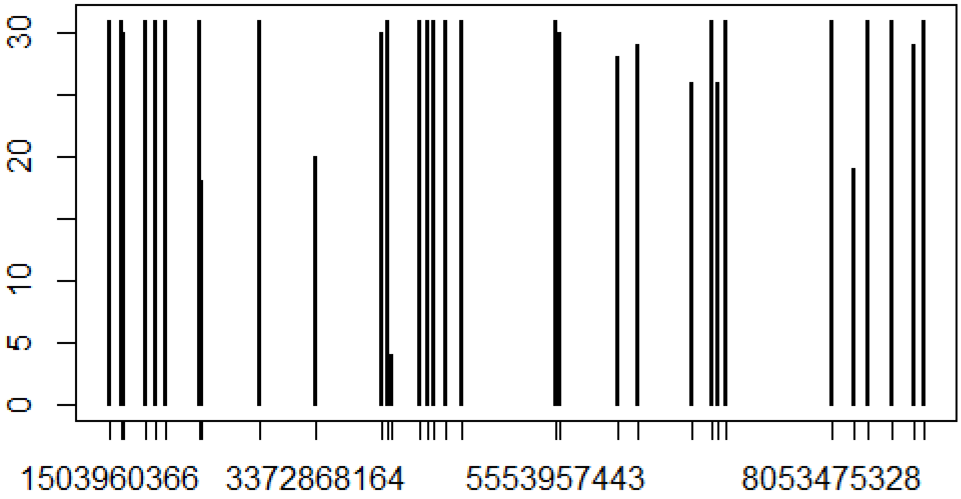
```
# Load all 18 Files

getwd()
setwd("D:/workin R/Copy1")
#Importing each file for initial exploration

df_dam <- read_csv("dailyActivity_merged.csv")
df_dcm <- read_csv("dailyCalories_merged.csv")
df_dim <- read_csv("dailyIntensities_merged.csv")
df_dsm <- read_csv("dailySteps_merged.csv")
df_hrsm <- read_csv("heartrate_seconds_merged.csv")
df_hcm <- read_csv("hourlyCalories_merged.csv")
df_him <- read_csv("hourlyIntensities_merged.csv")
df_hsm <- read_csv("hourlySteps_merged.csv")
df_mcnarrow <- read_csv("minuteCaloriesNarrow_merged.csv")
df_mcwide <- read_csv("minuteCaloriesWide_merged.csv")
df_minarrow <- read_csv("minuteIntensitiesNarrow_merged.csv")
df_miwide <- read_csv("minuteIntensitiesWide_merged.csv")
df_mMETnarrow <- read_csv("minuteMETsNarrow_merged.csv")
df_msleep <- read_csv("minuteSleep_merged.csv")
df_mstnarrow <- read_csv("minuteStepsNarrow_merged.csv")
df_mstwide<- read_csv("minuteStepsWide_merged.csv")
df_sldm<- read_csv("sleepDay_merged.csv")
df_wli<- read_csv("weightLogInfo_merged.csv")




########Start exploring each file   ###########
###Find what varables are there in files
colnames(df_dam)
summary(df_dam)
view(df_dam)


### Do this for all other files
## How Data is organized?

#There are 18 CSV files.
###18 files arranged under 4 major category
## daily data, hourly data, minute data , miscellaneous

#zeroing on daily data files

# The Csv file named daily activity merged appears to superset of other daily files.

#There is column in dsm (daily step merged)   "stepTotal" and "Totalsteps" in dam (daily activity
merged)  .....lets compare both

view(df_dam$TotalSteps-df_dsm$StepTotal)    #both are same
mean(df_dam$TotalSteps-df_dsm$StepTotal)
```

Comparing variables of merged files versus variables of individual files with help of summary functions.

```
summary(df_dam)
summary(df_dcm)

##Check identical or not
identical(df_dam$Id,df_dcm$Id)
identical(df_dam$Id,df_dim$Id)
identical(df_dam$Id,df_dsm$Id)


#We found that, indeed our idea is correct.  Therefore, for further analysis, I am going to ignore other
#daily files.

#why there is 940 observations?

# lets us count distinct values in ID
library(dplyr)
n_distinct(df_dam$Id)


## We have 33 unique ids

n_distinct(df_dam$ActivityDate)

##  we have 31 unique  days

#33 women *31 days = 1043

## Number of observation per woman is not same. In others words, not all women recorded their
measurements regularly.

#lets see count of each id

table(df_dam$Id)
plot(table(df_dam$Id))
```



# Yes , Not the same for each ID

# So , 15 Variables recorded on every day for single id
# 33 ids
# 31 days

### Next going on hourly data

```
##check for identical
identical(df_hcm$Id,df_him$Id)
identical(df_hcm$Id,df_hsm$Id)
identical(df_hcm$ActivityHour,df_hsm$ActivityHour)
```

##same activity hours and ids in all 3 hourly data files

## Why 22099 observations ?

```
## 24 Hr in single day
## 31 days
##ideally  24*31= 744 observations for each IDs

## BUT
table(df_hcm$Id)
```

```
##Not every Ids has 744 obseervation

## Thus, idealy 744*33 IDS= 24552 but we have 22099



####  Calory, intesity, steps  at hourly level  33 women
##Minutes wide file


##Check same Ids or not
identical(unique(df_hcm$Id),unique(df_mcwide$Id))
identical(unique(df_hcm$Id),unique(df_miwide$Id))
identical(unique(df_hcm$Id),unique(df_mstwide$Id))

### These minutes wide file contains data for same IDS


###These files contain calory, intesity and step data measured at minute level


### These minutes  wide file contains data for same IDS
## 21645 observations
###These files contain calory,intensity and step data measured at minute level
##Data is in wide format , 60 values for every activity hrs
```

```
##Heart Rate Data
##This is mean heart rate value at 5/10/15 seconds

##check IDS

unique(df_hrsm$Id)
unique(df_hcm$Id)


n_distinct(df_hrsm$Id)
##  we have heart rate data for only 14 IDS




##Weight  Log Data

n_distinct(df_wli$Id)

##For Only 8 IDS

colnames(df_wli)
summary(df_wli)
```

####   We have weight Data for 8 IDS
### we have heart rate data for 14 IDS
## We have sleep data 24 IDS
## We have Calory/Steps/Intensity for 33 IDS

Sr.No	File Name	Variables
1.	dailyActivity_merged	[1] "Id" "Activity Date" "Total Steps" [4] "Total Distance" "TrackerDistance" "LoggedActivitiesDistance" [7] "VeryActiveDistance" "ModeratelyActiveDistance" "LightActiveDistance" [10] "SedentaryActiveDistance" "VeryActiveMinutes" "FairlyActiveMinutes" [13] "LightlyActiveMinutes" "SedentaryMinutes" "Calories"
2.	dailyCalories_merged	"Id" "Activity Day" "Calories"
3.	dailyIntensities_merged	[1] "Id" "Activity Day" "Sedentary Minutes" [4] "LightlyActiveMinutes" "FairlyActiveMinutes" "VeryActiveMinutes" [7] "SedentaryActiveDistance" "LightActiveDistance" "ModeratelyActiveDistance" [10] "VeryActiveDistance"
4.	dailySteps_merged	[1] "Id" "Activity Day" "Step Total"
5.	heartrate_seconds_merged	Id" "Time" "Value
6.	hourlyCalories_merged	Id" "Activity Hour" "Calorie
7.	hourlyIntensities_merged	"Id" "Activity Hour" "Total Intensity" "Average Intensity"
8.	hourlySteps_merged	"Id" "Activity Hour" "Step Total"
9.	minuteCaloriesNarrow_merged	"Id" "ActivityMinute" "Calories"
10.	minuteCaloriesWide_merged	"Id" "ActivityHour" "Calories00"----Calories 59

11.	minuteIntensitiesNarrow_merged	"Id" "ActivityMinute" "Intensity"
12.	minuteIntensitiesWide_merged	"Id" "ActivityHour" "Intensity00" "Intensity01"… …Intensity59
13.	minuteMETsNarrow_merged	"Id" "ActivityMinute" "METs"
14.	minuteSleep_merged	"Id" "date" "value" "logId"
15.	minuteStepsNarrow_merged	"Id" "ActivityMinute" "Steps"
16.	minuteStepsWide_merged	"Id" "ActivityHour" "Steps00" "Steps01"……Step59
17.	sleepDay_merged	[1] "Id" "SleepDay" "TotalSleepRecords " "TotalMinutesAsleep" [5] "TotalTimeInBed"
18.	weightLogInfo_merged	[1] "Id" "Date" "WeightKg" "WeightPounds" "Fat" [6] "BMI" "IsManualReport" "LogId"

Variables

Sr.No	Variables	Explanation
1	ID	A unique id is assigned to each participating woman. There are 33 unique IDs
2.	Activity Day	Day of activity
3.	Total Distance	Total Distance travelled in kilometers
4	Total Steps	Total steps taken in 24 hr
5	Tracked Distance	Distance tracked by device
6	Logged Activity Distance	
7-10	Very active\Moderate active\light active \sedentry active distance	Kilometres tracked during very active\moderate active\light active\sedentary mode
11-14	Light active\very active\moderate active\sedanty active minutes	Minutes spent in 4 different intensity of activity
15	Calorie	Energy spent in kilocalories during a day
16	Time	Time of day
17.	Value	Heart-rate
18.	Activity Hr	The hour of activity
	Calorie	Calorie spent in a hour
20	Steps_total	Total steps taken I hr

21	Activity Minute	The time in minutes of activity
22	Calorie	Calorie consumed in that minute
23	Calorie 00-Calorie59	Calorie consumed in minute 00 , minute 01 and so on
24	Intensity	Intensity 0,1,2,3 sedentary-light-active-very active
25	Intesity 00-Intensity 59	Intensity at a particular minute
26	MET	Equivalent Metabolic activity
27	logID	The login Id of user
28	Value	Value
29	Sleep Day	The day of sleep
30	TotalTimeInBed	Total time spent in BED
31	WeightKg	Weight in Kg
32	TotalMinutesAsleep	Total Minutes spent sleeping
33	WeightPounds	Weight in Pounds
34	Fat	Fat of the participant
35	BMI	BMI
36	Is Manual Report	Is Manual Report

## Data cleaning

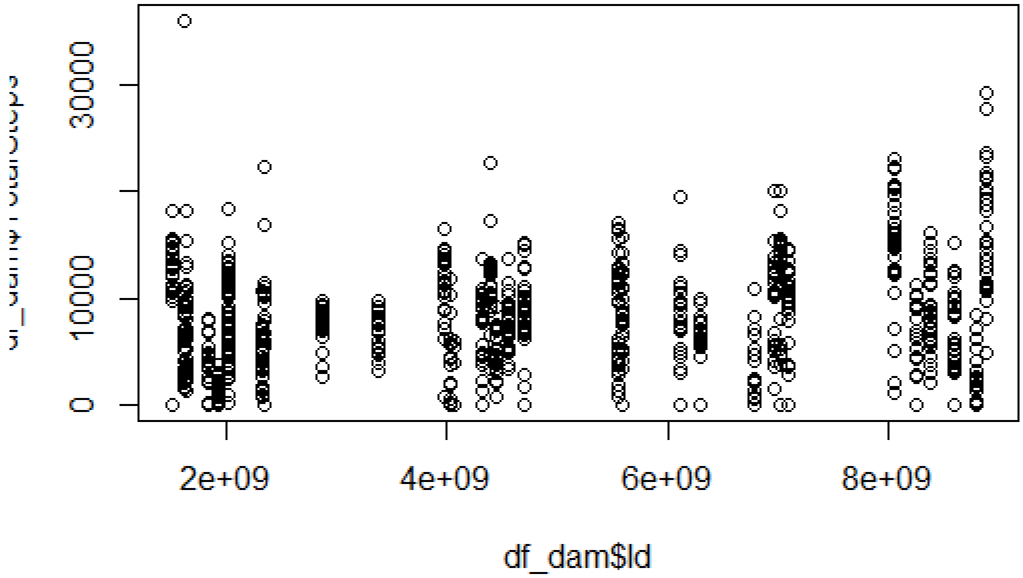
# For This stage, I am using daily data , Daily activity merged, other daily data files are not needed as all variables are included in this file .

# Not using minute data files, for this project information from minute data files can be representated by hourly data files.

##Starting with Daily file of 15 variables

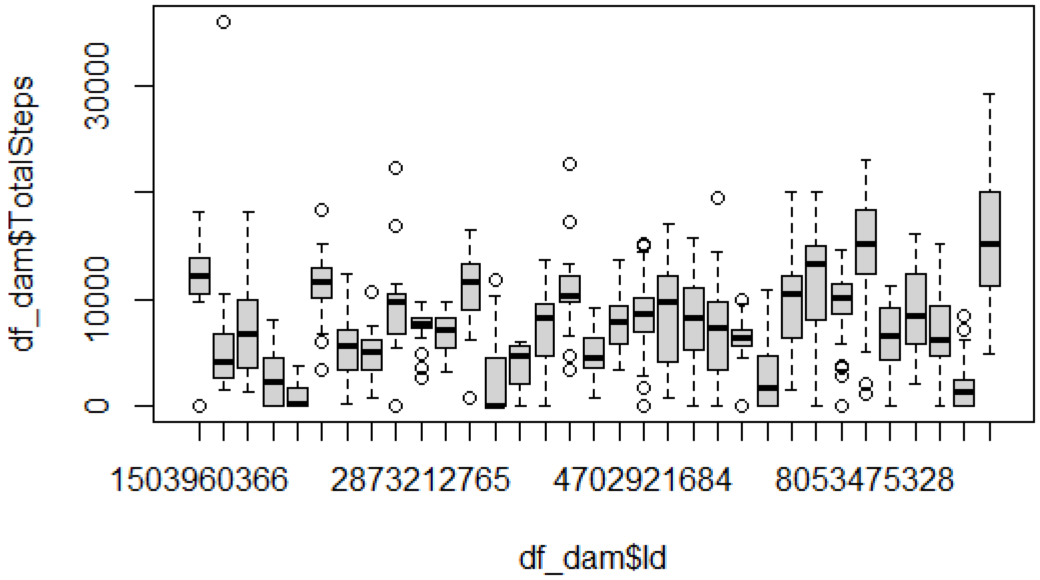
```
Str(df_dam)
```

```
plot(df_dam$Id,df_dam$TotalSteps)
```



```
##  It looks like Box plot for variables will be best
```

```
##Plot Box plot of total steps wrt ID
boxplot(df_dam$TotalSteps ~ df_dam$Id)
```



## This is interesting, Steps covered by 33 IDS. Remarkable Variation. The Minimum value of steps take is Zero.

```
####Filtering IDS which have less than 10 steps
v=df_dam[df_dam$TotalSteps <  10,]
table(v$Id)
```

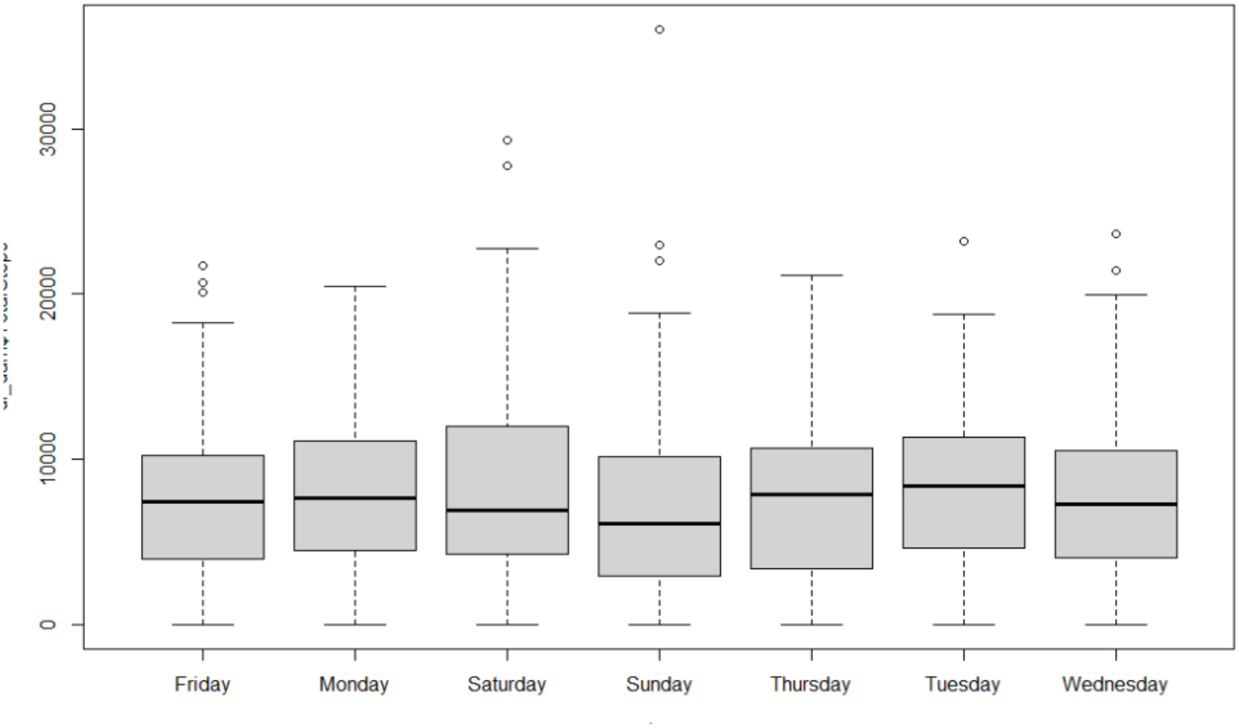
## There are 80 instances when less than 10 steps were covered.

```
print(v$Id)
print(table(v$Id))
```

## We have identified problematic IDS , which have low number of steps. What is wrong ? Has device malfunctioned? Participants were Sick? Some Accidents.? Ideally , an alert message should be generated in such cases.

```
1503960366 1844505072 1927972279 4020332650 4057192912 4702921684 5577150313
6117666160
         1         12         14         14          1          1          2
5
6290855005 6775888955 7007744171 7086361926 8253242879 8583815059 8792009665
         5         10          2          1          1          1         10
```

```
## Lets add weekday in the analysis
## the date format is not good, need to change
df_dam$ActivityDate <- as.Date.character(df_dam$ActivityDate, format="%m/%d/%Y")

#Add weekday column
df_dam<- transform(df_dam, Weekday=weekdays(df_dam$ActivityDate))
glimpse(df_dam)
head(df_dam)

# Draw steps vs weekdays
boxplot(df_dam$TotalSteps ~ df_dam$Weekday)
table(v$Weekday)
```



```
###Next Vriable Sedantry Minutes


# number of times 0 for sedantry minutes
table(df_dam$SedentaryMinutes==0)
```

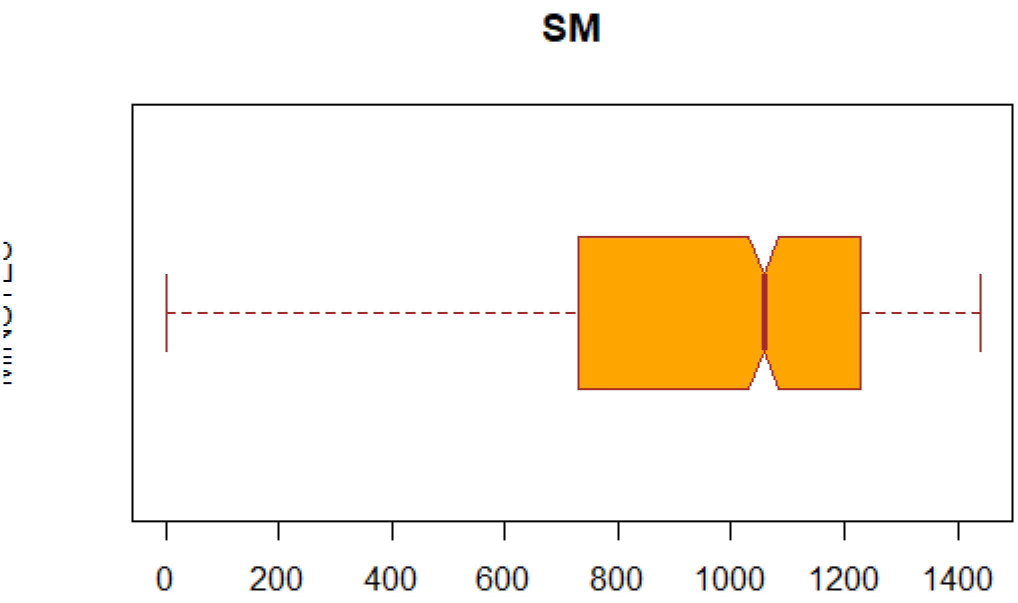
```
FALSE  TRUE 
  939     1 
```

## That’s Good , People take rest

```
summary(df_dam$SedentaryMinutes)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
    0.0   729.8  1057.5   991.2  1229.5  1440.0 
```

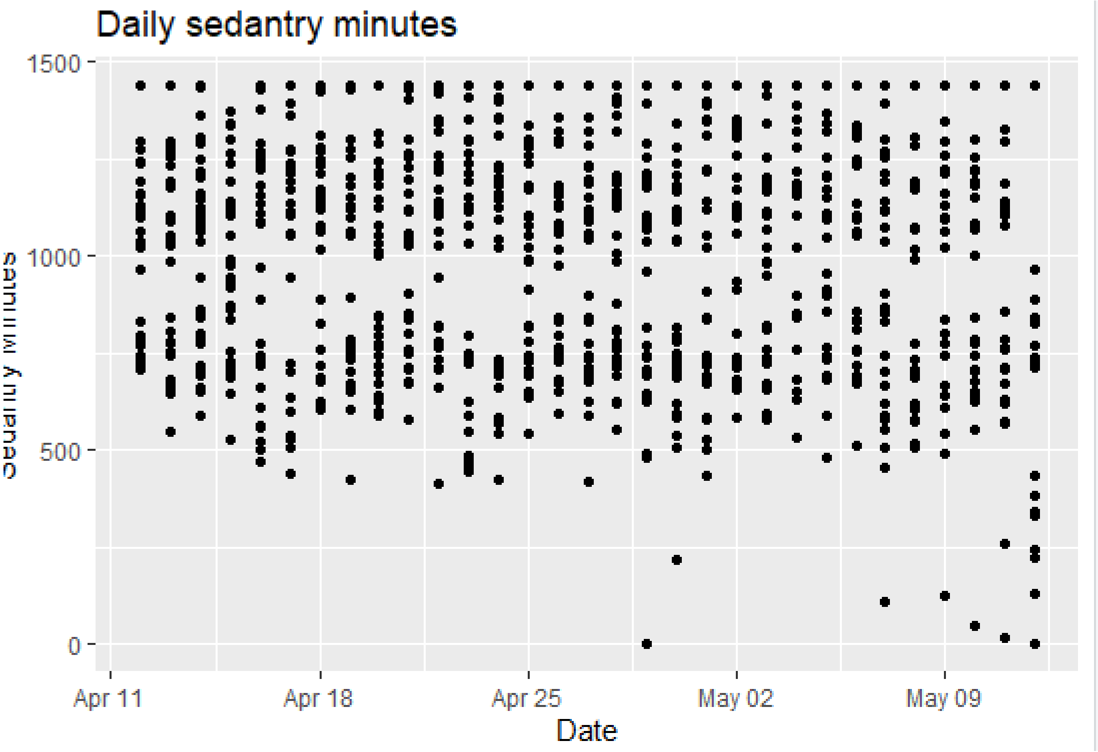
## Mean is 999 minutes , that is 16.5 Hr , 1st Quadrant is 729 i.e 12 hrs.

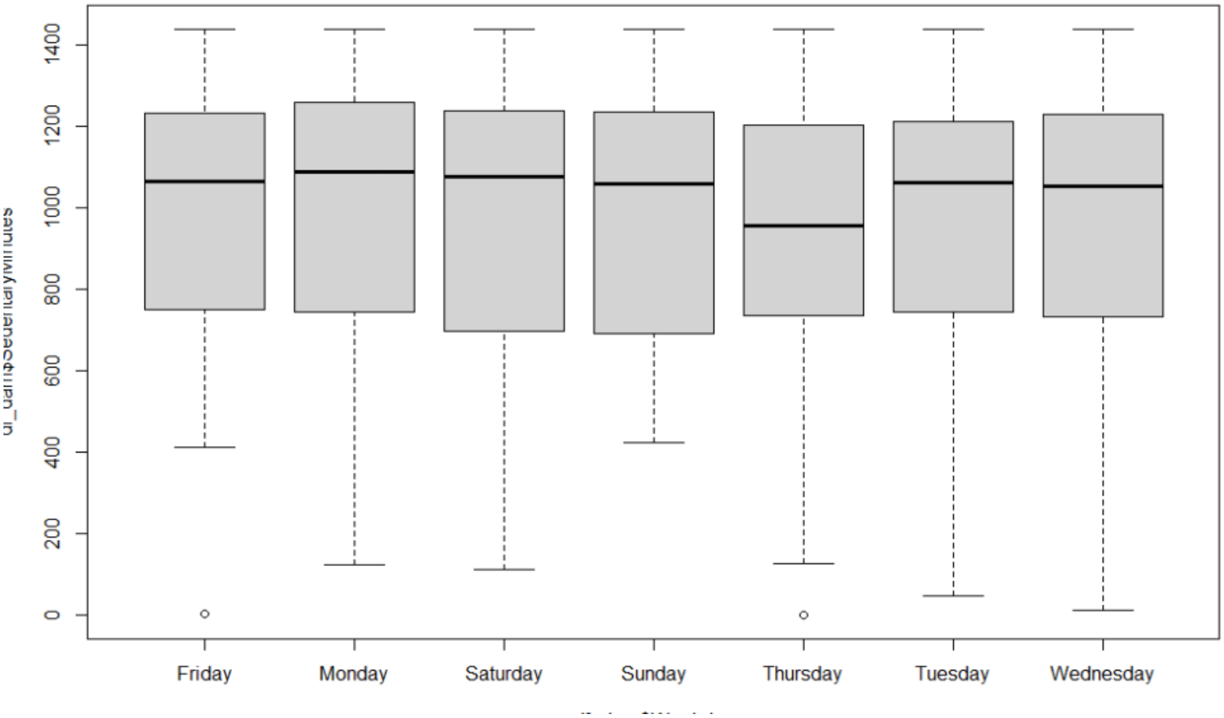
## People taking lots of rest ….
## ARE they elderly ? ARE THEY SICK? Device is OK?



# Plot   Sedantry minutes wrt Date

```
ggplot(data = df_dam, aes(x = df_dam$ActivityDate, y = df_dam$SedentaryMinutes)) +
  geom_point() +
  labs(x = "Date",
       y = "Sedantry Minutes",
       title = "Daily sedantry minutes")
```


Daily sedantry minutes



```
##  Lets finds Ids  which  are  dispalying high sedantry

k=df_dam[df_dam$SedentaryMinutes > 1200,]
table(k$Id)   ## IDS with more than 20 hr sedantry minutes
```

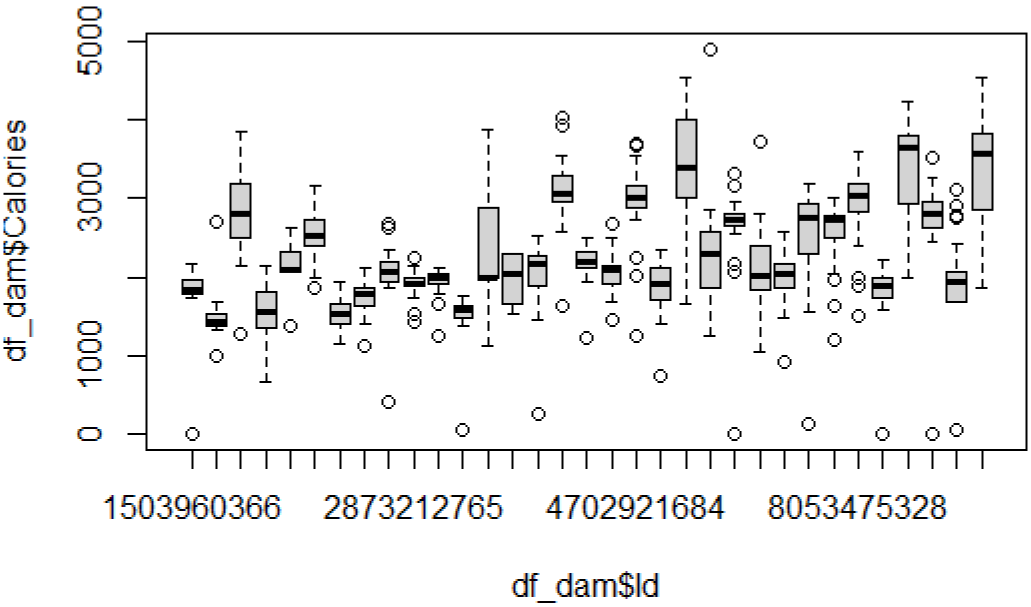
```
1503960366 1624580081 1644430081 1844505072 1927972279 2022484408 2320127002
2873212765
         3         28         14         21         25          3         18
4
3372868164 4020332650 4057192912 4319703577 4388161847 4445114986 4558609924
4702921684
         1         23          3          4          2          3          3
2
5577150313 6117666160 6290855005 6775888955 7007744171 7086361926 8053475328
8253242879
         3          7          8         21          3          4         11
17
8583815059 8792009665 8877689391
        22         13          4
```

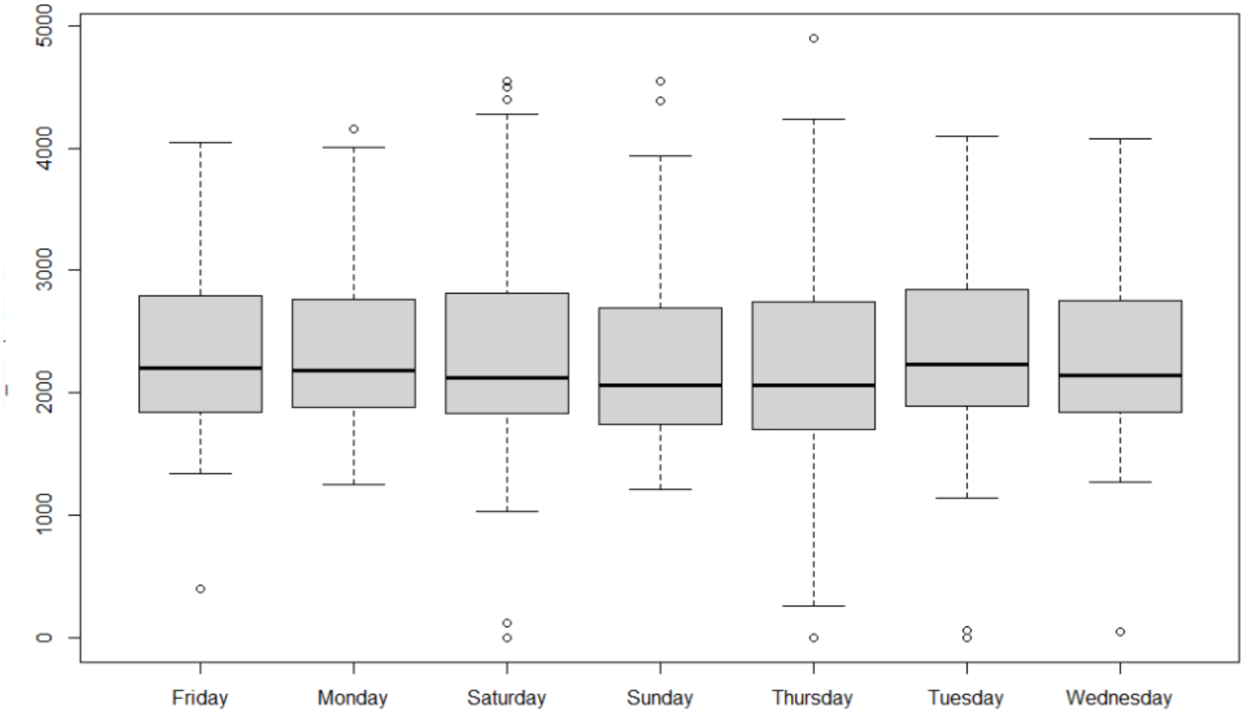
```
###   This is again problematic.
```

```
### Next variable I chose is Calorie

boxplot(df_dam$Calories)
boxplot(df_dam$Calories ~ df_dam$Weekday)
boxplot(df_dam$Calories ~ df_dam$Id)
boxplot(df_dam$Calories ~ df_dam$ActivityDate)

table(df_dam$Calories > 1000)

table(df_dam$Calories < 3000)
```

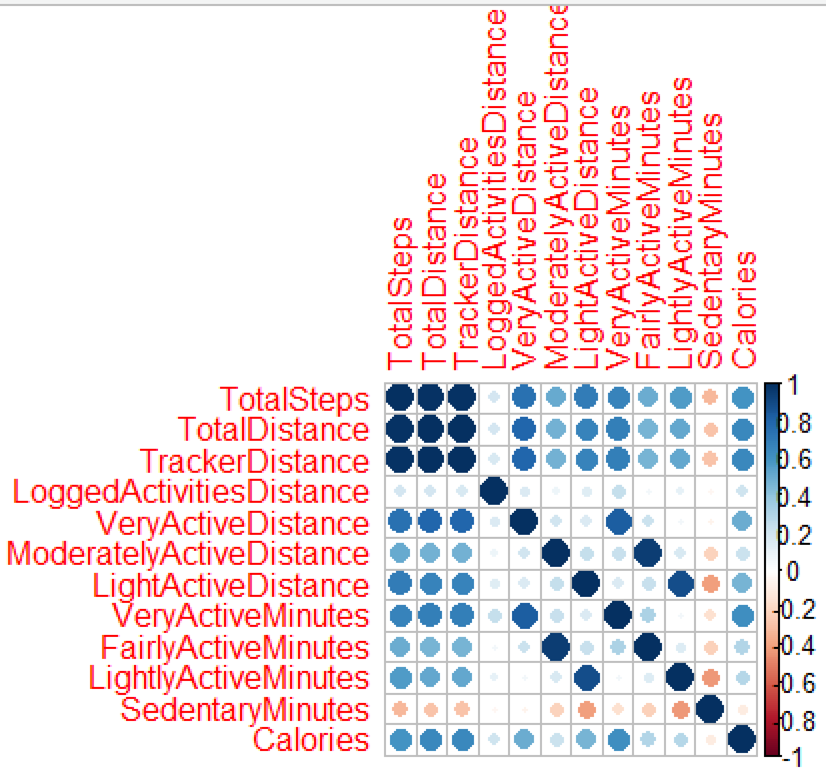
```
FALSE  TRUE
   12   928
> table(df_dam$Calories < 3000)

FALSE  TRUE
  153   787
```

```
### 787  instances  of going over 3000 calorie, 12 instances of going below 1000
```


df_dam$Calories
0
1000
3000
5000
1503960366
2873212765
4702921684
8053475328
df_dam$Id


0
1000
2000
3000
4000
5000
Friday
Monday
Saturday
Sunday
Thursday
Tuesday
Wednesday

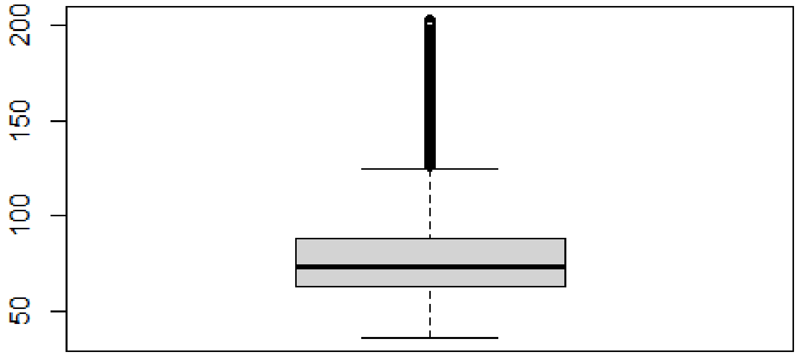
```
library("corrplot")

str(df_dam)
df_dam2 = subset(df_dam, select = -c(ActivityDate,Id,SedentaryActiveDistance))


str(df_dam2)
cor(df_dam2)
library(corrplot)
corrplot(cor(df_dam2))
```



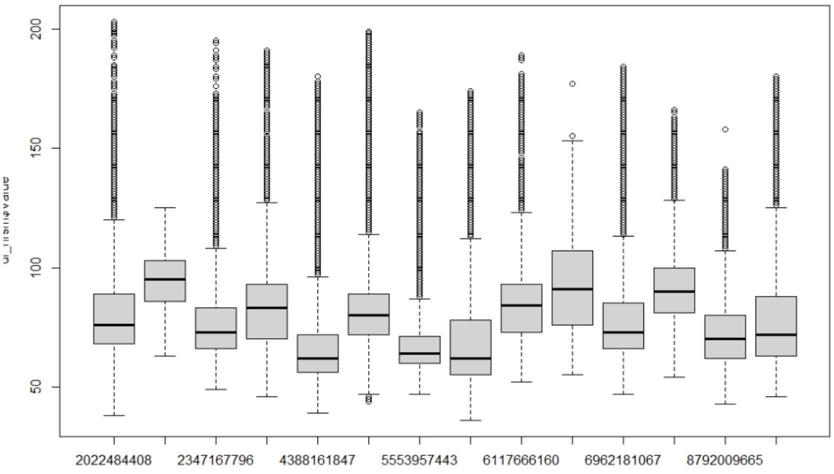
```
###  There is nothing surprising in correlation.



## Very Active minutes

table(df_dam$VeryActiveMinutes> 60)  #112
table(df_dam$VeryActiveMinutes> 120) #14
table(df_dam$VeryActiveMinutes> 180)  #6
table(df_dam$VeryActiveMinutes> 240)  #0
```

##moving to heart rate data

## Plot box plot



## How can heart rate less than 50?   More than 150 ?



##Plot Heart rate with respect to IDS

### Only 14 IDS …..why?

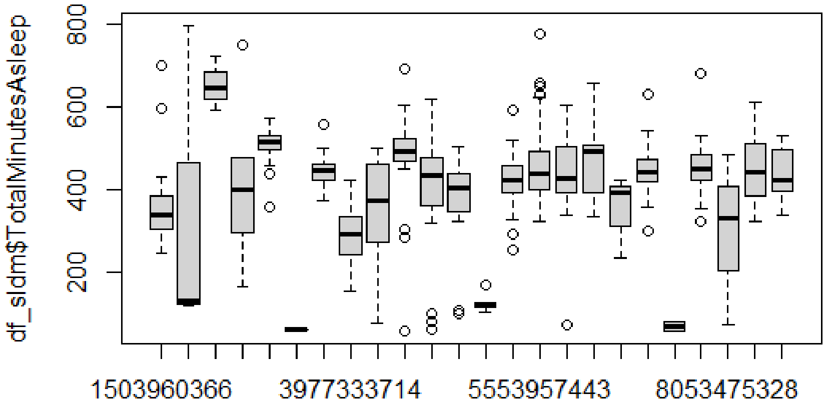
```
## Why Heart rate so low and high?


table(df_hrsm$Value < 50)   ##39134
## Brady cardia  or someone is very fit

table(df_hrsm$Value > 101 ) ##254130

39134/2483658      #1.5%
254130/2483658      #10.2%
```

```
##Sleep data
table(df_sldm$TotalSleepRecords)
unique(df_sldm$Id)    ## 24 ID
```

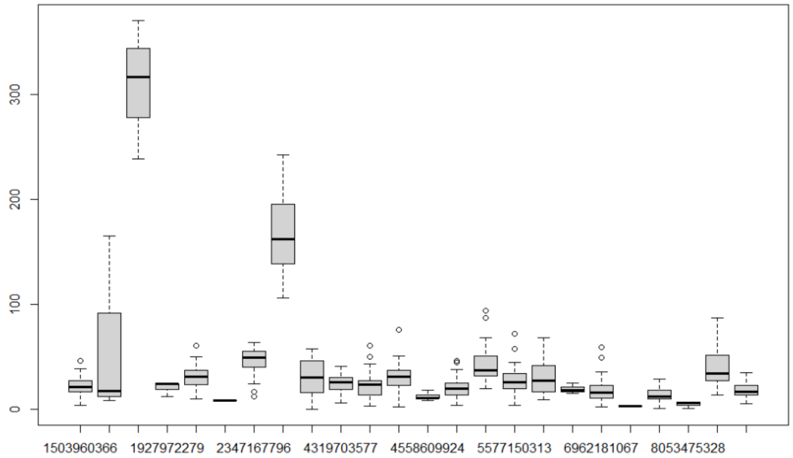
```
  1    2    3 
367   43    3 
```

```
[1] 1503960366 1644430081 1844505072 1927972279 2026352035 2320127002 2347167
796
 [8] 3977333714 4020332650 4319703577 4388161847 4445114986 4558609924 470292
1684
[15] 5553957443 5577150313 6117666160 6775888955 6962181067 7007744171 708636
1926
[22] 8053475328 8378563200 8792009665
```

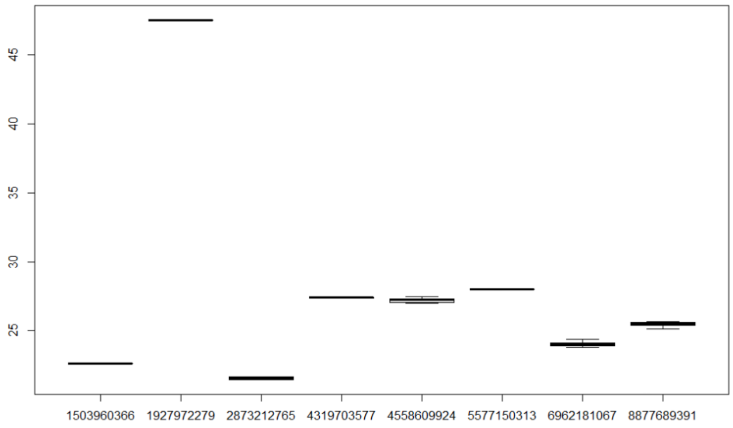
```
boxplot(df_sldm$TotalMinutesAsleep ~ df_sldm$Id)
```



```
table(df_sldm$TotalMinutesAsleep <  60 )   ##2
table(df_sldm$TotalMinutesAsleep < 120)   ##15
table(df_sldm$TotalMinutesAsleep < 180)  #22
table(df_sldm$TotalMinutesAsleep <  240) #30
table(df_sldm$TotalMinutesAsleep <  300)  ##50

table(df_sldm$TotalMinutesAsleep >  480)   ##115

## Too little sleepers ....alert message should be sent, stay away from
##dangerous jobs

##Oversleepers   more than 8 hrs ..should be sent ....



##Plotting diffrence between sleep time and total time in bed

x=df_sldm$TotalTimeInBed - df_sldm$TotalMinutesAsleep
boxplot(x ~ df_sldm$Id)
```

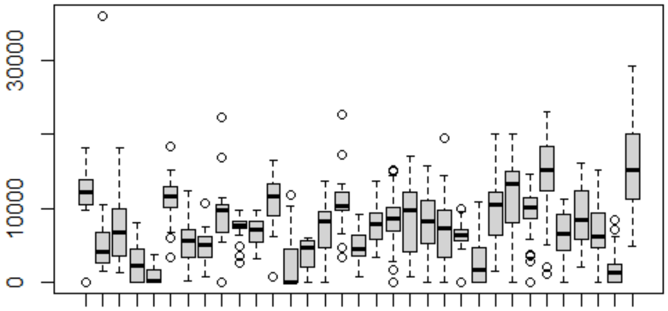


```
# we see Two ID have huge difference.  Spending 2-4 hrs in bed , without sleeping.



## Weight data

summary(df_wli)

unique(df_wli$Id)    #8 IDS

##  BMI is more relevant as only Weight data without height and age is irrelevant
#boxplot(df_wli$BMI ~ df_wli$Id)
```

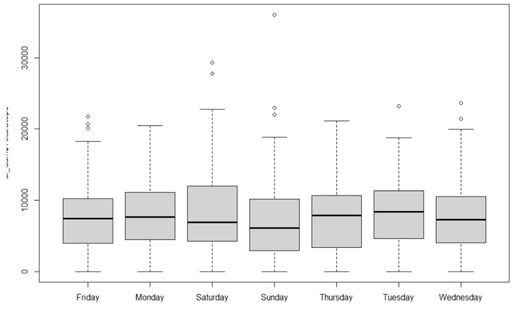


```
##   BMI   MIN : 21.45   MAX  47.54
##  18 --  24    Normal  BMI  range
table(df_wli$BMI >  24) ##44
## Several  values in FAT  as NA, better sensor/reporting
```

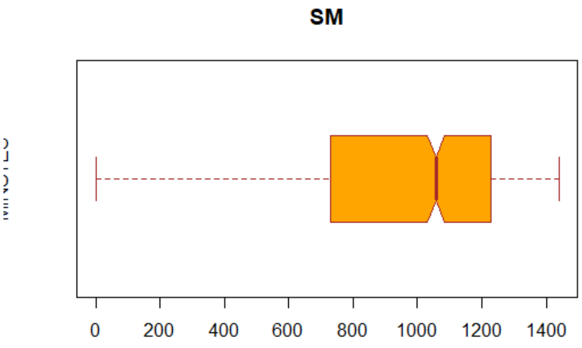
1. **Analyze the distribution of Variables.**



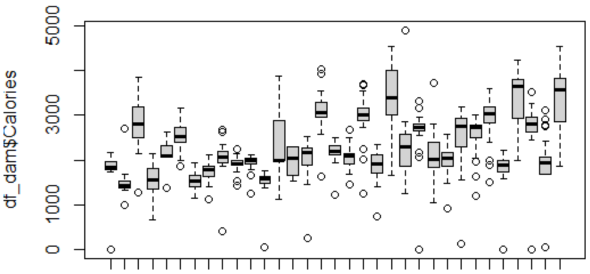
Plot1: Number of Steps taken in a day Vs Ids.

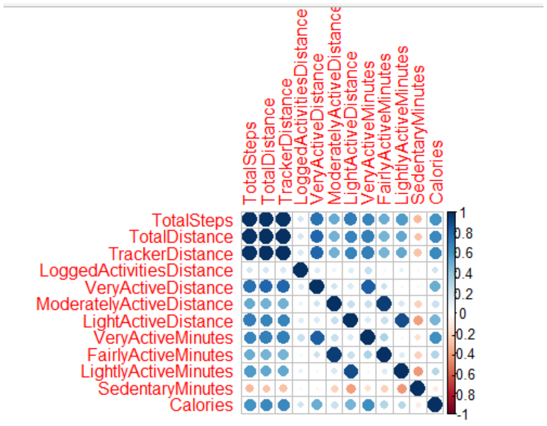


Plot2: Sedentary Minutes vs Day.



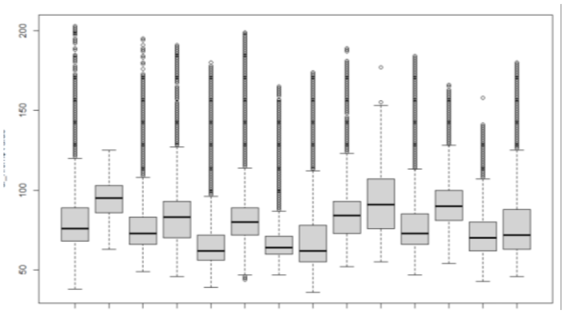
Plot3: Sedentary Minutes distribution

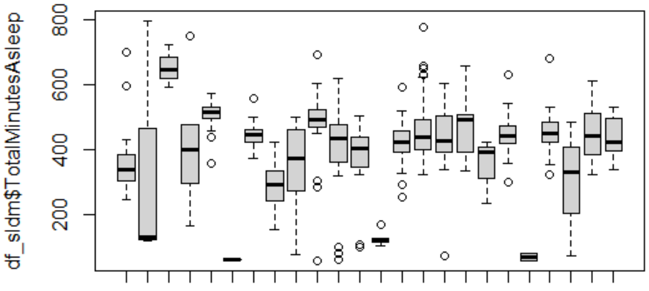




Calories Vs IDs

Correlation Plot





Heartbeat Vs ID

Sleep minutes vs IDs