STAT 542

Homework 3 Example Solutions

Problem 2, Chapter 4:

(a) In the ‘select’ function, we list columns to be selected, not a logical condition (as we would in the ‘filter’ function).

(b) Nothing wrong with this; it selects all columns except ‘year’.

(c) Should be a singles equals-sign, not a double equals-sign.

(d) Need filter(sex==’M’)

(e) No error caused, but it will only select ‘year’ once, not twice, so we don’t need to list year twice.

(f) No error, but each level of ‘n’ is just a single number, so we are averaging each single value of n, which is kind of pointless.

(g) Nothing wrong with this. This code provides counts of people with “uncommon names” and counts of people with “common names”, where a common name is defined as one shared by more than 100 people.

Problem 3, Chapter 4:

The line

filter(am == 1)

will work on the original data frame, but not on the summary statistic output created by the ‘summarize’ function. Should move the

filter(am == 1)

to the second line of code in the pipeline.

Problem 6, Chapter 4:

(a) summarize (b) mutate (c) arrange (d) filter (e) select

Problem 11, Chapter 4:

There does not appear to be much association between the number of violations and the median violation score, for zip codes with at least 50 violations (the picture is nearly identical if we instead consider zip codes with at least 50 *inspections*). See the scatterplot below:



If we focus attention on only zip codes in Manhattan (boro == “MANHATTAN”), then the plot looks slightly different, but still little association (maybe weak positive association):



Problem 13, Chapter 4:

(a):

 yearID teamID lgID W L WSWin WinPct

1 2008 LAA AL 100 62 N 0.6172840

2 2002 ANA AL 99 63 Y 0.6111111

3 2014 LAA AL 98 64 N 0.6049383

4 2009 LAA AL 97 65 N 0.5987654

5 2005 LAA AL 95 67 N 0.5864198

6 2007 LAA AL 94 68 N 0.5802469

7 1982 CAL AL 93 69 N 0.5740741

8 1986 CAL AL 92 70 N 0.5679012

9 2004 ANA AL 92 70 N 0.5679012

10 1989 CAL AL 91 71 N 0.5617284

(b): Yes, the WSWin columns shows they won the World Series in 2002.

Problem 14, Chapter 4:

The plane with the most flights out of NYC in 2013 is the one with talinum ‘N725MQ’.

Here is the plot of the number of trips per month for this plane:



Problem 2, Chapter 5 exercises:

(a)

 playerID nameFirst nameLast tHR tSB

 <chr> <chr> <chr> <int> <int>

1 bondsba01 Barry Bonds 762 514

2 rodrial01 Alex Rodriguez 696 329

3 mayswi01 Willie Mays 660 338

4 dawsoan01 Andre Dawson 438 314

5 beltrca01 Carlos Beltran 435 312

6 bondsbo01 Bobby Bonds 332 461

7 sandere02 Reggie Sanders 305 304

8 finlest01 Steve Finley 304 320

(b)

 playerID nameFirst nameLast tW tSO

 <chr> <chr> <chr> <int> <int>

 1 johnswa01 Walter Johnson 417 3509

 2 maddugr01 Greg Maddux 355 3371

 3 clemero02 Roger Clemens 354 4672

 4 carltst01 Steve Carlton 329 4136

 5 ryanno01 Nolan Ryan 324 5714

 6 suttodo01 Don Sutton 324 3574

 7 niekrph01 Phil Niekro 318 3342

 8 perryga01 Gaylord Perry 314 3534

 9 seaveto01 Tom Seaver 311 3640

10 johnsra05 Randy Johnson 303 4875

(c) From this table, Pete Alonso in 2019 has the worst Batting Average of any player who hit 50 home runs in a season.

 playerID yearID nameFirst nameLast HR BatAvg

1 ruthba01 1921 Babe Ruth 59 0.3777778

2 ruthba01 1920 Babe Ruth 54 0.3763676

3 foxxji01 1932 Jimmie Foxx 58 0.3641026

4 ruthba01 1927 Babe Ruth 60 0.3555556

5 wilsoha01 1930 Hack Wilson 56 0.3555556

6 mantlmi01 1956 Mickey Mantle 52 0.3527205

7 foxxji01 1938 Jimmie Foxx 50 0.3486726

8 bondsba01 2001 Barry Bonds 73 0.3277311

9 sosasa01 2001 Sammy Sosa 64 0.3275563

10 gonzalu01 2001 Luis Gonzalez 57 0.3251232

11 ruthba01 1928 Babe Ruth 54 0.3227612

12 fostege01 1977 George Foster 52 0.3203252

13 sosasa01 2000 Sammy Sosa 50 0.3195364

14 mayswi01 1955 Willie Mays 51 0.3189655

15 rodrial01 2001 Alex Rodriguez 52 0.3180380

16 mayswi01 1965 Willie Mays 52 0.3172043

17 mantlmi01 1961 Mickey Mantle 54 0.3171206

18 belleal01 1995 Albert Belle 50 0.3168498

19 greenha01 1938 Hank Greenberg 58 0.3147482

20 rodrial01 2007 Alex Rodriguez 54 0.3138937

21 kinerra01 1947 Ralph Kiner 51 0.3132743

22 howarry01 2006 Ryan Howard 58 0.3132530

23 mcgwima01 1996 Mark McGwire 52 0.3120567

24 judgeaa01 2022 Aaron Judge 62 0.3105263

25 kinerra01 1949 Ralph Kiner 54 0.3096539

26 sosasa01 1998 Sammy Sosa 66 0.3079316

27 griffke02 1997 Ken Griffey 56 0.3042763

28 thomeji01 2002 Jim Thome 52 0.3041667

29 mizejo01 1947 Johnny Mize 51 0.3020478

30 rodrial01 2002 Alex Rodriguez 57 0.2996795

31 mcgwima01 1998 Mark McGwire 70 0.2986248

32 anderbr01 1996 Brady Anderson 50 0.2970639

33 sosasa01 1999 Sammy Sosa 63 0.2880000

34 fieldpr01 2007 Prince Fielder 50 0.2879581

35 ortizda01 2006 David Ortiz 54 0.2867384

36 davisch02 2013 Chris Davis 53 0.2859589

37 griffke02 1998 Ken Griffey 56 0.2843602

38 judgeaa01 2017 Aaron Judge 52 0.2841328

39 olsonma02 2023 Matt Olson 54 0.2828947

40 stantmi03 2017 Giancarlo Stanton 59 0.2814070

41 mcgwima01 1999 Mark McGwire 65 0.2783109

42 fieldce01 1990 Cecil Fielder 51 0.2774869

43 vaughgr01 1998 Greg Vaughn 50 0.2722513

44 marisro01 1961 Roger Maris 61 0.2694915

45 jonesan01 2005 Andruw Jones 51 0.2627986

46 bautijo02 2010 Jose Bautista 54 0.2601054

47 alonspe01 2019 Pete Alonso 53 0.2596315

Problem 4, Chapter 5 exercises:

1. The oldest plane that flew from NYC in 2013 has tailnum N381AA and is from 1956.
2. Using n\_distinct we see the total number of distinct planes flying from NYC is 3322.