

MAKING SENSE OF DATA

Traceis™ Data Exploration Studio Clean

1 Select the prepare step

This example uses the baseball hitters dataset. To clean any of the variables, first select the prepare step.

2 Select the clean tab

3 To clean the variable

To clean an individual variable, first click on the name of the variable.

The screenshot shows the Traceis Data Exploration Studio 2007 beta 1.1 interface. The 'Prepare' step is selected in the left sidebar (indicated by a red circle 1). The 'Clean' tab is active in the top menu bar (indicated by a red circle 2). A variable is selected in the 'Select variable' list (indicated by a red circle 3). The main window displays a table of data with columns: Hitter's name, Bat times (36), Hits (36), Home runs (36), Runs (36), Runs batted (36), Walks (36), Player team (36), Player's position (36), and Put outs (36). The table lists various players and their statistics. The bottom status bar shows 'Total observations: 263; Selected observations: 0'.

Hitter's name	Bat times (36)	Hits (36)	Home runs (36)	Runs (36)	Runs batted (36)	Walks (36)	Player team (36)	Player's position (36)	Put outs (36)
Al Newman	185	37	1	23	8	21	Mon.	2B	76
Hitter's name	315	81	7	24	38	39	Hou.	C	632
Bat times (36)	574	159	21	107	75	69	Det.	SS	338
Hits (36)	239	60	0	30	11	22	Bal.	2B	121
Home runs (36)	202	55	4	31	26	27	L.A.	C	304
Runs (36)	594	169	4	74	51	35	Oak.	SS	282
Runs batted (36)	479	130	18	66	72	76	Sea.	1B	880
Walks (36)	496	141	20	65	78	37	Mon.	RF	200
Player team (36)	401	92	17	49	66	65	Chi.	DH	0
Player's position (36)	401	87	10	39	42	30	Mon.	1B	805
Put outs (36)	323	81	6	26	32	8	Atl.	SS	143
Hitter's name	418	113	13	48	61	47	S.L.	RF	211
Bat times (36)	266	73	0	24	24	7	K.C.	SS	121
Hits (36)	413	92	16	72	48	65	Ph.	OF	280
Home runs (36)	196	43	7	29	27	30	Ph.	UT	80
Runs (36)	629	168	18	73	102	40	Bos.	1B	1067
Runs batted (36)	550	152	6	92	37	81	Hou.	2B	262
Walks (36)	379	106	10	36	69	30	L.A.	3B	72
Player team (36)	217	46	7	32	19	9	MI.	UT	307
Player's position (36)	419	106	6	55	36	22	Hou.	OF	226
Put outs (36)	181	41	1	15	21	33	MI.	1B	326
Hitter's name	474	129	10	50	56	40	Chi.	C	732
Bat times (36)	472	116	16	60	62	74	S.F.	C	518
Hits (36)	324	73	4	32	18	22	Chi.	OF	222
Home runs (36)	204	49	6	23	25	12	Sea.	C	419
Runs (36)	268	60	5	24	25	15	S.F.	C	442
Runs batted (36)	426	109	3	55	43	62	Chi.	OH	361
Walks (36)	587	163	4	92	51	70	Chi.	OF	434
Player team (36)	513	137	20	90	95	90	Cal.	LF	287
Player's position (36)	583	166	17	83	80	56	Chi.	3B	198
Put outs (36)	127	32	8	16	22	14	S.D.	C	202
Hitter's name	568	158	20	89	75	73	Chi.	3B	105

MAKING SENSE OF DATA

1

Count of missing and non-numeric observations

The dialog displays (1) the number of numeric observations in the variable, (2) the number of non-numeric (i.e., text values), and (3) the number of observations with no values for the observations.

2

Options for cleaning the data

Dependent on whether there are non-numeric or missing observations, a series of options is presented for cleaning the data set. If there are non-numeric values, these observations may be removed from the data set or replaced with a numeric value. If there are missing values, then options for either removing the observations or replacing them with a value are presented.

