Assignment 1 - R/Git

msasnur@kent.edu

28/10/2019

```
# Data set has been downoaded from Kaggle.com
# (https://www.kaggle.com/heesoo37/120-years-of-olympic-history-athletes-and-results/downloads/120-years-of-olympic-history-athletes-and-results.zip/2)
# Data set is about 120 years of Olympic history on athletes
# Data set contains basic bio data of athletes from Athens 1896 to Rio 2016

library(readr)

#downloaded dataset (in .csv file) is assigned to athletes variable

athletes<-read_csv("athlete_events.csv",col_names = TRUE)
```

```
## Parsed with column specification:
## cols(
##   ID = col_double(),
##   Name = col_character(),
##   Sex = col_character(),
##   Age = col_double(),
##   Height = col_double(),
##   Weight = col_double(),
##   Team = col_character(),
##   NOC = col_character(),
##   Games = col_character(),
##   Year = col_double(),
##   Season = col_character(),
##   City = col_character(),
##   Sport = col_character(),
##   Event = col_character(),
##   Medal = col_character()
## )
```

```
View(athletes)

#Printing first 6 records
head(athletes,6)
```

```
## # A tibble: 6 x 15
##      ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##   <dbl> <chr> <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>
## 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
## # ... with 4 more variables: City <chr>, Sport <chr>, Event <chr>,
## #   Medal <chr>
```

```
#Descriptive Statistics performed on athletes Age
#To find out Mean Age
print(mean(athletes$Age, na.rm = TRUE))
```

```
## [1] 25.5569
```

```
#To find out Median Age
print(median(athletes$Age, na.rm = TRUE))
```

```
## [1] 24
```

```
#to find out range in Age by knowing max and min value
print(min(athletes$Age, na.rm = TRUE))
```

```
## [1] 10
```

```
print(max(athletes$Age, na.rm = TRUE))
```

```
## [1] 97
```

```
print(range(athletes$Age, na.rm = TRUE))
```

```
## [1] 10 97
```

```
#to find quantile range for Age COLumn
print(quantile(athletes$Age, na.rm = TRUE))
```

```
##   0%  25%  50%  75% 100%
##   10   21   24   28   97
```

```
print(IQR(athletes$Age, na.rm = TRUE))
```

```
## [1] 7
```

```
#to find variance and standard deviation for Age Column
print(var(athletes$Age, na.rm = TRUE),10)
```

```
## [1] 40.8776203
```

```
print(sd(athletes$Age, na.rm = TRUE),10)
```

```
## [1] 6.393560847
```

```
#Using factors and levels fuctions to find unique values in team column
teams<-factor(athletes$Team)
head(levels(teams),10)
```

```
##  [1] "30. Februar"           "A North American Team"
##  [3] "Acipactli"             "Acturus"
##  [5] "Afghanistan"           "Akatonbo"
##  [7] "Alain IV"              "Albania"
##  [9] "Alcaid"                "Alcyon-6"
```

```
head(athletes[athletes$Team == "India",],10)
```

```
## # A tibble: 10 x 15
##       ID Name  Sex     Age Height Weight Team  NOC   Games  Year Season
##    <dbl> <chr> <chr> <dbl>  <dbl>  <dbl> <chr> <chr> <chr> <dbl> <chr>
##  1   281 S. A~ M        NA     NA     NA India IND   1928~  1928 Summer
##  2   281 S. A~ M        NA     NA     NA India IND   1928~  1928 Summer
##  3   512 Shin~ F        19    167     53 India IND   1984~  1984 Summer
##  4   512 Shin~ F        19    167     53 India IND   1984~  1984 Summer
##  5   512 Shin~ F        23    167     53 India IND   1988~  1988 Summer
##  6   512 Shin~ F        23    167     53 India IND   1988~  1988 Summer
##  7   512 Shin~ F        27    167     53 India IND   1992~  1992 Summer
##  8   512 Shin~ F        31    167     53 India IND   1996~  1996 Summer
##  9   663 Shar~ M        22    186     85 India IND   2004~  2004 Summer
## 10   663 Shar~ M        26    186     85 India IND   2008~  2008 Summer
## # ... with 4 more variables: City <chr>, Sport <chr>, Event <chr>,
## #   Medal <chr>
```

```
head(unique(athletes$Name),10)
```

```
##  [1] "A Dijiang"
##  [2] "A Lamusi"
##  [3] "Gunnar Nielsen Aaby"
##  [4] "Edgar Lindenau Aabye"
##  [5] "Christine Jacoba Aaftink"
##  [6] "Per Knut Aaland"
##  [7] "John Aalberg"
##  [8] "Cornelia \"Cor\" Aalten (-Strannood)"
##  [9] "Antti Sami Aalto"
## [10] "Einar Ferdinand \"Einari\" Aalto"
```

```
#Applying transformation on Weights variable after assigning Weight column
data to it.
Weights<-athletes$Height
head(sqrt(Weights),10)
```

```
##  [1] 13.41641 13.03840       NA       NA 13.60147 13.60147 13.60147
##  [8] 13.60147 13.60147 13.60147
```

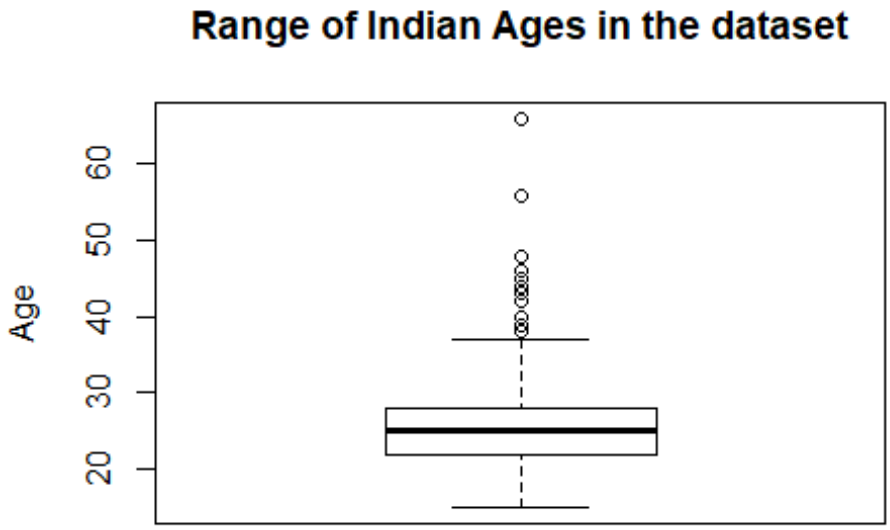
```
head(log(Weights),10)
```

```
##  [1] 5.192957 5.135798       NA       NA 5.220356 5.220356 5.220356
##  [8] 5.220356 5.220356 5.220356
```

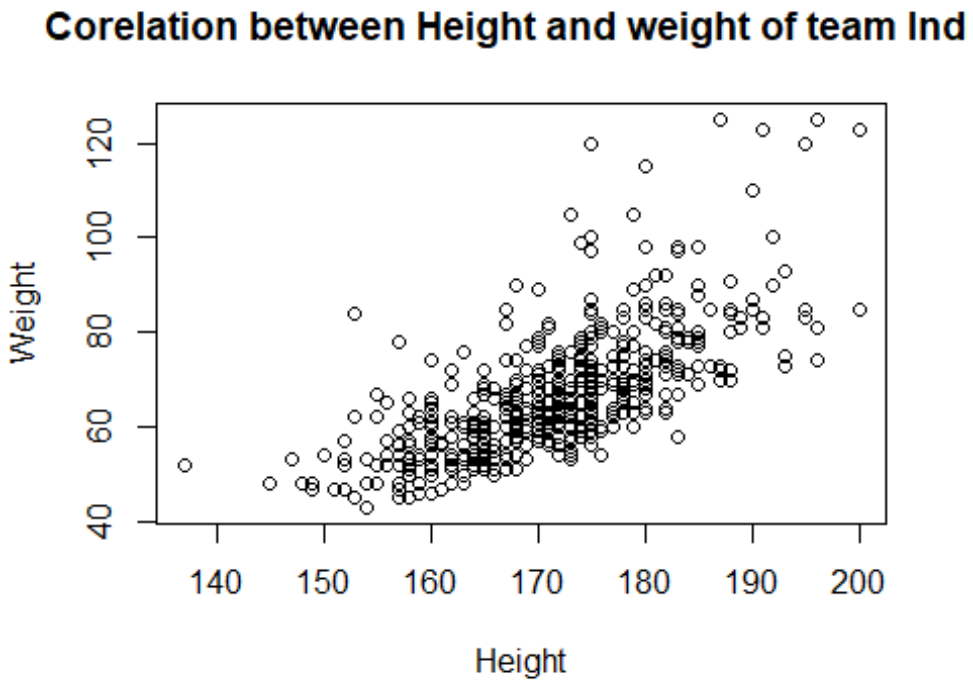
```
head(exp(Weights),10)
```

```
##  [1] 1.489384e+78 6.761794e+73           NA           NA 2.210442e+80
##  [6] 2.210442e+80 2.210442e+80 2.210442e+80 2.210442e+80 2.210442e+80
```

```r
#Using boxplot function to draw Box Plot Graph for Athletes Age in Team India
boxplot(athletes$Age[athletes$Team=="India"], ylab = "Age", main = "Range of
Indian Ages in the dataset")
```




```r
#Scatter Plot showing corelation between height and weight of all the
athletes in team India
x<-athletes$Height[athletes$Team=="India"]
y<-athletes$Weight[athletes$Team=="India"]
plot(x,y, xlab = "Height", ylab = "Weight", main="Corelation between Height
and weight of team India")
```




```
#Corelation observed in this scatter plot between height and weight shows
that as Height increase for an athlete, their weight also increases
```