Yuji\_Hayashi\_R

Yuji Hayashi

2020/5/21

Table of Contents

[Import libraries and Datasets 2](#_Toc41640898)

[COURSERA FINAL PROJECT 2](#_Toc41640899)

[The objective of Coursera Final Assignment 2](#_Toc41640900)

[Check the data 3](#_Toc41640901)

[Preprocessing 49](#_Toc41640902)

[PCA 64](#_Toc41640903)

[Split the data 65](#_Toc41640904)

[Base model 67](#_Toc41640905)

[Predict 69](#_Toc41640906)

[Accuracy measures 69](#_Toc41640907)

[Features Importance 69](#_Toc41640908)

[Predict the given test data 70](#_Toc41640909)

[Financial Application 71](#_Toc41640910)

[The objective of this experiment 71](#_Toc41640911)

[Check the data 71](#_Toc41640912)

[Preprocessing 73](#_Toc41640913)

[Split the data 75](#_Toc41640914)

[Base model 77](#_Toc41640915)

[Predict 78](#_Toc41640916)

[Accuracy measures 78](#_Toc41640917)

[Features Importance 79](#_Toc41640918)

[Predict the new client’s charges 80](#_Toc41640919)

# Import libraries and Datasets

At first, let’s install and import packages needed in my analysis!

#install.packages("caret")  
#install.packages("kernlab")  
#install.packages("e1071")  
#install.packages("ISLR")  
#install.packages("ggplot2")  
#install.packages("PreProcess")  
#install.packages(c("data.table", "dplyr"))  
#install.packages("dplyr")  
#install.packages("randomForest", quite=TRUE)  
#install.packages("xgboost")  
#install.packages("gbm")  
#install.packages("Matrix")  
  
library("caret"); library("kernlab"); library("e1071");  
library("ISLR"); library("ggplot2"); library("PreProcess");  
library("dplyr"); library("readr");library("data.table");  
library(randomForest); library(xgboost); library(Matrix); library(miscTools)

Set working directory and import the assignment datasets I use here.

setwd("C:/Users/Yuzi/OneDrive/M1\_EDHEC/Financial Econometrics/Final Assignment/EDHEC\_FinalAssignment")  
  
# COURSERA project  
df <- read\_csv("pml-training.csv")  
df\_test <- read\_csv("pml-testing.csv")  
  
# Financial Application  
data <- read\_csv("Yuji\_Hayashi\_data\_file.csv")

# COURSERA FINAL PROJECT

## The objective of Coursera Final Assignment

The objective is to predict the response variable classe in the case of pml-testing.csv by building a machine learning model which is made with the training data pml-training.csv. Finally, I will make predictions of classe for 20 users in the case given as Final Assignment of Coursera. For the details, I chose Random Forest model to predict it because the performance is nicer than other models in terms of accuracy for this dataset.

## Check the data

Check the shapes of each dataset.

dim(df)

## [1] 19622 160

dim(df\_test)

## [1] 20 160

The train dataset is composed of about 20,000 data points, 160 variables. On the other hand, the test dataset is composed of 20 data points, 160 variables.

See the above and below 5 rows of each dataset.

head(df, 5)

## # A tibble: 5 x 160  
## id user\_name raw\_timestamp\_p~ raw\_timestamp\_p~ cvtd\_timestamp new\_window  
## <dbl> <chr> <dbl> <dbl> <chr> <chr>   
## 1 1 carlitos 1323084231 788290 05/12/2011 11~ no   
## 2 2 carlitos 1323084231 808298 05/12/2011 11~ no   
## 3 3 carlitos 1323084231 820366 05/12/2011 11~ no   
## 4 4 carlitos 1323084232 120339 05/12/2011 11~ no   
## 5 5 carlitos 1323084232 196328 05/12/2011 11~ no   
## # ... with 154 more variables: num\_window <dbl>, roll\_belt <dbl>,  
## # pitch\_belt <dbl>, yaw\_belt <dbl>, total\_accel\_belt <dbl>,  
## # kurtosis\_roll\_belt <chr>, kurtosis\_picth\_belt <chr>,  
## # kurtosis\_yaw\_belt <chr>, skewness\_roll\_belt <chr>,  
## # skewness\_roll\_belt.1 <chr>, skewness\_yaw\_belt <chr>, max\_roll\_belt <dbl>,  
## # max\_picth\_belt <dbl>, max\_yaw\_belt <chr>, min\_roll\_belt <dbl>,  
## # min\_pitch\_belt <dbl>, min\_yaw\_belt <chr>, amplitude\_roll\_belt <dbl>,  
## # amplitude\_pitch\_belt <dbl>, amplitude\_yaw\_belt <chr>,  
## # var\_total\_accel\_belt <dbl>, avg\_roll\_belt <dbl>, stddev\_roll\_belt <dbl>,  
## # var\_roll\_belt <dbl>, avg\_pitch\_belt <dbl>, stddev\_pitch\_belt <dbl>,  
## # var\_pitch\_belt <dbl>, avg\_yaw\_belt <dbl>, stddev\_yaw\_belt <dbl>,  
## # var\_yaw\_belt <dbl>, gyros\_belt\_x <dbl>, gyros\_belt\_y <dbl>,  
## # gyros\_belt\_z <dbl>, accel\_belt\_x <dbl>, accel\_belt\_y <dbl>,  
## # accel\_belt\_z <dbl>, magnet\_belt\_x <dbl>, magnet\_belt\_y <dbl>,  
## # magnet\_belt\_z <dbl>, roll\_arm <dbl>, pitch\_arm <dbl>, yaw\_arm <dbl>,  
## # total\_accel\_arm <dbl>, var\_accel\_arm <dbl>, avg\_roll\_arm <dbl>,  
## # stddev\_roll\_arm <dbl>, var\_roll\_arm <dbl>, avg\_pitch\_arm <dbl>,  
## # stddev\_pitch\_arm <dbl>, var\_pitch\_arm <dbl>, avg\_yaw\_arm <dbl>,  
## # stddev\_yaw\_arm <dbl>, var\_yaw\_arm <dbl>, gyros\_arm\_x <dbl>,  
## # gyros\_arm\_y <dbl>, gyros\_arm\_z <dbl>, accel\_arm\_x <dbl>, accel\_arm\_y <dbl>,  
## # accel\_arm\_z <dbl>, magnet\_arm\_x <dbl>, magnet\_arm\_y <dbl>,  
## # magnet\_arm\_z <dbl>, kurtosis\_roll\_arm <dbl>, kurtosis\_picth\_arm <chr>,  
## # kurtosis\_yaw\_arm <chr>, skewness\_roll\_arm <dbl>, skewness\_pitch\_arm <chr>,  
## # skewness\_yaw\_arm <chr>, max\_roll\_arm <dbl>, max\_picth\_arm <dbl>,  
## # max\_yaw\_arm <dbl>, min\_roll\_arm <dbl>, min\_pitch\_arm <dbl>,  
## # min\_yaw\_arm <dbl>, amplitude\_roll\_arm <dbl>, amplitude\_pitch\_arm <dbl>,  
## # amplitude\_yaw\_arm <dbl>, roll\_dumbbell <dbl>, pitch\_dumbbell <dbl>,  
## # yaw\_dumbbell <dbl>, kurtosis\_roll\_dumbbell <dbl>,  
## # kurtosis\_picth\_dumbbell <dbl>, kurtosis\_yaw\_dumbbell <chr>,  
## # skewness\_roll\_dumbbell <dbl>, skewness\_pitch\_dumbbell <dbl>,  
## # skewness\_yaw\_dumbbell <chr>, max\_roll\_dumbbell <dbl>,  
## # max\_picth\_dumbbell <dbl>, max\_yaw\_dumbbell <dbl>, min\_roll\_dumbbell <dbl>,  
## # min\_pitch\_dumbbell <dbl>, min\_yaw\_dumbbell <dbl>,  
## # amplitude\_roll\_dumbbell <dbl>, amplitude\_pitch\_dumbbell <dbl>,  
## # amplitude\_yaw\_dumbbell <dbl>, total\_accel\_dumbbell <dbl>,  
## # var\_accel\_dumbbell <dbl>, avg\_roll\_dumbbell <dbl>,  
## # stddev\_roll\_dumbbell <dbl>, var\_roll\_dumbbell <dbl>, ...

tail(df, 5)

## # A tibble: 5 x 160  
## id user\_name raw\_timestamp\_p~ raw\_timestamp\_p~ cvtd\_timestamp new\_window  
## <dbl> <chr> <dbl> <dbl> <chr> <chr>   
## 1 19618 adelmo 1322832937 588376 02/12/2011 13~ no   
## 2 19619 adelmo 1322832937 596287 02/12/2011 13~ no   
## 3 19620 adelmo 1322832937 636283 02/12/2011 13~ no   
## 4 19621 adelmo 1322832937 964299 02/12/2011 13~ no   
## 5 19622 adelmo 1322832937 972293 02/12/2011 13~ yes   
## # ... with 154 more variables: num\_window <dbl>, roll\_belt <dbl>,  
## # pitch\_belt <dbl>, yaw\_belt <dbl>, total\_accel\_belt <dbl>,  
## # kurtosis\_roll\_belt <chr>, kurtosis\_picth\_belt <chr>,  
## # kurtosis\_yaw\_belt <chr>, skewness\_roll\_belt <chr>,  
## # skewness\_roll\_belt.1 <chr>, skewness\_yaw\_belt <chr>, max\_roll\_belt <dbl>,  
## # max\_picth\_belt <dbl>, max\_yaw\_belt <chr>, min\_roll\_belt <dbl>,  
## # min\_pitch\_belt <dbl>, min\_yaw\_belt <chr>, amplitude\_roll\_belt <dbl>,  
## # amplitude\_pitch\_belt <dbl>, amplitude\_yaw\_belt <chr>,  
## # var\_total\_accel\_belt <dbl>, avg\_roll\_belt <dbl>, stddev\_roll\_belt <dbl>,  
## # var\_roll\_belt <dbl>, avg\_pitch\_belt <dbl>, stddev\_pitch\_belt <dbl>,  
## # var\_pitch\_belt <dbl>, avg\_yaw\_belt <dbl>, stddev\_yaw\_belt <dbl>,  
## # var\_yaw\_belt <dbl>, gyros\_belt\_x <dbl>, gyros\_belt\_y <dbl>,  
## # gyros\_belt\_z <dbl>, accel\_belt\_x <dbl>, accel\_belt\_y <dbl>,  
## # accel\_belt\_z <dbl>, magnet\_belt\_x <dbl>, magnet\_belt\_y <dbl>,  
## # magnet\_belt\_z <dbl>, roll\_arm <dbl>, pitch\_arm <dbl>, yaw\_arm <dbl>,  
## # total\_accel\_arm <dbl>, var\_accel\_arm <dbl>, avg\_roll\_arm <dbl>,  
## # stddev\_roll\_arm <dbl>, var\_roll\_arm <dbl>, avg\_pitch\_arm <dbl>,  
## # stddev\_pitch\_arm <dbl>, var\_pitch\_arm <dbl>, avg\_yaw\_arm <dbl>,  
## # stddev\_yaw\_arm <dbl>, var\_yaw\_arm <dbl>, gyros\_arm\_x <dbl>,  
## # gyros\_arm\_y <dbl>, gyros\_arm\_z <dbl>, accel\_arm\_x <dbl>, accel\_arm\_y <dbl>,  
## # accel\_arm\_z <dbl>, magnet\_arm\_x <dbl>, magnet\_arm\_y <dbl>,  
## # magnet\_arm\_z <dbl>, kurtosis\_roll\_arm <dbl>, kurtosis\_picth\_arm <chr>,  
## # kurtosis\_yaw\_arm <chr>, skewness\_roll\_arm <dbl>, skewness\_pitch\_arm <chr>,  
## # skewness\_yaw\_arm <chr>, max\_roll\_arm <dbl>, max\_picth\_arm <dbl>,  
## # max\_yaw\_arm <dbl>, min\_roll\_arm <dbl>, min\_pitch\_arm <dbl>,  
## # min\_yaw\_arm <dbl>, amplitude\_roll\_arm <dbl>, amplitude\_pitch\_arm <dbl>,  
## # amplitude\_yaw\_arm <dbl>, roll\_dumbbell <dbl>, pitch\_dumbbell <dbl>,  
## # yaw\_dumbbell <dbl>, kurtosis\_roll\_dumbbell <dbl>,  
## # kurtosis\_picth\_dumbbell <dbl>, kurtosis\_yaw\_dumbbell <chr>,  
## # skewness\_roll\_dumbbell <dbl>, skewness\_pitch\_dumbbell <dbl>,  
## # skewness\_yaw\_dumbbell <chr>, max\_roll\_dumbbell <dbl>,  
## # max\_picth\_dumbbell <dbl>, max\_yaw\_dumbbell <dbl>, min\_roll\_dumbbell <dbl>,  
## # min\_pitch\_dumbbell <dbl>, min\_yaw\_dumbbell <dbl>,  
## # amplitude\_roll\_dumbbell <dbl>, amplitude\_pitch\_dumbbell <dbl>,  
## # amplitude\_yaw\_dumbbell <dbl>, total\_accel\_dumbbell <dbl>,  
## # var\_accel\_dumbbell <dbl>, avg\_roll\_dumbbell <dbl>,  
## # stddev\_roll\_dumbbell <dbl>, var\_roll\_dumbbell <dbl>, ...

head(df\_test, 5)

## # A tibble: 5 x 160  
## id user\_name raw\_timestamp\_p~ raw\_timestamp\_p~ cvtd\_timestamp new\_window  
## <dbl> <chr> <dbl> <dbl> <chr> <chr>   
## 1 1 pedro 1323095002 868349 05/12/2011 14~ no   
## 2 2 jeremy 1322673067 778725 30/11/2011 17~ no   
## 3 3 jeremy 1322673075 342967 30/11/2011 17~ no   
## 4 4 adelmo 1322832789 560311 02/12/2011 13~ no   
## 5 5 eurico 1322489635 814776 28/11/2011 14~ no   
## # ... with 154 more variables: num\_window <dbl>, roll\_belt <dbl>,  
## # pitch\_belt <dbl>, yaw\_belt <dbl>, total\_accel\_belt <dbl>,  
## # kurtosis\_roll\_belt <lgl>, kurtosis\_picth\_belt <lgl>,  
## # kurtosis\_yaw\_belt <lgl>, skewness\_roll\_belt <lgl>,  
## # skewness\_roll\_belt.1 <lgl>, skewness\_yaw\_belt <lgl>, max\_roll\_belt <lgl>,  
## # max\_picth\_belt <lgl>, max\_yaw\_belt <lgl>, min\_roll\_belt <lgl>,  
## # min\_pitch\_belt <lgl>, min\_yaw\_belt <lgl>, amplitude\_roll\_belt <lgl>,  
## # amplitude\_pitch\_belt <lgl>, amplitude\_yaw\_belt <lgl>,  
## # var\_total\_accel\_belt <lgl>, avg\_roll\_belt <lgl>, stddev\_roll\_belt <lgl>,  
## # var\_roll\_belt <lgl>, avg\_pitch\_belt <lgl>, stddev\_pitch\_belt <lgl>,  
## # var\_pitch\_belt <lgl>, avg\_yaw\_belt <lgl>, stddev\_yaw\_belt <lgl>,  
## # var\_yaw\_belt <lgl>, gyros\_belt\_x <dbl>, gyros\_belt\_y <dbl>,  
## # gyros\_belt\_z <dbl>, accel\_belt\_x <dbl>, accel\_belt\_y <dbl>,  
## # accel\_belt\_z <dbl>, magnet\_belt\_x <dbl>, magnet\_belt\_y <dbl>,  
## # magnet\_belt\_z <dbl>, roll\_arm <dbl>, pitch\_arm <dbl>, yaw\_arm <dbl>,  
## # total\_accel\_arm <dbl>, var\_accel\_arm <lgl>, avg\_roll\_arm <lgl>,  
## # stddev\_roll\_arm <lgl>, var\_roll\_arm <lgl>, avg\_pitch\_arm <lgl>,  
## # stddev\_pitch\_arm <lgl>, var\_pitch\_arm <lgl>, avg\_yaw\_arm <lgl>,  
## # stddev\_yaw\_arm <lgl>, var\_yaw\_arm <lgl>, gyros\_arm\_x <dbl>,  
## # gyros\_arm\_y <dbl>, gyros\_arm\_z <dbl>, accel\_arm\_x <dbl>, accel\_arm\_y <dbl>,  
## # accel\_arm\_z <dbl>, magnet\_arm\_x <dbl>, magnet\_arm\_y <dbl>,  
## # magnet\_arm\_z <dbl>, kurtosis\_roll\_arm <lgl>, kurtosis\_picth\_arm <lgl>,  
## # kurtosis\_yaw\_arm <lgl>, skewness\_roll\_arm <lgl>, skewness\_pitch\_arm <lgl>,  
## # skewness\_yaw\_arm <lgl>, max\_roll\_arm <lgl>, max\_picth\_arm <lgl>,  
## # max\_yaw\_arm <lgl>, min\_roll\_arm <lgl>, min\_pitch\_arm <lgl>,  
## # min\_yaw\_arm <lgl>, amplitude\_roll\_arm <lgl>, amplitude\_pitch\_arm <lgl>,  
## # amplitude\_yaw\_arm <lgl>, roll\_dumbbell <dbl>, pitch\_dumbbell <dbl>,  
## # yaw\_dumbbell <dbl>, kurtosis\_roll\_dumbbell <lgl>,  
## # kurtosis\_picth\_dumbbell <lgl>, kurtosis\_yaw\_dumbbell <lgl>,  
## # skewness\_roll\_dumbbell <lgl>, skewness\_pitch\_dumbbell <lgl>,  
## # skewness\_yaw\_dumbbell <lgl>, max\_roll\_dumbbell <lgl>,  
## # max\_picth\_dumbbell <lgl>, max\_yaw\_dumbbell <lgl>, min\_roll\_dumbbell <lgl>,  
## # min\_pitch\_dumbbell <lgl>, min\_yaw\_dumbbell <lgl>,  
## # amplitude\_roll\_dumbbell <lgl>, amplitude\_pitch\_dumbbell <lgl>,  
## # amplitude\_yaw\_dumbbell <lgl>, total\_accel\_dumbbell <dbl>,  
## # var\_accel\_dumbbell <lgl>, avg\_roll\_dumbbell <lgl>,  
## # stddev\_roll\_dumbbell <lgl>, var\_roll\_dumbbell <lgl>, ...

tail(df\_test, 5)

## # A tibble: 5 x 160  
## id user\_name raw\_timestamp\_p~ raw\_timestamp\_p~ cvtd\_timestamp new\_window  
## <dbl> <chr> <dbl> <dbl> <chr> <chr>   
## 1 16 eurico 1322489713 706637 28/11/2011 14~ no   
## 2 17 pedro 1323094971 920315 05/12/2011 14~ no   
## 3 18 carlitos 1323084285 176314 05/12/2011 11~ no   
## 4 19 pedro 1323094999 828379 05/12/2011 14~ no   
## 5 20 eurico 1322489658 106658 28/11/2011 14~ no   
## # ... with 154 more variables: num\_window <dbl>, roll\_belt <dbl>,  
## # pitch\_belt <dbl>, yaw\_belt <dbl>, total\_accel\_belt <dbl>,  
## # kurtosis\_roll\_belt <lgl>, kurtosis\_picth\_belt <lgl>,  
## # kurtosis\_yaw\_belt <lgl>, skewness\_roll\_belt <lgl>,  
## # skewness\_roll\_belt.1 <lgl>, skewness\_yaw\_belt <lgl>, max\_roll\_belt <lgl>,  
## # max\_picth\_belt <lgl>, max\_yaw\_belt <lgl>, min\_roll\_belt <lgl>,  
## # min\_pitch\_belt <lgl>, min\_yaw\_belt <lgl>, amplitude\_roll\_belt <lgl>,  
## # amplitude\_pitch\_belt <lgl>, amplitude\_yaw\_belt <lgl>,  
## # var\_total\_accel\_belt <lgl>, avg\_roll\_belt <lgl>, stddev\_roll\_belt <lgl>,  
## # var\_roll\_belt <lgl>, avg\_pitch\_belt <lgl>, stddev\_pitch\_belt <lgl>,  
## # var\_pitch\_belt <lgl>, avg\_yaw\_belt <lgl>, stddev\_yaw\_belt <lgl>,  
## # var\_yaw\_belt <lgl>, gyros\_belt\_x <dbl>, gyros\_belt\_y <dbl>,  
## # gyros\_belt\_z <dbl>, accel\_belt\_x <dbl>, accel\_belt\_y <dbl>,  
## # accel\_belt\_z <dbl>, magnet\_belt\_x <dbl>, magnet\_belt\_y <dbl>,  
## # magnet\_belt\_z <dbl>, roll\_arm <dbl>, pitch\_arm <dbl>, yaw\_arm <dbl>,  
## # total\_accel\_arm <dbl>, var\_accel\_arm <lgl>, avg\_roll\_arm <lgl>,  
## # stddev\_roll\_arm <lgl>, var\_roll\_arm <lgl>, avg\_pitch\_arm <lgl>,  
## # stddev\_pitch\_arm <lgl>, var\_pitch\_arm <lgl>, avg\_yaw\_arm <lgl>,  
## # stddev\_yaw\_arm <lgl>, var\_yaw\_arm <lgl>, gyros\_arm\_x <dbl>,  
## # gyros\_arm\_y <dbl>, gyros\_arm\_z <dbl>, accel\_arm\_x <dbl>, accel\_arm\_y <dbl>,  
## # accel\_arm\_z <dbl>, magnet\_arm\_x <dbl>, magnet\_arm\_y <dbl>,  
## # magnet\_arm\_z <dbl>, kurtosis\_roll\_arm <lgl>, kurtosis\_picth\_arm <lgl>,  
## # kurtosis\_yaw\_arm <lgl>, skewness\_roll\_arm <lgl>, skewness\_pitch\_arm <lgl>,  
## # skewness\_yaw\_arm <lgl>, max\_roll\_arm <lgl>, max\_picth\_arm <lgl>,  
## # max\_yaw\_arm <lgl>, min\_roll\_arm <lgl>, min\_pitch\_arm <lgl>,  
## # min\_yaw\_arm <lgl>, amplitude\_roll\_arm <lgl>, amplitude\_pitch\_arm <lgl>,  
## # amplitude\_yaw\_arm <lgl>, roll\_dumbbell <dbl>, pitch\_dumbbell <dbl>,  
## # yaw\_dumbbell <dbl>, kurtosis\_roll\_dumbbell <lgl>,  
## # kurtosis\_picth\_dumbbell <lgl>, kurtosis\_yaw\_dumbbell <lgl>,  
## # skewness\_roll\_dumbbell <lgl>, skewness\_pitch\_dumbbell <lgl>,  
## # skewness\_yaw\_dumbbell <lgl>, max\_roll\_dumbbell <lgl>,  
## # max\_picth\_dumbbell <lgl>, max\_yaw\_dumbbell <lgl>, min\_roll\_dumbbell <lgl>,  
## # min\_pitch\_dumbbell <lgl>, min\_yaw\_dumbbell <lgl>,  
## # amplitude\_roll\_dumbbell <lgl>, amplitude\_pitch\_dumbbell <lgl>,  
## # amplitude\_yaw\_dumbbell <lgl>, total\_accel\_dumbbell <dbl>,  
## # var\_accel\_dumbbell <lgl>, avg\_roll\_dumbbell <lgl>,  
## # stddev\_roll\_dumbbell <lgl>, var\_roll\_dumbbell <lgl>, ...

See the column names.

colnames(df)

## [1] "id" "user\_name"   
## [3] "raw\_timestamp\_part\_1" "raw\_timestamp\_part\_2"   
## [5] "cvtd\_timestamp" "new\_window"   
## [7] "num\_window" "roll\_belt"   
## [9] "pitch\_belt" "yaw\_belt"   
## [11] "total\_accel\_belt" "kurtosis\_roll\_belt"   
## [13] "kurtosis\_picth\_belt" "kurtosis\_yaw\_belt"   
## [15] "skewness\_roll\_belt" "skewness\_roll\_belt.1"   
## [17] "skewness\_yaw\_belt" "max\_roll\_belt"   
## [19] "max\_picth\_belt" "max\_yaw\_belt"   
## [21] "min\_roll\_belt" "min\_pitch\_belt"   
## [23] "min\_yaw\_belt" "amplitude\_roll\_belt"   
## [25] "amplitude\_pitch\_belt" "amplitude\_yaw\_belt"   
## [27] "var\_total\_accel\_belt" "avg\_roll\_belt"   
## [29] "stddev\_roll\_belt" "var\_roll\_belt"   
## [31] "avg\_pitch\_belt" "stddev\_pitch\_belt"   
## [33] "var\_pitch\_belt" "avg\_yaw\_belt"   
## [35] "stddev\_yaw\_belt" "var\_yaw\_belt"   
## [37] "gyros\_belt\_x" "gyros\_belt\_y"   
## [39] "gyros\_belt\_z" "accel\_belt\_x"   
## [41] "accel\_belt\_y" "accel\_belt\_z"   
## [43] "magnet\_belt\_x" "magnet\_belt\_y"   
## [45] "magnet\_belt\_z" "roll\_arm"   
## [47] "pitch\_arm" "yaw\_arm"   
## [49] "total\_accel\_arm" "var\_accel\_arm"   
## [51] "avg\_roll\_arm" "stddev\_roll\_arm"   
## [53] "var\_roll\_arm" "avg\_pitch\_arm"   
## [55] "stddev\_pitch\_arm" "var\_pitch\_arm"   
## [57] "avg\_yaw\_arm" "stddev\_yaw\_arm"   
## [59] "var\_yaw\_arm" "gyros\_arm\_x"   
## [61] "gyros\_arm\_y" "gyros\_arm\_z"   
## [63] "accel\_arm\_x" "accel\_arm\_y"   
## [65] "accel\_arm\_z" "magnet\_arm\_x"   
## [67] "magnet\_arm\_y" "magnet\_arm\_z"   
## [69] "kurtosis\_roll\_arm" "kurtosis\_picth\_arm"   
## [71] "kurtosis\_yaw\_arm" "skewness\_roll\_arm"   
## [73] "skewness\_pitch\_arm" "skewness\_yaw\_arm"   
## [75] "max\_roll\_arm" "max\_picth\_arm"   
## [77] "max\_yaw\_arm" "min\_roll\_arm"   
## [79] "min\_pitch\_arm" "min\_yaw\_arm"   
## [81] "amplitude\_roll\_arm" "amplitude\_pitch\_arm"   
## [83] "amplitude\_yaw\_arm" "roll\_dumbbell"   
## [85] "pitch\_dumbbell" "yaw\_dumbbell"   
## [87] "kurtosis\_roll\_dumbbell" "kurtosis\_picth\_dumbbell"   
## [89] "kurtosis\_yaw\_dumbbell" "skewness\_roll\_dumbbell"   
## [91] "skewness\_pitch\_dumbbell" "skewness\_yaw\_dumbbell"   
## [93] "max\_roll\_dumbbell" "max\_picth\_dumbbell"   
## [95] "max\_yaw\_dumbbell" "min\_roll\_dumbbell"   
## [97] "min\_pitch\_dumbbell" "min\_yaw\_dumbbell"   
## [99] "amplitude\_roll\_dumbbell" "amplitude\_pitch\_dumbbell"  
## [101] "amplitude\_yaw\_dumbbell" "total\_accel\_dumbbell"   
## [103] "var\_accel\_dumbbell" "avg\_roll\_dumbbell"   
## [105] "stddev\_roll\_dumbbell" "var\_roll\_dumbbell"   
## [107] "avg\_pitch\_dumbbell" "stddev\_pitch\_dumbbell"   
## [109] "var\_pitch\_dumbbell" "avg\_yaw\_dumbbell"   
## [111] "stddev\_yaw\_dumbbell" "var\_yaw\_dumbbell"   
## [113] "gyros\_dumbbell\_x" "gyros\_dumbbell\_y"   
## [115] "gyros\_dumbbell\_z" "accel\_dumbbell\_x"   
## [117] "accel\_dumbbell\_y" "accel\_dumbbell\_z"   
## [119] "magnet\_dumbbell\_x" "magnet\_dumbbell\_y"   
## [121] "magnet\_dumbbell\_z" "roll\_forearm"   
## [123] "pitch\_forearm" "yaw\_forearm"   
## [125] "kurtosis\_roll\_forearm" "kurtosis\_picth\_forearm"   
## [127] "kurtosis\_yaw\_forearm" "skewness\_roll\_forearm"   
## [129] "skewness\_pitch\_forearm" "skewness\_yaw\_forearm"   
## [131] "max\_roll\_forearm" "max\_picth\_forearm"   
## [133] "max\_yaw\_forearm" "min\_roll\_forearm"   
## [135] "min\_pitch\_forearm" "min\_yaw\_forearm"   
## [137] "amplitude\_roll\_forearm" "amplitude\_pitch\_forearm"   
## [139] "amplitude\_yaw\_forearm" "total\_accel\_forearm"   
## [141] "var\_accel\_forearm" "avg\_roll\_forearm"   
## [143] "stddev\_roll\_forearm" "var\_roll\_forearm"   
## [145] "avg\_pitch\_forearm" "stddev\_pitch\_forearm"   
## [147] "var\_pitch\_forearm" "avg\_yaw\_forearm"   
## [149] "stddev\_yaw\_forearm" "var\_yaw\_forearm"   
## [151] "gyros\_forearm\_x" "gyros\_forearm\_y"   
## [153] "gyros\_forearm\_z" "accel\_forearm\_x"   
## [155] "accel\_forearm\_y" "accel\_forearm\_z"   
## [157] "magnet\_forearm\_x" "magnet\_forearm\_y"   
## [159] "magnet\_forearm\_z" "classe"

colnames(df\_test)

## [1] "id" "user\_name"   
## [3] "raw\_timestamp\_part\_1" "raw\_timestamp\_part\_2"   
## [5] "cvtd\_timestamp" "new\_window"   
## [7] "num\_window" "roll\_belt"   
## [9] "pitch\_belt" "yaw\_belt"   
## [11] "total\_accel\_belt" "kurtosis\_roll\_belt"   
## [13] "kurtosis\_picth\_belt" "kurtosis\_yaw\_belt"   
## [15] "skewness\_roll\_belt" "skewness\_roll\_belt.1"   
## [17] "skewness\_yaw\_belt" "max\_roll\_belt"   
## [19] "max\_picth\_belt" "max\_yaw\_belt"   
## [21] "min\_roll\_belt" "min\_pitch\_belt"   
## [23] "min\_yaw\_belt" "amplitude\_roll\_belt"   
## [25] "amplitude\_pitch\_belt" "amplitude\_yaw\_belt"   
## [27] "var\_total\_accel\_belt" "avg\_roll\_belt"   
## [29] "stddev\_roll\_belt" "var\_roll\_belt"   
## [31] "avg\_pitch\_belt" "stddev\_pitch\_belt"   
## [33] "var\_pitch\_belt" "avg\_yaw\_belt"   
## [35] "stddev\_yaw\_belt" "var\_yaw\_belt"   
## [37] "gyros\_belt\_x" "gyros\_belt\_y"   
## [39] "gyros\_belt\_z" "accel\_belt\_x"   
## [41] "accel\_belt\_y" "accel\_belt\_z"   
## [43] "magnet\_belt\_x" "magnet\_belt\_y"   
## [45] "magnet\_belt\_z" "roll\_arm"   
## [47] "pitch\_arm" "yaw\_arm"   
## [49] "total\_accel\_arm" "var\_accel\_arm"   
## [51] "avg\_roll\_arm" "stddev\_roll\_arm"   
## [53] "var\_roll\_arm" "avg\_pitch\_arm"   
## [55] "stddev\_pitch\_arm" "var\_pitch\_arm"   
## [57] "avg\_yaw\_arm" "stddev\_yaw\_arm"   
## [59] "var\_yaw\_arm" "gyros\_arm\_x"   
## [61] "gyros\_arm\_y" "gyros\_arm\_z"   
## [63] "accel\_arm\_x" "accel\_arm\_y"   
## [65] "accel\_arm\_z" "magnet\_arm\_x"   
## [67] "magnet\_arm\_y" "magnet\_arm\_z"   
## [69] "kurtosis\_roll\_arm" "kurtosis\_picth\_arm"   
## [71] "kurtosis\_yaw\_arm" "skewness\_roll\_arm"   
## [73] "skewness\_pitch\_arm" "skewness\_yaw\_arm"   
## [75] "max\_roll\_arm" "max\_picth\_arm"   
## [77] "max\_yaw\_arm" "min\_roll\_arm"   
## [79] "min\_pitch\_arm" "min\_yaw\_arm"   
## [81] "amplitude\_roll\_arm" "amplitude\_pitch\_arm"   
## [83] "amplitude\_yaw\_arm" "roll\_dumbbell"   
## [85] "pitch\_dumbbell" "yaw\_dumbbell"   
## [87] "kurtosis\_roll\_dumbbell" "kurtosis\_picth\_dumbbell"   
## [89] "kurtosis\_yaw\_dumbbell" "skewness\_roll\_dumbbell"   
## [91] "skewness\_pitch\_dumbbell" "skewness\_yaw\_dumbbell"   
## [93] "max\_roll\_dumbbell" "max\_picth\_dumbbell"   
## [95] "max\_yaw\_dumbbell" "min\_roll\_dumbbell"   
## [97] "min\_pitch\_dumbbell" "min\_yaw\_dumbbell"   
## [99] "amplitude\_roll\_dumbbell" "amplitude\_pitch\_dumbbell"  
## [101] "amplitude\_yaw\_dumbbell" "total\_accel\_dumbbell"   
## [103] "var\_accel\_dumbbell" "avg\_roll\_dumbbell"   
## [105] "stddev\_roll\_dumbbell" "var\_roll\_dumbbell"   
## [107] "avg\_pitch\_dumbbell" "stddev\_pitch\_dumbbell"   
## [109] "var\_pitch\_dumbbell" "avg\_yaw\_dumbbell"   
## [111] "stddev\_yaw\_dumbbell" "var\_yaw\_dumbbell"   
## [113] "gyros\_dumbbell\_x" "gyros\_dumbbell\_y"   
## [115] "gyros\_dumbbell\_z" "accel\_dumbbell\_x"   
## [117] "accel\_dumbbell\_y" "accel\_dumbbell\_z"   
## [119] "magnet\_dumbbell\_x" "magnet\_dumbbell\_y"   
## [121] "magnet\_dumbbell\_z" "roll\_forearm"   
## [123] "pitch\_forearm" "yaw\_forearm"   
## [125] "kurtosis\_roll\_forearm" "kurtosis\_picth\_forearm"   
## [127] "kurtosis\_yaw\_forearm" "skewness\_roll\_forearm"   
## [129] "skewness\_pitch\_forearm" "skewness\_yaw\_forearm"   
## [131] "max\_roll\_forearm" "max\_picth\_forearm"   
## [133] "max\_yaw\_forearm" "min\_roll\_forearm"   
## [135] "min\_pitch\_forearm" "min\_yaw\_forearm"   
## [137] "amplitude\_roll\_forearm" "amplitude\_pitch\_forearm"   
## [139] "amplitude\_yaw\_forearm" "total\_accel\_forearm"   
## [141] "var\_accel\_forearm" "avg\_roll\_forearm"   
## [143] "stddev\_roll\_forearm" "var\_roll\_forearm"   
## [145] "avg\_pitch\_forearm" "stddev\_pitch\_forearm"   
## [147] "var\_pitch\_forearm" "avg\_yaw\_forearm"   
## [149] "stddev\_yaw\_forearm" "var\_yaw\_forearm"   
## [151] "gyros\_forearm\_x" "gyros\_forearm\_y"   
## [153] "gyros\_forearm\_z" "accel\_forearm\_x"   
## [155] "accel\_forearm\_y" "accel\_forearm\_z"   
## [157] "magnet\_forearm\_x" "magnet\_forearm\_y"   
## [159] "magnet\_forearm\_z" "problem\_id"

We can see the similar column names. Here, we do not have to know the details.

Here, see the data types.

sapply(df, class)

## id user\_name raw\_timestamp\_part\_1   
## "numeric" "character" "numeric"   
## raw\_timestamp\_part\_2 cvtd\_timestamp new\_window   
## "numeric" "character" "character"   
## num\_window roll\_belt pitch\_belt   
## "numeric" "numeric" "numeric"   
## yaw\_belt total\_accel\_belt kurtosis\_roll\_belt   
## "numeric" "numeric" "character"   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt   
## "character" "character" "character"   
## skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt   
## "character" "character" "numeric"   
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## "numeric" "character" "numeric"   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt   
## "numeric" "character" "numeric"   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt   
## "numeric" "character" "numeric"   
## avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt   
## "numeric" "numeric" "numeric"   
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt   
## "numeric" "numeric" "numeric"   
## avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## "numeric" "numeric" "numeric"   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z   
## "numeric" "numeric" "numeric"   
## accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## "numeric" "numeric" "numeric"   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z   
## "numeric" "numeric" "numeric"   
## roll\_arm pitch\_arm yaw\_arm   
## "numeric" "numeric" "numeric"   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm   
## "numeric" "numeric" "numeric"   
## stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm   
## "numeric" "numeric" "numeric"   
## stddev\_pitch\_arm var\_pitch\_arm avg\_yaw\_arm   
## "numeric" "numeric" "numeric"   
## stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## "numeric" "numeric" "numeric"   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x   
## "numeric" "numeric" "numeric"   
## accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## "numeric" "numeric" "numeric"   
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm   
## "numeric" "numeric" "numeric"   
## kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm   
## "character" "character" "numeric"   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm   
## "character" "character" "numeric"   
## max\_picth\_arm max\_yaw\_arm min\_roll\_arm   
## "numeric" "numeric" "numeric"   
## min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm   
## "numeric" "numeric" "numeric"   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## "numeric" "numeric" "numeric"   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell   
## "numeric" "numeric" "numeric"   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell   
## "numeric" "character" "numeric"   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell   
## "numeric" "character" "numeric"   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell   
## "numeric" "numeric" "numeric"   
## min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell   
## "numeric" "numeric" "numeric"   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell   
## "numeric" "numeric" "numeric"   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell   
## "numeric" "numeric" "numeric"   
## var\_roll\_dumbbell avg\_pitch\_dumbbell stddev\_pitch\_dumbbell   
## "numeric" "numeric" "numeric"   
## var\_pitch\_dumbbell avg\_yaw\_dumbbell stddev\_yaw\_dumbbell   
## "numeric" "numeric" "numeric"   
## var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## "numeric" "numeric" "numeric"   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y   
## "numeric" "numeric" "numeric"   
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y   
## "numeric" "numeric" "numeric"   
## magnet\_dumbbell\_z roll\_forearm pitch\_forearm   
## "numeric" "numeric" "numeric"   
## yaw\_forearm kurtosis\_roll\_forearm kurtosis\_picth\_forearm   
## "numeric" "character" "character"   
## kurtosis\_yaw\_forearm skewness\_roll\_forearm skewness\_pitch\_forearm   
## "character" "character" "character"   
## skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm   
## "character" "numeric" "numeric"   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm   
## "character" "numeric" "numeric"   
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm   
## "character" "numeric" "numeric"   
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm   
## "character" "numeric" "numeric"   
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm   
## "numeric" "numeric" "numeric"   
## avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## "numeric" "numeric" "numeric"   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm   
## "numeric" "numeric" "numeric"   
## gyros\_forearm\_x gyros\_forearm\_y gyros\_forearm\_z   
## "numeric" "numeric" "numeric"   
## accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z   
## "numeric" "numeric" "numeric"   
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z   
## "numeric" "numeric" "numeric"   
## classe   
## "character"

sapply(df\_test, class)

## id user\_name raw\_timestamp\_part\_1   
## "numeric" "character" "numeric"   
## raw\_timestamp\_part\_2 cvtd\_timestamp new\_window   
## "numeric" "character" "character"   
## num\_window roll\_belt pitch\_belt   
## "numeric" "numeric" "numeric"   
## yaw\_belt total\_accel\_belt kurtosis\_roll\_belt   
## "numeric" "numeric" "logical"   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt   
## "logical" "logical" "logical"   
## skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt   
## "logical" "logical" "logical"   
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## "logical" "logical" "logical"   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt   
## "logical" "logical" "logical"   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt   
## "logical" "logical" "logical"   
## avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt   
## "logical" "logical" "logical"   
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt   
## "logical" "logical" "logical"   
## avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## "logical" "logical" "logical"   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z   
## "numeric" "numeric" "numeric"   
## accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## "numeric" "numeric" "numeric"   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z   
## "numeric" "numeric" "numeric"   
## roll\_arm pitch\_arm yaw\_arm   
## "numeric" "numeric" "numeric"   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm   
## "numeric" "logical" "logical"   
## stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm   
## "logical" "logical" "logical"   
## stddev\_pitch\_arm var\_pitch\_arm avg\_yaw\_arm   
## "logical" "logical" "logical"   
## stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## "logical" "logical" "numeric"   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x   
## "numeric" "numeric" "numeric"   
## accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## "numeric" "numeric" "numeric"   
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm   
## "numeric" "numeric" "logical"   
## kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm   
## "logical" "logical" "logical"   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm   
## "logical" "logical" "logical"   
## max\_picth\_arm max\_yaw\_arm min\_roll\_arm   
## "logical" "logical" "logical"   
## min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm   
## "logical" "logical" "logical"   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## "logical" "logical" "numeric"   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell   
## "numeric" "numeric" "logical"   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell   
## "logical" "logical" "logical"   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell   
## "logical" "logical" "logical"   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell   
## "logical" "logical" "logical"   
## min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell   
## "logical" "logical" "logical"   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell   
## "logical" "logical" "numeric"   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell   
## "logical" "logical" "logical"   
## var\_roll\_dumbbell avg\_pitch\_dumbbell stddev\_pitch\_dumbbell   
## "logical" "logical" "logical"   
## var\_pitch\_dumbbell avg\_yaw\_dumbbell stddev\_yaw\_dumbbell   
## "logical" "logical" "logical"   
## var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## "logical" "numeric" "numeric"   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y   
## "numeric" "numeric" "numeric"   
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y   
## "numeric" "numeric" "numeric"   
## magnet\_dumbbell\_z roll\_forearm pitch\_forearm   
## "numeric" "numeric" "numeric"   
## yaw\_forearm kurtosis\_roll\_forearm kurtosis\_picth\_forearm   
## "numeric" "logical" "logical"   
## kurtosis\_yaw\_forearm skewness\_roll\_forearm skewness\_pitch\_forearm   
## "logical" "logical" "logical"   
## skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm   
## "logical" "logical" "logical"   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm   
## "logical" "logical" "logical"   
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm   
## "logical" "logical" "logical"   
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm   
## "logical" "numeric" "logical"   
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm   
## "logical" "logical" "logical"   
## avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## "logical" "logical" "logical"   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm   
## "logical" "logical" "logical"   
## gyros\_forearm\_x gyros\_forearm\_y gyros\_forearm\_z   
## "numeric" "numeric" "numeric"   
## accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z   
## "numeric" "numeric" "numeric"   
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z   
## "numeric" "numeric" "numeric"   
## problem\_id   
## "numeric"

Fortunately, almost all of the variables are numeric. But there are a few columns having character values. I need to change them into numeric or factor values before running a machine learning model.

As to columns having character values, here check the exact values.

head(Filter(is.character, df), 5)

## # A tibble: 5 x 28  
## user\_name cvtd\_timestamp new\_window kurtosis\_roll\_b~ kurtosis\_picth\_~  
## <chr> <chr> <chr> <chr> <chr>   
## 1 carlitos 05/12/2011 11~ no <NA> <NA>   
## 2 carlitos 05/12/2011 11~ no <NA> <NA>   
## 3 carlitos 05/12/2011 11~ no <NA> <NA>   
## 4 carlitos 05/12/2011 11~ no <NA> <NA>   
## 5 carlitos 05/12/2011 11~ no <NA> <NA>   
## # ... with 23 more variables: kurtosis\_yaw\_belt <chr>,  
## # skewness\_roll\_belt <chr>, skewness\_roll\_belt.1 <chr>,  
## # skewness\_yaw\_belt <chr>, max\_yaw\_belt <chr>, min\_yaw\_belt <chr>,  
## # amplitude\_yaw\_belt <chr>, kurtosis\_picth\_arm <chr>, kurtosis\_yaw\_arm <chr>,  
## # skewness\_pitch\_arm <chr>, skewness\_yaw\_arm <chr>,  
## # kurtosis\_yaw\_dumbbell <chr>, skewness\_yaw\_dumbbell <chr>,  
## # kurtosis\_roll\_forearm <chr>, kurtosis\_picth\_forearm <chr>,  
## # kurtosis\_yaw\_forearm <chr>, skewness\_roll\_forearm <chr>,  
## # skewness\_pitch\_forearm <chr>, skewness\_yaw\_forearm <chr>,  
## # max\_yaw\_forearm <chr>, min\_yaw\_forearm <chr>, amplitude\_yaw\_forearm <chr>,  
## # classe <chr>

unique(df$classe)

## [1] "A" "B" "C" "D" "E"

unique(df$new\_window)

## [1] "no" "yes"

The response variable classe is composed of A, B, C, D, E.

See the descriptive statistics of the dataset. Generally, it is tough to follow all of statistical values, but this is an important step to grasp the data.

summary(df)

## id user\_name raw\_timestamp\_part\_1 raw\_timestamp\_part\_2  
## Min. : 1 Length:19622 Min. :1.322e+09 Min. : 294   
## 1st Qu.: 4906 Class :character 1st Qu.:1.323e+09 1st Qu.:252912   
## Median : 9812 Mode :character Median :1.323e+09 Median :496380   
## Mean : 9812 Mean :1.323e+09 Mean :500656   
## 3rd Qu.:14717 3rd Qu.:1.323e+09 3rd Qu.:751891   
## Max. :19622 Max. :1.323e+09 Max. :998801   
##   
## cvtd\_timestamp new\_window num\_window roll\_belt   
## Length:19622 Length:19622 Min. : 1.0 Min. :-28.90   
## Class :character Class :character 1st Qu.:222.0 1st Qu.: 1.10   
## Mode :character Mode :character Median :424.0 Median :113.00   
## Mean :430.6 Mean : 64.41   
## 3rd Qu.:644.0 3rd Qu.:123.00   
## Max. :864.0 Max. :162.00   
##   
## pitch\_belt yaw\_belt total\_accel\_belt kurtosis\_roll\_belt  
## Min. :-55.8000 Min. :-180.00 Min. : 0.00 Length:19622   
## 1st Qu.: 1.7600 1st Qu.: -88.30 1st Qu.: 3.00 Class :character   
## Median : 5.2800 Median : -13.00 Median :17.00 Mode :character   
## Mean : 0.3053 Mean : -11.21 Mean :11.31   
## 3rd Qu.: 14.9000 3rd Qu.: 12.90 3rd Qu.:18.00   
## Max. : 60.3000 Max. : 179.00 Max. :29.00   
##   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt skewness\_roll\_belt.1  
## Length:19622 Length:19622 Length:19622 Length:19622   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## skewness\_yaw\_belt max\_roll\_belt max\_picth\_belt max\_yaw\_belt   
## Length:19622 Min. :-94.300 Min. : 3.00 Length:19622   
## Class :character 1st Qu.:-88.000 1st Qu.: 5.00 Class :character   
## Mode :character Median : -5.100 Median :18.00 Mode :character   
## Mean : -6.667 Mean :12.92   
## 3rd Qu.: 18.500 3rd Qu.:19.00   
## Max. :180.000 Max. :30.00   
## NA's :19216 NA's :19216   
## min\_roll\_belt min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt  
## Min. :-180.00 Min. : 0.00 Length:19622 Min. : 0.000   
## 1st Qu.: -88.40 1st Qu.: 3.00 Class :character 1st Qu.: 0.300   
## Median : -7.85 Median :16.00 Mode :character Median : 1.000   
## Mean : -10.44 Mean :10.76 Mean : 3.769   
## 3rd Qu.: 9.05 3rd Qu.:17.00 3rd Qu.: 2.083   
## Max. : 173.00 Max. :23.00 Max. :360.000   
## NA's :19216 NA's :19216 NA's :19216   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt avg\_roll\_belt   
## Min. : 0.000 Length:19622 Min. : 0.000 Min. :-27.40   
## 1st Qu.: 1.000 Class :character 1st Qu.: 0.100 1st Qu.: 1.10   
## Median : 1.000 Mode :character Median : 0.200 Median :116.35   
## Mean : 2.167 Mean : 0.926 Mean : 68.06   
## 3rd Qu.: 2.000 3rd Qu.: 0.300 3rd Qu.:123.38   
## Max. :12.000 Max. :16.500 Max. :157.40   
## NA's :19216 NA's :19216 NA's :19216   
## stddev\_roll\_belt var\_roll\_belt avg\_pitch\_belt stddev\_pitch\_belt  
## Min. : 0.000 Min. : 0.000 Min. :-51.400 Min. :0.000   
## 1st Qu.: 0.200 1st Qu.: 0.000 1st Qu.: 2.025 1st Qu.:0.200   
## Median : 0.400 Median : 0.100 Median : 5.200 Median :0.400   
## Mean : 1.337 Mean : 7.699 Mean : 0.520 Mean :0.603   
## 3rd Qu.: 0.700 3rd Qu.: 0.500 3rd Qu.: 15.775 3rd Qu.:0.700   
## Max. :14.200 Max. :200.700 Max. : 59.700 Max. :4.000   
## NA's :19216 NA's :19216 NA's :19216 NA's :19216   
## var\_pitch\_belt avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## Min. : 0.000 Min. :-138.300 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.000 1st Qu.: -88.175 1st Qu.: 0.100 1st Qu.: 0.010   
## Median : 0.100 Median : -6.550 Median : 0.300 Median : 0.090   
## Mean : 0.766 Mean : -8.831 Mean : 1.341 Mean : 107.487   
## 3rd Qu.: 0.500 3rd Qu.: 14.125 3rd Qu.: 0.700 3rd Qu.: 0.475   
## Max. :16.200 Max. : 173.500 Max. :176.600 Max. :31183.240   
## NA's :19216 NA's :19216 NA's :19216 NA's :19216   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z accel\_belt\_x   
## Min. :-1.040000 Min. :-0.64000 Min. :-1.4600 Min. :-120.000   
## 1st Qu.:-0.030000 1st Qu.: 0.00000 1st Qu.:-0.2000 1st Qu.: -21.000   
## Median : 0.030000 Median : 0.02000 Median :-0.1000 Median : -15.000   
## Mean :-0.005592 Mean : 0.03959 Mean :-0.1305 Mean : -5.595   
## 3rd Qu.: 0.110000 3rd Qu.: 0.11000 3rd Qu.:-0.0200 3rd Qu.: -5.000   
## Max. : 2.220000 Max. : 0.64000 Max. : 1.6200 Max. : 85.000   
##   
## accel\_belt\_y accel\_belt\_z magnet\_belt\_x magnet\_belt\_y   
## Min. :-69.00 Min. :-275.00 Min. :-52.0 Min. :354.0   
## 1st Qu.: 3.00 1st Qu.:-162.00 1st Qu.: 9.0 1st Qu.:581.0   
## Median : 35.00 Median :-152.00 Median : 35.0 Median :601.0   
## Mean : 30.15 Mean : -72.59 Mean : 55.6 Mean :593.7   
## 3rd Qu.: 61.00 3rd Qu.: 27.00 3rd Qu.: 59.0 3rd Qu.:610.0   
## Max. :164.00 Max. : 105.00 Max. :485.0 Max. :673.0   
##   
## magnet\_belt\_z roll\_arm pitch\_arm yaw\_arm   
## Min. :-623.0 Min. :-180.00 Min. :-88.800 Min. :-180.0000   
## 1st Qu.:-375.0 1st Qu.: -31.77 1st Qu.:-25.900 1st Qu.: -43.1000   
## Median :-320.0 Median : 0.00 Median : 0.000 Median : 0.0000   
## Mean :-345.5 Mean : 17.83 Mean : -4.612 Mean : -0.6188   
## 3rd Qu.:-306.0 3rd Qu.: 77.30 3rd Qu.: 11.200 3rd Qu.: 45.8750   
## Max. : 293.0 Max. : 180.00 Max. : 88.500 Max. : 180.0000   
##   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm stddev\_roll\_arm   
## Min. : 1.00 Min. : 0.00 Min. :-166.67 Min. : 0.000   
## 1st Qu.:17.00 1st Qu.: 9.03 1st Qu.: -38.37 1st Qu.: 1.376   
## Median :27.00 Median : 40.61 Median : 0.00 Median : 5.702   
## Mean :25.51 Mean : 53.23 Mean : 12.68 Mean : 11.201   
## 3rd Qu.:33.00 3rd Qu.: 75.62 3rd Qu.: 76.33 3rd Qu.: 14.921   
## Max. :66.00 Max. :331.70 Max. : 163.33 Max. :161.964   
## NA's :19216 NA's :19216 NA's :19216   
## var\_roll\_arm avg\_pitch\_arm stddev\_pitch\_arm var\_pitch\_arm   
## Min. : 0.000 Min. :-81.773 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 1.898 1st Qu.:-22.770 1st Qu.: 1.642 1st Qu.: 2.697   
## Median : 32.517 Median : 0.000 Median : 8.133 Median : 66.146   
## Mean : 417.264 Mean : -4.901 Mean :10.383 Mean : 195.864   
## 3rd Qu.: 222.647 3rd Qu.: 8.277 3rd Qu.:16.327 3rd Qu.: 266.576   
## Max. :26232.208 Max. : 75.659 Max. :43.412 Max. :1884.565   
## NA's :19216 NA's :19216 NA's :19216 NA's :19216   
## avg\_yaw\_arm stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## Min. :-173.440 Min. : 0.000 Min. : 0.000 Min. :-6.37000   
## 1st Qu.: -29.198 1st Qu.: 2.577 1st Qu.: 6.642 1st Qu.:-1.33000   
## Median : 0.000 Median : 16.682 Median : 278.309 Median : 0.08000   
## Mean : 2.359 Mean : 22.270 Mean : 1055.933 Mean : 0.04277   
## 3rd Qu.: 38.185 3rd Qu.: 35.984 3rd Qu.: 1294.850 3rd Qu.: 1.57000   
## Max. : 152.000 Max. :177.044 Max. :31344.568 Max. : 4.87000   
## NA's :19216 NA's :19216 NA's :19216   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x accel\_arm\_y   
## Min. :-3.4400 Min. :-2.3300 Min. :-404.00 Min. :-318.0   
## 1st Qu.:-0.8000 1st Qu.:-0.0700 1st Qu.:-242.00 1st Qu.: -54.0   
## Median :-0.2400 Median : 0.2300 Median : -44.00 Median : 14.0   
## Mean :-0.2571 Mean : 0.2695 Mean : -60.24 Mean : 32.6   
## 3rd Qu.: 0.1400 3rd Qu.: 0.7200 3rd Qu.: 84.00 3rd Qu.: 139.0   
## Max. : 2.8400 Max. : 3.0200 Max. : 437.00 Max. : 308.0   
##   
## accel\_arm\_z magnet\_arm\_x magnet\_arm\_y magnet\_arm\_z   
## Min. :-636.00 Min. :-584.0 Min. :-392.0 Min. :-597.0   
## 1st Qu.:-143.00 1st Qu.:-300.0 1st Qu.: -9.0 1st Qu.: 131.2   
## Median : -47.00 Median : 289.0 Median : 202.0 Median : 444.0   
## Mean : -71.25 Mean : 191.7 Mean : 156.6 Mean : 306.5   
## 3rd Qu.: 23.00 3rd Qu.: 637.0 3rd Qu.: 323.0 3rd Qu.: 545.0   
## Max. : 292.00 Max. : 782.0 Max. : 583.0 Max. : 694.0   
##   
## kurtosis\_roll\_arm kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm  
## Min. :-1.809 Length:19622 Length:19622 Min. :-2.541   
## 1st Qu.:-1.345 Class :character Class :character 1st Qu.:-0.561   
## Median :-0.894 Mode :character Mode :character Median : 0.040   
## Mean :-0.366 Mean : 0.068   
## 3rd Qu.:-0.038 3rd Qu.: 0.671   
## Max. :21.456 Max. : 4.394   
## NA's :19294 NA's :19293   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm max\_picth\_arm   
## Length:19622 Length:19622 Min. :-73.100 Min. :-173.000   
## Class :character Class :character 1st Qu.: -0.175 1st Qu.: -1.975   
## Mode :character Mode :character Median : 4.950 Median : 23.250   
## Mean : 11.236 Mean : 35.751   
## 3rd Qu.: 26.775 3rd Qu.: 95.975   
## Max. : 85.500 Max. : 180.000   
## NA's :19216 NA's :19216   
## max\_yaw\_arm min\_roll\_arm min\_pitch\_arm min\_yaw\_arm   
## Min. : 4.00 Min. :-89.10 Min. :-180.00 Min. : 1.00   
## 1st Qu.:29.00 1st Qu.:-41.98 1st Qu.: -72.62 1st Qu.: 8.00   
## Median :34.00 Median :-22.45 Median : -33.85 Median :13.00   
## Mean :35.46 Mean :-21.22 Mean : -33.92 Mean :14.66   
## 3rd Qu.:41.00 3rd Qu.: 0.00 3rd Qu.: 0.00 3rd Qu.:19.00   
## Max. :65.00 Max. : 66.40 Max. : 152.00 Max. :38.00   
## NA's :19216 NA's :19216 NA's :19216 NA's :19216   
## amplitude\_roll\_arm amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## Min. : 0.000 Min. : 0.000 Min. : 0.00 Min. :-153.71   
## 1st Qu.: 5.425 1st Qu.: 9.925 1st Qu.:13.00 1st Qu.: -18.49   
## Median : 28.450 Median : 54.900 Median :22.00 Median : 48.17   
## Mean : 32.452 Mean : 69.677 Mean :20.79 Mean : 23.84   
## 3rd Qu.: 50.960 3rd Qu.:115.175 3rd Qu.:28.75 3rd Qu.: 67.61   
## Max. :119.500 Max. :360.000 Max. :52.00 Max. : 153.55   
## NA's :19216 NA's :19216 NA's :19216   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell  
## Min. :-149.59 Min. :-150.871 Min. :-2.174   
## 1st Qu.: -40.89 1st Qu.: -77.644 1st Qu.:-0.682   
## Median : -20.96 Median : -3.324 Median :-0.033   
## Mean : -10.78 Mean : 1.674 Mean : 0.452   
## 3rd Qu.: 17.50 3rd Qu.: 79.643 3rd Qu.: 0.940   
## Max. : 149.40 Max. : 154.952 Max. :54.998   
## NA's :19221   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell  
## Min. :-2.200 Length:19622 Min. :-7.384   
## 1st Qu.:-0.721 Class :character 1st Qu.:-0.581   
## Median :-0.133 Mode :character Median :-0.076   
## Mean : 0.286 Mean :-0.115   
## 3rd Qu.: 0.584 3rd Qu.: 0.400   
## Max. :55.628 Max. : 1.958   
## NA's :19218 NA's :19220   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell  
## Min. :-7.447 Length:19622 Min. :-70.10   
## 1st Qu.:-0.526 Class :character 1st Qu.:-27.15   
## Median :-0.091 Mode :character Median : 14.85   
## Mean :-0.035 Mean : 13.76   
## 3rd Qu.: 0.505 3rd Qu.: 50.58   
## Max. : 3.769 Max. :137.00   
## NA's :19217 NA's :19216   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell min\_pitch\_dumbbell  
## Min. :-112.90 Min. :-2.20 Min. :-149.60 Min. :-147.00   
## 1st Qu.: -66.70 1st Qu.:-0.70 1st Qu.: -59.67 1st Qu.: -91.80   
## Median : 40.05 Median : 0.00 Median : -43.55 Median : -66.15   
## Mean : 32.75 Mean : 0.45 Mean : -41.24 Mean : -33.18   
## 3rd Qu.: 133.22 3rd Qu.: 0.90 3rd Qu.: -25.20 3rd Qu.: 21.20   
## Max. : 155.00 Max. :55.00 Max. : 73.20 Max. : 120.90   
## NA's :19216 NA's :19221 NA's :19216 NA's :19216   
## min\_yaw\_dumbbell amplitude\_roll\_dumbbell amplitude\_pitch\_dumbbell  
## Min. :-2.20 Min. : 0.00 Min. : 0.00   
## 1st Qu.:-0.70 1st Qu.: 14.97 1st Qu.: 17.06   
## Median : 0.00 Median : 35.05 Median : 41.73   
## Mean : 0.45 Mean : 55.00 Mean : 65.93   
## 3rd Qu.: 0.90 3rd Qu.: 81.04 3rd Qu.: 99.55   
## Max. :55.00 Max. :256.48 Max. :273.59   
## NA's :19221 NA's :19216 NA's :19216   
## amplitude\_yaw\_dumbbell total\_accel\_dumbbell var\_accel\_dumbbell  
## Min. :0 Min. : 0.00 Min. : 0.000   
## 1st Qu.:0 1st Qu.: 4.00 1st Qu.: 0.378   
## Median :0 Median :10.00 Median : 1.000   
## Mean :0 Mean :13.72 Mean : 4.388   
## 3rd Qu.:0 3rd Qu.:19.00 3rd Qu.: 3.434   
## Max. :0 Max. :58.00 Max. :230.428   
## NA's :19221 NA's :19216   
## avg\_roll\_dumbbell stddev\_roll\_dumbbell var\_roll\_dumbbell avg\_pitch\_dumbbell  
## Min. :-128.96 Min. : 0.000 Min. : 0.00 Min. :-70.73   
## 1st Qu.: -12.33 1st Qu.: 4.639 1st Qu.: 21.52 1st Qu.:-42.00   
## Median : 48.23 Median : 12.204 Median : 148.95 Median :-19.91   
## Mean : 23.86 Mean : 20.761 Mean : 1020.27 Mean :-12.33   
## 3rd Qu.: 64.37 3rd Qu.: 26.356 3rd Qu.: 694.65 3rd Qu.: 13.21   
## Max. : 125.99 Max. :123.778 Max. :15321.01 Max. : 94.28   
## NA's :19216 NA's :19216 NA's :19216 NA's :19216   
## stddev\_pitch\_dumbbell var\_pitch\_dumbbell avg\_yaw\_dumbbell   
## Min. : 0.000 Min. : 0.00 Min. :-117.950   
## 1st Qu.: 3.482 1st Qu.: 12.12 1st Qu.: -76.696   
## Median : 8.089 Median : 65.44 Median : -4.505   
## Mean :13.147 Mean : 350.31 Mean : 0.202   
## 3rd Qu.:19.238 3rd Qu.: 370.11 3rd Qu.: 71.234   
## Max. :82.680 Max. :6836.02 Max. : 134.905   
## NA's :19216 NA's :19216 NA's :19216   
## stddev\_yaw\_dumbbell var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## Min. : 0.000 Min. : 0.00 Min. :-204.0000 Min. :-2.10000   
## 1st Qu.: 3.885 1st Qu.: 15.09 1st Qu.: -0.0300 1st Qu.:-0.14000   
## Median : 10.264 Median : 105.35 Median : 0.1300 Median : 0.03000   
## Mean : 16.647 Mean : 589.84 Mean : 0.1611 Mean : 0.04606   
## 3rd Qu.: 24.674 3rd Qu.: 608.79 3rd Qu.: 0.3500 3rd Qu.: 0.21000   
## Max. :107.088 Max. :11467.91 Max. : 2.2200 Max. :52.00000   
## NA's :19216 NA's :19216   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y accel\_dumbbell\_z   
## Min. : -2.380 Min. :-419.00 Min. :-189.00 Min. :-334.00   
## 1st Qu.: -0.310 1st Qu.: -50.00 1st Qu.: -8.00 1st Qu.:-142.00   
## Median : -0.130 Median : -8.00 Median : 41.50 Median : -1.00   
## Mean : -0.129 Mean : -28.62 Mean : 52.63 Mean : -38.32   
## 3rd Qu.: 0.030 3rd Qu.: 11.00 3rd Qu.: 111.00 3rd Qu.: 38.00   
## Max. :317.000 Max. : 235.00 Max. : 315.00 Max. : 318.00   
##   
## magnet\_dumbbell\_x magnet\_dumbbell\_y magnet\_dumbbell\_z roll\_forearm   
## Min. :-643.0 Min. :-3600 Min. :-262.00 Min. :-180.0000   
## 1st Qu.:-535.0 1st Qu.: 231 1st Qu.: -45.00 1st Qu.: -0.7375   
## Median :-479.0 Median : 311 Median : 13.00 Median : 21.7000   
## Mean :-328.5 Mean : 221 Mean : 46.05 Mean : 33.8265   
## 3rd Qu.:-304.0 3rd Qu.: 390 3rd Qu.: 95.00 3rd Qu.: 140.0000   
## Max. : 592.0 Max. : 633 Max. : 452.00 Max. : 180.0000   
##   
## pitch\_forearm yaw\_forearm kurtosis\_roll\_forearm  
## Min. :-72.50 Min. :-180.00 Length:19622   
## 1st Qu.: 0.00 1st Qu.: -68.60 Class :character   
## Median : 9.24 Median : 0.00 Mode :character   
## Mean : 10.71 Mean : 19.21   
## 3rd Qu.: 28.40 3rd Qu.: 110.00   
## Max. : 89.80 Max. : 180.00   
##   
## kurtosis\_picth\_forearm kurtosis\_yaw\_forearm skewness\_roll\_forearm  
## Length:19622 Length:19622 Length:19622   
## Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character   
##   
##   
##   
##   
## skewness\_pitch\_forearm skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm  
## Length:19622 Length:19622 Min. :-66.60 Min. :-151.00   
## Class :character Class :character 1st Qu.: 0.00 1st Qu.: 0.00   
## Mode :character Mode :character Median : 26.80 Median : 113.00   
## Mean : 24.49 Mean : 81.49   
## 3rd Qu.: 45.95 3rd Qu.: 174.75   
## Max. : 89.80 Max. : 180.00   
## NA's :19216 NA's :19216   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm min\_yaw\_forearm   
## Length:19622 Min. :-72.500 Min. :-180.00 Length:19622   
## Class :character 1st Qu.: -6.075 1st Qu.:-175.00 Class :character   
## Mode :character Median : 0.000 Median : -61.00 Mode :character   
## Mean : -0.167 Mean : -57.57   
## 3rd Qu.: 12.075 3rd Qu.: 0.00   
## Max. : 62.100 Max. : 167.00   
## NA's :19216 NA's :19216   
## amplitude\_roll\_forearm amplitude\_pitch\_forearm amplitude\_yaw\_forearm  
## Min. : 0.000 Min. : 0.0 Length:19622   
## 1st Qu.: 1.125 1st Qu.: 2.0 Class :character   
## Median : 17.770 Median : 83.7 Mode :character   
## Mean : 24.653 Mean :139.1   
## 3rd Qu.: 39.875 3rd Qu.:350.0   
## Max. :126.000 Max. :360.0   
## NA's :19216 NA's :19216   
## total\_accel\_forearm var\_accel\_forearm avg\_roll\_forearm stddev\_roll\_forearm  
## Min. : 0.00 Min. : 0.000 Min. :-177.234 Min. : 0.000   
## 1st Qu.: 29.00 1st Qu.: 6.759 1st Qu.: -0.909 1st Qu.: 0.428   
## Median : 36.00 Median : 21.165 Median : 11.172 Median : 8.030   
## Mean : 34.72 Mean : 33.502 Mean : 33.165 Mean : 41.986   
## 3rd Qu.: 41.00 3rd Qu.: 51.240 3rd Qu.: 107.132 3rd Qu.: 85.373   
## Max. :108.00 Max. :172.606 Max. : 177.256 Max. :179.171   
## NA's :19216 NA's :19216 NA's :19216   
## var\_roll\_forearm avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## Min. : 0.00 Min. :-68.17 Min. : 0.000 Min. : 0.000   
## 1st Qu.: 0.18 1st Qu.: 0.00 1st Qu.: 0.336 1st Qu.: 0.113   
## Median : 64.48 Median : 12.02 Median : 5.516 Median : 30.425   
## Mean : 5274.10 Mean : 11.79 Mean : 7.977 Mean : 139.593   
## 3rd Qu.: 7289.08 3rd Qu.: 28.48 3rd Qu.:12.866 3rd Qu.: 165.532   
## Max. :32102.24 Max. : 72.09 Max. :47.745 Max. :2279.617   
## NA's :19216 NA's :19216 NA's :19216 NA's :19216   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm gyros\_forearm\_x   
## Min. :-155.06 Min. : 0.000 Min. : 0.00 Min. :-22.000   
## 1st Qu.: -26.26 1st Qu.: 0.524 1st Qu.: 0.27 1st Qu.: -0.220   
## Median : 0.00 Median : 24.743 Median : 612.21 Median : 0.050   
## Mean : 18.00 Mean : 44.854 Mean : 4639.85 Mean : 0.158   
## 3rd Qu.: 85.79 3rd Qu.: 85.817 3rd Qu.: 7368.41 3rd Qu.: 0.560   
## Max. : 169.24 Max. :197.508 Max. :39009.33 Max. : 3.970   
## NA's :19216 NA's :19216 NA's :19216   
## gyros\_forearm\_y gyros\_forearm\_z accel\_forearm\_x accel\_forearm\_y   
## Min. : -7.02000 Min. : -8.0900 Min. :-498.00 Min. :-632.0   
## 1st Qu.: -1.46000 1st Qu.: -0.1800 1st Qu.:-178.00 1st Qu.: 57.0   
## Median : 0.03000 Median : 0.0800 Median : -57.00 Median : 201.0   
## Mean : 0.07517 Mean : 0.1512 Mean : -61.65 Mean : 163.7   
## 3rd Qu.: 1.62000 3rd Qu.: 0.4900 3rd Qu.: 76.00 3rd Qu.: 312.0   
## Max. :311.00000 Max. :231.0000 Max. : 477.00 Max. : 923.0   
##   
## accel\_forearm\_z magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z  
## Min. :-446.00 Min. :-1280.0 Min. :-896.0 Min. :-973.0   
## 1st Qu.:-182.00 1st Qu.: -616.0 1st Qu.: 2.0 1st Qu.: 191.0   
## Median : -39.00 Median : -378.0 Median : 591.0 Median : 511.0   
## Mean : -55.29 Mean : -312.6 Mean : 380.1 Mean : 393.6   
## 3rd Qu.: 26.00 3rd Qu.: -73.0 3rd Qu.: 737.0 3rd Qu.: 653.0   
## Max. : 291.00 Max. : 672.0 Max. :1480.0 Max. :1090.0   
##   
## classe   
## Length:19622   
## Class :character   
## Mode :character   
##   
##   
##   
##

summary(df\_test)

## id user\_name raw\_timestamp\_part\_1 raw\_timestamp\_part\_2  
## Min. : 1.00 Length:20 Min. :1.322e+09 Min. : 36553   
## 1st Qu.: 5.75 Class :character 1st Qu.:1.323e+09 1st Qu.:268655   
## Median :10.50 Mode :character Median :1.323e+09 Median :530706   
## Mean :10.50 Mean :1.323e+09 Mean :512167   
## 3rd Qu.:15.25 3rd Qu.:1.323e+09 3rd Qu.:787738   
## Max. :20.00 Max. :1.323e+09 Max. :920315   
## cvtd\_timestamp new\_window num\_window roll\_belt   
## Length:20 Length:20 Min. : 48.0 Min. : -5.9200   
## Class :character Class :character 1st Qu.:250.0 1st Qu.: 0.9075   
## Mode :character Mode :character Median :384.5 Median : 1.1100   
## Mean :379.6 Mean : 31.3055   
## 3rd Qu.:467.0 3rd Qu.: 32.5050   
## Max. :859.0 Max. :129.0000   
## pitch\_belt yaw\_belt total\_accel\_belt kurtosis\_roll\_belt  
## Min. :-41.600 Min. :-93.70 Min. : 2.00 Mode:logical   
## 1st Qu.: 3.013 1st Qu.:-88.62 1st Qu.: 3.00 NA's:20   
## Median : 4.655 Median :-87.85 Median : 4.00   
## Mean : 5.824 Mean :-59.30 Mean : 7.55   
## 3rd Qu.: 6.135 3rd Qu.:-63.50 3rd Qu.: 8.00   
## Max. : 27.800 Max. :162.00 Max. :21.00   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt skewness\_roll\_belt.1  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## skewness\_yaw\_belt max\_roll\_belt max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## Mode:logical Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt amplitude\_pitch\_belt  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## amplitude\_yaw\_belt var\_total\_accel\_belt avg\_roll\_belt stddev\_roll\_belt  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## var\_roll\_belt avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt avg\_yaw\_belt   
## Mode:logical Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## stddev\_yaw\_belt var\_yaw\_belt gyros\_belt\_x gyros\_belt\_y   
## Mode:logical Mode:logical Min. :-0.500 Min. :-0.050   
## NA's:20 NA's:20 1st Qu.:-0.070 1st Qu.:-0.005   
## Median : 0.020 Median : 0.000   
## Mean :-0.045 Mean : 0.010   
## 3rd Qu.: 0.070 3rd Qu.: 0.020   
## Max. : 0.240 Max. : 0.110   
## gyros\_belt\_z accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## Min. :-0.4800 Min. :-48.00 Min. :-16.00 Min. :-187.00   
## 1st Qu.:-0.1375 1st Qu.:-19.00 1st Qu.: 2.00 1st Qu.: -24.00   
## Median :-0.0250 Median :-13.00 Median : 4.50 Median : 27.00   
## Mean :-0.1005 Mean :-13.50 Mean : 18.35 Mean : -17.60   
## 3rd Qu.: 0.0000 3rd Qu.: -8.75 3rd Qu.: 25.50 3rd Qu.: 38.25   
## Max. : 0.0500 Max. : 46.00 Max. : 72.00 Max. : 49.00   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z roll\_arm   
## Min. :-13.00 Min. :566.0 Min. :-426.0 Min. :-137.00   
## 1st Qu.: 5.50 1st Qu.:578.5 1st Qu.:-398.5 1st Qu.: 0.00   
## Median : 33.50 Median :600.5 Median :-313.5 Median : 0.00   
## Mean : 35.15 Mean :601.5 Mean :-346.9 Mean : 16.42   
## 3rd Qu.: 46.25 3rd Qu.:631.2 3rd Qu.:-305.0 3rd Qu.: 71.53   
## Max. :169.00 Max. :638.0 Max. :-291.0 Max. : 152.00   
## pitch\_arm yaw\_arm total\_accel\_arm var\_accel\_arm   
## Min. :-63.800 Min. :-167.00 Min. : 3.00 Mode:logical   
## 1st Qu.: -9.188 1st Qu.: -60.15 1st Qu.:20.25 NA's:20   
## Median : 0.000 Median : 0.00 Median :29.50   
## Mean : -3.950 Mean : -2.80 Mean :26.40   
## 3rd Qu.: 3.465 3rd Qu.: 25.50 3rd Qu.:33.25   
## Max. : 55.000 Max. : 178.00 Max. :44.00   
## avg\_roll\_arm stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm stddev\_pitch\_arm  
## Mode:logical Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## var\_pitch\_arm avg\_yaw\_arm stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## Mode:logical Mode:logical Mode:logical Mode:logical Min. :-3.710   
## NA's:20 NA's:20 NA's:20 NA's:20 1st Qu.:-0.645   
## Median : 0.020   
## Mean : 0.077   
## 3rd Qu.: 1.248   
## Max. : 3.660   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x accel\_arm\_y   
## Min. :-2.0900 Min. :-0.6900 Min. :-341.0 Min. :-65.00   
## 1st Qu.:-0.6350 1st Qu.:-0.1800 1st Qu.:-277.0 1st Qu.: 52.25   
## Median :-0.0400 Median :-0.0250 Median :-194.5 Median :112.00   
## Mean :-0.1595 Mean : 0.1205 Mean :-134.6 Mean :103.10   
## 3rd Qu.: 0.2175 3rd Qu.: 0.5650 3rd Qu.: 5.5 3rd Qu.:168.25   
## Max. : 1.8500 Max. : 1.1300 Max. : 106.0 Max. :245.00   
## accel\_arm\_z magnet\_arm\_x magnet\_arm\_y magnet\_arm\_z   
## Min. :-404.00 Min. :-428.00 Min. :-307.0 Min. :-499.0   
## 1st Qu.:-128.50 1st Qu.:-373.75 1st Qu.: 205.2 1st Qu.: 403.0   
## Median : -83.50 Median :-265.00 Median : 291.0 Median : 476.5   
## Mean : -87.85 Mean : -38.95 Mean : 239.4 Mean : 369.8   
## 3rd Qu.: -27.25 3rd Qu.: 250.50 3rd Qu.: 358.8 3rd Qu.: 517.0   
## Max. : 93.00 Max. : 750.00 Max. : 474.0 Max. : 633.0   
## kurtosis\_roll\_arm kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm max\_picth\_arm   
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## max\_yaw\_arm min\_roll\_arm min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm  
## Mode:logical Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell pitch\_dumbbell   
## Mode:logical Mode:logical Min. :-111.118 Min. :-54.97   
## NA's:20 NA's:20 1st Qu.: 7.494 1st Qu.:-51.89   
## Median : 50.403 Median :-40.81   
## Mean : 33.760 Mean :-19.47   
## 3rd Qu.: 58.129 3rd Qu.: 16.12   
## Max. : 123.984 Max. : 96.87   
## yaw\_dumbbell kurtosis\_roll\_dumbbell kurtosis\_picth\_dumbbell  
## Min. :-103.3200 Mode:logical Mode:logical   
## 1st Qu.: -75.2809 NA's:20 NA's:20   
## Median : -8.2863   
## Mean : -0.9385   
## 3rd Qu.: 55.8335   
## Max. : 132.2337   
## kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell skewness\_pitch\_dumbbell  
## Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## skewness\_yaw\_dumbbell max\_roll\_dumbbell max\_picth\_dumbbell max\_yaw\_dumbbell  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## min\_roll\_dumbbell min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell  
## Mode:logical Mode:logical Min. : 1.0   
## NA's:20 NA's:20 1st Qu.: 7.0   
## Median :15.5   
## Mean :17.2   
## 3rd Qu.:29.0   
## Max. :31.0   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell var\_roll\_dumbbell  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## avg\_pitch\_dumbbell stddev\_pitch\_dumbbell var\_pitch\_dumbbell avg\_yaw\_dumbbell  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## stddev\_yaw\_dumbbell var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## Mode:logical Mode:logical Min. :-1.0300 Min. :-1.1100   
## NA's:20 NA's:20 1st Qu.: 0.1600 1st Qu.:-0.2100   
## Median : 0.3600 Median : 0.0150   
## Mean : 0.2690 Mean : 0.0605   
## 3rd Qu.: 0.4625 3rd Qu.: 0.1450   
## Max. : 1.0600 Max. : 1.9100   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y accel\_dumbbell\_z  
## Min. :-1.180 Min. :-159.00 Min. :-30.00 Min. :-221.0   
## 1st Qu.:-0.485 1st Qu.:-140.25 1st Qu.: 5.75 1st Qu.:-192.2   
## Median :-0.280 Median : -19.00 Median : 71.50 Median : -3.0   
## Mean :-0.266 Mean : -47.60 Mean : 70.55 Mean : -60.0   
## 3rd Qu.:-0.165 3rd Qu.: 15.75 3rd Qu.:151.25 3rd Qu.: 76.5   
## Max. : 1.100 Max. : 185.00 Max. :166.00 Max. : 100.0   
## magnet\_dumbbell\_x magnet\_dumbbell\_y magnet\_dumbbell\_z roll\_forearm   
## Min. :-576.0 Min. :-558.0 Min. :-164.00 Min. :-176.00   
## 1st Qu.:-528.0 1st Qu.: 259.5 1st Qu.: -33.00 1st Qu.: -40.25   
## Median :-508.5 Median : 316.0 Median : 49.50 Median : 94.20   
## Mean :-304.2 Mean : 189.3 Mean : 71.40 Mean : 38.66   
## 3rd Qu.:-317.0 3rd Qu.: 348.2 3rd Qu.: 96.25 3rd Qu.: 143.25   
## Max. : 523.0 Max. : 403.0 Max. : 368.00 Max. : 176.00   
## pitch\_forearm yaw\_forearm kurtosis\_roll\_forearm  
## Min. :-63.500 Min. :-168.000 Mode:logical   
## 1st Qu.:-11.457 1st Qu.: -93.375 NA's:20   
## Median : 8.830 Median : -19.250   
## Mean : 7.099 Mean : 2.195   
## 3rd Qu.: 28.500 3rd Qu.: 104.500   
## Max. : 59.300 Max. : 159.000   
## kurtosis\_picth\_forearm kurtosis\_yaw\_forearm skewness\_roll\_forearm  
## Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## skewness\_pitch\_forearm skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm min\_yaw\_forearm  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## amplitude\_roll\_forearm amplitude\_pitch\_forearm amplitude\_yaw\_forearm  
## Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## total\_accel\_forearm var\_accel\_forearm avg\_roll\_forearm stddev\_roll\_forearm  
## Min. :21.00 Mode:logical Mode:logical Mode:logical   
## 1st Qu.:24.00 NA's:20 NA's:20 NA's:20   
## Median :32.50   
## Mean :32.05   
## 3rd Qu.:36.75   
## Max. :47.00   
## var\_roll\_forearm avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm  
## Mode:logical Mode:logical Mode:logical Mode:logical   
## NA's:20 NA's:20 NA's:20 NA's:20   
##   
##   
##   
##   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm gyros\_forearm\_x   
## Mode:logical Mode:logical Mode:logical Min. :-1.0600   
## NA's:20 NA's:20 NA's:20 1st Qu.:-0.5850   
## Median : 0.0200   
## Mean :-0.0200   
## 3rd Qu.: 0.2925   
## Max. : 1.3800   
## gyros\_forearm\_y gyros\_forearm\_z accel\_forearm\_x accel\_forearm\_y   
## Min. :-5.9700 Min. :-1.2600 Min. :-212.0 Min. :-331.0   
## 1st Qu.:-1.2875 1st Qu.:-0.0975 1st Qu.:-114.8 1st Qu.: 8.5   
## Median : 0.0350 Median : 0.2300 Median : 86.0 Median : 138.0   
## Mean :-0.0415 Mean : 0.2610 Mean : 38.8 Mean : 125.3   
## 3rd Qu.: 2.0475 3rd Qu.: 0.7625 3rd Qu.: 166.2 3rd Qu.: 268.0   
## Max. : 4.2600 Max. : 1.8000 Max. : 232.0 Max. : 406.0   
## accel\_forearm\_z magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z  
## Min. :-282.0 Min. :-714.0 Min. :-787.0 Min. :-32.0   
## 1st Qu.:-199.0 1st Qu.:-427.2 1st Qu.:-328.8 1st Qu.:275.2   
## Median :-148.5 Median :-189.5 Median : 487.0 Median :491.5   
## Mean : -93.7 Mean :-159.2 Mean : 191.8 Mean :460.2   
## 3rd Qu.: -31.0 3rd Qu.: 41.5 3rd Qu.: 720.8 3rd Qu.:661.5   
## Max. : 179.0 Max. : 532.0 Max. : 800.0 Max. :884.0   
## problem\_id   
## Min. : 1.00   
## 1st Qu.: 5.75   
## Median :10.50   
## Mean :10.50   
## 3rd Qu.:15.25   
## Max. :20.00

select(df, classe)

## # A tibble: 19,622 x 1  
## classe  
## <chr>   
## 1 A   
## 2 A   
## 3 A   
## 4 A   
## 5 A   
## 6 A   
## 7 A   
## 8 A   
## 9 A   
## 10 A   
## # ... with 19,612 more rows

Although vizualization of the histgrams or distributions is better to be done here, I skip it because there are so many variables

Check the columns having missing values. These missing values should be avoided before running a machine learning model.

sapply(df, function(y) sum(is.na(y)))

## id user\_name raw\_timestamp\_part\_1   
## 0 0 0   
## raw\_timestamp\_part\_2 cvtd\_timestamp new\_window   
## 0 0 0   
## num\_window roll\_belt pitch\_belt   
## 0 0 0   
## yaw\_belt total\_accel\_belt kurtosis\_roll\_belt   
## 0 0 19216   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt   
## 19216 19216 19216   
## skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt   
## 19216 19216 19216   
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## 19216 19216 19216   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt   
## 19216 19216 19216   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt   
## 19216 19216 19216   
## avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt   
## 19216 19216 19216   
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt   
## 19216 19216 19216   
## avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## 19216 19216 19216   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z   
## 0 0 0   
## accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## 0 0 0   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z   
## 0 0 0   
## roll\_arm pitch\_arm yaw\_arm   
## 0 0 0   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm   
## 0 19216 19216   
## stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm   
## 19216 19216 19216   
## stddev\_pitch\_arm var\_pitch\_arm avg\_yaw\_arm   
## 19216 19216 19216   
## stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## 19216 19216 0   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x   
## 0 0 0   
## accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## 0 0 0   
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm   
## 0 0 19294   
## kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm   
## 19216 19216 19293   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm   
## 19216 19216 19216   
## max\_picth\_arm max\_yaw\_arm min\_roll\_arm   
## 19216 19216 19216   
## min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm   
## 19216 19216 19216   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## 19216 19216 0   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell   
## 0 0 19221   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell   
## 19218 19216 19220   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell   
## 19217 19216 19216   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell   
## 19216 19221 19216   
## min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell   
## 19216 19221 19216   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell   
## 19216 19221 0   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell   
## 19216 19216 19216   
## var\_roll\_dumbbell avg\_pitch\_dumbbell stddev\_pitch\_dumbbell   
## 19216 19216 19216   
## var\_pitch\_dumbbell avg\_yaw\_dumbbell stddev\_yaw\_dumbbell   
## 19216 19216 19216   
## var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## 19216 0 0   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y   
## 0 0 0   
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y   
## 0 0 0   
## magnet\_dumbbell\_z roll\_forearm pitch\_forearm   
## 0 0 0   
## yaw\_forearm kurtosis\_roll\_forearm kurtosis\_picth\_forearm   
## 0 19216 19216   
## kurtosis\_yaw\_forearm skewness\_roll\_forearm skewness\_pitch\_forearm   
## 19216 19216 19216   
## skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm   
## 19216 19216 19216   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm   
## 19216 19216 19216   
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm   
## 19216 19216 19216   
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm   
## 19216 0 19216   
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm   
## 19216 19216 19216   
## avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## 19216 19216 19216   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm   
## 19216 19216 19216   
## gyros\_forearm\_x gyros\_forearm\_y gyros\_forearm\_z   
## 0 0 0   
## accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z   
## 0 0 0   
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z   
## 0 0 0   
## classe   
## 0

sapply(df\_test, function(y) sum(is.na(y)))

## id user\_name raw\_timestamp\_part\_1   
## 0 0 0   
## raw\_timestamp\_part\_2 cvtd\_timestamp new\_window   
## 0 0 0   
## num\_window roll\_belt pitch\_belt   
## 0 0 0   
## yaw\_belt total\_accel\_belt kurtosis\_roll\_belt   
## 0 0 20   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt   
## 20 20 20   
## skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt   
## 20 20 20   
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## 20 20 20   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt   
## 20 20 20   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt   
## 20 20 20   
## avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt   
## 20 20 20   
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt   
## 20 20 20   
## avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## 20 20 20   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z   
## 0 0 0   
## accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## 0 0 0   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z   
## 0 0 0   
## roll\_arm pitch\_arm yaw\_arm   
## 0 0 0   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm   
## 0 20 20   
## stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm   
## 20 20 20   
## stddev\_pitch\_arm var\_pitch\_arm avg\_yaw\_arm   
## 20 20 20   
## stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## 20 20 0   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x   
## 0 0 0   
## accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## 0 0 0   
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm   
## 0 0 20   
## kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm   
## 20 20 20   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm   
## 20 20 20   
## max\_picth\_arm max\_yaw\_arm min\_roll\_arm   
## 20 20 20   
## min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm   
## 20 20 20   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## 20 20 0   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell   
## 0 0 20   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell   
## 20 20 20   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell   
## 20 20 20   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell   
## 20 20 20   
## min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell   
## 20 20 20   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell   
## 20 20 0   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell   
## 20 20 20   
## var\_roll\_dumbbell avg\_pitch\_dumbbell stddev\_pitch\_dumbbell   
## 20 20 20   
## var\_pitch\_dumbbell avg\_yaw\_dumbbell stddev\_yaw\_dumbbell   
## 20 20 20   
## var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## 20 0 0   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y   
## 0 0 0   
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y   
## 0 0 0   
## magnet\_dumbbell\_z roll\_forearm pitch\_forearm   
## 0 0 0   
## yaw\_forearm kurtosis\_roll\_forearm kurtosis\_picth\_forearm   
## 0 20 20   
## kurtosis\_yaw\_forearm skewness\_roll\_forearm skewness\_pitch\_forearm   
## 20 20 20   
## skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm   
## 20 20 20   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm   
## 20 20 20   
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm   
## 20 20 20   
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm   
## 20 0 20   
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm   
## 20 20 20   
## avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## 20 20 20   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm   
## 20 20 20   
## gyros\_forearm\_x gyros\_forearm\_y gyros\_forearm\_z   
## 0 0 0   
## accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z   
## 0 0 0   
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z   
## 0 0 0   
## problem\_id   
## 0

## Preprocessing

Next step is to delete the columns having missing values. In the below process, I transform the data for finding missing values and unlist them to create the new dataframe df and df\_test.

df %>% lapply(.,anyNA) %>% unlist %>% !.

## id user\_name raw\_timestamp\_part\_1   
## TRUE TRUE TRUE   
## raw\_timestamp\_part\_2 cvtd\_timestamp new\_window   
## TRUE TRUE TRUE   
## num\_window roll\_belt pitch\_belt   
## TRUE TRUE TRUE   
## yaw\_belt total\_accel\_belt kurtosis\_roll\_belt   
## TRUE TRUE FALSE   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt   
## FALSE FALSE FALSE   
## skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt   
## FALSE FALSE FALSE   
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## FALSE FALSE FALSE   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt   
## FALSE FALSE FALSE   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt   
## FALSE FALSE FALSE   
## avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt   
## FALSE FALSE FALSE   
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt   
## FALSE FALSE FALSE   
## avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## FALSE FALSE FALSE   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z   
## TRUE TRUE TRUE   
## accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## TRUE TRUE TRUE   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z   
## TRUE TRUE TRUE   
## roll\_arm pitch\_arm yaw\_arm   
## TRUE TRUE TRUE   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm   
## TRUE FALSE FALSE   
## stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm   
## FALSE FALSE FALSE   
## stddev\_pitch\_arm var\_pitch\_arm avg\_yaw\_arm   
## FALSE FALSE FALSE   
## stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## FALSE FALSE TRUE   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x   
## TRUE TRUE TRUE   
## accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## TRUE TRUE TRUE   
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm   
## TRUE TRUE FALSE   
## kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm   
## FALSE FALSE FALSE   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm   
## FALSE FALSE FALSE   
## max\_picth\_arm max\_yaw\_arm min\_roll\_arm   
## FALSE FALSE FALSE   
## min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm   
## FALSE FALSE FALSE   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## FALSE FALSE TRUE   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell   
## TRUE TRUE FALSE   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell   
## FALSE FALSE FALSE   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell   
## FALSE FALSE FALSE   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell   
## FALSE FALSE FALSE   
## min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell   
## FALSE FALSE FALSE   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell   
## FALSE FALSE TRUE   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell   
## FALSE FALSE FALSE   
## var\_roll\_dumbbell avg\_pitch\_dumbbell stddev\_pitch\_dumbbell   
## FALSE FALSE FALSE   
## var\_pitch\_dumbbell avg\_yaw\_dumbbell stddev\_yaw\_dumbbell   
## FALSE FALSE FALSE   
## var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## FALSE TRUE TRUE   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y   
## TRUE TRUE TRUE   
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y   
## TRUE TRUE TRUE   
## magnet\_dumbbell\_z roll\_forearm pitch\_forearm   
## TRUE TRUE TRUE   
## yaw\_forearm kurtosis\_roll\_forearm kurtosis\_picth\_forearm   
## TRUE FALSE FALSE   
## kurtosis\_yaw\_forearm skewness\_roll\_forearm skewness\_pitch\_forearm   
## FALSE FALSE FALSE   
## skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm   
## FALSE FALSE FALSE   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm   
## FALSE FALSE FALSE   
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm   
## FALSE FALSE FALSE   
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm   
## FALSE TRUE FALSE   
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm   
## FALSE FALSE FALSE   
## avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## FALSE FALSE FALSE   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm   
## FALSE FALSE FALSE   
## gyros\_forearm\_x gyros\_forearm\_y gyros\_forearm\_z   
## TRUE TRUE TRUE   
## accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z   
## TRUE TRUE TRUE   
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z   
## TRUE TRUE TRUE   
## classe   
## TRUE

df <- df %>% select\_if(lapply(.,anyNA) %>% unlist %>% !.)  
dim(df)

## [1] 19622 60

df\_test %>% lapply(.,anyNA) %>% unlist %>% !.

## id user\_name raw\_timestamp\_part\_1   
## TRUE TRUE TRUE   
## raw\_timestamp\_part\_2 cvtd\_timestamp new\_window   
## TRUE TRUE TRUE   
## num\_window roll\_belt pitch\_belt   
## TRUE TRUE TRUE   
## yaw\_belt total\_accel\_belt kurtosis\_roll\_belt   
## TRUE TRUE FALSE   
## kurtosis\_picth\_belt kurtosis\_yaw\_belt skewness\_roll\_belt   
## FALSE FALSE FALSE   
## skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt   
## FALSE FALSE FALSE   
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt   
## FALSE FALSE FALSE   
## min\_pitch\_belt min\_yaw\_belt amplitude\_roll\_belt   
## FALSE FALSE FALSE   
## amplitude\_pitch\_belt amplitude\_yaw\_belt var\_total\_accel\_belt   
## FALSE FALSE FALSE   
## avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt   
## FALSE FALSE FALSE   
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt   
## FALSE FALSE FALSE   
## avg\_yaw\_belt stddev\_yaw\_belt var\_yaw\_belt   
## FALSE FALSE FALSE   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z   
## TRUE TRUE TRUE   
## accel\_belt\_x accel\_belt\_y accel\_belt\_z   
## TRUE TRUE TRUE   
## magnet\_belt\_x magnet\_belt\_y magnet\_belt\_z   
## TRUE TRUE TRUE   
## roll\_arm pitch\_arm yaw\_arm   
## TRUE TRUE TRUE   
## total\_accel\_arm var\_accel\_arm avg\_roll\_arm   
## TRUE FALSE FALSE   
## stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm   
## FALSE FALSE FALSE   
## stddev\_pitch\_arm var\_pitch\_arm avg\_yaw\_arm   
## FALSE FALSE FALSE   
## stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x   
## FALSE FALSE TRUE   
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x   
## TRUE TRUE TRUE   
## accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## TRUE TRUE TRUE   
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm   
## TRUE TRUE FALSE   
## kurtosis\_picth\_arm kurtosis\_yaw\_arm skewness\_roll\_arm   
## FALSE FALSE FALSE   
## skewness\_pitch\_arm skewness\_yaw\_arm max\_roll\_arm   
## FALSE FALSE FALSE   
## max\_picth\_arm max\_yaw\_arm min\_roll\_arm   
## FALSE FALSE FALSE   
## min\_pitch\_arm min\_yaw\_arm amplitude\_roll\_arm   
## FALSE FALSE FALSE   
## amplitude\_pitch\_arm amplitude\_yaw\_arm roll\_dumbbell   
## FALSE FALSE TRUE   
## pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell   
## TRUE TRUE FALSE   
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell   
## FALSE FALSE FALSE   
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell   
## FALSE FALSE FALSE   
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell   
## FALSE FALSE FALSE   
## min\_pitch\_dumbbell min\_yaw\_dumbbell amplitude\_roll\_dumbbell   
## FALSE FALSE FALSE   
## amplitude\_pitch\_dumbbell amplitude\_yaw\_dumbbell total\_accel\_dumbbell   
## FALSE FALSE TRUE   
## var\_accel\_dumbbell avg\_roll\_dumbbell stddev\_roll\_dumbbell   
## FALSE FALSE FALSE   
## var\_roll\_dumbbell avg\_pitch\_dumbbell stddev\_pitch\_dumbbell   
## FALSE FALSE FALSE   
## var\_pitch\_dumbbell avg\_yaw\_dumbbell stddev\_yaw\_dumbbell   
## FALSE FALSE FALSE   
## var\_yaw\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## FALSE TRUE TRUE   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y   
## TRUE TRUE TRUE   
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y   
## TRUE TRUE TRUE   
## magnet\_dumbbell\_z roll\_forearm pitch\_forearm   
## TRUE TRUE TRUE   
## yaw\_forearm kurtosis\_roll\_forearm kurtosis\_picth\_forearm   
## TRUE FALSE FALSE   
## kurtosis\_yaw\_forearm skewness\_roll\_forearm skewness\_pitch\_forearm   
## FALSE FALSE FALSE   
## skewness\_yaw\_forearm max\_roll\_forearm max\_picth\_forearm   
## FALSE FALSE FALSE   
## max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm   
## FALSE FALSE FALSE   
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm   
## FALSE FALSE FALSE   
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm   
## FALSE TRUE FALSE   
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm   
## FALSE FALSE FALSE   
## avg\_pitch\_forearm stddev\_pitch\_forearm var\_pitch\_forearm   
## FALSE FALSE FALSE   
## avg\_yaw\_forearm stddev\_yaw\_forearm var\_yaw\_forearm   
## FALSE FALSE FALSE   
## gyros\_forearm\_x gyros\_forearm\_y gyros\_forearm\_z   
## TRUE TRUE TRUE   
## accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z   
## TRUE TRUE TRUE   
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z   
## TRUE TRUE TRUE   
## problem\_id   
## TRUE

df\_test <- df\_test %>% select\_if(lapply(.,anyNA) %>% unlist %>% !.)  
dim(df\_test)

## [1] 20 60

After deleting the columns having missing values, only 60 explanatory variales are still available! Let’s use them to build a prediction model.

Here, again check the columns I will use to create a model.

colnames(df)

## [1] "id" "user\_name" "raw\_timestamp\_part\_1"  
## [4] "raw\_timestamp\_part\_2" "cvtd\_timestamp" "new\_window"   
## [7] "num\_window" "roll\_belt" "pitch\_belt"   
## [10] "yaw\_belt" "total\_accel\_belt" "gyros\_belt\_x"   
## [13] "gyros\_belt\_y" "gyros\_belt\_z" "accel\_belt\_x"   
## [16] "accel\_belt\_y" "accel\_belt\_z" "magnet\_belt\_x"   
## [19] "magnet\_belt\_y" "magnet\_belt\_z" "roll\_arm"   
## [22] "pitch\_arm" "yaw\_arm" "total\_accel\_arm"   
## [25] "gyros\_arm\_x" "gyros\_arm\_y" "gyros\_arm\_z"   
## [28] "accel\_arm\_x" "accel\_arm\_y" "accel\_arm\_z"   
## [31] "magnet\_arm\_x" "magnet\_arm\_y" "magnet\_arm\_z"   
## [34] "roll\_dumbbell" "pitch\_dumbbell" "yaw\_dumbbell"   
## [37] "total\_accel\_dumbbell" "gyros\_dumbbell\_x" "gyros\_dumbbell\_y"   
## [40] "gyros\_dumbbell\_z" "accel\_dumbbell\_x" "accel\_dumbbell\_y"   
## [43] "accel\_dumbbell\_z" "magnet\_dumbbell\_x" "magnet\_dumbbell\_y"   
## [46] "magnet\_dumbbell\_z" "roll\_forearm" "pitch\_forearm"   
## [49] "yaw\_forearm" "total\_accel\_forearm" "gyros\_forearm\_x"   
## [52] "gyros\_forearm\_y" "gyros\_forearm\_z" "accel\_forearm\_x"   
## [55] "accel\_forearm\_y" "accel\_forearm\_z" "magnet\_forearm\_x"   
## [58] "magnet\_forearm\_y" "magnet\_forearm\_z" "classe"

colnames(df\_test)

## [1] "id" "user\_name" "raw\_timestamp\_part\_1"  
## [4] "raw\_timestamp\_part\_2" "cvtd\_timestamp" "new\_window"   
## [7] "num\_window" "roll\_belt" "pitch\_belt"   
## [10] "yaw\_belt" "total\_accel\_belt" "gyros\_belt\_x"   
## [13] "gyros\_belt\_y" "gyros\_belt\_z" "accel\_belt\_x"   
## [16] "accel\_belt\_y" "accel\_belt\_z" "magnet\_belt\_x"   
## [19] "magnet\_belt\_y" "magnet\_belt\_z" "roll\_arm"   
## [22] "pitch\_arm" "yaw\_arm" "total\_accel\_arm"   
## [25] "gyros\_arm\_x" "gyros\_arm\_y" "gyros\_arm\_z"   
## [28] "accel\_arm\_x" "accel\_arm\_y" "accel\_arm\_z"   
## [31] "magnet\_arm\_x" "magnet\_arm\_y" "magnet\_arm\_z"   
## [34] "roll\_dumbbell" "pitch\_dumbbell" "yaw\_dumbbell"   
## [37] "total\_accel\_dumbbell" "gyros\_dumbbell\_x" "gyros\_dumbbell\_y"   
## [40] "gyros\_dumbbell\_z" "accel\_dumbbell\_x" "accel\_dumbbell\_y"   
## [43] "accel\_dumbbell\_z" "magnet\_dumbbell\_x" "magnet\_dumbbell\_y"   
## [46] "magnet\_dumbbell\_z" "roll\_forearm" "pitch\_forearm"   
## [49] "yaw\_forearm" "total\_accel\_forearm" "gyros\_forearm\_x"   
## [52] "gyros\_forearm\_y" "gyros\_forearm\_z" "accel\_forearm\_x"   
## [55] "accel\_forearm\_y" "accel\_forearm\_z" "magnet\_forearm\_x"   
## [58] "magnet\_forearm\_y" "magnet\_forearm\_z" "problem\_id"

Here, let’s set an index in the dataset. the most left column id is better to be used as an index.

rownames(df) <- df$id  
df$id <- NULL  
  
rownames(df\_test) <- df\_test$id  
df\_test$id <- NULL

Delete the columns user\_name and cvtd\_timestamp because they look unrelated to my objective in the prediction.

df = subset(df, select = -c(user\_name,cvtd\_timestamp))  
df\_test = subset(df\_test, select = -c(user\_name, cvtd\_timestamp, problem\_id))  
dim(df)

## [1] 19622 57

dim(df\_test)

## [1] 20 56

The reason why the number of columns in df is more than that of df\_test is that the dataset df\_test does not have the column classe which is the response variable in this analysis.

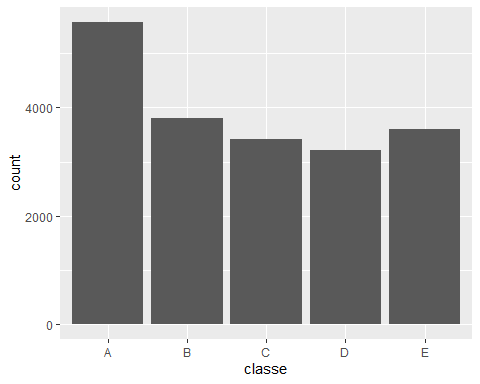
To make a prediction model, I need to convert character values into factor values. In R language, factor columns are used for classification problems while numeric columns are used for regression problems.

df$new\_window <- as.factor(df$new\_window)  
df\_test$new\_window <- as.factor(df\_test$new\_window)  
  
df$classe <- as.factor(df$classe)  
  
# Adjust the level of df\_test to that of df  
levels(df\_test$new\_window) <- levels(df$new\_window)

summary(df\_test)

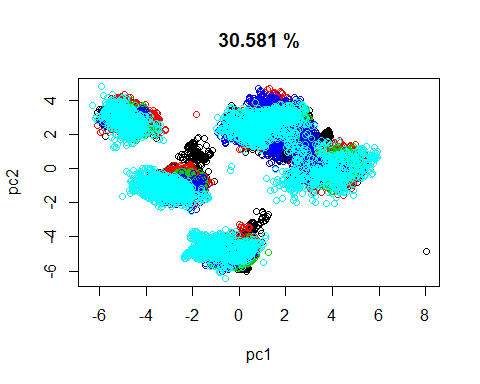
## raw\_timestamp\_part\_1 raw\_timestamp\_part\_2 new\_window num\_window   
## Min. :1.322e+09 Min. : 36553 no :20 Min. : 48.0   
## 1st Qu.:1.323e+09 1st Qu.:268655 yes: 0 1st Qu.:250.0   
## Median :1.323e+09 Median :530706 Median :384.5   
## Mean :1.323e+09 Mean :512167 Mean :379.6   
## 3rd Qu.:1.323e+09 3rd Qu.:787738 3rd Qu.:467.0   
## Max. :1.323e+09 Max. :920315 Max. :859.0   
## roll\_belt pitch\_belt yaw\_belt total\_accel\_belt  
## Min. : -5.9200 Min. :-41.600 Min. :-93.70 Min. : 2.00   
## 1st Qu.: 0.9075 1st Qu.: 3.013 1st Qu.:-88.62 1st Qu.: 3.00   
## Median : 1.1100 Median : 4.655 Median :-87.85 Median : 4.00   
## Mean : 31.3055 Mean : 5.824 Mean :-59.30 Mean : 7.55   
## 3rd Qu.: 32.5050 3rd Qu.: 6.135 3rd Qu.:-63.50 3rd Qu.: 8.00   
## Max. :129.0000 Max. : 27.800 Max. :162.00 Max. :21.00   
## gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z accel\_belt\_x   
## Min. :-0.500 Min. :-0.050 Min. :-0.4800 Min. :-48.00   
## 1st Qu.:-0.070 1st Qu.:-0.005 1st Qu.:-0.1375 1st Qu.:-19.00   
## Median : 0.020 Median : 0.000 Median :-0.0250 Median :-13.00   
## Mean :-0.045 Mean : 0.010 Mean :-0.1005 Mean :-13.50   
## 3rd Qu.: 0.070 3rd Qu.: 0.020 3rd Qu.: 0.0000 3rd Qu.: -8.75   
## Max. : 0.240 Max. : 0.110 Max. : 0.0500 Max. : 46.00   
## accel\_belt\_y accel\_belt\_z magnet\_belt\_x magnet\_belt\_y   
## Min. :-16.00 Min. :-187.00 Min. :-13.00 Min. :566.0   
## 1st Qu.: 2.00 1st Qu.: -24.00 1st Qu.: 5.50 1st Qu.:578.5   
## Median : 4.50 Median : 27.00 Median : 33.50 Median :600.5   
## Mean : 18.35 Mean : -17.60 Mean : 35.15 Mean :601.5   
## 3rd Qu.: 25.50 3rd Qu.: 38.25 3rd Qu.: 46.25 3rd Qu.:631.2   
## Max. : 72.00 Max. : 49.00 Max. :169.00 Max. :638.0   
## magnet\_belt\_z roll\_arm pitch\_arm yaw\_arm   
## Min. :-426.0 Min. :-137.00 Min. :-63.800 Min. :-167.00   
## 1st Qu.:-398.5 1st Qu.: 0.00 1st Qu.: -9.188 1st Qu.: -60.15   
## Median :-313.5 Median : 0.00 Median : 0.000 Median : 0.00   
## Mean :-346.9 Mean : 16.42 Mean : -3.950 Mean : -2.80   
## 3rd Qu.:-305.0 3rd Qu.: 71.53 3rd Qu.: 3.465 3rd Qu.: 25.50   
## Max. :-291.0 Max. : 152.00 Max. : 55.000 Max. : 178.00   
## total\_accel\_arm gyros\_arm\_x gyros\_arm\_y gyros\_arm\_z   
## Min. : 3.00 Min. :-3.710 Min. :-2.0900 Min. :-0.6900   
## 1st Qu.:20.25 1st Qu.:-0.645 1st Qu.:-0.6350 1st Qu.:-0.1800   
## Median :29.50 Median : 0.020 Median :-0.0400 Median :-0.0250   
## Mean :26.40 Mean : 0.077 Mean :-0.1595 Mean : 0.1205   
## 3rd Qu.:33.25 3rd Qu.: 1.248 3rd Qu.: 0.2175 3rd Qu.: 0.5650   
## Max. :44.00 Max. : 3.660 Max. : 1.8500 Max. : 1.1300   
## accel\_arm\_x accel\_arm\_y accel\_arm\_z magnet\_arm\_x   
## Min. :-341.0 Min. :-65.00 Min. :-404.00 Min. :-428.00   
## 1st Qu.:-277.0 1st Qu.: 52.25 1st Qu.:-128.50 1st Qu.:-373.75   
## Median :-194.5 Median :112.00 Median : -83.50 Median :-265.00   
## Mean :-134.6 Mean :103.10 Mean : -87.85 Mean : -38.95   
## 3rd Qu.: 5.5 3rd Qu.:168.25 3rd Qu.: -27.25 3rd Qu.: 250.50   
## Max. : 106.0 Max. :245.00 Max. : 93.00 Max. : 750.00   
## magnet\_arm\_y magnet\_arm\_z roll\_dumbbell pitch\_dumbbell   
## Min. :-307.0 Min. :-499.0 Min. :-111.118 Min. :-54.97   
## 1st Qu.: 205.2 1st Qu.: 403.0 1st Qu.: 7.494 1st Qu.:-51.89   
## Median : 291.0 Median : 476.5 Median : 50.403 Median :-40.81   
## Mean : 239.4 Mean : 369.8 Mean : 33.760 Mean :-19.47   
## 3rd Qu.: 358.8 3rd Qu.: 517.0 3rd Qu.: 58.129 3rd Qu.: 16.12   
## Max. : 474.0 Max. : 633.0 Max. : 123.984 Max. : 96.87   
## yaw\_dumbbell total\_accel\_dumbbell gyros\_dumbbell\_x gyros\_dumbbell\_y   
## Min. :-103.3200 Min. : 1.0 Min. :-1.0300 Min. :-1.1100   
## 1st Qu.: -75.2809 1st Qu.: 7.0 1st Qu.: 0.1600 1st Qu.:-0.2100   
## Median : -8.2863 Median :15.5 Median : 0.3600 Median : 0.0150   
## Mean : -0.9385 Mean :17.2 Mean : 0.2690 Mean : 0.0605   
## 3rd Qu.: 55.8335 3rd Qu.:29.0 3rd Qu.: 0.4625 3rd Qu.: 0.1450   
## Max. : 132.2337 Max. :31.0 Max. : 1.0600 Max. : 1.9100   
## gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y accel\_dumbbell\_z  
## Min. :-1.180 Min. :-159.00 Min. :-30.00 Min. :-221.0   
## 1st Qu.:-0.485 1st Qu.:-140.25 1st Qu.: 5.75 1st Qu.:-192.2   
## Median :-0.280 Median : -19.00 Median : 71.50 Median : -3.0   
## Mean :-0.266 Mean : -47.60 Mean : 70.55 Mean : -60.0   
## 3rd Qu.:-0.165 3rd Qu.: 15.75 3rd Qu.:151.25 3rd Qu.: 76.5   
## Max. : 1.100 Max. : 185.00 Max. :166.00 Max. : 100.0   
## magnet\_dumbbell\_x magnet\_dumbbell\_y magnet\_dumbbell\_z roll\_forearm   
## Min. :-576.0 Min. :-558.0 Min. :-164.00 Min. :-176.00   
## 1st Qu.:-528.0 1st Qu.: 259.5 1st Qu.: -33.00 1st Qu.: -40.25   
## Median :-508.5 Median : 316.0 Median : 49.50 Median : 94.20   
## Mean :-304.2 Mean : 189.3 Mean : 71.40 Mean : 38.66   
## 3rd Qu.:-317.0 3rd Qu.: 348.2 3rd Qu.: 96.25 3rd Qu.: 143.25   
## Max. : 523.0 Max. : 403.0 Max. : 368.00 Max. : 176.00   
## pitch\_forearm yaw\_forearm total\_accel\_forearm gyros\_forearm\_x   
## Min. :-63.500 Min. :-168.000 Min. :21.00 Min. :-1.0600   
## 1st Qu.:-11.457 1st Qu.: -93.375 1st Qu.:24.00 1st Qu.:-0.5850   
## Median : 8.830 Median : -19.250 Median :32.50 Median : 0.0200   
## Mean : 7.099 Mean : 2.195 Mean :32.05 Mean :-0.0200   
## 3rd Qu.: 28.500 3rd Qu.: 104.500 3rd Qu.:36.75 3rd Qu.: 0.2925   
## Max. : 59.300 Max. : 159.000 Max. :47.00 Max. : 1.3800   
## gyros\_forearm\_y gyros\_forearm\_z accel\_forearm\_x accel\_forearm\_y   
## Min. :-5.9700 Min. :-1.2600 Min. :-212.0 Min. :-331.0   
## 1st Qu.:-1.2875 1st Qu.:-0.0975 1st Qu.:-114.8 1st Qu.: 8.5   
## Median : 0.0350 Median : 0.2300 Median : 86.0 Median : 138.0   
## Mean :-0.0415 Mean : 0.2610 Mean : 38.8 Mean : 125.3   
## 3rd Qu.: 2.0475 3rd Qu.: 0.7625 3rd Qu.: 166.2 3rd Qu.: 268.0   
## Max. : 4.2600 Max. : 1.8000 Max. : 232.0 Max. : 406.0   
## accel\_forearm\_z magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z  
## Min. :-282.0 Min. :-714.0 Min. :-787.0 Min. :-32.0   
## 1st Qu.:-199.0 1st Qu.:-427.2 1st Qu.:-328.8 1st Qu.:275.2   
## Median :-148.5 Median :-189.5 Median : 487.0 Median :491.5   
## Mean : -93.7 Mean :-159.2 Mean : 191.8 Mean :460.2   
## 3rd Qu.: -31.0 3rd Qu.: 41.5 3rd Qu.: 720.8 3rd Qu.:661.5   
## Max. : 179.0 Max. : 532.0 Max. : 800.0 Max. :884.0

ggplot(df, aes(classe)) + geom\_bar()



## PCA

numerical\_df <- df[, -which (colnames(df) %in% c("new\_window", "classe"))]  
prcomp.obj <- prcomp(numerical\_df, scale=TRUE) # PCA  
  
pc1 <- prcomp.obj$x[,1] # PC1  
pc2 <- prcomp.obj$x[,2] # PC2  
  
label <- as.factor(df$classe) # Class lavel  
percent <- summary(prcomp.obj)$importance[3,2] \* 100 # Cuml  
  
plot(pc1, pc2, col = label, main = paste(percent, "%"))



Finally, the 2 principle components only explain 30% of the result and it does not look effective in this context.

## Split the data

Finally, let’s prepare for building a machine learning model!

set.seed(0)  
inTrain <- createDataPartition(y=df$classe, p=0.8, list=FALSE)  
training <- df[inTrain,]  
testing <- df[-inTrain,]  
  
dim(training)

## [1] 15699 57

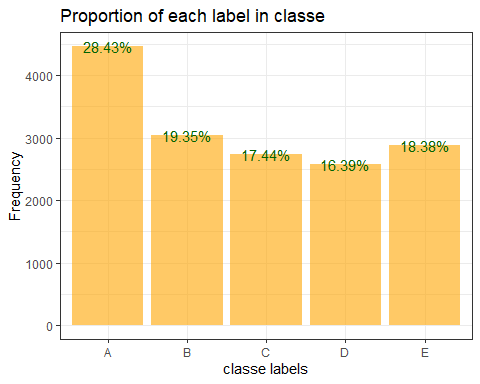
dim(testing)

## [1] 3923 57

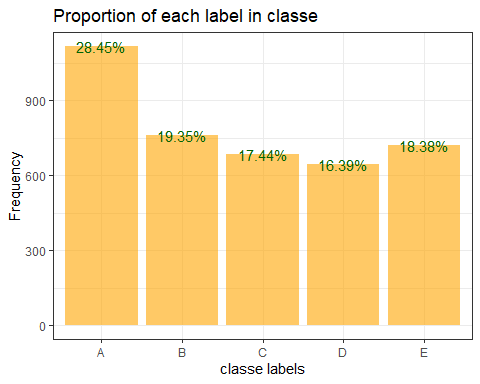
Although this point is confusing, I split the “train” dataset into “training” and “testing”. This “testing” is not the same as “df\_test” dataset on the above. This split is for measuring the performance of the model I create.

Let’s check the ditribution of classe in training is nearly equal to that in testing.

# Training  
ggplot(training, aes(classe)) +   
 geom\_bar(aes(y = (..count..)),fill="orange", alpha=0.6) +   
 geom\_text(aes(y = (..count..),label = ifelse((..count..)==0,"",scales::percent((..count..)/sum(..count..)))), stat="count",colour="darkgreen") +   
 theme\_bw() +  
 xlab("classe labels") +  
 ylab("Frequency") +   
 #scale\_x\_discrete(labels=c("yes"="NCR", "no"="Outside NCR")) +   
 ggtitle("Proportion of each label in classe")



# testing  
ggplot(testing, aes(classe)) +   
 geom\_bar(aes(y = (..count..)),fill="orange", alpha=0.6) +   
 geom\_text(aes(y = (..count..),label = ifelse((..count..)==0,"",scales::percent((..count..)/sum(..count..)))), stat="count",colour="darkgreen") +   
 theme\_bw() +  
 xlab("classe labels") +  
 ylab("Frequency") +   
 #scale\_x\_discrete(labels=c("yes"="NCR", "no"="Outside NCR")) +   
 ggtitle("Proportion of each label in classe")



They look almost similar, so no problem in further analysis.

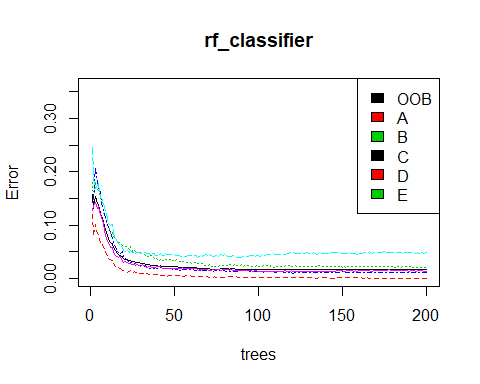
Create a base model. Before parameter tuning or something, let’s see the result as soon as possible. If the accuracy is enough to achieve my objective, that is enough!

## Base model

# Define the control  
trControl <- trainControl(method = "cv", number = 10, search = "grid")  
  
# RF  
# Perform training:  
set.seed(123)  
rf\_classifier = randomForest(training$classe ~ ., data=training, ntree=200, mtry=3, importance=TRUE, maxnodes = 500, trControl=trControl, metrics="accuracy")  
rf\_classifier

##   
## Call:  
## randomForest(formula = training$classe ~ ., data = training, ntree = 200, mtry = 3, importance = TRUE, maxnodes = 500, trControl = trControl, metrics = "accuracy")   
## Type of random forest: classification  
## Number of trees: 200  
## No. of variables tried at each split: 3  
##   
## OOB estimate of error rate: 1.64%  
## Confusion matrix:  
## A B C D E class.error  
## A 4462 1 0 1 0 0.0004480287  
## B 30 2974 34 0 0 0.0210664911  
## C 0 27 2709 2 0 0.0105916728  
## D 0 0 122 2450 1 0.0478041197  
## E 0 0 13 27 2846 0.0138600139

plot(rf\_classifier, ylim=c(0,0.36))  
legend('topright', colnames(rf\_classifier$err.rate), col=1:3, fill=1:3)



## XGB  
#gbm\_classifier <- train(training$classe ~ ., method="gbm", data=training, verbose=FALSE)  
#gbm\_classifier  
#plot(gbm\_classifier, ylim=c(0,0.36))  
#legend('topright', colnames(gbm\_classifier$err.rate), col=1:3, fill=1:3)

I just set up ntree, mtry, maxnodes, trControl, metrics here. They determine how much deep the trees become, how many counts the experiment iterates, and how we evaluate the model as being “good”. I did cv 10 times as default setting. The parameter should be determined by grid search, but the model here seems enough to analyze the data I deal with in this experiment because the performance is very nice as I will explain later.

## Predict

Because the base model seems enough fitted to the training data, let’s see the accuracy when applicating it to the testing data.

## Accuracy measures

# confusion matrix  
table(pred, testing$classe)

##   
## pred A B C D E  
## A 1116 8 0 0 0  
## B 0 741 6 0 0  
## C 0 10 678 33 5  
## D 0 0 0 610 5  
## E 0 0 0 0 711

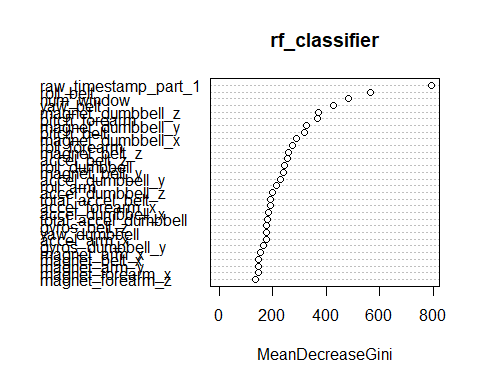
# Accuracy rate  
correctAns <- 0  
for ( i in 1:nrow( table( pred, testing$classe )))  
correctAns <- correctAns + table(pred, testing$classe)[i,i]  
correctAns / nrow( testing )

## [1] 0.9829212

In the testing data, the model achieved its high performance; accuracy is98.3% !

## Features Importance

varImpPlot(rf\_classifier,type=2)



The variables raw\_timestamp\_part\_1, roll-bell, and num\_window are important according to the above visualization.

## Predict the given test data

Finally, let’s apply my accurate model to prediction of the assignment task!

pred\_test = predict(rf\_classifier, newdata=df\_test)  
pred\_test

## 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20   
## B A B A A E D B A A B C B A E E A B B B   
## Levels: A B C D E

I got a full score in the final quiz given on Coursera platform by answering the above results!! C’est parfait!

# Financial Application

## The objective of this experiment

In this dataset, the objective is to predict the insurance charges by new clients when deciding on whether to admit applicants as new clients.

## Check the data

Set working directory and import the assignment datasets I use here.

Check the shapes of each dataset.

dim(data)

## [1] 1338 7

The train dataset is composed of 1338 data points, 7 variables.

See the above and below 5 rows of each dataset.

head(data, 5)

## # A tibble: 5 x 7  
## age sex bmi children smoker region charges  
## <dbl> <chr> <dbl> <dbl> <chr> <chr> <dbl>  
## 1 19 female 27.9 0 yes southwest 16885.  
## 2 18 male 33.8 1 no southeast 1726.  
## 3 28 male 33 3 no southeast 4449.  
## 4 33 male 22.7 0 no northwest 21984.  
## 5 32 male 28.9 0 no northwest 3867.

tail(data, 5)

## # A tibble: 5 x 7  
## age sex bmi children smoker region charges  
## <dbl> <chr> <dbl> <dbl> <chr> <chr> <dbl>  
## 1 50 male 31.0 3 no northwest 10601.  
## 2 18 female 31.9 0 no northeast 2206.  
## 3 18 female 36.8 0 no southeast 1630.  
## 4 21 female 25.8 0 no southwest 2008.  
## 5 61 female 29.1 0 yes northwest 29141.

See the column names.

colnames(data)

## [1] "age" "sex" "bmi" "children" "smoker" "region" "charges"

We can see the 6 explanatory variables and one response variable.

Here, see the data types.

sapply(data, class)

## age sex bmi children smoker region   
## "numeric" "character" "numeric" "numeric" "character" "character"   
## charges   
## "numeric"

Fortunately, almost all of the variables are numeric. But there are a few columns having character values. I need to change them into numeric or factor values before running a machine learning model.

As to columns having character values, here check the exact values.

head(Filter(is.character, data), 5)

## # A tibble: 5 x 3  
## sex smoker region   
## <chr> <chr> <chr>   
## 1 female yes southwest  
## 2 male no southeast  
## 3 male no southeast  
## 4 male no northwest  
## 5 male no northwest

unique(data$sex)

## [1] "female" "male"

unique(data$smoker)

## [1] "yes" "no"

unique(data$region)

## [1] "southwest" "southeast" "northwest" "northeast"

Check the columns having missing values. These missing values should be avoided before running a machine learning model.

sapply(data, function(y) sum(is.na(y)))

## age sex bmi children smoker region charges   
## 0 0 0 0 0 0 0

There is no missing value in this dataset.

Nest, see the descriptive statistics of the dataset. Generally, it is tough to follow all of statistical values, but this is an important step to grasp the data.

summary(data)

## age sex bmi children   
## Min. :18.00 Length:1338 Min. :15.96 Min. :0.000   
## 1st Qu.:27.00 Class :character 1st Qu.:26.30 1st Qu.:0.000   
## Median :39.00 Mode :character Median :30.40 Median :1.000   
## Mean :39.21 Mean :30.66 Mean :1.095   
## 3rd Qu.:51.00 3rd Qu.:34.69 3rd Qu.:2.000   
## Max. :64.00 Max. :53.13 Max. :5.000   
## smoker region charges   
## Length:1338 Length:1338 Min. : 1122   
## Class :character Class :character 1st Qu.: 4740   
## Mode :character Mode :character Median : 9382   
## Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

## Preprocessing

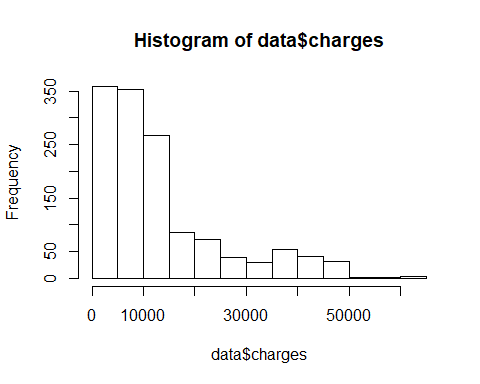
To make a prediction model, I need to convert character values into factor values. In R language, factor columns are used for classification problems while numeric columns are used for regression problems.

data$sex <- as.factor(data$sex)  
data$smoker <- as.factor(data$smoker)  
data$region <- as.factor(data$region)

summary(data)

## age sex bmi children smoker   
## Min. :18.00 female:662 Min. :15.96 Min. :0.000 no :1064   
## 1st Qu.:27.00 male :676 1st Qu.:26.30 1st Qu.:0.000 yes: 274   
## Median :39.00 Median :30.40 Median :1.000   
## Mean :39.21 Mean :30.66 Mean :1.095   
## 3rd Qu.:51.00 3rd Qu.:34.69 3rd Qu.:2.000   
## Max. :64.00 Max. :53.13 Max. :5.000   
## region charges   
## northeast:324 Min. : 1122   
## northwest:325 1st Qu.: 4740   
## southeast:364 Median : 9382   
## southwest:325 Mean :13270   
## 3rd Qu.:16640   
## Max. :63770

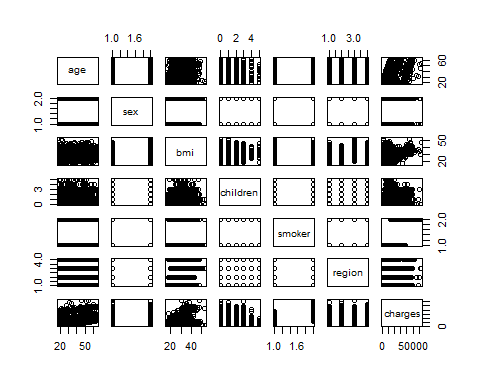
hist(data$charges)



As we know, the insurance charges are distributed as right-skewed.

Let’s visualize the relation between each variable.

pairs(data)



It looks that the response variable charges is correlated with the variables age, bmi, and smoker.

## Split the data

Finally, let’s prepare for building a machine learning model!

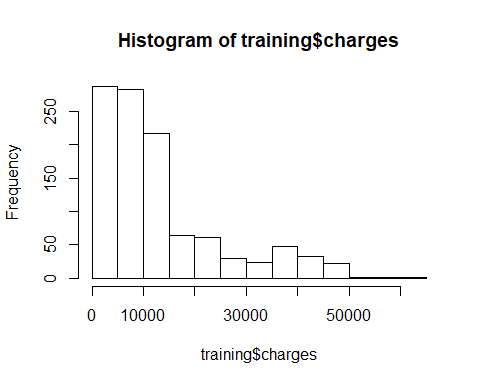
set.seed(0)  
inTrain <- createDataPartition(y=data$charges, p=0.8, list=FALSE)  
training <- data[inTrain,]  
testing <- data[-inTrain,]  
  
dim(training)

## [1] 1072 7

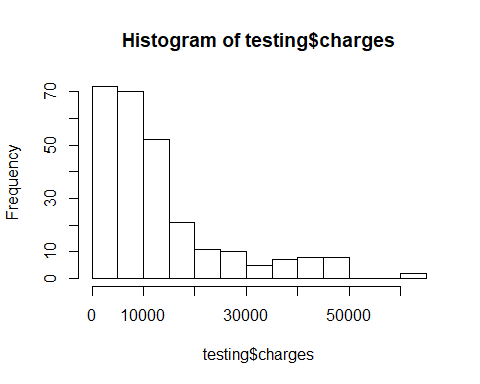
dim(testing)

## [1] 266 7

# Training  
hist(training$charges)



# testing  
hist(testing$charges)



I splitted tha data here. Let’s check the ditribution of charges in training dataset is nearly equal to that in testing dataset. They look almost similar, so no problem in further analysis.

Create a base model of regression. Before parameter tuning or something, let’s see the result as soon as possible. If the accuracy is enough to achieve my objective, that is enough!

## Base model

# Define the control  
trControl <- trainControl(method = "cv", number = 10, search = "grid")  
  
# RF  
# Perform training:  
set.seed(1)  
rf\_regressor = randomForest(training$charges ~ ., data=training, ntree=400, mtry=3, importance=TRUE, maxnodes = 500, trControl=trControl, metrics="accuracy")  
rf\_regressor

##   
## Call:  
## randomForest(formula = training$charges ~ ., data = training, ntree = 400, mtry = 3, importance = TRUE, maxnodes = 500, trControl = trControl, metrics = "accuracy")   
## Type of random forest: regression  
## Number of trees: 400  
## No. of variables tried at each split: 3  
##   
## Mean of squared residuals: 21413482  
## % Var explained: 85.19

I just set up ntree, mtry, maxnodes, trControl, metrics here. They determine how much deep the trees become, how many counts the experiment iterates, and how we evaluate the model as being “good”. I did cv 10 times as default setting. The parameter should be determined by grid search, but the model here seems enough to analyze the data I deal with in this experiment because the performance is very nice as I will explain later.

## Predict

Because the base model seems enough fitted to the training data, let’s see the accuracy when applicating it to the testing data.

The graph shows the deviations of each data from the redline meaning predicted values are equal to the actual values.

## Accuracy measures

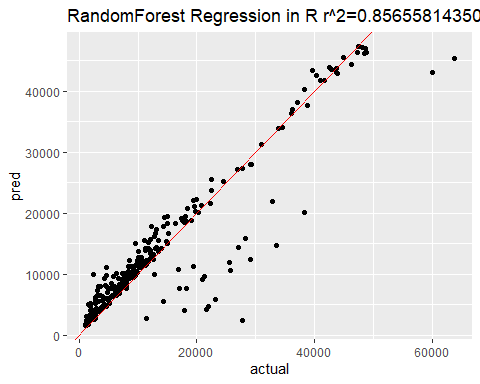
# R2  
r2 <- rSquared(testing$charges, testing$charges - rf\_pred)  
r2

## [,1]  
## [1,] 0.8565581

# RMSE  
rmse <- (mean((testing$charges - rf\_pred)^2))^(1/2)  
rmse

## [1] 4704.389

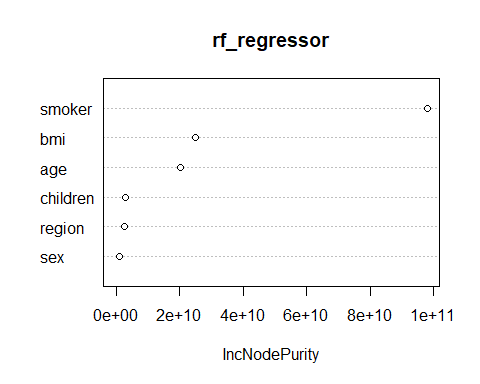
# Viz  
p <- ggplot(aes(x=actual, y=pred),  
 data=data.frame(actual=testing$charges, pred=rf\_pred))  
p + geom\_point() +  
 geom\_abline(color="red") +  
 ggtitle(paste("RandomForest Regression in R r^2=", r2, sep=""))



In the testing data, the model achieved its performance; accuracy is 85.6% !

## Features Importance

varImpPlot(rf\_regressor,type=2)



It looks that the variables smoker, bmi, age are important according to the above visualization.

## Predict the new client’s charges

Finally, let’s apply my accurate model to prediction of a new client!

Here, let’s assume the client name is Donald Trump. He is a 73 years old man, having 5 children. His bmi may be 30 and he is well known to hate smoking. He is living in Washington DC, which is located in Northeast of America.

# generate the new dataframe including the new client's information.  
Donald <- data.frame("age" = 73, "sex" = "male", "bmi" = 30, "children" = 5,  
 "smoker" = "no", "region" = "northeast")  
  
# Adjust the level of df\_test to that of df  
levels(Donald$sex) <- levels(data$sex)  
levels(Donald$smoker) <- levels(data$smoker)  
levels(Donald$region) <- levels(data$region)  
  
# Predict the Donald's charges per year.  
pred\_Donald = predict(rf\_regressor, newdata=Donald)  
pred\_Donald

## 1   
## 20113.29

Finally, we predict the charges Donald requests us to pay back should be 20113 USD.