Final Project - GSBA 524

The objective of this exploratory analysis is to dig deeper into seven years worth of Olympic games winners data and to find meaningful insights

The Olympic games conducts two cycles of international sporting events in the Summer and the Winter. By segregating the data into these two cycles, we can identify the top ten sports and the the top ten nations with the most medals. We also analyzed the wins by continent to better understand which part of the world are most olympic medals going to.

```
olympics = read.csv('Olympics Data.csv')
olympics$Olympics[olympics$Year =="2000" | olympics$Year =="2004"
| olympics$Year == "2008" | olympics$Year == "2012"] = "Summer"
olympics$Olympics[olympics$Year =="2002" | olympics$Year =="2006"
| olympics$Year == "2010" | olympics$Year == "2014"] = "Winter"
Summer = subset(olympics, Olympics == "Summer")
Winter = subset(olympics, Olympics == "Winter")
Summer = na.omit(Summer)
summer_sports= aggregate(Total.Medals~ Sport,Summer, sum)
summer_sports = data.frame(Sport =summer_sports$Sport,
                Medals = summer_sports$Total.Medals)
summer_sports = summer_sports[order(-summer_sports$Medals),]
summer_sports[1:10,]
```

```
##         Sport Medals
## 24   Swimming    765
## 2   Athletics    753
## 20      Rowing    576
## 13   Football    407
## 16     Hockey    386
## 15   Handball    351
## 8    Canoeing    333
## 9     Cycling    306
## 32  Waterpolo    306
## 5  Basketball    287
```

```
winter_sports= aggregate(Total.Medals~ Sport,Winter, sum)
winter_sports = data.frame(Sport =winter_sports$Sport,
                           Medals = winter_sports$Total.Medals)
winter_sports = winter_sports[order(-winter_sports$Medals),]
winter_sports[1:10,]
```

```
##                         Sport Medals
## 8                  Ice Hockey    384
## 4        Cross Country Skiing    174
## 15              Speed Skating    140
## 2                    Biathlon    138
## 11 Short-Track Speed Skating    138
## 1               Alpine Skiing     90
## 5                     Curling     82
## 3                   Bobsleigh     72
## 6              Figure Skating     54
## 10            Nordic Combined     54
```

```
summer_winners = aggregate(Total.Medals ~ Competing.Country, Summer, sum)
summer_winners = data.frame(Country = summer_winners$Competing.Country,
                            Medals = summer_winners$Total.Medals)
summer_winners = summer_winners[order(-summer_winners$Medals),]
summer_winners[1:10,]
```

```
##           Country Medals
## 105 United States   1079
## 81         Russia    664
## 5       Australia    602
## 19          China    482
## 38        Germany    460
## 39  Great Britain    314
## 69    Netherlands    286
## 51          Italy    280
## 35         France    274
## 53          Japan    272
```

```
winter_winners = aggregate(Total.Medals ~ Competing.Country, Winter, sum)
winter_winners = data.frame(Country = winter_winners$Competing.Country,
                            Medals = winter_winners$Total.Medals)
winter_winners = winter_winners[order(-winter_winners$Medals),]
winter_winners[1:10,]
```

```
##          Country Medals
## 5         Canada    233
## 28 United States    233
## 12       Germany    169
## 25        Sweden    108
## 21        Russia    104
## 19        Norway    103
## 10       Finland    101
## 2        Austria     76
## 26   Switzerland     57
## 14         Italy     51
```

```
continents = read.csv('Continents.csv')
colnames(continents)[1] = "C"
colnames(olympics)[3] ="C"
olympics =merge(olympics,continents)
cont = aggregate(Total.Medals ~ Continent,olympics,sum)
cont
```

```
##       Continent Total.Medals
## 1        Africa          198
## 2          Asia         1377
## 3        Europe         4829
## 4 North America         2042
## 5       Oceania          661
## 6 South America          422
```

```
table(olympics$Continent, olympics$Olympics)
```

```
##
##                  Summer Winter
##   Africa            182      0
##   Asia             1162     76
##   Europe           3627    781
##   North America    1369    427
##   Oceania           568      7
##   South America     419      0
```

We observe that USA, Russia and Germany dominate in both, the Summer and Winter olympics. Additionally, Australia is a high performer in the Summer Olympics and Canada in the winter olympics.Among the continents, Europe has the most medals, A large contributer to this count being the winter olympic games where the sporting events are more suited for the climate of European and North American countries than the more tropical countries in Asia

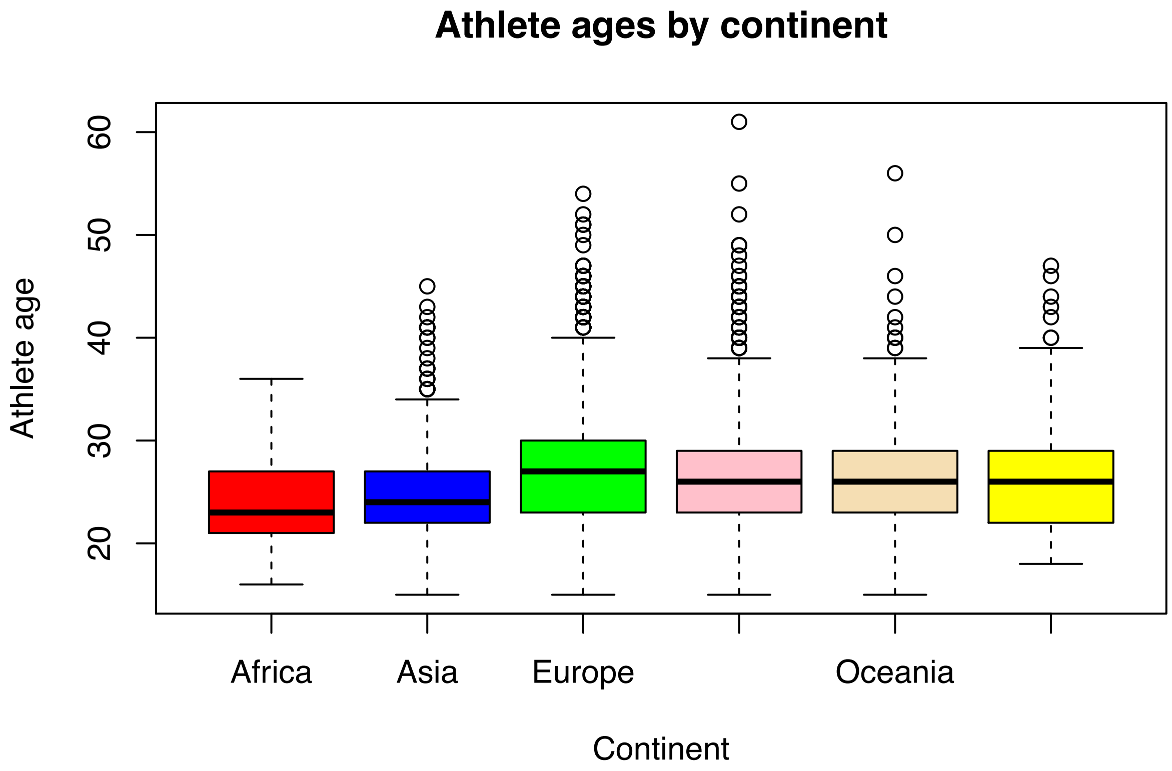
Next, we'll try to analyze the performance of the athletes with respect to ages and find answers to some questions like Do younger athletes perform better? Do olympics winner ages vary across continents? Which age group has multiple medal winners?

NOTE : Since the purpose of this analysis is to identify the ages at which wins occur, we are not concerned about athlete ages as they participate in the games over the course of several years and their individual wins. For eg. Michael Phelp's wins at ages 19 and 27 are treated as separate entities.

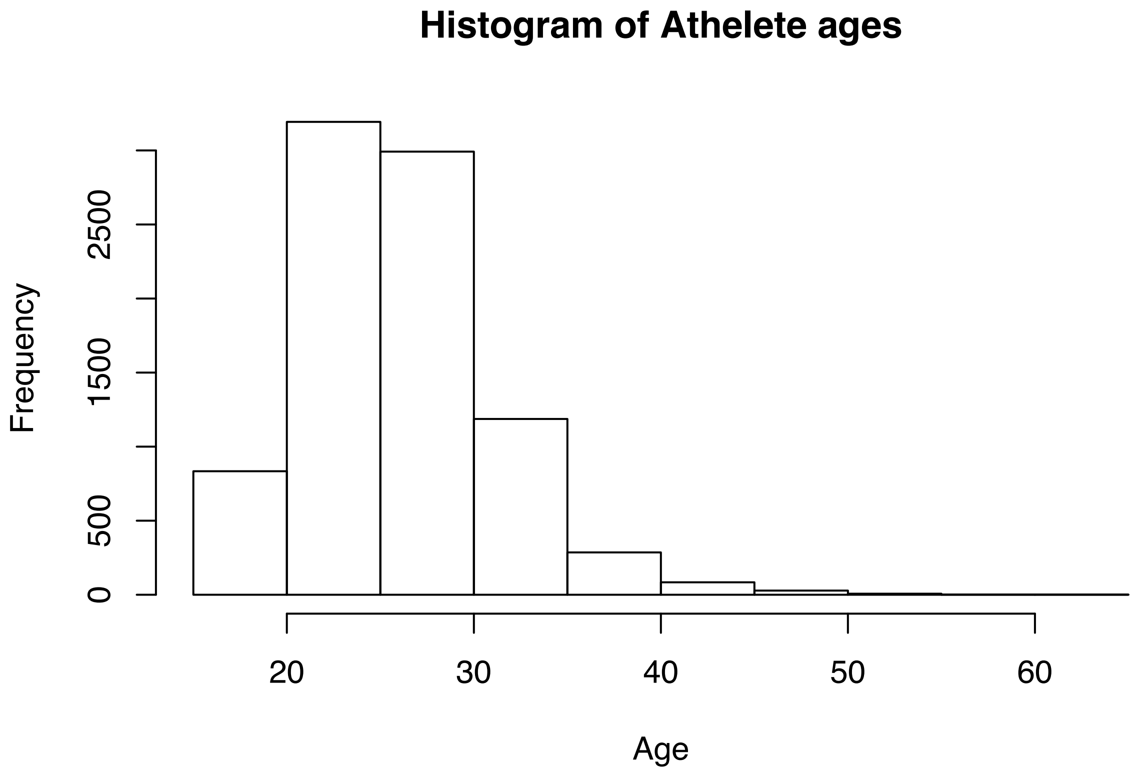
```
olympics$agegroup[olympics$Age <= 24 ] = "Young"
olympics$agegroup[olympics$Age > 24 & olympics$Age <= 30] = "Average"
olympics$agegroup[olympics$Age > 30] ="Old"




boxplot(Age~Continent,data=olympics,
        main = "Athlete ages by continent",xlab ="Continent",ylab = "Athlete age",
        col = c("red","blue","green","pink","wheat","yellow"))
```

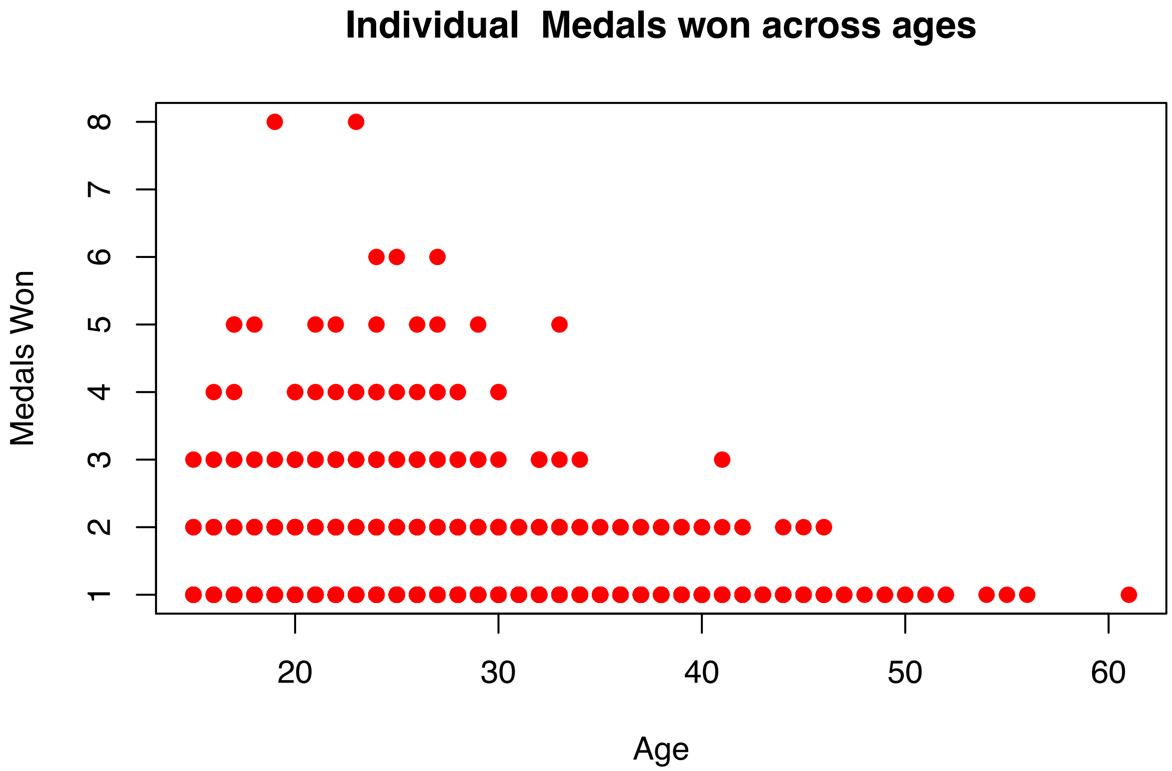


```r
hist(olympics$Age, xlab = "Age", main = "Histogram of Athelete ages")
```



```
hdi = read.csv("HDI.csv")
colnames(hdi)[2] = "CC"
colnames(olympics)[4] = "CC"
olympics = merge(olympics,hdi)
#str(olympics)
colnames(olympics)[23] = "GNI"
olympics$gn1 = as.character(olympics$GNI)
olympics$gn1 = as.numeric(olympics$gn1)
#attach(olympics)
#str(olympics)

plot(olympics$Age, olympics$Total.Medals, col = "red", pch = 19,
     xlab = "Age", ylab= "Medals Won", main = "Individual  Medals won across ages")
```



```r
agewins = aggregate(Age~Sport, olympics, mean)
agewins = data.frame(Sport = agewins$Sport, Avg.age = agewins$Age)
agewins = agewins[order(-agewins$Avg.age),]
head(agewins,10)
```

```
##                   Sport  Avg.age
## 16           Equestrian 37.92357
## 13              Curling 32.85366
## 32             Shooting 31.03371
## 31              Sailing 29.95238
## 7      Beach Volleyball 29.85417
## 9             Bobsleigh 29.66667
## 34             Skeleton 29.61111
## 26                 Luge 28.83333
## 8              Biathlon 28.81915
## 12 Cross Country Skiing 28.39062
```

```r
tail(agewins, 10)
```

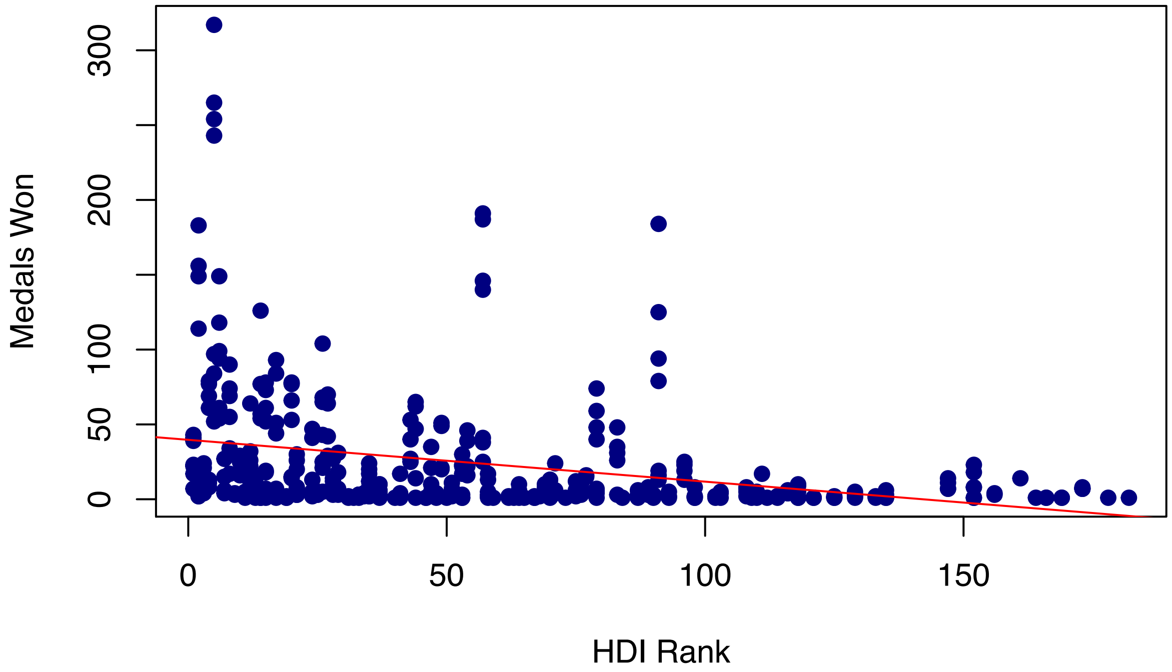
```
##                        Sport  Avg.age
## 10                    Boxing 24.24731
## 19                  Football 24.12285
## 44                Trampoline 23.91667
## 40     Synchronized Swimming 23.85321
## 42                 Taekwondo 23.75000
## 33 Short-Track Speed Skating 22.69792
## 39                  Swimming 22.61807
## 15                    Diving 22.53982
## 21                Gymnastics 21.35233
## 29       Rhythmic Gymnastics 18.90476
```

We observe that the Asian and African athletes tend to be younger when they win. In contrast Europe,North America and Australia have more athletes with olympic winners at age 40 or more. Also, Multiple medal winners are between ages 20 and 35. Sports like Equestrian and Curling have the oldest winners and swimming and gymnastics have youngest winners.

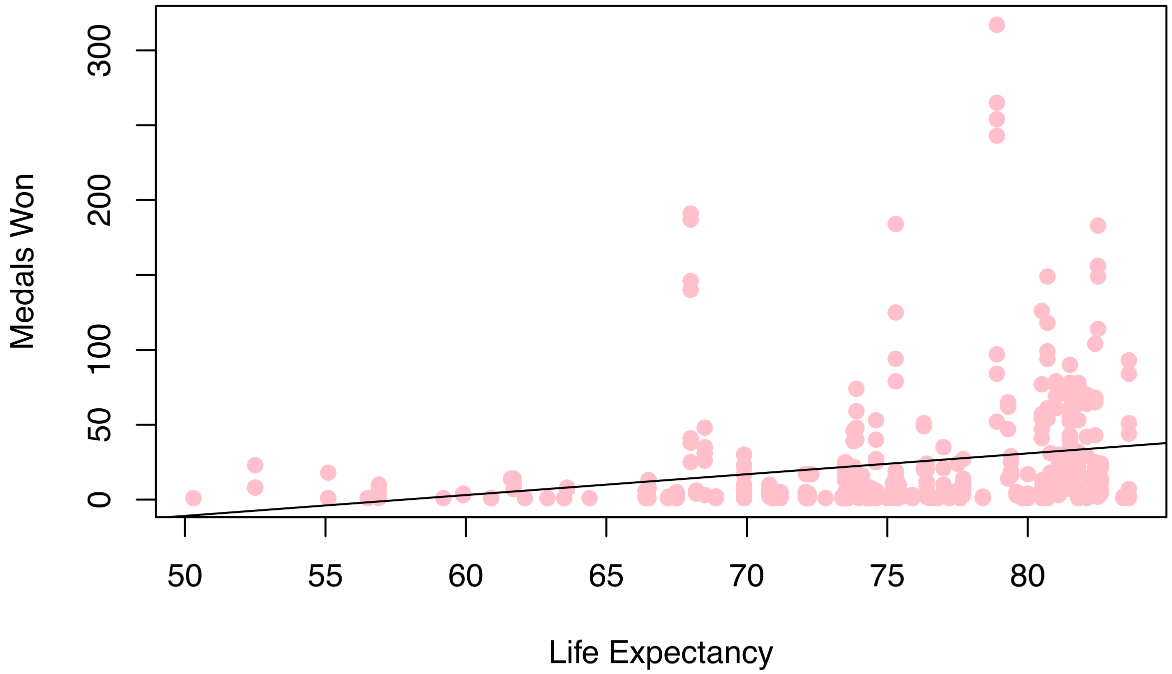
It is reasonable to assume that a nation's economy and human development conditions decide the training facilities and other resources for athletes. Let's try to investigate if there's any relation between each of these factors like schooling, human development and life expectancy affect the wins at the games.

Note : We are using data from all seven years and assuming the above factors for each country remain the same over the years. The idea is to understand given these conditions, how many medals can be won by a country at any given year.

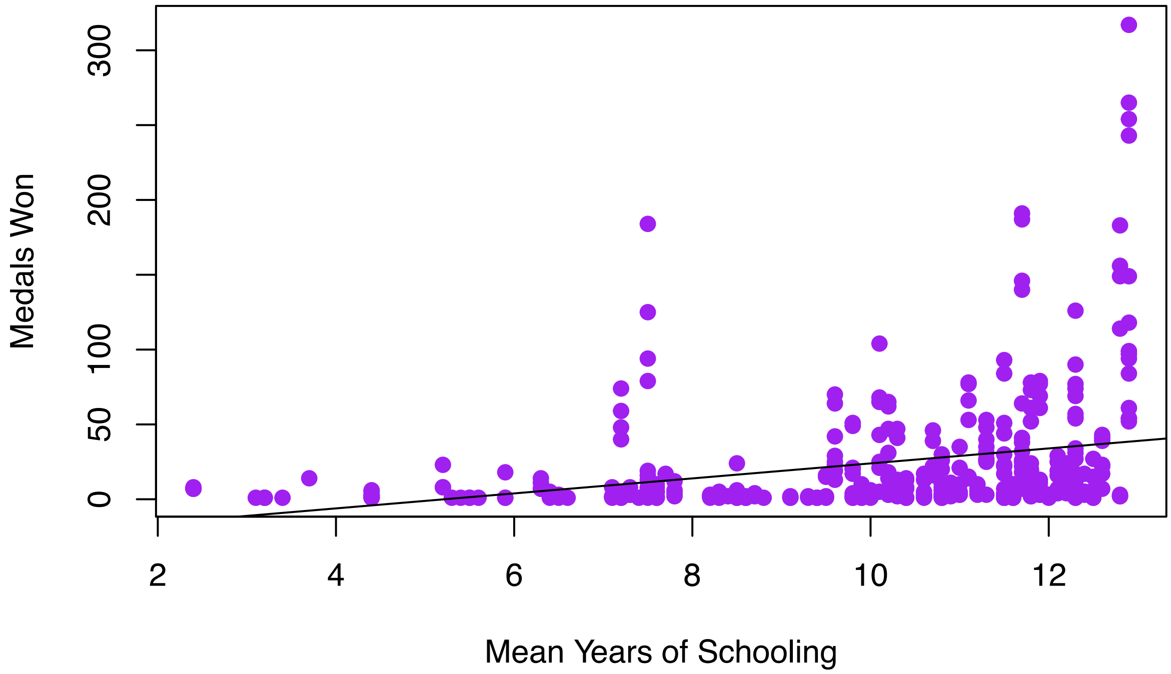
```r
medals = aggregate(Total.Medals ~ CC+Year, olympics, sum)
#medals
gdp = read.csv('CountryGDP.csv')
colnames(gdp)[2] ="CC"
medalsgdp = merge(medals,gdp)
medalshdi = merge(medals,hdi)
colnames(medalshdi)[10] = "GNI"
plot(medalshdi$HDI.rank, medalshdi$Total.Medals, col ="navy", pch = 19,
     xlab = "HDI Rank", ylab = "Medals Won")
reg1 = lm(Total.Medals~HDI.rank, data = medalshdi)
abline(reg1, col = "red")
```



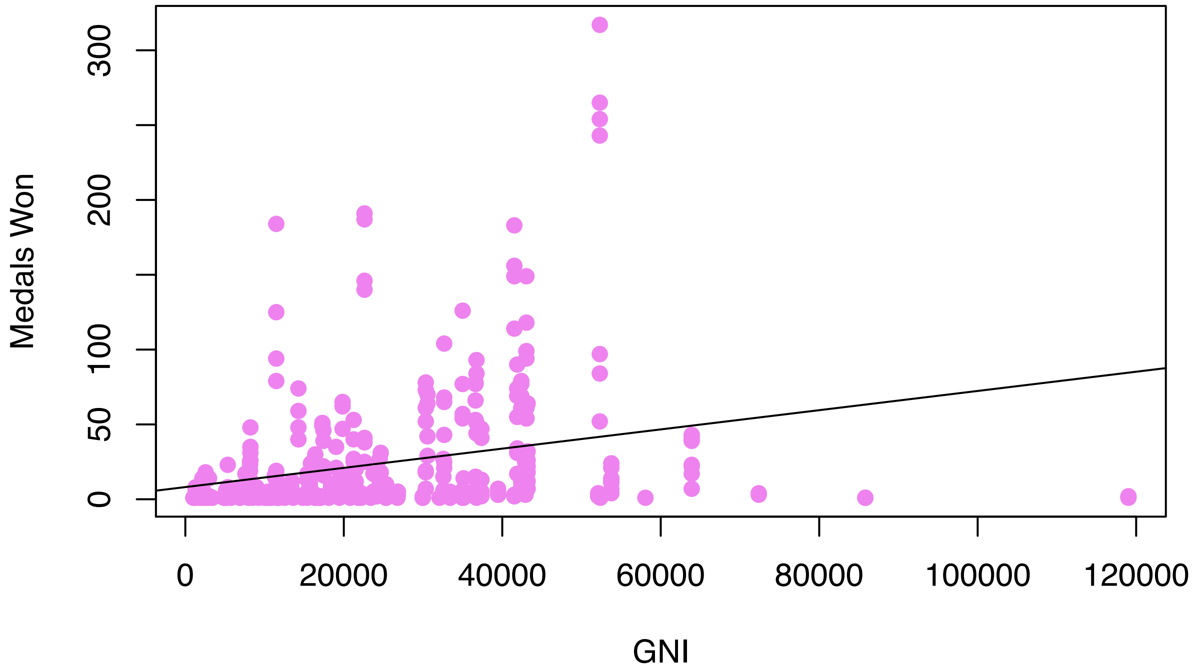
```r
plot(medalshdi$Life.expectancy.at.birth, medalshdi$Total.Medals,
     col ="pink", pch = 19, xlab = "Life Expectancy", ylab= "Medals Won")
reg2 = lm(Total.Medals~ Life.expectancy.at.birth, data = medalshdi)
abline(reg2)
```



```
plot(medalshdi$Mean.years.of.schooling,medalshdi$Total.Medals,
     col = "purple", pch = 19, xlab ="Mean Years of Schooling",
     ylab = "Medals Won")
reg3 = lm(Total.Medals~ Mean.years.of.schooling, data = medalshdi)
abline(reg3)
```

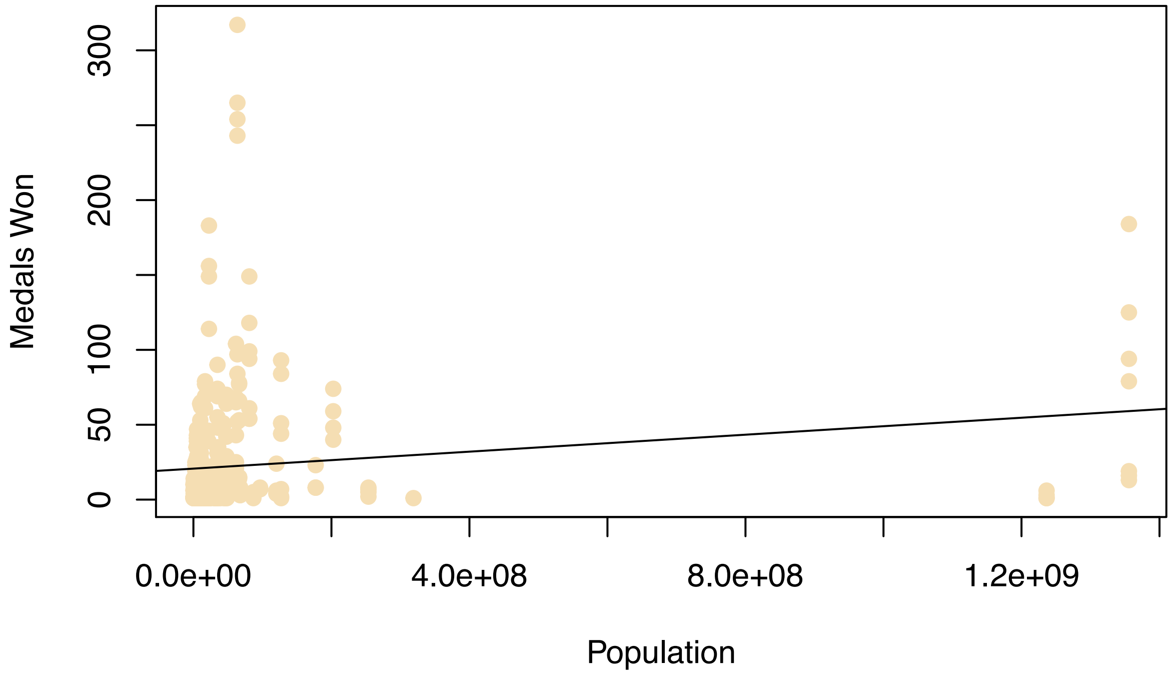


```r
plot(medalshdi$GNI,medalshdi$Total.Medals, col = "violet",
     pch = 19, xlab ="GNI", ylab = "Medals Won")
reg4 = lm(Total.Medals~ GNI, data = medalshdi)
abline(reg4)
```



Note :Due to the fact that the population data in the given file was incorrect, we've used the population data from the World Bank for the population analysis

```
pop = read.csv('Pop.csv')
pop = pop[1:3]
colnames(pop)[2] = "CC"
medalshdi =merge(medalshdi,pop)
plot(medalshdi$Population,medalshdi$Total.Medals,
     col = "wheat", pch = 19, xlab ="Population", ylab = "Medals Won")
reg5 = lm(Total.Medals~ Population, data = medalshdi)
abline(reg5)
```



From the graphs we see that the countries that perform best at the games have higher mean years of schooling, life expectancy and rank higher on the HD index as compared to those which perform poorly.Population and GNI are less indicative factors.

These results can be sumamrized by a correlation matrix, which shows the correlation between the Total Medals and Mean Years of schooling. Life expectancy, Population, GNI and HDI Rank.

```
c= data.frame(Medals = medalshdi$Total.Medals,
              Education =medalshdi$Mean.years.of.schooling,
              Expectancy = medalshdi$Life.expectancy.at.birth,
              Population = medalshdi$Population, GNI = medalshdi$GNI,
              HDI = medalshdi$HDI.rank)
cor(c)
```

```
##                 Medals  Education  Expectancy  Population        GNI
## Medals      1.0000000  0.2629496  0.27712237  0.16195733  0.2834151
## Education   0.2629496  1.0000000  0.62258784 -0.32230301  0.5366120
## Expectancy  0.2771224  0.6225878  1.00000000 -0.09660907  0.6645058
## Population  0.1619573 -0.3223030 -0.09660907  1.00000000 -0.1847297
## GNI         0.2834151  0.5366120  0.66450585 -0.18472971  1.0000000
## HDI        -0.3097215 -0.8468943 -0.87868175  0.23345613 -0.7771446
##                   HDI
## Medals     -0.3097215
## Education  -0.8468943
## Expectancy -0.8786818
## Population  0.2334561
## GNI        -0.7771446
## HDI         1.0000000
```

Continuing with our analysis, we categorize the nations as per the World Bank income classification by GNI per capita as follows:

Low income: $1,035 or less Lower middle income: $1,036 to $4,085 Upper middle income: $4,086 to $12,615 High income: $12,616 or more

We also classified the countries by the education level as follows:

Poor Education : 6 years of schooling or less Avg Education : 6 years to 10 yeats Good Education : 10 years or more

We then tried to analyze the countries performance by Income level and Education level.

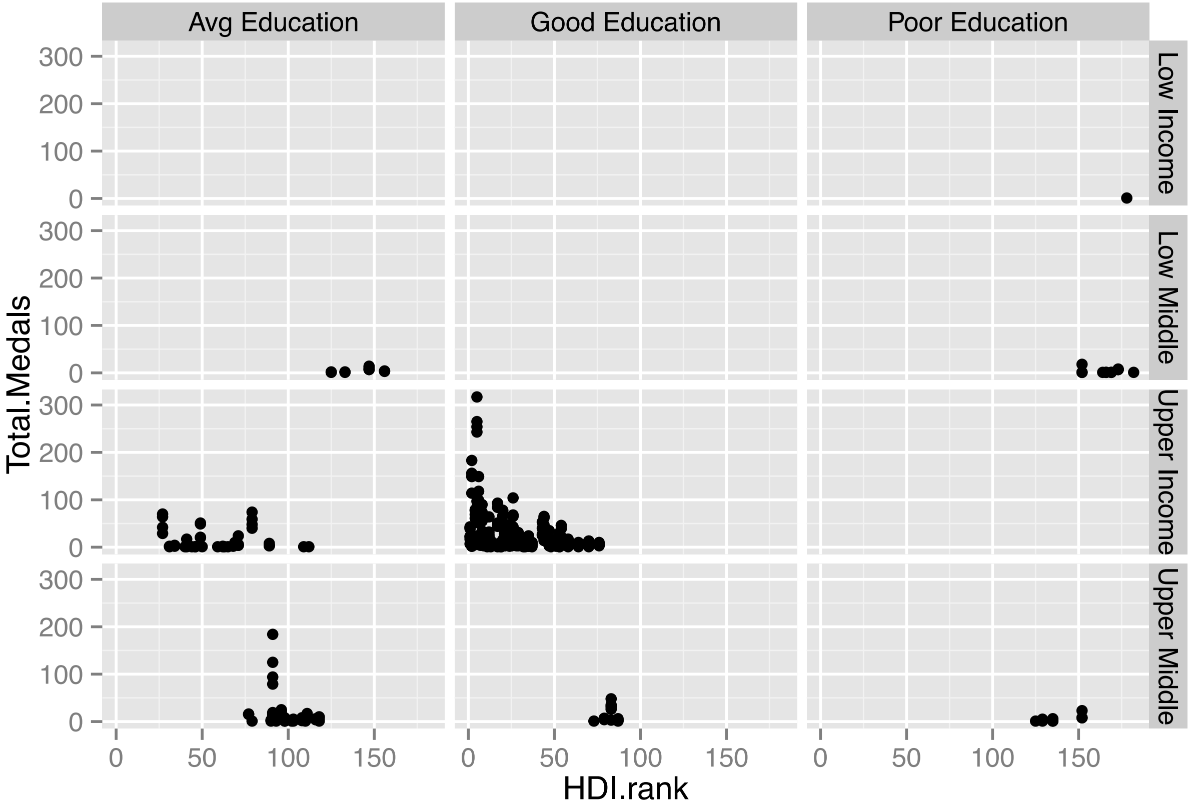
```
medalshdi$incomecat[medalshdi$GNI <= 1035] <- "Low Income"
medalshdi$incomecat[medalshdi$GNI > 1035 & medalshdi$GNI <= 4085] <- "Low Middle"
medalshdi$incomecat[medalshdi$GNI > 4085 & medalshdi$GNI < 12615] <-"Upper Middle"
medalshdi$incomecat[medalshdi$GNI > 12615] <-"Upper Income"
medalshdi$schoolcat[medalshdi$Mean.years.of.schooling <= 6] = "Poor Education"
medalshdi$schoolcat[medalshdi$Mean.years.of.schooling > 6 &
                      medalshdi$Mean.years.of.schooling <= 10] = "Avg Education"
medalshdi$schoolcat[medalshdi$Mean.years.of.schooling  >10] = "Good Education"

library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.1.3
```

```
b = ggplot(data = medalshdi, aes(x=HDI.rank, y = Total.Medals))

b = b +geom_point(size = 2)
b = b + facet_grid(incomecat~schoolcat)
b
```



We observe that the low income, low education countries rarely performed well at the games. Although countries like India and Guatemala which have a poor education level but better income level perform slightly better

Not surprisngly, the countries that perform the best have the highest education and income levels.

SUMMARY

Based on the above findings from the initial exploratory data analysis, we'd like to summarize the conclusions as follows 1. European and North American contries are generally more succesfull at the olympics. These could be due to a combination of reasons including infrastructure, training, nation's development. By conducting a multiple linear regression analysis as part of the next phase of analysis, we can identify the factors that contribute to wins at the olympics. 2. European and North American countries also have greater number of older athletes winning medals. We can explore the factors like healthcare, diet, physical traits of athletes from different nations as well as nature of sports they contest to understand why older athletes in some countries perform better than others. 3. By analyzing data from more previous games and also collecting more information about the athletes, more meaningful and actionable insights can be gathered.