encounter <- read.csv("/Users/liyi/Desktop/RP/day\_in\_the\_life.csv")

# How many unique subjects in this data set

index <- duplicated(encounter[,1])  
encounter\_subset <- encounter[!index,]  
(nrow(encounter\_subset))

## [1] 1307

# How many males / females are in the data set?

# # The number of male in this dataset

(nrow(encounter\_subset[encounter\_subset$Gender == 1,]))

## [1] 514

# # The number of female in this dataset

(nrow(encounter\_subset[encounter\_subset$Gender == 2,]))

## [1] 793

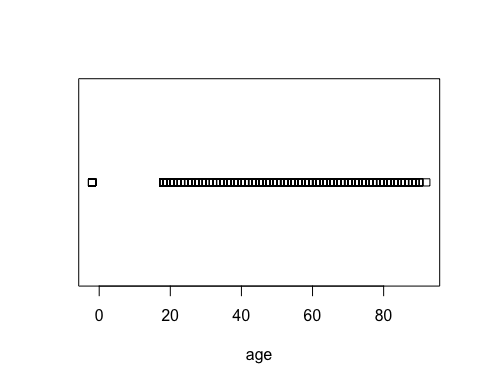
# Age distribution for all of the subjects

# # mean and sd of age

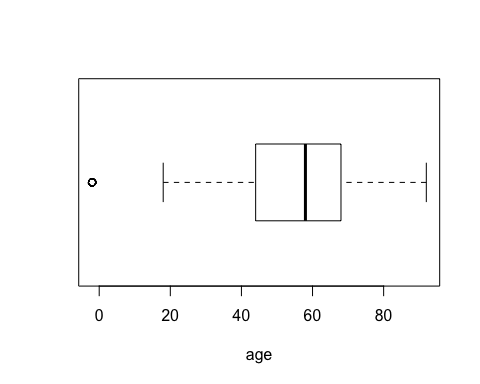
meanage = mean(encounter\_subset$Age,rm.na = TRUE)  
sdage = sd(encounter\_subset$Age)

## plots of raw data of age

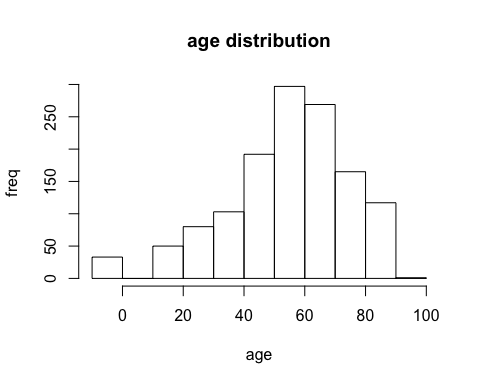
stripchart(encounter\_subset$Age, xlab = "age")



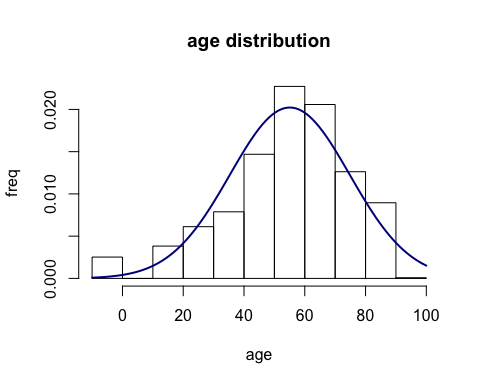
boxplot(encounter\_subset$Age,horizontal = TRUE,xlab = "age")

 ## histgram

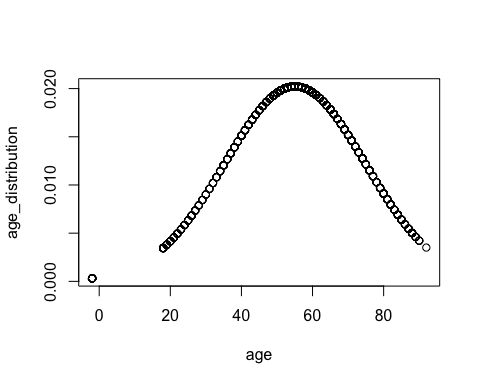
hist(encounter\_subset$Age,xlab = "age",ylab = "freq",main = "age distribution")



hist(encounter\_subset$Age,xlab = "age",ylab = "freq",main = "age distribution",freq = FALSE)  
curve(dnorm(x,   
 mean=meanage,   
 sd=sdage),   
 add=TRUE,   
 col="darkblue",   
 lwd=2)

 ## Plot distribution

age\_distribution <- dnorm(encounter\_subset$Age, mean(encounter\_subset$Age), sd = sd(encounter\_subset$Age))  
plot(encounter\_subset$Age,age\_distribution, xlab = "age")

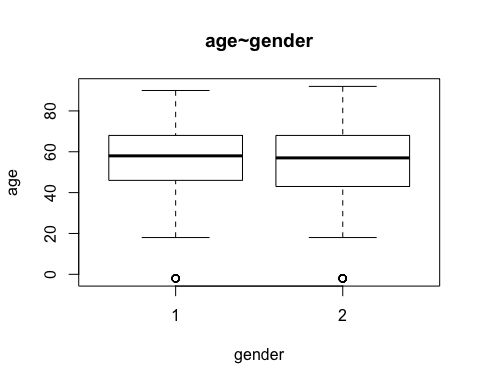
 ## The summary of age distribution

summary(encounter\_subset$Age)

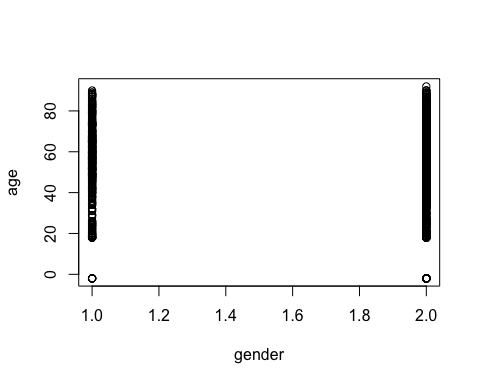
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.00 44.00 58.00 55.08 68.00 92.00

# Age distribution bases on gender

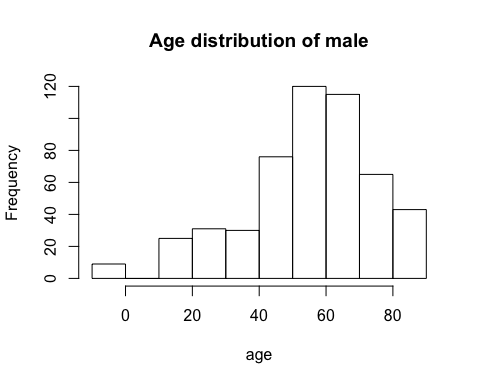
par(mfrow=c(1,1))  
boxplot(Age~Gender,data = encounter\_subset,xlab = "gender",ylab = "age",main = "age~gender")



plot(Age~Gender,data = encounter\_subset,xlab = "gender",ylab = "age")



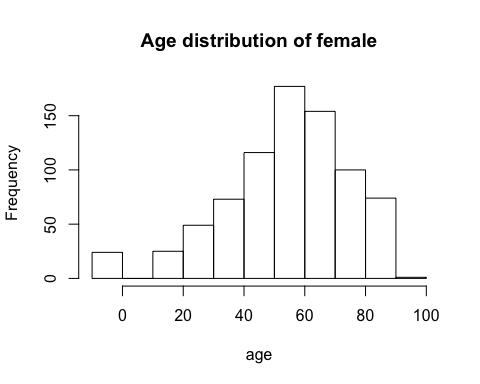
hist(encounter\_subset$Age[which(encounter\_subset$Gender==1)],xlab = "age", main = "Age distribution of male")



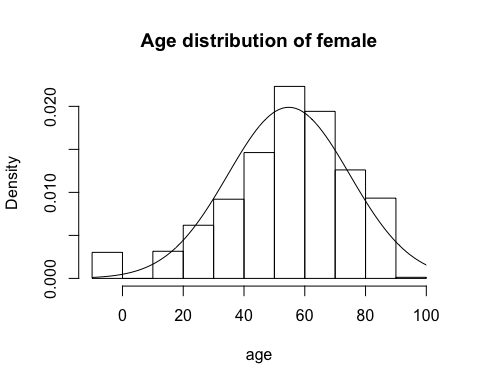
hist(encounter\_subset$Age[which(encounter\_subset$Gender==1)],freq = FALSE, xlab = "age", main = "Age distribution of male")  
curve(dnorm(x,  
 mean = mean(encounter\_subset$Age[which(encounter\_subset$Gender==1)]),  
 sd = sd(encounter\_subset$Age[which(encounter\_subset$Gender==1)])),  
 add = TRUE)



hist(encounter\_subset$Age[which(encounter\_subset$Gender==2)],xlab = "age", main = "Age distribution of female")



hist(encounter\_subset$Age[which(encounter\_subset$Gender==2)],freq = FALSE, xlab = "age", main = "Age distribution of female")  
curve(dnorm(x,  
 mean = mean(encounter\_subset$Age[which(encounter\_subset$Gender==2)]),  
 sd = sd(encounter\_subset$Age[which(encounter\_subset$Gender==2)])),  
 add = TRUE)

 ## The summary of age based on gender

summary(encounter\_subset$Age[which(encounter\_subset$Gender==1)])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.0 46.0 58.0 55.6 68.0 90.0

summary(encounter\_subset$Age[which(encounter\_subset$Gender==2)])

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -2.00 43.00 57.00 54.74 68.00 92.00

# The distribution of the number of locations visited by each study participant

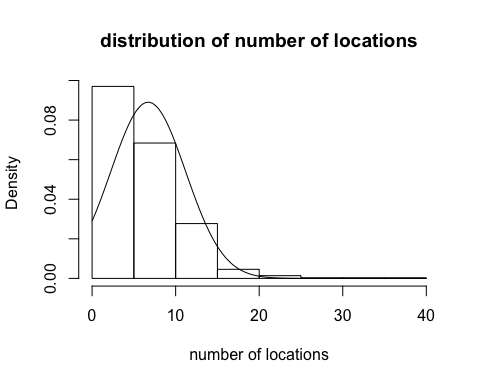
num\_of\_locations <- data.frame(num = numeric(0))  
for(subject in encounter\_subset$Subject){  
 #print(subject)  
 num\_of\_locations <- rbind(num\_of\_locations, nrow(encounter[which(encounter$Subject == subject), ]))  
 #nrow(encounter[which(encounter$Subject == subject), ])  
}  
nrow(num\_of\_locations)

## [1] 1307

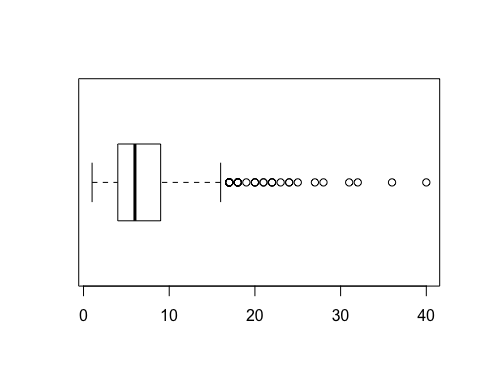
hist(num\_of\_locations$X1L,xlab = "number of locations",main = "distribution of number of locations")



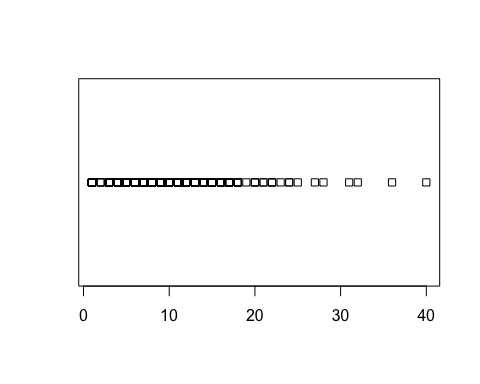
hist(num\_of\_locations$X1L,freq = FALSE,xlab = "number of locations",main = "distribution of number of locations")  
curve(dnorm(x,  
 mean = mean(num\_of\_locations$X1L),  
 sd = sd(num\_of\_locations$X1L)),  
 add = TRUE)



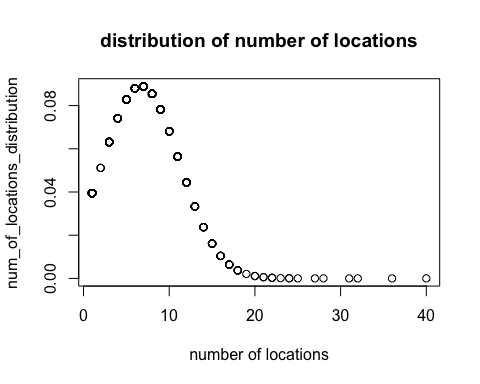
boxplot(num\_of\_locations$X1L,horizontal = TRUE)



stripchart(num\_of\_locations$X1L)



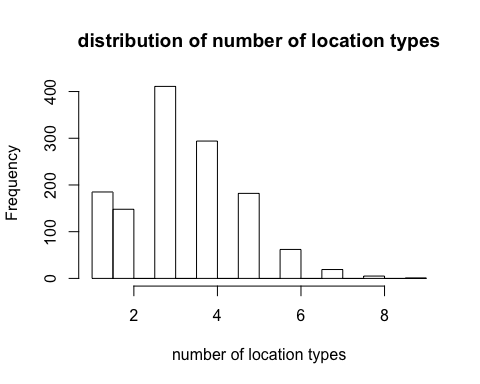
num\_of\_locations\_distribution <- dnorm(num\_of\_locations$X1L, mean=mean(num\_of\_locations$X1L), sd = sd(num\_of\_locations$X1L))  
plot(num\_of\_locations$X1L,num\_of\_locations\_distribution,xlab = "number of locations",main = "distribution of number of locations")

 #What is the distribution of number of different *types* of locations visited by each study participant?

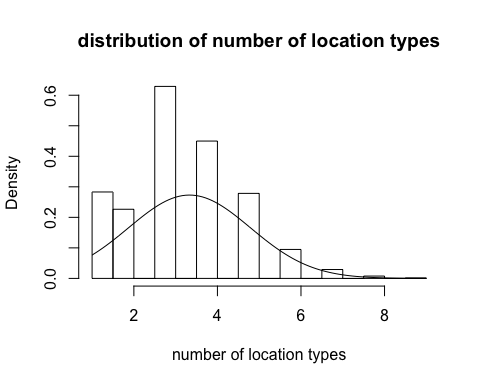
index <- duplicated(encounter[,c(1,6)])  
encounter\_subset\_locationType <- encounter[!index,]  
encounter\_subset\_locationType[c(1,2,3,4,5),]

## Subject Age Gender Location.ID Description  
## 1 49 65 1 10001 NULL  
## 2 51 35 2 10002 NULL  
## 3 51 35 2 10003 NULL  
## 4 51 35 2 10004 The Reject Shop  
## 15 53 -2 2 10015 NULL  
## Location.Type Suburb  
## 1 Home Greenvale  
## 2 Home Sunbury  
## 3 Private Transport (private or rented car, motorbike) NULL  
## 4 Retail and hospitality (bars, cafes, shops, hair dressing, etc.) Sunbury  
## 15 Home Kew East  
## Post.Code Lat Lon Time.of.Arrival Time.of.Departure Duration  
## 1 3059 -37.640616 144.8834858 2013/2/24 7:00 2013/2/24 23:00 960  
## 2 3429 -37.582356 144.70768 2013/2/24 7:00 2013/2/24 10:30 210  
## 3 NULL NULL NULL 2013/2/24 10:30 2013/2/24 10:35 5  
## 4 3429 -37.580002 144.727982 2013/2/24 10:35 2013/2/24 10:55 20  
## 15 3102 -37.79788 145.05404 2013/3/9 7:00 2013/3/9 9:00 120

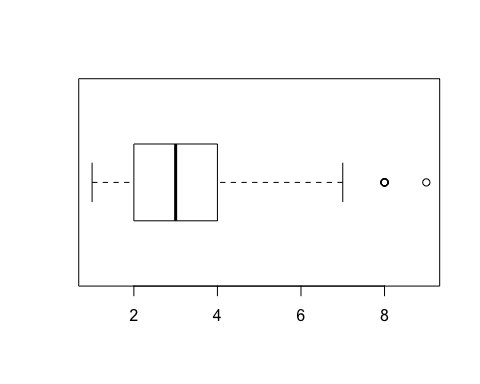
num\_of\_locationTypes <- data.frame(num = numeric(0))  
for(subject in encounter\_subset$Subject){  
 num\_of\_locationTypes <- rbind(num\_of\_locationTypes, nrow(encounter\_subset\_locationType[which(encounter\_subset\_locationType$Subject == subject), ]))  
}  
hist(num\_of\_locationTypes$X1L,xlab = "number of location types",main = "distribution of number of location types")



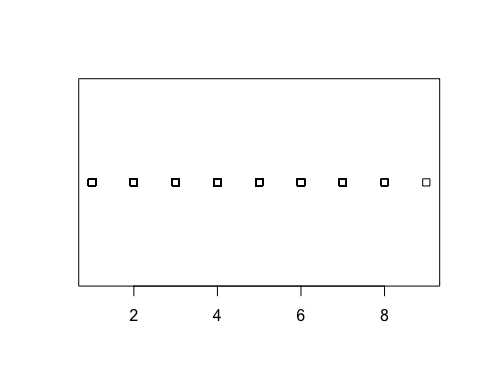
hist(num\_of\_locationTypes$X1L,freq = FALSE,xlab = "number of location types",main = "distribution of number of location types")  
curve(dnorm(x,  
 mean = mean(num\_of\_locationTypes$X1L),  
 sd = sd(num\_of\_locationTypes$X1L)),  
 add = TRUE)



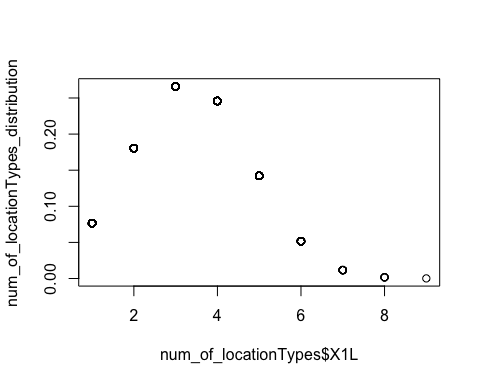
boxplot(num\_of\_locationTypes$X1L,horizontal = TRUE)



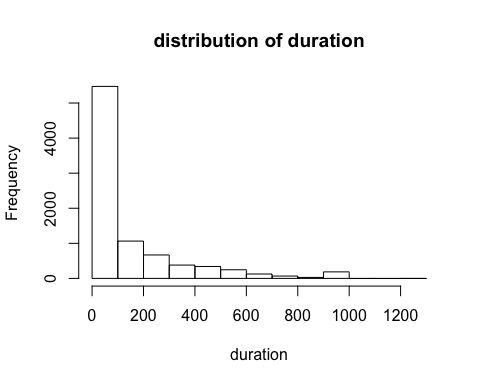
stripchart(num\_of\_locationTypes$X1L)



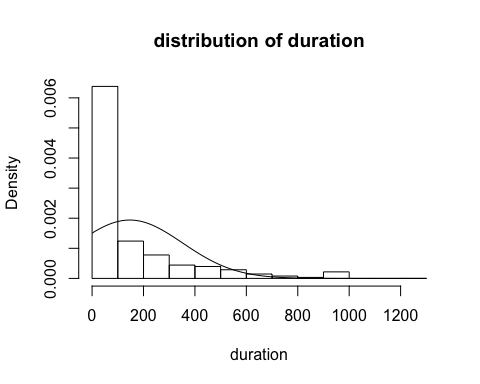
num\_of\_locationTypes\_distribution <- dnorm(num\_of\_locationTypes$X1L, mean=mean(num\_of\_locationTypes$X1L), sd = sd(num\_of\_locationTypes$X1L))  
plot(num\_of\_locationTypes$X1L,num\_of\_locationTypes\_distribution)

 # What is the distribution of times spent at locations?

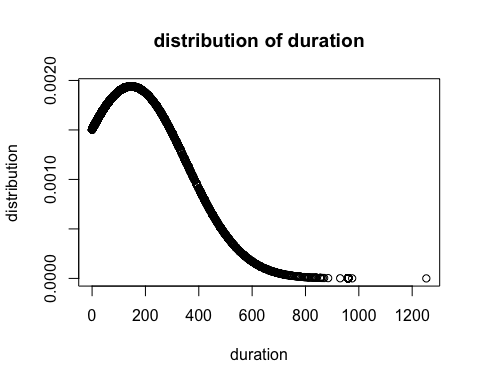
encounter\_duration <- read.csv("/Users/liyi/Desktop/RP/day\_in\_the\_life.csv")  
encounter\_with\_duration\_subset <- encounter\_duration[which(encounter\_duration$Duration!="NULL"), ]  
hist(as.numeric(as.character(encounter\_with\_duration\_subset$Duration)),xlab = "duration",main = "distribution of duration")



hist(as.numeric(as.character(encounter\_with\_duration\_subset$Duration)),freq = FALSE,xlab = "duration",main = "distribution of duration")  
meanvalue <- mean(as.numeric(as.character(encounter\_with\_duration\_subset$Duration)),na.rm=TRUE)  
sdvalue <- sd(as.numeric(as.character(encounter\_with\_duration\_subset$Duration)),na.rm=TRUE)  
curve(dnorm(x,  
 mean = meanvalue,  
 sd = sdvalue),  
 add = TRUE)



# encounter\_with\_duration\_subset\_single <- na.omit(encounter\_with\_duration\_subset$Duration)  
duration\_location\_distribution <- dnorm(na.omit(as.numeric(as.character(encounter\_with\_duration\_subset$Duration))), mean = meanvalue, sd = sdvalue)  
plot(na.omit(as.numeric(as.character(encounter\_with\_duration\_subset$Duration))),duration\_location\_distribution,xlab = "duration",ylab = "distribution",main = "distribution of duration")



Charactering trajectories

I rewrite “Public Transport”, “Private Transport” and “Car Journey” to “M”, which means movement, and use the first character to represent the location type. When two or more adjacent location types are the same of one subject, I merge them to one. To reduce the number of different trajectories, I set a threshold 60 mins for the duration of one staying at a place. Finally, I concatenate location types to a trajectory. There are totally 1307 subjects and 337 different sequences.