Project 1

Deepak Kumar

July 10, 2016

watch\_acc <- read.csv("C:/Users/Deepak Rathi/Downloads/Watch\_accelerometer/Watch\_accelerometer.csv")

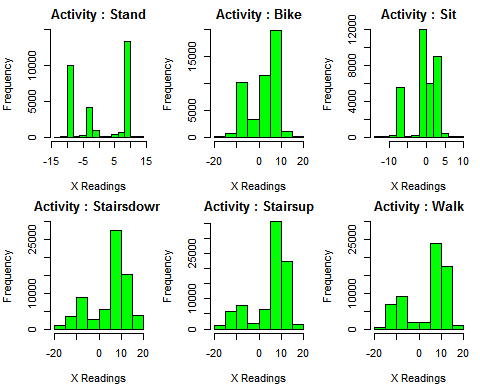
Comments: Here we are subsetting the Data. As we are restricting ourselves to Samsung Gear Data.

gear <- subset(watch\_acc, Model == "gear")  
gear1 <- subset(gear, Device == "gear\_1")  
gear2 <- subset(gear, Device == "gear\_2")  
gear1 <- gear1[2:10]  
gear2 <- gear2[2:10]

we create a Seperate multi-panel Graph that shows the distribution of "X" watches grouped by six different activites.

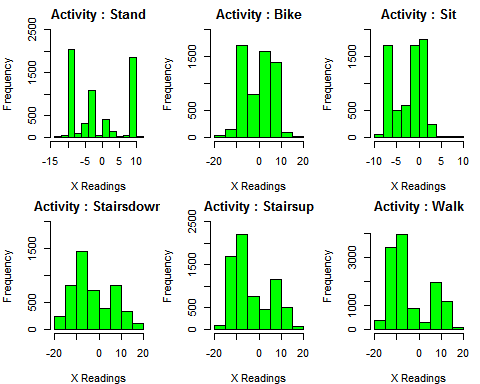
**Histogram for Samsung Gear 1 device Grouped By 6 Different Activites**

##Histogram 1  
par(mfrow = c(2, 3), mar = c(4, 4, 2, 1))  
hist(subset(gear1, gear1$gt=="stand")$x, col="green", breaks = 10, ylim = c(0,15000), xlab = "X Readings", main = "Activity : Stand" )  
hist(subset(gear1, gear1$gt=="bike")$x, col="green" , breaks = 10, ylim = c(0,20000),xlab = "X Readings", main = "Activity : Bike")  
hist(subset(gear1, gear1$gt=="sit")$x, col="green", breaks = 10, xlab = "X Readings", main = "Activity : Sit")  
  
hist(subset(gear1, gear1$gt=="stairsdown")$x, col="green" , breaks = 10, ylim = c(0,30000), xlab = "X Readings", main = "Activity : Stairsdown")  
hist(subset(gear1, gear1$gt=="stairsup")$x, col="green" , breaks = 10, xlab = "X Readings", main = "Activity : Stairsup")  
hist(subset(gear1, gear1$gt=="walk")$x, col="green", breaks = 10,ylim = c(0,30000), xlab = "X Readings", main = "Activity : Walk")



Histogram for Samsung Gear 2 device Grouped By 6 Different Activites

#Histogram 2  
par(mfrow = c(2, 3), mar = c(4, 4, 2, 1))  
hist(subset(gear2, gear2$gt=="stand")$x, col="green",breaks = 10, ylim = c(0,2500), xlab = "X Readings", main = "Activity : Stand" )  
hist(subset(gear2, gear2$gt=="bike")$x, col="green" , breaks = 10, ylim = c(0,2000), xlab = "X Readings", main = "Activity : Bike" )  
hist(subset(gear2, gear2$gt=="sit")$x, col="green" , breaks = 10, ylim = c(0,2000), xlab = "X Readings", main = "Activity : Sit" )  
hist(subset(gear2, gear2$gt=="stairsdown")$x, col="green" , breaks = 10, ylim = c(0,2000), xlab = "X Readings", main = "Activity : Stairsdown" )  
hist(subset(gear2, gear2$gt=="stairsup")$x, col="green" , breaks = 10, ylim = c(0,2500), xlab = "X Readings", main = "Activity : Stairsup" )  
hist(subset(gear2, gear2$gt=="walk")$x, col="green" , breaks = 10, ylim = c(0,4500), xlab = "X Readings", main = "Activity : Walk" )



Comparision of Samsung Gear 1 and Samsung Gear 2 Devices.

The Histogram of Stand Activity is showing that in Gear 1 and Gear 2 the maximum frequency of values are on the -10 and 10. while the second highest frequency values are on -5 in both graph. The largest frequency values in Gear 1 on 10, while in the Gear2 largest frequenct values are on -10. During the Bike Activity most of the frequency of distribution of data lies on between 0 to 10. while in Gear 2 data -10 to -5 and 0 to 10. The largest frequncey in Gear 1 is on 10 while n the case of Gear 2 is opposite, it is on -10. During the "Sit" Activity the Gear 1 reading is showing that most of the frequency lies between -2.5 to 4, while the Gear 2 reading is showing that most of the data lies between -8 to 4. The difference between both devices readings is that largest frequency in Gear 1 is on -2 and Gear2 is on 2. During the Starisdown, stairsup and Walk activity the largest frequncy of the readings in Gear1 device are between 5 to 15. while in Ger2 Device largest frequency of reading in Gear2 device are between -15 to 0. In conclusion we can say that largest frequency of data shown in both devices is opposite if the largest frequent reading in Gear1 during the "Sit" activity is recorded on -10 than in Gear 2 it will be on 10.

str(gear)

## 'data.frame': 449077 obs. of 10 variables:  
## $ Index : int 0 1 2 3 4 5 6 7 8 9 ...  
## $ Arrival\_Time : num 1.42e+12 1.42e+12 1.42e+12 1.42e+12 1.42e+12 ...  
## $ Creation\_Time: num 2.79e+13 2.79e+13 2.79e+13 2.79e+13 2.79e+13 ...  
## $ x : num -0.565 -0.833 -1.018 -1.223 -1.577 ...  
## $ y : num -9.57 -9.71 -9.94 -10.14 -10.48 ...  
## $ z : num -0.614 -0.607 -0.544 -0.566 -0.403 ...  
## $ User : Factor w/ 9 levels "a","b","c","d",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Model : Factor w/ 2 levels "gear","lgwatch": 1 1 1 1 1 1 1 1 1 1 ...  
## $ Device : Factor w/ 4 levels "gear\_1","gear\_2",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ gt : Factor w/ 7 levels "bike","null",..: 6 6 6 6 6 6 6 6 6 6 ...

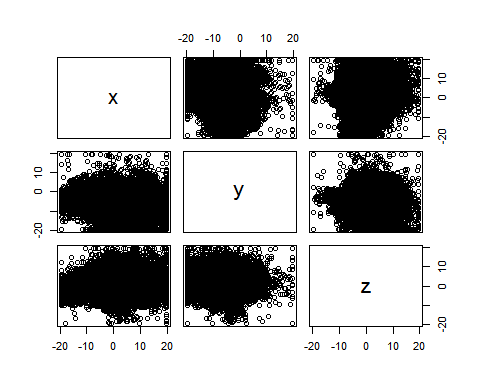
levels(gear$gt)

## [1] "bike" "null" "sit" "stairsdown" "stairsup"   
## [6] "stand" "walk"

gear$Device <- factor(gear$Device)  
gear <- subset(gear, gear$gt !="null")  
gear$gt <- factor(gear$gt)  
  
## There is a null which is of no use, so we will drop the all rows will null value in gt column  
## Drop rows with Null value in Gear1 device Data  
gear1 <- subset(gear1, gear1$gt !="null")  
gear1$gt <- factor(gear1$gt)  
## Drop rows with Null value in Gear2 device Data  
gear2 <- subset(gear2, gear2$gt !="null")  
gear2$gt <- factor(gear2$gt)

**Here we will check the Co- relation between three Numerical variables of both devices**

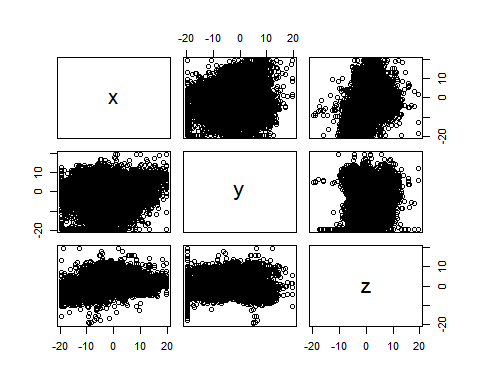
pairs(gear1[,3:5])



cor(gear1[,3:5])

## x y z  
## x 1.00000000 0.09471302 -0.0190213  
## y 0.09471302 1.00000000 -0.1180640  
## z -0.01902130 -0.11806398 1.0000000

pairs(gear2[,3:5])



cor(gear1[,3:5])

## x y z  
## x 1.00000000 0.09471302 -0.0190213  
## y 0.09471302 1.00000000 -0.1180640  
## z -0.01902130 -0.11806398 1.0000000

Both Graphs and co –relation table is showing no strong correlation between any numerical variables.

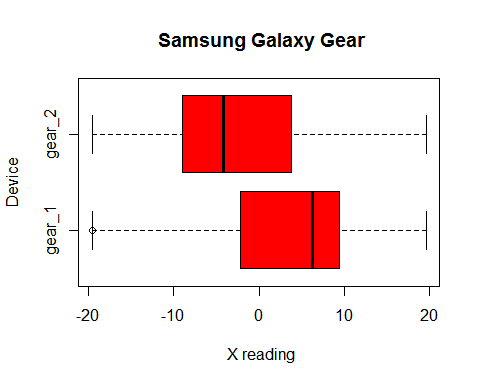
#Table   
aggregate(x ~ Device , data= gear, FUN = summary)

## Device x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.  
## 1 gear\_1 -19.610 -2.160 6.270 3.498 9.430 19.610  
## 2 gear\_2 -19.610 -8.959 -4.172 -2.664 3.750 19.610

#Table  
aggregate(x ~ Device , data= gear, FUN = sd)

## Device x  
## 1 gear\_1 7.971189  
## 2 gear\_2 7.763091

#boxplot  
boxplot(gear$x ~gear$Device, col= "red", horizontal = TRUE, ylab = "Device", xlab = " X reading", main=" Samsung Galaxy Gear")



Insight: The box plot shows that gear2 is skewed towards Right while the Gear 1 is skewed towards left. The box plot for Gear 1 is showing one extreme value, while both devices inter quantile range is almost same.

Here we want to calculate that mean and standard deviation

require(ggplot2)

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.2.5

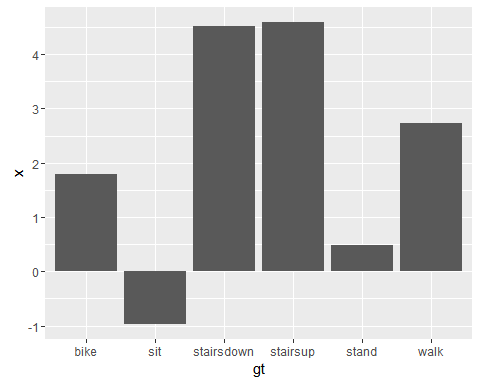
#Table  
aggregate(x ~ gt , data= gear, FUN = sd)

## gt x  
## 1 bike 6.172026  
## 2 sit 3.703329  
## 3 stairsdown 8.544426  
## 4 stairsup 8.372862  
## 5 stand 8.212414  
## 6 walk 9.413649

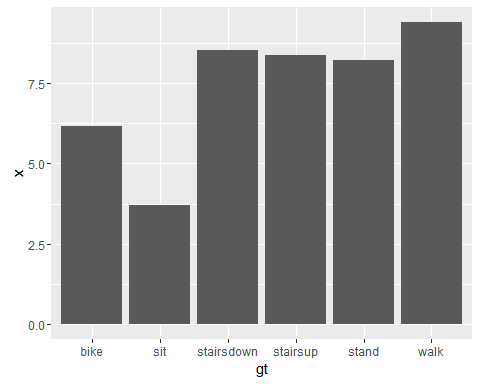
#Table  
aggregate(x ~ gt , data= gear, FUN = mean)

## gt x  
## 1 bike 1.7894958  
## 2 sit -0.9653837  
## 3 stairsdown 4.5224975  
## 4 stairsup 4.5913889  
## 5 stand 0.4782217  
## 6 walk 2.7385756

#Barplot showing mean of Both Gear Devices  
activitymean <- data.frame(aggregate(x ~ gt , data= gear, FUN = mean))  
ggplot(activitymean, aes(x = gt, y = x)) + geom\_bar(stat= "identity", position = "identity")



#Barplot showing standard Deviation of Both Gear Devices  
  
activitysd <- data.frame(aggregate(x ~ gt , data= gear, FUN = sd))  
ggplot(activitysd, aes(x = gt, y = x)) + geom\_bar(stat= "identity", position = "identity")



This chart and table is showing that data of “sit”, “stairsdown”, “stairsup” are closer to mean, while rest of away from mean

#Table  
aggregate(x ~ gt , data= gear1, FUN = median)

## gt x  
## 1 bike 4.6537294  
## 2 sit -0.5955577  
## 3 stairsdown 7.4714058  
## 4 stairsup 8.0391310  
## 5 stand -1.3551182  
## 6 walk 7.8894930

# Table  
aggregate(x ~ gt , data= gear1, FUN = summary)

## gt x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.  
## 1 bike -19.6100 -4.1930 4.6540 2.0390 6.3450 19.6100  
## 2 sit -13.6500 -1.3740 -0.5956 -0.6666 2.2340 9.3870  
## 3 stairsdown -19.6100 2.3050 7.4710 5.0510 10.4000 19.6100  
## 4 stairsup -19.6100 3.9960 8.0390 5.2630 10.4000 19.6100  
## 5 stand -12.6000 -9.1960 -1.3550 0.8346 9.3580 13.0000  
## 6 walk -19.6100 -5.8710 7.8890 4.0570 10.5200 19.6100

# Table  
  
aggregate(x ~ gt , data= gear1, FUN = mean)

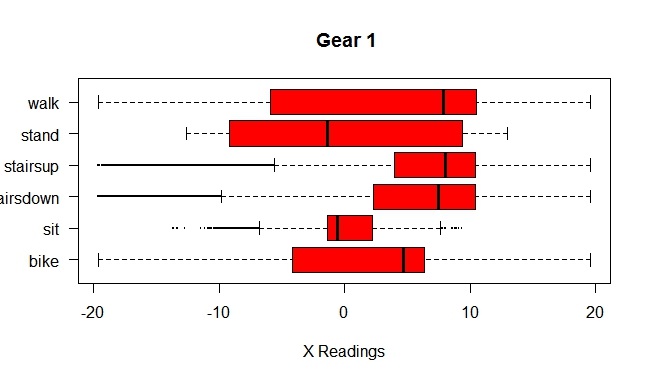
## gt x  
## 1 bike 2.0387386  
## 2 sit -0.6666481  
## 3 stairsdown 5.0510945  
## 4 stairsup 5.2627622  
## 5 stand 0.8345651  
## 6 walk 4.0573863

# Table  
aggregate(x ~ gt , data= gear1, FUN = sd)

## gt x  
## 1 bike 6.121920  
## 2 sit 3.648678  
## 3 stairsdown 8.278069  
## 4 stairsup 8.009458  
## 5 stand 8.287565  
## 6 walk 8.932245

Insight: In the Standard Deviation and Mean table it is showing that the data of Samsung Gear1 device of these activities “Sit”, “Stairsup” and “Stairsdown” is dispersed near to mean, while the data of “ Stand” activity is to away from mean.

#boxplot   
boxplot(gear1$x ~ gear1$gt, col="red" , xlab = "Activities", ylab=" X Readings" , main = " Gear 1" , las = 1, cex=0.2, horizontal= TRUE)



Insight: This histogram is showing that Stairsup and stairsdown and sit activity has more extreme values(outliers), The walk, bike, stairsup and starisdown activity is skewed towards left, while stand and sit activity is skewed towards right.

# Table  
aggregate(x ~ gt , data= gear2, FUN = median)

## gt x  
## 1 bike 1.371279  
## 2 sit -1.176152  
## 3 stairsdown -5.260360  
## 4 stairsup -6.994062  
## 5 stand -3.791817  
## 6 walk -7.326856

#Table  
aggregate(x ~ gt , data= gear2, FUN = summary)

## gt x.Min. x.1st Qu. x.Median x.Mean x.3rd Qu. x.Max.  
## 1 bike -19.6100 -6.2580 1.3710 -0.2282 5.0800 19.6100  
## 2 sit -9.6670 -6.1610 -1.1760 -2.4630 0.5866 9.0150  
## 3 stairsdown -19.6100 -9.5540 -5.2600 -2.9270 5.3860 19.6100  
## 4 stairsup -19.6100 -10.0600 -6.9940 -3.4700 4.9170 19.6100  
## 5 stand -13.8500 -9.0570 -3.7920 -1.2960 8.6960 10.7800  
## 6 walk -19.6100 -10.8900 -7.3270 -4.0520 6.3500 17.5900

# Table  
aggregate(x ~ gt , data= gear2, FUN = mean)

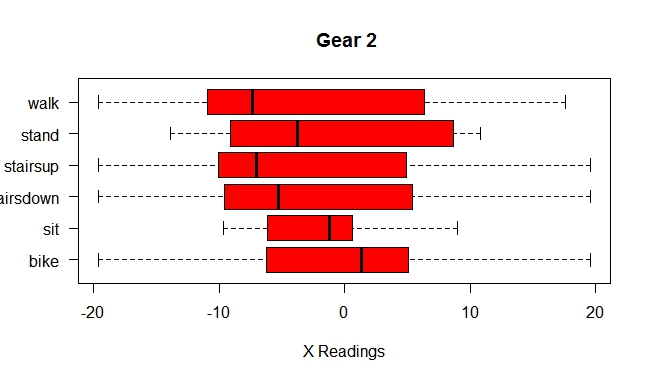
## gt x  
## 1 bike -0.2282395  
## 2 sit -2.4628045  
## 3 stairsdown -2.9268133  
## 4 stairsup -3.4704512  
## 5 stand -1.2964490  
## 6 walk -4.0520271

#Table  
aggregate(x ~ gt , data= gear2, FUN = sd)

## gt x  
## 1 bike 6.205900  
## 2 sit 3.611230  
## 3 stairsdown 8.760049  
## 4 stairsup 8.437817  
## 5 stand 7.582624  
## 6 walk 8.889881

Insight: In the Standard Deviation and Mean table it is showing that the data of Samsung Gear2 device of these activities “Sit” and “Stand” is dispersed near to mean, while the data of “Stairsup” “starisdown” and “walk” activity is to highly disperse from mean.

#BoxPlot  
  
boxplot(gear2$x ~ gear2$gt, col="red", xlab = "Activities", xpd= TRUE, ylab=" X Readings" , main = " Gear 2" , las =1, cex = 0.2, horizontal = TRUE )



Insight: The gear2 data has no extreme values. The walk, stand stairsup, stairsdown are skewed towards right, while “sit” and “bike” activity is skewed towards right. Conclusion: The difference b/w gear1 and gear2 data is that gear 1 data contain most outliers, while Gear2 has no outliers. The gear2 device Inter quantile range is also more in comparison to gear1 devices.

A paragraph reflecting on the issues, challenges, or learning from this project.

The project was little bit challenging, because I always worked under someone who always gave guidance to what to do next, or work in a group, so this give little bit assurance that I am on right path, so the second part of the project was little bit challenging. I have to choose appropriate method, which is appropriate for this. Even I completed this project, but I am still little bit unsure that, the summaries and visualization I did, Ss that correct or not? The learning from this project was using new libraries, there are so many things I did, which are not in this report, but I learnt, if I get any error in code, so what type of error is this and how to debug this.