

# Package ‘RSADBE’

January 20, 2025

**Type** Package

**Title** Data related to the book ``R Statistical Application Development by Example''

**Version** 1.0

**Date** 2013-05-13

**Author** Prabhanjan Tattar

**Maintainer** Prabhanjan Tattar <prabhanjannt@gmail.com>

**Description** The package contains all the data sets related to the book written by the maintainer of the package.

**License** GPL-2

**NeedsCompilation** no

**Repository** CRAN

**Date/Publication** 2013-06-04 09:00:58

## Contents

|                                |    |
|--------------------------------|----|
| RSADBE-package . . . . .       | 2  |
| Bug_Metrics_Software . . . . . | 3  |
| CART_Dummy . . . . .           | 3  |
| CT . . . . .                   | 4  |
| DCD . . . . .                  | 5  |
| employ . . . . .               | 5  |
| galton . . . . .               | 6  |
| Gasoline . . . . .             | 7  |
| GC . . . . .                   | 8  |
| IO_Time . . . . .              | 9  |
| lowbwt . . . . .               | 10 |
| MDR . . . . .                  | 11 |
| octane . . . . .               | 12 |
| OF . . . . .                   | 12 |
| PW_Illus . . . . .             | 13 |
| resistivity . . . . .          | 14 |

|                           |    |
|---------------------------|----|
| Samplez . . . . .         | 14 |
| sat . . . . .             | 15 |
| SCV . . . . .             | 16 |
| SCV_Modified . . . . .    | 16 |
| SCV_Usual . . . . .       | 17 |
| Severity_Counts . . . . . | 18 |
| simpledata . . . . .      | 18 |
| SPD . . . . .             | 19 |
| SQ . . . . .              | 19 |
| TheWALL . . . . .         | 20 |
| VD . . . . .              | 21 |

## Index 22

---

|                |  |
|----------------|--|
| RSADBE-package | <i>Data Sets for the "R Statistical Application Development by Example" Book</i> |
|----------------|--|

---

### Description

The RSADBE package contains all the data sets used in the book "R Statistical Application Development by Example". Data sets have been collected from various sources and an attempt has been made to ensure that all the right credits are given. If some omissions are there, kindly accept the current work as a compliment for your work.

### Details

Package: RSADBE  
 Type: Package  
 Version: 1.0  
 Date: 2013-05-13  
 License: GPL-2

This package is aimed to complement the book. Any data set required in the book may simply loaded using `data(GC)` as an example.

### Author(s)

Prabhanjan  
 Maintainer: Prabhanjan Tattar <prabhanjannt@gmail.com>

### References

Tattar, P.N. (2013). R Statistical Application Development by Example. Packt Publication.

### Examples

`data(GC)`

---

Bug\_Metrics\_Software    *Bug Metrics Data*

---

**Description**

A data set which reports the 5 different type of bugs for 5 software. The count frequencies are available for pre- and post- release of the data.

**Usage**

```
data(Bug_Metrics_Software)
```

**Format**

A three dimensional array on the bug counts of 5 software at 5 severity levels.

**Source**

<http://www.eclipse.org/jdt/core/index.php>

**Examples**

```
data(Bug_Metrics_Software)
```

---

CART\_Dummy    *A cooked-data set for illustration of the partitions of CART concept*

---

**Description**

Partitions play a very important aspect of CART methodology. This data set has been cooked to translate the intuitions into partitions!

**Usage**

```
data(CART_Dummy)
```

**Format**

A data frame with 54 observations on the following 3 variables.

X1 Input variable 1

X2 Input variable 2

Y category of the output

**References**

Berk, R. A. (2008). Statistical Learning from a Regression Perspective. Springer.

**Examples**

```

data(CART_Dummy)
CART_Dummy$Y = as.factor(CART_Dummy$Y)
par(mfrow=c(1,2))
plot(c(0,12),c(0,10),type="n",xlab="X1",ylab="X2")
points(CART_Dummy$X1[CART_Dummy$Y==0],CART_Dummy$X2[CART_Dummy$Y==0],pch=15,col="red")
points(CART_Dummy$X1[CART_Dummy$Y==1],CART_Dummy$X2[CART_Dummy$Y==1],pch=19,col="green")
title(main="A Difficult Classification Problem")
plot(c(0,12),c(0,10),type="n",xlab="X1",ylab="X2")
points(CART_Dummy$X1[CART_Dummy$Y==0],CART_Dummy$X2[CART_Dummy$Y==0],pch=15,col="red")
points(CART_Dummy$X1[CART_Dummy$Y==1],CART_Dummy$X2[CART_Dummy$Y==1],pch=19,col="green")
segments(x0=c(0,0,6,6),y0=c(3.75,6.25,2.25,5),x1=c(6,6,12,12),y1=c(3.75,6.25,2.25,5),lwd=2)
abline(v=6,lwd=2)
title(main="Looks a Solvable Problem Under Partitions")

```

CT

*The Cow Temperature Data***Description**

The data set is adapted from Velleman and Hoaglin (1984). The body temperature of a cow is measured at 6:30am on 75 consecutive days. We use this data set with the intent of explaining the concept of "data smooting". The data appears on page 165 where we have 30 days body temperature.

**Usage**

```
data(CT)
```

**Format**

A data frame with 30 observations on the following 2 variables.

Day day number

Temperature temperature at 6:30am

**Source**

The entire classic book of Velleman and Hoaglin is available at [http://dspace.library.cornell.edu/bitstream/1813/78/2/A-B-C\\_of\\_EDA\\_040127.pdf](http://dspace.library.cornell.edu/bitstream/1813/78/2/A-B-C_of_EDA_040127.pdf)

**References**

Velleman, P.F., and Hoaglin, D. (1984). Applications, Basics, and Computing of Exploratory Data Analysis.

**Examples**

```

data(CT)
plot.ts(CT$Temperature,col="red",pch=1)

```

---

`DCD`*Understanding Drain Current Vs Ground-to-Source Voltage*

---

**Description**

The data pertains to an experiment where the drain current is measured against the ground-to-source voltage. We use this data set for understanding of a simple scatterplot.

**Usage**

```
data(DCD)
```

**Format**

A data frame with 10 observations on the following 2 variables.

GTS\_Voltage The voltage

Drain\_Current Drain in the current

**References**

Montgomery, D. C., and Runger, G. C. (2007). Applied Statistics and Probability for Engineers, (With CD). J.Wiley.

**Examples**

```
data(DCD)
plot(DCD)
```

---

`employ`*A data set used for understanding the very basic steps in R*

---

**Description**

The data set is used to simply understand the working of read.table, View, class and supply R functions

**Usage**

```
data(employ)
```

**Format**

A data frame with 60 observations on the following 3 variables.

Trade a numeric vector

Food a numeric vector

Metals a numeric vector

**Examples**

```
data(employ)
```

---

```
galton
```

*The famous Galton data set*

---

**Description**

Sir Francis Galton used this data set for understanding the (linear) relationship between the height of parent and its effect on the height of child.

**Usage**

```
data(galton)
```

**Format**

A data frame with 928 observations on the following 2 variables.

child children's height

parent parent's height

**Details**

A scatter plot may be used for preliminary investigation of the kind of relationship between parent's height and their children. A simple linear regression model may also be built for quantifying the relationship.

**References**

[http://en.wikipedia.org/wiki/Francis\\_Galton](http://en.wikipedia.org/wiki/Francis_Galton)

**Examples**

```
data(galton)
plot(galton)
```

---

Gasoline

*Car Mileage Dataset*

---

### **Description**

This data set has been used primarily for understanding a multivariate data set. Multiple regression model is also introduced and discussed completely through this example.

### **Usage**

```
data(Gasoline)
```

### **Format**

A data frame with 25 observations on the following 12 variables.

- y Miles per gallon
- x1 Displacement (cubic inches)
- x2 Horsepower (foot-pounds)
- x3 Torque (foot-pounds)
- x4 Compression ratio
- x5 Rear axle ratio
- x6 Carburetor (barrels)
- x7 Number of transmission speeds
- x8 Overall length (inches)
- x9 Width (inches)
- x10 Weight (pounds)
- x11 Type of transmission (A-automatic, M-manual)

### **References**

Montgomery, D. C., Peck, E.A., and Vining, G.G. (2012). Introduction to linear regression analysis. Wiley.

### **Examples**

```
data(Gasoline)
```

---

GC

*German Credit Screening Data*

---

### **Description**

Loans are an asset for the banks! However, not all the loans are promptly returned and it is thus important for a bank to build a classification model which can identify the loan defaulters from those who complete the loan tenure.

### **Usage**

`data(GC)`

### **Format**

A data frame with 1000 observations on the following 21 variables.

`checking` Status of existing checking account  
`duration` Duration in month  
`history` Credit history  
`purpose` Purpose of loan  
`amount` Credit amount  
`savings` Savings account or bonds  
`employed` Present employment since  
`installp` Installment rate in percentage of disposable income  
`marital` Personal status and sex  
`coapp` Other debtors or guarantors  
`resident` Present residence since  
`property` Property  
`age` Age in years  
`other` Other installment plans  
`housing` Housing  
`existcr` Number of existing credits at this bank  
`job` Job  
`depends` Number of people being liable to provide maintenance for  
`telephon` Telephone  
`foreign` foreign worker  
`good_bad` Loan Defaulter

### **Source**

<http://www.stat.auckland.ac.nz/~reilly/credit-g.arff> and [http://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))



**References**

[cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf](http://cran.r-project.org/doc/contrib/Sharma-CreditScoring.pdf)

**Examples**

```
data(GC)
```

---

IO\_Time

*CPU Time and IO Processes Relationship*

---

**Description**

The CPU is known to depend on the number of active IO processes. This data set will be used for the purposes of understanding scatterplots, resistant lines, and simple linear regression model.

**Usage**

```
data(IO_Time)
```

**Format**

A data frame with 10 observations on the following 2 variables.

No\_of\_IO Number of IO Processes

CPU\_Time The CPU time

**Source**

<http://www.cs.gmu.edu/~menasce/cs700/files/SimpleRegression.pdf>

**Examples**

```
data(IO_Time)  
plot(IO_Time)
```

---

lowbwt

*Low Birth Weight*

---

### **Description**

A consolidation of the concepts learnt the later half of the book is worked through using this example.

### **Usage**

```
data(lowbwt)
```

### **Format**

A data frame with 189 observations on the following 10 variables.

LOW indicator of birth weight less than 2.5kg

AGE mother's age in years

LWT mother's weight in pounds at last menstrual period

RACE mothers race ("white", "black", "other")

SMOKE smoking status during pregnancy

PTL number of previous premature labours

HT history of hypertension

UI presence of uterine irritability

FTV number of physician visits during the first trimester

BWT birth weight in grams

### **Source**

<http://www.statlab.uni-heidelberg.de/data/linmod/birthweight.html>

### **References**

Hosmer, D.W. and Lemeshow, S. (2001). Applied Logistic Regression. New York: Wiley.

### **Examples**

```
data(lowbwt)
plot(lowbwt)
```

---

MDR

*Male Death Rates*

---

### **Description**

The problem is to understand the effect of the average amount of tobacco smoked and the cause of death on the male death rates per 1000.

### **Usage**

```
data(MDR)
```

### **Format**

A data frame with 15 observations on the following 5 variables.

X Death Causes

G0 No smoking

G14 Between 1-14 grams

G24 Between 15-24 grams

G25 More than 25 grams

### **Source**

[http://dspace.library.cornell.edu/bitstream/1813/78/2/A-B-C\\_of\\_EDA\\_040127.pdf](http://dspace.library.cornell.edu/bitstream/1813/78/2/A-B-C_of_EDA_040127.pdf)

### **References**

Velleman, Paul F., and David C. Hoaglin. Applications, basics, and computing of exploratory data analysis. Vol. 142. Boston: Duxbury Press, 1981.

### **Examples**

```
data(MDR)  
boxplot(MDR)
```

---

octane

*Octane Rating Data set*

---

### Description

An experiment is conducted where the octane rating of gasoline blends can be obtained using two methods. Two samples are available for testing each type of blend, and Snee (1981) obtains 32 different blends over an appropriate spectrum of the target octane ratings.

### Usage

```
data(octane)
```

### Format

A data frame with 32 observations on the following 2 variables.

Method\_1 Ratings under Method 1

Method\_2 Ratings under Method 2

### References

Vining, G.G., and Kowalski, S.M. (2011). *Statistical Methods for Engineers*, 3e. Brooks/Cole.

### Examples

```
data(octane)
par(mfrow=c(1,2))
hist(octane$Method_1)
hist(octane$Method_2)
## maybe str(octane) ; plot(octane) ...
```

---

OF

*Understanding the Overfitting Problem*

---

### Description

This is a data set cooked up by the author to highlight the problem of overfitting. The variables have no physical meaning.

### Usage

```
data(OF)
```

**Format**

A data frame with 10 observations on the following 2 variables.

X Just another covariate

Y Just another output

**Examples**

```
data(OF)
plot(OF)
```

---

PW\_Illus

*A data set for illustrating "Piecewise Linear Regression Model"*

---

**Description**

As with the "OF" data set, this data set has been created by the author to build up the ideas leading up to piecewise linear regression model.

**Usage**

```
data(PW_Illus)
```

**Format**

A data frame with 100 observations on the following 2 variables.

X an input vector

Y an output vector

**Examples**

```
data(PW_Illus)
plot(PW_Illus)
```

---

|             |                             |
|-------------|-----------------------------|
| resistivity | <i>Resistivity of wires</i> |
|-------------|-----------------------------|

---

**Description**

The resistivity of wires is known to depend on its manufacturing process. The data set is used primarily to understand the boxplot.

**Usage**

```
data(resistivity)
```

**Format**

A data frame with 8 observations on the following 2 variables.

Process.1 Resistivity of wires under process 1

Process.2 Resistivity of wires under process 2

**References**

Gunst, R. F. (2002). Finding confidence in statistical significance. *Quality Progress*, 35 (10), 107-108.

**Examples**

```
data(resistivity)
boxplot(resistivity)
```

---

|         |                                |
|---------|--------------------------------|
| Samplez | <i>A hypothetical data set</i> |
|---------|--------------------------------|

---

**Description**

This data set shows that data may also have skewness inherent in them!

**Usage**

```
data(Samplez)
```

**Format**

A data frame with 2000 observations on the following 2 variables.

Sample\_1 a numeric vector

Sample\_2 a numeric vector

**Examples**

```
data(Samplez)
hist(Samplez$Sample_1)
hist(Samplez$Sample_2)
```

---

sat

*SAT-M marks and its impact on the final exams of a course*

---

**Description**

The final completion of a stat course is believed to depend on the marks scored by the student during his qualifying SAT-M marks. This data set is used to setup the motivation for binary regression models such as probit and logistic regression models.

**Usage**

```
data(sat)
```

**Format**

A data frame with 30 observations on the following 5 variables.

Student.No Student number

Grade Grade of the student

Pass Pass-Fail indicator in the final exam

Sat The SAT-M marks

GPP The GPP group

**References**

Johnson, Valen E., and James H. Albert. Ordinal data modeling. Springer, 1999.

**Examples**

```
data(sat)
```

---

|     |   |
|-----|---|
| SCV | <i>An illustrative data set where the "Response" depends on four variables A-D and a fifth categorical variable</i> |
|-----|---|

---

**Description**

This data set is primarily used to illustrate some basic R functions.

**Usage**

```
data(SCV)
```

**Format**

A data frame with 16 observations on the following 6 variables.

Response an output vector

A variable A

B variable B

C Variable C

D variable D

E a factor with two levels Modified Usual

**Examples**

```
data(SCV)
```

---

|              |  |
|--------------|--|
| SCV_Modified | <i>SCV data set by category "Modified"</i> |
|--------------|--|

---

**Description**

This data set is a part of the SCV dataset.

**Usage**

```
data(SCV_Modified)
```



**Format**

A data frame with 8 observations on the following 6 variables.

Response an output vector

A variable A

B variable B

C Variable C

D variable D

E a factor with two levels Modified

**Examples**

```
data(SCV_Modified)
```

---

SCV\_Usual

*SCV data set with caterogy "Usual"*

---

**Description**

This data set is part of the SCV data set.

**Usage**

```
data(SCV_Usual)
```

**Format**

A data frame with 8 observations on the following 6 variables.

Response an output vector

A variable A

B variable B

C Variable C

D variable D

E a factor with two levels Usual

**Examples**

```
data(SCV_Usual)
```

---

|                 |   |
|-----------------|---|
| Severity_Counts | <i>Severity counts for the JDT software</i> |
|-----------------|---|

---

**Description**

The software system Eclipse JDT Core has 997 different class environments related to the development. The bug identified on each occasion is classified by its severity as Bugs, NonTrivial, Major, Critical, and High. We need to understand the bug counts before- and after- software release.

**Usage**

```
data(Severity_Counts)
```

**Format**

Before and after release bug counts at five severity levels for the JDT software.

**Source**

<http://www.eclipse.org/jdt/core/index.php>

**Examples**

```
data(Severity_Counts)
barplot(Severity_Counts,xlab="Bug Count",xlim=c(0,12000), col=rep(c(2,3),5))
```

---

|               |  |
|---------------|--|
| simplifiedata | <i>A simulated data set for illustrating the ROC concept</i> |
|---------------|--|

---

**Description**

ROC is an important tool for comparing different models for the same classification problem. This data set comes with barebones infrastructure and is simply complementary in nature towards setting up a clear understanding the ROC construction.

**Usage**

```
data(simplifiedata)
```

**Format**

A data frame with 200 observations on the following 2 variables.

Predictions Predicted probabilities

Label True class of the observations

**Examples**

```
data(simplifiedata)
```

---

SPD

*The supervisor performance data*

---

**Description**

This data is used to check your understanding of the multiple linear regression model.

**Usage**

```
data(SPD)
```

**Format**

A data frame with 30 observations on the following 7 variables.

Y Supervisors performance

X1 Aspect 1

X2 Aspect 2

X3 Aspect 3

X4 Aspect 4

X5 Aspect 5

X6 Aspect 6

**References**

"Regression analysis by example" by Samprit Chatterjee and Ali S. Hadi, Wiley

**Examples**

```
data(SPD)
pairs(SPD)
```

---

SQ

*Sample Questionnaire Data*

---

**Description**

The sample questionnaire data is simply used to familiarize the reader with data and statistical terminologies.

**Usage**

```
data(SQ)
```

**Format**

A data frame with 20 observations on the following 12 variables.

Customer\_ID Customer ID

Questionnaire\_ID Questionnaire ID

Name Customers Name

Gender Customers gender Female Male

Age Age of the customer

Car\_Model Car Model's name

Car\_Manufacture\_Year Month and year of car's manufacturing

Minor\_Problems Minor problems were fixed by the workshop center indicator No Yes

Major\_Problems Major problems were fixed by the workshop center indicator No Yes Yes

Mileage The overall mileage of the car (kms/litre)

Odometer The overall kilometers travelled by the car

Satisfaction\_Rating How satisfied was the customer Very Poor < Poor < Average < Good < Very Good

**Examples**

```
data(SQ)
```

---

TheWALL

*Test centuries of Rahul Dravid*

---

**Description**

Rahul Dravid has been a modern arthictet of Indian test cricket team. His resilient centuries and holding the wicket at one end of the cricket pitch has earned him the name "The Wall". We analyze his centuries at "Home" and "Away" test matches.

**Usage**

```
data(TheWALL)
```

**Format**

A data frame with 36 observations on the following 11 variables.

Sl\_No An indicator

Score The century scores

Not\_Out\_Indicator Indicates whether Dravid remained unbeaten at the end of the team innings

Against The teams against whom Dravid scored the century

Position Dravid's batting position, out of 11

Innings An indicator of the first to fourth innings  
Test Test number  
Venue Venue of the test match  
HA\_Ind Match was in home country or away  
Date Date on the which the test began  
Result Did India won the match?

**Examples**

```
data(TheWALL)
```

---

VD

*Voltage Drop Dataset*

---

**Description**

The voltage is known to drop in a guided missile after a certain time. The data has been to illustrate certain cubic spline models.

**Usage**

```
data(VD)
```

**Format**

A data frame with 41 observations on the following 2 variables.

Time Time of missile

Voltage\_Drop Drop in the voltage

**References**

Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. Wiley, 2012.

**Examples**

```
data(VD)
```

# Index

- \* **Bar plot**
  - Bug\_Metrics\_Software, 3
- \* **Basic Tools**
  - employ, 5
- \* **Box Plot**
  - resistivity, 14
- \* **Box plot**
  - MDR, 11
- \* **CART, partitions**
  - CART\_Dummy, 3
- \* **Histogram, Stem-and-leaf plots**
  - octane, 12
- \* **Linear multiple regression model**
  - Gasoline, 7
- \* **Logistic Regression, Credit data**
  - GC, 8
- \* **Logistic Regression**
  - sat, 15
- \* **Logistic regression**
  - lowbwt, 10
- \* **Multiple linear regression**
  - SPD, 19
- \* **Overfitting**
  - OF, 12
- \* **Piece-wise Linear Regression**
  - PW\_Illus, 13
- \* **Piecewise linear regression model**
  - VD, 21
- \* **RSADBE**
  - RSADBE-package, 2
- \* **Sample Data**
  - SQ, 19
- \* **Scatter plot**
  - DCD, 5
- \* **Simple regression model**
  - IO\_Time, 9
- \* **datasets**
  - galton, 6
  - Samplez, 14
  - SCV, 16
  - SCV\_Modified, 16
  - SCV\_Usual, 17
  - Severity\_Counts, 18
  - simplifiedata, 18
  - TheWALL, 20
- \* **smoothing, hanning**
  - CT, 4
- Bug\_Metrics\_Software, 3
- CART\_Dummy, 3
- CT, 4
- DCD, 5
- employ, 5
- galton, 6
- Gasoline, 7
- GC, 8
- IO\_Time, 9
- lowbwt, 10
- MDR, 11
- octane, 12
- OF, 12
- PW\_Illus, 13
- resistivity, 14
- RSADBE (RSADBE-package), 2
- RSADBE-package, 2
- Samplez, 14
- sat, 15
- SCV, 16
- SCV\_Modified, 16
- SCV\_Usual, 17

Severity\_Counts, 18

simplifiedata, 18

SPD, 19

SQ, 19

TheWALL, 20

VD, 21