gesis Leibniz-Institut für Sozialwissenschaften

Workflow Datenaufbereitung

Version 1, 11.12.2023, Beck, Kerstin/Heizmann, Boris

Zugehörig zu:

Behrens, K., Quandt, M., Zenk-Möltgen, W., Beck, K., & Brück, R. (2023). KonsortSWD Measure TA.2-M.12: FDMontheground - Projektbericht. Zenodo. https://doi.org/10.5281/zenodo.10260847.

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Organisation der Daten und Dokumente

Version 4, 11.12.2023, Änderungen von Boris Heizmann
Version 3, 14.09.2023, Änderungen von Kerstin Beck
Version 2, Mai 2023, Änderungen von ET
Version 1, 14.09.2023, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Im Rahmen des Forschungsdatenmanagements empfiehlt es sich, im Vorfeld Absprachen über die Dateiablage zu treffen. Aufbereitungsschritte und Sicherung werden dadurch für alle Beteiligten transparenter und die Koordination vereinfacht.

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Laufwerke, in diesem Fall „I:“, muss vorhanden sein.

3 Voraussetzungen

-

4 Beschreibung der Workflows

4.1 Ordnerstruktur anlegen

Bevor die eigentliche Aufbereitung beginnt, sind einige vorbereitende Schritte notwendig. Zunächst sollte im individuellen Unterordner der bearbeitenden Person (i.e. da, wo Schreibrechte vorliegen) im Laufwerk I: ein Ordner erstellt werden, in dem drei weitere Unterordner Platz für die Dateien bieten, die zur Verfügung gestellt werden. Dabei empfiehlt sich die folgende Ordnerstruktur:

I:\...\aufbereitung_eigener Name\Studienname\NUMMER ZAxxxx

Der Name des Ordners setzt sich aus dem Namen der Studie, ggfs. einer Ordnungsnummer sowie der GESIS-Studiennummer (ZAXXXX) zusammen.

In diesem Ordner sollten dann die Unterordner „Aufbereitung“, „Originale“ und „Service“ erstellt werden.

Der Ordner „Originale“ sollte alle Originaldateien enthalten, die zu Beginn der Aufbereitung zur Verfügung gestellt werden. Die Dateien im „Originale“-Ordner sollten nicht verändert werden und in ihrer originalen Form erhalten bleiben.

4.2 Dateien sammeln / erstellen

Die Dateien, die im weiteren Aufbereitungsprozess notwendig sind und in den Ordner „Aufbereitung“ kopiert (nicht verschoben!) werden sollten, sind

- der Datensatz selbst (in der englischsprachigen Version) (.sav),
- der Fragebogen (z.B. name_bilingual matrix.pdf) und
- alle Dokumentationsdateien, Methodenberichte, Berichte etc. (meist als .pdf)

Der Ordner „Aufbereitung“ ist der Ordner, in dem die Aufbereitungsarbeit stattfindet und in dem die Dateien verändert werden. Er enthält also auch die Aufbereitungssyntax.

Im Ordner „Service“ werden später nach der Bearbeitung der drei oben genannten Dateien diese in ihrer fertigen Version (nachdem sie nicht mehr bearbeitet werden müssen) abgespeichert. Im Ordner „Originale“ UND im Ordner „Service“ sollten keine Veränderungen mehr stattfinden.

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Archivvariablen erstellen / Syntax erarbeiten

Version 4, 11.12.2023, Änderungen von Boris Heizmann
Version 3, 14.09.2023, Änderungen von Kerstin Beck
Version 2, Mai 2023, Änderungen von ET
Version 1, 14.09.2023, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Jedem Datensatz müssen zusätzlich zu den Datensatzvariablen sogenannte Archivvariablen zugeordnet werden. Hierbei handelt es sich z.B. um Studiennummer, DOI, Version, Edition, Survey.

Im Folgenden wird der Aufbau der dafür notwendigen Syntaxdatei und deren Einsatzweise erläutert sowie die benötigten Befehle beispielhaft aufgeführt (s. Kapitel 4.4).

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Arbeitsordner, in diesem Fall auf dem Laufwerk „I:“, und Programme (SPSS und Excel) muss vorhanden sein.

3 Voraussetzungen

Die Ordnerstruktur wird entsprechend der Anleitung (How-to_Ordnerstruktur.docx) erstellt.

4 Beschreibungen der Workflows

4.1 Syntaxfile erstellen

Wenn alle notwendigen Dateien im Ordner „Aufbereitung“ gespeichert sind, muss eine weitere Datei erstellt werden. Dafür wird der Datensatz (die .sav-Datei) geöffnet. Über das Auswahlmenü

Datei > Neu > Syntax

wird eine Syntax erstellt, in der die Aufbereitung stattfindet. Sie sollte, wie auch der Datensatz, mit der ZA-Nummer und ggfs. weiteren Ordnungsnummern der Datenkollektion benannt werden. Für die gesamte Aufbereitung wird diese einzige Syntax verwendet (in der alle notwendigen Überarbeitungsschritte gespeichert werden), die dann am Ende komplett über den Quelldatensatz gelaufen lassen werden kann, um aus dem Original die Version zu erstellen, die veröffentlicht wird.

Weitere nötige Daten werden in der Regel vom Kollektionsverantwortlichen per Mail über einen Ordnerlink geliefert. Hier werden neben dem Pfad zu den Dateien auch der doi des Datensatzes, sowie die ZA-Nummer mitgeteilt, die in den Datensatz, die Dateinamen sowie Readme etc. eingepflegt werden.

In einigen Fällen kann es passieren, dass dem neuen Datenprodukt noch kein doi zugewiesen wurde. Dieser wird entweder im Verlauf der Bearbeitungsphase zugesendet, oder die entsprechenden Felder für den doi bleiben zunächst frei und vom Kollektionsverantwortlichen eingefügt, sobald der doi feststeht (die doi-Felder in der Syntax aber nicht herauslöschen, sondern mit einem * als Kommentar kennzeichnen, damit die Aktualisierung des doi nicht vergessen wird).

4.2 Basis der Syntax kopieren

Es lohnt sich, zu Beginn der Aufbereitung einen älteren Datensatz zur Hand zu nehmen und die Syntax zu kopieren und in die jetzt zu bearbeitende Syntax einzufügen. Orientieren kann man sich hierbei, sofern zutreffend, an den Vorgänger-Datensätzen. Es kann entweder Schritt für Schritt vorgegangen werden oder es kann eine Kopie einer kompletten Syntax angelegt werden.

Fragen aus einem alten Fragebogen können sich durchaus wiederholen. Ist das der Fall, lohnt es sich, die Syntax des damaligen Datenproduktes („Vorgängerstudie“) als Basis zu nehmen. Für die Demographie- und die technischen Variablen ist es jedoch besser, eine Studie als Vorgänger zu wählen, welche zeitlich kurz vor der zu bearbeitenden erhoben wurde, und die entsprechenden Variablen ebenfalls enthält.

Beim Kopieren von Syntaxkomponenten aus Word können Formatierungsschwierigkeiten z.B. bei Anführungszeichen auftreten, daher ist das Kopieren aus diesem Dokument nicht empfehlenswert. Kopiert werden sollte also nur von Syntax zu Syntax in SPSS.

Hinweis für die Aufbereitung: mit dem Vorzeichen * lässt sich eine eingetragene Zeile zum Kommentar machen. Mit der Kommentarfunktion sollte jeder neue Abschnitt indiziert werden, der eingeleitet wird.

4.3 Syntaxkopf bearbeiten

Der erste Part, der bearbeitet wird, ist der Syntaxkopf. Zwar nimmt der Kopf keinen Einfluss auf die veröffentlichte Version des Datensatzes, er enthält jedoch relevante Informationen für alle, die zu einem späteren Zeitpunkt auf die Syntax zugreifen.

Alle Stellen, die im Folgenden farbig markiert sind, müssen auf den aktuellen Stand gebracht werden.

Der Aufbau des Syntaxkopfes gestaltet sich wie folgt:

*Encoding: UTF-8.

*archive study id: GESIS ZAxxxx=> Ersetzen durch die aktuelle ZA-Nummer (s. E-Mail)
*survey: NAME NUMBER => Name des Datenproduktes, ggfs. mit Nummer
*lifecycle stage: archiving
*lifecycle event: processing.ingress
*software: SPSS
*software version:
*processor: xx => Ersetzen durch die individuellen Initialen des Bearbeitenden

*date: 20xx-xx-xx => Ersetzen durch das aktuelle Datum
*input datafile local: => Ersetzen durch den Dateipfad zur Datei, die durch die Syntax bearbeitet wird
*output datafile: (see SAVE OUTFILE)

4.4 Archivvariablen bearbeiten

Wichtig: JEDER Befehl muss nach der Ausführung noch einmal kontrolliert werden: Es kann passieren, dass sich im Fragebogen kleinste Details geändert haben, wie z.B. die Bezeichnungen der Arbeitsfelder oder dass beim Geschlecht nicht mehr nur männlich und weiblich, sondern auch etwas anderes wählbar wird. Um zu vermeiden, dass Variablen nach veralteten Standards aufbereitet werden, muss also stets die jeweilige Variable mit dem Fragebogen bzw. den Originalvariablen abgeglichen und auf Korrektheit geprüft werden.

4.4.1 Study-Nr.

Nachdem der Kopf bearbeitet ist, geht es an die tatsächliche Aufbereitung des Datensatzes. Dafür werden zunächst die Archivvariablen bearbeitet. Die Archivvariablen beinhalten alle formellen Informationen zum Datensatz an sich. Anhand der Syntax gestaltet sich die Aufbereitung wie folgt:

```
COMPUTE studyno = XXXX.
FORMATS studyno (F4.0).
VARIABLE LABELS
 studyno "ARCHIVE STUDY NUMBER".
VALUE LABELS
 studyno XXXX "GESIS STUDY ID ZAXXXX".
EXECUTE.
```

4.4.2 DOI

Die zweite Variable, die erstellt wird, beinhaltet die doi-Nummer des Datensatzes. Die doi-Nummer –oder ‚Digital Object Identifier'-Nummer, die eine eindeutige Zuteilung in einem standardisierten System erlaubt - wird dabei wie folgt zugeordnet:

```
STRING doi (A20).
COMPUTE doi='DOI: 10.4232/1.XXXXX'.
VARIABLE LABELS
        doi "DIGITAL OBJECT IDENTIFIER".
EXECUTE.
```

Dabei wird mit dem Befehl STRING eine neue String-Variable erstellt. Im Gegensatz zu numerischen Variablen werden hier keine Zahlenwerte zugeordnet, sondern eine Zeichenkette, wofür die meisten statistischen Verfahren nicht anwendbar sind. Im vorliegenden Fall soll die Stringvariable 20 Zeichen lang sein (A20).

4.4.3 Version

Das Vorgehen bei der Variable ‚Version' gleicht dem Vorgehen bei der doi-Variable. Hier muss das Datum an das aktuelle Datum angepasst werden. Die Versionsnummer ist bei der ersten Aufbereitung immer 1.0.0 und wird erst in späteren Aktualisierungen angepasst.

```
STRING version (A25).
VARIABLE LABELS
 version
 'GESIS ARCHIVE VERSION & DATE'.
COMPUTE version ='1.0.0 (YYYY-MM-DD)'. => Ersetzen durch das Release-Datum
EXECUTE.
```

4.4.4 Edition

Das Vorgehen bei der Edition-Variable gleicht dem bei der *studyno*-Variable. Sie beschreibt die Art der Veröffentlichung, was in diesem Falle die Archivversion ist. In der Regel kann dieser Befehlsblock einfach aus dem Vorgänger übernommen werden und bedarf keinerlei Anpassung.

```
COMPUTE edition=1.
FORMATS edition (F1.0).
VARIABLE LABELS
 edition "DATASET EDITION".
VALUE LABELS
 edition 1 "ARCHIVE RELEASE".
EXECUTE.
```

4.4.5 Survey

Unter der Annahme, dass die Studie eine Ordnungsnummer hat, wird die *Survey*-Nummer auf die gleiche Art und Weise zugeordnet wie die Edition. Außerdem muss der Zeitraum angepasst werden, in dem die Erhebung durchgeführt wurde.

```
COMPUTE survey=NUMMER.
FORMATS survey (F3.0).
VARIABLE LABELS
        survey "SURVEY IDENTIFICATION".
VALUE LABELS
        survey NUMMER "NAME NUMMER (Monat Jahr)".
EXECUTE.
```

4.4.6 Caseid

Die *Caseid*-Variable wird aus der bereits vorhandenen *Respondent-ID*-Variablen erstellt. Da die Nummern teilweise von Land zu Land einzeln vergeben wurden, kann es sein, dass einzelne Nummern doppelt vorkommen.

4.4.7 Serialid

Die *Serialid*-Variable wird – wie auch die *caseid* – auf Basis einer bereits vorhandenen Variable erstellt. Sie zählt die Position im Datensatz. Diese wird gebraucht, um die ‚uniqid' zu bilden.

4.4.8 Uniqid

Für manche Auswertungen und für den späteren Verlauf der Aufbereitung kann es notwendig sein, jedem Fall eine eindeutige Nummer zuzuordnen. Hierfür wird der folgende Befehl verwendet:

```
COMPUTE uniqid=LAND*1000000+serialid.
FORMATS uniqid (F8.0).
VARIABLE LABELS uniqid "UNIQUE RESPONDENT ID (CASEID BY COUNTRY CODE)".
EXECUTE.
```

Auch hier ist es in der Regel so, dass dieser Befehlsblock von einer Vorgängerstudie übernommen werden kann und sich nicht ändert. Es muss dennoch geprüft oder mit dem Kollektionsverantwortlichen von Zeit zu Zeit abgesprochen werden, ob sich etwas geändert hat. Es kann z. B. erforderlich sein, weitere Nullen anzuhängen bei der Befehlskomponente LAND*1000000: Die Variable LAND ist die noch unbearbeitete Ländervariable, die den einzelnen Personen ihren Herkunftsstaat zuordnen. Anhand dieses Befehls wird jedem Staat durch die Multiplikation mit einer Zahl, die größer ist, als alle vorhandenen caseids, ein Präfix zugeschrieben, welches dem Ländercode entspricht. Die Addition der caseid sorgt dafür, dass jedem Fall eine individuelle Zahl zugeordnet wird. Diese setzt sich aus dem zweistelligen Ländercode und der caseid zusammen. Eine Person aus dem Land „9" mit der caseid „1234" hätte somit die uniqid „9001234".

4.4.9 Mode

Die Variable ‚Mode' beschreibt, auf welchem Weg die befragten Personen erreicht wurden. Sofern telefonbasierte Interviews geführt wurden, bestehen hier meist die Optionen Festnetz oder Mobiltelefon. Auch die ‚Mode'-Variable basiert auf einer bereits vorhandenen Variable:

```
RENAME VARIABLES ph1 = mode.
VARIABLE LABELS mode "INTERVIEW: TYPE OF PHONE LINE (PH1)".
EXECUTE.
```

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Dublettencheck auf Fallebene

Version 4, 11.12.2023, Änderungen von Boris Heizmann
Version 3, 14.09.2023, Änderungen von Kerstin Beck
Version 2, Mai 2023, Änderungen von ET
Version 1, 14.09.2023, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Die Kontrolle der doppelten Fälle stellt einen wichtigen Prüfschritt dar, in dem überprüft wird, ob die Antworten eines Befragten (oder ein Fall) mehrfach in den Datensatz aufgenommen wurden.

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Arbeitsordner auf dem Laufwerk „I:“ sowie SPSS muss vorhanden sein.

3 Voraussetzungen

Ordnerstruktur angelegt und Arbeitsdateien entsprechend gespeichert.

4 Beschreibungen der Workflows

4.1 Doppelte Fälle ermitteln

a) In SPSS: Reiter „Daten“, Unterpunkt „Doppelte Fälle ermitteln“

b) Auswahl von bis zu 64 Variablen und Einfügen in das Fenster „Übereinstimmende Fälle definieren durch“. Wichtig ist hier: Es dürfen nur Fragen gewählt werden, die von den Befragten beantwortet wurden, also Fragen des eigentlichen Fragebogens sowie die demographischen Daten der Befragten, sowie das Land. Dabei sollten vor allem solche Variablen gewählt werden, die viele Ausprägungen haben, da bei Dummyvariablen Unterschiede per se unwahrscheinlicher bzw. zufällige Dubletten wahrscheinlicher sind.

c) Bestätigen der Auswahl, indem man auf ‚Einfügen‘ klickt und den Befehl dann noch in der Syntax ausführt. Der Befehl muss auskommentiert (mit *) gespeichert werden. Das ist wichtig, damit man im Nachhinein nachvollziehen kann, welche Variablen Teil des Befehls waren. Im Ausgabefenster wird dann geprüft, ob es unter „Doppelte Fälle“ angezeigte Fälle gibt.

Statistiken

Indikator jeder letzten Fallübereinstimmung als Primär

N	Gültig	12832
	Fehlend	0

Indikator jeder letzten Fallübereinstimmung als Primär

		Häufigkeit	Prozent	Gültige Prozente	Kumulierte Prozente
Gültig	Doppelter Fall	23	,2	,2	,2
	Primärer Fall	12809	99,8	99,8	100,0
	Gesamt	12832	100,0	100,0	

Beispiel:

> Hier gibt es 23 doppelte Fälle. Falls es keine doppelten Fälle geben sollte, wäre die Zeile „Doppelter Fall“ nicht vorhanden.

d) Sollte es doppelte Fälle geben: Wechsel im SPSS-Hauptansichtsfenster in die Datenansicht.

e) Durch das bisherige Vorgehen wurden die doppelten Fälle in Paaren an den Anfang des Datensatzes sortiert. Jetzt folgt die Kontrolle, welche der Fälle doppelt vorkommen.

f) Die Anzahl der doppelten Fälle, die in der Tabelle angezeigt werden, zeigt die Hälfte (der Anzahl) der Fälle an, die jetzt kontrolliert werden. Werden also vier doppelte Fälle angezeigt, werden die ersten acht Fälle des Datensatzes kontrolliert (die doppelten Fälle und die dazugehörigen „Originale“). Die Fälle, die sich möglicherweise doppeln, folgen unmittelbar aufeinander (also Zeile 1 und 2, Zeile 3 und 4 etc.). Es kann auch dreifache etc. Fälle geben.

g) Jetzt werden die einzelnen Personen angeschaut, die doppelt vorkommen könnten. Dafür wird nach rechts durch die Ausprägungen gescrollt und kontrolliert, ob sich tatsächlich alle Antworten der demographischen und der im Fragebogen erhobenen Fragen gleichen. Dies ist notwendig, da SPSS nur maximal 64 Variablen einbeziehen kann, was oftmals nicht den gesamten Datensatz umfasst.

h) Gleichen sich alle Ausprägungen, werden die uniqids und die serialids der beiden Fälle notiert und bei Abschluss der Aufbereitung im Readme vermerkt. Manche Studien haben jedoch nur sehr wenige Fragen, sodass Dubletten im Sinne von übereinstimmenden Fällen bei gleichzeitig ca. 1.000 Fällen pro Land quasi unvermeidbar sind bzw. in großer Zahl auftreten und somit nicht notwendigerweise ein Problem darstellen. Der Kollektionsverantwortliche entscheidet dann, ob der Hinweis im Readme sinnvoll ist.

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Transformation auf Standardindices/abgeleitete Variablen

Version 4, 11.12.2023, Änderungen von Boris Heizmann
Version 3, 14.09.2023, Änderungen von Kerstin Beck
Version 2, Mai 2023, Änderungen von ET
Version 1, 14.09.2023, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Um eine geographische Zuordnung der Fälle zu erlauben, werden Ländervariablen erstellt, die nach unterschiedlichen Standards definiert werden.

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Arbeitsordner auf dem Laufwerk „I:“ sowie SPSS muss vorhanden sein.

3 Voraussetzungen

Ordnerstruktur angelegt und Arbeitsdateien entsprechend gespeichert.

4 Beschreibungen der Workflows

Wichtig: JEDER Befehl muss nach der Ausführung noch einmal kontrolliert werden: Es kann passieren, dass sich im Fragebogen kleinste Details geändert haben, wie z.B. die Bezeichnungen der Arbeitsfelder oder dass beim Geschlecht nicht mehr nur männlich und weiblich, sondern auch etwas anderes wählbar wird. Um zu vermeiden, dass Variablen nach veralteten Standards aufbereitet werden, muss also stets die jeweilige Variable mit dem Fragebogen bzw. den Originalvariablen abgeglichen und auf Korrektheit geprüft werden.

4.1 Ländervariablen erstellen / anpassen

Die für die Ländervariablen grundlegende Variable ist bereits im Datensatz vorhanden. Sie wird nach den Standards des Erhebungsunternehmens durchführt.

4.1.1 oricntry

Die erste Ländervariable namens *oricntry wird* durch die einfache Umbenennung der Variable *b* erstellt.

```
RENAME VARIABLES (b=ipscntry).
VARIABLE LABELS
oricntry "ORIGINAL COUNTRY/SAMPLE ID".
```

```
EXECUTE.
```

In der Regel kann dieser Befehlsblock so übernommen werden, da er sich nicht ändert.

4.1.2 country

Die Ausprägungen der *country*-Variable orientieren sich an den Beitrittsjahren in die EU bzw. dem Kandidatenstatus. Somit werden erst die sechs Gründungsmitglieder gruppiert, dann die Länder der Norderweiterung und so weiter. Die Benennungen erfolgen dabei nach dem ISO-Standard 3166.

Da die Variable *oricntry* nach den immer gleichen Standards erstellt wird, ist auch hier ein bereits vorgefertigtes Skript vorhanden (das jedoch natürlich kontrolliert werden muss). Zur Kontrolle kann man die Werte der *oricntry* vor dem Ausführen mit dem Befehl und den untenstehenden Wertelabels vergleichen und eine Kreuztabellierung beider Variablen vornehmen.

```
COMPUTE country=oricntry.
FORMATS country (F2.0).
VARIABLE LABELS
        country "COUNTRY/SAMPLE ID (SERIES STANDARD)".
RECODE country (1=2) (2=29) (3=20) (4=7) (5=4) (6=21) (7=8) (8=11) (9=12) (10=1)
    (11=32) (12=5) (13=19) (14=23) (15=24) (16=6) (17=22) (18=25) (19=3) (20=18)
    (21=26) (22=13) (23=30) (24=28) (25=27) (26=16) (27=17) (28=9) (29=43) (30=41)
    (31=42) (32=35).
EXECUTE.
```

So wird die neue Variable *country* erstellt und rekodiert. Dabei wird jedem Land der vorgeschriebene Wert zugewiesen, den es in der neuen Variablen erhalten soll.

Es sind nicht immer alle Länder Teil eines jeden Datensatzes, deswegen lohnt es sich auch hier, mehrere ältere Ausgaben zu konsultieren, um ggf. herauszufinden, welcher Wert einem Land in der Regel zugewiesen wird. Dabei ist es wichtig, auch die jeweiligen Originaldaten zu betrachten, da hier auch Veränderungen vorliegen können.

Danach werden den jeweiligen Ausprägungen die Labels hinzugefügt. Sollten Staaten nicht im Datensatz enthalten sein, die im Vorgänger enthalten waren, werden sie mit (NOT INCLUDED) vermerkt. In einem letzten Schritt erfolgt die Kreuztabellierung mit oricntry zur Überprüfung.

```
VALUE LABELS
country
1 "FR - France"
2 "BE - Belgium"
3 "NL - The Netherlands"
4 "DE - Germany"
```

```
5 "IT - Italy"
6 "LU - Luxembourg"
7 "DK - Denmark"
8 "IE - Ireland"
 9 "GB - United Kingdom (NOT INCLUDED)"
10 "GBNIR - Northern Ireland (NO SEPARATE SAMPLE)"
11 "GR - Greece"
12 "ES -Spain"
13 "PT - Portugal"
14 "DE-E Germany East (NO SEPARATE SAMPLE)"
15 " (OUTDATED CODE FOR NORWAY)"
16 "FI - Finland"
17 "SE - Sweden"
18 "AT - Austria"
19 "CY - Cyprus (Republic)"
20 "CZ - Czech Republic"
21 "EE - Estonia"
22 "HU - Hungary"
23 "LV - Latvia"
24 "LT - Lithuania"
25 "MT - Malta"
26 "PL - Poland"
27 "SK - Slovakia"
28 "SI - Slovenia"
29 "BG - Bulgaria"
30 "RO - Romania"
31 "TR - Turkey (NOT INCLUDED)"
32 "HR - Croatia"
33 "CY-TCC - Cyprus TCC (NOT INCLUDED)"
34 "MK - Makedonia/FYROM (NOT INCLUDED)"
35 "ME - Montenegro (NOT INCLUDED)"
36 "RS - Serbia (NOT INCLUDED"
37 "AL - Albania (NOT INCLUDED)"
39 "MD - Moldavia (NOT INCLUDED)"
41 "NO - Norway (NOT INCLUDED)"
42 "CH - Switzerland (NOT INCLUDED)"
43 "IS - Iceland (NOT INCLUDED)"
44 "LI - Liechtenstein (NOT INCLUDED)"
49 "IL - Israel (NOT INCLUDED)"
61 "US - USA (NOT INCLUDED)".
```

4.1.3 isocntry

Die erstellte *country*-Variable wird jetzt als Grundlage genutzt, um die *isocntry*-Variable zu erstellen. Hierbei ist der Unterschied zur *country*-Variable, dass es sich um eine Stringvariable handelt, welche als Werte die nach ISO-3166-Standard formulierten Länderkürzel beinhaltet.

```
STRING isocntry (A6).
if (country eq 1) isocntry='FR'.
```

```
if (country eq 2) isocntry='BE'.
if (country eq 3) isocntry='NL'.
if (country eq 4) isocntry='DE'.
if (country eq 5) isocntry='IT'.
if (country eq 6) isocntry='LU'.
if (country eq 7) isocntry='DK'.
if (country eq 8) isocntry='IE'.
if (country eq 9) isocntry='GB'.
*if (country eq 10) isocntry='GB-NIR'.
if (country eq 11) isocntry='GR'.
if (country eq 12) isocntry='ES'.
if (country eq 13) isocntry='PT'.
*if (country eq 14) isocntry='DE-E'.
if (country eq 16) isocntry='FI'.
if (country eq 17) isocntry='SE'.
if (country eq 18) isocntry='AT'.
if (country eq 19) isocntry='CY'.
if (country eq 20) isocntry='CZ'.
if (country eq 21) isocntry='EE'.
if (country eq 22) isocntry='HU'.
if (country eq 23) isocntry='LV'.
if (country eq 24) isocntry='LT'.
if (country eq 25) isocntry='MT'.
if (country eq 26) isocntry='PL'.
if (country eq 27) isocntry='SK'.
if (country eq 28) isocntry='SI'.
if (country eq 29) isocntry='BG'.
if (country eq 30) isocntry='RO'.
if (country eq 32) isocntry='HR'.
EXECUTE.

*if (country eq 31) isocntry='TR'.
*if (country eq 34) isocntry='MK'.
*if (country eq 35) isocntry='ME'.
*if (country eq 39) isocntry='MD'.
*if (country eq 43) isocntry='IS'.
*if (country eq 36) isocntry='RS'.
*if (country eq 41) isocntry='NO'.
*if (country eq 49) isocntry='IL'.
*if (country eq 61) isocntry='US'.
*if (country eq 65) isocntry='BR'.
*if (country eq 83) isocntry='CN'.
*if (country eq 84) isocntry='IN'.
*if (country eq 33) isocntry='CY-TCC'.

VARIABLE LABELS isocntry
        "COUNTRY CODE - ISO 3166".
```

Da auch hier als Basis, die bereits standardisierte *country*-Variable fungiert, ist der vorformulierte Befehl ausreichend, um die Variable zu erstellen. Ab der Variable Türkei sind die Kürzel hier als

Kommentar vermerkt, damit sie bei Bedarf genutzt werden können. Auch diese Variable wird mittels Kreuztabellierung mit oricntry überprüft.

4.2 EU-Gruppierungsvariablen standardisieren

Die EU-Gruppierungsvariablen werden erstellt, um in der Auswertung zwischen den einzelnen Untergruppen einfach differenzieren zu können. So werden die einzelnen Staaten in inhaltlich relevante Gruppen geordnet, die sich zumeist durch EU-Erweiterungen ergeben. Im Normalfall sind die für die Gruppierungsvariablen grundlegenden Variablen bereits im Datensatz vorhanden. Sie müssen nur noch durch Kreuztabellierung mit oricntry geprüft und dann in eine standardisierte Form gebracht werden. Dafür werden jeweils neue Variablen erstellt, die dann rekodiert und deren Labels angepasst werden. Dabei gestaltet sich der gesamte Befehl für jede einzelne Gruppe wie folgt:

```
COMPUTE eu6=VEU6. => neuer Variablenname=Name der im Datensatz vorliegenden Var.
RECODE eu6 (2=0).
VALUE LABELS
 eu6
    0          "Not country group" => Standardisierte Ausprägung für 0, die für alle
Gruppierungsvariablen gilt
    1     "EU-6". =>Standardisiertes Label der Gruppe 1, das beschreibt, welche Staaten in
die Gruppe eingeordnet sind
VARIABLE LABELS
        eu6   "NATION GROUP EU-6". => Standardisiertes Label der Variable
FORMATS eu6 (F1.0).
EXECUTE.
```

Dieses Vorgehen wird mit allen weiteren Gruppen wiederholt. Das genaue Skript sollte dabei aus einer älteren Syntax kopiert und dann Gruppe für Gruppe kontrolliert werden. Dabei ist es wichtig, jede einzelne Gruppe zu überprüfen, da es auch vorkommen kann, dass unterschiedliche oder neue Gruppen gebildet werden (s. Brexit).

Beim Verwenden einer älteren Syntax können zwei problematische Situationen auftreten (abgesehen von veränderten Definitionen bzw. Variablennamen seitens des Datengebers):

a) Im Datensatz vorkommende Variablen wurden nicht in das standardisierte Format übertragen, da die Gruppe neu ist und die Vorgängersyntax keine entsprechenden Befehle enthielt.

   In diesem Fall muss für die jeweiligen Variablen ein neuer Befehl geschrieben werden, dessen Aufbau dem der anderen Variablen gleicht. Das Variablenlabel und das Ausprägungslabel zur 1 müssen angepasst und in ein möglichst kurzes Format gebracht werden. Dabei kann man sich in den meisten Fällen aber an den bereits vorhandenen Labels orientieren, in jedem Fall muss dies aber mit dem Kollektionsverantwortlichen abgesprochen werden.

b) Befehle, die in der Syntax vorkommen, wurden nicht ausgeführt.

Der Grund hierfür ist meist ein vorgestelltes „*“. Damit werden die Befehle in Kommentare umgeformt. So bleiben sie erhalten und können in einer späteren Version wieder genutzt werden.

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Gewichtungsvariablen prüfen

Version 3, 11.12.2023, Änderungen von Boris Heizmann
Version 2, 14.09.2023, Änderungen von Kerstin Beck
Version 1, Juli 2023, Änderungen von ET, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Die Gewichte können vor der Auswertung aktiviert werden, um für die jeweilige Gruppe das Verhältnis der Stichprobengröße auf die tatsächlichen Populationsgrößen der Länder herzustellen sowie um die Repräsentativität zu erhöhen.

Die einzige Aufgabe, die hier abseits von der Benennung der Gewichte besteht, ist es, die Gewichtungsvariablen zu überprüfen, da in den meisten Fällen nicht alle Gewichte tatsächlich Teil des Datensatzes sind.

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Arbeitsordner auf dem Laufwerk „I:“ sowie SPSS muss vorhanden sein.

3 Voraussetzungen

Ordnerstruktur angelegt und Arbeitsdateien entsprechend gespeichert.

4 Beschreibungen der Workflows

Die Gewichtungsvariablen werden jeweils unter dem Namen *wxx* geführt, wobei *xx* für Ziffern zwischen *1* und meist *99*, oder der Endung *ex* stehen kann. Hier muss nur das Variablenlabel angepasst werden entsprechend der jeweiligen Vorgängerstudie.

Es muss einzeln kontrolliert werden, ob die jeweilige Variable vorhanden und inhaltlich entsprechend gelabelt ist, oder ob sich die Definition eines Gewichts verändert hat. Dazu wird mittels des weight-Befehls in SPSS eine Auszählung der Ländervariablen vorgenommen. Im Falle neuer oder anderweitig veränderter Gewichte bzw. Gewichtungsdefinitionen muss der Kollektionsverantwortliche informiert werden, sodass ggfs. eine Anpassung des Labels des Gewichtes oder eine Rückfrage beim Datengeber durchgeführt bzw. gestellt werden kann.

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Erkennen und Bearbeiten von Missings

Version 4, 11.12.2023, Änderungen von Boris Heizmann
Version 3, 14.09.2023, Änderungen von Kerstin Beck
Version 2, Mai 2023, Änderungen von ET
Version 1, 14.09.2023, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Aufgrund von Fehlern bei der Datenübertragung, aber auch durch die im Fragebogen angelegte Filterführung kann es bei einzelnen Fragen zu fehlenden Werten kommen. Diese gilt es zu finden und gemäß den entsprechenden Konventionen zu rekodieren und zu kennzeichnen.

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Arbeitsordner auf dem Laufwerk „I:“ sowie SPSS muss vorhanden sein.

3 Voraussetzungen

Ordnerstruktur angelegt und Arbeitsdateien entsprechend gespeichert.

4 Beschreibungen der Workflows

Die inhaltlichen Fragen sind - im Gegensatz zu den demographischen Fragen - nicht sehr stark standardisiert. Daher gibt es kein festes Vorgehen zu den jeweiligen Fragen, es kommt jedoch oft vor, dass Variablen aus verschiedenen Gründen fehlende Werte aufweisen. Diese sind so zu bearbeiten, dass Nutzende direkt ablesen können, warum fehlende Werte existieren, sofern diese Information vorhanden ist.

Wichtig: JEDER Befehl muss nach der Ausführung noch einmal kontrolliert werden: Es kann passieren, dass sich im Fragebogen kleinste Details geändert haben, wie z.B. die Bezeichnungen der Arbeitsfelder oder dass beim Geschlecht nicht mehr nur männlich und weiblich, sondern auch etwas anderes wählbar wird. Um zu vermeiden, dass Variablen nach veralteten Standards aufbereitet werden, muss also stets die jeweilige Variable mit dem Fragebogen bzw. den Originalvariablen abgeglichen und auf Korrektheit geprüft werden.

4.1 System- Missings

Es kann aufgrund von Datenübertragungsfehlern, aufgrund nicht gelabelter Filter oder aufgrund anderweitig fehlender Informationen zu fehlenden Werten kommen. Daher muss der gesamte Datensatz auf system missings geprüft werden. Dies kann in SPSS z. B. einfach über die Codebuch-Funktion von SPSS (Analyze > Reports > Codebook) und die Suche (Edit > Find) nach dem Wort „system“ in dem erzeugten Codebuch vorgenommen werden, oder durch ein Markieren der

Variablen in der Variablenansicht und per Rechtsklick angeforderte „Deskriptive Statistiken“, in welchen ebenfalls fehlende Werte ersichtlich werden.

4.2 Filter-Missings

Zusätzlich müssen auch die Filter-Missings geprüft werden. Wenn im Fragebogen ein Filter formuliert ist, werden einzelne Personen von einer Frage ausgeschlossen, zum Beispiel: Die Frage, welchen Studiengang eine Person gewählt hat, wird nur denjenigen Befragten gestellt, welche angaben, dass sie studiert haben. Immer, wenn Personen durch einen solchen Filter ausgeschlossen werden, müssen die dadurch entstandenen fehlenden Werte inhaltlich als solche ausgewiesen werden. Dafür wird die Variable rekodiert und das entsprechende Wertelabel der Variable angepasst. Die Festlegung des Wertes, dem die üblicherweise als system-missing vorliegenden fehlenden Werte zugewiesen werden, hängt davon, wie viele Stellen die höchste Ausprägung hat. Ist es nur eine Ziffer, werden die fehlenden Werte mit 9 kodiert, sind es zwei, werden sie mit 99 kodiert usw. Es kann jedoch bei einstelligen Variablen sein, dass die 9 anderweitig belegt ist, z.B. mit „Don‘ t know“. Auch dann müssen die missings die 99 (oder auch mit „9999“ falls „99“ und auch „999“ bereits belegt sind) kodiert werden.

Im nächsten Schritt werden die Wertelabels definiert. Dafür muss untersucht werden, wie der Filter genau formuliert wird. Wenn im Fragebogen beispielsweise „ASK Q1 IF D8=1, OTHERS GO TO Q2“ geschrieben steht, muss dieser Filter in der am kürzesten möglichen Form in das Label formuliert werden. Dabei stellt man sich die Frage, welche Personen von der Frage ausgeschlossen werden, was hier all jene sind, die nicht die Ausprägung 1 in der Variable d8 aufweisen. Es wird also das Gegenteil des Filters formuliert und dann wie folgt in das Label gesetzt:

```
ADD VALUE LABELS q1.1 to q1.8
    9 "Inap. (not 1 in d8)".
```

Das Kürzel Inap. steht für inappropriate (unpassend) und kündigt an, dass die in Klammern benannten Personen herausgefiltert wurden.

Es kann auch vorkommen, dass mehrere Variablen innerhalb eines Filters stehen (z.B.: „ASK Q2, Q3, Q4 ONLY IF D2=2 TO 4 OR D9=2“). In diesem Fall müssen die beiden Komponenten miteinander verknüpft werden, sodass hier die Formulierung „Inap. (not 2 to 4 in d2 and not 2 in d9)“ die ideale Formulierung wäre. Hier ist darauf zu achten, ob die einzelnen Teile mit einem „und“ oder einem „oder“ verbunden werden müssen. Gelegentlich kann es auch übersichtlicher sein, den Filter ohne „not“ zu formulieren. Im Beispiel von eben wäre das „Inap. (1 in d2 and 1 in d9)“, doch nur dann, wenn d2 nur Werte von 1-4 und d9 nur die Werte 1 und 2 aufweist. Bei der Entscheidung für eine Variante kann auch relevant sein, dass SPSS nur Wertelabel von einer Länge bis zu 120 Zeichen unterstützt, und bei sehr komplexen Filterführungen daher die im Sinne der verwendeten Zeichen kürzere Variante gewählt werden muss. In jedem Fall muss die Korrektheit des Filters durch Kreuztabellierung der entsprechenden Variablen überprüft werden. Es können sowohl in den Daten als auch in dem im Fragebogen selbst genannten Filter Fehler auftreten, welche dann dokumentiert bzw. korrigiert werden müssen.

Sind die fehlenden Werte rekodiert und gelabelt, müssen sie noch im Datensatz als fehlend markiert werden. Dafür findet sich in der Variablenansicht die Spalte mit dem Titel „Fehlend“. Die Markierung im Datensatz geschieht über den Befehl (hier nur für den Wert „9“)

```
MISSING VALUES q1.1 to q1.8 (9).
```

Es kann auch vorkommen, dass ohne ersichtlichen Grund fehlende Werte im Datensatz auftreten, entweder durch einen Fehler im Datensatz oder durch einen nicht dokumentierten Filter. In diesen und ähnlichen Fällen muss mit der kollektionsverantwortlichen Person Rücksprache gehalten werden.

gesis Leibniz-Institut für Sozialwissenschaften

How-to – Meta- und Demographiedaten im Datensatz (Paradaten wie IDs etc., Variablennamen und -label)

Version 4, 11.12.2023, Änderungen von Boris Heizmann
Version 3, 14.09.2023, Änderungen von Kerstin Beck
Version 2, Mai 2023, Änderungen von ET
Version 1, 14.09.2023, unter Mitwirkung von L-M F, LL und MM

Inhaltsverzeichnis

1 Einleitung

Für den zu veröffentlichenden Datensatz müssen für eine reibungslose und sinnvolle Nutzung die enthaltenen Variablen entsprechend aufbereitet sein. Aus diesem Grunde werden die Variablen- und Valuelabels aller Variablen (Survey- und Demographiefragen) geprüft und, falls notwendig, überarbeitet. Zusätzlich werden auch neue Variablen aus bereits vorhandenen gebildet oder Recodierungen vorgenommen. Eine Überprüfung bzw. Anpassung der verwendeten Skalenniveaus findet auch statt.

Die zu verwendenden Syntaxteile werden im Folgenden beispielhaft vorgestellt. Es ist zu beachten, dass beim Kopieren von Syntax aus Word Formatierungsschwierigkeiten z.B. bei Anführungszeichen auftreten, daher ist das Kopieren aus diesem Dokument nicht empfehlenswert. Kopiert werden sollte also nur von Syntax zu Syntax in SPSS.

2 Zuständigkeiten und Berechtigungen

Die Bearbeitung der Datensätze wird vom Kollektionsverantwortlichen und ggfs. weiteren Personen vorgenommen.

Zugriff auf die entsprechenden Arbeitsordner auf dem Laufwerk „I:“ sowie SPSS muss vorhanden sein.

3 Voraussetzungen

Ordnerstruktur angelegt und Arbeitsdateien entsprechend gespeichert.

Variablen aus dem Fragebogen mit ‚RGPS‘ im Namen müssen nicht aufbereitet werden, da diese nicht in die Veröffentlichung aufgenommen werden.

4 Beschreibungen der Workflows

Wichtig: JEDER Befehl muss nach der Ausführung noch einmal kontrolliert werden: Es kann passieren, dass sich im Fragebogen kleinste Details geändert haben, wie z.B. die Bezeichnungen der Arbeitsfelder oder dass beim Geschlecht nicht mehr nur männlich und weiblich, sondern auch etwas anderes wählbar wird. Um zu vermeiden, dass Variablen nach veralteten Standards aufbereitet werden, muss also stets die jeweilige Variable mit dem Fragebogen bzw. den Originalvariablen abgeglichen und auf Korrektheit geprüft werden.

4.1 Aufbereitung der Fragen aus dem Fragebogen

Die in diesem Kapitel (4.1) vorgestellten Vorgehensweisen dienen nur der Erläuterung und Hintergrundinformation. Die eigentlichen Bearbeitungsschritte werden im nächsten Kapitel (4.2) zu den jeweiligen Variablen erläutert.

Die Aufbereitung der Fragen aus dem Fragebogen ist der am wenigsten regelmäßige Teil, da sich hier am seltensten Fragen doppeln. Es kann vorkommen, dass einzelne Fragestellungen oder ganze Blöcke dem entsprechen, was in vorherigen Studien abgefragt wurde. In dem Fall kann die Syntax zu der entsprechenden Frage aus der Syntax des Vorgängers (also zu dem Datensatz gehörend, der die gleiche Fragestellung beinhaltet) übernommen werden. Es kann jedoch immer vorkommen, dass einzelne Elemente der Frage oder der Antwortmöglichkeit modifiziert wurden, weshalb eine detaillierte Prüfung zwingend notwendig ist. Speziell, wenn im Datensatz oder Fragebogen vermerkt ist, dass eine modifizierte Version (M) der Frage gestellt wurde oder neue Antwortmöglichkeiten (N) eingeführt wurden, muss diese Kontrolle erfolgen, weil unter Umständen Labels verändert, oder Ausprägungen hinzugefügt oder gestrichen wurden. Es kann sich jedoch nicht auf (M) und (N) verlassen werden, sodass eine genaue Prüfung immer notwendig ist.

Ansonsten ist hier folgendermaßen vorzugehen:

a) Einige der Variablen werden nur mit den Ausprägungen „*Not mentioned*“ auf der 0 und der Antwort auf die Frage auf der 1 gelabelt. Hier kann ein einzelner Befehl genutzt werden, der all diese Variablen zusammenfasst und den Ausprägungen *0* jeweils das Label „*Not mentioned*“ zuordnet. Diese Veränderung soll alle Labels betreffen, die bei der Ausprägung *0* das Label „*no [Ausprägung bei 1]*“ aufweisen. Der fertige Befehl kann dann so aussehen:

```
ADD VALUE LABELS
VARIABLENNAME to VARIABLENNAME
0 "Not mentioned".
```

Beachten: Vorsicht bei der Verwendung von „to“: Hier werden alle Variablen, die zeilenweise im Datensatz zwischen q10.1 und q12.8 liegen (vgl. Variablenansicht), mit dem Label versehen. Es muss also sichergestellt werden, dass bspw. q11 korrekterweise ebenfalls ein 0 „Not mentioned“ erhalten muss!

[Mit „add value labels“ werden einzelne Werte gelabelt, der Befehl „value labels“ (ohne „add“) definiert alle Werte einer Variablen komplett neu. Würde hier also das „add“ fehlen, dann wäre auf allen genannten Variablen nur noch die 0 gelabelt.]

Eine weitere Anpassung der Wertelabels ergibt sich, wenn „Don’t know“-Antworten im Fragebogen mit (DO NOT READ OUT) vermerkt sind. Diese Anweisung wird ebenfalls in die Wertelabels eingefügt, falls sie in den Quellvariablenlabels nicht vorhanden ist.

```
ADD VALUE LABELS
VARIABLENNAME to VARIABLENNAME
DK-WERT "Don't know (DO NOT READ OUT)".
```

b) Manchmal sind die Wertelabels ab einer bestimmten Zeichenzahl abgeschnitten. Daraus ergibt sich die Notwendigkeit, dass alle Labels kontrolliert und mit dem Fragebogen verglichen werden müssen. Sollte ein Label nicht vollständig angezeigt werden, muss es neu

angelegt werden, gegebenenfalls mit … am Ende, falls das Label mehr als 120 Zeichen (SPSS-Limit) aufweist. Dies geschieht wiederum mit dem add value labels-Befehl.

c) Neben der Kontrolle der inhaltlichen Fragen müssen auch gebildeten Variablen wie totals im Rahmen der Aufbereitung überprüft werden. Für diese Kontrolle bieten sich Kreuztabellen an, um zu prüfen, welche Variablen in die jeweiligen totals eingeflossen sind. Für die Überprüfung komplexerer totals, in die beispielsweise zwei verschiedene Variablen eingeflossen sind, sind dreidimensionale Kreuztabellen geeignet. Mit diesen kann erkannt werden, wie die totals zugeordnet sind und ob diese Zuordnung korrekt ist, ob z.B. an den richtigen Stellen nur Nullen in den Zellen bzw. leere Zellen vorhanden sind. Für noch komplexere totals müssen Hilfsvariablen gebildet werden. Dies ist im Zweifel mit der kollektionsverantwortlichen Person abzustimmen.

4.2 Aufbereitung der demographischen Variablen

Für die demographischen Variablen ähnelt das Vorgehen in quasi allen Fällen stark dem der Vorgängerstudie. Hier ist es möglich, die Zeilen aus einer alten Syntax zu kopieren, an einigen, wenigen Stellen zu modifizieren und dann auszuführen. Es empfiehlt sich, sich an einer möglichst neuen Ausgabe der Studie zu orientieren.

Die folgenden Variablen beziehen sich auf Personenbefragungen. Studien mit anderen Analyseeinheiten, etwa Betriebe oder anderen Organisationen, enthalten u. U. andere „Demographie“-Variablen.

4.2.1 Alter

Die für die Alter-Variablen notwendige Variable ist im Datensatz vorhanden und wird in den für die Datenkollektion üblichen Standardnamen umbenannt. Sie wird als Grundlage genommen, um rekodierte Altersvariablen zu erzeugen. Die Veränderung, die erfolgen muss, ist die Anpassung der Value Labels der d1-Variable. Hier muss das Alter der jüngsten und der ältesten Befragungsperson angepasst werden, die mit dem FREQUENCIES-Befehl angezeigt werden können. Sollte es untypischerweise vorkommen, dass der jüngste Teilnehmende älter ist als 15 Jahre, müssen auch die d1r1- und d1r2-Labels angepasst werden. Zum Schluss wird die Originalvariable mit den rekodierten Variablen kreuztabelliert, um deren Korrektheit sicherzustellen.

```
RENAME VARIABLES (ORIGINALNAME=NEUERNAME).
FREQ NEUERNAMEN.
VALUE LABELS NEUERNAME
    15    "15 years" => Anpassen an den jüngsten Teilnehmenden der Erhebung
    98    "98 years". => Anpassen an den ältesten Teilnehmenden der Erhebung
VALUE LABELS NEUERNAMEr1
    1       "15 - 24 years" => Sollte der jüngste Teilnehmende keine 15 Jahre alt sein: bitte anpassen!
```

```
    2     "25 - 39 years"
    3     "40 - 54 years"
    4     "55 years and older".
VALUE LABELS NEUERNAMEr2
    1     "15 - 24 years" => Sollte der jüngste Teilnehmende keine 15 Jahre alt sein: anpassen!
    2     "25 - 34 years"
    3     "35 - 44 years"
    4     "45 - 54 years"
    5     "55 - 64 years"
    6     "65 years and older".
VARIABLE LABELS
    NEUERNAME    "AGE EXACT"
    NEUERNAMEr1  "AGE RECODED - 4 CATEGORIES"
    NEUERNAMEr2  "AGE RECODED - 6 CATEGORIES".
EXECUTE.
```

4.2.2 Geschlecht

Zumeist ist der Befehl für die Geschlechtsvariable bisher unverändert geblieben und im originalen Variablennamen belassen. Bei manchen Umfragen kann die Geschlechtsvariable jedoch mehr als zwei Ausprägungen haben:

```
variable labels ORIGINALNAME "GENDER".
recode ORIGINALNAME (4=3). => Hiermit werden der Wert 4 „Prefer not to say“ mit dem von 3 „In another way“ kombiniert, um ein etwaiges Reidentifikationsrisiko zu verringern.
execute.
value labels ORIGINALNAME
    1 "Male"
    2 "Female"
    3 "In another way / Prefer not to say".
    execute.
```

4.2.3 Nationalität

Die Nationalitätsvariable bildet die Staatsangehörigkeiten der Befragten ab. Es ist nicht zwingend notwendig, dass die Staatsangehörigkeiten den Staaten entsprechen, in denen die Personen befragt wurden. Auch hier gibt es eine festgelegte Zuordnung der Werte, die sich nicht ändern sollte. Es werden lediglich die Variablenlabels angepasst ("NATIONALITY: LÄNDERNAME"). Hier ist dann noch eine Kontrolle (Kreuztabellierung mit der Ländervariablen) durchzuführen, um zu überprüfen, ob die Fälle korrekt in die Variable eingeordnet wurden, ob also die Mehrheit einer Nationalität sich tatsächlich im entsprechenden Land befindet.

4.2.4 Bildungsjahre

Die Aufbereitung ähnelt der Aufbereitung der Altersvariable: Auch hier müssen der geringste und der höchste Wert in den value labels angepasst werden. Die rekodierten Varianten verändern sich hier im Normalfall nicht, werden jedoch durch Kreuztabellierung mit der Originalvariablen geprüft.

```
FREQ ORIGINALNAME.
ADD VALUE LABELS ORIGINALNAME
    0    "Refusal"
    2    "2 years"
    72    "72 years".
```

4.2.5 Berufsgruppe

Diese Variablen ordnen den Befragten die Berufsgruppe zu, in der sie beschäftigt sind. Auch hier ist die Syntax festgelegt. Wie es auch bei einigen q-Fragen nötig ist, ist hier ein ADD VALUE LABELS Teil der Syntax, da die Labels einiger Ausprägungen standardmäßig zu lang sind, um korrekt eingetragen zu werden. Ebenfalls ist es hier allgemein nötig, jedes Mal zu kontrollieren, ob die Gruppen korrekt benannt wurden, d.h. ob sich im Erhebungsinstrument im Fragebogen etwas verändert hat. Etwaige mitgelieferte Gruppierungsvariablen werden übernommen, nachdem ihre Korrektheit durch Kreuztabellierung sichergestellt wurde.

4.2.6 Art des Wohnorts

In der im Normalfall unveränderten Variable wird die Größe der Ortschaft zugeordnet.

4.2.7 Telefonische Erreichbarkeit

Diese Variablen behandeln die Wege, auf denen die Befragten telefonisch erreichbar sind. Auch hier ändert sich zwischen den Befragungen im Normalfall nichts. Es wird jedoch eine Variable gebildet, welche Festnetz und Mobilfunk kombiniert. Diese Variablen enthalten keine fehlenden Werte, sollte dies abweichend anders sein, ist die Generierung entsprechend anzupassen; in jedem Fall ist auch hier eine Prüfung anhand der Quellvariablen erforderlich.

```
COMPUTE NEUEVARIABLE=0.
FORMATS NEUEVARIABLE (F1.0).
IF (FESTNETZ eq 1 and MOBILFUNK eq 2) NEUEVARIABLE =1.
IF (FESTNETZ eq 2 and MOBILFUNK eq 1) NEUEVARIABLE =2.
IF (FESTNETZ eq 1 and MOBILFUNK eq 1) NEUEVARIABLE =3.
VALUE LABELS
    NEUEVARIABLE
    1 "Mobile only"
    2 "Landline only"
    3 "Mobile and landline".
```

4.2.8 Haushaltsgröße

Bei den Haushaltsgrößen verhält es sich wie bei den Alters- und Bildungsjahrvariablen. Die Labels müssen dahingehend angepasst werden, dass die geringste und die höchste Ausprägung ausgezeichnet werden. Im Normalfall ist hier nur die Abänderung des höchsten Werts nötig.

4.2.9 Regionsvariable (basiert auf NUTS Nomenclature of Territorial Units for Statistics der EU)

Die NUTS-Variablen sind Regionalvariablen, die für jedes Land und teilweise auf mehreren Ebenen separat angelegt werden (NUTS 1 umfassen z.B. Bundesländer, NUTS 2 dann Städte und Landkreise, vgl. https://ec.europa.eu/eurostat/de/web/nuts/background für mehr Informationen zu diesem Klassifikationsschema)

Darüber hinaus werden die Labels so angepasst, dass besondere Zeichen der einzelnen Sprachen (z.B. ä, ö, ü im deutschen) so umformatiert werden, dass sie auch für diejenigen korrekt angezeigt werden, die die jeweiligen Zeichen nicht nutzen. Dafür werden sie vereinfacht dargestellt (ä → ae, ö → oe, ü → ue, außerdem werden Diakritika wie accents etc. entfernt: é → e, außerdem: æ → ae; ø → o; å → a; ç, č → c.

Auch hier verändern sich die Befehle im Normalfall nicht. Die NUTS-Klassifikation wird durch EUROSTAT festgelegt und ändert sich in einem Rhythmus von ca. drei Jahren in unterschiedlich großem Ausmaß. Diese Veränderungen sind jedoch nicht notwendigerweise in den Erhebungen umgesetzt. Oft geschieht dies mit Verzögerung und nicht für alle Länder gleichzeitig. Daher wird stets geprüft, ob die Benennungen und Verteilungen der fertig aufbereiteten Regionsvariablen denjenigen aus der originalen Quelldatei entsprechen. Es wird also für jede Variable geprüft, ob die Aufbereitungssyntax noch aktuell ist. Zusätzlich werden die Variablen für den Quell- und den finalen Datensatz ausgezählt und verglichen.

Ein weiterer Schritt, der vollzogen wird, bezieht sich auf die nichtbinäre Geschlechterkategorie. Für Befragte in dieser Kategorie wird aus Datenschutzgründen die Regionsinformation entfernt, und auch Fälle gelöscht, wenn sie nur in einer einzigen Region im jeweiligen Land vorfindlich und damit über Aggregatdarstellungen der Rohdaten zuordenbar wären.

Zunächst wird eine Kreuztabelle erstellt, die Regionen und die Geschlecht-Variable kreuztabelliert. Bei Befragten mit der Ausprägung 3 auf der Gendervariablen werden die Regionsvariablen auf die Ausprägung „997“ gesetzt.

Anschließend wird die Ausgabe der Kreuztabelle überprüft und wenn in einem Land nur eine Region Personen mit gender=3 enthält, dann werden diese Fälle (ggfs. auch nur ein einzelner Fall) mit einem geeigneten SELECT IF-Befehl entfernt. Die Anzahl der durch diesen Vorgang entfernten Fälle wird später Stelle im Readme vermerkt.

4.3 Anpassung der Skalenniveaus

Im vorletzten Schritt werden mittels des VARIABLE LEVEL-Befehls die Skalenniveaus der Variablen festgelegt. Da die meisten Variablen erfahrungsgemäß nominalskaliert sind, werden zunächst alle Variablen auf ein nominales Skalenniveau gesetzt. Danach wird die Liste Variable für Variable durchsucht. Die ordinalskalierten bzw. die metrischen Variablen werden dabei in jeweils einen einzelnen Befehl zusammengefasst und ihren Skalenniveaus zugeordnet.

4.4 Speichern des fertigen Datensatzes

Im letzten Schritt wird der fertige Datensatz gespeichert. Dafür wird ein Befehl angelegt, in dem alle relevanten Variablen sowie der Dateipfad, unter dem der Datensatz gespeichert werden soll, verzeichnet sind. In dem in diesem Beispiel genutzten Datensatz sähe das wie folgt aus:

```
SAVE OUTFILE = " PFAD\STUDIENNUMMER_v1-0-0.sav"
            /keep
VARIABLENLISTE.
```

Statt VARIABLENLISTE sind alle Variablen, die im finalen Datensatz vorkommen sollen, einzugeben. Sie wird einer Vorgängerstudie entnommen, aber auf den aktuellen Datensatz angepasst. Das sind im Detail die Archivvariablen, gefolgt von den inhaltlichen Fragen und den demographischen Fragen, in der Reihenfolge der Quelldaten (es muss auch geprüft werden, ob alle relevanten Variablen der Quelldatei mit abgespeichert wurden). Die Variablen, die auf das Kürzel RGPS enden, werden jeweils nicht in den Speicherbefehl aufgenommen. Die letzten Variablen, die gelistet werden, sind in jedem Fall die Gewichtungsvariablen. Speziell bei den Gewichten ist es nicht gefordert, sie in chronologischer Reihenfolge zu speichern, sondern in der Reihenfolge, in der sie bereits im Quelldatensatz abgelegt sind. Eine Veränderung ist demnach nicht nötig.

Das Ziel ist es, eine Syntax erstellt zu haben, welche komplett fehlerfrei durchläuft, vom Öffnen der Quelldatei bis zum Speichern der finalen Zieldatei. Außerdem muss eine Übersicht über den finalen Datensatz mithilfe des CODEBOOK-Befehls ausgegeben und durchgegangen werden, um auf diese Weise nach Auffälligkeiten zu suchen.