ADS2 Practical 5: Getting and cleaning Data

Chaochen Wang

Semester 1, 2019/20

Work through this guide alone or in groups. Facilitators are here to help. The time it takes to complete this practical can vary between individuals - this is OK. Do not worry if you do not finish within the session.

Learning Objectives

- Use different data types and data structures in R
- Explain advantages of tidy datasets
- Clean a real-world dataset according to tidy principles

1. West Nile Virus (WNV) Mosquito Test

Background

List of locations and test results for pools of mosquitoes tested through the Chicago Department of Public Health Environmental Health program. The Chicago Department of Public Health maintains an environmental surveillance program for West Nile Virus (WNV). This program includes the collection of mosquitoes from traps located throughout the city; the identification and sorting of mosquitoes collected from these traps; and the testing of specific species of mosquitoes for WNV. Source: data.cityofchicago.org

The dataset is trimmed and monitored from the original dataset.

1) Input the "WNV_mosquito_test_results.csv" using *read.csv*, pay attention to the missing values in the original dataset, use the argument "na.strings" to convert them to NA's. Afterwards, please drop the incomplete records from the data frame. You may find *drop_na* command from *tidyr* package very useful. You can use *anyNA* to check if the records are successfully removed.
2) Check the data types (class) of the variables in the data frame. Do you need to convert any of them?
3) The name of the first variable "SEASON.YEAR" is not accurate, please change it to "YEAR".
4) Pay attention to the datatype of the variable TEST.DATE. There is specific datatype to manipulate datetime, please try *as.POSIXct* to convert them, be careful of the "format" and "tz" arguments (Tip: the time zone of Chicago is "America/Chicago". Check the change of the class afterwards. Assign the first datetime to "dat1", check the attributes of POSIXct datatype. Then alter the timezone attribute to "America/Los_Angeles", and see what happens to dat1.
5) The LOCATION variable consists of two elements LATITUDE and LONGITUDE, try command *gsub* to remove "(" and ")" first and then use command separate to generate two new variables based on LOCATION (Try settings in arguments "remove" and "convert" to see the differences of output).

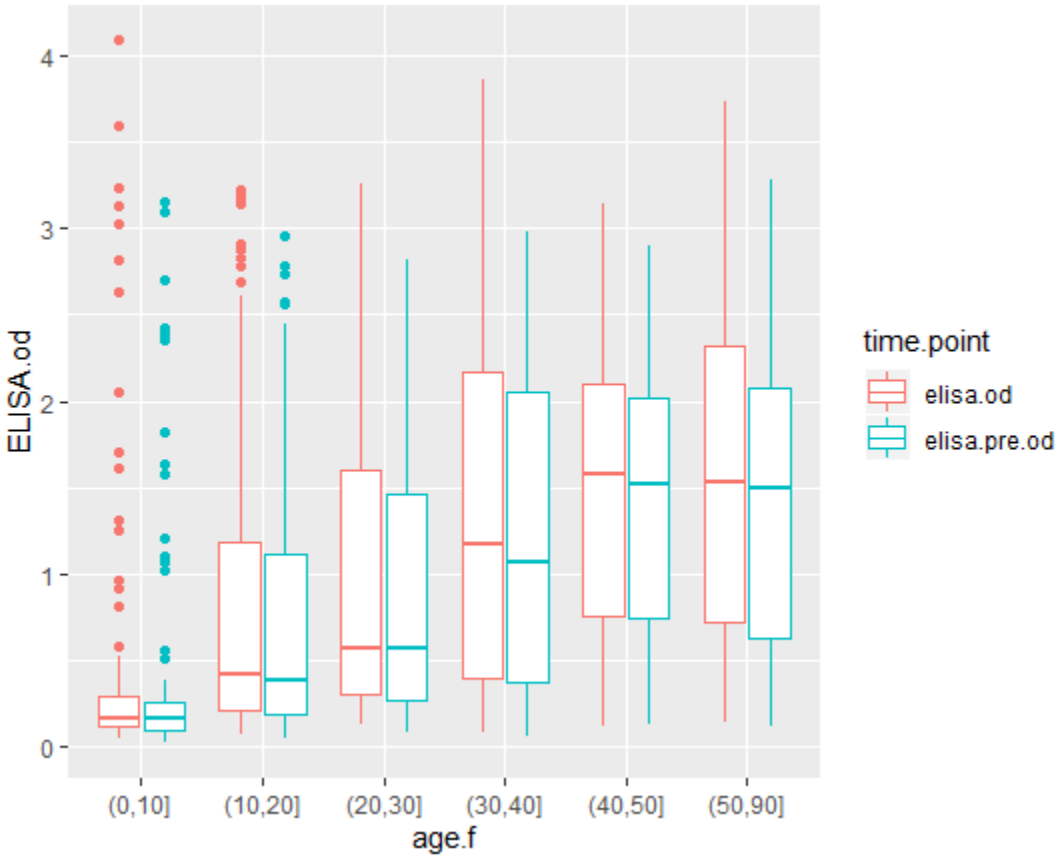
2. Tests for antibodies to trachoma PGP3 antigen

Background

This set includes data used in a latent class model to compare testing platforms for detection of antibodies against the Chlamydia trachomatis antigen Pgp3. Source: data.cdc.gov.

The dataset is trimmed from original dataset.

1) Input dataset from "Tests_PGP3.csv", pay attention that there are two types of missing values blank and “NA”. Convert variables to factors as needed.
2) The variable “sex” is not readable, convert to more readable format (“1” as “M”, “2” as “F”).
3) Plot boxplots to view the relationships of elisa.od and sex and age.f. What is wrong? Drop the incomplete records and plot again.
4) The variables “elisa.od” and “elisa.pre.od” are both measurements at different time points with ELISA. Try to use command *gather* in *tidyr* package to reshape the data frame so that the two measurements are combined into one variable “ELISA.od” (key=”time.point”, try the argument “factor_key”). Use ggplot to plot a boxplot to view the relationships between “ELISA.od” and “age.f”, use “color” argument to group “time.point”. You should see a plot like this.



5) Use command *spread* in *tidyr* package to separate the ELISA.od back to two variables based on time.point.