Statistics for Linguists
08 July 2022

10:00	Workshop introduction
10:15	Loading and exploring datasets
10:45	Data transformation and coding
11:15	Practical exercise
12:15	Review of practical
12:30 - 13:30	LUNCH BREAK
13:30	lmer and glmer
14:30	Post-hoc analysis and model visualization
15:00	Practical exercise
16:00	Review of practical
16:15	Model building
17:00	End of workshop

Statistics for Linguists

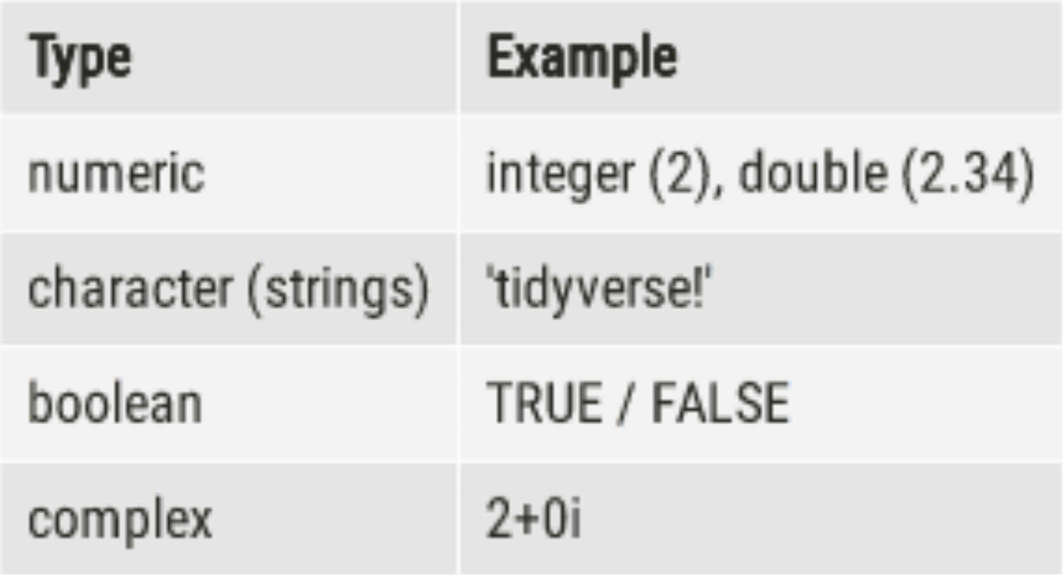
Loading and exploring datasets

Margreet Vogelzang – mv498@cam.ac.uk

Learning objectives

- **You will learn to load/import data**
- **Explore a dataset and create descriptive statistics**
- Transform a dataset (if needed)
- Code your factors
- Build a mixed model
- Perform post-hoc statistics
- Visualize your data and your model

4 main types of data



Type	Example
numeric	integer (2), double (2.34)
character (strings)	'tidyverse!'
boolean	TRUE / FALSE
complex	2+0i

Special types:

NA	# missing data
NULL	# empty
-Inf/Inf	# infinite values
NaN	# Not a Number

Structures

- Vectors

With c() you can concatenate to make a vector: c(43, 5.6, 2.90)

- Lists: can contain anything

```
list(f = factor(c("AA", "BB")),
     v = c(43, 5.6, 2.90),
     s = 4)
```

- Factors: are used to represent categorical data

```
> factor(c("AA", "BB", "AA", "CC")
[1] AA BB AA CC
Levels: AA BB CC
```

Structures

- data.frame: Similar to lists but all objects must have the same length. Defaut data structure in R

  ```
  data.frame(
          f = factor(c("AA", "AA", "BB")),
          v = c(43, 5.6, 2.90),
          s = rep(4, 3))
  ```

- data.table: enhanced, optimized version of the data.frame

- tibble: *tidyverse* data structure

Importing data

- Represents probably the first step of your work
- R can handle multiple data types
  - Flat files (.csv, .tsv, ...)
  - Excel files (.xls, .xlsx)
  - Foreign statistical formats
    - .sas from SAS
    - .sav from SPSS
    - .dta from Stata
  - databases (SQL, SQLite ...)

Importing data

- R base already provides functions for text files
  - read.csv()
  - read.delim()
  - ...
- Additional packages to load your data:
  - readr package:
    - read_csv(): comma separated (,)
    - read_csv2(): separated (;)
    - read_tsv(): tab separated
    - read_delim(): general delimited files, auto-guesses delimiter
    - read_table(): columns separated by white-space(s)

Importing data

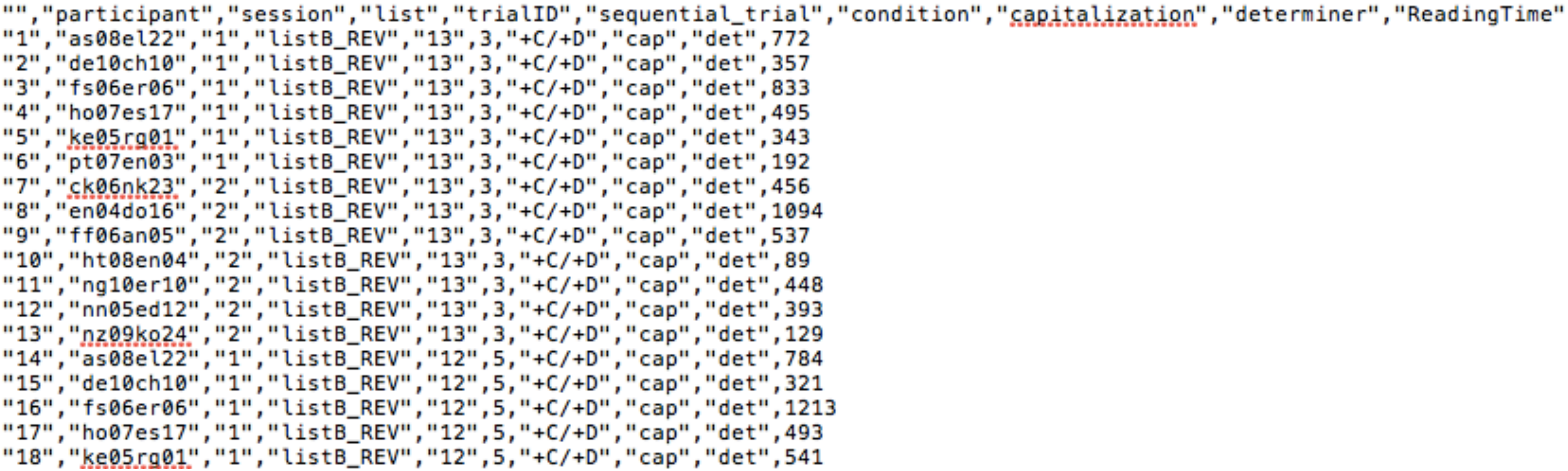
- R base already provides functions for text files
  - read.csv()
  - read.delim()
  - ...
- Additional packages to load your data:
  - readxl package:
    - read_excel()
    - read_xls()
    - read_xlsx()
  - haven package:
    - read_sas() for SAS
    - read_sav() for SPSS
    - read_dta() for Stata

Example: reading a .csv file

- Download pycholinguistics_data.csv to your computer
    - File can be found at https://margreetvogelzang.github.io/
- Open the file with a text viewer and have a look at its content
- Does the delimiter fit the file extension?

Example: reading a .csv file

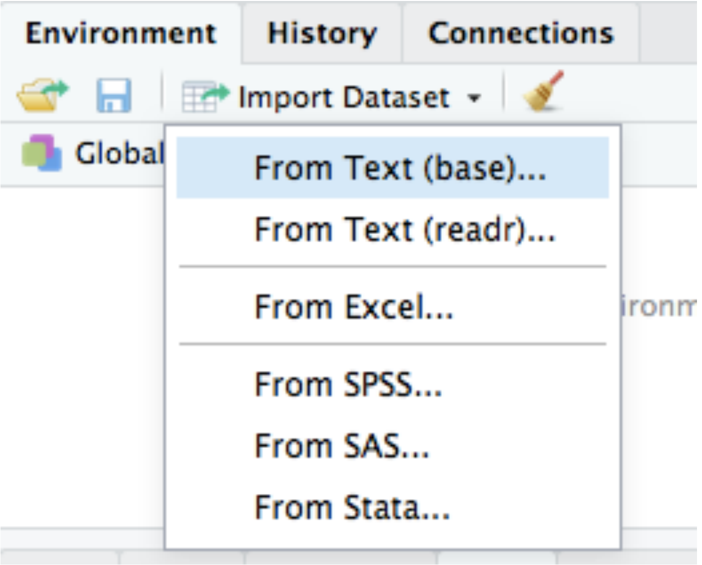
- Download pycholinguistics_data.csv to your computer
  - File can be found at https://margreetvogelzang.github.io/
- Open the file with a text viewer and have a look at its content
- Does the delimiter fit the file extension?

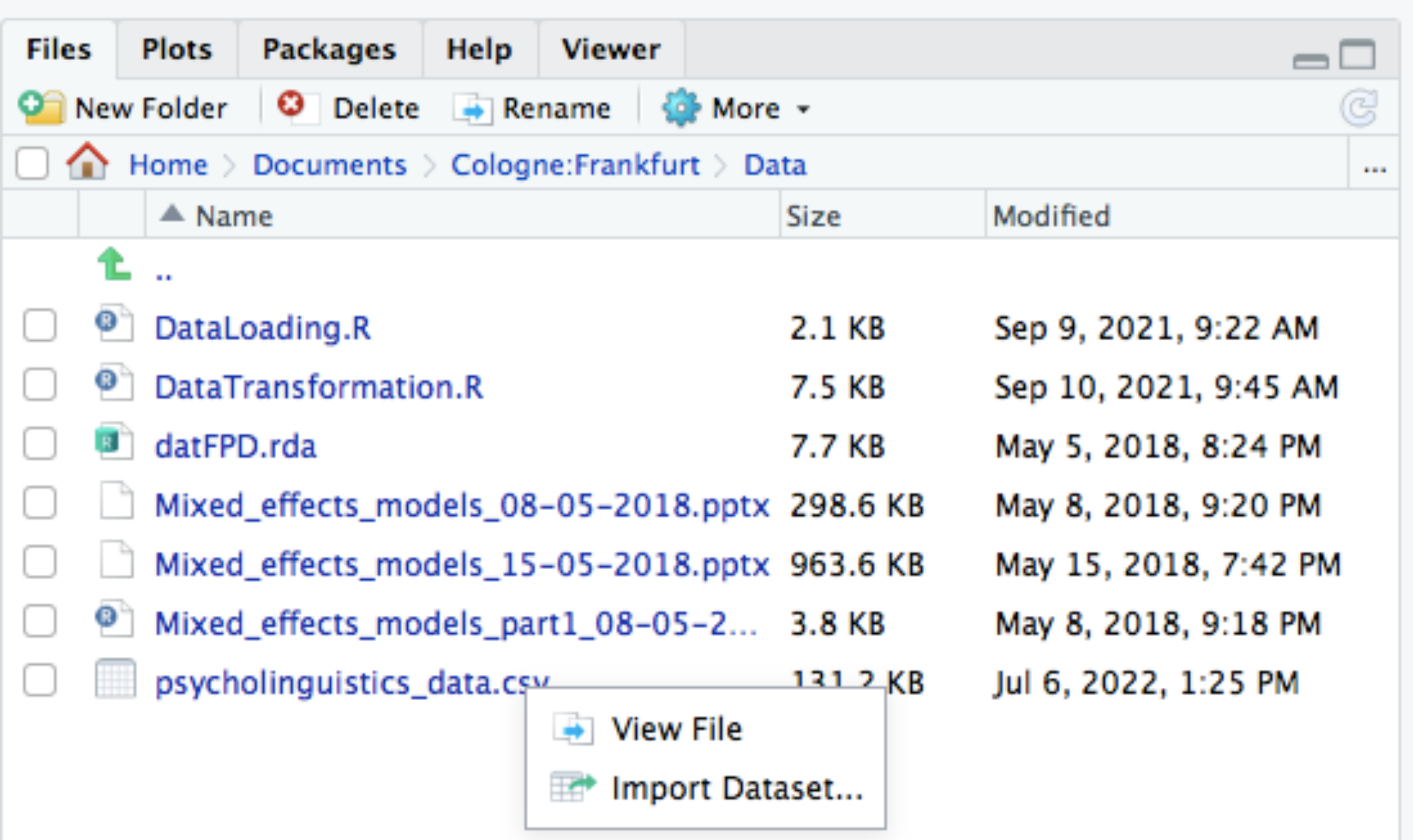


```
"","participant","session","list","trialID","sequential_trial","condition","capitalization","determiner","ReadingTime"
"1","as08el22","1","listB_REV","13",3,"+C/+D","cap","det",772
"2","de10ch10","1","listB_REV","13",3,"+C/+D","cap","det",357
"3","fs06er06","1","listB_REV","13",3,"+C/+D","cap","det",833
"4","ho07es17","1","listB_REV","13",3,"+C/+D","cap","det",495
"5","ke05rg01","1","listB_REV","13",3,"+C/+D","cap","det",343
"6","pt07en03","1","listB_REV","13",3,"+C/+D","cap","det",192
"7","ck06nk23","2","listB_REV","13",3,"+C/+D","cap","det",456
"8","en04do16","2","listB_REV","13",3,"+C/+D","cap","det",1094
"9","ff06an05","2","listB_REV","13",3,"+C/+D","cap","det",537
"10","ht08en04","2","listB_REV","13",3,"+C/+D","cap","det",89
"11","ng10er10","2","listB_REV","13",3,"+C/+D","cap","det",448
"12","nn05ed12","2","listB_REV","13",3,"+C/+D","cap","det",393
"13","nz09ko24","2","listB_REV","13",3,"+C/+D","cap","det",129
"14","as08el22","1","listB_REV","12",5,"+C/+D","cap","det",784
"15","de10ch10","1","listB_REV","12",5,"+C/+D","cap","det",321
"16","fs06er06","1","listB_REV","12",5,"+C/+D","cap","det",1213
"17","ho07es17","1","listB_REV","12",5,"+C/+D","cap","det",493
"18","ke05rg01","1","listB_REV","12",5,"+C/+D","cap","det",541
```

Example: reading a .csv file

- In RStudio, you can use the interface…





Name	Size	Modified
..		
DataLoading.R	2.1 KB	Sep 9, 2021, 9:22 AM
DataTransformation.R	7.5 KB	Sep 10, 2021, 9:45 AM
datFPD.rda	7.7 KB	May 5, 2018, 8:24 PM
Mixed_effects_models_08-05-2018.pptx	298.6 KB	May 8, 2018, 9:20 PM
Mixed_effects_models_15-05-2018.pptx	963.6 KB	May 15, 2018, 7:42 PM
Mixed_effects_models_part1_08-05-2…	3.8 KB	May 8, 2018, 9:18 PM
psycholinguistics_data.csv	131.2 KB	Jul 6, 2022, 1:25 PM

Example: reading a .csv file

- …or you can use code (recommended) to import it programmatically
  - Ensures reproducibility
  - Enables sharing with others

- Import the file, for example with read.csv()

```
> read.csv("data/pycholinguistics_data.csv")
```

- Make sure to assign the new dataset to a name

Example: reading a .csv file

- …or you can use code (recommended) to import it programmatically
  - Ensures reproducibility
  - Enables sharing with others

- Import the file, for example with read.csv()

```
> read.csv("data/pycholinguistics_data.csv")
```

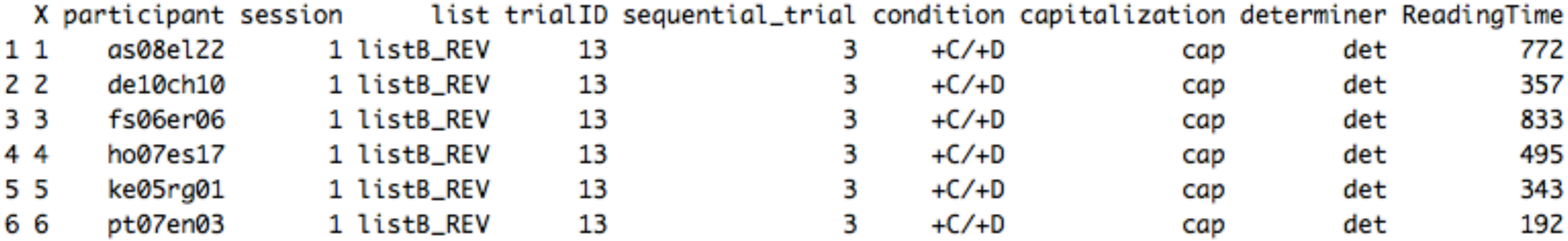
Check the documentation if your data file has particularities

- Make sure to assign the new dataset to a name

Exploring our dataset

- Data exploration functions:
  - head()

    > head(psycholinguistics_data)

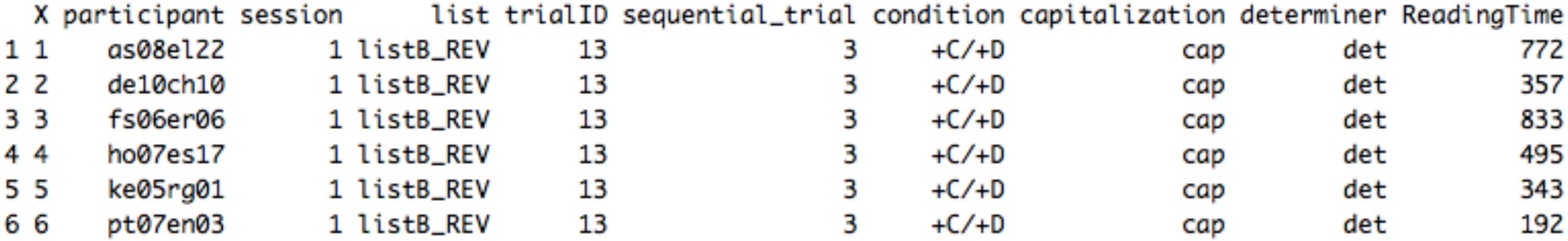


```
  X participant session      list trialID sequential_trial condition capitalization determiner ReadingTime
1 1    as08el22       1 listB_REV      13                3     +C/+D            cap        det         772
2 2    de10ch10       1 listB_REV      13                3     +C/+D            cap        det         357
3 3    fs06er06       1 listB_REV      13                3     +C/+D            cap        det         833
4 4    ho07es17       1 listB_REV      13                3     +C/+D            cap        det         495
5 5    ke05rg01       1 listB_REV      13                3     +C/+D            cap        det         343
6 6    pt07en03       1 listB_REV      13                3     +C/+D            cap        det         192
```

Exploring our dataset

- Data exploration functions:
  - head()

    > head(psycholinguistics_data)



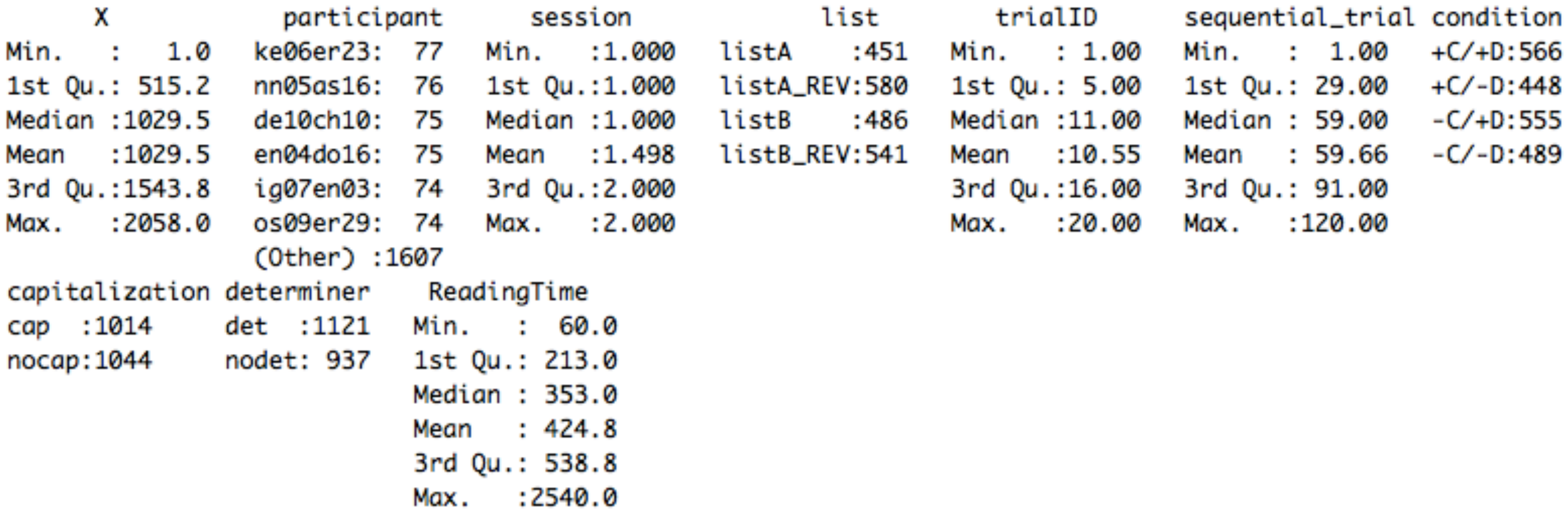
```
  X participant session      list trialID sequential_trial condition capitalization determiner ReadingTime
1 1    as08el22       1 listB_REV      13                3     +C/+D            cap        det         772
2 2    de10ch10       1 listB_REV      13                3     +C/+D            cap        det         357
3 3    fs06er06       1 listB_REV      13                3     +C/+D            cap        det         833
4 4    ho07es17       1 listB_REV      13                3     +C/+D            cap        det         495
5 5    ke05rg01       1 listB_REV      13                3     +C/+D            cap        det         343
6 6    pt07en03       1 listB_REV      13                3     +C/+D            cap        det         192
```

> tail(psycholinguistics_data)

Exploring our dataset

- Data exploration functions:
  - summary()

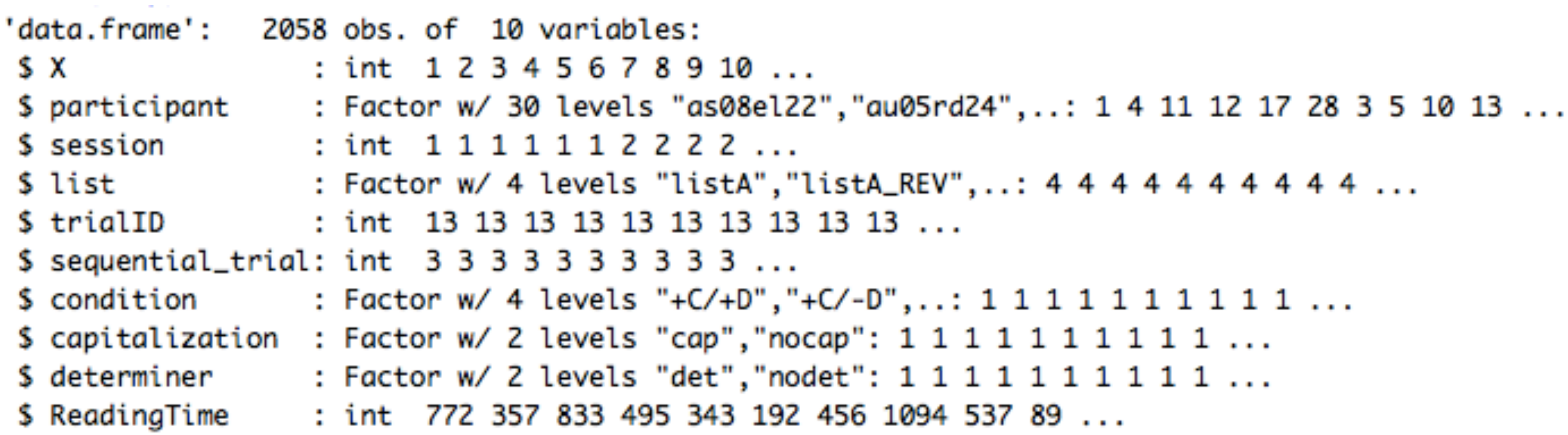
    > summary(psycholinguistics_data)



```
       X             participant      session               list          trialID       sequential_trial condition
 Min.   :   1.0   ke06er23:  77   Min.   :1.000   listA    :451   Min.   : 1.00   Min.   :  1.00   +C/+D:566
 1st Qu.: 515.2   nn05as16:  76   1st Qu.:1.000   listA_REV:580   1st Qu.: 5.00   1st Qu.: 29.00   +C/-D:448
 Median :1029.5   de10ch10:  75   Median :1.000   listB    :486   Median :11.00   Median : 59.00   -C/+D:555
 Mean   :1029.5   en04do16:  75   Mean   :1.498   listB_REV:541   Mean   :10.55   Mean   : 59.66   -C/-D:489
 3rd Qu.:1543.8   ig07en03:  74   3rd Qu.:2.000                   3rd Qu.:16.00   3rd Qu.: 91.00
 Max.   :2058.0   os09er29:  74   Max.   :2.000                   Max.   :20.00   Max.   :120.00
                  (Other) :1607
 capitalization determiner   ReadingTime
 cap  :1014     det  :1121   Min.   :  60.0
 nocap:1044     nodet: 937   1st Qu.: 213.0
                             Median : 353.0
                             Mean   : 424.8
                             3rd Qu.: 538.8
                             Max.   :2540.0
```

Exploring our dataset

- Data exploration functions:
  - str: display the internal **str**ucture of an R object
    > str(psycholinguistics_data)



```
'data.frame':   2058 obs. of  10 variables:
 $ X              : int  1 2 3 4 5 6 7 8 9 10 ...
 $ participant    : Factor w/ 30 levels "as08el22","au05rd24",..: 1 4 11 12 17 28 3 5 10 13 ...
 $ session        : int  1 1 1 1 1 1 2 2 2 2 ...
 $ list           : Factor w/ 4 levels "listA","listA_REV",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ trialID        : int  13 13 13 13 13 13 13 13 13 13 ...
 $ sequential_trial: int  3 3 3 3 3 3 3 3 3 3 ...
 $ condition      : Factor w/ 4 levels "+C/+D","+C/-D",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ capitalization : Factor w/ 2 levels "cap","nocap": 1 1 1 1 1 1 1 1 1 1 ...
 $ determiner     : Factor w/ 2 levels "det","nodet": 1 1 1 1 1 1 1 1 1 1 ...
 $ ReadingTime    : int  772 357 833 495 343 192 456 1094 537 89 ...
```

Exploring our dataset: some functions

- summary(psycholinguistics_data[10])
- summary(psycholinguistics_data$ReadingTime)
- names(psycholinguistics_data)
- unique(psycholinguistics_data$participant)
- length(unique(psycholinguistics_data$participant))
- nrow(psycholinguistics_data)
- ncol(psycholinguistics_data)

- tapply(psycholinguistics_data$ReadingTime,
  psycholinguistics_data$participant, mean)

Exploring our dataset: some functions

Multiple ways to achieve the same thing:

- tapply(psycholinguistics_data$ReadingTime,
  psycholinguistics_data$participant, mean)
- data.table(psycholinguistics_data)[,list(ReadingTime = mean(
  ReadingTime)),by=list(participant)] # requires data.table package
- psycholinguistics_data %>% group_by(participant) %>%
  summarise(mean_ReadingTime = mean(ReadingTime),
  sd_ReadingTime = sd(ReadingTime)) # requires tidyverse package

Statistics for Linguists
08 July 2022

10:00	Workshop introduction
10:15	Loading and exploring datasets
10:45	Data transformation and coding
11:15	Practical exercise
12:15	Review of practical
12:30 - 13:30	LUNCH BREAK
13:30	lmer and glmer
14:30	Post-hoc analysis and model visualization
15:00	Practical exercise
16:00	Review of practical
16:15	Model building
17:00	End of workshop