Data Science Workflows course,

1. {desc} , makkelijkste manier om info te krijgen over de auteurs
2. *#install.packages("desc")*
3. **library**(desc)
4. desc <- description$**new**()
5. desc$**get\_authors**()

## werkt niet om een of andere reden

Exel bestand downloading

library(readxl)

FancySuppTable2 <- read\_excel("Lesson1/FancySuppTable2.xls")

View(FancySuppTable2)

De website die we gebruiken is gemaakt met bookdown

Kijken of ik het volgende allemaal heb:

1. Docker desktop - Gedaan
2. Visual studio code microsoft - Gedaan
3. Visual studio coder docker plugin Gedaan
4. R- at least version 4.0 - Ja
5. Rstudio IDE - Ja
6. Git for windows - ja
7. DBeaver - ja
8. CyberDuck - ja
9. Microsoft teams - ja

Hoe kan je iemand resulaten laten

Beginnen met synthetische data zou een goede manier zijn om te beginnen met werken. Omdat echte data altijd vervuilt is.

Het concept van reproduceerbaarheid is direct gelinkt met de scientific methode, the cornerstone of science en de volgende vijf stappen

1. Formulating a hypothesis
2. Designing the study
3. Running the study and collecting the data
4. Analyzing the data
5. Reporting the study

Je moet elke exercise in ene RMarkdown document maken

Sla het op ergens in de course project

Les 1 heeft 2 assignments sla ze op als verschillende files

1. Noem ze
2. Dsfb2\_workflows exercises
3. Dsfb2\_workflows\_portfolio
4. Dsfb2\_research\_project
5. Sla ze op locaal of Rstudio environment

--------------------------------------------------tip toevoegen van lees eerst alles door voordat je gaat werken

--------------------------------------------------noem de bestanden die je bruikt iets anders dan here: en This file

Vraag J

**Reading a strangely formatted column from an Excel file**

1. From within R, try downloading the following file: <http://genesdev.cshlp.org/content/suppl/2009/06/11/gad.1806309.DC1/FancySuppTable2.xls>
   1. Lukt me niet. Ik weet niet welke code ik hiervoor moet ggebruiken
2. What problems do you encounter, when trying to import this Excel file using {readxl}?
   1. Ik weet niet welke code ik moet gebruiken of waarvandaan ik het xl bestand moet halen
3. Perform a manual download of this file by going [here:](http://genesdev.cshlp.org/content/suppl/2009/06/11/gad.1806309.DC1). You need the file under the link Supp Table 2.xls
   1. Wat bedoel je hiermee
4. Why is manually downloading this file a problem with regard to open science?
   1. Je wilt dat het automatisch in R kunnen zetten. Als je manualyy het in R moet zetten elke keer en je hebt 1000 files werkt het niet.
   2. Inv met open Science is het een extra stap om bij de files te komen en open science is voor makkelijk acces voor alle bestanden
5. Import the manually downloaded Excel file into R
   1. Gedaan
6. Download [this file](https://ndownloader.figstatic.com/files/7148432)
   1. Gedaan
7. Open the file and compare the contents of the file to the Supplementary table shown [here](https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7#Sec1). Look specifically at the column Example Gene Name Conversion. What do you observe? Why do you think this is happening?
   1. Bij eentje werden de nummers een datum
   2. Bij andere zijn het nummers of een naam
8. Look at this article in the Microsoft help: <https://support.microsoft.com/en-us/office/stop-automatically-changing-numbers-to-dates-452bd2db-cc96-47d1-81e4-72cec11c4ed8>
   1. ok
9. Change the column Example Gene Name Conversion in the file [here:](http://genesdev.cshlp.org/content/suppl/2009/06/11/gad.1806309.DC1) to text.
   1. Ok
10. Change the column named GENE SYMBOL in [this file](http://genesdev.cshlp.org/content/suppl/2009/11/03/23.21.2484.DC2/Bilodeau_Supp_Table_11.xls) to text.
    1. Ook gedaan maar hoe helpt mij dit de gene called 40059 te vinden?
11. Now try to find the wrongly converted gene called 40059 in this data file
    1. Ik kan het vinden met grep
12. On the basis of your observations in this exercise: what are your conclusions about using Excel. How could you make the use of Excel *safer* for data collection and analysis? What would you recommend your peers? Write (in Rmd) a short recommendation on the safe use of Excel for data collection (min 200 words).
    1. Exel kan soms raare bug opleveren doordat het datums van nummers maakt die nummbers horen te zijn. Dit probleem kan ervoor zorgen dat de data niet bereikbaar is met algoritmes of het zorgt voor exstra werk. Dit probleem zou makkelijk kunnen worden opgelost door een optie in excel toe te brengen die er voor zorgt dat het automatisch omzetten naar datums niet meer wordt gedaan. Totdat excel dit probleem oplost zijn er mogelijkheden om deze automatische convertie tegen te gaan. Dit kan het makkelijkst gedaan worden door een spatie in te typen elke keer doordat er een nummer wordt opgeschreven of een (‘) te gebruiken voor elk getypte nummer.

Excel heeft een feature die voor veel problemen leid onder analisten. Deze feature zorgt er voor dat nummers die datums zouden kunnen zijn automatische worden geconvert naar datums. Dit is voor de algemene gebruiker soms handig maar voor analisten leid dit tot frustratie. Excel zou dit kunnen tegengaan door deze feature een optionele functie te maken die je kan uit en aanschakelen. Totdat excel dit heeft gedaan zijn er andere mogelijkheden om deze feature tegen te gaan. Door iets toe te voegen voordat er een nummer wordt ingetypt zet Excel het nummer niet om in een datum. De makkelijkste manier is een spatie te plaatsen elke keer voordat er een nummer wordt ingetypt. Wat ook kan is een ‘ te zetten voor elk nummer of een 0 en een spatie. Als je ziet dat je met data werkt waar mensen dit niet gedaan hebben kan je de cellen omzetten in tekst. Dit kan je doen door de row cellen met onbedoelde datums te selecteren, cntrl + 1 in te drukken en daarna tekst te selecteren en hierna ok te klikken . Buiten excel kan je de functie Paste met collapse = “, “ gebruiken om te zorgen dat de Data bruikbaar is.