Cluster Analysis on Australian Athletes Data

Yehia Abdelmohsen (900193174)

4/24/2022

Contents

Introduction

The data presented contains information for athletes from different types of sports. The purpose of collecting the data was to test if blood hemoglobin levels are different when comparing endurance-related athletes to those in power-related events. In this analysis, we will aim to cluster the athletes into their respective events. We will also check if the clustering algorithm splits the athletes into male and female. Finally, we will add a new class variable which represents whether the athlete participates in endurance/power events and check if the clustering algorithm followed that idea.

About the Data

The data presented contains information related to Australian athletes. It contains 202 observations where each observation represents an athlete. There are 13 variables include the class variable which represents the sport the athlete plays.

The data is obtained from R, its documentation is linked below.

Source: https://vincentarelbundock.github.io/Rdatasets/doc/DAAG/ais.html

Variables

- Rcc - Red blood cell count. This is a quantitative variable.
- Wcc - White blood cell count, per liter. This is a quantitative variable.
- Hematocrit - Percent of hematocrit. Hematocrit: The ratio of the volume of red blood cells to the total volume of blood. This is a quantitative variable.
- Hg - Hemaglobin concentration in g per decaliter (10 liters). Hemaglobin: A protein that carries oxygen to organs. This is a quantitative variable.
- Ferr - Plasma ferritins in ng. Measures the amount of ferritin in blood. Ferritin is a blood protein that contains iron. This is a quantitative variable.
- Bmi - Body mass index, in kg/m^2. Body mass divided by the square of the height m^2. This is a quantitative variable.
- Ssf - Sum of skin folds. Estimates the percentage of body fat by measuring skin fold thickness. This is a quantitative variable.
- PcBfat - Percentage of body fat. This is a quantitative variable.
- Lbm - Lean body mass in kg. Total body weight minus body fat weight. This is a quantitative variable.
- Ht - Height in cm. This is a quantitative variable.
- Wt - Weight in kg. This is a quantitative variable.
- Sex - A factor representing the sex of the athlete: female and male. This is a qualitative categorical variable.
- Sport - A factor representing the sport the athlete plays B Ball, Field, Gym, Netball, Row, Swim, T 400m, T Sprnt, Tennis, and W Polo. This is a qualitative categorical variable and will be used as the class variable.

Some Remarks Regarding the Data

Below is an initial look at the data.

rcc	wcc	hc	hg	ferr	bmi
3.96	7.5	37.5	12.3	60	20.56
4.41	8.3	38.2	12.7	68	20.67

ssf	pcBfat	lbm	ht	wt	sex	sport
109.1	19.75	63.32	195.9	78.9	f	B_Ball
102.8	21.30	58.55	189.7	74.4	f	B_Ball

Here is a look at the sports the athletes participate in.

```
##  [1] "B_Ball"  "Field"   "Gym"     "Netball" "Row"     "Swim"    "T_400m" 
##  [8] "T_Sprnt" "Tennis"  "W_Polo" 
```

Problem Statement

The data contains quantitative variables we can use to cluster the data. The aim of this analysis:

- Firstly, we will see if we can cluster the data into ten classes each one representing a sport.
- If we are not able to cluster into the ten sports. We will try to either cluster into female and male or power-related and endurance-related.

Some Graphs

Clustering into Sports

In Figure 1, the various sports are colored differently. It is obvious that the sports are not easily distinguishable from each other using two variables.

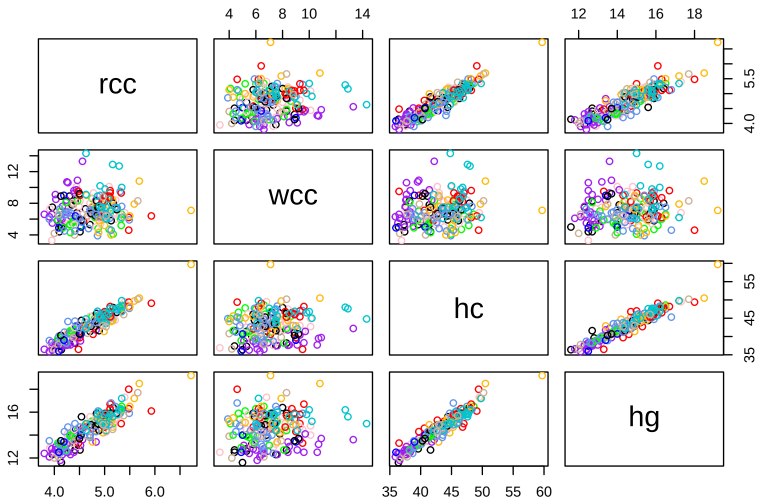

Figure 1: Pairs Plot

Again, in Figure 2 we are not able to distinguish between the sports using any of the two variables. The only points that we are able to distinguish from the rest are the red points in the BMI graphs, which represent athletes in field sports. Field sports in Track and Field are those that include jumping (e.g. long jump, high jump) and throwing (e.g. shot put, javelin). Usually throwing athletes are very heavy athletes due to the nature of their sport. This explains the clear distinction between them and the rest of the sports.

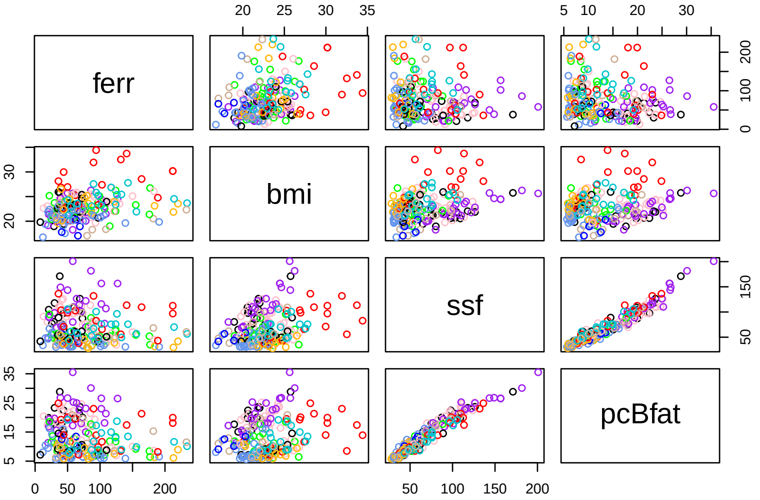

Figure 2: Pairs Plot

Other than the point mentioned regarding field athletes, the sports overlap in most cases and it is hard to distinguish between them in 2D. We will see what the clustering algorithms can do in higher dimensions.

Clustering into Male and Female

In Figure 3, we can see that the two clusters are clearly distinguishable. Men are represented by red points and women are represented by black points. If I were to use one variable to distinguish between the two it would be Hemoglobin (hg) since there is a very clear distinction. Hemotacrit (hc) is a close second to distinguish between the variables. To view more pair plots jump to appendix section: Pair Plots.

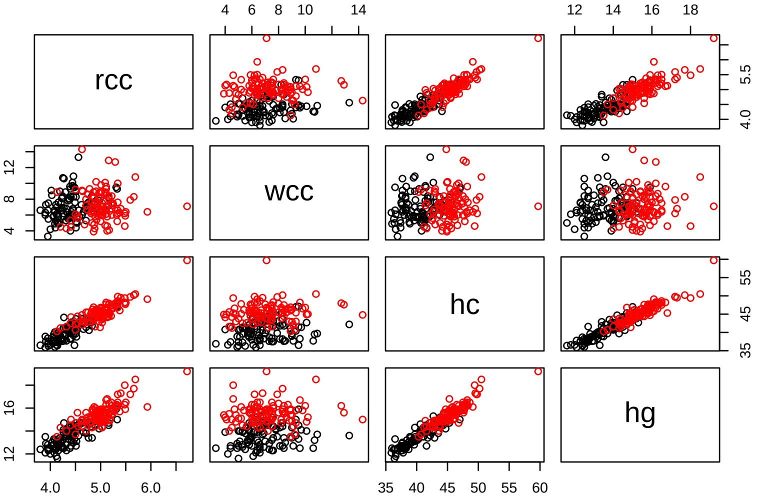

Figure 3: Pairs Plot

Clustering into Power-related and Endurance-related sports

Given that the data related to the sports being characterized as power or endurance sports is not available, I had to research the topic. Some sources said that a sport was endurance based and some power based. Therefore, I used some common knowledge and the sources information to split the sports. I decided to split them in the following way.

Endurance sports: Basketball, Rowing, Sprint (400m), Tennis, and Water Polo.
Power sports: Gym, Netball, Swimming, Sprints (<400m), and Field.

Obviously, we can see in Figure 4 that it is not easy to distinguish between the sports in 2D. That is of course under the assumption that my splitting of the sports is correct. We will move forward with this assumption, and we will see if the clustering algorithms give us any interesting results.

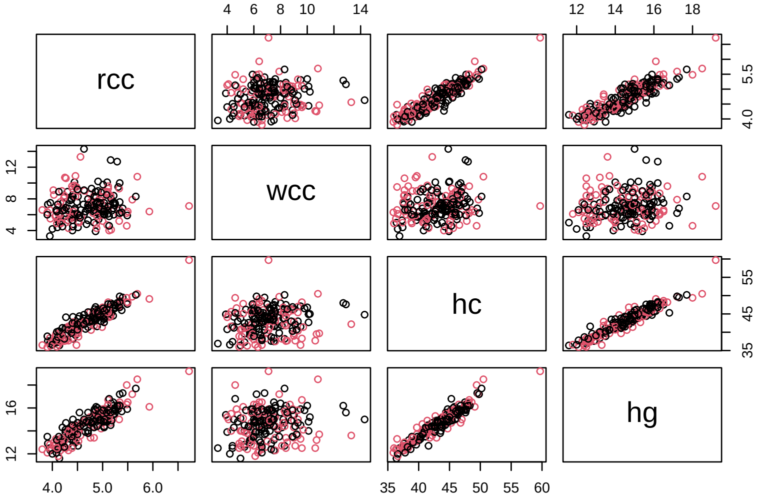

Figure 4: Pairs Plot

Clustering

Hierarchical

Below are the sports with their corresponding number equivalents. In addition, the last column represents the number of athletes in the data from the corresponding sport.

Table 3: Cluster Numbers and Count

	sports	df.sport	count
1	1	B_Ball	25
14	5	Row	37
36	4	Netball	23
59	6	Swim	22
68	2	Field	19
69	7	T_400m	29
77	8	T_Sprnt	15
90	9	Tennis	11

	sports	df.sport	count
97	3	Gym	4
182	10	W_Polo	17

Male and Female

The output in Figure 5 shows results when using euclidian distance for distances between points and Ward D for distances between clusters. The output will be tested to see if the clusters represent male and female.

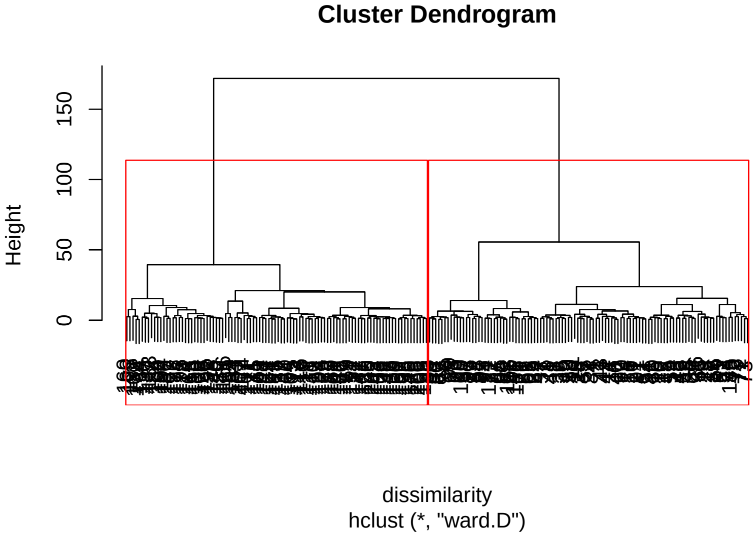

Figure 5: Hierarchical Clustering (k=2)

Another iteration was run where euclidian distance was used for distances between points and Ward D^2 for distances between clusters (Figure 6). However, the result of using euclidian distance for distances between points and Ward D for distances between clusters was more accurate when computing the classification error. Other iterations were examined (e.g. manhattan and complete linkage, euclidian and single linkage), but none really provided results like euclidian and Ward D (in terms of classification error), therefore, we will move forward with this combination.

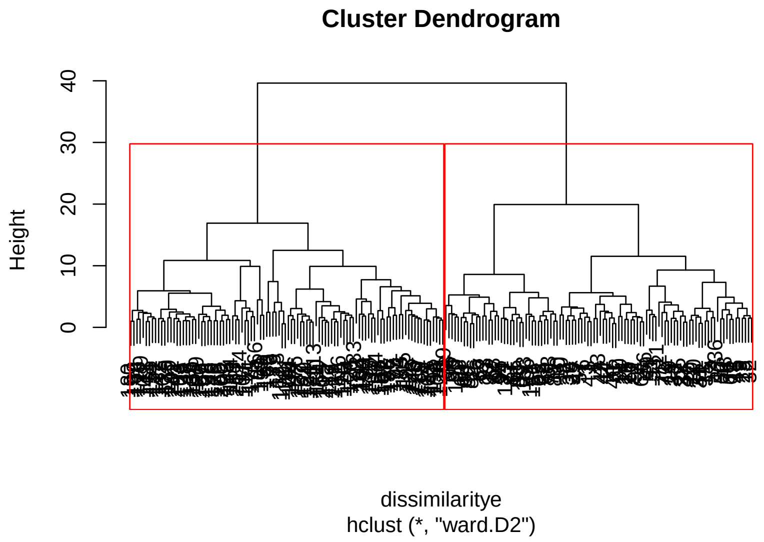

Figure 6: Hierarchical Clustering (k=2)

We then computed the confusion matrix for the case where the two clusters are assumed to be male and female. An error rate of 2.9% is observed. This is a very convincing error rate. It shows that we can distinguish between male and female athletes given that we have information related to their blood and body (e.g. Hemoglobin, BMI, Weight). The success of clustering into male and female foreshadows the failure of clustering into ten sports. Nevertheless, we will still give it a shot.

Table 4: Confusion Matrix (Clusters Represent Gender)

1	2
99	5
1	97

```
##  Error rate = 0.02970297
```

```
## R2 =  0.3470876
```

```
## [1] 0.3470876
```

Sports

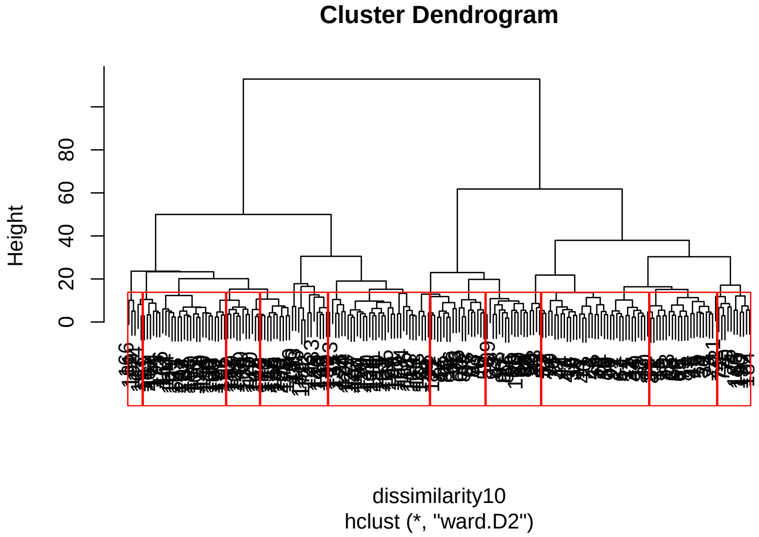

Figure 7: Hierarchical Clustering (k=10)

Figure 7 shows the clusters when $k = 10$. The output below shows an error rate of about 60%. This is very high and suggests that the clusters do not represent the sports. The idea that male and female athletes aren't likely to be in the same clusters shows that we may need 17 clusters. 8 for male athlete sports and 9 for female athlete sports. This leads us to believe that splitting the data into male and female, then clustering may improve our chances in reaching a convincing error rate.

```
## Loading required package: clue
```

Table 5: Confusion Matrix (Clusters Represent Sports)

4	5	3	10	8	7	1	6	2	9
17	2	0	0	0	0	5	1	0	1
2	19	0	0	0	2	5	5	0	1
3	1	4	0	0	7	2	3	0	3
0	1	0	4	0	0	1	0	6	0
1	1	0	0	4	4	0	0	1	2
0	0	0	4	5	15	5	7	4	2
0	11	0	5	1	0	6	4	1	0
0	0	0	1	3	1	0	2	0	1
0	2	0	3	0	0	1	0	7	0
0	0	0	0	2	0	0	0	0	1

```
##  Error rate = 0.6089109
```

```
## R2 =  0.6568497
```

```
## [1] 0.6568497
```

After splitting the data into male and female we clustered using euclidian distance and Ward D^2. The output shown in Figure 8 represents the male subset in the data. There are 102 male observations in the data.

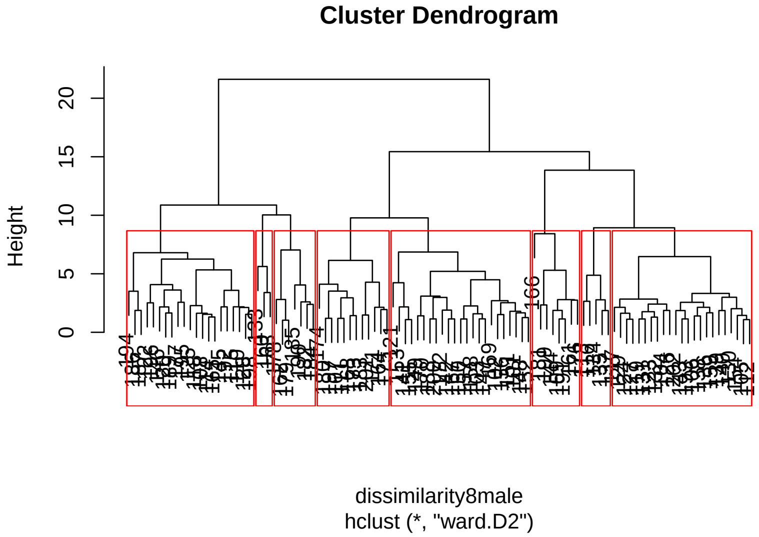

Figure 8: Hierarchical Clustering (k=8)

After looking at the error rate for multiple combinations of distance between point measures and distance between clusters measures, we decided on euclidian distance and Ward D^2. It had the smallest error rate. This error rate 83%, however, is not very convincing. It's important to note that the small size of the data may be a factor that influences the clustering algorithm, and that more data may help decrease the error rate.

Table 6: Confusion Matrix (Clusters Represent Male Sports)

9	1	8	7	10	5	6	2
1	0	4	12	1	1	3	1
1	8	0	2	1	6	4	1
1	0	3	3	2	0	2	1
0	0	1	0	8	7	2	3
1	0	3	1	1	0	1	1
0	3	0	0	0	1	1	0
0	1	0	0	0	0	0	2
0	0	0	0	4	0	0	3

```
##  Error rate = 0.8333333
```

```
## R2 =  0.5811915
```

```
## [1] 0.5811915
```

Figure 9 shows the clustering results when clustering $k = 9$ for female athletes.

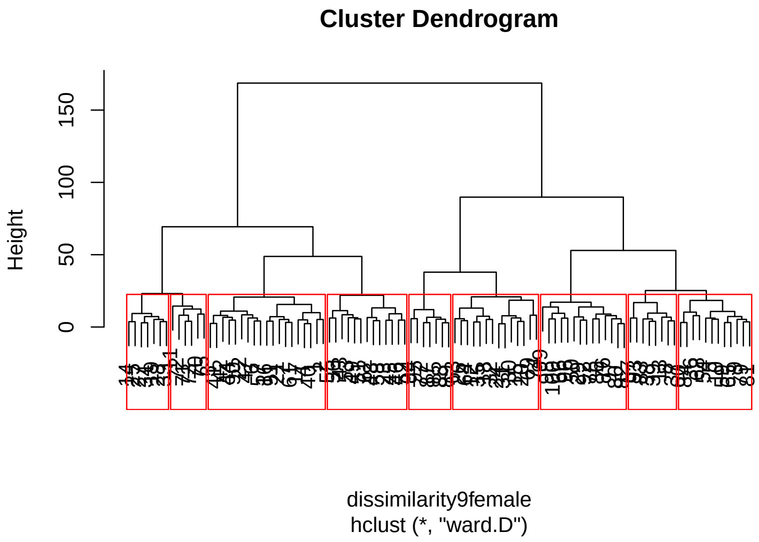

Figure 9: Hierarchical Clustering (k=8)

The error rate is 56%, which is better than 83%. However, it is not too far off from 60% when the male and female athletes were in one set. This leads us to the conclusion that separating the data into male and female is not the best idea.

Table 7: Confusion Matrix (Clusters Represent Female Sports)

1	6	9	2	5	3	4	7	8
6	1	2	0	4	0	6	0	0
3	3	0	1	0	0	2	3	0
2	1	1	0	0	0	3	1	0
1	0	0	5	0	0	0	0	0
1	2	0	1	9	0	0	1	0
0	0	0	0	7	0	0	0	0
0	1	0	0	1	0	11	0	0
0	1	2	0	1	4	1	5	0
0	0	2	0	0	0	0	1	4

```
##  Error rate = 0.56
```

```
## R2 =  0.6057922
```

```
## [1] 0.6057922
```

Power/Endurance Sports

Now looking at the error rate when power/endurance sports are assumed to be the classes. It is considerably high 46%. This leads us to believe that Power/Endurance sports do not represent the clusters. However, this may also be the result of incorrect allocation of the sports into power/endurance.

Table 8: Confusion Matrix (Clusters Represent Power/Endurance Sports

2	1
47	57
36	62

```
##  Error rate = 0.460396
```

```
## R2 =  0.3470876
```

```
## [1] 0.3470876
```

K-Means

Note: Results in this section are different with every run due to the non-deterministic nature of k-Means. Therefore, explanations may not align exactly with the numbers available in the code.

We will now be using k-Means to cluster the data. We can determine the number of clusters using the elbow curve shown in Figure 9. In this case, I would choose $k = 4$. However, since we are assuming a k already we will not be using it.

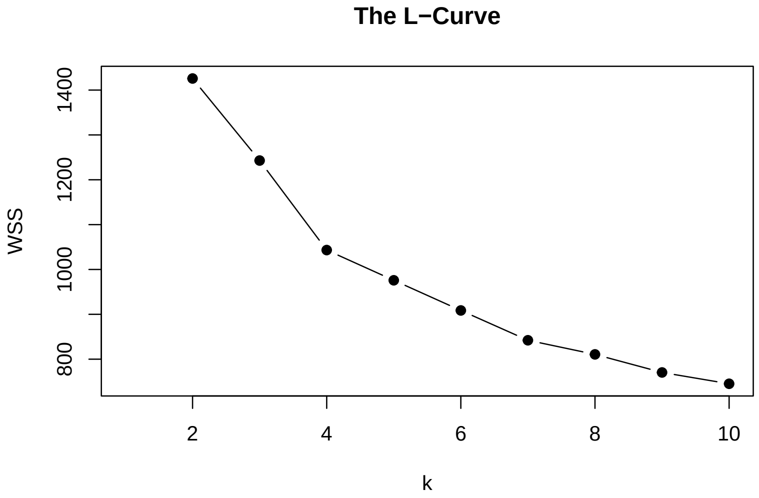

Figure 10: Determining the Number of Clusters

Male and Female

Table 9: The Number of Individuals in Each Cluster

clusters	Freq
1	100
2	102

When clustering using k-Means, we obtain similar results to Hierarchical clustering. Results show an error rate of 3.9% when the clusters are assumed to be male and female athletes. Out of 202 observations there are 8 misclassifications. These are very convincing results and are almost identical to Hierarchical (Euclidian and Ward D2) available in the appendix.

Table 10: Confusion Matrix (Clusters Represent Gender - k=2)

1	2
96	4
4	98

```
##  Error rate = 0.03960396
```

```
## R2 =  0.3551394
```

```
## [1] 0.3551394
```

Sports

When clustering using k-Means where $k = 10$, results are again similar to that of Hierarchical clustering. The error rate is 67%. This shows that the clusters do not represent the sports.

Table 11: Confusion Matrix (Clusters Represent Sports)

2	4	6	10	1	9	3	8	5	7
8	0	0	1	0	0	0	0	0	0
1	9	0	0	1	0	0	0	0	0
1	0	6	4	8	0	0	1	9	0
2	0	1	7	0	0	0	1	4	0
0	11	2	0	8	3	0	0	5	0
1	0	0	0	0	1	0	2	0	0
1	2	5	0	0	4	4	4	2	11
0	0	2	2	0	1	0	3	1	1
3	1	2	0	4	0	0	0	16	2
2	0	4	3	4	2	0	4	0	15

```
##  Error rate = 0.6188119
```

```
## R2 =  0.9543323
```

```
## [1] 0.9543323
```

We then split the data exactly like we did in Hierarchical clustering into male and female athletes. The results were quite similar. For male athletes, the classification of sport resulted in an error rate of 87%, which is quite high. In this same case Hierarchical clustering resulted in an error rate of 83%.

Table 12: Confusion Matrix (Clusters Represent Male Sports - k=8)

2	1	5	6	7	9	10	8
5	1	0	0	0	0	4	0
3	5	6	5	2	1	1	1
0	3	2	1	0	0	0	0
1	0	1	0	0	1	1	2
2	0	5	2	0	0	7	1
0	3	0	0	9	1	2	3
0	0	0	2	1	1	1	3
1	0	1	3	6	0	1	1

```
##  Error rate = 0.8529412
```

```
## R2 =  0.6824868
```

```
## [1] 0.6824868
```

Again, we did the same thing for female athletes. The results were also similar to Hierarchical clustering. The error rate was 60%. In this case Hierarchical clustering resulted in an error rate of 56%.

Table 13: Confusion Matrix (Clusters Represent Female Sports - k=9)

3	5	2	6	7	4	1	9	8
0	3	0	1	1	2	1	0	1
0	9	1	1	0	0	1	0	0
0	6	2	0	0	0	1	0	0
0	1	0	4	3	2	5	1	0
4	1	1	2	5	1	0	2	0
0	0	1	0	0	8	0	0	0
0	2	0	1	0	10	5	2	0
0	0	2	0	0	0	0	0	0
0	0	0	0	2	0	0	2	3

```
##  Error rate = 0.64
```

```
## R2 =  0.8383301
```

```
## [1] 0.8383301
```

Power/Endurance Sports

Clustering into power and endurance related sports results in an error rate of 47%, which is similar to the results obtained by Hierarchical clustering. This leads us to conclude that the clusters don't represent power/endurance sports.

Table 14: Confusion Matrix (Clusters Represent Gender)

1	2
63	39
56	44

```
##  Error rate = 0.470297
```

```
## R2 =  0.3551394
```

```
## [1] 0.3551394
```

Conclusion

To conclude, it's obvious that the data is most suitable for clustering male and female athletes. This is because the classification error is the lowest when $k = 2$ and when we compare the clusters to the class variable sex in the data. Even though we failed to cluster the data into sports, I'm still convinced it is possible with more data.

Also, I think clustering the sports into power and endurance would work well, however, my judgment of which sports were power/endurance heavy was not ideal. It is possible to look at all combinations and picking the one with the lowest error rate, however, that would take a lot of computation and time.

Finally, more data would provide more convincing results as in all clustering and classification problems.

Table 15: Comparing Methods Error Rate

	Sex	Power/Endurance	Sports	Sports (Male)	Sports (Female)
Hierarchical	2.9%	46%	60%	83%	56%
k-Means	3.9%	47%	65%	87%	56%
k	2	2	10	8	9
n	202	202	202	102	100

Appendix

Pair Plots

Clustering into sports

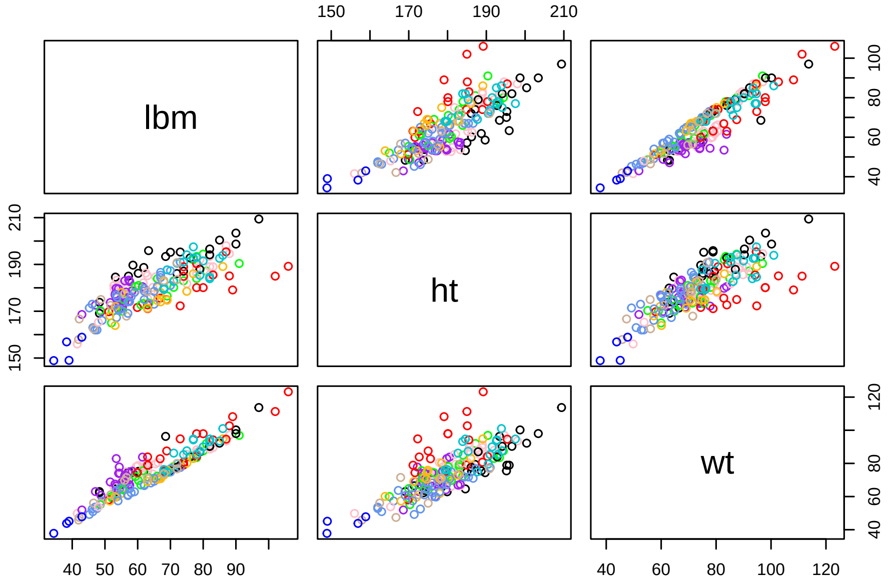

Figure 11: Pairs Plot

Clustering into male and female

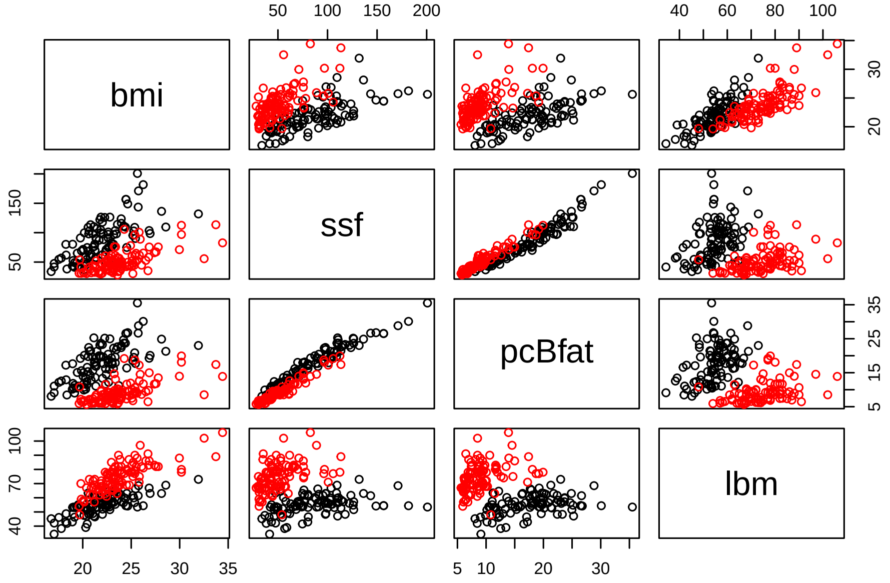

Figure 12: Pairs Plot

Clustering into Power-related and Endurance-related sports

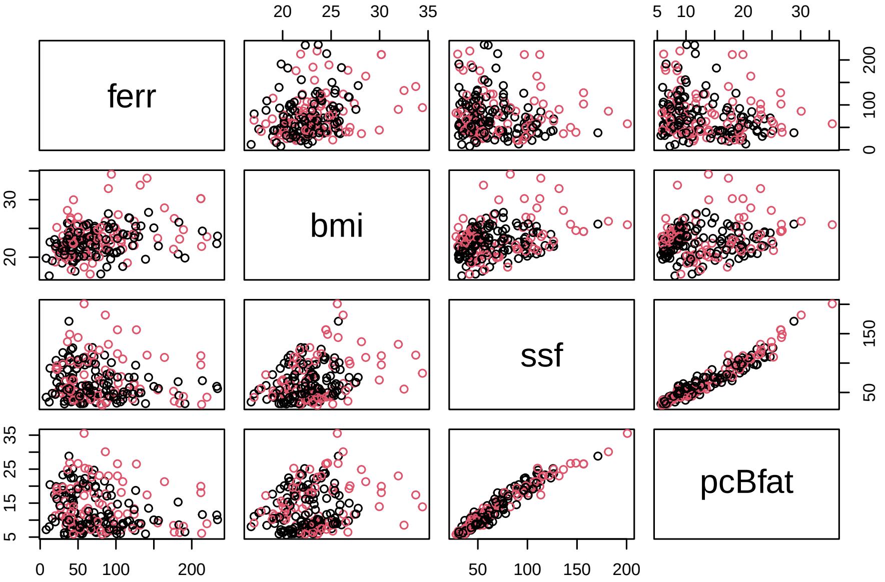

Figure 13: Pairs Plot

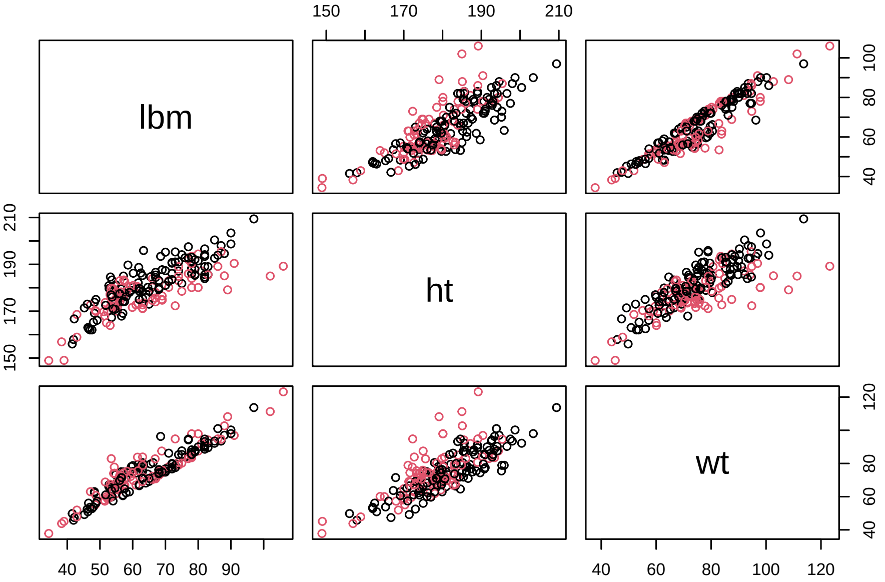

Figure 14: Pairs Plot

Clustering Trials

Hierarchical (Euclidian and Ward D2)

Table 16: Confusion Matrix (Clusters Represent Gender)

1	2
96	4

1	2
4	98

```
##  Error rate = 0.03960396
```