

# Data Exploration: Making Decisions

Yao Yu

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
<code>id</code>	Unique ID for each respondent
<code>rare_disease_prog</code>	From the rare disease problem, the program chosen by the respondent (either 'Program A' or 'Program B')
<code>rare_disease_cond</code>	From the rare disease problem, the framing condition to which the respondent was assigned (either 'save' or 'die')
<code>linda</code>	From the Linda problem, the option the respondent thought most probable, either "teller" or "teller and feminist"
<code>cab</code>	From the cab problem, the respondent's estimate of the probability the car was blue
<code>gender</code>	One of "man", "woman", "non-binary", or "other"
<code>year</code>	Year at Harvard
<code>college_stats</code>	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.)

```
# load the class-generated bias data
bias_data <- read_csv("data/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?

**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.

```
# Calculating the proportions of respondents who chose Program A under the save
# and die conditions
bias_data %>%
  filter(rare_disease_prog == "Program A") %>%
  count(rare_disease_cond) %>%
  mutate(n = round(n / sum(n), digits = 3)) %>%
  gt() %>%
  tab_header(title = "Respondents who chose Program A") %>%
  cols_label(
    rare_disease_cond = "Condition",
    n = "Proportion of responses"
  )
```

Respondents who chose Program A

Condition	Proportion of responses
die	0.349
save	0.651

```
# Calculating the 95% confidence interval
prop_test(bias_data, rare_disease_prog ~ rare_disease_cond,
  conf_level = .95, 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
```

From the table, we can see that a higher proportion of respondents chose Program A under the save condition. This reflects the same pattern that Kahneman described in Prospect Theory that people tend to be more loss adverse. The 95% confidence interval supports this theory further because it tells us that at a 95% confidence level, we can say that the difference in proportion between participants who choose Program A under the "save" and "die" conditions is between 0.543 and 0.093.

## 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  $P(B) \geq P(A \wedge B)$ <sup>1</sup>.

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

```
# Calculating the proportion of the class answered this question correctly
bias_data %>%
  count(linda) %>%
  mutate(n = round(n / sum(n), digits = 3)) %>%
  gt() %>%
  tab_header(title = "How the class responded to the Linda problem") %>%
  cols_label(
    linda = "Choice",
    n = "Proportion of responses"
  )
```

How the class responded to the Linda problem

Choice	Proportion of responses
teller	0.706
teller and feminist	0.294

A fairly high proportion of students in the class, around 70%, answered the Linda problem correctly. I think most people tend to choose the wrong option because they use the representativeness heuristic to make a quick judgment. Because Linda appears to be more representative of a feminist, people are naturally drawn to connect her as a feminist even though it is statistically not as likely as Linda being just a bank teller.

---

<sup>1</sup>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."

### 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.

```
# Evaluating if the responses are different among students who have studied
# stats and those have not
bias_data %>%
  count(college_stats, linda) %>%
  group_by(college_stats) %>%
  mutate(n = round(n / sum(n), digits = 3)) %>%
  arrange(desc(college_stats)) %>%
  pivot_wider(names_from = "college_stats", values_from = "n") %>%
  gt() %>%
  tab_header(title = "Responses to the Linda problem based on stats background") %>%
  cols_label(
    linda = "Choice",
    Yes = "Stats",
    No = "No Stats"
  )
```

Responses to the Linda problem based on stats background

Choice	Stats	No Stats
teller	0.759	0.613
teller and feminist	0.241	0.387

I hypothesized that students who have a stats background are less likely to be fooled by the Linda problem. The data provides initial support for my hypothesis in that a higher proportion of students with a stats background correctly chose “teller” compared to students without a stats background.

## 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?**

$$\begin{aligned} P(A|B) &= \frac{P(B|A)P(A)}{P(B)} \\ &= \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|A^c)P(A^c)} \end{aligned}$$

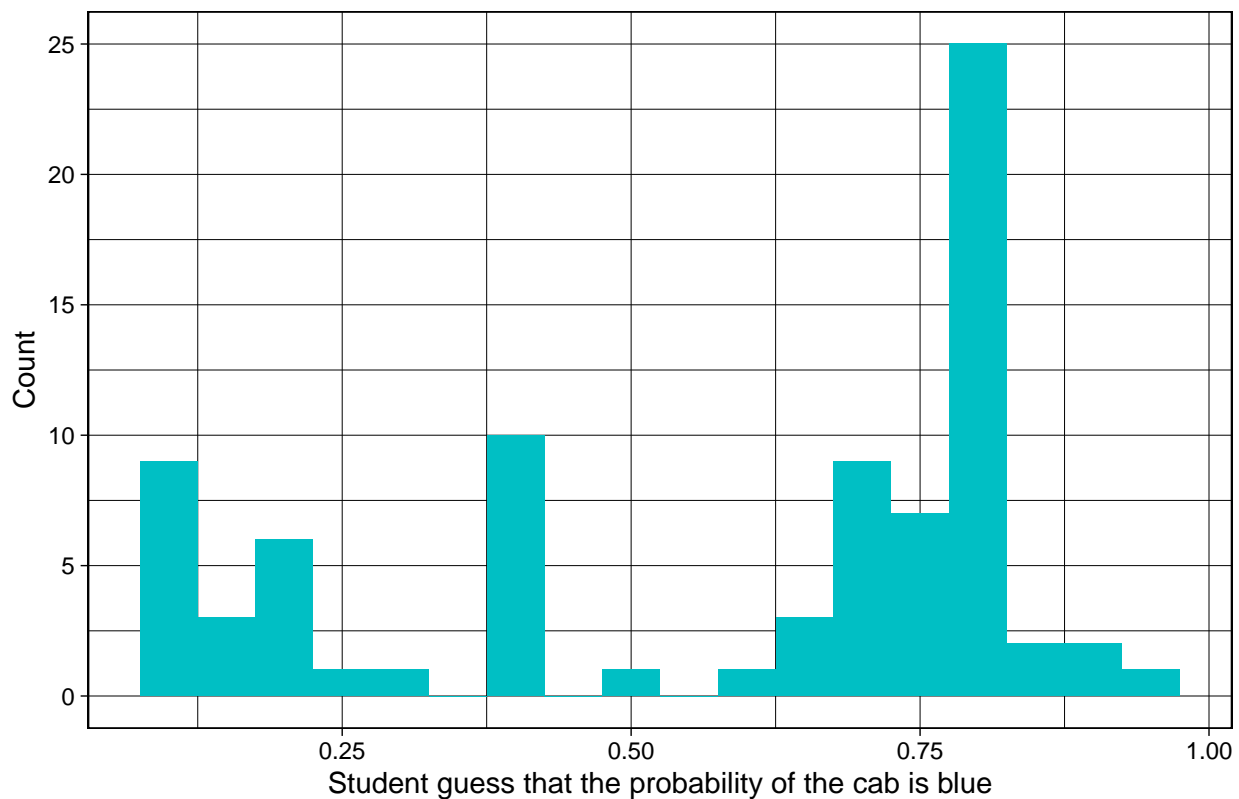
$$P(A) = \text{Blue}$$

$$P(B) = \text{Witness says Blue}$$

$$\begin{aligned} P(A|B) &= \frac{(0.8)(0.15)}{(0.8)(0.15) + (0.2)(0.85)} \\ &= \frac{0.12}{0.12 + 0.17} \\ &\approx 0.414 \end{aligned}$$

```
# Creating a histogram showing the distribution of class responses
bias_data %>%
  drop_na(cab) %>% # Removing NA guesses
  ggplot(aes(x = cab)) +
  geom_histogram(binwidth = .05, fill = "#00bfc4") +
  theme_linedraw() +
  labs(
    title = "Distribution of student responses to the cab problem",
    x = "Student guess that the probability of the cab is blue",
    y = "Count"
  )
```

Distribution of student responses to the cab problem



```
# Finding the most common Guess
most_common_guess <- bias_data %>%
  count(cab) %>%
  arrange(desc(n)) %>%
  head(1) %>%
  pull(cab)

print(most_common_guess)
```

```
## [1] 0.8
```

In the cab problem, we can see that the correct answer is about 41.4% that the cab was actually blue based on the witness's testimony. However, the most common response among the students was 80%.

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

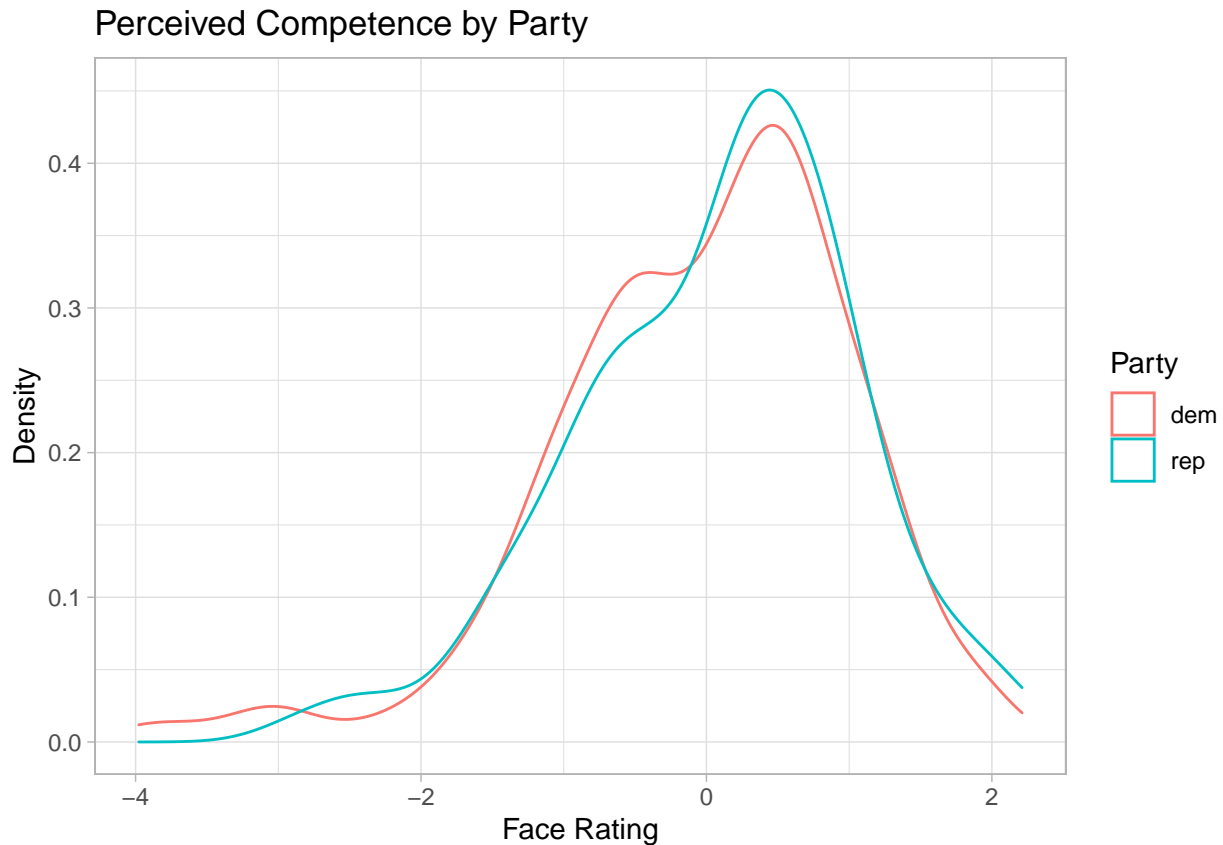
As before, make sure you replace "file\_name\_here\_2.csv" with the name of the file.

```
face_data <- read_csv("data/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
feature_img <- 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 be displayed in different colors
  theme_light() +
  labs(title = "Perceived Competence by Party",
       x = "Face Rating",
       y = "Density",
       color = "Party")

feature_img
```



```
# png("cognition.png", units="in", width=8, height=5, res=300)
# print(feature_img)
# dev.off()
```

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)?

```
# Calculating some summary statistics between incumbent vs. non-incumbent
face_data %>%
  group_by(incumbent) %>%
  summarize(min = round(min(face_rating), digits = 3),
            median = round(median(face_rating), digits = 3),
            mean = round(mean(face_rating), digits = 3),
            max = round(max(face_rating), digits = 3),
            .groups = "drop") %>%
  arrange(desc(incumbent)) %>%
  gt() %>%
  tab_header(title = "Difference between perceived competence of incumbent vs. non-incumbent")
```

Difference between perceived competence of incumbent vs. non-incumbent

	incumbent	min	median	mean	max
TRUE		-2.491	0.363	0.284	1.984
FALSE		-3.978	0.033	-0.164	2.212

```
# Using a t-test to calculate the significance of the difference between
# incumbents vs. non-incumbents
difference_in_means(face_rating ~ incumbent, data = 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
```

From the t-test, we can say that at a 95% confidence level, the difference between perceived competence of incumbent vs. non-incumbent is between an interval of 0.269 and 0.626. In other words, incumbents tend to be perceived as being more competent. This supports the results and theory of Atkinson et al. (2009) that physical appearance matters in how people perceive politicians.

## 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)?

```
# Filtering for only non-incumbents
face_data_non_incumbents <- face_data %>%
  filter(!incumbent)

# Calculating some summary statistics
face_data_non_incumbents %>%
  group_by(tossup) %>%
  summarize(min = round(min(face_rating), digits = 3),
            median = round(median(face_rating), digits = 3),
            mean = round(mean(face_rating), digits = 3),
            max = round(max(face_rating), digits = 3),
            .groups = "drop") %>%
  arrange(desc(tossup)) %>%
  gt() %>%
  tab_header(title = "Difference between perceived competence of non-incumbent candidates in tossup vs. non-tossup races")
```

Difference between perceived competence of non-incumbent candidates in tossup vs. non-tossup races

tossup	min	median	mean	max
TRUE	-2.491	0.268	-0.015	1.335
FALSE	-3.978	-0.009	-0.189	2.212

```
# Using a t-test to calculate the significance of the difference in perceived
# competence between tossups and non-tossup races among non-incumbent candidate
difference_in_means(face_rating ~ tossup, data = face_data_non_incumbents)
```

```
## 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
```

When evaluating the perceived competence of non-incumbent candidate faces in tossup vs. non-tossup races, there is not a statistically significant difference. However, in the summary statistics, we can see that the perceived competence of non-incumbent candidate faces in tossup races on average is still higher than the perceived competence of non-incumbent candidate faces in non-tossup races. This still supports the results and theory of Atkinson et al. (2009) in that physical appearance may still be one factor when deciding which candidates run in closer races.

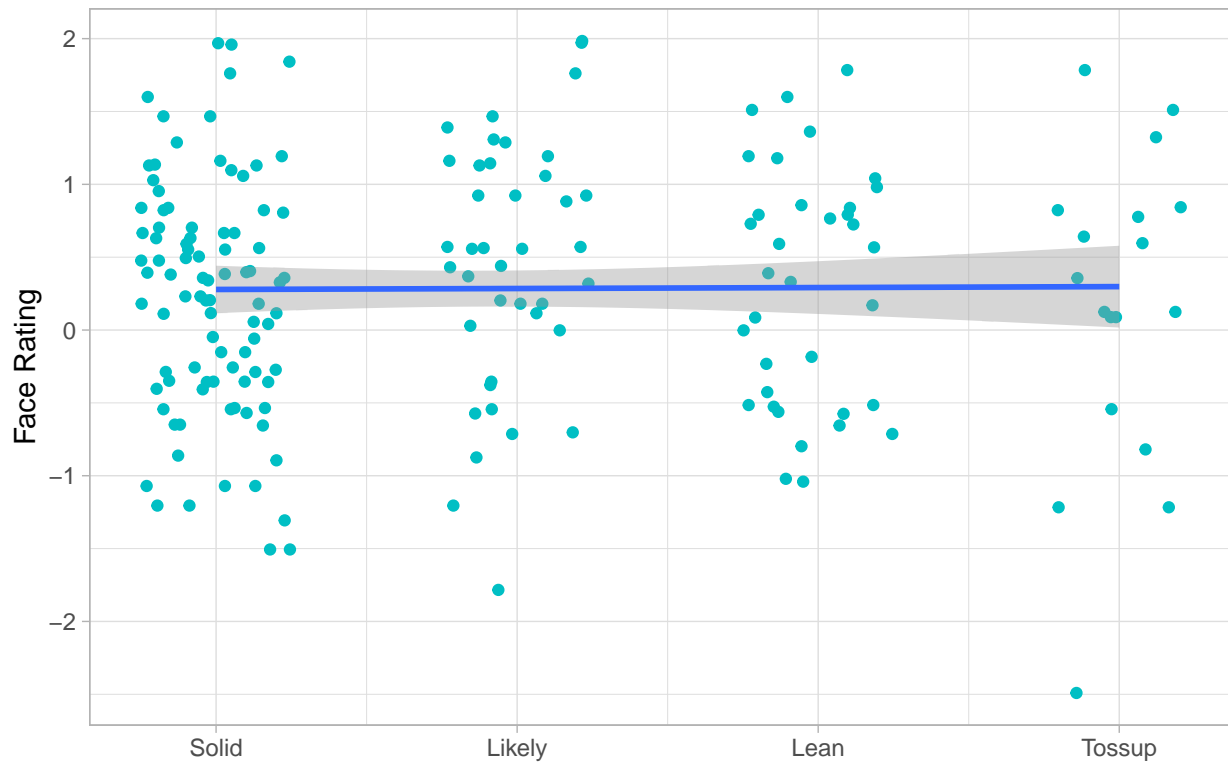
## Question 7: Data Science Question

Atkinson et al. (2009, 236) suggest that "...incumbents from the most competitive districts would have higher facial quality than incumbents from the most safe incumbent districts due to the selection process of better faces to competitive districts, inducing a negative relationship between 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.

```
# Plotting face_rating vs competitive among incumbents
# I'm defining competitiveness using four categories based on the Cook rankings:
# 0: Solid
# 1: Likely
# 2: Lean
# 3: Tossup
seat_safety <- face_data %>%
  filter(incumbent) %>%
  mutate(competitive = case_when(
    str_detect(cook, "Solid") ~ 0,
    str_detect(cook, "Likely") ~ 1,
    str_detect(cook, "Lean") ~ 2,
    str_detect(cook, "Toss") ~ 3,
    TRUE ~ NA_real_
  )) %>%
  ggplot(aes(x = competitive, y = face_rating)) +
  geom_jitter(width = 0.25, color = "#00bfc4") +
  geom_smooth(formula = y ~ x, method = "lm") +
  scale_x_continuous(breaks = c(0, 1, 2, 3), labels = c("Solid", "Likely", "Lean", "Tossup")) +
  theme_light() +
  labs(
    title = "Face Rating scores vs race competitiveness among incumbents",
    x = "",
    y = "Face Rating"
  )

seat_safety
```

Face Rating scores vs race competitiveness among incumbents



```
# png("seat_safety.png", units="in", width=8, height=5, res=300)
# print(seat_safety)
# dev.off()
```

Under my definition of competitiveness based on the Cook rankings for races, I do not find evidence to support the idea that seat safety is negatively correlated with incumbent facial quality. From my plot, it can be seen that the face rating scores remain relatively steady across all types of competitiveness among incumbents.

### 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.