MSCS 264: Exam 2 Key

Take-Home Exam Guidelines.

By signing the Pledge below, you certify that you conformed to the following guidelines for this take-home exam:

- You may use all materials from this class (textbook, class notes, Moodle posts, Rmd files, answer keys, etc.). In addition, google searches are okay; if your search takes you to outside sources that you use, please list the websites' URLs here:
-
- Obviously, no consulting with anyone else, either in our class or not, and either in-person or electronically (e.g. no posting questions online). Avoid even comments like "#1 is hard".
- Questions to Prof Roback are allowed – I may not be able to answer everything, but I'll answer what I can.
- If you have any questions about what is appropriate, please ask!

Exams are due **before class** on Tuesday, Nov 13th. (NO exceptions unless cleared with me before exams passed out.) You should submit a knitted pdf file on Moodle, but be sure to show all of your R code, in addition to your output, plots, and written responses.

PLEDGE: By typing my full name below, I pledge on my honor that I have neither received nor given assistance during this exam nor have I witnessed others receiving assistance, and I have followed the guidelines as described above.

SIGNATURE: (type full name)

We will focus on data from APM Reports podcast series "In the Dark" about a controversial Mississippi death penalty case and potential racial discrimination in jury selection. This link provides background about the case in question and the analysis of jury selection data over 25 years in Mississippi's Fifth Court District. This link provides descriptions of the 3 data sets and their variables that we will be examining in this exam. The 3 data sets (jurors.csv, trials.csv, and voir_dire_answers.csv) have been downloaded to the Class > Data folder on the R server.

1. There are two trials in voir_dire_answers.csv that contain data on only a single juror: trials 78 and 270. We wish to examine these jurors more closely. First, create the following tibble containing data from the two jurors in trials 78 and 270. Note that I renamed `juror_id__trial__id` as `trial_id`, `fam_law_enforcement` as `fam_law_enf`, and `death_hesitation` as `death_hes`.

```
# A tibble: 2 x 7
  juror_id trial_id accused fam_accused know_def fam_law_enf death_hes
     <int>    <int> <lgl>   <lgl>       <lgl>    <lgl>       <lgl>
1     3842       78 FALSE   FALSE       FALSE    FALSE       FALSE
2    13121      270 FALSE   FALSE       FALSE    FALSE       FALSE
```

Next, create a longer version of that same data:

```
# A tibble: 10 x 4
   juror_id trial_id question    answer
      <int>    <int> <chr>       <lgl>
 1     3842       78 accused     FALSE
 2    13121      270 accused     FALSE
 3     3842       78 fam_accused FALSE
 4    13121      270 fam_accused FALSE
```

```
 5     3842       78 know_def    FALSE
 6    13121      270 know_def    FALSE
 7     3842       78 fam_law_enf FALSE
 8    13121      270 fam_law_enf FALSE
 9     3842       78 death_hes   FALSE
10    13121      270 death_hes   FALSE
```

and then return the longer data set back into its original form:

```
# A tibble: 2 x 7
  juror_id trial_id accused fam_accused know_def fam_law_enf death_hes
     <int>    <int> <lgl>   <lgl>       <lgl>    <lgl>       <lgl>
1     3842       78 FALSE   FALSE       FALSE    FALSE       FALSE
2    13121      270 FALSE   FALSE       FALSE    FALSE       FALSE
```

Finally, form the following tibble for the two jurors from trials 78 and 270. Note the following features:

- `numTrue` is the total number of TRUE responses to the 5 questions in the previous tibble
- the variables `year`, `casenum`, and `defendant` all come from the `trial` variable in jurors.csv
- you can strip off anything in square brackets in the case number portion of `trial`. `str_sub` can be helpful here.
- `year` is double precision

```
# A tibble: 2 x 6
  juror_id trial_id  year casenum defendant    numTrue
     <int>    <int> <dbl> <chr>   <chr>          <int>
1     3842       78  1995 7009    Ricky Lenard       0
2    13121      270  1992 4419    Jerry Holmes       0
```

```
# Q1

twojurors <- voirdire %>%
  rename(trial_id = juror_id__trial__id,
         fam_law_enf = fam_law_enforcement,
         death_hes = death_hesitation) %>%
  filter(trial_id == 78 | trial_id == 270) %>%
  select(juror_id, trial_id, accused, fam_accused, know_def, fam_law_enf, death_hes)
twojurors
```

```
## # A tibble: 2 x 7
##   juror_id trial_id accused fam_accused know_def fam_law_enf death_hes
##      <dbl>    <dbl> <lgl>   <lgl>       <lgl>    <lgl>       <lgl>
## 1     3842       78 FALSE   FALSE       FALSE    FALSE       FALSE
## 2    13121      270 FALSE   FALSE       FALSE    FALSE       FALSE
```

```
twojurors_long <- twojurors %>%
  gather(key = "question", value = "answer", accused:death_hes)
twojurors_long
```

```
## # A tibble: 10 x 4
##    juror_id trial_id question    answer
##       <dbl>    <dbl> <chr>       <lgl>
##  1     3842       78 accused     FALSE
##  2    13121      270 accused     FALSE
##  3     3842       78 fam_accused FALSE
##  4    13121      270 fam_accused FALSE
##  5     3842       78 know_def    FALSE
```

```
##  6    13121      270 know_def    FALSE
##  7     3842       78 fam_law_enf FALSE
##  8    13121      270 fam_law_enf FALSE
##  9     3842       78 death_hes   FALSE
## 10    13121      270 death_hes   FALSE
```

```
twojurors_wide <- twojurors_long %>%
  spread(key = "question", value = "answer")
twojurors_wide
```

```
## # A tibble: 2 x 7
##   juror_id trial_id accused death_hes fam_accused fam_law_enf know_def
##      <dbl>    <dbl> <lgl>   <lgl>     <lgl>       <lgl>       <lgl>
## 1     3842       78 FALSE   FALSE     FALSE       FALSE       FALSE
## 2    13121      270 FALSE   FALSE     FALSE       FALSE       FALSE
```

```
twojurors_wide %>%
  left_join(jurors, by = c("juror_id" = "id")) %>%
  mutate(numTrue = accused + fam_accused + know_def + fam_law_enf + death_hes) %>%
  select(juror_id, trial_id, trial, numTrue) %>%
  separate(trial, into = c("year_num", "defendant"), sep = "--") %>%
  mutate(year_num = str_sub(year_num, 1, 9)) %>%
  separate(year_num, into = c("year", "casenum")) %>%
  mutate(year = parse_number(year))
```

```
## # A tibble: 2 x 6
##   juror_id trial_id  year casenum defendant    numTrue
##      <dbl>    <dbl> <dbl> <chr>   <chr>          <int>
## 1     3842       78  1995 7009    Ricky Lenard       0
## 2    13121      270  1992 4419    Jerry Holmes       0
```

2. voir_dire_answers.csv contains data from 89 trials (after filtering out one row corresponding to a trial id of NA). We wish to find summary statistics and create plots to compare strike rates by the State prosecutor for black and white potential jurors.

a) When tidying your data, be sure to:

- exclude jurors with unknown race
- only include jurors who are eligible to be struck by the State (i.e. `strike_eligibility` is either "Both State and Defense" or "State")
- create a new variable which specifies if a juror's `race` is the same as the `defendant_race`
- create a new variable which specifies if a juror was "Struck by the state" or not (as recorded in `struck_by`)

b) Produce the following summary table:

```
# A tibble: 2 x 4
  race  prop_struck num_struck total
  <chr>       <dbl>      <int> <int>
1 Black       0.534        396   741
2 White       0.114        175  1541
```

c) Produce a segmented (filled) bar chart illustrating the `prop_struck` comparison in the table above.

d) Produce a faceted (side-by-side) bar chart illustrating black vs. white strike rates for potential jurors who are the same race as the defendant and those who are a different race than the defendant.

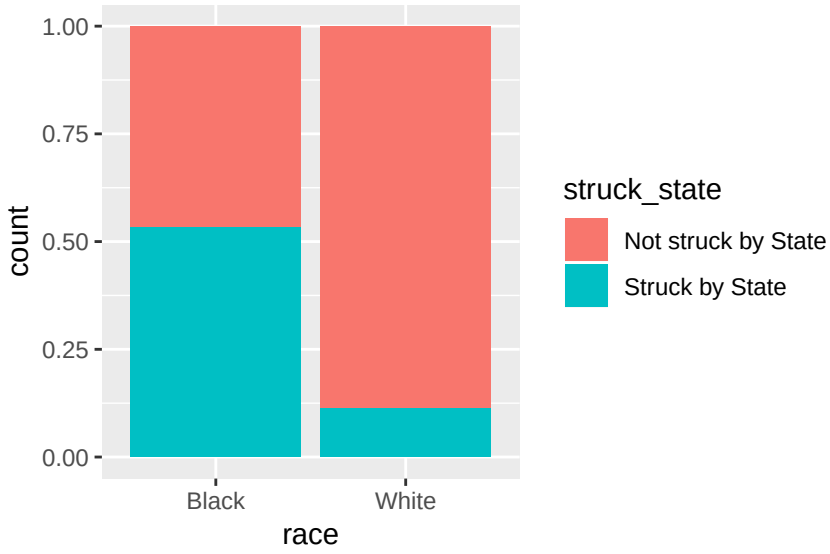
e) Produce a faceted (side-by-side) bar chart illustrating black vs. white strike rates for potential jurors who have been accused of crimes in the past and those who have not.

f) Comment on how your plots in (d) and (e) help illustrate how the black vs. white difference in strike rates persists even after controlling for other factors that may affect strike decisions.

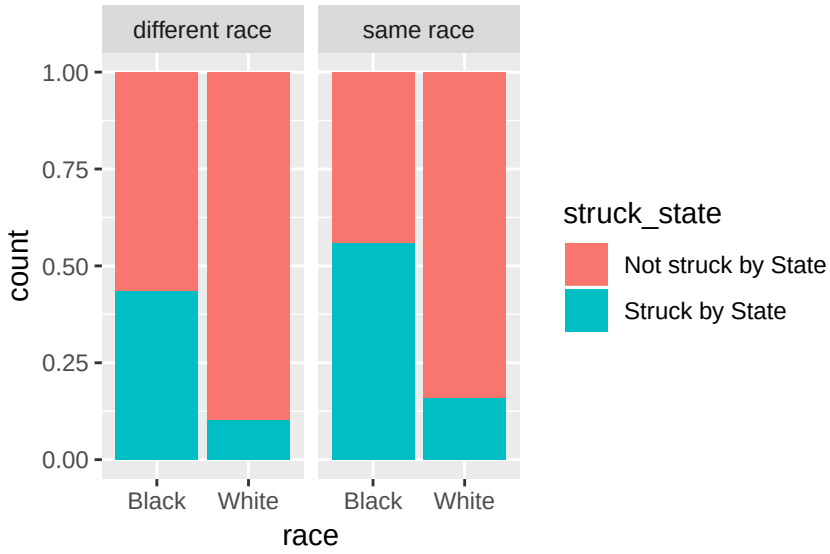
```
# Q2

# 89 trials represented in voirdire (not counting 1 with trial_id = NA),
#   although 2 have n=1, while 305 trials represented in jurors and trials
# print(voirdire %>% count(juror_id__trial__id) %>% arrange(n), n = Inf)

# all data for jurors in 89 trials with complete voir dire transcript
master <- voirdire %>%
  left_join(trials, by = c("juror_id__trial__id" = "id")) %>%
  filter(!is.na(juror_id__trial__id)) %>%
  left_join(jurors, by = c("juror_id" = "id"))

# Smaller data set created to investigate strike rates by the state by race
#   and other confounder variables, some of which are found in voirdire
master_small <- master %>%
  select(juror_id, trial__id, struck_by, race, gender, defendant_race, accused,
         fam_accused, know_def, fam_law_enforcement, death_hesitation,
         strike_eligibility, notes, cause_number) %>%
  filter(race != "Unknown") %>%
  filter(strike_eligibility == "Both State and Defense" |
           strike_eligibility == "State") %>%
  mutate(same_race = ifelse(race == defendant_race,
                            "same race", "different race"),
         struck_state = ifelse(struck_by == "Struck by the state",
                               "Struck by State", "Not struck by State"),
         year = parse_number(str_sub(cause_number)))

master_small %>%
  group_by(race) %>%
  summarise(prop_struck = mean(struck_state == "Struck by State"),
            num_struck = sum(struck_state == "Struck by State"),
            total = n())
```

```
## # A tibble: 2 x 4
##   race  prop_struck num_struck total
##   <chr>       <dbl>      <int> <int>
## 1 Black       0.534        396   741
## 2 White       0.114        175  1541
```

```
ggplot(master_small) +
  geom_bar(aes(x = race, fill = struck_state), position = "fill")
```




```
ggplot(master_small) +
  geom_bar(aes(x = race, fill = struck_state), position = "fill") +
  facet_grid(. ~ same_race)
```



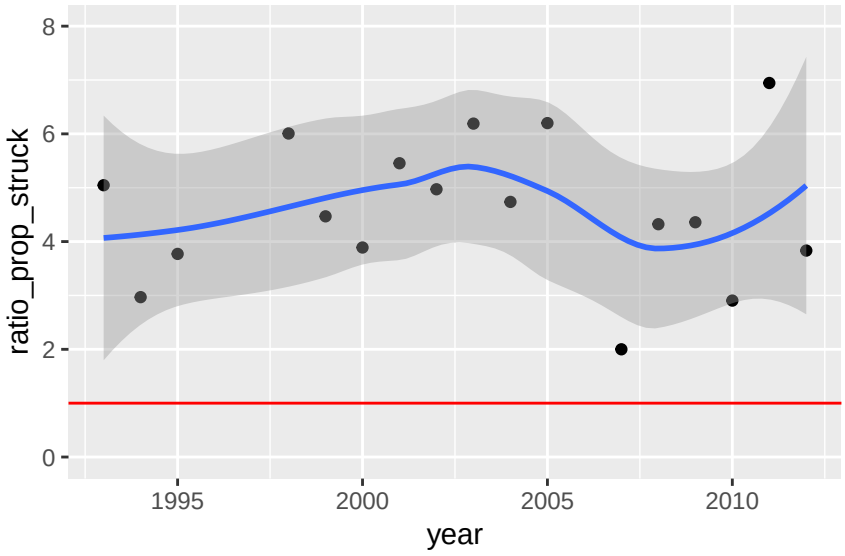

```
ggplot(master_small) +
  geom_bar(aes(x = race, fill = struck_state), position = "fill") +
  facet_grid(. ~ accused)
```

3. Create a plot with year on the x-axis and the ratio of proportion of black jurors struck by the state to the proportion of white jurors struck by the state on the y-axis (e.g. if 60% of black jurors are struck in a given year and 20% of white jurors, the ratio would be 3). When answering this question, be sure to:

- use `spread` if at all possible
- filter out years with ratios above 10, since they tend to overwhelm the plot
- include a red horizontal line where the ratio is 1, which would indicate that black and white jurors are struck at the same rate
- comment on conclusions you can draw from your plot

```
# Q3

master_small %>%
  group_by(year, race) %>%
  summarise(prop_struck = mean(struck_state == "Struck by State")) %>%
  spread(key = "race", value = "prop_struck") %>%
  mutate(ratio_prop_struck = Black / White) %>%
  mutate(ratio_prop_struck = ifelse(ratio_prop_struck < 10,
                                    ratio_prop_struck, NA)) %>%
  ggplot(aes(x = year, y = ratio_prop_struck)) +
    geom_point() +
    geom_smooth() +
    ylim(c(0,8)) +
    geom_hline(yintercept = 1, color = "red")
```

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

## Warning: Removed 2 rows containing non-finite values (stat_smooth).

## Warning: Removed 2 rows containing missing values (geom_point).
```



4.

a) Set `x <- (trials %>% filter(voir_dire_present) %>% distinct(def_attny_1))[[1]]`. This should produce a list of 95 defense attorneys from trials where a jury selection transcript existed. Then use `str_subset` to find attorneys whose names meet the following criteria (do 5 independent searches):

- has same name as one's father (e.g. Jr, II, III, IV, etc.)
- has a middle initial instead of a middle name
- has two repeated letters in their name (e.g. "tt" or "ee" but not "II" or "III"; if a name has both "tt" and "II" that counts)
- has a last name that ends with r or s
- has a first name that starts with 1 vowel but not 2 vowels

b) Create a table of the approximate number of potential jurors in each county who were struck because they were related to the defendant or another potential juror. To do this, search for the terms "relat", "cousin", or "kin" in the `notes` column of voir_dire_answers.csv.

```
# Q4

x <- (trials %>% filter(voir_dire_present) %>% distinct(def_attny_1))[[1]]

str_subset(x, "Jr|II")            # same name as father (Jr, II, III, IV, ...)
```

```
##  [1] "James H. Powell, III"    "Grady F. Tollison, Jr."
##  [3] "Richard Carter, III"     "Bennie L. Jones, Jr."
##  [5] "Joey Hood, II"           "Mitchell M. Lundy Jr."
##  [7] "R.T. Laster, Jr."        "Robert T. Laster, Jr."
##  [9] "Johnnie E. Walls, Jr."   "Ross R. Barnett, Jr"
## [11] "Bernard C. Jones, Jr."   "H. Lee Bailey, Jr."
## [13] "Hugh Lee Bailey, Jr."    "Thomas M. Flanagan, Jr."
## [15] "Edwin A. Flint, Jr."     "Leland H. Jones, III"
## [17] "Pearson Liddell, Jr."    "Victor W. Carmody, Jr."
```

```
str_subset(x, ".[A-Z]\\. ")       # middle initial instead of middle name
```

```
##  [1] "James H. Powell, III"    "Rosalind H. Jordan"
##  [3] "Raymond M. Baum"         "Grady F. Tollison, Jr."
##  [5] "Jackson M. Brown"        "Edward C. Fenwick"
```

```
##  [7] "Louis F. Coleman"        "Bennie L. Jones, Jr."
##  [9] "Thomas A. Coleman"       "Steven E. Farese"
## [11] "William L. Maxey"        "Mitchell M. Lundy Jr."
## [13] "R.T. Laster, Jr."        "Robert T. Laster, Jr."
## [15] "Johnnie E. Walls, Jr."   "Bradley S. Peeples"
## [17] "Ross R. Barnett, Jr"     "Bernard C. Jones, Jr."
## [19] "Jeff G. Houston"         "Thomas M. Flanagan, Jr."
## [21] "Aelicia L. Thomas"       "Edwin A. Flint, Jr."
## [23] "James C. Mayo"           "John M. Colette"
## [25] "Chatwin M. Jackson"      "Kevin D. Camp"
## [27] "James G. Mcintyre"       "Leland H. Jones, III"
## [29] "James P. Vance"          "John H. Gilmore"
## [31] "Victor W. Carmody, Jr."
```

```
str_subset(x, "([^I])\\1")        # two repeated letters (but not II, III, etc.)
```

```
##  [1] "James H. Powell, III"   "Grady F. Tollison, Jr."
##  [3] "Eddie Fenwick"          "Keith Ball"
##  [5] "Mickey Mallette"        "Kevin Null"
##  [7] "Bennie L. Jones, Jr."   "J. Niles McNeel"
##  [9] "Joey Hood, II"          "William L. Maxey"
## [11] "Mitchell M. Lundy Jr."  "Johnnie E. Walls, Jr."
## [13] "Bradley S. Peeples"     "Ross R. Barnett, Jr"
## [15] "Johnnie McDaniels"      "J. Stewart Parrish"
## [17] "Jeff G. Houston"        "H. Lee Bailey, Jr."
## [19] "Jimmy Vance"            "Hugh Lee Bailey, Jr."
## [21] "Webb Franklin"          "Stephanie Mallette"
## [23] "Caroline Moore"         "John M. Colette"
## [25] "Austin Vollor"          "Jim Davis Hull"
## [27] "Jannie Lewis"           "Johnnie Walls"
## [29] "Jeffery Waldo"          "Lee Jones"
## [31] "Lee Bailey"             "Mitchell Lundy, Sr."
## [33] "Billie Jo White"        "Pearson Liddell, Jr."
## [35] "Pearson Liddell"        "Kenneth Bridges"
## [37] "Michael Goggans"        "Brian Neely"
## [39] "David Tisdell"          "W. Mitchell Moran"
```

```
str_subset(x, "[^J][rs]$|([rs],.*)$") # last name ends with r or s
```

```
##  [1] "Richard Carter, III"   "Richard Carter"
##  [3] "K. Elizabeth Davis"    "Bennie L. Jones, Jr."
##  [5] "R.T. Laster, Jr."      "Robert T. Laster, Jr."
##  [7] "Johnnie E. Walls, Jr." "Bradley S. Peeples"
##  [9] "Johnnie McDaniels"     "Bernard C. Jones, Jr."
## [11] "Ray Charles Carter"    "Aelicia L. Thomas"
## [13] "Mark Majors"           "Austin Vollor"
## [15] "Jannie Lewis"          "Johnnie Walls"
## [17] "Elizabeth Davis"       "Lee Jones"
## [19] "Leland H. Jones, III"  "Alison Steiner"
## [21] "Kenneth Bridges"       "Michael Goggans"
## [23] "Andy Davis"
```

```
str_subset(x, "^[AEIOU][^aeiou]") # first name starts with 1 vowel but not 2
```

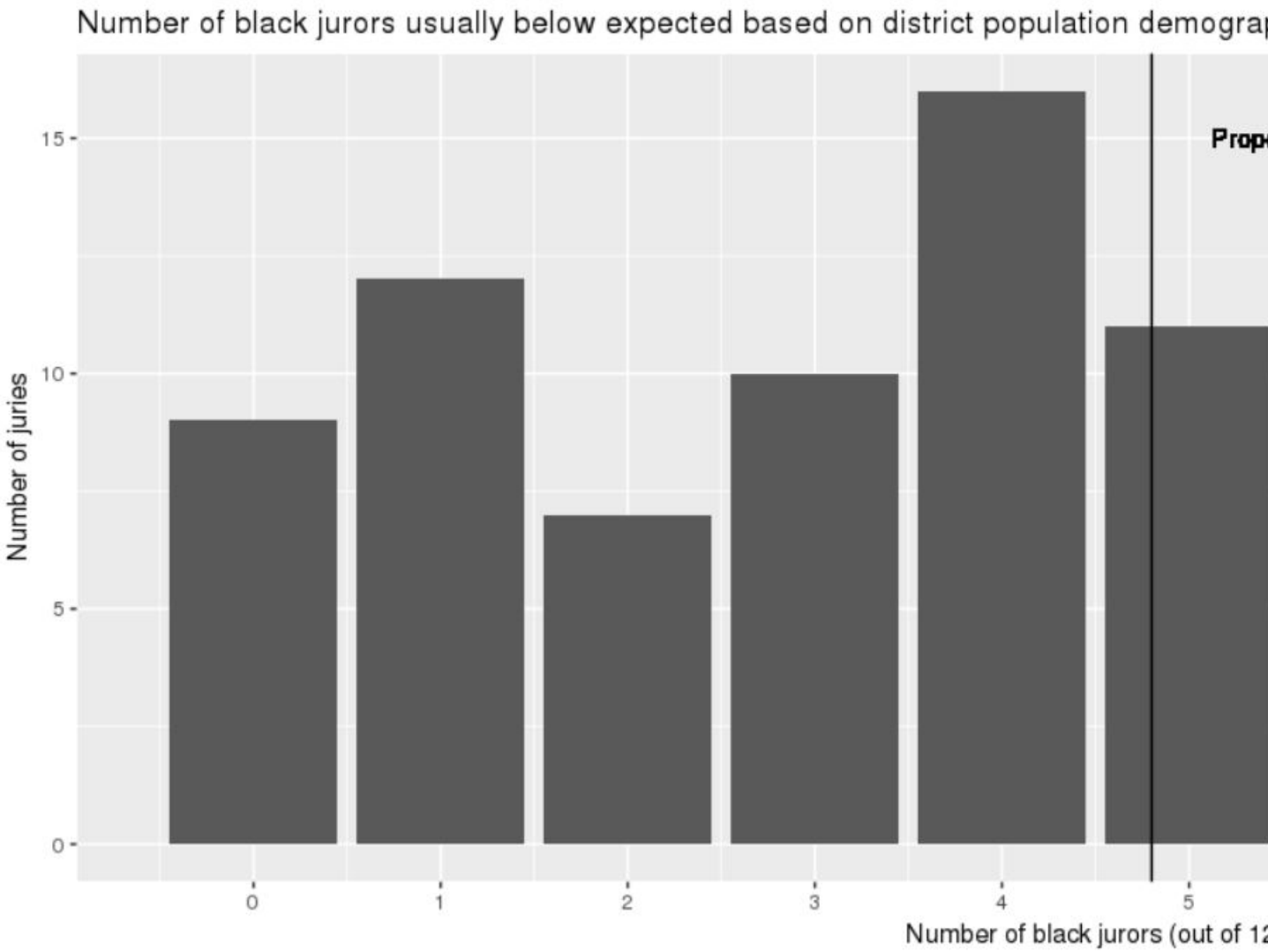
```
##  [1] "Edward C. Fenwick"   "Eddie Fenwick"       "Antwayn Patrick"
##  [4] "Imhotep Alkebu-lan"  "Azki Shah"           "Edwin A. Flint, Jr."
```

```
##  [7] "Elizabeth Davis"     "Alison Steiner"      "Andy Davis"
## [10] "Andre' de Gruy"
```

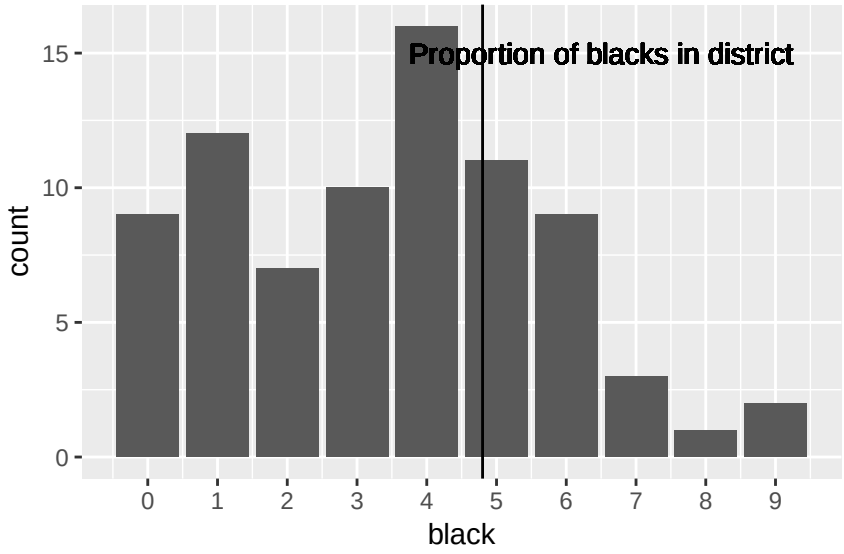
```
# don't need master_small for this analysis
#  - can use all jurors with voirdire data
master %>%
  select(notes, county) %>%
  filter(!is.na(notes)) %>%
  mutate(relative = str_detect(notes, "relat|cousin|kin")) %>%
  group_by(county) %>%
  summarise(relatives = sum(relative))
```

```
## # A tibble: 7 x 2
##   county     relatives
##   <chr>          <int>
## 1 Attala             0
## 2 Carroll            0
## 3 Choctaw            0
## 4 Grenada            0
## 5 Montgomery        12
## 6 Webster           28
## 7 Winston            4
```

5. Recreate the plot below. Only consider trials where there are exactly 12 jurors in our database who were selected to serve (i.e. where `struck_by` is "Juror chosen to serve on jury"). Also note that the vertical line is based on 40% black population in the district.




```
# Q5

# To recreate the plot, must use master data = trials in voirdire
master %>%
  filter(struck_by == "Juror chosen to serve on jury") %>%
  group_by(trial__id) %>%
  summarise(jurors = n(),
            black = sum(race == "Black"),
            white = sum(race == "White")) %>%
  filter(jurors == 12) %>%
  ggplot(aes(x = black)) +
    geom_bar() +
    geom_vline(xintercept = 4.8) +
    scale_x_continuous(breaks = seq(0, 12, 1)) +
    geom_text(x = 6.5, y = 15, label = "Proportion of blacks in district")
```




```
# If use all trials instead
jurors %>%
  filter(struck_by == "Juror chosen to serve on jury") %>%
  group_by(trial__id) %>%
  summarise(jurors = n(),
            black = sum(race == "Black"),
            white = sum(race == "White")) %>%
  filter(jurors == 12) %>%
  ggplot(aes(x = black)) +
    geom_bar() +
    geom_vline(xintercept = 4.8) +
    scale_x_continuous(breaks = seq(0, 12, 1)) +
    geom_text(x = 6.5, y = 75, label = "Proportion of blacks in district")
```