Lab 05 - Describing Data

Francisco Santamarina

October 13, 2016

Load the necessary packages and dataset.

```
##load data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pander)
library("Lahman")
data("Teams")
```

Question 1

1. Drop all of the data before 1990.

```
postteams <- filter(Teams, yearID>=1990)
summary(postteams)
```

```
##      yearID     lgID         teamID       franchID      divID
##  Min.   :1990   AA:  0   ATL    : 26   ANA    : 26   Length:758
##  1st Qu.:1996   AL:367   BAL    : 26   ATL    : 26   Class :character
##  Median :2003   FL:  0   BOS    : 26   BAL    : 26   Mode  :character
##  Mean   :2003   NA:  0   CHA    : 26   BOS    : 26
##  3rd Qu.:2009   NL:391   CHN    : 26   CHC    : 26
##  Max.   :2015   PL:  0   CIN    : 26   CHW    : 26
##                 UA:  0   (Other):602   (Other):602
##       Rank            G             Ghome             W
##  Min.   :1.0   Min.   :112.0   Min.   :44.00   Min.   : 43.00
##  1st Qu.:2.0   1st Qu.:162.0   1st Qu.:81.00   1st Qu.: 71.00
##  Median :3.0   Median :162.0   Median :81.00   Median : 80.00
##  Mean   :3.1   Mean   :159.5   Mean   :79.75   Mean   : 79.75
##  3rd Qu.:4.0   3rd Qu.:162.0   3rd Qu.:81.00   3rd Qu.: 89.00
##  Max.   :7.0   Max.   :163.0   Max.   :84.00   Max.   :116.00
##
```

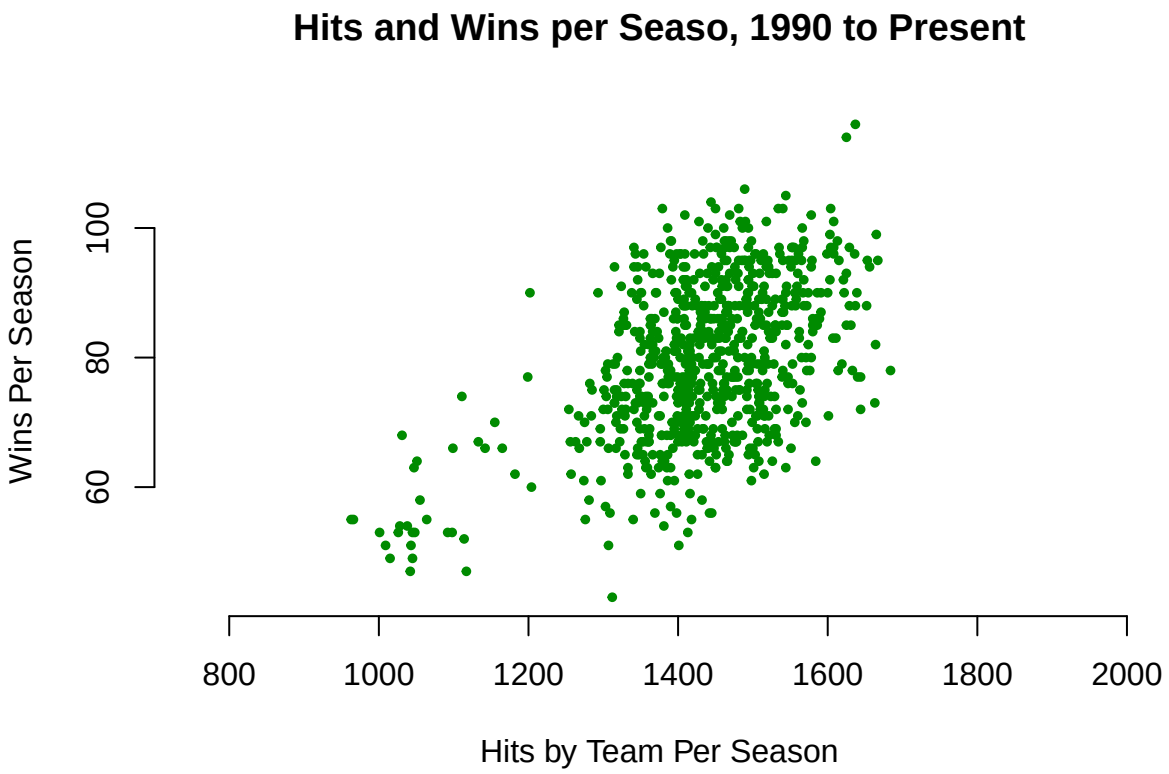
```
##        L              DivWin              WCWin              LgWin
##  Min.   : 40.00   Length:758         Length:758         Length:758
##  1st Qu.: 71.00   Class :character   Class :character   Class :character
##  Median : 79.00   Mode  :character   Mode  :character   Mode  :character
##  Mean   : 79.75
##  3rd Qu.: 88.75
##  Max.   :119.00
##
##    WSWin                  R                AB             H
##  Length:758         Min.   : 466.0   Min.   :3856   Min.   : 963
##  Class :character   1st Qu.: 671.2   1st Qu.:5474   1st Qu.:1381
##  Mode  :character   Median : 733.0   Median :5531   Median :1441
##                     Mean   : 734.8   Mean   :5461   Mean   :1433
##                     3rd Qu.: 793.0   3rd Qu.:5593   3rd Qu.:1503
##                     Max.   :1009.0   Max.   :5781   Max.   :1684
##
##       X2B              X3B              HR              BB
##  Min.   :159.0   Min.   :11.00   Min.   : 68.0   Min.   :319.0
##  1st Qu.:261.0   1st Qu.:24.00   1st Qu.:134.2   1st Qu.:472.2
##  Median :281.0   Median :30.00   Median :159.0   Median :523.0
##  Mean   :279.7   Mean   :30.52   Mean   :160.1   Mean   :526.2
##  3rd Qu.:301.0   3rd Qu.:36.00   3rd Qu.:183.8   3rd Qu.:577.0
##  Max.   :376.0   Max.   :61.00   Max.   :264.0   Max.   :775.0
##
##        SO              SB               CS              HBP
##  Min.   : 568   Min.   : 25.0   Min.   : 12.00   Min.   : 26.0
##  1st Qu.: 955   1st Qu.: 76.0   1st Qu.: 33.00   1st Qu.: 47.0
##  Median :1056   Median : 97.0   Median : 42.00   Median : 54.0
##  Mean   :1057   Mean   :101.3   Mean   : 43.25   Mean   : 56.1
##  3rd Qu.:1156   3rd Qu.:123.0   3rd Qu.: 51.00   3rd Qu.: 64.0
##  Max.   :1535   Max.   :256.0   Max.   :118.00   Max.   :103.0
##                                                  NA's   :278
##        SF              RA               ER              ERA
##  Min.   :24.00   Min.   : 448.0   Min.   : 407.0   Min.   :2.940
##  1st Qu.:38.00   1st Qu.: 668.2   1st Qu.: 608.2   1st Qu.:3.840
##  Median :44.00   Median : 731.5   Median : 671.0   Median :4.215
##  Mean   :44.68   Mean   : 734.8   Mean   : 672.8   Mean   :4.257
##  3rd Qu.:50.00   3rd Qu.: 798.8   3rd Qu.: 735.0   3rd Qu.:4.640
##  Max.   :75.00   Max.   :1103.0   Max.   :1015.0   Max.   :6.380
##  NA's   :278
##        CG              SHO              SV            IPouts
##  Min.   : 0.000   Min.   : 0.000   Min.   :20.00   Min.   :2952
##  1st Qu.: 4.000   1st Qu.: 6.000   1st Qu.:35.00   1st Qu.:4305
##  Median : 6.000   Median : 9.000   Median :41.00   Median :4333
##  Mean   : 7.599   Mean   : 8.823   Mean   :40.47   Mean   :4274
##  3rd Qu.:10.000   3rd Qu.:11.000   3rd Qu.:45.00   3rd Qu.:4361
##  Max.   :29.000   Max.   :24.000   Max.   :68.00   Max.   :4485
##
##        HA             HRA             BBA             SOA
##  Min.   : 929   Min.   : 76.0   Min.   :288.0   Min.   : 560.0
##  1st Qu.:1371   1st Qu.:141.0   1st Qu.:478.0   1st Qu.: 960.2
##  Median :1442   Median :160.5   Median :526.0   Median :1054.0
##  Mean   :1433   Mean   :160.1   Mean   :526.2   Mean   :1056.9
##  3rd Qu.:1509   3rd Qu.:180.0   3rd Qu.:574.0   3rd Qu.:1163.8
```

```
##  Max.   :1734   Max.   :241.0   Max.   :784.0   Max.   :1450.0
##
##        E               DP              FP             name
##  Min.   : 54.0   Min.   : 82.0   Min.   :0.9700   Length:758
##  1st Qu.: 93.0   1st Qu.:135.0   1st Qu.:0.9800   Class :character
##  Median :105.0   Median :148.0   Median :0.9810   Mode  :character
##  Mean   :106.2   Mean   :147.5   Mean   :0.9809
##  3rd Qu.:118.0   3rd Qu.:161.0   3rd Qu.:0.9840
##  Max.   :173.0   Max.   :204.0   Max.   :0.9910
##
##      park             attendance          BPF             PPF
##  Length:758         Min.   : 642745   Min.   : 88.0   Min.   : 88.0
##  Class :character   1st Qu.:1782273   1st Qu.: 97.0   1st Qu.: 97.0
##  Mode  :character   Median :2309898   Median :100.0   Median :100.0
##                     Mean   :2348487   Mean   :100.2   Mean   :100.2
##                     3rd Qu.:2887454   3rd Qu.:102.0   3rd Qu.:103.0
##                     Max.   :4483350   Max.   :129.0   Max.   :129.0
##
##    teamIDBR         teamIDlahman45     teamIDretro
##  Length:758         Length:758         Length:758
##  Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character
##
##
##
##
```

Question 2

2. Examine the relationship between Hits and Wins using a scatterplot. Use meaningful labels and a visually appealing style.

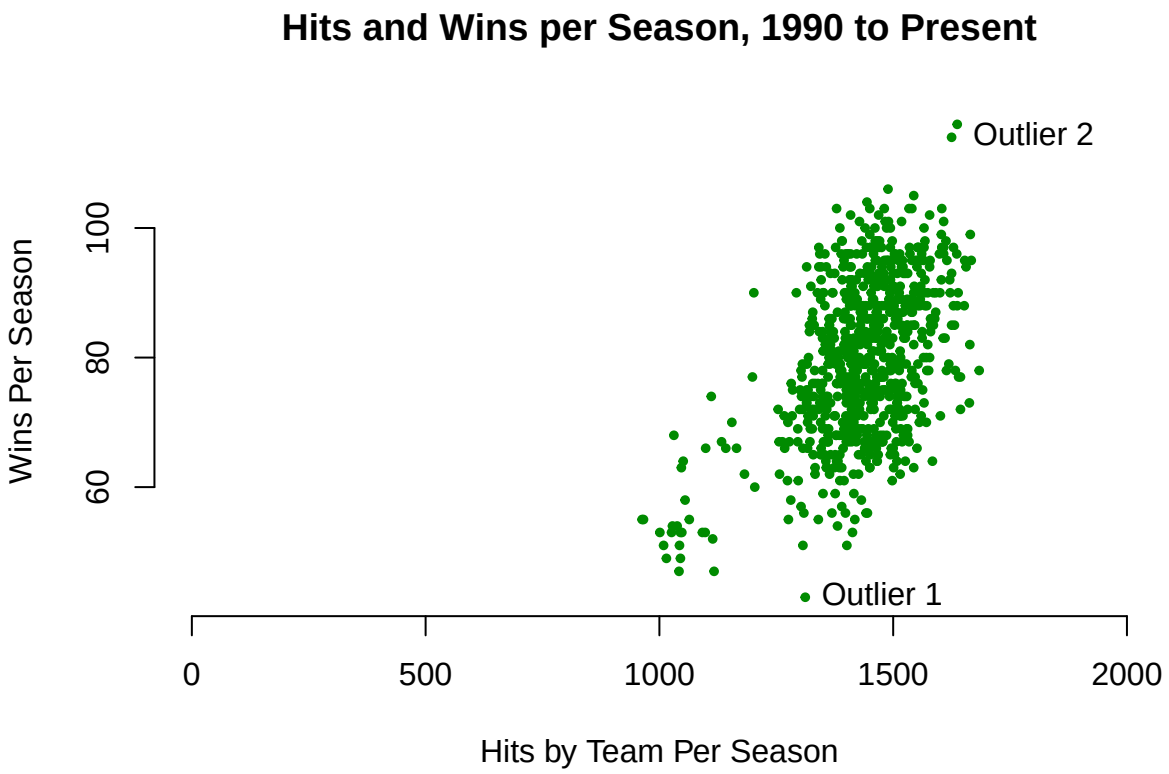
```r
plot( x=postteams$H, y=postteams$W,
      xlim=c(750,2000), #limits the range of data for variable x
      main="Hits and Wins per Seaso, 1990 to Present", #title
      xlab="Hits by Team Per Season", #x-axis label
      ylab="Wins Per Season", #y-axis label
      col="green4", #color
      pch=20, #indicates the plot symbol to use
      cex=.8, #determines size of the symbols/plots
      bty="n" #gets rid of the box around the plot
    )
```



Question 3

3. Annotate two outliers on the graph with the teamID. You can identify the points using the identify function and clicking on the graph near the points to return their positions: identify(x, y)

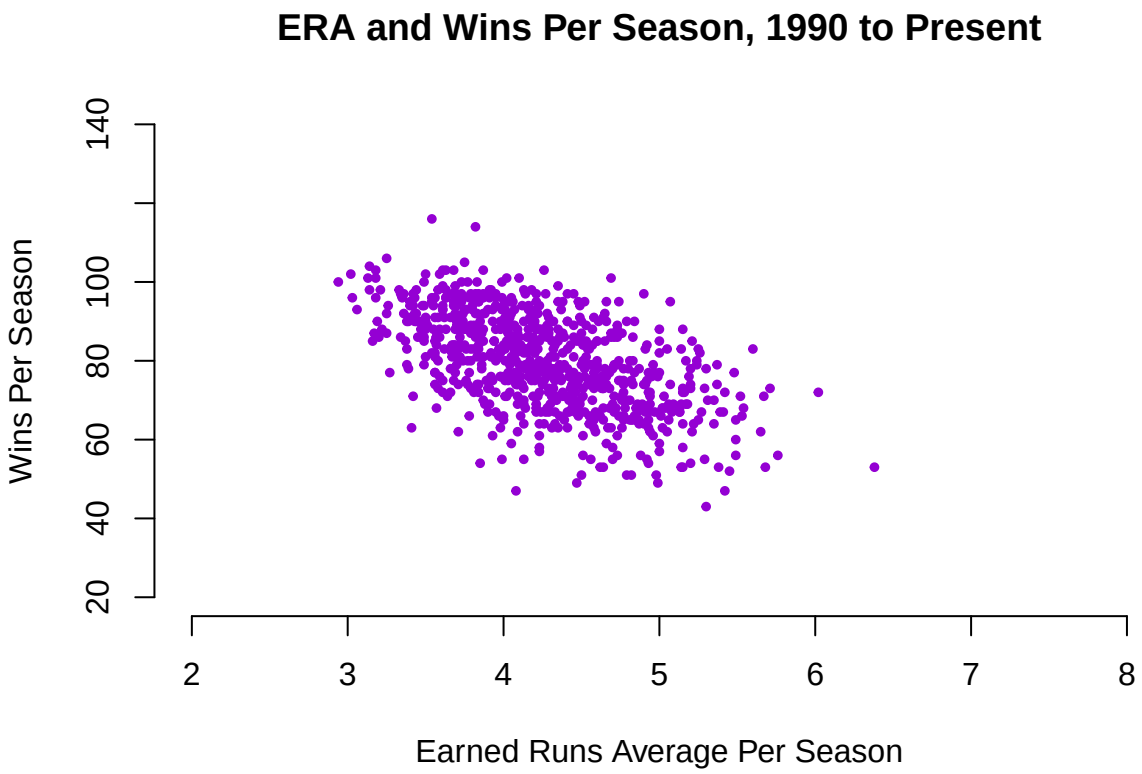
```
plot( x=postteams$H, y=postteams$W,
      xlim=c(0000,2000), #limits the range of data for variable x
      main="Hits and Wins per Season, 1990 to Present", #title
      xlab="Hits by Team Per Season", #x-axis label
      ylab="Wins Per Season", #y-axis label
      col="green4", #color
      pch=20, #indicates the plot symbol to use
      cex=.8, #determines size of the symbols/plots
      bty="n" #gets rid of the box around the plot
    )

text( x= 1306, y=43, labels= "Outlier 1", pos=4, col="gray0" )
text( x= 1630, y=114, labels="Outlier 2", pos=4, col="gray0" )
```



Question 4

4. Examine the relationship between ERA and Wins using a scatterplot. Use meaningful labels and a visually appealing style.

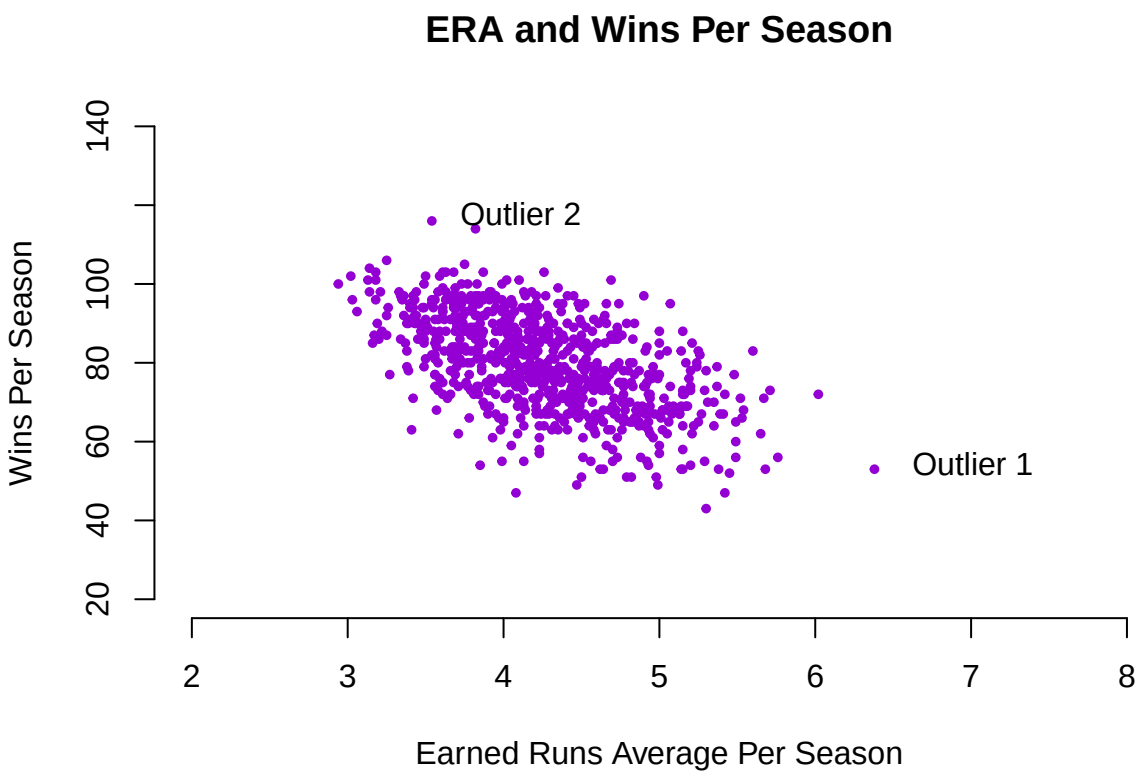
```
plot( x=postteams$ERA, y=postteams$W,
      xlim=c(2,8), #limits the range of data for variable x
      ylim=c(20,140), #limits the range of data for variable x
      main="ERA and Wins Per Season, 1990 to Present", #title
      xlab="Earned Runs Average Per Season", #x-axis label
      ylab="Wins Per Season", #y-axis label
      col="darkviolet", #color
      pch=20, #indicates the plot symbol to use
      cex=.8, #determines size of the symbols/plots
      bty="n" #gets rid of the box around the plot
    )
```



Question 5

5. Annotate two outliers on the graph with the teamID.

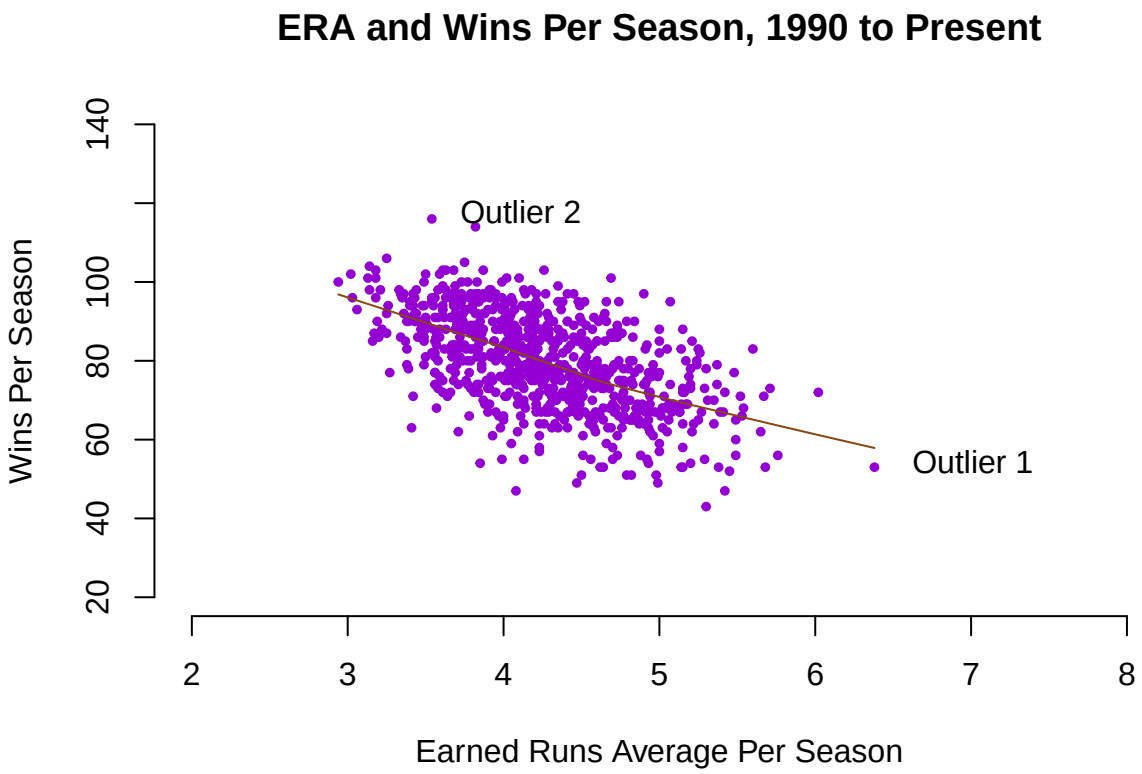
```
plot( x=postteams$ERA, y=postteams$W,
      xlim=c(2,8), #limits the range of data for variable x
      ylim=c(20,140), #limits the range of data for variable x
      main="ERA and Wins Per Season", #title
      xlab="Earned Runs Average Per Season", #x-axis label
      ylab="Wins Per Season", #y-axis label
      col="darkviolet", #color
      pch=20, #indicates the plot symbol to use
      cex=.8, #determines size of the symbols/plots
      bty="n" #gets rid of the box around the plot
    )
text( x= 6.5, y=53.4, labels= "Outlier 1", pos=4, col="gray0" )
text( x= 3.6, y=116.8, labels="Outlier 2", pos=4, col="gray0" )
```



Question 6

6. BONUS - add a trend line to the scatterplot to highlight the relationship.

```
plot( x=postteams$ERA, y=postteams$W,
      xlim=c(2,8), #limits the range of data for variable x
      ylim=c(20,140), #limits the range of data for variable x
      main="ERA and Wins Per Season, 1990 to Present", #title
      xlab="Earned Runs Average Per Season", #x-axis label
      ylab="Wins Per Season", #y-axis label
      col="darkviolet", #color
      pch=20, #indicates the plot symbol to use
      cex=.8, #determines size of the symbols/plots
      bty="n" #gets rid of the box around the plot
    )
text( x= 6.5, y=53.4, labels= "Outlier 1", pos=4, col="gray0" )
text( x= 3.6, y=116.8, labels="Outlier 2", pos=4, col="gray0" )

lines( lowess(postteams$ERA, postteams$W), col="chocolate4", lwd=1)
```


ERA and Wins Per Season, 1990 to Present
Wins Per Season
20
40
60
80
100
140
Outlier 2
Outlier 1
2
3
4
5
6
7
8
Earned Runs Average Per Season