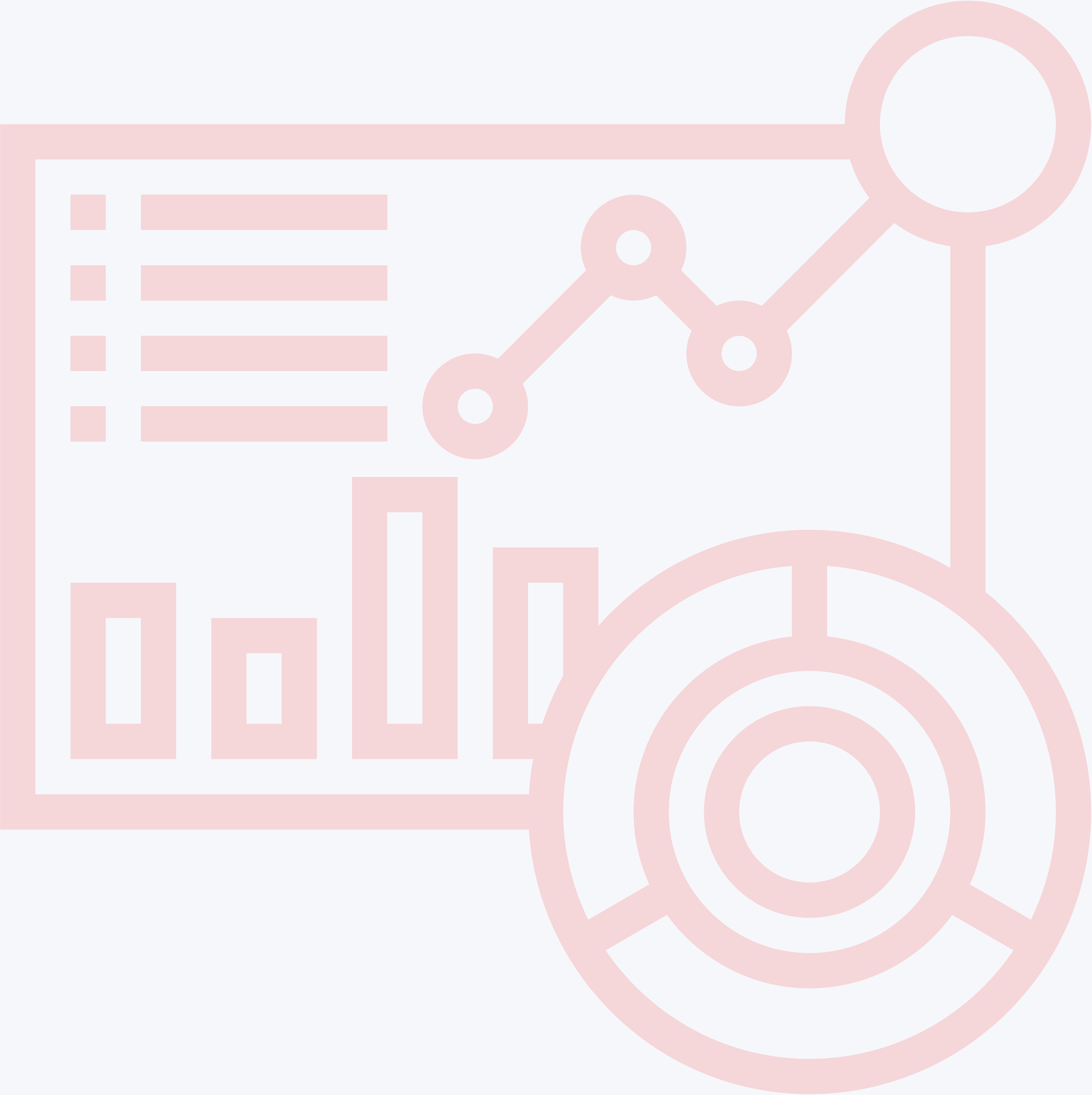


Summarizing Data

Statistics Tutorial

Day 3

Prabesh Dhakal

2020 April 23

WHAT ARE WE DOING TODAY?

RECAP + Q&A

We briefly revisit the contents from last week.

SUMMARIZING DATA

We talk about how we can summarize data.

EXERCISE

We analyze some data.

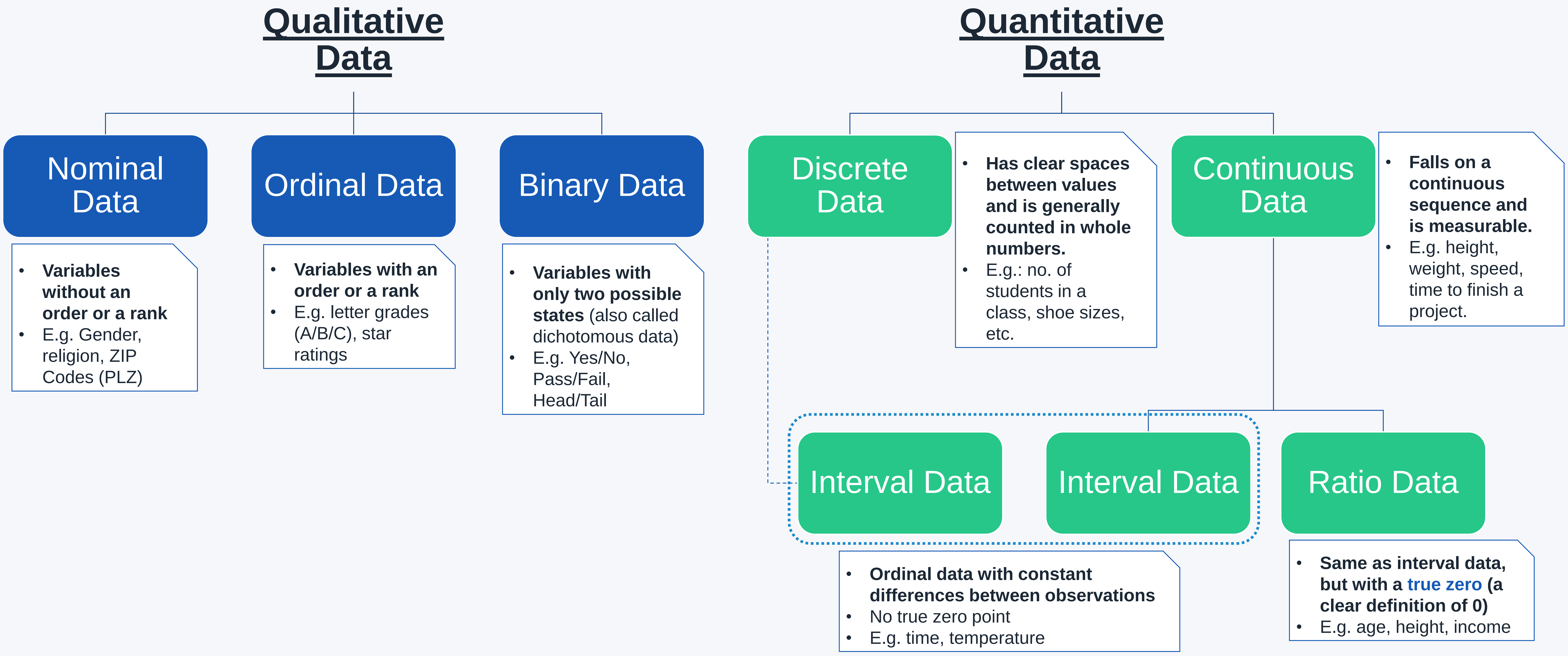
Q&A and Recap

Please ask if you have any questions now.

Otherwise, we can move on to the recap.

Note: you might want to grab a pen, paper & calculator for today’s session.

BASIC TYPES OF DATA IN STATISTICS



Qualitative Data

Nominal Data

- **Variables without an order or a rank**
- E.g. Gender, religion, ZIP Codes (PLZ)

Ordinal Data

- **Variables with an order or a rank**
- E.g. letter grades (A/B/C), star ratings

Binary Data

- **Variables with only two possible states** (also called dichotomous data)
- E.g. Yes/No, Pass/Fail, Head/Tail

Quantitative Data

Discrete Data

- **Has clear spaces between values and is generally counted in whole numbers.**
- E.g.: no. of students in a class, shoe sizes, etc.

Continuous Data

- **Falls on a continuous sequence and is measurable.**
- E.g. height, weight, speed, time to finish a project.

Interval Data

Interval Data

- **Ordinal data with constant differences between observations**
- No true zero point
- E.g. time, temperature

Ratio Data

- **Same as interval data, but with a true zero (a clear definition of 0)**
- E.g. age, height, income

🛠 CLASS EXERCISE - 1 🛠

TASK:

Please identify the type of data each column from the data set on the left side contains:

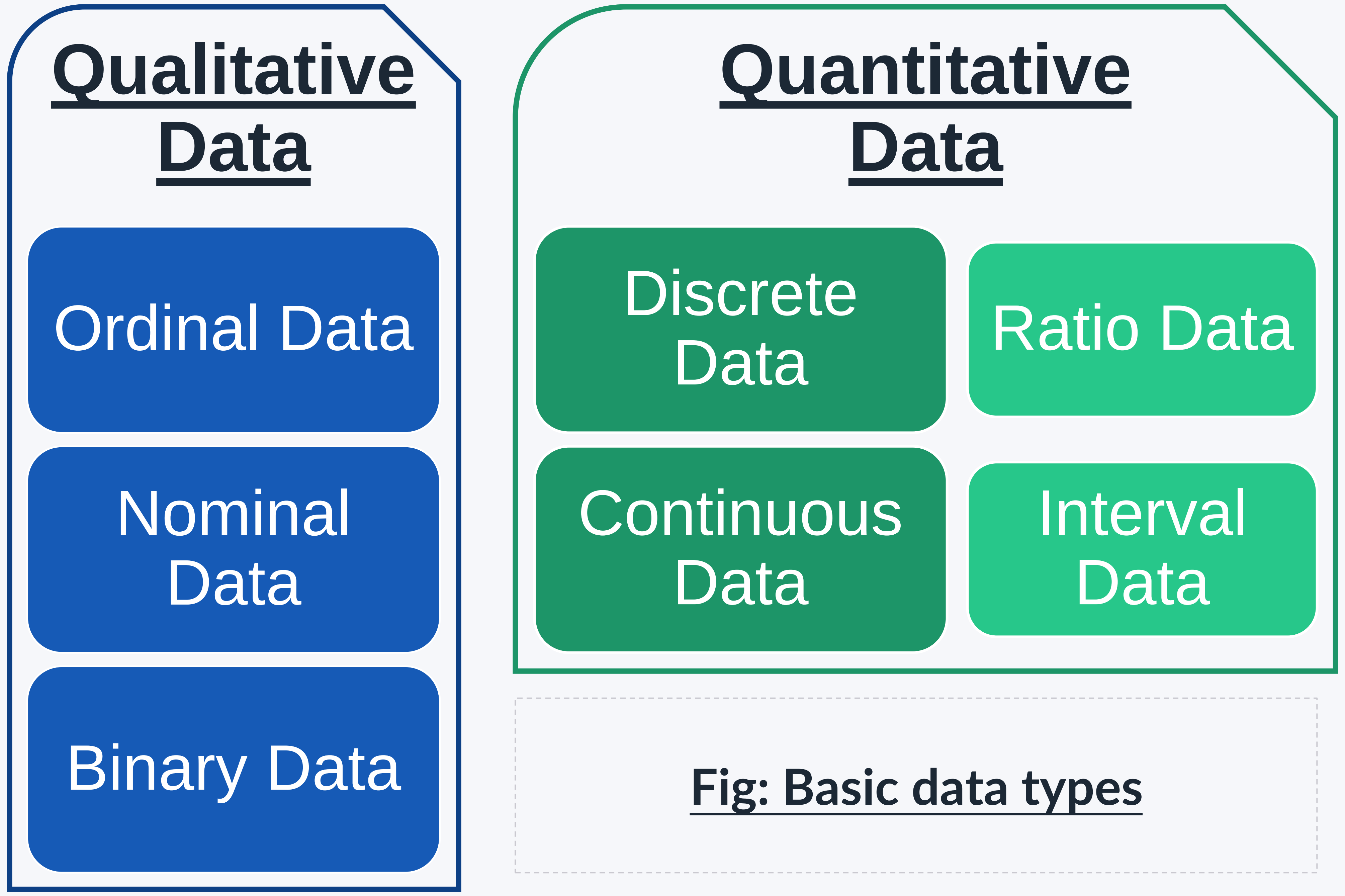
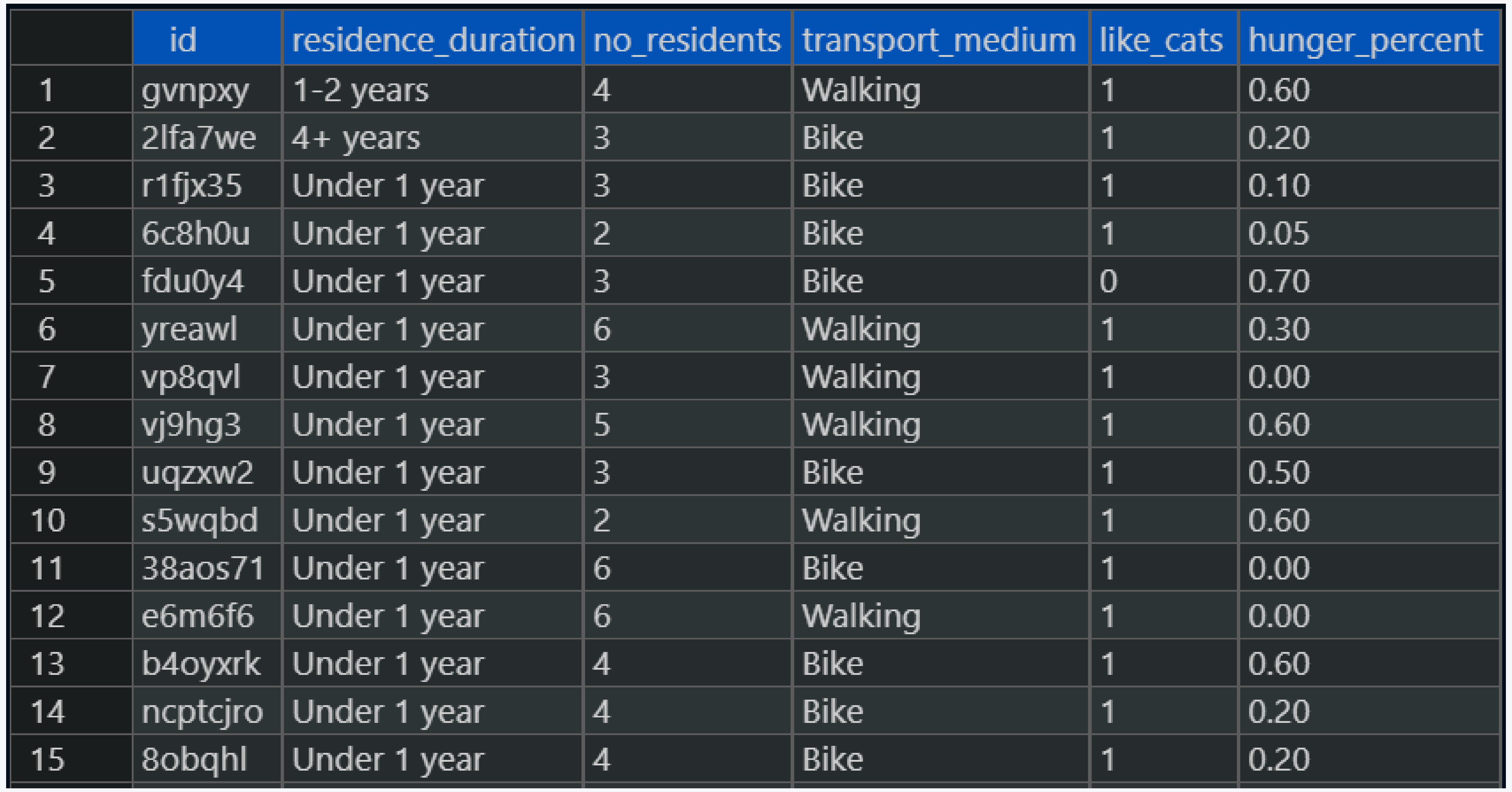


Fig: Basic data types



	id	residence_duration	no_residents	transport_medium	like_cats	hunger_percent
1	gvnpxy	1-2 years	4	Walking	1	0.60
2	2lfa7we	4+ years	3	Bike	1	0.20
3	r1fjx35	Under 1 year	3	Bike	1	0.10
4	6c8h0u	Under 1 year	2	Bike	1	0.05
5	fdu0y4	Under 1 year	3	Bike	0	0.70
6	yreawl	Under 1 year	6	Walking	1	0.30
7	vp8qvl	Under 1 year	3	Walking	1	0.00
8	vj9hg3	Under 1 year	5	Walking	1	0.60
9	uqzxw2	Under 1 year	3	Bike	1	0.50
10	s5wqbd	Under 1 year	2	Walking	1	0.60
11	38aos71	Under 1 year	6	Bike	1	0.00
12	e6m6f6	Under 1 year	6	Walking	1	0.00
13	b4oyxrk	Under 1 year	4	Bike	1	0.60
14	ncptcjro	Under 1 year	4	Bike	1	0.20
15	8obqhl	Under 1 year	4	Bike	1	0.20

* You can download the slides on MyStudy

Data Summarization

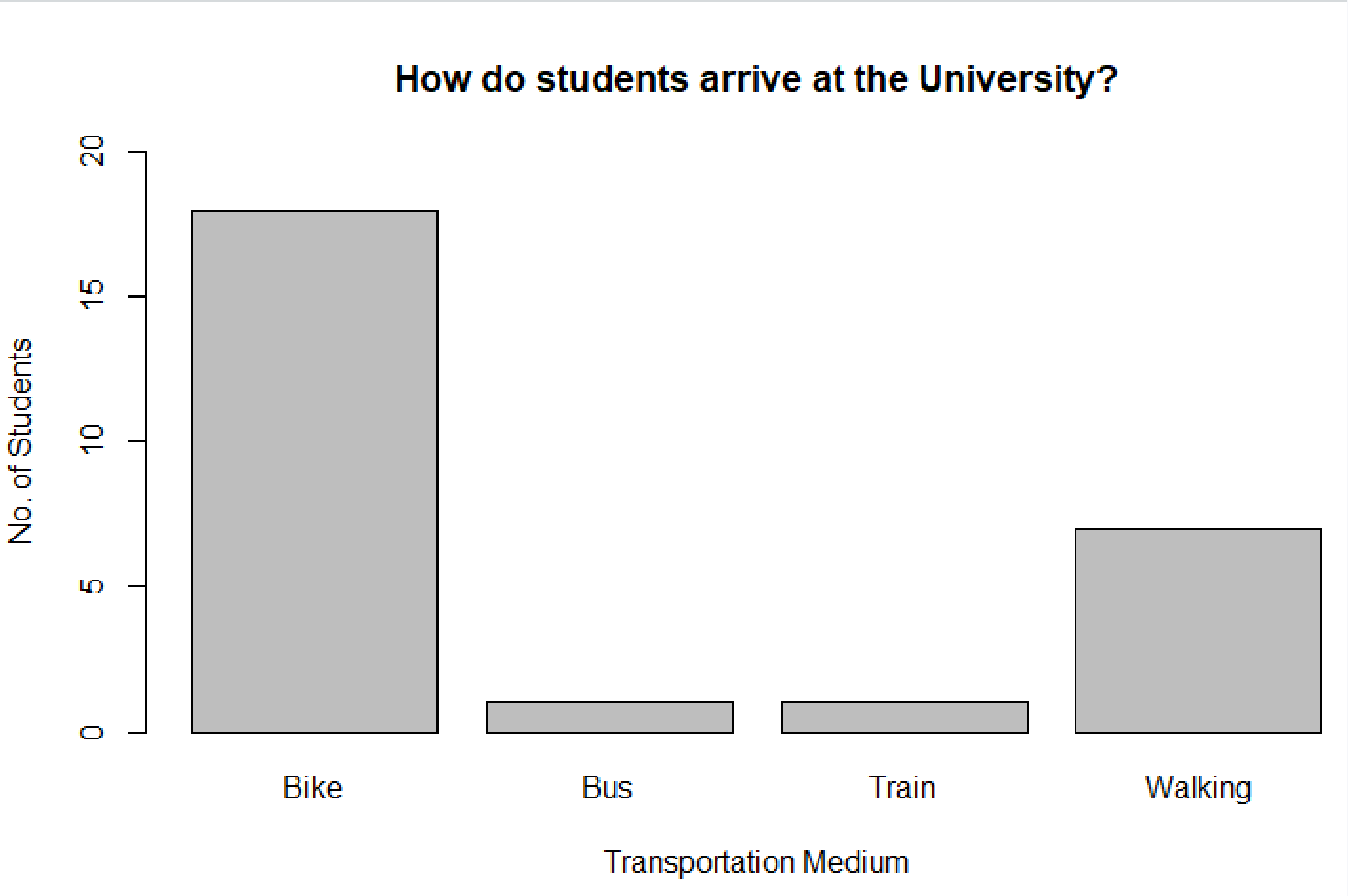
1. Introduction to data distribution
2. Measures of central tendency and dispersion
3. Box plots and Outliers

DISTRIBUTION OF THE DATA

1. **What?**
   - An arrangement of values of a variable showing their observed or theoretical frequency of occurrence
2. **Why?**
   - Shows how frequent each value is in a given data set
   - Enables us to get a better sense of the data than what just the numbers in the tables suggest
3. **How?**
   - *Discrete data*: bar chart
   - *Continuous data*: histogram

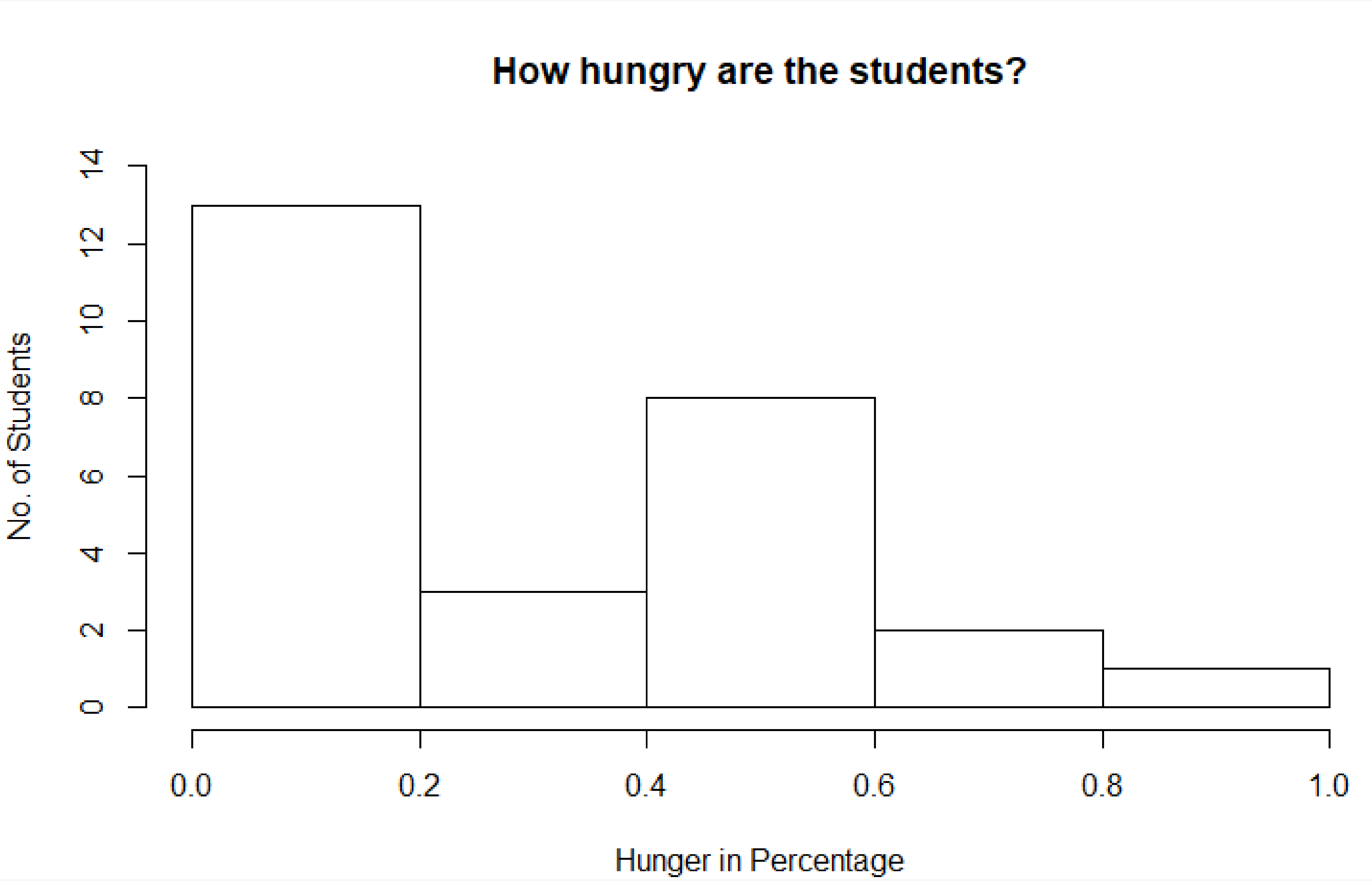
DISCRETE DATA: BAR PLOTS

- Takes only certain values (discrete values)
- Are represented by *bar charts*
  - There are gaps between the bars



CONTINUOUS DATA: HISTOGRAM

- Takes any value within some range
- Are represented by *histograms*
  - There are no gaps between the bars, and the distribution will look a little smoother for a larger N.



BASIC PROPERTIES OF DISTRIBUTION

- All statistical distributions have inherent properties, the most basic of which are:
  - Mean
  - Median
  - Mode
  - Variance
  - Standard deviation

Good news: most of these concepts are intuitive to understand

MEASURES OF CENTRAL TENDENCIES

- Central tendencies signify the “average” of the data
  - Mode, mean, and median
- **Mode** = the most frequent value in the data
- **Mean** = arithmetic average of a set of numeric values

$$\boldsymbol{mean = \bar{x} \; = \frac{\sum x}{N}}$$

$$where, x = each\ data\ point\ and$$

$$N = total\ number\ of\ data\ points$$

MEDIAN (CENTRAL TENDENCY)

- The value whose occurrence lies in the middle of a set of observations (divides the data into two “equal” parts)
- Steps:
  1. Arrange the data in an ascending order
  2. If N is odd:

$$\boldsymbol{median = \left(\frac{N+1}{2}\right)^{th} item}$$

  3. If N is even:
     - Identify the middle two numbers and take their average

$$\boldsymbol{median = \frac{\left(\frac{N}{2}\right)^{th} item + \left(\frac{N}{2}+1\right)^{th} item}{2}}$$

QUARTILE

- Quartiles divide the data into 4 “equal” parts
- Median is the second quartile
- 1st Quartile = Lower Quartile: $\boldsymbol{Q_1} = \left(\frac{\boldsymbol{N+1}}{\boldsymbol{4}}\right)^{th} \, term$
- 2nd Quartile = $\boldsymbol{Q_2} = median$
- 3rd Quartile = Upper Quartile: $\boldsymbol{Q_3} = \left(\frac{\boldsymbol{3(N+1)}}{\boldsymbol{4}}\right)^{th} \, term$

MEASURES OF DISPERSION: RANGE & IQR

- Dispersion = measure of how much the data varies from the mean; e.g. range, variance, standard deviation, interquartile range

  - **Range** = $largest\ value - smallest\ value = L - S$

  - **Interquartile range** = where the middle 50% of the data lies

$$\boldsymbol{IQR = Q_3 - Q_1}$$

MEASURES OF DISPERSION: VARIANCE

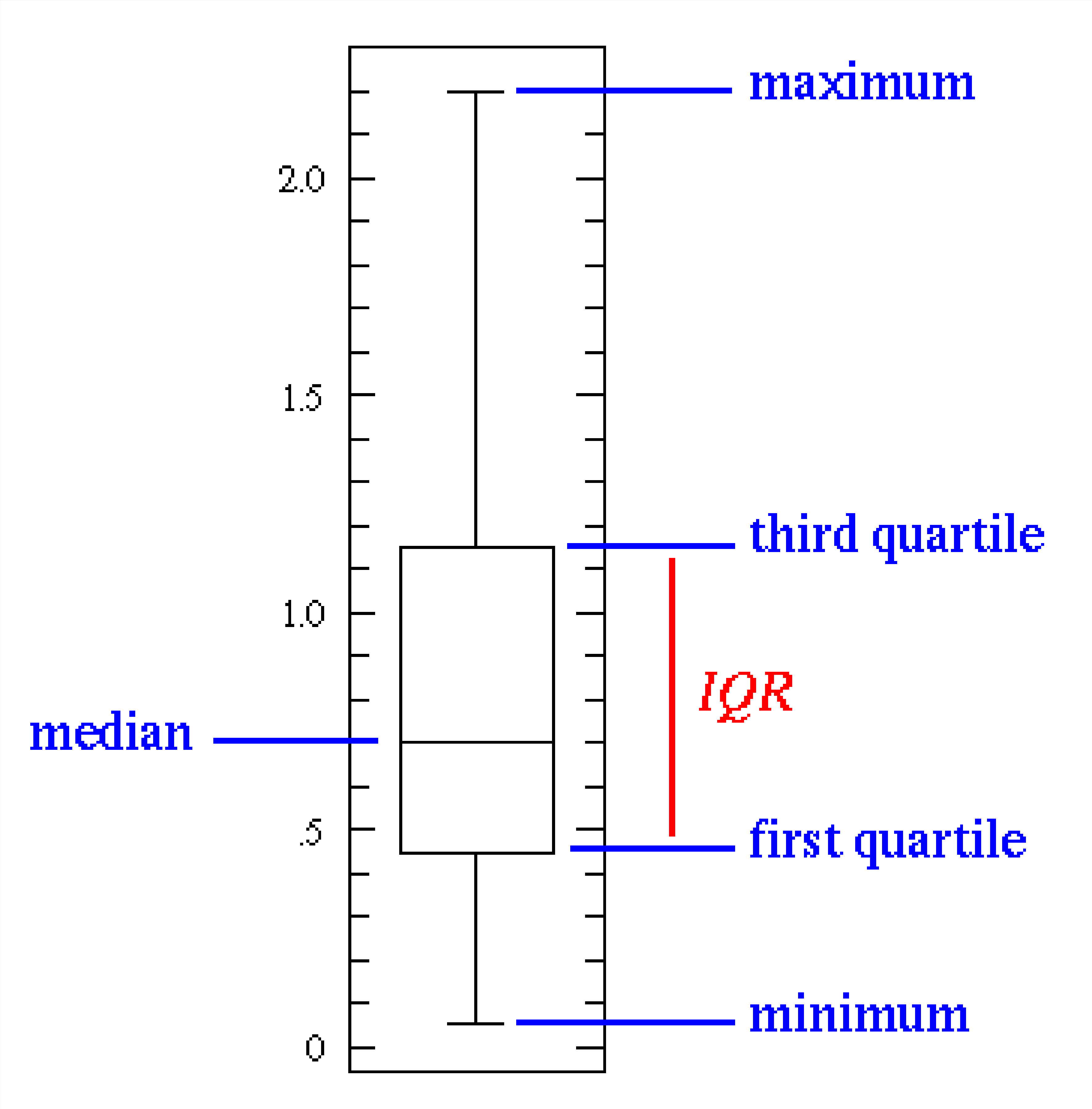
- **Variance** = a more robust, and widely accepted, measure of dispersion, and is defined as:

$$\boldsymbol{sample\ variance = s^2 = \frac{\sum(x_i - \ \overline{x})^2}{N-1}}$$

$$\boldsymbol{population\ variance = \sigma^2 = \frac{\sum(x_i - \ \mu)^2}{N}}$$

- **Standard deviation** (SD) = $\sqrt{variance} = \ \boldsymbol{\sigma}$ or $\boldsymbol{s}$
  - Measures the variability in the observations
  - Is easier to interpret because the values’ unit is in the scale of the data points

BOX PLOTS



- Summarize many measures of central tendencies and dispersion
- Learn more: http://www.physics.csbsju.edu/stats/box2.html

Exercise

1. Apply what we learned earlier to a small data.
2. Use R for simple data analysis.

A SMALL EXERCISE

Let’s take these numbers:

17, 12, 14, 7, 8,
19, 23, 19, 10, 7,
12, 7, 12

and calculate mean, mode, median, range, variance, S.D., quartiles…

PLAN FOR NEXT WEEK

That’s it for today! :-)

Next week, we are going to discuss:

1. Normal Distribution, Probability
2. Hypothesis Testing

If you want to reach me, mail me at:

`prabesh.dhakal@stud.leuphana.de`