Statistical Moments and Beyond

Outline: Statistical Moments and Beyond

- What is statistics, and what are the importance of statistics?
- Data exploration, manipulation.
- Examining distributions
- Basic descriptive data visualisation

An introduction to

tidy data management **with R**

{Tidy} data management

An Introduction and The Why

{Tidy} data management: The Why

{Tidy} data management: The Why

O1_LONGITUDE	O2_LATITUDE	One of main goals in life has been to make my parents proud	Feeling of happiness	State of health (subjective)	How often do you pray	Year of birth	Age recoded (6 intervals)	Employment status	country code
-1.537885	55.078311	Agree strongly	Very happy	Very good	Never, practically never	2000	16-24	Part time (less than 30 hours a week)	GBR
-1.532713	55.074637	Agree	Quite happy	Poor	Once a day	1950	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.608998	55.132695	Agree	Very happy	Poor	Never, practically never	1971	45-54	Part time (less than 30 hours a week)	GBR
-1.608338	55.1344	Don't know	Very happy	Poor	Several times each week	1988	25-34	Unemployed	GBR
-1.602372	55,128,665	Agree	Quite happy	Fair	Never, practically never	1966	55-64	Self employed	GBR
-1.603687	55.12734	Agree	Very happy	Good	Never, practically never	1947	65 and more years	Retired/pensioned	GBR
-1.601168	55.134646	Disagree	Not very happy	Good	Once a day	1944	65 and more years	Retired/pensioned	GBR
-1.602418	55.130346	Agree	Quite happy	Fair	Never, practically never	1954	65 and more years	Retired/pensioned	GBR
-1.608852	55.13645	Agree	Very happy	Good	Several times each week	1949	65 and more years	Retired/pensioned	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.458406	55,046.29	Disagree	Quite happy		Several times each week	1937	65 and more years	Retired/pensioned	GBR
-1.461515	55.051.383	Agree	Not very happy		Several times each week	1960	55-64	Retired/pensioned	GBR
-1.465077	55.051893	Agree	N/A	Good	Less often	1947	N/A	Retired/pensioned	GBR
-1.467017	55.054274	Agree	Quite happy	Good	Never, practically never	1978	35-44	Part time (less than 30 hours a week)	GBR
-1.464305	55.054585	Agree strongly	Very happy	Very good	Never, practically never	1951	65 and more years	Retired/pensioned	GBR
-1.461513	55.044554	Strongly disagree	Quite happy	Good	Less often	1982	35-44	Full time (30 hours a week or more)	GBR
-1.452616	55.043975	NA	Very happy	Good	Less often	1939	65+	Retired/pensioned	GBR
-1.454677	55.043175	Agree	Very happy	Good	Never, practically never	1950	65 and more years	Retired/pensioned	GBR
-1.505645	54.995998	Agree	Very happy	Fair	Never, practically never	1906	65 and more years	Retired/pensioned	GBR

An irrelevant column

{Tidy} data management: The Why

O1_LONGITUDE	O2_LATITUDE	One of main goals in life has been to make my parents proud	Feeling of happiness	State of health (subjective)	How often do you pray	Year of birth	Age recoded (6 intervals)	Employment status	country code
-1.537885	55.078311	Agree strongly	Very happy	Very good	Never, practically never	2000	16-24	Part time (less than 30 hours a week)	GBR
-1.532713	55.074637	Agree	Quite happy	Poor	Once a day	1950	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.608998	55.132695	Agree	Very happy	Poor	Never, practically never	1971	45-54	Part time (less than 30 hours a week)	GBR
-1.608338	55.1344	Don't know	Very happy	Poor	Several times each week	1988	25-34	Unemployed	GBR
-1.602372	55,128,665	Agree	Quite happy	Fair	Never, practically never	1966	55-64	Self employed	GBR
-1.603687	55.12734	Agree	Very happy	Good	Never, practically never	1947	65 and more years	Retired/pensioned	GBR
-1.601168	55.134646	Disagree	Not very happy	Good	Once a day	1944	65 and more years	Retired/pensioned	GBR
-1.602418	55.130346	Agree	Quite happy	Fair	Never, practically never	1954	65 and more years	Retired/pensioned	GBR
-1.608852	55.13645	Agree	Very happy	Good	Several times each week	1949	65 and more years	Retired/pensioned	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.458406	55,046.29	Disagree	Quite happy		Several times each week	1937	65 and more years	Retired/pensioned	GBR
-1.461515	55.051.383	Agree	Not very happy		Several times each week	1960	55-64	Retired/pensioned	GBR
-1.465077	55.051893	Agree	N/A	Good	Less often	1947	N/A	Retired/pensioned	GBR
-1.467017	55.054274	Agree	Quite happy	Good	Never, practically never	1978	35-44	Part time (less than 30 hours a week)	GBR
-1.464305	55.054585	Agree strongly	Very happy	Very good	Never, practically never	1951	65 and more years	Retired/pensioned	GBR
-1.461513	55.044554	Strongly disagree	Quite happy	Good	Less often	1982	35-44	Full time (30 hours a week or more)	GBR
-1.452616	55.043975	NA	Very happy	Good	Less often	1939	65+	Retired/pensioned	GBR
-1.454677	55.043175	Agree	Very happy	Good	Never, practically never	1950	65 and more years	Retired/pensioned	GBR
-1.505645	54.995998	Agree	Very happy	Fair	Never, practically never	1906	65 and more years	Retired/pensioned	GBR

Duplicate observations

{Tidy} data management: The Why

O1_LONGITUDE	O2_LATITUDE	One of main goals in life has been to make my parents proud	Feeling of happiness	State of health (subjective)	How often do you pray	Year of birth	Age recoded (6 intervals)	Employment status	country code
-1.537885	55.078311	Agree strongly	Very happy	Very good	Never, practically never	2000	16-24	Part time (less than 30 hours a week)	GBR
-1.532713	55.074637	Agree	Quite happy	Poor	Once a day	1950	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.608998	55.132695	Agree	Very happy	Poor	Never, practically never	1971	45-54	Part time (less than 30 hours a week)	GBR
-1.608338	55.1344	Don't know	Very happy	Poor	Several times each week	1988	25-34	Unemployed	GBR
-1.602372	55,128,665	Agree	Quite happy	Fair	Never, practically never	1966	55-64	Self employed	GBR
-1.603687	55.12734	Agree	Very happy	Good	Never, practically never	1947	65 and more years	Retired/pensioned	GBR
-1.601168	55.134646	Disagree	Not very happy	Good	Once a day	1944	65 and more years	Retired/pensioned	GBR
-1.602418	55.130346	Agree	Quite happy	Fair	Never, practically never	1954	65 and more years	Retired/pensioned	GBR
-1.608852	55.13645	Agree	Very happy	Good	Several times each week	1949	65 and more years	Retired/pensioned	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.458406	55,046.29	Disagree	Quite happy		Several times each week	1937	65 and more years	Retired/pensioned	GBR
-1.461515	55.051.383	Agree	Not very happy		Several times each week	1960	55-64	Retired/pensioned	GBR
-1.465077	55.051893	Agree	N/A	Good	Less often	1947	N/A	Retired/pensioned	GBR
-1.467017	55.054274	Agree	Quite happy	Good	Never, practically never	1978	35-44	Part time (less than 30 hours a week)	GBR
-1.464305	55.054585	Agree strongly	Very happy	Very good	Never, practically never	1951	65 and more years	Retired/pensioned	GBR
-1.461513	55.044554	Strongly disagree	Quite happy	Good	Less often	1982	35-44	Full time (30 hours a week or more)	GBR
-1.452616	55.043975	NA	Very happy	Good	Less often	1939	65+	Retired/pensioned	GBR
-1.454677	55.043175	Agree	Very happy	Good	Never, practically never	1950	65 and more years	Retired/pensioned	GBR
-1.505645	54.995998	Agree	Very happy	Fair	Never, practically never	1906	65 and more years	Retired/pensioned	GBR

Outliers

{Tidy} data management: The Why

Structural Errors

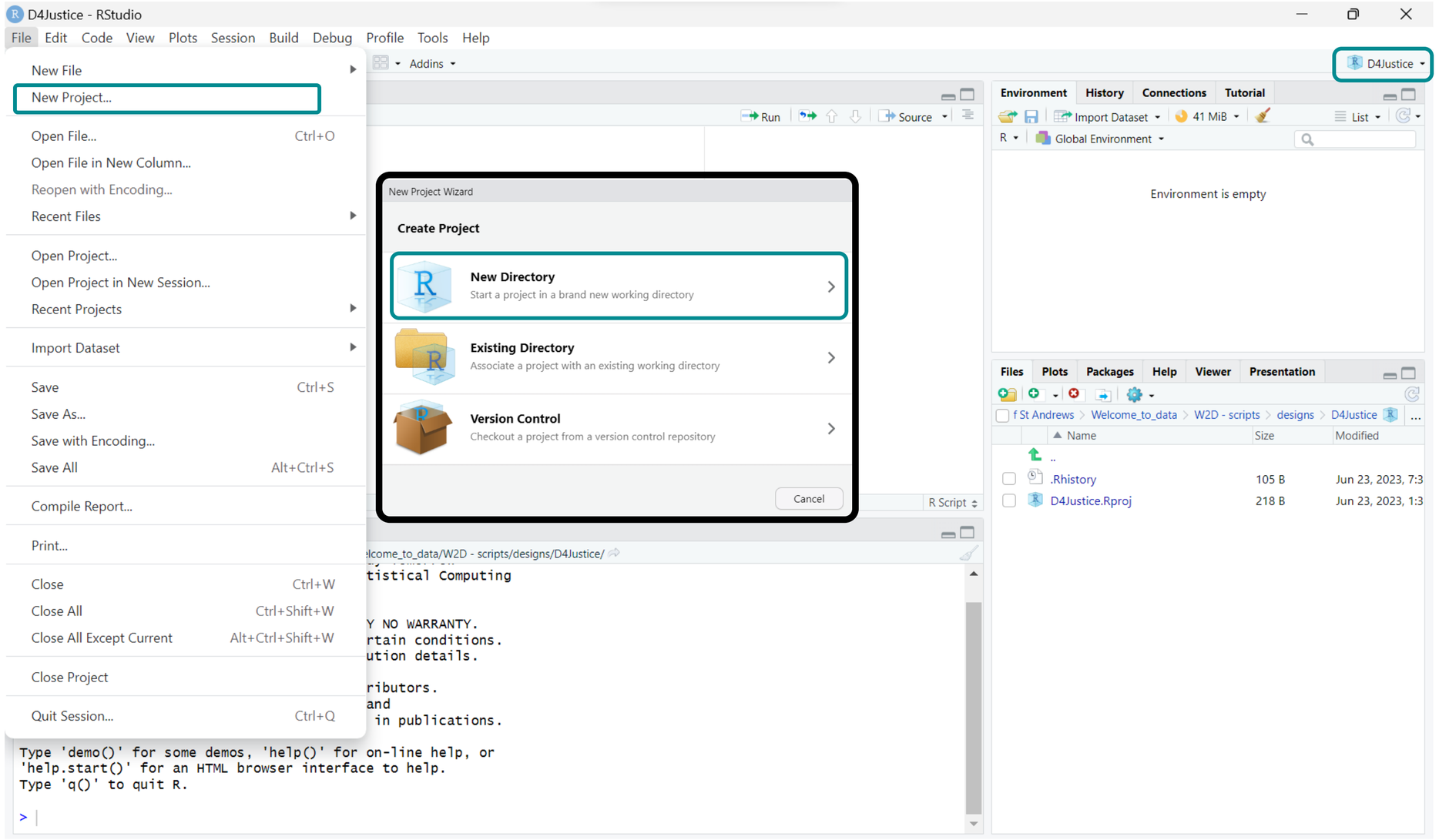
O1_LONGITUDE	O2_LATITUDE	One of main goals in life has been to make my parents proud	Feeling of happiness	State of health (subjective)	How often do you pray	Year of birth	Age recoded (6 intervals)	Employment status	country code
-1.537885	55.078311	Agree strongly	Very happy	Very good	Never, practically never	2000	16-24	Part time (less than 30 hours a week)	GBR
-1.532713	55.074637	Agree	Quite happy	Poor	Once a day	1950	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.608998	55.132695	Agree	Very happy	Poor	Never, practically never	1971	45-54	Part time (less than 30 hours a week)	GBR
-1.608338	55.1344	Don't know	Very happy	Poor	Several times each week	1988	25-34	Unemployed	GBR
-1.602372	55,128,665	Agree	Quite happy	Fair	Never, practically never	1966	55-64	Self employed	GBR
-1.603687	55.12734	Agree	Very happy	Good	Never, practically never	1947	65 and more years	Retired/pensioned	GBR
-1.601168	55.134646	Disagree	Not very happy	Good	Once a day	1944	65 and more years	Retired/pensioned	GBR
-1.602418	55.130346	Agree	Quite happy	Fair	Never, practically never	1954	65 and more years	Retired/pensioned	GBR
-1.608852	55.13645	Agree	Very happy	Good	Several times each week	1949	65 and more years	Retired/pensioned	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.458406	55,046.29	Disagree	Quite happy		Several times each week	1937	65 and more years	Retired/pensioned	GBR
-1.461515	55.051.383	Agree	Not very happy		Several times each week	1960	55-64	Retired/pensioned	GBR
-1.465077	55.051893	Agree	N/A	Good	Less often	1947	N/A	Retired/pensioned	GBR
-1.467017	55.054274	Agree	Quite happy	Good	Never, practically never	1978	35-44	Part time (less than 30 hours a week)	GBR
-1.464305	55.054585	Agree strongly	Very happy	Very good	Never, practically never	1951	65 and more years	Retired/pensioned	GBR
-1.461513	55.044554	Strongly disagree	Quite happy	Good	Less often	1982	35-44	Full time (30 hours a week or more)	GBR
-1.452616	55.043975	NA	Very happy	Good	Less often	1939	65+	Retired/pensioned	GBR
-1.454677	55.043175	Agree	Very happy	Good	Never, practically never	1950	65 and more years	Retired/pensioned	GBR
-1.505645	54.995998	Agree	Very happy	Fair	Never, practically never	1906	65 and more years	Retired/pensioned	GBR

{Tidy} data management: The Why

Missing Values

O1_LONGITUDE	O2_LATITUDE	One of main goals in life has been to make my parents proud	Feeling of happiness	State of health (subjective)	How often do you pray	Year of birth	Age recoded (6 intervals)	Employment status	country code
-1.537885	55.078311	Agree strongly	Very happy	Very good	Never, practically never	2000	16-24	Part time (less than 30 hours a week)	GBR
-1.532713	55.074637	Agree	Quite happy	Poor	Once a day	1950	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.535722	55.073129	Agree	Quite happy	Good	Several times each week	1952	65 and more years	Retired/pensioned	GBR
-1.608998	55.132695	Agree	Very happy	Poor	Never, practically never	1971	45-54	Part time (less than 30 hours a week)	GBR
-1.608338	55.1344	Don't know	Very happy	Poor	Several times each week	1988	25-34	Unemployed	GBR
-1.602372	55,128,665	Agree	Quite happy	Fair	Never, practically never	1966	55-64	Self employed	GBR
-1.603687	55.12734	Agree	Very happy	Good	Never, practically never	1947	65 and more years	Retired/pensioned	GBR
-1.601168	55.134646	Disagree	Not very happy	Good	Once a day	1944	65 and more years	Retired/pensioned	GBR
-1.602418	55.130346	Agree	Quite happy	Fair	Never, practically never	1954	65 and more years	Retired/pensioned	GBR
-1.608852	55.13645	Agree	Very happy	Good	Several times each week	1949	65 and more years	Retired/pensioned	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.200647	54.688283	Agree	Quite happy	Poor	Never, practically never	1993	25-34	Full time (30 hours a week or more)	GBR
-1.458406	55,046.29	Disagree	Quite happy		Several times each week	1937	65 and more years	Retired/pensioned	GBR
-1.461515	55.051.383	Agree	Not very happy		Several times each week	1960	55-64	Retired/pensioned	GBR
-1.465077	55.051893	Agree	N/A	Good	Less often	1947	N/A	Retired/pensioned	GBR
-1.467017	55.054274	Agree	Quite happy	Good	Never, practically never	1978	35-44	Part time (less than 30 hours a week)	GBR
-1.464305	55.054585	Agree strongly	Very happy	Very good	Never, practically never	1951	65 and more years	Retired/pensioned	GBR
-1.461513	55.044554	Strongly disagree	Quite happy	Good	Less often	1982	35-44	Full time (30 hours a week or more)	GBR
-1.452616	55.043975	NA	Very happy	Good	Less often	1939	65+	Retired/pensioned	GBR
-1.454677	55.043175	Agree	Very happy	Good	Never, practically never	1950	65 and more years	Retired/pensioned	GBR
-1.505645	54.995998	Agree	Very happy	Fair	Never, practically never	1906	65 and more years	Retired/pensioned	GBR

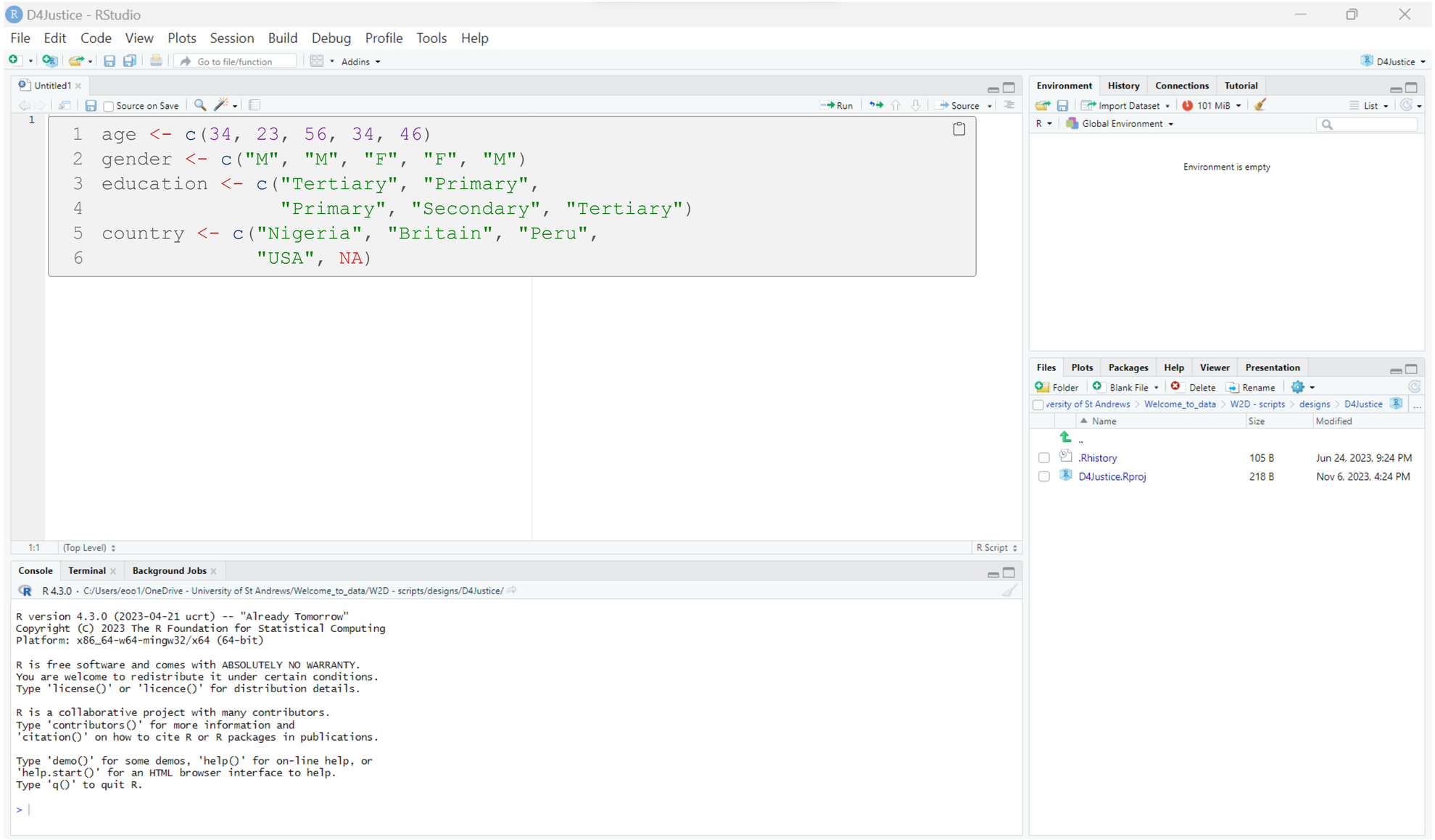
{Tidy} data management: The Why



{Tidy} data management

The How

Tidy data management: The How



Tidy data management: The How

```r
age <- c(34, 23, 56, 34, 46)
gender <- c("M", "M", "F", "F", "M")
education <- c("Tertiary", "Primary",
               "Primary", "Secondary", "Tertiary")
country <- c("Nigeria", "Britain", "Peru",
             "USA", NA)

participants <- data.frame (age, gender,
                            education, country)
```

Tidy data management: The How

```r
age <- c(34, 23, 56, 34, 46)
gender <- c("M", "M", "F", "F", "M")
education <- c("Tertiary", "Primary",
               "Primary", "Secondary", "Tertiary")
country <- c("Nigeria", "Britain", "Peru",
             "USA", NA)

participants <- data.frame (age, gender,
                            education, country)

dim (participants)
```

```
[1] 5 4
```

Tidy data management: The How

```r
age <- c(34, 23, 56, 34, 46)
gender <- c("M", "M", "F", "F", "M")
education <- c("Tertiary", "Primary",
               "Primary", "Secondary", "Tertiary")
country <- c("Nigeria", "Britain", "Peru",
             "USA", NA)

participants <- data.frame (age, gender,
                            education, country)

head (participants, 3)
```

```
  age gender education country
1  34      M  Tertiary Nigeria
2  23      M   Primary Britain
3  56      F   Primary    Peru
```

Tidy data management: The How

```r
age <- c(34, 23, 56, 34, 46)
gender <- c("M", "M", "F", "F", "M")
education <- c("Tertiary", "Primary",
               "Primary", "Secondary", "Tertiary")
country <- c("Nigeria", "Britain", "Peru",
             "USA", NA)

participants <- data.frame (age, gender,
                            education, country)

head (participants, 3)
```

```
  age gender education country
1  34      M  Tertiary Nigeria
2  23      M   Primary Britain
3  56      F   Primary    Peru
```

```r
tail (participants, 3)
```

```
  age gender education country
3  56      F   Primary    Peru
4  34      F Secondary     USA
5  46      M  Tertiary    <NA>
```

Tidy data management: The How

```r
age <- c(34, 23, 56, 34, 46)
gender <- c("M", "M", "F", "F", "M")
education <- c("Tertiary", "Primary",
               "Primary", "Secondary", "Tertiary")
country <- c("Nigeria", "Britain", "Peru",
             "USA", NA)

participants <- data.frame (age, gender,
                            education, country)

head (participants, 3)
```

```
  age gender education country
1  34      M  Tertiary Nigeria
2  23      M   Primary Britain
3  56      F   Primary    Peru
```

```r
tail (participants, 3)
```

```
  age gender education country
3  56      F   Primary    Peru
4  34      F Secondary     USA
5  46      M  Tertiary    <NA>
```

```r
str (participants)
```

```
'data.frame':   5 obs. of  4 variables:
 $ age      : num  34 23 56 34 46
 $ gender   : chr  "M" "M" "F" "F" ...
 $ education: chr  "Tertiary" "Primary" "Primary" "Secondary"
...
 $ country  : chr  "Nigeria" "Britain" "Peru" "USA" ...
```

Tidy data management: The How

```r
age <- c(34, 23, 56, 34, 46)
gender <- c("M", "M", "F", "F", "M")
education <- c("Tertiary", "Primary",
               "Primary", "Secondary", "Tertiary")
country <- c("Nigeria", "Britain", "Peru",
             "USA", NA)

participants <- data.frame (age, gender,
                            education, country)

head (participants, 3)
```

```
  age gender education country
1  34      M  Tertiary Nigeria
2  23      M   Primary Britain
3  56      F   Primary    Peru
```

```r
tail (participants, 3)
```

```
  age gender education country
3  56      F   Primary    Peru
4  34      F Secondary     USA
5  46      M  Tertiary    <NA>
```

```r
str (participants)
```

```
'data.frame':   5 obs. of  4 variables:
 $ age      : num  34 23 56 34 46
 $ gender   : chr  "M" "M" "F" "F" ...
 $ education: chr  "Tertiary" "Primary" "Primary" "Secondary"
...
 $ country  : chr  "Nigeria" "Britain" "Peru" "USA" ...
```

```r
str (participants$age)
```

```
 num [1:5] 34 23 56 34 46
```

{Tidy} data management

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

Package	Functions
readxl	read_excel('my-spreadsheet.xls', sheet = 1), read_xls('my-spreadsheet.xls'), read_xlsx('my-spreadsheet.xlsx')
readstata13	read.dta13('my-stata-data.dta')
readr	read_csv('my-csv-file.csv'), read_csv2('my-csv-file.csv'), read_delim(), read_rds()
vroom	vroom('my-csv-file.csv')
tidyxl	xlsx_cells('my_nightmare_file.xlsx')
haven	read_dta(), read_sas(), read_sav(), read_spss(), read_stata()
utils	read.csv, read.delim, read.table

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
install.packages("haven")
install.packages("readstata13")
install.packages("tidyxl")
install.packages("readxl")
```

- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
install.packages("haven")
install.packages("readstata13")
install.packages("tidyxl")
install.packages("readxl")

library (haven)
library (readstata13)
library (tidyxl)
library (readxl)
```

- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```r
install.packages("haven")
install.packages("readstata13")
install.packages("tidyxl")
install.packages("readxl")

library (haven)
library (readstata13)
library (tidyxl)
library (readxl)

?read_dta
?read_xls
?read.dta13
```

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
1  library (readxl)
2  library(dplyr)
3
4  wvs_data <- read_xlsx("wvs_greatBritain.xlsx")
5  glimpse(wvs_data)
```

Assess the structure of the data with `glimpse()` from `{dplyr}`

```
Rows: 2,609
Columns: 368
$ `version: Version of Data File`
<chr> …
$ `doi: Digital Object Identifier`
<chr> …
$ `A_YEAR: Year of survey`
<chr> …
$ `B_COUNTRY: ISO 3166-1 numeric country code`
<chr> …
$ `B_COUNTRY_ALPHA: ISO 3166-1 alpha-3 country code`
<chr> …
$ `C_COW_NUM: CoW country code numeric`
<chr> …
$ `C_COW_ALPHA: CoW country code alpha`
<chr> …
$ `D_INTERVIEW: Interview ID`
<chr> …
$ `J_INTDATE: Date of interview`
<chr> …
$ `FW_START: Year/month of start-fieldwork`
<chr> …
$ `FW_END: Year/month of end-fieldwork`
<chr> …
```

Numeric

Integer

Character

Factor

Logical

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
1 library (readxl)
2 library(dplyr)
3
4 wvs_data <- read_xlsx("wvs_greatBritain.xlsx")
5 head(wvs_data, 10)
```

Assess the structure of the data with `head()` from `{utils}`

A tibble: 10 × 368

Q38: It is children duty to take care of ill parent <chr>	Q39: People who don't work turn lazy <chr>	Q40: Work is a duty towards society <chr>
Disagree	Strongly disagree	Strongly disagree
Disagree	Neither agree or disagree	Agree
Agree	Disagree	Agree
Disagree	Disagree	Disagree
Neither agree nor disagree	Neither agree or disagree	Agree
Disagree	Disagree	Agree
Disagree	Neither agree or disagree	Don't know
Disagree	Disagree	Agree
Disagree	Disagree	Disagree
Disagree strongly	Disagree	Neither agree or disagree

1-10 of 10 rows | 73-75 of 368 columns

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Inspect the first 10 rows in the data with `head()` from `{utils}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
1 library (readxl)
2 library(dplyr)
3 library (janitor)
4
5 wvs_data <- read_xlsx("wvs_greatBritain.xlsx")
6 head(wvs_data, 10)
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Fix column names with `clean_names()` from `{janitor}` 📦

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
library (readxl)
library(dplyr)
library (janitor)

wvs_data <- read_xlsx("wvs_greatBritain.xlsx")
head(wvs_data, 10)

?clean_names

wvs_clean_data <- clean_names(wvs_data)
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Assess the structure of the data with `glimpse()` *from* `{dplyr}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
1  library (readxl)
2  library(dplyr)
3  library (janitor)
4
5  wvs_data <- read_xlsx("wvs_greatBritain.xlsx")
6  head(wvs_data, 10)
7
8  ?clean_names
9
10 wvs_clean_data <- clean_names(wvs_data)
11 glimpse(wvs_clean_data)
```

Cleaned column names

```
Rows: 2,609
Columns: 368
$ version_version_of_data_file
<chr> …
$ doi_digital_object_identifier
<chr> …
$ a_year_year_of_survey
<chr> …
$ b_country_iso_3166_1_numeric_country_code
<chr> …
$ b_country_alpha_iso_3166_1_alpha_3_country_code
<chr> …
$ c_cow_num_co_w_country_code_numeric
<chr> …
$ c_cow_alpha_co_w_country_code_alpha
<chr> …
$ d_interview_interview_id
<chr> …
$ j_intdate_date_of_interview
<chr> …
$ fw_start_year_month_of_start_fieldwork
<chr> …
```

Uncleaned column names

```
Rows: 2,609
Columns: 368
$ `version: Version of Data File`
<chr> …
$ `doi: Digital Object Identifier`
<chr> …
$ `A_YEAR: Year of survey`
<chr> …
$ `B_COUNTRY: ISO 3166-1 numeric country code`
<chr> …
$ `B_COUNTRY_ALPHA: ISO 3166-1 alpha-3 country code`
<chr> …
$ `C_COW_NUM: CoW country code numeric`
<chr> …
$ `C_COW_ALPHA: CoW country code alpha`
<chr> …
$ `D_INTERVIEW: Interview ID`
<chr> …
$ `J_INTDATE: Date of interview`
<chr> …
$ `FW_START: Year/month of start-fieldwork`
<chr> …
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Keep only the relevant columns with `select()` *from* `{dplyr}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
library (readxl)
library(dplyr)
library (janitor)

wvs_data <- read_xlsx("wvs_greatBritain.xlsx")
head(wvs_data, 10)

?clean_names

wvs_clean_data <- clean_names(wvs_data)

sub_wvs_data <- wvs_clean_data %>%
                ## Select a few columns
                select(a_year_year_of_survey,
                       q261_year_of_birth,
                       q260_sex,
                       h_urbrural_urban_rural,
                       q269_respondent_citizen,
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Rename columns with long or complicated names with `rename()` from `{dplyr}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```r
sub_wvs_data <- wvs_clean_data %>%
                ## Select a few columns
                select(a_year_year_of_survey,
                       q261_year_of_birth,
                       q260_sex,
                       h_urbrural_urban_rural,
                       q269_respondent_citizen,
                       q223_local_party_preference_local_name,
                       q165_believe_in_god,
                       q191_justifiable_violence_against_other_people,
                       q275_highest_educational_level_respondent_isced_2011) %>%
                ## Rename columns
                rename (survey_yr =  a_year_year_of_survey,
                        party_pref = q223_local_party_preference_local_name,
                        violence_just = q191_justifiable_violence_against_other_peopl
                        education = q275_highest_educational_level_respondent_isced_2
                        residence = h_urbrural_urban_rural)
```

```
Rows: 2,609
Columns: 9
$ survey_yr              <chr> "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", "2022", …
$ q261_year_of_birth     <chr> "1967", "1980", "2000", "1950", "1952", "1971", "1988", "1966", "1947", "1944", "1954", "1949", "1993", "1937", …
$ q260_sex               <chr> "Female", "Female", "Female", "Female", "Male", "Female", "Female", "Female", "Female", "Female", "Male", "Male"…
$ residence              <chr> "Rural", "Rural", "Rural", "Rural", "Rural", "Urban", "Urban", "Urban", "Urban", "Urban", "Urban", "Urban", "Urb…
$ q269_respondent_citizen <chr> "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", "Yes", …
$ party_pref             <chr> "4", "GBR: Labour Party", "GBR: Labour Party", "GBR: Liberal Democrats", "GBR: Conservative and Unionist Party",…
$ q165_believe_in_god    <chr> "Yes", "Don't know", "No", "No", "Yes", "No", "Yes", "Yes", "Yes", "Yes", "No", "Yes", "No", "Yes", "Yes", "Yes"…
$ violence_just          <chr> "Never justifiable", "Never justifiable", "Never justifiable", "Never justifiable", "Never justifiable", "Never …
$ education              <chr> "Upper secondary education (ISCED 3)", "Master or equivalent (ISCED 7)", "Post-secondary non-tertiary education …
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Tabulate a few columns to understand the structure and identify potential structural errors with `table()` from `{base}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
1 table (sub_wvs_data$q260_sex)
```

```
                                  Female
                                    1471
                                    Male
                                    1105
                                No answer
                                      17
Other missing; Multiple answers Mail (EVS)
                                      16
```

```
1 table (sub_wvs_data$party_pref)
```

```
                                    -1                                    -2
                                   261                                   111
                                    -5                                     4
                                    19                                   271
                                     5          GBR: British National Party
                                    19                                     3
GBR: Conservative and Unionist Party       GBR: Democratic Unionist Party
                                   552                                     2
                   GBR: Green Party            GBR: Independence Party
                                   178                                    20
                  GBR: Labour Party             GBR: Liberal Democrats
                                   702                                   229
                   GBR: Plaid Cymru                      GBR: Reform UK
                                    40                                    24
         GBR: Scottish National Party                      GBR: Sinn Féin
                                   175                                     1
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Create a new object `wvs_clean_1` from `wvs_clean_1`

```
wvs_clean_1 <- sub_wvs_data
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Create a variable (replace if existing) with `mutate()` *from* `{dplyr}`

```
wvs_clean_1 <- sub_wvs_data %>%
               mutate (q260_sex = if_else((q260_sex != "Female" & q260_sex != "Male"),
                                          true = NA,
                                          false = q260_sex,
                                          missing = NA))
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Replace values that don't meet a condition to NA with `ifelse()` from `{dplyr}` 🧰

```
wvs_clean_1 <- sub_wvs_data %>%
               mutate (q260_sex = if_else((q260_sex != "Female" & q260_sex != "Male"),
                                          true = NA,
                                          false = q260_sex,
                                          missing = NA))
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Replace all less meaningful values to missing (NA) with `ifelse()` from `{dplyr}` 📦

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
wvs_clean_1 <- sub_wvs_data %>%
               mutate (q260_sex = if_else((q260_sex != "Female" & q260_sex != "Male"),
                                          true = NA,
                                          false = q260_sex,
                                          missing = NA),

                       residence = if_else((residence == "No answer; Missing"),
                                          true = NA,
                                          false = residence,
                                          missing = NA),

                       q269_respondent_citizen = if_else((q269_respondent_citizen != "No" &
                                                            q269_respondent_citizen != "Yes"),
                                                         true = NA,
                                                         false = q269_respondent_citizen,
                                                         missing = NA),

                       q261_year_of_birth = if_else((q261_year_of_birth == "No answer" |
                                                       q261_year_of_birth == "Other missing;
                                                    true = NA,
                                                    false = q261_year_of_birth,
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Assess the structure of the data, again with `glimpse()` *and tabulate with* `table()`

```
glimpse(wvs_clean_1$q261_year_of_birth)
```

```
chr [1:2609] "1967" "1980" "2000" "1950" "1952" "1971" "1988" "1966" ...
```

```
table(wvs_clean_1$violence_just)
```

```
                   2                    3                    4                    5
                 277                  168                   63                  107
                   6                    7                    8 Always justifiable
                  22                   21                    5                   12
Never justifiable
                1913
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Create a variable (replace if existing) with `mutate()` *from* `{dplyr}`

```
wvs_clean_2 <- wvs_clean_1 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence))
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Convert values in a variable to numeric with `as.numeric()` *and to categories with* `as.factor()` *from* `{base}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
wvs_clean_2 <- wvs_clean_1 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence))
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Recode values in a variable with `case_when()` *from* `{dplyr}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
wvs_clean_3 <- wvs_clean_2 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence)) %>%
               mutate (violence_just = case_when(violence_just == "Always justifiable" ~ 9,
                                                 violence_just == "Never justifiable" ~ 1,
                                                 .default = as.numeric(violence_just)))
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Assess the structure of the data with `glimpse()` *from* `{dplyr}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
wvs_clean_3 <- wvs_clean_2 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence)) %>%
               mutate (violence_just = case_when(violence_just == "Always justifiable" ~ 9,
                                                 violence_just == "Never justifiable" ~ 1,
                                                 .default = as.numeric(violence_just)))

str (wvs_clean_3$violence_just)
```

```
num [1:2609] 1 1 1 1 1 1 1 1 1 ...
```

```
table (wvs_clean_3$violence_just)
```

```
   1    2    3    4    5    6    7    8    9
1913  277  168   63  107   22   21    5   12
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

```r
wvs_clean_3 <- wvs_clean_2 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence)) %>%
               mutate (violence_just = case_when(violence_just == "Always justifiable" ~ 9,
                                                 violence_just == "Never justifiable" ~ 1,
                                                 .default = as.numeric(violence_just)))
```

- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Create a variable-age (replace if existing) with `mutate()` from `{dplyr}` 📦

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```r
wvs_clean_3 <- wvs_clean_2 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence)) %>%
               mutate (violence_just = case_when(violence_just == "Always justifiable" ~ 9,
                                                 violence_just == "Never justifiable" ~ 1,
                                                 .default = as.numeric(violence_just))) %>%

               mutate (age = survey_yr - q261_year_of_birth)
```

{Tidy} data management:

A case study using the `World Value Survey`

- **Fix structural errors**

  Assess the distribution of age with `table()` *from* `{base}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

```
wvs_clean_3 <- wvs_clean_2 %>%
               mutate (q261_year_of_birth = as.numeric(q261_year_of_birth),
                       survey_yr = as.numeric(survey_yr),
                       q260_sex = as.factor(q260_sex),
                       q269_respondent_citizen = as.factor(q269_respondent_citizen),
                       residence = as.factor(residence)) %>%
               mutate (violence_just = case_when(violence_just == "Always justifiable" ~ 9,
                                                 violence_just == "Never justifiable" ~ 1,
                                                 .default = as.numeric(violence_just))) %>%

               mutate (age = survey_yr - q261_year_of_birth)

table (wvs_clean_3$age)
```

```
19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44
25 15 16 30 33 33 22 33 36 27 29 40 48 58 41 45 36 32 45 37 43 35 46 33 45 42
45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70
39 43 29 40 30 34 35 43 38 41 47 40 46 39 43 52 37 43 22 45 47 49 34 42 37 47
71 72 73 74 75 76 77 78 79 80 81 82 83 84 85 86 87 88 89 90 91 92 94 95 96 99
42 31 42 39 39 38 22 40 18 27 19 13 17 19 18 15  5 10  7  6  7  3  1  3  3  1
```

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- *Remove duplicate or irrelevant observations*
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
wvs_clean_4 <- wvs_clean_3 %>%
               filter (age >= 18 & age <= 34)

str (wvs_clean_4)
```

```
tibble [531 × 10] (S3: tbl_df/tbl/data.frame)
 $ survey_yr              : num [1:531] 2022 2022 2022 2022 2022 ...
 $ q261_year_of_birth     : num [1:531] 2000 1988 1993 1996 1990 ...
 $ q260_sex               : Factor w/ 2 levels "Female","Male": 1 1 2 2 2 1 1 1 1 1 ...
 $ residence              : Factor w/ 2 levels "Rural","Urban": 1 2 2 2 2 1 1 1 1 2 ...
 $ q269_respondent_citizen: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ party_pref             : chr [1:531] "GBR: Labour Party" NA "GBR: Labour Party" "GBR: Conservative
and Unionist Party" ...
 $ q165_believe_in_god    : chr [1:531] "No" "Yes" "No" "No" ...
 $ violence_just          : num [1:531] 1 1 5 3 6 1 1 1 3 3 ...
 $ education              : chr [1:531] "Post-secondary non-tertiary education (ISCED 4)" "Upper
secondary education (ISCED 3)" "Bachelor or equivalent (ISCED 6)" "Bachelor or equivalent (ISCED 6)"
...
```

- ***Remove duplicate or irrelevant observations***

  Keep only data (or observations) from young adults aged 18-34 years with `filter()` from `{dplyr}`

- Handle (remove) unwanted outliers

- Handle (remove) missing data

```
tibble [2,609 × 10] (S3: tbl_df/tbl/data.frame)
 $ survey_yr              : num [1:2609] 2022 2022 2022 2022 2022 ...
 $ q261_year_of_birth     : num [1:2609] 1967 1980 2000 1950 1952 ...
 $ q260_sex               : Factor w/ 2 levels "Female","Male": 1 1 1 1 2 1 1 1 1 1 ...
 $ residence              : Factor w/ 3 levels "No answer; Missing",..: 2 2 2 2 2 3 3 3 3 3 ...
 $ q269_respondent_citizen: Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ party_pref             : chr [1:2609] NA "GBR: Labour Party" "GBR: Labour Party" "GBR: Liberal Democrats" ...
 $ q165_believe_in_god    : chr [1:2609] "Yes" NA "No" "No" ...
 $ violence_just          : num [1:2609] 1 1 1 1 1 1 1 1 1 1 ...
 $ education              : chr [1:2609] "Yes" "Yes" "Yes" "Yes" ...
 $ age                    : num [1:2609] 55 42 22 72 70 51 34 56 75 78 ...
```

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- ***Handle (remove) unwanted outliers***
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
1  table(wvs_clean_4$education)
```

```
                                  Bachelor or equivalent (ISCED 6)
                                                               150
                                  Doctoral or equivalent (ISCED 8)
                                                                 4
Early childhood education (ISCED 0) / no education
                                                                 4
                              Lower secondary education (ISCED 2)
                                                               105
                                   Master or equivalent (ISCED 7)
                                                                75
   Post-secondary non-tertiary education (ISCED 4)
                                                                11
                                     Primary education (ISCED 1)
                                                                 3
                          Short-cycle tertiary education (ISCED 5)
                                                                50
```

- Remove duplicate or irrelevant observations
- ***Handle (remove) unwanted outliers***

  mmmCheck if there are any missing values in the data with `anyNA()` from `{base}`

  
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

- Remove duplicate or irrelevant observations

- ***Handle (remove) unwanted outliers***

  mmmCheck if there are any missing values in the data with `anyNA()` from `{base}`

- Handle (remove) missing data

- Validate

```
wvs_clean_4 <- wvs_clean_4 %>%
               mutate (education = case_when((education == "Early childhood education (ISCED 0) /
                                                education == "Primary education (ISCED 1)" |
                                                 education == "Upper secondary education (ISCED 3)
                                                education == "Lower secondary education (ISCED 2)"


                                              (education == "Short-cycle tertiary education (ISCED
                                                 education == "Post-secondary non-tertiary educati
                                                 education == "Bachelor or equivalent (ISCED 6)" |
                                                 education == "Master or equivalent (ISCED 7)" |
                                                 education == "Doctoral or equivalent (ISCED 8)")

                                              .default = factor(education)))

table (wvs_clean_4$education)
```

```
Post-secondary Secondary or less
           290                221
```

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
1  table(wvs_clean_4$violence_just)
```

```
  1   2   3   4   5   6   7   8   9
330  69  51  25  32   5   9   2   2
```

- Remove duplicate or irrelevant observations

- ***Handle (remove) unwanted outliers***

  mmmCheck if there are any missing values in the data with `anyNA()` from `{base}`

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

- Remove duplicate or irrelevant observations

- ***Handle (remove) unwanted outliers***

  mmmCheck if there are any missing values in the data with `anyNA()` from `{base}`

- Handle (remove) missing data

- Validate

```
wvs_clean_4 <- wvs_clean_4 %>%
               mutate (violence_just = if_else((violence_just >= 2),
                                               true = "Justified",
                                               false = "Never justified",
                                               missing = NA)) %>%
               mutate (violence_just = as.factor(violence_just))

table (wvs_clean_4$violence_just)
```

```
Justified Never justified
      195             330
```

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- *Handle (remove) missing data*
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
1 anyNA(wvs_clean_4)
```

```
[1] TRUE
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- ***Handle (remove) missing data***

  Check if there are any missing values in the data with `anyNA()` from `{base}`

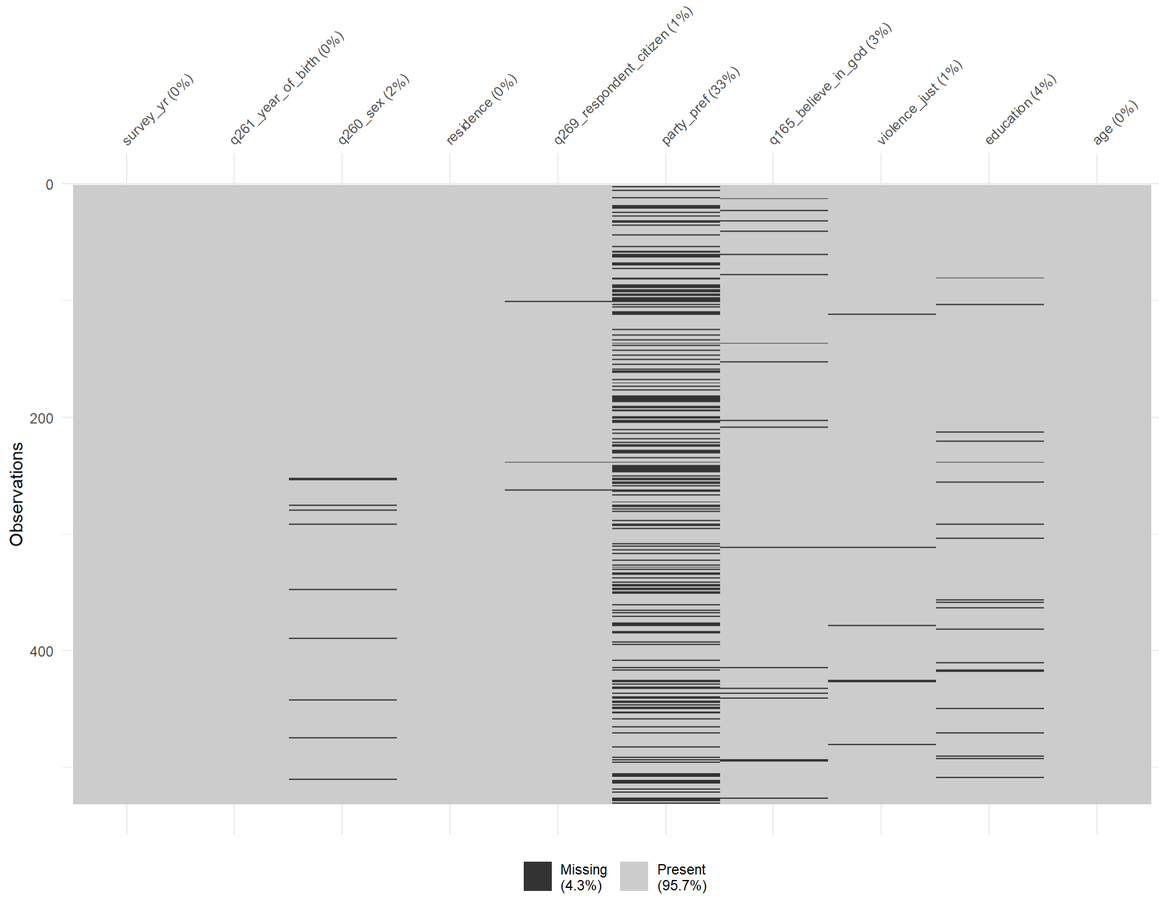
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
library (visdat)
vis_miss(wvs_clean_4)
```



- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- ***Handle (remove) missing data***

  Visualise missing values across the dataset with `vis_miss()` from `{visdat}`

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

- Remove duplicate or irrelevant observations

```
library (visdat)
vis_miss(wvs_clean_4)

wvs_clean_5 <-  wvs_clean_4 %>%
                filter (!is.na (q260_sex) &
                          !is.na(residence) &
                          !is.na(q269_respondent_citizen))
```

- Handle (remove) unwanted outliers

- ***Handle (remove) missing data***

  Remove missing values in each column with `filter()` *from* `{dplyr}` 📦

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
library (visdat)
vis_miss(wvs_clean_4)

wvs_clean_5 <-  wvs_clean_4 %>%
                filter (!is.na (q260_sex) &
                          !is.na(residence) &
                          !is.na(q269_respondent_citizen))
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- ***Handle (remove) missing data***

  Check whether a column has missing values with `is.na()` from `{base}`

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
library (visdat)
vis_miss(wvs_clean_5)
```

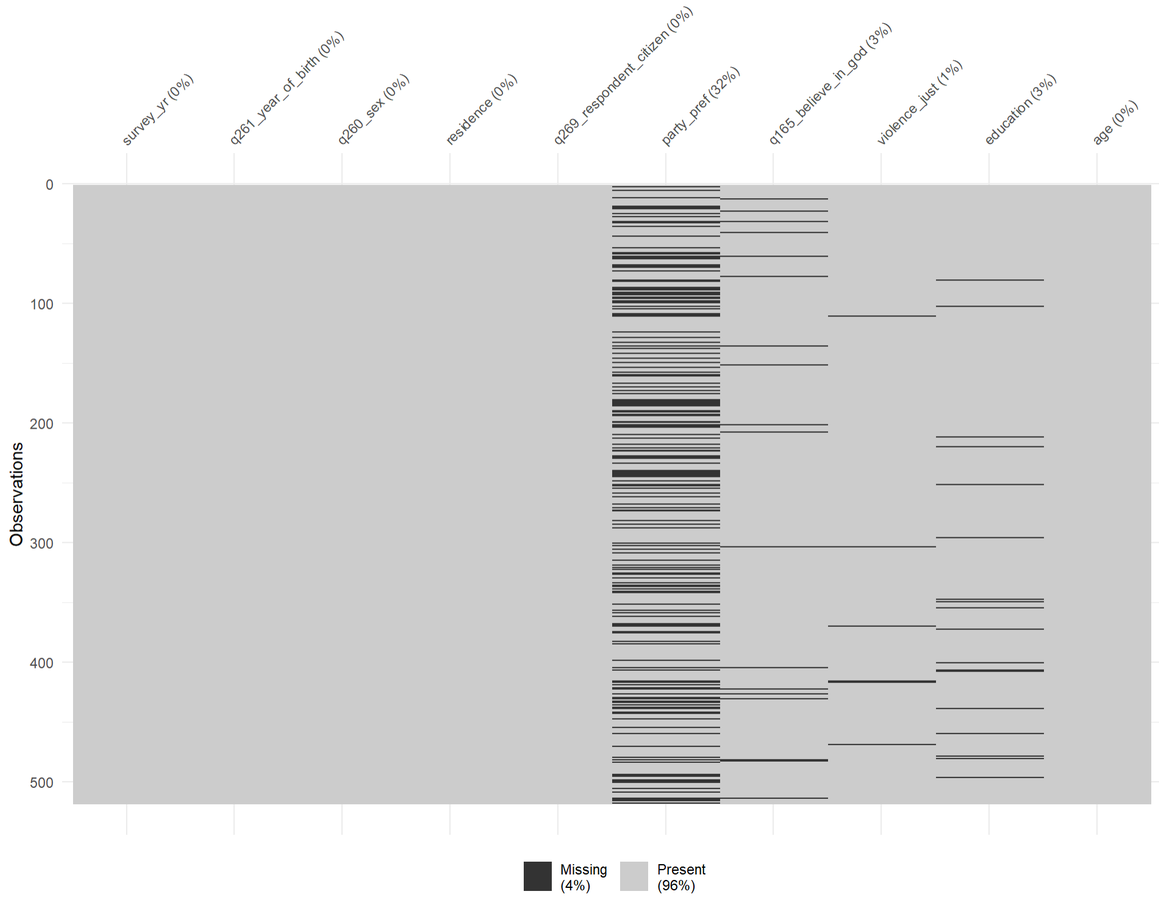
- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- ***Handle (remove) missing data***

  Visualise missing values across the dataset with `vis_miss()` from `{visdat}`

- Validate



{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

- Remove duplicate or irrelevant observations

```
library (visdat)
vis_miss(wvs_clean_4)

wvs_clean_5 <-  wvs_clean_4 %>%
                filter (!is.na (q260_sex) &
                          !is.na(residence) &
                          !is.na(q269_respondent_citizen))
```

- Handle (remove) unwanted outliers

- ***Handle (remove) missing data***

  Remove missing values in each column with `filter()` *from* `{dplyr}` 📦

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

- Remove duplicate or irrelevant observations

```
library (visdat)
vis_miss(wvs_clean_4)

wvs_clean_5 <-  wvs_clean_4 %>%
                filter (!is.na (q260_sex) &
                          !is.na(residence) &
                          !is.na(q269_respondent_citizen)) %>%
                filter (!is.na (party_pref) &
                          !is.na(q165_believe_in_god) &
                          !is.na(violence_just) &
                          !is.na(education))
```

- Handle (remove) unwanted outliers

- ***Handle (remove) missing data***

  Remove missing values in each column with `filter()` *from* `{dplyr}` 📦

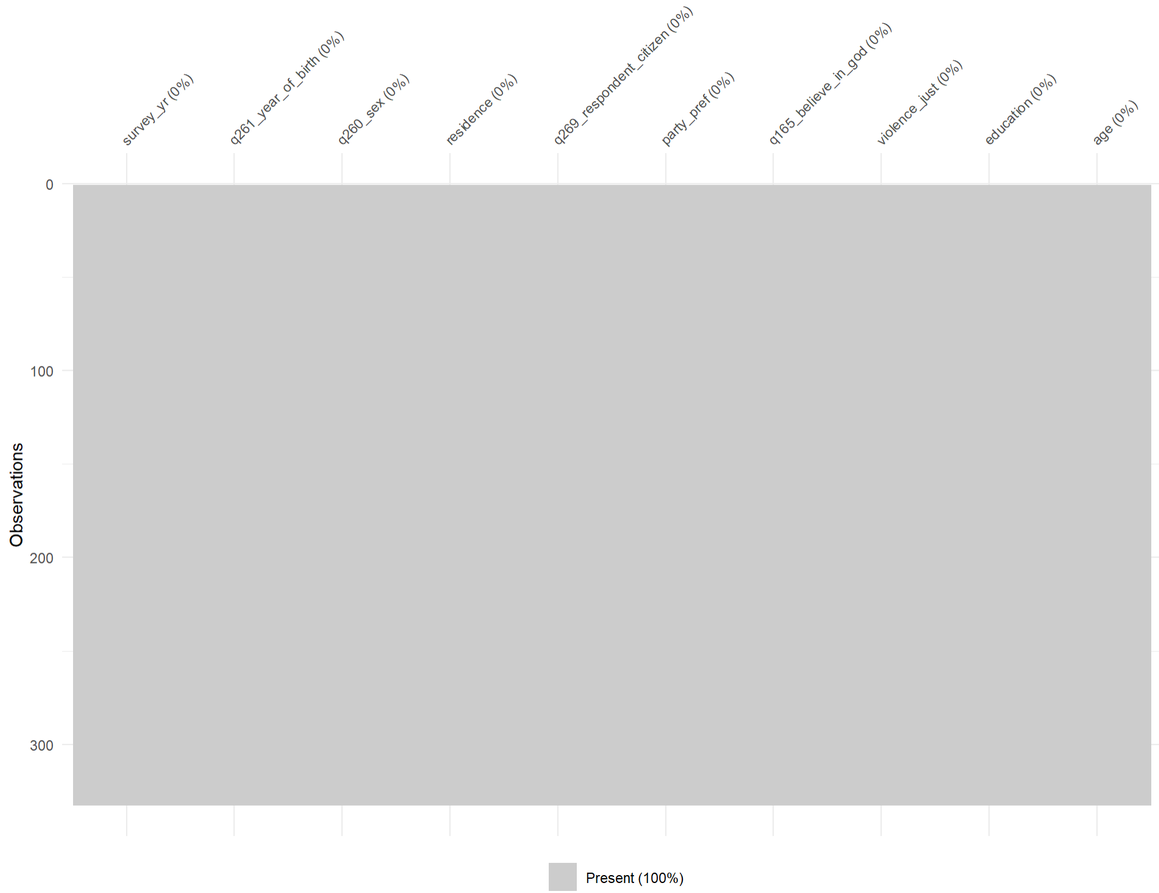
- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
1 library (visdat)
2 vis_miss(wvs_clean_5)
```



- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- ***Handle (remove) missing data***

  Visualise missing values across the dataset with `vis_miss()` from `{visdat}`

- Validate

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- *Validate*

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
## Sample of young people with complete cases
dim(wvs_clean_5)
```

```
[1] 332  10
```

- Remove duplicate or irrelevant observations

```
## Full sample of young people with missing cases
dim(wvs_clean_4)
```

```
[1] 531  10
```

- Handle (remove) unwanted outliers

```
## Full adult sample in the dataset with missing cases
dim(wvs_clean_3)
```

```
[1] 2609   10
```

- Handle (remove) missing data

- ***Validate***
  Assess the data dimensions with `dim()` from `{base}`

{Tidy} data management:

A case study using the `World Value Survey`

- Fix structural errors

```
## Sample of young people with complete cases
dim(wvs_clean_5)
```

```
[1] 332  10
```

- Remove duplicate or irrelevant observations

```
## Full sample of young people with missing cases
dim(wvs_clean_4)
```

```
[1] 531  10
```

```
## Full adult sample in the dataset with missing cases
dim(wvs_clean_3)
```

```
[1] 2609   10
```

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- ***Validate***
  Assess the data dimensions with `dim()` from `{base}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
## Sample of young people with complete cases
dim(wvs_clean_5)
```

```
[1] 332  10
```

- Remove duplicate or irrelevant observations

```
## Full sample of young people with missing cases
dim(wvs_clean_4)
```

```
[1] 531  10
```

```
## Full adult sample in the dataset with missing cases
dim(wvs_clean_3)
```

```
[1] 2609   10
```

- Handle (remove) unwanted outliers

- Handle (remove) missing data

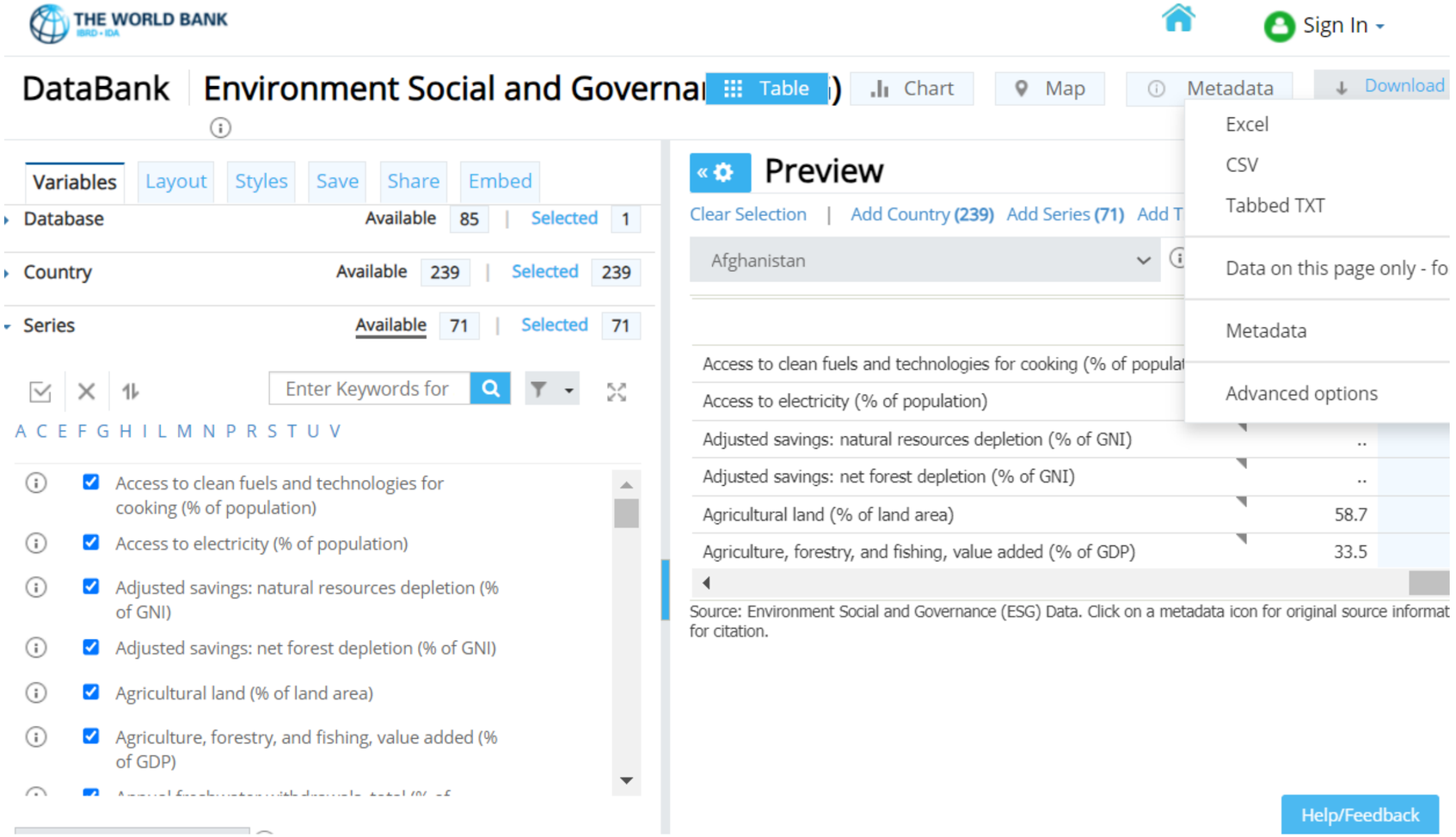
- ***Validate***
  Assess the data dimensions with `dim()` from `{base}`

{Tidy} data management

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate



{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
1  library(dplyr)
2  wb_dt <- read.csv("wb_databank.csv")
3
4  head(wb_dt, 10)
5  tail(wb_dt, 10)
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

	Country.Name	Country.Code	Series.Name	Series
16964	World	WLD	Strength of legal rights index (0=weak to 12=strong)	IC.LGL
16965	World	WLD	Terrestrial and marine protected areas (% of total territorial ...	ER.PTI
16966	World	WLD	Tree Cover Loss (hectares)	AG.LN
16967	World	WLD	Unemployment, total (% of total labor force) (modeled ILO ...	SL.UEI
16968	World	WLD	Unmet need for contraception (% of married women ages 1...	SP.UW
16969	World	WLD	Voice and Accountability: Estimate	VA.ES
16970				
16971				
16972				
16973	Data from database: Environment Social and Governance (E...			
16974	Last Updated: 10/02/2023			
16975	Code	License Type	Indicator Name	Short
16976	EG.CFT.ACCS.ZS	CC BY-4.0	Access to clean fuels and technologies for cooking (% of po...	
16977	EG.ELC.ACCS.ZS	CC BY-4.0	Access to electricity (% of population)	
16978	NY.ADJ.DRES.GN.ZS	CC BY-4.0	Adjusted savings: natural resources depletion (% of GNI)	
16979	NY.ADJ.DFOR.GN.ZS	CC BY-4.0	Adjusted savings: net forest depletion (% of GNI)	
16980	AG.LND.AGRI.ZS	CC BY-4.0	Agricultural land (% of land area)	
16981	NV.AGR.TOTL.ZS	CC BY-4.0	Agriculture, forestry, and fishing, value added (% of GDP)	
16982	ER.H2O.FWTL.ZS	CC BY-4.0	Annual freshwater withdrawals, total (% of internal resources)	
16983	SI.SPR.PCAP.ZG	CC BY-4.0	Annualized average growth rate in per capita real survey me...	The g
16984	SH.DTH.COMM.ZS	CC BY-4.0	Cause of death, by communicable diseases and maternal, pr...	
16985	SL.TLF.0714.ZS	CC BY-4.0	Children in employment, total (% of children ages 7-14)	
16986				

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
1  library(dplyr)
2
3  wb_dt <- read.csv("wb_databank.csv")
4
5  head(wb_dt, 10)
6
7  tail(wb_dt, 10)
8
9  wb_data <- wb_dt[1:16969,]
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
wb_data <- wb_dt[1:16969,]

wb_data_2 <-  wb_data %>%
              janitor::clean_names()
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wb_data <- wb_dt[1:16969,]

wb_data_2 <-  wb_data %>%
              janitor::clean_names() %>%
              pivot_longer(cols = contains("_yr"),
                           values_to = "values",
                           names_to = "year")
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wb_data <- wb_dt[1:16969,]

wb_data_2 <-  wb_data %>%
              janitor::clean_names() %>%
              pivot_longer(cols = contains("_yr"),
                           values_to = "values",
                           names_to = "year") %>%
              mutate (values = ifelse(values == "..",
                                      NA, values)) %>%
              mutate (values = as.numeric(values))
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wb_data <- wb_dt[1:16969,]

wb_data_2 <-  wb_data %>%
              janitor::clean_names() %>%
              pivot_longer(cols = contains("_yr"),
                           values_to = "values",
                           names_to = "year") %>%
              mutate (values = ifelse(values == "..",
                                      NA, values)) %>%
              mutate (values = as.numeric(values)) %>%
              mutate (period = str_extract(year, "[0-9]+")) %>%
              mutate (year = as.numeric (period))
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
new_dta <-  wb_data_2
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
new_dta <-  wb_data_2 %>%
            filter (year >= 2000 & year <= 2020) %>%
            filter (!is.na (values))
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
data_key <- new_dta %>%
            select (series_name, series_code)
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
data_key <- new_dta %>%
            select (series_name, series_code) %>%
            reframe(series_code = unique(series_code),
                    series_name = unique(series_name))
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
data_key <- new_dta %>%
            select (series_name, series_code) %>%
            reframe(series_code = unique(series_code),
                    series_name = unique(series_name)) %>%
            mutate (series_code = str_to_lower(series_code)) %>%
            mutate (series_code = str_replace_all(series_code, "\\.",
```

Inspect the first 10 rows in the data with `head()` from `{utils}` 📦

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors

```
wide_dta <- new_dta %>%
            select (country_name, year,
                    series_code, values)
```

- Remove duplicate or irrelevant observations

- Handle (remove) unwanted outliers

- Handle (remove) missing data

- Validate

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wide_dta <- new_dta %>%
            select (country_name, year,
                    series_code, values) %>%
            pivot_wider(names_from = series_code,
                        values_from = values)
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wide_dta <- new_dta %>%
            select (country_name, year,
                    series_code, values) %>%
            pivot_wider(names_from = series_code,
                        values_from = values) %>%
            janitor::clean_names()
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wide_dta <- new_dta %>%
            select (country_name, year,
                    series_code, values) %>%
            pivot_wider(names_from = series_code,
                        values_from = values) %>%
            janitor::clean_names() %>%
            select (country_name, year, eg_cft_accs_zs,
                    eg_elc_accs_zs, en_atm_co2e_pc,
                    en_clc_heat_xd, sp_dyn_tfrt_in )
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wide_dta <- new_dta %>%
            select (country_name, year,
                    series_code, values) %>%
            pivot_wider(names_from = series_code,
                        values_from = values) %>%
            janitor::clean_names() %>%
            select (country_name, year, eg_cft_accs_zs,
                    eg_elc_accs_zs, en_atm_co2e_pc,
                    en_clc_heat_xd, sp_dyn_tfrt_in ) %>%
            mutate (eg_cft_accs_zs = round (eg_cft_accs_zs, 2),
                    eg_elc_accs_zs = round (eg_elc_accs_zs, 2),
                    en_atm_co2e_pc = round (en_atm_co2e_pc, 2),
                    en_clc_heat_xd = round (en_clc_heat_xd, 2),
                    sp_dyn_tfrt_in = round (sp_dyn_tfrt_in, 2))
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
wide_dta <- new_dta %>%
            select (country_name, year,
                    series_code, values) %>%
            pivot_wider(names_from = series_code,
                        values_from = values) %>%
            janitor::clean_names() %>%
            select (country_name, year, eg_cft_accs_zs,
                    eg_elc_accs_zs, en_atm_co2e_pc,
                    en_clc_heat_xd, sp_dyn_tfrt_in ) %>%
            mutate (eg_cft_accs_zs = round (eg_cft_accs_zs, 2),
                    eg_elc_accs_zs = round (eg_elc_accs_zs, 2),
                    en_atm_co2e_pc = round (en_atm_co2e_pc, 2),
                    en_clc_heat_xd = round (en_clc_heat_xd, 2),
                    sp_dyn_tfrt_in = round (sp_dyn_tfrt_in, 2))
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

{Tidy} data management:

A case study using the `World Bank Data`

- Fix structural errors
- Remove duplicate or irrelevant observations
- Handle (remove) unwanted outliers
- Handle (remove) missing data
- Validate

```
1 final <-  w
```

```
1 Tips for findind data
2 ├── What type of data is needed?
3 ├── Who could have collected the data?
4 │
5 Where to Find Data by:
6 ├── Academics
7 ├── Not-for-profit Organizations
8 ├── Government Departments
9 ├── Private/Comercial Companies
```

Inspect the first 10 rows in the data with `head()` from `{utils}`

Statistical Foundations