Course Final Assignment 'ToxCast'

Marc Teunis

2017-07-13

# Directions

## Package suggestion, you may need others!

library(tidyverse)

*Use ggplot2 syntaxis for all the figures in this assigment study*

*Use packages of the tidyverse, or any other package you think you need to solve the questions below*

*Use this Rmarkdown template to create a report adressing the assignment below*

*Hand in the assignment at the before deadline*

## Introduction

For the assessment of potential harmful exffects of chemicals, animal testing is required. Recently, a large study was conducted called "ToxCast".

## The data

The data can be found in the folder "./assignments/toxcast/"

Remember rprojroot?

## defines the root of the project for later use  
require("rprojroot") || utils::install.packages("rprojroot")

## [1] TRUE

library(rprojroot)  
root <- find\_root\_file(criterion = is\_rstudio\_project)

The file that you need to read into R is:

data\_file <- list.files(file.path(root, "assignments", "toxcast"), pattern = "^toxrefdb\*", full.names = TRUE)

Once loaded into R the data should have the dimensions 11815, 53, rows by columns.

For more details on the data see: <ftp://newftp.epa.gov/comptox/High_Throughput_Screening_Data/Animal_Tox_Data/> The data was downloaded from this website on 11 July 2017

## General TIPS

* Inspect the data
* Look at complete cases
* Remove NAs
* Tidy data if necessary
* Solve overplotting issues by using alpha and/or facet\_wrap
* The Rmd file you write must include all the steps of the analysis, including data load and graphics
* Provide proper graph titles and labesl for axes in all plots
* For this case you need to know about importing data in R. The file is csv format, remember to set the na.strings. In this dataset NA's are indicated by the term NULL Determine the delimiter used in the datafile.
* Remember the programming rules: use snake\_case for all variables in the data and in your script.
* Use the syntax from the {tidyverse} as much a possible
* You are allowed to use google/stack overflow, whatever you can find to solve the questions
* All graphics must be made with {ggplot2}
* Remember {dplyr} to select(), filter(), group\_by(), mutate() and summarize() data
* Check the type of each variable in the data. Do you need to change the variable type like data$var\_1 <- as.factor(data$var\_1)?
* All graphs need to have proper titles and axis labelling

## The case assignment

1. Create a github repository "toxcast\_assignment" in your account.
2. Read the data into R
3. Inspect the data
4. Describe missing values and the type of variables
5. Generate **at least 4** different graphs that describe the data
6. Describe all the steps you preformed to tidey and clean the data
7. Explain why you select certain variables or make other selections or changes
8. Answer the case questions below.
9. Upload a rendered (github\_document) and the associated \*.Rmd file (name the files both: "toxcast\_assignment") to your github account with the solutions to the assignment
10. Sent the teacher an email with the link to the files (the repository)

# Case questions

Try to answer the next 4 questions with a graph and narratives explaining what can be seen in the graph:

Write a general conclusion on the four graphs you create.

## Graph 1) Plotting two categorical variables, and prevent overplotting.

## 1A) How does the relationship between the chemicals ("chemical\_name") and the "effect\_category" variable look?

**TIPS**

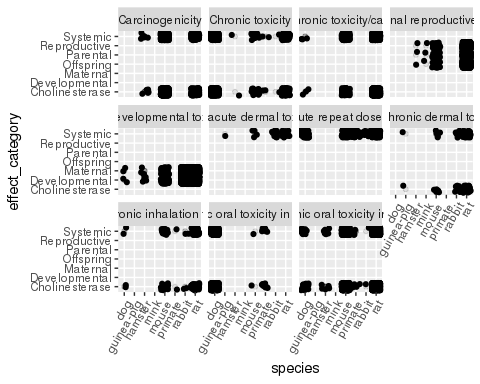
* This is a plot with two categorical variables.
* To prevent overplotting, do you need to use geom\_jitter(possition = "jitter")?
* Remember setting alpha()?
* Try using facets.

**ANSWER**

#?sample  
names(toxcast)

## [1] "X1" "chemical\_id"   
## [3] "chemical\_casrn" "chemical\_name"   
## [5] "chemical\_sets" "data\_source"   
## [7] "entry\_status\_id" "entry\_status"   
## [9] "entry\_level\_id" "entry\_level"   
## [11] "usability" "usability\_desc"   
## [13] "study\_id" "source\_study\_numeric\_id"   
## [15] "source\_study\_alphanumeric\_id" "year"   
## [17] "citation" "guideline\_no"   
## [19] "guideline\_name" "study\_type\_id"   
## [21] "species\_id" "strain"   
## [23] "comments\_animal" "admin\_method"   
## [25] "admin\_route" "dose\_start"   
## [27] "dose\_start\_unit" "dose\_end"   
## [29] "dose\_end\_unit" "lot\_batch"   
## [31] "purity" "source"   
## [33] "ldt" "hdt"   
## [35] "dose\_unit" "no\_doses\_tested"   
## [37] "study\_type" "species"   
## [39] "effect\_category" "study\_level\_lel\_dose\_level"   
## [41] "lel\_qualifier" "lel\_dose\_level"   
## [43] "lel\_dose" "nel\_qualifier"   
## [45] "nel\_dose\_level" "nel\_dose"   
## [47] "study\_level\_loael\_dose\_level" "loael\_qualifier"   
## [49] "loael\_dose\_level" "loael\_dose"   
## [51] "noael\_qualifier" "noael\_dose\_level"   
## [53] "noael\_dose"

ggplot(data = sample\_frac(toxcast, 1), aes(x = species, y = effect\_category)) +  
 geom\_point(alpha = 1/9) +  
 geom\_jitter(position = "jitter") +  
 facet\_wrap( ~ guideline\_name) +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1))



## 1B) Which species is most frequently represented in the data?

Write a number of code lines to determine the number of animals used for the species that is the most frequently represented in the data. Use the variable species to substantiate you claim.

**ANSWER**

str(toxcast$species)

## chr [1:11815] "dog" "dog" "rat" "rat" "rat" "dog" "dog" "mouse" ...

toxcast$species <- as.factor(toxcast$species)  
  
(species <- toxcast %>%  
 count(species))

## # A tibble: 8 × 2  
## species n  
## <fctr> <int>  
## 1 dog 1486  
## 2 guinea-pig 3  
## 3 hamster 37  
## 4 mink 3  
## 5 mouse 2102  
## 6 primate 77  
## 7 rabbit 1346  
## 8 rat 6761

## or  
pander::pander(table(toxcast$species))

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| dog | guinea-pig | hamster | mink | mouse | primate | rabbit | rat |
| 1486 | 3 | 37 | 3 | 2102 | 77 | 1346 | 6761 |

## Graph 2) A categorical and a continuous variable

**Use only the data for the dog in this graph, and only for the chemicals below:**

* "Sulfentrazone"
* "Cyproconazole"
* "Thiamethoxam"
* "Glufosinate-ammonium"
* "Flusilazoleand"

## 2A) Plot the relationships between the chemical\_name of the selected chemicals above and the noael\_dose variable for the **rats**?

**TIPS**

* Select the variables in the dataset that you need, generate a subset of the data in a new dataframe
* Filter the data for only the **rat**-data
* You need to group\_by the species and the chemical\_name
* You will need to summarize the noael\_dose
* Maybe you can use facets, for example to see information on the chemicals (variable guideline\_no?)
* Do you need to remove NAs? with na.omit()
* try geom\_jitter(position = "jitter")

**ANSWER**

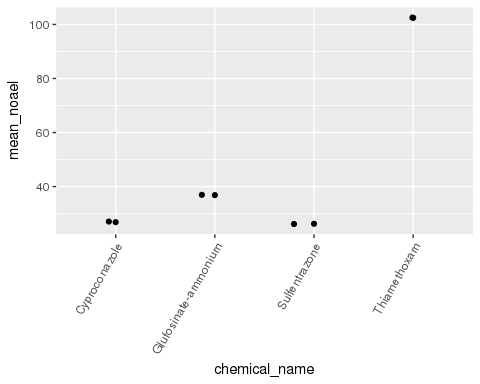
table\_chemicals <- as\_tibble(table(toxcast$chemical\_name))  
toxcast <- na.omit(toxcast)  
class(toxcast)

## [1] "tbl\_df" "tbl" "data.frame"

toxcast$chemical\_name <- as.factor(toxcast$chemical\_name)  
levels(toxcast$species)

## [1] "dog" "guinea-pig" "hamster" "mink" "mouse"   
## [6] "primate" "rabbit" "rat"

## subsetting the data  
toxcast\_selected <- toxcast %>%  
 dplyr::select(chemical\_name,   
 species,   
 noael\_dose) %>%  
 filter(species == "rat",   
 chemical\_name == "Sulfentrazone" |  
 chemical\_name == "Cyproconazole" |  
 chemical\_name == "Thiamethoxam" |  
 chemical\_name == "Glufosinate-ammonium" |  
 chemical\_name == "Flusilazoleand") %>%  
 group\_by(chemical\_name, species) %>%  
 summarise(mean\_noael = mean(noael\_dose))   
  
## graph  
toxcast\_selected %>%  
 ggplot(aes(x = chemical\_name, y = mean\_noael)) +  
 geom\_point() +  
 theme(axis.text.x = element\_text(angle = 60, hjust = 1)) +  
 geom\_jitter(position = "jitter") #+



# facet\_wrap( ~ guideline\_no)  
# ?geom\_jitter

## 2B) Toxicity?

Which of the chemicals under 2A is the most harmful. Remember that noael stand for "no-observed adverse effect level" which is the concentration for which no effect was observed.

**ANSWER**

## the chemicals with the lowest noael are the most harmful, Cypro and Sulfe are approx. equally harmful according the noael levels in the graph

## Graph 3) Two continuous variables - a panel with two plots

Look at the relationship between the loael\_dose and the noael\_dose. Plot a graph That shows the correlation between these variables for all the data in the ToxCast dataset. You will discover that there are two outliers for the Mouse. In a second graph, plot the same data but zoom in to the bulk of the data. Use x = 3000 (loael\_dose) as a cut-off for the x-axis and y = 1000 (noael\_dose) for the y-axis

**TIPS**

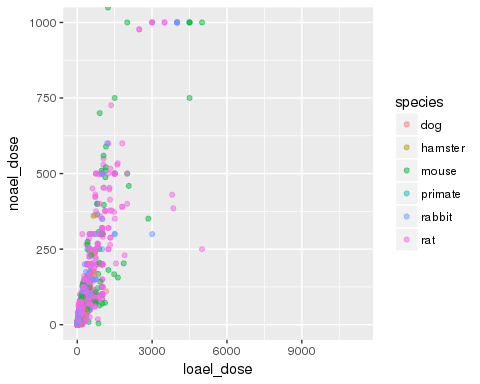
* To zoom in on data use the layer coord\_cartesian(ylim=c(0, 1000)) and coord\_cartesian(xlim=c(0, 3000)). Add these to you plot
* You can use cowplot::plot\_grid() to plot graphs in a panel
* You will get a bonus if you plot the names of the chemicals at the two outlier point in the first graph of the panel

**ANSWER**

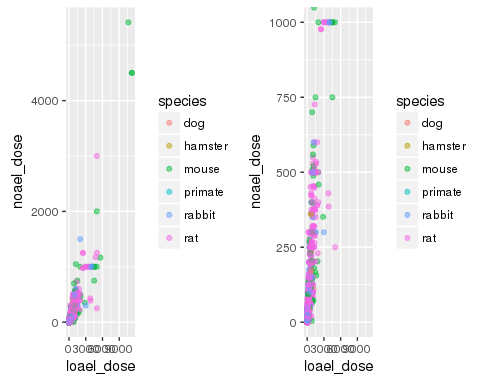
names(toxcast)

## [1] "X1" "chemical\_id"   
## [3] "chemical\_casrn" "chemical\_name"   
## [5] "chemical\_sets" "data\_source"   
## [7] "entry\_status\_id" "entry\_status"   
## [9] "entry\_level\_id" "entry\_level"   
## [11] "usability" "usability\_desc"   
## [13] "study\_id" "source\_study\_numeric\_id"   
## [15] "source\_study\_alphanumeric\_id" "year"   
## [17] "citation" "guideline\_no"   
## [19] "guideline\_name" "study\_type\_id"   
## [21] "species\_id" "strain"   
## [23] "comments\_animal" "admin\_method"   
## [25] "admin\_route" "dose\_start"   
## [27] "dose\_start\_unit" "dose\_end"   
## [29] "dose\_end\_unit" "lot\_batch"   
## [31] "purity" "source"   
## [33] "ldt" "hdt"   
## [35] "dose\_unit" "no\_doses\_tested"   
## [37] "study\_type" "species"   
## [39] "effect\_category" "study\_level\_lel\_dose\_level"   
## [41] "lel\_qualifier" "lel\_dose\_level"   
## [43] "lel\_dose" "nel\_qualifier"   
## [45] "nel\_dose\_level" "nel\_dose"   
## [47] "study\_level\_loael\_dose\_level" "loael\_qualifier"   
## [49] "loael\_dose\_level" "loael\_dose"   
## [51] "noael\_qualifier" "noael\_dose\_level"   
## [53] "noael\_dose"

plot\_a <- toxcast %>%   
ggplot(aes(x = loael\_dose, y = noael\_dose)) +  
 geom\_point(aes(color = species), alpha = 1/2)   
  
plot\_b <- toxcast %>%   
ggplot(aes(x = loael\_dose, y = noael\_dose)) +  
 geom\_point(aes(color = species), alpha = 1/2) +  
 coord\_cartesian(xlim = c(0, 3000)) +  
 coord\_cartesian(ylim = c(0, 1000))  
plot\_b



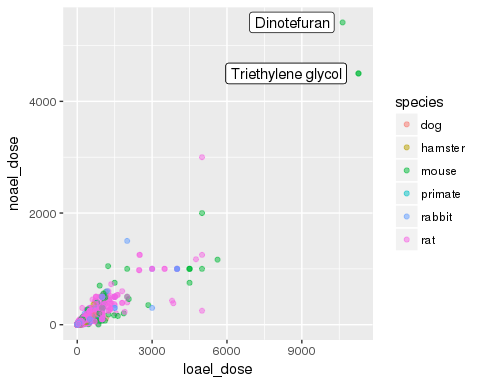
panel <- cowplot::plot\_grid(plot\_a, plot\_b)  
panel



# the outliers  
(outliers <- toxcast %>% select(chemical\_name, loael\_dose, noael\_dose) %>%  
 filter(noael\_dose > 3000, loael\_dose > 4000))

## # A tibble: 3 × 3  
## chemical\_name loael\_dose noael\_dose  
## <fctr> <dbl> <dbl>  
## 1 Triethylene glycol 11270 4500  
## 2 Triethylene glycol 11270 4500  
## 3 Dinotefuran 10635 5414

labels <- outliers$chemical\_name  
  
# add labels to plot  
plot\_a <- toxcast %>%   
ggplot(aes(x = loael\_dose, y = noael\_dose)) +  
 geom\_point(aes(color = species), alpha = 1/2) +  
 geom\_label(data = outliers, aes(x = loael\_dose, y = noael\_dose, label = labels), hjust = 1.1 )  
plot\_a



#?geom\_label

## Graph 4) Free choice

Generate a plot of your choosing. It can cantain any data from the ToxCast dataset. You may plot whatever you think is useful to study effects of the chemicals on the animals in the dataset. Be sure to tell a nice story about what can be discovered from your plot.

**ANSWER**