Data Wrangling I Activity

QBS Bootcamp 2025

In Class Activity

Working in groups, define a new variable for hypertension in our original dataset (randomData). Here we will define hypertension as systolic blood pressure over 130 or a diastolic blood pressure over 80. Plot the distribution in age for individuals with and without hypertension using boxplots.

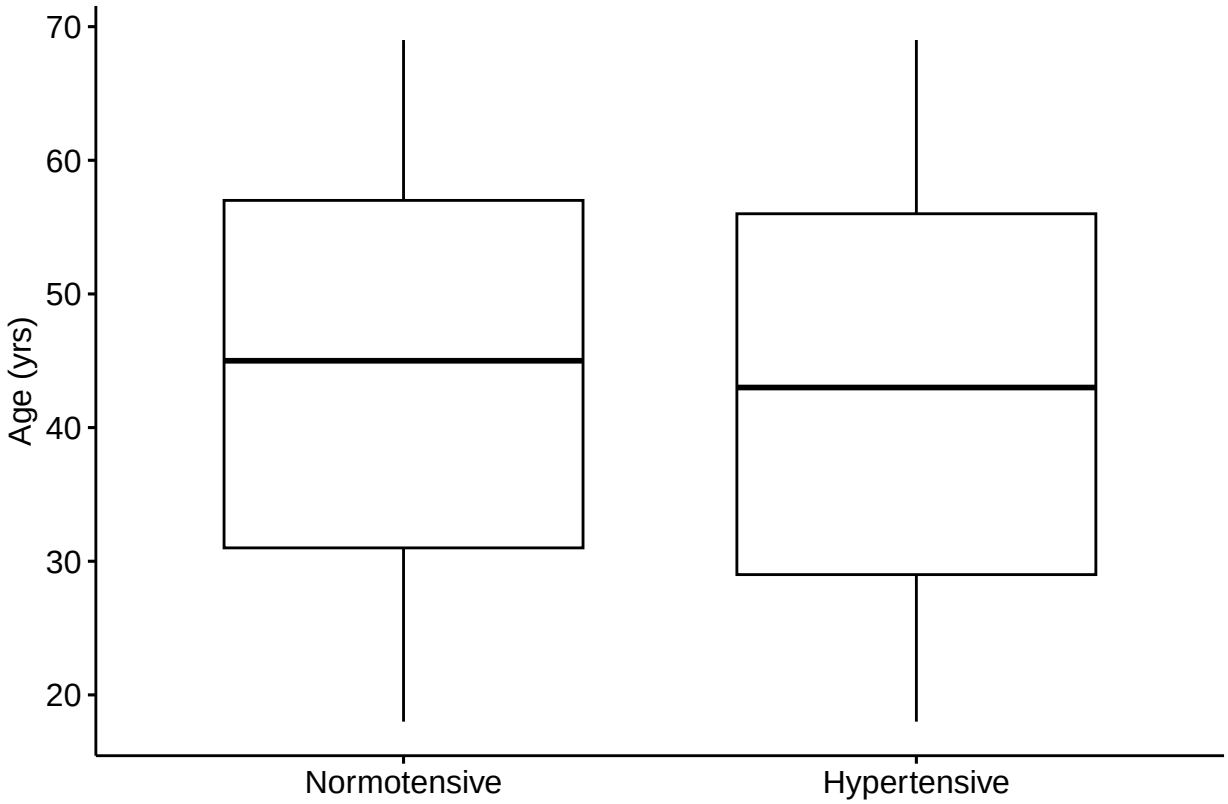
Use the *melt* function to generate boxplots of the distribution of systolic and diastolic blood pressure in hypertensive vs. normotensive individuals (color should be based on hypertension status).

Use the *dcast* function to generate a table summarizing the mean age, systolic, and diastolic BP for males and females, separately, with and without hypertension. Your table should have 4 rows. Order your table output such that it lists values for normotensive individuals first and hypertensive individuals second.

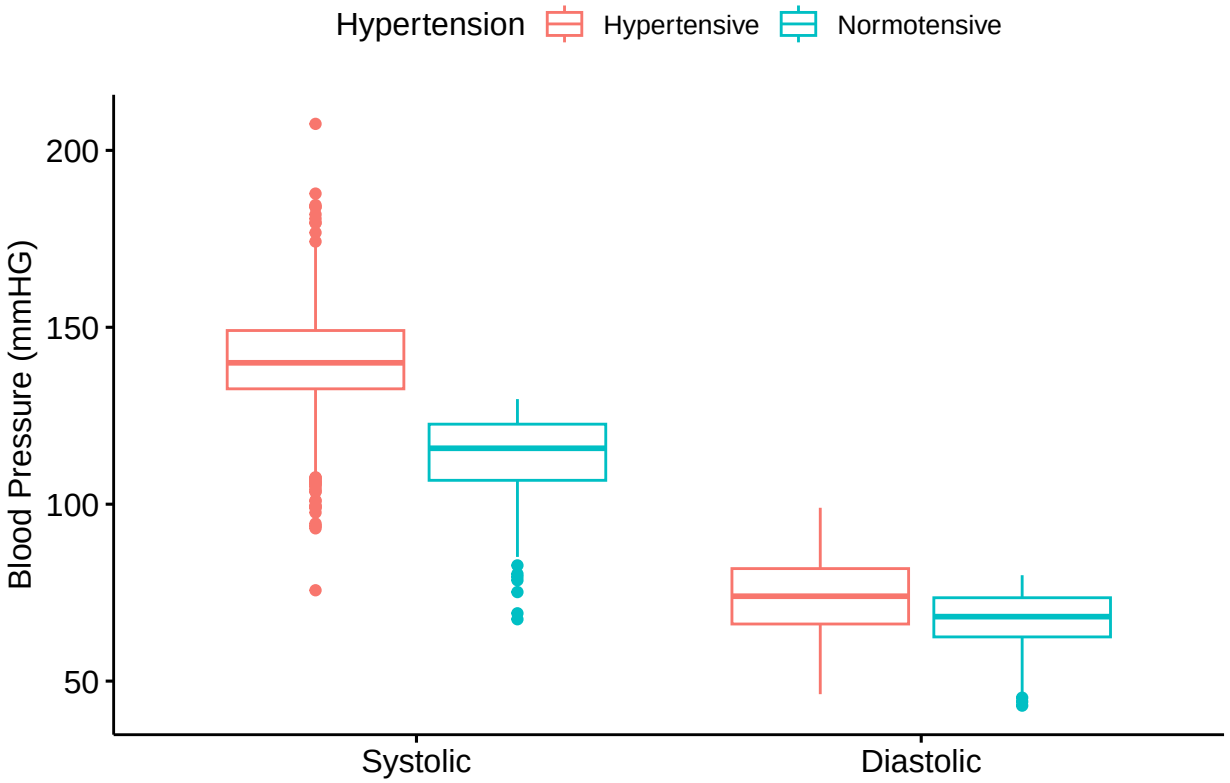
Sample Solution

Remember: There are many ways to solve any problem in R. As long as you get the same end product, what you did is likely just as valid as what I did.

```
# Set a random seed
set.seed(103)

# Define a data frame with our randomly generated data
randomData <- data.frame('SubjectID' = seq(1:1000),
                         'systolicBP' = rnorm(n = 1000,mean = 128,sd = 20),
                         'diastolicBP' = rnorm(n = 1000,mean = 71,sd = 10),
                         'Age' = trunc(runif(n = 1000,min = 18,max = 70)),
                         'Male' = rbinom(n = 1000,size = 1,prob = 0.5))
# Define a new factor variable from an old binary
randomData$BiologicalSex <- factor(ifelse(randomData$Male == 1,'Male','Female'))

# Define variable for hypertension
randomData$Hypertension <- ifelse(randomData$systolicBP > 130 | randomData$diastolicBP > 80,
                                  'Hypertensive','Normotensive')
# Plot distribution of age based on hypertension status
ggpubr::ggboxplot(randomData,x = 'Hypertension',y = 'Age',xlab = '',ylab = 'Age (yrs)')
```




```
# Generate long format data
longData <- reshape2::melt(randomData,id.vars = c('SubjectID','Age','Male','BiologicalSex','Hypertensio
                           value.name = 'BP',variable.name = 'BP.Type')
# Reformat names for bp type
longData$BP.Measure <- ifelse(longData$BP.Type == 'systolicBP','Systolic','Diastolic')

# Plot BP by hypertensive classification
ggpubr::ggboxplot(longData,x = 'BP.Measure',y = 'BP',color = 'Hypertension',
                  xlab = '',ylab = 'Blood Pressure (mmHG)')
```




```
# Regenerate long data to also include age
longData2 <- reshape2::melt(randomData,id.vars = c('SubjectID','Male','BiologicalSex','Hypertension'),
                            value.name = 'Cont.Value',variable.name = 'Var.Name')
# Generate summary table
sumTab <- reshape2::dcast(longData2,formula = BiologicalSex + Hypertension ~ Var.Name,
                          fun.aggregate = mean,value.var = 'Cont.Value')
sumTab[order(sumTab$Hypertension,decreasing = T),]
```

```
##   BiologicalSex Hypertension systolicBP diastolicBP      Age
## 2        Female Normotensive   114.7668    67.23993 43.64706
## 4          Male Normotensive   112.6688    67.64147 44.62162
## 1        Female Hypertensive   138.8948    73.64332 43.98425
## 3          Male Hypertensive   141.1750    73.96578 42.28713
```