R - In class assignment #2

Chris Montgomery

Jan 23, 2019

R Markdown

Complete these tasks: 1. Read in the file with all the baseball players height, weight and age info (BaseballHeightWeight.csv) as a dataframe. 2. View the data 3. Change the headers/column names appropriately. 4. Print the "head" of the data set. 5. Print the "tail" of the data set. 6. Find the statistics (summary) of the heights, weights, and ages of the players. Using complete sentences, give the reader these summary.

Make Charts: 7. make a boxplot of the weight of the players. Does it look normal? 8. make a histogram of the height of the players. Does the distribution looks normal? 9. Make a plot with weights vs heights of the players, color by Teams 10. Make a plot with weights vs heights of the players, color by age

Subsetting: 11. Obtain a subset of the data with only Team Washington. Using complete sentences, give the summary statistics on height of Team Washington. 12. Obtain another subset with Team Washington and only players older than 25. Again, give the summary of the statistics on height.

1. Read in the file with all the baseball players height, weight and age info

```
# This is coder's comments
df <- read.csv("BaseballHeightWeight.csv")
# baseballdf <- data.frame(read.csv("BaseballHeightWeight.csv"))
```

2. View the data

```
#View the dataframe. the command df prints the entire dataframe.
head(df)
```

```
##              Name Team       Position Height.inches. Weight.pounds.   Age
## 1   Adam_Donachie  BAL        Catcher             74            180 22.99
## 2       Paul_Bako  BAL        Catcher             74            215 34.69
## 3 Ramon_Hernandez  BAL        Catcher             72            210 30.78
## 4    Kevin_Millar  BAL  First_Baseman             72            210 35.43
## 5     Chris_Gomez  BAL  First_Baseman             73            188 35.71
## 6   Brian_Roberts  BAL Second_Baseman             69            176 29.39
```

3. Change the headers/column names appropriately.

```
#View column names of DF
colnames(df)
```

```
## [1] "Name"           "Team"           "Position"       "Height.inches."
## [5] "Weight.pounds." "Age"
```

```
#Create a vector of new, more appropriate column names
new.cols <- c("name", "team", "position", "height", "weight", "age")
```

```
#Replace old column names with vector of new column names, print to confirm
colnames(df) <- new.cols
head(df, n = 2)
```

```
##            name team position height weight   age
## 1 Adam_Donachie  BAL  Catcher     74    180 22.99
## 2     Paul_Bako  BAL  Catcher     74    215 34.69
```

4. Print the "head" of the data set.

```
#Print the first 10 observations of the dataset
head(df, n = 10)
```

```
##                name team       position height weight   age
## 1     Adam_Donachie  BAL        Catcher     74    180 22.99
## 2         Paul_Bako  BAL        Catcher     74    215 34.69
## 3   Ramon_Hernandez  BAL        Catcher     72    210 30.78
## 4      Kevin_Millar  BAL  First_Baseman     72    210 35.43
## 5       Chris_Gomez  BAL  First_Baseman     73    188 35.71
## 6     Brian_Roberts  BAL Second_Baseman     69    176 29.39
## 7     Miguel_Tejada  BAL      Shortstop     69    209 30.77
## 8       Melvin_Mora  BAL  Third_Baseman     71    200 35.07
## 9       Aubrey_Huff  BAL  Third_Baseman     76    231 30.19
## 10       Adam_Stern  BAL     Outfielder     71    180 27.05
```

5. Print the "tail" of the data set.

```
#Print last 6 observations of the dataframe

tail(df)
```

```
##                 name team       position height weight   age
## 1029    Josh_Hancock  STL Relief_Pitcher     75    205 28.89
## 1030   Brad_Thompson  STL Relief_Pitcher     73    190 25.08
## 1031   Tyler_Johnson  STL Relief_Pitcher     74    180 25.73
## 1032 Chris_Narveson  STL Relief_Pitcher     75    205 25.19
## 1033   Randy_Keisler  STL Relief_Pitcher     75    190 31.01
## 1034     Josh_Kinney  STL Relief_Pitcher     73    195 27.92
```

6. Find the statistics (summary) of the heights, weights, and ages of the players. Using complete sentences, give the reader these summary.

```
#summary stats of the dataframe columns
summary(df[c("height", "weight", "age")])
```

```
##      height         weight           age
##  Min.   :67.0   Min.   :150.0   Min.   :20.90
##  1st Qu.:72.0   1st Qu.:187.0   1st Qu.:25.44
##  Median :74.0   Median :200.0   Median :27.93
##  Mean   :73.7   Mean   :201.7   Mean   :28.74
##  3rd Qu.:75.0   3rd Qu.:215.0   3rd Qu.:31.23
##  Max.   :83.0   Max.   :290.0   Max.   :48.52  
##                 NA's   :1
```

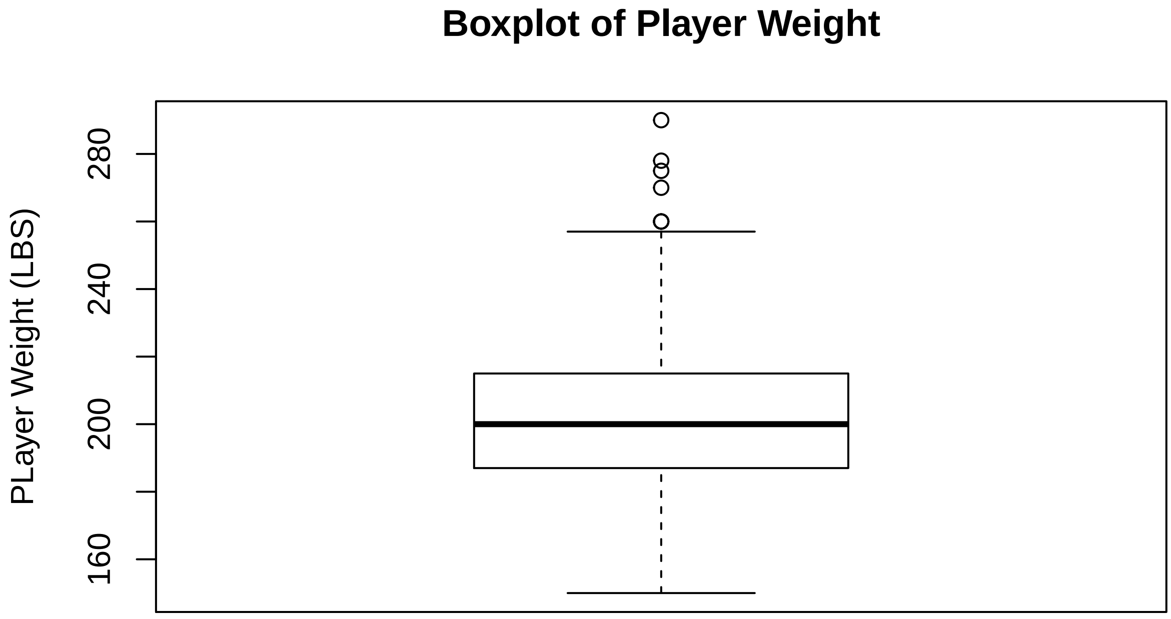
```
#find who's weight is not featured in the dataset.
df[is.na(df$weight),]
```

```
##              name team         position height weight   age
## 641 Kirk_Saarloos  CIN Starting_Pitcher     72     NA 27.77
```

The heights of players in the data frame ranged from a minimum of 67" (5' 7") to a max of 83" (6' 11"). The average (mean) height was nearly 6 feet 7 inches, with a median of 6 feet 7 inches. The interquartile range for height is seemingly small, with only 3 inches separating the first and third quartiles. In terms of weight, players in the dataframe ranged between 150 and 290 pounds. The mean and median weight were similar, both at approximately 200 pounds. The weight of one player, Kirk Saarloos, was not featured in the dataframe. Finally, players in the dataframe spanned ages 20.9 to 48.5, with an average age of 28.74. Like height and weight, age the spread between the median and mean age in the dataframe was relatively small. This could suggest the above variables are normally distributed, or at least not skewed.

7. make a boxplot of the weight of the players. Does it look normal?

```
#make a boxplot of the weight of the players. Does it look normal?
boxplot(df$weight, main = "Boxplot of Player Weight", ylab = "PLayer Weight (LBS)")
```



A simple boxplot of player weights suggest that several upper bound outliers exist in the data. Visual inspection show that at least 5 players have weights above the third quartile + 1.5 * IQR. Further inspection shows which players represent the outliers.

```
#Let's identify which players are upper bound outliers in weight
outlier <- 1.5 * (quantile(df$weight, .75, na.rm = TRUE) - quantile(df$weight, .25, na.rm = TRUE)) +  q

outlier
```

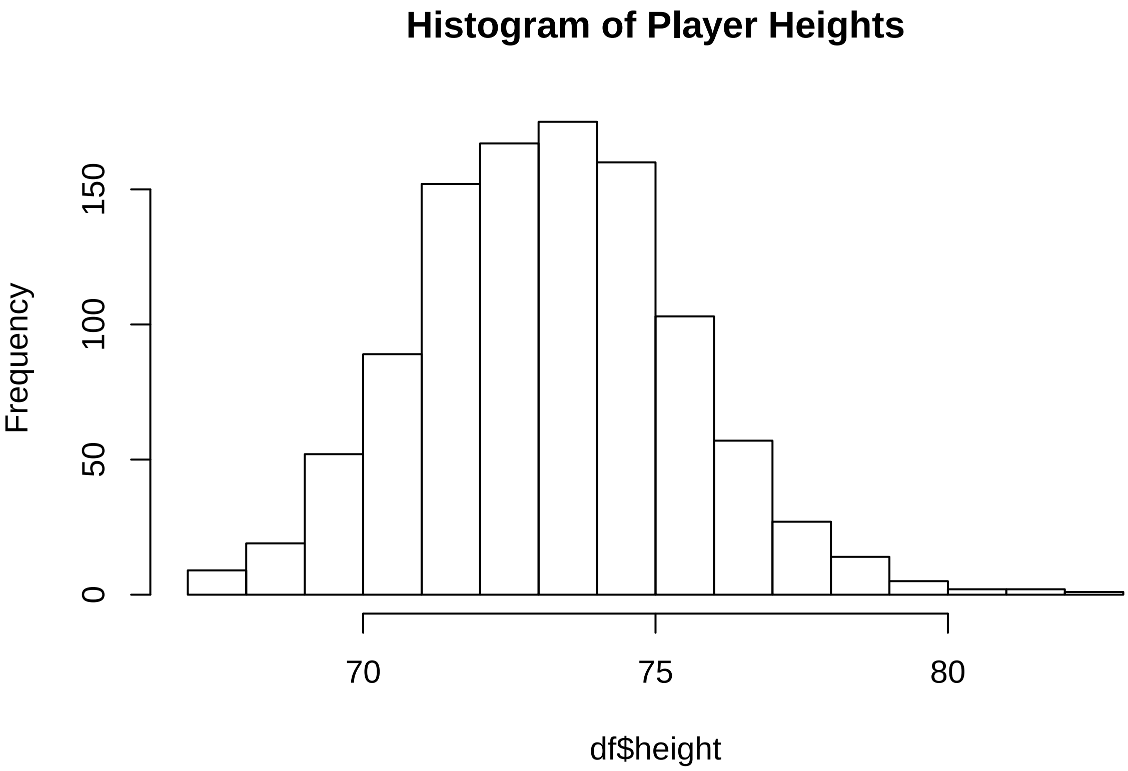
```
## 75% 
## 257
```

```
df[df$weight > outlier,]
```

```
##                name team          position height weight   age
## 63    Andrew_Sisco  CWS    Relief_Pitcher     81    260 24.13
## 65     Bobby_Jenks  CWS    Relief_Pitcher     75    270 25.96
## 160  C.C._Sabathia  CLE  Starting_Pitcher     79    290 26.61
## 237  Chris_Britton  NYY    Relief_Pitcher     75    278 24.21
## 431   Frank_Thomas  TOR Designated_Hitter     77    275 38.76
## 474    Boof_Bonser  MIN  Starting_Pitcher     76    260 25.38
## NA            <NA> <NA>              <NA>     NA     NA    NA
## 834 Prince_Fielder  MLW     First_Baseman     72    260 22.81
## 929      Jon_Rauch  WAS    Relief_Pitcher     83    260 28.42
```

8. make a histogram of the height of the players. Does the distribution looks normal?

```
#make a histogram of the height of the players. Does the distribution looks normal?
hist(df$height, main = "Histogram of Player Heights")
```



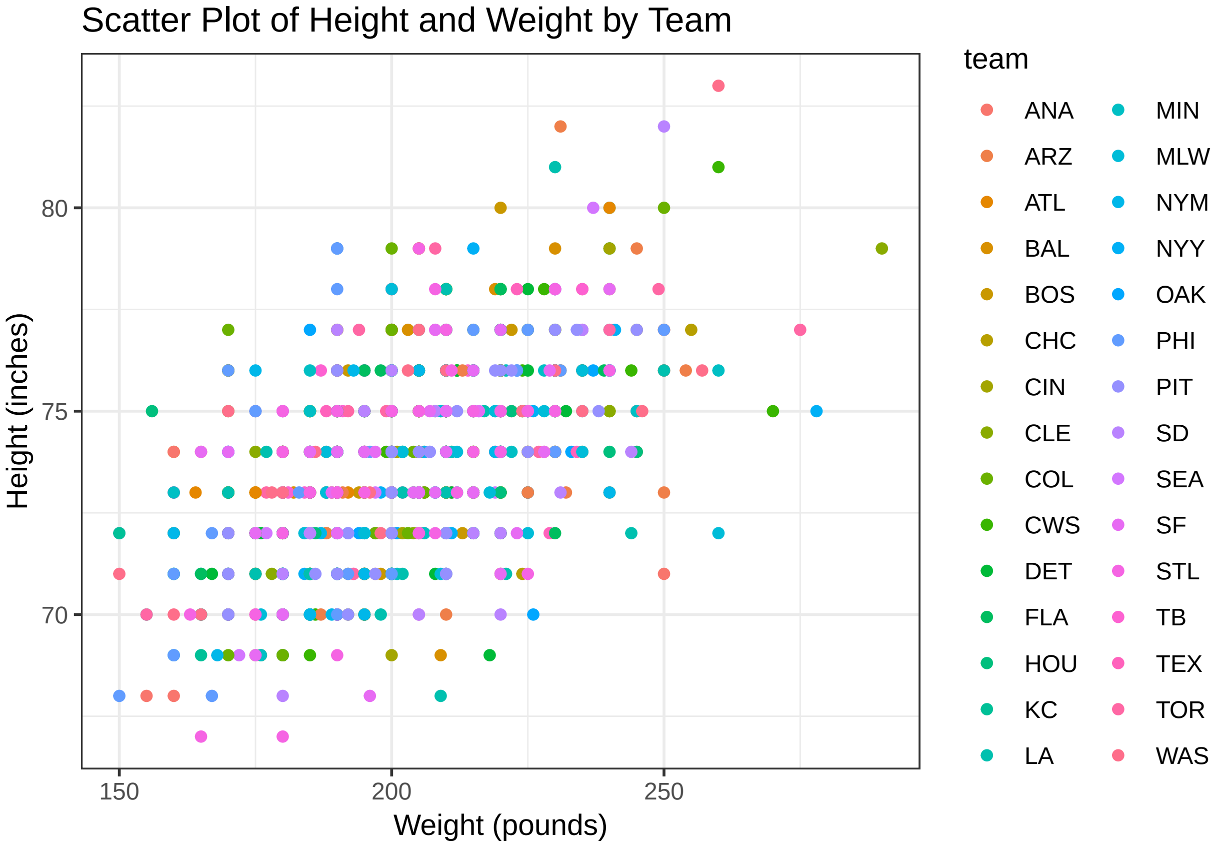
The above histogram of the player heights at first glance appears to be well approximated by a normal distribution. However, there does appear to be a few high values which could indicate rightward skew.

9. Make a plot with weights vs heights of the players, color by Teams

```
#Make a plot with weights vs heights of the players, color by Teams

library(ggplot2)
ggplot(df, aes(x = weight, y = height, color = team))+ theme_bw()+
  geom_point() + labs(title = "Scatter Plot of Height and Weight by Team")+
                    xlab ("Weight (pounds)") + ylab ("Height (inches)")
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



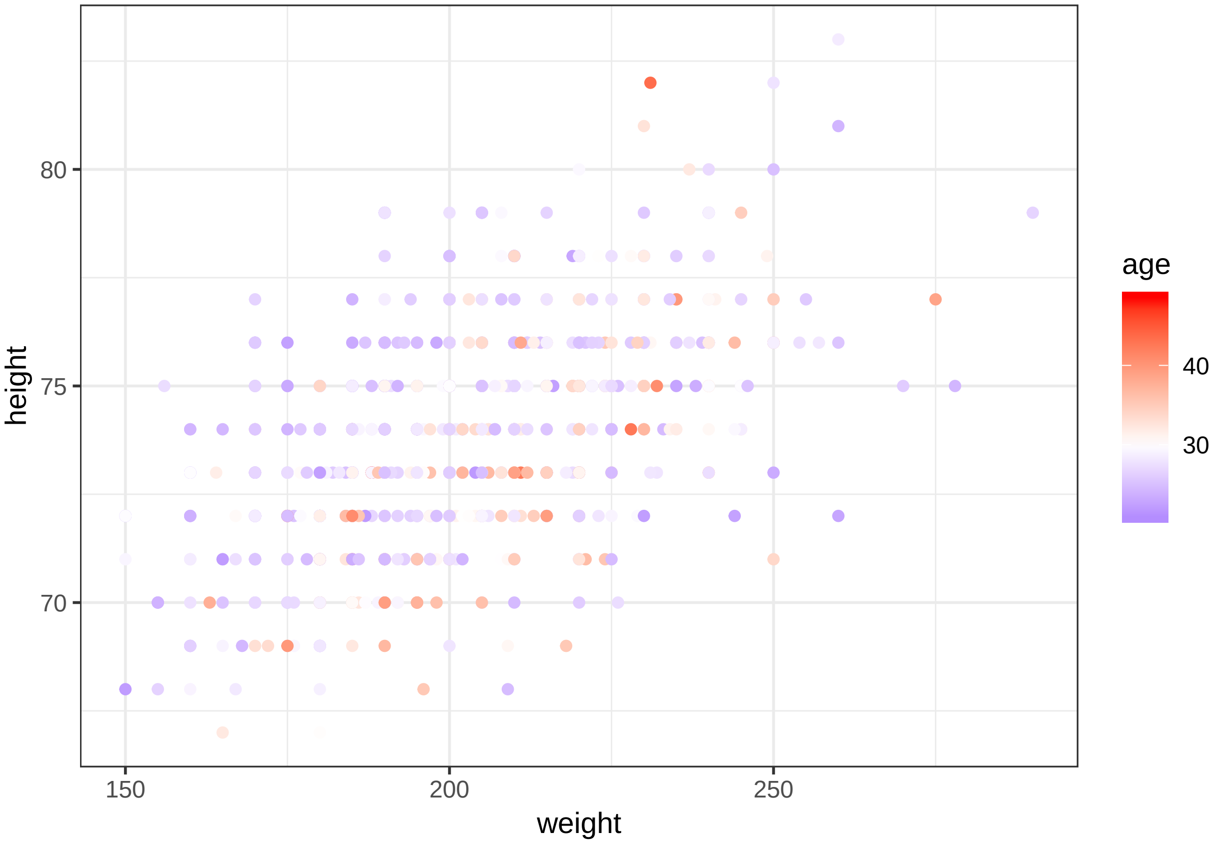
Unfortunately, I think segmenting the data by team introduces too much noise in the above plot. Further analysis could attempt to aggregate up to American vs. National league to determine whether different trends exist between the two leagues. Regardless, there appears to be a clear positive linear relationship between height and weight.

10. Make a plot with weights vs heights of the players, color by age

```
#10. Make a plot with weights vs heights of the players, color by age

ggplot(df, aes(x = weight, y = height, color = age)) + theme_bw()+
  geom_point()+ scale_colour_gradient2(midpoint=30, low="blue",
                    high="red" )
```

```
## Warning: Removed 1 rows containing missing values (geom_point).
```



At first glance, it may seam that older players could have a lower intercept in the above plot compared to younger players. Further analysis could include age interaction terms in any sort of modeling betweein height and weight.

11. Obtain a subset of the data with only Team Washington. Using complete sentences, give the summary statistics on height of Team Washington.

```
#Obtain a subset of the data with only Team Washington. Using complete sentences, give the summary stat

was <- subset(df, team == "WAS")
summary(was)
```

```
##              name          team                 position      height     
##  Alex_Escobar  : 1   WAS    :36   Relief_Pitcher  :14   Min.   :70.00  
##  Austin_Kearns : 1   ANA    : 0   Outfielder      : 7   1st Qu.:73.00  
##  Beltran_Perez : 1   ARZ    : 0   Starting_Pitcher: 6   Median :74.00  
##  Bernie_Castro : 1   ATL    : 0   Shortstop       : 3   Mean   :74.14  
##  Billy_Traber  : 1   BAL    : 0   Catcher         : 2   3rd Qu.:75.00  
##  Brett_Campbell: 1   BOS    : 0   First_Baseman   : 2   Max.   :83.00  
##  (Other)       :30   (Other): 0   (Other)         : 2                  
##      weight           age       
##  Min.   :150.0   Min.   :22.34  
##  1st Qu.:180.0   1st Qu.:25.36  
##  Median :199.0   Median :26.79  
##  Mean   :199.8   Mean   :26.94  
##  3rd Qu.:211.2   3rd Qu.:28.49  
##  Max.   :260.0   Max.   :32.30
##
```

The above summary describes information related to all of the Washington Nationals players contained in the dataset. In total, there are 20 pitchers, 7 outfielders, and 9 position players, totaling 36 players. The team appears young by MLB standards, with a mean age of approximately 27 and a max age of 32.3. Compared to the MLB summary table, the nationals appear to be quite average in terms of height and weight, with averages of 74.14 inches and 199.8 pounds respectively. The height range of Nationals players appears to be pretty small, with a minimum height of 70 inches and maximum of 83. The interquartile range was quite compact, with only 2 inches separating the first and third quartiles.

12. Obtain another subset with Team Washington and only players older than 25. Again, give the summary of the statistics on height.

```
# Obtain another subset with Team Washington and only players older than 25. Again, give the summary of

was.older <- was[was$age > 25,]

summary(was.older[,c("height")])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   70.00   73.00   74.00   74.13   75.00   83.00
```

The older players on the Nationals have a wide range of heights, ranging from 67 inchess to 83 inches. The median and mean height for Washington players over 25 were similar, at approximately 74 inches. The interquartile range was likewise compact, with only three inches separating the 1st and 3rd quartiles.