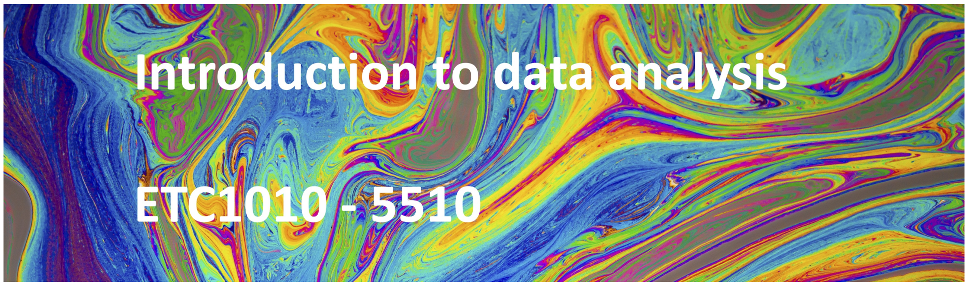

Introduction to Data Analysis

Week 2 - Tutorial

Contents

Workshop objectives

✓ Changing the form of a dataset using pivot_wider and pivot_longer
✓ Altering the dataset using filter and select
✓ Adding new variables using mutate
✓ Reporting using summarise, count and group_by
✓ Data cleaning and performing data analysis using tidyverse

Instructions

1. Follow this link here and log in into Rstudio Cloud using google and enter your monash user details (for help see the week 1 workshop 2 video in Moodle).
2. In each question you will replace '_____' with your answer, please note that the Rmd will not knit until you've answered all of the questions.

1 Exercise 1: Use wide, long format and separate

Resources

✓ koalabilby.rmd
✓ data/koala_bilby.csv

1. Open the file **koala-bilby.Rmd**
2. Read the data in and explore the variables
3. Pivot the data into long form, naming the two new variables, `label` and `count`
4. Separate the labels into two new variables, `animal`, `state`
5. Pivot the long form data into wide form, where the columns are `state`s.
6. Convert the long form data into wide form, where the columns are the `animal`s.

2 Exercise 2: Worked example of data cleaning

Resources

✓ pisa.Rmd
✓ data/pisa_au.rds
✓ data/Codebook_CMB.xlsx

1. Open the file **pisa.Rmd**, read and execute the preparation steps that were taken to clean the data
2. Explore how the STRATUM variable is processed to create three new variables: `state`, `schtype` and `yr`
   a. Take a sample of the stratum in the object `strat_slice` to see what the data originally look like
   b. Compare what the results of `strat_slice` are compare to using `str_sub` below
   c. Write what you think `str_sub` does, and what the `start` and `end` arguments represent
   d. Write how the STRATUM variable is processed to create the three new variables: `state`, `schtype` and `yr`?"
3. Explain what the `rename` operation is doing (around line 100)

4. Perform the following summaries:

    a. Compute the average of math scores by state.
    b. Which state does best, on average, on math?

5. Compute the difference in average male and female math scores by state by:

    a. Calculating the mean for math by state and gender
    b. Pivoting one of the columns using `pivot_wider`
    c. Now calcluate the difference between male and female scores and arrange
    d. Which state has the smallest average gender difference?

6. Does test anxiety have an effect on math score?

3 Exercise 3: Data exploration and manipulation

Resources

✓ frenchfries.rmd
✓ data/french_fries.csv

1. Open **french-fries.rmd** and load the tidyverse pacage by replacing the ’_____’ with the correct package

2. Read in the french fries csv data file, is `french_fries` in long or wide format, how can you tell?

3. The next code chunk converts the data to long form. What do you notice about each observation?

4. Filter french fries data using `filter()` to have:

    a. only week 1
    b. weeks 1-4 only
    c. oil type 1 (oil type is called treatment)
    d. oil types 1 and 3 but not 2

5. Show the following variables using `select()`:

    a. choose time, treatment and rep
    b. choose subject through to rating
    c. show everything except subject (drop subject)

6. For the french fries data compute a new variable called lrating by taking a log of the rating using `mutate`

7. Use `group_by()` and `summarise()` to do the following:

    a. Compute the average rating by subject
    b. Compute the average rancid rating per week

8. Use `count()` to count the following:

    a. the number of subjects
    b. the number of types

4 Exercise 4: Data analysis

Resources

✓ frenchfries.rmd
✓ data/french_fries.csv

1. Are the `rating`s for each `type` similar?
2. Are the replicates (`reps`) like each other?
   a. Pivot the fries_long dataset wide to show the `rating` for `rep` 1 and `rep` 2 in their own column
   b. Summarise your pivoted wide data to report the correlation between the ratings in rep 1 and rep 2
   c. Plot the ratings for rep 1 and 2 and add a title to the plot with a short summary of the data quality
3. Does `rep` data quality differ by `type`?
   a. Summarise your fries_spread to report the correlation between rep 1 and rep 2 for each `type`
   b. Plot the rating for each `rep` by `type`, what do you find?