Using joins to create MLB tables

Group 9

M7 ICA1

Introduction

Today you will get practice merging data frames with inner and outer join functions available in package `dplyr`. To get started, load packages `tidyverse` and `Lahman`.

Package `Lahman` has numerous data frames about Major League Baseball. Type `help(package = "Lahman")` in your Console to see everything available.

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

Data

You will work with data frames in package `Lahman`. When needed, utilize R's help to get an understanding of what the variables are in a given data frame. For example, `?Salaries` will provide a short description for each of the variables in data frame `Salaries`.

Questions

Question 1

Select three data frames from package `Lahman`. Identify what variables are in common between any pair of the three data frames, and identify what variables are in common between all three data frames. What are the primary keys for each data frame?

```
names(Batting)
```

```
 [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "G"
 [7] "AB"       "R"        "H"        "X2B"      "X3B"      "HR"
[13] "RBI"      "SB"       "CS"       "BB"       "SO"       "IBB"
[19] "HBP"      "SH"       "SF"       "GIDP"
```

```
names(Pitching)
```

```
 [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "W"
 [7] "L"        "G"        "GS"       "CG"       "SHO"      "SV"
```

```
[13] "IPouts"   "H"        "ER"       "HR"       "BB"       "SO"
[19] "BAOpp"    "ERA"      "IBB"      "WP"       "HBP"      "BK"
[25] "BFP"      "GF"       "R"        "SH"       "SF"       "GIDP"
```

```
names(Fielding)
```

```
 [1] "playerID" "yearID"   "stint"    "teamID"   "lgID"     "POS"
 [7] "G"        "GS"       "InnOuts"  "PO"       "A"        "E"
[13] "DP"       "PB"       "WP"       "SB"       "CS"       "ZR"
```

All 3 data frames have: playerID, yearID, stint, teamID, lgID, G

Batting and Pitching share: HBP, SF, H Batting and Fielding share: CS Pitching and Fielding share: GS

Batting primary key for the data frames is Player ID Pitching primary key for the data frames is Player ID Fielding primary key for the data frames is Player ID

Question 2

Use data frames `Managers` and `AwardsManagers` to reproduce the data frame given in the html output.

```
full_join(Managers, AwardsManagers) %>%
  filter(yearID==2016)
```

```
    playerID yearID teamID lgID inseason   G   W   L rank plyrMgr
1  showabu99   2016    BAL   AL        1 162  89  73    2       N
2  farrejo03   2016    BOS   AL        1 162  93  69    1       N
3  venturo01   2016    CHA   AL        1 162  78  84    4       N
4  francte01   2016    CLE   AL        1 161  94  67    1       N
5  ausmubr01   2016    DET   AL        1 161  86  75    2       N
6  hinchaj01   2016    HOU   AL        1 162  84  78    3       N
7   yostne01   2016    KCA   AL        1 162  81  81    3       N
8  sciosmi01   2016    LAA   AL        1 162  74  88    4       N
9  molitpa01   2016    MIN   AL        1 162  59 103    5       N
10 girarjo01   2016    NYA   AL        1 162  84  78    4       N
11 melvibo01   2016    OAK   AL        1 162  69  93    5       N
12 servasc01   2016    SEA   AL        1 162  86  76    2       N
13  cashke01   2016    TBA   AL        1 162  68  94    5       N
14 banisje01   2016    TEX   AL        1 162  95  67    1       N
15 gibbojo02   2016    TOR   AL        1 162  89  73    2       N
16  halech01   2016    ARI   NL        1 162  69  93    4       N
17 gonzafr99   2016    ATL   NL        1  37   9  28    5       N
18 snitkbr99   2016    ATL   NL        2 124  59  65    5       N
19 maddojo99   2016    CHN   NL        1 162 103  58    1       N
20 pricebr99   2016    CIN   NL        1 162  68  94    5       N
21 weisswa01   2016    COL   NL        1 162  75  87    3       N
22 roberda07   2016    LAN   NL        1 162  91  71    1       N
23 mattido01   2016    MIA   NL        1 161  79  82    3       N
24 counscr01   2016    MIL   NL        1 162  73  89    4       N
25 collite99   2016    NYN   NL        1 162  87  75    2       N
26 mackape01   2016    PHI   NL        1 162  71  91    4       N
27 hurdlcl01   2016    PIT   NL        1 162  78  83    3       N
```

```
28 greenan01   2016    SDN   NL        1 162  68  94    5       N
29 bochybr01   2016    SFN   NL        1 162  87  75    2       N
30 mathemi01   2016    SLN   NL        1 162  86  76    2       N
31 bakerdu01   2016    WAS   NL        1 162  95  67    1       N
                     awardID  tie notes
1                       <NA> <NA>    NA
2                       <NA> <NA>    NA
3                       <NA> <NA>    NA
4  BBWAA Manager of the Year <NA>    NA
5                       <NA> <NA>    NA
6                       <NA> <NA>    NA
7                       <NA> <NA>    NA
8                       <NA> <NA>    NA
9                       <NA> <NA>    NA
10                      <NA> <NA>    NA
11                      <NA> <NA>    NA
12                      <NA> <NA>    NA
13                      <NA> <NA>    NA
14                      <NA> <NA>    NA
15                      <NA> <NA>    NA
16                      <NA> <NA>    NA
17                      <NA> <NA>    NA
18                      <NA> <NA>    NA
19                      <NA> <NA>    NA
20                      <NA> <NA>    NA
21                      <NA> <NA>    NA
22 BBWAA Manager of the Year <NA>    NA
23                      <NA> <NA>    NA
24                      <NA> <NA>    NA
25                      <NA> <NA>    NA
26                      <NA> <NA>    NA
27                      <NA> <NA>    NA
28                      <NA> <NA>    NA
29                      <NA> <NA>    NA
30                      <NA> <NA>    NA
31                      <NA> <NA>    NA
```

Question 3

Use data frames `Managers`, `AwardsManagers`, and `Master` to reproduce the data frame given in the html output.

```
mawards <- inner_join(Managers, AwardsManagers)
full_join(mawards, Master) %>%
select(yearID,teamID,  nameFirst, nameLast, awardID) %>%
  filter(yearID >= 2000)
```

```
   yearID teamID nameFirst   nameLast                  awardID
1    2000    CHA     Jerry     Manuel BBWAA Manager of the Year
2    2000    CHA     Jerry     Manuel   TSN Manager of the Year
3    2000    SFN     Dusty      Baker BBWAA Manager of the Year
4    2000    SFN     Dusty      Baker   TSN Manager of the Year
5    2001    SEA       Lou   Piniella BBWAA Manager of the Year
```

6	2001	SEA	Lou	Piniella	TSN	Manager of the Year
7	2001	PHI	Larry	Bowa	BBWAA	Manager of the Year
8	2001	PHI	Larry	Bowa	TSN	Manager of the Year
9	2002	ANA	Mike	Scioscia	BBWAA	Manager of the Year
10	2002	ANA	Mike	Scioscia	TSN	Manager of the Year
11	2002	ATL	Bobby	Cox	TSN	Manager of the Year
12	2002	SLN	Tony	LaRussa	BBWAA	Manager of the Year
13	2003	KCA	Tony	Pena	BBWAA	Manager of the Year
14	2003	KCA	Tony	Pena	TSN	Manager of the Year
15	2003	ATL	Bobby	Cox	TSN	Manager of the Year
16	2003	FLO	Jack	McKeon	BBWAA	Manager of the Year
17	2004	MIN	Ron	Gardenhire	TSN	Manager of the Year
18	2004	TEX	Buck	Showalter	BBWAA	Manager of the Year
19	2004	ATL	Bobby	Cox	BBWAA	Manager of the Year
20	2004	ATL	Bobby	Cox	TSN	Manager of the Year
21	2005	CHA	Ozzie	Guillen	BBWAA	Manager of the Year
22	2005	CHA	Ozzie	Guillen	TSN	Manager of the Year
23	2005	ATL	Bobby	Cox	BBWAA	Manager of the Year
24	2005	ATL	Bobby	Cox	TSN	Manager of the Year
25	2006	DET	Jim	Leyland	BBWAA	Manager of the Year
26	2006	DET	Jim	Leyland	TSN	Manager of the Year
27	2006	FLO	Joe	Girardi	BBWAA	Manager of the Year
28	2006	FLO	Joe	Girardi	TSN	Manager of the Year
29	2007	CLE	Eric	Wedge	BBWAA	Manager of the Year
30	2007	CLE	Eric	Wedge	TSN	Manager of the Year
31	2007	ARI	Bob	Melvin	BBWAA	Manager of the Year
32	2007	ARI	Bob	Melvin	TSN	Manager of the Year
33	2008	TBA	Joe	Maddon	BBWAA	Manager of the Year
34	2008	TBA	Joe	Maddon	TSN	Manager of the Year
35	2008	CHN	Lou	Piniella	BBWAA	Manager of the Year
36	2008	FLO	Fredi	Gonzalez	TSN	Manager of the Year
37	2009	LAA	Mike	Scioscia	BBWAA	Manager of the Year
38	2009	LAA	Mike	Scioscia	TSN	Manager of the Year
39	2009	COL	Jim	Tracy	BBWAA	Manager of the Year
40	2009	COL	Jim	Tracy	TSN	Manager of the Year
41	2010	MIN	Ron	Gardenhire	BBWAA	Manager of the Year
42	2010	MIN	Ron	Gardenhire	TSN	Manager of the Year
43	2010	SDN	Buddy	Black	BBWAA	Manager of the Year
44	2010	SDN	Buddy	Black	TSN	Manager of the Year
45	2011	TBA	Joe	Maddon	BBWAA	Manager of the Year
46	2011	TBA	Joe	Maddon	TSN	Manager of the Year
47	2011	ARI	Kirk	Gibson	BBWAA	Manager of the Year
48	2011	ARI	Kirk	Gibson	TSN	Manager of the Year
49	2012	BAL	Buck	Showalter	TSN	Manager of the Year
50	2012	OAK	Bob	Melvin	BBWAA	Manager of the Year
51	2012	WAS	Davey	Johnson	BBWAA	Manager of the Year
52	2012	WAS	Davey	Johnson	TSN	Manager of the Year
53	2013	BOS	John	Farrell	TSN	Manager of the Year
54	2013	CLE	Terry	Francona	BBWAA	Manager of the Year
55	2013	PIT	Clint	Hurdle	BBWAA	Manager of the Year
56	2013	PIT	Clint	Hurdle	TSN	Manager of the Year
57	2014	BAL	Buck	Showalter	TSN	Manager of the Year
58	2014	BAL	Buck	Showalter	BBWAA	Manager of the Year
59	2014	WAS	Matt	Williams	TSN	Manager of the Year

```
60   2014    WAS      Matt   Williams BBWAA Manager of the Year
61   2015    MIN      Paul    Molitor   TSN Manager of the Year
62   2015    TEX      Jeff   Banister BBWAA Manager of the Year
63   2015    CHN       Joe     Maddon BBWAA Manager of the Year
64   2015    NYN     Terry    Collins   TSN Manager of the Year
65   2016    CLE     Terry   Francona BBWAA Manager of the Year
66   2016    LAN      Dave    Roberts BBWAA Manager of the Year
```

Question 4

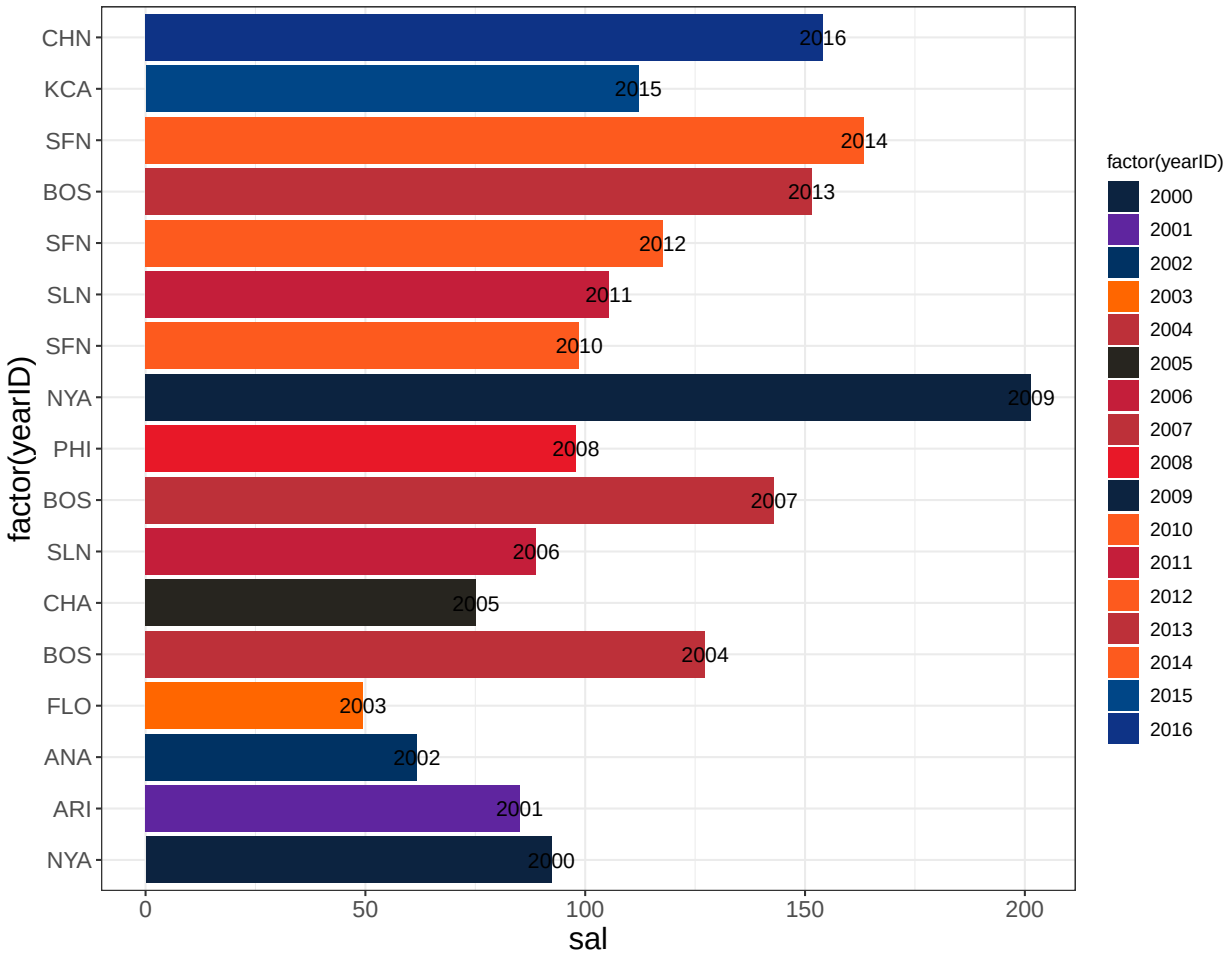
Use data frames `Teams` and `Salaries` to reproduce the plot given in the html output. Bar colors represent each team's primary color.

```
Winners <- Teams%>%
  filter(WSWin == "Y")%>%
  select("teamID", "yearID")

winner.salaries <- Salaries%>%
  semi_join(Winners, by = c("teamID", "yearID"))

totals <- winner.salaries%>%
  filter(yearID > 1999)%>%
  group_by(teamID, yearID)%>%
  summarise(sal = sum(salary)/1000000)%>%
  arrange(yearID)

  ggplot(totals, aes(x = factor(yearID), y = sal, fill = factor(yearID))) +
  geom_bar(stat = "identity", position = "dodge") +
    theme_custom() +
  scale_x_discrete(labels = totals$teamID) +
    coord_flip() +
  scale_fill_manual(values = c("#0C2340", "#5F259F", "#003263", "#FF6600", "#BD3039", "#27251F", "#C41E3
    geom_text(aes(label = yearID))
```



Question 5

Adjust your plot in Question 4 for inflation with year 2000 as baseline. Comment on the differences between the plots.

Year	USD Value	Inflation Rate
2000	$1.00	3.36%
2001	$1.03	2.85%
2002	$1.04	1.58%
2003	$1.07	2.28%
2004	$1.10	2.66%
2005	$1.13	3.39%
2006	$1.17	3.23%
2007	$1.20	2.85%
2008	$1.25	3.84%
2009	$1.25	-0.36%
2010	$1.27	1.64%
2011	$1.31	3.16%
2012	$1.33	2.07%
2013	$1.35	1.46%
2014	$1.37	1.62%

Year	USD Value	Inflation Rate
2015	$1.38	0.12%
2016	$1.39	1.26%

```
usd = c(1.00, 1.03, 1.04, 1.07, 1.10, 1.13, 1.17, 1.20, 1.25, 1.25, 1.27, 1.31, 1.33, 1.35, 1.37, 1.38,
length(usd)
```

```
[1] 17
```

```
length(Winners$sal*usd)
```

```
[1] 0
```

```
n.salary = Winners$sal*usd

Winners%>%
  ggplot(mapping = aes(y = n.salary, x = as.factor(yearID), fill = teamID)) +
  geom_bar(stat = "identity") +
  scale_x_discrete(labels = totals$teamID) +
  scale_fill_manual(values = c("#0C2340", "#5F259F", "#003263", "#FF6600", "#BD3039", "#27251F", "#C41E3
  coord_flip() +
  theme_custom() +
  geom_text(aes(label = yearID), color = "white", position = position_stack(vjust = 0.9)) +
  labs(title = "Payroll of world series winners", x = "World Series Winners", y = "Salary in millions")
```

```
Error: Aesthetics must be either length 1 or the same as the data (121): y
```

References

1. Lahman, S. (2017) Lahman's Baseball Database, 1871-2016.
2. RStudio Cheatsheets