STAT 231: Problem Set 2B

Sean Wei

due by 5 PM on Friday, September 11

Series B homework assignments are designed to help you futher ingest and practice the material covered in class over the past week(s). You are encouraged to work with other students, but all code must be written by you and you must indicate below who you discussed the assignment with (if anyone).

Steps to proceed:

1. In RStudio, go to File > Open Project, navigate to the folder with the course-content repo, select the course-content project (course-content.Rproj), and click "Open"
2. Pull the course-content repo (e.g. using the blue-ish down arrow in the Git tab in upper right window)
3. Copy ps2B.Rmd from the course repo to your repo (see page 6 of the GitHub Classroom Guide for Stat231 if needed)
4. Close the course-content repo project in RStudio
5. Open YOUR repo project in RStudio
6. In the ps2B.Rmd file in YOUR repo, replace "YOUR NAME HERE" with your name
7. Add in your responses, committing and pushing to YOUR repo in appropriate places along the way
8. Run "Knit PDF"
9. Upload the pdf to Gradescope. Don't forget to select which of your pages are associated with each problem. *You will not get credit for work on unassigned pages (e.g., if you only selected the first page but your solution spans two pages, you would lose points for any part on the second page that the grader can't see).*

If you discussed this assignment with any of your peers, please list who here:

ANSWER:

MDSR Exercise 4.14 (modified)

Use the `Pitching` data frame from the `Lahman` package to identify every pitcher in baseball history who has accumulated at least 300 wins (`W`) and at least 3,000 strikeouts (`SO`).

a. How many pitchers meet this criteria?

ANSWER: There are 10 players that meet this criteria.

```
library(Lahman)
data(Pitching)
stats <- Pitching %>%
  group_by(playerID) %>%
  summarise(total_wins = sum(W), total_strikeouts = sum(SO)) %>%
  filter(total_wins >= 300 & total_strikeouts >= 3000)
nrow(stats)
```

```
## [1] 10
```

b. Which of these pitchers had the most accumulated strikeouts? How many strikeouts had he accumulated? What is the most strikeouts he had in one season?

ANSWER: The ID of the pitcher who had the most accumulated strikeouts is ryanno01. The total amount of strikeouts he has accumulated is 5,714. The most strikeouts he had in a season was 383 in 1973.

```
#finding the most accumulated strikeouts
stats %>%
  arrange(desc(total_strikeouts)) %>%
  head(1)
```

```
## # A tibble: 1 x 3
##   playerID total_wins total_strikeouts
##   <chr>         <int>            <int>
## 1 ryanno01        324             5714
```

```
#finding out the most strikeouts in a season
ryan_stats <- filter(Pitching, playerID == "ryanno01")
ryan_stats %>%
  arrange(desc(SO)) %>%
  head(1) %>%
  select(playerID, yearID, SO)
```

```
##   playerID yearID  SO
## 1 ryanno01   1973 383
```

MDSR Exercise 4.17 (modified)

a. The Violations data set in the `mdsr` package contains information regarding the outcome of health inspections in New York City. Use these data to calculate the median violation score by zipcode and dba for zipcodes 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.

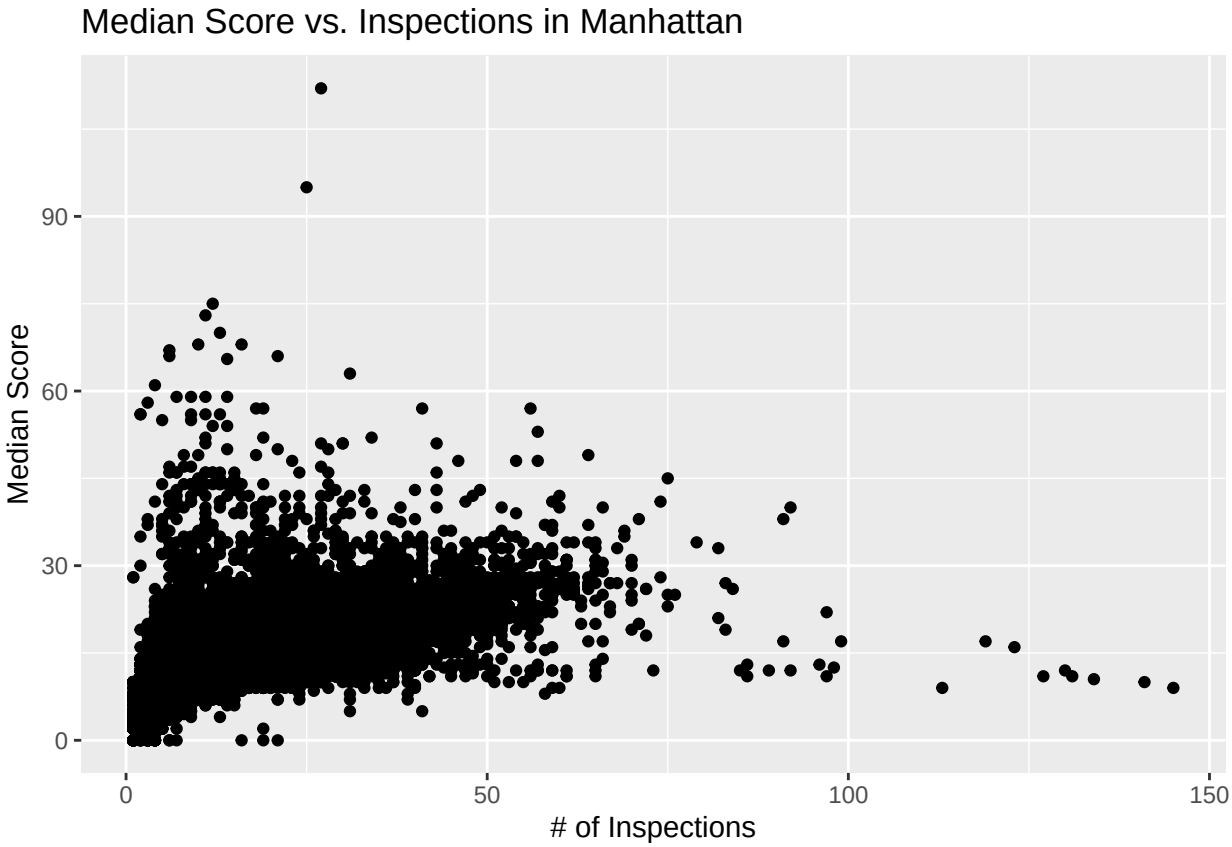
ANSWER: Looking at the scatterplot below, there does not seem to be any clear pattern between the number of inspections and the median score. Although, it could be argued that there might be some positive association between the number of inspections and median score.

```
data(Violations)

#creating appropriate data table
median_scores <- Violations %>%
  filter(is.na(score) == FALSE & boro == "MANHATTAN") %>%
  group_by(zipcode, dba) %>%
  summarise(median_score = median(score), inspections = n())
median_scores
```

```
## # A tibble: 9,359 x 4
## # Groups:   zipcode [81]
##    zipcode dba                          median_score inspections
##      <int> <chr>                               <dbl>       <int>
##  1   10001 10TH AVENUE PIZZA & CAFE               17          30
##  2   10001 16 HANDLES                              2           3
##  3   10001 230 FIFTH                              23          29
##  4   10001 33 Gourmet                             26          49
##  5   10001 35 DUET                                 9          11
##  6   10001 5 SENSES                               32           7
##  7   10001 5BAR KARAOKE                           31          22
##  8   10001 7 GRAMS CAFFE                           5           5
##  9   10001 876 MARKET DELI                        15          22
## 10   10001 99 CENTS BEST & FRESH PIZZA            11          12
## # ... with 9,349 more rows
```

```
#create scatterplot
ggplot(median_scores, aes(x = inspections, y = median_score)) +
  geom_point() +
  labs(x = "# of Inspections", y = "Median Score") +
  ggtitle("Median Score vs. Inspections in Manhattan")
```



b. In your visualization in part (a), there should be at least a few points that stand out as outliers. For *one of the outliers*, add text to the outlier identifying what business it is and an arrow pointing from the text to the observation. First, you may want to `filter` to identify the name of the business (so you know what text to add to the plot).

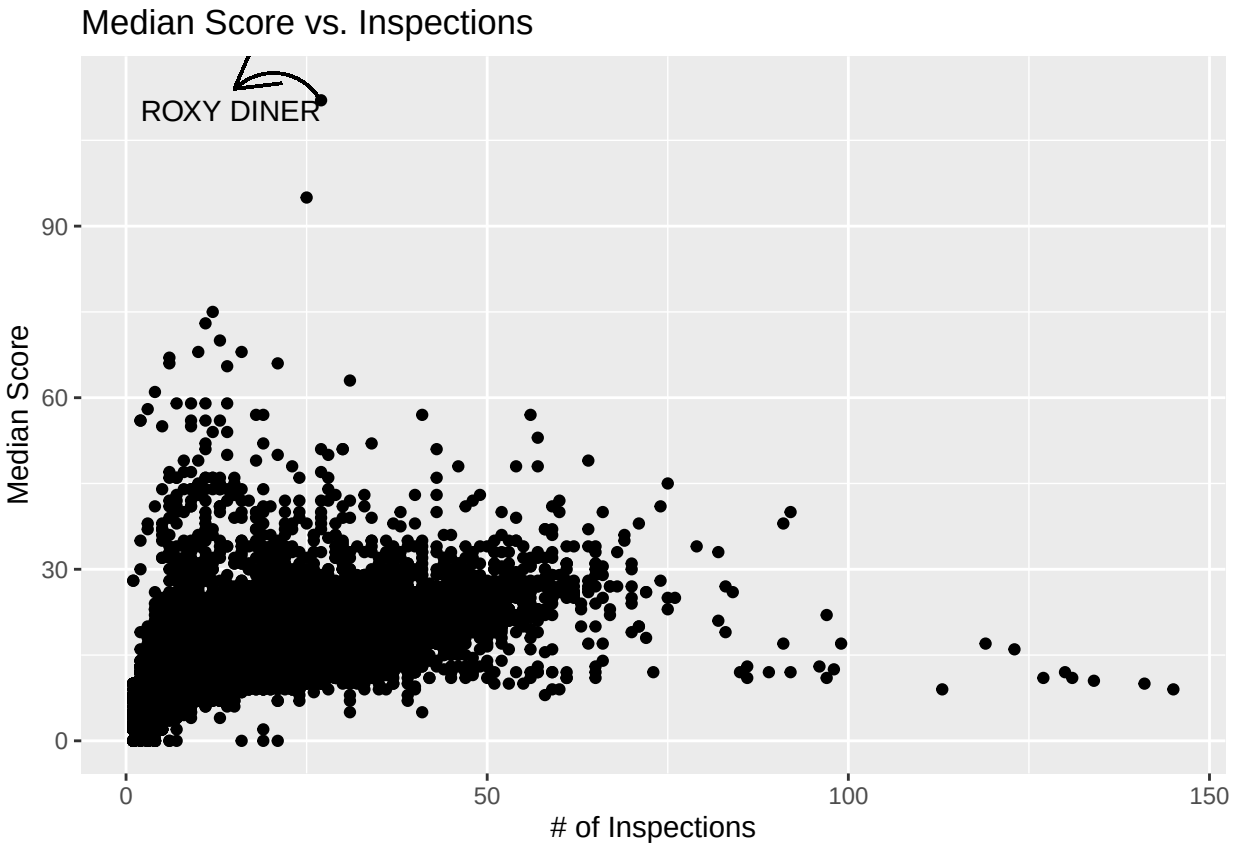
(Can't remember how to create a curved arrow in `ggplot`? Can't remember how to add text to the plot in `ggplot`? Check out the answers to questions #5 and #8, respectively, in the Moodle R Q&A forum!)

```
#want to get the point of the highest median score
outlier <- median_scores %>%
  arrange(desc(median_score)) %>%
  head(1)
outlier
```

```
## # A tibble: 1 x 4
## # Groups:   zipcode [1]
##   zipcode dba        median_score inspections
##     <int> <chr>             <dbl>       <int>
## 1   10036 ROXY DINER          112          27
```

```
#create scatterplot w/ arrow and name
ggplot(median_scores, aes(x = inspections, y = median_score)) +
  geom_point() +
  geom_text(aes(label=ifelse(median_score==112, as.character(outlier$dba),'')),
            hjust = 1, vjust=1) +
```

```
geom_curve(aes(x = 27, y = 112, xend = 15, yend = 114), arrow = arrow()) +
labs(x = "# of Inspections", y = "Median Score") +
ggtitle("Median Score vs. Inspections")
```



MDSR Exercise 5.7

Generate the code to convert the data frame shown with this problem in the textbook (on page 130) to wide format (e.g. see result table). Hint: use `gather()` in conjuction with `spread()`; OR `pivot_longer()` in conjuction with `pivot_wider()`.

```
FakeDataLong <- data.frame(grp = c("A","A","B", "B")
                           , sex = c("F", "M", "F", "M")
                           , meanL = c(0.22, 0.47, 0.33, 0.55)
                           , sdL = c(0.11, 0.33, 0.11, 0.31)
                           , meanR = c(0.34, 0.57, 0.40, 0.65)
                           , sdR = c(0.08, 0.33, 0.07, 0.27))
Data_Wide <- pivot_wider(FakeDataLong, names_from = "sex",
                         values_from = c("meanL", "sdL", "meanR", "sdR"))
Data_Wide
```

```
## # A tibble: 2 x 9
##   grp   meanL_F meanL_M sdL_F sdL_M meanR_F meanR_M sdR_F sdR_M
##   <fct>   <dbl>   <dbl> <dbl> <dbl>   <dbl>   <dbl> <dbl> <dbl>
## 1 A        0.22    0.47  0.11  0.33    0.34   0.570  0.08  0.33
## 2 B        0.33    0.55  0.11  0.31    0.4    0.65   0.07  0.27
```

PUG Post

What topics or questions are you interested in exploring related to your PUG theme? Dream big here. Don't worry about whether there is data out there that's available and accessible that you could use to address your questions/topics. Just brainstorm some ideas that get you excited. In your PUG team discussion forum on GitHub, start a thread called "Brainstorming" (or, if another team member has already started the thread, reply to their post) with your ideas.

> ANSWER: Do not write anything here. Write down your ideas in your PUG team's discussion thread titled "Brainstorming" on GitHub.