

# Data Science for Doctors – Part 2 : Descriptive Statistics



Data science enhances people's decision making. Doctors and researchers are making critical decisions every day. Therefore, it is absolutely necessary for those people to have some basic knowledge of data science. This series aims to help people that are around medical field to enhance their data science skills.

We will work with a health related database the famous “Pima Indians Diabetes Database”. It was generously donated by Vincent Sigillito from Johns Hopkins University. Please find further information regarding the dataset [here](#).

This is the second part of the series, it will contain the main descriptive statistics measures you will use most of the time. Those measures are divided in measures of central tendency and measures of spread. Moreover, most of the exercises can be solved with built-in functions, but I would encourage you to solve them “by hand”, because once you know the mechanics of the measures, then you are way more confident on using those measures. On the “solutions” page, I have both methods, so even if you didn't solve them by hand, it would be nice if you check them out.

Before proceeding, it might be helpful to look over the help pages for the mean, median, sort , unique, tabulate, sd, var, IQR, mad, abs, cov, cor, summary, str, rcorr.

You also may need to load the Hmisc library.  
`install.packages('Hmisc')`

```
library(Hmisc)
```

In case you haven't solve the [part 1](#), run the following [script](#) to load the prerequisites for this part.

Answers to the exercises are available [here](#).

If you obtained a different (correct) answer than those listed on the solutions page, please feel free to post your answer as a comment on that page.

### **Exercise 1**

Find the [mean](#) of the mass variable.

### **Exercise 2**

Find the [median](#) of the mass variable.

### **Exercise 3**

Find the [mode](#) of the mass.

### **Exercise 4**

Find the [standard deviation](#) of the age variable.



**Learn more** about descriptive statistics in the online courses [Learn by Example: Statistics and Data Science in R](#) (including 8 lectures specifically on descriptive statistics), and [Introduction to R](#).

### **Exercise 5**

Find the [variance](#) of the mass variable.

**Unlike the popular mean/standard deviation combination, interquartile range and median/mean absolute deviation are not sensitive to the presence of outliers. Even though it is recommended to go for MAD because they can approximate the standard deviation.**

### **Exercise 6**

Find the [interquartile range](#) of the age variable.

### **Exercise 7**

Find the [median absolute deviation](#) of age variable. Assume that the age follows a normal distribution.

### **Exercise 8**

Find the [covariance](#) of the variables age, mass.

### **Exercise 9**

Find the [spearman](#) and [pearson](#) correlations of the variables age, mass.

### **Exercise 10**

Print the summary statistics, and the structure of the data set. Moreover construct the correlation matrix of the data set.