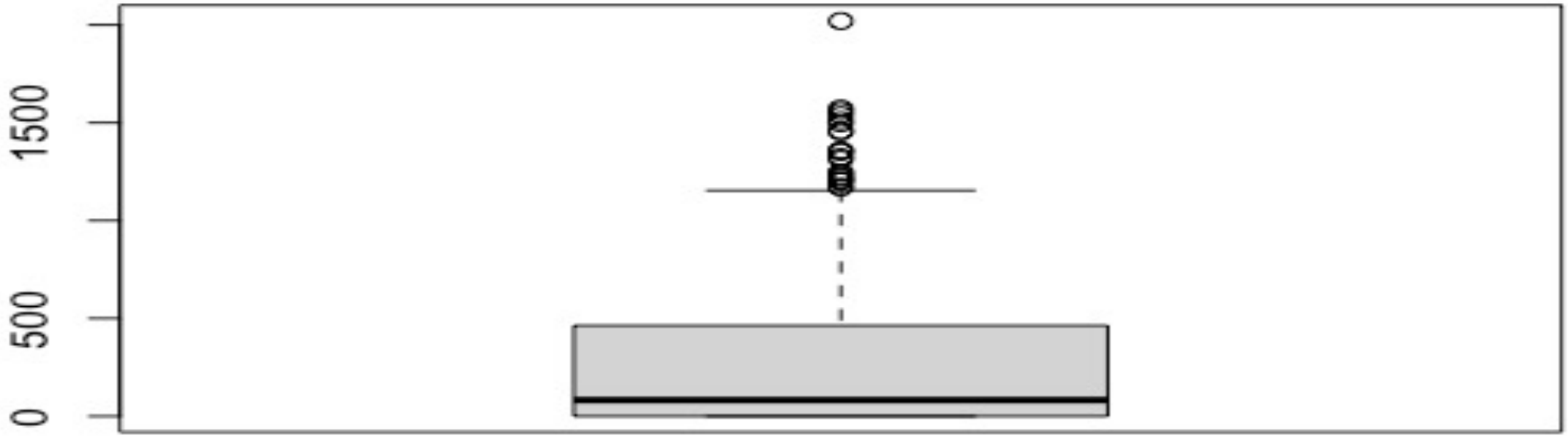
# Then get this chart of ‘newCases’ in R Studio:



# 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 |



# Work 3: Import “SAQ8.sav” data and replicate these tables in R Studio with your own code

| 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 | |

**.sav is a SPSS data file**

# Work 4: Import “MR_drugs.xls” file and replicate the following table in R Studio with you own code

| $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. | | | | |

# Submit these works/assignments here:

- We will use Google Classroom for now as MS Team will take time
- Please send a black email to me so that I can add you in the Google Classroom:
- shitalbhandary@gmail.com
- Our next class will be on Wednesday 29 March 2023 from 6:30 am till 9:30 am

# Question/queries?

# Thank you!

@shitalbhandary