Portfolio Ghislaine

Ghislaine van Gilse

2023-01-03

Contents

Prerequisites

This portfolio was created as an assignment for Data Science for Biology 2 at the HU University of Applied Sciences Utrecht.

14 Nov. 2022 - 8 Jan. 2023

Curriculum Vitae

[**Insert Name here**]

Adres postcode

Telefoon: Email: Geboortedatum: Geboorteplaats:

Werkervaring

Tijdens mijn middelbare school en HBO-opleiding heb ik enkele bijbanen gehad:

Action Augustus 2015 - februari 2016, Hilversum

Hier heb ik als vakkenvuller medewerkster gewerkt.

Sociëteit de Unie Augustus 2018 - februari 2020, Hilversum Bij deze herensociëteit heb ik als ampulante kracht in de bediening gewerkt. Daarnaast heb ik ook achter de bar gestaan en incidenteel in de spoelkeuken gewerkt.

Kinderoppas September 2018 - mei 2020, Hilversum Opgepast op basis- en onderbouw (middelbare) schoolkinderen. Deze heb ik onder meer geholpen bij het maken van hun huiswerk en bij het leren voor toetsen.

Blokker Februari 2022 - heden, Hilversum Hier werk ik achter de kassa, maar ben ik ook in de winkel bezig met klanten, vakkenvullen en vracht.

Opleiding

Alberdingk Thijm College September 2013 - juli 2018, Hilversum Op deze school heb ik vijf jaar HAVO gedaan. Vakkenpakket Natuur & Gezondheid met natuurkunde en economie als keuzevakken.

ROC Midden-Nederland - VAVO Lyceum Utrecht / HAVO-diploma September 2018 - juli 2019, Utrecht Het vak natuurkunde gevolgd om mijn HAVO-diploma af te ronden.

Hogeschool Utrecht / Propedeuse September 2019 - heden, Utrecht Opleiding Life Science - Biologie en Medisch laboratorium onderzoek Propedeuse behaald: Januari 2021

Honours: Februari 2021: Mensenrechten & duurzame ontwikkeling

Vaardigheden

Labvaardigheden Onder andere: Celkweek, (licht- en fluorescentie)microscopie, SDS-page, Western-blot, (q)PCR, ELISA, (Agarosegel) elektroforese, kloneerwerk

Data verwerking (Basis)kennis van excel en SPSS Basiskennis van het werken met Rstudio waaronder, Bash R

Talen Nederlands (moedertaal) Engels B1

Hobby's & Interesses

(Berg)wandelen Paardrijden Skiën Trainen met de hond Bakken Puzzelen

Nevenactiviteiten

Gedurende drie jaar op de middelbare school van het Alberdingk Thijm College ben ik actief geweest als lid van de leerlingenraad.

Vrijwilligerswerk Comic Con November 2022

Chapter 1

Reproducible Research

1.1 C.elegans plate experiment

For this experiment, adult C.elegans nematodes have been exposed to varying concentrations of different compounds. The data that is used for this exercise is supplied y J. Louter (INT/ILC).

At first we had to review the Excel file with all the data. The file is called ./data.CE.LIQ.FLOW.062__Tidydata.xlsx. Something that stood out to me was that the different compounds were measured in nM. But the ethanol and the meduim of the cells was measured in percentage.

The compounds used in this experiment are: 2,6-diisopropylnaphthalene, Decane and Naphthalene with as positive control group 1,5% Ethanol in S-Medium and as negative control group just S-Medium.

```
elegansData <- read_excel(here("port_data/CE.LIQ.FLOW.062_Tidydata.xlsx"))
tibble(elegansData)
```

```
## # A tibble: 360 x 34
##    plateRow plateCo~1 vialNr dropC~2 expType expRe~3 expName expDate
##    <lgl>    <lgl>      <dbl> <chr>   <chr>     <dbl> <chr>   <dttm>
##  1 NA       NA             1 a       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  2 NA       NA             1 b       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  3 NA       NA             1 c       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  4 NA       NA             1 d       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  5 NA       NA             1 e       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  6 NA       NA             2 a       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  7 NA       NA             2 b       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  8 NA       NA             2 c       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
##  9 NA       NA             2 d       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
## 10 NA       NA             2 e       experi~       3 CE.LIQ~ 2020-11-30 00:00:00
## # ... with 350 more rows, 26 more variables: expResearcher <chr>,
## #   expTime <dbl>, expUnit <chr>, expVolumeCounted <dbl>, RawData <dbl>,
## #   compCASRN <chr>, compName <chr>, compConcentration <chr>, compUnit <chr>,
## #   compDelivery <chr>, compVehicle <chr>, elegansStrain <chr>,
## #   elegansInput <dbl>, bacterialStrain <chr>, bacterialTreatment <chr>,
## #   bacterialOD600 <dbl>, bacterialConcX <dbl>, bacterialVolume <dbl>,
## #   bacterialVolUnit <chr>, incubationVial <chr>, incubationVolume <dbl>, ...
```

After loading the data into Rstudio we were asked to inspect the data types of the columns: RawData, compName and compConcentration and what types we expected. I expect RawData: Numeric/integer, compName: factor and compConcentration: numeric/double. The typeof function tells me that RawData is a double. It doesn't contain decimals so it should be integer but this is not a big problem for further analyses. CompName is set as character which should be factor. CompConcentration is character. This is odd since the column contains numbers with decimals so this has to be double. This means the data types have not correctly been assigned while importing the data into R.

```
typeof(elegansData$RawData)
```

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

```
typeof(elegansData$compName)
```

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

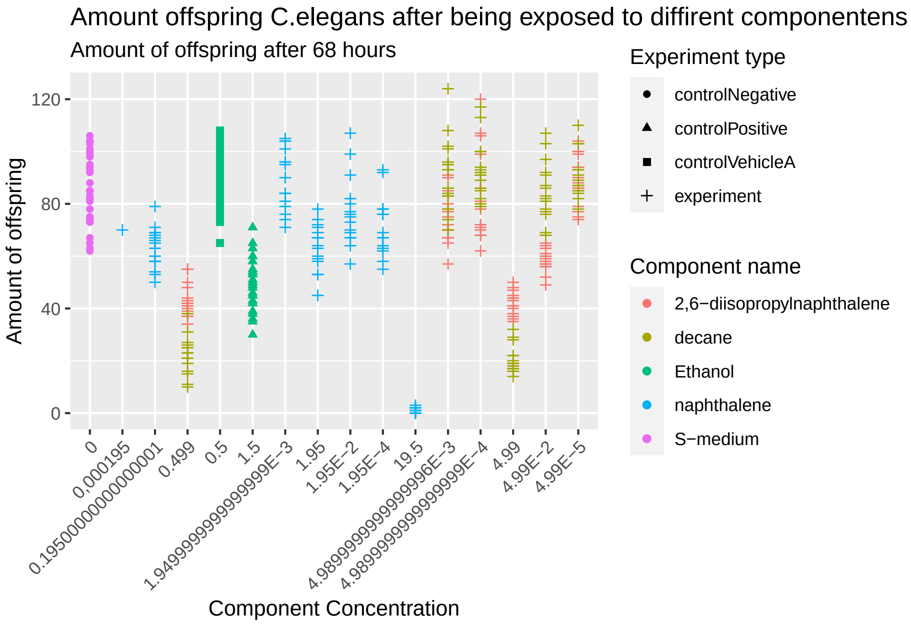
```
typeof(elegansData$compConcentration)
```

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

Not reassigning the datatypes while plotting the first plot. With compConcentration on the x-axis and RawData on the y-axis. Also each level of compName got its own colour each level in the expType got its own shape.

```
  ggplot(elegansData, aes(x=compConcentration, y=RawData, shape=expType, colour=compNam
  geom_point()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1, hjust=1))+
  labs(colour="Component name", shape="Experiment type",
       x="Component Concentration",
       y="Amount of offspring",
       title="Amount offspring C.elegans after being exposed to diffirent componentens"
       subtitle= "Amount of offspring after 68 hours"
)
```

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



As you can see the x-axis has been placed on alphabetic order and not numerical order. This is because the type of compConcentration is 'character' and not 'double'.

Let's reassign the datatypes with the correct ones

```
elegansData$RawData <- as.integer(elegansData$RawData)
elegansData$compName <- as.factor(elegansData$compName)
elegansData$compConcentration <- as.numeric(elegansData$compConcentration)
```

And let's check the types again!

```
class(elegansData$RawData)
```

```
## [1] "integer"
```

```
class(elegansData$compName)
```

```
## [1] "factor"
```

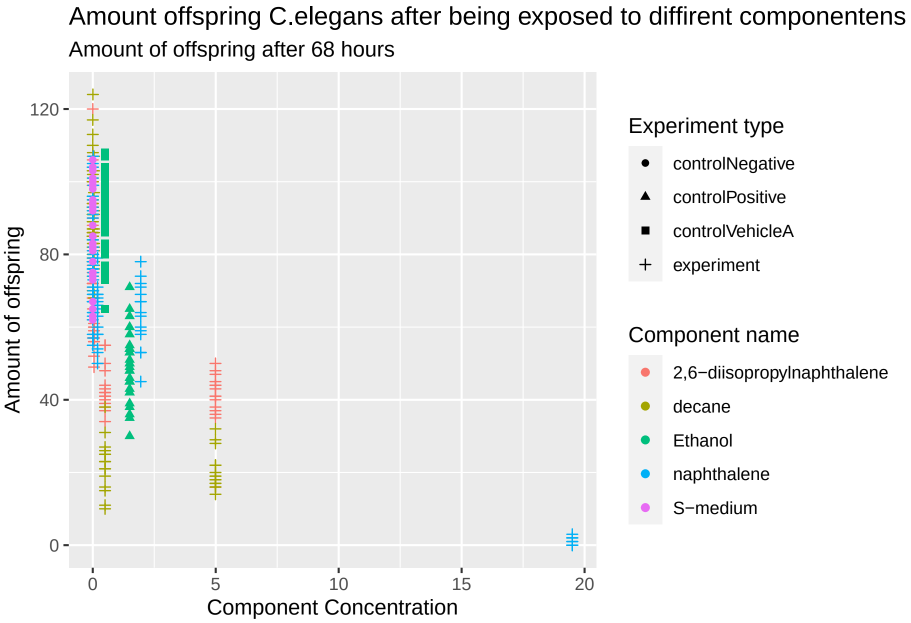
```
typeof(elegansData$compConcentration)
```

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

Now this has been fixed, let's see what the plot looks like.

```
ggplot(elegansData, aes(x=compConcentration, y=RawData, shape=expType, colour=compName
  geom_point()+
  labs(colour="Component name", shape="Experiment type",
       x="Component Concentration",
       y="Amount of offspring",
       title="Amount offspring C.elegans after being exposed to diffirent componentens'
       subtitle= "Amount of offspring after 68 hours"
)
```

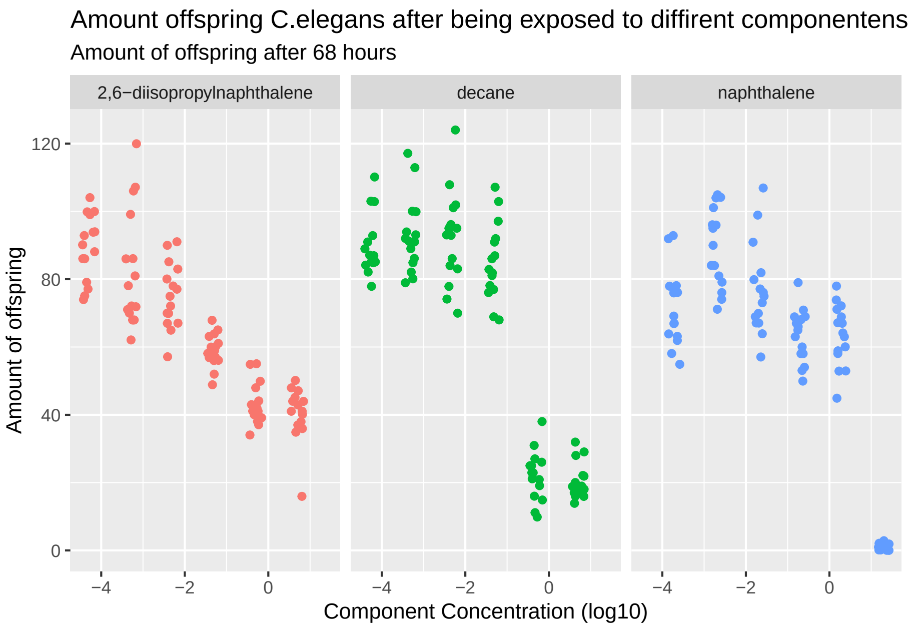
```
## Warning: Removed 6 rows containing missing values (`geom_point()`).
```



To get a clearer view, a log10 transformation is added on the x-axis. Also a bit of jitter is added so you can see the individual dots better. I've also separated the component.

```
elegansData_nM <- elegansData %>% filter(compUnit == "nM")

  ggplot(elegansData_nM, aes(x=log10(compConcentration), y=RawData, colour=compName))+
  geom_point(position = position_jitter(h=0.15,w=0.15), show.legend=FALSE)+
  scale_x_continuous()+
  labs(colour="Component name", shape="Experiment type",
       x="Component Concentration (log10)",
       y="Amount of offspring",
       title="Amount offspring C.elegans after being exposed to diffirent componentens",
       subtitle= "Amount of offspring after 68 hours",
       legend
)+
  facet_wrap(~compName)
```

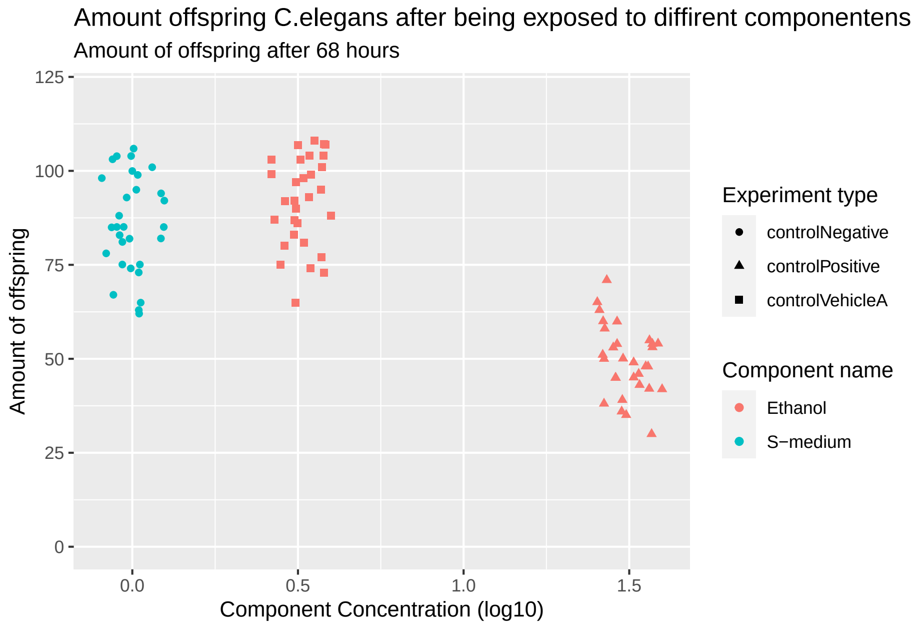
```
## Warning: Removed 6 rows containing missing values (`geom_point()`).
```



Amount offspring C.elegans after being exposed to diffirent componentens

Amount of offspring after 68 hours

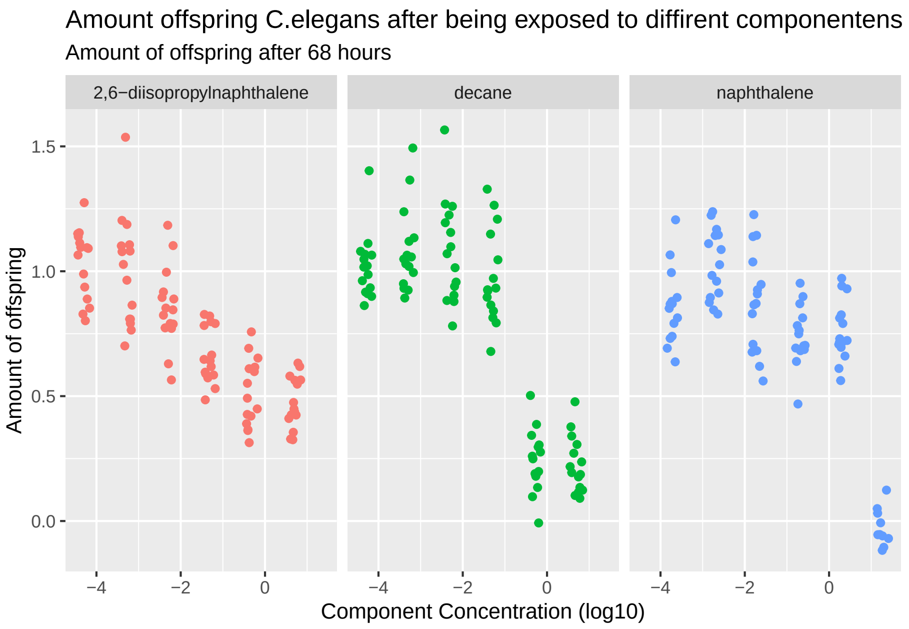
```
elegansData_pct <- elegansData %>% filter(compUnit == "pct")

ggplot(elegansData_pct, aes(x=compConcentration, y=RawData, shape=expType, colour=compName))+
  geom_point(position = position_jitter(h=0.1,w=0.1))+
  coord_cartesian(ylim = c(0, 120))+
  labs(colour="Component name", shape="Experiment type",
       x="Component Concentration (log10)",
       y="Amount of offspring",
       title="Amount offspring C.elegans after being exposed to diffirent componentens"
       subtitle= "Amount of offspring after 68 hours"
)
```



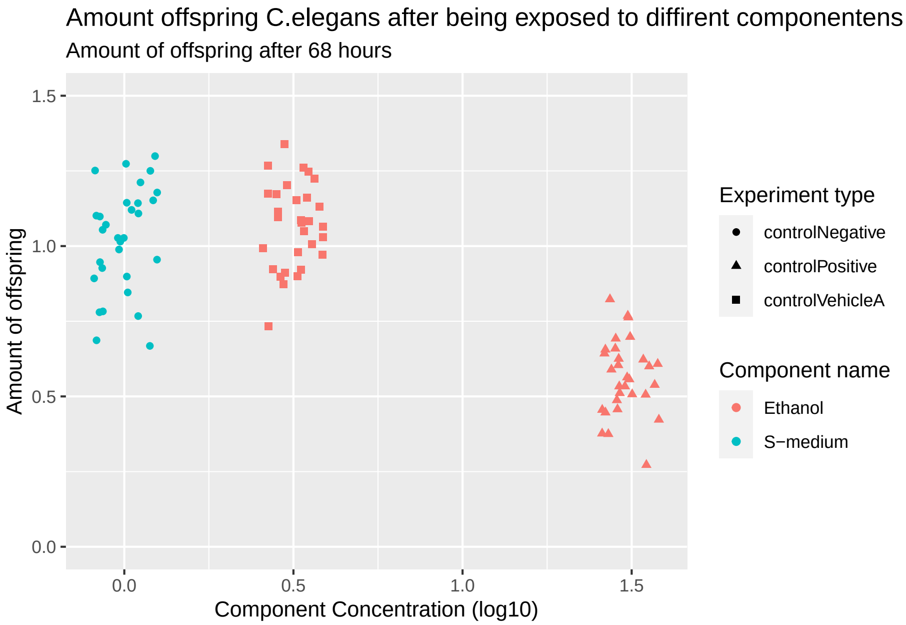
The positive control for this experiments is Ethanol (expType controlPositive in excel file). The negative control for this experiment is S-Medium (expType controlNegative in excel file).

For further possible analysis; To see if there is indeed an effect of different concentrations in offspring count I'd start with a Shapiro-Wilk test and normalize the data

```
MeanOfDataCtrlNeg <- elegansData %>% filter(expType == "controlNegative") %>% summariz

NormCelegansData <- elegansData %>%
  select(RawData, compName, compConcentration, expType, compUnit) %>% na.omit() %>%
  mutate(normalized = RawData/MeanOfDataCtrlNeg$mean)


elegansData_nM <- NormCelegansData %>% filter(compUnit == "nM")

  ggplot(elegansData_nM, aes(x=log10(compConcentration), y=normalized, colour=compName
  geom_point(position = position_jitter(h=0.15,w=0.15), show.legend=FALSE)+
  scale_x_continuous()+
  labs(colour="Component name", shape="Experiment type",
       x="Component Concentration (log10)",
       y="Amount of offspring",
       title="Amount offspring C.elegans after being exposed to diffirent componentens",
       subtitle= "Amount of offspring after 68 hours",
       legend
)+
  facet_wrap(~compName)
```



```
elegansData_pct <- NormCelegansData %>% filter(compUnit == "pct")

ggplot(elegansData_pct, aes(x=compConcentration, y=normalized, shape=expType, colour=compName))+
  geom_point(position = position_jitter(h=0.1,w=0.1))+
  coord_cartesian(ylim = c(0, 1.5))+
  labs(colour="Component name", shape="Experiment type",
       x="Component Concentration (log10)",
       y="Amount of offspring",
       title="Amount offspring C.elegans after being exposed to diffirent componentens",
       subtitle= "Amount of offspring after 68 hours"
)
```



The data has been normalized so it's easier to understand and work with.

Chapter 2

Open Peer Review

Portfolio assignment 1.2

This assigment is about identifying reproducibility issues in a scientific publication. Therefor we must use the criteria cited in table 1.

Work in progress

Chapter 3

Guerilla Principles

Here is how to install the fs package to make the dir tree for if needed

```
install.packages("fs")
```

If installed, this is the code to use to print the dir tree.

```
fs::dir_tree(here::here("rstudio-exportDaurII"))
```

This is what my daurII looks like in the Geurilla principles.

Chapter 4

Looking Ahead

For portfolio assigment 3.2 we were asked to answer the following quesitons.

- Where do I want to be in ~2 years time?
- How am I doing now with respect to this goal?
- What would be the next skill to learn?

In two years I hope to be done with school. I still have some subjects left and the internship is a period of a year. It all should be doable but of course we don't know what the future brings. I'd also like to start a job but I'm not sure what I like to do yet, so I keep an open mind. I'd like to combine data science with the practical work. If I could I'd like to work with DNA or genetics.

How I am doing with respect to this goal? Hmm it could be better. I've been having troubles with focussing so I end up starting at my screen all day and not doing anything productive. I still hope I can be done with school by February 2024.

So for a next skill it would be nice to learn more about genetics. So maybe trying to find if some mutations occur ofter in specific diseases.

Chapter 5

Portfolio assigment 5

Work in progress