DATA TYPES

Statistics Tutorial
Day 2

Prabesh Dhakal
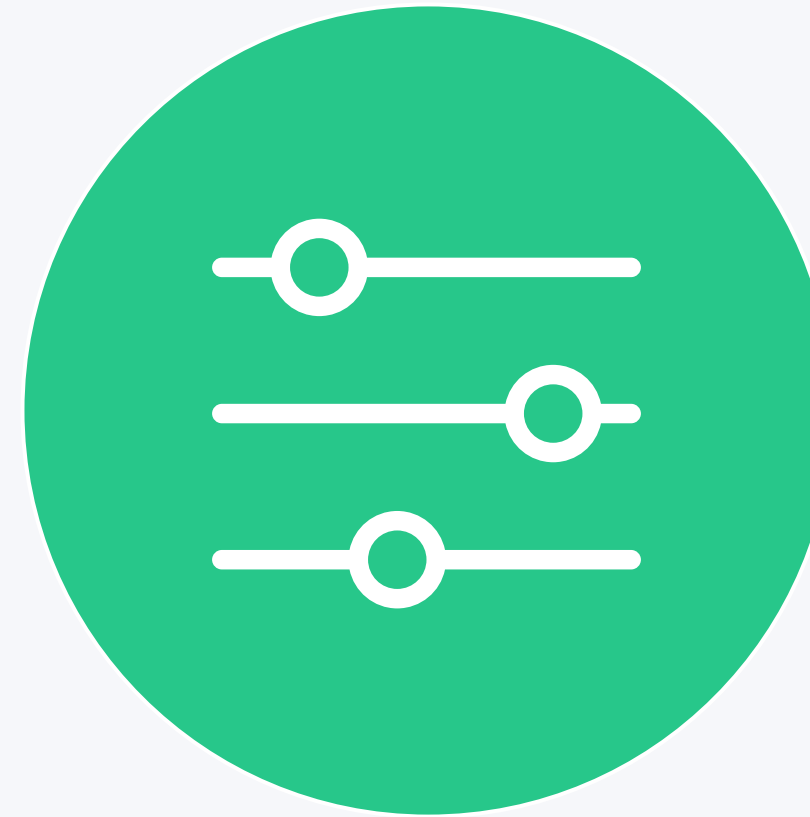2020 April 16

Recap

1. Statistics: what and why?

2. Refresher on RStudio

🛠️ CLASS EXERCISE - 1 🛠️

Poll on Zoom.

REVIEW FROM LAST SESSION

• • •

1. Statistics is the science of collecting, preparing, analyzing, and interpreting data.

2. Statistics allows us to:
   (i) make sense of the world around us,
   (ii) make decisions, and
   (iii) make predictions.

3. The tool we will use to work with the data and perform calculations is going to be R-Studio.
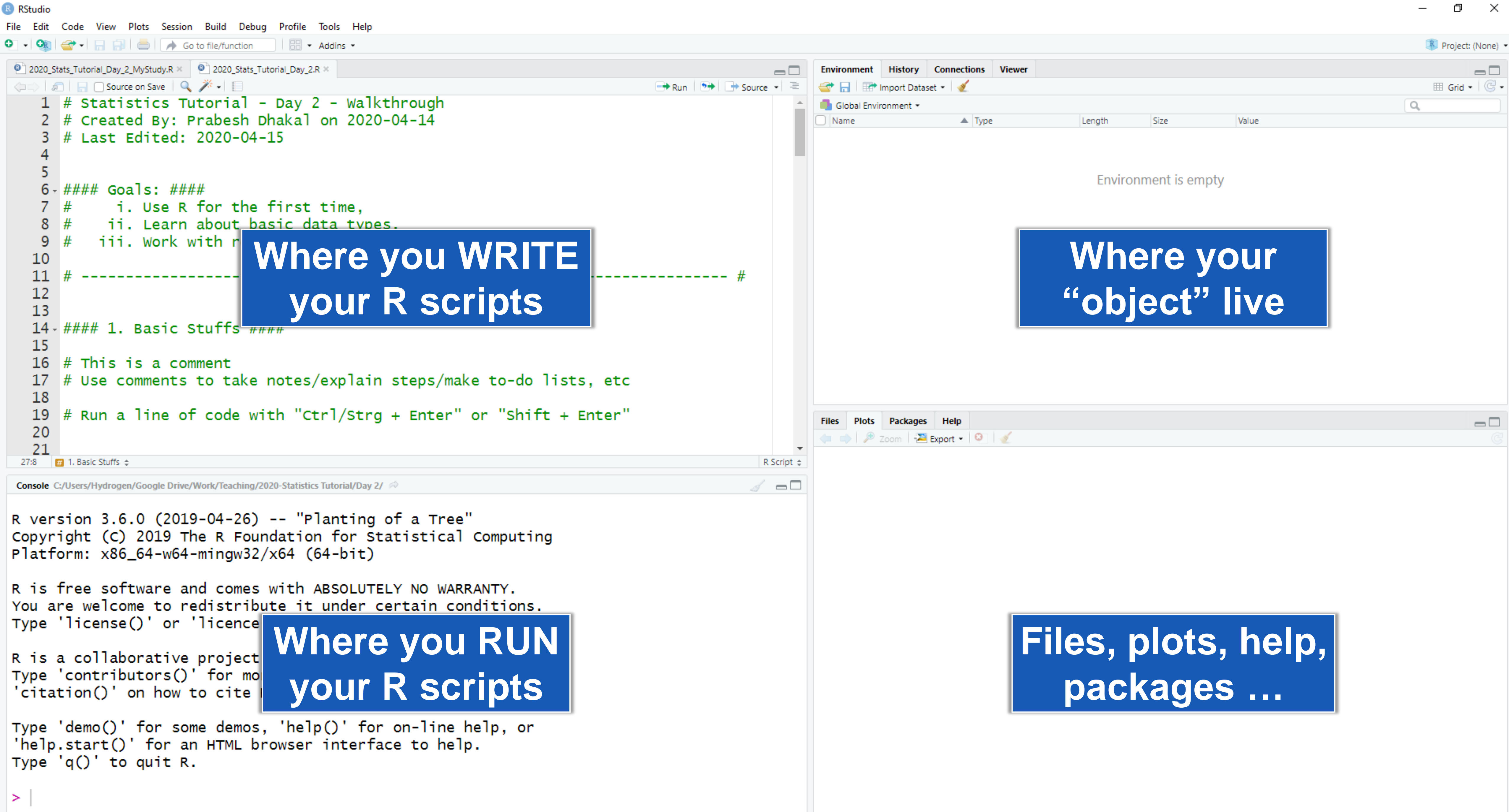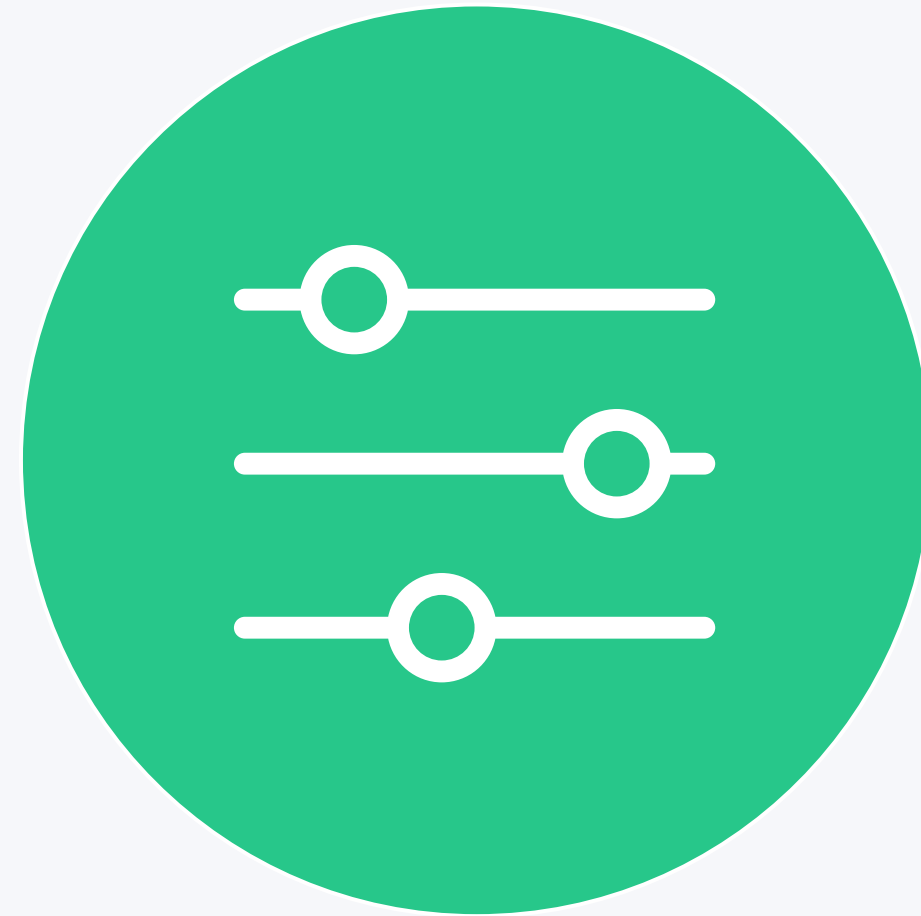
Data Types

1. Differentiate between the types of data

2. Determine how we can approach them

🛠️ CLASS EXERCISE - 2 🛠️

Please go to this link on your web browser and fill in the form:

www.t1p.de/do6x

BASIC TYPES OF DATA IN STATISTICS

Qualitative Data

Nominal Data

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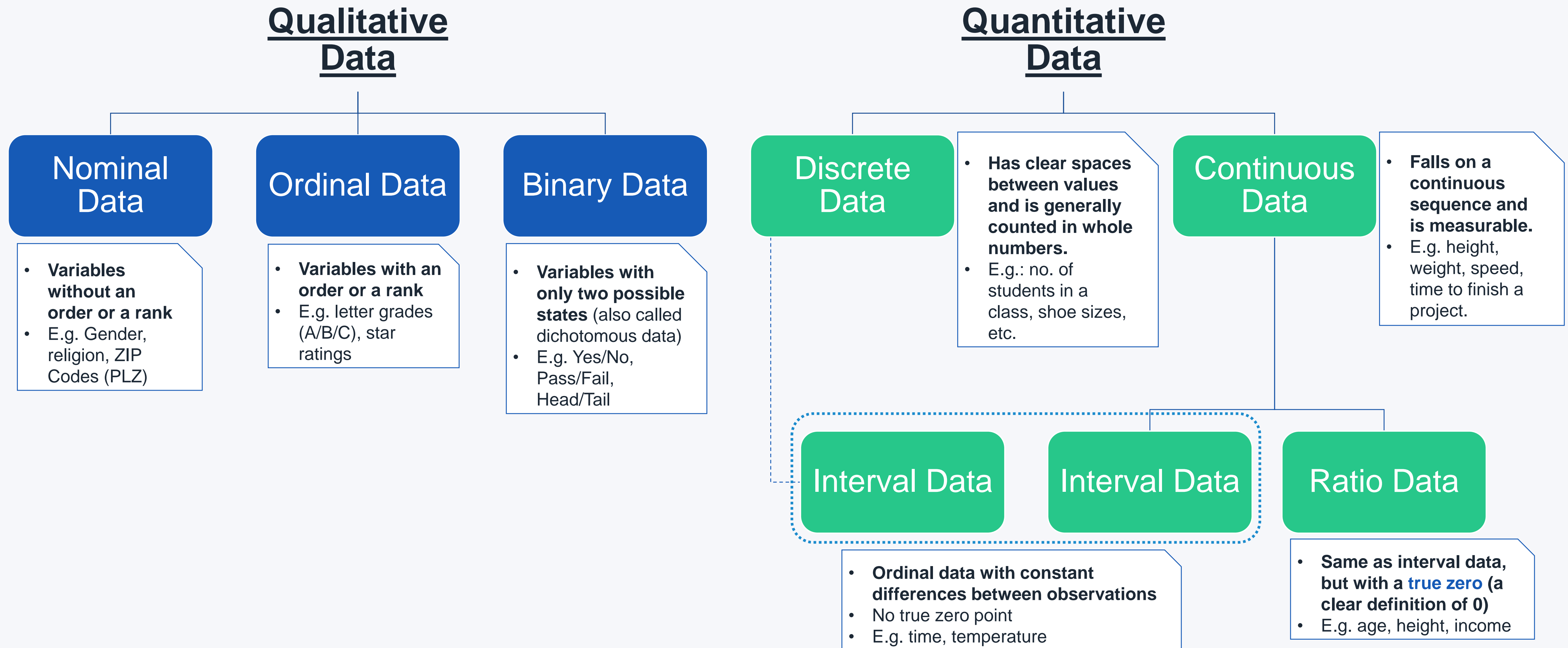- E.g. Gender, religion, ZIP Codes (PLZ)

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

WHAT DATA LOOKS LIKE

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
1	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
2	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
3	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
4	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
5	7.2	0.23	0.32	8.5	0.058	47	186	0.9956	3.19	0.4	9.9	6
6	8.1	0.28	0.4	6.9	0.05	30	97	0.9951	3.26	0.44	10.1	6
7	6.2	0.32	0.16	7	0.045	30	136	0.9949	3.18	0.47	9.6	6
8	7	0.27	0.36	20.7	0.045	45	170	1.001	3	0.45	8.8	6
9	6.3	0.3	0.34	1.6	0.049	14	132	0.994	3.3	0.49	9.5	6
10	8.1	0.22	0.43	1.5	0.044	28	129	0.9938	3.22	0.45	11	6
11	8.1	0.27	0.41	1.45	0.033	11	63	0.9908	2.99	0.56	12	5
12	8.6	0.23	0.4	4.2	0.035	17	109	0.9947	3.14	0.53	9.7	5
13	7.9	0.18	0.37	1.2	0.04	16	75	0.992	3.18	0.63	10.8	5
14	6.6	0.16	0.4	1.5	0.044	48	143	0.9912	3.54	0.52	12.4	7
15	8.3	0.42	0.62	19.25	0.04	41	172	1.0002	2.98	0.67	9.7	5
16	6.6	0.17	0.38	1.5	0.032	28	112	0.9914	3.25	0.55	11.4	7
17	6.3	0.48	0.04	1.1	0.046	30	99	0.9928	3.24	0.36	9.6	6
18	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
19	7.4	0.34	0.42	1.1	0.033	17	171	0.9917	3.12	0.53	11.3	6
20	6.5	0.31	0.14	7.5	0.044	34	133	0.9955	3.22	0.5	9.5	5
21	6.2	0.66	0.48	1.2	0.029	29	75	0.9892	3.33	0.39	12.8	8
22	6.4	0.31	0.38	2.9	0.038	19	102	0.9912	3.17	0.35	11	7
23	6.8	0.26	0.42	1.7	0.049	41	122	0.993	3.47	0.48	10.5	8
24	7.6	0.67	0.14	1.5	0.074	25	168	0.9937	3.05	0.51	9.3	5
25	6.6	0.27	0.41	1.3	0.052	16	142	0.9951	3.42	0.47	10	6

WHAT DATA LOOKS LIKE

● ● ●
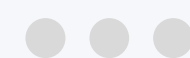
	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa
9	4.4	2.9	1.4	0.2	setosa
10	4.9	3.1	1.5	0.1	setosa
11	5.4	3.7	1.5	0.2	setosa
12	4.8	3.4	1.6	0.2	setosa
13	4.8	3.0	1.4	0.1	setosa
14	4.3	3.0	1.1	0.1	setosa
15	5.8	4.0	1.2	0.2	setosa

BASIC CALCULATIONS FOR DATA TYPES

	Nominal Data	Ordinal Data	Interval Data	Continuous Data
Frequency distribution	✔	✔	✔	✔
Median, quartiles, percentiles	✖	✔	✔	✔
Add or subtract	✖	✖	✔	✔
Mean, standard deviation	✖	✖	✔	✔
Ratio, coefficient of variation	✖	✖	✖	✔

BASIC STATISTICAL TESTS

●●●

		Outcome Variables		
	Binary	**Categorical**	**Ordinal**	**Continuous**
Binary	Chi-square, Fisher-Exact Tests	Chi-square Test	Wilcoxon	T-test/ Wilcoxon
Categorical	Chi-square, Fisher-Exact Tests	Chi-square Test	Kruskal-Wallis	ANOVA
Ordinal	Wilcoxon	Kruskal-Wallis	Spearman Correlation	ANOVA
Continuous	T-test/Wilcoxon/ Logistic Regression	ANOVA/ Class Prediction	Ordinal Regression	Correlation/ Linear Regression

Predictor Variables

R Exercises

1. Basic data types in R

2. Basic data structures in R

BASIC DATA TYPES IN R

● ● ●

	Example	R Script
Character (or string)	'apple'	a <- 'apple'
Numeric	1.618	b <- 1.618
Integer	3L, -5L	c <- 3L
Booleans	TRUE, FALSE	d <- TRUE
Missing Values	NA	e <- NA

🛠️ VECTOR OBJECTS IN R 🛠️

	Example	R Script
Vectors	**c(1, 3, 4)**	**eg_vec <- c(1, 3, 4)**

Run the following in R Studio:

1. `eg_vec <- c(1, 3, 4)`
2. `class(eg_vec)`
3. `typeof(eg_vec)`
4. `eg_vec * 3`
5. `eg_vec * eg_vec`
6. `View(eg_vec)`
7. `eg_vec_1 <- eg_vec * 3`
8. `View(eg_vec_1)`

GENERAL R TIPS

● ● ●

Here are some functions that you will find useful:

- **help(x), builtins(), ?x**

- **typeof(x), class(x)**

- **str(x), summary(x)**

- **getwd(), setwd()**

- **list.files()**

- **length(x)**

CHECKING DATA TYPES & STRUCTURES

• • •

1. Know the basic data types and data structures.

2. Check the structure of the data structure using: `str()`

3. Check the summary of the data structure using: `summary()`

4. Useful functions: `help()` , `class()` : when in doubt, use them both.

PLAN FOR NEXT WEEK

● ● ●

That's it for today! :-)

Next week, we are going to discuss:

1. Descriptive Statistics

2. Data Distribution

If you want to reach me, mail me at:
`prabesh.dhakal@stud.leuphana.de`