HW 1 solutions

Stat 371

For this lab, we will be using the `mlb11` data set, as per Lab 1. Please submit a completed HTML file, due at the beginning of class.

Question 1

Write the code do upload the data into `R` and show the first 6 rows of the data set.

```
download.file("http://www.openintro.org/stat/data/mlb11.RData", destfile = "mlb11.RData")
load("mlb11.RData")
head(mlb11)
```

```
##                 team runs at_bats hits homeruns bat_avg strikeouts
## 1       Texas Rangers  855    5659 1599      210   0.283        930
## 2      Boston Red Sox  875    5710 1600      203   0.280       1108
## 3      Detroit Tigers  787    5563 1540      169   0.277       1143
## 4  Kansas City Royals  730    5672 1560      129   0.275       1006
## 5 St. Louis Cardinals  762    5532 1513      162   0.273        978
## 6       New York Mets  718    5600 1477      108   0.264       1085
##   stolen_bases wins new_onbase new_slug new_obs
## 1          143   96      0.340    0.460   0.800
## 2          102   90      0.349    0.461   0.810
## 3           49   95      0.340    0.434   0.773
## 4          153   71      0.329    0.415   0.744
## 5           57   90      0.341    0.425   0.766
## 6          130   77      0.335    0.391   0.725
```

Question 2

What element of the `mlb11` data set corresponds to the 10th row and 6th column? What team and variable is referred to in this cell?

```
mlb11[10,6]
```

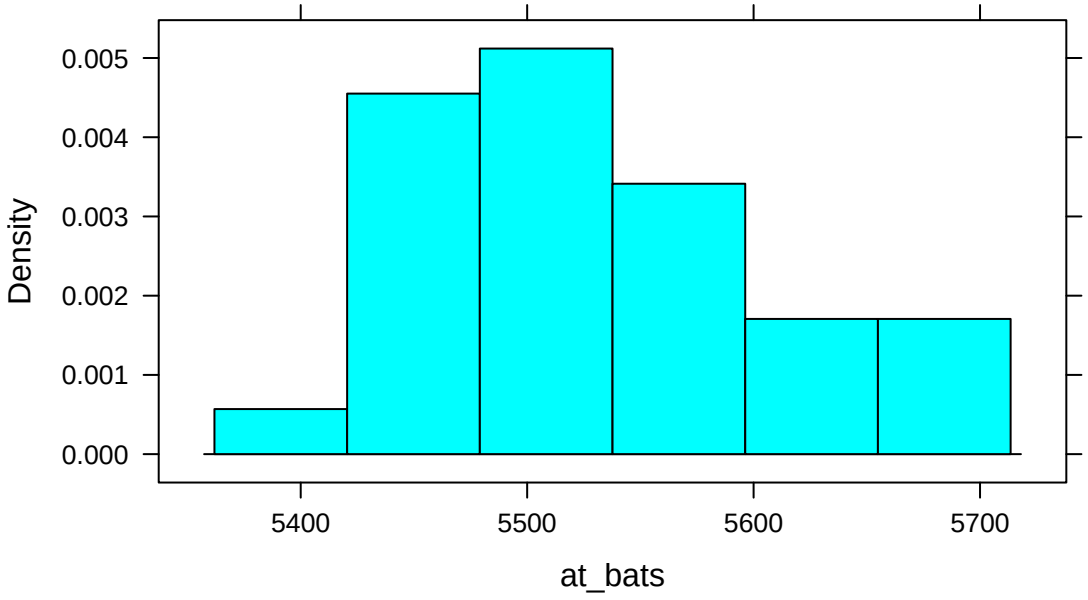
```
## [1] 0.258
```

This number, 0.258, represents the Houston Astros batting average.

Question 3

Describe the distribution of on at bats (`at_bats`) by considering its center, shape, and spread. Back up your claims by identifying the mean, median, standard deviation and interquartile range of this variable.

```
histogram(~at_bats, data = mlb11)
```




```
favstats(~at_bats, data = mlb11)
```

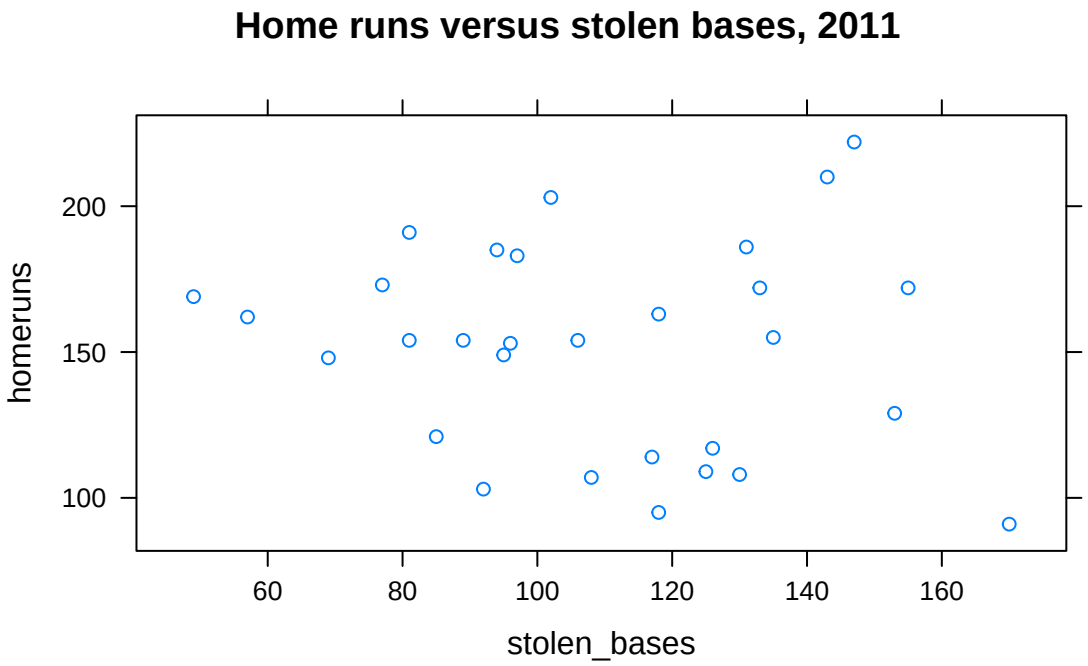
```
##   min      Q1 median   Q3  max   mean       sd  n missing
##  5417 5448.25 5515.5 5575 5710 5523.5 79.87307 30       0
```

The distribution of at bats is centered at roughly 5500 (mean 5523, median 5515), and is unimodal with perhaps a slight skew right. The standard deviation is 79.87 and the IQR is 126.75.

Question 4

A coach is curious if stealing more bases is linked with more home runs. Make and describe a scatter plot of team home runs versus stolen bases. Then, add a title to your plot.

```
xyplot(homeruns ~ stolen_bases, data = mlb11, main = "Home runs versus stolen bases, 2011")
```



There doesn't appear to be any link between home runs and stolen bases. Most of the plot looks like uneven scatter.

Question 5.

Create a new variable, `AboveAveWins`, representing whether or not the team won more than half of its games (there are 162 games in a season). How many teams won 82 games or more?

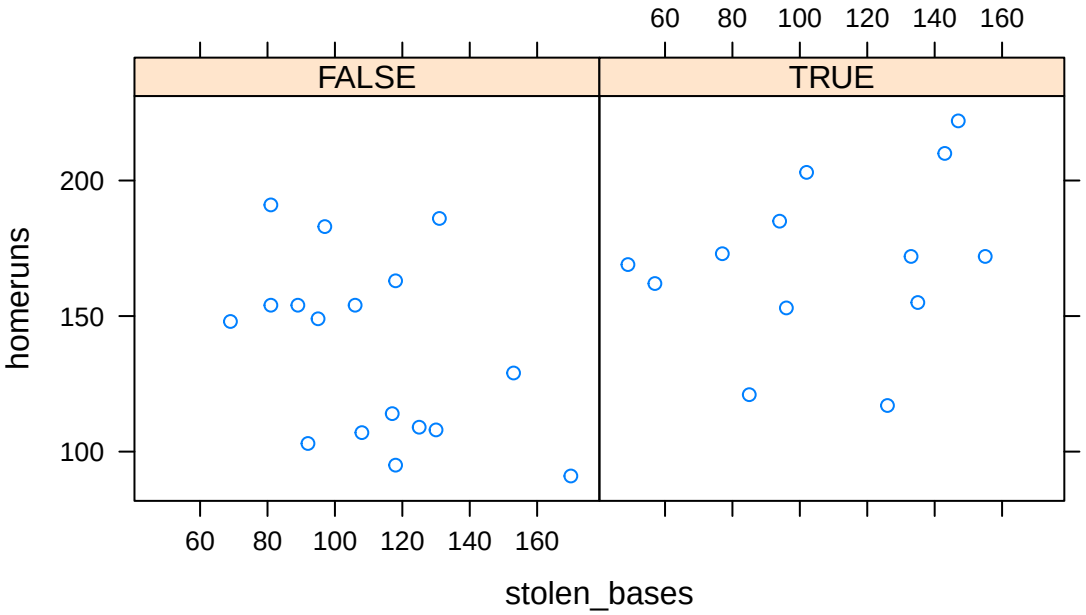
```
mlb11 <- mutate(mlb11, AboveAveWins = wins >=82)
table(mlb11$AboveAveWins)
```

```
##
## FALSE  TRUE
##    17    13
```

Question 6.

You can create separate scatter plots within groups using the code below.

```
xyplot(homeruns ~ stolen_bases | AboveAveWins, data = mlb11)
```



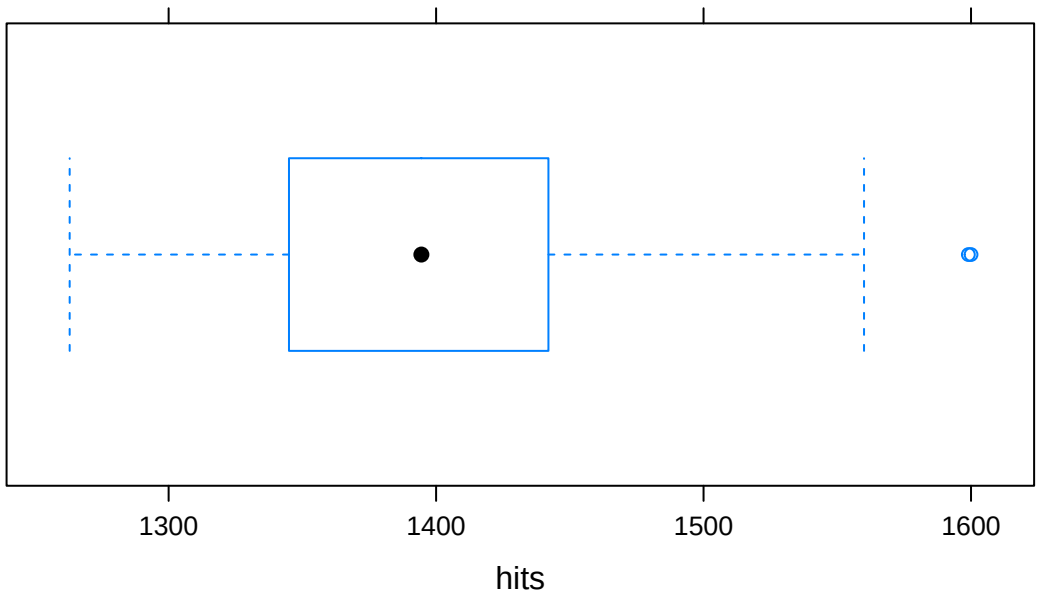
Does the association between home runs and stolen bases differ within the groups of `AboveAveWins`?

Possibly. In the `False` panel, there appears to be a negative link between stolen bases and home runs, while there appears to be a positive association between stolen bases and home runs.

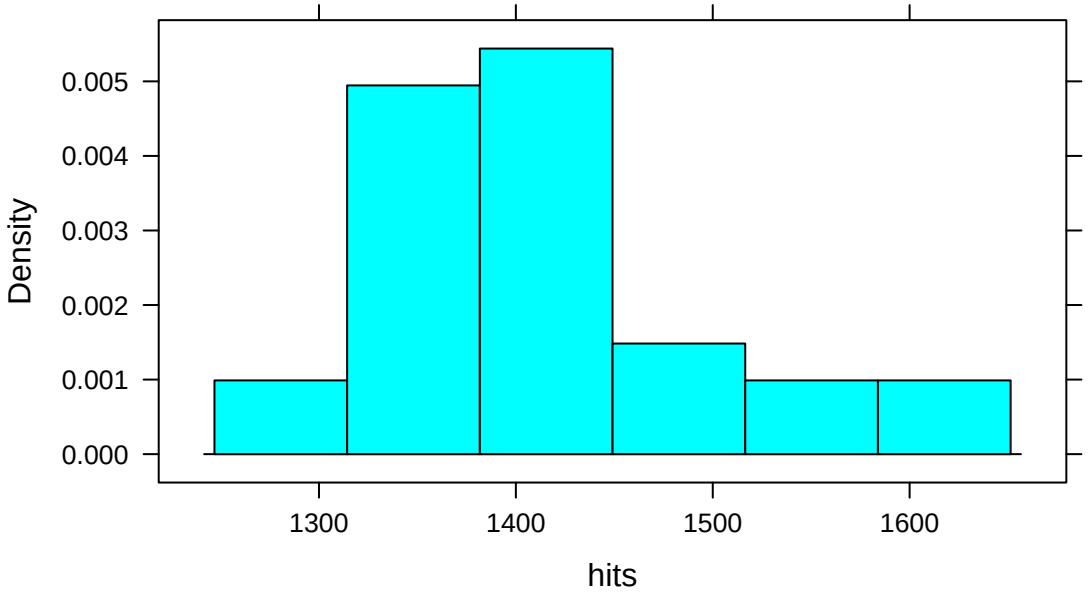
Question 7.

Make both a histogram and a boxplot of `hits`. What features are apparent in the histogram that aren't apparent in the boxplot? What features are apparent in the boxplot that aren't apparent in the histogram?

```
bwplot(~hits, data =mlb11)
```




```
histogram(~hits, data = mlb11)
```



The extent of outlier(s) are more obvious in the boxplot, while the unimodal shape is more apparent in the histogram.