STAT 231: Problem Set 2B

Jack Dove

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: 10 pitchers.

```
library(Lahman)
Pitching1<- Pitching %>%
  select(playerID, yearID, W, SO)

Pitching1 %>%
  group_by(playerID) %>%
  summarize(
    totalW = sum(W), totalSO = sum(SO)
  ) %>%
  filter(totalW >299 & totalSO > 2999) %>%
  arrange(desc(totalSO))
```

```
## # A tibble: 10 x 3
##    playerID  totalW totalSO
##    <chr>      <int>   <int>
##  1 ryanno01     324    5714
##  2 johnsra05    303    4875
##  3 clemero02    354    4672
##  4 carltst01    329    4136
##  5 seaveto01    311    3640
##  6 suttodo01    324    3574
##  7 perryga01    314    3534
##  8 johnswa01    417    3509
##  9 maddugr01    355    3371
## 10 niekrph01    318    3342
```

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: Nolan Ryan accumulated 5714 total strikeouts. In 1973, he struck out (wow) 383 batters.

```
Pitching2 <- Pitching1 %>%
  filter(playerID =="ryanno01") %>%
  arrange(desc(SO))
Pitching2
```

```
##    playerID yearID  W  SO
## 1  ryanno01   1973 21 383
## 2  ryanno01   1974 22 367
## 3  ryanno01   1977 19 341
## 4  ryanno01   1972 19 329
```

```
## 5  ryanno01   1976 17 327
## 6  ryanno01   1989 16 301
## 7  ryanno01   1987  8 270
## 8  ryanno01   1978 10 260
## 9  ryanno01   1982 16 245
## 10 ryanno01   1990 13 232
## 11 ryanno01   1988 12 228
## 12 ryanno01   1979 16 223
## 13 ryanno01   1985 10 209
## 14 ryanno01   1991 12 203
## 15 ryanno01   1980 11 200
## 16 ryanno01   1984 12 197
## 17 ryanno01   1986 12 194
## 18 ryanno01   1975 14 186
## 19 ryanno01   1983 14 183
## 20 ryanno01   1992  5 157
## 21 ryanno01   1981 11 140
## 22 ryanno01   1971 10 137
## 23 ryanno01   1968  6 133
## 24 ryanno01   1970  7 125
## 25 ryanno01   1969  6  92
## 26 ryanno01   1993  5  46
## 27 ryanno01   1966  0   6
```

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: I see a medium positive relationship between number of inspections and median violation score by zipcode and by DBA. I generated a dual-facet smoothed scatterplot, which, using the log of inspections, shows an initial strong positive relationship between the two variables, then tapers off at about 9 total inspections. These results are also supported by moderate correlation coefficients, both between 0.4-0.6 and suggesting a medium link between log n inspections and median score for both categories.

```
data(Violations)
head(Violations, 10)
```

```
## # A tibble: 10 x 16
##     camis dba   boro  building street zipcode  phone inspection_date     action
##     <int> <chr> <chr>    <int> <chr>    <int>  <dbl> <dttm>              <chr>
##  1 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2015-02-09 00:00:00 Viola~
##  2 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2014-03-03 00:00:00 Viola~
##  3 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-10-10 00:00:00 No vi~
##  4 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-09-11 00:00:00 Viola~
##  5 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-09-11 00:00:00 Viola~
##  6 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-08-14 00:00:00 Viola~
##  7 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-08-14 00:00:00 Viola~
##  8 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-08-14 00:00:00 Viola~
##  9 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-08-14 00:00:00 Viola~
## 10 3.01e7 MORR~ BRONX     1007 MORRI~   10462 7.19e9 2013-08-14 00:00:00 Viola~
## # ... with 7 more variables: violation_code <chr>, score <int>, grade <chr>,
## #   grade_date <dttm>, record_date <dttm>, inspection_type <chr>,
## #   cuisine_code <dbl>
```

```
#By Zipcode
Violations1 <- Violations %>%
  filter(is.na(score)==FALSE) %>%
  group_by(zipcode) %>%
  summarise(medianscore = median(score), ninspections = log(n()))
Violations1
```

```
## # A tibble: 229 x 3
##    zipcode medianscore ninspections
##      <int>       <dbl>        <dbl>
##  1   10001          15         8.98
##  2   10002          18         9.04
##  3   10003          17         9.44
##  4   10004          14         7.68
##  5   10005          17         7.04
##  6   10006          17         6.83
##  7   10007          16         7.69
##  8   10009          17         8.63
```

```
##  9   10010          17         8.39
## 10   10011          17         9.01
## # ... with 219 more rows
```

```
plot1<-ggplot(data=Violations1) +
  geom_point(mapping = aes(x = ninspections, y = medianscore)) +
  geom_smooth(mapping= aes(x = ninspections, y = medianscore), type = "loess") +
  labs(
    x="Log Number of Inspections",
    y="Median Violation Score",
    title= "NYC DBAs"
)
```

```
## Warning: Ignoring unknown parameters: type
```

```
#By DBA
Violations2 <- Violations %>%
  filter(is.na(score)==FALSE) %>%
  group_by(dba) %>%
  summarise(medianscore = median(score), ninspections = log(n()))
Violations2
```

```
## # A tibble: 19,758 x 3
##    dba                                                medianscore ninspections
##    <chr>                                                    <dbl>        <dbl>
##  1 ''W'' CAFE                                                  22         3.14
##  2 (LEWIS DRUG STORE) LOCANDA VINI E OLII                      20         2.83
##  3 (LIBRARY)  FOUR & TWENTY BLACKBIRDS                          9         2.20
##  4 (PUBLIC FARE) 81st street and central park west (De~        19         2.94
##  5 @NINE                                                       14         3.91
##  6 / L'ECOLE                                                   19         2.71
##  7 #1 GARDEN CHINESE                                           21         3.18
##  8 #1 SABOR LATINO RESTAURANT                                  21         3.66
##  9 $1 PIZZA $2 BEER                                            17         3.69
## 10 1 2 3 BURGER SHOT BEER                                      20         2.89
## # ... with 19,748 more rows
```

```
plot2 <- ggplot(data=Violations2) +
  geom_point(mapping = aes(x = ninspections, y = medianscore)) +
  geom_smooth(mapping= aes(x = ninspections, y = medianscore), type = "loess") +
  labs(
    x="Log Number of Inspections",
    y="Median Violation Score",
    title= "NYC Zipcodes"
)
```

```
## Warning: Ignoring unknown parameters: type
```

```
library(cowplot)
```

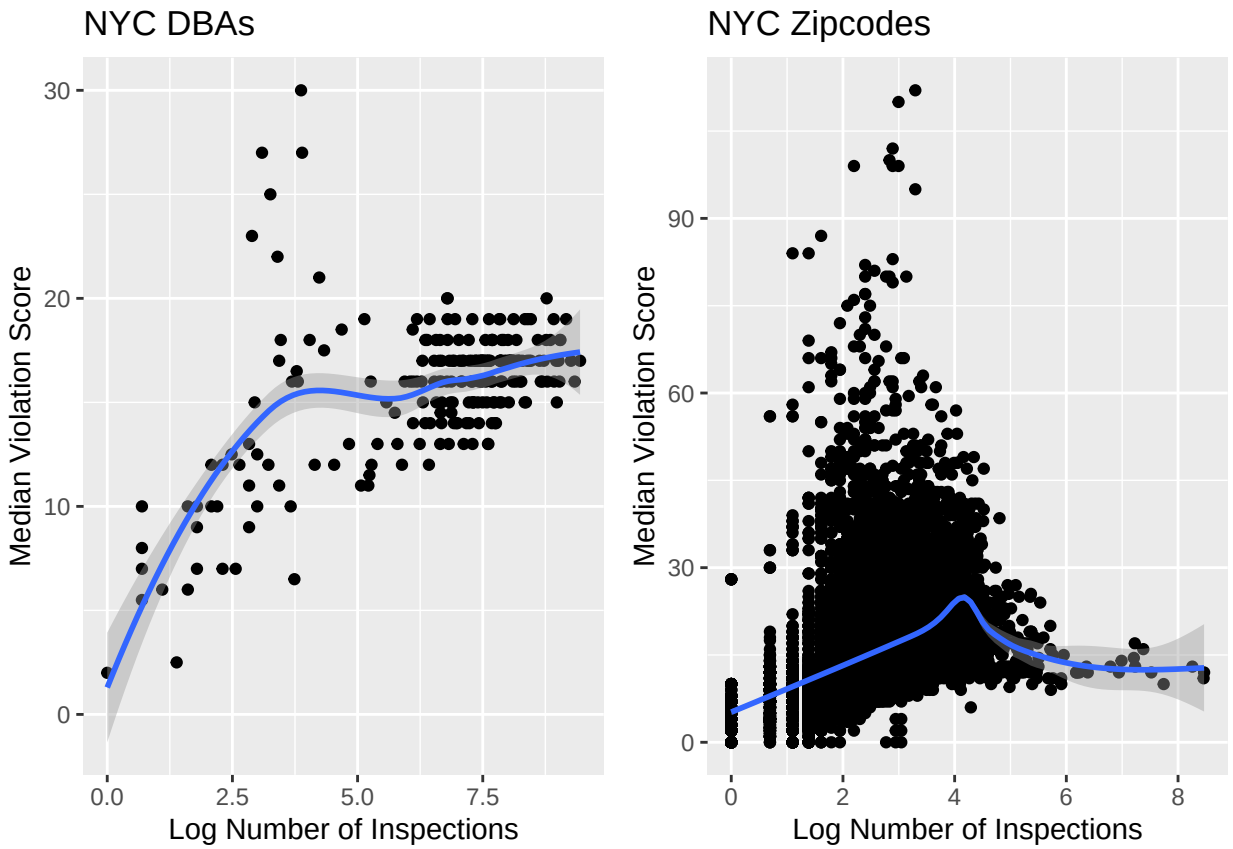
```
##
## Attaching package: 'cowplot'
```

```
## The following object is masked from 'package:mosaic':
##
##     theme_map
```

```
#Put the charts together
plot_grid(plot1, plot2)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
cor(medianscore~ninspections, data=Violations1)
```

```
## [1] 0.5264116
```

```
cor(medianscore~ninspections, data=Violations2)
```

```
## [1] 0.4604565
```

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).

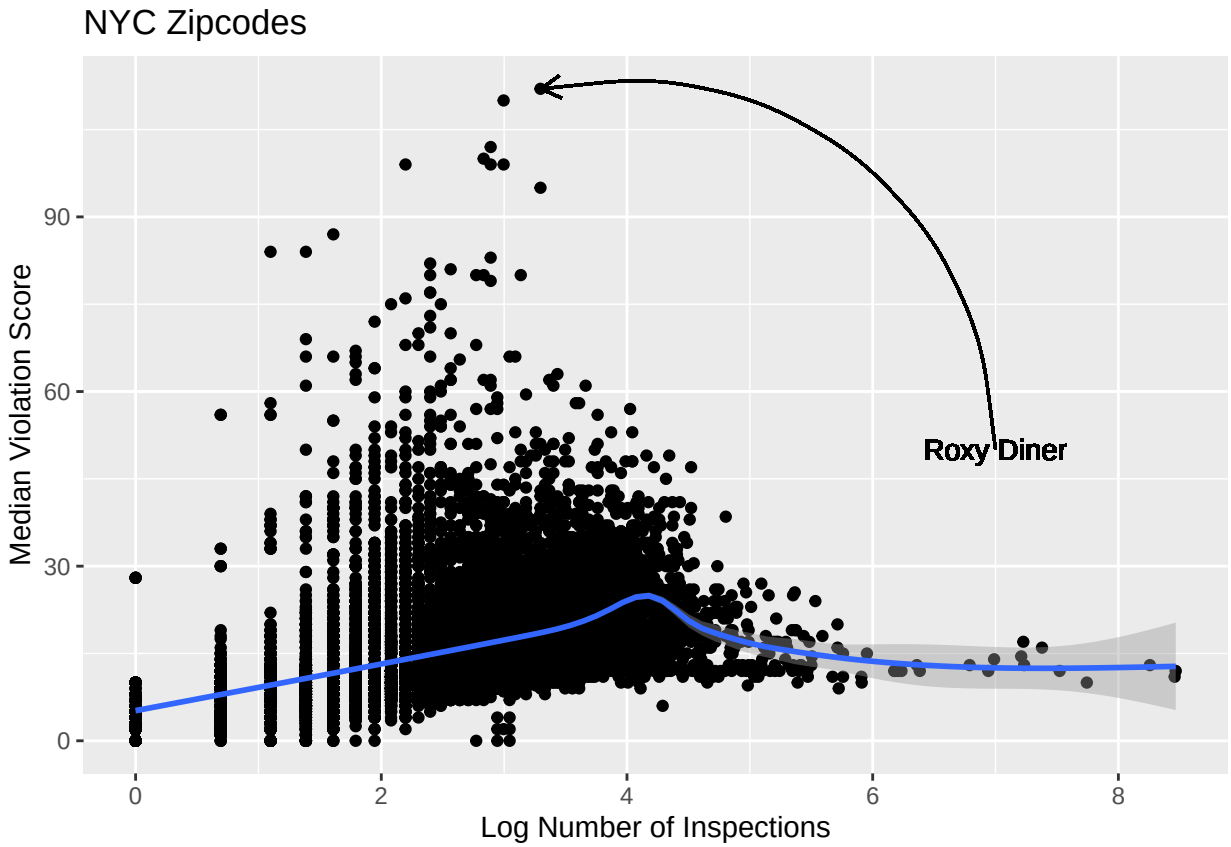
(Can't remember how to create a curved arrow in `ggplot`? Can't remember how to add text to the plot in `ggplot`? Check out the answers to questions #5 and #8, respectively, in the Moodle R Q&A forum!)

```
outlier <- Violations2 %>%
  filter(medianscore > 110)
outlier
```

```
## # A tibble: 1 x 3
##   dba        medianscore ninspections
##   <chr>            <dbl>        <dbl>
## 1 ROXY DINER         112         3.30
```

```
plot2 + geom_curve(x = 7, xend = 3.3, y = 50, yend = 112, arrow = arrow(length = unit(0.3,"cm")), curva
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



MDSR Exercise 5.7

Generate the code to convert the data frame shown with this problem in the textbook (on page 130, and shown below) to wide format (i.e., the 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))
# Looked up unite function on internet
Table <- FakeDataLong %>%
  gather(key = "x", value = "y", meanL, meanR, sdL, sdR) %>%
  unite(col = "x1", sex, x, sep = ".") %>%
  spread(x1, y)
Table
```

```
##   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
```

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.