**Codebook**

**Course project**

**Coursera Getting and Cleaning Data (in R)**

**Johns Hopkins University.**

Antonio Rubiera

August 28, 2019

[Description 2](#_Toc17893181)

[Data Description (Metadata) 3](#_Toc17893182)

[Subjects and Activities 3](#_Toc17893183)

[Data Column Descriptions 4](#_Toc17893184)

# Description

In this codebook we describe the data contained in the file **selected\_data\_grouped.txt**. The original dataset this data is derived is from the dataset made public by Davide Anguita, Alessandro Ghio, Luca Oneto, Xavier Parra2 and Jorge L. Reyes-Ortiz as documented in their research paper “A Public Domain Dataset for Human Activity Recognition Using Smartphones.”[[1]](#footnote-2) The original data comes in two parts: a training dataset, labeled ‘train’ in the original dataset, containing 7,352 observations, and a test dataset, labeled ‘test’ in the original dataset, containing 2,947 observations.

There are two categories of data in the original dataset. The first category is composed of a sample of the raw inertial signal data for three variables. The second category is composed of a processed sample of 561 variables. In this codebook we describe a selection from the **second** category for 79 fields measuring mean values and standard deviations. The procedure we follow is as required for this assignment:

1. We merge the training and the test data sets to create one data set.
2. We extract 79 measurements on the mean and standard deviation for each measurement from the 561 measurements in the original dataset.
3. We use descriptive activity names to name the activities in the data.
4. We label the data set columns appropriately with descriptive variable names.
5. From the data set in step 4, we extract a second, independent tidy data set with the average (men) of each variable (the 79 columns) grouped by activity and subject. It is this extracted, tidy data set that is contained in the file **selected\_data\_grouped.txt**.

The github repository for this assignment is found at: <https://github.com/rubiera/GitHubLink>

# Data Description (Metadata)

The dataset contained in the file selected\_data\_grouped.txt has four columns for subjects and activities as group by variables for the mean and standard deviation of the acceleration (m/s^2) and gyroscope measurement (m) extracted from the original dataset

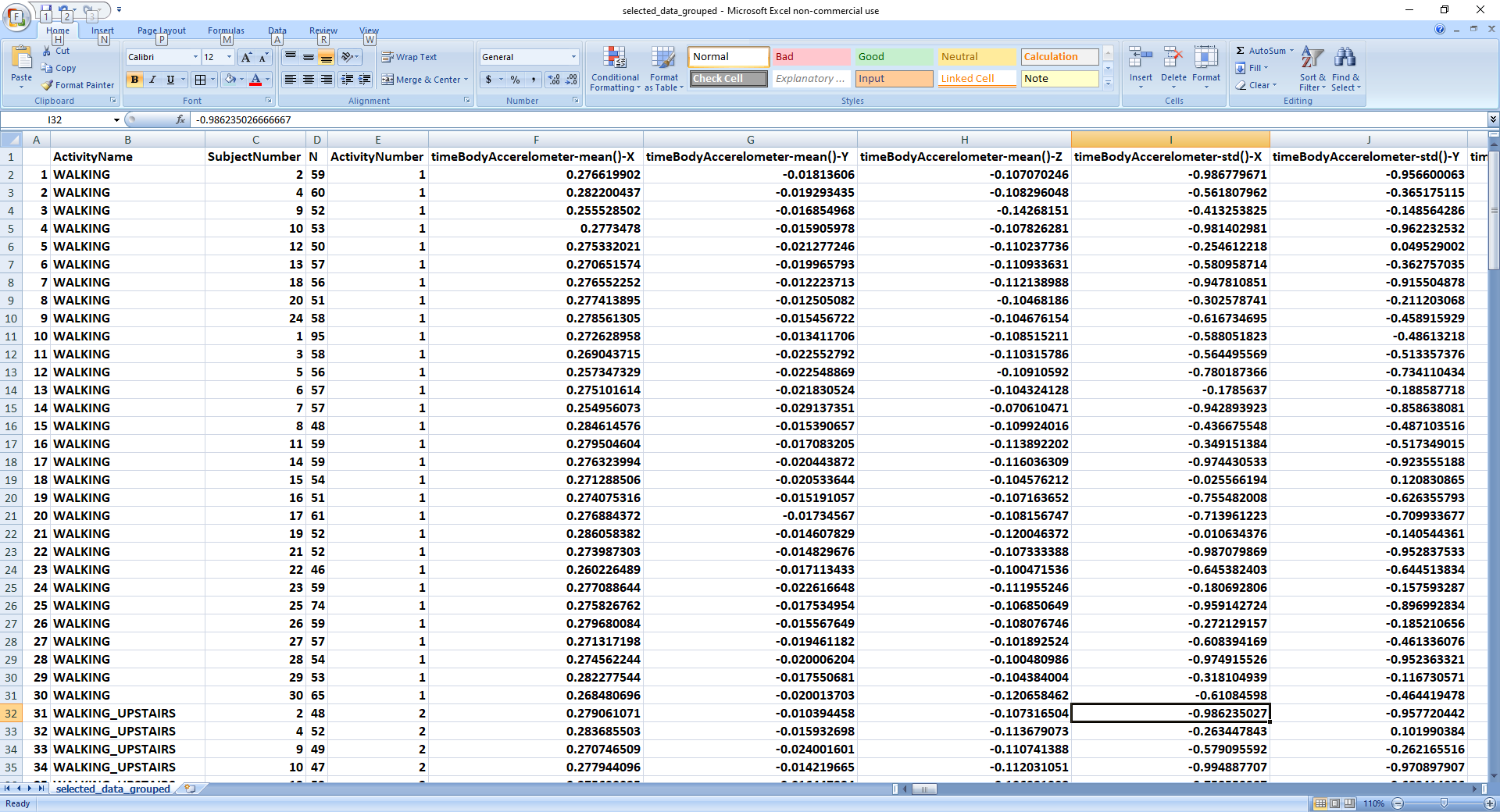
## Subjects and Activities

Subject Number is in the range 1 to 30 and found in the data as the column SubjectNumber. Activity descriptions are in the column ActivityName and activity numbers are in the column ActivityNumber.

Activity descriptions and their respective numbers are:

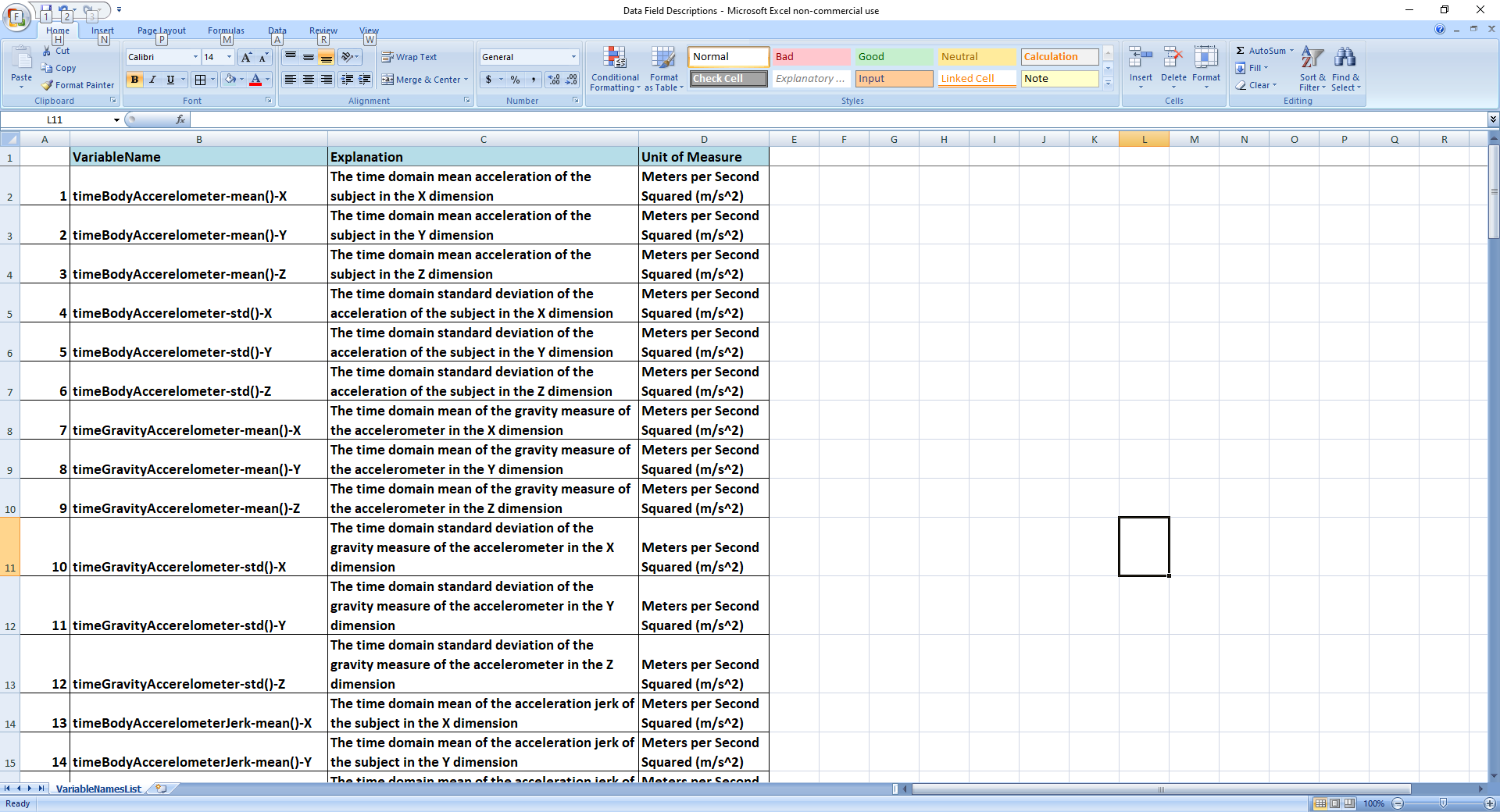
1. Activity Labels and Activity Numbers
2. WALKING
3. WALKING\_UPSTAIRS
4. WALKING\_DOWNSTAIRS
5. SITTING
6. STANDING
7. LAYING

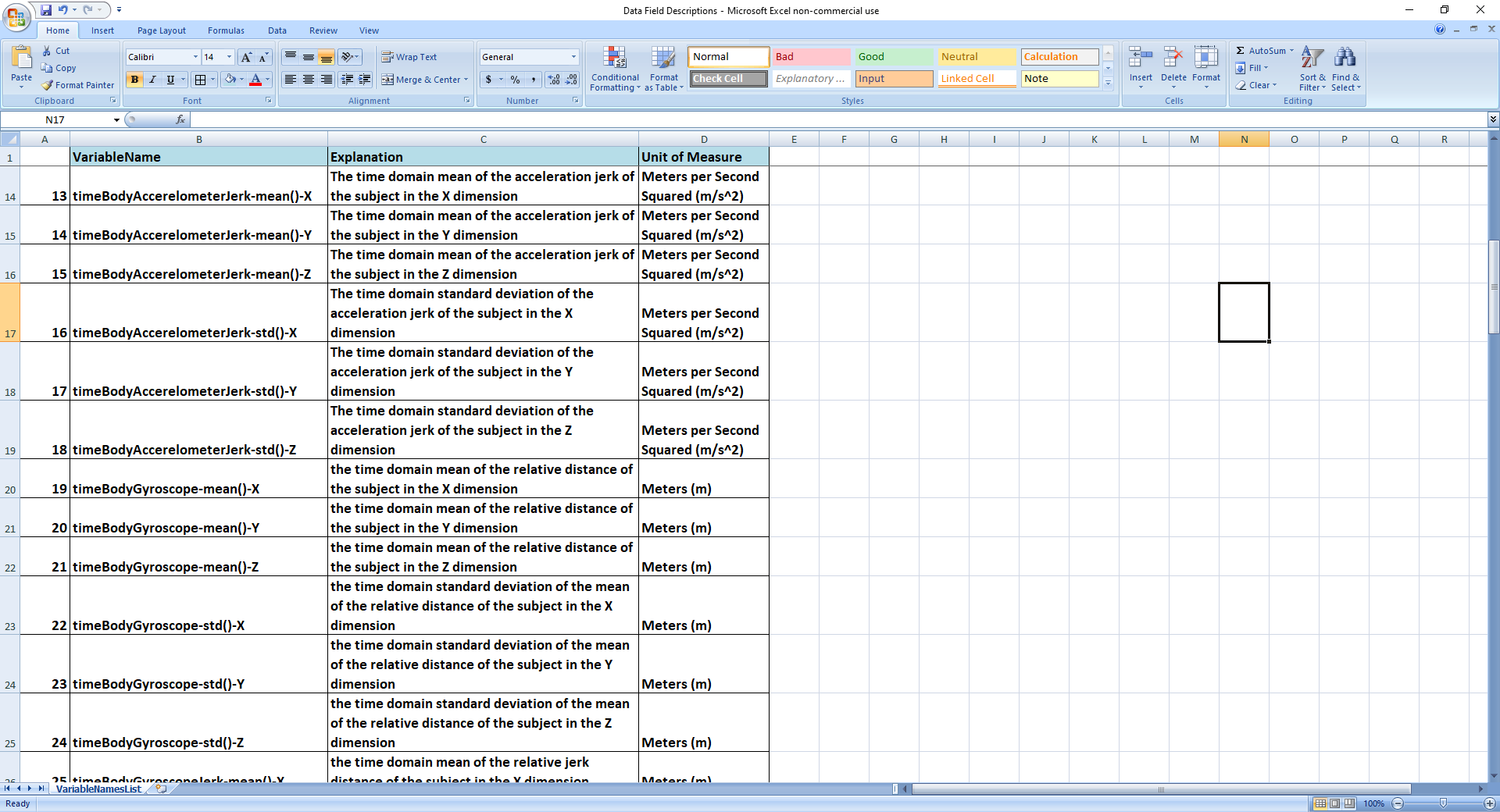
The data has a count for the number of observations N (in the range 36 to 95 Observations) for a given subject and a matching activity (subject 1 to 30 X activity 1 to 6, for a total of 180 possible groups). A snipped of the leftmost columns of the dataset (as extracted from txt into a csv file) is shown below.

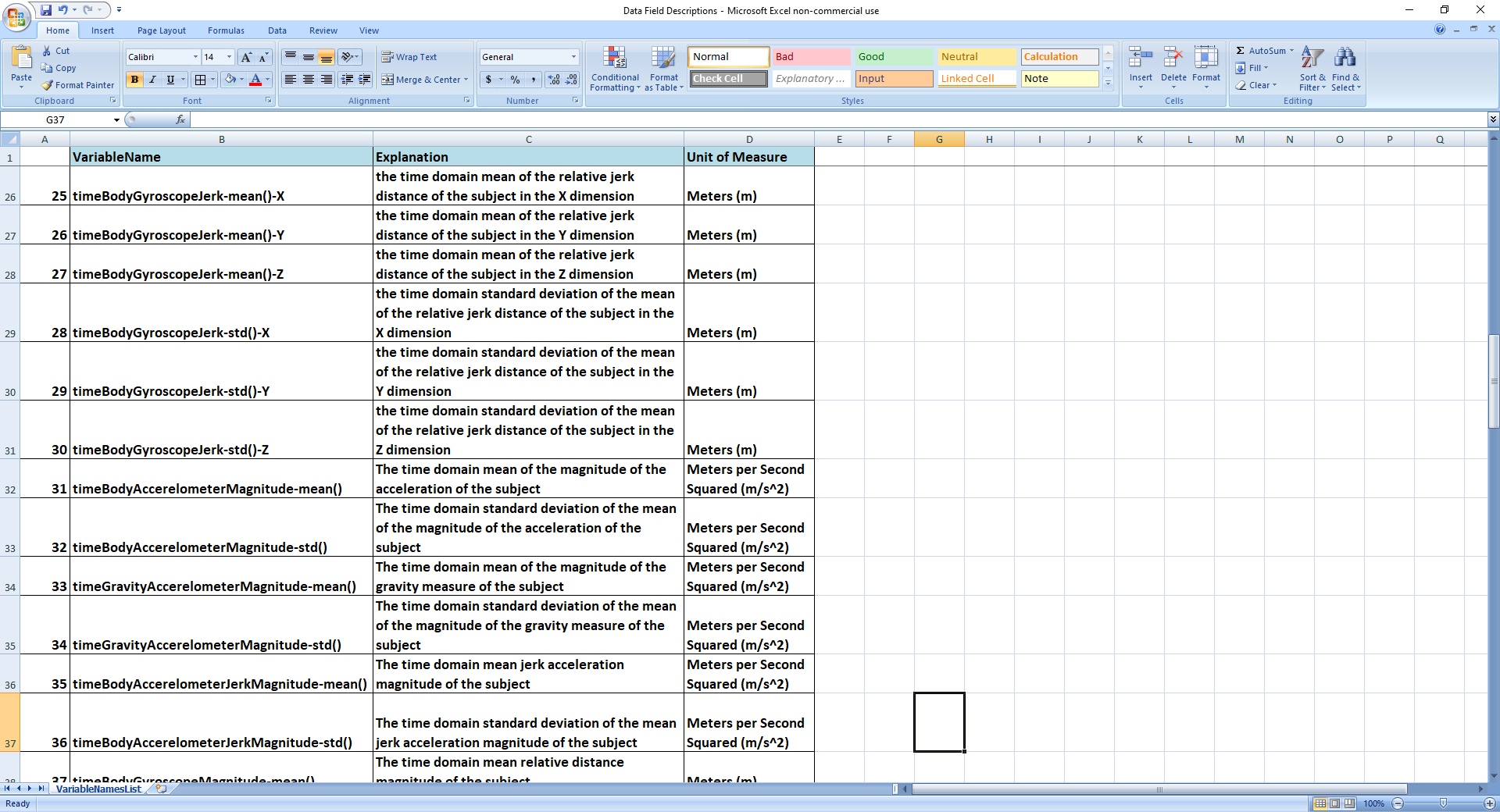


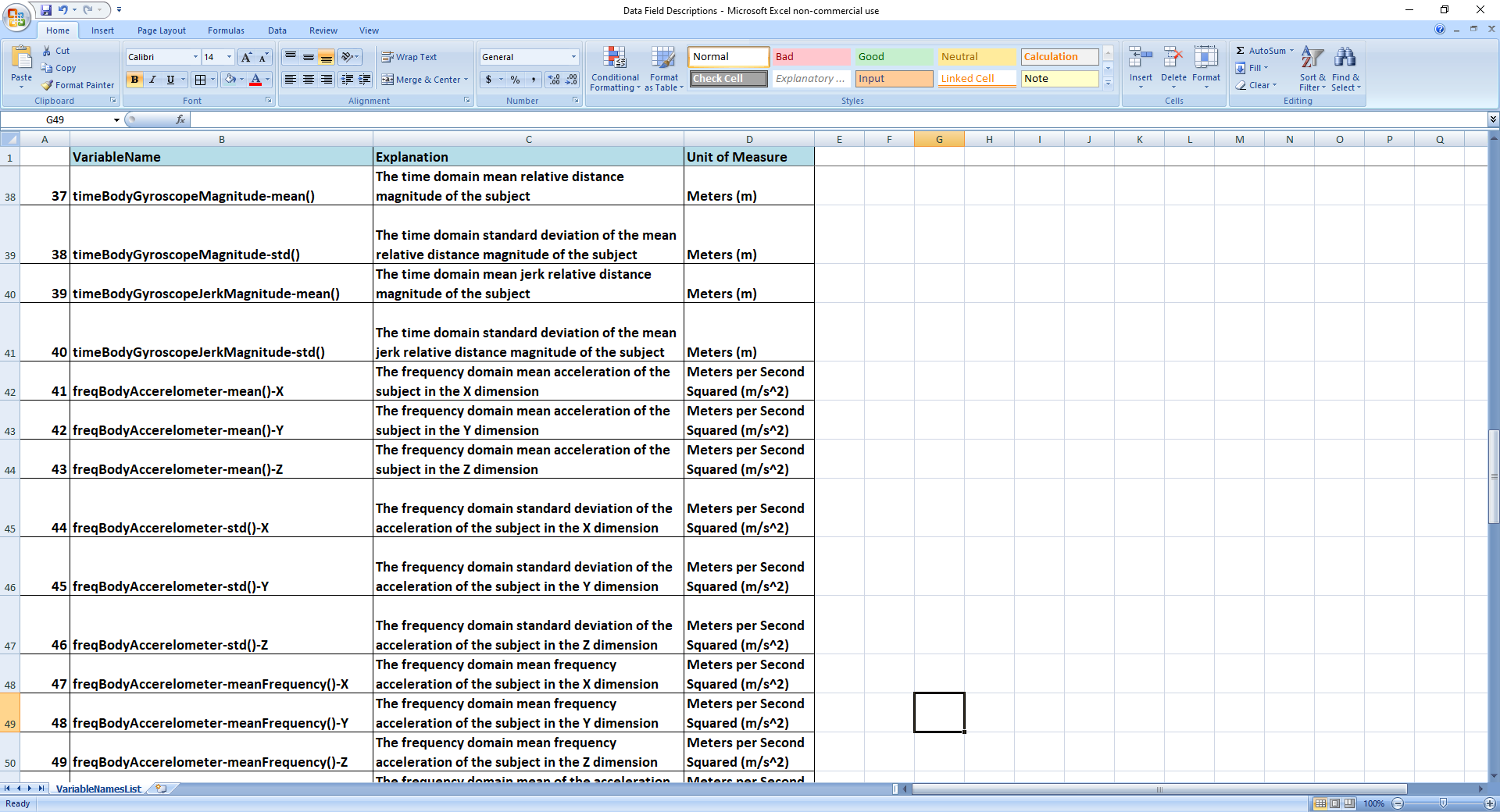
## Data Column Descriptions

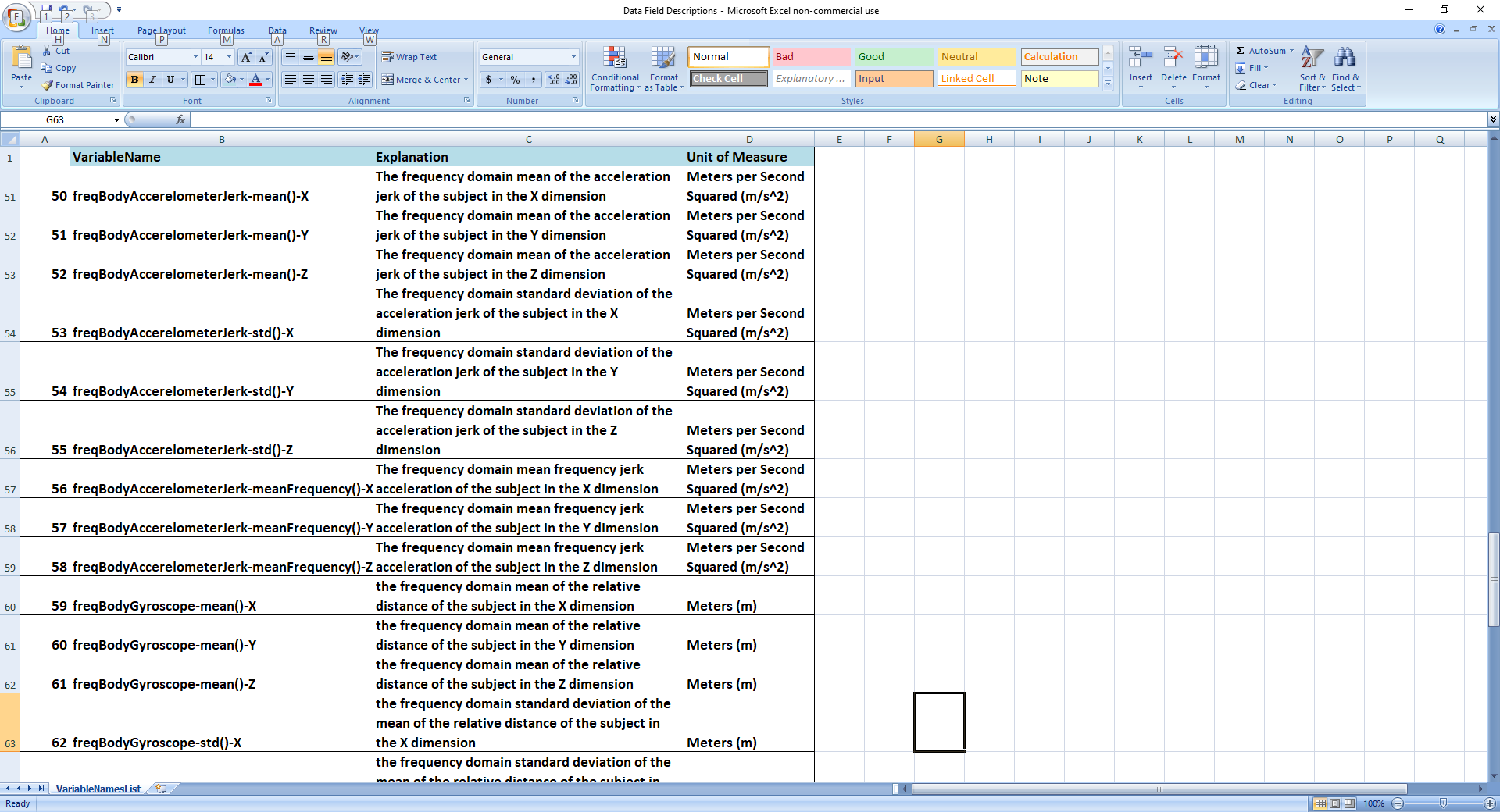
The data is grouped by subject and activity, and shows the mean of each of the 79 measurement columns. There are therefore 180 categories of means, each containing 79 measurements. The data columns are described in the following figures.

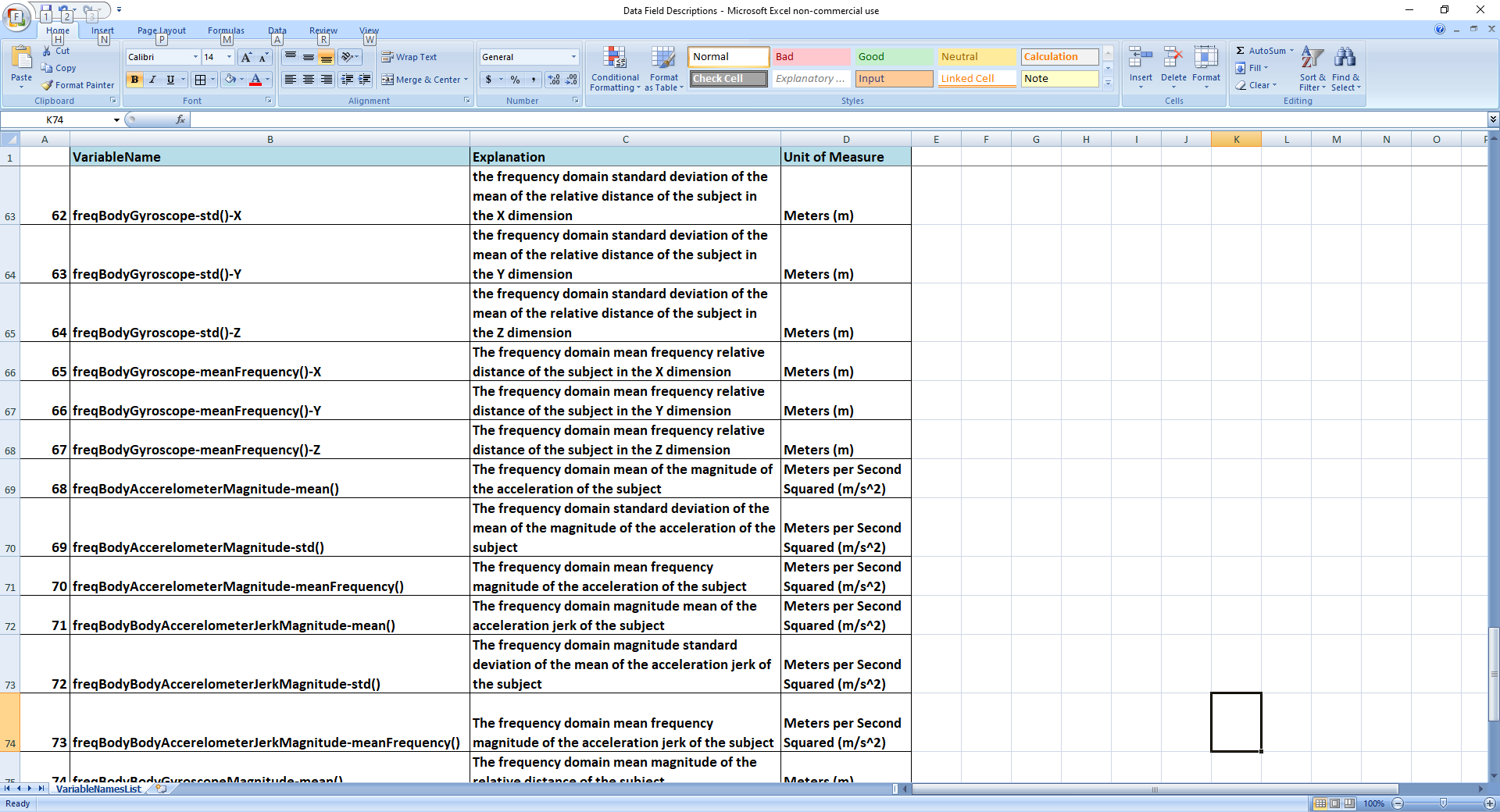


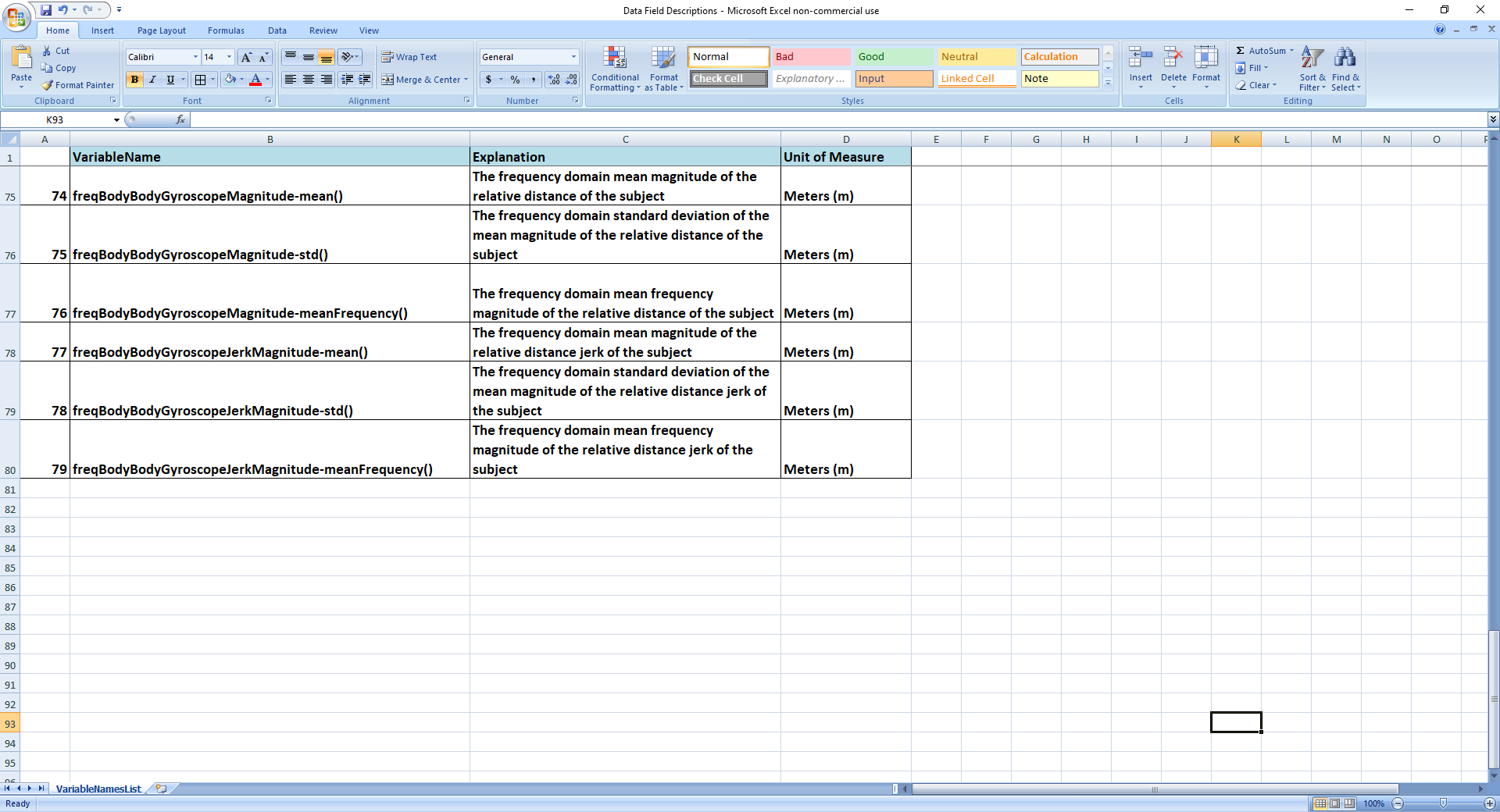
****

****

****

****

****

****

1. ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0.

   Available from <http://www.i6doc.com/en/livre/?GCOI=28001100131010> [↑](#footnote-ref-2)