1. **Before the homework, I read the file:**

Read a csv file:

bodyfat <- read.csv('bodyfat.csv')

1. **Read in the bodyfat.csv data file and generate a variable "bodycat" to categorize body fat into the three categories above. Make sure all 252 observations are categorized into either athlete, average, or obese.**

bodyfat$bodycat[bodyfat$bodyfat < 14 ] <- 'Athlete'

bodyfat$bodycat[bodyfat$bodyfat >= 14 & bodyfat$bodyfat <= 25 ] <- 'Average'

bodyfat$bodycat[bodyfat$bodyfat > 25 ] <- 'Obese'

bodyfat$bodycat

Table

Description automatically generated

**Figure 1**

1. **Using summarize to identify the four height quartiles, create a new variable "htcat" to categorize height into "short", "below average", "above average", and "tall".**

summary(bodyfat$height)



**Table 1**

bodyfat$htcat[bodyfat$height < 68.25 ] <- ‘Short’

bodyfat$htcat[bodyfat$height >= 68.25 & bodyfat$height < 70.00] <- ‘Below average’

bodyfat$htcat[bodyfat$height >= 70.00 & bodyfat$height < 72.25] <- ‘Above average’

bodyfat$htcat[bodyfat$height >=72.25 ] <- ‘Tall’

bodyfat$htcat

Table

Description automatically generated

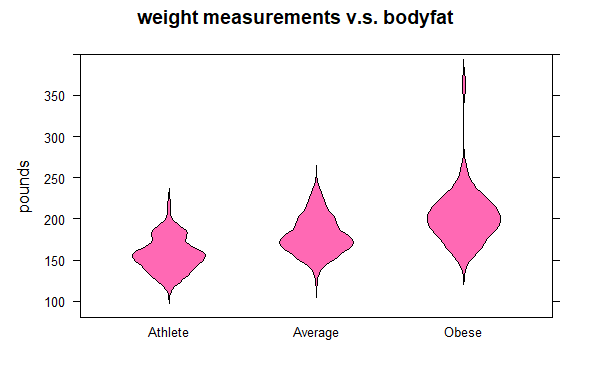
**Figure 2**

1. **Create a violin plot of weight separated by bodycat. Make sure your plots show up in some kind of order that makes sense. In complete sentences, summarize what the violin plots tell you. Are the weights evenly distributed within a range for all categories? Do athletes tend to be within a certain weight range? You may use summarize() to help you. Rough estimates are also okay.**

bwplot(weight~bodycat, data = bodyfat, panel = panel.violin,

ylab = 'pounds', col = 'hotpink',

ylim = c(80,400), main = "weight measurements v.s. bodyfat")



**Figure 3**

The violin plot tells me that when body fat is higher, people will be heavier. The median of three different group categories slightly increases from left to right and all the groups will be distributed in their own range. That’s why we always have a stereotype that people who are heavier have higher body fat. Besides that, the ‘Obese’ group has a taller range from ~120-370 pounds. That’s strange because I do not find it from the dataset, so I do my investigation into the differences between violin and boxplot.

**3. Extended:**

I consider the Figure 3 did not provide me with enough information, so I search the information to learn further details. Luckily, I can use the ‘function’ code to combine violin and box plot. As a result, the final violin plot will look like as the below:

**bwplot (weight~bodycat, data = bodyfat, horizontal=FALSE,**

**panel = function(..., box.ratio)**

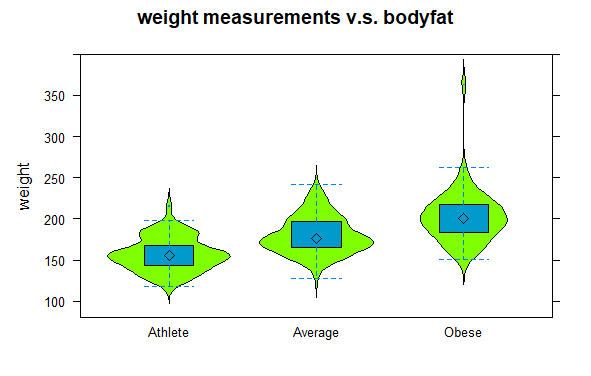
**{panel.violin(..., col = 'chartreuse', varwidth = T, box.ratio = 5)**

**panel.bwplot(..., col='black', cex=1, pch=23, fill="deepskyblue3", box.ratio = 0.5)},**

**par.settings = list(box.rectangle=list(col='black'),**

**plot.symbol = list(pch=22, cex = 0.1)), ylim= c(80,400),**

**main = "weight measurements v.s. bodyfat")**

****

**Figure 3.2**

1. **Create a stem-and-leaf plot for weight. Be sure to find an appropriate scale for the data.**

library(multcomp)

stem(bodyfat$bodyfat, scale = 1)

Text

Description automatically generated

**Figure 4**

stem(bodyfat$bodyfat, scale = 2)

**A picture containing table

Description automatically generated**

**A picture containing table

Description automatically generated**

**Figure 5**

When compared with **scale = 2** and **scale =1**, the two stem-and-leaf plots can tell us what the true number is. However**, scale = 2** is obviously more complicated than **scale = 1**, and too many datasets in a plot are not easily accepted by people at the first spot. Therefore, scale = 1 is apparently better.

1. **A. Create overlapping histograms of neck for the three body categories. For this exercise, do not use the default breaks. Use breaks that you think make sense. Remember to make sure that the first histogram is an appropriate window size so when you "add" the other graphs, those histograms aren't cut-off. Also, remember to use the same break widths for overlapped histograms.**

hist(bodyfat$neck[bodyfat$bodycat == 'Athlete'], breaks = seq(25,55,2),

col = rgb(1,0,1,0.25), xlab = " Neck(cm)",

ylim = c(-1,50), main = 'Histogram of neck circumcenter difference:\ngroup by bodyfat')

hist(bodyfat$neck[bodyfat$bodycat == 'Average'], breaks = seq(25,55,2),

col = rgb(0,1,1,0.25), add = T)

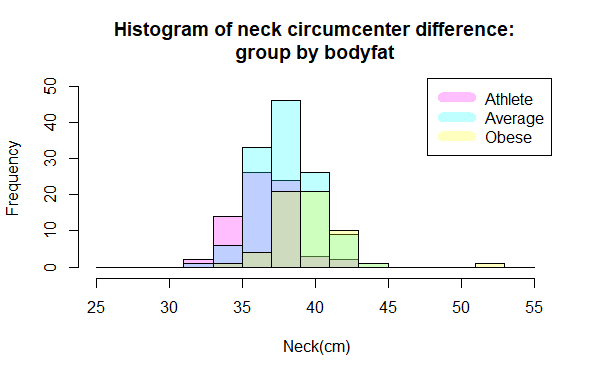
hist(bodyfat$neck[bodyfat$bodycat == 'Obese'], breaks = seq(25,55,2),

col = rgb(1,1,0,0.25), add = T)

legend ('topright', c('Athlete', 'Average', 'Obese'),

col = c(rgb(1,0,1,0.25), rgb(0,1,1,0.25), rgb(1,1,0,0.25)),

lwd = 10)



**Figure 6**

Chart, histogram, box and whisker chart

Description automatically generated

**Figure 6.1**

**B. In the same window, add 3 density plots—1 for each body category. Do not use the default bandwidth. Use a bandwidth that you think makes sense. (Note, you will need to have used freq = F in your histograms.**

1. **'Athete’:**

hist(bodyfat$neck[bodyfat$bodycat == 'Athlete'], breaks = seq(25,55,2),

col = rgb(1,0,1,0.25), xlab = "Neck(cm)", ylim = c(0,0.2),

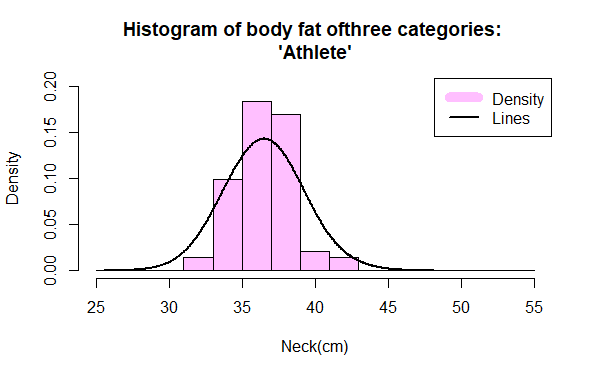
freq = F, main = "Histogram of body fat ofthree categories: \n'Athlete'")

dens1 <- density (bodyfat$neck[bodyfat$bodycat == 'Athlete'], bw = 2)

lines(dens1, lwd=2)

legend ('topright', c('Density', 'Lines'),

col = c(rgb(1,0,1,0.25), 'Black'), lwd = c(10,2))

****

**Figure 7**

1. **'Average':**

**hist(bodyfat$neck[bodyfat$bodycat == 'Average'], breaks = seq(25,55,2),**

**col = rgb(0,1,1,0.25), xlab = "Neck(cm)", ylim = c(0,0.2),**

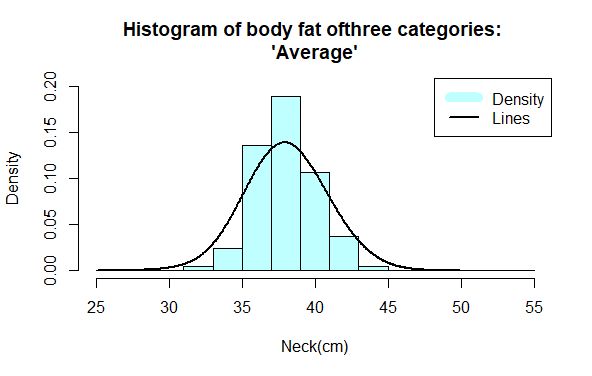
**freq = F, main = "Histogram of body fat ofthree categories: \n'Average'")**

**dens2 <- density (bodyfat$neck[bodyfat$bodycat == 'Average'], bw = 2)**

**lines(dens2, lwd=2)**

**legend ('topright', c('Density', 'Lines'),**

**col = c(rgb(0,1,1,0.25), 'Black'), lwd = c(10,2))**

****

**Figure 8**

1. **'Obese':**

**hist(bodyfat$neck[bodyfat$bodycat == 'Obese'], breaks = seq(25,55,2),**

**col = rgb(1,1,0,0.25), xlab = "Neck(cm)", ylim = c(0,0.2),**

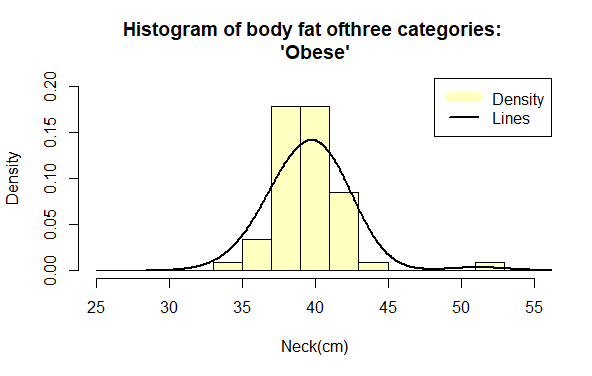
**freq = F, main = "Histogram of body fat ofthree categories: \n'Obese'")**

**dens3 <- density (bodyfat$neck[bodyfat$bodycat == 'Obese'], bw = 2)**

**lines(dens3, lwd=2)**

**legend ('topright', c('Density', 'Lines'),**

**col = c(rgb(1,1,0,0.25), 'Black'), lwd = c(10,2))**

****

**Figure 9**

**C. In complete sentences, compare neck circumference across the three body categories using your histograms and density plots.**

My first anticipation before analyzing the data is that the higher body fat person will be fatter. However, when I check the combined histogram plot with the neck circumcenter, it did not have an apparent trend. (Figure 6). Luckily, when I put separated them as a cartoon (Figure 6.1), it clearly tells me that the neck circumcenter migrates when people have higher body fat. When the category is ‘Athlete’, it has the most count in 35~37; the category is ‘Average’ has the mode in 37-39 and the category is ‘Obese’ has the mode in 39~41.

Besides that, when the category is ‘Obese’, it has an outliner in the neck circumcenter (~50cm). Hence, I write a code to check the person’s data.

heavilyobese <- bodyfat[(bodyfat$neck > 50),]

It tells me this person will be categorized as the ‘Tall’, which is rational (The taller person will have higher body fat).

The density plot (**Figure 7~9**) shows all the three categories are normal distributions. (Not right skew or left skew).

1. **Repeat Question 5 for abdomen.**

**6A:**

hist(bodyfat$abdomen[bodyfat$bodycat == 'Athlete'], breaks = seq(65,155,4),

col = rgb(1,0,1,0.25), xlab = "Abdomen(cm)",

ylim = c(-1,30), main = 'Histogram of body fat of three categories')

hist(bodyfat$abdomen[bodyfat$bodycat == 'Average'], breaks = seq(65,155,4),

col = rgb(0,1,1,0.25), add = T)

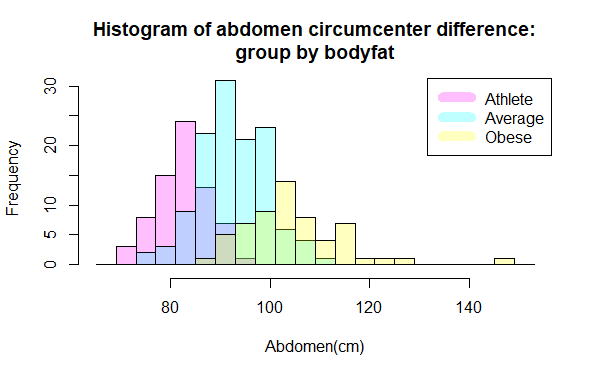
hist(bodyfat$abdomen[bodyfat$bodycat == 'Obese'], breaks = seq(65,155,4),

col = rgb(1,1,0,0.25), add = T)

legend ('topright', c('Athlete', 'Average', 'Obese'),

col = c(rgb(1,0,1,0.25), rgb(0,1,1,0.25), rgb(1,1,0,0.25)),

lwd = 10)



**Figure 10**

**6B:**

par(mfrow= c(1,3))

hist(bodyfat$abdomen[bodyfat$bodycat == 'Athlete'], breaks = seq(65,155,4),

col = rgb(1,0,1,0.25), xlab = "Abdomen(cm)", ylim = c(0,0.1),

freq = F, main = "Histogram of body fat ofthree categories: \n'Atlete'")

dens4 <- density (bodyfat$abdomen[bodyfat$bodycat == 'Athlete'], bw = 2)

lines(dens4, lwd=2)

legend ('topright', c('Density', 'Lines'),

col = c(rgb(1,0,1,0.25), 'Black'), lwd = c(10,4))

##6B. 'Average'

hist(bodyfat$abdomen[bodyfat$bodycat == 'Average'], breaks = seq(65,155,4),

col = rgb(0,1,1,0.25), xlab = "Abdomen(cm)", ylim = c(0,0.1),

freq = F, main = "Histogram of body fat ofthree categories: \n'Average'")

dens5 <- density (bodyfat$abdomen[bodyfat$bodycat == 'Average'], bw = 2)

lines(dens5, lwd=2)

legend ('topright', c('Density', 'Lines'),

col = c(rgb(0,1,1,0.25), 'Black'), lwd = c(10,4))

##6B. 'Obese'

hist(bodyfat$abdomen[bodyfat$bodycat == 'Obese'], breaks = seq(65,155,4),

col = rgb(1,1,0,0.25), xlab = "Abdomen(cm)", ylim = c(0,0.1),

freq = F, main = "Histogram of body fat ofthree categories: \n'Obese'")

dens6 <- density (bodyfat$abdomen[bodyfat$bodycat == 'Obese'], bw = 1)

lines(dens6, lwd=2)

legend ('topright', c('Density', 'Lines'),

col = c(rgb(1,1,0,0.25), 'Black'), lwd = c(10,4))

Chart, diagram, histogram

Description automatically generated

**Figure 11**

**6C:**

**Figure 10** meets my anticipation: people who have higher body fat have more cross-sections on their bodies (the mode moves to the right when people have more body fat). We can also observe the same result in the density plot (**Figure 11**).

There also can find the outliner in the ‘Obese’ category, which tells me that this person is very tall, heavy, and normal, not a person who has a special thick neck but normal in the abdomen.

~The End~