main

# Check

##################################################################################  
#  
# Course: 32513 - Advanced Data Analytics Algorithms  
# Week:   
# data source: Kaggle  
# dataset: Stay Alert! Ford Challenge  
# Filters:   
#  
# Comments:   
# From:   
#  
##################################################################################  
  
#######################################################################  
# #  
# 1 - LOADING DATA #  
# #  
#######################################################################  
  
##### Reading the file #####  
data <- read.csv("Data/fordTrain.csv",  
 header=TRUE, stringsAsFactors=FALSE, na.strings = c("NA", ""),  
 strip.white = TRUE, blank.lines.skip=TRUE, skip=0)  
  
validatedata <- read.csv("Data/fordTest.csv",  
 header=TRUE, stringsAsFactors=FALSE, na.strings = c("NA", ""),  
 strip.white = TRUE, blank.lines.skip=TRUE, skip=0)  
  
summary(data)

## TrialID ObsNum IsAlert P1   
## Min. : 0.0 Min. : 0.0 Min. :0.0000 Min. :-22.48   
## 1st Qu.:125.0 1st Qu.: 302.0 1st Qu.:0.0000 1st Qu.: 31.76   
## Median :250.0 Median : 604.0 Median :1.0000 Median : 34.15   
## Mean :250.2 Mean : 603.8 Mean :0.5788 Mean : 35.45   
## 3rd Qu.:374.0 3rd Qu.: 906.0 3rd Qu.:1.0000 3rd Qu.: 37.31   
## Max. :510.0 Max. :1210.0 Max. :1.0000 Max. :101.35   
## P2 P3 P4 P5   
## Min. :-45.629 Min. : 504 Min. : 23.89 Min. : 0.03892   
## 1st Qu.: 9.904 1st Qu.: 792 1st Qu.: 49.18 1st Qu.: 0.09211   
## Median : 11.400 Median :1000 Median : 60.00 Median : 0.10508   
## Mean : 11.997 Mean :1027 Mean : 64.06 Mean : 0.17892   
## 3rd Qu.: 13.644 3rd Qu.:1220 3rd Qu.: 75.76 3rd Qu.: 0.13881   
## Max. : 71.174 Max. :2512 Max. :119.05 Max. :27.20220   
## P6 P7 P8 E1   
## Min. : 128.0 Min. : 0.2622 Min. :0 Min. : 0.00   
## 1st Qu.: 668.0 1st Qu.: 66.6667 1st Qu.:0 1st Qu.: 0.00   
## Median : 800.0 Median : 75.0000 Median :0 Median : 0.00   
## Mean : 845.4 Mean : 77.8876 Mean :0 Mean : 10.51   
## 3rd Qu.: 900.0 3rd Qu.: 89.8204 3rd Qu.:0 3rd Qu.: 28.24   
## Max. :228812.0 Max. :468.7500 Max. :0 Max. :243.99   
## E2 E3 E4 E5   
## Min. : 0.0 Min. :0.0000 Min. :-250.00 Min. :0.00800   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: -8.00 1st Qu.:0.01569   
## Median : 0.0 Median :0.0000 Median : 0.00 Median :0.01600   
## Mean :102.8 Mean :0.2906 Mean : -4.23 Mean :0.01626   
## 3rd Qu.:211.6 3rd Qu.:0.0000 3rd Qu.: 6.00 3rd Qu.:0.01669   
## Max. :360.0 Max. :4.0000 Max. : 260.00 Max. :0.02394   
## E6 E7 E8 E9   
## Min. :260.0 Min. : 0.000 Min. :0.000 Min. :0.0000   
## 1st Qu.:348.0 1st Qu.: 0.000 1st Qu.:0.000 1st Qu.:1.0000   
## Median :365.0 Median : 1.000 Median :1.000 Median :1.0000   
## Mean :358.7 Mean : 1.757 Mean :1.383 Mean :0.8768   
## 3rd Qu.:367.0 3rd Qu.: 2.000 3rd Qu.:2.000 3rd Qu.:1.0000   
## Max. :513.0 Max. :25.000 Max. :9.000 Max. :1.0000   
## E10 E11 V1 V2   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. :-4.79500   
## 1st Qu.: 52.00 1st Qu.: 0.000 1st Qu.: 41.93 1st Qu.:-0.17500   
## Median : 67.00 Median : 0.000 Median :100.40 Median : 0.00000   
## Mean : 63.31 Mean : 1.315 Mean : 76.97 Mean :-0.03771   
## 3rd Qu.: 73.00 3rd Qu.: 0.000 3rd Qu.:108.50 3rd Qu.: 0.07000   
## Max. :127.00 Max. :52.400 Max. :129.70 Max. : 3.99000   
## V3 V4 V5 V6   
## Min. : 240.0 Min. : 0.000 Min. :0.0000 Min. : 0   
## 1st Qu.: 255.0 1st Qu.: 1.488 1st Qu.:0.0000 1st Qu.:1259   
## Median : 511.0 Median : 3.019 Median :0.0000 Median :1994   
## Mean : 573.8 Mean : 19.961 Mean :0.1798 Mean :1716   
## 3rd Qu.: 767.0 3rd Qu.: 7.481 3rd Qu.:0.0000 3rd Qu.:2146   
## Max. :1023.0 Max. :484.488 Max. :1.0000 Max. :4892   
## V7 V8 V9 V10 V11   
## Min. :0 Min. : 0.00 Min. :0 Min. :1.000 Min. : 1.677   
## 1st Qu.:0 1st Qu.: 0.00 1st Qu.:0 1st Qu.:3.000 1st Qu.: 7.948   
## Median :0 Median :12.80 Median :0 Median :4.000 Median : 10.773   
## Mean :0 Mean :12.71 Mean :0 Mean :3.312 Mean : 11.668   
## 3rd Qu.:0 3rd Qu.:21.90 3rd Qu.:0 3rd Qu.:4.000 3rd Qu.: 15.271   
## Max. :0 Max. :82.10 Max. :0 Max. :7.000 Max. :262.534

summary(validatedata)

## TrialID ObsNum IsAlert P1   
## Min. : 0.0 Min. : 0.0 Length:120840 Min. :17.78   
## 1st Qu.:24.0 1st Qu.: 302.0 Class :character 1st Qu.:33.46   
## Median :49.0 Median : 604.0 Mode :character Median :34.88   
## Mean :49.5 Mean : 603.7 Mean :35.45   
## 3rd Qu.:74.0 3rd Qu.: 906.0 3rd Qu.:36.86   
## Max. :99.0 Max. :1210.0 Max. :81.82   
## P2 P3 P4 P5   
## Min. :-25.912 Min. : 504 Min. : 26.00 Min. :0.04022   
## 1st Qu.: 9.601 1st Qu.: 788 1st Qu.: 49.18 1st Qu.:0.08951   
## Median : 11.289 Median :1000 Median : 60.00 Median :0.10119   
## Mean : 12.008 Mean :1027 Mean : 64.15 Mean :0.12234   
## 3rd Qu.: 13.543 3rd Qu.:1220 3rd Qu.: 76.14 3rd Qu.:0.11546   
## Max. : 39.757 Max. :2308 Max. :119.05 Max. :4.66777   
## P6 P7 P8 E1   
## Min. : 468.0 Min. : 49.34 Min. :0 Min. : 0.00   
## 1st Qu.: 636.0 1st Qu.: 76.92 1st Qu.:0 1st Qu.: 0.00   
## Median : 700.0 Median : 85.71 Median :0 Median : 0.00   
## Mean : 707.7 Mean : 86.17 Mean :0 Mean :10.65   
## 3rd Qu.: 780.0 3rd Qu.: 94.34 3rd Qu.:0 3rd Qu.:28.54   
## Max. :1216.0 Max. :128.21 Max. :0 Max. :64.69   
## E2 E3 E4 E5   
## Min. : 0.0 Min. :0.0000 Min. :-250.0000 Min. :0.00800   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: -6.0000 1st Qu.:0.01480   
## Median : 0.0 Median :0.0000 Median : 0.0000 Median :0.01588   
## Mean : 98.2 Mean :0.1912 Mean : -0.5721 Mean :0.01595   
## 3rd Qu.:211.3 3rd Qu.:0.0000 3rd Qu.: 6.0000 3rd Qu.:0.01739   
## Max. :360.0 Max. :4.0000 Max. : 254.0000 Max. :0.02394   
## E6 E7 E8 E9   
## Min. :260.0 Min. : 0.000 Min. :0.000 Min. :0.000   
## 1st Qu.:303.0 1st Qu.: 0.000 1st Qu.:0.000 1st Qu.:1.000   
## Median :356.0 Median : 1.000 Median :2.000 Median :1.000   
## Mean :347.7 Mean : 1.358 Mean :1.614 Mean :0.868   
## 3rd Qu.:367.0 3rd Qu.: 2.000 3rd Qu.:3.000 3rd Qu.:1.000   
## Max. :513.0 Max. :25.000 Max. :9.000 Max. :1.000   
## E10 E11 V1 V2   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. :-3.60500   
## 1st Qu.: 65.00 1st Qu.: 0.000 1st Qu.: 56.54 1st Qu.:-0.17500   
## Median : 71.00 Median : 0.000 Median :101.80 Median : 0.00000   
## Mean : 68.66 Mean : 1.771 Mean : 78.70 Mean :-0.03602   
## 3rd Qu.: 77.00 3rd Qu.: 0.000 3rd Qu.:107.65 3rd Qu.: 0.07000   
## Max. :127.00 Max. :36.800 Max. :122.65 Max. : 3.11500   
## V3 V4 V5 V6   
## Min. : 240.0 Min. : 0.000 Min. :0.0000 Min. : 0   
## 1st Qu.: 255.0 1st Qu.: 1.488 1st Qu.:0.0000 1st Qu.:1489   
## Median : 497.0 Median : 3.019 Median :0.0000 Median :2015   
## Mean : 562.1 Mean : 32.215 Mean :0.1794 Mean :1732   
## 3rd Qu.: 767.0 3rd Qu.: 5.994 3rd Qu.:0.0000 3rd Qu.:2123   
## Max. :1023.0 Max. :478.494 Max. :1.0000 Max. :3408   
## V7 V8 V9 V10 V11   
## Min. :0 Min. : 0.00 Min. :0 Min. :1.000 Min. : 9.538   
## 1st Qu.:0 1st Qu.: 0.00 1st Qu.:0 1st Qu.:4.000 1st Qu.:12.775   
## Median :0 Median :12.30 Median :0 Median :4.000 Median :13.471   
## Mean :0 Mean :12.84 Mean :0 Mean :3.379 Mean :13.900   
## 3rd Qu.:0 3rd Qu.:22.80 3rd Qu.:0 3rd Qu.:4.000 3rd Qu.:15.224   
## Max. :0 Max. :82.10 Max. :0 Max. :7.000 Max. :18.323

#######################################################################  
# #  
# 2 - DATA UNDERSTANDING #  
# #  
#######################################################################  
  
#To ensure steps are repeatable  
set.seed(131)  
  
#No of records  
nrow(data)

## [1] 604329

#No of attributes  
ncol(data)

## [1] 33

#No missing values  
summary(data)

## TrialID ObsNum IsAlert P1   
## Min. : 0.0 Min. : 0.0 Min. :0.0000 Min. :-22.48   
## 1st Qu.:125.0 1st Qu.: 302.0 1st Qu.:0.0000 1st Qu.: 31.76   
## Median :250.0 Median : 604.0 Median :1.0000 Median : 34.15   
## Mean :250.2 Mean : 603.8 Mean :0.5788 Mean : 35.45   
## 3rd Qu.:374.0 3rd Qu.: 906.0 3rd Qu.:1.0000 3rd Qu.: 37.31   
## Max. :510.0 Max. :1210.0 Max. :1.0000 Max. :101.35   
## P2 P3 P4 P5   
## Min. :-45.629 Min. : 504 Min. : 23.89 Min. : 0.03892   
## 1st Qu.: 9.904 1st Qu.: 792 1st Qu.: 49.18 1st Qu.: 0.09211   
## Median : 11.400 Median :1000 Median : 60.00 Median : 0.10508   
## Mean : 11.997 Mean :1027 Mean : 64.06 Mean : 0.17892   
## 3rd Qu.: 13.644 3rd Qu.:1220 3rd Qu.: 75.76 3rd Qu.: 0.13881   
## Max. : 71.174 Max. :2512 Max. :119.05 Max. :27.20220   
## P6 P7 P8 E1   
## Min. : 128.0 Min. : 0.2622 Min. :0 Min. : 0.00   
## 1st Qu.: 668.0 1st Qu.: 66.6667 1st Qu.:0 1st Qu.: 0.00   
## Median : 800.0 Median : 75.0000 Median :0 Median : 0.00   
## Mean : 845.4 Mean : 77.8876 Mean :0 Mean : 10.51   
## 3rd Qu.: 900.0 3rd Qu.: 89.8204 3rd Qu.:0 3rd Qu.: 28.24   
## Max. :228812.0 Max. :468.7500 Max. :0 Max. :243.99   
## E2 E3 E4 E5   
## Min. : 0.0 Min. :0.0000 Min. :-250.00 Min. :0.00800   
## 1st Qu.: 0.0 1st Qu.:0.0000 1st Qu.: -8.00 1st Qu.:0.01569   
## Median : 0.0 Median :0.0000 Median : 0.00 Median :0.01600   
## Mean :102.8 Mean :0.2906 Mean : -4.23 Mean :0.01626   
## 3rd Qu.:211.6 3rd Qu.:0.0000 3rd Qu.: 6.00 3rd Qu.:0.01669   
## Max. :360.0 Max. :4.0000 Max. : 260.00 Max. :0.02394   
## E6 E7 E8 E9   
## Min. :260.0 Min. : 0.000 Min. :0.000 Min. :0.0000   
## 1st Qu.:348.0 1st Qu.: 0.000 1st Qu.:0.000 1st Qu.:1.0000   
## Median :365.0 Median : 1.000 Median :1.000 Median :1.0000   
## Mean :358.7 Mean : 1.757 Mean :1.383 Mean :0.8768   
## 3rd Qu.:367.0 3rd Qu.: 2.000 3rd Qu.:2.000 3rd Qu.:1.0000   
## Max. :513.0 Max. :25.000 Max. :9.000 Max. :1.0000   
## E10 E11 V1 V2   
## Min. : 0.00 Min. : 0.000 Min. : 0.00 Min. :-4.79500   
## 1st Qu.: 52.00 1st Qu.: 0.000 1st Qu.: 41.93 1st Qu.:-0.17500   
## Median : 67.00 Median : 0.000 Median :100.40 Median : 0.00000   
## Mean : 63.31 Mean : 1.315 Mean : 76.97 Mean :-0.03771   
## 3rd Qu.: 73.00 3rd Qu.: 0.000 3rd Qu.:108.50 3rd Qu.: 0.07000   
## Max. :127.00 Max. :52.400 Max. :129.70 Max. : 3.99000   
## V3 V4 V5 V6   
## Min. : 240.0 Min. : 0.000 Min. :0.0000 Min. : 0   
## 1st Qu.: 255.0 1st Qu.: 1.488 1st Qu.:0.0000 1st Qu.:1259   
## Median : 511.0 Median : 3.019 Median :0.0000 Median :1994   
## Mean : 573.8 Mean : 19.961 Mean :0.1798 Mean :1716   
## 3rd Qu.: 767.0 3rd Qu.: 7.481 3rd Qu.:0.0000 3rd Qu.:2146   
## Max. :1023.0 Max. :484.488 Max. :1.0000 Max. :4892   
## V7 V8 V9 V10 V11   
## Min. :0 Min. : 0.00 Min. :0 Min. :1.000 Min. : 1.677   
## 1st Qu.:0 1st Qu.: 0.00 1st Qu.:0 1st Qu.:3.000 1st Qu.: 7.948   
## Median :0 Median :12.80 Median :0 Median :4.000 Median : 10.773   
## Mean :0 Mean :12.71 Mean :0 Mean :3.312 Mean : 11.668   
## 3rd Qu.:0 3rd Qu.:21.90 3rd Qu.:0 3rd Qu.:4.000 3rd Qu.: 15.271   
## Max. :0 Max. :82.10 Max. :0 Max. :7.000 Max. :262.534

#Formatting Dataset  
dput(names(data))

## c("TrialID", "ObsNum", "IsAlert", "P1", "P2", "P3", "P4", "P5",   
## "P6", "P7", "P8", "E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8",   
## "E9", "E10", "E11", "V1", "V2", "V3", "V4", "V5", "V6", "V7",   
## "V8", "V9", "V10", "V11")

data<-data[c("TrialID", "ObsNum", "P1", "P2", "P3", "P4", "P5",   
 "P6", "P7", "P8", "E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8",   
 "E9", "E10", "E11", "V1", "V2", "V3", "V4", "V5", "V6", "V7",   
 "V8", "V9", "V10", "V11", "IsAlert")]  
  
#Descriptive Statistics  
psych::describe(data, fast = FALSE )

## vars n mean sd median trimmed mad min  
## TrialID 1 604329 250.17 145.45 250.00 249.49 185.32 0.00  
## ObsNum 2 604329 603.84 348.93 604.00 603.83 447.75 0.00  
## P1 3 604329 35.45 7.48 34.15 34.56 4.00 -22.48  
## P2 4 604329 12.00 3.76 11.40 11.74 2.62 -45.63  
## P3 5 604329 1026.67 309.28 1000.00 1008.35 320.24 504.00  
## P4 6 604329 64.06 19.76 60.00 62.45 20.53 23.89  
## P5 7 604329 0.18 0.37 0.11 0.13 0.03 0.04  
## P6 8 604329 845.38 2505.34 800.00 794.79 177.91 128.00  
## P7 9 604329 77.89 18.58 75.00 77.20 17.35 0.26  
## P8 10 604329 0.00 0.00 0.00 0.00 0.00 0.00  
## E1 11 604329 10.51 14.05 0.00 9.06 0.00 0.00  
## E2 12 604329 102.79 127.26 0.00 85.89 0.00 0.00  
## E3 13 604329 0.29 1.01 0.00 0.00 0.00 0.00  
## E4 14 604329 -4.23 35.51 0.00 -1.26 11.86 -250.00  
## E5 15 604329 0.02 0.00 0.02 0.02 0.00 0.01  
## E6 16 604329 358.67 27.40 365.00 360.09 19.27 260.00  
## E7 17 604329 1.76 2.85 1.00 1.02 1.48 0.00  
## E8 18 604329 1.38 1.61 1.00 1.10 1.48 0.00  
## E9 19 604329 0.88 0.33 1.00 0.97 0.00 0.00  
## E10 20 604329 63.31 18.89 67.00 63.49 11.86 0.00  
## E11 21 604329 1.32 5.25 0.00 0.00 0.00 0.00  
## V1 22 604329 76.97 44.39 100.40 81.48 17.11 0.00  
## V2 23 604329 -0.04 0.40 0.00 -0.03 0.26 -4.80  
## V3 24 604329 573.79 298.41 511.00 559.53 379.55 240.00  
## V4 25 604329 19.96 63.27 3.02 5.23 2.27 0.00  
## V5 26 604329 0.18 0.38 0.00 0.10 0.00 0.00  
## V6 27 604329 1715.69 618.18 1994.00 1769.39 302.45 0.00  
## V7 28 604329 0.00 0.00 0.00 0.00 0.00 0.00  
## V8 29 604329 12.71 11.53 12.80 11.78 18.09 0.00  
## V9 30 604329 0.00 0.00 0.00 0.00 0.00 0.00  
## V10 31 604329 3.31 1.24 4.00 3.50 0.00 1.00  
## V11 32 604329 11.67 9.93 10.77 11.35 4.94 1.68  
## IsAlert 33 604329 0.58 0.49 1.00 0.60 0.00 0.00  
## max range skew kurtosis se  
## TrialID 510.00 510.00 0.02 -1.17 0.19  
## ObsNum 1210.00 1210.00 0.00 -1.20 0.45  
## P1 101.35 123.83 2.45 17.77 0.01  
## P2 71.17 116.80 0.86 10.06 0.00  
## P3 2512.00 2008.00 0.51 -0.28 0.40  
## P4 119.05 95.16 0.64 -0.30 0.03  
## P5 27.20 27.16 20.22 937.06 0.00  
## P6 228812.00 228684.00 89.79 8164.54 3.22  
## P7 468.75 468.49 1.94 22.75 0.02  
## P8 0.00 0.00 NaN NaN 0.00  
## E1 243.99 243.99 0.77 0.15 0.02  
## E2 360.00 360.00 0.75 -1.00 0.16  
## E3 4.00 4.00 3.35 9.42 0.00  
## E4 260.00 510.00 -2.47 19.76 0.05  
## E5 0.02 0.02 0.18 3.82 0.00  
## E6 513.00 253.00 -0.47 2.95 0.04  
## E7 25.00 25.00 2.84 8.13 0.00  
## E8 9.00 9.00 2.20 6.34 0.00  
## E9 1.00 1.00 -2.29 3.26 0.00  
## E10 127.00 127.00 -0.43 1.11 0.02  
## E11 52.40 52.40 4.06 15.66 0.01  
## V1 129.70 129.70 -0.94 -0.85 0.06  
## V2 3.99 8.79 -0.76 14.69 0.00  
## V3 1023.00 783.00 0.26 -1.42 0.38  
## V4 484.49 484.49 5.14 27.41 0.08  
## V5 1.00 1.00 1.67 0.78 0.00  
## V6 4892.00 4892.00 -0.87 -0.79 0.80  
## V7 0.00 0.00 NaN NaN 0.00  
## V8 82.10 82.10 0.41 -0.70 0.01  
## V9 0.00 0.00 NaN NaN 0.00  
## V10 7.00 6.00 -1.17 -0.16 0.00  
## V11 262.53 260.86 21.17 527.15 0.01  
## IsAlert 1.00 1.00 -0.32 -1.90 0.00

# Charts

hist\_box\_plots <- function(dat, x){  
 hist(dat[, x], main = paste0("Histogram of ",x), xlab=x, ylab="Frequency", col = "blue")  
 boxplot(dat[, x], main = paste0("Boxplot of ",x), xlab=x, col = "blue")  
}  
  
variables <- c("P1", "P2", "P3", "P4", "P5",   
 "P6", "P7", "P8", "E1", "E2", "E3", "E4", "E5", "E6", "E7", "E8",   
 "E9", "E10", "E11", "V1", "V2", "V3", "V4", "V5", "V6", "V7",   
 "V8", "V9", "V10", "V11")  
  
layout\_matrix <- matrix(c(1:4), nrow = 2, ncol = 2, byrow = T)  
layout(mat = layout\_matrix)  
lapply(variables, hist\_box\_plots, dat=data)

