STA130 Winter 2020 R Tutorial 3 Sample Solutions

Vivian Ngo

In this tutorial, we will experiment with the oly12 dataset in VGAMdata

```
library(tidyverse) # load the tidyverse package
library(VGAMdata) # load the VGAMdata package so that you can access the datasets inside of it
```

```
glimpse(oly12) # brief summary of the oly12 dataframe
```

```
## Observations: 10,384
## Variables: 14
## $ Name    <fct> Lamusi A, A G Kruger, Jamale Aarrass, Abdelhak Aatakni, Mar...
## $ Country <fct> People's Republic of China, United States of America, Franc...
## $ Age     <int> 23, 33, 30, 24, 26, 27, 30, 23, 27, 19, 37, 28, 28, 28, 22,...
## $ Height  <dbl> 1.70, 1.93, 1.87, NA, 1.78, 1.82, 1.82, 1.87, 1.90, 1.70, N...
## $ Weight  <int> 60, 125, 76, NA, 85, 80, 73, 75, 80, NA, NA, NA, 60, 64, 62...
## $ Sex     <fct> M, M, M, M, F, M, F, M, M, M, M, M, F, F, M, F, M, M, M, M,...
## $ DOB     <date> 1989-02-06, NA, NA, 1988-09-02, NA, 1984-06-09, NA, 1989-0...
## $ PlaceOB <fct> NEIMONGGOL (CHN), Sheldon (USA), BEZONS (FRA), AIN SEBAA (M...
## $ Gold    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Silver  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Bronze  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Total   <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Sport   <fct> Judo, Athletics, Athletics, Boxing, Athletics, Handball, Ro...
## $ Event   <fct> "Men's -60kg", "Men's Hammer Throw", "Men's 1500m", "Men's ...
```

```
# oly12 # calling the dataframe itself will display the dataframe, but not a summary

# View(oly12) # the View function (with a capital "v") will open another tab in R and show you the data
# if you view this dataframe, you will see that each row corresponds to an athlete that participated in
```

To find Canadian athletes:

```
# Using filter to keep only canadian athletes,
# then glimpse to view the number of observations
oly12 %>% filter(Country == "Canada") %>%
  glimpse()
```

```
## Observations: 274
## Variables: 14
## $ Name    <fct> Jennifer Abel, Natalie Achonwa, Mohammed Ahmed, Dylan Armst...
## $ Country <fct> Canada, Canada, Canada, Canada, Canada, Canada, Canada, Can...
```

```
## $ Age     <int> 20, 19, 21, 31, 28, 24, 20, 28, 23, 22, 21, 56, 29, 24, 23,...
## $ Height  <dbl> 1.60, 1.92, 1.90, 1.93, 1.85, 1.83, 1.68, 1.86, 1.86, 1.68,...
## $ Weight  <int> 62, 83, 60, 139, 82, 78, 150, 90, 80, 58, 75, 78, 98, 48, 6...
## $ Sex     <fct> F, F, M, M, F, F, M, M, M, F, M, M, M, F, F, F, M, M, F, F,...
## $ DOB     <date> NA, NA, 1991-05-01, NA, NA, 1988-06-05, 1992-11-03, NA, NA...
## $ PlaceOB <fct> Montreal (CAN), , Mogadishu (SOM), Kamloops (CAN), , , West...
## $ Gold    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Silver  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ Bronze  <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,...
## $ Total   <int> 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,...
## $ Sport   <fct> Diving, Basketball, Athletics, Athletics, Basketball, Baske...
## $ Event   <fct> "Women's 3m Springboard, Women's Synchronised 3m Springboar...
```

```
# Using filter to keep only canadian athletes,
# then count the number of rows in the resulting data frame
oly12 %>% filter(Country == "Canada") %>%
  nrow()
```

```
## [1] 274
```

```
# Use summarise to calculate the number of athletes for each country,
# then filter to keep only the row for Canada
oly12 %>% group_by(Country) %>%
  summarise(team_size = n()) %>%
  filter(Country=="Canada")
```

```
## # A tibble: 1 x 2
##   Country team_size
##   <fct>       <int>
## 1 Canada        274
```

```
# Sum up the number of observations where Country is Canada
sum(oly12$Country=="Canada")
```

```
## [1] 274
```

```
# Filter to get Canadian countries and then sum up the rows
oly12 %>% filter(Country =="Canada")%>%
  summarize(n=n())
```

```
##     n
## 1 274
```

274 athletes represented Canada at the 2012 Olympic Games.

Create a new dataframe called `oly12_selectedSports` which contains only data for athletes who competed in Weightlifting and Badminton

```
oly12 %>% filter(Sport == "Weightlifting" | Sport == "Badminton") %>% head()
```

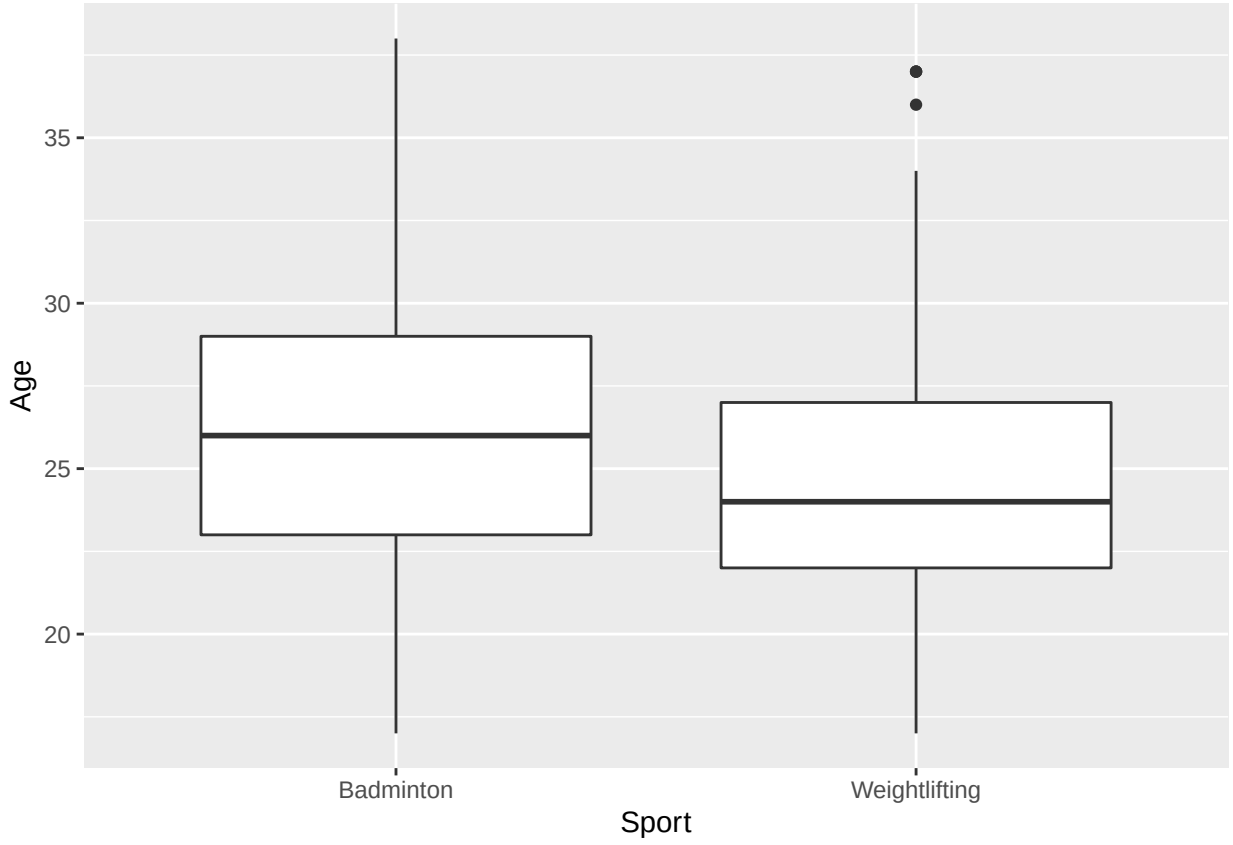
```
##                                Name       Country Age Height Weight Sex
## 1                Mohamed Abdel Baki         Egypt  25   1.62     69   M
## 2                   Tarek Abdelazim         Egypt  25   1.75     85   M
## 3                       Pablo Abian         Spain  27   1.77     78   M
```

```
## 4 Khalil Mahmoud K Abir Abdelrahman         Egypt  20   1.62     75   F
## 5        Luz Mercedes Acosta Valdez        Mexico  31   1.66     63   F
## 6                     Chris Adcock Great Britain  23   1.83     80   M
##          DOB    PlaceOB Gold Silver Bronze Total         Sport         Event
## 1       <NA>     FAYOUM    0      0      0     0 Weightlifting    Men's 69kg
## 2       <NA>    ELMENIA    0      0      0     0 Weightlifting    Men's 85kg
## 3 1985-12-06  CALATAYUD    0      0      0     0     Badminton Men's Singles
## 4       <NA> Alexandria    0      0      0     0 Weightlifting  Women's 75kg
## 5       <NA>     Sonora    0      0      0     0 Weightlifting  Women's 63kg
## 6       <NA>  Leicester    0      0      0     0     Badminton Mixed Doubles
```

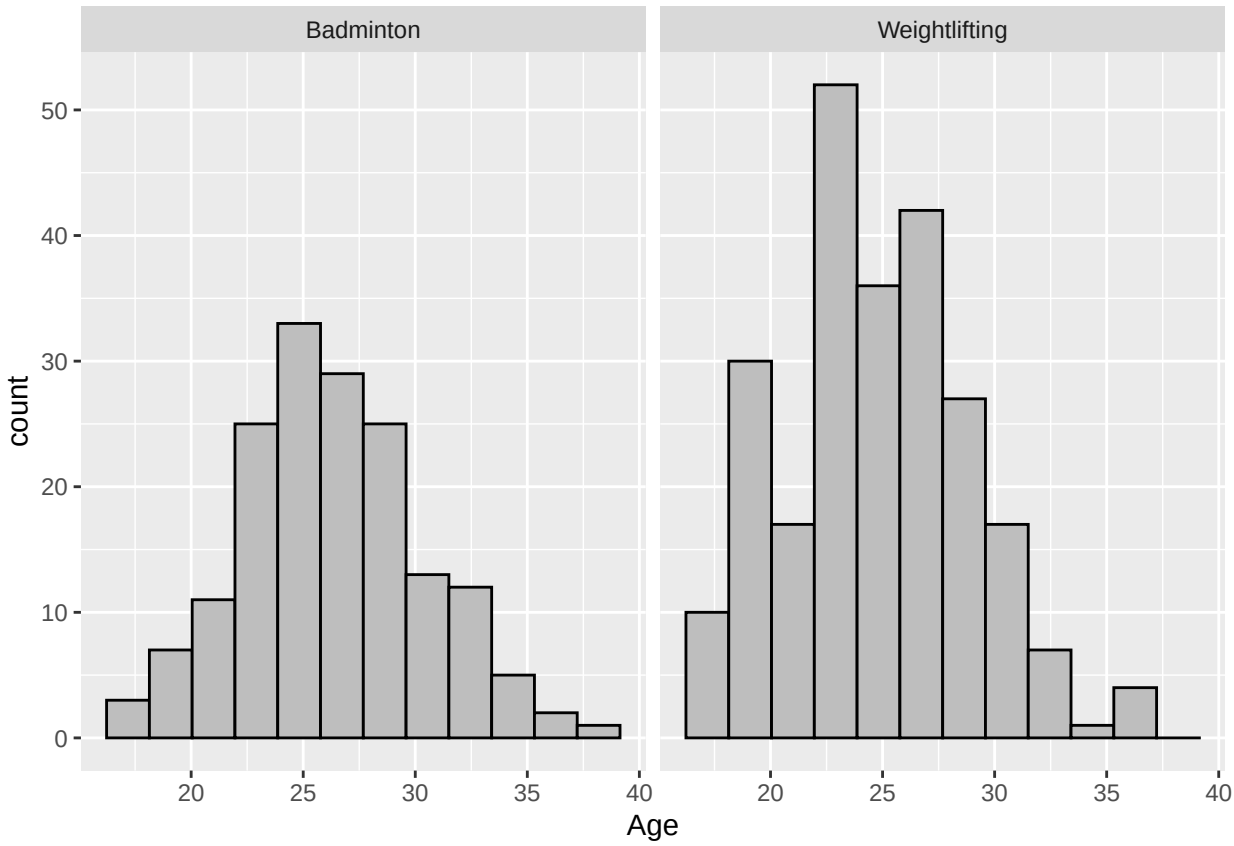
```
# !!! remember to CREATE a new dataframe!
oly12_selectedSports <- oly12 %>% filter(Sport == "Weightlifting" | Sport == "Badminton")
```

Compare the age distribution for olympic athletes competing in weightlifting and badminton using both boxplots and histograms.

```
oly12_selectedSports %>% ggplot(aes(x=Sport, y=Age)) +
  geom_boxplot()
```




```
oly12_selectedSports %>% ggplot(aes(x=Age)) +
  geom_histogram(bins=12, color="black", fill="gray") + facet_wrap(~Sport)
```



Answer the following questions:

(i) Are the age distributions of badminton players and weightlifters symmetrical or skewed?

From the histograms, we can see that the age distribution of badminton players is approximately symmetric but the age distribution of weightlifters is slightly skewed to the right. This can also be seen in the boxplots of the age distributions - in particular, we see there are two outliers in the right tail of the age distribution of weightlifters, corresponding to two weightlifters who are much older than most of the weightlifters.

(ii) Is the median age higher for badminton players or weightlifters?

```
# look back at the boxplots

# or, calculate the medians
oly12_selectedSports %>% filter(Sport =="Badminton") %>%
  summarize(median=median(Age))
```

```
##   median
## 1     26
```

```
oly12_selectedSports %>% filter(Sport =="Weightlifting") %>%
  summarize(median=median(Age))
```

```
##   median
## 1     24
```

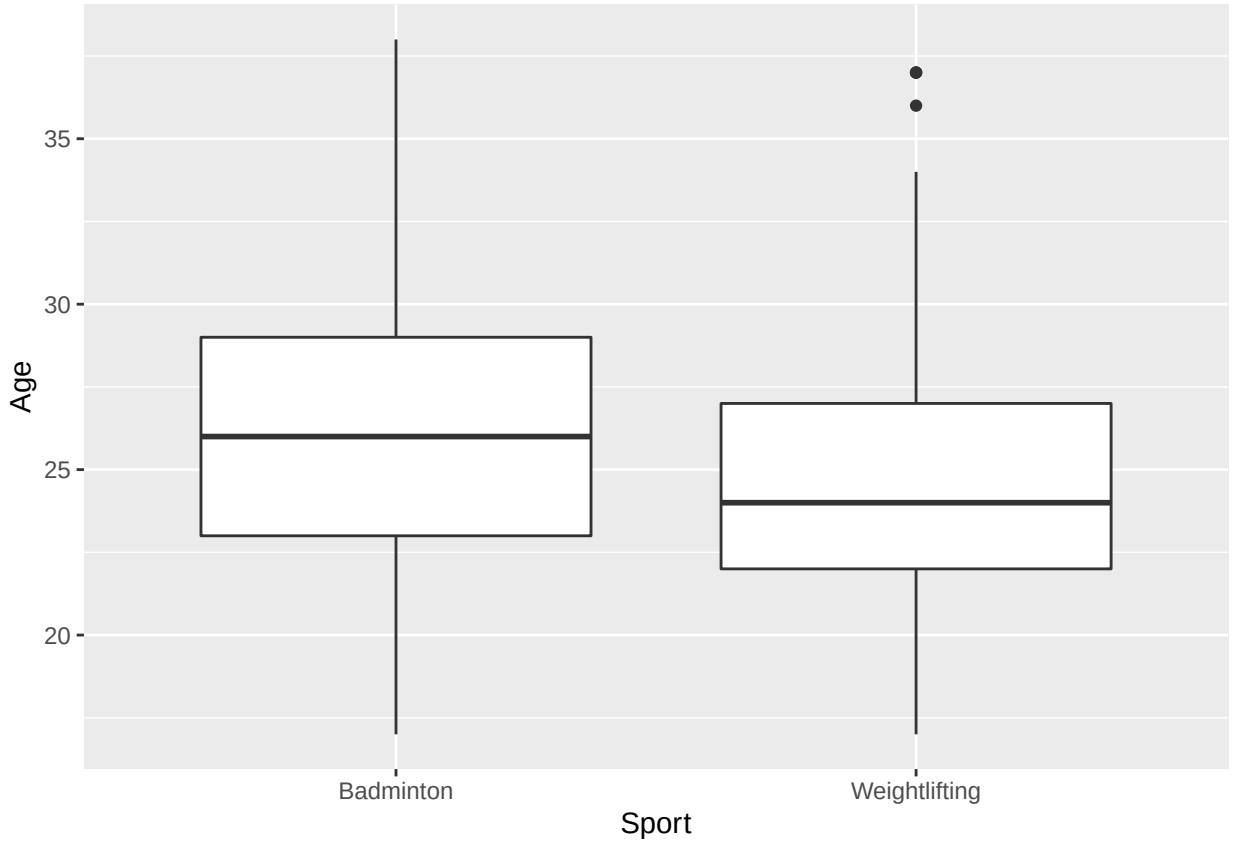
```
# group by!
oly12_selectedSports %>% group_by(Sport) %>%
  summarize(median_age=median(Age))
```

```
## # A tibble: 2 x 2
##   Sport         median_age
##   <fct>              <dbl>
## 1 Badminton             26
## 2 Weightlifting         24
```

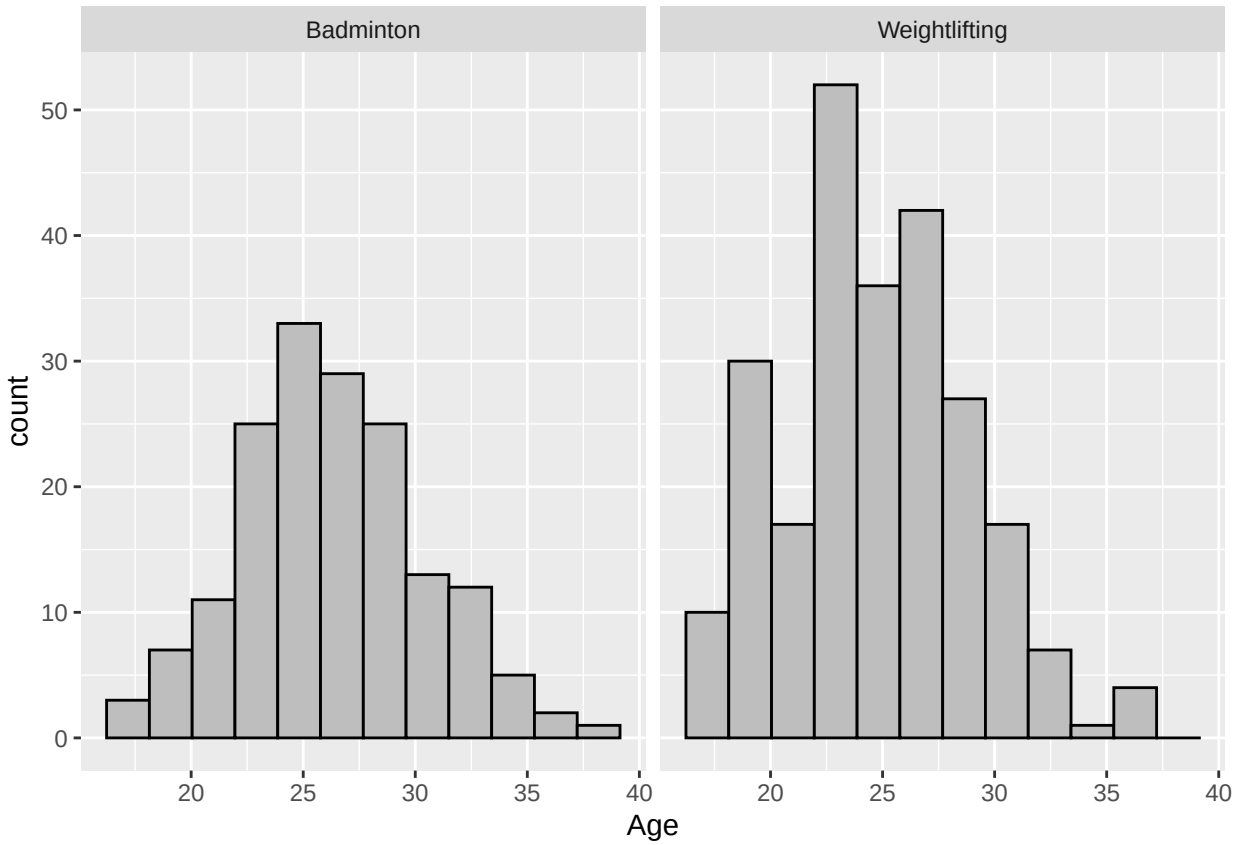
From the boxplots, we can see that the median age of badminton players is higher than the median age of weightlifters (~26 vs ~24).

(iii) Based only on the histogram and boxplots, predict whether the standard deviation of the ages is similar or different.

```
# look at the boxplot and histogram again
oly12_selectedSports %>% ggplot(aes(x=Sport, y=Age)) +
  geom_boxplot()
```




```
oly12_selectedSports %>% ggplot(aes(x=Age)) +
  geom_histogram(bins=12, color="black", fill="gray") + facet_wrap(~Sport)
```



I predict that the standard deviatiation of ages for badminton players will be a little bit larger than that of weightlifters since the IQR and whiskers are both a bit longer. However, the range of the age distributions (max - min) are similar for both sports.

Create a summary table reporting the minimum, maximum, mean, median, and standard deviation of ages for badminton players and weightlifters. Compare these values to the prediction you made in (e-iii)

```
oly12_selectedSports %>% group_by(Sport) %>%
  summarise(min=min(Age), max=max(Age), mean=mean(Age), median=median(Age), sd=sd(Age))
```

```
## # A tibble: 2 x 6
##   Sport           min   max  mean median    sd
##   <fct>         <int> <int> <dbl>  <dbl> <dbl>
## 1 Badminton        17    38  26.2     26  4.12
## 2 Weightlifting    17    37  24.6     24  4.06
```

As predicted in (e-iii) the standard deviation of ages is slightly higher for badminton players than for weightlifters (4.12 vs 4.06), but they are very similar.

Use the `arrange` function to find the name and age of the 6 oldest athletes who competed in the 2012 Olympics.

```
oly12  %>%
  arrange(desc(Age)) %>%
  head(6) %>% # default for head is 6 as well
  select(Name, Age, Sport, Event)
```

```
##                Name Age      Sport                                        Event
## 1   Hiroshi Hoketsu  71 Equestrian                   Individual Dressage, WHISPER
## 2 Afanasijs Kuzmins  65   Shooting                    Men's 25m Rapid Fire Pistol
## 3        Ian Millar  65 Equestrian Individual Jumping, Team Jumping, STAR POWER
## 4    Carl Bouckaert  58 Equestrian Individual Eventing, Team Eventing, CYRANO Z
## 5  Andrei Kavalenka  57   Shooting                                    Men's Trap
## 6        Mary Hanna  57 Equestrian Individual Dressage, Team Dressage, SANCETTE
```

```
oly12  %>%
  arrange(-Age) %>%
  head(6) %>%
  select(Name, Age, Sport, Event)
```

```
##                Name Age      Sport                                        Event
## 1   Hiroshi Hoketsu  71 Equestrian                   Individual Dressage, WHISPER
## 2 Afanasijs Kuzmins  65   Shooting                    Men's 25m Rapid Fire Pistol
## 3        Ian Millar  65 Equestrian Individual Jumping, Team Jumping, STAR POWER
## 4    Carl Bouckaert  58 Equestrian Individual Eventing, Team Eventing, CYRANO Z
## 5  Andrei Kavalenka  57   Shooting                                    Men's Trap
## 6        Mary Hanna  57 Equestrian Individual Dressage, Team Dressage, SANCETTE
```

Modify your code from (f) to find the name, Age, and event for the 6 oldest competitors who won gold medals at the 2012 olympics

```
oly12  %>%
  filter(Gold > 0) %>%
  arrange(desc(Age)) %>%
  head(6) %>%
  select(Name, Age, Sport, Event)
```

```
##                 Name Age          Sport
## 1      Peter Thomsen  51     Equestrian
## 2      Ingrid Klimke  44     Equestrian
## 3    Sergei Martynov  44       Shooting
## 4  Kristin Armstrong  38 Cycling - Road
## 5  Valentina Vezzali  38        Fencing
## 6 Alexandr Vinokurov  38 Cycling - Road
##                                                Event
## 1          Individual Eventing, Team Eventing, BARNY
## 2 Individual Eventing, Team Eventing, BUTTS ABRAXXAS
## 3                              Men's 50m Rifle Prone
## 4   Women's Individual Time Trial, Women's Road Race
## 5         Women's Individual Foil, Women's Team Foil
## 6       Men's Individual Time Trial, Men's Road Race
```

```
# why is the code below not okay?
oly12  %>%
  arrange(desc(Age)) %>%
  head(6) %>%
```

```
  select(Name, Age, Sport, Event) %>%
  filter(Gold > 0)
```

Create a new variable called `total_medals` and find the name of the athlete who won the most medals at the 2012 Olympics.

```
oly12 %>%
  mutate(total_medals = Gold + Silver + Bronze) %>%
  arrange(desc(total_medals)) %>%
  head() %>%
  select(Name, Country, Sport, total_medals)
```

```
##              Name                  Country    Sport total_medals
## 1     Ryan Lochte United States of America Swimming            5
## 2   Alicia Coutts                Australia Swimming            4
## 3  Michael Phelps United States of America Swimming            4
## 4 Allison Schmitt United States of America Swimming            4
## 5   Yannick Agnel                   France Swimming            3
## 6  Missy Franklin United States of America Swimming            3
```