

# Peer Assignment Week4 Coursera

DJ Edwards

9/13/2020

## Context

This is a peer graded assignment for the John Hopkins University R Foundations Course on Coursera. It represents a demonstration of the R programming skills discussed focused on getting and cleaning data using **R**.

For this assignment I will:

- 1) create a tidy data set
- 2) create a Github repository to store/share my R scripts used for the analysis
- 3) create a code book (this markdown) and a README file that describes the variables, the data, and any transformations.

## Data

For this assignment we will use data from a 2012 experiment concerning wearable technology. Specifically, we will look at data collected from the accelerometers of the Samsung Galaxy S2 smartphone of 30 (19-48 yr old) humans during daily activities.

A full description is available at the site where the data was obtained: <http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones>

## Pull the Data

```
#create a filename for us to store the file/zip
filename<-"Smartphone_Data.zip"

#uses the course suggested if statement to see if you have the zip'file. Load if you don't/curl for http
if(!file.exists(filename)){
  fileURL<-"https://d396qusza40orc.cloudfront.net/getdata%2Fprojectfiles%2FUCI%20HAR%20Dataset.zip"
  download.file(fileURL,filename, method="curl")
}

#unzips the files from the UCI HAR Dataset folder into my Smartphone folder
if(!file.exists("UCI HAR Dataset")){
  unzip(filename)
}
```

## Wrangle the labels and ids.

The README.txt file inside of the UCI HAR Dataset provides the information for how the user, activity and features information relates to the test and training set data. We will have to bring in the labels, features and user ids and bind this information with the test and training set information to make a tidy data frame.

```
#read in all of the txt files associated with the features and their ids
featureName<-read.table("UCI HAR Dataset/features.txt", col.names=c("featureId","featureLabel"))
#these are the recorded variables

#read in the activity labels and their ids
activityLabel<-read.table("UCI HAR Dataset/activity_labels.txt", col.names=c("activityId","activityLabel"))

#read in the user id numbers for the training and test set information
userId_Test<-read.table("UCI HAR Dataset/test/subject_test.txt", col.names=c("userId"))
userId_Train<-read.table("UCI HAR Dataset/train/subject_train.txt", col.names=c("userId"))

#read in the test set data for the experiment (X is the feature, y is the activity)
feature_Test<-read.table("UCI HAR Dataset/test/X_test.txt", col.names=featureName$featureLabel)
activity_Test<-read.table("UCI HAR Dataset/test/y_test.txt", col.names=c("activityId"))

#read in the training set data for the experiment
feature_Train<-read.table("UCI HAR Dataset/train/X_train.txt", col.names=featureName$featureLabel)
activity_Train<-read.table("UCI HAR Dataset/train/y_train.txt", col.names=c("activityId"))
```

## Construct the smartphone data frame

This takes the 10299 rows and adds the user information and the model prediction (activity-(X) based on the recorded variable( feature-(y)). We should expect 10299 rows and 563 columns.

```
feature_df<-rbind(feature_Train,feature_Test)
activity_df<-rbind(activity_Train,activity_Test)
user_df<-rbind(userId_Train,userId_Test)
Smartphone_df<-cbind(user_df,feature_df,activity_df)

#check dimensions of the data frame
dim(Smartphone_df)
head(Smartphone_df)
```

## Make the names more user friendly

This section provides more readable variable names based on the experiment

```
names(Smartphone_df)<-gsub("^t","Time",names(Smartphone_df))
names(Smartphone_df)<-gsub("^f","Frequency",names(Smartphone_df))
names(Smartphone_df)<-gsub("-freq()", "Frequency", names(Smartphone_df))
names(Smartphone_df)<-gsub("-mean()", "Mean",names(Smartphone_df))
names(Smartphone_df)<-gsub("-std()", "Standard Deviation", names(Smartphone_df))
names(Smartphone_df)<-gsub("BodyBody", "Body", names(Smartphone_df))
names(Smartphone_df)<-gsub("tBody","TimeBody",names(Smartphone_df))
names(Smartphone_df)<-gsub("Acc", "Acelerometer", names(Smartphone_df))
names(Smartphone_df)<-gsub("Gyro", "Gyroscope", names(Smartphone_df))
names(Smartphone_df)<-gsub("Mag", "Magnitude", names(Smartphone_df))
```

## Subset the data frame for only mean and standard deviation variable and add activity name

This uses dplyr to subset the Smartphone experiment into only those variables that involve the mean or std. It also replaces activityId with the activity name (label) and then renames the column

```
library("dplyr")
Smartphone_subset<-Smartphone_df %>% select(userId,contains("mean"),contains("std"),activityId) %>%
mutate(activityId=activityLabel[activityId,2]) %>%
rename(activityLabel=activityId)
```

## Makes a tidy data frame for the averages of the experiment subset by userId and activity

```
Tidy_Result<-Smartphone_subset %>%
  group_by(userId,activityLabel) %>%
  summarise(across(where(is.numeric),~mean(.x,na.rm=TRUE)))

#output and show structure of result
Tidy_Result
str(Tidy_Result)
```

## Output the Tidy\_Result

```
#write the output to the required txt file
write.table(Tidy_Result,file="./Tidy_Result.txt",row.names=FALSE, col.names = TRUE)
```