Data Exploration: Making Decisions

Kayla Huang

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.)

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

```
## # A tibble: 85 x 7
##    rare_disease_prog rare_disease_co~ linda       cab year  gender college_stats
##    <chr>             <chr>            <chr>     <dbl> <chr> <chr>  <chr>
##  1 Program B         die              teller     0.2  4+    Man    No
##  2 Program B         save             teller     0.17 2     Man    Yes
##  3 Program A         save             teller     0.8  3     Man    Yes
##  4 Program B         die              teller    NA    2     Woman  Yes
##  5 Program B         save             teller     0.2  3     Man    Yes
##  6 Program A         save             teller     0.15 3     Woman  Yes
##  7 Program A         die              teller     0.7  3     Woman  No
##  8 Program B         die              teller a~  0.85 3     Man    No
##  9 Program B         die              teller a~  0.75 3     Man    Yes
## 10 Program A         die              teller    NA    4+    Man    Yes
## # ... with 75 more rows
```

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?

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

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

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

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?

```
library(infer)
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
```

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.

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?

```
mean(bias_data$linda == "teller")
```

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

Maybe people thought there was some correlation between being a bank teller and being a feminist. Perhaps this correlation was overestimated.

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.

```
# do the responses vary by gender of respondent?
bias_data %>%
  group_by(gender) %>%
  summarize(stat = mean(linda == "teller"))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 3 x 2
##   gender      stat
##   <chr>      <dbl>
## 1 Man        0.75
## 2 Non-binary 0.5
## 3 Woman      0.657
```

Hypothesis: Women are more likely than men to assume that Linda is both a bank teller and a feminist. Results: This was proven correct by the analysis of the survey data. Women were more likely to assume "teller and feminist" over "teller" compared to men. This might be because women are more likely to assume other women are feminists?

[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.

Question 4: Data Science Question

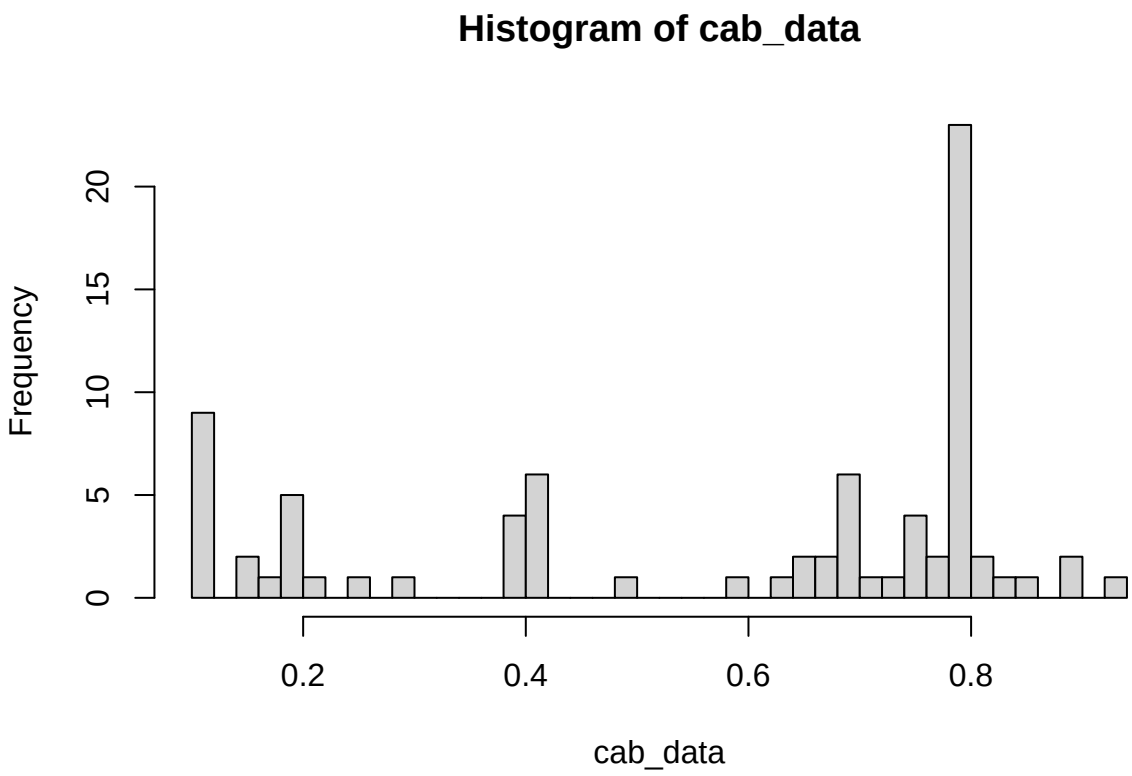
Now we will take a look at the taxi cab problem. This problem, originally posed by Tversky and Kahneman in 1977, is intended to demonstrate what they call a "base rate fallacy". To refresh your memory, here is the text of the problem, as you saw it on the survey last week:

> 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?

```
cab_data <- bias_data$cab
hist(cab_data, breaks=30)
```



```r
getmode <- function(v) {
   uniqv <- unique(v)
   uniqv[which.max(tabulate(match(v, uniqv)))]
}
getmode(cab_data)
```

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

The most common guess in the class was 0.8, as noted in the example problem.

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.

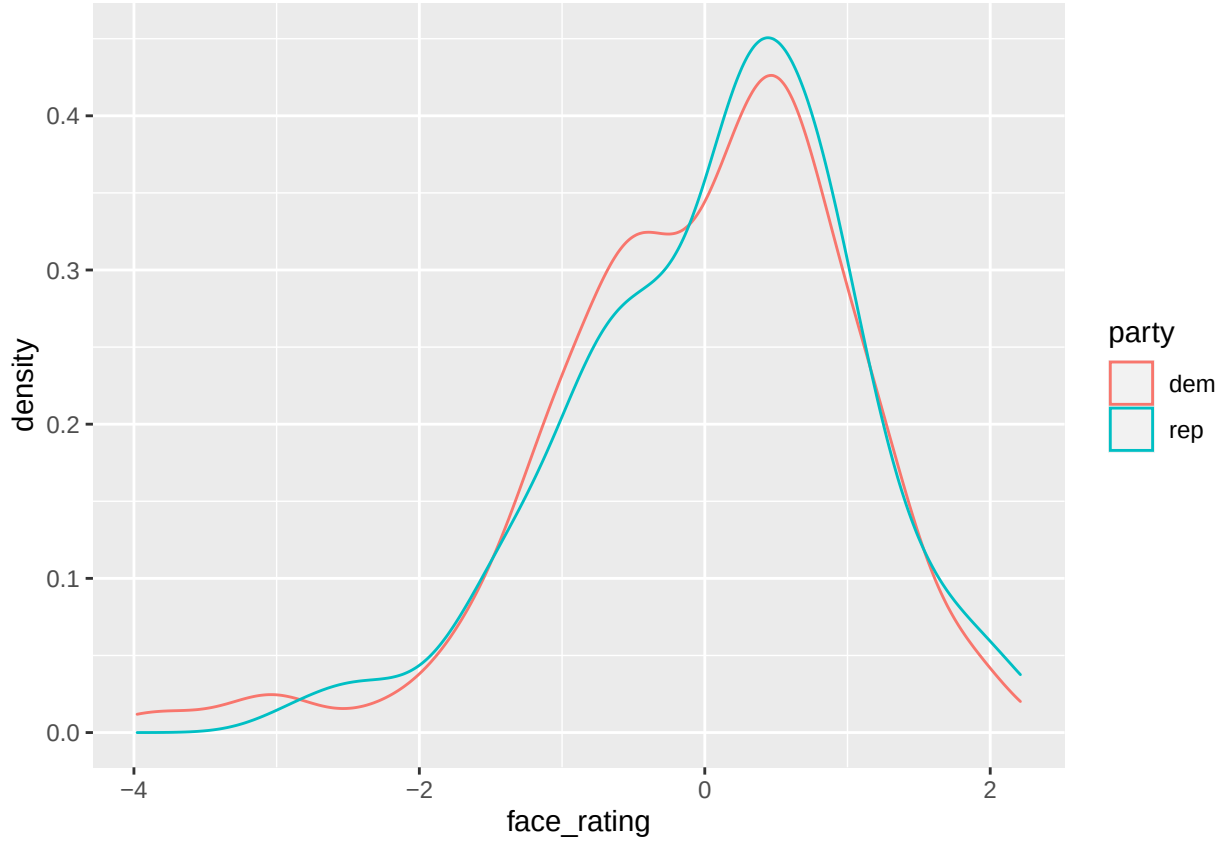
```r
face_data <- read_csv("senate_data.csv")
face_data
```

```
## # A tibble: 444 x 9
##    cook     year state face_rating incumbent candidate        party tossup   jpg
##    <chr>   <dbl> <chr>       <dbl> <lgl>     <chr>            <chr> <lgl>  <dbl>
##  1 LeanRep  1992 AK          1.60  TRUE      Frank H. Murkow~ rep   FALSE    537
##  2 Likely~  1992 AL          1.16  TRUE      Richard C. Shel~ dem   FALSE    105
##  3 SolidD~  1992 AR          1.97  TRUE      Dale Bumpers     dem   FALSE    445
##  4 SolidD~  1992 AR          0.214 FALSE     Mike Huckabee    rep   FALSE    446
##  5 Tossup~  1992 CA         -1.37  FALSE     Bruce Herschens~ rep   TRUE     447
```

```
##  6 LeanDem  1992 CO         -1.02  FALSE     Ben Nighthorse ~ dem   FALSE    543
##  7 Likely~  1992 CT         -0.544 TRUE      Christopher J. ~ dem   FALSE    114
##  8 SolidD~  1992 FL          0.563 TRUE      Bob Graham       dem   FALSE    545
##  9 LeanDem  1992 GA          0.170 TRUE      Wyche Fowler     dem   FALSE    448
## 10 LeanDem  1992 GA          0.319 FALSE     Paul Coverdell   rep   FALSE    548
## # ... with 434 more rows
```

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)
```

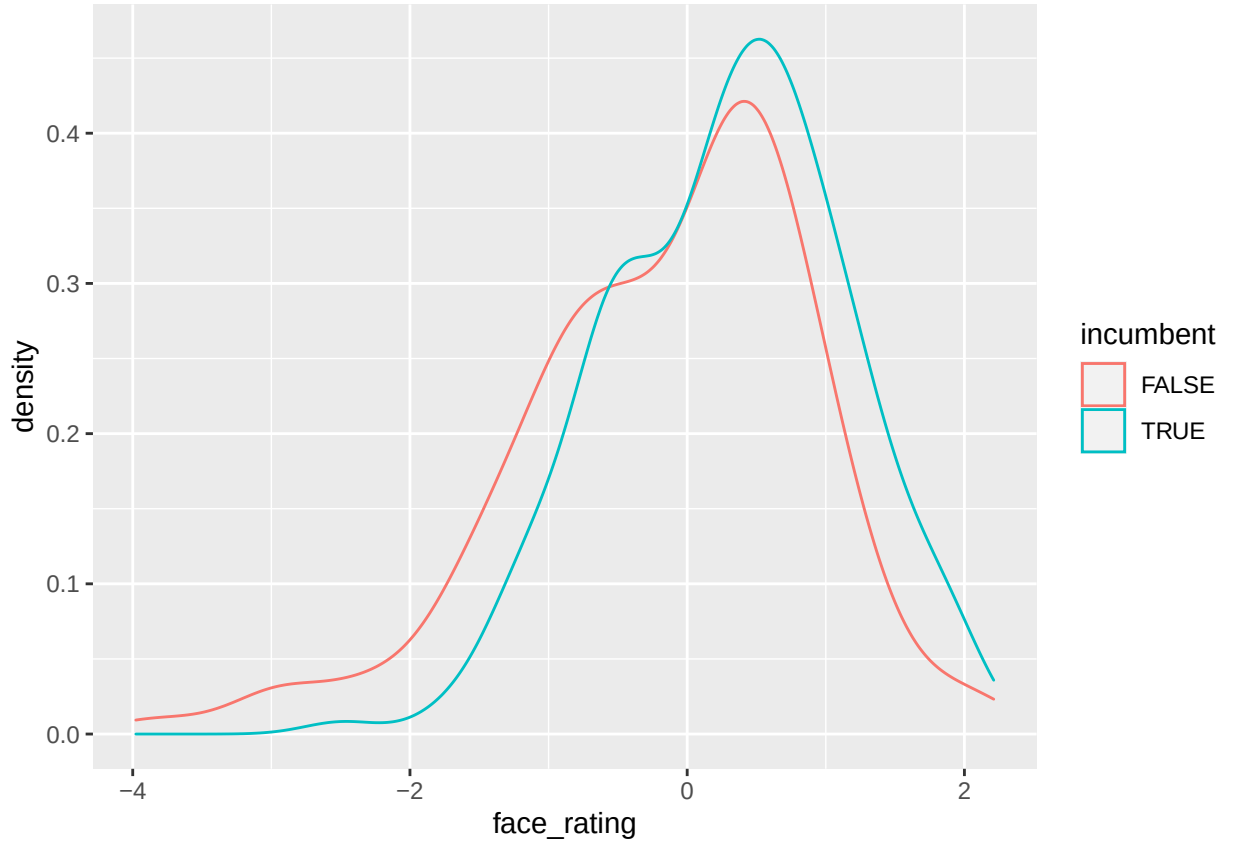
```
## Design:  Standard
##              Estimate Std. Error  t value  Pr(>|t|)    CI Lower  CI Upper       DF
## partyrep 0.1044044 0.09565385 1.091482 0.2756698 -0.08360089 0.2924098 431.5741
```

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)?

```
ggplot(data = face_data, aes(x = face_rating, color = incumbent)) +
  geom_density()
```




```
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
```

Yes, the data suggests that there is some advantage incumbents have over non-incumbents in terms of face ratings. Both analyses show this, in fact. This is discussed in the Atkinson et al. reading in the form of something called "selection effects." In this case, this might apply because incumbents may have been originally elected due, in some part, to the positive associations with their face. This may carry over after they serve and become incumbents.

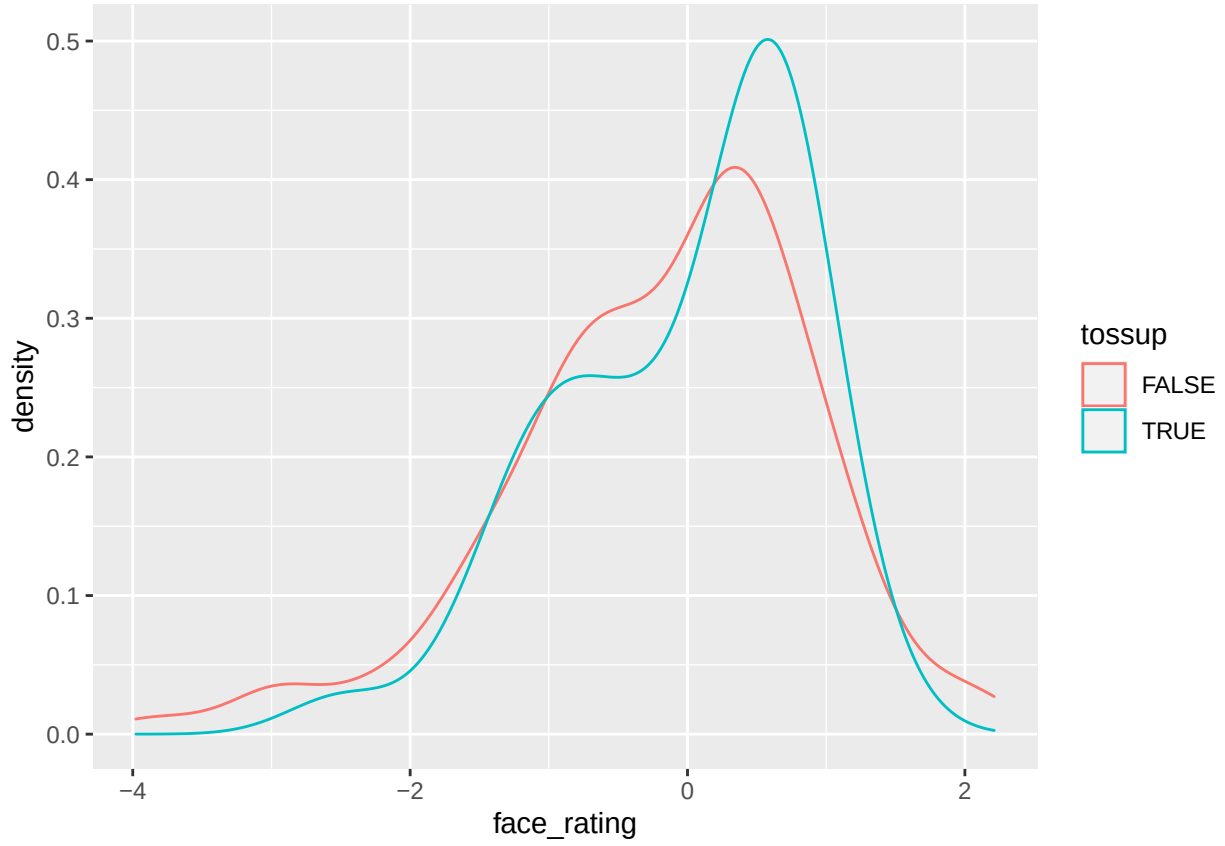
Question 6

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)?

```
noninc_data <- filter(face_data, incumbent == FALSE)
noninc_data
```

```
## # A tibble: 260 x 9
##    cook     year state face_rating incumbent candidate        party tossup   jpg
##    <chr>   <dbl> <chr>       <dbl> <lgl>     <chr>            <chr> <lgl>  <dbl>
##  1 SolidD~  1992 AR          0.214 FALSE     Mike Huckabee    rep   FALSE    446
##  2 Tossup~  1992 CA         -1.37  FALSE     Bruce Herschens~ rep   TRUE     447
##  3 LeanDem  1992 CO         -1.02  FALSE     Ben Nighthorse ~ dem   FALSE    543
##  4 LeanDem  1992 GA          0.319 FALSE     Paul Coverdell   rep   FALSE    548
##  5 SolidD~  1992 HI         -1.99  FALSE     Rick Reed        rep   FALSE    449
##  6 SolidR~  1992 IA         -1.87  FALSE     Jean Lloyd-Jones dem   FALSE    450
##  7 Tossup~  1992 ID         -1.04  FALSE     Richard Stallin~ dem   TRUE     452
##  8 Tossup~  1992 ID          0.600 FALSE     Dirk Kempthorne  rep   TRUE     453
##  9 SolidD~  1992 IL          0.641 FALSE     Carol Moseley B~ dem   FALSE    554
## 10 SolidD~  1992 IL         -0.210 FALSE     Richard S. Will~ rep   FALSE    454
## # ... with 250 more rows
```

```
ggplot(data = noninc_data, aes(x = face_rating, color = tossup)) +
  geom_density()
```




```
difference_in_means(face_rating ~ tossup, data = noninc_data)
```

```
## 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
```

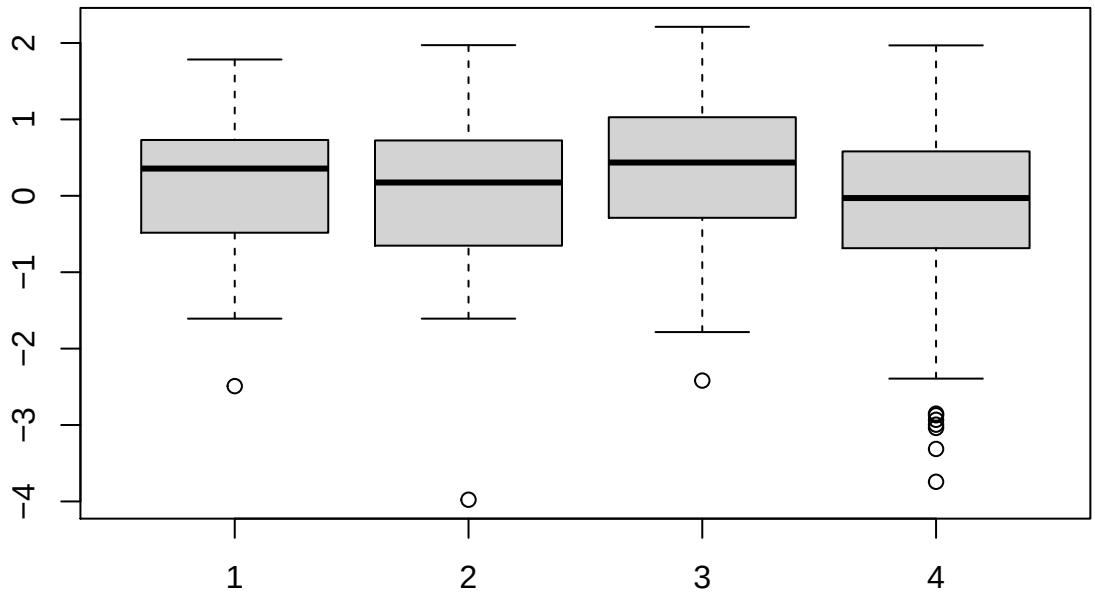
Here, the density plot seems to suggest that, in tossup races, non-incumbents have a bigger advantage in terms of facial rating than non-incumbents in non-tossup races do. However, looking at the t-test, we see that this difference is not statistically significant. This may have to do with how large our sample sizes are. These findings, though, if they were true, are in line with what Atkinson et al. suggests—that facial ratings are more important in tossup 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 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.

```
# split into four dataframes
# Ask about [-4, 4] safety range
lean <- filter(face_data, cook == "LeanRep" | cook == "LeanDem")
```

```
likely <- filter(face_data, cook == "LikelyRep" | cook == "LikelyDem")
solid <- filter(face_data, cook == "SolidRep" | cook == "SolidDem")
tossup <- filter(face_data, cook == "TossupRep" | cook == "TossupDem")
boxplot(tossup$face_rating, lean$face_rating, likely$face_rating, solid$face_rating)
```



There does not seem to be a negative relationship between

```
face_data[face_data == "TossupRep"] <- "1"
assign("TossupDem", -1)
assign("LeanRep", 2)
assign("LeanDem", -2)
assign("LikelyRep", 3)
assign("LikelyDem", -3)
assign("SolidRep", 4)
assign("SolidDem", -4)
assign("dem", -1)
assign("rep", 1)

face_data
```

```
## # A tibble: 444 x 9
##    cook     year state face_rating incumbent candidate        party tossup   jpg
##    <chr>   <dbl> <chr>       <dbl> <lgl>     <chr>            <chr> <lgl>  <dbl>
##  1 LeanRep  1992 AK          1.60  TRUE      Frank H. Murkow~ rep   FALSE    537
##  2 Likely~  1992 AL          1.16  TRUE      Richard C. Shel~ dem   FALSE    105
##  3 SolidD~  1992 AR          1.97  TRUE      Dale Bumpers     dem   FALSE    445
```

```
##  4 SolidD~  1992 AR          0.214 FALSE     Mike Huckabee    rep   FALSE    446
##  5 1        1992 CA         -1.37  FALSE     Bruce Herschens~ rep   TRUE     447
##  6 LeanDem  1992 CO         -1.02  FALSE     Ben Nighthorse ~ dem   FALSE    543
##  7 Likely~  1992 CT         -0.544 TRUE      Christopher J. ~ dem   FALSE    114
##  8 SolidD~  1992 FL          0.563 TRUE      Bob Graham       dem   FALSE    545
##  9 LeanDem  1992 GA          0.170 TRUE      Wyche Fowler     dem   FALSE    448
## 10 LeanDem  1992 GA          0.319 FALSE     Paul Coverdell   rep   FALSE    548
## # ... with 434 more rows
```

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. Note: sitting this question out since I joined the course three weeks late and did not have time to speak to a TF about this while catching up with the class.