Final Project Report - Fall Term 2023

Analyzing Olympic Events Data

MGSC 661: Multivariate Statistics

Yifan Lu (261144056)

1．Introduction

The Olympic Games is a worldwide combined games organized by the International Olympic Committee. It originated from the Olympia Games in ancient Greece and was a symbol of the ancient Greeks' display of strength and spirit to the gods. The modern Olympic Games were initiated by French educator Pierre de Coubertin in the late 19th century, and the first modern Olympic Games were held in Athens, Greece in 1896. The Olympic Games are held every four years and are one of the most influential and popular sporting events in the world.

I use basic biographical data and medal results of athletes from the 1896 Athens Olympics to the 2016 Rio Olympics, which contain personal information such as gender, age, height, weight, and nationality of the athletes, as well as information about the competitions and medals in which they participated.

By analyzing this data, we hope to address the following set of questions:

(1) I wish to determine whether the game is fair.

(2) I wish to explore whether the physical fitness requirements of athletes differ from sport to sport.

(3) I wish to be able to predict whether an athlete will be able to win a medal by their physical fitness.

2．Data description

In this chapter, we first perform some checks on the data, deal with missing values, and then carry out the underlying analysis.

2.1. Missing value processing

The raw data contains a total of 15 columns of data including player number, name, gender, height, weight, team, sport, and medal information. Through inspection, we found that the missing values mainly exist in age, height, weight, and medals. Due to the randomness of the medal type, in this study, the medalist is labeled as 1. For the missing data, it means that the medalist is a non-winning player, so it is labeled as 0. At the same time, I deleted other data with missingness. The percentage of people who got medals in the processed data is about 15%.

2.2. Data distribution

I did some preliminary analysis of the data and obtained the following information:

(1) Female athletes are approximately half the size of their male counterparts in all data sets.

(2) In terms of average data, female athletes are lower in height and weight than male athletes.

(3) There are also cases where some female athletes are physically stronger than most male athletes.

3．Model selection & methodology

3.1.Fairness of play

In this section, I plan to study the effect of gender and country on the probability of winning the award for several reasons:

(1) To verify the fairness and objectivity of the Olympics.

(2) To understand the differences between countries, to improve the training of athletes, and to develop competition strategies.

(3) To promote cultural and sports exchanges.

The following methods are used:

(1) Count the number of prizes won by athletes of different genders and calculate the probability of winning and plot the graphs using ggplot2.

(2) Counting the probability of winning for different countries and plotting the graphs using ggplot2.

3.2.Physical Fitness Requirements for Different Sports

In this section, I plan to examine the differences in physical fitness of athletes who have won awards in different sports. There are several main objectives:

(1) To understand the characteristics and requirements of different sports.

(2) To assist in scientific training and selection.

(3) To promote the all-round development of athletes.

The main methods used are:

Firstly, I use unsupervised learning method to cluster the winning athletes. I use the K-Means method to classify the winning athletes into 4 categories. Secondly, I analyze the data of the four classes statistically and mine the data features. Finally, I use word frequency analysis and draw word clouds to analyze the information on different types of athletes' adaptation to sports.

3.3.Prediction of Basketball Program Awards

In this section, I would like to build predictive models to predict whether or not the athlete is likely to get a medal in basketball by using information about the athlete's physical fitness. This method can help coaches to improve to make business decisions, improve team sports, and also guide players training to some extent.

In order to fully improve the precision of the model, I divided the basketball players' data into training set and validation set according to the ratio of 8:2. At the same time, I constructed an evaluation function with recall, precision, and F1 score as indicators, and selected three models, namely, random forest, logistic regression, and decision tree, for testing.

4．Results

4.1.Fairness of play

Comparisons show that the Olympics are more equitable in terms of gender, with male and female athletes having almost the same probability of winning. However, when the athletes' countries are different, the probability of winning fluctuates greatly. This problem may be because different countries have different economic and sports levels, which leads to great differences in the selection and training strategies of participating athletes. Athletes from some countries may have a stricter selection system and more scientific training, and therefore have a higher chance of winning awards.

4.2.Physical Fitness Requirements for Different Sports

I used the K-Means method to compare and analyze the data after dividing them into four groups. The results show that among the four groups of athletes, the first and third groups have a higher percentage of male athletes, which is about 97% and 100% respectively, the second group is female athletes, and the fourth group has a percentage of male athletes of about 87%.

Group Number	Male player rate
1	97.4%
2	0%

3	100%
4	87.3%

By comparing the ages of the four groups of data we found that the average age of the athletes in the fourth group was significantly higher than that of the other three groups of athletes, while the physical fitness of the athletes in the first group was significantly higher than that of the other three groups of athletes.

Therefore, I think the characteristics of the four groups of athletes are respectively:

Type 1: Physically dominant

Type 2: average female

Type 3: average male

Type 4: Veteran type

Further, I conducted word frequency statistics and drew word clouds for the programs in which the four groups of award-winning athletes participated respectively, and according to the results I found the following patterns:

- sports such as water polo and basketball are more physically demanding
- fencing, rowing and other programs require more experience
- swimming, gymnastics, soccer, etc. do not require strong physical confrontation and are more skill-oriented.

- Athletics covers a large number of different events, which cannot be analyzed for the moment due to the level of detail of the data, but most of the winners are either physically strong or experienced.

4.3.Prediction of Basketball Program Awards

I selected the basketball program and constructed a model to determine whether an athlete has the potential to win an award. The athlete's gender, age, height, and weight data are used as independent variables, and whether or not they win awards is used as the dependent variable.

Comparing the three different methods of random forest, logistic regression, and decision tree, the results are as follows.

Methodology	Recall	Precision	F1 Score
Random Forest	0.2	0.404	0.268
Logistic Regression	0	0	0
Decision Tree	1	1	1

Therefore, the decision tree method has a very high performance in mining potential players.

5．Conclusion

Through the research of this project, I have come to the following main conclusions:

1. The Olympic Games are a relatively fair sporting competition in which both male and female athletes have an almost equal chance of winning. However, due to different underlying conditions, athletes from some countries have relatively higher chances of winning, despite the fact that the Games give most countries the opportunity to participate. Therefore, I believe that we should strengthen the exchange of sports culture and training methods between different countries, especially between disadvantaged countries and strong countries, to enhance the fairness of the competition and improve the level of sports around the world.

2. Through analyzing the data of the winning athletes, we can find that different sports have different requirements for athletes' physical quality. Water polo, basketball and other sports have higher physical requirements; fencing, rowing and other sports have higher experience requirements; while swimming, gymnastics, soccer and other sports do not require strong physical confrontation and pay more attention to skills. Therefore, when selecting athletes, sports with higher physical requirements should focus on selecting athletes with advantageous innate conditions for training. For sports with higher experience requirements, athletes with better mental qualities can be appropriately selected to focus on training and increase their experience through a large number of competitions. It is recommended that athletes in such sports should not choose to retire because of the difficulty in winning medals at a young age.

3. The decision tree model used in this study can determine whether a basketball player has the potential to win awards, and I think this model can help coaches choose athletes with more potential. Of course, in conjunction with reality, any model may have errors, and coaches are also encouraged to choose athletes who are not physically outstanding, but whose skills, mentality, enthusiasm and other factors are suitable for participating in the competition.

6．Appendices

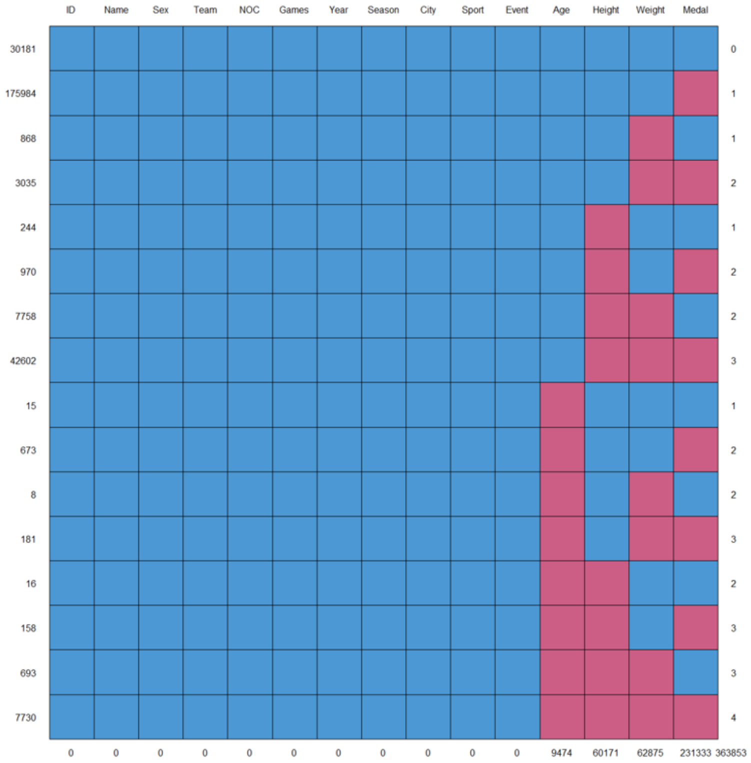


Figure 1 Missing Data

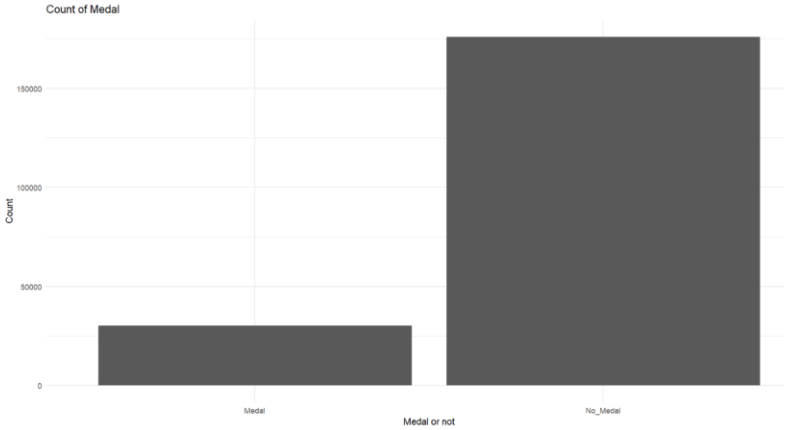


Figure 2 Comparison of the amount of data on whether medals were won or not

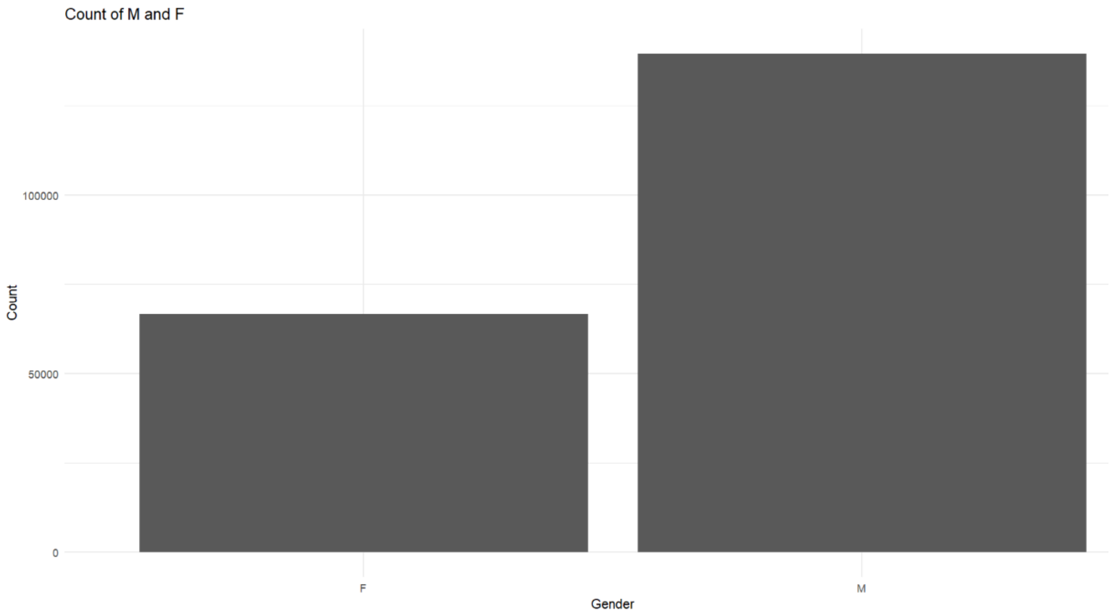


Figure 3 Percentage of athletes of different genders

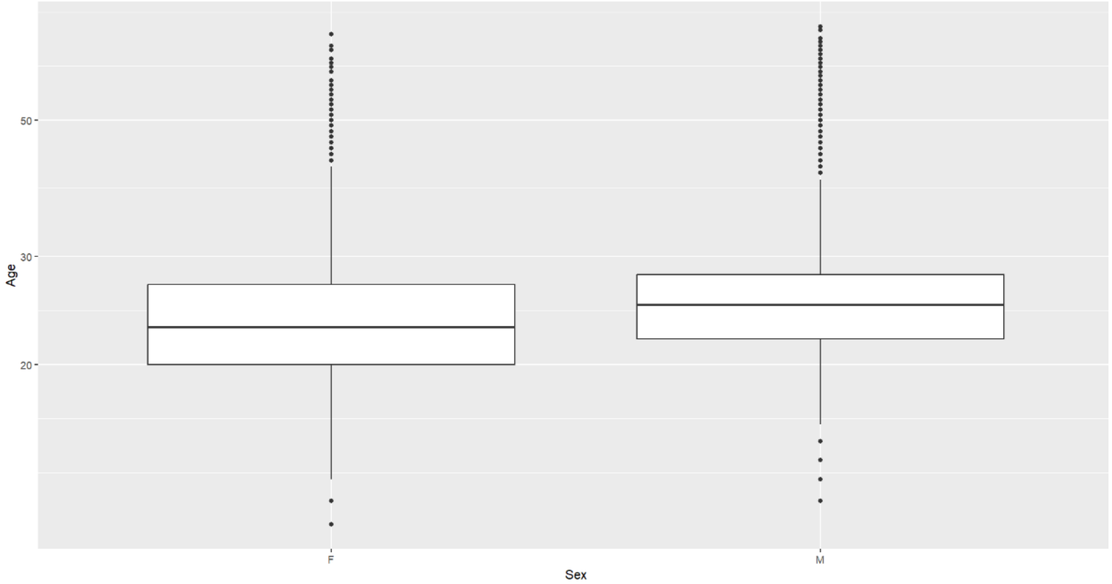


Figure 4 Age difference of athletes of different genders

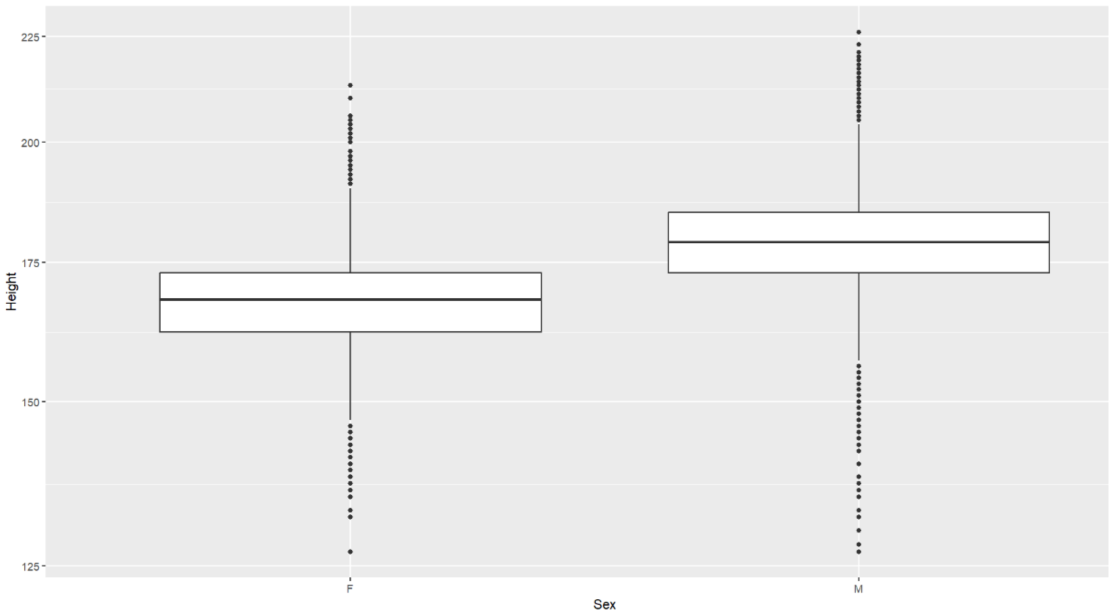


Figure 5 Difference in height of athletes of different genders

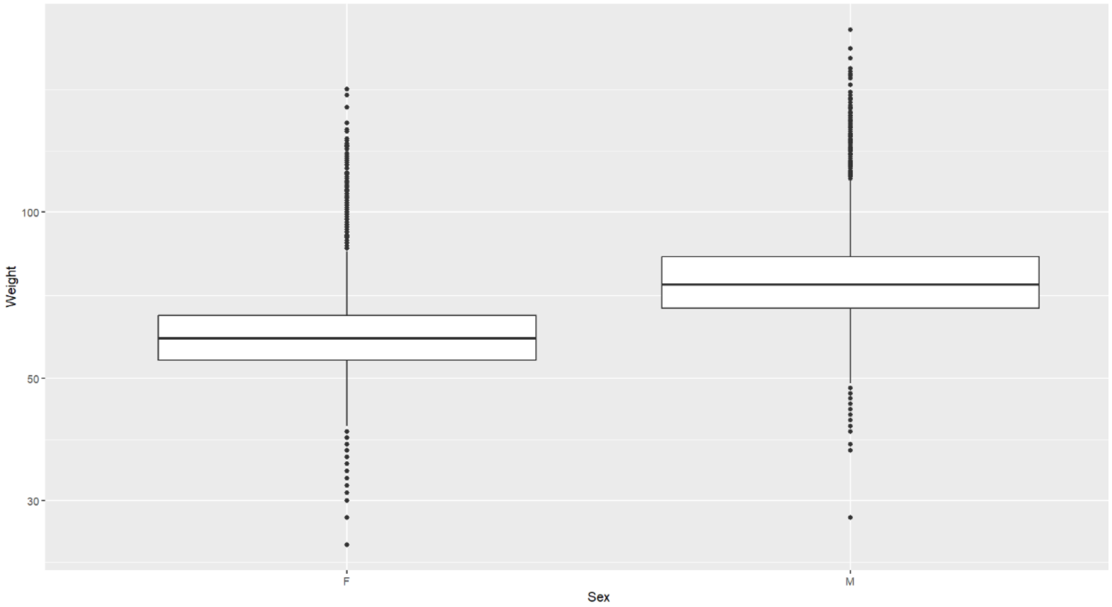


Figure 6 Difference in weight of athletes of different genders

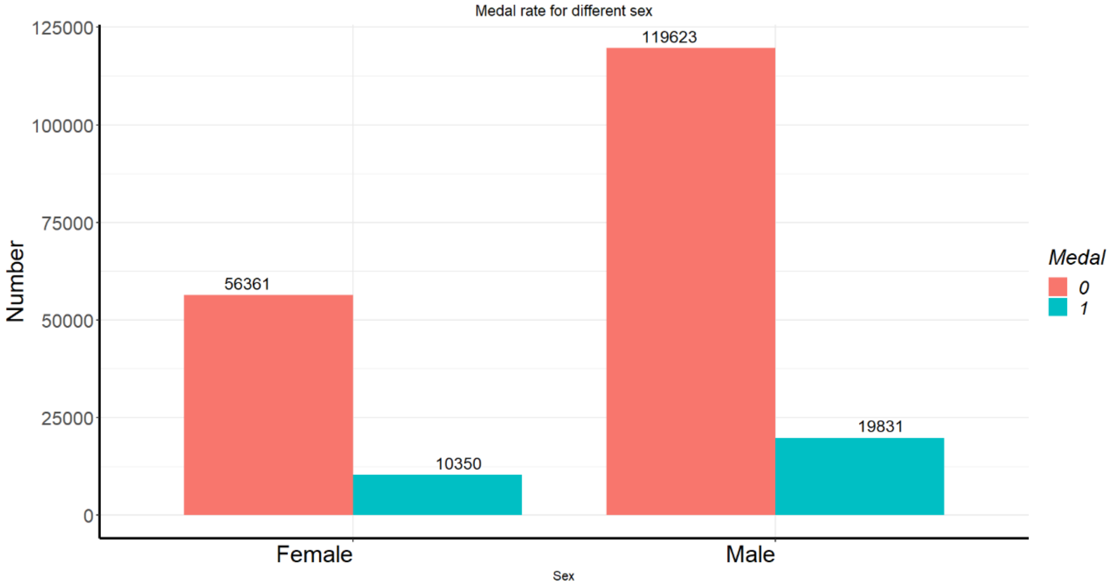


Figure 7 Probability of winning an award by sex

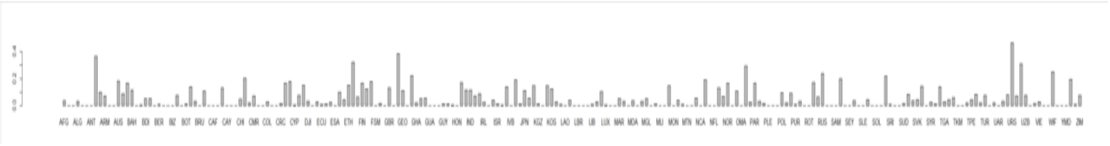
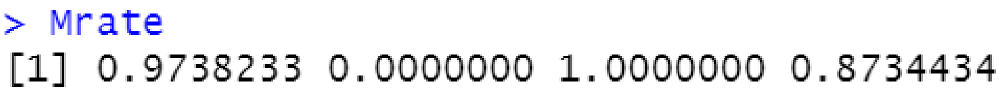


Figure 8 Probability of winning an award by country



```
> Mrate
[1] 0.9738233 0.0000000 1.0000000 0.8734434
```

Figure 9 Percentage of males for the four groups of data

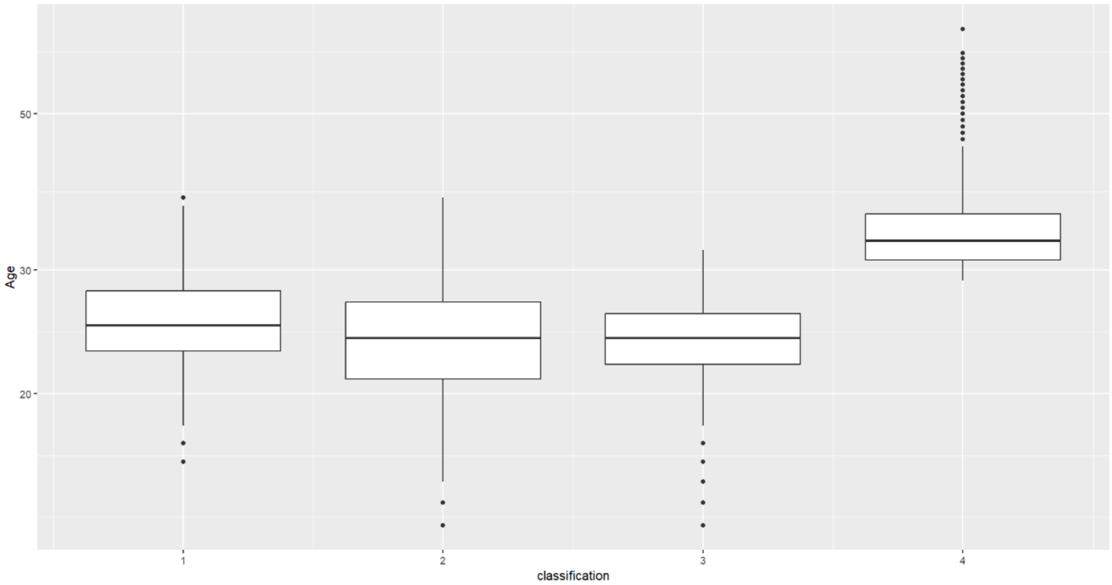


Figure 10 Age distribution of the four groups of data

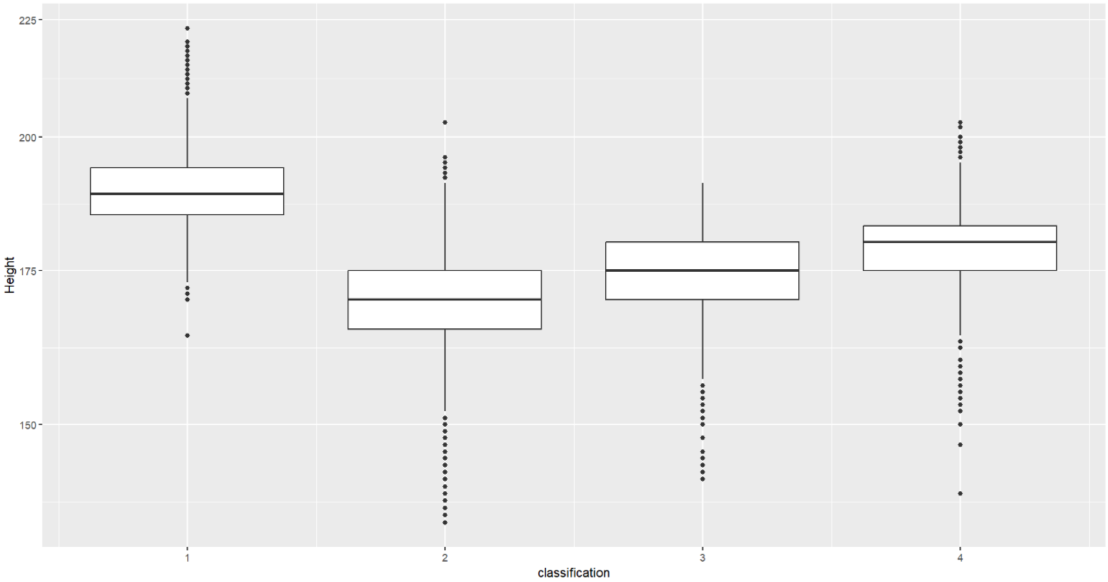


Figure 11 Height distribution of the four data sets

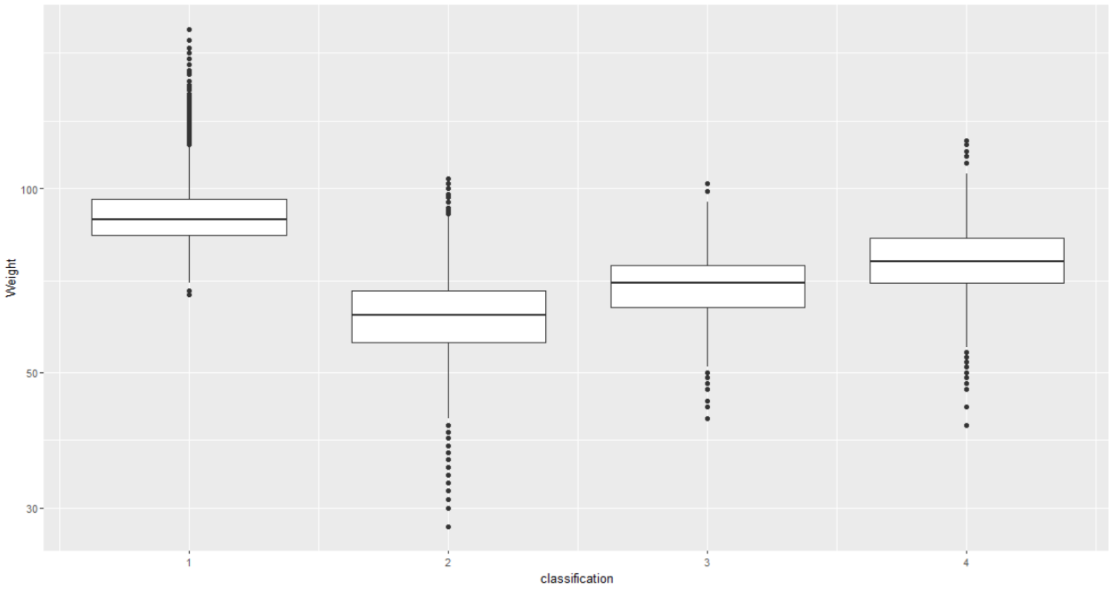


Figure 12 Weight distribution of the four data sets

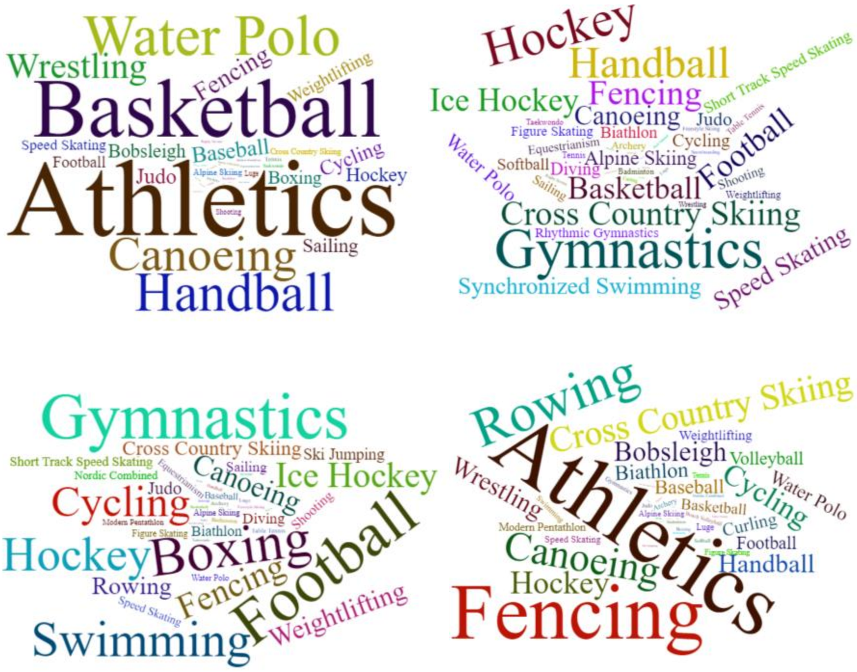
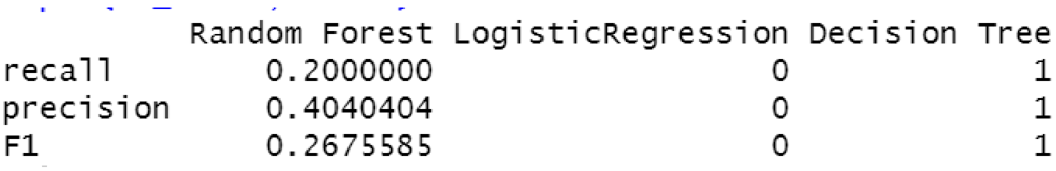


Figure 13 Word cloud statistics for the four data sets



```
          Random Forest LogisticRegression Decision Tree
recall        0.2000000                  0             1
precision     0.4040404                  0             1
F1            0.2675585                  0             1
```

Figure 14 Comparison of the performance of the three models