Olympic Porjct - GSBA Final Project

Introduction

This report will analyze the Olympic data from 2000 to 2012, and only the participants – who earned metal(s) – are included in this dataset. The aggregation data only includes 2012 Olympic Games (aggreatedinfo.csv), because indexes of Human Development Index (HDI), Population, and Gross Domestic Product (GDP) are approximately from year 2012 or 2013.

All the data are available here:https://github.com/DylanYileWu/gsba-final.git

Clarification: Given the limited time and dataset, I aggregated the year 2012 data by the countries of the athletes (combinedata.csv). If more time are given, data like yearly HDI indexes for each country, yearly populations for each country, yearly GDP for each country can be used to better study the relation among HDI indexes, GDP, population,and athletes who won the trophies, aggregation will base on the countries of the athletes and the year of the Olympic Games.

Data columns definition(combinedata.csv):

competing_country: Country name

competing_country_code: Country code

sports_involve: How many sports this country participant

mean_age: The average age for all the participants from this country

totalgold: Total earned gold metals for this country

totalsilver: Total earned silver medals for this country

totalbronze: Total earned bronze medals for this country

totalmedal: Total medals for this country(including gold, solver, and bronze)

top_perf_sport: The best performance sport for this country

top_perf_sport_medal: The corresponding number of medals for the top performance sport

hdi: Human development index

life_expectancy_at_birth:Life expectancy

mean_years_of_schooling: Average years in school

expected_years_of_schoolings: Expeced years in school

gross_national_income_per_captia: GNI per captia

hdiranking: Ranking by HDI, the lower the better

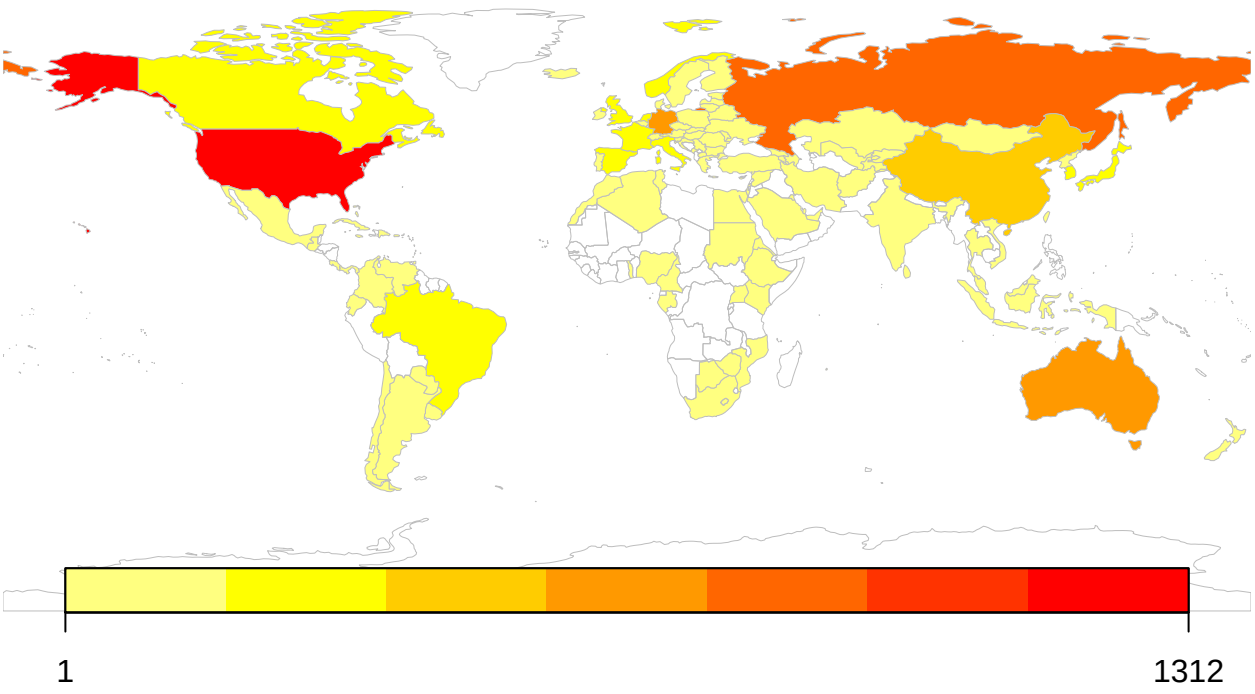
population: Total population

GDP: GDP for this country

Data Analysis - Original Data

Medals earned by each country

```
library(rworldmap)
total_madel=0
total_madel=tapply(origindata$Tota_Medals,origindata$Competing.Country, FUN=sum)
#make the data as table
total_madel=as.table(total_madel)
total_madel=data.frame(total_madel)
#assign column names
colnames(total_madel)=c("Country","Total_medal")
#create a map-shaped window
mapDevice('x11')
#create the map
capture.output(spdf <- joinCountryData2Map(total_madel, joinCode="NAME",
              nameJoinColumn="Country"), file='NUL')
mapCountryData(spdf, nameColumnToPlot="Total_medal", catMethod="fixedWidth",
               mapTitle="Total medal(s) for each participant country")
```

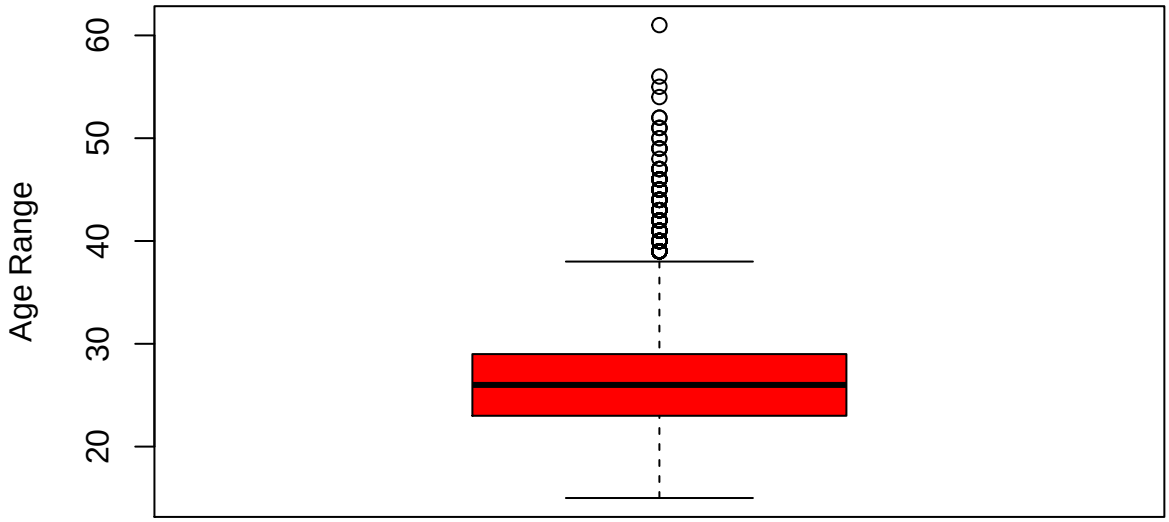


Athlete Age

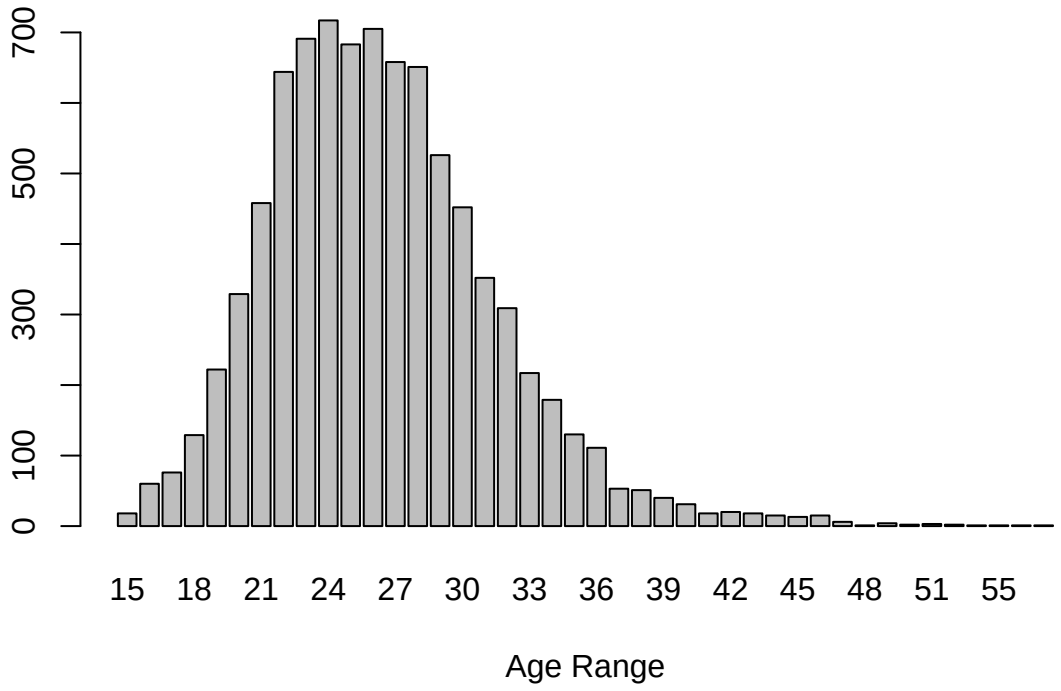
```
#Summary of age
summary(origindata$Age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   15.00   23.00   26.00   26.41   29.00   61.00       5
```

```
#The boxplot chart
boxplot(origindata$Age,col=c("red"),ylab="Age Range")
```




```
#Age freq bar chart
agecount.freq=table(origindata$Age)
barplot(agecount.freq,xlab="Age Range")
```

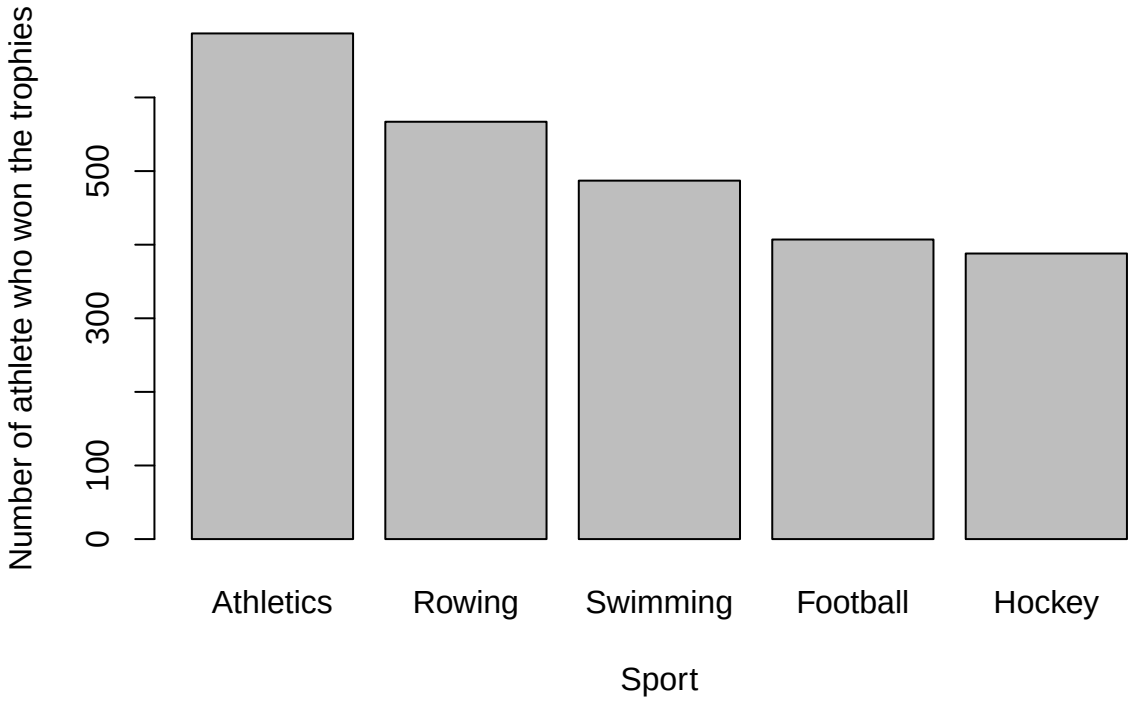


Summary-Athlete Age

The winning athletes' ages ranging from 15 to 61, with the mean age 26. Nearly half of the athletes age from 23 to 29, which represent the middle fifty percent of all the participants. Overall, the participants age is normally distributed, and there are few old athletes as well - right skewed.

Sports

```
sport.freq=data.frame(origindata$Sport)
#sort the sport frequency in decreasing order
sport.freq=sort(table(origindata$Sport),decreasing=T)
#show the top 5 popular sports
barplot(sport.freq[1:5],xlab="Sport",ylab="Number of athlete who won the trophies")
```



Summary-Sports

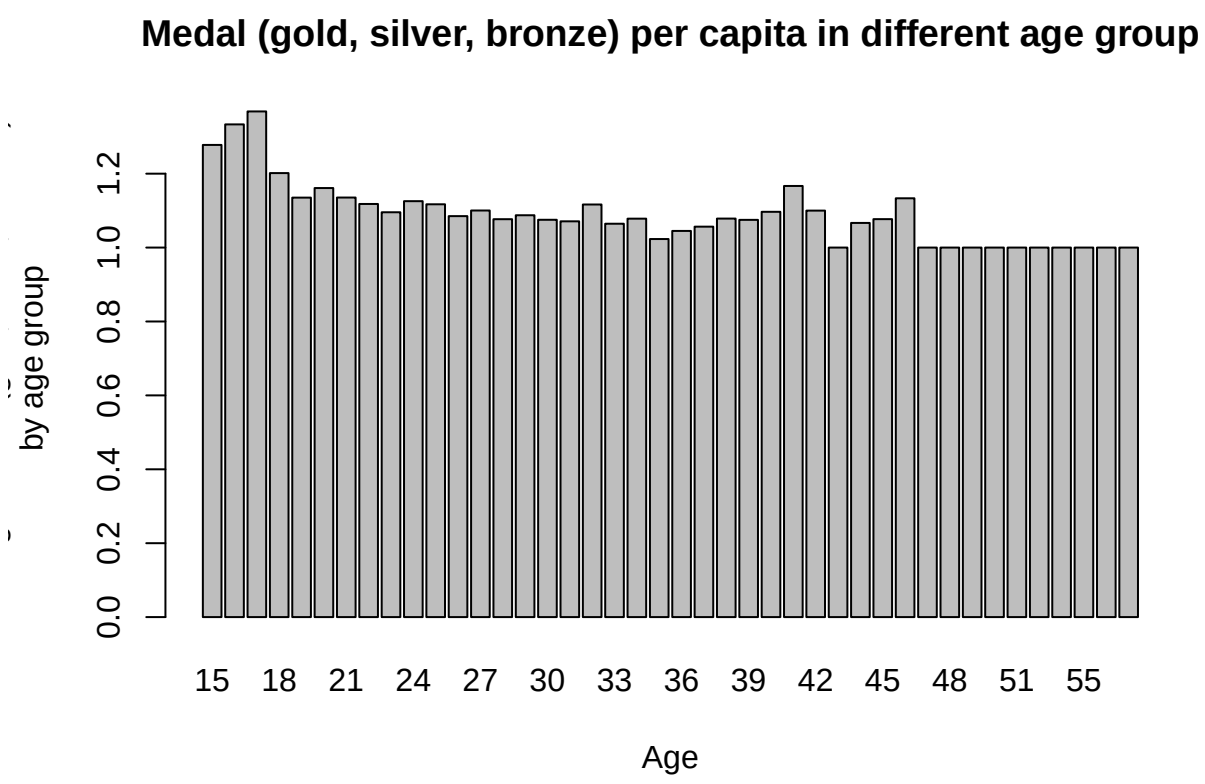
The top five most-attended sports are Athletics, Rowing, Swimming, Football, and Hockey.

```
sport.freq[1:5]
```

```
##
## Athletics    Rowing  Swimming  Football    Hockey
##       687       567       487       407       388
```

Does Age affect athlete to obtain medals

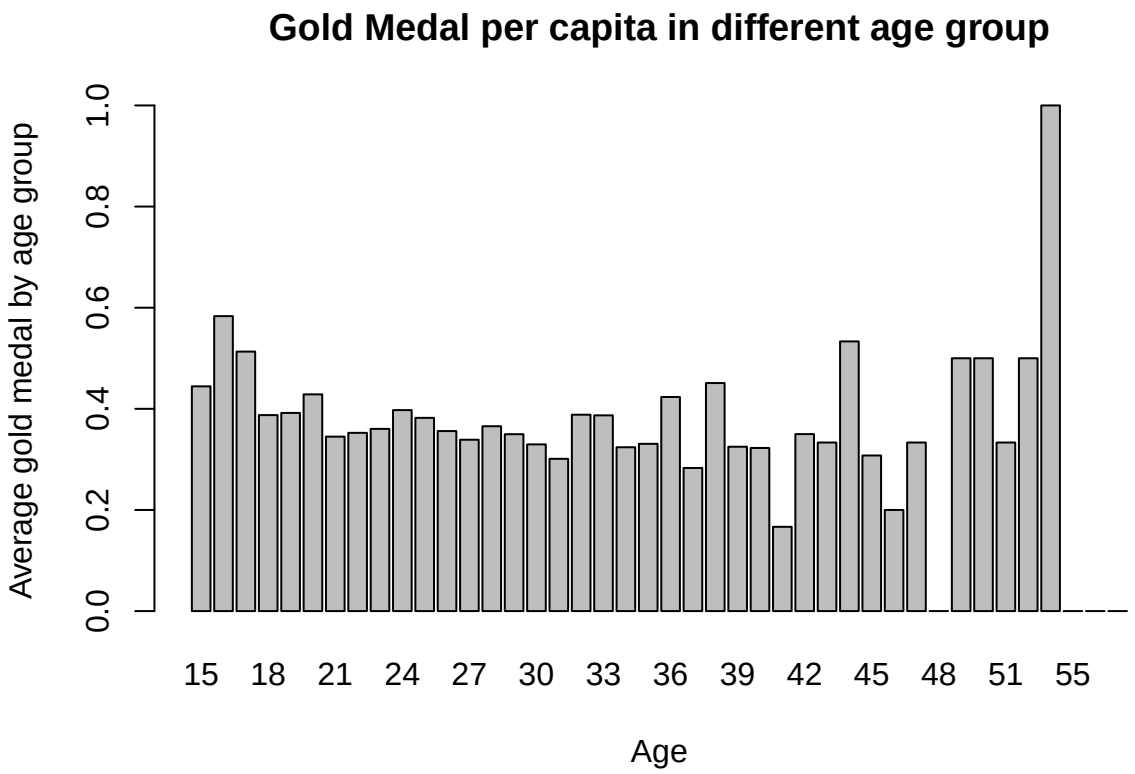
```
#number of medals by age
num_medals=tapply(origindata$Tota_Medals, origindata$Age, FUN=sum)
#number of gold medals by age
num_gold_medals=tapply(origindata$Gold_Medals, origindata$Age, FUN=sum)
#number of athelete by age is calculated earlier (agecount.freq)
medal_per_person_by_age=num_medals/agecount.freq
gold_per_person_by_age=num_gold_medals/agecount.freq
#Access the values
barplot(medal_per_person_by_age,xlab="Age",ylab="Average medal (gold, silver, bronze)
  by age group",main="Medal (gold, silver, bronze) per capita in different age group")
```




```
medal_per_person_by_age
```

```
##       15       16       17       18       19       20       21       22
## 1.277778 1.333333 1.368421 1.201550 1.135135 1.161094 1.135371 1.118012
##       23       24       25       26       27       28       29       30
## 1.095514 1.125523 1.117130 1.085106 1.100304 1.076805 1.087452 1.075221
##       31       32       33       34       35       36       37       38
## 1.071023 1.116505 1.064516 1.078212 1.023077 1.045045 1.056604 1.078431
##       39       40       41       42       43       44       45       46
## 1.075000 1.096774 1.166667 1.100000 1.000000 1.066667 1.076923 1.133333
##       47       48       49       50       51       52       54       55
## 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000 1.000000
##       56       61
## 1.000000 1.000000
```

```
barplot(gold_per_person_by_age,xlab="Age",ylab="Average gold medal by age group",
        main="Gold Medal per capita in different age group")
```




```
gold_per_person_by_age
```

```
##        15        16        17        18        19        20        21
## 0.4444444 0.5833333 0.5131579 0.3875969 0.3918919 0.4285714 0.3449782
##        22        23        24        25        26        27        28
## 0.3524845 0.3603473 0.3974895 0.3821376 0.3560284 0.3389058 0.3655914
##        29        30        31        32        33        34        35
## 0.3498099 0.3296460 0.3011364 0.3883495 0.3870968 0.3240223 0.3307692
##        36        37        38        39        40        41        42
## 0.4234234 0.2830189 0.4509804 0.3250000 0.3225806 0.1666667 0.3500000
##        43        44        45        46        47        48        49
## 0.3333333 0.5333333 0.3076923 0.2000000 0.3333333 0.0000000 0.5000000
##        50        51        52        54        55        56        61
## 0.5000000 0.3333333 0.5000000 1.0000000 0.0000000 0.0000000 0.0000000
```

Summary - Does Age affect athlete to obtain medals

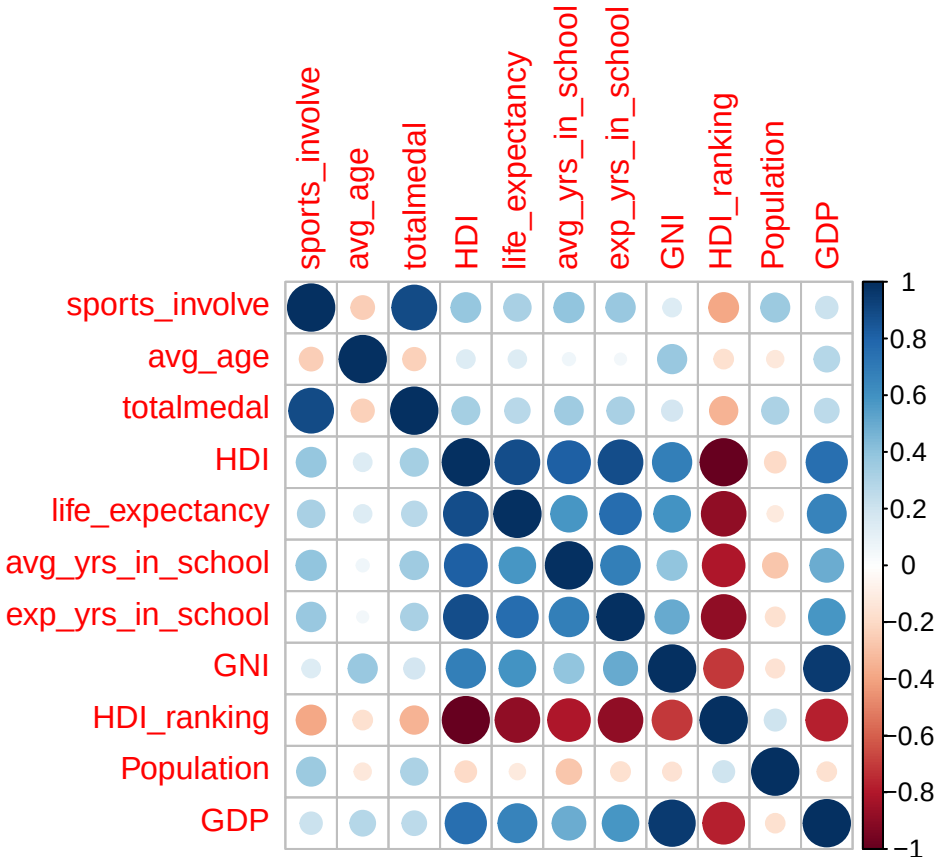
Age does affect the overall performance in each age group. From the "Medal (gold, silver, bronze) per capita in different age group" chart, Young athletes (from age 15 to 17) achieve higher medals per capita than the rest age groups. However, Age has less effects on athlete obtaining gold medals, young athlete still shows higher achievement. Moreover, Athlete from age 49 to 54 shows high rate of getting gold medals. Such phenomenon can be caused by the small amount of old participants, and it may also mean that those old athletes has long time practice in the sports field they compete.

Data Analysis - Combine Data

Correlation among data

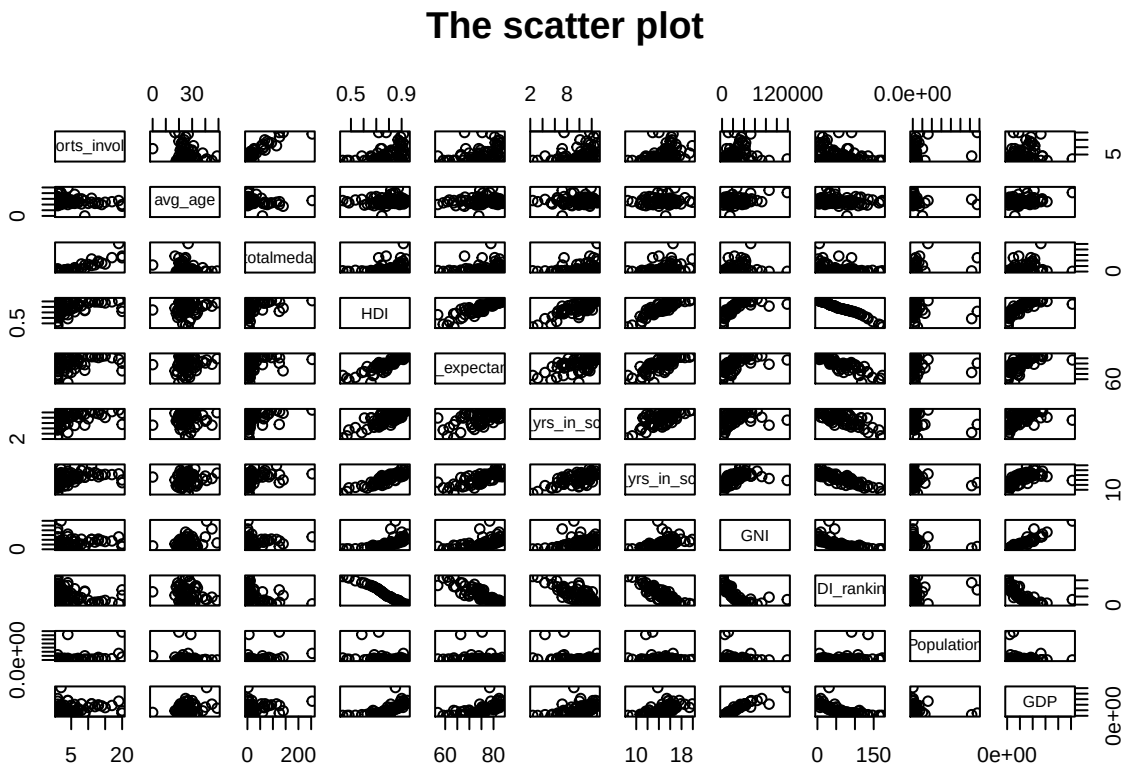
visualized correlation table

```
attach(combinedata)
d <- data.frame(sports_involve=sports_involve,avg_age=mean_age,totalmedal=totalmedal,
                HDI=hdi,life_expectancy=life_expectancy_at_birth,
                avg_yrs_in_school=mean_years_of_schooling
                ,exp_yrs_in_school=expected_years_of_schoolings,
                GNI=gross_national_income_per_captia,
                HDI_ranking=hdiranking,Population=population,GDP=GDP)
# delete all the missing value
d=na.omit(d)
#create the correlation table
dtable=cor(d)
# visualize the correlation table
library('corrplot') #package corrplot
corrplot(dtable, method = "circle")
```



scatter plot

```
#the scatter plot
pairs(d, data=mtcars,main="The scatter plot")
```



Summary-Correlation among data

From the correlation table, It is clear that all the indexes from the Human Development Index (HDI.csv) table have strong positive relationship with each other, and the high negative relationship between HDI ranking and rest indexes, which is expected since higher HDI will be ranked as smaller number. Moreover, those HDI indexes show strong relationship with GDP, which is also expected since high GDP will contribute to human living quality. The population overall has negative correlation with most indexes from HDI table and GDP, which means high population becomes a burden to a high population country. Overall, most the indexes have either weak or median positive relation with number of sports (sports_involve) that a country participates, which further influence the medals (totalmedal) that a country can get in the end.

Top performance sport for each country

```
#the scatter plot
dt=data.frame(combinedata)
cols <- c(1, 8:10)
dt[, cols]
```

```
##      competing_country totalmedal   top_perf_sport top_perf_sport_medal
## 1                Qatar          2        Athletics                    1
## 2            Singapore          4     Table Tennis                    4
## 3               Norway         17         Handball                   14
## 4          Switzerland          4          Cycling                    1
## 5        United States        254         Swimming                   68
## 6            Hong Kong          1          Cycling                    1
## 7          Netherlands         69           Hockey                   33
## 8               Canada         55         Football                   18
## 9            Australia        114         Swimming                   32
## 10              Kuwait          1         Shooting                    1
## 11             Ireland          5           Boxing                    4
## 12              Sweden         22         Handball                   14
## 13      Chinese Taipei          2        Taekwondo                    1
## 14             Germany         94           Rowing                   17
## 15             Belgium          3             Judo                    1
## 16             Denmark         16           Rowing                    7
## 17       Great Britain        126           Rowing                   28
## 18               Japan         84         Football                   18
## 19             Finland          5          Sailing                    4
## 20              France         78         Swimming                   21
## 21         South Korea         61         Football                   18
## 22             Bahamas          4        Athletics                    4
## 23        Saudi Arabia          4       Equestrian                    4
## 24         New Zealand         27           Rowing                    9
## 25               Spain         64         Handball                   15
## 26             Bahrain          1        Athletics                    1
## 27               Italy         68          Fencing                   16
## 28            Slovenia          5           Rowing                    2
## 29      Czech Republic         14         Canoeing                    5
## 30            Slovakia          5         Canoeing                    3
## 31              Cyprus          1          Sailing                    1
## 32              Greece          3           Rowing                    2
## 33            Portugal          2         Canoeing                    2
## 34           Lithuania          5           Boxing                    1
## 35             Estonia          2        Athletics                    1
## 36              Poland         12        Athletics                    2
## 37 Trinidad and Tobago         10        Athletics                   10
## 38             Hungary         25         Canoeing                   14
## 39               Gabon          1        Taekwondo                    1
## 40              Latvia          3 Beach Volleyball                    2
## 41           Argentina         21           Hockey                   17
## 42              Russia        140        Athletics                   23
## 43             Croatia         35         Handball                   15
## 44            Malaysia          2        Badminton                    1
## 45            Botswana          1        Athletics                    1
## 46         Puerto Rico          2        Athletics                    1
## 47             Belarus         23         Canoeing                    8
## 48              Mexico         24         Football                   16
## 49              Turkey          5        Athletics                    2
## 50            Bulgaria          2           Boxing                    1
## 51             Romania         16       Gymnastics                    7
## 52          Kazakhstan         13           Boxing                    4
## 53             Grenada          1        Athletics                    1
```

```
## 54           Venezuela          1          Fencing                 1
## 55                Iran         12        Wrestling                 6
## 56              Brazil         59       Volleyball                24
## 57        South Africa          9           Rowing                 4
## 58            Colombia          8          Cycling                 3
## 59              Serbia         16       Waterpolo                13
## 60          Azerbaijan         10        Wrestling                 7
## 61                Cuba         14           Boxing                 4
## 62            Thailand          3           Boxing                 1
## 63             Tunisia          3         Swimming                 2
## 64               China        125         Swimming                15
## 65  Dominican Republic          2        Athletics                 2
## 66             Jamaica         25        Athletics                25
## 67             Algeria          1        Athletics                 1
## 68             Ukraine         26        Athletics                 6
## 69               Egypt          2          Fencing                 1
## 70             Armenia          3        Wrestling                 2
## 71             Georgia          7        Wrestling                 6
## 72             Morocco          1        Athletics                 1
## 73           Guatemala          1        Athletics                 1
## 74           Indonesia          2    Weightlifting                 2
## 75               India          6         Shooting                 2
## 76            Mongolia          5           Boxing                 2
## 77             Moldova          2    Weightlifting                 2
## 78          Uzbekistan          4        Wrestling                 2
## 79          Tajikistan          1           Boxing                 1
## 80          Montenegro         14         Handball                14
## 81               Kenya         11        Athletics                11
## 82         North Korea          6    Weightlifting                 4
## 83              Uganda          1        Athletics                 1
## 84            Ethiopia          7        Athletics                 7
## 85         Afghanistan          1        Taekwondo                 1
```

Summary-Top performance sport

For many countries, the medals from top performance sport usually consisit most of the total medals earned for 2012 Oliypic Game