CodeBook for Week 3 Project

Getting and Cleaning Data Course Coursera

Date: 21 Feb 2015

Author: Randeep Grewal

Project summary Given a set of raw data files comprising of Samsung Galaxy S measurements, in two data sets (training and test) combine them and process them to a tidy dataset.

Then select the columns comprising mean or standard deviation data and average them by subject and group.

Further information on the data is at [http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones]

Raw Data Files The raw data is downloaded from the following url [http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones]

After unzipping it created a directory 'UCI HAR Dataset'

Task details The following is from the project task website

You should create one R script called run analysis. R that does the following.

- 1. Merges the training and the test sets to create one data set.
- 2. Extracts only the measurements on the mean and standard deviation for each measurement.
- 3. Uses descriptive activity names to name the activities in the data set
- 4. Appropriately labels the data set with descriptive variable names.
- 5. From the data set in step 4, creates a second, independent tidy data set with the average of each variable for each activity and each subject.

Code

My code is broken into steps as above plus step 0. However my steps are in a slightly different order from the task as I felt it was more logical (but for consistency I kept the original task step numbers).

Step 0 - Downloading raw data, unzipping and loading into R This is fairly self-explanatory from the code. I use the following variables:

activity_labels -this provides the label variables we will need later

features -this is the vector which contains the column labels for the X datasets

Steps 1,3,4 - Merge training and test sets, use descriptive activity names and label with descriptive variable names The two subdirectories (test and training) each contain 3 critical files subject, X and Y These are loaded to subject_test, x_test, y_test or subject_training, x_training and y_training accordingly.

It is important to recognise that the final dataframe we want comprises a cbind of subject, y, x after the test and training datasets have gone through rbind. However the values in y need to be replaced with the values in activity_labels. (By x, y and subject I mean x_test and x_training etc)

The features vector contains the labels of the x datasets.

For step 4 I chose to use CamelCases. Furthermore I convert the leading 'f' to Freq, and the leading 'T' to Time. For further tidiness and to ensure that the column names are acceptable syntax in R we convert '-std()' to 'Std' and '-mean()' to 'Mean'. The exact code is here: (experimenting re inserting code!)

```
col_names <- names(data)
col_names <- gsub("^t","Time", col_names)
col_names <- gsub("^f","Freq", col_names)
col_names <- gsub("-std\\(\\)-|-std\\(\\)","Std",col_names)
col_names <- gsub("-mean\\(\\)-|-mean\\(\\)","Mean", col_names)
colnames(data) <- col_names</pre>
```

Step 2 - Extract only the measurements on the mean and the standard deviation for each measurement I chose to do this step relatively late in my code as it helped me understand what was going on by labelling the dataset first.

There is some discussion in the forums about whether this means we extract every occurrance of mean and standard deviation or only those where the label is mean() or std(). Based on the discussion in the forum I chose to chose only the latter option. See [https://class.coursera.org/getdata-011/forum/thread?thread_id=19]

My key decision is reflected in these line:

By changing the first of these two lines we could change the number of columns the code returns.

By using the above options I get 68 columns - of which 66 are the data from the X dataframes and the other two are subject (= the previous subject dataset) and activity (= the previous Y dataset). Others, by using more permissive searches for 'mean' and 'std' have obtained upto 79 columns (which I did before I looked at the data columns and decided to exclude columns such as 'angle(tBodyGyroJerkMean,gravityMean)')

Step 5 - Independent tidy data set with the average of each varible for each activity and each subject My interpretation of this is that this requires applying the mean function on a dataset that is grouped by both activity and subject. Hence given that there are 30 subjects, each of whom does 6 activities we will get 180 values which indeed the output dataset delivers.

The key lines of code are:

```
newdata <- group_by(data, Subject,Activity)
data_output <- summarise_each(newdata,funs(mean))</pre>
```

The dataset data_output is written to a file using write.table. I believe it is a tidy dataset as it has 180 rows (ie 30 subjects x 6 activities). There are 68 columns with each column representing the mean of a measured variable.

Code Book

Summary of input The input file files are explained in my text up. The mappings from eg tBodyAccmean()-z to TimeBodyAccMeanZ should be fairly obvious from the context. Nevertheless for completeness I list the input fields from the *features* file (at the end of this document) as they give the body of variables for the X data sets. The subject and y datasets are self explanatory from my explanations above.

(See end of document for input fields - put there to make it look tidier and more readable)

Summary of output The list below provides the variables that the code writes to 'tidy_data.txt'. The first variable *Subject* is the id for the subject. The second varible *Activity* describes the activity ("WALK-ING","WALKING_UPSTAIR","WALKING_DOWNSTAIRS","SITTING","STANDING","LAYING").

For all the other variables the data is the means of the Time or Freq based data from the accelerometer or gyroscope within the phone.

The following is the complete list of the variables (the variable number is for convenience only and not part of the variable name):

- 1 Subject
- 2 Activity
- 3 TimeBodyAccMeanX
- 4 TimeBodyAccMeanY
- 5 TimeBodyAccMeanZ
- 6 TimeBodyAccStdX
- 7 TimeBodyAccStdY
- 8 TimeBodyAccStdZ
- 9 TimeGravityAccMeanX
- 10 TimeGravityAccMeanY
- 11 TimeGravityAccMeanZ
- 12 TimeGravityAccStdX
- 13 TimeGravityAccStdY
- 14 TimeGravityAccStdZ
- 15 TimeBodyAccJerkMeanX
- 16 TimeBodyAccJerkMeanY
- 17 TimeBodyAccJerkMeanZ
- 18 TimeBodyAccJerkStdX
- 19 TimeBodyAccJerkStdY
- 20 TimeBodyAccJerkStdZ
- 21 TimeBodyGyroMeanX
- 22 TimeBodyGyroMeanY
- 23 TimeBodyGyroMeanZ
- 24 TimeBodyGyroStdX
- 25 TimeBodyGyroStdY

- 26 TimeBodyGyroStdZ
- 27 TimeBodyGyroJerkMeanX
- 28 TimeBodyGyroJerkMeanY
- 29 TimeBodyGyroJerkMeanZ
- 30 TimeBodyGyroJerkStdX
- 31 TimeBodyGyroJerkStdY
- 32 TimeBodyGyroJerkStdZ
- 33 TimeBodyAccMagMean
- 34 TimeBodyAccMagStd
- $35~{\rm Time Gravity Acc Mag Mean}$
- $36~{\rm TimeGravityAccMagStd}$
- $37~{\rm TimeBodyAccJerkMagMean}$
- 38 TimeBodyAccJerkMagStd
- $39~{\rm TimeBodyGyroMagMean}$
- 40 TimeBodyGyroMagStd
- 41 TimeBodyGyroJerkMagMean
- 42 TimeBodyGyroJerkMagStd
- 43 FreqBodyAccMeanX
- 44 FreqBodyAccMeanY
- 45 FreqBodyAccMeanZ
- 46 FreqBodyAccStdX
- 47 FreqBodyAccStdY
- 48 FreqBodyAccStdZ
- 49 FreqBodyAccJerkMeanX
- 50 FreqBodyAccJerkMeanY
- 51 FreqBodyAccJerkMeanZ
- 52 FreqBodyAccJerkStdX
- 53 FreqBodyAccJerkStdY
- 54 FreqBodyAccJerkStdZ
- 55 FreqBodyGyroMeanX
- 56 FreqBodyGyroMeanY
- 57 FreqBodyGyroMeanZ
- 58 FreqBodyGyroStdX
- 59 FreqBodyGyroStdY
- 60 FreqBodyGyroStdZ
- 61 FreqBodyAccMagMean

- 62 FreqBodyAccMagStd
- 63 FreqBodyBodyAccJerkMagMean
- 64 FreqBodyBodyAccJerkMagStd
- 65 FreqBodyBodyGyroMagMean
- 66 FreqBodyBodyGyroMagStd
- 67 FreqBodyBodyGyroJerkMagMean
- 68 FreqBodyBodyGyroJerkMagStd

Listing of original input variables The original input features list (which gives the variable names for X) is below: (In order to save space not very neat)

V1 V2

1 1 tBodyAcc-mean()-X 2 2 tBodyAcc-mean()-Y 3 3 tBodyAcc-mean()-Z 4 4 tBodyAcc-std()-X 5 5 tBodyAccstd()-Y 6 6 tBodyAcc-std()-Z 7 7 tBodyAcc-mad()-X 8 8 tBodyAcc-mad()-Y 9 9 tBodyAcc-mad()-Z 10 10 tBodyAcc-max()-X 11 11 tBodyAcc-max()-Y 12 12 tBodyAcc-max()-Z 13 13 tBodyAcc-min()-X 14 14 tBodyAcc-min()-Y 15 15 tBodyAcc-min()-Z 16 16 tBodyAcc-sma() 17 17 tBodyAcc-energy()-X 18 18 tBodyAcc-energy()-Y 19 19 tBodyAcc-energy()-Z 20 20 tBodyAcc-iqr()-X 21 21 tBodyAcc-iqr()-Y 22 22 tBodyAcc-iqr()-Z 23 23 tBodyAcc-entropy()-X 24 24 tBodyAcc-entropy()-Y 25 25 tBodyAcc-entropy()-Z 26 26 tBodyAcc-arCoeff()-X,1 27 27 tBodyAcc-arCoeff()-X,2 28 28 tBodyAcc-arCoeff()-X,3 29 29 arCoeff()-Y,3 33 33 tBodyAcc-arCoeff()-Y,4 34 34 tBodyAcc-arCoeff()-Z,1 35 35 tBodyAcc-arCoeff()-Z,2 36 36 tBodyAcc-arCoeff()-Z,3 37 37 tBodyAcc-arCoeff()-Z,4 38 38 tBodyAcc-correlation()-X,Y 39 39 tBodyAcccorrelation()-X,Z 40 40 tBodyAcc-correlation()-Y,Z 41 41 tGravityAcc-mean()-X 42 42 tGravityAcc-mean()-Y 43 43 tGravityAcc-mean()-Z 44 44 tGravityAcc-std()-X 45 45 tGravityAcc-std()-Y 46 46 tGravityAcc-std()-Z 47 47 tGravityAcc-mad()-X 48 48 tGravityAcc-mad()-Y 49 49 tGravityAcc-mad()-Z 50 50 tGravityAccmax()-X 51 51 tGravityAcc-max()-Y 52 52 tGravityAcc-max()-Z 53 53 tGravityAcc-min()-X 54 54 tGravityAcc-min()-Y 55 55 tGravityAcc-min()-Z 56 56 tGravityAcc-sma() 57 57 tGravityAcc-energy()-X 58 58 tGravityAcc-energy()-Y 59 59 tGravityAcc-energy()-Z 60 60 tGravityAcc-iqr()-X 61 61 tGravityAcc-iqr()-Y 62 62 tGravityAcc-iqr()-Z 63 63 tGravityAcc-entropy()-X 64 64 tGravityAcc-entropy()-Y 65 65 tGravityAccentropy()-Z 66 66 tGravityAcc-arCoeff()-X,1 67 67 tGravityAcc-arCoeff()-X,2 68 68 tGravityAcc-arCoeff()-X,3 69 69 tGravityAcc-arCoeff()-X,4 70 70 tGravityAcc-arCoeff()-Y,1 71 71 tGravityAcc-arCoeff()-Y,2 72 72 tGravityAcc-arCoeff()-Y,3 73 73 tGravityAcc-arCoeff()-Y,4 74 74 tGravityAcc-arCoeff()-Z,1 75 75 tGravitvAcc-arCoeff()-Z.2 76 76 tGravitvAcc-arCoeff()-Z.3 77 77 tGravitvAcc-arCoeff()-Z.4 78 78 tGravityAcc-correlation()-X,Y 79 79 tGravityAcc-correlation()-X,Z 80 80 tGravityAcc-correlation()-Y,Z 81 81 tBodyAccJerk-mean()-X 82 82 tBodyAccJerk-mean()-Y 83 83 tBodyAccJerk-mean()-Z 84 84 tBodyAccJerkstd()-X 85 85 tBodyAccJerk-std()-Y 86 86 tBodyAccJerk-std()-Z 87 87 tBodyAccJerk-mad()-X 88 88 tBodyAccJerk-mad()-Y 89 89 tBodyAccJerk-mad()-Z 90 90 tBodyAccJerk-max()-X 91 91 tBodyAccJerkmax()-Y 92 92 tBodyAccJerk-max()-Z 93 93 tBodyAccJerk-min()-X 94 94 tBodyAccJerk-min()-Y 95 95 tBodyAccJerk-min()-Z 96 96 tBodyAccJerk-sma() 97 97 tBodyAccJerk-energy()-X 98 98 tBodyAccJerkenergy()-Y 99 99 tBodyAccJerk-energy()-Z 100 100 tBodyAccJerk-iqr()-X 101 101 tBodyAccJerk-iqr()-Y 102 102 tBodyAccJerk-iqr()-Z 103 103 tBodyAccJerk-entropy()-X 104 104 tBodyAccJerk-entropy()-Y 105 105 tBodyAccJerk-entropy()-Z 106 106 tBodyAccJerk-arCoeff()-X,1 107 107 tBodyAccJerk-arCoeff()-X,2 108 108 tBodyAccJerk-arCoeff()-X,3 109 109 tBodyAccJerk-arCoeff()-X,4 110 110 tBodyAccJerk-arCoeff()-Y,1 111 111 tBodyAccJerk-arCoeff()-Y,2 112 112 tBodyAccJerk-arCoeff()-Y,3 113 113 tBodyAccJerk-arCoeff()-Y,4 114 114 tBodyAccJerk-arCoeff()-Z,1 115 115 tBodyAccJerk-arCoeff()-Z,2 116 116 tBodyAccJerk-arCoeff()-Z,3 117 117 tBodyAccJerk-arCoeff()-Z,4 118 118 tBodyAccJerk-correlation()-X,Y 119 119 tBodyAccJerkcorrelation()-X,Z 120 120 tBodyAccJerk-correlation()-Y,Z 121 121 tBodyGyro-mean()-X 122 122 tBodyGyro-mean()-Y 123 123 tBodyGyro-mean()-Z 124 124 tBodyGyro-std()-X 125 125 tBodyGyro-std()-Y

Y 448 448 fBodyGyro-entropy()-Z 449 449 fBodyGyro-maxInds-X 450 450 fBodyGyro-maxInds-Y 451 451 fBodyGyro-maxInds-Z 452 452 fBodyGyro-meanFreq()-X 453 453 fBodyGyro-meanFreq()-Y 454 454 fBodyGyro-meanFreq()-Z 455 455 fBodyGyro-skewness()-X 456 456 fBodyGyro-kurtosis()-X 457 457 fBodyGyro-skewness()-Y 458 458 fBodyGyro-kurtosis()-Y 459 459 fBodyGyro-skewness()-Z 460 460 fBodyGyro-kurtosis()-Z 461 461 fBodyGyro-bandsEnergy()-1,8 462 462 fBodyGyro-bandsEnergy()-9,16 463 463 fBodyGyro-bandsEnergy()-17,24 464 464 fBodyGyro-bandsEnergy()-25,32 465 465 fBodyGyrobandsEnergy()-33.40 466 466 fBodyGyro-bandsEnergy()-41.48 467 467 fBodyGyro-bandsEnergy()-49.56 468 468 fBodyGyro-bandsEnergy()-57,64 469 469 fBodyGyro-bandsEnergy()-1,16 470 470 fBodyGyrobandsEnergy()-17,32 471 471 fBodyGyro-bandsEnergy()-33,48 472 472 fBodyGyro-bandsEnergy()-49,64 473 473 fBodyGyro-bandsEnergy()-1,24 474 474 fBodyGyro-bandsEnergy()-25,48 475 475 fBodyGyrobandsEnergy()-1,8 476 476 fBodyGyro-bandsEnergy()-9,16 477 477 fBodyGyro-bandsEnergy()-17,24 478 478 fBodyGyro-bandsEnergy()-25,32 479 479 fBodyGyro-bandsEnergy()-33,40 480 480 fBodyGyrobandsEnergy()-41,48 481 481 fBodyGyro-bandsEnergy()-49,56 482 482 fBodyGyro-bandsEnergy()-57,64 483 483 fBodyGyro-bandsEnergy()-1,16 484 484 fBodyGyro-bandsEnergy()-17,32 485 485 fBodyGyrobandsEnergy()-33,48 486 486 fBodyGyro-bandsEnergy()-49,64 487 fBodyGyro-bandsEnergy()-1,24 488 488 fBodyGyro-bandsEnergy()-25,48 489 489 fBodyGyro-bandsEnergy()-1,8 490 490 fBodyGyrobandsEnergy()-9,16 491 491 fBodyGyro-bandsEnergy()-17,24 492 492 fBodyGyro-bandsEnergy()-25,32 493 493 fBodyGyro-bandsEnergy()-33.40 494 494 fBodyGyro-bandsEnergy()-41.48 495 495 fBodyGyrobandsEnergy()-49,56 496 496 fBodyGyro-bandsEnergy()-57,64 497 497 fBodyGyro-bandsEnergy()-1,16 498 498 fBodyGyro-bandsEnergy()-17,32 499 499 fBodyGyro-bandsEnergy()-33,48 500 500 fBodyGyrobandsEnergy()-49,64 501 501 fBodyGyro-bandsEnergy()-1,24 502 502 fBodyGyro-bandsEnergy()-25,48 503 503 fBodyAccMag-mean() 504 504 fBodyAccMag-std() 505 505 fBodyAccMag-mad() 506 506 fBodyAccMag-max() 507 507 fBodyAccMag-min() 508 508 fBodyAccMag-sma() 509 509 fBodyAccMagenergy() 510 510 fBodyAccMag-iqr() 511 511 fBodyAccMag-entropy() 512 512 fBodyAccMag-maxInds 513 513 fBodyAccMag-meanFreq() 514 514 fBodyAccMag-skewness() 515 515 fBodyAccMag-kurtosis() 516 516 fBodyBodyAccJerkMag-mean() 517 517 fBodyBodyAccJerkMag-std() 518 518 fBodyBodyAccJerkMag-mad() 519 519 fBodyBodyAccJerkMag-max() 520 520 fBodyBodyAccJerkMag-min() 521 521 fBodyBodyAccJerkMagsma() 522 522 fBodyBodyAccJerkMag-energy() 523 523 fBodyBodyAccJerkMag-iqr() 524 524 fBodyBodyAccJerkMag-entropy() 525 525 fBodyBodyAccJerkMag-maxInds 526 526 fBodyBodyAccJerkMagmeanFreq() 527 527 fBodvBodvAccJerkMag-skewness() 528 528 fBodvBodvAccJerkMag-kurtosis() 529 529 fBodyBodyGyroMag-mean() 530 530 fBodyBodyGyroMag-std() 531 531 fBodyBodyGyroMag-mad() 532 532 fBodyBodyGyroMag-max() 533 533 fBodyBodyGyroMag-min() 534 534 fBodyBodyGyroMag-sma() 535 535 fBodyBodyGyroMag-energy() 536 536 fBodyBodyGyroMag-iqr() 537 537 fBodyBodyGyroMag-entropy() 538 538 fBodyBodyGyroMag-maxInds 539 539 fBodyBodyGyroMag-meanFreq() 540 540 fBodyBodyGyroMagskewness() 541 541 fBodyBodyGyroMag-kurtosis() 542 542 fBodyBodyGyroJerkMag-mean() 543 543 fBodyBodyGyroJerkMag-std() 544 544 fBodyBodyGyroJerkMag-mad() 545 545 fBodyBodyGyroJerkMagmax() 546 546 fBodyBodyGyroJerkMag-min() 547 547 fBodyBodyGyroJerkMag-sma() 548 548 fBodyBodyGyroJerkMag-energy() 549 549 fBodyBodyGyroJerkMag-iqr() 550 550 fBodyBodyGyroJerkMagentropy() 551 551 fBodyBodyGyroJerkMag-maxInds 552 552 fBodyBodyGyroJerkMag-meanFreq() 553 553 fBodyBodyGyroJerkMag-skewness() 554 554 fBodyBodyGyroJerkMag-kurtosis() 555 555 angle(tBodyAccMean,gravity) 556 556 angle(tBodyAccJerkMean),gravityMean) 557 557 angle(tBodyGyroMean,gravityMean) 558 558 angle(tBodyGyroJerkMean,gravityMean) 559 559 angle(X,gravityMean) 560 560 angle(Y,gravityMean) 561 561 angle(Z,gravityMean)

8