



## A Capitalized Title: Something about a Package cleanR

Achim Zeileis  
Universität Innsbruck

Second Author  
Plus Affiliation

---

### Abstract

The abstract of the article.

*Keywords:* keywords, comma-separated, not capitalized, Java.

---

## 1. Introduction

Statisticians and data analysts spend a large portion of their time on data cleaning and on data wrangling. Packages such as `tidyverse`, `dplyr`, and `data.table` make data wrangling a lot easier in R, but there are only a few options available for data cleaning.

Data cleaning is a time consuming endeavour and it inherently requires human interaction since every dataset is different and the variables in the dataset can only be understood in the proper context of the experiment. While each dataset is different and require ... (“simple questions”

... there are a number of common occurrences that .... This is particularly Automated  
... partly be done by ... However ... enough ... similar tasks that pop  
However it is not uncommon for ... to run a set of ....

## 2. Checking a dataset for errors

```
> library(cleanR)
> data(testData)
> testData
```

```
charVar factorVar numVar intVar boolVar keyVar emptyVar _joeVar jack__var
```

1	a	a	1	1	TRUE	1	1	1	1
2	b	b	2	2	FALSE	2	1	2	2
3	c	c	3	3	TRUE	3	1	3	3
4	a	a	4	4	TRUE	4	1	4	4
5	b	b	5	5	TRUE	5	1	5	5
6	d	d	6	6	FALSE	6	1	6	6
7	a	a	7	7	TRUE	7	1	7	7
8	a	a	8	8	FALSE	8	1	8	8
9	b	b	9	9	TRUE	9	1	9	9
10	c	c	10	10	TRUE	10	1	10	10
11	a	a	1	1	TRUE	11	1	11	11
12	b	b	1	1	FALSE	12	1	12	12
13	d	d	1	1	NA	13	1	13	13
14	a	a	5	5	NA	14	1	14	14
15	<NA>	999	5	5	NA	15	1	15	15

	numOutlierVar	smartNumVar	cprVar	cprKeyVar	miscodedMissingVar
1	1	0	010101-1111	010101-1111	.
2	2	0	020102-2929	020102-2929	
3	3	0	121201-1902	121201-1902	nan
4	4	0	030729-2222	030729-2222	NaN
5	5	0	080909-1212	080909-1212	NAN
6	6	0	010101-1111	020202-0101	na
7	7	0	020102-2929	030303-0101	NA
8	8	1	121201-1902	040404-0101	Na
9	9	1	030729-2222	050505-0101	Inf
10	10	1	080909-1212	060606-0101	inf
11	11	1	010101-1111	070707-0101	-Inf
12	12	1	020102-2929	080808-0101	-inf
13	13	1	121201-1902	090909-0101	-
14	14	1	030729-2222	020202-0202	9
15	100	1	080909-1212	030303-0202	9

```
> clean(testData)
```

Arguments

### 3. Extending the error checks

### 4. About Java

**Affiliation:**

Achim Zeileis  
Department of Statistics and Mathematics  
Faculty of Economics and Statistics  
Universität Innsbruck  
6020 Innsbruck, Austria  
E-mail: [Achim.Zeileis@uibk.ac.at](mailto:Achim.Zeileis@uibk.ac.at)  
URL: <http://eeecon.uibk.ac.at/~zeileis/>