Title: The Height and Weight of Junior High School Students

Data Summary

The *jhs* dataset consisted of 19 observations and 5 variables. The data recorded the name of the student, the age of the student, the sex of the student (1 for male and 2 for female) and the height and weight of the student.

The average weight of Junior High Student by for boys and girls is:

```
> round(tapply(WEIGHT, SEX, mean),2)
     1      2
108.95  90.11
```

The average height of Junior High Student by for boys and girls is:

```
> round(tapply(HEIGHT, SEX, mean),2)
    1     2
63.91 60.59
```

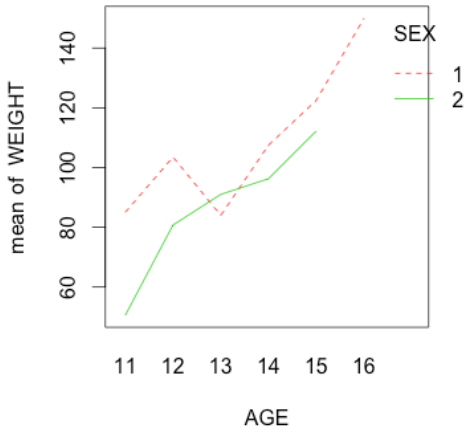
The average weight of Junior High Student by age is:

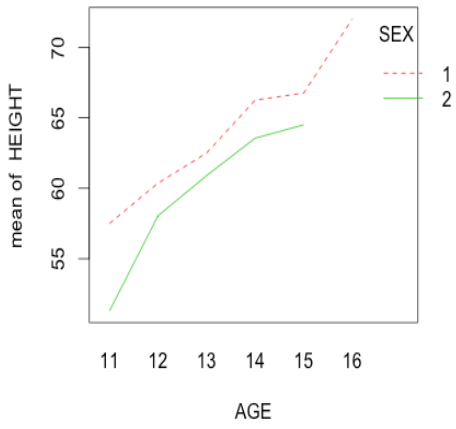
```
> round(tapply(WEIGHT, AGE, mean),2)
    11     12     13     14     15     16
 67.75  94.40  88.67 101.88 117.38 150.00
```

The average hight of Junior High Student by age is:

```
> round(tapply(HEIGHT, AGE, mean),2)
   11    12    13    14    15    16
54.40 59.44 61.43 64.90 65.62 72.00
```

```
> cor(jhs[2:4], use="complete.obs", method="kendall")
          HEIGHT    WEIGHT       AGE
HEIGHT 1.0000000 0.7142984 0.6260352
WEIGHT 0.7142984 1.0000000 0.6169202
AGE    0.6260352 0.6169202 1.0000000
```





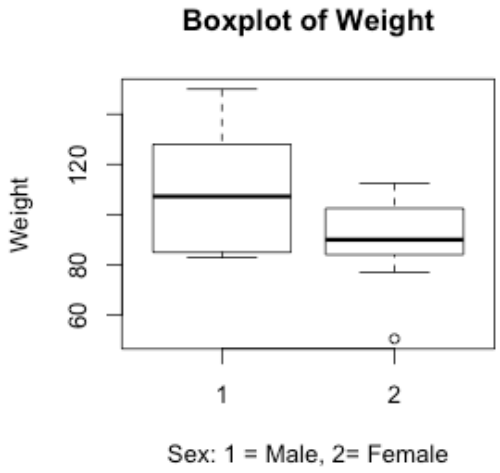
From the correlation test, we can see that age is strongly correlated with height and weight. We can also see that from the interaction plots above, as the age increases the weight and the height of student of different sex also increase.

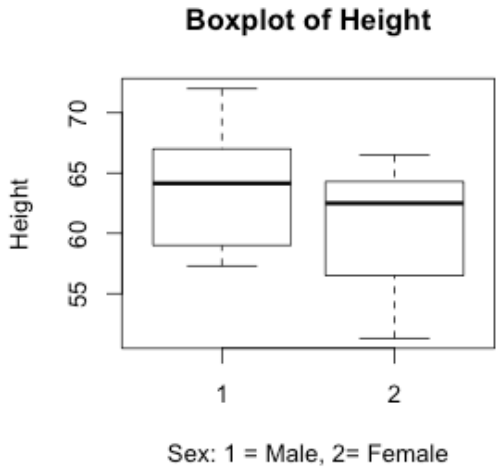
Statement of Problem

For this dataset, we want to test for the significant different in height and weight is due to sex or age. And if there is difference, we want to group students into same cluster of similar height and weight using hierarchical cluster analysis.

Analysis

Boxplot





From the boxplot, weight and the height do not seen to be different between male and female. However, weight and height increased by age increment. Now, we want to test if there is true differences in height and weight according to manova.

MANOVA

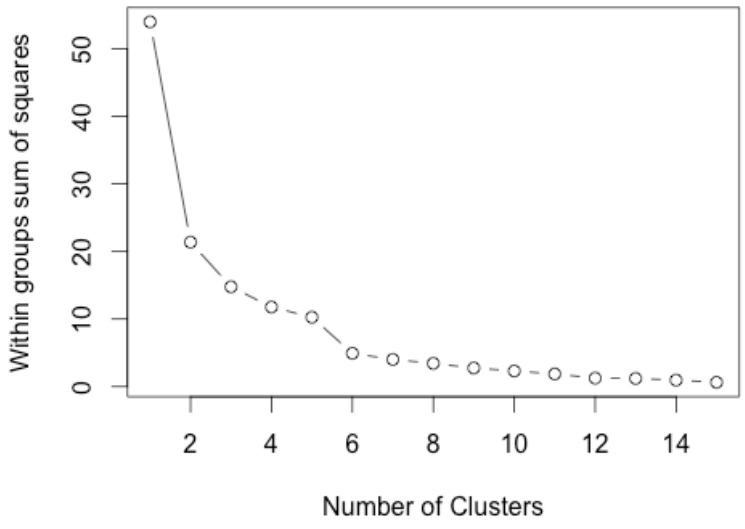
```
> ma = manova(cbind(WEIGHT,HEIGHT)~ AGE + SEX, jhs)

> summary(ma, test="Hotelling-Lawley")
          Df Hotelling-Lawley approx F num Df den Df    Pr(>F)
AGE        1          2.75574   20.668      2     15 4.895e-05 ***
SEX        1          0.50267    3.770      2     15   0.04716 *
```

Result of Hotelling-Lawley test in MANOVA suggests that there is true difference in height and weight between sexes and ages. Therefore we want to group student according to their differences.

Hierarchical Cluster Analysis

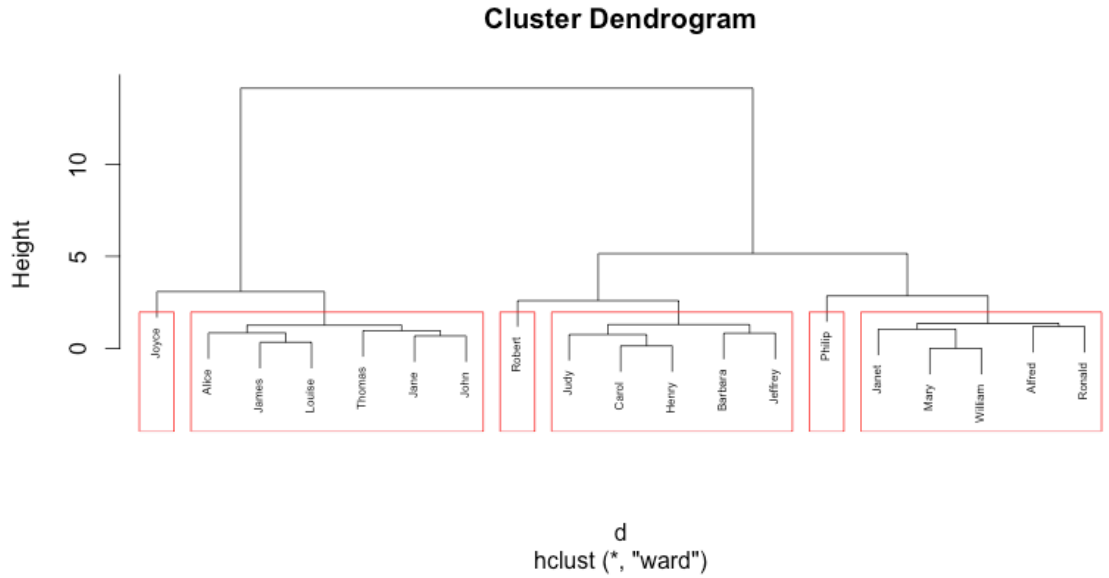
Frist, we need to determine the number of clusters (n) minimizes the sum of square within groups. Then find the Euclidean distance between two clusters to be the maximum distance between their individual components. Lastly, plot dendrogram with n groups.

```
scd.st= scale(na.omit(jhs[2:4]))
wss <- (nrow(scd.st)-1)*sum(apply(scd.st,2,var))
for (i in 2:15) wss[i] <- sum(kmeans(scd.st,centers=i)$withinss)
plot(1:15, wss, type="b", xlab="Number of Clusters",ylab="Within groups sum of squares")
```



The sum of square within groups test suggests we use 6 clusters.

```
> d = dist(scd.st, method = "euclidean")##distance matrix
> fit = hclust(d, method="ward")
> plot(fit, cex = .5, labels = NAME)
> groups = cutree(fit,k=6)## cut tree into 6 clusters
> rect.hclust(fit, k=6, border="red")
```




```
> GROUP = as.factor(groups)
> clus = cbind(jhs, GROUP)
> for(i in 1:6){print(subset(jhs, GROUP == i ))}
```

```
      NAME HEIGHT WEIGHT AGE SEX
1   Alfred   69.0  112.5  14   1
8    Janet   62.5  112.5  15   2
14    Mary   66.5  112.0  15   2
17  Ronald   67.0  133.0  15   1
19 William   66.5  112.0  15   1
      NAME HEIGHT WEIGHT AGE SEX
2   Alice   56.5   84.0  13   2
6   James   57.3   83.0  12   1
7    Jane   59.8   84.5  12   2
10   John   59.0   99.5  12   1
13 Louise   56.3   77.0  12   2
18 Thomas   57.5   85.0  11   1
      NAME HEIGHT WEIGHT AGE SEX
3  Barbara   65.3   98.0  13   2
4    Carol   62.8  102.5  14   2
5    Henry   63.5  102.5  14   1
9  Jeffrey   62.5   84.0  13   1
12    Judy   64.3   90.0  14   2
      NAME HEIGHT WEIGHT AGE SEX
11 Joyce   51.3   50.5  11   2
      NAME HEIGHT WEIGHT AGE SEX
15 Philip     72    150  16   1
      NAME HEIGHT WEIGHT AGE SEX
16 Robert   64.8    128  12   1
```

Conclusion

According to a MANOVA test, we find out that there is different in height and weight of different sexes and ages. Then we group students into similar height and weight.

The first cluster consist of Alfred, Janet, Mary, Ronald, and William with height range from 62.5 to 66.5 and weight range from 112.4 to 133.0 and age 14, and 15.

The second cluster is Alice, James, Jane, John, Louise, and Thomas with height range from 56.3 to 59.8 and weight from 77 to 99 and age 11 to 13.

The third cluster is Barbaca, Carol, Henry, Jeffrey, and Judy with height range from 62.5 to 65.3 and weight from 84 to 102.5 and age of 13, and 14. This cluster is similar to the first cluster except the weight is much lower.

The fourth, fifth and sixth clusters are some extreme cases. Joyce, in the fourth cluster, is the shortest and lightest of all. Philip, in fifth cluster, is the tallest and heaviest of all. And Robert, in the sixth cluster is the heaviest given in height and age.