CPSC 260
Spring 2019

Ben's Time with the Mets

Practice with the R package: dplyr

We are going to explore Sean Lahman's historical baseball database, which contains complete seasonal records for all players on all Major League Baseball teams going back to 1871. These data are made available in R via the Lahman package.

1. Download the package Lahman. Then load it into R/retrieve it from the library so you can access the data set called Teams.

   ```
   install.packages("Lahman")
   library(Lahman)
   ```

2. There are 2835 rows and 48 columns in the data set, so don't try to view the whole thing!!!!! Instead take a quick look at just the first 10 rows. Also determine what type of object Teams is.

```
head(Teams)
```

```
  yearID lgID teamID franchID divID Rank  G Ghome  W  L
1   1871   NA    BS1      BNA  <NA>    3 31    NA 20 10
2   1871   NA    CH1      CNA  <NA>    2 28    NA 19  9
3   1871   NA    CL1      CFC  <NA>    8 29    NA 10 19
4   1871   NA    FW1      KEK  <NA>    7 19    NA  7 12
5   1871   NA    NY2      NNA  <NA>    5 33    NA 16 17
6   1871   NA    PH1      PNA  <NA>    1 28    NA 21  7
  DivWin WCWin LgWin WSWin   R   AB   H X2B X3B HR BB
1   <NA>  <NA>     N  <NA> 401 1372 426  70  37  3 60
2   <NA>  <NA>     N  <NA> 302 1196 323  52  21 10 60
3   <NA>  <NA>     N  <NA> 249 1186 328  35  40  7 26
4   <NA>  <NA>     N  <NA> 137  746 178  19   8  2 33
5   <NA>  <NA>     N  <NA> 302 1404 403  43  21  1 33
6   <NA>  <NA>     Y  <NA> 376 1281 410  66  27  9 46
  SO SB CS HBP SF  RA  ER  ERA CG SHO SV IPouts  HA HRA
1 19 73 16  NA NA 303 109 3.55 22   1  3    828 367   2
2 22 69 21  NA NA 241  77 2.76 25   0  1    753 308   6
3 25 18  8  NA NA 341 116 4.11 23   0  0    762 346  13
4  9 16  4  NA NA 243  97 5.17 19   1  0    507 261   5
5 15 46 15  NA NA 313 121 3.72 32   1  0    879 373   7
6 23 56 12  NA NA 266 137 4.95 27   0  0    747 329   3
  BBA SOA   E DP    FP                    name
1  42  23 243 24 0.834    Boston Red Stockings
2  28  22 229 16 0.829 Chicago White Stockings
3  53  34 234 15 0.818  Cleveland Forest Citys
4  21  17 163  8 0.803    Fort Wayne Kekiongas
5  42  22 235 14 0.840        New York Mutuals
6  53  16 194 13 0.845  Philadelphia Athletics
                        park attendance BPF PPF
1          South End Grounds I         NA 103  98
2      Union Base-Ball Grounds         NA 104 102
3 National Association Grounds         NA  96 100
4               Hamilton Field         NA 101 107
5     Union Grounds (Brooklyn)         NA  90  88
6     Jefferson Street Grounds         NA 102  98
  teamIDBR teamIDlahman45 teamIDretro
1      BOS            BS1         BS1
2      CHI            CH1         CH1
3      CLE            CL1         CL1
4      KEK            FW1         FW1
5      NYU            NY2         NY2
6      ATH            PH1         PH1
```

CPSC 260
Spring 2019

Ben's Time with the Mets

```
> class(Teams)
[1] "data.frame"
```

Ben worked for the New York Mets from 2004 to 2012. We are going to take a look at how well the team performed during those years. We'll start by asking: How many wins and losses did the Mets have during each of these years?

3. First we need to look at the column names and determine which ones hold the information we're looking for. In R print a list of column names, then look at the document explaining the variables in the data set.

   help(Teams)
   colnames(Teams)

```
[1] "yearID"          "lgID"            "teamID"
 [4] "franchID"        "divID"           "Rank"
 [7] "G"               "Ghome"           "W"
[10] "L"               "DivWin"          "WCWin"
[13] "LgWin"           "WSWin"           "R"
[16] "AB"              "H"               "X2B"
[19] "X3B"             "HR"              "BB"
[22] "SO"              "SB"              "CS"
[25] "HBP"             "SF"              "RA"
[28] "ER"              "ERA"             "CG"
[31] "SHO"             "SV"              "IPouts"
[34] "HA"              "HRA"             "BBA"
[37] "SOA"             "E"               "DP"
[40] "FP"              "name"            "park"
[43] "attendance"      "BPF"             "PPF"
[46] "teamIDBR"        "teamIDlahman45"  "teamIDretro"
```

4. What are the names for the columns that show the year, team name, number of wins, and number of losses?

yearID, teamID, W, L

5. Use the filter() and select() commands to quickly create a subset containing only this data for the years that Ben worked for the Mets.
   Note: NYN is the abbreviation used for the theam. It stands for New York National League Club.

   a) Do this using the filter and select commands separately.See guide below.

```
Mets = filter(Teams, teamID =='NYN', yearID>=2004, yearID<=2012)
> Mets2=select(Mets, yearID, teamID, W, L)
> Mets2
  yearID teamID  W  L
1   2004    NYN 71 91
2   2005    NYN 83 79
3   2006    NYN 97 65
4   2007    NYN 88 74
5   2008    NYN 89 73
6   2009    NYN 70 92
7   2010    NYN 79 83
```

Ben's Time with the Mets

```
8   2011    NYN 77 85
9   2012    NYN 74 88
```

b) Do this in one line, either nesting the commands or piping the commands together.

```
Mets3=Teams %>%
+   filter(teamID=="NYN", yearID>=2004, yearID<=2012) %>%
+   select(yearID, teamID, W, L)
> Mets3

  yearID teamID  W  L
1   2004    NYN 71 91
2   2005    NYN 83 79
3   2006    NYN 97 65
4   2007    NYN 88 74
5   2008    NYN 89 73
6   2009    NYN 70 92
7   2010    NYN 79 83
8   2011    NYN 77 85
9   2012    NYN 74 88
```

Guide for part (a): Filter the rows of the Teams data frame so that you only have the rows that correspond to the New York Mets. There are 54 of those, since the Mets joined the National League in 1962.

Next, filter these data so as to include only those seasons in which Ben worked for the team—those with yearID between 2004 and 2012.

Finally, select only those columns that were relevant to our question.

We've answered the simple question of how the Mets performed during the time that Ben was there, but since we are data scientists, we are interested in deeper questions. For example, some of these seasons were subpar—the Mets had more losses than wins. Did the team just get unlucky in those seasons? Or did they actually play as badly as their record indicates?

In order to answer this question, we need a model for the expected number of wins per season (or expected win percentage). It turns out that one of the most widely used contributions to the field of baseball analytics (courtesy of Bill James) is exactly that!

The simplest version of this model is: $\frac{1}{1+\left(\frac{RA}{R}\right)^2}$ ← expected win percentage

6. What does RA and R stand for? Look at the document explaining the variables in the data set.

R = runs scored, RA = opponents runs scored

7. Create a new subset of data containing everything the previous one did, plus RA and R.

Ben's Time with the Mets

```
myMets = Teams %>%
+   filter(teamID=="NYN", yearID %in% 2004:2012) %>%
+   select(yearID, teamID, W, L, R, RA)
> myMets

  yearID teamID  W  L   R  RA
1   2004    NYN 71 91 684 731
2   2005    NYN 83 79 722 648
3   2006    NYN 97 65 834 731
4   2007    NYN 88 74 804 750
5   2008    NYN 89 73 799 715
6   2009    NYN 70 92 671 757
7   2010    NYN 79 83 656 652
8   2011    NYN 77 85 718 742
9   2012    NYN 74 88 650 709
```

8. Compute the *actual* win percentage, call it WPct. This will be the number of wins divided by the total number of games: W/(W+L). Use the mutate command to add WPct to your data frame.

```
myMets2 = myMets %>% mutate(WPct = W/(W+L))
> myMets2

  yearID teamID  W  L   R  RA      WPct
1   2004    NYN 71 91 684 731 0.4382716
2   2005    NYN 83 79 722 648 0.5123457
3   2006    NYN 97 65 834 731 0.5987654
4   2007    NYN 88 74 804 750 0.5432099
5   2008    NYN 89 73 799 715 0.5493827
6   2009    NYN 70 92 671 757 0.4320988
7   2010    NYN 79 83 656 652 0.4876543
8   2011    NYN 77 85 718 742 0.4753086
9   2012    NYN 74 88 650 709 0.4567901
```

9. Now compute the *expected* win percentage (formula given above), call it E_WPct, and add it to your data frame.

```
myMets2 = myMets2 %>% mutate(expected_WPct = 1/(1+(RA/R)^2))
> myMets2

  yearID teamID  W  L   R  RA      WPct expected_WPct
1   2004    NYN 71 91 684 731 0.4382716     0.4668211
2   2005    NYN 83 79 722 648 0.5123457     0.5538575
3   2006    NYN 97 65 834 731 0.5987654     0.5655308
4   2007    NYN 88 74 804 750 0.5432099     0.5347071
5   2008    NYN 89 73 799 715 0.5493827     0.5553119
6   2009    NYN 70 92 671 757 0.4320988     0.4399936
7   2010    NYN 79 83 656 652 0.4876543     0.5030581
8   2011    NYN 77 85 718 742 0.4753086     0.4835661
9   2012    NYN 74 88 650 709 0.4567901     0.4566674
```

10. In how many seasons did the Mets perform as expected or better? Which seasons were these? Use filter() to print just these rows of data.

```
filter(myMets2, WPct>=expected_WPct)

yearID teamID  W  L   R  RA      WPct expected_WPct
```

Ben's Time with the Mets

```
1   2006    NYN 97 65 834 731 0.5987654     0.5655308
2   2007    NYN 88 74 804 750 0.5432099     0.5347071
3   2012    NYN 74 88 650 709 0.4567901     0.4566674
```

3 seasons

11. Ok, so which seasons were worst? We can simply sort the rows of the data frame using arrange(). But first, we need to define what is meant by "worst."

    a) Define worst as having the lowest *actual* win percentage.

```
arrange(myMets2, WPct)
  yearID teamID  W  L   R  RA      WPct expected_WPct
1   2009    NYN 70 92 671 757 0.4320988     0.4399936
2   2004    NYN 71 91 684 731 0.4382716     0.4668211
3   2012    NYN 74 88 650 709 0.4567901     0.4566674
4   2011    NYN 77 85 718 742 0.4753086     0.4835661
5   2010    NYN 79 83 656 652 0.4876543     0.5030581
6   2005    NYN 83 79 722 648 0.5123457     0.5538575
7   2007    NYN 88 74 804 750 0.5432099     0.5347071
8   2008    NYN 89 73 799 715 0.5493827     0.5553119
9   2006    NYN 97 65 834 731 0.5987654     0.5655308

arrange(myMets2, -WPct)
  yearID teamID  W  L   R  RA      WPct expected_WPct
1   2006    NYN 97 65 834 731 0.5987654     0.5655308
2   2008    NYN 89 73 799 715 0.5493827     0.5553119
3   2007    NYN 88 74 804 750 0.5432099     0.5347071
4   2005    NYN 83 79 722 648 0.5123457     0.5538575
5   2010    NYN 79 83 656 652 0.4876543     0.5030581
6   2011    NYN 77 85 718 742 0.4753086     0.4835661
7   2012    NYN 74 88 650 709 0.4567901     0.4566674
8   2004    NYN 71 91 684 731 0.4382716     0.4668211
9   2009    NYN 70 92 671 757 0.4320988     0.4399936
```

    b) Now let's sort by how much worse they did than expected, i.e. looking at the difference between actual win percent and expected win percent. (Negative values would indicate that they did worse than expected, while positive values would indicate that they did better than expected.)

```
arrange(myMets2, WPct - expected_WPct)
1 -0.0285494528
2 -0.0415117862
3  0.0332345873
4  0.0085027505
5 -0.0059291854
6 -0.0078947848
7 -0.0154037544
8 -0.0082574426
9  0.0001227254
```

12. If we want we can add this difference to the data frame, too! Add it and call it diff.

```
myMets3 = myMets2 %>% mutate(diff = WPct-expected_WPct)
```

Ben's Time with the Mets

```
> myMets3
  yearID teamID  W  L   R  RA      WPct expected_WPct
1   2004    NYN 71 91 684 731 0.4382716     0.4668211
2   2005    NYN 83 79 722 648 0.5123457     0.5538575
3   2006    NYN 97 65 834 731 0.5987654     0.5655308
4   2007    NYN 88 74 804 750 0.5432099     0.5347071
5   2008    NYN 89 73 799 715 0.5493827     0.5553119
6   2009    NYN 70 92 671 757 0.4320988     0.4399936
7   2010    NYN 79 83 656 652 0.4876543     0.5030581
8   2011    NYN 77 85 718 742 0.4753086     0.4835661
9   2012    NYN 74 88 650 709 0.4567901     0.4566674
           diff
1 -0.0285494528
2 -0.0415117862
3  0.0332345873
4  0.0085027505
5 -0.0059291854
6 -0.0078947848
7 -0.0154037544
8 -0.0082574426
9  0.0001227254
```

13. Let's also add what diff equates to in terms of the number of games. Determine this by multiplying diff by total number of games. Call this new variable Game_diff.

```
myMets3 = myMets3 %>% mutate(Game_diff = diff*(W+L))
> myMets3
  yearID teamID  W  L   R  RA      WPct expected_WPct
1   2004    NYN 71 91 684 731 0.4382716     0.4668211
2   2005    NYN 83 79 722 648 0.5123457     0.5538575
3   2006    NYN 97 65 834 731 0.5987654     0.5655308
4   2007    NYN 88 74 804 750 0.5432099     0.5347071
5   2008    NYN 89 73 799 715 0.5493827     0.5553119
6   2009    NYN 70 92 671 757 0.4320988     0.4399936
7   2010    NYN 79 83 656 652 0.4876543     0.5030581
8   2011    NYN 77 85 718 742 0.4753086     0.4835661
9   2012    NYN 74 88 650 709 0.4567901     0.4566674
           diff   Game_diff
1 -0.0285494528 -4.62501135
2 -0.0415117862 -6.72490937
3  0.0332345873  5.38400315
4  0.0085027505  1.37744558
5 -0.0059291854 -0.96052803
6 -0.0078947848 -1.27895513
7 -0.0154037544 -2.49540821
8 -0.0082574426 -1.33770571
9  0.0001227254  0.01988152
```

14. Now that we have two diffs, should be more specific with the first one we created. Rename the diff column pct_diff using the rename() command.

```
> myMets3 = rename(myMets3, pct_diff = diff)
```

Ben's Time with the Mets

You can see that 2006 was the Mets' most fortunate year—since they won five more games than our model predicts—but 2005 was the least fortunate—since they won almost seven games fewer than our model predicts.

This type of analysis helps us understand how the Mets performed in individual seasons, but we know that any randomness that occurs in individual years is likely to average out over time. So while it is clear that the Mets performed well in some seasons and poorly in others, what can we say about their overall performance?

15. Let's use summarize() to find the Met's average *actual* win percentage.

```
myMets3 %>% summarise(avg_WPct = mean(WPct))
   avg_WPct
1 0.4993141
```

16. In addition to average actual win percentage, also use summarize to count the number or seasons we are looking at, and the total number of wins and losses during these seasons.

```
myMets3 %>% summarise(avg_WPct = mean(WPct), N_seasons=n(), total_W=sum(W),
total_L=sum(L))

avg_WPct N_seasons total_W total_L
1 0.4993141         9     728     730
```

Usually, when we are summarizing a data frame, like we did above, it is more interesting to consider different groups. In this case, we can discretize these years into three chunks: one for each of the three general managers under whom Ben worked. Jim Duquette was the Mets' general manager in 2004, Omar Minaya from 2005 to 2010, and Sandy Alderson from 2011 to 2012.

17. Add a column to your data frame that specifies the manager for each year using the given code:

```
Name of your data frame %>% mutate(gm = ifelse(yearID==2004,"Duquette",
                                               ifelse(yearID>2010, "Alderson","Minaya")))
```

```
myMets4 = myMets3 %>% mutate(gm = ifelse(yearID==2004, "Duquette",
                                         ifelse(yearID>2010, "Alderson", "Minaya")))
myMets4

yearID teamID  W  L   R  RA      WPct expected_WPct
1   2004    NYN 71 91 684 731 0.4382716     0.4668211
2   2005    NYN 83 79 722 648 0.5123457     0.5538575
3   2006    NYN 97 65 834 731 0.5987654     0.5655308
4   2007    NYN 88 74 804 750 0.5432099     0.5347071
5   2008    NYN 89 73 799 715 0.5493827     0.5553119
6   2009    NYN 70 92 671 757 0.4320988     0.4399936
7   2010    NYN 79 83 656 652 0.4876543     0.5030581
8   2011    NYN 77 85 718 742 0.4753086     0.4835661
9   2012    NYN 74 88 650 709 0.4567901     0.4566674
       pct_diff   Game_diff       gm
```

Ben's Time with the Mets

```
1 -0.0285494528 -4.62501135 Duquette
2 -0.0415117862 -6.72490937   Minaya
3  0.0332345873  5.38400315   Minaya
4  0.0085027505  1.37744558   Minaya
5 -0.0059291854 -0.96052803   Minaya
6 -0.0078947848 -1.27895513   Minaya
7 -0.0154037544 -2.49540821   Minaya
8 -0.0082574426 -1.33770571 Alderson
9  0.0001227254  0.01988152 Alderson
```

~~18. Look at the details for creating the general manager (gm) variable:~~

~~gm=ifelse(yearID==2004,"Duquette", ifelse(yearID>2010, "Alderson","Minaya")))~~

~~Write a few statements explaining what this code has done.~~

19. Now compute the same summaries as in #16, but use group_by() to compare based on who the general manager was at the time.

```
myMets4 %>% group_by(gm) %>%
  summarise(avg_WPct = mean(WPct), N_seasons=n(), total_W=sum(W), total_L=sum(L))
myMets4
```

```
gm       avg_WPct N_seasons total_W total_L
  <chr>       <dbl>     <int>   <int>   <int>
1 Alderson    0.466         2     151     173
2 Duquette    0.438         1      71      91
3 Minaya      0.521         6     506     466
> myMets4
  yearID teamID  W  L   R  RA      WPct expected_WPct
1   2004    NYN 71 91 684 731 0.4382716     0.4668211
2   2005    NYN 83 79 722 648 0.5123457     0.5538575
3   2006    NYN 97 65 834 731 0.5987654     0.5655308
4   2007    NYN 88 74 804 750 0.5432099     0.5347071
5   2008    NYN 89 73 799 715 0.5493827     0.5553119
6   2009    NYN 70 92 671 757 0.4320988     0.4399936
7   2010    NYN 79 83 656 652 0.4876543     0.5030581
8   2011    NYN 77 85 718 742 0.4753086     0.4835661
9   2012    NYN 74 88 650 709 0.4567901     0.4566674
       pct_diff   Game_diff       gm
1 -0.0285494528 -4.62501135 Duquette
2 -0.0415117862 -6.72490937   Minaya
3  0.0332345873  5.38400315   Minaya
4  0.0085027505  1.37744558   Minaya
5 -0.0059291854 -0.96052803   Minaya
6 -0.0078947848 -1.27895513   Minaya
7 -0.0154037544 -2.49540821   Minaya
8 -0.0082574426 -1.33770571 Alderson
9  0.0001227254  0.01988152 Alderson
```

20. Our summaries don't all make sense now that we're making comparisons. This is because the general managers each worked for a different number of seasons! List at least 3 summaries that would make sense to compare across general managers.

CPSC 260
Spring 2019

Ben's Time with the Mets

21. Compute the summaries that you listed in #20. Also include the number of seasons that each general manager worked.

#Number 20 and 21

#Compare the runs scored, opponents runs scored, and the team W % for their first year.

#2004 - Duquette = 684 runs, 731 opponent runs, .4382716 W %

#2005 - Minaya = 722 runs, 648 opponent runs, .5123457 W %

#2011 - Alderson = 718 runs, 742 opponent runs, .4753086 W %

22. So who was the best general manager? Write a few statements based on your summaries to back up your claim.

Minaya would be the best general manager in my opinion as he has the most runs earned while having the least opponent runs scored. He also has the highest win percentage based off of each manager's first year.

Challenge!

Complete all of your work in one long command. The full power of the chaining/piping operator is revealed when we do all the analysis at once. We will still retain the step-by-step logic.

1. You will not do *everything* above, but do the following in one long command:
    A. Specify you want the columns: yearID, teamID, W, L, R, RA – but only for the Mets during the years that Ben worked with them.
    B. Add columns for: WPct, E_WPct, gm, and a new variable: *expected* number of games the Mets should win (E_W).
    C. Summarize the data you have by computing the number of seasons per gm, and the average for the following variables per gm: W, E_W, WPct
    D. Arrange the summarized results so that the highest average WPct is in the top row.