Practice Set 2

Ziji Zhou

Due by 10pm ET on Friday

Practice Set Information

During the week, you will get further practice with the material by working through the Practice Set, a set of problems designed to give you practice beyond the examples produced in the text.

You may work through these problems with peers, but all work must be completed by you (see the Honor Code in the syllabus) and you must indicate who you worked with below.

Even then, the best approach here is to try the problems on your own before discussing them with peers, and then write your final solutions yourself.

GitHub Workflow

1. Before editing this file, verify you are working on the copy saved in *your* repo for the course (check the filepath and the project name in the top right corner).
2. Before editing this file, make an initial commit of the file to your repo to add your copy of the problem set.
3. Change your name at the top of the file and get started!
4. You should *save, knit, and commit* the .Rmd file each time you've finished a question, if not more often. You should also *push* your commits back onto GitHub occasionally (you can do this after each commit).
5. When you think you are done with the assignment, save the pdf as "*Name_thisfilename_date*.pdf" before committing and pushing (this is generally good practice but also helps me in those times where I need to download all student homework files).

Gradescope Upload

For each question (e.g., 3.1), allocate all pages associated with the specific question. If your work for a question runs onto a page that you did not select, you may not get credit for the work. If you do not allocate *any* pages when you upload your pdf, you may get a zero for the assignment.

You can resubmit your work as many times as you want before the deadline, so you should not wait until the last minute to submit some version of your work. Unexpected delays/crises that occur on the day the assignment is due do not warrant extensions (please submit whatever you have done to receive partial credit).

Practicing Academic Integrity

If you worked with others or used resources outside of provided course material (notes, textbook, etc) to complete this assignment, please acknowledge them below using a bulleted list.

I acknowledge the following individuals with whom I worked on this assignment:

Name(s) and corresponding problem(s)

-

I used the following sources to help complete this assignment:

Source(s) and corresponding problem(s)

- Problem 3: https://www.dummies.com/programming/r/how-to-create-a-data-frame-from-scratch-in-r/
- Problem 2: https://stackoverflow.com/questions/38008863/how-to-draw-a-nice-arrow-in-ggplot2/61383034
- Problem 2: https://ggplot2.tidyverse.org/reference/annotate.html
- Problem 3: https://www.youtube.com/watch?v=gnUgSkKEW5c

Problem 1 **MDSR 5.2** Use the `Batting`, `Pitching`, and `Master` tables in the **Lahman** package to answer the following questions.

1.1 List the name of every player in baseball history who has accumulated at least 300 home runs (HR) and at least 300 stolen bases (SB). You can find the first and last name of the player in the `Master` data frame. Join this to your result along with the total home runs and total bases stolen for each of these elite players.

```
mlb_HR_SB <- Batting %>%
  group_by(playerID) %>%
  summarise(HR_total = sum(HR), SB_total = sum(SB))%>%
  filter(HR_total >= 300 & SB_total >= 300) %>%
  inner_join(Master, by = c("playerID" = "playerID")) %>%
  select(nameFirst,nameLast,playerID,HR_total,SB_total)
kable(mlb_HR_SB, booktabs = TRUE)
```

nameFirst	nameLast	playerID	HR_total	SB_total
Carlos	Beltran	beltrca01	435	312
Barry	Bonds	bondsba01	762	514
Bobby	Bonds	bondsbo01	332	461
Andre	Dawson	dawsoan01	438	314
Steve	Finley	finlest01	304	320
Willie	Mays	mayswi01	660	338
Alex	Rodriguez	rodrial01	696	329
Reggie	Sanders	sandere02	305	304

1.2 Similarly, list the names every pitcher in baseball history who has accumulated at least 300 wins (W) and at least 3,000 strikeouts (SO).

```
mlb_W_SO <- Pitching %>%
  group_by(playerID) %>%
  summarise(W_total = sum(W), SO_total = sum(SO))%>%
  filter(W_total >= 300 & SO_total >= 3000) %>%
  inner_join(Master, by = c("playerID" = "playerID")) %>%
  select(nameFirst,nameLast,playerID,W_total,SO_total) %>%
  arrange(desc(W_total))
kable(mlb_W_SO, booktabs = TRUE)
```

nameFirst	nameLast	playerID	W_total	SO_total
Walter	Johnson	johnswa01	417	3509
Greg	Maddux	maddugr01	355	3371
Roger	Clemens	clemero02	354	4672
Steve	Carlton	carltst01	329	4136
Nolan	Ryan	ryanno01	324	5714
Don	Sutton	suttodo01	324	3574
Phil	Niekro	niekrph01	318	3342
Gaylord	Perry	perryga01	314	3534
Tom	Seaver	seaveto01	311	3640
Randy	Johnson	johnsra05	303	4875

1.3 Finally, list the name and year of every player who has hit at least 50 home runs in a single season. Which player had the lowest batting average in that season? Note: Batting average is calculated as the number of hits (H) divided by the number of at bats (AB).

```
mlb_HR_BA <- Batting %>%
  filter(HR >= 50) %>%
  inner_join(Master, by = c("playerID" = "playerID")) %>%
  mutate(battingAverage = H/AB) %>%
  select(nameFirst, nameLast, playerID, yearID, battingAverage, HR) %>%
  arrange(desc(battingAverage))
kable(mlb_HR_BA, booktabs = TRUE)
```

nameFirst	nameLast	playerID	yearID	battingAverage	HR
Babe	Ruth	ruthba01	1921	0.3777778	59
Babe	Ruth	ruthba01	1920	0.3763676	54
Jimmie	Foxx	foxxji01	1932	0.3641026	58
Babe	Ruth	ruthba01	1927	0.3555556	60
Hack	Wilson	wilsoha01	1930	0.3555556	56
Mickey	Mantle	mantlmi01	1956	0.3527205	52
Jimmie	Foxx	foxxji01	1938	0.3486726	50
Barry	Bonds	bondsba01	2001	0.3277311	73
Sammy	Sosa	sosasa01	2001	0.3275563	64
Luis	Gonzalez	gonzalu01	2001	0.3251232	57
Babe	Ruth	ruthba01	1928	0.3227612	54
George	Foster	fostege01	1977	0.3203252	52
Sammy	Sosa	sosasa01	2000	0.3195364	50
Willie	Mays	mayswi01	1955	0.3189655	51
Alex	Rodriguez	rodrial01	2001	0.3180380	52
Willie	Mays	mayswi01	1965	0.3172043	52
Mickey	Mantle	mantlmi01	1961	0.3171206	54
Albert	Belle	belleal01	1995	0.3168498	50
Hank	Greenberg	greenha01	1938	0.3147482	58
Alex	Rodriguez	rodrial01	2007	0.3138937	54
Ralph	Kiner	kinerra01	1947	0.3132743	51
Ryan	Howard	howarry01	2006	0.3132530	58
Mark	McGwire	mcgwima01	1996	0.3120567	52
Ralph	Kiner	kinerra01	1949	0.3096539	54
Sammy	Sosa	sosasa01	1998	0.3079316	66
Ken	Griffey	griffke02	1997	0.3042763	56
Jim	Thome	thomeji01	2002	0.3041667	52
Johnny	Mize	mizejo01	1947	0.3020478	51
Alex	Rodriguez	rodrial01	2002	0.2996795	57
Mark	McGwire	mcgwima01	1998	0.2986248	70
Brady	Anderson	anderbr01	1996	0.2970639	50
Sammy	Sosa	sosasa01	1999	0.2880000	63
Prince	Fielder	fieldpr01	2007	0.2879581	50
David	Ortiz	ortizda01	2006	0.2867384	54
Chris	Davis	davisch02	2013	0.2859589	53
Ken	Griffey	griffke02	1998	0.2843602	56
Aaron	Judge	judgeaa01	2017	0.2841328	52
Giancarlo	Stanton	stantmi03	2017	0.2814070	59
Mark	McGwire	mcgwima01	1999	0.2783109	65
Cecil	Fielder	fieldce01	1990	0.2774869	51
Greg	Vaughn	vaughgr01	1998	0.2722513	50
Roger	Maris	marisro01	1961	0.2694915	61
Andruw	Jones	jonesan01	2005	0.2627986	51
Jose	Bautista	bautijo02	2010	0.2601054	54
Pete	Alonso	alonspe01	2019	0.2596315	53

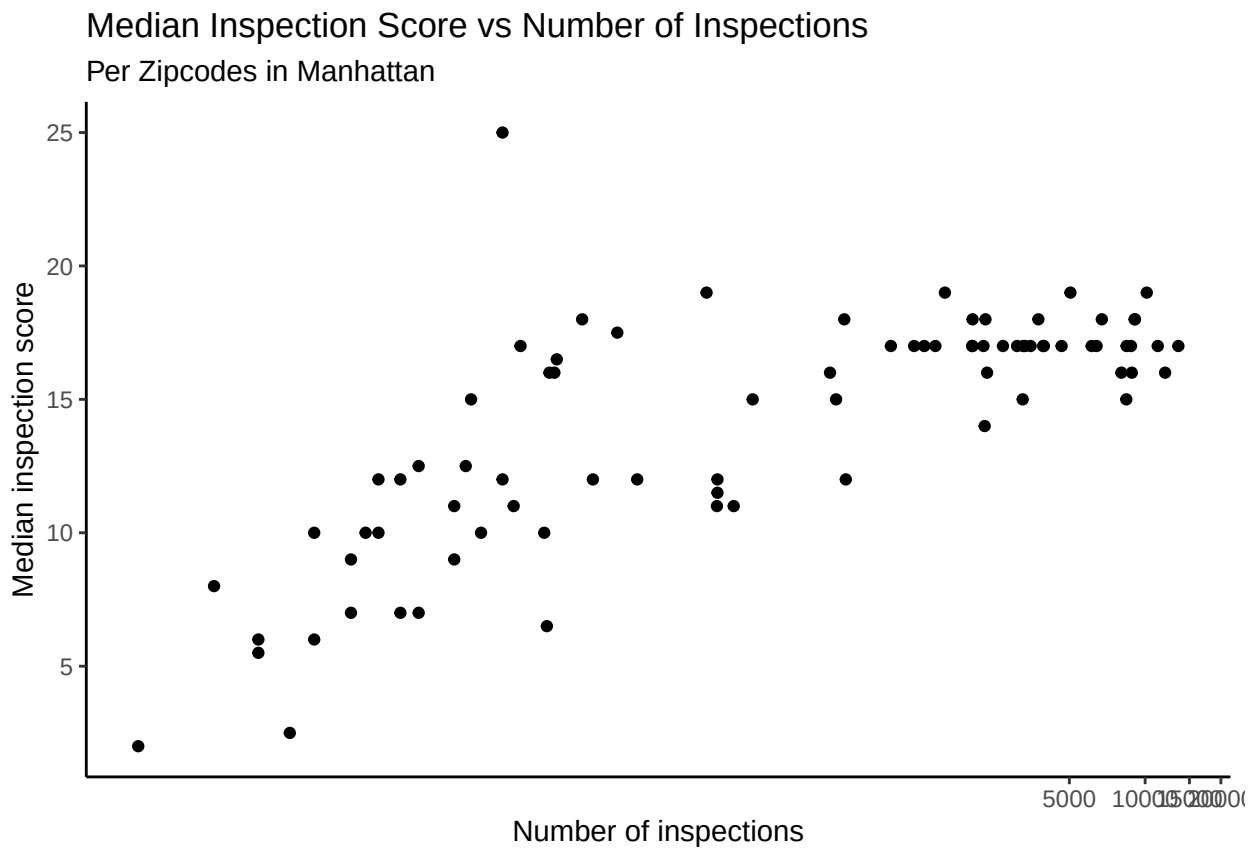
Pete Alonso had the worst batting average of 0.260 for a player hitting 50 or more HR in a season.

Problem 2 **MDSR 4.11 (modified)** The `Violations` data set in the **mdsr** package contains information regarding the outcome of health inspections of restaurants in New York City. Note that higher inspection scores indicate worse violations: "restaurants with an inspection score between 0 and 13 points earn an A, those with 14 to 27 points receive a B and those with 28 or more a C" (nyc.gov).

2.1 Use these data to calculate the median violation score by zip code for zip codes in Manhattan. What pattern, if any, do you see between the number of inspections and the median score? Generate a visualization to support your response.

```
violations_median <- Violations %>%
  filter(boro == "MANHATTAN") %>%
  group_by(zipcode) %>%
  summarize(median_violation = median(score, na.rm = TRUE), inspections = n()) %>%
  arrange(desc(median_violation))

ggplot(violations_median) +
  geom_point(aes(x = inspections,
                 y = median_violation)) +
  coord_trans(x = "log") +
  labs(title = "Median Inspection Score vs Number of Inspections",
       subtitle = "Per Zipcodes in Manhattan",
       x = "Number of inspections",
       y = "Median inspection score")
```

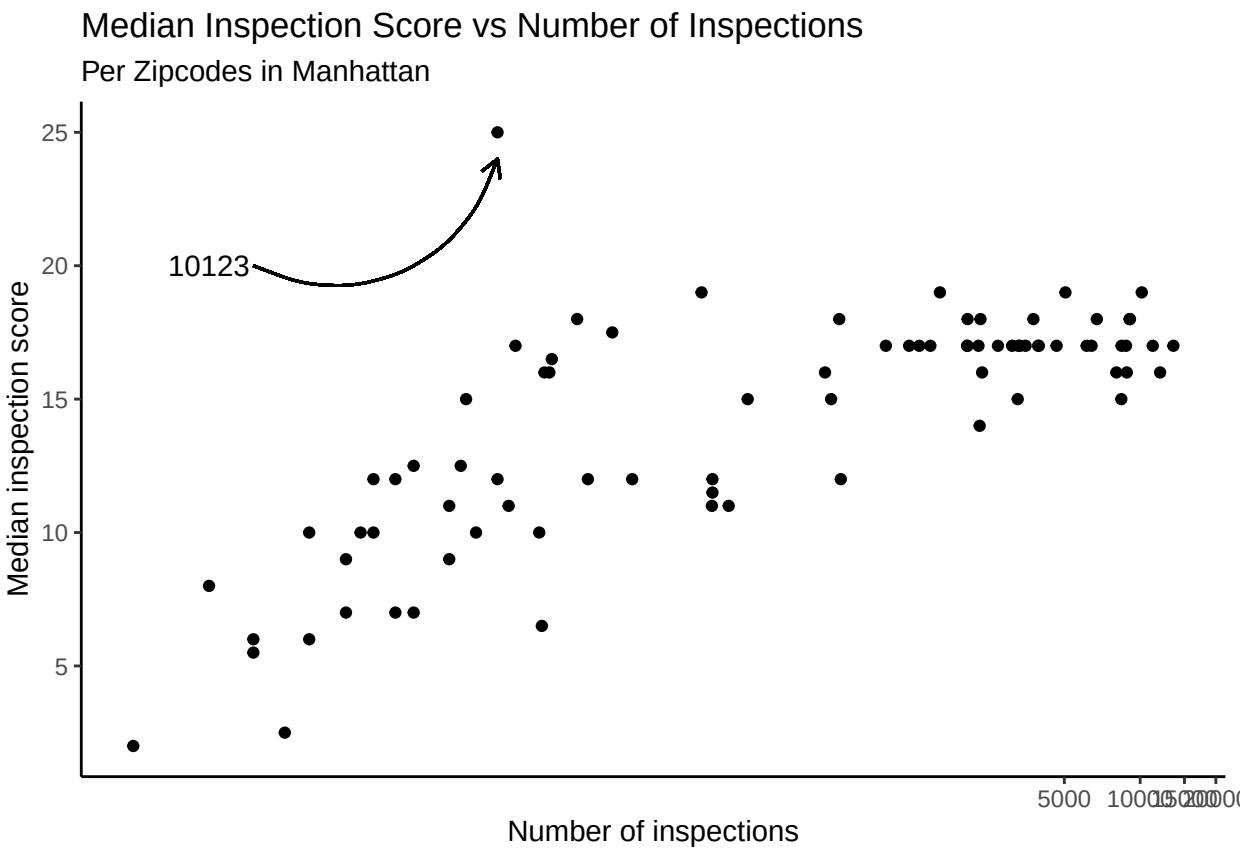


There is a positive correlation between inspections and median inspection score. Do bad scores motivate more inspections? Or do more inspections expose more violations?

2.2 In your visualization above, there are several potential outliers but there is one zipcode in particular that does not seem to fall along the general trend. Add text to the outlier identifying what zipcode it is, and add an arrow pointing from the text to the observation. Note: first, you may want to `filter()` to identify the zipcode (so you know what text to add to the plot).

```
#outliers in 10123
ggplot(violations_median) +
  geom_point(aes(x = inspections,
                 y = median_violation)) +
  coord_trans(x = "log") +
  labs(title = "Median Inspection Score vs Number of Inspections",
       subtitle = "Per Zipcodes in Manhattan",
       x = "Number of inspections",
       y = "Median inspection score") +
  annotate("text", x = 2, y = 20, label = "10123") +
  geom_curve(
  x = 3,
  y = 20,
  xend = 28,
  yend = 24,
```

```
  arrow = arrow(length = unit(0.03, "npc")))
```



Problem 3 **MDSR 6.5** Generate the code to convert the data frame from the starting point (Figure 1) to the results (Figure 2). Hint: use `pivot_longer()` in conjunction with `pivot_wider()`.

grp	sex	meanL	sdL	meanR	sdR
A	F	0.225	0.106	0.340	0.085
A	M	0.470	0.325	0.570	0.325
B	F	0.325	0.106	0.400	0.071
B	M	0.547	0.308	0.647	0.274

Figure 1: Starting point

	grp	F.meanL	F.meanR	F.sdL	F.sdR	M.meanL	M.meanR	M.sdL	M.sdR
1	A	0.22	0.34	0.11	0.08	0.47	0.57	0.33	0.33
2	B	0.33	0.40	0.11	0.07	0.55	0.65	0.31	0.27

Figure 2: Results

```
#create the vectors
grp <- c("A","A","B","B")
sex <- c("F","M","F","M")
meanL <- c(0.225,0.470,0.325,0.547)
sdL <- c(0.106,0.325,0.106,0.308)
meanR <- c(0.340,0.570,0.400,0.647)
sdR <- c(0.085,0.325,0.071,0.274)
data_long <- data.frame(grp,sex,meanL,sdL,meanR,sdR)
kable(data_long, booktabs = TRUE)
```

grp	sex	meanL	sdL	meanR	sdR
A	F	0.225	0.106	0.340	0.085
A	M	0.470	0.325	0.570	0.325
B	F	0.325	0.106	0.400	0.071
B	M	0.547	0.308	0.647	0.274

```
data_wide <- data_long %>%
  pivot_longer(c("meanL","sdL","meanR","sdR"), names_to = "Type of Measurement", values_to = "Number") %
  pivot_wider(names_from = c("sex","Type of Measurement"), values_from = Number, names_sep = ".")
kable(data_wide, booktabs = TRUE, digits = 2)
```

grp	F.meanL	F.sdL	F.meanR	F.sdR	M.meanL	M.sdL	M.meanR	M.sdR
A	0.22	0.11	0.34	0.09	0.47	0.32	0.57	0.32
B	0.32	0.11	0.40	0.07	0.55	0.31	0.65	0.27