Project 1: Olmypic History Trend Analytics

Group 9 - Thomas George Thomas, Yang Liu, Pratyush Pothuneedi

10/28/2021

```
#Importing Required packages
library(tidyverse)
library(reshape2)
library(dplyr)
library(knitr)

## Uncomment below to set the working directory.
##setwd("C:/Users/Docs")
```

1. Data Acquisition

importing the datasets

```
# Data of the athelets and countries
atheletes_df <- read.csv('athlete_events.csv', header = TRUE, sep = ',')
head(atheletes_df, 5) # structure of the dataset
```

```
##   ID                     Name Sex Age Height Weight           Team NOC
## 1  1                A Dijiang   M  24    180     80          China CHN
## 2  2                 A Lamusi   M  23    170     60          China CHN
## 3  3      Gunnar Nielsen Aaby   M  24     NA     NA        Denmark DEN
## 4  4     Edgar Lindenau Aabye   M  34     NA     NA Denmark/Sweden DEN
## 5  5 Christine Jacoba Aaftink   F  21    185     82    Netherlands NED
##         Games Year Season      City         Sport
## 1 1992 Summer 1992 Summer Barcelona    Basketball
## 2 2012 Summer 2012 Summer    London          Judo
## 3 1920 Summer 1920 Summer Antwerpen      Football
## 4 1900 Summer 1900 Summer     Paris    Tug-Of-War
## 5 1988 Winter 1988 Winter   Calgary Speed Skating
##                              Event Medal
## 1      Basketball Men’s Basketball  <NA>
## 2     Judo Men’s Extra-Lightweight  <NA>
## 3          Football Men’s Football  <NA>
## 4      Tug-Of-War Men’s Tug-Of-War  Gold
## 5 Speed Skating Women’s 500 metres  <NA>
```

```
# Importing Data of the regions tied with the NOC code
regions_df <- read.csv('noc_regions.csv', header= TRUE, sep =',')
head(regions_df, 5)
```

```
##   NOC      region                notes
## 1 AFG Afghanistan
## 2 AHO     Curacao Netherlands Antilles
## 3 ALB     Albania
## 4 ALG     Algeria
## 5 AND     Andorra
```

2. Data Wrangling

2.1 Data Discovery

Summary Statistics

```
summary(atheletes_df)
```

```
##        ID             Name              Sex                 Age
##  Min.   :     1   Length:271116      Length:271116      Min.   :10.00
##  1st Qu.: 34643   Class :character   Class :character   1st Qu.:21.00
##  Median : 68205   Mode  :character   Mode  :character   Median :24.00
##  Mean   : 68249                                         Mean   :25.56
##  3rd Qu.:102097                                         3rd Qu.:28.00
##  Max.   :135571                                         Max.   :97.00
##                                                         NA’s   :9474
##      Height          Weight          Team               NOC
##  Min.   :127.0   Min.   : 25.0   Length:271116      Length:271116
##  1st Qu.:168.0   1st Qu.: 60.0   Class :character   Class :character
##  Median :175.0   Median : 70.0   Mode  :character   Mode  :character
##  Mean   :175.3   Mean   : 70.7
##  3rd Qu.:183.0   3rd Qu.: 79.0
##  Max.   :226.0   Max.   :214.0
##  NA’s   :60171   NA’s   :62875
##     Games                Year         Season              City
##  Length:271116      Min.   :1896   Length:271116      Length:271116
##  Class :character   1st Qu.:1960   Class :character   Class :character
##  Mode  :character   Median :1988   Mode  :character   Mode  :character
##                     Mean   :1978
##                     3rd Qu.:2002
##                     Max.   :2016
##
##     Sport              Event              Medal
##  Length:271116      Length:271116      Length:271116
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

We can see that there are NA's in the numerical fields of Age, Height, Weight which we will be handling

```
summary(regions_df)
```

```
##      NOC               region             notes          
##  Length:230         Length:230         Length:230        
##  Class :character   Class :character   Class :character  
##  Mode  :character   Mode  :character   Mode  :character  
```

Discovering Discrete Data

```
kable(
  atheletes_df %>%
    summarise(total_records=n()),
  caption = "Total Records for Athletes"

)
```

Table 1: Total Records for Athletes

total_records
271116

```
kable(
  regions_df %>%
    summarise(total_records=n()),
  caption = "Total Records in Regions"

)
```

Table 2: Total Records in Regions

total_records
230

Looking for NA's in all the columns

```
colnames(atheletes_df)[apply(atheletes_df, 2, anyNA)]
```

```
## [1] "Age"    "Height" "Weight" "Medal"
```

We already knew that Age, Height & Weight have NA's as seen above in the summary. We found that categorical Medal's have NA, taking a closer look:

```
kable(
  atheletes_df %>%
    group_by(Medal) %>%
    summarise(total_records=n())
    ,caption="Records by Medal Count"

)
```

Table 3: Records by Medal Count

Medal	total_records
Bronze	13295
Gold	13372
Silver	13116
NA	231333

We can see that there are 231333 NA’s for Medals which happens to be categorical data and we need to handle this in the cleaning part

2.2 Structuring

```
head(atheletes_df,5)
```

```
##   ID                     Name Sex Age Height Weight           Team NOC
## 1  1                A Dijiang   M  24    180     80          China CHN
## 2  2                 A Lamusi   M  23    170     60          China CHN
## 3  3      Gunnar Nielsen Aaby   M  24     NA     NA        Denmark DEN
## 4  4     Edgar Lindenau Aabye   M  34     NA     NA Denmark/Sweden DEN
## 5  5 Christine Jacoba Aaftink   F  21    185     82    Netherlands NED
##         Games Year Season      City         Sport
## 1 1992 Summer 1992 Summer Barcelona    Basketball
## 2 2012 Summer 2012 Summer    London          Judo
## 3 1920 Summer 1920 Summer Antwerpen      Football
## 4 1900 Summer 1900 Summer     Paris    Tug-Of-War
## 5 1988 Winter 1988 Winter   Calgary Speed Skating
##                                 Event Medal
## 1      Basketball Men’s Basketball  <NA>
## 2     Judo Men’s Extra-Lightweight  <NA>
## 3          Football Men’s Football  <NA>
## 4      Tug-Of-War Men’s Tug-Of-War  Gold
## 5 Speed Skating Women’s 500 metres  <NA>
```

We can see that we don’t need to do additional restructuring as columns like “Games” is already split and available as Year and Season

2.3 Cleaning

Handling Missing Data

We can't filter out the NA values since the columns that exhibit them are required for our analysis. We will be filling the NA values in Age, Height, Weight with Mean values and Medals with 'None'. Here, Medals with 'None' would signify that the athletes simply didn't win any of the categories of Medals (Gold, Silver, Bronze). We didn't filter out the NA records in Age, Height and Weight because that would mean that crucial data would be missing leading to data skewness, we are using the Mean values to reduce the degree of skewness while maintaining data integrity

Replacing NA's in Medals

```
atheletes_df$Medal %>%
  replace_na("None")
```