Assignment 3 - Data wrangling

R workshop

Fall 2021

Guidelines

Use R markdown to complete your assignment. Please provide all the code to make your work 100% reproducible.

1. Data wrangling

Read the file "lotus.csv".

```
lotus <- read.csv("../data/lotus.csv")
head(lotus)
```

```
##            planta trat doy tallo_1  tallo_2  tallo_3 tallo_4 tallo_5 flores
## 1 Trébol frutilla    c 155   38.00 34.75000 31.50000      NA      NA      0
## 2 Trébol frutilla    c 155   35.75 34.00000 31.50000      NA      NA      0
## 3 Trébol frutilla    c 155   29.50 27.50000 26.00000      NA      NA      0
## 4 Trébol frutilla    c 155   24.50 22.83333 21.16667      NA      NA      0
## 5 Trébol frutilla    c 155   31.00 30.75000 27.75000      NA      NA      0
## 6 Trébol frutilla    c 155   37.00 35.25000 30.25000      NA      NA      0
##          hv hm   tallo       h_t        pa          c     pac         r ra
## 1 0.3682500  0 0.55485 0.6636929 0.9231000 0.07790000 1.00100 0.2131500  0
## 2 0.3029500  0 0.50410 0.6009720 0.8070500 0.13545000 0.94250 0.2245500  0
## 3 0.2808500  0 0.48945 0.5738073 0.7703000 0.09545000 0.86575 0.2147500  0
## 4 0.1630667  0 0.18470 0.8828731 0.3477667 0.03933333 0.38710 0.1516333  0
## 5 0.2701000  0 0.37530 0.7196909 0.6454000 0.07990000 0.72530 0.2276000  0
## 6 0.2532000  0 0.42385 0.5973811 0.6770500 0.04925000 0.72630 0.1143500  0
##        rsum
## 1 0.2131500
## 2 0.2245500
## 3 0.2147500
## 4 0.1516333
## 5 0.2276000
## 6 0.1143500
```

1.1.

Generate a new dataframe with the following columns: plant (not "planta"), trt, plant_id (factor with levels 1, 2, 3, 4 and 5), tallo_cm, tallo_g.

1.2.

Take the dataframe `lotus` again, make 5 relevant questions. Write the questions and design correct visualizations. Remember to name the axes and specify the units properly.

1.3

2. Reshaping

2.1.

Take the data frame from USDA and keep only the data from the counties from states where the historical mean yield of corn is 12 tn ha^{-1}. Create a dataframe `df_21` that has a column for year and one for each one of the selected states, and the data in each cell is the average corn yield of that year, in tons per hectare. See example:

```
##    year       IA       IN         KS
## 1  2011 17.00318 11.63601  3.0377514
## 2  2012 15.71114 16.66661 15.0912290
## 3  2013 11.21801 14.80106 14.1320054
## 4  2014 15.18659 10.13095 16.4479753
## 5  2015 20.65186 11.56645  6.7838426
## 6  2016 15.15499 13.88790  9.9941334
## 7  2017 17.28446 16.75124  0.2576570
## 8  2018 11.99512 15.60973 13.4047475
## 9  2019 15.24229 16.89521 11.3711676
## 10 2020 12.53335 13.73771  0.4715854
```

3. Think

You are taking measurements of a 2x3x6 factorial experiment. The variables you are measuring are incident and intercepted radiation (top and bottom of the canopy), leaf area of 5 plants (individually measured), stem diameter and plant height. You go to the field 5 times in the growing cycle. At the end, you also measure kernel number, kernel weight and yield. How would you organize your data?

Also, from class:

Read the file "Data file of Heat tolerance of chickpea genotypes in thermal zone of Ethiopia.xlsx" (also, check the metadata) and then:
a. Create the columns: * "FPI" that is the Flowering-Podding Interval (i.e. the difference (in days) between 50% flowering and 50% podding); "PD", Podding duration, as the difference (in days) between start and end of podding.
b. Read the file "chickpea_weather_madeup.csv" and pair each treatment with its data and create a new dataframe containing all the information, but only with the treatments that have crop data.
c. Create a new dataframe that is the "longer" version of that "wide" dataframe.
d. Visualize Yield versus PD, and visualize the average temperature july-september.

Extra and super fun: Pick 2 treatments of your choice and make a timeline with x = days, y = treatment, and in text show the event happening at which moment.

Discuss

You are working on a model to predict yield using soil, weather and crop data. You want to try the following variables to your model:
* Yield
* Density
* Genotype
* N fertilization
* P fertilization
* Planting date
* Soil water content
* Sand %
* Clay %
* pH

- Precipitation (induvidual for April, May, June, July, Sept and Oct)
- Temperature (induvidual for April, May, June, July, Sept and Oct)

How would you arrange your data frame?

Cheatsheets