Clustering Analysis for Athletes Data

Sharmin Akhter

Memorial University of Newfoundland

April 17, 2023

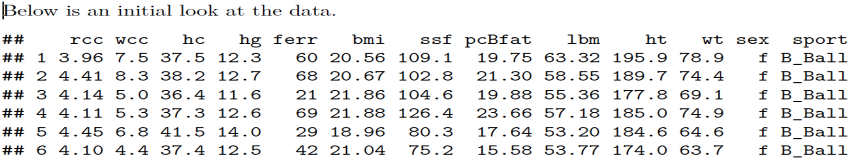
Aim of the Analysis

- Data contains information for athletes from various sports.
- Test if blood hemoglobin levels differ between endurance and power athletes.
- Analysis aims to cluster athletes into their respective events.
- Check if clustering algorithm separates athletes by gender.
- Evaluate if the clustering algorithm follows the endurance/power classification.

Problem Statement

- Cluster the data into ten classes each one representing a sport.
- Cluster into female and male or power-related and endurance-related.

Data Information



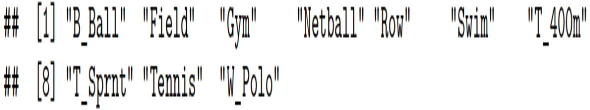
Below is an initial look at the data.

```
##      rcc wcc   hc   hg ferr   bmi   ssf pcBfat   lbm    ht   wt sex  sport
## 1 3.96 7.5 37.5 12.3   60 20.56 109.1  19.75 63.32 195.9 78.9   f B_Ball
## 2 4.41 8.3 38.2 12.7   68 20.67 102.8  21.30 58.55 189.7 74.4   f B_Ball
## 3 4.14 5.0 36.4 11.6   21 21.86 104.6  19.88 55.36 177.8 69.1   f B_Ball
## 4 4.11 5.3 37.3 12.6   69 21.88 126.4  23.66 57.18 185.0 74.9   f B_Ball
## 5 4.45 6.8 41.5 14.0   29 18.96  80.3  17.64 53.20 184.6 64.6   f B_Ball
## 6 4.10 4.4 37.4 12.5   42 21.04  75.2  15.58 53.77 174.0 63.7   f B_Ball
```

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

1. 202 observations, 13 variables
2. **Source:** R document

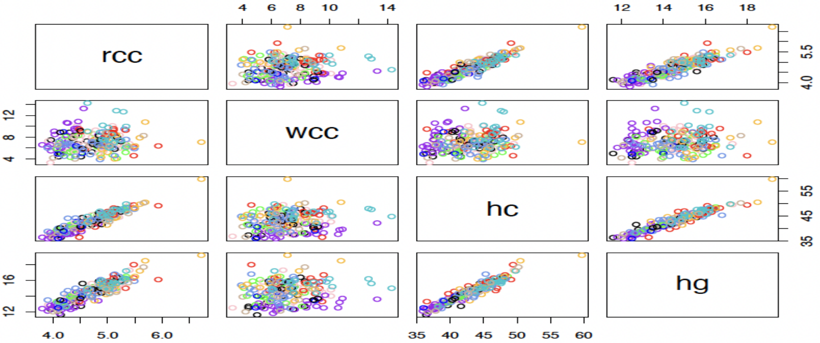


```
## [1] "B_Ball"  "Field"   "Gym"     "Netball" "Row"     "Swim"    "T_400m"
## [8] "T_Sprnt" "Tennis"  "W_Polo"
```

Outline

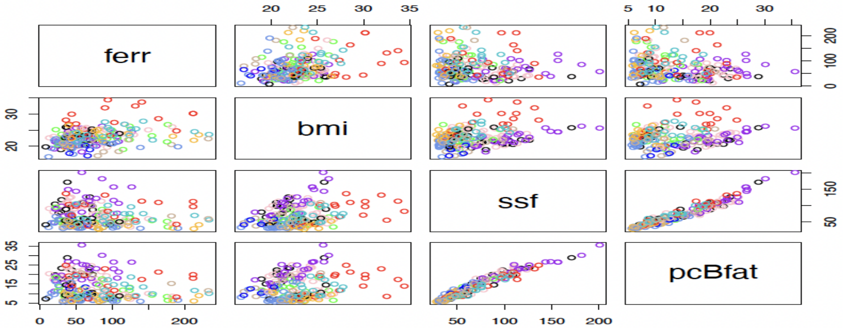
- Clustering through some pair plots
- Hierarchical
- K-means
- Conclusion

Clustering into Sports



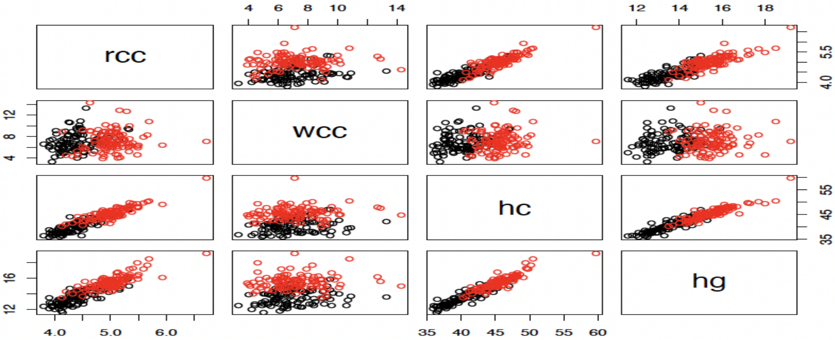
- Variables are not distinguishable between the sports

Clustering into Sports



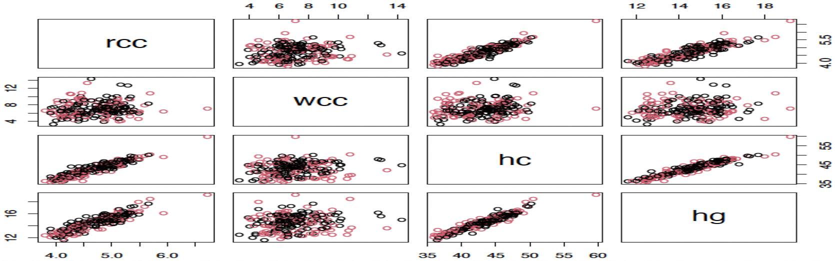
- Only BMI is distinguishable which are in red points and represents athletes in sports

Clustering into Gender



- Clearly distinguishable(Men, Women)
- Hemoglobin and Hematocrit are the most effective variables for differentiating between Men and Women

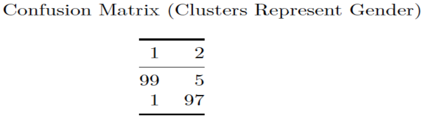
Clustering into two sports group



- **Endurance sports:** Basketball, Rowing, Sprint (400m), Tennis, and Water Polo.
- **Power sports:** Gym, Netball, Swimming, Sprints (¡400m), and Field.
- Not distinguishable

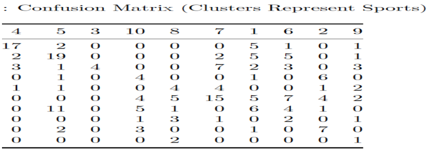
Hierarchical Clustering k=2, k=10, k=2

Confusion Matrix for Gender, Sports, Power/Endurance Sports



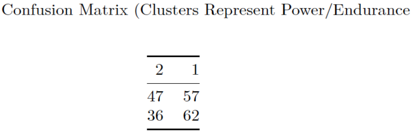
Confusion Matrix (Clusters Represent Gender)

1	2
99	5
1	97



: 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



Confusion Matrix (Clusters Represent Power/Endurance

2	1
47	57
36	62

Hierarchical Clustering and K-means

- **Error Rate and R^2 for Hierarchical**
  1. Cluster Represent Gender 2.9% and 0.3470876
  2. Cluster Represent Sports 60% and 0.6568497
  3. Cluster Represent Power/Endurance Sports 46% and 0.3470876
- **Error Rate and R^2 for K-mean**
  1. Cluster Represent Gender 3.9% and 0.3551394
  2. Cluster Represent Sports 61% and 0.9355204
  3. Cluster Represent Power/Endurance Sports 47% and 0.3551394

Summary

- Data is best suited for clustering male and female athletes due to the lowest classification error when k = 2.
- Clustering based on sports was not successful, but may be possible with more data.
- Clustering sports into power and endurance categories has potential, but requires better judgment of which sports belong to each category.
- Evaluating all combinations to find the lowest error rate is possible, but computationally expensive and time-consuming.

Thank you!!!