STAT 1293 Assignment 2

Gordon Lu

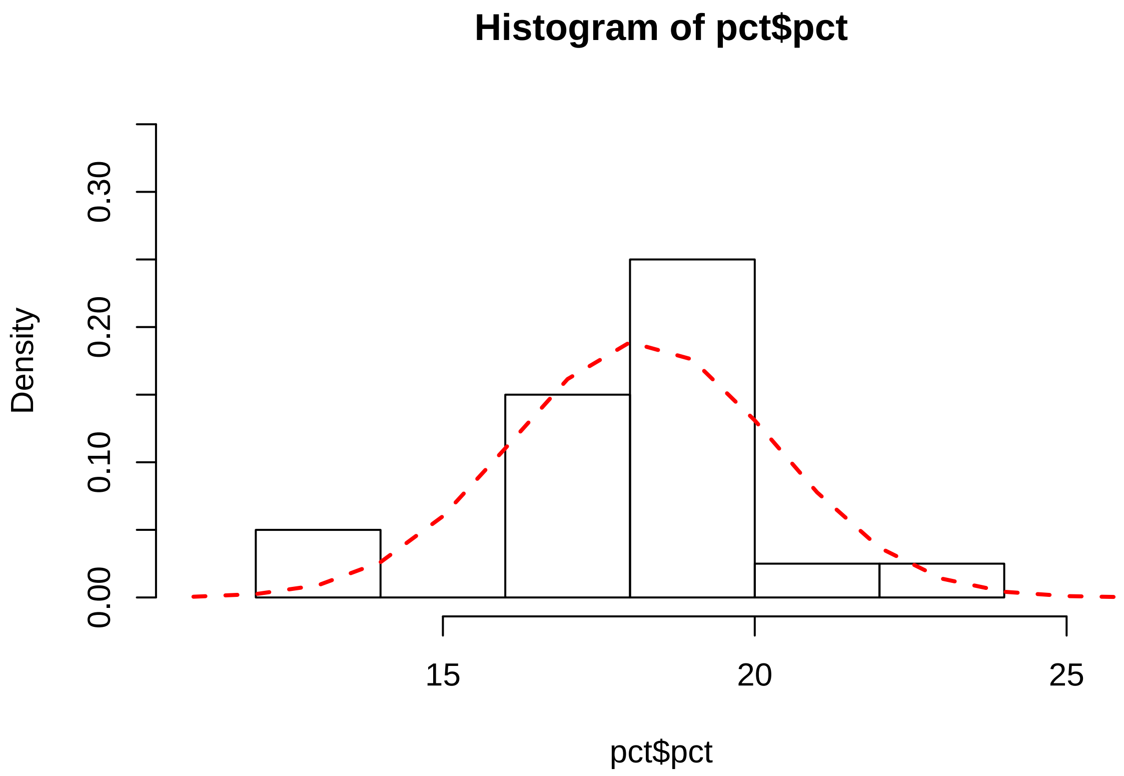
7/7/2020

Problem 1: Bad weather, bad tips? (20 points)

1a) Create a histogram of the percent of tips (pct). Overlay the histogram with a normal density curve (red, dashed). (4 points)

Solution:

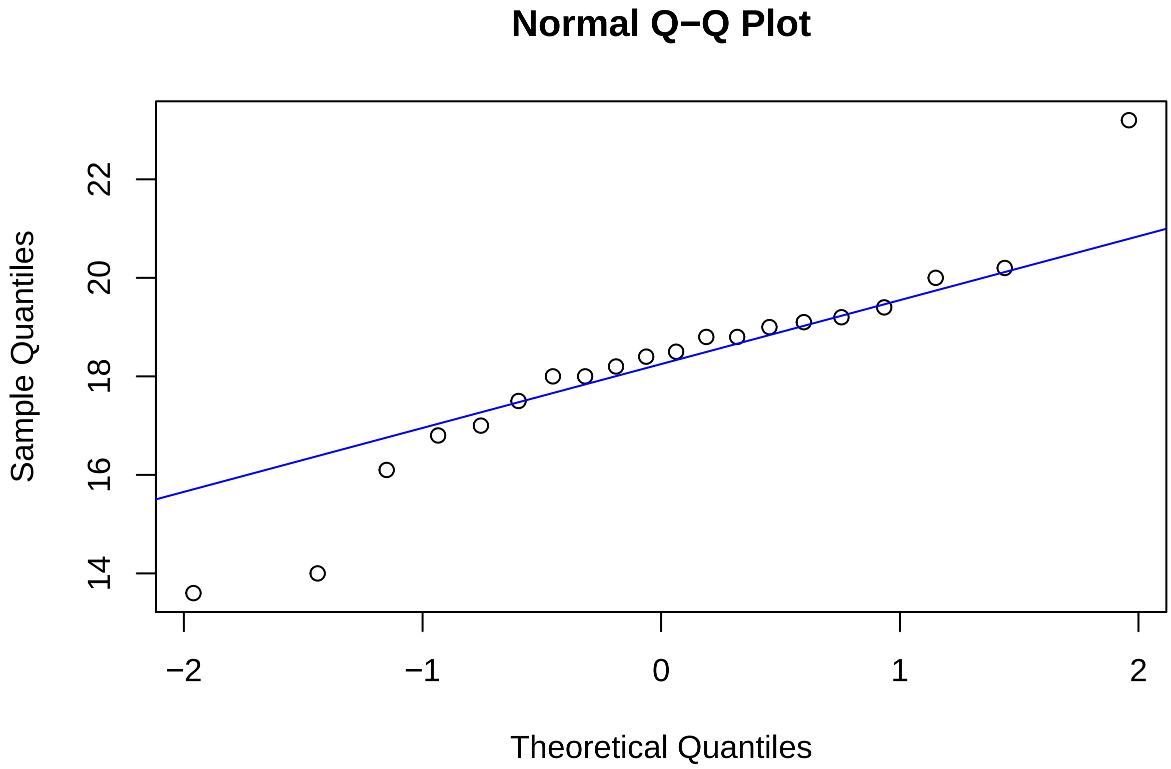
```r
pct <- read.table("C:/Users/gordo/Desktop/tip3.txt", header = TRUE) #read in tip3
hist(pct$pct, freq = F, xlim = c(11, 26), ylim = c(0, 0.35))
y = seq(11, 26)
lines(y, dnorm(y, mean(pct$pct), sd(pct$pct)), col = 2, lwd = 2, lty = 2)
```



1b) Create a Q-Q plot of pct. Add a reference line (blue, solid). (4 points)

Solution:

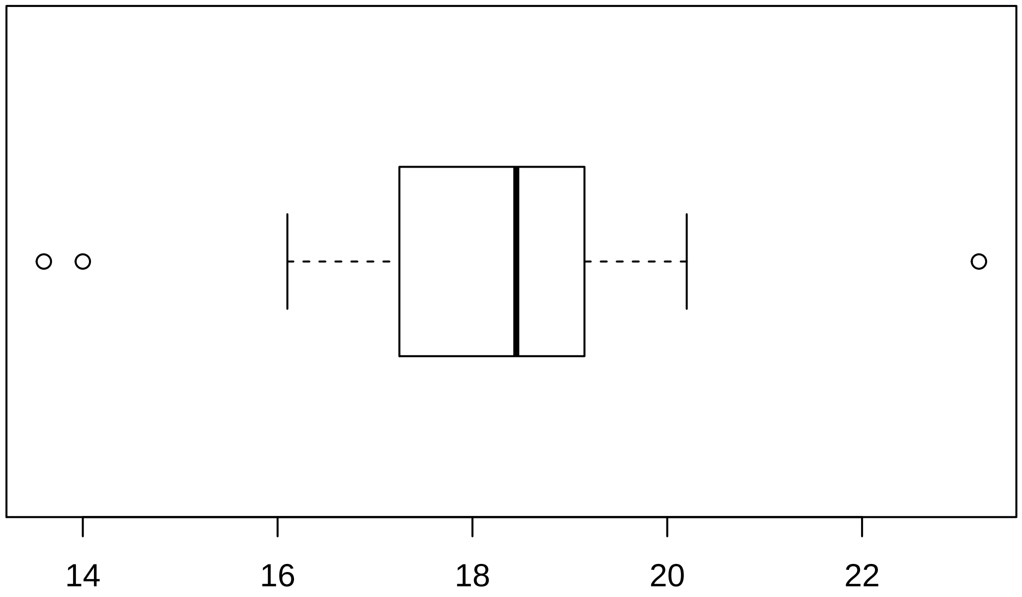
```
qqnorm(pct$pct)
qqline(pct$pct, col = 4)
```



1c) Create a horizontal boxplot of pct. Are there any outliers? (4 points)

Solution:

```
boxplot(pct$pct, horizontal = T)
```



Yes, there appear to be 3 outliers. 2 on the lower end, and 1 on the upper end.

1d) Calculate the 5-number summary (Min, Q1, Median, Q3, and Max) of pct. (4 points)

Solution:

```
summary(pct$pct)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   13.60   17.38   18.45   18.19   19.12   23.20
```

1e) Calculate the mean and standard deviation of `pct`. (4 points)

Solution:

```
mean(pct$pct)
```

```
## [1] 18.19
```

```
sd(pct$pct)
```

```
## [1] 2.104606
```

Problem 2: E.coli in swimming areas (10 points)

2a) Create a stem plot of the E. coli levels (Ecolil). (3 points)

Solution:

```
ecoli <- read.table("C:/Users/gordo/Desktop/ecoli.txt", header = TRUE) #read in ecoli
stem(ecoli$Ecolil)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 01112223345559
##   1 | 9
##   2 | 9
```

The data seems to be right-skewed, this is apparent through the two upper outliers, `19` and `29`.

2b) Split the each stem to two stems. (3 points)

Solution:

```
stem(ecoli$Ecolil, 2)
```

```
##
##   The decimal point is 2 digit(s) to the right of the |
##
##   0 | 0111222334
##   0 | 5559
##   1 |
##   1 | 9
##   2 |
##   2 | 9
```

2c) Calculate the descriptive statistics using `summary()`. (4 points)

Solution:

```
summary(ecoli$Ecolil)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   14.72   31.25   56.28   47.75  291.00
```

Any evidence of skewness? Yes. The minimum value of `Ecolil` is 1, which is pretty far away from the `25%` quantile, and the rest of the data. Also, notice that the maximum value of `Ecolil` is `291` is pretty far away from the `mean`, and `75%` quantile and thus is a good indication that the data is right-skewed.

Problem 3: Daily Activity and Obesity (30 points)

3a) Transform the variable Group to a factor, using labels Lean and Obese. (4 points)

Solution:

```
obese <- read.table("C:/Users/gordo/Desktop/obese.txt", header = TRUE) #read in obese
obese <- transform(obese, Group = factor(Group, labels = c("Lean", "Obese")))
#transform Group variable as factor-type vector
obese
```

```
##    Group Subject   Stand     Sit     Lie
## 1   Lean       1 511.100 370.300 555.500
## 2   Lean       2 607.925 374.512 450.650
## 3   Lean       3 319.212 582.138 537.362
## 4   Lean       4 584.644 357.144 489.269
## 5   Lean       5 578.869 348.994 514.081
## 6   Lean       6 543.388 385.312 506.500
## 7   Lean       7 677.188 268.188 467.700
## 8   Lean       8 555.656 322.219 567.006
## 9   Lean       9 374.831 537.031 531.431
## 10  Lean      10 504.700 528.838 396.962
## 11 Obese      11 260.244 646.281 521.044
## 12 Obese      12 464.756 456.644 514.931
## 13 Obese      13 367.138 578.662 563.300
## 14 Obese      14 413.667 463.333 532.208
## 15 Obese      15 347.375 567.556 504.931
## 16 Obese      16 416.531 567.556 448.856
## 17 Obese      17 358.650 621.262 460.550
## 18 Obese      18 267.344 646.181 509.981
## 19 Obese      19 410.631 572.769 448.706
## 20 Obese      20 426.356 591.369 412.919
```

3b) Calculate and compare the descriptive statistics of standing time (stand) between the two groups. (4 points)

Solution:

```
stand.lean <- obese$Stand[obese$Group == "Lean"] #get standing time for those in Lean group
stand.obese <- obese$Stand[obese$Group == "Obese"] #get standing time for those in the Obese group.

summary_stand.lean <- summary(stand.lean) #store summary in a variable
summary_stand.obese <- summary(stand.obese)

print("Summary of Standing Time by Lean is:") #print out
```

```
## [1] "Summary of Standing Time by Lean is:"
```

```
summary_stand.lean
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   319.2   506.3   549.5   525.8   583.2   677.2
```

```
print("Summary of Standing Time by Obese is:")
```

```
## [1] "Summary of Standing Time by Obese is:"
```

```
summary_stand.obese
```

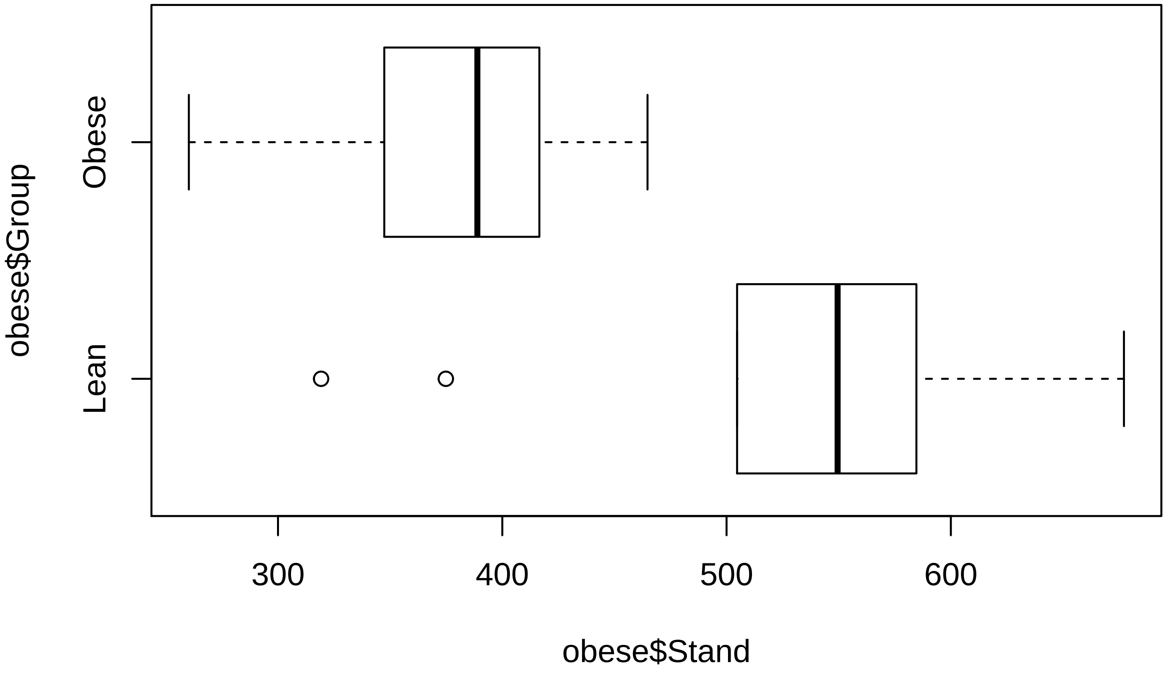
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   260.2   350.2   388.9   373.3   415.8   464.8
```

We can see that on average, those in the `Lean` group typically spend more time standing than those in the `Obese` group.

3c) Create a horizontal side-by-side boxplot for the standing time of the two groups. (3 points)

Solution:

```
boxplot(obese$Stand ~ obese$Group, horizontal = TRUE) #Generate side-by-side boxplot
```



```
#standing time desribed by groups.
```

3d) Compare the descriptive statistics between the two groups with regard to `sit` and `lie`. (4 points)

Solution:

```
sit.lean <- obese$Sit[obese$Group == "Lean"] #get sitting time for those in Lean group
sit.obese <- obese$Sit[obese$Group == "Obese"] #get sitting time for those in the Obese group.

summary_sit.lean <- summary(sit.lean) #store summary in a variable
summary_sit.obese <- summary(sit.obese)

print("Summary of Sitting Time by Lean is:") #print out
```

```
## [1] "Summary of Sitting Time by Lean is:"
```

```
summary_sit.lean
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   268.2   351.0   372.4   407.5   493.0   582.1
```

```
print("Summary of Sitting Time by Obese is:")
```

```
## [1] "Summary of Sitting Time by Obese is:"
```

```
summary_sit.obese
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   456.6   567.6   575.7   571.2   613.8   646.3
```

```
lie.lean <- obese$Lie[obese$Group == "Lean"] #get lying time for those in Lean group
lie.obese <- obese$Lie[obese$Group == "Obese"] #get lying time for those in the Obese group.

summary_lie.lean <- summary(lie.lean) #store summary in a variable
summary_lie.obese <- summary(lie.obese)

print("Summary of Lying Time by Lean is:") #print out
```

```
## [1] "Summary of Lying Time by Lean is:"
```

```
summary_lie.lean
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   397.0   473.1   510.3   501.6   535.9   567.0
```

```
print("Summary of Lying Time by Obese is:")
```

```
## [1] "Summary of Lying Time by Obese is:"
```

```
summary_lie.obese
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   412.9   451.8   507.5   491.7   519.5   563.3
```

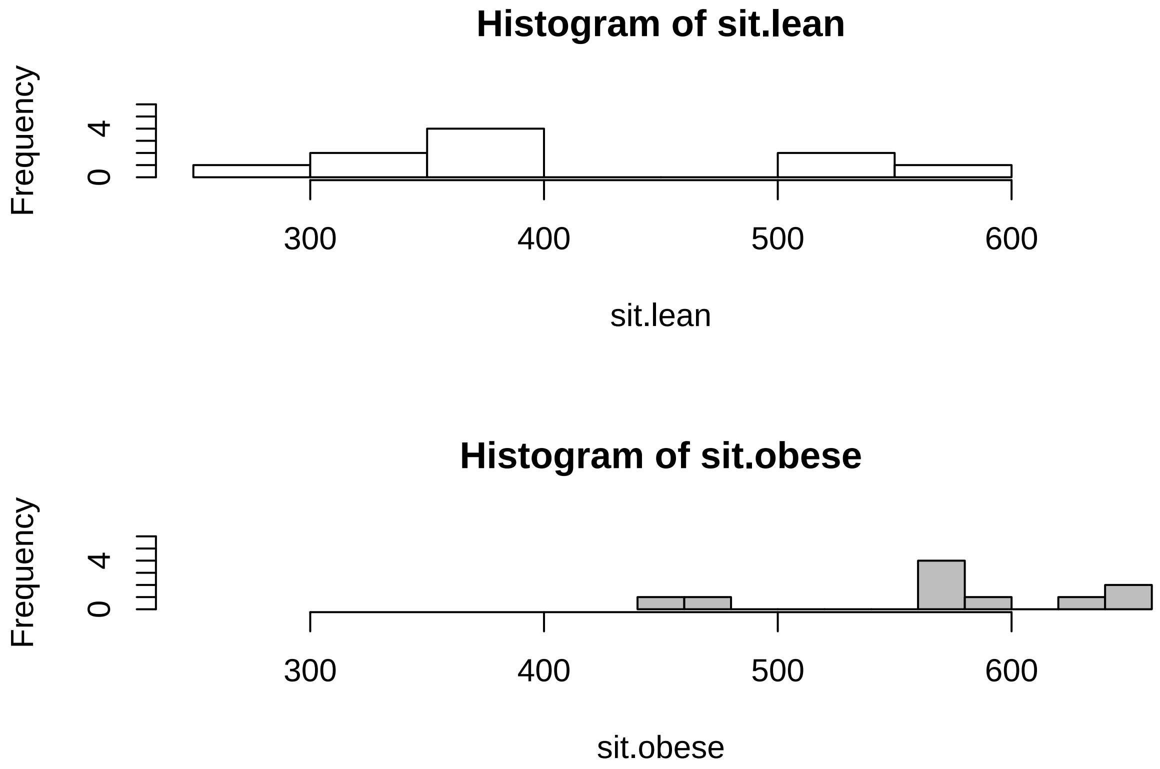
In comparing the sitting time between the `lean` and `obese` groups, it appears that, on average, the sitting time for the `obese` group is significantly greater than that of the `lean` group.

As for the lying time for the `lean` and the `obese` groups, it appears that, on average, the lying time for `lean` group is greater than that of the `obese` group. This could be attributed to the idea that those in the `lean` group may tend to spend more time on exercise, and thus may spend more time resting.

3e) Create histograms of Sit for the two groups. Let the two histograms have the same x limits in order to do comparison. (4 points)

Solution:

```
par(mfrow = c(2,1)) #To easily compare the two histograms
hist(sit.lean, breaks = 10, xlim = c(250, 650), ylim = c(0, 6), col = "white")
#plot histogram of sit by lean
hist(sit.obese, breaks = 10, xlim = c(250, 650), ylim = c(0, 6), col = "grey")
```



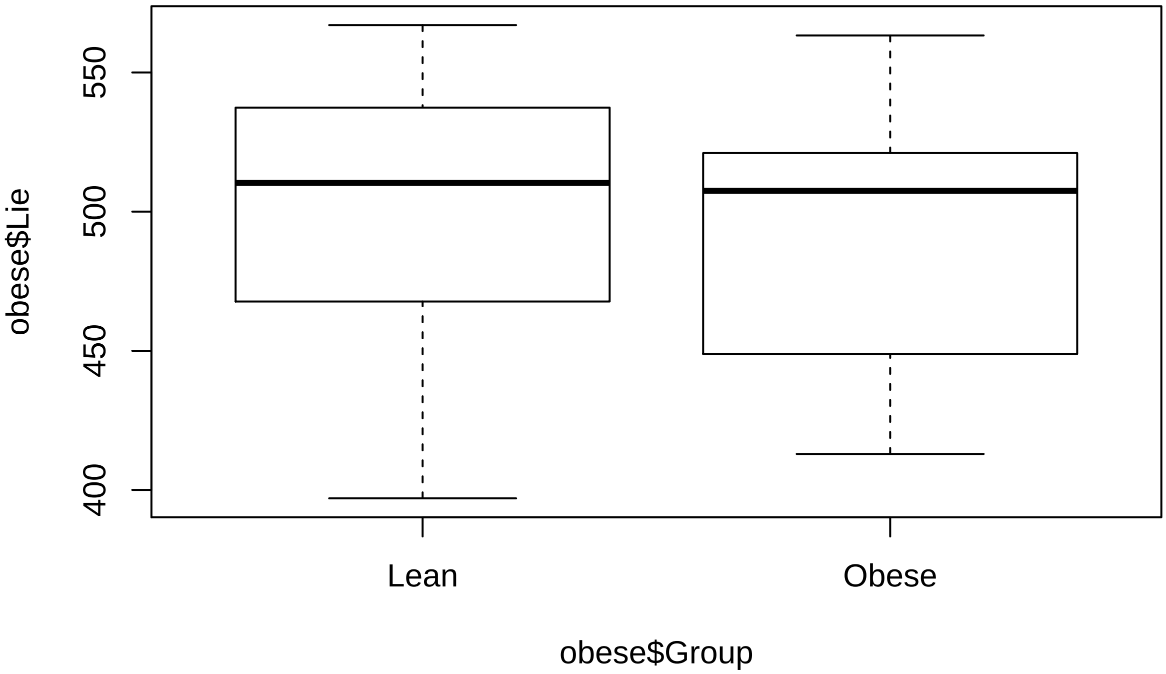
```
#plot histogram of sit by obese
```

It is apparent that the `obese` group spends more time sitting than the `lean` group.

3f) Create a vertical side-by-side boxplot for the time spent on lying down of the two groups. (4 points)

Solution:

```
boxplot(obese$Lie ~ obese$Group, horizontal = FALSE) #Generate side-by-side boxplot
```



No, there is no obvious difference. It is difficult to tell that there is a difference in lying time based on group.

3g) Compare the summary statistics of all variables between the two groups using the function by. (4 points)

Solution:

```
obese_summary <- by(obese, obese["Group"], summary)
obese_summary
```

```
## Group: Lean
##    Group       Subject          Stand            Sit              Lie       
##  Lean :10   Min.   : 1.00   Min.   :319.2   Min.   :268.2   Min.   :397.0  
##  Obese: 0   1st Qu.: 3.25   1st Qu.:506.3   1st Qu.:351.0   1st Qu.:473.1  
##             Median : 5.50   Median :549.5   Median :372.4   Median :510.3  
##             Mean   : 5.50   Mean   :525.8   Mean   :407.5   Mean   :501.6  
##             3rd Qu.: 7.75   3rd Qu.:583.2   3rd Qu.:493.0   3rd Qu.:535.9  
##             Max.   :10.00   Max.   :677.2   Max.   :582.1   Max.   :567.0  
## ------------------------------------------------------------ 
## Group: Obese
##    Group       Subject          Stand            Sit              Lie       
##  Lean : 0   Min.   :11.00   Min.   :260.2   Min.   :456.6   Min.   :412.9  
##  Obese:10   1st Qu.:13.25   1st Qu.:350.2   1st Qu.:567.6   1st Qu.:451.8  
##             Median :15.50   Median :388.9   Median :575.7   Median :507.5  
##             Mean   :15.50   Mean   :373.3   Mean   :571.2   Mean   :491.7  
##             3rd Qu.:17.75   3rd Qu.:415.8   3rd Qu.:613.8   3rd Qu.:519.5  
##             Max.   :20.00   Max.   :464.8   Max.   :646.3   Max.   :563.3  
```

From the results of the `by()` function, it is apparent that for `Standing`, those in the `Lean` group tend to spend more time in comparison to the `Obese` group. As for `Sitting`, those in the `Obese` group tend to spend more time in comparison to the `Lean` group. For `Lying` time, although the `Lean` group does expend more time than the `Obese` group, the differences aren't that significant to draw a massive conclusion from `Lying Time`. It appears that for more menial tasks, the `Obese` group tends to expend more time, while the `Lean` group tends to expend more time on tasks that require more energy.

3h) What conclusion can you make from the previous analysis? (3 points)

Solution:

Based on the previous analysis, it is apparent for more `menial` tasks, such as `sitting`, those in the `obese` group spend more time in comparison to tasks that require more energy such as `standing`. As for the `lean` group, it appears that they spend more time on tasks that require more energy such as `standing` and `lying down`, and less on more menial tasks such as `sitting`.