Homework 7 *Biostatistics 203A* *Yenlin Lai*

Due date: Thursday December 2nd @ 6PM

1. Use `tapply` to calculate mean price by neighborhood and write the results of this function out to a new data frame called `avg.price`. List the name of the most expensive neighborhood:

```
library(dplyr)

airbnb <- read.csv("D:/UCLA Biostat/Fall 2021/Biostat 203A/Assignment7/airbnb_los_angeles_2017_03_10.csv", header = T)

avg.price <- tapply(airbnb$price, airbnb$neighborhood, mean)

avg.price <- as.data.frame(avg.price)

head(avg.price)
```

```
> head(avg.price)
                 avg.price
Adams-Normandie   75.92105
Agoura Hills     138.61111
Alhambra          87.36424
Alondra Park     113.60000
Altadena         145.86391
Arcadia           94.32768
```

```
avg.price %>%

 filter(avg.price == max(avg.price))
```

```
> avg.price %>%
+   filter(avg.price == max(avg.price))
                avg.price
Tujunga Canyons      2595
```

The most expensive neighborhood is Tujunga Canyons, which has mean price 2595.

2. The `dplyr` package is extremely powerful and useful for both data manipulation and summarization. We will review 5 main commands available in the `dplyr` package.

As done previously using `tapply`, calculate mean overall satisfaction by room type using `summarize`.

```
 library(dplyr)
airbnb %>%
 group_by(room_type) %>%
 summarise(mean_of_satisfaction = mean(overall_satisfaction))
```

```
# A tibble: 3 x 2
  room_type       mean_of_satisfaction
  <chr>                          <dbl>
1 Entire home/apt                 3.03
2 Private room                    2.69
3 Shared room                     2.08
```

3. The Major League Baseball Data

To practice summarizing data in R, we will familiarize ourselves with a new data set containing team information by year for each of the existing 30 teams from 1876 to 2016. This data set was originally compiled for an analysis of coaching records and to attempt to answer the question of why managers change jobs. The data was originally extracted from https://www.baseball-reference.com (https://www.baseball-reference.com).
The following variables are included in the **baseballdata.csv** file:

Year	**Calendar year**
Tm	Team name in the calendar year
Lg	League
G	Total games played
W	Total games won
L	Total games lost
Ties	Total games tied
WL	Win loss percentage
Finish	Standing at end of season
GB	Games back relative to the team in first place
Playoff	Information about how the team finished the playoffs, if they participated
R	Total runs earned
RA	Total runs allowed
Attendance	Annual attendance at games
BatAge	Average age of all batters on the team
PAge	Average age of all pitchers on the team
TopPlayer	The best player on the team in that calendar year
Managers	The team's manager or managers in that calendar year

```
bball <- read.csv("baseballdata.csv", sep = ",", header = TRUE, stringsA
sFactors = FALSE)
head(bball)
```

Exercises

1. How many years did the Dodgers have a winning record? What percentage does this constitute?

```
win <- bball %>%
  filter(Tm == "Los Angeles Dodgers" & WL >= 0.5) %>%
  nrow()
win
[1] 46
```

Dodgers had a winning record for 46 years.

```
complete <- bball %>%
  filter(Tm == "Los Angeles Dodgers") %>%
  nrow()
win/complete
[1] 0.779661
```

The percentage of the number of years the team won out of the total number of years the team played is 77.97%.

2. Which team has the most world series wins?

```
wswin <- bball %>%
 filter(grepl("Won WS",Playoffs)) %>%
 arrange(Year) %>%
 select(Tm, Year, Playoffs)
names(which.max(table(wswin$Tm)))
```

[1] "New York Yankees"

New York Yankees has the most world series wins.

Exercise [~HW]:

On how many occassions was the team who won the world series in the current year the same team as the team that won the world series the previous year?

```
for (n in 1:nrow(wswin)+1) {
 if(wswin$Tm[n]==wswin$Tm[n+1]) {
  wswin$strike[n] <- "Y"
 }
 else {wswin$strike[n] <- "N"
 }
}
table(wswin$strike)
```

```
> table(wswin$strike)

 N  Y
93 22
```

There were 22 occasions the team who won the world series in the current year the same team as the team that won the world series the previous year.

3. Which team has competed in the most World Series?

```
wsattend <- bball %>%
 filter(grepl("WS",Playoffs)) %>%
 arrange(Year) %>%
 select(Tm, Year, Playoffs)
names(which.max(table(wsattend$Tm)))
```

[1] "New York Yankees"

New York Yankees has competed in the most World Series.

4. Create a new variable containing the run differential (runs earned minus runs allowed) and calculate the average run differential across each of the 30 current teams.

```
bball_R <- bball %>%
 mutate(RD = R-RA)
```

```
head(bball_R)
```

```
> head(bball_R)
  Year                   Tm      Lg   G  W  L Ties    WL pythWL   Finish GB
1 2016 Arizona Diamondbacks NL West 162 69 93    0 0.426  0.424 4th of 5 22
2 2015 Arizona Diamondbacks NL West 162 79 83    0 0.488  0.504 3rd of 5 13
3 2014 Arizona Diamondbacks NL West 162 64 98    0 0.395  0.415 5th of 5 30
4 2013 Arizona Diamondbacks NL West 162 81 81    0 0.500  0.493 2nd of 5 11
5 2012 Arizona Diamondbacks NL West 162 81 81    0 0.500  0.530 3rd of 5 13
6 2011 Arizona Diamondbacks NL West 162 94 68    0 0.580  0.545 1st of 5 --
        Playoffs   R  RA Attendance BatAge PAge           TopPlayer
1                752 890  2,036,216   26.7 26.4      J.Segura (5.7)
2                720 713  2,080,145   26.6 27.1 P.Goldschmidt (8.8)
3                615 742  2,073,730   27.6 28.0 P.Goldschmidt (4.5)
4                685 695  2,134,895   28.1 27.6 P.Goldschmidt (7.1)
5                734 688  2,177,617   28.3 27.4        A.Hill (5.0)
6 Lost LDS (3-2) 731 662  2,105,432   28.2 27.4       J.Upton (6.1)
                            Managers              current   RD
1                       C.Hale (69-93) Arizona Diamondbacks -138
2                       C.Hale (79-83) Arizona Diamondbacks    7
3 K.Gibson (63-96) and A.Trammell (1-2) Arizona Diamondbacks -127
4                     K.Gibson (81-81) Arizona Diamondbacks  -10
5                     K.Gibson (81-81) Arizona Diamondbacks   46
6                     K.Gibson (94-68) Arizona Diamondbacks   69
```

```
meanRD <- bball_R %>%
 group_by(current) %>%
 summarise(meanRD = mean(RD))
as.data.frame(meanRD)
```

```
> as.data.frame(meanRD)
                          current     meanRD
1            Arizona Diamondbacks -19.052632
2                  Atlanta Braves   4.865248
3                Baltimore Orioles -40.034483
4                   Boston Red Sox  27.543103
5                     Chicago Cubs  22.730496
6                Chicago White Sox   5.612069
7                  Cincinnati Reds   8.585185
8                Cleveland Indians  13.422414
9                 Colorado Rockies -38.916667
10                  Detroit Tigers  10.405172
11                  Houston Astros -12.381818
12              Kansas City Royals -27.229167
13 Los Angeles Angels of Anaheim  -2.750000
14             Los Angeles Dodgers  33.263158
15                   Miami Marlins -51.333333
16               Milwaukee Brewers -27.062500
17                 Minnesota Twins -30.301724
18                   New York Mets -16.272727
19               New York Yankees 105.403509
20               Oakland Athletics -22.965517
21           Philadelphia Phillies -43.447761
22              Pittsburgh Pirates   5.318519
23                San Diego Padres -57.229167
24           San Francisco Giants  56.641791
25                Seattle Mariners -40.075000
26             St. Louis Cardinals  27.348148
27                  Tampa Bay Rays -57.947368
28                  Texas Rangers -30.035714
29              Toronto Blue Jays  10.750000
30           Washington Nationals -25.187500
```

5. Was the mean average age of batters higher from 1990-1999 or from 2000-2009?

```
mage1990s <- bball %>%
 filter(Year >= 1990 & Year <= 1999) %>%
 summarise(meanage = mean(BatAge))
mage2000s <- bball %>%
 filter(Year >= 2000 & Year <= 2009) %>%
 summarise(meanage = mean(BatAge))
mage1990s > mage2000s
    meanage
[1,]   FALSE
```

No, the mean average age of batters was not higher from 1990-1999 than from 2000-2009.

6. Among the teams who won a world series, which team had the worst record and in what year?

```
wswinwl <- bball %>%
 filter(grepl("Won WS",Playoffs)) %>%
 arrange(Year) %>%
 select(Tm, Year, WL, Playoffs)
wswinwl %>%
 filter(WL==min(WL))
```

```
> wswinwl %>%
+   filter(WL==min(WL))
                    Tm Year    WL      Playoffs
1 St. Louis Cardinals 2006 0.516 Won WS (4-1)
```

St. Louis Cardinals had the worst record and in 2006 among the teams which won a world series.