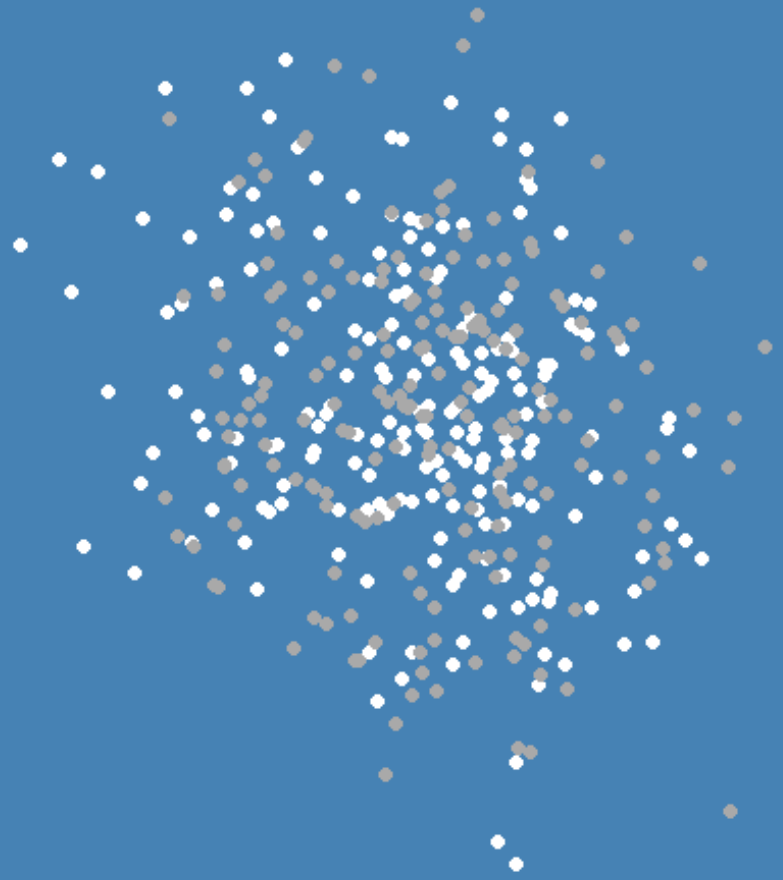
Introduction to R

3 Exploratory Data Analysis

Summary statistics, (Cross-)Tabulations, Correlation Matrices

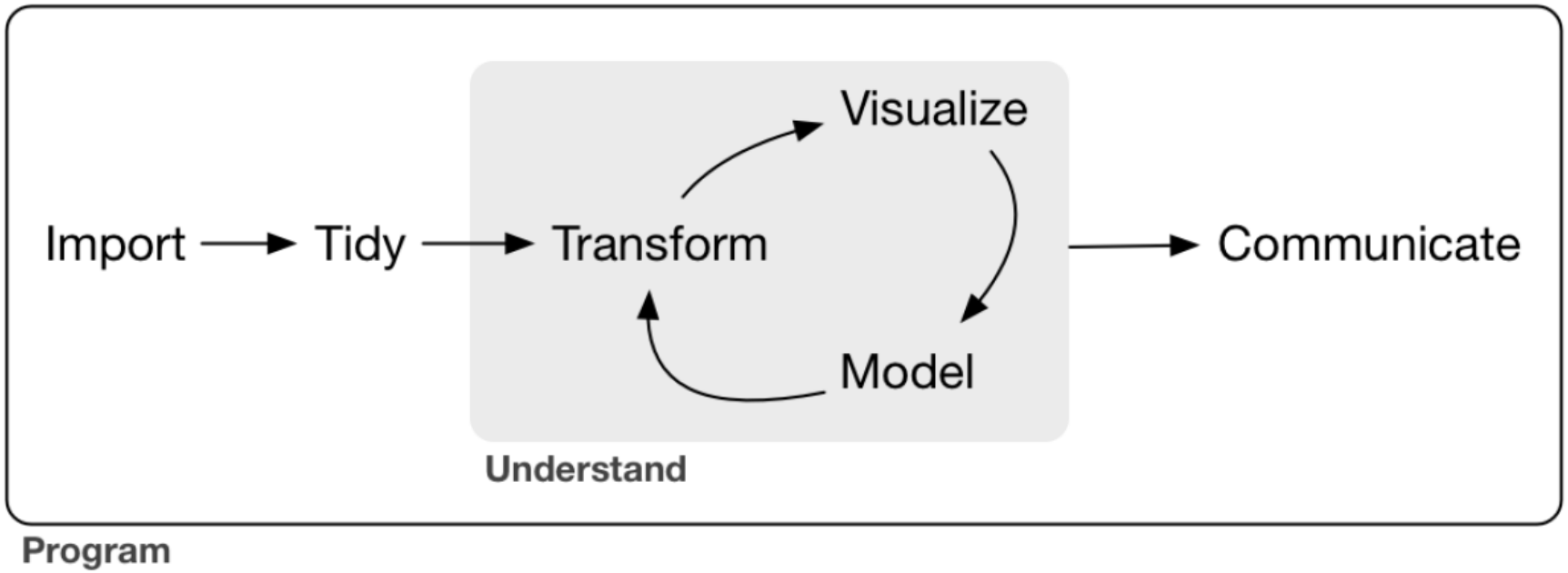
Lion Behrens, M.Sc.



UNIVERSITY OF MANNHEIM

University of Mannheim
Chair of Social Data Science and Methodology
Chair of Quantitative Methods in the Social Sciences

Data Project Flow



Note: Figure from Wickham and Grolemund, 2017. R For Data Science. Sebastopol: O' REILLY.

What is Exploratory Data Analyis?

Before you are modeling your data using methods from statistical modeling or machine learning, you typically first want to understand your data.

Exploratory data analysis typically encompasses:

- summary statistics
  - minimum and maximum values
  - measures of central tendency (mean, median, mode)
  - measures of dispersion (standard deviation, variance)
  - the shape of univarie distributions (skewness, kurtosis)
  - extent of missing values
- frequency tables, proportion tables, cross-tabulations
- correlation matrices

Note: Often, getting to know your data includes data visualization as well. Visualizing data will be covered separately in Module 4 of this course, but feel free to already jump ahead and go through the material on (exploratory) data visualization.

Prerequisite: Data Wrangling Pipeline (I/III)

```
library(tidyverse)
ess10 <- haven::read_dta("./dat/ESS10.dta")
ess10 <- ess10 %>% # subset variables
  select(country = cntry, # sociodemographics
         gender = gndr,
         education_years = eduyrs,
         trust_social = ppltrst, # multidimensional trust
         trust_parliament = trstprl,
         trust_legalSys = trstlgl,
         trust_police = trstplc,
         trust_politicians = trstplt,
         trust_parties = trstprt,
         trust_EP = trstep,
         trust_UN = trstun,
         left_right = lrscale, # attitudes
         life_satisfaction = stflife,
         pol_interest = polintr,
         voted = vote, # turnout
         party_choice = prtvtefr # party choice
         ) %>%
  mutate_at(c("country", "gender", "voted", "party_choice"), as.character) %>% # change types
  mutate_at("pol_interest", as.numeric) %>% # change types
  filter(country == "FR") # subset cases (only include France)
```

Prerequisite: Data Wrangling Pipeline (II/III)

```
ess10 <- ess10 %>%
  mutate(gender = recode_factor(gender,
                                `1` = "Male",
                                `2` = "Female"),
         voted = recode_factor(voted,
                               `1` = "Yes",
                               `2` = "No",
                               `3` = "Not eligible"),
         party_choice = recode_factor(party_choice,
                                      `1`  = "Lutte Ouvriére",
                                      `2`  = "Nouv. Parti Anti-Capitaliste",
                                      `3`  = "Parti Communiste Français",
                                      `4`  = "La France Insoumise",
                                      `5`  = "Parti Socialiste",
                                      `6`  = "Europe Ecologie Les Verts",
                                      `7`  = "La République en Marche",
                                      `8`  = "Mouvement Démocrate",
                                      `9`  = "Les Républicains",
                                      `10` = "Debout la France",
                                      `11` = "Front National",
                                      `12` = "Other",
                                      `13` = "Blank",
                                      `14` = "Null")
         )
```

Prerequisite: Data Wrangling Pipeline (III/III)

```
ess10 <- ess10 %>%
  mutate(education_years = na_if(education_years, 114), # set 114 to missing
         pol_interest = (pol_interest * -1) + 5, # invert scale
         life_satisfaction = life_satisfaction + 1 # change scale to [1, 11]
         ) %>%
  drop_na(trust_politicians, gender, education_years,
          life_satisfaction, pol_interest) # list-wise deletion of missings
```

Inspecting Labels

Inspecting Labels

Now all of our variables are named intuitively and our categorical variables are in nice and handy (factor) format. To get value labels for continuous variables, use sjlabelled::get_labels():

```
table(ess10$left_right)
```

```
##
##   0   1   2   3   4   5   6   7   8   9  10
##  71  39  90 209 159 538 162 198 145  39  63
```

```
library(sjlabelled)
get_labels(ess10$left_right)
```

```
##  [1] "Left"       "1"          "2"          "3"          "4"          "5"           "6"
##  [8] "7"          "8"          "9"          "Right"      "Refusal"    "Don't know" "No answer"
```

(Cross-)Tabulations

Frequency Tables, Table of Proportions, Crosstabulations

Very simply, we could be interested in inspecting the frequencies of certain events. For instance, let's inspect how many of our respondents have voted in the last national elections.

```
# frequency table
table(ess10$voted)
```

```
##
##          Yes           No Not eligible
##         1003          568          289
```

```
# table of relative frequencies (proportions)
prop.table(table(ess10$voted))
```

```
##
##          Yes           No Not eligible
##    0.5392473    0.3053763    0.1553763
```

Frequency Tables, Table of Proportions, Crosstabulations

What about vote choice?

```
# frequency table
table(ess10$party_choice)
```

```
##
##               Lutte Ouvriére Nouv. Parti Anti-Capitaliste    Parti Communiste Français
##                            9                            6                           16
##          La France Insoumise            Parti Socialiste    Europe Ecologie Les Verts
##                           44                          135                          123
##      La République en Marche         Mouvement Démocrate           Les Républicains
##                          220                           15                          131
##              Debout la France              Front National                       Other
##                           12                          105                           11
##                        Blank                         Null
##                           30                            5
```

Frequency Tables, Table of Proportions, Crosstabulations

What about vote choice?

```
# table of relative frequencies (proportions)
prop.table(table(ess10$party_choice))
```

```
##
##              Lutte Ouvriére Nouv. Parti Anti-Capitaliste    Parti Communiste Français
##                  0.010440835                 0.006960557                 0.018561485
##          La France Insoumise            Parti Socialiste    Europe Ecologie Les Verts
##                  0.051044084                 0.156612529                 0.142691415
##      La République en Marche          Mouvement Démocrate           Les Républicains
##                  0.255220418                 0.017401392                 0.151972158
##             Debout la France               Front National                       Other
##                  0.013921114                 0.121809745                 0.012761021
##                        Blank                         Null
##                  0.034802784                 0.005800464
```

Frequency Tables, Table of Proportions, Crosstabulations

Crosstabulations are performed by providing various variables as arguments to table().

Absolute frequencies:

```
# cross-tab between turnout and life satisfaction
table(ess10$voted, ess10$life_satisfaction)
```

```
##
##                  1   2   3   4   5   6   7   8   9  10  11
##   Yes           15   7  18  21  39  93  91 180 298 135 106
##   No            15   5  18  18  31  70  72 109 114  52  64
##   Not eligible   5   2   4  11   9  22  20  54  78  35  49
```

Frequency Tables, Table of Proportions, Crosstabulations

Crosstabulations are performed by providing various variables as arguments to table().

Relative frequencies (proportions):

```
# cross-tab between turnout and life satisfaction
crosstab <- prop.table(table(ess10$voted, ess10$life_satisfaction),
                       margin = 2 # for column percentages
                       )
```

```
# round to two digits
round(crosstab, digits = 2)
```

```
##
##                    1    2    3    4    5    6    7    8    9   10   11
##   Yes           0.43 0.50 0.45 0.42 0.49 0.50 0.50 0.52 0.61 0.61 0.48
##   No            0.43 0.36 0.45 0.36 0.39 0.38 0.39 0.32 0.23 0.23 0.29
##   Not eligible 0.14 0.14 0.10 0.22 0.11 0.12 0.11 0.16 0.16 0.16 0.22
```

Get To Know Your Data

Get a global look at your data: base R

There are several options to get a global impression of your data. This is important to spot whether there are any problems, such as mis-coding of variables or an unusual number of missing values.

```
summary(ess10)
```

```
  country            gender     education_years   trust_social    trust_parliament trust_legalSys
Length:1977        Male  : 974   Min.    :  0.00   Min.    : 0.000   Min.    : 0.000   Min.    : 0.000
Class :character   Female:1003   1st Qu.: 11.00   1st Qu.: 3.000   1st Qu.: 3.000   1st Qu.: 4.000
Mode  :character                 Median : 13.00   Median : 5.000   Median : 5.000   Median : 5.000
                                 Mean   : 13.42   Mean   : 4.687   Mean   : 4.543   Mean   : 5.211
                                 3rd Qu.: 16.00   3rd Qu.: 6.000   3rd Qu.: 6.000   3rd Qu.: 7.000
                                 Max.   :114.00   Max.   :10.000   Max.   :10.000   Max.   :10.000
                                 NA's   :44       NA's   :2        NA's   :62       NA's   :17
 trust_police    trust_politicians trust_parties      trust_EP         trust_UN        left_right
Min.    : 0.000   Min.    : 0.000   Min.    : 0.000   Min.    : 0.000   Min.    : 0.000   Min.    : 0.000
1st Qu.: 5.000   1st Qu.: 2.000   1st Qu.: 2.000   1st Qu.: 3.000   1st Qu.: 4.000   1st Qu.: 4.000
Median : 7.000   Median : 4.000   Median : 3.000   Median : 5.000   Median : 5.000   Median : 5.000
Mean   : 6.363   Mean   : 3.896   Mean   : 3.394   Mean   : 4.395   Mean   : 5.175   Mean   : 5.071
3rd Qu.: 8.000   3rd Qu.: 5.000   3rd Qu.: 5.000   3rd Qu.: 6.000   3rd Qu.: 7.000   3rd Qu.: 7.000
Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000   Max.   :10.000
NA's   :7        NA's   :26       NA's   :41       NA's   :123      NA's   :141      NA's   :223
life_satisfaction  pol_interest            voted                         party_choice
Min.    : 0.000    Min.   :1.000   Yes         :1025   La République en Marche  : 223
1st Qu.: 6.000    1st Qu.:2.000   No          : 590   Parti Socialiste         : 136
Median : 7.000    Median :3.000   Not eligible: 307   Les Républicains         : 132
Mean   : 7.023    Mean   :2.657   NA's        :  55   Europe Ecologie Les Verts: 126
3rd Qu.: 8.000    3rd Qu.:3.000                       Front National           : 107
Max.   :10.000    Max.   :4.000                       (Other)                  : 150
NA's   :6         NA's   :2                           NA's                     :1103
```

Note: summary() can also be applied to individual vectors (columns of your dataframe).

Get a global look at your data: psych

```
library(psych)
psych::describe(ess10)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
country*	1	1977	1.00	0.00	1	1.00	0.00	1	1	0	NaN	NaN	0.00
gender*	2	1977	1.51	0.50	2	1.51	0.00	1	2	1	-0.03	-2.00	0.01
education_years	3	1933	13.42	4.39	13	13.34	2.97	0	114	114	6.29	141.53	0.10
trust_social	4	1975	4.69	2.10	5	4.79	1.48	0	10	10	-0.37	-0.16	0.05
trust_parliament	5	1915	4.54	2.41	5	4.62	2.97	0	10	10	-0.23	-0.56	0.06
trust_legalSys	6	1960	5.21	2.47	5	5.35	2.97	0	10	10	-0.40	-0.51	0.06
trust_police	7	1970	6.36	2.17	7	6.54	1.48	0	10	10	-0.79	0.57	0.05
trust_politicians	8	1951	3.90	2.18	4	3.93	1.48	0	10	10	-0.09	-0.52	0.05
trust_parties	9	1936	3.39	2.09	3	3.38	2.97	0	10	10	0.03	-0.55	0.05
trust_EP	10	1854	4.39	2.44	5	4.46	2.97	0	10	10	-0.20	-0.63	0.06
trust_UN	11	1836	5.17	2.43	5	5.31	2.97	0	10	10	-0.39	-0.36	0.06
left_right	12	1754	5.07	2.25	5	5.10	1.48	0	10	10	-0.06	-0.03	0.05
life_satisfaction	13	1971	7.02	2.22	7	7.25	1.48	0	10	10	-1.00	0.95	0.05
pol_interest	14	1975	2.66	0.97	3	2.70	1.48	1	4	3	-0.28	-0.87	0.02
voted*	15	1922	1.63	0.74	1	1.53	0.00	1	3	2	0.73	-0.86	0.02
party_choice*	16	874	7.37	2.56	7	7.29	2.97	1	14	13	0.33	-0.21	0.09

Get a global look at your data: psych

```
library(psych)
psych::describe(ess10)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
country*	1	1977	1.00	0.00	1	1.00	0.00	1	1	0	NaN	NaN	0.00
gender*	2	1977	1.51	0.50	2	1.51	0.00	1	2	1	-0.03	-2.00	0.01
education_years	3	1933	13.42	4.39	13	13.34	2.97	0	114	114	6.29	141.53	0.10
trust_social	4	1975	4.69	2.10	5	4.79	1.48	0	10	10	-0.37	-0.16	0.05
trust_parliament	5	1915	4.54	2.41	5	4.62	2.97	0	10	10	-0.23	-0.56	0.06
trust_legalSys	6	1960	5.21	2.47	5	5.35	2.97	0	10	10	-0.40	-0.51	0.06
trust_police	7	1970	6.36	2.17	7	6.54	1.48	0	10	10	-0.79	0.57	0.05
trust_politicians	8	1951	3.90	2.18	4	3.93	1.48	0	10	10	-0.09	-0.52	0.05
trust_parties	9	1936	3.39	2.09	3	3.38	2.97	0	10	10	0.03	-0.55	0.05
trust_EP	10	1854	4.39	2.44	5	4.46	2.97	0	10	10	-0.20	-0.63	0.06
trust_UN	11	1836	5.17	2.43	5	5.31	2.97	0	10	10	-0.39	-0.36	0.06
left_right	12	1754	5.07	2.25	5	5.10	1.48	0	10	10	-0.06	-0.03	0.05
life_satisfaction	13	1971	7.02	2.22	7	7.25	1.48	0	10	10	-1.00	0.95	0.05
pol_interest	14	1975	2.66	0.97	3	2.70	1.48	1	4	3	-0.28	-0.87	0.02
voted*	15	1922	1.63	0.74	1	1.53	0.00	1	3	2	0.73	-0.86	0.02
party_choice*	16	874	7.37	2.56	7	7.29	2.97	1	14	13	0.33	-0.21	0.09

Get a global look at your data: psych

```
library(psych)
psych::describe(ess10)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
country*	1	1977	1.00	0.00	1	1.00	0.00	1	1	0	NaN	NaN	0.00
gender*	2	1977	1.51	0.50	2	1.51	0.00	1	2	1	-0.03	-2.00	0.01
education_years	3	1933	13.42	4.39	13	13.34	2.97	0	114	114	6.29	141.53	0.10
trust_social	4	1975	4.69	2.10	5	4.79	1.48	0	10	10	-0.37	-0.16	0.05
trust_parliament	5	1915	4.54	2.41	5	4.62	2.97	0	10	10	-0.23	-0.56	0.06
trust_legalSys	6	1960	5.21	2.47	5	5.35	2.97	0	10	10	-0.40	-0.51	0.06
trust_police	7	1970	6.36	2.17	7	6.54	1.48	0	10	10	-0.79	0.57	0.05
trust_politicians	8	1951	3.90	2.18	4	3.93	1.48	0	10	10	-0.09	-0.52	0.05
trust_parties	9	1936	3.39	2.09	3	3.38	2.97	0	10	10	0.03	-0.55	0.05
trust_EP	10	1854	4.39	2.44	5	4.46	2.97	0	10	10	-0.20	-0.63	0.06
trust_UN	11	1836	5.17	2.43	5	5.31	2.97	0	10	10	-0.39	-0.36	0.06
left_right	12	1754	5.07	2.25	5	5.10	1.48	0	10	10	-0.06	-0.03	0.05
life_satisfaction	13	1971	7.02	2.22	7	7.25	1.48	0	10	10	-1.00	0.95	0.05
pol_interest	14	1975	2.66	0.97	3	2.70	1.48	1	4	3	-0.28	-0.87	0.02
voted*	15	1922	1.63	0.74	1	1.53	0.00	1	3	2	0.73	-0.86	0.02
party_choice*	16	874	7.37	2.56	7	7.29	2.97	1	14	13	0.33	-0.21	0.09

Get a global look at your data: psych

```
library(psych)
psych::describe(ess10)
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
country*	1	1977	1.00	0.00	1	1.00	0.00	1	1	0	NaN	NaN	0.00
gender*	2	1977	1.51	0.50	2	1.51	0.00	1	2	1	-0.03	-2.00	0.01
education_years	3	1933	13.42	4.39	13	13.34	2.97	0	114	114	6.29	141.53	0.10
trust_social	4	1975	4.69	2.10	5	4.79	1.48	0	10	10	-0.37	-0.16	0.05
trust_parliament	5	1915	4.54	2.41	5	4.62	2.97	0	10	10	-0.23	-0.56	0.06
trust_legalSys	6	1960	5.21	2.47	5	5.35	2.97	0	10	10	-0.40	-0.51	0.06
trust_police	7	1970	6.36	2.17	7	6.54	1.48	0	10	10	-0.79	0.57	0.05
trust_politicians	8	1951	3.90	2.18	4	3.93	1.48	0	10	10	-0.09	-0.52	0.05
trust_parties	9	1936	3.39	2.09	3	3.38	2.97	0	10	10	0.03	-0.55	0.05
trust_EP	10	1854	4.39	2.44	5	4.46	2.97	0	10	10	-0.20	-0.63	0.06
trust_UN	11	1836	5.17	2.43	5	5.31	2.97	0	10	10	-0.39	-0.36	0.06
left_right	12	1754	5.07	2.25	5	5.10	1.48	0	10	10	-0.06	-0.03	0.05
life_satisfaction	13	1971	7.02	2.22	7	7.25	1.48	0	10	10	-1.00	0.95	0.05
pol_interest	14	1975	2.66	0.97	3	2.70	1.48	1	4	3	-0.28	-0.87	0.02
voted*	15	1922	1.63	0.74	1	1.53	0.00	1	3	2	0.73	-0.86	0.02
party_choice*	16	874	7.37	2.56	7	7.29	2.97	1	14	13	0.33	-0.21	0.09

Get a global look at your data: skimr

```
library(skimr)
skimr::skim_without_charts(ess10)
```

```
-- Data Summary ------------------------
                           Values
Name                       ess10
Number of rows             1977
Number of columns          16
_________________________
Column type frequency:
  character                1
  factor                   3
  numeric                  12
_________________________
Group variables            None

-- Variable type: character ----------------------------------------------------------
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 country               0             1   2   2     0        1          0

-- Variable type: factor -------------------------------------------------------------
  skim_variable n_missing complete_rate ordered n_unique
1 gender                0         1     FALSE          2
2 voted                55         0.972 FALSE          3
3 party_choice       1103         0.442 FALSE         14
  top_counts
1 Fem: 1003, Mal: 974
2 Yes: 1025, No: 590, Not: 307
3 La : 223, Par: 136, Les: 132, Eur: 126

-- Variable type: numeric ------------------------------------------------------------
   skim_variable     n_missing complete_rate  mean    sd p0 p25 p50 p75 p100
 1 education_years          44         0.978 13.4  4.39   0  11  13  16  114
 2 trust_social              2         0.999  4.69 2.10   0   3   5   6   10
 3 trust_parliament         62         0.969  4.54 2.41   0   3   5   6   10
 4 trust_legalSys           17         0.991  5.21 2.47   0   4   5   7   10
 5 trust_police              7         0.996  6.36 2.17   0   5   7   8   10
 6 trust_politicians        26         0.987  3.90 2.18   0   2   4   5   10
 7 trust_parties            41         0.979  3.39 2.09   0   2   3   5   10
 8 trust_EP                123         0.938  4.39 2.44   0   3   5   6   10
 9 trust_UN                141         0.929  5.17 2.43   0   4   5   7   10
10 left_right              223         0.887  5.07 2.25   0   4   5   7   10
11 life_satisfaction         6         0.997  7.02 2.22   0   6   7   8   10
12 pol_interest              2         0.999  2.66 0.965  1   2   3   3    4
```

Get a global look at your data: skimr

```
library(skimr)
skimr::skim_without_charts(ess10)
```

```
-- Data Summary ------------------------
                           Values
Name                       ess10
Number of rows             1977
Number of columns          16
_________________________
Column type frequency:
  character                1
  factor                   3
  numeric                  12
_________________________
Group variables            None

-- Variable type: character ------------------------------------------------------------
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 country               0             1   2   2     0        1          0

-- Variable type: factor ---------------------------------------------------------------
  skim_variable n_missing complete_rate ordered n_unique
1 gender                0         1     FALSE          2
2 voted                55         0.972 FALSE          3
3 party_choice       1103         0.442 FALSE         14
  top_counts
1 Fem: 1003, Mal: 974
2 Yes: 1025, No: 590, Not: 307
3 La : 223, Par: 136, Les: 132, Eur: 126

-- Variable type: numeric --------------------------------------------------------------
   skim_variable     n_missing complete_rate  mean    sd p0 p25 p50 p75 p100
 1 education_years          44         0.978 13.4  4.39   0  11  13  16  114
 2 trust_social              2         0.999  4.69 2.10   0   3   5   6   10
 3 trust_parliament         62         0.969  4.54 2.41   0   3   5   6   10
 4 trust_legalSys           17         0.991  5.21 2.47   0   4   5   7   10
 5 trust_police              7         0.996  6.36 2.17   0   5   7   8   10
 6 trust_politicians        26         0.987  3.90 2.18   0   2   4   5   10
 7 trust_parties            41         0.979  3.39 2.09   0   2   3   5   10
 8 trust_EP                123         0.938  4.39 2.44   0   3   5   6   10
 9 trust_UN                141         0.929  5.17 2.43   0   4   5   7   10
10 left_right              223         0.887  5.07 2.25   0   4   5   7   10
11 life_satisfaction         6         0.997  7.02 2.22   0   6   7   8   10
12 pol_interest              2         0.999  2.66 0.965  1   2   3   3    4
```

Get a global look at your data: skimr

```
library(skimr)
skimr::skim_without_charts(ess10)
```

```
-- Data Summary ------------------------
                           Values
Name                       ess10
Number of rows             1977
Number of columns          16
_________________________
Column type frequency:
  character                1
  factor                   3
  numeric                  12
_________________________
Group variables            None

-- Variable type: character ------------------------------------------------------
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 country               0             1   2   2     0        1          0

-- Variable type: factor ---------------------------------------------------------
  skim_variable n_missing complete_rate ordered n_unique
1 gender                0         1     FALSE          2
2 voted                55         0.972 FALSE          3
3 party_choice       1103         0.442 FALSE         14
  top_counts
1 Fem: 1003, Mal: 974
2 Yes: 1025, No: 590, Not: 307
3 La : 223, Par: 136, Les: 132, Eur: 126

-- Variable type: numeric --------------------------------------------------------
   skim_variable     n_missing complete_rate  mean    sd p0 p25 p50 p75 p100
 1 education_years          44         0.978 13.4  4.39   0  11  13  16  114
 2 trust_social              2         0.999  4.69 2.10   0   3   5   6   10
 3 trust_parliament         62         0.969  4.54 2.41   0   3   5   6   10
 4 trust_legalSys           17         0.991  5.21 2.47   0   4   5   7   10
 5 trust_police              7         0.996  6.36 2.17   0   5   7   8   10
 6 trust_politicians        26         0.987  3.90 2.18   0   2   4   5   10
 7 trust_parties            41         0.979  3.39 2.09   0   2   3   5   10
 8 trust_EP                123         0.938  4.39 2.44   0   3   5   6   10
 9 trust_UN                141         0.929  5.17 2.43   0   4   5   7   10
10 left_right              223         0.887  5.07 2.25   0   4   5   7   10
11 life_satisfaction         6         0.997  7.02 2.22   0   6   7   8   10
12 pol_interest              2         0.999  2.66 0.965  1   2   3   3    4
```

Get a global look at your data: skimr

```
library(skimr)
skimr::skim_without_charts(ess10)
```

```
-- Data Summary ------------------------
                           Values
Name                       ess10
Number of rows             1977
Number of columns          16
_________________________
Column type frequency:
  character                1
  factor                   3
  numeric                  12
_________________________
Group variables            None

-- Variable type: character ------------------------------------------------------
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 country               0             1   2   2     0        1          0

-- Variable type: factor ---------------------------------------------------------
  skim_variable n_missing complete_rate ordered n_unique
1 gender                0         1     FALSE          2
2 voted                55         0.972 FALSE          3
3 party_choice       1103         0.442 FALSE         14
  top_counts
1 Fem: 1003, Mal: 974
2 Yes: 1025, No: 590, Not: 307
3 La : 223, Par: 136, Les: 132, Eur: 126

-- Variable type: numeric --------------------------------------------------------
   skim_variable     n_missing complete_rate  mean    sd p0 p25 p50 p75 p100
 1 education_years          44         0.978 13.4  4.39   0  11  13  16  114
 2 trust_social              2         0.999  4.69 2.10   0   3   5   6   10
 3 trust_parliament         62         0.969  4.54 2.41   0   3   5   6   10
 4 trust_legalSys           17         0.991  5.21 2.47   0   4   5   7   10
 5 trust_police              7         0.996  6.36 2.17   0   5   7   8   10
 6 trust_politicians        26         0.987  3.90 2.18   0   2   4   5   10
 7 trust_parties            41         0.979  3.39 2.09   0   2   3   5   10
 8 trust_EP                123         0.938  4.39 2.44   0   3   5   6   10
 9 trust_UN                141         0.929  5.17 2.43   0   4   5   7   10
10 left_right              223         0.887  5.07 2.25   0   4   5   7   10
11 life_satisfaction         6         0.997  7.02 2.22   0   6   7   8   10
12 pol_interest              2         0.999  2.66 0.965  1   2   3   3    4
```

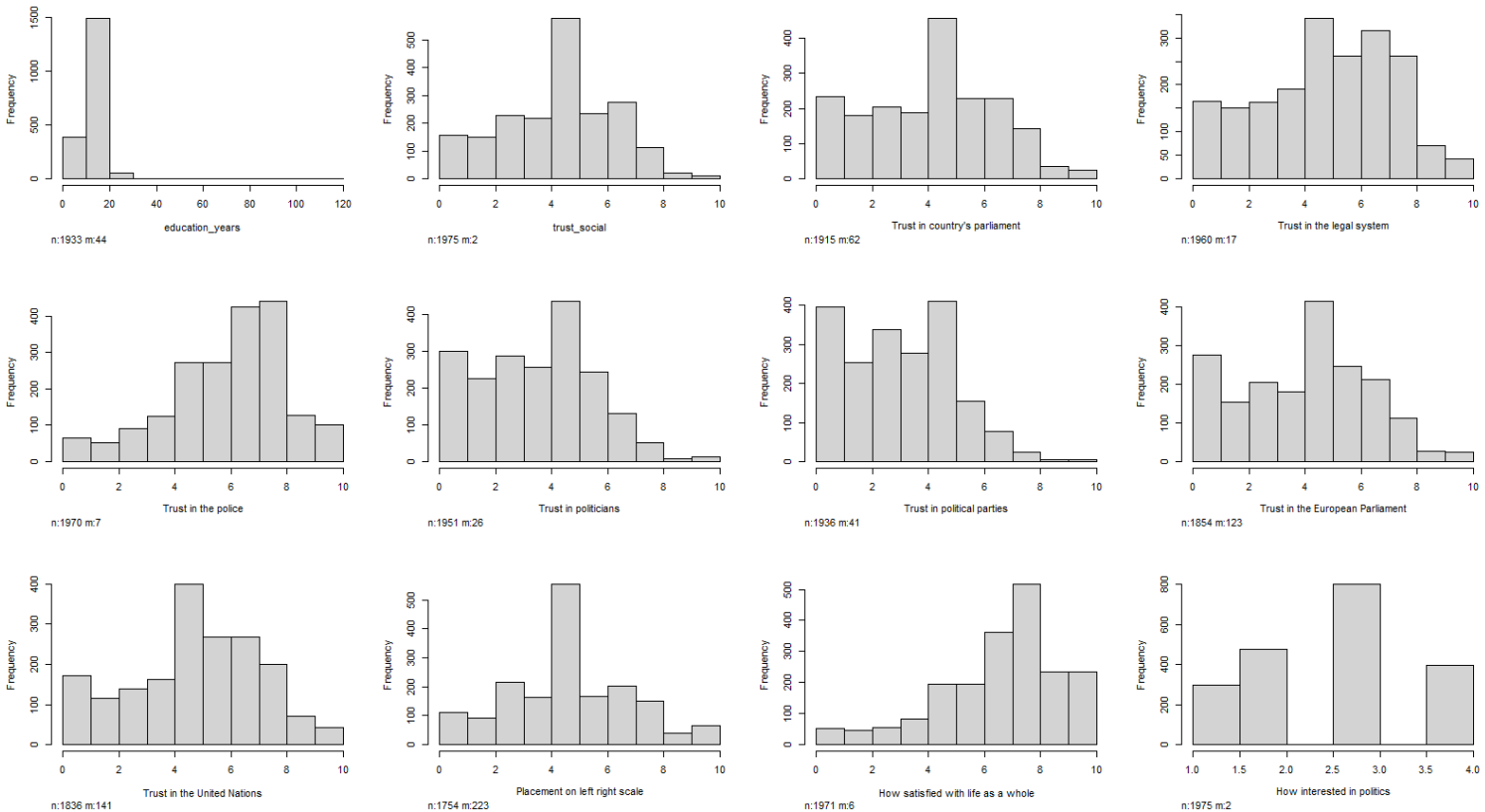
Get a global look at your data: skimr

```
library(skimr)
skimr::skim_without_charts(ess10)
```

```
-- Data Summary ------------------------
                           Values
Name                       ess10
Number of rows             1977
Number of columns          16
_________________________
Column type frequency:
  character                1
  factor                   3
  numeric                  12
_________________________
Group variables            None

-- Variable type: character ------------------------------------------------------
  skim_variable n_missing complete_rate min max empty n_unique whitespace
1 country               0             1   2   2     0        1          0

-- Variable type: factor ---------------------------------------------------------
  skim_variable n_missing complete_rate ordered n_unique
1 gender                0         1     FALSE          2
2 voted                55         0.972 FALSE          3
3 party_choice       1103         0.442 FALSE         14
  top_counts
1 Fem: 1003, Mal: 974
2 Yes: 1025, No: 590, Not: 307
3 La : 223, Par: 136, Les: 132, Eur: 126

-- Variable type: numeric --------------------------------------------------------
   skim_variable     n_missing complete_rate  mean    sd p0 p25 p50 p75 p100
 1 education_years          44         0.978 13.4  4.39   0  11  13  16  114
 2 trust_social              2         0.999  4.69 2.10   0   3   5   6   10
 3 trust_parliament         62         0.969  4.54 2.41   0   3   5   6   10
 4 trust_legalSys           17         0.991  5.21 2.47   0   4   5   7   10
 5 trust_police              7         0.996  6.36 2.17   0   5   7   8   10
 6 trust_politicians        26         0.987  3.90 2.18   0   2   4   5   10
 7 trust_parties            41         0.979  3.39 2.09   0   2   3   5   10
 8 trust_EP                123         0.938  4.39 2.44   0   3   5   6   10
 9 trust_UN                141         0.929  5.17 2.43   0   4   5   7   10
10 left_right              223         0.887  5.07 2.25   0   4   5   7   10
11 life_satisfaction         6         0.997  7.02 2.22   0   6   7   8   10
12 pol_interest              2         0.999  2.66 0.965  1   2   3   3    4
```

(Quick) Visual Inspection of Distributions

```
library(Hmisc)
Hmisc::hist.data.frame(ess10[,sapply(ess10, is.numeric)], # only apply to numeric variables
                       nclass = 10)
```



Describing Different Groups

Summary Statistics by Group

Let's say we are interested in the summary statistics of some variables by group rather than globally on all observations. We can make use of the group_by() function that we already got to know in Module 2.

The code chunk below explores trust in politicians by party choice in France.

```
library(dplyr)
ss_byparty <- ess10 %>%
  group_by(party_choice) %>%
  filter(!is.na(trust_politicians)) %>% # exclude missing values
  dplyr::summarize(n = length(trust_politicians), # number of observations
                   min = min(trust_politicians), # minimum value
                   q1 = quantile(trust_politicians, 0.25), # 25% quantile
                   median = median(trust_politicians), # median
                   mean = mean(trust_politicians), # arithmetic mean
                   q3 = quantile(trust_politicians, 0.75), # 75% quantile
                   max = max(trust_politicians), # maximum value
                   sd = sd(trust_politicians) # standard deviation
                   )
```

Summary Statistics by Group

Let's say we are interested in the summary statistics of some variables by group rather than globally on all observations. We can make use of the group_by() function that we already got to know in Module 2.

The code chunk below explores trust in politicians by party choice in France.

```
library(dplyr)
ss_byparty <- ess10 %>%
  group_by(party_choice) %>%
  filter(!is.na(trust_politicians)) %>% # exclude missing values
  dplyr::summarize(n = length(trust_politicians), # number of observations
                   min = min(trust_politicians), # minimum value
                   q1 = quantile(trust_politicians, 0.25), # 25% quantile
                   median = median(trust_politicians), # median
                   mean = mean(trust_politicians), # arithmetic mean
                   q3 = quantile(trust_politicians, 0.75), # 75% quantile
                   max = max(trust_politicians), # maximum value
                   sd = sd(trust_politicians) # standard deviation
                   )
```

Summary Statistics by Group

Let's inspect the levels of trust that different party supporters place in politicians.

Rows are ordered by their mean level of trust.

```
print(ss_byparty[order(ss_byparty$mean),])
```

Summary Statistics by Group

Let's inspect the levels of trust that different party supporters place in politicians.

Rows are ordered by their mean level of trust.

```
print(ss_byparty[order(ss_byparty$mean),])
```

```
##                    party_choice    n min   q1 median     mean   q3 max       sd
## 1                Lutte Ouvriére   11   0 0.00    2.0 2.454545 3.50   8 2.910795
## 13                        Blank   30   0 1.00    2.0 2.533333 3.75   6 1.888866
## 12                        Other   11   0 2.00    3.0 2.818182 3.50   5 1.328020
## 11               Front National  107   0 1.00    3.0 2.981308 5.00  10 2.306600
## 2  Nouv. Parti Anti-Capitaliste    6   0 1.50    3.5 3.500000 5.50   7 2.738613
## 14                         Null    5   0 0.00    5.0 3.600000 6.00   7 3.361547
## 4            La France Insoumise   44   0 2.00    3.0 3.636364 5.00   7 1.805559
## 15                         <NA> 1077   0 2.00    4.0 3.701021 5.00  10 2.263688
## 6     Europe Ecologie Les Verts  126   0 3.00    4.0 3.984127 5.00   8 1.824211
## 5               Parti Socialiste  136   0 3.00    4.0 4.154412 5.25   8 1.924034
## 10              Debout la France   12   2 2.75    4.5 4.333333 6.00   7 1.874874
## 3     Parti Communiste Français   16   2 3.00    5.0 4.375000 5.25   7 1.543805
## 9               Les Républicains  132   0 4.00    5.0 4.636364 6.00   9 1.887171
## 7       La République en Marche  223   0 4.00    5.0 4.856502 6.00  10 1.812628
## 8            Mouvement Démocrate   15   0 4.00    6.0 5.333333 7.00   8 2.093072
```

Summary Statistics by Group

Let's inspect the levels of trust that different party supporters place in politicians.

Rows are ordered by their mean level of trust.

```
print(ss_byparty[order(ss_byparty$mean),])
```

```
##                    party_choice    n min   q1 median     mean   q3 max       sd
## 1                 Lutte Ouvriére   11   0 0.00    2.0 2.454545 3.50   8 2.910795
## 13                         Blank   30   0 1.00    2.0 2.533333 3.75   6 1.888866
## 12                         Other   11   0 2.00    3.0 2.818182 3.50   5 1.328020
## 11                Front National  107   0 1.00    3.0 2.981308 5.00  10 2.306600
## 2  Nouv. Parti Anti-Capitaliste    6   0 1.50    3.5 3.500000 5.50   7 2.738613
## 14                          Null    5   0 0.00    5.0 3.600000 6.00   7 3.361547
## 4           La France Insoumise   44   0 2.00    3.0 3.636364 5.00   7 1.805559
## 15                         <NA> 1077   0 2.00    4.0 3.701021 5.00  10 2.263688
## 6     Europe Ecologie Les Verts  126   0 3.00    4.0 3.984127 5.00   8 1.824211
## 5               Parti Socialiste  136   0 3.00    4.0 4.154412 5.25   8 1.924034
## 10              Debout la France   12   2 2.75    4.5 4.333333 6.00   7 1.874874
## 3     Parti Communiste Français   16   2 3.00    5.0 4.375000 5.25   7 1.543805
## 9               Les Républicains  132   0 4.00    5.0 4.636364 6.00   9 1.887171
## 7       La République en Marche  223   0 4.00    5.0 4.856502 6.00  10 1.812628
## 8           Mouvement Démocrate   15   0 4.00    6.0 5.333333 7.00   8 2.093072
```

Summary Statistics by Group

Let's inspect the levels of trust that different party supporters place in politicians.

Rows are ordered by their mean level of trust.

```
print(ss_byparty[order(ss_byparty$mean),])
```

```
##                    party_choice    n min   q1 median     mean   q3 max       sd
## 1                Lutte Ouvriére   11   0 0.00    2.0 2.454545 3.50   8 2.910795
## 13                        Blank   30   0 1.00    2.0 2.533333 3.75   6 1.888866
## 12                        Other   11   0 2.00    3.0 2.818182 3.50   5 1.328020
## 11               Front National  107   0 1.00    3.0 2.981308 5.00  10 2.306600
## 2  Nouv. Parti Anti-Capitaliste    6   0 1.50    3.5 3.500000 5.50   7 2.738613
## 14                        Null    5   0 0.00    5.0 3.600000 6.00   7 3.361547
## 4           La France Insoumise   44   0 2.00    3.0 3.636364 5.00   7 1.805559
## 15                         <NA> 1077   0 2.00    4.0 3.701021 5.00  10 2.263688
## 6     Europe Ecologie Les Verts  126   0 3.00    4.0 3.984127 5.00   8 1.824211
## 5              Parti Socialiste  136   0 3.00    4.0 4.154412 5.25   8 1.924034
## 10             Debout la France   12   2 2.75    4.5 4.333333 6.00   7 1.874874
## 3     Parti Communiste Français   16   2 3.00    5.0 4.375000 5.25   7 1.543805
## 9              Les Républicains  132   0 4.00    5.0 4.636364 6.00   9 1.887171
## 7       La République en Marche  223   0 4.00    5.0 4.856502 6.00  10 1.812628
## 8           Mouvement Démocrate   15   0 4.00    6.0 5.333333 7.00   8 2.093072
```

Correlations

Relations Between Variables: Correlations

To get an impression of how different variables are related to each other, we can compute bivariate correlations and store them in a matrix.

Simple correlation matrix:

```
cor(ess10[, c("education_years", "life_satisfaction", "left_right", "trust_politicians")],
    method = "pearson",  # calculate Pearson's correlation coefficient
    use = "complete.obs" # list-wise deletion
    )
```

```
##                   education_years life_satisfaction  left_right trust_politicians
## education_years        1.00000000         0.1423890 -0.09307594        0.00520711
## life_satisfaction      0.14238902         1.0000000  0.13817270        0.24775920
## left_right            -0.09307594         0.1381727  1.00000000        0.05051487
## trust_politicians      0.00520711         0.2477592  0.05051487        1.00000000
```

Relations Between Variables: Correlations

To get an impression of how different variables are related to each other, we can compute bivariate correlations and store them in a matrix.

Rounded to two digits:

```
round(cor(ess10[, c("education_years", "life_satisfaction", "left_right", "trust_politicians")],
    method = "pearson",  # calculate Pearson's correlation coefficient
    use = "complete.obs" # list-wise deletion
    ), digits=2)
```

```
##                   education_years life_satisfaction left_right trust_politicians
## education_years              1.00              0.14      -0.09              0.01
## life_satisfaction            0.14              1.00       0.14              0.25
## left_right                  -0.09              0.14       1.00              0.05
## trust_politicians            0.01              0.25       0.05              1.00
```

Relations Between Variables: Correlations

To get an impression of how different variables are related to each other, we can compute bivariate correlations and store them in a matrix.

If you want to get p-values for your correlation coefficients, try out the Hmisc package again.

First, the correlation coefficients again:

```
library(Hmisc)
cormatrix <- Hmisc::rcorr(as.matrix(ess10[, c("education_years", "life_satisfaction", "left_right",
cormatrix$r
```

```
##                   education_years life_satisfaction  left_right trust_politicians
## education_years        1.00000000         0.1399378 -0.09307594        0.02181401
## life_satisfaction      0.13993784         1.0000000  0.13817270        0.26954998
## left_right            -0.09307594         0.1381727  1.00000000        0.05051487
## trust_politicians      0.02181401         0.2695500  0.05051487        1.00000000
```

Relations Between Variables: Correlations

To get an impression of how different variables are related to each other, we can compute bivariate correlations and store them in a matrix.

If you want to get p-values for your correlation coefficients, try out the Hmisc package again.

Afterwards, the p-values:

```
cormatrix$P
```

```
##                   education_years life_satisfaction   left_right trust_politicians
## education_years                NA      8.399463e-10 1.143984e-04        0.34105177
## life_satisfaction    8.399463e-10                NA 9.348994e-09        0.00000000
## left_right           1.143984e-04      9.348994e-09           NA        0.03656945
## trust_politicians    3.410518e-01      0.000000e+00 3.656945e-02                NA
```

Overview of Packages for Exploratory Data Analysis

Covered in this material

- dplyr using group_by()
- sjlabelled
- psych
- skimr
- Hmisc

Not covered in this material

- summarytools
- corrr
- correlation

and many more...

References

Parts of this course are inspired by the following resources:

- Wickham, Hadley and Garrett Grolemund, 2017. *R for Data Science - Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly.
- Bahnsen, Oke and Guido Ropers, 2022. *Introduction to R for Quantitative Social Science*. Course held as part of the GESIS Workshop Series.
- Breuer, Johannes and Stefan Jünger, 2021. *Introduction to R for Data Analysis*. Course held as part of the GESIS Summer School in Survey Methodology.
- Teaching material developed by Verena Kunz, David Weyrauch, Oliver Rittmann and Viktoriia Semenova.