# Codebook for Output of run\_analysis.R

This file is an overview of the meaning of the variables and data stored in the output from the run\_analysis.R program as specified by the final project instructions for the Getting and Cleaning Data course (or my interpretation of the instructions). This data was derived from a data set generated by having people perform certain activities while wearing a cell phone with an accelerometer and gyroscope. These two sensors were used to generate motion information which was then condensed in certain ways to produce the original data set for this assignment.

The original data comes with several files with much more detailed explanation (in particular, readme.txt and features\_info.txt). If there is any confusion in the information contained in this file, please look at these original files.

Overall, the output data file consists of two data columns to differentiate the data – Subject and Activity – and many columns of averages of various data values (not original data). All of the numeric data in the file are either based on mean values or standard deviation values of a larger data set (not even the original data that this output was generated from). Before getting into an explanation of the different numeric data, the first two variables should be explained.

“Subject” is a value to differentiate data from different individuals who performed the activities used to gather the original data. There were thirty individuals and they are numbered from 1 to 30. “Activities” is a list of six different activities that were performed in order to take measurements. In output data each different activity is given a fairly descriptive name. These are:

1. WALKING
2. WALKING\_UPSTAIRS
3. WALKING\_DOWNSTAIRS
4. SITTING
5. STANDING
6. LAYING

Remember that we have 30 individuals performing 6 different tasks. Multiply those together and we get 180, which is the number of records in our output data, meaning that we have results for every combination of individual and activity. For each of those activities numeric values are recorded for each of 66 different parameters. From the original data set, these values are the mean of all records that fall into the specific (Subject,Activity) pair in the output file. For example, for Subject 1 and the activity of Walking, the analysis function takes all of the records that match that combination and for each variable takes the mean of all of the values for that particular variable. Note that taking the mean of a mean seems odd and the mean of a standard deviation may not even be meaningful, but it is a way to condense our existing data for certain specific analysis (i.e. variation between Subjects and Activities).

Now, for an explanation of the variable names. First of all, the names are fundamentally derived (stolen) from the variable names in the original data. The fact that the variable names do not include an indication that they are in fact averages is a deficiency. The names can be found in a file called NeatDataEFCVars.txt (ignore the X at the top which is a vestigial column name). The variable names have several different parts each with a limited number of names, although not all names include all parts. In general each name can be broken down as follows

1. ‘t’ or ‘f’ – indicates time domain or frequency domain
2. “Body” or “Gravity” – indicating whether a measurement (force) is either due to body motion or gravity (the only other force acting on the Subject.
3. “Acc” or “Gyro” – indicating whether the data was coming from an accelerometer or a gyroscope.
4. “Jerk” – not always present; indicating whether the measurement had something to do with sudden movement (I think)
5. “Mag” – indicates a single magnitude, rather than vector (X,Y,Z) data
6. “mean” or “std” – the only two types of data pulled from the original data set. The mean of some data collection or the Standard Deviation around that mean. Usually the two can be matched up fairly clearly.
7. “X” – indicates the X component of some vector data
8. “Y” – indicates the Y component of some vector data
9. “Z” – indicates the Z component of some vector data

There are some “.” symbols (periods) in the variable names at various places. This is a default character that was substituted for other characters (usually ‘(‘ and ‘)’) that could not be read in correctly when the variable names were read from features.txt. It is another deficiency that could be cleaned up with more work on the analysis program. For the most part the periods do not really obscure the variable name but just make it a look a little odd.

As an example, the variable name “tBodyGyroJerk.mean…X” is a time domain variable (t) from the Gyroscopic sensor (Gyro), meaning It’s a record of angular acceleration, of the body (Body) and it has something to do with the jerking force (Jerk) of the subjects motion. This particular variable is part of a vector and records the X dimension and, finally, the data is a mean value (remember, this means that the original data was a mean; all of the values in the final output file are means).

Look at the variable file to get an idea of the various combination of the parts listed above. The resulting names are not an exhaustive permutation of the parts described, but it is beyond the scope of this assignment to explain why certain combinations are useful whereas other combinations are not.