Lab 4

Christopher Loan

October 14, 2020

Install and load the package *Lahman*, which will give you access to the dataset *Teams*

```
library(Lahman)
library(tidyverse)
library(janitor)
```

- Produce a subset of the data (as a new object) that has the following characteristics:
  - Only one team (your choice)
  - data from 1980 to present (or as present as the dataset gets)
  - Includes 5 columns: name, yearID, W, L, R, RA

(The variables above correspond to the team name, the year, wins, losses, runs scored, and runs allowed)

- Make sure you select a team that is currently still around, or it probably won't be interesting (see a list of current at http://www.espn.com/mlb/teams).
- Create a new variable corresponding to the winning percentage for the team you chose over time

$$w_{pct} = \frac{wins}{wins + losses}$$

- Order by winning percentage: Least to greatest
- Order by winning percentage: greatest to least
- Compute the mean and standard deviation of winning percentage
- With the full dataset
  - compute the average and standard deviation of winning percentage for each team.
  - Order by highest winning percentage
- Use the full data to reproduce the plot below

```
subset1 <- Teams %>% filter(teamID == teamID[length(Teams$teamID)], yearID >= 1980) %>% select(yearID, W
subset1
```

```
##    yearID  W   L   R  RA
## 1    2005 81  81 639 673
## 2    2006 71  91 746 872
## 3    2007 73  89 673 783
## 4    2008 59 102 641 825
## 5    2009 59 103 710 874
```

```
## 6    2010 69  93 655 742
## 7    2011 80  81 624 643
## 8    2012 98  64 731 594
## 9    2013 86  76 656 626
## 10   2014 96  66 686 555
## 11   2015 83  79 703 635
## 12   2016 95  67 763 612
## 13   2017 97  65 819 672
## 14   2018 82  80 771 682
## 15   2019 93  69 873 724
```

```
subset1 <- subset1 %>% mutate(win_perc = W/(W+L))
subset1
```

```
##    yearID  W   L   R  RA  win_perc
## 1    2005 81  81 639 673 0.5000000
## 2    2006 71  91 746 872 0.4382716
## 3    2007 73  89 673 783 0.4506173
## 4    2008 59 102 641 825 0.3664596
## 5    2009 59 103 710 874 0.3641975
## 6    2010 69  93 655 742 0.4259259
## 7    2011 80  81 624 643 0.4968944
## 8    2012 98  64 731 594 0.6049383
## 9    2013 86  76 656 626 0.5308642
## 10   2014 96  66 686 555 0.5925926
## 11   2015 83  79 703 635 0.5123457
## 12   2016 95  67 763 612 0.5864198
## 13   2017 97  65 819 672 0.5987654
## 14   2018 82  80 771 682 0.5061728
## 15   2019 93  69 873 724 0.5740741
```

```
low_to_high <- subset1 %>% arrange(win_perc)
low_to_high
```

```
##    yearID  W   L   R  RA  win_perc
## 1    2009 59 103 710 874 0.3641975
## 2    2008 59 102 641 825 0.3664596
## 3    2010 69  93 655 742 0.4259259
## 4    2006 71  91 746 872 0.4382716
## 5    2007 73  89 673 783 0.4506173
## 6    2011 80  81 624 643 0.4968944
## 7    2005 81  81 639 673 0.5000000
## 8    2018 82  80 771 682 0.5061728
## 9    2015 83  79 703 635 0.5123457
## 10   2013 86  76 656 626 0.5308642
## 11   2019 93  69 873 724 0.5740741
## 12   2016 95  67 763 612 0.5864198
## 13   2014 96  66 686 555 0.5925926
## 14   2017 97  65 819 672 0.5987654
## 15   2012 98  64 731 594 0.6049383
```

```r
high_to_low <- subset1 %>% arrange(desc(win_perc))
high_to_low
```

```
##    yearID  W   L   R  RA  win_perc
## 1    2012 98  64 731 594 0.6049383
## 2    2017 97  65 819 672 0.5987654
## 3    2014 96  66 686 555 0.5925926
## 4    2016 95  67 763 612 0.5864198
## 5    2019 93  69 873 724 0.5740741
## 6    2013 86  76 656 626 0.5308642
## 7    2015 83  79 703 635 0.5123457
## 8    2018 82  80 771 682 0.5061728
## 9    2005 81  81 639 673 0.5000000
## 10   2011 80  81 624 643 0.4968944
## 11   2007 73  89 673 783 0.4506173
## 12   2006 71  91 746 872 0.4382716
## 13   2010 69  93 655 742 0.4259259
## 14   2008 59 102 641 825 0.3664596
## 15   2009 59 103 710 874 0.3641975
```

```r
subset1 %>% summarize(avg_wins = mean(win_perc))
```

```
##    avg_wins
## 1 0.5032359
```

```r
subset1 %>% summarize(sd_of_wins = sd(win_perc))
```

```
##   sd_of_wins
## 1 0.08075644
```

```r
Teams %>% group_by(teamID) %>% mutate(win_perc = W/(W+L)) %>% summarize(avg_wins = mean(win_perc))
```

```
## # A tibble: 149 x 2
##    teamID avg_wins
##    <fct>     <dbl>
##  1 ALT       0.24
##  2 ANA       0.512
##  3 ARI       0.495
##  4 ATL       0.515
##  5 BAL       0.507
##  6 BFN       0.487
##  7 BFP       0.273
##  8 BL1       0.482
##  9 BL2       0.424
## 10 BL3       0.484
## # ... with 139 more rows
```

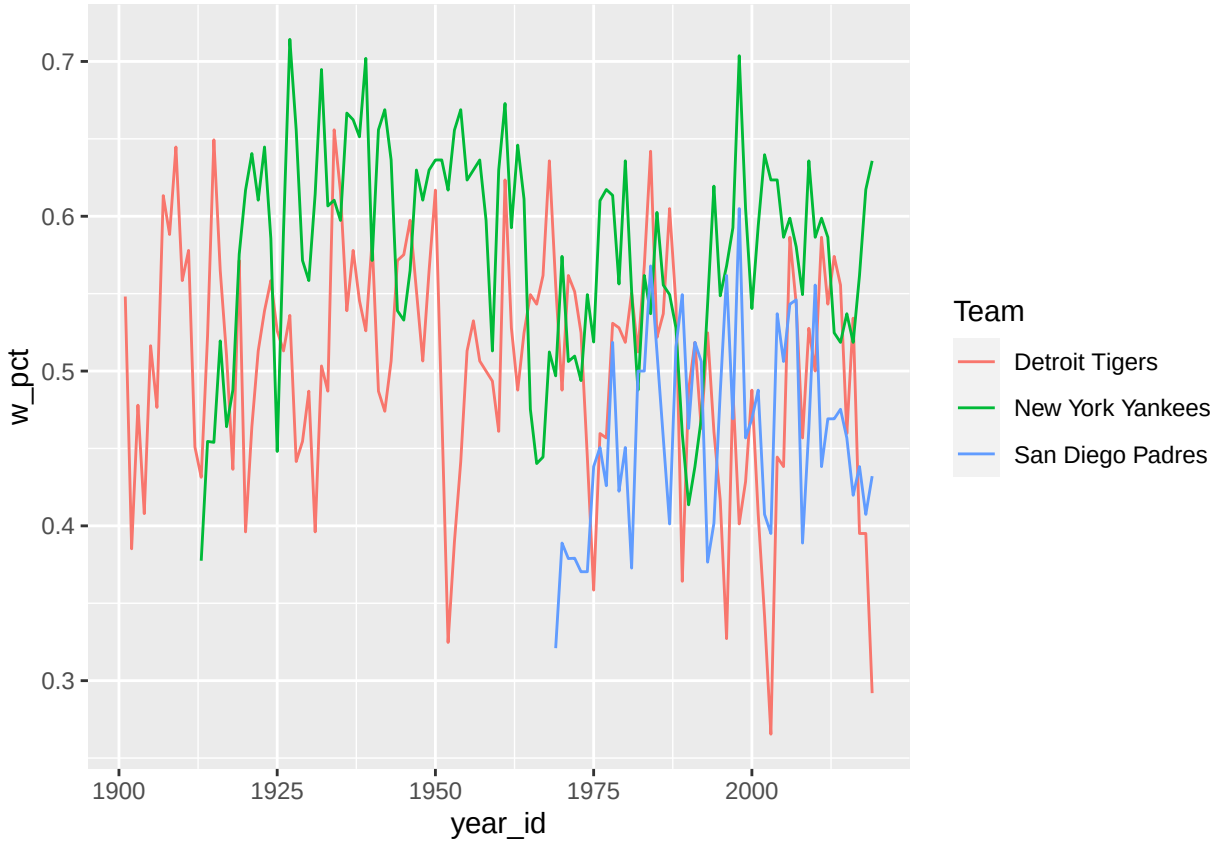
```r
Teams %>% group_by(teamID) %>% mutate(win_perc = W/(W+L)) %>% summarize(sd_of_wins = sd(win_perc))
```

```
## # A tibble: 149 x 2
##    teamID sd_of_wins
##    <fct>       <dbl>
##  1 ALT        NA
##  2 ANA         0.0577
##  3 ARI         0.0762
##  4 ATL         0.0808
##  5 BAL         0.0859
##  6 BFN         0.120
##  7 BFP        NA
##  8 BL1         0.253
##  9 BL2         0.125
## 10 BL3         0.0599
## # ... with 139 more rows
```

```
Teams %>% group_by(teamID) %>% mutate(win_perc = W/(W+L)) %>% summarize(avg_wins = mean(win_perc)) %>% a
```

```
## # A tibble: 149 x 2
##    teamID avg_wins
##    <fct>     <dbl>
##  1 SLU       0.832
##  2 BS1       0.773
##  3 BS2       0.689
##  4 CH1       0.679
##  5 MLU       0.667
##  6 PH1       0.661
##  7 CNU       0.657
##  8 SL4       0.637
##  9 BSP       0.628
## 10 HAR       0.613
## # ... with 139 more rows
```

```
# Please put the code for the plot in this chunk.
Teams %>%
  filter(name == 'Detroit Tigers'|
           name == 'New York Yankees'|
           name == 'San Diego Padres') %>%
  mutate(win_perc = W/(W+L)) %>%
  ggplot(mapping = aes(x = yearID,
                       y = win_perc,
                       color = name)) +
  geom_line() + labs(color = 'Team',
                     x = 'year_id',
                     y = 'w_pct')
```


0.7
0.6
0.5
0.4
0.3
w_pct
1900
1925
1950
1975
2000
year_id
Team
Detroit Tigers
New York Yankees
San Diego Padres