Problematic Pupil Prejudice and Paritality Pertaining to Professor Performance

Vincent Robinson, Alex Heller & Luke Knoblich

2023-04-28

```
library(tidyverse)
library(openintro)
library(gridExtra)
```

Introduction

Dataset overview

Our primary data frame entitled "Professor evaluations and beauty", under the variable `evals`, is an observational study taking a sample of 463 classes taught by 94 professors at the University of Texas. It contains data about the professor's evaluation by the students along with some other attributes, such as the professor's beauty rating according to a random sample of students, rank/tenure status, and ethnicity.

The data was taken with a variety of levels of sampling. First, all the data is taken from the University of Texas, which is an example of clustered sampling. For practical reasons, the study could not collect data from more than a single university, decreasing the generalizability of the results to all academic institutions. Some stratification was also done for beauty in specific, as beauty ratings were done specifically by equal numbers of males and females, one of each of lower class level and two of upper class level.

Finally, to select the courses and professors to study, the study eliminated all faculty who did not have photographs on the University of Texas's website since their beauty would be impossible to evaluate. This has a possible bias since certain departments of the university did not have photographs of any of their faculty, and hence entire fields were not included in the data. However, despite that, it effectively results in a form of simple random sampling for selection the courses and professors to study.

Column descriptions

Professor columns

- `prof_id` (discrete numeric):
  - Variable identifying the professor who taught the course (out of 94 professors).
- `score` (continuous numeric):
  - Average professor evaluation score: (1) very unsatisfactory - (5) excellent.
- `rank` (nominal categorical):
  - Rank of professor: `teaching`, `tenure track`, `tenured`.
- `ethnicity` (nominal categorical):
  - Ethnicity of professor: `not minority`, `minority`.
- `gender` (nominal categorical):
  - Gender of professor: `female`, `male`.

- `language` (nominal categorical):
  - Language of school where professor received education: `english` or `non-english`.
- `age` (discrete numeric):
  - Age of professor.

Course columns

- `course_id` (discrete numeric):
  - Variable identifying the course (out of 463 courses).
- `cls_perc_eval`: (continuous numeric):
  - Percent of students in class who completed evaluation.
- `cls_did_eval`: (discrete numeric):
  - Number of students in class who completed evaluation.
- `cls_students` (discrete numeric):
  - Total number of students in class.
- `cls_level` (nominal categorical):
  - Class level: `lower`, `upper`.
- `cls_profs` (nominal categorical):
  - Number of professors teaching sections in course in sample: `single`, `multiple`.
- `cls_credits` (nominal categorical):
  - Number of credits of class: `one credit` (lab, PE, etc.), `multi credit`.

Beauty columns

- `bty_f1lower/bty_f1upper/bty_f2upper` (discrete numeric):
  - Beauty rating of professor from female of lower/upper class level: (1) lowest - (10) highest.
- `bty_m1lower/bty_m1upper/bty_m2upper` (discrete numeric):
  - Beauty rating of professor from male of lower/upper class level: (1) lowest - (10) highest.
- `bty_avg` (continuous numeric):
  - Average beauty rating of professor.
- `pic_outfit` (nominal categorical):
  - Outfit of professor in picture: `not formal`, `formal`.
- `pic_color` (nominal categorical):
  - Color of professor's picture: `color`, `black&white`.

Research Questions

Given the previous data explorations, this project plans to explore the following question:

Do any of the following variables cause a significant positive or negative impact on evaluation score?

a. Gender
b. Ethnicity
c. Tenure status
d. Combinations of the above

Background research

A very large corpus of research has been done on how different attributes affect professor evaluations. Beauty, the primary focus of the study that the data for this project come from, has been widely studied already; for instance, Wolbring & Riordan (2016) and Ponzo & Scoppa (2013) both study the effects of beauty on professor evaluations and conclude that, even when adjusted for other variables, beauty has a not insignificant positive impact on professor ratings. A few studies, such as Campbell et al. (2005), claim that beauty has no effect, but the overwhelming majority of papers seem to agree that the impact is positive.

Similarly, the other variables in this study have also been studied extensively, especially gender bias in evaluations. Nearly all research on this, such as Arrona-Palacios et al. (2020), Mitchell & Martin (2018), and Chávez & Mitchell (2020), conclude that male professors get better evaluations than female professors. Ethnicity studies are also prolific: Chávez & Mitchell (2020) and Bavishi et al. (2010) agree that minority professors get lower ratings on average. Indeed, the data from the initial exploration charts above seem to agree with these conclusions at a cursory glance.

Finally, studies looking at evaluations of tenured faculty are less common, at least those that focus on whether tenure status helps or hurts evaluations scores. However, one study, Murray et al. (2020) seems to indicate that tenured faculty usually have better evaluations, but that this impact is less noticeable than many other variables.

Hypotheses

Given the trends observed in the analysis of the exploration, we make the following hypotheses concerning the entire set of data. For simplicity because the multitude of variables being analyzed simultaneously, we place related hypotheses on a single line. Additionally, we box either the null or alternative hypothesis depending on which we predict to be true.

1. Gender and ethnicity:

   a. Professors who are either male or non-minorities or both do not receive higher scores on average.

   $\boxed{H_0}$ $\mu_{male} = 0$ $\mu_{not\ min} = 0$ $\mu_{fem/not\ min} = 0$ $\mu_{male/not\ min} = 0$ $\mu_{male/min} = 0$
   H_A $\mu_{male} > 0$ $\mu_{not\ min} > 0$ $\mu_{fem/not\ min} > 0$ $\mu_{male/not\ min} > 0$ $\mu_{male/min} > 0$

   b. Females and minorities in general and female minorities in specific receive lower scores on average.

   H_0 $\mu_{female} = 0$ $\mu_{minority} = 0$ $\mu_{fem/min} = 0$
   $\boxed{H_A}$ $\mu_{female} < 0$ $\mu_{minority} < 0$ $\mu_{fem/min} < 0$

2. Teaching:

   a. Teaching professors in general receive higher scores on average.

   H_0 $\mu_{teaching} = 0$
   $\boxed{H_A}$ $\mu_{teaching} > 0$

   b. Teaching professors who are female receive lower scores on average.

   H_0 $\mu_{fem/teach} = 0$
   $\boxed{H_A}$ $\mu_{fem/teach} < 0$

   c. Teaching professors who are not minorities do not receive higher scores on average.

   $\boxed{H_0}$ $\mu_{not\ min/teach} = 0$
   H_A $\mu_{not\ min/teach} > 0$

   d. All other teaching professors receive higher scores on average.

   H_0 $\mu_{male/teach} = 0$ $\mu_{male/min/teach} = 0$ $\mu_{male/not\ min/teach} = 0$
   $\boxed{H_A}$ $\mu_{male/teach} > 0$ $\mu_{male/min/teach} > 0$ $\mu_{male/not\ min/teach} > 0$

3. Tenure track/tenured:

   a. Tenure track/tenured professors in general do not receive lower scores than average.

   $\boxed{H_0}$ $\mu_{ten\ track} = 0$ $\mu_{tenured} = 0$ $\mu_{not\ teach} = 0$
   H_A $\mu_{ten\ track} < 0$ $\mu_{tenured} < 0$ $\mu_{not\ teach} < 0$

   b. Tenure track/tenured professors who are female do not receive lower scores on average.

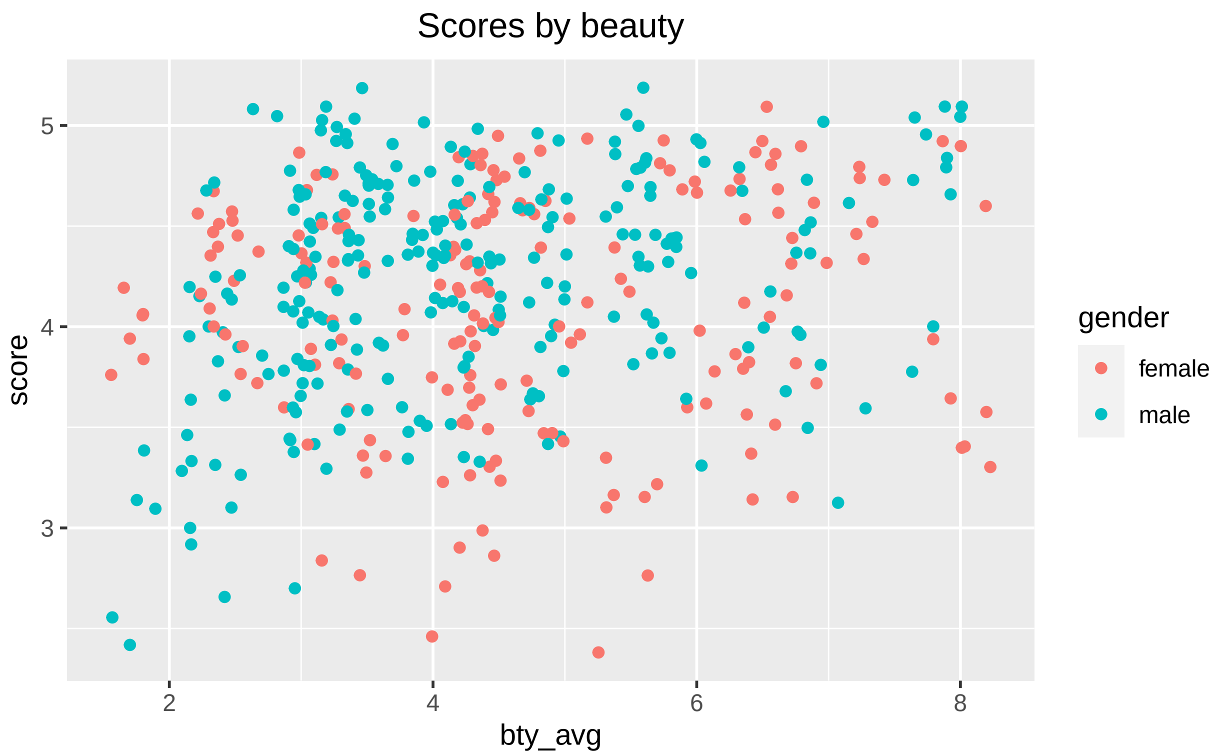
   $\boxed{H_0}$ $\mu_{fem/not\ teach} = 0$
   H_A $\mu_{fem/not\ teach} < 0$

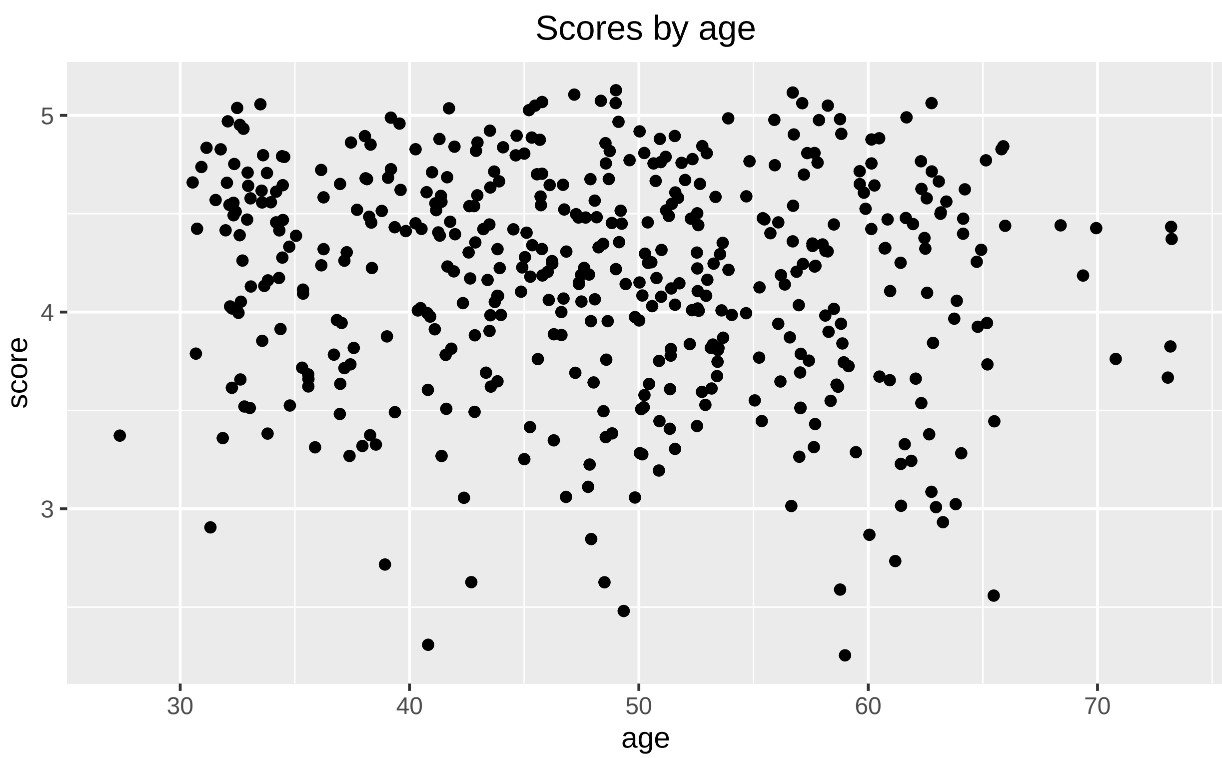
   c. Tenure track/tenured professors who are minorities receive lower scores on average.

   H_0 $\mu_{min/not\ teach} = 0$
   $\boxed{H_A}$ $\mu_{min/not\ teach} < 0$

Exploration

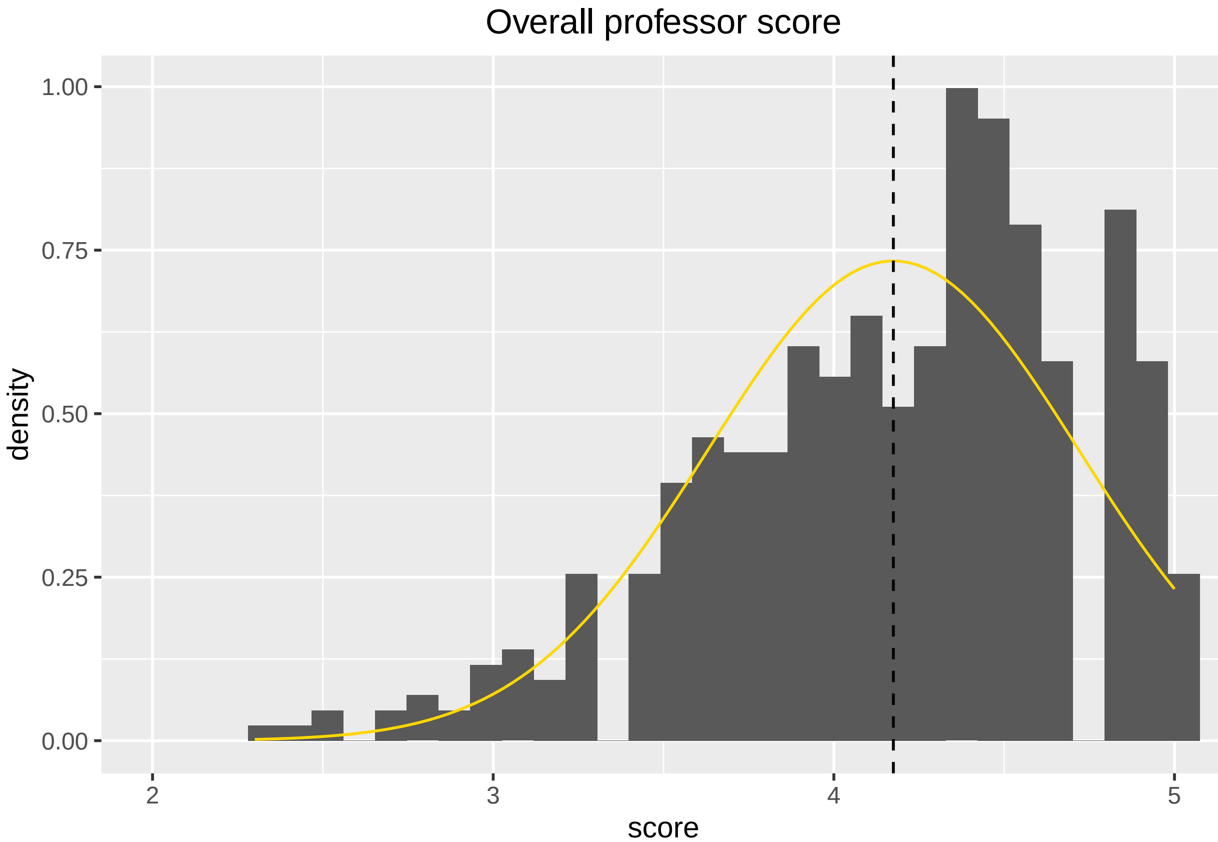
Scatterplots

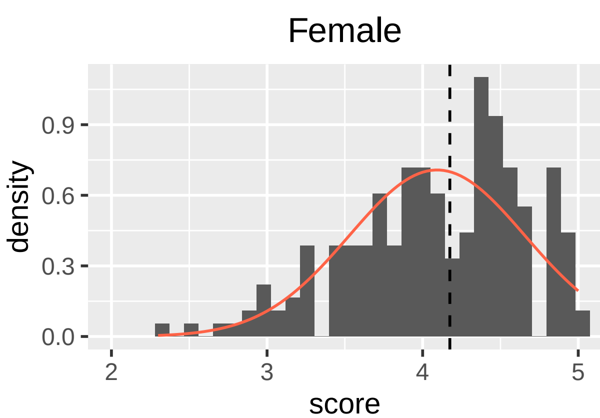


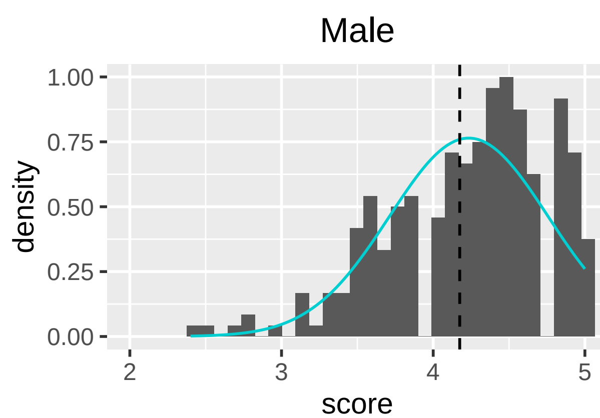


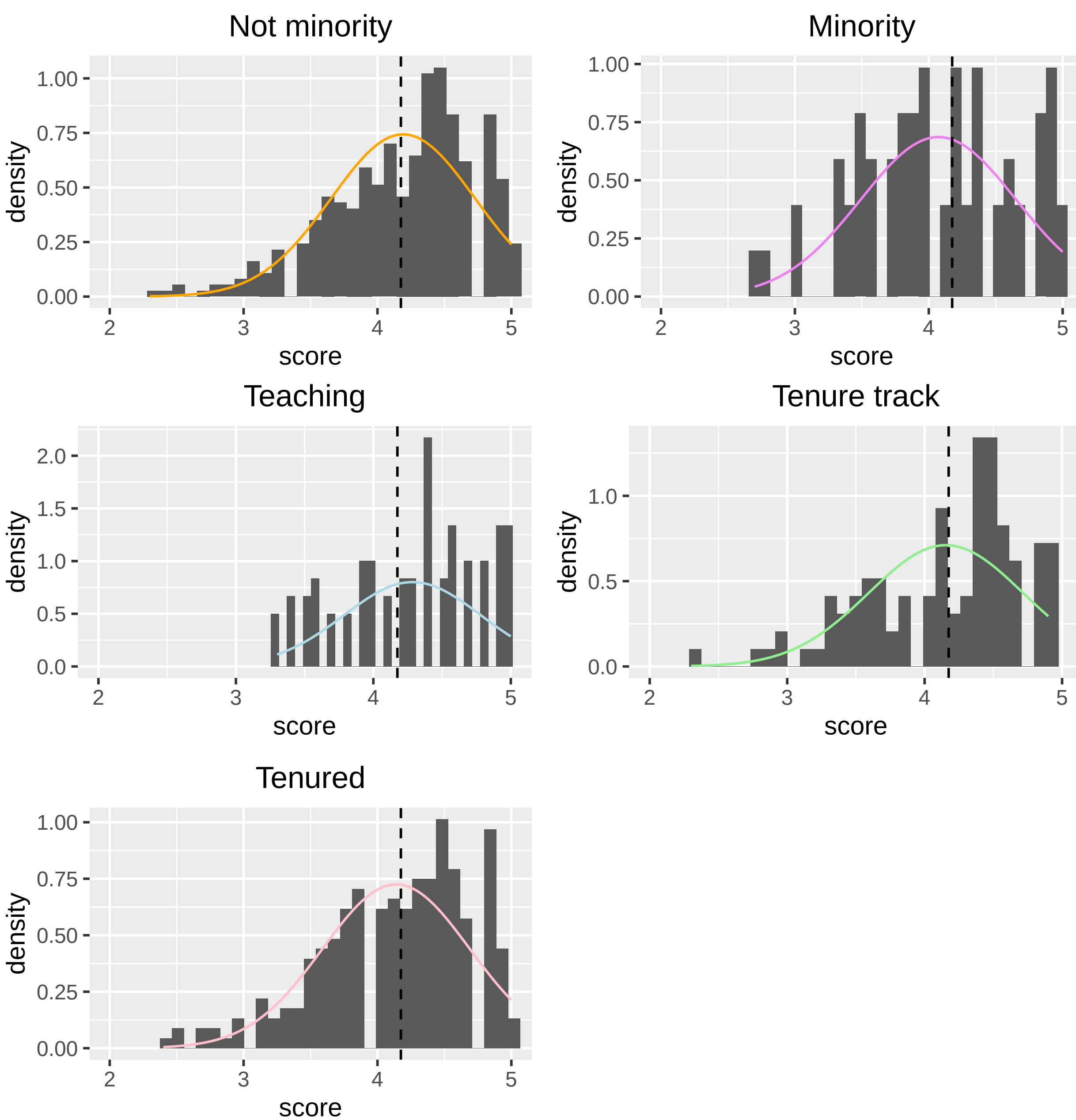
Histograms

All histograms are based on professor score. Each plot has a vertical line showing the mean score of the entire data, giving an easy reference point to see if that particular statistic is above or below average.



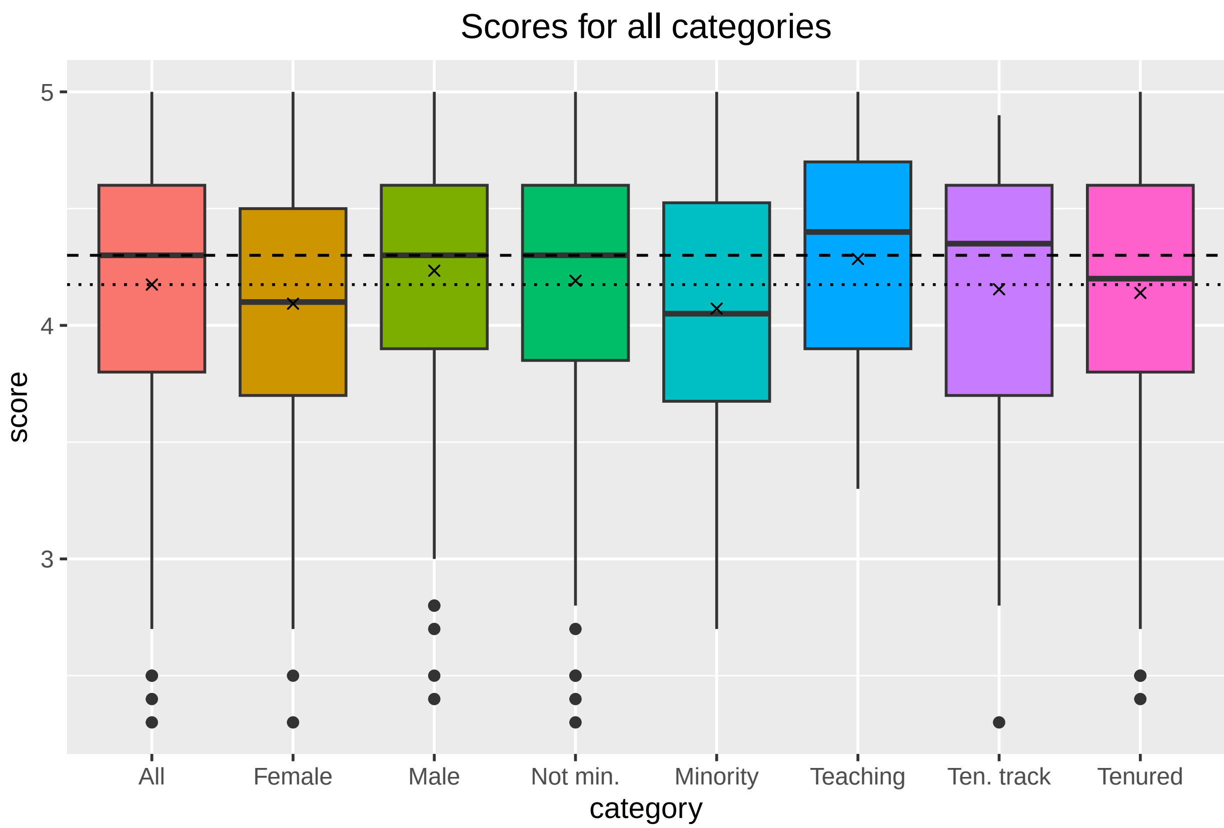


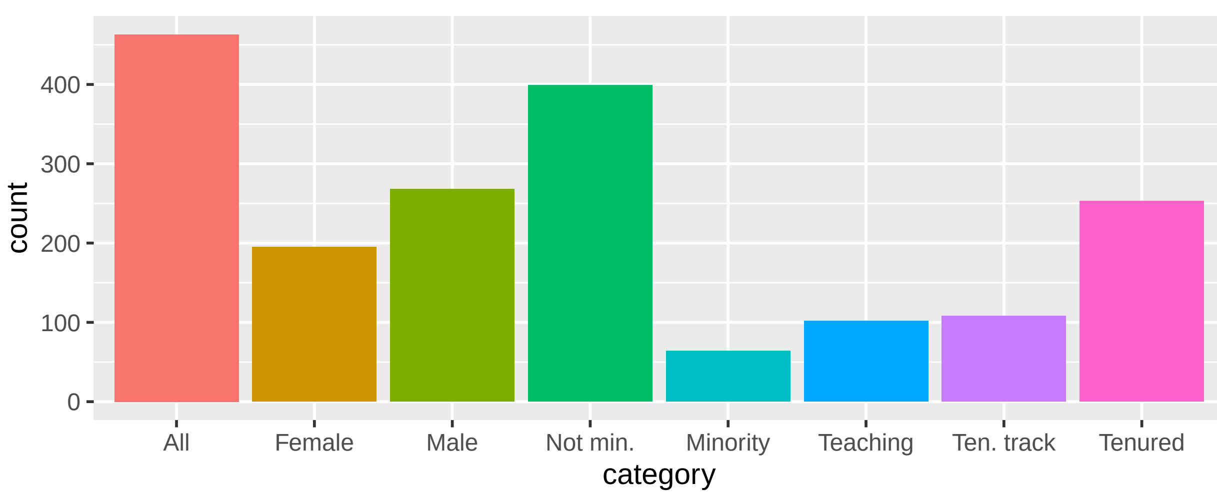


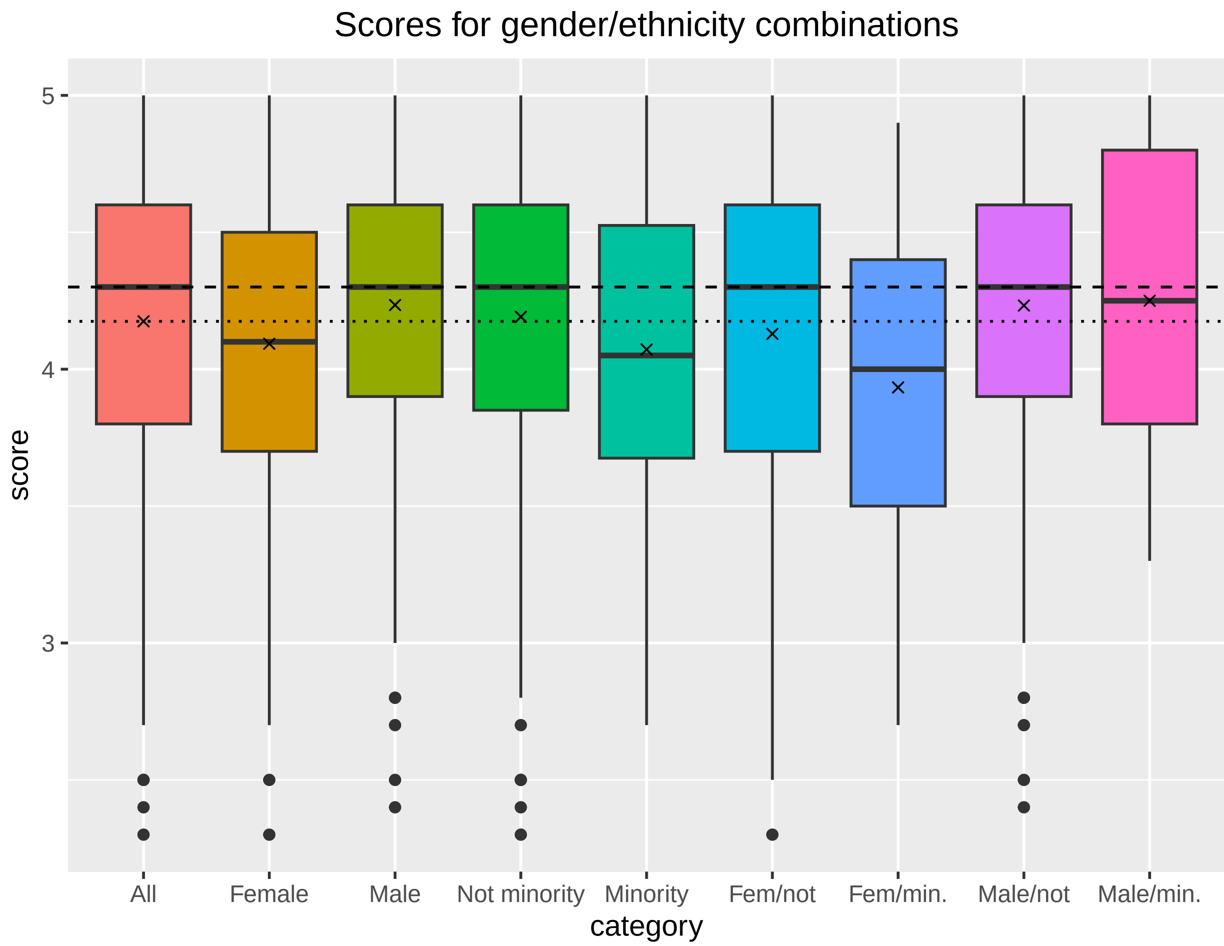

Not minority
Minority
Teaching
Tenure track
Tenured
density
score
1.00
0.75
0.50
0.25
0.00
2.0
1.5
1.0
0.5
0.0
2
3
4
5

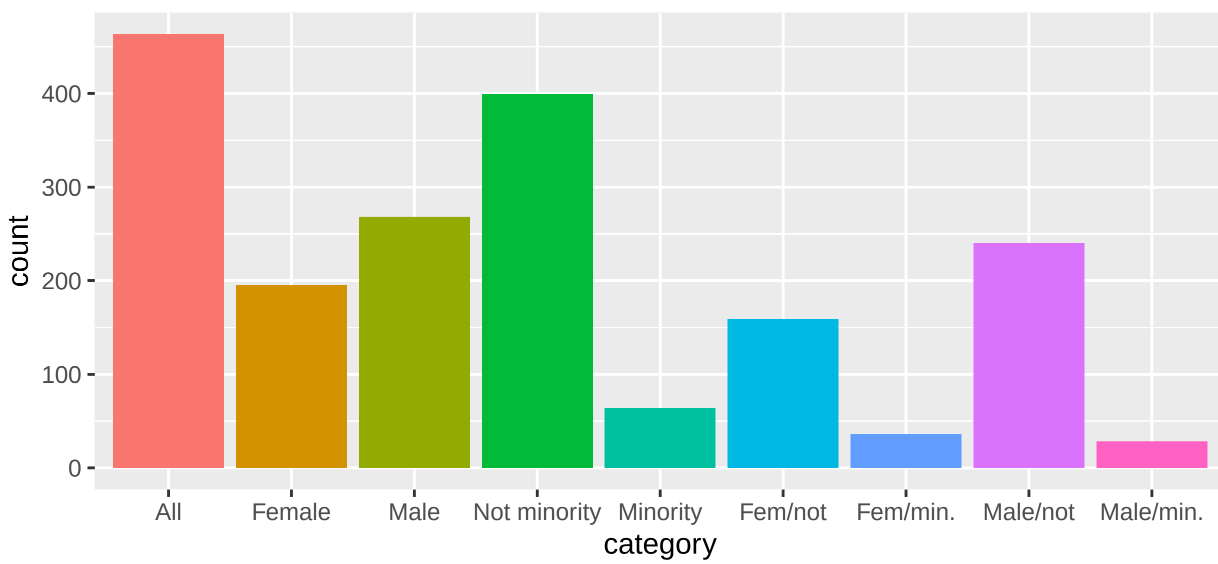
Box and Bar Plots

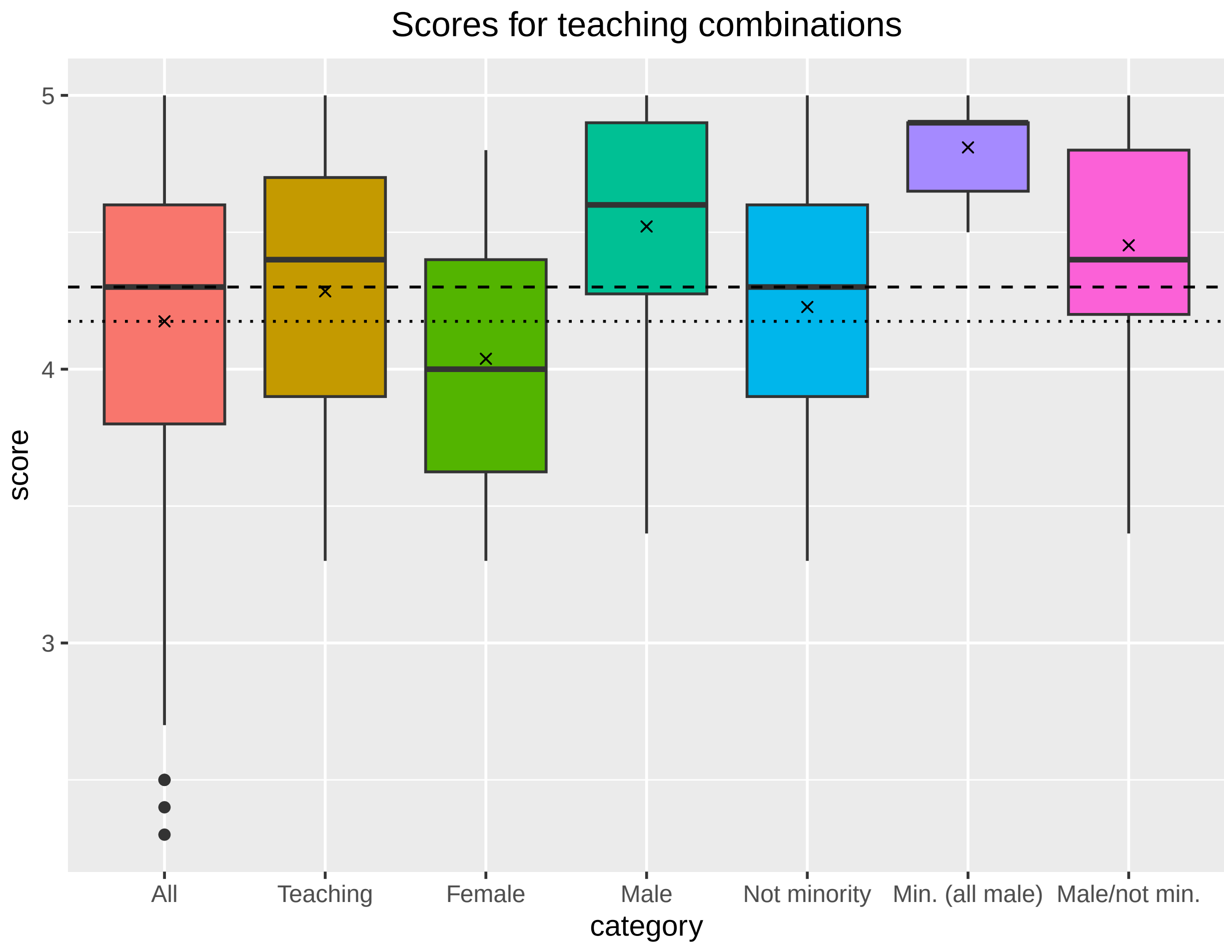
All box plots based on professor score, just like the histograms. They are arranged to make comparisons between groups of data simple, and so show many box plots of refined scope side by side. The dashed line shows the median score of all the professors, and the dotted line shows the mean score. Below each of the box plots is a bar chart showing the number of professors in that group, allowing judgement of whether the results of the above box plot seem significant given the sample size.

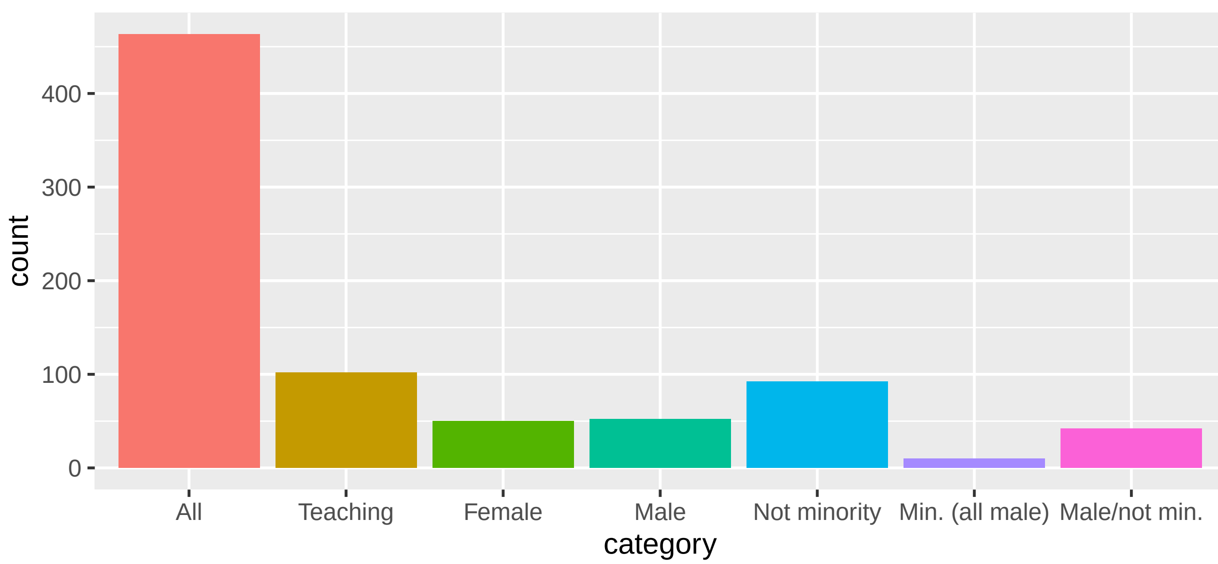


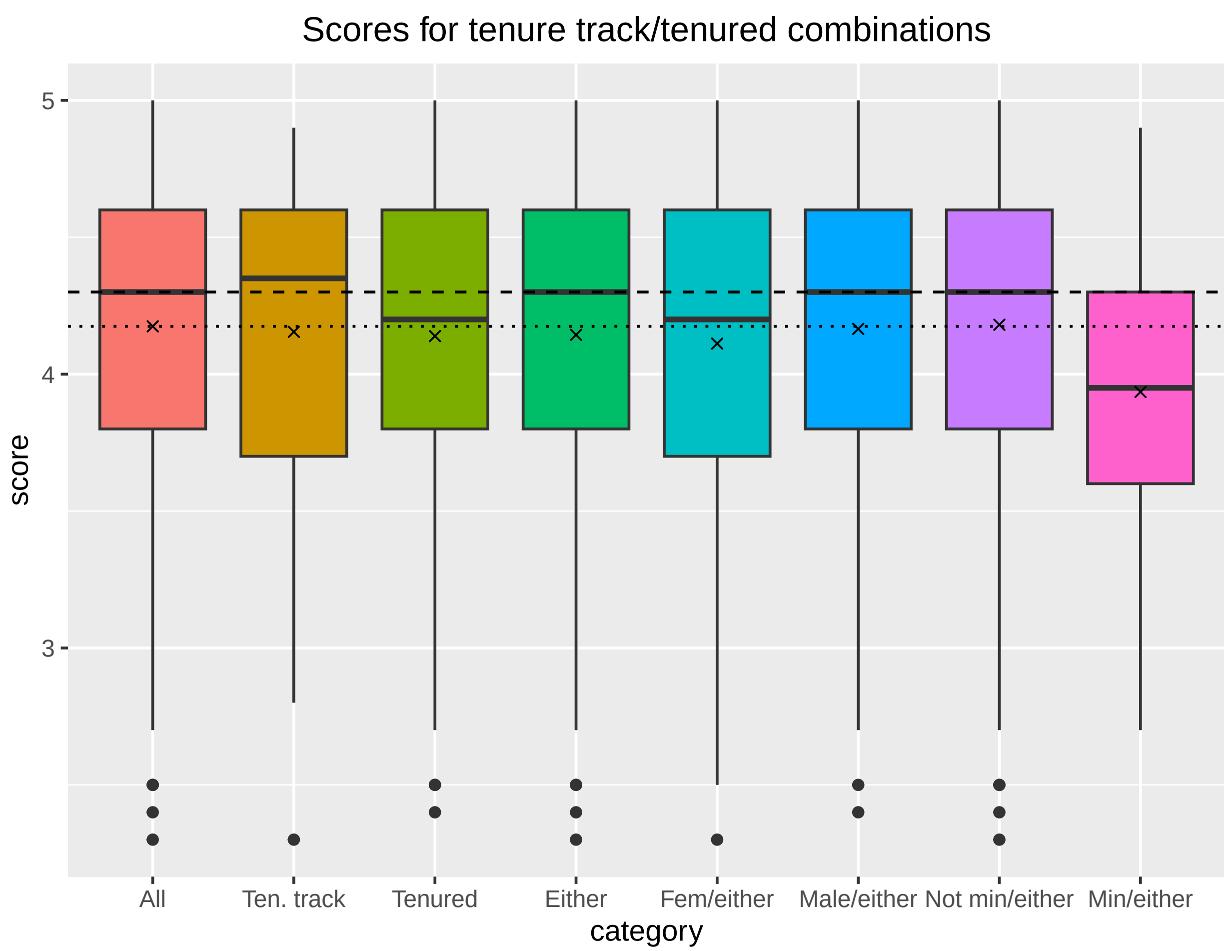


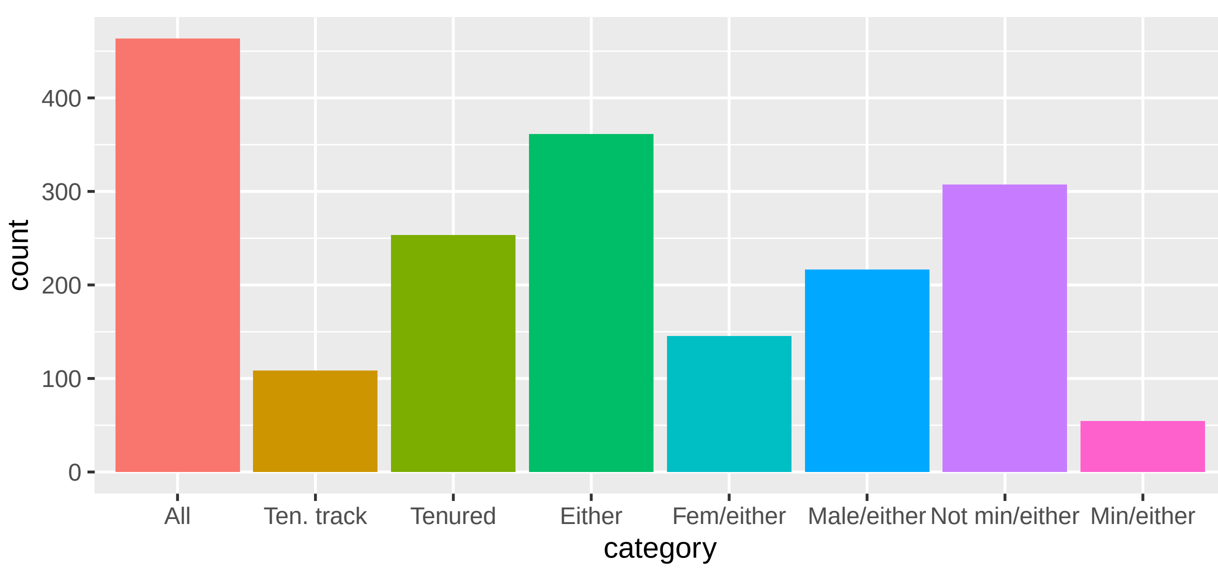

Scores for gender/ethnicity combinations
score
5
4
3
All
Female
Male
Not minority
Minority
Fem/not
Fem/min.
Male/not
Male/min.
category


count
400
300
200
100
0
All
Female
Male
Not minority
Minority
Fem/not
Fem/min.
Male/not
Male/min.
category


Scores for teaching combinations
score
5
4
3
All
Teaching
Female
Male
Not minority
Min. (all male)
Male/not min.
category


count
400
300
200
100
0
All
Teaching
Female
Male
Not minority
Min. (all male)
Male/not min.
category


Scores for tenure track/tenured combinations
score
5
4
3
All
Ten. track
Tenured
Either
Fem/either
Male/either
Not min/either
Min/either
category


count
400
300
200
100
0
All
Ten. track
Tenured
Either
Fem/either
Male/either
Not min/either
Min/either
category

Interpretations

Based on our box plots, a few notable trends can be seen:

- It seems that males and non-minorities are only slightly above average, even when combined with other variables. It is hard to tell if the background research agrees or disagrees with this in particular, since they compare male/non-minority scores directly to female/minority scores, not both to the average score.
- Females and minorities seem below average, but particularly female minorities, who are far below average (as opposed to male minorities, who are slightly above). This is clearly supported by the background research, since females and minorities get lower scores across the board.
- Tenure track/tenured professors all look slightly below average except for minorities of those groups, who are further below average. The background research seems to indicate that the opposite is true, namely that they are slightly above average, but since the research seems uncertain about these statistics, it is not surprising that our data is slightly in the opposite direction.
- Teaching professors appear to be above average, but when combined with other variables, they seem all over the board; however, all are above average except females, who are below. Again, it's hard to say anything about this in context of the background research, but it seems that students may not have a unified opinion on teaching professors.

Results

Throughout this analysis of the data, a significance value of $\alpha = 0.05$ is used.

```
alpha <- 0.05
```

See Appendix B for the list of helper functions used in these analyses.

Assumptions and limitations

As mentioned in the overview of the data, the professors chosen for the study were chosen primarily using simple random sampling, which ensures that we have independent random samples. Additionally, the histograms in the exploration demonstrate that the data is reasonably normal; although some of the smaller groups are less normal, they are close enough for our analysis. To help compensate for this, we use the t distribution exclusively to ensure the accuracy of our analysis. Both of these satisfy the conditions necessary to analyze our data using difference of means.

Gender and ethnicity tests

```
rbind(
  test_diff_means(es, female$score,       "lower") %>% group("Female"),
  test_diff_means(es, male$score,         "upper") %>% group("Male"),
  test_diff_means(es, not_minority$score, "upper") %>% group("Not minority"),
  test_diff_means(es, minority$score,     "lower") %>% group("Minority"),
  test_diff_means(es, female_not$score,   "lower") %>% group("Female/not min."),
  test_diff_means(es, female_min$score,   "lower") %>% group("Female/min."),
  test_diff_means(es, male_not$score,     "upper") %>% group("Male/not min."),
  test_diff_means(es, male_min$score,     "upper") %>% group("Male/min.")
) %>% grid.table()
```

	name	int_low	int_high	p_value	reject
1	Female	−0.17585734	0.01203832	0.043555097	TRUE
2	Male	−0.02050368	0.13970036	0.072060580	FALSE
3	Not minority	−0.05600897	0.08900507	0.327441831	FALSE
4	Minority	−0.25666125	0.05095121	0.093120371	FALSE
5	Female/not min.	−0.14641618	0.05481777	0.185002475	FALSE
6	Female/min.	−0.44119965	−0.04159372	0.009652441	TRUE
7	Male/not min.	−0.02490532	0.14044528	0.084976025	FALSE
8	Male/min.	−0.14626474	0.29680469	0.245836635	FALSE

Rank tests

Teaching groups

```
rbind(
  test_diff_means(es, teaching$score,       "upper") %>% group("Teaching"),
  test_diff_means(es, female_tch$score,     "lower") %>% group("Female"),
  test_diff_means(es, male_tch$score,       "upper") %>% group("Male"),
  test_diff_means(es, not_min_tch$score,    "upper") %>% group("Not minority"),
  test_diff_means(es, min_tch$score,        "upper") %>% group("Minority (all male)"),
  test_diff_means(es, male_not_tch$score,   "upper") %>% group("Male, not minority")
) %>% grid.table()
```

	name	int_low	int_high	p_value	reject
1	Teaching	−0.0003807914	0.219548199	2.539178e−02	TRUE
2	Female	−0.2745473370	0.001087294	2.588208e−02	TRUE
3	Male	0.2179480736	0.474899576	8.373813e−07	TRUE
4	Not minority	−0.0605085273	0.165396310	1.794116e−01	FALSE
5	Minority (all male)	0.4949074847	0.775632472	1.470079e−06	TRUE
6	Male, not minority	0.1320973132	0.423204548	2.016647e−04	TRUE

Tenure track/tenured groups

```
rbind(
  test_diff_means(es, tenure_track$score, "lower") %>% group("Tenure track"),
  test_diff_means(es, tenured$score,      "lower") %>% group("Tenured"),
  test_diff_means(es, not_teaching$score, "lower") %>% group("Either"),
  test_diff_means(es, female_ntch$score,  "lower") %>% group("Female, either"),
  test_diff_means(es, min_ntch$score,     "lower") %>% group("Minority, either")
) %>% grid.table()
```

	name	int_low	int_high	p_value	reject
1	Tenure track	−0.1382814	0.09808060	0.368325892	FALSE
2	Tenured	−0.1199783	0.04877915	0.203407231	FALSE
3	Either	−0.1067545	0.04482912	0.211139320	FALSE
4	Female, either	−0.1731600	0.04714823	0.130059873	FALSE
5	Minority, either	−0.3913376	−0.08775208	0.001283846	TRUE

Conclusions

The results from the gender and ethnicity tests confirmed our first predicted hypothesis that professors who are either male or non-minorities or both do not receive higher scores on average. Our second predicted hypothesis was partly correct in that professors that fall under the categories of female and female/minority receive lower scores on average but the general minority population does not.

All of our predicted hypothesis for the teaching group were correct. These are that teaching professors in general receive higher scores on average, teaching professors who are female receive lower scores on average, and all other teaching professors including all male, male/minority, and male/non-minority receive higher scores on average. We also fail to reject the hypothesis that teaching professors who are not minorities receive higher scores on average meaning there is not sufficient evidence to support the claim.

Our predicted hypothesis for tenure and track/tenured groups were correct once again. We failed to reject the null hypothesis tenure and track/tenured professors in general do not receive lower scores than average, meaning there is not sufficient evidence to support that this group receives lower scores than the average. We also failed to reject the hypothesis that tenure and track/tenured professors who are female do not receive lower scores on average. Once again there is not sufficient evidence to support the statement that female tenure and track/tenured professors receive lower scores on average. Finally, we reject our final null hypothesis in favor of the alternative hypothesis that tenure and track/tenured professors who are minorities receive lower scores on average.

To summarize, female and female/minority professors generally receive lower scores than all groups. The minority population of the tenure and track/tenured group also in general received lower scores on average.

References

Arrona-Palacios, A., Okoye, K., Camacho-Zuñiga, C., Hammout, N., Luttmann-Nakamura, E., Hosseini, S., & Escamilla, J. (2020). Does professors' gender impact how students evaluate their teaching and the recommendations for the best professor? *Heliyon*, *6*(10), e05313.

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education*, *3*(4), 245.

Campbell, H. E., Gerdes, K., & Steiner, S. (2005). What's looks got to do with it? Instructor appearance and student evaluations of teaching. *Journal of Policy Analysis and Management: The Journal of the Association for Public Policy Analysis and Management*, *24*(3), 611–620.

Chávez, K., & Mitchell, K. M. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, *53*(2), 270–274.

Mitchell, K. M., & Martin, J. (2018). Gender bias in student evaluations. *PS: Political Science & Politics*, *51*(3), 648–652.

Murray, D., Boothby, C., Zhao, H., Minik, V., Bérubé, N., Larivière, V., & Sugimoto, C. R. (2020). Exploring the personal and professional factors associated with student evaluations of tenure-track faculty. *PLoS One*, *15*(6), e0233515.

Ponzo, M., & Scoppa, V. (2013). Professors' beauty, ability, and teaching evaluations in italy. *The BE Journal of Economic Analysis & Policy*, *13*(2), 811–835.

Wolbring, T., & Riordan, P. (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Social Science Research*, *57*, 253–272.

Appendix A: Plot generation code

To keep the exploration section clean and easy to follow, we place all code that creates plots in this appendix.

```
# Make plots have centered titles by default
theme_update(plot.title = element_text(hjust = 0.5))

ggplot(evals, aes(x = bty_avg, y = score)) +
  ggtitle("Scores by beauty") +
  geom_jitter(aes(color = gender), width = 0.2, height = 0.2)

ggplot(evals, aes(x = age, y = score)) +
  ggtitle("Scores by age") +
  geom_jitter(width = 2, height = 0.2)

# Make a shortcut to evals$score since we use it so often.
es <- evals$score

# Find important statistics for the entire data so we don't have to recalculate
# them every time we plot something.
evals_mean = mean(es)
evals_median = median(es)
evals_sd = sd(es)

# Function that plots a histogram of a data frame with scores, including a
# title, vertical line for the global mean, and a specific color for the
# histogram curve.
plot_score <- function(data, title, color) {
  args <- c(mean = mean(data$score), sd = sd(data$score))
  return(
    ggplot(data, aes(x = score)) +
      geom_blank() +
      ggtitle(title) +
      geom_histogram(aes(y = after_stat(density)), bins = 30) +
      stat_function(fun = dnorm, args = args, col = color) +
      geom_vline(xintercept = evals_mean, linetype = "dashed") +
      coord_cartesian(xlim = c(2, 5))
  )
}

plot_score(evals, "Overall professor score", "gold")
female <- evals %>%
  filter(gender == "female")
female_plot <- plot_score(female, "Female", "tomato")

male <- evals %>%
  filter(gender == "male")
male_plot <- plot_score(male, "Male", "darkturquoise")

grid.arrange(female_plot, male_plot, nrow = 2, ncol = 2)

not_minority <- evals %>%
  filter(ethnicity == "not minority")
not_minority_plot <- plot_score(not_minority, "Not minority", "orange")
```

```r
minority <- evals %>%
  filter(ethnicity == "minority")
minority_plot <- plot_score(minority, "Minority", "violet")

grid.arrange(not_minority_plot, minority_plot, nrow = 2, ncol = 2)

teaching <- evals %>%
  filter(rank == "teaching")
teaching_plot <- plot_score(teaching, "Teaching", "lightblue")

tenure_track <- evals %>%
  filter(rank == "tenure track")
tenure_track_plot <- plot_score(tenure_track, "Tenure track", "lightgreen")

tenured <- evals %>%
  filter(rank == "tenured")
tenured_plot <- plot_score(tenured, "Tenured", "pink")

grid.arrange(teaching_plot, tenure_track_plot, tenured_plot, nrow = 2, ncol = 2)

# Function that plots combined box and bar plots for scores, including a title,
# colors, points for the mean of each box plot, and lines for the global mean
# and median.
plot_score <- function(data, title) {
  box_plot <- ggplot(data, aes(x = category, y = score)) +
    ggtitle(title) +
    geom_boxplot(aes(fill = category), show.legend = FALSE) +
    geom_hline(yintercept = evals_median, linetype = "dashed") +
    geom_hline(yintercept = evals_mean, linetype = "dotted") +
    stat_summary(fun = mean, geom = "point", shape = 4)

  bar_plot <- ggplot(data, aes(x = category)) +
    geom_bar(aes(fill = category), show.legend = FALSE)

  grid.arrange(box_plot, bar_plot, heights = c(5, 3), ncol = 1)
}

# Function that takes the "category" column of a data frame and converts it from
# a character into a factor without changing the organization.
factor_category <- function(data) {
  data$category <- factor(data$category, levels = unique(data$category))
  return(data)
}

box_category <- rbind(
  evals        %>% add_column(category = "All"),
  female       %>% add_column(category = "Female"),
  male         %>% add_column(category = "Male"),
  not_minority %>% add_column(category = "Not min."),
  minority     %>% add_column(category = "Minority"),
  teaching     %>% add_column(category = "Teaching"),
  tenure_track %>% add_column(category = "Ten. track"),
  tenured      %>% add_column(category = "Tenured")
```

```r
) %>% factor_category()

plot_score(box_category, "Scores for all categories")

female_not <- female %>%
  filter(ethnicity == "not minority")
female_min <- female %>%
  filter(ethnicity == "minority")
male_not <- male %>%
  filter(ethnicity == "not minority")
male_min <- male %>%
  filter(ethnicity == "minority")

box_gender_ethnicity <- rbind(
  evals        %>% add_column(category = "All"),
  female       %>% add_column(category = "Female"),
  male         %>% add_column(category = "Male"),
  not_minority %>% add_column(category = "Not minority"),
  minority     %>% add_column(category = "Minority"),
  female_not   %>% add_column(category = "Fem/not"),
  female_min   %>% add_column(category = "Fem/min."),
  male_not     %>% add_column(category = "Male/not"),
  male_min     %>% add_column(category = "Male/min.")
) %>% factor_category()

plot_score(box_gender_ethnicity, "Scores for gender/ethnicity combinations")

female_tch <- teaching %>%
  filter(gender == "female")
male_tch <- teaching %>%
  filter(gender == "male")
not_min_tch <- teaching %>%
  filter(ethnicity == "not minority")
min_tch <- teaching %>%
  filter(ethnicity == "minority")

# Note: Out of all the teaching female professors, none of them are minorities.
# That makes all the triple teaching statistics redundant except for male
# non-minorities, which is a unique non-empty group.
male_not_tch <- male_tch %>%
  filter(ethnicity == "not minority")

box_teaching <- rbind(
  evals        %>% add_column(category = "All"),
  teaching     %>% add_column(category = "Teaching"),
  female_tch   %>% add_column(category = "Female"),
  male_tch     %>% add_column(category = "Male"),
  not_min_tch  %>% add_column(category = "Not minority"),
  min_tch      %>% add_column(category = "Min. (all male)"),
  male_not_tch %>% add_column(category = "Male/not min.")
) %>% factor_category()

plot_score(box_teaching, "Scores for teaching combinations")
```

```
not_teaching <- evals %>%
  filter(rank != "teaching")
female_ntch <- not_teaching %>%
  filter(gender == "female")
male_ntch <- not_teaching %>%
  filter(gender == "male")
not_min_ntch <- not_teaching %>%
  filter(ethnicity == "not minority")
min_ntch <- not_teaching %>%
  filter(ethnicity == "minority")

box_not_teaching <- rbind(
  evals        %>% add_column(category = "All"),
  tenure_track %>% add_column(category = "Ten. track"),
  tenured      %>% add_column(category = "Tenured"),
  not_teaching %>% add_column(category = "Either"),
  female_ntch  %>% add_column(category = "Fem/either"),
  male_ntch    %>% add_column(category = "Male/either"),
  not_min_ntch %>% add_column(category = "Not min/either"),
  min_ntch     %>% add_column(category = "Min/either")
) %>% factor_category()

plot_score(box_not_teaching, "Scores for tenure track/tenured combinations")
# Function that formats a named list as a tibble with one row to be used as a
# row in a visual table. It will have an extra column "name" at the beginning
# for the name of the row.
group <- function(list, name) {
  return(list %>%
      as_tibble() %>%
      add_column(.before = 1, name = name)
  )
}
```

Appendix B: Helper functions

We define a set of helper functions for finding inferences for a single mean and the difference of two means using the t distribution to make for simpler, less cluttered analyses.

```
# ---------------------------------------------------------------------------
# Takes a vector of data and calculates statistical data together into a single
# convenient list useful for analyzing inferences for single or multiple means.
#
# Params:
#   data  A vector of numbers to calculate the information from.
#
# Returns:
#   A named list with the following fields:
#     $n     The number of items in the data.
#     $mean  The mean of the data.
#     $sd    The standard deviation of the data.
# ---------------------------------------------------------------------------
calc_mean_data <- function(data) {
  return(list(
    data = data,
    n    = length(data),
    mean = mean(data),
    sd   = sd(data)
  ))
}

# ---------------------------------------------------------------------------
# More powerful version of the pt() function that can specify whether to
# calculate for the lower, upper, or both tails of the distribution for less
# than, greater than, or not equal to alternate hypotheses, respectively.
#
# Regardless of whether the t-score happens to be positive or negative, the
# function handles it correctly.
#
# Params:
#   q     The t-score to pass to pt().
#   df    The degrees of freedom to pass to pt().
#   tail  Which tail to compute for. May be "lower", "upper", or "both".
# ---------------------------------------------------------------------------
pt_tail <- function(q, df, tail) {
  if (tail == "lower") {
    # For the lower tail, use pt() verbatim since it calculates that directly.
    p_value <- pt(q, df)
  } else if (tail == "upper") {
    # For the upper tail, subtract the result from one to flip the tail.
    p_value <- 1 - pt(q, df)
  } else if (tail == "both") {
    # For both tails, take the absolute value of the t-score so the sign doesn't
    # mess things up. Since that gives us the upper tail, subtract the result of
    # pt() from one. Then, multiply by two to get both the value for tails.
    p_value <- 2 * (1 - pt(abs(q), df))
  }

  return(p_value)
```

```r
}

# ------------------------------------------------------------------------------
# Calculates a confidence interval and a hypothesis test for an inference for
# one mean given sample information, a null value, and information about which
# tail to compute it for.
#
# Params:
#   test       The sample size, mean, and standard deviation in a named list
#              with the same fields as calc_mean_data().
#   null_mean  The null value for the mean to compare against.
#   tail       The tail to compute for; see pt_tail().
#
# Returns:
#   A named list with the following fields:
#     $int_low   The lower bound of the confidence interval.
#     $int_high  The upper bound of the confidence interval.
#     $p_value   The p-value calculated for the hypothesis test.
#     $reject    Whether the null hypothesis should be rejected based on alpha.
# ------------------------------------------------------------------------------
infer_one_mean <- function(test, null_mean, tail) {
  # First, compute our degrees of freedom for the t distribution.
  df <- test$n - 1

  # Find the standard error and calculate the t-score. Run it through the t
  # distribution to get the p-value.
  std_err <- test$sd / sqrt(test$n)

  t_score <- (test$mean - null_mean) / std_err
  p_value <- pt_tail(t_score, df, tail)

  # Find the t* and the margin of error for the confidence interval.
  t_star <- qt(1 - alpha / 2, df)
  margin_err <- t_star * std_err

  # We now have all the data we need, so return it in our list. We reject the
  # null hypothesis if the p-value is strictly less than the currently defined
  # alpha variable.
  return(list(
    int_low  = test$mean - margin_err,
    int_high = test$mean + margin_err,
    p_value  = p_value,
    reject   = p_value < alpha
  ))
}

# ------------------------------------------------------------------------------
# Same as infer_one_mean(), but takes a vector of data to be passed through
# calc_mean_data() instead of taking the named list of statistics.
# ------------------------------------------------------------------------------
test_one_mean <- function(test_data, null_mean, tail) {
  return(infer_one_mean(
    calc_mean_data(test_data),
```

```
    null_mean,
    tail
  ))
}

# ------------------------------------------------------------------------------
# Calculates a confidence interval and a hypothesis test for an inference for
# the difference of two means given information for both samples and information
# about which tail to compute it for.
#
# Params:
#   ctrl       Statistics for first mean; this contains the sample size, mean,
#              and standard deviation in a named list with the same fields as
#              calc_mean_data().
#   expr       Statistics for the second mean in the same format as 'ctrl'.
#   tail       The tail to compute for; see pt_tail().
#
# Returns:
#   A named list with the same fields as infer_one_mean().
# ------------------------------------------------------------------------------
infer_diff_means <- function(ctrl, expr, tail) {
  # Calculate our degrees of freedom for the t distribution, which we derive
  # from the lesser of the two samples.
  df <- min(ctrl$n, expr$n) - 1

  # Find the standard error and calculate the t-score. Run it through the t
  # distribution to get the p-value.
  std_err <- sqrt((ctrl$sd ^ 2 / ctrl$n) + (expr$sd ^ 2 / expr$n))

  diff_mean <- expr$mean - ctrl$mean
  p_value <- pt_tail(diff_mean / std_err, df, tail)

  # Find the t* and the margin of error for the confidence interval.
  t_star <- qt(1 - alpha / 2, df)
  margin_err <- t_star * std_err

  # We now have all the data we need, so return it.
  return(list(
    int_low  = diff_mean - margin_err,
    int_high = diff_mean + margin_err,
    p_value  = p_value,
    reject   = p_value < alpha
  ))
}

# ------------------------------------------------------------------------------
# Same as infer_diff_means(), but takes two vectors of data to be passed through
# calc_mean_data() instead of taking the named list of statistics.
# ------------------------------------------------------------------------------
test_diff_means <- function(ctrl, expr, tail) {
  return(infer_diff_means(
    calc_mean_data(ctrl),
    calc_mean_data(expr),
```

```
    tail
  ))
}
```