Zeru-Zhou-project03

September 13, 2021

1 Project 3 – Zeru Zhou

TA Help: NA

Collaboration: NA

- Get help at piazza
- Get help from videos posted by Dr. Ward

1.1 Question 1

[1]:
```
list.files("/depot/datamine/data/olympics")
```

1. 'athlete_events.csv' 2. 'regions.csv'

[3]:
```
dim(read.csv("/depot/datamine/data/olympics/athlete_events.csv"))
```

1. 271116 2. 15

[4]:
```
dim(read.csv("/depot/datamine/data/olympics/regions.csv"))
```

1. 230 2. 3

[1]:
```
olympics <- read.csv("/depot/datamine/data/olympics/athlete_events.csv")
```

[8]:
```
dim(olympics)
```

1. 271116 2. 15

[6]:
```
head(olympics)
```

A data.frame: 6 x 15

ID <int>	Name <chr>	Sex <chr>	Age <int>	Height <int>	Weight <dbl>	Team <chr>
1	A Dijiang	M	24	180	80	China
2	A Lamusi	M	23	170	60	China
3	Gunnar Nielsen Aaby	M	24	NA	NA	Denmark
4	Edgar Lindenau Aabye	M	34	NA	NA	Denmark/Sweden
5	Christine Jacoba Aaftink	F	21	185	82	Netherlands
5	Christine Jacoba Aaftink	F	21	185	82	Netherlands

[7]:
```
str(olympics)
```

```
'data.frame':   271116 obs. of  15 variables:
 $ ID    : int  1 2 3 4 5 5 5 5 5 5 …
 $ Name  : chr  "A Dijiang" "A Lamusi" "Gunnar Nielsen Aaby" "Edgar Lindenau
Aabye" …
 $ Sex   : chr  "M" "M" "M" "M" …
 $ Age   : int  24 23 24 34 21 21 25 25 27 27 …
 $ Height: int  180 170 NA NA 185 185 185 185 185 185 …
 $ Weight: num  80 60 NA NA 82 82 82 82 82 82 …
 $ Team  : chr  "China" "China" "Denmark" "Denmark/Sweden" …
 $ NOC   : chr  "CHN" "CHN" "DEN" "DEN" …
 $ Games : chr  "1992 Summer" "2012 Summer" "1920 Summer" "1900 Summer" …
 $ Year  : int  1992 2012 1920 1900 1988 1988 1992 1992 1994 1994 …
 $ Season: chr  "Summer" "Summer" "Summer" "Summer" …
 $ City  : chr  "Barcelona" "London" "Antwerpen" "Paris" …
 $ Sport : chr  "Basketball" "Judo" "Football" "Tug-Of-War" …
 $ Event : chr  "Basketball Men's Basketball" "Judo Men's Extra-Lightweight"
"Football Men's Football" "Tug-Of-War Men's Tug-Of-War" …
 $ Medal : chr  NA NA NA "Gold" …
```

Olympics has 271116 rows and 15 columns, and there are 3 types of data: integer, numeric and character. Each row contains basic information about a specific athlete.

1.2 Question 2

[10]:
```
length(unique(olympics$Sport))
```

66

[13]:
```
unique(olympics$Sport)
```

1. 'Basketball' 2. 'Judo' 3. 'Football' 4. 'Tug-Of-War' 5. 'Speed Skating' 6. 'Cross Country Skiing' 7. 'Athletics' 8. 'Ice Hockey' 9. 'Swimming' 10. 'Badminton' 11. 'Sailing' 12. 'Biathlon' 13. 'Gymnastics' 14. 'Art Competitions' 15. 'Alpine Skiing' 16. 'Handball' 17. 'Weightlifting' 18. 'Wrestling' 19. 'Luge' 20. 'Water Polo' 21. 'Hockey' 22. 'Rowing' 23. 'Bobsleigh' 24. 'Fencing' 25. 'Equestrianism' 26. 'Shooting' 27. 'Boxing' 28. 'Taekwondo' 29. 'Cycling' 30. 'Diving' 31. 'Canoeing' 32. 'Tennis' 33. 'Modern Pentathlon' 34. 'Figure Skating' 35. 'Golf' 36. 'Softball' 37. 'Archery' 38. 'Volleyball' 39. 'Synchronized Swimming' 40. 'Table Tennis' 41. 'Nordic Combined' 42. 'Baseball' 43. 'Rhythmic Gymnastics' 44. 'Freestyle Skiing' 45. 'Rugby Sevens' 46. 'Trampolining' 47. 'Beach Volleyball' 48. 'Triathlon' 49. 'Ski Jumping' 50. 'Curling' 51. 'Snowboarding' 52. 'Rugby' 53. 'Short Track Speed Skating' 54. 'Skeleton' 55. 'Lacrosse' 56. 'Polo' 57. 'Cricket' 58. 'Racquets' 59. 'Motorboating' 60. 'Military Ski Patrol' 61. 'Croquet' 62. 'Jeu De Paume' 63. 'Roque' 64. 'Alpinism' 65. 'Basque Pelota' 66. 'Aeronautics'

There are 66 unique sports in the dataset olympics. I use 'unique' function to show unique individual sports included in the dataset. There are some sports I never expected like 'Motorboating' since I thought it is more like a hobby and just for fun. Never expected it is on olympics.

1.3 Question 3

[3]:
```
us_athletes <- subset(olympics, NOC=='USA')
```

[4]:
```
dim(us_athletes)
```

1. 18853 2. 15

[5]:
```
china_athletes <- subset(olympics, NOC=='CHN')
```

[6]:
```
dim(china_athletes)
```

1. 5141 2. 15

[7]:
```
both <- subset(olympics, NOC=='USA'| NOC=='CHN')
```

[8]:
```
dim(both)
```

1. 23994 2. 15

There are 18853 rows in 'us_athletes' dataset; There are 5154 rows in dataset contains athletes from China named 'china_athletes'; There are 23994 rows in 'both' dataset.

1.4 Question 4

[9]:
```
prop.table(table(olympics$Sex[olympics$NOC=='USA']))
```

```
        F         M
0.2934811 0.7065189
```

[8]:
```
prop.table(table(olympics$Sex[(olympics$NOC=='USA') &␣
 ↪(olympics$Medal=='Gold')]))
```

```
        F         M
0.3229719 0.6770281
```

[11]:
```
prop.table(table(olympics$Sex[olympics$NOC=='CHN']))
```

```
        F         M
0.5388057 0.4611943
```

[9]:
```
prop.table(table(olympics$Sex[(olympics$NOC=='CHN') &␣
 ↪(olympics$Medal=='Gold')]))
```

```
  F   M
0.6 0.4
```

29.35% athletes in the US are women; 32.30% athletes in the US with gold medals are women. 53.88% athletes in China are women; 60% athletes in China with gold medals are women.

1.5 Question 5

[2]:
```
us_athletes <- subset(olympics, NOC=='USA')
```

[16]:
```
us_age <- us_athletes$Age
```

[12]:
```
which.max(us_age)
```

17647

[14]:
```
us_athletes[which.max(us_age), c('Age', 'Sport', 'Year')]
```

A data.frame: 1 x 3

	Age	Sport	Year
	<int>	<chr>	<int>
257055	97	Art Competitions	1928

[18]:
```
china_athletes <- subset(olympics, NOC=='CHN')
```

[20]:
```
ch_age <- china_athletes$Age
```

[21]:
```
which.max(ch_age)
```

1130

[22]:
```
china_athletes[which.max(ch_age), c('Age', 'Sport', 'Year')]
```

A data.frame: 1 x 3

	Age	Sport	Year
	<int>	<chr>	<int>
100160	45	Equestrianism	2008

For the US, the oldest athlete is 97 years old, the sport is 'Art Competitions', and the olympics year is 1928. For China, the oldest athlete is 45 years old, the sport is 'Equestrianism', and the olympics year is 2008.

1.6 Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together – We are Purdue.