Tidy Data Practice

Christopher Han

February 20, 2019

This is a simple exercise of importing raw data and transforming it into tidy data and is part of the week 4 assignment in Getting and Cleaning Data from Data Science Specialization by Johns Hopkins University on Coursera.

The dataset used is from UCI Machine Learning Repository on the topic of "Human Activity Recognition Using Smartphones Data Set".

There are two files used for this exercise: functions.R and run_analysis.R

functions.R

This script contains two functions used in run_analysis.R.

1. cleandata() - imports the raw files, combines them into one data frame and formatted to be tidy.

```
cleandata <- function(){
        # Read data
        setwd("C:/Users/Chris Han/tidy-data-practice/UCI HAR Dataset")
        features <- read.table("features.txt", stringsAsFactors = FALSE)
        train <- read.table("train/X_train.txt")
        trainlabels <- read.table("train/y_train.txt")
        trainsubject <- read.table("train/subject_train.txt")
        test <- read.table("test/X_test.txt")
        testlabels <- read.table("test/y_test.txt")
        testsubject <- read.table("test/subject_test.txt")

        # Assign feature names to columns, choose only features with mean/std
        colnames(train) <- features$V2
        trainmeanstdcol <- grepl("mean()\\b|std()\\b", colnames(train))
        sum(trainmeanstdcol)
        train <- train[,trainmeanstdcol]
        train <- cbind(trainlabels, train)
        train <- cbind(trainsubject, train)

        # Repeat for test data
        colnames(test) <- features$V2
        testmeanstdcol <- grepl("mean()\\b|std()\\b", colnames(test))
        sum(testmeanstdcol)
        test <- test[,testmeanstdcol]
        test <- cbind(testlabels, test)
        test <- cbind(testsubject, test)

        # Combine train and test data
        fulldata <- rbind(train, test)
        colnames(fulldata)[1] <- "subject"
        colnames(fulldata)[2] <- "activity"
        fulldata
}
```

2. labeldescriptivename() - modifies the column names to be more descriptive and easier to interpret.

```r
labeldescriptivename <- function(fulldata){
        renamedata <- fulldata
        colnames(renamedata)
        colnames(renamedata) <- tolower(colnames(renamedata))
        colnames(renamedata) <- gsub("-", "", colnames(renamedata))
        colnames(renamedata) <- sub("^t", "TimeDomain", colnames(renamedata))
        colnames(renamedata) <- sub("^f", "FrequencyDomain", colnames(renamedata))
        colnames(renamedata) <- sub("acc", "Acceleration", colnames(renamedata))
        colnames(renamedata) <- sub("()", "", colnames(renamedata), fixed = TRUE)
        colnames(renamedata) <- sub("mag", "Magnitude", colnames(renamedata))
        colnames(renamedata) <- gsub("body", "Body", colnames(renamedata))
        colnames(renamedata) <- sub("gravity", "Gravity", colnames(renamedata))
        colnames(renamedata) <- sub("x$", "_X_Axis", colnames(renamedata))
        colnames(renamedata) <- sub("y$", "_Y_Axis", colnames(renamedata))
        colnames(renamedata) <- sub("z$", "_Z_Axis", colnames(renamedata))
        colnames(renamedata) <- sub("mean", "_Mean", colnames(renamedata))
        colnames(renamedata) <- sub("std", "_StandardDeviation", colnames(renamedata))
        colnames(renamedata) <- sub("gyro", "Gyroscope", colnames(renamedata))
        colnames(renamedata) <- sub("jerk", "Jerk", colnames(renamedata))
        colnames(renamedata) <- sub("BodyBody", "Body", colnames(renamedata))
        colnames(renamedata)[2] <- "activity"
        renamedata
}
```

run_analysis.R

This script accomplishes 5 things.

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.

```r
library(dplyr)
setwd("C:/Users/Chris Han/tidy-data-practice")
source("functions.R")
setwd("C:/Users/Chris Han/tidy-data-practice/UCI HAR Dataset")

# 1 and 2
# Merges the training and the test sets to create one data set.
# Extracts only the measurements on the mean and standard deviation for each measurement.

fulldata <- cleandata()

# 3
# Uses descriptive activity names to name the activities in the data set
fulldata$activity <- factor(fulldata$activity, levels = c(1,2,3,4,5,6),
                            labels = c("walking", "walking upstairs",
                                       "walking downstairs", "sitting",
                                       "standing", "laying"))
```

```
# 4
# Appropriately labels the data set with descriptive variable names.

renamedata <- labeldescriptivename(fulldata)

# 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.

newdata <- renamedata %>%
                group_by(subject, activity) %>%
                        summarise_all(mean)

newdata
```

```
## # A tibble: 180 x 68
## # Groups:   subject [?]
##    subject activity TimeDomainBodyA~ TimeDomainBodyA~ TimeDomainBodyA~
##      <int> <fct>               <dbl>            <dbl>            <dbl>
##  1       1 walking             0.277         -0.0174           -0.111
##  2       1 walking~            0.255         -0.0240           -0.0973
##  3       1 walking~            0.289         -0.00992          -0.108
##  4       1 sitting             0.261         -0.00131          -0.105
##  5       1 standing            0.279         -0.0161           -0.111
##  6       1 laying              0.222         -0.0405           -0.113
##  7       2 walking             0.276         -0.0186           -0.106
##  8       2 walking~            0.247         -0.0214           -0.153
##  9       2 walking~            0.278         -0.0227           -0.117
## 10       2 sitting             0.277         -0.0157           -0.109
## # ... with 170 more rows, and 63 more variables:
## #   TimeDomainBodyAcceleration_StandardDeviation_X_Axis <dbl>,
## #   TimeDomainBodyAcceleration_StandardDeviation_Y_Axis <dbl>,
## #   TimeDomainBodyAcceleration_StandardDeviation_Z_Axis <dbl>,
## #   TimeDomainGravityAcceleration_Mean_X_Axis <dbl>,
## #   TimeDomainGravityAcceleration_Mean_Y_Axis <dbl>,
## #   TimeDomainGravityAcceleration_Mean_Z_Axis <dbl>,
## #   TimeDomainGravityAcceleration_StandardDeviation_X_Axis <dbl>,
## #   TimeDomainGravityAcceleration_StandardDeviation_Y_Axis <dbl>,
## #   TimeDomainGravityAcceleration_StandardDeviation_Z_Axis <dbl>,
## #   TimeDomainBodyAccelerationJerk_Mean_X_Axis <dbl>,
## #   TimeDomainBodyAccelerationJerk_Mean_Y_Axis <dbl>,
## #   TimeDomainBodyAccelerationJerk_Mean_Z_Axis <dbl>,
## #   TimeDomainBodyAccelerationJerk_StandardDeviation_X_Axis <dbl>,
## #   TimeDomainBodyAccelerationJerk_StandardDeviation_Y_Axis <dbl>,
## #   TimeDomainBodyAccelerationJerk_StandardDeviation_Z_Axis <dbl>,
## #   TimeDomainBodyGyroscope_Mean_X_Axis <dbl>,
## #   TimeDomainBodyGyroscope_Mean_Y_Axis <dbl>,
## #   TimeDomainBodyGyroscope_Mean_Z_Axis <dbl>,
## #   TimeDomainBodyGyroscope_StandardDeviation_X_Axis <dbl>,
## #   TimeDomainBodyGyroscope_StandardDeviation_Y_Axis <dbl>,
## #   TimeDomainBodyGyroscope_StandardDeviation_Z_Axis <dbl>,
## #   TimeDomainBodyGyroscopeJerk_Mean_X_Axis <dbl>,
## #   TimeDomainBodyGyroscopeJerk_Mean_Y_Axis <dbl>,
## #   TimeDomainBodyGyroscopeJerk_Mean_Z_Axis <dbl>,
## #   TimeDomainBodyGyroscopeJerk_StandardDeviation_X_Axis <dbl>,
## #   TimeDomainBodyGyroscopeJerk_StandardDeviation_Y_Axis <dbl>,
## #   TimeDomainBodyGyroscopeJerk_StandardDeviation_Z_Axis <dbl>,
## #   TimeDomainBodyAccelerationMagnitude_Mean <dbl>,
## #   TimeDomainBodyAccelerationMagnitude_StandardDeviation <dbl>,
## #   TimeDomainGravityAccelerationMagnitude_Mean <dbl>,
## #   TimeDomainGravityAccelerationMagnitude_StandardDeviation <dbl>,
## #   TimeDomainBodyAccelerationJerkMagnitude_Mean <dbl>,
## #   TimeDomainBodyAccelerationJerkMagnitude_StandardDeviation <dbl>,
## #   TimeDomainBodyGyroscopeMagnitude_Mean <dbl>,
## #   TimeDomainBodyGyroscopeMagnitude_StandardDeviation <dbl>,
## #   TimeDomainBodyGyroscopeJerkMagnitude_Mean <dbl>,
## #   TimeDomainBodyGyroscopeJerkMagnitude_StandardDeviation <dbl>,
## #   FrequencyDomainBodyAcceleration_Mean_X_Axis <dbl>,
## #   FrequencyDomainBodyAcceleration_Mean_Y_Axis <dbl>,
## #   FrequencyDomainBodyAcceleration_Mean_Z_Axis <dbl>,
## #   FrequencyDomainBodyAcceleration_StandardDeviation_X_Axis <dbl>,
## #   FrequencyDomainBodyAcceleration_StandardDeviation_Y_Axis <dbl>,
## #   FrequencyDomainBodyAcceleration_StandardDeviation_Z_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationJerk_Mean_X_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationJerk_Mean_Y_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationJerk_Mean_Z_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationJerk_StandardDeviation_X_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationJerk_StandardDeviation_Y_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationJerk_StandardDeviation_Z_Axis <dbl>,
## #   FrequencyDomainBodyGyroscope_Mean_X_Axis <dbl>,
## #   FrequencyDomainBodyGyroscope_Mean_Y_Axis <dbl>,
## #   FrequencyDomainBodyGyroscope_Mean_Z_Axis <dbl>,
## #   FrequencyDomainBodyGyroscope_StandardDeviation_X_Axis <dbl>,
## #   FrequencyDomainBodyGyroscope_StandardDeviation_Y_Axis <dbl>,
## #   FrequencyDomainBodyGyroscope_StandardDeviation_Z_Axis <dbl>,
## #   FrequencyDomainBodyAccelerationMagnitude_Mean <dbl>,
## #   FrequencyDomainBodyAccelerationMagnitude_StandardDeviation <dbl>,
## #   FrequencyDomainBodyAccelerationJerkMagnitude_Mean <dbl>,
## #   FrequencyDomainBodyAccelerationJerkMagnitude_StandardDeviation <dbl>,
## #   FrequencyDomainBodyGyroscopeMagnitude_Mean <dbl>,
## #   FrequencyDomainBodyGyroscopeMagnitude_StandardDeviation <dbl>,
## #   FrequencyDomainBodyGyroscopeJerkMagnitude_Mean <dbl>,
## #   FrequencyDomainBodyGyroscopeJerkMagnitude_StandardDeviation <dbl>
```