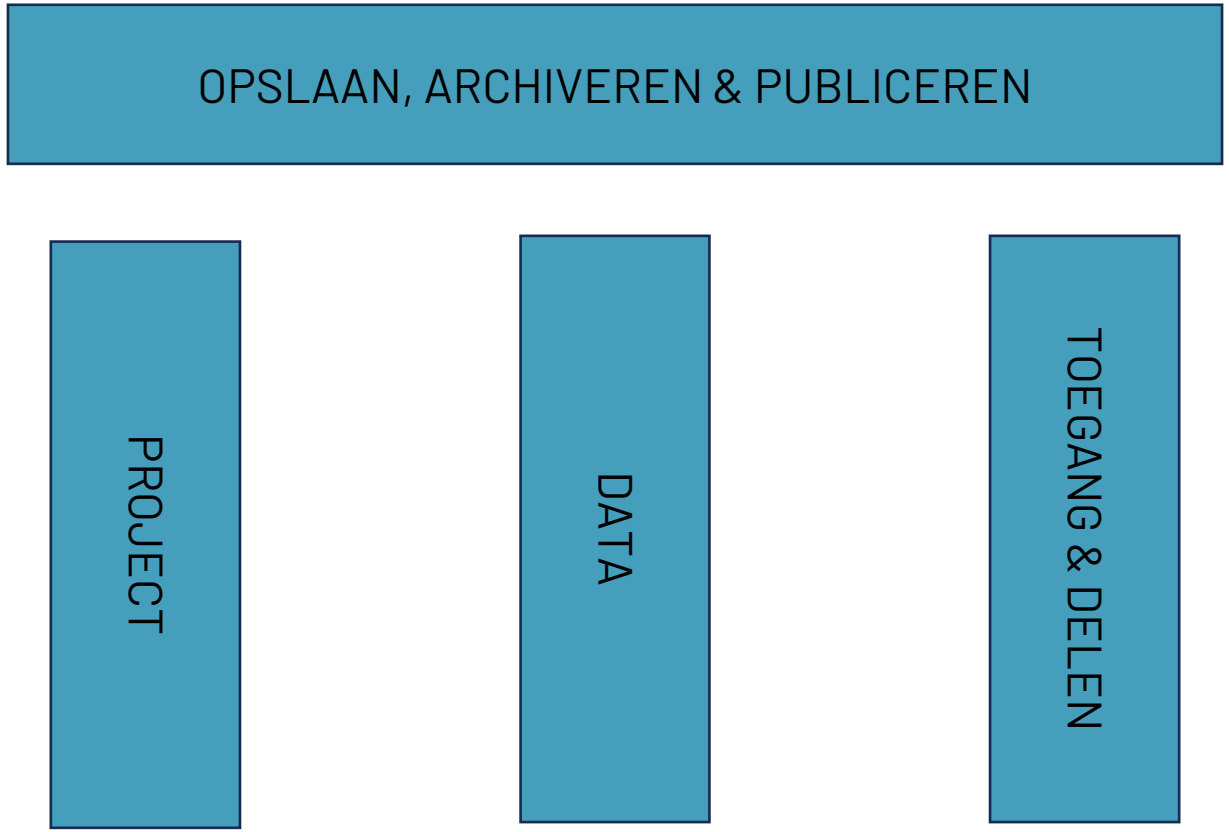
DCC-PO

Research Support for Applied Sciences

WAT IS DATADOCUMENTATIE?

DATADOCUMENTATIE



OPSLAAN, ARCHIVEREN EN PUBLICEREN

De vindbaarheid, toegankelijkheid en herbruikbaarheid van de data wordt door het goed documenteren verhoogd, zowel bij de opslag tijdens het project als bij het archiveren en publiceren.

- stel de mappenstructuur en bestandsnamen vast zodat de data in het onderzoeksproject efficiënt te vinden en herkennen zijn
- bepaal waar je efficiënt je datadocumentatie tijdens je onderzoek vastlegt en verzamelt
- hou de datadocumentatie bij tijdens je onderzoek. Zorg dat je een procedure maakt, zodat je op de juiste momenten hier aan herinnerd wordt en aan werkt
- werk tijdens je onderzoek je datamanagementplan steeds bij

PROJECT

Informatie over het project.

- Projectgegevens
  - medewerkers
  - instelling
- Financier
- Gerelateerde projecten en publicaties
  - projecten
  - publicaties uit de data

- Datamanagementplan

DATA

Beschrijving van ontstaan en bewerking van dataset.

- Omschrijving variabelen
  - Codeboek
  - (Electronic) Lab Notebooks
  - Scripts
  - SPSS en ATLAS.ti
- Omschrijving analyse
  - SPSS (Syntax)
  - R
  - Python
- Software ontwikkeld voor dit onderzoek
- Onderzoeksprotocol
- Informed consent
- Informatiebrief
- Vragenlijst
- Stimuli

TOEGANG EN DELEN

Informatie over hoe de onderzoeksdata toegankelijk wordt gemaakt.

- README-file over structuur bestanden
- Metadatastandaard
- Vocabulaire
- Persistent Identifier
- Licenties
- Bestandsformaten

PROJECTGEGEVENS

- naam project Nederlandstalig en Engelstalig
- afkorting/acroniem project
- korte samenvatting project
- looptijd project
- naam en contactgegevens hoofdonderzoeker/ORCID en/of DAI
- overige projectleden/rol/ORCID en/of DAI
- instelling
- lectoraat
- discipline
- trefwoorden project Nederlandstalig en Engelstalig

FINANCIER

- naam financier
- naam subsidie
- project/subsidienummer

GERELATEERDE PROJECTEN - PUBLICATIES

Hier noem je publicaties of projecten die de onderzoeksdata gebruiken.

- vermeld de volledige referentie van de publicatie

Voorbeeld:

Bovens, J., Moresi, S., Van den Langenberg, N., & Snoeren, M. (2022). De sleutel naar grensoverstijgend samenwerken in learning communities: De rol van bruggenbouwer. *Tijdschrift voor HRM*, *25*(3), 85-103. https://doi.org/10.5117/THRM2022.3.003.BOVE

DATAMANAGEMENTPLAN

Efficiënt omgaan met onderzoeksgegevens vereist planning. Steeds meer subsidieverstrekkers stellen eisen aan datamanagement en ook de Nederlandse gedragscode wetenschappelijke integriteit hanteert richtlijnen voor goed databeheer. Het helpt onderzoekers echter ook bij het inventariseren van risico's bij het beheer van onderzoeksgegevens gedurende het hele onderzoeksproces. Het datamanagementplan (DMP) is een levend document dat je voortdurend blijft updaten. Pas na afloop van je onderzoeksproject is het definitief. Een DMP schrijf je altijd met ondersteuning van een datasteward. De volgende punten komen aan bod in een DMP.

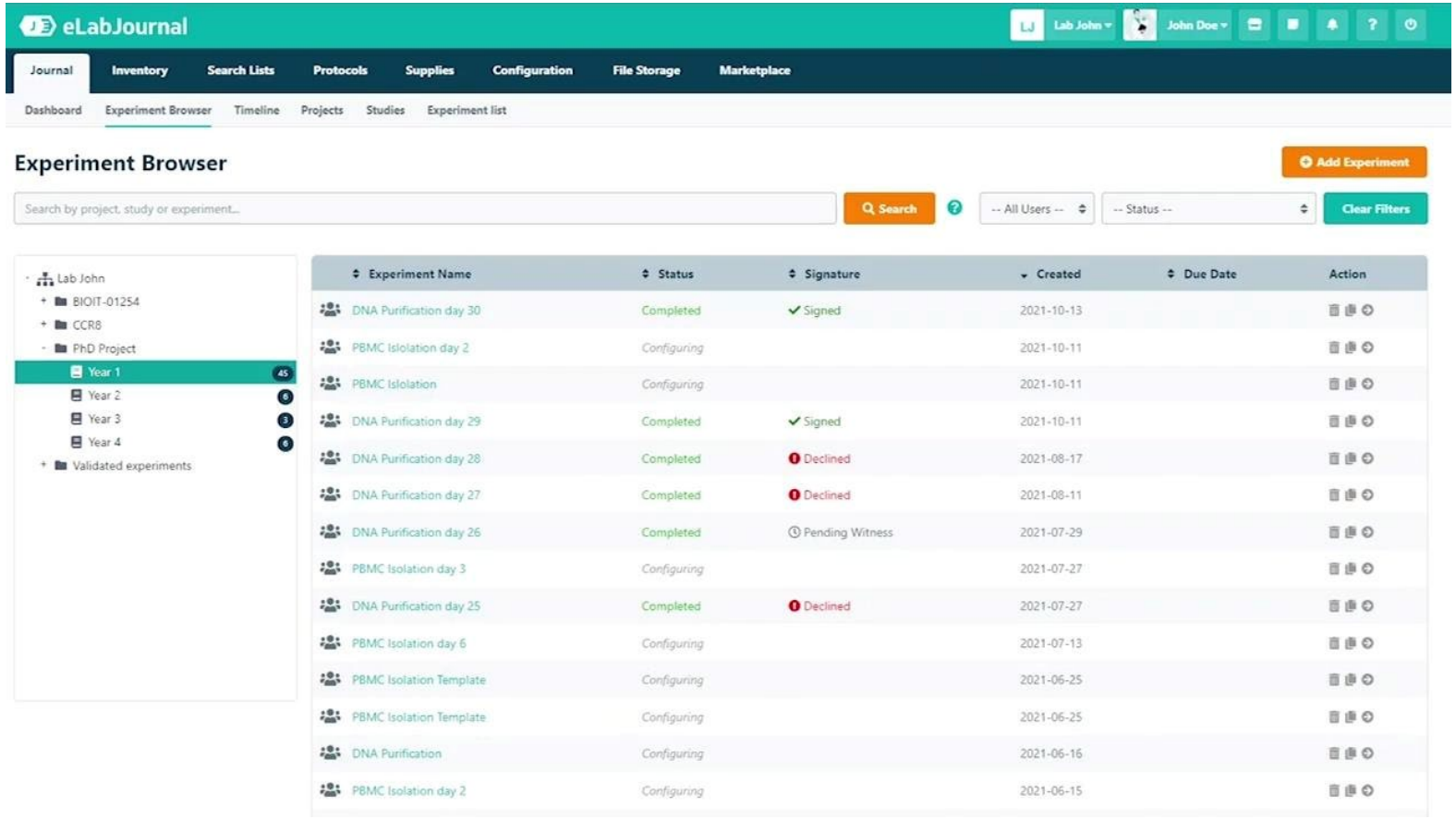
- administratieve gegevens
- beschrijving van de dataverzameling
- standaarden en metadata
- ethisch en juridisch
- opslag en autorisatie
- archivering en hergebruik
- bewaartermijn
- beschrijving software en tools

OMSCHRIJVING VARIABELEN CODEBOEK

- Contextbeschrijving: wanneer, waar en door wie
- Beschrijving van onderzoekspopulatie
- Selectiemethode
- Tool voor dataverzameling
- Codering van de variabelen
  - Naam en labels van de variabelen van de dataset
  - Codering antwoordopties
- Gerelateerde publicaties
- Meeteenheden
- Uitleg van afkortingen, concepten, categorieën
- Uitleg van codes en symbolen die gebruikt worden bij missing data
- Frequentietellingen, aantekeningen
- Voorbeelden: readable codebook in PDF format. Dit is hetzelfde codeboek in machine-actionable DDI-formaat (XML-formaat). Deze bestanden zijn gegenereerd met behulp van de DDI-Codebook-standaard en zijn gemarkeerd met Nesstar Publisher. Meer voorbeelden hier

Question		Variable Name	Value Codes for Responses
ID Number		ID	001-624
Q1. Is curbside pick-up available at your residence?		Curbpick	1 = No 2 = Yes 9 = Missing
Q2. Does your household currently participate in the city's curbside recycling program?		Curbsidp	1 = No 2 = Yes 9 = Missing
Q3. Does your household currently participate in the city's drop-off recycling program?		Dropoffp	1 = No 2 = Yes 9 = Missing
Q3a. What suggestions do you have for improving the drop-off recycling facility that you use?		Doimprov	1 = Increase collection freq. 2 = Clean up spilled materials 3 = Add more bins 4 = Mark bins more clearly 5 = Aluminum bins need larger openings 6 = Spray for bees/wasps 7 = Other 8 = Not applicable 9 = Missing
Q4. Right now, only some neighborhoods get their recyclables picked up at the curb as part of a pilot program. Would you favor or oppose expanding curbside recycling to all city households?		Expncurb	1 = Oppose 2 = Favor 3 = Not Sure/Don't Know 9 = Missing
•			
•			
•			
Q7. Would you be willing to take any of the following materials at least monthly to a drop-off center?	(Newspaper)	NEWS	0 = No, 1 = Yes, 9 = Missing
	(Plastic bottles)	PLASTIC	0 = No, 1 = Yes, 9 = Missing
	(Aluminum cans)	ALUM	0 = No, 1 = Yes, 9 = Missing
	(Glass bottles/jars)	GLASS	0 = No, 1 = Yes, 9 = Missing

OMSCHRIJVING VARIABELEN (ELECTRONIC) LAB NOTEBOOKS



OMSCHRIJVING VARIABELEN SCRIPTS

```r
# R program to add two numbers

# Assigning values to variables
a <- 9
b <- 4

# Printing sum
print(a + b)
```

rapportools (version 1.1)

max: Maximum

Description

Returns the maximum of all values in a vector by passing {codemax as `fn` argument to `univar` function.

Usage

```
max(...)
```

Arguments

...
parameters to be passed to `univar` function

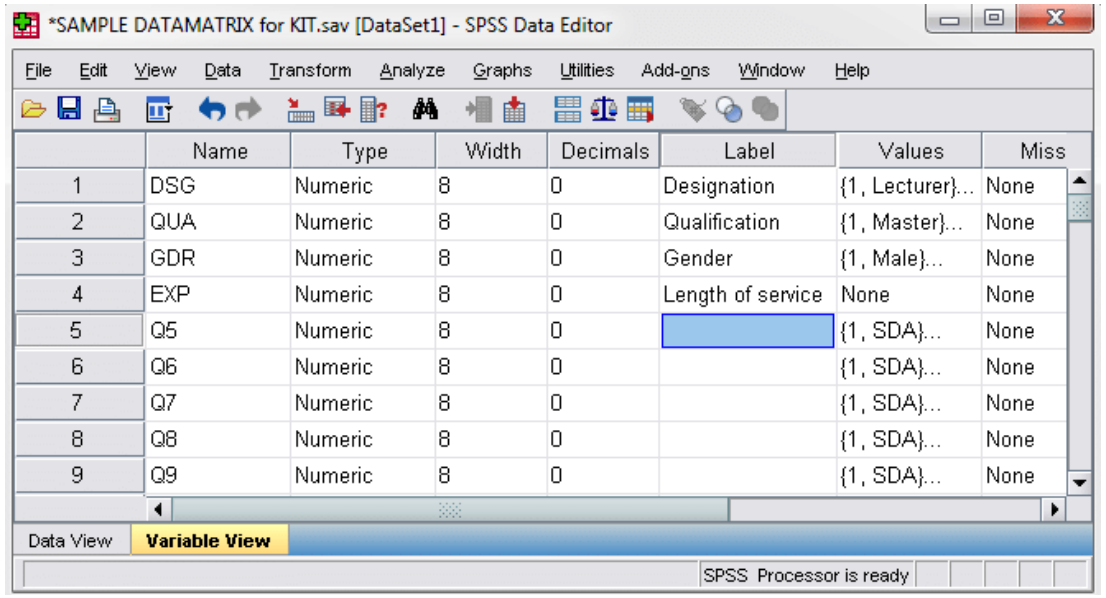
Value

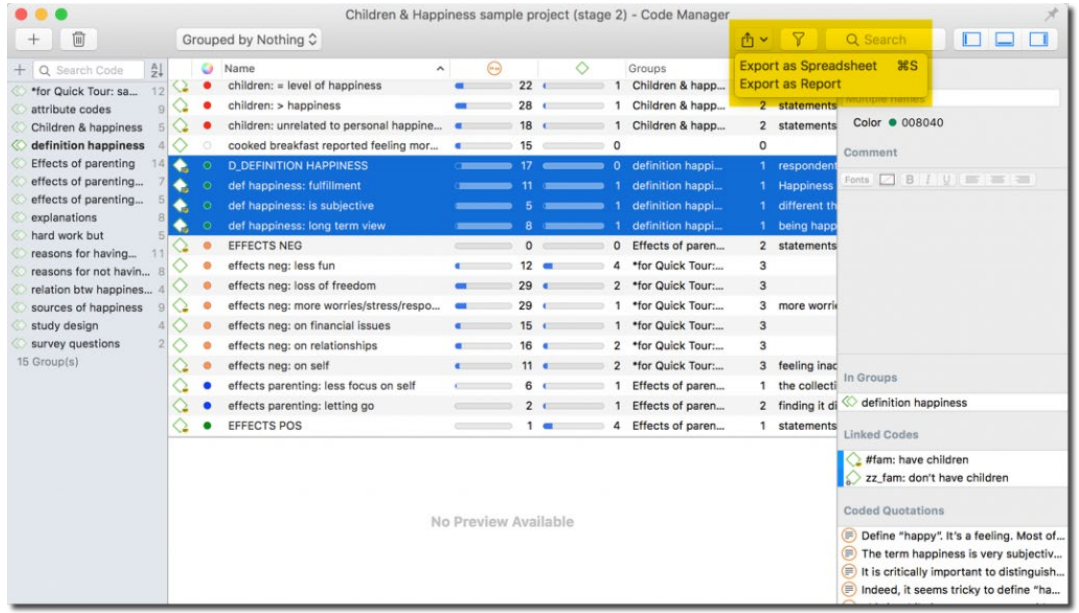
a numeric value with maximum value

```python
"""
Simple Python Program to Add Two Numbers

This script takes two numbers as input, adds them, and prints the result.

"""

def add_two_numbers(num1, num2):
    """
    Adds two numbers and returns the result.

    Parameters:
    - num1 (float): The first number.
    - num2 (float): The second number.

    Returns:
    float: The sum of num1 and num2.
    """
    result = num1 + num2
    return result
```

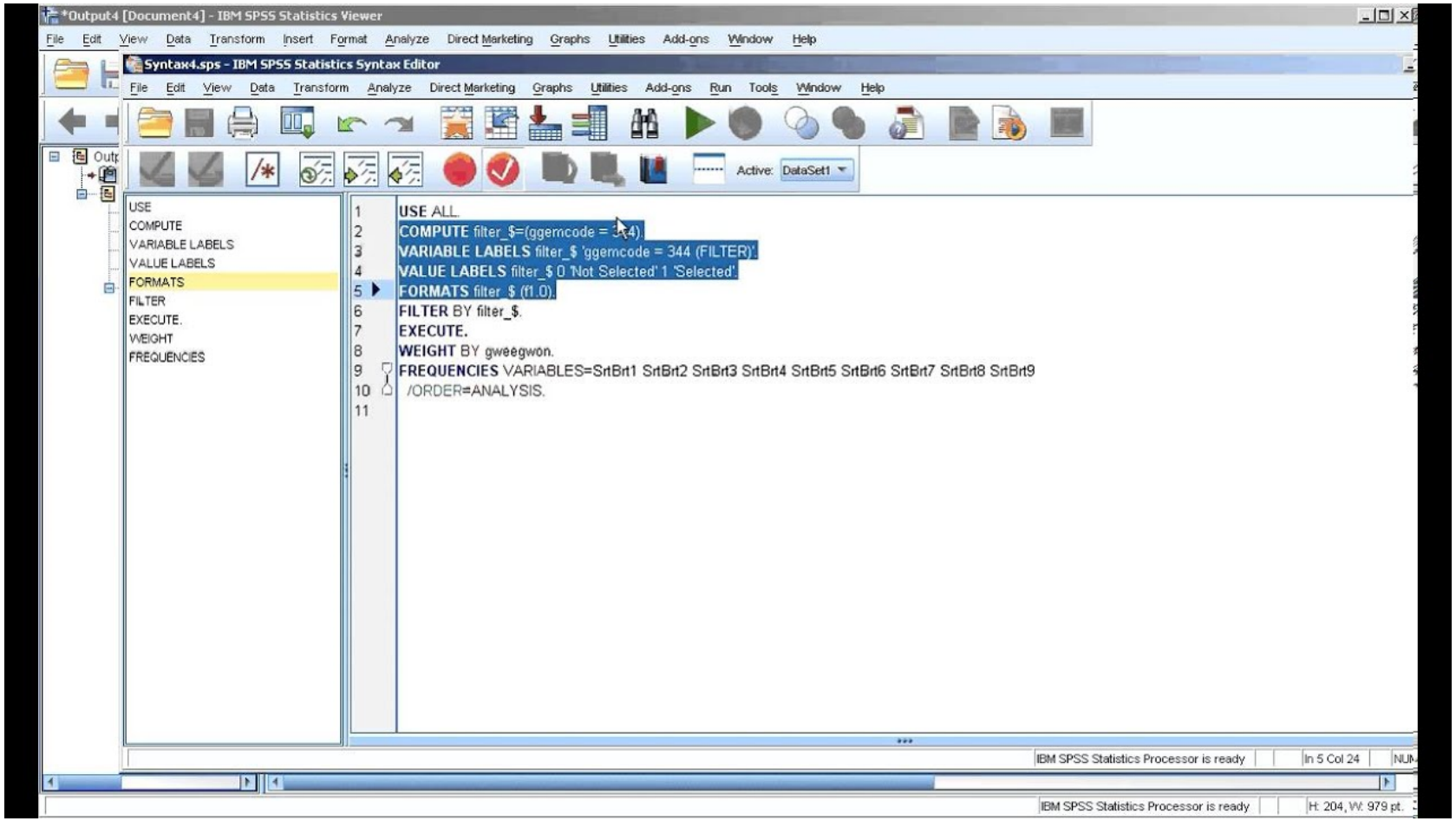
```python
def main():
    # Get user input for the two numbers
    num1 = float(input("Enter the first number: "))
    num2 = float(input("Enter the second number: "))

    # Call the add_two_numbers function
    sum_result = add_two_numbers(num1, num2)

    # Display the result
    print(f"The sum of {num1} and {num2} is: {sum_result}")

if __name__ == "__main__":
    # Call the main function when the script is executed
    main()
```

OMSCHRIJVING VARIABELEN SPSS EN ATLAS.TI

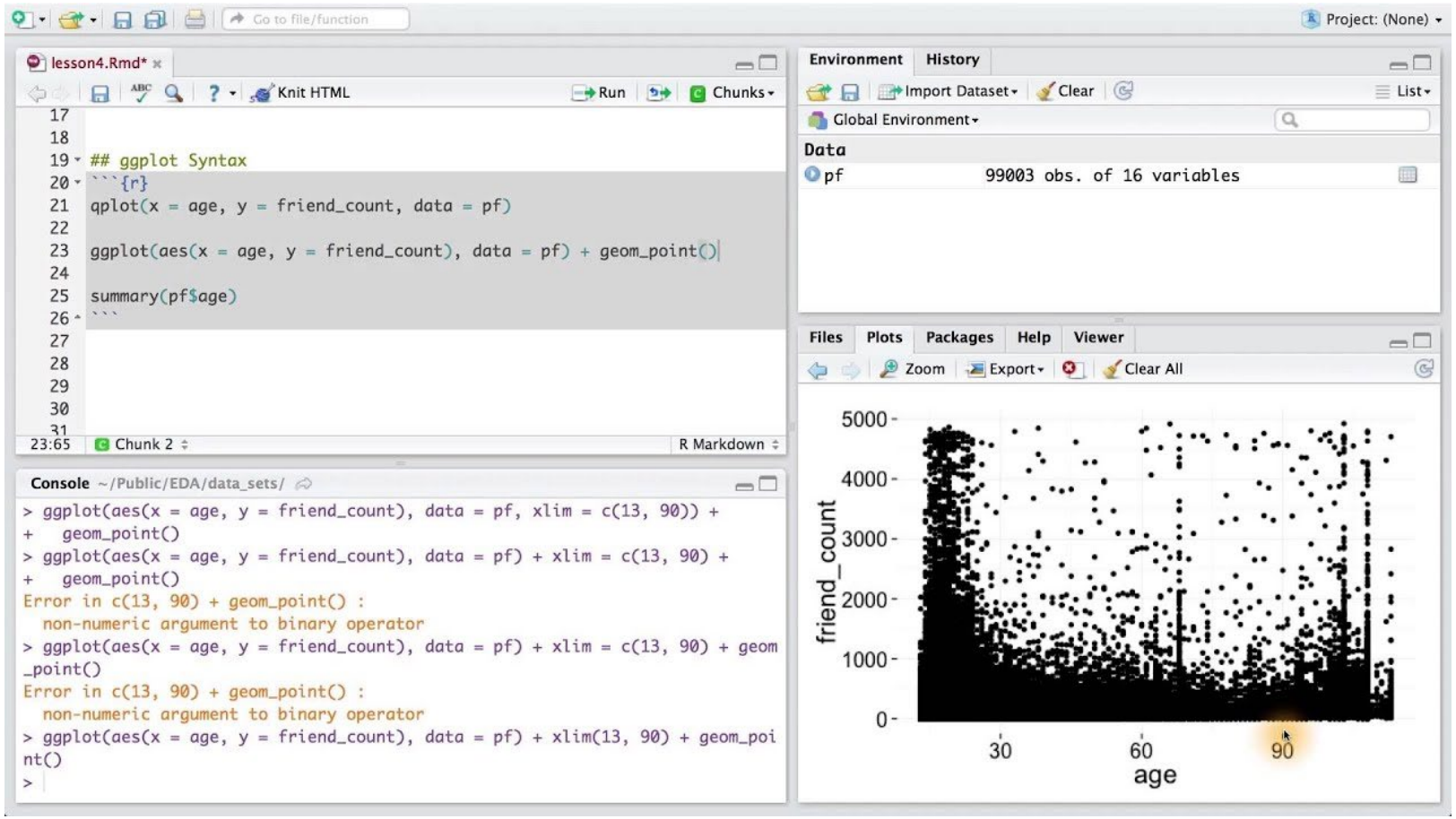




OMSCHRIJVING ANALYSE SPSS (SYNTAX)



OMSCHRIJVING ANALYSE R



OMSCHRIJVING ANALYSE PYTHON

```python
# importing libraries and modules
import catboost as cb
from sklearn.model_selection import train_test_split
import pandas as pd

# reading from a data source

train_df = pd.read_csv('heart_disease.csv')

# selecting a set of features and specifying the ground truth
train_df2 = train_df.iloc[:, 3:].values
train_x = train_df2.drop(['SSN', 'Target'], axis=1)
train_y = train_df2['Target']

# splitting the training data to train and validation sets
train_x2, val_x2, train_y2, val_y2 = train_test_split(
    train_x, train_y, test_size=0.20)

# initializing a model
clf = cb.CatBoostClassifier(eval_metric="AUC", iterations=40)

# training the model
clf.fit(train_x2, train_y2, eval_set=(val_x2, val_y2))
```

```python
import numpy as np
import random

def generate_float_list(lwr, upr, num):
    """
    Return a list of num random decimal floats ranged between␣
↪lwr and upr.

    Range(lwr, upr) creates a list of every integer between␣
↪lwr and upr.
    random.sample takes num integers from the range list,␣
↪chosen randomly.
    """
    int_list = random.sample(range(lwr, upr), num)
    return [x/100 for x in int_list]

# Create two lists
height = generate_float_list(100, 220, 10)
weight = generate_float_list(5000, 20000, 10)

# Convert these to Numpy arrays
np_height = np.array(height)
np_weight = np.array(weight)

print(np_height)
print(np_weight)
```

SOFTWARE

Wordt er onderzoekssoftware gemaakt binnen het onderzoek? Denk aan een software management plan, zie Practical guide to Software Management Plans.

Software publiceren kan via Research Software Directory.

ONDERZOEKSPROTOCOL

Een onderzoeksprotocol bevat het volgende:

- introductie (onderzoeksvraag)
- methoden
  - design
  - werving deelnemers
  - dataverzameling
  - data-analyse
  - datamanagement
- ethische overwegingen
- toestemming proefpersoon
- valorisatie en publicatie
- financiering

INFORMED CONSENT

- voor deelname aan onderzoek is geïnformeerde toestemming (informed consent) nodig van de deelnemer

- in je datadocumentatie neem je een niet-ingevuld toestemmingsformulier (informed consent) op

INFORMATIEBRIEF

Met een informatiebrief worden deelnemers voor het onderzoek geïnformeerd over:

- het doel van het onderzoek
- de opdrachtgever
- wie het onderzoek uitvoert
- wat er van de deelnemers verwacht wordt (welke handelingen verricht worden/welke vragenlijsten ingevuld moeten worden etc.)
- rechten van de deelnemer
- wat de voor- en nadelen van deelname aan het onderzoek zijn
- waar eventueel meer informatie over het onderzoek op te vragen is

VRAGENLIJST

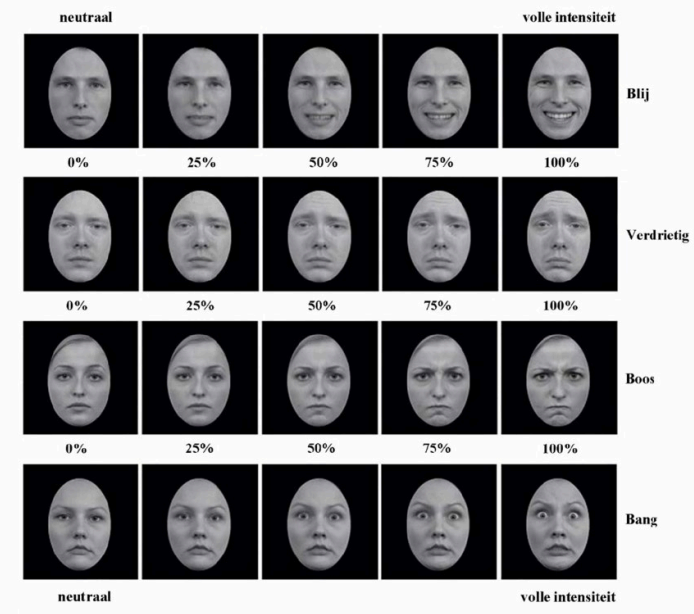
In je datadocumentatie neem je (een) niet-ingevulde vragenlijst(en) op.

Voorbeelden zijn:

- digitale vragenlijst
- vragen die je tijdens interview wil stellen
- overzicht met onderwerpen die je tijdens een interview of met een focusgroep wil bespreken
- observatieschema

STIMULI

Alles wat wordt aangeboden aan een deelnemer in een experimentele conditie om een bepaalde response te ontlokken.



Demonstratie Stroop-effect

Noem zo snel mogelijk de kleuren van de woorden.

Groen	Blauw
Blauw	Rood
Paars	Groen
Blauw	Rood
Rood	Paars
Congruent	Incongruent

Het Stroop-effect houdt in dat het meer tijd kost om de kleuren te benoemen in de incongruente dan in de congruente conditie.

VOCABULAIRE

Vocabulaires zijn woordenlijsten die gebruikt kunnen worden voor labelen, indexeren of categoriseren. Vocabulaires zorgen voor consistentie en interoperabiliteit en zijn machine leesbaar. Er zijn verschillende soorten:

- gecontroleerde vocabulaire: woordenlijst. Beleid bepaalt wie, wanneer en hoe woorden aan de lijst worden toegevoegd. Alleen woorden uit de lijst mogen gebruikt worden.
- taxonomie: classificatieschema, vaak in een boomstructuur. Entiteiten worden beschreven onderverdeeld en geordend in hiërarchische groepen en categorieën. Bevat criteria waaraan moet worden voldaan om tot de categorie te horen. Komt tot stand op basis van metingen en experimenten.
- thesaurus: gestructureerde gecontroleerde woordenlijst, waarin hiërarchische relaties worden gelegd naast gerelateerde termen en synoniemen.
- ontologie: woordenlijst, bevat abstracte en concrete concepten maar ook processen en taken. Tussen de concepten worden relaties aangegeven die zorgen voor structuur en ordening. Bevat ook de regels waaraan de concepten en relaties binnen een bepaald vakgebied moeten voldoen.
- typologie: een conceptuele classificatie (ontologie) op basis van logisch redeneren.

README-FILE

Een README-file kan voor het hele onderzoeksproject geschreven worden, maar als de dataset complex is ook per werkpakket of bestand.

Het bevat bijvoorbeeld:

- omschrijving van de mappenstructuur
- omschrijving van de totstandkoming van de benaming
- inhoud van de files en hun format
- relatie met andere datasets
- methodologische informatie
- informatie over de rechten en het al dan niet kunnen delen van (vertrouwelijke) data
- je kunt een README-file maken met het Windows Kladblok of met een MAC TextEdit (kies “Make plain tekst”). Meer complexe README-files zijn bv Markdown codeboeken
- Een voorbeeld template voor een README-file

METADATASTANDAARD

Metadatastandaarden zijn vastgestelde schema's van metadata bestaande uit verschillende elementen met informatie. Vaak heeft een repository al vastgesteld welke metadatastandaard je moet gebruiken.

- DataverseNL gebruikt bijvoorbeeld de Dublin Core Metadata Standard, bestaande uit 15 elementen DCMS_15 elements
- als er geen metadatastandaard gespecificeerd is in de geselecteerde repository, kun je zelf zoeken naar een geschikte metadatastandaard Guidance metadata standard

PERSISTENT IDENTIFIER

Een Persistent Identifier (PID of PI) is een permanente verwijzing en uniek label naar een digitale bron. De digitale bron kan een persoon, plaats of ding (bv. artikel of dataset) zijn. Het doel van een PID is dat met deze code de digitale bron altijd vindbaar blijft, ook als de URL (het webadres) van de digitale bron of de naam wordt gewijzigd. Een PID is dus een stabiele link naar een digitale bron.

Voorbeelden van een PID zijn:

- ORCID (Open Researcher and Contributor ID) is een PID voor onderzoekers. Om te waarborgen dat je publicaties daadwerkelijk aan jou gelinkt worden, is het van belang dat je met jouw ORCID in de onderzoekersdatabase van de Repository geregistreerd staat. Je kunt je ORCID aanvragen op de website van ORCID .
- DOI (Digital Object Identifier) is een PID voor een dataset en wordt bij deponeren in een repository automatisch toegekend. Dit hoef je dus niet zelf aan te vragen.
- RAiD (Research Activity Identifier) is een PID voor een onderzoeksproject
- ROR (Research Organization Registry) is een PID voor een onderzoeksinstelling

LICENTIES

Door je onderzoeksdata te voorzien van een licentie bepaal je vooraf in welke mate de onderzoeksdata hergebruikt mag worden door anderen.

- Creative Commons wordt vaak gebruikt voor data uit kwantitatief en kwalitatief onderzoek. Er zijn verschillende CC licenties waar je uit kunt kiezen
- voorbeelden van licenties voor software en code vind je bij het Open Source Initiative

BESTANDSFORMATEN

Voorkeursformaten zijn de bestandsformaten die te lezen zijn in vrij verkrijgbare software, waarvan goede en open documentatie beschikbaar is, en die veel gebruikt worden (binnen een bepaalde discipline).

- door te kiezen voor voorkeursformaten maak je je onderzoeksdata en datadocumentatie herbruikbaar, toegankelijk en duurzaam. Bestanden zijn dan over een aantal jaren nog te openen.
- repositories stimuleren het gebruik van voorkeursformaten door een overzicht te bieden, zoals 4TU.ResearchData en DANS. Repositories verschillen in hun aanbeveling.
- repositories maken meestal onderscheid tussen voorkeursformaten (preferred of recommended) en acceptabele formaten (non-preferred formats of acceptable formats).

MAPPENSTRUCTUUR

Maak je bestanden vindbaar door op het volgende te letten:

- laat de fase of type van onderzoek terugkomen in de mappenstructuur
- bewaar ruwe data in een aparte map
- maak niet te veel niveaus in je mappenstructuur
- documenteer de mappenstructuur in een README-file
- hou met de opzet van je mappenstructuur rekening met het delen van data

BESTANDSNAMEN

Maak je bestanden vindbaar door op het volgende te letten (zie ook RDNL)

- geef het bestand een betekenisvolle naam (Project1_Manuscript)
- hou een consistente benaming aan (Project1_Manuscript Project2_Manuscript_YYYYMMDD)
- gebruik geen spaties (maar _ of -) en geen speciale tekens
- benoem het versie nummer (Project1_Data_V1, Project1_Data_V2)
- maak de benaming niet te lang