Data Exploration: Making Decisions

Luke Kolar

September 9, 2021

In this Data Exploration assignment, you have two separate data sets with which you will work. The first involves the data generated by you and your classmates last week when you took the in-class survey. The second involves some of the data used in the Atkinson et al. (2009) piece that you read for class this week. Both data sets are described in more detail below.

If you have a question about any part of this assignment, please ask! Note that the actionable part of each question is **bolded**.

Part 1: Cognitive Biases

You may have noticed that the questions on the survey you took during class last week were based on the Kahneman (2003) reading you did for this week. The goal for this set of questions is to examine those data to see if you and your classmates exhibit the same cognitive biases that Kahneman wrote about. The data you generated is described below.

Data Details:

- File Name: `bias_data.csv`
- Source: These data are from the in-class survey you took last week.

Variable Name	Variable Description
`id`	Unique ID for each respondent
`rare_disease_prog`	From the rare disease problem, the program chosen by the respondent (either 'Program A' or 'Program B')
`rare_disease_cond`	From the rare disease problem, the framing condition to which the respondent was assigned (either 'save' or 'die')
`linda`	From the Linda problem, the option the respondent thought most probable, either "teller" or "teller and feminist"
`cab`	From the cab problem, the respondent's estimate of the probability the car was blue
`gender`	One of "man", "woman", "non-binary", or "other"
`year`	Year at Harvard
`college_stats`	Indicator for whether or not the respondent has taken a college-level statistics course

Before you get started, make sure you replace "file_name_here_1.csv" with the name of the file. (Also, remember to make sure you have saved the .Rmd version of this file and the file with the data in the same folder.)

```r
# load the class-generated bias data
bias_data <- read_csv("bias_data.csv")
```

Question 1

First, let's look at the rare disease problem. You'll recall from the Kahneman (2003) piece that responses to this problem often differ based on the framing (people being saved versus people dying), despite the fact that the two frames are logically equivalent. This is what is called a 'framing bias'.

Did you all exhibit this bias? Since the outcomes for this problem are binary, we need to test to see if the proportions who chose Program A under each of the conditions are the same. Report the difference in proportions who chose Program A under the 'save' and 'die' conditions. Do we see the same pattern that Kahneman described?

```r
bias_data %>%
  group_by(rare_disease_cond) %>%
  summarize(program_a_prop = mean(rare_disease_prog == "Program A"))
```

```
## # A tibble: 2 x 2
##   rare_disease_cond program_a_prop
##   <chr>                     <dbl>
## 1 die                       0.349
## 2 save                      0.667
```

ANSWER: *Yes, only about one third of respondents chose Program A under the "die" framing, whereas two-thirds chose Program A when the question used "save" framing. These are similar to the results observed by Kahneman, who posited that people are more loss-averse.*

EXTENSION: Report the 95% confidence interval for the difference in proportions you just calculated. Hint: the `infer` package has a function that is useful here. What does the 95% confidence interval mean?

Note that extensions to questions are not the same as data science questions. Complete this question if you like, but it is not required for data science students like actual data science questions.

```r
prop_test(bias_data, rare_disease_prog ~ rare_disease_cond, order = c("die", "save"))
```

```
## # A tibble: 1 x 6
##   statistic chisq_df p_value alternative lower_ci upper_ci
##       <dbl>    <dbl>   <dbl> <chr>          <dbl>    <dbl>
## 1      7.36        1 0.00666 two.sided     -0.543  -0.0928
```

ANSWER: *The 95% confidence interval for the difference in proportions is (-0.543, -0.093). This means we can be 95% confident that the true difference in proportions is between these values.*

Question 2

Now let's move on to the Linda problem. As we read in Kahneman (2003), answers to this problem tend to exhibit a pattern called a "conjunction fallacy" whereby respondents overrate the probability that Linda is a bank teller *and* a feminist rather than just a bank teller. From probability theory, we know that the

conjunction of two events A and B can't be more probable than either of the events occurring by itself; that is, $P(A) \geq P(A \wedge B)$ and $P(B) \geq P(A \wedge B)$[1].

What proportion of the class answered this question correctly? Why do you think people tend to choose the wrong option?

```
nrow(bias_data %>% filter(linda == "teller")) / nrow(bias_data)
```

```
## [1] 0.7058824
```

ANSWER: *About 70% of the class answered this correctly. People tend to get this wrong because of the "conjunction fallacy," whereby it's easier for the mind to link feminist stereotypes to Linda instead of calculating the probabilities of each (which would reveal that it's more logical to guess she's a bank teller).*

Question 3

What attributes of the respondents do you think might affect how they answered the Linda problem and why? Using the data, see if your hypothesis is correct.

```
bias_data %>%
  group_by(year) %>%
  summarize(stat = mean(linda == "teller"))
```

```
## # A tibble: 4 x 2
##   year   stat
##   <chr> <dbl>
## 1 1     0.5
## 2 2     0.696
## 3 3     0.707
## 4 4+    0.765
```

```
bias_data %>%
  group_by(college_stats) %>%
  summarize(stat = mean(linda == "teller"))
```

```
## # A tibble: 2 x 2
##   college_stats  stat
##   <chr>         <dbl>
## 1 No            0.613
## 2 Yes           0.759
```

```
bias_data %>%
  group_by(year, college_stats) %>%
  summarize(stat = mean(linda == "teller"), n = n())
```

```
## # A tibble: 7 x 4
## # Groups:   year [4]
##   year  college_stats  stat     n
```

[1]The symbol $\wedge$ is used in logical expressions to mean "AND". If there are two conditions, A and B, then $A \wedge B$ is true only when both A and B are separately true. The expression $P(A) \geq P(A \wedge B)$ is therefore interpreted as: "The probability A is true is greater than or equal to the probability that both A and B are true.

```
##   <chr> <chr>         <dbl> <int>
## 1 1     Yes           0.5       4
## 2 2     No            0.5      10
## 3 2     Yes           0.846    13
## 4 3     No            0.533    15
## 5 3     Yes           0.808    26
## 6 4+    No            1         6
## 7 4+    Yes           0.636    11
```

ANSWER: *I would guess - and the data show - that education level (which we can measure using the* `year` *variable) and whether or not a student has taken a college statistics course (the* `college_stats` *variable) would both positively correlate with correctness on the problem. Both are true, as is, for the most part, grouping by both variables. See above tables.*

Question 4: Data Science Question

Now we will take a look at the taxi cab problem. This problem, originally posed by Tversky and Kahneman in 1977, is intended to demonstrate what they call a "base rate fallacy". To refresh your memory, here is the text of the problem, as you saw it on the survey last week:

> A cab was involved in a hit and run accident at night. Two cab companies, the Green and the Blue, operate in the city. 85
>
> A witness identified the cab as Blue. The court tested the reliability of the witness under the same circumstances that existed on the night of the accident and concluded that the witness correctly identified each one of the two colours 80
>
> What is the probability that the cab involved in the accident was Blue rather than Green knowing that this witness identified it as Blue?

The most common answer to this problem is .8. This corresponds to the reliability of the witness, without regard for the base rate at which Blue cabs can be found relative to Green cabs. In other words, respondents tend to disregard the base rate when estimating the probability the cab was Blue.

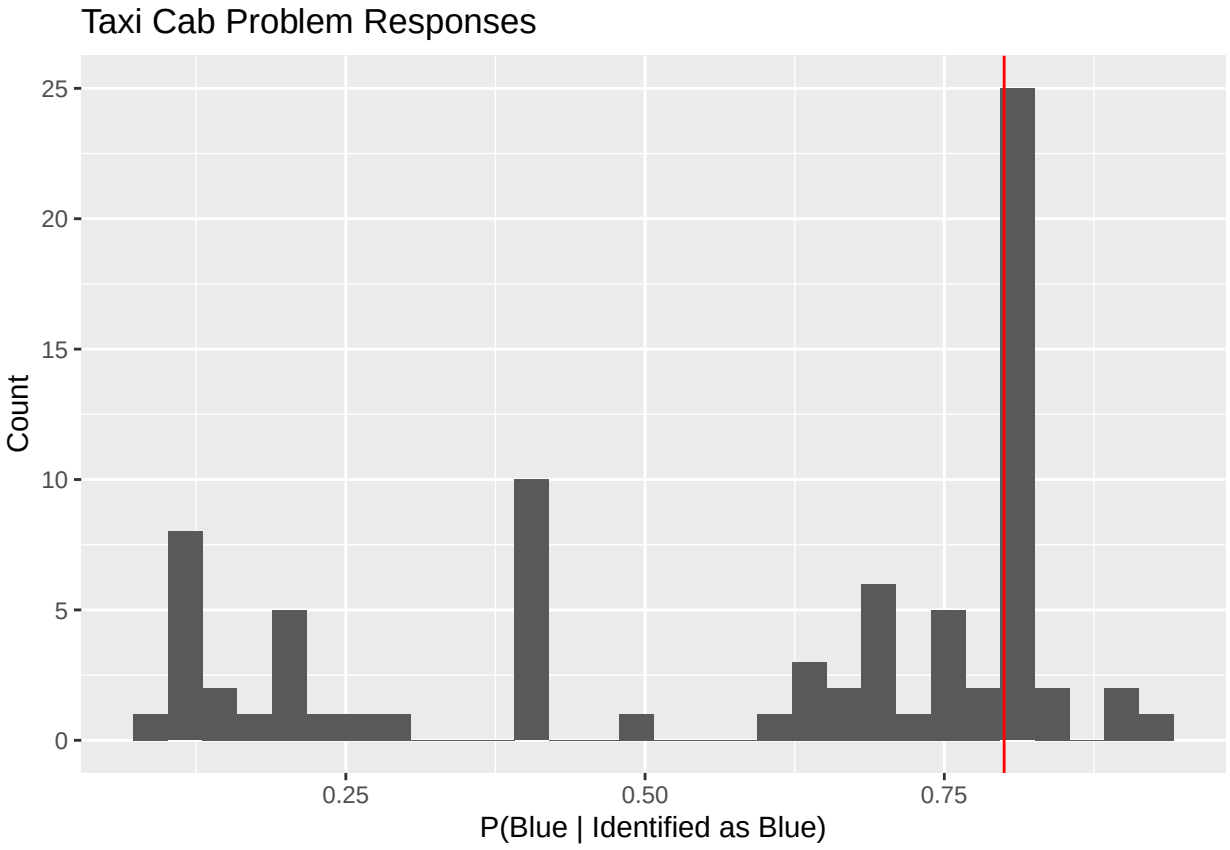
What is the true probability the cab was Blue? Visualize the distribution of the guesses in the class using a histogram. What was the most common guess in the class?

ANSWER: *Bayes' Rule says that to find the true probability, we must take the probability that the car was both identified as blue and was blue (0.8) times the probability the car was blue (0.15). Then, we divide this by the product of the same two probabilities (0.8 and 0.15) plus the product of the probability the car was identified as blue but was green (0.2) and the probability the car was green (0.85). The math is in the code chunk below, but the answer is about 0.414. The distribution of guesses is also below, with the red line being the most common answer.*

```
# P(iB|B)*P(B) / [P(iB|B)*P(B) + P(iB|G)*P(G)]
(0.8 * 0.15) / ((0.8 * 0.15) + (0.2 * 0.85))
```

```
## [1] 0.4137931
```

```
bias_data %>%
  ggplot(aes(x = cab)) + geom_histogram() + geom_vline(aes(xintercept = .8), col = "red") +
  labs(x = "P(Blue | Identified as Blue)", y = "Count", title = "Taxi Cab Problem Responses")
```



Part 2: Political Faces

Now you will investigate some of the data used in Atkinson et al. (2009). These data cover Senate candidates from 1992-2006 and include face ratings, partisanship, incumbent status, and other variables.

Data Details:

- File Name: `senate_data.csv`
- Source: These data are condensed and adapted from the replication data for Atkinson et al. (2009).

Variable Name	Variable Description
`cook`	The assessment of the Senate race from the Cook Political Report in the year prior to the election
`year`	The year of the election
`state`	The state in which the candidate was running
`face_rating`	The normalized rating of the candidate's perceived competence based on an image of the face
`incumbent`	An indicator variable for whether the candidate was an incumbent
`candidate`	The candidate's name
`party`	The candidate's political party
`tossup`	An indicator variable for whether the race was one of two "tossup" categories according to Cook
`jpg`	A unique identifier for the photo of the candidate

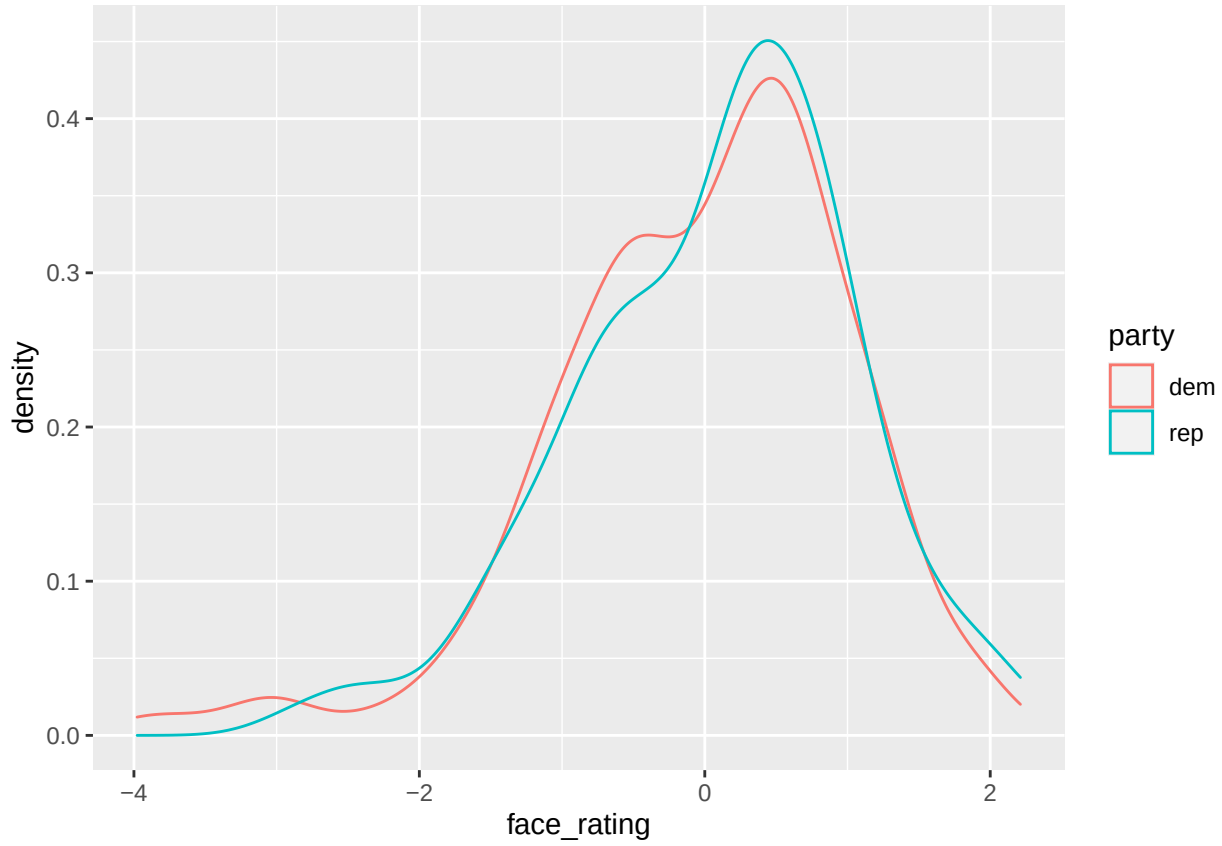
As before, make sure you replace "file_name_here_2.csv" with the name of the file.

```
face_data <- read_csv("senate_data.csv")
```

As an example of how you might write your own code to analyze these data, let's take a look at whether there was a difference in the perceived competence of Democratic and Republican candidates' faces. We can examine this question graphically using a density plot.

```
# make density plot of perceived competence by party

ggplot(data = face_data, aes(x = face_rating, color = party)) + # note that by setting color = party,
  geom_density()                                                # the face ratings of each party will b
```




```
                                                              # displayed in different colors
```

We can also consider this statistically using a t-test for whether or not the mean face ratings are significantly different across parties.

```
# conduct a t-test of difference-in-means
difference_in_means(face_rating ~ party, data = face_data)
```

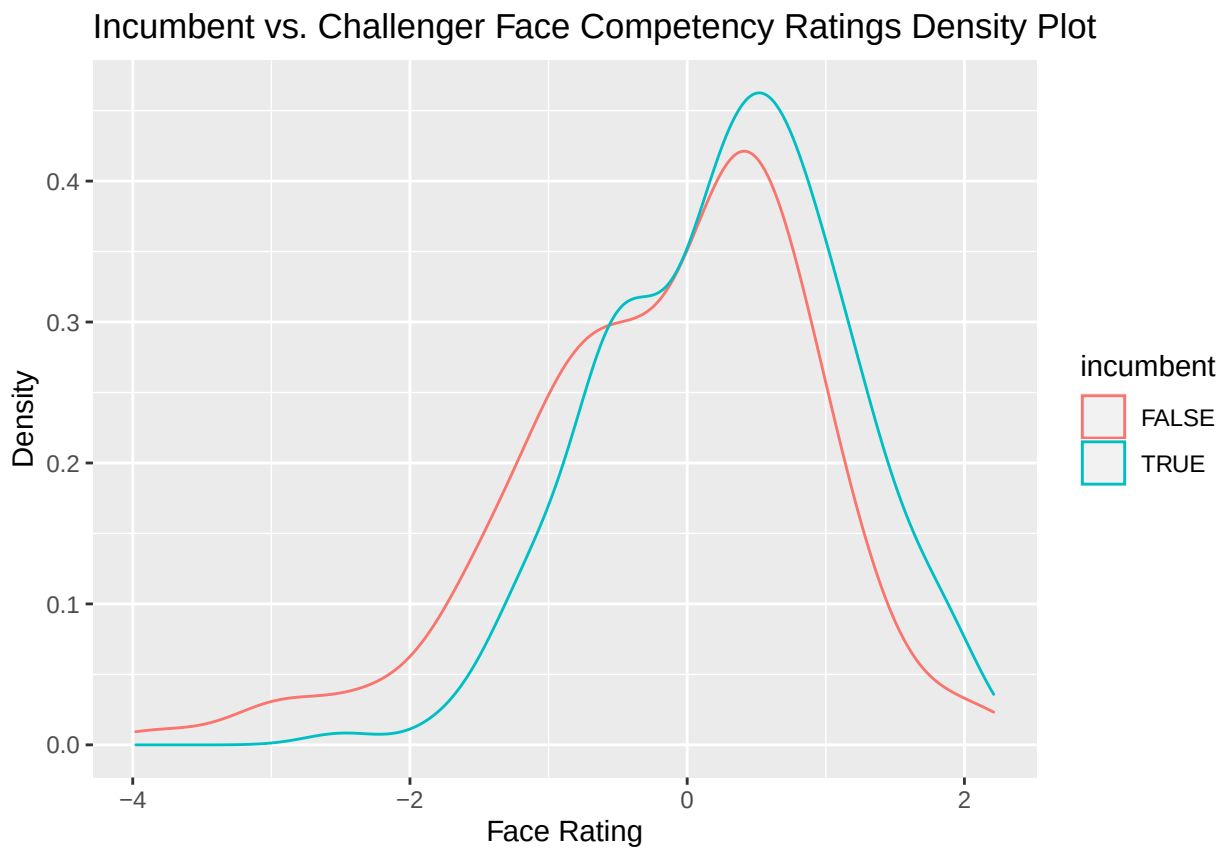
```
## Design:  Standard
##           Estimate Std. Error  t value  Pr(>|t|)    CI Lower  CI Upper       DF
## partyrep 0.1044044 0.09565385 1.091482 0.2756698 -0.08360089 0.2924098 431.5741
```

Neither the graphical nor the statistical approaches suggest a significant difference in perceived competence of candidate faces by party.

Question 5

Do the data suggest a significant difference between perceived competence of incumbent vs. non-incumbent candidate faces? How do your findings relate to the results and theory of Atkinson et al. (2009)?

```
face_data %>%
  ggplot(aes(x = face_rating, color = incumbent)) + geom_density() +
  labs(x = "Face Rating", y = "Density",
       title = "Incumbent vs. Challenger Face Competency Ratings Density Plot")
```




```
difference_in_means(face_rating ~ incumbent, face_data)
```

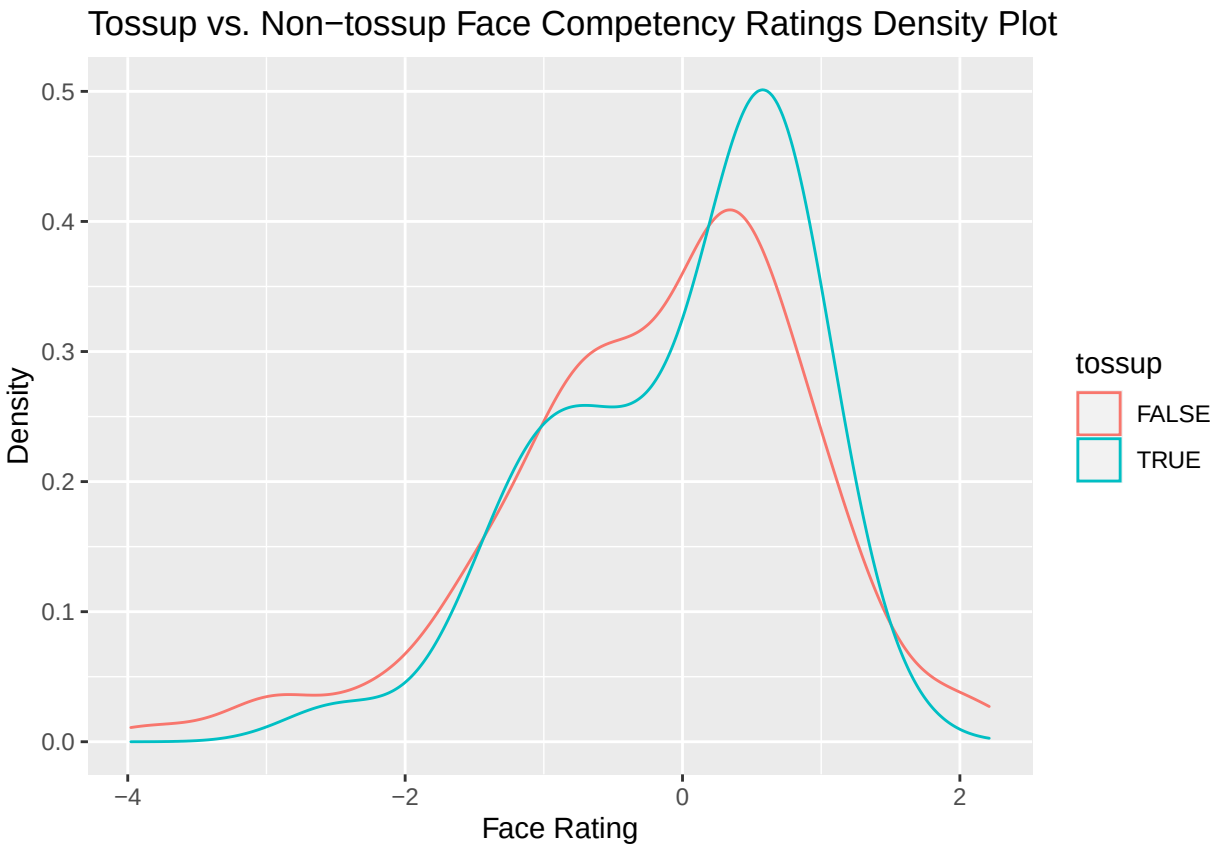
```
## Design:  Standard
##             Estimate Std. Error t value     Pr(>|t|)  CI Lower  CI Upper
## incumbent 0.4480374 0.09084939 4.93165 1.161294e-06 0.2694804 0.6265944
##                  DF
## incumbent 436.1783
```

ANSWER: *The density plot and the t-test of difference in means both suggest that incumbents were perceived to have more competent faces on average. The incumbent face ratings density peaks at a higher rating than for non-incumbents on the plot, and the t-test confidence interval's lower and upper bounds are greater than zero, suggesting a statistically significant difference. Atkinson et. al (2009) suggest that incumbent face ratings are perceived as more competent because of "selection effects" relating to why these politicians were elected in the first place (voters' preferences and party strategy).*

Question 6

Do the data suggest a significant difference between perceived competence of non-incumbent candidate faces in tossup vs. non-tossup races? What might explain any similarities or differences between these results and those from the previous question? How do your findings relate to the results and theory of Atkinson et al. (2009)?

```
face_data %>%
  filter(incumbent == "FALSE") %>%
  ggplot(aes(x = face_rating, color = tossup)) + geom_density() +
  labs(x = "Face Rating", y = "Density",
       title = "Tossup vs. Non-tossup Face Competency Ratings Density Plot")
```




```
difference_in_means(face_rating ~ tossup, face_data %>% filter(incumbent == "FALSE"))
```

```
## Design:  Standard
##         Estimate Std. Error  t value  Pr(>|t|)   CI Lower  CI Upper       DF
## tossup 0.1740081  0.1612837 1.078894 0.2850889 -0.1488172 0.4968334 58.16194
```

```
face_data %>%
  filter(incumbent == "FALSE") %>%
  group_by(tossup, incumbent) %>%
  summarize(n = n())
```

```
## # A tibble: 2 x 3
```

```
## # Groups:   tossup [2]
##   tossup incumbent     n
##   <lgl>  <lgl>     <int>
## 1 FALSE  FALSE       222
## 2 TRUE   FALSE        38
```

ANSWER: *While a glance at the density plot suggests a difference in average face ratings for tossup versus non-tossup races, since the former's peak has greater density at a higher face rating, the results of the difference in means t-test were not statistically significant, as zero is included in the confidence interval. But, looking at the data, only 38 of the non-incumbents were in tossup races, compared with 222 non-incumbents in non-tossup races. The former may be too small a sample size for a t-test.*

Question 7: Data Science Question

Atkinson et al. (2009, 236) suggest that "...incumbents from the most competitive districts would have higher facial quality than incumbents from the most safe incument districts due to the selection process of better faces to competitive districts, inducing a negative relationship betwen incumbent face and incumbent vote." **Do the data support the idea that seat safety is negatively correlated with incumbent facial quality? Make a plot to visualize this relationship.** Note that this question may require you to define at least one new variable.

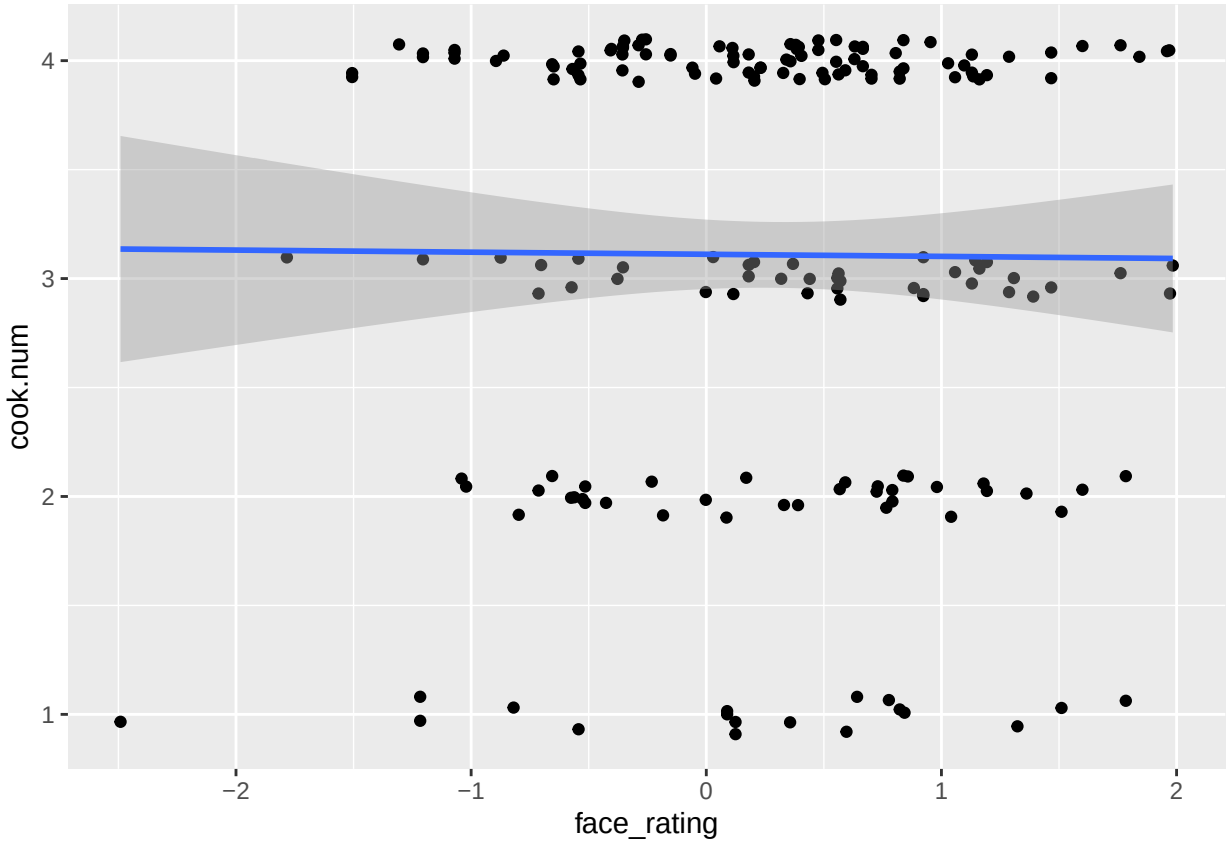
```
glimpse(face_data)
```

```
## Rows: 444
## Columns: 9
## $ cook        <chr> "LeanRep", "LikelyDem", "SolidDem", "SolidDem", "Tossup...
## $ year        <dbl> 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1992, 1...
## $ state       <chr> "AK", "AL", "AR", "AR", "CA", "CO", "CT", "FL", "GA", "...
## $ face_rating <dbl> 1.5996333, 1.1614910, 1.9686832, 0.2143184, -1.3698259,...
## $ incumbent   <lgl> TRUE, TRUE, TRUE, FALSE, FALSE, FALSE, TRUE, TRUE, TRUE...
## $ candidate   <chr> "Frank H. Murkowski", "Richard C. Shelby", "Dale Bumper...
## $ party       <chr> "rep", "dem", "dem", "rep", "rep", "dem", "dem", "dem",...
## $ tossup      <lgl> FALSE, FALSE, FALSE, FALSE, TRUE, FALSE, FALSE, FALSE, ...
## $ jpg         <dbl> 537, 105, 445, 446, 447, 543, 114, 545, 448, 548, 292, ...
```

```
new_face <- face_data %>%
  mutate(cook.num = ifelse(cook == "SolidDem", 4,
                    ifelse(cook == "LikelyDem", 3,
                    ifelse(cook == "LeanDem", 2,
                    ifelse(cook %in% c("TossupDem", "TossUpDem"), 1,
                    ifelse(cook %in% c("TossupRep", "TossUpRep"), 1,
                    ifelse(cook == "LeanRep", 2,
                    ifelse(cook == "LikelyRep", 3,
                    ifelse(cook == "SolidRep", 4, 5)))))))) %>%
  filter(incumbent == TRUE) %>%
  mutate(cook.num = as.numeric(cook.num))

new_face %>% ggplot(aes(y = cook.num, x = face_rating)) + geom_jitter(height = 0.1, width = 0) +
  stat_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(new_face$cook.num, new_face$face_rating)
```

```
## [1] -0.00789173
```

Question 8

Is there something else interesting or informative that you could explore using either of these datasets? If so, run it by a TF and try it out here.