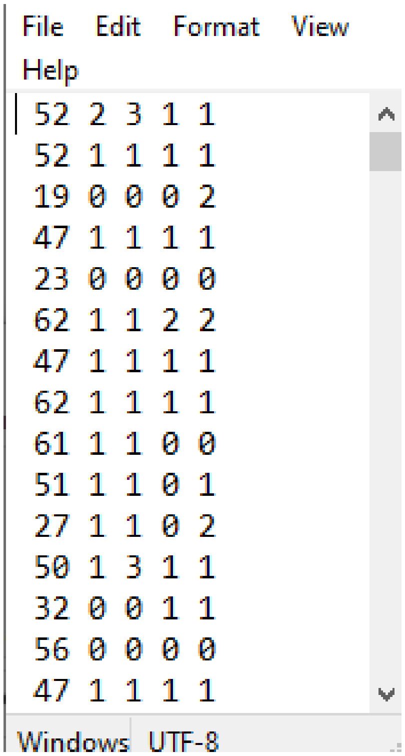
Twin data: descriptive statistics

In this example, we show how to assign variable names and category labels. The data we are using is for twins. Each line of the data contains information on a pair of twins. For each pair, we have the following information:

- Their year of birth
- Whether the first twin had an asthma attack prior to the age of 14
- Whether the first twin had an asthma attack after the age of 14
- Whether the second twin had an asthma attack prior to the age of 14
- Whether the second twin had an asthma attack after the age of 14

The data are contained in a text file with the name **asthma.raw**. All files for this example can be found in the **Prelis examples\Descriptive stats and regression** folder. The first few lines of the data are shown below.



```
52 2 3 1 1
52 1 1 1 1
19 0 0 0 2
47 1 1 1 1
23 0 0 0 0
62 1 1 2 2
47 1 1 1 1
62 1 1 1 1
61 1 1 0 0
51 1 1 0 1
27 1 1 0 2
50 1 3 1 1
32 0 0 1 1
56 0 0 0 0
47 1 1 1 1
```

For the variables describing asthma attacks, three codes were assigned:

- 1 = Never
- 2 = Occasionally
- 3 = Frequently.

We will now show how to assign variable names and category labels to these data and run descriptive statistics on it.

The most basic form of syntax to run data screening on this data set is encapsulated in the PRELIS syntax

```
Data Ninputvariables = 5
Rawdata = ASTHMA.RAW
Output
```

It gives the name of the data, and the number of variables. However, the output obtained is hard to interpret as default variable names are assigned:

```
Univariate Distributions for Ordinal Variables

 VAR 2            Frequency Percentage Bar Chart
      0     462         30.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
      1     909         60.7    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
      2      83          5.5    ⍰⍰⍰⍰
      3      44          2.9    ⍰⍰
```

As a first step, let us provide names for the variables by using the Labels command (see **asthma1.prl**)

```
!Data Screening of ASTHMA.RAW
!Reading Data in Free Format
!
!Variables: birth_year = Year of Birth
!           before14_twin1 = Asthma prior age 14 in Twin 1
!           After14_twin1  = Asthma after age 14 in Twin 1
!           before14_twin2 = Asthma prior age 14 in Twin 2
!           After14_twin2  = Asthma after age 14 in Twin 2
! Codes: 0=missing 1=never 2=occasional 3=frequently
!
Data Ninputvariables = 5
Labels
Birth_year before14_twin1 After14_twin1 before14_twin1 After14_twin1 !These are
the names of my variables
Rawdata = ASTHMA.RAW
Output
```

The output for this run gives output that is a lot easier to interpret, but we still need to assign category labels for the 4 variables.

```
Univariate Distributions for Ordinal Variables

 before14_twin1   Frequency Percentage Bar Chart
             0     462        30.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             1     909       60.7    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             2      83         5.5    ⍰⍰⍰⍰
             3      44         2.9    ⍰⍰

 After14_twin1    Frequency Percentage Bar Chart
             0     459        30.6    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             1     859       57.3    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             2     129         8.6    ⍰⍰⍰⍰⍰⍰⍰
             3      51         3.4    ⍰⍰⍰

 before14_twin1   Frequency Percentage Bar Chart
             0     477        31.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             1     910       60.7    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             2      70         4.7    ⍰⍰⍰⍰
             3      41         2.7    ⍰⍰

 After14_twin1    Frequency Percentage Bar Chart
             0     471        31.4    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             1     851       56.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
             2     143         9.5    ⍰⍰⍰⍰⍰⍰⍰⍰
             3      33         2.2    ⍰⍰

 Univariate Summary Statistics for Continuous Variables

 Variable            Mean  St. Dev.  Skewness  Kurtosis  Minimum Freq.  Maximum Freq.
 ---------------     ----  --------   -------  --------  ------- -----  ------- -----
       Birth_year  45.884    14.351    -0.850     0.254    0.000     1   98.000     1
```

To do this, we use the Clabels command, as shown below in **Asthma2.prl**:

```
!Data Screening of ASTHMA.RAW
!Reading Data in Free Format
!Variables: Birth_year = Year of Birth
!           before14_twin1 = Asthma prior age 14 in Twin 1
!           After14_twin1  = Asthma after age 14 in Twin 1
!           before14_twin2 = Asthma prior age 14 in Twin 2
!           After14_twin2  = Asthma after age 14 in Twin 2
! Codes: 0=missing 1=never 2=occasional 3=frequently
Data Ninputvariables = 5
Labels
Birth_year before14_twin1 After14_twin1 before14_twin2 After14_twin2
RA = ASTHMA.RAW
CO Birth_year
Clabels:  before14_twin1 After14_twin1 before14_twin2 After14_twin2 0=MISS 1=NVER
2=OCCL 3=FREQ
Output
```

The results of the data screening is now a lot easier to interpret:

```
Total Sample Size(N) =   1498

 Univariate Distributions for Ordinal Variables

 before14_twin1   Frequency Percentage Bar Chart
          MISS     462        30.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          NVER     909        60.7    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          OCCL      83         5.5    ⍰⍰⍰⍰
          FREQ      44         2.9    ⍰⍰

 After14_twin1    Frequency Percentage Bar Chart
          MISS     459        30.6    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          NVER     859        57.3    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          OCCL     129         8.6    ⍰⍰⍰⍰⍰⍰⍰
          FREQ      51         3.4    ⍰⍰⍰

 before14_twin2   Frequency Percentage Bar Chart
          MISS     477        31.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          NVER     910        60.7    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          OCCL      70         4.7    ⍰⍰⍰⍰
          FREQ      41         2.7    ⍰⍰

 After14_twin2    Frequency Percentage Bar Chart
          MISS     471        31.4    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          NVER     851        56.8    ⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰⍰
          OCCL     143         9.5    ⍰⍰⍰⍰⍰⍰⍰⍰
          FREQ      33         2.2    ⍰⍰
```

However, we note that there is a lot of missing data, as indicated by the MISS category for each variable. We opt to apply listwise deletion to remove the missing cases. This is done by using the Missing keyword on the Data command, as shown below (**asthma3.prl**):

```
!Data Screening of ASTHMA.RAW
!Reading Data in Free Format
!Variables: Birth_year = Year of Birth
!           before14_twin1 = Asthma prior age 14 in Twin 1
!           After14_twin1  = Asthma after age 14 in Twin 1
!           before14_twin2 = Asthma prior age 14 in Twin 2
!           After14_twin2  = Asthma after age 14 in Twin 2
! Codes: 0=missing 1=never 2=occasional 3=frequently
Data Ninputvariables = 5 missing = 0
Labels
Birth_year before14_twin1 After14_twin1 before14_twin2 After14_twin2
RA = ASTHMA.RAW
CO Birth_year
Clabels:  before14_twin1 After14_twin1 before14_twin2 After14_twin2 0=MISS 1=NVER
2=OCCL 3=FREQ
Output
```

When we now inspect the data, we have successfully discarded missing data. The output also provides information on how many observations were deleted and which patterns it pertained to:

```
Total Sample Size(N) =   1498

 Number of Missing Values     0     1     2     3     4     5
          Number of Cases   815   125   217    54   286     1

 Effective Sample Sizes
 Univariate (in Diagonal) and Pairwise Bivariate (off Diagonal)

                        Birth_year        before14_twin1         After14_twin1        before14_twin2
After14_twin2
                  ----------------   ----------------   ----------------   ----------------   ----------
    Birth_year            1497
before14_twin1            1036              1036
After14_twin1             1039               986                1039
before14_twin2            1021               879                 885                1021
After14_twin2              1027               876                 883                 973                1027

 Percentage of Missing Values
 Univariate (in Diagonal) and Pairwise Bivariate (off Diagonal)

                     Birth_year    before14_twin1     After14_twin1    before14_twin2     After14_twin2
                  ----------------   ----------------   ----------------   ----------------   --------------
    Birth_year              0.07
before14_twin1             30.84             30.84
 After14_twin1             30.64             34.18             30.64
before14_twin2             31.84             41.32             40.92             31.84
 After14_twin2             31.44             41.52             41.05             35.05             31.44

 Missing Data Map

 Frequency Percent   Pattern
       815    54.4   0 0 0 0 0
        32     2.1   0 1 0 0 0
        27     1.8   0 0 1 0 0
        99     6.6   0 1 1 0 0
        30     2.0   0 0 0 1 0
         6     0.4   0 1 0 1 0
         4     0.3   0 0 1 1 0
        14     0.9   0 1 1 1 0
        36     2.4   0 0 0 0 1
         2     0.1   0 1 0 0 1
         1     0.1   0 0 1 0 1
         9     0.6   0 1 1 0 1
       105     7.0   0 0 0 1 1
        13     0.9   0 1 0 1 1
        18     1.2   0 0 1 1 1
       286    19.1   0 1 1 1 1
         1     0.1   1 1 1 1 1
```