Assignment 1 (A1): Git Hub Configuration and Data Cleaning

Instructions

Contents

0.1 A1.1: Introduction to version control 1
0.2 A1.2: Introduction to the dataset 2
0.3 A1.3: Data Manipulation Refresher 2

1 References **9**

The objective of this warm-up assignment is to brush up your programming and data wrangling skills as well as to introduce you to the data set you will be working on in Assignment 1. More precisely, you will here be asked here to do the following:

1. Introduction to version control
    - Create a **Github** (or **gitlab**) account
    - Link your **Git** account to your **RStudio**
    - Create a new repository/project
2. Introduction to the data set
3. Data Manipulation refresher

- Transform several data sets into a single one
    - Filter data so as to keep only the variables needed for next assignments
    - Tidyverse packages (**stringr** and **dplyr**)

0.1 A1.1: Introduction to version control

In the assignments you will be asked to upload your code on **Github** and the **GitHub** repositories will be part of the portfolio, therefore all students must make an account and link it to their **RStudio** (you'll thank us later for this!).

- Install latest **R** version
- Install latest **RStudio** version
- Create a Git account and link it to your RStudio

N.B. Create a GitHub repository for the Assignment 1 and link it to a project on your **RStudio**.

0.2 A1.2: Introduction to the dataset

Project: Language development in Autism Spectrum Disorder (ASD)

Source: (**fusaroli2019hearing?**)

Background: Autism Spectrum Disorder (ASD) is often related to language impairment, and language impairment strongly affects the patients ability to function socially (maintaining a social network, thriving at work, etc.). It is therefore crucial to understand how language abilities develop in children with ASD, and which factors affect them (to figure out e.g. how a child will develop in the future and whether there is a need for language therapy).

However, language impairment is always assessed by relying on the parent, teacher or clinician's subjective judgment of the child, and measured very sparsely (e.g. at 3 years old and again at 6). To help address this gap in clinical practice, we videotaped around 30 kids with ASD and around 30 comparison kids (matched by linguistic performance at visit 1) for approximately 30 minutes of naturalistic interactions with a parent. We repeated the data collection 6 times per kid at an interval of 4 months between each visit.

We transcribed the data and counted:

1. the amount of words that each kid uses in each video. Same for the parent.
2. the amount of unique words that each kid uses in each video. Same for the parent.
3. the amount of morphemes per utterance (Mean Length of Utterance) displayed by each child in each video. Same for the parent.

Different data sets were built by the researchers involved in the project:

1) demographic and clinical data about the children (recorded by a clinical psychologist)
2) length of utterance data (calculated by a linguist)
3) amount of unique and total words used

Your job in this assignment is to double-check the data and make it analysis-ready for the next assignment, in which we will try to understand how children language development is a function of cognitive and social factors and how the latter can be used as "cues" to likely future language impairments.

0.3 A1.3: Data Manipulation Refresher

If you have created a Rstudio project for this assignment, the working directory for this assignment (the default path for any function, procedure or command that works on files) is the project directory. If you have decided otherwise, first make sure that you set your working directory to the path of the current RMarkdown file. You can also use the following code chunk to install and/or load the packages you will need for the tasks below.

Load the following three data sets, after downloading them from dropbox and saving them in your working directory:

- Demographic data of the participants
- Length of utterance data
- Word data

```
participants <- read.csv("data/demo_train.csv")
LU <- read.csv("data/LU_train.csv")
W_data <- read.csv("data/token_train.csv")
```

Using various visualisation and aggregation procedures, explore these data sets and try to compare them to each other and draw parallel. You'll quickly notice that this is not an easy task. This is in part due to the fact that the different data sets were built by different researchers at different points in time. In particular, you'll find out that:

- the same variables might have different names (e.g. participant and visit identifiers)
- the same variables might report the values in different ways (e.g. participant and visit IDs)

Given this, it is important to make sure that all the relevant variables are identical in both name, type, and possible values.

```
#visualizations
glimpse(LU)
```

```
## Rows: 352
## Columns: 12
## $ SUBJ      <chr> "A.D.", "A.D.", "A.D.", "A.D.", "A.D.", "A.D.", "A.H.", "A.H~
## $ VISIT     <chr> "visit1.", "Visit2.", "visit3.", "visit4.", "visit5.", "visi~
## $ MOT_MLU   <dbl> 3.621993, 3.857367, 4.321881, 4.415330, 5.209615, 4.664013, ~
## $ MOT_LUstd <dbl> 2.164553, 2.417939, 2.517464, 2.449573, 2.814165, 2.765261, ~
## $ MOT_LU_q1 <dbl> 2, 2, 2, 2, 3, 2, 2, 2, 3, 3, 2, 3, 1, 2, 3, 3, 2, 1, 2, 3, ~
## $ MOT_LU_q2 <dbl> 4, 4, 4, 4, 5, 5, 3, 4, 5, 5, 4, 5, 3, 4, 5, 4, 4, 3, 4, 4, ~
## $ MOT_LU_q3 <dbl> 5.00, 6.00, 6.00, 6.00, 7.25, 7.00, 5.00, 6.00, 7.00, 8.00, ~
## $ CHI_MLU   <dbl> 1.252252, 1.013605, 1.556886, 2.251572, 3.238095, 2.865169, ~
## $ CHI_LUstd <dbl> 0.4739801, 0.1158462, 0.7470885, 1.5780274, 2.3559400, 2.247~
## $ CHI_LU_q1 <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 2, 1, 1, 1, 1, 1, 1, 1, 1, ~
## $ CHI_LU_q2 <dbl> 1, 1, 1, 2, 3, 2, 2, 2, 2, 4, 3, 4, 1, 1, 3, 1, 1, 1, 1, 1, ~
## $ CHI_LU_q3 <dbl> 1.00, 1.00, 2.00, 3.00, 5.00, 4.00, 2.00, 4.00, 4.00, 5.00, ~
```

```
#change VISIT to visit and cut visit and . out of the string, remove X column
glimpse(participants)
```

```
## Rows: 372
## Columns: 36
## $ Child.ID                   <chr> "A.A.", "A.D.", "A.D.", "A.D.", "A.D.", "A.~
## $ Visit                      <int> 1, 1, 2, 3, 4, 5, 6, 1, 2, 3, 4, 5, 6, 1, 2~
## $ Ethnicity                  <chr> "White", "White", "White", "White", "White"~
## $ Diagnosis                  <chr> "B", "B", "B", "B", "B", "B", "B", "B", "B"~
## $ ASD_check                  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ASD2                       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ Gender                     <int> 1, 1, 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2~
## $ Birthdate                  <chr> "18/02/09", "20/12/04", "20/12/04", "20/12/~
## $ Age                        <dbl> 18.07, 19.80, 23.93, 27.70, 32.90, 35.90, 4~
## $ Total..Understands...Says. <int> 0, 16, NA, NA, NA, NA, NA, 28, NA, NA, NA, ~
## $ Total..Understands.        <int> 183, 245, NA, NA, NA, NA, NA, 266, NA, NA, ~
## $ Total.of.Both              <int> 183, 261, 133, 563, 76, 95, 94, 294, 303, 5~
## $ Age2                       <dbl> 18.07, 19.80, 23.93, 27.70, 32.53, 35.90, 4~
## $ ADOS                       <int> 15, 0, NA, NA, NA, 0, NA, 1, NA, NA, NA, 0,~
## $ CARS                       <dbl> 29.0, 16.0, NA, NA, NA, 15.0, NA, 15.5, NA,~
## $ CDI1                       <int> 183, 261, NA, NA, NA, NA, NA, 294, NA, NA, ~
## $ VinelandStandardScore      <int> 90, 100, 100, 103, 107, 110, 106, 100, 103,~
## $ VinelandReceptive          <int> 20, 26, 27, 26, 29, 29, 31, 24, 25, 25, 27,~
```

```
## $ VinelandExpressive         <int> 19, 19, 31, 57, 73, 80, 87, 30, 60, 70, 75,~
## $ VinelandWritten            <int> NA, NA, NA, NA, NA, NA, 6, NA, NA, NA, NA, ~
## $ DailyLivingSkills          <int> 121, 119, 113, 109, 102, 107, 109, 91, 95, ~
## $ Socialization              <int> 104, 108, 110, 109, 102, 107, 100, 88, 89, ~
## $ MotorSkills                <int> 111, 111, 113, 114, 105, 114, 107, 84, 88, ~
## $ MullenRaw                  <int> NA, 28, NA, NA, 33, NA, 42, 29, NA, NA, 39,~
## $ MullenTScore               <int> NA, 66, NA, NA, 53, NA, 57, 59, NA, NA, 60,~
## $ MullenAge                  <int> NA, 26, NA, NA, 33, NA, 46, 27, NA, NA, 41,~
## $ FineMotorRaw               <int> NA, 22, NA, NA, NA, NA, 39, 21, NA, NA, NA,~
## $ FineMotorTScore            <int> NA, 52, NA, NA, NA, NA, 59, 36, NA, NA, NA,~
## $ FIneMotorAge               <int> NA, 21, NA, NA, NA, NA, 45, 20, NA, NA, NA,~
## $ ReceptiveLanguageRaw       <int> NA, 28, NA, NA, NA, NA, 40, 27, NA, NA, NA,~
## $ ReceptiveLanguageTScore    <int> NA, 72, NA, NA, NA, NA, 61, 59, NA, NA, NA,~
## $ ReceptiveLanguageAge       <int> NA, 30, NA, NA, NA, NA, 49, 28, NA, NA, NA,~
## $ ExpressiveLangRaw          <int> NA, 14, NA, NA, NA, NA, 44, 18, NA, NA, NA,~
## $ ExpressiveLangTScore       <int> NA, 33, NA, NA, NA, NA, 68, 38, NA, NA, NA,~
## $ ExpressiveLangAge          <int> NA, 14, NA, NA, NA, NA, 55, 18, NA, NA, NA,~
## $ EarlyLearningComposite     <int> NA, 112, NA, NA, NA, NA, 122, 96, NA, NA, N~
```

```
#nothing
glimpse(W_data)
```

```
## Rows: 352
## Columns: 8
## $ SUBJ         <chr> "A.D.", "A.D.", "A.D.", "A.D.", "A.D.", "A.D.", "A.H.", "~
## $ VISIT        <chr> "visit1.", "Visit2.", "visit3.", "visit4.", "visit5.", "v~
## $ types_MOT    <int> 378, 403, 455, 533, 601, 595, 334, 464, 482, 449, 534, 48~
## $ types_CHI    <int> 14, 18, 97, 133, 182, 210, 51, 149, 164, 206, 207, 173, 5~
## $ types_shared <int> 9, 15, 82, 113, 156, 181, 19, 130, 146, 165, 185, 145, 48~
## $ tokens_MOT   <int> 1835, 2160, 2149, 2260, 2553, 2586, 2674, 2694, 2630, 239~
## $ tokens_CHI   <int> 139, 148, 255, 321, 472, 686, 260, 530, 542, 754, 588, 46~
## $ X            <lgl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N~
```

```
#cut X column, change VISIT to visit, cut Visit and . out of the string
```

1. Renaming variables

The first task thus consists in identifying which variable names are spelled differently and rename them accordingly.

Tip: To find the right procedures and functions, you can:

- look through the chapter on data transformation in the book *R for data science* (**wickham2023?**)
- familiarize yourself with the package **dplyr** (which is part of the tidyverse)
- google "how to rename variables in R"
- check the **janitor** R package.

Many different procedures can be used here, and unless you have to deal with gigantic data sets and have severe limitations in terms of computation time and space, there is no need for optimization here. Whatever works works.

```r
#Fixing LU (wrong visit name, wrong child.id name) and W_data (useless column, wrong visit and child.id
names(LU)[names(LU) == "VISIT"] <- "Visit"
names(LU)[names(LU) == "SUBJ"] <- "Child.ID"

W_data <- subset(W_data, select = -X)
names(W_data)[names(W_data) == "VISIT"] <- "Visit"
names(W_data)[names(W_data) == "SUBJ"] <- "Child.ID"


#just manually removing the 2 versions of the visit string and the ., can definitely be done smarter
LU <- LU %>%
    mutate(Visit = gsub("visit", "", Visit)) %>%
    mutate(Visit = gsub("Visit", "", Visit)) %>%
    mutate(Visit = gsub("\\.", "", Visit))

W_data <- W_data %>%
    mutate(Visit = gsub("visit", "", Visit)) %>%
    mutate(Visit = gsub("Visit", "", Visit)) %>%
    mutate(Visit = gsub("\\.", "", Visit))
```

2. Renaming values

Find a way to homogenize the way "visit" is reported amongst the different datasets. The function **str_extract ()** from the package **stringr** can help you here: with the use of the regular expression '//d', this function allows you to capture the first occurrence of a number (*d* is for *digit*) within a string.

```r
### Write your code here
names(LU)[names(LU) == "VISIT"] <- "Visit"

LU <- LU %>%
    mutate(Visit = gsub("visit", "", Visit)) %>%
    mutate(Visit = gsub("Visit", "", Visit)) %>%
    mutate(Visit = gsub("\\.", "", Visit))

names(W_data)[names(W_data) == "VISIT"] <- "Visit"
W_data <- W_data %>%
    mutate(Visit = gsub("visit", "", Visit)) %>%
    mutate(Visit = gsub("Visit", "", Visit)) %>%
    mutate(Visit = gsub("\\.", "", Visit))
```

A similar task needs to be done regarding the value names of the Child.ID variable in the demographic data set. The values of this variable that are not abbreviated do not end with "." (i.e. Adam), whereas they do in the other two data sets (i.e. Adam.). Key merges, that is, merging of data sets based on shared variables, can only be done is the latter have overlapping value names; if no identical value names can be found, nothing will be merged. In the present case, a simple way to ensure this consists in removing all points from the values of the corresponding variables. The package **stringr** can again be of help here, notably the function **str_replace_all()**.

Tip: You can either have one line of code for each child name that needs to be changed (easier, more typing) or specify the pattern that you want to replace (more complicated: look up "regular expressions", but more generic and thus less typing)

```r
#Adding a . to the end of each name, unless there already is one
participants <- participants %>%
    mutate(Child.ID = gsub("(?<!\\.)$", ".", Child.ID, perl = TRUE))
#?<! is a negative lookbehind assertion, "what follows should not be preceded by a certain pattern
#$ the end of line
#\\. the dot character
#from ChatGPT
```

4. Data subsetting

This task consists in transforming the three data sets so as to keep only the variables that are of relevance for the project. For this purpose, the function **select()** from the **tidyverse** package **dplyr** will do nicely.

The variables of relevance here are the following:

- Child.ID,
- Visit,
- Diagnosis,
- Ethnicity,
- Gender,
- Age,
- ADOS,

- MullenRaw,
- ExpressiveLangRaw,
- Socialization
- MOT_MLU,
- CHI_MLU,
- types_MOT,
- types_CHI,
- tokens_MOT,
- tokens_CHI.

Most variables names should make sense, that is, should be straightforward as regards to the information it contains. Here are the less intuitive ones among the ones listed above:

- *ADOS* (Autism Diagnostic Observation Schedule) indicates the severity of the autistic symptoms (the higher the score, the worse the symptoms). Source: (**lord2000autism?**)
- *MLU* stands for 'Mean Length of Utterance'
- *types* stands for unique words. For example, the same word appearing multiple times only accounts for 1 word type.
- *tokens* stands for overall amount of words. Each occurrence of any word is a token. This example will help better understand the type/token distinction: in the sentence "the horse is a horse, of course, of course", there are 9 tokens (9 words in all), but only 6 types (the, horse, is, a, of, course).
- *MullenRaw* indicates non verbal IQ, as measured by the Mullen Scales of Early Learning (MSEL)(**lee2013mullen?**)
- *ExpressiveLangRaw* indicates verbal IQ, as measured by MSEL
- *Socialization* indicates social interaction skills and social responsiveness, as measured by (**volkmar1987social?**)

Feel free to rename the variables into something you think is more intuitive (i.e. *nonVerbalIQ*, *verbalIQ*)

```r
#joining dataframes
final <- inner_join(LU,W_data)
```

```
## Joining with `by = join_by(Child.ID, Visit)`
```

```r
#making them both integer to join
final$Visit <- as.integer(final$Visit)
participants$Visit <- as.integer(participants$Visit)

#joining dataframes
final <- inner_join(final, participants)
```

```
## Joining with `by = join_by(Child.ID, Visit)`
```

```r
#checking types
typeof(participants$Visit)
```

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

```r
typeof(final$Visit)
```

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

```r
#removing the useless or only keeping the useful
final_df <- final %>%
    select(Child.ID, Visit,Diagnosis, Ethnicity, Gender, Age, ADOS, MullenRaw, ExpressiveLangRaw, Social

#renaming to be more understandable, after merging
names(final_df)[names(final_df) == "ExpressiveLangRaw"] <- "Verbal_IQ"
names(final_df)[names(final_df) == "MullenRaw"] <- "NonVerbal_IQ"
```

- Child.ID,
- Visit,
- Diagnosis,
- Ethnicity,
- Gender,
- Age,
- ADOS,

- MullenRaw, –> “NonVerbal_IQ”
- ExpressiveLangRaw, –> “Verbal_IQ”
- Socialization
- MOT_MLU,

- CHI_MLU,
- types_MOT,
- types_CHI,
- tokens_MOT,
- tokens_CHI.

5. Data merge

Following completion of the previous cleaning procedures, the different data sets can now be merged into a single one.

It is important here to check if merging the data sets:

- has resulted in any loss of relevant data
- has resulted in the creation of NAs within the merged data set. If this is so, it is important to understand why these NAs were created (e.g. some measures were not taken at all visits, some recordings were lost or permission to use was withdrawn).

```
#they have been merged already --> final_df
```

6. Data filtering

In order for our models to be useful, we want to minimize the need to actually test children as they develop. In other words, we would like to be able to predict the children's linguistic development after only having tested them once. Therefore we need to make sure that our *ADOS*, *MullenRaw*, *ExpressiveLangRaw* and *Socialization* variables are reporting (for all visits) only the scores from visit 1.

A possible way to do so:

- create a new data set having only first visits as rows as well as *child.ID* and the 4 relevant clinical variables as columns
- rename the clinical variables in a way that clearly indicates they relate to the first visit (e.g. *ADOS* to *ADOS1*)
- merge the new data set with the old

only reporting the scores from visit 1! in *ADOS Non_VerbalIQ Verbal_IQ Socialization*, alongside *Child.ID*, *Visit*, *Gender*

```
#changing all parameter values to visit 1 parameter values
first_values <- final_df %>%
  group_by(Child.ID) %>%
  mutate(
    ADOS = ADOS[Visit == 1],
    NonVerbal_IQ = NonVerbal_IQ[Visit == 1],
    Verbal_IQ = Verbal_IQ[Visit == 1],
    Socialization = Socialization[Visit == 1]
  )

#just for convenience, but way too many variables
cleaned <- first_values
```

7. Reverse Encoding

An important part of data cleaning is making sure both variable names and value allow for easy and intuitive analysis and interpretation. For example, the different values for the categorical variable *gender* have been encoded to '1' and '2', but these numbers are useless to anyone who doesn't have the proper encoding dictionary, which maps each number to the corresponding category, namely 'male' or 'female'. In the following code chunk, resolve this ambiguity by reversing the encoding of the gender variable, so that 1 and 2 are changed to M and F respectively. In the same vein, transform the 'Diagnosis' variable so that 'A's and 'B's are reverted back to ASD (Autism Spectrum Disorder) and TD (typically developing) respectively.

```
#revert them to Gender 1=M, 2=F and Diagnosis A=ASD, B=TD
cleaned <- cleaned %>%
    mutate(Gender = ifelse(Gender == 1, "M", "F")) %>%
    mutate(Diagnosis = ifelse(Diagnosis == "A", "ASD", "TD"))
```

8. Saving cleaned data

Finally, save the cleaned data in Comma-Separated Values (CSV) format, under a name and at a location of your own choosing.

```
write.csv(cleaned, file = 'data/cleaned-IngridB.csv', row.names = FALSE)
```

Your data is now ready for the next assignment!

1 References