My portfolio

Claudia van der Zijden

2021-06-20

Contents

Chapter 1

Introduction

Welkom to my portfolio!

In my portfolio I try to give an impression of my programming skills. This is mainly in r but I also have some experience with bash.

This portfolio contains a number of chapters with assignments that I have made, it also contains my resume. The last chapter called machine learning is about a tutorial assignment in which I tried to learn more about machine learning.

I hope this portfolio will give you a good idea of my skills.

For further questions you can always email claudiavanderzijden@hotmail.nl

Chapter 2

Reproducible research

C. elegans plate experiment

The data for this exercise was kindly supplied by J. Louter (INT/ILC) and was derived from an experiment in which adult C.elegans nematodes were exposed to varying concentrations of different compounds. The variables RawData (the outcome - number of offspring counted as an integer value, after incubation time), compName (the generic name of the compound/chemical), the compConcentration (the concentration of the compound), and the expType are the most important variables in this dataset.

A typical analysis with this data would be to run a dose-response analysis using a log-logistic model with estimates for the maximal, the minimal, the IC50 concentration and the slope at IC50. We will not go into the details but a good package to run such computations and create graphs in R is the {drc} package. See: and:. In the exercise below we will create some visualizations using {ggplot2}.

Before we start, we will inspect the dataset. We do this by opening it in Excel. When you look at this dataset, a few things stand out. Among other things, there are many tabs with very large tables without an explanation. This makes it difficult for outsiders to use this data.

Then we will load the data into rstudio.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v ggplot2 3.3.3     v purrr   0.3.4
## v tibble  3.1.2     v dplyr   1.0.6
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
```

```
## -- Conflicts ---------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)
ce_liq_flow_062 <- read_excel("data/CE.LIQ.FLOW.062_Tidydata.xlsx", sheet = 1)
```

Now we can look at the data types. we will do this for the columns rawData, compName and compConcentration.

```
typeof(ce_liq_flow_062$RawData)
```

```
## [1] "double"
```

```
typeof(ce_liq_flow_062$compName)
```

```
## [1] "character"
```

```
typeof(ce_liq_flow_062$compConcentration)
```

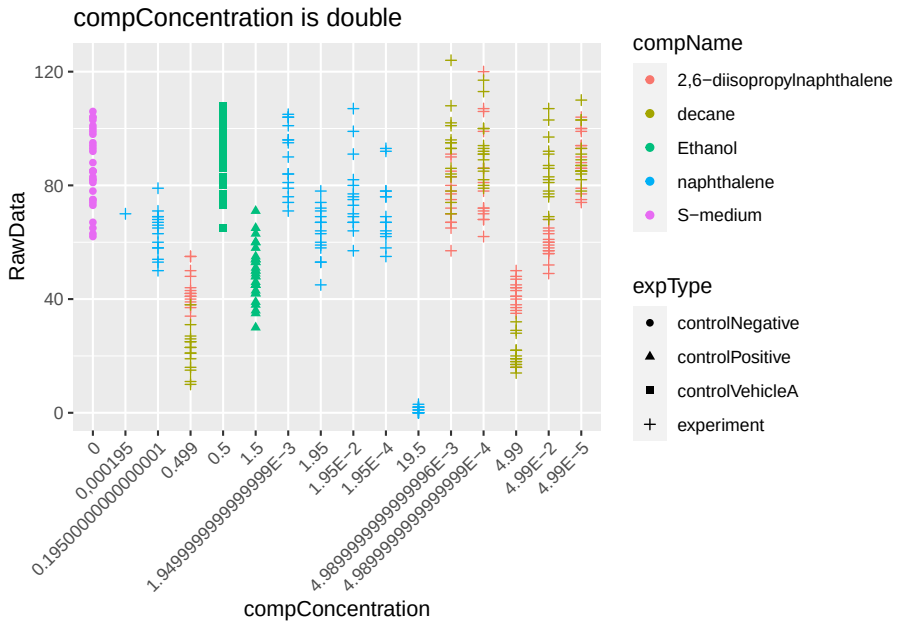
```
## [1] "character"
```

You would expect comConcentration to be numeric but as you can see this is character.

Now we are going to make a scatter plot of the data. We put compconcentration on the x-axis and DataRaw on the y-axis. We give a different color to the levels of compname and a different shape to the levels of expType. In addition, we ensure that the numbers below the x-axis are rotated 45 degrees so that we can read those.

```
ggplot(data = ce_liq_flow_062, aes(x = compConcentration, y = RawData)) +
  geom_point(aes(colour = compName, shape = expType)) +
   scale_x_discrete(guide = guide_axis(angle = 45)) +
  labs(title = "compConcentration is double")
```

```
## Warning: Removed 5 rows containing missing values (geom_point).
```



If we now look at this plot, you can see that the scale of the x-axis is not linearly distributed. This is probably due to the data type of comcondition. So we're going to change it to numeric. Then we will plot the data again. We now use a log10 transformation to improve the distribution of the x-axis. We also use jitter to avoid overlapping data points.

```
ce_liq_flow_062$compConcentration <- as.numeric(as.character(ce_liq_flow_062$compConcentration))
```

```
## Warning: NAs introduced by coercion
```

```
typeof(ce_liq_flow_062$compConcentration)
```

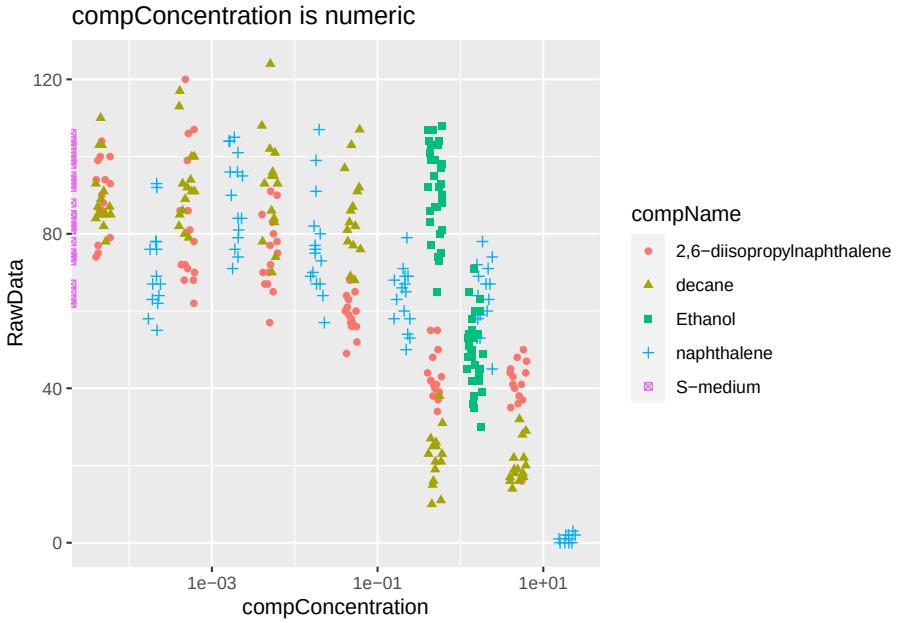
```
## [1] "double"
```

```
log10_scatter <-ggplot(data = ce_liq_flow_062, aes(x = compConcentration, y = RawData)) +
  geom_point(position=position_jitter(width=.1,height=0),aes(colour = compName, shape = compName)
   scale_x_discrete(guide = guide_axis(angle = 45))+
   labs(title = "compConcentration is numeric")


log10_scatter + scale_x_log10()
```

```
## Scale for 'x' is already present. Adding another scale for 'x', which will
## replace the existing scale.
```

```
## Warning: Transformation introduced infinite values in continuous x-axis
```

```
## Warning: Removed 6 rows containing missing values (geom_point).
```



The positive control for this experiments is **naphthale**. The negative control for this experiment is **S-medium**.

After reviewing the data, we could proceed with the analysis of the data to find out whether there is indeed an effect of different concentrations on offspring count and whether the different compounds have a different curve. To find out, first check whether the data is normally distributed. This can be done with the shapio-wilk test. This can be used to determine whether a parametric or non-parametric test can be used to see if there is a statistically significant difference between the different groups.

Finaly we normalize the data for the controlNegative in such a way that the mean value for controlNegative is exactly equal to 1 and all other values are expressed as a fraction thereof. Than we rerun the graph with the normalized data.

#“{r 1.1J} normalize <- function(x) { return ((x - min(x)) / (max(x) - min(x))) }

ce_liq_flow_062*compVehicle*compVehicle == “controlNegative”))

H.Why would you want to take the step under J?

<!--chapter:end:01-Reproducible_Research.Rmd-->

Open Peer Review

In this assisgment w arw goning to find a scientific article ourself, using PubMed or another dat

This is the link to the article I use
https://www.biorxiv.org/content/10.1101/2020.10.02.322917v2.full

The title of this article is: Leveraging high-throughput screening data and conditional generativ
The authors of this article are: Adrian J. Green, Martin J. Mohlenkamp, Jhuma Das, et al.

Peer revieuw part 1

__Study Purpose__ : the summary briefly explains what is more important to conduct this research
__Data Availability Statement__: not present
__Data Location__: it does describe what the data should look like and there are references to ar
__Study Location__: there is no information about where the study was conducted in the material a
__Author Review__: the details of the authors are not easy to obtain, the names of the authors ar
__Ethics Statement__: the introduction briefly mentions ethics
__Funding Statement__: nothing is said about funding
__Code Availability__: no code is shared in the article

Open peer review part 2
Next we are going to try to find a article with R code. We do this on the OSF website.
We are going to try to get the code working in our r studio.

This is the link to the code we will use
https://osf.io/gkcn7/

To make this code work we only have to change the way to load the data, out commend the effect si

You can find the working script in the appendix, chapther 11

It took little effort to get this script working. On a scale of 0 to 5 I would give it a 4

<!--chapter:end:02-Open_peer_review.Rmd-->

Guerrilla analytics

In this assignment I cleaned up my projects according to the Guerrilla analytics.The result can b

Daur2 project

```
![Claudia](data/Gurilla/Daur2.png){ width=70%}


## Portfolio project
![Claudia](data/Gurilla/portfolio.png){ width=70%}


## Project project
![Claudia](data/Gurilla/project.png){ width=70%}




<!--chapter:end:03-Guerrilla_analytics.Rmd-->

# Curriculum vitae

![Claudia](data/CV/cvfoto.png){ width=100%}

<!--chapter:end:04-Curriculum_vitae.Rmd-->

# Mendaly

In practice, many use has been made of RNA sequencing (RNA-seq) methods. With RNA-seq,
In our project we also want to pay attention to a new shiny app. Although the ISEE app
Ultimately, it would be nice if you only had to fill in a dataset and you would then re

<!--chapter:end:05-Mendely.Rmd-->

# Relational databases

TIPS

Be aware, the flu and dengue data contains metadata that should be stripped from the da
Think of a way to create valid country names that fit with the gapminder data.
Remember (!) that in the end, this assignment needs to be reported by a .Rmd file for y
Assignment

Load the flu ("./data/flu_data.csv), the dengue (."/data/dengue_data.csv) and the gapmi

Check if they are in the right shape. Is the data in the 'tidy' format? If not change t

Change the country and date variables of the three tables so that they coincide in terr

Store the three tables as separate (so six in total) .csv and .rds files.
```

In Dbeaver create a new PostgreSQL database "workflowsdb"

Using RPostgreSQL, insert the tables into the database.

Inspect the contents of the tables with SQL (in DBeaver) and save the SQL script.

Inspect the contents of the tables with dplyr (in R) and save a RMarkdown showing what you are do

Load the gapminder data in R and change the dataframe in such as way that you could join it to de

Save this clean gapminder data in the "workflowsdb" database

Perform some joins (your choice) with SQL (can be done in DBeaver or with dplyr.

Generate a joined table, and export this from the database to R.

Show some descriptive statistics with this table, and at least 3 visualisations using ggplot2.

show all of your actions in this assignment in a Rmd file, perhaps with pictures and provide text

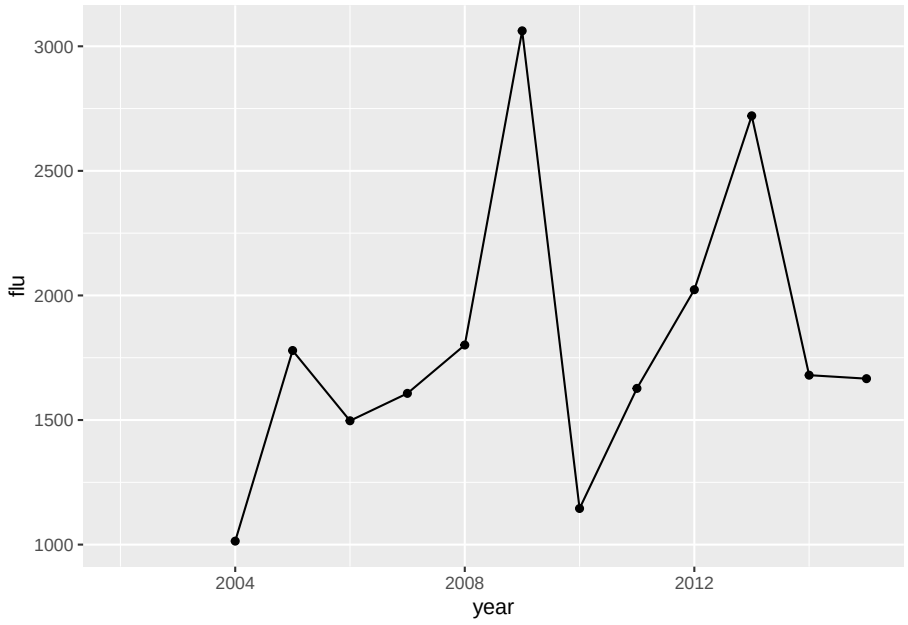
```r
library(tidyverse)
library(dslabs)
gapminder <- as_tibble(gapminder)
flu_data<- read.csv(url("https://raw.githubusercontent.com/ClaudiavdZ/tlsc-dsfb26v-20_workflows/m
flu_data <- as_tibble(flu_data)
dengue_data<- read.csv(url("https://raw.githubusercontent.com/ClaudiavdZ/tlsc-dsfb26v-20_workflow

write.table(dengue_data , file = "dengu_data.csv")
write.table(dengue_data , file = "dengu_data.RDS")
write.table(flu_data , file = "flu_data.csv")
write.table(flu_data , file = "flu_data.RDS")
write.table(gapminder , file = "gapminder.csv")
write.table(gapminder , file = "gapminder.RDS")

library(DBI)
con <- dbConnect(RPostgres::Postgres(),
                 dbname = "myfirstdb",
                 host="localhost",
                 port="5432",
                 user="postgres",
                 password="Veroni36")
dbListTables(con)

## [1] "test"        "gapminder"   "flu_data"    "dengue_data"
```

```
#dbWriteTable(con, "dengue_data", dengue_data)
#dbWriteTable(con, "flu_data", flu_data)
#dbWriteTable(con, "gapminder", gapminder)

# library(janitor)
# gapminder_usd <- as.data.frame(t(gapminder))
# gapminder_usd <- gapminder_usd %>% row_to_names(row_number = 1)


flu_usd <- gather(
  flu_data,
  key = "country",
  value = "flu",
  Argentina:Uruguay
)
#seperate year from month and day
flu_usd <- separate(flu_usd, Date, into = c("year", "month", "day"),  sep = "-")
#count sum of flu
flu_usd <- aggregate(flu_usd$flu, by=list(year=flu_usd$year, country=flu_usd$country),
flu_usd <- flu_usd %>% rename(flu = x)
flu_usd$year <- as.integer(flu_usd$year)

dengue_usd <- gather(
  dengue_data,
  key = "country",
  value = "dengue",
  Argentina:Venezuela
)
dengue_usd <- separate(dengue_usd, Date, into = c("year", "month", "day"),  sep = "-")
dengue_usd <- aggregate(dengue_usd$dengue, by=list(year=dengue_usd$year, country=dengue
dengue_usd <- dengue_usd %>% rename(dengue = x)
dengue_usd$year <- as.integer(dengue_usd$year)

alltogether <- left_join(flu_usd, gapminder, by = c("country", "year"))
alltogether <- left_join(alltogether, dengue_usd , by = c("country", "year"))



#infant_mortelity firtelety life expantie door flu and dengue in verschillende jaren i
#en beetje statestiek
flu_plot <- function(dataframe, land){
  dataframe %>% filter(country == land) %>%
    ggplot(aes(x = year, y = flu)) +
  geom_line() +
  geom_point()
```

```
  }

flu_plot(alltogether,"Netherlands")
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```

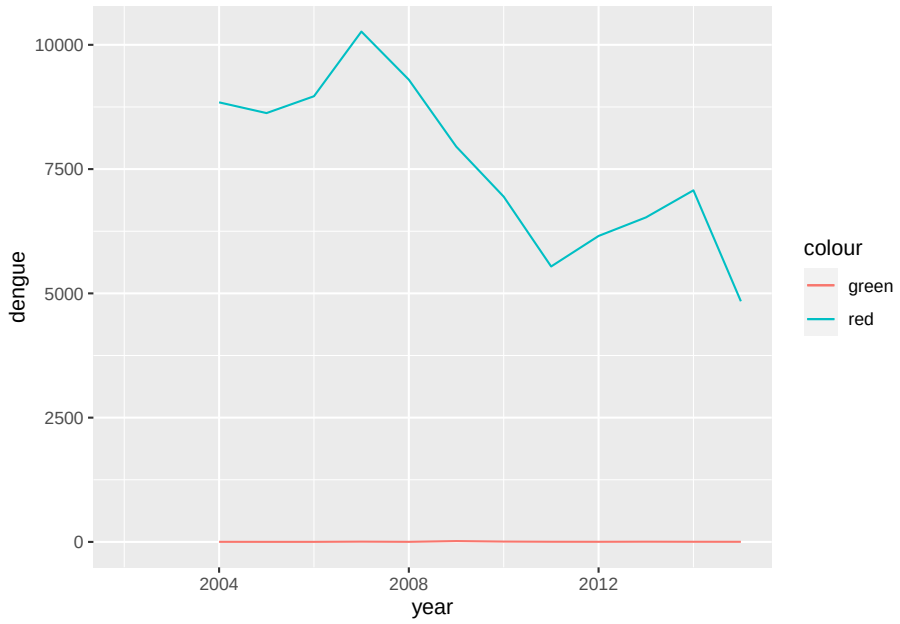
```
## Warning: Removed 2 rows containing missing values (geom_point).
```




```
alltogether %>% filter(country == "Argentina") %>%
  ggplot() +
  geom_line(aes(y = dengue,x=year, colour = "green"),) +
  geom_line(aes(y = flu,x=year, colour = "red"))
```

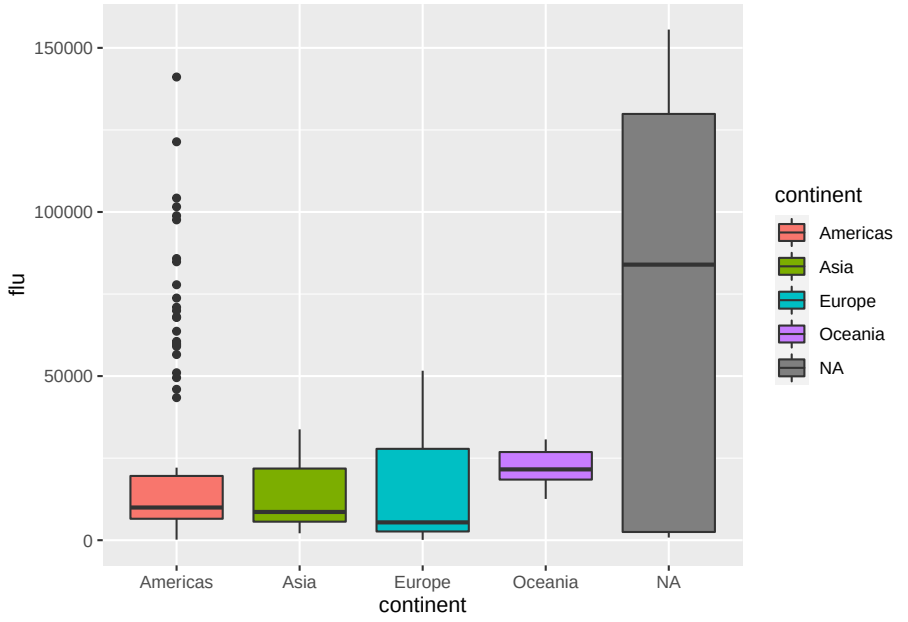
```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 2 row(s) containing missing values (geom_path).
```




```
ggplot(data = alltogether, aes(x = continent, y = flu)) +
  geom_boxplot(aes(fill = continent))
```

```
## Warning: Removed 72 rows containing non-finite values (stat_boxplot).
```




```
shapiro.test(alltogether$fertility)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$fertility
## W = 0.84528, p-value < 2.2e-16
```

```
shapiro.test(alltogether$flu)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$flu
## W = 0.70363, p-value < 2.2e-16
```

```
shapiro.test(alltogether$dengue)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$dengue
## W = 0.91218, p-value = 0.0009743
```

```
shapiro.test(alltogether$infant_mortality)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$infant_mortality
## W = 0.75988, p-value < 2.2e-16
```

```
shapiro.test(alltogether$life_expectancy)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$life_expectancy
## W = 0.93264, p-value = 9.214e-12
```

```
shapiro.test(alltogether$gdp)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$gdp
## W = 0.53327, p-value < 2.2e-16
```

```
shapiro.test(alltogether$population)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  alltogether$population
## W = 0.74646, p-value < 2.2e-16
```

Chapter 3

My own package

Chapter 4

Parameters

Chapter 5

Looking ahead

Chapter 6

Appendix

This code belongs to chapter 3 I dont let this script run because then I get a error because of the plots are to large

R code for: L?pez Steinmetz L.C., Dutto Florio M.A., Leyes C.A., Fong S.B., Rigalli A. & Godoy J.C. Levels and predictors of depression, anxiety, and suicidal risk during COVID-19 pandemic in Argentina: The impacts of quarantine extensions on mental health state. #"'{r}

library(tidyverse) library(readxl)
# Load the dataset: table<-read_excel("data/Peer/dataset.xlsx") summary(table)

6.0.0.0.0.1 SUB-TITLE: METHODS > Sample and procedure

Chapter 7

SAMPLE N = 1100