Assignment 3

Evan Hollier

2023-03-01

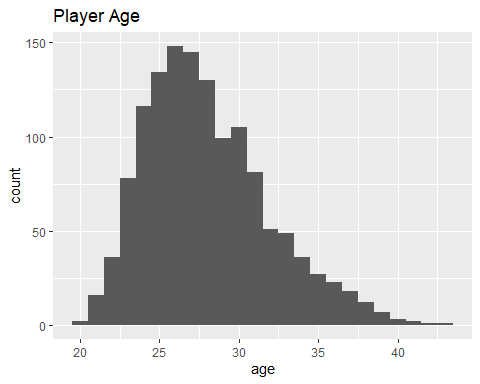
Connect to the mlb database.

db = dbConnect(MySQL(),   
 user='root',   
 password='mysql',   
 dbname='mlb',   
 host='localhost')  
  
dbListTables(db)

## [1] "mlb\_batting" "mlb\_manager" "mlb\_master" "mlb\_pitching" "mlb\_team"

**1) Plot a histogram of the players age as of January 1st, 2014. When the question asks for players, be sure and exclude the managers.**

rs = dbSendQuery(db, 'SELECT player\_id, birth\_year, birth\_month, birth\_day   
 FROM mlb\_master   
 WHERE player\_id NOT IN  
 (SELECT player\_id FROM mlb\_manager)'  
 )  
dat = dbFetch(rs, n = -1) # n = -1 fetches all rows  
dat$age = as.integer(difftime(  
 "2014-1-1",  
 paste(dat$birth\_year,dat$birth\_month,dat$birth\_day, sep='-')))  
dat$age = dat$age / 365  
ggplot(dat, aes(x=age)) +  
 geom\_histogram(binwidth=1) +  
 ggtitle("Player Age")



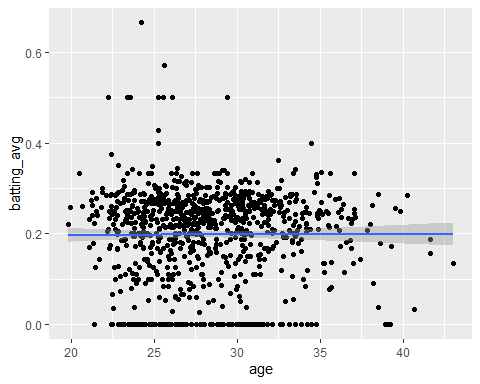
**2) Do a regression on players heights and weights to get the mathematical relationship between the two. (get the slope and intercept of the regression).**  
(Done in Python)  
Y = -154.89 + 4.9455X  
  
**3) Plot the relationship between players age (as of Jan. 1, 2014) and batting average. Batting average is determined by dividing a player’s hits by his total at-bats for a number between zero (shown as .000) and one (1.000). Do players get better or worse with age? What is the best age for a player?**

rs = dbSendQuery(db, 'SELECT player\_id, hits/at\_bats AS batting\_avg, team\_id,  
 hits, at\_bats   
 FROM mlb\_batting   
 WHERE at\_bats>1 AND  
 player\_id NOT IN  
 (SELECT player\_id FROM mlb\_manager)'  
 )

## Warning in .local(conn, statement, ...): Decimal MySQL column 1 imported as  
## numeric

# Players whose have never been at bat are filtered.  
dat1 = dbFetch(rs, n = -1)  
dat1 <- merge(dat[c("player\_id","age")], dat1, by="player\_id")  
  
ggplot(dat1, aes(x = age, y = batting\_avg)) +  
 geom\_point() +  
 stat\_smooth(method='lm')

## `geom\_smooth()` using formula = 'y ~ x'



# Determine best age:  
dat1$flat\_age = floor(dat1$age)  
batavg\_age <- aggregate(hits~flat\_age, dat1, sum)  
batavg\_age <- merge(aggregate(at\_bats~flat\_age, dat1, sum),   
 batavg\_age, by="flat\_age")  
batavg\_age$batting\_avg = batavg\_age$hits / batavg\_age$at\_bats  
  
batavg\_age[batavg\_age$batting\_avg == max(batavg\_age$batting\_avg),]

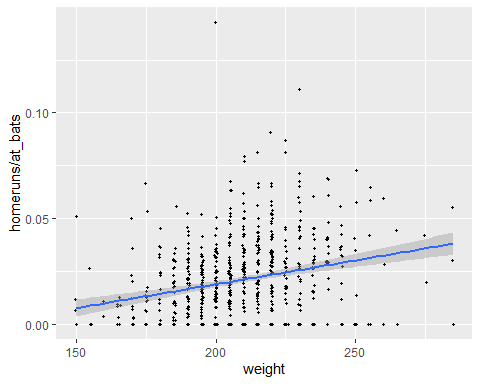
## flat\_age at\_bats hits batting\_avg  
## 2 20 121 32 0.2644628

Using linear regression, there doesn’t seem to be any relationship between age and batting average. The line of regression is horizontal, meaning as age increases, batting average doesn’t change. The age with the best batting average is 20.

**4) Do heavier players hit more home runs per at bat than lighter players? Plot the relationship between weight vs home runs per at bat. Compare home runs per at bat for above average weight players to below average weight players? Is the difference statistically significant? Remember when calculating the average, don’t average the averages. Sum up the home runs, sum up the at bats, then divide them.**

rs = dbSendQuery(db, 'SELECT weight,   
 homeruns, at\_bats  
 FROM mlb\_master  
 INNER JOIN mlb\_batting  
 ON mlb\_master.player\_id = mlb\_batting.player\_id  
 WHERE at\_bats>20 AND  
 mlb\_master.player\_id NOT IN  
 (SELECT player\_id FROM mlb\_manager)'  
 )  
dat2 = dbFetch(rs, n = -1)  
  
ggplot(dat2, aes(x = weight, y = homeruns/at\_bats)) +  
 geom\_jitter(size=0.6) +  
 stat\_smooth(method='lm')

## `geom\_smooth()` using formula = 'y ~ x'



# Compare above average to below average:  
avg\_weight <- mean(dat2$weight)  
homerunavg\_weight <- aggregate(homeruns~weight>avg\_weight, dat2, sum)  
homerunavg\_weight <- merge(aggregate(at\_bats~weight>avg\_weight, dat2, sum),   
 homerunavg\_weight)  
homerunavg\_weight$homerun\_avg = homerunavg\_weight$homeruns / homerunavg\_weight$at\_bats  
  
# Determine statistical significance  
true\_homerun\_avg = sum(homerunavg\_weight$homeruns) / sum(homerunavg\_weight$at\_bats)  
dat2.below <- dat2[dat2$weight < avg\_weight,]  
dat2.below$homerun\_avg = dat2.below$homeruns / dat2.below$at\_bats  
dat2.above <- dat2[dat2$weight > avg\_weight,]  
dat2.above$homerun\_avg = dat2.above$homeruns / dat2.above$at\_bats  
  
t.test(dat2.below$homerun\_avg,   
 dat2.above$homerun\_avg,  
 mu=true\_homerun\_avg)

##   
## Welch Two Sample t-test  
##   
## data: dat2.below$homerun\_avg and dat2.above$homerun\_avg  
## t = -23.739, df = 659.49, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0.02550918  
## 95 percent confidence interval:  
## -0.011297015 -0.005673299  
## sample estimates:  
## mean of x mean of y   
## 0.01646245 0.02494760

Using linear regression, the relationship between weight and home run average is shown to be positive. The line of regression slopes upward, meaning as weight increases, home run average increases. The home run average is 0.01906026 for below average weight players, and 0.03176666 for above average. The home run average for all players is 0.02550918. Performing a Two Sample t-test on the home run averages above and below average weight indicates the difference in means is statistically significant.

**5) Which team has the oldest average age for players? Is there a relationship between average age of players and winning percentage?**

rs = dbSendQuery(db, 'SELECT player\_id, mlb\_batting.team\_id FROM mlb\_batting  
 INNER JOIN mlb\_team  
 ON mlb\_team.team\_id = mlb\_batting.team\_id  
 WHERE mlb\_batting.player\_id NOT IN  
 (SELECT player\_id FROM mlb\_manager)'  
 )  
dat3 = dbFetch(rs, n = -1)  
dat3 <- merge(dat[c("player\_id","age")], dat3, by="player\_id")  
dat3 <- aggregate(age~team\_id, dat3, mean)  
dat3[dat3$age == max(dat3$age),]

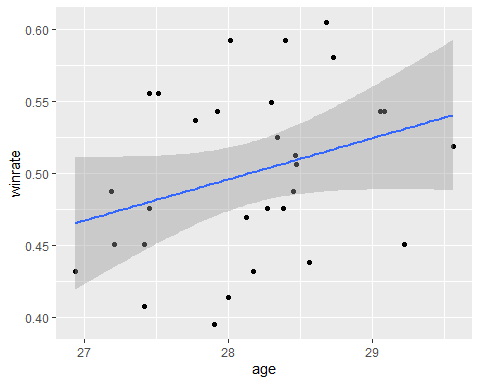
## team\_id age  
## 18 NYA 29.565

rs = dbSendQuery(db, 'SELECT team\_id, wins/games as winrate FROM mlb\_team')

## Warning in .local(conn, statement, ...): Decimal MySQL column 1 imported as  
## numeric

dat3.2 = dbFetch(rs, n = -1)  
dat3 <- merge(dat3, dat3.2, by="team\_id")  
  
ggplot(dat3, aes(x = age, y = winrate)) +  
 geom\_point() +  
 stat\_smooth(method='lm')

## `geom\_smooth()` using formula = 'y ~ x'



New York Yankees (NYA) have the oldest average age for players. Using linear regression, the relationship between a team’s average age and winning percentage is shown to be positive. The line of regression slopes upward, meaning as average age increases, winning percentage increases.

**6) Calculate the percentage of times a player is caught stealing (from batting table: caught\_stealing/(stolen\_bases + caught\_stealing) \* 100.**  
**Is there a statistically significant difference between the best manager and the worst manager? Best manager being defined as the manager with the lowest percentage of times their players get caught stealing.**

rs = dbSendQuery(db, 'SELECT team\_id, stolen\_bases, caught\_stealing  
 FROM mlb\_batting  
 WHERE mlb\_batting.player\_id NOT IN  
 (SELECT player\_id FROM mlb\_manager)'  
 )  
dat4 = dbFetch(rs, n = -1)  
dat4.ag <- aggregate(stolen\_bases~team\_id, dat4, sum)  
dat4.ag <- merge(aggregate(caught\_stealing~team\_id, dat4, sum),   
 dat4.ag)  
# Impossible to differentiate player stats based on manager stint,   
# so stick with just team\_id to represent a manager.  
dat4.ag$caught\_rate = dat4.ag$caught\_stealing / (dat4.ag$stolen\_bases + dat4.ag$caught\_stealing) \* 100  
dat4.ag

## team\_id caught\_stealing stolen\_bases caught\_rate  
## 1 ARI 33 86 27.73109  
## 2 ATL 33 95 25.78125  
## 3 BAL 20 44 31.25000  
## 4 BOS 25 63 28.40909  
## 5 CHA 36 85 29.75207  
## 6 CHN 40 65 38.09524  
## 7 CIN 52 122 29.88506  
## 8 CLE 27 104 20.61069  
## 9 COL 48 85 36.09023  
## 10 DET 41 106 27.89116  
## 11 HOU 37 122 23.27044  
## 12 KCA 36 153 19.04762  
## 13 LAA 39 81 32.50000  
## 14 LAN 50 138 26.59574  
## 15 MIA 21 58 26.58228  
## 16 MIL 43 102 29.65517  
## 17 MIN 36 99 26.66667  
## 18 NYA 26 112 18.84058  
## 19 NYN 34 101 25.18519  
## 20 OAK 20 83 19.41748  
## 21 PHI 26 109 19.25926  
## 22 PIT 47 104 31.12583  
## 23 SDN 34 91 27.20000  
## 24 SEA 42 96 30.43478  
## 25 SFN 27 56 32.53012  
## 26 SLN 32 57 35.95506  
## 27 TBA 27 63 30.00000  
## 28 TEX 59 105 35.97561  
## 29 TOR 21 78 21.21212  
## 30 WAS 23 101 18.54839

min(dat4.ag$caught\_rate)

## [1] 18.54839

max(dat4.ag$caught\_rate)

## [1] 38.09524

t.test(dat4.ag$caught\_rate)

##   
## One Sample t-test  
##   
## data: dat4.ag$caught\_rate  
## t = 26.852, df = 29, p-value < 2.2e-16  
## alternative hypothesis: true mean is not equal to 0  
## 95 percent confidence interval:  
## 25.42073 29.61248  
## sample estimates:  
## mean of x   
## 27.51661

There is a statistically significant difference between the best manager and the worst manager. The best and worst managers’ percentages of times their players get caught stealing are both well outside the 95 percent confidence interval for all managers.

**7) For each team, calculate the batting average (hits/at\_bats) for the team. Do all teams with better than average batting also have better than 50% winning percentage? Print a table with all the teams that have above average batting and less than 50% winning percentage.**

dat5 <- aggregate(hits~team\_id, dat1, sum)  
dat5 <- merge(aggregate(at\_bats~team\_id, dat1, sum), dat5, by="team\_id")  
dat5$batting\_avg = dat5$hits / dat5$at\_bat  
dat5 <- merge(dat5[c("team\_id", "batting\_avg")],   
 dat3[c("team\_id", "winrate")],   
 by="team\_id")  
  
(true\_batting\_avg <- mean(dat5$batting\_avg))

## [1] 0.2511381

dat5 <- dat5[dat5$batting\_avg > true\_batting\_avg,]  
dat5 <- dat5[dat5$winrate < .5,]  
dat5

## team\_id batting\_avg winrate  
## 5 CHA 0.2526164 0.4506  
## 9 COL 0.2765692 0.4074  
## 15 MIA 0.2527095 0.4753  
## 17 MIN 0.2536831 0.4321  
## 28 TEX 0.2565982 0.4136

No, there are some teams with above average batting and less than 50% winning percentage. These teams are: CHA, COL, MIA, MIN, and TEX.