Table of Contents

[Preparing the data 3](#_Toc532752253)

[Analyzing the data 4](#_Toc532752254)

[A) Do the analysis as in Module 3 for at least one categorical variable and at least one numerical variable. Show appropriate plots for your data. 4](#_Toc532752255)

[B) Do the analysis as in Module 3 for at least one set of two or more variables. Show appropriate plots for your data. 9](#_Toc532752256)

[C) Pick one variable with numerical data and examine the distribution of the data. 11](#_Toc532752257)

[D) Draw various random samples of the data and show the applicability of the Central Limit Theorem for the variable 13](#_Toc532752258)

[E) Show how various sampling methods can be used on your data. What are your conclusions if these samples are used instead of the whole dataset. 15](#_Toc532752259)

[Implementation of any feature(s) not mentioned in the specification 17](#_Toc532752260)

[References: 21](#_Toc532752261)

**Preparing the data**

Dataset“120 years of Olympic history: athletes and results” is basic biographical data on athletes and medal results from Athens 1896 to Rio 2016. This dataset file “athlete\_events.csv” was taken from [www.kaggle.com](http://www.kaggle.com) site (<https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete_events.csv>). Since I have decided to analyze the data only for Kazakhstan team, I subset the specific data and exported it into "kazAthlete.csv" file. Kazakhstan started to participate in Olympic games from 1994 after becoming an independent country.

*> athlete\_events <- read.csv("~/Desktop/CS544Final\_Kalikova/athlete\_events.csv")*

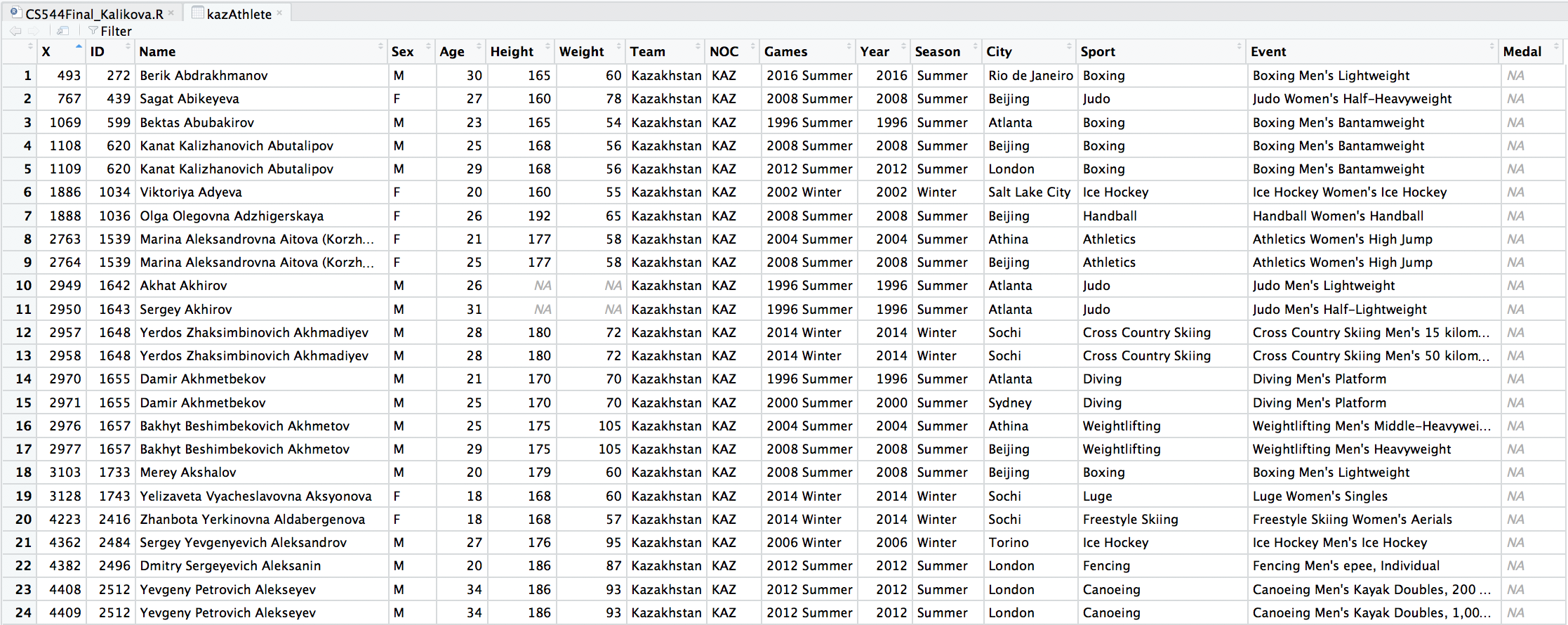
*> athletesKazakhstan = subset(athlete\_events, Team == "Kazakhstan")*

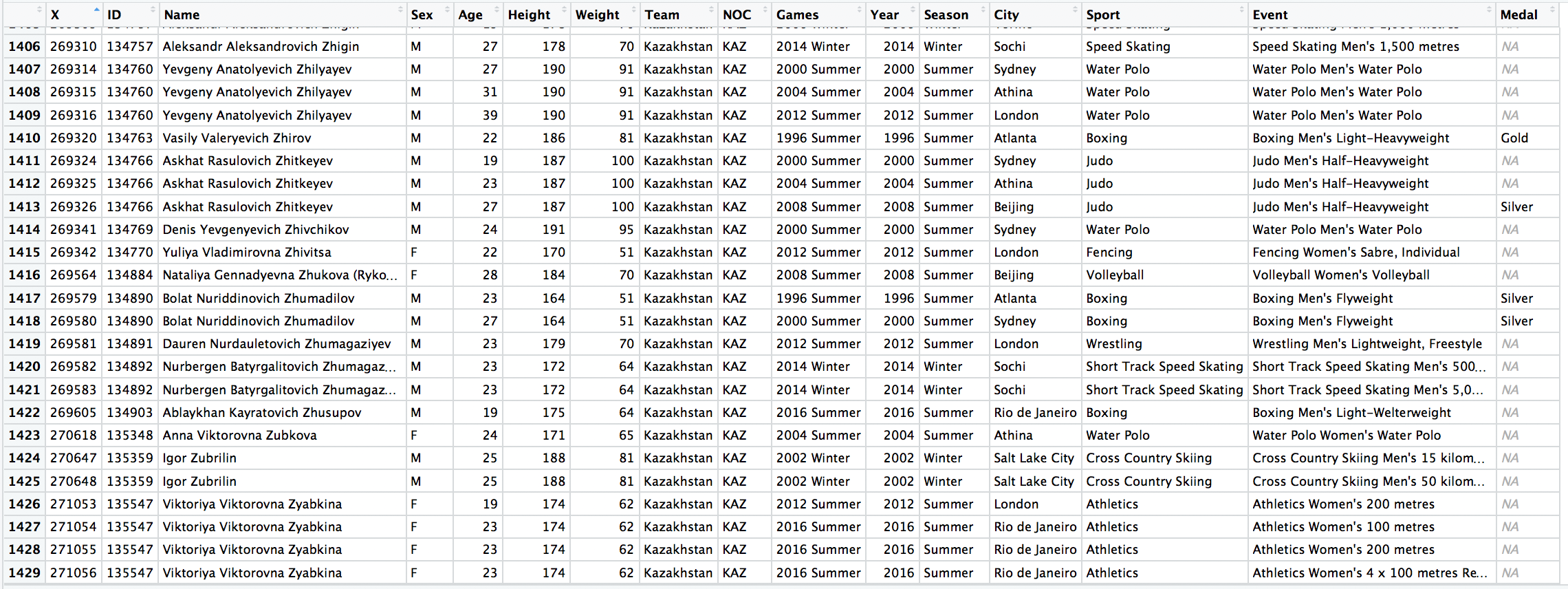
*> write.csv(athletesKazakhstan, "kazAthlete.csv")*

The “kazAthlete” dataset contains 1429 rows and 16 variables. The columns that mostly have been analyzing are Name, Sex, Age, Height, Weight, Year, Season, City, Sport, Medal. The “kazAthlete” dataset was imported into R.

*> kazAthlete <- read.csv("~/Desktop/CS544Final\_Kalikova/kazAthlete.csv")*

The first and last 24 rows of the “kazAthlete” dataset are shown below:

**

**

# **Analyzing the data**

**A) Do the analysis as in Module 3 for at least one categorical variable and at least one numerical variable. Show appropriate plots for your data.**

Categorical variables

The variables Sex and Sport were taken into analysis for categorical variables. There are 549 female athletes and 880 male athletes from Kazakhstan, which means that from all athletes from Kazakhstan who participated in Olympic games from 1994 till 2016, 38% was women and 62% men.

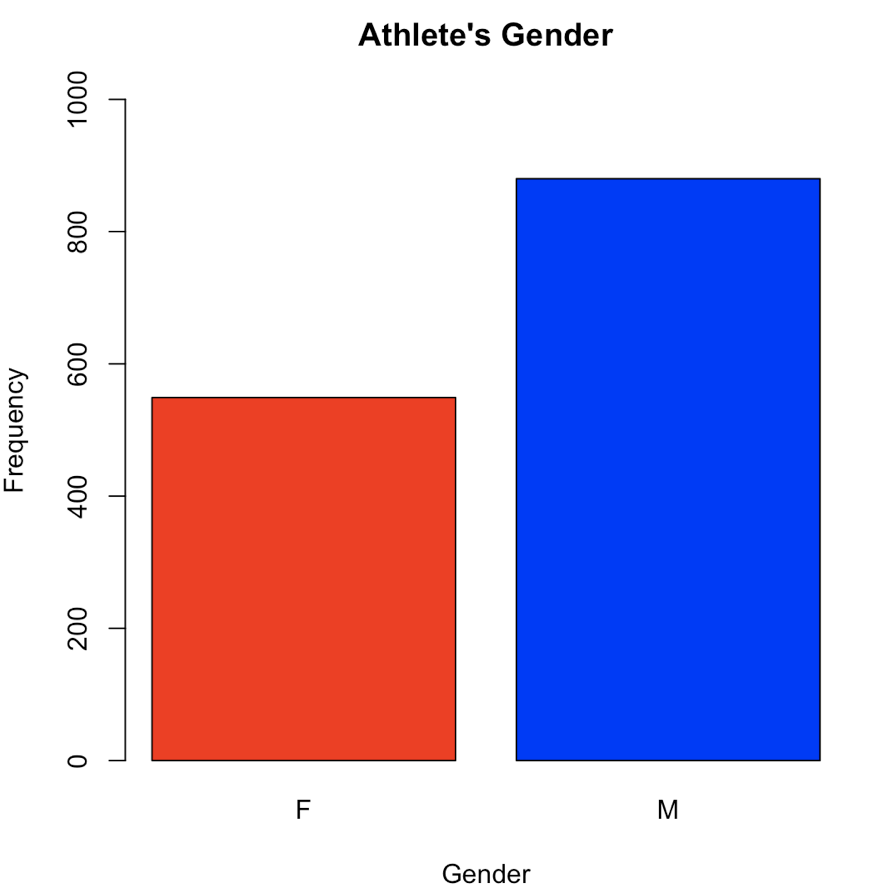
*> gender = table(kazAthlete$Sex, exclude = FALSE)*

*> gender*

F M

549 880

*> gender/length(kazAthlete$Sex)*



F M

0.3841847 0.6158153

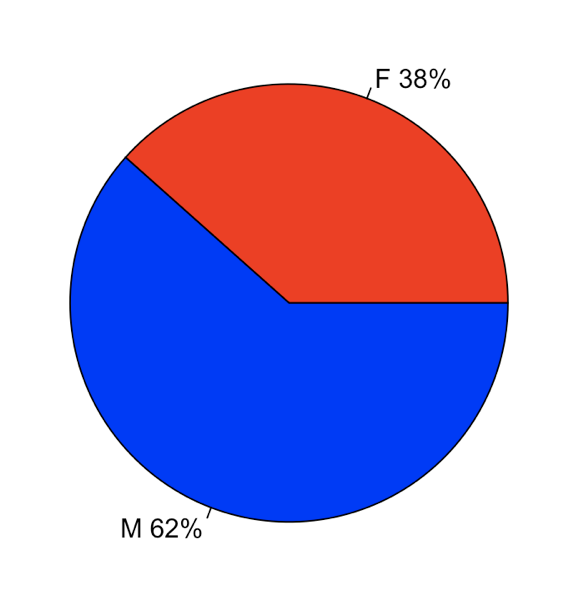
The barplot and piechart shown for Sex variable.

*> barplot(gender,*

*+ col = c("red", "blue"), ylim = c(0, 1000),*

*+ main = "Athlete's Gender",*

*+ xlab = "Gender", ylab = "Frequency")*

**

*> slice.labels = names(gender)*

*> slice.percents = round(gender/sum(gender) \*100)*

*> slice.labels = paste(slice.labels, slice.percents)*

*> slice.labels = paste(slice.labels, "%", sep = "")*

*> pie(gender, labels = slice.labels,*

*+ col = c("red", "blue"))*

*>*

The barplot for categorical variable Sport is shown below. The athletes from Kazakhstan performed in 36 Sport. The higher number of athletes (240) was presented in Cross Country Skiing.

*> sports = table(kazAthlete$Sport)*

*> sports*

Alpine Skiing Archery Athletics Biathlon

31 27 138 94

Boxing Canoeing Cross Country Skiing Cycling

56 68 240 41

Diving Fencing Figure Skating Freestyle Skiing

13 10 11 22

Gymnastics Handball Ice Hockey Judo

45 14 64 49

Luge Modern Pentathlon Rhythmic Gymnastics Rowing

1 10 4 8

Shooting Short Track Speed Skating Ski Jumping Snowboarding

54 15 50 2

Speed Skating Swimming Synchronized Swimming Table Tennis

68 68 10 2

Taekwondo Tennis Trampolining Triathlon

9 8 1 6

Volleyball Water Polo Weightlifting Wrestling

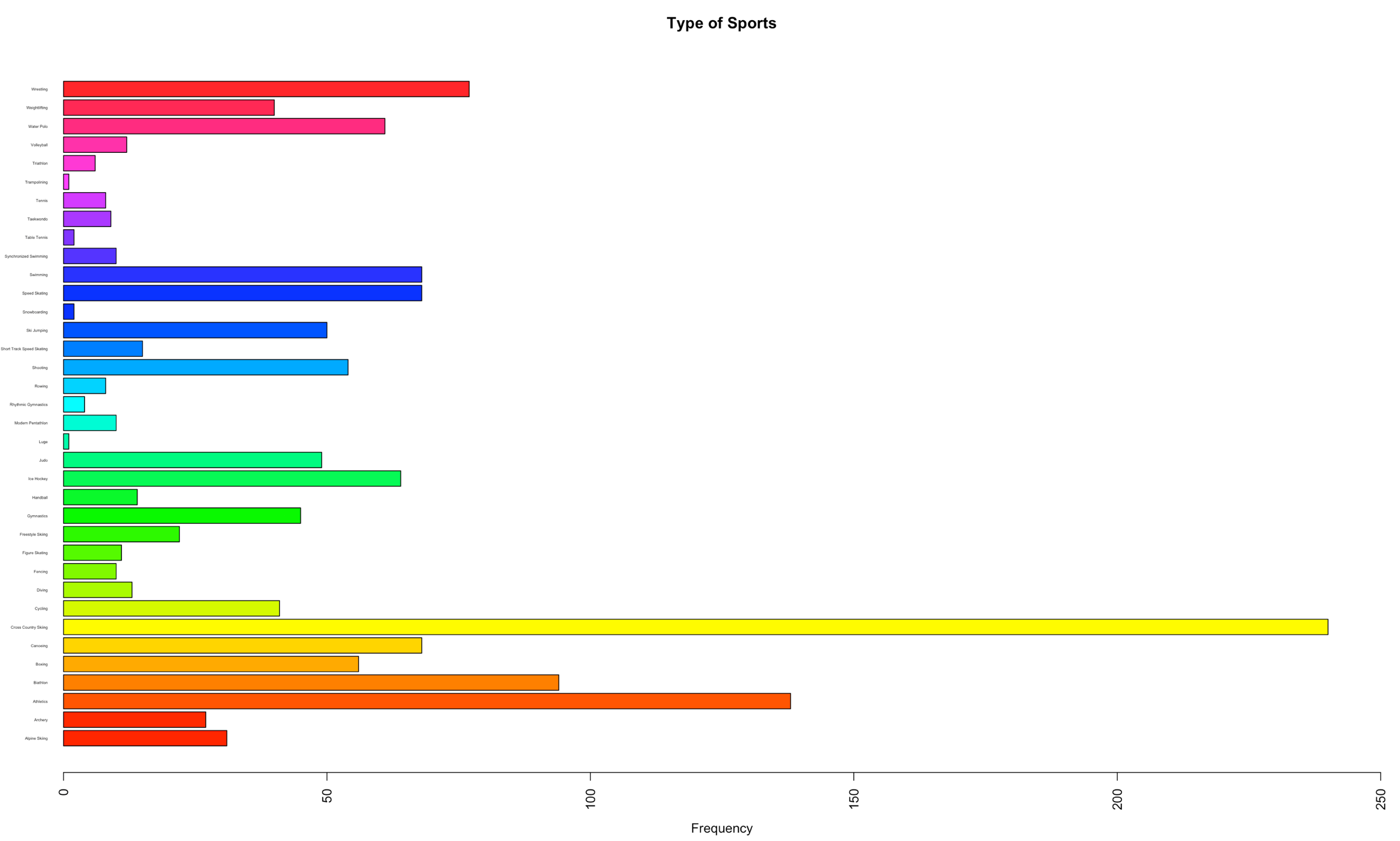
12 61 40 77

*> barplot(sports, col = rainbow(36), xlim = c(0, 250),*

*+ horiz = TRUE, las = 2, cex.names=0.3,*

*+ main = "Type of Sports", xlab = "Frequency")*

*>*



Numerical variables

The numerical variables Year, Height, and Weight were analyzed. Kazakhstan team started to participate in Olympic games starting from 1994 year. Year 2000 was the year when the higher number (158) of Kazakhstan’s athletes were presented on Olympic games.

*> table(kazAthlete$Year)*

1994 1996 1998 2000 2002 2004 2006 2008 2010 2012 2014 2016

71 142 114 158 101 130 97 146 100 132 115 123

*> #analyzing which year had the higher number of athletes*

> which(table(kazAthlete$Year) == max(table(kazAthlete$Year)))

2000

4

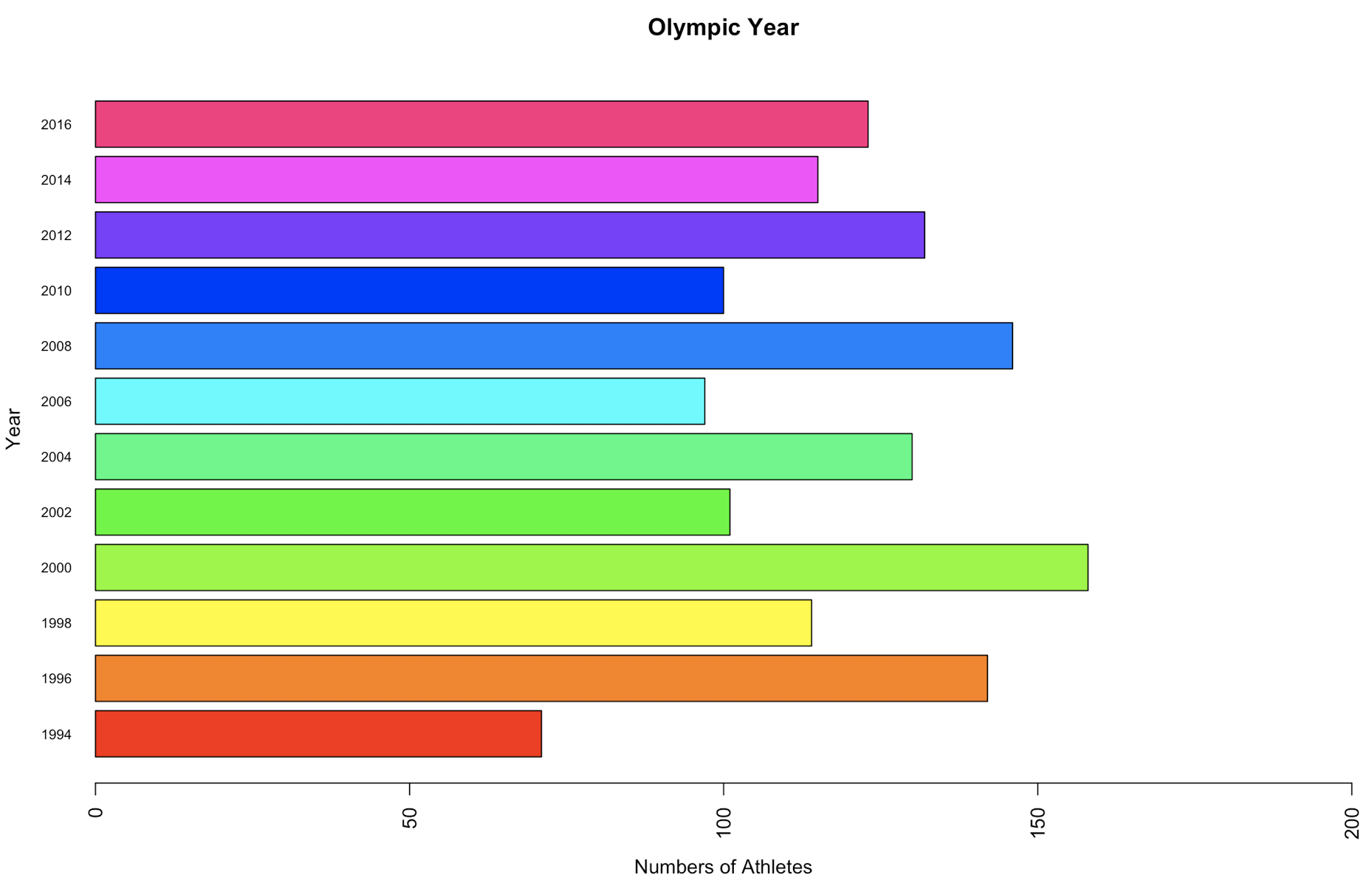
*> barplot(table(kazAthlete$Year), horiz = TRUE,*

*+ col = rainbow(12), xlim = c(0, 200),*

*+ main = "Olympic Year", las = 2, cex.names=0.7,*

*+ ylab = "Year", xlab = "Numbers of Athletes")*

*>*



The barplot for variable Height is shown below. The min height is 150 cm (3 athletes), whereas the max height is 205 cm (2 athletes). There is no information about heights of 74 athletes. 101 athletes have the same height which is 178.

*> table(kazAthlete$Height, exclude = FALSE)*

150 152 153 154 155 156 157 158 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176

3 8 5 4 7 10 8 31 33 29 32 11 28 46 36 26 34 9 70 34 24 62 40 50 52

177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 201 202 205

33 101 25 98 29 30 38 46 49 29 40 26 13 38 10 8 7 11 8 7 3 7 4 1 2

<NA>

74

*> which(table(kazAthlete$Height) == max(table(kazAthlete$Height)))*

178

27

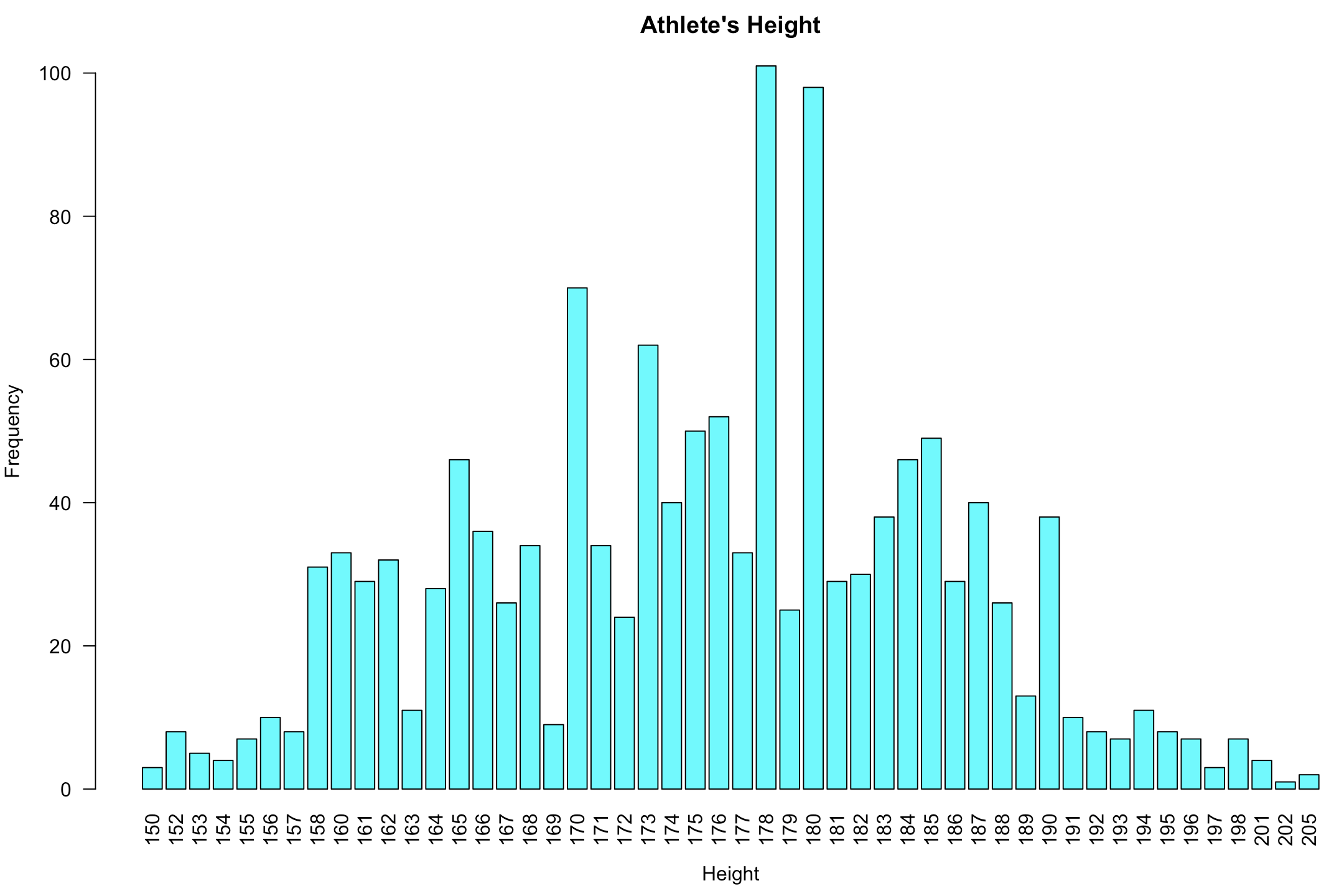
*> barplot(table(kazAthlete$Height),*

*+ col = "cyan", ylim = c(0, 100),*

*+ main = "Athlete's Height", las = 2,*

*+ xlab = "Height", ylab = "Frequency")*

>



The barplot for variable Weight is shown below. The min weight is 42 kg (1 athlete), whereas the max weight is 150 kg (2 athletes). There is no information about weights of 80 athletes. 85 athletes have the same weight which is 70.

*> table(kazAthlete$Weight, exclude = FALSE)*

42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66

1 5 5 3 2 3 21 3 30 16 7 25 56 28 42 21 27 11 53 3 49 44 46 41 20

67 68 69 70 71 72 73 74 75 76 76.5 77 78 79 80 81 82 83 84 85 86 87 88 89 90

50 49 22 85 13 44 25 20 61 20 2 17 51 11 34 17 5 25 15 38 6 14 12 1 23

91 92 93 94 95 96 97 98 99 100 101 102 104 105 107 113 118 120 125 130 138 144 150 <NA>

9 13 4 8 17 6 10 12 1 14 1 2 4 7 1 1 2 6 2 2 2 1 2 80

*> #analyzing the higher number of athletes' weight*

*> which(table(kazAthlete$Weight) == max(table(kazAthlete$Weight)))*

70

29

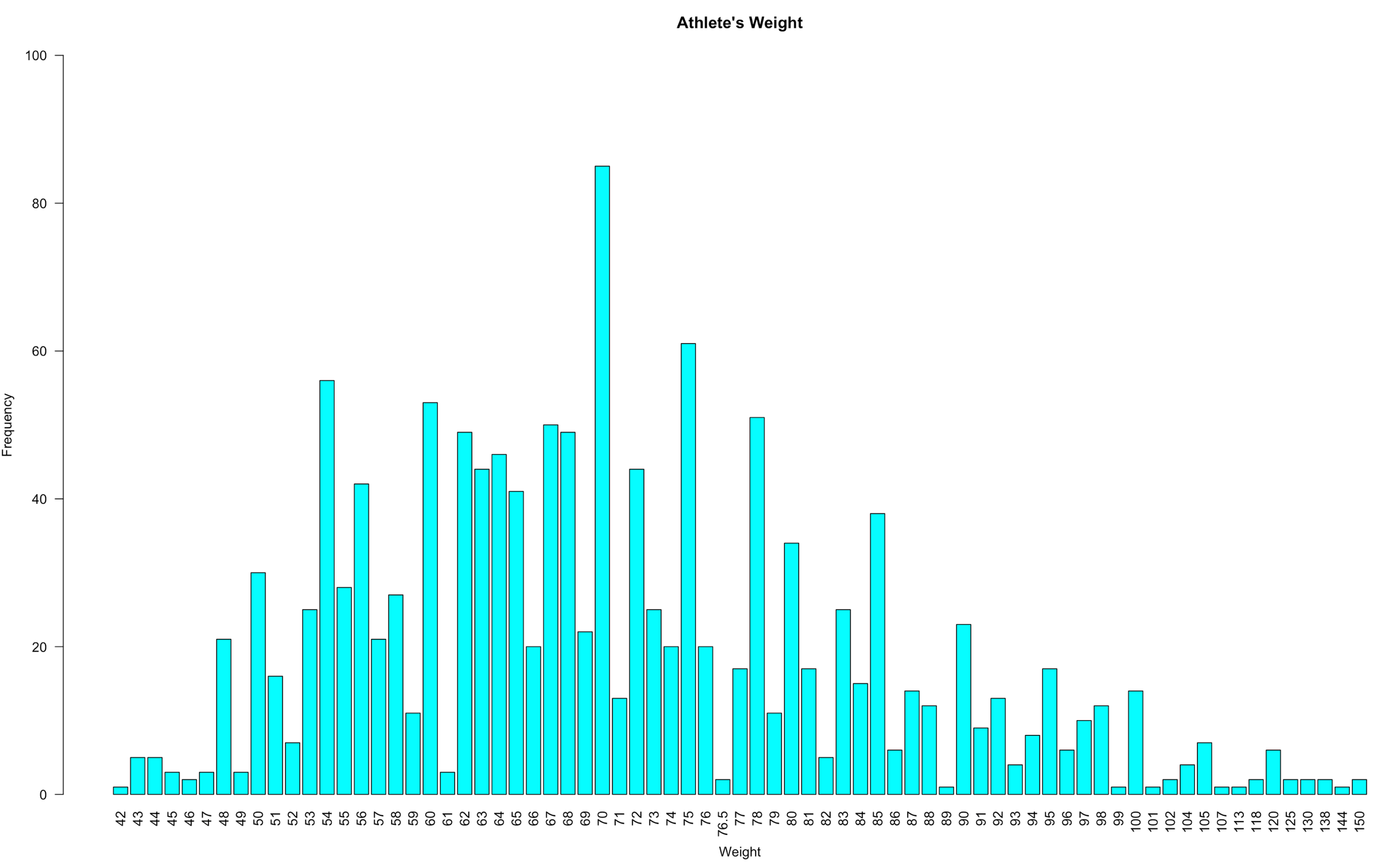
*> barplot(table(kazAthlete$Weight),*

*+ col = "cyan", ylim = c(0, 100),*

*+ main = "Athlete's Weight", las = 2,*

*+ xlab = "Weight", ylab = "Frequency")*

*>*

**

**B) Do the analysis as in Module 3 for at least one set of two or more variables. Show appropriate plots for your data.**

The set of Sex and Medal variables were analyzed and showed in mosaicplot and barplots.

There are 20 gold, 25 silver, and 32 bronze medals, which is overall 77 medals were won by athletes from Kazakhstan. Female athletes have won 21 medals: gold – 5, silver – 5, and bronze – 11. Male athletes have won 21 medals: gold – 15, silver – 20, and bronze – 21.

*> attach(kazAthlete)*

*The following objects are masked from kazAthlete (pos = 3):*

*Age, City, Event, Games, Height, ID, Medal, Name, NOC, Season, Sex, Sport, Team, Weight, X, Year*

*>*

*> table(Medal, exclude = FALSE)*

Medal

Bronze Gold Silver <NA>

32 20 25 1352

*> medals = table(Sex, ordered(Medal, levels = c("Gold", "Silver", "Bronze")))*

*> medals*

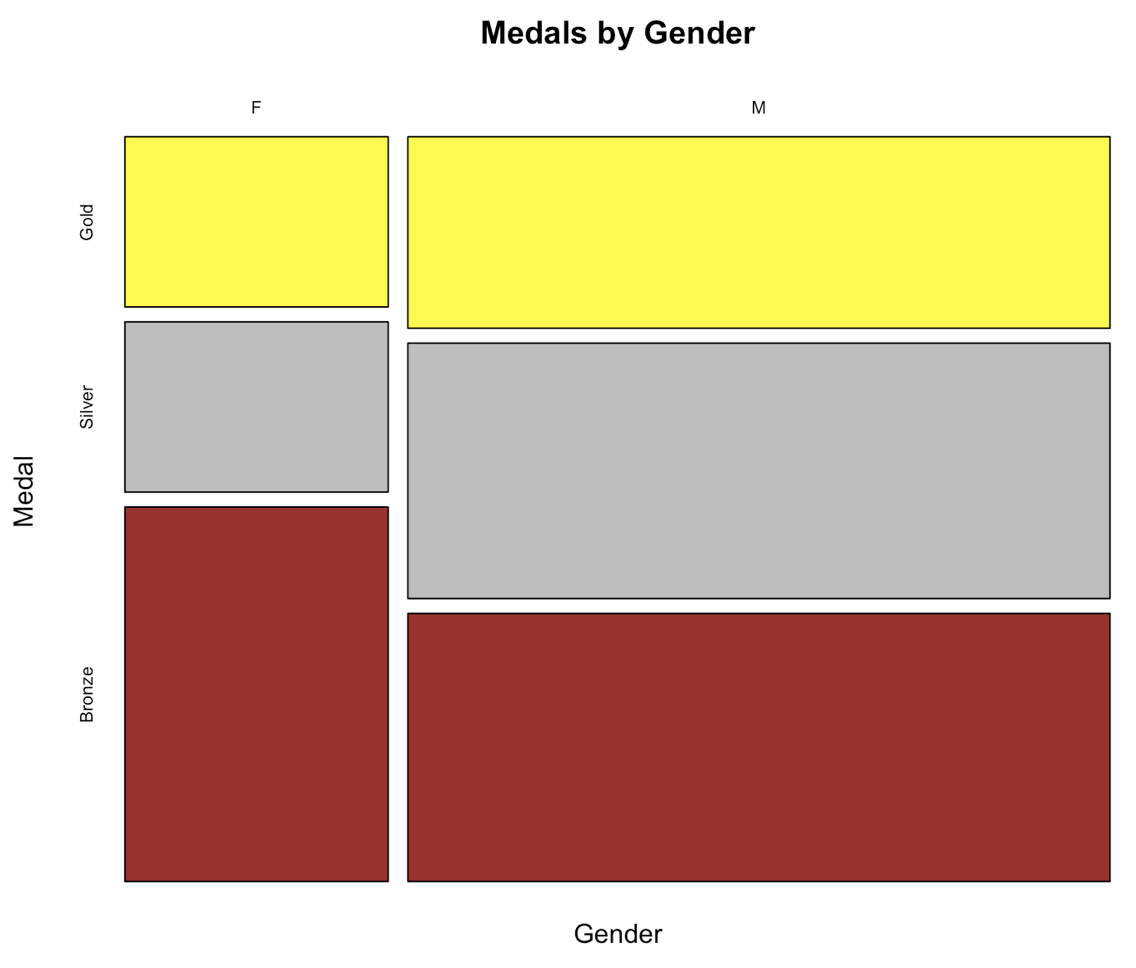
Sex Gold Silver Bronze

F 5 5 11

M 15 20 21

*> addmargins(medals)*

Sex Gold Silver Bronze Sum

 F 5 5 11 21

M 15 20 21 56

Sum 20 25 32 77

*> prop.table(medals)*

Sex Gold Silver Bronze

F 0.06493506 0.06493506 0.14285714

M 0.19480519 0.25974026 0.27272727

*> mosaicplot(medals, col= c("yellow", "grey", "brown"),*

*+ main = "Medals by Gender",*

*+ ylab = "Medal", xlab = "Gender")*

*>*

*> par(mfrow = c(1,2))*

*> barplot(medals, beside = TRUE, ylim = c(0, 25),*

*+ legend.text = TRUE, args.legend = list(x = "top"),*

*+ col = c("red", "blue"), xlab = "Medal")*

*>*

*> barplot(t(medals), beside = TRUE, ylim = c(0, 30),*

*+ legend.text = TRUE, args.legend = list(x = "top"),*

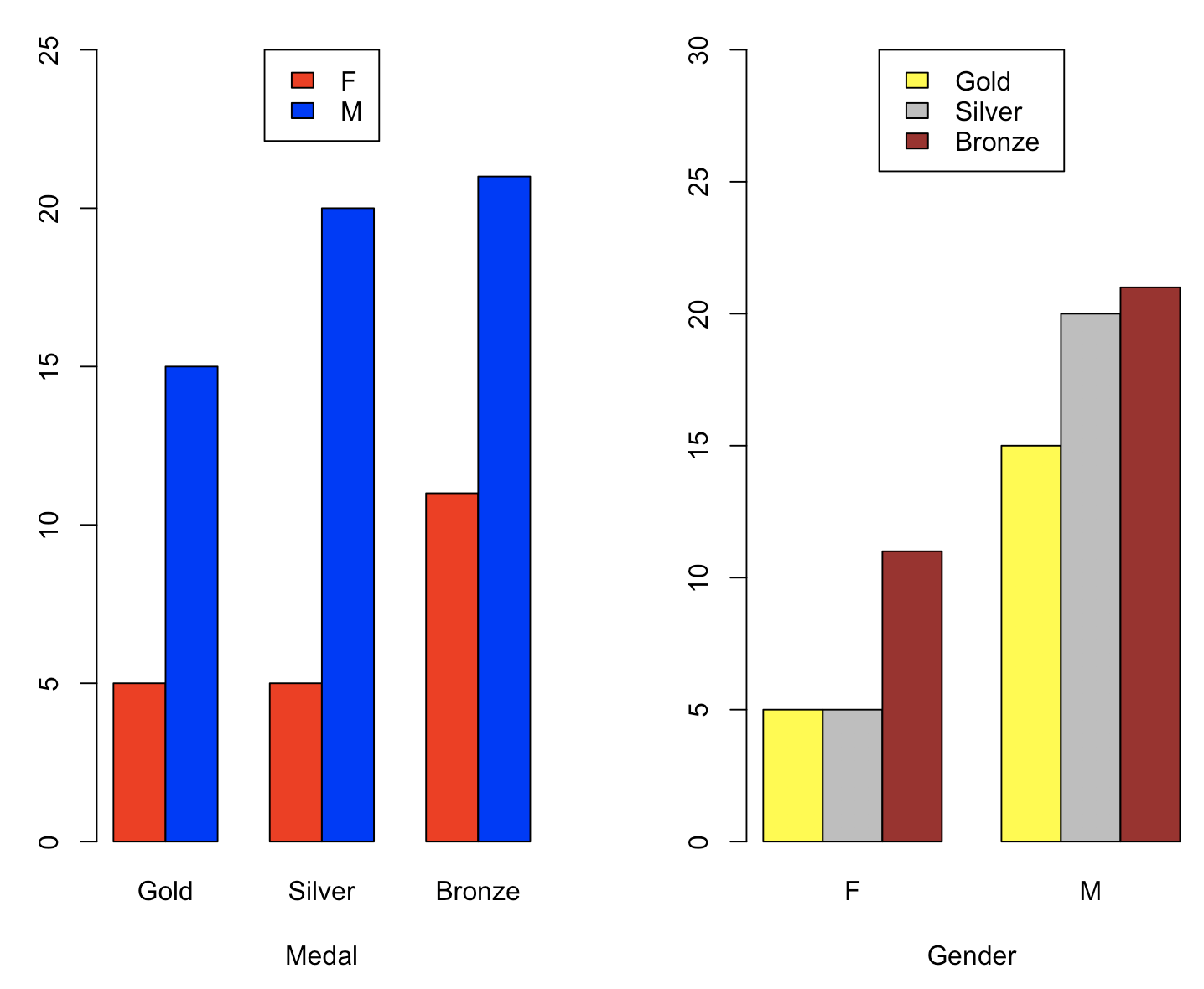
*+ col = c("yellow", "grey", "brown"), xlab = "Gender")*

*> par(mfrow = c(1,1))*

*>*

*> detach(kazAthlete)*

*>*



**C) Pick one variable with numerical data and examine the distribution of the data.**

The numerical variable Age was analyzed and the distribution of the Age data was examined. The average age among athletes is 25. The younger athlete’s age is 13, whereas the oldest athletes is 52 years old.

*> age = kazAthlete$Age*

*> mean(age)*

[1] 25.40938

*> median(age)*

[1] 25

*> var(age)*

[1] 24.16633

*> sd(age)*

[1] 4.915926

*> fivenum(age)*

[1] 13 22 25 28 52

*> summary(age)*

Min. 1st Qu. Median Mean 3rd Qu. Max.

13.00 22.00 25.00 25.41 28.00 52.00

*> f = fivenum(age)*

*> c(f[2] - 1.5\*(f[4] - f[2]),*

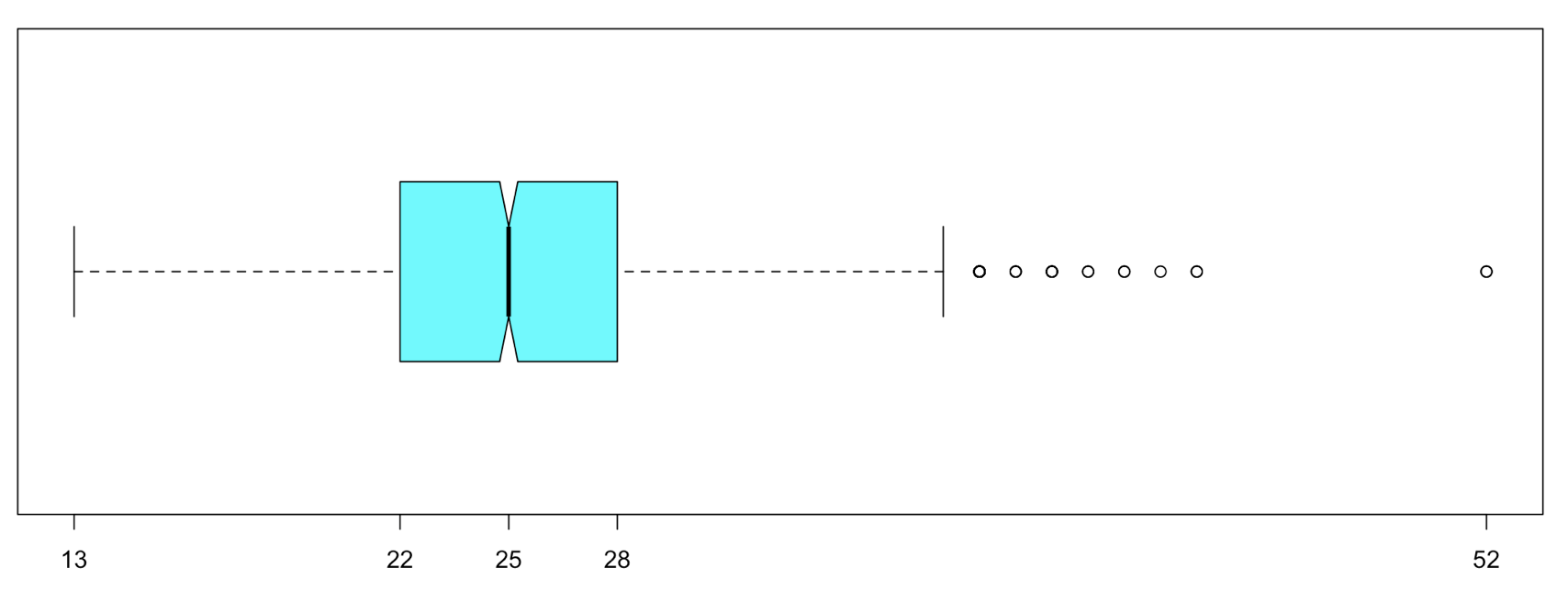
*+ f[4] + 1.5\*(f[4] - f[2]))*

[1] 13 37

*> boxplot(age, horizontal = TRUE, xaxt = "n", notch = TRUE, col = "cyan")*

*> axis(side = 1, at = fivenum(age), labels = TRUE)*

*>*



288 athletes from Kazakhstan are 22 and 24 years old, in each category 144 athletes. The distribution of athletes age are shown in barplot.

*> table(kazAthlete$Age)*

13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 52

1 1 5 9 14 34 61 68 92 144 131 144 111 99 83 106 70 49 43 34 40 19 14 16 13 12 3 4 2 2 1 2 2

*> which(table(kazAthlete$Age) == max(table(kazAthlete$Age)))*

22 24

10 12

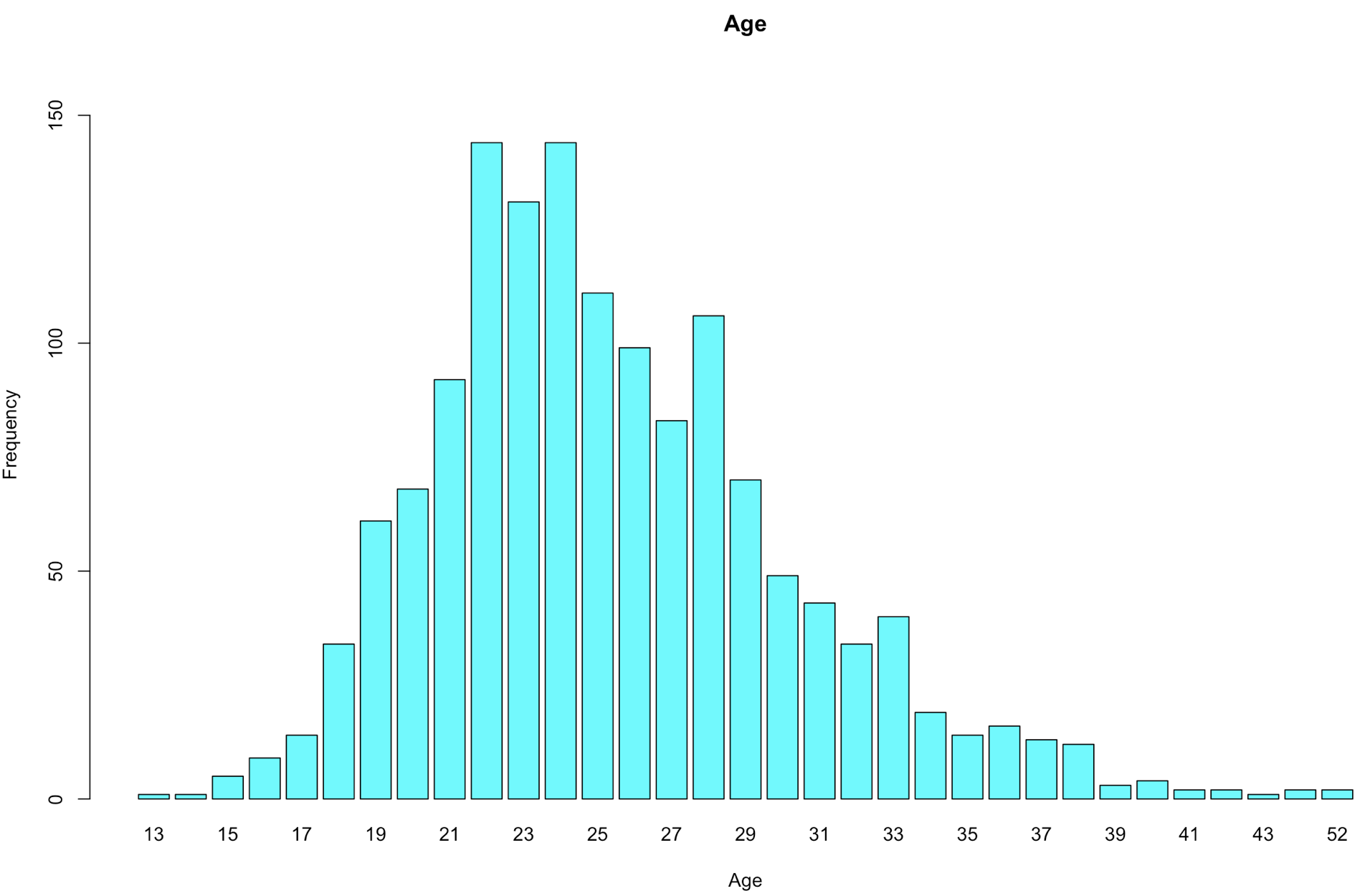
*> barplot(table(kazAthlete$Age),*

*+ col = "cyan", ylim = c(0, 160),*

*+ main = "Age",*

*+ xlab = "Age", ylab = "Frequency")*

*>*



**D) Draw various random samples of the data and show the applicability of the Central Limit Theorem for the variable**

The various random samples of the data and the applicability of the Central Limit Theorem for the variable Age were drawn. The mean of age is 25.4, whereas standard deviation equals 4.9. A histogram of the sample means distribution is shown below.

*> age = kazAthlete$Age*

*> mean(age)*

[1] 25.40938

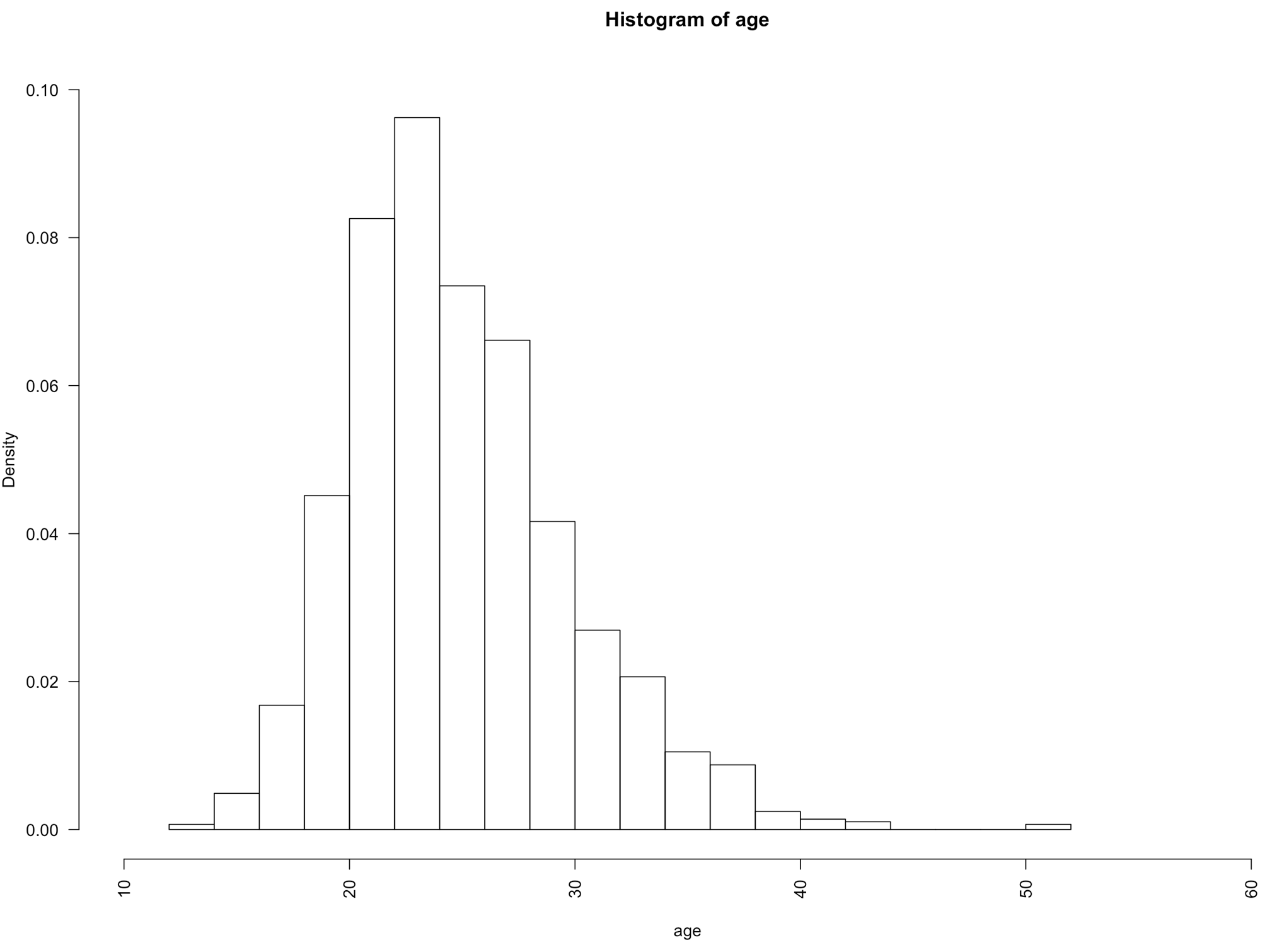
*> sd(age)*

[1] 4.915926

*> hist(age, prob = TRUE,*

*+ xlim = c(10, 60), ylim = c(0, 0.1),*

*+ breaks = 15, las = 2)*



With samples sizes of 5, 10, 20, 30, generated the data for 9000 samples using the normal distribution. The histogram of the densities of the sample means is shown below.

*> samples<- 9000*

*> xbar <- numeric(samples)*

*> par(mfrow = c(2,2))*

*> for (size in c(5, 10, 20, 30)) {*

*+ for (i in 1:samples) {*

*+ xbar[i] <- mean(sample(age, size = size, replace = FALSE))*

*+ }*

*+ hist(xbar, prob = TRUE,*

*+ breaks = 15, las = 2, ylim = c(0, 0.5),*

*+ main = paste("Sample Size =", size))*

*+ cat("Sample Size = ", size, " Mean = ", mean(xbar),*

*+ " SD = ", sd(xbar), "\n")*

*+ }*

Sample Size = 5 Mean = 25.39844 SD = 2.165085

Sample Size = 10 Mean = 25.41863 SD = 1.557535

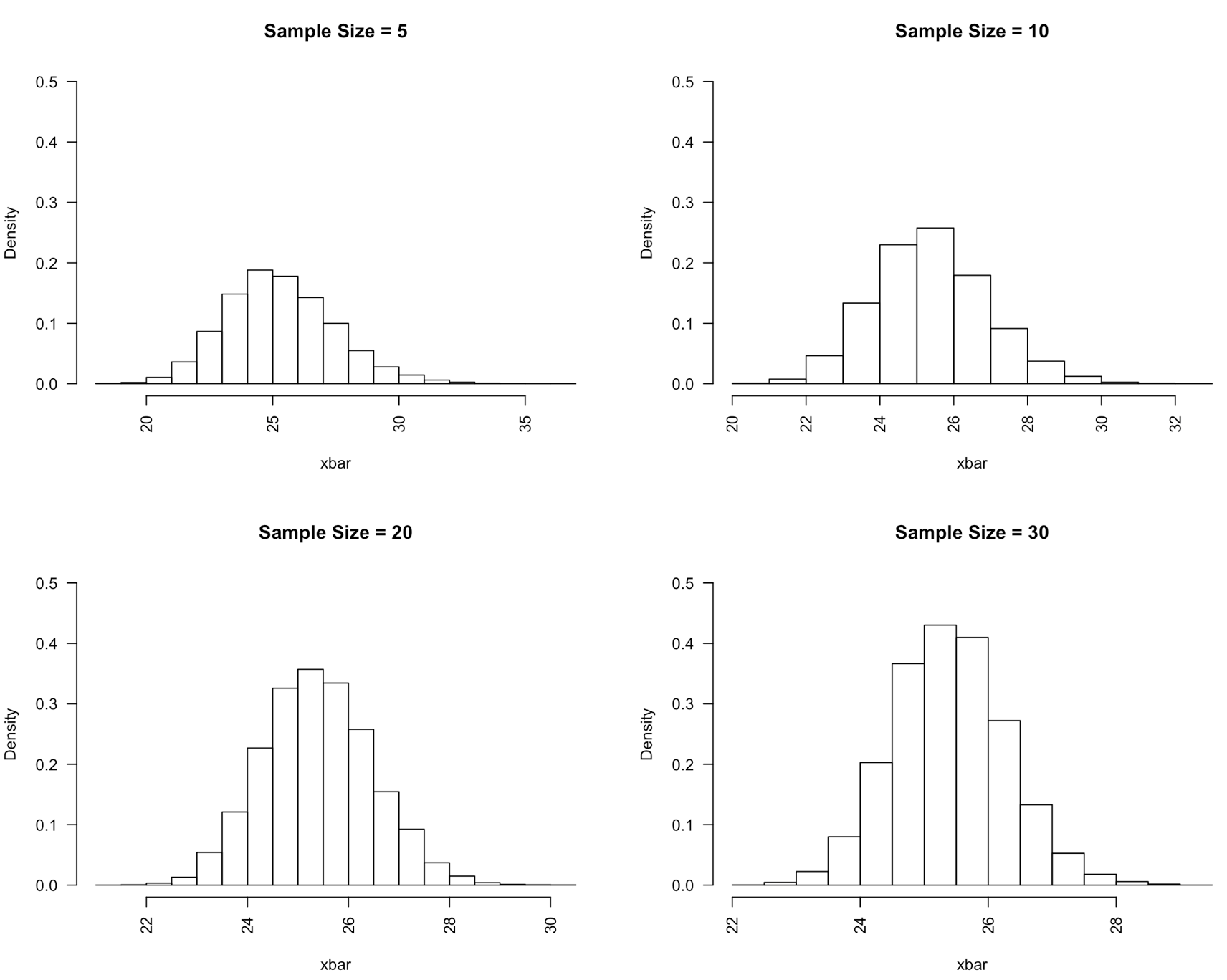
Sample Size = 20 Mean = 25.41458 SD = 1.086812

Sample Size = 30 Mean = 25.41037 SD = 0.8907015

*> par(mfrow = c(1,1))*

*>*

When the sample size getting larger, the distribution getting less right skewed, and approaching to the normal distribution shape. Also, when sample size is getting larger, then value of standard deviation is decreasing.



**E) Show how various sampling methods can be used on your data. What are your conclusions if these samples are used instead of the whole dataset.**

3 various probability sampling techniques were used on the data: simple random sampling, systematic sampling, and stratified sampling. The dataset has information about 1,429 athletes from Kazakhstan.

*> library(sampling)*

*> dataSample = kazAthlete*

*> head(dataSample[c(2,3,4,13)])*

ID Name Sex City

1 272 Berik Abdrakhmanov M Rio de Janeiro

2 439 Sagat Abikeyeva F Beijing

3 599 Bektas Abubakirov M Atlanta

4 620 Kanat Kalizhanovich Abutalipov M Beijing

5 620 Kanat Kalizhanovich Abutalipov M London

6 1034 Viktoriya Adyeva F Salt Lake City

*> nrow(dataSample)*

[1] 1429

Simple Random Sampling

A simple random sample of size 70 is drawn from the athletes with replacement. The athletes that are included in the above random sample correspond to the rows of the dataset as shown below. The data of the selected sample and the frequency of athletes’ age are shown below.

*> s = srswr(70, nrow(dataSample))*

*> s [s !=0]*

[1] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

[63] 1 1 1 1 1 1 1

*> rows = (1:nrow(dataSample))[s!=0]*

*> rows = rep(rows, s[s !=0])*

*> rows*

[1] 27 38 51 102 105 125 128 145 146 152 174 192 212 238 262 275 290 295 334 348 368 421 437 449 459

[26] 459 461 487 496 537 557 565 576 657 690 723 734 737 755 783 787 819 836 844 862 864 918 942 949 980

[51] 982 991 995 1015 1029 1053 1055 1123 1230 1240 1257 1276 1289 1324 1345 1346 1356 1361 1368 1403

*> sample.1 = dataSample[rows, ]*

*> head(sample.1[c(2,3,4,13)])*

ID Name Sex City

27 2606 Rezeda Aleyeva F Sydney

38 4527 Vladimir Yuryevich Antipin M Nagano

51 5650 Dina Aspandiyarova F Sydney

102 9765 Vladimir Belonogov M Sydney

105 9832 Sergey Nikolayevich Belyayev M Atlanta

125 12378 Olga Yevgenyevna Bludova-Safronova F Rio de Janeiro

*> table(sample.1$Age)*

18 19 20 21 22 23 24 25 26 27 28 29 30 32 33 36 37 38 43

3 2 2 3 2 9 6 7 3 6 9 3 3 2 3 2 2 2 1

Systematic sampling

For a sample of size 70, the data is divided into 21 groups. From the first group, a random item is selected. The rows of the systematic sample are now computed by taking every 21st item. The selected sample is indexed from these rows. The frequency of athletes’ age is shown below.

*> N = nrow(dataSample)*

*> N*

[1] 1429

*> n = 70*

*> k = ceiling(N/n)*

*> k*

[1] 21

*> r = sample(k,1)*

*> r*

[1] 5

*> s = seq(r, by = k , length = n)*

*> s*

[1] 5 26 47 68 89 110 131 152 173 194 215 236 257 278 299 320 341 362 383 404 425 446 467 488 509

[26] 530 551 572 593 614 635 656 677 698 719 740 761 782 803 824 845 866 887 908 929 950 971 992 1013 1034

[51] 1055 1076 1097 1118 1139 1160 1181 1202 1223 1244 1265 1286 1307 1328 1349 1370 1391 1412 1433 1454

*> sample.2 = dataSample[s, ]*

*> head(sample.2[c(2,3,4,13)])*

ID Name Sex City

5 620 Kanat Kalizhanovich Abutalipov M London

26 2514 Lyubov Alekseyeva F Salt Lake City

47 5442 Sergey Arzamasov M Sydney

68 6447 Dmitry Aleksandrovich Babenko M Torino

89 8885 Yermek Satybaldyyevich Bayduashov M London

110 9834 Galina Vasilyevna Belyayeva F Athina

*> table(sample.2$Age)*

17 19 20 21 22 23 24 25 26 27 28 29 30 31 32 35 37 52

1 4 3 7 7 8 9 2 3 3 4 5 4 2 2 2 1 1

Stratified Sampling with Two Variables

Each season is stratified by the gender of the athletes. The following data frame shows the data for the two seasons organized first by season and then by the gender (female and male) within each season. The contingency table showing the tabulation of the data by season and gender is shown below. For a desired sample size of 70, the number of athletes from each stratum (season and gender) were calculated proportional to their frequencies.

The extracted data shows the stratified sample organized by season and then by gender within each season.

*> data = data.frame(*

*+ Season = kazAthlete$Season,*

*+ Gender = kazAthlete$Sex)*

*> head(data)*

Season Gender

1 Summer M

2 Summer F

3 Summer M

4 Summer M

5 Summer M

6 Winter F

*> data = data[order(data$Season, data$Gender), ]*

*> freq <- table(data$Season, data$Gender)*

*> freq*

F M

Summer 314 517

Winter 235 363

*> st.sizes <- 70 \* freq / sum(freq)*

*> st.sizes = as.vector(t(st.sizes))*

*> st.sizes = st.sizes[st.sizes != 0]*

*> st.sizes*

[1] 15.38139 25.32540 11.51155 17.78167

*> st.1 <- strata(data, stratanames = c("Season", "Gender"),*

*+ size = st.sizes, method = "srswor")*

*> sample.3 <- getdata(data, st.1)*

*> head(sample.3)*

Season Gender ID\_unit Prob Stratum

135 Summer F 42 0.0489853 1

195 Summer F 48 0.0489853 1

385 Summer F 87 0.0489853 1

734 Summer F 168 0.0489853 1

813 Summer F 182 0.0489853 1

830 Summer F 190 0.0489853 1

If these 3 various probability sampling techniques were used instead of the whole dataset we can make a conclusion that most athletes ages are between 22 and 28. Also, the number of male athletes is greater than the number of female athletes.

**Implementation of any feature(s) not mentioned in the specification**

The library tidyverse were used to convert kazAthlete dataframe into athletes\_kaz tibble. A tibble is more suitable and easier to handle than the data frame. By default, the tibble displays the first 10 rows of data and leaves a comment on how many more rows are there. As shown there are 1,429 rows and 16 columns.

By using the summarize function, the maximum age recorded for each sport category were calculated and the plot of these values is shown below.

*> library(tidyverse)*

*> athletes\_kaz = as\_tibble(kazAthlete)*

*> athletes\_kaz*

# A tibble: 1,429 x 16

X ID Name Sex Age Height Weight Team NOC Games Year Season City Sport Event Medal

<int> <int> <fct> <fct> <int> <int> <dbl> <fct> <fct> <fct> <int> <fct> <fct> <fct> <fct> <fct>

1 493 272 Berik Abdrakhman… M 30 165 60 Kazakh… KAZ 2016 S… 2016 Summer Rio de … Boxing Boxing Men'… NA

2 767 439 Sagat Abikeyeva F 27 160 78 Kazakh… KAZ 2008 S… 2008 Summer Beijing Judo Judo Women'… NA

3 1069 599 Bektas Abubakirov M 23 165 54 Kazakh… KAZ 1996 S… 1996 Summer Atlanta Boxing Boxing Men'… NA

4 1108 620 Kanat Kalizhanov… M 25 168 56 Kazakh… KAZ 2008 S… 2008 Summer Beijing Boxing Boxing Men'… NA

5 1109 620 Kanat Kalizhanov… M 29 168 56 Kazakh… KAZ 2012 S… 2012 Summer London Boxing Boxing Men'… NA

6 1886 1034 Viktoriya Adyeva F 20 160 55 Kazakh… KAZ 2002 W… 2002 Winter Salt La… Ice H… Ice Hockey … NA

7 1888 1036 Olga Olegovna Ad… F 26 192 65 Kazakh… KAZ 2008 S… 2008 Summer Beijing Handb… Handball Wo… NA

8 2763 1539 Marina Aleksandr… F 21 177 58 Kazakh… KAZ 2004 S… 2004 Summer Athina Athle… Athletics W… NA

9 2764 1539 Marina Aleksandr… F 25 177 58 Kazakh… KAZ 2008 S… 2008 Summer Beijing Athle… Athletics W… NA

10 2949 1642 Akhat Akhirov M 26 NA NA Kazakh… KAZ 1996 S… 1996 Summer Atlanta Judo Judo Men's … NA

# ... with 1,419 more rows

*> #showing the maximum age recorded for each sport category*

*> maxAgeSport = athletes\_kaz %>%*

*+ group\_by(Sport) %>%*

*+ summarise(maxAge = max(Age))*

*> maxAgeSport*

# A tibble: 36 x 2

Sport maxAge

<fct> <dbl>

1 Alpine Skiing 27

2 Archery 42

3 Athletics 40

4 Biathlon 33

5 Boxing 32

6 Canoeing 40

7 Cross Country Skiing 38

8 Cycling 38

9 Diving 26

10 Fencing 32

# ... with 26 more rows

*> # showing the plot for these values*

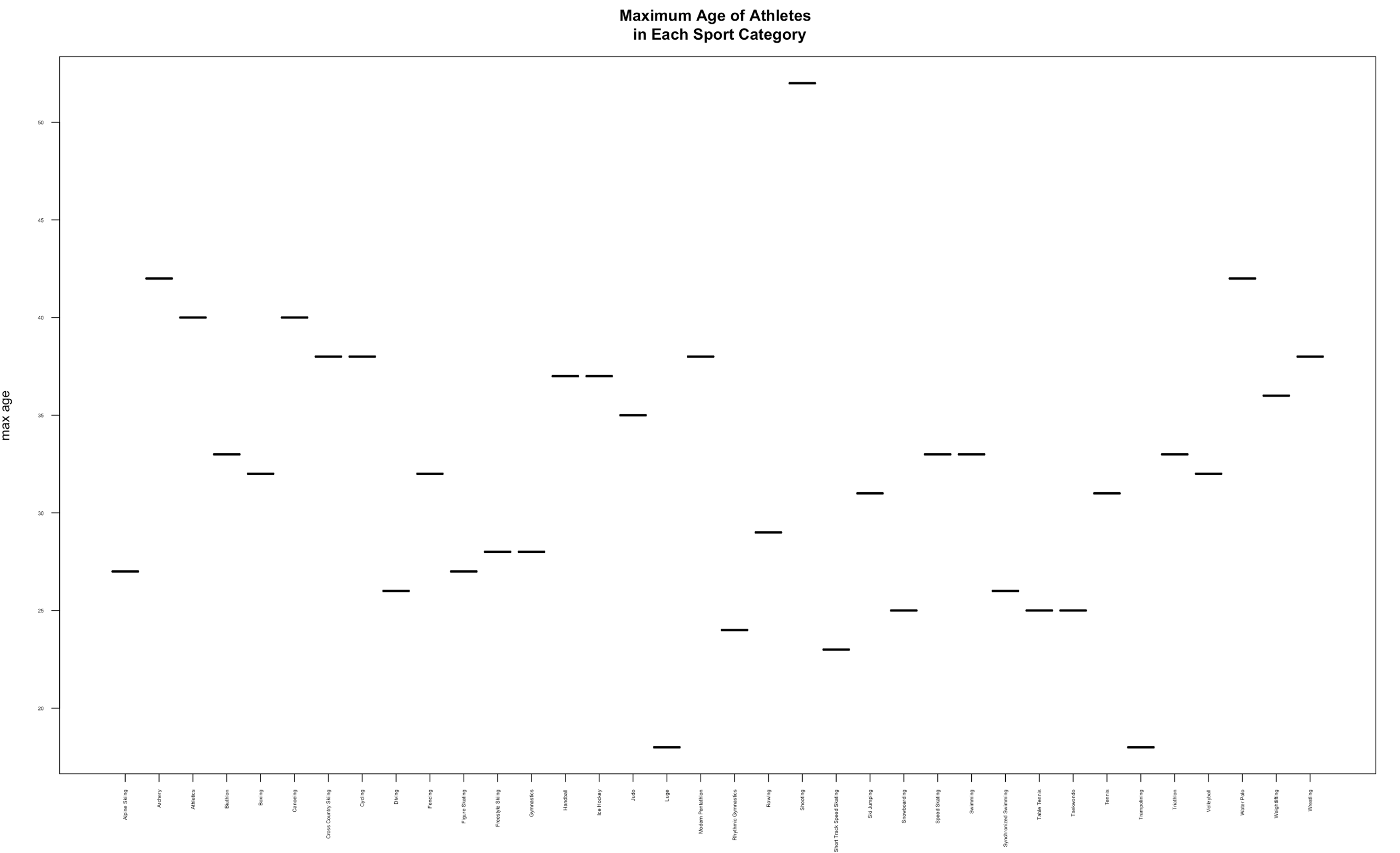
*> plot(maxAgeSport$Sport, maxAgeSport$maxAge,*

*+ ylab = "max age", pch = 16,*

*+ las = 2, cex.axis = 0.4,*

*+ main = "Maximum Age of Athletes \n in Each Sport Category")*

*>*

**

The filter function selects the rows from the dataset that match the specified conditions involving the variables in the data. The following example filters all rows for the season of Summer. By using the summarize function, the average athletes' heights of Summer Season were calculated and the plot of these values is shown below.

*> athletesSummer = filter(athletes\_kaz, Season == "Summer")*

*> athletesSummer*

# A tibble: 831 x 16

X ID Name Sex Age Height Weight Team NOC Games Year Season City Sport Event Medal

<int> <int> <fct> <fct> <int> <int> <dbl> <fct> <fct> <fct> <int> <fct> <fct> <fct> <fct> <fct>

1 493 272 Berik Abdrakhman… M 30 165 60 Kazakh… KAZ 2016 S… 2016 Summer Rio de … Boxing Boxing Men'… NA

2 767 439 Sagat Abikeyeva F 27 160 78 Kazakh… KAZ 2008 S… 2008 Summer Beijing Judo Judo Women'… NA

3 1069 599 Bektas Abubakirov M 23 165 54 Kazakh… KAZ 1996 S… 1996 Summer Atlanta Boxing Boxing Men'… NA

4 1108 620 Kanat Kalizhanov… M 25 168 56 Kazakh… KAZ 2008 S… 2008 Summer Beijing Boxing Boxing Men'… NA

5 1109 620 Kanat Kalizhanov… M 29 168 56 Kazakh… KAZ 2012 S… 2012 Summer London Boxing Boxing Men'… NA

6 1888 1036 Olga Olegovna Ad… F 26 192 65 Kazakh… KAZ 2008 S… 2008 Summer Beijing Handb… Handball Wo… NA

7 2763 1539 Marina Aleksandr… F 21 177 58 Kazakh… KAZ 2004 S… 2004 Summer Athina Athle… Athletics W… NA

8 2764 1539 Marina Aleksandr… F 25 177 58 Kazakh… KAZ 2008 S… 2008 Summer Beijing Athle… Athletics W… NA

9 2949 1642 Akhat Akhirov M 26 NA NA Kazakh… KAZ 1996 S… 1996 Summer Atlanta Judo Judo Men's … NA

10 2950 1643 Sergey Akhirov M 31 NA NA Kazakh… KAZ 1996 S… 1996 Summer Atlanta Judo Judo Men's … NA

# ... with 821 more rows

*> avgHeightSummer = athletesSummer %>%*

*+ filter(Height > 0) %>%*

*+ group\_by(Age) %>%*

*+ summarise(avgHeight = mean(Height))*

*> avgHeightSummer*

# A tibble: 33 x 2

Age avgHeight

<int> <dbl>

1 13 180

2 14 161

3 15 152

4 16 162

5 17 179.

6 18 176

7 19 171.

8 20 178.

9 21 178.

10 22 173.

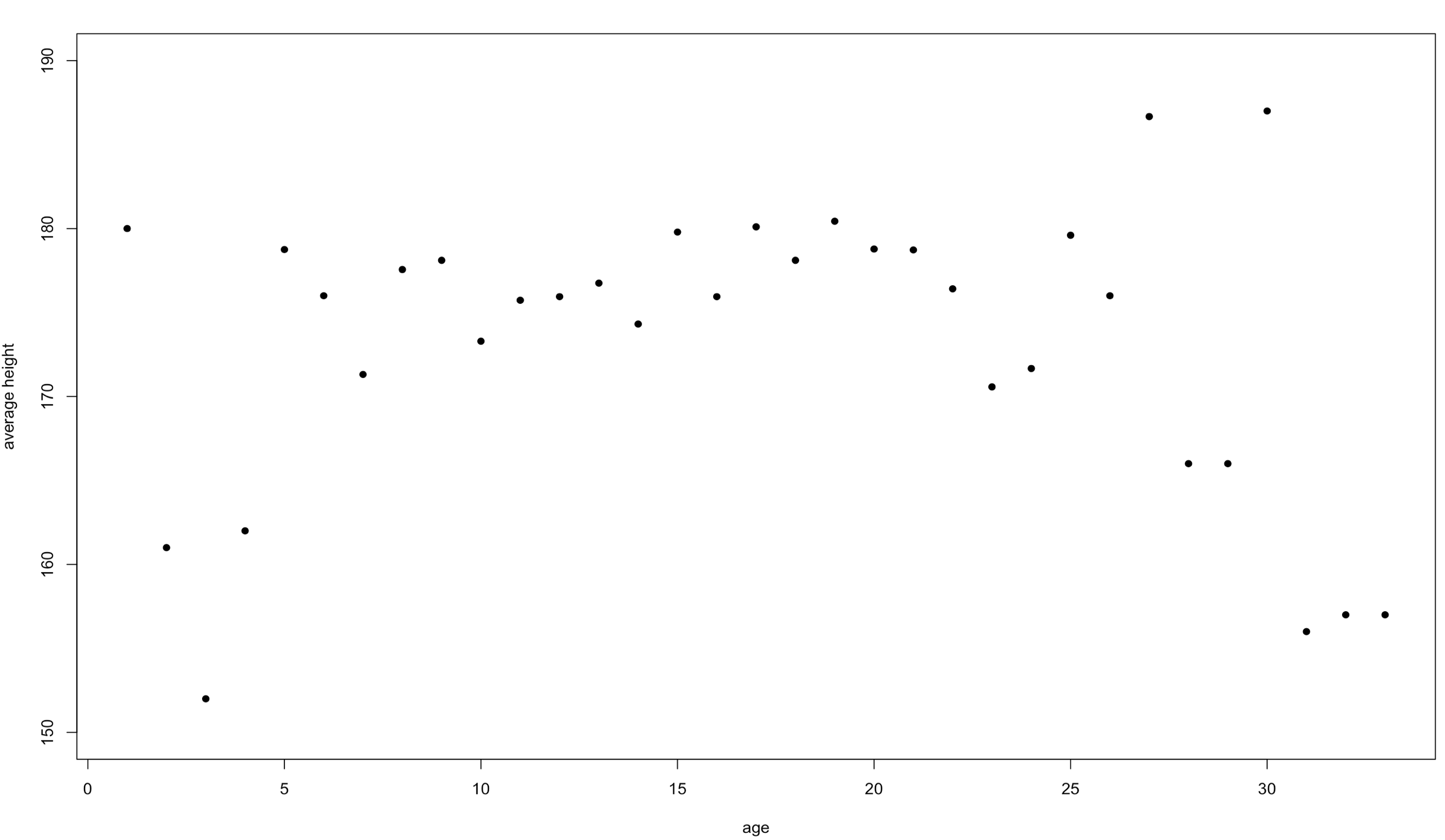
# ... with 23 more rows

*> # showing the plot for these values*

*> plot(avgHeightSummer$avgHeight, ylim = c(150, 190),*

*+ xlab = "age", ylab = "average height", pch = 16)*

*>*

**

# **References:**

1) “120 years of Olympic history: athletes and results” dataset, <https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results#athlete_events.csv>.

2) Printable Lectures for MET CS544 A1 Foundation of Analytics