Laboratory Exercise Week 3

Brian Tipton STAT-380-001

9/5/23

Directions:

- Write your R code inside the code chunks after each question.
- Write your answer comments after the `#` sign.
- To generate the word document output, click the button `Knit` and wait for the word document to appear.
- RStudio will prompt you (only once) to install the `knitr` package.
- Submit your completed laboratory exercise using Blackboard's Turnitin feature. Your Turnitin upload link is found on your Blackboard Course shell under the Laboratory folder.

For this exercise, you will need to use the package `mosaic` to find numerical and graphical summaries.

```
# install mosaic package if necessary
if (!require(mosaic)) install.packages(`mosaic`)
# load the package in R
library(mosaic) # load the package mosaic to use its functions
```

My Custom functions used from my local lab projects .Rprofile

```
source("../../.Rprofile", chdir = TRUE)
```

```
catXWithString
```

```
## function (string, x, nl = TRUE, sep = " ")
## {
##     if (nl) {
##         cat(paste(string, toString(x), "\n", sep = sep))
##     }
##     else {
##         cat(paste(string, toString(x), sep = sep))
##     }
## }
```

1. Recall the `iris` data set from last week's exercise. The `iris` data set is already pre-loaded in R - look at the help file using `?iris` for more information on this data set.
   i) Check the structure of the data using the function `str(iris)`.

   ii) Find the average (or mean) measurement of the variable `Sepal.Length`. Do this in two ways as described in the lesson.

   iii) Find the average `Sepal.Length` for the different flower `Species`. Give a brief comment on the averages.

iv) Repeat (ii) and (iii) but use the summary standard deviation `sd()` which describes the spread of the variable.

v) Describe the shape of the variable `Sepal.Length` by creating a histogram using `histogram()`. Write your description outside the code chunk.

vi) Compare the `Sepal.Length` of the three species of flowers by creating a side-by-side boxplot using `bwplot()`. Write your description outside the code chunk.

Code chunk

```
str(iris)
```

```
## 'data.frame':    150 obs. of  5 variables:
##  $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
##  $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
##  $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
##  $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
##  $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
```

```
mean(iris$Sepal.Length) |> catXWithString(string = "Mean of Sepal.Length using 'mean(iris$Sepal.Length)
```

```
## Mean of Sepal.Length using 'mean(iris$Sepal.Length)':  5.84333333333333
```

```
mean(~ Sepal.Length , data = iris) |> catXWithString(string = "Mean of Sepal.Length using 'mean(~ Sepal
```

```
## Mean of Sepal.Length using 'mean(~ Sepal.Length , data = iris)':  5.84333333333333
```

```
mean(Sepal.Length ~ Species, data= iris) |> catXWithString(string = "Mean of all species")
```

```
## Mean of all species 5.006, 5.936, 6.588
```

```
print("The averages seem to be lining up close to the middle ground of these 3 values")
```

```
## [1] "The averages seem to be lining up close to the middle ground of these 3 values"
```

```
sd(~ Sepal.Length, data = iris) |> catXWithString(string = "Standard Deviation: ")
```

```
## Standard Deviation:  0.828066127977863
```

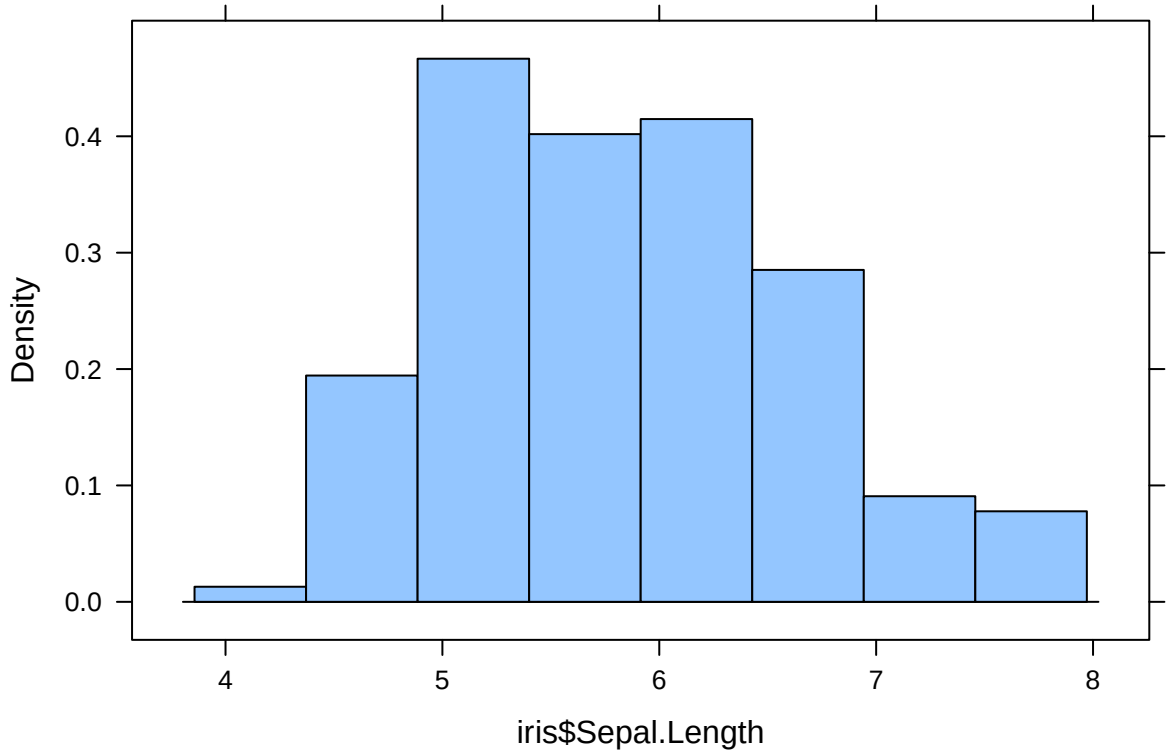
```
sd(Sepal.Length ~ Species, data = iris) |> catXWithString(string = "Standard Deviation all species: ")
```

```
## Standard Deviation all species:  0.352489687213451, 0.516171147063863, 0.635879593274432
```

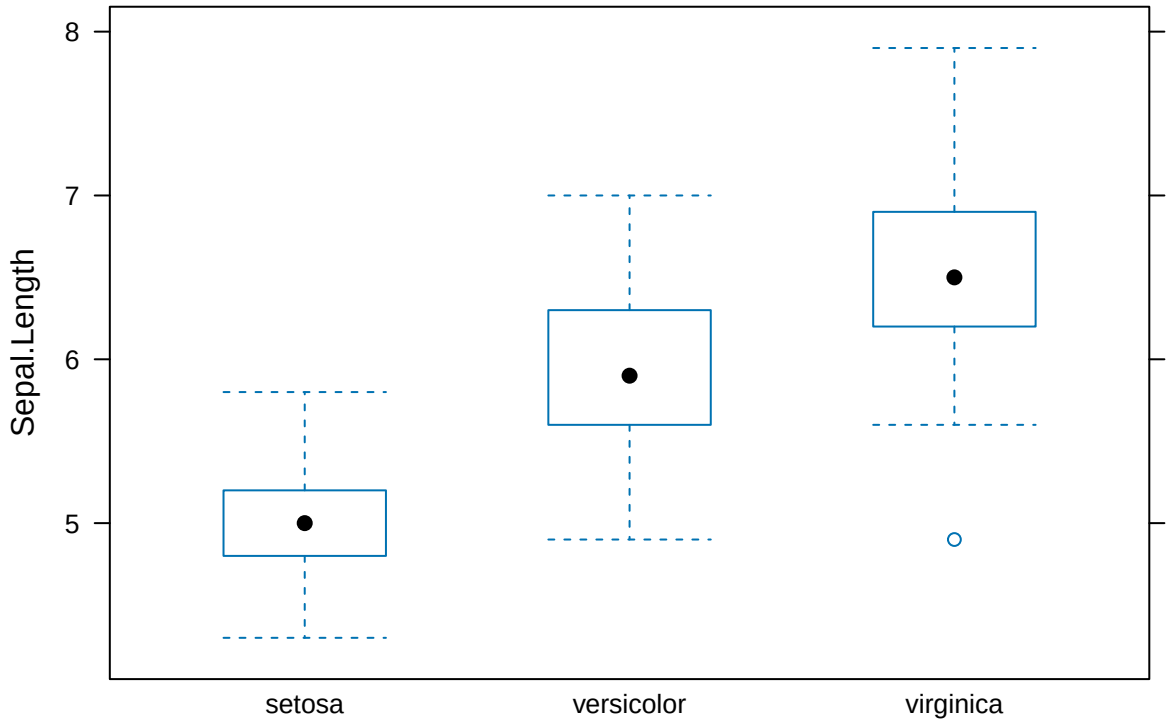
```
print("The standard deviation has more variance in it from species to species")
```

```
## [1] "The standard deviation has more variance in it from species to species"
```

```
histogram(iris$Sepal.Length)
```




```
bwplot(Sepal.Length ~ Species, data = iris)
```



Note

- The Sepal.Length has the highest amount of density roughly from 5-6

2. The data set `MLB-TeamBatting-S16.csv` contains MLB Team Batting Data for selected variables.

Load the data set from the given url using the code below. This data set was obtained from Baseball Reference.

- Tm - Team
- Lg - League: American League (AL), National League (NL)
- BatAge - Batters' average age
- RPG - Runs Scored Per Game
- G - Games Played or Pitched
- AB - At Bats
- R - Runs Scored/Allowed
- H - Hits/Hits Allowed
- HR - Home Runs Hit/Allowed
- RBI - Runs Batted In
- SO - Strikeouts
- BA - Hits/At Bats
- SH - Sacrifice Hits (Sacrifice Bunts)
- SF - Sacrifice Flies

i) Find the average measurement for the following variables `BatAge`, `RPG`, `R`, `H` and `BA`.

ii) Create dotplot's or histogram's for each variable in (i).

iii) Using your own words, describe the distribution of each variable in (i). Write your answer outside the code chunk.

iv) Find the average and the standard deviation of the variables `RPG`, `H` and `BA` for each league.

v) Describe any differences or similarities between the leagues. Write your comment outside the code chunk.

Code chunk

```
# load the data set
mlb16.data <- read.csv("https://raw.githubusercontent.com/jpailden/rstatlab/master/data/MLB-TeamBatting
str(mlb16.data) # check structure
```

```
## 'data.frame':    30 obs. of  14 variables:
##  $ Tm    : chr  "ARI" "ATL" "BAL" "BOS" ...
##  $ Lg    : chr  "NL" "NL" "AL" "AL" ...
##  $ BatAge: num  26.7 28.9 28.4 28.5 27.4 28.3 27.8 28.9 27.8 29.8 ...
##  $ RPG   : num  4.64 4.03 4.59 5.42 4.99 4.23 4.42 4.83 5.22 4.66 ...
##  $ G     : int  162 161 162 162 162 162 162 161 162 161 ...
```

```
##  $ AB    : int  5665 5514 5524 5670 5503 5550 5487 5484 5614 5526 ...
##  $ R     : int  752 649 744 878 808 686 716 777 845 750 ...
##  $ H     : int  1479 1404 1413 1598 1409 1428 1403 1435 1544 1476 ...
##  $ HR    : int  190 122 253 208 199 168 164 185 204 211 ...
##  $ RBI   : int  709 615 710 836 767 656 678 733 805 719 ...
##  $ SO    : int  1427 1240 1324 1160 1339 1285 1284 1246 1330 1303 ...
##  $ BA    : num  0.261 0.255 0.256 0.282 0.256 0.257 0.256 0.262 0.275 0.267 ...
##  $ SH    : int  43 64 17 8 42 29 58 31 54 17 ...
##  $ SF    : int  38 52 36 40 37 44 44 60 34 38 ...
```

```
head(mlb16.data)  # show first six rows
```

```
##    Tm Lg BatAge  RPG   G   AB   R    H  HR RBI   SO    BA SH SF
## 1 ARI NL   26.7 4.64 162 5665 752 1479 190 709 1427 0.261 43 38
## 2 ATL NL   28.9 4.03 161 5514 649 1404 122 615 1240 0.255 64 52
## 3 BAL AL   28.4 4.59 162 5524 744 1413 253 710 1324 0.256 17 36
## 4 BOS AL   28.5 5.42 162 5670 878 1598 208 836 1160 0.282  8 40
## 5 CHC NL   27.4 4.99 162 5503 808 1409 199 767 1339 0.256 42 37
## 6 CHW AL   28.3 4.23 162 5550 686 1428 168 656 1285 0.257 29 44
```

```
mean_BatAge <- mean(mlb16.data$BatAge, na.rm = TRUE)
mean_RPG <- mean(mlb16.data$RPG, na.rm = TRUE)
mean_R <- mean(mlb16.data$R, na.rm = TRUE)
mean_H <- mean(mlb16.data$H, na.rm = TRUE)
mean_BA <- mean(mlb16.data$BA, na.rm = TRUE)

# Display the mean values
mean_BatAge
```

```
## [1] 28.43
```

```
mean_RPG
```

```
## [1] 4.478333
```

```
mean_R
```

```
## [1] 724.8
```

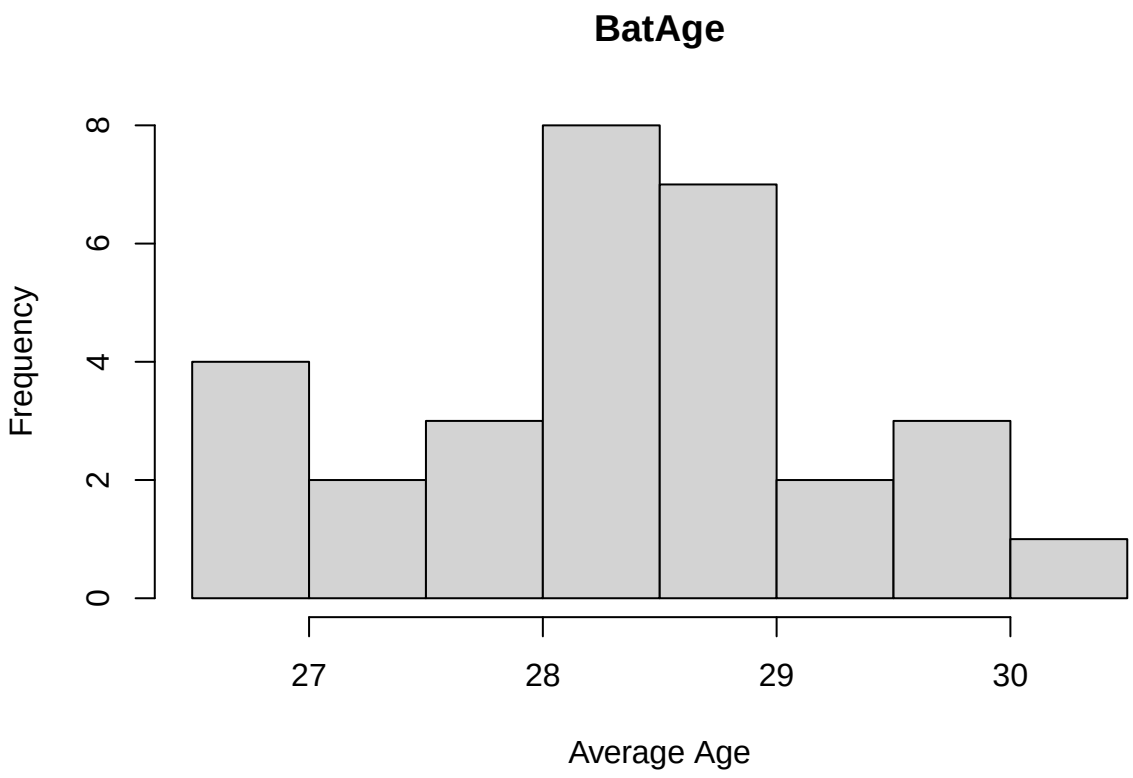
```
mean_H
```

```
## [1] 1409.2
```

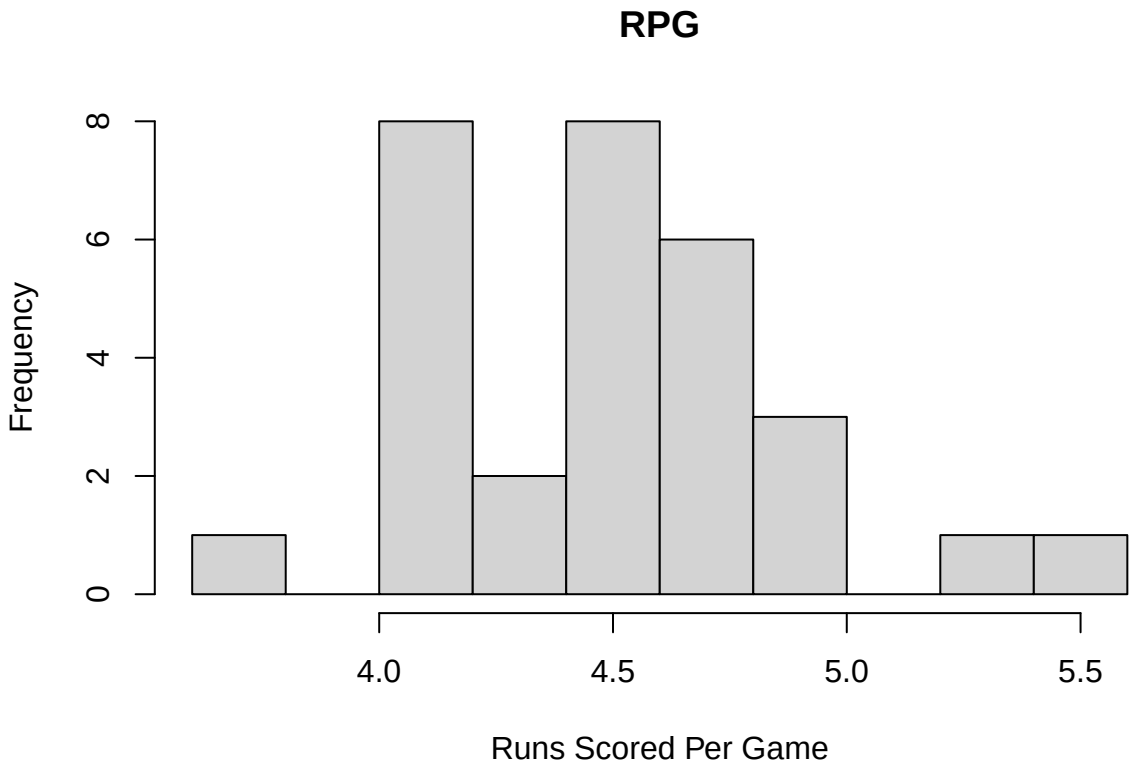
```
mean_BA
```

```
## [1] 0.2553
```

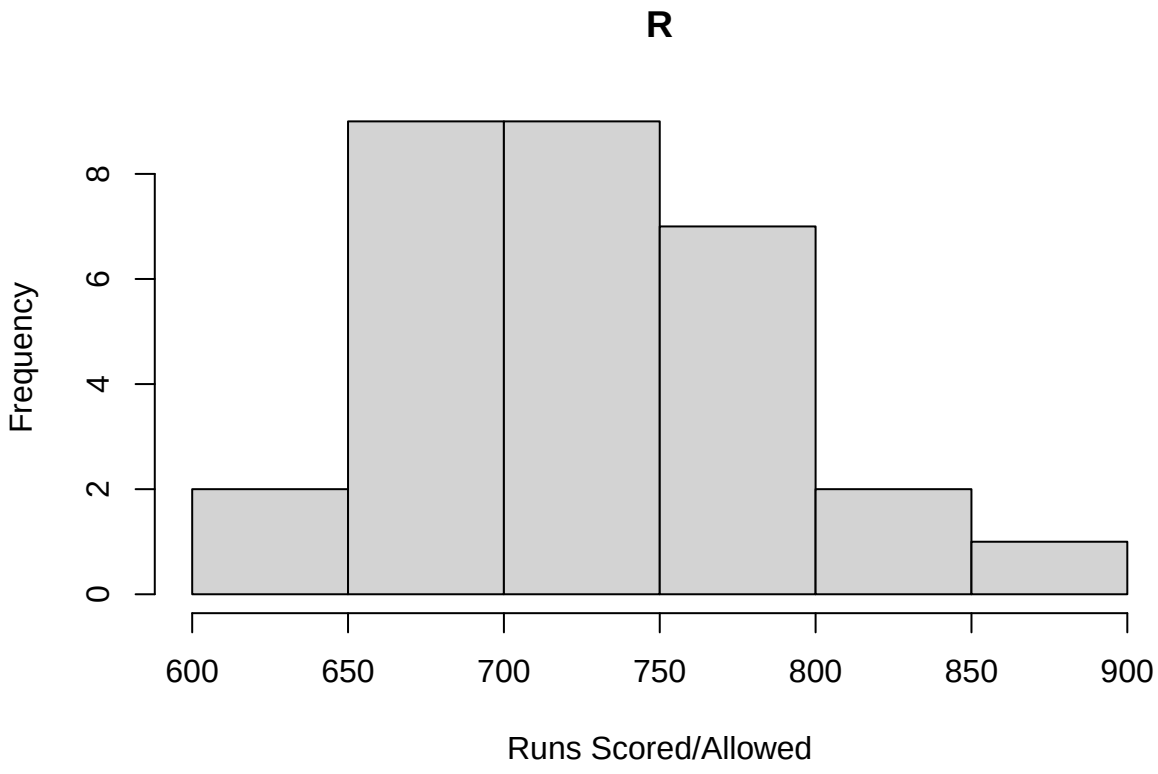
```
hist(mlb16.data$BatAge, main="BatAge", xlab="Average Age")
```



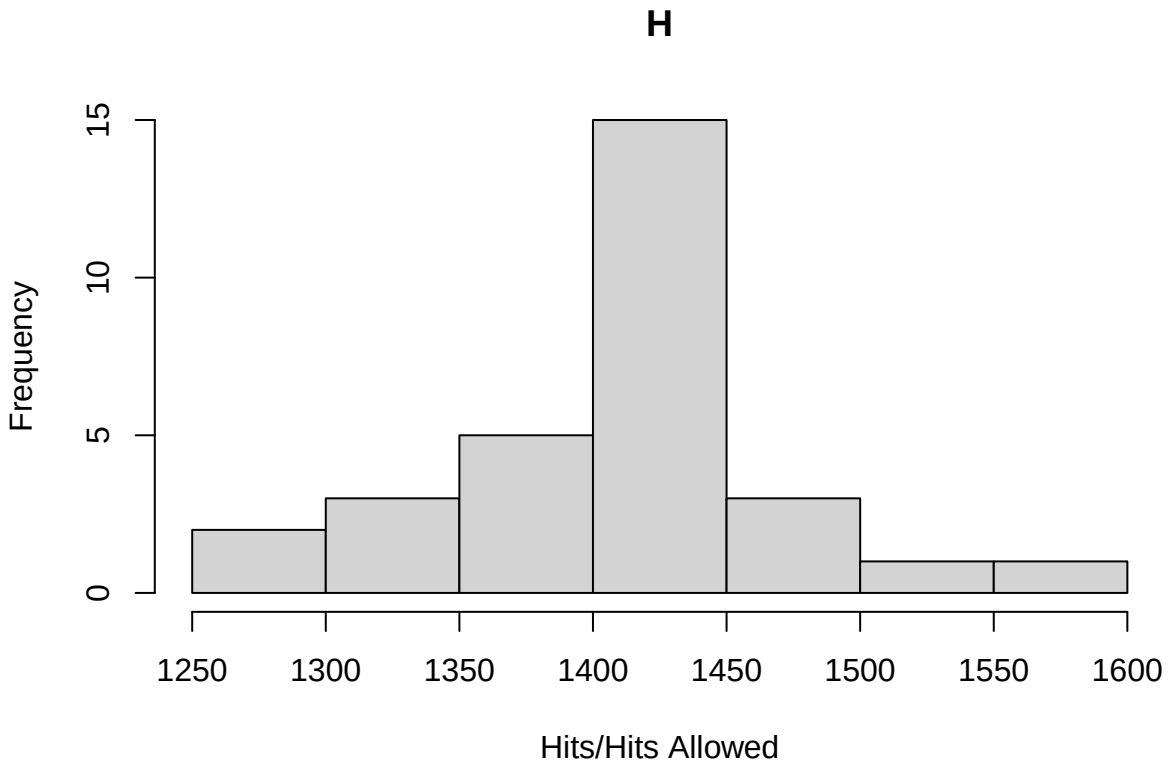

```
hist(mlb16.data$RPG, main="RPG", xlab="Runs Scored Per Game")
```



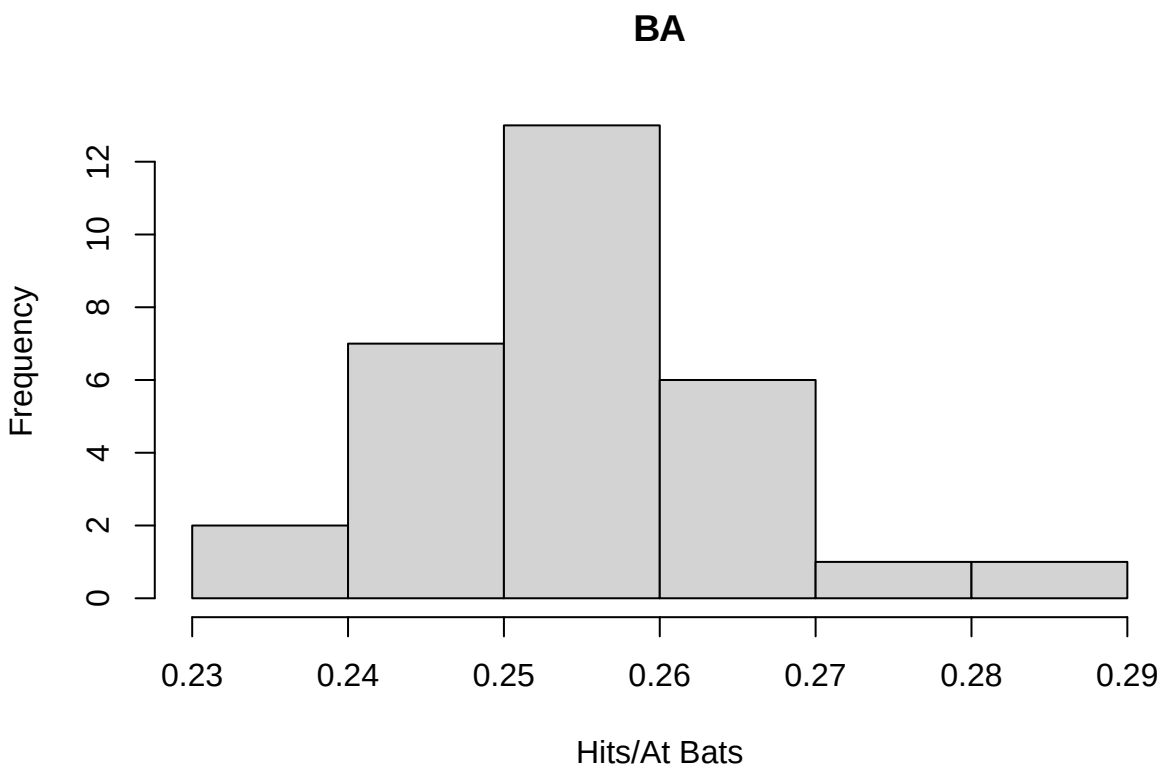

```
hist(mlb16.data$R, main="R", xlab="Runs Scored/Allowed")
```




```
hist(mlb16.data$H, main="H", xlab="Hits/Hits Allowed")
```




```
hist(mlb16.data$BA, main="BA", xlab="Hits/At Bats")
```




```
mean(mlb16.data$RPG ~ Lg ,data = mlb16.data)
```

```
##       AL       NL
## 4.519333 4.437333
```

```
sd(mlb16.data$RPG ~ Lg ,data = mlb16.data)
```

```
##        AL        NL
## 0.3533607 0.3914345
```

```
mean(mlb16.data$H ~Lg,data = mlb16.data)
```

```
##       AL       NL
## 1419.933 1398.467
```

```
sd(mlb16.data$H ~Lg,data = mlb16.data)
```

```
##       AL       NL
## 64.49858 71.57301
```

```
mean(mlb16.data$BA ~Lg,data = mlb16.data)
```

```
##        AL        NL
## 0.2568667 0.2537333
```

```
sd(mlb16.data$BA ~Lg,data = mlb16.data)
```

```
##          AL          NL
## 0.009869626 0.009837731
```

Both of the leagues were all similar in the standard deviation and mean for both. Although there was some discrepancies