Data Cleaning of Olympic Athletes Project 10/7/2018

## Load Data

setwd("/Users/ltl/Desktop/2018-Fall/INLS641/Project")  
olympics <- read.csv("athlete\_events.csv")  
olympics <- as.tibble(olympics)  
olympics

## # A tibble: 271,116 x 15  
## ID Name Sex Age Height Weight Team NOC Games Year Season  
## <int> <fct> <fct> <int> <int> <dbl> <fct> <fct> <fct> <int> <fct>   
## 1 1 A Di… M 24 180 80 China CHN 1992… 1992 Summer  
## 2 2 A La… M 23 170 60 China CHN 2012… 2012 Summer  
## 3 3 Gunn… M 24 NA NA Denm… DEN 1920… 1920 Summer  
## 4 4 Edga… M 34 NA NA Denm… DEN 1900… 1900 Summer  
## 5 5 Chri… F 21 185 82 Neth… NED 1988… 1988 Winter  
## 6 5 Chri… F 21 185 82 Neth… NED 1988… 1988 Winter  
## 7 5 Chri… F 25 185 82 Neth… NED 1992… 1992 Winter  
## 8 5 Chri… F 25 185 82 Neth… NED 1992… 1992 Winter  
## 9 5 Chri… F 27 185 82 Neth… NED 1994… 1994 Winter  
## 10 5 Chri… F 27 185 82 Neth… NED 1994… 1994 Winter  
## # ... with 271,106 more rows, and 4 more variables: City <fct>,  
## # Sport <fct>, Event <fct>, Medal <fct>

## Remove Winter Olympic Data & Null data

# Remove Winter Olympic Data  
# The size of dataset was reduced from 271,116 to 222,552 rows.  
SummerData <- olympics %>% filter(Season == "Summer")  
SummerData

## # A tibble: 222,552 x 15  
## ID Name Sex Age Height Weight Team NOC Games Year Season  
## <int> <fct> <fct> <int> <int> <dbl> <fct> <fct> <fct> <int> <fct>   
## 1 1 A Di… M 24 180 80 China CHN 1992… 1992 Summer  
## 2 2 A La… M 23 170 60 China CHN 2012… 2012 Summer  
## 3 3 Gunn… M 24 NA NA Denm… DEN 1920… 1920 Summer  
## 4 4 Edga… M 34 NA NA Denm… DEN 1900… 1900 Summer  
## 5 8 "Cor… F 18 168 NA Neth… NED 1932… 1932 Summer  
## 6 8 "Cor… F 18 168 NA Neth… NED 1932… 1932 Summer  
## 7 10 "Ein… M 26 NA NA Finl… FIN 1952… 1952 Summer  
## 8 12 Jyri… M 31 172 70 Finl… FIN 2000… 2000 Summer  
## 9 13 Minn… F 30 159 55.5 Finl… FIN 1996… 1996 Summer  
## 10 13 Minn… F 34 159 55.5 Finl… FIN 2000… 2000 Summer  
## # ... with 222,542 more rows, and 4 more variables: City <fct>,  
## # Sport <fct>, Event <fct>, Medal <fct>

# Remove rows with null values in Height or Weight or Age  
# The size of dataset was reduced from 222,552 to 166,706 rows.  
SummerData <-   
 SummerData %>% drop\_na(Height, Weight, Age)  
SummerData

## # A tibble: 166,706 x 15  
## ID Name Sex Age Height Weight Team NOC Games Year Season  
## <int> <fct> <fct> <int> <int> <dbl> <fct> <fct> <fct> <int> <fct>   
## 1 1 A Di… M 24 180 80 China CHN 1992… 1992 Summer  
## 2 2 A La… M 23 170 60 China CHN 2012… 2012 Summer  
## 3 12 Jyri… M 31 172 70 Finl… FIN 2000… 2000 Summer  
## 4 13 Minn… F 30 159 55.5 Finl… FIN 1996… 1996 Summer  
## 5 13 Minn… F 34 159 55.5 Finl… FIN 2000… 2000 Summer  
## 6 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 7 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 8 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 9 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 10 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## # ... with 166,696 more rows, and 4 more variables: City <fct>,  
## # Sport <fct>, Event <fct>, Medal <fct>

## Select the Will-be-excluded Sports

We need to decide and select the sports that will be excluded from our project. Those sports are discontinued sports or sports with fewer participant countries. We have no bias on those sports. But in order to keep consistent with the scope of our project, we will remove them.

First, we extract two tables from our original dataset. One is spt\_year\_num (sport year information) which contains the sport name, first held year, last held year and held times. Another is spt\_ctry\_num (sport country number) which contains the sport name and the number of participant country.

# spt\_held\_number  
spt\_year <-   
 SummerData %>%  
 group\_by(Sport, Year) %>%  
 count()  
  
spt\_held\_number <-  
 spt\_year %>%  
 group\_by(Sport) %>%  
 count() %>%  
 arrange(nn)  
  
# spt\_year\_info (firstY, lastY)  
spt\_year\_info <-  
 spt\_year %>%  
 group\_by(Sport) %>%  
 summarise(lastY = max(Year), firstY = min(Year)) %>%  
 arrange(lastY)  
  
# spt\_year\_num (combine upper two tables' info)  
spt\_year\_num <-  
 inner\_join(spt\_year\_info, spt\_held\_number, by = "Sport") %>%  
 arrange(nn, lastY) %>%  
 rename(held\_times = nn)  
  
spt\_year\_num

## # A tibble: 43 x 4  
## Sport lastY firstY held\_times  
## <fct> <dbl> <dbl> <int>  
## 1 Lacrosse 1904 1904 1  
## 2 Motorboating 1908 1908 1  
## 3 Figure Skating 1920 1920 1  
## 4 Ice Hockey 1920 1920 1  
## 5 Golf 2016 2016 1  
## 6 Rugby Sevens 2016 2016 1  
## 7 Tug-Of-War 1908 1904 3  
## 8 Rugby 1924 1900 3  
## 9 Softball 2008 1996 4  
## 10 Art Competitions 1948 1924 5  
## # ... with 33 more rows

# spt\_ctry\_num  
ctr\_event <-   
 SummerData %>%   
 select(Sport, NOC)  
  
spt\_ctry\_num <-   
 ctr\_event[!duplicated(ctr\_event),] %>%  
 group\_by(Sport) %>%  
 count() %>%  
 arrange(n) %>%  
 rename(number\_of\_countries = n)  
  
spt\_ctry\_num

## # A tibble: 43 x 2  
## # Groups: Sport [43]  
## Sport number\_of\_countries  
## <fct> <int>  
## 1 Figure Skating 1  
## 2 Lacrosse 1  
## 3 Motorboating 1  
## 4 Ice Hockey 2  
## 5 Rugby 3  
## 6 Tug-Of-War 6  
## 7 Art Competitions 7  
## 8 Softball 12  
## 9 Rugby Sevens 14  
## 10 Baseball 15  
## # ... with 33 more rows

Then use the upper two tables, we selected the candidates of the will-be-excluded sports using two principles and combine their results.

* a: according to the number of held times of a Sport program. We found the sports held less than 10 times in Olympic history and not held in recent 5 Olympics. Besides, we also selected the special sport of one Olympic which has the characteristic that lastY == firstY.
* b: according to the number of participant countries. We found the sports have less than 20 participant countries.

# Sports held less than 10 times in Olympic history and not held in recent 5 Olympics  
# Besides, "lastY == firstY" implies the special sport events of each Olympic.  
# So, they will be removed too.  
a <- filter(spt\_year\_num, (lastY == firstY) | ((held\_times <= 40) & (lastY < 1996)))  
a$Sport

## [1] Lacrosse Motorboating Figure Skating Ice Hockey   
## [5] Golf Rugby Sevens Tug-Of-War Rugby   
## [9] Art Competitions  
## 66 Levels: Aeronautics Alpine Skiing Alpinism Archery ... Wrestling

# Sports have less than 20 participant countries  
b <- filter(spt\_ctry\_num, number\_of\_countries <= 20)  
b$Sport

## [1] Figure Skating Lacrosse Motorboating Ice Hockey   
## [5] Rugby Tug-Of-War Art Competitions Softball   
## [9] Rugby Sevens Baseball   
## 66 Levels: Aeronautics Alpine Skiing Alpinism Archery ... Wrestling

# Sports we will remove from the dataset  
(c <- union(a$Sport, b$Sport))

## [1] "Lacrosse" "Motorboating" "Figure Skating"   
## [4] "Ice Hockey" "Golf" "Rugby Sevens"   
## [7] "Tug-Of-War" "Rugby" "Art Competitions"  
## [10] "Softball" "Baseball"

## Get Updated Dataset

Finally, remove data of those sports and store our data into a new csv file.

# Remove rows with Sport in Sports  
# The size of dataset was reduced from 166,706 to 164,913 rows  
ProjectData <-   
 SummerData %>%  
 filter(!Sport %in% c)  
ProjectData

## # A tibble: 164,913 x 15  
## ID Name Sex Age Height Weight Team NOC Games Year Season  
## <int> <fct> <fct> <int> <int> <dbl> <fct> <fct> <fct> <int> <fct>   
## 1 1 A Di… M 24 180 80 China CHN 1992… 1992 Summer  
## 2 2 A La… M 23 170 60 China CHN 2012… 2012 Summer  
## 3 12 Jyri… M 31 172 70 Finl… FIN 2000… 2000 Summer  
## 4 13 Minn… F 30 159 55.5 Finl… FIN 1996… 1996 Summer  
## 5 13 Minn… F 34 159 55.5 Finl… FIN 2000… 2000 Summer  
## 6 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 7 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 8 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 9 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## 10 17 Paav… M 28 175 64 Finl… FIN 1948… 1948 Summer  
## # ... with 164,903 more rows, and 4 more variables: City <fct>,  
## # Sport <fct>, Event <fct>, Medal <fct>

# Output the new dataset  
write.csv(ProjectData, './Project\_Data.csv')