IMT 573 Lab: Data Summarization

Naga Soundari Balamurugan

October 11th, 2018

Don't forget to list the full names of your collaborators!

Collaborators: Jayashree Raman

Instructions:

Before beginning this assignment, please ensure you have access to R and/or RStudio.

1. Download the `week3b_lab_OPTIONAL.Rmd` file from Canvas. Open `week3a_lab.Rmd` in RStudio (or your favorite editor) and supply your solutions to the assignment by editing `week3b_lab_OPTIONAL.Rmd`. You will also want to download the `athlete_events.csv` data file, containing historical data from the modern Olympic Games, including all the Games from Athens 1896 to Rio 2016. (see https: //www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/home)
2. Replace the "Insert Your Name Here" text in the `author:` field with your own full name.
3. Be sure to include code chucks, figures and written explanations as necessary. Any collaborators must be listed on the top of your assignment. Any figures should be clearly labeled and appropriately referenced within the text.
4. When you have completed the assignment and have **checked** that your code both runs in the Console and knits correctly when you click `Knit`, rename the R Markdown file to `YourLastName_YourFirstName_lab3b.Rmd`, and knit it into a PDF. Submit the compiled PDF or Word doc on Canvas.

In this lab, you will need access to the following R packages:

```
# Load some helpful libraries
library(tidyverse)
library(dplyr)
```

Problem 1: Data Cleaning

In this problem we will use the `athlete_events.csv` data. Import the data in **R** using: - `read.csv()` - `read.table()` - `read.delim()`

Look at the head or tail each time you try a new import method. Answer the following questions.

```
#Read the athlete events data using read.csv
athleteEventsData_CSV <- read.csv("athlete_events.csv", sep=",")
head(athleteEventsData_CSV)
```

```
##   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
## 6  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
## 6 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>
## 6 Speed Skating Women's 1,000 metres  <NA>
```

```
tail(athleteEventsData_CSV)
```

```
##            ID                   Name Sex Age Height Weight     Team NOC
## 271111 135568 Olga Igorevna Zyuzkova   F  33    171     69  Belarus BLR
## 271112 135569             Andrzej ya   M  29    179     89 Poland-1 POL
## 271113 135570               Piotr ya   M  27    176     59   Poland POL
## 271114 135570               Piotr ya   M  27    176     59   Poland POL
## 271115 135571     Tomasz Ireneusz ya   M  30    185     96   Poland POL
## 271116 135571     Tomasz Ireneusz ya   M  34    185     96   Poland POL
##              Games Year Season           City       Sport
## 271111 2016 Summer 2016 Summer Rio de Janeiro  Basketball
## 271112 1976 Winter 1976 Winter      Innsbruck        Luge
## 271113 2014 Winter 2014 Winter          Sochi Ski Jumping
## 271114 2014 Winter 2014 Winter          Sochi Ski Jumping
## 271115 1998 Winter 1998 Winter         Nagano   Bobsleigh
## 271116 2002 Winter 2002 Winter Salt Lake City   Bobsleigh
##                                        Event Medal
## 271111            Basketball Women's Basketball  <NA>
## 271112               Luge Mixed (Men)'s Doubles  <NA>
## 271113 Ski Jumping Men's Large Hill, Individual  <NA>
## 271114       Ski Jumping Men's Large Hill, Team  <NA>
## 271115                     Bobsleigh Men's Four  <NA>
## 271116                     Bobsleigh Men's Four  <NA>
```

```
#Read the athlete events data using read.csv
athleteEventsData_Table <- read.table("athlete_events.csv", header = TRUE, sep=",")
head(athleteEventsData_Table)
```

```
##   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
## 6  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
## 6 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>
## 6 Speed Skating Women's 1,000 metres  <NA>
```

```
tail(athleteEventsData_Table)
```

```
##            ID                    Name Sex Age Height Weight     Team NOC
## 271111 135568 Olga Igorevna Zyuzkova   F  33    171     69  Belarus BLR
## 271112 135569             Andrzej ya   M  29    179     89 Poland-1 POL
## 271113 135570               Piotr ya   M  27    176     59   Poland POL
## 271114 135570               Piotr ya   M  27    176     59   Poland POL
## 271115 135571     Tomasz Ireneusz ya   M  30    185     96   Poland POL
## 271116 135571     Tomasz Ireneusz ya   M  34    185     96   Poland POL
##              Games Year Season           City       Sport
## 271111 2016 Summer 2016 Summer Rio de Janeiro  Basketball
## 271112 1976 Winter 1976 Winter      Innsbruck        Luge
## 271113 2014 Winter 2014 Winter          Sochi Ski Jumping
## 271114 2014 Winter 2014 Winter          Sochi Ski Jumping
## 271115 1998 Winter 1998 Winter         Nagano   Bobsleigh
## 271116 2002 Winter 2002 Winter Salt Lake City   Bobsleigh
##                                         Event Medal
## 271111            Basketball Women's Basketball  <NA>
## 271112               Luge Mixed (Men)'s Doubles  <NA>
## 271113 Ski Jumping Men's Large Hill, Individual  <NA>
## 271114       Ski Jumping Men's Large Hill, Team  <NA>
## 271115                     Bobsleigh Men's Four  <NA>
## 271116                     Bobsleigh Men's Four  <NA>
```

```
#Read the athlete events data using read.csv
athleteEventsData_Delim <- read.delim("athlete_events.csv", sep=",")
head(athleteEventsData_Delim)
```

```
##   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
## 6  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
## 6 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>
## 6 Speed Skating Women's 1,000 metres  <NA>
```

```
tail(athleteEventsData_Delim)
```

```
##            ID                    Name Sex Age Height Weight     Team NOC
## 271111 135568 Olga Igorevna Zyuzkova   F  33    171     69  Belarus BLR
## 271112 135569             Andrzej ya   M  29    179     89 Poland-1 POL
## 271113 135570               Piotr ya   M  27    176     59   Poland POL
## 271114 135570               Piotr ya   M  27    176     59   Poland POL
## 271115 135571     Tomasz Ireneusz ya   M  30    185     96   Poland POL
## 271116 135571     Tomasz Ireneusz ya   M  34    185     96   Poland POL
##              Games Year Season           City       Sport
## 271111 2016 Summer 2016 Summer Rio de Janeiro  Basketball
## 271112 1976 Winter 1976 Winter      Innsbruck        Luge
## 271113 2014 Winter 2014 Winter          Sochi Ski Jumping
## 271114 2014 Winter 2014 Winter          Sochi Ski Jumping
## 271115 1998 Winter 1998 Winter         Nagano   Bobsleigh
## 271116 2002 Winter 2002 Winter Salt Lake City   Bobsleigh
##                                        Event Medal
## 271111            Basketball Women's Basketball  <NA>
## 271112               Luge Mixed (Men)'s Doubles  <NA>
## 271113 Ski Jumping Men's Large Hill, Individual  <NA>
## 271114       Ski Jumping Men's Large Hill, Team  <NA>
## 271115                     Bobsleigh Men's Four  <NA>
## 271116                     Bobsleigh Men's Four  <NA>
```

(a) What are the differences in the import functions? Did you try different arguments (parameters) within the functions? What did you notice?

The data type varies for different import functions. The parameters also adds variation to the data. Using a 'header = TRUE' parameter reads the header in the data or else a new header say(V1,V2, etc) is added to the data. The parameter 'StringAsFactors = TRUE' converts all the data to factor type.

(b) Continue with the imported version you found most useful. Change any data types that don't seem correct. What (if any) did you change and why?

```
#Change the datatype of weight to integer
athleteEventsData_CSV$Weight <- as.integer(athleteEventsData_CSV$Weight)
```

As all the rows in the weight column are whole numbers, I found it more appropriate to convert it to integer type.

(c) Find the mean height and weight of all participants.

```
#Mean of height
heightMean <- mean(athleteEventsData_CSV$Height, na.rm = TRUE)
heightMean
```

```
## [1] 175.339
```

```
#Mean of Weight
weightMean <- mean(athleteEventsData_CSV$Weight, na.rm = TRUE)
weightMean
```

```
## [1] 70.69985
```

(d) Find the mean height and weight of all participants separated by season and sex.

```
#Mean height by Season
heightMeanBySeason <- athleteEventsData_CSV %>% group_by(Season) %>%
  dplyr::summarize(meanheight = mean(Height, na.rm = TRUE))

heightMeanBySeason
```

```
## # A tibble: 2 x 2
##   Season meanheight
##   <fct>       <dbl>
## 1 Summer       176.
## 2 Winter       175.
```

```
#Mean weight by Season
weightMeanBySeason <- athleteEventsData_CSV %>% group_by(Season) %>%
  dplyr::summarize(meanweight = mean(Weight, na.rm = TRUE))

weightMeanBySeason
```

```
## # A tibble: 2 x 2
##   Season meanweight
##   <fct>       <dbl>
## 1 Summer       70.7
## 2 Winter       70.8
```

```
#Mean height by Sex
heightMeanBySex <- athleteEventsData_CSV %>% group_by(Sex) %>%
  dplyr::summarize(meanheight = mean(Height, na.rm = TRUE))

heightMeanBySex
```

```
## # A tibble: 2 x 2
##   Sex   meanheight
##   <fct>      <dbl>
## 1 F           168.
## 2 M           179.
```

```
#Mean weight by Season
weightMeanBySex <- athleteEventsData_CSV %>% group_by(Sex) %>%
  dplyr::summarize(meanweight = mean(Weight, na.rm = TRUE))

weightMeanBySex
```

```
## # A tibble: 2 x 2
##   Sex   meanweight
##   <fct>      <dbl>
```

```
## 1 F           60.0
## 2 M           75.7
```

```
#Mean height by Sex
heightMeanBySexSeason <- athleteEventsData_CSV %>% group_by(Sex, Season) %>%
  dplyr::summarize(meanheight = mean(Height, na.rm = TRUE))

heightMeanBySexSeason
```

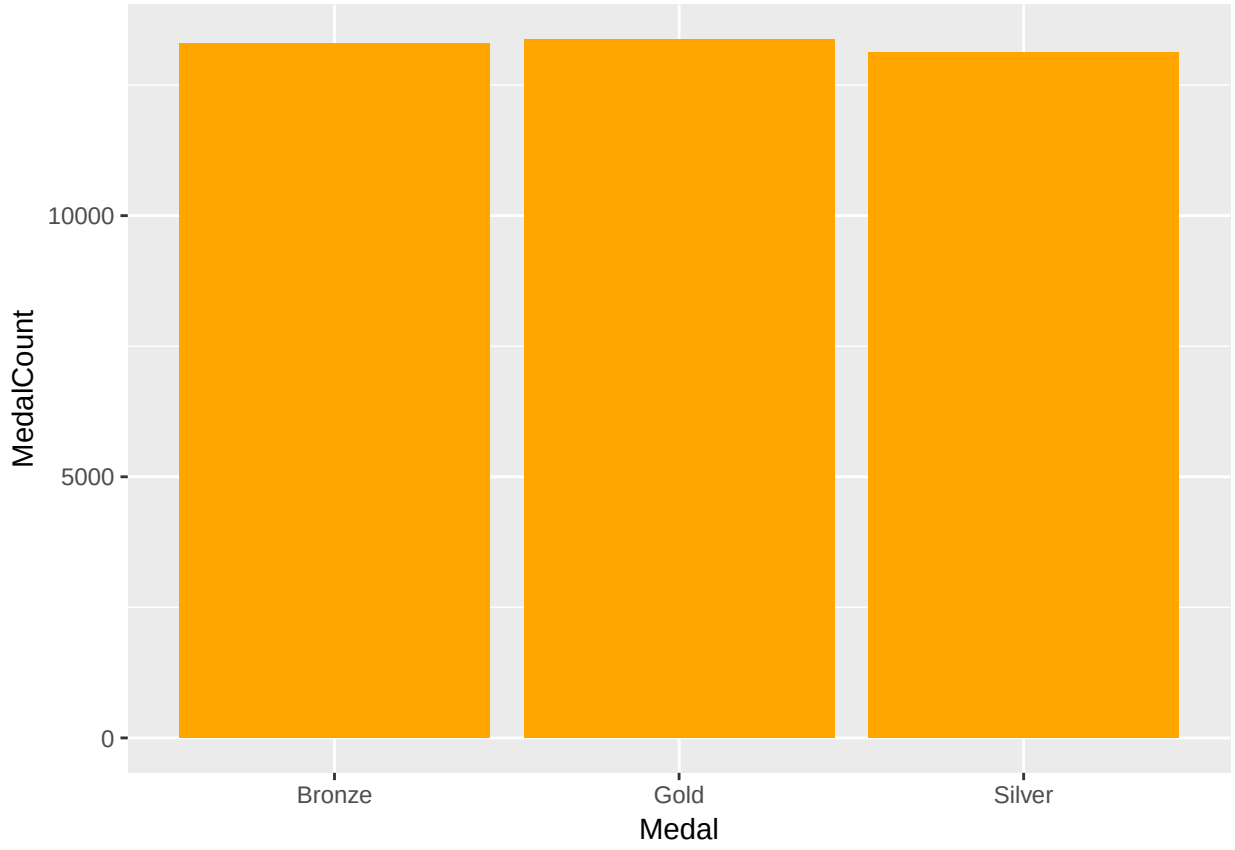
```
## # A tibble: 4 x 3
## # Groups:   Sex [?]
##   Sex   Season meanheight
##   <fct> <fct>       <dbl>
## 1 F     Summer       168.
## 2 F     Winter       167.
## 3 M     Summer       179.
## 4 M     Winter       179.
```

```
#Mean weight by Season
weightMeanBySexSeason <- athleteEventsData_CSV %>% group_by(Sex, Season) %>%
  dplyr::summarize(meanweight = mean(Weight, na.rm = TRUE))

weightMeanBySexSeason
```

```
## # A tibble: 4 x 3
## # Groups:   Sex [?]
##   Sex   Season meanweight
##   <fct> <fct>       <dbl>
## 1 F     Summer       60.1
## 2 F     Winter       59.8
## 3 M     Summer       75.6
## 4 M     Winter       76.4
```

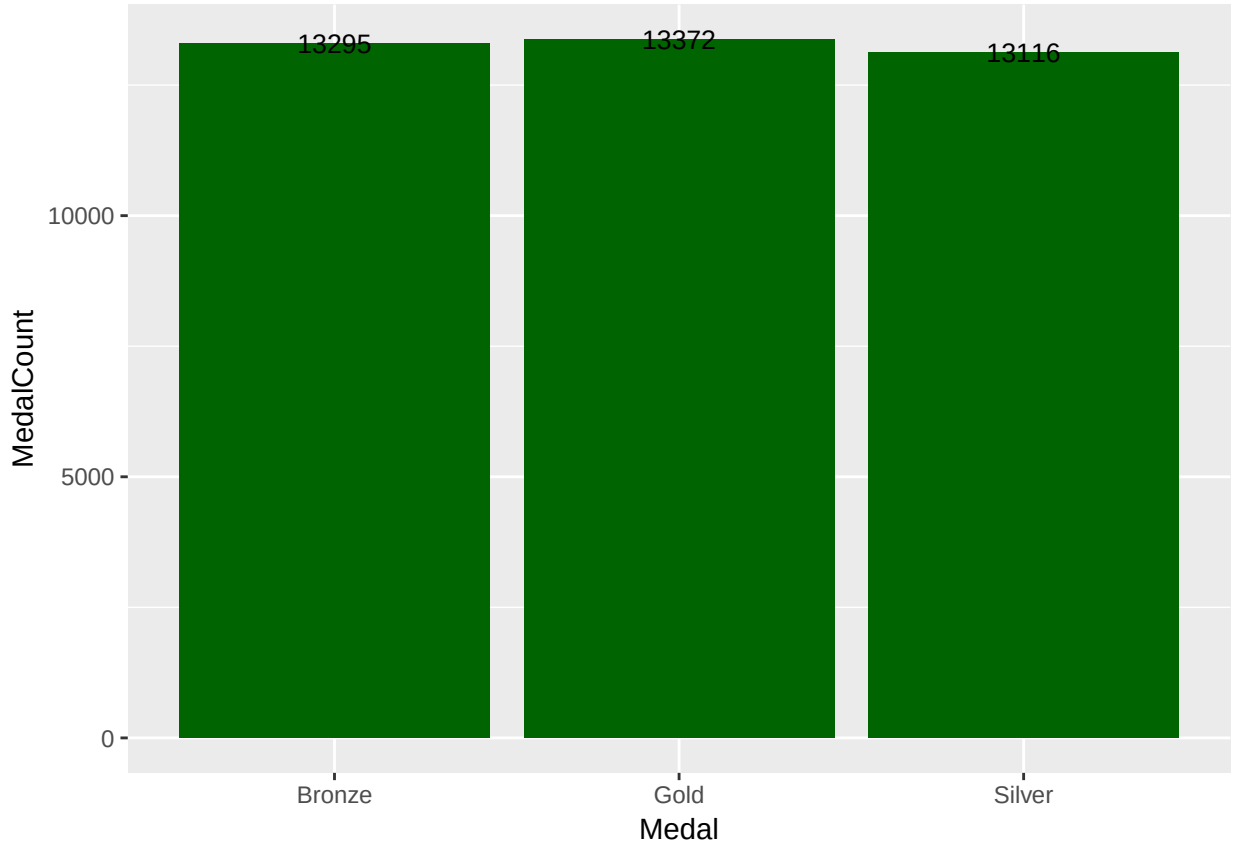
(d) Produce a histogram of medals won.

```
#Histogram of medals won
medalsWon <- athleteEventsData_CSV %>% group_by(Medal) %>% dplyr::summarize(MedalCount = n())
medalsWon <- na.omit(medalsWon)

medalsHist <- ggplot(data = medalsWon, aes(x = Medal, y = MedalCount)) +
  geom_bar(stat="identity", fill = "orange")
medalsHist
```



(e) The result may surprise you. Explore the count of medals.

```
#Histogram with count
medalsHistCount <- ggplot(data = medalsWon, aes(x = Medal, y = MedalCount)) +
  geom_bar(stat="identity", fill = "darkgreen") + geom_text(aes(label=MedalCount), color="black", size=3

medalsHistCount
```



(e) Count medals by age of participant.

```
#Count by age of participants
medalsWonByAge <- athleteEventsData_CSV %>% group_by(Medal, Age) %>% dplyr::summarize(MedalCount = n())
medalsWonByAge <- na.omit(medalsWonByAge)
medalsWonByAge
```

```
## # A tibble: 165 x 3
## # Groups:   Medal [3]
##    Medal    Age MedalCount
##    <fct>  <int>      <int>
##  1 Bronze    10          1
##  2 Bronze    12          3
##  3 Bronze    13          2
##  4 Bronze    14         18
##  5 Bronze    15         54
##  6 Bronze    16        105
##  7 Bronze    17        172
##  8 Bronze    18        286
##  9 Bronze    19        469
## 10 Bronze    20        692
## # ... with 155 more rows
```

Continue exploring the dataset with summary statistics and visualizations. Did you discover anything interesting that you would explore further if given time?

The data can be explored by Medal by city Medal by sport Medal by year Medal by Team etc.,