Data Science

Codebook(Mid I)

Muhammad Hamza

16F-8345

Process:

1. **Download the dataset:**

   Dataset downloaded and extracted under the folder called **'forDataScienceMid'**

2. **Assigning values to variables**

   X_test <- read.table("X_test.txt") : *2947 rows, 561 columns*

   contains recorded features test data

   Y_test <- read.table("Y_test.txt"): 2947 rows, 1 columns

   contains test data of activities code labels

   X_train <- read.table("X_train.txt") : 7352 rows, 561 columns

   contains recorded features train data

   Y_train <- read.table("Y_train.txt"):7352 rows, 1 columns

   contains train data of activities code labels

   features <- read.table("features.txt") : 561 rows, 2 columns

   activity <- read.table("activity_labels.txt") : 6 rows, 2 columns

   test_sub <- read.table("subject_test.txt") :2947 rows, 1 column

   contains test data of 9/30 volunteer test subjects being observed

   train_sub <- read.table("subject_train.txt"):7352 rows, 1 column

   contains train data of 21/30 volunteer subjects being observed

3. Merging training and test sets to create one dataset

```
Subjects <- rbind(test_sub,train_sub)

y <- rbind(Y_train,Y_test)

x <- rbind(X_train,X_test)

final <- cbind(Subjects,y)

final <- cbind(final,x)
```

- **x** (10299 rows, 561 columns) is created by merging X_train and X_test using rbind() function
- **y** (10299 rows, 1 column) is created by merging Y_train and Y_test using rbind() function
- **Subjects** (10299 rows, 1 column) is created by merging subject_train and subject_test using rbind() function
- **final** (10299 rows, 563 column) is created by merging Subject, Y and X using cbind() function

4. Extracts only the measurements on the mean and standard deviation for each measurement

```
indexes <- grep("mean|std",features[,2])
indexes <- indexes+2
indexes <- c(1,2,indexes)
# remove unnecessary columns
f_dataset <- final[,indexes]
indexes <- grep("mean|std", features[,2])
feature_titles <- features[,2]
features_titles <- feature_titles[indexes]
features_titles <- lapply(features_titles, as.character)
features_titles <- c("Subjects", "Activity", features_titles)
colnames(f_dataset) <- features_titles
```

f_dataset (10299 rows, 88 columns) is created by subsetting final data,

selecting only columns: subject, code and the measurements on the mean and standard deviation (std) for each measurement.

5. **Uses descriptive activity names to name the activities in the data set**

```
activity_labels <- read.table("activity_labels.txt")
activity_titles <- activity_labels[,2]
activities_titles <- lapply(activity_titles, as.character)
f_dataset <- within(f_dataset, Activity <- factor(Activity, labels = activities_titles))
F_dataset
```

F_dataset contains the descriptive activity name of all activities such as standing,walking etc

6. **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**

```
melting <- melt(f_dataset, id=c("Subjects","Activity"))
tidying <- dcast(melting, Subjects+Activity ~ variable, mean)
Tidying
```

7. **Write this into a csv file**

```
write.csv(tidying, "Data_Science.csv", row.names=FALSE)
```

Data_Science.csv contains all the data of tidy f_dataset.