Der ALLBUS Datensatz

Cornelius Puschmann

Table of contents

Die Allgemeine Bevölkerungsumfrage der Sozialwissenschaften (ALLBUS) ist eine standardisierte Befragung der deutschen Bevölkerung, die seit 1980 regelmäßig durch das GESIS Leibniz Institut für Sozialforschung durchgeführt wird. Im ALLBUS werden in der Regel alle zwei Jahre Daten über Einstellungen, Verhaltensweisen und Sozialstruktur der Bevölkerung in der Bundesrepublik Deutschland gesammelt. Dafür wird in persönlichen Interviews jeweils eine repräsentative Stichprobe aus der Bevölkerung Deutschlands befragt (jeweils ca. 2.800 bis 3.500 Befragte).

Abgefragt werden u.a. Einschätzungen und Einstellungen in den Bereichen:

- Wirtschaft
- Umwelt
- Immigration
- Politische Einstellungen und Partizipation
- Mediennutzung
- Einstellungen zu Ehe, Familie und Partnerschaft
- Einstellungen zu und Kontakte mit Behörden

Figure 1: Wir haben unsere Daten, Bild generiert von Midjourney

- Freizeitaktivitäten

- Gesundheit und gesundheitsrelevantes Verhalten

Folgend arbeiten wir im Rahmen dieses Moduls durchgängig mit dem ALLBUS, speziell mit der Erhebnungswelle des Jahres 2021.

1 Data Management & Einlesen des ALLBUS

Wir beginnen damit, die notwendigen Pakete zu laden, die wir für die ersten Schritte mit den Daten benötigen. Das sind hier die Pakete haven (für das Einlesen der ALLBUS-Daten im Stata-Format) und das Pakete readr (für das Einlesen einiger Vorab vorbereiteter Samples aus dem Gesamtdatensatz), sowie das Paket dplyr, mit dem wir am Schluss einen Beispielhaften Teildatensatz bilden.

```
if(!require("pacman")) {install.packages("pacman");library(pacman)}
p_load(haven, readr, dplyr)

options(scipen = 999)
```

Nun lesen wir den ALLBUS-Datensatz mittels der Stata-Importfunktion `read_dta` aus dem Paket `haven` ein.

```
daten <- read_dta("Datensatz/Allbus_2021.dta")
```

Als nächstes laden wir zudem noch drei zuvor erstellte Zufallssamples im Umfang von 20, 200 und 500 Zeilen aus dem Gesamtdatensatz. Diese enthalten weiterhin eine deutlich kleinere Anzahl relevanter Variablen und sind daher etwas übersichtlicher als der sehr große Hauptdatensatz.

```
sample_klein <- read_rds("Datensatz/ALLBUS_sample_klein.rds")
sample_mittel <- read_rds("Datensatz/ALLBUS_sample_mittel.rds")
sample_gross <- read_rds("Datensatz/ALLBUS_sample_gross.rds")
```

2 Erste Schritte mit dem ALLBUS

Zunächst schauen wir uns die Daten an. Dies geschieht entweder daturch, dass man den Objektnamen verwendet (also im folgende Beispiel einfach sample_klein) oder indem man mit einem Klick in RStudio unter dem Reiter Environment oder mit dem Befehl `View()` aufruft. Bei diesem Vorgehen zeigt RStudio die Daten an, was i.d.R. den praktischsten Zugang darstellt.

```
sample_klein
```

```
# A tibble: 20 x 4
   alter geschlecht bildung            fernsehkonsum
   <dbl> <fct>      <fct>                      <dbl>
 1    37 MANN       FACHHOCHSCHULREIFE             3
 2    38 MANN       MITTLERE REIFE                 7
 3    47 FRAU       VOLKS-,HAUPTSCHULE             6
 4    66 MANN       HOCHSCHULREIFE                 2
 5    47 FRAU       HOCHSCHULREIFE                 7
 6    75 MANN       VOLKS-,HAUPTSCHULE             7
 7    41 FRAU       MITTLERE REIFE                 4
 8    18 MANN       NOCH SCHUELER                  7
 9    91 MANN       <NA>                          NA
10    56 MANN       HOCHSCHULREIFE                 4
11    58 MANN       HOCHSCHULREIFE                 7
12    32 MANN       MITTLERE REIFE                 2
13    47 MANN       FACHHOCHSCHULREIFE             1
```

```
14    49 MANN       MITTLERE REIFE                1
15    23 MANN       MITTLERE REIFE                0
16    48 FRAU       MITTLERE REIFE                0
17    49 MANN       HOCHSCHULREIFE                0
18    36 MANN       FACHHOCHSCHULREIFE            7
19    70 MANN       HOCHSCHULREIFE                7
20    43 FRAU       MITTLERE REIFE                7
```

Der kleine Beispieldatensatz illustriert den grundlegeenden Aufbau des ALLBUS. Dieser folgt (beim ALLBUS, aber auch den meisten anderen Befragungen) den folgenden Prinzipien

- jede Befragungswelle ist ein einzeles Data Frame-Objekt (= eine große Tabelle)
- die Zeilen der Tabelle sind die Beobachtungen (= RespondentInnen)
- die Spalten der Tabelle sind die Variablen (= Antworten auf Survey-Fragen, oder bei der Möglichkeit zur Mehrfachnennung, die einzelnen Antwort*optionen*)
- die Zelleninhalte sind i.d.R. (Dummy)Zahlenwerte (= etwa "1" für wenig Zustimmung und "5" für hohe Zustimmung)

3 Einen Überblick über den ALLBUS gewinnen

Im Beispieldatensatz sind die Werte der Variablen *alter*, *geschlecht* und *bildung* recht leicht nachvollziebar, wobei sie etwas unterschiedliche Datentypen aufweisen, wie man mit Hilfe der Funktion `str` ermitteln kann.

```
str(sample_klein)
```

```
tibble [20 x 4] (S3: tbl_df/tbl/data.frame)
 $ alter        : num [1:20] 37 38 47 66 47 75 41 18 91 56 ...
 $ geschlecht   : Factor w/ 4 levels "KEINE ANGABE",..: 2 2 3 2 3 2 3 2 2 2 ...
 $ bildung      : Factor w/ 9 levels "NICHT BESTIMMBAR",..: 6 5 4 7 7 4 5 9 NA 7 ...
 $ fernsehkonsum: num [1:20] 3 7 6 2 7 7 4 7 NA 4 ...
```

Bei der Variable *alter* handelt es sich schlicht um eine Zahl (num), während *geschlecht* und *bildung* sogenannte Faktoren sind. Faktoren nutzt man in R, um wiederholende nicht numerische (typischerweise nominal oder ordinalskalierte) Werte zu speichern. Praktisch jeder Faktor könnte genausogut eine Zeichenkette (chr) sein, aber oftmals sind Faktoren praktischer, weil sie eine festen Reihenfolge haben können ("ranked factors"), die sich bei Bedarf auch in Zahlen umwandel lassen. Im konkreten Beispiel ist das Geschlecht ein ungerankter Faktor, der Bildungsgrad hat hingegen einen Rang. Der Fernsehkonsum ist schließlich eine Likert-skalierte

Variable, die wir hier und auch an anderer Stelle als metrische Variable behandeln (und dafür den R-Datentypen numeric verwenden), auch wenn das strikt genommen nicht immer zulässig ist – zumindest dann nicht, wenn man nur auf Grundlage eines einzelnen Items misst und eine 5- oder 7-Punkt Skala verwendet (vgl etwa hier).

Wie sehen die anderen Samples aus? Wir sehen uns im nächsten Schritt das große Zufallssample (500 Fälle) an.

```
sample_gross
```

```
# A tibble: 500 x 29
   alter geschlecht fernsehkonsum politisches_interesse links_rechts_einordnung
   <dbl> <fct>              <dbl> <fct>                                    <dbl>
 1    50 FRAU                   7 MITTEL                                       5
 2    77 MANN                   7 STARK                                        3
 3    64 FRAU                   7 <NA>                                         5
 4    53 MANN                  NA UEBERHAUPT NICHT                             8
 5    29 MANN                   0 SEHR STARK                                   2
 6    44 FRAU                   3 MITTEL                                       7
 7    79 MANN                   5 MITTEL                                       4
 8    32 MANN                   7 MITTEL                                       5
 9    36 FRAU                   0 STARK                                        6
10    66 MANN                   7 STARK                                        3
# i 490 more rows
# i 24 more variables: wahlabsicht_partei <fct>,
#   zufriedenheit_demokratie <fct>, entwicklung_kriminalitaet <fct>,
#   social_media_nachrichtenquelle <dbl>, glaubwuerdigkeit_oer_tv <fct>,
#   glaubwuerdigkeit_privat_tv <fct>, glaubwuerdigkeit_zeitungen <fct>,
#   glaubwuerdigkeit_social_media <fct>, vertrauen_mitmenschen <dbl>,
#   vertrauen_gesundheitswesen <dbl>, ...
```

Da wir es jetzt mit einer größeren Zahl an Beobachtungen and Variablen zu tun haben kann es nützlich sein, sich einen Überblick zu verschaffen.

Zunächst lassen wir uns die Variablennamen (also die Spalten) ausgeben. Dies geschieht mit der Funktion `colnames`.

```
colnames(sample_gross)
```

```
 [1] "alter"                          "geschlecht"
 [3] "fernsehkonsum"                  "politisches_interesse"
 [5] "links_rechts_einordnung"        "wahlabsicht_partei"
```

```
 [7] "zufriedenheit_demokratie"           "entwicklung_kriminalitaet"
 [9] "social_media_nachrichtenquelle"     "glaubwuerdigkeit_oer_tv"
[11] "glaubwuerdigkeit_privat_tv"         "glaubwuerdigkeit_zeitungen"
[13] "glaubwuerdigkeit_social_media"      "vertrauen_mitmenschen"
[15] "vertrauen_gesundheitswesen"         "vertrauen_bundesverfassungsgericht"
[17] "vertrauen_bundestag"                "vertrauen_stadt_gemeindeverwaltung"
[19] "vertrauen_katholische_kirche"       "vertrauen_evangelische_kirche"
[21] "vertrauen_justiz"                   "vertrauen_fernsehen"
[23] "vertrauen_zeitungswesen"            "vertrauen_hochschulen"
[25] "vertrauen_bundesregierung"          "vertrauen_polizei"
[27] "vertrauen_parteien"                 "vertrauen_eu_kommission"
[29] "vertrauen_eu_parlament"
```

Eine etwas detailliertere Beschreibung erhalten wir durch die Funktion `str`. Diese liefert uns auch die Dimensionen (Anzahl der Zeilen und Spalten) des Data Frames, sowie die Variablen, deren Datentyp und die ersten zehn Ausprägungen.

```
str(sample_gross)
```

```
tibble [500 x 29] (S3: tbl_df/tbl/data.frame)
 $ alter                               : num [1:500] 50 77 64 53 29 44 79 32 36 66 ...
 $ geschlecht                          : Factor w/ 4 levels "KEINE ANGABE",..: 3 2 3 2 2 3 2 2
 $ fernsehkonsum                       : num [1:500] 7 7 7 NA 0 3 5 7 0 7 ...
 $ politisches_interesse               : Factor w/ 7 levels "DATENFEHLER: MFN",..: 5 4 NA 7 3 5
 $ links_rechts_einordnung             : num [1:500] 5 3 5 8 2 7 4 5 6 3 ...
 $ wahlabsicht_partei                  : Factor w/ 13 levels "NICHT WAHLBERECHTIGT",..: NA 9 NA
 $ zufriedenheit_demokratie            : Factor w/ 10 levels "DATENFEHLER: MFN",..: 5 6 7 8 6 5
 $ entwicklung_kriminalitaet           : Factor w/ 8 levels "DATENFEHLER: MFN",..: 6 5 4 4 6 5
 $ social_media_nachrichtenquelle      : num [1:500] 0 7 0 7 7 6 0 7 NA 0 ...
 $ glaubwuerdigkeit_oer_tv             : Factor w/ 7 levels "DATENFEHLER: MFN",..: 4 4 5 5 4 NA
 $ glaubwuerdigkeit_privat_tv          : Factor w/ 7 levels "DATENFEHLER: MFN",..: 5 NA 5 5 5 N
 $ glaubwuerdigkeit_zeitungen          : Factor w/ 7 levels "DATENFEHLER: MFN",..: 5 4 5 5 5 NA
 $ glaubwuerdigkeit_social_media       : Factor w/ 7 levels "DATENFEHLER: MFN",..: 6 5 NA 5 6 4
 $ vertrauen_mitmenschen               : num [1:500] 3 3 3 2 3 3 2 NA 3 3 ...
 $ vertrauen_gesundheitswesen          : num [1:500] 6 6 5 7 6 5 5 NA 5 5 ...
 $ vertrauen_bundesverfassungsgericht: num [1:500] 6 6 4 1 6 6 2 NA 5 5 ...
 $ vertrauen_bundestag                 : num [1:500] 5 5 5 1 2 7 1 NA 5 4 ...
 $ vertrauen_stadt_gemeindeverwaltung: num [1:500] 5 5 5 1 5 6 4 NA 5 3 ...
 $ vertrauen_katholische_kirche        : num [1:500] 2 1 1 1 2 5 1 NA 2 1 ...
 $ vertrauen_evangelische_kirche       : num [1:500] 3 3 1 1 4 2 1 NA 2 2 ...
 $ vertrauen_justiz                    : num [1:500] 5 3 3 4 4 1 3 NA 4 4 ...
 $ vertrauen_fernsehen                 : num [1:500] 4 6 3 3 3 6 1 NA 4 4 ...
```

```
 $ vertrauen_zeitungswesen           : num [1:500] 4 6 3 3 3 6 1 NA 4 4 ...
 $ vertrauen_hochschulen             : num [1:500] 5 6 5 1 6 7 3 NA 5 5 ...
 $ vertrauen_bundesregierung         : num [1:500] 5 5 5 1 3 7 1 NA 5 4 ...
 $ vertrauen_polizei                 : num [1:500] 5 5 4 5 4 7 3 NA 5 4 ...
 $ vertrauen_parteien                : num [1:500] 4 4 3 1 3 4 1 NA 2 3 ...
 $ vertrauen_eu_kommission           : num [1:500] 5 5 3 1 3 5 3 NA 4 4 ...
 $ vertrauen_eu_parlament            : num [1:500] 5 5 3 1 3 6 3 NA 4 4 ...
```

Eine alternative (aber etwas ordentlichere) Ansicht erhält man mit dem Befehl `glimpse` aus dem Paket `tibble` (im tidyverse enthalten).

```
tibble::glimpse(sample_gross)
```

```
Rows: 500
Columns: 29
$ alter                                <dbl> 50, 77, 64, 53, 29, 44, 79, 32, 36,~
$ geschlecht                           <fct> FRAU, MANN, FRAU, MANN, MANN, FRAU,~
$ fernsehkonsum                        <dbl> 7, 7, 7, NA, 0, 3, 5, 7, 0, 7, 7, 7~
$ politisches_interesse                <fct> MITTEL, STARK, NA, UEBERHAUPT NICHT~
$ links_rechts_einordnung              <dbl> 5, 3, 5, 8, 2, 7, 4, 5, 6, 3, 4, 7,~
$ wahlabsicht_partei                   <fct> NA, DIE GRUENEN, NA, WUERDE NICHT W~
$ zufriedenheit_demokratie             <fct> SEHR ZUFRIEDEN, ZIEMLICH ZUFRIEDEN,~
$ entwicklung_kriminalitaet            <fct> IST GLEICH GEBLIEBEN, HAT ETWAS ZUG~
$ social_media_nachrichtenquelle       <dbl> 0, 7, 0, 7, 7, 6, 0, 7, NA, 0, 0, N~
$ glaubwuerdigkeit_oer_tv              <fct> SEHR GLAUBWUERDIG, SEHR GLAUBWUERDI~
$ glaubwuerdigkeit_privat_tv           <fct> EHER GLAUBWUERDIG, NA, EHER GLAUBWU~
$ glaubwuerdigkeit_zeitungen           <fct> EHER GLAUBWUERDIG, SEHR GLAUBWUERDI~
$ glaubwuerdigkeit_social_media        <fct> EHER N. GLAUBWUERDIG, EHER GLAUBWUE~
$ vertrauen_mitmenschen                <dbl> 3, 3, 3, 2, 3, 3, 2, NA, 3, 3, 1, 3~
$ vertrauen_gesundheitswesen           <dbl> 6, 6, 5, 7, 6, 5, 5, NA, 5, 5, 3, N~
$ vertrauen_bundesverfassungsgericht   <dbl> 6, 6, 4, 1, 6, 6, 2, NA, 5, 5, 7, N~
$ vertrauen_bundestag                  <dbl> 5, 5, 5, 1, 2, 7, 1, NA, 5, 4, 4, N~
$ vertrauen_stadt_gemeindeverwaltung   <dbl> 5, 5, 5, 1, 5, 6, 4, NA, 5, 3, 5, N~
$ vertrauen_katholische_kirche         <dbl> 2, 1, 1, 1, 2, 5, 1, NA, 2, 1, 1, N~
$ vertrauen_evangelische_kirche        <dbl> 3, 3, 1, 1, 4, 2, 1, NA, 2, 2, 4, N~
$ vertrauen_justiz                     <dbl> 5, 3, 3, 4, 4, 1, 3, NA, 4, 4, 5, N~
$ vertrauen_fernsehen                  <dbl> 4, 6, 3, 3, 3, 6, 1, NA, 4, 4, 5, N~
$ vertrauen_zeitungswesen              <dbl> 4, 6, 3, 3, 3, 6, 1, NA, 4, 4, 6, N~
$ vertrauen_hochschulen                <dbl> 5, 6, 5, 1, 6, 7, 3, NA, 5, 5, 6, N~
$ vertrauen_bundesregierung            <dbl> 5, 5, 5, 1, 3, 7, 1, NA, 5, 4, 2, N~
$ vertrauen_polizei                    <dbl> 5, 5, 4, 5, 4, 7, 3, NA, 5, 4, 6, N~
$ vertrauen_parteien                   <dbl> 4, 4, 3, 1, 3, 4, 1, NA, 2, 3, 2, N~
```

```
$ vertrauen_eu_kommission           <dbl> 5, 5, 3, 1, 3, 5, 3, NA, 4, 4, 2, N~
$ vertrauen_eu_parlament            <dbl> 5, 5, 3, 1, 3, 6, 3, NA, 4, 4, 2, N~
```

Die hier verwendete Syntax PAKETNAME::FUNKTION ist vielleicht zunächst etwas irritierend. Mit ihr rufen wir ein Paket auf, welches wir nicht vorher geladen haben. Das ist mitunter nützlich und kommt hier zur Anwendung, weil wir das Paket `tibble` hier ansonsten nicht benutzen.

Schließlich lassen sich mit dem Befehle `summary` auch noch Eckwerte wie die Ausprägung von Faktorenstufen (bei Faktoren) oder Lageparameter (bei metrischen Variable) ermitteln.

```
summary(sample_gross)
```

```
     alter              geschlecht  fernsehkonsum        politisches_interesse
 Min.   :18.00   KEINE ANGABE:  0   Min.   :0.000   MITTEL          :231
 1st Qu.:38.00   MANN        :245   1st Qu.:4.000   STARK           :138
 Median :56.00   FRAU        :254   Median :7.000   WENIG           : 58
 Mean   :53.78   DIVERS      :  1   Mean   :5.305   SEHR STARK      : 53
 3rd Qu.:68.00                      3rd Qu.:7.000   UEBERHAUPT NICHT: 17
 Max.   :93.00                      Max.   :7.000   (Other)         :  0
 NA's   :3                          NA's   :8       NA's            :  3
 links_rechts_einordnung   wahlabsicht_partei       zufriedenheit_demokratie
 Min.   : 1.000          CDU-CSU    : 97      ZIEMLICH ZUFRIEDEN:151
 1st Qu.: 4.000          DIE GRUENEN: 89      ETWAS ZUFRIEDEN   : 63
 Median : 5.000          SPD        : 51      ETWAS UNZUFRIEDEN : 46
 Mean   : 4.935          FDP        : 50      SEHR ZUFRIEDEN    : 32
 3rd Qu.: 6.000          AFD        : 33      ZIEML. UNZUFRIEDEN: 25
 Max.   :10.000          (Other)    : 56      (Other)           :  3
 NA's   :23              NA's       :124      NA's              :180
       entwicklung_kriminalitaet social_media_nachrichtenquelle
 HAT ETWAS ZUGENOMMEN:168         Min.   :0.000
 HAT STARK ZUGENOMMEN:163         1st Qu.:0.000
 IST GLEICH GEBLIEBEN:118         Median :1.000
 HAT ETWAS ABGENOMMEN: 34         Mean   :2.921
 HAT STARK ABGENOMMEN:  3         3rd Qu.:7.000
 (Other)             :  0         Max.   :7.000
 NA's                : 14         NA's   :45
        glaubwuerdigkeit_oer_tv       glaubwuerdigkeit_privat_tv
 EHER GLAUBWUERDIG   :233        EHER GLAUBWUERDIG   :232
 SEHR GLAUBWUERDIG   :177        EHER N. GLAUBWUERDIG:154
 EHER N. GLAUBWUERDIG: 62        SEHR GLAUBWUERDIG   : 40
 GAR NICHT GLAUBWUERD: 13        GAR NICHT GLAUBWUERD: 23
```

```
DATENFEHLER: MFN    :  0        DATENFEHLER: MFN    :  0
(Other)             :  0        (Other)             :  0
NA's                : 15        NA's                : 51
       glaubwuerdigkeit_zeitungen      glaubwuerdigkeit_social_media
EHER GLAUBWUERDIG   :273          EHER N. GLAUBWUERDIG:230
SEHR GLAUBWUERDIG   :124          GAR NICHT GLAUBWUERD:109
EHER N. GLAUBWUERDIG: 53          EHER GLAUBWUERDIG   : 74
GAR NICHT GLAUBWUERD: 11          SEHR GLAUBWUERDIG   : 11
DATENFEHLER: MFN    :  0          DATENFEHLER: MFN    :  0
(Other)             :  0          (Other)             :  0
NA's                : 39          NA's                : 76
vertrauen_mitmenschen vertrauen_gesundheitswesen
Min.   :1.000         Min.   :1.000
1st Qu.:2.000         1st Qu.:4.000
Median :2.000         Median :5.000
Mean   :2.225         Mean   :4.975
3rd Qu.:3.000         3rd Qu.:6.000
Max.   :3.000         Max.   :7.000
NA's   :15            NA's   :175
vertrauen_bundesverfassungsgericht vertrauen_bundestag
Min.   :1.000                      Min.   :1.000
1st Qu.:4.000                      1st Qu.:3.000
Median :6.000                      Median :4.000
Mean   :5.212                      Mean   :4.076
3rd Qu.:7.000                      3rd Qu.:5.000
Max.   :7.000                      Max.   :7.000
NA's   :179                        NA's   :183
vertrauen_stadt_gemeindeverwaltung vertrauen_katholische_kirche
Min.   :1.000                      Min.   :1.000
1st Qu.:4.000                      1st Qu.:1.000
Median :5.000                      Median :2.000
Mean   :4.495                      Mean   :2.259
3rd Qu.:5.000                      3rd Qu.:3.000
Max.   :7.000                      Max.   :7.000
NA's   :177                        NA's   :179
vertrauen_evangelische_kirche vertrauen_justiz vertrauen_fernsehen
Min.   :1.000                 Min.   :1.000    Min.   :1.000
1st Qu.:1.000                 1st Qu.:4.000    1st Qu.:3.000
Median :3.000                 Median :5.000    Median :4.000
Mean   :3.016                 Mean   :4.567    Mean   :3.642
3rd Qu.:4.000                 3rd Qu.:6.000    3rd Qu.:5.000
Max.   :7.000                 Max.   :7.000    Max.   :7.000
NA's   :182                   NA's   :179      NA's   :176
```

```
 vertrauen_zeitungswesen vertrauen_hochschulen vertrauen_bundesregierung
 Min.   :1.000           Min.   :1.000         Min.   :1.00
 1st Qu.:3.000           1st Qu.:5.000         1st Qu.:3.00
 Median :4.000           Median :5.000         Median :4.00
 Mean   :4.006           Mean   :5.176         Mean   :4.08
 3rd Qu.:5.000           3rd Qu.:6.000         3rd Qu.:5.00
 Max.   :7.000           Max.   :7.000         Max.   :7.00
 NA's   :177             NA's   :177           NA's   :177
 vertrauen_polizei vertrauen_parteien vertrauen_eu_kommission
 Min.   :1.000     Min.   :1.000      Min.   :1.000
 1st Qu.:4.000     1st Qu.:2.000      1st Qu.:2.000
 Median :5.000     Median :3.000      Median :4.000
 Mean   :4.994     Mean   :3.205      Mean   :3.516
 3rd Qu.:6.000     3rd Qu.:4.000      3rd Qu.:5.000
 Max.   :7.000     Max.   :6.000      Max.   :7.000
 NA's   :177       NA's   :178        NA's   :178
 vertrauen_eu_parlament
 Min.   :1.000
 1st Qu.:2.000
 Median :4.000
 Mean   :3.575
 3rd Qu.:5.000
 Max.   :7.000
 NA's   :178
```

Nun schauen wir uns den ALLBUS selbst – also den Gesamtdatensatz – genauer an.

```
daten
```

```
# A tibble: 5,342 x 544
   za_nr           doi   version respid substudy mode    splt21  eastwest german
   <dbl+lbl>       <chr> <chr>    <dbl> <dbl+lb> <dbl+l> <dbl+l> <dbl+lb> <dbl+>
 1 5280 [ALLBUS 2~ http~ 2.0.0 ~      1 1 [SIMU~ 4 [MAI~ 2 [SPL~ 1 [ALTE~ 1 [JA]
 2 5280 [ALLBUS 2~ http~ 2.0.0 ~      2 1 [SIMU~ 4 [MAI~ 1 [SPL~ 1 [ALTE~ 1 [JA]
 3 5280 [ALLBUS 2~ http~ 2.0.0 ~      3 1 [SIMU~ 4 [MAI~ 1 [SPL~ 1 [ALTE~ 1 [JA]
 4 5280 [ALLBUS 2~ http~ 2.0.0 ~      4 1 [SIMU~ 4 [MAI~ 2 [SPL~ 2 [NEUE~ 1 [JA]
 5 5280 [ALLBUS 2~ http~ 2.0.0 ~      5 1 [SIMU~ 4 [MAI~ 2 [SPL~ 1 [ALTE~ 1 [JA]
 6 5280 [ALLBUS 2~ http~ 2.0.0 ~      6 1 [SIMU~ 3 [CAW~ 2 [SPL~ 1 [ALTE~ 1 [JA]
 7 5280 [ALLBUS 2~ http~ 2.0.0 ~      7 2 [SEQU~ 3 [CAW~ 3 [SPL~ 1 [ALTE~ 1 [JA]
 8 5280 [ALLBUS 2~ http~ 2.0.0 ~      8 1 [SIMU~ 4 [MAI~ 2 [SPL~ 1 [ALTE~ 1 [JA]
 9 5280 [ALLBUS 2~ http~ 2.0.0 ~      9 2 [SEQU~ 3 [CAW~ 3 [SPL~ 2 [NEUE~ 1 [JA]
10 5280 [ALLBUS 2~ http~ 2.0.0 ~     10 1 [SIMU~ 4 [MAI~ 2 [SPL~ 1 [ALTE~ 1 [JA]
```

```
# i 5,332 more rows
# i 535 more variables: ep01 <dbl+lbl>, ep03 <dbl+lbl>, ep04 <dbl+lbl>,
#   ep06 <dbl+lbl>, lm01 <dbl+lbl>, lm02 <dbl+lbl>, lm19 <dbl+lbl>,
#   lm20 <dbl+lbl>, lm21 <dbl+lbl>, lm22 <dbl+lbl>, lm14 <dbl+lbl>,
#   xr19 <dbl+lbl>, xr20 <dbl+lbl>, lm27 <dbl+lbl>, lm28 <dbl+lbl>,
#   lm29 <dbl+lbl>, lm30 <dbl+lbl>, lm31 <dbl+lbl>, lm32 <dbl+lbl>,
#   lm33 <dbl+lbl>, lm34 <dbl+lbl>, lm35 <dbl+lbl>, lm36 <dbl+lbl>, ...
```

Es wird schnell klar, das weniger die Anzahl der Beobachtungen als vielmehr die Anzahle der Variablen (544) eine Herausforderung darstellt, zumal diese eher kryptische Namen wie `hp06` haben. Wie sich also zurechfinden?

4 Die Variablen des ALLBUS

Zum Glück lassen sich die sog. *Labels*, also die sprechenden Beschriftungen die sowohl Fragen and auch Antwortoptionen im ALLBUS-Stata-Datensatz haben, mittels R extrahieren (dies geschieht mit dem Paket `labelled`). Wir haben dies bereits vorbereitend für den ALLBUS gemacht und laden die entsprechenden Tabellen nun nur noch.

```
variablen <- read_csv2("Datensatz/ALLBUS_2021_variablen.csv", show_col_types = FALSE)
variablen_optionen <- read_csv2("Datensatz/ALLBUS_2021_variablen_optionen.csv", show_col_type
```

Es lohnt sich, beide Objekte mittels `View()` oder durch eine Klick ob die beiden Objekte *variablen* und *variablen_optionen* in RStudio anzuschauen. Interessant sind die Felder variable (der Name der Variable im ALLBUS) und label (eine sprechende Beschreibung).

Suchen wir etwa nach `hp06` finden wie die Label-Beschreibung "EPIDEMIE: STAAT DARF KRANKE ISOLIEREN", die schon deutlich besser interpretierbar ist als `hp06`. Eine genaue Dokumentation und (vor allem wichtig) der genaue Fragetext findet sich in den Dokumenten ZA5280_fb_CAWI.pdf (Fragebogen) und ZA5280_cdb.pdf (Variablenreport). Beide sind wie der ALLBUS selbst abgelegt im Ordner Datensatz, werden aber ausserhalb von RStudio geöffnet.

Der Fragebogen reicht normalerweise aus, um sich einen Überblick zu verschaffen, aber der Variablenreport ist dann nützlich, wenn man den Zusammenhang zwischen einem Dummywert (bspw. "4") und dessen Bedeutung in Verbindung mit einer bestimmten Variable herausfinden möchte. Die Fragen lauter bei `hp06`

Und was denken Sie über folgende Maßnahmen: Sollte in Deutschland in Zeiten schwerer Epidemien der Staat das Recht haben, Folgendes zu tun?

Nachweislich infizierte Personen isolieren

Und der Dummy-Wert "4" steht bei dieser Frage für die Antwort *Auf keinen Fall*.

Wenn man die drei kleinen Zufallssamples mit dem Hauptdatensatz vergleicht, fällt schnell auf, dass die Samples ausschließlich (echte) Zahlenfelder für (vor allem) ordinale Likert-skalierte Variablen enthalten. Bei diesen bedeutet ein höherer Wert i.d.R. mehr Zustimmung oder eine ausgeprägtere Verhaltensausprägung gegenüber geringeren Werten. Es gibt aber auch Fälle in denen diese sog. Polarität der Variablenwerte umgedreht ist und geringe Werte "stärker" sind als hohe, order solche, in denen wir es mit nominalen Skalen zu tun haben, die Zahlen also in keinerlei logischem Zusammenhang stehen.

Was heißt das konkret? Zur Sicherheit sollte man im Rahmen einer eigenen Analyse in den Hauptdatensatz und in die Dokumentatuin schauen um absolut sicher zu sein, dass man keine unzulässigen Umformungen oder Berechnungen vornimmt (und etwa den Mittelwert einer nominalskalierten Variable bestimmt), oder die Ergebnisse misinterpretiert (etwa wenn die Polarität einer Variablen umgedreht codiert wurde). Es gilt immer: *know your data*.

Zunächst ist es aber vollkommen legetim, um die Variablenliste oder das Befragungsdokument nach interessanten Variablen zu durckforsten. Wir können in der `View()`-Ansicht des Objekts variable nach Begriffen suchen, die in den Labels vorkommen. Beispielsweise finden wir mit einer Suche nach dem Begriff 'medien' die Variablen `lm35` (Nutzung von sozialen Medien als Nachrichtenquelle) und `lm39` (Glaubwürdigkeit sozialer Medien mit Blick auf Kriminalität), die uns vielleicht interessieren.

5 Bildung eines Teilsamples

Ein Schritt, der praktisch für alle Studienprojekte im Verlauf des Semesters relevant sein wird, ist die Bildung eines Teildatensatzes, welcher die Variablen (und ggf. Fälle) enthält, die für Ihre Analyse relevant sind.

Technisch gesehen ist das gar nicht unbedingt notwendig – wir können jeder Zeit Berechnungen am Gesamtdatensatz anstellen. Aber oft ist ein Teildatensatz übersichtlicher und ermöglicht ein besseres Verständnis der Daten.

Wie bildet man ein solches Teilsample? Entscheidend ist hier die Funktion `select`.

```
fernsehkonsum <- daten %>%
  select(age, sex, lm02)
```

Wir extrahieren hier mittels select drei Variablen, nämlich Alter (`age`), Geschlecht (`sex`) und den Fernsehkonsum in Minuten (`lm02`).

Das Ergegnis ist ein Datensatz, der weiterhin alle 5.342 Fälle, aber nur drei (statt 544) Variablen enthält.

```
fernsehkonsum
```

```
# A tibble: 5,342 x 3
   age       sex       lm02
   <dbl+lbl> <dbl+lbl> <dbl+lbl>
 1 54        2 [FRAU]  210
 2 53        1 [MANN]   90
 3 89        2 [FRAU]  135
 4 79        1 [MANN]   60
 5 62        2 [FRAU]  180
 6 23        1 [MANN]   45
 7 31        2 [FRAU]   30
 8 57        2 [FRAU]   -9 [KEINE ANGABE]
 9 68        1 [MANN]  180
10 51        2 [FRAU]  180
# i 5,332 more rows
```

Eine gewisse Komplikation ist allerdings die Tatsache, dass in diesem Ausschnitt die Variablen `sex` noch eine Dummy-Zahl ist (1 = männlich, 2 = weiblich, 3 = divers) und zudem die Variable `lm02` negative Werte enthät. Diese zeigen keinen negativen Fernsehkonsum an, sondern werden für Spezialwerte verwendet ("keine Angabe", "durch Filterbedingung weggefallen"). Als Faustregel gilt: **Negative Werte im ALLBUS sollten praktisch immer durch NAs ersetzt werden**. Das ist unbedingt von "0" als Wert zu unterscheiden. Mit einer "echten" Null kann ebenso wie mit "echten" Negtivwerten gerechnet werden, dies führt aber zu substantiellen Verzerrungen, wenn es sich um Dummy-Werte handelt.

Der folgenden Codeblock bereinigt die Daten zum Fernsehkomsum. Dazu benennte er die drei Variablen znnächst in transparentere Were um. Anschlißend werden negative Werte in NAs umgewandelt (hier mit der Funktion replace_with_na_all aus dem Paket naniar). Dann wird das Geschlecht faktorisiert, was die Zahlen durch das Label (also MANN / FRAU / DIVERS) ersetzt. Und schließlich werden die Labels und Attribute entfernt, die nun nicht mehr benötigt werden.

```
fernsehkonsum_bereinigt <- fernsehkonsum %>%
  rename(alter = age,
         geschlecht = sex,
         fernsehkonsum_minuten = lm02) %>%
  naniar::replace_with_na_all(condition = ~.x < 0) %>%
  mutate(geschlecht = as_factor(geschlecht)) %>%
  labelled::remove_labels() %>%
  labelled::remove_attributes("format.stata")
```

```
fernsehkonsum_bereinigt
```

```
# A tibble: 5,342 x 3
   alter geschlecht fernsehkonsum_minuten
   <dbl> <fct>                      <dbl>
 1    54 FRAU                         210
 2    53 MANN                          90
 3    89 FRAU                         135
 4    79 MANN                          60
 5    62 FRAU                         180
 6    23 MANN                          45
 7    31 FRAU                          30
 8    57 FRAU                          NA
 9    68 MANN                         180
10    51 FRAU                         180
# i 5,332 more rows
```

6 Zusammenfassung

Wir sind jetzt in einer guten Position, um mit der praktischen Arbeit am ALLBUS zu beginnnen, also der Analyse und Interpretation konkreter Daten.