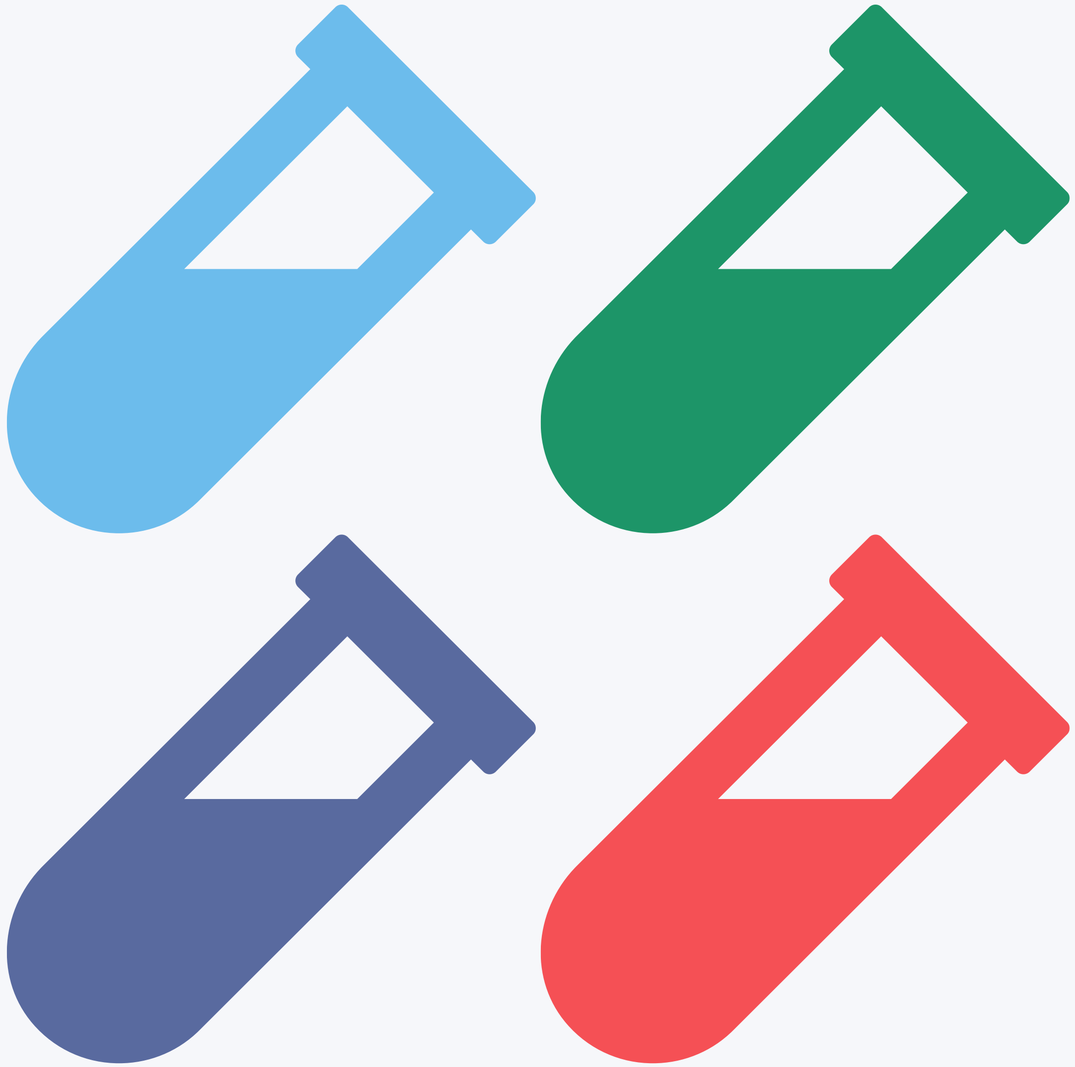


ANOVA

Statistics Tutorial

Day 9

Prabesh Dhakal

2020 June 11

WHAT ARE WE DOING TODAY?

RECAP + Q&A

We briefly revisit the contents from last week.

ANOVA

EXERCISE

We apply what we learned.

Q&A and Recap

Please ask if you have any questions now.

Otherwise, we can move on to the recap.

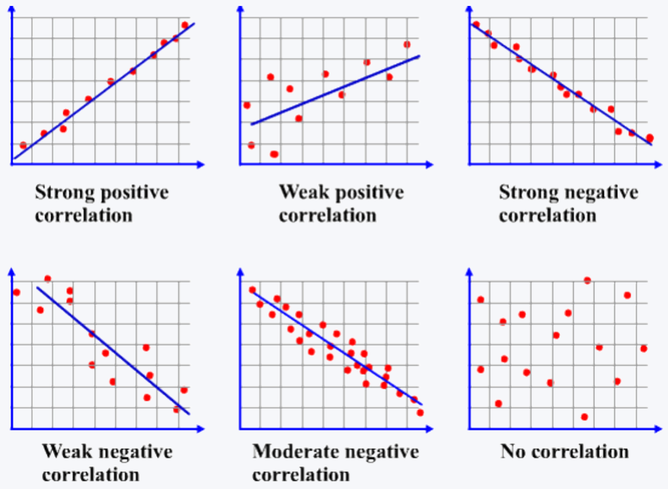
(PEARSON'S) CORRELATION

$$cor(x, y) = \frac{cov(x, y)}{\sigma_x * \sigma_y}$$

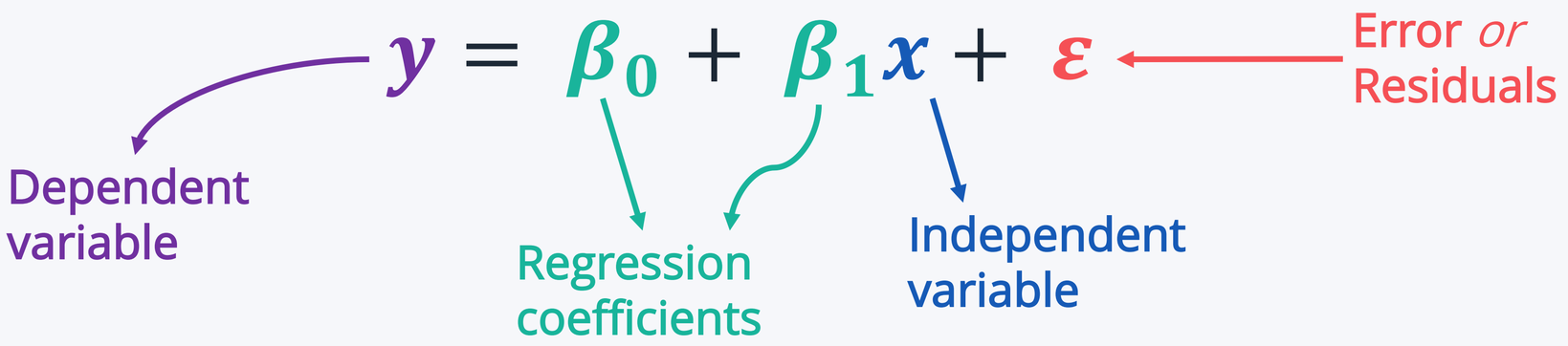
A measure of strength of linear relationship between two quant. Variables.

Value lies between [-1, +1].

What is high vs medium vs low correlation?



REGRESSION



Objective:

estimating the "**right**" regression coefficients

The model with smallest **error** is the best model

=

the straight line that best fits the data.

```
> summary(model_cars)

Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

IT'S BEEN A LONG JOURNEY

1. **Descriptive Data**
    - Summary statistics: central tendencies / dispersion / outliers …
    - Data visualization: bar plot / histogram / box plot / scatter plot …
2. **Data Distributions**: normal / uniform / poisson / …
3. **Hypothesis Testing**
    - Shapiro-Wilk's Test of Normality
    - Chi-square Tests: independence / goodness of fit
    - One Sample t-Test / Paired t-Test / Independent t-Test
    - F-Test for equality of variance
4. **Pearson's Correlation**
5. **Simple Linear Regression**

ANOVA

ANOVA

Applies to cases with more than 2 groups.

H_0: $Means\ of\ the\ groups\ are\ equal$

H_1: $Means\ of\ at\ least\ one\ group\ is\ not\ the\ same$

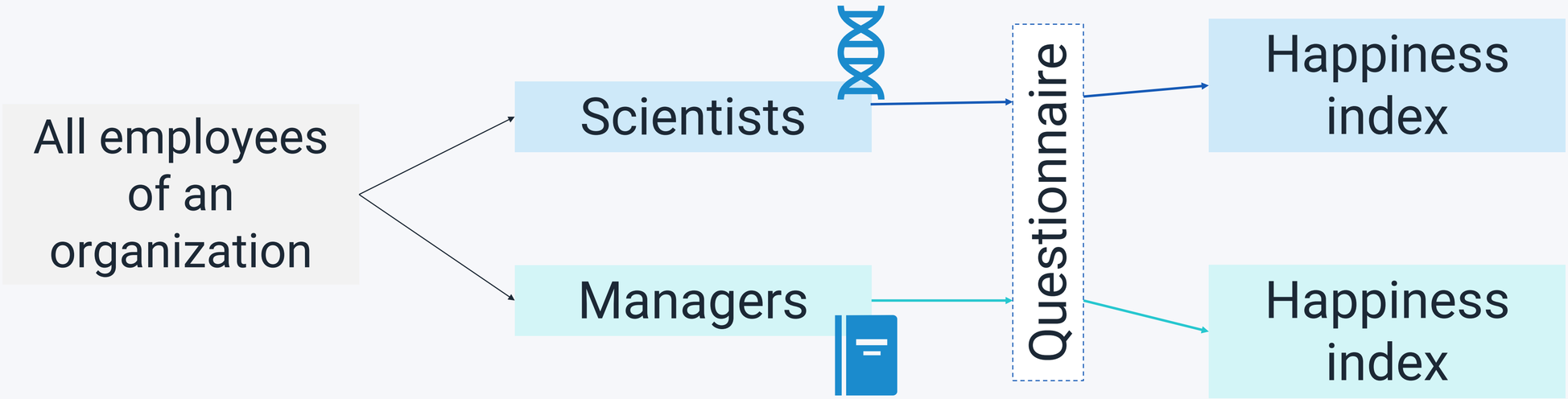
Consider two research questions:

Are scientists happier than managers?

Which of the 4 fertilizers work best?

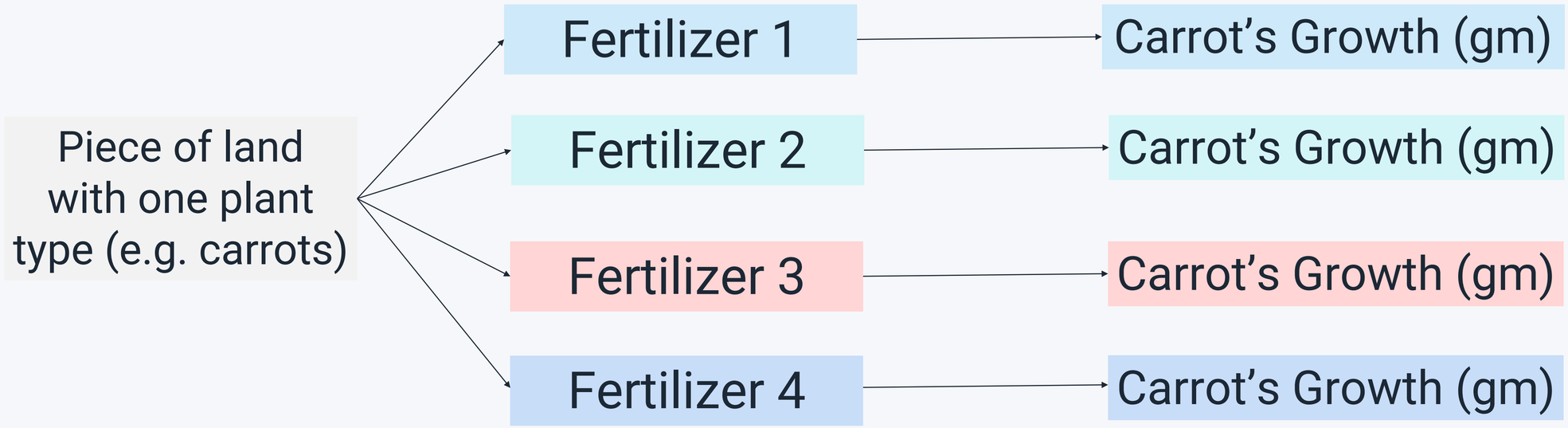
ARE SCIENTISTS HAPPIER THAN MANAGERS?

This is a case of **comparison of two groups → t-Test**



WHICH OF THE 4 FERTILIZERS WORK THE BEST?

This is a case of **comparison of more than 2 groups → ANOVA**



ONE-WAY VS TWO-WAY ANOVAS

One-way ANOVA:

- one factor investigated
- e.g. Does type of sand affect the yield of crop?

Two-way ANOVA:

- two factors investigated concurrently
- e.g. Does type of sand and type of fertilizer affect the yield?

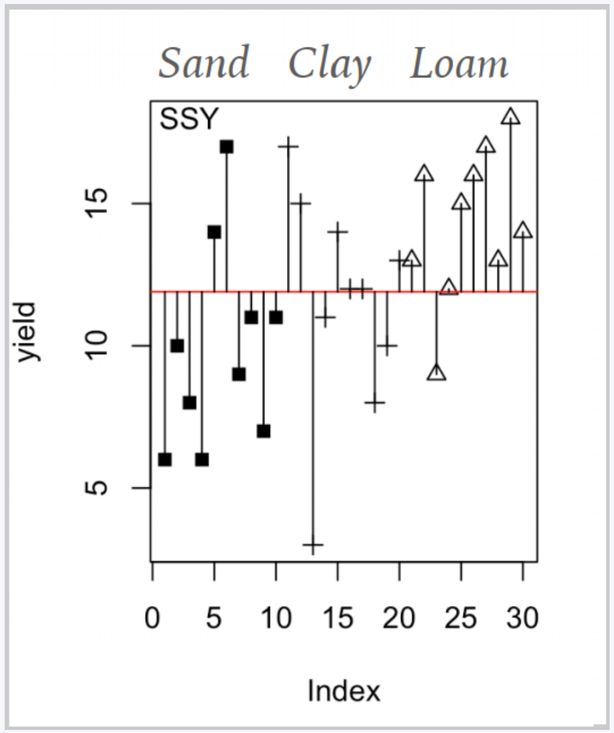
ONE-WAY VS TWO-WAY ANOVAS

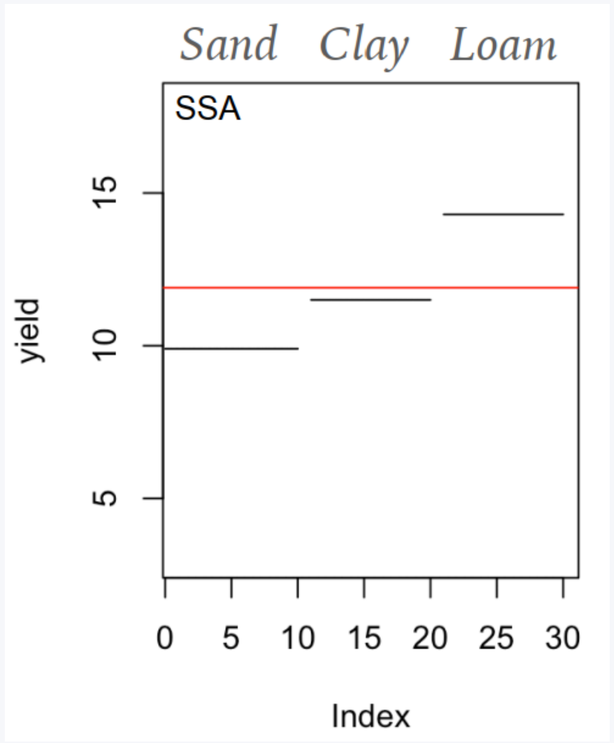
	One-way ANOVA	Two-way ANOVA
Basis of comparison	3 or more *levels* of *one factor*	Effect of *multiple levels* of *two factors*
Independent Vars	1	2
Sample size	Can be unequal in each group	Needs to be equal in each group

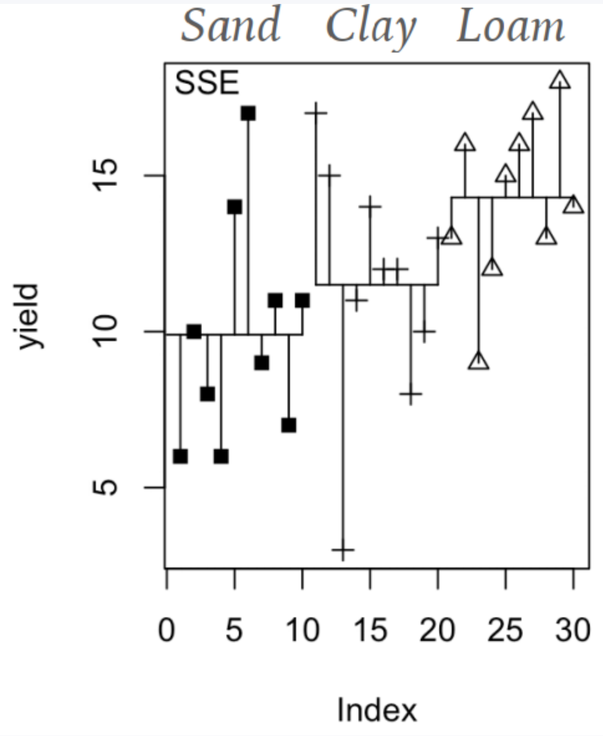
One-way ANOVA Intuition

Total Variation (SSY) = **Between-groups Variation (SSA)** + **Within-groups Variation (SSE)**

$data\ point - overall\ mean$ $treatment\ mean - overall\ mean$ $data\ points - treatment\ means$







ANOVA Test Statistic (F Ratio)

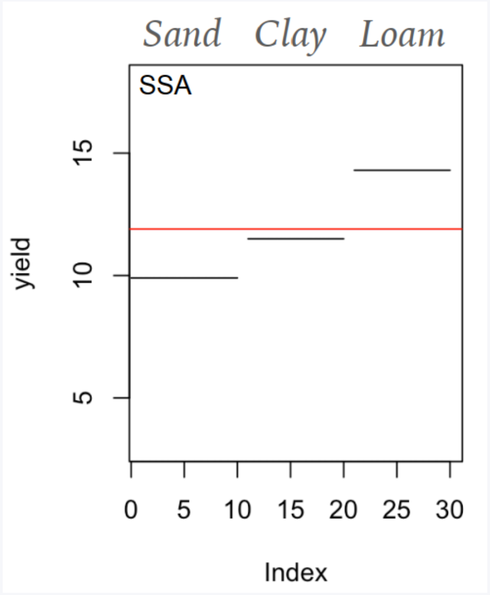
$$F - ratio = \frac{SS_{between\_groups}}{df_{treatment}} \div \frac{SS_{within\_groups}}{df_{residuals}}$$

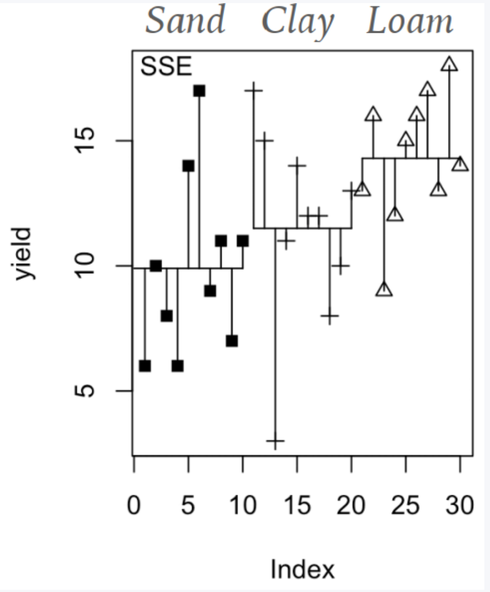
$k - 1$ $\quad$ $k * (n - 1)$

$(test\ statistic)$

$(treatment\ means - overall\ mean)$

$(data\ points - treatment\ means)$





ONE-WAY ANOVA RESULT FROM R

```
> model <- lm(yield ~ soil)
> anova(model)
Analysis of Variance Table
Response: yield
          Df  Sum Sq  Mean Sq F value  Pr(>F)
soil       2   99.2   49.600  4.2447 0.02495 *
Residuals 27  315.5   11.685  ---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1
```

$$F-ratio \quad = \quad \frac{SS_{between\_groups}}{df_{treatment}} \quad \div \quad \frac{SS_{within\_groups}}{df_{residuals}}$$

test statistic

treatment mean – overall mean **(Soil)**

data points – treatment means **(Residuals)**

TWO-WAY ANOVA RESULT FROM R

```
> anova(lm(mpg ~ cyl + am, data=mt_df))
Analysis of Variance Table

Response: mpg
           Df  Sum Sq Mean Sq F value    Pr(>F)
cyl         2 824.78  412.39 43.6566 2.477e-09 ***
am          1  36.77   36.77  3.8922   0.05846 .
Residuals  28 264.50    9.45
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

(Videos on ANOVA)

- One way: *https://www.youtube.com/watch?v=-yQb_ZJnFXw*
- Two way: *https://www.youtube.com/watch?v=cNIIn9bConY*

Exercise

Download the R file for Day 9 and open it on RStudio. ☺

PLAN FOR NEXT WEEK

That’s it for today! :-)

Tasks: `Freeform Exercise.zip` on MyStudy

Next week, we are going to discuss:

- End-to-end Data Analysis Workflow

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